Exploratory Search in Time-Oriented Primary Data

vom Fachbereich Informatik
der Technischen Universität Darmstadt
genehmigte

DISSERTATION

zur Erlangung des akademischen Grades eines
Doktor-Ingenieurs (Dr.-Ing.)
von

Dipl.-Inform. Jürgen Bernard

geboren in Lohr am Main, Deutschland

Referenten der Arbeit:	Prof. Dr. techn. Dieter W. Fellner Technische Universität Darmstadt Prof. Dr. rer. nat. Tobias Schreck Technische Universität Graz
Tag der Einreichung:	08/10/2015
Tag der mündlichen Prüfung:	20/11/2015

Darmstädter Dissertation
D 17
Darmstadt, 2015

Erklärung zur Dissertation

Hiermit versichere ich die vorliegende Dissertation selbständig nur mit den angegebenen Quellen und Hilfsmitteln angefertigt zu haben. Alle Stellen, die aus Quellen entnommen wurden, sind als solche kenntlich gemacht. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Darmstadt, den 08.Oktober 2015 Jürgen Bernard

Abstract

In a variety of research fields, primary data that describes scientific phenomena in an original condition is obtained. *Time-oriented primary data*, in particular, is an indispensable data type, derived from complex measurements depending on time. Today, time-oriented primary data is collected at rates that exceed the domain experts' abilities to seek valuable information undiscovered in the data. It is widely accepted that the magnitudes of uninvestigated data will disclose tremendous knowledge in data-driven research, provided that domain experts are able to gain insight into the data. Domain experts involved in data-driven research urgently require analytical capabilities. In scientific practice, predominant activities are the generation and validation of hypotheses. In analytical terms, these activities are often expressed in confirmatory and exploratory data analysis. Ideally, analytical support would combine the strengths of both types of activities.

Exploratory search (ES) is a concept that seamlessly includes information-seeking behaviors ranging from search to exploration. ES supports domain experts in both gaining an understanding of huge and potentially unknown data collections and the drill-down to relevant subsets, e.g., to validate hypotheses. As such, ES combines predominant tasks of domain experts applied to data-driven research. For the design of useful and usable ES systems (ESS), data scientists have to incorporate different sources of knowledge and technology. Of particular importance is the state-of-the-art in interactive data visualization and data analysis. Research in these factors is at heart of Information Visualization (IV) and Visual Analytics (VA). Approaches in IV and VA provide meaningful visualization and interaction designs, allowing domain experts to perform the information-seeking process in an effective and efficient way. Today, best-practice ESS almost exclusively exist for textual data content, e.g., put into practice in digital libraries to facilitate the reuse of digital documents. For time-oriented primary data, ES mainly remains at a theoretical state.

Motivation and Problem Statement This thesis is motivated by two main assumptions. First, we expect that ES will have a tremendous impact on data-driven research for many research fields. In this thesis, we focus on time-oriented primary data, as a complex and important data type for data-driven research. Second, we assume that research conducted to IV and VA will particularly facilitate ES. For time-oriented primary data, however, novel concepts and techniques are required that enhance the design and the application of ESS. In particular, we observe a *lack of methodological research* in ESS for time-oriented primary data. In addition, the size, the complexity, and the quality of time-oriented primary data hampers the *content-based access*, as well as the design of visual interfaces for gaining an *overview of the data content*. Furthermore, the question arises how ESS can incorporate techniques for *seeking relations between data content and metadata* to foster data-driven research. Overarching challenges for data scientists are to create usable and useful designs, urgently requiring the *involvement of the targeted user group* and support techniques for choosing meaningful *algorithmic models and model parameters*. Throughout this thesis, we will resolve these challenges from conceptual, technical, and systemic perspectives. In turn, domain experts can benefit from novel ESS as a powerful analytical support to conduct data-driven research.

Concepts for Exploratory Search Systems (Chapter 3) We postulate concepts for the ES in time-oriented primary data. Based on a survey of analysis tasks supported in IV and VA research, we present a comprehensive selection of tasks and techniques relevant for search and exploration activities. The assembly guides data scientists in the choice of meaningful techniques presented in IV and VA. Furthermore, we present a reference workflow for the design and the application of ESS for time-oriented primary data. The workflow divides the data processing and transformation process into four steps, and thus divides the complexity of the design space into manageable parts. In addition, the reference workflow describes how users can be involved in the design. The reference workflow is the framework for the technical contributions of this thesis.

Visual-Interactive Preprocessing of Time-Oriented Primary Data (Chapter 4) We present a visual-interactive system that enables users to construct workflows for preprocessing time-oriented primary data. In this way, we introduce a means of providing content-based access. Based on a rich set of preprocessing routines, users can create individual solutions for data cleansing, normalization, segmentation, and other preprocessing tasks. In addition, the system supports the definition of time series descriptors and time series distance measures. Guidance concepts support users in assessing the workflow generalizability, which is important for large data sets. The execution of the workflows transforms time-oriented primary data into feature vectors, which can subsequently be used for downstream search and exploration techniques. We demonstrate the applicability of the system in usage scenarios and case studies.

Content-Based Overviews (Chapter 5) We introduce novel guidelines and techniques for the design of content-based overviews. The three key factors are the creation of meaningful data aggregates, the visual mapping of these aggregates into the visual space, and the view transformation providing layouts of these aggregates in the display space. For each of these steps, we characterize important visualization and interaction design parameters allowing the involvement of users. We introduce guidelines supporting data scientists in choosing meaningful solutions. In addition, we present novel visual-interactive quality assessment techniques enhancing the choice of algorithmic model and model parameters. Finally, we present visual interfaces enabling users to formulate visual queries of the time-oriented data content. In this way, we provide means of combining content-based exploration with content-based search.

Relation Seeking Between Data Content and Metadata (Chapter 6) We present novel visual interfaces enabling domain experts to seek relations between data content and metadata. These interfaces can be integrated into ESS to bridge analytical gaps between the data content and attached metadata. In three different approaches, we focus on different types of relations and define algorithmic support to guide users towards most interesting relations. Furthermore, each of the three approaches comprises individual visualization and interaction designs, enabling users to explore both the data and the relations in an efficient and effective way. We demonstrate the applicability of our interfaces with usage scenarios, each conducted together with domain experts. The results confirm that our techniques are beneficial for seeking relations between data content and metadata, particularly for data-centered research.

Case Studies - Exploratory Search Systems (Chapter 7) In two case studies, we put our concepts and techniques into practice. We present two ESS constructed in design studies with real users, and real ES tasks, and real time-oriented primary data collections. The web-based *VisInfo* ESS is a digital library system facilitating the visual access to time-oriented primary data content. A content-based overview enables users to explore large collections of time series measurements and serves as a baseline for content-based queries by example. In addition, VisInfo provides a visual interface for querying time oriented data content by sketch. A result visualization combines different views of the data content and metadata with faceted search functionality. The *MotionExplorer* ESS supports domain experts in human motion analysis. Two content-based overviews enhance the exploration of large collections of human motion capture data from two perspectives. MotionExplorer provides a search interface, allowing domain experts to query human motion sequences by example. Retrieval results are depicted in a visual-interactive view enabling the exploration of variations of human motions. Field study evaluations performed for both ESS confirm the applicability of the systems in the environment of the involved user groups. The systems yield a significant improvement of both the effectiveness and the efficiency in the day-to-day work of the domain experts. As such, both ESS demonstrate how large collections of time-oriented primary data can be reused to enhance data-centered research.

In essence, our contributions cover the entire time series analysis process starting from accessing raw time-oriented primary data, processing and transforming time series data, to visual-interactive analysis of time series. We present visual search interfaces providing content-based access to time-oriented primary data. In a series of novel exploration-support techniques, we facilitate both gaining an overview of large and complex time-oriented primary data collections and seeking relations between data content and metadata. Throughout this thesis, we introduce VA as a means of designing effective and efficient visual-interactive systems. Our VA techniques empower data scientists to choose appropriate models and model parameters, as well as to involve users in the design. With both principles, we support the design of usable and useful interfaces which can be included into ESS. In this way, our contributions bridge the gap between search systems requiring exploration support and exploratory data analysis systems requiring visual querying capability. In the ESS presented in two case studies, we prove that our techniques and systems support data-driven research in an efficient and effective way.

Zusammenfassung

Primärdaten beschreiben Phänomene in ihrer ursprünglichen Form und unterliegen damit keiner Veränderung oder Manipulation. So darf stets vermutet werden, dass zeitbasierte Primärdaten unerforschtes Wissen birgen, welches insbesondere für die datenzentrierte Forschung von großem Interesse ist. In aufwändigen Projekten werden zeitbasierte Primärdaten erhoben und anschließend persistiert. Die Größe, die Heterogenität, sowie der Zeitbezug zeitbasierter Primärdaten stellt die datenzentrierte Forschung vor große Herausforderungen. Um unerforschtes Wissen abzurufen bedarf es geeigneter Werkzeuge aus den Bereichen der konfirmativen und vor allem der explorativen Datenanalyse. Eine Vision in der Forschungslandschaft ist Wiederverwendung von persistierten Primärdaten. So könnten auch andere Forscher an der datenzentrierten Forschung teilhaben. Insbesondere zeitbasierte Daten sind häufig unwiederbringlich, was deren Wiederverwendung weiter motiviert. Eine der entscheidenden Fragen besteht darin wie Forschern ein intuitiver und effektiver Zugang zu zeitbasierten Primärdaten gewährt werden kann, selbst wenn das Informationsbedürfnis der Forscher zunächst unbestimmt ist.

In dieser Dissertation habe ich es mir zur Aufgabe gemacht die datenzentrierte Forschung bei der Wiederverwendung und der Analyse von zeitbasierten Primärdaten zu unterstützen. Dazu setze ich das Konzept der Explorativen Suche (ES) erstmals für zeitbasierte Primärdaten in die Praxis um. Grundsätzlich repräsentiert die ES die Idee, verschiedene Informationsbedürfnisse des Nutzers in einem System vereint zu unterstützen. Dabei sollen Aktivitäten vom Abrufen von Faktenwissen (Suche) bis hin zur Erkundung völlig neuer Such- und Informationsräume (Exploration) unterstützt werden. Um die ES erstmals für zeitbasierte Primärdaten umzusetzen, bediene ich mich der Techniken der Informationsvisualisuerung und der Visual Analytics. Die Informationsvisualisierung ist die Lehre der visuell-interaktiven Repräsentierung von abstrakten Daten, Visual Analytics erforscht das geeignete Zusammenspiel zwischen automatischer Datenanalyse und visueller Datenexploration. Eine Recherche verwandter Arbeiten ergab insbesondere folgende ungelöste Probleme. Zunächst existierte die ES nur als Konzept, mit der Ausnahme von Systemen für Textdaten. Es fehlte an Strategien, um das Design geeigneter Systeme auch methodisch zu unterstüzen. Der inhaltsbasierte Zugang zu zeitbasierten Primärdaten stellte ein zentrales technisches Problem dar. So war die Suche bisher nur über Metadaten (Daten über Daten) möglich. Zur Unterstützung der explorativen Datenanalyse lag eine Schwierigkeit darin, einen Überblick über große Mengen an zeitbasierten Primärdaten in einem visuellen Suchsystem anzubieten. Des Weiteren bestand ein Defizit in Suchsystemen darin, dass die Identifikation von Zusammenhängen zwischen Zeitseriendaten (dem Datencontent) und Metadaten nicht Teil des analytischen Repertoires war.

In dieser Dissertation beschäftige ich mich mit diesen Herausforderungen und entwickle Methoden, Techniken, und Systeme für die ES in zeitbasierten Primärdaten. Es werden Methoden für das Design von explorativen Suchsystemen für zeitbasierte Primärdaten aufgezeigt (Kapitel 3). Darauf aufbauend stellen die Kapitel 4, 5, und 6 die technischen Schwerpunkte der Disseration dar. Zunächst löst das erste Visual Analytics System für das visuell-interaktive Preprocessing von Zeitseriendaten das Problem des inhaltsbasierten Zugangs zu zeitbasierten Primärdaten. Ein weiteres Kapitel stellt Richtlinien und Techniken für das Design von Überblicksvisualisierungen für Zeitseriendaten vor. Schließlich werden drei neuartige Techniken für die kombinierte Analyse von Datencontent und Metadaten vorgestellt. Die technischen Beiträge dieser Dissertation berücksichtigen explizit die Herausforderung, geeignete algorithmische Modelle in der richtigen Reihenfolge und mit den richtigen Parametern zu wählen. Des Weiteren wird für alle Techniken beschrieben, wie Nutzer in das Design involviert werden können. In Kapitel 7 validiere die Methoden und Techniken anhand zweier explorativer Suchsysteme für zeitbasierter Primärdaten. Mit den Ergebnissen dieser Dissertation leiste ich einen Beitrag zur Wiederverwendung von zeitbasierten Primärdaten, insbesondere zur Unterstützung der datenzentrierten Forschung. Nutzer können durch die Definition von visuell-interaktiven Suchanfragen (query-by-sketch, query-by-example) direkt im Datencontent suchen. Mit visuell-interaktiven Überblicksdarstellungen sind Nutzer zudem in der Lage unbekannte Zusammenhänge im Suchraum zu explorieren und diese für die Wissenserweiterung zu nutzen. Durch die Öffnung des Designprozesses für den Nutzer und die strikt visuelle Art der Datenrepräsentierung leistet diese Dissertation zudem einen Beitrag zum User-centered Design, sowie zur Kommunikation von Information und Wissen aus zeitbasierten Primärdaten.

Acknowledgements

This thesis would not have been possible without the support of mentors, colleagues, students, friends, family, and my girlfriend, to all of whom I am very thankful.

I would like to thank my primary PhD advisor Prof. Dr. Dieter W. Fellner for his confidence in my work and his support throughout the entire process. I want to thank my secondary PhD advisor Prof. Dr. Tobias Schreck, who was also my supervisor in the first two years of my time as a PhD student at the Interactive Graphics Systems Group (GRIS). Even after his call to Konstanz, he was still very supportive, inspiring, and encouraging which led to a variety of great publications and finally to this thesis. I am also very thankful to Prof. Dr. Jörn Kohlhammer, the head of the Information Visualization and Visual Analytics department at Fraunhofer IGD, where I have worked for the last years to finish this thesis. Throughout several projects, he granted me access to different fields related to data-centered research. Together with Tobias Schreck, Jörn Kohlhammer was a true mentor for me. Thus, I had the pleasure to be trained in both basic research and in applied research, which was highly valuable for both my personal development and for this thesis. I would also like to thank Dr. Thorsten May and Prof. Dr. Arjan Kuijper. In their roles as a supervisor and a research coach I could always ask them for feedback. Thorsten May is an excellent reviewer deeply involved into the subject, whereas Arjan Kuijper did a great job in organizational and strategical matters.

I would like to express my sincere thanks to all people involved in my publications. In particular, I would like to name all 30 co-authors of my first-author publications. Dieter Fellner, Tobias Schreck, Jörn Kohlhammer, Tobias Ruppert, Maximilian Scherer, Tatiana von Landesberger, Sebastian Bremm, Thorsten May, Nils Wilhelm, Irina Sens, Jan Brase, Oliver Koepler, Martin Steiger, Oliver Goroll, Sven Widmer, Hendrik Lücke-Tieke, Björn Krüger, David Sessler, James Davey, Marco Hutter, Arjan Kuijper, Debora Daberkow, Mila Runnwerth, Katrin Fischer, Daniel Keim, Sebastian Mittelstädt, Michael Behrisch, Simon Thum, Thorsten Schlomm, and Dirk Pehrke.

I would also like to thank the administrative teams and secretariats at GRIS and IGD namely Carola Eichel, Silke Romero, Nils Balke, Gabriele Knöß, and Patricia Häg, who contributed to a productive and friendly working environment. A special thank-you goes to my colleagues and trainees with whom I had a great time at work and beyond. In particular, I would like to thank Sven Widmer for his continuous feedback as an external but highly-experienced discussion partner. I thank Tobias Ruppert, Martin Steiger, and Nils Wilhelm for loads of intensive discussions and close collaboration leading to dozens of publications. My thank goes to Marco Hutter, Andreas Bannach, Martin Steiger, Sebastian Maier, Hendrik Lücke-Tieke, Alex Ulmer, and David Sessler for all the fruitful endeavors for technical improvement. Once more, I would like to thank Noel Stanton and Carina Fath, who volunteered to review this thesis with a focus on grammar, spelling, and native-speaking. You did a great job, particularly since this thesis became comparatively long.

I thank my friends for their patience throughout the writing process. I am looking forward to spend more time with you in future again! Finally, I would like to thank my entire, beloved family for their support in hard situations and for being proud of me no matter what. My special thanks goes to my mother Leonore Bernard, my father Rudolf Bernard, and my sister Nadine Bernard. I would like to conclude my acknowledgments with my girlfriend Carina Fath. Carina, we went through a tough time and you have always been so patient and supportive! I'm looking forward to share the rest of my life with you, I love you!

Jürgen Bernard
November 2015

Contents

CHAPTER 1

Introduction

Contents

To this day, mankind has seen different paradigms of scientific discovery. In ancient times, the main paradigm was *experimental science* which described natural phenomena. In medieval times, the predominant paradigm was *theoretical science* in which scientists sought models and generalizations. Beginning with the age of computers, complex models and phenomena are calculated and simulated with *computational science*.

Today, data has become scientific capital. We experience the fourth paradigm of science, the era of *data-driven research* involving sensors, data storage, data processing, and data exploration [HTT09]. Data-driven research is also referred to as "the use of massive data sets to find patterns as the basis of research" [The12]. Data-driven research fits well in these days since data is gathered at continuously increasing speed. New technologies for collecting valuable primary data led to vast repositories and data warehouses which are, to some extent, able to manage today's huge data collections. However, the pure quantity of primary data exceeds scientists' abilities to analyze data; and to gain valuable insight. Thus, the majority of the stored primary data still remains unexploited. This information overload problem (also called data deluge) is still one of the most challenging research questions for data analysis in general. The analysis and exploration of primary data urgently requires new sophisticated solutions. In this thesis, **we**[1] present concepts, guidelines, techniques and real-world systems supporting domain experts in data-driven research. For this purpose, we employ *Exploratory Search* as a guiding concept to approach the value of time-oriented primary data.

Three key requirements for effective and efficient data-driven research include (a) strategies to make these complex data types usable, (b) visual-interactive interfaces using supportive analytical technologies and (c) collaborations between domain experts and data scientists. Data scientists are specialists in the analysis of data and the design of sophisticated analytical tools, using, e.g., *Information Visualization* (hereafter, IV) and *Visual Analytics* (hereafter, VA) techniques. Domain experts are specialists in application domains and have tremendous domain knowledge, which they can employ to carry out data-driven research. The golden mean would combine the strengths of both as collaborators in a meaningful way, to yield powerful analytical tools for data-driven research.

Two main activities in the scientific practice of domain experts are the *exploration* of large data collections and the *search* for relevant sub-collections. These two activities are closely related to the process of hypotheses formulation and hypotheses testing, to exploratory analysis and confirmatory analysis, as well as to browsing and querying in large data collections. *Exploratory search* (hereafter, ES) is a concept combining the complementary strengths of search and exploration activity. In this connection, ES is promising to support data-driven research. Domain experts can greatly benefit from efficient and effective *ES systems* (hereafter, ESS) supporting their scientific practice. For the

[1] Throughout this thesis I use the *we*-form as a tribute to all the co-authors who enriched the body of my publications.

(a) *Data-driven research in Earth observation. At the Neumayer Station located in Antarctica, weather phenomena are measured with different sensors. For more than 30 years, different stations all over the world have provided valuable time-oriented primary data for reuse.*

(b) *Recording of human motion. Domain experts are interested in variations of different actors performing different motions with multiple repetitions.*

Figure 1.1 *Research domains where time-oriented primary data is measured, processed, and subsequently stored for scientific reuse. Important research goals drawing on time-oriented primary data are based on the identification of structural information, such as frequent patterns, periodic behaviors, outliers, or trends. Similarly, domain experts are interested in the lookup of previously known phenomena, e.g., to validate hypotheses. The left figure shows Earth observation measurements which we used in the first case study of this thesis (see Section 7.1). On the right, the recording of human motion capture data is shown as used in the second case study (see Section 7.2).*

design of ESS, design study methodology from IV and VA can be applied to foster the collaboration between data scientists and domain experts. For the design of useful IV and VA solutions, it is most important to characterize (a) the users (domain experts), (b) the user tasks (exploratory search), and (c) the involved data. Throughout this thesis, we focus on time-oriented primary data, a data type with special characteristics which are highly appropriate for capturing scientific phenomena depending on time. We present concepts, guidelines, techniques, and example systems showing how domain experts can be supported in the ES in time-oriented primary data.

1.1. The Value of Primary Data

Many research projects include the collection of data, either to answer direct research questions, or to amplify contextual information about complex phenomena. In data-driven research, this data is often referred to as primary data. Primary data can be characterized as a *direct* product from a source. In contrast to secondary data, primary data comprises the original condition of a phenomenon without being processed, transformed, or manipulated into other forms. The unaffected nature of primary data makes it particularly valuable for data-driven research. In general, it can be assumed that primary data contains *undiscovered knowledge*. Common sources for the collection of primary data are interviews, experiments, observations, simulations, or other types of first-hand experiences. Primary data is collected in a variety of application areas, such as Earth and environmental science, physics, medicine, biology, or social sciences. In some cases, the terms raw data, sensor data, measurement data, scientific data, or research data are used interchangeably, all having a strong association with the notion of primary data. An illustration of two different sources of primary data is presented in Figure 1.1 using the examples of Earth observation and human motion analysis.

Primary data is further augmented with explanatory *metadata* (data about data) occurring in scientific practice, e.g., to further characterize experiment conditions. Thus, data provides two bodies of valuable information; the *data content* and *metadata*. This multi-modal nature of primary data yields undiscovered *relations* between data content

and metadata, which additionally enhance the value of primary data. Domain experts spend much time in seeking relations between data content and metadata. Sophisticated relation-seeking techniques, such as presented in IV and VA, can enhance the relation-seeking process tremendously.

Two of the widely applied working methods in data-driven research are (1) the formulation of hypotheses and (2) the validation of hypotheses. The formulation of hypotheses is an exploratory analysis process where domain experts have to browse through large collections of primary data, e.g., to reveal structural information and patterns. The data exploration process enables domain experts to formulate new hypotheses. The validation of hypotheses is a confirmatory analysis process where domain experts first seek to find appropriate data subsets, followed by data processing and statistical testing. For both exploring large primary data collections and searching for interesting data subsets, domain experts are reliant on sophisticated computational support.

The two conducted research projects presented in Figure 1.1 have in common that the domain experts are interested in collecting and analyzing primary data. In addition, both projects make primary data collections publicly available for scientific reuse. Primary data is typically passed through different steps of the so-called *life-cycle* to exploit its value more effectively. At first, primary data is *created* and augmented with explanatory metadata, e.g, in specialized research projects. Next, primary data is *processed* and curated for data management. In the subsequent *analysis* phase, the primary data is interpreted, published, and prepared for *preservation*, i.e., to avoid accidental loss of data. In the *access* phase, the data is distributed, shared and promoted. Finally, in the *reuse* phase, primary data may be subject to reanalysis and validation approaches, or to follow-up research. In the course of the life-cycle of primary data, different stakeholders contribute to the data life-cycle. Among others, data authors, domain experts, data curators, database managers, data scientists, and digital librarians interact with primary data.

> *Ideally, the scientist should be able to plug-in almost any scientific data resource and computational service into a scientific workflow, inspect and visualize data on the fly as it is computed, make parameter changes when necessary and re-run only the affected downstream components, and capture sufficient metadata (...).*
>
> Ludäscher et al. [LAB*06], *2006*

For the *analysis* and exploration of primary data, scientists rely on computer-supported tools enabling scalable data processing. New scientific knowledge is often gained by domain experts putting together *pipelines* for the analysis of data. We echo the quote of Ludäscher et al. [LAB*06] which nicely outlines the potential effectiveness and the efficiency of future analysis capability. Domain experts are starting to recognize the value of analytical methods that go beyond statistics. The technology in data-driven research has advanced from simple batch executions to complex *scientific workflows* [DGST09] where different computer-supported analysis steps are put together into a workflow. Domain experts urgently require support in the visual-interactive construction of such analytical pipelines to leave tedious batch processing behind. Similarly, a massive use of visualization may enhance the domain experts' data analysis capability. This applies to the visual representation of end products of the analytical process, but particularly to the visualization of the analytical process and its intermediate results. In this connection, scientific workflows can greatly benefit from IV, VA, and other research fields related to exploratory data analysis. Likewise, the role of data scientists becomes increasingly important to support data-driven research.

The effective *access* and *reuse* of primary data is an overall goal for data-centered research [HTT09, The12]. In this way, not only research projects producing primary data themselves, but also the remaining research community consuming primary data, can benefit from the growing magnitudes of data. Benefits are an enhanced accessibility for sharing, the validation of findings leading to an increased transparency, the reduction of costs for data recreation, implications facilitating science as an open enterprise, and building on the works of others. Many scientific observations and experiments cannot be repeated at all, particularly, if the primary data measurements are dependent on absolute time. To support the access and the reuse of primary data, domain experts require additional infrastructures. As an example, *Digital Libraries* (hereafter, DL, DLs) facilitate the collection, the storage, and the retrieval of digital documents, and thus play an important role in data-driven research. To support domain experts in the identification of relevant data subsets, DLs require meaningful techniques to facilitate the search for primary data.

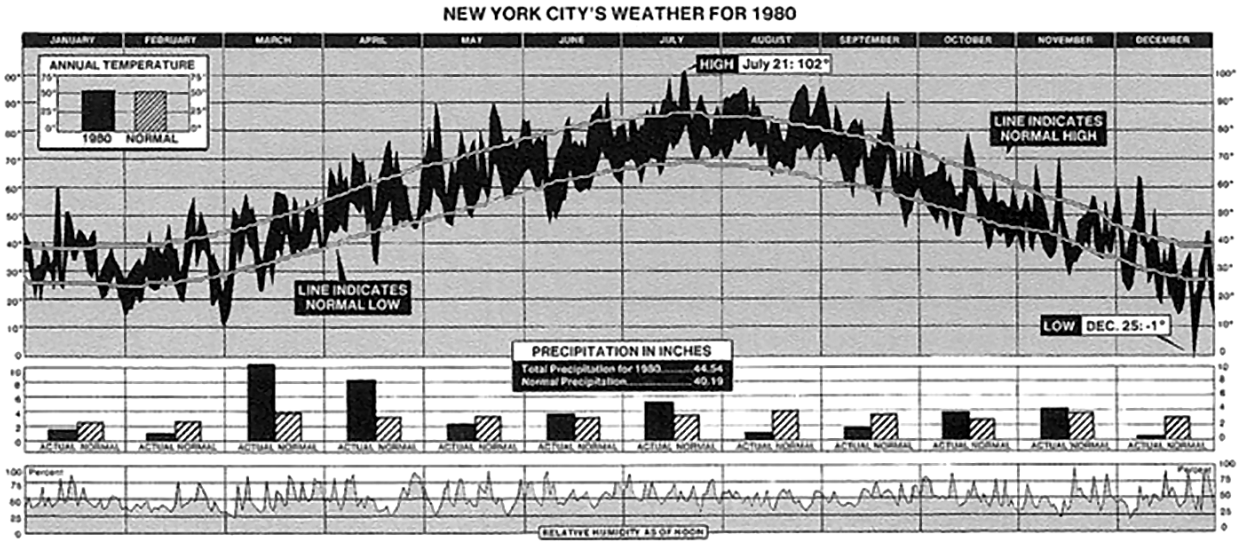


Figure 1.2 *New York City's weather in 1980 including progressions of temperature, precipitation and humidity. Popular example of historical time series visualization. New York Times, January 11, 1981, p.32, adapted in the book of Aigner et al. [AMST11] (used with permission).*

The value of primary data coexists with a variety of data-centered challenges distracting domain experts from their core research goals. Similar to other data types, primary data collections pose different types of complexity. A specific challenge particularly related to the complexity of primary data regards the *quality*. In contrast to secondary data which may be a processed (condensed, cleaned, distilled) derivate with a certain extent of quality management, primary data is a raw type containing potentially many different (unexpected) quality leaks. Thus, primary data has to be cleansed and be manipulated into a usable form. A further associated challenge is the *size* of today's primary data collections. Effective data abstraction and data aggregation techniques are required to provide overviews and summaries of large data collections. In addition, the *heterogeneity* of primary data is a challenge for many data-driven research approaches. Primary data may include different types, like spatio-temporal, multivariate, and multi-modal data [KH13]. The heterogeneity of primary data is increased by different data formats, missing data standards or data not matching existing standards. Furthermore, the phenomena described by primary data may be extremely complex and require significant time and effort to be explored and understood [PVW09]. Another factor contributing to an enhanced heterogeneity of primary data is the explanatory *metadata*. Sophisticated techniques are needed to exploit the complete body of knowledge including both data content and metadata.

1.2. Time-Oriented Data

In this thesis, we focus on primary data with a specific type of data content, i.e., time-oriented data. Time has always played an important role in scientific discovery. In *empirical sciences* many natural phenomena had a dependency on time, such as observations during entire days or seasonal effects within years. In *theoretical sciences*, the first visualizations of time-oriented observations were provided, such as illustrations of planetary orbits. Many complex phenomena simulated in *computational sciences* have a dependency on time, e.g., simulations on flow, thermal behavior, or societal development. In the age of *data-driven science*, the role of time-oriented data may even exceed its importance for earlier paradigms. Primary data depending on time occurs in virtually every scientific discipline, e.g., in the assessment of climatic trends, in the observation of movement, or in recoding the development of patient well-being. Simultaneously to the scientific use of time-oriented data, a great variety of artworks illustrating time-dependent data have been developed. Examples are the visualizations of military campaigns, histories of music genre development, or charts of train schedules (see, e.g., the surveys of Tufte [Tuf90] and Bertin [Ber83]). In Figure 1.2, we refer to the visualization 'New York City's weather in 1980', as it was re-illustrated by Aigner et al. [AMST11].

The unique characteristic of time-oriented data is the existence of a *value domain* and an additional *temporal domain*. Variables stored in the value domain are dependent on the time. Both domains have intrinsic characteristics that require the treatment of time-oriented data as a special data type. Time-oriented data is appropriate to capture complex

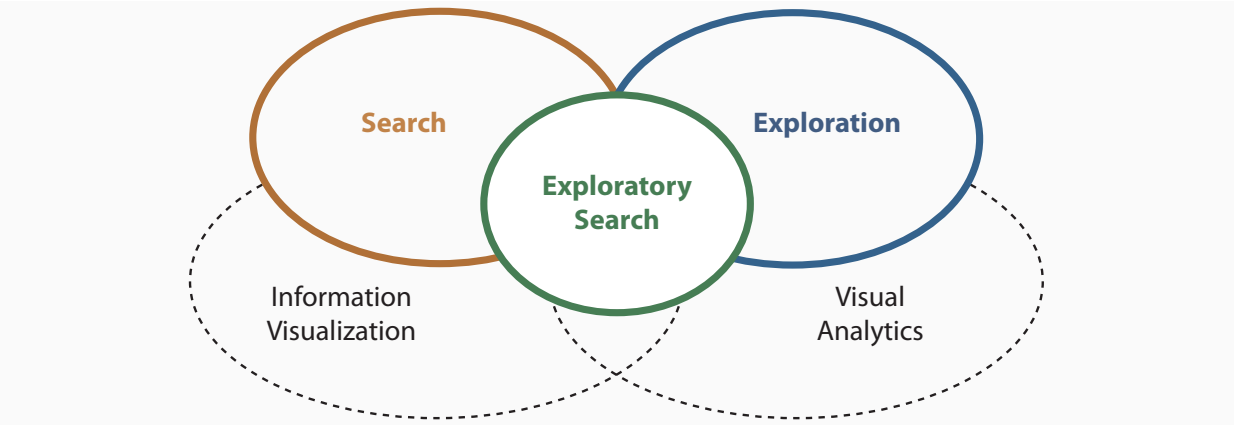


Figure 1.3 *Positioning ES. Obviously, ES benefits from advances made in search and exploration support. The research fields of IV and VA contribute to these developments, and thus are also highly relevant to facilitate ES.*

phenomena including periodic and cyclic behavior, seasonal effects, trends, noise, frequent patterns, outliers, events, intervals, durations, and many others (see, e.g., the survey of Aigner et al. [AMST11]). However, the specific treatment of time-oriented data poses challenges in modeling data structures for time-oriented data, processing and transforming time-oriented data, and designing effective visualizations and interactions for time-oriented data. All these factors are subject to ongoing research in associated basic research domains, such as *Information Retrieval* (hereafter, IR), *Data Mining* (hereafter, DM), IV, and VA. A particular research question for this thesis regards the challenge of supporting domain experts in their data-driven research activities with meaningful time series analysis applications. In this connection, the unique characteristics of *time-oriented primary data* pose additional challenges. Essentially, the quality of primary data affects the quality of the data content. These quality issues have to be resolved with specific data cleansing techniques, to support domain experts using time-oriented primary data. Yet another characteristics of time-oriented data is its tendency to specific analysis tasks. Domain experts may want to *localize* known values in the temporal domain (what -> when?), *identify* values for a given temporal domain (when -> what?), or *explore* both the value and the temporal domain to reveal unknown patterns and relations (what? / when?) [AA06, BM13]. Especially for explorers the determination of the temporal domain as the *dependent variable* for the validation of hypotheses may be insufficient. A particular challenge in supporting exploratory data analysis tasks involves questions concerning the dependent variable. Domain experts may want to adapt the dependent variable in the course of the analysis, or may even want to formulate new hypotheses without a prior determination of dependent variables at all.

1.3. Exploratory Search in Time-Oriented Primary Data

At a core level, the effectiveness and the efficiency of data-driven research in time-oriented primary data depends on at least two key activities, which have to be supported by sophisticated tools. One key activity regards the retrieval of relevant subsets in large collections of time-oriented primary data, e.g., in the *access* and *reuse* step of the data life-cycle. Retrieving relevant data subsets leads to an enhanced hypotheses testing process, in combination with downstream data analysis and statistical testing. The other key activity considers gaining an understanding of the structures and patterns of large document collections at a glance, e.g., in the *analysis* step of the life-cycle of primary data. This exploratory activity has to be supported with meaningful solutions. As a result, the process of hypotheses formulation can be enhanced. The activities of searchers and explorers reflect two salient information-seeking behaviors of humans in the process of acquiring knowledge [Mar95, Shn96, Mar06, WR09].

Search The retrieval of relevant data subsets assumes that at least some information of the targeted subset is already known. For time-oriented data, knowing both the temporal and the value domain is referred to a *lookup* task. Finding the point in time or points in time when a given object occurred, is typically referred to as a *localization* task. Finding the object or set of objects at a given point in time is described as an *identification* task [AA06, AMST11]. The effectiveness of systems supporting these tasks depends on two assumptions. First, the data collection requires the

existence of relevant objects for the seeker. Second, the seeker must be enabled to formulate well-formed queries, which, in turn lead to relevant results through the retrieval system. The predominant retrieval paradigm of this activity is described as "query and response" [WR09], based on the alternating participation of users formulating queries and waiting for response of the system. This type of information seeking is referred to as known-item search, lookup, or fact retrieval [Shn96] [Mar06, p. 29 ff.]. In the course of this thesis, we refer to this type of activity as *search*.

> *Everything on Earth can be found, if only you do not let yourself be put off searching.*
>
> Philemon of Syracuse, *(c. 362 BC - c. 262 BC)*

Exploration The need for gaining an understanding of large document collections at a glance assumes that at least some information of the targeted data set is unknown. Thus, gaining an understanding of large document collections is rather associated with the process of hypotheses formulation than to hypotheses testing, and thus associated with an undirected search. To a certain extent, the information need of the seeker is ill-defined or unknown in the first place. The activity can be described as the process of seeking latent but potentially useful information in large data collections [Kei02, KMS*08], i.e., exploratory data analysis. Brehmer and Munzner assign the task of neither knowing what to seek, nor where to seek as *explore* [BM13]. The effectiveness of systems supporting this type of information seeking depends on the usefulness of the provided content summaries (overviews), on the ability to browse through the data, and on the manner of how local aspects of the data can be achieved and exploited. Many of these systems use the Information-Seeking Mantra ("Overview first, zoom and filter, then details-on-demand") by Ben Shneiderman [Shn96]. In contrast to the query and response paradigm, the human participation in these types of systems shows a continuous active engagement. In the course of this thesis, we refer to this type of activity as *exploration*.

> *Exploration is essentially the construction of a workflow as a cascade of operations that filter, summarize, and analyze the data.*
>
> Jean-Daniel Fekete, *Visual Analytics Infrastructures (...) [Fek13]*, 2013

Exploratory Search Both the search in and the exploration of time-oriented primary data is highly relevant for data-driven research and related application fields. Ideally, analytical systems would support domain experts in both search and exploration, leading to efficient and effective hypotheses testing and hypotheses formulation. In this way, analytical systems would be able to cover large parts of the data life-cycle, particularly the *analysis*, *access*, and *reuse* of time-oriented primary data. On the one hand, the characterized search process corresponds to a directed search, supporting confirmatory analysis. On the other hand, exploration depicts an undirected search supporting exploratory analysis. To better support these different types of information-seeking, they should be conflated to single systems. Domain experts would be able to carry out information-seeking activities from simple lookups to enhanced learning, to complex investigation [Mar06]. Search activity can be supported with visual-interactive querying and result analysis, while for exploration activity content-based overviews, information drill-down interaction, and *Details-on-Demand* [2] functionality can be provided. Metadata attributes provided with the time-oriented primary data can be used with techniques, such as (faceted) search, dynamic queries, and exploratory relation seeking.

In this thesis, we characterize the concept of *Exploratory Search* (ES) as the combination of search and exploration activities, as advocated by Marchionini [Mar06]. A classification of ES with respect to IV and VA is shown in Figure 1.3. We comply with the notion of ES presented by White and Roth [WR09] where ES is described as an extension of

[2] In this thesis, Details-on-Demand is a strategic term. While different scientific notations exist, we use the upper-case variant.

the search activity ("beyond the query-response paradigm"). Similarly, exploratory data analysis can benefit from ES by extending the functionality with search support, such as visual querying and query result exploration. *Exploratory Search Systems* (ESS) combine the techniques known from search systems with techniques from exploratory data analysis to support both activities search and exploration. ESS can greatly benefit from techniques presented in Information Visualization (IV) and Visual Analytics (VA). IV is defined as "the use of computer-supported, interactive, visual representations of abstract data to amplify cognition" [CMS99]. VA is "the science of analytical reasoning facilitated by interactive visual interfaces" [TC05]. According to Keim et al. "Visual Analytics combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making based on very large and complex data sets" [KAF*08]. As such, VA is not only relevant for the *use* of ESS, but also for the *design* of ESS since VA supports the process of making meaningful design decisions. A variety of concepts, guidelines, techniques, and best-practice approaches have been presented in IV and VA, inspiring the design and the use of ESS for time-oriented primary data.

However, the number of existing ESS is comparably small. The most considerable amount of ESS was presented for textual data content (see the work of Herrmannova and Knoth for an overview [HK12]). For non-textual documents only few approaches exist. For time-oriented primary data, the number of ESS approaches is particularly limited. This may be due to the fact that research in ES is still a young discipline and many works still address ES from a conceptual perspective. This is contrasted with the huge design space for ESS, especially for time-oriented primary data. In fact, supporting domain experts with usable and useful ESS for time-oriented data is a challenging task. In the following, we outline a set of research challenges which have to be resolved to design usable and useful ESS for time-oriented primary data.

1.4. Challenges

We briefly outline the six most challenging factors for the design and the use of ESS for time-oriented primary data. Each of the challenges is described in detail in Section 2.6 where we summarize the related work.

C_{MES} Missing Methodology for the Design of ESS The combination of search and exploration activity in a single system is very promising for data-driven research. In this way, a variety of information needs can be supported in a single system. Best-practice ESS make massive use of visualization and interaction designs which evidences the association of ES with IV and VA. However, only a few best-practice ESS already exist, and thus the methodology for the design of ESS remains widely uninvestigated. In addition, most best-practice ESS are limited to textual data content. For non-textual data content the number of approaches is scarce, particularly for time-oriented primary data. Challenges for non-textual documents are caused by the complexity of these data types, which impedes design of query formulation techniques, overviews of the content, and other visual-interactive interfaces. The question arises whether a more transparent and target-oriented summarization of IV and VA techniques can be achieved to support the process of designing enhanced ESS for non-textual data content. A variety of surveys and taxonomies for tasks and techniques in IV and VA exist, which have to be condensed and mapped to search and exploration activity, required for implementations of the ES concept. Furthermore, missing methodology for the design of ESS also applies to the analytical workflows required for powerful ESS. Data scientists can rely on a variety of reference models, frameworks, and reference architectures presented for scientific workflows, *Knowledge Discovery in Databases* (hereafter, KDD) [FPS96], IV, and VA. However, the methodological adoption of these general concepts to the specific challenges and requirements in the design of ESS remains unsolved. ESS require to cope with a huge design space posed by the complex data type and the variety of algorithmic models relevant for ES. It remains a challenge to identify different canonical steps in the workflow to divide the problem into manageable parts. Finally, the involvement of domain experts in the design requires the consideration of yet another type of conceptual methodology; the connection to design studies and user-centered design known from IV and VA. However, while design study methodology supports data scientists in involving users, design studies abstract from the targeted data, tasks, and algorithmic models required in the design process.

C_{CBA} Content-Based Access to Time-Oriented Primary Data Content-based access to time-oriented primary data is a key feature for ESS. The majority of existing ESS, however, focuses on textual data content. For non-textual document types the number of ESS is low, particularly for time-oriented primary data. Search-oriented fields, such as DLs, are confronted with providing content-based access to new non-textual data types. Similarly, from an exploration perspective, it is challenging to gain insight into the structures of large and complex time-oriented primary data

collections. For time-oriented primary data, most challenging factors of complexity are the size, the heterogeneity, and the quality. An indispensable aspect in content-based access regards the transformation of raw data into formats that analysis and visualization techniques can address. For both search and exploration activity, approaches using *feature vectors* (hereafter, FV) have proven to be very effective. We characterize FVs as compact and yet precise representations of complex data objects. However, to transform time-oriented primary data into usable formats, data scientists need to apply cascades of preprocessing routines assembled to a workflow. Important routines are data cleansing models, normalizations, segmentations, and time series descriptors generating FVs. Data scientists also have to define similarity functions which must to stick to the users' notion of similarity. Assumed the challenges of the content-based access to time-oriented primary data are resolved, downstream algorithmic models, such as effective IR strategies and content-based overviews, can benefit from both content-based access and meaningful similarity measures.

$\mathbf{C}_{\text{CBO}}$ Gaining an Overview of the Data Content Content-based overviews of large data collections are a particularly appropriate starting point for exploratory information seeking. Overviews reveal structural information of the data collection, such as frequently occurring patterns, or interesting relations between patterns. In addition, content-based overviews are a powerful basis for the formulation of content-based queries. However, designing content-based overviews is a challenging task. Data scientists have to gain an understanding of the data set for being able to make meaningful design decisions. This task is associated with upstream challenges of the content-based access to time-oriented primary data $\mathbf{C}_{\text{CBA}}$. Moreover, the complexity of the algorithmic workflow required for content-based overviews contributes to the challenge. In essence, a multitude of high-dimensional input data has to be represented in display space in a meaningful and intuitive way. Important steps in the workflow include the aggregation of the multitude of high-dimensional data, the visual representation of these high-dimensional aggregates, and the layout of these visual aggregates in 2D. An additional challenge is induced by the targeted user group. Individual user tasks, notions of similarity and interestingness, and other types of user preferences must to be considered and be harmonized with technical requirements. Yet another issue is the integration of powerful content-based visual-interactive querying concepts on top of content-based overviews. The formulation of visual queries by example and by sketch is greatly beneficial for the ES concepts, albeit this constitutes a challenging task for data scientists, especially for non-textual data content. *Query-by-Example* and *Query-by-Sketch* [3] interfaces require specific visualization and interaction designs. Furthermore, query interfaces have to be linked with retrieval algorithm. Finally, embedding content-based overviews in ESS adds to the challenges. Powerful ESS may consist of multiple views showing the data from different perspectives. Linking concepts have to be implemented, allowing the localization of objects across different views. These overall visualization and interaction designs also influence the design of content-based overviews.

$\mathbf{C}_{\text{C+M}}$ Challenges in Combining Data Content and Metadata Relation-seeking support enables data scientists and domain experts to gain insight into complex structures and relations of time-oriented primary data sets. Interesting relations may exist in the data content, or in metadata. In addition, relations between data content and metadata may be particularly interesting, e.g., to facilitate data-driven research. However, seeking relations between multiple attributes or even multi-modal data types is challenging. The same applies to the design of visual-interactive interfaces supporting domain experts in relation seeking. Since time-oriented primary data collections are assumed to be undiscovered in the first place, sophisticated analysis tools are challenged in autonomously revealing interesting relations hidden in the data. The heterogeneity of the data content and the attached explanatory metadata adds to the difficulty of the challenge. Functional definitions of interestingness and special visual encodings are required to guide users towards interesting relations, and thus to facilitate the computer-supported hypotheses generation process. Appropriate algorithmic models, model parameters in combination with user-centered design decisions are a prerequisite to generate usable and useful visual-interactive interfaces.

$\mathbf{C}_{\text{MPC}}$ Model and Parameter Choice The design of ESS requires the construction of meaningful workflows combining different algorithmic models. The design spaces for providing content-based access $\mathbf{C}_{\text{CBA}}$, creating content-based overviews $\mathbf{C}_{\text{CBO}}$, and facilitating relation seeking between data content and metadata $\mathbf{C}_{\text{C+M}}$ may serve as examples. At a core level, important steps in the workflow are the transformation of time-oriented primary data into a usable form, the extraction of features, as well as downstream models to facilitate search and exploration activity. In addition, the workflows include visualization and interaction designs, allowing their integration into ESS. For the design of ESS, data scientists have to face the challenges of finding right models, putting these into a workflow in the right order, and choosing right model parameters. These three challenges are at heart of VA. The choice of models and their

[3] In this thesis, Query-by-Example and Query-by-Sketch are strategic terms. While different notations exist, we use the upper-case variant.

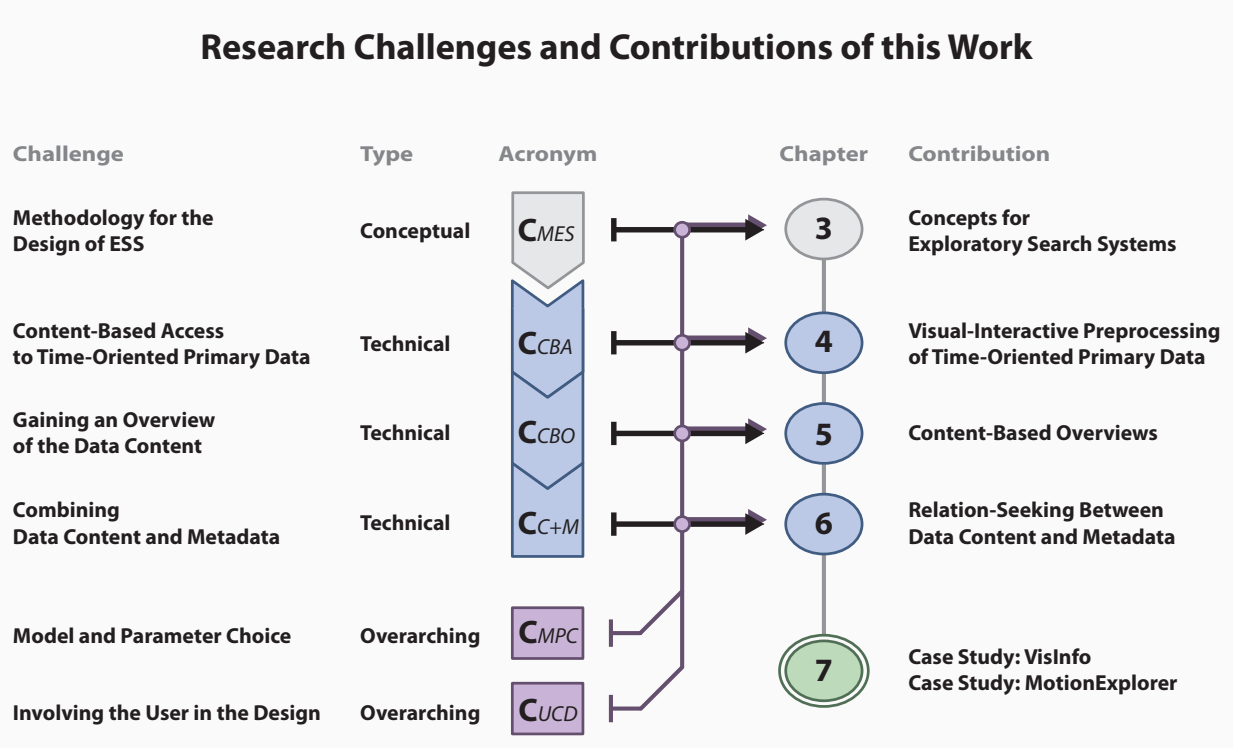


Figure 1.4 *Outline of major research challenges addressed in this thesis. For the conceptual challenge* $\mathbf{C}_{MES}$ *and the three technical challenges* $\mathbf{C}_{CBA}$, $\mathbf{C}_{CBO}$, *and* $\mathbf{C}_{C+M}$, *we provide explicit chapters including the solutions of the problems. Two overarching challenges* $\mathbf{C}_{MPC}$ *and* $\mathbf{C}_{UCD}$ *heavily influenced the concepts and techniques presented in this thesis.*

parameterizations have strong implications on the usefulness of the algorithmic support provided by the ESS. Key roles for selecting appropriate models and model parameters play the assessment of the quality and the involvement of domain experts in the design. We refer to $\mathbf{C}_{\text{MPC}}$ as an overarching challenge since it is relevant for the construction of virtually any data-centered analysis workflow. In this thesis, choosing appropriate models and model parameters is a particular concern for all conceptual and technical challenges.

$\mathbf{C}_{\text{UCD}}$ Involving the User in the Design The three main factors with an influence on the design of VA systems are *users*, *data*, and *tasks* [MA14]. While characterizations of data and tasks (techniques) are to some extent generalizable, the involved users make a VA project special, or even unique. For usable and useful ESS, the engagement of the users in the design is highly appropriate, or even essential. Different user roles may be involved, all having different requirements for the ESS. The collaboration of designers (referred to as data scientists) with users (referred to as domain experts) poses gaps in the knowledge/expertise and in the interest. A particular challenge for domain experts and data scientists is the targeted data collection which is assumed to be unknown in the first place. The missing awareness of the characteristics of the time-oriented primary data collection hampers the requirement definition process and the design of analytical support. Thus, gaining an early understanding of the intrinsic properties of the data set essential for both data scientists and domain experts. Similarly, it is highly appropriate to understand scientists' practices for being able to build good tools and services [BWE06]. Design study methods presented for IV and VA suggest the involvement of users in the design from the start of a project, including the characterization of the users' domain. Users should be involved in the design process, at least for major steps of the workflow. We exemplify the need for user engagement through the definition of similarity and interestingness functions. These functions are required for many powerful algorithmic models facilitating search and exploration activity. Important examples of models using similarity functions are retrieval and clustering algorithms. Throughout this thesis, we refer to *interestingness* as the degree to which *relations* between two objects or groups of objects are relevant for the user. Especially for data-driven research in time-oriented primary data, the notions of similarity and interestingness in the heads of domain experts are particularly important for the design of useful ESS. In general, mapping the notions of data and tasks in the heads of domain experts to functional implementations is a challenge at different steps of the workflow, especially for complex

and previously unknown data sets. We refer to $\mathbf{C}_{\text{UCD}}$ as an overarching challenge. Involving the user in the design is important for most design projects, especially if user needs are complex and possibly ill-defined in the first place. In this thesis, involving the user in the design adds to the difficulty of all conceptual and technical challenges.

1.5. Contribution

A careful reflection of related works reveals ES as a promising concept to facilitate data-driven research in time-oriented primary data. In essence, we postulate six challenges impeding ES in time-oriented primary data. In this thesis, we resolve these research challenges. We introduce concepts, guidelines, techniques, and systems for the ES in time-oriented primary data. In essence, we present a conceptual framework, three main technical contributions, and two case studies for the design and the application of ESS. Figure 1.4 outlines the contributions of this thesis.

Concepts for Exploratory Search Systems With the concepts presented in Chapter 3, we face the challenge of missing methodology for the design of ESS $\mathbf{C}_{\text{MES}}$. In a survey of task taxonomies, we present an overview of IV and VA tasks assembled in a single diagram. The structure of the assembly is characterized by Marchionini's Information-Seeking Process [Mar95] as a common basis for ES, IV, and VA. In addition, we assign these tasks and associated techniques to search and exploration activity. Hence, we bridge the gap between the ES concept on the one hand and the rich set of tasks and techniques of IV and VA on the other hand. Moreover, we present a reference workflow for the design and the application of visual-interactive interfaces for ESS. The reference workflow consists of four main steps, which serve as a baseline for the construction of workflows for ESS. For this purpose, we reflect, adopt, and extend existing reference models, frameworks, and reference architectures presented in scientific workflows, KDD, IV, and VA towards the design of ESS. The reference workflow also comprises the involvement of the user in the design. To achieve this, the reference workflow distinguishes between the design phase and the application phase of visual-interactive interfaces. The indication of the design process reflects and adopts existing user-centered design and design study methodology. Our workflow is applied to the techniques and systems presented throughout this thesis. All three technical contributions and the two case studies build upon the workflow.

Visual-Interactive Preprocessing of Time-Oriented Primary Data We present guidelines and techniques for the visual-interactive preprocessing of time-oriented primary data in Chapter 4. We support domain experts in the construction of a preprocessing pipeline for time-oriented data, and thus provide content-based access $\mathbf{C}_{\text{CBA}}$. The visual-interactive system for the construction of the pipeline allows confronting quality challenges to transform the data into a usable format. Domain experts can add algorithmic models to the pipeline to adapt the data towards individual user needs. Examples are models for the segmentation of time-oriented data to temporal patterns, or normalization models to make the data comparable. With the definition of a time series descriptor the time-oriented data content can be transformed into the feature space. In this way, downstream models of the reference workflow can be executed with a FV representation of the time-oriented primary data. The definition of a similarity function concludes the construction of the preprocessing pipeline. Domain experts are able to express their notion of similarity; the functional definition of similarity then serves as a by-product for downstream models of the reference workflow. Our system is equipped with techniques from VA to enhance the construction of preprocessing pipelines. Different techniques guide domain experts in selecting meaningful models and in defining appropriate parameters (cf. $\mathbf{C}_{\text{MPC}}$). Furthermore, we present a guidance which allows testing the workflow with most diverse input data to assess the generalizability of the workflow for large data collections. The visual-interactive means of the system also enhance the collaboration between domain experts and data scientists, and fosters the transparency of the workflow construction (cf. $\mathbf{C}_{\text{UCD}}$). We prove the usefulness of the techniques in the context of the VisInfo case study, which is presented in Section 7.1 in detail.

Content-Based Overviews of Time-Oriented Primary Data The second technical contribution regards content-based overviews, presented in Chapter 5. We explicitly resolve the challenge of gaining an overview of large collections of data content, e.g., of time-oriented primary data $\mathbf{C}_{\text{CBO}}$. Again, we use our reference workflow as a framework. In three main steps, we present (1) techniques for the visual-interactive aggregation of large time-oriented primary data collections, (2) guidelines and techniques for the visual mapping of high-dimensional data objects, and (3) guidelines and techniques for the layout of aggregated data in the display space. As a result, data scientists and domain experts using our techniques can design meaningful content-based overviews, which can be integrated in ESS. For the visual-interactive aggregation of large time-oriented primary data collections (1), we use the FVs as an input. These FVs may be the product of content-based access strategies (cf. $\mathbf{C}_{\text{CBA}}$), e.g., as presented in Chapter 4. We present

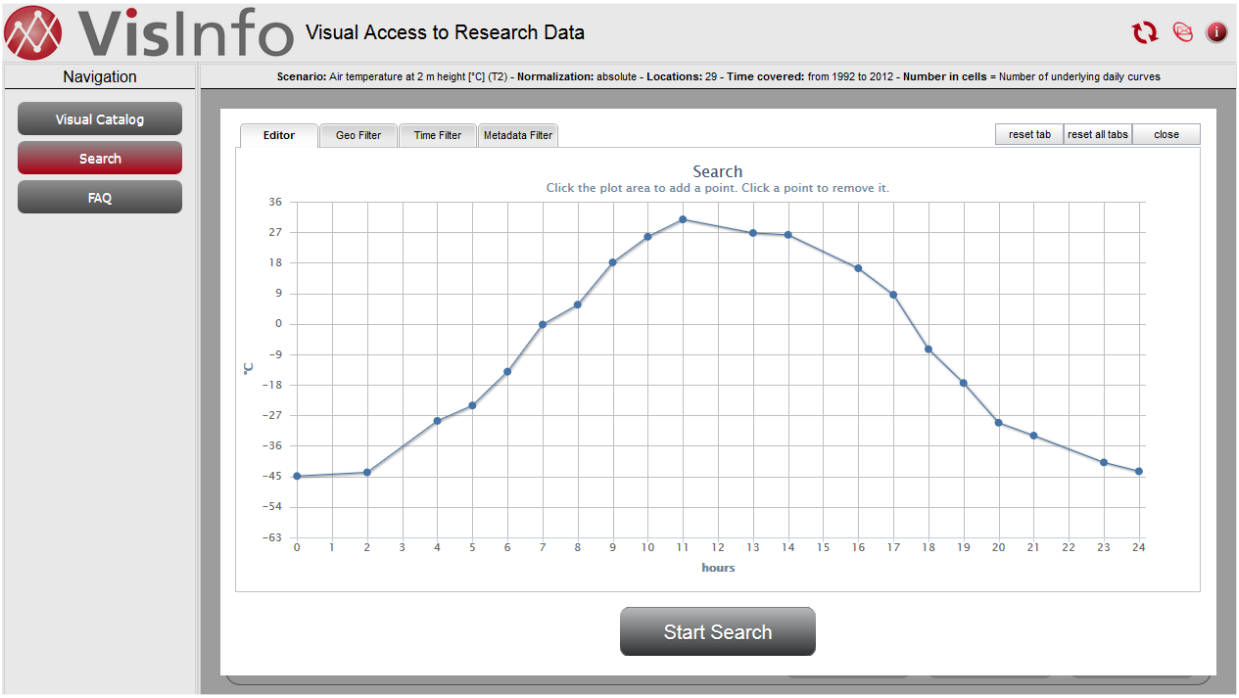


Figure 1.5 *Search in time-oriented primary data. The VisInfo digital library system allows domain experts to sketch a temperature curve which is subsequently used for the retrieval of similar time-oriented data content.*

visual-interactive cluster analysis techniques as a means for the aggregation of the data collections. For this purpose, we use quality measures for cluster analysis to guide the user in the selection of meaningful cluster algorithms and algorithm parameters (cf. $\mathbf{C}_{\text{MPC}}$). The quality-driven analysis techniques cover four different levels of granularity from coarse (macro) levels to fine (micro) levels. Our guidelines and techniques for the visual mapping of high-dimensional data objects (2) emphasize two main aspects. First, we present solutions for glyph designs of high-dimensional data objects. Second, we present solutions for the use of color as a visual variable to encode high-dimensional data objects visually. In the section about layouts for aggregated data (3), we show how high-dimensional data objects can be allocated in 2D in a meaningful way. We present different classes and techniques of layout algorithms, and assess their advantages and disadvantages. In addition, we draw a connection to the data aggregation method used, which has a significant influence on the design choice of meaningful layout algorithms. The guidelines and techniques of the three steps (1), (2), and (3) comply with the overarching challenge of involving the user in the design (cf. $\mathbf{C}_{\text{UCD}}$). Our line of approach includes a transparent, visual, and iterative design process. All guidelines and techniques presented in the chapter are evaluated in association with the two case study ESS presented in Chapter 7 (VisInfo and MotionExplorer).

Relation Seeking Between Data Content and Metadata In Chapter 6, we focus on the challenging task of seeking interesting relations between data content and explanatory metadata $\mathbf{C}_{\text{C+M}}$. In three different approaches, we present technical solutions for different types of relation-seeking behaviors. The three approaches focus on (1) mapping metadata onto content-based overviews, (2) mapping data content onto metadata-based overviews, and (3) revealing relations in visual interfaces for multi-attribute data. The first approach maps metadata attributes onto content-based overviews. For this purpose, we presume a content-based overview solution (cf. $\mathbf{C}_{\text{CBO}}$) as, e.g., presented in Chapter 5. We use the distribution of metadata on the content-based overview as a means for assessing the interestingness of relations between metadata and data content. The second approach maps data content onto metadata-based overviews, Figure 1.7 illustrates the technique. Based on a novel similarity concept for the entities of metadata attributes, we provide layouts of entities in 2D. For each entity in the metadata layout, a small content summary solution is provided. Consequently, domain experts are empowered to reveal interesting relations between metadata entities and the time-oriented primary data content. The third approach abstracts from the distinction between data content and metadata, but rather supports relation seeking in multi-attribute data. Attributes of different type (numerical, categorical, etc.) are unified to sets of bins which build the basis for algorithmic models yielding interesting relations between bins. Different user-defined interestingness measures enhance the relation-seeking process and support domain experts

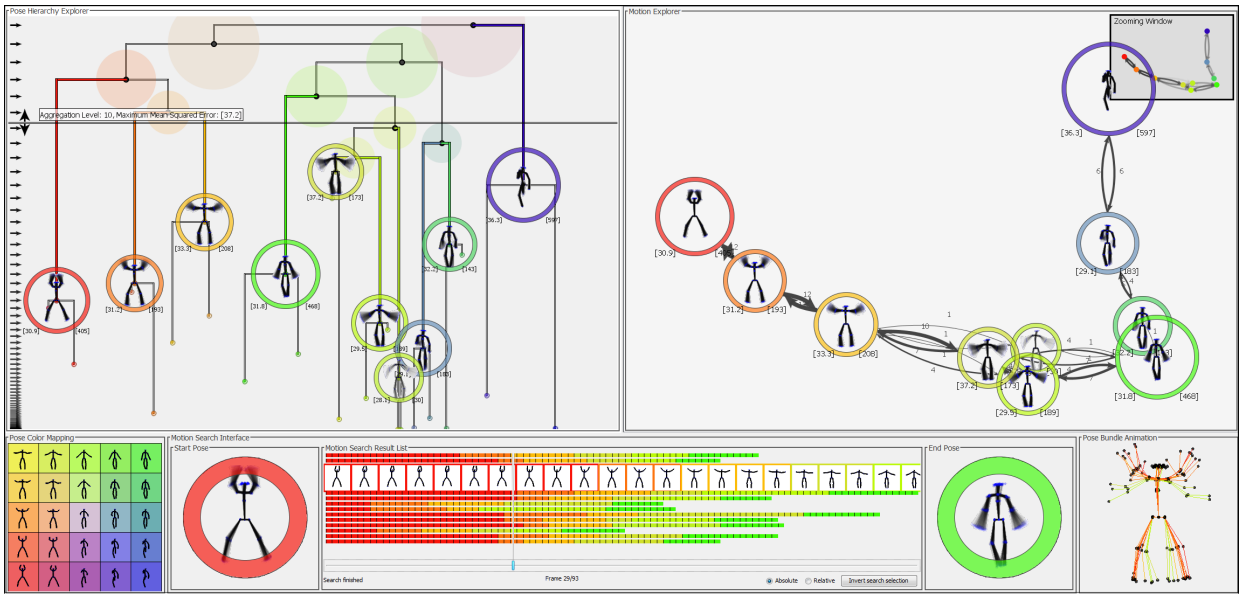


Figure 1.6 *Exploratory search in human motion capture data. Two content-based overviews support users in the exploration of human poses (top left) and human motion (top right). A Query-by-Example technique supports the content-based search in the human motion capture data. Retrieved motion sequences can be explored in a list-based interface (bottom). In the example, a query was submitted consisting of the red and green pose. The search retrieved 12 different jumping jack motions.*

in the formulation of new hypotheses in large and potentially unknown data collections. In all three solutions, we discuss and justify model and parameter choices (cf. $\mathbf{C}_{\text{MPC}}$), and highlight design parameters which should be defined by involving the targeted user group (cf. $\mathbf{C}_{\text{UCD}}$). One beneficial by-product of the solutions is the identification of interesting metadata attributes, which can subsequently be used in faceted search interfaces. We refer to VisInfo case study in Section 7.1 for details.

Case Studies — Exploratory Search Systems: VisInfo and MotionExplorer *VisInfo* is a DL system allowing the ES in time-oriented primary data (see Section 7.1). The Earth observation domain serves as an example application in combination with a data warehouse with a large collection of accessible time-oriented primary data. VisInfo was designed in a design study method in close collaboration with digital librarians and Earth observation scientists. User-centered design concepts and a variety of evaluation techniques helped to design and refine VisInfo towards a usable and useful ESS. Users of VisInfo have the means to gain an overview of hundred thousands of daily time series patterns based on four user-defined notions of similarity. The overview is provided based on a content-based overview of temporal FVs. The temporal FVs are the result of a content-based access strategy, the preprocessing workflows were designed together with the domain experts. The interaction designs of VisInfo support the identification of local aspects of interest and the selection of example patterns. Query-by-Example and Query-by-Sketch interfaces enable users to search for relevant data concent. An example of sketched-based querying of a temperature progression is presented in Figure 1.5. The exploration of the retrieval results includes a visualization of the raw content of the retrieved measurement documents, the utilization of meaningful facets, and the visualization of the search results with different views (geographical and calendar-based).

MotionExplorer is an ESS for large collections of human motion capture data (see Section 7.2). MotionExplorer was designed together with experts from human motion analysis and synthesis domain in a design study project. A careful domain and data characterization and an iterative visualization and interaction design process led to an effective and efficient system where domain experts can search for human motion sequences with only 5 clicks. Based on two content-based overview concepts, experts can gain an overview of both human poses and human motions existing in large data collections (see Figure 1.6). Based on an inquiry of the domain experts within the design phase, the level of abstraction (the number of displayed aggregates/clusters) is user-steerable. Thus, if users change the level of abstraction, the models in the MotionExplorer workflow automatically adapt the content-based overviews, including the associated visual representations of human poses and motions. MotionExplorer uses the Query-by-Example

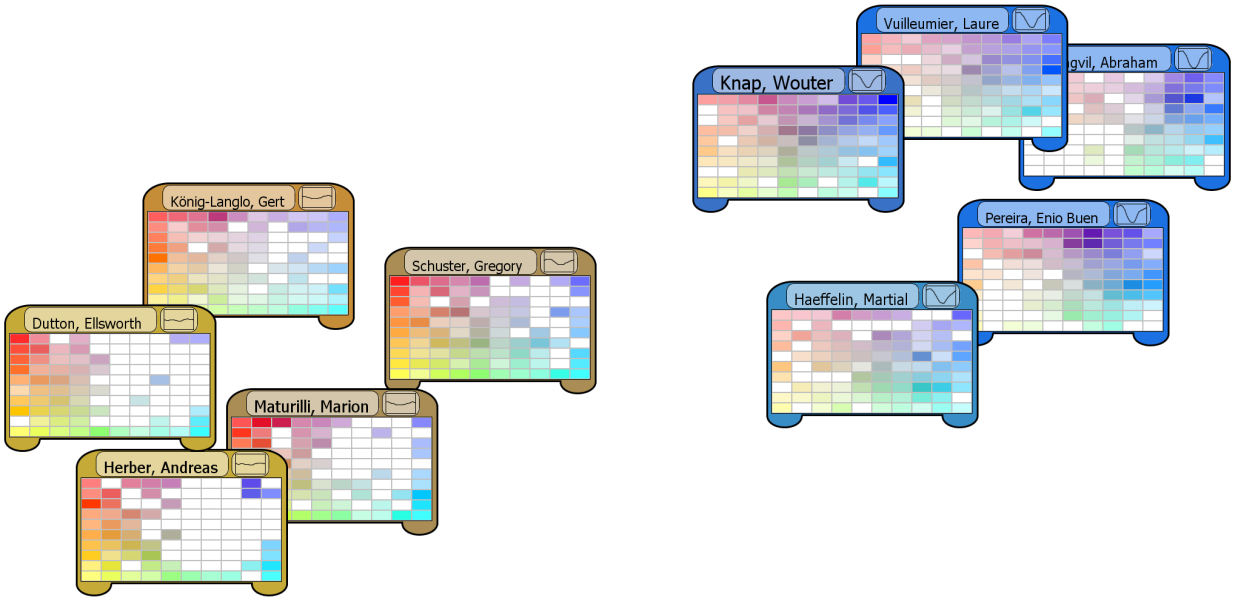


Figure 1.7 *Seeking for relations between metadata and data content. A layout algorithm aligns ten entities (here: domain experts) in 2D based on the pairwise distances of their measured data content (here: daily patterns of relative humidity). Interestingly, the distance relations of the data content divide the domain experts into two groups. Domain experts can employ insight gained from this novel relation-seeking technique in data-driven research.*

concept to support users in searching for human motion sequences. A retrieval algorithm suggested by the domain experts retrieves subsequences of human motions, which can subsequently be explored in the result visualization. The exploration of retrieved motion sequences facilitates the identification of so-called micro variations in human motion. The exploration process is supported by different visualization and interaction designs, such as focus+context, linked views, and an interface for the comparative analysis of multiple search results.

1.6. Thesis Structure

Chapter 2 provides an overview of the related work, organized by the factors data, users, and tasks. We start from a task-based perspective and characterize ES, as well as best-practice ESS. In this connection, we also define IV and VA and indicate the domains' value for ES. The introduction in primary data and time-oriented data represents the data perspective of this thesis. In the following, we combine the data and the task perspective with an overview of scientific workflows and analysis frameworks. Afterwards, the user perspective is represented with a review of relevant stakeholders, methodology on user-centered design and design studies, and evaluation strategies. Finally, we present the research challenges remaining for this thesis in detail.

Chapter 3 describes concepts for the design of ESS. In essence, two main contributions resolve the lack of methodology for the design of ESS $\mathbf{C}_{\text{MES}}$. First, we present an assembly of IV and VA tasks relevant for ES, including a mapping to search and exploration activity. Second, we present a novel reference workflow for the design and the application of visual-interactive interfaces for ESS.

Chapters 4, 5, and 6 comprise the three main technical contributions of this thesis. With these contributions, we resolve the challenges of providing content-based access to time-oriented primary data $\mathbf{C}_{\text{CBA}}$, gaining an overview of the data content $\mathbf{C}_{\text{CBO}}$, and combining data content and metadata $\mathbf{C}_{\text{C+M}}$. The three contributions build on the reference workflow presented in Chapter 3. The order of the chapters reflects the downstream application of the presented techniques. All three chapters share a common structure. An introduction section sharpens the respective research goal to be solved. A brief overview of baseline techniques introduces specific techniques for reuse, integration, or

extension. Afterwards, the main body of each chapter describes the technical contribution in detail. We emphasize the overarching research goals of choosing appropriate models and parameters $\mathbf{C}_{\text{MPC}}$ and the involvement of the user in the design $\mathbf{C}_{\text{UCD}}$ both having a strong influence on the proposed solutions. As the next step, we present usage scenarios and case studies which prove the validity of the approaches. In a summary section, we discuss individual aspects of our contributions and conclude the chapter.

Chapter 7 depicts two ESS for time-oriented primary data built on projects drawing on real data, real users, and real problems. Both ESS, VisInfo and MotionExplorer, implement the reference workflow for the design and the application of visual-interactive interfaces for ESS, presented in Chapter 3. In addition, both case studies make massive use of the technical contributions presented in Chapters 4, 5, and 6. With the two real-world ESS, we prove the validity of the concepts and techniques presented in this thesis. Not least, the ESS demonstrate the applicability of the ES concept for the data-driven research in time-oriented primary data.

Chapter 8 concludes the thesis in a brief overview of the main research challenges and a summary of the contributions solving the challenges. Finally, we outline possible extensions and future work.

1.7. List of Abbreviations

In this thesis, we abbreviate a set of frequently occurring terms. All of these terms are well-known in respective research fields. Table 1.1 details.

Abbreviation	Term
IV	Information Visualization
VA	Visual Analytics
ES	Exploratory Search
ESS	Exploratory Search System
DL	Digital Libraries
IR	Information Retrieval
DM	Data Mining
KDD	Knowledge Discovery in Databases
FV	Feature Vector
HCI	Human Computer Interaction
SciVis	Scientific Visualization
BMU	Best-Matching Unit
QE	Quantization Error
TE	Topographic Error
ML	Machine Learning

Table 1.1 *Abbreviations used in this thesis, listed in the order of occurrence.*

CHAPTER 2

Related Work

In this chapter, we present an overview of related works. The table of contents provides an overview of the structure. Most distinguishable aspects are concepts and techniques, as well as the focus on the factors of the design triangle (data, users, and tasks) for interactive visualizations [MA14]. In Section 2.1, we characterize ES, drawing on definitions for information seeking, IV, and VA. In Sections 2.2 and 2.3, we characterize primary data and time-oriented data. In Section 2.4, we review (scientific) workflows, as well as reference models and frameworks presented in KDD, IV, and VA. Section 2.5 presents an overview of methods and methodologies for the involvement of users in the design. Section 2.6 concludes the related work with a summarization of open challenges which will be resolved in this thesis.

Contents

2.1. Exploratory Search

In this section, we review the scope of ES from a task-based perspective. First, we introduce information seeking referring to as an overcharging concept of ES in Section 2.1.1. Second, we present definitions of IV and VA in Section 2.1.2. On this basis, we review tasks and task taxonomies applied in IV and VA. Third, we characterize search and exploration activity in Section 2.1.3, followed by characterizations of ES and and overview of best-practice ESS.

2.1.1. Information Seeking

In this section, we provide an overview of information seeking. We characterize information seeking, outline environmental factors, and illustrate the information-seeking process. The most relevant works for this review are the book about information seeking in electronic environments by Gary Marchionini [Mar95], the Visual Information-Seeking Mantra by Ben Shneiderman [Shn96], and the work of White and Roth [WR09].

Characterization of Information Seeking Humans were always concerned with the process of seeking information (e.g., material objects, sensual experiences, ethereal objects). Information seeking can be defined as "the process on which humans purposefully engage to change their state of knowledge" [Mar95]. According to Marchionini, the term *search* can be used for both the behavioral manifestation of humans engaged in information seeking and the actions taken by computers to match and display information objects. "Uncertainty is a natural user experience within the process of information seeking and acquiring meaning" [WR09]. Information seeking is to a high level a

cognitive process; it connotes with the process of acquiring knowledge [Mar95]. In this way, information seeking is closely related to concepts like learning or problem solving. As such, information seeking is more human-oriented and open ended compared with IR (see, e.g., [BYRN*99]). IR targets query-document matching under the assumption that relevant information exists to formulate a well-formed query, i.e., that the object must have been known at some point [WR09, p. 10]. If information seekers need to use rather informal and opportunistic seeking strategies, Marchionini uses the term *browsing*; a verbalization we will keep with. Traditionally, the term browsing is inspired from activities, such as navigating across documents, crawling through records, or scanning literature to find items to examine more closely [Mar95]. In an early characterization, browsing is referred to as "an exploratory, information-seeking strategy that depends on serendipity. It is especially appropriate for ill-defined problems and for exploring new task domains." [MS88].

Environmental Factors of Information Seeking Marchionini characterizes a number of personal and environmental factors which influence information seeking. Essential factors are the *information seeker*, the *task*, the *search system*, the *domain*, the *setting*, and the *search outcomes*. The framework of factors, all influencing information seeking, is highly relevant for this thesis.

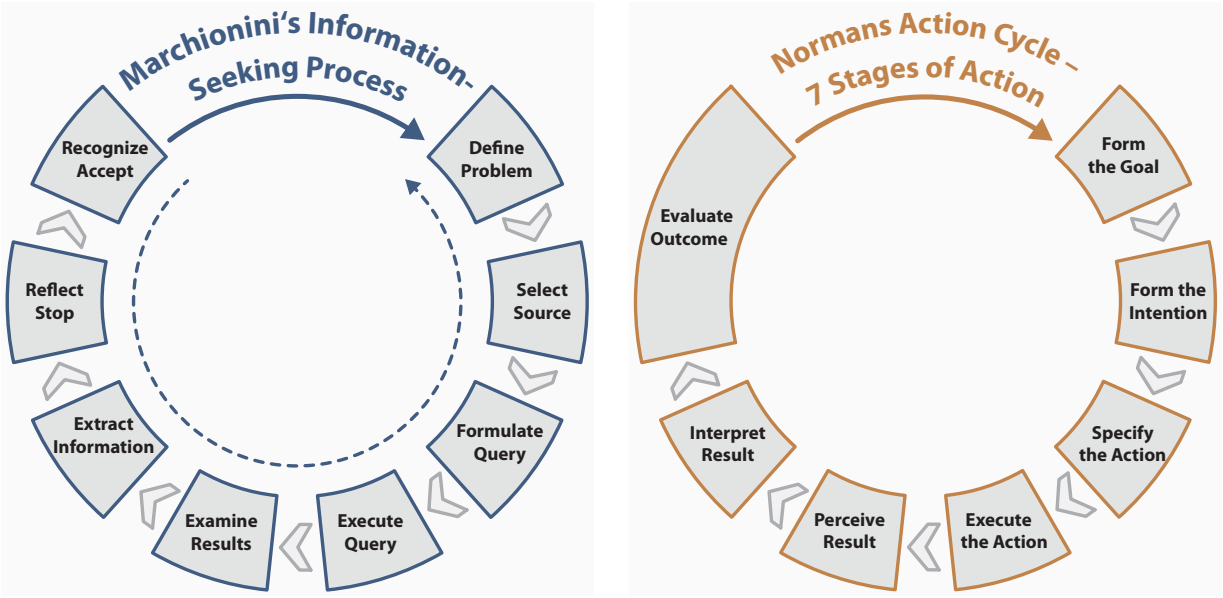
The *information seeker* "defines the task, controls the interaction with the search system, examines and extracts relevant information, assesses the progress, and determines when the information-seeking process is complete" [Mar95]. Information seekers can be distinguished by their *information-seeking behavior*. A variety of definitions and characterizations of humans' information-seeking behavior have been proposed which we will not describe in detail. In this thesis, we adhere to the characterization, e.g., used in [WR09], referring information-seeking behavior to as the way people need, seek, and use information in different context. In Section 2.5.1, we characterize the roles of different collaborators involved in the design and the use of ESS. We assign information seekers the role of *domain experts*.

Marchionini refers a *task* to as the "manifestation of an information seeker's problem and as what drives information-seeking actions" [Mar95]. Information seeking is often motivated by an incompleteness or problematic situation, a gap between what the seeker knows and what he wants to know [WR09, p. 11, p. 32]. This phenomenon is also referred to as the *information need* (see, e.g., [Mar06]). Marchionini distinguishes between the two general seeking strategies of *searching* and *browsing*, depending on how good a query (i.e., a seeking strategy) can be formalized (formal, analytical vs. informal, opportunistic) [Mar95]. Simiarly, Andrienko and Andrienko describe information need by stating what is known and what needs to be found [AA06]. A variety of task taxonomies in IV and VA are based on the types of questions users ask (i.e., *queries*, *constraints*) and seek to answer with a visualization (see, e.g., [Mac95, AA06, BM13]). In Section 2.1.2, we present an in-depth overview of tasks applied in IV and VA.

"A *search system* is a source that represents knowledge and provides tools and rules for accessing and using that knowledge" [Mar95]. Throughout this thesis, we particularly focus on the design and the use of *exploratory* search systems. For the design of ESS, we echo Miksch and Aigner's design triangle [MA14] dividing the design space into thee factors *users*, *data*, and *tasks*. We refer to Section 2.5 for an overview of user-centered activities.

Marchionini describes a *domain* as a "body of knowledge composed of entities and relationships". While the technical contributions of this thesis are domain-independent, the two case studies presented in Chapter 7 are targeted towards specific domains, i.e., DLs and Earth observation, and human motion analysis and synthesis. The *setting* describes the situation of the information seeker, including physical settings like the amount of time. In addition, the setting includes the psychological and social ecology of information seeking. The *search outcomes* include both products and the process. As products, "outcomes are the results using an information system, i.e., feedback from the system" [Mar95]. Especially if the information-seeking behavior is to a high degree exploratory, the process can also be considered a valuable outcome. In this thesis, the latter is of particular importance for the design of useful ESS.

The Information-Seeking Process by Marchionini In Figure 2.1, we illustrate Marchionini's Information-Seeking Process. The eight steps of the process are (1) the problem definition, (2) the selection of a source, (3) the query formulation, (4), the query execution, (5) the result examination, (6) the information extraction, (7) the reflection, and (8) the recognize/accept step [Mar95]. By default, the transactions between the steps are directed forward. However, Marchionini also emphasizes the iterative nature of the process including local cycles. We draw a connection to Norman's Action Cycle [Nor02, p. 46] describing the interaction between a user and the world, using seven steps. Both the Information-Seeking Process and Norman's action cycle are useful for the characterization of user *intents*, user *(inter)-actions*, and *visualizations/representations* from a machine. From a Human Computer Interaction (hereafter, HCI) and a IV perspective, these three factors *user*, *interaction*, and *visualization* are appropriate for the characterization of *tasks*. For designers (data scientists), both Marchionini's Information-Seeking Process and



(a) *The Information-Seeking Process adapted from Gary Marchionini [Mar95]. Marchionini differentiates eight steps. While the main step transitions (gray) point forwards, the process provides a variety of iterative transitions pointing backwards. Obviously, users carry out multiple iterations of the process. The insight gained in latter iterations enhances the query formulation.*

(b) *Norman's Action Cycle [Nor02] describes interaction between users and the world in seven stages. At a coarse level, human action has two aspects: execution and evaluation. Execution "involves doing something" and evaluation "is the comparison of what happened in the world with what we wanted to happen (our goal)." The action cycle is iterative.*

Figure 2.1 *Two process models describing interaction between users and sources of information. Large parts of the two processes show associations with each other. Both process models reflect the principle of humans interacting with visual-interactive interfaces. In addition, both illustrations describe the iterative nature of the process.*

Norman's Action Cycle contribute to the awareness of typical user actions and behaviors, and thus guide the design of computer-supported systems.

We briefly describe the states of the Information-Seeking Process illustrated in Figure 2.1. Information seeking begins with the *recognition and acceptance* of a problem, i.e., a person becomes aware of an information need [WR09]. In the *problem definition* step the information seeker may hypothesize what the answer will be, but at least creates an expectation of what the answer will look like. The *selection* of a search system is often based on the proximity of the information system to the problem and its ease of use. However, Marchionini also points out that "naive information seekers" have default search systems which they use for far to many analysis tasks [Mar95]. We will recall this concern in the related work section about primary data, where we emphasize the missing trust of scientists in unfamiliar (but possibly more powerful) systems, revealed by several studies (cf. Section 2.2). The *formulation* of an initial *query* may serve as an entry point to the system followed by browsing or query reformulations [Mar95]. From a data scientist's perspective, the challenge of this step is the semantic mapping of the information seeker's 'vocabulary' to the 'vocabulary' of the system. The solutions presented in this thesis resolve the semantic mapping challenge with visual-interactive interfaces for the formulation of content-based queries. Related work for the formulation of visual queries for time-oriented data is reviewed in Section 2.3.3. In the *query execution* step both the user and the system put the formulated query into action. If the query implies a search action most of the work is taken over by the system. If the information-seeking behavior of the user is to a high degree exploratory, meaningful interaction designs (e.g., for browsing) allow a continuous active engagement of the seeker [Mar06, WR09]. One challenge of the *query execution* step is to adopt the users' notion of similarity to a functional definition. The solutions of this thesis involve VA techniques to carry out visual-interactive time series preprocessing workflows, including the user-based definition of time series similarity. In the *result examination* step the information seeker interprets the outcome of the search process. The information seeker assesses the progress towards completing the information-seeking task [Mar95]. In the *information extraction* step information is judged by its relevance. However, the gathered information may not fully meet the conditions of the information seekers' goal. In the *reflect/stop* step the information seeker decides

Goal/Task/Action	Description
Overview Gain an overview	Gaining an *overview* of the data set is one of the most important user goals in exploratory data analysis approaches. Based on the overview, the user is able to drill down to local aspects of interest.
Zoom Zoom to interesting items	*Zooming* can be seen as an example interaction technique to drill down in the displayed information. Consequently, zooming interaction supports users to focus on local aspects of interest.
Filter Rem. the uninteresting	*Filtering* interaction allowing users to remove uninteresting items from the display. Shneiderman refers to dynamic queries as an example of filtering interaction.
Details-on-demand Select item(s) and get details	The *details-on-demand* task completes Shneiderman's Visual Information-Seeking Mantra. Relevant types of interesting properties are single values, ranges, or distributions. In many cases, the interestingness of such properties is higher if they occur either most frequently (patterns, structural aspects) or almost never (anomalies, outliers).
Relate View relationships	The *relate* task describes the user activity of identifying relations like e.g., dependencies, associations, or patterns in general.
History Undo/replay/refinement	*History* can be facilitated by visualization designs showing previous user actions, or allowing to step back in the information-seeking process.
Extract Save/share data/params	*Extract* is referred to saving both relevant data subsets and control parameters for reuse or sharing. Related goals of extraction are data provenance and collaboration.

Table 2.1 *The Task by Data Type Taxonomy by Ben Shneiderman. [Shn96].*

whether an additional iteration is necessary. Furthermore, the Information-Seeking Process itself may be subject to reflection [Mar95].

The Visual Information-Seeking Mantra In his Visual Information-Seeking Mantra ("Overview first, zoom and filter, then details-on-demand"), Ben Shneiderman advocates a basic principle for designing advanced graphical user interfaces [Shn96]. First, the graphical system should support users in gaining an overview of the entire data collection. Second, meaningful visualization and interaction designs should support users in drilling down the data collection to local aspects. Finally, the analysis local aspects in detail should be supported.

Ben Shneiderman also echoes Marchionini's distinction of information-seeking behaviors in searching and browsing. On the one hand, Shneiderman describes searching as the process of looking up known things (*fact retrieval*, *known item search*). On the other hand, browsing describes the process of developing an understanding of unexpected patterns within the collection. The distinction between searching and browsing plays a key role in this thesis. Shneiderman maps the concept of information-seeking to seven most relevant data types known in IV an VA [Shn96]. The data types are *one-*, *two-*, *tree-dimensional* data, *multi-dimensional* data, *trees*, *networks*, and *time-oriented* data. Shneiderman further presents one of the pioneer task taxonomies for IV, including seven low-level tasks for designing advanced graphical user interfaces. The taxonomy contains the tasks *overview*, *zoom*, *filter*, *details-on-demand*, *relate*, *history*, and *extract* (see Table 2.1). For this thesis, Ben Shneiderman's work is most important for four factors. First, it describes searching and browsing as two complementing information-seeking behaviors, both highly relevant for our definition of ES. Second, it brings together information seeking and IV (and VA, successively), i.e., it combines the cognitive process of humans, seeking information with techniques to help seekers attain new insights with visualization. Third, it poses guidelines for designing advanced graphical user interfaces, which we will echo for the design of ESS. Finally, it emphasizes time-oriented data as an important data type with specific characteristics which need to be taken into account (see Section 2.3 for a review of time-oriented data).

Related Models and Theories We briefly review baseline models and theories related to information seeking.

Sense-making is "is the creation of situational awareness and understanding in situations of high complexity or uncertainty to make decisions" [WR09]. Klein et al. define sense-making as a "motivated, continuous effort to understand connections (which can be among people, places, and events) to anticipate their trajectories and act effectively" [KMH06]. In sense-making, *search* is considered a part of a larger process [Hea09, p. 80] [WR09]. The process can be divided into two main components: "information retrieval through searching and browsing, and analysis and synthesis of results" [Hea09, p. 80]. At a more detailed level, sense-making typically starts with the identification of a need for information, followed by a specification of the information's structure. The process is continued by searching and browsing for information and the downstream analysis and synthesis of information to create insight. In most cases, the term search is understood in an abstract sense, e.g., as the activity for finding help or make sense of the current situation [WR09]. The sense-making process typically ends with some action deduced from the gained insight. "Sense-making is most often applied in information-intensive tasks, such as intelligence analysis, scientific research, and the legal discovery process" [Hea09, p. 80]. The sense-making process is iterative and often contains internal loops. Similarly, sense-making does not always have clear beginning and ending points. ES supports sense-making through visual representations of large and complex data. These visual representations support revealing structures or

capturing salient patterns of data [WR09]. From a data-centered, more technical perspective, sense-making is mirrored by Fayyad's survey of KDD for making sense of data [FPS96]. Our definition of ES will adopt the main components of the sense-making process: searching and browsing, as well as analysis and synthesis of results.

Information Foraging (theory) attempts to explain information-seeking behavior in humans. The theory adopts food-foraging mechanisms of living organisms as a metaphor to understand how humans *search* for information. The theory assumes that humans transfer these mechanisms over to tasks, such as exploring, finding, and exploiting information (see [Hea09, p. 73] for an overview). Information foraging theory divides the information-seeking process in two activities, i.e., gathering new information patches and consuming available information. As such, information foraging responds to humans facing the "payoff in finding new kinds of information versus the cost of getting to a new patch of information" [Hea09, p. 73]. In this way, information foraging aims at gaining a better understanding of the search behavior of users, e.g., to improve the usability of web search engines. White and Roth assign information foraging as one overcharging concept of ES [WR09]. ESS can support the exploration and identification of information patches and the information gain.

Berrypicking is a model for *searching* in information systems. Bates [Bat89] introduced berrypicking as an information-seeking process analogous to picking berries in a forest. Since berries are typically dispersed rather than bundled, information seekers move through an information space gathering for fragments of information similar to information foraging [WR09, p. 28]. New encountered information gives the berrypicker new ideas and directions to follow and new conceptions of queries, respectively. The process of berrypicking is often illustrated with the trajectory of an information seeker through an information space starting with an initial query and an initial set of retrieved documents. This query-response strategy is iterated leading the berrypicker to different locations of the search space. In the end, the berrypicker has archived two types of results a final set of documents and a path through the search space. Berrypicking is meaningful if the size of the document collection needs to be decreased with every query. Hence, the document collection is reduced, based on the partial information collected in the course of the information-seeking process. Bates refers *browsing* to as a type of search activity supported by berrypicking and respective search interfaces. We acknowledge the notion of browsing as an activity of navigating through a search space, especially in the context of reducing the search space to sets of relevant documents. However, the berrypicking model (and Bate's notion of browsing) lacks an incorporation of exploratory data analysis capability. From a IR perspective, berrypicking inspires the notion ES [WR09]. However, to comply with the definition of ES in this thesis, the berrypicking model lacks concepts and techniques based on exploratory data analysis.

2.1.2. Information Visualization and Visual Analytics

We next introduce and define IV and VA. Afterwards, we gain an overview of the diversity of tasks and task taxonomies in the context of IV and VA.

Introduction to Information Visualization An overall goal of visualization in general is gaining insight [CMS99]. IV is defined as "the use of computer-supported, interactive, visual representations of abstract data to amplify cognition" [CMS99]. In the following paragraphs, we characterize the terms of the latter definition in detail. IV is related to a variety of research fields, such as computer science, computer graphics, information design, and HCI. IV has its beginnings in the late eighties when it emerged from *Scientific Visualization* (hereafter, SciVis). While SciVis focuses on the visualization (rendering, simulation) of scientific data with an inherent physical component, IV involves abstract, non-spatial data [TM04]. Early works in IV are, e.g., visual-interactive browsing systems for hypertext [MS88], visual information-seeking systems using dynamic queries [AS94], or search interfaces for DLs [FHN*93].

Abstract data refers to data that is not connected per se to some spatial location [AMST11], i.e., it does not support the inherent mapping to any geometry. IV includes the research on *visual representations* to map abstract data to computer-supported visual interfaces. One of the most relevant reference models for the design of IV systems is the Information Visualization Reference Model [CMS99], see Section 2.4.3 for an in-depth description. At the coarsest level, raw data is transformed in a structured form (tables), mapped to visual structures, and successively transformed into views of the visual-interactive system [CMS99]. For the visual mappings and the view transformations, IV adopts human perception and design principles, e.g., surveyed by Jacques Bertin [Ber83], Edward R. Tufte [Tuf90], or Colin Ware [War12]. One of the most relevant concepts for mapping abstract data to the visual space is the use of *visual variables* (visual attributes, visual encodings). Visual variables are a set of symbols that can be applied to encode data visually. Important visual variables are, e.g., *position*, *size*, *shape*, *value*, *color*, *orientation*, or *texture*, *aspect ratio*, and *curvature* [Ber83, BKC*13]. Depending on the *type* of the underlying data attribute (numerical, ordinal, categorical),

individual visual variables are more or less accurate. *Interaction* in general can be described as different kinds of actions of subjects or objects having an effect on each other. Interaction in IV and HCI refers to as the communication between human and the machine [YKSJ07]. Similar to HCI, IV uses interaction (techniques) to change and adjust the visual representation of the data. One distinction between IV and HCI is the "asymmetry in data rates" [War12], i.e., in IV the main focus is on the visual representation of data and the interactive adaption of representations, in contrast to entering (new) data into systems. The latter aspect is more important in HCI in comparison to IV [YKSJ07].

Introduction to Visual Analytics More than one decade ago, the research field of *Visual Analytics* (VA) emerged from exploratory data analysis [Tuk77], visual DM, IV, HCI, and other related fields. VA is "the science of analytical reasoning facilitated by interactive visual interfaces" [TC05]. According to Keim et al. "Visual Analytics combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making based on very large and complex data sets" [KAF*08]. VA particularly supports exploratory information-seeking behavior, as postulated in the goals of VA defined by Keim et al. [KAF*08].

- Synthesize information and derive insight from massive, dynamic, ambiguous, and often conflicting data
- Detect the expected and discover the unexpected
- Provide timely, defensible, and understandable assessments
- Communicate assessment effectively for action

Different factors influence the design of VA solutions. We echo the design triangle for the design of VA systems for time-oriented data presented by Miksch and Aigner [MA14]. At this core level, mandatory factors are *data*, the *users*, and *tasks*. VA research has created a variety of methods and methodology for involving the user in the design to yield usable and useful VA solutions. We refer to Section 2.5 where we elaborate user-centered design and design study methodology for VA in detail.

Different Types of Task Taxonomies The combination of visual representations and interaction techniques is beneficial for seeking information in different types of abstract data. IV (and VA) research has created a variety of techniques to support different information-seeking behaviors including search and exploration. In IV the term *task* is often used to describe different types of 'activities'. However, different notions of tasks exist, depending on the perspective. In this section, we shed light on different notions of tasks and *task taxonomies* existing in IV and VA.

Taxonomy: The Task by Data Type Taxonomy We have already referred to the task taxonomy presented in Shneiderman's Visual Information-Seeking Mantra [Shn96]. Shneiderman combines the perspectives of users, interactions, and visualizations to present seven tasks essential for *designing* advanced graphical user interfaces. The task taxonomy for visual information seeking is presented in Table 2.1.

Taxonomy: The Task Taxonomy by Card et al. One of the most relevant baseline works for IV is presented by Card et al. [CMS99], extending Shneiderman's data by task taxonomy [Shn96]. The taxonomy of Card et al. is relevant for this thesis as it explicitly names *Browse* and *Search*. Other additional tasks are *Read Fact*, *Read Comparison*, *Read Pattern*, *Manipulate*, and *Create*. The taxonomy was adopted by many succeeding taxonomies.

Taxonomy: A Task Taxonomy for Spatio-Temporal Data In their book about the VA of spatio-temporal analysis, Andrienko and Andrienko present a taxonomy for spatio-temporal data [AA06]. The authors build up their taxonomy from a user's perspective. Spatio-temporal data has specific data characteristics that influence visualization and analysis tasks [Shn96, AMST11]. Important tasks are the *identification* of data based on their value domain, the *localization* of data based on their temporal (geo) domain, the *comparison* of data elements, and *seeking relations* between data elements. An important distinction is made between user tasks at the *elementary* level (single elements) and at the *synoptic* level (stets of elements, clusters). We present an in-depth overview the task taxonomy of Andrienko and Andrienko in Section 2.3 on time-oriented data.

Taxonomy: The Task Taxonomy of User Interactions by Yi et al. Yi et al. [YKSJ07] present a task taxonomy with the focus on high-level user interactions and associated interaction techniques. Yi et al. survey former works and classify tasks by their degree of abstraction from high-level to low-level. High-level tasks are referred to as goals/intents of users while low-level tasks are associated with techniques to achieve user goals/intents. On this basis, Yi et al. present their taxonomy of high-level user interactions (intents) associated with low-level interaction techniques, as shown in Table 2.2.

Taxonomy: Categorization of IV and VA Techniques for Multi-faceted Scientific Data by Kehrer and Hauser In a recent survey, Kehrer and Hauser present a categorization of techniques for IV and VA based on multiple facets of

Goal/Task/Action	Description
Select Mark something as interesting	*Select* interaction techniques enabling users to mark data elements as interesting. A special visual encoding supports the identification of selected data elements. The enhanced identification also enables users to follow selected data elements when the visual representation is subject to change.
Explore Show me something else	*Explore* describes the set of interaction techniques allowing users to show different data elements, i.e. to carry out breadth first search traversal. The authors exemplify panning as a relevant exploration technique where the user typically grabs the scene and moves it by mouse dragging or with a scrollbar.
Reconfigure Different arrangement	*Reconfigure* interaction allowing users to change the way data items are arranged at the display. The reconfiguration of the visual representation provides different perspectives on the data.
Encode Different representation	*Encode* describes the set of interaction techniques allowing users to show the data with a different visual representation, including changing visual variables. Changing the type of representation provides different aspects of the data and its relationships.
Abstract, Elaborate More or less detail	*Abstract/Elaborate* enabling users to adjust the level of abstraction of a data representation. This class of interaction techniques echoes the Information-Seeking Mantra showing data representations from an abstract (overview) to a fine-gained level of detail (detail-on-demand).
Filter Show me something conditionally	*Filter* interaction involves the concept of conditional data representation. Many filter techniques enable users to specify ranges or conditions meaning that only data elements are displayed meeting the criteria. Dynamic query techniques are among the most relevant filter interaction techniques.
Connect Show me related items	*Connect* interactions enabling users to reveal associations and relationships (relations) between data elements. Relations may exist between already shown items or may include hidden items. The highlighting technique is an example of emphasizing data elements within different views.

Table 2.2 *Task taxonomy by Yi et al. [YKSJ07] with high-level user interactions and associated techniques.*

scientific data [KH13]. The authors explicitly use the term techniques, i.e., the taxonomy describes how IV and VA tasks can be supported at a low, technical level. The categorization is relevant for this thesis for the following reasons. First, the focused data types are highly related to primary data. Second, Kehrer and Hauser's categorization reflects a broad scope including techniques from KDD, HCI, IV, and VA. Third, many of the described techniques in the context of KDD and VA are not only relevant for the use, but also for the design of ESS. The six categories in Table 2.3 show the broad spectrum of the taxonomy, ranging from techniques that mainly address visual representations to methods that focus on computational analysis. In addition, many of the surveyed approaches rely on interaction concepts, such as linking and brushing, zooming, panning, or view reconfiguration [YKSJ07]. Many of the analytical techniques are associated with KDD. We review techniques from KDD in Section 2.3.2, in which we provide an overview of temporal analysis tasks. The KDD process in general is reviewed in Section 2.4 with an emphasis on workflows for the design and the application of ESS.

Taxonomy: The Multi-Level Typology of Abstract Visualization Tasks by Brehmer and Munzner Finally, we review a recent typology of abstract visualization tasks presented by Brehmer and Munzner [BM13]. Similar to Yi et al. [YKSJ07], the authors address the problem that gaps exist between the multi-faceted concepts of tasks in existing task taxonomies. In particular, Brehmer and Munzner's task typology responds to the gap of different levels of abstraction ranging from low-level to high-level tasks. The multi-level typology is based on the differentiation *why* and *how* visualization tasks are performed, and *what* (data) constitutes the input and the output of visualization tasks. The three questions why, how, and what are similar to the characterization of the visualization problem presented by Aigner et al. [AMST11]. This distinction is highly relevant for the design of (time-oriented) visualization systems, and thus relevant for this thesis. An additional benefit in the work of Brehmer and Munzner is a survey of dozens of task taxonomies and typologies.

The survey strengthens the use of some termini employed throughout this thesis. We give a brief overview of the typology. *Why* describes why a task is performed in different scopes from *consume*, to *search*, to *query* (from high-level to low-level). At the search-level, Brehmer and Munzner distinguish between known/unknown target identities and known/unknown locations (in the visualization). This differentiation echoes the taxonomy of Andrienko and Andrienko [AA06] where spatio-temporal references (locations) are opposed to characteristics (target identities), see Section 2.3.2 for an in-depth description of temporal analysis tasks. If both the target and the location of a data object are known, the user is able to execute a *lookup* task in the visualization. If a user knows the target identity but is unaware of the location of a data object, a *locate* (localization) task needs to be supported by the visualization system. Localization tasks are also known as known-item search [Mar06]. *Browse* is the task required if the user knows the location of an unknown target identity. For time-oriented data, this task is described as a direct lookup, or as an *identification* task [AA06, AMST11]. Finally, *explore* entails searching for characteristics (target identities) without regard of their location. Exploration tasks are also referred to as overview tasks or foraging tasks [CMS99]. Browse and explore might include discovering data characteristics like particular values, extreme values, anomalies, trends, or ranges [BM13]. The characterization of explore (as a task based on ill-defined information need) reflects the notion of exploration activity focused in this thesis.

Category	Description of Techniques
Visual Data Fusion	e.g., maps, glyph designs, color & texture, transparency, layering techniques, fusion of multiple runs, summary statistics.
Relation and Comparison	e.g., side-by-side comparison (juxtaposition), overlay (superposition), explicit encoding of differences/correlations (e.g. by difference, or by time-warp), see also [GAW*11].
Navigation	interactive search, zooming, panning, ranking, automated viewpoint selection, parameter space navigation, see e.g. [BPFG11].
Focus + Context, Overview + Detail	e.g., different graphical resources focus specification, clustering and outlier preservation, brushing combined with overview and detail and linked views.
Interactive feature specification	e.g., pattern sketch by user, supervised machine learning, transfer functions.
Data aggregation and abstraction	e.g., pattern and value extraction, dimension reduction, aggregation and summary statistics, clustering, outliers.

Table 2.3 *Techniques for the IV and VA of multi-faceted scientific data by Kehrer and Hauser [KH13].*

2.1.3. Characterization of Exploratory Search

Search From the earliest days, search for information has been a fundamental application of humankind, e.g., for research or development [BWE06, Mar06]. At a rough estimate, the history of computer-supported search can be divided into two eras; before and after the Web became a worldwide phenomenon [Hea09, p. 3]. Before the Web, computer-supported search and retrieval was predominantly executed by small groups of highly educated and trained users, such as librarians, journalists, and other search intermediaries [Hea09, p. 3]. The search corpora were specialized, information-oriented text collections of high quality, such as bibliographic records for university libraries, newswire articles, or legal cases of opinions. Older search systems were based on metadata (document abstracts, titles, etc.) rather than on the document content itself. "Search was usually used to find the name and location of a source containing this information, and then a physical paper copy would have to be obtained to see the full text" [Hea09]. Teletype displays typically posed the user interface, often provided with command-line interfaces. These interfaces usually required complex operators for the specification of queries which were difficult to learn.

The shift to the era of the Internet exposed a new type of search. In contrast to the earlier era, everyone was now able to use computer-supported search systems from both perspectives their physical accessibility and their required user expertise. The search space changed from specific high quality collections to large and heterogeneous document corpora including a multiplicity of information types. With the Internet era, the full-text of documents formed the new basis for the retrieval rather than metadata. However, the content-based access to complex data types, such as time-oriented primary data, remains a challenge to this day [Hea09] (see the related work on primary data in Section 2.2). Finally, graphical displays were introduced, which replaced the Teletype displays. Factors of usability were taken into account more seriously, such as the enhanced formulation of queries. However, the visual-interactive formulation of queries for time-oriented data is still scarcely provided in today's search systems (see the related work on time-oriented data and applications in Section 2.3).

In the last decade, search concepts were further extended with navigation and browsing activity, leading to the notion of ES [Mar06, WR09]. The overall trend toward more active engagement of the user in the search process is associated with the progress made in HCI, IR, and IV research [Mar06]. For textual data content the incorporation of IV techniques has helped to improve search interfaces [Hea09], and thus to better support search-like information-seeking behavior. For textual data content there even exists a counterpart to the Visual Information-Seeking Mantra, postulated by van Ham and Perer and referred to as "Search, Show Context, Expand on Demand" [vHP09]. The principle assumes that users are able to start the information-seeking process with the formulation with a query. Based on the result set of a textual query, users are enabled to expand the retrieved network, successively. While this concept is convincing, it still requires users being able to formulate their information need, which is difficult for many data types. The visual-interactive formulation of queries for non-textual data types is still subject to ongoing research [vLSFK12]. Similarly, it remains a challenge to reflect and condense the rich set of search-support techniques and make them available for designers of ESS.

Exploration In the current language, exploration is associated with terms like investigation, expedition, analysis, discovery, research, diagnosis, or inquiry. Together with exploration, these concepts share the need for implicit information, the curiosity for something unexploited, or the possibility for revealing of something unexpected. These

paraphrases also match the notion of exploration in the context of exploratory data analysis, visual data exploration and VA in general. In this section, we characterize exploration in the context of data analysis, present related user tasks, and review interaction and visualization techniques supporting exploration.

Data exploration and *exploratory data analysis* can be described as the process of seeking latent but potentially useful information in large data collections (see, e.g., [Tuk77, Kei02, KMS*08]). Data exploration is associated with DM [HKP11] and KDD [FPS96]. The field of data exploration research was formed far back in the history of the computer era. An early conceptual work addressing the scientific data analysis process is presented by Springmeyer et al. [SBM92]. The work is particularly relevant since it explicitly focuses on the characterization of data-centered research practices. Early techniques for the exploration of data were, e.g., the k-means clustering algorithm presented by MacQueen [Mac67] or the Principal Component Analysis (PCA) projection algorithm by Jolliffe [Jol02] also called discrete Karhunen-Loéve Transform (KLT). At a coarse level, the exploratory data analysis process starts with an at least partially unknown data collection. The explorer has only a vague idea of the data content, and thus starts the exploration with no prior assumptions or hypotheses about the data. In the iterative data analysis phase, the explorer usually carries out a set of undirected search activities, usually to detect structures, patterns, trends, or outliers. The data exploration process ends with result visualizations which possibly lead to new hypotheses about the data.

A paradigm shift and a most motivating aspect for VA was forging ahead with visualization and interaction designs to support the entire data exploration process. *Visual data exploration* (visual DM, interactive exploration) describes the combination of traditional DM techniques and IV techniques [Kei02]. Visual data exploration has the advantage that the user is directly involved in the process. In opposition to DM where a data set is analyzed automatically, visual data exploration is a human-guided process. However, it is not the aim of visual data exploration to replace DM, but to couple the rich set of DM routines, enhancing the exploratory data analysis process. It combines the strength of automatic analysis techniques with the strength of human perceptual and cognitive abilities. Pioneering works for the visual exploration of data from a visualization perspective are the works of Bertin [Ber83] and Tufte [Tuf90]. A large variety of information visualization techniques has been developed to support visual data exploration, see, e.g., the taxonomy by Keim [Kei02].

The introduction of VA further contributed to the process of enhanced data exploration, especially for complex data types. The goal of exploration in the context of IV, VA, and data management was recently described by Jean-Daniel Fekete. [Fek13] "Exploration is essentially the construction of a workflow as a cascade of operations that filter, summarize, and analyze the data". In Section 2.4, we review workflows for KDD, IV, and VA in detail.

Exploratory Search One of the first works describing the term ES is presented by Marchionini in 2006 [Mar06]. Marchionini postulates the need for new search interfaces which "move the process beyond predictable fact retrieval" - "from finding to understanding". Based on studies conducted in an earlier works [MS88], Marchionini emphasizes the existence of different analytical search strategies of different complexity. Using the example of web searches, he describes the expectations of users spanning from searching to learning and exploratory discovery, drawing on a more active engagement of users in the search process as opposed to classical search theory. "Exploratory search makes us all pioneers and adventurers in a new world of information riches awaiting discovery along with new pitfalls and costs" [Mar06]. Marchionini characterizes ES as a combination of search activities that go beyond simple *lookup* tasks. In particular, he emphasizes *learning* and *investigation* as the two key concepts exceeding fact retrieval and known-item search. Learning can be facilitated by techniques supporting comparison tasks, data aggregation and integration. Investigation can be provided with concepts and techniques, such as analysis, navigation, evaluation and discovery. Marchionini postulates to "combine the work in HCI and IR" as a beneficial means to facilitate ES. While Marchionini emphasizes the need for new information tools, content, and pattern mining support, he makes no explicit reference to IV, let alone VA.

In the work of White and Roth [WR09], IV is explicitly named as a means of enhancing ES, together with sense-making, interactive IR, berrypicking, information foraging, exploratory behavior, and information seeking. The authors emphasize the applicability of ES for information seekers who are generally [WR09]:

- unfamiliar with the domain of their goal (i.e., need to learn about a topic)
- unsure about the ways to achieve their goals (either the technology or the process)
- or even unsure about their goal in the first place

In the first place, the information need of users is underspecified and associated with some degree of uncertainty. To reduce the uncertainty, users can carry out ES activity to explore the search space iteratively. This is why ES is often referred to as a concept where the route becomes the destination. The definition of ES, however, is complex and

multifaceted [WR09]. "Although there may be circumstances where exploratory strategies are used continually to allow people to discover new associations, kinds of knowledge, and decision making, they are often motivated by a complex information problem, a poor understanding of terminology and information space structure, and a desire to learn" [WR09]. The authors postulate a model combining two important elements of ES, i.e. the *problem context* and the *search process*. The problem context in ES is ill-structured and often includes complex situations. Users require additional contextual information to clarify their goals and actions. During the ES process, "it is likely that the problem context will become better understood by the searcher, allowing them to make more informed decisions about interaction or information use" [WR09]. "The problem solution can be constructed from information within relevant documents and knowledge accumulated during the search, including the examination of partially relevant and irrelevant documents" [WR09]. The search process may involve multiple query iterations and result sets which require interpretation, information extraction, and reflection (cf. Marchionini's Information-Seeking Process [Mar95] and Norman's Action Cycle [Nor02] presented in Figure 2.1). "Much of the search time in learning tasks is devoted to examining and comparing results, as well as reformulating queries to discover the boundaries of key concept definitions. Learning search tasks are best suited to combinations of browsing and analytical strategies, with lookup searches embedded to locate the correct neighborhood for browsing" [WR09]. In their model of ES, White and Roth illustrate the interaction between the search process and the problem context. The authors postulate the complementary means of the two predominant actions *exploratory browsing* and *focused searching*. "Exploratory browsing exposes users to collection content to help relate the problem context to similar documented experiences and promote information discovery. Focused searching may include some degree of navigation, but is generally intended to help the user follow a known or expected trail rather than forging new ground. Effective ESS will maintain a balance between analytical and browsing activities and support a symbiotic search relationship between searcher and system" [WR09].

We conclude the review of White and Roth's characterization of ES with an overview of mandatory features which should be provided in ESS. The design of ESS poses new requirements which involve new research (e.g., in IV), and collaborative efforts among data scientists, domain experts, digital librarians and users in general. The list of required features is as follows [WR09].

1. Support querying and rapid query refinement: Systems must help users formulate queries and adjust queries and views on search results in real time.
2. Offer facets and metadata-based result filtering: Systems must allow users to explore and filter results through the selection of facets and document metadata.
3. Leverage search context: Systems must leverage available information about their user, their situation, and the current ES task.
4. Offer visualizations to support insight and decision making: Systems must present customizable visual representations of the collection being explored to support hypothesis formulation and trend spotting.
5. Support learning and understanding: Systems must help users acquire both knowledge and skills by presenting information in ways amenable to learning given the users' current knowledge/skill level.
6. Facilitate collaboration: Systems must facilitate synchronous and asynchronous collaboration between users in support of task division and knowledge sharing.
7. Offer histories, workspaces, and progress updates: Systems must allow users to backtrack quickly, store and manipulate useful information fragments, and provide updates of progress toward an information goal.
8. Support task management: Systems must allow users to store, retrieve, and share search tasks in support of multi-session and multi-user ES scenarios.

Exploratory Search Systems We conclude the review of ES with an overview of best-practice ESS. The structure is based on the predominant data types used in the ESS.

Textual documents are maybe the most frequently used data type supported by ESS. We refer to the work of Herrmannova and Knoth for a recent overview [HK12]. A welcome early example is the Envision DL system [FHN*93]. A content-based access strategy supports querying and browsing in scientific literature. A query history enables information seekers to move backwards in the process. Search results are visualized at a 2D display where the axis are spanned by relevant metadata. A similar result visualization technique is presented by Shneiderman et al. [SFRG00]. The DL system distributes thousands of retrieved documents on a 2D display. The axes are defined based on categorical or hierarchical data structures. The Jingsaw tool by Stasko et al. supports querying and relation seeking in text documents [SGL08]. In multiple views, information seekers can draw connections between documents and document groups. The latter is facilitated by aggregation techniques for the text documents. A recent system for searching and

browsing in large textual document collections presented by Nocaj and Brandes uses hierarchical data aggregation and layout algorithms to align query results in 2D [NB12]. In a usage scenario for newsletter articles, the authors demonstrate how meaningful topics and key terms support information seekers in gaining an overview of the document collection represented by a map metaphor.

A number of ESS focus on *metadata-rich* document types. In this connection, effective, faceted browsing and filtering techniques are particularly beneficial. The FacetMap approach overcomes drawbacks of scrollable list-based result visualizations with multiple views containing metadata facets [SCM*06]. The facets support different data types, such as temporal or hierarchical data, most of them aligned in 2D views. FacetMap supports both querying and browsing. The VisGets ESS follows a similar strategy [DCCW08]. Different IV techniques are applied in multiple coordinated views to support visual information seeking and ES. Several types of web-based information retrieved with textual queries is visualized in temporal, geographical, or word cloud-based views supporting browsing and filtering interactions. A more recent ES approach uses the semantic structure from Wikipedia together with textual and spatio-temporal information [HCQ*12]. Different visualization techniques support relation-seeking tasks between query 'concepts', Wikipedia documents, and additional metadata. Example visualizations are geo-referenced or temporal-annotated views, as well as general techniques, such as heatmaps.

An early best-practice example of *movie* documents is the FilmFinder tool by Ahlberg and Shneiderman [AS94]. The visual-interactive system extends classical IR concepts by massive use of dynamic querying for a variety of metadata attributes about movies. A 2D arrangement of movies at the center of the display provides an overview of the movie collection. The temporal domain of the movie documents is mapped to the x-axis, the y-axis represents the movie rating. A categorical colormap encodes movie genres visually. A DL system for *music* collections facilitating content-based music retrieval and visual-interactive browsing is presented by Merkl et al. [MPR02]. A content-based overview visualization based on data aggregation summarizes large music archives. Users can use the overview to drill down in the search space towards different groups of music documents. For that purpose, the data aggregation algorithm provides a hierarchical structure. The MusicGalaxy system by Stober and Nürnberger takes up the concept on content-based overviews for large music archives [SN11]. The approach makes use of data projection algorithms for the creation of similarity-preserving layouts which, in return, provide the overview of the music collection. Different interaction designs, such as querying, zooming, and filtering enable information seekers to facilitate ES activities.

The Timebox widgets technique by Hochhauser and Shneiderman enables users to draw rectangular areas and angular constructs in the display to apply content-based dynamic queries for *time-oriented* data [HS04]. Consequently, large numbers of time-oriented data are mapped into a query interface by means of superposition. Usage scenarios indicate the benefit of these content-based querying techniques to enhance ESS for time-oriented data. In their exploratory data analysis approach for *microarray* data, Seo and Shneiderman combine dynamic queries with browsing techniques [SS02]. Hierarchical clustering results of genomic microarray data can be explored visual-interactively. Multiple views and different dynamic query controls supplement the exploratory data analysis process. A profound overview of ESS for *electronic health care* records is provided by Rind et al. [RWA*13]. Various tools and systems are presented supporting physicians in the ES in *patient histories*. A survey of search and exploration approaches for different data types in the *biological* and *chemical* domains is presented by Attwood et al. [AKM*09]. The authors emphasize the need for more powerful tools to access and extract the knowledge hidden in databases and articles. In addition, a variety of approaches for finding and extracting valuable information is presented, partially implementing ES concepts.

Open Research Questions in Exploratory Search ES is a convincing concept to overcome current practice approaches in different ways. A variety of open research questions exist some of which are outlined in the following. For an overview of the diversity of research questions for ES, we refer to the works of Marchionini [Mar06] and White and Roth [WR09].

A widely accepted research question is the *need for better tools* to support ES in general [Mar06]. Most of the current tools supporting information seeking either focus on search, or exploration, but do not provide ES activity. The varieties of different information-seeking behaviors (spanning from lookup to learning to investigation [Mar06]) are only sparsely supported in single systems. The difficulty of this research question becomes worse if one considers the multitude of ES solutions required to meet the different combinations of included data, users, and tasks (cf. the design triangle by Miksch and Aigner [MA14]). Especially in cases where complex data has to be associated with complex but ill-defined user needs, it is not yet foreseeable to what extent ES can be put into practice. In this connection, further research is needed in the application and reflection of ESS design projects with different combinations of the design triangle. Closely associated is the need for *interdisciplinary* and *collaborative* approaches combining the strength of different collaborator roles, such as data scientists and domain experts (cf. Section 2.5).

A central research question regards the *content* of the included data collection. *Content-based access* plays a key role for many tools facilitating search and exploration support [MRC05, SN11, HK12, vLSFK12]. ES systems exploiting the data content in a meaningful way may include effective and efficient computational support, which in return requires careful designs. Example approaches for content-based access exist in domains where IR is a predominant access paradigm (see, e.g. DLs [BWE06, CCF*08, HPK08, HTT09]), as well as domains where exploratory data analysis and VA is frequently applied (see, e.g., the work of Keim et al. [KMS*08] for an overview). To cope with the data content a variety of technical challenges have to be resolved. Examples are the complexity of the underlying data content (c.f. Sections 2.2.1 and 2.3.1) and challenges associated with providing the mandatory processing and analysis support (cf. Section 2.4.5). Furthermore, ESS need to provide enhanced visualization and interaction designs to support users in accessing and exploiting the data content.

An additional objective in ES is the combined exploitation of *data content and metadata* of document collections. Provided that access strategies exist to both data content and metadata, further research is required to support users in revealing the value of both data 'spaces'. The combined search and exploration in these two multi-modal spaces enables users to identify interesting *relations* between data content and metadata, and thus to gain new insight. Technical research challenges are coping with heterogeneous data types, estimating the interestingness of relations with algorithmic support, providing meaningful visual interfaces, and involving users in the design.

ES requires further research in *search interfaces* [Mar06, Hea09]. In general, searchers need to be supported in an enhanced query formulation, in the modification of queries, and in the exploration of search results [vLSFK12]. All three research goals can greatly benefit from visualization and interaction techniques presented in HCI, IV, and VA research. Relevant examples of "fluid user interfaces" [Mar06] include faceted search and dynamic query tools [HS04, Hea09]. Both techniques are beneficial means of bridging search and exploration activity in single systems. In addition, powerful content-based querying concepts, such as Query-by-Sketch and Query-by-Example, can be put into practice to further support searchers.

Other open research questions in ES apply to supporting *exploratory activity*. Explorers may require breadth-first search as the preferred traversal strategy, e.g., to gain an overview of the data collection. Similarly, ESS need to support the hypotheses formulation process. White and Roth describe the need for "what-if-analyses" [WR09] in ES and the associated requirement to deduce structural information about the data set, which in turn supports sound decision making. From a technical perspective, providing powerful content-based summaries (overviews) of the data collection [SS02, AA06, HK12, KH13] poses one of the most challenging tasks. The data content needs to be accessed, processed, aggregated, and visually represented in a meaningful way. In this connection, research in ES can greatly benefit from exploratory data analysis and VA (see, e.g., [Kei02, KMS*08]). Another technical challenge is supporting explorers in browsing through huge information spaces, e.g., to drill down to meaningful subsets as an alternative to querying (cf. Section 2.1.1). For this purpose, enhanced visualization and interaction designs need to be provided. Finally, ES needs to support the detailed analysis of relevant small subsets (details-on-demand).

We conclude the outline of open research questions with an emphasis on *evaluation* strategies for ES and ESS. The implementation of this new search paradigm including new data processing techniques, new functional support, and new visual-interactive means requires new strategies to assess the usefulness and the usability. In the following, we outline possible research directions for the evaluation of ESS.

- Conduct research in mining and comparing user behavior patterns [Mar06]
- Assess the information novelty gathered by users. The amount of new information encountered by users is a core value of ES [WR09]
- Measuring the effectiveness [Mun09, LBI*12], e.g., with task success measures. In ES, however, it is also important to evaluate in which way users reach their goals [WR09]
- Measuring the efficiency [TC05, LBI*12]. Measures, such as the task completion time, are an effective way, e.g., to compare ESS with classical query-response tools [WR09].
- Measuring the cognitive and mental load of users is an additional means to evaluate ES, e.g., by counting the number of insights, etc. [WR09]
- Apply user-centered design. To ensure both usable and useful ESS, it is highly appropriate to involve the user in the design [TM04, PVW09]
- Apply a design study method. Research in design study methodology has revealed a variety of approaches for collaborative design projects including case studies and other evaluation strategies [Mun09, SMM12]

2.2. Scientific Primary Data

In this section, we survey primary data, as well as associated concepts, techniques, and user groups. First, we characterize (scientific) primary data in Section 2.2.1, describe its value for data-driven research, and draw a connection between primary data objects and the more general class of 'documents'. Next, we show challenges associated with primary data and outline the solutions provided in this thesis. Section 2.2.2 describes the six phases of the *data life-cycle* for primary data. We present an overview of the data life-cycle and summarize challenges associated with ES. Finally, in Section 2.2.3, we outline *content-based access* as a most motivating concept for this thesis and present strategies for primary data.

2.2.1. Characterization of Primary Data

Scientific *primary data* can be seen as a direct data product collected from a source. Common sources for the collection of primary data are interviews, experiments, observations, computations, or simulations. A variety of application domains exist where primary data is collected. An illustration of different sources of time-oriented primary data was presented in Figure 1.1 in the introduction chapter. In contrast to *secondary data* (derivative data), primary data comprises the original condition of a phenomenon without being processed, transformed, or manipulated other forms. Scientists use primary data to represent complex real-world phenomena and as a means of providing evidence. The unaffected nature of primary data makes it particularly valuable for data-driven research. In some cases, the terms raw data, sensor data, measurement data, scientific data, or research data are used interchangeably. Many of the domains differ in the nature of their data, their methods, and their conventions about data use and reuse [CIZW13, KH13]. From an IV and VA perspective, we refer to Kehrer and Hauser's classification of scientific data [KH13]. The authors differentiate between:

- Spatio-temporal data (spatial structures and/or dynamic processes)
- Multi-variate data (multiple data attributes, e.g., temperature or pressure)
- Multi-modal data (data stemming different sources like CT, MRI, large-scale measurements, etc.)
- Multi-run data (stemming from multiple simulation runs repeated with varied parameter settings)
- Multi-model scenarios (resulting from coupled simulation models, e.g., coupled climate models)

The Value of Primary Data We take a closer look at the benefits of primary data. To start with, primary data may lead to actual, appropriate, and specific data collections for the particular subject/object of interest. This is especially the case when primary data is collected in a scientific context. Here, (scientific) primary data can be seen as the result of experimental procedures, scientific measurements, or other research-driven ventures. However, primary data is not only the product of scientific data collections, but has become scientific capital [BWE06]. Scientific research is associated with the production, analysis, storage, management, and reuse of data [CIZW13]. Primary data plays a key role in a paradigm shift towards data-intensive scientific discovery [HTT09], assigning primary data the basis for data-driven research. We refer to the ongoing discussion about the quote of Clive Humby in 2006: 'data is the new oil' [1]. Being a new source of knowledge [The12], it can be assumed that primary data contains large quantities of *undiscovered knowledge* not yet sufficiently investigated, understood, or researched. This undiscovered knowledge may come to light if meaningful data subsets can be retrieved out of typically large primary data repositories, and if the required analytical capability is available for the scientist for subsequent sense-making. Different *information needs* [Mar95, Mar06] of researchers based on the complexities of their search activities also make primary data attractive for exploratory data analysis. An additional issue that amplifies the value of primary data are Digital Object Identifiers (DOI), which are increasingly attached to primary data, making the data documents citable [Bra04]. The ongoing registration of primary data can be seen as a success for citation and long term integration. The GetInfo [2] portal provided by the German National Library of Science and Technology (TIB) may serve as an example of the utilization of DOI. In addition, scientific data management systems [AKD10] incorporating the *data life-cycle* principle foster the storage and the accessibility of primary data. Existing frameworks provide solutions for data storage, management and dissemination of complex objects and relationships (see, e.g., the Fedora architecture [LPSW06]). Not least, the value of primary data is particularly increased by the synergy effects between primary data and *scientific workflow*

[1] http://ana.blogs.com/maestros/2006/11/data_is_the_new.html, last accessed on September 26th, 2015
[2] GetInfo, German National Library of Science and Technology (TIB), https://getinfo.de/app, last accessed on September 26th, 2015.

systems. The latter allow researchers for the definition and execution of *scientific workflows* on the underlying primary data (see Section 2.4.1 for an overview). Moreover, data sharing is a valuable principle for the scientific progress and the advancement of knowledge discovery. The value of primary data is enhanced by the *open access*, *open data*, and *open science* paradigms proclaimed by various initiatives (see, e.g., [MH10, The12, CIZW13]). The open access paradigm also contributes to an enhanced *reuse* of primary data which in combination facilitates science as an open enterprise [The12]. We refer to the survey of Costas et al. for an overview of open access data repositories, open data journals, and recommendations for involved stakeholders [CIZW13]. Almost any scientific domain provides scientific literature often containing both scientific publications and (open) data publications [CIZW13]. Finally, a large number of (open) primary data collections, data repositories, and data warehouses contribute to an enhanced accessibility of primary data [MH10]. In the following, we present some of the described initiatives.

- *SDSS - Sloan Digital Sky Survey* [3] for physics and astronomy, mapping visible objects in the universe.
- *GigaScience - GigaDB* [4] for life and biomedical sciences
- *DRYAD* [5] digital repository for generic data underlying publications in the natural sciences
- *PANGAEA* [PAN] information system for geo-referenced data from Earth system research
- *ICPSR* - Inter-university Consortium for Political and Social Research [6] for social and behavioral research
- *JOPHD - Journal of Open Public Health Data* [7] for public health data sets
- *DataONE - Data Observation Network for Earth* [8] for environmental science
- *PsychData* [9] for all areas of psychology and social sciences
- *PubChem* [10] free, open database of chemical molecules and information about their biological activities
- *PubMed* [11] more than 24 million citations for biomedical literature
- *HDM05 - Motion Capture Data Base* [MRC*07] for motion capture data base free for research purposes
- *DKRZ - Deutsches Klimarechenzentrum* [12] for climate data
- *RDA - Research Data Alliance* [13] building social and technical bridges that enable sharing of data
- *UK Biobank* [14] various attributes of health data of 500,000 voluntary people (blood, urine, saliva samples, etc.)

We emphasize PANGAEA and HDM05 as the sources of time-oriented primary data applied in this thesis.

Primary Data Objects and Documents in General Regardless the field of research where primary data is collected, it is often stored in digital environments, e.g., scientific data management systems [AKD10]. Subsequent steps applied to primary data, such as the construction and execution of scientific workflows, require an access strategy to primary data collections. The access strategy to primary data complies with the access strategies for electronic *documents* in general, described, e.g., by Marchionini [Mar95]. Documents can be seen as containers carrying the content of single entities, such as primary data objects. Furthermore, documents can contain explanatory metadata (data about data) (see, e.g., [The12]). Prominent standards (schemes) for metadata are the *Dublin Core* [15], or the Data Cite kernel [Bra04]. While Dublin Core is more general, Data Cite is especially targeted towards scientific primary data documents. In many scientific publications the term 'document' is used to describe different primary data objects in a broader sense. For example, documents are stored and handled with data management systems [AKD10, CIZW13], collected, archived and organized in DLs [WBBM00, GMPS00, Bea07, ABB*07, CCF*08, HPK08], or used in IV, VA, and ES methodology [Mar95, Shn96, HK12]. Sometimes the term 'document-like objects' is used to cover the broad range of digital objects (e.g., voice, video, email, images, and data sets) [Bor10, p. 41].

[3] SDSS - Sloan Digital Sky Survey, http://www.sdss.org/, last accessed on September 26th, 2015.
[4] GigaDB, http://gigadb.org/, last accessed September 26th, 2015.
[5] DRYAD, http://datadryad.org/, last accessed on September 26th, 2015.
[6] ICPSR - Inter-university Consortium for Political and Social Research, https://www.icpsr.umich.edu, last accessed on September 26th, 2015.
[7] JOPHD - Journal of Open Public Health Data, http://openhealthdata.metajnl.com/, last accessed on September 26th, 2015.
[8] DataONE - Data Observation Network for Earth, https://www.dataone.org/, last accessed on September 26th, 2015.
[9] PsychData - Center for Research Data in Psychology, http://www.psychdata.de/, last accessed on September 26th, 2015.
[10] PubChem - National Center for Biotechnology Information, https://pubchem.ncbi.nlm.nih.gov/, last accessed on September 26th, 2015.
[11] PubMed - US National Library of Medicine, http://www.ncbi.nlm.nih.gov/pubmed, last accessed on September 26th, 2015.
[12] DKRZ - Deutsches Klimarechenzentrum ,https://www.dkrz.de/daten, last accessed on September 26th, 2015.
[13] RDA - Research Data Alliance, https://rd-alliance.org/, last accessed on September 26th, 2015.
[14] UK Biobank, http://www.ukbiobank.ac.uk/, last accessed on September 26th, 2015.
[15] The Dublin Core Metadata Initiative http://dublincore.org, last accessed on September 27th, 2015.

Defining Data Content, Metadata, Attributes, Entities, and Mixed Data In this thesis, we subdivide primary data objects into their data content and their explanatory metadata. The differentiation between data content and metadata is of particular importance. In Chapters 4 and 5, we contribute guidelines and techniques considering the (time-oriented) data content, while Chapter 6 contributes techniques for revealing relations between the (time-oriented) data content and metadata. In many cases, the distinction between data content and metadata is made by the data creators or the data curators. However, a general distinction between data content and metadata is not obvious in many cases, and may depend on the eye of the beholder. One approach is the referral of all 'intrinsic' primary values of a data source as the data content. Thus, secondary data and other attached information about the data would be metadata. An additional perspective is also referring secondary data to as the data content. The rationales are that secondary still reflects the data structure of the primary data and carries large parts of the information of the primary data content. Similarly, in the context of library systems, all information stemming from the source of a document may be defined as the content in general, while all information, additionally attached for library use, may be defined as metadata. Yet another approach focuses on a specific type of data (such as time-oriented data) which is subsequently referred to as the data content. Similarly, a distinguishing criterion may be the textual or non-textual type of document parts. From an IR perspective (see, e.g., [BYRN*99]), the data content may be the equivalent to the data subset where the retrieval strategy is applied to, again irrespective of whether it is based on primary, secondary data, or based on attached metadata. Depending on the analysis goal of researchers, the division into data content and explanatory metadata may also be subject to change to test new hypotheses, e.g., when dependent and independent variables are shifted between various attributes of a primary data set. Finally, a variety of approaches in IV and VA completely abstain from the distinction, but rather incorporate both data content and metadata in the analysis simultaneously.

Mixed data consists of attributes with different data types. Mixed data sets comprise quantitative and qualitative (numerical and nominal) attributes. Consequently, mixed data may carry both data content and metadata. Prominent approaches for the exploratory analysis of mixed data sets are, e.g., [KBH06, JJJ08, LSS*11]. The analysis of mixed data raises the question whether the differentiation between data content and explanatory metadata is even necessary, from an analytical perspective. Our contributions presented in Chapter 6 resolve challenges of how ESS can be equipped with techniques for revealing relations between the time-oriented data content and metadata. Two techniques explicitly focus on data content and metadata (see Sections 6.3 and 6.4). The third contribution emphasizes the exploratory analysis for mixed data possibly including both data content and metadata (see Section 6.5).

Throughout this thesis, we describe the time-oriented primary data and all secondary data (products of transformations on the time-oriented data) as the *data content*. We divide the time-oriented data content into the temporal domain (the characterization of the temporal component) and into the value domain (consisting of progressions of univariate, bivariate, or multivariate values). The VisInfo case study presents an ESS for *univariate* time-oriented data (see Section 7.1). Various usage scenarios presented in Section 5.3 are based on *bivariate* time-oriented data. The MotionExplorer case study presents an ESS for *multivariate* time-oriented data (see Section 7.2). In Section 2.3, we characterize time-oriented data in detail.

We define the attached information about the data and the provenance information gained within the transformation and analysis process as *metadata*. Metadata consists of (various) *attributes* each describing an individual property of the data, e.g., the 'location on Earth', the 'month of the measurement', or the 'sensor device'. Every metadata attribute consists of at least two different *entities* (bins, categories, values) defining the observations made for an attribute. Examples are 'spring', 'summer', 'autumn', and 'winter' for the metadata attribute 'season'. In addition, we describe *mixed data* as the data type where both data content and metadata can be combined and be used for exploratory analysis tasks, such as seeking relations.

Challenges in the Use of Primary Data for carrying out 'science as an open enterprise', data should be accessible, intelligible, assessable, usable [The12]. The exploitation of the value of primary data, however, coexists with a variety of challenges. In general, coping with (time-oriented) primary data requires dealing with at least four major challenges, i.e., the *heterogeneity* (diversity) of primary data, the *quality* of primary data, the *size* of large data collections, and the *time-varying* behavior of the data values. Solving challenges posed by the time-varying behavior of primary data are at heart of this thesis. We discuss challenges associated with the time-orientation in detail when we characterize time-oriented data in Section 2.3.

The *heterogeneity* of primary data is a great challenge for many data-driven research approaches [BWE06, AKD10, CIZW13]. Difficulties exist in the definition of data standards for individual data types. In addition, the dependencies to specific experiment conditions make a 'unification' of different primary data sources difficult. We refer to Keim et al. [KMSZ06] where the 'synthesis of heterogeneous types of data' is described as a main technical challenge for VA research. Moreover, Kehrer and Hauser's taxonomy of different types of heterogeneous scientific data clarifies

the difficulty of the problem [KH13]. We come back to this challenge when we present techniques for a combined analysis of data content and metadata, which is one of the three technical contributions described in this thesis (see Chapter 6). An aspect that adds to the challenge of data heterogeneity is the existence of different formats for identical data values. Prominent examples are different units for physical quantities, such as temperatures (Farnheit, Celsius) or distances (miles, kilometers). Yet another factor of different formats is posed by the data structures the primary data are stored in. These types of differences can be on the granularity of different separators between values in the data (comma, tab), or on the specification of the data structure in general (e.g., ISO standards). Being direct product from a source, primary data may not even comply with standardized data structures or formats.

This leads the discussion to the *quality* of primary data, which is an important concern on its own. In contrast to secondary data, which may be a processed (condensed, cleaned, distilled) derivate with a certain extent of quality management, primary data is a raw type containing potentially many types of (unexpected) quality leaks. The data quality depends on the environment, the sensor, or the measurement technique. Also human activities have an impact on the quality, e.g., when human judgment is involved in the data gathering process. In the context of time-oriented data, a taxonomy of 'dirty' data exists, which is beneficial for the characterization and selection of 'data cleansing' strategies [GGAM12]. For the application of sophisticated analysis capability it is crucial to meet certain data quality aspects. It should be frequent practice that data quality aspects are established in a series of additional preprocessing steps [KHP*11]. However, the use cases and usability examples presented in a vast number of today's approaches consider clean data as a prerequisite. Thus, the data cleansing process is not part of the proposed solution. In this thesis, we present the process of establishing quality standards for time-oriented primary data in one of three technical contributions (see Chapter 4). We discuss guidance concepts to support users in establishing data quality criteria and apply respective techniques in a visual-interactive system.

Finally, the *size* of today's collections contributes to the *complexity* of primary data, and thus to the challenges associated with primary data. As Borgman et al. state: new technologies for collecting data are leading to data production at rates that exceed scientists abilities to analyze, interpret, and draw conclusions [BWE06]. Nowadays the storage of large primary data volumes is feasible in most cases. However, dealing with this 'information overload' problem [KMSZ06] is still one of the most difficult challenges for data-driven research. One technical contribution of this thesis overcomes this challenge with guidelines and techniques for the design of content-based overviews (see Chapter 5). Furthermore, in both case studies of this thesis, we demonstrate how the integration of data abstractions in ESS supports scientists in coping with large primary data volumes (see Sections 7.1 and 7.2).

2.2.2. The Data Life-Cycle of Primary Data

Introduction We take a closer look at the different phases in the use and reuse of primary data. The *data life-cycle* is a conceptual representation of the data flow through different phases, from its creation to its long-term archival or, in some cases, its destruction. The data life-cycle influences many infrastructural and environmental elements in data-driven research. Prominent examples are 'e-science' environments distributed over the Internet [DGST09] (cyberinfrastructures) [BWE06], DLs [HTT09, FGS12], scientific data management systems [AKD10], or scientific workflow systems [DF08]. The concept of the data life-cycle contributes, and likewise benefits from the latter elements. Successively, the compliance with the life-cycle has auxiliary means for the value of primary data. In most works, the basic principles of the data life-cycle are characterized quite similarly. However, for individual phases of the life-cycle the terms used and the assignment of functionality varies in some cases. In principal, six different phases are passed through. We present a schematic illustration of a typical data life-cycle in Figure 2.2. The illustration condenses the information presented in related works [BWE06, AKD10, CIZW13] and the following organizations: Boston University Library [16], University of Pittsburg [17], DataONE [18], and DKRZ [19]. The six phases are the *creation*, the *processing*, the *analysis*, the *preservation*, the *access*, and the *reuse* of primary data. Some conceptual representations of the data life-cycle differ in some phases. For example, sometimes a *planning* step for the creation of primary data is added. Another example regards secure data life-cycles where typically an additional *deletion* phase is mandatory. An additional distinction can be made between the first and the last three phases of the data life-cycle. The first three phases are primarily applied by the domains where the data is created. The last three phases, however, can also be executed by external domains for data access and reuse. The latter aspect is most motivating for library treatment and also for this thesis.

[16] Boston University Libraries, Research Data Management - Data Life Cycle, http://www.bu.edu/datamanagement/background/data-life-cycle/, last accessed on September 24th, 2015.
[17] University of Pittsburg, Health sciences library system, http://info.hsls.pitt.edu/updatereport/?p=6057, last accessed on September 24th, 2015.
[18] DataONE - Data Observation Network for Earth, Best practices, https://www.dataone.org/best-practices, last accessed on September 24th, 2015.
[19] DKRZ - Deutsches Klimarechenzentrum, https://www.dkrz.de/daten, last accessed on September 26th, 2015.

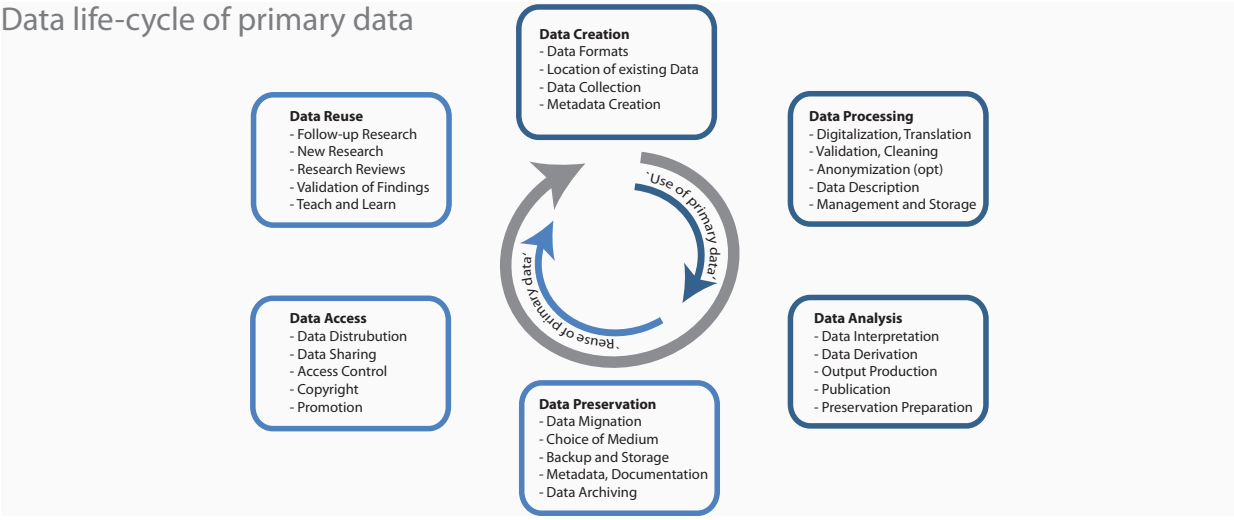


Figure 2.2 *The data life-cycle of primary data, assembled from related works [BWE06, AKD10, CIZW13]. In six phases the data is created, processed, analyzed, preserved, accessed, and reused. In many cases, the first and the last three phases can further be condensed to the 'use' and the 'reuse of primary data.*

Phases of the Data Life-Cycle We review the six phases of the data life-cycle. We also refer to the works of Borgmann et al. [BWE06], Ailamaki et al. [AKD10], and Costas et al. [CIZW13] for overviews of different phases.

In the data *creation* phase, primary data is collected and provided for subsequent processing. As an example, the primary data in the case studies of this thesis stem from human motion capture devices, as well as from series of sensors at stations for Earth observation. For the design and the use of ESS for time-oriented primary data, we make no assumptions or restrictions to particular domains, sensor techniques, or data standards. We refer to the latter Section 2.2.1 where characteristics of primary data, relevant for the creation phase, are described in detail.

Actions performed in the data *processing* phase are the digitalization and translation of primary data, specific data curation operations, or optional anonymization steps. In addition, achieving data quality standards is an important concern. One of the stakeholders deeply engaged in this phase are data curators, also called resource managers in the DL domain [CCF*08]. As a result of this step, the primary data is stored in data repositories, and administrated by scientific data management systems [AKD10]. Managing the enormous amounts of primary data collected from different sources is crucial for the scientific progress [MH10, AKD10].

The data *analysis* phase is of particular importance for this thesis. The techniques presented in our three technical contributions and the two ESS in our case studies enable domain experts carrying out (exploratory) data analysis. From a high-level perspective, our techniques support a) the drill-down process to sets of relevant documents possibly distributed over several repositories and b) the enhanced knowledge discovery process in data-driven research. In the remainder of the related work chapter, we shed light on the analysis phase from the perspectives of underlying data (see Section 2.3), exploratory search tasks (see Section 2.1), content-based access (see Section 2.2.3), workflows for carrying out data analysis (see Section 2.4), and roles of involved users (see Section 2.5).

The technical details of long-term archiving in the data *preservation* phase, like using robust storage technology [AKD10], are beyond the scope of this thesis. Similar to the techniques applied in the data processing phase, the data preservation phase is also related to data repository and data management technology [AKD10]. In the first place, scientists archive their data to facilitate a possible re-examination (reuse) in the future, which is a motivating aspect for this thesis. Moreover, these repositories may create and store additional metadata, provenance information, and result documentations gained in the analysis phase. Our techniques incorporate this information as an additional input to facilitate the reuse of primary data, e.g., with the combined analysis of data content and metadata (see Chapter 6). Similar to the techniques applied in the data processing phase, the data preservation phase is also related to data repository and data management technology [AKD10].

The data *access* phase has a strong interrelationship with this thesis. On the one hand, the accessibility of primary data must be provided from a financial and political perspective. This is a precondition for the execution of ES techniques and for the incorporation of respective primary data for reuse. On the other hand, the accessibility of large sets of primary data must be feasible from a data complexity point of view [TC06, vLSFK12] [HTT09, p. 114]. Hence,

an increase in efficiency and effectiveness for accessing large data repositories with ES techniques may contribute to the data access, and foster the public accessibility of primary data. In this connection, we also refer to Section 2.2.3 on content-based access. Effective data management plays a key role for providing (content-based) access to primary data. Scientific data management aims at providing technology for data storage, DM, data transformation, data access, and data indexing, to mention the most relevant [AKD10]. To cope with the heterogeneous technical requirements different stakeholders are involved in data management infrastructures (see, e.g., the DELOS reference model [CCF*08]). Relevant stakeholders are data curators, database managers, data scientists, DL designers, and domain experts. Data warehouses and data repositories [MH10], and DLs [FGS12] are environments associated with scientific data management infrastructures. Moreover, scientific data management serves as a base technology for scientific workflows and scientific workflow systems, such as the Kepler system [LAB*06]. In this thesis, two main contributions enhance the *content-based access* to the time-oriented primary data (cf. Section 4) and the *visual access* to the data content with content-based overviews (cf. Section 5).

The data *reuse* phase completes the data life-cycle. Reuse of primary data is desirable to increase the transparency of research and research results, and to lower the cost of research by sharing of data [CIZW13]. The cost of preserving data is often considerably smaller than the costs of collecting new data [The12]. For many time-oriented primary data collections the reuse of earlier measurements is sort of obligatory, since measurements depending on absolute time cannot be replicated. In addition, many research goals drawing on time-oriented primary data are targeted towards the temporal comparison, e.g., for the identification of temporal developments. As a consequence, analytical support is required leading to an effective reuse of primary data. Objectives like validating data-driven research results, or carrying out new analytical activity based on data reuse directly associate the data reuse phase with the data analysis phase described earlier. Due to the particular importance of data reuse for this thesis, we recall sections in the related works chapter associated with data reuse. We shed light on the perspectives of the focused data type (see Section 2.3), exploratory search tasks (see Section 2.1), content-based access (see Section 2.2.3), workflows for carrying out data analysis and reuse (see Section 2.4), and involved users (see Section 2.5).

Open Research Questions within the Data Life-Cycle We conclude the review of the data life-cycle with an outline of research challenges. For a broader picture, we refer to the works of Borgman et al. [BWE06], Marcial and Hemminger [MH10], Ailamaki et al. [AKD10], and Costas et al. [CIZW13].

One research question in the data life-cycle relates to the *different involved stakeholders*. We already listed data curators, database managers, data scientists, DL designers, and domain experts explicitly. Moreover, other stakeholders may exist, e.g. with a political or financial stance [Sub12, CIZW13]. On the one hand, data-driven research typically involves the expertise of these different stakeholders in the course of the data life-cycle. On the other hand, research may also be carried out in cross-domain projects [AKD10], e.g., by incorporating various domain experts from biology, chemistry, and medicine. In both described scenarios, gaps exist on the knowledge-level, but also on the system-level. Examples for the latter are missing data standards [BWE06], data abstractions [Fek13], or visual interfaces [RBK13]. In this connection, we refer to the recent work of Jean-Daniel Fekete. Fekete outlines the need for more collaboration among researchers and practitioners across the visualization, analytics, and the data management domains [Fek13]. This issue is discussed in the related work section on user-centered design in detail (see Section 2.5).

The complexity of time-oriented primary data impedes *finding relevant data* (sub-)sets. As a result of the 'data deluge', scientists urgently require assistance to identify and select data for current use [BWE06]. Presuming the data was shared (on the Web), studies report that difficulties exist in locating relevant information [MH10]. The reasons may be due to isolated (or unknown) repositories which need to be linked together, in insufficiently curated or indexed metadata, or in search capability which does not comply with the different information needs and the exploratory seeking behavior of researchers [Mar06]. Since the identification and extraction of relevant data is one of the classical activities of researchers, supporting such activities is important [BWE06], and most motivating for this thesis. Finding data is (still) performed by systems based on *fact retrieval* [Mar95, p. 29 ff.] and *known-item search* [WR09, p. 14]. As a rule, having a clear notion of targeted results within large search spaces is indispensable to perform such search procedures successfully. However, since the sense-making process in data-driven research is to large parts exploratory, known fact retrieval systems often are inadequate. We refer to Section 2.1.1 for an overview information-seeking behaviors.

The complexity of time-oriented primary data not only causes challenges for finding data, but also for *exploring* and *analyzing* data. Data-driven research requires an efficient and effective content-based access to relevant primary data, which is cumbersome for large, heterogeneous, low-quality, and time-varying primary data sets. Not only the data access, but also sophisticated analytical capability is necessary to perform in-depth analysis of huge amounts of data [AKD10]. Automated data processing would be an efficient means for coping with the complexity, and

thus for the analysis of primary data. However, the complexity of primary data makes the automation difficult, in particular. The phenomena described by the data are extremely complex and require significant time and effort to understand [PVW09]. Automated data processing is only reasonable in scenarios where the data is entirely understood and no human judgment is needed [KMS*08]. Finally, sophisticated visual-interactive interfaces would support scientists in the exploitation of the value of scientific data collections. One promising method to resolve these challenges is the collaboration among experts from IV, VA, and data management [Fek13]. In this way, visual access to the data content can be provided, and thus the exploration of data be supported. Jean-Daniel Fekete also emphasizes the importance of the workflow construction as a cascade of operations that filter, summarize, and analyze/explore the data (visually) [Fek13]. We review concepts and techniques for the construction of (scientific) workflows in Section 2.4. One of the main contributions of this thesis is the visual-interactive support for the guided construction of preprocessing workflows for time-oriented primary data (see Chapter 4).

2.2.3. Content-Based Access to Primary Data

Providing access to documents via the data content, e.g., to (time-oriented) primary data, is a key feature to exploit the potential value of the given document collection. Regardless the quality and the amount of provided metadata, large quantities of information can be expected in the (time-oriented) data content. Content-based access to primary data raises challenges, as described in the characterization of primary data in Section 2.2.1. In the related work section about time-series data, we describe specific challenges associated with the time-oriented data content (see Section 2.3). In this section, we first define content-based access for this thesis. Second, we describe content-based access strategies from the perspectives of three stakeholders *data scientists*, *domain experts*, and *librarians*.

Defining Content-Based Access *Content-Based Access (CBA)* is the principle of accessing documents and document collections by means of their data content. To exploit the data content as an access strategy, the potentially complex data content itself needs to be represented in a format enabling effective and efficient computational processing. The predominant representation of various types of data content are so-called FVs, often characterized as compact and yet faithful representations of complex objects. Downstream operations can be applied to use CBA strategies. Example operations are search functionalities (e.g., Query-by-Sketch or Query-by-Example) and exploration functionalities, such as content-based overviews. Challenging tasks for providing CBA strategies are the identification and extraction of appropriate features, as well as addressing the gap between the notion of similarity in the heads of domain experts and computable numerical similarity measures. CBA is subject to active research in many application domains, such as DLs and IR. While CBA strategies to textual content play to some extent a pioneering role, research on CBA strategies to non-textual documents is an emerging field for various data types. Examples are audio, image, and video content (often referred to as multimedia), as well as other complex types of data objects, such as time-oriented data or primary data.

Data Scientists working in research fields, such as KDD, IR, IV, and VA are used to access the data content in their daily work to improve their analytical capability. A great variety of models, techniques, visualizations, and interaction designs has been researched today. Some of them are already employed in analytical systems in the everyday life of the society:

- finviz [20], maps and scatterplots showing stock market data
- Google Maps [21], for geo-referenced data
- Many Eyes [22], a platform where people can upload data sets and create interactive visualizations
- Tableau [23], easy-to-use software for the creation of lightweight IV solutions

However, these best-practice examples do hardly rely on (time-oriented) primary data. Today, data scientists still are confronted with a variety of factors making the design of individual content-based access solutions for time-oriented primary data challenging. For primary data, in particular, the data quality and the different types of complexity are most influencing factors (cf. Section 2.2.1). Ben Shneiderman outlined seven basic types of data (1-dimensional,

[20] finviz - Financial Visualizations, https://finviz.com/map.ashx, last accessed on September 23th, 2015.
[21] Google Maps, https://www.google.de/maps/, last accessed on September 23th, 2015.
[22] Google Maps, http://www-969.ibm.com/software/analytics/manyeyes/, last accessed on September 23th, 2015.
[23] Tableau, http://www.tableau.com/, last accessed on September 23th, 2015.

2-dimensional, 3-dimensional, multi-dimensional, tree-based, network-based, and time-oriented data), all of which expect individual treatment [Shn96] (cf. Section 2.3 for a survey on individual properties of time-oriented data). Aside from data characteristics, the data scientists' content-based access solutions typically rely on requirements defined by external application domains. As already outlined, domain experts carry out varieties of analysis tasks in different application fields. These tasks need to be considered carefully for the content-based access strategy (see Section 2.5 for a review of user-centered design). Regardless the design parameters and the series of challenges in content-based access, some inspiring solutions for time-oriented primary data have been presented in the past. For instance, we refer to the approaches presented in the Earth observation domain by Steinbach et al. [STK*03], Nocke et al. [NSBW08], Kehrer et al. [KLM*08], and Tominski et al. [TDN11]. In the VisInfo case study in Chapter 7, we come back to the Earth observation domain.

Domain experts involved in data-driven research obviously have an interest in content-based access. In their field of research, domain experts are typically involved in large parts of the data life-cycle (see Section 2.2.2). In particular, domain experts may take part in the creation process of (time-oriented) primary data, e.g., in conducting experiments. In addition, domain experts may play an important role in the data curation process. Thus, domain experts are often involved in the early phases of the data life-cycle. Domain experts also play an active role in the analysis, preservation, access and reuse phase. Within the later phases of the data life-cycle, the data content serves as the basis for scientific workflows. Domain experts create scientific workflows for the identification of new information and the discovery of new insight (cf. Section 2.4.1). However, domain experts still underestimate the value of advanced visual-interactive steps within their scientific workflows [NSBW08,TDN11]. Designing sophisticated analytical support to solve complex domain-specific problems is neither the core responsibility of domain experts, nor their core area of competence. This particularly applies to research in content-based access strategies. Domain experts often rely on basic toolkits available in their field of research, accepting the fact hat their data access and analysis process may be tedious in comparison to sophisticated external analytical capability. As a result, many researchers execute at least parts of their scientific workflows manually. The missing 'trust' of domain experts in unfamiliar but possibly meaningful techniques, e.g., exposed by data scientists, is one possible reason. Refusing changes in their working routines yet constitutes another cause [BWE06, HPK08, KMS*08]. A most motivating implication for this thesis is the need for collaborative approaches where domain experts and data scientists work together for the construction of sophisticated workflows [vW06, SMM12]. A motivating example of the VisInfo case study presented in Chapter 7 is the PanPlot [24] tool. PanPlot supports the visualization of time-oriented primary data exported from the PANGAEA [PAN] repository. The standalone tool is beneficial for the visualization of single data sets covering a time duration of one month at one measurement station for Earth observation. Remaining challenges in the researchers' workflow are *searching* for relevant data sets in large repositories for the detailed visualization in PanPlot, as well as *exploratory analysis* capability to identify, compare, or relate patterns in large data collections.

Librarians and related stakeholders represent another type of stakeholder relevant for this thesis. Librarians and library designers need content-based data access to support the functionality of late phases of the data life-cycle. The motivation of librarians is to support users in the identification of relevant subsets, and thus to facilitate data access and reuse. To exploit the potential value of primary data, information seekers typically steer their attention to relevant data subsets corresponding to their *information needs* [MS88, AS94, Mar95, GMPS00, Mar06]. Due to the complexity of primary data, but also to the exploratory nature of data-driven research, visual-interactive search and exploration concepts are particularly suited to support DL systems for time-oriented primary data. We refer to the book of Marti A. Hearst for an overview of search systems and search user interfaces [Hea09], a characterization of DLs is presented in the VisInfo case study in Section 7.1.1.

Metadata has ever played an important role in DL systems facilitating search and retrieval. Accompanied with that, a most widespread querying method is based on textual queries. Still, entering text in query dialogs is a common subject of search practice, especially for known-item search and fact retrieval. With new types of data content emerging in DL research, librarians are encouraged to access the data content to enhance library support. Content-based access should not only be applied to processing and analyzing primary data, but also for providing (exploratory) search operations. Ideally, DL systems facilitating ES operations might be an integral part of the scientific workflows of domain experts. The sophisticated incorporation of the data content into the information-seeking process can expedite the search tremendously, not only from an IR [BYRN*99], but also from a visual search perspective [HK12]. However, experts in search systems and library services are neither experts in data science, nor experts in the targeted domains

[24] PanPlot, PanPlot2, visualization of data versus time or space, http://wiki.pangaea.de/wiki/PanPlot, last accessed on September 28th, 2015.

for library services. To some extent this is a tragic situation since librarians urgently require the expertise of data scientists and domain experts to provide enhanced search support. In the following, we present a brief overview of advances for library services facilitating content-based access.

Textual content is maybe the most widespread data type used for search and retrieval techniques in the DL domain. Varieties of full-text search techniques based on complex indexing mechanisms enable users to efficiently retrieve documents. Baseline web search engines may serve as popular example solutions, indexing the textual content of millions of websites. In addition, the Lucene [25] framework may serve as an example of a text indexing and retrieval, from a data scientist's perspective. Finally, applications with enhanced visual search and exploration support for large textual document collections exist, all of them mastering content-based access. An overview of ESS for text-based corpora is presented by Herrmannova and Knoth [HK12]. Text, however, is only one basic data type among 1-dimensional, 2-dimensional, 3-dimensional, multi-dimensional, tree-based, network-based, and time-oriented data [Shn96]. Thus, DL interfaces for visual search and exploration should not be limited to textual content. However, the progress made in content-based access for *non-textual* documents in general is still not reflected by DL systems to the same extent. Positive examples exist for image, or audio-visual data content [MPR02, SN11]. In these cases, results from multimedia processing and multimedia retrieval are used to facilitate content-based visual search. We refer to a content-based retrieval system for the reuse of human motion capture data [MRC05] as an example relevant for the MotionExplorer case study presented in Section 7.2. However, for other complex non-textual data types only few approaches have been equipped with visual search capability. In particular, the DL domain has seen comparably few content-based visual search approaches for time-oriented primary data to this day; a motivating aspect for this thesis. Examples for other (complex) data types where content-based access solutions are presented in visual search scenarios are architectural model data [BBC*10], multivariate research data [Sch13], or systems providing varieties of data types including 3D data [ABB*07]. In the basic research field for time-oriented data, progress was made in the past as we will show in Section 2.3. However, content-based visual search and exploration systems for time-oriented data have not been exhaustive subject of practice in the DL domain. In the same manner, content-based visual search systems supporting data-driven research for time-oriented primary data are still an emerging research topic.

2.3. Time-Oriented Data

Time has ever played an important role in human history, just like the observation of time-oriented phenomena. We have already emphasized the value of time-oriented primary data in the age of data-driven research associated with the fourth paradigm of science [HTT09]. Within the last decades, the research in KDD, IV, and VA has made tremendous progress in the visual analysis of time-oriented data [AA06, AMST11]. An early best-practice example of the visualization of time-oriented data is the visualization of New York's Weather in 1980 presented in the New York Times (see Figure 1.2). Progressions of different attributes (temperature, precipitation and humidity) are represented in a single visualization. Different temporal analysis tasks are supported, such as the comparison of monthly patterns, or seeking relations between the attributes.

In this section, we review time-oriented data, as well as associated analysis tasks and visual-interactive analysis techniques. We start with a characterization of time-oriented data in Section 2.3.1, followed by an overview of temporal analysis tasks in Section 2.3.2 from the perspectives KDD, time series DM, IV, and VA. Finally, Section 2.3.3 provides an overview of techniques and applications for the search in and the exploration of time-oriented data.

2.3.1. Characterization of Time-Oriented Data

A characteristic property of the data focused in this thesis is its dependency on time. Time-oriented data is among the most relevant and most often applied data types. Time-oriented data has a variety of intrinsic characteristics that expect specific treatment [Shn96, AA06, AMST11, GGAM12]. In general, time-oriented data can be described as a set of tuples (objects, elements, items, records). Every tuple consists of a temporal element (temporal domain) and at least one data value (value domain). The set of tuples with necessarily unique temporal elements (primitives) is ordered by the temporal domain [AA06]. In return, the value domain of the tuples contains the information about the phenomenon, related to time.

In the following, we review characteristic properties of the temporal domain, the value domain, and relations between time and data. For both domains different mathematical properties, design aspects and abstractions exist

[25] Lucene - The Apache Lucene open-source search software, http://lucene.apache.org/, last accessed on September 17th, 2015.

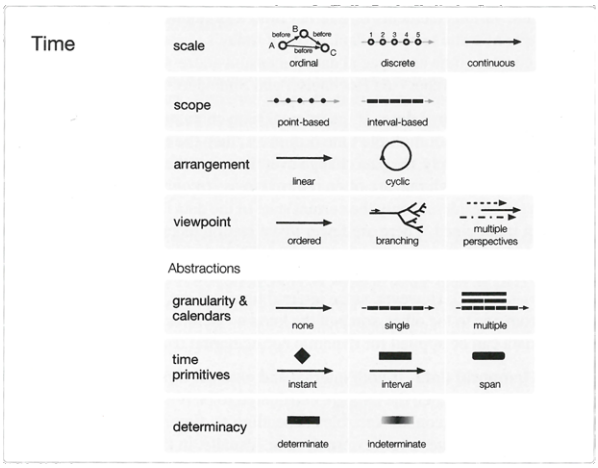
which need to be taken into consideration [AA06, AMST11]. Examples are discrete or continuous temporal domains, quantitative or qualitative value domains, as well as univariate, or multivariate value domains. Taking these individual properties into account is most relevant for the design of appropriate data models, processing time-oriented data, and time series visualization. Our characterization of time-oriented data corresponds with earlier surveys on time-oriented data presented by Fabian Mörchen [Mör06] and Aigner et al. [AMST11, pp. 45-68]. We particularly recommend the in-depth descriptions of modeling time, characterizing data, and relating data and time.

Modeling Time We start with an overview of different characteristics of the *temporal domain* (see Figure 2.3a). One magnitude for the characterization of time is the *scale* which can be ordinal, discrete, and continuous. In an ordinal temporal model the time is described in relative terms (e.g., before, after) [Mör06]. Discrete time domains are applicable for the definition of quantifiable distances measured in integers (e.g., $3,600,000ms$). For continuous time models, between any two points in time, there always exists an additional point. The scale is also relevant for the visualization of time-oriented data. Ordinal models enable the indication of relations between elements. Discrete models can be mapped to integers, while in continuous models the elements can be mapped to real numbers. The scale of both case studies focused in this thesis (Earth observation and human motion capture data, cf. Chapter 7) are modeled discrete with a high resolution allowing a quasi-continuous interpretation. A second magnitude for modeling time regards the *scope* of the temporal elements. Point-based temporal models have an expansion in time equal to zero (e.g., 11th September 2001). Interval-based temporal models relate to subsections of time with a temporal extent greater than zero (e.g., 11th. September 2001 00:00:00 to 11th September 2001 23:59:59). The scope of the data used in our case studies is point-based. A third magnitude of modeling time is the *arrangement* which can be linear and cyclic. Linear temporal models consider time proceeding constantly from past to future. Cyclic models consist of a set of recurring temporal patterns (e.g., seasons of a year). Similarly, the concept periodicity can be described as a repetition of at least similar values with an almost constant time period [Mör06]. The arrangement of time-oriented data used in our case studies varies between linear and cyclic. In this connection, the definition of (cyclic) temporal patterns is one of the most relevant design decisions. As an example, the content-based overview in VisInfo case study (cf. Section 7.1) builds up on daily patterns covering instances of recurrent phenomena.

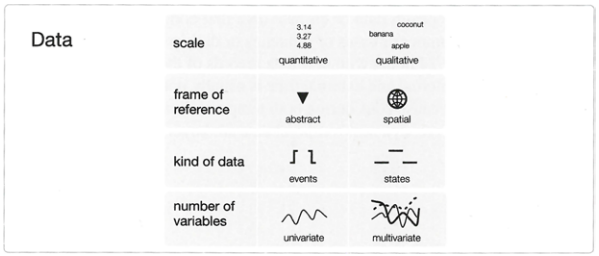
For many temporal data models it is important to consider different *granularities* of time. This is especially necessary if the temporal model also needs to provide 'date-like' characteristics in calendar systems (e.g., days and weeks). Typical temporal models with different granularities are based on a discrete time domain with smallest possible units (chronons) [Mör06] (e.g., days). In addition, these temporal models consist of coarser granularities on top of the discrete time domain (e.g., weeks or fortnights) [ABM*07]. A specific problem in modeling multiple granularities is the existence of irregular mappings between different granularities (e.g., days and months). *Time primitives* are different types of basic elements of the temporal domain used to relate data to time. We distinguish absolute and relative points in time. Absolute points are located at a fixed position along the time domain, this type can further be distinguished in instants and intervals [AMST11]. Relative (span) time primitives have no absolute position in time. For the different time primitives different relations exist. For example, the three existing relations for instants are *before*, *equals*, and *after*. We refer to [Mör06, AMST11] for characterizations of more complex relations for intervals. In this thesis, we use the hierarchical organization of time for the extraction of temporal metadata. In the VisInfo case study presented in Section 7.1, we add the *Month*, the *Season*, and the *Year* of a measurement to the set of metadata attributes. In this way, we are able to identify interesting relations between daily time-oriented patterns and metadata attributes. In this connection, we also support hypotheses testing where the temporal domain does not necessarily represent the dependent variable.

Characterizing Data In this section, we survey different characteristics of the *value domain*. The value domain considers the data values associated with time primitives (see Figure 2.3b).

The *scale* of the value domain can be distinguished in qualitative and quantitative variables. Qualitative values describe either unordered (nominal) or ordered (ordinal) sets of elements. Quantitative (numerical) values are based on a metric scale (discrete or continuous). Numerical data models are commonly used in statistics [Mör06]. Symbolical (non-numerical, qualitative) data models are commonly used in bioinformatics, e.g., for the analysis of gene data [Mör06]. Symbolical data models can be obtained from numerical data models by binning or data aggregation approaches, the Symbolic Aggregate Approximation (SAX) descriptor [LKLC03] may serve as an example. Prominent approaches for non-numerical (symbolic) time series are the VisTree tool for monitoring and mining symbolic time series [LKL*04] and the time series bitmaps technique for the visual comparison of mosaic-based glyphs [KLK*05]. In this thesis, we focus on time-oriented data with quantitative value domains, i.e., our data models support numerical value domains.



(a) *Design aspects of the temporal (time) domain.*



(b) *Design aspects of the value (data) domain.*

(c) *Relating the data domain and the time domain.*

Figure 2.3 *Modeling time-oriented data. Characteristics of the time domain and the data domain, as well as relations between data and time. Illustration by Aigner et al. [AMST11] (used with permission).*

The *frame of reference* regards the aspect whether data values are connected to a *spatial* location, or the data values are *abstract* [AA06]. The data models used throughout this thesis have no restrictions for the frame of reference. In the VisInfo case study in Section 7.1, we include spatial primary data from over 50 measurement stations all over the world. We preserve the spatial information in additional metadata attributes (*Location*, *Longitude*, *Latitude*, and *Hemisphere*) and use the univariate time-oriented data in an abstract context, e.g., for the design of content-based overviews (cf. Chapter 5). In addition, we show how the spatial information of time-oriented data can be used to facilitate relation-seeking support. In the MotionExplorer case study in Section 7.2, we work with motion capture data. Multiple 3D markers capture spatial locations in absolute coordinates. We use the spatial information of 3D markers for glyph designs of human poses, while the content-based overview solution is based on a data abstraction.

For the interpretation of the value domain it is important to distinguish between states and events (*kind of data*). Events depict the markers of state changes while states are typically described as a continuity between events. Finally, a distinctive property of the value domain is the *number of variables* (dimensionality). We distinguish between univariate, bivariate and multivariate domains.

Relating the Temporal Domain and the Value Domain We next characterize the *relation* between time and data. Aigner et al. distinguish between the *internal time* and the *external time* [AMST11] (see Figure 2.3b). The internal time is the time inherent in the data model. If the data model provides values that change over time the data is *temporal* (dependent on time), otherwise the data is non-temporal. The external time is the time extrinsic to the data model. The external time is *dynamic* if the data (not merely the data values) change over time, otherwise the data is static. *Static non-temporal data* does neither contain changes of the value domain over time, nor contain changes of the data itself. Thus, static non-temporal data is out of the scope for this thesis. *Static temporal data* consist of a value domain with multiple internal time primitives. This type of data is dependent on time. Approaches based on static temporal data relations are most common for the visualization of time-oriented data [AMST11]. Static temporal data models support values changing over time, while the data itself is static. Both of our case studies are based on static temporal data, while our techniques are not limited to static temporal data. *Dynamic non-temporal data* provides multiple external time primitives, i.e., the data depend on the external time. Streaming data is a typical relation of dynamic non-temporal data drawing on tasks, such as visual monitoring. In many cases, it is possible to map static temporal to dynamic non-temporal data models [AMST11]. *Dynamic temporal data* comprise multiple time primitives of the internal and the external time. In this case, internal and external time points are strongly coupled and can often be mapped into each other [AMST11]. Dynamic temporal data also influences the techniques and ESS presented in this thesis. In the VisInfo case study, we support the ES of daily Earth observation patterns. The curve progression within each day corresponds to the internal time while the large set of daily patterns represents the dependency to the external time. In combination with metadata attributes, our techniques enable relation seeking between both internal and external time dependencies. In the MotionExplorer case study, we analyze the motion capture data of artists performing multiple

Techniques in the KDD workflow	Taxonomies of Tasks/Techniques	Search	Explor.
KDD Reference Workflow			
├─Preprocessing			
│ ├─Cleansing	[Mör06, GGAM12]	yes	yes
│ └─Wrangling	[Mör06, KHP*11]	yes	yes
├─Transformation			
│ ├─Representation, Descriptors	[KK03, LKLC03, Mör06, Fu11]	yes	yes
│ └─Similarity	[KK03, Mör06, Fu11]	yes	yes
├─Data Mining			
│ ├─Indexing, Retrieval by Content	[KK03, LKLC03, Mör06, Fu11, AMST11]	yes	no
│ ├─Pattern, Motif Discovery	[LKL*04, Mör06, Fu11, AMST11]	no	yes
│ ├─Anomaly Detection	[LKLC03, LKL*04, Mör06]	no	yes
│ ├─Clustering	[KK03, LKLC03, Mör06, Fu11, AMST11]	no	yes
│ ├─Classification	[KK03, LKLC03, Mör06, Fu11, AMST11]	no	no
│ ├─Rule Discovery	[Mör06, Fu11]	no	no
│ ├─Summarization	[LKLC03, Fu11]	no	yes
│ ├─Subspace Matching	[LKL*04]	yes	yes
│ ├─Segmentation	[KK03, Mör06, Fu11]	yes	yes
│ └─Prediction	[Mör06, AMST11]	no	no
└─Interpretation			
└─Visualization	[Mör06, Fu11]	yes	yes

Table 2.4 *Overview of techniques/tasks and taxonomies for time-oriented data applied in the KDD reference workflow [FPS96].*

repetitions of (periodic) motions. Consequently, our solution supports both the exploration of internal and external dependencies in time.

2.3.2. Temporal Analysis Tasks

In the following, we characterize time-oriented data from a task-based perspective. First, we present a brief overview of the KDD process for time-oriented data. KDD in general will be reviewed in Section 2.4 in detail. Second, we extend the review of task taxonomies for IV and VA presented in Section 2.1.2 with task taxonomies explicitly presented for the analysis of time-oriented data. Finally, we summarize the section about temporal data analysis and define terms relevant for the remainder of this thesis.

The KDD Perspective on Temporal Analysis Tasks Analysis tasks can be surveyed from different perspectives and on several layers of abstraction. In the section about IV and VA, the perspectives in the task taxonomies focused on the user, the interaction and the visualization (cf. Section 2.1.2). The KDD perspective forms an additional point of view complementing the three perspectives mentioned above. In general, IV and VA uses the techniques known from KDD to facilitate analytical support. Similarly, we will combine the technical repertoire from KDD with techniques from IV and VA to support temporal analysis tasks in downstream ESS. Our technical contributions in Chapters 4, 5, and 6 make massive use of algorithmic models for time-oriented data. First, we apply techniques for time series preprocessing, time series descriptors, and time series similarity. In the following, we provide an overview of associated steps and techniques borrowed from the KDD workflow [FPS96]. Second, we use KDD techniques for the design of content-based overviews. This is why we review different classes of models, such as clustering, projections, and layout algorithms in detail. Third, we present techniques supporting the relation-seeking process between the time-oriented data content and attached metadata. Finally, we use the concept of constructing and executing workflows with a specialization on temporal data analysis. We present an overview of the KDD reference workflow [FPS96], as well as related workflow concepts in Section 2.4.

A variety of task taxonomies have been presented, subdividing the steps of the abstract KDD reference workflow into classes of techniques for time-oriented data [KK03, LKLC03, LKL*04, HKP11, Mör06, Fu11, KHP*11, GGAM12]. In Table 2.4, we provide an overview of salient steps in the KDD process for time-oriented data. The reference

workflow is subdivided into preprocessing, transformation, DM, and interpretation. Individual classes of techniques define the finest subdivision in the tree-view. In the last two columns of Table 2.4, we outline the relevance of individual KDD techniques for exploration and search support for ESS. *Preprocessing of time-oriented data* poses the first main step in the KDD reference workflow. A most relevant task is to provide a certain extent of data quality needed for downstream models in the workflow. Surveys for dirty data [GGAM12] and data wrangling [KHP*11] outline the challenges associated with providing data quality, and provide an overview of the different models involved in the data cleansing process. Preprocessing of time-oriented data is of particular importance for this thesis. One reason is the central challenge of providing content-based access to time-oriented primary data (cf. Sections 2.2.1 and 2.2.3). In general, preprocessing time series is most relevant for downstream models of the data analysis workflow. *Transformations of time-oriented data* transform the data content into formats which can be addressed by downstream DM techniques. A most common format is a FV describing the time-oriented data content in a compact and representative way. Capturing the relevant information of complex data and representing it with features is deemed particularly challenging in general [FH09]. The model for generating FVs is referred to as time series representations, or time series descriptors [KK03, LKLC03, Mör06, Fu11]. Depending on the targeted DM tasks, most approaches require the definition similarity as an additional step. *Data Mining* (DM) [HKP11] includes a variety of models applied in the KDD reference workflow. Most relevant for providing search support is indexing, retrieval by content, subspace matching, and segmentation [KK03, LKLC03, LKL*04, Mör06, Fu11]. For exploratory analysis support, most relevant models are pattern discovery, motiv discovery, anomaly detection, clustering, summarization, subspace matching, and segmentation [KK03, LKLC03, LKL*04, Mör06, Fu11]. *Interpretation* of KDD results includes visualization strategies, e.g., for the evaluation of results [Mör06, Fu11]. This also shows one of the differences between KDD and VA. While the KDD reference workflow uses visualization predominantly for the interpretation of final results, VA makes use of visualization throughout the entire process. In this connection, it can be noted that our first technical contribution in Chapter 4 is one of the very first VA approaches for preprocessing time-oriented data with visual-interactive means.

Taxonomies for Temporal Analysis Tasks We review temporal analysis tasks with an emphasis of search and exploration activity. Supplementary to the questions of *what* the focused data is like and *how* the data is presented in a visualization system, user tasks contribute to the question *why* time-oriented data is provided in a visualization [AMST11]. Until today, only few taxonomies for user tasks applied to time-oriented data have been presented [MA14].

In his task by data type taxonomy, **Ben Shneiderman** listed a set of seven basic tasks *Overview*, *Zoom*, *Filter*, *Details-on-Demand*, *Relate*, *History*, and *Extract* [Shn96]. These functional/operational tasks can be seen as requirements for designers of exploratory tools [AA06]. In addition, Shneiderman extended the taxonomy with three specific tasks for the analysis of time-oriented data.

- Finding all events *before* a moment or period, plus the basic tasks
- Finding all events *during* a moment or period, plus the basic tasks
- Finding all events *after* a moment or period, plus the basic tasks

It can be seen that Shneidermans temporal tasks are associated with the relations of instant time primitives *before*, *equals*, and *after* for time-oriented data [Mör06, AMST11]. Shneiderman‘s task taxonomy focuses on answering questions for the design of analysis systems, which is most relevant for this thesis. However, Shneiderman’s taxonomy does not further characterize *why* and *how* information seekers carry out temporal analysis tasks, a distinction which will be included in future taxonomies [AA06, BM13].

An additional baseline taxonomy for time-oriented data is the low-level task description by **Alan MacEachren** [Mac95]. The temporal analysis tasks are defined by a set of relevant questions that users may seek to answer with the time series visualization and analysis system [AMST11]. We list MacEachren‘s tasks for time-oriented data as reviewed by Andrienko and Andrienko [AA06] and Aigner et al. [AMST11].

- Existence of data element (if?)
- Temporal location (when?)
- Temporal duration (how long?)
- Temporal pattern / temporal texture (how often?)
- Rate of change (how fast?)
- Sequence of entities (what order?)
- Synchronization (do entities occur together?)

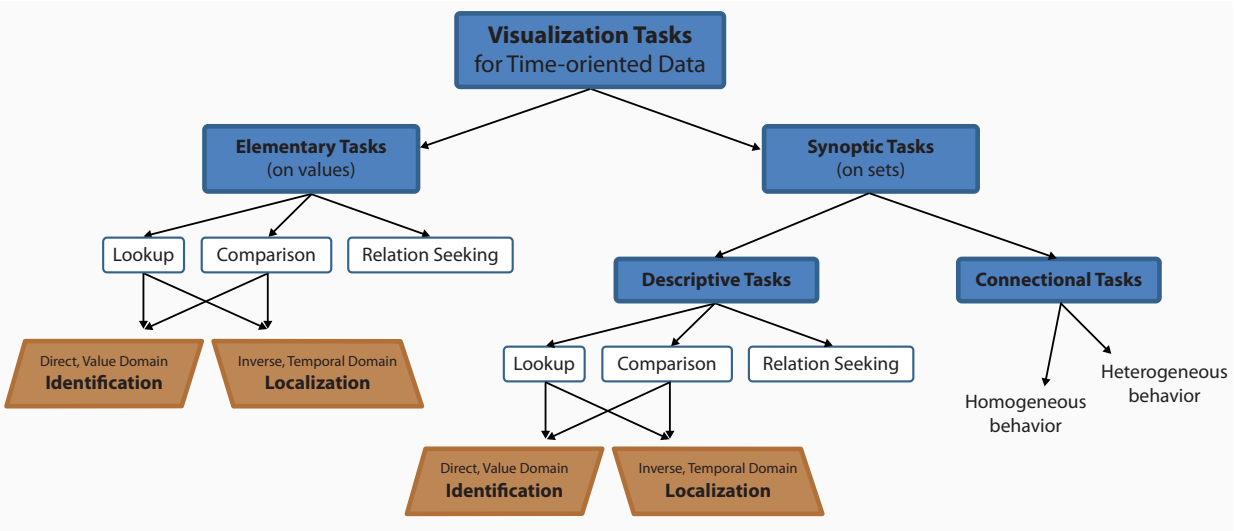


Figure 2.4 *Taxonomy of visualization tasks for time-oriented data. Adapted from [AA06, TFS08].*

Andrienko and Andrienko describe the seven tasks as query types, addressing different types of information. The queries can be seen as an elaboration of a more general task, i.e., to describe the times or set of times when a given object occurred (what -> when), or to describe the object or set of objects at a given time (when -> what). The distinction between known / unknown information in the temporal / value domain is reconsidered in later taxonomies partially building upon the taxonomy of MacEachren [AA06, AMST11, BM13]. Seeking for information in the value domain is typically referred to as *identification* tasks, while searching in the temporal domain is typically described as a *localization* task [AMST11]. The verbal task descriptions by MacEachren are easy to understand and can serve as a guideline when designing visual representations of time or time-oriented data. However, a more formal notation is desirable when shifting to a more scientific or theoretical point of view [AMST11].

The formal task taxonomy of **Andrienko and Andrienko** addresses this concern [AA06]. The taxonomy is based two formal models. First, a data model defines the correspondence between *references* (referrer, the temporal domain) and *characteristics* (attribute values, the value domain), similar to the characterization of time-oriented data presented by Aigner et al [AMST11] (cf. Section 2.3.1). Second, a task model represents a task as a combination of a *target* and *constraints* (unknown and known information). In terms of Jacques Bertin's *Semiology of Graphics* a target is called the "question type" [Ber83]. A target may be one or more characteristics, or one or more references. The goal is to find the initially unknown information corresponding to the known information. The definitions of a target and constraints depend on the types of information related to the structure of the data. Andrienko and Andrienko differentiate between tasks responding to specific data elements (*elementary*) and sets or groups (*synoptic*) of data. An illustration of the task taxonomy is presented in Figure 2.4. Elementary tasks regard individual data points separately, while synoptic tasks factor a set of data points as a whole [AMST11]. Elementary tasks are further divided into *lookup*, *comparison*, and *relation-seeking* tasks. Lookup tasks are divided into direct lookup (the references are known and the data characteristics are unknown, also called identification), and inverse lookup (data characteristics are known and references are searched, also called localization). Relation-seeking tasks search for occurrences of relations specified between data characteristics and references [AMST11]. Comparison tasks differ from relation-seeking tasks in a way that relations to be searched are not specified beforehand. Comparison tasks are further subdivided into direct and inverse comparison tasks, depending on whether the reference or the data characteristics are interrelated. For synoptic tasks, Andrienko and Andrienko consider the notion of a *behavior*. A behavior is a configuration of data characteristics corresponding to some reference set. A behavior is similar to what analysts also call a 'pattern'. Synoptic tasks are divided into descriptive and connectional tasks. Descriptive tasks specify properties of either a set of references or data characteristics, connectional tasks connect at least two sets of homogeneous or heterogeneous behavior. Descriptive tasks are divided into *lookup*, *relation seeking*, and *comparison* tasks. Lookup tasks are referred to as the identification and localization of patterns. Direct lookup corresponds to the definition of patterns where a set of known references yields unknown data characteristics (identification). Inverse lookup corresponds to pattern search where unknown references are searched based on known data characteristics (localization). Comparison tasks at the synoptic level apply to the comparison of patterns. Direct and inverse comparison tasks differs in their dependency on whether a set

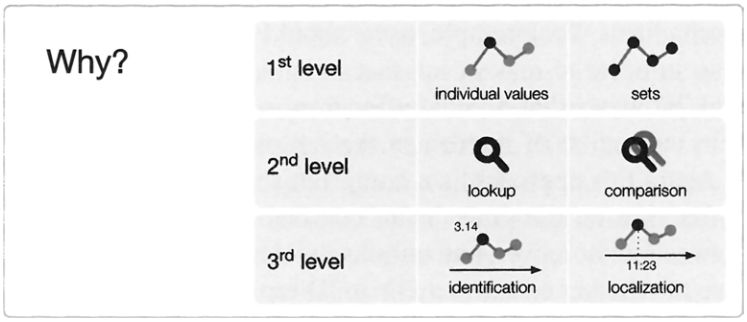


Figure 2.5 *An abstract view of the temporal analysis task taxonomy by Andrienko and Andrienko [AA06] presented by Aigner et al. [AMST11] (used with permission). We adopt the definitions of lookup, comparison, identification, and localization tasks.*

of references or a set of data characteristics is compared [AMST11]. Finally, synoptic relation-seeking tasks consider two sets of characteristics or references to come up with relationships between these sets.

Task Definitions used in this Work We conclude the section about temporal analysis tasks with definitions of terms used in this thesis. In KDD, DM, and IR, techniques as reviewed in Table 2.4 are often described as tasks. To avoid confusion of the term task, we refer techniques known from KDD, DM, and IR to the terms *operations*, *routines*, *models*, or *steps*, depending on the context. For example, we are talking about black-box *operations*, preprocessing *routines*, *steps* within a workflow, or analytical *models*. All these operations support data analyis tasks and can be applied in tools enabling the formulation of analytical questions a user might seek to answer.

We continue with an overview of tasks answering the question *why* [AMST11, BM13] a user is seeking information (see Figure 2.5).

- The **localization** task is carried out by a user knowing the characteristics of the value domain and searching for references in the temporal domain (inverse lookup [AA06]).
- The **identification** task is defined by reading characteristics in the value domain for given references in the temporal domain (direct lookup [AA06]).
- A **comparison** task is carried out if a) for fixed references different target entities are to be compared (direct comparison [AA06]), or b) if for a fixed target identities different references are to be compared (inverse comparison [AA06]).
- A **relation** is described as an association between different types of data. The most relevant example is the relation between the temporal domain and the value domain forming tuples as elements of time-oriented data. Other relevant relations addressed in this thesis are associations between different attributes of a data set, e.g., between data content and metadata.

Localization, identification, comparison and relation-seeking tasks can be carried out at the elementary level or at the synoptic level. In an exploratory context is more likely to execute the analysis tasks at a synoptic level [AA06].

2.3.3. Visual-Interactive Search and Exploration

We review visual-interactive search and exploration approaches for time-oriented data. We start with a brief overview of classical time series retrieval approaches, followed by an overview of visual query interfaces for time-oriented data. Finally, we present exploration-based time series approaches, with an emphasis on time series clustering.

Time Series Retrieval Research in content-based time series retrieval has a long history. Time series retrieval is typically based on time series descriptor techniques. The application of descriptors leads to compact but still representative FVs of the time-oriented data. For an efficient time series retrieval, these FVs are typically managed in index structures. The goal of index structures is enabling the fast execution of queries. Feature-based retrieval approaches typically use distance measures to be able to compare queries with the indexed FVs. In this way, time series

retrieval algorithms calculate sets of nearest neighbors for a given query. For an in-depth overview of content-based time series retrieval approaches, we refer to the survey of Fabian Mörchen [Mör06, p. 40].

One of the first time series retrieval approaches was based on the Discrete Fourier Transform (DFT) descriptor presented by Agrawal et al. [AFS93]. The Fourier coefficients are indexed in a R* tree, time series features are queried with the Euclidean distance measure. In the following, a great variety of descriptor approaches were presented, e.g., including the Discrete Wavelet Transform (DWT) [CF99]. In the same way, research on distance measures created a variety of techniques for time-oriented data. Both descriptors and distance measures serve as important baseline techniques for one of our technical contributions, the visual-interactive preprocessing of time-oriented data (see Chapter 4). A time series retrieval approach particularly relevant for this thesis is presented by Eamonn Keogh and his group. The VizTree system is one of the first systems with a visual representation of the underlying index structures which also facilitates visual browsing [LKL05]. The tree-based index visualization supports the discovery of frequently occurring patterns ('motiv discovery'), surprising patterns ('anomaly detection'), and query-by-content. VizTree is encouraging for this thesis as one of the first examples showing how visualization can provide access to time-oriented content, and likewise enable new analysis tasks beyond the classical time series retrieval paradigm. The work of Scherer et al. explicitly focuses on time-oriented scientific primary data [SBS11]. In contrast to many univariate approaches, the work presents a regression-based descriptor for bivariate time-oriented data. In his thesis, Maximilian Scherer further extends the approach with a measure of similarity and a benchmark for retrieval in bivariate and multivariate data collections [Sch13].

Visual-Interactive Querying of Time-Oriented Data Simultaneously with time series retrieval, progress was made in the research on visual-interactive querying methods. Querying is a process in which computer-supported tools provide answers to users' questions about the underling data [AA06]. Visual-interactive querying can be referred to as enabling users to formulate or characterize the properties of the content to be retrieved by visual-interactive means. Due to its importance for ESS, visual-interactive querying of time-oriented data is further illustrated in an example. A researcher is interested in daily patterns of temperature curves all over the world. At an appropriate DL system, she draws the temperature pattern of interest, for example, a linear upward trend. With the execution of the drawn query the DL system retrieves time-oriented primary data content with linear upward trends, allowing the researcher to analyze a meaningful subset of the data collection in detail.

One of the earliest and best known techniques for visual-interactive time series querying are Hochheiser and Shneiderman's *Timebox widgets* [HS04] (see Table 2.5). Based on the visualization of superimposed time series, users are able to drag rectangular boxes directly into the visual interface. The boxes serve as filtering functions, time series not contained in the boxes are filtered out. In addition, a second technique is presented where users can define *angular queries* (see Table 2.5). Time series whose progression does not correspond with the defined angular direction are filtered out. Both query types are considered early proposals for dynamic querying of time-oriented data (see the FilmFinder tool as pioneer work for dynamic queries in general [AS94]). Challenges for the design of visual-interactive querying systems using Timebox widgets are the design parameters which need to be addressed. First, all time series need to be aligned to a common temporal domain (represented as the x-axis of the interface). This may be trivial for time-oriented data where the internal time is modeled linear with absolute points in time (e.g., for stock charts). However, additional preprocessing operations my be executed on the time series to align cyclic or relative temporal domains in a meaningful way. Similarly, the value domain yields design parameters. One of the most relevant decisions of the value domain regards the normalization strategy to make time series visually comparable. Finally, a challenge arises if the temporal domain is non-equidistant (non-uniformly sampled). In this case, the visual interface and the querying techniques need to be adapted to non-equidistant time series representations.

Directly *sketching* time series is one of the most prominent alternatives to specifying Timebox widgets. Query-by-Sketch enables users to draw the shape of the time series. While TimeBox widgets retrieve time series which *exactly* match the specification of a given surrounding box, Query-by-Sketch retrieves times series *similar* to the shape of the drawn query. Thus, Query-by-Sketch is less appropriate for finding exact matches [AA06]. The power of Query-by-Sketch depends on at least two criteria. First, the appropriateness of query results depends on the users' knowledge on a particular data set. If users are not aware of the search space, the questions users may ask a tool may be ill-defined [WR09]. Our solutions for content-based overviews presented in Chapter 5 resolve the challenge of ill-defined query formulation by providing an overview of the search space. Second, the retrieved results depend on the definition of similarity implemented by the system. Time-oriented data have various characteristics and the transformations to be applied to time series are manifold. It is a particular challenge to design time series descriptors and definitions of similarity which meet the notion of similarity in the heads of users [BSR*14].

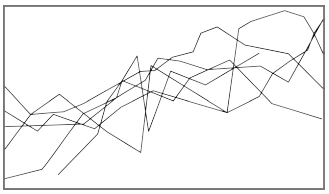
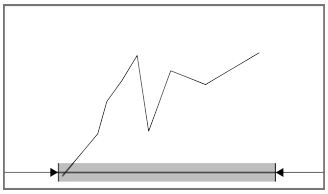
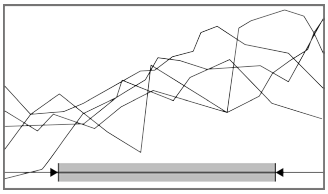
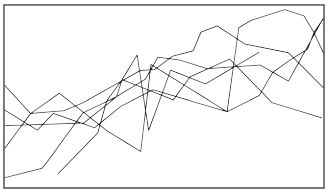
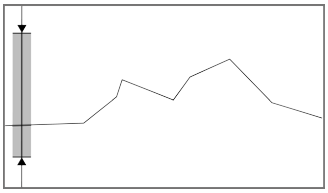
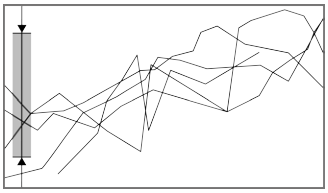
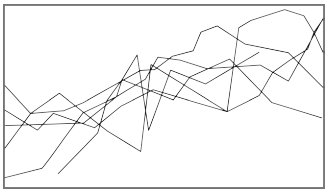
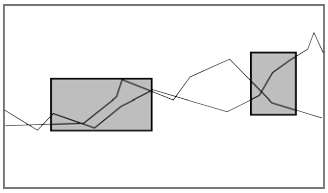
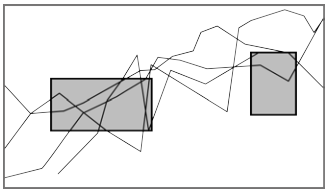
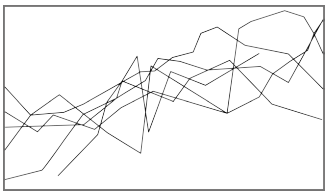
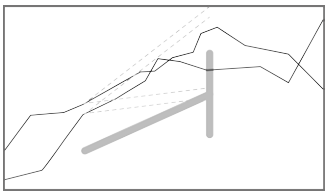
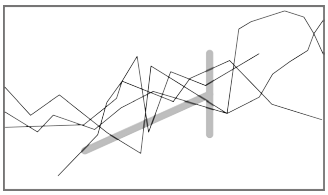
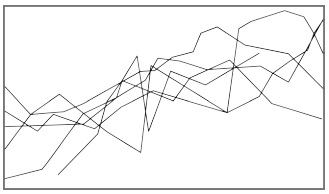
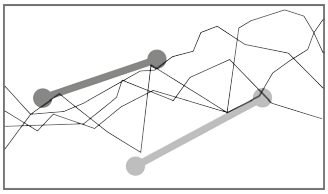
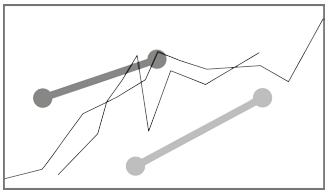
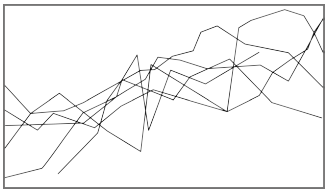
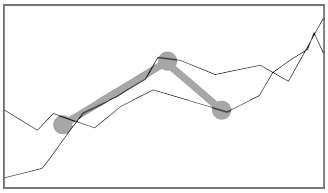
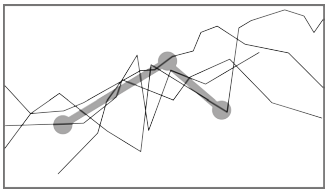
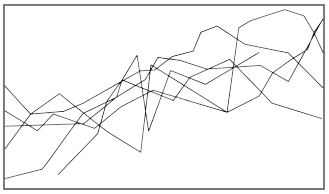
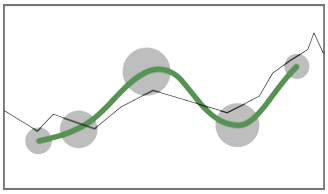
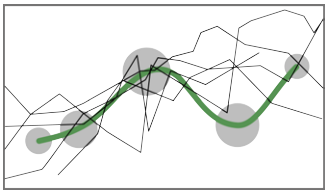
Visual Query Technique	Set of Time Series Example time series for the application of a visual query	Included by the Query Time series matching a given query	Excluded by the Query Time series not matching a given query
Dynamic Query (temporal domain) e.g., [AS94]			
Dynamic Query (value domain) e.g., [AS94]			
Timebox Widgets [HS04]			
Angular Query [HS04]			
Min Max Query [RLL*05]			
Goal Query (similar to query-by-sketch) [RLL*05]			
Relaxed Selection Query [HF09]			

Table 2.5 *Visual-interactive techniques for querying time-oriented data content.*

A querying technique inspired by Timebox widgets and Query-by-Sketch is the QueryLines approach [RLL*05]. The technique enables users to define soft constraints and preferences for querying time series. Four different types of queries are provided with the technique (see Table 2.5). The *minimum* and the *maximum* query can be used to filter time series outside a user-defined value domain. The *goal* specifies the shape of a time series to be retrieved similar to Query-by-Sketch. The *trend* query allows the definition of time series matching a direction (similar to the angular query presented by Hochheiser and Shneiderman [HS04]). If a user over-constraints a query, the soft constraints of the QueryLines technique can be used to present near misses (also called soft matches) enabling users to refine queries.

More recently, a Query-by-Sketch technique was presented where not only the sketch, but also the dynamics of the drawing motion is used for the query specification [HF09] (see Table 2.5). The visual interface of the *relaxed selection technique* supports both the definition of the shape of the query and the local accuracy. The latter is achieved by the size of points indicating a change in the direction of the sketched query. A matching algorithm automatically aligns the time series of the data set with the given query.

Finally, we review *Query-by-Example*, a visual querying technique where the system provides meaningful shapes of the time series and the user executes queries of interest. The main challenge of Query-by-Sketch is associated with the notion of interest of the user, i.e., what the user considers to be a meaningful shape for a time series example. In this connection, content-based overviews have proven to be highly beneficial [vLBBS10, vLSFK12]. In the two case studies in Chapter 7, we present two Query-by-Example solutions for time-oriented data in combination with content-based overviews.

Visual-Interactive Exploration of Time-Oriented Data In this section, we present techniques and applications supporting the exploration of time-oriented data. The goal of explorers is to discover interesting patterns in the temporal domain, the value domain, or in both domains without a priori assumptions [BM13]. To extract useful information from time-oriented data, a variety of analytical models have been presented. We refer to Section 2.3.2 where we present an overview of temporal analysis tasks and involved operations borrowed from time series DM and KDD [Mör06, FPS96]. Examples supporting exploration are classification, clustering, or pattern discovery techniques. For the exploratory analysis of time-oriented data, these models share a common goal: the *abstraction* of data to reduce the workload when computing and interpreting visual representations [AMST11]. Classification and clustering techniques are applied to abstract from raw data objects to meaningful groups of data. Thus, the data abstraction transfers the data from the elementary to the synoptic level which is most relevant for exploratory data analysis [AA06]. For pattern discovery, including the discovery of particular values, extreme values, anomalies, trends, or ranges, the goal is to emphasize relevant patterns by de-emphasizing irrelevant data [AMST11]. Many visual exploratory analysis approaches use the Information-Seeking Mantra ("Overview first, zoom and filter, then details-on-demand") by Ben Shneiderman [Shn96, AA06], "often beginning at an overview level of the visualization" [BM13]. This is why one of our three technical contributions explicitly focuses on the design of content-based overviews (see Chapter 5). In the following, we focus on visual-interactive clustering and pattern discovery approaches facilitating the exploration of time-oriented data. We refer to the books of Andrienko and Andrienko [AA06] and Aigner et al. [AMST11] for overviews of time series visualization approaches, partially including exploratory analysis support.

One of the most prominent and inspiring approaches is the Calendar View by van Wijk and van Selow [VWVS99]. Similar to our techniques the authors focus on the visualization of time series clusters to gain an overview of time-oriented data sets. The Calendar View clusters temporal patterns of equal duration to reveal different 'profiles', e.g., in the application field of power demand. These profiles are represented as superimposed linecharts for the enhanced comparison of the temporal patterns. Colors are used to link the different daily profiles to a calendar-based view. Here, juxtaposition of colored daily profiles is applied to identify weekly, monthly, and yearly patterns. One drawback of the approach is the use of categorical colors for the different cluster profiles. As a consequence, colors used for linking profiles in different views do not reflect the similarity of the clusters. In an exploratory analysis approach similar to the Calendar View, Steiger et al. [SBM*14] recently overcame the shortcoming with a similarity-preserving 2D colormap. Steiger et al also present additional interaction designs for the exploration of the temporal data content, such as the selection of subsets with a lasso-like interaction.

A clustering-based approach related to the DL domain is presented by Merkl et al. [MPR02]. A content-based similarity concept is used to present a hierarchically-structured map of music documents. A more recent exploratory data analysis approach based on audio documents is presented by Stober and Nürnberger [SN11]. The exploration system makes use of different interaction techniques (e.g., distortion, filtering) to support users in browsing through the music collection. Projection techniques are applied to map music documents into the screen space in a similarity-preserving way (see Section 5.2.3 for an overview of projection and layout techniques).

The clustering-based technique presented by Fu et al. responds to the discovery of stock time series patterns [FCNL01]. Time series data sets are abstracted with the Perceptually Important Points (PIP) descriptor (cf. [ZJGK10]) to preserve the overall shape of a time series, while reducing the number of tuples. The Self-organizing Maps (SOM) [KSH01] algorithm clusters the temporal patterns in a topology-preserving way enabling users to gain an overview of the data set. The SOM algorithm is a prominent means for the exploration of high-dimensional data in general. The algorithm is special in the way that it combines data aggregation (clustering) with data projection (layout). We refer to the work of Vesanto for an overview of SOM-based data visualization techniques for exploratory data analysis in general [Ves99]. In a VA approach for 2D time series, Schreck et al. demonstrate how unsupervised iterative clustering algorithms (here: SOM) can be adapted to semi-supervised variants by means of VA [SBVLK09]. The possibility to re-execute the clustering to optimize the clustering result is convincing for the design of ESS, and thus for our technical contribution about content-based overviews presented in Chapter 5.

A tool for the exploration of visually aggregated time-oriented data is the Time-Series bitmaps approach presented by Kumar et al. [KLK*05]. The tool enables users to browse through large collections of time series to discover clusters, anomalies, and other regularities. The approach uses the Symbolic Aggregate Approximation (SAX) descriptor [LKLC03] to visually represent time series subsequences as a bitmap metaphor. An additional approach presented by the same group is VizTree [LKL05]. Time series are aggregated to symbolic motifs, and visually represented as a tree. VisTree supports the discovery of frequent patterns (here called motives), the detection of surprising patterns (anomalies), and the formulation of content-based queries.

For the use case of data exploration in ecological research, Ahmed et al. [AYMW11] present a system with linked views providing a flexible interactive control of clustering inputs and outputs. Ecologists are able to use combinations of DM techniques and visual data representations. Color is used for linking views, however, the colors are chosen in a categorical rather than in a similarity-preserving way. The LifeFlow system visualizes large amounts of temporal clinical data [WGGP*11]. Based on the provided overviews, medical researchers are able to explore patient event sequences and discover patterns of interest. The overview of temporal events is provided by means of visual data aggregation allowing for an intuitive structure of large collections of event sequences. The KronoMiner tool provides a rich set of interaction designs for the exploration of time-oriented data [ZCB11]. The authors postulate a set of requirements for the exploration of time-oriented data as a result of informal studies conducted in different application domains. Core requirements are the applicability for complex input time series, linked views and multiple perspectives on the data, interaction designs reflecting Shneiderman's Information-Seeking Mantra [Shn96], and analytical methods, such as comparison and relation seeking.

The overview-first principle is also addressed in the LiveRAC system, where system management time series are explored visual-interactively [MMKN08]. LiveRAC is one of the first systems developed with a design study method in VA research (see Section 2.5). Besides the rigorous utilization of user-centered design principles, LiveRAC is special in the way that the system supports the exploration of multivariate time-oriented data. The multivariate value domain in the time-oriented data is an additional challenge for both the functional support and the visualization and interaction design. The MotionExplorer case study presented in Section 7.2 focuses on challenges with multivariate time series in the field of human motion capture data analysis.

2.4. Workflows and Frameworks - Combining Data and Tasks

In the last sections, we illustrated tasks associated with ES activity within the information-seeking process in Section 2.1.2, described the process of primary data within the data life-cycle in Section 2.2.2, outlined the value of scientific workflows in Section 2.2.3, and reviewed the process of time-oriented data in the KDD workflow in Section 2.3.2. In the following, we combine the data-centered and task-based perspectives in a review of workflows, pipelines, reference models, and analytical frameworks. As a result, we will be able to deduce a reference workflow for the design and the application of ESS systems, presented in Chapter 3. We start with a detailed review of scientific workflows in Section 2.4.1, followed by an overview of the KDD reference workflow in Section 2.4.2. In Sections 2.4.3 and 2.4.4, we present important reference models for IV and VA. Section 2.4.5 concludes with an overview of VA frameworks for time-oriented data.

Definition A workflow is a high-level specification of a set of steps (also called activities, or tasks) and dependencies between them that must be satisfied to accomplish a specific goal in a working environment [DGST09]. Each step represents the execution of a combinational unit, such as running a program, submitting a query to a database, submitting a job to a compute cloud or grid, or invoking a service over the Web to use a remote source [HTT09, p. 138].

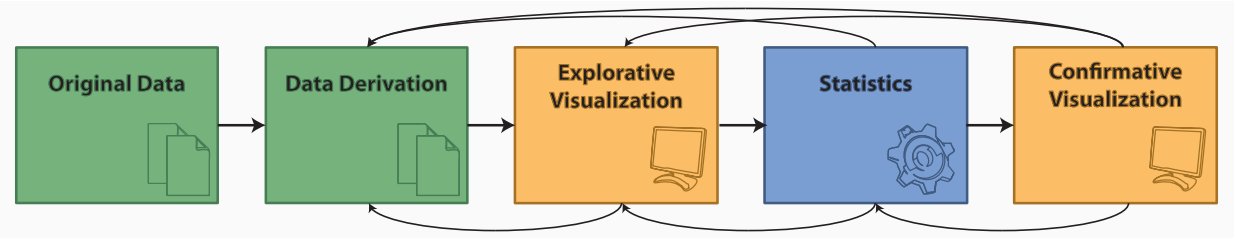


Figure 2.6 *Scientific workflow, here for the exploratory analysis of climate data. A series of subsequent steps is processed starting from the original (primary) data, over (visual) exploratory analysis, towards (visual) confirmatory analysis. Dependencies can also reflect backwards to previous steps [KLM*08]. We illustrate the workflow with the colors and icons used in this thesis.*

The data output from one step is consumed by subsequent steps according to a predefined graph topology that orchestrates the flow of data.

2.4.1. Scientific Workflows for Primary Data

In the following, we shed light on synergy effects between primary data and scientific workflows. We echo the trend towards enhanced exploratory analysis tasks in data-driven research. Furthermore, we outline associated challenges, such as coping with the scale and the diversity of today's primary data. The increase in data generation must be matched by scalable processing methods. Thus, new scientific knowledge is often gained by scientists putting together 'pipelines' for the analysis of data and the discovery of knowledge [LAB*06]. It is a long-term goal to support scientists with easy-to-use tools where primary data resources can be combined with computational services, visualizations and analysis techniques [LAB*06]. The technologies of such 'e-Science' applications have advanced from simple batch executions of data analysis. As described by Ludäscher et al., "scientific workflows are increasingly adopted across many natural science and engineering disciplines, spanning all conceivable dimensions and scales, from particle physics and computational chemistry simulations, to bio-informatics, medicine, environmental sciences, engineering, geology, phylogeny, all the way to astronomy and cosmology, to name a few" [LAB*09]. The time-oriented data produced in these disciplines can be plugged into scientific workflows supporting the automation of repetitive tasks, as well as the execution of complex analysis processes [DF08].

Scientific Workflows A scientific workflow can be seen as a sophisticated process of multiple coordinated steps in a scientific application. For example, a data analysis protocol consisting of a sequence of preprocessing, analysis, simulation and postprocessing steps is a typical workflow scenario in e-Science applications [DGST09]. The goal of e-Science workflow systems is to provide a specialized programming environment to simplify the programming effort required by scientists to orchestrate a computational science experiment. Scientific workflow systems generally have three components: an execution platform, a visual design suite, and a development kit. We describe the three components in more detail, according to Goble and Roure [HTT09, p. 138 ff.]. The *platform* executes the workflow on behalf of applications and handles common crosscutting concerns. Included activities are (1) the invocation of service applications, (2) monitoring and recovery from failures, (3) the optimization of memory, storage and execution, (4) data handling, (5) process logging and data provenance, and (6) security and access policy modeling. The *design suite* provides a visual scripting application for authoring and sharing workflows and preparing the components that are to be incorporated as executable steps. In this way, the creation of workflows can also be managed by users who are no experts in programming languages, or data science in general. The *development kit* enables developers to extend the capabilities of the system and enables workflows to be embedded into applications, web portals, or databases. This embedding has the potential to incorporate sophisticated knowledge seamlessly and invisibly into the tools that scientists use routinely.

The Life-Cycle of Scientific Workflows According to Deelman et al., the life-cycle of scientific workflows typically consists of four different phases [DGST09]. In the *composition* (construction) phase, the steps of workflows are specified, edited, and orchestrated as a graph topology of subsequent steps. Typical means for the composition are scripting languages, graphical interfaces, or automated workflow generation support techniques. Workflows are either described resource-independent (also called abstract [LAB*09]) or executable based on a given (mapped) resource. In

the *mapping* phase, (abstract) workflows are mapped to underlying resources. This is either performed directly by users, or by the system. In the first case, users select appropriate resources manually, while in the second case, users design workflows on a level of abstraction above the targeted execution environment. The result of the mapping phase is an executable workflow. In the *execution* phase, workflows are carried out by execution engines or enactment subsystems. Factors to be addressed are execution models, fault tolerance mechanisms, and the ability to adapt workflows based on previous steps in the cycle. Finally, in the *metadata and provenance* phase, data (and associated metadata) and provenance information created during various stages of the workflow lifecycle is recorded (see the survey Davidson and Freire on provenance in scientific workflows [DF08]). The placement of such information in a variety of registries supports the reuse when new workflows are designed.

In our technical contributions, we focus on graphical interfaces for the construction of workflows. The workflows are to some degree resource-independent as any source of time-oriented primary data can be plugged in which can be transformed into the form of the provided data models. However, for the construction of workflows, we enable users to map the workflow onto appropriate resources manually. In this way, we provide a) visual feedback with an instant workflow execution, b) data-centered parameter guidance, and c) techniques to guarantee generalizability of the workflow for large data collections. While the execution of the workflows can already be triggered in the construction and mapping phase, we provide a persistence concept to store and reload workflows for reuse.

The Value of Scientific Workflows Scientific workflow systems (SWS) provide a number of benefits for data-driven research. We highlight characteristics of SWS supporting the scholarly activities of researchers.

- SWS have become a necessary tool for many applications, enabling the construction and execution of complex analysis tasks on distributed resources [DGST09]. In many cases, SWS are designed in close collaboration with the scientists. As a consequence, resulting systems are well designed to handle the domain-specific use cases.
- SWS facilitate the automation of data-centered analysis goals. Based on the level of abstraction provided by SWS, scientists are able to conceptualize and manage the analysis process. As a result, SWS enable scientists to focus on the science driving the experiment instead of data and process management [LAB*09]. A limitation of the automation is the complexity of the targeted primary data and the analysis task which often requires additional human judgment [KMS*08, PVW09, BAF*13].
- Workflows are rapidly replacing primitive shell scripts [DF08]. While alternative ad-hoc approaches for constructing computational scientific experiments typically rely on imperative languages, SWS are rather based on dataflow languages. The workflow representation as directed graphs with nodes (steps) and connections (data dependencies, data flow) supports the creation and modification of workflows using graphical interfaces [DF08]. As a result, the programing model of such workflows can be kept simple. In many cases, sequences of steps can be composed by connecting the outputs of one step to the inputs of another. This visual abstraction hides implementation details and makes workflows more suitable for users who do not have substantial programming expertise. [DF08] Thus, this shift of workflow definition towards graphical user interfaces empowers scientists to build their own pipelines when they need them and how they need them [HTT09, p. 138].
- The graph-based structure of the workflow also provides beneficial means for the exploration of data. The visual abstraction facilitates monitoring of the workflow execution. SWS (more recently) record the workflow configuration, parameter sets of single steps and data analysis result of workflow executions [DF08]. This provenance information is kept available for later use, e.g., to optimize the data analysis results based on quality criteria, or for documentation [LAB*09].
- Closely related to the latter is the concept of involving visualization for intermediate steps of the workflow. With visualization becoming a more integral part of the workflow, users can visually comprehend not only the result of the workflow execution, but also the intermediate results of individual steps [FH11].
- The modularity of scientific workflows also facilitates an efficient reuse of components [DF08]. The reuse of knowledge artifacts (actors, workflows, etc.) is encouraged for both within and across disciplines [LAB*09]. Consequently, a combined reuse of both SWS components and primary data [CIZW13, KH13] has high potential for a synergetic interaction, depending on the technical infrastructure and the involved stakeholders.

Scientific Workflow Systems and Applications In the review on SWS, we distinguish between research efforts for providing SWS and SWS implementations, i.e. instances of a specific scientific workflow application. We recommend the work of Deelman et al. for a survey SWS [DGST09] in general.

A convincing early work associated with research on scientific workflows is presented by Springmeyer et al. [SBM92]. The authors present a characterization of the scientific data analysis process including an operation taxonomy grounded in observations of lab scientists studying physical phenomena. Many recent SWS are developed independently of particular application domains or single data science projects. Examples are *Kepler* [LAB*06], *VisTrails* [BCS*05], *SciRun* [26], *Taverna* [27], *Pegasus* [28], or *ICENI* [29]. Virtually all of these SWS emphasize the visual-interactive workflow composition. The Kepler SWS is (like VisTrails) an open source and provenance management system. It supports the effective reuse of existing workflow components, e.g., by providing interfaces to existing tools and applications. Likewise, many external follow-up modules build up on Kepler's baseline framework. The execution of workflows can either be performed locally, or by invoking remote services, such as web services and grid environments. Other SWS with grid computation support are Taverna, ICENI, or Pegasus. SCIRun and VisTrails explicitly focus on the exploration and optimization of parameter spaces. In addition, VisTrails supports exploratory computational tasks, such as the visualization and analysis of data in the field of SciVis. A system directly originating from IV and VA research is *DimStiller* enabling users to create workflows for an enhanced dimension reduction of multi-dimensional data [IMI*10]. The provided guidance concepts and the parameter optimization support enable the creation of data transformation and dimension reduction pipelines not only for experts in IV and VA. As such, Dimstiller is particluarly inspiring for our solution for visual-interactive preprocessing of time-oriented data presented in Chapter 4.

We proceed with the review of SWS applications and scientific workflow instances. In this connection, we neglect the large class of non-visual SWS approaches but rather emphasize approaches including the visual interactive construction and execution of scientific workflows. An extensive survey of SWS directly building on IV and VA concepts (and SciVis) is provided by Kehrer and Hauser [KH13]. The DimStiller system provides a case study with a 30-dimensional data set with physical attributes of chemicals, such as molecular weight and the number of bonds they possess. The constructed workflow includes the assessment of the quality of a given clustering result. The creation of views showing quality metrics and the spatial distribution of projected data empowered the users to assess that the given clustering was suboptimal. In climate research, prominent analysis goals are the generation and verification of hypotheses, e.g., for climate change. Steinbach et al. present a clustering-based approach with geo-spatial visualizations [STK*03]. Measurements based on sea level pressure (SLP) and sea surface temperature (SST) are clustered with a so-called 'shared nearest neighbor' clustering algorithm and correlated with existing climate indices. The representation of results on a map visualization enables Earth scientists to localize known behavior but also to identify new climate indices which may support a better prognosis of potentially unknown Earth phenomena. An additional clustering-based visual exploration tool for primary data is presented by Scherer et al. [SBS11]. A similarity concept for bivariate time-oriented data based on a functional dependency descriptor serves as a basic step in the pipeline. A visual representation of clustering results supports the exploration of primary data documents.

In the work of Kehrer et al. scientists are able to explore indicators for climate change in a visual-interactive system with linked views [KLM*08]. The underlying multivariate and time-oriented data sets provide climate models and observational primary data. In geosciences scientists aim at understanding ecological systems to predict changes and responses in space and time. For that purpose, Zafar et al. present visual-interactive analysis techniques including linked views for the exploration of steerable clustering results of multi-dimensional ecological data [AYMW11]. The tool facilitates the analysis of both primary data observed by remote sensing towers and secondary data calculated as functions of the observations. In molecular biology the analysis of protein and DNA data is supported with the Line Graph Explorer tool by Kincaid and Lam [KL06]. The time-oriented data stemming from electropherograms is presented in a list-based view, enhancing the analysis of value distributions and complex patterns. Moreover, relations between the time-oriented data and additional metadata can be explored in a linked view. The interactive visual exploration of medical data is facilitated by Turkay et al. [TLLH13]. The tool supports scientists to cope with large heterogeneous data sets, and thus to formulate new hypotheses. The main goal is the identification of relations between age, sex, neuropsychological test scores, and the statistics for the segmented brain regions. Finally, the authors discuss how interactive analysis methods can help analysts to cope with challenges of large and heterogeneous data.

[26] D. Weinstein, S. Parker, J. Simpson, K. Zimmerman, and G. Jones. Visualization in the SCIRun Problem-Solving Environment. In C. Hansen and C. Johnson, editors, Visualization Handbook, pp. 615-632. Elsevier, 2005.

[27] T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M. R. Pocock, A. Wipat, and P. Li. Taverna: A tool for the composition and enactment of bioinformatics workflows. Bioinformatics Journal, 20(17):3045-3054, 2004.

[28] E. Deelman, J. Blythe, Y. Gil, C. Kesselman, G. Mehta, K. Vahi, K. Blackburn, A. Lazzarini, A. Arbree, R. Cavanaugh, and S. Koranda. Mapping Abstract Complex Workflows onto Grid Environments. Journal of Grid Computing, 1(1):25-39, 2003.

[29] Mayer A., Mcgough S., Furmento N., Lee W., Newhouse S., Darlington J.: ICENI Dataflow and Workflow: Composition and Scheduling in Space and Time. In UK e-Science All Hands Meeting (2003), IOP Publishing Ltd, pp. 894-900. 14

Open Research Questions for Scientific Workflows The value of scientific workflows for data-driven research is challenged with various problems and remaining research questions. We review general challenges concerning the data, the users, and the tasks. In addition, we outline research questions of designing meaningful visualizations and interactions, choosing appropriate models and parameters, and drawing on concepts, such as uncertainty, user guidance, and provenance information.

From a *data-centered perspective*, a general challenge exists in the size of today's data collections. Managing the enormous amount of primary data being collected is considered the key to scientific progress. [AKD10] Today, the data collection rates still outperform the quantities of data exploitation, which makes efficient processing of huge primary data collections necessary [BWE06]. Taking time-oriented primary data into account, the heterogeneity, the quality, and the dependency on time pose additional challenges for scientific workflows. In Sections 2.2 and 2.3, we review the challenges of heterogeneous, low quality data, as well as time-oriented data in detail. A variety of operations exist which can be included into the workflow to resolve the outlined challenges.

Moreover, it is a widely accepted challenge that *real-world applications* often consist of a series of heterogeneous problems [KMS*08]. The dynamic nature of scientific analyzes needs to be supported by scientific workflows, allowing domain experts to execute real-world applications. In this connection, the visual-interactive construction of workflows is beneficial, however, it is only one side of the coin. The other side is to "lower the cost of visualization generation" in general, and allow it to become a more integral part of current SWS and involved analyses [FH11]. However, in SWS visualization is still too often just an end product of the scientific analysis [FH11]

The designs of *meaningful visualizations* for the included data and data transformations in different steps of the workflow are challenges on their own. Visual representations are crucial for the data exploration and the knowledge discovery process [TC06, KMS*08]. Visualizations make the input/output of different steps more transparent and enable scientists to communicate their results visually. The quality of such data representations and the ability of scientists to perceive respective visualizations is a concern for future research in general [TM04, KAF*08]. Different design considerations and guidelines have to be addressed to yield meaningful visualization and interaction designs [Ber83, Tuf86, Tuf90, War12]. Furthermore, interaction design principles within the user interfaces need to be addressed [AS94, CMS99, HCL05]. Rapid feedback for the interactive behavior of scientists may help to further improve current analysis capability [KKEM10, p. 147].

An additional challenge regards the *role of domain experts* within the design and the use of scientific workflows, which needs a more in-depth investigation [BWE06, AKM*09, TDN11, MA14]. It is a particular challenge for scientific workflows that scientists still underestimate the value of advanced visual-interactive steps and their benefit for the enhanced hypotheses formulation/validation [NSBW08, TDN11]. One possible reason therefore is the missing 'trust' of researchers in unfamiliar techniques and the tendency to refuse to change their working routines [BWE06, HPK08, KMS*08]. This may also be a signal for the demand of more/better user-centered design approaches. The different skills in domain (data) knowledge, data transformation, and IV as they occur in the data-driven research expect the collaboration of *different stakeholders* for the construction of sophisticated scientific workflows [vW06, SMM12]. The ongoing research in evaluation and design study methodology in IV and VA has the potential to further improve the SWS design process and to enhance the collaboration of different stakeholders [KAF*08, Mun09, SMM12] [KKEM10, p. 146] (cf. Section 2.5).

From an analytical perspective, the construction of efficient and effective scientific workflows poses additional challenges concerning *models and parameters*. The success of a workflow design depends on the choice of the right steps in the right order, which requires expert knowledge about the underlying data and the data transformation techniques [FPS96, KLR04]. Identifying the best automated algorithm for a step remains a challenge in general [KAF*08]. In this connection, the application of various models on time-oriented data requires a close combination of human judgment and automated computation [BAF*13]. Moreover, many data transformation steps hold complex parameter spaces which have to be optimized [BCS*05, KHP*11]. Especially for users with no in-depth knowledge in parameter-laden of algorithms two main dangers are as follows. On the one hand, that incorrect settings may cause an algorithm to fail in finding the true patterns. [KLR04] On the other hand, the algorithm may report spurious patterns that do not really exist, or greatly overestimate the significance of the reported patterns. As stated in DimStiller providing *analytical guidance and support* capability within each of the applied steps of the workflow is crucial since at least some of the steps cannot be run without human interference [IMI*10].

An additional means to minimize cognitive and perceptual biases is a clear understanding of *uncertainties* within the scientific workflow, which is an additional desirable objective [KKEM10, p. 146]. Similarly, a challenge is supporting researchers in reporting the sensibility of their results and the validity of their findings [TLLH13]. Especially for exploratory analysis tasks the *classification of success* and the decision what makes a good solution is ill-defined [KKEM10, p. 147]. Quantitative statements, more advanced inferential statistics, and continuous feedback may

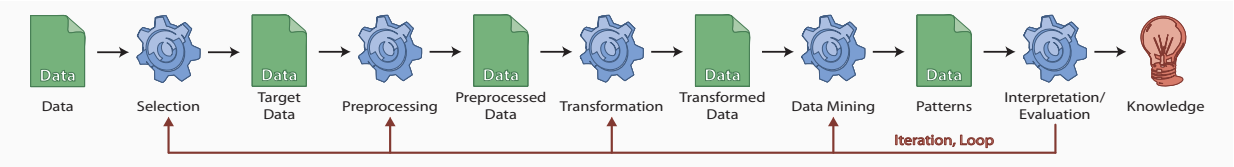


Figure 2.7 *The Knowledge Discovery (and DM) Process [FPS96]. We illustrate the KDD process with the colors and icons used in this thesis.*

help to verify the statistical significance of certain relations [TLLH13, Fek13]. Closely associated is the relevance of storing associated *metadata and provenance information* created during the steps of the workflow [BCS*05] [KKEM10, p. 146]. Hence, valuable information about SWS execution variables and the usage behavior of individuals can be gathered. Moreover, an analysis of provenance information and the creation of insightful visualizations of provenance data may help scientists to better understand their results [DF08, AKD10].

2.4.2. The Knowledge Discovery in Databases Process

In the latter sections, we reviewed different types of analysis tasks and algorithmic techniques to be applied to time-oriented primary data. In the following, we present an overview of the KDD process at a glance, i.e., the concatenation of single operations to a process for the enhanced data analysis. KDD is targeted towards large data sets with the goal to detect potentially interesting information, and thus to support sense-making and knowledge discovery. Fayyad et al. define KDD as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [FPS96]. The authors present pipeline concept of general steps within the KDD process with *data* as input and *knowledge* as output (see Figure 2.7). Initial steps of the pipeline include the *selection* of appropriate target data and *preprocessing* the raw data. In a *transformation* step the data is converted into a format which *DM* techniques can address. A prominent example of such formats are FVs which carry most of the information of the potentially large raw data in a compact representation. DM is a particular important step in the KDD process since the typical output of DM techniques are 'patterns' containing potentially valuable information. In an *interpretation and evaluation* step knowledge is discovered based on the provided pattern information. In summary, the KDD reference model is a pipeline of subsequently executed operational steps. In addition, the pipeline provides a feedback loop starting at the downstream evaluation step for iterative refinement. The DM step maybe consists of the most extensive set of operations (cf. Table 2.4). Relevant classes of techniques for this thesis are indexing, content-based retrieval, pattern and motif discovery, anomaly detection, and clustering, in particular. The KDD process is particularly relevant for this thesis since we elaborate the entire pipeline from time-oriented primary data to visual search and exploration interfaces. Particularly relevant for the time-oriented data are time series KDD steps like preprocessing and the transformation of time series into the feature space (cf. Section 4.2), as well as downstream techniques, such as clustering for the design of content-based overviews (cf. Section 5.2). General research and application challenges of the KDD process are the assessment of statistical significance, overfitting, noisy data, complex relationships, understanding patterns, and user interaction [FPS96].

2.4.3. The Information Visualization Reference Model

One of the most prominent reference models for the use of IV systems is the Information Visualization Reference Model by Card et al. [CMS99]. In the following, we call this reference model the Card pipeline. Figure 2.8 shows a variant recently presented by Jean-Daniel Fekete [Fek13]. At first, raw data needs to be transformed into a structured format until it can be provided in data tables. The *data transformation* step includes most of the phases presented in the KDD model. The subsequent *visual mapping* step shifts the abstract data to visual structures, and thus transforms the data into a visual form. In the *view transformation* step the visual structures are arranged to views. IV systems may implement the reference model multiple times in different ways if respective systems consist of different views. The user interacts with the views, e.g., by triggering the feedback loop back to the three upstream steps of the pipeline. Typical modifications triggered with the feedback loop are variations of underlying data collections by database queries, changes in the encodings including visual variables, or reconfigurations of data items at the display [YKSJ07]. In contrast to VA, IV only occasionally changes the underlying computational models and respective model parameters. A most fundamental concept of the Card pipeline is the clear distinction between the non-visual and the visual stage

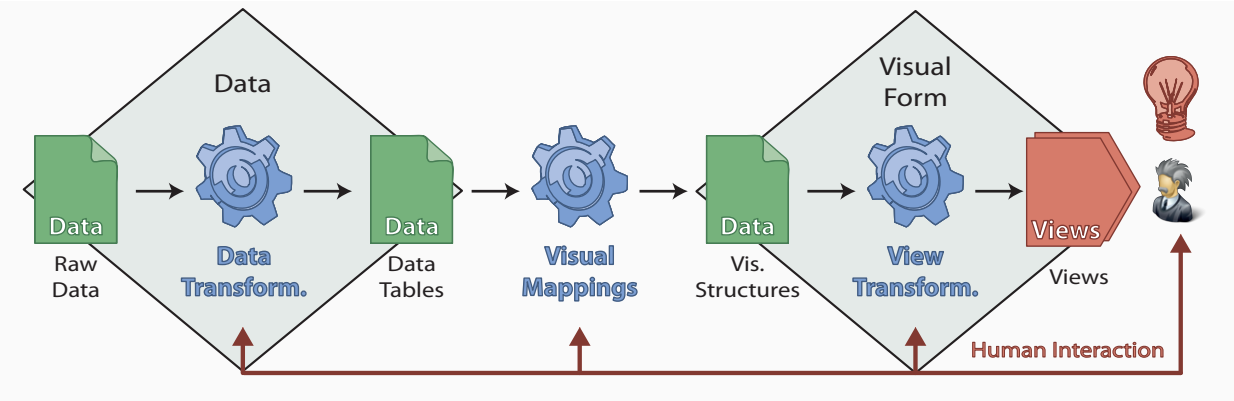


Figure 2.8 *The Information Visualization Reference Model by Card et al. [CMS99]. Shortcomings for this work are the missing user-centered design process and the lack of detail in the data transformation step.*

of the pipeline. While at least a portion of the human interaction triggers data transformations, the majority of the interaction focuses visual mappings and view transformations, respectively. However, the data transformation step yields most of the analytical parts relevant for the design of ESS components. This is why the Card pipeline is rather beneficial for the *use* of ESS than for its *design*. In this thesis, we extend the Card pipeline with VA concepts, especially for the data transformation stage early in the workflow.

A prominent example implementing the Card pipeline is the Prefuse information visualization framework [HCL05]. In the proposed framework, abstract data is transformed, filtered, and mapped to a visual form. These visual analogues are rendered at the interactive display. The user is able to execute UI controls to trigger I/O libraries, an action list of filtering and visual encoding operations, and renderers for the visual analogues. Similar to the Card pipeline, the transformations applied to abstract data are kept short, while the steps of visual encoding and interaction design receive more attention. While it is in accordance with the use of IV systems to keep the data transformation step abstract and compact, our model extends this early step to better cope with the challenges of ESS design. Dos Santos and Brodlie present a visualization workflow targeted towards multivariate and multi-dimensional data [dSB04]. The authors extend the transformation step of the Card pipeline by emphasizing the 'data problem' and by adding analysis and visualization steps. The results of these preliminarily steps are used for an enhanced elaboration of 'focus data' until the remaining steps are executed. From the design perspective, an interesting approach is working with (software) design patterns for IV as presented by Heer and Agrawala [HA06]. In contrast to the Card pipeline these design patterns are not bound to particular data-centered steps of the pipeline. On the contrary, these design patterns support the development of IV software from an object-oriented programming perspective, possibly at any (visual-interactive) step of the design workflow.

Finally, a convincing work for the design of visual-interactive systems is presented by Bertini et al. [BTK11]. The authors survey techniques that use quality metrics to help in the visual exploration of meaningful patterns in high-dimensional data. The overall goal is to automate the design process in a way that quality metrics are used to suggest most promising visualizations for a given analysis task. In their systematization, Bertini et al. present a reworked version of the Card pipeline extended by 'quality-metrics-driven automation' for each of the three steps (data transformation, visual mapping, and view transformation). In this way, the authors also reveal relations between existing quality metrics. The systematization is encouraging for our techniques in two ways. First, the authors postulate the use of quality metrics at different steps of the pipeline. We agree with the authors' statement who deem important the assessment of quality for any given step within a workflow. Second, with their extension of the Card pipeline, Bertini et al. recommend to provide users of IV systems with additional competences which facilitate the design of IV systems in a most meaningful way. In this thesis, we distinguish between the *design* and the *application* of visual-interactive systems. The design phase, the workflows for the visual-interactive systems are created. Data scientists are encouraged to incorporate external competences, such as domain experts (users), quality metrics, and other types of design support. In the application phase of visual-interactive systems, users are able to interact with the workable designs, just as presented in the Card pipeline.

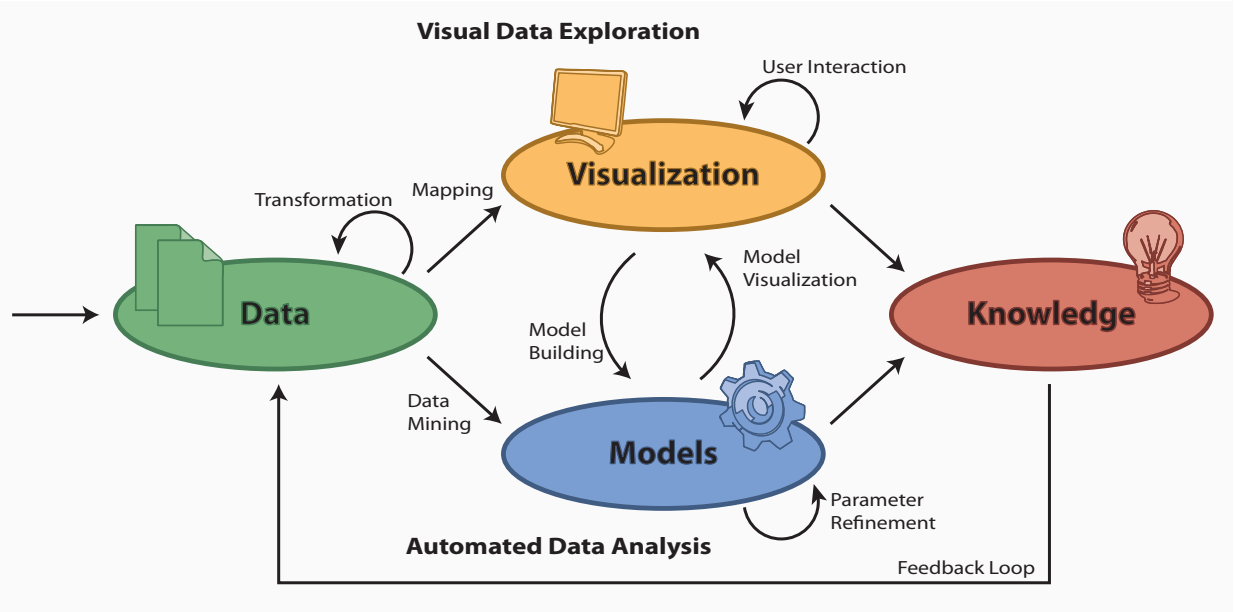


Figure 2.9 *The VA process presented by Keim et al. [KAF*08, KMS*08, KKEM10]. We illustrate the VA process with the colors and icons used in this thesis.*

2.4.4. The Visual Analytics Process

A widely accepted conceptualization of the VA process is presented by Keim et al. [KKEM10]. The VA process workflow consists of the four elements *data*, *visualization*, *model*, and *knowledge*, arranged in the shape of a diamond. In the following, we call this VA process the VA Diamond. A set of heterogeneous data sources (i.e., the internet, newspapers, books, scientific experiments, expert systems) serve as the input [KMS*08]. The output is the knowledge gained from the input source within the VA process. Insight is obtained either directly from the visualization, or through the confirmation of hypotheses, as a result of automated analysis methods (model). The VA process enables users to combine visualization with underlying models by means of interaction. A feedback loop spanning from the knowledge step back to the data, indicates the iterative nature of the VA process. In a broader sense, the interaction within the VA Diamond between the visualization and the model can be seen as a model modification (optimization), e.g., by variation of parameter values. Moreover, the feedback loop enables users to change the model (or the visualization) in itself, providing an additional powerful design parameter for both data analysis and system design.

The strength of the VA process is the tight coupling of visual-interactive data visualization and automated model calculation. The VA process scheme can be implemented for any step (or step combinations) of data-centered workflows. Especially the steps presented in the KDD workflow can benefit from visual-interactive VA capability. Classical black-box driven approaches only apply visualization as a means of representing the results of automatically generated results, e.g., provided with the execution of scientific workflows. On the contrary, VA also uses visual-interactive representations at large parts in the process of constructing workflows.

A shortcoming of the VA Diamond is the high level of abstraction. The concept of 'models' combines all the algorithmic models possibly included in a (scientific) workflow in a single element. As a result, analysis tasks and goals, such as the design of content-based overviews, or relation seeking between data content and metadata, cannot be mapped onto the VA process directly. In fact, the VA process may be duplicated for any major step of the targeted workflow. In contrast to the well understood Card pipeline, the VA process is still under development and the design of such a reference model still requires more time and experience [Fek13].

2.4.5. Time Series Visual Analytics Frameworks

A conceptual framework for the VA of time-oriented data is presented by Aigner et al. [ABM*07] (see Figure 2.10). Similar to the VA Diamond, it combines *visualization and interaction components* with *analytical & mining components*. Both components are embedded in a *graphical user interface*. In addition, *internal data structures* for

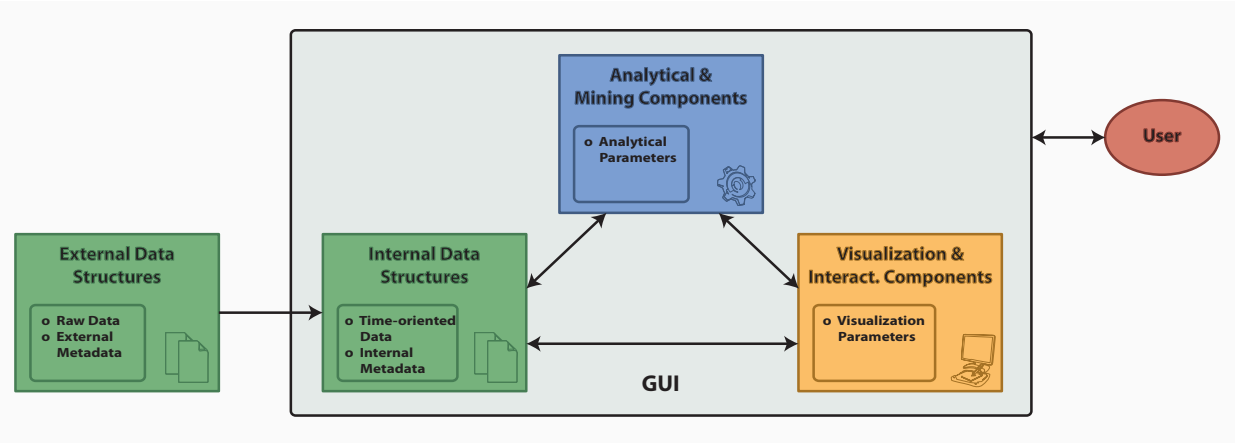


Figure 2.10 *A concept of designing VA frameworks and systems for time-oriented data [ABM*07]. We illustrate the framework with the colors and icons used in this thesis.*

time-oriented data content and metadata are provided. The GUI serves as an abstract layer between the user and the three involved components (visualization, model, and data). Finally, the GUI is connected to an *external data source* providing the raw data (e.g., primary data) including metadata. At this level of abstraction, this framework builds the basis for the different time-oriented VA tools that we present in our technical contributions. The distinction between internal and external data structures is a particularly influential factor for this thesis. In this connection, the framework is more precise than the VA process presented by Keim et al. since two mandatory elements of the framework address data structures. However, owed to the high level of abstraction, the framework still lacks an indication of *how* the transformation of external into internal data structures is to be performed. More precisely, the VA framework does not support data analysts and system designers in the transformation of external into internal data structures. In this connection, most severe challenges for time-oriented data are quality issues, normalizations, other preprocessing operations, as well as the definition of descriptors and time series similarity (see Sections 2.3.2 and 4.2). Surprisingly, in many time series applications, the transformation of external data structures into internal data structures is not part of the approach. Preprocessed time-oriented data is rather pre-given, possibly based on previously executed workflow steps that are not described in detail. For time series analysis applications, the exclusion of such preprocessing steps may be reasonable since the main contribution (from a scientific point of view) is based on performing analysis tasks, or visualization and interaction design. However, for the design of ESS, the transformation step from external raw data to internal time-oriented data is particularly important. The same applies to data-centered research approaches building up on primary data.

Similar to the VA framework for time-oriented data, Andrienko and Andrienko present a framework towards spatio-temporal data analysis as shown in Figure 2.11. The user directly interacts with *visual tools* and *interactive controls*, respectively. In turn, interactive controls trigger *data selection and grouping tools*. The data perspective is represented by a *data storage* assembling the primary data and different types of secondary data. The data storage can be manipulated by data selection and grouping tools, as well as by additional *computational tools*. Thus, computational tools, such as clustering, modeling, and other data transformations, are clearly separated from data selection and grouping tools. The framwework of Andrienko and Andrienko incorporates the data flow, the control flow, and the user perception and is, thus, related to the model-view-control pattern (MVC) known from design patterns in software engineering. The Andrienko and Andrienko's framework complements the VA process model presented by Keim et al. in a a more precise elaboration of relevant computational tools and different types of data structures including the original data and intermediate, such as clusters. However, similar to the framework of Aigner et al. the scheme does not provide means for the workflow construction of time-oriented ESS components.

2.5. User-Centered Design

Together with *data* and *tasks*, the *user* addresses the third magnitude in the design triangle for time-oriented VA applications [MA14]. In the course of the section, we survey principles, activities and models on interactions between designers (data scientists) and users. As a result, we will gain a clear understanding of user-centered concepts and

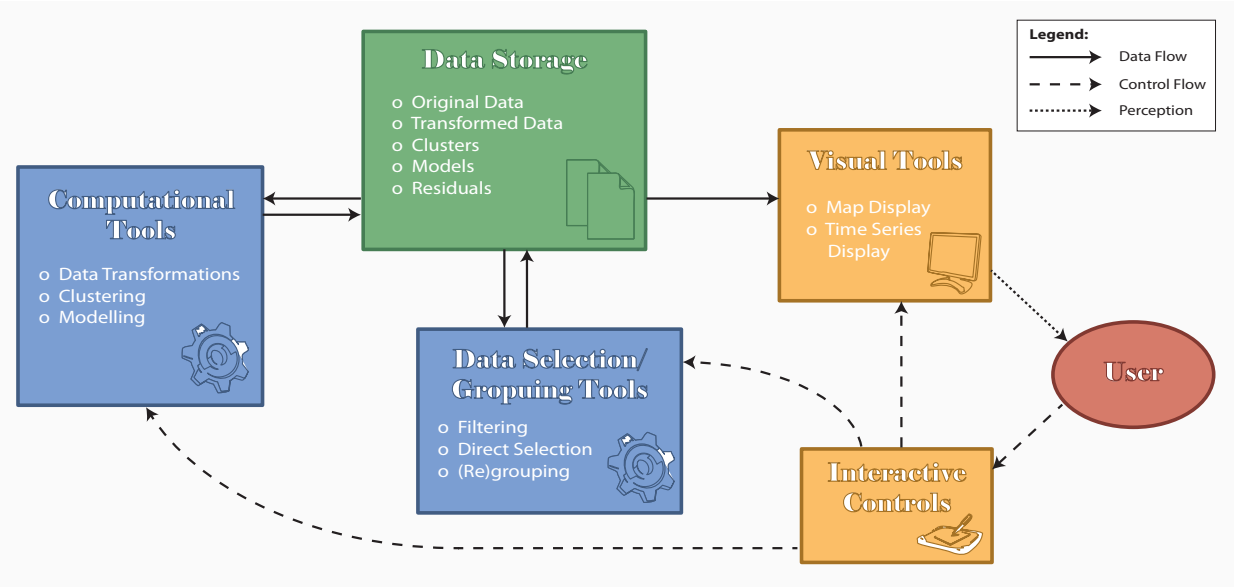


Figure 2.11 *A framework of spatio-temporal data analysis incorporating data flow, control flow and perception [AA13]. We illustrate the framework with the colors and icons used in this thesis.*

terms relevant for the design and the use of ESS. We start with a characterization of collaborator roles in Section 2.5.1. Next, we review user-centered design principles and activities from an IV and VA perspective, including an emphasis on design study methodology in Section 2.5.2. Finally, we present evaluation strategies from a methodological and a technical perspective in Section 2.5.3.

2.5.1. Different Collaborator Roles

We define a *role* (also called a stakeholder) as a person or group with a special expertise within a collaborative project. We expect this role to have an interest in a successful completion of the project. A role is not necessarily executed by a single person. Similarly, a person may be involved in several roles. Each collaborator role provides specific skills, which may be beneficial for the joint approach. However, stakeholders also have individual expectations from a project. The design of ESS for a targeted user group working in a specific application domain typically requires the collaboration of different stakeholders. The awareness of different roles in collaborations is a common theme in many research areas and user-centered design approaches [SMM12]. A survey of the biological and chemical domain may serve as an example [AKM*09]. Approaches drawing on a genuine collaboration between different stakeholders (including data scientists) are characterized "essential to stimulate constructive discussions and collaborations among all the relevant players" [AKM*09]. Similarly, in their survey of building DLs for primary data, Borgman et al. point out that it is "necessary to understand scientists' data practices for being able to build good tools and services" [BWE06]. We observe a similar situation in collaborative approaches in political decision making [RBU*14]. The essential roles (possibly) included in collaborative approaches for the design of ESS are as follows.

Collaborator Role: Domain Expert We describe the role of the *domain expert* in detail. Domain experts play the most critical role in the collaborative efforts targeted in this thesis. Domain experts are the end users doing the actual analysis [SMM12], i.e., domain experts are the people who will use the ESS. With the role of domain experts, we unite related roles, such as *front-line analysts* [SMM12], *information seekers* [Mar95], or more generally: *users*. In the context of primary data (cf. Section 2.2) and scientific workflows (cf. Section 2.4.1), we draw a parallel between domain experts and scientists. This is why in some paragraphs, we will use the terms *domain experts*, *scientists*, and *researchers* interchangeably. We echo Yi et al. [YKSJ07] to emphasize that *users* (rather than *viewers* or *people*) actively use and interact the visual-interactive techniques and systems. It is considered a pitfall to start a user-centered design approach (or a design study) without establishing the contact to at least one domain expert. [SMM12] However, additional domain experts can be identified within the project. Domain experts need useful tools [vW06] to conduct scientific practice. The expertise of domain experts has many facets including deep domain knowledge, the knowledge

of individual (time-oriented) data sets, the work practice including specific analysis tasks, and the cultivation of individual technical jargons (e.g., [TC05, vW06, MH10, PVW09, The12, MA14]). However, in many cases, domain experts only have limited knowledge in computer science or in batch-programming [DF08]. Domain experts often work informally with a high degree of independence. In many cases, the information-seeking behavior of domain experts is rather exploratory than based on lookup-search or fact retrieval [Mar06]. Domain experts may be involved in the data creation process of the data life-cycle (see Section 2.2.2). Regardless the fact that the data creation often requires technical expertise (e.g., in the utilization of sensor technology), we identify a strong relation between the role of *data creators* and domain experts. As an example, in the VisInfo case study in Section 7.1, one of the domain experts from Earth observation research spends several months in a year at a research station at Antarctica, leading the data creation department. Domain experts also play a role in the data analysis and the data reuse phase of the data life-cycle. In addition, in Section 2.2.3, we describe the role of domain experts in the context of content-based access.

Collaborator Role: Data Curator An additional role, relevant for data-centered collaborations, is the *data curator*, most relevant in the data processing phase of the data life-cycle (see Section 2.2.2). Data curators accept primary data and execute processing and storing tasks, such as translation, digitalization, formatting, or cleaning. In the DL domain, data curators are also called *resource managers* [CCF*08]. In the VisInfo case study, we had an active collaboration with data curators from the PANGAEA data warehouse [PAN]. In combination with data management technology [AKD10], the role of a data curator can be extended to a *data provider*, or *database manager*.

Collaborator Role: Digital Librarian The *digital librarian* is yet an additional role relevant for this thesis. Digital librarians support information seekers by facilitating library treatment and search support for digital documents (such as time-oriented primary data). Librarians play an overarching role in the data access and the data reuse phase of the data life-cycle (see Section 2.2.2). In this way, librarians also support the dissemination of (primary) data documents. Sharing (open) data is deemed important especially in scientific disciplines that consider primary data scientific capital [BWE06] [HTT09] [CIZW13]. In the VisInfo case study in Section 7.1, we present a DL system allowing domain experts carrying out ES tasks on time-oriented primary data. The overall goal of the project was to facilitate visual access to time-oriented primary data by means of a content-based access strategy. In this collaborative approach the role of digital librarians was equally important as the role of the involved domain experts from Earth observation. We refer to Section 2.2.3 for more details about the role of digital librarians in the context of content-based access in general.

Collaborator Role: Gatekeepter A further important role in a collaboration plays the *gatekeepter* [SMM12]. The gatekeeper has the power to approve or block the project. Similarly, the gatekeeper is authorized to distribute resources and work time on the project. In a scientific context, gatekeepers are often seniors or principal investigators.

Collaborator Role: Political & Financial Stakeholder In this connection, we briefly outline the influence of *political* and *financial stakeholders* on the project. Only if political and financial conditions are provided in an appropriate way, a project can be put into practice. In the context of primary data, the accessibility of primary data must be provided, which is often a political or financial concern [Sub12, CIZW13]. Taking the technical contributions of this thesis into account, we assign the non-technical stakeholders a subordinate role.

Collaborator Role: Data Scientist Finally, we characterize the role of a *data scientist*. With data scientists, we refer to computer scientists developing data-centered concepts and solutions from database technology to information science. The main functions of data scientists are the design and the application of a) data models, b) data-centered algorithms, c) visualizations and interaction designs, and d) the active collaboration with stakeholders in scientific practice. Data scientists combine expertise in computer science, HCI [vW06], (information) visualization design [Mar95, vW06, PVW09, Mun09, SMM12], and data analysis [FPS96, TC05, KKEM10]. Influential works including scope, challenges and future directions of IV and VA [CMS99, TC05, TC06, KMS*08, KAF*08, KKEM10] give a profound insight in the practices of data scientists. For more details about the role of data scientists in the context of content-based access to primary data, we refer to Section 2.2.3. In general, we identify two goals for data scientists. On the one hand, data scientists aim at creating new methods, e.g., to improve data-centered analysis capability [vW06]. This is especially the case if a data scientist plays an active role in research, (e.g., DM, IR, KDD, IV, or VA). On the other hand, data scientists represent an important role in collaborative projects [Mun09] [SMM12]. As a side note, we want to remark that it is also possible that data scientists with different focus and expertise can take part at a

collaborative effort. For instance, we refer to the MotionExplorer case study in Section 7.2, in which we collaborated with experts in human motion capture retrieval research.

Challenges in the Collaboration between Domain Experts and Data Scientists We conclude the section with an outline of challenges arising from collaborative efforts. In particular, we observe the relationship between domain experts and data scientists, defining the two most relevant roles for this thesis. Our characterization of challenges in collaboration echoes the work of van Wijk [vW06]. Data scientists aim at creating new methods, e.g., to play an active role in research. Domain experts need useful tools, e.g., to carry out data-driven research. As a result, two gaps in the collaboration between data scientists and domain experts can be identified. First, a knowledge gap exists between the two stakeholders resulting from the different expertises. Both stakeholders use different technical jargons and even if both use familiar names they can have a different meaning. The data scientist needs to understand at least the basics of the domain experts' field of research [vW06]. In return, the domain expert needs to find a way to communicate how new insight can be described allowing the data scientist to derive a characterization of technical solutions. However, due to complexity of the information-seeking behavior, it is not always easy for domain experts to express what they are looking for at the first place. Second, an interest gap exists between the two stakeholders. Data scientists aim at publishing in journals and leading conferences on visualization. In consequence, data scientists focus on the research of new interesting methods an techniques, but not primarily on their usability. [vW06] Domain experts, however, are interested in tools that will help to do work faster and better. Criteria on tools are their usability, their usefulness, or their applicability for specific data characteristics. Van Wijk refers user-centered design approaches to the "royal road" to resolve these challenges.

2.5.2. User-Centered Design and Design Study Methodology

User-centered Design User-centered design can be described as an iterative process including the characterization of the context of use, the specification of requirements, the development of design prototypes and solutions, and the evaluation against the requirements [TM04] [PVW09] [LD11]. The goal of user-centered design is "to lead to the development of usable designs" [Hea09]. In a user-centered design approach, the "needs, wants, and limitations of the end user are given extensive attention at each stage of the design process" [vW06]. In addition, usability is a highly relevant quality criterion. For this purpose, it is important to understand the context of use and to characterize who the users are, what they know, and what they want [Mun09]. Data scientists need to gain an understanding of the tasks domain experts execute in their work environment [MA14]. However, the specification of requirements based on the characteristics of the data, the users, and the users' tasks is non-trivial for exploratory visualization [TM04] [vW06] [Mun09], particularly if time-oriented data is involved [MA14]. For the problem characterization, a frequent practice in user-centered design is a combination of methods including interviews and observations [SMM12], such as conceptual inquiry [LBI*12]. Moreover, a close collaboration including the involvement of domain experts is relevant in the entire course of the project [TM04]. Van Wijk outlines the strength of domain experts in explaining what they do and don't like, while data scientists are good at mastering design challenges with new solutions [vW06]. The manner of how domain experts perceive and interact with a visualization design influences their understanding of the data, as well as the system's usefulness [TM04]. These human factors contribute to the visualization process and should play an important role in the design and evaluation phase. In this connection, in user-centered design approaches, it is desirable to iterate over design solutions. Moreover, providing multiple candidates at least for critical steps in the project contributes to the design process. Evaluation is yet another important principle and should be conducted at least in the late phases of the project. For the assessment of success, the requirements specified earlier, can be used as a baseline in the evaluation phase. As a result of these principles and activities, user-centered design has the potential to yield usable and useful solutions [TM04] [vW06] [LD11].

The user-centered design process has also been formalized in the ISO-standard 13407 "Human-centred design processes for interactive systems" [30]. ISO 13407 can be seen as a description of best practice in user-centered design. It provides a guideline for user-centered design, including principles and activities carried out throughout the life-cycle of interactive systems. The standard describes four human-centered design activities illustrated in Figure 2.12. These four activities are based on the following principles of human-centered design:

- active involvement of users
- appropriate allocation of function to system and to user

[30] Human-centred design processes for interactive systems, https://www.iso.org/obp/ui/#iso:std:52075:en, last accessed on October 08th, 2015.

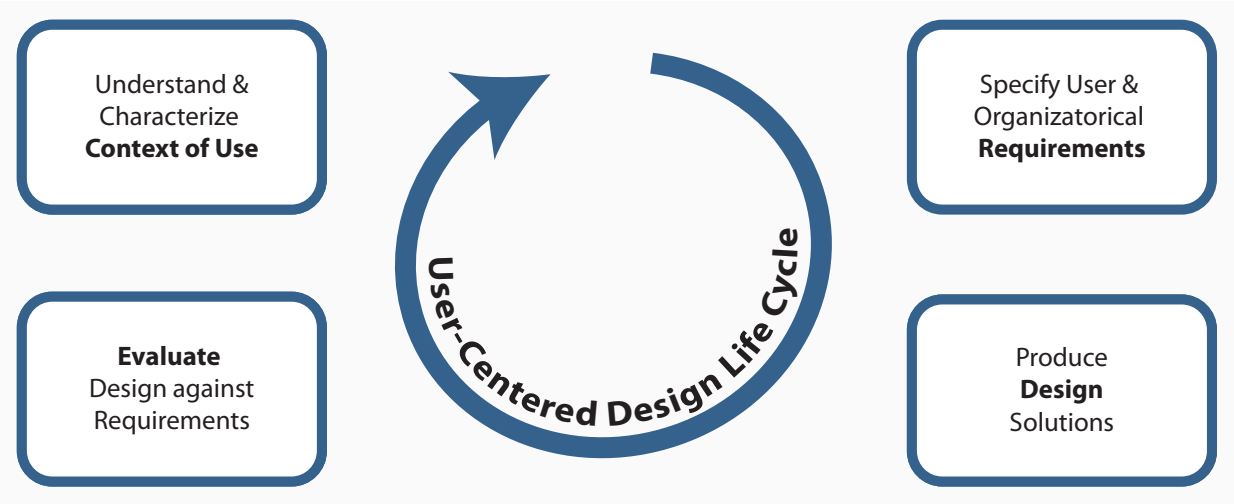


Figure 2.12 *User-centered design life-cycle, adapted from [LD11].*

- iteration of design solutions
- multi-disciplinary design

A pitfall of user-centered design is the knowledge gap between domain expert and the data scientist [vW06]. User-centered design requires amounts of time and energy to be successful [vW06]. A project is more likely to fail if data scientists miss to familiarize with the domain. This may contribute to the fact that the demand for multi-disciplinary collaborations and the necessity to understand scientists' practices, e.g., in chemistry, biology, or DLs, remains high [BWE06, AKM*09].

As another pitfall in user-centered design, domain experts do not necessarily need beyond-state-of-the-art solutions from IV and VA. This pitfall sticks well to the saying "if all you have is a hammer, you may be led to believe that every problem is a nail". In fact, a more appropriate solution is to adapt and combine established techniques, and supplement it with an easy-to-use interface tailored to the domain of the domain expert [vW06]. This type of approaches is in line with the design study methodology reviewed in the next section. In this connection, it is important to mention that the term established needs to be interpreted relative. In many successful design studies, established IV and VA techniques were confirmed to be very innovative in various application domains.

A third pitfall in user-centered design is neglecting the other two factors of the design triangle [MA14] (data and task). In Section 2.3, we elaborate that time-oriented data is a complex data type and associated user tasks need to be characterized carefully. Being focused on users too strongly may run the risk of missing important data and task characteristics. However, what a user wants may not always be what a user needs. We recall the claim of Pretorius et al. "what does the user want to see" and "what do the data want to be" [PVW09]. IV designers should not only rely on statements of users, but also on an in-depth examination of the data. Pretorius and van Wijk argue that especially the combination of both gaining insight by focusing on user data and carrying out requirement analysis results in more effective visualization techniques. "By switching between these perspectives, the information visualization designer can often address an impasse that would have been difficult otherwise. Also, one perspective often sheds light on the other" [PVW09]. We review data-centered challenges according to Fuchs and Hauser [FH09].

- How can we capture all the relevant information?
- How can we combine relevant quantities of information belonging to the same location in space?
- What quantities can be derived from the data that will help users to understand the data?
- What are the features in the data and how can we combine VA techniques for finding them?
- Which visualization technique is most appropriate to the given data?
- How can we combine different visualization techniques for efficient and effective visualization?

Yet another pitfall in user-centered design regards the domain experts. Not only the data scientists, but also domain experts need to contribute to the design process to guarantee a successful collaboration. Collaborative projects are more likely to fail if domain experts have no time or no interest. Similarly, to a certain extent, domain experts need

Nine-Stage Framework	Stage Description
PRECONDITION	
Learn	Providing solid knowledge of the visualization literature
Winnow	Identification of most promising collaborations
Cast	Awareness of different roles in collaborations
CORE	
Discover	Problem characterization & abstraction
Design	Data abstraction, visual encoding & Interaction
Implement	Prototypes, tool & usability
Deploy	Release & gather feedback
ANALYSIS	
Reflect	Confirm, refine, reject, propose guidelines
Write	Design study paper

Table 2.6 *Nine-stage design study methodology framework classified into three top-level categories presented by Sedlmair et al. [SMM12]. The authors emphasize the iterative dynamics of the process.*

to be open for new ideas, e.g., as a momentum of curiosity or the need for new solutions. However, a variety of studies discuss a missing trust of domain experts in unfamiliar techniques and a refusal to change their working routines [BWE06] [HPK08] [KMS*08]. Finally, it would be beneficial for the data scientist if the characterization of the domain experts' problems could be generalized [vW06]. In this case, the interest gap described earlier can be reduced significantly.

Design Study Methodology We review design studies based on two surveys on design study methodology; the *nested model* of Tamara Munzner [Mun09] and the *reflections from the trenches and the stacks* by Sedlmair et al [SMM12]. In addition, we retrospect Miksch and Aigner's design triangle which eases the design in distinguishing three main factors: the characteristics of data, users, and users' tasks [MA14]. The framework guides the design of VA methods resolving in particular time and time-oriented data.

A design study can be defined as a "project in which visualization researchers *analyze* a specific real-world problem faced by domain experts, *design* a visualization system that supports solving this problem, *validate* the design, and *reflect* about lessons learned to refine visualization design guidelines" [SMM12]. A design study approach is reasonable if the problem to be solved includes (1) a sufficient amount of data and (2) a task clarity too imprecise for being solved automatically by a machine. Design studies require a careful *analysis* of the targeted domain to translate data and tasks [MA14] into abstractions [Mun09] that can be used by domain experts. For design studies it is mandatory that real users and real data are involved allowing to solve real-world problems. A key role in design studies plays the *design* itself, requiring knowledge of the space of possible solutions [PVW09, MA14]. Design can be defined as "the creative process of searching through a vast space of possibilities to select one of many possible good choices from the backdrop of the far larger set of bad choices" [SMM12]. Especially the analysis and the design phases are highly iterative and cyclic [PVW09, SMM12]. Another important aspect in design studies regards *validation* [Mun09] (evaluation, e.g., [PVW09]). Validation is highly appropriate at different stages of the design process drawing on techniques like justification according to known principles, qualitative analysis, informal expert feedback, and post-deployment field studies [SMM12]. Munzner's nested model [Mun09] can serve as a guidance in the validation of the problem analysis and the visualization design. To carry out research on design studies, Sedlmair et al. emphasize the value of *reflection* by confirming, refining, rejecting, or by proposing guidelines.

Sedlmair et al. structure the process of a design study in three core phases; a precondition, a core and an analysis phase. At a more fine-gained level, the authors postulate a nine-step framework, shown in Table 2.6. The three main contributions of a design study are as follows [SMM12].

- Problem characterization and abstraction
- Validated visualization design
- Reflection on the design study

We next present an overview of Tamara Munzner's nested model for visualization design and validation [Mun09]. Both the VisInfo case study in Section 7.1 and the MotionExplorer case study in Section 7.2 are influenced by the nested model. Munzner uses the metaphor of a nested model since the four described layers have dependencies between each other: the output from a level above is input to the level below. In this way, Munzner emphasizes the problem of upstream errors which are inevitably cascaded to all downstream levels. To resolve these problems, the nested model provides guidance for individual validation strategies for each of the four levels. The validation strategy of the nested model is based on different *threats to validity* requiring individual means of validation and justification. Two different types of validation approaches exist. On the one hand, *immediate* validation can be carried out within a distinct layer without having dependencies to other layers. On the other hand, *downstream* validation requires validation from downstream levels nested with a particular layer. Downstream dependencies add to the challenge of validation, however, they are necessary to guarantee meaningful designs. The model consists of four nested layers.

In the **domain characterization** phase, the data scientist learns about the tasks and data of target users in some particular target domain (cf. [MA14] for time-oriented data). Typical challenges are to build a common vocabulary, or to get confident with existing workflows. In this connection, we refer to challenges resulting from gaps between data scientists and domain experts described in Section 2.5.1. A positive example regarding the domain characterization within a design study is the LiveRAC system [MMKN08] for the visual-interactive exploration of time-oriented system management data. The authors accurately describe the system management domain and the involved roles leading to a simplified characterization of user tasks.

The **data/operation abstraction design** phase is to "map problems and data from the vocabulary of the specific domain into a more abstract and generic description, i.e., in the vocabulary of computer science" [Mun09]. An operation is defined on a generic level. As such, an operation rather matches tasks described in abstract task taxonomies than in low-level taxonomies (cf. Section 2.3.2). A highly relevant aspect in the data/operation abstraction phase regards the transformation of raw data into the data types that visualization techniques can address. This goal is in line with our technical contribution for visual-interactive time series preprocessing presented in Chapter 4. Munzner criticizes the lack of discussions of the choices made in abstracting domain-specific tasks and data to generic operations and data types. A positive early exception is the characterization of the scientific data analysis process by Springmeyer et al. [SBM92]. The authors present an operation taxonomy based on an empirical study observing scientists of different disciplines while they were analyzing their data. The results of their characterization of scientific data analysis revealed that the data-centered research activity of scientists was partially beyond the data analysis capability supported by computers at the time.

The **encoding/interaction design** defines the third level of the nested model. From an IV and VA perspective, the encoding/interaction design phase is maybe the best explored layer. Works on graphics and perception [Ber83, Tuf86, Tuf90, War12], IV [CMS99, AMST11], and VA [TC06] [AA06] [KAF*08] [KKEM10] can serve as a profound set of guidelines. In this level, Munzner includes the steps of visualization and interaction design which both should be discussed explicitly and clearly. We present a positive example of visualization design and validation related to our technical contribution on content-based overviews (cf. Chapter 5). In his work on energy models for graph clustering, Noack presents two variants of LinLog energy models [Noa07]. LinLog models compute graph layouts focusing on revealing cluster structures of graphs. Noack distinguishes between energy models (defining the layout) and energy minimization algorithms (specifying how to compute these layouts). Noack clearly describes the designs of the two energy models and presents a validation of the models based on a qualitative discussion of the results in comparison to the state-of-the-art. The clear differentiation between the energy models (encoding design layer) and the energy minimization algorithms (algorithm design layer) uncouples his contribution from the downstream algorithm design layer. Furthermore, Noack clearly states his upstream assumptions to the data/operation abstraction design layer: The input for LinLog models requires the abstraction of the data to a node-link graph. As a consequence, the approach is also independent of a particular domain which makes a domain characterization obsolete.

Finally, the **algorithm design** represents the fourth layer. The context of algorithm design is based on an automation of the visualization and interaction designs. Munzner refers to the general computer science domain in which the topic of algorithm design is discussed in detail.

2.5.3. Evaluation

We conclude the related work section about user-centered activity with a review of evaluation methodology and techniques. While evaluation does not necessarily require the involvement of users, user-centered design and design study methods necessarily require evaluation. The technical contributions presented in Chapters 4, 5, and 6 rigorously advocate user-centered design. In addition, both the VisInfo and the MotionExplorer case study presented

Design Process	Category & Description	Surveys & Refs
Design & Eval.		
Nine Stages		
Learn	Visualization literature	[SMM12]
Winnow	Select promising collaborations	[SMM12]
Cast	Identify collaborator roles	[SMM12]
Discover	Problem characterization & abstraction	[SMM12] [Mun09]
Design	Data abstraction, visual encoding & interaction	[SMM12] [Mun09]
Implement	Prototypes, tool & usability	[SMM12]
Deploy	Release & gather feedback	[SMM12]
Reflect	Confirm, refine, reject, propose guidelines	[SMM12]
Write	Write design study paper	[SMM12]
Three Stages		
Exploratory	Often done before the design, "Can I understand more?"	[ED06] [PVW09]
Formative	Often done before the deployment, "Can I make it better?"	[ED06] [LBI*12]
Summative	Often done during deployment, "Is it correct?"	[ED06] [LBI*12]
Nested Model		
Domain Problem	Domain problem characterization	[Mun09]
Abstraction	Data/operation abstraction	[Mun09]
Design	Encoding/interaction technique design	[Mun09]
Algorithm	Algorithm design	[Mun09]
Five Stages		
Pre-design	E.g., understand users' work environment and workflow	[LBI*12] [PVW09]
Design	E.g., visual encoding and interaction design	[LBI*12] [PVW09]
Prototype	E.g., has visualization achieved its design goals?	[LBI*12] [PVW09]
Deployment	E.g., assess the visualization's effectiveness	[LBI*12]
Re-design	E.g., improve a current design	[LBI*12]

Table 2.7 *Categorizations for the classification of the design and evaluation process.*

in Sections 7.1 and 7.2 are built upon design study methods, and thus various evaluation strategies. In general, evaluation of designs, techniques, and applications is mandatory to support both: scientific evidence and user-centered design principles (such as design studies). Providing scientific evidence is a research field on its own, however, the applied methods heavily depend on the domains conducting evaluation. In IV, and more recently in VA, a variety of methodologies and techniques have been presented. However, the evaluation of visualization techniques is challenging [Pla04]. Contributing reasons are (a) the diversity of data, users, and tasks [MA14] and (b) the fuzzy nature of insight [PVW09], which is the goal of visualization [CMS99]. In the following, we review evaluation and validation in IV and VA based on important supplementing literature [Pla04, ED06, Car08, PVW09, Mun09, LBI*12, SMM12].

Desirable Factors in Evaluation Many evaluation strategies are carried out in awareness of three desirable principles [Car08]:

- **Generalizability** (applicability for other users, situations, etc.)
- **Precision** (degree to which one can be definite about the measurements taken)
- **Realism** (context of study should correspond to targeted application context)

The generalization principle aims at applying the results of an evaluation to other people and contexts not having been involved in the study [Car08]. The precision describes the degree to which one can be definite about the measurements that were taken in the evaluation [Car08]. The precision may suffer from environment factors not intended to be studied. The realism reflects the context of an evaluation with respect to the targeted application environment [Car08]. The realism of experiments may suffer from factors like synthetic instead of real-world data, too few data, non-experts, or laboratory environments. Ideally, evaluation techniques support all three desirable factors. However, these three factors are to a certain extent opposed to each other. For example, evaluation strategies following realistic measures taken in-field may lack of precision since many environmental factors cannot be excluded from the evaluation. In return,

Categorizations	Category & Description	Surveys & Refs
Categorizations		
Eval. Goal		
Predictive	Compare design alternatives, compute usability metrics	[PVW09] [LBI*12]
Observational	Understand user behavior and performance	[LBI*12]
Participative	Understand user behavior, performance, thoughts, experience	[LBI*12]
Eval. Criteria		
Functionality	Whether the system provides required functionalities	[TC05] [LBI*12]
Effectiveness	Whether the system provides value, e.g., new insight	[Mun09] [LBI*12]
Efficiency	Whether the system achieves a better performance	[TC05] [LBI*12]
Usability	Readability, how easy users interact with the system	[PVW09] [LBI*12]
Usefulness	Whether the system is useful and user may benefit from it	[Mun09]
Eval. Scope		
Work Envir.	Adoption, productivity, satisfaction	[TC05]
System	Efficiency, effectiveness, satisfaction, scalability	[TC05]
Components	Speed, accuracy, scalability, identification of limitations	[TC05]
Duration		
Short-term	E.g., focus on components, learning of novice users	[Pla04]
Longitudinal	E.g., capture and multiple setups, change variables	[Pla04] [LD11] [LBI*12]
Group Design		
In-group	Same subjects for multiple tests, (cf. longitudinal study)	[LBI*12]
Between-group	2+ groups of subjects, different factors tested simultaneously	[LBI*12]
Location		
Laboratory	Controlled conditions, lack of realism, more precise	[Car08] [LBI*12]
Field	Real world situation, hard to control, more realistic	[Car08] [LBI*12]
Formality		
Formal	Mathematically computed and summarized data/result	[Mun09]
Informal	Opportunistic, feedback, inform testing, anecdotal evidence	[Mun09]
Measure Type		
Qualitative	informal, interviews, requirements, feedback, insight	[Car08] [Mun09]
Quantitative	Statistics, measures, high accuracy & precision & significance	[ED06] [Car08]
Eval. Target		
Verification	"Whether one has built the product right"	[Mun09]
Validation	"Whether one has built the right product"	[Mun09]
User Focus		
Expert	Realistic, understand problem & data, but often too few	[Pla04] [ED06]
Non-expert	Little knowledge, many, interaction or perception experiments	[ED06]

Table 2.8 *Categorizations for the classification of evaluation techniques.*

laboratory experiments yield higher precision but may lack of realism. Furthermore, the generalizability depends on the magnitude of theoretical vs. experimental evaluation, as well as the degree of domain specificity. It is therefore necessary to chose the evaluation technique with respect to the particular evaluation goal [Car08].

Categorizations for the Classification of Evaluation Techniques A systematic review of evaluation papers, evaluation surveys and works on evaluation methodology revealed a variety of categorizations. Subsequently, we present different categorizations classifying evaluation techniques. In this connection, the work of Lam et al. is encouraging for this thesis [LBI*12]. We are aware of two things. First, we involve a variety of terms whose meaning may vary considerably in different application domains. We will use the termini as proposed in most relevant works on evaluation methodology [Pla04, ED06, Car08, PVW09, Mun09, LBI*12, SMM12]. Second, some of the categorizations may have dependencies between each other. In some cases, the categorizations may even be fully correlated. As an example, the distinction between formal and informal evaluations and the distinction between qualitative and quantitative measure types is highly correlated. However, we list the categorizations individually to avoid that relevant termini are disregarded. We present the categorizations of evaluation techniques in two tables. In Table 2.7,

we present categorizations depending on the design process. Thus, the different evaluation categories within each categorization differ by the chronological application. Table 2.7 is highly relevant to associate evaluation with design study methodology. In Table 2.8, we present categorizations revealing different technical evaluation approaches. This is beneficial for the classification of evaluation techniques by relevant factors.

Definitions of Evaluation Strategies used in this Thesis We conclude the survey of user-centered design and evaluation strategies with the definition of four evaluation and validation techniques relevant for this thesis. Once again, we emphasize the relevance of design study methodology. This is why we presented an in-depth characterization of design studies in Section 2.5.2. In the following, we define a *use case*, a *usage scenario*, and a *case study*.

Use Cases are sets of actions carried out by users interacting with systems to achieve a goal. In this thesis, we further characterize use cases as scenarios describing the specific combination of users, data, and tasks (problems). Data scientists aim at creating analytical systems allowing users to carry out use cases with a successful outcome.

Usage Scenarios may be based on real data, but are conducted by the developers themselves, rather than by domain experts [SMM12]. Usage scenarios are a means of demonstration of findings and a legitimate form of validation. However, they are not as convincing as studies reporting from the wild, involving real users with real problems.

Case Studies (field studies) are evaluation strategies aiming for the validation of tools [SMM12]. Necessarily, case studies include real users, real problems, and real data. Many design studies feature strong case studies as they provide a similar setting, including real users, real problems, and real data.

2.6. Research Challenges for this Thesis

We conclude the overview of related works with a condensed summarization of remaining challenges for the ES in time-oriented primary data. In the following chapters, we will confront these general challenges by transforming them into concrete research goals. These research goals build the basis for our concepts, guidelines, techniques, and systems presented in this thesis.

C_{MES} Methodology for the Design of Exploratory Search Systems The review of ES, IV, and VA in Section 2.1 revealed the common ground, joint techniques, and similar objectives of these fields. On the one hand, IV and VA induced a great variety of visual-interactive solutions, some of which were already adopted in implementations of the ES concept. On the other hand, ES is a valuable application field for IV and VA, resolving research challenges based on information seeking behaviors ranging from search to exploration. In general, information-seeking as, e.g., defined by Marchionini [Mar95], is a common denominator of ES and IV [WR09]. This also applies to the younger field of VA which, however, is significantly less conceptualized in the ES context so far. In the past, a variety of best-practice approaches for ESS have been presented, making massive use of IV and VA techniques. However, the vast majority of best-practice *ESS only accepts textual data content*. In fact, we identify a general lack of ESS solutions for non-textual documents. For time-oriented primary data, hardly any ESS have been presented thus far. The review of related works also reveals a missing reflection of ESS which certainly relies on the missing existence of best practice ESS for non-textual data. A rare example reviewing and reflecting existing ESS is the survey of Herrmannova and Knoth [HK12], even if it is limited to approaches focusing on textual data content. Since ES is also very promising for non-textual data types (cf. Sections 2.1, 2.2, and 2.3), the question arises why implementations of the ES concept for non-textual data types are particularly rare. We comply with the challenges emphasized in a review of visual search approaches for complex data types. According to the authors, challenges are based on difficulties in the (visual) formulation of queries, in the definition of similarity functions, as well as in providing meaningful visualization and interaction designs for the targeted data types [vLSFK12].

We deem the lack of missing best-practice ESS for non-textual documents a methodological challenge. The missing reflection of best practices contributed to the fact that ES has not yet induced a general methodology for the design of ESS. In turn, data scientists are left alone with the complex data types, the huge design space, and the requirements of the targeted user group. Subsequently, we examine three types of methodological challenges in detail all hampering the design of ESS.

- Missing summary of best-practice techniques from IV and VA applicable for ESS (cf. Section 2.1.2)
- Missing utilization of reference models, frameworks and workflows in KDD, IV, and VA (cf. Section 2.4)
- Missing involvement of design study methodology known in IV and VA (cf. Section 2.5)

Research in IV and VA has induced a great variety of visualizations and interaction techniques for both textual and non-textual data types. A significant part of these techniques is also relevant for the design of enhanced ESS. The question arises whether a more transparent and target-oriented *summarization of IV and VA techniques* can be achieved to support the process of designing enhanced ESS. The focus of interest is the variety of surveys and taxonomies of user tasks presented for IV and VA (cf. Section 2.1.2). These taxonomies provide valuable overviews of the rich set of tasks addressed in IV and VA, albeit these taxonomies differ in their perspectives and levels of abstraction. All taxonomies deal with the question of how users can be supported in their interactions with analytical systems, i.e., in the information-seeking process drawing on computer-supported tools [Mar95]. However, the identification of common features between tasks and techniques presented in IV and VA on the one hand, and implementations of the ES concept on the other hand, remains largely uninvestigated. Thus, we identify a gap between the ES concept and IV and VA tasks/techniques, which hampers the design process of ESS. This methodological challenge is associated with the challenge to "lower the cost of visualization generation" and allow it to become a more integral part of scientific workflows [FH11] for downstream ESS.

The second methodological challenge is the lack of analytical *workflows* presented for ESS. In turn in KDD, IV, and VA a variety of reference models and (scientific) workflows have been presented supporting data scientists in the design of data analysis workflows (cf. Section 2.4). Similarly, we reviewed the life-cycle of primary data including six reference stages (cf. Section 2.2.2). A conceptualization of the ES workflow for the design of ESS, however, is entirely missing as it can, e.g., be seen in the survey of White and Roth [WR09]. For data scientists, it remains a challenge to combine the various existing (scientific) workflows, methodologies and frameworks to generate a single, reproducible reference workflow for the design of ESS. We emphasize two factors which additionally contribute to the challenge for time-oriented primary data. First, the high level of abstraction of existing reference workflows and framework impedes the direct application for concrete data and tasks, such as the ES in time-oriented primary data. The VA process presented by Keim et al. [KKEM10] may serve as an example of a valuable, but too abstract process model. Second, the number of workflows and frameworks which explicitly focus on time-oriented (primary) data is scarce. The conceptual frameworkfs by Aigner et al. [ABM*07] and Andrienko and Andrienko [AA13] serve as welcome exceptions. It remains challenge to divide the entire workflow into meaningful steps which are a) specific to solve particular challenges within the workflow creation and b) general to be valid for a variety of ES implementations.

The third methodological challenge applies to *design* of analytical workflows and the engagement of users. For the design of ESS, e.g., to support data-driven research, it is highly important to involve the targeted domain experts (cf. Sections 2.2, 2.3, and 2.5). Gaining a clear understanding of the different information needs is a prerequisite for the design of usable and useful ESS. Despite this fact, we identify a lack of research approaches emphasizing the involvement of users in the design. In addition, existing (scientific) workflows, methodologies and frameworks for IV and VA solutions only focus on the data and control flow, but not on the design of respective systems (cf. Section 2.4). Most current reference models do not differentiate between designers and users, or between the design and the application phase, respectively. While users are typically described as actors in the application of systems (see, e.g., the Card pipeline [CMS99]), reference models lack of assigning users an active role in the design phase. In turn, research on design study methodology [Mun09, SMM12] poses a counterpart to the latter described frameworks (cf. Section 2.5.2). In contrast to the reference models described above, design study methodology explicitly emphasizes the design of IV and VA systems including user involvement and evaluation. As such, design study methodology complements the data and model-centered frameworks. However, one challenge of guidelines for design studies is the abstraction from the data, tasks, and algorithmic models included in the design process.

C_{CBA} Content-Based Access to Time-oriented Primary Data The review of ES and best-practice ESS in Section 2.1 revealed the key role of content-based access strategies to facilitate search and exploration support. Similarly, content-based access builds the baseline principle for the construction and execution of scientific workflows for primary data (cf. Sections 2.2 and 2.4.1). Finally, in time series retrieval, time series DM, IV, and VA content-based access forms the basis for most analytical approaches for time-oriented data (cf. Section 2.3). We reviewed content-based access from a data scientist's a domain expert's and a digital librarian's perspective in Section 2.2.3. We identified that search-oriented fields, such as the DL domain, IR, and research on visual search are confronted with challenges emerging from complex, non-textual data types. From an exploration perspective, it is challenging to gain insight into the structures of large and complex primary data. Data scientists have to face problems associated with various factors making the design of individual content-based access solutions challenging. In the following, we further elaborate general challenges in content-based access, specific challenges for primary data, and specific challenges associated with time-oriented data.

In general the utilization of *raw data content* for enhanced analytical models is hardly feasible. In fact, FV-based approaches have proven to be very effective. However, capturing the relevant features of complex and potentially unknown data objects is deemed particularly challenging in general [FH09, KH13]. The extraction of features is accompanied by challenges posed by preprocessing steps necessary to transform the data into appropriate formats. Example steps include data cleansing, normalization, sampling, or segmentation. In many real world use cases, the challenges induced by this design space should be resolved by involving domain experts in the process. Domain experts may provide, e.g., characterizations of the data, notions of data similarity, notions of interesting data relations, and domain knowledge for the design of useful data abstractions. Other challenges for content-based access strategies occur when the FVs serve as the input for models of complex analytical workflows. In many cases, such workflows require a certain level of data quality, the definition of similarity, or compact representations. Finally, the complexity of many non-textual data types adds to the challenges associated with content-based access.

The *heterogeneity* of primary data is a great challenge for many data-driven research approaches (cf. Section 2.2). Contributing factors are different data formats, missing data standards, or data not corresponding to existing standards. The phenomena described by primary data may be extremely complex and require significant time and effort to be explored and understood [PVW09]. It is challenging to derive the inherent information from the data and to apply meaningful preprocessing and feature extraction steps [FH09].

Diagnosing data *quality* issues and manipulating data into a usable form is yet an additional challenge for large and complex data in general [KHP*11]. For primary data in specific, this challenge is likely to be more severe in many cases. While quality leaks in secondary data may already be approached, raw primary data may contain a variety of issues related with data quality. Being a direct data product of a given source the data quality depends on factors, such as the environment, the sensor, or the measurement technique, all of which add to the challenge of data quality. Also human activities have an impact on the quality, e.g., when human judgment is involved in the data gathering process. For the exploitation of the value of primary data, it is most relevant to face the challenges of representing data quality [LK06] and manipulating data into a usable form [KHP*11].

The *size* of many data collections is the third major challenge for primary data we want to make explicit. Emerging technologies for collecting and storing primary data still exceed the scientists' abilities to process, analyze, interpret, and draw conclusions on data [BWE06, AKM*09, AKD10]. Dealing with this 'data deluge' ('information overload' problem) is still one of the most difficult challenges for data-driven research and for IV and VA in general [KMS*08]. To understand the complete body of information, enhanced analytical techniques are needed to condense the available information into manageable and human-readable parts.

Time-oriented data induces an additional magnitude of data complexity. Phenomena represented as time-oriented primary data consist of a *value* domain, which is associated with a *temporal* domain (cf. Section 2.3). Different mathematical properties, design aspects, and data abstractions require individual treatment (cf. Section 2.3.1). Providing meaningful data representations, data operations, and visual representations for time-oriented data is a challenging task on its own. In addition, the quality of time-oriented (primary) data requires individual treatment to gain accurate insight [GGAM12]. Similarly, time-oriented data requires a variety of individual steps for data processing, data transformation, feature extraction, and for the definition of similarity to match the information need of the users and downstream operations, respectively (see Section 2.3.2). Finally, effective visual representations and interaction designs for time-oriented data remain a future research challenge [LK06, FH09, AMST11], e.g., to cope with specific application challenges.

C_{CBO} Gaining an Overview of the Data Content Content-based overviews of large data collections are a particularly appropriate starting point for exploratory information seeking [Shn96, CMS99, GMPS00, YKSJ07, KMS*08, CKB09, KH13]. In this way, content-based access strategies can be used for enhanced ESS. Overviews reveal structural information of the data collection, such as frequently occurring patterns, or interesting relations between patterns. The value of content-based overviews is widely accepted for ESS, even if most best-practice approaches are limited to textual data content (cf. Section 2.1.3). Only a few of ESS approaches provide content-based overviews for non-textual data, particularly for time-oriented data. Some approaches based on music collections may serve as welcome examples (cf. Section 2.1.3). For time-oriented primary data, however, hardly any best-practice ESS exit. Most of the challenges of content-based overviews are associated with the huge design space. In the following, we elaborate important factors, all of which contribute to the challenge of designing powerful content-based overviews.

- Upstream data set (i.e., high-dimensional time series features)
- Different consecutive steps in the design process
- Visual representation of content-based overviews

- Preferences of the involved user group
- Ability of content-based overviews to formulate (visual) queries
- Downstream ESS and additional requirements

The challenge of *upstream data* sets is associated with the challenge of providing content-based access strategies $\mathbf{C}_{\text{CBA}}$. For the review of remaining challenges for content-based overviews, we assume effective and efficient content-based access strategies as an appropriate upstream prerequisite, such as the creation of FVs. However, understanding the high-dimensional data input (i.e., features) is still a challenge for the design of content-based overviews. At least in early phases of the design, data scientists are challenged to create content-based overviews without knowing whether the targeted solution matches the characteristics of the data. The challenge becomes severe if data scientists need to fully understand the input data to supplement missing user requirements [PVW09]. Thus, the design of content-based overviews is to some degree a hen-and-egg problem since an overview would be needed to create meaningful overviews. An additional challenge of content-based overviews is the *cascade of different models* (operations) required for the design. This process is similar to the construction of exploratory data analysis workflows in general [Fek13]. In order ensure scalability for large data sets effective data abstraction, aggregation and visualization concepts are highly important [LK06]. It remains a challenge to assess the quality of data abstractions and to factor user judgment. Similarly, the open research question remains to which extent data scientists can be *guided* towards appropriate solutions with algorithmic support. Closely related to data abstraction and aggregation is the need of explorers to interact with the data provided in content-based overviews at *different levels of details* [LK06].

It is a challenging task to provide meaningful *visualization and interaction designs* that facilitate the exploration of content-based overviews at different levels of detail. More specific, the visual representation of data abstractions poses a separate branch of design challenges. To provide meaningful overviews an overall structure needs to be imposed on the data, such as hierarchical structures [EF10], or spatial structures preserving the similarity of the data elements (i.e., layouts) [LK06]. However, *revealing structure from high-dimensional input data* to align visual data representations in the display space constitutes a subject to research in IV and related fields. To support the information drill-down to local aspects of interest, meaningful browsing techniques need to be provided. Interaction challenges for visual interfaces associated with browsing are, e.g., to abstract/elaborate, filter, select, connect data [YKSJ07]. The *involvement of users* in the design is another factor in general, we refer to challenge $\mathbf{C}_{\text{UCD}}$ elaborating this challenge in detail. One of the shortcomings of many existing content-based overview applications is the *lack of justification of design choices*. This impedes the validation and hampers reflections on the design process in general, e.g., to derive a general design methodology. In addition, merely presenting final results masks the iterative design process (cf. Section 2.5). Yet an additional challenge refers to the formulation of (visual) queries based on content-based overviews. Content-based overviews are highly appropriate for the *Query-by-Example* technique [vLSFK12], which is most beneficial for ES, not only for textual content. However, it remains a research challenge to implement the Query-by-Example concept for ESS for time-oriented primary data. Finally, the downstream analysis systems, such as ESS, pose an additional challenge for content-based overviews. Enhanced analysis systems provide sophisticated visualization and interaction designs which range over different (complementing) views. Linking concepts for the localization of data objects in *different views* include highlighting strategies or the use of color. It is an additional challenge for the design of content-based overviews to match these strategies of downstream ESS.

$\mathbf{C}_{\text{C+M}}$ Combining Data Content and Metadata In Section 2.2, we provided an overview of primary data including different characteristics of data content and metadata, as well as concepts and techniques for relation seeking. The temporal dependency of the data contributes to the complexity of the data type (cf. Section 2.3). For heterogeneous, multi-attributed and possibly multi-modal primary data, seeking relations between different attributes is most relevant to understand the data at a glance. Relation seeking is considered an exploratory activity, although revealed relations also facilitate the design of metadata facets, which in turn support search activity.

Seeking relations between multiple attributes in data content and metadata is challenging in general. These complex bodies of information yield various types of relations, e.g., correlations between two numeric attributes, or speaking more general, phenomena which occur more often than expected. The challenge is worse for complex time-oriented primary data collections where it must be assumed that explorers have only little knowledge about the data set, and thus have only few hypotheses in the first place. Thus, relation-seekers need to be supported with effective and efficient analytical systems allowing an enhanced hypotheses generation process. In addition, browsing through rich sets of data content and metadata needs to be supplemented with guidance concepts leading to interesting relations. It is a particular challenge to capture the notion of interestingness in the heads of domain experts and to put it into formal models. Concepts using frequent user behavior or semantic annotations (such as recommender systems) to assess

the interestingness of relations will not work since primary data is often less frequented by users, and ontologies are missing. The use of interestingness measures (e.g., statistical significance testing) needs to be evaluated for their applicability for relation seeking (cf. Section 2.3). This applies to time-oriented data, but also to the complete body of available information including metadata. The combination of both data content and metadata causes challenges since it must be assumed that these both data types are heterogeneous. Metadata typically consist of various attributes of numerical, ordinal, or nominal type. In return, the time-oriented data content is a specific data type with individual properties (cf. Section 2.3). The functional definition of interesting relations in heterogeneous data types is particularly challenging and user-dependent. Finally, the different information needs of domain experts contribute to the challenge. If the domain experts' intent is search activity, e.g., to carry out confirmatory analysis, relation seeking must support the validation of hypotheses. However, if the information-seeking behavior of domain experts is exploratory, relation seeking between data content and metadata should support the generation of new hypotheses. This also involves the factor that time-oriented data does not necessarily form the dependent variable in the analysis process. In fact, Andrienko and Andrienko refer the identification of temporal locations for a given set of target identities to as a localization task [AA06]. The assessment of dependencies between some given data subsets and the temporal information provided with the data set poses challenges to the design of algorithmic models, visualizations and interactions.

$\mathbf{C}_{\text{MPC}}$ Model and Parameter Choice Exploratory data analysis, research on scientific workflows, and other fields related to data analysis share the goal to maximize the degree of automation of the data-centered workflow execution. Automated data processing, however, is only reasonable in scenarios where the data is entirely understood and no human judgment is needed [KMS*08]. In Sections 2.1, 2.2, and 2.3, we provided an overview of analysis tasks and the large variety of models which can be applied to time-oriented primary data. The combination of models and their parameter sets leads to a large design space. The challenges on content-based access $\mathbf{C}_{\text{CBA}}$, content-based overviews $\mathbf{C}_{\text{CBO}}$, and combining data content and metadata $\mathbf{C}_{\text{C+M}}$ may serve as examples. Large design spaces bear the challenge of users and analysts getting lost in irrelevant or inappropriately processed or presented information [MA14]. At a glance, data scientists need to face the challenges of finding right models, putting these into a workflow in the right order, and choosing right parameter settings for each of the models.

The *order of models* contributes to the challenge of the complexity of workflows in exploratory data analysis [Fek13]. It is widely accepted that one challenge in real-world applications is the series of heterogeneous problems which need to be solved [KMS*08, p. 88]. For time-oriented primary data, we emphasized challenges in data cleansing, preprocessing, feature extraction, as well as downstream analysis techniques, such as data aggregation and retrieval (cf. Section 2.3). The correlations of solutions for individual steps, their assembly to a single process, and maintaining a meaningful order are challenges on their own. As an example, it is particularly difficult to assess the implications of early steps in the workflow for the entire analysis process. Another challenge to confront is supporting domain experts in reporting the sensibility of their results and the validity of their findings [TLLH13]. Especially for exploratory analysis tasks the classification of success and the decision what makes a good solution is ill-defined [KKEM10, p. 147]. Quantitative statements, more advanced inferential statistics, and continuous feedback may help to verify the statistical significance of certain relations [TLLH13, Fek13]. The classification of success is particularly important for the automation of data processing in a workflow. For data-driven research, human judgment is crucial for the construction of meaningful workflows, ideally in combination with guidance and quality assessment concepts provided by the analytical system. A similar supporting concept is the justification of generalizability of defined workflows. Human judgment is tedious and manual process steering not applicable for large data sets. Ideally, systems support users in testing workflow designs for large primary data collections. The identification of variations in the results caused by the diversity of the input data would be valuable to assess generalizability.

A challenging task at a more fine-gained granularity is *choosing an appropriate algorithmic model* for a given step within the workflow. Finding appropriate models requires a clear understanding of the data, the incorporation of domain knowledge, and the assessment of the model output. For data scientists it is not per se clear which technique will lead to the (user-intended) results [MA14]. In this connection, we echo the demand for solutions closely combining human judgment and automated computation [BAF*13]. In addition, VA concepts, such as visual quality assessment and user guidance, would support the design process [SS02]. Providing meaningful models (and model parameters) would also ensure that the workflow does not report spurious patterns that do not really exist, or greatly overestimates the significance of the reported patterns [KLR04].

Finally, the definition of appropriate *parameter settings for every model* of the workflow poses a particular challenge which often needs supervision by a domain expert [KAF*08]. Most models pose parameters in which the optimal effect of a routine can be achieved for a given goal. One example is the level of data abstraction which is arbitrary

from an implementation point of view. However, finding the optimal level of abstraction from a user's point of view is non-trivial [LK06]. Data scientists can highly benefit from parameter guidance concepts presented in VA research [SS02, IMI*10, SSW*12]. A welcome example visualizing the output of a model with respect to different parameter values is a time series segmentation approach presented by Keogh et al. [KCHP04]. However, this example is an assembly of different output visualizations. The use of a visual-interactive system to be able to compare parameter effects is not part of the approach.

C_{UCD} Involving the User in the Design The design of visual-interactive systems should be based on a detailed analysis of users, their information needs, and their tasks [BC02, vW06, Mun09, SMM12, Fek13, RBK13, MA14]. We characterized different stakeholders and collaborator roles associated with time-oriented primary data in different contexts, such as the information-seeking process, the data life-cycle, the content-based access to primary data, and user-centered design in general (cf. Sections 2.1, 2.2, and 2.5). In addition, we reviewed collaboration problems based on gaps, e.g., at the knowledge-level, at the interest-level, or at the system-level [vW06, BWE06, Fek13, RBK13] (cf. Section 2.5.1). Finally, we observed the relationship between domain experts and data scientists in detail. These two stakeholders define the two most relevant roles for this thesis. In the following, we list particularly important factors in the design of ESS where gaps between data scientists and domain experts need to be bridged to yield usable and useful tools.

An early challenge in the design process is to *gain an understanding of the underlying time-oriented primary data*, to familiarize with the involved users, and to understand their information-seeking behaviors. The missing awareness of the properties of the underlying data and of the involved user group hampers fully-automatic data processing and the usefulness of downstream solutions [KMS*08, SMM12, MA14]. In the language of design study methodology, this challenge is called the data, user, and task characterization/abstraction (cf. Section 2.5). Data-driven research contributes to the difficulty of the challenge since the underlying data set may to some degree be unknown by all stakeholders in the first place.

An additional challenge associated with the involvement of users is the *selection of meaningful models* and the definition of appropriate *model parameters*. For data scientists it is important to ensure the quality of models and model parameters which have a decisive influence on the output of the analytical workflow (cf. $\mathbf{C}_{MPC}$). However, the choice of the models also depends on the analysis tasks of domain experts. In addition, the preferences of domain experts influence the choice of models and model parameters. Especially for complex workflows drawing on large design spaces, it is a particular challenge to include both automated analysis techniques and human judgment in an efficient and effective way [KMS*08]. As an overall goal, data scientists and domain experts are able to fix rich sets of models and model parameters within the design phase, leading to both usable and simplified ESS.

A related user-centered challenge is the *definition of similarity and interestingness functions* required for various powerful algorithmic models. The classical example of searchers is the class of retrieval algorithms requiring carefully designed similarity functions in combination with compact and yet representative FVs. Models relevant for explorers rather focus on structural information of the underlying content supporting to gain an overview (clustering, layout algorithms). Finally, relation-seeking algorithms require interestingness measures to guide explorers towards interesting relations. With the definition of meaningful similarity and interestingness functions, data scientists will be able to provide models which lead domain experts towards relevant data subsets or interesting relations between data subsets. Thus, similarity and interestingness functions are important factors whose definition directly depends on the requirements of the involved user group. Designing similarity and interestingness functions according the notions in the heads of users is a particular challenge. In fact, meaningful definitions of similarity and interestingness require carefully created workflows for processing and transforming time-oriented data into a comparable format. For example, the inclusion of a simple normalization operation into the workflow has a tremendous effect on the similarity function. By using such operations, data scientists are able to adopt the users' notion of similarity.

User studies conducted with domain experts revealed yet another challenge associated with the engaged user group. *Rising the trust of domain experts* in highly sophisticated external solutions is a particular user-centered challenge [BWE06, HPK08, KMS*08]. Similarly, domain experts still underestimate the value of advanced visual-interactive steps within their scientific workflows [NSBW08, TDN11]. At least for parts of their workflows, domain experts still rely on basic analytical toolkits. It is important for the success of ESS approaches to include these domain experts in the construction phase of the workflows to raise the trust, include visualization into the design, and to replace deprecated solutions.

The final challenge we make explicit emphasizes the need for the involvement of users in the *design of visual-interactive interfaces*. While visualization and interaction per se is highly beneficial to execute the information-seeking

process effectively, challenges exist in tailoring visual-interactive interfaces towards the requirements of the engaged user group (cf. Section 2.5). For both visualizations and interactions, there is a large design space, often with multiple valid solutions. To design usable and useful interfaces, data scientists need to resolve challenges associated with providing (multiple) early prototypes, highly iterative design phases, and different evaluation strategies, all by involving users in the design.

2.7. Summary

We conclude the related work chapter with a brief summary. We reviewed the ES of time-oriented primary data from four different scopes. These scopes are conceptual, data-centered, task-centered, an user-centered. We started with an introduction to ES from a conceptual perspective, including overviews of the information-seeking search theory, as well as definitions for IV and VA (cf. Section 2.1). In addition, we reviewed task taxonomies presented in IV and VA. Finally, we emphasized search and exploration activity and presented an overview of best-practice ESS. The second scope of the chapter responded to time-oriented primary data. For the sake of clarity, we presented both primary data and time-oriented data in different sections (cf. Sections 2.2 and 2.3). We started with the characterization of primary data, followed by a review of the life-cycle of primary data. Finally, we emphasized content-based access to primary data from the perspectives of different stakeholders. In the section for time-oriented data, we started with a characterization of time-oriented data. We draw a connection to KDD and time series DM, and presented task taxonomies for time-oriented data. Finally, we presented visual-interactive systems for the search in and the exploration of time-oriented data. The third scope of the related work chapter emphasized the task perspective (cf. Section 2.4). First, we gained an overview of scientific workflows and scientific workflow systems, which are beneficial for data-driven research. Second, we reviewed the KDD process as a baseline concept for enhanced analytical exploitation of (time-oriented) data. Finally, we presented four most relevant reference models and frameworks presented in IV and VA. In particular, the Card pipeline, the VA Diamond, and the two VA frameworks for time-oriented data have a strong influence on the concepts presented in this thesis. The fourth scope of the related work chapter regarded the users' perspective (see Section 2.5). We gained an overview of different collaborator roles engaged in the design process of visual-interactive systems for time-oriented primary data. Moreover, we reviewed the user-centered design process and design study methodology presented in IV and VA. Finally, we presented summaries of evaluation strategies applied within the design process of IV and VA systems. We summarized the chapter with a reflection of related works yielding six central research challenges for this thesis (cf. Section 2.6).

CHAPTER 3

Concepts for Exploratory Search Systems

In the course of this thesis, we will present novel concepts, guidelines, techniques, and systems for the ES in time-oriented primary data. Careful assessment of the related work reveals six major challenges which currently impede the design of powerful ESS for time-oriented primary data (cf. Section 2.6). The technical challenges $\mathbf{C}_{\text{CBA}}$, $\mathbf{C}_{\text{CBO}}$, and $\mathbf{C}_{\text{C+M}}$ will be solved in Chapters 4, 5, and 6. The overarching challenges $\mathbf{C}_{\text{MPC}}$ and $\mathbf{C}_{\text{UCD}}$ define additional requirements for all contributions. In Chapter 7, we put ES into practice in two real-world case studies.

In this chapter, we form the conceptual basis of this thesis. The review of the related work reveals that best-practice ESS for textual data content greatly benefit from concepts and techniques presented in IV and VA. We assume that concepts and techniques presented in IV and VA can also facilitate the design of ESS for non-textual data content, particularly time-oriented primary data. In this way, drawing on concepts and techniques from IV and VA will facilitate the process of resolving the central research challenges outlined in the related work chapter. Due to the lack of existing approaches, especially for time-oriented primary data, we identify a gap between the ES concept on the one hand and the rich set of techniques presented in IV and VA on the other hand. Thus, our main methodological goals of this chapter are as follows.

- Bridging the gap between the ES concept and the rich set of solutions presented in IV and VA
- Adapting existing methodology from IV and VA to the requirements for ESS

In this chapter, we postulate two conceptual contributions. First, we strengthen the association of ES with IV and VA. We identify intersection points based on an assembly of IV and VA tasks relevant for ES and ESS, respectively. In addition, we map these tasks to search and exploration activity. On this basis, we present the definition of ES for this thesis, including an emphasis on IV and VA. Second, we present a reference workflow for the design and application of ESS for time-oriented primary data. The reference workflow divides the design process in four main steps. Moreover, the reference workflow differentiates between the design and the application of visual-interactive interfaces included in ESS. With the utilization of the reference workflow, data scientists generate views which can be integrated into operable ESS. We will use the reference workflow as a framework for guidelines and techniques, as well as for the ESS presented in this thesis. This chapter is influenced by the reflections of our design studies [BDF*15, BWK*13, WVZ*15, BSM*15a], our system for the construction of workflows [BRG*12], and our conceptual works with time-oriented primary data [BBF*10, BBF*11, SBS11].

Contents

3.1. Introduction

3.1.1. Motivation

ES is an emerging concept aiming to combine the strength of both search and exploration activities (cf. Section 2.1). Hence, different information-seeking behaviors can be supported ranging from the lookup and localization of known items to exploratory investigation and learning. Looking back to the review of primary data in Section 2.2, it becomes apparent that ES particularly qualifies for exploiting the value of primary data. Usable and useful ESS have the potential to support data-driven research and to unlock the undiscovered knowledge hidden in primary data. With the focus on time-oriented data, complex temporal real-world phenomena can be captured and be archived in primary data repositories for reuse. The review of time-oriented data in Section 2.3 reveals that in KDD, IV, and VA a multitude of concepts and techniques have been presented which are ready to be applied in both data-driven research approaches and in ESS. Similarly, the review of scientific workflows, reference models and analysis frameworks presented in Section 2.4 indicates the analytical power which can be included to facilitate ESS. Finally, Section 2.5 provides an overview of user-centered design and design study methodology which can be applied to generate both usable and useful visual-interactive systems.

Best-practice ESS show how concepts, workflows, and visual-interactive techniques from IV and VA enhance ES implementations. However, most of today's best-practice ESS are limited to textual data content. In fact, we identify a general lack of ESS solutions for non-textual documents. For time-oriented primary data, hardly any ESS have been presented thus far. On this basis and after a careful review of the related work, we postulate six main research challenges, all hampering the design of ESS for time-oriented primary data (see Section 2.6). From a technical perspective, content-based access to time-oriented data $\mathbf{C}_{\text{CBA}}$, content-based overviews $\mathbf{C}_{\text{CBO}}$, and relation seeking between data content and metadata $\mathbf{C}_{\text{C+M}}$ are among the most difficult challenges. In addition, two salient factors are the overarching challenges of choosing appropriate analytical models and model parameters $\mathbf{C}_{\text{MPC}}$ and involving users in the design $\mathbf{C}_{\text{UCD}}$. In the course of this thesis, we will resolve the technical challenges under consideration of the overarching challenges. Moreover, we will present two real-world ESS for time-oriented primary data making use of the technical contributions.

In this chapter, we focus on conceptual challenges for the design of ESS for time-oriented primary data, i.e., missing methodology for the design of ESS $\mathbf{C}_{\text{MES}}$. ES is a young concept and the number of contemporary ESS is comparably small. Especially the lack of non-textual ESS contributes to the missing reflection of ESS approaches which also explains the still missing design methodology. In IV and VA, however, we identify a wealth of methodology with various latent intersection points with ES.

- Methodology about tasks and visual-interactive techniques (cf. Sections 2.1 and 2.3)
- Methodology about workflows (cf. Section 2.4)
- Methodology about design studies (cf. Section 2.5)

The overarching research question addressed in this chapter is how these different methodologies from IV and VA can be adopted to ES and the design of ESS, to prevent reinventing the wheel. The result would serve as a concept for the design of ESS. Associated questions are how search activity is implemented in IV and VA, especially for time-oriented data. The same applies to exploration activity. Similarly, the (visual-interactive) construction of workflows and the involvement of users in the design of visual-interactive interfaces need to be surveyed. Our goal is to reflect and adopt best-practice concepts and techniques from IV and VA, and obtain a generalizable methodology for the design of ESS. In this way, we also aim to bridge the gap between ES and IV (and VA) by the identification and exploitation of intersection points. As a result, we form a framework with of concepts, tasks, and techniques relevant for the design of ESS.

3.1.2. Research Goals

The related work revealed six major research challenges (cf. Section 2.6) which we will address in the course of this thesis. In this chapter, we explicitly remedy the lack of methodology for the design of ESS $\mathbf{C}_{\text{MES}}$. The methodology has to reflect the three technical challenges of this thesis $\mathbf{C}_{\text{CBA}}$ $\mathbf{C}_{\text{CBO}}$ $\mathbf{C}_{\text{C+M}}$. The methodology also has to consider the associated challenges of choosing appropriate models and model parameters $\mathbf{C}_{\text{MPC}}$ and the challenges of involving the user in the design $\mathbf{C}_{\text{UCD}}$. Based on an assessment of the involved challenges, we postulate five *research goals* which need to be reached to yield a methodology for the design of ESS.

RG_{MES1} Exploitation of Intersection Points Between ES and IV Research in ES methodology can benefit from IV and VA solutions to avoid reinventing the wheel. In this connection, intersection points between ES and IV can play a key role, allowing to adopt existing solutions towards ES in an easy way. However, for non-textual document types these intersection points have to be made explicit. The great variety of solutions presented in IV and VA have to be identified, condensed, and be associated with the requirements for ES and ESS design. This applies the conceptual and the technical perspective. On the one hand, research in methodology for ESS remains an unsolved research goal, hampered by the lack of best-practice ESS for non-textual data and a missing reflection of these rare approaches $\mathbf{C}_{MES}$. On the other hand, research in IV and VA has produced a rich set of methodological contributions which may be adopted for ES and the design of ESS. From a technical perspective it still requires further research in which way best-practice approaches from IV and VA can be put into usable and useful ESS, especially for time-oriented primary data. We have already described tasks, visual-interactive techniques, workflows, and design studies as important factors to be considered. A promising intersection point is the emphasis on (analysis) tasks which are extensively discussed in ES, as well as in IV and VA. It remains a goal to combine the variety of IV and VA tasks presented in different task taxonomies for the identification of intersection points with ES.

RG_{MES2} Definition of ES for this Work The related work presented in Section 2.1 shows different descriptions and characterizations of ES. These works represent strong motivations for ES, e.g., in the way ES can support data-driven research. Today, ES is predominantly described from a conceptual and theoretical perspective, e.g., by characterizations of supported information-seeking behaviors. However, important technical factors for the design and the use of ESS are not included in the definition. In this connection, the definition of ES for this thesis should go beyond conceptual and theoretical perspectives, by making the latent connection of ES with IV and VA more explicit $\mathbf{C}_{MES}$. In addition, to shift the ES concept to a more technical level, the definition should outline mandatory visual interfaces of enhanced ESS.

RG_{MES3} Adoption of Existing Methodology from Scientific Workflows, KDD, IV, and VA Following powerful methodological workflows presented in KDD, IV, and VA, a reference workflow for the design of ESS would be beneficial. It is mandatory that the new reference workflow includes best-practice methodology presented for scientific workflows, KDD, IV, and VA. Together, the existing methodology of different research fields is highly beneficial to compensate missing methodology in ESS design $\mathbf{C}_{MES}$. To avoid reinventing the wheel, a careful review of related workflows, frameworks, and models is required. Hence, the new reference workflow for the design and the application of ESS would benefit from scientific workflows, KDD, IV, and VA.

RG_{MES4} Splitting the Workflow into Canonical Steps The design of ESS forms a huge design space including a variety of challenges for data scientists. The access to the data content $\mathbf{C}_{CBA}$ has to be provided in a meaningful way, a property which depends on the information needs of the targeted user group $\mathbf{C}_{UCD}$. Search and exploration activity must to be supported, e.g., by providing content-based overviews in combination with visual querying interfaces $\mathbf{C}_{CBO}$. In addition, supporting relation seeking between data content and metadata $\mathbf{C}_{C+M}$ adds to the challenge of splitting the workflow. All these functionalities require a rich set of algorithmic models $\mathbf{C}_{MPC}$ which need to be put into a cascaded workflow. Many models have parameters which multiply the complexity of the design space $\mathbf{C}_{MPC}$. Our solutions require a massive use of VA techniques to support the construction of meaningful workflows combining the strength of both human judgment and algorithmic power. To cope with the complexity of such workflows, the new reference workflow should split the design space into meaningful steps. The steps should be canonical, meaning that the steps should be applicable for a variety of ESS approaches drawing on different data, users and user tasks.

RG_{MES5} Combining the Data Transformation and User-Centered Design Process The review of related work poses two principal types of process models for analytical systems. The first type is defined by the data and the involved algorithmic models, referred to as workflows (cf. Section 2.4). Workflows are greatly beneficial to conceptualize the principal data transformation steps. Requirements for this type of process model are well covered with $\mathbf{RG}_{MES3}$ and $\mathbf{RG}_{MES4}$. The second type of process model focuses on user-centered design and design study methodology (cf. Section 2.5), and thus is beneficial to resolve challenges associated with involved users $\mathbf{C}_{UCD}$. The design process is emphasized, including canonical steps, such as the requirement analysis, the iterative design, or the deployment of analytical systems. Today, these two types of process models are still applied regardless of each other. It would be beneficial if the reference model would combine both the data transformation and user-centered design process.

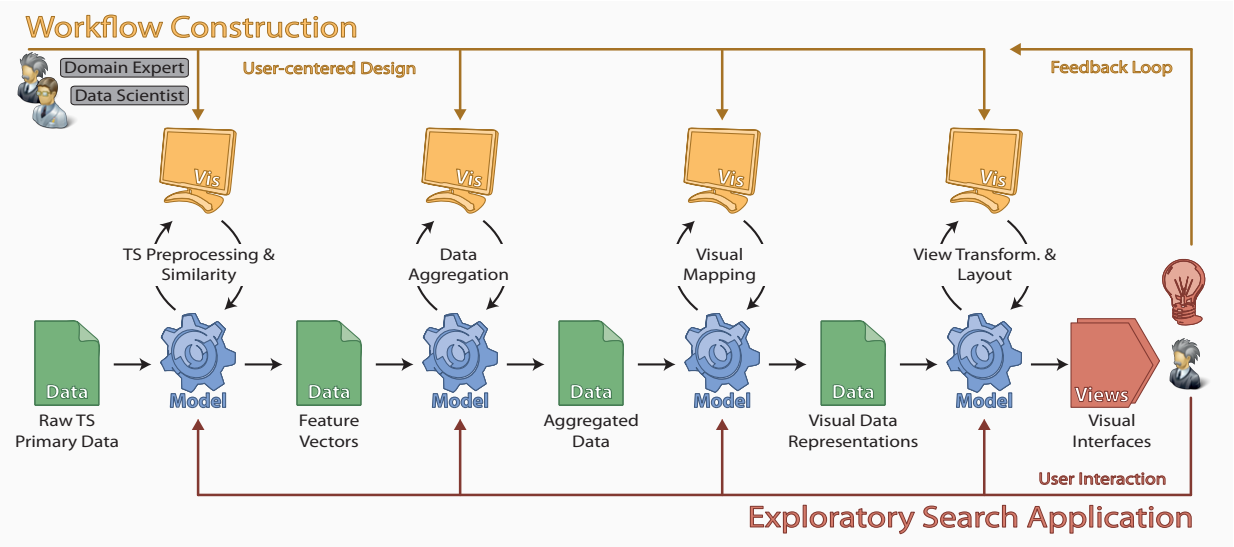


Figure 3.1 *Reference workflow for the design and the application of visual-interactive interfaces for ESS. From left to right the data-centered workflow is shown, divided into four main steps each containing an algorithmic model (blue). From above to below the user-centered design is shown: the reference workflow differentiates between the design phase (workflow construction) and the ESS application (workflow execution).*

3.1.3. Contribution

In this chapter, we propose two conceptual contributions both of which are relevant for the remaining work. Thereby, we confront the challenge of missing methodology for the design of ESS $\mathbf{C}_{\text{MES}}$. First, we present a *survey of IV and VA tasks* compiled as a single assembly. Based on the assembly, we associate IV and VA tasks with search and exploration activity. The search and exploration activities subsequently depict intersection points to associate ES with tasks and techniques presented in IV and VA. In this way, we accomplish the objective of bringing ES together with IV and VA $\mathbf{RG}_{\text{MES1}}$. As a result, designers of ESS are able to gain an overview of the rich set of tasks and techniques provided in IV and VA associated with search and exploration activity. On this basis, we define Exploratory Search for this thesis, with an emphasis on concepts and techniques presented in IV and VA $\mathbf{RG}_{\text{MES2}}$. Second, we present a *reference workflow* for the design and the application of ESS for time-oriented primary data (see Figure 3.1). The reference workflow describes the design of important visualizations (views) for ESS. In addition, the reference workflow depicts how the *design phase* of ESS, when workflows are constructed, can be intertwined with the *application phase*, when workflows are executed and tested. The results of the workflow construction phase are a visual-interactive interfaces which can be integrated into operable ESS. Throughout this thesis, we will design three different types of views for ESS based on the reference workflow:

- Content-based overviews
- Visual query interfaces and visualizations of search results
- Views for relation seeking between the data content and the associated metadata

The two contributions serve as baseline concepts for the guidelines, techniques, and systems presented in the following chapters. The survey of IV and VA tasks and techniques supports the process for the user-centered design of visual-interactive visualizations for ES. The reference workflow is implemented in the technical contributions of this thesis, i.e., preprocessing of time-oriented primary data, content-based overviews, and a combined exploration of data content and metadata (cf. Chapters 4, 5, and 6). Finally, in the case studies, we use the reference workflow for the design of usable and useful ESS systems (cf. Chapter 7).

3.1.4. Chapter Overview

The remainder of the chapter is as follows. First, we present the survey of search and exploration activity as known from IV and VA in Section 3.2. We review IV and VA tasks and techniques, assign the tasks to a single assembly,

and map search and exploration activity onto the assembly. On this basis, we present a definition of ES. Second, we present a reference workflow for the design and application of ESS for time-oriented primary data in Section 3.3. We describe both the data-centered workflow and the user-centered workflow along the two axes of the diagram. The workflow adopts important methodology from existing reference models which are extended and adapted towards the construction of workflows for time-oriented primary data. Third, in Section 3.4, we outline the three main technical contributions of this thesis. All of them reflect the reference workflow. In addition, we outline the two real-world case studies of this thesis. In both case studies, we construct workflows for the subsequent execution of visual-interactive ESS. Finally, Section 3.5 summarizes the concepts presented for this thesis.

3.2. Survey of Search and Exploration Activity

In Section 2.1, we presented the scope of ES. We first reviewed information seeking, a baseline theory for ES in Section 2.1.1. Second, we defined IV and VA and presented an overview of task taxonomies in Section 2.1.2. Finally, we reviewed different characterizations of ES in Section 2.1.3. Subsequently, we condense the variety of tasks and techniques presented in IV and VA which contribute to search and exploration activity. By taking various task taxonomies into account, we present a single assembly of tasks in Section 3.2.1. Moreover, we highlight tasks in the assembly relevant for a) search activity, b) exploration activity, or c) both activities. This leads to intersection points associating ES with the tasks and techniques presented in IV and VA $\mathbf{RG}_{\text{MES1}}$. The rich set of existing IV and VA solutions can subsequently be used as a toolkit for the design of ESS. In Sections 3.2.2 and 3.2.3, we further describe these existing tasks and techniques for search and exploration activity in detail. Finally, we present the definition of ES for this thesis in Section 3.2.4 in consideration of IV and VA tasks $\mathbf{RG}_{\text{MES2}}$.

3.2.1. Assembly of Information Visualization and Visual Analytics Tasks

The review of the different task taxonomies in Section 2.1.2 gives an idea of the heterogeneous scopes of task and task characterizations. Tasks are described on *different levels of complexity* and abstractions, or from *different perspectives*, such as users, interactions, and visualizations [YKSJ07]. In addition, distinctions are made between the high-level *intent* (goal) of a task and the low level *technique* for the execution of the task. As an example, Ben Shneiderman's task taxonomy reflects general user goals/tasks within the information-seeking process (e.g., overview, Details-on-Demand) [Shn96]. Similarly, some of the tasks (also) fit into taxonomies for low-level interaction techniques (zoom, filter) [YKSJ07]. Finally, Shneiderman's tasks can be classified by the intent of a visualization design to support the user in the information-seeking process (e.g., overview, relate). These heterogeneous scopes of tasks and task characterizations impede the reflection and the utilization of existing tasks. In the following, we survey different perspectives on tasks existing in IV and VA in a single assembly. With the assembly, we are able to identify search and exploration activity within the 'task-universe' of IV and VA. Figure 3.2 illustrates the assembly of IV tasks.

In the assembly, we apply two main distinguishing criteria adopted from Yi et al. [YKSJ07], Aigner et al. [AMST11], and Brehmer et al. [BM13]. First, we differentiate the levels of abstraction, i.e., high-level and low-level tasks. A high-level task corresponds to the superior *goal* or *intent*, while the lower task level corresponds to technical artifacts. Second, we distinguish three different perspectives, i.e., the *user*, the *interaction*, and the *visualization* perspective. User goals/tasks resemble the perspective of users, i.e., the different information needs of users. Interaction intents/techniques clarify how the information-seeking behavior of users can be communicated to the system. Visualization intents/techniques focus on visualization designs supporting users in the information-seeking process. In this connection, we borrow the common ground of Marchionini's Information-Seeking Process [Mar95] and Norman's Action Cycle [Nor02] describing interaction between users and sources of information (see Figure 2.1). The core 'spatialization' of the task assembly is a cycle which is inspired by the two process models.

On the left of the assembly in Figure 3.2, the *user* perspective is shown involving high-level user *goals* and low-level user *tasks* (embodied with a yellow shape). A high-level goal is something to be achieved, often vaguely stated and imprecisely specified [Nor02]. Depending on the information need, goals can typically be broken down to tasks, i.e., to a lower level of abstraction. Norman refers to this phase of the process as forming the intention and specifying the action(s). Similarly, the *interaction* perspective on tasks is structured in high-level *intents* and low-level *techniques* (orange shape). The interaction designs of visual-interactive systems yield sets of techniques addressing the interaction intents of users. The execution of an interaction causes a change in the state of the information source [Nor02] (here: the visual-interactive system). At this stage of the process, we refer to the variety of techniques available in the KDD process. These techniques can be included into analytical systems and be triggered by visual-interactive interfaces (see

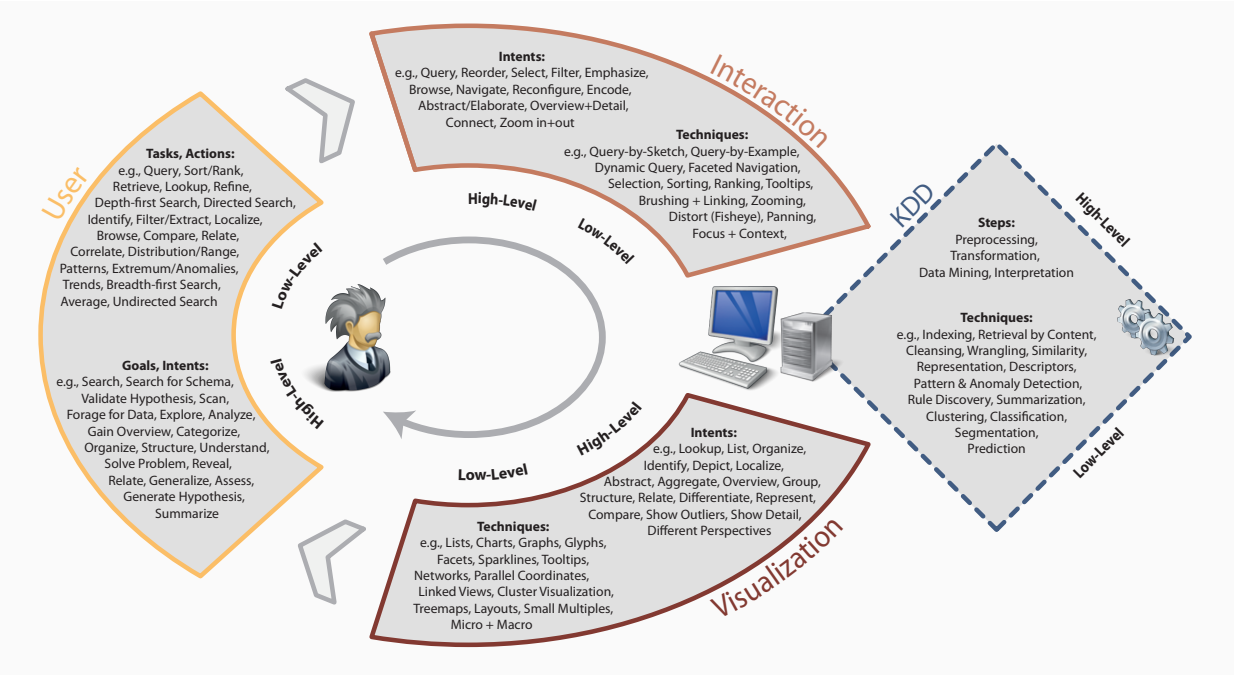


Figure 3.2 *Assembly of IV and VA tasks extracted from task taxonomies and put into a process (cf. [SBM92, Mar95, Shn96, CMS99, Kei02, AA06, YKSJ07, AMST11, BM13, KH13]). Following the categorizations of Yi et al. [YKSJ07], we distinguish between high-level and low-level tasks from three perspectives: user, interaction, and visualization (representation). On the right, we indicate an intersection point to steps and techniques from the KDD process [FPS96]. We review KDD techniques for time-oriented data in Section 2.3.2.*

Section 2.3.2 for a review of KDD techniques for time-oriented data). As a consequence of the execution, the system calculates a result which can be represented visually. A high-level goal of the *visualization* perspective is to amplify cognition [TC05] by visualization design (red shape). The user is able to perceive and interpret the visualization, to extract information, and to gain new insight [Nor02]. For this purpose, IV provides a variety of visualization techniques, see, e.g., the book of Aigner et al. [AMST11] for an overview of techniques for time-oriented data.

3.2.2. Survey of Search Activity in Information Visualization and Visual Analytics

We take a closer look at search activity and its association with tasks presented in IV and VA. To this end, we map search activity to the assembly of tasks. The result is presented in Figure 3.3. We survey search activity based on the three perspectives, presented in the assembly (user - interaction - visualization). In the survey, we combine technical factors from IV and VA with conceptual factors, e.g., characterized in ES. In this way, we expose intersecting points for search activity between the ES concepts and the techniques used in IV and VA $\mathbf{RG}_{\text{MES1}}$.

The User Perspective on Search Tasks We start the review from a user's perspective. The high-level goal of a searcher is to find relevant information for a predefined problem. Search results typically consist of a small set of potentially relevant documents containing potentially relevant information. In many cases, the searcher has a hypothesis which may be validated or disproved with the collected information. More generally, searchers feature a extent of known information which can be used to find unknown information. The distinction between known and unknown information for time-oriented data is described by Andrienko and Andrienko as follows [AA06].

- Searching for the time or the times when a given object occurred (what -> when)
- Searching for the object or set of objects at a given time (when -> what)

In low-level taxonomies of user tasks, finding the time and the times when a given object occurred is typically defined as a *localization* task [AA06, AMST11]. Finding the object or set of objects at a given time is described as an *identification* task. If both the objects and the set of times are known, searchers are able to execute a *lookup*

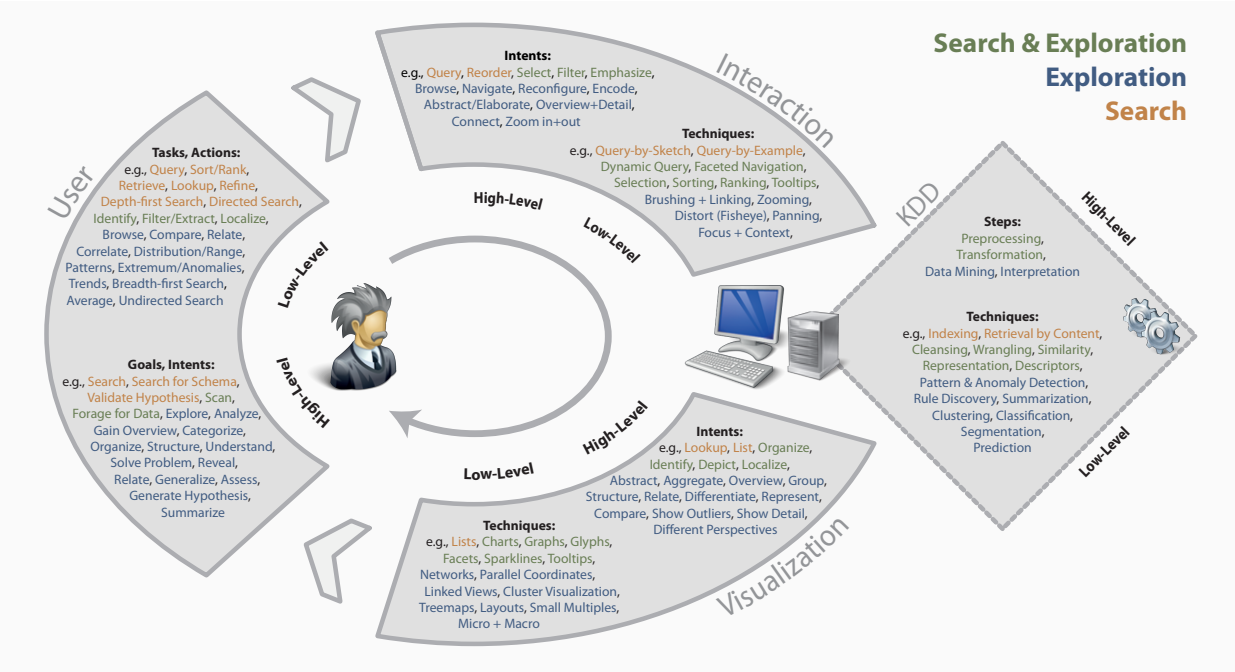


Figure 3.3 *Mapping of search and exploration activity to the assembly of IV and VA tasks (extracted from task taxonomies and put into a process). IV and VA tasks predominantly executed by searchers are colored orange, tasks associated with exploration are colored blue. Tasks related to both activities are colored green.*

task [Mar06, BM13]. Low-level tasks in which objects to be retrieved are known beforehand are also referred to as *fact retrieval* or *known-item search* [Mar06]. Concrete tasks and actions of searchers include querying, retrieving, sorting, and ranking document collections.

The Interaction Perspective on Search Tasks A most relevant interaction task for users with search intent is querying (see the interaction perspective in Figure 3.3). We characterize querying as the formulation of a precise specification of the information need, which results in a small subset of possibly relevant documents provided by the retrieval system. While in early search systems the formulation of queries was typically based on complex operators, IV has contributed to a simplification of query formulation with a variety of visual-interactive techniques. For instance, Venn diagrams are used for the visual specification of boolean queries (see, e.g., [Hea09, p. 47 ff.]). Visual query term suggestions and visualizations for auto-complete functionalities support users in formulating textual queries. Furthermore, dynamic querying is among the most prominent visual querying techniques for a variety of data types (see, e.g., [AS94, SS02]). Dynamic queries enable searchers to define and adapt search results by an interactive specification of borders (e.g., adaption of min-max for a continuous attribute). Hence, dynamic querying also supports the intents of users with exploratory information needs.

Another visual querying technique applicable for a variety of data types is the Query-by-Example concept. The user is provided with a set of potentially relevant query terms which can be used to start the retrieval. Advantages of querying by example are an enhancement of the search process by retrieving non-empty result sets. However, the set of provided example queries has to match the information need of the user. In this thesis, we present different techniques and example systems resolving this challenge. As an alternative, Query-by-Sketch techniques can be applied allowing users to draw queries based on their information need. Challenges of Query-by-Sketch techniques exist in the hand-crafted creation of potentially complex objects, as well as the interpretation and mapping of these sketches onto the feature space. In addition, the pure Query-by-Sketch concept does not involve an overview of the search space. Thus, the seeker may need several query-response trials with bad or empty results until the query can be adapted towards documents actually available in the document corpus. For time-oriented data a variety of visual querying techniques have been presented including dynamic queries, Query-by-Example and Query-by-Sketch. We present an in-depth review of visual querying techniques for time-oriented data in Section 2.3.3.

Apart from querying, the intents of searchers may require interaction techniques to select, focus, or highlight small subsets of documents. While the size of the displayed set of documents remains unchanged, users are able to emphasize documents corresponding to their information need. Tooltipping shows detailed information of single documents, or grouped document collections. Highlighting and selection techniques allow the searcher to emphasize relevant documents. These documents, in turn, may be the subject to downstream Query-by-Example techniques. Sorting interaction allows the re-arrangement of the data, and thus for focusing on relevant data based on a ranking criterion.

Finally, faceted search has become an important interaction technique for enhanced search practice. Faceted search and navigation supports "flexible movement within category hierarchies, seamless integration of browsing with keyword search, fluid alternation between refining and expanding, and avoidance of empty result sets" [Hea09]. Similar to dynamic queries, faceted search (faceted navigation) is useful in search scenarios with exploratory information-seeking behavior. A shortcoming of faceted search interfaces is its dependency on predefined structures. In most cases, facets have to be hand-crafted to meet certain quality levels. Our contributions in Chapter 6 for seeking relations between data content and metadata are motivated by the need for enhanced faceted search and navigation concepts. The VisInfo case study in Section 7.1 uses a set of appropriate metadata as faceted search.

The Visualization Perspective on Search Tasks We conclude with visualization intents and visualization techniques supporting search activities in IV (see the visualization perspective in Figure 3.3). From a high-level perspective, supporting search activities by visual representations can be described by the intents to provide organized and structured result sets. These result sets support the identification and localization of single elements.

Given the granularity of entire result sets, the salient visual representation techniques are list-based interfaces. A number of retrieved elements is represented in a ranked order. The ranking criteria differ by measures like precision, accuracy, relevance, popularity, or profitability (from a search provider's perspective). Many list-based search results allow switching the ranking criterion. One drawback of list-based search results is the comparably small number of shown elements. Especially for large document collections it is not guaranteed that the searcher identifies all relevant documents, i.e., precision may be preferred over recall. A further shortcoming of list-based search results is the mendable use of the of visual variable *position* (x, and y). Usually, the y-axis represents the ranking, while the x-axis does not carry structural information about the data set. In addition, adjacent list elements do not necessarily build a natural order (also called a topology), since the order of the ranking is based on the given query. As a result, result lists at a glance do not support gaining structural information of the collection and single list elements typically need to be examined and interpreted independently.

Given the granularity of single documents within a result set, the classical visual representation is a textual description. In many cases, the description includes various relevant metadata attributes. Moreover, these descriptions may contain links to related documents or document classes (also referred to as 'hypertext systems' [MS88]), as well as other domain-specific information. Today, functionality to support Details-on-Demand tasks is common practice to highlight query term occurrences within the content. Another extension of result element visualizations regards information about the relationship between the query and the retrieved document. Examples are small iconic graphics representing the relative length of documents and the relative positions of the query terms within the documents, referred to as TileBars, or heatmap-based techniques (see [Hea09, p. 254] for an overview). Still other types of graphical extensions are based on powerful glyph-based techniques, such as the Document Cards technique by Strobelt et al. [SOR*09] representing condensed information about scientific publications by small card metaphors. This class of techniques can be described as thumbnail images of documents. The advantages of these miniaturized images are that (a) users are able to gain a quick preview of the document content, (b) thumbnails enable recognition (re-identification - a user has previously seen the thumbnails in the past), and (c) similarity-preserving visual variables may facilitate the comparison of documents [Hea09]. Thumbnails and glyph designs in general are an important technique for mapping the data content to the visual space. Our contribution referring to content-based overviews (cf. Chapter 5) involves the careful design of glyphs for high-dimensional data objects supporting all three outlined advantages. The enhancement of search results with additional visual representations facilitates comparing documents, as well as seeking relations between documents and attached metadata. These types of search-result visualizations are especially helpful for information seekers with an exploratory information need with the aim to gain insight in structural properties of retrieved document collections. This is why graphical extensions in search results are useful for both searchers and explorers.

3.2.3. A Survey of Exploration Activity in Information Visualization and Visual Analytics

We survey exploration activity and its association with tasks presented in IV and VA. The result is presented in Figure 3.3. The illustration demonstrates how exploration activity is associated with task and techniques presented in

IV and VA $\mathbf{RG}_{\text{MES1}}$. It is remarkable that exploration and search activities have a small overlap. In the following, we survey exploration activity based on the three perspectives (user - interaction - visualization).

The User Perspective on Data Exploration Tasks At the beginning of the exploratory information-seeking process, the user has hardly any hypothesis about the data. It is important to support explorers in gaining insight in the structural information of the data collection [KMSZ06, KMS*08]. Information seekers with an exploratory information need may be motivated by various of general user goals. Exploration is related to the visual *analysis*, *summarization*, *categorization*, and *organization* of large data collections. Providing *overviews* of the data is one of the most relevant design goals of IV systems [Shn96]. In addition, explorers envisage to *reveal relations*, to *understand data characteristics*, and to seek *new hypotheses* about the underlying data. Low-level user tasks involve, e.g., *sorting*, *filtering*, *clustering*, *comparing*, *relating*, data elements or sets of data elements, to identify *extrema*, *anomalies*, *trends*, frequent *patterns*, or *associations* within the data.

The Interaction Perspective on Data Exploration Tasks To support the explorers' goals and tasks, a variety of interaction intends can be applied (see the interaction perspective in Figure 3.3). An indispensable interaction intent of explorers is *browsing*. Browsing is referred to as navigating through a search space [Bat89]. In contrast to direct search, browsing interaction refers to a mainly *undirected* navigation through of structures to explore possibly unknown targets [Hea09]. Browsing "is the process of developing an understanding of unexpected patterns within the collection" [Shn96]. Thus, it supports rather informal and opportunistic information-seeking strategies [Mar95]. Browsing activities can include following links, switching views, or scanning and selecting sets of operations [Hea09]. Browsing can be supplemented with other interaction techniques, such as *panning*, *zooming*, *sorting*, *highlighting*, *focus+context*, *history* or *extract* functionality [Shn96]. We emphasize *panning* as a most relevant activity where the explorer navigates laterally across a view of data [YKSJ07, Hea09] and *zooming* allowing explorers to navigate deeper into local aspects of interest [Shn96, EF10]. Some of these interaction techniques are also relevant for another, most relevant, interaction intent of explorers: the intent for *overview+detail* and *abstract+elaborate* interaction. These types of exploratory activities are well characterized with the Visual Information-Seeking Mantra by Ben Shneiderman [Shn96]. The explorer interacts with the system to gain an overview of the data collection at a coarse level with high data abstraction. A related interaction technique is *brushing*, possibly in combination with *linking* information in different views. Depending on the gained structures (relations, patterns), the explorer requires interaction techniques, such as zooming, panning, and filtering to drill down to local, more specific information. *Filtering* interaction supports the exploratory information-seeking behavior as a means of excluding needless information from the search space, and thus obtaining more display capacity for relevant information. Filtering is also a intersection point to enhanced search interaction techniques, i.e., *faceted search* and *dynamic querying*, described earlier in the section for search activity. Hochheiser and Shneiderman's Timebox widgets shall be named as a pioneer work to support interactive exploration of time-oriented data with dynamic querying. The Visual Information-Seeking Mantra concludes the exploration process with an elaboration of details, on demand. Related interaction techniques are *zooming*, *selection*, *highlighting*, *tooltipping*, or *focus+context*.

The Visualization Perspective on Data Exploration Tasks We conclude the review with visual representation techniques to support exploration intents (see the visualization perspective in Figure 3.3). Among the most relevant visualization intents is supporting explorers in gaining an *overview* of the entire data collection [Shn96, KH13]. Especially if the data collection is large and complex, this is a challenging task for visualization designers. One technical contribution of this thesis regards the design of content-based overviews (see Chapter 5). Important techniques for providing overviews are visualizations of data *aggregates*, such as *cluster visualizations* (see, e.g., [VWVS99, EF10]), designs for the visual representation of *high-dimensional data* objects and clusters, such as *glyphs* (see, e.g., [BKC*13]), and *layouts* for the representation of *structural* information of the data (e.g., [POM07, IMI*10]). We refer to Section 5.2 for an in-depth review of baseline visualization techniques related to content-based overviews. Yet another class of techniques useful for explorers is based on the *multiple linked views* concept. Linked views show sets of data elements from *different perspectives* in different views. A linking technique has to be provided, enabling looking up identical data objects in different views and seeking for relations between different data objects. An overview of linked views is presented, e.g., by Kehrer and Hauser [KH13], example techniques are demonstrated by Stasko et al. [SGL08], Kehrer et al. [KLM*08], or by Dork et al. [DCCW08]. Associated with linking views is the *small multiples* concept introduced by Edward Tufte [Tuf90]. A set of similar graphs or charts with identical axes and scales is represented in juxtaposition for an enhanced *comparison* of sets of objects [GAW*11]. Another visualization technique presented by Edward Tufte [Tuf90] is the *Micro-Macro* visualization where data is represented on two superimposed levels of

granularity. We use the technique for the design of content-based overviews in Chapter 5. A recent example of the visual exploration of small multiples and large singles is presented by van den Elzen and van Wijk [vdEvW13]. A variety of visualization techniques, such as *graphs*, *networks*, *treemaps*, or *parallel coordinates* can be used to support exploratory information-seeking. We refer to the work of Keim [Kei02] for an overview of pioneer works. Specific visualization techniques for the exploration of time-oriented data are reviewed in Section 2.3, a survey is provided by Aigner et al. [AMST11].

3.2.4. Exploratory Search

In Table 3.1, we summarize the review of different characterizations of search and exploration activity. It can be seen that search and exploration are to a high degree opposed to each other. Taken in isolation, both search and exploration are highly important types of information-seeking activities. However, it is recognized that many real-world problems, decisions, sources of information, data types, and information needs in general require both search and exploration activity [MS88, Mar95, Shn96, HS04, Hea06, Mar06, DCCW08, WR09, SBS11, HCQ*12, vLSFK12, HK12]. Information seekers will then be better supported in the process of gaining meaningful insight from time-oriented primary data. Based on the different characterizations of search and exploration, we define ES for this thesis as follows $\mathbf{RG}_{\text{MES2}}$.

Defining Exploratory Search *Exploratory Search (ES)* is the seamless combination of different information-seeking behaviors ranging from search to exploration activity. ES supports answering questions spanning from simple lookups and known item searches to complex learning and investigation tasks. ES includes seeking information in the complete body of the document collection, i.e., the data content, metadata, and relations in between. Visual interfaces facilitating ES require visualization and interaction designs make it possible to gain an overview of the document collection, to identify and search for relevant subsets, and to explore extracted information in detail.

Characteristics of Exploratory Search In Figure 3.3, we mapped search and exploration activity to the assembly of IV and VA tasks. Following this, we outline different characteristics of search and exploration activity, see Table 3.1 for an overview. In this way, we show how ES can combine the complementary theories and concepts for search and exploration activity. Table 3.1 can serve as a baseline for the definition of requirements for systems enabling users carrying out ES tasks.

The information need of searchers is referred to as known-item search and fact retrieval [Mar06, p. 29 ff.] [WR09, p. 14]. In general, the curiosity of searchers is rather specific and towards well-defined goals. The interaction intents of searchers typically require the definition of queries leading to directed searches. On the contrary, the curiosity of explorers is diverse and the goal is often ill-defined in the first place. The interaction intents of explorers include exploratory browsing leading to undirected searches. One of the goals of effective ESS is "moving seamlessly between browsing and searching, minimizing context switching, and keeping users focused" [FGS12].

The entry point characterizes the visual representation at start of the information-seeking process. Searchers typically directly begin with the formulation of queries. The predominant retrieval paradigm of searchers is described as "query and response" [WR09], based on the alternating participation of users formulating queries and waiting for response. The entry point of explorers is well-described with the Visual Information-Seeking Mantra by Ben Shneiderman [Shn96]. In the majority of cases, exploratory data analysis systems first present an overview. Depending on the interaction intents of the user, concepts such as overview+detail and abstract/elaborate need to be provided by the systems. Thus, the human partition in exploratory data analysis systems shows a continuous active engagement.

The information-seeking behavior of searchers within the process is considered static, the preferred activity corresponds with lookup tasks and known item searches. On the contrary, activity of exploratory searchers also involves learning and investigation [Mar06]. This is why the information-seeking behavior of explorers may be subject to change within the process. New patterns, structures, or trends may be revealed within the data, possibly facilitated by various comparison [GAW*11] and relation-seeking tasks [KH13]. Browsing through visual data representations supports gaining new insights into the underlying data collection. ES adopts the principal of changing information-seeking behavior [Mar06, WR09]. While lookup and localization activity may be included within the process, browsing and other exploratory interaction intents need to be provided in ESS to facilitate learning and investigation.

Hypotheses testing is among the activities of searchers while explorers focus on the formulation of new hypotheses. ES supports both the validation of hypotheses and the formulation of new hypotheses. In his way, ES also poses a intersection point for confirmatory analysis and exploratory analysis strategies. Searchers prefer confirmatory analysis, i.e., the goal-oriented examination of hypotheses including confirmatory result visualizations. Explorers focus on

Category Magnitudes for the characterization	**Search** Terms, concepts and techniques for searchers	**Exploration** Terms, concepts and techniques for explorers
Search Theory: Direction direction of the search activity, type of resource allocation	Directed, formal & analytical [Mar95]	Undirected [KMS*08], informal & opportunistic [Mar95]
Information Need the intent/curiosity/goal of seekers to obtain information	Known-item search, fact retrieval, [Mar06], specific [WR09]	Diverse, ill-defined [WR09]
Analysis Strategy distinction between confirmatory and exploratory analysis	Confirmatory Analysis [Shn02]	Exploratory Analysis [Kei02] [Shn02]
Complexity of Seeking information-seeking activities that need to be supported	Lookup [Mar06]	Learn, investigate [Mar06]
Dynamics of Seeking whether the information-seeking behavior changes over time	Static [Hea09]	Dynamic [Hea09]
Scientific Method type of research / investigation	Hypotheses testing/valid. [Kei02] [Shn02]	Hypotheses generation [Kei02, KH13]
Human Participation in the information-seeking process	Alternating (Query+Response) [Mar06, WR09]	Continuous process, continuous active engagement [Mar06, WR09]
Entry Point the visual representation at start	Query	Overview [Shn96]
Query Formulation predominant query type	Typing, creation, Query-by-Sketch	Query-by-Example
Drill-Down Strategy the way how sets of potentially relevant data are achieved	Create small subset	Reduce, filter, zoom, select, Details-on-Demand
Information Processing knowledge ordering strategy	Bottom-up	Top-down
Traversal Strategy predominant navigation type through the search space	Depth-first search	Breadth-first search
User Tasks predominant low-level tasks intended by the information seeker	Lookup [Mar06, BM13], identify & localize [AA06, AMST11]	Compare & relation seeking [AA06, AMST11], explore [BM13], summarize [BM13]
Interaction Tasks predominant interaction tasks intended by the information seeker	Querying, Iterative search, Focused searching [Mar06, WR09]	Browsing [MS88, Mar06, WR09], overview+detail & abstract/elaborate
Visualization Tasks predominant visualization intents/tasks to be supported	Visual querying, result representation	Overview, aggregation, linked views, small multiples, Details-on-Demand

Table 3.1 *Different characteristics of search and exploration activity. Summary of the related work on ES (cf. Section 2.1) and the survey of search and exploration activity in IV and VA.*

exploratory analysis, i.e., an undirected search in data with the goal to reveal patterns, structures, or trends without prior knowledge about the data. The insight gained within the exploratory data analysis contributes to the hypotheses formulation process. In combining confirmatory and exploratory analysis, ESS may contribute to an efficient and effective scientific process.

Another distinguishing characteristic is the information drill-down and search space traversal strategy of searchers and explorers. With the formulation and execution of queries, searchers obtain small data subsets of high precision. To this end, we refer to searching as depth-first search strategies for traversing data collections. On the contrary, exploratory browsing enables explorers to carve out relevant information with a variety of interaction techniques yielding subsets with high recall. Accordingly, we assign exploratory browsing a breadth-first search strategy. For the search space traversal strategy, ES may constitute an intersection point. ESS may provide various interaction techniques leading to seamless combinations of depth-first search and breadth-first search strategies.

3.3. A Reference Workflow for Exploratory Search Systems

As in the following, we present a reference workflow as a framework for this thesis. The reference workflow is shown in Figure 3.1. It describes the design and the application of visual-interactive interfaces relevant for ESS. Supported visual-interactive interfaces are, e.g., content-based overviews, visual-interactive query interfaces, as well as views for seeking relations between data content and metadata. The reference workflow explicitly factors time-oriented primary data, however, it may also be adapted to other data types.

The reference workflow divides the pipeline (x-axis) into four *steps* which transform and streamline the different data representations. Hence, the four steps divide the challenges of workflow construction into smaller and more specific parts $\mathbf{RG}_{\text{MES4}}$. The four steps are the result of an adoption of most prominent methodology from scientific workflows, KDD, IV, and VA $\mathbf{RG}_{\text{MES3}}$. For each step algorithmic *models* are required to facilitate data manipulation and transformation. Our approach postulates the massive use of VA support for the design of each of the four steps. Thus, data scientists yield powerful computational models based on human judgment. In addition, we are able to approach challenges caused by cascade effects, i.e., implications of routines building up on each other. With the visual-interactive support, our techniques allow the enhanced creation and validation of workflow steps.

The design of important visual-interactive interfaces of ESS is supported from both the data-centered and the user-centered perspective. Users can be involved in the design, to resolve usability and usefulness challenges within the design phase $\mathbf{RG}_{\text{MES5}}$. According to user-centered design and design study methodology, we distinguish between the design and the application phase (y-axis). Consequently, the reference workflow can guide data scientists towards the design of visual-interactive interfaces in a user-centered way. The design phase (above the workflow) comprises the domain and data characterization, as well as the workflow construction. The application phase (below the workflow) reflects the workflow execution, evaluation, and ESS application. Moreover, the application of ESS triggers a feedback loop to further improve the workflow, e.g., as a result of evaluation strategies. The reference workflow supports collaborative approaches by involving domain experts within the design. Combining the expertise of domain experts and data scientists fosters the choice of meaningful models and parameter values as a prerequisite for usable and useful ESS. As a fundamental difference between the two phases, the active role of data scientists only takes place in the design phase. One of our goals is to maximize the degree of automation of the workflow. At the same time, we support data scientists in the identification of parameters (and models) which should be subject to user adaption in the application phase.

3.3.1. Data-Centered Workflow

The review of related works showed that the direct utilization of the raw data content in analytical systems is hardly feasible. Instead, in the data-centered workflow (x-axis) the time-oriented primary data is transformed into different representations (formats) until it can be displayed in visual-interactive interfaces in a meaningful way. In this connection, the reference workflow adopts and extends existing methodology of reference models and frameworks presented in the related work. The KDD process by Fayyad et al. [FPS96], the Card pipeline [CMS99], as well as the time series VA frameworks by Aigner et al. [ABM*07] and Andrienko and Andrienko [AA13] serve as driving forces for the definition of data transformations and algorithmic models $\mathbf{RG}_{\text{MES3}}$. As a result, we propose a reference workflow with four mandatory steps/models within the workflow (see the blue gear wheel icons in Figure 3.1). In each model (blue), the underlying data is transformed into another data representation (green). The models divide the huge design space into smaller subspaces, each focusing on specific challenges, depending on *data*, *task*s, and involved

users. In this way, we reach the goal of coping with the complexity of the design space by splitting the workflow into a meaningful number of canonical steps $\mathbf{RG}_{\text{MES4}}$.

In addition, multiple complementing views may be designed for a single ESS and be linked together, e.g., by color coding or interaction designs. In this case, the reference workflow may be used several times for the different views. We postulate the incorporation of (a) VA methods and (b) user-centered design principles for the design of each of the four models. In this way, the workflow construction becomes visually comprehensible and interactively steerable. Moreover, the incorporation of visual-interactive interfaces in the design process supports the active user involvement. To support this, the reference workflow inscribes *four instances of the VA Diamond* [KKEM10] along the x-axis, one for each data transformation step. The four major steps of the reference workflow are as follows.

Time Series Preprocessing and Similarity Definition The first model accepts the raw time-oriented primary data as input $\mathbf{C}_{\text{CBA}}$. Challenges like establishing data quality, handling outliers, and coping with the temporal domain need to be met with meaningful *preprocessing* routines. Guidance concepts may be included to provide quality information, to estimate the degree of generalizability of the applied routines for large data sets, or to show the impact of applied routines for atypical time-oriented data. In addition, the first step of the workflow postulates the user-centered *definition of time series similarity* as a key functionality for search and exploration activity. The similarity concept is coupled with the definition of a time series descriptor to transform the time series into the feature space. The FVs and similarity functions can subsequently be used by downstream models, such as retrieval algorithms and data aggregation techniques. Thus, the FV-based approach is a means of providing search and exploration support.

Data Aggregation The second model in the workflow factors the *aggregation of data*. The goal is the summarization of data, i.e., to reveal groups of data which can subsequently be substituted by single objects or representatives. Meaningful aggregates enable data scientists to cope with large data collections, and thus support the exploratory information-seeking process. In particular, the aggregation of data is a key functionality for the design of content-based overviews $\mathbf{C}_{\text{CBO}}$ and interfaces for seeking relations $\mathbf{C}_{\text{C+M}}$. Many aggregation approaches are based on (visual) cluster analysis which facilitates the unsupervised assignment of the data objects to groups. However, the data aggregation step is not restricted to clustering models. Other implementations may involve statistical models, binning approaches, or other models expressing large data sets in a condensed form. Guidance concepts, quality assessment strategies, and other techniques drawing on VA may be used to reveal powerful data aggregates. Data structures containing the aggregated data information (i.e., sets, clusters) build the output of the data aggregation step.

Visual Mapping and Glyph Design For the next step, the aggregated data is transferred into the visual space in a *visual mapping* step, e.g., postulated in the Card pipeline [CMS99]. The visual mapping of data aggregates enables data scientists to display patterns in a visual form. Involving users in the design process leads to meaningful visual mappings. In this way, users are able to *localize* known items and to *identify* unknown patterns within the search space. Important visual mappings of data aggregates are e.g, *glyph design* techniques. In this thesis, visual data representations are required for the design of content-based overview visualizations $\mathbf{C}_{\text{CBO}}$, visual Query-by-Example interfaces, and visualizations supporting relation seeking between data content and metadata $\mathbf{C}_{\text{C+M}}$. Thus, the visual mapping step can be used for both search and exploration interfaces.

View Transformation and Layout The *view transformation and layout* step comprises the (a) arrangement of visual structures to views (visual interfaces), (b) enhancement of the visual interfaces with interaction designs, and (c) integration of the visual interfaces into ESS (cf. multiple coordinated and linked views). Layout techniques, such as data projections, facilitate the arrangement of visual data representations to one big picture reflecting structural information about the underlying data space $\mathbf{C}_{\text{CBO}}$ $\mathbf{C}_{\text{C+M}}$. The success of this visualization and interaction design step highly depends on properly gathered user requirements, on iterative design implementations, and on conducting appropriate evaluation procedures. As a result, the targeted ESS can be provided with views, such as content-based overviews, or visual-interactive interfaces facilitating relation seeking between the data content and associated metadata. The output of the view transformation and layout step is a single view, which can be integrated in the ESS.

3.3.2. User-Centered Workflow

The reference workflow exceeds existing workflow models by emphasizing the active involvement of the user within the design phase. The y-axis illustrates the two core phases of the user-centered design workflow. We divide the

process into a design phase (yellow) and an application phase (red). The design phase should include the domain, data and task characterization, followed by the definition of requirements, and an iterative design process. In the research field of scientific workflows the design phase is referred to as the workflow *construction* (cf. Section 2.4). Hence, the reference model explicitly reflects the phase, when data scientists and domain experts collaborate within the course of the workflow construction, to play an active role in the design process. The application phase involves the use of the workflows assembled to workable ESS. Furthermore, the application phase may be subject to testing and summative evaluation strategies. This is why we include a feedback loop at the upper right of the reference workflow. In the research field of scientific workflows the application phase is referred to as the workflow *execution*. Domain experts will be the stakeholders performing the workflow execution (ESS application), the active role of data scientists ends with the (successful) completion of the design phase. According to the design study methodology reviewed in Section 2.5.2, our reference workflow suggests user involvement at all mandatory steps of the design, i.e., the four core models of the data-centered workflow. We postulate the use of the VA process by Keim et al. [KAF*08] for each of the four models. Visual-interactive interfaces for the selection of appropriate models and model parameters facilitate the user-centered design process. In the following, we describe the user-centered phases in detail.

Design Phase In the design phase, data scientists and domain experts construct workflows which can be automatically executed after the ESS is deployed. An overall goal of the design phase is the identification of models and model parameters for the construction of meaningful workflows $\mathbf{C}_{\text{MPC}}$ in a user-centered way $\mathbf{C}_{\text{UCD}}$.

According to user-centered design and design study methodology surveyed in Section 2.5, we recommend a *domain and data characterization phase* in the beginning of a project. In this phase, domain experts and data scientists exchange vocabulary and communicate goals, intents, and visions. Data scientists learn about data characteristics, about the targeted user group, and their analysis tasks. In this phase, the assembly of IV and VA tasks presented in Figure 3.3 may be included to support the design process. Finally, data scientists accept requirements building the basis for the ESS design and for later evaluation strategies. Obviously, the domain and data characterization phase is dominated by verbal and conceptual thoughts. Nevertheless, the reference workflow should already be applied in this early stage of the design process. We postulate visual-interactive analysis tools as a most beneficial means for showing early results and rapid prototypes. For each of the four core steps, the reference workflow poses visual-interactive access (yellow icons) allowing data scientists and domain experts to collaborate and communicate relevant aspects in a transparent and comprehensible way.

In the *design phase*, the workflow is constructed in an iterative way. Data scientists and domain experts play an active role in the decision-making process which models and model parameters are most appropriate for the downstream ESS $\mathbf{C}_{\text{MPC}}$ $\mathbf{C}_{\text{UCD}}$. Visual-interactive tools can be used for the definition, the validation, and the optimization of models and parameters. VA-support techniques, such as guidance concepts, foster the selection of most appropriate models and parameters. An example case for the involvement of VA is the visual comparison of model outputs, contributing to an enhanced parameter steering process. In addition, collaboration is beneficial for the identification of models and model parameters which users would also like to steer in the application phase. As an example, domain experts may also want to steer the aggregation level of a content-based overview when using the final ESS. Hence, the design process facilitates the trade-off between the set of mandatory and powerful interaction designs and the ambition to keep (search) interfaces simple [Hea09]. Especially ESS which will also be used by non-experts can benefit from simple but powerful interfaces. DL systems may serve as examples, as our first case study shows (cf. Section 7.1).

Application Phase In the application phase domain experts are able to use the final ESS, and thus to execute the constructed workflows. Similarly, the application phase is a means for testing the workflow and for summative evaluation strategies. User interaction may trigger actions in the four main steps of the reference workflow, just as can be seen in the Card pipeline [CMS99]. We explicitly postulate that user interaction may trigger VA support, such as the exchange of models or the adaption of model parameters in the application phase. These changes are the result of the design phase, when data scientists identify which models and model parameters should also be modifiable in the application phase. Based on the underlying workflow, users are empowered to explore large search spaces in a visual-interactive way, e.g., by techniques for discovering interesting relations between data content and metadata. Depending on the degree of VA support in the application phase, users are able, e.g., to trigger the data aggregation step to adapt the content-based overview (see the MotionExplorer case study in Section 7.2). Search tasks are supported by visual query formulation whereon the search engine is executed (see both case studies in Chapter 7). Provided that most of the required models and model parameters are selected and calibrated within the user-centered design phase, domain experts are able to use powerful but yet simple user interfaces.

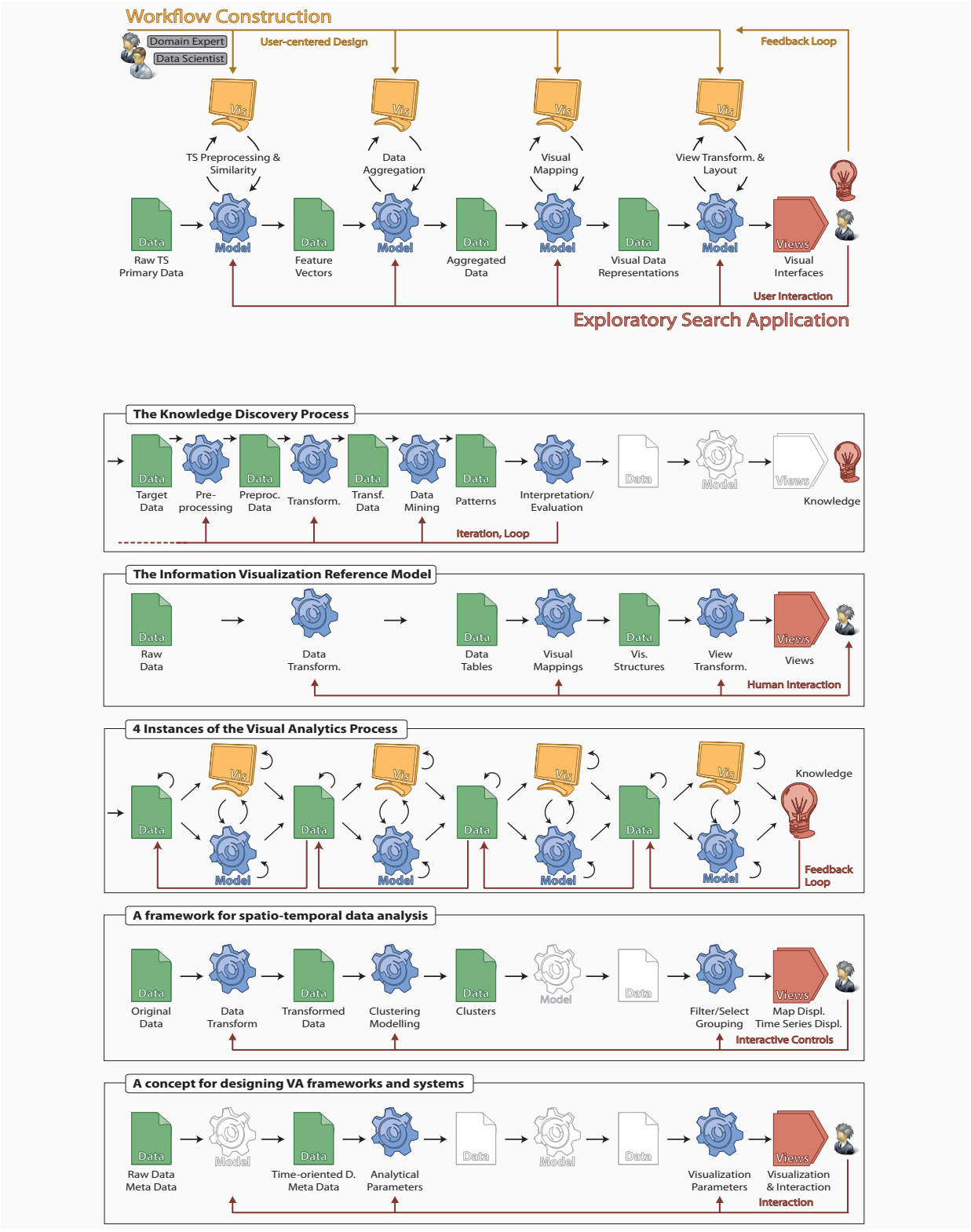


Figure 3.4 *The reference workflow presented in this thesis, in association with the adopted reference models and frameworks. Most relevant works are the KDD process [FPS96], the Information Visualization Reference Model (Card Pipeline) [CMS99], the Visual Analytics Process (VA Diamond) [KAF*08], the framework for spatio-temporal data analysis [AA13], and the concept for designing VA frameworks and systems [ABM*07].*

3.3.3. Relationship with Existing Reference Models

One of the research goals to be addressed with the reference workflow is the reuse of methodology presented in related works $\mathbf{RG}_{\text{MES3}}$. In this section, we draw a connection between the reference workflow on the one hand and existing reference models and frameworks on the other hand. In Figure 3.4, we associate the reference workflow with five most inspiring models reflecting the data-centered perspective. These models are:

- the Knowledge Discovery (KDD) Process [FPS96]
- the Information Visualization Reference Model (Card Pipeline) [CMS99]
- the Visual Analytics Process (VA Diamond) [KAF*08]
- the framework for spatio-temporal data analysis [AA13]
- the concept for designing VA frameworks and systems [ABM*07]

The most inspiring methodologies from a user-centered design perspective are the nested model by Tamara Munzner [Mun09] and the reflections from the trenches and the stacks by Sedlmair et al. [SMM12]. A review of design study methodology including the different phases in the design process is provided in Section 2.5.

3.4. Outlook for the Contributions of this Thesis

3.4.1. Three Main Technical Contributions

In the next three chapters, we present three technical contributions.

- Visual-Interactive Preprocessing of Time-Oriented Primary Data $\mathbf{C}_{\text{CBA}}$
- Content-Based Overviews $\mathbf{C}_{\text{CBO}}$
- Relation Seeking Between Data Content and Metadata $\mathbf{C}_{\text{C+M}}$

Each of the contributions supports the design of visual-interactive interfaces for ESS. The rationale of each of the contributions is to disclose relevant analytical models and to facilitate the definition of model parameters $\mathbf{C}_{\text{MPC}}$ in a user-centered way $\mathbf{C}_{\text{UCD}}$. The technical contributions build up on the four steps (models) presented in the reference workflow, an overview of the contributions is depicted in Figure 3.5, in association with the reference workflow. Each technical contribution provides guidance concepts, quality assessment strategies, and other principles known from VA. In addition, the technical contributions support the visual-interactive access to the models and the underlying data. In this way, the design process becomes more comprehensible and transparent, which also facilitates the involvement of users in the design process. Constructed workflows and visual-interactive interfaces can be integrated into downstream ESS, and thus support data-driven research in a usable and useful way. The user-centered automation of the workflows also has the advantage that important steering parameters for the workflows can be identified in the design phase. Users will then be able to interactively steer relevant models and parameters when using the ESS. Simultaneously, many models and parameters can be fixed in the design phase, leading to a simplification of the user interfaces.

Visual-Interactive Preprocessing of Time-Oriented Primary Data The first technical contribution is explicitly dedicated to the first step of the reference workflow $\mathbf{C}_{\text{CBA}}$ (see the first yellow workflow step in Figure 3.5). We present a system for the visual-interactive definition of time series preprocessing workflows (see Chapter 4). In this way, we elaborate the first of the four blue design spaces presented in the reference model in detail. A variety of algorithmic models exist, which are relevant for meaningful preprocessing workflows, including techniques for data cleansing, time series segmentation, and time series normalization. In addition, the visual-interactive system enables users to define time series descriptors transforming time-oriented primary data into the feature space. Finally, users are empowered to define distance measures which, in combination with the preprocessing, characterize the notion of time series similarity. FVs define the output of the preprocessing workflow. In consequence, downstream search engines and data aggregation concepts can be applied in an efficient and effective way.

Content-Based Overviews The second technical contribution considers the creation of content-based overviews for time-oriented data presented in Chapter 5 $\mathbf{C}_{\text{CBO}}$. As a result, the content-based overview visualization can be added into the ESS. On the basis of FVs as data input, the design of content-based overviews spans the remaining three steps of the reference workflow including the subtasks data aggregation, visual mapping, and data layout (see the second

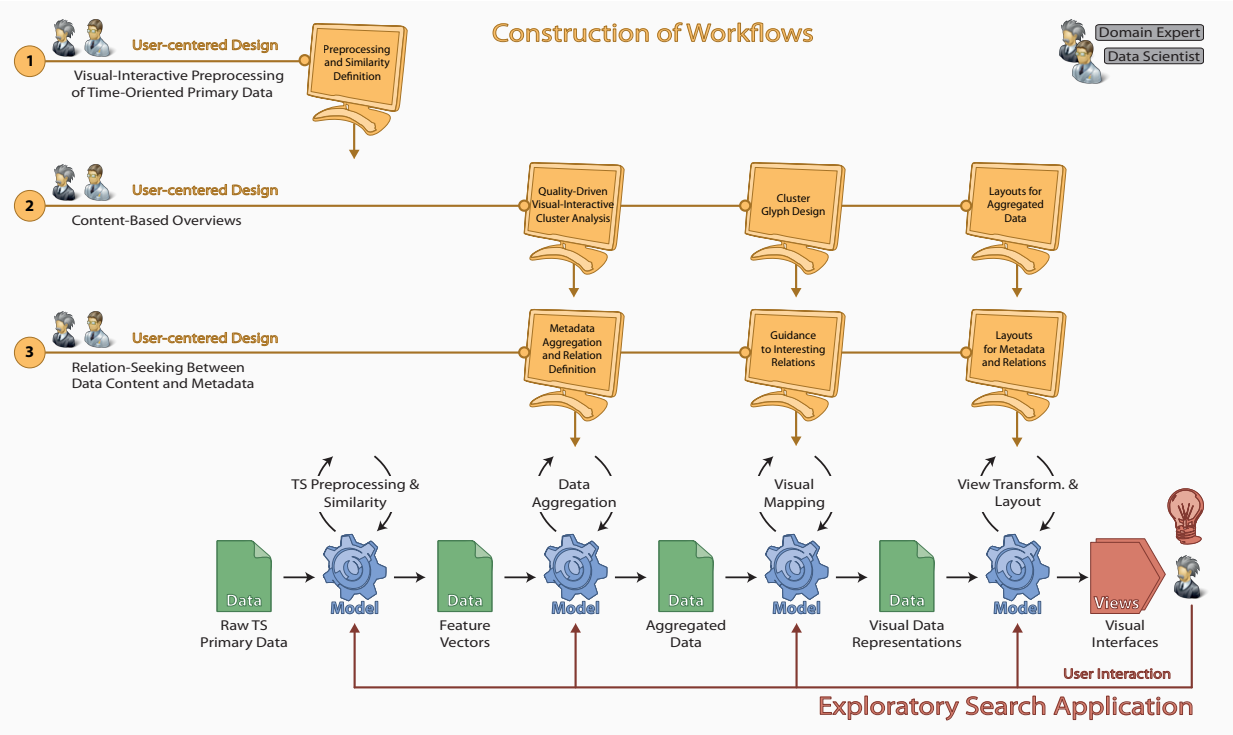


Figure 3.5 *The three main technical contributions of this thesis in association with the reference workflow.*

yellow workflow in Figure 3.5). We present visual-interactive techniques for the choice of appropriate data aggregation techniques and show how data aggregates can be visually encoded.Finally, we discuss view transformation and layout techniques for the arrangement of data aggregates on 2D displays.

Relation Seeking Between Data Content and Metadata With the third technical contribution, we support domain experts in seeking relations between data content and metadata $\mathbf{C}_{\text{C+M}}$ (see Chapter 6). We present different visual-interactive techniques, each with an individual focus on the identification of interesting relations. The all-embracing distinction between the techniques is the interactive definition of dependent variables for hypotheses formulation and testing, which may be based on the data content, metadata, or on both. To support exploratory data analysis, we also present a technique which reveals interesting relations without prior user assumptions or hypotheses. Guidance concepts automatically lead users to interesting relations hidden in possibly large sets of independent variables. The visual-interactive techniques can be integrated into ESS. To facilitate relation seeking between data content and metadata, we use the entire reference workflow (see the third yellow workflow in Figure 3.5). The first step of the reference workflow benefits from the results of the first contribution (visual-interactive time series preprocessing). Based on FVs as input data, we use the remaining reference workflow to carry out relation-seeking support. We use content-based overview solutions which will be created according to the second contribution (content-based overviews). Similarly, metadata is processed in a workflow, and thus made applicable for relation seeking. Finally, we present different visual-interactive techniques, relating the provided data content with metadata.

3.4.2. Two Case Studies

In Chapter 7, we present the two case studies of this thesis. In both cases, ESS are presented as the result of design study projects with real-world domain experts, real-world tasks, and real-world time-oriented primary data. In the first case study in Section 7.1, we present VisInfo, a DL system for time-oriented primary data in the Earth observation domain. VisInfo benefits from both the incorporation of the ES concept and techniques borrowed from IV and VA. In the second case study in Section 7.2, we present MotionExplorer, an ESS for the reuse of human motion capture data, supporting domain experts interested in, e.g., human motion synthesis. In both case studies, we use the reference workflow as a means of (1) constructing data-centered workflows for the downstream ESS and (2) carrying out

user-centered design. Both case study sections are structured as follows. First, we describe how we characterized the application domain, the data and the tasks. Second, we present results for the first step of the reference workflow, i.e., the visual-interactive time series preprocessing step. Third, we show how we constructed content-based overviews including data aggregation, visual mapping, and data layout, with the aid of the reference workflow. In this connection, we also present the interaction designs of the ESS which were elaborated in collaboration with the users. Furthermore, we describe the relation-seeking support between data content and metadata. Finally, we show additional evaluation strategies for both ESS. For more details, we refer to Chapter 7.

3.5. Summary

3.5.1. Discussion

Information Retrieval The reference workflow proposed in this chapter can be used to guide data scientists through the design process. In the next chapters, we will demonstrate the applicability of the workflow. We will present visual-interactive interfaces for ESS, such as content-based overviews, content-based query interfaces, views for reeking relations between data content and metadata, and representations of retrieval results. In this way, we support search and exploration activity in a common workflow. The second step in the workflow, however, focuses on the aggregation of data which mainly supports exploration activity. Thus, the data aggregation step is the only step in the reference workflow which does not guide both activities in an equally balanced way. The decision was made because, as opposed to data aggregation, this thesis does not contribute novel information retrieval techniques. This is why we emphasize data aggregation in the workflow as opposed to retrieval. Nevertheless, we consider the adaption of the step towards the terminology *data aggregation and retrieval* to provide gereralizability.

System Integration In the last step of the workflow, data scientists and domain experts create views (visualizations) which can be included into ESS. In IV and VA, it is common practice that powerful systems provide multiple views to enhance analytical capability. Concepts for linking views can be integrated to facilitate the lookup of objects in different views, possibly from different perspectives. However, we made the decision to end the reference workflow with the creation of single views. Thus, the integration of these views into systems is not represented. We recommend several instances of the reference workflow if a targeted ESS is intended to have multiple views. In a downstream *system integration* phase, these instances may be conflated to a single system. Additional visualization and interaction designs may be required to unify the visual interfaces and to facilitate usability. Again, we recommend the involvement of users in the process to foster the creation of useful and usable designs including, e.g., implementations for linking multiple views.

3.5.2. Conclusion

In this chapter, we introduced the concepts for this thesis. We presented two main contributions. First, we presented a *survey of IV and VA tasks* which we arranged to be a single assembly. In addition, we used the assembly to highlight search and exploration activity. The assembly provides an overview of the rich set of tasks and techniques presented in IV and VA and can support data scientists in the design process of ESS. The characterization of search and exploration activity in the context of IV and VA tasks allowed us to associate ES with techniques presented in IV and VA. Based on the characterization of search and exploration activity, we presented the definition of ES for this thesis. With the second contribution, we presented a *reference workflow* for the design and the use of ESS. The reference workflow adopts existing methodologies and reference frameworks from scientific workflows, KDD, IV, and VA. In addition, the reference workflow reflects the specific tasks (ES) and data (time-oriented primary data) applied in this thesis. Finally, the reference workflow combines a data-centered axis with a user-centered design axis in a single diagram. In other words, these two basic principles are conflated to a single process. The reference workflow describes the design of ESS by showing four main steps (models) in the design space, each of which can be implemented with IV and VA techniques in a user-centered way. Finally, we presented an outlook for the remainder of this thesis. In three following chapters, we will present guidelines and techniques for the design and the application of visual interfaces for ESS. The reference workflow will be employed as a baseline in all three chapters. Furthermore, we use the assembly of tasks and techniques presented in IV and VA for the design of visual interfaces for ESS. Similarly, the reference workflow will form the basis for two real-world ESS presented in the case study chapter.

CHAPTER 4

Visual-Interactive Preprocessing of Time-Oriented Primary Data

> *Every block of stone has a statue inside it and it is the task of the sculptor to discover it. Carving (a statue out of stone) is easy, you just go down to the skin and stop.*
>
> Michelangelo (1475 - 1564), *the creation of David between 1501 and 1504*

A variety of ESS for textual content exist, allowing users to formulate textual queries and explore textual document collections. For non-textual data types, however, the number of ESS approaches is scarce, especially for time-oriented primary data. Challenges in the content-based access to time-oriented primary data contribute to this lack of ESS approaches (cf. $\mathbf{C}_{\text{CBA}}$). The creation of FV, i.e., compact and yet precise data representations, is challenging, especially if the data collection is unknown. Data scientists have to select appropriate time series descriptors out of a huge set of candidates. In addition, data characteristics pose upstream challenges. The raw primary data (opposed to possibly curated secondary data) bear quality challenges which have to be faced. Similarly, the specific characteristics of time-oriented data have to be considered. Thus, workflows that provide content-based access to time-oriented primary data require preprocessing operations for data cleansing, normalization, sampling, and segmentation. Finally, specific user requirements add to the difficulty of the challenge. Domain experts may, e.g., expect data characteristics, require notions of similarity, or provide domain knowledge relevant for data abstractions. One goal in the construction of preprocessing workflows is the maximization of the degree of automation. However, due to the complexity of the design space an automation without human judgment is hardly feasible. It is challenging to support domain experts in assessing the sensibility and validity of the results. Moreover, cascades of multiple operations mask quality leaks caused by inappropriate model and parameter choices. Visualization is a means of resolving these challenges. Visual-interactive interfaces are greatly beneficial to facilitate the tight coupling of human inspection with automated model calculation. And still, visualization is too often just an end product of (scientific) workflows, especially for preprocessing tasks.

We present a visual-interactive system for preprocessing time-oriented primary data. Users are able to create complex workflows from a rich set of available routines. In addition, users are supported in the definition of time series descriptors and time series similarity. A guidance concept allows selecting appropriate model parameters. Moreover, the approach supports users in assessing the workflow generalizability for large and untested data. With the visual-interactive system, we provide a means for the content-based access to time-oriented primary data $\mathbf{C}_{\text{CBA}}$. This chapter is mainly based on [BRG*12] and partially based on [BBF*10, BBF*11, BRS*12a, BRS*12b, BSR*14, SBS11, SSW*12].

Contents

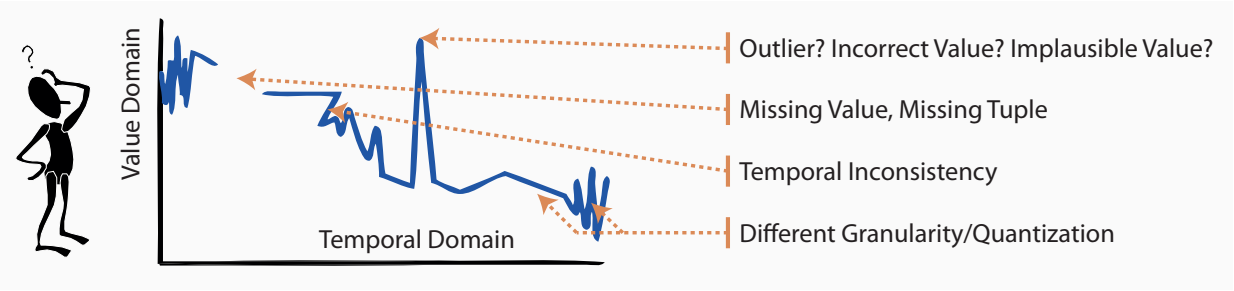


Figure 4.1 *Different quality leaks of time-oriented data. The illustration was used in the domain and data characterization phase of the VisInfo case study to discuss relevant data cleansing strategies (see Section 7.1).*

4.1. Introduction

4.1.1. Motivation

We have already emphasized the importance of content-based access strategies for the design of powerful ESS, e.g., for time-oriented primary data $\mathbf{C}_{\text{CBA}}$. In general, data preprocessing is a mandatory element in the analysis protocol of scientific workflows [DGST09]. In particular, this applies to primary data which are a direct product of a given source. The original condition of measured phenomena coexists with the increased risk of quality leaks. For the execution of (visual) search and exploration tasks the utilization of raw primary data is hardly feasible without prior processing and transformation (cf. Section 2.2). Preprocessing becomes necessary whenever the data does not match the requirements for the downstream analysis techniques. Time-oriented data has the special property of including a value domain and a temporal domain (cf. Section 2.3). Both domains have specific characteristics and require individual treatment. Relevant criteria to facilitate search and exploration tasks on time-oriented data are, e.g., *clean* data ensuring data quality, *compact* data providing scalability, and *comparable* data facilitating retrieval and data aggregation.

Providing data quality is one of the most important and tedious tasks in the preprocessing phase, until the data can be further processed by downstream visualization and analysis tools. *Dirty data* may lead to false assumptions and wrong statistical interpretation. An illustrative example of dirty time-oriented data is shown in Figure 4.1. Data cleansing (cleaning, scrubbing, wrangling) transforms raw data into usable data [KHP*11]. A variety of data quality aspects exist as shown in the comprehensive taxonomy of dirty time-oriented data by Gschwandtner et al. [GGAM12]. Important classes of dirty time-oriented data are, e.g., missing data, data duplicates, or incorrect values. However, it can be considered a chicken-egg dilemma that meaningful data cleansing capability requires (visual) data analysis, and the effectiveness of data analysis highly depends on the data quality. Kandel et al. [KHP*11] criticize a missing data wrangling capability in many existing tools. The authors identify the shortcoming that typical research papers tend to showcase their visualization results with previously cleaned data, but often neglect to mention how data errors were found and fixed. In fact, systems for visual-interactive preprocessing of time-oriented data are particularly scarce.

The scalability of downstream models depends on meaningful data abstractions. Various design choices for data transformations have to be made to yield compact and yet faithful time series representations. A prominent approach, e.g., inspired by retrieval methods, is the transformation of time-oriented data into FVs with time series descriptors (cf. Section 2.3.2). Descriptors generate FVs, preserving relevant properties and likewise neglecting irrelevant properties. In addition to data abstraction, many analytical models require similarity functions that make data comparable. Examples are retrieval algorithms calculating nearest neighbors, or clustering algorithms assigning large sets of data elements into groups. The design of useful similarity functions particularly relies on the choice of normalizations, descriptors, and similarity measures.

The usefulness of preprocessing workflows not only depends on the time-oriented primary data, but also on the targeted user group. For the discovery of comprehensible knowledge it is important to choose representations that can be interpreted by the domain experts [Mör06]. If it is not known whether the data matches the requirements of users and their tasks, it certainly is not advisable to apply an automated pre-process as a black-box operation. On the contrary, visualization enables humans to identify motifs in the data, and thus can serve as a complement to the computational power of machines. Moreover, visual abstractions hide implementation details and make workflows

more suitable for users who do not have substantial programming expertise (cf. Section 2.4). With visualization becoming a more integral part of the workflow, users can visually comprehend both the result of the workflow execution and the intermediate results of individual steps. The construction of workflows for time series preprocessing can be supported with interactive visual interfaces in combination with automatic techniques, as outlined in the VA agenda by Thomas and Cook [TC05]. An interactive adjustment of the pipeline facilitates learning about the characteristic data properties, checking the requirements, and testing the effects of suitable operations and their parameters. A best-practice VA example applied to the task of dimensionality reduction in high-dimensional data is the Dimstiller approach presented by Ingram et al. [IMI*10]. Understanding and transforming input dimensions is facilitated by a series of visual-interactive steps controlled by users.

We assume that visualization (of the model and the data) can facilitate the communication between different stakeholders involved in the process. The judgment of domain experts coupled with the expertise of data scientists fosters the construction of effective workflows. As an example, the data cleansing process requires domain knowledge to make informed decisions, and thus it is difficult to automate [KHP*11]. Likewise, assigning domain experts an active role in the process of implementing their notions of similarity may improve the usefulness of provided similarity functions [BSR*14]. In general, choosing appropriate models and model parameters can be improved by closer collaborations between domain experts and data scientists using visual-interactive interfaces.

4.1.2. Research Goals

The related work revealed six major research challenges (cf. Section 2.6), which we confront in this thesis. In this chapter, we meet the challenge of providing content-based access to time-oriented primary data $\mathbf{C}_{\text{CBA}}$. In addition, we face the associated challenges of choosing appropriate models and model parameters $\mathbf{C}_{\text{MPC}}$ and the challenges of involving users in the design $\mathbf{C}_{\text{UCD}}$. Based on a reflection of the involved challenges, we postulate nine *research goals* which have to be addressed to achieve useful and usable approaches for preprocessing time-oriented primary data. In this way, we provide a system for the effective and efficient content-based access to time-oriented primary data.

RG_{CBA1} Representation of Time-Oriented Primary Data In their survey, Kandel et al. characterize the challenge of choosing appropriate representations for raw data in general [KHP*11]. For time-oriented primary data different design choices exist for representing dirty data properties, such as outliers, missing values, duplicates, or data formatting issues $\mathbf{C}_{\text{CBA}}$. The challenge applies to both the underlying data structure (cf. Sections 2.2 and 2.3) and the visual representation of raw time-oriented data within the workflow (cf. Section 2.4). Both the data structure and the visual data representation should be general to support content-based access to different data sources and specific to cope with intrinsic properties and quality leaks of time-oriented primary data. An example of different quality leaks of time-oriented data is shown in Figure 4.1.

RG_{CBA2} Comparing Model Input with Model Output Opening the black-box of automated workflows enables visual inspection and direct manipulation [KHP*11]. However, closely combining human judgment and automated computation [BAF*13] requires observing the data manipulations made by every single method of the process. Choosing appropriate models in the appropriate order is a challenging task $\mathbf{C}_{\text{MPC}}$. The effects of different models to the downstream process are significant, just as the effect of propagated errors. The changes caused by an applied model must to be comprehensible. Thus, for a given model the direct comparison of input and output data is a challenge which must be resolved.

RG_{CBA3} Choosing Appropriate Parameter Values The challenge of choosing relevant models is accompanied with defining meaningful model parameter values [Fek13] $\mathbf{C}_{\text{MPC}}$. Most of the available methods have at least one parameter. Identifying suitable, if not optimal, parameter values is often a complex task [BRG*12]. It is important that users understand the role of parameters in order to avoid incorrect settings causing an algorithm to fail in finding true patterns, or to report serious patterns that do not really exist [KLR04]. The threshold value for assigning values as outliers may serve as an example. While data scientists often operate by rules-of-thumb, domain experts may judge outliers by more obvious criteria, depending on specific real-world occasions $\mathbf{C}_{\text{UCD}}$. Visualizations showing the *delta* of model outputs caused by different parameter values may have a positive impact on the decision-making process.

RG_{CBA4} Guaranteeing Data Quality The determination of when data is *clean* is a challenging task since there is not one definition of clean data [KHP*11]. The notion of data quality in itself as a human-centered challenge [LK06] $\mathbf{C}_{\text{UCD}}$.

Several data-centered challenges exist that influence the data quality $\mathbf{C}_{\text{CBA}}$. Taxonomies of *dirty* time-oriented data exist [GGAM12] providing information about the mere size of the design space of achieving data quality. We postulate that toolkits for time series preprocessing workflows should provide variety of *cleansing* techniques. This allows data scientists and domain experts to perform data cleansing based on the specific properties of the underlying data and the analysis tasks. A tight integration of visualization supports the diagnosis and problem solving process [KHP*11]. Furthermore, visualization is a beneficial means of including the user in the design process $\mathbf{C}_{\text{UCD}}$.

$\mathbf{RG}_{\text{CBA5}}$ Trade-off: Providing a Compact but Faithful Representation Downstream techniques require compact but yet representative transformations of the raw time-oriented data [Mör06], often referred to as FVs as a product of time series descriptors. A compact time series representation improves the performance of processing large repositories by eliminating irrelevant parts of the data [KK03]. However, choosing representative data abstractions may be arbitrary from a data scientists' perspective $\mathbf{C}_{\text{MPC}}$. Estimating the relevance of intrinsic data properties, and thus preserving representative information, is a challenge which should be resolved by involving the judgment of domain experts $\mathbf{C}_{\text{UCD}}$. Techniques are required to stay aware of how different preprocessing steps affect the outcome of the process $\mathbf{C}_{\text{MPC}}$. In this way, an important step in the content-based access to time-oriented primary data is addressed $\mathbf{C}_{\text{CBA}}$.

$\mathbf{RG}_{\text{CBA6}}$ Supporting the Users' Notion of Similarity Providing comparability by defining a similarity function can be seen as a useful by-product of time series preprocessing workflows for many applications, such as ESS. Similarity measures are most relevant for many downstream routines applied to time-oriented data $\mathbf{C}_{\text{MPC}}$. Measuring the similarity of two time series is typically provided with the definition of a distance measure, such as the Euclidean distance. However, in most cases, the users' notion of similarity is considerably more complex than simple distance measures can express $\mathbf{C}_{\text{UCD}}$. Supporting the users' notion of similarity may involve data quality aspects $\mathbf{RG}_{\text{CBA4}}$, meaningful time series representations $\mathbf{RG}_{\text{CBA5}}$, and other steps within the preprocessing workflow. As an example, Keogh and Kasetty emphasize the importance of additional normalization steps to "reveal the true similarity of two time series" [KK03]. Thus, the definition of similarity should comply with the users' notion of similarity $\mathbf{C}_{\text{UCD}}$.

$\mathbf{RG}_{\text{CBA7}}$ Generalizability of Workflow Configurations We have already argued that automated computation is efficient, but requires carefully designed workflows to guarantee the effectiveness $\mathbf{C}_{\text{MPC}}$. Human inspection and judgment, however, are a tedious manual process which is not applicable for large data sets. Consequently, the user is required to select a few time series to control and test the workflow. As a result, at least parts of the data set are left untested, especially if data sets are large or subject to change $\mathbf{C}_{\text{CBA}}$. The workflow execution with untested time-oriented data sets bears the risk that models and model parameters are chosen inappropriately to guarantee *generalizability* $\mathbf{C}_{\text{MPC}}$. A related challenge is not to *overfit* the pipeline when workflows are constructed with too small sets of testing data. To increase the robustness of the workflow, the workflow construction should at least approximate the diversity of the underlying data repository [BRG*12] $\mathbf{C}_{\text{CBA}}$.

$\mathbf{RG}_{\text{CBA8}}$ Reuse and Revision of Workflows Similar to the reuse of primary data and its enhanced value for data-driven research (cf. Section 2.2) workflows may be a subject of reuse and revision (cf. Section 2.4) . The reuse may resolve challenges when single domain experts want to repeat a data transformation process, or may foster collaborative approaches. Workflow revisions may facilitate coping with changing data efficiently, e.g., by editing the pipeline, or refining parameters [KHP*11] $\mathbf{C}_{\text{MPC}}$. Obvious actions are providing workflow visualizations and allowing interactive workflow adjustments. Challenges exist in the complexity of workflows potentially consisting of a variety of steps and step parameters which need to be visually represented $\mathbf{C}_{\text{MPC}}$. Moreover, the provenance information of constructed workflows needs to be preserved beyond single run-time sessions, and be re-stored at a later time. The VisTrails approach maintaining a history of the workflows serves as a convincing example [BCS*05].

$\mathbf{RG}_{\text{CBA9}}$ Involving the User in the Design Process Assigning domain experts an active role in the workflow construction provides transparency and facilitates collaboration $\mathbf{C}_{\text{UCD}}$. Visual-interactive capability has the potential to open data scientists' speciality to a broader audience by making the process observable and steerable. In the latter research goals, we outlined mandatory technical artifacts that will enhance preprocessing time-oriented primary data. In essence, involving VA techniques for the construction of workflows is a beneficial means of involving users in the process. However, visual-interactive interfaces require meaningful visualization and interaction designs for being usable and useful. More visual-interactive systems are needed for data preprocessing, since the number of inspiring

best-practice approaches is small [KHP*11]. General research challenges like closely combining human judgment and automated computation, or applying meaningful operations in a proper order [Fek13] prevail $\mathbf{C}_{\text{MPC}}$.

4.1.3. Contribution

We present a system for the visual-interactive definition of preprocessing workflows for time-oriented primary data. With this system, the user is able to select models from a large set of available preprocessing routines and drag them into a visual representation of the workflow. Visualization and interaction designs for the preprocessing pipeline enhance model control and parameter steering. The approach provides visual representations of the time-oriented primary data. The visual comparison of both the model input and the model output enables the assessment of effects for a given time series. In addition, a parameter guidance concept visualizes the output with alternative parameterizations for a given model and time series. As a result, users are to validate and, if necessary, improve the calibration of the parameter values. Finally, we contribute a guidance concept, supporting assumptions about the workflow generalizability for large sets of input time series. The system automatically captures the variation of the input data set and suggests additional, not yet considered time-oriented primary data for testing. As a result, the user is able to test the effect of the preprocessing workflow on a high variation of input data with only few test candidates.

A large set of models is provided for cleansing and wrangling raw time-oriented data. As a result, the user is able to improve the quality of the time series to a level expected, e.g., by downstream steps of the workflow. Moreover, different routines support the modification of both the temporal and the value domain of time series to put the time series in a comparable format. Examples are outlier reduction, normalization, or segmentation routines. The system supports users in facing the trade-off between a compact but yet faithful time series representation (FV). The definition of time series descriptors is supported in a visual-interactive way, including the guidance concept for descriptor parameter calibration. Finally, the approach enables users to define a distance measure which can subsequently be used by downstream models. With the construction of a workflow consisting of models for time series preprocessing, time series representation and distance measures, the system enables users to access the content of time-oriented primary data $\mathbf{C}_{\text{CBA}}$. The resulting time series FVs can be applied in downstream models, such as techniques for search and exploration support. Our system is one of the very first VA approaches enabling broader audiences, without expertise in data science, to construct preprocessing workflows for time-oriented data.

4.1.4. Relation to the Reference Workflow

In Figure 4.2, we draw the connection between the preprocessing pipeline presented in this chapter and the reference workflow for this thesis. It can be seen that the reference workflow addresses preprocessing of time-oriented data in an explicit step. In the first step of the reference workflow, raw time-oriented primary data is preprocessed and transformed into a usable format. In addition, a similarity function is defined. The output of the step is a set of FVs which can be applied by downstream models, such as retrieval, or data aggregation algorithms.

4.1.5. Chapter Overview

Section 4.2 discusses related work in the fields time series KDD and DM. Thus, we extent the review of general related work about temporal analysis tasks presented in Section 2.3.2 with specific techniques addressed in this chapter. Afterwards, in Section 4.3, we present our approach for the visual-interactive preprocessing of time-oriented primary data. In Section 4.4, we prove the usefulness of our approach. Section 4.5 discusses and concludes the approach.

4.2. Baseline Techniques

Time series preprocessing is of particular importance to transform time-oriented primary data into a usable format [KHP*11]. In this way, challenges regarding the *quality* of the primary data content can be addressed as a prerequisite to exploit the value of primary data (cf. Section 2.2). Other types of preprocessing techniques are the *normalization* of time series to foster comparability, the *segmentation* of time series to reveal patterns, or the *transformation* time series into the feature space to yield a compact and faithful representation. Typically, a set of these techniques is compiled as a (scientific) workflow, facilitating content-based access to time-oriented primary data. Preprocessing is considered the first fundamental step within the KDD reference workflow [FPS96], e.g., to perform downstream DM techniques, such as time series clustering [WL05]. A variety of surveys for time series preprocessing exist, many of them including

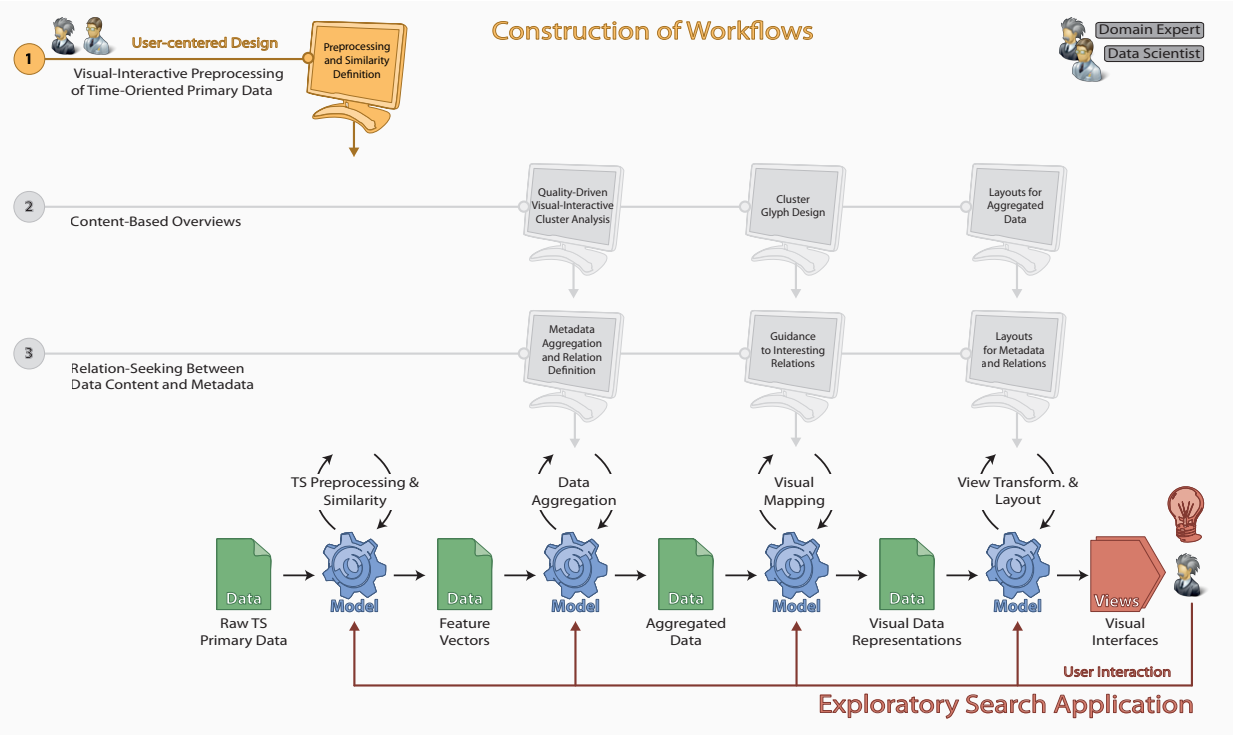


Figure 4.2 *The first technical contribution of this thesis is a visual-interactive system for the construction of preprocessing workflows for time-oriented primary data. In this way, we address the challenge of providing content-based access to time-oriented primary data* $\mathbf{C}_{CBA}$*. With this system, we present a means of handling the first step of the reference workflow in a user-centered way.*

taxonomies for the different classes of algorithms [FPS96, KK03, LKL*04, HKP11, Mör06, Fu11, AMST11, KHP*11]. In the following, we provide an overview of baseline techniques for cleansing, normalizing, sampling, and segmenting time-oriented data, as well as a review of time series descriptors and distance measures for time series.

4.2.1. Preprocessing Techniques

Cleansing of Time-oriented Data The overall goal of data cleansing techniques is to make data usable for downstream preprocessing and DM techniques. Individual cleansing strategies often consist of an error detection and an error handling phase. We review relevant techniques based on the taxonomy of dirty time-oriented data by Gschwandtner et al. [GGAM12]. The detection and removal of missing data is one of the most relevant and often applied data cleansing issues. Entries in the value domain (*missing values*), or complete *tuples* (elements) may be missing. Depending on the data source and the data curation policy, missing data is often tagged with a specific indicator, such as *NULL*, or *NaN*. However, time-oriented primary data may contain *dummy entries*, such as -999, or -1, causing specific interpreting effort. Other possible causes of dirty data are *duplicates* or redundant data stemming from multiple entries for the identical time stamp. These duplicates may be identical, or inconsistent. Typical causes of duplicates (and missing tuples) are data formats for regions on Earth where the time is changing two times a year. This issue actually occurred in the VisInfo case study presented in Section 7.1, and had to be cleansed, successively. *Implausible values*, *outdated temporal data*, *ambiguous data*, or different types of *wrong data* complete the taxonomy of dirty time-oriented data for single sources. Examples which occurred in our case studies in Chapter 7 are, e.g., a measurement taken in the year 2099 instead of 1999, or a value a thousand times as high as purposed (10,000 instead of 10.000). Such abnormally large values cause masking effects by affecting modeled distributions to an extent that other extreme values may be interpreted as normal [KHP*11]. Detecting and handling *outliers*, *anomalies*, or *novelties* are important data cleansing techniques. A standard outlier detection technique is the comparison of the value domain, a multiple of the standard deviation defining a steering parameter. An example of the visual identification of time series anomalies is presented by Kumar et al. [KLK*05]. Features extracted with the symbolic SAX descriptor [LKLC03] are visualized with a bitmap metaphor. Colors are used to encode the relative frequency of the features, and thus

enable the identification of clusters and anomalies. Prominent classes of outlier, anomaly, and missing value removal techniques are *interpolation* and *regression* models. However, the robust linear interpolation fails if the time interval including missing values is too sizable, e.g., larger than the intrinsic periodicity of the time series. Domain-specific solutions for missing value handling extend the validity of the last value until a new value is available as seen, e.g., in financial time series [Mör06, p. 23]. Many analysis tasks focus on global patterns in time-oriented data in favor of local properties. For the *reduction of noise*, *moving average* techniques are often employed, providing parameters for the moving average kernel function (see, e.g., [Mör06, p. 22]). Other causes of erroneous data are formating issues, like different date standards [KHP*11]. Especially interval-based data representations of the temporal domain often require special treatment since calendars constitute a non metric space (e.g., a month may have 28 or 31 days) [AMST11]. Finally, we emphasize another relevant data cleansing class for many downstream techniques: avoiding non-equidistant representations of the temporal domain [WL05]. Provided that the time intervals between all chronological time stamps are equidistant the time domain can be neglected and the value domain be processed as an array of (numeric) values.

Normalization of Time-oriented Data With normalization, we understand modifications on the temporal domain, or the value domain to adjust time-oriented data of different scales to a common scale. Normalization is relevant for unbiased, natural, and subjectively correct similarity calculations [KK03, WL05]. Thus, one of the most relevant goals of normalization methods applied in this work is to provide comparability. For the value domain, important techniques are the *min-max*, the *0-max*, the *offset translation*, the *amplitude scaling*, or the *quantile* normalization. Techniques for the removal of trends are often used as a normalization step. Linear or higher order functions can be fitted to the time series to keep the residuals [Mör06, p. 22]. However, these methods tend to amplify noise which may require additional treatment, like the application of moving average techniques. An example of *linear trend removal* is Earth observation. When the analytical focus is on the comparison short time periods (e.g., daily patterns) it is relevant to previously differentiate global trends possibly caused by climate change. A normalization technique for the temporal domain is temporal offset translation which is applied to compensate external effects, e.g., caused by different climate zones. Finally, metadata can be transformed into a comparable state. For example, ZIP-codes can be converted into the latitude-longitude centroids based on the geo locations. A unique property of data normalization is their applicability at virtually all positions within the preprocessing workflow. An example of an early utilization of normalization within the workflow is the translation of value offsets. The average value of any time series is subtracted which puts the focus on variations within the value domain. In this way, seasonal patterns or trends can be revealed. In a later phase of the pipeline, normalization might be applied to more fine-gained patterns, like subsequences. To this end, a local min-max normalization applied to segmented time series subsequences provides time series patterns in a relative scale. Finally, normalization techniques are used at a late state of the time series preprocessing pipeline, e.g., to adapt the value domain of time series features [FPS96, HKP11]. Naturally, multiple normalizations at different positions of the pipeline can be used in a single time series preprocessing workflow.

Segmentation of Time-oriented Data Many indexing, classification, and clustering approaches are applied to time series subsequences [KK03, Fu11] requiring the segmentation of time series. For a comparison of different segmentation techniques drawing on scalability benchmarks, we refer to the work of Keogh et al. [KK03]. Other relevant surveys are presented by Mörchen [Mör06, p. 35] and Fu [Fu11]. In general, the output of time series segmentation models differs between patterns of equal length and of unequal length. Segmentation techniques can further be subdivided into three classes. *Sliding window* approaches traverse time series linearly for the extraction of segments based on different domain-specific criteria and thresholds. For example, in Earth observation research, segments with a pattern length of one day or one year may be relevant to observe natural phenomena [BBF*11], while in stock chart analysis weekly patterns (Monday to Friday) are segmented [SBVLK09]. As a result, in domain-specific issues in which no fixed time interval for the time series segmentation is appropriate, the segmentation models may become more complex. *Top-down* approaches recursively partition time series until a termination criterion is met. One example is the Perceptual Important Points (PIP) algorithm where the number of important points is successively increased until the targeted parameter value of interesting points is achieved. *Bottom-up* approaches start with a configuration where every data element is an individual segment. In the course of the calculation, segments are merged subsequently until a stop criterion is reached. Keogh et al. approve linear scaling bottom-up approaches with a quality above average [KCHP04]. However, the performance of segmentation techniques of all three classes differ, particularly with respect to the series of real-world data sets [KK03]. From an implementation point of view the outputs of time series segmentation pose a special characteristic. While many other preprocessing models accept one input time series to reveal one output time series, segmentation typically constitutes a one-to-many operation.

4.2.2. Time Series Similarity

Descriptors for Time-oriented Data Time series descriptors are compact representations of time-oriented data preserving relevant data characteristics [KK03, LKLC03, DTS*08, Fu11]. Descriptors often *transform* the time-oriented data into another data (vector) space, which is why we use the term *FV* to describe the output of descriptor models. The notion of relevance for the preservation of data characteristics is highly data and use-case dependent, which may explain the great variety of descriptors and steering parameters. Descriptors with high compression rates have been presented abstracting the data to compression rates above 95%. Benchmark tests for descriptors, distance measures, and different real-world data sets are provided in the survey of Keogh et al. [KK03]. Providing both a good compression and a high fidelity of the time-oriented data is a trade-off [KCHP04], requiring careful design considerations. As a general rule, distorting the shape of the time series too much creates problems, e.g., through to high compression rates [Fu11]. We divide time series descriptors into the three classes of *model-based* [Mör06], *non-data adaptive*, and *data adaptive* descriptors [DTS*08].
In *model-based* approaches, models are fitted to the time-oriented data. The resulting model parameters are used for further DM tasks [Mör06]. Hidden Markov Models may serve as examples. Descriptors based on time series statistics (means, standard deviation, linear trend, periodicity, etc.) can also be assigned to the model-based class. We apply such a time series descriptor to estimate the dissimilarity of time series, and thus to guide users in the choice of untested time-oriented data. *Non-data adaptive* descriptors transform the time-oriented data into another space, properties of the value domain have no influence on the procedure. The Discrete Fourier Transform (DFT) transforms the time-oriented data into the spectral space, also called the 'frequency domain'. DFT-based approaches were combined with R* tree index structures, e.g., presented in the work of Agrawal et al. [AFS93]. With the Discrete Wavelet Transform (DWT) the output is transformed into wavelets, preserving both spatial and frequency information [CF99]. Another descriptor relevant for this thesis is the Piecewise Aggregate Approximation (PAA) descriptor which aggregates time series subsequence patterns of equal length to a single numeric value [KCPM01]. A prominent method transforming the time-oriented data into the symbolic space is the Symbolic Aggregate Approximation (SAX) [LKLC03]. Based on a discretization of the value domain into symbols usually provided by applying PAA, the SAX descriptor produces subsequences of symbols by using a sliding window approach. A visualization tool applying the SAX descriptor is VisTree [LKL*04] allowing visually mining and monitoring of large time-oriented data collections. *Data adaptive* time series descriptors consider both the time domain and the value domain in their procedure. As a consequence, these class of techniques can better approximate the time-oriented data (e.g., by allowing non-equidistant patterns), but the comparison of the output data is more difficult [Mör06]. Some approaches are based on piecewise segments, like the Piecewise Linear Approximation (PLA) descriptor [KP98]. Other techniques apply the Singular Value Decomposition (SVD) as a integral part of the Principal Component Analysis (PCA) [Jol02] providing the optimal linear transform in the sense of energy preservation [Mör06]. Moreover, the Perceptual Important Points (PIP) algorithm was applied as a data-adaptive time series descriptor. One example is the work of Ziegler et al. [ZJGK10] where large collections of financial time series data are analyzed in a visual-interactive system.

Distance Measures and Similarity Definition Measures for time series similarity (or distance measures respectively) are required for a variety of downstream analysis tasks. Distance measures need to be chosen with care to reflect the similarity of the underlying time-oriented data [KK03, WL05, DTS*08]. In addition, the involved user group has a significant influence on the appropriateness of distance measures. Downstream applications using time series distance measures are, e.g., *retrieval*, *classification*, or *clustering* [HKP11, Mör06]. A variety of different time series distance measures have been presented, relevant surveys are, e.g., [KK03, WL05] [Mör06, p. 23] [DTS*08]. We divide time series distance measures into the classes of lock-step measures and elastic measures.

Lock-step measures compare the i-th points of two time series [DTS*08]. Important representatives are the L_p-norms, especially L_1 (Manhattan distance), L_2 (Euclidean distance), and L_{inf} (infinity norm). The Euclidean distance is maybe the most commonly applied time series distance measure. L_p-norms are easy to implement and to evaluate. Especially for large data sets the L_p-norms are competitive in comparison with more complex techniques [DTS*08]. Moreover, L_p-norms are parameter-free. However, the comparison method based on fixed indices is highly vulnerable to noisy data and inapplicable for phase-invariant comparison. Upstream preprocessing transformations are required to make L_p distances invariant to amplitude scales and translations [Mör06]. Considerable downsides of lock-step measures are that similar segments with small temporal distortions are overlooked and moreover the underlying time series need to have the same length. One of the first examples using lock-step measures is the work of Agrawal et al. applying the Euclidean distance on DFT features [AFS93]. *Elastic measures* allow time series to be stretched or compressed to provide a better match [DTS*08]. In this way, small distortions of the time axis can be addressed [Mör06]. The

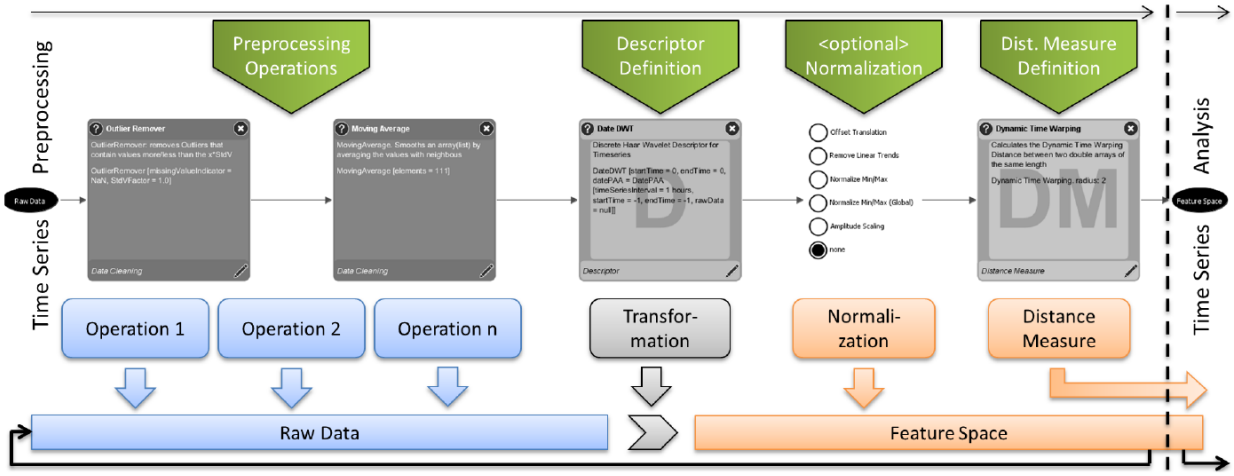


Figure 4.3 *Principal steps of the pipeline for time series preprocessing [BRG*12]. At a glance, raw time-oriented primary data is transformed into the feature space. Gray rectangles illustrate the visual representation of individual models of the pipeline in the visual-interactive system.*

Dynamic Time Warping (DTW) algorithm is one of the most important elastic time series distance measures finding the optimal alignment between two time series [SC07]. As a consequence, DTW can calculate distances for time-oriented data of different length. However, while the original version of the algorithm is also parameter-free, it suffers from a quadratic time and space complexity $O(n^2)$. For DTW different improvements have introduced incorporating parameters, like the warping window size. Moreover, several lower bounding measures have been presented to speed up similarity search using DTW [DTS*08]. As an example, the FastDTW algorithm introduced by Salvador and Chan approximating DTW has linear time and space complexity $O(n)$ [SC07]. A further subclass of elastic measures is the edit distance for strings. A prominent representative is using the longest common subsequence model [DTS*08]. A threshold parameter ε serves as a matching criterion if the distance of two time series is below ε.

4.3. Visual-Interactive Preprocessing of Time-Oriented Primary Data

In the following, we present a visual-interactive system achieving the nine research goals postulated in Section 4.1.2. For this purpose, the system uses baseline techniques for preprocessing time series and for defining time series similarity, as reviewed in Section 4.2. In addition, the system adopts guidance concepts and techniques from VA research to enhance the construction process of preprocessing workflows. The principle steps of the preprocessing pipeline covered by the system are shown in Figure 4.3. At a glance raw time-oriented primary data can be manipulated by a variety of preprocessing routines until it is transformed into the feature space. Together with a user-defined similarity measure these time series FVs define the output of such preprocessing pipelines. In this connection, the system covers the first of the four major design spaces of the reference workflow proposed for the design and application of ESS (cf. Figure 4.2). An overview of the visual-interactive system is presented in Figure 4.4. Following, we present six integral parts of the system which address the nine research goals for preprocessing time-oriented primary data.

1. Concepts and building blocks describe the overall pipeline, the data abstraction, and the visual representation
2. The Preprocessing Toolbox is a visual component providing baseline preprocessing techniques
3. The Workflow View visually represents the constructed preprocessing workflow
4. The Raw Data Selection View supports users in gaining an overview of the input data
5. The Detail View allows the input-output comparison of models and respective parameters
6. The Statistics View supports the workflow construction with statistical information

4.3.1. Concepts and Building Blocks

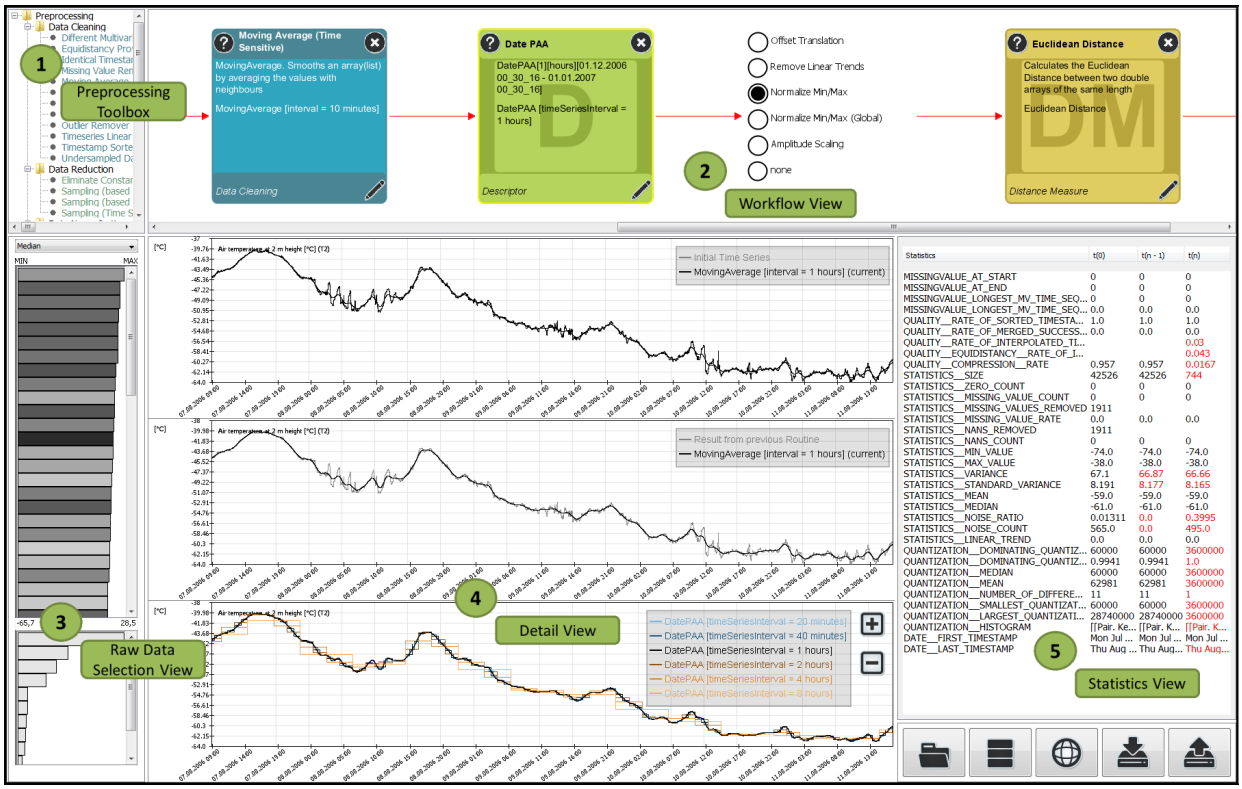


Figure 4.4 *The visual-interactive system for time series preprocessing. The visible part of the preprocessing workflow (top) consists of a moving average routine, a Date PAA descriptor, a min-max normalization, and a distance measure (Euclidean). Different views provide details about the data and models used in the preprocessing workflow.*

Refining the Step for Time Series Preprocessing and Similarity Definition As a first step, we refine the reference workflow for the design of ESS components (cf. Figure 4.2). In this chapter, we focus on the first of the four data transformations outlined in the reference workflow. We elaborate the step *Preprocessing and Similarity Definition* with a pipeline for the visual-interactive specification of time series preprocessing applications (workflows).

A schematic illustration of the preprocessing pipeline is shown in Figure 4.3. The overall goal of the preprocessing pipeline is the transformation of a time series into the feature space for being able to carry out downstream analytical models. In this way, the pipeline is a means of coping with a variety of time-oriented primary data sources $\mathbf{RG}_{\text{CBA1}}$ providing access to time-oriented primary data. We divide the pipeline into a phase when raw time-oriented data is processed and a phase which directly applies to the FV space. In between these two phases, the time series descriptor transforms the raw time-oriented primary data into the FV space. A variety of preprocessing operations can be added to the pipeline and be (re-)arranged in arbitrary order. Important goals are to provide data quality, to sample the data, and to segment the time series into relevant patterns (cf. Section 4.2). In addition, the preprocessing workflow requires the definition of a time series descriptor. For example, the Discrete Fourier Transformation may be applied to transform the time series into the signal space. The choice of the descriptor depends on the targeted trade-off between compactness and fidelity. After the transformation of the time-oriented data into the feature space, the system provides an optional normalization step to make different FVs comparable. The definition of a distance measure concludes the reference pipeline. The time series representation is then applicable for downstream DM techniques which can directly be applied to the FVs.

Representation of Time-Oriented Primary Data For the construction of preprocessing workflows a careful abstraction of time-oriented primary data is required. On the one hand, the abstraction should reflect specific characteristics of primary data (cf. Section 2.2). On the other hand, the data abstraction should be general to enable content-based access to heterogeneous primary data sources. Similarly, the visual representation of the time-oriented data should reflect the characteristics of primary data. In the following, we define a set of requirements for the visualization of time-oriented primary data for visual-interactive systems for time series preprocessing. These

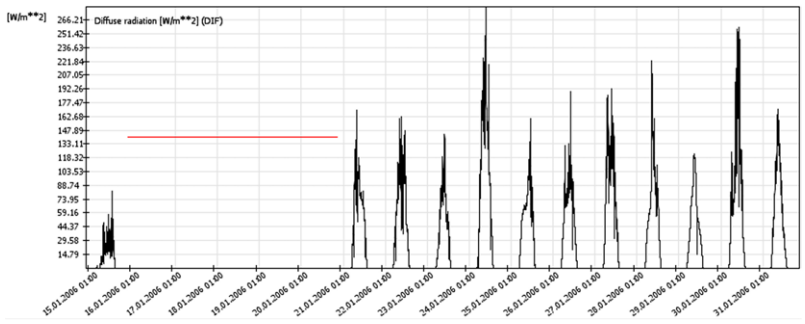
requirements can serve as a guideline for resolving the research goal of representing time-oriented primary data $\mathbf{RG}_{\text{CBA1}}$. As opposed to general time series visualization systems, the visual encodings for raw time-oriented data in preprocessing systems have to master additional complexities. From a high-level perspective, these additional complexities can be divided into two groups. First, visual representations need to be applicable for raw time-oriented primary data possibly containing a variety of quality leaks (cf. Sections 2.2.1 and 2.3.1, $\mathbf{RG}_{\text{CBA1}}$). Second, the visual representations have to facilitate the comparison of multiple time series as a baseline for addressing the model and model parameter choice $\mathbf{RG}_{\text{CBA2}}$ and $\mathbf{RG}_{\text{CBA3}}$. We list a set of functional requirements for visual encodings of time series visualizations.

- Visual encodings should be applicable for various *data abstractions* for time-oriented data
- Visual encodings should provide a solution for the *identification of missing values*
- Visual encodings should support the adaption of the *scale of the value domain*
- Visual encodings should *represent the physical unit* of the value domain of the primary data, if necessary.
- Visual encodings should enhance the *comparison of input and output* time series of given models
- Visual encodings should enable the visualization of multiple time series to facilitate *comparison* tasks

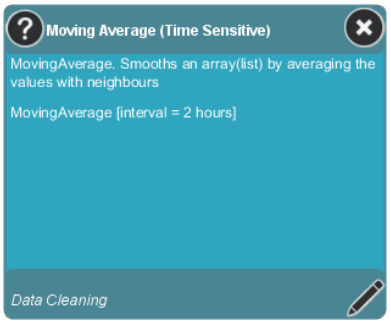
Data Model In the related work, we characterized time-oriented data (cf. Section 2.3.1). We reviewed models for the temporal domain and the value domain. In addition, we characterized relations between the temporal domain and the value domain. The challenge that can be infered from characterization is the impossibility to simultaneously cover all aspects of time-oriented data within a single visualization process [AMST11]. This is why our solutions for the visual representation of time-oriented primary data require a data model in full awareness of the characteristics of time-oriented data. In particular, we model time values in a discrete way of milliseconds with a point-based scope in a linear arrangement (cf. Section 2.3.1). The high temporal resolution of the data model enables the user of our system to model quasi-continuous time-oriented data. The value domain of our data model is quantitative with no restrictions to the frame of reference. For the visual-interactive preprocessing of time series, we interpret the data in an abstract way. This data model supports both accessing large varieties of time-oriented primary data and representing most of the intrinsic properties of time-oriented data in general.

Visual Time Series Representation We refer to the book of Aigner et al. for a comprehensive survey of visualizations for time series [AMST11]. In accordance with many approaches presented in the survey, our visual encoding uses a baseline linechart for the visualization of the temporal and the value domain. Our solution also meets the requirements for the representation and the visual-interactive preprocessing of time-oriented primary data, presented above. Figure 4.5a shows our solution for the visual encoding of (raw) time-oriented primary data. The visualization on the left indicates missing values with red dots at the center of the y-axis. In this way, missing values can be identified even if the line chart display is overplotted, i.e., if the number of data points exceeds the number of available pixels in x-direction significantly. Another relevant property for the visualization of raw time-oriented data is the *scale of the value domain*. In many cases, a linear scale of the y-axis may be most relevant [AMST11]. However, especially for science-driven applications and in data-driven research, a logarithmic scale of the y-axis is desirable on demand. Providing both scale types in an alternating or even comparative manner provides different views on the same underlying data. The definition of the scale is a user parameter. In addition, while many time-oriented data in time series analysis systems are shown dimensionless, we recommend providing a visual time series representation that shows the *physical unit* and, if necessary, a short description of the physical unit. This is especially helpful to relate and compare multiple time series visualizations with different physical units. The physical unit in Figure 4.5a is *[W/m**2]*, the physical unit description is *Diffuse radiation [W/m**2] (DIF)*. Another important property of time series visualizations is the applicability for comparison tasks [AA06]. A good example of how visual comparison tasks can foster time series preprocessing is presented by Keogh et al. [KCHP04]. In a juxtaposition setting a time series segmentation routine is shown with different results based on user-defined thresholds. A space-saving alternative to juxtaposed visualizations is superposition that visualizes multiple time series on top of each other with different layers [GAW*11], which we refer to as *bundling*. Thus, the time series visualization applied in our visual-interactive system uses the bundling technique to plot multiple overlaying time series in the same space. Finally, we use different colors to distinguish individual time series shown in a bundle. As an example, Figure 4.6 shows multiple time series with different colors including black, brown, and blue. While the bundle visualization enables users to differentiate the shapes of different time series, the use of color additionally facilitates the identification and localization of individual time series [AA06]. Obviously, we provide a legend to indicate the time series and their respective color values.

Model Output Representation We present the representation of the model output provided in the approach, relevant to be able to compare changes caused by models and by different model parameters. For the exploration



(a) *Visual encodings of raw time-oriented primary data including the indication of missing values, the adaption of the y-scale (between linear and logarithmic), and the visualization of the physical unit.*



(b) *The visual representation of a preprocessing routine. Here: moving average with a 2-hour kernel function.*

Figure 4.5 *Baseline techniques for visual-interactive time series preprocessing. Left: visual encodings of raw time-oriented data. Right: visual representation of a routine for the visual-interactive preprocessing pipeline.*

of model outputs based on parameter spaces, we draw inspiration from the visual encodings presented by Berger et al. [BPFG11]. As it can be seen in the upper two examples of Figure 4.6 the bundling capability can be applied to the *comparison of input and output* time series of a given preprocessing routine. In these examples the result of a previous routine in the pipeline (the gray time series containing noise) is adapted by a moving average routine, here with a kernel interval of 1 hour. The output of the routine is shown with black color. A smoothened time series can be seen following local trends, but suppressing the noisy behavior of the input time series. Of course the estimation of noise is up to the user and the usefulness of the applied routine depends on the use case. In any case, with the comparison of input and output time series by means of superposition, we provide the basis to address research goal $\mathbf{RG}_{\text{CBA2}}$. The second research goal achieved by the bundling capability combined with different color codings is the parameter guidance $\mathbf{RG}_{\text{CBA3}}$. Choosing appropriate parameter values can be facilitated by showing colored time series bundles of different output solutions of a preprocessing routine based on parameter modifications. With the bundles, users are empowered to *compare different model outputs*. In combination with a legend visualization showing the parameter configuration of different colored time series outputs, an effective visualization is provided to reach research goal $\mathbf{RG}_{\text{CBA3}}$. The plus and the minus icons in the legend enable users to change the number of alternative parameterizations. A click on a particular alternative parameterization triggers an event which can be used by the system to interactively adapt the parameterization. The example visualization at the bottom of Figure 4.6 demonstrates the effectiveness of the presented output comparison technique. The output time series of the moving average routine with the current parameter value with an kernel interval of 1 hour is shown with black color. Additionally, three alternative parameterizations with lower kernel intervals (15, 30, and 45 minutes) and four alternative parameterizations with higher kernel intervals (2, 4, 8, and 16 hours) are shown with different colors in a superposition setting. A bipolar colormap from brown to blue allows differentiating the output time series, drawing on modifications in the color saturation. Users can compare value progressions of the different outputs and additionally identify the parameterization of respective time series by looking up the linked color values in the provided legend.

With the visual encodings, we provide a visual representation of raw time-oriented data which achieves the research goal of the raw time series representation $\mathbf{RG}_{\text{CBA1}}$. Simultaneously, the visual representation builds the baseline to address the research goals of the model input-output comparison $\mathbf{RG}_{\text{CBA2}}$ and the parameter guidance $\mathbf{RG}_{\text{CBA3}}$.

4.3.2. Preprocessing Toolbox

The component at the upper left of the system (cf. Figure 4.4 (1)) provides a tree-view of baseline techniques for preprocessing time-oriented data. The *Preprocessing Toolbox* is structured in six main classes of techniques according to the review of baseline techniques for time series preprocessing in Section 4.2.

- data cleansing (e.g., missing values, outlier detection, moving average, etc.)
- data reduction (e.g., sampling, PIP sampling, etc.)
- data normalization (e.g., min-max normalization, offset translation, amplitude scaling, etc.)
- data segmentation (e.g., interval-based, index-based, PIP-based, etc.)

- descriptors (e.g., PAA, DTW, DFT, PIP, etc.)
- distance measures (e.g., Euclidean distance, Dynamic Time Warping, etc.)

As an example, the Preprocessing Toolbox provides a rich set of data cleansing routines which serve as a basis to address the research goal of providing data quality for time-oriented primary data $\mathbf{RG}_{\text{CBA4}}$. Every subtree in the Preprocessing Toolbox contains a set of routines, all of which can be applied to the workflow construction. The tree-based Preprocessing Toolbox provides tooltip capability showing the name and a short description of the routine. Drag-and-drop functionality enables users to add preprocessing routines to the workflow presented in the Workflow View on the right of the Preprocessing Toolbox. Every routine is an implementation of a software interface which we call a *PreprocessingRoutine*. Mandatory properties of a *PreprocessingRoutine* are the name, the description, the parameter overview, the parameter steering, the class of the technique, and naturally the functionality of the preprocessing routine. We make this technical artifact explicit for three reasons. First, because our approach is one of the first VA systems for time-oriented primary data in general. Second, the functional implementation of the *PreprocessingRoutine* allows integrating different technical solutions, such as external libraries, internal frameworks, or algorithms as a result of user-centered design. In this connection, we want to point out the leverage point to algorithmic libraries for time-oriented data, such as the *TimeBench* [1] software library. Third, the interface defines the principal properties of a preprocessing routine required for the visual representation of the routines in the pipeline. For instance, the visual representation of a Moving Average routine is presented in Figure 4.5b.

4.3.3. Workflow View

The *Workflow View* is the visual representation of the preprocessing pipeline (cf. Figure 4.4 (2)). Mandatory steps in the pipeline are preprocessing routines on the left, a time series descriptor, an optional FV normalization step, and a distance measure on the right. The visual representation of the preprocessing makes the process of the workflow construction visually comprehensible and facilitates the involvement of domain experts $\mathbf{RG}_{\text{CBA9}}$. The structure of the pipeline design is based on the illustration shown in Figure 4.3. A primary data source can be defined and be loaded into the system. Different parsers are provided to access multiple primary data sources. As a result, different raw primary data sources can be represented in an internal data format $\mathbf{RG}_{\text{CBA1}}$. The system provides different icons at the lower right for loading time series from file, from a database connection, or from a web-repository. After a primary data source is loaded, preprocessing routines can be dragged and dropped from the Preprocessing Toolbox to a particular position within the workflow. The number and the order of routines is not restricted and can be defined by the user. By adding a routine into the pipeline the workflow is executed for a selected time series up to the stage of the routine in the workflow. Thus, the effect of changes on the pipeline can be analyzed on-the-fly, which is one of the most desirable goals for scientific workflows in general (cf. Section 2.4). Users are able to observe the modifications on the time series for every single routine $\mathbf{RG}_{\text{CBA2}}$ $\mathbf{RG}_{\text{CBA3}}$. Changing the order of the routines in the pipeline is also possible by drag-and-drop interaction. In this way, users can analyze the impact of step permutations in the workflow, and thus validate and further enhance the workflow. Similarly, users are able to add a time series descriptor to the pipeline. An additional step in the workflow regards the normalization of FVs after the time series descriptor was applied (cf. Figure 4.3). The FV normalization is located after the descriptor calculation step. The final step in the preprocessing workflow regards the definition of a distance measure for downstream routines. Similar to the other steps of the pipeline, the definition of a distance measure can be performed visual-interactively, and thus does not require any programming knowledge or batch scripting. Hence, we encourage users to define similarity functions by themselves, based on their notion of time series similarity $\mathbf{RG}_{\text{CBA6}}$.

4.3.4. Raw Data Selection View

We demonstrate how the system assesses the workflow generalizability for large data sets $\mathbf{RG}_{\text{CBA7}}$. The *Raw Data Selection View* provides a list-based visual interface of the loaded time-oriented primary data (cf. Figure 4.4 (3)). Users can select single primary data to test the constructed workflow for different data characteristics. Each raw data object is visually represented as a horizontal bar. By dragging the bar into the Detail View the particular time series is visualized and the modifications of the current workflow on the time series can be analyzed. The bar size of each raw data object corresponds to a statistical property selectable with a combobox. Users are able to compare the statistical values of different raw data, and thus to assess the variability of respective statistical properties. As an

[1] TimeBench, https://github.com/ieg-vienna/TimeBench, last accessed on Sept 30th, 2015.

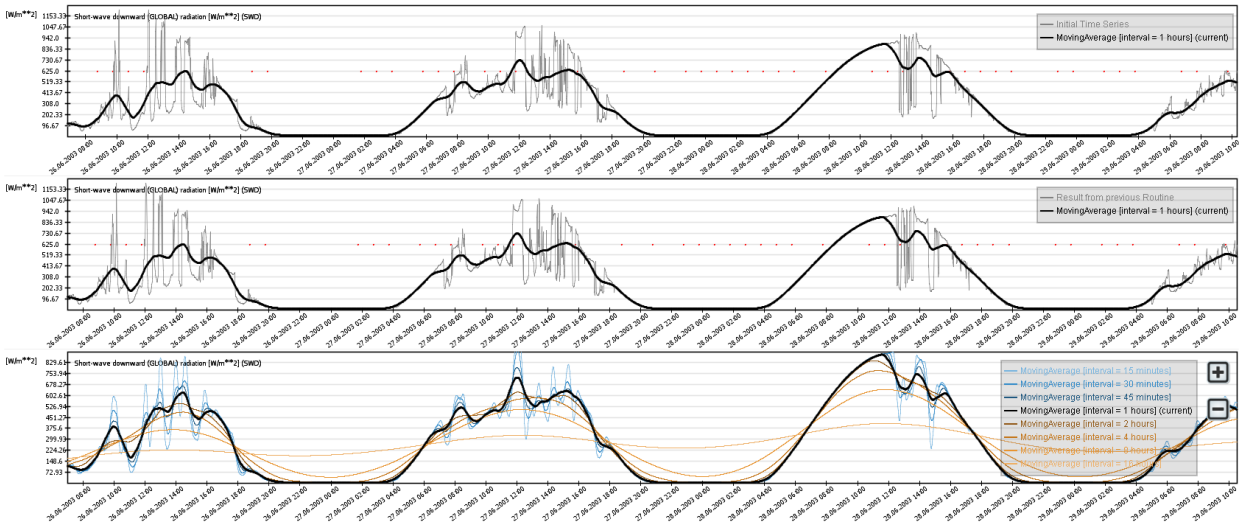


Figure 4.6 *Model input-output visualization (upper two linecharts) and parameter guidance visualization (lower linechart). Removal of missing values and calibration of a moving average routine. A kernel parameter of 1 hour sufficiently removes local noise provided by the raw Shortwave-Downward radiation measurement.*

example, the median of the raw time-oriented data is used in Figure 4.4. The upper time series with the longest bar has the highest median value in the entire data collection. The remaining primary data is arranged in descending order based on the user-selectable statistical property. A scrollbar interaction allows access to large primary data collections. Other relevant statistical properties are the number of missing values, the linear trend, the noise rate, or the min, max, mean, and standard deviation values. This rank-based visualization of statistical properties is a means of estimating the variability of the data collection. Consequently, users can estimate the workflow generalizability for the tested time series.

To further enhance the workflow generalizability of the workflow, the data Raw Data Selection View offers an additional guidance concept. We indicate the degree of *dissimilarity of unused primary data* compared to the used data. To this end, we use gray-value colormap (from black to light gray) to encode the degree of dissimilarity of every horizontal time series bar. Primary data with black colors are most dissimilar to all data previously tested by the user in the preprocessing pipeline. Consequently, black bars are mostly recommended by the system as test time series to enhance the workflow generalizability. Raw data with light gray colors are most similar to time series already used. To better distinguish between used and unused data, we divide the raw time-oriented data into two lists. The upper list represents the unused raw data, the lower list contains the time series already tested by the user. Thus, all time series in the lower list necessarily need to be colored with light gray color values.

We take a closer look at the functionality of the guidance concept. The guidance concept is based on a model-based statistical descriptor for assessing the distance between individual time series (cf. Section 4.2). With the statistical model, we estimate the dissimilarity of unused raw data in contrast to the ones already visualized and tested. We define three requirements for the statistical model.

1. Robustness to low quality raw time-oriented data
2. Value and shape-based data discrimination
3. Low redundancy between the model features

In the style of statistical time series descriptors (cf. Section 4.2), we use a combined model based on (a) two statistical properties that reflect the values of a distinct time series (median, and standard deviation) and (b) three properties based on the decomposition of time series (trend, periodicity, and noise). For the calculation of the periodicity (seasonality), we apply the autocorrelation of time series (the cross-correlation with itself). With the median and the standard deviation, we provide two robust features for preserving the value domain of a time series. The shape of a time series is represented by the decomposition of the time series. All five features can be extracted from raw time-oriented data, possibly showing quality leaks. To unify the five dimensions of the feature space, we apply a min-max normalization on every dimension. As a consequence, every primary data object can be assigned to a particular point in the feature space. The final output of the guidance model is a list of tuples. For every time series the distance to the nearest used

time series is returned. We use the unweighted Euclidean distance to define the distance between the points. As a result, in the Raw Data Selection View the bar with the maximum distance is colored black since the corresponding time series provides the maximum distance to the previously used time series. The guidance concept supports users in the choice of most dissimilar time series to test the pipeline. In this way, the risk of overfitting the pipeline is reduced by a means of assessing the workflow generalizability $\mathbf{RG}_{\text{CBA7}}$.

4.3.5. Detail View

In the *Detail View* at the center of the system, users can trace the workflow creation and execution (cf. Figure 4.4 (4)). The Detail View enables users to analyze intermediate preprocessing results by providing direct feedback for selected preprocessing routines. The output of the current model is compared with initial raw data (upper), the output of the last model (middle), and to alternative parameterizations of the current model (bottom). The linechart visualization described in Section 4.3.1 represents raw time-oriented data and facilitates visual comparison tasks. In this way, the system accomplishes the comparison of input and output data for any given model $\mathbf{RG}_{\text{CBA2}}$. Similarly, the ability to compare different outputs supports users in choosing appropriate parameter values $\mathbf{RG}_{\text{CBA3}}$.

In the example shown in Figure 4.4 a Piecewise Aggregate Aggregation (PAA) descriptor (cf. the green box in the Workflow View) is tested. A moving average model is the upstream *PreprocessingRoutine* in the pipeline (cf. the blue box in the Workflow View) before the PAA descriptor transforms the time-oriented data into the feature space. In the upper and in the center linechart of the Detail View, a moving average routine is compared with the initial raw time-oriented data (upper) and the output of the last routine (center). In the time series visualization at the bottom, the bundling technique is applied to be able to compare different parameter values. The output of the PAA descriptor is shown for different parameters (kernel intervals from 20 minutes to 8 hours). The current kernel interval of 1 hour (black curve) produces a time series representation with a shape similar to the output of the moving average routine. Orange and brown line charts indicate parameter values that cause a deviation of the model output from the model input. This example demonstrates how the output comparison supports users in the choice of parameter values. Users are able to assess whether a parameterization induces a meaningful trade-off between compactness and fidelity $\mathbf{RG}_{\text{CBA5}}$.

A second example is shown in Figure 4.7. The user analyzes the modifications made by a noise keeper routine. In contrast to data cleansing strategies where noise is usually reduced, in this example a moving average is subtracted from the time series to preserve the noise. The result of the previous routine is shown in a gray linechart, the output of the noise keeper routine is visualized with black linecharts in both visualizations. The lower visualization additionally shows six alternative parameterizations each with different effects on the time series modification.

4.3.6. Statistics View

Finally, we present the *Statistics View* at the lower right of the system (cf. Figure 4.4 (5)). The Statistics View enables users to trace the effects of the models of the workflow on the underlying time series. Varieties of statistical properties of time series are visualized in a tabular layout. Every line represents the progression of a statistical property in the course of the workflow. Three columns are provided showing the statistical values of the initial raw time-oriented data t(0), the previous preprocessing routine t(n-1), and the statistical values of the current model output t(n). The Statistics View highlights changes in the statistical values with red color. In this way, users can lookup modifications on statistical properties in an intuitive way, which improves both the effectiveness and the efficiency of the workflow analysis process. The quantitative assessment of statistical properties provided with the Statistics View supports the model input-output comparison $\mathbf{RG}_{\text{CBA2}}$, the choice of appropriate parameter values $\mathbf{RG}_{\text{CBA3}}$, the improvement of data quality $\mathbf{RG}_{\text{CBA4}}$, and the trade-off between compactness and fidelity $\mathbf{RG}_{\text{CBA5}}$. As an example of $\mathbf{RG}_{\text{CBA5}}$, we refer to the statistical property 'QUALITY_COMPRESSION_RATE' in the Statistics View presented in Figure 4.4. The property assesses the ratio of time-value pairs remaining in the process in comparison to the original number of pairs. For the currently selected model (the green PAA descriptor), the ratio decreased to 0.0167 meaning that the number of time-value pairs of the timeseries was reduced to approximately 1% of the original size. While this assesses the compactness of the time series representation, the Detail View also confirms a high fidelity of the result. This demonstrates that the PAA descriptor with the current parameterization produces a FV representing a given time series in a compact and faithful way $\mathbf{RG}_{\text{CBA7}}$.

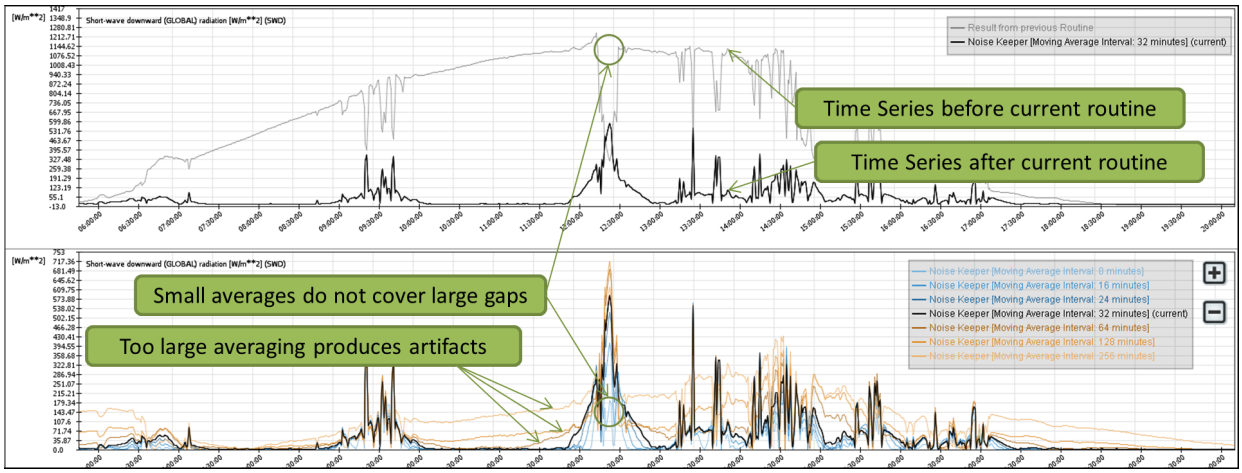


Figure 4.7 *Illustrative example of model input-output comparison and parameter guidance. The input time series is shown at the top (gray), the output time series is shown in black. A noise keeper routine is integrated in the pipeline. The model subtracts a moving average from the time series. Thus, only the noise of the time series remains. At the bottom seven different parameterizations of the model can be compared. Currently a 32 minutes kernel parameter is selected (black time series), smaller and higher parameter values are encoded with different blue and brown colors.*

4.3.7. Interaction Techniques

We have already highlighted interaction designs of individual views. In this section, we describe the interaction techniques of the system for the construction and modification of workflows at a glance. The principal steps of the pipeline (cf. Figure 4.3) serve as a baseline for the visual workflow representation shown in the Workflow View. Users can interactively add preprocessing routines to the pipeline, change parameters, modify the order of routines, and delete obsolete routines as necessary. This process can be carried out by data scientists, by domain experts, or in collaborative approaches $\mathbf{RG}_{\text{CBA9}}$. The Detail View enables users to track the effects of models $\mathbf{RG}_{\text{CBA2}}$ and alternative parameter values $\mathbf{RG}_{\text{CBA3}}$. As a result, the visual-interactive system provides means of improving the data quality $\mathbf{RG}_{\text{CBA4}}$ and managing the trade-off between compact and yet precise time series representations $\mathbf{RG}_{\text{CBA5}}$. The Statistics View provides additional support by monitoring quantitative statistical properties of time series. With the guided selection of unused raw time-oriented data, users can efficiently test the workflow with heterogeneous time series, and thus foster the generalizability $\mathbf{RG}_{\text{CBA7}}$. When the user has finished the workflow construction, the workflow is ready for execution. With the definition of a FV and a distance measure, downstream analytical models, such as data aggregation and search algorithms, can be addressed. Both the creation of time series FVs and the definition of distance measures are provided systematically with visual-interactive techniques. In this way, users are able to construct powerful workflows without the need of programming or batch scripting. This also achieves the research goal of supporting users in the definition of similarity functions based on their notion of similarity $\mathbf{RG}_{\text{CBA6}}$. At the lower right of the system, we provide a control that triggers the storage of the workflow, e.g., for future execution. The workflow can subsequently be executed as a batch process. In addition, the system allows users to load workflows for a reuse and a revision $\mathbf{RG}_{\text{CBA8}}$. This is especially important if the data is subject to change, e.g., if the data set is extended or if the external time of the data is dynamic (cf. Section 2.3.1). Loading workflows enables domain experts to make changes to existing workflows, or to use old workflows as templates.

4.4. Usage Scenario

4.4.1. Data and Domain Characterization

In Section 7.1, we present VisInfo, a DL system for the content-based access to time-oriented primary data gathered in the field of Earth observation. Within the case study, we created a set of different preprocessing workflows for time-oriented primary data. The overall goal of the workflows was to provide a means of content-based access to facilitate data-driven research in the Earth observation domain. The visual-interactive system for time series preprocessing

presented in this chapter was used for the workflow creation in VisInfo. In the following, we present the results of the workflow construction phase, conducted in a collaborative effort together with digital librarians and Earth observation scientists. For an in-depth overview of the VisInfo case study, we refer to Section 7.1. Particularly relevant for the construction of the preprocessing workflows was a data, task, and domain characterization carried out in an early phase of the design study (cf. Sections 7.1.1, 7.1.1, and 7.1.1).

We provide a brief introduction to the characteristics of the time-oriented primary data of the usage scenario. The data consists of different radiation-based measurements recorded at 55 stations all over the globe. The main quantization of the time-oriented data is one minute, however, the data also contains heterogeneous and occasionally much longer quantizations. The measurements are recorded over 21 years yielding an overall number of about 500,000,000 measurements for every provided parameter. The parameters factored in VisInfo are the *Shortwave-Downward radiation (SWD)* and the *2 meter air temperature*. The SWD measurements are most relevant for Earth observation scientists, while the temperature measurements are a plausible parameter enabling an intuitive access to the data content even for non-experts. For both parameters, we carried out individual workflows, each of which we will call a 'scenario' in the VisInfo system. Together with the domain experts, we carried out a collaborative study for the creation of a time series preprocessing pipeline and for the definition of the similarity function for the time-oriented primary data. Most important was the discussion and communication of different notions of similarity based on expert knowledge and user preference. Together with the meta data the disc space of the input data allocates the size of above 20 gigabytes. This is why data compression was an important objective in every scenario of the usage scenario. The overall goal of the workflow was the calculation of compact but yet precise time series representations (FVs), as well as similarity functions that reflect the notion of similarity suggested by the domain experts. Moreover, we aimed at providing workflow configurations which could be generalized for the entire data repository since the size of the document collection made it impossible to test the workflow for every single document.

4.4.2. Construction of a Time Series Preprocessing Workflow

While accessing the primary data source, we focused on two relevant parameters, the SWD and the temperature. We started the construction of the preprocessing workflow by adding a *Metadata Filter* routine to the pipeline. The routine removes time series not matching the required filter criteria in the metadata attribute *Physical Unit*. Together with the domain experts, we made the decision to branch the preprocessing workflow for the SWD and the temperature. The rationales rely on the value domains of both data sets which differ significantly, as well as the different intents domain experts focus on the two data sources. As an advantage, individual properties in the value domain of the two data sources can be considered individually in a best possible manner, leading to two different scenarios for the VisInfo system. In the VisInfo ESS, users have the means to choose the desired scenario when using the system.

To become familiar with the value domain of the two data sources, we carried out statistical analyses of the raw time-oriented primary data. We expected to gain an overview of the data quality which was necessary for developing a data cleansing strategy. The value domain of the temperature measurements ranges from -85.2°C (actually measured at the south pole) to +224.5°C which in all probability is an implausible value. The domain experts characterized the highest temperatures of all measurement stations at a level of +60°C in desert regions. Similarly, the value domain of the SWD raw data is between -72 W/m^2 and 3,627 W/m^2 even if the realistic value domain is between 0 W/m^2 and 1,200 W/m^2. To this end, we applied data cleansing for both value domains of the time-oriented primary data. The dominating quantization of the time series is one minute (over 91% of all time intervals), the ratio of missing values is below 10%. The salient periodicity provided by many primary data has an interval length of one day. This periodicity was expected by the domain experts. Both the SWD and the temperature measurements depend on the sun, and thus should show a daily-recurring periodic behavior called diurnal variation by the Earth observation experts. In addition, annual periodic behavior should be recognizable in the data set, as confirmed by the domain experts. Together with the experts, we decided to focus on daily patterns which should be more diverse and manifold compared to annual patterns. As a result, the VisInfo ESS will allow the exploration of daily patterns and execution of visual queries, e.g., by sketching daily curve progressions.

We aimed at resolving data quality aspects related to implausible values. Implausible values have fatal effects on downstream routines, such as normalizations and distance measures. In Figure 4.8 the result of the *Outlier Remover* routine is shown for the raw time-oriented data containing the peak value of +224.5°C. The domain experts explained that theoretically, it may be possible that measurements show variations of up to 3 standard deviations. Consequently, we calibrated the Outliner Remover routine to remove outliers above 3 standard deviations. In consequence, the data cleansing model resolves the implausible value of +224.5°C. We added further relevant data cleansing routines to the pipeline. A *Time Stamp Sorter* guarantees the consistency of the temporal domain, an indispensable requirement

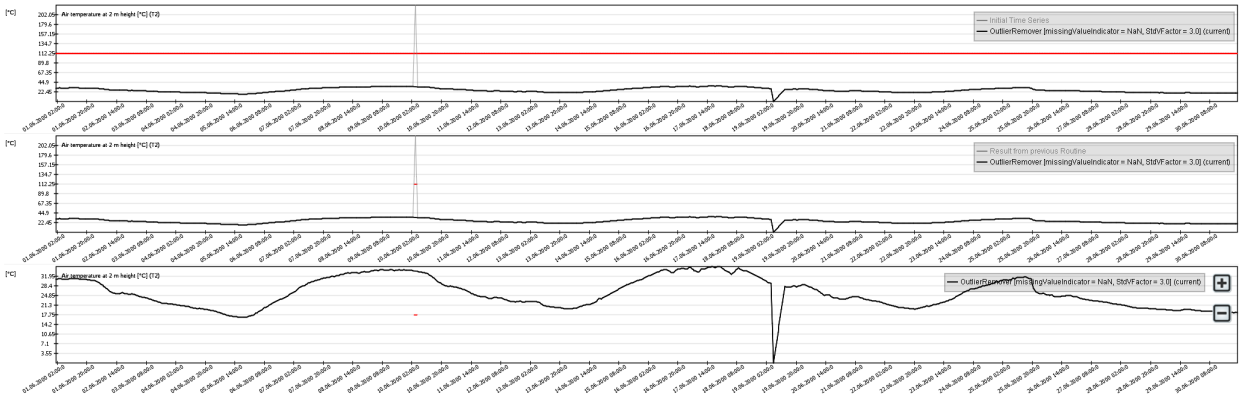


Figure 4.8 *Outlier removal of an implausible value (+224.5° C). A second implausible value (0.0° C) still remains.*

for many subsequent routines of the workflow. An *Identical Time Stamp Merger* extends the set of routines for achieving consistency of the temporal domain. A *Missing Value Remover* routine mitigates quality issues caused by missing values. We chose a variant where time-value pairs containing missing values are completely removed from the time series. While other routines for missing value handling are conceivable (e.g., interpolation-based models), we preferred to neglect the introduction of additional assumptions about the data in such an early phase of the workflow. In Figure 4.5a, a missing value interval of about five days can be identified. Such a long time period (compared with the periodicity of one day) will cause serious problems for many smarter missing value removal models.

Another concern relates to the shape of the value domain. Together with the digital librarians and the Earth observation researchers, we made the decision to focus on the general shape of the time series patterns instead of the local noisy behavior. In fact, the domain experts stated that analysis tasks exist where the noise in the measurements is of importance. As an example, solarthermic analysis requires the analysis of noise in SWD measurements. To remove the noise and preserve the overall shape of the time series, we added a *Moving Average* routine. The result of the moving average model is a smoothed curve progression. Furthermore, potentially remaining implausible values, such as the plunge down to 0.0°C in Figure 4.8 are wiped out. Figure 4.6 shows a moving average routine with the final kernel value of 1 hour. The calibration of the parameter value mitigates the noise of the underlying primary data significantly. The overall shape of the time series, however, remains broadly unchanged. As the next step, we resolved a specific problem of the provided primary data source, i.e., the temporal domain is based on Greenwich Mean Time (GMT). As a consequence, measurements taken at other longitudes on the globe have an offset between the 'true local time' and GMT. As an example, daily measurements taken from Australia starting at midnight are represented by a temporal domain starting roughly at noon time. To remove this temporal offset, we applied a *True Local Time Normalizer* routine which makes the temporal domain of all stations around the globe comparable. It is obvious that we implemented this routine specifically for the usage scenario. For this purpose, we received a conversion rule from the domain experts based on the provided metadata (longitude and latitude). The temporal conversion is parameter-free and has an accuracy of at least $+-5min$.

As a result, we achieved a robust representation of the raw time-oriented primary data. Both the temporal and the value domain of the time series is cleansed from general and rather specific quality leaks contained in the raw data. In addition, we ensured that the constructed data cleansing workflow is generalizable for the great variety of value progressions hidden in the 6,813 monthly time-oriented primary data provided with the data source. On this basis, we were able to add enhanced downstream steps in the workflow, such as the definition of time series similarity. A remaining shortcoming of the current time series representation is that the temporal domain may contain different quantizations and temporal gaps. This will be of concern in the following definition of a time series descriptor.

4.4.3. Time Series Similarity Definition

We next discuss the definition of the user-centered time series similarity function. The similarity function is applied by downstream steps, such as the retrieval and the data aggregation component. At first, we provided routines to prepare the time series for the feature extraction. Next, we added a time series descriptor to the pipeline. Based on the specification of FVs, we finally defined a distance measure.

Together with the domain experts, we discussed how to develop a compact and still precise representation of the time-oriented data. Based on the recommendation of the experts, an aggregation of the time series to a temporal resolution of one hour generates a sufficiently accurate representation of the primary data. The current quantization of most time series measurements, however, is one minute. Far more serious, some time series still contain gaps of arbitrary length. Some measurements were incomplete for a length of days, such as the time series shown in Figure 4.5a. The first challenge concerned the creation of a meaningful format of the temporal domain which could subsequently serve as the input for the descriptor transformation. To this end, we added a *Linear Interpolation* routine to the pipeline. An interpolation model was chosen providing two parameters. The first parameter is the targeted quantization, which is one hour in our case. This is the temporal resolution that a time series should at least have when it is transformed with the descriptor. Quantizations with smaller time intervals than one hour are untouched. The second parameter regards the maximum length of a gap in the temporal domain. If the time interval of a gap is longer than the parameter threshold, no linear interpolation will be calculated. Together with the domain experts, we decided that gaps of more than four hours should not be the basis for a linear interpolation scheme. For the next step, we segmented the times to the targeted pattern length of one day. A time series *Segmentation* routine cuts the monthly time-oriented data to daily patterns starting at midnight. The latter parameter value was a suggestion by the domain experts. In consequence, the number of 6,813 documents with time-oriented data content yields a sum of roughly 200,000 time series patterns for both the temperature and the SWD measurements. We agreed with the domain experts that daily patterns containing too sizable gaps should not be part of the search space. As a consequence, we removed daily patterns with temporal gaps above 4 hours. In our respective publication, this routine is called *UnderSampledDataRemover* [BRG*12].

The choice of a time series descriptor was made in favor of the *Piecewise Aggregate Approximation* (PAA) (cf. Section 4.2). The result of the parameter determination phase is shown in Figure 4.4 where the interval of one hour is compared with alternative kernel values from 20 minutes to 8 hours. As already stated by the domain experts, the one hour interval is a good approximation of the original time series while a single value for every hour also generates a compact representation (compression ratio 60:1). In combination with the previously applied moving average routine the output of the PAA descriptor is both robust and precise. The calibration of the parameter value to one hour for every aggregate is also generalizable for the entire data collection. As a result of the applied model, a FV with 24 dimensions is produced representing a daily time series pattern.

For the next step, we focused on the definition of a distance measure for the FVs. In this context, one of the most relevant results of the user-centered design process was the distinction between the absolute and the relative value domain of the daily time series patterns. Absolute values of the domain (e.g., 1,000 W/m^2) enable the comparison of different time series patterns in a global sense. It will then be possible to distinguish between high and low value progressions, e.g., to be able to compare patterns within the annual cycle. Similarly, it should be easily possible to distinguish between desert regions and polar regions. However, the domain experts unmistakably confirmed that the absolute value progression is not the only property of the primary data which discriminates different stations in a meaningful way. In addition, the relative curve progression of daily patterns should pose interesting temporal patterns. The relative value domain is achieved by an additional normalization routine applied to the FVs. The shape of every curve is transformed into the value domain between 0.0 and 1.0 meaning that every shape reaches an absolute minimum and an absolute maximum in the course of the daily progression. For the domain experts, these relative time series features enable the comparison of different intra-day volatilities, whatever their absolute values. Thus, domain experts are able to assess increases and decreases in the value domain within single days. Patterns like linear trends or bell-curves may serve as examples. For instance, we enabled domain experts to identify and compare patterns measured in maritime and in continental climates [BRS*12a]. These sort of findings are further illustrated in Chapter 5, where we show how we designed a content-based overview solution for VisInfo (see, e.g., Figure 5.18). In the conversations with the domain experts it quickly became clear that "both the absolute and the relative patterns can be important". As a consequence, we branched the two pipelines a second time and added a *Local min-max Normalization* step to the relative pattern comparison scenario. At the end, we received four different similarity scenarios for VisInfo which can be selected by the user when using the VisInfo ESS. In particular, for both parameters (SWD and temperature), we provide a preprocessing workflow for absolute and relative features.

As a final step in the visual-interactive pipeline, we added a distance measure for the downstream retrieval and data aggregation models. The decision was made in favor of the Euclidean distance since it is fast, parameter-free, and well accepted by the domain experts (e.g., for facilitating the maritime vs. continental scenario). We also introduced the domain experts to the concept of warping-based distance measures (cf. Section 4.2). However, the objective that in one pattern the measure at 11 a.m. may be compared with the measure at 1 p.m. in another pattern confused the domain experts. Even if it might make sense in some cases, the domain experts suggested that a bin-to-bin comparison

method will be more generalizable. It is needless to say that warping-based distances may constitute an interesting variation of the similarity definition concept. Together with other alternative models and model parameters, warping distances may be an interesting subject of future work.

4.4.4. Reflection on the Usage Scenario

The usage scenario amplified several benefits of the approach. We were able to create workflows for a large collection of time-oriented primary data together with domain experts in an efficient and effective way. A variety of criteria in the process were subject to collaborative debates. The visual interfaces of the system served as a means of communicating the effects of models and parameter values. Further, we were able to accept suggestions from the domain experts for models and parameter values, and adapt the workflow, accordingly. The usage scenario also demonstrated how different notions of similarity can be supported. At last, we created four workflows leading to four different sets of FVs and definitions of similarity (scenarios). In this regard, we also demonstrated the applicability of the approach to facilitate content-based access for a downstream ESS (VisInfo). We resolved the challenge of transforming time-oriented primary data into a compact and faithful FV representation. Another aspect regards the length of the four preprocessing pipelines. At least ten models are assembled to a preprocessing pipeline. In this way, we were able to confront various quality challenges and domain-specific issues. And still, we were able to assess the quality of the workflow for any model, and thus ensured the effectiveness of the approach.

4.5. Summary

4.5.1. Discussion

Time Series Similarity For both search and exploration activity the definition of distance measures is an important design parameter for many downstream models. While the workflow of our visual-interactive system includes the definition of similarity metrics, it currently does not support the assessment of effects of different similarity measures. Obviously, downstream approaches with visual-interactive means can be used to assess the effects of different definitions of similarity. It would be beneficial if these approaches facilitate the comparison of multiple results to assess these differences. As an alternative, a visual-interactive guidance concept for the effects of different similarity measures on time series data could be followed. Similarly, user feedback can be gathered to support the decision-making process of choosing the most appropriate similarity measures. Inspiration for this idea was one of our recent works where we presented a user-based visual-interactive similarity definition solution for mixed data objects [BSR*14].

Uncertainty Another discussion issue is dealing with uncertainty. The visual-interactive means of our system to cope with uncertainty can be further investigated. In fact, our data model for time-oriented data provides a concept for uncertainty. Our data model supports the representation of the temporal domain, the value domain, and a third, the uncertainty domain. With this domain, we are able to characterize the uncertainty of the data, but also the output of models applied to the data. However, for many models applied in time series KDD, it remains an issue of future work to support the assessment of uncertainty. From a visualization perspective, uncertainty is one of the long term challenges in general [KKEM10].

User Guidance Our novel system enables users to construct preprocessing workflows for time series data in a visual-interactive way. However, the models to be included in the workflows need to be selected by hand. A possible future work approach is a wizard guiding non-experts to most relevant preprocessing workflows without the need of defining most appropriate models in the most appropriate order with most appropriate parameters. The wizard can include combinations of high-level questions to assess the 'mental notion' of the preprocessing workflow in the heads of non-experts or domain experts. Ideally, the wizard is sensitive to a predefined data source to reduce the design space to meaningful solutions for targeted data sets.

4.5.2. Conclusion

In this chapter, we presented a novel method for the content-based access to time-oriented primary data $\mathbf{C}_{\text{CBA}}$. We divided the challenge into nine research goals for preprocessing time-oriented primary data, all of which were

addressed in the course of this chapter. Our solution is a system for the visual-interactive construction of preprocessing workflows for time-oriented primary data. Data scientists, domain experts, and collaborative research efforts are now able to compose preprocessing pipelines. The construction of preprocessing pipelines is based on a large set of individual routines drawn from an extensible library. Hence, we provide a means for the first of four steps of the reference workflow for the design and the application of ESS for time-oriented primary data (cf. Figure 4.2).

To summarize, the visual-interactive system for preprocessing of time-oriented primary data facilitates the content-access to different primary sources. The visual-interactive means and the guidance concepts support both data scientists and domain experts in the construction of meaningful preprocessing workflows. The output of the system is a set of time series FVs which can be applied by downstream algorithms. In addition, data scientists and domain experts are able to define similarity concepts for time-oriented data. With the visual-interactive system for preprocessing of time-oriented primary data, we contribute a solution for the first step of the reference workflow for the design and the application of visual-interactive interfaces relevant for ESS (cf. Figure 4.2). By mastering this step, we are now able to focus on the remaining three steps of the reference workflow. In the next chapters, we present guidelines and techniques for the design of content-based overviews in Chapter 5, and for relation seeking between data content and metadata in Chapter 6. In both cases, we present novel methods which benefit from the output of the approach presented in this chapter, i.e., a visual-interactive system for preprocessing of time-oriented primary data.

CHAPTER 5

Content-Based Overviews

> *Overview first,*
> *zoom and filter,*
> *then details-on-demand.*

The Visual Information-Seeking Mantra, by Ben Shneiderman [Shn96], *1996*

> *Analyze First -*
> *Show the Important -*
> *Zoom, Filter and Analyze Further -*
> *Details-on-Demand.*

The Visual Analytics Mantra, according to Keim et al. [KMS*08], *2008*

With the Visual Information-Seeking Mantra, Ben Shneiderman presented one of the most influential guidelines for designing visual-interactive interfaces. Meaningful content-based overviews of large data collections are a particularly appropriate starting point, not only for explorers, but also for searchers. Content-based overviews disclose structural information, such as unknown patterns or interesting relations between patterns. Based on visual-interactive overviews of the data content, users can zoom, filter and analyze further, e.g., to identify details on demand. In addition, content-based overviews support the identification and selection of meaningful example queries, and can be used in combination with dynamic query and faceted search interfaces. To summarize, content-based overviews are an essential part of powerful ESS. In this chapter, we face challenges associated with providing overviews of large and complex data collections (cf. $\mathbf{C}_{\text{CBO}}$). A careful characterization of the design space yields various influencing factors. Among the most relevant factors are the underlying data content, the requirements necessary for downstream ESS, the preferences of the involved user group, and different steps in the workflow of data transformations. Our approaches are based on two key principles. First, we show how VA can enhance the design process in an effective and efficient way. Second, our techniques support the user-centered design principle as a means of involving users in the design process. For this purpose, we divide the workflow of data transformations for content-based overviews into three mandatory steps. These are the *aggregation* of high-dimensional data, the *visual mapping* of the data aggregates, and the *layout* of the data within the overview visualization. First, we introduce novel quality-driven visual-interactive cluster analysis as a means of facilitating data aggregation. Second, we show how high-dimensional data objects can be encoded visually in a meaningful way. Finally, we depict different layout strategies to represent the visual data aggregates in the display space. We prove the usefulness of our contributions with various real-world examples. This chapter is mainly based on [BvLBS09, BvLBS10, BvLBS11a, BWS*12, BRS*12a, BRG*12, BWK*13, BDF*15, BSM*15b] and partially based on [Ber09, SBVLK09, SBM*14, SBMK14, WVZ*15].

Contents

5.1. Introduction

5.1.1. Motivation

Gaining an overview of the underlying data collection is a prerequisite for many analysis tasks. Content-based overviews (content summaries, content summary solutions, previews) visualize the search space at a glance and build a baseline for user interaction. Typical downstream interaction techniques empowered by overview visualizations are browsing, filtering, panning, zooming, and content-based visual querying. Thus, content-based overviews are most beneficial for enhanced ESS. The basic principle in the design of content-based overviews is providing visual representations of information abstracted from primary information objects, relating and displaying information objects at the screen, and manipulating information objects [GMPS00]. An effective approach for the abstraction from primary information objects is *data aggregation*, assigning the objects of possibly large data sets to groups (clusters). The visual representation of clusters is often achieved by glyph designs providing *visual mappings* of the high-dimensional information. These visual mappings can further be exploited when content-based overviews are subject to Query-by-Example interaction. Finally, relating and displaying objects at the screen can be provided by *layout* techniques mapping high-dimensional data elements (or clusters) to the 2D display space. Subject to the condition that a meaningful content-based accesses strategy is provided $\mathbf{C}_{\text{CBA}}$, data scientists have the means to create powerful content-based overviews leading to enhanced ESS. Both explorers and searchers can greatly benefit from content-based overview solutions, enabling users carrying out different information-seeking activities in an effective and efficient way.

However, the cascade of operations necessary for the design of content-based overviews entails various challenges that data scientists have to resolve. The large varieties of clustering techniques, glyph designs, and layouts yield a huge design space $\mathbf{C}_{\text{CBO}}$ $\mathbf{C}_{\text{MPC}}$. This is why the reference workflow of this thesis provides an explicit step for any of the three factors (cf. Figure 3.1). In addition, data scientists have to consider the characteristics of the underlying data $\mathbf{C}_{\text{CBA}}$ and the preferences of involved users $\mathbf{C}_{\text{UCD}}$, both having a decisive influence on whether a content-based overview will be assessed useful and usable. Important design choices in the (visual) *cluster analysis* step are the specification of the most appropriate class of clustering algorithms and the most appropriate clustering algorithm, respectively. Moreover, virtually all clustering algorithms provide parameters which need to be defined $\mathbf{C}_{\text{MPC}}$. A prominent example of such a parameter is the number of clusters a user wants to explore. At least (a) the intrinsic number of clusters in the underlying data and (b) the preference of the targeted user group influence this parameter choice. The *visual representation* of high-dimensional data aggregates yet bears other challenges. Users should have an influence on decisions which of the data aggregates' properties are most important, and should be emphasized with visual encodings in a cluster glyph. These visual encodings will enable users to identify and localize clusters efficiently. Other influencing factors for the visual mapping are perceptual rules a data scientists has to be aware of. Finally, the *layout* of the visual aggregates in the 2D space depends on the applied clustering routine, the number of elements shown in the display, and the preference of the user. Depending on the choice of layout, different interaction techniques have to be provided. Basically, most relevant classes of layout techniques are:

- data projections mapping high-dimensional input data to the display space
- force-directed layouts optimizing the positions of cluster nodes based on a set of allocation constraints
- layouts directly relying on the structure of the clustering result (e.g., a dendrogram for hierarchical clusterings)

In IV and VA research, a variety of techniques and applications have been presented for visual cluster analysis, glyph designs, and layouts. However, only few approaches address design choices for all three factors explicitly. In particular, for time-oriented data the number of user-centered design study approaches is scarce. In other words, a shortcoming of many existing approaches is a missing justification of the design choices made for clustering, glyph design, and layouts. These justifications are required for the generalization of results and for the reflection of the design process (cf. Section 2.5). In addition, many approaches for visual cluster analysis fall short of incorporating the quality of the obtained clustering results. However, quality plays an important role for cluster analysis approaches as a means of comparing different clustering results and selecting meaningful candidates. And still, existing search systems only make limited use of content-based overviews, especially approaches for non-textual data content. In this connection, the visual-interactive formulation of content-based queries *by example* poses a promising enhancement of classical search activity by means of content-based overviews.

5.1.2. Research Goals

The related work revealed six major research challenges (cf. Section 2.6) which we will address in the course of this thesis. In this chapter, we explicitly resolve the challenge of designing content-based overviews $\mathbf{C}_{\text{CBO}}$. In addition, we face the associated challenges of choosing appropriate models and model parameters $\mathbf{C}_{\text{MPC}}$ and the challenges of involving the user in the design $\mathbf{C}_{\text{UCD}}$. To this aim, we reconsider the content-based access (to time-oriented primary data) $\mathbf{C}_{\text{CBA}}$ as an upstream challenge. Based on a review of the involved challenges, we postulate eight *research goals* for the design of content-based overviews.

RG_{CBO1} Gaining an Understanding of the Underlying Data Set Not only for the *application* of content-based overviews but also for its *design* it is crucial to gain an overview of the underlying data content. Thus, for the design of meaningful content-based overviews an overview of the data content is necessary. In other words, making sense of the data in the design phase is to some degree a hen-and-egg problem. Hence, especially for complex data objects it is indispensable to apply meaningful content-based access strategies $\mathbf{C}_{\text{CBA}}$. The result of the contributions presented in Chapter 4 is a shift from coping with raw time-oriented primary data $\mathbf{C}_{\text{CBA}}$ to using high-dimensional time series FV in a meaningful way. A particular research goal of this chapter is gaining an understanding of collections of FVs to design meaningful content-based overviews. Understanding these data representations helps data scientists to define and implement appropriate decisions. Similarly, gaining an in-depth understanding of the data characteristics complements the requirements posed by the involved users [PVW09] $\mathbf{C}_{\text{UCD}}$. Preliminary content-based overview solutions may only be based on hypotheses and user requirements but not on the awareness of the structures of the data collection. Multiple iterations may be necessary to identify and justify appropriate design choices. Enhanced content-based overviews may be based on further research in iterative design concepts including feedback loops, in a tighter integration of visualization, or in VA concepts, such as quality assessment and user guidance.

RG_{CBO2} Choice of an Appropriate Clustering Algorithm Research in DM has produced a variety of clustering algorithms, each with specific characteristics posing different strengths and weaknesses [JMF99]. The research community has come to the agreement that "there is no best clustering algorithm" per se [Jai10]. On the contrary, the question of the most appropriate clustering algorithm highly depends on the data, the user, and the tasks [MA14] and has to be considered by the data scientist. The choice of a clustering algorithm is at heart of the overarching challenge of selecting appropriate models $\mathbf{C}_{\text{MPC}}$. As a general rule, the *quality* of clustering algorithms can be assessed, allowing data scientists to make deliberated decisions. Challenges refer to the (visual) quality assessment of clustering results and to techniques required to make different clustering results comparable [SS02]. In this connection, visual cluster analysis including guidance concepts is greatly beneficial to support the decision-making process. In addition, the visualization intents of data scientists and domain experts affect the choice of clustering algorithms. Some algorithms require an additional layout strategy for the visualization of the results. Other algorithms provide hierarchical and network-like structures which can be visualized directly. Taking interaction designs and other user preferences into account, experienced users may want to define the level of data aggregation visual-interactively $\mathbf{C}_{\text{UCD}}$. Other user groups may prefer a simplistic interaction design based on static cluster visualizations. The latter challenges may be characterized as the "need to achieve a tighter integration between clustering algorithms and the application needs" [Jai10]

RG_{CBO3} Choice of Appropriate Model Parameters The choice of meaningful parameter values constitutes another research goal in the design of content-based overviews. Together with the choice of appropriate algorithms, it poses one of the overarching challenges of this thesis $\mathbf{C}_{\text{MPC}}$. Almost every clustering algorithm provides parameters which need to be calibrated for the considered use case. We have already named the number of clusters as an important example. This choice of the optimal level of abstraction depends on the intrinsic properties of the data. However, to some extent different levels of abstraction may be valid form a data scientist's perspective. From a domain expert's perspective, finding the optimal level of abstraction is a non-trivial challenge [LK06]. In many cases, the number of parameters of a clustering algorithm is not limited to the question of the level of abstraction. As an example, the Self-organizing Maps algorithm (SOM) as one of the most appropriate variants for ESS provides a set of multiple parameter values, depending on the implementation variant. A possible approach is the application of quality measures, assessing the quality of clustering results. These measures may build the basis for the visual comparison of clustering results.

RG_{CBO4} Visual Representation of High-Dimensional Data Elements and Clusters With the selection of an appropriate clustering algorithm the data collection can be aggregated to a set of high-dimensional clusters. A downstream

research goal is the definition of meaningful visual representations for high-dimensional data and data aggregates. Challenging tasks in this visual mapping step are the choice of most descriptive data attributes and choice of visual variables. In an ideal case, the visual representations reflect the pairwise similarities of the high-dimensional data, e.g., based on the users' notion of similarity $\mathbf{C}_{\text{UCD}}$. In many cases, glyph designs are carried out for the visualization of high-dimensional data elements [BKC*13]. Meaningful design choices enable domain experts to execute elementary tasks, such as the identification, the localization, and the comparison of single data elements [AA06]. However, the visual representation of *clusters* poses additional challenges since clusters consist of multiple data elements. Cluster glyph designs should at least approximate the number of elements and the variance of descriptive data attributes. In this way, glyph designs also facilitate synoptic tasks, such as comparison and relation-seeking tasks [AA06].

RG_{CBO5} Choice of Layout Technique The layout of multiple data elements or clusters in 2D is a non-trivial research goal in its own. Different quality and design criteria exist, all influencing each other. A challenging task for aligning multivariate data in 2D is the *preservation of the structure* of the input space. Quality criteria, such as the preservation of pairwise distances of data elements can be used, often based on a predefined notion of similarity $\mathbf{C}_{\text{UCD}}$. Structure-preserving layouts enable the visual representation of data elements in a topological order, similar to a map or a landscape metaphor, sometimes called data-landscape. The class of *projection-based* algorithms is particularly suited for layouts with an emphasis on topology preservation. Another, often contrasting quality criterion is the avoidance of overplotting. Overplotting occurs if (similar) data elements overlap each other in the display. To avoid overplotting, the class of *force-directed* layout algorithms (spring layouts) can be applied. Finally, some data aggregation methods offer *clusters with structural information*, such as trees or networks, may be used in individual layout solutions. Many layout techniques pose additional parameters increasing the design space $\mathbf{C}_{\text{MPC}}$. In addition, factors such as user preferences, interaction designs, and the display size influence the choice of layout techniques $\mathbf{C}_{\text{UCD}}$. To conclude, the underlying data aggregation method, the trade-off between different quality criteria, and other use case-specific requirements constitute the choice of layout techniques to be a challenging research goal.

RG_{CBO6} Different Levels of Abstraction In the course of the information-seeking process users are interested in different levels of detail [Shn96]. While specific details of the data are often presented individual views on demand, content-based overviews can also support users in providing different levels of abstraction [CKB09]. In this connection, ESS can benefit from concepts and techniques presented in IV and VA (cf. Chapter 3). Important factors of the design are the degree of interactivity of the ESS, the preference of the user $\mathbf{C}_{\text{UCD}}$, and the flexibility of the underlying data aggregation model $\mathbf{C}_{\text{MPC}}$. At a core level, the assessment of related works yields four technical branches of providing different levels of data abstraction.

- Interactive steering of the clustering algorithm to change the aggregation level [EF10]
- Zooming to enable the interactive drill-down to local aspects within the layout [CKB09, EF10]
- Semantic zooming to show more or less details of the same entity (level of detail) [CKB09]
- Overview+context and Micro-Macro views showing both abstracted and detailed information [CKB09, Tuf90]

RG_{CBO7} Content-Based Querying The formulation of content-based queries is one of the most important interactive features in ESS since it enables domain experts to search in the data content. In this chapter, we assume that the ESS approach is based on a meaningful content-based access strategy $\mathbf{C}_{\text{CBA}}$. Content-based overviews allow the identification of relevant characteristics of the search space, and thus can serve as a meaningful basis for content-based querying. A particular research goal is the visual-interactive formulation of queries to further enhance the search activity in ESS. An encouraging concept is *Query-by-Example*, which can be directly provided in combination with content-based overviews. However, enabling Query-by-Example requires meaningful interaction designs, as well as connecting the content-based overview with the retrieval component of the system. In addition, individual views need to be provided by the ESS to visually represent the query interface. Finally, the design of highly interactive interfaces should be in accordance with the engaged user group $\mathbf{C}_{\text{CBA}}$. In the context of content-based overviews, facilitating Query-by-Example is a considerable challenge.

RG_{CBO8} Involving the User in the Design Process An overarching research goal regards the involvement of users in the design process $\mathbf{C}_{\text{UCD}}$. In fact, it is related to rather technical research goals presented earlier. In general, the involvement of users is an active subject to research in IV and VA, e.g., in design study approaches (cf. Section 2.5). Some inspiring design studies have been conducted providing insight in domain characterizations, system designs,

summative evaluations, and reflections on the results. However, methodological support for the involvement of users in the design of content-based overviews for ESS is missing to high degrees, in particular for time-oriented primary data. The design study of the LiveRAC system may serve as a positive example [MMKN08] for multivariate time-oriented data. Since best-practice experiences are missing, the automatic construction of such works without user-involvement is hardly feasible. On the contrary, visual-interactive interfaces for the design of the most mandatory steps within the workflow may be beneficial for involving users in the design process. The active role of users facilitates the model and parameter selection $\mathbf{C}_{\text{MPC}}$, as well as the visualization and interaction design based on preference information and expert feedback $\mathbf{C}_{\text{UCD}}$. Hence, content-based overviews may become usable and useful.

5.1.3. Contribution

In this chapter, we present guidelines and techniques for the design of content-based overviews. Figure 5.1 shows the three main contributions with respect to the reference workflow of this thesis. We show how visual-interactive cluster analysis can facilitate the design process in the *data aggregation* step. In particular, we contribute techniques for quality-driven visual-interactive cluster analysis. For the *visual mapping* step, we show how high-dimensional data elements and data aggregates can be represented visually. For this purpose, we emphasize the user-centered design of cluster glyphs. In this connection, we make use of color as a similarity-preserving visual variable for high-dimensional data objects, which is greatly beneficial for linking objects in multiple views. We show how the *view transformation* step of the reference workflow is facilitated with layouts for aggregated data. We show how different outputs of the data aggregation step can be used for the design of content-based layouts in 2D. In addition, we present techniques to support the user with different levels of data abstraction. Finally, we bridge the gap between exploration and search support with *Query-by-Example* techniques that can directly be applied to content-based overviews.

5.1.4. Relation to the Reference Workflow

In Figure 5.1, we draw the connection between the contributions of this chapter and the reference workflow of this thesis, presented in Chapter 3. The reference workflow poses three consecutive steps which are explicitly covered by the three technical contributions of this chapter. A FV builds the input data for the process, the final visual interface is a content-based overview as a result of three major design steps. The postulated solutions for the three steps are *quality-driven visual-interactive cluster analysis*, *cluster glyph design*, and *layouts for aggregated data*.

5.1.5. Chapter Overview

Section 5.2 reviews specific techniques in the fields cluster analysis, visualization of high-dimensional data, and layouts for high-dimensional data. In Section 5.3, we present our approaches for quality-driven visual-interactive cluster analysis. Afterwards, in Section 5.4, we present guidelines and techniques for the visual mapping of high-dimensional data objects. Section 5.5, shows guidelines and techniques for the creation of layouts for aggregated data in 2D. Section 5.6 summarizes the chapter with a discussion and a conclusion.

5.2. Baseline Techniques

In the following, we discuss baseline techniques in (visual) cluster analysis, visual representations of high-dimensional data objects, and layout techniques for high-dimensional data.

5.2.1. Cluster Analysis

Overview of Clustering Techniques Data clustering can be referred to as unsupervised techniques for building meaningful groups of data. As such, clustering is most suitable for the aggregation of large data collections. For in-depth information about data clustering, we refer to the survey of Jain et al. [JMF99] and the recent work by Jain [Jai10]. Clustering can be sub-divided into several classes of algorithms. With (1) partitioning-based, (2) density-based, (3) hierarchical, and (4) neural network-based algorithms, we introduce four classes of baseline techniques for the design of content-based overviews.

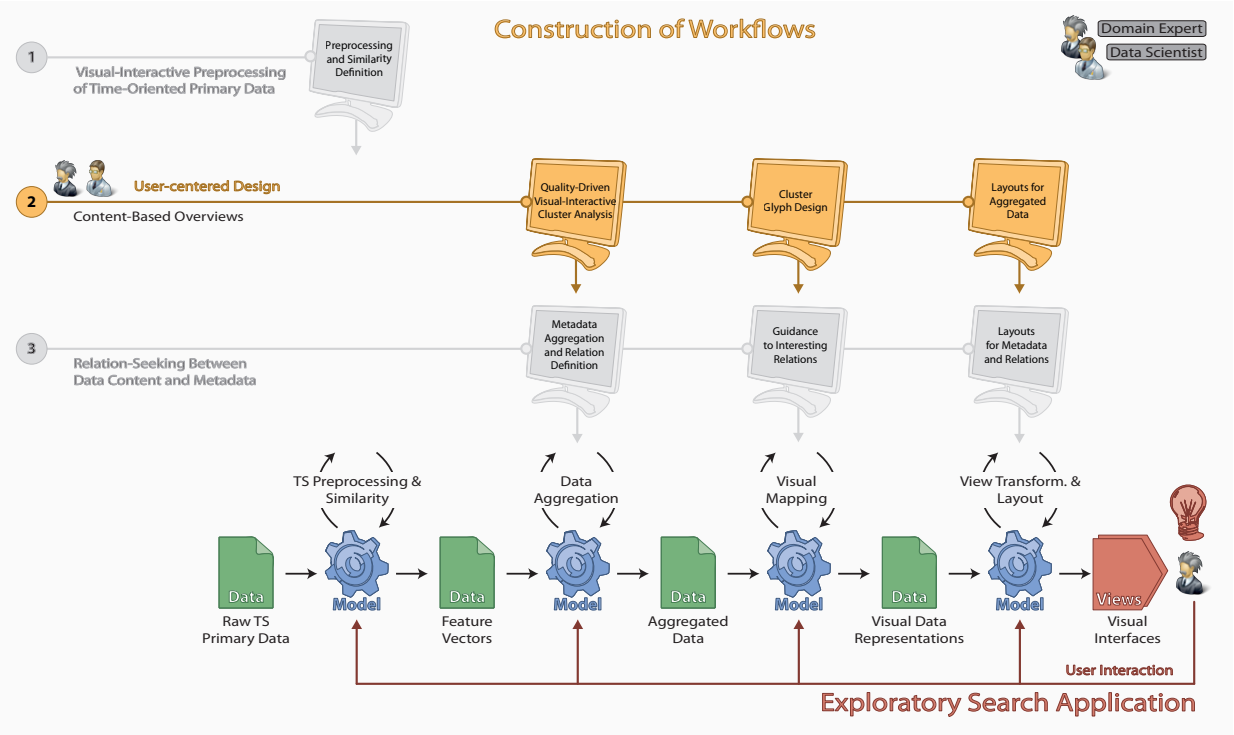


Figure 5.1 *The second technical contribution of this thesis proposes guidelines and techniques for the design of content-based overviews. The design of content-based overviews require data aggregation, visual mapping, and layout techniques, which cover the last three steps of the reference workflow.*

Partitioning-based (centroid-based) cluster algorithms divide the data set into k different clusters whereby k is an input parameter. As a result, a number of n data elements is assigned to k clusters. The choice of k with no a-priori knowledge of the intrinsic properties of the data set is one of the issues of partitioning clustering algorithms [JMF99]. A popular algorithm is the k-means algorithm presented by MacQueen [Mac67]. In every iteration, the algorithm assigns every data element e to the most proximate centroid $\mu(c)^{(t)}$ of cluster $c^{(t)}$. Every iteration is completed by an update function whereby the k centroids are recalculated. In contrary to medoids-based algorithms, the centroids of k-means clusters are necessarily members of the clusters. The update function transforms every centroid $c^{(t)}$ to a new centroid $\mu(c)^{(t+1)}$. K-means aims at solving the NP-hard optimization problem of finding k cluster centroids and assigning all objects to clusters with average minimized squared Euclidean distances. The iterative algorithm terminates after a number of i iterations, or if a (local) minimum is reached. K-means is easy to implement and with the linear complexity of $O(k*n*i)$ a fast clustering algorithm. Clusters of the k-means algorithm have a spherical shape, according to the cells in a Voronoi diagram. As a consequence, k-means lacks of identifying clusters of arbitrary shape. A further problem of the algorithm is its dependency to the initial partition. To ensure a good quality the k-means algorithm needs to be executed with different initializations (seeds). Another shortcoming of the k-means algorithm is that an adaption of k generates entirely different clustering results. Thus, if users change the 'aggregation level' in the course of the analysis (e.g., by increasing/decreasing k), all structural information about the data set becomes obsolete and additional cognitive load is required to gain an overview of new clustering results.

Density-based cluster algorithms identify dense regions of a data set. Data elements in sparsely populated regions of a data set (e.g., outliers) may remain unassigned. Density-based cluster algorithms enable the identification of dense regions of arbitrary shape. The DBScan algorithm is a prominent representative for the class of density-based clustering algorithms [Jai10]. Typical algorithm parameters are the minimum number of elements n, qualifying dense region a cluster, and epsilon ε which defines the proximity constraint for a dense region. These two parameters for the definition of density pose one of the drawbacks of density-based clustering algorithms. For a given parameter set it is (a) not possible to predict the number of clusters and (b) not guaranteed that any clusters are returned at all. Different strategies have been pursued to overcome the latter drawbacks, e.g., by methods based on parameter sampling, or by implementations only requiring a single parameter for density-based algorithms.

Hierarchical clustering algorithms calculate tree-based cluster structures. Two different types of hierarchical clustering algorithms exist. *Divisive* (top down) approaches first assign the entire data set to a single cluster node (root) and then split the nodes, successively. Different criteria for the splitting order and the splitting algorithm exist. In a recursive splitting order, nodes of every sub-tree are traversed and split independently of the remaining tree leading to a complexity of $O(n^2)$. *Agglomerative* clustering (bottom up) approaches start with n leaf clusters each containing a single data element. Within the process, cluster nodes are linked by a specific criterion. Reaching a single remaining root node terminates the algorithm. Popular linkage criteria are based on single linkage, or complete linkage calculations executable with a complexity of at least $O(n^2)$. As a general rule, nodes within the tree structure of hierarchical clustering results contain nested (sub-)groupings of a data subset which are expanded in respective subtrees. A frequently used visual representation of hierarchical clustering results is the dendrogram (see, e.g., [SS02]). A beneficial property of hierarchical structures is that they provide an inherent level-of-detail concept. If a user aims at changing the number of clusters k, it is possible to apply local split/merge operations and leave the remaining cluster hierarchy unchanged. Thus, the structure of the clustering result is preserved while a local cluster is adapted. Traversing the tree-based cluster structure can be performed by the user (e.g., by local drill-down) possibly leading to non-evenly balanced tree representations. As an alternative, changing the number of clusters can also be provided by an automated routine. In this case, the user changes k, and the system decides which clusters are split/merged by some quality criterion, leading to even-balanced tree representations. The MotionExplorer ESS presented in Section 7.2 uses the benefits of this seamless level-of-detail adaption. A drawback of hierarchical cluster algorithms is the complexity in the calculation, e.g., as opposed to partitioning-based clustering. Yet another shortcoming of hierarchical clusterings is that calculated cluster hierarchies cannot be used for a recalculation.

Artificial neural network-based clustering algorithms are motivated by biological neural networks. The network learns the information contained in the data set by grouping similar patterns to single neurons (units, cells, clusters, nodes). As a result, a high-dimensional input data set is associated with a low-dimensional network of output nodes. In the learning phase, the weights between the input and the output space are changed subsequently. The algorithms terminate once a termination criterion is satisfied. Termination criteria may be a number of i iterations, or reaching a (local) minimum. The structure of the low-dimensional output network is beneficial for cluster visualization. In contrast to many partitioning-based algorithms, artificial neural network approaches are difficult to parameterize. The Self-organizing Maps algorithm (SOM) [KSH01] may serve as a most prominent representative, having at least five parameters. In the basic implementation the SOM algorithm has a 2D output network, which can be visualized to facilitate visual cluster analysis. Different techniques exist for the initialization of the SOM network. A best-practice approach is the initialization with an upstream PCA [Jol02] execution. In the same way, less complex initialization variants like random sampling of data elements are used in practice. The learning phase of the SOM can be sub-divided into a re-allocation phase and a training phase for each iteration. We briefly outline the characteristics of the learning phase of SOMs, for an in-depth overview, we refer to the book of Teuvo Kohonen [KSH01]. First, every data element e is re-allocated to its most proximate node in the network, the so called Best-Matching Unit (hereafter, BMU). Second, the nodes of the network are re-trained based on the new BMU allocation. A specific property of network is that not only the nodes are trained with the vector information of the data elements, but also the neighbor nodes. In this way, many neural network-based approaches generate a topological ordering of the input data within the network of nodes in the output space. In consequence, these neural networks provide a means for both: data clustering and data projection.

Cluster Quality Measures Cluster quality measures assess to which extent a clustering result poses meaningful groupings for a given data set. It is generally recognized that the result of an initial clustering rarely reflects the perfect solution. In fact, the cluster analysis process is a subject of iterative refinement. That is why quantitative assessments about the quality of clustering results are mandatory. The development of cluster quality measures is a separate, yet important research field related to cluster analysis. The overall goal for the use of cluster quality measures is an effective evaluation of clustering results (cluster validity), facilitating the identification of a partition that fits best for a given data set. For VA systems the results of cluster quality measures can be used to provide guidance concepts for improving the cluster quality.

Intra-cluster quality measures assess the quality of single clusters. Thus, relations in between different clusters are neglected. A typical quality goal of intra-cluster measures is the *compactness* of clusters commonly used for density-based and centroid-based clustering algorithms. Different interpretations of the compactness of a cluster exist (see, e.g., [Ber09, BvLBS11a]). A measure relevant for this thesis is the *Quantization Error* (hereafter, QE) of a cluster (see, e.g., [KSH01]). The QE assesses the average mean squared error for a cluster c based on the data elements e and the cluster centroid $\mu(c)$.

Inter-cluster quality measures focus on relations between clusters. A prominent quality goal of inter-cluster measures is the degree of *separation* of clusters. According to separation measures a clustering result has a good quality if the clusters have high vector distance to each other. Enhanced separation quality measures incorporate both cluster centroids and data elements of the clusters. As a consequence, good cluster separation is achieved if the distance between the data elements of a particular cluster to all other clusters (centroids or respective data elements) is high. Many cluster algorithms pose inter-cluster quality measures. In particular, the nodes of artificial network-based algorithms not only assess the vector distances of the high-dimensional data space, but also take the grid distances of the low-dimensional manifold into account. This is why a variety of quality measures can be used to assess the *preservation of topology* (see, e.g., [BvLBS11a]). Topology preservation is achieved if clusters of similar high-dimensional input data are also aligned close to each other in the output space. In this thesis, we use the *Topographic Error* (hereafter, TE) as a means of assessing the quality of the topology preservation. A TE occurs when for a data element the BMU and the second BMU have a grid distance $\geq k$. The parameter k is a user parameter. As a result, the measure assesses the number of violations of the topology preservation on the granularity of single data elements. A variety of cluster quality measures assess both intra and inter-cluster quality. Prominent examples are Dunn's Index and Dunn-like indices, the Davies Bouldin Index, and the modified Hubert Statistic [Ber09, BvLBS11a].

Orthogonal to intra-cluster and inter-cluster quality, a distinction can be made between global, cluster-based and data point-based quality measures. These different types of measures enable the comparison of clustering results on different levels of granularity. *Global cluster quality measures* assess the quality of a clustering result in a single value. The advantage of global quality measures is that the visualization and the interpretation of the measures is easily feasible. As such global quality measures are highly appropriate to be able to compare different clustering results. The relative context of different clustering results reveals the quality of an individual result. Inter-cluster quality measures are commonly represented as a one-value result since inter-cluster measures assess the global structure of a clustering. The Dunn's Index, the Davies Bouldin Index, and the modified Hubert Statistic can all be represented as global quality measures. Similarly, the average QE of a clustering result can be aggregated to a global quality measure. *Cluster-based quality measures* assess the quality of single clusters. A single clustering result reveals a number of quality values, typically depending on the number of clusters k. As a consequence, the quality of individual clusters of a clustering result can be compared. For the visual cluster analysis the results of cluster-based quality measures are used for the identification of local patterns. As an example, if a k-means clustering reveals clusters with significantly different compactness scores the analyst may assume that the data set consists of regions with different densities. *Data element-based cluster quality measures* pose the finest granularity of quality measures. A variety of quality measures exist revealing quality scores for every single data element. Data element-based quality measures provide beneficial quality statemens for local structures of the input space. As a result, individual data elements can be visually encoded by their relative quality. Thus, users have the means to identify these local structures of the input space. For instance, the QE measure is calculated on the granularity of data elements. In this context, we draw a connection between the QE and one of its baseline applications; as a measure for vector quantization [KSH01]. The TE measure is another example of a data element-based quality measure in the context of topology-preservation measures.

Cluster Visualization Applications We review clustering approaches using IV techniques for the representation of clustering results. The TopicNets approach applies the Latent Dirichlet Allocation (LDA) algorithm for the calculation of topic models for large corpora of text documents [GOB*12]. The LDA algorithm assigns documents to topics by means of a probability distribution function. As such, LDA is a so-called fuzzy clustering algorithm, which is a subgroup of partitioning-based algorithms. The clustering results serve as the basis for different graph-based overview visualizations, some of which provide information drill-down interactions. A survey of hierarchical clustering and aggregation approaches in IV and VA is provided by Elmqvist and Fekete [EF10]. Prominent techniques for the visualization of hierarchical aggregates are treemaps, starplots, adjacency matrices, scatter plots, parallel coordinates, node-link diagrams, and glyph designs. Elmqvist and Fekete also present design guidelines for aggregated visualizations and suggest a basic vocabulary on interaction techniques. A best-practice approach for hierarchical cluster visualization on microarray data is presented by Seo and Shneiderman [SS02]. The Hierarchical Clustering Explorer combines dendrogram-based visualizations with multiple linked views. The system provides a variety of interaction techniques, such as dynamic querying controls to facilitate exploration tasks. Another visualization approach for hierarchical clusters is presented by Sprenger et al. [SBG00]. The H-BLOB algorithm groups cluster hierarchies at multiple levels of detail. Based on a hierarchy of implicit surfaces, clusters of different granularity are visualized in node-link diagrams. An approach for the visual organization of search results is introduced by Nocaj and Brandes [NB12]. The authors present a map-based visualization of hierarchically clustered collections of interrelated items as they exist in DLs or knowledge bases. The approach uses the Micro-Macro view metaphor presented by Tufte [Tuf90] to explore search results. In

a similar way, neural network-based approaches provide a map-based overview of the clustering result. One example is the SOM-based DL system for music collections presented by Merkl et al. [MPR02]. A hierarchical SOM algorithm is applied to structure document clusters in a topology-preserving manner. An approach stemming from visual DM is presented by Vesanto [Ves99]. Vesanto presents visual clustering techniques based on the visualization of characteristic properties of SOM results. A SOM-based approach for the visual comparison of descriptors is presented by Bremm et al. [BvLBS11b]. Based on a comparative visualization the technique supports users in the selection of different descriptors.

5.2.2. Visual Representations of High-Dimensional Data

In Section 2.1.2, we define IV and outline the Card pipeline by Card et al. [CMS99], which is described in Section 2.4.3 in detail. The *visual mappings* step in the Card pipeline maps the data into a visual form using visual variables. We review the characterization of visual variables as a set of symbols that can be applied to encode data visually. Important visual variables are *position*, *size*, *shape*, *value*, *color*, *orientation*, or *texture* [Ber83]. As an example, single numerical attributes are encoded most accurately with the visual variable *position* [CMS99], i.e., the coordinate within the x-axis and/or the y-axis. In the following, we review glyph design as a means of visually representing high-dimensional data. Afterwards, we refer to the visual variable color which will play a special role for visually encoding high-dimensional data throughout this thesis.

Encoding High-Dimensional Data Objects with Glyphs *Glyph-based visualizations* are a common form for depicting multivariate information. Data elements are transformed into a collection of visual objects referred to as glyphs [BKC*13]. Following the definition of Borgo et al. "a glyph is a small independent visual object depicting attributes of a data record" [BKC*13]. Glyphs can be placed discretely in a view independently from others. As a result, spatial relationships between data elements can be represented by the distribution of glyphs on the display and by connections drawn in between. Glyphs can be composed by various visual variables, icons, and symbols. For an in-depth survey of glyphs and glyph design guidelines, we refer to the state of the art report of Borgo et al. [BKC*13]. Visual mappings for scientific data with an emphasis on glyph design are discussed in the survey of Kehrer and Hauser [KH13].

We review best-practice approaches for glyph designs in IV and VA with an emphasis on time-oriented data. The time series bitmaps technique presented by Kumar et al. [KLK*05] visually represents the results of Keogh's SAX descriptor for time-oriented data. The SAX descriptor transforms the value domain of time-oriented data into a symbolic alphabet. The time series bitmaps visualize small sequences of the symbols in a matrix-based heatmap visualization. The resulting glyph design facilitates the visual comparison of large time series data collections. The Calendar View provides a glyph design for the alignment of daily patterns in a grid-based calendar metaphor [VWVS99]. The visual variable position is used to support the comparison of weekly patterns. An example of a complex glyph design is the Document Cards metaphor presented by Strobelt et al. [SOR*09]. Textual and image-based content of scientific publications is visually represented in a compact interactive card metaphor. Prominent terms of the documents are highlighted, different sizes indicate the term frequency. With a click on a term related images and pages of the paper are highlighted with a heatmap metaphor. The KronoMiner tool uses a radial layout for the visualization of small sets of time series [ZCB11]. The different radial segments enable the comparison of the time-oriented data.

A convincing radial glyph design for the visualization of periodical time series is the ClockMap technique [FFM12]. Color is used to map the aggregated value domain of a time series to arc segments. A hierarchical treemap layout allocates large numbers of clockmaps at the screen to support comparison tasks. In a controlled experiment Fuchs et al. [FFM*13] evaluate the performance of ClockMaps in comparison to line glyphs, stripe glyphs, and star glyphs. Three different analysis tasks related to time series analysis are evaluated. The authors assess a good performance to line glyphs for the detection of peaks and trends, while the radial layout of ClockMap is beneficial for the temporal location of single values. The DICON framework is another promising approach for the glyph-based visualization of high-dimensional data objects [CGSQ11]. Icon-based glyphs are combined with force-directed layouts to provide high-level statistical information of multi-dimensional clusters. An abstract but effective, and visually beautiful glyph design is the figurative visualization of software metrics by Fabian Fischer [Bec14]. Ten mixed data attributes extracted from classes of source code are visually encoded with characteristic properties of a contour feather. The glyph design supports discerning categories of code entities, detecting problems in the code, and studying the evolution of code. An example of the glyph-based visual aggregation of event sequences is the LifeFlow system [WGGP*11]. LifeFlow provides an overview of possibly large event sequence data sets of different event types. A visual aggregation concept groups identical event sequences, and thus structures the data set visually. Differences within the visual aggregates can be explored on demand with a temporal spacing metaphor. We conclude with a VA system making use of a

variety of glyph designs. The Small Multiple, Large Singles approach presented by van den Elzen and van Wijk provides a variety of standard visualization techniques, such as barcharts, node-link diagrams, scatter plots, or parallel coordinates [vdEvW13]. Multiples of small visualizations are provided as alternatives to the current state (large single). Color is used to link data elements in different views.

Encoding High-Dimensional Data Objects with Color *Color* is another means of depicting information about high-dimensional data. Color is one of the most accurate visual variables for linking objects in different views. Thus, using color is greatly beneficial for effective ESS with multiple linked views, showing data from different perspectives. In most cases, the visual encoding of data objects is achieved with colormaps. Colormaps assign colors to objects based on a given functional specification. In this connection, categorical data is well represented by qualitative colormaps while numerical data is encoded with quantitative colormaps (see, e.g., [TFS08]). For numerical values, data scientists need to consider whether the value range of a particular attribute is unipolar or bipolar, a property which may also be subject to user preference. Most of the varieties of existing colormaps assign univariate data to color.

High-dimensional data objects are more difficult to encode with color in a meaningful way. Approaches exist using color in a qualitative way to link clusters in different views (see, e.g., [AYMW11]). However, qualitative colormaps lack of preserving the relations between data objects, e.g., the reflection of pairwise similarities. To preserve the characteristics of the high-dimensional data set it is mandatory for colormaps to take the pairwise similarities of the high-dimensional data into account. From a coarse perspective, two criteria need to be considered. On the one hand, similar objects should receive a similar color. On the other hand, dissimilar objects items should be colored with dissimilar colors. Two types of approaches exist fulfilling this criterion. *Data-dependent* approaches optimize a transformation function for a given high-dimensional data set into a targeted color space. An overview of existing approaches and design guidelines is provided by Mittelstädt et al. [MBS*14]. The drawback of data-dependent approaches is that every data set requires an individual optimizations step. In addition, the optimization includes various models and model parameters. Finally, different color spaces exist, potentially serving as output spaces for the optimization. *Data-independent* approaches make use of static 2D colormaps. Static 2D colormaps are predefined functions mapping normalized 2D values to colors. As an upstream technique data projection can be applied to map the high-dimensional data elements into 2D. In Section 5.4.2, we present a survey of presented data-independent 2D colormaps. An overview of relevant 2D colormaps is presented in Table 5.2. The challenges of the data-independent color mapping strategy are the choice of the upstream data projection and the choice of the most appropriate static 2D colormap for a given analysis task. In this thesis, we exploit the value of static 2D colormaps in various case studies (see, e.g., the MotionExplorer case study in Section 7.2). This is why in Section 5.4.2, we present important 2D colormaps, quality considerations, and design guidelines for the use of 2D colormaps.

5.2.3. Layout Techniques for High-Dimensional Data

Layout techniques determine the positions of data elements on the display. In this way, high-dimensional data elements and clusters can be represented in 2D. Thus, layout techniques are appropriate to perform the view transformation step in the Card pipeline [CMS99] and in our reference workflow (cf. Figure 3.1). We divide layout techniques into the three different classes projection-based, force-directed, and cluster structure-based techniques.

Projection-Based Layouts The projection of high-dimensional data elements to low dimensional representations is an important subject of research in KDD and VA. Projection-based approaches are applied to facilitate multi-dimensional data visualization and exploratory data analysis [Ves99, POM07]. The process of representing the information of high-dimensional data elements with reduced dimensionality is also known as dimension reduction (see, e.g., [IMI*10] for a VA approach). Representing high-dimensional information in a lower space cannot be carried out without information loss. A variety of measures to assess the quality of projections exist each focusing on the preservation of some of the properties of high-dimensional data sets. Virtually all projection techniques focus on the preservation of at least some of the quality aspects, depending on the class of projection algorithm. Projections can be subdivided into the classes of linear and non-linear projections. *Linear projections* calculate a functional mapping of each high-dimensional data element to the low-dimensional space. As a consequence, the results of linear projections can also be applied to new data elements, or other data sets, respectively. A widely applied linear projection technique is Principal Component Analysis (PCA) [Jol02] also called discrete Karhunen-Loéve Transform (KLT) in signal processing. PCA uses an orthogonal transformation to convert the high-dimensional data space into a set of linearly uncorrelated dimensions (the principal components). The first principal component covers the largest possible variance of the data set. Each succeeding component in turn has the largest remaining variance while being uncorrelated with the preceding components. The

latter is achieved with the constraint that all components are orthogonal to each other. As a consequence, layouts based on a 2D PCA show the variance of the data set spanned by the first two principal components. A drawback of 2D PCA is that the variance information of the remaining $n-2$ output dimensions is not considered in the result. As a consequence, a visible position of the 2D output typically has different places of origin in the high-dimensional input data space. *Non-linear projections* assume that the shape of the high-dimensional data set can be represented by a non-linear manifold in a low-dimensional output space. Relevant non-linear projection techniques are the class of Multidimensional Scaling (MDS) algorithms [Kru64], or the class of neural network-based techniques, such as the Self-organizing Maps (SOM) [KSH01] algorithm. Moreover, a variety of other non-linear projection algorithms have been presented, many of which are combinations or extensions of existing techniques. Non-linear variants prefer local over global properties of the data set. As a consequence, non-linear projections distort the data space, i.e., the global distribution of data elements in the output space may not faithfully represent the distribution of the input space. As an example, the MDS algorithm optimizes a function which assigns objects with low pairwise distances close to each other in the output space.

Projection-based VA applications have been presented for a variety of data types and application fields. The Projection Explorer is a projection-based visualization tool for the exploration of high-dimensional data elements [POM07]. Color is used to depict the similarity between user-selectable data elements. The approach also facilitates clustering of the projection output. Projection Explorer provides application examples based on structured tabular and unstructured text data. The MusicGalaxy system uses a neighborhood-preserving projection to provide an overview of large collections of music tracks to facilitate exploratory music retrieval [SN11]. The visual interface (a map metaphor) supports zooming, a multi-focal zoom lens enables users to explore local details. Furthermore, MusicGalaxy provides mechanisms to explicitly address problems caused by projection errors. The approach of Steiger et al. enables experts in energy networks to detect patterns and anomalies in time-oriented data [SBM*14]. The MDS algorithm is used to represent large numbers of daily power consumption curves in 2D. In addition, the approach uses a clustering concept in combination with a static 2D colormap to emphasize prominent patterns in the data set in a similarity-preserving way.

Force-Directed Layouts *Pairwise relations* between data objects are a popular means of representing the structure of objects. Force-directed layouts are applied to represent these structures visually. The type of these objects is not relevant for force-directed layouts, as long as pairwise relations are specified in between objects. For time series data and time series clusters different types of pairwise relations exist.

A widely used type of relation is the *pairwise distance between data objects*. If the distances between any two objects a and b are symmetric ($dist(a,b) = dist(b,a)$), force-directed layouts can use the pairwise distances and generate a layout of the objects for the visual representation. For time-oriented data objects the temporal domain and the value domain is typically involved in the definition of similarity. As an example, our visual-interactive time series preprocessing techniques, presented in Chapter 4, also provide a means of defining time series similarity. An alternative strategy for revealing pairwise relations in time series data is the *comparison of very short temporal segments*, such as single time stamps. In this case, only the (multivariate) value domain is used for the calculation of similarity. The time series data is segmented to (multivariate) states of a very short duration. For instance, in case of human motion capture data analysis, the segments often consist of single poses. Thus, the relations are defined on the granularity of single segments (states) instead of time series objects at a glance. In turn, the relations of global time series may be a product of generalization concepts, e.g., provided with sequence alignment algorithms.

Layout algorithms facilitate the visual representation of pairwise relations. As such, layout algorithms are the subject of graph drawing, resulting in node-link diagrams or other graph-based visualization techniques. Force-directed layout algorithms assign any given data object a 2D coordinate in the display. Different criteria for the allocation of data items on the display exist, such as the relative item positions, the minimization of edge crossings, or the aesthetically pleasing appearance. Most algorithms optimize these criteria in an iterative manner. An iteration usually consists of the phases of force definition and force execution. The definition of forces is similar to the physical properties of springs. Attraction and repulsion forces are defined by the optimization criteria based on node or edge properties of the network. In the force execution phase, the defined forces are transmitted. An overview of force-directed layout algorithms is presented in the work of Andreas Noack [Noa07]. Noack also presents the LinLog layout algorithm, which is relevant for overview visualizations presented in this thesis (cf. Section 6.4). As a special property, the energy model of the LinLog algorithm considers criteria for graph clustering.

VA applications using force-directed layouts exist for various of application domains and data types. A survey of tools facilitating exploratory search for large text document collections is provided by Herrmannova and Knoth [HK12]. The majority of presented approaches rely on graph structures and node-link diagrams to provide content-based overviews. The FacetAtlas tool applies text mining and entity extraction to transform a set of multi-faceted medical text documents into an entity-rational data model [CSL*10]. A Kernel Density Estimation (KDE) is applied to provide

a basic layout for document clusters. A force-directed layout balances the dynamics of node visibility caused by interactive data exploration. Important criteria for the layout are the balance between the readability and stability. The TopicNets approach computes graph data structures based on text documents aggregated with the Latent Dirichlet Allocation (LDA) topic modeling algorithm [GOB*12]. A dissimilarity matrix based on the symmetric Kullback-Leibler divergence between topics provides the input for a MDS projection step. The MDS algorithm assigns every topic to a preliminarily position on the display. Finally, a force-directed layout algorithm determines the final positions for every topic. The DICON framework provides different layout techniques for the visualization of cluster glyphs [CGSQ11]. A graph-based visualization is provided based on a force-directed layout using the relations between pairwise data elements and clusters. Moreover, DICON provides an MDS-based layout technique to represent the similarity between data elements and clusters. To reduce overplotting a local force-directed layout is applied.

Cluster Structure-Based Layouts We conclude the review of layout techniques with a brief overview of cluster structure-based layouts. We use the term cluster structure-based layout to combine all visualization techniques explicitly exploiting structural properties of clustering results.

Hierarchical data aggregations play a crucial role for cluster structure-based layouts. Different visualization techniques exist utilizing the structure of hierarchical clustering results. An overview of visualization and interaction designs for hierarchical aggregation is provided by Elmqvist and Fekete [EF10]. We have already emphasized the relevance of dendrograms for the visualization of hierarchies. For instance, the exploration tool for hierarchical clustering results presented by Seo and Shneiderman makes use of dendrogram visualizations [SS02]. Different interaction techniques and linked views facilitate the information drill-down to local aspects of the hierarchical clustering result. With the H-BLOB algorithm a graph-based hierarchical layout is provided [SBG00]. Implicit surfaces show nested groupings of clusters according to the tree-structure of the hierarchical clustering result. A popular approach for hierarchical data visualization is the treemap. As an example, the Clockmap technique presented by Fischer et al. shows a circular treemap variant with ClockMap glyphs representing periodical time series patterns [FFM12]. An inspiring approach facilitating a treemap based on a Voronoi metaphor is shown by Nocaj and Brandes [NB12]. Based on a MDS-based topology documents of a search space are grouped within nested Voronoi cells.

Grid-based structures are another structural output of clustering algorithms. An important example is the Self-organizing Maps (SOM) algorithm [KSH01]. For the exploratory data analysis of high-dimensional data most SOM implementations provide a 2D regular grid structure. The grid consists of a variety of cells (clusters, nodes, units) providing individual parts of the high-dimensional input data set. As such, SOMs can be seen as a mix of vector quantization, partitioning-based clustering, and non-linear projection algorithm. A variety of visual DM approaches exist using the grid structure of the SOM output (see [Ves99] for an overview of SOM-based visualizations). In addition, VA systems have been presented facilitating information seeking, ES, and exploratory data analysis. Examples are the SOM-enhanced Jukebox by Merkl et al. using a hierarchical SOM [MPR02], the descriptor selection tool by Bremm et al. [BvLBS11b], and the visual cluster analysis system by Schreck et al. [SBVLK09].

5.3. Quality-Driven Visual-Interactive Cluster Analysis

> “ *(...) there is no single clustering algorithm that has been shown to dominate other algorithms across all application domains.* ”
>
> Anil K. Jain [Jai10], *2010*

In this section, we present VA techniques for the aggregation of high-dimensional data with cluster analysis. In particular, we present novel visualizations of cluster quality which enhance the cluster analysis process. On this basis, we apply visual-interactive cluster analysis for the user-centered design of cluster visualization applications. In turn, cluster visualizations serve as the basis for content-based overviews, e.g., for time-oriented primary data. When applied within an ESS, users can directly benefit from the exploratory nature of content-based overviews. Structural

information of the high-dimensional data content can be revealed and be used for subsequent drill-down and visual querying interactions. As fundamental prerequisites, content-based overviews should reflect the properties of the data set and comply with domain knowledge and application needs. Thus, gaining an in-depth understanding of the data set is also an essential objective in the design phase of ESS and a basis for making meaningful decisions. Data scientists have to choose cluster algorithms with parameterizations which comply with the data characteristics and the preferences of involved users. Finally, a meaningful clustering result depicts a solution for the data aggregation step of the reference workflow (cf. Figure 5.1).

We present a novel set of techniques facilitating quality-driven visual-interactive cluster analysis. To this end, we systematically characterize different notions of clustering quality and incorporate these factors into the visual clustering analysis process. We define a hierarchy of abstractions for measuring the quality of cluster analysis results. Our hierarchy of abstractions includes the cluster quality assessment based on:

- global measures, for an entire clustering result, presented in Section 5.3.2
- cluster-based measures, for individual clusters, presented in Section 5.3.3
- data element-based measures, for single data elements, presented in Section 5.3.4
- comparing the results of different clustering algorithms in Section 5.3.5

For each quality assessment strategy, we show how VA techniques can facilitate the optimization of clustering quality. With our techniques, we approach various research goals outlined for content-based overviews. First, the quality-driven visual cluster analysis process supports data scientists and domain experts in gaining an in-depth understanding of the data set $\mathbf{RG}_{\text{CBO1}}$. Second, we introduce VA techniques for the *justification of design choices* for clustering algorithms $\mathbf{RG}_{\text{CBO2}}$. Third, our techniques facilitate the calibration of algorithm (model) parameters $\mathbf{RG}_{\text{CBO3}}$. Fourth, our hierarchy of abstractions supports the analysis on different levels of abstraction $\mathbf{RG}_{\text{CBO6}}$. Finally, visual-interactive means enable the involvement of users $\mathbf{RG}_{\text{CBO8}}$. Collaborative working of data scientists and domain experts may help to capture domain knowledge and the intuition of the domain experts. The final clustering result provides the solution for the data aggregation step of the workflow (cf. Figure 5.1). It can subsequently be used as a basis for the design of downstream steps in the reference workflow for the design of content-based overviews.

The remainder of this section is as follows. In Section 5.3.1, we outline our reference implementation for semi-supervised visual-interactive cluster analysis. In Sections 5.3.2, 5.3.3, and 5.3.4, we present techniques for the visual quality assessment of clustering results on a global, a cluster-based, and a data element-based granularity. Section 5.3.5 introduces the notion of quality based on the comparison of results of different clustering algorithms.

5.3.1. Semi-Supervised Visual-Interactive Clustering

To unfold the benefits of VA, we postulate the adaption of unsupervised cluster algorithms from black-box techniques to visual-interactive semi-supervised variants. For iterative clustering algorithms, branching and *visually monitoring* of intermediate results can serve as a powerful basis for an enhanced cluster analysis process [AYMW11]. In addition, *steering* the algorithm parameters based on the analyzed intermediate results may reveal meaningful clustering results for the user. In a reference publication, we have presented a semi-supervised clustering algorithm, using the example of the Self-organizing Maps (SOM) algorithm [SBVLK09]. Please note that parts of this contribution are presented in the thesis of Tatiana von Landesberger [vA10]. This is why we refer the work to as a baseline technique for this thesis. The technique supports *initializing*, *monitoring* and *steering* the clustering process using visual-interactive means. During the clustering process, the user is able to pause the training, update parameters, and resume the process. While the fully automatic calculation may generate meaningful clustering results, we also recognized the need to more closely integrate the expert user in the clustering process (cf. research goal $\mathbf{RG}_{\text{CBO8}}$). With our reference technique users are able to incorporate domain knowledge, application needs, and user preferences. As such, the technique implements the general VA principle of combining automated analysis with human expert supervision. In general, candidates for our semi-supervised, visual-interactive clustering technique fulfill at least several of the following properties:

- **Initialization** - starting the algorithm with an initial clustering helps to resolve cold start problems and supports the reuse and refinement of clustering results
- **Parameters** - the clustering algorithm should at least provide one parameter
- **Steering** - the clustering algorithm (the parameters) should be steerable in the course of the calculation
- **Sensors** - the clustering algorithms should enable to plug in sensors for assessing intermediate results; the algorithm does not necessarily need be iterative for sensors

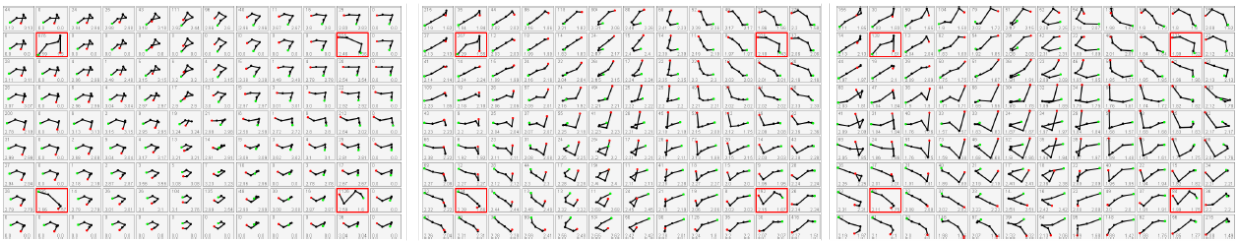


Figure 5.2 *SOM training with four predefined fixed-points (red) reused from a preceding training result. From left to right the initialization and two intermediate results of the training are shown. The topology of the map adopts the vector information of the four fixed-points.*

- **Monitoring** - the intermediate result of the algorithm and/or the optimization model should be observable by visual means; adaptions and extensions of the algorithm included
- **Quality** - the clustering algorithm should be applicable for quality measures

The SOM, as an example, fulfills all of these properties. In the reference publication the SOM algorithm has been chosen with respect to its iterative nature and its relevance in visual cluster analysis. In addition, the SOM is also one of the most challenging algorithms for data scientists, due to its non-deterministic behavior, its large number of input parameters, and its different training objectives. Most relevant objectives are revealing intrinsic properties of the data set by means of clustering, the preservation of topology, and the quantization of input vectors [KSH01]. Finally, the SOM algorithm does not provide a cost function for the optimization of the training. The training of the clustering algorithm is only based on the data set, a predefined similarity measure (cf. Chapter 4), and the parameterization. In most implementations the only termination criterion of the algorithm is the completion of the predefined number of iterations. These criteria qualify the SOM to be a suitable non-trivial candidate for a semi-supervised visual-interactive clustering variant. However, providing semi-supervised variants to facilitate the full potential of VA can be generalized to many other candidates. For non-iterative algorithms, we suggest the integration of other types of *sensors* in the calculation process. We assume that most clustering algorithms provide intermediate results which can be used for monitoring (and steering) the clustering process. Our semi-supervised variant provides visual-interactive means for the initialization, the training, and the postprocessing phase of the cluster analysis process.

Initialization For a variety of clustering algorithms the quality of the results depends on its initialization. We propose visual-interactive techniques for the initialization of the clustering algorithm. In general both the definition of the algorithm parameters and the high-dimensional vector information of the clustering can be a subject to interactive initialization. Our reference technique enables users to control the initialization process by explicitly defining individual cluster values, e.g., by sketching visual representations of cluster information. In addition, the user can use the results of earlier clustering results as a pool of example patterns. We refer to Section 5.4, where we discuss techniques for the meaningful visual representation of high-dimensional clusters. The initialization of the SOM manifold can also be achieved by the standard variants of random and PCA-based initialization. A final step in the initialization phase of the SOM regards the definition of fixed-points. By fixing a SOM cluster the vector information of the cluster is preserved within the training. Proximate clusters on the SOM grid are adapted by the fixed-point information. As a result, domain knowledge and user preference information can be assembled within the clustering process. Two examples can be seen in Figures 5.2 and 5.3, where fixed-points are defined and observed within the training for example data elements and for sketched data.

Training The training process of the clustering algorithm can be visually observed (see again Figures 5.2 and 5.3). Control functionality enables steering of the process, on demand. Within the training, users are empowered to pause the training, to update parameters, and to resume the training. In the course of the training the user can monitor the vector distribution of the cluster visualization. The visualization of intermediate results of the clustering and the corresponding data distribution reveals an early impression of the dense regions of the high-dimensional data set. An example of the observation of a SOM training can be seen in Figure 5.6. The system provides a number of steering interactions to guide the learning process towards user-desired results. For these interactions, the user is able to suspended the training process. First, the user can adjust single clusters, e.g., by using a sketch editor tool. Second, adapting training parameters is supported. For the SOM, relevant parameters are, e.g., the number of integrations,

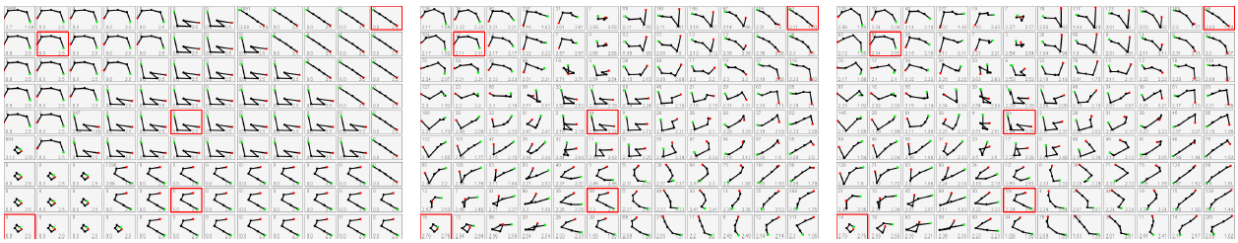


Figure 5.3 *SOM training with six abstract predefined fixed-points (red). From left to right the initialization and two intermediate results of the training are shown. The topology of the map adopts the vector information of the six fixed-points. However, the abstract means of the fixed-points compromises the quality of the clustering process.*

or the kernels for the neighborhood and learning, see Section 5.2.1 for an overview of all SOM parameters. Third, in case of the SOM, regions of the map can be adjusted by defining fixed-points in combination with interpolation mechanisms. Finally, we support reinforcing the training of selected patterns.

Postprocessing In the postprocessing phase, users can analyze the clustering results. The visual-interactive means of the technique enable an in-depth exploration of patterns contained in the data set. In addition, the technique provides interactive capability to support the postprocessing process. A relevant aspect in this context is the *iterative refinement* of clustering results as a preprocessing step for downstream clusterings. For this purpose, the user is able to merge multiple clusters, and thus adapt the level of aggregation locally. In the same way the expansion of clusters for a finer grained analysis is supported. Here, the set of the clusters, to be expanded, serves as the input for another training for an iterative refinement. Other postprocessing interactions are creating, deleting, and editing clusters. In this way, these interactions support the individualization of the clustering result. From a task-based perspective, possible examples are the emphasis of underrepresented patterns or the reduction of frequent but rather uninteresting patterns. For candidates yielding a structure, such as the grid-based SOM algorithm, the optimization of these structures may be of interest. Our reference implementation allows an interactive re-arrangement of the SOM grid as a beneficial means of facilitating the iterative refinement. In the next sections, we show how visual quality assessment strategies can foster visual-interactive semi-supervised clustering. As a result, the justification of quality can be transferred to the calculation process to support semi-supervised clustering.

5.3.2. Quality Assessment with Global Quality Measures

Characterizing the clustering quality by a scalar global measure is a simple and straightforward approach. Detailed information about available one-value measures is provided in Section 5.2.1. Deriving a scalar value for the quality of an entire clustering result is a classical approach to be able to compare *multiple clustering results*. In Section 5.3.1, we characterized properties for semi-supervised clustering approaches. Building on this, global quality measures not only facilitate the comparison of clustering results but also for the comparison of *multiple trainings* and intermediate results. Our techniques support both concepts. This enhancement of the visual cluster process is at heart of the research goal for global quality measures. Both the comparison of multiple clustering results and of multiple trainings can be seen in Figure 5.4. We provide a linechart view where the quality progressions of multiple cluster trainings can be analyzed. The x-axis represents the training progress in a relative manner, enabling the comparison of trainings with different numbers of iterations. The y-axis represents the cluster quality. By the use of colors, we are able to show different quality measures in a single display. Thus, the linechart view facilitates the comparison of both different trainings and different quality measures at once. We use color brightness to differentiate between recent (dark) and early (bright) trainings. On the right, a control panel supports the user in the selection and the control of available quality measures. A legend on the right assigns selected global quality measures to color and brightness values. In addition, the legend provides statistical information about the global minimum and maximum values of the quality measures. The example in Figure 5.4 shows two quality error measures with opposing tendency. One error measure (blue) decreases indicating an improvement in the quality in the course of the training. However, the other error measure (green) increases meaning that in this respect the quality of the cluster algorithm worsens. This example shows that data scientists typically have to consider a trade-off between different quality criteria when designing appropriate clusterings. In this example, the average QE (blue) assesses the vector quantization of the underlying clustering algorithm (here: SOM). The QE typically decreases within the SOM training since the SOM network increasingly adapts local structures of

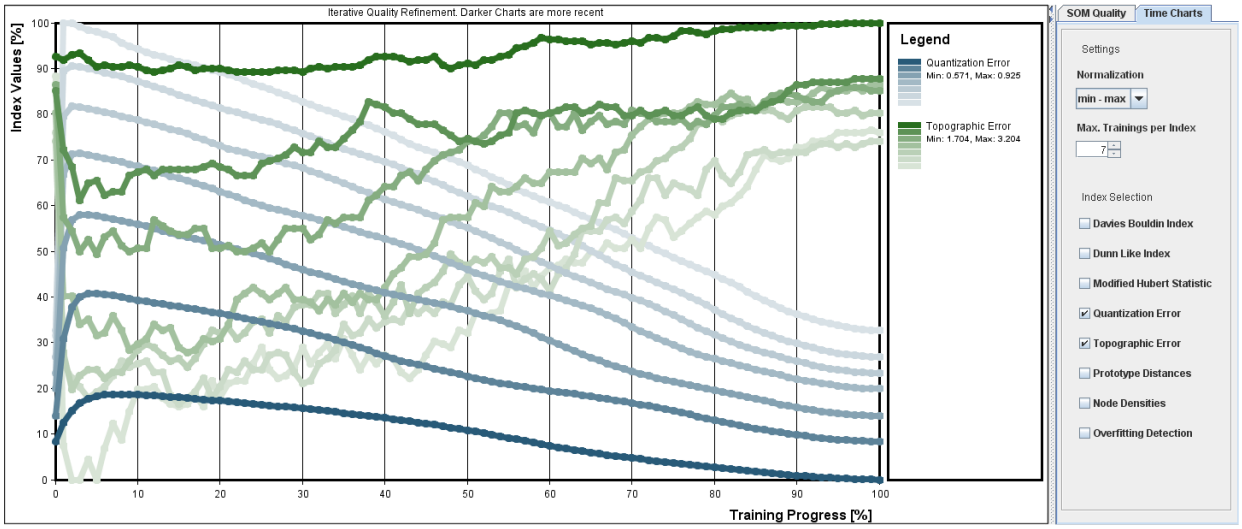


Figure 5.4 *Observation of the SOM training process. Comparison of two quality measures (green and blue) with opposing tendency. The results of seven clusterings are visualized, dark colors represent more recent results.*

the data set. The TE (green) assesses the preservation of the topology of the SOM. The TE typically increases within the SOM training since the non-linear SOM forms local structures which may violate the global structure (topology). The trade-off between vector quantization and topology preservation is of particular concern for SOMs [KSH01]. Our technique supports facing these types of design challenges visually, in a supervised way.

The visual-interactive interface also facilitates the comparison of multiple trainings of a clustering algorithm. We use superposition for the visual comparison of the individual quality progressions (linecharts). In Figure 5.4 the linecharts of seven successively executed SOM trainings are presented, the darkest linechart represents the most recent training. The user can change training parameters and observe the effects on the diverse quality criteria on-the-fly. For example, balancing the QE and the TE yielded by the respective results is provided in an efficient and effective way. This example shows how the user is able to directly assess and validate a) the global quality of different clustering results and b) the influences based on parameter changes $\mathbf{RG}_{\text{CBO3}}$. Moreover, the visual-interactive means easily support the involvement of users to discuss quality trade-offs, design aspects, and user preferences $\mathbf{RG}_{\text{CBO8}}$.

Our technique supports the observation of parameter changes. In addition, the effect of a direct manipulation of the cluster vectors by user interaction can be observed visually. Within the latter Section 5.3.1, we have also shown how the presented semi-supervised cluster analysis technique supports editing SOM clusters and for defining fixed-points. We next observe the progression of global quality measures in case when users directly manipulate the vector information of the clustering. The observation of the manipulated training in Figure 5.5 is demonstrated with two measures. The red linecharts represent an *error* measure punishing bad quality with high values (here: an overfitting detection measure [Ber09]). With the blue measure, a *quality* measure is shown assigning bad quality with low values. In total the progressions of four different trainings are shown. The measures assess similar quality results for all four clusterings. However, In the last run the cluster vector information was manipulated directly. As a result, the quality measures provide peak values at about 50% of the training progress. It can be observed that the red error measure shows an upward spike while the blue quality measure suffers a drop. Towards the end of the training, the measures indicate that the quality situation relaxes in the course of the training. However, the resulting quality values indicate significantly worse quality compared to the other three earlier SOM trainings. We briefly describe the two quality measures shown in the example. The modified Hubert statistic (blue) is an established cluster quality measure (cf. Section 5.2.1). The measure calculates the correlation between pairwise distances of the data elements with the pairwise distance of respective clusters of the data elements. With the overfitting detection measure (red), we are able to assess the relative probability that single clusters of the SOM grid suffer from overfitting effects caused by less generalizable parameterizations, or by violations of local vector information in the SOM grid [Ber09].

We summarize the technique for using global quality measures to enhance visual-interactive cluster analysis. The technique allows the visual assessment of quality implications of iterative refinement strategies. The presented examples show that the technique supports the assessment of both the adaption of training parameters and the modification of the clusters. With the technique, we achieve two research goals for the design of content-based

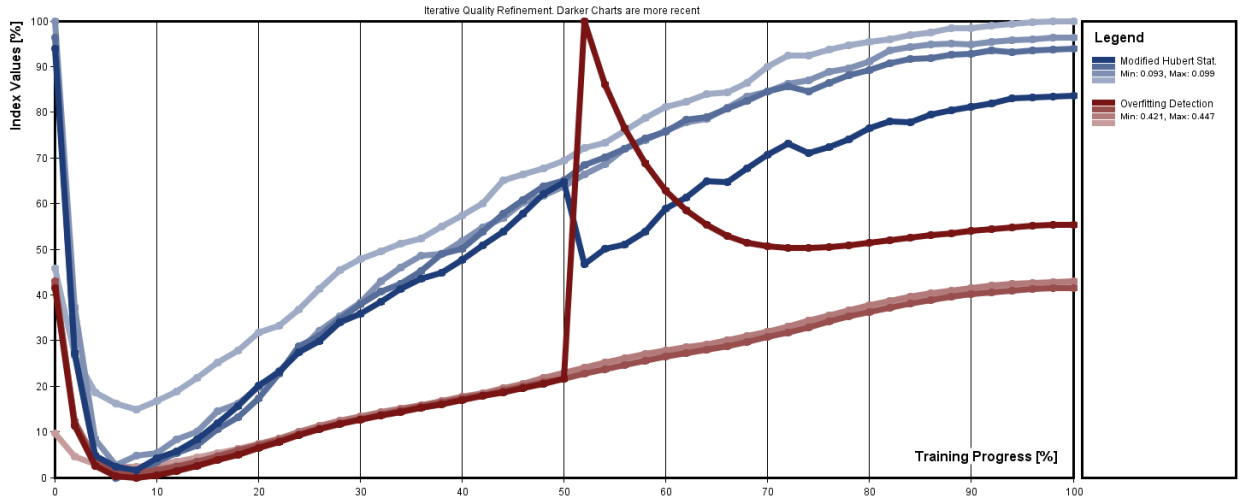


Figure 5.5 *Observation of the SOM training process. An intervention in the semi-supervised clustering process (the vector information of a cluster was subject to change) has a significant impact on the two displayed measures. The red measure is an error measure while the blue measure is a quality measure. This is why the red linechart shows a peak and the blue lineachart shows a sharp decrease both indicating a quality leak.*

overviews. On the one hand, this enables data scientists to improve the clustering results towards important quality aspects $\mathbf{RG}_{\text{CBO3}}$. On the other hand, change requests made by the user based on domain knowledge, application need, or preference can be assessed visually $\mathbf{RG}_{\text{CBO8}}$.

5.3.3. Quality Assessment with Cluster-Based Quality Measures

Cluster-based quality measures assign every single cluster a quality value. With this change in the granularity of quality assessment, we put the focus on the comparison of single clusters. For the design of enhanced content-based overviews, we use cluster-based quality measures for three use cases.

- The observation of a single cluster training
- The comparison of clusters within a single clustering result, possibly with multiple quality measures
- The comparison of multiple clustering results with a single measure

The two most widespread visualization techniques for cluster-based quality measures are color-based heatmaps and shape-based connector diagrams. An overview of available visual quality assessment techniques is presented in [Ber09, BvLBS11a]. In the following, we present color-based and shape-based visualization techniques using cluster-based quality measures.

Color-based visual quality assessment Color-based visual quality assessment techniques assign every cluster an individual color based on a predefined colormap. Typically, the set of quality values is normalized (min-max) to encode the quality by means of color. The SOM algorithm is beneficial for the visualization of the cluster-based quality. The grid-based nature of the SOM supports coloring the cells based on the quality values. As a result, the quality of a cluster visualization can be assessed by a heatmap metaphor. In addition, other cluster algorithms can be visualized in combination with a cluster layout strategy. We address the research goal of cluster layouts in 2D in Section 5.5. A SOM-based example presented in our reference publication [SBVLK09] is shown in Figure 5.6. Three different iterations of a SOM training are shown that assess the cluster quality visually. The QE quylity measure is used for every single cluster. A global min-max normalization of the quality values of all iterations facilitates the quality comparison of different steps within the training process. It can be seen that in the beginning of the training the QE measure has high (bad) values for most of the clusters. In the course of the training the QE decreases, and thus the quality of most of clusters improves. However, the clusters still have different quality values. Some of the clusters in the right image still perform badly. Local bad performances may have different reasons. First, the training of the clustering may still have the potential of improvement. This aspect can further be assessed in subsequent iterative

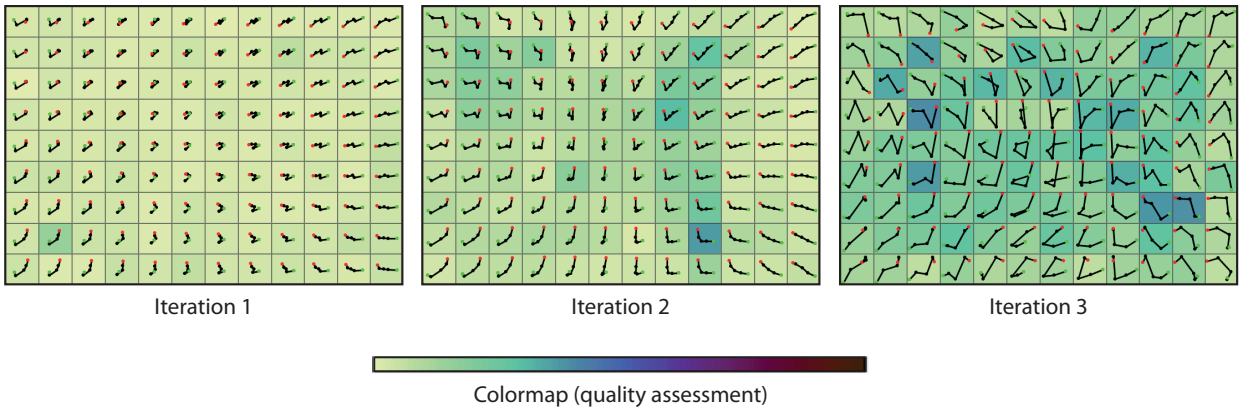


Figure 5.6 *Semi-supervised SOM algorithm. Visualization of three different iterations of the online learning process. The clusters are colored by a cluster-based quality measure. A global normalization of the quality measure values enables the visual comparison of clustering results within the training process.*

refinement steps. Second, high local QEs often occur for clusters covering regions of sparse data population. As such the QE measure can be applied as a visual means of outlier analysis.

Shape-based visual quality assessment Shapes are yet another visual variable which can be used for the visual quality assessment of single clusters. Shape-based visualizations typically link clusters with visual connectors to indicate some sort of relation. As an alternative, vector-based visualizations exist indicating a direction for a cluster with respect to an underlying association algorithm. The most popular example using this vector fields metaphor assigns every cluster in the clustering result visualization to a super-cluster (see [Ber09, BvLBS11a]). A connector-based variant is shown in Figure 5.7. The six images show the connectors of the TE measure. For the cluster-based quality visualization every connector assesses topology violations of involved BMUs. In the upper and the lower half of Figure 5.7 two SOM trainings are shown. The two trainings differ in their parameter calibration. On the left in the images 1a) and 2a), random initializations of the SOM manifold are illustrated. It is no surprise that random initializations bear a multitude of TEs (blue connectors). The images 1b) and 2b) show intermediate results of the SOM trainings after 10% of all iterations. In both images the numbers of TEs have decreased significantly. However, the result in the upper figure has more TEs than the result in the lower image. This trend continues in the images 1c) and 2c) where the final clustering results are shown. The upper figure still has some large TE connectors across the SOM grid while the training of the lower figure only has small local TEs.

Combined visualizations of color-based and shape-based visual quality assessment Cluster-based quality visualizations can also be based on a combination of color-based and shape-based techniques. Due to the different visual encodings two different quality measures can be composed in a single display. As a result, the complementing information of two quality measures be factored for the enhanced visual quality assessment. An example is shown in Figure 5.8 where the shape-based TE connectors (black) are visualized on top of a RGB similarity coloring. The RGB similarity coloring uses algorithms for the similarity-preserving visual representation of cluster vectors [Ber09]. Thus, clusters with similar vector information are visualized with similar colors. In Section 5.4.2, we present guidelines and techniques for the use of color to represent the similarity of high-dimensional data in detail. Figure 5.8 shows a usage scenario where a high-dimensional spoken letter data set (617 dimensions) was the input for a SOM-based vector quantization approach. On the left, the predominant letters for any cluster are visualized for an intermediate result. On the right, the combined color-based and shape-based visualization of the intermediate result is shown. Users are able to identify different color regions representing the vector properties of the high-dimensional data set. In addition, localizing identical colors at different regions of the SOM grid is easily possible. The color-based quality visualization is complemented by black connectors revealing TEs on the SOM manifold. The combined analysis of colors and error connectors in this usage scenario facilitates the identification of topology violations on the SOM manifold. For example, the lower left clusters of the grid (letters Q, W, H, and U - highlighted blue) share similar values with the cluster region at the center, right showing identical letters. The quality assessment visualization reveals that in the intermediate result still needs improvement before the clustering result can be applied as a content-based overview. In

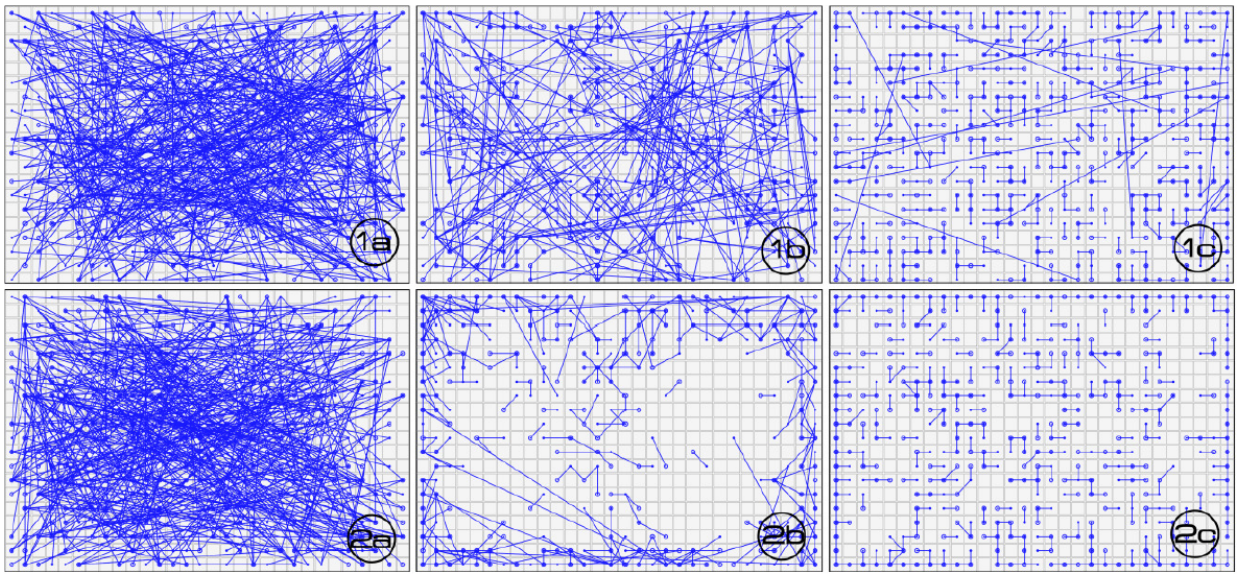


Figure 5.7 *Quality visualization showing TEs. Every blue connector indicates a TE, i.e., a data element having its two BMUs far away from each other on the SOM grid. Two SOM trainings are shown, the training at the bottom shows the better quality.*

this case, a quality goal in the layout of the clustering is not fulfilled sufficiently. We provide an overview of cluster layout strategies in Section 5.5.

To summarize, the visual quality assessment on the granularity of single clusters has several advantages. First, the visual comparison of different clusters within a clustering result reveals differences in the quality. This gives detailed information about the local quality of the clustering result. Moreover, this may disclose local properties of the high-dimensional data set, like dense regions or outliers. Second, the juxtaposition of several intermediate results of a clustering shows how the quality of every cluster develops within the training progress. Important quality aspects can be observed in detail, like the preservation of topology or the vector quantization. Third, the comparison of different clustering results enables users to relate (local) quality outcomes with parameterizations and modifications of the clustering result. As a result, cluster-based visual quality assessment particularly facilitates the justification of parameter values $\mathbf{RG}_{\text{CBO3}}$. In combination with a meaningful visual representation of cluster information, cluster-based visual quality assessment reaches the goal of gaining an understanding of the data set $\mathbf{RG}_{\text{CBO1}}$. In addition, the visual means of the techniques also enable the involvement of users in the design phase $\mathbf{RG}_{\text{CBO8}}$.

5.3.4. Quality Assessment with Data Element-Based Quality Measures

In the following, we present techniques for the visual quality assessment of clustering results on the granularity of single data elements. These techniques enable a fine-granular analysis of the distribution of samples in terms of the spatial cluster structure. In addition, the focus on single data elements also supports the analysis of relations between data elements and associated clusters based on different quality aspects. We show how a combined visualization of clusters and data elements can support the visual cluster analysis process and enhance visual quality assessment.

Supporting Technique: Micro-Macro Views In this section, we use the layout of clusters in 2D to create a second visualization layer on a finer granularity. The second layer visualizes the individual data elements based on the spatial distribution of the associated clusters. This idea is inspired by Edward R. Tufte‘s concept of Micro-Macro views [Tuf90]. Tufte shows designs displaying information on two different aggregation levels. The user is able to identify micro properties which can easily be related to the provided macro information. This concept is also promising for providing content-based overviews at different levels of detail. Micro-Macro views show the distribution of data elements based on the location of the cluster prototypes in 2D [BvLBS09]. The Micro-Macro views technique can be applied to any topology-preserving layout for clustering results. The SOM algorithm has the advantage to combine both the clustering and the layout step in a single routine. This is why we present the Micro-Macro views through the

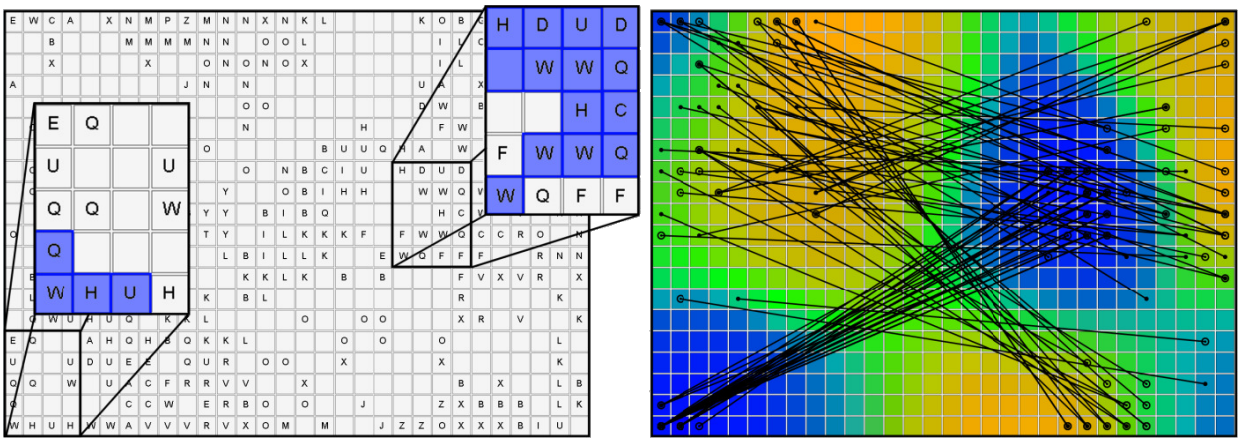


Figure 5.8 *Visual cluster quality assessment of a spoken-letter data set with 617 dimensions. On the left, the predominant spoken letters are visualized for each SOM cell. Two areas are enlarged where phonetically dissimilar letters are located next to each other on the SOM grid (e.g., W, H). On the right, we demonstrate two techniques allowing users to assess this friction in the SOM topology. The visual variable color is used to show the distribution of the high-dimensional input space in a similarity-preserving way. Black connector shapes encode the occurrence of TEs visually. It can be seen that the two areas, enlarged in the image, share similar vector information exposing various TEs.*

example of SOMs. For cluster layouts in general, we refer to Section 5.5 where we present guidelines and techniques for enhanced cluster layouts in 2D.

Figure 5.9 shows three different visual representations of the technique using the SOM as an example. At the macro level all three examples show a visual cluster representation (here: trajectory patterns). The micro level provides different visual encodings, depending on the properties of interest. The micro level on the left image shows the individual data elements of a data set in a scatter view. The user is empowered to identify the distribution of data elements in relation to the cluster representations. The micro level in the center image shows a density visualization based on a heatmap metaphor. The fine-granular density information of the data distribution can be related to the cluster representations. The star view in the right image links the entities of the micro level with the aggregated macro level. A star metaphor is chosen to explicitly show the relation of every data element to the spatial location of the cluster prototype. These three micro levels can be dynamically selected by the user in combination with the provided macro level.

Based on the cluster layout Micro-Macro views require an additional calculation step. The rationale is to refine the 2D mesh provided by the clusters (supporting points) to reveal a fine-gained projection layer for single data elements. The fitness of this projection layer can be at the pixel-level, depending on the parameterization. The calculation of the positions for every data element is described in previous works in detail [Ber09, BvLBS11a]. As a rule, the discrete position of clusters on the layout is replaced by a continuous spatial location of every data element. The mapping algorithm consists of two steps. First, we increase the number of supporting points by interpolating the cluster vector information based on the spatial distribution. We apply a cubic spline interpolation corresponding to Kohonen's suggestion of adequate local interpolation schemes [KSH01]. Second, we calculate the new screen position by weighing the probability function of a data element (the BMU including the neighborhood) on the new manifold [KSH01]. Main parameters of the Micro-Macro view technique (i.e., refinement resolution, color mapping normalization options, etc.) can be interactively steered by the user.

Data element-based visual quality assessment In the following, we use the Micro-Macro views technique for the visual quality assessment of clustering results on the granularity of data elements. At first, we prove the usefulness of the interpolation technique for the supporting points. For that purpose, we start with a cluster layout in 2D and successively increase the fitness parameter of the interpolation technique. Two proof-of-concept examples are shown in 5.10. In both examples a SOM-based cluster layout is provided. On the left, the effect of increased numbers of supporting points can be seen. From left to right the number of supporting points (clusters) on the 2D layout is increased, successively. Finally, the number of supporting points is 256 times as high as in the standard clustering (a SOM with 30×20 clusters). The increased number of supporting points also enables users to visually assess the clustering quality more precisely. The separation of good quality regions (blue) to bad quality regions (red) on

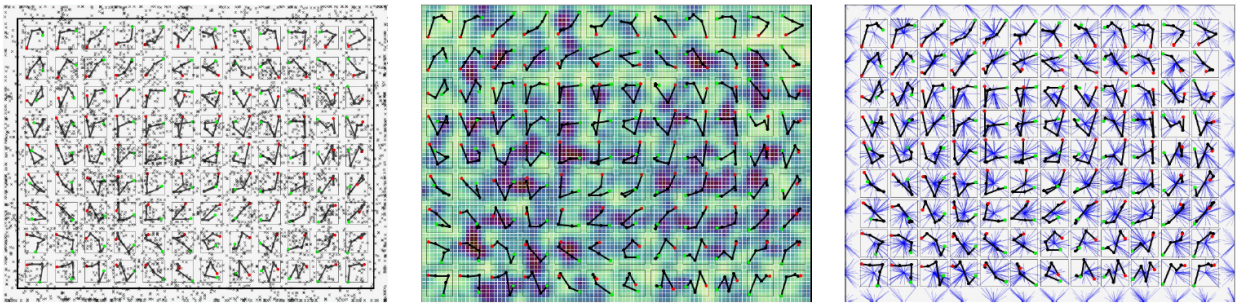


Figure 5.9 *Combined micro (data element) and macro (cluster) views can support the visual cluster analysis process. Implementations of the concept are a scatter-based view (left), a density-based view (center), and a star view (right).*

the manifold is more distinctly. On the right, the precision of a cluster-based quality visualization (small image) is compared with a data element-based visualization (large image). In the data element-based visualization the scatter view of the data elements is combined with a density map. Local properties of dense data regions (colored dark blue) can be analyzed in detail.

A variety of cluster quality measures can be calculated at the granularity of single data elements (cf. Section 5.2). In combination with the chosen cluster algorithm, Micro-Macro views use the quality information of data elements to support fine-granular visual cluster analysis. An important example is the QE of data elements. While for more coarse granularities of quality assessment the QE is the product of an aggregation scheme, the Micro-Macro view is able to visualize every single error value individually. In other words, local quality aspects in the data set can be explored in detail. We present a usage scenario where a DBScan clustering is presented in combination with the Micro-Macro views technique. Figure 5.11 shows three examples of the resulting Micro-Macro views. A trajectory data set with bivariate time-oriented data is chosen. In the three images different results of the DBScan cluster algorithm are shown with data element-based quality visualizations. The layouts of the DBScan clusters in 2D are represented at the macro level. In the left image the micro level uses the density view. Blue color values represent dense data regions enabling users to identify data clusters. Red areas of the layout represent sparsely populated regions of the data set. It is an interesting finding that virtually all cluster prototypes of the macro level are located in proximity to the dense data regions. The center and the right image contain similar visual encodings. The scatter view is chosen at the micro level. Every data point is visualized as a single data point colored by the QE. Green values indicate a good vector quantization. In these examples the user can easily identify areas where the majority of data elements show a good vector quantization. The proximity of such data elements in the layout and the consistently good quantization values support the user in the identification of cluster regions. The distribution of the DBScan cluster prototypes shows a similar picture. Virtually all of the DBScan clusters are located at dense regions of the data set. With the visual cluster representation at the macro level (the trajectories) the user can easily identify the cluster patterns contained in the data set. Red color values indicate data elements with a comparably worse vector quantization. In the right image this information is highlighted with an interactive user control, red data elements are encoded with a thick stroke width. With this interactive adaption the user is supported with a Micro-Macro technique for the identification and analysis of outliers.

We conclude the description of data element-based quality assessment techniques. We showed how visual cluster analysis and visual cluster quality assessment can be carried out at the granularity of data points. The user is able to identify fine-granular patterns of the data set. This information complements the overview gained at the global and the cluster-based granularity and supports the in-depth analysis of the data set $\mathbf{RG}_{\text{CBO1}}$. The assessment of local quality aspects also facilitates postprocessing steps, like merging similar clusters. The insight gained on the data set and the applied clustering algorithm also enables users to balance parameter values $\mathbf{RG}_{\text{CBO3}}$. The Micro-Macro technique implements Keogh's concept of shown different granularities of information by means of superposition. Referring to this, we also showed how visual cluster analysis can support interactive level-of-detail concepts for content-based overviews $\mathbf{RG}_{\text{CBO6}}$. Concerning the design of content-based overviews, we suggest data scientists and domain experts to carry out data element-based visual cluster quality assessment in a collaborative way. The different expertises may contribute to appropriate content-based overview solutions facilitating user-centered design $\mathbf{RG}_{\text{CBO8}}$.

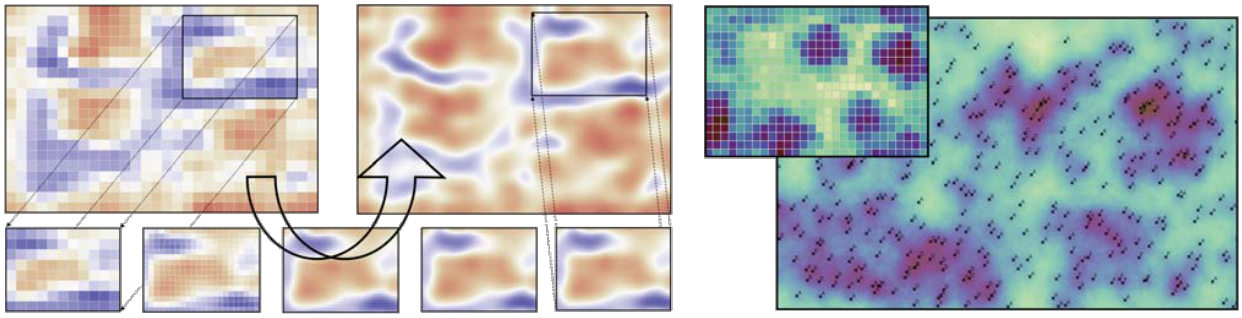


Figure 5.10 *Proof-of-concept examples of the new micro layer. In the left image the number of supporting points (interpolated cluster vectors) is increased, successively. With every iteration the display becomes more fine gained, here to reveal cluster structures (blue). In the right image a combination of the scatter view and the density view is shown. The spatial distribution of dense regions on the manifold can be analyzed at high precision.*

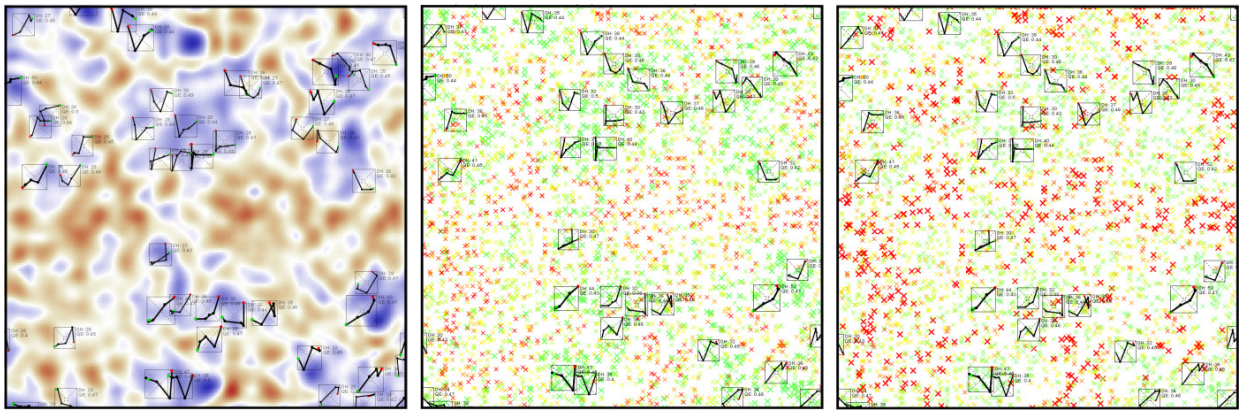


Figure 5.11 *Cluster correspondence views with clustering results of the DBScan algorithm. Three data element-based visualizations show relations between dense data regions and the locations of the DBScan clusterings. The trajectory visualizations reflect the vector information of the DBScan clusters. In the left image blue areas indicate dense data regions. In the two remaining images dominating green data elements (good QEs) indicate dense areas of the data set. In the right image a user control highlighted outliers (red data elements).*

5.3.5. Quality Assessment with Cluster Correspondence Views

Up to this point, we presented visual quality assessment techniques for different results of a *single clustering algorithm*. Our final concept supports the visual comparison of clustering results of *multiple clustering algorithms*. We introduce novel correspondence visualization techniques mapping the output of supportive clustering algorithms to a targeted clustering result visualization. An overview of illustrative examples is presented in Figure 5.12. For the targeted clustering result a 2D cluster layout serves as the baseline for mapping the clusters of supportive clustering algorithms onto the screen space. Similar to our Micro-Macro views concept, the cluster correspondence visualization can be applied to any topology-preserving layout for clustering results. We again refer to Section 5.5, where we present practical guidelines and techniques for the layout of clusters in 2D in detail.

With the cluster correspondence visualization technique, we support users in two complex cluster analysis tasks. First, we enable the visual comparison of the results of different clustering algorithms. As a result, users are able to relate the clusters of different algorithms with each other. Individual properties of the different clustering algorithms can be explored, exploited, and balanced against each other based on the analysis goal. For the design of content-based overviews this visualization technique allows choosing appropriate cluster algorithms $\mathbf{RG}_{\text{CBO2}}$. Second, we support the validation of a clustering result by comparing the output with contesting alternative algorithms. Thus, we use the individual properties of different clustering solutions to visually assess the quality of a targeted clustering algorithm. Similarly, the visual quality assessment of the targeted clustering algorithm also enables the validation of parameter values $\mathbf{RG}_{\text{CBO3}}$.

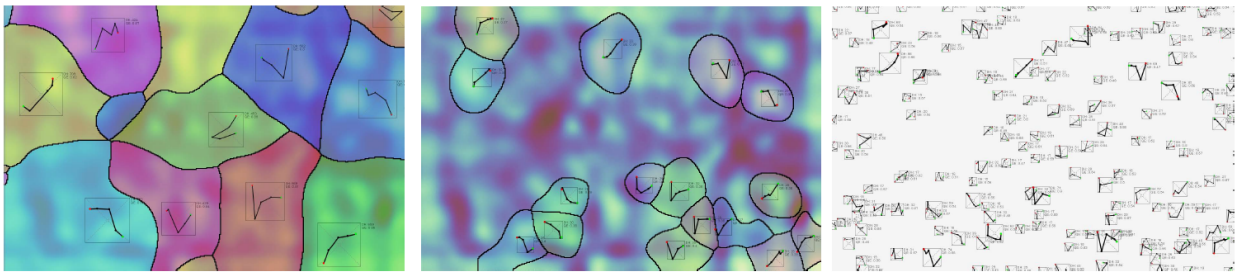


Figure 5.12 *Overview of illustrative examples using the cluster correspondence visualization technique.*

For our techniques, we chose the SOM as the targeted clustering and layout algorithm. In the latter section, we briefly described the algorithm mapping data elements to a targeted cluster layout. For an in-depth description, we refer to our corresponding publications [BvLBS09, BvLBS11a]. For the cluster correspondence visualization, we also map the vector information of supportive cluster centroids onto the layout. The centroids are displayed with visual cluster representations (in the example Figure 5.12: trajectories). A qualitative color mapping assigns cluster colors to single data elements. As an alternative the map manifold can be colored with respect to the affiliation to the most similar cluster.

On the left of Figure 5.12, the correspondence visualization of a k-means clustering result is shown. The centroids of the k-means clusters are visualized. The qualitative colormap partitions the display by the Voronoi cells of the k-means clusters. In the center image a DBScan clustering result is mapped onto a SOM-based data cluster heatmap. Light green background colors reveal dense regions of the data set. The mapping of the DBScan clusters predominantly matches the dense regions of the data set. On the right of Figure 5.12, an overview of a DBScan clustering with an increased number of clusters is shown. For every cluster the number of elements and the QE is visualized in addition to the trajectory visualization. The number of elements of a cluster is additionally encoded with the size of the visual representation. The example shown on the right may be an appropriate candidate for a content-based overview based on a DBScan clustering.

Illustrative Example I — Cluster Correspondence Between the SOM and a Supportive Clustering In the Figures 5.11 and 5.12 the results of single supportive clustering results are shown with the correspondence visualization. Based on the SOM cluster layout individual results of the k-means clustering and the DBScan clustering algorithm are mapped onto the layout. In Figure 5.11, we combine the data element-based Micro-Macro visualization with the cluster correspondence view. At the macro layer DBScan results are shown while the micro layer consists of density-based and scatter-based visualizations. The user is able to relate the corresponding clusters to interesting properties of the data set. Color value is used to a) reveal dense data regions on the manifold (blue color, left image) and to b) assess the QE quality of data elements (green color in the two remaining images). In Figure 5.12, different implementations of the cluster correspondence visualization technique are shown. In the left image, the manifold is colored with respect to the qualitative coloring of a mapped k-means result. At the center image a DBScan correspondence visualization is shown. The density-based DBScan clustering algorithm only assigns dense regions of the data set to clusters while neglecting the remaining data elements. In the example, the DBScan clustering calculated 16 clusters which are visualized on top of a density heatmap (bright colors indicate dense regions). The right image shows the correspondence visualization of a DBScan clustering result with multiple clusters on top of a SOM-based layout (the latter is not visualized). This image exposes how the correspondence views technique provides a layout of a supportive clustering algorithm based on a given cluster layout (here: SOM-based). In this way, cluster correspondence views can be used to provide content-based overviews for any supportive clustering result.

Illustrative Example II — Comparison of Multiple Supportive Clustering Results We recall a usage scenario of one of our publications, in which cluster correspondence views were demonstrated for an enhanced visual cluster analysis scenario [BvLBS11a]. In the usage scenario, we resolve two analytical challenges. The first challenge considers one of the most sustainable questions in cluster analysis, i.e., the question of how many ‘true’ clusters are contained in a given data set. The second challenge relates to the question of choosing appropriate parameters for a clustering algorithm to reveal these ‘true’ clusters.

In the two Figures 5.13 and 5.14, we show how cluster correspondence visualizations can support the user in resolving these two fundamental questions. In both Figures, a SOM-based layout is used to show supportive clustering

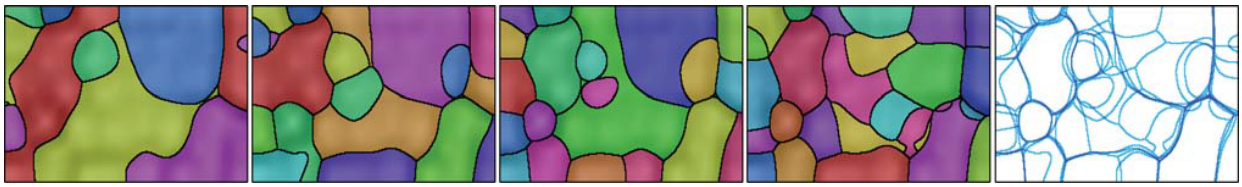


Figure 5.13 *Four cluster correspondence visualizations facilitating the k-means algorithm. The clustering results of different parameterizations ($k = 5$, $k = 10$, $k = 20$, and $k = 30$) can be compared. On the right, the cluster borders of all four clusterings are shown via superposition, revealing recurring cluster border patterns. The user is able to estimate a) the number of clusters of the data set and b) the appropriate parameterization for the k-means algorithm.*

results. In Figure 5.13, we apply the k-means clustering algorithm, while Figure 5.14 describes the workflow with a DBScan clustering algorithm. In both workflows four different parameterizations for the supportive clustering algorithms are applied. For the k-means workflow, we choose $k = 5$, $k = 10$, $k = 20$, and $k = 30$. The four different clustering results are shown by means of juxtaposition. The most interesting aspects of the four clustering results are the colored shapes of every cluster and the black cluster borders, respectively. For both clustering algorithms users can easily identify recurring cluster shapes and cluster borders. To make this findings more explicit, we condense the border information of the four clustering results in a layered image on the right. In Figure 5.13, the k-means cluster borders are colored blue, while in Figure 5.14 the DBScan cluster borders are colored green. Regardless of concrete parameterizations, users are able to identify recurring cluster borders. This is also evidenced by the finding that large white areas on the layout remain for all four parameterizations. Especially for the k-means workflow it is interesting to see that results with different k seem to share a common border surface reminding to a hierarchical clustering concept. It may be a subject of a more in-depth investigation in how the border structures of k-means results with different k are assignable to a hierarchical structures.

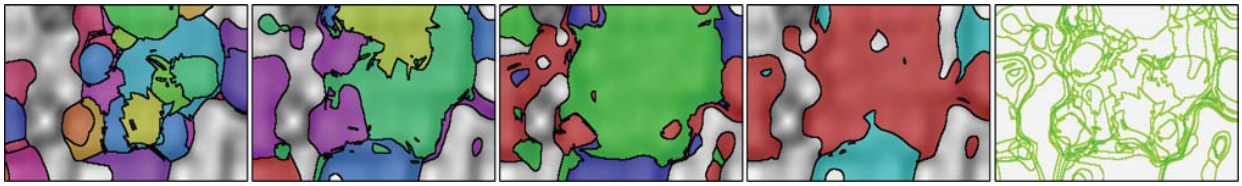


Figure 5.14 *Four cluster correspondence visualizations facilitating the DBScan algorithm. The clustering results of different parameterizations can be compared* $(\varepsilon, minPts)$ = $(5,5.00)$, $(10,5.50)$, $(15,5.75)$, $(20,5.50)$. *On the right, the cluster borders of all four clusterings are shown via superposition, revealing recurring cluster border patterns. The user is able to estimate a) the number of clusters of the data set and b) the appropriate parameterization for the DBScan algorithm.*

We summarize the usefulness of cluster correspondence visualizations. Cluster correspondence visualizations combine the results of different cluster algorithms in a single visualization by sharing a common cluster layout. This novel technique supports the user in assessing the results of different clustering algorithms in a single visualization. Specific properties of clustering algorithms can be compared and the best algorithm be selected for a particular part $\mathbf{RG}_{\text{CBO2}}$. Cluster correspondence visualizations also reveal the dependency of clustering results to their parameter values $\mathbf{RG}_{\text{CBO3}}$. In illustrative examples, we showed how different results of the k-means and the DBScan clustering algorithm can be compared with each other. This also enables users to gain an in-depth understanding of properties of a data set, such as the number of clusters $\mathbf{RG}_{\text{CBO1}}$. Finally, we suggest to use the visual means of the techniques for involving the user in the cluster analysis workflow $\mathbf{RG}_{\text{CBO8}}$ and for developing enhanced content-based overview visualizations.

5.4. Visual Mapping of High-Dimensional Data Objects

> *Like all coding schemes, a well-designed glyph-based visualization can facilitate efficient and effective information encoding and visual communication. (...) In dealing with the ever-increasing problem of data deluge, it is a technique that is not to be overlooked.*
>
> according to Borgo et al. [BKC*13], *2013*

In this section, we address the research goal of representing high-dimensional data elements and clusters visually $\mathbf{RG}_{\text{CBO4}}$. In the visual mapping step of the conceptual reference workflow (cf. Figure 5.1) the output of the data aggregation step is transformed into the visual space. As a general rule, the data aggregates are patterns of high-dimensional data, e.g., the product of visual-interactive cluster analysis (cf. Section 5.3). Encoding high-dimensional data objects with visual variables is crucial to facilitate visual data exploration [Kei02]. The visualization of data aggregates gives an insight into patterns of the high-dimensional data space. In addition, visualizing multiple data aggregates helps to gain an in-depth understanding of the relations of patterns in the high-dimensional data space. As such, visual mapping of high-dimensional data objects can serve as a profound basis for an efficient and effective knowledge generation process typically applied as a postprocessing step of cluster analysis $\mathbf{RG}_{\text{CBO1}}$. Similarly, for content-based overviews, visual data representations directly enable the visual access to the search space. In content-based overview, these visual data representations can be used for the formulation of queries by example $\mathbf{RG}_{\text{CBO7}}$. In any case, visual representations of the high-dimensional data aggregates are greatly beneficial for exploration activity.

Research in visually encoding high-dimensional data objects is conducted for a long time, an overview is provided in Section 5.2.2. A variety of encodings for high-dimensional data objects has been presented focusing on time-oriented data. In this section, we briefly reflect the related work and sharpen the eyesight for visual mappings for particularly suited for content-based overviews. Many inspiring concepts and techniques can be adopted for the design of enhanced ESS. In this section, we investigate two mandatory factors in more detail. First, we investigate the utilization of visual encodings for both high-dimensional clusters *and* for high-dimensional data. In many existing approaches the $1 : n$ relation between a cluster and the associated data elements is not considered. Visual encodings that solely include the data elements do hardly support synoptic analysis tasks [AA06]. On the contrary, visual encodings that only show cluster prototypes (e.g., the centroid) do not reflect the high-dimensional data characteristics of a cluster. We postulate that only visual data representations consisting of both the cluster information and associated data elements actually reveal the properties of a high-dimensional data subset. In this case, both elementary and synoptic tasks are supported [AA06]. Second, we emphasize the value of the visual variable color. Color can be used to visually encode high-dimensional data objects, even if rarely applied in the related work. The two benefits of using color are a) the ability of a continuous, similarity-preserving encoding of high-dimensional data objects and b) the ability to link data elements and clusters in different views provided by the ESS.

5.4.1. Glyph-Designs for High-Dimensional Data Clusters

Especially for large data sets solely the visualization of single high-dimensional data elements does not yield a usable content-based overview. As we will further discuss in the layout Section 5.5, with an increasing number of elements the display becomes cluttered and the visual representations will be overplotted. This is why (visual) data aggregation is often used to reduce the number of elements to a set of most relevant aggregates (cf. Section 5.3). In many cases, the information of these data aggregates can be transformed to the visual space, similar to the techniques for visually mapping high-dimensional data elements in general.

For the design of ESS, we suggest not only to encode cluster prototypes, but also the information about the data elements. On the one hand, for providing content-based overviews, it is highly appropriate to visually represent data aggregates. We recommend directly using visual cluster representations for querying interesting data subsets by example $\mathbf{RG}_{\text{CBO7}}$. On the other hand, the data elements (provided by data aggregates in an $1 : n$ relation) directly represent the search (result) space on an entity-level. As such, the information of the data set is presented at different levels of abstraction $\mathbf{RG}_{\text{CBO6}}$. Showing overview information about the granularity of data elements directly enables users to anticipate the results of potential search queries. As a consequence, content-based overviews providing information of clusters and data elements provide a preview-functionality even before a cluster was used as a Query-by-Example. At this point, we want to recall Tufte's concept of Micro-Macro views [Tuf90] and our technical adoptions presented in Section 5.3.

Requirements for Visual Cluster Representations Carrying out a glyph design for both high-dimensional clusters and high-dimensional data elements is challenging $\mathbf{RG}_{\text{CBO4}}$. The typical $1 : n$ relation between clusters and data elements requires specific visual encodings. An obvious design pitfall is to overload the visual cluster representation. As a consequence, the design principle of building simple interfaces may be violated (see, e.g., the work of Marti A. Hearst [Hea09]). We suggest to face a trade-off between the degree of provided information and the complexity of the visual cluster representation. It is not a surprise that we made good experiences in directly involving the user in the design process $\mathbf{RG}_{\text{CBO8}}$. In fact, carrying out cluster glyph designs is an iterative step in the ESS workflow. Similarly, non-experts can be incorporated in the design process in additional studies to improve the intuitiveness of the glyph design. In this way the interpretability of the glyph design can be evaluated and the generalizability of the design be guaranteed.

Glyph design is an iterative process [BKC*13]. One best-practice approach for every iteration may include a) to design several prototypes, b) to carry out an in-group design evaluation, to identify the best fit(s), and c) to gather additional qualitative feedback which can be adopted in the next iteration. See, e.g., one of our recent works for a real-world example [BSW*14]. In many cases, several meaningful solutions for a glyph design exist. One aspect to ensure clarity may be to recall the most relevant analysis tasks to be supported by the cluster glyph. Including such tasks in the evaluation process may help to further assess the appropriateness of different glyph designs for the ESS. An example of the evaluation of different glyph designs for time-oriented data is presented by Fuchs et al. [FFM*13]. In a controlled experiment, four glyphs (line glyph, star glyph, stripe glyph, and clock glyph) are evaluated. In the awareness of existing analysis tasks for single data elements (elementary tasks [AA06]) the participants of the study were asked to carry out tasks relevant for the temporal glyph designs (peak detection, temporal location, and trend detection). The three quantitative measures gathered in the course of the evaluation were the ratio of correct answers, the average completion time, and the confidence scores in percent. As a result, line glyphs are well suited for the detection of peaks and trends, while the radial layout of clock glyphs is beneficial for the temporal location of single values. For the evaluation of cluster glyphs, synoptic tasks [AA06] need to be included. Tasks relevant for the design of content-based overviews are *comparison*, *relation seeking*, *identification*, and *localization*.

Aside from the underlying data and the analysis tasks a cluster glyph also has to comply with user preferences. To a certain extent data and tasks may allow multiple design choices. Designing visual encodings then relies on the expertise of the designers and naturally on the preference of the users. However, as briefly discussed in Section 5.2.2, assigning visual variables to high-dimensional multi-faceted data is not arbitrary. Designers must be aware of perceptual principles of visual variables, gestalt psychology, and semiology [Ber83, Tuf86, Tuf90]. It is mandatory to choose the visual variables with respect to their ability to encode data.

From a functional perspective, we postulate a guideline for cluster glyph designs for content-based overviews.

- $\mathbf{R}_1$ The glyph design has to be applicable for time-oriented data if time-oriented data is provided. Hence, the temporal domain of the cluster information should be represented visually.
- $\mathbf{R}_2$ The value domain of the cluster information has to be represented visually. The most crucial aspect is that the user is able to *identify* the high-dimensional vector information.
- $\mathbf{R}_3$ The number of data elements of a cluster (the size) must be recognizable with the cluster glyph. In this way, the user gains an impression of the size relationships of the displayed data subsets.
- $\mathbf{R}_4$ The variations of the data elements within the cluster need to be encoded visually. As a result, the user is, e.g., able to assess the compactness (the homogeneity) of the represented data subset.
- $\mathbf{R}_5$ The visual encodings of the cluster glyph should support *comparison and relation-seeking* tasks. The perceived similarity between two glyphs should comply with the similarity in the high-dimensional space.

Visual Cluster Representation Examples Design is a creative process. Typically, different approaches including variations in data, tasks, and users lead to individual solutions. In the following, we present a summary of visual cluster designs carried out in some of our publications. Table 5.1 shows an overview of the different approaches. The table columns are structured as follows. We present three columns showing an example of the cluster glyph design followed by a similar and a dissimilar cluster glyph. Finally, five additional columns show details in how the requirements $\mathbf{R}_1$, $\mathbf{R}_2$, $\mathbf{R}_3$, $\mathbf{R}_4$, and $\mathbf{R}_5$ are implemented.

In most examples of univariate time-oriented data the temporal domain $\mathbf{R}_1$ is visually encoded with the position information of the x-axis. For bivariate data the temporal domain can be encoded with connectors between temporally adjacent data elements in 2D. With the MotionExplorer case study in Section 7.2, we present an example of multivariate

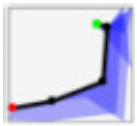
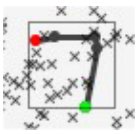
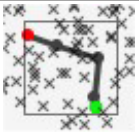
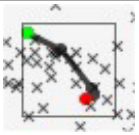
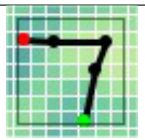
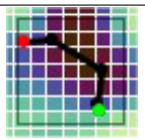
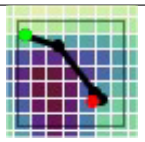
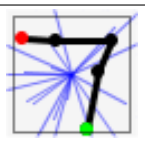
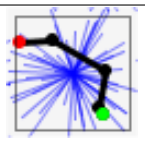
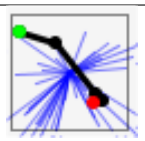
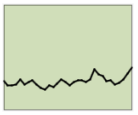
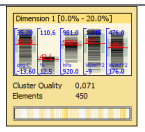
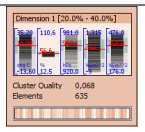
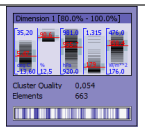
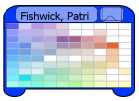
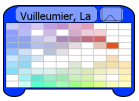
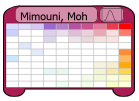
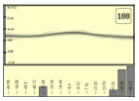
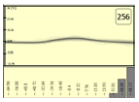
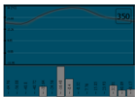
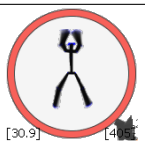
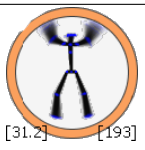
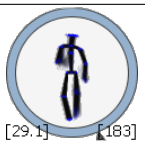
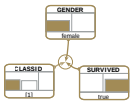
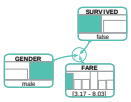
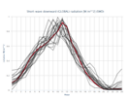
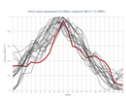
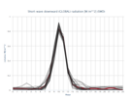
Reference Glyph publication	Glyph Example cluster glyph design of the reference	Similar Cluster Cluster glyph with similar content like the example glyph	Dissimilar Cluster Cluster glyph with dissimilar content like the example glyph	Temporal Encoding Visual encoding of the temporal domain	Cluster Encoding Visual encoding of the value domain of the cluster	Size Encoding Visual encoding of the size of the cluster	Variation Encoding Encoding of the variation of the cluster	Relation Encoding Encoding of the of the similarity preservation
[SBVLK09]				Position	Position	Implicit (opacity)	Position + opacity	Cluster enc.
[BvLBS09]				Position	Position	Implicit (position)	Position (scatter plot)	Cluster enc.
[BvLBS09]				Position	Position	Color, heatmap	Position, heatmap	Cluster enc.
[BvLBS09]				Position	Position	Implicit (position)	Position (star plot)	Cluster enc.
[BBF*11]				Position, linechart	Position, linechart	Not provided	Not provided	Cluster enc. + Color
[BWS*12]				Timeline (position)	5 Bars	Label	Variance bars (position)	Cluster enc. + color
[BRS*12a]				Position	Position	Implicit	Position + transparency	Cluster enc. + color
[BRS*12b]				Position, linechart	Position, linechart	Label	Position (opacity bands)	Color (interestingness)
[BWK*13]				Not provided	Stick figure (position)	Label, optional: size	Opacity stick figures, label	Cluster enc. + color
[BSW*14]				Not provided	Position, barchart	Position, barchart	Position, barchart	Color
[BDF*15]				Position, linechart	Position, linechart	Label	Position, linecharts	Cluster enc.

Table 5.1 *Examples of visual cluster encodings developed for different applications with time-oriented data.*

data (human motion capture data). In this case, a node-link diagram is chosen to show temporal transitions between the high-dimensional cluster vectors (cf. Figure 5.25).

In many cases, the value domain $\mathbf{R}_2$ of univariate cluster glyphs is visually represented by the position information of the y-axis. Together with temporal domain, shown at the x-axis, this combination results in a linechart metaphor. In the bivariate examples the *position* information of the x-axis and the y-axis is chosen. As a consequence, the value domain is represented in a scatter metaphor. These examples showing multivariate clusters have various encodings. In this complex design space the number of general diagram types is scarce. With the two examples presented in Table 5.1, we present a general technique and a specific solution as the result of a user-centered design approach. In the five-dimensional usage scenario the y-axis is used multiple times with individual bar-line metaphors [BWS*12]. A possible extension of the solution would be to connect the data in the five diagrams to reveal a parallel coordinates metaphor. In the human motion analysis scenario [BWK*13] the position information of the x-axis and the y-axis is chosen to draw the joints of a 48 dimensional human pose cluster. With the knowledge of the involved user group the joints were connected to a stick-figure metaphor for an enhanced identification of human body configurations.

For the visual representation of the cluster size $\mathbf{R}_3$ different visual variables are chosen. First, labels are included to show the exact number of elements in a textual way. Second, the data elements of a cluster are visualized by means of superposition. Finally, the size of some cluster glyphs is chosen to represent individual cluster sizes.

A scalable metaphor for the visual representation of the variation $\mathbf{R}_4$ of univariate data elements is the boxplot metaphor. The boxplot shows quantile information of a set of numerical values, and thus serves as a visual aggregation of a possibly large number of data elements. The opacity-bands technique is another means of showing the variation of the data elements within a cluster. In this case, the data elements of a cluster visualized in superposition are additionally furnished with a semi-transparent bend metaphor. In an example of bivariate data [SBVLK09], opacity bands connect trajectory clusters with the individual trajectory data elements. In the examples of multivariate data solely the visualization of the data elements shows the variation of the cluster. Due to the complexity of position information for multiple data elements and multiple dimensions an additional explicit encoding of variation might induce too much confusion.

The preservation of similarity between clusters $\mathbf{R}_5$ is a high-level requirement. In most examples the visual encodings for $\mathbf{R}_1$, $\mathbf{R}_2$, $\mathbf{R}_3$, and $\mathbf{R}_4$ contribute to the ability to estimate the similarity between any two clusters. In accordance with the design guidelines for visual variables [Ber83] the *position* variable plays an important role in many examples of Table 5.1. The relative positions within every cluster glyph are highly relevant to represent high-dimensional cluster information, and thus to facilitate the comparison of different clusters. The second visual variable in these examples used for relating different clusters with each other is *color*. Especially for multivariate cluster glyphs the preservation of similarity between clusters can be supported with color information. The MotionExplorer case study [BWK*13] serves as an example. High-dimensional vectors representing human poses are encoded with color in a similarity-preserving way. Both the red cluster and the orange cluster in Table 5.1 show human poses having their arms above their head. The two colors are comparably similar. On the contrary, a blue cluster shows dissimilar human poses having their arms going downwards. Likewise, the blue color is very different compared with the red and orange color.

Summary We presented guidelines and techniques for the visual mapping of high-dimensional objects by means of glyph designs. In this connection, the five requirements mentioned above foster the use of glyph designs for content-based overviews. Hence, we postulated a solution to represent high-dimensional data elements and clusters visually $\mathbf{RG}_{\text{CBO4}}$. With the eleven glyph designs from our recent works presented in Table 5.1, we presented different examples of how glyph designs can support users in gaining an understanding of the data set $\mathbf{RG}_{\text{CBO1}}$. Finally, the different examples presented in Table 5.1 also show how user preference may be included in the design process. Provided that design principles are considered, different solutions may exist for a given data set which can be factored in user-centered design approaches $\mathbf{RG}_{\text{CBO8}}$.

5.4.2. The Use of Color

For the design of ESS the visual variable color plays a key role. ESS typically consist of different visualizations each representing the data from a *different perspective*. Most relevant visualizations are content-based overviews, visual query interfaces, and the search-result visualizations. A powerful technique in analytical systems providing different views is the *linking-views* concept. Linking views supports the enhanced localization of objects in different views. Hence, user are empowered to seek relations of similar objects with respect to the specific properties displayed in different views. For example, if users are able to localize a cluster in a calendar-based and a geo-based view, they

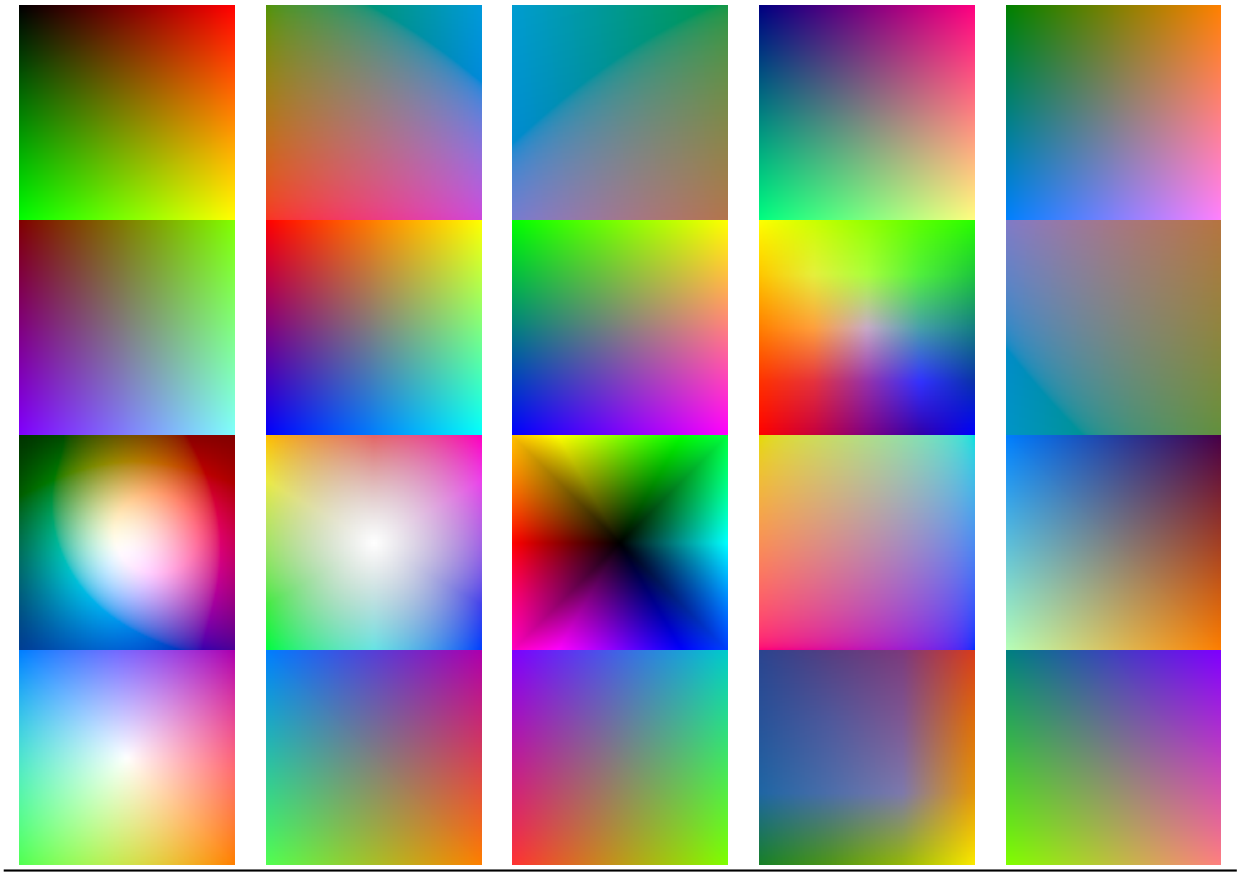


Table 5.2 *Static 2D colormaps for mapping high-dimensional data elements to the visual variable color. The colormaps differ in their applicability to reflect pairwise similarity and thus for content-based overviews.*

can reveal spatio-temporal relations. From an interaction perspective, it is always possible to provide linking by highlighting focused objects in different views, on-demand. However, we argue that user interaction should not remain the only means for linking views. We advocate that linking concepts based on (static) visual encodings are more powerful for gaining an understanding of the entire data collection at a glance. In Table 5.1, we show examples in which the visual variables *position* and *color* are used to relate different cluster glyphs in the content-based overview component. Likewise, these visual variables enable the localization of known cluster glyphs in multiple views. However, not every view of an ESS is supposed to show cluster glyphs in large scale. Reasons may be limitations in the display space, substantially higher numbers of objects, or simply an inappropriate visualization technique. In these cases, the visual variable position does not suffice linking views. In return, the visual variable *color* is still highly appropriate. Looking-up colors in different views is an easily feasible task for human perception. The variable color enables linking similar data elements down to the granularity of single pixels. In Section 5.2.2, we briefly reviewed different design guidelines and techniques for univariate colormaps. We also described why univariate colormaps hardly preserve the pairwise similarity of high-dimensional data objects, such as time series FVs. However, for the design of ESS a solution for coloring high-dimensional data objects is desirable. In this section, we present a guideline for the use of the visual variable *color* for high-dimensional data objects.

2D Colormaps for Content-Based Overviews As a basic principle for the use of the visual variable color in a quantitative way, similar data should be colored with similar colors while dissimilar data should be encoded with dissimilar colors. In contrast to univariate colormaps, 2D colormaps enable mapping high-dimensional data objects to color. The static 2D colormap approach is divided into two steps (see Figure 5.15). First, a data projection technique maps the high-dimensional data elements into 2D. In this step, it is important that the projection algorithm preserves the pairwise distances of the data elements in a best possible manner. We will come back to the challenge of choosing appropriate projection algorithms in Section 5.5. Second, the projected 2D data receives a color by a predefined (static)

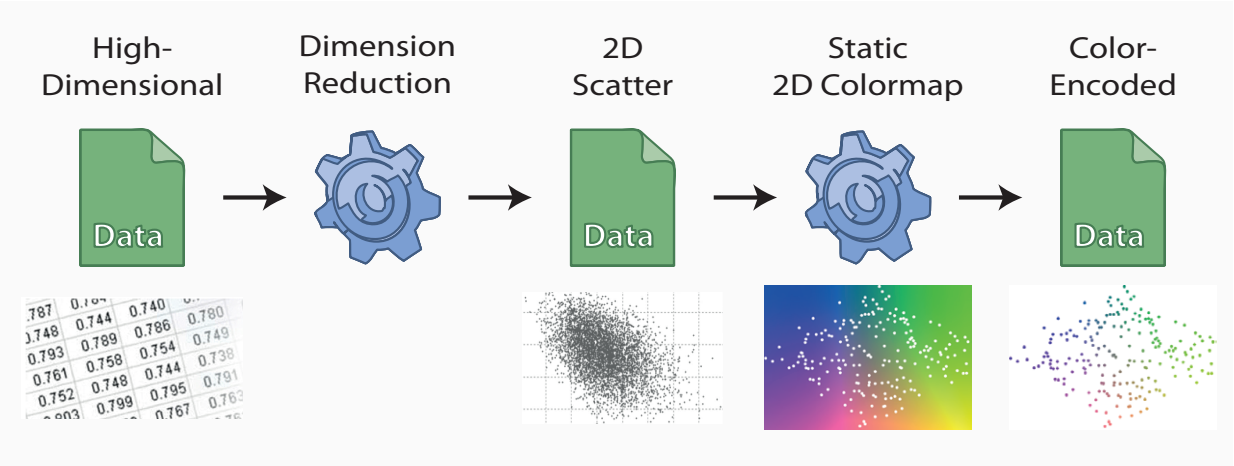


Figure 5.15 *2D Colormaps: workflow for visually mapping high-dimensional data with similarity-preserving colors.*

2D colormap. In a previous publication, we presented a survey of existing static 2D colormaps [BSM*15b], an overview of relevant colormaps is provided in Table 5.2. In addition, we presented seven quantitative *quality measures* for 2D colormaps. Finally, we provided a *guideline for the design and the use* of 2D colormaps with respect to analysis tasks.

The quality measures allow to assess the goodness of fit of static 2D colormaps for different design criteria. For example, for being able to distinguish many different data properties the number of available colors in a 2D colormap should be as high as possible (*color exploitation*). An additional design criterion regards the *perceptual linearity* of the 2D colormap. It is important that the perceived color distances of any two points of the 2D colormap correspond to the spatial distances of respective points. This quality criterion can further be sub-divided in the preservation of local and the preservation of global perceptual aspects. Violations of this quality criterion will either induce non-existing relationships, or disguise existing relationships between information units. Other quality measures assess the distances to black and to white color (*black, white distance*) to avoid conflicts with background or contours colors. Yet another quality criterion regards the *attention steering* capability of different colors. For many analysis tasks it is important that all provided colors are equally salient for the human perception.

In the following, we present the set of most appropriate static 2D colormaps for content-based overviews. To achieve this, we recall the salient analysis tasks provided with content-based overviews according to the task taxonomy of Andrienko and Andrienko [AA06]. First, content-based overviews allow the *identification* of (unknown) information. Second, content-based overviews facilitate the *localization* of (known) information. Finally, content-based overviews support the *comparison/relation* of objects with other objects or additional metadata. Altogether, with these analysis tasks content-based overviews support exploration activity within the search space. If the content-based overview provides the exploration of both data elements and data aggregates, these tasks can be carried out on a elementary and a synoptic granularity.

A Guideline for Similarity-Preserving Color Mappings We present a guideline for the use of existing static 2D colormaps based on their goodness-of-fit for content-based overviews. For that purpose, we apply a mapping of the analysis tasks provided with content-based overviews to the quality criteria for the design of static 2D colormaps. For an in-depth description of quality criteria, analysis tasks, and a mapping between these two, we refer to our survey publication on static 2D colormaps [BSM*15b]. We briefly present the results of the mapping of the relevant analysis tasks to the provided design criteria. The salient quality criteria are the *color exploitation* and the *perceptual linearity*. In addition, the *attention steering* quality should be considered. Finally, we recommend colormaps to adhere to a certain distance to contours and background colors used in a system *white distance* and *black distance*. We treat these two quality measures separately to better distinguish between use cases with black and white backgrounds/contours colors. To preserve the relative importance, we define a weighting of the presented quality criteria. The color exploitation and the perceptual linearity are considered with the weight 2.0. We divide the latter quality criterion in the preservation of local and global perceptual linearity. The attention steering criterion and the distances to white and black color receive the weight of 1.0. However, other weightings of the quality criteria for static 2D colormaps in individual use cases may also be appropriate.

Title	Color Space	Image	Black Background 2x Color Exploitation 1x Perceptual Linearity (global) 1x Perceptual Linearity (local) 1x Attention Steering	White Background 2x Color Exploitation 1x Perceptual Linearity (global) 1x Perceptual Linearity (local) 1x Attention Steering
Bremm et al. (regular)	L*a*b		+	+
Constant Blue	sRGB		−	∅
Cube Diagonal Cut BCYR	sRGB		++	+
Ramirez et al.	HSV		++	∉
Mittelstädt et al.	L*a*b		++	+
TeulingFig2	sRGB		∉	+
TeulingFig3	sRGB		++	∉
TeulingFig3NoWhitening	sRGB		+	++
TeulingFig4a	sRGB		+	+
Yeo et al.	sRGB		−	−

Table 5.3 *Overview of the 10 most applicable static 2D colormaps for analysis tasks relevant in content-based overviews. We conduct the study for both black and white background colors. The analysis tasks require different quality criteria for static colormaps such as the exploitation of colors, the perceptual linearity and attention steering [BSM* 15b]. The Borda count election method provides the overall scores aggregated to ++, +, ∅, and −. We exclude 12 of the 22 available colormaps since they performed very weak in at least one quality citerion (∉).*

Case Study — the MotionExplorer System To prove the usefulness of the guideline, we demonstrate the results of a real-world case study. In Section 7.2, we present the MotionExplorer system enabling domain experts in the ES of human motion capture data. MotionExplorer makes massive use of the visual variable color for linking the seven views. To this end, high-dimensional human motion poses are encoded with color in a similarity-preserving way. The question arises which of the many existing static 2D colormaps is the most appropriate one for the analysis tasks in the MotionExplorer ESS. We use the guideline for similarity-preserving color mappings. Based on 22 candidates, we will identify the colormap fitting best for the given tasks. The quality of each colormap is assessed with the quality measures described in the guideline. We double the weight of the color exploitation and the perceptual linearity scores. Moreover, we take the attention steering and the distances to the background/contours into account. For the MotionExplorer system, we select the distance to the white background as another criterion. Based on the Borda count

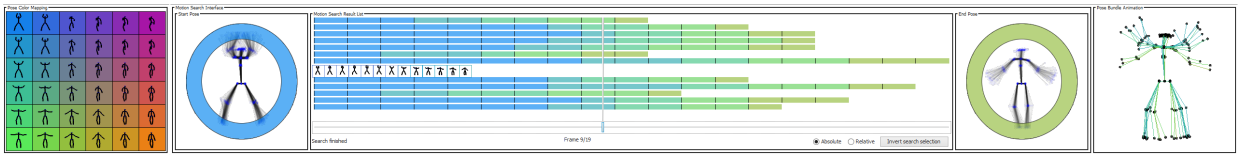


Figure 5.16 *The MotionExplorer system as an example of a system using a static 2D colormap. This example shows an appropriate colormap for the tasks provided by content-based overviews and the linking-views concept. The colormap has a large number of distinguishable colors and is particularly suitable with respect to the perceptual linearity, the attention steering, and the distances to white and black color.*

election method, we calculate an overall ranking of the 22 colormaps. The result is presented in Table 5.3. The ten best colormaps are listed in alphabetical order, for the sake of illustration, we conduct the study for both black and white color distances. The aggregated Borda ranks (++, + , ∅, and −) are presented for each of the candidates. It can be seen that three colormaps perform well (at least +) for both black and white backgrounds. In the end, we chose the colormap 'TeulingFig3NoWhitening' as the most appropriate candidate for white backgrounds. Figure 5.16 presents the winner candidate included in the MotionExplorer system. Table 5.3 neglects 12 of the 22 available colormaps. These 12 colormaps scored worse in at least one of the quality measures. In this way, we avoid that a candidate wins even if it has design flaws in at least one important factor. One of the less inappropriate colormaps is shown in Figure 5.16. This colormap makes only limited use of the available colors. In addition, the colormap has very bright and very dark colors which would be difficult to differentiate from the background and the contours of the MotionExplorer system.

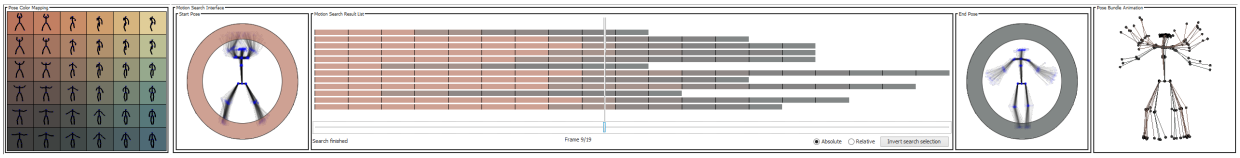


Figure 5.17 *The MotionExplorer system as an example of a system using a static 2D colormap for linking views. In this example, an inappropriate colormap for content-based overviews is provided. The colormap has a low number of distinguishable colors, weak distances to white and black backgrounds, and a poor perceptual linearity.*

Summary We summarize the guideline for the use of color for content-based overviews. Static 2D colormaps are a powerful means of representing high-dimensional objects in the visual space in a similarity-preserving way. The large number of existing variants, the number of different quality goals, and the variety of analysis tasks to be supported, pose a huge design space. To this end, we presented a guideline which identifies appropriate colormaps for a given set of quality goals or analysis tasks. In a recent publication, we also draw a connection between quality goals and analysis tasks. This is why we are able to suggest a weighting of the different quality goals to support the design of visual representations of high-dimensional objects in a best possible manner $\mathbf{RG}_{\text{CBO4}}$. By using color in a similarity preserving way, users are able to carry out analysis tasks, such as identification or comparison tasks. In this way, we improve the process of gaining an understanding of the underlying data set $\mathbf{RG}_{\text{CBO1}}$, e.g., in content-based overviews. The guideline and the overview of existing colormap approaches also allows involving users in the design process. Typically, different colormaps are appropriate the targeted approach, enabling data scientists to consider the preference of domain experts $\mathbf{RG}_{\text{CBO8}}$.

5.5. Layouts for Aggregated Data

> *Visual transformations are used to modify and augment the static visual structures to create views into the visual structures. Visualizations exist in space-time. View transformations exploit time to extract more information from the visualization than would be possible statically.*
>
> according to Card et al. [CMS99] p. 31, *1999*

The final step in the reference workflow for the design of ESS regards the layout of aggregated data (cf. Figure 5.1). Layouts provide the position information for data elements and clusters in the display space. We distinguish between projection-based, force-directed, and cluster structure-based layout approaches, see Section 5.2.3 for an overview of baseline techniques. In the following, we show how each of the three types of layouts can be used for content-based overviews for ESS. First, we present solutions for the choice of appropriate layouts in Section 5.5.1 $\mathbf{RG}_{\text{CBO5}}$. Second, we show how the three layout types can be provided with concepts for different levels of data abstraction in Section 5.5.2 $\mathbf{RG}_{\text{CBO6}}$. Finally, we demonstrate how the layouts can facilitate visual querying by example in Section 5.5.3 $\mathbf{RG}_{\text{CBO7}}$.

5.5.1. Choice of Layout Technique

Choosing an appropriate layout for the underlying data (or data aggregation) is an important challenge in the design of content-based overviews. Influencing criteria are, e.g., the chosen data aggregation method, the trade-off between topology preservation and overplotting avoidance, as well as other system and user requirements $\mathbf{RG}_{\text{CBO8}}$.

Cluster Structure-Based Solutions For cluster structure-based layouts the design decision directly depends on the chosen data aggregation technique. If a hierarchical or a grid-based (network-based) clustering algorithm was chosen, the resulting cluster structure can directly be used for the layout design. As a consequence, for cluster structure-based solutions no additional projection or force-directed layout technique needs to be included. In many cases, design decisions for the clustering step and the layout step are a joint effort. In the section about quality-driven visual-interactive cluster analysis, we showed how such clustering algorithms can be selected and be parameterized appropriately (cf. Section 5.3). As a result, the remaining design space for cluster structure-based layout solutions relates to the specific characteristics of the clustering output, system requirements, and the preference of the user.

We take a closer look at *grid-based* clustering results. The regular structure of the clustering result enhances the view transformation of the output, i.e., the position information of the clusters can directly be used for the layout in the visual space. The arrangement of clusters in a grid is only one side of the coin. In addition, we recommend this type of clustering result to provide a second important property, i.e., the preservation of topology. Neighboring clusters in the input space also need to be aligned close to each other in the output space. The cluster layout will then be particularly usable for content-based overviews, e.g., to support users in gaining an understanding of the structural information of the underlying data set $\mathbf{RG}_{\text{CBO1}}$. In Section 5.3, we presented techniques for the visual assessment of cluster quality including measures for the topology preservation.

The SOM algorithm may serve as an example of the design of layouts for grid-based cluster structures. The output of a SOM clustering is a regular grid structure which can directly be visualized as a layout in 2D. The number of cells depends on the size of the targeted overview, the size of the glyph design, and on the preference of the user. In a user-centered design approach, the user should be able to determine the grid resolution, possibly based on multiple prototypical solutions. In combination with visual-interactive cluster analysis (cf. Section 5.3) and iterative glyph designs (cf. Section 5.4), this provides a powerful basis for involving the user in the design process $\mathbf{RG}_{\text{CBO8}}$. An alternative design decision could be that users have the means to interactively refine the grid resolution when using the ESS. In the VisInfo case study in Section 7.1, we present an ESS with a SOM-based content-based overview. Figure 5.18 shows the content of Shortwave-Downward radiation measurements (SWD). The daily curve progressions are aggregated to 80 clusters structured in a $8x10$ grid. The provided grid resolution is the result of a requirement analysis phase, conducted with domain experts from Earth observation research. The cluster glyph shows the mean

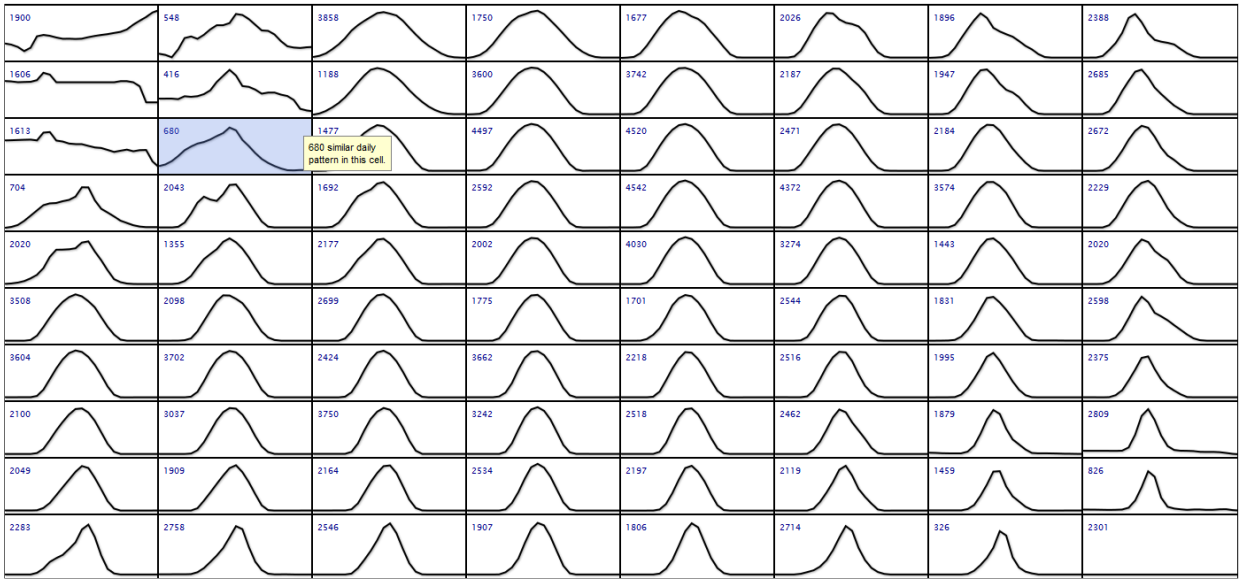


Figure 5.18 *SOM-based grid layout providing a content-based overview. About 200,000 daily curve progressions of Shortwave-Downward radiation are visually represented in the grid-based cluster layout. It can be seen that many solar-dependent radiation measurements have a peak in the middle of the temporal domain (roughly noon). However, the peaks of the upper right patterns occur earlier in the day compared to the measurements at the lower left. The domain experts reason that this effect depends on maritime and continental climate influences.*

daily curve progression for every cluster cell. The temporal domain of the data is aggregated to a 24-dimensional vector (one vector for every hour) each containing a numeric value. A click on a cell opens a Details-on-Demand view showing individual data elements. We will take an in-depth look at the Details-on-Demand view in Section 5.5.3 on visual querying. In addition, for every cluster we show label at the upper left representing the cluster size. Overall, the grid structure provides an overview of almost 200,000 daily patterns provided in the data set. Based on the intrinsic property of the SOM, the patterns in the grid-based layout are aligned in a topology-preserving way. The user is able to identify a global order of the high-dimensional information. All in all, the user is able to gain an overview of 4,500,000 numeric values in an intuitive way $\mathbf{RG}_{\text{CBO1}}$. The SOM grid also resolves the challenge of overplotting. In return, the SOM output maximizes both a non-overlapping alignment of clusters and the exploitation of the display space (100% of the display space is allocated by the rectangular grid layout).

Choosing a layout for grid-based clustering results is significantly easier than for arbitrary clustering results. The remaining design decision is the parameterization of the SOM (cf. Section 5.5.3) including the definition of the grid resolution. In the VisInfo case study the choice of the resolution was made by incorporating the user in the design process $\mathbf{RG}_{\text{CBO8}}$. To summarize, the combination of grid-based cluster structures and user-centered design is beneficial for providing meaningful layouts $\mathbf{RG}_{\text{CBO5}}$.

Hierarchical clustering results pose another type of clustering result which can be used for the layout step in the design workflow. Moreover, the nested structure of the hierarchical clustering output enhances the visualization of the clustering result at different levels of detail. The upstream steps in the reference workflow pose a variety of parameters which we discussed in Sections 5.5.3 and 5.4. In the layout of hierarchical clustering results, important design parameters are the size of the visualization and the glyphs, the visualization and interaction design, and particularly the targeted level of abstraction which determines the depth of the hierarchy visualization $\mathbf{RG}_{\text{CBO6}}$. The method to specify the aggregation level is subject to user level design. One approach is to provide a static aggregation level defined by user preference. In this case, an iterative approach may foster the choice of the most meaningful aggregation level. Other, more dynamic solutions include an interactive adjustment of the aggregation level, either at a global level, or at different local levels of the hierarchy. In any case, we recommend the involvement of users in the process of choosing different levels of abstraction $\mathbf{RG}_{\text{CBO8}}$. The level of abstraction is not the only design parameter for hierarchical visualizations. The visualization and interaction design poses additional factors requiring careful design choices. Important parameters are the intended dimensionality, the representation type and the alignment strategy in the display. For an in-depth overview of convincing examples of visualization and interaction techniques for hierarchical data aggregates, we refer to the design guideline by Elmqvist and Fekete [EF10].

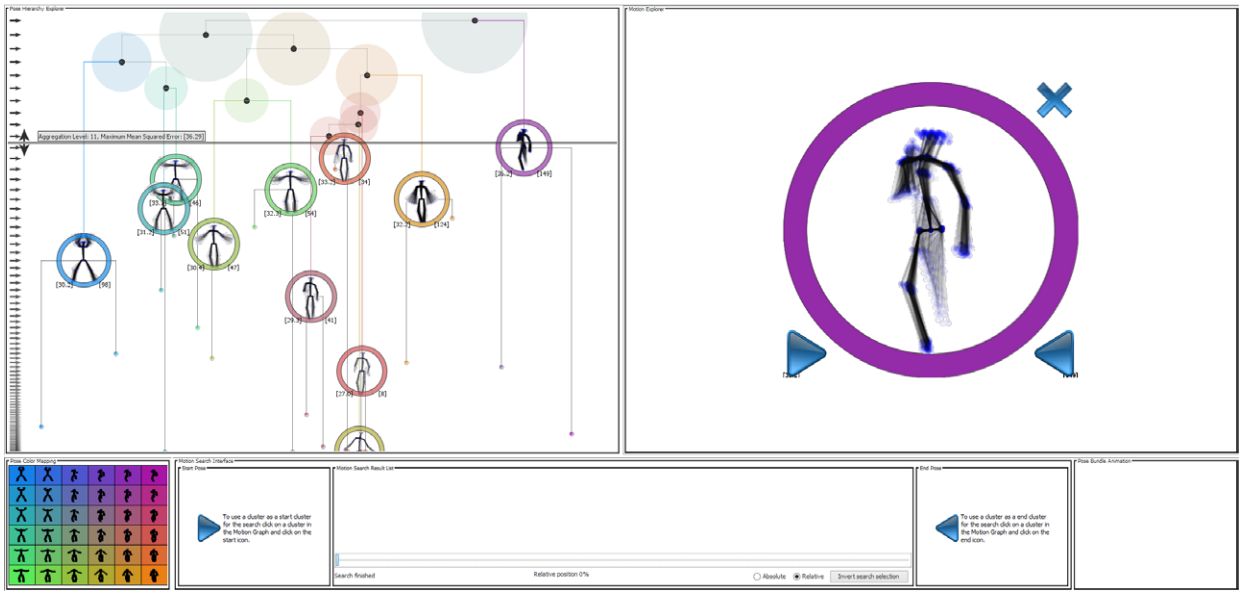


Figure 5.19 *The MotionExplorer system for the ES in human motion capture data presented in the case study in Section 7.2. A tree-based layout (top left) provides an overview of the data content based on a hierarchical clustering result. A global slider control enables users to control the aggregation level. At the current aggregation level 11 clusters of human poses are shown. Changes of the aggregation level are triggered to other components of the ESS. In the illustrative example the user clicked a purple kickboxing pose which was subsequently shown in a Details-on-Demand view (top right). By clicking the blue buttons at the lower left and right of the Details-on-Demand visualization, the selected pose is set as a start (end) pose of the query interface (bottom).*

As an example, the MotionExplorer ESS in Section 7.2 uses a tree-based axis-parallel 2D layout. In this way, the ESS for human motion capture data supports domain experts in gaining an overview of human poses. Figure 5.19 (top left) shows the tree-based hierarchy of poses. Together with domain experts, we decided to provide a hierarchical clustering as an upstream data aggregation step. The k-means algorithm was used to split individual subtrees into two parts. Based on an inquiry of the domain experts, we assigned the splitting order of clusters with respect to the cluster variance. As a consequence, upper branches in the tree structure divide the most heterogeneous sub-structures of the data set while lower branches support for the identification of local differentiations. The tree-based visualization shows the hierarchical clustering results allowing the domain experts to gain an overview of a large number of human poses $\mathbf{RG}_{\text{CBO1}}$. The tree-layout visualization also allows steering the aggregation level. A horizontal control bar defines the number of displayed aggregates in individual subtrees (the aggregation level). As a rule, every node in the tree above the control bar is split into its child nodes while every node below the control bar is shown at a glance. The aggregation level control bar can be dragged by the user along the y-axis to adapt the aggregation level. Changing the aggregation level automatically triggers the underlying data model to adapt the granularity of the displayed clustering result. Different linked views of the system are sensitive to this user interaction. Hence, we provide a means of supporting different levels of abstraction $\mathbf{RG}_{\text{CBO6}}$. To conclude this tree-based layout example, the challenge of choosing the layout technique for a content-based overview was simplified significantly $\mathbf{RG}_{\text{CBO5}}$. The output of the hierarchical clustering algorithm can directly be used for a hierarchical layout.

Projection-Based Solutions Many clustering algorithms do not create additional low-dimensional output structures which can be applied to downstream layout approaches. In this case, projection-based solutions can be applied. Projection-based layouts map the high-dimensional input space to the 2D display space. An overview of relevant projection techniques is presented in Section 5.2.3. Likewise, we briefly introduced the class of projection quality measures to validate design decisions for projections. For content-based overviews a most relevant design criterion is the preservation of structure provided in the high-dimensional input space. The class of linear projection techniques tends to preserve the global structure of a numerical input data set more accurate than non-linear projections. On the contrary, non-linear projections tend to show a better performance in the preservation of local structures. In many cases, the structure of the input space is defined in terms of the distribution of data elements in a high-dimensional

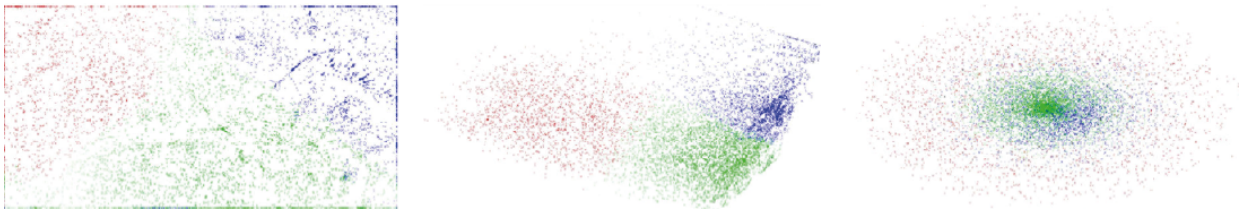


Figure 5.20 *Applicability of different projection techniques for a data set with three class labels. The first technique exploits the display space best (SOM). The second technique preserves the global structure (PCA). With the third technique green and blue class labels are not well separated.*

space or, as a consequence, in pairwise distances between input data elements. In addition, other structural aspects may be due to cluster affiliations and class labels, or on other relations of the data elements, like attached metadata.

The choice of an appropriate projection technique depends on a clearly defined specification of structure and the trade-off in the preservation of local or global structural aspects $\mathbf{RG}_{\text{CBO5}}$. Other design criteria for the choice of the projection techniques rely on the output space. For the evaluation of the output space the preference of the user may be included in the design process $\mathbf{RG}_{\text{CBO8}}$. One design criterion is the exploitation of the available display space, an example is shown in Figure 5.20. The three projection techniques significantly differ in the distribution of the data elements in the output space and in the exploitation of the display space. Similarly, the avoidance of overplotting effects is a subject of projection output comparison. Especially if the projection-based overview is meant to provide glyph designs, the avoidance of overplotting is a most relevant design criterion. The tree projection-based candidates in Figure 5.20 are the product of a prototype-based approach applied in one of our recent publications [BWS*12]. In the usage scenario for projection-based layouts, we explored a five-dimensional data set with ten years of weather observation data measured every three hours. The five measurements are the temperature, the relative humidity, the air pressure, the Shortwave-Downward radiation, and the Longwave-Upward radiation of a weather station in Payerne, Switzerland. With the projection-based layout, we supported various analysis tasks, like the identification of periodic seasonal effects, frequent patterns, and outliers. Consequently, the design criterion to avoid overplotting was less relevant since dense regions in the input space should also be recognizable as such in the display space. To reveal periodic patterns, we were interested in the preservation of the global structure of the high-dimensional input space, and thus chose the linear PCA projection (the center image of Figure 5.20. The projection-based overview of the data content is shown in Figure 5.21. Measurements of ten years are projected into 2D, a path metaphor connects temporally adjacent measurements (here: all selected data elements of the yellow cluster.

Force-Directed Solutions Similar to projection-based solutions force-directed solutions do not necessarily rely on the low-dimensional structure of a clustering output. On the contrary, structural aspects of the (high-dimensional) input data set determine the forces of the layout. For instance, the pairwise relations between nodes (data elements, clusters) serve as a possible basis for the calculation of the layout. As a consequence, these types of force-directed solutions accept the information of distance matrices. Other types of input for the definition of forces between nodes include cluster affiliations and class labels, or other relations of the data elements like attached metadata. This examples of input types have in common that force-directed layouts tend to prefer local structures of the input data set in favor of global structures. The definition of relevant forces and the interplay of different types of attracting and repulsing forces is one of the challenges of force-directed layouts. Likewise, these parameters enable the engagement of users in the design process. One decisive advantage of force-directed layouts as opposed to projection-based approaches is that overplotting can be avoided easily. The reason for this is the definition of (additional) local repulsion forces enabling the designer to steer the degree of overplotting in the layout. Especially if the number of data elements is low, it is important that the layout design considers the reduction of superimposed visual structures. In this way, the size of the glyph design for every included object can be enlarged, and thus the level of detail within the glyph designs be maximized. As a result, force-directed layouts are particularly appropriate if the number of nodes is countable and if the nodes are visualized with large glyph designs $\mathbf{RG}_{\text{CBO5}}$. The trade-off between the size of the glyph designs and the degree of overplotting is another subject of user involvement. Together with the design of the glyphs (cf. Section 5.4) and the specification of the forces, this complements possible user-centered design aspects $\mathbf{RG}_{\text{CBO8}}$. Meeting both design criteria minimizing the degree of overplotting and maximizing the size of the cluster glyphs leads to useful overviews of the data set. For that purpose, force-directed layouts are appropriate, especially if the preservation of

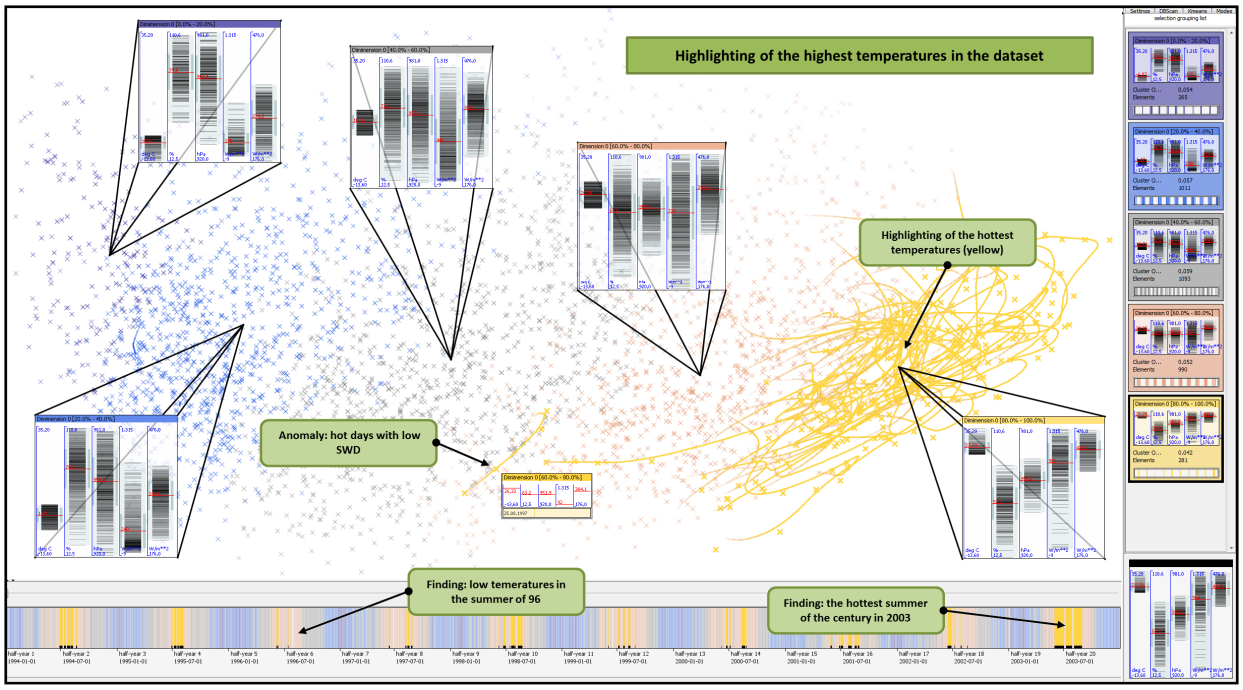


Figure 5.21 *Projection-based overview of a five-dimensional data set of weather observations over ten years. The projection maps every data element to 2D, temporally adjacent data elements are connected with a path metaphor. Users are able to aggregate the data, here, by means of a binning technique. The resulting groups of data are shown in the list-based interface on the right. In this example, the user chose the temperature dimension of the data set (the first dimension) as the binning criterion. Measurements with the hottest 20% percentile are encoded with yellow color; the cluster is currently selected. It can be seen that the projection predominantly aligns the yellow data on the right of the content-based overview. The global time axis at the bottom reveals a seasonal pattern: yellow data points always occur every summer, especially in the summer 2003, which is referred to as "the summer of the century".*

local structure is more relevant than the preservation of global structure. As a result, force-directed layouts also enable to gain an understanding of the provided data content $\mathbf{RG}_{\text{CBO1}}$.

We present an example of a content-based overview based on a force-directed layout. In the associated usage scenario the content of a high-dimensional temporal data collection is related to the additional metadata. For this purpose, a data set with daily temperature curve progressions is aggregated by metadata attributes, like the 'Location' on Earth. In other words not the content is used for the data aggregation but the locations of the measurements on Earth. In the example, the data set is aggregated to 21 clusters each providing the measurements of a single location on Earth. In this way, users are able to facilitate exploratory relation seeking between the data content and additional metadata. In this section, we use the described usage scenario as an example of gaining an overview of the content based on force-directed layouts. In Chapter 6, we present an in-depth description of how a combined analysis of the data content and the attached metadata can facilitate ESS. The LinLog layout algorithm is used as an example of force-directed layouts. The forces of the layout algorithm are based on the pairwise distances of the 21 locations. To this end, we require a measure assessing the distance between any two given locations. The distance measure is based on the data content, i.e., the daily temperature measurements taken at any location. We compare the set of time series FVs of station *a* with the FVs of station *b* and calculate the average linkage distance as known from hierarchical clustering. The Euclidean distance metric is used for the calculation of pairwise distances between high-dimensional data objects. As a result, we reveal pairwise distances of 21 locations on Earth based on the collected time-oriented data. The driving question of this example is whether the force-directed layout will calculate a meaningful overview of the 21 locations on Earth. For this purpose, we conducted the case study together with domain experts from Earth observation. It was interesting to see whether the domain experts would confirm a meaningful structure of 21 locations aligned in 2D by the force-directed layout. For the researchers in the usage scenario it is interesting to see whether the structure of the locations on Earth is related with the structure of the temperature data content. Figure 5.22 shows the output of the layout. A glyph design is provided showing the name of the location at the top and the dominating temperature curve at the top right. A similarity-preserving color coding indicates the similarity of different locations

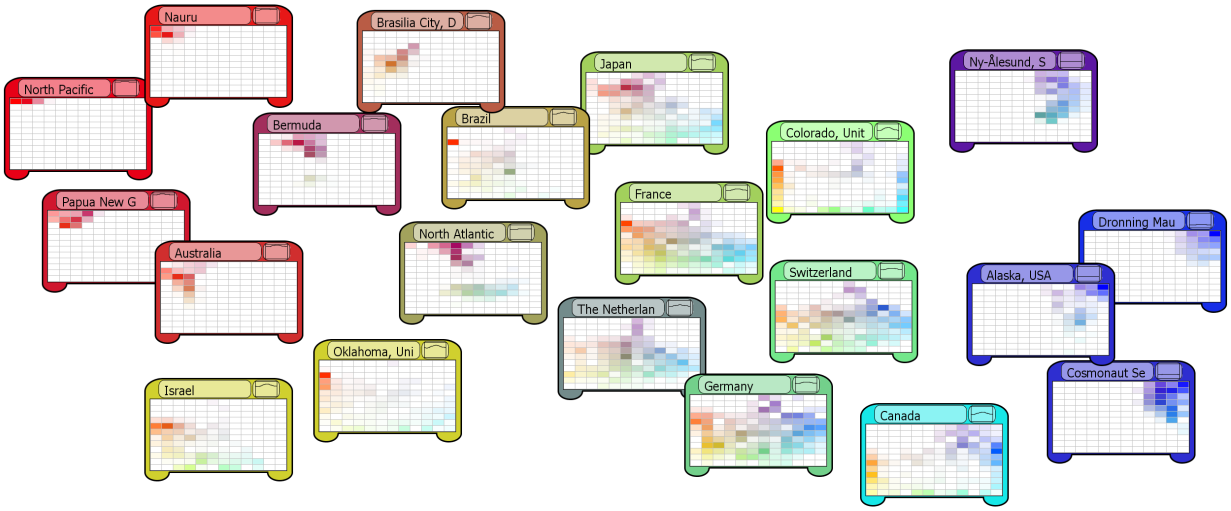


Figure 5.22 *Content-based overview provided with a force-directed layout. The 21 cluster glyphs are aligned in such a way that overplotting is nearly avoided. In this usage scenario the daily progression of absolute temperature patterns is related to the metadata property 'Location' on Earth. Both the data content and the metadata are related with respect to the predominant climate. Hot temperature regions are aligned at the upper left, the coldest temperatures can be identified at locations aligned on the right.*

on Earth based on the distance calculation. The remaining visual encodings will be further described in Chapter 6, in which we investigate research goals of relating data content to metadata in detail. The result of the force-directed layout shows red clusters in a display region at the upper left, where the content of locations on Earth with warm temperature measurements is aligned. On the contrary, the blue clusters at the lower right pose measurements taken at Arctic and Antarctic locations. At the center of the layout a variety of green and yellow clusters indicate locations with moderate temperatures including measurement stations in Europe. The layout reveals that the structure of the temporal data content and the structure of the metadata attribute 'Location' on Earth are related with each other. Such a relation between data content and metadata can serve as a basis to prove existing hypotheses, or to formulate new hypotheses. If an *interesting* relation is identified, researchers are able to use this new information as a starting point for an in-depth analysis. In this connection, the domain experts confirm the usefulness of the structural information provided by the force-directed layout. For in-depth details of the approach, we refer to our corresponding publication, depicting several examples of force-directed layouts including case studies with domain experts [BRS*12a]. A precondition for the usefulness of the display was the parameterization of the force-directed layout in a way that all provided information is visible. The Linlog layout is parameterized in a way that overplotting was mitigated almost completely. In the same way, the glyphs of the 21 locations on Earth are enlarged to an extent that a variety of details within each cluster glyph can be identified. This shows how the force-directed LinLog layout can be used to provide a content-based overview for an ESS.

Facing the Trade-off Between Projection-based and Force-Directed Solutions Many projection-based and force-directed solutions share a common data input. Both solutions are able to use the pairwise distances or relations of high-dimensional data objects. However, the output of the two types of approaches is different, depending on the different classes of underlying algorithms. Most projection-based layouts preserve the global structure of the more accurately than force-directed layouts. On the contrary, force-directed layouts are better suited to optimize local layout criteria, like the avoidance of overplotting. In addition, force-directed layouts can be provided with additional forces to optimize the output of the layout. As an example, an additional postprocessing step may ensure that all nodes of the layout are keeping a minimum distance to the borders of the display. In this section, we show how projection-based and force-directed solutions can be combined to benefit from the strength of both individual solutions. We suggest to execute a projection-based layout to achieve a global order of the data set in 2D, and thus a meaningful initialization of the layout. In an additional step, a force-directed solution optimizes the layout based on individual design criteria. These design criteria can be defined and weighted by the data scientist, or by the involved user $\mathbf{RG}_{\text{CBO8}}$. As a result, the best combined solution can be achieved $\mathbf{RG}_{\text{CBO5}}$ and a trade-off between the different model parameters be faced $\mathbf{RG}_{\text{CBO3}}$.

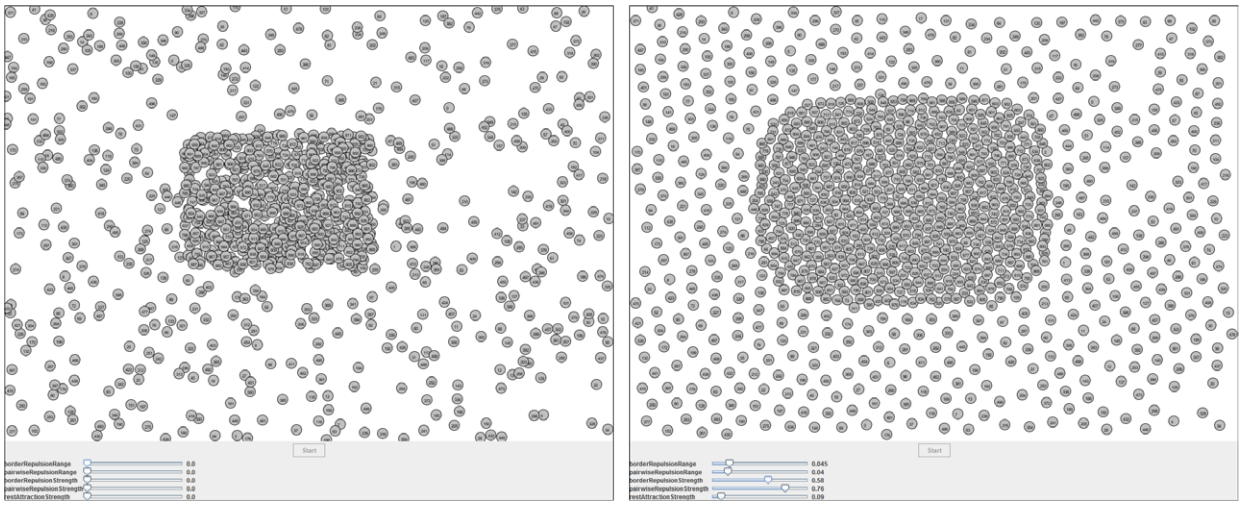


Figure 5.23 *Trade-off between different conflicting quality goals in the layout design. A force-directed layout is applied as a postprocessing routine of a projection result (left). As a result, the number of overplotting data objects is reduced significantly. However, the structure of the projection is disrupted. Quality measures for each of the quality goals an visual-interactive tools for steering the forces of the layout may help to face the trade-off.*

An illustrative example is shown in Figure 5.23. A synthetic data set with 1,000 high-dimensional data elements is used. The data set combines randomly distributed elements along the entire display with data elements aligned in a small rectangular shape in a hyperplane. The data set is projected in 2D yielding a rectangular area at the center of the layout with a high density of data objects (left image). The structure of the high-dimensional data set is well preserved in 2D. However, the data elements within the rectangular region of the layout are highly overplotted while most remaining regions of the display are only sparsely distributed. Another disruptive fact is that some of the projected data elements are aligned at the border of the display, meaning that large parts of the circular glyphs are outside the display. With the additional force-directed layout step, forces can be defined to a) reduce overplotting, b) clear the display borders, and c) preserve the structure provided by the projection algorithm. The forces for the preservation of the structure ensure that the positions of the data projection in 2D are adhered. In other words, the data scientist is able to steer the influence of additional forces. In the right image of Figure 5.23, the additional force-directed layout is executed. The layout of the data elements has changed significantly. The degree of overplotted data elements has decreased significantly and the data elements at the borders have moved inwards the display. However, the changes of the positions also led to a reduced preservation of the structure of the high-dimensional data input. For every design criterion a numeric quality measure can be provided to assess the impact of changes made by the force-directed layout. Both the data scientist and the user can steer the force-directed optimization process of the layout in a visual-interactive way. In this way, it is possible to face the trade-off between different design criteria for the layout of data elements in 2D and to define the most appropriate model parameters $\mathbf{RG}_{\text{CBO3}}$. In consequence, the benefits of a projection-based and a force-based solution can be combined in a user-centered way $\mathbf{RG}_{\text{CBO5}}$ $\mathbf{RG}_{\text{CBO8}}$.

5.5.2. Different Levels of Abstraction

We face the challenge of providing different levels of data abstraction (LOA) in content-based overviews $\mathbf{RG}_{\text{CBO6}}$. The possibility to present the data at different LOA supports the Information-Seeking Mantra (“Overview first, zoom and filter, then details-on-demand”) presented by Ben Shneiderman [Shn96]. Subsequently, we show how different visualization and interaction designs can support users in browsing at different LOA. Both the visualization and the interaction design is carried out in the design phase. User interaction, however, is executed in the application phase of ESS. In the course of the user-centered design phase, we recommend to determine model parameters whose values can be fixed. Fixed parameters reduce the number of visualization controls, and thus the complexity of the ESS’ interface. However, some parameters need to be user-steerable in the application phase of the system. One of the reasons for providing users with steering capability is the involvement of VA capability in the application phase of the ESS. Different factors, such as the data, the user, and the task reveal whether parameters are fixed in the *design phase* or are made steerable in the *application phase*. Depending on these influencing factors, different LOA solutions can be

Figure 5.24 *The result of a hierarchical clustering is shown in a tree-based visualization. Domain experts can change the level of abstraction by dragging the aggregation level slider. Here, poses of skiing movement are drilled-down to the aggregation levels of 5, 15, and 30.*

possible. In the following, we recall four main solutions that resolve the challenge of providing different levels of data abstraction (LOA) in content-based overviews $\mathbf{RG}_{\text{CBO6}}$.

- (1) Interactive steering of the clustering algorithm to change the aggregation level
- (2) Zooming to enable the interactive drill-down to local aspects within the layout
- (3) Semantic zooming to show more or less details of the same object (level of detail)
- (4) Overview+context and Micro-Macro views showing both abstracted and detailed information

Interactive Steering of the Clustering Algorithm LOA concepts based on interactively steering the clustering algorithm are at heart of VA. In this case, users are able to change model parameters in the application phase of the content-based overview. As a result, a new clustering result is calculated and subsequently visualized in 2D based on the included layout strategy. The probably most relevant steering parameter the number of desired clusters which has a direct effect on the provided LOA strategy. One challenging task for LOA strategies based on interactively steering the clustering algorithm is the significant change of the visual clustering result which may be disturbing for the user. For instance, the partitioning-based k-means algorithm calculates an entirely different clustering result if the number of clusters k is adapted. As a consequence, users are required to familiarize with the new cluster layout each time the model parameter is adapted. This is why we recommend data aggregation techniques preserving large parts of the cluster information. In fact, these techniques may be applicable for the local adaption of clusters. A particularly suitable class of aggregation techniques is hierarchical clustering. In this case, the subtrees of the cluster tree can be adapted while the remaining cluster hierarchy remains unchanged. We refer to the work of Elmqvist and Fekete for an in-depth overview of visualization and interaction designs for hierarchical data aggregation [EF10].

We present an example of a LOA strategy based on interactive steering of the clustering algorithm. The corresponding usage scenario is presented in Figure 5.24. The figure shows one of the two content-based overview visualizations provided in the MotionExplorer case study (cf. Section 7.2). In the course of the design phase, the decision was made to provide an interactive steering parameter enabling the domain experts to adapt the LOA in the application phase. In Figure 5.24, a tree-based layout shows the different LOA of a hierarchical clustering result of human poses. A horizontal bar enables the involved user group to steer the aggregation level by dragging the control bar to different vertical positions. Poses of skiing movement are drilled-down to the levels of 5, 15, and 30 data aggregates, respectively. This supports domain experts in gaining an overview of the underlying data collection $\mathbf{RG}_{\text{CBO1}}$ by interactively steering the LOA $\mathbf{RG}_{\text{CBO6}}$. This usage scenario serves as an example of solution (1) carried out in the application phase of the ESS. This example also shows how user-centered design can be carried out in the design phase to generate useful visual interfaces $\mathbf{RG}_{\text{CBO8}}$.

Zooming to Enable the Interactive Drill-down Solution (2) postulates interactively zooming in content-based layouts. Zooming is one of the most prominent interaction techniques, especially in of map-based visual interfaces. With the layout step of the reference workflow, we also propose spatial distributions of the data, also referred to as ‘information landscapes’. The layout can be created based on a cluster structure, a projection, or on force-direction. The structure of the spatial distribution of content-based overviews directly depends on high-dimensional data properties.

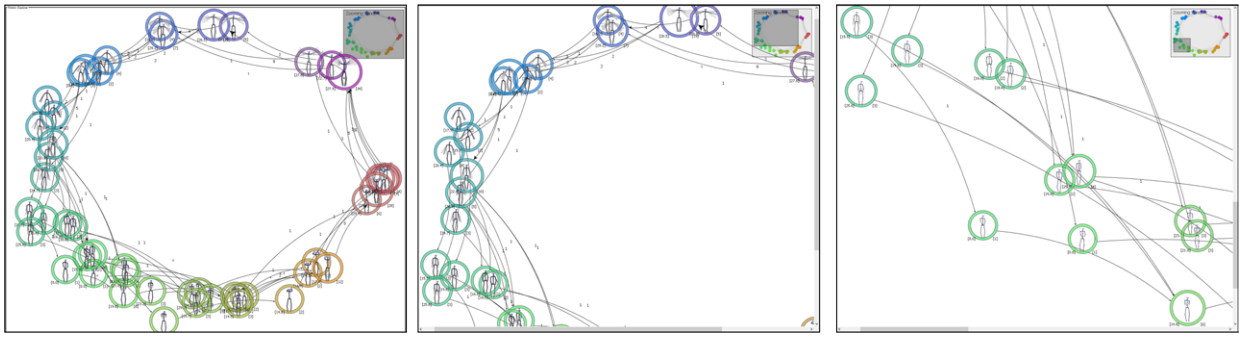


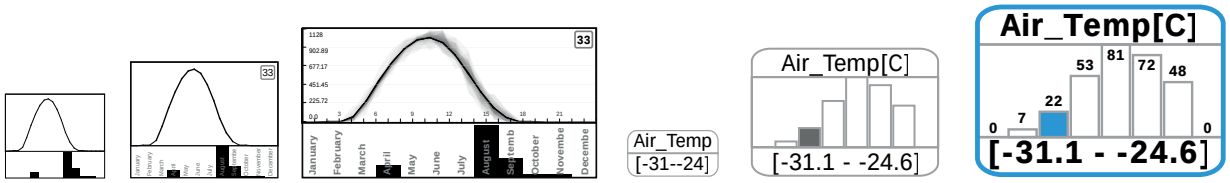
Figure 5.25 *Overview of motion sequences provided by a node-link diagram with 50 clusters. A rotating-arms motion is shown in multiple repetitions. However, the amount of 50 clusters leads to overplotting, hampering the identification of local structures. For this purpose the node-link diagram supports zooming and panning interactions. At the upper right of the diagram a small overview display shows the data at a glance including the current viewport of the diagram. Filtering can be applied to reduce the number of displayed nodes to the set of relevant ones. Style variations in human motion can be explored in detail.*

In this connection, many content-based layout solutions are appropriate for zooming interaction as a means for LOA. Zooming interaction enables users to drill down to local aspects of the data set. Solution (2) not only factors details of data aggregates (nodes), but also local relations between nodes. The distance of aggregates in the layout yields information about the strength of the relation. For multivariate time-oriented data, path and edge metaphors can additionally be used to represent the temporal domain (cf. Figure 5.21). Drill-down interaction enables users to identify these relations in more detail.

Figure 5.25 shows an example with a projection-based layout strategy. This example exposes the other content-based overview presented in the MotionExplorer ESS (cf. Section 7.2). An overview of human motion capture data is represented as a node-link diagram. Every node shows a cluster of human poses while the edges between nodes indicate the human motion sequences. A small overview window at the upper right of the node-link diagram shows the node-link structure at a glance. Zooming (and panning) interaction in the diagram triggers an adaption in the current viewport shown in the overview window. In each of the three images 50 human poses are shown. The user executed zooming and panning interaction to reveal local aspects of the data set. The chosen zooming-levels enable the visualization of the most data aggregates without overplotting. Taking the example of the MotionExplorer case study, we demonstrate the usefulness of zooming as a means for steering the LOA $\mathbf{RG}_{\text{CBO6}}$.

Semantic Zooming to show more or less Details Semantic zooming (3) allows objects to be represented differently at different scales [CKB09]. In this way, we emphasize semantic zooming as type of LOA appropriate for content-based overviews. Apart from the layout step, we focus on the visual mapping step of the reference workflow, necessary for visually encoding high-dimensional data objects (cf. Section 5.4). Especially glyph designs combining the information of clusters and data elements are a beneficial means for providing semantic zooming. A limiting factor in the glyph design is the available display space, depending on the display size and the layout. In any case, if the user applies semantic zooming the available display space of individual glyphs increases significantly. These different scales allow glyphs to be represented differently. To facilitate LOA with semantic zooming, we advocate the design of glyphs at multiple levels of detail. The more display space is provided, the more details can be released in the visual data representation.

We present two examples of semantic zooming in Figure 5.26. In each of these examples a cluster glyph is shown at different scales. Depending on the size, more or less detailed information is shown. At the highest resolution both examples pose a variety of visual encodings in a single cluster glyph. On the left, the cluster of a daily radiation pattern is shown in combination with an associated metadata attribute. The temporal information is visualized with a linechart visualization, the metadata attribute ‘Month’ is shown with a barchart. The combination of these two data types in a single visual encoding enables users to relate the time-oriented data content with the metadata attribute for the given cluster. However, at small scale only a few abstract visual encodings can be shown. The displayed level of detail can be increased with the increasing resolution, successively. At the largest semantic zooming level, users have the means to identify the value domain of the daily time series pattern on the y-axis and the hour within the day on the x-axis. In addition, the bundling technique is applied showing the cluster prototype *and* the multitude of associated



(a) *A visualization of a cluster with daily time series patterns in combination with a metadata attribute depicted as a bar chart.*

(b) *Visualization of the distribution of the attribute air temperature in degrees the second bin is emphasized.*

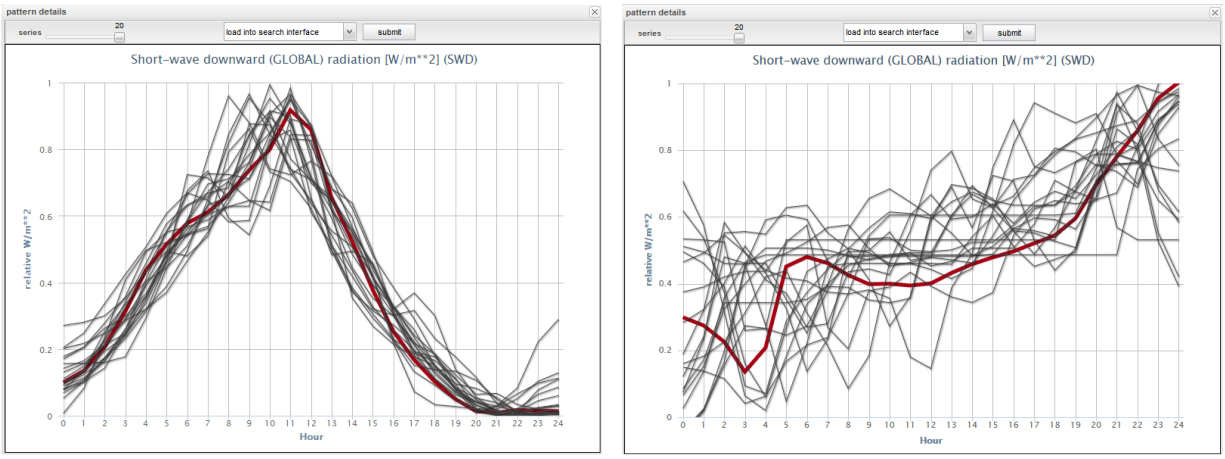
Figure 5.26 *Two examples of semantic zooming. The size of a cluster glyph determines the level of abstraction of the visual representation. With the increasing size the glyphs present change in the stroke and text sites. In addition, more details are presented, such as the bundle visualization of the time series cluster (left).*

time series patterns. Finally, label information in the barchart reduces the interpretation effort from an identification task along the x-axis to a simple lookup task. As a result, this semantic zooming level supports users in complex analysis tasks, such as seeking relations between the data content and additional metadata. The example glyph of the left example provides the insight that the radiation pattern with peak in the value domain of ĩ,000 before noon time is predominantly measured in August. The example on the right shows the metadata attribute 'Air_Temperature in C'. More precisely, it emphasizes the bin with the distribution in the value domain between '-31.1 to -24.6'. If the scale of the glyph is low, only the labels of metadata and the bin name are shown. With the increase of the zooming level the glyph presents additional details, such as the allocation of the highlighted bin within the value domain. The final level of the semantic zoom allows marking the numbers observations for each bin.

To summarize, semantic zooming is a powerful interaction technique to facilitate different LOA $\mathbf{RG}_{\text{CBO6}}$. Provided that a meaningful visual mapping and layout is designed, users are supported in gaining an understanding of the data set $\mathbf{RG}_{\text{CBO1}}$. Enhanced analysis tasks can be addressed, such as comparison and relation-seeking tasks, which particularly supports exploration activity. For both the visual mapping step and the layout step, we presented guidelines and techniques enabling the involvement of users in the design. Hence, useful semantic zooming techniques may be designed together with the user, which can subsequently applied in the content-based overview $\mathbf{RG}_{\text{CBO8}}$.

Overview+Context and Micro-Macro Views In the final LOA strategy for content-based overviews, we combine overview+context and Micro-Macro views. Both techniques have in common that they provide an overview of the data collection which makes them highly appropriate for the use in content-based overviews. In addition, both techniques present an additional detailed view of the data collection. The difference of the both techniques is that overview+content prefers distinct presentation spaces (juxtaposition) [CKB09] while Micro-Macro views are based on an overlay of the two layers (superposition) [Tuf90]. The combination of both an overview and local details makes the two techniques particularly appropriate as an LOA concept. Prominent examples of overview+context are (geographical) map interfaces, allowing users to drill down to local 'spaces' while an overview map in an additional view shows the local space in combination with the global context. In Figure 5.25, we used overview-context for the visual representation of human motion sequences in 2D. Users are able to zoom into the content-based overview to identify local aspects of interest. The small overview display at the upper right of the view shows the data at a glance including an indication of the current local viewport. An example of Micro-Macro view was presented in Figure 5.9 the section on quality-driven visual-interactive cluster analysis (cf. Section 5.3). In three images, we showed how an abstract cluster layout (macro) can be combined with a fine-gained layer of data points (micro). We showed micro layers based on a scatter-based view (left), a density-based view (center), and a star view (right).

A real-world example of the Micro-Macro views technique is incorporated in Figure 5.21 which was introduced in the guideline on projection-based layouts in Section 5.5.1. We demonstrated how a 5-dimensional Earth observation data set is aligned in 2D with a PCA projection. Every data point is visualized at a specific display coordinate in a similarity-preserving way. The user is able to aggregate the data collection while using the visual-interactive system. The aggregation results are visualized in the projection view as an additional macro-layer. The display coordinate of the aggregates is calculated also calculated in a similarity-preserving way, based on the vector information of the centroids. The different LOA consisting of an abstract layer with clusters and a fine-gained layer of associated data elements support users in carrying out identification and comparison tasks. In this way, the distribution of the data points, the density information, and the distribution of clusters can be used to gain an understanding of the underlying data set $\mathbf{RG}_{\text{CBO1}}$.



(a) *An example pattern with a peak radiation at about 11 a.m.* **(b)** *An example pattern with a linear upward trend.*

Figure 5.27 *In the VisInfo case study the Details-on-Demand visualization of the content-based overview is used as the visual interface for the formulation of a content-based Query-by-Example. The user is empowered to further adapt the query pattern in an additional search interface (cf. Figure 5.28). Finally, the pattern is executed by the retrieval algorithm. The comparison of the two illustrative examples reveals that the variation of data elements with respect to the red cluster centroids is not always the same. The number of linear upward trends seems to be small in comparison to the pattern on the left.*

In summary, overview+context and Micro-Macro views are beneficial to support users in the analysis of different LOA. For the design of these techniques different models and parameters have to be defined starting in the data aggregation step, the visual mapping step, and the layout step of the reference workflow (cf. Figure 5.1). In this chapter about content-based overviews, we presented guidelines and techniques for designing solutions for all of these steps in the workflow, by involvement of the user $\mathbf{RG}_{\text{CBO8}}$. As a result, both overview+context and Micro-Macro can be designed as LOA techniques within the content-based overview $\mathbf{RG}_{\text{CBO6}}$.

5.5.3. Visual Querying

With the successful completion of the layout step in the reference workflow, the content-based overview visualization can be applied, e.g., in an ESS. The layout allows gaining an overview of the data content. Glyph designs can be carried out to represent the high-dimensional data elements and data aggregates visually. The content-based overview visualization may facilitate different levels of abstraction, possibly by means of user interaction. The final research goal which can be resolved with content-based overviews is visual querying $\mathbf{RG}_{\text{CBO7}}$. We propagate the utilization of visual cluster representations for querying by example. Content-based overviews not only provide an overview of the data set, they can also serve as a baseline for the visual query formulation by example. As such, with the Query-by-Example technique the strength of content-based overviews and content-based retrieval can be combined.

Query-by-Example is a query technique enabling users to select a data element (or data aggregate) of interest and use the data element as a visual query. We recommend combining the Query-by-Example technique with the Details-on-Demand concept suggested by the Information-Seeking Mantra [Shn96]. Details-on-Demand solutions can, e.g., be achieved with strategies providing different LOA. LOA concepts coupled with meaningful interaction designs facilitate the drill-down to local aspects of the data set. Finally, these local aspects can be analyzed in detail, on demand. We have already recommended the incorporation of glyphs as a means of showing high-dimensional objects in detail. Similarly, glyphs are greatly beneficial to submit queries by example. For time-oriented data, two different types of queries are conceivable. First, a single item series pattern may be identified and be used as a meaningful query object. The user expects a search result where a set of data elements similar to the queried pattern are retrieved. For that purpose, the pattern is transformed into the FV space and the retrieval algorithm is executed. Second, the user may be interested in the retrieval of time series (sub-)sequences possibly consisting of multiple time series patterns. In this occasion an interval query needs to be defined visually. One possibility is the definition of a start and an end

Figure 5.28 *Search interface of the VisInfo ESS facilitating visually querying by sketch. The left example pattern shown in Figure 5.27 was loaded into the interface. The user made some adoptions in the evening hours (a bump). Sketching a visual query from-scratch is also possible.*

pattern to specify the subsequences to be retrieved. The visual query interface then provides two slots which need to be set before a query is executed by example.

Visually Querying Time Series Patterns The VisInfo case study in Section 7.1 applies the single pattern-based querying approach. FVs for daily curve progressions are presented in a SOM-based content-based overview (cf. Figure 5.18). By clicking a single cell a Details-on-Demand view appears at the center of the ESS. The detail-in-demand view is presented in Figure 5.27. The glyph design consists of a red linechart representing the curve progression of the SOM cell. In addition, the web-based interface enables users to visualize up to 20 data elements of the cluster, visually represented by black linecharts. At the top of the Details-on-Demand view a control element allows loading the pattern into the query editor. The query editor is shown in Figure 5.28. A linechart is chosen for the visualization of the query pattern in detail, including axis labels for both the temporal and the value domain. Moreover, the editor enables users to adapt the query pattern. This query technique is called Query-by-Sketch and is particularly suitable if users have a clear notion of the content-based query they want to execute. Accordingly, Query-by-Sketch and Query-by-Example are complementing visual query techniques. Both techniques enhance the search capability of the content-based overview. The example of the query editor demonstrated in Figure 5.28 shows an example pattern selected from the SOM-based overview (cf. Figure 5.27 (left)). However, for the sake of plausibility, we used the query editor to modify the value domain of the daily pattern in the hours 20 and 21. The search scenario comprises a relative context, meaning that the value domain of the time series is between 0% and 100%. The result of the retrieval will yield patterns preserving the shape of the curve whatever the absolute values of the value domain.

Visually Querying Pattern Sequences The MotionExplorer case study uses the Query-by-Example concept for the retrieval of human motion sequences. The MotionExplorer ESS is described in Section 7.2. With a click on pose aggregates in one of the two content-based overviews, a Details-on-Demand view is presented to the user. The Details-on-Demand view is illustrated in Figure 5.19 showing a purple kickboxing pose. An enlarged pose cluster glyph is shown at full level-of-detail (top right of the MotionExplorer ESS). To facilitate the query formulation for subsequence search the enlarged human pose can be set as the start pose or as the end pose. For that purpose, two blue buttons are provided in the level-of-detail view. In Figure 5.19 both the start pose and the end pose of the search interface are still undefined. In Figure 5.29 the start pose is set with the orange pose representing an upright standing human pose. For the end pose the purple kickboxing body configuration shown in Figure 5.19 is used. The retrieval result is visualized in between the two query poses. The data set contains six kickboxing motions which can be explored by the domain expert in detail.

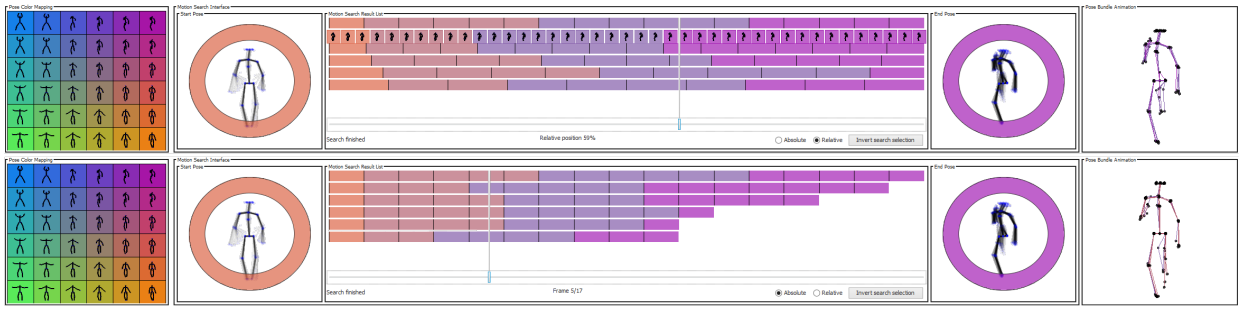


Figure 5.29 *The search interface of the MotionExplorer system presented in Section 7.2. The user is able to define a start and an end pose. As a result, the retrieval algorithm of the MotionExplorer system is executed revealing a set of human motion sequences. The visualization and interaction design of MotionExplorer allows exploring the result set in detail. Users can switch between an absolute time axis meaning that the temporal domain of every retrieved sequence has an equal length or relative time axis where all sequences are stretched to the full length of the result visualization.*

5.6. Summary

5.6.1. Discussion

Closing the Feedback Loop In Section 5.3, we presented novel techniques to enhance the cluster analysis process. In particular, our techniques allow the visual quality assessment of clustering results on different levels of abstraction. The system enables data scientists to validate clustering results by visual comparison, e.g., with other clustering results. While the comparison of clustering results was an important obstacle to overcome [SS02], our techniques currently do not replay the feedback of analysts gained from the visual comparison. With the information gained from our techniques, analysts still select models and model parameters by themselves. A promising extension of our techniques would consider labeling information provided by analysts to improve the model and parameter choice. *Machine Learning* (hereafter, ML) techniques, such as semi-supervised learning or active learning, could be integrated in the process to *learn* the preference information of analysts, and thus to anticipate models and model parameters generating appropriate clustering results.

Enhanced User Guidance A quite similar issue refers to enhanced user guidance concepts [SS02]. With our guidelines and techniques, we presented various ways to guide users through the huge design space relevant to design content-based overviews. However, the idea behind user guidance leaves several potentials for other points of emphasis. Complementary to the idea of closing the feedback loop, the system could autonomously suggest model and parameter choices. The precondition for this concept would require the automation of the cluster analysis process, e.g., in combination with the quality measures postulated and applied with our techniques. As such, this enhanced guidance concept could also facilitate the choice of meaningful layout strategies (cf. Section 5.4). It remains subject to future work, whether such guidance concepts would be useful to support analysts in the cluster analysis process.

Glyph Design In Section 5.4, we presented a guideline for the design of glyphs for high-dimensional data aggregates including associated data elements. One of the criteria directly considered the visual representation of time-oriented data. Obviously, this criterion could be adapted towards other types of data. In general, glyph design is a research field on its own allowing a variety of other solutions, e.g., besides our eleven glyph designs presented in Table 5.1. From a conceptual perspective, it would be interesting to evaluate the intuitiveness and effectiveness of other combinations of semiotic systems and models [Ber83]. Other visual variables could be included in the design process, the summarization presented by Borgo et al. [BKC*13] provides an overview (location, size, color hue, color value, color saturation, orientation, grain, arrangement, shape, fuzziness, transparency).

Uncertainty Understanding and visualizing uncertainties is an important subject to future research in many fields involved in data analysis [KKEM10]. With regard to models for data aggregation and layouts, uncertainty can be characterized in different ways. First, parameters of given models induce different results, leaving analysts uncertain which result solves a problem best. Second, the same applies to different models included in the process. Finally,

the data contributes to uncertainty effects, e.g., by different extents of quality. With our techniques, we postulated quality measures as a means of uncertainty assessment. We demonstrated that visual quality assessment is highly appropriate to validate the usefulness of models and model parameters. However, our techniques could be extended to depict uncertainty effects even more explicitly. What works well for the comparison of different parameter values could also be extended for the comparison of multiple model outputs. In this way, the combinations of uncertainty posed by the included data and the applied models would come to light more clearly. Similarly, it would be interesting to investigate the output of models exposing uncertainty as a means of their nature. Fuzzy clustering is a welcome example, generating clustering results with a probability distribution for the assignment of any given data element to any cluster.

5.6.2. Conclusion

In this chapter, we presented guidelines and techniques for the design of content-based overviews. According to the reference workflow for the design and application of ESS (cf. Figure 5.1), we split the design space into three steps. First, we showed how quality-driven visual-interactive cluster analysis can facilitate the *data aggregation* step. With the data aggregation step, we reduced the number of data elements to a meaningful amount. Basically, we incorporated two different types of external feedback for the creation of enhanced data aggregates. On the one hand, we presented guidelines and techniques for the involvement of the users in the data aggregation step. Thus, domain knowledge and preference information can now be included in the design process. On the other hand, we showed how visual cluster quality assessment strategies can ensure the validity of data aggregates. Second, for the *visual mapping* step of the reference workflow, we showed how high-dimensional data elements and data aggregates can visually be represented in the display space. For the design of content-based overviews for ESS, our main focus was on glyph design and on the use of color as a similarity-preserving visual variable. We presented a guideline for the user-centered design of visual cluster representations and pointed out eleven examples demonstrating how the guideline can be implemented in real-world case studies. For the use of color as a similarity-preserving variable, we presented a set of quality criteria and mapped the criteria to ES tasks. Finally, we presented a list of most appropriate static 2D colormaps. Third, for the *view transformation* step of the reference workflow, we addressed three different factors. We showed how different outputs of the data aggregation step can be used for the design of content-based layouts in 2D. Hence, high-dimensional objects can be represented in the low-dimensional display space in a meaningful way. Next, we presented four different solutions to support users in interacting at different levels of data abstraction. Finally, we showed how content-based overviews can be used as a basis for visual querying.

With the content-based access to time-oriented primary data $\mathbf{C}_{\text{CBA}}$ (cf. Chapter 4) and the novel methods for providing content-based overviews $\mathbf{C}_{\text{CBO}}$, data scientists are able to elaborate the complete reference workflow for the design and application of ESS (cf. Figure 5.1). With content-based overviews and visual-interactive querying techniques, we provide two most relevant visual interfaces for enhanced ESS. In addition, our techniques can serve as a baseline for the design of visual interfaces for yet another important class of analysis tasks to enhance future ESS: seeking relations between data content and metadata. However, relation seeking between data content and metadata yields individual challenges that need to be overcome $\mathbf{C}_{\text{C+M}}$. In the next chapter, we present three novel techniques, all addressing this type of relation-seeking task. The contributions will be based on the guidelines and techniques for the design of content-based overviews.

CHAPTER 6

Relation Seeking Between Data Content and Metadata

> *As the number of documents grows larger, however, it becomes increasingly difficult for an investigator to track the connections between data and make sense of it all. The sheer number of entities involved may make it very difficult for a person to form a clear understanding of the underlying concepts and relationships in the document collection.*
>
> according to Stasko et al. [SGL08], *2008*

In the Chapters 4 and 5, we presented guidelines and techniques for providing content-based access (cf. $\mathbf{C}_{\text{CBA}}$) and content-based overviews (cf. $\mathbf{C}_{\text{CBO}}$) to time-oriented primary data. However, data content poses only one part of the available information in many document collections. In this chapter, we include metadata within the analytical process. ES can greatly benefit from solutions providing the combined search *and* exploration of data content *and* metadata. This enables users to identify *relations* between these heterogeneous types of data. The *relation space* forms a new basis for exploratory data analysis, and thus for the design of visual-interactive interfaces, e.g., for the use in enhanced ESS. The exploration of relations facilitates both the formulation and the validation of hypotheses. To this end, the *interestingness* of relations plays a key role as the number of relations may become huge for large and complex data sets. Visualization and interaction designs from IV and guidance concepts borrowed from VA can be applied to lead users to the most interesting relations. The guided exploration of interesting relations can further enhance ESS, e.g., in combination with content-based access strategies and content-based overviews. Throughout this chapter, we present techniques for the relation seeking between data content and metadata. At a core level, our contributions can be differentiated as follows. First, we present techniques for relating metadata attributes to content-based layouts. Second, we present techniques for relating data content to metadata-based layouts. Finally, we present techniques for relating multiple (mixed) data attributes, consisting of both data content and metadata, to each other. This chapter is mainly based on [BBF*10, BBF*11, BRS*12b, BRS*12a, BSW*14] and partly based on [SBS11, BSR*14, SBM*14, SBSK15].

Contents

6.1. Introduction

6.1.1. Motivation

In Sections 2.2.1 and 2.3, we characterized primary data and time-oriented data. Based on the widely accepted assumption that primary data contains undiscovered knowledge, one may have a presentiment of the value of (time-oriented) primary data for data-driven research. Specific benefits of primary data are the large number of different sources yielding a variety of opportunities for analytical and collaborative approaches. In addition, the life-cycle of primary data gives an idea of the different stakeholders, applications, and exploitation processes. In the following, we further investigate the two most relevant types of data provided with primary data. The *data content* on the one hand contains primary information about collected experiments, interviews, observations, simulations, or phenomena. In Chapters 4 and 5, we have already presented solutions for the content-based access to time-oriented primary data and for designing content-based overviews. On the other hand, virtually any primary data collection provides valuable *metadata*. This metadata may be attached in the data creation phase, e.g., to store different experiment conditions. Additional metadata may be attached in preprocessing and analysis phases, e.g., to store statistical information or provenance information about transformations applied to the data. Finally, in the data preservation, access, and reuse phases, additional metadata may be attached to primary data to facilitate DL support for example.

To underline this point, it is generally accepted that most bodies of information consisting of both primary data content and metadata yield valuable but undiscovered knowledge, which is especially relevant for data-driven research (cf. Section 2.2). A beneficial advantage can be expected by using both data content and metadata in a joint approach. Domain experts will then be able to test hypotheses, e.g., to assess the effect of different experiment conditions. In addition, novel analysis techniques will enable domain experts to explore the complete body of information, e.g., by seeking new hypotheses. For example, for complex phenomena it will then be possible to identify previously undiscovered causes and effects. However, the question arises of how data content and metadata can be combined in a meaningful way. In particular, confronting the challenge of relation seeking between data content and metadata is uncharted land for the great majority of existing ESS. In this connection, we briefly echo the shortcomings of exploratory data analysis approaches and search systems, which are to some extent opposed to each other (cf. Section 2.1.2).

Exploratory data analysis provides valuable tools helping analysts to understand the content of large data sets, e.g., in terms of representative clusters. Beyond data content, information seekers are also interested in the relationships between data content and metadata. Revealing these relations may lead to new insights and undiscovered knowledge. The need of generating and validating hypotheses based on metadata shifts the attached metadata towards the exploratory analysis process. In this connection, relation seeking is a among the most relevant tasks in the analysis of time-oriented data (cf. Section 2.3.2). However, adding metadata into the exploratory analysis process is challenging. A particular challenge is the combined analysis of these two data types posing entirely different characteristics. Novel techniques are required to facilitate enhanced relation seeking between data content and metadata.

Search systems, such as DLs, incorporate (textual) metadata into the search and retrieval process for a long time. Especially for document types with complex data content, the access to such documents is typically provided solely with the aid of metadata attributes. Searchers are able to query for attributes like authors, titles, publishers, and other types of metadata for a targeted document collection. Similarly, faceted search techniques enable searchers to drill down in the document collection by selecting metadata entities of interest. For that purpose, individual hand-curated metadata is made visually accessible to the user in terms of facets. However, search systems often fall short in incorporating non-textual content in the search and exploration process. Content-based access strategies and downstream analysis techniques for the data content are required to enrich the capability of today's search systems.

Both exploratory data analysis and search systems motivate the *combined analysis of data content and metadata*. Similarly, we support the idea that future ES approaches should incorporate data content *and* explanatory metadata. It is desirable to support users in validating hypotheses and in providing exploratory means to formulate new hypotheses on the complete body of available information. Appropriate overview and navigation techniques for data content and metadata will allow users to obtain a better understanding of the overall data space, before narrowing down the search to more specific queries. Metadata-based techniques, like faceted search approaches, have proven to be very useful in the ES process and may be extended by content-based access. In return, content-based overviews may be equipped with interesting relations to additional metadata attributes. Both techniques relating metadata to data content and relating data content to metadata are promising concepts to further enhance ES concepts.

However, the combined exploration of heterogeneous data types consisting of data content and attached metadata remains difficult and the number of best-practice solutions is scarce. Different concepts exist for relating heterogeneous types of data to each other. The applicability of such solutions and the interestingness of the revealed relations depends on the data, the user, and the analysis task. One of the most fundamental factors regards the question whether a relation may be defined in a meaningful way. In this connection, it must be pointed out that an interesting relation does not necessarily apply to the entire data set. In turn, many relations regard a data subset which has something specific in common, often referred to as a cohort, or a bin. Building on the definition of a relation, it remains a challenge to shed light on what makes a relation interesting for the user. Yet another concern regards the question of how users can be guided towards these interesting relations. Especially for large and complex data sets the number of relations (the relation space) is huge. An essential component of the challenge to provide overviews of both the data space *and* the relation space. In addition, the question arises to what extent these approaches allow the user to participate in the design. Finally, appropriate solutions may support the visual communication of findings.

6.1.2. Research Goals

The related work revealed six major research challenges (cf. Section 2.6) which we will address in this thesis. In this chapter, we explicitly resolve the challenges of combining data content and metadata $\mathbf{C}_{\text{C+M}}$. Furthermore, we face the associated challenges of choosing appropriate models and model parameters $\mathbf{C}_{\text{MPC}}$ and involving the user in the design $\mathbf{C}_{\text{UCD}}$. To this effect, we reconsider the upstream challenges of providing content-based access to time-oriented primary data $\mathbf{C}_{\text{CBA}}$ and content-based overviews $\mathbf{C}_{\text{CBO}}$. Based on a reflection of the involved challenges, we postulate eight central *research goals* for relation seeking between data content and metadata.

$\mathbf{RG}_{\text{C+M1}}$ Relating Different Types of Data Seeking relations between data content and metadata requires the ability to relate these different data types with each other. Thus, we are confronted with two data types of different structure, also referred to as multi-modal data $\mathbf{C}_{\text{C+M}}$. Meaningful content-based access strategies are required $\mathbf{C}_{\text{CBA}}$, e.g., with visual-interactive preprocessing techniques (cf. Chapter 4). In this connection, specific challenges associated with the underlying data type can be resolved, e.g., with time-oriented primary data $\mathbf{C}_{\text{CBA}}$. Likewise, metadata has to be accessed and be prepared for relation seeking. Metadata typically consists of various different nominal or numerical (quantitative and qualitative) attributes. Datasets with mixed attribute types will be referred to as *mixed data* throughout this thesis (cf. Section 2.2.1). The analysis of mixed data is considered difficult in general [JJJ08]. Finally, establishing a connection between the two multi-modal data types (data content and metadata) constitutes a specific research goal $\mathbf{C}_{\text{C+M}}$. Unification strategies have to be applied to make these different attribute types comparable and to enable relation seeking. However, the characteristics of different data types need to be reflected by the definition of a relation, and by the visual-interactive interface. In addition, the definition of a relation has to match the information-seeking behavior of users $\mathbf{C}_{\text{UCD}}$. To discover relations between data content and the metadata, appropriate DM, ML, and statistical analysis techniques need to be incorporated.

$\mathbf{RG}_{\text{C+M2}}$ Assessing the Interestingness of Relations Not every relation between different attributes of a data set is interesting for the user per se $\mathbf{C}_{\text{UCD}}$. In fact, the interestingness of a relation depends on the data, the analysis task and the user, in particular. The definition of interestingness must be adequate for the types of underlying data. The time-oriented primary data content and the different types of available metadata attributes have a strong influence on the appropriateness of different interestingness measures $\mathbf{C}_{\text{CBO}}$. As an example many measures for assessing the interestingness of a relation consider either numerical, ordinal, or categorical data. The exploratory nature of relation-seeking tasks contribute to the difficulty of defining the interestingness of relations. An interesting relation might be something entirely new to the expert $\mathbf{C}_{\text{UCD}}$. Instead, it might be a relation which satisfies or disproves the users' expectations. Specific analytical solutions may consider similarity measures or correlation measures to identify interesting relations while more generic approaches may be based on dependency measures. Of course, specific measures, like linear correlations, generate more specific conclusions. However, we need to ensure that more general patterns are not overlooked in the first place [Shn02]. A possible solution is providing a variety of interestingness measures each with a different specialization for the data, user, and task $\mathbf{C}_{\text{UCD}}$ involved.

$\mathbf{RG}_{\text{C+M3}}$ Multiple Granularity Problem An interesting relation can be identified for the entire data set meaning that some attributes of the data set share a relation. We call this granularity the *attribute-level*. For instance, one might detect that the top speed of cars is related to their horse power. A widely used measurement for the linear correlation (dependence) between two numerical attributes is the Pearson correlation coefficient. However, this class of

measurements may mask the specific behavior of different subsets (observations, cohorts) in the data. To continue the example, cars exist which have tremendous horse power, but are specified for cruising (low top speeds) instead of racing. We call this analysis granularity the *bin-level*, where observations are made for subsets of the data, defined by one, two, or multiple attributes. An additional example may be a relation that significantly applies to all female participants of a study, but not to male participants. This example demonstrates how categorical metadata attributes can be included in the exploratory relation-seeking process. However, different data types increase the complexity of defining relations and their respective interestingness $\mathbf{C}_{\text{C+M}}$. To support relation seeking data-driven research, an efficient approach should provide analytical capabilities for the identification of relations at both the attribute-level and the bin-level.

$\mathbf{RG}_{\text{C+M4}}$ Overview of the Relation Space In Chapter 5, we presented guidelines and techniques for content-based overviews, e.g., for time-oriented primary data $\mathbf{C}_{\text{CBA}}$ $\mathbf{C}_{\text{CBO}}$. We now shift the research goal towards providing overviews of the complete body of information including data content and metadata. $\mathbf{C}_{\text{C+M}}$. In addition, the relations between different types of data need to be visually represented. Providing an overview of the relation space is an IV challenge. Hunting for interesting relations in complex data sets may be time-consuming, especially if single relations have to be validated one by one in a batch-process. On the contrary, an overview of the relations will provide a broad range of available relations from start. Showing a variety of relations in an intuitive way facilitates overview-first exploration. This change in the analytical workflow would shift the batch process problem towards a relation-space overview problem. For providing an overview of the relation space, the structure of the relations between different types of data has to be presented visually. A careful choice of data abstractions is needed to provide a) an overview of the focused data attributes, b) an overview of available relations, and c) tight coupling of both. Thus, visual representations must be easy to translate to the domain knowledge of the user and vice versa, while remaining generic enough to cover different types of findings $\mathbf{C}_{\text{UCD}}$. In this connection, a particular challenge is revealing structure in (categorical) metadata attributes. Providing structure in metadata requires carrying out different steps of the reference workflow for ESS (cf. Figure 3.1), i.e., the data aggregation, the visual mapping, and the view transformation. For metadata without any structure it is hardly possible to provide layouts for an overview-first exploration. In current ESS the structure of metadata for the visual representation of, e.g., facets often must to be created by hand [Hea06]. Approaches that create structure for metadata layouts automatically are needed.

$\mathbf{RG}_{\text{C+M5}}$ Guiding Users Towards Interesting Relations Manually inspecting individual relations is tedious. Especially for large and complex data sets a search problem arises. Fully interactive search for potentially useful or interesting relations between data content and metadata may constitute a cumbersome and long process $\mathbf{C}_{\text{C+M}}$. Appropriate guidance concepts would lead the user through the potentially huge search space towards the most interesting relations. Emphasizing most interesting relations describes a depth-first search approach to keep track of the relation space. As such, guidance concepts are an appropriate means for generating new hypotheses and gaining new insight. A prerequisite for being able to guide users towards interesting relations is the assessment of the users' notion of interestingness $\mathbf{C}_{\text{UCD}}$, which in turn can be represented visually. However, visually encoding the interestingness score of a relation (cf. research goal $\mathbf{RG}_{\text{C+M2}}$) is only part of the coin. Guidance concepts should also reveal groups of interesting relations or even multivariate relations. In this connection, it is important that guidance concepts are able to work at multiple levels of granularity (cf. research goal $\mathbf{RG}_{\text{C+M3}}$). Moreover, technical solutions must be provided to rank the interestingness of different relations, and combinations of relations. As an example, a clustering result may have interesting relations with multiple metadata attributes. In that case, users should be guided towards metadata attributes or sets of metadata attributes providing the most interesting relations.

$\mathbf{RG}_{\text{C+M6}}$ Interaction Design According to the Information-Seeking Mantra [Shn96] the relation-seeking process should be supported with information drill-down capability. The user may want to steer the number of shown relations. Furthermore, the user may want to focus on subsets of the provided information to match specific domain knowledge or to prove expectations of the relations. A solution to this requires flexibility in terms of dimensionality and complexity. Interaction designs supporting information drill-down are, e.g., zooming, panning, and filtering operations. To solve this research goal, filters are needed for both to adapt the shown data subset and to drill down the relation-space. These interaction designs are associated with the challenges of exploring time-oriented data $\mathbf{C}_{\text{CBA}}$ $\mathbf{C}_{\text{CBO}}$ and of combining data content with metadata $\mathbf{C}_{\text{C+M}}$. From a VA perspective, especially the experienced user may want to steer relevant parameters of the provided models $\mathbf{C}_{\text{MPC}}$. In this sense, this research goal is associated with the specification of the interestingness of a relation $\mathbf{RG}_{\text{C+M2}}$ based on the users' notion of interestingness and with the goal to guide users towards the most interesting relations $\mathbf{RG}_{\text{C+M5}}$. These types of algorithmic and visual-interactive support may take additional model parameters which may be steerable by the user.

RG_{C+M7} Involving the User in the Design Process In Chapter 4, we showed how the user can be involved in the process of providing content-based access C_{CBA}. Similarly, in Chapter 5, we presented guidelines and techniques facilitating the user-centered design of content-based overviews C_{CBO}. Relating the individual attribute types of complex data elements with each other also forms a large design space where user involvement is desirable C_{UCD}. Important factors are the selection of sets of relevant attributes and coping with the individual attributes. In addition, engaging users is beneficial for the definition of a relation, the assessment of interesting relations, and the visual-interactive design of the relation-seeking application. In this context, we draw a connection to the design of content-based overviews (cf. Chapter 5). Similarly, to reach this research goal, the three steps of the reference workflow (data aggregation, visual mapping, and layout) need to be approached. However, the existence of an additional *relation space* (cf. research goal RG_{C+M4}) increases the design's complexity and, consequently, recommends involving users C_{UCD}. In many cases, users may want to choose between different implementations of the available models, and respective model parameter values C_{MPC}. As a result, some of these parameters may already be fixed in the design phase, reducing the complexity of the interactive solution. For the remaining parameters useful interaction designs enabling users to steer the models in the application phase need to be carried out.

RG_{C+M8} Visual Communication of Interesting Relations We have already described the benefits of content-based overviews for the visual communication of data-centered information and insight. The visual communication of the relation space is yet another promising, but also challenging task C_{C+M}. A particular problem is the visual complexity of data content, metadata, *and* mutual relations in between. In contrast to the visual communication of data, the communication of relations between data attributes is per se a communication of information. However, one challenge regards the communication of information about relations in an intelligible way. Moreover, the question arises how non-experienced audiences can gain insight based on the communicated information. For being able to communicate knowledge gained during the ESS application, intuitive solutions may be beneficial to keep the cognitive load of external audiences as low as possible. In this connection, involving users in the design C_{UCD} is greatly beneficial. Other requirements for the usefulness of such visual-interactive displays may be based on the targeted media of communication. Some devices may only have a limited display or print resolution. Another concern may be the use of misleading colors, such as white color, which is already reserved in print media.

6.1.3. Contribution

In this chapter, we present three different techniques supporting users in seeking interesting relations between data content and metadata. First, we present a technique mapping metadata attributes onto data content. The data content is structured in a content-based overview layout and serves as the targeted variable for relation seeking. The technique visually assesses which metadata attributes share interesting relations with the data content. In addition, users can analyze which clusters of the data content can be described in terms of mapped metadata. Second, we show how data content can be mapped onto the layout of a metadata attribute. In contrast to the first contribution, metadata is defined as the target attribute and is represented in an overview visualization. Data content is mapped onto the layout to support the user in seeking relations between metadata and data content. With the final contribution, we neglect the terms of data content and metadata. Instead, we assign both data content and metadata to a mixed data set. We present an approach where multiple attributes of different types of data (mixed data) can be explored at a glance. The definition of a target variable in advance is not necessary. Based on overview visualizations of both the mixed data set and the most interesting relations, the user is able to formulate new hypotheses about the data. Interaction designs enable users to drill down to interesting relations and the associated data, respectively. With the incorporation of multiple data attributes the approach also supports the identification of interesting multivariate relations. We present usage scenarios for all three contributions in which we validate of the techniques's usefulness. The three technical contributions are based on [BRS*12b, BRS*12a, BSW*14].

All three techniques implement the reference workflow for the design of visual-interactive interfaces for ESS. For the first step of the workflow (preprocessing of time-oriented primary data, we build upon the solutions presented in Chapter 4. The remaining three steps (data aggregation, visual mapping, and layout) are most relevant for the presented techniques. In this connection, we also use the guidelines and techniques presented in Chapter 5.

6.1.4. Relation to the Reference Workflow

In Figure 6.1, we draw the connection between the contributions of this chapter and the reference workflow for the design and the application of ESS for time-oriented primary data. Relation seeking between data content and metadata

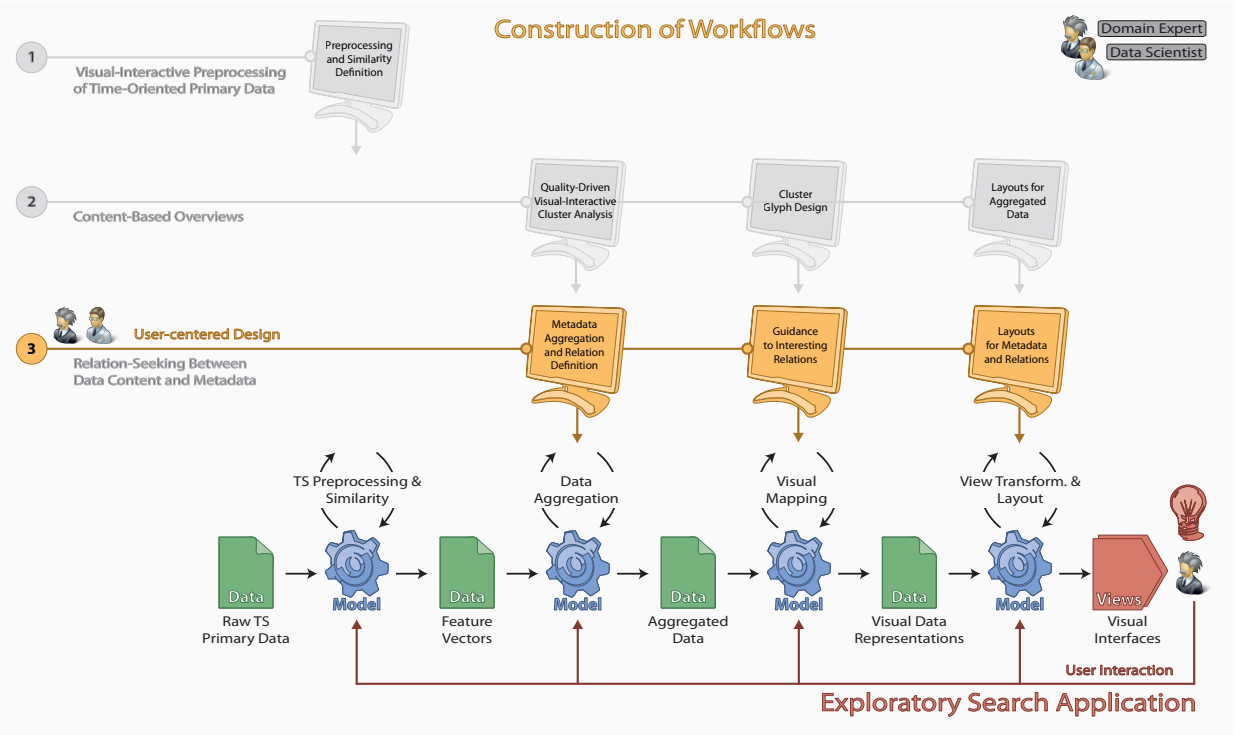


Figure 6.1 *In this chapter, we present techniques for relation seeking between data content and metadata. We use three steps of the reference workflow, i.e., the data aggregation, the visual mapping and the layout step.*

builds on content-based access and visual overviews of data content. This is why this chapter benefits from the upstream Chapters 4 and 5. In addition, we use the reference workflow for the design of visual-interactive interfaces for ESS for the contributions presented in this chapter. We implement the data aggregation step to aggregate metadata and data content. On this basis, we are able to define different types of relations between these two data abstractions. In the visual mapping step, we present different visual encodings of the aggregates and the relations in between. Moreover, we present techniques guiding users towards most interesting relations. Finally, in the view transform and layout step, the body of information is represented in the display space. The structures within the data abstractions of both data content and metadata are used for the layout of the data in 2D. Furthermore, we present interaction designs for browsing, zooming, and filtering of both the data space and the relation space.

6.1.5. Chapter Overview

Section 6.2 reviews specific techniques for seeking relations with an emphasis on the combined analysis of data content and metadata. In Sections 6.3, 6.4, and 6.5, we present three approaches for seeking relations between data content and metadata. We present techniques for mapping metadata onto content-based overviews, mapping data content onto metadata-layouts, and relation seeking in multi-attribute data. Section 6.6 summarizes the chapter with a discussion and a conclusion.

6.2. Baseline Techniques

6.2.1. Mapping Metadata onto the Data Content

In Chapter 5, we presented guidelines and techniques for the design of content-based overviews. We next show existing baseline techniques for approaches mapping metadata attributes onto content-based overviews to support relation seeking. Kehrer and Hauser divide solutions supporting relation seeking and comparison tasks for *multifaceted scientific (primary) data* [KH13] in three visualization techniques. These techniques are *side-by-side* comparison, *overlay* in same coordinate system, and *explicit encoding* of differences and correlations. The techniques match

the concepts of juxtaposition, superposition, and explicit encoding presented in the survey of visual comparison presented by Gleicher et al. [GAW*11]. A variety of cluster visualizations map additional (metadata) attributes onto the display. If a cluster layout in 2D is provided, the same coordinate system can also be used in combination with additional visual variables, such as (background) colors to support relation seeking. As an example, Vesanto uses heatmap metaphors to represent the density/proximity of individual clusters in a cluster visualization (here: the Self-organizing Maps algorithm [KSH01]) [Ves99]. For the visualization of high-dimensional data elements (and additional attributes) glyph designs are among the most widespread visualization techniques [BKC*13]. Glyph designs represent high-dimensional data attributes with various visual variables. We presented guidelines and techniques for glyph designs for content-based overviews in Section 5.4.1. A classical overlay-based approach for mapping additional metadata onto the data content is (cluster) labeling. The SOM-based music retrieval approach by Rauber et al. may serve as an example [MPR02]. Clusters of music tracks are labeled with title and artist attributes. In the approach of Nocaj and Brandes hierarchically clustered document collections are visually represented in a so-called reference-map [NB12]. The layout includes labels for the textual description of the hierarchical document clusters, as well as a link-based overlay technique for the visual representation of query hits. The FacetAtlas tool layouts document clusters in 2D and additionally shows multifaceted relations of documents within or cross the document clusters [CSL*10]. Finally, widely applied relation-seeking techniques in exploratory data analysis are (a) linking views and (b) showing data from different perspectives. An overview of linked views is presented, e.g., by Kehrer and Hauser [KH13], example techniques are demonstrated by Stasko et al. [SGL08], Kehrer et al. [KLM*08], or by Dork et al. [DCCW08]. Visual encodings enable the localization of data elements in different views, whereas each view shows a different perspective of the data. An effective visual variable for linking views is color. Relation-seekers are able to link objects of identical colors in different views, and correlate the information provided by different perspectives. Lacks of techniques in the related work remain in the condensed/aggregated visualization of both the data content and the associated metadata information. In addition, most techniques neglect to support users in assessing the interestingness of a given relation between data content and metadata.

6.2.2. Mapping Data Content onto Metadata

Another class of existing baseline techniques maps the data content onto metadata which is distributed in a spatial arrangement on the display. We first present an overview of metadata-based layouts, before we show approaches mapping data content onto metadata.

Metadata Layouts In many search systems individual documents are visually represented in a list-based view. We echo Smith et al. in that the list-based paradigm has the "opportunity for improvement" for heterogeneous data and rich metadata [SCM*06]. A number of layouts for metadata exists which go beyond a list-based spatial allocation of the display space (see again Smith et al. for an overview [SCM*06]). An important concept for the layout and utilization of metadata in search systems is providing facets (also called category systems). Facets consist of concepts (terms, entities, items) of a particular category (property, metadata attribute) typically represented in list-based or hierarchical structures. Marti A. Hearst describes metadata facets as a "set of meaningful labels organized in such a way as to reflect the concepts relevant to a domain" [Hea06]. Facets usually need to be created manually to preserve the concepts of a respective domain. However, this process can be automated to a certain extent of accuracy. The strength of metadata facets is their coherency and their relative completeness. It is still an unanswered question what kind of structure (layout) is most effective for exploration and browsing of information collections [Hea06]. The FacetMap approach [SCM*06] (not to be confused with the FacetAtlas tool [CSL*10]) provides a layout of nested metadata facets to support both searching and browsing. The approach supports hierarchical facets and metadata attributes with domain-specific concepts. The approach, however, does not support relation-seeking tasks in combination with the data content. Other types of metadata layouts are based on the scatterplot metaphor. In the Envision DL [FHN*93] retrieved documents are arranged by two axes spanned by two metadata attributes. The approach uses the intrinsic order of the two metadata attributes, in the given example an alphabetical order and a chronological order. In summary, it is still a subject to research of how many metadata attributes can be arranged in meaningful layouts in 2D. Our techniques overcome the challenge of providing meaningful layouts for metadata in different ways.

Mapping Approaches For search systems it is frequent practice to enrich the search support with quantitative information about documents, such as the number of documents referring to a facet, or the number of pages of a document. However, in most cases, these types of additional mappings more often comply with additional metadata than with the data content of the document collection. From an exploratory analysis perspective, mapping data content onto metadata can be seen as a process to support the formulation of new hypotheses. A dependent attribute (here: a metadata attribute) is selected and the analyst is interested in relations between the dependent metadata attribute

and the data content to be mapped. However, in time series data analysis many approaches define the temporal domain of the time series data content as the dependent variable (cf. Section 2.3). The number of approaches in which the primary time axis is decoupled and other dependent variables can be defined is scarce. This shortcoming motivates our techniques enabling relation-seekers to shift the dependent variable between the time-oriented data content and different attributes in the metadata. A best-practice example from Earth observation and climate research is presented by Kehrer et al. [KLM*08]. The user is able to select a subset of measurement data based on some non-temporal attribute. In different linked views the subset can be visualized with respect to the primary time axis of the measurement data.

6.2.3. Relation Seeking in Multiple Heterogeneous Attributes

In many exploratory data analysis applications the classification of data attributes into data content and metadata is not entirely clear or even undesired. Instead, many approaches rather distinguish between numerical (quantitative) and nominal (qualitative, categorical) attributes (cf. Section 2.2). These heterogeneous data types are referred to as mixed data. Similarly, many exploratory data analysis approaches loose themselves from a fixed determination of the dependent (targeted) attribute and the independent attributes. Instead, exploration systems aim at supporting the user in generating new hypotheses on interesting and possibly unknown relations between various attributes. We present an in-depth overview of relevant concepts and techniques in [BSW*14]. The work of Kehrer et al. [KLM*08] again serves as an example. A variety of multivariate data attributes (primary and secondary data) can be explored in combination with time series data. The multiple linked views technique is applied to support relation seeking. Different target variables can be defined by an interactive selection of data subsets based on individual attribute visualizations. The Jingsaw system supports the user in the identification of relations (connections) between entities of different documents [SGL08]. The multiple linked views technique is applied to illustrate possibly complex attribute spaces. Other approaches for relation seeking in mixed data sets use the parallel coordinates metaphor, such as Parallel Sets [KBH06] and VisBricks [LSS*11]. These approaches enable gaining an overview of relations in multivariate data sets including the output of clustering or binning techniques. A VA system for large-scale categorical data is presented by Alsallakh et al. [AAMG12]. Based on an aggregation concept, their Contingency Wheel++ is adequate for millions of data records as shown for movie ratings. The user is able to define a single categorical attribute as the target variable. A wheel metaphor aligns the data set on the display. Additional independent variables can be mapped onto the wheel for the identification of interesting relations. From an exploratory data analysis perspective, one drawback of the approach is the condition that the user has to define a fixed targeted variable. A variety of approaches exist for the analysis of categorical data sets. An overview is provided in the state-of-the-art report on the visualization and analysis of sets, presented by Alsallakh et al. [AMA*14]. The authors outline the analysis of relations between different sets over a collection of data elements as one of the future challenges. Johannson et al. present a technique for the quantification of categorical attributes [JJJ08]. Multiple Correspondence Analysis (MCA) is applied to analyze a set of numerical variables and to suggest a quantification of a categorical attribute.

6.2.4. Assessing the Interestingness of Relations

We present an overview of baseline techniques assessing the interestingness of relations, e.g., between data content and metadata. Assessing the interestingness of a given relation is most relevant for guiding users towards the most interesting relations, based on the functional definition of interestingness. Relations exist at different levels of granularity, important granularities are the object-level (very fine-gained), the bin-level (for categories of an attribute), the attribute-level (attributes at a glance), or the cluster-level (clusters). In addition, relations may exist between heterogeneous levels of granularity. An example of the latter is the class label of data elements grouped in a data cluster. A great variety of interestingness measures for relations exist. The data types covered by a relation have a great influence on the reliability and the meaningfulness of interestingness measures. In the following, we review interestingness measures for metadata attributes with an emphasis on nominal and categorical data types, followed by content-based interestingness measures focusing on numerical attributes. Finally, we emphasize interestingness measures for multivariate data content.

Interestingness Measures for Relations in Metadata One type of metadata-based interestingness measures is based on annotations, e.g., provided with tagging-based approaches or metadata attachments performed in the data processing step of the data life-cycle. As a rule, two documents are considered interesting (similar) if their annotations are similar. Thus, in these types of approaches the interestingness between entities can be defined in terms of similarity, see [BRS*12a] for an overview. Another means of assessing the interestingness of metadata-based relations can be borrowed from recommender systems. In these statistical approaches the interestingness of a relation between two entities is increased, whenever they are being bought together, looked at the same time, have the same

co-citation patterns, etc. Interestingness can also be deduced from ontologies and other semantic data structures. Assumed an ontology is provided for a given domain, the incorporation of semantic relations for the definition of interestingness is possible. A wide span of areas, such as DLs, bioinformatics, financial services, web services, business intelligence, and national security exist, yielding new types of applications to extract (or analyze) semantic relations [KMS*08, BRS*12a, NRB*13]. Two prominent semantic applications DBPedia [1] and Wolfram Alpha [2] may serve as examples. An example in the ES context is presented by Hecht et al. [HCQ*12]. The authors introduce explanatory semantic relatedness measures, i.e., estimations of the degree of 'relatedness' between two concepts, assessed by mining additional Wikipedia [3] resources. The Contingency Wheel++ approach presented by Alsallakh et al. reveals relations in multiple categorical attributes [AAMG12]. A circular layout aggregates documents of large collections with respect to a categorical target attribute as a basis for relation seeking between categorical attributes. For the assessment of interestingness a measure of association based on Pearson's residuals is applied. In other words, the significance of associations between the categories of two categorical attributes can be assessed. The rationale of such statistical tests is the comparison of something assumed about the data with something actually measured in the data.

Interestingness Measures for Relations between Metadata and Data Content Approaches exist that define interestingness measures for relations between metadata and the non-textual data content, or between mixed data attributes. For instance, the distribution of labeled data elements within a cluster may be a subject of diversity measures for the assessment of interestingness. For cluster analysis approaches the degree of cluster separation can be assessed with cluster quality measures revealing interesting relations between clusters. We refer to Section 5.3 in which we present techniques for quality-driven visual-interactive cluster analysis including an overview of baseline quality measures. The approach by Johansson et al. combines the relations of numerical attributes with domain knowledge to quantify categorical attributes supporting the assessment of interesting relations [JJJ08]. The counterpart of quantification concepts is the categorization of numerical data attributes. Several binning techniques exist assigning every data element to a bin based on the numeric value domain. Salient concepts are the preservation of the value domain and the frequency domain. The result of binning approaches is a unification of mixed data attributes. In this connection, statistical dependency testing can be applied to identify interesting relations between groups of data elements (bin to bin comparison) as outlined for the Contingency Wheel++ approach. An overview of statistical dependency testing including different implementation and application examples is presented in [BSW*14]. Contingency tables visually represented by, e.g., mosaic plots, are a prominent example of the assessment of the interestingness by the absolute count of observations. Finally, similarity measures may be defined for mixed data sets to assess the interestingness between mixed data elements. Most recently, we presented a visual-interactive approach for the definition of similarity for mixed data sets [BSR*14]. Users are able to give similarity-preserving feedback for individual objects based on their domain knowledge (we call it: mental similarity notion).

Interestingness Measures for Relations in the Data Content For the interestingness measures for relations regarding the data content, we primarily focus on numerical attributes as typically provided in feature-based approaches followed in this thesis. We start with examples of bivariate relations. Many approaches for the visual assessment of interestingness between sets of bivariate attributes are based on scatter plot visualizations. A scatter plot is an intuitive visualization for humans to judge if the data relations between two numerical attributes are interesting. A variety of computational methods have been postulated to automatically analyze the interestingness of sets of bivariate data. The rationale is to analyze the distribution of points in 2D, where usually a narrow distribution is considered interesting (cf. [Shn02]). The Pearson correlation coefficient may serve as a prominent example. An alternative class of techniques defines the interestingness of bivariate sets of data elements by assessing goodness-of-fit parameters of different regression models [SBS11]. The approach includes seven different regression models yielding a seven-dimensional FV for scatted data distributions. We continue the overview of interestingness measures for relations with representatives used for multivariate data content. From a DM perspective, several measures support interestingness-based VA, see [BRS*12b] for a brief overview. Complementary to the latter, projection techniques can be applied to reduce the dimensionality of the multivariate data set. As a result, techniques such as presented for bivariate data can be applied to the projection output. In their quantification approach for categorical data, Johansson et al. [JJJ08] apply Multiple Correspondence Analysis (MCA) to analyze a set of numerical variables. In this way, interesting relations in the data can be identified supporting the quantification process of categorical attributes. Due to its particular importance for this thesis, we emphasize the role of similarity measures for the assessment of interestingness of time series data. This process usually involves a time series descriptor and a distance measure to assess the dissimilarity of any two data objects. For in-depth details of various definitions of time series similarity (and distances), we refer to Chapter

[1] DBPedia http://de.dbpedia.org/, last accessed on September 24th, 2015
[2] Wolfram Alpha: Computational Knowledge Engine http://www.wolframalpha.com, last accessed on September 24th, 2015
[3] Wikipedia - http://en.wikipedia.org, last accessed on September 24th, 2015

4. For groups of unlabeled data elements cluster quality measures can be used, such as the cluster compactness or the separation of the cluster within a clustering result (cf. Section 5.2.1). As described earlier, for labeled data the result of such quality measures is informative for the interestingness of single clusters or clustering results. The DL approach of Rauber et al. may serve as an example [MPR02]. A hierarchical SOM clustering of a music collection is additionally labeled with the QE values for each cell of the SOM. In consequence, the user can identify clusters of particularly similar music. Rauber et al. also use the quantification error as a split criterion for the SOM hierarchy.

6.3. Mapping Metadata onto Content-Based Overviews

Visual-interactive cluster analysis techniques enable users to understand the content of large data sets in terms of representative clusters and cluster relations. In this way, the output of visual cluster analysis tools can build the basis for content-based layouts and overviews (cf. Chapter 5). To understand the complete body of information, the clustering results are to be understood in context of the attached metadata given for the data collection. Especially for data-driven research with primary data, it may be of particular relevance to consider the conditions under which the data was measured. Examples may be the age of patients in electrocardiogram data, or the location on Earth in the context of weather measurements. As a result, ESS incorporating the complete body of information would allow users to (1) understand the main characteristics of the data content itself and (2) the relations between data content and attached metadata. The latter can be considered a correlation problem, where the degree of variation of one variable (data content) with others (metadata attributes) is assessed.

We recall the central challenges of existing techniques aiming at solving both (1) and (2). In Section 6.2, we outlined linked views as a means of relation seeking. However, linked views have the drawback that relation-seekers need to identify related objects across different views, e.g., a cluster in a content-based overview and a metadata attribute in a metadata visualization. For this purpose, highlighting techniques can be applied to facilitate the localization of single relations in different views. However, intensive user interaction is required until the overall structural information of multiple views can be comprehended. These approaches hamper the comparison of multiple relations at a glance, and thus the identification of the most interesting relations. Another common technique is the visualization of metadata on top of content-based overviews, e.g., by means of labeling techniques. However, the pure visualization of metadata labels on top of content-based overviews neither reveals the interestingness of the visualized relations, nor guides users towards the most interesting relations. For both types of existing approaches manually inspecting the distribution of metadata for each cluster is tedious, especially for large and complex data sets. Fully interactive search for potentially useful or interesting relations may constitute a cumbersome and long information-seeking process. In addition, the problem worsens if the number of available metadata attributes is large. Nonetheless, for primary data a multitude of metadata attributes can be expected, all being potentially relevant for deriving hypotheses.

We contribute a technique allowing users to seek interesting relations between content-based overview and additional metadata attributes. The technique supports guiding users through potentially huge search spaces. To this end, the approach automatically identifies potentially interesting metadata attributes and highlights interesting relations to foster an in-depth exploration. For each cluster, the *distribution of the metadata* is summarized. We guide users to interesting relations between data content and metadata by visualizing a histogram of the metadata distribution for every cluster. Moreover, for the histogram of every cluster, we compute measures of interestingness defined on the metadata distribution. More precisely, the measures of interestingness are based on the homogeneity of the metadata distribution for a given cluster, as well as the dispersion of the metadata distribution among the neighborhood of similar clusters. These measures are used to automatically score and rank relations between clusters and associated metadata based on their relative interestingness. The interestingness of every relation is displayed in the content-based layout to facilitate both the easy identification of individuals and the comparison of many relations. Appropriate visual representations for interesting relations, as well as a ranking of interestingness scores, guide users towards the most interesting relations in the possibly large and complex relation search space. We test our approach on a large real-world data set with time-oriented primary data. The usage scenario demonstrates how our approach automatically identifies most interesting metadata attributes and how the user is supported in discovering interesting and visually understandable relations.

6.3.1. Relation Definition

At first, we describe how a relation between data content and metadata is defined. We denote time-oriented data as data content which is to be clustered, e.g., in an instantiation of the reference workflow for the design of ESS (see

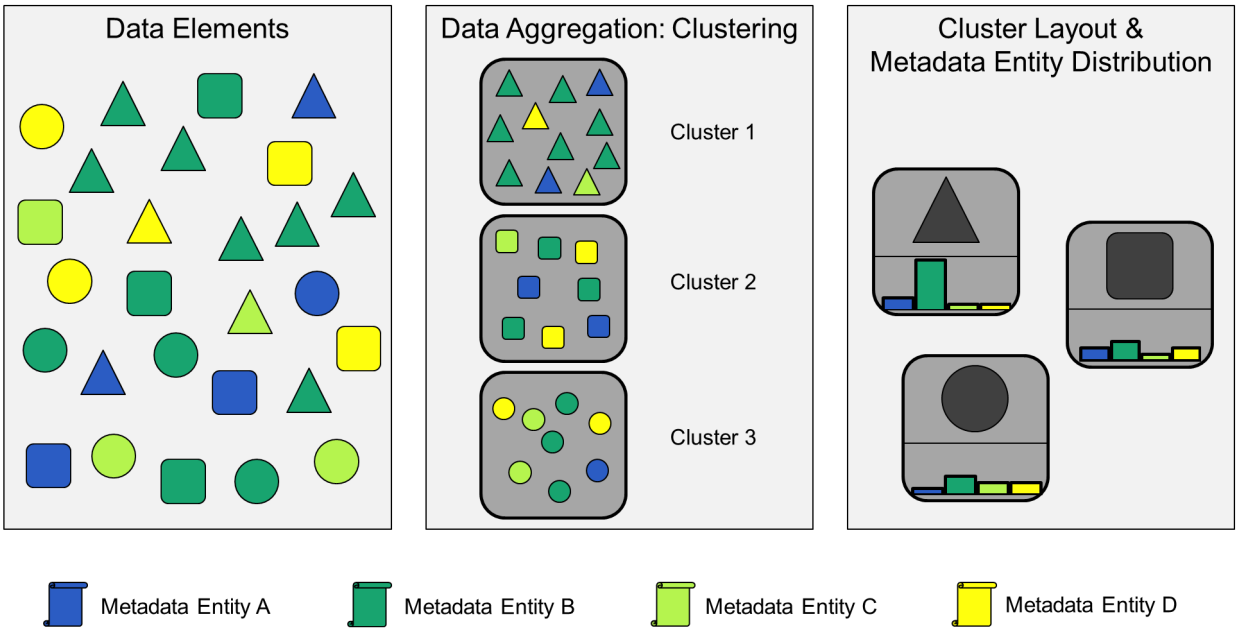


Figure 6.2 *Definition of a relation between data content and a metadata attribute. The schematic illustration shows a set of data elements on the left where the shape of a data element shall depict the data content and the color of data elements indicates a metadata attribute. A clustering routine at the center reveals three shape-based clusters. On the right, a cluster layout in 2D is shown. The distribution of the metadata attribute color is shown as a histogram metaphor for every cluster. A cluster and its associated metadata defines a relation.*

Figure 6.1). In Chapter 4, we presented techniques for user-centered, visual-interactive time series preprocessing and for the definition of appropriate similarity measures. In Chapter 5, we presented guidelines and techniques for the user-centered design of content-based overviews. We recall these contributions for content-based access and content-based overview strategies.

The available metadata information may consist of a rich set of attributes, such as the creator or the publisher of a scientific document, among many other attributes. Our approach accepts metadata attributes of different data type, i.e. binary, categorical, ordinal, and numerical attributes. We suggest the involvement of the user for the characterization of the potentially large set of metadata attributes. User feedback may be incorporated to ensure a certain extent of data quality in the metadata attributes. As an example, quality leaks may be caused by categorical metadata attributes consisting of hundreds of different but rarely populated observations which hamper the calculation of significant interestingness scores. Metadata attributes providing weaknesses in the data quality may be subject to additional preprocessing routines. For every metadata attribute, we are interested in the distribution of available entities (bins, observations). Hence, we aggregate numerical attributes and generate a set of bins reflecting the distribution within an attribute. As a positive side effect, we achieve a unification of binary, categorical, ordinal and numerical data types. The relation of a cluster in the data set and its metadata distribution defines the relation of this approach.

In Figure 6.2, we show a conceptual illustration of the relation definition. The data content is illustrated with the shape of the shown data elements, a metadata attribute is depicted with the color of the shown data elements. On the basis of the shapes of the data content, we design a content-based layout as described in Chapter 5. As the next step, we compute a histogram of the metadata distribution for every cluster of the content-based layout. We store the relations between clusters and histograms as a basis for downstream interestingness measures. The image on the right shows a content-based overview of the three shape-based clusters and the associated histograms. The distribution of metadata entities is depicted with colors. As a result of the cluster layout and the cluster glyph, all relations between clusters and metadata histograms can be represented visually. With the definition of a relation between metadata and data content, we provide a solution for research goal $\mathbf{RG}_{\text{C+M1}}$. A relation is defined on the granularity of single bins, i.e., for single clusters. To accomplish the research goal of multiple granularities $\mathbf{RG}_{\text{C+M3}}$, we will show how the provided content-based overview and the given relations can also foster the exploration at the attribute-level.

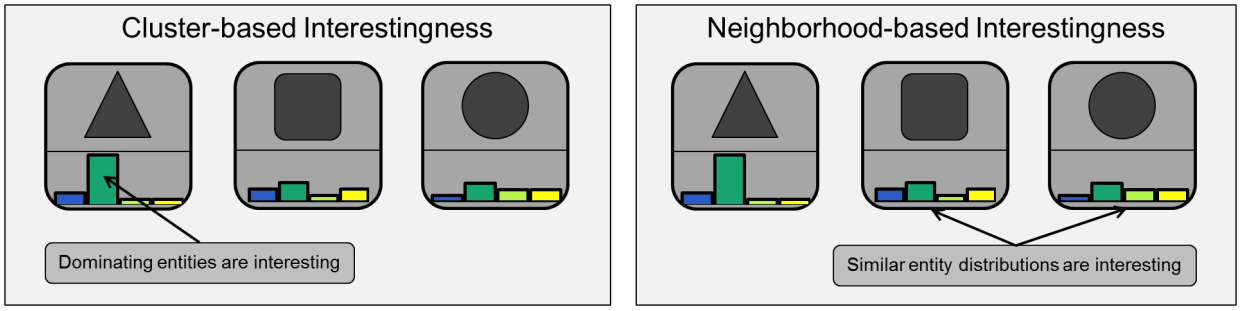


Figure 6.3 *Two complementary concepts for the definition of interesting relations. Left: the cluster-based measure marks relations as interesting if the diversity of the metadata distribution is low. Right: the neighborhood-based measure assesses a relation as interesting if the metadata distributions of similar clusters are also similar.*

6.3.2. Assessing the Interestingness of a Relation

Two of the metadata histograms on the right of Figure 6.2 are very similar. However, they are associated with two different clusters. Obviously, the distributions of metadata histograms do not discriminate these two clusters. On the contrary, a third metadata histogram with a peak of green colored data elements is quite unique. It occurs in one of the clusters exclusively. In the following, we describe how the interestingness of these types of findings can be assessed algorithmically $\mathbf{RG}_{\text{C+M2}}$. We present two complementary concepts for the characterization of interesting relations. For both concepts, we present implementations of measures and outline alternative solutions. The chosen measures are proposed heuristically and expected to be a good starting point. They can easily be extended by more specific measures. In particular, we advocate the involvement of the targeted user group for the design of meaningful interestingness measures for a given real-world example $\mathbf{RG}_{\text{C+M7}}$. All interestingness scores $i_M(C)$ considered here are functions of a cluster C and a metadata attribute M.

$$i_M : C \rightarrow \mathbb{R}^+, C \subseteq E, |C| \neq \emptyset, M \in \mathcal{M} \tag{6.1}$$

E is the number of data elements and $\mathcal{M}$ is the set of metadata attributes. One interestingness concept is based on the granularity of single clusters and the other measure is based on the neighborhood of sets of similar clusters. With the distinction between a cluster-based and neighborhood-based measures, we also confront the challenge of supporting the exploration of relations on multiple granularities $\mathbf{RG}_{\text{C+M3}}$. A schematic illustration of the two interestingness measures is presented in Figure 6.3.

Cluster-Based Interestingness Definition The objective of cluster-based interestingness measures is to describe the *diversity of the metadata distribution* for each individual cluster. This enables users to assess the relation between single clusters and their metadata distribution more precisely. A high interestingness score between a cluster and a metadata attribute is obtained if small subsets of metadata entities (observations) are highly prevalent. Especially dominating metadata entities in the metadata histogram are considered interesting since they directly support sense-making. The dominating green bin in the histogram of colors on the left of Figure 6.3 serves as an illustrative example. Hence, a measure is needed assessing the diversity of a population. Reflecting this characteristic for categorical metadata attributes, we choose Simpson's diversity index D (see Equation 6.2) as a relevant measure of diversity, and thus for the interestingness function i_M (see [BRS*12b] for further references). Let $P_1, P_2, \ldots P_n$ be a partition of the metadata attribute M (observations) into n value sets. The corresponding element set can be defined as follows. $B_k = \{b \in E | M(b) \in P_k\}$. Here, B is referred to as a bin.

$$D_M(C) = 1 - \sum_{k=1}^{n} \frac{|B_k \bigcap C|(|B_k \bigcap C| - 1)}{|C|(|C| - 1)} \tag{6.2}$$

The output of the Simpson's index takes values between 0 and 1. Low index values reflect a high diversity of the bin distribution, which is considered less interesting from a cluster-based perspective. The Simpson's index is computed for every cluster leading to a set of independent interestingness measures for a clustering result. This procedure is repeated for every metadata attribute available in the body of information. The Simpson's index is an appropriate measure for discrete distributions. This is why we aggregate numerical metadata attributes with binning

techniques. Different classes of binning algorithms exist, prominent representatives are frequency-preserving and domain-preserving techniques. The choice of a binning technique influences the histogram creation. We suggest to involve the targeted user group for the choice of the binning technique $\mathbf{RG}_{C+M7}$. For alternative diversity measures, we propose entropy-based measures, such as introduced by Claude Shannon, who invented the information theory.

Neighborhood-Based Interestingness Definition The objective of neighborhood-based interestingness measures is assessing the *homogeneity of metadata distributions among similar clusters*. On the one hand, these measures regard the content-based similarity of clusters. On the other hand, the homogeneity of metadata distributions is taken into account. As an assumption, the relations between clusters and a given metadata attribute are interesting whenever similar metadata histograms are also assigned to clusters of similar data content. An illustrative example is shown in Figure 6.3 (right). The metadata distributions of two clusters are very similar. Assumed that the data content of such clusters is also similar, the neighborhood-based interestingness measure will calculate interestingness scores above average. One of the analysis tasks facilitated with this type of interestingness measures is the aggregation of clusters in additional postprocessing steps of the workflow. In combination with the cluster-based interestingness measure, the assessment of interestingness on the granularity of multiple clusters achieves the multi granularity goal $\mathbf{RG}_{C+M3}$.

We again apply binning techniques to discretize numerical metadata attributes. The discretized histogram information is used to compute the similarity between two distributions via similarity measures for 'cross-bin' comparisons. Measures, such as dynamic time warping and the earth mover distance, may serve as examples (see [BRS*12b] for further references.) We prefer 'cross-bin' comparison since the discretization of numerical values might result in inaccurate bin distributions of similar values. For categorical metadata attributes, we use the Euclidean distance as a representative for 'bin-to-bin' comparisons. With the 'bin-to-bin' comparison the distance of the relative frequencies of identical categories can be assessed and compared easily. Based on the relative frequencies both the measures for numerical and for categorical attributes are in the interval [0;1]. Values around 0 result from almost identical distributions, while higher values indicate decreasing similarity. With the definition of similarity measures for both the cluster content and the associated metadata distributions, we are able to define the interestingness of the neighborhood of every cluster. We compute the homogeneity of the metadata distributions of clusters providing similar data content. Then, for each individual cluster, an average is calculated by weighting the neighbors' content-based distances. As a result, clusters with metadata distributions very similar to the metadata distributions of similar clusters produce the highest homogeneity scores, and thus depict most interesting relations.

6.3.3. Overview of the Relation Space

Our approach relies on the Self-organizing Maps algorithm as a visual clustering and cluster layout solution for the data content. We made the design decision of choosing the SOM approach for its convenience for visual representation. The SOM-based layout of the data content maximizes the exploitation of the display space without overplotting. Furthermore, the preservation of topology aligns clusters similar in the input space next to each other in the output space; a highly relevant property for mapping additional metadata information about top of the content-based overview. However, our technique is not limited to SOM, see Section 5.5 for alternative layout solutions preserving the structure of the high-dimensional input data.

We choose a superposition strategy for the visual representation of relations, i.e., the visualization of the metadata distribution for every cluster on top of the content-based layout. The output of the SOM algorithm provides a grid-based cluster layout. For every SOM cell a glyph design combines the information of the data content and the metadata distribution. At the top of every cluster the relevant information of the data content is illustrated, as described in the guideline on user-centered visual mapping of high-dimensional objects in Section 5.4. For the visual representation of the metadata distribution, we use a histogram metaphor at the bottom of every cluster glyph. For numerical metadata attributes, we apply an additional binning step for the purpose of discretization.

Figure 6.4 shows the glyph design for case data set applied in this approach. We divide the display space in two parts where the upper part shows the data content and the lower part shows the metadata distribution. Similar to the VisInfo case study presented in Chapter 7, we carry out a glyph design for daily measurements of Earth observation measurement data. The cluster centroid is depicted with a black line, the opacity-bands technique reveals the variation of the possibly large number of data elements. A label at the upper right shows the size of the cluster. The example in Figure 6.4 temperature measurements are clustered showing a curve progression between 0°C and 17°C. The temporal domain represents a full day starting and ending at midnight. The progression of the daily temperature curves show an increase of the temperature in the mid of the day. A labeled histogram shows the distribution of metadata. In the example figure, the metadata attribute *Location* of the measurement is chosen. It can be seen that the temperature

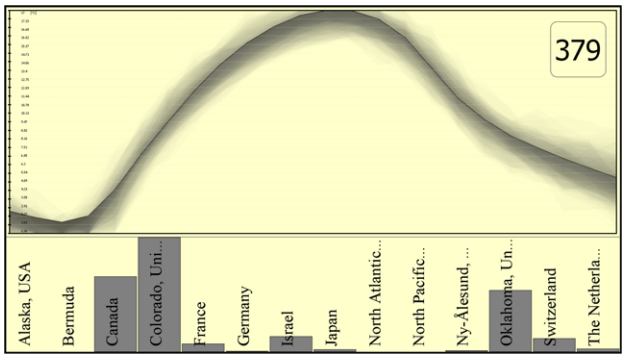


Figure 6.4 *Visual representation of a relation. A glyph design combines the most relevant information of the high-dimensional data content at the top and the metadata distribution at the bottom. In this example, a cluster of daily temperature curves is shown. Most of the measurements were observed at only three locations on Earth.*

curve of the cluster only occurs at some of the locations. In particular, the measurement stations in Canada, Colorado, and Oklahoma show most of the temperature patterns of the cluster.

With the glyph design combining a visual representation of the cluster *and* the metadata distribution, we reveal detailed information of every single relation. Together with distribution of cluster information in the layout, the user can identify variances in the distribution of the metadata attribute. Hence, we establish a solution for the relation space overview problem $\mathbf{RG}_{\text{C+M4}}$. Examples for overview visualizations are presented in Figures 6.5 and 6.6. With the overview of the data content, the additional attribute and the relations between the data content and the attribute, we also provide a means for the communication of results and insights $\mathbf{RG}_{\text{C+M8}}$.

6.3.4. Guiding Users Towards Interesting Relations

Our approach provides an overview of the data content and the associated relations with a targeted metadata attribute. With the SOM-based layout a variety of clusters and superimposed relations are presented to the user at a glance. With the two complementary concepts of interestingness measures, we have a means for assessing of the most interesting relations in the body of information (cf. $\mathbf{RG}_{\text{C+M2}}$). In the following, we present two visual-interactive techniques that directly exploit the results of these interestingness calculations. First, we show how the user is guided towards the most interesting relations between the data content and a *single metadata attribute*. Second, we present a technique enabling users to rank and compare the interestingness values of *multiple metadata attributes*. In combination, these techniques achieve the goal of guiding users towards most interesting relations $\mathbf{RG}_{\text{C+M5}}$. In addition, with these techniques, we highlight interesting relations at the bin-level and the attribute-level. That is why the guidance concepts also contribute a solution for the multiple granularity challenge $\mathbf{RG}_{\text{C+M3}}$.

Guiding Users towards Interesting Relations in a Single Metadata Attribute Cluster-based and the neighborhood-based interestingness measures generate valuable information about the multitude of existing relations. Visual mappings help to represent these interestingness scores visually, and thus to guide users towards interesting relations. For a single metadata attribute, we achieve this by coloring the background of cluster cells. The color scale ranges from dark green corresponding to low interestingness, to bright yellow depicting high interestingness scores [BRS*12b]. The colormap is shown in Figure 6.7. However, other colormap solutions are conceivable, especially if the user is involved in the design process (cf. $\mathbf{RG}_{\text{C+M7}}$). Basically, we recommend utilizing a quantitative, unipolar 1D colormap. For mapping the interestingness values of a single metadata attribute M to color value, we apply a *local min-max normalization* of the interestingness values. The local normalization scales the interestingness scores $i_M(C)$ according to their respective maximum and minimum and maps them to a color range. The interestingness measures for cluster-based and the neighborhood-based relations constitute a user parameter provided with a user control. Extensions of the set of provided interestingness measures, e.g., to cope with individual user needs, can easily be included in the user control.

The relation shown in Figure 6.4 has a high interestingness score represented by bright yellow color. The cluster-based interestingness measure is demonstrated, revealing that the cluster has a low diversity in the histogram (high interestingness) compared with other clusters of the given metadata attribute. The temperature pattern of the cluster only occurs at few locations on Earth. In Figure 6.5, we present the guidance technique for an entire content-based

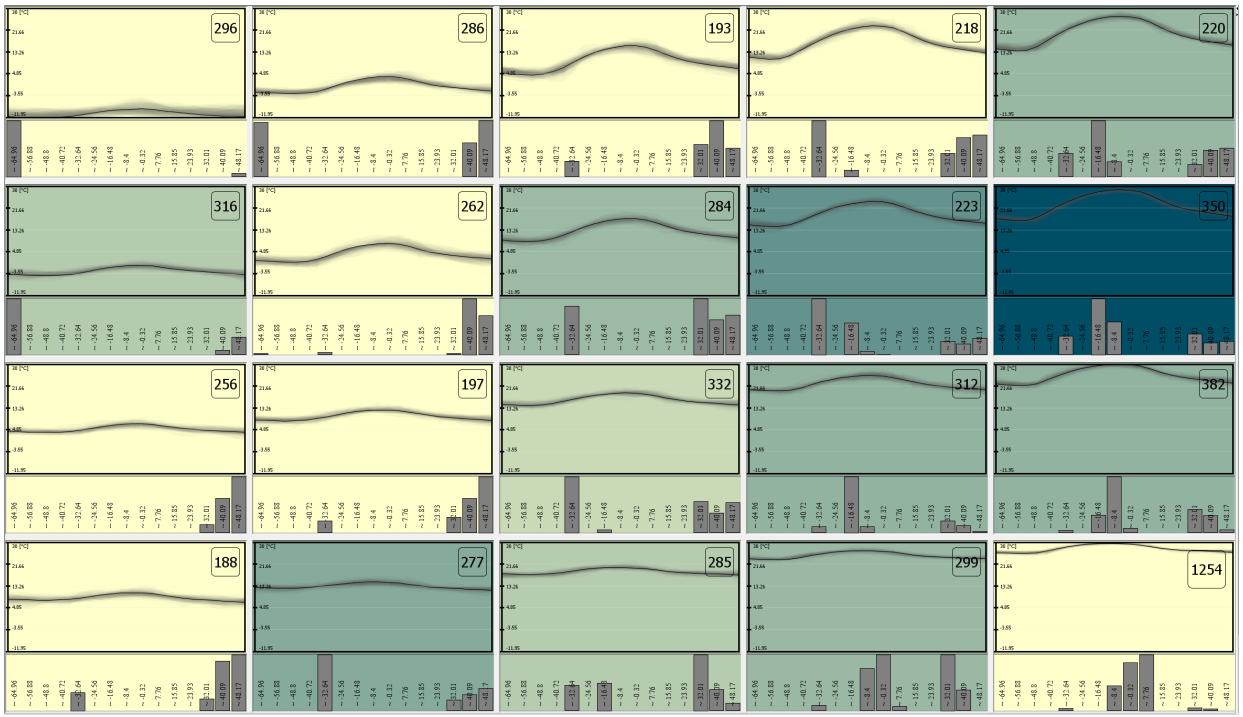


Figure 6.5 *Guided discovery of interesting relations for a single metadata attribute (local normalization). The clustering result of daily temperature patterns are related to the attribute Latitude of the measurement station. The cell-based interestingness measure is shown assigning high interestingness scores to histogram distributions of low diversity depicted with bright yellow colors. Highest interestingness scores occur for very low temperature curves which are typically measured at latitudes of Arctic and Antarctic measurement stations (latitudes* $< -50°$ *and* $> 50°$*).*

layout with the *local min-max normalization* for a single metadata attribute. The cluster-based interestingness measure is shown for daily temperature curves related to the metadata attribute *Latitude*. At the upper left several clusters with light yellow background colors reveal high interestingness scores attracting the attention of the user. Another cluster with a high interestingness score is located at the lower right. In this case, the cluster with the highest temperature patterns of the data set is shown. These temperature measurements predominantly occur near the equator (see the metadata histogram). The remaining clusters are considered less interesting since the diversity of the associated metadata histograms is higher. Thus, the backgrounds of these clusters are colored with greenish colors. The most uninteresting cluster colored dark green is associated with a histogram including six bins all containing a considerable percentage of the cluster pattern. This constitutes the highest diversity of all interestingness scores of the particular metadata attribute. An investigation of this cluster reveals that the temperature pattern occurs at both the northern and the southern hemisphere, in latitudes between the equator and 50° latitude.

Guiding Users towards Interesting Relations in Multiple Metadata Attributes With the local normalization and the visual encoding of interestingness in the cluster glyph, we reach the goal of guiding users towards interesting relations for single metadata attributes. However, due to high quantities of metadata attributes, it might be infeasible to visually explore every possible metadata view at large. Instead, an overview of all available metadata attributes and their interestingness would be beneficial. For this purpose, we present a technique for guiding users towards the most interesting metadata attributes and to the responsible relations. We extend the local normalization function from a single metadata attribute M to a global normalization for the set of all metadata attributes $\mathcal{M}$. The global normalization scales the interestingness scores $i_M(C)$ according to their respective maximum and minimum in the entire set of $\mathcal{M}$ and maps them to a color range. With the global normalization, the interestingness scores for every cluster and every metadata attribute can be compared algorithmically. This has an implication for the visual encoding of the cluster glyphs. For a visualized metadata attribute, the encoding can be used to guide users with both local and global interestingness scores. Consequently, the user can directly identify whether a metadata attribute either contains an above-average number of interesting relations, or may indeed be excluded from the exploration process due to low

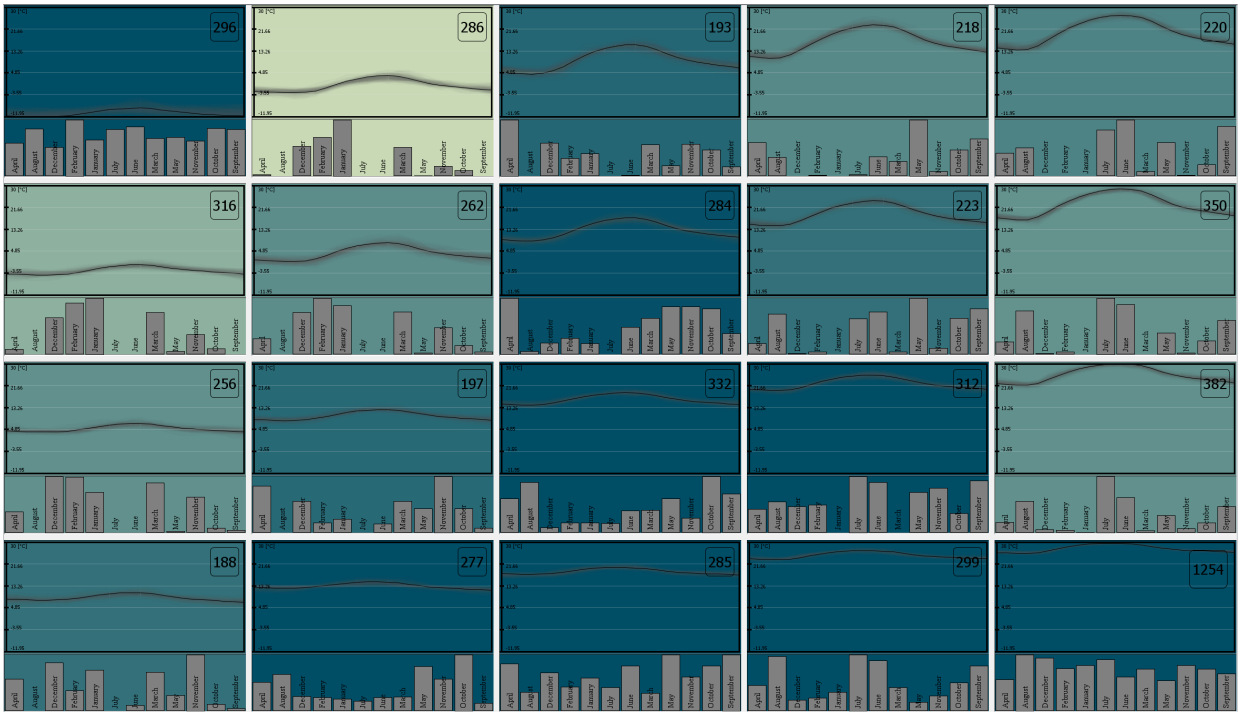


Figure 6.6 *The global normalization of the interestingness scores of all relations also allows the identification of rather uninteresting metadata attributes. The metadata attribute Month achieves many dark green cluster colorings revealing uninteresting relations compared with relations of other metadata attributes. The attribute Month is only weakly related to the data content.*

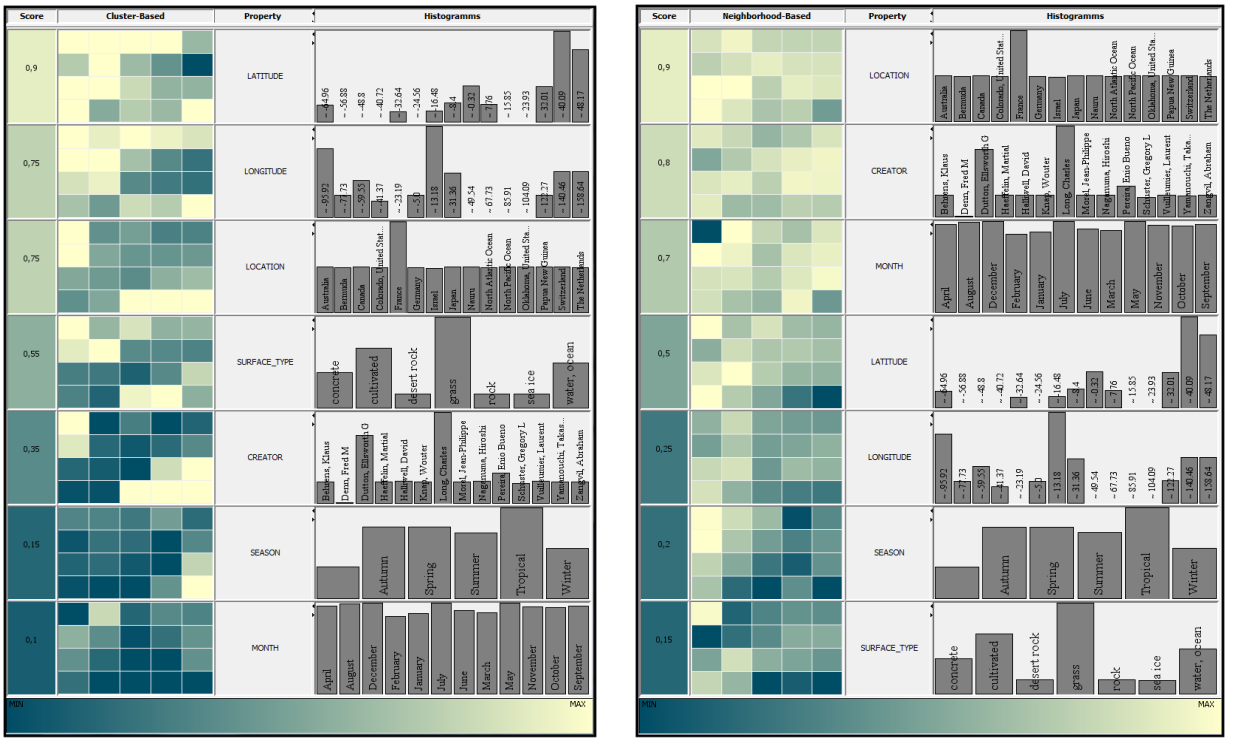
interestingness scores. An example of a rather uninteresting metadata attribute is presented in Figure 6.6. The twelve bins of the metadata attribute *Month* are assigned to the clusters in the content-based overview. However, most of the metadata histograms show homogeneous distributions leading to low interestingness scores for the cluster-based interestingness measure. None of the relations is colored light yellow indicating that not a single relation is interesting compared with interestingness scores of other metadata attributes of the data set. On the contrary, some of the relations are colored dark green indicating that these relations are most uninteresting compared to relations of all metadata attributes.

Our final guidance concept considers a fully automatic ranking for all metadata attributes. The overall goal is to guide users towards the most interesting metadata attributes of the entire data set. In the same breath, we aim at providing an overview of all interesting relations of all metadata attributes. Our solution involves a list-based visualization of all metadata attributes including a preview of the interestingness scores of the individual relations. The order of the list corresponds to the overall interestingness score of every metadata. An example of the metadata ranking is presented in Figure 6.7.

The calculation of the global interestingness score is as follows. First, we calculate the *median* of all interestingness scores $i_M(C)$ (including all combinations of clusters C and metadata attributes M). For the next step, the interestingness values of all metadata attributes M are compared with the *median* interestingness value. For each M, the number of interesting relations above the *median*, divided by the number of clusters $number(C)$, determines the global interestingness of the metadata attribute. The overall score $Score_M$ of each M is calculated as follows:

$$Score(M) = \frac{||\{C : i_M(C) > median\}||}{number(C)}, C \subseteq E, C \neq \emptyset, M \in \mathcal{M} \tag{6.3}$$

As an alternative to the presented scoring function the quantiles information of the entire set of interestingness values can be used to assess the overall score of metadata attributes. As the next step, sorting metadata attributes by their overall interestingness scores provides the attribute ranking as presented in Figure 6.7. In the first column, the overall interestingness score for each metadata attribute is shown. The value range of the score is limited by 1.0 if all relations of an attribute are above the *median*. Accordingly, an attribute described with 0.0 does not contain a



(a) *Metadata ranking for cluster-based interestingness scores.* **(b)** *Metadata ranking for neighborhood-based interestingness scores.*

Figure 6.7 *List of metadata attributes ranked by their interestingness. Each attribute is represented by an overall score, a preview of the colored cluster layout, and a histogram showing the global metadata distribution. The colormap at the bottom depicts the interestingness from low (dark green) to high (bright yellow). In image (a) a ranking of cluster-based interestingness scores is shown. The guidance concept assigns the highest score to the metadata attribute Latitude. Image (b) shows the ranking of neighborhood-based interestingness scores. The highest interestingness score is assigned to the metadata attribute Location.*

singe interesting relation above the *median*. The colormap of the ranking column corresponds to the colormap for the indication of interestingness. In the second column a compact visualization of the content-based layout provides a preview of the associated interestingness values. In the third column the name of the metadata attribute is shown. In the last column a histogram of the global metadata distribution is visualized. To approach the multiple granularity problem $\mathbf{RG}_{C+M3}$, the calculation can be executed for both cluster-based and neighborhood-based interestingness measures. The user can interactively choose which of the two measures is used for calculating the ranking.

6.3.5. Interaction Design

As follows, we present the set of interaction designs provided in our approach. We differentiate between interaction techniques relevant predominantly (a) for the *design* phase and (b) for the phase when our approach is actually *used* in a particular ESS.

Interaction Techniques for the Design of the Approach We provide a set of interaction designs to enhance the usability of the interface. The interactions are demonstrated in a visual-interactive design prototype. On the one hand, this supports the identification of the set of interactive controls which the user also wants to steer in the *use* phase of the ESS. On the other hand, we are able to define the set of parameters which can already be fixed in the design phase. In this way, we are able to provide both a maximum of flexibility and a minimum of remaining interaction complexity. In Figure 6.8 four different configurations of the visual-interactive prototype are shown. On the right of every prototype, a variety of control elements is provided for the adjustment of both visualization and model parameters. An interactive

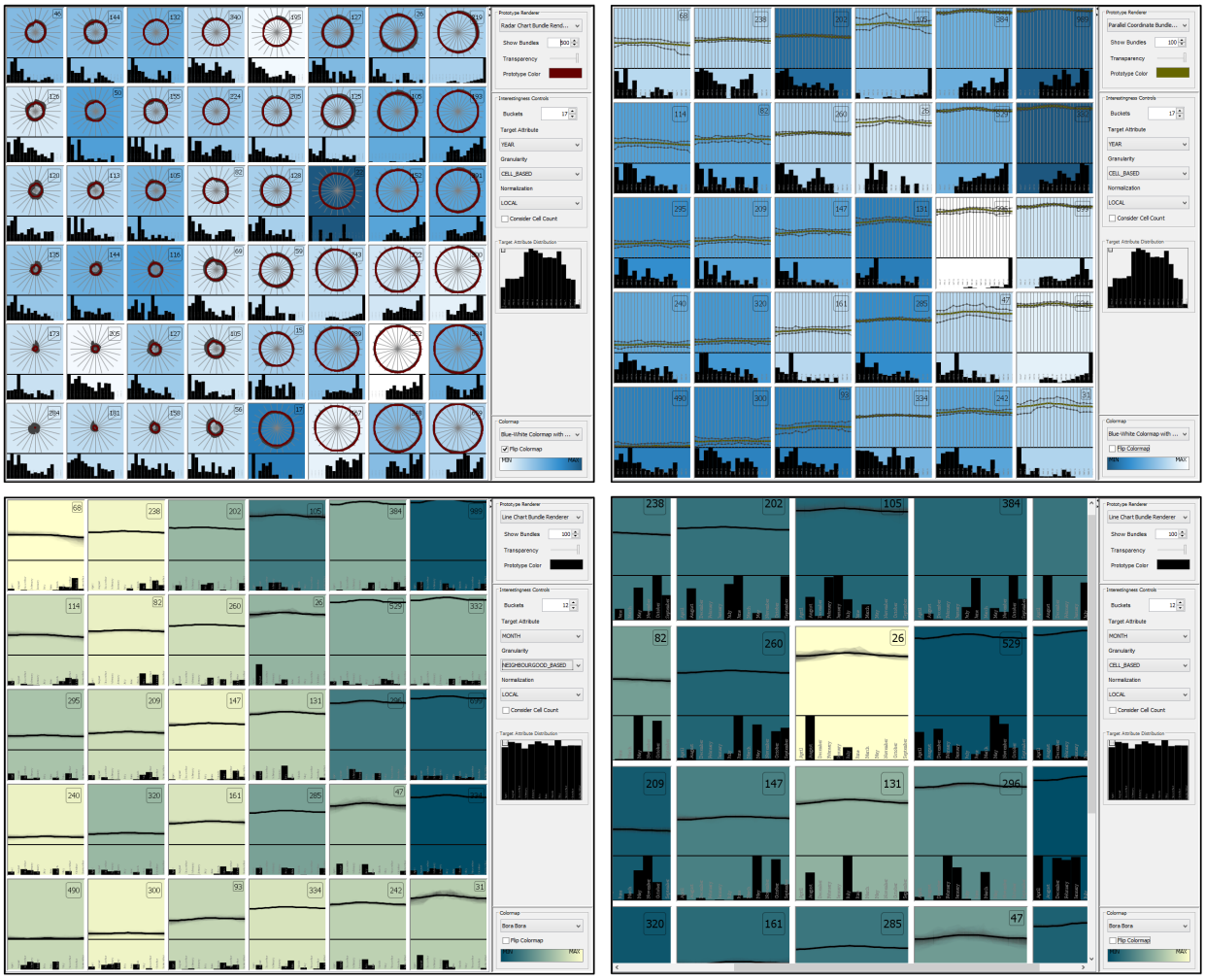


Figure 6.8 *Our approach with four different configurations of model and visualization parameters. Our visual-interactive prototype provides flexible and user-centered application designs. Moreover, we are able to fix a subset of the steering parameters in the design phase, and thus reduce the complexity of the interaction design for the use of the technique.*

parameter, most relevant in the design phase, steers the visual representation of the cluster content. The user can chose between different visual encodings for the high-dimensional time series vector data, as described in our guideline for glyph design presented in Section 5.4.1. In the images at the top a radial design and a parallel coordinates-like design is shown. Similarly, the user can steer the number of shown opacity bands. Hence, the system supports steering the number of shown data elements for each cluster. Interactively coloring the cluster prototype vector is also possible. A further parameter allowing user engagement is the choice of the colormap used for guiding users towards the most interesting relations. In the upper two images a blue-to-white colormap is shown, the left image shows the colormap in a flipped color order assigning the most interesting relations to blue color values. Another difference between the images regards the grid resolution of the SOM cluster layout. The upper left image shows a SOM manifold with 8×6 cells while the remaining images have a SOM grid with a resolution of 6×5 cells.

While the interactive controls, as presented, influence the usability and the usefulness of the technique, the majority of the factors may already be fixed in the design phase of the ESS. This depicts how user-centered design makes room for flexible solutions and helps to reduce the number of interactive controls to a minimum. With the concept of providing a visual-interactive design prototype of the application, we particularly aim at addressing the research goal of user involvement in the design $\mathbf{RG}_{\text{C+M7}}$.

Interaction Designs for the Application of the Technique in an ESS With the integration of the visual-interactive interface into ESS, the user will then be able to reveal interesting relations between data content and attached metadata.

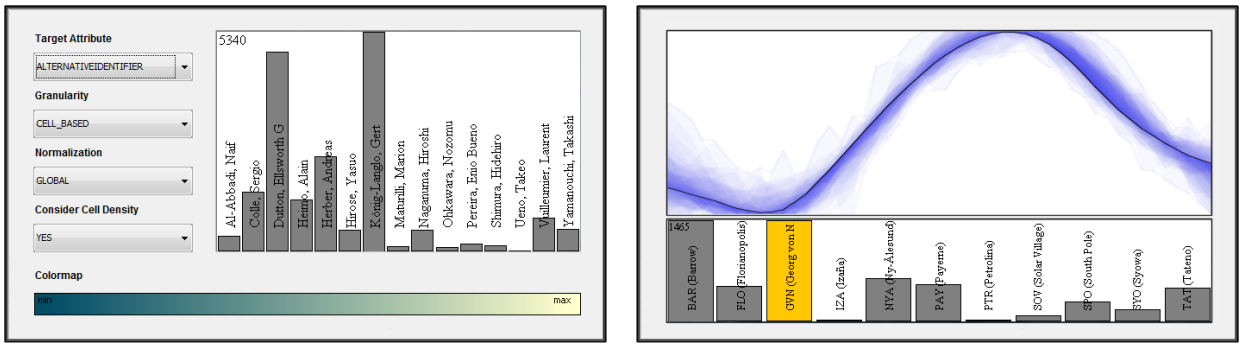


Figure 6.9 *Visualization design for user-steerable components. Left: Controls for the model parameters including the choice of the metadata attribute, the interestingness measure, the normalization of interestingness scores, and the cell density. A histogram shows the global distribution of the selected metadata attribute. At the bottom the colormap is shown. Right: Interaction design for a cluster glyph. The bins of the metadata histogram are selectable. A selection model triggers the bin selection to all cluster glyphs for an enhanced comparison of bins.*

For that purpose, we provide interactive controls allowing users to steer relevant model and visualization parameters in the application phase. A highly relevant interaction design for the use in an ESS regards zooming and panning. As a result, the user is able to facilitate interactive information drill-down. In the lower right image of Figure 6.8, the user applied zooming and panning on the content-based overview. The 6×5 grid of the SOM-based layout is enlarged for the detailed exploration of local relations. In the example, a light yellow cell is centered showing an interesting cell-based relation. The enlargement of the layout by the zooming interaction can be carried out until a single cluster fills the entire display space. In this case, the glyph-based visual representation of a single cluster is shown at the full level-of-detail. An essential user parameter is the choice of the targeted metadata attribute in combination with the number of categories of the metadata histogram visualization. In the upper two examples of Figure 6.8 the metadata attribute *Year* is selected. The data set consists of 17 consecutive years of data all assigned to a single bin. In the lower two examples the metadata attribute *Month* is selected showing 12 bins, accordingly.

Other relevant steerable parameters of the model are the choice of the interestingness measure and the local and global normalization applied to the guidance concept. We present an example of a compact visual-interactive interface for steering model parameters in the left image of Figure 6.9. The image on the right shows the selection concept for single bins of metadata histograms. In this example, the third bin is selected including the measurements taken at the station (GVN-Georg von Neumayer, Antarctica). The selection is triggered to all metadata histograms allowing the enhanced localization of the selected bins. An example showing metadata histograms including selections is presented in Figure 6.10. The metadata attribute *Creator* is selected showing all researchers who contributed measurements to the collection of primary data. The bin of *Dr. Ellsworth G. Dutton* is selected, a famous researcher in Earth observation who unfortunately passed away during the collaboration in the VisInfo case study. It can be seen that the data content contributed by Dr. Dutton is predominantly located at the top left corner of the content-based overview. In other words, we identify an interesting relation between the contributor and his measurements in comparison to the remaining data content. These type of findings can easily be identified with the selection concept. The afro mentioned region in the content-based overview characterized by the measurements of Dr. Dutton is also marked as highly interesting. According to the neighborhood-based interestingness measure these relations consist of both similar data content and similar metadata histograms revealing ‘super-cluster’. The global normalization technique of the interestingness values is chosen, indicating that the neighborhood-based interestingness scores of these relations are among the most interesting values in the complete body of information. The latter example concludes the set of provided interactive capability of our approach. With the individual steering parameters and the involvement of the user in the interaction design, we achieve the research goal of meaningful interaction designs $\mathbf{RG}_{\text{C+M6}}$ and the research goal of involving the user in the design $\mathbf{RG}_{\text{C+M7}}$.

6.3.6. Evaluation

In the course of the technical description of our approach, we presented a variety of images to demonstrate our concepts and techniques using examples. In particular, we demonstrated our cluster-based interestingness measure in Figure 6.5 for daily temperature patters which we related with the metadata attribute *Latitude*. This example shows how users are

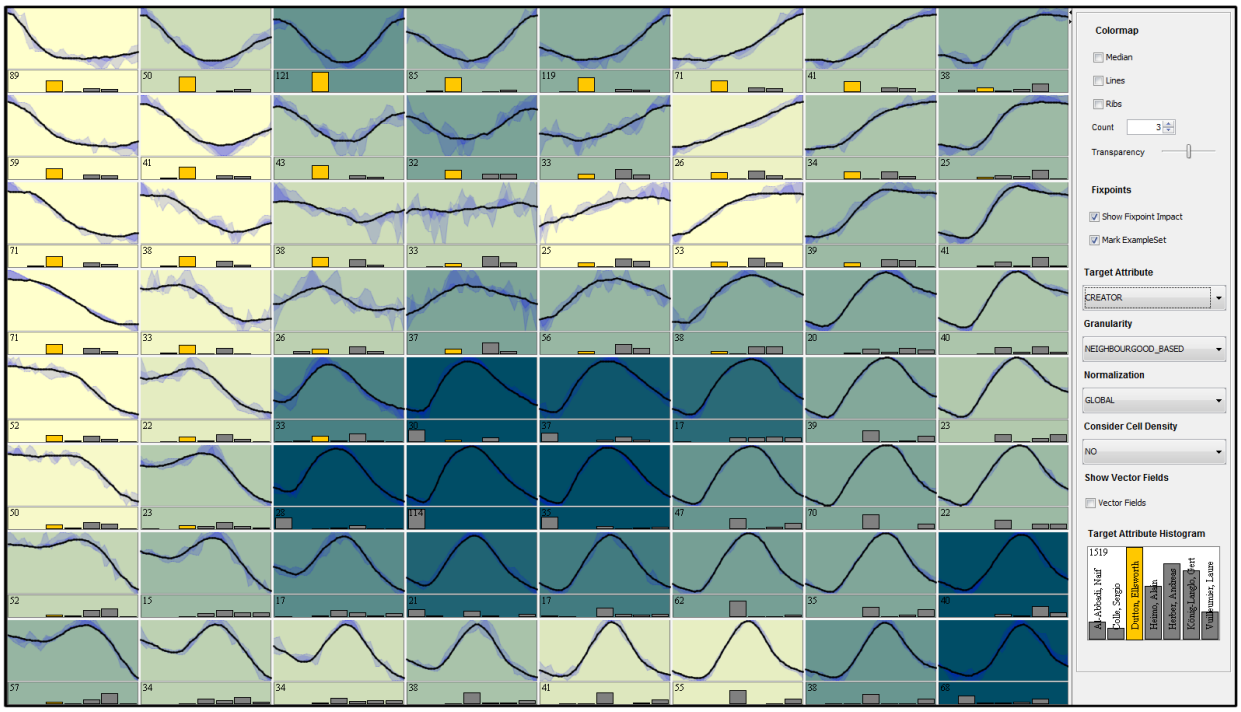


Figure 6.10 *Example showing the guidance concept based on the neighborhood-based interestingness. The upper left region of the layout is marked as most interesting (light yellow) revealing that this region consists of similar content and similar metadata histograms. The global normalization of the interestingness coloring expresses that this finding is most interesting with respect to all metadata attributes of the data set. One of the bins is selected, enabling users to localize the bin in all histograms without mental effort. This also introduces the insight that the respective bin of the metadata attribute only occurs at the upper left of the content-based layout.*

guided in the identification of most interesting relations. Various clusters share interesting relations with the chosen metadata attribute. In addition, we demonstrated the usefulness of the guidance concept for neighborhood-based interestingness in Figure 6.6. This example shows that the metadata attribute *Month* is rather uninteresting with respect to the given data content. Our hypothesis for this insight was that multiple measurement locations all over the world are included in the data set. In additional tests, we validated that our hypothesis was correct. When using our technique for a single location on Earth, we reveal high interestingness scores for the metadata attribute *Month*. In this case, seasonal effects can be observed, detecting that the temperature content is dependent on the month of the year, as expected. In Figure 6.7, we presented a proof-of-concept example of the list-based guidance concept. It can be observed that both interestingness measures assigned high scores to spatial metadata attributes, i.e., the *Latitude*, the *Longitude*, and the *Location*. Recalling the dependency between temperature measurements and the spatial distribution of the measurements on Earth the interface ranked relevant attributes to the top. Finally, we recall Figure 6.8 where four configurations of the visual-interactive prototype are shown. The images show the high degree of individualization which can be obtained within the design phase, by involving the user.

In our corresponding publication, we carried out two real-world usage scenarios showing the usefulness of our approach [BRS*12b]. We included an Earth observation data set, similar to the primary data set applied in the VisInfo case study presented in Section 7.1. The usage scenario was carried out (1) to prove our two interestingness concepts and (2) to give illustrative examples of how our technique can be used by domain experts to seek interesting relations in a complex and previously undiscovered data set. In the course of the usage scenarios, we identified a variety of interesting relations which gave rise to an in-depth analysis. We proved the correctness of our data findings by collaborating with domain experts from the Alfred Wegener Institute (AWI) for Polar and Marine Research in Bremerhaven, Germany.

6.4. Mapping Data Content onto Metadata Layouts

In today's search systems, such as DLs, information is organized, preserved, and made publicly available. Depending on the quality and the quantity of the attached metadata, many document types can be a subject of query-response concepts. In metadata-based search systems, metadata information is used as a means of full-text search allowing the user to exercise known-item search and fact retrieval (cf. Sections 2.1.3 and 2.2.1). Today, many data repositories (e.g., for time-oriented primary data) are mainly accessed by the metadata attributes of the documents, e.g., time or location of observation, or the name of the creator of the respective document. In Chapters 4 and 5, we presented guidelines and techniques about how content-based access and content-based overviews can facilitate ES for such complex data types. Subsequently, we take the relations between metadata attributes and the data content additionally into account. We present a novel technique supporting users in seeking relations between metadata layouts by including the data content in the exploration process.

In many search systems most of the attached metadata do not show any relations to each other. As a consequence, only the documents matching a textual query can be considered a valid result. For example, searching for a specific content creator will result in a list of all library objects composed by the content creator. Relations between different metadata entities are neglected. Yet, it would be very beneficial to have a means for the identification of related library objects. To achieve this, large online shops use recommender-based systems or semantical annotations to enrich the set of provided items and relations between items. These concepts require tedious manual annotation or high visitor volume in combination with tracking and analyzing user behavior. For time-oriented primary data, these concepts are not feasible due to the size and the complexity of the data collections, as well as due to the various information-seeking behaviors of domain experts, especially in data-driven research. One solution to reveal relations is the incorporation of data content in the search process. Especially ES concepts can greatly benefit from including data content to support both search and exploration. The research goal which we aim to achieve here is the integration of the structural information hidden in the data content into the metadata-based search process to enhance this important access strategy. In today's search systems faceted search approaches have proven to be very useful. They group documents according to categories of potentially hand-curated metadata, such as 'Author', 'Year', or 'Location'. However, techniques relating metadata (facets) with the associated data content are scarce. Moreover, the appropriateness of current metadata layouts for representing relations to data content is still subject to research. As an example, a list-based arrangement of metadata entities will hardly reflect the intrinsic structure of the associated data content. In this connection, providing appropriate structures for metadata is of particular concern for the representation of relations between metadata and data content.

We present designs for visual-interactive metadata layouts allowing users to navigate the metadata search space, discover relations between metadata and data content, and access individual primary data documents. Based on the notion of metadata *attributes*, users are able to select a specific metadata field of interest. Our approach is based on the definition of similarity between individual metadata *entities*. We automatically compute the similarity of metadata entities based on the underlying data content. As the next step, we create similarity-preserving layouts of metadata entities in 2D, based on a force-directed layout. As a result, these metadata layouts can be used as a map-metaphor to interactively browse the metadata search space. The visual exploration of interesting relations is additionally supported by showing a content summary of the data elements for each metadata entity. Our approach supports new information-seeking behaviors which make it appropriate for the integration in ESS. As an example, our technique supports analytical questions, like "which domain experts measured Earth observation patterns most similar to mine?", or "which locations on Earth produce the most similar measurements?". We show the applicability of our technique for a data set from the PANGAEA data warehouse [PAN]. In a case study, we show the practical applicability of our search and exploration display by incorporating domain experts from the Earth observation field (cf. [BRS*12a]).

6.4.1. Relation Definition

Our approach postulates the incorporation of data content to reveal interesting relations between the entities of metadata attributes. The structural information gained from the data content is used for the design of metadata layouts. In this way, our visual-interactive solution reveals relations between metadata entities, as well as relations between metadata entities and associated data content.

The entities of a given metadata attribute serve as the basis for the partition of the data collection. Thus, while we applied clustering for content-based overviews in Chapter 5, in this chapter a metadata-based binning technique implements the data aggregation step of the reference workflow (cf. Figure 6.1). As a result of the binning process,

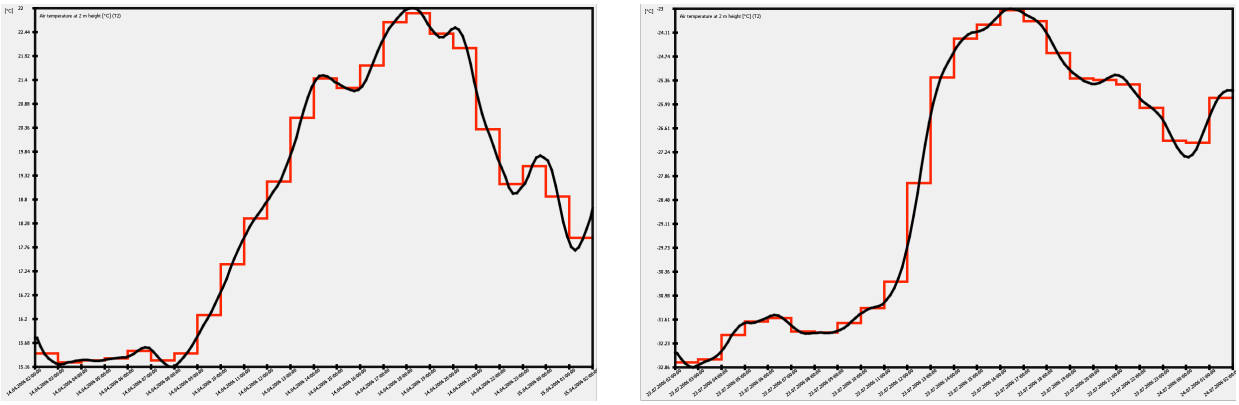


Figure 6.11 *Visual prototype for the content-based access to time-oriented data. Two daily temperature patterns are shown. While the shapes of the time series look similar the value domains are entirely different (left: around* $20°C$*, right: below* $-20°C$*). The involved domain experts confirmed that the comparison of both the absolute value domain and the relative shape progression may lead to interesting relations and insights.*

we reveal a data aggregation result where each data subset corresponds to a metadata entity. Each bin can be described by a set of associated time series FVs. This association between metadata entities and these time series FV defines the relation between metadata and data content. As a precondition for working with FVs, we require a content-based access strategy. This is why we use the visual-interactive system for time series preprocessing presented in Chapter 4. Salient factors to achieve are the quality of the time series, the compact representation of time series FVs and the definition of a similarity measure. As described in Chapter 4 the step of time series preprocessing and similarity definition allows a high degree of user involvement $\mathbf{RG}_{\text{C+M7}}$. The examples shown for this technique correspond to the data set of the VisInfo case study presented in Section 7.1. For this technique, the actual solution for time series preprocessing and similarity definition is inspired by the usage scenario presented in Section 4.4. We segment the time-oriented data content into daily patterns starting at midnight (true local time). In Figure 6.11 two daily time series patterns are shown (black linechart). While the shape of the patterns looks similar the value domain of the patterns is very different. For the domain experts engaged in this approach it is relevant to consider both the relative and the absolute values of a daily time series pattern. That is why we apply two different normalization strategies in the final visual-interactive solution leading to two different preprocessing workflows. In Figure 6.11, we also visualize a preview of the applied time series descriptor output. The red lines indicate the result of the Piecewise Aggregate Approximation (PAA) descriptor [KCPM01]. With the assignment of the data content to the individual metadata entities and the transformations applied to the raw time-oriented data content, we provide a solution for relating metadata and data content $\mathbf{RG}_{\text{C+M1}}$.

6.4.2. Assessing the Interestingness of a Relation

A prerequisite for the design of meaningful metadata layouts is revealing some sort of structure from the body of available information. It follows that we describe how our technique enables the definition of interestingness between two metadata entities. The association between a metadata entity and the binned data content corresponds to a one-to-many relation. As an example, multiple measurements may exist for a given location on Earth. Our solution uses this one-to-many relation between metadata entities and the sets of data content. From the data aggregations' perspective, the sets of data content are equal to the output of a clustering algorithm. Thus, we can borrow distance measures for data elements and clusters as an implementation of the definition of interestingness. In this way, the interestingness of relations between metadata entities can be expressed by the similarity of the associated data content. Figure 6.12 illustrates the concept of content-based similarity for metadata entities. In the orange part of the illustration, the question of similarity between a 'Scientist A' and a 'Scientist B' arises. For both metadata entities a set of associated documents exists, which means that both scientists have created some data content. In the blue box the concept of content-based similarity is explained. The data elements of both metadata entities form individual clusters which can be compared by similarity measures. The output of the similarity measures is assigned to the relation between the two factored metadata entities. To measure content-based similarity between two data elements, we chose the Euclidean distance in combination with the PAA time series descriptor (cf. Section 4.4). Other combinations of measures and descriptors are possible. For example, if the users' notion of similarity prefers a 'cross-bin' similarity measure the

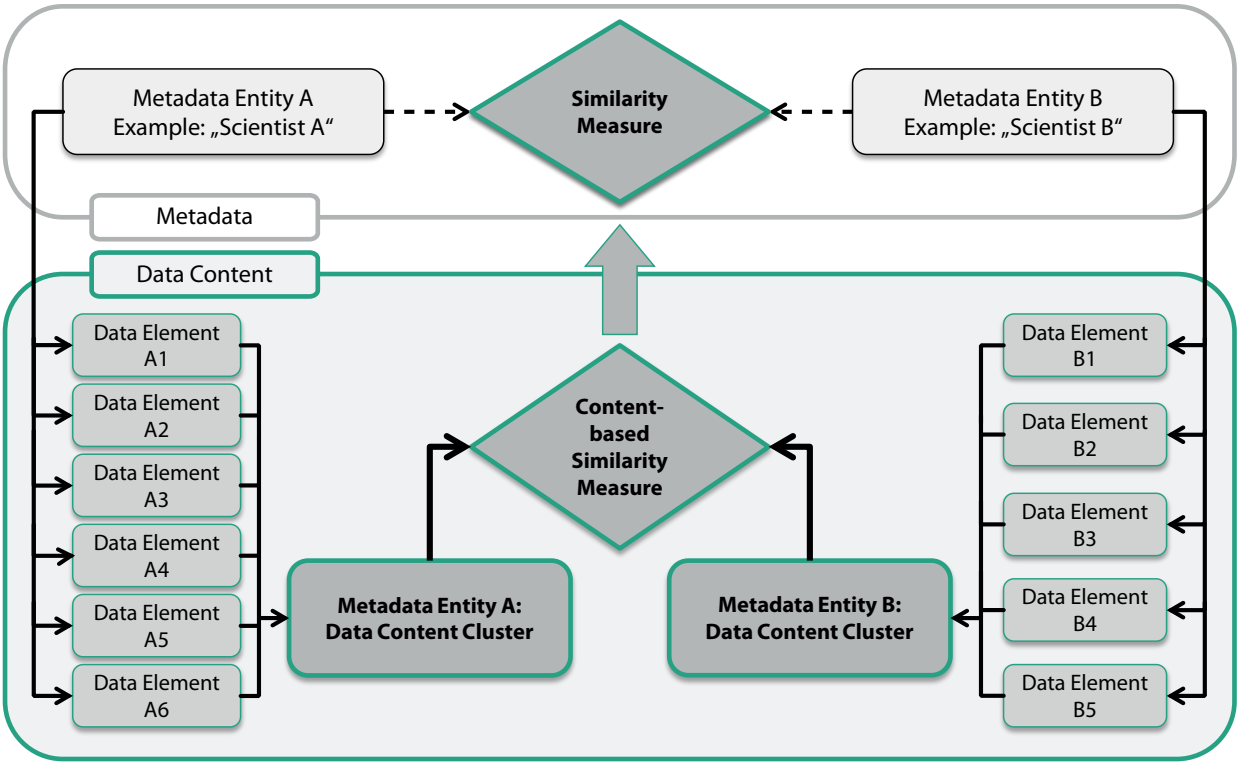


Figure 6.12 *The concept of metadata similarity which is based on content-based similarity measures. The content of any two metadata entities is compared with a content-based similarity (distance) measure. The output defines the relative similarity of two given metadata entities. Using the example of scientists, 'Scientist A' and 'Scientist B' can be compared and related to each other on the basis of their measured data content.*

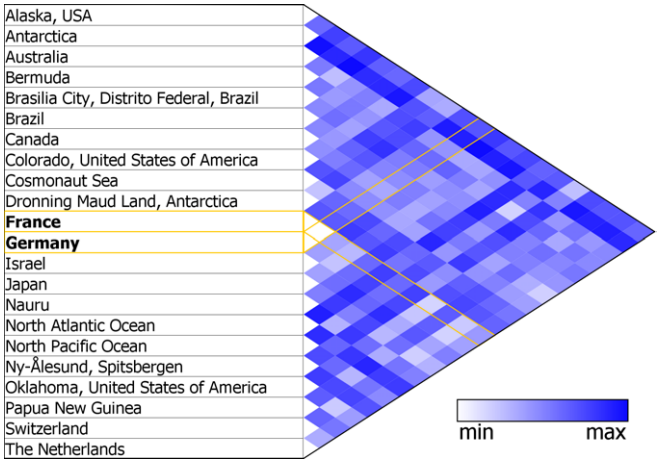
Dynamic Time Warping (DTW) algorithm can be used to find a more flexible alignment between two time series FVs. Involving users in the design of the similarity concept supports the creation of meaningful solutions $\mathbf{RG}_{C+M7}$.

The complexity of defining content-based similarity measures, however, is greater than comparing single FVs with each other. Similar to hierarchical clustering algorithms a similarity model is needed to be able to compare $m : n$ data elements (cf. Section 5.2.1). Similar to the linkage criteria for hierarchical clustering, we can (a) apply single linkage, (b) compute the average of each pair-wise distance, (c) average the descriptors for each set and then compute the distance, (d) determine the median descriptor for each set and compute their distances. We chose option (c) as a robust and yet discriminative variant of comparing two sets of data elements. The choice of the linkage criterion also allows involving users. Other models may also be relevant and individual solutions should correspond to the expectations of the included user group $\mathbf{RG}_{C+M7}$. Our concept of content-based similarity allows different implementations, as long as the descriptor computation result is a vector and the applied distance measure is a true metric. With our implementation, we are able to assess the interestingness of relations $\mathbf{RG}_{C+M2}$.

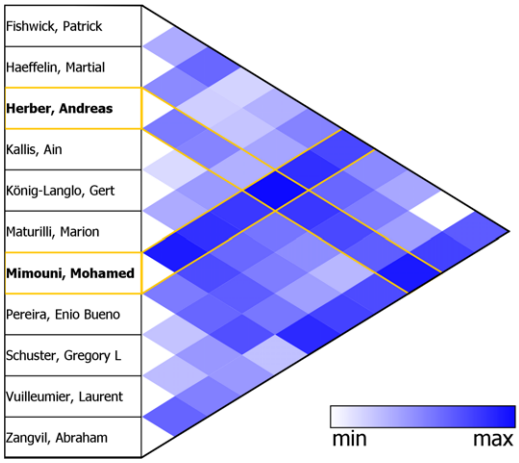
6.4.3. Overview of the Relation Space

We first provide an overview of the pairwise relations between metadata entities. Second, we present the result of our content-based layout technique for metadata entities. Finally, we show how data content can additionally be visualized on top of the metadata layout to gain a full overview of the relation space.

Visualizing Pairwise Distances of Metadata Entities As a result of the interestingness definition, we are able to express the relations between any two metadata entities with the distance between their respective features. A visualization of these pairwise entity distances is presented in Figure 6.13 using the example of the metadata attributes *Location* and *Creator*. A colormap represents the relative interestingness of any pair of metadata entities. White diamonds in the distance matrix indicate the most similar (interesting) entity relations. In return, blue diamonds



(a) *Distance matrix showing the interestingness of entity relations of the metadata attribute* Location*. With the content-based distance measure (temperature measurements), the metadata entities* France *and* Germany *have the strongest relation to each other in the entire metadata space. The white cell in the distance matrix indicates the most similar temperature measurements.*



(b) *Distance matrix showing the interestingness of entity relations of the metadata attribute* Creator*. Based on Shortwave-Downward radiation (SWD), the two selected data creators* Herber, Andreas *and* Mimouni, Mohamed *measured the most dissimilar data content (blue cell in the distance matrix).*

Figure 6.13 *Visualization of the pairwise distances of entity relations based on the content-based similarity measure. The most interesting relations between metadata entities are defined by small distances. The visualization of the distance matrix supports the visual exploration of these similarities. In addition, the distance matrix serves as the basis for the design of metadata layouts.*

represent most dissimilar entity relations. With the visualization of the distance matrix, we are able to a) explore these similarities visually, b) validate the similarity measure, and c) involve the user in the design of the content-based similarity measure. Please note that the user can freely choose which metadata property is subject to exploration. Our concept can be applied to any metadata attribute provided in the data set.

Content-Based Layouts for Metadata Entities With the entity distance matrix, we generate the missing structural information for a given metadata attribute. This structural information facilitates the design of layouts for metadata entities on the screen space. The layout aligns metadata entities in a 2D map, and thus serves as a baseline for both meaningful visualization and interaction designs. A preview of a metadata layout is presented in Figure 6.14. The user is able to browse the map of metadata entities, which are represented by large glyph designs. Recalling the guideline for layouts of aggregated data (cf. Section 5.5), two different classes of layout techniques are conceivable. First, projection-based layouts can be applied to the provided distances matrix. Second, the pairwise relations between metadata entities can serve as a basis for force-directed layouts. We define a set of requirements to balance the applicability of the two classes of layouts. An important requirement is the preservation of topology, i.e., that similar entities are depicted close to each other while dissimilar entities should have a greater visual distance. That way, the user can explore similar metadata entities in the vicinity of a focused entity. While the preservation of topology is highly relevant, it does not support the decision process since both projection-based and force-directed layouts produce topology-preserving solutions. The low number of expected metadata entities (smaller than 100) shifts the decision towards force-directed layouts. The decisive requirement for the class of layout techniques is the avoidance of overplotting. As it can be seen in Figure 6.14, it is desirable to maximize the size of any metadata entity glyph. Thus, the layout has to arrange the visual representations of metadata entities with a minimum of overplotting. Users will then be able to localize related entities close to each other with a minimum of overlap. To this end, we chose a force-directed layout in favor of a projection-based technique. Consequently, we interpret metadata entities as nodes, and their pairwise similarities as edges, respectively. The weighted edge-repulsion LinLog model serves as layout algorithm for the metadata entity graph [Noa07].

A Visual Summary of the Data Content With the content-based metadata layout, we are able to provide an overview of the entities of a given metadata attribute. The overview of the metadata attribute *Creator* in Figure

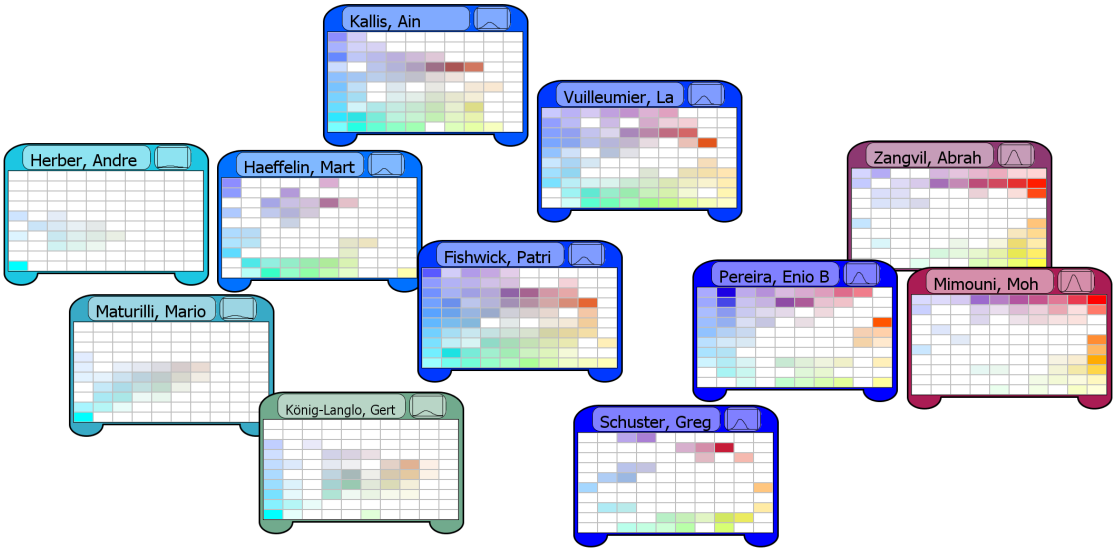


Figure 6.14 *Content-based layout for eleven entities of the metadata attribute 'Creator'. A layout algorithm aligns entities in 2D in a topology-preserving way. The relations between entities are based on a content-based distance measure. A glyph design with a mosaic metaphor enables users to relate metadata entities with their data content. The data creators* Herber, Andreas *and* Mimouni, Mohamed *who measured the most dissimilar data content (cf. Figure 6.13b) are aligned farthest away from each other. A similarity-preserving colormap additionally supports the identification of similar entities.*

6.14 may again serve as an illustrative example. As already outlined, we advocate carrying out a glyph design for metadata entities provided in the layout. Hence, our technique will enable domain experts to gain an overview of the entire relation space (cf. research goal $\mathbf{RG}_{\text{C+M4}}$). On the one hand, the metadata-based overview layout supports the identification of entity-relations. On the other hand, each metadata entity glyph provides an overview of the data content associated with its metadata entity.

To facilitate the latter, we require a content-based overview concept nested in any given metadata entity glyph in the metadata layout. For that purpose, we use the guidelines and techniques for content-based overviews presented in Chapter 5. The design concept of the content summary is presented in Figure 6.15. In our usage scenario, the data content consists of time-oriented measurement curves, each with the duration of one day. Our solution is based on the SOM approach since it provides a topology-preserving layout of the data content and avoids overplotting effects. In addition, the SOM approach exploits 100% of the display space which is higly appropriate for the glyph designs. The SOM consists of a 10×10 grid of cells, each representing a cluster of daily curve patterns. Within each cell, a representative curve pattern and the cluster size is visualized, as suggested in the section about glyph designs for high-dimensional data (cf. Section 5.4.1). For justification of the SOM parameters, we refer to our solutions on quality-based visual cluster analysis presented in Section 5.3. Moreover, the user may be involved in the glyph design process, e.g., for the choice of the SOM grid resolution $\mathbf{RG}_{\text{C+M7}}$. As a last step in the design of the content-based overview, we apply the guidelines and techniques for the utilization of 2D colormaps presented in Section 5.4.2. We map a 2D colormap onto the SOM grid. As a result, we are able to represent the SOM cells with similarity-preserving colors. The color coding of individual cells facilitates localization tasks carried out with the metadata entity glyph. The 2D colormap can be adapted to specific use case scenarios $\mathbf{RG}_{\text{C+M7}}$. For our approach, we choose a color map supporting the notion of warm (red) and cold (blue) temperatures. Note that such color semantics may also lead to false assumptions if the colormap is applied inappropriately.

Enriching the Metadata Entity Glyphs with Content Summaries With the content summary, users are empowered to gain an overview of all time series patterns of the underlying data content. The remaining task is a content summary for every single metadata entity. Every metadata entity defines a subset of the data collection. Based on the global content summary, we calculate a derivate for any given metadata entity. To represent these data subsets visually, visual variables are required supporting the identification these subsets. For this purpose, the design of the entity glyph uses the position information of the SOM grid, the color information of the colored cells. In addition,

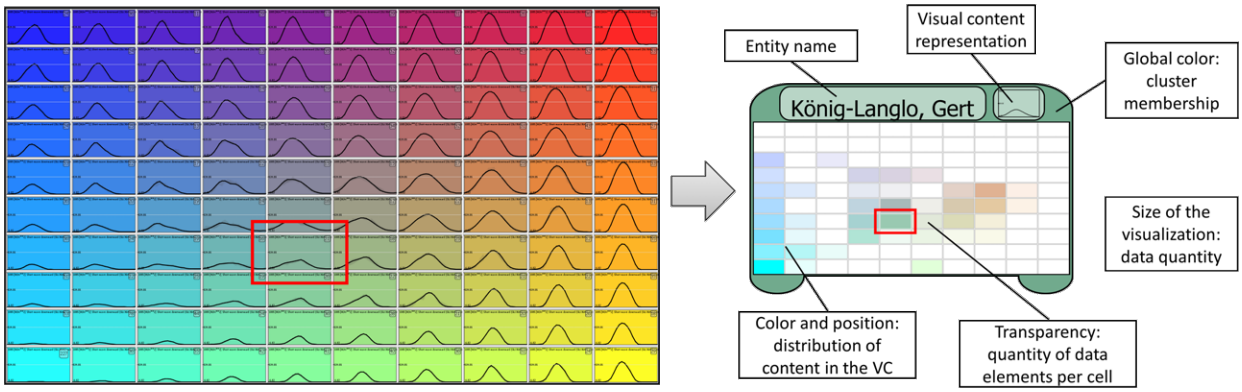


Figure 6.15 *Left: the content summary of all measurements of the data set. Right: the summary of the content of a single entity. A SOM-based overview reveals the most prominent patterns of the time-oriented data collection. A mosaic metaphor is applied in combination with color and transparency value to summarize small proportions of the content summary in every entity glyph. As an example, the data content measured by the data creator Dr. Gert König-Langlo is predominantly located in the lower left of the content summary (colored from green to light blue).*

color transparency is used to indicate the relative frequency of the data subset for the SOM grid. The result of the visual mapping is a mosaic-like glyph design enabling users to identify meaningful subsets. An example of the glyph design for a metadata entity is shown on the right of Figure 6.15. For a particular metadata entity (here the *Creator* Dr. Gert König-Langlo), a mosaic-based fingerprint of the associated data content is shown in the individual content summary. Accordingly, the fingerprints of all entity glyphs again result in the entire content summary as presented on the left of Figure 6.15. Another visual encoding regards the border of an entity glyph. The border is colored by the color of the most representative pattern in the content summary. In the example of Dr. Gert König-Langlo, the most representative pattern is green. Additional visual encodings in the entity glyph are the name of the entity and the most representative curve pattern associated with the metadata entity (the cluster centroid).

The entity layout at a glance allows relation seeking between metadata entities. Together with the matrix visualization, domain experts are provided with the means of gaining an overview of entities and entity relations. The SOM-based content summary additionally provides an overview of the data content. Furthermore, the entity glyph design supports the identification of relations between the entities and their associated data content. Finally, the entity glyphs enable users to compare different entities and their data content. In this way, users are provided with a seamless transition between information about metadata and data content. In summary, these techniques demonstrate how users are able to gain an overview of the relation space $\mathbf{RG}_{\text{C+M4}}$.

6.4.4. Guiding Users Towards Interesting Relations

We provide different views on metadata, the data content, and the relations between. These views guide users in the identification of relevant patterns and relations $\mathbf{RG}_{\text{C+M5}}$. Some of the visual encodings in these views are explicitly designed to facilitate user guidance. The *distance matrix* enables users to identify relations between metadata entities with different content-based similarities. A colormap facilitates the assessment of pairwise distances. The color mapping guides users towards different types of patterns in the relation space, such as clusters of ‘dense’ entities or outlier entities. As an example, on the left of Figure 6.13, the location *Antarctica* has very high distances to virtually all other locations. The color coding of the content-based similarity measure allows the identification of the measurement station *Antarctica* as an outlier in the set of metadata entities. As an additional benefit, the distance matrix enables users to localize expected patterns to prove the design of the content-based similarity measure. In addition, the user is supported in the exploration of unexpected patterns.

The *metadata entity layout* aligns similar entities next to each other in a 2D map metaphor. As such, the visualization enables the identification of similar metadata entities and groups of entities. Similarly, the map metaphor enables users to identify most dissimilar metadata entities. Figure 1.7 in the introduction chapter may serve as an example. The *glyph design* further enhances the exploratory means of the metadata entity layout. The color coding of every entity glyph guides users towards similar metadata entities. Moreover, the color distribution in the mosaic metaphor reveals relations between a metadata entity and the underlying data content. In the mosaic plots the transparency value

is yet another important visual variable for guiding users to relations between entities and the data content. These visual encodings are combined with the color coding provided with the static 2D colormap. The colormap facilitates linking the global content summary (the SOM) with the individual content summaries of the entity glyphs. With the 2D colormap, we also achieve the research goal of multiple granularities $\mathbf{RG}_{\text{C+M3}}$. At the granularity of *bin-level* (single entities) the user can reveal relations of every metadata entity in detail. An interesting relation between an entity and the underlying data content exists if the content summary of the entity glyph (the mosaic) shows a distribution of low diversity. Consequently, the user can gain insights, like "the curve progressions of the data content at location X all show bell-shapes", or "the temperature progression at the Antarctica are hardly sun-dependent". The latter finding is a proof-of-concept example carried out with the involved domain experts within the design phase. Furthermore, relations at the *attribute-level* can be identified. This class of relations includes the comparison of different entity glyphs revealing relations like "in tropical regions the daily peak temperature is earlier in the day than in other regions of the Earth", which was another insight carried out with the engaged domain experts for Figure 6.16. Finally, a relation on the attribute-level can be revealed if the distribution of the data content in most entity glyphs is low. In this case, the data content discriminates metadata entities very well; a relation between the data content and the metadata entity can be assessed.

An example is presented in Figure 5.22 in the guideline for choosing an appropriate layout technique (cf. Section 5.5.1). The 21 metadata entities in the layout show a significant partition of the data content. In the content summaries some of the locations have only red data content, others only have yellow content, and again others have data content colored blue. There seems to be a global (co-)relation between the metadata entity *Location* on Earth and the temperature patterns of the measured data content. Only at European locations a variety of data content was measured revealing heterogeneous temperature progressions. The explanation of the involved domain experts was two-fold. On the one hand, European temperatures are moderate, and thus may collide with warm and cold climate zones. On the other hand, the annual temperature amplitude in Europe is large compared with other locations in the data set.

6.4.5. Interaction Design

We briefly present the most relevant steering parameters. In the design phase of the technique, the definition of the content-based similarity measure is of particular importance. We recommend the utilization of visual-interactive time series preprocessing and similarity definition techniques as described in Chapter 4. As a result, the tight coupling of user interaction and model calculation can facilitate the iterative design process $\mathbf{RG}_{\text{C+M7}}$. Similarly, the steps of data aggregation, visual mapping, and layout can be carried out in a visual-interactive and user-centered way (cf. Chapter 5). As an example, the SOM-based content summary provides zooming and panning interaction. The user is able to focus on specific regions of the content-based layout. The glyph design of the high-dimensional data content is level-of-detail capable.

In the application phase of our technique, we provide a user control for the selection of the targeted metadata attribute. This enables users to explore the possibly large set of metadata attributes in the ESS and identify the most relevant attributes for faceted search interaction. A model parameter that can be subject to interaction is the choice of the content-based similarity model. Depending on the expertise of the targeted user group, this VA parameter may be greatly beneficial. Our approach allows domain experts to adjust the content-based similarity model in a way that relative or absolute values can be compared. The matrix visualization showing the pairwise content-based distances of metadata entities (cf. Figure 6.13), provides two principle user interactions. First, the diamonds of the distance matrix can be clicked. The selection highlights the corresponding metadata entities and facilitates the lookup of selected entities. In this way, the user is supported in relating interesting color codings with the particular metadata entities. Second, a user may want to identify the similarities of one or many particular entities. By selecting these entities in the list, the corresponding diamonds are highlighted. In this case, metadata information serves as a starting point for the exploration and the corresponding similarities between particular entities are highlighted. Finally, the metadata entity layout supports zooming and panning interaction. Thus, the user is able to narrow down the metadata space to metadata entities of interest. The level-of-detail concept of the entity glyph supports the analysis of metadata entities on different levels of abstraction (cf. Section 5.5.2). Depending on the zooming level, the glyph supports the in-depth analysis of single metadata entities and the relation between the entity with the data content. With the presented steering parameters, we accomplish the research goal of including interaction designs $\mathbf{RG}_{\text{C+M6}}$.

6.4.6. Evaluation

We briefly demonstrate the usefulness of our technique with the primary data set applied in the VisInfo case study presented in Section 7.1. For two in-depth real world application examples, we refer to our publication [BRS*12a]. The example data collection of this usage scenario is provided by the PANGAEA [PAN] data repository, a data warehouse for time-oriented Earth observation measurements in the research areas of water, ice, sediment and atmosphere. Our data set focuses on atmospheric weather measurements, gathered in the scope of the Baseline Surface Radiation Network (BSRN) [BKLS12], a PANGAEA compartment. In general, data of BSRN regards the development of radiation and meteorological measurements over time, expressed by up to 100 physical parameters, recorded up to a temporal resolution of one measurement per minute. Common physical units include atmospheric pressure, relative air humidity, temperature and a variety of radiation-based measurements, such as Shortwave-Downward radiation (SWD). The targeted user group consists of domain experts working in the field of Earth observation. The domain experts are from the Alfred Wegener Institute (AWI) for Polar and Marine Research in Bremerhaven, Germany. According to the usage scenario described in the chapter about time series preprocessing and similarity definition (cf. Section 4.4), we handled quality considerations for the time-oriented primary data set. In addition, we applied the PAA descriptor as a compact and yet precise FV transformation. Based on an inquiry of the domain experts, we provided solutions for relative and absolute measurements of daily temperature patterns. In this usage scenario, all example images are based on relative curve progressions.

Figure 6.16 presents an overview of the workflow. At the upper left, the visual content-summary is shown. At the lower left, the visualization of pairwise distances between metadata entities of a given metadata attribute is depicted. Finally, the content-based metadata layout for the metadata attribute is presented.

The SOM-based content summary provides an overview of the variety of daily temperature curve progressions. Due to the chosen relative scenario, we focus on the shape of the daily curves in a relative context in favor of the absolute values of the value domain. Every pattern has a duration of 24 hours starting at midnight. Most of the cells show a temperature progression with a peak somewhere in the middle, i.e., at noon time. The green cells at the upper part of the SOM have a comparably early peak; the highest temperatures are already reached before noon. From the domain experts, we learn that these types of patterns predominantly occur in tropical regions. In these climates it often starts to rain at noon leading to a decrease in temperature. On the right of the SOM in the yellow and orange region, the domain experts identify temperature patterns with a peak in the afternoon (after 50% of the x-axis). These sort of temperature progressions is more often observed in moderate and continental climate regions of the Earth. The left side of the SOM (blue region) reveals temporal patterns following an entirely different progression. One of the hypotheses of the domain experts was that this measurements may originate from locations where the sun does not have such a strong impact on the temperature as we are used to. Especially Arctic and Antarctic regions typically have such differing temperature patterns.

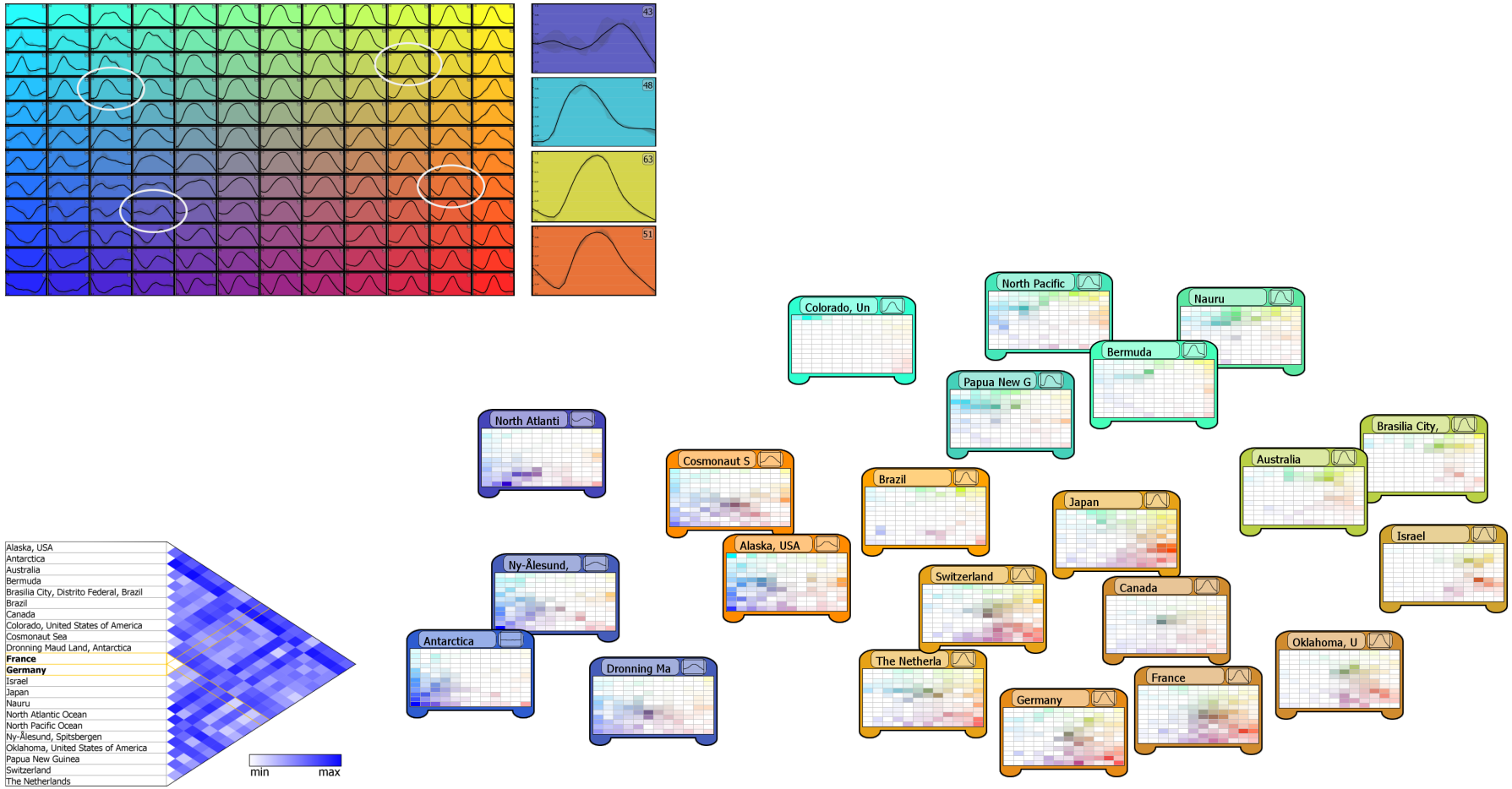


Figure 6.16 *Usage scenario showing the usefulness of the approach. At the upper left the content summary of time-oriented data set is shown. The content-based similarity measure reveals the distance matrix for the entities of a user-selectable metadata attribute. On the right, the content-based layout of metadata entities is provided with a force-directed layout. The 22 metadata entities are aligned in a topology-preserving way while simultaneously avoiding overplotting. In this usage scenario relations between the metadata attribute 'Location' on Earth and the daily progression of relative temperature patterns are demonstrated.*

The distance matrix visualization shows the pairwise distances of metadata entities of the attribute *Location* on Earth. Altogether, 22 measurement stations (metadata entities) are included in the data collection. The whitest cell, and thus the most similar data content is assessed between the entities *France* and *Germany*. This may serve as a proof the concept since the two locations also share spatial proximity. Other very similar entities are, e.g., *Bermuda* and *North Pacific Ocean*. Note that we expect these interesting relations to be aligned close to each other in the downstream metadata layout. The most distant entity relations are between the location *Antarctica* and many other locations. The measurements from the station at *Antarctica* seem to define an outlier pattern.

The metadata on the right proves these expectations. The tuples *France* and *Germany*, as well as *Bermuda* and *North Pacific Ocean* are both aligned close to each other. In addition, the layout allocated the outlier pattern *Antarctica* to a border region at the lower left of the display. With the metadata entity layout an overview of the entity relation space is provided. As a result, the domain experts are not only able to lookup the pairwise relations between metadata entities, but can also identify groups of metadata entities sharing similar data content. In other words, we reveal a topology-preserving visual clustering of categorical metadata entities. At the lower left blue measurements are located. The lookup in SOM-based content summary reveals that these metadata entities consist of many measurements with a low sun-dependency, as discussed earlier. At the center of the layout a large cluster of brown entities can be identified. The comparison of the mosaic plots reveals that many individual content summaries (especially the ones of *France* and *Germany*) look very similar. The domain experts interpret the brown entity group as the cluster of moderate and continental climates, associated with many countries from Europe and USA. An interesting insight provides a group of five green entities at the top of the layout. The temperature progressions assigned to these metadata entities have their peak before noon. In fact, all 5 entities represent locations on Earth where tropical climates can be assessed. The domain experts also identified unexpected patterns. For example, the location *North Atlantic Ocean* predominantly consists of blue and purple measurements. However, some of the daily curve progressions are assigned to the yellow and orange region of the SOM grid. Such a relation between data content and metadata can serve as a basis to prove existing hypotheses, or to formulate new hypotheses. If an *interesting* relation is identified, researchers are able to use this new information as a starting point for an in-depth analysis.

With the overview of the workflow as provided with Figure 6.16, domain experts also have a means of communicating data-centered research results $\mathbf{RG}_{C+M8}$. Different relations between metadata and the associated data content be explored, looked-up, and communicated at a glance. While the interaction designs support users in the exploration of the provided body of information, the visualizations enable the communication of gained insights.

6.5. Relation Seeking in Multi-Attribute Data

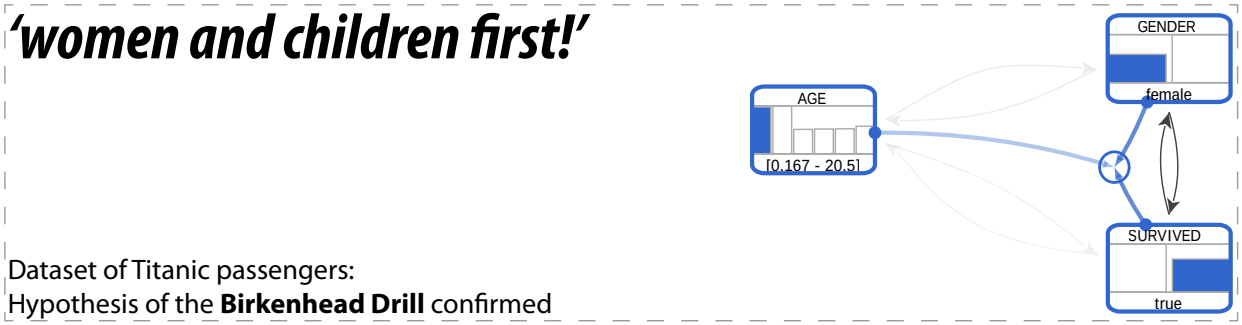


Figure 6.17 *Relation seeking in multi-attribute data. With our techniques, we are able to confirm the expected Birkenhead Drill: 'women and children first!' in a data set of titanic passengers.*

In Sections 6.3 and 6.4, we presented two techniques for a) mapping metadata onto content-based layouts and b) mapping data content onto metadata-based layouts. These techniques enable users to reveal relations between this heterogeneous types of data, i.e., the data content and the attached metadata. With the contribution presented in this section, we neglect the distinction between data content and metadata. In return we resolve challenges of multi-attribute data possibly drawing on the data content and varieties of metadata attributes. In Section 2.2, we have already outlined *mixed data* as a data type of multiple heterogeneous attributes. Almost every research domain generates mixed primary data sets. Primary data sets consisting of mixed data may contain multiple attributes of different data types. Note that it is also possible to include time-oriented data content into mixed data sets. This is

either achieved by aggregating the temporal domain to categories (e.g., the month of a year, or the hour of a day), or by assigning the dimensions of a time series FV to the set of mixed data attributes.

Mixed data sets have the characteristic that interesting relations may exist between multiple attributes. It can be assumed that mixed primary data contains undiscovered knowledge in many attributes, which makes it attractive for a variety of exploratory data analysis approaches. That is why we neglect the definition of an a priori target attribute as carried out for the data content in Section 6.3 and for metadata attributes in Section 6.4. Instead, we face the challenge of revealing interesting relations between all attributes of the entire body of information.

For specific domain problems, exploratory data analytics approaches are already successfully developed. However, state-of-the-art visual data representation techniques from the IV and VA domains are still rarely applied in many data-centered research domains (cf. Sections 2.2.1 and 2.4.1). On the contrary, many domain experts actually perform at least parts of their analyses using general purpose tools - most notably, Excel. This mismatch leaves the question, whether VA techniques can be used as a generic baseline technique for mixed primary data. Similar to the two latter techniques, the solution should achieve the eight research goals outlined in Section 6.1.2. In addition, the technique needs to cope with the heterogeneity of mixed data types and the multitude of available attributes. Finally, interesting relations revealed in the data set may cover more than two attributes of the data set. Instead, the technique should also guide domain experts to the most interesting relations between multiple attributes. An illustrative example is presented in Figure 6.17, showing the analysis of the passengers' data set of the Titanic disaster. The displayed finding confirms the 'Birkenhead Drill': "women and children first!". It is remarkable that this interesting relation between three attributes (*Gender*, *Survived?*, and *Age*) is only relevant for one observation for each attribute (*female*, *true*, and *0.167-20.51*). Thus, the essential research challenge for relation seeking in mixed data is considering both attributes at a glance and their respective observations (bins). That is why our novel technique is also inspired by feature selection and subspace clustering approaches (see our particular publication for a more in-depth discussion [BSW*14]).

We contribute a technique which enables users to identify direct and indirect relations between multiple attributes in mixed data subsets. Our techniques support the analysis of relations on the attribute-level and on the bin-level, a relation between bins of two attributes represents the finest granularity. The interestingness of bin relations is based on statistical dependency measures. The approach provides complementary linked views showing bin relations in a matrix, a node-link, and a list metaphor. The number of displayed bin relations is interactively steerable. Interactive drill-down based on expert knowledge and algorithmic support is provided in the matrix view where the user can apply filters to focus on most interesting bins and attributes. The node-link view enables users to analyze multivariate bin relations. For that purpose, highlighting of multivariate bin relations is supported by interactive subspace clustering. To support different information-seeking behaviors of the user, three clustering techniques are provided. We demonstrate the applicability of the technique for the data set on Titanic passengers as a proof-of-concept example.

6.5.1. Relation Definition

The atomic data object of our technique is a single *bin* of a single attribute. In this way, we abstract the data from single data elements to sets of data elements sharing some common property, such as metadata entity. We assign a relation between any two bins of two different attributes. Every bin represents a set of data elements grouped by the value domain (the observations) of a given attribute. For example, a bin can be the group of people between 40 and 50 years, all female participants of a study, or a cluster/class of time series patterns with a constant linear upward trend. For all n attributes of the mixed data set, the data is aggregated to bins revealing n different partitions of the data set. We define a set of requirements relevant for the data aggregation step. First, the individual aggregations support the *unification* of the input attributes, and thus for coping with data sets of heterogeneous attribute types. Second, we achieve a basis for the class of algorithms calculating the *interestingness* of a relations (cf. research goal $\mathbf{RG}_{\text{C+M2}}$). Third, bins enable the exploration on multiple granularities (the bin-level and the attribute-level) (cf. research goal $\mathbf{RG}_{\text{C+M3}}$). Finally, the abstraction from individual data elements to bins makes the approach scalable for large primary data collections. Based on these requirements, we reveal bins which can serve as a solid basis for the approach. The association between any two bins (of different attributes) defines the relation of the technique $\mathbf{RG}_{\text{C+M1}}$.

The binning process has a great influence on the calculation of interesting relations. Depending on the attribute type, different binning approaches exist. As an example, for numerical attributes domain-preserving or frequency-preserving approaches are widely applied. Binning of individual attributes does not necessarily take place in the application phase of our techniques, but can also be carried out as a preprocessing step in the design phase. In any case, we highly recommend to involve the user in the design of appropriate data aggregates $\mathbf{RG}_{\text{C+M7}}$. We store binning results in a config file to accelerate the application of our technique in the targeted ESS. As a consequence, our approach requires

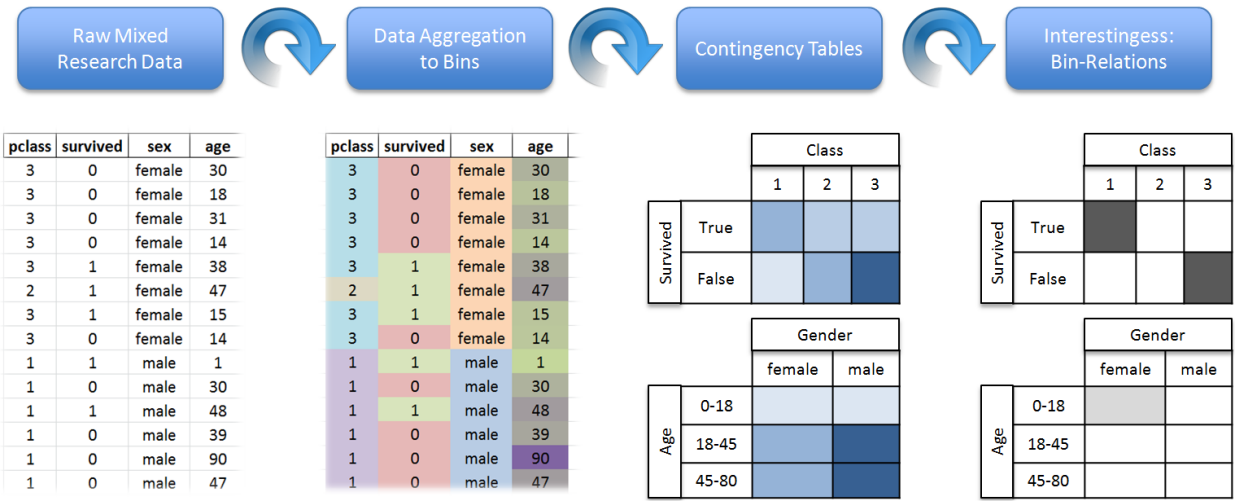


Figure 6.18 *Illustrative example of the relation definition and the assessment of a relations' interestingness. All attributes of a mixed primary data set are aggregated to bins. We define the intersection of two bins as a relation. In contingency tables the number of occurrences of every two intersected bins can be seen. For the assessment of interesting relations, we apply statistical dependency measures which compare something that was assumed about a intersection population and the actual measured population. The example data set shows passengers of the Titanic. The output of our interestingness measure reveals that more people from the 1st class actually survived the disaster as assumed. As a basic assumption the ratio of survivors having traveled in the 1st class should have been equals to the ratio of 1st class passengers to all passengers.*

a set of partitions of the mixed data set as the default input. For an in-depth review of binning techniques and a detailed description on binning techniques provided in our approach, we refer to publication [BSW*14].

6.5.2. Assessing the Interestingness of a Relation

In the following, we characterize the interestingness of a relation between two bins, i.e., between two subsets of a data collection $\mathbf{RG}_{C+M2}$. We require an interestingness measure satisfying two properties. First, it must be applicable for relations between bins. Second, it has to support the exploration of deviations between *assumptions* and *measurements* in the data. The measurements in the data are the actual occurrences of observations, e.g., the number of women having survived the Titanic disaster. From a statistical perspective, a robust assumption about a bin relation can be based on the populations of the involved bins. We give an illustrative example of two bins aggregated from different attributes. If a bin $b1$ of allocates 60% of the elements of the data set and a bin $b2$ occupies a subset of 40% of the data, a general assumption about the intersection of the to bins would estimate a population of 24% of the data set. A high population of intersections alone may be interesting. Approaches focusing on the absolute count of bin intersections use are, e.g., contingency table visualizations. However, our approach does not focus on the absolute count of observation, but on the deviations between assumptions and the measurements. High deviations between assumptions and the measurements are assessed most interesting. In this connection, our definition of interestingness can be referred to as the notion unexpected discoveries since relations will be marked interesting if they occur more or less often *than expected*. In this case, the user of the ESS has an enhanced tool to explore a mixed data sets for unexpected and possibly unknown relations. However, something assumed about a bin relation may be a subjective matter. Consequently, the design of assumptions about a data set may benefit from an incorporation of domain knowledge $\mathbf{RG}_{C+M7}$. As opposed to robust statistical calculations of something assumed, the assumptions of the users may vary and may further enhance the usefulness of the approach.

In Figure 6.18, we characterize the workflow for assessing the interestingness of bin relations. We use the data set of Titanic passengers to further illustrate the workflow. On the left, an extract of the raw mixed data set is shown, visualized in a tabular format with a general purpose tool. In a preprocessing step, binning approaches aggregate the data to bins with respect to the attributes of the mixed data set. The pairwise bin relations of any two attributes are stored in data structures (contingency tables). Here, the measured (observed) populations of the bin relations can be looked-up for the calculation of interestingness. We apply statistical dependency measures for the assessment of

interestingness. In our publication [BSW*14], we provide the Mutual Information measure and the Chi Square measure χ^2 as two well known and robust variants. However, our approach is not limited to the two measures. Depending on the application domain, other interestingness measures may be required by the user $\mathbf{RG}_{C+M7}$ and be included into the workflow. In our approach, we provide a steering control allowing users to select the most relevant interestingness measure $\mathbf{RG}_{C+M6}$. As a result of this workflow step, we are able to assess the interestingness of the relation between any two bins $\mathbf{RG}_{C+M2}$. The values (scores) of the interestingness measures can be interpreted in a relative way. Thus, users are able to identify most interesting relations without effort. In this connection, the assessment of interestingness enhances the depth-first-search for most interesting relations. We use the scores to a) guide users towards individual, most interesting relations and to b) neglect the tedious search in uninteresting relations. However, we want to point out that many statistical dependency measures are not designed for the comparison of the scores of different bin relations. This is why we refuse the application of the scores as a means of ranking, or to be able to compare of the strength of the dependency of individual bin relations.

6.5.3. Overview of the Relation Space

In the following, we present an overview of the three visual-interactive views of the approach as illustrated in Figure 6.22. At a glance, these views enable users to gain an overview of the data content, metadata, and the relation space $\mathbf{RG}_{C+M4}$. The matrix on the left supports the analysis on the attribute-level, the node-link diagram at the center focuses on the analysis of bins. On the right, we support subspace cluster analysis in a visual-interactive way. A glyph design represents and links bins and attributes in all views. All views are sensitive to an interestingness filter control at the top of the display enabling users to filter bin relations below an interestingness threshold.

A Glyph for Attributes and Bins We present a glyph design for the visual representation of bins and associated attributes. A barchart metaphor shows the distribution of a binned attribute (see Figure 6.19). As a result, we can use the glyph design for the visualization of attributes at a glance. In addition, highlighting single bins within the distribution emphasizes the representation of bins in the context of the corresponding attribute. We conducted a laboratory design study with 14 non-experts for the iterative design of the visual encodings. Within the design process, most of the participants were asked for feedback several times $\mathbf{RG}_{C+M7}$. In earlier phases, we identified that the visual complexity weakened the usefulness of the prototype. In consequence, we kept the visual encodings as simple as possible, with the option of interactive adaption. A barchart is aligned horizontally, accompanied by an attribute and a bin label (see Figure 6.19a). Different coloring concepts (one color per attribute, one color per bin) were withdrawn due to the cause of confusion. In the end, we use the glyph on combination with color for explicitly linking bin relations in different views. Different outlines of the glyph were compared within the study. The final representation keeps the glyph compact and supports cluster color encoding. Concepts for indicating an additional interestingness score per bin are set to optional. According to the multitude of possibly interesting bin relations, a bin can be assigned to multiple colors at the same time. The design choice for assigning multiple colors to a single bin was also part of the user study (see Figure 6.19b). On the left, two insufficient variants are shown. The concept of outlines with different sizes caused unintended ranking indication, while the second variant caused more distractions. Moreover, both variants lack a missing color orientation. Finally, we implemented the third design concept resolving the described drawbacks. The glyph includes the following interaction designs (cf. research goal for interaction designs $\mathbf{RG}_{C+M6}$). First, single bins can be filtered, e.g. if a bin is not interesting for the user. Second, categorical attributes and neighboring bins of numerical attributes can be merged. Third, bins with more than one distinct value can be split. Fourth, bins for categorical attributes can be reordered. Finally, a set of bins (or attributes) can be selected. The selection assigns the set of items as target variables. A highlighting of interesting relations is triggered allowing users to validate or formulate new hypotheses. Splitting and merging interactions triggers the model for a recalculation of affected bins and bin relations. As a result, the user can interactively adjust the binning results and define the set of involved bins, according to the information need $\mathbf{RG}_{C+M7}$.

Gaining an Overview of Attributes and Bins A matrix-based visualization enables users to explore bin relations at two granularities, i.e., the attribute-level and the bin-level $\mathbf{RG}_{C+M3}$. The supported analysis tasks are as follows: First, the user is empowered to gain an overview of attributes and bins of the mixed primary data set. Second, bin relations can be compared with each other. Finally, the matrix visualization allows an information drill-down. To this end, the interactive capability of the glyph design can be applied. The matrix visualization represents all pairwise bin relations in a single view (see Figure 6.20). The quadratic layout supports the visualization of asymmetric relations. Please note that many interestingness measures are asymmetric, in contrast to, e.g., distance metrics. The matrix is structured

(a) *The final glyph design in different levels of detail (LOD). Depending on the screen space and the desired complexity, the glyph can be set to different LOD.*

(b) *Study for the design of multi colorings. Several insufficient designs were withdrawn in the course of the iterative design until a final design solving all drawbacks was implemented.*

Figure 6.19 *Glyph design for the visual representation of single bins within the context of their attributes. The final glyph design is used in all three main views of the system. Color coding enhances the subspace clustering process.*

in major divisions, defined by the attributes. Furthermore, minor divisions represent the bins of each attribute, using the interactive glyph design. The order of attributes is based on the number of interesting bin relations. We assume that a large number of interesting relations also classifies the attribute as interesting, even if the interestingness scores of individual bin relations should not be subject of algorithmic comparison. Depending on the exploratory context and the data characteristics, the user can switch between different presets (means, median and maximum scores). Cells of the matrix represent the interestingness values between two particular bin relations. We use alpha value for the identification of the interestingness score of relations, uninteresting relations hardly differ from the color of the background. On the contrary, interesting relations are highlighted for an enhanced lookup. The matrix visualization allows filtering attributes. As a result, the user can reduce the number of visualized attributes to the most interesting or relevant attributes. The information-drill-down is further enhanced by the interactive bin filter provided with the bin glyph design. In other words, the user can individualize the set of bins of any attribute. Filtered bins and attributes are visualized in a separate view at the bottom of the approach. Consequently, the user is able to lookup filtered items and re-include these on demand. With the matrix-based attribute visualization, we provide a set of relevant interaction designs, e.g., for individualizing and filtering the displayed information. In this way, we demonstrate how the research goal of meaningful interaction designs can be achieved $\mathbf{RG}_{\text{C+M6}}$.

Gaining an Overview of Interesting Bin Relations A node-link visualization of bins and bin relations at the center of the approach (cf. Figure 6.22) enables users to explore multivariate bin relations. A large-scale image of the node-link visualization is presented in Figure 7.6a. In accordance with the guidelines and techniques for layouts presented in Section 5.5.1, we use structural information for the alignment of bins in 2D. The structure for the topology-preserving layout is based on the interestingness scores of pairwise bin relations. The interestingness scores serve as a means of representing pairwise distances, similar to a distance matrix. The MDS projection algorithm uses the pairwise interestingness scores for the calculation of a interestingness-preserving layout in 2D. In addition to MDS, we apply a downstream force-directed layout algorithm to mitigate local overplotting effects. Hence, we use the benefits of both classes of layout algorithms in a single visualization, as suggested in Section 5.5.1. It is also possible to add additional attraction and repulsion forces to the layout design to meet individual user needs, if required $\mathbf{RG}_{\text{C+M7}}$. In our usage scenario, we added repulsion forces between bins and the display border to avoid out-of-boundary effects.

For the visual representation of single bins, we reuse the interactive bin glyph. Interesting relations between bins are indicated with an edge metaphor, the alpha value of an edge assesses the interestingness of the particular relation. As a result of the layout strategy, bins with interesting relations are aligned close to each other. Thus, the topology of the bins in combination with the edge information reveals interesting subspace structures of the multivariate data set. In the Titanic example data set shown in Figure 7.6a, an interesting subspace is, e.g., defined by the bins *GENDER=male*, *CLASSID=3*, *FARE='lowest'*, and *SURVIVED=false* on the lower left of the node-link visualization (yellow). Obviously, the subspace defined by these bins was observed more often than assumed by the model (and possibly expected by the user). This example also shows how indirect relations can be uncovered. While the model of our approach assesses relations between two bins, the proximity of multiple bins and multitude of edges reveals relations of higher dimensionality. As a result, relations between multiple attributes can be observed at a glance, just as demonstrated for the Titanic example data set. The node-link visualization complements the matrix visualization, showing the same data through a different perspective and based on different analysis tasks. The user is supported in gaining an overview of bins and multivariate bin relations. Furthermore, different subspaces spanned by the bin relations can be compared visually. The visualization provides an interactive legend for the individualization of the bin glyphs and the node-link structure, on demand. In this connection, in Figure 6.22, the visual representation of edges (interestingness of bin relations) is activated. Similar to the matrix visualization, the node-link diagram enables the comparison of different entities (here: bins and bin clusters), linking entities interactively (by hovering), and

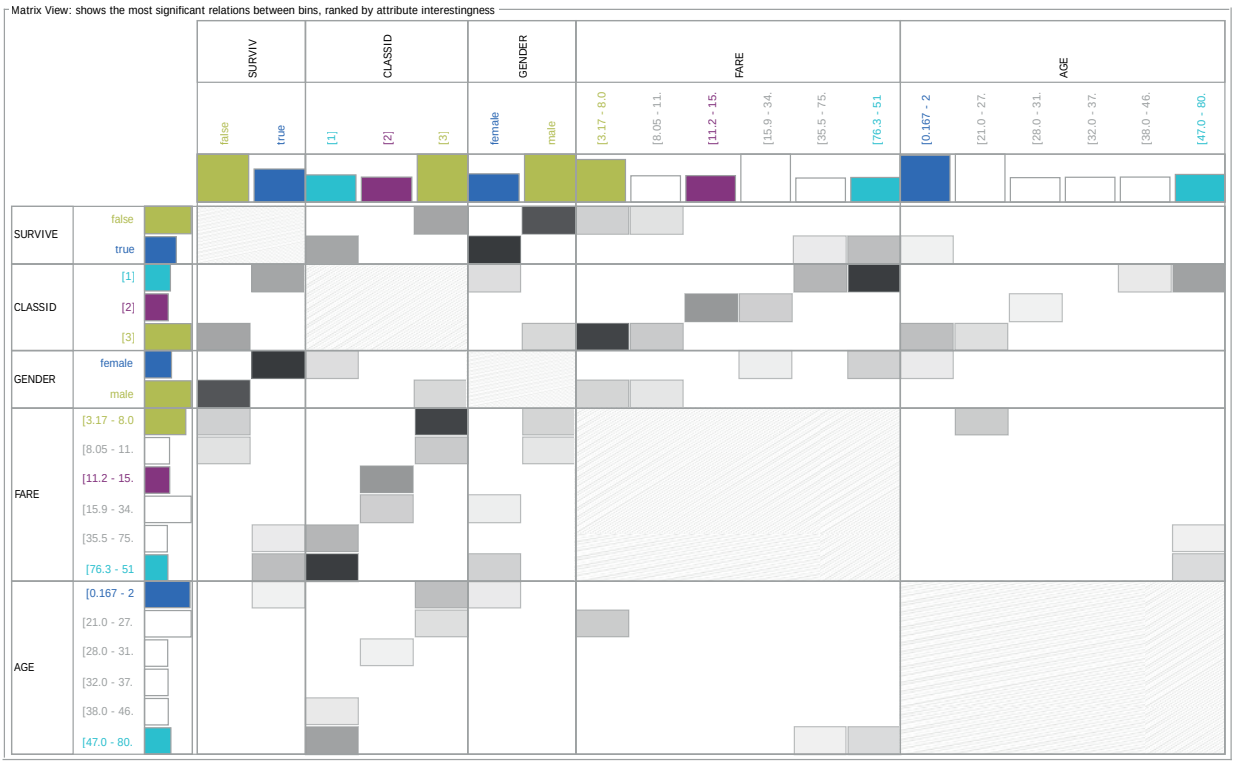
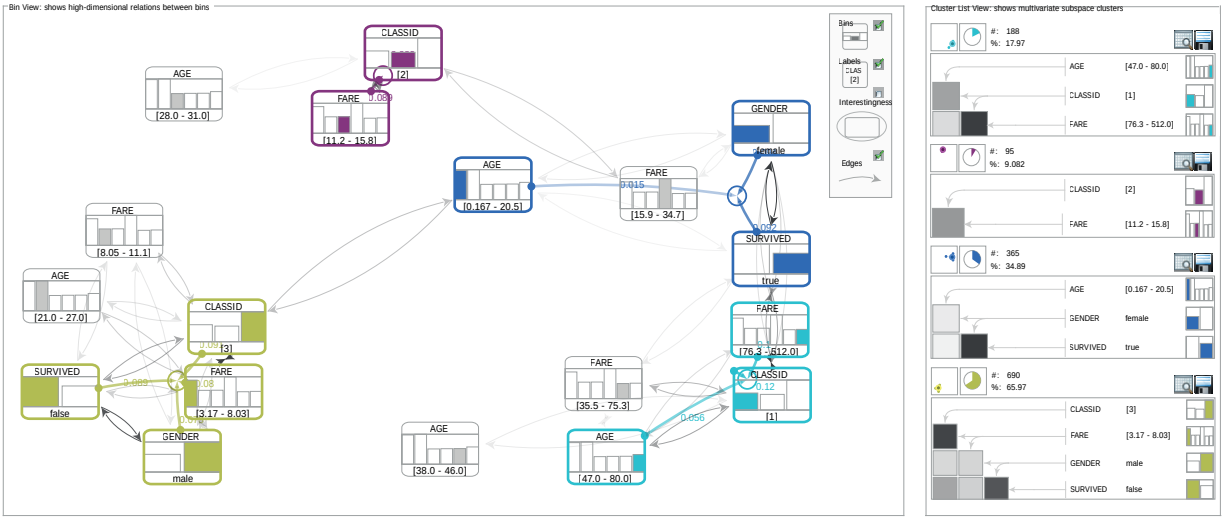


Figure 6.20 *Matrix visualization depicting the interestingness scores of all pairwise bin relations. All bins of all attributes are visualized in a single visualization. Dark colors of cells within the matrix indicate most interesting bin relations. In the Titanic example, most interesting relations exist for women surviving the disaster, old-aged 1st class travelers, or children traveling with the 3rd class.*

the reduction of visual complexity by drill-down interaction (right-click for filtering bins). The node-link view is sensitive to the global interestingness filter. Relations below a particular interestingness threshold are removed from the visualization. Likewise, bins without a single relation above the interestingness threshold are removed. The selection of bins triggers the bin clustering algorithm, which is described in the subsequent section in detail. Finally, the node-link diagram provides zooming and panning interaction allowing users to drill down to local aspects of interest. With the interactive capabilities of the view, we demonstrate how meaningful interaction designs can be provided for interfaces facilitation relation seeking between data content and metadata $\mathbf{RG}_{\text{C+M6}}$.

Exploration of Multivariate Relations Our approach provides another model that supports users in identifying multivariate relations. We provide visual-interactive subspace cluster analysis techniques to guide users towards the most interesting multivariate subspaces. Users can select from three different clustering algorithms; each clustering algorithm focuses on a different information-seeking behavior. We briefly outline the different subspace clustering strategies, for an in-depth description, we refer to our publication [BSW*14]. The *bin clustering* algorithm accepts a set of selected bins. The algorithm provides information about interesting relations of single bins. In this way, bin clustering is valuable for users who test hypotheses against a specific data subspace. For a targeted bin, the multivariate subspace is revealed based on interesting relations. Furthermore, multiple bins can be clustered at the same time. If two of the targeted bins have a direct relation, however, a single subspace structure is calculated. The *attribute clustering* algorithm accepts a set of target attributes selected by the user. The algorithm identifies all subspaces in the data set containing at least one bin of every targeted attribute. Indirect relations between revealed subspaces can be identified by bins with multi-cluster assignment. The *exploratory clustering algorithm* emphasizes the most interesting bin relations of the data set without any target variable. It guides users towards the most interesting multivariate bin relations without supervision. Thus, users are able to gain an overview of the most interesting subspaces. Each of these subspaces may serve as a basis to formulate new hypotheses based on new insights. For this purpose, we use hierarchical, agglomerative clustering of bins with a user-defined aggregation level. Supported merge criteria include single-, median-, average- and complete-linkage (cf. Section 5.2.1).



(a) *Node-link diagram showing bins and bin relations. The layout is chosen to reflect the pairwise interestingness scores. Thus, proximate bins also share interesting relations. In this way, indirect clusters calculated by one of three (multivariate) bin relations can be identified. The blue cluster shows the Birkenhead Drill: "women provided clustering algorithms. and children first!"* **(b)** *List of multivariate subspace*

Figure 6.21 *Visualizations of multivariate bin relations. A node-link diagram and a list-based cluster visualization.*

No matter which of the clustering technique is applied, the clustering result is not limited to bivariate bin relations. Instead, the resulting data subspaces can be based on multivariate bin relations. In this connection, a bin can also be assigned to multiple clusters if the bin belongs to multiple clustered subspaces. Clusters are represented in a list-based cluster visualization, as presented in Figure 7.6b. The visualization of a single cluster consists of four elements describing the multivariate bin relation in detail, a) the label of the attribute, b) the label of the bin, c) the glyph of the bin, and d) a diagonal matrix to represent the internal bin relations. The bin glyphs support the quick localization of bins. In addition, the label information enables user to identify multivariate bin relations. As an example, the yellow cluster in the cluster list visualization reveals the following association. "Male passengers, traveling in the 3rd class, with lowest fares, and passed away in the disaster" - share a relation with each other. For a further analysis of the subset in detail a tabular view can be opened, the respective icon is located at the upper right. Finally, to support the export of findings, subspaces can be stored in a separate file by clicking the disc icon.

6.5.4. Guiding Users Towards Interesting Relations

We present the visualization and interaction designs guiding users towards interesting relations $\mathbf{RG}_{C+M5}$. The interestingness of a relation serves as the functional basis. With the alpha value encoding for cells of the matrix and for the edges of the node-link diagram, we encode the interestingness values of individual relations visually. The user is able to identify the interesting relations without effort. This concept is supported by a filter control helping users to reduce the amount of visualized relations. With the control an interestingness threshold value can be defined, less interesting relations are excluded from the visualization. This interestingness slider control guides users towards the most interesting relations in a possibly large relation space. The subspace clustering techniques guide users towards the most interesting multivariate bin relations. We use the visual variable color explicitly for linking clustering results in the three provided views. The color enhances the identification and localization of cluster structures in every view. In this way, we combine the individual advantages of each view to facilitate reasoning about the clustered subspace structures. The choice of colors needs to discriminate dissimilar clusters, and simultaneously to indicate similar (and possibly intersecting) clusters. To this end, we use a static 2D colormap based on the topology-preserving node-link layout (cf. Section 5.4.2). The 2D colormap describes a diagonal cut of the RGB cube to exploit large parts of the available color space without colliding with the background and foreground colors of the approach. In the usage scenarios presented in the following evaluation section, we show examples of how the technique supports gaining new hypotheses, supported by guiding users towards interesting relations.

6.5.5. Evaluation

We used the Titanic data set to show the visual-interactive encodings of this approach. The Titanic data set is easy to understand and serves as a proof-of-concept example. In the following, we present another illustration of the Titanic data set showing whether the Birkenhead Drill ("women and children first!") really took place on board of the Titanic. However, the Titanic data set is only a lightweight variant of what we call a mixed primary data set with multiple attributes, relevant for the exploration of unexpected knowledge. This is why we briefly echo the two real world usage scenarios presented in our respective publication [BSW*14]. In both usage scenarios, we show the applicability and the usefulness of our approach for data-centered research.

Usage Scenario: the Titanic Data Set In Figure 6.22, the user can gain an overview of the three linked views of the approach. In combination, the linked views present an overview of the attributes, the bins, and the (multivariate) bin relations of the data set. Most interesting bin relation are visually encoded with black cell colors in the matrix and the cluster list, and with black edges in the node-link diagram. At the top the interestingness filter control is shown. The threshold is at a level where the 60% of the relations are filtered (the less interesting proportion, respectively). With the exploratory clustering, we identify four different clusters of the Titanic data set. At the lower left of the node-link diagram, the yellow cluster shows a subspace with male passengers having traveled in the 3rd class with cheap fares, who did not survive the journey. The remaining three clusters show old passengers traveling in the expensive 1st class (cyan), and 2nd class passengers paying for average priced tickets (purple). Finally, we identify that the blue cluster contains the group of survivors, including more women and children than expected. Note that all clusters can also be looked up in the matrix view and in the cluster list view. Each view shows a different perspective on the data. With the cluster list view on the right, the user is directly able to communicate the findings $\mathbf{RG}_{\text{C+M8}}$. The list of explored subsets can be used as provenance information, for subsequent discussions with researchers, or for a publication of new insights. We conclude the usage scenario of the Titanic data set with an in-depth exploration of the childrens' bin. In Figure 6.23, we present the result of a bin clustering. The red subspace is related to the 3rd class (male passengers, cheap fares, diseased) while the blue subspace includes the survivors (1st class, high fares, female). The interesting finding in the visualization is the role of the childrens' bin (*AGE=0.167-20.5*). It is assigned to both opposing subspaces. The children reflect both statements: more children traveled in the 3rd class as expected (ratio higher than in the 1st and 2nd class), and more children survived as expected (ratio higher than the ratio of diseased children). In the childrens' bin the two subspaces have an intersection. The layout factors this circumstance by assigning the childrens' bin at the center of the display between the two subspaces.

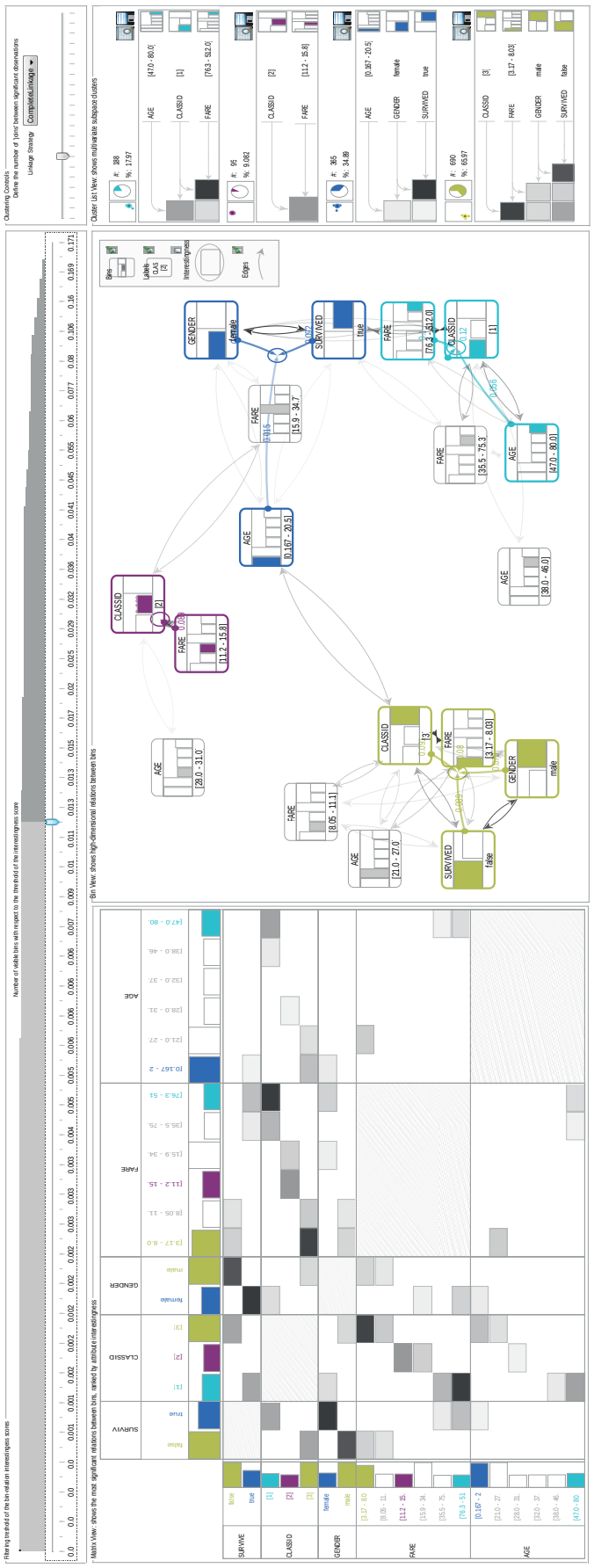


Figure 6.22 *The Titanic data set as a proof of concept. Left: matrix visualization with ranked attributes showing all pairwise bin relations above the interestingness filter (black cells indicate the most interesting bin relations). The survived status of passengers is aligned left, referring this attribute to as the most interesting one in the attribute set. A particularly interesting bin relation is exposed by females having survived the catastrophe. The interestingness filter is shown at the top, more than half of all relations are filtered out. Center: a node-link diagram depicts bin relations in a similarity-preserving (interestingness-preserving) way. Bins sharing interesting relations with each other are aligned in immediate vicinity. Consequently, domain experts are able to gain insight into multivariate bin relations. As an example, the bins at the lower left represent a multivariate subspace, characterized by male passengers, who paid very low fares, traveled in the 3rd class, and unfortunately did not survive. Another proximal bin associates this 'cohort' to the 21-27 age-group. Right: the list-based cluster view provides compact representations of subspace clustering results. The blue cluster relates to the Birkenhead Drill: "women and children first!".*

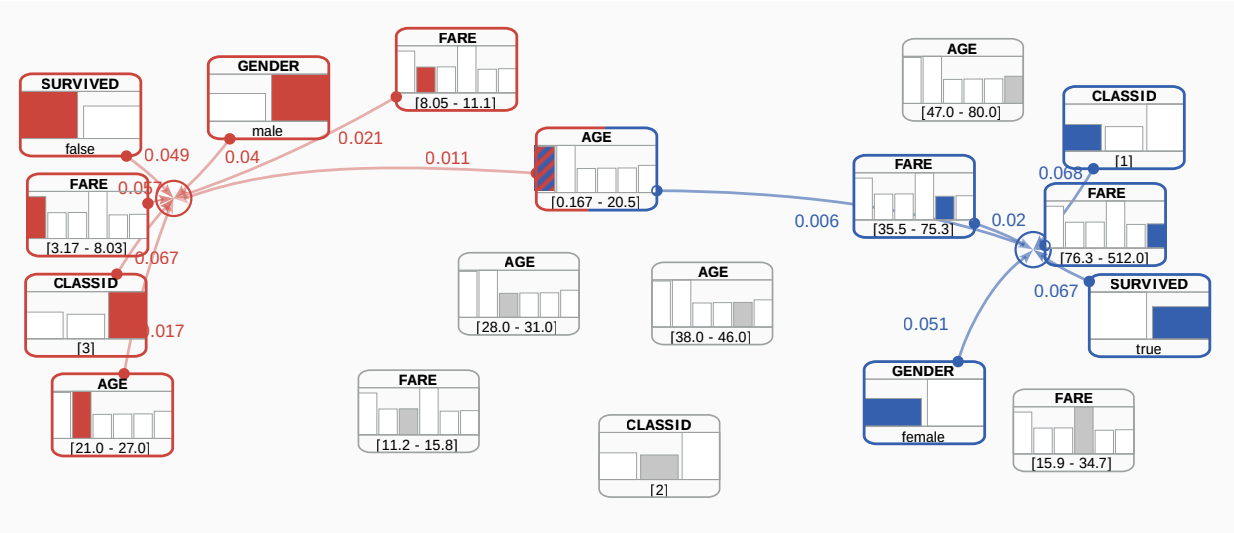


Figure 6.23 *Bin clustering of the Titanic data set. The target bins 'Class=3' (red) and 'Survived=true' (blue) reveal separated clusters, except of the childrens' bin 'AGE [0-20.5]'.*

Case Study: Meteorological Synoptical Observations In this case study, we supported domain experts carrying out meteorological synoptical observations for a climate observation station in Antarctica. The mixed primary data set covered 30 years of measurements taken every three hours (92902 time stamps). The data set contains 26 attributes, most relevant representatives are *Air Temperature*, *Wind Direction*, *Wind Speed*, *Cloud Height*, *Cloud Amount*, *Horizontal Visibility*, *Humidity*, and *Weather Influence*. For the exploration of temporal relations, we integrated the temporal aggregations *Year*, *Month*, *Season*, *Weekday*, and *Hour of Day*. We applied frequency-preserving and domain-preserving binning techniques for numerical attributes as requested by the domain experts.

The domain experts had hypotheses about wind directions, seasonal influences, and the resulting weather situations. We were able to gain the domain experts' trust with the detection of complex but expected weather phenomena. As an example, our exploratory clustering algorithm obtained multivariate subspaces containing popular weather phenomena, such as a bad weather situation. The revealed subspace consists of bins with low horizontal visibility, low cloud height, high wind speeds, obscured sky conditions, synoptic weather influence (wind coming from eastern) and blowing snow drifts. The result can be seen in Figure 7 in our respective publication [BSW*14]. For the discovery of unexpected findings, we took the foggy weather situations into account, which were not yet fully understood in the scientific community. One of our discoveries regards the fog behavior in winter, which is influenced by the continental climate of the inner Antarctica, represented by katabatic winds, (so called radiation fog).

Case Study: Prostate Cancer Research For the second case study, we accessed a mixed primary data set from prostate cancer research. Together with physicians involved in prostate cancer research, we explored one of the largest (anonymized) patient record data sets in the world. The mixed data set is special as it includes attributes of clinical treatment and genomic indicators of patients. With the combination of these two heterogeneous data sources, the domain experts hope to explore undiscovered knowledge to improve patient treatment. Altogether, the data set consists of 13,571 anonymized patient records containing 41 mixed data attributes. We carried out preprocessing steps to treat data quality issues, to convert temporal attributes, and to create meaningful binnings.

Again, we gained the confidence of the domain experts with the detection of expected domain knowledge. The domain experts acknowledged a variety of relations for the attribute *Biological Recurrence* (BCR), i.e., one of the strongest indicators for patient well-being. We were also able to support the physicians in the discovery of unexpected findings. We detected strong relations between genomic indicators, the age of patients, and the surgery. Another discovery was the multivariate relation of the genomic 'deletion' *ERG_IHC* to the deletions *PTEN*, *MAP3K7*, *FOXP1*, *CHD1*, and *WWOX*. For more details of the case study, the terminology, and the visual result representation, we refer to our corresponding publication [BSW*14]). We are happy about the discovery of new findings which will be published in the prostate cancer research community soon.

6.6. Summary

6.6.1. Discussion

Dependency on Upstream Results One point of discussion regards our techniques' dependency on the output of early steps of the reference workflow (cf. Figure 3.1), particularly the preprocessing and similarity definition step for the time-oriented data content, as well as the steps for the design of content-based layouts. In fact, we presented visual-interactive solutions for these design spaces (see Chapters 4 and 5). However, our techniques for seeking relations between data content and metadata form a specific setting where additional requirements exist. In this connection, a visual-interactive system would be beneficial including the entire reference workflow. In this way, effects of upstream models on the relation-seeking capability can be assessed more easily. Users were then able to adapt the access strategy to the data content, based on new insights gained with our relation-seeking techniques.

Formal Definition of a Relation Another discussion issue is the formal definition of a relation. Indeed, we presented different types of relations in the three techniques and faced the challenge of revealing interesting relations on multiple granularities. However, we cannot guarantee that the techniques will allow users to identify all intrinsic relations of a data set in an efficient and effective way. Many data-centered research challenges may be based on specific considerations and requirements. The trade-off between the generalizability and the specificity of relations may be subject to change. A promising solution would be an interactive editor for the specification and adaption of relations and interestingness measures. These specifications may even determine which of our presented concepts are most appropriate for the information-seeking behavior of users. The latter may be solved with a wizard functionality estimating the user need. The model selection for different steps of the reference workflow may be adapted, respectively (cf. Figure 3.1).

Glyph Design We presented different glyph designs for the data content, metadata distributions, and relations. The appropriateness of such visual mappings is relevant to support the user in the identification of insightful relations. We executed multiple iterations for the glyph design, which expected the active engagement of users. Today, the development of visual-interactive glyph design techniques is still an open research question [BKC*13]. Ideally the techniques include relevant design principles, such as perceptual and semantic implications of visual variables and mappings. For our relation-centered glyph designs, it would be interesting to carry out studies assessing possible impacts on the techniques' usefulness. Guidance concepts making the actual process of iterative user-centered glyph design more efficient may be included.

Multiple Content-Based Access Strategies A further discussion issue is directly coupled with a possible extension of the techniques. Currently, we include a single content-based access strategy for the identification of interesting relations. It would be interesting to see different content-based access strategies in a combined approach. Similar to the techniques for cluster correspondence views (cf. Section 5.3.5), relations between multiple clustering results *and* multiple metadata attributes may be revealed. Essential parameters in creating alternative content-based access strategies are time series preprocessing, time series similarity, and the cluster analysis process (cf. Chapters 4 and 5.3).

Scalability The final point of discussion regards scalability issues. From a content-based perspective, both the descriptor approach and the data aggregation step contribute to compact and scalable data abstractions for large data sets. From a metadata perspective, the number of metadata attributes influences the provided models' scalability. The models for revealing interesting relations applied in the three techniques are able to cope with about 100 metadata attributes in real time. If the number of attributes increases significantly, we suggest a persistent concept based on pre-calculation. Many of the automated processes of the models can already be scheduled in the design phase. From a visual-interactive perspective, our techniques are limited by the display resolution. Again, solutions need to be provided if the metadata exceeds 100 attributes. We suggest a pre-filtering concept showing only the subset of most interesting relations. The filtering concept should be interactively steerable allowing users to adapt the focus.

6.6.2. Conclusion

In this chapter, we presented three different techniques for the identification of interesting relations between data content and metadata. With novel visual interfaces, we address three different types of relation seeking. Each visual

Research Goal	Mapping Metadata → Content-based Overviews	Mapping D. Content → Metadata Layouts	Seeking Relations in Multi-Attribute Data
Relating Different Types of Data Ability to relate data content with metadata **RG**$_{C+M1}$	Relation between content-based cluster and distribution of metadata entities	Aggregation of data by metadata entities, relation between entities and associated data content	One aggregation/discretization/binning for each attribute in the data set, relation: bin to bin
Assessing the Interestingness of Relations Ability to calculate an interestingness score **RG**$_{C+M2}$	Cluster: Diversity of metadata / Cluster Result: Homogeneity of metadata distributions in clusters	Between entities: content-based distances + 2D colormap, between entities and content: glyph design	Deviations between assumptions and measurements, Mutual Information and Chi Square χ^2 measure
Multiple Granularity Problem Consider both the bin-level and the attribute level **RG**$_{C+M3}$	Bin-level: cluster-based interestingness, attribute-level: neighborhood-based interestingness	Bin-level: glyph designs for any entity cluster, attribute level: similarity of glyphs and colors	Bin-level: bin relations, matrix view, node-link diagram, cluster list view, attribute-level: matrix view
Overview of the Relation Space Overview of the data content, metadata, and relations in between **RG**$_{C+M4}$	SOM-based content-based overview, Glyph design for metadata distribution, Interestingness visualization	Novel metadata layout, content-based summary/overview, glyph design for the relations	Matrix view: all metadata attributes, bins, bin relations, node-link/cluster list view: bin relations
User Guidance Guide users towards most interesting relations **RG**$_{C+M5}$	List-based view of metadata attributes ranked by interestingness, colormap showing interesting clusters	Distance matrix showing interesting entities, metadata layout showing groups and outliers, glyph design	Baseline: interestingness score of bin relations highlighted with color, bin layout, clustering
Meaningful Interaction Design Ability to zoom, filter, pan, etc. **RG**$_{C+M6}$	Design: steerable model parameters, application: adjustment of measure, metadata, normalization	Selection of metadata attribute, content-based similarity model, zooming, panning	Filtering of attributes and bins, filtering of bin relations, zooming, panning, cluster parameters
Involving the User in the Design Identify and utilize relevant design and steering parameters **RG**$_{C+M7}$	Visual-interactive design prototype, similarity measures, interestingness measures, glyph design	Content-based similarity, glyph design, content-based overview, 2D colormap, projection, layout	Binning, interestingness measure, glyph design, 2D colormap, cluster algorithms
Visual Communication of Relations Ability to communicate relation-seeking results **RG**$_{C+M8}$	List-based visualization of metadata attribute ranking, visualization of the relations including glyph	Metadata layout including glyph designs, content-based overview, entity distance matrix	Complementing views: matrix with attributes and bins, node-link with bin relations, cluster list

Table 6.1 *Summary of the eight central research goals relevant for seeking relations between data content and metadata. Each of the research goals is accomplished in the three different techniques.*

interface is based on the reference workflow (cf. Figure 6.1) and can be integrated in ESS. While not being limited to, we focused on the content of time-oriented primary data. The first approach is based on a content-based overview. Metadata attributes are mapped on top of the similarity-preserving layout. The second approach considers a metadata-based layout, the topology of the layout is based on the similarity of the associated data content. The data content is mapped on top of the metadata layout. In the third approach, we abstract from the concepts of data content and metadata. Instead, we provide a layout of bins of multiple mixed data attributes. The topology of the bin layout is based on a statistical significance measures. For all three concepts, we contribute implementations each in combination with a real-world data set, involved domain experts, and their respective analytical problems. With the new techniques, we resolved research challenge $\mathbf{C}_{C+M}$, which we divided into eight research goals. Table 6.1 demonstrates how the three techniques comply with the research goals.

CHAPTER 7

Case Studies — Exploratory Search Systems

> *Exploratory search makes us all pioneers and adventurers in a new world of information riches awaiting discovery along with new pitfalls and costs.*

according to Gary Marchionini [Mar06], *2006*

We demonstrate the usability and the usefulness of our guidelines and techniques presented in Chapters 4, 5, and 6 in two real-world case studies. Both scenarios include large time-oriented primary data sets, users of different research domains with an affinity to data-driven research, and analysis tasks involving ES. Both case studies make use of the entire reference workflow for the design and application of visual-interactive interfaces for ESS. The presented ESS are the result of two design studies based on the user-centered design principle. In the first case study, we present VisInfo, a DL system supporting domain experts from the Earth observation domain to access large data collections. The second case study refers to the MotionExplorer system, an ESS supporting domain experts in the field of human motion capture data analysis. The structure of the case studies is as follows. First, we present a characterization of the application domains, the involved data, and the tasks. We outline analytical challenges and derive requirements for the ESS. Second, we demonstrate how the techniques described in the latter chapters are applied to the design of the respective ESS. Third, we show the results of the evaluation strategies conducted in the real-world setups. Our primary validation criteria are the usability and the usefulness of the ESS. In addition, we present real-world examples showing the applicability of the ESS. Finally, we conclude the case studies in a summary section.

Contents

Figure 7.1 *The web-based VisInfo ESS allows the selection of different data sources and notions of similarity. The currently selected similarity notion executes primary data source with temperature measurements with a similarity notion on relative daily curve shapes.*

7.1. VisInfo — A Visual-Interactive Digital Library System for Time-Oriented Primary Data

> *“The goal is to have a world in which all science literature is online, all of the science data is online, and they interoperate with each other.”*
>
> according to Jim Gray [HTT09], *2009*

We present VisInfo, a web-based DL system providing visual access to time-oriented primary data. The primary data collection is derived from PANGAEA [PAN], a data repository for geo-referenced scientific Earth observation data. Based on an ES concept, VisInfo at first provides a content-based overview for the exploration of large collections of time-oriented primary data. Furthermore, the system supports the interactive definition of visual queries by example and by sketch. Finally, VisInfo enables users to explore retrieved search results from different perspectives including visualizations of the data content and the attached metadata. The design process of the VisInfo ESS was based on the user-centered design principle including all four steps of the reference workflow (cf. Chapter 3). Experts in data and computer science, experts in usability engineering, a scientific DL, and domain experts from Earth observation were involved in the interdisciplinary and collaborative approach. We discuss comprehensive user studies in the requirement analysis phase based on paper prototyping, user interviews, screen casts, and user questionnaires. Heuristic evaluations and two usability testing rounds were applied during the design phase of the DL system. The results of the evaluations certify measurable improvements in the course of the iterative design, as well as a usable and useful DL system for the visual access to time-oriented primary data. The VisInfo system is mainly based on [BBF*10, BBF*11, BRG*12, BDF*15].

7.1.1. Characterization of the Domain, Data, and User Tasks

Characterization of the Domain — Earth Observation The Earth observation domain gathers information about the Earth system. One of the goals of the Earth observation domain is the exploration and characterization of Earth phenomena. Prominent examples are the El Niño ocean-atmosphere phenomenon, the Gulf Stream warm Atlantic Ocean current, or the role of the oceans in the Earth system in general. The characterization of Earth phenomena includes the specification of these phenomena with measurement parameters. A similar scope regards the identification of dependencies between different phenomena and the involved sets of parameters, respectively. Examples are seasonal effects showing a dependency on the solar radiation, the air pressure with a functional dependency to the height in the atmosphere, or the dependency of photosynthesis on the ratio of carbon dioxide in the air. Moreover, Earth observation focuses on the identification of changes within Earth system, such as the greenhouse effect, glacial melting, and other indicators of climate change. The impact of human behavior on the Earth' environment is only one reason why Earth observation has become more and more important. Earth observation research is carried out in many different fields focusing on physical, biological, or chemical aspects. The applied sensor technology assesses phenomena in the air, in the water, in the sediment, in the ice, or in the atmosphere. The variety of application fields combined with the multitude of collected parameters considers Earth observation a prominent representative for the creation of primary data at large scale. A variety of data repositories and data warehouses provide primary data from the Earth observation domain, partially according to the open data principle, Section 2.2.1 illustrates. Today, modern data-centered technology applied in Earth observation covers large parts of the data life-cycle (cf. Section 2.2.2). In particular, steps like the creation of primary data in various formats, the processing of primary data, and the storage of primary data in data repositories are common practices. An increasing amount of primary data is provided with DOIs [Bra04] helping to improve the visibility and dissemination of documents. In the same way, data-driven research is common practice in various fields of Earth observation. Scientific workflows are increasingly adopted across many natural science and engineering disciplines (cf. Section 2.4.1). This development also led to an increased demand for enhanced analytical capability. Convincing examples of IV and VA, applied to Earth observation, are presented by Steinbach et al. [STK*03], Nocke et al. [NSBW08], Kehrer et al. [KLM*08], and Tominski et al. [TDN11].

The domain experts from Earth observation, involved in the VisInfo case study, work for the Alfred-Wegener Institute (AWI) for Polar and Marine Research in Bremerhaven. A core project of AWI most relevant for VisInfo is the Baseline Surface Radiation Network (BSRN). The network creates, processes, and analyzes over 100 parameters (time-oriented attributes) relevant for Earth observation research considering various aspects of radiation. The domain experts consider the Shortwave-Downward radiation (SWD) as one of the most relevant parameters in the radiation context. SWD is a special type of radiation-based energy emitted from the sun. SWD affects many processes investigated by Earth observation research. Together with other parameters SWD is considered in climate analysis in many different ways. As an example, in combination with cloud data SWD enhances the analysis of cloud effects. In a similar way, SWD is relevant for research on the ozone layer hole. SWD is also applied for building models in agriculture, hydrology, and solarthermics. The BSRN supports these types of research fields based on the measured data. In particular, BSRN network aims at providing both information about the scientists who measured radiation data and information about the scientists who will reuse the radiation data. This concerns the BSRN network, a most relevant user group for the VisInfo DL system.

Characterization of the Domain — Digital Libraries Of those engaged in VisInfo, the second expert group works in the field of DLs. DLs support the collection, the storage, and the retrieval of digital documents. Digital librarians are considered a key user group of information providers. DLs can contribute to the data-driven research paradigm [HTT09] substantially. For surveys on DLs, reference models, and DL systems, we refer to [BYRN*99, Bea07, CCF*08, HTT09], and [FGS12]. DLs play an important role in the data life-cycle (cf. Section 2.2.2). In particular, DLs can support the data preservation, the data access, and the data reuse. Thus, DLs predominantly support the later steps of the data life-cycle. DLs and Earth observation are complimentary elements in the area of data-driven research. On the one hand, primary data benefits from DL support in a way that the data becomes publicly available for access and reuse. On the other hand, for DLs primary data is a valuable source for providing a new area of library service for new user groups. A widespread approach in DLs is the utilization of metadata attached to documents to make document collections searchable. Faceted search support and full-text search on the explanatory metadata are basic skills supported by DLs. The Greenstone software system for the design of DLs including full-text searching and metadata-based browsing may serve as an example [WBBM00]. Beyond the utilization of attached metadata, DLs can greatly benefit from ES capability incorporating the data content itself [WR09]. As a result, the user will then be able to gain an overview of the data content [HK12] and to formulate content-based (visual) queries [vLSFK12]. We have already emphasized the

value of the content-based access to time-oriented primary data in Sections 2.2.1 and 2.3. The data content has to be processed, analyzed, and used for search and exploration support to expose the hidden information and undiscovered knowledge. Not only from an IR [BYRN*99], but also from a visual search perspective [HK12], the incorporation of data content can facilitate the information-seeking process tremendously. This is why one of the most relevant challenges of modern DLs is content-based access to the underlying document collections. In fact, DLs made good progress in the content-based access to textual data collections in the past (cf. Section 2.2.1). A number of approaches combine DL support with ES capability on textual data content [HK12]. In addition, DL systems with content-based access solutions for different types of multimedia content already exist [MPR02, SN11]. For other types of data content, however, the DL community has seen comparably few content-based access approaches to this day. A visual search approach for architectural model data [BBC*10] and a DL system providing a series of data types including 3D data [ABB*07] may serve as positive exceptions. In particular, visual search and data exploration capability known from IV and VA has hardly been put into practice in the field of DLs thus far.

Challenges of the Domains For both the Earth observation and the DLs domain, various challenges exist that need to be overcome before content-based access strategies for time-oriented primary data can be put into practice. Most of these challenges can be described in terms of (a) the time-oriented primary data per se, (b) the data life-cycle, and (c) scientific workflows. In addition, many of these challenges apply to both of Earth observation and DL. In our characterization, we also echo the challenges presented in Section 2.2.

The complexity of time-oriented primary data is a predominant challenge not only for DLs [BWE06] and data-centered research fields, like Earth observation [AKD10, CIZW13, KH13], but also for VA research [KMS*08, TC06] (cf. Section 2.2.1). The complexity can be based on the size of the collection, on the heterogeneity of the data, and particularly important for this thesis: on the time-varying behavior. Moreover, time-oriented primary data requires data cleansing [KHP*11, GGAM12] before the value of the data can be exploited. The data life-cycle of the involved domain experts poses additional types of challenges (cf. Section 2.2.2). An example is the identification and specification of relevant data subsets for a subsequent analysis [BWE06, MH10]. Approaches based on fact retrieval [Mar95, p. 29 ff.] and known-item search [WR09, p. 14] may not be sufficient when data is searched for exploratory research efforts. This challenge worsens if the underlying data collection (the search space) is large, unknown, or insufficiently understood [BWE06, KMS*08]. In each instance, additional exploration-first approaches [KMS*08] are to be recommended. Furthermore, different stakeholders must to be involved in the data life-cycle to avoid gaps in data standards [BWE06, FGS12], data abstractions [Fek13], or visual interfaces [RBK13]. Finally, the construction of scientific workflows poses a number of challenges (cf. Section 2.4.1). Most relevant for the construction of scientific workflows is the choice of appropriate algorithmic models in a carefully considered order [FPS96, KAF*08]. Moreover, many models provide parameters which need to be defined in a meaningful way [FPS96]. For both challenges, it is relevant that data scientists incorporate the knowledge of different expert groups, such as DLs and Earth observation [BWE06, vW06, AKM*09, TDN11, SMM12, Fek13, MA14]. Ideally, a scientific workflow can be created in such a way that it can be executed fully automatically. In this case, a scientific workflow covers the strength of both humans in the creation and machines in the execution. Still other important challenges in the construction of scientific workflows regard (a) visual feedback strategies for intermediate results [TC06, KMS*08, FH11, TLLH13, Fek13], (b) the utilization of advanced visual-interactive workflow steps in general [NSBW08, TDN11], and (c) understanding and visualizing uncertainties [KKEM10]. Finally, the incorporation of explanatory metadata [Bra04] in the analytical process is a challenge on its own.

Characterization of the Data Our time-oriented primary data set is a subset from the scientific data repository PANGAEA [PAN]. PANGAEA is operated by the Alfred-Wegener Institute (AWI) for Polar and Marine Research in Bremerhaven and the MARUM-Center for Marine Environmental Sciences in Bremen. PANGAEA archives, publishes, and distributes geo-referenced scientific Earth observation data. Data in PANGAEA comprises observations from four main research areas, i.e., water (e.g., temperature, salinity, oxygen), sediment (e.g., total organic carbon TOC), ice (e.g., chemical composition, dust concentration), and atmosphere (e.g., temperature, humidity). PANGAEA supports the export of data for access and reuse in a variety of plain text document types. A metadata-based search interface allows defining filters, and thus specifying the targeted data subsets. The data export covers raw time-oriented primary data content, e.g., in ASCII table format. A primary data document consists of a large data table, consisting of measurements taken at discrete points in time. The temporal domain defines the order of the measurements in the documents. The documents also contain a header with explanatory metadata information about citations, originating project names, spatial and temporal conditions, and parameter descriptions. The metadata corresponds to the Data Cite kernel [Bra04], a metadata standard especially targeted towards scientific primary data documents.

The PANGEA subset of our case study is a collection of time-oriented primary data documents [BKLS12]. For dissemination and reproducibility reasons, our data subset is certified with an own DOI and indexed in the PANGAEA catalog. The primary data documents of the collection cover measurements taken from the years 1992 to 2012 at 55 stations worldwide. For every month and every BSRN station a single document is stored at the PANGEA repository leading to 6,813 documents in total. The predominant temporal resolution of the time-oriented data content is one minute. The high temporal resolution, the duration of Earth observation measurements over 21 years, and the variety of stations all over the world yields a data set of approximately 500,000,000 individual records. In addition, the data tables in the documents have up to 100 columns, each representing an individual measurement parameter. At the request of the involved domain experts, we learned that the most relevant parameters for the VisInfo DL system are the temperature progression at 2m height, and the Shortwave-Downward radiation (SWD). We chose the BSRN data set because it (a) is relevant to a large research community, (b) provides a substantial amount of data content, and (c) features a rich set of explanatory metadata. The presence of attached metadata requires relation-seeking tasks combining the time-oriented data content and metadata (cf. Chapter 6).

Characterization of the Tasks We recall the goals of the Earth observation and the DL domain to derive analysis tasks for the targeted ES system. In addition, we echo the identified challenges of the involved domains. On this basis, we outline the desirable functionality of VisInfo as a result of the requirement analysis phase. For a more in-depth description of the task characterization phase of this design study, we refer to our respective publications [BBF*10, BBF*11, BDF*15].

Domain experts in Earth observation explore and characterize phenomena of the Earth system. Part of the work can be described as the identification of dependencies between different parameters, or parameter spaces. The special characteristics of time-oriented data also enable the domain experts to observe time-dependent parameters, and thus to identify changes over time. These research efforts are to a high degree of an exploratory nature. The collections of vast amounts of time-oriented primary data enable the domain experts to formulate and validate scientific hypotheses. The data-driven research paradigm shifts the research towards exploratory data analysis known from IV and VA. Taking the time series measurements of the targeted SWD parameter into account, we identify several analysis tasks.

First, if the data collection is previously unknown, the domain experts need tools for the *exploration* and *identification* of interesting patterns. An interesting pattern can be an anomaly in the data (such as an outlier in the SWD progression) or something occurring frequently (a cluster). The functionality that allows the identification of interesting patterns supports the domain experts in the formulation of new hypotheses.

Moreover, the existence of large primary data collections requires domain experts to carry out *localization* tasks of already known patterns classified as interesting earlier on. The localization of known patterns is most likely a search task as known from search systems, such as DLs. These known patterns may be due to both metadata attributes or the time-oriented primary data content. The contextual analysis of localized patterns allows the domain expert to validate hypotheses, or to formulate subsequent hypotheses. An interesting temporal pattern may be localized periodically (e.g., seasonal) when the SWD pattern is dependent on the temporal domain. Popular examples are high values of SWD measurements in the summer months (when the sun is closer to its zenith). Furthermore, an interesting pattern may be identified rather coincidentally when the pattern is not dependent on the temporal domain. An example for such coincidental occurrences of SWD patterns is a so-called clear-sky condition, when the sky is free of any clouds. In this cases, the observed parameter may rather be dependent on other parameters (e.g., weather conditions) or to attributes in the attached metadata.

In this connection, *relation seeking* and *comparison* tasks gain priority. The process of relation seeking, especially on ill-defined information need, is to a large extent exploratory. Interesting relations in different parameters, or between parameters and the attached metadata attributes can be revealed with the respective exploratory analysis capability. Yet another class of research questions can be answered with analytical capability enabling comparison tasks. We have already outlined the example of high values of SWD measurements. However, high values can only be identified as such in comparison to lower values. In the SWD example, low values predominantly occur in winter when the influence of the sun is less important. The involved domain experts want to compare patterns in the value domain of SWD measurements and patterns based on the temporal domain. Comparing the value domain of patterns is often carried out in cluster analysis approaches, provided that the overview of the clustering result is comprehensible for the domain experts. The comparison of temporal patterns is frequently applied by Earth observation scientists, e.g., when changes over time are to be assessed.

7.1.2. Requirements

In the domain characterization phase, we echoed the challenges about the primary data, the data life-cycle, and scientific workflows. Both the Earth observation scientists and the digital librarians are confronted with these problems. These shortcomings precisely motivate the VisInfo approach. For the domain experts from Earth observation it is crucial to apply the concepts of the data life-cycle and scientific workflows. The product of the scientific workflow should be a content-based access and data abstraction strategy that sticks to the domain knowledge of the domain experts and to the targeted analysis tasks. As an example, the similarity concept for the time-oriented data content should be defined in accordance with the domain experts' notion of similarity. This is why the design of the workflow calls for tight coupling with the domain experts which in turn raises the trust of the domain experts in the approach. The domain experts consider visual feedback from intermediate results as a most beneficial functionality for the design of the scientific workflow. As a result of the workflow construction, the search system will provide a meaningful retrieval system for the underlying data content. In addition, the workflow design will create a useful content-based overview solution for the vast amounts of data.

The digital librarians engaged in VisInfo aim at providing a DL system for the content-based access to time-oriented Earth observation data. The time-oriented primary data collection of the data set should be made applicable for second-party data access and reuse. The DL system should support data-driven research. The main focus of the digital librarians is on the ease of use of the system. Any additional barrier could potentially cause the user to turn away from the system before even fully assessing the possibilities of the paradigm shift towards data-driven research. While the Earth observation scientists are the targeted expert user group, the DL system should also be useful for non-experts. This is another requirement for why the visualization and interaction design of VisInfo should be intuitive and usable. The data collection should consist of measurements of the SWD parameter, which the domain experts consider highly relevant. Moreover, it would be beneficial if temperature measurements would also be a part of the search space. In consequence, non-experts using the DL system will have a plausible parameter for intuitive access to the data content.

The VisInfo approach has to comply with the scientific workflows executed by the targeted domain experts. To this end, existing workflows need to be identified, reflected, and adapted to the functionality of the DL system. In return the DL system has to provide new visual-interactive means to interact with the data content and the associated metadata. The retrieval process should be extended to visual-interactive querying functionality, based on the content-based data access strategy. As a result, domain experts will then be supported in the translation of their similarity notion to a functional query. In this connection, the formulation of a *Query-by-Sketch* would be a highly beneficial interaction concept. In return, metadata can be used for interactively filtering the collection to meaningful subsets. In addition, the content-based access strategy should provide the means to gain a *visual overview* of the large data content. A content-based overview, in VisInfo referred to as a *visual catalog*, will support the domain experts in carrying out exploratory data analysis tasks, such as the identification and comparison of patterns. In this way, the individual characteristics of the search space can be revealed. If an identified pattern is considered interesting, a *Query-by-Example* concept would be greatly beneficial. Being able to execute queries be-example would allow domain experts to localize interesting patterns and associated data aspects in detail.

The *visualization of search results* should show different perspectives of the data. A visualization of the time-oriented data content, facets of most relevant metadata, and other specific views will allow domain experts to reveal interesting relations in the retrieved data subsets. The relevance of the individual metadata attributes will be assessed in two ways. First, it depends on insights to be gained by a relation-seeking task carried out within the design phase. Second, the experts from Earth observation and from DLs will then be involved in the identification of relevant metadata attributes. As a consequence of the search-result exploration, the domain experts can develop a clearer understanding of the intrinsic properties of the result space and the superordinate search space, respectively. If a domain expert is confident with the retrieved data subset the DL system should provide external links to the original data warehouse where the raw primary data is preserved. In case a domain expert wants to adapt a retrieved subset the DL system should support the iterative refinement of formulated queries.

7.1.3. Time Series Preprocessing and Similarity Definition

As a first main step in the reference workflow for the creation of ESS (cf. Figure 3.1), we require a solution for time series preprocessing and for the definition of time series similarity. In consequence, we will then be able to design downstream steps, such as the retrieval component and the data aggregation step of the reference workflow. In Chapter 4, we explicitly elaborated solutions for the visual-interactive preprocessing of time-oriented data and for the visual-interactive definition of time series similarity. Taking the VisInfo case study into consideration, we

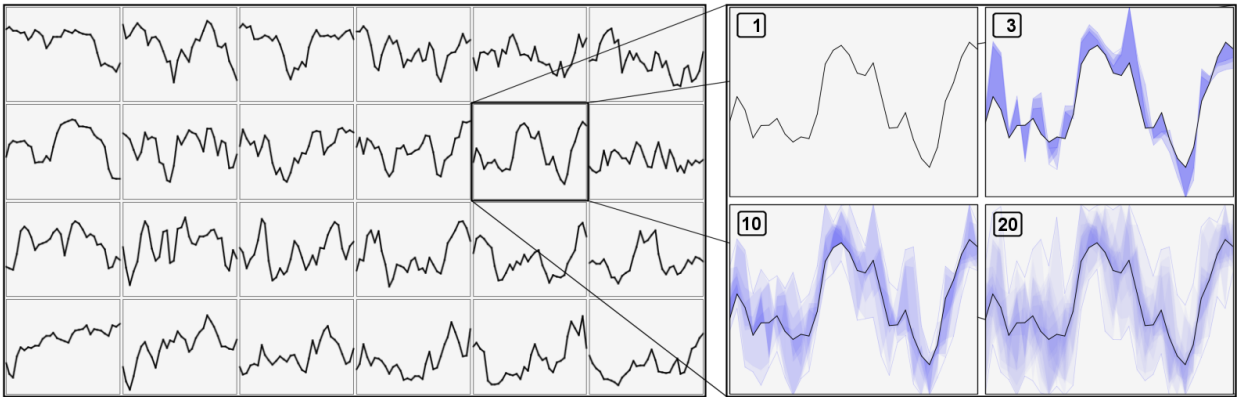


Figure 7.2 *Visual design prototype of the 'Visual Catalog', the content-based overview of the VisInfo ESS.*

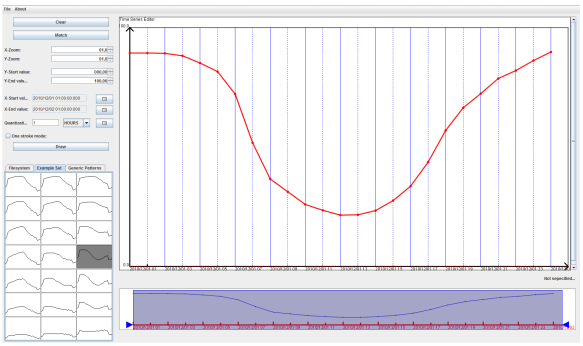
used the visual-interactive system of Chapter 4 and carried out a collaborative design study together with the domain experts from Earth observation. As a result, we created four preprocessing workflows each generating a different feature set. Two workflows are based on Shortwave Downward radiation (SWD) and two workflows factor temperature measurements. In both cases, we provide an absolute and a relative workflow scenario. The PAA [KCPM01] descriptor transforms the time-oriented data into the feature space, the Euclidean distance defines the similarity of time series features. We provide an in-depth description of the workflow construction phase in Section 4.4, the usage scenario of Chapter 4. Additional information about the construction of the preprocessing and the similarity definition step can be seen in our corresponding publications [BRG*12, BBF*11]. In VisInfo, explicit user parameters in the time series preprocessing phase are the choice of the targeted data set (SWD, temperature), as well as the notion of similarity (absolute, relative). These parameters help to access the data content in a meaningful way and to define appropriate functions for time series similarity. Furthermore, the visual-interactive VA system for time series preprocessing allows involving users and creating workflows in a user-centered way.

7.1.4. Content-Based Overviews of Time-Oriented Data

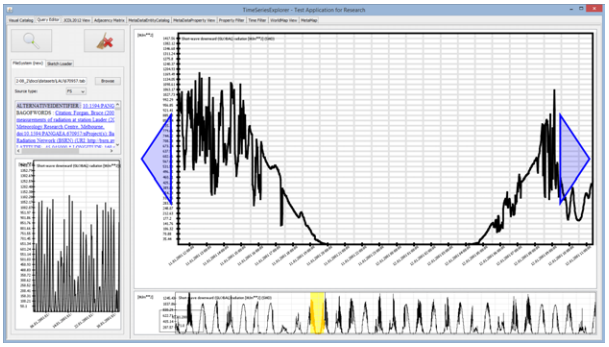
With the time series FV and the definition of similarity, we approached data-specific challenges for both time series and primary data. We next discuss the designs for the three remaining steps of the reference workflow, the data aggregation, the visual mapping, and the content-based layout (cf. Figure 3.1). To a large extent, the design process is reflected in Chapter 5, in which we present guidelines and techniques for the design of content-based overviews. Many of the presented images refer to the VisInfo case study.

Data Aggregation Together with the domain experts, we created the concept of a 'Visual Catalog'. The Visual Catalog will enable the users to gain an overview of the time-oriented data content, most likely to our techniques for the design of content-based overviews presented in Chapter 5. Based on the concept, we developed visual prototypes for all three steps of the reference workflow. A prototype of the design of the content-based overview is presented in Figure 7.2. An early visual-interactive system including a prototype of the Visual Catalog was presented in a concept paper of the VisInfo approach [BBF*10]. In several iterations, we discussed different revisions of the visual prototype together with the digital librarians and the Earth observation scientists. A refined version of the visual prototype was subsequently published in another publication [BBF*11]. At that time, we also fixed the SOM algorithm as our targeted clustering model. In subsequent iterations, we elaborated different parameter values of the SOM algorithm as described in our section about quality-driven visual cluster analysis (cf. Section 5.3). Most relevant from a user-centered design perspective was the calibration of the grid resolution of the SOM layout. In the end, the resolution of the final VisInfo was set to 8×10 cells (see, e.g., Figure 5.18).

Visual Mapping Simultaneously to the design of the clustering model and the definition of most relevant parameter values, we carried out different iterations of the glyph design for the visual representation of single data aggregates (cf. Section 5.4). In Table 5.1, we show the cluster glyph of the concept paper [BBF*11] and the final glyph design applied to the VisInfo system [BDF*15]. The final glyph design uses a linechart for the encoding of the temporal and the value



(a) *Sketch editor including a variety of parameters for both the temporal and the value domain. Example patterns from an example set can be used by drag-and-drop. The interface provides zooming, a global time axis at the bottom always represents the global shape of the sketch.*



(b) *Sketch editor including a file input functionality. Loaded files are presented in a list and can be dragged into the sketch interface. A global overview of the loaded monthly time series at the bottom is linked with the current curve pattern.*

Figure 7.3 *Design prototypes of the visual Query-by-Sketch interface of VisInfo. The user is able to sketch (modify) time series patterns. With the execution of the search process the sketch is transformed into the feature space.*

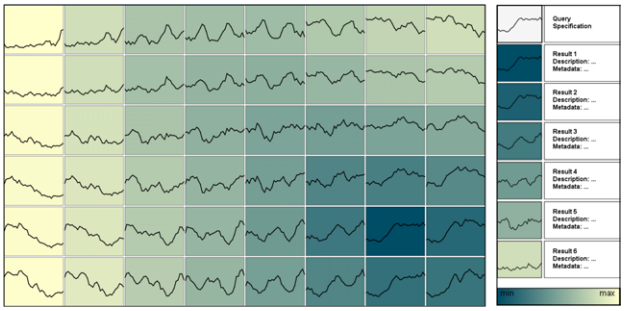
domain of the data. The size of a cluster is represented with a label. A subset of the most similar data elements can be interactively added by users to assess the variation of the cluster.

Layout, Levels of Abstraction The grid-based layout of the content-based overview is predefined due to the utilization of the SOM clustering algorithm. With this cluster structure-based layout solution, no additional projection or force-directed layout technique needed to be included into the workflow (cf. Section 5.5.1). In the final VisInfo system, we provide a Details-on-Demand concept for single cells of the SOM catalog. If a user clicks a single cell, an enlarged visual representation of the respective cluster is shown. In contrast to the visualization of the clusters in the SOM grid, the detail visualization also shows a bundle of most similar data elements (with respect to the cluster prototype), the number of elements is interactively steerable by the user. The Details-on-Demand visualization is presented in Figure 5.27 showing two different clusters. It can be seen that the shapes of the cluster centroids and the most similar data elements are considerably different.

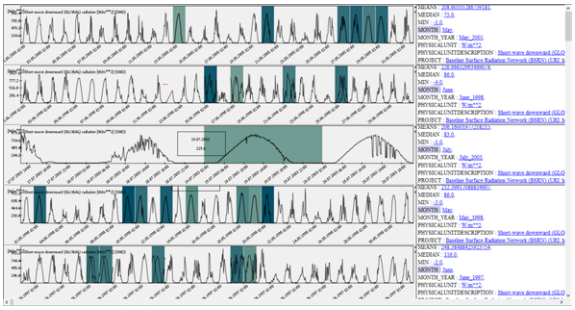
With the content-based overview, the VisInfo ESS provides a visual representation of the Visual Catalog. The essential research challenge was choosing appropriate models in a meaningful order with suitable parameters. In addition, we approached challenges associated with exploring time-oriented primary data. For this purpose, we faced the guidelines and techniques presented in Chapter 5 and emphasized the involvement of users in the design.

7.1.5. Visual Interactive Querying and Search Result Exploration

A most relevant interaction design for the targeted ES capability is visual querying. We provide a visual-interactive editor allowing users of the DL system to query time series patterns by sketch. An early prototype of the sketch editor is shown in Figure 7.3. In Section 5.5.3, we described in detail two most relevant visual querying techniques *Query-by-Example* and *Query-by-Sketch*. In the VisInfo ESS, we include both querying techniques. Based on the content-based overview, the user is able to focus on local aspects of interest with the Details-on-Demand visualization of single clusters. In addition, cluster patterns can be used for querying by example. The Query-by-Example concept in Section 5.5.3 is illustrated with the visual-interactive designs of the VisInfo case study. We refer to the respective section for an in-depth description of the Query-by-Example solution of VisInfo. The second visual querying concept provided in VisInfo is Query-by-Sketch. The visual-interactive sketch editor of the VisInfo system is presented in Figure 5.28. Users can sketch a curve progression which is subsequently transformed into a content-based query for the retrieval algorithm. As a consequence, the sketch editor supports two different information-seeking behaviors. First, it serves as a visual interface for users with a clear notion of the search pattern. In this case, a new query can be created literally 'from scratch'. Second, the editor supports the modification of previously selected example patterns and queries. As a result, the patterns of the content-based overview can serve as the basis for modifications by the users, carried out in the sketch editor. The sketch interface also serves as a means of refinement of previously executed



(a) *Mockup design for the result visualization based on an example query selected in the content-based overview. Color is used to assess the distance between retrieved documents and the example pattern. The colormap is shown on the lower right.*



(b) *Early design of the search-result view. A list of time-oriented primary data documents can be explored in detail (Details-on-Demand is active for the 3rd data content). The metadata visualization supports seeking relations.*

Figure 7.4 *Design prototypes of the visual search-result exploration. A list of retrieved time series patterns is visualized in the raw time series format of the primary data document. In addition, attached metadata is visualized.*

queries. In this connection, VisInfo addresses an iterative information-seeking process facing the challenge associated with searching in the data content of time-oriented primary data. For more details of the Query-by-Sketch technique, we refer to Section 5.5.3.

In VisInfo, the result of the content-based retrieval constitutes a second exploration space. The user is able to explore the retrieved subset of the search space in detail. Early designs of the search-result representation are shown in Figure 7.4. An illustration of the final result visualization is presented in Figure 7.5. VisInfo provides different views for the analysis of retrieved data content and for the identification of relations of the content to the associated metadata. The retrieved documents are shown in a list-based visualization providing information about the time-oriented data and metadata. A list of metadata attributes for every retrieved document is shown on the right. The set of provided metadata attributes was suggested by the digital librarians. The visual representation of the time-oriented data content eases the way for a detailed analysis. The content of the monthly measurements is shown as a whole, the retrieved daily pattern within the document is highlighted. In the course of the study, we carried out different interaction designs for the visualization of the raw time-oriented primary data. The final linechart visualization presented in Figure 7.5 supports zooming and panning interactions enabling users to identify local aspects of interest. For the visual representation of the raw time-oriented data in the result view, we carried out an individual preprocessing workflow. Most challenging was the pure amount of raw data for every time series document which has to be transferred from the VisInfo server to the web client. The overall goal of the workflow was a compact and yet precise representation of the raw time-oriented primary data content. The applied time series descriptor preserves the perceptual important time-value pairs of the raw time series. For more information about the additional workflow including the Perceptually Important Points (PIP) descriptor, we refer to our corresponding publication [BRG*12].

We further provide two views (perspectives), assembling the retrieved data by their geographical reference and to the temporal occurrence. Both the geographical view and the time view enable revealing relations between the daily time series patterns and the attached metadata attributes. Furthermore to the three different views of the search-result presentation, we provide a set of relevant metadata facets. Since a content-based retrieval in hundred of thousands of search patterns typically retrieves a variety of relevant items [WR09], metadata facets are most reasonable to reduce the resulting space in a meaningful way. Together with the digital librarians and the Earth observation experts, we made the decision in favor of ‘Location’, ‘Year’, ‘Month’, ‘Season’ and ‘Surface Type’. These metadata attributes are shown on the left of the retrieval result visualization in a faceted search view (see again Figure 7.5). After the search-result exploration and a potential information gain, the user can redefine the query and start another search. Concluding the search process, the user can follow external links to the data warehouses to download the retrieved data sets.

7.1.6. Identification of Relations Between the Time-Oriented Data Content and Metadata

VisInfo also benefits from the concepts and techniques for the identification of relations between the time-oriented data content and attached metadata in Chapter 6. The three different concepts are (1) mapping metadata on top of

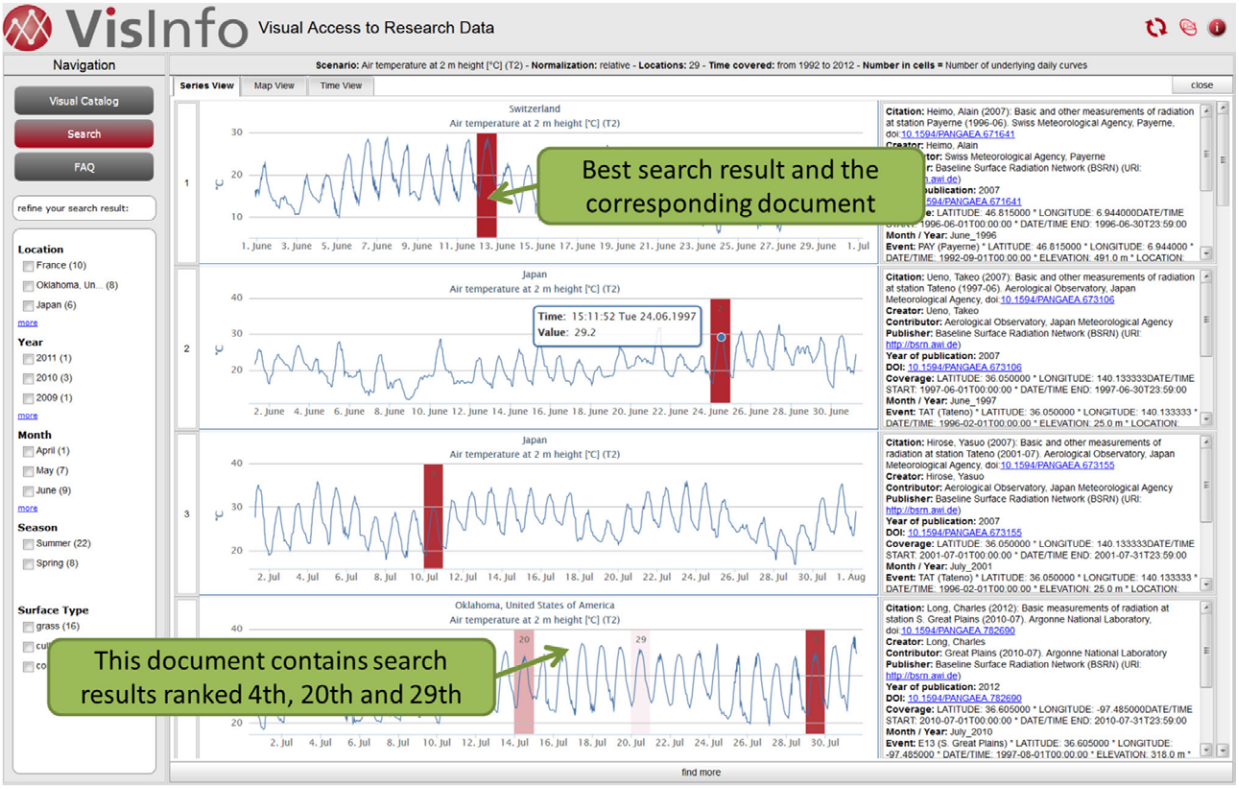
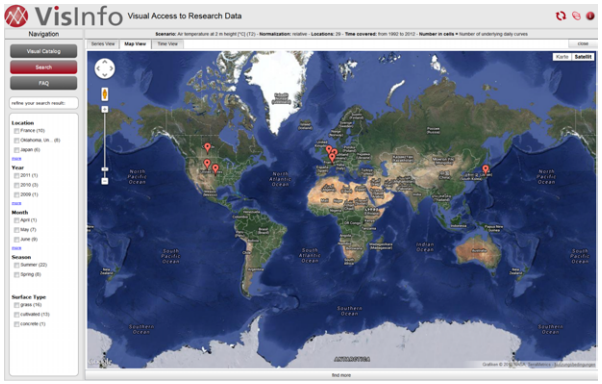
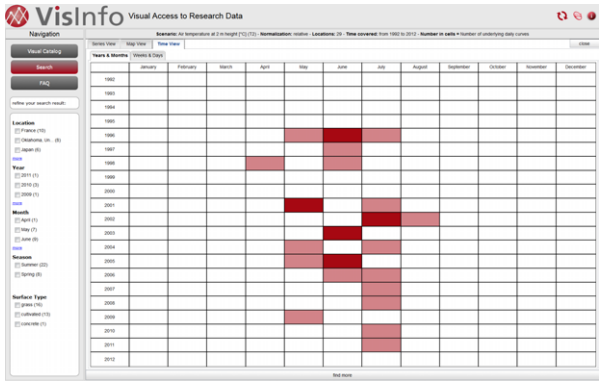


Figure 7.5 *The visual representation of the retrieved data content in VisInfo. The best matching daily search patterns are highlighted red. Metadata facets (left) enable the information drill-down to the result space. Details of multiple metadata attributes are provided on the right. Moreover, to the view a map view and a time view is provided which enables revealing geo-spatial relations within the retrieved documents.*

a content-based layout for time series, (2) mapping time-oriented data content on top of a metadata layout, and (3) revealing relations between multiple data attributes of mixed data type. The implemented techniques for (1) and (2) are directly associated with the VisInfo case study. The usage scenarios of both techniques include parts of the VisInfo data set, as well as the domain experts from Earth observation engaged in the VisInfo design study. For an in-depth description of the techniques and the applied usage scenarios, we refer to Sections 6.3 and 6.4. Both techniques helped us to gain a better understanding of the data, the application domain, and respective analysis tasks. In particular, the techniques allowed us to identify the most interesting relations between the time-oriented data content and the attached metadata. The insights gained with the techniques enabled us to include the most relevant metadata attributes into the VisInfo ESS. With the Geo view and the Time view in Figure 7.6, we explicitly addressed the visualization of the most interesting metadata attributes. The geographical view and the time view include the metadata attributes *Location*, *Longitude*, *Latitude*, *Month*, and *Year*. Hence, the visualization of search results also enables users carrying out relation-seeking tasks. For the identification of interesting metadata attributes, we used our techniques presented in Chapter 6, together with the domain experts. Analysis results from the design phase can, e.g., be seen in Figures 6.7 and 6.16. In addition to the Geo view and the Time view, we provide metadata facets based on the most interesting metadata attributes. We included the metadata attributes *Location*, *Year*, and *Month* to the set of metadata facets (see again Figure 7.5). Together with the domain experts, we identified interesting relations between the data content and seasonal effects. This is why we added the additional metadata attribute *Season* to the set of metadata attributes and to the metadata facets, respectively. The season attribute has two advantages over the other temporal attributes. First, it is more robust than temporal attributes at a coarser granularity. Second, we created the *Season* attribute in a way that it covers temporal effects of the different hemispheres. As an example, the entity *Summer* regards measurements from, loosely speaking, the Months June, July, and August on the northern hemisphere and measurements from the Months December, January, and February on the southern hemisphere. Finally, we added a metadata attribute called *Surface Type* to the set of metadata facets. The surface type facilitates gaining insight into the terrain and the vegetation around measurement stations. In essence, with our techniques presented in Chapter 6, we were able to enhance the VisInfo



(a) *Geo view showing results with respect to their location. Most retrieved documents are measured in the Northern hemisphere.*



(b) *Time view showing results in a calendar. The majority of the retrieved documents contain data measured in the summer period.*

Figure 7.6 *Geographical and temporal metadata build the basis for alternative perspectives on the retrieval result.*

ESS in seeking relations between data content and metadata. In the design phase, this enabled us to gain an overview of the complete body of information and interesting relations. In the application phase, users are now able to seek relations in most interesting metadata, which we included in the workable ESS.

7.1.7. Evaluation and Application

The VisInfo system was developed according to the guidelines for design studies and the principles for user-centered design (cf. Section 2.5). Experts from both the field of DLs and the Earth observation domain were involved in the design of the ESS. With the characterization of the domain, the data, and the tasks, we ensured of deriving relevant system requirements. In addition, the characterizations helped us to validate the progress made in the design study, i.e., to prove whether we built the right solution. In Chapters 4, 5, and 6, we described how the user-centered design principle can be integrated into the workflow for designing visual-interactive interfaces, e.g., for ESS. In each of the chapters, we also presented evaluation strategies to prove the usability and the usefulness of design choices. For each of the four steps of the reference workflow, we involved the domain experts and carried out validation strategies for the technical artifacts of the ESS. The evaluation strategy of VisInfo is also presented in our publications [BRG*12, BBF*11, BDF*15]. We started the evaluation strategy at the very beginning of the design study. An overview of the strategies conducted within the domain and problem characterization of the design study can be seen in [BDF*15]. We started with *paper prototypes*, followed by *informal interviews* with the involved stakeholders. Next, we designed a *rapid prototype*, and presented it to the community via *screencasts* to gather additional feedback. Finally, we carried out a *questionnaire* with digital librarians and Earth observation experts to assess the usefulness and the usability of the intermediate result. The evaluation strategy of the design phase is also presented in [BDF*15]. We chose a set of evaluation steps, each followed by a design iteration of the web-based prototype. On a broad level the design of the final VisInfo system can be subdivided into four cycles. After the first design, we invited usability experts to perform *heuristic evaluation* methods, including a *cognitive walkthrough*. After the second and the third design iterations, we carried out *usability testing* rounds with users of the DL system. Based on the feedback gathered in the iterative evaluation strategies, we successfully improved the Visinfo ESS. Finally, we deployed the web-based VisInfo ESS. We refer to our publication for an in-depth overview of the evaluation strategies [BDF*15].

In addition to the core evaluation strategies, we present usage scenarios to assess the applicability of VisInfo. In the course of the guidelines and techniques described in Chapters 4, 5, and 6, we outlined various use cases and example images applied in the VisInfo case study. Moreover, we present usage scenarios for the application of the VisInfo ESS in several publications [BBF*10, BBF*11, BDF*15]. To conclude the review of evaluation strategies, we echo a typical usage scenario of the VisInfo ESS. Interested readers are invited to test VisInfo by following this link [1]. The repetition of the usage scenario briefly describes the typical steps of an ES process. At the beginning, the user chooses a data and similarity scenario in a dialog presented in Figure 7.1. The user is able to choose between SWD and temperature data,

[1] VisInfo http://demo.vis-info.info/, last accessed on October 08th, 2015.

as well as between absolute and relative curve patterns, depending on the data of interest and their notion of similarity. Afterwards, the SOM-based Visual Catalog is presented allowing users to gain an overview of the data content. The SOM grid with its 8×10 cells represents the data content in a similarity-preserving way; this preservation of topology facilitates the identification of interesting patterns in an intuitive way. We point to Figure 5.18 in Section 5.5.1 on content-based layouts for a SOM-based Visual Catalog example. The user selects a SOM pattern for the detailed analysis, defines the pattern as the visual Query-by-Example and uses the sketch editor to make local adaptions to the curve progression. We refer to Figures 7.5 and 7.6 for different visual representations of a typical search result. The list-based result visualization enables users to explore the data content of retrieved documents in detail. In addition, a list of metadata is shown for each of the retrieved documents. The metadata facets support the drill-down of the search result to the set of relevant documents. A 'find more' button at the bottom allows users to extend the set of displayed retrieval results. In the Geo view in 7.6, the user recognizes that the majority of the retrieved documents are measured in the Northern hemisphere. Moreover, in the Time view, the user recognizes that the majority of the retrieved documents contain data measured in the summer period. According to different insights gained in the course of the information-seeking process, the user is able to focus on other patterns of the Visual Catalog, refine a previously submitted query, or to follow external links provided for relevant documents. This concludes the typical usage scenario of VisInfo — and the section on the VisInfo case study. We Summarize both the VisInfo case study and the following MotionExplorer case study in Chapter 7.3.

7.1.8. Reflection on VisInfo

Discussion and Possible Extensions We briefly review the lessons learned, look at issues for further discussion, and future research challenges. In the course of the design study, we identified some insights of the design of ESS which may be beneficial for new approaches.

First, we can confirm the *missing 'trust'* of scientists in unfamiliar techniques and the refusal to change their working routines [BWE06, HPK08, KMS*08]. Naturally, informal interviews with the domain experts have shown that bringing ES into their fields is of great value for their work. It is widely recognized that for data-driven scientific discovery, new analytical solutions are required, implementing content-based search and exploration in large collaborative DLs going forward. However, many domain experts prefer working with their own tools, even if they are aware that the analytical capability of these techniques is limited, to some extent. One possible solution for this concern is user-centered design. Raising the trust for solutions provided by external data scientists can be facilitated by collaborative data cleansing, similarity definition, and content-based overview approaches. It may play a key role in building trust if visualization is included in the design of workflow's mandatory steps.

Second, *facilitating usability* is most relevant for the design of innovative visual-interactive ESS. Otherwise, a user with no prior exposure to visual search may face a barrier to entry before she can use the system effectively. Again, including domain experts in the design from the very beginning may help to resolve this challenge. In this way, the familiarization efforts of the users will be mitigated.

Third, we identified the *need for various, complementing evaluation strategies*. Choosing appropriate evaluation methodologies in the course of the implementation phase was important. Based on the iterative evaluation strategy, most usability barriers could be identified and subsequently overcome. The design study methodology including the characterization of data, user, and task [MA14], as well as carrying out design workflows [Mun09, SMM12] helps to build the right solution and to build the solution right.

Another point of discussion is the *variety of analysis tasks and research goals* associated with data-driven research. We presented four different similarity scenarios including different data, data representations, and similarity functions. However, many other similarity notions exist which would allow the domain experts carrying out other research goals. Thus, for expert users manipulating parameters for the most relevant models within the scientific workflow are highly appropriate. We expect VA to play a key role in many future ESS approaches.

Another discussion item is the *retrieval algorithm*. In VisInfo, we developed an index for the retrieval of high-dimensional vector data. The implementation is based on the concept of vector quantization similar to the SOM algorithm for the aggregation of data. The index enables search result calculations with nearest neighbor queries in under one second. The increase in speed compared to a sequential scan search algorithm is measured at 1,250% on average. We calibrated the index in such a way that the search result matches the search in more than 99.9% of all searches. However, other index structures, such as tree-based algorithms exist, possibly leading to faster searches. It would be interesting to compare the performance of different retrieval algorithms for larger data sets possibly consisting of millions of daily curve progressions.

We identified various possible extensions of the VisInfo ESS. Among others, we identify the following research challenges:

- Advanced search interfaces may go beyond Query-by-Sketch or -Example, but allow domain experts to formulate hypotheses in other useful and appropriate ways. Examples are the specification of correlations, or lags in time and space, helping to discover interesting data sets.
- A deeper integration of the content-based search with published research papers. Techniques from Natural Language Processing (NLP) can be used to automatically link textual descriptions extracted from research papers to phenomena observed in the non-textual primary data, possibly enhanced by cross-referencing the DOIs. From a VA perspective a combination of our techniques with, e.g., the TopicNets approach for large text corpora using the LDA topic modeling algorithm is promising [GOB*12].
- We currently see a lack of collaborative *usage*. The visual search system can be extended to an annotation, rating and referencing scheme to improve the exchange of data among scientists. This may include recommender functionality which correlates with, e.g., user profiles, or search sessions.

Conclusion We presented VisInfo, an ESS providing visual access to time-oriented primary data. Based on a content-based layout, the system enables users to gain an overview of large collections of primary data. The overview visualization allows carrying out exploratory analysis tasks including an interactive drill-down to local and detailed aspects of interest. To provide efficient retrieval, VisInfo supports the visual-interactive definition of content-based queries by sketch or by example. Additional metadata-based filters enable users to reduce the search space to meaningful subsets. The visual search-result representation provides different views, including a visual representation of the raw time-oriented data content, a geographical visualization, and a calendar-based visualization. The VisInfo design study was a collaborative approach between digital librarians, data scientists, usability experts and Earth observation scientists. The development process of VisInfo was performed with the user-centered design principle. In the domain and problem characterization phase, different evaluation strategies were conducted, including paper prototyping, user interviews, screen casts, and user questionnaires. The design phase was accompanied by three evaluation phases all supporting the iterative improvement of the system. In particular, a heuristic evaluation and two usability testing rounds with expert users helped to improve VisInfo with respect to its usability and usefulness. The web-based VisInfo system is online, interested readers are invited to test VisInfo by following this link [2].

[2] VisInfo, http://demo.vis-info.info, last accessed on October 08th, 2015.

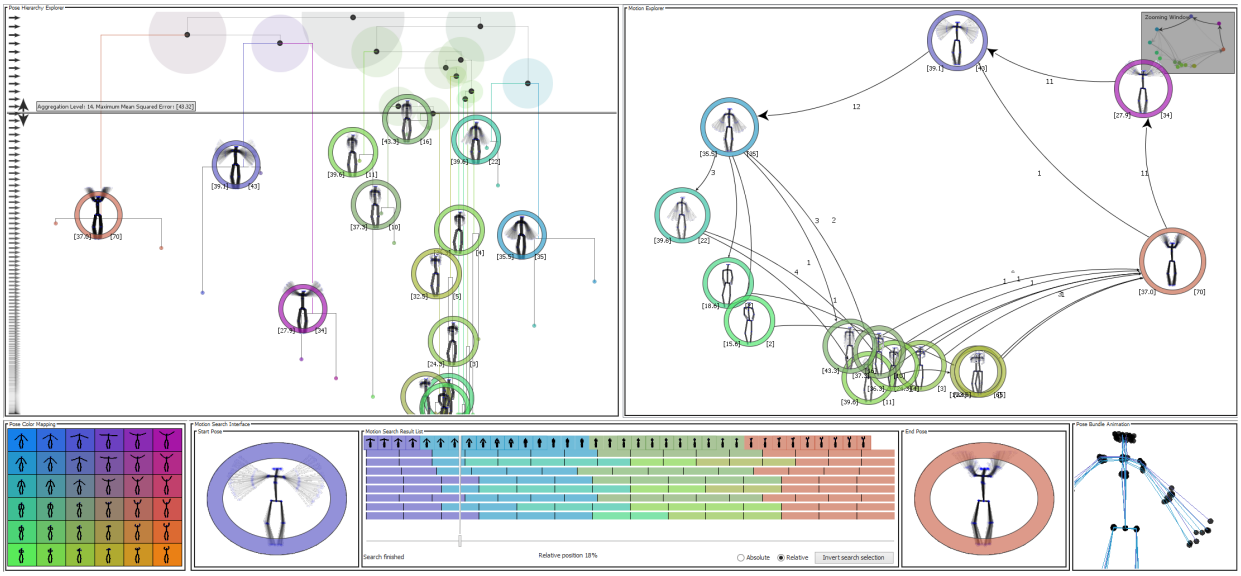


Figure 7.7 *MotionExplorer facilitates the ES in human motion capture data. A content-based overview based on a user-steerable clustering algorithm enables users to adjust the aggregation level (top left). A second overview supports the exploration of human poses and motions (top right). Both views enable users to define queries by example. The search interface (bottom) facilitates the interactive exploration of search results including the identification of style variations in human motion.*

7.2. MotionExplorer — Exploratory Search in Human Motion Capture Data

> *Can I take screen shots with this tool? I would like to show my analysis results to my colleagues.*
>
> *I've been working with motion capture data for eight years now, but this perspective of the data was really enriching. I had a lot of fun with the tool.*
>
> *No need to analyze poses and sequences by hand. Producing an effective result in only a few minutes. I like that.*

user feedback for the MotionExplorer system [BWK*13], *2013*

We present MotionExplorer [BWK*13], an ESS for sequences of human motion in large motion capture data collections (see Figure 1.6). This special type of multivariate time-oriented data is relevant in many research fields including medicine, sports and animation. Key goals of our targeted domain experts are the exploration of motion states (poses) and motions (transitions between poses), as well as the search for relevant motion subsequences. A downstream task performed by the domain experts is the synthesis of motion vectors by interpolation and combination. In the practice of research and application of human motion capture data, challenges exist in providing visual summaries and drill-down functionality for handling large collections of motion data. MotionExplorer provides a content-based *overview of human poses* including an interactive data aggregation concept. A second content-based overview sheds light on the *motion sequences* available in the data collection. The overviews serve as a basis for exploratory browsing and an information drill-down. Moreover, a *Query-by-Example* implementation enables domain experts to select visual queries, and thus to search for motion sequences. A third component combines the visual query interface and the visual result visualization for the *exploration of retrieved motion sequences*. Human poses are linked in all provided views with a similarity-preserving color mapping. We developed MotionExplorer in a

design study setup in close collaboration with the targeted users including a summative field study. Additionally, we conducted a laboratory design study to substantially improve MotionExplorer towards a usable, useful and robust design. MotionExplorer enables a search in human motion capture data with only a few mouse clicks. The domain experts unanimously confirm that the system can efficiently support their work. In the following section, we will look at the MotionExplorer design study with respect to the reference workflow, and the technical contributions presented in this thesis. The MotionExplorer ESS is presented in [BWK*13]. Furthermore, the system is based on techniques presented in [BRS*12a, BWS*12, BSM*15b, WVZ*15, BSM*15b].

7.2.1. Characterization of the Domain, Data, and User Tasks

Characterization of the Domain We briefly describe the human motion synthesis domain. The overall goal of domain experts working in the field of human motion synthesis is an effective reuse of existing motion capture data to synthesize new human motions. Thus, an overview of existing data is important to understand the potential building blocks for new motions. However, domain experts have difficulties in gaining an overview of large collections of human motion capture. Our involved user group reported on in-depth inspections of data collections which lasted for days. Our users distinguish between motion on the macro and micro level. On the macro level, different types of motions are identified (e.g., activities, such as cross-country-skiing in the classical or the skating technique). On the micro level, style variations of the same motion type are determined (e.g., different speeds, dynamics, or skill levels of the skating technique). Motion synthesis typically involves identifying first the motion type of interest and then selecting an appropriate style to use or adapt. For human motion capture data, a variety of content-based retrieval techniques exist. Visually querying human motion capture data by example, however, is difficult. As an example, the approach by Müller et al. supports the content-based retrieval of human motion capture data based on short motion clip queries and a query-dependent specification of the notion of similarity [MRC05]. For more details of the goals, the state-of-the-art and the open challenges of the research domain on human motion synthesis, we refer to [BWK*13].

Characterization of the Data Human motion capture data is an instance of multivariate time-oriented data. It is applied in various research fields, such as medicine, sports, and animation. It can be acquired from human actors marked with detectable vector nodes using video tracking. As an alternative, motion data can be obtained synthetically by simulation. In the case described by the involved domain experts, the motion sequences were performed by actors, recorded by a multi-camera system. The 3D positions of the markers are reconstructed from the 2D images via triangulation for each time frame. Consequently, motion capture primary data typically consist of frames containing a high-dimensional vector representation of a human pose. Depending on the number of markers, the resulting multivariate time-oriented data can contain hundreds of dimensions (for every time stamp). The frame rate of camera takes, imported in the MotionExplorer ESS, have a temporal resolution (frame rate) of 120 Hz. Hence, every second the amount of scalar values contained in the raw primary data is approximately 15,000. It is obvious that the manual analysis of the raw time-oriented primary data is very time-consuming. However, the retrieval of a motion sequence of interest often has to be done by manually screening the set of motion sequences. For research purposes, many test data sets have to be generated. Usually, data is recorded only for a specific project. The reuse of data is often only feasible as long as those people are available who recorded the data. Fortunately, some motion capture collections are also stored in large primary data warehouses where they are available for re-usage. In our experiments, the domain experts recommend obtaining primary data from the HDM05 motion capture database [MRC*07]. This database is a systematic recording of a wide variety of motions performed by various actors, in multiple repetitions. Many instances of similar human movement are also available as sets of extremely varied behaviors. Thus, the database provides a good basis for investigating both the effectiveness and the efficiency of our approach.

Characterization of the User Tasks The overall goal of our domain experts is the synthesis of human motion data based on an effective reuse of existing motion segments. A variety of algorithms for both the retrieval of human motion capture data and the synthesis of segments exist. However, ES capability for the utilization of the retrieval algorithms has not been presented thus far. Subsequently, we illustrate the missing functionality in the analytical workflow of the domain experts. In particular, we identified four high-level questions that will be addressed with MotionExplorer.

- Which variations of *poses* are available in the data set?
- How can *queries* be specified intuitively?
- Does a *motion* exist for reuse, or do I need to record new camera takes?
- Which *style variations* exist from start pose A to end pose B?

In the course of the domain and task characterization phase, we were able to formulate the analysis tasks and the central challenges of the domain experts more precisely. First, as motion data collections are typically large, an overview is needed. Visual access to large collections of motion capture data has not become a common application field for VA thus far. In addition, the spatio-temporal variations in existing human motion data make this complex data type inappropriate for many existing tools. Due to a lack of specialized tools, the domain experts report on working through their data collections, manually in many cases. Second, visual-interactive functionality to obtain meaningful data subsets efficiently is scarce. While algorithmic search methods for indexing and matching of motion data exist, we identify a lack of visual query formulation interfaces in practice. Third, identifying interesting sequences from the retrieved search results remains challenging. While in large motion databases style variations typically exist, these are hard to distinguish automatically but often require subjective interpretation by the expert. Visual representations for the exploration of retrieved sequences are expected to be useful, but have not yet been put often into practice.

7.2.2. Requirements

As a result of the domain, data, and task characterization, we discovered a set of 10 functional requirements to resolve the described challenges. We briefly summarize the requirements as follows.

- *Overview*: The system should provide the big picture of the data. A global ordering would be nice to access similar data.
- *Aggregation*: To reduce the complexity, large data sets can be aggregated. The level of aggregation should be adjustable.
- *Cluster glyph*: Data aggregations should be displayed as visual structures. These should support an intuitive assessment.
- *Motion graph*: This metaphor should be integrated. Poses should be displayed as nodes and motion sequences as edges.
- *Level of detail*: Local behavior of the data should be explorable. This enables the domain experts to analyze details.
- *Filtering*: It should be possible to exclude and re-include data to focus on relevant subsets.
- *Visual querying*: Formulating example queries for motion sequence search should be as intuitive as possible.
- *Path search*: A state of the art retrieval technique should be integrated to search for motion sequences.
- *Search result exploration*: Style variations of different retrieved sequences are interesting and should be recognizable.
- *Multiple views*: Different aspects of the data should be shown side-by-side in the same system.

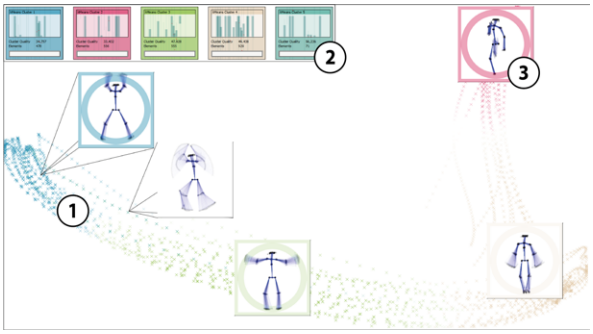
7.2.3. Time Series Preprocessing and Similarity Definition

The content-based access strategy of MotionExplorer is based on the recommendations of the domain experts and the best practices known in the motion data analysis domain. We present a brief overview of the data preprocessing and the similarity definition process. Based on a recommendation of the domain experts, we defined a single human pose as one 'atomic' data object. The HDM05 motion capture database [MRC*07] provides the raw time-oriented primary data as binary C3D-files where motions are represented as sequences of single poses. The domain experts suggested to shift the pelvis of each human pose to the origin of the coordinate system. To obtain poses that are always viewed head-on, we normalized each pose by a rotation around the z-axis. In this way, the human poses become comparable. As the next step, we extracted a compact and yet precise FV for downstream overview and search tasks. Based on a proposal of the domain experts, we reduced the 3D markers to a relevant subset. As a result, the MotionExplorer system obtained a FV with 48 dimensions containing 16 3D marker coordinates in absolute scale. In agreement with the domain experts, we use the Euclidean distance measure as the default similarity function. We refer to [BWK*13] for an in-depth description of the data abstraction and the preprocessing workflow.

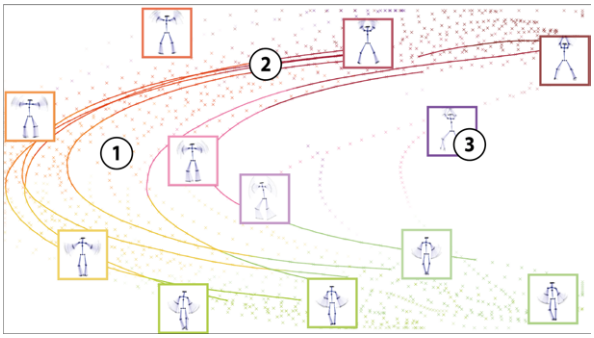
7.2.4. Content-Based Overviews of Time-Oriented Data

For the MotionExplorer system, we used the guidelines and techniques for content-based overviews (cf. Chapter 5) twice. The rationale for this strategy was to present two overviews, one for individual poses and one for human motions as available in the data set. Both overviews share a common data aggregation and visual mapping step of the reference workflow (cf. Figure 3.1). The layout step, however, is based on two different techniques.

Data Aggregation A cluster of human poses is the targeted data aggregation concept in MotionExplorer. A pose cluster is the principal data structure for the content-based overviews and the visual querying concepts. Every cluster contains large numbers of similar human poses, i.e., the spatial aggregation of the poses, in the vocabulary of our targeted domain. The data aggregation step was inspired by a previous work on multivariate Earth observation



(a) *Prototype of a content-based overview using the results of earlier works: (1) individual data elements (poses) visualized as scatter diagram, (2) a generic glyph design for high-dimensional clusters, (3) pose aggregates in a preliminary glyph design.*



(b) *Prototype of a content-based overview combining visualizations of poses and motions: (1) individual data elements (poses) visualized as scatter diagram, (2) individual motions visualized as splines, (3) pose aggregates in a preliminary glyph design.*

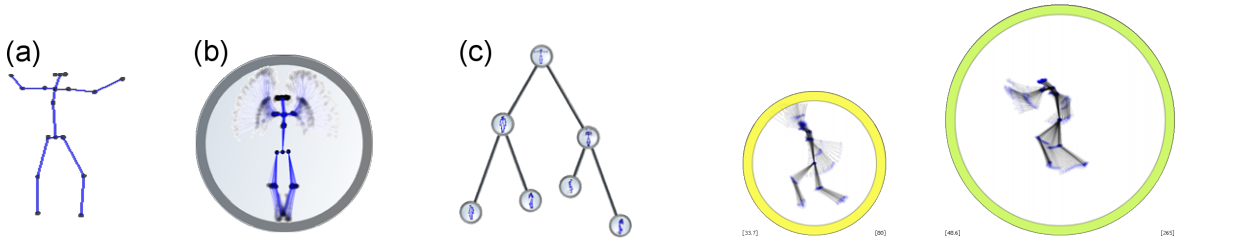
Figure 7.8 *Two early prototypes of the design phase. The k-means algorithm was used for the aggregation of human poses (left: $k = 5$, right: $k = 12$). Different stages of the glyph designs are shown. The micro-macro views concept is applied to show both the data elements and the data aggregates [BvLBS09, BWS*12]. While the coloring is still categorical on the left, the design iteration on the right already provides a similarity-preserving colormap.*

data [BWS*12]. In this publication, we used the micro-macro views technique to present both data elements and clusters in a combined content-based overview [BvLBS09]. By default, we visualized single data elements (micro level) with a scatterplot metaphor. In addition, we provided steering parameters enabling users to calculate and visualize data aggregates as a second layer on top of the visualization. An early prototype of MotionExplorer adopting the proposed micro-macro views technique is shown in the left image of Figure 7.8. Already in the domain characterization and data abstraction phase, we identified the curiosity of the domain experts for the aggregation of human poses. Particularly, the ability to aggregate human poses 'at runtime' would be greatly desirable. This is why we made the design decision to neglect the micro level, when we discussed the early prototypes. On the contrary, we rather focused on the exploration of pose aggregates, including the visual-interactive adaption of the level of data aggregation. This is one of the decisive design considerations why MotionExplorer provides a hierarchical clustering concept for the data aggregation step. A cluster hierarchy enhances the adjustment of the number of presented clusters k without the need of a recalculation of the data aggregation (cf. Section 5.2.1). Moreover, the nested concept of the cluster hierarchy allows splitting individual branches of clusters while other data aggregates remain unchanged. For the calculation of the cluster hierarchy, we use a divisive clustering algorithm. The algorithm supports multi-threading to scale for large data sets. The divisive nature of the algorithm at first reveals the coarsest clusters of the hierarchy, which in turn are instantly available for the content-based overview visualization. The domain experts may control two clustering parameters:

- (a) the principle which cluster is to be split next when k is increased
- (b) the splitting strategy for a particular cluster

Since we want to obtain compact clusters, we define (a) splitting the cluster with the maximum standard deviation as a default. As an alternative, we suggest splitting the cluster with the highest number of elements. Concerning (b), we follow a representative approach and apply a k-means clustering with $k = 2$ as a default.

Visual Mapping Inspired by the human anatomical drawing by Leonardo da Vinci (Vitruvian Man), we chose a circular design for the cluster glyph (see Figure 7.9). The glyph design was carried out based on the requirements of the domain experts and according to the guideline presented in Section 5.4. The cluster glyph is also shown in Table 5.1 of real-world examples using our guideline for cluster glyph designs. The cluster glyph represents the value domain of the cluster (the nearest neighbor of the centroid) as a human stick-figure pose, and the set of poses in the cluster as deviating, transparent figures. The number of displayed poses for each cluster is limited to a maximum of 500. We provide two additional labels for every cluster glyph; one label displays the compactness of the cluster while the other represents the cluster size. The relative cluster size is displayed by the size of a surrounding circle. Finally, we color the cycle of each cluster glyph to illustrate similarity among the clusters. For this purpose, we use the guideline on 2D colormaps for the similarity-preserving coloring of multivariate objects presented in Section



(a) *Data structures designed in the data characterization and aggregation phase. A single human pose/frame consists of a FV with 48 dimensions (163D marker positions) (a), a pose aggregate/cluster consisting of multiple (similar) poses (b), and the hierarchical clustering tree (c).*

(b) *Final design of the cluster glyph for human poses. The style variations of human poses around the cluster centroid are shown by opacity bands. In addition, we apply similarity-preserving colors and indications for the cluster size.*

Figure 7.9 *Results of the data aggregation and the visual mapping step of the reference workflow (cf. Figure 3.1).*

5.4.2. In MotionExplorer, we particularly make use of the color coding concept by linking all seven views with the similarity-preserving color information (see the title Figure 1.6). To this end, we provide a similarity-preserving color legend at the bottom left corner of the system for a straightforward lookup of poses and respective colors. The grid of the color legend is the result of a SOM using all pose FVs in the manner of a vector quantization scheme.

View Transformation / Layout We present two different content-based overviews based on two different layouts. While the layout at the top left of the system (cf. Figure 1.6) provides an overview of human poses the layout at the top right of the system allows gaining an overview of human motions.

The first overview is a cluster structure-based layout directly using the hierarchical structure of the clustering result. The view meets the follow requirements:

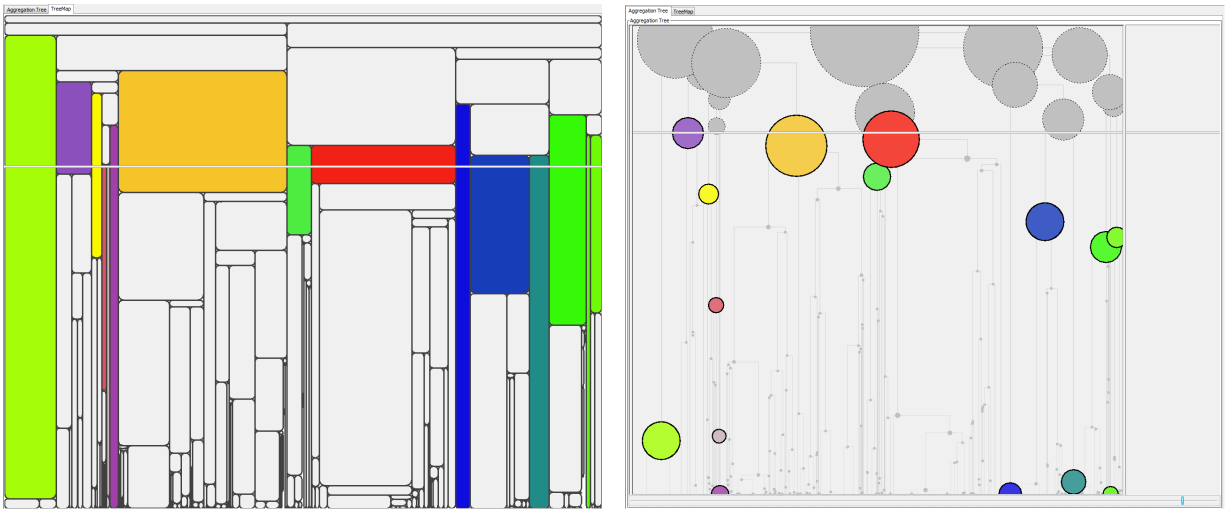
- it serves as a global overview of human poses by representing the tree structure of the cluster hierarchy
- it enables filtering to narrow down the number of elements
- it provides interactive level of detail functionality
- it supports the formulation of visual queries

Early prototypes of the pose cluster hierarchy visualization are shown in Figure 7.10. The first prototype is based on a treemap metaphor while the second prototype uses a tree/dendrogram metaphor. The two designs were evaluated in an in-group user study, conducted together with experts in data science within the design phase of MotionExplorer. As a result of the study, we neglected the treemap-like approach. The decisive advantage of the tree-based variant is the applicability of the y-axis for the slider control using the level-of-detail concept required by the domain experts. The final tree-based layout of the cluster hierarchy is presented in Section 5.5.1 as an illustrative example of cluster structure-based layouts (cf. Figure 5.24). Additional illustrations including usage scenarios are presented in the MotionExplorer title Figure 1.6 (top left). The visual cluster hierarchy allows the interactive adaption of the aggregation level. To this end, the visualization provides a slider control bar which can be dragged in vertical direction. A change event triggered by the slider interaction automatically adapts the number of displayed clusters in both content-based overviews.

The second overview for the visual representation of the human motions is based on a projection-based layout. The functional support of this content-based overview is as follows:

- it provides an overview of all poses and motions at the current aggregation level following a node-link metaphor
- it preserves the global order of the poses to intuitively assess their similarity
- it provides interactive means of avoiding local overplotting
- it supports the visual query formulation

Early prototypes of the content-based overview for human motions are presented in Figure 7.8. In the left design, the human motion sequences were only implicitly visible. In this early prototype, the path lines of the individual pose data elements indicated the human motion. In the right design, individual human motions were additionally drawn with spline-based path lines. Together with the domain experts, we finally decided to choose a node-link diagram for



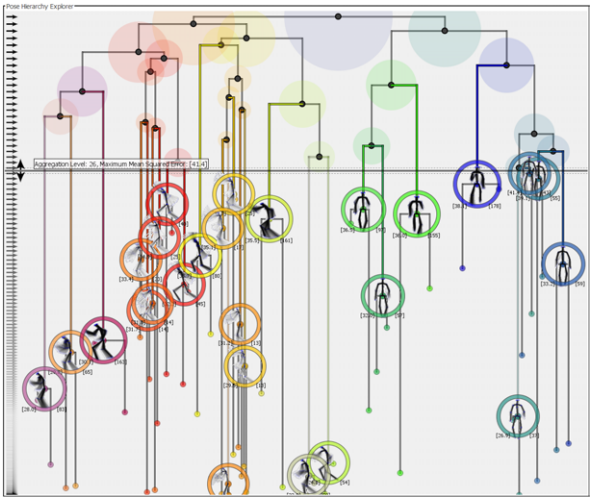
(a) *Treemap-like visualization.* **(b)** *Tree/dendrogramm visuzalization.*

Figure 7.10 *Two early prototypes for the visual representation of the pose cluster hierarchy. In both prototypes the final cluster glyph including stick-figure poses was not included yet. In a case study conducted with experts in data science, we identified that the intuitiveness of the second variant for the required interactive level-of-detail concept is significantly better.*

the visualization of both human poses and human motion. This visual metaphor is very close to the domain experts' notion of a 'motion graph', which is a highly relevant data and index structure in their domain. As a consequence, the content-based overview provides pose clusters as nodes and aggregated motions between the human poses as links. In other words, we abstract single motions to motion aggregates, which mitigates overplotting problems significantly, especially if the data collection is large. This is why the visual representations of both poses and motions are based on aggregation concepts in the final MotionExplorer ESS.

We present different examples showing the final design of the content-based overview of human motions. In the introduction chapter, we illustrated the MotionExplorer ESS through the example of jumping jack motions and kick motions in Figure 1.6. In Figure 7.7, the title figure of this case study, the MotionExplorer system is demonstrated supporting the ES of rotating arms motions. In Figure 7.11, we demonstrate the final content-based overviews with two different types of cross-country skiing. The layout strategy for the content-based overview is based on the PCA projection, mapping the current set of pose clusters into the display space. The linear PCA preserves the global structure of the high-dimensional poses, and thus provides a similarity-preserving layout in 2D (cf. Section 5.5.1). Within the design phase, we also assessed the value of non-linear projections, as well as force-directed layout alternatives. Our rationale was to balance the trade-off between the preservation of the structure and the avoidance of overplotting. The decisive argument of the domain experts for the PCA-based approach was that overplotting poses are far less severe than losing the preservation of topology (the global structure of the high-dimensional feature space). In fact, the identification of superimposed clusters is a valuable information for the domain experts for revealing densely populated cluster regions in the data space. MotionExplorer provides two interactive overplotting avoidance strategies. The first is a technique based on a so-called 'flower metaphor' shifting superimposed clusters aside in combination with mouse interaction. The flower metaphor can be seen for the motion at the lower right in Figure 7.11b. All glyphs are interactively aligned next to each other without local overplotting. The second technique contained in the content-based overview is zooming and panning. A small overview window at the upper right of the display preserves the global overview. The zooming interaction can either be used to avoid local overplotting or as a means for information drill-down towards local aspects of interest. The latter supports the visualization of the overview at different levels of abstraction. This is in accordance with the proposed guidelines and techniques for different levels of abstractions presented in Section 5.5.2.

The clusters shown in both content-based overviews serve as the basis for querying by example. Two interaction designs are provided to facilitate Query-by-Example. The first variant is dragging a pose into the search interface, dropping the pose starts the retrieval algorithm. The second variant is based on a Details-on-Demand concept. A click on a pose opens a detailed view enabling the analysis of local aspects. The Details-on-Demand view also provides



(a) *Exploration of a hierarchy of human poses. A dendrogram visualization shows the result of a hierarchical clustering algorithm. Domain experts can adapt the number of shown clusters (currently 26) by dragging the aggregation level slider. Changes in the aggregation level are delegated to the other views of the system*

(b) *The motion explorer shows human poses as nodes and motion sequences as edges. In this example, skiing poses are represented from a lateral and a frontal view. The projection-based layout clearly distinguishes between the two postures. The prevent-overlap function (referred to as 'flower metaphor') is active in the sub-graph at the bottom right.*

Figure 7.11 *The two complementing overviews of the MotionExplorer system. On the left, a hierarchy of poses allows steering the aggregation level. On the right, a node-link diagram represents poses and motions. Zooming and panning supports the exploration of human motions, filtering reduces the number of shown items.*

two blue buttons for assigning the displayed pose as a Query-by-Example. We refer to Section 5.5.3 for an in-depth description of the Query-by-Example technique including an illustration in Figure 5.19. Finally, we briefly describe the filtering concept implemented in MotionExplorer. In both content-based overviews a right click on a human pose triggers the filter routine. In the tree-based overview (top left) the pose cluster colored gray, a right click on a filtered pose re-includes the pose to the set of considered poses. In the node-link (top right) the pose is excluded from the visualization, human motions connected with respective poses are removed as well. The retrieval algorithm also neglects filtered poses and motions.

7.2.5. Visual Interactive Querying and Search Result Exploration

The MotionExplorer ESS provides a visual search interface enabling the domain experts to:

- run the human motion retrieval algorithm of the domain experts based on querying by example
- analyze style variations in the search result based on exploratory analysis capability
- drill down the number of retrieved sequences to the number of relevant results

The search interface is displayed at the bottom of the MotionExplorer ESS. The interface is divided into the following components:

- a grid-based color legend facilitating the mapping between the human stick-figure poses and the similarity-preserving 2D colormap
- a component that represents the start pose of a visual query
- the search-result list of retrieved human motion sequences at the center
- a component that represents the end pose of a visual query
- the pose bundle animation view for the detailed exploration of style variations in retrieved motion

The list-based result interface at the center shows retrieved sequences with the granularity of single frames. The frames are separated by black vertical lines. Again, the similarity-preserving 2D colormap is used to link motion

Figure 7.12 *Collaboration with users in the course of this design study.*

segments to the pose color mapping. In the usage scenario illustrated in the title Figure 1.6 12 motion sequences were retrieved holding poses with colors from red to green. In consultation with the domain experts, resulting motion sequences are warped to fit with the horizontal display space per default. This enables an easy comparison of retrieved sequences along the relative time axis. The display of sequences in an absolute arrangement of time is also possible. An illustration of the interface for both the relative and the absolute time axis was presented in Section 5.5.3 as an example of visual querying in Figure 5.29. Six kickboxing motions were retrieved with the MotionExplorer system following a color progression from orange to purple. Depending on the visual representation of the temporal domain in relative or the absolute mode, different style variations can be observed at the micro-level. A time slider enables the exploration of style variations in an interactive way. By dragging the slider, an animation of all corresponding result poses is shown in the pose bundle animation view (bottom right). A filter function enables the domain experts to reduce the number of style variations, inverting the filter status is possible. For additional information about single subsequences, a focus-and-context interaction is implemented. By focusing the result list, the currently accessed subsequence is enriched by a key frame representation of the particular human motion. This search interface enables the domain experts to find interesting motion style variations across different results. In the title Figure 1.6 the focus-and-context interaction is active for the third motion sequence.

7.2.6. Evaluation and Application

MotionExplorer was developed in a design study setup. The experts from the human motion synthesis domain were involved in the design of the ESS from start. In addition, in the design phase of the system, we additionally evaluated design considerations together with groups of non-experts. A coarse overview of the collaborations with users is presented in Figure 7.12.

In an early phase, we conducted formative field observations and interviews with the domain experts. This helped us to characterize the domain of the involved expert group, their data and their considered tasks. In the course of this familiarization phase, we also introduced the domain experts to the domain of IV and VA. As a result of this early phase of the study, we were able to define the functional requirements for MotionExplorer. The functional requirements helped us to develop and validate MotionExplorer in the subsequent phases of the study.

The system design phase of MotionExplorer was carried out in an iterative and user-centered way. We conducted regular consultations with the domain experts including visual prototyping and design weighing strategies. The construction of the workflow of MotionExplorer was carried out in accordance with the reference workflow (cf. Figure 3.1) presented in Chapter 3. We adopted functional artifacts of the domain, such as the feature extraction and the retrieval concept. Previous works enabled us to present rapid prototypes and to re-factor useful components, such as the micro-macro views concept implemented in [BWS*12]. The data abstraction, visual mapping and layout step of the workflow were designed with respect to the techniques and guidelines presented in Chapter 5.

Within the design phase, we also conducted a user study together with non-experts. The user study helped us in three different aspects. First, we were able to assess the intuitiveness of the design decisions made together with the domain experts. The results helped us to make assumptions on the generalizability of the designs. Second, we were able to increase the number of design iterations, even if the domain experts were not involved. Third, we were able to gain additional feedback, recommendations, and usability considerations. The most relevant issues covered by the user study were as follows:

- The cluster structure-based layout of poses including the interaction design for steering the aggregation level
- The content-based overview of motions including the interaction design for the avoidance of overplotting
- The glyph design for individual human pose clusters
- The visualization and interaction design for the visual-interactive querying and search-result exploration

The final evaluation of MotionExplorer included case studies conducted together with the domain experts. The motivation for the case studies was the question whether MotionExplorer helps the domain experts with their research.

Figure 7.13 *Two of the pictures taken at the observational, summative field study conducted together with domain experts with the primary data of their choice (used with permission).*

We wanted to assess to which extent the domain experts would be able to identify their data collections and if the visual encodings and the interactive functionality is intuitive and usable. We were also interested in whether the domain experts would accept the system as a tool for decision making in their scientific work. For this purpose, we conducted an insight-based summative field study with five domain experts in human motion synthesis at their lab. Due to the small number of participants, we neglected our findings based on quantitative test variables. The procedure was as follows: After a familiarization, we introduced the domain experts to the purpose of the field study. We presented an introduction to the tool and to the supported analysis tasks. We divided the study into two parts. Each participant was asked to (a) perform four specific tasks and after that (b) was invited to run the system in an exploratory manner on her own data sets without supervision. The tasks covered the main workflow of MotionExplorer, i.e., the exploration of human poses and motions based on a gained overview, the formulation of a visual query, and the exploration of retrieved motion sequences. The exploratory part of the field study was conducted on data collections recommended by the domain experts. Thus, we were able to observe the domain experts using the system in a real-world setup. Most interesting for us was determining how fast the domain experts would familiarize themselves with the system, and what data findings or insights would occur. We briefly review the results of the case study in Table 7.1, profound details about the case study's setup and the evaluation results are presented in [BWK*13].

7.2.7. Reflection on MotionExplorer

Discussion and Possible Extensions To the best of our knowledge, our system depicts one of the very first ESS for large motion data based on a customized and interactive data clustering approach. The layout of the content-based overview of human poses is based on a hierarchical clustering result. The layout of the content-based overview of human motions uses the output of a data projection algorithm for a similarity-preserving visualization of human poses and motions, respectively. By taking new technologies for motion tracking into account, we expect much more motion data in future. This development induces the need of scalable retrieval and exploration techniques to become available. As shown qualitatively, our system could be operated by expert and non-expert users in an intuitive and effective way. The received comments indicate that the approach can support various information-seeking behaviors in (human) motion analysis. The motion data representation in our case is an instance of multivariate, time-dependent data. This data is typically hard to visualize. A glyph design in combination with a similarity-preserving color encoding facilitates the identification and localization of poses (data aggregates) in different linked views. In our case, we could draw on representing the motion poses by a stick-figure metaphor, which is easy to interpret by users.

We also identified a number of limitations of our current study that should be investigated as a next step. First, pose clustering in our approach is done by interactive clustering. This relies on the domain experts being able to find useful cluster parameters and interpret the obtained cluster. Also, automatic classification of poses should be integrated. Based on training data, one may train classifiers to automatically segment different poses. Similarly, the identification of style variations at the micro level may be difficult to do only visually, especially if variations are all mapped to

Eval. Criterion	Description
New Insights	The domain experts reported on surprisingly large numbers of style variations in their data. Filtering and zooming helped to drill down the displayed information to (local) aspects of interest.
Communication	The domain experts identified the ability of the system to help communicating search and analysis results. The possibility to visually verify hypotheses and to communicate outcomes was appreciated.
Usability	The domain experts were pleasantly surprised that they learned to use the system very quickly. They welcomed the intuitive description of the cluster glyph of human poses and the interplay of the different views. After the task completion tests, all domain experts familiarized with the system and were, thus, able to work autonomously.
Effectiveness	We recognized a curiosity about the system among the domain experts, who recognized the components of the system and their interplay as useful. The visualization and interaction designs form a helpful solution to carry out ES tasks with a previously unexpected effectiveness. As demonstrated in the case study, domain experts successfully performed a re-identification of known data aspects.
Efficiency	A reduced effort in performing analysis tasks was recognized by all domain experts. Gaining fast insight into the complex data type of human motion sequences by the overview visualizations was widely accepted, especially if the data collection is large. The identification of human poses, motion sequences and style variations thereof was considered simple, which was quoted as very time-consuming in the past. However, we acknowledge that not all participants were able to perform a search with the minimal costs of five clicks.

Table 7.1 *Summary of results of the insight-based summative field study conducted with real users and real data.*

the same glyph. In densely populated glyphs it may be difficult to visually discriminate the styles. Again, automatic classification can be useful to this end.

As another possible extension, the poses and motion (pose transitions) can be enriched with meta data (if available) of the motion recordings. If certain poses have been pre-classified, this should be included in the visualization. Interactive capability for efficiently labeling glyphs and transitions is also possible.

Finally, in our field study, we only valued the qualitative findings because the population of the participants was low. More formal studies, including also quantitative measurements are desirable to further assess the system capabilities. For example, we may more closely incorporate our ESS into workflows for interactive motion synthesis and observe the improvement potential.

Conclusion We presented MotionExplorer, an ESS for large data collections of human motion capture data. The system provides an overview of human poses in a tree-based visualization as a result of a hierarchical clustering. A node-link diagram enables users to gain an overview of human motions where the nodes represent human poses and the links indicate existing motion sequences. The visual mapping of data aggregates is carried out with a glyph design based on a stick-figure metaphor including a similarity-preserving color encoding. Visual interfaces for query formulation and the exploration of search results support the user in retrieving human motion sequences and style variations thereof. MotionExplorer was developed in a design study setup together with domain experts from the start. In addition, we evaluated the system by laboratory design interviews with non-experts. Finally, we conducted a field study with the domain experts. Based on the field study, the experts confirm the usability and efficiency of the system to support their day-to-day work. Analysis results within the field study also presented new aspects of the data that had not been taken into consideration or had not been visually communicated before. Finally, we discussed a number of limitations and extension possibilities of our work that should be considered in future.

7.3. Summary

In this chapter, we presented the case studies of this thesis. In two ESS, we showed how the ES concept can be applied to time-oriented primary data. To the best of our knowledge, the two approaches are among the first ESS for time-oriented primary data. Both case studies were conducted with principles from design study methodology and user-centered design. Based on characterizations of the domain, the data and the user tasks, we identified analytical challenges of the involved domains, and put these into functional system requirements. On this basis, we applied the reference workflow for the design and application of visual-interactive interfaces for ESS $\mathbf{C}_{\text{MES}}$. The challenge of a

meaningful choice of models and model parameters $\mathbf{C}_{\text{MPC}}$ was of particular concern in all main technical components of the two ESS. Similarly, we forced the involvement of users in any important step of the design $\mathbf{C}_{\text{UCD}}$. In this connection, the incorporation of the guidelines and techniques presented in Chapters 4, 5, and 6 was greatly beneficial. We employed the visual-interactive preprocessing system presented in Chapter 4 facilitating the content-based access to time-oriented primary data $\mathbf{C}_{\text{CBA}}$. Together with the domain experts, we resolved quality issues, applied normalization and segmentation routines, and agreed on meaningful time series descriptors. In addition, we conducted user studies to assess the time series similarity function which matched the notions of similarity in the heads of the domain experts. As a result, we obtained FVs and similarity functions which posed the input for the data aggregation step of the reference workflow. We designed content-based overviews $\mathbf{C}_{\text{CBO}}$ in the terms of the guidelines and techniques presented in Chapter 5. Together with the domain experts, we discussed different techniques for the data aggregation, the visual mapping, and the layout step of the reference workflow. We designed appropriate techniques to provide different levels of data abstraction and created visual-interactive means of formulating queries. Moreover, we discussed relation-seeking tasks including the data content and associated metadata $\mathbf{C}_{\text{C+M}}$. In consequence, we obtained two usable and useful ESS. Each ESS enables users to gain an overview of large collections of time-oriented primary data. Furthermore, users have the means to formulate queries in a visual-interactive way. Finally, the visualization of search results is provided, enriched with visual-interactive means for the exploration of search results from different perspectives.

Both design studies were accompanied by evaluation strategies from start. User studies, interviews and observations helped us to characterize the domain, the data and the analysis tasks appropriately, and to define requirements for the ESS. In the design phases of the two ESS, we used the reference workflow for the design and the application of visual-interactive interfaces for ESS (cf. Chapter 3). Our visual-interactive techniques for the construction of the workflow allowed the involvement of the domain experts. Various design decisions of both ESS were made in favor or user preferences or specific analysis tasks. In addition, we conducted usability studies with non-experts during the design phase to enhance the intuitiveness of the systems. As summative evaluation strategies, we performed field studies together with the domain experts to validate the usefulness of the approaches. We gathered qualitative and quantitative feedback to further improve the systems. In essence, the domain experts of both ESS attested the systems to be both usable and useful. Based on the field studies, the experts confirmed the efficiency of the system to support their day-to-day work, especially for large collections of time-oriented primary data. Analysis results within the field studies also presented new aspects of the data that had not been taken into consideration or had not been visually communicated before. To summarize, both the VisInfo and the MotionExplorer ESS support the effective reuse of time-oriented primary data.

CHAPTER 8

Thesis Conclusions and Future Challenges

Throughout this thesis, we presented concepts, guidelines, techniques, and systems for the Exploratory Search (ES) in time-oriented primary data. We focused on application areas involved in data-driven research. In essence, our contributions cover the entire time series processing and transformation workflow from raw time-oriented primary data to visual-interactive interfaces for time series analysis. These interfaces can subsequently be integrated into Exploratory Search Systems (ESS). We support visual search in time-oriented primary data with interactive queries by sketch and by example. Content-based overviews and novel techniques for seeking relations between data content and metadata create the exploration support provided in the course of this thesis. We introduced VA as a means of designing effective and efficient visual-interactive systems. Including VA in the workflow construction enables data scientists to choose appropriate models and model parameters. Moreover, visual access to intermediate steps of the workflow facilitates user involvement leading to usable and useful solutions. As a result, domain experts benefit from data analysis workflows with a high degree of both automation and quality. Likewise, the workflow automation in combination with user-centered design leads to visual interfaces for downstream ESS that can be kept simple and intuitive. Henceforth, data scientists and domain experts can utilize our contributions as a framework for building effective and efficient ESS for time-oriented primary data. By putting the ES concept into practice, we combine two information-seeking activities which beforehand were predominantly supported with specific tools. As a result, we bridge the gap between search systems that require exploration support and exploratory data analysis systems that require visual search support. In the two ESS presented in the case studies, we demonstrated that our techniques and systems support data-driven research in an efficient and effective way. In particular, we took a step forward to automate the hypotheses formulation and validation process. In combination with the user-centered definition of similarity and interestingness, we introduced a series of techniques that automatically reveal the most similar patterns and interesting relations autonomously.

This chapter starts with a summarization of the challenges for the ES in time-oriented primary data in Section 8.1. In Section 8.2, we recall the main results of this thesis. We outline future challenges in Section 8.3.

Contents

8.1. Summarization of Challenges

A careful reflection of the related work in Chapter 2 revealed six challenging factors which hampered ES in time-oriented primary data.

- $\mathbf{C}_{\text{MES}}$ Missing Methodology for the Design of Exploratory Search Systems (ESS)
- $\mathbf{C}_{\text{CBA}}$ Content-Based Access to Time-Oriented Primary Data
- $\mathbf{C}_{\text{CBO}}$ Gaining an Overview of the Data Content
- $\mathbf{C}_{\text{C+M}}$ Combining Data Content and Metadata
- $\mathbf{C}_{\text{MPC}}$ Model and Parameter Choice
- $\mathbf{C}_{\text{UCD}}$ Involving the User in the Design

We started with addressing methodological challenges $\mathbf{C}_{\text{MES}}$ in Chapter 3. Today, the number of best-practice examples of ESS is still scarce, particularly for non-textual data types. For time-oriented primary data virtually no best-practice ESS has been presented yet. While the wealth of the ES concept is beyond dispute, the lack of best-practice approaches and the resulting lack of reflection leaves a gap in methodology for the design of ESS. In addition, data scientists are confronted with a huge design space, formed by complex factors, such as time-oriented primary data, users engaged in data-centered research, and complex analysis tasks ranging from search to exploration. The involvement of concepts and techniques presented Information Visualization (IV) and Visual Analytics (VA) is a promising approach to enhance design of ESS. However, research in intersection points between ES and IV and VA was required to avoid reinventing the wheel in the design of ESS. We accepted challenge $\mathbf{C}_{\text{MES}}$ with the definition of five research goals which were accomplished in Chapter 3. Section 8.2.1 summarizes our respective contributions.

The challenge of content-based access to time-oriented primary data $\mathbf{C}_{\text{CBA}}$ was resolved in Chapter 4. It is widely accepted that the utilization of the data content is most promising to enhance the computer-supported search and exploration process. While research in the content-based access to textual documents had a pioneering role to a certain extent, content-based access is still an extensive subject of investigation for complex non-textual data types. For time-oriented primary data, content-based access remains a particular challenge especially for application fields providing search support, such as DLs. Similarly, gaining insight into the unknown structures of complex data is still a central research challenge. Other factors contributing to the challenge are the size of today's data collections, the heterogeneity, and the data quality. The data quality and the additional temporal dimension are particular challenges for time-oriented primary data. We confronted challenge $\mathbf{C}_{\text{CBA}}$ by proposing nine research goals which we achieved in Chapter 4. In Section 8.2.2, we summarize how our novel approach obtains these research goals.

Chapter 5 resolves the challenge of providing overviews of the data content $\mathbf{C}_{\text{CBO}}$. It is recognized that gaining an overview of the structural information is most desirable. However, providing content-based overviews for complex data is a challenging task. Influencing factors are the specific characteristics of the underlying data collections and the upstream content-based access strategy. At heart of the challenge is to condense the available body of information to a set of relevant representatives while neglecting less important information (data aggregation). To support the visual information-seeking process, the resulting data aggregates have to be represented visually (visual mapping). The arrangement of these visual aggregates in the display space in a meaningful way yet constitutes an additional challenge (layout). Finally, appropriate interaction designs and coupling content-based overviews with visual querying concepts contribute to the challenge. In Chapter $\mathbf{C}_{\text{CBO}}$, we divided challenge $\mathbf{C}_{\text{CBO}}$ into eight research questions, which we solved with our guidelines and techniques. Section 8.2.3 recalls these main results.

In Chapter 6, we faced the challenge of combining data content and metadata $\mathbf{C}_{\text{C+M}}$. Supporting this type of relation seeking in complex data is yet another promising but challenging task to facilitate ES. Assumed that upstream challenges, such as content-based access and content-based overviews, are approached, data scientists have to include metadata in the analysis process. The association of data content to metadata may lead to beneficial insights, e.g., for data-driven research. However, to reveal relations, data scientists at first have to provide a means for coping with these different types of data. In addition, the multitude of resulting relations has to be qualified to guide users towards the most interesting relations. Finally, meaningful visualization and interaction designs, as well as the integration of the interface into the targeted ESS, pose challenges for the data scientist. We divide challenge $\mathbf{C}_{\text{C+M}}$ into eight research goals which we achieve with each of three novel techniques presented in Chapter 6. In Section 8.2.4, we turn to detailed conclusions.

Finally, this thesis discussed two central problems hampering the design of usable and useful visual-interactive analysis techniques, i.e., choosing meaningful models and model parameters $\mathbf{C}_{\text{MPC}}$ and involving users in the design $\mathbf{C}_{\text{UCD}}$. These two overarching challenges set the course of all contributions in Chapters 3, 4, 5, and 6. The design of ESS for time-oriented primary data leads to a huge design space posed by this specific configuration of involved users, data, and tasks. The choice of meaningful algorithmic models in a meaningful order is a difficult task for virtually any data analysis approach $\mathbf{C}_{\text{MPC}}$. Similarly, the challenge affects any scientific workflow, e.g., created to facilitate data-driven science. Obviously the challenge scales with the complexity of the design space. Moreover, defining meaningful parameters for given models contributes to the difficulty of the problem. The involvement of users in the design $\mathbf{C}_{\text{UCD}}$ requires data scientists to grasp methodology from user-centered design and design studies. The design process should be performed in an iterative way. Furthermore, especially for data-driven approaches it is a challenge to carry out the process of designing complex workflows in a transparent and interactive way. A variety of design variables should be defined by users to comply with specific user needs, the definition of time series similarity may serve as an example. In the following conclusion section, we summarize the concepts, guidelines, techniques, and systems presented throughout this thesis. For each chapter, we highlight how we achieved the overarching goals of choosing meaningful models and parameters $\mathbf{C}_{\text{MPC}}$ and involving users in the design $\mathbf{C}_{\text{UCD}}$.

8.2. Conclusions

In this thesis, we outlined a set of research challenges which needed to be resolved to design usable and useful ESS for time-oriented primary data. Throughout thus thesis, we presented concepts, guidelines, techniques, and systems for the ES in time-oriented primary data. In the following subsections, we summarize how our contributions solve the research challenges of this thesis.

8.2.1. Concepts for Exploratory Search Systems

Research Goal	Description
$\mathbf{RG}_{\text{MES1}}$	Exploitation of Intersection Points Between ES and IV
$\mathbf{RG}_{\text{MES2}}$	Definition of ES for this Work
$\mathbf{RG}_{\text{MES3}}$	Adoption of Existing Methodology from Scientific Workflows, KDD, IV, and VA
$\mathbf{RG}_{\text{MES4}}$	Splitting the Workflow into Canonical Steps
$\mathbf{RG}_{\text{MES5}}$	Combining the Data Transformation and User-Centered Design Process

Table 8.1 *Research goals for the challenge* $\mathbf{C}_{MES}$ ***Methodology for the Design of Exploratory Search Systems***.

We presented the concepts of this thesis in Chapter 3. At a glance, we postulated two methodological contributions. First, we presented a *survey of IV and VA tasks* which we arranged to a single assembly. The assembly provides an overview of the rich set of tasks and techniques presented in IV and VA. In addition, we marked the assembly with labels for *Search*, *Search & Exploration*, and *Exploration*. In this way, data scientists are able to lookup appropriate tasks and techniques for the design process of usable and useful ESS. The task assembly also formed the basis for the definition of Exploratory Search for this thesis. Second, we presented a *reference workflow* for the design and the application of ESS for time-oriented primary data. The reference workflow adopts methodology from scientific workflows KDD, IV, and VA. Four main steps (models) describe the design process of visual-interactive views for ESS, each of the steps can be implemented with IV and VA techniques in a user-centered way. We employed these two contributions as a framework for the techniques and systems presented in the remainder of this thesis. The visual-interactive techniques presented in the Chapters 4, 5, and 6 reflect the four steps of the reference workflow. The design studies in Chapter 7 implement the entire pipeline presented in the workflow.

With our novel concepts, we resolved the challenge of missing methodology for the design of ESS $\mathbf{C}_{\text{MES}}$. We divided the challenge into five factors, each of which was mapped to a research goal (see Table 8.1). With the arrangement of tasks and techniques from IV and VA in a single assembly, we explicitly highlighted intersection points between ES, IV and VA $\mathbf{RG}_{\text{MES1}}$. Designers of ESS now have a condensed overview of existing tasks and techniques presented in IV and VA. Hence, we avoid reinventing the wheel for future ES implementations. In this connection, we also presented the definition of ES for this thesis $\mathbf{RG}_{\text{MES2}}$. The definition expands earlier conceptual characterizations of ES by a focus on mandatory visual-interactive interfaces. In addition to tasks and techniques, we reflected the methodological research produced in scientific workflows, KDD, IV, and VA. Our goal was to condense and adopt existing best-practice methodology for the creation of a reference workflow for visual-interactive interfaces for ESS $\mathbf{RG}_{\text{MES3}}$. As a result, the reference workflow presented in this thesis adopts the KDD process [FPS96], the Information Visualization Reference Model (Card Pipeline) [CMS99], the Visual Analytics Process (VA Diamond) [KAF*08], the framework for spatio-temporal data analysis [AA13], and the concept for designing VA frameworks and systems [ABM*07]. To cope with the wealth of existing data transformations, models, and design factors, we split the workflow into four consecutive steps (models) $\mathbf{RG}_{\text{MES4}}$. The division of the huge design space into modular parts enables data scientists to focus on specific and independent challenges. Each step (model) accepts a specific type of data, applies a set of analytical processes, and finally transforms the data into an output format. In this way, the complexity of the design space and the induced challenges are is dispensed. As an example, the challenge of providing content-based access to time-oriented primary data $\mathbf{C}_{\text{CBA}}$ is explicitly addressed to with the first step of the reference workflow, leading to a FV representation of the time-oriented primary data. Finally, a prime objective of the reference workflow is to support the involvement of users in the design process. To achieve this, the reference workflow combines the data transformation process with the user-centered design process $\mathbf{RG}_{\text{MES5}}$. In each of the four steps of the workflow, we included an adoption of the VA Diamond to support (a) visual-interactive interfaces for effective user steering and (b) algorithmic models for efficient data processing and transformation.

8.2.2. Visual-Interactive Preprocessing of Time-Oriented Primary Data

Research Goal	Description
$\mathbf{RG}_{\text{CBA1}}$	Representation of Time-oriented Primary Data
$\mathbf{RG}_{\text{CBA2}}$	Comparing Model Input with Model Output
$\mathbf{RG}_{\text{CBA3}}$	Choosing Appropriate Parameter Values
$\mathbf{RG}_{\text{CBA4}}$	Guaranteeing Data Quality
$\mathbf{RG}_{\text{CBA5}}$	Trade-off: Providing a Compact but Faithful Representation
$\mathbf{RG}_{\text{CBA6}}$	Supporting the Users' Notion of Similarity
$\mathbf{RG}_{\text{CBA7}}$	Generalizability of Workflow Configurations
$\mathbf{RG}_{\text{CBA8}}$	Reuse and Revision of Workflows
$\mathbf{RG}_{\text{CBA9}}$	Involvement of Domain Experts

Table 8.2 *Research goals for the challenge* $\mathbf{C}_{CBA}$ ***Content-Based Access to Time-Oriented Primary Data.***

We presented a novel approach for the visual-interactive creation of preprocessing workflows for time-oriented data in Chapter 4. The system enables users to transform time-oriented primary data into the feature space. In addition, the visual-interactive system allows the determination of distance measures for the definition of time series similarity. As a result, the FVs and the definitions of time series similarity can be applied in downstream steps of the reference workflow, e.g., to facilitate search and exploration support.

With the new system, we resolved the challenge of content-based access to time-oriented primary data $\mathbf{C}_{\text{CBA}}$. We divided challenge $\mathbf{C}_{\text{CBA}}$ into eight research goals (see Table 8.2), all of which were solved in the course of the chapter. The system provides visual representations for both time-oriented primary and preprocessing routines. Thus, users of the system are able to directly interact with representations of the raw input data $\mathbf{RG}_{\text{CBA1}}$ and with algorithmic models to be applied to the data. These visual representations of the time-oriented primary data are also used for showing intermediate results. In this way, we achieve the goal of providing a visual means for the input-output comparison of models in the preprocessing pipeline $\mathbf{RG}_{\text{CBA2}}$. In addition, the approach enables users to compare multiple model outputs in a bundle visualization for time-oriented primary data. As a result, the effect of different parameter values becomes comparable $\mathbf{RG}_{\text{CBA3}}$. To cope with diverse factors of dirty data, the system provides a large and yet extensible library of data cleansing routines. In combination with the model input-output comparison and the parameter support, the system provides solutions to improve the data quality $\mathbf{RG}_{\text{CBA4}}$. In the usage scenarios, we showed that data scientists and domain experts were able to identify quality leaks and develop meaningful cleansing strategies. The system provides the means of transforming time-oriented (primary) data into the FV space in a meaningful way. Time series descriptors can be selected, compared, and adjusted. Consequently, users can face the trade-off between compact but yet precise FV representations of raw data with visual-interactive means $\mathbf{RG}_{\text{CBA5}}$. To support the users' notion of similarity, the system allows defining distance measures. In combination with the preprocessing strategy, the definition of time series descriptors, and the choice of normalization strategies users have the means to factor their notion of similarity $\mathbf{RG}_{\text{CBA6}}$. A guidance concept supports users in the selection of testing data. The technique approximates the diversity of time-oriented data, and thus supports users in avoiding overfitting. Maximizing the diversity of the test set contributes to the construction of generalizable preprocessing workflows $\mathbf{RG}_{\text{CBA7}}$. The system supports the reuse and revision of preprocessing pipelines. Users are able to store and load existing workflows. The reuse facilitates the effective and efficient refinement of preprocessing pipeline for changing data (sources), as well as for the collaboration with other users $\mathbf{RG}_{\text{CBA8}}$. Another objective of the system was to support the active involvement of users to combine both human expert judgment and automated computation. In this way, the visual-interactive system opens the process of time series preprocessing for a broader audience. Former lacks of black-box approaches, such as requiring programming expertise and batch scripting, are overcome with meaningful visualization and interaction designs of the entire process. The abstraction from implementation details with visual-interactive interfaces supports collaborative working involving data scientists and domain experts. Parameter settings are now open for explanation and discourse. Different stakeholders have the opportunity to exchange knowledge and optimize the preprocessing pipeline. By assigning domain experts an active role in the construction of preprocessing pipelines, we achieved research goal $\mathbf{RG}_{\text{CBA9}}$.

8.2.3. Content-Based Overviews

Research Goal	Description
$\mathbf{RG}_{\text{CBO1}}$	Gaining an Understanding of the Underlying Data Set
$\mathbf{RG}_{\text{CBO2}}$	Choice of an Appropriate Clustering Algorithm
$\mathbf{RG}_{\text{CBO3}}$	Choice of Appropriate Model Parameters
$\mathbf{RG}_{\text{CBO4}}$	Visual Representation of High-Dimensional Data Elements and Clusters
$\mathbf{RG}_{\text{CBO5}}$	Choice of Layout Technique
$\mathbf{RG}_{\text{CBO6}}$	Different Levels of Abstraction
$\mathbf{RG}_{\text{CBO7}}$	Content-Based Querying
$\mathbf{RG}_{\text{CBO8}}$	Involving the User in the Design Process

Table 8.3 *Research goals for the challenge* $\mathbf{C}_{CBO}$ ***Gaining an Overview of the Data Content.***

In Chapter 5, we faced the challenge of gaining an overview of the data content $\mathbf{C}_{\text{CBO}}$. To resolve the challenge, we presented novel guidelines and techniques for the design of content-based overviews. Based on these guidelines and techniques, we presented visual-interactive interfaces for content-based overviews which can be employed in ESS. Our approach builds on a content-based access strategy, such as presented in Chapter 4. According to the reference workflow we divided the remaining design space into three steps. First, we showed how visual-interactive cluster analysis can facilitate the design process in the *data aggregation* step. Second, for the *visual mapping* step, we showed how high-dimensional data elements and data aggregates can be represented visually. Finally, we discussed how the *view transformation* step of the reference workflow can be facilitated with layouts for aggregated data.

At a more fine-gained level, we divided challenge $\mathbf{C}_{\text{CBO}}$ into eight parts, each of which defined a research goal to be addressed in the chapter (see Table 8.3). Research goal $\mathbf{RG}_{\text{CBO1}}$ exposed that data scientists (and domain experts) need to gain an understanding of the underlying data set, e.g., the time-oriented primary data collection. It was our particular concern that this understanding of the data set can already be gained within the design phase. For this purpose, we exploited the complementary strengths of different factors, such as visual-interactive data analysis techniques, user-centered design, and cluster quality assessment, enabling data scientists and domain experts to gain a profound understanding of complex data sets. With the research goal $\mathbf{RG}_{\text{CBO2}}$, we emphasized the challenge of choosing appropriate clustering algorithms in the data aggregation step. Closely related is research goal $\mathbf{RG}_{\text{CBO3}}$, responding to the challenge of choosing meaningful parameters for a given algorithmic model. To achieve these goals, we presented novel visual-interactive techniques for quality-driven cluster analysis. We divided the problem by introducing visual-interactive quality assessment strategies on four levels of granularity. Coarsely speaking, we introduced visual quality assessment for 1) the global clustering result, 2) any given cluster of a clustering result, 3) any single data point of the underlying data set, and 4) the comparison of different clustering results of different clustering algorithms. Research goal $\mathbf{RG}_{\text{CBO4}}$ depicted challenges in visually representing high-dimensional objects. For content-based overviews, we particularly emphasized the need for visual representations including both the cluster information and the information of associated data elements. We postulated a guideline with five mandatory factors and presented eleven examples demonstrating how we addressed this guideline in our previous works in a variety of real-world case studies. Research goal $\mathbf{RG}_{\text{CBO5}}$ addressed the challenge of choosing appropriate layouts for aggregated data. We presented guidelines and techniques showing how meaningful layouts can be designed. At a core level, we distinguished between 1) layouts exploiting the structure of a given clustering result (such as hierarchies, or networks), 2) projection-based layouts (including linear and non-linear techniques), and 3) force-directed layouts (with various attracting and repulsing forces). In this connection, we emphasized the important role of the user in the design process and showed how user involvement can enhance the usefulness of the solutions. With research goal $\mathbf{RG}_{\text{CBO6}}$, we pointed out that usable and useful content-based overview solutions should provide different levels of abstraction to support different information-seeking behaviors. We presented four different strategies for providing different levels of abstraction for content-based overviews. For each of the four strategies, we presented solutions associated with the two case studies of this thesis (cf. Chapter 7). Research goal $\mathbf{RG}_{\text{CBO7}}$ addressed visual querying, based on content-based overviews. We showed how the Query-by-Example strategy can be utilized for single patterns (clusters, features, data elements) and temporal sequences (subsequences). Again, we refer to the two case studies where the applicability of visual querying time-oriented primary data is proven (cf. Chapter 7). Research goal $\mathbf{RG}_{\text{CBO8}}$ emphasized the challenge of involving users in the design process. To this aim, we consider the active engagement of

users in all three steps of the reference workflow. In the data aggregation step our visual-interactive techniques directly allow the involvement of the user, e.g., to understand the underlying data, or to choose meaningful models and model parameters. For the visual mapping step, our guidelines and techniques recommend user-centered design principles to generate meaningful glyphs and an appropriate choices of similarity-preserving colormaps. For the layout step, we showed how the user can be involved in the process of choosing a meaningful layout strategy, leading to usable and useful layouts for content-based overviews.

8.2.4. Relation Seeking Between Data Content and Metadata

Research Goal	Description
$\mathbf{RG}_{\text{C+M1}}$	Relating Different Types of Data
$\mathbf{RG}_{\text{C+M2}}$	Assessing the Interestingness of Relations
$\mathbf{RG}_{\text{C+M3}}$	Multiple Granularity Problem
$\mathbf{RG}_{\text{C+M4}}$	Overview of the Relation Space
$\mathbf{RG}_{\text{C+M5}}$	Guiding Users Towards Interesting Relations
$\mathbf{RG}_{\text{C+M6}}$	Interaction Design
$\mathbf{RG}_{\text{C+M7}}$	Involving the User in the Design Process
$\mathbf{RG}_{\text{C+M8}}$	Visual Communication of Interesting Relations

Table 8.4 *Research goals for the challenge* $\mathbf{C}_{C+M}$ ***Challenges in Combining Data Content and Metadata***.

In Chapter 6, we discussed solutions combining the analysis of data content and metadata in a single visual interface. We presented three different techniques supporting users in seeking different types of relations between data content and metadata. In addition, we presented different methods for the definition interestingness for different types of relations. On this basis, we introduced visual encodings guiding the users towards interesting relations, visualizations showing overviews of the data and the relations, and interaction designs enabling users to seek interesting relations in an efficient and effective way. Hence, carrying out tedious batch processes for single tests has become obsolete. As a consequence, hypotheses can be derived and validated more efficiently and effectively. To summarize, our techniques mitigate the problem of defining and testing hypotheses to the definition of meaningful relations and interestingness measures. Consequently, the approaches are particularly useful in an ES context when users aim at browsing large data spaces for which the a-priori knowledge is rare. Moreover, all implementations provide an individual perspective of the data which qualifies the presented techniques for multiply linked view systems. Apart from that, ESS can benefit from the techniques in other fields. As an example, faceted search support can greatly benefit from the presented metadata layouts.

The first technique presented in Section 6.3 assigns metadata onto the data content. The data content is structured in a content-based overview layout and serves as the targeted variable for relation seeking. Metadata attributes represent independent variables which are subsequently mapped onto the content-based overview. Users can seek interesting metadata attributes and their relations with the clusters of the content-based overview. The second technique in Section 6.4 follows the opposite strategy. We present a novel approach showing how the data content can be mapped onto the layout of a metadata attribute. The technique combines concepts of a) faceted search in metadata and b) VA for the content-based exploration of time-oriented data. Our approach is novel in that it derives a notion of similarity for metadata entities from similarity measures applied to the associated data content. Dependent on the metadata layout, different subsets of the data content are associated with individual metadata entities. The third technical contribution in Section 6.5 abstracts from the differentiation between data content and metadata. In fact, the technique unifies the complete body of information to a multi-attributed mixed data set. The technique identifies the most interesting (statistically dependent) relations between the mixed data attributes, and exploits this information for the creation of layouts. With this novel technique, users are able to identify the most interesting relations in a multi-attributed data set at a glance.

We challenged each of the three techniques to solve eight research goals, summarized in Table 8.4. We defined these research goals in Chapter 6 as prerequisites for facilitating relation seeking between data content and metadata with visual-interactive interfaces in a meaningful way. At heart of most relation-seeking tasks is the meaningful definition of a relation $\mathbf{RG}_{\text{C+M1}}$. In this thesis the combination of data content and metadata was in the focus of interest.

The first technique maps distributions of metadata onto clusters of the data content. The second technique maps clusters of the data content onto the entities of metadata attributes. The third technique focuses on relations between data subsets defined by the entities of any given attribute. For each of the three techniques, we presented individual functional definitions of interestingness $\mathbf{RG}_{C+M2}$. This enabled us to the (relative) interestingness of any given relation. While the presented functional definitions constitute a starting point, we also paid attention to exchangeability of interestingness measures. Interesting relations may exist between data subsets (bins, clusters) or between entire data attributes $\mathbf{RG}_{C+M3}$. This is why our techniques achieve the research goal to reveal relations at both levels of granularity. Another challenging goal was to provide an overview of the relation space in each technique $\mathbf{RG}_{C+M4}$. In this way, users are empowered to reveal structural information about the data set and respective relations. For this purpose, we made use of the contributions presented in Chapter 5 on content-based overviews and structure-preserving layouts. In addition, we presented different visual means to include the relation space on top of the data overview visualizations. We presented additional visual encodings to highlight most interesting relations in each of the techniques. These encodings support users in the identification of most interesting relations and avoid getting lost in the huge relation space $\mathbf{RG}_{C+M5}$. Additional interaction designs enable users to drill down to local aspects of interest, or to filter the set of displayed information by different criteria $\mathbf{RG}_{C+M6}$. For each of the three techniques, we exposed several design parameters. These design parameters are a beneficial way to involve users in the design process and to individualize our techniques by the special user needs $\mathbf{RG}_{C+M7}$. As a result, some of the models and model parameters may already be fixed in the design phase, reducing the complexity of the resulting interactive solutions. Furthermore, the user-centered design approach also regards the identification of models and model parameters relevant for the user as steering parameters in the application phase of our techniques. With the combinations of overviews, interestingness measures, guidance concepts, and interaction techniques our implementations support the communication of revealed relations $\mathbf{RG}_{C+M8}$. The user benefits from this capability by (1) visually preserving gained insight for a future application of the system (provenance), (2) being able to exchange insights with other users peer-to-peer (collaboration), or (3) being able to publish the insights in respective scientific literature visually (publication).

8.2.5. Case Studies — Exploratory Search Systems

In Chapter 7, we put our concepts, guidelines, and techniques into practice. Both VisInfo in Section 7.1 and the MotionExplorer system in Section 7.2 employ time-oriented primary data and make it visual-interactively accessible. Domain experts have the means to use (reuse) large collections of data by applying ES tasks with different visual-interactive means. Both ESS provide content-based overviews enabling users to reveal structural information of the data collection. The visual-interactive specification queries supports users to search for specific information in the data content. Similarly, both ESS support the information drill-down and provide details on demand. Finally, the visual-interactive exploration of search results enables users to compare search results and to seek relations between the data content and additional metadata. In each of the case studies, we collaborated with domain experts to design usable and useful visual interfaces. The summative field studies conducted in for both ESS confirmed the efficiency and the effectiveness of the systems.

The two case studies also demonstrate the applicability of our concepts, guidelines and techniques presented for the ES in time-oriented primary data. To the best of our knowledge, VisInfo and MotionExplorer are among the very first ESS for time-oriented primary data. To achieve this goal, we applied our novel concepts presented in Chapter 3 in both case studies. In addition, we used the guidelines and techniques to provide content-based access, content-based overviews, and views for the combined analysis of data content and metadata (cf. Chapters 4, 5, and 6). In this connection, the two case studies also provide different means of resolving the six main research challenges of this thesis. The summary of the case study section describes how both ESS resolve the six main research challenges in detail (cf. Section 7.3).

We conclude the summary with a brief reflection of the key factors addressed in the two case studies. Table 8.5 provides a condensed overview. The order of occurrence basically reflects the chronological steps carried out in the case studies. For the sake of simplification, we abstract from the iterative nature of the design process. Both case studies were conducted with principles from design study methodology and user-centered design. Based on characterizations of domain, data, and users' tasks $\mathbf{F}_1$, we identified analytical challenges of the involved domains, and put these into functional system requirements $\mathbf{F}_2$. In the VisInfo ESS, we characterized the domains of Earth observation and DLs, the MotionExplorer system targeted the application of human motion capturing data analysis and synthesis. In both design studies, we identified central challenges in the data life-cycle and in the construction of scientific workflows. In essence, content-based access, the abstraction of data, and the need for visual-interactive means of searching and exploring large data collections affected the requirements. For the VisInfo ESS, our novel visual-interactive system

Factor	Description
$\mathbf{F}_1$	Characterization of the Domain, Data, and User Tasks
$\mathbf{F}_2$	Requirements
$\mathbf{F}_3$	Time Series Preprocessing and Similarity Definition
$\mathbf{F}_4$	Content-Based Overviews of Time-Oriented Data
$\mathbf{F}_5$	Visual Interactive Querying and Search Result Exploration
$\mathbf{F}_6$	Identification of Relations Between the Time-Oriented Data Content and Metadata
$\mathbf{F}_7$	Evaluation and Application

Table 8.5 *Key factors of the two case studies.*

to preprocess time-oriented primary data (cf. Chapter 4) particularly solved the challenge to gain access to the data content $\mathbf{F}_3$. For the MotionExplorer system, we extended the preprocessing library for the application of multivariate time series. In both case studies, we collaborated with the domain experts when we constructed the preprocessing workflows, leading to meaningful FV representations of the time-oriented primary data. In an analogous manner, we defined similarity functions reflecting the experts' notion of time series similarity. The four remaining key factors of the case studies ($\mathbf{F}_4$-$\mathbf{F}_7$) were addressed as follows.

VisInfo The VisInfo ESS uses a SOM-based data aggregation solution $\mathbf{F}_4$. We optimized the SOM parameters with our novel techniques for the quality-driven visual-interactive cluster analysis in Section 5.3. In several iterations, we designed a cluster glyph for daily curve patterns (cf. Section 5.4). A click on a single cluster in the web-based application exposes a Details-on-Demand view where the cluster glyph is enriched with associated data elements. The layout of the clusters in the display space is provided with the 2D output of the SOM algorithm. Referring to Section 5.5.1 this equates to a cluster structure-based solution. Overview+context is implemented as a means for different levels of abstraction (cf. Section 5.5.2). Taking the visual-interactive formulation of queries into account $\mathbf{F}_5$, VisInfo provides both the Query-by-Sketch and the Query-by-Example technique (cf. Section 5.5.3). The visualization of search results makes use of the techniques for combining data content and metadata (cf. Chapter 6). A faceted search interface is combined with the visualization of the retrieved data content and multiple views showing interesting metadata attributes $\mathbf{F}_6$. In different evaluation strategies and application examples, we proved the usability and the usefulness of VisInfo (cf. Section 7.1.7) $\mathbf{F}_7$.

MotionExplorer The MotionExplorer ESS provides two content-based overviews $\mathbf{F}_4$. The first is based on a steerable hierarchical clustering algorithm which aggregates human poses. A tree visualization represents the hierarchy of pose clusters in the ESS. Thus, we applied a cluster structure-based layout of the data aggregates (cf. Section 5.5.1). In several iterations, we designed a cluster glyph for human motion. We used a stick figure metaphor to represent clusters and associated data elements visually (cf. Section 5.4). A circular outline represents the pairwise similarity between cluster by using a 2D colormap. The second content-based overview depicts both human poses and motion. To this end, we applied a projection-based layout preserving the pairwise similarities of the cluster vector information in the display space (cf. Section 5.5.1). Users can interactively remove local overplotting by means of a force-directed layout algorithm applied to the output of the projection-based layout (see again Section 5.5.1). Together with the domain experts, we decided to make the hierarchical clustering algorithm interactively steerable. In this way, domain experts are empowered to adapt the level of data abstraction 'at runtime' (cf. Section 5.5.2). MotionExplorer supports the formulation of visual queries, the Query-by-Example technique is provided $\mathbf{F}_5$. Users can query motion sequences by selecting start and end poses (cf. Section 5.5.3). In a search-result visualization, users are able to explore micro variations of the retrieved motion subsequences. Together with the domain experts, we decided to neglect relation seeking between data content and metadata $\mathbf{F}_6$. In turn, we emphasized relation-seeking activity within the data content (human poses) by linking all seven views of the ESS with similarity-preserving colors. Finally, we presented different usage scenarios and the results of a summative field study conducted with the domain experts to assess the applicability of MotionExplorer $\mathbf{F}_7$. The domain experts unanimously confirm that the system can efficiently support their day-to-day work.

8.3. Future Challenges

In this thesis, we carried out research in concepts, techniques and systems to facilitate the ES in time-oriented primary data. Our contributions offer new research potential for future investigation. In the following, we turn to possible future challenges drawing on the results of this thesis.

Involvement of other Data, Users, and User Tasks It is obvious that the huge design space investigated in this thesis allows many specific solutions. Other users, data, and information-seeking tasks yield other promising configurations for future ES approaches, each requiring a careful problem characterization and implementation of required features. Data scientists can benefit from our concepts, guidelines, and techniques, to generate usable and useful ESS in an effective and efficient way. It will be an interesting future work to compare future ESS approaches and to reflect on individual lessens learned.

Tasks Depending on the information-seeking behavior of the involved user group, future ESS may balance visual-interactive means for search or exploration differently. Search-related fields, such as DLs, or research in IR, may further extend search capability with novel exploratory means, such as content-based overviews (cf. Chapter 5), or interfaces for seeking relations between data content and metadata (cf. Chapter 6). In turn, exploratory data analysis approaches may integrate search functionality into existing exploration systems, such as content-based visual query interfaces (cf. Chapter 5). Throughout this thesis, we followed the approach of a uniform distribution of search and exploration support. With our ESS, users are able to alternate between search or exploration activity virtually at any time of the information-seeking process. Since ES is still a young discipline, it will be interesting to develop other visual-interactive systems with a different emphasis on search and exploration activity.

Users We emphasized the applicability of our solutions for data-driven research. Beyond Earth observation, DLs, and human motion analysis, a variety of other research fields exist which would be worth to collaborate. In addition, many application fields beyond data-driven research can benefit from the solutions presented in this thesis. Example applications are decision support, financial analysis, power network analysis, logistics, service network analysis, and process optimization. One research challenge addresses possible adoptions of our concepts and techniques, necessary to ease the usefulness and the usability of ESS beyond the case studies presented in this thesis. In essence, it will be very interesting to collaborate with other user groups and assess the generalizability of the presented solutions.

Data We emphasized time-oriented primary data as the specific data type of this thesis. Time-oriented primary data yields particular challenges, many of which we resolved in the course of this thesis. One subject to future work will be to abstract our solutions and make them applicable for other data types. In particular, the *content-based access* by means of other complex data types $\mathbf{C}_{\text{CBA}}$ will pose new challenges. While the generation of other powerful FVs is one part of the challenge, other data content also influences the visual representations provided in ESS. The *visual access* to other types of data content yet constitutes another future challenge [vLSFK12].

Collaborative Information Seeking In this thesis, we emphasized the need for collaborations between data scientists and domain experts in the *design* phase, to generate usable and useful tools. The design study methodology in VA provides profound guidelines which we adopted towards the data and tasks used in this thesis. Beyond the design of ESS, collaboration can also be a means of the *application* of respective systems. ESS, including powerful visual-interactive interfaces, are a beneficial platform to include interfaces that allow collaborative information seeking and decision making. While search is often a solitary activity, many search tasks may involve multiple individuals [WR09]. In the latter case, the information space can be explored collaboratively. Multiple individuals in the collaboration would be part of shared learning approaches. Different expertises and competences can be conflated to solve complex tasks effectively. Our novel techniques, enabling domain experts to visually communicate their results of search and exploration tasks, can be seen as a step towards more collaborative approaches. For instance, the visual representation of results can directly be used in research publications. In future, our approaches can be extended by annotation, rating and referencing schemes to improve the exchange of data among scientists. This may include recommender functionality correlating, e.g., user profiles, or search session. In essence, future challenges to facilitate enhanced collaboration are manifold. First, dividing complex tasks has the potential to hinder learning for individual team members [WR09]. Second, collaborative systems will require different means of summarizing insight for sharing. This objective is related to the emerging research in gathering provenance information conducted in various fields [BCS*05, LAB*06, DF08, The12]. Third, the different expertises of the individuals may require different visual representations of the same information. The line of approach to keep search interfaces simple [Hea09] may contribute to this challenge. Finally, collaborative ESS have to face various technical challenges, such as the

participation of different distributed users, shared visualizations and interactions, and the synchronization of data flows and multimedia communication in real-time. It would be interesting to assess the applicability of our guidelines and techniques for the design and application of collaborative information seeking approaches.

Relation Seeking and Interestingness Measures Users with exploratory information-seeking behavior are typically interested in different types of relations in the data. Relations can be expected between clusters (data subsets), between single data elements, or between metadata. Obviously, relations may also exist between combinations of the latter types of data. As an example, for data-driven research relations between data content and metadata are of particular interest. This is why our novel techniques include three different types of relations between data content and metadata. However, we cannot guarantee that our techniques enable users to identify all relations of a data set in an efficient and effective way. Many data-centered research challenges may be based on specific considerations and requirements [Shn02]. The trade-off between the generalizability and the specificity of relations must be taken into account. In addition, the definition of a relation may be subject to change in the course of the analysis process.

Another important factor is the data's perspective, irrespective of user needs. The actual existence of different types of intrinsic relations in the data still requires further investigation. It is common ground that what users want to see is not necessarily what the data want to be [PVW09]. Needless to mention is the importance of addressing users' requirements. Besides, another important goal is to reveal virtually any type of existing relation hidden in the data autonomously. However, this endeavor is especially challenging for complex and unknown data collections.

For both challenges a promising solution would be an interactive editor for the efficient specification and adaption of relations. Similarly, providing sets of useful interestingness measures for any formal definition of a relation would be beneficial. Instant feedback from the system may guide the analyst towards most promising definitions of relations and interestingness measures. These specifications may even determine which of our presented concepts are most appropriate for the information-seeking behavior of the user. The latter may be solved with a wizard functionality estimating the user need and subsequently adapts the model selection for relations and interestingness measures. Our reference workflow (cf. Figure 3.1) can serve as a framework for construction of such workflows.

Similarity We take a closer look at an elementary question in many fields related to data analysis - the question what makes two objects similar? Similarity measures are a salient criterion for a vast amount of (unsupervised) DM, ML, and IR techniques. The most prominent example employed throughout this thesis is cluster analysis. Besides, definitions of similarity form the baseline functionality for the application of nearest neighbor searches. Furthermore, similarity is relevant for the use of color as a similarity-preserving visual variable. Finally, many layouts strategies arrange complex objects in 2D in a similarity-preserving way, depending on a functional specification of proximity. In this thesis, we defined time series similarity with a visual-interactive system. However, it currently does not support the assessment of effects of different similarity measures. A possible future work approach is a visual-interactive guidance concept for the effects of different similarity measures on time-oriented data. To our knowledge, this has not been a subject of VA approaches for time-oriented data. Future work includes extending our approaches to further domain-specific similarity notions in time-oriented data, such as similarity based on time series motif analysis, or the correlation between measurements.

Beyond time-oriented data, the research challenge of defining meaningful similarity measures can be raised to a new level, especially for complex data objects. While various similarity measures for numerical vector representations exist (FVs - in combination with preprocessing strategies and descriptors), assessing the similarity of complex real-world objects with multiple mixed attributes remains a future challenge in general. It would be very useful to incorporate the feedback from domain experts for the assessment of similarity in a visual-interactive way. Today, supervised learning and active learning approaches enable data scientists to understand complex phenomena, a concept which can also be applied to learning similarity functions for complex data objects. In a series of conceptual approaches, we postulated visual-interactive learning strategies to assess the similarity of complex objects by learning user feedback [BSR*14, BSB*14, SBKK14]. In a first implementation, we learned the similarity of countries on Earth. Here, the countries objects were composed of multiple mixed attributes, such as, the population, the dominating religion, the national language, or the gross domestic product. Users were able to arrange countries in a 2D display to express their individual notions of similarity. In turn, the learning system interpreted the arrangement of the objects, and learned a functional representation of the similarity. This concept and first implementation only constitutes a starting point for more extensive research in this topic. In turn, the results of such approaches can be applied to validate and optimize the similarity functions applied in this thesis.

Publications and Talks

The thesis is partially based on the following publications and talks:

Publications

Journal Publications

1. **BERNARD J.**, SESSLER D., MAY T., SCHLOMM T., PEHRKE D., KOHLHAMMER J.: A visual-interactive system for prostate cancer cohort analysis. *Computer Graphics and Applications (CG&A), IEEE 35*, 3, (2015), 44–55. doi:10.1109/MCG.2015.49.
2. **BERNARD J.**, DABERKOW D., FELLNER D., FISCHER K., KOEPLER O., KOHLHAMMER J., RUNNWERTH M., RUPPERT T., SCHRECK T., SENS I.: VisInfo: a digital library system for time series research data based on exploratory search - a user-centered design approach. *International Journal on Digital Libraries (IJoDL) 16*, 1, (2015), 37–59. doi:10.1007/s00799-014-0134-y.
3. STEIGER M., **BERNARD J.**, MITTELSTÄDT S., LÜCKE-TIEKE H., KEIM D. A., MAY T., KOHLHAMMER J.: Visual Analysis of Time-Series Similarities for Anomaly Detection in Sensor Networks. *Computer Graphics Forum (CGF) 33*, 3, (2014), 401–410. doi:10.1111/cgf.12396.
4. **BERNARD J.**, STEIGER M., WIDMER S., LÜCKE-TIEKE H., MAY T., KOHLHAMMER J.: Visual-interactive Exploration of Interesting Multivariate Relations in Mixed Research Data Sets. *Computer Graphics Forum (CGF) 33*, 3, (2014), 291–300. doi:10.1111/cgf.12385.
5. **BERNARD J.**, WILHELM N., KRUGER B., MAY T., SCHRECK T., KOHLHAMMER J.: Motionexplorer: Exploratory search in human motion capture data based on hierarchical aggregation. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 12, (2013), 2257–2266. doi:10.1109/TVCG.2013.178.
6. **BERNARD J.**, WILHELM N., SCHERER M., MAY T., SCHRECK T.: Timeseriespaths: Projection-based explorative analysis of multivariate time series data. *Journal of WSCG 20*, 2, (2012), 97–106.
7. **BERNARD J.**, BRASE J., FELLNER D., KOEPLER O., KOHLHAMMER J., RUPPERT T., SCHRECK T., SENS I.: A visual digital library approach for time-oriented scientific primary data. *Springer International Journal of Digital Libraries (IJoDL), ECDL 2010 Special Issue 11*, 2, (2011), 111–123. doi:10.1007/978-3-642-15464-5_35.
8. SCHRECK T., **BERNARD J.**, VON LANDESBERGER T., KOHLHAMMER J.: Visual cluster analysis of trajectory data with interactive kohonen maps. *Information Visualization 8*, 1, (2009), 14–29. doi:10.1057/ivs.2008.29.

Papers

1. **BERNARD J.**, STEIGER M., MITTELSTÄDT S., THUM S., KEIM D., KOHLHAMMER J.: A survey and task-based quality assessment of static 2D colormaps. In *SPIE, Visualization and Data Analysis (VDA)*, (2015), vol. 9397, pp. 93970M–93970M–16. doi:10.1117/12.2079841.
2. WILHELM N., VÖGELE A., ZSOLDOS R., LICKA T., KRÜGER B., **BERNARD J.**: Furyexplorer: visual-interactive exploration of horse motion capture data. In *SPIE, Visualization and Data Analysis (VDA)*, (2015), pp. 93970F–93970F–15. doi:10.1117/12.2080001.

3. **BERNARD J.**, SESSLER D., RUPPERT T., DAVEY J., KUIJPER A., KOHLHAMMER J.: User-based visual-interactive similarity definition for mixed data objects-concept and first implementation. *Journal of WSCG 22*, (2014).
4. RUPPERT T., **BERNARD J.**, MAY T., KOHLHAMMER J.: Combining computational models and interactive visualization to support rational decision making. In *Advances in Visual Computing (ISVC'13)*, vol. 8887 of *Lecture Notes in Computer Science*. Springer International Publishing, (2014), pp. 345–356. doi:10.1007/978-3-319-14249-4_33.
5. RUPPERT T., **BERNARD J.**, ULMER A., LÜCKE-TIEKE H., KOHLHAMMER J.: Visual access to an agent-based simulation model to support political decision making. In *International Conference on Knowledge Technologies and Data-driven Business (i-KNOW)*, (2014), ACM, pp. 16:1–16:8. doi:10.1145/2637748.2638410.
6. STEIGER M., **BERNARD J.**, MAY T., KOHLHAMMER J.: A survey of direction-preserving layout strategies. In *Spring Conference on Computer Graphics (SCCG)*, (2014), ACM, pp. 21–28. doi:10.1145/2643188.2643189.
7. NAZEMI K., RETZ R., **BERNARD J.**, KOHLHAMMER J., FELLNER D.: Adaptive semantic visualization for bibliographic entries. In *Advances in Visual Computing (ISVC)*, (2013), vol. 8034 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2013, pp. 13–24. doi:10.1007/978-3-642-41939-3_2.
8. RUPPERT T., **BERNARD J.**, ULMER A., KUIJPER A., KOHLHAMMER J.: Visual access to optimization problems in strategic environmental assessment. In *Advances in Visual Computing (ISVC)*. (2013), vol. 8034 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 361–372. doi:10.1007/978-3-642-41939-3_35.
9. RUPPERT T., **BERNARD J.**, KOHLHAMMER J.: Bridging knowledge gaps in policy analysis with information visualization. In *Conference on Electronic Government (EGOV/ePart Ongoing Research)* (2013), vol. 221 of *LNI*, GI, pp. 92–103.
10. **BERNARD J.**, RUPPERT T., GOROLL O., MAY T., KOHLHAMMER J.: Visual-interactive preprocessing of time series data. In *SIGRAD*, (2012), vol. 81 of *Linköping Electronic Conference Proceedings*, Linköping University Electronic Press, pp. 39–48.
11. **BERNARD J.**, RUPPERT T., SCHERER M., KOHLHAMMER J., SCHRECK T.: Content-based layouts for exploratory metadata search in scientific research data. In *ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, (2012), ACM, pp. 139–148. doi:10.1145/2232817.2232844.
12. **BERNARD J.**, RUPPERT T., SCHERER M., SCHRECK T., KOHLHAMMER J.: Guided discovery of interesting relationships between time series clusters and metadata properties. In *International Conference on Knowledge Management and Knowledge Technologies (i-KNOW)*, (2012), ACM, pp. 22:1–22:8. doi:10.1145/2362456.2362485.
13. **BERNARD J.**, VON LANDESBERGER T., BREMM S., SCHRECK T.: Multiscale visual quality assessment for cluster analysis with Self-Organizing Maps. In *IS&T/SPIE Conference on Visualization and Data Analysis (VDA)*, (2011), SPIE Press, pp. 78680N.1 – 78680N.12. doi:10.1117/12.872545.
14. BREMM S., VON LANDESBERGER T., **BERNARD J.**, SCHRECK T.: Assisted descriptor selection based on visual comparative data analysis. In *IEEE - VGTC Conference on Visualization (EuroVis)*, (2011), Eurographics Association, pp. 891–900. doi:10.1111/j.1467-8659.2011.01938.x.
15. SCHERER M., **BERNARD J.**, SCHRECK T.: Retrieval and exploratory search in multivariate research data repositories using regressional features. In *ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, (2011), ACM, pp. 363–372. doi:10.1145/1998076.1998144.
16. **BERNARD J.**, BRASE J., FELLNER D., KOEPLER O., KOHLHAMMER J., RUPPERT T., SCHRECK T., SENS I.: A visual digital library approach for time-oriented research data. In *European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, (2010), Springer-Verlag, pp. 352–363. doi:10.1007/978-3-642-15464-5_35.
17. SCHRECK T., **BERNARD J.**, TEKUŠOVÁ T., KOHLHAMMER J.: Visual cluster analysis in trajectory data using editable Kohonen maps. In *IEEE Symposium on Visual Analytics Science and Technology (VAST)*, (2008), IEEE Computer Society, pp. 3–10.

Short Papers, Posters

1. **BERNARD J.**, SESSLER D., BANNACH A., MAY T., KOHLHAMMER J.: A Visual Active Learning System for the Assessment of Patient Well-Being in Prostate Cancer Research. In *IEEE Vis Workshop on Visual Analytics in Healthcare (VAHC)*, (2015), IEEE Computer Society. doi:10.1145/2836034.2836035.
2. GSCHWANDTNER T., SCHUMAN H., **BERNARD J.**, MAY T., BÖGL M., MIKSCH S., KOHLHAMMER J., RÖHLIG M., ALSALLAKH B.: Enhancing time series segmentation and labeling through the knowledge generation model. In *Eurographics Conference on Visualization (EuroVis) (Poster)*, (2015), Eurographics Association.
3. KOLDIJK S., **BERNARD J.**, RUPPERT T., KOHLHAMMER J., NEERINCX M., KRAAIJ W.: Visual Analytics of Work Behavior Data - Insights on Individual Differences. In *Eurographics Conference on Visualization (EuroVis) (Short Paper)*, (2015), Eurographics Association. doi:10.2312/eurovisshort.20151129.
4. RUPPERT T., **BERNARD J.**, LÜCKE-TIEKE H., MAY T., KOHLHAMMER J.: Visual-Interactive Text Analysis to Support Political Decision Making - From Sentiments to Arguments to Policies. In *EuroVis Workshop on Visual Analytics (EuroVA)*, (2015), The Eurographics Association. doi:10.2312/eurova.20151101.
5. STEIGER M., **BERNARD J.**, SCHADER P., KOHLHAMMER J.: Visual Analysis of Relations in Attributed Time-Series Data. In *EuroVis Workshop on Visual Analytics (EuroVA)*, (2015), Eurographics Association. doi:10.2312/eurova.20151105.
6. **BERNARD J.**, SESSLER D., BERISCH M., HUTTER M., SCHRECK T., KOHLHAMMER J.: Towards a User-Defined Visual-Interactive Definition of Similarity Functions for Mixed Data. In *IEEE Symposium on Visual Analytics Science and Technology (VAST) (Poster)*, (2014).
7. **BERNARD J.**, SESSLER D., MAY T., SCHLOMM T., PEHRKE D., KOHLHAMMER J.: A visual-interactive system for prostate cancer stratifications. In *IEEE Vis Workshop on Visualizing Electronic Health Record Data (EHRVis)*, (2014), IEEE Computer Society.
8. HUTTER M., STEIGER M., **BERNARD J.**, ZURLOH C., KOHLHAMMER J.: Interactive multi-criteria optimization of 2d color maps. In *Vision, Modelling and Visualization (VMV)*, (2014).
9. MITTELSTÄDT S., **BERNARD J.**, SCHRECK T., STEIGER M., KOHLHAMMER J., KEIM D. A.: Revisiting Perceptually Optimized Color Mapping for High-Dimensional Data Analysis. In *Eurographics Conference on Visualization (EuroVis) (Short Paper)*, (2014), 91–95. doi:10.2312/eurovisshort.20141163.
10. RUPPERT T., **BERNARD J.**, LÜCKE-TIEKE H., KOHLHAMMER J.: Towards a Tighter Coupling of Visualization and Public Policy Making . In *IEEE Symposium on Visual Analytics Science and Technology (VAST) (Poster)*, (2014).
11. SESSLER D., **BERNARD J.**, KUIJPER A., KOHLHAMMER J.: Adopting mental similarity notions of categorical data objects to algorithmic similarity functions. In *Vision, Modelling and Visualization (VMV)*, (2014). Poster Paper.
12. SCHRECK T., SHARALIEVA L., WANNER F., **BERNARD J.**, RUPPERT T., VON LANDESBERGER T., BUSTOS B.: Visual Exploration of Local Interest Points in Sets of Time Series. In *IEEE Symposium on Visual Analytics Science and Technology (VAST) (Poster)*, (2012).
13. VON LANDESBERGER T., BREMM S., **BERNARD J.**, SCHRECK T.: Smart query definition for content-based search in large sets of graphs. In *International Symposium on Visual Analytics Science and Technology (Short Paper)*, (2010), Eurographics Association, pp. 7–12.
14. **BERNARD J.**, VON LANDESBERGER T., BREMM S., SCHRECK T.: Cluster correspondence views for enhanced analysis of som displays. In *IEEE Symposium on Visual Analytics Science and Technology (Poster)*, (2010), pp. 217–218. doi:10.1109/VAST.2010.5651676.
15. **BERNARD J.**, VON LANDESBERGER T., BREMM S., SCHRECK T.: Micro-macro views for visual trajectory cluster analysis. In *Eurographics/IEEE Symposium on Visualization (Poster)*, (2009). Poster.

Talks

1. **Visual Analysis of Time-oriented Data**. Value and Challenges of Exploratory Search Concepts in User-centered Design Approaches, British Telecom, Ipswich, UK, 2013
2. **Visual Analysis of Time-oriented Data**. Value and Challenges of Exploratory Search Concepts in User-centered Design Approaches. Workshop on Visual Analytics — Present and Future of Human-Computer Interaction in Data Analytics - Conference on Artificial Intelligence (AI), Cambridge, UK, 2013
3. **Content-based vs. Metadata-based Search and Analysis**. Universität Konstanz, Konstanz, DE, 2012
4. **Visual Access to Time-Oriented Scientific Primary Data**. DKRZ, Hamburg, DE, 2011
5. **A Visual Digital Library Approach for Time-Oriented Scientific Primary Data**. A WGL-TIB Project. DataCite Workshop, Hannover, DE, 2010

Supervising Activities

The following list summarizes the academic accreditations supervised by the author. Some of the results of these works were partially used as an input into this thesis.

Bachelor Thesis

- David Sessler - User-centered Interactive Similarity Definition for Complex Data Objects
- Alex Ulmer - Visuelle Analyse multidimensionaler Optimierungsprobleme (supervisor: Tobias Ruppert)

Internships

- David Sessler - Nutzerzentrierte visuell-interaktive Ähnlichkeitsdefinition komplexer Datenobjekte
- Eduard Dobermann - Segmentierung Multivariater Zeitserien
- Saskia Koldijk - Visual Analytics of Work Behavior Data-Insights on Individual Differences
- Peter Sheldrik - Timebox widgets for interactive time series querying
- Tobias Stoll - Explorative Literatursuche
- Oliver Goroll - Visual-Interactive Time Series Preprocessing
- Nils Wilhelm - Visuelle Analyse von Multivariaten Zeitserien am Beispiel von Motion Capturing Data
- Jan Riemann - Implementierung eines symbolischen Zeitseriendeskriptors
- Martin Weigel - Implementierung eines symbolischen Zeitseriendeskriptors
- Michael Stoica - Implementierung eines symbolischen Zeitseriendeskriptors

Curriculum Vitae

Personal Data

Name	Jürgen Bernard
Birth date	January 21st, 1981. In Lohr am Main, Bavaria, Germany
Family status	Unmarried
Nationality	German

Education

2015	PhD Defense: 2015-11-20, Interactive Graphics Systems Group, Technische Universität Darmstadt, Germany
2010 – 2015	Cand. Dr.-Ing (PhD), Computer Science Faculty, Interactive Graphics Systems Group, Technische Universität Darmstadt, Germany
2009	Graduation (Diploma - 'Diplom Informatiker') in Computer Science at the Technische Universität Darmstadt, Germany
2009	Diploma Thesis 'Methoden zur Qualitätsbewertung von Self-organizing Maps zur Unterstützung des visuellen Analyseprozesses' at Technische Universität Darmstadt, Germany
2001 – 2009	Study at the Technische Universität Darmstadt, Germany
2000	High School Graduation (Abitur) in Lohr am Main, Germany

Work Experience

2012 – 2015	Researcher, Fraunhofer Institute for Computer Graphics Research IGD, Germany. Focus: Visual Analytics, Exploratory Search, Visual Cluster Analysis, Time-oriented Primary Data, User-centered Design, Digital Libraries
2008 – 2011	Researcher, Interactive Graphics Systems Group, Technische Universität Darmstadt, Germany. Focus: Visual Analytics, Visual Cluster Analysis, Time Series Analysis
2001 – 2007	Work Student (Software Engineer) at Bosch Rexroth AG, Lohr am Main, Germany. Focus: Visual thermal load analysis of industrial drive amplifiers
2000 – 2001	Alternative non-military Service at a Geriatric Nursing Center in Lohr am Main, Germany. Focus: Male Nurse, health care
1997 – 2000	Temporary Employee at Mannesmann Rexroth AG, Lohr am Main, Germany. Focus: Assembly and bench testing of industrial drive amplifiers

Bibliography

[AA06] ANDRIENKO N., ANDRIENKO G.: *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. doi:10.1007/3-540-31190-4. 5, 16, 20, 21, 26, 35, 36, 37, 39, 40, 41, 42, 44, 59, 66, 74, 79, 97, 112, 133, 134, 138

[AA13] ANDRIENKO N., ANDRIENKO G.: A visual analytics framework for spatio-temporal analysis and modelling. *Data Mining and Knowledge Discovery 27*, 1 (2013), 55–83. doi:10.1007/s10618-012-0285-7. 54, 63, 80, 83, 84, 223

[AAMG12] ALSALLAKH B., AIGNER W., MIKSCH S., GROLLER M.: Reinventing the contingency wheel: Scalable visual analytics of large categorical data. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 18*, 12 (2012), 2849–2858. doi:10.1109/TVCG.2012.254. 162, 163

[ABB*07] AGOSTI M., BERRETTI S., BRETTLECKER G., BIMBO A., FERRO N., FUHR N., KEIM D., KLAS C.-P., LIDY T., MILANO D., NORRIE M., RANALDI P., RAUBER A., SCHEK H.-J., SCHRECK T., SCHULDT H., SIGNER B., SPRINGMANN M.: Delosdlms - the integrated delos digital library management system. In *Digital Libraries: Research and Development*, vol. 4877 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2007, pp. 36–45. doi:10.1007/978-3-540-77088-6_4. 28, 35, 200

[ABM*07] AIGNER W., BERTONE A., MIKSCH S., TOMINSKI C., SCHUMANN H.: Towards a conceptual framework for visual analytics of time and time-oriented data. In *Conference on Winter Simulation: 40 Years! The Best is Yet to Come (WSC)* (Piscataway, NJ, USA, 2007), IEEE Press, pp. 721–729. 36, 52, 53, 63, 80, 83, 84, 223

[AFS93] AGRAWAL R., FALOUTSOS C., SWAMI A. N.: Efficient similarity search in sequence databases. In *International Conference on Foundations of Data Organization and Algorithms (FODO)* (London, UK, UK, 1993), Springer-Verlag, pp. 69–84. 42, 94

[AKD10] AILAMAKI A., KANTERE V., DASH D.: Managing scientific data. *Communications of the ACM 53*, 6 (2010), 68–78. doi:10.1145/1743546.1743568. 27, 28, 29, 30, 31, 32, 49, 50, 55, 64, 200

[AKM*09] ATTWOOD T., KELL D., MCDERMOTT P., MARSH J., PETTIFER S., THORNE D.: Calling international rescue: knowledge lost in literature and data landslide! *Biochem. J 424* (2009), 317–333. doi:10.1042/BJ20091474. 25, 49, 54, 57, 64, 200

[AMA*14] ALSALLAKH B., MICALLEF L., AIGNER W., HAUSER H., MIKSCH S., RODGERS P.: Visualizing sets and set-typed data: State-of-the-art and future challenges. In *IEEE Eurographics Conference on Visualization (EuroVis)– State of The Art Reports* (2014), Borgo R., Maciejewski R.,, Viola I., (Eds.), Eurographics Association, Eurographics Association, pp. 1–21. doi:10.2312/eurovisstar.20141170. 162

[AMST11] AIGNER W., MIKSCH S., SCHUMANN H., TOMINSKI C.: *Visualization of Time-Oriented Data*, 1st ed. Human-Computer Interaction. Springer Verlag, 2011. doi:10.1007/978-0-85729-079-3. 4, 5, 19, 20, 21, 35, 36, 37, 38, 39, 40, 41, 44, 59, 64, 73, 74, 78, 79, 92, 93, 97

[AS94] AHLBERG C., SHNEIDERMAN B.: Visual information seeking: Tight coupling of dynamic query filters with starfield displays. In *SIGCHI Conference on Human Factors in Computing Systems (CHI)* (New York, NY, USA, 1994), ACM, pp. 313–317. doi:10.1145/191666.191775. 19, 25, 34, 42, 43, 49, 75

[AYMW11] AHMED Z., YOST P., MCGOVERN A., WEAVER C.: Steerable Clustering for Visual Analysis of Ecosystems. In *International Workshop on Visual Analytics (EuroVA)* (2011), Eurographics Association, pp. 49–52. doi:10.2312/PE/EuroVAST/EuroVA11/049-052. 45, 48, 118, 121

[BAF*13] BÖGL M., AIGNER W., FILZMOSER P., LAMMARSCH T., MIKSCH S., RIND A.: Visual analytics for model selection in time series analysis. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 19*, 12 (2013), 2237–2246. doi:10.1109/TVCG.2013.222. 47, 49, 66, 89

[Bat89] BATES M. J.: The design of browsing and berrypicking techniques for the online search interface. *Online review 13*, 5 (1989), 407–431. 19, 77

[BBC*10] BERNDT R., BLÜMEL I., CLAUSEN M., DAMM D., DIET J., FELLNER D., FREMEREY C., KLEIN R., KRAHL F., SCHERER M., ET AL.: The probado project-approach and lessons learned in building a digital library system for heterogeneous non-textual documents. In *Research and Advanced Technology for Digital Libraries*. Springer, 2010, pp. 376–383. 35, 200

[BBF*10] BERNARD J., BRASE J., FELLNER D., KOEPLER O., KOHLHAMMER J., RUPPERT T., SCHRECK T., SENS I.: A visual digital library approach for time-oriented research data. In *European Conference on Research and Advanced Technology for Digital Libraries (ECDL)* (Berlin, Heidelberg, 2010), Springer-Verlag, pp. 352–363. doi:10.1007/978-3-642-15464-5_35. 69, 87, 155, 198, 201, 203, 207

[BBF*11] BERNARD J., BRASE J., FELLNER D., KOEPLER O., KOHLHAMMER J., RUPPERT T., SCHRECK T., SENS I.: A visual digital library approach for time-oriented scientific primary data. *Springer International Journal of Digital Libraries, ECDL 2010 Special Issue 11*, 2 (2011), 111–123. doi:10.1007/s00799-011-0072-x. 69, 87, 93, 135, 155, 198, 201, 203, 207

[BC02] BÖRNER K., CHEN C.: Visual interfaces to digital libraries: Motivation, utilization, and socio-technical challenges. In *Visual interfaces to digital libraries*. Springer, 2002, pp. 1–9. 67

[BCS*05] BAVOIL L., CALLAHAN S. P., SCHEIDEGGER C. E., VO H. T., CROSSNO P., SILVA C. T., FREIRE J.: Vistrails: Enabling interactive multiple-view visualizations. In *IEEE Visualization* (2005), IEEE Computer Society, p. 18. doi:10.1109/VIS.2005.113. 48, 49, 50, 90, 229

[BDF*15] BERNARD J., DABERKOW D., FELLNER D., FISCHER K., KOEPLER O., KOHLHAMMER J., RUNNWERTH M., RUPPERT T., SCHRECK T., SENS I.: VisInfo: a digital library system for time series research data based on exploratory search - a user-centered design approach. *International Journal on Digital Libraries 16*, 1 (2015), 37–59. doi:10.1007/s00799-014-0134-y. 69, 109, 135, 198, 201, 203, 207

[Bea07] BEARMAN D.: Digital libraries. *Annual review of information science and technology 41*, 1 (2007), 223–272. 28, 199

[Bec14] BECK F.: Software feathers: Figurative visualization of software metrics. In *International Conference on Information Visualization Theory and Application (IVAPP)* (2014), IEEE, pp. 5–16. doi:10.5220/0004650100050016. 117

[Ber83] BERTIN J.: *Semiology of Graphics*. University of Wisconsin Press, 1983. 4, 19, 23, 40, 49, 59, 117, 134, 136, 153

[Ber09] BERNARD J.: *Methoden zur Qualitätsbewertung von Self-organizing Maps zur Unterstützung des visuellen Analyseprozesses*. Diploma thesis (Diplomarbeit), Technische Universität Darmstadt, Graphisch-Interaktive Systeme (GRIS), Darmstadt, Germany, 2009. 109, 115, 116, 124, 125, 126, 128

[BKC*13] BORGO R., KEHRER J., CHUNG D. H., MAGUIRE E., LARAMEE R. S., HAUSER H., WARD M., CHEN M.: Glyph-based visualization: Foundations, design guidelines, techniques and applications. In *Eurographics State of the Art Reports* (2013), EG STARs, Eurographics Association, pp. 39–63. 19, 77, 112, 117, 133, 134, 153, 161, 194

[BKLS12] BERNARD J., KÖNIG-LANGLO G., SIEGER R.: Time-oriented earth observation measurements from the Baseline Surface Radiation Network (BSRN) in the years 1992 to 2012, reference list of 6813 datasets, 2012. PANGAEA - Data Publisher for Earth and Environmental Science. doi:10.1594/PANGAEA.787726. 182, 201

[BM13] BREHMER M., MUNZNER T.: A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 19*, 12 (2013), 2376–2385. doi:10.1109/TVCG.2013.124. 5, 6, 16, 21, 39, 40, 41, 44, 73, 74, 75, 79

[Bor10] BORGMAN C. L.: *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. The MIT Press, 2010. 28

[BPFG11] BERGER W., PIRINGER H., FILZMOSER P., GRÖLLER E.: Uncertainty-aware exploration of continuous parameter spaces using multivariate prediction. In *IEEE Eurographics Conference on Visualization*

(EuroVis) (2011), Eurographics Association, pp. 911–920. doi:10.1111/j.1467-8659.2011.01940.x. 22, 98

[Bra04] BRASE J.: Using digital library techniques–registration of scientific primary data. In *Research and advanced technology for digital libraries*. Springer, 2004, pp. 488–494. 27, 28, 199, 200

[BRG*12] BERNARD J., RUPPERT T., GOROLL O., MAY T., KOHLHAMMER J.: Visual-interactive preprocessing of time series data. In *SIGRAD* (2012), Kerren A., Seipel S., (Eds.), vol. 81 of *Linköping Electronic Conference Proceedings*, Linköping University Electronic Press, pp. 39–48. 69, 87, 89, 90, 95, 105, 109, 198, 203, 205, 207

[BRS*12a] BERNARD J., RUPPERT T., SCHERER M., KOHLHAMMER J., SCHRECK T.: Content-based layouts for exploratory metadata search in scientific research data. In *ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)* (New York, NY, USA, 2012), ACM, pp. 139–148. doi:10.1145/2232817.2232844. 87, 105, 109, 135, 146, 155, 159, 162, 163, 175, 182, 211

[BRS*12b] BERNARD J., RUPPERT T., SCHERER M., SCHRECK T., KOHLHAMMER J.: Guided discovery of interesting relationships between time series clusters and metadata properties. In *International Conference on Knowledge Management and Knowledge Technologies (i-KNOW)* (New York, NY, USA, 2012), ACM, pp. 22:1–22:8. doi:10.1145/2362456.2362485. 87, 135, 155, 159, 163, 166, 167, 168, 174

[BSB*14] BERNARD J., SESSLER D., BERISCH M., HUTTER M., SCHRECK T., KOHLHAMMER J.: Towards a User-Defined Visual-Interactive Definition of Similarity Functions for Mixed Data. In *IEEE Symposium on Visual Analytics Science and Technology (Poster Paper)* (2014). 230

[BSM*15a] BERNARD J., SESSLER D., MAY T., SCHLOMM T., PEHRKE D., KOHLHAMMER J.: A visual-interactive system for prostate cancer cohort analysis. *IEEE Computer Graphics and Applications (CG&A) 35*, 3 (2015), 44–55. doi:10.1109/MCG.2015.49. 69

[BSM*15b] BERNARD J., STEIGER M., MITTELSTÄDT S., THUM S., KEIM D., KOHLHAMMER J.: A survey and task-based quality assessment of static 2D colormaps. In *SPIE, Visualization and Data Analysis (VDA)* (2015), vol. 9397, pp. 93970M–93970M–16. doi:10.1117/12.2079841. 109, 138, 139, 211

[BSR*14] BERNARD J., SESSLER D., RUPPERT T., DAVEY J., KUIJPER A., KOHLHAMMER J.: User-based visual-interactive similarity definition for mixed data objects-concept and first implementation. *Journal of WSCG 22* (2014). 42, 87, 89, 106, 155, 163, 230

[BSW*14] BERNARD J., STEIGER M., WIDMER S., LÜCKE-TIEKE H., MAY T., KOHLHAMMER J.: Visual-interactive Exploration of Interesting Multivariate Relations in Mixed Research Data Sets. *Computer Graphics Forum (CGF) 33*, 3 (2014), 291–300. doi:10.1111/cgf.12385. 134, 135, 155, 159, 162, 163, 185, 186, 187, 189, 191, 193

[BTK11] BERTINI E., TATU A., KEIM D.: Quality metrics in high-dimensional data visualization: An overview and systematization. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 17*, 12 (2011), 2203–2212. doi:10.1109/TVCG.2011.229. 51

[BvLBS09] BERNARD J., VON LANDESBERGER T., BREMM S., SCHRECK T.: Micro-macro views for visual trajectory cluster analysis. In *IEEE Eurographics Conference on Visualization (EuroVis) (Poster Paper)* (2009), Eurographics Association. Poster. 109, 127, 131, 135, 213

[BvLBS10] BERNARD J., VON LANDESBERGER T., BREMM S., SCHRECK T.: Cluster correspondence views for enhanced analysis of som displays. In *IEEE Symposium on Visual Analytics Science and Technology (Poster Paper)* (2010), pp. 217–218. doi:10.1109/VAST.2010.5651676. 109

[BvLBS11a] BERNARD J., VON LANDESBERGER T., BREMM S., SCHRECK T.: Multiscale visual quality assessment for cluster analysis with Self-Organizing Maps. In *IS&T/SPIE Conference on Visualization and Data Analysis (VDA)* (2011), SPIE Press, pp. 78680N.1 – 78680N.12. doi:10.1117/12.872545. 109, 115, 116, 125, 126, 128, 131

[BvLBS11b] BREMM S., VON LANDESBERGER T., BERNARD J., SCHRECK T.: Assisted descriptor selection based on visual comparative data analysis. In *IEEE Eurographics Conference on Visualization (EuroVis)* (Aire-la-Ville, Switzerland, Switzerland, 2011), Eurographics Association, pp. 891–900. doi:10.1111/j.1467-8659.2011.01938.x. 117, 120

[BWE06] BORGMAN C., WALLIS J. C., ENYEDY N.: Building digital libraries for scientific data: An exploratory study of data practices in habitat ecology. In *European Conference on Research and Advanced*

Technology for Digital Libraries (ECDL) (Berlin, Heidelberg, 2006), Springer-Verlag, pp. 170–183. doi:10.1007/11863878_15. 9, 22, 26, 27, 29, 30, 31, 32, 34, 49, 54, 55, 57, 58, 64, 67, 200, 208

[BWK*13] BERNARD J., WILHELM N., KRUGER B., MAY T., SCHRECK T., KOHLHAMMER J.: MotionExplorer: Exploratory search in human motion capture data based on hierarchical aggregation. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 19*, 12 (2013), 2257–2266. doi:10.1109/TVCG.2013.178. 69, 109, 135, 136, 210, 211, 212, 218

[BWS*12] BERNARD J., WILHELM N., SCHERER M., MAY T., SCHRECK T.: Timeseriespaths: Projection-based explorative analysis of multivariate time series data. *Journal of WSCG 20*, 2 (2012), 97–106. 109, 135, 136, 144, 211, 213, 217

[BYRN*99] BAEZA-YATES R., RIBEIRO-NETO B., ET AL.: *Modern information retrieval*, vol. 463. ACM press New York, 1999. 16, 29, 34, 199, 200

[Car08] CARPENDALE S.: Information visualization. Springer-Verlag, Berlin, Heidelberg, 2008, ch. Evaluating Information Visualizations, pp. 19–45. doi:10.1007/978-3-540-70956-5_2. 60, 61

[CCF*08] CANDELA L., CASTELLI D., FERRO N., KOUTRIKA G., MEGHINI C., PAGANO P., ROSS S., SOERGEL D., AGOSTI M., DOBREVA M. (Eds.): *The DELOS Digital Library Reference model. Foundations for digital Libraries (Version 0.98)*. ISTI-CNR at Gruppo ALI, Pisa, 2008. 26, 28, 31, 32, 55, 199

[CF99] CHAN K.-P., FU A.-C.: Efficient time series matching by wavelets. In *Conference on Data Engineering(ICDE)* (Washington, DC, USA, 1999), IEEE Computer Society, pp. 126–133. doi:10.1109/ICDE.1999.754915. 42, 94

[CGSQ11] CAO N., GOTZ D., SUN J., QU H.: Dicon: Interactive visual analysis of multidimensional clusters. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 17*, 12 (2011), 2581–2590. doi:10.1109/TVCG.2011.188. 117, 120

[CIZW13] COSTAS R., I. M., ZAHEDI Z., WOUTERS P.: The Value of Research Data Metrics for datasets from a cultural and technical point of view. A Knowledge Exchange Report. 2013. 27, 28, 29, 30, 31, 32, 47, 55, 200

[CKB09] COCKBURN A., KARLSON A., BEDERSON B. B.: A review of overview+detail, zooming, and focus+context interfaces. *ACM Comput. Surv. 41*, 1 (2009), 2:1–2:31. doi:10.1145/1456650.1456652. 64, 112, 149, 150

[CMS99] CARD S. K., MACKINLAY J. D., SHNEIDERMAN B. (Eds.): *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999. 7, 19, 20, 21, 49, 50, 51, 55, 59, 60, 63, 64, 74, 80, 81, 82, 83, 84, 117, 118, 141, 223

[CSL*10] CAO N., SUN J., LIN Y.-R., GOTZ D., LIU S., QU H.: Facetatlas: Multifaceted visualization for rich text corpora. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 16*, 6 (2010), 1172–1181. 119, 161

[DCCW08] DORK M., CARPENDALE S., COLLINS C., WILLIAMSON C.: Visgets: Coordinated visualizations for web-based information exploration and discovery. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 14*, 6 (2008), 1205–1212. 25, 77, 78, 161

[DF08] DAVIDSON S. B., FREIRE J.: Provenance and scientific workflows: Challenges and opportunities. In *ACM International Conference on Management of Data (SIGMOD)* (New York, NY, USA, 2008), ACM, pp. 1345–1350. doi:10.1145/1376616.1376772. 30, 46, 47, 50, 55, 229

[DGST09] DEELMAN E., GANNON D., SHIELDS M., TAYLOR I.: Workflows and e-science: An overview of workflow system features and capabilities. *Future Gener. Comput. Syst. 25*, 5 (2009), 528–540. doi:10.1016/j.future.2008.06.012. 3, 30, 45, 46, 47, 88

[dSB04] DOS SANTOS S., BRODLIE K.: Gaining understanding of multivariate and multidimensional data through visualization. *Computers & Graphics 28*, 3 (2004), 311–325. 51

[DTS*08] DING H., TRAJCEVSKI G., SCHEUERMANN P., WANG X., KEOGH E.: Querying and mining of time series data: Experimental comparison of representations and distance measures. *VLDB Endowment 1*, 2 (2008), 1542–1552. doi:10.14778/1454159.1454226. 94, 95

[ED06] ELLIS G., DIX A.: An explorative analysis of user evaluation studies in information visualisation. In *AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization (BELIV)* (New York, NY, USA, 2006), ACM, pp. 1–7. doi:10.1145/1168149.1168152. 60, 61

[EF10] ELMQVIST N., FEKETE J.-D.: Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 16*, 3 (2010), 439–454. doi:10.1109/TVCG.2009.84. 65, 77, 112, 116, 120, 142, 148

[FCNL01] FU T.-C., CHUNG F.-L., NG V., LUK R.: Pattern discovery from stock time series using self-organizing maps. In *Workshop Notes of KDD2001 Workshop on Temporal Data Mining* (2001). 45

[Fek13] FEKETE J.-D.: Visual analytics infrastructures: From data management to exploration. *Computer 46*, 7 (2013), 22–29. 6, 23, 32, 33, 50, 52, 65, 66, 67, 89, 91, 200

[FFM12] FISCHER F., FUCHS J., MANSMANN F.: ClockMap: Enhancing Circular Treemaps with Temporal Glyphs for Time-Series Data. In *IEEE Eurographics Conference on Visualization (EuroVis) (Short Papers)* (2012), Meyer M., Weinkauf T., (Eds.), Eurographics Association, pp. 97–101. 117, 120

[FFM*13] FUCHS J., FISCHER F., MANSMANN F., BERTINI E., ISENBERG P.: Evaluation of alternative glyph designs for time series data in a small multiple setting. In *SIGCHI Conference on Human Factors in Computing Systems (CHI)* (New York, NY, USA, 2013), ACM, pp. 3237–3246. doi:10.1145/2470654.2466443. 117, 134

[FGS12] FOX E. A., GONÇALVES M. A., SHEN R.: *Theoretical Foundations for Digital Libraries: The 5S (Societies, Scenarios, Spaces, Structures, Streams) Approach*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2012. doi:10.2200/S00434ED1V01Y201207ICR022. 30, 32, 78, 199, 200

[FH09] FUCHS R., HAUSER H.: Visualization of multi-variate scientific data. *Computer Graphics Forum (CGF) 28*, 6 (2009), 1670–1690. 39, 57, 64

[FH11] FOX P., HENDLER J.: Changing the equation on scientific data visualization. *Science (New York, N.Y.) 331*, 6018 (2011), 705–708. doi:10.1126/science.1197654. 47, 49, 63, 200

[FHN*93] FOX E. A., HIX D., NOWELL L. T., BRUENI D. J., RAO D., WAKE W. C., HEATH L. S.: Users, user interfaces, and objects: Envision, a digital library. *Journal of the American Society for Information Science 44*, 8 (1993), 480–491. doi:10.1002/(SICI)1097-4571(199309)44:8<480::AID-ASI7>3.0.CO;2-B. 19, 24, 161

[FPS96] FAYYAD U. M., PIATETSKY-SHAPIRO G., SMYTH P.: From data mining to knowledge discovery in databases. *AI Magazine 17*, 3 (1996), 37–54. 7, 19, 23, 38, 44, 49, 50, 55, 74, 80, 83, 84, 91, 92, 93, 200, 223

[Fu11] FU T.-C.: A review on time series data mining. *Engineering Applications of Artificial Intelligence 24*, 1 (2011), 164–181. doi:10.1016/j.engappai.2010.09.007. 38, 39, 92, 93, 94

[GAW*11] GLEICHER M., ALBERS D., WALKER R., JUSUFI I., HANSEN C. D., ROBERTS J. C.: Visual comparison for information visualization. *Information Visualization 10*, 4 (2011), 289–309. doi:10.1177/1473871611416549. 22, 77, 78, 97, 161

[GGAM12] GSCHWANDTNER T., GÄRTNER J., AIGNER W., MIKSCH S.: A taxonomy of dirty time-oriented data. In *CD-ARES* (2012), Quirchmayr G., Basl J., You I., Xu L.,, Weippl E., (Eds.), vol. 7465 of *Lecture Notes in Computer Science*, Springer, pp. 58–72. 30, 35, 38, 39, 64, 88, 90, 92, 200

[GMPS00] GREENE S., MARCHIONINI G., PLAISANT C., SHNEIDERMAN B.: Previews and overviews in digital libraries: Designing surrogates to support visual information seeking. *Journal of the American Society for Information Science 51*, 4 (2000), 380–393. 28, 34, 64, 110

[GOB*12] GRETARSSON B., O'DONOVAN J., BOSTANDJIEV S., HÖLLERER T., ASUNCION A., NEWMAN D., SMYTH P.: TopicNets: Visual Analysis of Large Text Corpora with Topic Modeling. *ACM Trans. Intell. Syst. Technol. 3*, 2 (2012), 23:1–23:26. doi:10.1145/2089094.2089099. 116, 120, 209

[HA06] HEER J., AGRAWALA M.: Software design patterns for information visualization. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 12*, 5 (2006), 853–860. doi:10.1109/TVCG.2006.178. 51

[HCL05] HEER J., CARD S. K., LANDAY J. A.: Prefuse: A toolkit for interactive information visualization. In *SIGCHI Conference on Human Factors in Computing Systems (CHI)* (New York, NY, USA, 2005), ACM, pp. 421–430. doi:10.1145/1054972.1055031. 49, 51

[HCQ*12] HECHT B., CARTON S. H., QUADERI M., SCHÖNING J., RAUBAL M., GERGLE D., DOWNEY D.: Explanatory semantic relatedness and explicit spatialization for exploratory search. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)* (New York, NY, USA,

2012), ACM, pp. 415–424. doi:10.1145/2348283.2348341. 25, 78, 163

[Hea06] HEARST M. A.: Clustering versus faceted categories for information exploration. *Communications of the ACM 49*, 4 (2006), 59–61. 78, 158, 161

[Hea09] HEARST M. A.: *Search User Interfaces*, 1st ed. Cambridge University Press, New York, NY, USA, 2009. 18, 19, 22, 26, 34, 56, 75, 76, 77, 79, 82, 134, 229

[HF09] HOLZ C., FEINER S.: Relaxed selection techniques for querying time-series graphs. In *ACM Symposium on User Interface Software and Technology* (2009), ACM, pp. 213–222. 43, 44

[HK12] HERRMANNOVA D., KNOTH P.: Visual search for supporting content exploration in large document collections. *D-Lib Magazine 18*, 7/8 (2012). 7, 24, 26, 28, 34, 35, 62, 78, 119, 199, 200

[HKP11] HAN J., KAMBER M., PEI J.: *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2011. 23, 38, 39, 92, 93, 94

[HPK08] HULL D., PETTIFER S. R., KELL D. B.: Defrosting the digital library: bibliographic tools for the next generation web. *PLoS computational biology 4*, 10 (2008), e1000204. 26, 28, 34, 49, 58, 67, 208

[HS04] HOCHHEISER H., SHNEIDERMAN B.: Dynamic query tools for time series data sets: Timebox widgets for interactive exploration. *Information Visualization 3*, 1 (2004), 1–18. doi:10.1145/993176.993177. 25, 26, 42, 43, 44, 78

[HTT09] HEY A. J. G., TANSLEY S., TOLLE K. M.: *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009. 1, 3, 26, 27, 30, 31, 35, 45, 46, 47, 55, 198, 199

[IMI*10] INGRAM S., MUNZNER T., IRVINE V., TORY M., BERGNER S., MÖLLER T.: Dimstiller: Workflows for dimensional analysis and reduction. In *IEEE VAST* (2010), IEEE, pp. 3–10. doi:10.1109/VAST.2010.5652392. 48, 49, 67, 77, 89, 118

[Jai10] JAIN A. K.: Data clustering: 50 years beyond k-means. *Pattern Recognition Letters 31*, 8 (2010), 651–666. 111, 113, 114, 120

[JJJ08] JOHANSSON S., JERN M., JOHANSSON J.: Interactive quantification of categorical variables in mixed data sets. In *IEEE Information Visualisation (IV)* (2008), IEEE, pp. 3–10. doi:10.1109/IV.2008.33. 29, 157, 162, 163

[JMF99] JAIN A. K., MURTY M. N., FLYNN P. J.: Data clustering: A review. *ACM Comput. Surv. 31*, 3 (1999), 264–323. doi:10.1145/331499.331504. 111, 113, 114

[Jol02] JOLLIFFE I. T.: *Principal Component Analysis*, 3rd ed. Springer, 2002. 23, 94, 115, 118

[KAF*08] KEIM D., ANDRIENKO G., FEKETE J.-D., GÖRG C., KOHLHAMMER J., MELANÇON G.: Information visualization. Springer-Verlag, Berlin, Heidelberg, 2008, ch. Visual Analytics: Definition, Process, and Challenges, pp. 154–175. doi:10.1007/978-3-540-70956-5_7. 7, 20, 49, 52, 55, 59, 66, 82, 83, 84, 200, 223

[KBH06] KOSARA R., BENDIX F., HAUSER H.: Parallel sets: Interactive exploration and visual analysis of categorical data. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 12*, 4 (2006), 558–568. doi:10.1109/TVCG.2006.76. 29, 162

[KCHP04] KEOGH E. J., CHU S., HART D., PAZZANI M.: Segmenting Time Series: A Survey and Novel Approach. In *Data Mining In Time Series Databases*, vol. 57 of *Series in Machine Perception and Artificial Intelligence*. World Scientific Publishing Company, 2004, ch. 1, pp. 1–22. 67, 93, 94, 97

[KCPM01] KEOGH E., CHAKRABARTI K., PAZZANI M., MEHROTRA S.: Dimensionality reduction for fast similarity search in large time series databases. *Knowledge and Information Systems 3*, 3 (2001), 263–286. doi:10.1007/PL00011669. 94, 176, 203

[Kei02] KEIM D. A.: Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 8*, 1 (2002), 1–8. doi:10.1109/2945.981847. 6, 23, 26, 74, 78, 79, 133

[KH13] KEHRER J., HAUSER H.: Visualization and visual analysis of multifaceted scientific data: A survey. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 19*, 3 (2013), 495–513. doi:10.1109/TVCG.2012.110. 4, 21, 22, 26, 27, 30, 47, 48, 64, 74, 77, 78, 79, 117, 160, 161, 200

[KHP*11] KANDEL S., HEER J., PLAISANT C., KENNEDY J., VAN HAM F., RICHE N. H., WEAVER C., LEE B., BRODBECK D., BUONO P.: Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization 10*, 4 (2011), 271–288.

doi:10.1177/1473871611415994. 30, 38, 39, 49, 64, 88, 89, 90, 91, 92, 93, 200

[KK03] KEOGH E., KASETTY S.: On the need for time series data mining benchmarks: A survey and empirical demonstration. *Data Min. Knowl. Discov. 7*, 4 (2003), 349–371. doi:10.1023/A:1024988512476. 38, 39, 90, 92, 93, 94

[KKEM10] KEIM D., KOHLHAMMER J., ELLIS G., MANSMANN F. (Eds.): *Mastering the Information Age: Solving Problems with Visual Analytics*. VisMaster, http://www.vismaster.eu/book/, 2010. 49, 50, 52, 55, 59, 63, 66, 81, 106, 153, 200

[KL06] KINCAID R., LAM H.: Line graph explorer: Scalable display of line graphs using focus+context. In *Working Conference on Advanced Visual Interfaces (AVI)* (New York, NY, USA, 2006), ACM, pp. 404–411. doi:10.1145/1133265.1133348. 48

[KLK*05] KUMAR N., LOLLA N., KEOGH E., LONARDI S., RATANAMAHATANA C. A.: Time-series bitmaps: a practical visualization tool for working with large time series databases. In *SIAM 2005 Data Mining Conference* (2005), SIAM, pp. 531–535. 36, 45, 92, 117

[KLM*08] KEHRER J., LADSTADTER F., MUIGG P., DOLEISCH H., STEINER A., HAUSER H.: Hypothesis generation in climate research with interactive visual data exploration. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 14*, 6 (2008), 1579–1586. doi:10.1109/TVCG.2008.139. 34, 46, 48, 77, 161, 162, 199

[KLR04] KEOGH E., LONARDI S., RATANAMAHATANA C. A.: Towards parameter-free data mining. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)* (New York, NY, USA, 2004), ACM, pp. 206–215. doi:10.1145/1014052.1014077. 49, 66, 89

[KMH06] KLEIN G., MOON B., HOFFMAN R. R.: Making sense of sensemaking 1: Alternative perspectives. *IEEE Intelligent Systems 21*, 4 (2006), 70–73. doi:10.1109/MIS.2006.75. 18

[KMS*08] KEIM D. A., MANSMANN F., SCHNEIDEWIND J., THOMAS J., ZIEGLER H.: Visual data mining. Springer-Verlag, Berlin, Heidelberg, 2008, ch. Visual Analytics: Scope and Challenges, pp. 76–90. doi:10.1007/978-3-540-71080-6_6. 6, 23, 26, 33, 34, 47, 49, 52, 55, 58, 64, 66, 67, 77, 79, 109, 163, 200, 208

[KMSZ06] KEIM D. A., MANSMANN F., SCHNEIDEWIND J., ZIEGLER H.: Challenges in visual data analysis. In *Conference on Information Visualization (IV)* (Washington, DC, USA, 2006), IEEE Computer Society, pp. 9–16. doi:10.1109/IV.2006.31. 29, 30, 77

[KP98] KEOGH E., PAZZANI M.: An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. In *Conference on Knowledge Discovery and Data Mining (KDD)* (New York City, NY, 1998), ACM Press, pp. 239–241. 94

[Kru64] KRUSKAL J. B.: Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika 29*, 1 (1964), 1–27. 119

[KSH01] KOHONEN T., SCHROEDER M. R., HUANG T. S. (Eds.): *Self-Organizing Maps*, 3rd ed. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2001. 45, 115, 116, 119, 120, 122, 124, 128, 161

[LAB*06] LUDÄSCHER B., ALTINTAS I., BERKLEY C., HIGGINS D., JAEGER E., JONES M., LEE E. A., TAO J., ZHAO Y.: Scientific workflow management and the kepler system: Research articles. *Concurr. Comput. : Pract. Exper. 18*, 10 (2006), 1039–1065. doi:10.1002/cpe.v18:10. 3, 32, 46, 48, 229

[LAB*09] LUDÄSCHER B., ALTINTAS I., BOWERS S., CUMMINGS J., CRITCHLOW T., DEELMAN E., ROURE D. D., FREIRE J., GOBLE C., JONES M., KLASKY S., MCPHILLIPS T., PODHORSZKI N., SILVA C., TAYLOR I., VOUK M.: Scientific process automation and workflow management. Computational Science Series. Chapman & Hall, 2009, pp. 476–508. 46, 47

[LBI*12] LAM H., BERTINI E., ISENBERG P., PLAISANT C., CARPENDALE S.: Empirical studies in information visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 18*, 9 (2012), 1520–1536. doi:10.1109/TVCG.2011.279. 26, 56, 60, 61

[LD11] LLOYD D., DYKES J.: Human-centered approaches in geovisualization design: Investigating multiple methods through a long-term case study. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 17*, 12 (2011), 2498–2507. doi:10.1109/TVCG.2011.209. 56, 57, 61

[LK06] LARAMEE R. S., KOSARA R.: Challenges and unsolved problems. In *Human-Centered Visualization Environments* (2006), Kerren A., Ebert A.,, Meyer J., (Eds.), vol. 4417 of *Lecture Notes in Computer Science*, Springer, pp. 231–254. doi:10.1007/978-3-540-71949-6. 64, 65, 67, 89, 111

[LKL*04] LIN J., KEOGH E., LONARDI S., LANKFORD J. P., NYSTROM D. M.: Visually mining and monitoring massive time series. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2004), ACM, pp. 460–469. doi:10.1145/1014052.1014104. 36, 38, 39, 92, 94

[LKL05] LIN J., KEOGH E., LONARDI S.: Visualizing and discovering non-trivial patterns in large time series databases. *Information Visualization 4*, 2 (2005), 61–82. doi:10.1057/palgrave.ivs.9500089. 42, 45

[LKLC03] LIN J., KEOGH E., LONARDI S., CHIU B.: A symbolic representation of time series, with implications for streaming algorithms. In *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD)* (New York, NY, USA, 2003), ACM, pp. 2–11. doi:10.1145/882082.882086. 36, 38, 39, 45, 92, 94

[LPSW06] LAGOZE C., PAYETTE S., SHIN E., WILPER C.: Fedora: An architecture for complex objects and their relationships. *Int. J. Digit. Libr. 6*, 2 (2006), 124–138. doi:10.1007/s00799-005-0130-3. 27

[LSS*11] LEX A., SCHULZ H.-J., STREIT M., PARTL C., SCHMALSTIEG D.: Visbricks: Multiform visualization of large, inhomogeneous data. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 17*, 12 (2011), 2291–2300. doi:10.1109/TVCG.2011.250. 29, 162

[MA14] MIKSCH S., AIGNER W.: A matter of time: Applying a data-users-tasks design triangle to visual analytics of time-oriented data. *Computers & Graphics, Special Section on Visual Analytics 38* (2014), 286–290. doi:10.1016/j.cag.2013.11.002. 9, 15, 16, 20, 25, 39, 49, 53, 55, 56, 57, 58, 59, 60, 66, 67, 111, 200, 208

[Mac67] MACQUEEN J. B.: Some methods for classification and analysis of multivariate observations. In *Berkeley Symposium on Mathematical Statistics and Probability* (1967), Cam L. M. L., Neyman J., (Eds.), vol. 1, University of California Press, pp. 281–297. 23, 114

[Mac95] MACEACHREN A.: *How Maps Work: Representation, Visualization and Design*. Guilford Press, New York, 1995. 16, 39

[Mar95] MARCHIONINI G.: *Information Seeking in Electronic Environments*. Cambridge University Press, New York, NY, USA, 1995. 5, 10, 15, 16, 17, 18, 24, 27, 28, 32, 34, 54, 55, 62, 63, 73, 74, 77, 78, 79, 200

[Mar06] MARCHIONINI G.: Exploratory search: From finding to understanding. *Commun. ACM 49*, 4 (2006), 41–46. doi:10.1145/1121949.1121979. 5, 6, 16, 17, 21, 22, 23, 25, 26, 27, 32, 34, 55, 75, 78, 79, 197

[MBS*14] MITTELSTÄDT S., BERNARD J., SCHRECK T., STEIGER M., KOHLHAMMER J., KEIM D. A.: Revisiting Perceptually Optimized Color Mapping for High-Dimensional Data Analysis. *IEEE Eurographics Conference on Visualization (EuroVis) (Short Paper)* (2014), 91–95. doi:10.2312/eurovisshort.20141163. 118

[MH10] MARCIAL L. H., HEMMINGER B. M.: Scientific data repositories on the web: An initial survey. *J. Am. Soc. Inf. Sci. Technol. 61*, 10 (2010), 2029–2048. doi:10.1002/asi.v61:10. 28, 31, 32, 55, 200

[MMKN08] MCLACHLAN P., MUNZNER T., KOUTSOFIOS E., NORTH S.: Liverac: Interactive visual exploration of system management time-series data. In *SIGCHI Conference on Human Factors in Computing Systems (CHI)* (New York, NY, USA, 2008), ACM, pp. 1483–1492. doi:10.1145/1357054.1357286. 45, 59, 113

[Mör06] MÖRCHEN F.: *Time series knowledge mining*. Citeseer, 2006. 36, 38, 39, 42, 44, 88, 90, 92, 93, 94

[MPR02] MERKL D., PAMPALK E., RAUBER A.: Using Psycho-Acoustic Models and Self-Organizing Maps to Create a Hierarchical Structuring of Music by Sound Similarity. In *ISMIR 2002 Proceedings* (2002). 25, 35, 44, 117, 120, 161, 164, 200

[MRC05] MÜLLER M., RÖDER T., CLAUSEN M.: Efficient content-based retrieval of motion capture data. *ACM Trans. Graph. 24*, 3 (2005), 677–685. doi:10.1145/1073204.1073247. 26, 35, 211

[MRC*07] MÜLLER M., RÖDER T., CLAUSEN M., EBERHARDT B., KRÜGER B., WEBER A.: *Documentation Mocap Database HDM05*. Tech. Rep. CG-2007-2, Universität Bonn, 2007. 28, 211, 212

[MS88] MARCHIONINI G., SHNEIDERMAN B.: Finding facts vs. browsing knowledge in hypertext systems. *Computer 21*, 1 (1988), 70–80. doi:10.1109/2.222119. 16, 19, 23, 34, 76, 78, 79

[Mun09] MUNZNER T.: A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 15*, 6 (2009), 921–928. doi:10.1109/TVCG.2009.111. 26, 49, 55, 56, 58, 59, 60, 61, 63, 67, 84, 208

[NB12] NOCAJ A., BRANDES U.: Organizing search results with a reference map. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 18*, 12 (2012), 2546–2555. doi:10.1109/TVCG.2012.250. 25, 116, 120, 161

[Noa07] NOACK A.: Energy models for graph clustering. *J. Graph Algorithms Appl. 11*, 2 (2007), 453–480. 59, 119, 178

[Nor02] NORMAN D. A.: *The Design of Everyday Things*, reprint paperback ed. Basic Books, New York, 2002. 16, 17, 24, 73, 74

[NRB*13] NAZEMI K., RETZ R., BERNARD J., KOHLHAMMER J., FELLNER D.: Adaptive semantic visualization for bibliographic entries. In *Advances in Visual Computing (ISVC)*, Bebis G., Boyle R., Parvin B., Koracin D., Li B., Porikli F., Zordan V., Klosowski J., Coquillart S., Luo X., Chen M.,, Gotz D., (Eds.), vol. 8034 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2013, pp. 13–24. doi:10.1007/978-3-642-41939-3_2. 163

[NSBW08] NOCKE T., STERZEL T., BÖTTINGER M., WROBEL M.: Visualization of climate and climate change data: An overview. *Digital Earth Summit on Geoinformatics 2008: Tools for Global Change Research* (2008), 226–232. 34, 49, 67, 199, 200

[PAN] PANGAEA - DATA PUBLISHER FOR EARTH AND ENVIRONMENTAL SCIENCE:. http://www.pangaea.de/. Last accessed on June 19th, 2014. doi:10.1594/pangaea. 28, 34, 55, 175, 182, 198, 200

[Pla04] PLAISANT C.: The challenge of information visualization evaluation. In *Working Conference on Advanced Visual Interfaces (AVI)* (New York, NY, USA, 2004), ACM, pp. 109–116. doi:10.1145/989863.989880. 60, 61

[POM07] PAULOVICH F. V., OLIVEIRA M. C. F., MINGHIM R.: The projection explorer: A flexible tool for projection-based multidimensional visualization. In *Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI)* (Washington, DC, USA, 2007), IEEE Computer Society, pp. 27–36. doi:10.1109/SIBGRAPI.2007.39. 77, 118, 119

[PVW09] PRETORIUS A. J., VAN WIJK J. J.: What does the user want to see?: What do the data want to be? *Information Visualization 8*, 3 (2009), 153–166. doi:10.1057/ivs.2009.13. 4, 26, 33, 47, 55, 56, 57, 58, 60, 61, 64, 65, 111, 230

[RBK13] RUPPERT T., BERNARD J., KOHLHAMMER J.: Bridging knowledge gaps in policy analysis with information visualization. In *Conference on Electronic Government (EGOV/ePart Ongoing Research)* (2013), vol. 221, GI, pp. 92–103. 32, 67, 200

[RBU*14] RUPPERT T., BERNARD J., ULMER A., LÜCKE-TIEKE H., KOHLHAMMER J.: Visual access to an agent-based simulation model to support political decision making. In *International Conference on Knowledge Technologies and Data-driven Business (i-KNOW* (New York, NY, USA, 2014), ACM, pp. 16:1–16:8. doi:10.1145/2637748.2638410. 54

[RLL*05] RYALL K., LESH N., LANNING T., LEIGH D., MIYASHITA H., MAKINO S.: Querylines: Approximate query for visual browsing. In *Extended Abstracts on Human Factors in Computing Systems (CHI EA)* (New York, NY, USA, 2005), ACM, pp. 1765–1768. doi:10.1145/1056808.1057017. 43, 44

[RWA*13] RIND A., WANG T. D., AIGNER W., MIKSCH S., WONGSUPHASAWAT K., PLAISANT C., SHNEIDERMAN B., ET AL.: Interactive information visualization to explore and query electronic health records. *Foundations and Trends in Human-Computer Interaction 5*, 3 (2013), 207–298. 25

[SBG00] SPRENGER T. C., BRUNELLA R., GROSS M. H.: H-blob: A hierarchical visual clustering method using implicit surfaces. In *Conference on Visualization (VIS)* (Los Alamitos, CA, USA, 2000), IEEE Computer Society Press, pp. 61–68. 116, 120

[SBKK14] SESSLER D., BERNARD J., KUIJPER A., KOHLHAMMER J.: Adopting mental similarity notions of categorical data objects to algorithmic similarity functions, 2014. Vision, Modelling and Visualization (VMV) (Poster Paper). 230

[SBM92] SPRINGMEYER R. R., BLATTNER M. M., MAX N. L.: A characterization of the scientific data analysis process. In *Conference on Visualization (VIS)* (Los Alamitos, CA, USA, 1992), IEEE Computer Society Press, pp. 235–242. 23, 48, 59, 74

[SBM*14] STEIGER M., BERNARD J., MITTELSTÄDT S., LÜCKE-TIEKE H., KEIM D. A., MAY T., KOHLHAMMER J.: Visual Analysis of Time-Series Similarities for Anomaly Detection in Sensor

Networks. *Computer Graphics Forum (CGF) 33*, 3 (2014), 401–410. doi:10.1111/cgf.12396. 44, 109, 119, 155

[SBMK14] STEIGER M., BERNARD J., MAY T., KOHLHAMMER J.: A survey of direction-preserving layout strategies. In *Spring Conference on Computer Graphics (SCCG)* (New York, NY, USA, 2014), ACM, pp. 21–28. doi:10.1145/2643188.2643189. 109

[SBS11] SCHERER M., BERNARD J., SCHRECK T.: Retrieval and exploratory search in multivariate research data repositories using regressional features. In *International ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (New York, NY, USA, 2011), ACM, pp. 363–372. doi:10.1145/1998076.1998144. 42, 48, 69, 78, 87, 155, 163

[SBSK15] STEIGER M., BERNARD J., SCHADER P., KOHLHAMMER J.: Visual Analysis of Relations in Attributed Time-Series Data. In *EuroVis Workshop on Visual Analytics (EuroVA)* (2015), Bertini E., Roberts J. C., (Eds.), Eurographics Association. doi:10.2312/eurova.20151105. 155

[SBVLK09] SCHRECK T., BERNARD J., VON LANDESBERGER T., KOHLHAMMER J.: Visual cluster analysis of trajectory data with interactive kohonen maps. *Information Visualization 8*, 1 (2009), 14–29. doi:10.1057/ivs.2008.29. 45, 93, 109, 120, 121, 125, 135, 136

[SC07] SALVADOR S., CHAN P.: Toward accurate dynamic time warping in linear time and space. *Intell. Data Anal. 11*, 5 (2007), 561–580. 95

[Sch13] SCHERER M.: *Information Retrieval for Multivariate Research Data Repositories*. PhD thesis, Technische Universität, Darmstadt, 2013. 35, 42

[SCM*06] SMITH G., CZERWINSKI M., MEYERS B., ROBBINS D., ROBERTSON G., TAN D. S.: Facetmap: A scalable search and browse visualization. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 12*, 5 (2006), 797–804. doi:10.1109/TVCG.2006.142. 25, 161

[SFRG00] SHNEIDERMAN B., FELDMAN D., ROSE A., GRAU X. F.: Visualizing digital library search results with categorical and hierarchical axes. In *ACM Conference on Digital Libraries* (2000), ACM, pp. 57–66. 24

[SGL08] STASKO J., GÖRG C., LIU Z.: Jigsaw: Supporting investigative analysis through interactive visualization. *Information Visualization 7*, 2 (2008), 118–132. doi:10.1145/1466620.1466622. 24, 77, 155, 161, 162

[Shn96] SHNEIDERMAN B.: The eyes have it: A task by data type taxonomy for information visualizations. In *IEEE Symposium on Visual Languages (VL)* (Washington, DC, USA, 1996), IEEE Computer Society, pp. 336–. 5, 6, 15, 18, 20, 28, 34, 35, 39, 44, 45, 64, 73, 74, 77, 78, 79, 109, 112, 147, 151, 158

[Shn02] SHNEIDERMAN B.: Inventing discovery tools: Combining information visualization with data mining. *Information Visualization 1*, 1 (2002), 5–12. doi:10.1057/palgrave/ivs/9500006. 79, 157, 163, 230

[SMM12] SEDLMAIR M., MEYER M. D., MUNZNER T.: Design study methodology: Reflections from the trenches and the stacks. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 18*, 12 (2012), 2431–2440. doi:10.1109/TVCG.2012.213. 26, 34, 49, 54, 55, 56, 58, 60, 61, 62, 63, 67, 84, 200, 208

[SN11] STOBER S., NÜRNBERGER A.: MusicGalaxy: A Multi-focus Zoomable Interface for Multi-facet Exploration of Music Collections. In *Exploring Music Contents* (2011), vol. 6684 of *Lecture Notes in Computer Science*, Springer Verlag, pp. 273–302. 25, 26, 35, 44, 119, 200

[SOR*09] STROBELT H., OELKE D., ROHRDANTZ C., STOFFEL A., KEIM D., DEUSSEN O.: Document Cards: A Top Trumps Visualization for Documents. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 15*, 6 (2009), 1145–1152. doi:10.1109/TVCG.2009.139. 76, 117

[SS02] SEO J., SHNEIDERMAN B.: Interactively exploring hierarchical clustering results. *Computer 35*, 7 (2002), 80–86. doi:10.1109/MC.2002.1016905. 25, 26, 66, 67, 75, 111, 115, 116, 120, 153

[SSW*12] SCHRECK T., SHARALIEVA L., WANNER F., BERNARD J., RUPPERT T., VON LANDESBERGER T., BUSTOS B.: Visual exploration of local interest points in sets of time series. In *2012 IEEE Conference on Visual Analytics Science and Technology, VAST 2012, Seattle, WA, USA, October 14-19, 2012* (2012), pp. 239–240. doi:10.1109/VAST.2012.6400534. 67, 87

[STK*03] STEINBACH M., TAN P.-N., KUMAR V., KLOOSTER S., POTTER C.: Discovery of climate indices using clustering. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*

(KDD) (New York, NY, USA, 2003), ACM, pp. 446–455. doi:10.1145/956750.956801. 34, 48, 199

[Sub12] SUBER P.: *Open Access*. The MIT Press, 2012. 32, 55

[TC05] THOMAS J. J., COOK K. A.: *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Ctr, 2005. 7, 20, 26, 55, 61, 74, 89

[TC06] THOMAS J. J., COOK K. A.: A visual analytics agenda. *IEEE Comput. Graph. Appl. 26*, 1 (2006), 10–13. doi:10.1109/MCG.2006.5. 31, 49, 55, 59, 200

[TDN11] TOMINSKI C., DONGES J. F., NOCKE T.: Information visualization in climate research. *International Conference on Information Visualization 0* (2011), 298–305. doi:10.1109/IV.2011.12. 34, 49, 67, 199, 200

[TFS08] TOMINSKI C., FUCHS G., SCHUMANN H.: Task-driven color coding. In *12th International Conference on Information Visualisation, IV 2008, 8-11 July 2008, London, UK* (2008), pp. 373–380. doi:10.1109/IV.2008.24. 40, 118

[The12] THE ROYAL SOCIETY: *Science as an open enterprise*. Tech. Rep. June, 2012. 1, 3, 27, 28, 29, 32, 55, 229

[TLLH13] TURKAY C., LUNDERVOLD A., LUNDERVOLD A., HAUSER H.: Hypothesis generation by interactive visual exploration of heterogeneous medical data. In *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*, vol. 7947 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2013, pp. 1–12. doi:10.1007/978-3-642-39146-0_1. 48, 49, 50, 66, 200

[TM04] TORY M., MOLLER T.: Human factors in visualization research. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 10*, 1 (2004), 72–84. 19, 26, 49, 56

[Tuf86] TUFTE E. R.: *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, USA, 1986. 49, 59, 134

[Tuf90] TUFTE E.: *Envisioning Information*. Graphics Press, Cheshire, CT, USA, 1990. 4, 19, 23, 49, 59, 77, 112, 116, 127, 133, 134, 150

[Tuk77] TUKEY J. W.: *Exploratory Data Analysis*. Addison-Wesley, 1977. 20, 23

[vA10] VON ANTBURG T. L.: *Visual Analytics of Large Weighted Directed Graphs and Two-Dimensional Time-Dependent Data*. PhD thesis, Technische Universität, Darmstadt, 2010. 121

[vdEvW13] VAN DEN ELZEN S., VAN WIJK J. J.: Small multiples, large singles: A new approach for visual data exploration. In *Computer Graphics Forum (CGF)* (2013), vol. 32, Wiley Online Library, pp. 191–200. 78, 118

[Ves99] VESANTO J.: Som-based data visualization methods. *Intelligent Data Analysis 3* (1999), 111–126. 45, 117, 118, 120, 161

[vHP09] VAN HAM F., PERER A.: "search, show context, expand on demand": Supporting large graph exploration with degree-of-interest. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 15*, 6 (2009), 953–960. doi:10.1109/TVCG.2009.108. 22

[vLBBS10] VON LANDESBERGER T., BREMM S., BERNARD J., SCHRECK T.: Smart query definition for content-based search in large sets of graphs. In *International Symposium on Visual Analytics Science and Technology (Short Paper)* (2010), Eurographics Association, pp. 7–12. 44

[vLSFK12] VON LANDESBERGER T., SCHRECK T., FELLNER D., KOHLHAMMER J.: Visual Search and Analysis in Complex Information Spaces-Approaches and Research Challenges. In *Expanding the Frontiers of Visual Analytics and Visualization*, Dill J., Earnshaw R., Kasik D., Vince J.,, Wong P. C., (Eds.). Springer London, 2012, pp. 45–67. doi:10.1007/978-1-4471-2804-5_4. 22, 26, 31, 44, 62, 65, 78, 199, 229

[vW06] VAN WIJK J. J.: Bridging the gaps. *IEEE Comput. Graph. Appl. 26*, 6 (2006), 6–9. doi:10.1109/MCG.2006.120. 34, 49, 54, 55, 56, 57, 58, 67, 200

[VWVS99] VAN WIJK J. J., VAN SELOW E. R.: Cluster and calendar based visualization of time series data. In *IEEE Symposium on Information Visualization (InfoVis* (Washington, DC, USA, 1999), IEEE Computer Society, pp. 4–. 44, 77, 117

[War12] WARE C.: *Information Visualization: Perception for Design*, 3 ed. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2012. 19, 20, 49, 59

[WBBM00] WITTEN I. H., BODDIE S. J., BAINBRIDGE D., MCNAB R. J.: Greenstone: A comprehensive open-

source digital library software system. In *ACM Conference on Digital Libraries (DL)* (New York, NY, USA, 2000), ACM, pp. 113–121. doi:10.1145/336597.336650. 28, 199

[WGGP*11] WONGSUPHASAWAT K., GUERRA GÓMEZ J. A., PLAISANT C., WANG T. D., TAIEB-MAIMON M., SHNEIDERMAN B.: Lifeflow: Visualizing an overview of event sequences. In *SIGCHI Conference on Human Factors in Computing Systems (CHI)* (New York, NY, USA, 2011), ACM, pp. 1747–1756. doi:10.1145/1978942.1979196. 45, 117

[WL05] WARREN LIAO T.: Clustering of time series data-a survey. *Pattern Recogn. 38*, 11 (2005), 1857–1874. doi:10.1016/j.patcog.2005.01.025. 91, 93, 94

[WR09] WHITE R. W., ROTH R. A.: Exploratory search: Beyond the query-response paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services 1*, 1 (2009), 1–98. doi:10.2200/S00174ED1V01Y200901ICR003. 5, 6, 15, 16, 17, 18, 19, 22, 23, 24, 25, 26, 32, 42, 62, 63, 78, 79, 199, 200, 205, 229

[WVZ*15] WILHELM N., VÖGELE A., ZSOLDOS R., LICKA T., KRÜGER B., BERNARD J.: Furyexplorer: visual-interactive exploration of horse motion capture data. In *SPIE, Visualization and Data Analysis (VDA)* (2015), pp. 93970F–93970F–15. doi:10.1117/12.2080001. 69, 109, 211

[YKSJ07] YI J. S., KANG Y. A., STASKO J., JACKO J.: Toward a deeper understanding of the role of interaction in information visualization. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 13*, 6 (2007), 1224–1231. doi:10.1109/TVCG.2007.70515. 20, 21, 50, 54, 64, 65, 73, 74, 77

[ZCB11] ZHAO J., CHEVALIER F., BALAKRISHNAN R.: Kronominer: Using multi-foci navigation for the visual exploration of time-series data. In *SIGCHI Conference on Human Factors in Computing Systems (CHI)* (New York, NY, USA, 2011), ACM, pp. 1737–1746. doi:10.1145/1978942.1979195. 45, 117

[ZJGK10] ZIEGLER H., JENNY M., GRUSE T., KEIM D.: Visual market sector analysis for financial time series data. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on* (2010), pp. 83–90. doi:10.1109/VAST.2010.5652530. 45, 94