mastering the information age

solving problems with visual analytics

Edited by Daniel Keim, Jörn Kohlhammer, Geoffrey Ellis and Florian Mansmann

Mastering the Information Age
Solving Problems with
Visual Analytics

Edited by Daniel Keim, Jörn Kohlhammer,
Geoffrey Ellis and Florian Mansmann

Published by the Eurographics Association
–Postfach 8043, 38621 Goslar, Germany–

Printed in Germany, Druckhaus “Thomas Müntzer” GmbH, Bad Langensalza

Cover image: © iStockphoto.com/FotoMak

ISBN 978-3-905673-77-7

The electronic version of this book is available from the Eurographics Digital Library at `http://diglib.eg.org`

In Memoriam of Jim Thomas[1],
a visionary and inspiring person, innovative researcher, enthusiastic leader
and excellent promoter of visual analytics.

[1] Jim Thomas passed away on August 6, 2010.

Preface

Today, in many spheres of human activity, massive sets of data are collected and stored. As the volumes of data available to lawmakers, civil servants, business people and scientists increase, their effective use becomes more challenging. Keeping up to date with the flood of data, using standard tools for data management and analysis, is fraught with difficulty. The field of visual analytics seeks to provide people with better and more effective ways to understand and analyse these large datasets, while also enabling them to act upon their findings immediately, in real-time. Visual analytics integrates the analytic capabilities of the computer and the abilities of the human analyst, thus allowing novel discoveries and empowering individuals to take control of the analytical process. Visual analytics sheds light on unexpected and hidden insights, which may lead to beneficial and profitable innovation.

This book is one of the outcomes of a two-year project called VisMaster CA, a coordination action funded by the European Commission from August 2008 to September 2010. The goal of VisMaster was to join European academic and industrial R&D excellence from several individual disciplines, forming a strong visual analytics research community. An array of thematic working groups was set up by the consortium, which focused on advancing the state of the art in visual analytics. These working groups joined research excellence in the fields of data management, data analysis, spatial-temporal data, and human visual perception research with the wider visualisation research community.

This Coordination Action successfully formed and shaped a strong European visual analytics community, defined the research roadmap, exposed public and private stake-holders to visual analytics technology and set the stage for larger follow-up visual analytics research initiatives. While there is still much work ahead to realise the visions described in this book, Europe's most prestigious visual analytics researchers have combined their expertise to determine the next steps.

This research roadmap is the final delivery of VisMaster. It presents a detailed review of all aspects of visual analytics, indicating open areas and strategies for the research in the coming years. The primary sources for this book are the final reports of the working groups, the cross-community reports as well as the resources built up on the Web platform[2].

The VisMaster consortium is confident that the research agenda presented in this book, and especially the recommendations in the final chapter, will help to support a sustainable visual analytics community well beyond the duration of VisMaster CA, and also serves as the reference for researchers in related

[2] http://www.vismaster.eu

scientific disciplines, which are interested to join and strengthen the community. This research roadmap does not only cover issues that correspond to scientific challenges: it also outlines the connections to sciences, technologies, and industries for which visual analytics can become an 'enabling technology'. Hence, it serves as a reference for research program committees and researchers of related fields in the ICT theme and beyond, to assess the possible implications for their respective field.

Structure

Chapter 1 motivates the topic of visual analytics and presents a brief history of the domain. Chapter 2 deals with the basis of visual analytics including its current application areas, the visual analytics process, its building blocks, and its inherent scientific challenges.

The following Chapters 3 to 8 were written by respective working groups in VisMaster, assisted by additional partners of the consortium and community partners. Each of these chapters introduces the specific community that is linked to visual analytics (e.g., data mining). It then outlines the state of the art and the specific challenges and opportunities that lie ahead for this field with respect to visual analytics research. In particular, Chapter 3 deals with data management for visual analytics, Chapter 4 covers aspects of data mining, Chapter 5 outlines the application of visual analytics to problems with spatial and temporal components, Chapter 6 considers infra-structural issues, Chapter 7 looks at human aspects and Chapter 8 discusses evaluation methodologies for visual analytics.

The final chapter presents a summary of challenges for the visual analytics community and sets out specific recommendations to advance visual analytics research. These recommendations are a collaborative effort of all working groups and specifically address different target groups: the European Commission, the visual analytics research community, the broader research community, industry and governments, together with other potential users of visual analytics technology.

Acknowledgements

We would like to thank all the partners of VisMaster (including community partners) who have contributed to creating this book. Whilst some have produced chapters (authors of each chapter are shown overleaf), others have been involved with the reviewing process and/or coordinating their work groups.

Special thanks goes to Bob Spence and Devina Ramduny-Ellis for their most helpful comments and contributions.

We are appreciative of the excellent technical and creative support given by Florian Stoffel, Juri Buchmüller and Michael Regenscheit. We are truly grateful

once more for the excellent support of Eurographics, and in particular Stefanie Behnke, for publishing this work.

Last but not least, we are indebted to the European Commission, and especially, the project officer of VisMaster CA, Dr. Teresa de Martino, for supporting us throughout; her efforts have contributed appreciably to the success of this project.

This project was funded by the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 225924.

We hope that the content of this book will inspire you to apply current visual analytics technology to solve your real-world data problems, and to engage in the community effort to define and develop visual analytics technologies to meet future challenges.

Daniel Keim (Scientific Coordinator of VisMaster), **Jörn Kohlhammer** (Coordinator of VisMaster), **Geoffrey Ellis**, and **Florian Mansmann**

September 2010

Contents

List of Authors

Chapters 1 & 2	Daniel A. Keim	University of Konstanz
	Jörn Kohlhammer	Fraunhofer IGD
	Florian Mansmann	University of Konstanz
	Thorsten May	Fraunhofer IGD
	Franz Wanner	University of Konstanz
Chapter 3	Giuseppe Santucci	Sapienza Università di Roma
	Helwig Hauser	University of Bergen
Chapter 4	Kai Puolamäki	Aalto University
	Alessio Bertone	Danube University Krems
	Roberto Therón	Universidad de Salamanca
	Otto Huisman	University of Twente
	Jimmy Johansson	Linköping University
	Silvia Miksch	Danube University Krems
	Panagiotis Papapetrou	Aalto University
	Salvo Rinzivillo	Consiglio Nazionale delle Ricerche
Chapter 5	Gennady Andrienko	Fraunhofer IAIS
	Natalia Andrienko	Fraunhofer IAIS
	Heidrun Schumann	University of Rostock
	Christian Tominski	University of Rostock
	Urska Demsar	National University of Ireland
	Doris Dransch	German Research Centre for Geosciences
	Jason Dykes	City University London
	Sara Fabrikant	University of Zurich
	Mikael Jern	Linköping University
	Menno-Jan Kraak	University of Twente
Chapter 6	Jean-Daniel Fekete	INRIA
Chapter 7	Alan Dix	Lancaster University
	Margit Pohl	Vienna University of Technology
	Geoffrey Ellis	Lancaster University
Chapter 8	Jarke van Wijk	Eindhoven University of Technology
	Tobias Isenberg	University of Groningen
	Jos B.T.M. Roerdink	University of Groningen
	Alexandru C. Telea	University of Groningen
	Michel Westenberg	Eindhoven University of Technology
Chapter 9	Geoffrey Ellis	University of Konstanz
	Daniel A. Keim	University of Konstanz
	Jörn Kohlhammer	Fraunhofer IGD

1 Introduction

1.1 Motivation

We are living in a world which faces a rapidly increasing amount of data to be dealt with on a daily basis. In the last decade, the steady improvement of data storage devices and means to create and collect data along the way, influenced the manner in which we deal with information. Most of the time, data is stored without filtering and refinement for later use. Virtually every branch of industry or business, and any political or personal activity, nowadays generates vast amounts of data. Making matters worse, the possibilities to collect and store data increase at a faster rate than our ability to use it for making decisions. However, in most applications, raw data has no value in itself; instead, we want to extract the information contained in it.

Raw data has no value in itself, only the extracted information has value

The information overload problem refers to the danger of getting lost in data, which may be:

- irrelevant to the current task at hand,
- processed in an inappropriate way, or
- presented in an inappropriate way.

Due to information overload, time and money are wasted, scientific and industrial opportunities are lost because we still lack the ability to deal with the enormous data volumes properly. People in both their business and private lives, decision-makers, analysts, engineers, emergency response teams alike, are often confronted with large amounts of disparate, conflicting and dynamic information, which are available from multiple heterogeneous sources. There is a need for effective methods to exploit and use the hidden opportunities and knowledge resting in unexplored data resources.

Time and money are wasted and opportunities are lost

In many application areas, success depends on the right information being available at the right time. Nowadays, the acquisition of raw data is no longer the main problem. Instead, it is the ability to identify methods and models, which can turn the data into reliable and comprehensible knowledge. Any technology, that claims to overcome the information overload problem, should answer the following questions:

Success depends on availability of the right information

- Who or what defines the 'relevance of information' for a given task?
- How can inappropriate procedures in a complex decision making process be identified?
- How can the resulting information be presented in a decision-oriented or task-oriented way?

With every new application, processes are put to the test, possibly under circumstances totally different from the ones they have been designed for. The awareness of the problem of how to understand and analyse our data has greatly increased in the last decade. Even though we implement more powerful tools for automated data analysis, we still face the problem of understanding and 'analysing our analyses' in the future – fully automated search, filter and analysis only work reliably for well-defined and well-understood problems. The path from data to decision is typically fairly complex. Fully automated data processing methods may represent the knowledge of their creators, but they lack the ability to communicate their knowledge. This ability is crucial. If decisions that emerge from the results of these methods turn out to be wrong, it is especially important to be able to examine the processes that are responsible.

Visual analytics aims at making data and information processing transparent

The overarching driving vision of visual analytics is to turn the information overload into an opportunity: just as information visualisation has changed our view on databases, the goal of visual analytics is to make our way of processing data and information transparent for an analytic discourse. The visualisation of these processes will provide the means of examining the actual processes instead of just the results. Visual analytics will foster the constructive evaluation, correction and rapid improvement of our processes and models and ultimately the improvement of our knowledge and our decisions.

Visual analytics combines the strengths of humans and computers

On a grand scale, visual analytics provides technology that combines the strengths of human and electronic data processing. Visualisation becomes the medium of a semi-automated analytical process, where humans and machines cooperate using their respective, distinct capabilities for the most effective results. The user has to be the ultimate authority in directing the analysis. In addition, the system has to provide effective means of interaction to focus on their specific task. In many applications, several people may work along the processing path from data to decision. A visual representation will sketch this path and provide a reference for their collaboration across different tasks and at different levels of detail.

The diversity of these tasks cannot be tackled with a single theory. Visual analytics research is highly interdisciplinary and combines various related research areas such as visualisation, data mining, data management, data fusion, statistics and cognition science (among others). One key idea of visual analytics is that integration of all these diverse areas is a scientific discipline in its own right. Application domain experts are becoming increasingly aware that visualisation is useful and valuable, but often ad hoc solutions are used, which rarely match the state of the art in interactive visualisation science, much less the full complexity of the problems, for which visual analytics aims to seek answers. Even if the awareness exists, that scientific analysis and results must be visualised in one way or the other. In fact, all related research areas in the context of visual analytics research conduct rigorous science, each in their vibrant research communities. One main goal of this book is to demonstrate that collaboration can lead to novel, highly effective analysis tools, contributing solutions to the information overload problem in many important domains.

Because visual analytics is an integrating discipline, application specific research areas can contribute existing procedures and models. Emerging from highly application-oriented research, research communities often work on specific solutions using the tools and standards of their specific fields. The requirements of visual analytics introduce new dependencies between these fields.

The integration of the previously mentioned disciplines into visual analytics will result in a set of well-established and agreed upon concepts and theories, allowing any scientific breakthrough in a single discipline to have a potential impact on the whole visual analytics field. In return, combining and upgrading these multiple technologies onto a new general level will have a great impact on a large number of application domains.

1.2 An Historical Perspective on Visual Analytics

Automatic analysis techniques such as statistics and data mining developed independently from visualisation and interaction techniques. However, some key thoughts changed the rather limited scope of the fields into what is today called visual analytics research. One of the most important steps in this direction was the need to move from confirmatory data analysis (using charts and other visual representations to just present results) to exploratory data analysis (interacting with the data/results), which was first stated in the statistics research community by John W. Tukey in his 1977 book, Exploratory Data Analysis[116].

Early visual analytics: exploratory data analysis

With improvements in graphical user interfaces and interaction devices, a research community devoted their efforts to information visualisation[25, 27, 104, 122]. At some stage, this community recognised the potential of integrating the user in the knowledge discovery and data mining process through effective and efficient visualisation techniques, interaction capabilities and knowledge transfer. This led to *visual data exploration* and *visual data mining*[64]. This integration considerably widened the scope of both the information visualisation and the data mining fields, resulting in new techniques and many interesting and important research opportunities.

Visual data exploration and visual data mining

Two of the early uses of the term *visual analytics* were in 2004[125] and a year later in the research and development agenda, Illuminating the Path[111]. More recently, the term is used in a wider context, describing a new multidisciplinary field that combines various research areas including visualisation, human-computer interaction, data analysis, data management, geo-spatial and temporal data processing, spatial decision support and statistics[67, 5].

Since 2004: visual analytics

Despite the relatively recent use of the term visual analytics, characteristics of visual analytics applications were already apparent in earlier systems, such as the CoCo system created in the early 1990s to achieve improvement in the design of a silicon chip[32]. In this system, numerical optimisation algorithms alone were acknowledged to have serious disadvantages, and it was found that some of these could be ameliorated if an experienced chip designer continually

Some earlier systems exhibited the characteristics of visual analytics

monitored and guided the algorithm when appropriate. The Cockpit interface supported this activity by showing, dynamically, hierarchically related and meaningful indications of chip performance and sensitivity information, as well as on-the-fly advice by an artificial intelligence system, all of which information could be managed to interactively.

1.3 Overview

This book is the result of a community effort of the partners of the VisMaster Coordinated Action funded by the European Union. The overarching aim of this project was to create a research roadmap that outlines the current state of visual analytics across many disciplines, and to describe the next steps to take in order to form a strong visual analytics community, enabling the development of advanced visual analytic applications. The first two chapters introduce the problem space and define visual analytics. Chapters 3 to 8 present the work of the specialised working groups within the VisMaster consortium. Each of these chapters follow a similar structure – the motivation section gives an outline of the problem and relevant background information; the next section presents an overview of the state of the art in the particular domain, with reference to visual analytics; challenges and opportunities are then identified; and finally in the next steps section, suggestions, pertinent to the subject of the chapter, are put forward for discussion. Higher level recommendations for the direction for future research in visual analytics, put forward by the chapter authors are collectively summarised in the final chapter. We now outline the chapters in more detail.

Daniel A. Keim
Jörn Kohlhammer
Florian Mansmann
Thorsten May
Franz Wanner

Chapter 2 describes some application areas for visual analytics and puts the size of the problem into context, and elaborates on the definition of visual analytics. The interdisciplinary nature of this area is demonstrated by considering the scientific fields that are an integral part of visual analytics.

Giuseppe Santucci
Helwig Hauser

Chapter 3 reviews the field of data management with respect to visual analytics and reviews current database technology. It then summarises the problems that can arise when dealing with large, complex and heterogeneous datasets or data streams. A scenario is given, which illustrates tight integration of data management and visual analytics. The state of the art section also considers techniques for the integration of data and issues relating to data reduction, including visual data reduction techniques and the related topic of visual quality metrics. The challenges section identifies important issues, such as dealing with uncertainties in the data and the integrity of the results, the management of semantics (i.e., data which adds meaning to the data values), the emerging area of data streaming, interactive visualisation of large databases and database issues concerning distributed and collaborative visual analytics.

Chapter 4 considers data mining, which is seen as fundamental to the automated analysis components of visual analytics. Since today's datasets are often extremely large and complex, the combination of human and automatic analysis is key to solving many information gathering tasks. Some case studies are presented which illustrate the use of knowledge discovery and data mining (KDD) in bioinformatics and climate change. The authors then pose the question of whether industry is ready for visual analytics, citing examples of the pharmaceutical, software and marketing industries. The state of the art section gives a comprehensive review of data mining/analysis tools such as statistical and mathematical tools, visual data mining tools, Web tools and packages. Some current data mining/visual analytics approaches are then described with examples from the bioinformatics and graph visualisation fields. Technical challenges specific to data mining are described such as achieving data cleaning, integration, data fusion etc. in real-time and providing the necessary infrastructure to support data mining. The challenge of integrating the human into the data process to go towards a visual analytics approach is discussed together with issues regarding its evaluation. Several opportunities are then identified, such as the need for generic tools and methods, visualisation of models and collaboration between the KDD and visualisation communities.

Kai Puolamäki
Alessio Bertone
Roberto Therón
Otto Huisman
Jimmy Johansson
Silvia Miksch
Panagiotis Papapetrou
Salvo Rinzivillo

Chapter 5 describes the requirements of visual analytics for spatio-temporal applications. Space (as in for example maps) and time (values change over time) are essential components of many data analysis problems; hence there is a strong need for visual analytics tools specifically designed to deal with the particular characteristics of these dimensions. Using a sizeable fictitious scenario, the authors guide the reader towards the specifics of time and space, illustrating the involvement of various people and agencies, and the many dependencies and problems associated with scale and uncertainties in the data. The current state of the art is described with a review of maps, geographic information systems, the representation of time, interactive and collaborative issues, and the implication of dealing with massive datasets. Challenges are then identified, such as dealing with diverse data at multiple scales, and supporting a varied set of users, including non-experts.

Gennady Andrienko
Natalia Andrienko
Heidrun Schumann
Christian Tominski
Urska Demsar
Doris Dransch
Jason Dykes
Sara Fabrikan
Mikael Jern
Menno-Jan Kraak

Chapter 6 highlights the fact that currently most visual analytics application are custom-built stand-alone applications, using for instance, in-memory data storage rather than database management systems. In addition, many other common components of visual analytics applications can be identified and potentially built into a unifying framework to support a range of applications. The author of this chapter reviews architectural models of visualisation, data management, analysis, dissemination and communication components and outlines the inherent challenges. Opportunities and next steps for current research are subsequently identified which encourage a collaborative multidisciplinary effort to provide a much needed flexible infrastructure.

Jean-Daniel Fekete

Chapter 7 discusses visual perception and cognitive issues - human aspects of visual analytics. Following a review of the psychology of perception and cognition, distributed cognition, problem solving, particular interaction issues, the authors suggest that we can learn much from early application

Alan Dix
Margit Pohl
Geoffrey Ellis

examples. Challenges identified, include the provision of appropriate design methodologies and design guidelines, suitable for the expert analyst as well as the naive users; understanding the analysis process, giving the user confidence in the results, dealing with a wide range of devices and how to evaluate new designs.

Jarke van Wijk
Tobias Isenberg
Jos B.T.M. Roerdink
Alexandru C. Telea
Michel Westenberg

Chapter 8 explains the basic concept of evaluation for visual analytics, highlighting the complexities for evaluating systems that involve the close coupling of the user and semi-automatic analytical processes through a highly interactive interface. The exploratory tasks associated with visual analytics are often open ended and hence it is difficult to assess the effectiveness and efficiency of a particular method, let alone make comparisons between methods. The state of the art section outlines empirical evaluation methodologies, shows some examples of evaluation and describes the development of contests in different sub-communities to evaluate visual analytics approaches on common datasets. The authors then argue that a solid evaluation infrastructure for visual analytics is required and put forward some recommendations on how to achieved this.

Geoffrey Ellis
Daniel A. Keim
Jörn Kohlhammer

Chapter 9 summarises the challenges of visual analytics applications as identified by the chapter authors and presents concrete recommendations for funding agencies, the visual analytics community, the broader research community and potential users of visual analytics technology in order to ensure the rapid advancement of the science of visual analytics.

2 Visual Analytics

Visual analytics is not easy to define, due to its multi-disciplinary nature involving multiple processes and the wide variety of application areas. An early definition was "the science of analytical reasoning facilitated by interactive human-machine interfaces"[125]. However, based on current practice, a more specific definition would be: "Visual analytics combines automated analysis techniques with interactive visualisations for an effective understanding, reasoning and decision making on the basis of very large and complex datasets".

Visual analytics combines automated analysis with interactive visualisations

So, in terms of the goal of visual analytics, we can elaborate on this definition to state that visual analytics is the creation of tools and techniques to enable people to:

- Synthesise 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 these assessment effectively for action.

In Section 2.2 we will look at how visual analytics strives to achieves these goals in terms of the high-level processes required to generate knowledge from data, and then in Section 2.3 in terms of the many scientific disciplines that contribute to visual analytics. But firstly, in order to give a sense of the social and economic importance of visual analytics, as well as the scale of the data being dealt with, we will look at some typical uses.

2.1 Application of Visual Analytics

Visual analytics is essential in application areas where large information spaces have to be processed and analysed. Major application fields are physics and astronomy. For example, the discipline of astrophysics offers many opportunities for visual analytics techniques: massive volumes of unstructured data, originating from different directions of space and covering the whole frequency spectrum, from continuous streams of terabytes of data that can be recorded and analysed. With common data analysis techniques, astronomers can separate relevant data from noise, analyse similarities or complex patterns, and gain useful knowledge about the universe, but the visual analytics approach can significantly support the process of identifying unexpected phenomena inside the massive and dynamic data streams that would otherwise not be found by standard algorithmic means. Monitoring climate and weather is also a domain which involves huge amounts of data collected by sensors throughout the world and from satellites, in short time intervals. A visual approach can help

Monitoring the climate involves huge amounts of data from many different sources

to interpret these massive amounts of data and to gain insight into the dependencies of climate factors and climate change scenarios that would otherwise not be easily identified. Besides weather forecasts, existing applications visualise global warming, melting of the poles, the stratospheric ozone depletion, as well as hurricane and tsunami warnings.

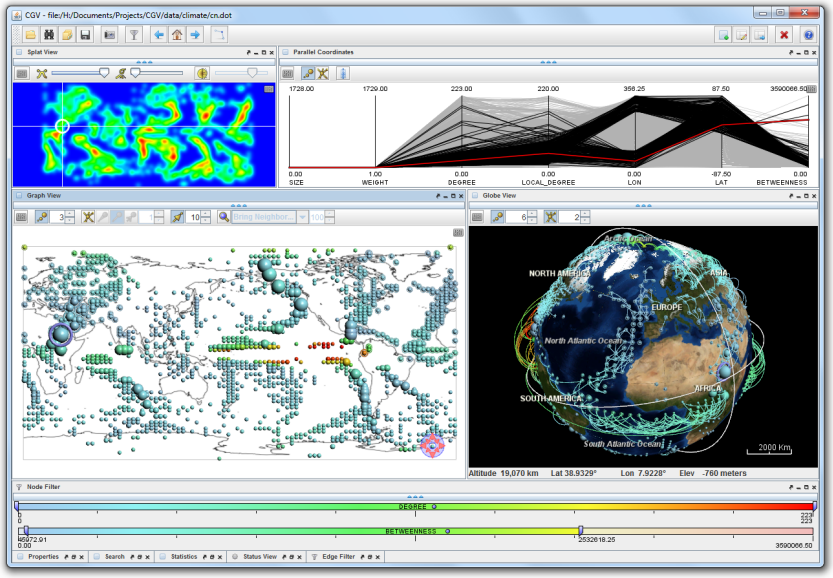

Figure 2.1: Visual analytics in action: Visual support for the simulation of climate models provided by CGV[113] (Coordinated Graph Visualization), a highly interactive graph visualisation system. To support different visualisation tasks, view ensembles can be created dynamically with the help of a flexible docking framework. CGV includes enhanced dynamic filtering, graph lenses, edge-based navigation, in addition to augmented navigation with infinite grid and radar view. Data source: Potsdam Institute for Climate Impact Research

More than 210 billion emails, 4 billion SMS and 50 million tweets per day

In the domain of emergency management, visual analytics can help determine the on-going progress of an emergency and identify the next countermeasures (e.g., construction of physical countermeasures or evacuation of the population) that must be taken to limit the damage. Such scenarios can include natural or meteorological catastrophes like flood or waves, volcanoes, storm, fire or epidemic growth of diseases (e.g. N1H1 virus), but also human-made technological catastrophes like industrial accidents, transport accidents or pollution. Visual analytics for security and geo-graphics is an important research topic. The application field in this sector is wide, ranging from terrorism informatics, border protection, path detection to network security. Visual analytics supports investigation and detection of similarities and anomalies in very large datasets. For example, on a worldwide scale, per day there are upwards of 210 billion emails, 4 billion SMS messages, 90 million tweets and the number of IP data packets exceeds 9000 billion. As an example

of document processing on a European level, the Europe Media Monitor collects news documents from 2,500 news sources: media portals, government websites, and news agencies and processes 80,000-100,000 articles per day in 43 languages.

Europe Media Monitor collects and processes 100,000 news articles per day in 43 languages

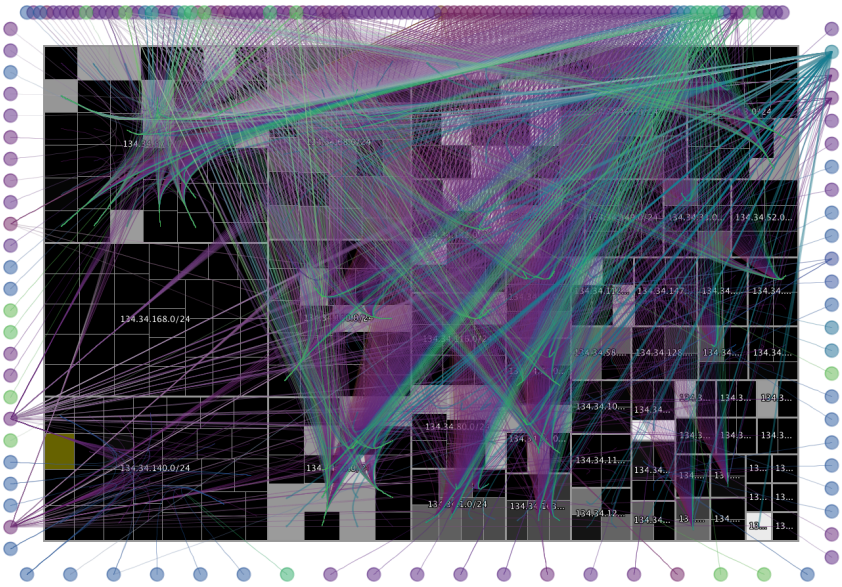

Figure 2.2: Visual analytics in action: Analysis of a distributed network attack on the SSH service of a university network using NFlowVis[76]. The TreeMap in the background represents the internal network structure with hosts as rectangles on the lowest level and external hosts as coloured dots on the outside. Hierarchical edge bundles reveal communication patterns such as the distributed attack from the hosts on the upper side

In biology and medicine, computer tomography, and ultrasound imaging for 3-dimensional digital reconstruction and visualisation produce gigabytes of medical data. The application area of bio-informatics uses visual analytics techniques to analyse large amounts of biological data. From the early beginning of sequencing, scientist in these areas face unprecedented volumes of data, like in the human genome project with three billion base pairs per human. Other new areas like proteomics (studies of the proteins in a cell), metabolomics (systematic study of unique chemical fingerprints that specific cellular processes leave behind) or combinatorial chemistry with tens of millions of compounds, add significant amounts of data every day. A brute-force computation of all possible combinations is often not possible, but interactive visual approaches can help to identify the main regions of interest and exclude unpromising areas.

Another major application domain for visual analytics is business intelligence. The financial market with its hundreds of thousands of assets generates large amounts of data on a daily basis, which results in extremely high data volumes

More than 300 million VISA credit card transaction per day

over the years. For example it is estimated that there are more than 300 million VISA credit card transaction per day. The main challenge in this area is to analyse the data under multiple perspectives and assumptions to understand historical and current situations, and then monitoring the market to forecast trends or to identify recurring situations. Other key applications in this area are fraud detection, the analysis of consumer data, social data and data associated with health care services.

Further application examples of visual analytics are shown in Figures 2.5 and 2.6 at the end of this chapter.

2.2 The Visual Analytics Process

Tight coupling of automated and visual analysis through interaction

The visual analytics process combines automatic and visual analysis methods with a tight coupling through human interaction in order to gain knowledge from data. Figure 2.3 shows an abstract overview of the different stages (represented through ovals) and their transitions (arrows) in the visual analytics process.

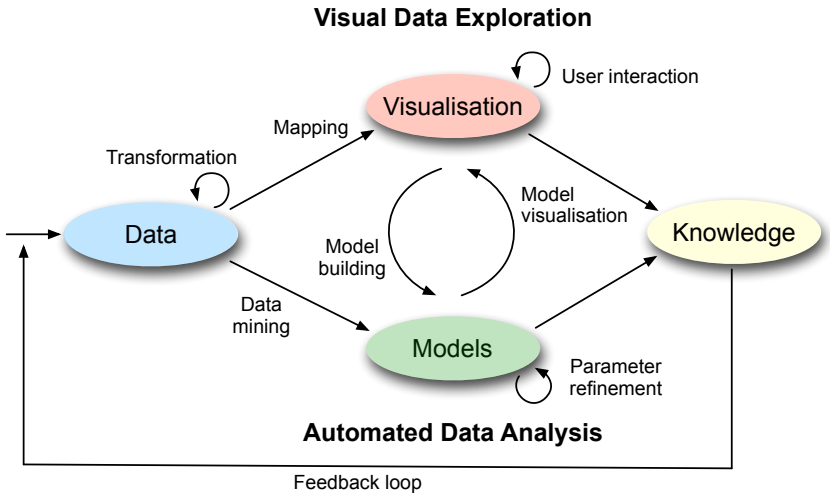

Figure 2.3: The visual analytics process is characterised through interaction between data, visualisations, models about the data, and the users in order to discover knowledge

In many application scenarios, heterogeneous data sources need to be integrated before visual or automatic analysis methods can be applied. Therefore, the first step is often to preprocess and transform the data to derive different representations for further exploration (as indicated by the *Transformation* arrow in Figure 2.3). Other typical preprocessing tasks include data cleaning, normalisation, grouping, or integration of heterogeneous data sources.

After the transformation, the analyst may choose between applying visual or automatic analysis methods. If an automated analysis is used first, data mining methods are applied to generate models of the original data. Once a model is created the analyst has to evaluate and refine the model, which can best be done by interacting with the data. Visualisations allow the analysts to interact with the automatic methods by modifying parameters or selecting other analysis algorithms. Model visualisation can then be used to evaluate the findings of the generated models. Alternating between visual and automatic methods is characteristic for the visual analytics process and leads to a continuous refinement and verification of preliminary results. Misleading results in an intermediate step can thus be discovered at an early stage, leading to better results and a higher confidence. If visual data exploration is performed first, the user has to confirm the generated hypotheses by an automated analysis. User interaction with the visualisation is needed to reveal insightful information, for instance by zooming in on different data areas or by considering different visual views on the data. Findings in the visualisations can be used to steer model building in the automatic analysis. In summary, in the visual analytics process, knowledge can be gained from visualisation, automatic analysis, as well as the preceding interactions between visualisations, models, and the human analysts.

Steer model building with visual findings

The visual analytics process aims at tightly coupling automated analysis methods and interactive visual representations. The guide to visually exploring data “Overview first, zoom/filter, details on demand”, as proposed by Shneiderman[98] in 1996 describes how data should be presented on screen. However, with massive datasets at hand, it is difficult to create an overview visualisation without losing interesting patterns, which makes zooming and filtering techniques effectively redundant as the users is given little information of what to examine further. In the context of visual analytics, the guide can usefully be extended to "Analyse first, show the important, zoom/filter, analyse further, details on demand”[65] indicating that it is not sufficient to just retrieve and display the data using a visual metaphor; rather, it is necessary to analyse the data according to its value of interest, showing the most relevant aspects of the data, and at the same time providing interaction models, which allow the user to get details of the data on demand.

Analyse first, show the important, zoom/filter, analyse further, details on demand.

2.3 Building Blocks of Visual Analytics Research

Visual analytics integrates science and technology from many disciplines, as shown in Figure 2.4. Visualisation is at the heart of the system, not only is it the means to communicate data values or the results of some analysis, but it is also increasingly being used to monitor processes in other disciplines, such as data management and data mining. We will now briefly consider the disciplines that contribute towards visual analytics.

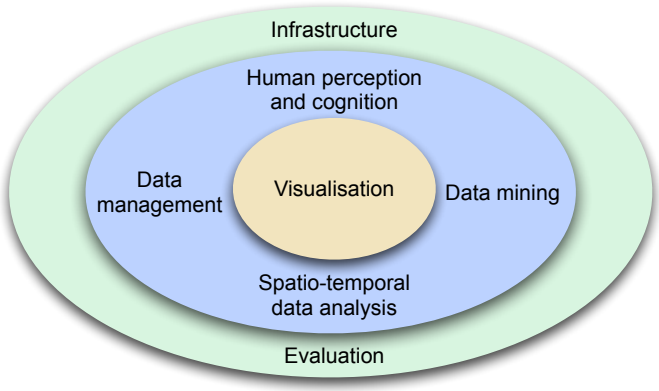

Figure 2.4: Visual analytics integrates visualisation with core adjacent disciplines and depends on the availability of appropriate infrastructure and evaluation facilities

Visualisation

Visualisation has emerged as a new research discipline during the last two decades. It can be broadly classified into scientific and information visualisation.

Scientific visualisation for 3D phenomena, such as fluid flow or molecular structures

Scientific visualisation is primarily concerned with visualising 3-dimensional (3D) data from the world of engineering, biology (whole body scans down to molecular structures), meteorology, cosmology, and so on, with the aim to represent the data, often temporal, as physical entities, such as surfaces, volumes and flows. A survey of current visualisation techniques can be found in the 'visualization handbook'[56]. Often, 3D scalar fields are visualised by iso-surfaces (3D contour) or semi-transparent point clouds. Also, in recent years, significant work has focused on the visualisation of complex 3D flow data, such as in aerospace engineering[114]. While current research has concentrated mainly on improving the efficiency of the visualisation techniques in enabling interactive exploration, more and more methods have been developed to automatically derive relevant visualisation parameters. In addition, interaction techniques such as focus & context[70] have gained importance in scientific visualisation.

Information visualisation for abstract data, often with many dimensions

Information visualisation has developed methods for the visualisation of abstract data where no explicit spatial references are given[104]. Typical examples include business data, demographics data, social networks and scientific data. Not only are we having to deal with huge volumes but the data often comprises of hundred of dimensions. Also, in addition to standard numeric and textual data types, some of these dimensions may be complex data types such as graphic, video, sound, and sophisticated data types now defined for the semantic web. The data values cannot be naturally mapped to 2D or 3D display space, as with scientific visualisation, and standard charting techniques such as x-y plots, line graphs and bar-charts are ineffective with large multi-dimensional

datasets. Moreover, as mentioned earlier, the capacity to interact with the data is extremely important. Novel visualisations have been developed such as parallel coordinates, treemaps, glyph and pixel-based visual data representations, to name just a few, together with a variety of techniques to reduce display clutter[41]. There are also special techniques for visualising structured data, such as graph-based approaches for networks, and for visualising spatial and temporal dimensions as found in geo-visualisation (described later in more detail). Furthermore, some visualisations make use of automatic data analysis techniques such as clustering or dimensional reduction as a preprocessing step prior to visualisation.

Data Management

The efficient management of data of various types and qualities is a key component of visual analytics, as it typically provides the input of the data, which is to be analysed. Generally, a necessary precondition to perform any kind of data analysis is an integrated and consistent database. Database research has, until the last decade, focused mainly on aspects of efficiency and scalability of exact queries on uniform, structured data. With the advent of the Internet and the easy access it provides to all kinds of diverse data sources, the focus of database research has shifted towards integration of this heterogeneous data. Finding effective representations for different data types such as numeric data, graphs, text, audio and video signals, semi-structured data, semantic representations and so on is a key problem of modern database technology. But the availability of heterogeneous data not only requires the integration of many different data types and formats but also necessitates data cleansing - such as dealing with missing and inaccurate data values. Modern applications require such intelligent data fusion to be feasible in near real-time and as automatic as possible. Also, new forms of information sources such as streaming data sources, sensor networks or automatic extraction of information from large document collections (e.g., text, HTML) result in a difficult data analysis problem; supporting this is currently the focus of database research[124]. Data management techniques increasingly make use of intelligent data analysis techniques and also on visualisation to optimise processes and inform the user.

Diverse data from the Internet imposes novel challenges to database research.

Data Mining

The discipline of data mining develops computational methods to automatically extract valuable information from raw data by means of automatic analysis algorithms[75]. There are various approaches; one is supervised learning from examples, where, based on a set of training samples, deterministic or probabilistic algorithms are used to learn models for the classification (or prediction) of previously unseen data samples. Decision trees, support vector machines and neural networks are examples of supervised learning. Another approach is unsupervised learning, such as cluster analysis[54], which aims to extract structure from data without prior knowledge being available. Solutions

Data mining: automatic extraction of valuable information from raw data

in this class are employed to automatically group data instances into classes based on mutual similarity, and to identify outliers in noisy data during data preprocessing. Other approaches include association rule mining (analysis of co-occurrence of data items) and dimensionality reduction. While data analysis was initially developed for structured data, recent research aims at analysing semi-structured and complex data types such as Web documents or multimedia data. In almost all data analysis algorithms, a variety of parameters needs to be specified, a problem which is usually not trivial and often needs supervision by a human expert. Interactive visualisation can help with this, and can also be used in presenting the results of the automatic analysis – so called 'visual data mining'.

Spatio-temporal Data Analysis

Finding relations and patterns in spatial and/or temporal data requires special techniques

Spatial data, is data with references in the real world, such as geographic measurements, GPS position data, and data from remote sensing applications; essentially, data that can be represented on a map or chart. Finding spatial relationships and patterns within this data is of special interest, requiring the development of appropriate management, representation and analysis functions (for example, developing efficient data structures or defining distance and similarity functions). Temporal data, on the other hand, is a function of time, that is the value of data variables may change over time; important analysis tasks here include the identification of patterns, trends and correlations of the data items over time. Application-dependent analysis functions and similarity metrics for time-related data have been proposed for a wide range of fields, such as finance and engineering.

Scale and uncertainty impose challenges on spatio-temporal data analysis

The analysis of data with references both in space and in time, spatial-temporal data, has added complexities of scale and uncertainty. For instance, it is often necessary to scale maps to look for patterns over wide and also localised areas, and similarly for time, we may wish to look for trends that occurs during a day and others that occurs on a yearly basis. In terms of uncertainty, spatio-temporal data is often incomplete, interpolated, collected at different times or based upon different assumptions. Other issues related to spatial-temporal data include complicated topological relations between objects in space, typically very large datasets and the need for specialised data types. In addition, more and more geo-spatial data is now accessible to non-expert communities and these 'analysts' need tools to take advantage of this rich source of information.

Perception and Cognition

Design of user interfaces needs to take perception and cognition into account

Perception and cognition represent the more human side of visual analytics. Visual perception is the means by which people interpret their surroundings and for that matter, images on a computer display. Cognition is the ability to understand this visual information, making inferences largely based on prior learning. The whole system is extremely complex, and it has taken decades

of research in fields such as psychology, cognitive science and neuro-science to try to understand how the visual system achieves this feat so rapidly. For many years it was thought that 'seeing' was a generally passive activity with a detailed 'map of the world', whereas now we recognise that it is very active, only searching for and selecting visual information, which is pertinent to the current task. Knowledge of how we 'think visually'[123] is important in the design of user interfaces and together with the practical experience from the field of human computer interaction, will help in the creation of methods and tools for design of perception-driven, multimodal interaction techniques for visualisation and exploration of large information spaces, as well as usability evaluation of such systems[36, 100].

Visual analytics relies on an efficient infrastructure to bind together many of the functions supplied by the various disciplines, in order to produce a coherent system. In addition, evaluation is critical in assessing both the effectiveness and usability of such systems. We will now consider these enabling technologies.

Infrastructure

Infrastructure is concerned with linking together all the processes, functions and services required by visual analytic applications so they work in harmony, in order to allow the user to undertake their data exploration tasks in an efficient and effective manner. This is difficult as the software infrastructures created by the different technologies are generally incompatible at a low level and this is further complicated as one of the fundamental requirement of visual analytics applications is high interactivity. For this reason, most visual analytics applications are currently custom-built stand-alone applications, using for example, in-memory data storage rather than database management systems. The design of system and software architectures is paramount in enabling applications to successfully utilise the most appropriate technologies. In addition, the reuse of many common components will result in applications being more adaptable and built much quicker.

Appropriately designed infrastructure is vital to the success of visual analytics

Evaluation

Researchers and developers continue to create new techniques, methods, models and theories, but it is very important to assess the effectiveness, efficiency and user acceptance of these innovations in a standard way, so they can be compared and potential problems can be identified. However, as demonstrated in Chapter 8, evaluation is very difficult given the explorative nature of visual analytics, the wide range of user experience, the diversity of data sources and the actual tasks themselves. In the field of information visualisation, evaluation has only recently become more prominent[13]. It has been recognised that a general understanding of the taxonomies regarding the main data types and

Rigorous assessment of current and innovative solutions across all disciplines is imperative

user tasks[4] to be supported are highly desirable for shaping visual analytics research.

The current diversification and dispersion of visual analytics research and development has focused on specific application areas. While this approach may suit the requirements of each of these applications, a more rigorous and scientific perspective based on effective and reproducible evaluation techniques, will lead to a better understanding of the field and more successful and efficient development of innovative methods and techniques.

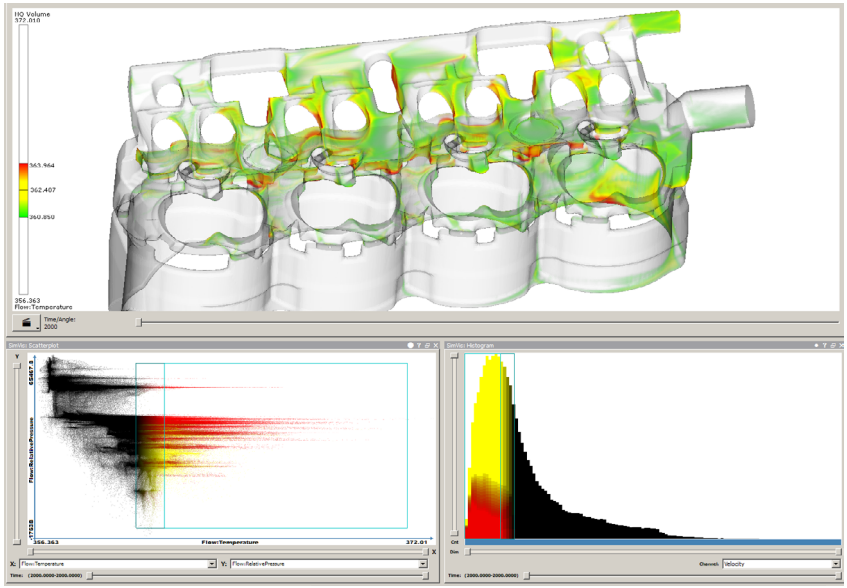

Figure 2.5: Visual analytics in action: Interactive visual analysis of a cooling jacket simulation. User has focused on critical regions of high temperatures and low flow velocities by brushing the two views (velocity histogram and temperature versus relative pressure) as they may indicate locations of insufficient cooling. Dataset is courtesy of AVL List GmbH, Graz, Austria; Interactive Visual Analysis © SimVis GmbH, 2010

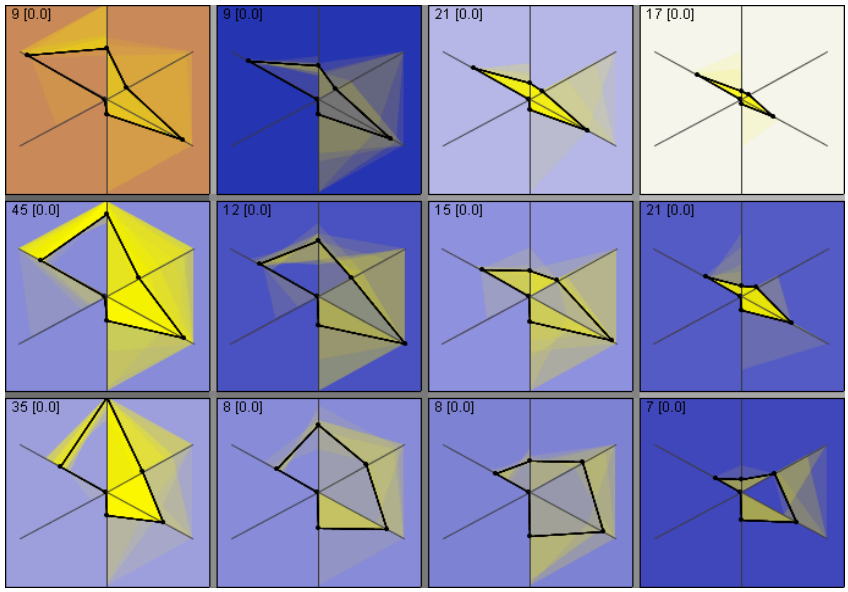

Figure 2.6: Visual analytics in action: Helping demography researchers to effectively analyse multivariate datasets. Six-dimensional demographic dataset was clustered into twelve groups, and the distribution shown by radial parallel coordinate plots. Yellow opacity bands illustrate the variance within the individual clusters and background colour coding correlates cluster with a specific target variable. Technique by Bak et al.[10]

3 Data Management

3.1 Motivation

One of the most exciting opportunities of the emerging Information Age is to extract useful findings from the immense wealth of data and information acquired, computed, and stored by modern information systems. This is witnessed by both professionals and single users that every day extract valuable pieces of information from very different kinds of data sources, e.g., files and emails on their laptops, data coming from their company databases, or data available on the Internet.

The big opportunity of the Information Age

Unfortunately, as described in Chapter 1, there are many obstacles, which impede the effective exploitation of such an opportunity: users and analysts may get overwhelmed by irrelevant, or inappropriately processed or presented information – the information overload problem.

Obstacles come from the fact that datasets are often very large and growing incrementally, data sources are heterogeneous and are typically distributed. As a consequence, it is necessary to take this into account when studying, assessing, and giving recommendations about techniques for managing data. This is particularly challenging and the following issues need to be considered:

Many obstacles need to be overcome

- **Heterogeneity of data sources.** In a number of applications, it is necessary to integrate and query data coming from diverse data sources; this is inherently difficult and not especially well researched. Logic based systems, balancing expressive power and computational cost, represent the state of the art solutions; however such approaches are neither well understood nor easy to use.
- **Different data types.** Data comes in a variety of types and with different structures. It is challenging to analyse, in an integrated fashion, numeric and non-numeric data, together with images, videos, models, and data presenting particular entities as found in geographic and temporal data (as discussed in more detail in Chapter 5).
- **Data streams.** In many application areas, the data is in the form of streams, that is, data from a source that frequently produces new pieces of information (sensor data, stock market data, news data, etc.). Further investigation is required to deal with the conceptual and technical issues.
- **Working under pressure.** In some applications, such as emergency management, the analytical process must be performed as quickly as possible in order to make timely critical decisions. In such cases, 'classical' data management flow methods, involving data experts are not appropriate and 'traditional' data activities like data querying, cleaning, integration, etc. need to be accelerated.

- **Time consuming activities.** Managing different data formats or measurement units, null values, column names, etc. can be a complex and time consuming activity, even with small and simple datasets.

In the last decades, significant research effort has been directed towards managing and exploring large amounts of data, and two robust disciplines have emerged: data management and visual analytics.

Data management ensures data consistency and standards

Data management is a well understood field, researched over the past 30 years, and provides methods for effectively dealing with large datasets. The techniques aim to ensure data consistency, avoiding duplication and handling data transactions in a formal way. They rely on a common and well understood model, the relational model, useful to exchange and integrate data, and they exploit a highly optimised and standardised data access interface, which is called the SQL query language.

Figure 3.1: Visual analytics: a visionary scenario. Excerpt from the VisMaster Video, http://videotheque.inria.fr/videotheque/doc/635

Visual analytics is interactive and allows for exploratory analysis

Visual analytics has emerged only recently compared to the related topics of information visualisation and data mining. The advantages of visual analytics are that it deeply involves the user in the analysis loop, exploiting his perceptive and cognitive capabilities. It can be employed in a dynamic manner, with quick visual interaction and switching of analysis paradigms, and it is intended for exploratory analysis, especially when the goals are not clearly defined.

However, in spite of the strong advances in these two synergetic fields, a big gap exists between them, which obstructs the integration of these two disciplines. The main issues are:

- **Dynamicity.** Classical data management activities rely on the relational model and on the SQL query language and are highly optimised for a simple and inherently static two step interaction: query formulation and collecting results. With large datasets (billions of items), this approach is

unlikely to provide the response (approximately 100msec) necessary for good interaction[99].
- **Standards**. While data management is based on well-known and accepted standards (i.e., the relational model and the SQL query language) visual analytics applications tend to access and handle data in a proprietary way, lacking a shared, proven, and efficient solution.
- **User interaction life-cycle.** From the end user's point of view, who is only interested in finding information, data management interactions are essentially single user, quick, and one shot: the user expresses a query against the data, collects the results and analyses it. In contrast to this, visual analytics activities last a long time and may involve several users. Thus, assistance is required for long-term activities and collaborative work that are currently poorly supported by classical data management techniques.

Integration of data management and visual analytics is important, as illustrated by this scenario

The following scenario illustrates a tight integration of data management and visual analytics capabilities. It describes the research activities of several doctors working in different hospitals across Europe.

Doctors are coordinating their efforts to achieve a better understanding of several new allergy cases that have been reported in different European cities. The new allergy mainly affects the hands of 5-9 year old children and while it is very irritating it is not serious: it resolves itself spontaneously in about two weeks, or in a few days if treated with a common antihistamine. What puzzles the doctors is that the disease appeared at the same time in different locations across Europe and that a reasonable explanation is not available.

Integration and analysis of different data sources

A smart integration engine allows for seamless integration of data coming from different sources and in different formats, including patients' location and personal data, pictures about the allergy, notes from doctors, and case histories. Several interactive visualisations are available in the system, tightly integrated with automatic analytical tools. Exploring the data structure and content, helps doctors in choosing the most appropriate ones.

Collaboration among users

Data and findings are shared among the doctors, and the system allows for collaborative work and for saving and reopening the analytical processes. Using such a system, the doctors are able to select all the cases that belong to the new allergy, discarding similar but not related situations. After that, they start to investigate the environment in which the children live, searching for some common patterns (alimentation, dressing, pollution, climatic situation, etc.). Again, this requires new complex integration activities and analysis tools. After two weeks of research they conclude that there are not relevant similar patterns.

New analysis directions

One doctor starts to compare the temporal evolution of the allergy and its response to medicines by linking to a large medical dataset describing allergy cases. He discovers a strong pattern similarity with some relatively rare contact allergies generated by a kind of rigid plastic largely used for toys and food containers; this allergy usually manifests itself after prolonged contact with the substance. The doctor shares these findings through the system, but some research on toys and food containers fail to find that substance. Another doctor points out a fact that is rather obvious but has previously gone unnoticed: while

the allergy affects both right and left hands, most cases involve the right hand. A quick analysis reveals that the less frequent left hand cases correspond to left-handed children. The analysis moves again to the alimentation of the children focusing, this time, not on the food components but on the plastic associated with the food (i.e., boxes, bags, etc.) and on the probability of touching plastic parts.

The cause is discovered

Eventually a doctor discovers that a European company is marketing a new brand of lollipop, quite popular among children, and that the lollipop's plastic stick contains the allergenic component.

To summarise, nowadays, analysts and end users have the opportunity of extracting useful pieces of information from a wealth of data. However, several obstacles stand in the way and we have seen how data management and visual analytics need to address different, and sometimes complementary, facets of the problem. In order to effectively exploit this challenging situation, an integration between these two approaches is required, reducing the gap that exists between them. Such a process requires the solution of several theoretical and practical issues that, if not adequately addressed, could seriously compromise the opportunity that the new Information Age offers.

3.2 State of the Art

3.2.1 Data Management

This section focuses on the main research fields active in the context of data management, emphasising activities and results that are particularly relevant for visual analytics; aspects associated with visualisation issues will be discussed in Section 3.2.2.

Relational Technology

The relational technology[44] is based on research from the 1970s: Ted Codd's visionary paper introduces the relational model and the System R research project at IBM's San Jose Research Lab, in which the SQL query language appeared. In the relational data model, data is represented in tables that are connected to each other by attribute values, without any explicit navigational link in the data. The flexibility offered by this feature and SQL meant that the relational model rapidly replaced the now largely obsolete network and hierarchical data models.

Relational DBMSs dominate the market

Nowadays, relational systems dominate the market and rely on a very mature computer science technology. Modern RDBMSs (Relational Database Management Systems) allow for accessing the data in a controlled and managed fashion. They present a clear separation between data structure and content, and incorporate robust means of handling security and data consistency that is ensured by arranging data management in Atomic, Consistent, Isolated,

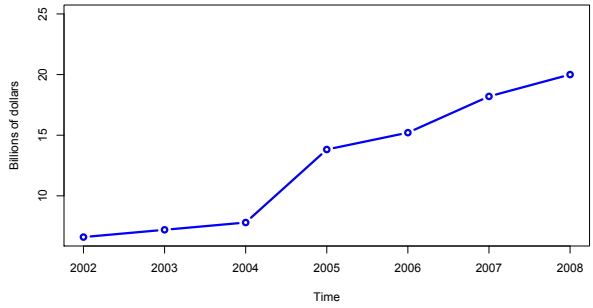

Figure 3.2: Purchases of relational database licenses in the last years (in billions of $)

and Durable transactions (so called transactions' ACID property). This permits seamless concurrent data access and data recovery in a collection of databases that is physically distributed across sites in a computer network (Distributed RDBMS), hiding the distribution details from the users that access the data through a common interface, using the widely accepted SQL query language. A coherent set of theorems and practical research on query optimisation and data indexing allows relational systems to deal with very large datasets.

The market of RDBMS is still growing: the worldwide sales of new licenses of relational database management systems (RDBMS) totalled about $20 billion dollars in 2008, increasing about three times the 2002 revenue of $6.6 billions, according to Gartner, Inc. as shown in Figure 3.2.

Data Integration

Providing unified and transparent access to a set of heterogeneous sources

Data integration is the problem of providing unified and transparent access to a set of autonomous and heterogeneous sources, in order to allow for expressing queries that could not be supported by the individual data sources alone. There is a big and still growing need for systems and techniques that support such a process, and very likely it is one of the major challenges for the future of IT. The problem is ubiquitous in modern software systems, and comes in different forms: data sources are characterised by a high degree of heterogeneity (e.g., different data models, different data types, different nomenclature, different data units, etc.), raising many challenges, and a number of methodologies, architectures, and systems have been developed to support it.

Data integration can be centralised, that is being performed within the same organisation (e.g., Enterprise Information Integration) or can be decentralised, involving two or more organisations, usually based on a peer-to-peer architecture. The latter assumes a data-centric coordination among the autonomous

organisations to dynamically expose a view of their data using an agreed data schema.

The integration can be virtual or materialised. In the first case, the data does not move from the original source and the integration is performed at query time; in the second case chunks of data are physically exchanged before the query process and collected in a single place (e.g., data warehousing).

The most relevant approach for visual analytics is the centralised, virtual information integration that represents an evolution of ideas dating back to the 80s. A collection of theoretical results is available, but a robust and definitive solution is still far from being reached. The available solutions foresee several tools for data source wrapping and database federation (e.g., DB2 Information Integrator), providing a common model for exchanging heterogeneous data and allowing physical transparency (i.e., masking from the user the physical characteristics of the sources), handling heterogeneity (federating highly diverse types of sources), preserving the autonomy of the data sources, and ensuring scalability (distributed query optimisation).

Semantic integration

However, these tools do not provide conceptual data transparency, i.e., they present the data as it is stored within the sources, leaving the heterogeneity arising from different naming, data representation, etc., unsolved. The most promising solution to this problem is called semantic integration[23] and is based on the idea of computing queries using a logic based engine that exploits a conceptual view of the application domain (i.e., an ontology), rather than a flat description of the data sources. Such a description, called a global schema, is independent from the sources that are mapped through a logic language into concepts of the global schema. A solution that is being adopted more often is to use, as a logic language the so called 'description logics' that are a subset of the first order logic and balance expressive power and computational cost.

Data Warehousing, OLAP and Data Mining

Data warehousing, OLAP (On-Line Analytical Processing), and data mining share many of the goals of visual analytics: they are intended for supporting, without the explicit use of visualisations, strategic analysis and decision-supporting processes.

Data warehousing for decision making

A data warehouse[62] is an integrated repository of data that can be easily understood, interpreted, and analysed by the people who need to use it to make decisions. It is different from a classical database for the following reasons: it is designed around the major entities of interests of an organisation (e.g., customers, sales, etc.), it includes some related external data not produced by the organisation and it is incremental, meaning that data, once added, is not deleted, allowing for analysing temporal trends, patterns, correlations etc. Moreover it is optimised for complex decision-support queries (vs. relational transactions). The different goals and data models of data warehousing

stimulated research on techniques, methodologies and methods, which differ from those used for relational DBMS.

The term OLAP[31] refers to end-user applications for interactive exploration of large multidimensional datasets. OLAP applications rely on a multidimensional data model created to explore the data from different points of view through so called data cubes (or data hypercubes), i.e., measures arranged through a set of descriptive categories, called dimensions (e.g., sales for city, department, and week). Hierarchies are defined on dimensions, (e.g., week ... month ... year) to enable additional aggregation levels. A data cube may hold millions of entries characterised by tens of dimensions and one of the challenges is to devise methods that ensure a high degree of interactivity. One solution is to pre-compute and store aggregated values for different levels of the hierarchies and reduce the size of the data (see below), thus sacrificing precision for speed. Another consideration is system usability. The user can only explore a small number of dimensions at any one time (i.e. the hypercube needs to be projected onto two-dimensional or three-dimensional spaces) and hence to gain insights into high dimensional data may require long and sometimes frustrating explorations.

Mining insights from large datasets

Data mining is the process of discovering knowledge or patterns from massive amounts of data through ad hoc algorithms. Data mining can be categorised based on the kinds of data to be analysed, such as relational data, text, stream, Web data, multimedia (e.g., image, video), etc. Its relationship with visualisations became more prevalent in the 90s when the term 'visual data mining' emerged, denoting techniques for making sense of data mining algorithms through different visualisations, built on both the mined data and on the results produced by the algorithms. The topic of data mining is further discussed in Chapter 4.

Data Reduction and Abstraction

In the context of data management, data reduction techniques have been used to obtain summary statistics, mainly for estimating costs (time and storage) of query plans in a query optimiser. The precision is usually adequate for the query optimiser and is much cheaper than a full evaluation of the query.

Data reduction can improve query optimisation and interaction

More recently the focus has moved onto data reduction techniques to improve the interactivity for OLAP applications operating on large amounts of data stored in the organisation data warehouse. Due to the analytical and exploratory nature of the queries, approximate answers are usually acceptable.

In summary, the purpose of data reduction in the context of data management is to save computational or disk access costs in query processing or to increase the systems responsiveness during interactive analysis. Data reduction relies on various techniques, like histograms, clustering, singular value decomposition, discrete wavelet transforms, etc. A comprehensive summary of data reduction techniques for databases can be found in the New Jersey data reduction

report[11]. Data reduction techniques can be usefully exploited in the context of visual analytics by reducing the number of dimensions and/or the complexity of relationships.

Data Quality

Databases often have to deal with data coming from multiple sources of varying quality - data could be incomplete, inconsistent, or contain measurement errors. To date, several research lines and commercial solutions have been proposed to deal with these kinds of data errors, in order to improve data quality.

Linking different views of the same data

Data conflicts have been studied by statisticians that needed to resolve discrepancies rising from large statistical surveys. One of the first problems of this kind was the presence of duplicated records of a person[43], and the devised practical and theoretical solution, called record linkage, allowed the collection and linkage of all the related data records, producing a unique and consistent view of the person. It was quickly understood that record linkage was only one of a larger set of problems, such as wrong, missing, inaccurate, and contradicting data, and in the late 1980's, researchers started to investigate all problems related to data quality. This line of research was advanced by both the increasing number of scientific applications based on large, numerical datasets and by the need to integrate data from heterogeneous sources for business decision making.

Restoring missing data

The problem of missing data was initially studied in the context of scientific/numerical datasets, relying on curative methods and algorithms able to align scientific data. More recently, the focus has moved on to non-numerical data and in particular, dealing with inherently low quality datasets such as information extracted from Web and sensor networks. MystiQ[19] is an example of research into building general purpose tools to management uncertain data.

Polishing the data

Dealing with missing data and duplicate records is only part of the overall process of data cleansing. We also need to identify and either correct or reject data that is incorrect or inaccurate, possibly through the use of aggregate data statistics. Additionally, the data many need to be standardised by, for example, adjusting the data format and measurement units.

3.2.2 Data Management and Information Visualisation

The data management research field acknowledges the key role that information visualisation can play in enhancing data management activities through ad hoc visualisation. In the following section, we describe some examples, which show the synergy that exists between data management and information visualisation.

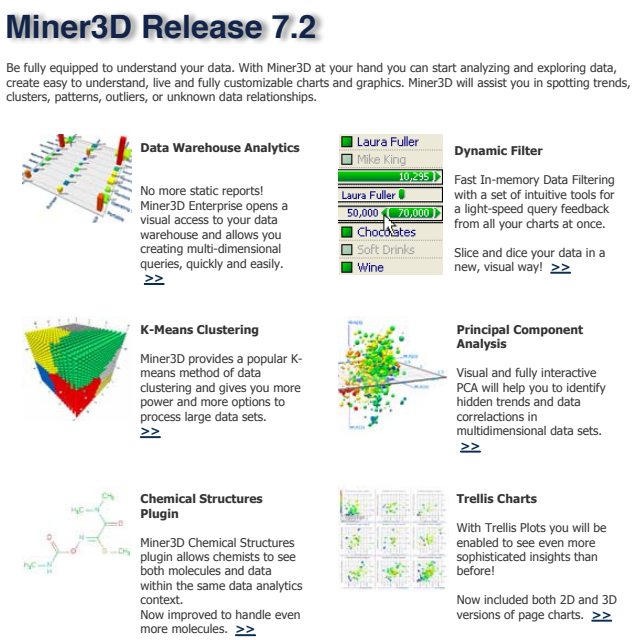

Figure 3.3: The commercial system Miner3D

Visual Data Mining

Using visualisation to enhance data mining

The inherent difficulties associated with data mining algorithms together with the need to justify the data mining results, has stimulated the development of integrated environments in which suitable visualisations are used as a complementary technique to support data mining. The combination of visualisation and data mining is known as 'visual data mining'.

This new field presents strong correlation with several synergic fields, i.e., knowledge discovery in databases, information visualisation and human-computer interaction. Common elements have been recognised that need to be specified when developing methodologies for visual data mining. These include: a) the initial assumptions posed by the respective visual representations; b) the set of interactive operations over the respective visual representations and guidelines for their application and interpretation, and c) the range of applicability and limitations of the visual data mining methodology. Several research projects have dealt with this new challenging research field, like the 3D Visual Data Mining (3DVDM) project in Aalborg University representing the dataset in various stereoscopic virtual reality systems. Commercial products such as VMiner3D[35] (see Figure 3.3), demonstrate the usefulness of the combination of data mining and visualisation.

Visual OLAP

A clear trend in business visualisation software exists, showing a progression from basic and well-understood data visualisations to advanced ones, and this is the case of OLAP applications that present emerging techniques for advanced interaction and visual querying.

To date, the commonly used OLAP interfaces enhance the traditional way of arranging data on a spreadsheet through automatic aggregation and sort functions that store the result in a second table (called a pivot table). This allows the end user to explore data cubes through traditional visualisation techniques such as time series plots, scatterplots, maps, treemaps, cartograms, matrices etc., as well as more specialised visualisations such as decomposition trees and fractal maps. Some applications also integrate advanced visualisation techniques developed by information visualisation researchers.

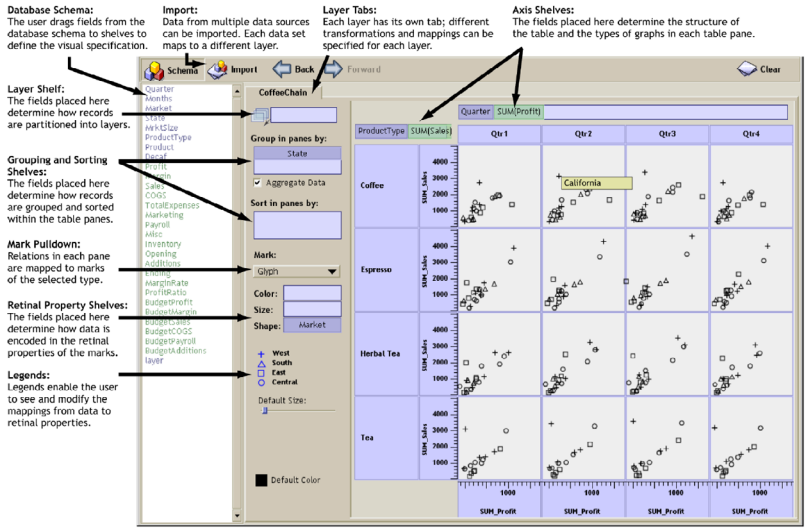

Figure 1: *The Polaris user interface. Analysts construct table-based displays of relational data by dragging fields from the database schema onto shelves throughout the display. A given configuration of fields on shelves is called a visual specification. The specification unambiguously defines the analysis and visualization operations to be performed by the system to generate the display.*

Figure 3.4: The Polaris interface as presented in the original paper[107] © 2002 IEEE

Two pioneering systems in visual online analytical processing

Polaris and ADVIZOR are among the first attempts in such a direction. Polaris is a visual tool for multidimensional analysis developed at Stanford University[107] and inherits the basic idea of the classical pivot table interface, using embedded graphical marks rather than textual numbers in the table cells. The types of supported graphics are arranged into a taxonomy, comprising of rectangle, circle, glyph, text, Gantt bar, line, polygon and image layouts (See Figure 3.4). Currently, Tableau Software commercialises the pioneering Polaris work. ADVIZOR represents the commercialisation of 10 years of research in Bell Labs on interactive data visualisation and in-memory data management[39].

It arranges multidimensional data from multiple tables onto a series of pages, each one containing several linked charts (from 15 chart types), which facilitates ad hoc exploration of the data.

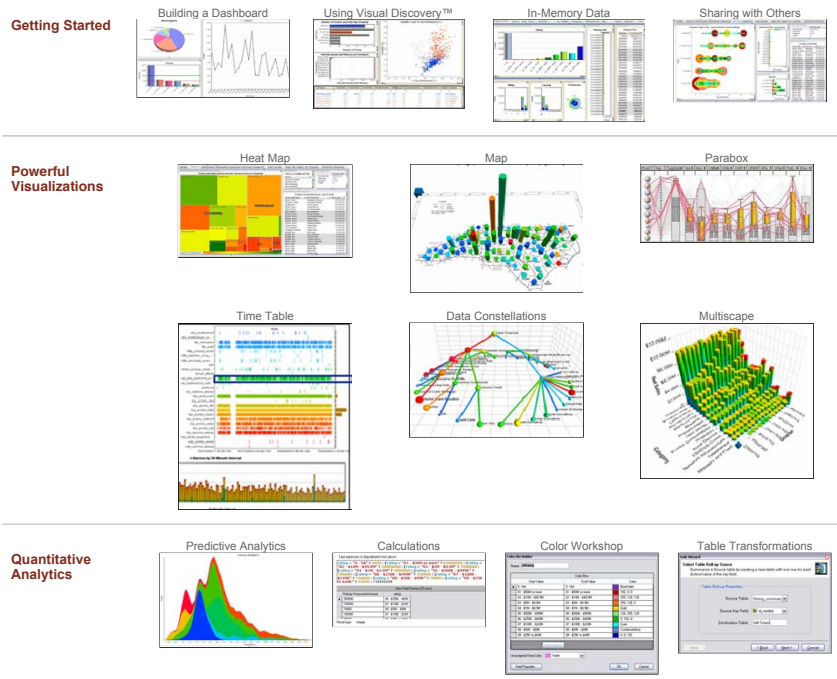

Figure 3.5: The rich set of linked visualisations provided by ADVISOR include barcharts, treemaps, dashboards, linked tables and time tables

Visual Data Reduction

Scaling visual presentations to high dimensional data or to billions of records is one of the challenges that visual analytics has to deal with, and requires a tight collaboration between data management and visual analytics. Visualising large datasets, often produces cluttered images, and hence various clutter reduction techniques[41] have been developed by the information visualisation community to ameliorate this. However, as mentioned previously, visual analysis is limited if the system does not allow for quick interaction while exploring the data. This requires new scalable data structures and algorithms for data reduction and/or innovative hierarchical data management. While several proposals are available, e.g., sampling, density-based hierarchical data aggregation and multi-resolution representation, a common understanding on how to interactively visualise vast datasets does not exist.

Visualising large datasets often produces cluttered images

Some clutter reduction techniques are fairly complex (e.g., dimension reduction), while other ones are relatively straightforward (e.g., uniform sampling).

However, disregarding the technical details and the computational aspects, all of them share a set of common problems:

- When is the technique needed?
- Which technique is the most appropriate for a given dataset and visualisation?
- To which extent do we have to apply it (e.g., how much do we have to sample)?
- How can we evaluate the quality of the reduced dataset?

Measuring the visualisation quality with quality metrics

Addressing these questions highlights some complicated issues, involving the data structure and cardinality, the user task, the chosen visualisation as well as perceptual and cognitive aspects. Several proposals addressed these issues using objectives *quality metrics* that capture some relevant data reduction measure. For example, the number of data items being displayed on a scatterplot can be used as a trigger to decide *when* to apply a reduction technique (e.g., sampling), *how much* to sample, and *to compare* the final result with another technique (e.g., data clustering). Obviously, several, non trivial parameters affect this process, like threshold values, data distribution, image size, etc. However, the matter still deserves further study as several issues are still far from being solved, such as how to assess the quality of a visualisation produced from the application of a data reduction technique.

Figure 3.6: Visual data reduction preserving density differences through visual quality metrics. The image shows a curative sampling algorithm applied to the topmost scatterplot producing a more comprehensible image (below). The measurements on the right allow the user to formally assess the performance of the algorithm

Data reduction techniques belong to two different categories:

- those that use quality metrics to optimise non visual aspects, e.g., time, space, tree balancing, etc;

- those that use quality metrics to optimise the *visualisation* of some data aspects relevant to the analytical process. We call this activity *visual data reduction*.

Visual data reduction perfectly fits the visual analytics philosophy: a) an automated analysis is performed on the data, discovering and measuring different relevant data aspects (e.g., strong correlation, outliers, density differences) and b) such measures are used as quality metrics to drive and evaluate the data reduction activity, presenting the end user with the visualisation that best conveys these interesting data aspects. But we must ensure that these facets are effectively *presented* to the end user as image degradation or perceptual issues could hide the precious insights highlighted by the automatic analytical process.

Visual quality metrics: size, visual effectiveness and feature preservation

Quality metrics adopted in the visual data reduction process must encompass data issues and visualisation issues in a strongly integrated fashion. In an example of this[17] the authors categorise these *visual quality metrics* in three classes: size metrics, visual effectiveness metrics, and feature preservation metrics. Size metrics have a descriptive character (e.g., number of data points) and are the basis for all further calculations. Visual effectiveness metrics are used to measure the image degradation (e.g., collisions, occlusions, etc.) or outliers. Examples are data density, collisions, and the data-ink-ratio. Feature preservation metrics are the core of visual quality metrics and are intended for measuring how correctly an image represents some data characteristics, taking into account data issues, visual aspects and perceptual aspects. Figure 3.6 illustrates such a feature preserving data reduction technique[16] (non uniform sampling) for 2D scatterplots, driven and evaluated by visual quality metrics that use data, visualisation and perceptual parameters, with the main goal of discovering and preserving data density differences.

Visualisation for the Masses

Visualising data on the Web has improved considerably in the last few years

Tools supporting search, browsing, and visualisation have dramatically improved in the past decade, so that it is possible to design and implement new Web based system integrating data management technologies and interactive information visualisation. An example of these new opportunities comes from the Swedish non-profit company Gapminder[1], acquired by Google that designed a system to make world census data available to a wider audience. The relational data of the census can be explored with two linked visualisations, a geographical map plus a two-dimensional scatterplot that uses colour and size to visualise two additional attribute values (see Figure 3.7). The system is easy to use and by allowing the user to explore the change of the variables over time, its effectiveness is enhanced considerably.

In summary, data management and visual analytics are two disciplines that, together, are able to exploit the opportunities coming from the Information Age. This survey of the state of the art shows, however, that while strong results have

[1] `http://graphs.gapminder.org/world/`

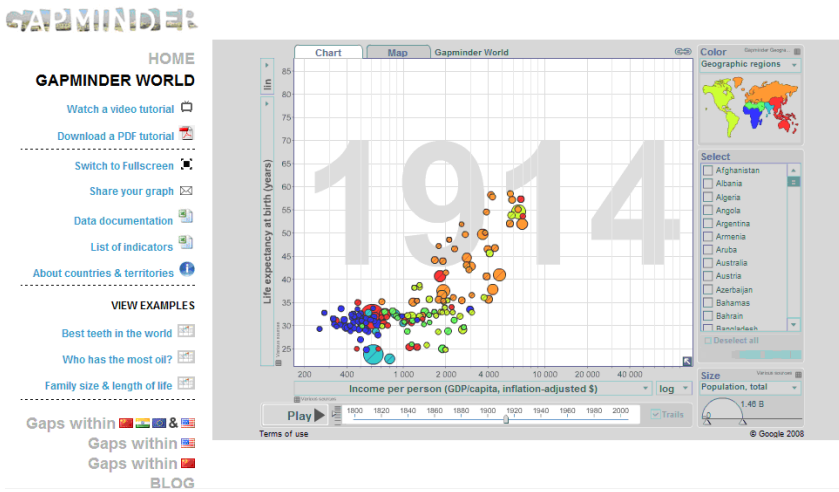

Figure 3.7: The GapMinder visualisation clearly shows the correspondence that exists between life expectancy at birth and income per person. Moreover, it shows the dramatic differences that exist between India, North Europe and North America

been reached within these two fields there are still some unexplored issues and some requirements that have not been addressed that impede the integration of fields. In particular, we need better logic based integration systems, a tighter integration between visualisation and data, more precise quality indicators, visual oriented data reduction techniques, and new interaction paradigms for large sets of users. In order to progress, it is mandatory to make these two disciplines more aware of each other's results and needs, addressing both common problems and specific issues.

3.3 Challenges and Opportunities

Symbiotic dependency of visual analytics and data management

In this section, we highlight the most important challenges that data management and visual analytics have to deal with to better exploit the opportunity of the Information Age. As, on the one hand, data management is intrinsic to visual analytics, solving some of data management's open problems will enhance visual analytics applications. On the other hand, specific visual analytics issues will pose new challenges for the data management community. Therefore, it is important to reflect upon the different perspectives and the inherent relationships when considering the mutual roles played by data management and visual analytics.

Moreover, as a general consideration, it is worth noting that some critical visual analytics issues are the subject of research effort in different scientific areas, often with different names. For example, the activity of data sampling is performed with different goals in both data management and information

visualisation research activities, but a shared view of problems and techniques between these two research fields does not exist. Therefore, it is essential to find and encourage such potential synergies.

Uncertainty

How to visualise missing data?

Solving issues resulting from incomplete, inconsistent, or erroneous data is crucial for both visual analytics and data management. Therefore, both robust and agreed methodologies are required. However, visual analytics looks at these issues in a different way and the straightforward adoption of the solutions proposed in the data management field could be either a valid solution or an obstacle to the analysis process. For example, assume that we are dealing with a missing or erroneous value. The data management techniques may use some curative algorithms, providing an alternative (e.g., interpolated or statistically computed) value for the bad data, but this solution can hide important facts; perhaps the value is empty because a person omitted to enter a value on the form to evade paying tax or an out of range value indicates a faulty sensor?

Data visualisation also has methods of dealing with missing data and so it has to be decided whether data management or the visualisation has responsibility for managing this. Whatever subsystem takes charge, it is necessary to remember the decisions made during the cleaning activities so that the user can be made aware of any uncertainties in the data.

Data Integration

Need for new integration systems

The integration of heterogeneous data is a core data management activity and its importance and use are increasing. Logic based systems, balancing expressive power and computational cost represent state of the art solutions. Visual analytics can greatly benefit from such an approach and does not raise particular issues in such a context, apart from situations that require quick decision making (e.g., emergency management) or upon performing data integration without expert support. In such cases, the integration engine should present an interface intended for non expert users and be able to make decisions with incomplete information, e.g., while integrating data coming from the Web. This is a new and challenging requirement, not addressed by traditional data management research.

Semantics Management

Making semantics a first class citizen

Associated with data integration activities, is the need for managing all the data semantics in a centralised way, for example, by adding a virtual logic layer on the top of the data itself. For example, data semantics could be used to describe synonyms such as 'is-a' relationships (e.g., a student is-a person and, as a consequence, everything holds for person holds for a student as well)

and constraints (e.g., you must be at least 18 years old to hold an Italian car driving license). This is useful not only for data integration, but also for dealing with all the semantic issues involved in the analytical process, like metadata management, abstraction levels, hierarchical structures and externalisation. Visual analytics applications should manage all the available semantics at one point, under the responsibility of the database management system. That includes also the semantics that are discovered during analytical (manual and automatic) activities – once discovered it should be added to the top virtual logic layer.

Such a challenging kind of semantic integration has been not researched in both visual analytics and data management fields and could represent an interesting starting point for cooperation between the two disciplines. This also represents a strong opportunity: semantics discovered during the analytical process could be usefully exploited for improving database activities and database performances, e.g., improving data cleaning and the query optimisation process.

Data Provenance and Integrity of Results

Where does the data come from?

While performing visual analytics and data management activities, the end user may need to inspect and understand the path associated with some data items. That includes, at least, a) the physical source hosting the original data, b) the reason why the data belongs to the observed dataset (that is useful when the query process is performed using some kind of logical deduction), and c) a way for better understanding the results of an automatic analysis step (e.g., a data mining classification or a record linkage). Moreover, while performing long and complex visual analytics activities, it could be useful to view the series of actions that generated the current display, i.e., what data and what transformations have been used to generate the actual visualisation?

Data Streaming

Visual analytics applications sometimes have to deal with dynamic data (i.e., new data is received on a regular basis) whilst the analysis process is running. For instance, a visual analysis of a social network, based on a live feed of their data. The analysis has to gracefully adjust to the updates; stopping the process and triggering a total re-computation would not be appropriate.

Continuous flows of data require special study

The following three aspects of data streams require further study, at a conceptual and technical level, in order to address the visual analytics and data management objectives:

- **Building data stream management systems.** That implies studying architectures and prototypes, stream-oriented query languages and operators, stream processing and efficient algorithms to keep an up-to-date online connection to the data sources.

- **Designing efficient algorithms for stream analysis.** In particular, we need algorithms that are able to proceed in an incremental way, mining information from the stream and capturing both trends and overall insights.
- **Change detection analysis.** Sometimes the analysis looks for relevant changes that happen within the stream, allowing for the fast detection of new or unexpected behaviours.

Time Consuming Low Level Activities

Managing diverse data types can be time consuming for the analyst

Data management and visual analytics problems are not always due to the large size of the dataset. Dealing with small details such as data heterogeneity, data formats and data transformation can be a time consuming and hence an unwelcome burden on the analyst. In these cases, new consistency checking languages could offer assistance, relieving the analyst of coding in SQL. In general, there needs to be a better comprehension of the role of low-level data management activities in the visual analytics process.

Further time consuming activities include selecting the appropriate view on the data, joining relevant tables and data sources, selecting the best visualisation for exploring the dataset, and associating data values to visual attributes. These call for some form of automation, which is able to assist the analyst in speeding up the overall analysis process. This issue is strongly connected with heterogeneous data integration and semantics management, as mentioned earlier, and researchers should address logic based solutions, capturing both predefined and discovered semantics in order to drive automatic or semi-automatic algorithms.

Interactive Visualisation of Large Databases

Visualising billions of items

Whilst the storage and retrieval of data from very large datasets is well understood, supporting effective and efficient data visualisations with, say billions of items and/or hundreds of dimensions, is still a challenging research activity. In particular, we need to provide the user with rapid feedback while exploring the data. Promising solutions come from different techniques of (visual) data reduction, able to scale on both data cardinality and data dimensions. Additionally, there are proposals to pre-compute metadata, e.g., indexing or aggregating data for visualisation purposes. However, the field is still a challenging one, and more formal approaches are needed, e.g., using formal definition of quality and visual quality metrics.

Researching this topic is crucial for visual analytics and its importance is also being acknowledged in the data management area. This suggests the pursuit of joint research efforts in areas such as new scalable data structures, novel algorithms for data reduction, innovative hierarchical data management and supporting visual analytics applications to adopt the data models of data management.

Distributed and Collaborative Visual Analytics

Visual analytics activities are more complex and longer than issuing a single query

Visual analytics activities are longer and more complex than issuing a single query against a dataset and exploring the result; moreover, they often involve several users at different sites. The process itself requires intermediate steps, saving and presenting partial results, annotating data and insights, resuming previous analysis, as well as sharing findings with the different users. Also, it is beneficial to be able to automatically reapplying the same visual analytics steps on different datasets or on a changed one, as with streaming data.

Long term and collaborative activities are poorly supported by classical data management techniques, and in order to reduce this gap, new research directions should be instigated, exploring collaborative systems explicitly designed to help the visual analytical process.

Visual Analytics for the Masses

Managing personal data is increasingly prevalent

The volume of personal digital data (i.e., emails, photos, files, music, movies, etc.) is increasing rapidly and with the availability of new Web based systems integrating data management technologies and information visualisation, this opens up new opportunities for visual analytics applications and new challenges. The home user becomes a naive analyst and this requires different interaction modalities for non-expert people and raises heterogeneity (data source and devices), scalability, and user acceptance issues.

Summary

Challenges for both the visual analytics and data management communities

Many challenges and opportunities associated with data management and visual analytics exist. They are related to solving basic data management problems that will help visual analytics activities, or to addressing problems arising from specific requirements of visual analytics. On the other hand, data management could fruitfully exploit by some results coming from visual analytics research. However, in order to make progress in the visual analytics field, we need to address some critical issues such as uncertainty problems, semantic data integration and semantics management, data provenance, data streaming, interactive visualisation of huge datasets, solving process intensive activities, and designing visual analytics systems intended for the general public. Dealing with these issues is a challenge that both communities have to take up, in order to take advantage of the increasing information opportunities available today.

3.4 Next Steps

By examining the state of the art in data management and the requirements of visual analytics, it is clear that the two displines would mutually benefit

from increased cooperative research and have subsequently identified particular challenges faced by these communities. We now describe what we would consider to be useful next steps towards stimulating future developments in data management and visual analytics research that will eventually enable new solutions that exploit the strengths of modern data management technology in the context of advanced visual analytics scenarios.

We recommend the activation of research projects bringing together the interdisciplinary competencies of visual analytics and data management, in order to progress and to gain a better understanding of the problems associated with the challenges described in the previous section. In particular, we point out three main areas that should to be addressed:

- The development of a new generation of data integration systems, based on the most advanced data management results and targeted to the specific, compelling visual analytics requirements, like high heterogeneity of data sources and data types, critical time constraints, and methods for effectively managing inconsistent and missing data.
- The development of new data reduction and analysis techniques for dealing with the modern complex data, such as high dimensional data and data streams.
- The development of new algorithms, data structures and visual data reduction techniques to facilitate the interactive visualisation of extremely large datasets.

Effort is required to distribute the ideas discussed in this chapter to potentially interested colleagues. This involves the dissemination of information about the potential advantages of enabling visual analytics in the data management research field, as well as the dissemination of information about the state of the art in data management and its according promises to visual analytics researchers. In addition, we suggest that steps are taken to improved literacy in each of the two fields in respective of other field, and as a longer term goal, update the educational curriculum to reflect the interdisciplinary nature of this topic.

4 Data Mining

4.1 Motivation

In recent years, researchers of different fields have identified a phenomenon that has been coined as information tsunami or data tsunami – we live in a world where the capacity of producing and storing data is increasing daily at a very fast pace, however, our ability, as human beings, to understand such an overwhelming amount of data has not grown at the same rate. In order to deal with this problem, we undoubtedly need new technologies to unite the seemingly conflicting requirements of scalability and usability in making sense of the data.

Humans are required in the data analysis process

In the last decades, several analysis methods have been developed which were purely automatic or purely visual, but to deal with the complexity of the problem space, humans need to be included at an early stage of the data analysis process[66]. We will now consider two examples of particularly complex problems that affect us: understanding the function of genes (e.g., how can devastating diseases be cured), and understanding earth dynamics (e.g., how can natural disasters be predicted).

The 21^{st} century has witnessed rapid development within the field of genomics. Initiatives such as the Human Genome Project and similar projects for other organisms, have begun to establish the genetic structure by identifying and locating genes in DNA sequences. Although far from perfect, these sequence-to-gene mappings will dramatically increase our understanding of genomics.

New tools and methodologies are necessary to help experts extract relevant information

At the same time, the world has been affected by some of the most catastrophic natural disasters in recent history. Some of these are of geologic origin, such as the recent L'Aquila earthquake (2009) or the Sumatra-Andaman earthquake (2004), which triggered the single worst tsunami in history; the majority are related to climatic dynamics. For example, Hurricane Katrina (2005), one of the costliest and deadliest hurricanes in American history; or El Niño (El Niño Southern Oscillation, ENSO), whose erratic periodicity cost hundreds of lives and caused billions in damage worldwide, partly through flooding in South America and partly through failed harvests in South East Asia. Natural and man-made catastrophes, coupled with increased security needs have triggered the improvement of monitoring systems (e.g., the Global Monitoring for Environment and Security, GMES[1]), capable of compiling data gathered from different sources (on the ground, from the depths of the oceans, by aircraft or balloon, or by satellites) and assembling them into usable, compatible and comparable information services.

[1] http://ec.europa.eu/gmes/

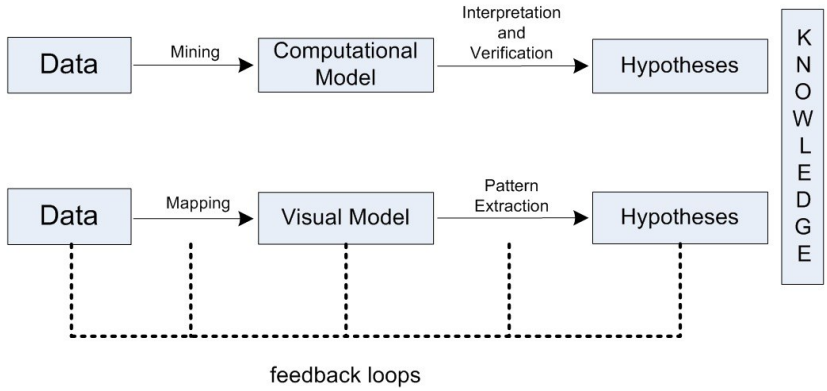

Figure 4.1: Comparing traditional data mining (top) and information visualisation (bottom) analytic processes[14]

Computers have played a key role in improving data acquisition methods thus providing us with the necessary depth of information to diagnose and prevent both diseases and natural disasters. Experts are required to assess current data sources and make predictions. Although massive amounts of data are available, it is imperative that new tools and new methodologies are developed to help these experts extract the relevant information.

KDD is useful but still limited

Knowledge discovery and data mining (KDD) is about semi or fully automated analysis of massive datasets and is therefore central to the problems at hand. Such automatic analysis methods are part of a discipline with a long tradition and solid, theoretical foundations. They are not focused on one application area, and the contributions of the field are more about general methodologies. KDD methods are especially suitable for analytical problems in which there exist means for assessing the quality of the proposed solutions. However, very often they become black-box methods in the hands of the end users (e.g., the prostate cancer physicians) or the algorithms provide results that do not lead to a solution to the problem, because they do not take into account relevant expert knowledge.

Limitations of visualisation methods

In contrast, visualisation methods use background knowledge, creativity and intuition to solve the problem at hand. While these approaches often give acceptable results for small datasets, they fail when the supplied data is too large to be captured by a human analyst[66]. Figure 4.1 compares the KDD and information visualisation processes.

Visual analytics approach is the third way

Nowadays, a third approach has begun to emerge, i.e., the visual analytics approach, which brings the experts' background knowledge back into the analysis process, together with the ability to interact and steer the analysis process.

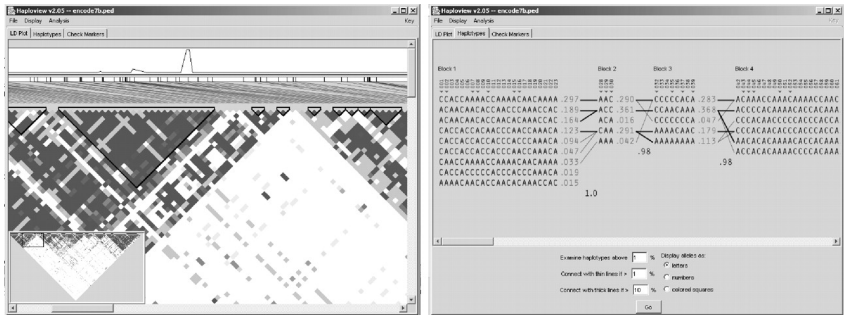

Figure 4.2: Haploview LD display[12] with recombination rate plotted above (left) and haplotypes display (right)

4.1.1 Visual Analytics as a Combination of Automated and Visual Analysis – Success Stories

There exist a number of successful application areas in which the visual analytics approach has been used together with KDD methods. Four notable examples are discussed; bioinformatics and climate change (mentioned already in the motivation section), the pervasive problem of finding patterns in data, and spatio-temporal data mining (discussed in Chapter 5).

Bioinformatics. Bioinformatics is one of the areas where KDD methods have been used extensively in combination with visualisation methods. In fact, bioinformatics is arguably one of the great successes in the field of computational data analysis – the combination of biology and KDD has produced a whole new area of research. The multidisciplinary approach that combines biology, medicine and visualisation with advanced KDD methods have resulted to new scientific knowledge and has led to understanding and treatments for serious diseases such as cancer. The fact that KDD methods and algorithms are central in the bioinformatics field is recognised by the scientific community. The importance of the combination of such methods with visualisation can be concluded from the fact that, ten out of the fifty most-frequently cited articles in the Bioinformatics journal, currently the leading reference in the field, propose visual analysis tools or methods (see, for example, Figure 4.2, where an interactive visual interface is used for computation and analysis of linkage disequilibrium statistics and population haplotype patterns from primary genotype data). As the complexity of research increases, more and more researchers and companies are relying on visual analytics as an indispensable aid for decision making in bioinformatics. Another example of this trend is the widely use of BioConductor[2] for computational biology and bioinformatics that provides access to a large collection of KDD, machine learning and statistics methods together with advanced visualisation techniques.

Ten out of the fifty most-frequently cited articles in the Bioinformatics journal propose visual analysis tools or methods

Climate change. KDD is becoming increasingly important for measuring the impact of climate change. The massive volume of climate-related data gathered

[2] http://www.bioconductor.org/

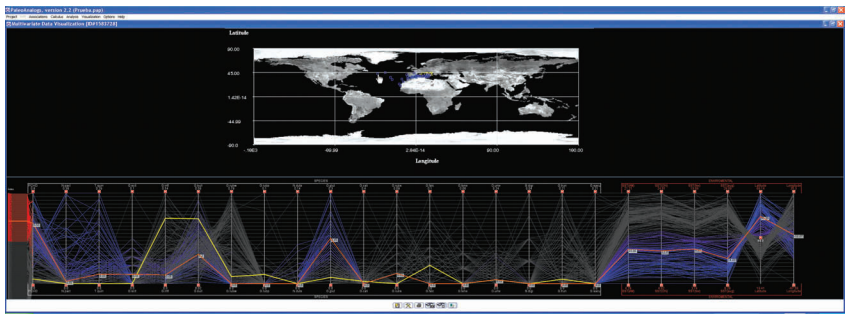

Figure 4.3: By means of a combination of an automatic pattern matching algorithm and an interactive visual interface, the expert is able to understand sea surface temperature changes over the past millions of years and use this to help predict future changes[109]

Visual analytics for predicting climate extremes

from remote and in-situ sensors is increasing rapidly. This vast climate database is augmented with proxy observations from the past and with data coming from simulations of global or regional climate models. In order to gain predictive insights on climate extremes and foresee events with potential impact, all these spatio-temporal data sources must be integrated, mined and presented in an understandable way. KDD methods can extract novel insights about climate extremes and regional change, while geographical information systems and multidimensional visualisation techniques can relate climate change and extremes to societal and ecological impacts. To illustrate this process, Figure 4.3 shows the distribution of micro-fossil species at different sites of the world through millions of years. These are used to reconstruct environmental features of the past by means of expert-steered k-nearest neighbour prediction; the use of linked parallel coordinate plots, maps and animations enables further analysis of the model.

Combining KDD and visualisation methods

Pattern identification. Searching for patterns is one of the main goals in KDD and it is applied to many varying domains such as medical, biological, financial and linguistic. Novel, exploratory data analysis tools and adaptive user interfaces have been developed by tailoring and combining existing KDD and visualisation methods. Variations of scatterplots, parallel coordinate plots, dendograms, heatmaps and many other visualisation techniques are used in combination with clustering, self organising maps, principal components analysis and other pattern extraction algorithms using colour linking and/or interactive brushing with excellent results. In the last five years, the success of this integration has contributed significantly to the use of visual analytics.

Spatio-temporal data mining. The availability of large repositories of spatial and spatio-temporal data has triggered the interest of the data mining community to the opportunities presented with these new data resources. However, this field presents new challenges and complexities: both the raw data (e.g., the traces of people moving in a city or flocks of animal migrating from one continent to another), and the extracted pattern (e.g., the aggregated flow from

one zone of a city to another), may be too complex to be interpreted effectively by the analyst[79]. A new research field, identified by the European Project GeoPKDD[3][47], is emerging from the interaction of data mining technique with visual analytics tools for spatio-temporal data. An example of this interaction is presented in Andrienko et al.[6], where the knowledge extraction process is driven by the analyst, enabling efficient management of large datasets through stepwise refinement of the extracted model.

Combining visualisation and data mining for analysing mobility

4.1.2 Is Industry Ready for Visual Analytics?

Generally, the use of visual analytics has been well received by industry. Several companies have embraced this business model and are selling visual analytics tools and/or offering consultancy services to different industries. Arguably, the main reason to adopt this novel approach is that business users have witnessed the success stories of data mining, but they need to understand its results. Few KDD models are easy to understand and techniques need to be developed to explain or visualise existing ones. Furthermore, there is a need for techniques to translate the user's questions into the appropriate input for the data mining algorithms. Industry representatives see the need for intuitive and interactive KDD/visual analytics methods by which they can readily interact with the data and the underlying KDD models.

Techniques are required to understand the resulting KDD models

Due to its generality, KDD can be used in most visual analytics scenarios. Some good examples of its use are given below.

Marketing data. Data mining has appeared often in the media as an artificial intelligence technique capable of extracting interesting patterns out of customer activity, allowing effective marketing campaigns to launch new products and acquire new customers (see, e.g., Xtract Ltd[4]). With the rapid development of IT, exploring and analysing the vast volumes of commercial data is becoming increasingly difficult. Visual analytics can help to deal with the flood of information, since it provides a means of dealing with highly non-homogeneous and noisy data and involves the user in the data mining process (see, e.g., Visual Analytics Inc.[5]).

Process industry. The problem is that manufacturing systems are much better at collecting data than they are at helping one understand it (see, e.g., Spotfire[6]). In this context, visual analytics provides a way of making sense of the very large volume of data generated by factories related to quality parameters, process trends, maintenance events, etc. Thus, visual analysis can help solving problems, such as detecting anomalies and analysing their causes that, in turn, will lead to the development of more efficient and reliable processes.

Software industry. The complexity and size of industrial projects is currently growing rapidly, and hence there is a clear need for tools that assist during

[3] http://www.geopkdd.eu/
[4] http://www.xtract.com/
[5] http://www.visualanalytics.com/
[6] http://spotfire.tibco.com/Solutions/Manufacturing-Analytics/

the development, testing and deployment cycles. Currently, understanding the evolution of software has become a crucial aspect in the software industry. In the case of large software systems, gaining insight into the evolution of a project is challenging. Retrieving, handling and understanding the data poses problems that can only be solved by tightly coupling data mining and visualisation techniques. Thus, visual analytics can be effectively applied to support decision making in the software industry.

Pharmaceutical industry. The drug discovery process is very complex and demanding and often requires a cooperative, interdisciplinary effort. Despite the considerable methodological advances achieved through the years and the huge resources devoted to this enterprise, the results are disappointing. The recent completion of the human genome project has not only unearthed a number of new possible drug targets but has also highlighted the need for better tools and techniques for the discovery and improvement of new drug candidates. The development of these new tools will benefit from a deeper understanding of the drugs' molecular targets as well as from more friendly and efficient computational tools. With the flood of data across all aspects of the pharmaceutical industry, visual analytics is emerging as a critical component of knowledge discovery, development, and business[94].

4.2 State of the Art

The focus of the visual analytics community has been on interactive visual representation and exploration of data. But, the aim of the KDD community has focussed on developing computational methods that can be used to extract knowledge from data. There is a general awareness of the need to integrate visual analytics and KDD, but relatively few efforts have been made to address this issue. In this section, we present an overview of research and commercial systems in the following categories: statistical and mathematical tools, visually supported tools and combined methods. At the end of this section, we present several examples of KDD/visual analytics approaches from the fields of bioinformatics and graph visualisation.

As we have seen, the objective of knowledge discovery and data mining is to extract information from large datasets[55, 108]. This process is characterised by a series of operations (i.e. data pre-processing, data mining, data cleaning) that transform the data in various ways to obtain patterns and models that represent the implicit information within the data. Usually, the pre-processing steps produce a dataset in a suitable format for the data mining algorithms. The post processing steps transform the output of the mining into a form that can be understood by the analyst.

Data mining tasks are classified as predictive or descriptive

Data mining tasks can be divided into predictive tasks (e.g., classification, regression) and descriptive tasks (clustering, pattern mining, association rule discovery, etc.). In the former case, the data is analysed to build a global model, which is able to predict the value of target attributes based on the observed values of the explanatory attributes. In descriptive tasks, the objective

is to summarise the data using local patterns that describe the implicit relationship and characteristics of the data itself. However, as discussed earlier, existing methods support limited user interaction and are mainly designed for homogeneous data sources. Some attempts have been made to enhance data mining with visualisation providing advanced interactive interfaces. A survey of the state of the art of current and proposed solutions that facilitate sense-making for interactive visual exploration of billion record datasets, is provided in 'Extreme visualization'[99]. Several interactive tools for information visualisation, designed for specific data types have been presented in the literature. These include graph visualisation[1], time series interactive search[20] and network visualisation[9].

We now give an overview of some research and commercial systems in the context of data mining and visualisation, categorised as follows:

- Statistical and mathematical tools
- Specific algorithmic tools
- Visual analytics libraries
- Visual data mining tools
- Web tools and packages
- Scientific visualisation tools
- Combined methods
- Computational information design

Statistical and mathematical tools. Statistical analysis has a long history of visualising the results as time series, bar charts, plots and histograms. Examples of tools providing statistical and mathematical visualisation are R[7], Matlab[8],Mathematica[9] and SAS[10] tools for statistical computing and graphics.

Specific algorithmic tools. Algorithmic tools have been developed by the research communities for a specific task or problem. Examples are Graphviz[11](see Figure 4.4), open source graph visualisation software, or Pajek[12], which is more focused on the analysis of social and complex network data by taking advantage of network/graph visualisation.

Visual analytics libraries. One example, originally aimed at providing expertise in data visualisation and visual design is BirdEye[13], a community project to advance the design and development of a comprehensive open source information visualisation and visual analytics library.

Visual data mining tools. Visual data mining creates visualisations to reveal hidden patterns from datasets. The need of new methods in data analysis has

[7]http://www.r-project.org/
[8]http://www.mathworks.com/
[9]http://www.wolfram.com/
[10]http://www.sas.com/technologies/bi/visualization/visualbi/index.html
[11]http://www.graphviz.org/
[12]http://vlado.fmf.uni-lj.si/pub/networks/pajek/
[13]http://code.google.com/p/birdeye/

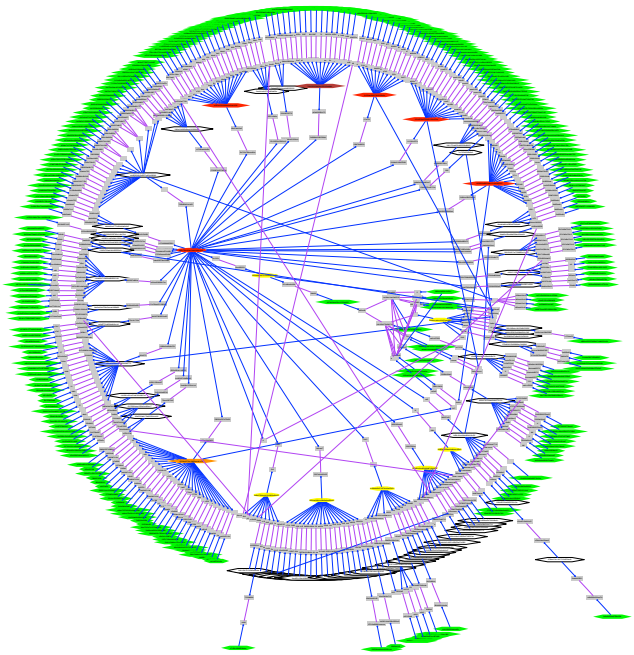

Figure 4.4: Radial layout graph visualisation made using Graphviz. A real-world network containing 300 sites over 40 countries. The diagram was made to trace network incidents and to support maintenance. Used with permission of AT&T

launched the field. Several products are on the market; often focused on 'business intelligence' such as marketing, risk analysis, sales analyses and customer relationship management. Some examples are:

KNIME[14] is a modular data exploration platform that enables the user to visually create data flows (or pipelines), selectively execute some or all analysis steps, and later investigate the results through interactive views on data and models.

Weka[15] is a collection of machine learning algorithms for data mining tasks, which allows the user to create pipelines in order to perform data pre-processing, classification, regression, clustering, association rules, and visualisation. It is open source code, developed in Java.

Similarly to Weka, RapidMiner[16] is an environment for machine learning and data mining tasks, which allows the user to create data flows, including input and output, data pre-processing and visualisation. It also integrates learning schemes and attribute evaluators from the Weka learning environment.

[14] http://www.knime.org/
[15] http://www.cs.waikato.ac.nz/ml/weka/
[16] http://rapid-i.com/

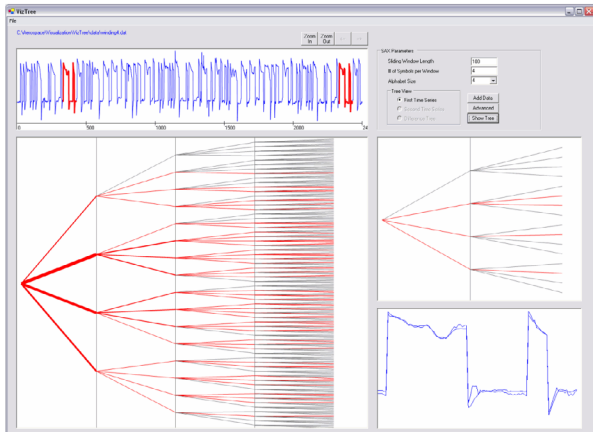

Figure 4.5: VizTree: The top panel is the input time series. The bottom left panel shows the subsequence tree for the time series. The top right window shows a zoomed in region of the tree, and the bottom window plots the actual subsequences when the user clicks on a branch

Web tools and packages. An increasing number of tools are available online, but user interaction becomes more complicated and difficult to model and optimise, when used remotely. With these tools, users can create visualisations using their own data. An example of an online social data analysis tool is ManyEyes[17], an IBM application for social data analysis.

Scientific visualisation tools. Scientific visualisation is the representation of data graphically as a means to gain understanding and insight into the data. It involves research in computer graphics, image processing, high performance computing, and many other areas. Scientific visualisation tools are often adopted for modelling complicated physical phenomenon. An example in the field of natural science is Gravity waves[18], where the Globus Toolkit has been used to harness the power of multiple supercomputers and simulate the gravitational effects of black-hole collisions. Other examples come from geography (e.g., terrain rendering) and ecology (e.g., climate visualisation).

Combined Methods. There have been some attempts to combine data mining and visualisation. For example, some concentrate on the analysis of time series by using tree visualisations and interactions (VizTree, see Figure 4.5), or propose a combination of visual data mining and time series (Parallel Bar Chart, see Figure 4.6), or combine KDD concepts and visualisations (Statigrafix[19], see Figure 4.7). However, each one lacks either effective visualisation, automatic data mining or requires a strong expertise in the application field.

[17] http://manyeyes.alphaworks.ibm.com
[18] http://www.anl.gov/Media_Center/logos20-2/globus01.htm
[19] http://statigrafix.com

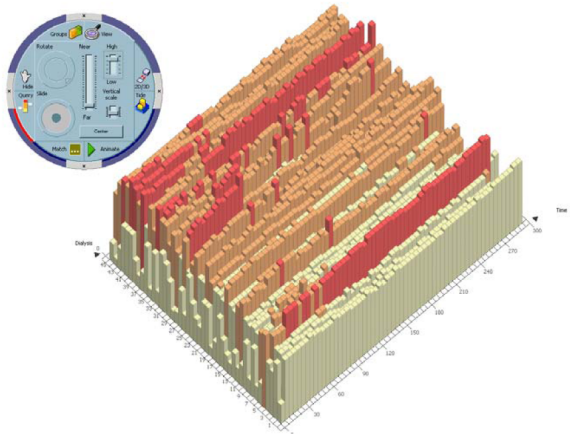

Figure 4.6: Parallel Bar Chart[30] visually represents each time-series in a bar chart format where the X axis is associated with time (the axis on the right), the Y axis with the value (height of a bar) of the series at that time and the axis on the left identifies the different time-series, ordered by date

Finally, in the bioinformatics and graph visualisation fields there are several examples of KDD/visual analytics approaches. For instance,

JUNG (Java Universal Network/Graph Framework[20]) is a software library that provides a common and extensible language for the modelling, analysis, and visualisation of data that can be represented as a graph or network. It is written in Java and it includes implementations of a number of algorithms from graph theory, data mining, and social network analysis, such as routines for clustering, decomposition, optimisation, random graph generation, statistical analysis, and calculation of network distances, flows, and importance measures.

HCE[21] (Hierarchical Clustering Explorer) for interactive exploration of multidimensional data. Genome researchers adopt cluster analysis to find meaningful groups in microarray data. Some clustering algorithms, such as k-means, require users to specify the number of clusters as an input, but users rarely know the right number beforehand. Other clustering algorithms automatically determine the number of clusters, but users may not be convinced of the result since they had little or no control over the clustering process. To avoid this dilemma, the Hierarchical Clustering Explorer (HCE, see Figure 4.8) applies the hierarchical clustering algorithm without a predetermined number of clusters, and then enables users to determine the natural grouping with interactive visual feedback (dendrogram and colour mosaic) and dynamic query controls.

[20] http://jung.sourceforge.net
[21] http://www.cs.umd.edu/hcil/hce/

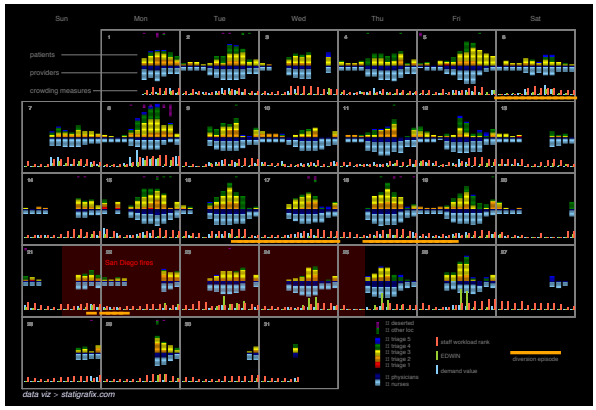

Figure 4.7: Calendar-template data visualisation of datasets captured at the visual analytics San Diego Health Service's Emergency Dept in Oct 2007 (Source: Alan Calvitti, statigrafix.com)

BicOverlapper[22] is a framework to support visual analysis of gene expression by means of biclustering. In order to improve the visualisation of biclusters, a visualisation technique (Overlapper) is proposed to simultaneously represent all biclusters from one or more biclustering algorithms, based on a force-directed layout. This visualisation technique is integrated in BicOverlapper, along with several other visualisation techniques and biclustering algorithms.

Computational Information Design. Similarly to the previous category, Computational Information Design has been suggested by Ben Fry from the Massachusetts Institute of Technology[23]. In an attempt to gain better understanding of data, fields such as information visualisation, data mining and graphic design are employed, each solving an isolated part of the specific problem, but failing in a broader sense: there are too many unsolved problems in the visualisation of complex data.

4.3 Challenges

4.3.1 Introduction

The developers of visual analytics applications face several fundamental challenges when attempting to develop integrated iterative methodologies that involve information gathering, data pre-processing, knowledge representation, interaction and decision making. One of the main purposes of this chapter is to establish the degree to which existing techniques and approaches can be integrated, and, in a wider sense, how the human-computer integration might be facilitated[14].

[22] http://vis.usal.es/bicoverlapper/
[23] http://benfry.com/phd/

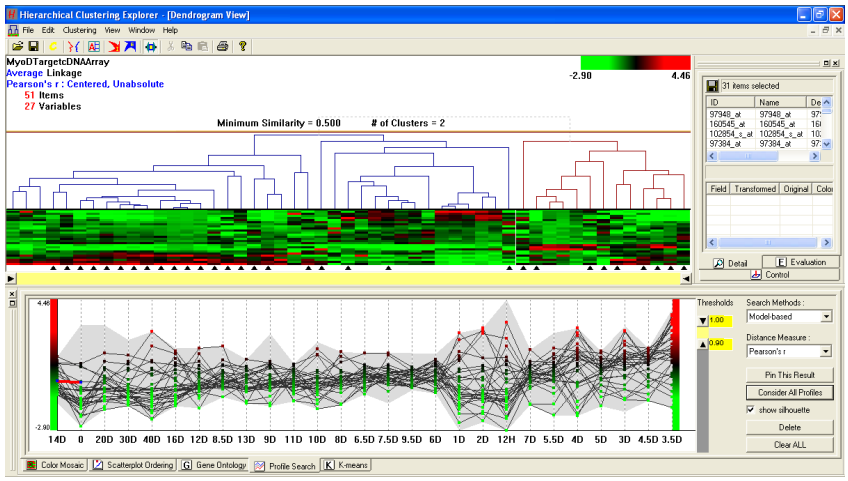

Figure 4.8: Hierarchical Clustering Explorer for interactive exploration of multidimensional data[97]

According to Thomas and Cook[111], the so-called 'Grand Challenges' faced by visual analytics can be grouped into five categories: *analytical reasoning*, *visual representations and interaction techniques*, *data representations and transformations*, *production, presentation and dissemination*, and *moving research into practice*. The first category (analytical reasoning) refers to the reasoning frameworks by which users derive insights or discover knowledge to support the decision making process. These frameworks provide the foundation for applying specific transformations, visual techniques or other operations, on the data. The second category (visual representations) covers all interactive means, methods and techniques that enable visual representation of data. The third category (data representations and transformations) refers more to the specific ways that data is represented, as well as the operations upon data (which might be noisy, incomplete, or uncertain). Representations refer to the fundamental 'structure' of the data within an application, usually non-intuitive to users, but responsible for facilitating data transformations, calculations, etc. The fourth category (production, presentations and dissemination) refers to user activity and interaction. Finally, the fifth category (moving research into practice) refers to the practical application of methods and techniques.

The five categories of Grand Challenges

In KDD, analytic reasoning, and data representations and transformations are highly relevant

In terms of KDD, the first (analytical reasoning) and third (data representations and transformations) categories are highly relevant. Next we analyse in more detail several specific technical challenges in both of these categories relating to KDD.

4.3.2 Data Issues

While data types, formats and characteristics are a key part of the motivation underlying visual analytics approaches, they also represent a key challenge

to operational implementations. Here we focus specifically on data mining issues. For visual analytics to be able to fulfil its true promise, we need the capability to integrate both heterogeneous and large datasets. This includes:

- (qualitative) textual data,
- data stored in (distributed) databases,
- data received from sensors,
- spatial data such as satellite imagery,
- audio and video.

KDD approaches tend to focus on specific types of quantitative data

KDD approaches tend to focus more on specific types of (quantitative) data, however, approaches for other data types are emerging[93, 40]. There are several levels of complication:

- Some of the data could be arriving in real-time, so that ways to manage this (storage, management and interactive analysis and visualisation) are required.
- Some of the data will be of variable quality, therefore we need to know as much as possible about the data itself.
- Data may be incomplete, so we need to know what is missing, as well as having ways to handle or manage the missing data.
- Data may be of different (spatial) scales, and therefore require transformations/mappings to be compatible with other data.

The means by which data should be managed and distributed is addressed in more detail in Chapter 3, and for spatio-temporal data in Chapter 5. However, to support the data mining initiatives in visual analytics, we require methods for data cleaning, integration, data fusion etc. If we are to achieve 'real time' analytics, then the cleaning and integration methods should be automated and fast. These problems are non-trivial and significant developments are required before data mining can be integrated into a visual analytics platform.

Data standards or metadata are required

A necessary feature of these developments will be the adoption of standards across different visual analytics toolsets and software environments. These data standards[24] do not just concern data formats. More fundamentally, we require *metadata*, or documentation of the data itself: lineage/source(s), formats, method of collection, accuracy and completeness in order to support data mining approaches.

4.3.3 Visual Analytics Platforms

One of the main goals of KDD is pattern extraction. This can be applied in many application domains, as discussed in the following section. Most of the existing visual analytics related software provides some common functionality (statistical analysis, graphing tools, algorithms, visualisation), but as noted in the previous section, data needs to be represented in a format suitable for the analysis algorithms.

[24] See for example http://www.iso.org/

Most software is developed for a specific task

Functions such as linking and brushing, scatterplots and clustering are basic functions, yet are missing from many software environments. A key reason for this is that most software has developed out of the specific needs of a particular discipline, and therefore is geared towards specific types of decision making. As noted in Section 4.2, a variety of tools and environments exist which address different aspects of visual analytics. Examples include KNIME[25] and OECD explorer[26], which are developed specifically for geographic data. These are significantly different from business intelligence tools, which focus specifically on marketing and management strategies and risk analysis, and differ significantly from bioinformatics tools.

The fundamental challenge, given that we are likely to see ongoing development of these heterogeneous toolsets, is to provide the functionality so that users can easily switch between visual analytics tools and data sources. To achieve this, data sources will have to be integrated directly using applications programming interfaces (APIs). Clearly, building specific visualisation tools for every use case is not a feasible solution. Generic tools are required that can be customised with appropriate algorithms and visual tools.

Many of the commonly used data mining algorithms are already well-developed and do not require expert users in order to be applied. For example, even a novice user can use a clustering algorithm, provided it has adequate documentation. This chapter has identified a wide variety of emerging software platforms both within, and closely related to visual analytics. Many of these have their own implementation of various algorithms. It has also been noted that an initial community repository for information visualisation and visual analytics algorithms is already underway (BirdEye). In order to facilitate KDD and data mining approaches, cross-disciplinary initiatives are required. Not simply to provide algorithms, but to inform the wider community (KDD, information visualisation and visual analytics) about their functionality and requirements. Cross-platform standards could also play an important role in this, in terms of defining a core set of widely used algorithms, as well as frequently used visualisation techniques.

Interdisciplinary initiatives are required

One further issue is the provision of distributed collaboration between disciplinary experts. This has a major implication for visual analytics platform in sharing very large datasets over the Internet. Further investigation is required into the kinds of technologies that can facilitate this.

4.3.4 Towards Visually Controlled Data Mining

Advanced KDD methods require expertise

The current data mining methods support only limited user interaction. Also, existing KDD methods are not directly applicable to visual analytics scenarios. This is essentially because the more advanced KDD methods are often non-intuitive, in that a significant degree of experience is required for their successful application. As well as user expertise, many KDD methods

[25] http://www.knime.org
[26] http://www.oecd.org/gov/ regional/statisticsindicators/explorer

require substantial processing time and therefore place significant demands on computer hardware.

In complex domains, the models and patterns extracted by traditional KDD approaches may also be difficult to interpret, and relevant information may be hidden within large results sets. It is envisaged that visual analytics methods may simplify the presentation and evaluation of the models extracted. These issues should be addressed if KDD is to be able to make a significant contribution to visual analytics (and vice-versa). Work is required on identifying and implementing means by which this might occur.

In a review of visual analytics, information visualisation and data mining literature, Bertini and Lalanne[14] classify recent literature within these disciplines along a 'continuum' of approaches, ranging from pure data mining to pure visualisation and propose new research questions and directions. Puolamäki et al.[91] identifies a new class of data mining methods, *visually-controlled data mining*.

Towards 'visually controlled mining'

For a data mining method to be useful in visual analytics it should be:

1. Fast enough – sub-second response is needed for efficient interaction.
2. Parameters of the method should be representable and understandable using visualisations.
3. Parameters should be adjustable by visual controls.

Efficient interaction represents a significant hurdle in bringing KDD to visual analytics, as noted above. In terms of the second and third requirements, further investigation into what types of 'visual controls' are required to manage and adjust the algorithms is required.

There are hardware, software, and algorithmic issues involved in developing the kind of mixed-initiative approach identified above. From a hardware point of view, machine specifications should be able to handle the computations adequately. The software should be as application-independent as possible, perhaps following the plugin topology favoured by many open source research tools. These algorithms must be both efficient and robust. One could conceive of a repository for plugins to various existing and emerging platforms (similar to BirdEye as noted above), maintained for quality control and ongoing community development.

Hardware and software issues are still open

The research on visual analytics, using visualisation and interaction methods to analyse large datasets, and data mining have evolved separately. However, at the current time, communication and interaction between both research communities has just started in the form of workshops under the umbrella of their main international conferences (such as SIGKDD and VisWeek). The success of these events has confirmed that there are significant benefits from bringing these communities together. A challenge lies in establishing collaboration between these research communities, so that we can focus on applications. This requires that domain experts from the data mining/KDD, visual analytics and information visualisation communities, collaborate on the specific ways that the two approaches can complement one another.

KDD and information visualisation communities should collaborate more

4.3.5 Research and Evaluation

It is possible to identify three general categories relating to research and evaluation from the perspective of KDD and data mining. These relate to *evaluation*, *research development* and *collaboration*.

Evaluation is difficult - it is unclear what a good solution is

The evaluation of visual analytics approaches is regarded as difficult. It requires specific criteria on how to judge a visual analytics solution or application. The evaluation also requires new measures. While significant criteria exist in the separate fields which visual analytics seeks to draw together, it is difficult to envisage how these might fit together in some unified way. For example, in the discipline of visualisation, a number of techniques and criteria exist for evaluation of results such as assessment of the effectiveness of the result (through user evaluations). Similarly, it is relatively easy to judge the outcome of traditional KDD approaches through validation of the results with reference data. However, in terms of combined KDD/visual analytics solutions, it is still unclear what a 'good' solution or application should look like. We therefore expect to see ongoing development of (design and implementation) guidelines, to help identify a base upon which we can build further.

Collaboration requires workflow sharing

In terms of research collaboration, significant technical challenges exist. Several of these were identified above. The general question is "how will collaborative data mining/visual analytics approaches work?" They would require facilities for transfer of data, but also of custom algorithms or even better, entire data workflows in some way. Some collaborative approaches are currently underway, but these are by no means well developed in terms of the requirements of a mixed-initiative *visually-controlled mining* approach. More work is required to investigate the possibilities of data, software, and even full workflow-sharing approaches and their respective practical limitations.

In terms of development of the research field itself, this brings about a sociological and very practical question: how to get the referees to accept visual analytics/KDD papers? Special issues are perhaps a temporary solution, but ultimately, alongside the rapid development of software and integrated solutions, we would expect to see several dedicated academic journals to support the research discipline.

4.4 Opportunities

While the key issues identified in the previous section are significant barriers to progress, several of these also represent major opportunities. Below we discuss four general categories of these: the development of generic tools and methods, regulation and quality control, visualisation of models, and linkage of KDD and visualisation communities.

Need for a repository of generic tools and methods

Firstly, generic components are needed in order to stimulate research. This obviously includes algorithms, i.e., methods, and software libraries (preferably

open source for maximal spread). It is possible to envisage some kind of 'repository' for things like plugins and software libraries with associated documentation to promote access to a range of research communities. It has already been identified that here will need to be some kind of regulation and quality control for this to develop in a controlled manner. The major opportunity in this sense is to provide the guidelines and framework for these components to develop.

Visualisation of models could be useful

In addition to the visualisation of the data we should move to *visualisation of models*. For example, why are two points clustered together? If we know some groups of people and their social interaction network, what kind of an interaction model would help to explain the data? The initial steps in achieving this are relatively simple: just bring the basic methods to visualisation of model spaces. Data mining models contain information about the phenomena. As discussed earlier in this chapter, initial approaches are already underway.

Collaboration between KDD and visualisation communities should be encouraged

The final opportunity, already identified above, relates directly to the above issue and involves collaboration between KDD and visualisation communities. The two communities certainly share an awareness that their approaches have significant overlap. While also a cultural challenge, there are significant opportunities for cross-pollination of approaches, methods and techniques. Ways to encourage and stimulate this might be through for example expert groups or mixed-initiative 'challenges' at key research conferences. From the review in Section 4.2, as well as the VAKD '09 Workshop[91], it would appear that we are close to a breakthrough.

4.5 Next Steps

Visual analytics is an emerging research field that combines the strengths of information visualisation, knowledge discovery in databases, data analysis and mining, data management and knowledge representation, human perception and user interaction. In this report we discussed the scope of visual analytics and analysed several challenges and opportunities that stem from this very promising field. Our investigation and analysis suggest that there is a clear need for integration of visual analytics and knowledge discovery and for building a community. The merging of the KDD and visual analytics communities could be achieved by two main approaches: bottom-up and top-down.

A bottom-up approach would include several dissemination activities, such as workshops, conferences and journal special issues. The VAKD '09 Workshop on Visual Analytics and Knowledge Discovery, organised by us, was a great success. The second VAKD workshop[27] will be organised in Sydney in conjunction with the 10th IEEE International Conference on Data Mining (ICDM 2010). A series of VAKD workshops will promote the development of novel visual analytics ideas and bring visual analytics research communities

[27] http://www.mpi-inf.mpg.de/conferences/VAKD10/

closer. Further, we should organise several collaborative research projects that would involve leading research groups.

Historically, challenges have been traditionally a good way to catalyse research. In VAKD '09 workshop, the authors were encouraged to address the tasks of the IEEE VAST 2008 visual analytics challenge[50], which contain both visual analytics and KDD angles in the performance evaluation. We should organise KDD challenges in the spirit of visual analytics. For example, the evaluation of a classification algorithm should not just be the classification accuracy but should also involve several other factors, such as, user interaction, visualisation, etc. It would be essential to include both visual analytics and KDD aspect in the Grand Challenges stated in Section 4.3.1.

Knowledge discovery approach should be reconsidered and data mining processes should evolve in the direction of visual analytics processes. As part of this process, we should consider new performance evaluation measures, as it is clear that we will need more than just algorithmic measures.

One major contribution would be to develop novel visual analytics approaches that enable visualisation for both the data and the underlying model. So far, standard visual analytics only allowed visualisation of the data. For this purpose, several existing information visualisation techniques could be used and further extended and tailored, with the help of data analysis methods, to produce useful and usable data model representations.

Current data mining methods support limited user interaction. For a data mining method to be optimal in a visual analytics application, it should be fast (sub-second response is needed for efficient interaction) and the parameters of the method should be understandable and adjustable by visual controls. By using visual interaction, the visually-controlled data mining process will be more efficient than by 'blindly' applying some data mining method, or by just interactively visualising data.

Another challenge for visual analytics is scalability of algorithms and heterogeneous data. Special emphasis should be given to methods that scale well and are applicable for indexing, accessing, analysing and visualising huge datasets. At the same time, a new trend in the area of data mining is being able to handle and combine data from large and possibly conflicting sources. Developing visual analytics algorithms that can handle this information overload and ambiguity efficiently would be another major contribution to the visual analytics community.

It is important to consider the application aspect of visual analytics. As also mentioned by Keim et al.[66], for the advance of visual analytics, several application challenges should be mastered including physics, astronomy, business, security, economics, biology and health, engineering and mechanics and GIS. Visual analytics applies to a wide range of different application fields and for our part we should encourage and enforce interdisciplinary collaboration. All the aforementioned communities should be investigated extensively and visual analytics algorithms should be developed that are tailored to their needs.

5 Space and Time

5.1 Motivation

Space and time are essential

People have to solve problems in time and space every minute of the day. Most of our decisions and actions depend on where we are and when. They also usually involve where other people or significant things are and how we expect the situation around us to develop. When the spatial scope of such decision-making exceeds an area that is directly observed or well-known people have traditionally used maps. These imperfect representations of reality serve as adequate models to solve problems and support decisions. Maps not only help people to orient themselves in geographical space but also to gain an understanding of events and evolving phenomena and to make discoveries – indeed, much map use can be considered (geo)visual analysis.

An historical example

An historic example is the discovery of the relationship between locations of reported cholera deaths in London and the location of a contaminated water pump. This geographic relationship was established in 1854 by plotting the locations of the deaths and water pumps on a map (see Figure 5.1). Doing so enabled Dr. John Snow to infer that the water source was contaminated. There is a temporal component to the story too. Decrease of the number of the cholera deaths after removing the pump handle subsequently confirmed the causal link.

More mundane problem solving in time and space is frequent and personal – avoiding slow traffic on the way to work, watching the weather forecast to decide what to wear, finding a nice place to live or go on holiday.

Complexity of spatio-temporal analyses increases

More complex systems, and their representation in larger datasets with more detail, greater scope, and higher resolution, enable us to address more complex spatial and spatio-temporal problems. To address the current issues faced by international society, people need to consider and understand global processes involving demography, economy, environment, energy, epidemic, food, international relationships, and other factors. They need to know how characteristics of these processes vary, develop and relate in time and space. People must strive for sustainable life, whilst preserving the environment and using the resources in a manner that is wise and just. People also need to know the risks from possible natural and man-made disasters and be prepared to protect themselves and to deal with the consequences. In these circumstances, simple maps with points and crosses are insufficient. To approach these problems, people need sophisticated maps and advanced computational techniques for data analysis. These must be interdependent and synergetic, accessible and usable, and must

support decision making in ways that take advantage of the kind of visual thinking deployed by Dr. Snow.

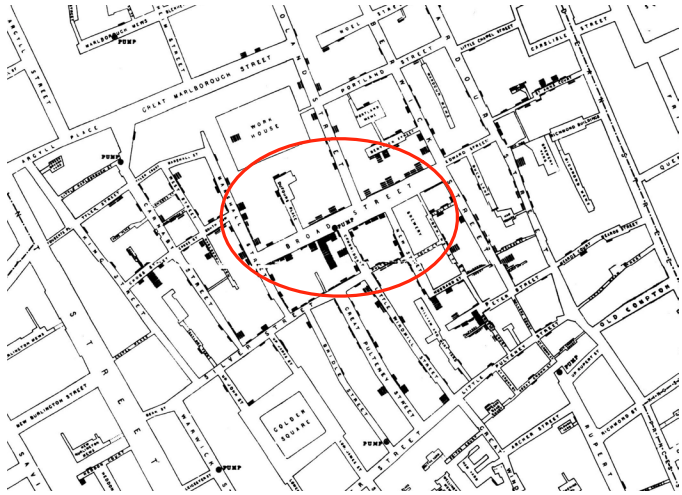

Figure 5.1: In the map made by Dr. John Snow in 1844 the death cases (indicated by bars) are clustered around the location of one of the water pumps. The cluster is encircled. Cited and analysed by Gilbert[49] and Tufte[115] (Source: http://johnsnow.matrix.msu.edu/)

This need refers not only to scientists and highly qualified experts. Nowadays, more and more citizens become spatio-temporal analysts when planning their journeys, looking for jobs, or searching for suitable places to live and visit. Those concerned about the development of their communities, regions, and countries want to understand the current situation and how this might be changing. They want to compare possible options and to take part in choosing the right strategies. They want to take advantage of the mass of data that is being collected and to which they are contributing. To do so, they take advantage of their visual and spatial capabilities in using maps. These maps show the locations of things and also reveal spatial and spatio-temporal patterns. They may be weather maps, election maps or disease maps and include maps reproduced in traditional ways on paper and their more sophisticated electronic descendants, which provide interactive tools for analysis. Our task is to give them such tools and help them use these tools to meet their information needs. The following scenario describes one of the possible ways in which visual analytics tools for spatio-temporal analyses could be developed and deployed to address a complex and changing process that affects large numbers of people.

Visual analytics tools are needed

5.2 A Scenario for Spatio-Temporal Visual Analytics

Late on Tuesday afternoon in mid-summer a severe thunderstorm passed through The City.

The Insurance Analyst

A number of reports of large hailstones mean that an insurance company requires a rapid overview of the damage incurred. To run an initial damage assessment, the insurance analysts need information about where the hail events occurred and about the things that are damaged. They therefore look for information from weather services, which provide data from different weather stations. Since hailstorms are very local, their exact locations cannot be detected entirely from existing sensor networks – storms often fall between the sensors. Therefore, the analysts make use of information from affected citizens provided on the Internet. Searching blogs, micro-blogs, photo sharing sites and other services where users make personal information available (including Flickr, Twitter and RSS feeds) reveals more detailed information about the spatial and temporal distributions of the hail events. The analysts use an interactive map to position the reported observations and transform them into structured, spatially and temporally referenced data, which are added to the database and simultaneously visualised. Spatial statistics are then used to identify possible tracks of the hailstorm derived from the data and probabilities associated with each. The results are also added to the map interface and data points are visually differentiated from the tracks that are derived from them. By combining these with the depicted observations, the analysts determine areas that are probably affected.

Analysis of storm tracks

Next, the analysts are interested in the things that were damaged during the storm. Those most vulnerable to hail include cars and agricultural areas. Cars are not static in time and place; therefore, data depicting traffic flow is considered. Such data is available from roadside sensors and increasing numbers of vehicle mounted devices. The analysts do not have access to 'live' data but quickly extract typical usage patterns for Tuesday rush hour in the summertime from a traffic database. By applying spatial and temporal filters to this dataset, the analysts can estimate the number of cars that passed through areas affected by the hailstorm during the time when it occurred, and make an initial assessment of the damage. The analysts put the filtered traffic flow data on the map and look at the typical destinations of the flows, to see in which districts the car owners live and to compare this with the spatial distribution of the clients of the insurance company.

Estimate disaster consequences

In order to detect the agricultural damage, satellite images showing information about the present status of agricultural areas are considered in combination with a land use database. One of the analysts recalls driving through the affected area some time ago and noticing strawberry fields. At this time of the year, the strawberries should have been already harvested. The analysts locate these areas on the map display and remove them. They also look at the other fields

and exclude those where no real damage from the hailstorm is expected. For the remaining fields, they calculate the estimated damage using the data about the types of the crops, the productivity of the fields, and the prices for agricultural products.

Report results of analysis by visual aids

Using interactive visual aids for report generation, the analysts report their findings. Besides a printable illustrated document, a series of annotated snapshots are developed from the visual displays. These are interactive and have links to the corresponding data and analysis artefacts, which are stored in the database. This report is forwarded to other working groups in the insurance company.

Development of long-term strategies

One of these groups examines long-term trends in hazard development and damage distribution. They investigate whether the frequency of hail events, their intensity or the associated damage are changing. Are hail events concentrated in certain areas? Another group deals with insurance contracts and customer issues and examines whether the spatial distribution of hail insurance customers is related to the spatial pattern of hail events. How many people in the most affected areas have an insurance policy? Should the insurance conditions be changed? How can exposure to risk be reduced?

The Family

A family living in The City has been affected by the hail. Their car was damaged whilst the father drove home from work. They are very upset about this and want to get more information about hazards in The City. They also want to know what they could do to protect themselves against hazards. They do this through a 'risk explorer' on the Internet. This interactive application enables citizens to examine their exposure to different hazards at different times and places according to different assumptions and levels of uncertainty. They are able to simulate different hazard events, such as historical or recent storms or floods and extremes with particular return frequencies to get an impression of their exposure to this type of hazard and the likely consequences. The risk explorer includes a discussion forum and a story-telling facility where people can place information about local hazard events on a map and an associated timeline. They can also post descriptions, annotations, and photos. People can report hazard events and discuss their occurrence and protection measures. The family subscribes to a warning service that will inform them about hazardous events more precisely in the future. The service provides information tailored to their situation. It derives the family's current and predicted location from an electronic diary, GPS-enabled mobile device, or cell-phone. If it coincides in time and space with predicted hazards, personal warnings are sent and alternative routing options and travel times are provided that account for the hazard. A visual display, which is adapted to the available device (PC, netbook, or mobile phone), explains why the warning has been sent and what the options are. By interacting with the display, people can enter additional facts about the current situation and their planned movements and ask the service to update predictions and recommendations. It is possible to compare the suggested options, choose the most appropriate one and,

Access to individually adapted information

Figure 5.2: A flooded chemical factory. © Greenpeace/Vaclav Vasku

if necessary, further adjust it interactively according to personal needs and priorities.

The Decision-Makers

Simulation of scenarios and consequences

Although the hail was a heavy and damaging event in The City, floods are the predominant problem. Politicians and local authorities have heard about an increase of heavy rainfall events and related flash floods as likely effects of climate change. They have to decide how to protect their community from floods in the future. To support decisions and develop strategies they need scientifically derived information that is presented clearly with assumptions, uncertainties, and alternative outcomes at the fore. Thus, expressive models are needed to simulate different situations related to different local conditions and climate parameters (see also Chapter 4). Scientists apply such models to calculate possible scenarios for The City and explain to the authorities the implications for their community.

Collaborative decision making

An Industrial Town is upstream of The City in a neighbouring country. The River that flows through The City originates in this country and passes through the Industrial Town on its way downstream. In the past floods have inundated factories in the Industrial Town (see Figure 5.2) and resulted in toxic material reaching The City and adjacent municipalities. Close collaboration between local and national governments is necessary in order to discuss safety precautions, to access and share relevant data, and to rapidly exchange information for early warning and protection. The Town's authorities have established contact and working relationships with neighbouring local and national authorities.

Now they start a collaborative decision finding process where all stakeholders are involved: different authorities, scientific advisors, the public, and several interest groups as well as stakeholders from the neighbouring country. The goal of this process is to establish risk and develop a pragmatic flood prevention strategy to protect future interests. Interactive visual tools facilitate the collaborative process. Analysts may look at specific aspects in detail by issuing interactive queries. Individual insights can be placed as annotations on the map. Annotation can be made visible to other analysts to initiate discussion. The arguments made are automatically tracked and a visualisation of the discussion flow helps in finding a good compromise for the discussed matter.

The Community

Share insight with the community

Since local authorities know that successful risk management requires not only technical and planning measures but also well-informed people with high risk awareness, they have also started a risk-awareness campaign in the schools. Teachers and school children work with the Internet 'risk explorer'. They explore the risk in their home area and also in other areas around the world. They can apply simulation models in a user-friendly manner to get a better impression about the effects of hazardous events and protection measures. A 'serious game' allows them to take the role of hazard defenders or decision makers to learn about the complexity of risk management; with a high score they can win a prize.

The Spatio-Temporal Analyst

Visual analytics methods must handle space and time appropriately

Note that all actors in these linked scenarios are spatio-temporal analysts: the insurance experts, administrators, politicians, scientists, engineers, insured and affected citizens, and school children. As spatio-temporal analysts, they must be enabled to find, see, summarise, relate, and comprehend changing and alternative relevant information effectively and efficiently and to record, report upon, and share discoveries. Sophisticated analytical tools with appropriate interactive visual interfaces for discovering relationships, synthesising knowledge, and making decisions can support this activity by providing the right people with the right information at the right time. Providing these is a challenging task, but one that can take advantage of a number of recent and developing technologies and scientific knowledge. To work effectively and enable beneficial decisions to be made, these tools must deal appropriately with the specifics of time and space.

5.3 Specifics of Time and Space

Most of the existing techniques for computational analysis, such as statistics and mathematical modelling have been developed to deal with numbers.

Temporal and spatial data have a number of properties that distinguish them from other types of data[8]. Unfortunately, their specifics are often ignored. Thus, temporal references and spatial coordinates are often treated just in the same way as ordinary numeric variables. This approach cannot yield valid analytical results. Time and space require special treatment and specific analysis methods.

Time and space are more than just numbers

5.3.1 Dependencies Between Observations

The processing, integration, and analysis of spatio-temporal data is both constrained and underpinned by the fundamental concept of spatial and temporal dependence. In the spatial domain, this is often referred to as 'the first law of geography' or 'Tobler's first law': "everything is related to everything else, but near things are more related than distant things"[112]. According to this law, characteristics at proximal locations tend to be correlated, either positively or negatively. In statistical terms, this is called spatial autocorrelation. Similar concepts of temporal dependence and temporal autocorrelation exist for relationships in time. Spatial and temporal dependencies forbid the use of standard techniques of statistical analysis, which assume independence among observations, and require specific techniques, such as spatial regression models, that take the dependencies into account.

The first law of geography

However, spatial and temporal dependence not only set constraints but also serve as sources of information and give important opportunities for data processing and analysis. Thus, spatial and temporal dependence enable:

- interpolation and extrapolation, which can be used to fill gaps in incomplete data,
- integration of information of different types and/or from different sources using references to common locations (spatial overlay),
- spatial and temporal inference,
- and many other operations (e.g., spatial and temporal navigation).

However, the effect of the first law is not absolute. In geography for instance, the law is weakened by the heterogeneity of the geographical space, where water differs from land, mountain range from valley, forest from meadow, seashore from inland, city centre from suburbs, and so on. Moreover, every location has some degree of uniqueness relative to the other locations. Spatial dependence is also affected by natural or artificial barriers. For example, the climate may significantly differ in two neighbouring valleys separated by a mountain range, and people's lives in two villages separated by a state border may also differ quite a lot. Similarly, temporal dependence may be interrupted by events; for example, radical changes may be caused by storms or floods. Relatedness between things may depend, not only on their distance (proximity) but also on direction. Thus, a flood or water pollution spreads downstream along a river. Events in time have an effect on future rather than past events. The notion of proximity is also phenomenon-dependent. It may be defined spatially, for example, in terms of distance by roads, rather than the straight line distance or distance on the Earth's surface. Temporal distances may be measured, for

Applicability of the first law

instance, in terms of ‘working days’ or ‘number of hours under particular conditions’ – inundation for example.

Some of these discontinuities, complexities and characteristics can be modelled and accounted for in informed spatio-temporal analysis. But it is impossible to account for all diverse factors affecting spatial and temporal dependence in developing fully automatic methods for analysis. Instead, visual analytics techniques may allow the analyst to see where and how the effect of the first law is modified by particular local conditions and to make necessary adjustments in the analysis, e.g., by varying parameters of analytical methods or choosing other methods (see also Chapter 4).

5.3.2 Uncertainty

Different sources and aspects of uncertainty

Unfortunately, in real world scenarios data is not always 100% perfect. The quality of data is often decreased due to errors, missing values, deviations, or other sources of uncertainty (see Chapter 3). Reasons might be, for instance, inaccurate data acquisition methods, data transmission problems, or even analytical processes such as spatial interpolation or temporal aggregation that result in loss of information. As of today, there is no consensus on the definition of uncertainty (often also denoted as ’data quality problem’); a universal way to visually represent uncertain data does not exist. One of the few closed definitions explains uncertainty as the “degree to which the lack of knowledge about the amount of error is responsible for hesitancy in accepting results and observations without caution”[60]. More generally, uncertainty can be considered a composition of different aspects such as:

- error – outlier or deviation from a true value,
- imprecision – resolution of a value compared to the needed resolution (e.g., values are highly accurately given for countries but are needed for states),
- accuracy – size of smallest interval for which data values exist,
- lineage – source of the data (e.g., raw satellite images or processed images),
- subjectivity – degree of subjective influence in the data,
- non-specificity – lack of distinctions for objects (e.g., an area is known to be used for growing crops, but not its specific kind), or
- noise – undesired background influence.

From an application oriented perspective, one can distinguish between different geometric uncertainties; geospatial, time, and thematic data uncertainty. Some of these concepts are quite different from others and might therefore require special treatment. What we are lacking is a unified term that subsumes the relevant kinds of distrust in some data.

To allow for effective analysis of spatio-temporal data, uncertainty has to be considered. Analytical methods must be tuned to the uncertainty in the data and visual representation have to convey inherently different kinds of uncertainty. Only if people are made aware of data quality problems and understand their implications, can visual analytics methods help them make informed decisions.

5.3.3 Scale

Spatio-temporal phenomena and processes exist and operate at different spatial and temporal extents. Thus, we say that a hail storm is a local, short-term phenomenon while climate change is global and temporally extended.

Scales of time

The dimension of time can include a single or multiple levels of scale (also called granularity of time). Temporal primitives can be aggregated or disaggregated into larger or smaller conceptual units. For example, 60 consecutive seconds are aggregated to one minute or five time steps in a discrete simulation model may correspond to one second in physical time. Most of the current tools for analysis and visualisation use models where the data is sequences of simple ⟨time-point, value⟩ pairs; only one level of granularity is considered. However, this is inadequate for a wide range of applications. For instance, in analyses related to hazard protection, it may be necessary to combine time scales with different granularities. For instance, the Decision Makers in our scenario would need to concurrently analyse outputs of simulation models with monthly resolution, data from weather forecast services specified for days, and annual estimates coming from prediction models of changing climate conditions (which in turn might have been mined from data based on decades or even centuries, see also Chapter 4). Developing methods and interfaces that achieve this is a challenging task, that is inadequately addressed by current methods of visualisation and analysis.

Scales of space

The scale of spatial analysis is reflected in the size of the units in which phenomena are measured and the size of the units in which the measurements are aggregated. It is well known in geography that the scale of analysis may significantly affect the results. For instance, patterns or relationships discerned at one scale may not be detected when examined at another scale. In extreme cases, opposite relationships may occur in the same place or time when different scales are considered. Such results can be regarded as highly scale dependent. Some phenomena and some places are more scale dependent than others. Representing this information numerically and graphically is a complex process.

Finding an appropriate scale is difficult

In order to observe and study a phenomenon most accurately, the scale of analysis must match the scale of phenomenon under consideration. Identifying the correct scale of phenomena is therefore a key problem for analysts. It is not always easy, however. In order to understand what scale of analysis would be adequate, analysts may need to use 'trial-and-error' approaches. Given spatial and/or temporal units of a particular size available in the original data, they can be aggregated into larger units in various ways. The opposite operation, decreasing the unit size, is only possible with involvement of additional data. Thus, in our example scenario, the scale of the data provided by the weather sensors was too large for examining the hail storm phenomenon. The analyst had to involve additional data to perform the analysis at an appropriate scale.

On the other hand, the scale of analysis should also be chosen according to the goals of analysis. As an example, in Figure 5.3, traffic data is visualised at

Figure 5.3: Analytic results are very dependent on the spatial scale used. At different scales, detailed or only very coarse traffic patterns can be made visible. (Source: produced using the CommonGIS visual analytics toolkit described in Andrienko & Andrienko[7], pp. 657-658)

different spatial scales and levels of aggregation: from individual trajectories of cars and aggregated flows between crossings and turns to large-scale aggregated flows between districts. The appropriate scale depends on whether the analyst needs to investigate the movement at a specific crossing and the adjacent streets, to detect the major routes of the traffic and to assess the traffic intensity on the major roads, or to consider the amount of movement between larger areas.

In aggregation, it is essential to be aware about the modifiable areal unit problem, which means that the analysis results may depend on how the units are aggregated. This refers not only to the sizes of the aggregates (scale effects) but also to their locations and composition from the smaller units (the delineation of the zones). Therefore, it is always necessary to test the sensitivity of any findings to the means of aggregation.

Furthermore, it is widely recognised that various scales of geographic and/or temporal phenomena interact, or that phenomena at one scale emerge from smaller or larger phenomena. This is captured by the notion of a hierarchy of scales, in which smaller phenomena are nested within larger phenomena. Local economies are nested within regional economies, rivers are nested within larger hydrologic systems, and so on. This means that analytical tools must adequately support analyses at multiple scales considering the specifics of space and time. Since time is still too often considered just as ordinary numbers, we next shed some light on what makes time such a special attribute.

5.3.4 Time

Granularity of time

In contrast to common data dimensions, which are usually 'flat', time has an inherent semantic structure, which is one source of increased complexity. By convention, time has a hierarchical system of granularities, including seconds, minutes, hours, days, weeks, months, years, centuries, and so on. These granularities are organised in different calendar systems. Furthermore, time contains natural cycles and re-occurrences. Some of these are regular and relatively predictable such as seasons, others are less regular such as social cycles like holidays or school breaks or economic cycles. In particular, two specific aspects of the dimensions of time have to be taken into account when devising analytical methods for temporal and spatio-temporal data.

Temporal primitives: time points or intervals

First, the temporal primitives that make up the temporal dimension must be considered. The temporal dimension can be viewed as composed of time points or time intervals. A time point is an instant in time. In contrast, a time interval is a temporal primitive with an extent. The choice of appropriate primitives must depend on the properties of the data and the problem at hand. Most of today's visual representations and analytical techniques do not differentiate between point-based and interval-based temporal data and do not represent the validity ranges of the data appropriately; and we know little about how to do this effectively.

Different structures of time

Secondly, the structural organisation of the temporal dimension is a relevant aspect. Three different types of temporal structures exist: ordered time, branching time, and multiple perspectives. Ordered time can be subdivided into two further subcategories: linear and cyclic time. Linear time corresponds to our natural perception of time as being a continuous sequence of temporal primitives, i.e., time proceeds from the past to the future. A cyclic time axis is composed of a finite set of recurring temporal primitives (e.g., the times of the day, the seasons of the year). Natural hazards such as flood events can also exhibit cyclic behaviour. To communicate the time patterns of such hazardous events and to allow for appropriate crisis management, this cyclic behaviour has to be represented. The concept of branching time facilitates the description and comparison of alternative scenarios, which is particularly relevant for planning or prediction. Time with multiple perspectives allows more than one point of view at observed facts. This type of time-related data is generated, in particular, when people describe their observations about hazard events via blogs or other

online means: each reporting person may have a distinct perspective on the events. While linear and cyclic time have already been addressed by existing visual analytics approaches, methods for analysing data related to branching time and time with multiple perspectives are still scarce. There is a need for methods that allow analysts to consider, compare and report upon different types of time in combination. Without such consideration, the complexities and subtleties of spatio-temporal data will not be accessible to analysts. This is important for risk management but also for other areas and problems. There may be hidden patterns in Dr. John Snow's data that would only be revealed through these perspectives.

Let us now look at the existing disciplines and technologies addressing the specifics of time and space.

5.4 State of the Art

5.4.1 Representation of Space

Cartography

Cartography has a long and venerable history

Cartography is the discipline dealing with the conception, production, dissemination and study of maps. Geographers and other professionals working with spatial data don't have to be convinced of the unique qualities of maps. They use them to express their ideas, to make a point, to obtain new knowledge and communicate among colleagues, and of course along with almost everyone else they use them to orientate and navigate. Outside the professional community, maps are also very much appreciated.

Maps represent spatio-temporal phenomena visually

Maps have the ability to present, synthesise, analyse and explore the real world. Maps do this well because they present a selection of the complexity of reality and visualise it in an abstract way. The cartographic discipline has developed a whole set of design alternatives and guidelines to realise the most suitable map that offers insight in spatial patterns and relations in particular contexts. The guidelines are partly based on conventions and partly on human perception. Examples of conventions are the use of blue hues to indicate water on maps in Western societies or the use of a colour scale from greens for lowland, via yellows to browns for mountains in topographic maps. Often these conventions are universal, but local exceptions do exist. Examples of perceptual design rules are the application of big symbols to represent large amounts and small symbols to represent a few items and legends that are designed to account for the non-linear perception of visual variables such as size.

Much has been done in cartography to address the issues of spatial scale. There is a dedicated sub-area of cartography called cartographic generalisation. Cartographic generalisation is the process of reducing multidimensional real-world complexity for depiction in a typically lower-dimensional map and entails reduction in detail as a function of depicting the world at a smaller scale. Cartographic generalisation is not just about filtering unnecessary

Figure 5.4: Maps show different graphical elements and different semantic content depending on the presentation's current scale. Reproduced by permission of swisstopo (BA100617) (Source: http://www.swisstopogeodata.ch/geodatenviewer/)

information, or information loss. It includes condensing the essential attributes (semantic generalisation) and preserving the geometric characteristics (graphic generalisation) of the depicted features. An example is given in Figure 5.4. As one moves from a larger scale representation across the scales to a small-scale representation, not only does the graphic density change, but also the meaning associated with the graphic marks. Thus, individual buildings are visible at the highest level of detail (large scale, high resolution), whereas only the size of the urban area, its shape and major transportation routes associated with the city may be relevant at lowest level-of-detail (small scale, low resolution)

Maps are very suitable for visual analysis. Co-location of patterns such as those between population density and recreation areas can often be seen at a glance. Cartographic theory and practice, much of which is based upon the interpretation of experimental results, enables us to show multiple themes in a single map. Cartography has developed techniques for the individual representation of particular types of phenomena and data and their effective combination enabling us to make use of the human perceptual and cognitive system to visualise several characteristics concurrently. We can, for example, compare terrain characteristics and land use, or use techniques for relief representation that can show key characteristics of topography such as slope, aspect and form concurrently. Initially, such maps were produced manually, but recently automated analytical techniques have been developed (see Figure 5.5).

Maps on the Internet

The Internet is changing the way that maps are produced, disseminated and used. Web maps are available to a wide and diverse population. They can be linked to a variety of sensors that make it possible to observe the current weather, traffic or water levels at any time during the day. Mobile devices ensure that these interactive real-time maps can be queried and contributed to anytime and anywhere. The cartographic discipline has also put lots of effort in usability research to determine whether maps deliver particular messages or achieve particular aims effectively. The existing design guidelines have been tested, but new technological developments continuously challenge these guidelines because new representations and interaction options become available. How, for example, do we make the best use of mobile devices to contribute to spatial

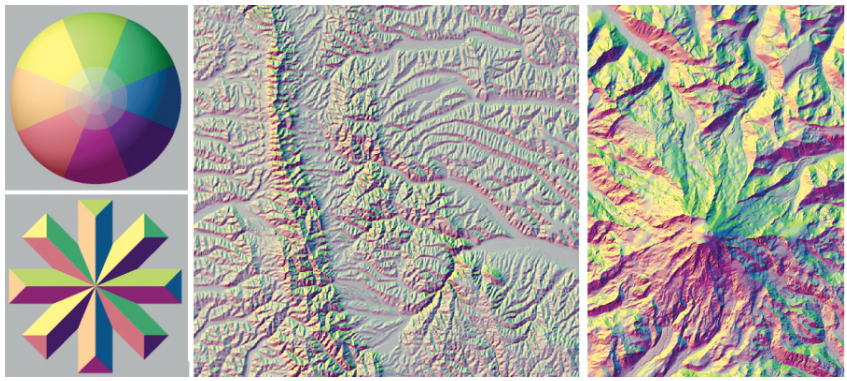

Figure 5.5: Multiple characteristics of topographic surfaces are visualised concurrently using combinations of hue, saturation and lightness. Image created with Landserf http://landserf.org/. Reproduced with permission of Jo Wood, giCentre, City University London. http://www.soi.city.ac.uk/~jwo/relief/

databases? How do we take advantage of opportunities for augmented and mixed reality applications?

Maps for the Information Age

The traditional role of a map is to 'present', but today the map should also be seen as a flexible interface to spatial data, since maps offer interaction with the data behind the visual representation. Additionally, maps are instruments that encourage exploration. As such, they are used to stimulate (visual) thinking about geospatial patterns, relationships and trends. In modern software systems, maps are combined with other types of graphical displays by dynamic coordination mechanisms, allowing, for instance, interactive probing for accessing multivariate data at different locations (see Figure 5.6).

Geographic Information Systems

GIS – a technology for millions

Most professional geographical analyses are undertaken with the use of geographic information systems or GIS. These systems combine data management, computational analysis, and map displays. GIS are widely used: the leading GIS vendor, Environmental Systems Research Institute of Redlands, California supports over 1 million users in 200 countries with more than 4000 employees. Recent reports of annual revenue are in the order of more than $600 million. Commercial GIS make steady incremental advances by incorporating cutting-edge research results from relevant scientific domains. These include GIScience through which a whole host of useful approaches that model, manipulate, summarise, project, generalise, relate and analyse geographic information have been developed. However, the main emphasis of GIS is on data management, transformation and computation and subsequent mapping. Their initial design deals well with (in today's terms) small, static spatial datasets and produces high quality static cartography that replicates and automates traditional paper-based mapping. Current GIS are weak in terms of the way in which they deal with

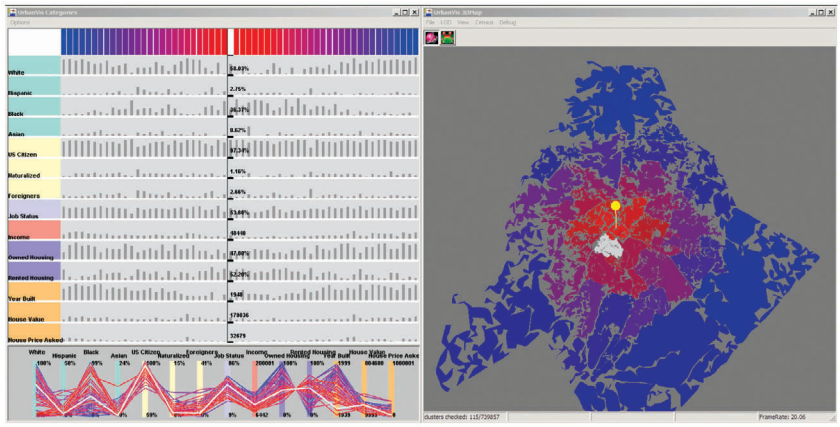

Figure 5.6: Multivariate socio-economic data associated with locations are explored by means of probes interactively placed on the map display. The visualisation on the left hand side is updated according to the probes location automatically through dynamic coordination[22] © 2008 IEEE

the temporal nature of geographic data. Time is routinely modelled as a high-level linear characteristic of spatial entities and maps and other analyses simply compare a limited number of particular moments rather than take advantage of the full structure of time.

The heritage of GIS means that they are not designed to support map use for interactive collaborative exploratory visual analysis. They are not designed to effectively deal with large dynamic datasets through a multitude of dynamic and novel displays that are considered by a range of disparate users. This legacy can be considered a significant hindrance to spatio-temporal visual analytics where dynamic maps are essential to the exploratory processes.

Geographic Information Science (GIScience)

Geographic information science, also known as geomatics and geoinformatics, is the academic theory behind the development, use, and application of geographic information systems. GIScience studies the fundamental issues arising from the creation, handling, storage, and use of geographical information. In particular, it deals with the representation of geographical information for computer processing, database design, efficient information retrieval, transformation of geographical information, and computational methods for analysis such as spatial statistics (see also Chapter 4). It also deals with the visual representation of geographical information; therefore, cartography can be considered as part of GIScience.

GIScience develops academic theories

GIScience does not deal well with space and time concurrently. Space always

comes first - due to the geographic and cartographic heritage. Geographers tend to think spatially ahead of temporally. There is a need to change this way of thinking.

Geovisualisation

Techniques and tools for interactive visual analysis of spatial and spatio-temporal data and for spatio-temporal decision-making are designed, developed and evaluated predominantly in the field of geographic visualisation, or geovisualisation. This developing research domain addresses the visual exploration, analysis, synthesis, and presentation of geographic data, information, and knowledge[38] and focuses on dynamic maps that are used to support exploratory processes. A characteristic feature of geovisualisation research is integration of approaches from multiple disciplines, including geography, geographic information science, cartography, information visualisation, data mining, and other cognate disciplines. The need for cross-disciplinary efforts to support the visual exploration and analysis of spatio-temporal data is a function of the growing size and complexity of the datasets that need to be analysed. The main achievements in the field of geovisualisation include developing cartography and GIScience in the contexts of large dynamic datasets. There is also the need for exploratory approaches through:

Geovisualisation for exploratory use of dynamic maps

- novel methods of visual representation for particular tasks, phenomena and data types;
- effective means of interacting with such displays that not only enable various kinds of visual queries but can rapidly change their appearance in response to user's manipulations;
- the development of knowledge and theory relating to responses to particular methods.

5.4.2 Representation of Time

Irrespective of the presence of a spatial component, data that embodies change over time poses challenges to all disciplines related to data visualisation and analysis. A wide repertoire of interactive techniques for visualising datasets with temporal components is available in the field of information visualisation. Figure 5.7 shows an example in which temporal patterns can be analysed at multiple scales. However, because it is difficult to consider all aspects of the dimension of time in a single visualisation, the majority of available methods address specific cases only – mostly the visualisation of data with a linear time axis. Moreover, as is the case in GIS, many of the current visual analytics and information visualisation systems do not include any special functions and techniques for dealing with all aspects of time but rather treat time as one among many other numerical variables.

Only isolated solutions for selected aspects exist

Visualisation of time and temporal data

The existing approaches can basically be categorised as techniques that visualise time-related data and techniques that visualise time per se. In the first case,

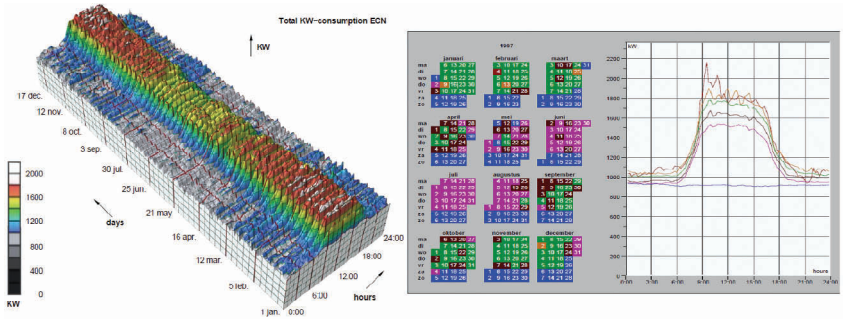

Figure 5.7: These visualisations of energy consumption data allow the analyst to investigate daily and seasonal temporal patterns[119] © 1999 IEEE

the focus is set on representing data, that is, quantitative or qualitative time-dependent attributes are represented with respect to a rather simple time axis (e.g., multivariate data represented with respect to linear time). The second case focuses on representing the characteristics of the time domain and its temporal primitives, while the representation of data is kept to a necessary minimum (for example, Gantt charts to represent relations between time intervals). In general, there are two options for visualising time and temporal data. Either we create a spatial arrangement of the time axis on the display or we utilise real world time, so that an animation shows visual representations of different time steps in quick succession.

Secondly, visual methods for temporal data can be categorised based on the time characteristics they were developed for:

- linear time vs. cyclic time,
- time points vs. time intervals, and
- ordered time vs. branching time vs. time with multiple perspectives.

Visualisation can reveal cyclic patterns

Figure 5.8 demonstrates the difference between linear and cyclic representations through an example related to patterns in human health data. While common line graphs are useful to show general trends and outliers, spiral visualisations address cyclic aspects of time-related data. The spiral's main purpose is the detection of previously unknown periodic behaviour of the data. This requires appropriate parametrisation of the visualisation method. Usually, it is difficult to find suitable parameter settings for unknown datasets. Therefore, it makes sense to support the detection of patterns either by applying analytical methods or by animating smoothly through different cycle lengths. In the latter case, periodic behaviour of the data becomes immediately apparent by the emergence of a pattern. Interaction facilities are needed to allow users to fine-tune the visualisation. Only then can we take full advantage of our perceptual system, e.g., in recognising patterns and motion.

Visualisation of temporal uncertainty

Whether temporal attributes are conceptually modelled as time points or time intervals, is another important characteristic that influences visualisation methods. Most of the known visualisation techniques that represent time-oriented data consider time points. Other approaches focus on representing temporal

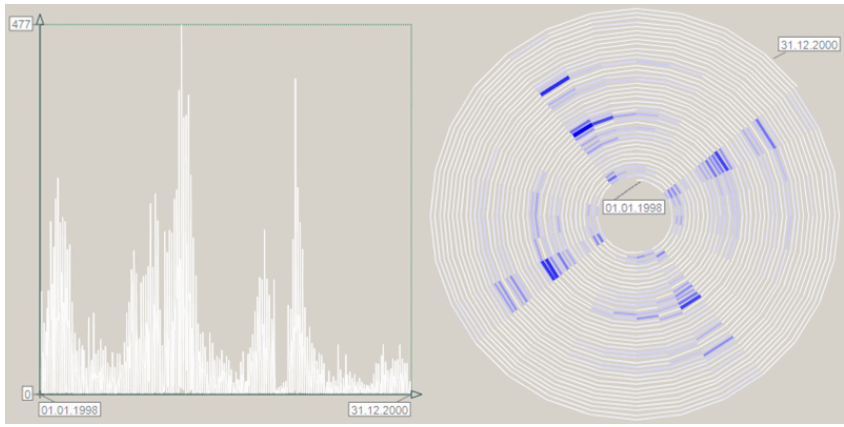

Figure 5.8: Two views of the same health related time series: In the linear plot (left) patterns can hardly be discerned. Switching to a cyclic spiral representation (right) makes an inherent cyclic (weekly) pattern apparent[3]

intervals and their interrelations. A particular challenge is the representation of uncertain temporal primitives, be it imprecise specifications of time points or fuzzy interval boundaries. Uncertainty might be introduced by explicit specification usually connected with future planning (e.g., "The meeting will start at 11 a.m. and will take approximately one hour" – which means that it is not quite clear when the meeting will be over) or is implicitly present in cases where data is given with respect to different temporal granularities (e.g., days vs. hours).

Visualising structures of time

Most of the visualisation techniques for time-related data known in the literature are suited to represent ordered time. Branching time and time with multiple perspectives, however, are definitely relevant types of time in visual analytics, especially when it comes to analysing data from heterogeneous sources like different sensor networks or public online forums, and when predictions of possible future scenarios are required. The few techniques for representing the latter types of time are capable of depicting only univariate qualitative data, or even visualise temporal primitives only; they can neither represent multiple time-dependent attributes nor are they combinable with visual representations of space, predominantly geographic maps. There is a strong need for advanced techniques to effectively visualise multivariate data exhibiting these specific time characteristics.

Aspects of time must be addressed adequately

The characteristics of the dimension of time have to be considered when devising new visual analytics methods for spatio-temporal data; integrating appropriate interaction methods is a key concern. Casual users and expert analysts must be allowed to adapt visual representations and analytical processes to a variety of tasks, including exploration in time, search, comparison, prediction, and manipulation. Only adequately adapted visual analytics techniques can fully support a broad range of users in reasoning about time- and space-

dependent data.

5.4.3 Interactive Methods for Visual Exploration

Time and space visually combined

Various applications and prototypes have been developed to facilitate and, moreover, stimulate data exploration. Figure 5.9 gives an example of an interactive map that dynamically changes its appearance to support visual detection of spatial patterns. Animated maps portray time-dependent data and dynamic phenomena by mapping the temporal dimension in the data to the physical time as it is experienced by the onlookers. The interactive space-time cube is an important visualisation technique for spatio-temporal data (see Figure 5.10). It implements one of the ideas of time geography[53], which considered space and time as inseparable and suggested a three-dimensional visual representation where two dimensions encode spatial aspects of the data and the third dimension represents time. This is contrary to commercial GIS architectures which fundamentally separate the spatial and temporal aspects of geographic information.

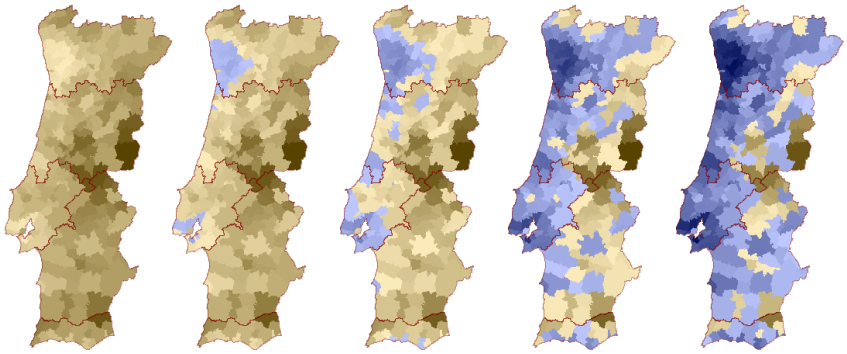

Figure 5.9: An interactive map display can dynamically change its appearance (here, the encoding of the data by colour shades) and in this way support perception and exploration of patterns of the spatial distribution. (Source: produced using the CommonGIS visual analytics toolkit described in Andrienko & Andrienko[7], pp. 657-658)

Coordinated multiple views visualisation

Typically, the size, dimensionality, and other complexities of data being analysed preclude the simultaneous representation of all data items, dimensions and relationships in a single display. Hence, the analyst has to understand the whole by looking at subsets, components, projections and selected aspects of the data. Some aspects of spatio-temporal data cannot be effectively visualised by means of maps alone and require other display types. Therefore, the effective combination of maps with statistical graphics and other visualisation techniques for temporal or structural aspects of the data are required. Since any view can only convey partial information, the analyst needs multiple views. These need to be linked so that the pieces of information contained in them can be related.

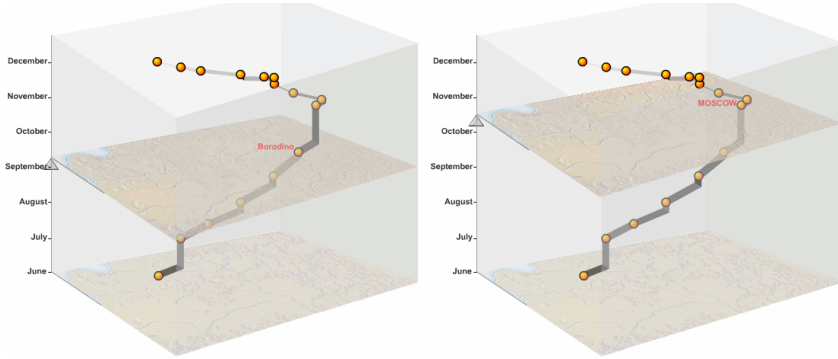

Figure 5.10: This interactive space-time cube portrays the movement of the Napoleon's army during the Russian campaign of 1812. By interactively moving the horizontal plane, the user can better see the positions of the army at different time moments. (Source: http://www.itc.nl/personal/kraak/1812/minard-itc.htm)

Means of enabling and symbolising selections and links are at the core of much visualisation activity.

Interactive dynamic filtering

One of the possible mechanisms for linking multiple displays is through 'dynamic filtering' (see Figure 5.11). The map (A) and the space-time cube (B) show the same subset of a dataset about 10,560 earthquakes selected by means of three different filters: 1) spatial window, which has been drawn by the user within the map display; 2) attribute filter (C), which selects the earthquakes with the magnitudes 4 or more; 3) temporal filter (D), which selects the earthquakes that occurred in the period from the beginning of 1995 until the end of 1999. All three filters may be interactively changed by the user; in response, the displays will immediately change their content to satisfy the new filter conditions.

5.4.4 Effectiveness of Visual Techniques

Empirical evidence through user studies

Visual analytics is different from 'standard' approaches to analysis. It is based on the assumption that interactive visual representations can amplify human natural capabilities for detecting patterns, establishing links, and making inferences. This assumption, however, needs to be empirically validated in order to arrive at effective visual analytics approaches for spatio-temporal data. Particularly in cartography there is a tradition for obtaining empirical evidence by means of experiments in which people use different variants of maps and graphics to find the information necessary for answering certain questions.

In some experiments, the measurements of the accuracy of the answers and the time spent seeking information are combined with methods that track the eye movements of those being tested. In this way, for example, it was found that people have difficulties in retrieving relevant information from colourful

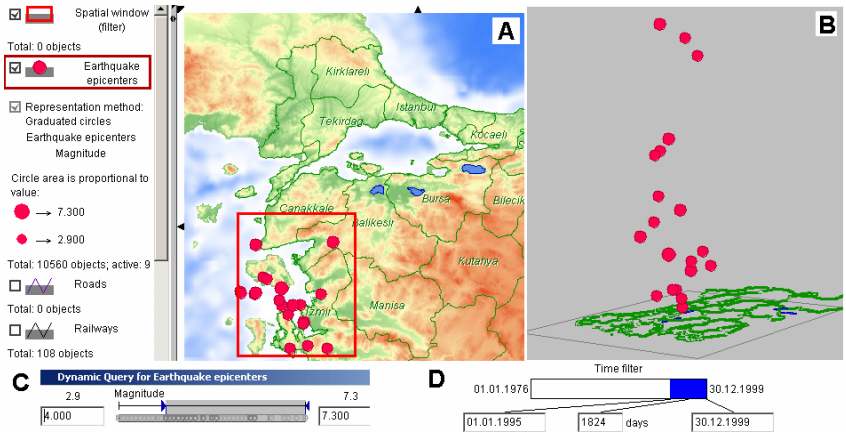

Figure 5.11: Linking multiple displays via 'dynamic filtering'. (Source: produced using the CommonGIS visual analytics toolkit described in Andrienko & Andrienko[7], pp. 657-658)

weather maps, which are often published in mass media. They are more able to detect what is relevant on a carefully designed, cognitively adequate map that uses established cartographic and visualisation principles to depict the same information content (see Figure 5.12).

However, the empirical studies conducted so far have addressed only a small fraction of existing techniques. We still know very little about the perception and use of interactive dynamic maps, different representations of time, three-dimensional and large scale displays, and maps combined with other graphics or multimedia content. Furthermore, new interaction devices and corresponding interaction methods need to be evaluated with regard to their usefulness for spatio-temporal visual analytics, for instance, map navigation and temporal browsing. Chapter 8 discusses general aspects of evaluation in more detail.

5.4.5 Dealing with Larger Data Sets

Large and complex data is difficult to deal with

The traditional visualisation approach involves the direct depiction of each record in a dataset so as to allow the analyst to extract noteworthy patterns by looking at the displays and interacting with them, as illustrated in Figures 5.6 ,5.9, and 5.10. However, these techniques may not be effective when applied to very large and complex datasets that are increasingly common. The displays may become illegible due to visual clutter and massive overplotting associated with large numbers of cases – would Dr. Snow have noticed the relationship between cholera deaths and pump location if he had access to the locations of all deaths and all pumps in London? For example, the Times Labs blog asks "How perilous is it to cycle where you live?" alongside a typical online map in which overplotted symbols render the question unanswerable (see Figure 5.13). Users may also have difficulty perceiving, tracking and

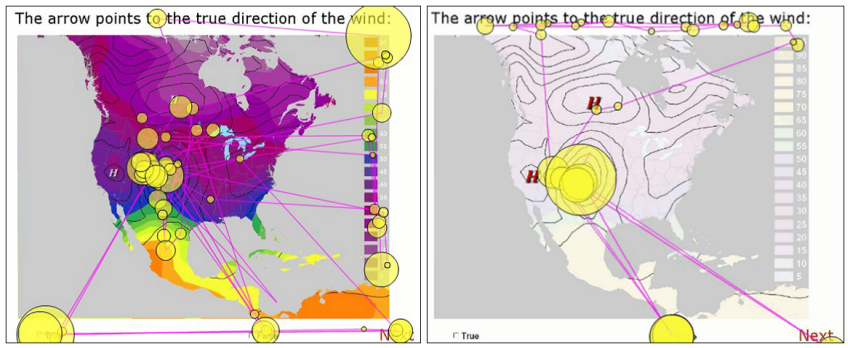

Figure 5.12: The yellow circles represent the locations and durations of eye fixation of the test participants seeking relevant information in two variants of a weather map with the same content. Those who used the cognitively more adequate map (on the right) detected the relevant information (located in the centre of the map) immediately

comprehending numerous visual elements that change simultaneously. The technology may not be sufficiently powerful to update the display fast enough or respond quick enough to user interactions, to enable efficient inference making.

Aggregation in the data domain

Two alternative approaches are being increasingly utilised in response to the current challenges. One modifies the traditional visualisation approach by involving methods for data aggregation and summarisation prior to graphical depiction and visualisation (see Chapter 3). The other approach involves applying more sophisticated computational techniques, such as those developed in data mining, to semi or fully automatically extract specific types of feature or pattern from data prior to visualisation. This visual data mining approach may apply to summaries and along with the visualisation of summaries, may take advantage of ideas and advances developed in direct depiction. Figure 5.14 gives an example of combining geovisualisation with data mining (see Chapter 4), specifically, the method known as 'Self-Organising Map', or SOM, which reduces the dimensionality of multivariate data to two dimensions and simultaneously groups items with similar characteristics. In this example, SOM has been applied to aggregate data about companies that operated in the USA during 12 years. By grouping the data, SOM has derived a number of distinctive economical profiles in terms of the activities of different industries, such as computer hardware, computer software energy, telecommunications, etc. The map series, where the states are coloured according to these profiles, allows the analyst to investigate how the profiles of the states changed over time. Thus, the analyst can notice that by the end of this relatively short time period no states remain where software companies would play a leading role. For larger time series or more complexly structured time axes or maps, currently existing techniques reach their limits. Future visual analytics methods will still have to face the challenge of dealing with immensely large datasets.

Figure 5.13: The overplotted symbols on the map hinder the analysis. © 2009 Google, Map Data © 2009 Tele Atlas. (Source:http://labs.timesonline.co.uk/blog/2009/03/11/uk-cycling-accidents/)

5.4.6 Collaborative Visualisation

Support people in working together

A currently emerging and very important research direction is collaborative visualisation[74] – design and use of technologies to enable groups of analysts to work productively with spatial and temporal information. The need for such approaches in which tacit knowledge is pooled is evident from our scenario.

Collaboration research addresses the following issues:

- collaboration: how interactive visual interfaces (in particular, map interfaces, which are essential for spatial problems) can enable many actors to work together in the same room, between rooms, between offices, between countries, or even between cultures;
- communication: how interactive visual interfaces can facilitate effective transfer of spatially and temporally-related information, knowledge, evidence, judgements, considerations, etc. from one actor to another.

5.4.7 Fundamental and Theoretical Research

Along with the progress in designing and developing innovative techniques and tools, a substantial amount of theoretical research has been undertaken

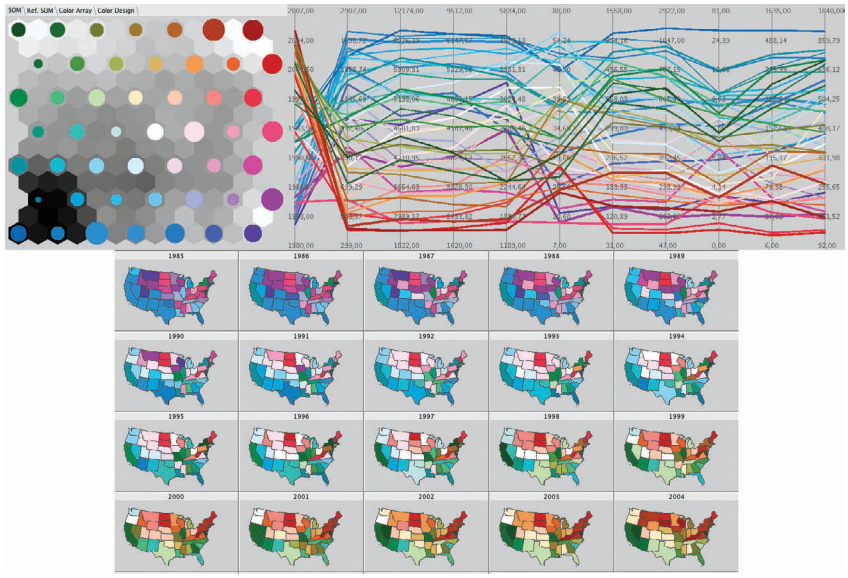

Figure 5.14: Space and time-referenced multivariate data is analysed with the use of SOM - Self-Organising Map, a computational method that groups and orders data according to the values of multiple variables[51]

A basis for spatio-temporal visual analytics research

in the areas of geovisualisation and information visualisation. The most essential monographs, which not only lay theoretical foundations and explain the principles of geovisualisation and information visualisation but also analyse the state of the art in the area and outline the main research directions, are:

- MacEachren (1995), How Maps Work: Representation, Visualization, and Design[73]
- Slocumet al. (2009), Thematic Cartography and Geovisualization[103]
- Spence (2006), Information Visualization - Design for Interaction[104]
- Ware (2004), Information Visualization: Perception for Design[122]
- Kraak and Ormeling (2003), Cartography: Visualization of Spatial Data[69]
- Andrienko and Andrienko (2006), Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach[7]

These monographs can be considered fundamental in their respective fields. In combination, they can serve as starting points for spatio-temporal visual analytics. However, more joint research is necessary to address the specifics of time and space in an holistic way.

5.5 Challenges and Opportunities

Whilst geovisualisation and information visualisation have developed some effective techniques for supporting the visual exploratory analysis of geographic and time-dependent information, it is much more complex to effectively support analysis of spatio-temporal data. Thus, the cyclical temporal patterns shown in Figure 5.3 may only occur in particular places - or may differ across space and spatial scale. Some steps have been made towards simultaneous handling of space and time in a more complete and sophisticated manner, but there is still much to do.

Besides, a number of other challenges face those analysing spatio-temporal phenomena, and these are the subject of the following sections.

5.5.1 Dealing with diverse data

Data from sensors, aerial and satellite imagery extend the opportunities for informed decision-making

We have seen that increasing amounts of spatio-temporal data is becoming available from various kinds of sensors, aerial and satellite imagery, statistical surveys, and many other sources. These datasets have the potential to significantly extend the opportunities for comprehensive analyses and informed decision-making. In parallel, data accessibility is improving. This is being achieved through the design and development of spatial-temporal information infrastructures (see also Chapter 6), standards for data, metadata, and services, and legislative regulations concerning the collection, quality, organisation, sharing, and use of data (see also Chapter 3). For example, OGC (Open Geospatial Consortium, Inc.) develops international standards to make complex spatial information and services accessible and useful with all kinds of applications. The INSPIRE initiative works to establish the Infrastructure for Spatial Information in the European Community, enabling spatial data from different sources across the Community to be combined in a consistent way and shared between several users and applications. Furthermore, a variety of models, concepts, algorithms, and data structures have been developed in the area of temporal databases.

However, this progress concerning the collection and accessibility of spatial and spatio-temporal data poses new challenges related to:

- new types of data, for which no analytical methods yet exist;
- large amounts of data, with which current analysis methods cannot cope,
- dynamic data arriving in real time, which require highly efficient methods capable to combine previous results with new data;
- data of diverse types, which need to be analysed in combination;
- data of diverse quality and inconsistent data from multiple sources, which need to be harmonised.

In our example scenario, the analysts combine official geographic information about The City with measurements from sensors, reports about incidents, trajectories of cars, phone call data, satellite images, outputs from simulation models,

and historical data about similar events in the past. This is not yet feasible, but the means to address this challenge are emerging.

Solutions needed to enable integrated processing and analysis of diverse data

Hence, visual analytics has to do more than just developing adequate methods to visualise and analyse different types of data, large amounts of data, and dynamic data. Visual analytics must also devise solutions for enabling integrated processing and analysis of diverse data.

As a prerequisite for any analytical task, analysts must first look at the data and identify uncertainties, inconsistencies and missing items. Only then can the data be pre-processed accordingly to make it suitable for analysis:

- ameliorate incomplete data by deriving missing parts from related data and from simulation models;
- harmonise inconsistent data by cross-checking with related data and knowledge;
- enrich and refine the data by deriving relevant new characteristics and constructs.

These preparatory operations have to be facilitated by appropriate visual analytics tools. In our scenario, the insurance analysts initially had incomplete data from the weather sensors. They used interactive visual methods to transform community-contributed unstructured information into structured data, which were fed into a statistical model for estimating the course of the storm. When the analysts viewed the model results and the observation data together they were able to derive a probable perimeter of the storm-affected area.

The prepared data is then subject to detailed analysis. At this stage too, analysts need to combine diverse data, for example, the estimated affected area, typical traffic flows, and the spatial distribution of insurance clients. Again, analysts have to be supported by visualisation and interactive tools working in synergy with appropriate computational techniques.

5.5.2 Support for Analysis at Multiple Scales

What visual representation is appropriate for different types of data at different spatial and temporal scales?

There is much to do for visual analytics in order to change the traditional practice in analysis, focusing on a single scale. As explained earlier, appropriate scales of analysis are not always clear in advance and single optimal solutions are unlikely to exist. Analysts may need to use ‘trial and error’ approaches. Interactive visual interfaces have a great potential for facilitating the empirical search for the acceptable scales of analysis and the verification of results by modifying the scale and the means of aggregation. To realise this potential, we need to know more about appropriate visual representation of different types of data at different spatial and temporal scales. We need to develop corresponding analysis-supporting interaction techniques, which should enable not only easy transitions from one scale or form of aggregation to another but also comparisons among different scales and aggregations.

Since various scales of geographic and temporal phenomena interact, analytical tools must also fully support analyses at multiple scales. Future research must answer the question: How do we help (a range of) analysts uncover and understand cross-scale relationships between phenomena?

The research on scale issues in visual analytics has to utilise and build upon a number of achievements from other disciplines. One is cartographic generalisation, including theory, best practices, and algorithms for automatic geometric and semantic generalisation of many types of data. However, cartographic generalisation is restricted to maps and does not give guidelines for other types of displays. Generalisation research has also been focussed on traditional cartography rather than on dynamic maps for exploration. Too little work has focussed on the generalisation of time-dependent data at different temporal scales. Generalisation usually considers representations only, rather than user-display interaction. How do we match interaction techniques to the scale of analysis? The concerted efforts of visual analytics researchers from various backgrounds will be required to address these issues.

5.5.3 Understand and Adequately Support Diverse Users

Non experts and experts alike need appropriately designed analytical tools

Professional analysts are usually specially trained. In particular, professional spatial analysts get training in the use of geographic information systems (GIS) and methods of spatial statistics. However, we argue that virtually everybody is now a spatio-temporal analyst. Of course, it cannot be expected that everyone receives special training before starting to analyse spatio-temporal data and making space and time-related decisions. Still, there is a need to provide this wide range of spatio-temporal analysts with adequate analytical tools that they are able to use effectively. How can this be achieved?

Fortunately, many potential users of visual analytics tools are relatively sophisticated in terms of their use of information systems. They are experienced in using computers and the Internet. They are familiar with dynamic displays of spatio-temporal information, such as weather maps shown on television. By playing video games, people become experienced from early childhood in interacting with dynamic visual displays. Adults often use online mapping services and have no problems with basic interactive operations such as zooming, panning, and selection. Virtual globes, in particular, Google Earth and Microsoft's Virtual Earth are increasingly popular and the globe is becoming a sufficiently important metaphor for manipulating spatial information to challenge the dominance of the map.

Hence, a certain level of computer and graphical competence can be expected from the potential users of visual analytics tools for spatio-temporal analysis and decision making. We can also expect that motivated users will not mind acquiring a reasonable amount of new knowledge and skills. The problem is how to appropriately convey this knowledge and these skills to the users?

On the other hand, visual analytics is different from 'standard' approaches to analysis. It is based on the assumption that interactive visual representations

can amplify human natural capabilities for detecting patterns, establishing links, and making inferences. The amplification of human perceptual and cognitive capabilities is not something achievable merely through training. While it is possible to explain to the users how to interpret a display and how to use interactive devices, the users can hardly be trained to gain insights from graphics and to reason more efficiently with the help of graphics. What matters here is the design of the visual representations and accompanying interaction techniques.

Design guidelines required for interactive, dynamic and multimedia maps

While a number of useful design rules and guidelines exist in cartography, the design of interactive maps, dynamic maps, three-dimensional displays, multimedia maps and maps combined with other graphics are still lacking any guidelines, and available empirical evidence is fragmentary and hard to generalise. Furthermore, we still know very little about the effectiveness of visual displays in supporting more sophisticated activities than answering simple questions typically used in experimental studies, specifically, exploratory data analysis, problem solving, knowledge synthesis, and decision making. These issues definitely require thorough research, which is vital for creating usable and useful visual analytics tools. This research requires interdisciplinary efforts involving computer scientists, cartographers, psychologists and cognitive scientists.

5.5.4 Reach the Users

Create GISs that are temporal and analytical, with an interactive visual emphasis

Geographic information systems (GIS) are and will remain in the future the main instrument for professional analysis of spatial information. The cutting-edge visualisation work being reported by research laboratories across Europe suggests possible solutions that can be adopted by the GIS industry. However, we should not just passively wait for this to happen. We can instead work on creating GISs that are temporal and analytical, with an interactive visual emphasis. We can realise the concept of spatio-temporal visual analytics as the new applied dynamic GIS that must take advantage of the range of useful algorithms and research in GIScience, GIS, geovisualisation, and information visualisation, the public interest in and experience of spatial and temporal data, the Internet, and the emerging display environments (e.g., multi-touch tables or smart display rooms), and overcome the legacy of static paper maps and traditional cartography that are based upon this model.

Spatio-temporal visual analytics draws from GIS, cartography and information visualisation, but needs to deal with the dimension of time much more effectively. Everything is geared towards the key objectives:

- deal with and make use of characteristics of time and
- deal with and make use of characteristics of space.

In the light of visual analytics, we have to develop approaches to support sense-making from new and large datasets and to allow users to generate evidence and communicate knowledge about the data. The solutions must be visual, interactive, exploratory, scalable, collaborative and lightweight. This ambitious

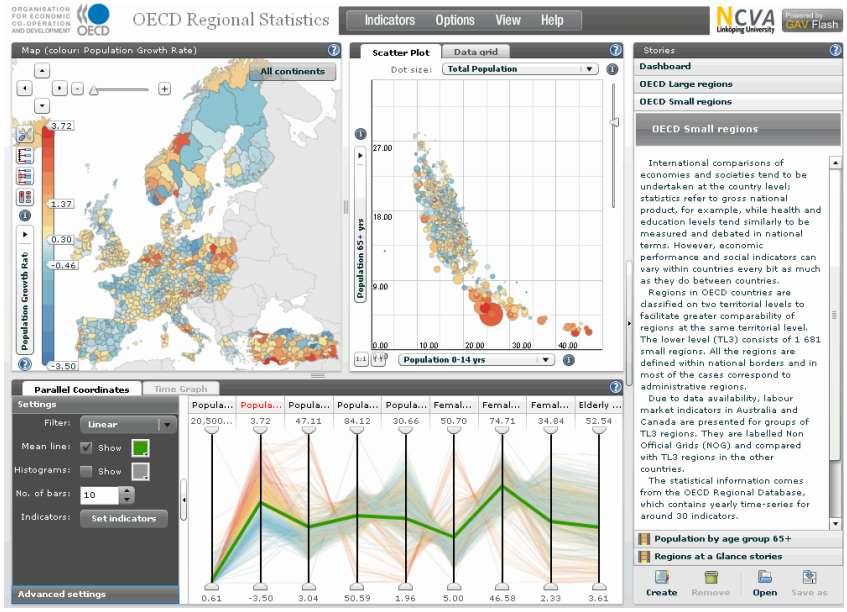

Figure 5.15: The interactive visual system OECD eXplorer allows specialists and general public to explore regional statistics data from OECD (Organisation for Economic Cooperation and Development, http://www.oecd.org/home/).

endeavour can only be solved through derivation of knowledge from and through close collaboration with other disciplines.

Software for spatio-temporal visual analytics should be lightweight, easily deployable and usable, rather than huge and complex like GIS, which require extensive training. Users may be especially happy if the analytical instruments they need are available as Web services or through an open APIs. The developers of visual analytics tools should strive to make their tools not only useful and usable but also accessible to users. A good example is OECD Explorer (Figure 5.15), a popular and impressive Web service that contains innovative means for recording and discussing findings. The system is implemented on the basis of the Flash/Flex platform, which is, on the one hand, suitable for enabling various interactive operations and dynamic displays, on the other hand, easily accessible to many Internet users through a Web browser plugin.

There are also other things to consider in implementing visual analytics tools:

- seamless integration of visualisations with computational techniques such as spatial statistics, time-series analysis, simulation models, spatio-temporal data mining, etc.,
- support for documenting the analysis process, keeping provenance of finding, reporting and storytelling,

- support for collaboration.

These requirements are not unique for tools dealing with spatio-temporal data but generally apply to all kinds of visual analytics software. However, the specifics of space and time definitely have an impact on implementing the requirements, which may be by itself a research topic.

5.6 Next Steps

In order to progress in the field of geo-spatial visual analytics, the following actions should undertaken:

- Develop approaches to support analysts in finding satisfactory scales of analysis, exploring and establishing scale dependency, verifying discovered patterns and relationships at different scales and with different aggregations, and understanding dependencies between phenomena operating at different scales in time and space.
- Develop scalable visual analytics solutions to enable integrated processing and analysis of multiple diverse types of spatial, temporal, and spatio-temporal data and information, including measured data, model outputs, and action plans from diverse official and more uncertain community contributed sources.
- Improve the understanding of human perceptual and cognitive processes in dealing with spatial and temporal information and visual displays of and interaction with such information. On this basis, develop appropriate design rules and guidelines for interactive displays of spatial and temporal information.
- Develop effective solutions for training both specialist and non-specialist users interested in undertaking spatio-temporal analysis.
- Develop a new generation of lightweight accessible dynamic visual analytics tools to support a range of personal and professional spatio-temporal analysts in the best possible way.
- Implement tools for spatio-temporal visual analytics in the way that allows rapid and easy deployment or online use through the Web. Make the tools compliant with the existing and emerging standards, interoperable and combinable; enable integration of the tools into user's existing workflows.

6 Infrastructure

6.1 Motivation

Supporting the strong demand in data storage, computation and interactive performances required by visual analytics applications is still a challenge. All currently existing visual analytics applications need to build their own specialised infrastructure for their specific problem. This is a normal stage of evolution in a domain of information science, as explained by Brian Gaines in his BRETAM model[46] (Figure 6.1). This model suggests that all research domains and fields run through the same stages. A new research domain or phenomenon starts by a Breakthrough – a discovery that can be physical or conceptual – followed by a Replication stage when the scientific community tries to replicate the initial discovery and test its limits. The next stage is Empiricism when researchers find empirical laws that can be used to apply the discovery. After that, some Theory is found that allows a deeper understanding and usually makes predictions about the phenomenon. The next stage is Automation when the phenomenon is totally accepted, followed by the Maturity stage when it is used routinely without question.

Visual analytics is only at the Replication stage

As the model describes, each domain should pass several stages before reaching maturity and this chapter plots a possible path to achieve this evolution successfully and effectively, when visual analytics as a whole is only at the Replication stage.

Visual analytics is both user-driven and data-driven

One of the most difficult issues of visual analytics is that it is both user-driven *and* data-driven. It is user-driven because during the interactive steps of the analysis, the user is specifying algorithms and parameters to explore the data. It is also data-driven because new data is made available to the user at unpredictable times, such as when algorithms run or databases are updated,. Traditionally, the domains described in chapters 3, 4 and 5 have created software infrastructures that are mostly data-driven with some exceptions for geographical visualisation systems. Conversely, visualisation systems are user-driven and manage mostly static data.

Software not designed for interaction or dynamic data is very difficult to adapt

Therefore, assembling technologies created by these multiple domains is a difficult challenge because the software infrastructures they currently rely on are incompatible at a deep level: when software is not designed to react to changes in the data or triggered by the user, it is very difficult to modify it later.

Interactive systems used to drive the analysis need to provide sub-second reactions to the user's actions. Furthermore, visualisation systems, required to understand large datasets visually, require the screen to be updated in less than 100ms following user action. In contrast, current databases serve transactions in

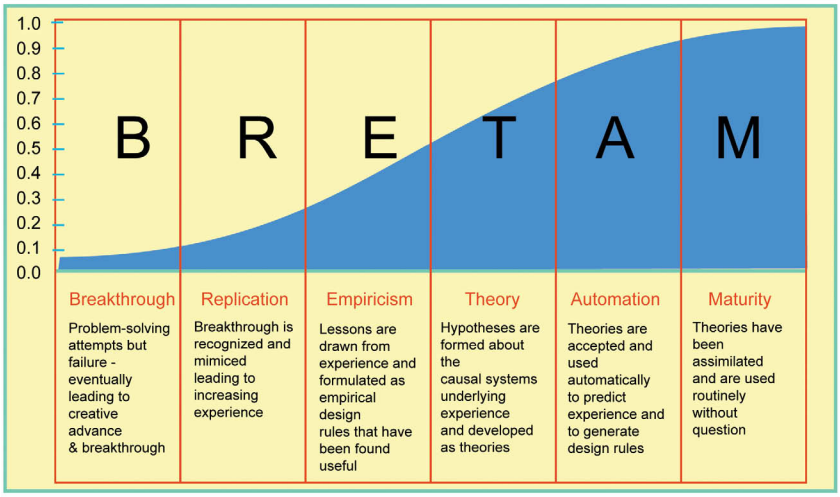

Figure 6.1: The BRETAM sequence plotted along the underlying logistic learning curve[46]

seconds and data mining algorithms available today are meant to run until completion, which can take minutes, hours or even days.

To overcome this situation, practitioners in visual analytics have started to implement ad-hoc systems, such as in-memory databases or user-steerable algorithms. However, this is not a sustainable solution in the long term for several reasons:

- **Loss in quality** When visualisation practitioners need to implement visual analytics systems, they cannot use off-the-shelf data-storage components or data mining components and hence need to implement them with their often limited skills of the domain. If data mining practitioners need to implement the same system, they will have problems integrating visualisation and interaction to a dynamically running analysis system. The problem is similar for a data-management practitioner.
- **Loss in resources** Since there is no accepted software architecture reference model for visual analytics, each system implements its software components in slightly different ways, leading to incompatibilities and no interoperability. This is becoming a bottleneck in the evolution of the field because most of the modules needed are difficult and expensive to implement and the lack of interoperability hinders sharing them within the community.
- **Loss in research time** Because research groups have to re-implement the visual analytics modules they require, they loose valuable time that would be better used for innovation.

Standards in visual analytics components will create a new market

- **Lack of component market** Since no standard exists, no commercial market can emerge for components. Several European companies sell visual analytics components but their market remains small at this level compared to other software components.

6.1.1 Examples

By taking the role of various actors in visual analytics, the software infrastructure issues are much easier to understand.

Exploration of Hierarchical Clustering from an Information Visualisation Viewpoint

Hierarchical clustering is one of the most popular clustering techniques used to make sense of large datasets. A large number of items (e.g., documents, genes, files, persons) are grouped according to a similarity criterion. Documents can be grouped according to the similarity of their textual contents, or simply because they share an author. Genes can be grouped because their DNA sequences are very similar, etc. The outcome of hierarchical clustering is a tree (or a direct acyclic graph) and the information visualisation community has a long tradition, as well as a collection of visual representations and interaction techniques, to navigate such trees. So, once the data has been hierarchical clustered, it can be visualised and explored using well-known and effective techniques.

However, in real life, computing good and meaningful hierarchical clustering is difficult and a push-button approach to clustering is likely to produce an incomprehensible hierarchy. Several issues should be considered when performing such clustering: what similarity measure to use, what attributes to select, how to deal with outliers and missing values, to name a few. The statistical analysis community has extensively studied these issues and also provide a wealth of quality measures to validate clustering, but choosing the similarity measures, the attributes and the validation method add extra complexity to the process that is now essentially made by trial and error.

Very few systems have effectively combined information visualisation with hierarchical clustering. HCE[96] is one example specialised for biological applications. It has required its author, a specialist in information visualisation, to re-implement popular hierarchical clustering algorithms and similarity metrics computation to offer the level of interaction required to achieve successful clustering. However this work is only applicable to one applied domain and therefore cannot be used in other domains. Breaking down such an application in modular components that could be assembled to suit other application domains in a modular, extensible and reusable way is currently not possible, due to the lack of a unified architectural model and appropriate software architecture to support it. Furthermore, to meet the interactive demands, the algorithm itself has to be programmed by an information visualisation specialist. Apart from the loss of time for the specialist, it may limit the level of sophistication of analytic components added to the information visualisation application since few information visualisation specialist are also specialists in statistical analysis.

Mining and Exploring Long Time Series from a Data Mining Viewpoint

VizTree[71] is a visual tool for mining large time series (Figure 4.5). Instead of using time dependent values directly, VizTree encodes it as a complete tree with a width and depth that can be specified interactively. The data mining part is very smart since the change from a long series of value into a tree simplifies greatly, many kinds of computation and allows for interactive search of patterns. However, the graphical interface of VizTree is very simple and the interaction is limited, with simple interactions such as selection of a time-range being done through form-based entries rather than by direct manipulation. Furthermore, VizTree is meant to mine long time-series, but as it reads flat files rather than make use of a database, its range is restricted. Again, the authors were specialised in one domain and did not use a ready-made software framework to implement their visualisation and interaction in a more effective way; they had to re-implement the missing parts in the best way they could, far from the state of the art in information visualisation, HCI and data management.

Database and Other Viewpoints

Good engineering practices imply separation of concerns without sacrificing quality

Further examples of this kind can be seen in the database field or from other kinds of analytical domains (video analysis, text analysis, etc.). The message here is that to build real applications, all these domains need each other's expertise, but currently, due to deep infrastructure model incompatibilities, they cannot connect the pieces of software together. Once all these domains agree on a conceptual model and implement it in their tools and modules, interoperability will become possible and cross fertilisation will become simpler.

6.1.2 Conclusion

To build demanding visual analytics applications, we need a new conceptual software architecture, a good separation of purpose between different stages of this software architecture and a good decomposition in components. Once we have agreed on this architectural model, we can create a new market of high-quality interoperable components to build the applications needed to transform the current flood of data into an opportunity for discoveries. These components, commercial or free, would allow researchers to focus on their domain of interest and skills and to push the research forward effectively. They will also increase the competitiveness of commercial companies by allowing them a better understanding of the market trends.

Designing the conceptual architecture is not simple because it is both user-driven and data-driven. Visual analytics is based on empowering human analysts and allowing them to apply complex analytical computations while maintaining interactive feedback and control. Most current analytical components are designed to run without interruption, delivering their results at the end.

For large datasets, this can take hours or days. Visual analytics needs analytical techniques that adapt to the human analysis process, not the other way around. As quoted by Thomas[111], chapter 5:

> Create visual analytics data structures, intermediate representations and outputs that support seamless integration of tools so that data requests and acquisition, visual analysis, note-taking, presentation composition, and dissemination all take place within a cohesive environment that supports around-the-clock operation and provides robust privacy and security control.

Even when these components are well understood, even standardised, more research on data typing is needed. Current databases only expose storage types (e.g., bytes, long integers, etc.) but not their meaning. Exposing the semantic type of a data is essential, in order to know what kind of analyse can be applied and what kind of visualisation is meaningful. An integer value can be used to represent a nominal identifier such as a hash value, it can also represent a day of the week or month or a true numeric value. SQL databases do not express the semantic of the numbers stored. Data mining systems usually classify data as nominal, ordered, numerical and ratio. This classification is rich enough for most statistical treatments but not sufficient for visualisation. The semantic web is an example of an application domain where sophisticated data types are being defined but there are also other initiatives and it is not clear how they will converge.

Visual analytics needs more expressive data types than provided by SQL or statistics

Since the requirements of visual analytics involve deep changes of the architectural models and implementations of several computer-science domains (e.g., databases, statistics, machine-learning, data analysis, visualisation), there is a strong need for all these domains to be aware of these demands to support exploratory and analytical capabilities in future systems.

6.2 State of the Art

Architectural models exist for all the domains related to visual analytics. We will briefly describe them and highlight the issues encountered when trying to incorporate them in visual analytics applications.

6.2.1 Visualisation Architecture and Data Structures

The domains of scientific visualisation and information visualisation have designed two reference architectural models that are slightly different but are now adopted in all the existing systems. The historic Visualisation Pipeline (Figure 6.2), proposed by Haber & McNabb[52] mainly describes the mapping of data space into visual space whereas the newer Information Visualisation Reference Model (Figure 6.3) as described by Card, Mackinlay and Shneiderman[25], which is a refinement of the Data State Model described by Ed Chi[29], refines the pipeline into a loop where user interaction can happen at all stages of the pipeline. All the well-known implementations of

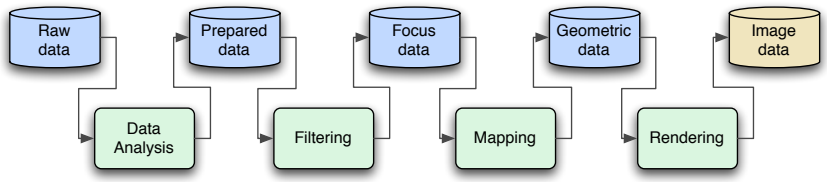

Figure 6.2: The Visualisation Pipeline, adapted from Dos Saltos and Brodlie[37]

information visualisation systems and toolkits adhere to this model and are mostly compatible conceptually, albeit slight implementations variations that give rise to some incompatibility problems, but efforts are ongoing to solve the interoperability issues.

While this model is useful for understanding the transformation from data to views and the processing of interactions back to the data, it fails to describe the analytical process of visual analytics.

Furthermore, the visualisation pipeline emphasises geometric data much more than information visualisation because much of its technical issues come from representing and optimising the geometry for rendering, which is of lesser concern to information visualisation.

Geographical visualisation reference model emphasises multi-scale representations at the data level and layering at the rendering level

Geographical visualisation is similar to scientific visualisation in the sense that geometry plays a very important role and that several methods have been used to model and encode the geography as geometric objects. Furthermore, most geographical visualisation systems are mostly 2D, so the final rendering stage is simple in principle but complex in practice due to the use of layers of information in most GIS systems. One important issue of geographical visualisation is the management of aggregation since maps show different levels of details with different forms depending on the zoom level. This issue of dynamic aggregation and multi-resolution modelling appears also in scientific visualisation but mainly for rendering issues. The problem of aggregation and multiple representations is much newer in information visualisation and has not been modelled in the existing architecture reference model. This is clearly

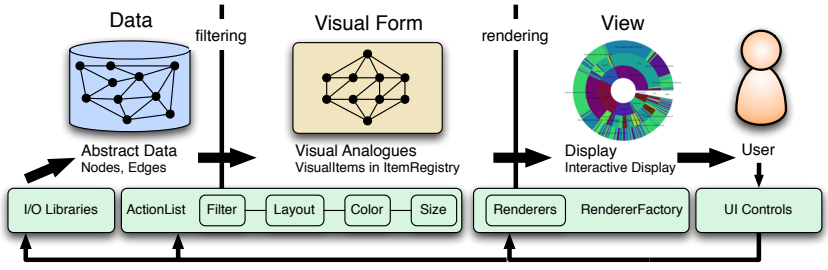

Figure 6.3: The Information Visualisation Reference Model, adapted from Heer et al.[57]

a visual analytics issue that should be better tackled by all the visualisation communities.

Blending different kinds of visualisations is currently difficult

Blending different kinds of visualisations in the same application is becoming more frequent. Scientific visualisation and geographic visualisation need information visualisation because they manage multi-valued data with complex topologies that can be visualised using their canonical geometry. In addition, they can also be explored with more abstract visual representations to avoid geometric artefacts. For example, census data can be visualised as a coloured map but also as a multi-dimensional dataset where the longitude and latitude are two attributes among others. Clustering this data by some similarity measure will then reveal places that can be far away in space but behave similarly in term of other attributes (e.g., level of education, level of income, size of houses etc.), similarity that would not be visible on a map.

On top of these visualisation systems, a user interface allows control of the overall application. User interfaces are well understood but they can be very different in styles. 3D systems use specific types of interfaces that are very different to traditional desktop interfaces. Moreover, information visualisation systems tend to deeply embed the interaction with the visualisation, offering special kinds of controls either directly inside the visualisations (e.g., range sliders on the axes of parallel coordinates) or around it but with special kinds of widgets (e.g., range sliders for performing range-queries). Interoperability can thus be described at several levels. At the data management level, at the architecture model level and at the interface level.

6.2.2 Data Management

All visual analytics applications start with data that can be either statically collected or dynamically produced. Depending on the nature of the data, visual analytics applications have used various ways of managing their storage. In order of sophistication, they are:

- Flat files using ad-hoc formats,
- Structured file formats such as XML,
- Specialised NoSQL systems, including Cloud Storage,
- Standard or extended transactional databases (SQL),
- Workflow or dataflow systems integrating storage, distribution and data processing.

Data Management for visual analytics can rely on different levels of sophistication

We will now consider these data storage methods, paying particular attention to the levels of service required by visual analytics, such as:

- Persistence (they all provide it by definition),
- Typing,
- Distribution,
- Atomic transactions,
- Notification,
- Interactive performance,
- Computation.

Ad-hoc flat files

Flat files, including XML, will only remain a commodity for interchange and high-performance acquisition of data

In the last 20 years, the most popular system for data analysis has been the spreadsheet calculator. Spreadsheets are ubiquitous and one of their strength is their simplicity and versatility, which comes partly from their lack of enforced strong typing and structuring. Most business and scientific data is available as spreadsheet files that are quite difficult to analyse automatically, due to this lack of typing and structuring. Therefore, practically all data analysis and visualisation systems provide extensive import/export support for spreadsheet files.

Variants of spreadsheet format files, such as the simple Comma Separated Values (CSV) files, are supported by almost all data-exchanging programs nowadays. The main pitfall of these spreadsheet formats is its lack of typing and metadata. These files require human interpretation before they can be used by an application.

Besides these well-known file formats, most data-oriented applications have used ad-hoc formats to save and load their internal state. The XML format has been designed to solve that problem and to offer an all-purpose file format for all the applications that require data storage and interchange. There are still three reasons not to adhere to standards: legacy programs that continue to use their ad-hoc formats, secrecy to hide details of internal structures of a system, and performance. XML adds sophisticated typing to flat files, which is very important, but no other services.

Highly demanding systems use ad-hoc architectures to collect data and analyse them quickly. Formats like the Hierarchical Data Format (HDF[1]), designed with performance in mind, are required in special situations, such as managing data returned from high-resolution high-throughput sensors in physics experiments, producing megabytes of data per second. However, this data is usually collected in short bursts and can be processed afterward using more standard formats. This step of data cleaning and filtering is an integral part of visual analytics and therefore, some visual analytics applications should be able to process and understand these formats, as well as the more standard ones. High-performance storage systems offer the same level of service as flat-files.

Traditional Databases (Row-based)

Extensions to traditional databases needed for typing, in-memory caching and fast notifications

Transactional databases have a long tradition of success and reliability. SQL is the standard and several products are currently available that implement different levels of SQL functionality for various prices, from free to thousands of Euros or more.

SQL technology is mature and implementations are usually robust – based on tables stored in row order. SQL provides atomic transactions (the well-known ACID properties). They provide most of the services required by visual

[1] `http://www.hdfgroup.org/`

analytics applications, except that the typing is not as expressive as needed. SQL types are related to their storage and to some extent to the operations that can be performed on them, but important properties of data cannot be expressed in a portable way using SQL alone. For example, standard SQL use integers for values and for categorical data (e.g., zip codes). It is essential in visual analytics (and statistics) to know precisely, the semantics of attributes in order to apply meaningful computations and visualisation techniques to them.

Since transactional databases implement all the data management services required for visual analytics, it would make sense for visual analytics systems to rely directly on them. However, they have several pitfalls:

- Interactively visualising data requires data to be in memory. With the exception of in-memory databases, standard transactional databases do not guarantee the sustained performance required by visualisation and analytical computations. Therefore, visual analytics components have to implement an in-memory version of the databases.
- The data types provided by SQL are mainly storage oriented, not semantic oriented. A value representing a latitude or longitude will be typed as Real. Visual analytics applications need to add more metadata and there is no widely adopted standard to do that.
- Notification is implemented through *triggers* in standard transactional databases. The trigger mechanism is very inefficient in most database implementation; some databases provide workarounds but they are not standard. Without an efficient notification mechanism implemented from the database layer, the visual analytics application needs to implement one on its own.

Analytical Databases (Column-based)

To address efficiency issues, both in terms of speed and memory, new databases architectures are column-based. For example, Kdb+ can handle streaming data and analysis on the fly; it has been experimented with in visual analytics by Chan et al. at Stanford. MonetDB[18] is a more general-purpose transactional database engine developed at CWI in Amsterdam that is also column-based. It implements most of the services required by visual analytics but has never been used as the default storage engine for visual analytics application so it remains to be seen if MonetDB delivers what it promises.

Specialised NoSQL Systems

Trendy NoSQL systems are spreading but their heterogeneity and short life-span are problematic

NoSQL systems are usually built to avoid the complexity of general transactional databases and provide faster, simpler or more specialised services. Google internally uses a very optimised file system called BigTable. Amazon internally uses a proprietary key-value structured storage system called Dynamo for its Web services. Several very different services are provided by NoSQL system, from document stores (e.g., CouchDB) to graphs (e.g., Neo4j), key-value store (e.g., BigTable) and hybrids.

NoSQL systems also include storage solutions on the Web or in 'Cloud Storage'. There is a trend in migrating resources on the Web through more than one provider. For example, several large online service providers (e.g., Amazon Simple Storage Service, Google Storage) provide Cloud Storage to allow outsourced storage and computations from Web services. Along the same line, new repositories on the Web offer high-level Web services to query their contents (e.g., Google and its visualisation API Data Source). However, ad-hoc storage management solutions do not provide any time performance guarantees for access or modification, so visual analytics applications need to build layers, such as caching, on top to deliver acceptable response.

Workflow and Dataflow Systems

According to the Workflow Management Coalition[2]:

> Workflow is concerned with the automation of procedures where documents, information or tasks are passed between participants according to a defined set of rules to achieve, or contribute to, an overall business goal. Whilst workflow may be manually organised, in practice most workflow is normally organised within the context of an IT system to provide computerised support for the procedural automation and it is to this area that the work of the Coalition is directed.

Scientific workflows have a great potential to become the backbone of visual analytic applications

In the recent years, several workflow systems have been designed to automate scientific processes; they are called 'scientific workflows' and since 2007 have their own workshop (IEEE International Workshop on Scientific Workflows)[3]. Although workflows are designed to apply a well-known process repeatedly, exploratory workflow systems are starting to appear, such as VisTrails [4]. VisTrails is system managing provenance and parameter setting for visualisation systems. A pipeline of processes is built and run interactively. Its results, in the form of visualisations, can be displayed in a table format, which allows multi-dimensional exploration by changing parameter values. The changes are recorded in a database, so later on, the user can explore their own construction of the pipeline or send it to another user for their interpretation. VisTrails is a very compelling system for exploration and visualisation of large-scale data. However, VisTrails has some weaknesses for visual analytics:

- It relies deeply on the Visualisation Toolkit (VTK): the visualisation pipeline is built directly as a VTK pipeline and parallel computation and rendering relies on the ParaView extension of VTK. Therefore, it relies heavily on a specific technology.
- It does not use a standard database for storing its state and data. It uses XML files in a directory to keep track of its history. VTK is neutral in term of data sources and can read from a large number of file formats and databases.

[2] http://www.wfmc.org/
[3] http://www.extreme.indiana.edu/swf-survey/
[4] http://www.vistrails.org

- It does not manage dynamic data: changing data sources does not trigger any re-computation and also each user initiated re-computation must start from scratch. VisTrails maintain a cache of computed results but the cache mechanism is not aware of dynamic data.
- It does not implement any protocol to manage the interaction among workflow/dataflow components. Only standard interactions are available.

Despite these weaknesses, VisTrails is a sophisticated scientific workflow system that allows exploration and provenance management. It should certainly be an inspiration for the future of visual analytics software infrastructures.

VisTrails should be an inspiration for future visual analytics software architectures

Data Management Conclusion

Ideally, the native storage management layer of a visual analytics application should provide all the services described in this section. Unfortunately, no existing storage management system currently offers all the required set of services. The visualisation community has started to design its own set of data management facilities that will not scale whereas the data management community is not yet aware of the new requirement for interaction and visualisation.

6.2.3 Data Analysis

Analytical systems usually implement a very simple architectural model: they read inputs and write outputs until their work is done. Several environments are available for analysis depending on the data types:

Analytical systems usually implement a very simple architectural model

- Statistical analysis (e.g., SPSS, SAS, R)
- Scientific computation (e.g., Matlab, Scilab)
- Machine learning toolkits (e.g., WEKA)
- Textual analysis (e.g., GATE, UIMA, SPSS/Text, SAS Text Miner)
- Video analysis (e.g., OpenCV)
- Image analysis (e.g., Khoros, IRIS Explorer)

Data analysis takes data as input (from a data management layer) and processes it to produce different kinds of interpretation. Analysis environments take several forms:

- Program libraries,
- Components,
- Toolkits,
- Simple applications,
- Integrated applications,
- Web-services.

Most of the analysis systems can be run from a database or flat files. They tend to be neutral in the form of the data they input, except for integrated applications

that internally manage some form of database. All of the analysis systems output their results in flat files, XML files or databases.

Data analysis and data management solutions are usually well integrated using open standards

Data analysis and data management solutions are usually well integrated and applications, that are performing analysis without user exploration, can be developed with a wealth of software solutions that can be combined, relatively easily in powerful ways; this combination is purely data driven: the program will run to completion to return a solution.

However, from a visual analytics standpoint, their main weakness lies in their 'architectural reference model'. Analysis systems read from data files, apply a computation and output to other data files. This is fine as long as interactivity and exploration are not required. For visual analytics, interactivity and exploration are essential. In the case of a large dataset and complex analysis, the analyst does not want to wait minutes or hours if they are not sure that the analysis is useful. In the past, systems such as Hive[92] have been designed for trading quality and speed, but they were only prototypes. To provide the right level of service, analytical components should change their reference model to be able to present an overview first and then allow for progressive refinement under the control of the analyst. More research is needed to better understand how to address these needs.

Analytical components should change their reference model to suit the needs of visual analytics

In the recent years, there have been several attempts at providing machine learning and data analysis as external services. For example, Microsoft has defined a set of data analysis protocols: XML for Analysis, DMX (Data Mining Models) and the Data Mining Group [5] has designed PMML (Predictive Model Markup Language) as a way to communicate and run data mining algorithms in a vendor neutral fashion.

These services are now available from several robust analytical platforms, either free such as the R statistical system or commercial such as SPSS, or toolkits such as Weka (a popular data mining toolkit in Java[6]). Web-based implementations are also available from well-known providers such as Google with its Google Prediction API[7]. From a visual analytics standpoint, these system offer a very rich set of capabilities but with crucial features missing:

- **Fast initial response with progressive refinement** Some analytical algorithms are incremental by nature. For example, computing eigenvectors with the largest eigen values (e.g., for Principal Component Analysis) uses a power-iteration that is incremental in nature; best heuristics for computing the 'travelling salesman algorithm' start from an initial tour and try to improve it incrementally. These algorithms and many others are routinely available in several analysis systems but they never provide any intermediate results due to the absence of a software protocol to send results on demand. A standardised protocol would be one way of overcoming the problem.

[5] http://www.dmg.org/
[6] http://www.cs.waikato.ac.nz/ml/weka/
[7] http://code.google.com/apis/predict/

- **Re-computation following small changes** Computing a hierarchical clustering is usually done in two steps: computing a distance matrix and iteratively extracting the two closest items until all items are extracted. The first operation has a quadratic complexity in the number of items to cluster, which can be very large. Once a distance matrix is computed, changing one item only means re-computing one line of the distance matrix, which is very fast. However, most clustering algorithms will not be able to store a distance matrix once the clustering has been computed in order to do this. Therefore, the penalty of changing a value (adding or removing) is quadratic instead of linear. Therefore, visual analytics systems need specialised dynamic hierarchical clustering functions (e.g., to remove outliers) rather than rely on standard library routines.
- **Steering the computation** Some algorithms are inherently slow, such as MDS (Multi-Dimensional Scaling). However, they can be steered to deliver faster results on a specified region of interest. This steering is not inherently complex to implement, but visual analytics practitioners need to do it themselves.

All these capabilities, and maybe some more, are required by visual analytics applications and are not provided by analysis systems. More research work is needed to find the right level of services and combination of algorithms to fit the needs of visual analytics.

Conclusion for Data Analysis

Services essential for visual analytics are missing from standard analytical architectures

In recent years, the different analysis communities have made a strong effort to facilitate the use of their systems and algorithms. Most advanced analytical systems can be connected to other applications through several mechanisms of communication: direct library linking, component integration, inter-process communication or Web services. However, there is a deep mismatch between the level of services they provide and the needs for visual analytics: they do not provide mechanisms for:

- Fast imprecise answers with progressive refinement;
- Incremental re-computation, either in the data (e.g., some data has been changed) or in the analysis parameters;
- Steering the computation towards data regions that are of higher interest to the user.

These newer requirements can be difficult or impossible to implement on top of the existing services. More research work is needed to understand how they can be supported in the future.

6.2.4 Dissemination and Communication

Currently, results of complex analysis are presented to decision makers using slide-shows such as Microsoft PowerPoint. A slide-show is a support for story-telling: analysts need to collect evidence of their findings to report them in

Few systems to manage dissemination and publication of visual analytics results

a meaningful order using text, still or animated images captured from their exploration and sources.

In the recent years, the GapMinder system designed by Hans Rosling [8] showed how effective visualisation and animation can be for telling compelling stories about data. However, existing visual analytics systems provide no mechanism to move from the analytical process to the presentation process, except for producing still images. Even these still images are not completely adequate for paper publication because the graphical characteristics of the printed medium are different from the screen.

Several systems have been designed to gather analysis information during an analytical process. The accumulated information can be revisited and kept for later use or archived. Oculus nSpace is one of them, designed for visual analytics applications. Although designed to help the exploration process, nSpace is also helpful to create presentations at the end of the analytical process. However, again, the created presentation is not interactive or linked to the actual exploration. Systems such as VisTrails (see Section 6.2.2) offer all the capabilities for linking back images to the exploration process. Explorations done with VisTrails can be distributed and replayed easily. However, they are currently not meant to be used for slide-shows style presentations.

To summarise, systems to manage dissemination and publication of visual analytics results are definitely lacking; this offers interesting opportunities for research and commercial products.

6.2.5 Cross-cutting Issues

Each domain has been exploring cross-cutting issues separately, they should now coordinate

Software infrastructure has been described above in the order of the pipeline process. However, issues that are common to all levels are now discussed.

Distribution

Distribution is an important aspect of visual analytics. The data management can be distributed, the analysis can be distributed and the rendering can be distributed. Therefore, several questions arise: is there one mechanism for distribution (for example, the database engine should be responsible for the distribution) or should there be one mechanism for each tool, or a general mechanism (for example multicast communication) so that all the tools can communicate using a common bus?

For now, each tool implements its distribution mechanism and visual analytics applications need to cope with all of them. Accessing a distributed resource, whether for storage, computation or rendering, is not particularly complicated and making use of several mechanisms is not an important issue, unless rapid interaction and coordination is involved. In that case, notification mechanisms should be used, which is complicated when several resources are involved

[8] http://www.gapminder.org/

because the mechanisms offered can be quite different. Standard SQL databases offer only triggers that are usually inefficient and no standard mechanism is provided to propagate notifications across a network. Analytical modules often do not provide any notification mechanism, with the exception of image processing systems, which usually do.

Even if one distribution mechanism could be used for all the parts of a visual analytics application, it might be less effective than several mechanisms well designed for each task. For example, the parallel visualisation rendering system ParaView[9] uses distributed memory, whereas most SQL databases use Internet network connections. There is no way to change either of these implementations for technical and historical reasons.

Finally, with the advent of computation in the Cloud, processing will also migrate to the Internet or to large computation grids. These technologies require special skills that are currently in short supply.

New Computing Paradigms

Beyond distribution, new programming paradigms are emerging. Cloud computing has already been mentioned, with its grid-computing variant, but GPU programming is also becoming more widespread and can be used for very demanding applications. Visualisation has been using GPU programming from early on, but the data analysis community is only starting to utilise this powerful computation resource.

All these new paradigms will evolve quickly in the forthcoming years and it is necessary for the visual analytics software infrastructure to keep pace with these developments and be compatible with them.

Language

Since visual analytics relies on several complex components to carry out potentially long computations, the programming language and interoperability between languages is very important. Currently, the choice of programming language used in a visual analytics project seriously restricts the choice of tools available. The information visualisation community has several toolkits programmed in Java. The scientific visualisation community generally uses C++. New environments such as Microsoft .NET allow programs written in different programming languages to interoperate but the Java language is not so well supported. New languages are now in use such as Microsoft F# for advanced functional programming, Scala for scalable computation and SVG-based JavaScript for Web application. New ones will eventually appear. How can visual analytics avoid constraining the software infrastructure landscape by programming languages? Two choices are possible:

Only research can teach us what combination of languages and mechanisms are best suited to develop and deploy visual analytics applications

[9] http://www.paraview.org/

- Rely on a virtual machine such as Microsoft CLR or the Java virtual machine, but there are still complex issues in term of code libraries that are not solved by a virtual machine.
- Use Web-based mechanisms such as Web services. However, whilst the current mechanisms can provide relatively high throughput it is usually at the expense of high latency, and therefore not suitable for interactive applications.

Only research can teach us what combination of languages and mechanisms are best suited to develop and deploy visual analytics applications beyond the current level of craftsmanship.

6.3 Challenges

Designing an accepted conceptual architectural model is difficult because it involves several well established domains

Designing an accepted conceptual architectural model for visual analytics is a difficult issue because it involves several domains that are already well established and hence will need a collaborative effort to understand cross-domain issues. Several workshops have started to tackle the problem but it can still take several years before reaching a consensus. More effort should be devoted to experiments in this domain so as to quickly agree on a recognised architectural model that all components comply with.

So far, visual analytics systems have been implemented by extending existing environments. Database practitioners have extended their database environment, machine-learning and data analysis practitioners have extended their analysis environments, and visualisation practitioners have extended their visualisation environments. The results are not satisfactory. This has led to work being done by non-experts in the fields, often leading to sub-optimal solutions; too many resources have been wasted to 'reinvent the wheel' and the solutions do not scale or do not provide good quality interaction.

A unified architectural model does not mean one unified implementation

A unified architectural model will involve fairly new programming paradigms such as asynchronous computing and the management of multi-scale data structures. It is important to emphasise that a unified architectural model does not mean one unified implementation. Several domains have found it necessary to deal with this issue in the past and have found several solutions without relying on one particular implementation. However, in contrast to previous standardisation work, visual analytics will involve much more diverse domains and some clear methodology should be devised to reach convergence and agreements among this diversity.

Once this conceptual phase is achieved, it will lead to a clear specification of software components and to the potential creation of a market for components. Practitioners of visual analytics applications will be able to reuse components implemented by others, whether commercial or free, whether for profit or for research. Designing analytical components that scale and provide capabilities for interaction is a difficult challenge. It will require new analysis methods, in addition to the adaptation of existing methods that have not been implemented with interaction in mind.

Moreover, the requirements of visual analytics will foster new interesting research in the domain of high-performance databases, analytical components that can provide results at any time and be steered interactively, and new visualisations that could scale to arbitrarily large sized datasets.

6.3.1 Visualisation

Visualisation architectures should merge; more research is needed to solve incompatibilities

Existing visualisation software infrastructures are quite different in capabilities. Scientific visualisation can manage terabytes of geometric data in real-time with special-purpose computers, as information visualisation can only deal with millions of data points. Geographical visualisation can only display a limited amount of information, usually less than a million items, but by using very sophisticated aggregation methods that can manage terabytes of actual data, users are able to navigate freely. The important challenges are thus:

- Allow all the visualisation domains to share a common rendering pipeline, where graphic acceleration can be used simply, multi-thread rendering is supported natively, and overlaying and other merging techniques can be used to blend images generated from all the visualisation domains (scientific, information-based or geographical).
- Improve research on data structures and algorithms for aggregation to try to unify the different facets currently used to aggregate visualisations. Historically, geometric aggregation is very different from data aggregation and geographic aggregation. Unifying them would facilitate the software integration of components from the different domains.
- Allow deeper integration of all the visualisation domains. Most existing systems use side-by-side views, barely coordinated. Adapting existing coordination mechanisms to work with all the visualisation domains would facilitate linked and coordinated views.
- Improve research on software architectures for collaborative visualisations to allow the software infrastructures to be usable in single-user and multi-user settings.

6.3.2 Data Management

Data management model should provide distribution, in-memory caching, notification management and expressive typing

Since all the components of visual analytics require data to be stored and distributed to other software components, the data management component seems to be a good candidate to be the central mechanism for data and, to some extent, for distribution.

Information visualisation systems rely on an in-memory database to maintain their data. Relying on a solid database system would allow the domain of visual analytics to grow to larger sizes and lead to more robust applications.

Looking at the services described in Section 6.2.2, we can list the most important features that a successful data management architecture should provide:

Data Typing

The standard typing provided by SQL is not sufficient; higher-level types should be supported, in particular those listed by Card and Mackinlay[24]. At the infrastructure level, these types can be seen as metadata: there is a need to support rich metadata to adapt to rich information associated with the data. More sophisticated types should also be supported at the storage level. For example, there are several ways to aggregate numerical values – currently, most databases support single-valued summarisation, such as average or median, but more sophisticated summarisation include min-max or distribution histograms. Supporting these types, among others, is essential for analysis and visualisation. Special types have already been specified for geographical databases, it is important to allow these extensions at the database infrastructure level.

Managing and indexing dynamic, streamed data requires new mechanisms

Managing dynamic data, including streamed data, is also very important and not standard in databases. Time-stamped and volatile data is becoming increasingly important to manage. One of the difficult issues associated with this kind of data is in indexing and summarisation. Depending on the application domain, streaming data can be summarised with simple statistical values or more complex types such as wavelet transforms. Current databases do not support these types of analysis on the fly.

Distribution

Distribution is needed at several points of visual analytics systems; unify it when possible

Most databases are distributed using simple network connections. However, the performance of streamed-network links is low compared to the processing power of existing hardware architectures. Newer database systems offer datagram distribution for fast replication. Allowing more flexible and faster distribution protocols will allow the overall visual analytics infrastructure to grow to larger sizes and higher processing power. A fast distributed database can become the central point to manage distributed processing using newer parallel architectures, such as computer grids and multi-core GPUs, and distributed rendering systems, such as wall-sized displays, large tabletops or collaborative environments.

Distribution should also involve caching mechanisms so that the same software infrastructure can be used to manage massive databases and in-memory databases in a consistent way. Current visual analytics applications manage the transfer of relevant data in ad-hoc ways with little cooperation between the central database and the in-memory one, and no compatibility at the programming level.

Atomic Transactions

Visual analytics requires long transactions that are not supported by standard databases. Since analytical components may run for hours, days and weeks, the data manager needs to support very long commit phases, probably with some reconciliation strategy to deal with errors instead of promoting a complete fail. If analytical components can save partial results, they can finish transactions at a faster pace but it can take minutes or hours before a meaningful cycle of operation is ready to be committed. Traditional databases do not support these long transactions, although some drafts have been submitted for standardisation by major vendors using 'snapshot' isolation. More research work should be devoted to specifying a semantic of long transactions compatible with analysis, and to designing mechanisms for interactive refresh of visualised structures.

Notification

Notification in databases is currently implemented through the trigger mechanism, which executes some code when the data is modified. The support for triggers is very heterogeneous from one database to the other. While Oracle supports general triggers and extend them to notify on structural changes (schema modification), others such as MySQL lack much of this functionality.

These weaknesses hamper the use of standard databases for visual analytics and force practitioners to implement some kind of in-memory database that are certainly not as powerful and reliable as the mature database engines, but they fulfil the important requirements of visual analytics.

Revisiting database mechanisms such as notification will improve visual analytics

Newer database systems such as MonetDB offer a low-level layer of implementation where new kinds of notification mechanisms can be implemented and experimented with. The view of MonetDB as a 'memory shared across the network' instead of a facility to store and query data appears to be suited to visual analytics.

Interactive Performance

Visualisation systems and analytical systems need optimised in-memory data structures. They also implement the standard visualisation pipeline 'backwards', meaning that it is the view that triggers the computation of visible parts, pulling computations from the pipeline instead of just displaying the already computed contents of the previous stages. This is very different from what current analytical systems and databases provide.

Currently, database systems are not designed to allow fast in-memory computation or rendering. Visual analytics require high performance and so finding mechanisms to unify fast memory management with persistence, a fast query mechanism and distribution, would allow visual analytics to work

on a solid base. If this is not possible, then more work is required on a good separation of issues between database technologies and analysis and visualisation technologies, to avoid duplicating design and implementation efforts.

Computation

Current workflow systems connected to databases work by computing 'forward', starting from the beginning of the dependencies to the end. As mentioned above, visualisation systems usually work backward by pulling the required data from the pipeline, computing it on demand, steered by the analyst. Can workflow systems be improved to support this pull mechanism, allow some steering and to provide on-demand approximate solutions quickly to improve them later when possible?

Finding mechanisms and policies to allow large-scale asynchronous pull computation needs more research and experiments before it can be specified and standardised.

6.3.3 Analysis and Data Mining

The analysis and data mining domains use a simple, yet effective software architecture. Several implementations now work on local machines between different products, through the Web or on the Cloud. However, this architecture is not suited to visual analytics applications as it is. The main issues are:

- **Progressive analysis:** provide quick answers first, then make improvements incrementally or on-demand;
- **Management of dynamic data:** incremental analysis instead of restarting it from the beginning;
- **Steerable analysis:** allow long-computations to be steered by users when possible.

Currently, no data analysis tools provide these services. There are two paths that can be pursued to solve this gap: a) combine existing services to try to obtain the desired results or b) re-implement existing systems to provide the services.

As a first step, specifying a consensual software model for an analytical component will be required. All the analytical communities should be gathered to find agreement and social acceptance since this new software model will certainly require considerable work to be fully implemented and functional.

Conclusion

To better understand the interdisciplinary software architectural issues of visual analytics, all the specialists of toolkits and tools from the domains involved should meet and publish a white paper on recognised issues and ways to solve them. The VisMaster project has made a start by organising two workshops but the scope of the problem is so broad that it will need several more workshops, involving more focused domains, in order to move towards a good understanding of interdisciplinary issues and ways to implement them in a modular and extensible fashion.

The diversity of problems advocates for open standard rather than proprietary solutions

The diversity of problems addressed by the visual analytics community advocates for open standard rather than proprietary solutions. While some proprietary solutions are already available, most visual analytics applications will certainly need several analytical modules from several vendors and trying to monopolise the market with proprietary interfaces will instead slow-down the growth of the visual analytics field and delay the creation of a market for rich analytical components that can be integrated in interactive applications.

6.4 Opportunities

Solving the architecture issues will open a new market for components

Despite the large number of challenges facing the design and development of software infrastructures for visual analytics, the opportunities are considerable and viable. They are both scientific and commercial. Scientifically, the increased production of data should be harnessed but this requires new methods. However, even if the principles of exploring and analysing data become better understood every day, benefiting from visual analytics will require well designed software infrastructures.

Once these infrastructures are available, scientists and practitioners will devote fewer resources to specific software developments, instead they will rely on sound infrastructures to build their visual analytics applications.

Commercially, the market of visual analytics components does not exist because the requirements for these components are not well understood. However, the demand for such components is already high. Once the requirements and specifications of these components (or abstract components) are known, several companies, with varying sizes, will be able to provide their expertise and added value to practitioners in visual analytics and more generally to improve the usability of data-intensive applications.

When analytical components are usually small, there is also a need for larger systems such as databases and visualisation systems. New databases will be able to cope with massive data interactively. Addressing data management issues will certainly lead to new database engines, faster, more scalable, more distributed and providing analytical and interactive capabilities. New visualisation systems will be able to cope with very large datasets and to allow newer kinds of

visualisations, combining or merging scientific visualisation, geographic visualisation and information visualisation when needed.

6.5 Next Steps

The topic of software architecture for visual analytics is broad. To come to a common understanding of the problems and reach a consensus, we need contact points: workshops and conferences gathering all practitioners, research and industry. In the area of software infrastructures and standardisation, industry is often ahead of researchers.

Research agencies should create incentives for researchers in databases, analysis, visualisation and communication to work together to iterate on the design of a conceptual architecture for visual analytics. If this problem is not considered as a whole for all domains of visual analytics, as opposed to specific in each and every domain, we will see an explosion of partial solutions to the overall infrastructure problem with issues in interoperability. Therefore, there should be a coordinated action aimed at solving the problem overall.

In practical terms, this could be done by funding a few applied-research projects to design and experiment software architectures for visual analytics. The outcomes of these projects should be partly open since the goal is to promote interoperability and compatibility. Since software architecture is an interdisciplinary issue, it cannot be effectively addressed by several smaller funded projects, gathering some experts in software architecture of visual analytics. Initially, the problem should be addressed as widely as possible. Later on, more targeted groups will certainly address more focused issues but they need to be aware of the big picture first.

Once some level of agreement is reached, some organisation should be created to promote and manage the specifications. It can be done in the context of an existing organisation such as the Object Management Group (OMG) or independently as a visual analytics alliance. In any case, it should gather interested parties from industry and research to help specify 'standard' software architectures and APIs for a visual analytics module to interoperate with the right level of functionality and provide clear semantics.

Such a coordination will not only be beneficial to visual analytics but also to the related domains. Data management will provide the right services to scale, data analysis will be easier to integrate in more interactive environments and visualisation will be easier to deploy.

7 Perception and Cognitive Aspects

7.1 Motivation

The human is at the heart of visual analytics human interaction, analysis, intuition, problem solving and visual perception. This chapter is entitled "perception and cognition", and it is possible to have a narrow focus of this looking purely at the perceptual and cognitive aspects during the time when a user interacts directly with a visualisation or adjusts parameters in a model. However, there are many human-related aspects of visual analytics beyond those involved in the direct interactions between a user and a visual representation of data. Figure 7.1 presents a simplified view of the broad visual analytics process that emphasises some of the wider context and the human issues involved.

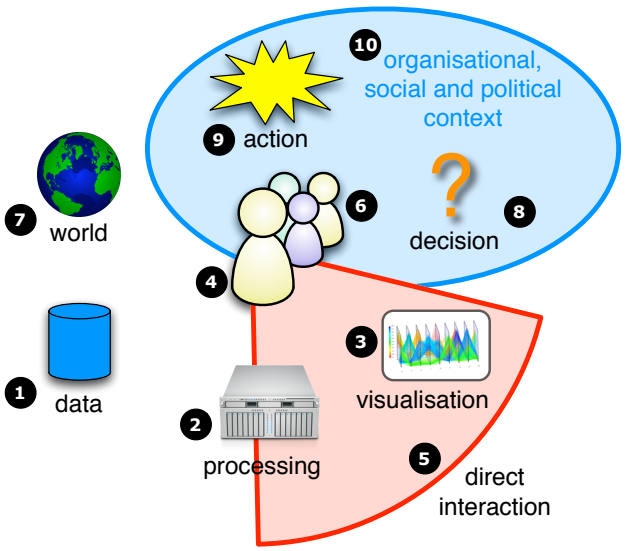

Figure 7.1: The human context of visual analytics

Working through the numbered parts of Figure 7.1, visual analytics involves some data (1), typically being processed (2) computationally (e.g., machine learning, statistics), then visualised (3) and interpreted by the user (4) in order to perform problem solving, analysis etc.. The pie-shaped region (5) represents the obvious direct interactions between the primary user, processing and visualisation. When multiple people are involved in this process, it can also be collaborative (6).

However, the role of people goes beyond direct interaction with visual analytics systems. The data being visualised comes from the world (7) (or some simulation of it) and is typically used by people, who may not be those involved in interacting with the visual analytics system, to make decisions (8) that influence actions (9) that ultimately affect the world.

This gives rise to a far broader organisational, social and political context (10): the stakeholders who use the outputs of visual analytics and those impacted by the decisions cannot be ignored by those using the systems and indeed, those involved in the technical process may be subject to social and political pressures as well as considering how the results of the visual analytics process can best be presented to others.

7.2 State of the Art

There is a substantial literature on specific techniques and systems for interactive visualisation in general, although fewer looking at human interaction when there is more complex non-visualisation processing as in visual analytics (with exceptions such as clustering or dimensional reduction). Looking beyond experience reports or simple user studies to detailed perceptual and cognitive knowledge the picture becomes more patchy. There is work on static visualisation (e.g., abilities to compare sizes), yet there is little on even simple interactive or dynamic visualisation let alone where this is combined with more complex processing. Again, whilst there is a longstanding literature of technical aspects of collaborative visualisation, social and organisational aspects are less well studied. For example, recent work on sales forecasting found that, perhaps unsurprisingly, issues of organisational context and politics were as important as statistical accuracy. Methodology is also important, even in more traditional visualisation areas issues, such as evaluation, are known to be problematic.

7.2.1 Psychology of Perception and Cognition

Psychological research on perception of visual information is based on the seminal work of Allan Paivio who asserted that the human perceptual system consists of two subsystems, one being responsible for verbal material and the other for all other events and objects (especially visual information). He emphasised the importance of mental images for human cognition. Even if some of his assumptions have been criticised, his considerations still provide an important reference point for psychologists investigating visual perception.

Distinction between high and low-level vision

Researchers in perceptual psychology usually distinguish between high and low-level vision. Activities related to low-level vision are usually associated with the extraction of certain physical properties of the visible environment, such as depth, three-dimensional shape, object boundaries or surface material properties. High-level vision comprises activities like object recognition and

classification. Results from low-level vision research are finding their way now in visualisation and visual analytics[122], but results from high-level vision research are not yet adopted.

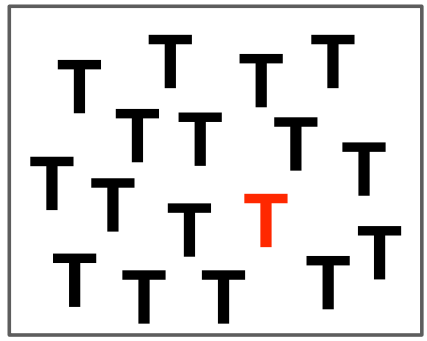

Figure 7.2: Preattentive processing – pop-out effect

Ware[122] discusses preattentive processing quite extensively. This theory tries to explain the fact that some elements of visual displays pop out immediately and can be processed almost automatically (see Figure 7.2). These elements can be distinguished easily from others, for example by their form, colour or orientation. Such processes are very important for visual perception because they support visual search considerably. Despite some criticism, this theory has been very influential for information visualisation and visual analytics because the quality of systems representing information on a computer screen depends to a considerable extent on whether they support search processes efficiently.

Preattentive processing makes items pop out the display automatically

The human visual system has by far the highest bandwidth (the amount of data in a given time interval) of any of our senses and there is considerable research into how we make use of this data about our immediate environment. Visual representations are generally very short lived (about 100msec) and consequently much of what we 'see' is discarded before it reaches consciousness. Evolution has given humans the ability to rapidly comprehend visual scenes, as well as text and symbols and much of this rapid, unconscious processing involves representations in our conceptual short-term memory[90] where small snippets of information (such as individual words) are consolidated into more meaningful structures. However, addition processing stages are required before we become aware of a particular stimulus and it survives in longer-term memory. Demands on this higher-level processing from rapidly presented sequences of visual stimuli can give rise to failures in retaining visual information, such as attentional blink and repetition blindness[33], and as such are important to designers of visual analytic systems.

Another theory of visual perception, which has some relevance for visual analytics, is Gestalt psychology. This assumes that visual perception is a holistic process and that human beings have a tendency to perceive simple geometric forms as illustrated by the examples in Figure 7.3. This implies that the structure underlying a visual display is more important than the elements

Humans tend to perceive simple geometric forms

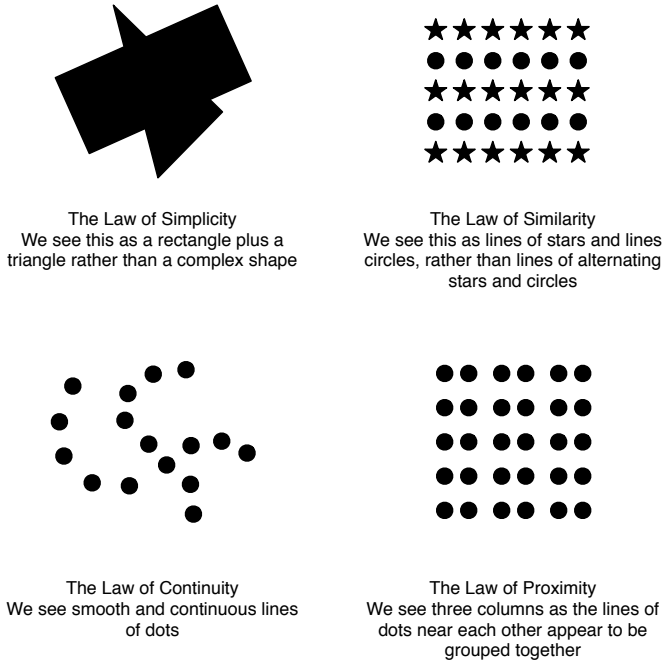

Figure 7.3: The Gestalt Laws imply that we tend to see simple, often connected structures within a scene. (Only a subset of the Laws is shown)

and is often summarised as 'The whole is more than the sum of its parts'. These principles can be used for guiding attention efficiently in visual displays in order to help reason through the data, although we need to be aware that strong visual characteristics, such as bright colours or joining lines, can dominate or influence one's reasoning processes.

Visual perception is an exploratory process

Recent research in the psychology of perception indicates that perception is an exploratory and active process. Gibson[48] pointed out that human perception is tied to the movement of the human body in a natural environment. We do not see a sequence of more or less static images but a continuous flow of changing scenes in this natural environment whilst we move around.

Neisser[81] developed a model of perception based on a cycle consisting of schemata, available information about objects and perceptual exploration (see Figure 7.4). The process described in this model is always influenced by past experience (schemata, expectations). Based on this experience, hypotheses are formulated which guide perceptual exploration. Our cognitive resources, especially our short term memory, are limited; therefore, we direct our attention only to objects we consider in advance to be interesting. If the information from the environment does not match these hypotheses, schemata in human memory are modified. This is an ongoing and iterative process.

In this context, the movement of the eyes, especially the so-called saccadic

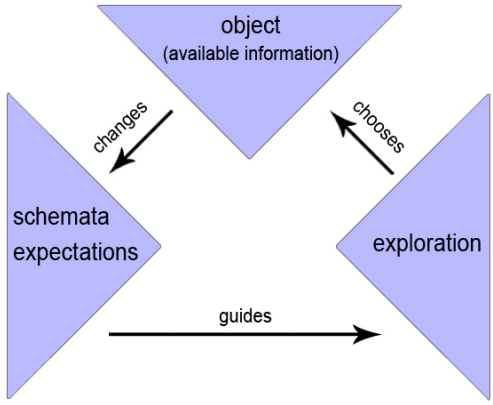

Figure 7.4: Model of Perception[81]

movements, plays an important role. The eyes do not move continuously, but in series of jumps (about four per second). Between these jumps, fixation occurs when people gaze at objects in the environment. Eye movements are especially important as peripheral human vision is rather inefficient. To resolve detail, an image has to be projected onto the fovea - a fairly small region on the retina, which is responsible for sharp central vision. Everything else in the visual field is quite blurred (see Figure 7.5). It is, therefore, not possible to get a comprehensive impression of the environment at one glance. In this context, eye movements play an important role. They enable human beings to see the necessary details in a series of several fixations. We have to look for information actively to get a fairly comprehensive image of the environment, in a process quite similar to the one described by Neisser (see above).

Eyes move in a series of jumps

Human peripheral vision is poor

Figure 7.5: Acuity is only high in the centre of the visual field

These and similar approaches in the psychology of perception are especially suited for modelling the interaction of users with visualisations. The usage of such tools is often described as an active and exploratory process yielding

complex insights[7]. In the course of this process, hypotheses are generated and tested on the basis of the data visualised by the tool. There is research that applies results from cognitive psychology to the design of visualisations, which adopts such an approach[102].

Change blindness means we often fail to see seemingly obvious changes

Recently, the phenomenon of change blindness has attracted much attention[102]. Change blindness describes the phenomenon that observers often fail to notice important changes in their environment, especially if they do not pay attention to these changes. Rensink also argues that humans do not possess a detailed, picture-like representation of the scenes they see. Nevertheless, observers gain the subjective impression that they have a stable representation of their environment. This is due to the fact that observers can get any information they need whenever they want it just by focusing their attention on the relevant object. It might be argued that observers use the environment as some kind of external memory to relieve their own limited capacities (especially short term memory and processing capacities). This approach also assumes that perception is active, not passive. Ware describes this as visual queries - the search for patterns in the outside world. This capacity of human information processing is very flexible and adaptive. Both Rensink and Ware argue that visual representations, especially visualisations on a computer screen, should be designed in a way to support these processes, and they both suggest a set of design guidelines for this.

Flexible and adaptive vision system searches for patterns

7.2.2 Distributed Cognition

Distributed cognition is a theoretical framework describing the interaction between (groups of) persons and artefacts[58, 61]. It builds on the information-processing concept of a problem space, but extends the boundaries of the problem space to incorporate knowledge in the mind of the user and knowledge in the world. It proposes that our everyday problem solving involves the coordinated use of knowledge structures in the mind, in our environment and from other individuals. The object of investigation is, therefore, not the single individual, but a system of cooperating individuals and artefacts. The model has been adopted in HCI to clarify problems of the interaction of users and computers. Distributed cognition argues that cognitive accomplishments are usually achieved in conjunction with artefacts. In these artefacts, representations of knowledge are embodied as, for example, in a thermometer or other measuring devices, which contain the accumulated information about this scientific area. Results achieved by using such cognitive tools emerge from the interaction between human and artefact and cannot only be attributed to human activity.

Representation of knowledge is embodied in everyday objects

Humans do not have to remember everything but extract visual clues from the environment

In many cases, human users of information technology do not possess coherent and comprehensive mental models of the problem at hand. Such mental models only emerge in the process of using the technology because the information relevant for the solution of these problems is distributed among humans and computers. O'Malley and Draper[83] argue that computer users do not possess and also do not need such coherent models because they can, in many situations, extract the relevant information from the environment. In this way, users

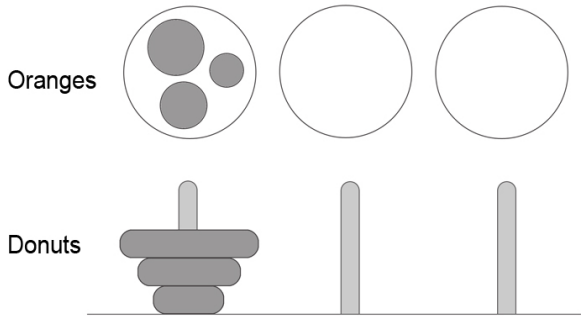

Figure 7.6: Oranges vs Donuts representation of Towers of Hanoi (adapted from Liu et al.[72])

can alleviate the burden on short term and long term memory. In free recall experiments, users of word processors are usually unable to recall many menu items. This low achievement is, according to O'Malley and Draper, due to the fact that users employ stimuli from the context of the interface to guide their search processes. They either use semantic relationships between header and menu item or information about spatial location to find relevant commands. The successful usage of a word processor (and similar programs) is, therefore, probably due to the interaction between user and system and depends on distributed representation of the relevant information on the computer as well as in the mind of the user. In this context, the strengths of each information processing system (humans and computers) are utilised and both systems complement each other. Hollan et al[58] argue that computer interfaces should be designed in a way to support this process efficiently.

In relation to visual artefacts, probably the most compelling work comes from Zhang & Normans theory of distributed representations. Central to the theory is the concept of the Representational Effect: "The phenomenon that different (visual) representations of a common abstract structure can generate dramatically different representational efficiencies, task complexities and behavioural outcomes"[128]. It has been argued that the design of visualisations should carefully consider this effect[72] as every representation offers various possibilities and has specific constraints. The Towers of Hanoi problem can, for example, be represented as different sized donuts on pegs or oranges on a plate (see Figure 7.6). The donuts-on-pegs representation is inherently easier because constraints of the problem are part of the analogy, as only the topmost and largest donut can be removed (only one can be moved at a time). Users adopt situated solution strategies using previous practical experience rather than abstract mental plans. This phenomenon is more consistent with distributed cognition than with traditional problem-solving theories. It seems to be a plausible assumption that similar strategies are used in interacting with information visualisation tools as every object on the screen offers specific perceived affordances (e.g., a button looks like an object that can be pressed).

Problem solving depends on context rather than abstract plans

Interaction is, therefore, very important although Liu et al.[72] point out that

interaction is still a concept which is not very well understood, and research is required into how people develop interaction strategies during sense-making and analytical reasoning.

7.2.3 Problem Solving

It has been argued that the exploration of data represented by visualisations is to a certain extent a problem solving activity. In problem solving, researchers usually distinguish between well-defined and ill-defined (or ill-structured) problems. The latter is where virtually no information about the problem and possible solutions are available, so the early stages of problem solving (recognition, definition, representation of a problem) are a challenging task. If the problem is well-defined, the emphasis of the problem solver's activity is on the later stages (development of a solution strategy, progress monitoring, evaluation of the solution). In addition, the solution path can often be described by an algorithm, which is not possible with ill-defined problems because they usually necessitate radical changes in problem representation.

Ill-defined problems are a challenge

An example for an ill-defined problem, which might necessitate radical change of representation, could be described as follows. Imagine a person going to work by car. One day, the car breaks down, and expensive repair is necessary. The person has to decide, whether they wants to repair the car or buy a new car. The problem to solve in this case, is the consideration of whether it is more expensive to repair the old car or buy a new (or used) car. But they might also consider not to buy a new car at all, but take the bus to go to work instead. Often, such radical reformulations of problem representations are not self-evident. In the case of ceasing to use a car, this has serious consequences for the life style of a person. This is, therefore, not an easy choice.

So far, research into problem solving (e.g., Simon's theory of problem solving) has concentrated on well-defined problems, although most problems in everyday life are ill-defined. Likewise, the problems for which interactive information visualisations are developed are often ill-defined. The Andrienkos[7] point out that a common goal in explorative data analysis is to 'get acquainted with the data'. This is a very general goal, and often more specific questions are only formulated after a general overview of the data. This usually is an iterative process of exploration. At the beginning, the problem is not defined in great detail, and radical changes of representation (e.g., another type of visualisation) in the course of the exploration of the data are possible.

Gaining insight is about discovery and is often unexpected

In this context, the concept of insight plays an important role. Increasingly, the term 'insight' is being used[82, 127] to denote that the exploration of information presented by visualisations is a complex process of sense-making. Saraiya et al.[95] define insight "as an individual observation about the data by the participant, a unit of discovery". They observe that the discovery of an insight is often a sudden process, motivated by the fact that the user observes something they have overlooked before. It is the purpose of visualisations to support this process and make the detection of insights easier. North[82] points out that the definition of insight used in information visualisation is fairly informal

and that researchers tend to use implicit conceptualisations. He posits that important characteristics of insights are that they are complex, deep (building up over time), qualitative (not exact), unexpected and relevant. Yi et al[127] also argue that there is no common definition of the term 'insight' in the information visualisation community. They point out that insights are not only end results, but might also be the source of further exploration processes. At the beginning of such exploration processes, there is often no clearly defined goal, and insights might be gained by serendipity. They assume that a vital question is how people gain insights, and they identify four distinctive processes how this might be done: provide overview (understand the big picture), adjust (explore the data by changing level of detail/selection, e.g., by grouping, aggregation, filtering), detect patterns and match the user's mental model (linking the presented information with real-world knowledge). The authors note that barriers to gaining insight include inappropriate visual encoding, poor usability and clutter.

How we gain insight is a vital question when designing visualisations

There is some similarity of the ideas about insight in information visualisation/visual analytics and the concept of insight proposed by psychology, especially in the area of human reasoning and problem solving[106]. The term insight was first used in psychology by Gestalt psychologists. Gestalt psychology conceptualises insight as a result of productive thinking, which goes beyond existing information. It often comes suddenly as a consequence of a complete restructuring of existing information. Gestalt psychology is based on holistic cognitive processes, which means that we do not solve problems by trial and error in a stepwise process (as behaviourism had assumed), but by detecting the meaningful overall structure of a situation.

Gestalt psychology suggests gaining insight is about restructuring existing information

Mayer[77] points out that research concerning insight concentrates on the first phases of the problem solving process (especially the representation of the problem) and on non-routine problems, that is problems, which problems solvers have not solved previously. He describes five interrelated views of insight based on the assumptions of Gestalt psychology:

- Insight as completing a schema
- Insight as suddenly reorganising visual information
- Insight as reformulation of a problem
- Insight as removing mental blocks
- Insight as finding a problem analogue

In principle, all of the above mentioned aspects are relevant for the clarification of the processes related to interaction with visualisations, but some of them seem to be especially important. 'Insight as suddenly reorganising visual information' is per se concerned about visual cues. It occurs when a person *looks* at a problem situation in a new way. Insight as the reformulation of a problem is related to that. In this case, a problem situation is represented in a completely new way. The suddenness of a solution is often seen as a characteristic of this theory of insights. It should be pointed out, however, that suddenness in this context does not mean that the solutions occur very quickly as restructuring may take some time, and even if a viable solution turns up, it usually requires some effort to realise it.

Insight may occur suddenly but often requires much unconscious effort

Whilst the importance of insight is for non-routine and ill-defined problems, in practice, laboratory experiments focus on well understood puzzles in order to make the empirical research more tractable. These puzzles are new to the subjects being studied, but typically have a single 'right' solution and all the information needed available (see for example, the puzzle in Figure 7.7).

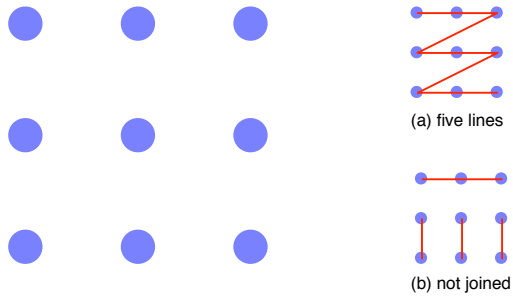

Figure 7.7: Nine dots puzzle: draw four straight lines that go through all nine dots, but without lifting pen from the paper. Note (a) and (b) show two incorrect solutions (a) has five lines not four and in (b) the lines cannot be drawn without lifting the pen. (see Figure 7.10 for solution)

Expert decision making often uses a highly developed mental model

Research into expert decision making in critical systems may provide an alternative path from understanding to insight. Klein has investigated how workers such as fire fighters, pilots and military personnel can resolve problems in high pressure environments[68]. He proposes that naturalistic decision making is often recognition primed, based on an individuals projected model of causal relationships. He provides a compelling example of how a naval officer was able to distinguish between an oncoming missile and friendly aircraft in a very primitive visual display. This difference would be impossible for a non-expert to identify as it involved the integration of both visual feedback and a highly developed mental model of the battlefield. This style of investigation is highly relevant for understanding the 'A-ha' moment that allows expert decision making to occur.

The usage of analogies also plays an important role for getting insights and is often mentioned as a source of creative thought[59]. In information visualisation, space is usually used as an analogy for other, more abstract phenomena (consider a scatterplot of engine size vs. miles per gallon). As human beings are highly capable of processing spatial information coming from their environment, space is a powerful analogy. In recent years, experimentation has taken place to clarify the concept of insight. The results of this research might form a valuable input for visual analytics, especially because it emphasises the reasoning processes associated with using information visualisations.

7.2.4 Interaction

The previous sections have concentrated on how humans perceive visual artefacts within abstract representation of data and try to make sense of these in order to gain information. We have also looked at work on modelling interaction and developing theoretical frameworks. The importance of interaction has been emphasised, as it is this that provides the opportunity for the user to explore the dataset. Whilst we can make good use of the large amount of research effort under the umbrella of HCI, there is not so much work focussed on visual analytics. Indeed, one of the recommendations from Illuminating the Path[111] was the creation of a new science of interaction to support visual analytics.

Interaction is vital in visual information discovery

A comprehensive review of the literature on interacting with visualisations is given by Fikkert et al.[45], although the authors do focus on virtual environments and associated display and interaction devices rather than information visualisations.

Attempts have been made to classify interaction for information visualisation[25]. More recently Yi et al.[126] identified the following categories of interaction:

We should think about the users' intentions when designing interactive systems

- select : mark data items of interest, possible followed by another operation,
- explore : show some other data e.g., panning, zoom, resampling,
- reconfigure : rearrange the data spatially e.g., sort, change attribute assigned to axis, rotate (3D), slide,
- encode : change visual appearance e.g., change type of representation (view), adjust colour/size/shape,
- abstract/elaborate : show more or less detail e.g., details on demand, tooltips, geometric zoom,
- filter : select or show data matching certain conditions,
- connect : highlight related data items e.g., brushing (selection shown in multiple views).

It useful to group together different interactive operations in this way, but possibly a more important outcome is a vocabulary to think about users' intentions when exploring datasets.

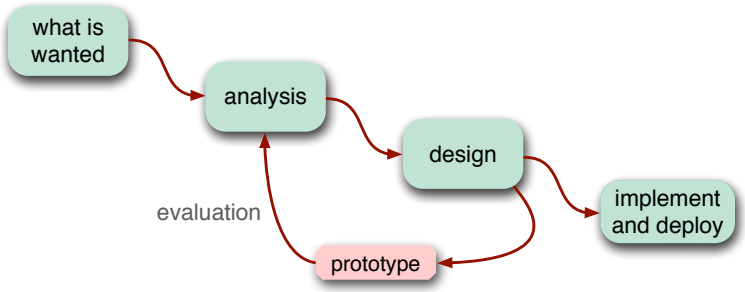

Figure 7.8: Typical user interface design process[36]

7.2.5 User Evaluation

User evaluation is at the heart of both research into human computer interaction and usability professional practice. This is because both researchers and practitioners recognise the limits of their ability to predict users responses to complex interactive systems and so there is always a need to test with real users, if possible, in real situations. For this reason, user interface design processes usually involve a tight cycle of prototyping and evaluation (see Figure 7.8).

Within HCI, evaluation techniques fall roughly into two styles:

- *quantitative evaluation* emphasises measurable outcomes, typically task completion time and error rate,
- *qualitative evaluation* emphasises more interpretative analysis of video and audio logs, or direct observations

Quantitative evaluation is often performed within well-controlled situations or laboratory settings, whereas qualitative evaluation is often performed 'in the wild' in real world situations, or artificial ones made closer to reality. Sometimes the two are seen as alternatives with strong proponents for each, but they can more productively be seen as complementary offering alternative insights.

While not diminishing the importance of effective evaluation, there is also a growing realisation that user evaluation, at least interpreted in a simplistic sense, is not always the most appropriate tool for all stages in the design and research processes. The tight cycle of prototyping and evaluation works well for refining and fixing the details of a design, especially in well understood domains. However, it is not so effective at arriving at novel designs or establishing the insights needed to drive new design ideas.

Within the information visualisation community, there is an ongoing discussion about methodological approaches for evaluation[13]. In this context, researchers argue that the measurement of time and error are insufficient to evaluate information visualisation techniques and tools because visualisation is typically exploratory in nature: interaction with information visualisation yields insights rather than information. This discussion is highly relevant for visual analytics because it emphasises the importance of the human reasoning processes as a whole, discovering new patterns within data rather than performing a known task in an 'optimal' way.

Quantitative evaluation is not appropriate for exploratory visualisation systems

One approach to this is to adopt more qualitative approaches. One example is grounded evaluation, an iterative design process that uses qualitative studies as a form of evaluation that can be carried out before initial design has been recommended. This is based on the recognition that to ensure the utility of visual analytics solutions it is necessary to ensure that the context of use is focused upon throughout the development life-cycle.

While this and other qualitative methods are better able to deal with the exploratory nature of the use of visual analytics, they still face the problem that users may not be able to appreciate the potential of radically new technology.

In such cases, it may be better to regard early prototypes as *technology probes*; that is being there to expose users to new ideas and then use this as the means to obtain rich, usually qualitative, feedback.

If evaluation is set within a wider context of 'validating' designs or research concepts, then these different approaches can be seen as building parts of a larger argument that may also include previous literature, published empirical data, theoretical models, and expert insights.

Chapter 8 discusses evaluation as an issue within visual analytics.

7.2.6 Early Application Examples

While basic theories and knowledge of human capabilities and behaviour can be applied from first principles, more applied knowledge is also needed. This is especially important when multiple factors come together. For example, complex interactions cannot be thought of as a combination of simple interactions often studied in pure science. In addition, it is only after protracted use that many issues become apparent, making it particularly hard to assess novel technology. This is of course the case for visual analytics as it is a new field. Happily it is possible to find much older systems and areas, which share essential characteristics with visual analytics and indeed would probably be termed as such if the phrase had been used when they were first established. Such systems are an opportunity to explore more applied issues with the benefit of hindsight and in some cases long-term use. They offer a wealth of existing knowledge to help us design for new visual analytics systems, and also a comparison point to assess the impact of changing factors such as massive data volumes or new interaction technology.

Can learn from early applications that share characteristics of visual analytics

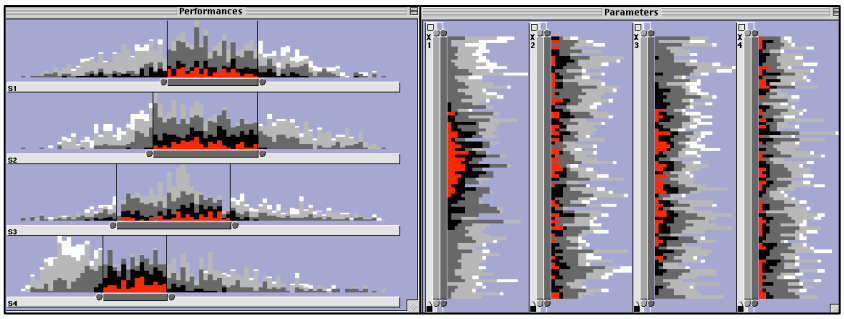

Figure 7.9: Parameter selection sliders of the Influence Explorer[117]

One example is the Influence Explorer shown in Figure 7.9[117], which used extensive computation to simulate the space of design parameters of light bulbs and then presented the results using interactive visualisations. Unlike much of the early work in visualisation, which was often framed around particular ideas for techniques, in the case of Influence Explorer the problem domain came first and the innovative interactive visualisation was developed in order to solve the problem. Influence Explorer embodies many critical features of visual

analytics, notably the fact that there is a joint activity of analysis involving human perception and insight as well as computation and visualisation. It made use of sampling, as the complete data space of design parameters was too large to simulate exhaustively. It also allowed the user to 'peek over the horizon'.

Influence Explorer embodies many critical features of visual analytics

Most visualisations show you the effect of the current viewing parameters and rely on the user to actively interact to see alternatives temporally; in contrast, by exploiting a technique first introduced in the Attribute Explorer[118], the Influence Explorer's parameter selection sliders include miniature histograms, which let you know what the impact would be of alternative or future selections. The Influence Explorer is highly unusual, it is the first system of which we are aware that actively used sampling in visualisation and HiBROWSE[42] (a largely text-based faceted browser) is the only similar system allowing this 'peeking'.

Business problems have long required complex analysis. One example is sales forecasting. In this case past sales data is typically modelled using various forms of time series analysis and the predictions from this visualised as simple graphs. However, the computer predictions cannot be used on their own as there are many additional internal and external factors such as sales promotions, advertising, competitors, and even the weather, all of which can influence future sales. Using the forecasting system involves the selection of data (e.g., do you include historic data on a product that had a major change?), the choice of forecasting algorithm (e.g., seasonal adjustments, kind of time series analysis), and then the inputting of 'adjustments', that is manual changes to the predicted outputs – effectively, an early example of visual analytics combining computation, visualisation and user interaction.

Sales forecasting is an early example of visual analytics

The analyst may face pressure from members of the organisation, for example, a product division may wish to see higher forecasts in their area, and the results are not the end point of the process as they feed into decision making meetings where the forecasts are used to form plans for stocking, pricing etc.. Even the notion of accuracy that is central to the study of this process is problematic as the predictions feed into the process that is being predicted; indeed, forecasters are often more concerned that their forecast are reasonable and make sense to the recipient, and are less concerned with an ideal 'best' prediction.

There is a tight integration between users, computation process, organisational influence and the reflexive nature of visual analytics

As we can see, sales forecasting emphasises the need to take into account the whole picture in Figure 7.1: the tight interaction between users and computational processes, the organisational and political pressures that influence the analysis, and the reflexive nature of visual analytics, where its outcomes may affect the data on which it depends. These are lessons for visual analytics more generally. For example, in the homeland security applications that are the focus of Illuminating the Path[111], there may be a predicted attack by a terrorist group and the suspects consequentially arrested; the attack will therefore not take place, but this does not mean the prediction was 'wrong'.

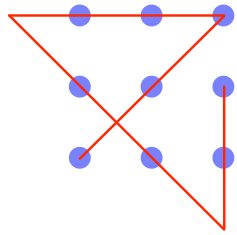

Figure 7.10: Solution to nine dots puzzle (see Figure 7.7)

7.3 Challenges and Opportunities

7.3.1 Stakeholders and Communities

The user facing a visualisation is not alone. As already noted, the results of visual analytics are often for other people, perhaps a manager making a decision or maybe for the general public. How can visual analytics systems support not just the problem solving and analytic task of the direct user, but also the production of static or interactive visualisation for these secondary users? In some situations, the end user may not be a visualisation expert but an ordinary day-to-day user of a computer. When issues of eDemocracy and open data are high on political agendas; we need to consider ways to give the general public tools to understand and to analyse vast volumes of data, in their own ways.

General public need tools to understand data in their own ways

Appropriate design methodologies have to be developed to support the reasoning processes of non-expert users. Simplicity, natural metaphors and intuitiveness will be important aspects in this context. Guidance is often required, especially for novice users, on what visualisations (scatterplot, parallel coordinates, treemaps, etc.) are appropriate for a given task; the focus should be on the problem to be solved rather than the type of raw data. Some visualisation tools (e.g., Tableau Software) attempt to do this by suggesting alternatives, albeit only simple visualisations, and some systems automatically generate the 'appropriate' visualisation.

Accepting that the immediate user is part of a community can help. As mash-up technology becomes commonplace, local experts could be one way in which more customised visualisations could be made available for those with less technical skills. For this we need to consider both the social means to share such components and also the technical means by which visualisations and processing can be slotted together in a plug-and-play manner. This is also an important requirement for effective research to prevent researchers wasting time implementing basic mechanisms in order to experiment with a single novel feature. While there are existing open-source visualisation systems a combination of poor documentation, reluctance to use external solutions, and lack of knowledge of what is available, makes reuse rare. Google's visualisation framework is a notable example that has allowed much

Community effort is required to share visualisation components

more mashing of visualisations – can we envisage richer variations of this approach?

7.3.2 Applying Psychological Theory to Real Applications

As discussed in Section 7.2, there are substantial amounts of existing and emerging perceptual and cognitive knowledge that is highly relevant to visual analytics. For example, in medical analysis, slides or X-ray plates are examined for signs of tumours. Eye gaze data of expert radiographers has shown that important artefacts were looked at but did not get reported by the person doing the analysis; that is the expert's level of visual attention was a better estimate of the presence of a tumour than the conscious choice. In other kinds of visual analytics is this likely to also be an issue? This example raises issues concerning foveal vs. peripheral vision, the importance of the latter in many forms of visual processing is only recently being understood and could be a powerful asset in visual presentation. Furthermore, implicit learning and more conscious reasoning are separate human processing systems. Often people have a gut reaction or make an instinctive decision but the process is unclear; modelling this type of decision making is a difficult problem, and this kind of issue is only recently being addressed in 'dual space' theories of cognition.

Make existing knowledge on perceptual and cognition more accessible as well as promoting new research in this area

At a more detailed level, this relates to design decisions such as ratios between blank space and 'used' space, text lists vs. graphic display, and display aesthetics. There is a real challenge to mine the literature to bring out these more general issues that are often buried in papers describing specific techniques and systems. However, some new fundamental knowledge is also needed. One such case arose in connection with dot densities as found in dense scatterplots. One of the key measures used in assessing perceptual stimuli is 'just noticeable difference', for example, when two slightly different sounds are played, when do subjects cease to notice that they are not the same. However, whilst the data was present for visual stimuli relating to solid blocks of colours or shades, no similar data was found in the base psychological literature for dot density. In order to assess the acceptability of differing sub-sampling regimes, Bertini and Santucci[15] had to perform fresh experiments to determine just noticeable difference for dot density.

7.3.3 Understanding the Analytical Process

We still do not have a complete understanding of information visualisation let alone visual analytics. There are examples: Illuminating the Path[111] cites several frameworks for understanding the sense-making and analysis process, Yi et al.[127] characterise the process of gaining insight, and work by de Bruijn and Spence[34] considers different classes of browsing (search, opportunistic, involuntary and perusal) and suggest interaction modes to support such behaviour.

A key challenge is to extend such frameworks to consider the entire analytic process. Most of the frameworks are focused on the stages of visual exploration, although the data/frame theory of Klein et al.[68] also considers the mental representations of the analyst and Pirolli and Card[85] have a dual loop model of the sense-making process, which takes into account the way mental schema give rise to hypothesis and interact with the exploration of external data. However, there seems to be an absence of a) consideration of the visualisation and understanding of the parameter setting of computational processes, and b) the externalisation of the analysts mental representations. Both are connected to distributed cognition and ecological perception (see Section 7.2) as a) they should take into account that the perception of a visualisation is related to the (interactive) setting of the parameters that gave rise to it, and b) the externalisation of mental constructs makes them available for perceptual and more explicit critique. Understanding this more clearly can help suggest appropriate interaction mechanisms, for example, making use of design rationale, annotations, history and provenance.

Further understanding of the analytic process is needed in order to design appropriate interaction mechanisms

7.3.4 The Need for Design Guidelines

There is a need to create clear design guidelines for designers of visual analytics systems and also the means to share practical design knowledge. Many writers on both visualisation and machine learning are wedded to their own particular techniques, so it is often hard to obtain unbiased views of the adequacy of techniques beyond the advocacy of their proponents. This is always a problem, but as the computational side of visual analytics is more complex, having clear advice is correspondingly more important.

There are some steps in this direction including work on design patterns for information visualisation and for visual analytics. However, given that visualisations are applied/developed to support specific contexts, how can design knowledge be re-used in alternative domains? As an example, whilst the concept of Fisheye interfaces goes back more than 20 years, there are still calls for clear design guidelines. This is emphasised in a recent review of challenges in information visualisation, which suggests that the entire field of information visualisation is in the pursuit of finding the most effective visual metaphors. The authors point out that one single metaphor is unlikely to overcome the problems of cognition (intuitive from a users point of view), very large datasets and/or a high number of dimensions. A closely related challenge is the choice of an optimal level of visual abstraction (e.g., from the low level 1-to-1 correspondence of a scatterplot to high levels that involve clustering); however, as with visual metaphors, the choice is very dependent on the user and their experience, knowledge and goals.

The effectiveness of different visual metaphors and levels of abstraction is very dependent on the user

To date, efforts to communicate design knowledge have tended to focus on the re-use of pre-designed solutions. However, the need to design visualisations that reflect contextual system constraints restricts the utility of this approach. Rather than prescribing design solutions the development of related taxonomies of cognitive work systems and appropriate design methodologies have been proposed. These can be used by a designer to classify contextual problems and

to identify relevant design artefacts that will support the overall visualisation design process. While these taxonomies are only briefly outlined, their extension could enable and encourage the re-use of design knowledge across different work domains.

Interaction designers need guidelines based on underlying perception and cognition research

Cognitive empirical and theoretical knowledge is continually developing, but we cannot wait for this to be fully developed, but instead must create good engineering advice based on existing knowledge and update this as knowledge develops. Spence[105] highlights the need for 'brokers', people who are able to identify important factors in the perception and cognition literature and interpret these for the benefit of interaction designers. One example of this are 'design actions' that provide guidelines for some specific cases[34]; and the design patterns mentioned above also can be seen in this light.

Another example is the Ecological Interface Design framework. This provides visual design guidelines that support specific levels of cognitive control, including diagnostic activities[21]. While the framework has been validated across a range of complex process control systems, its applicability to intentional decision making and analytical model building requires further investigation.

7.3.5 Defining the Language of Visual Analytics

Clear definitions are essential for the advancement of science, but many of the concepts used in visual analytics do not have precise definitions. While anything involving human capabilities inevitably has fuzzy edges, there is a clear need to attempt to develop clear definitions of core concepts, subject to understanding the limits of such definitions once formulated.

One example is the concept of insight. Insight is an important concept for the perceptual and cognitive analysis of interaction with visual analytics methodologies reflecting the importance of reasoning processes in the task of interpretating large amounts of data, however, there is as yet no precise and systematic definition of this concept. Valuable input can be gained from research in cognitive psychology[106]. This research is influenced by Gestalt psychology, which conceptualises reasoning as a holistic and structured process.

Definitions are important

Even the term 'visual analytics' is itself potentially problematic. The adjective 'visual' suggests the use of sight only whereas visualisation is the action of creating a mental model (in the user). Modalities other than visual are important as perception is a holistic process, encompassing sound, touch, smell and taste; these modalities should also be considered in visual analytics. It is evident that these other senses, most significantly aural representations, have a part to play, but this needs to be emphasised lest the term visual analytics accidentally marginalises them. There is also confusion between information visualisation and visual analytics suggesting that greater clarity is required to explain the new issues that arise.

One step in the direction of greater clarity is Thomas's[110] discussion paper on a proposed taxonomy for visual analytics. This effectively creates a lexicon of key terms and some structure to them, but each term really needs a complete definition.

7.3.6 Observability and Trust

There is a need for the user to be made more aware of the visual analytics process in order to gain confidence in the results; for example, business intelligence is commonly used in the context of a decision support system and suffers from poor user acceptance, with the user often ignoring evidence in favour of (potentially biased) past experience. It is suggested that this mistrust in the outputs of a decision support system may be overcome by making the users more aware of the automated decision making process. This is exacerbated by the fact that people are not necessarily good statistical thinkers and so a challenge of visual analytics is to make the statistical methods understandable so the user has enough confidence in the results to counter biased opinions. As a further example, when comparing dynamic queries[2] with Attribute Explorer[118], filtering throws away data and potentially makes it more difficult to obtain a mental model of the data. On the positive side, selecting a facet and filtering the results at least gives the user an idea of the amount of data related to that facet. Note that while faceted browsing allows the user to rapidly reduce the amount of data, the Attribute Explorer is unusual in that its miniature histograms allow you to see an overview of the complete dataset and also assess potentially what may happen as you make further parameter filtering selections. As noted earlier in Section 7.2.6, in general it is rare for visualisation to give a 'glimpse over the horizon' or some idea of the potential results of applying a filter before actually doing the filtering.

Understanding the analytic process can give confidence in the results

One particular issue is the visualisation of uncertainty. Uncertainty takes many forms and some can be estimated quantitatively such as the statistical variance of estimates, but others are more qualitative such as the level of confidence you have in a particular data source (e.g., BBC news vs. sales literature of a competitor). The dual space understanding of cognition is critical here as humans have a primitive-stimulus response learning system that learns through repeated exposure. This effectively learns probabilities, but very slowly. The other mechanism is our more explicit memories and reasoning over them through abduction. This higher-order memory gives us one step learning of new situations, but is relatively poor at probabilities, without explicit mathematical analysis. A challenge for visual analytics, is to use machine processing and visualisation, to complement the human analyst's abilities in understanding uncertainties.

Users should be made aware of sources of uncertainty in the data

7.3.7 Evaluation of Novel Designs

As discussed in Section 7.2 and Chapter 8, issues of evaluation are a hot topic within the information visualisation community with regular workshops on the

topic[13]. Many different methods are used to study information visualisation methodologies[26], but more work is required to determine which of these methods are especially appropriate for visual analytics. It is an open question whether traditional methods of cognitive psychology or HCI are appropriate for the investigation of perceptual and cognitive aspects of visualisation. It maybe that we should develop entirely new methodologies that take into account both complexity of detail and context.

Evaluation of visual analytics is particularly difficult as problems are often ill-defined and open ended

Part of the problem is that visual analytics is about solving open ended problems and so it is hard to create meaningful tests as almost by definition these will be for known solutions: puzzles not problems. This is a problem in both research and real world application. For example, the management science literature on sales forecasting systems focuses almost entirely on 'accuracy' as the key evaluation parameter. However, as discussed in Section 7.2, forecasts affect the stocking, placement and advertising decisions of a company and hence sales (the full outer circle in Figure 7.1). That is, the visual analytics within this is itself part of the process it is predicting.

Problem solving involves gaining insight, and this occurs at different levels during the problem solving process. So we need to think about assessing the effectiveness of a design (in terms of interactivity and visualisation) on the generation of insight in: a) assessing the data and finding relationships, b) the capability to support hypothesis formulation, and c) how well the conclusions reached by the user at each stage of analysis can be traced so they can be verified by others.

New technologies may help evaluation

New technologies such as eye-tracking and even brain scanning, offer the potential for radically different ways of approaching evaluation. However, these are themselves areas of substantial complexity, for example, one problem is the appropriate interpretation of the data gained from eye-tracking studies and the definition of clear variables that can be measured by eye-tracking. It maybe that in the short term we need visual analytics to actually address some of the research challenges in these areas as they offer visual analytics new tools of investigation. It is an open question whether these new techniques actually offer any additional information than more qualitative methods such as cooperative evaluation. In general, evaluation of all kinds is also expensive and so in the world of practical visual analytics system design we also need low cost/resources methods.

7.3.8 Designing for the Analyst

Flexible designs allow analysts to customise the way they work

Visualisation designers have their own ideas about what constitute good designs and visualisations, and build these assumptions into the tools created for the analysts. However, analysts often do not think the same way as designers. While there may be a need for some standard tools for standard tasks, the challenge for the community will be the development of advanced tool sets for the analysts. These would enable the analysts to bring different functional capabilities together, enabling them to create visualisations and interact with

them in ways that are flexible yet robust. This presumes we have a good understanding of the information handling strategies that users invoke when working with the different sorts of data and documents. Liu et al.[72] also points out that many of the current visualisation systems are not flexible enough to allow user customisation and hence may inhibit the analysts managing data sources and hypothesis creation in a way their feel as appropriate.

Dealing with biased opinions has already been noted as a problem. The work of an analyst is influenced by a host of cognitive biases (e.g., confirmatory bias and anchoring) and many of these biases are often set in motion by the way information is 'fed' into the perceptual-cognitive processes. How do interactive visualisations (designs) influence biases? We need to know and understand the effect of information designs that combine interactivity and visualisation on interpretation and analysis, and the inter-play of that with known cognitive and perceptual biases. The way in which a system presents patterns and cues, and how their significance and salience are rendered, can activate biases. Therefore, it is important to be aware of when they may occur and then develop appropriate controls to minimise such effects.

Need to be aware of and minimise cognitive and perceptual biases

7.3.9 Changing Interfaces: Users, Data and Devices

In current practice, the mathematical models used in decision support are processed offline and only the results are visualised by the user. There is a need to make this process more dynamic both in terms of parameter setting and also the choice of models; however, this will create demands on the underlying visual analytics architectures. Looking at the choice of visualisations, some are highly information intensive, but also very complex, whereas others give less information, but maybe more informative for a novice. There is a real challenge in adapting these visualisations to suit the user and the data, whether under direct user control or semi-automatically; and furthermore to transition smoothly between different levels of visualisation complexity. Similar issues arise when dealing with different devices and hardware from mobile phones to wall-sized multi-screen displays.

Systems need to adapt to a wide range of users, data types and sources, and input/output devices

In the business intelligence world, visual analytics is often presented as a set of visualisations (e.g., treemap, heatmap) from which people with 'data overload' can select an appropriate solution, with little consideration of either the problem to be solved or the process required. We clearly need to be able to offer more guidance as to which methods are better suited to particular classes of problems. The issue here is not the kinds of raw data (time series, categorical, network, etc.), but what we want to do with the data. Furthermore, there are different levels and timescales of problem solving in business (e.g., financial, sales) from everyday decision making to longer term corporate policymaking. Visual analytics is typically applied to 'bigger' decisions, but many systems do not take into account the long-term use and re-use, such as means to annotate past use to inform future interactive sessions. The use of visual analytics for much more moment to moment decision making is perhaps even more problematic and would likely require some automatic aid.

Users need guidance in choosing an appropriate visual analytic solutions for a given task

The Web was designed to ship fairly traditional data from CERN to physicists across the world. However, the Web has more recently given rise to very large-scale data such as folksonomies and tag data, co-occurrence data used in recommender systems and RDF ontologies for the semantic web. Web data presents new problems being both large scale, but also typically less-tabular, and more relational; in the case of semantic web there is the potential for inference and data to become intertwined. As with visual analytics itself, we can easily find 'Web-like' data before the Web, so there are places to look for inspiration, but certainly this is likely to pose fresh challenges for large scale visual analytics in the years to come.

7.4 Next Steps

From the previous sections, we can identify several necessary actions in order to progress understanding of human aspects of visual analytics:

- appropriate design methodologies need to be developed taking into account all the human issues impacting visual analytics as discussed in Section 7.2, the heterogeneity of devices and data as discussed in Section 7.3.9, and range of stakeholders (Section 7.3.1)
- these need to be backed up by design guidelines and clear definitions, especially for non-expert users of visual analytics systems (Sections 7.3.1, 7.3.4 and 7.3.5)
- of particular importance are the development of interaction and visualisation mechanisms that will enable analysis to assess more confidently the reliability of results of visual analytics systems, including issues of uncertainty and provenance of data (Section 7.3.6)
- these need to be backed up by appropriate evaluation mechanisms, potentially including emerging techniques such as eye tracking (Section 7.3.7)
- all of the above require an ongoing development of the basic human science of visual analytics including brokering existing fundamental psychological and social knowledge, generating new such knowledge and most importantly creating robust and applicable holistic models of the visual analytics process (Sections 7.3.2 and 7.3.3)

In general, the topic of perceptual and cognitive aspects of visual analytics is highly interdisciplinary and these very heterogeneous disciplines provide interesting input for visual analytics. Whilst we have gone some way in establishing contacts between these communities, there is much still to accomplish.

8 Evaluation

8.1 Motivation

Visual analytics is a promising and ambitious concept. The aims are to enable people to get insight in large amounts of heterogeneous data, understand the underlying phenomena described by the data, to smoothly integrate multiple data analysis methodologies, and to offer support for the complete knowledge discovery process. These aims are very challenging. For many practical instances, it is unknown how to reach these; for existing solutions it is often unknown how well they realise these aims; and overall, there is a lack of solid findings, models and theories. As a result, visual analytics still has a long way to go before it can be considered a mature technology.

In making progress towards meeting these aims, evaluation will play a crucial role, but the characteristics of visual analytics presents difficult problems for effective evaluation. In this chapter, we elaborate on this by examining particular problems, then give an overview of the state of the art in evaluation, and finally present some recommendations for the research roadmap.

Evaluation include techniques, methods, modes and theories as well as software tools

Evaluation concerns here the assessment of the quality of artefacts related to visual analytics. Both *quality* and *artefacts* should be considered as broad container terms. Artefacts are not limited to software tools, but also include, for example, techniques, methods, models and theories. As visual analytics is both a science and a technology, the key aspects of quality are *effectiveness*, *efficiency*, and *user satisfaction*. In other words, artefacts should be evaluated on whether they fulfil their aims, on the resources required, and whether they meet needs and expectations of users. Taking a broad view, this includes aspects such as degree of fit in current workflows, performance, and ease of use. As argued in the previous chapter, users are central in all this, and awareness of their importance is still increasing, not only in visual analytics, but also in related fields. One example from geovisualisation is that the International Cartographic Association (ICA) has established a committee on Use and User Issues[1]

Stakeholders include developers and end users

The results of evaluation are important for all stakeholders. Integrators and end-users of visual analytics need to know about the quality of artefacts. Put practically, the developer of a new system that takes advantage of visual analytics techniques needs to know which techniques to choose for the problem at hand; users who have to select a system, a method, or even a parameter-setting need information to make the best decision, in order to save time and to prevent themselves from the use of inappropriate techniques, leading to

[1] http://www.univie.ac.at/icacomuse

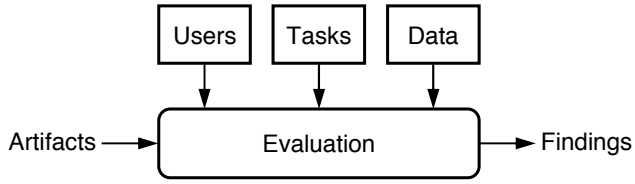

Figure 8.1: The main ingredients of evaluation

wrong results. Furthermore, good evaluation results are important to convince integrators and end-users to adopt novel techniques. Hence, evaluation is an important ingredient in the innovation process, from research to application on larger scales.

The task of researchers and developers is not just to develop new ideas and techniques; assessment of the quality, scope, and applicability of those innovations is equally important. As mentioned, those results are vital for end-users, but also for research itself. Evaluation can show which problems have and have not been solved, it provides benchmarks, against which new results can be compared.

However, for several reasons proper evaluation of visual analytics is not easy. First, visual analytics encompasses many different aspects and disciplines, which makes it hard to make generic statements; second, in visual analytics humans play a central role, in contrast to, say, statistics.

Evaluation involves users, tasks and data

Figure 8.1 shows a schematic overview of evaluation in visual analytics. Evaluation leads to findings on the quality of artefacts. Such findings are never absolute, but depend on *users*, *tasks*, and *data*, which taken together define the scope of the findings. To give a simple example, a finding could be that the use of scatterplots (artefact) is helpful to find clusters (task) in records with a limited number of real-valued attributes (data), provided that observers have had training in the proper interpretation (users). Such findings can be produced using relatively simple lab experiments, as all aspects are well-defined. Much more challenging is to obtain generic findings, such as when to use automated techniques instead of techniques with a human in the loop, for broad classes of users, tasks, and data. Another challenge is to obtain precise, quantitative findings, for instance on how much time is saved by adopting a technique. Again, solid findings would be highly useful, and to produce such findings is a major challenge for the field. However, an even more daunting challenge is to obtain findings that characterise the knowledge discovery process: the rationale behind the decisions taken by the user and the type (and quality and quantity) of insight being obtained.

Obtaining findings which can be applied generically is a daunting task

The complexity and diversity of users, tasks and data is high

The complexity and size of evaluation in visual analytics can be understood further by considering the ingredients (users, tasks, artefacts and data) in more detail. All these are complex in themselves. They are hierarchical, because different levels of abstraction can be distinguished; multivariate, because different properties can be distinguished; and heterogeneous, because in real-world scenarios, combinations of data, tasks, etc. usually have to be dealt with. This

complexity is within the core of the mission of visual analytics. Whereas other fields in visualisation often focus on specific user groups with well-defined tasks and standardised, homogeneous datasets, visual analytics aims at much more diversity. In the following, this diversity and complexity is discussed in more detail for users, tasks, artefacts, and data.

Users. The user community targeted at is large. In the ideal case, findings apply to the general user, but for specific problems specific users have to be targeted, and their capabilities, interests, and needs have to be taken into account (for more on this, see Chapter 7). At various levels of detail, a distinction can be made between professionals and a lay-audience; professionals can be split up into, for instance, scientists, data-analysts, managers, etc.; and of course, all these categories can be subdivided further, down to, for example, experts in European patents on laser-optics technology or bioinformatics researchers dealing with crop diseases. Furthermore, aspects like age, country, culture, gender, training, perceptual and cognitive skill levels, or motivation can have an influence on the performance obtained when an artefact is used.

Obtaining appropriate expert users is difficult; results from using students may not be representative

This leads to interesting problems for evaluation. For example, dealing with experts requires a thorough understanding of their needs and wishes, such that the appropriate aspects are evaluated; also, such experts are often scarce and have limited time available. One often used escape route is to replace the experts with undergraduate students and have them evaluate new methods and techniques, but it is unclear to what extent the results found carry over to real-world users.

Tasks. Users apply visual analytics to fulfil tasks, and here again complexity strikes. In information visualisation, often just low-level tasks are considered, such as spotting trends, clusters, and outliers. However, people that use visual analytics have to carry out tasks like protecting the safety of a computer network or a transportation system, manage a company, or decide on a policy. There are many levels between such complex responsibilities and the elementary tasks; and, given the ambition of visual analytics these fall within the scope. A practical and important issue here is that such more complex tasks do not lend themselves well to standard lab-experiments. They can require from days to months to complete, require in-depth expertise of the subjects, and these tasks are too important to allow wrong decisions to be made. In the following section, current approaches to handle this are discussed.

Complex and extended tasks are often not suitable for laboratory experiments

Artefacts. The artefacts of visual analytics can also be considered at various levels of detail. On a very detailed scale, one can study the effectiveness of, say, graphical representations or a specific technique. On a higher level are the software tools, to be compared with other tools. On a still higher level, one can study the suitability of such technologies in general. This implies that one also has to study aspects such as the available tutorial material, coupling with other systems, and the costs involved. Besides these levels, the scope of the artefacts varies greatly. Artefacts can relate to visualisation, automated analysis, knowledge management, presentation, data cleansing, etc., and in a full-blown

Artefacts for evaluation range from graphical representations to the suitability of particular technologies

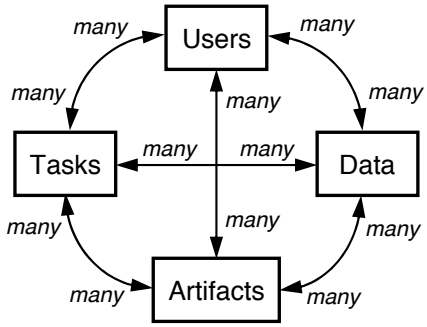

Figure 8.2: Relations between users, tasks, data, and artefacts

environment for visual analytics, all these issues have to be addressed in one way or another.

Data. The data to be considered is also complex (see Chapter 3 for a detailed discussion). Whereas a standard visualisation usually deals with homogeneous, single datasets (which is often difficult enough), visual analytics has to deal with combinations of heterogeneous data (for example, weather data, multi-media, written reports), huge amounts of data, requiring reduction via automated methods; and new data can arrive or is sought during the analysis.

In summary, we argued that users, tasks, artefacts, and data in visual analytics are complex and heterogeneous. In reality, it is even more complex, as all this complexity multiplies, as shown in Figure 8.2. In a simple laboratory experiment, one standard user evaluates a few variations of an artefact, for a small number of well-defined tasks, using similar datasets. In the real world, people use several tools simultaneously, have a variety of tasks, use many different datasets, cooperate with other people, and all this over extended periods of time in a flexible and dynamic setting. All this makes evaluation a difficult task, and shows that it is not easy to find generic and reliable answers to the question of which artefacts to use and when.

In the next sections, we describe the state of the art of evaluation methodologies in visual analytics and present recommendations for improvements of current approaches.

8.2 State of the Art

Visual analytics artefacts should be evaluated in terms of effectiveness, efficiency, and user satisfactions to assess their quality. This requires evaluation methodologies, covering a wide range of algorithmic performance measures to real-world technology adoption and utility metrics. Chapter 6 of Thomas and Cook's book[111] outlines evaluation approaches for visual analytics on three levels: *component*, *system*, and *environment*. With respect to components, there

exists a proliferation of isolated evaluations. On the system level, success is hard to quantify and difficult to trace back to individual components or computations. Here, it is important to track the history of investigation, e. g., in analytic workflows. Metrics are needed to address the learnability and utility of systems. Quantification of insights is essential (examples in bioinformatics have recently appeared[95]). On the environment level, evaluation needs to consider technology adoption. Across all levels, one needs to measure the benefit of the technology in producing an improved product.

Metrics are needed to measure usability, learnability, quantification of insights and technology benefits

Visual analytics technology is used by people who carry out their tasks with visualisation tools, sometimes over long periods of time, searching for information in various ways[88]. This means that, in addition to measures of performance and efficiency, there is a need to evaluate the interaction of people with the visualisation tools in order to understand their usability, usefulness, and effectiveness. Such aspects can be addressed by empirical evaluation methodologies, as often used in the fields of human-computer interaction (HCI) and computer-supported collaborative work (CSCW).

This section gives an overview of the state of the art of such methods for evaluating visual analytics.

8.2.1 Empirical Evaluation Methodologies

A range of evaluation methods exist for examining interactive techniques[26]. These include quantitative methods, qualitative methods, mixed method approaches, usability studies, and informal evaluation techniques (the classes not being mutually distinct). Depending on the chosen method one can, for example, examine in a controlled environment (such as a laboratory) very specific questions for which a testable hypothesis can be formulated, and this can lead to conclusions with high confidence. Another type of evaluation can look at broader questions using qualitative methods. Here, the focus is on data acquisition through observation and interviewing. A wide range of specific techniques exists in this area that can be used depending on the types of questions to be answered. In addition, there are also mixed-method techniques that combine aspects from both qualitative and quantitative evaluation. A separate category within evaluation techniques is usability evaluation, which deals specifically with the ease of use of interactive tools. Here, a combination of quantitative and qualitative evaluation can be employed. Finally, informal evaluations can be used. These involve fewer people who give feedback on a visualisation or the interactive system used to create them, providing anecdotal evidence for its usefulness or effectiveness, which can be useful for techniques that mainly focus on a technical contribution.

Range of evaluation methods include qualitative, quantitative, combined and informal

A number of papers discuss evaluation methodology in general. In his seminal paper, McGrath[78] identifies important factors that are all desired but not simultaneously realisable in evaluation studies: *generalisability*, *precision*, and *realism*. He also classifies specific evaluation approaches with respect to their abstractness and obtrusiveness and, on this continuum, indicates the respective position of the three aforementioned factors.

Evaluation strives to be generalisable, precise and realistic

Qualitative and longitudinal studies are particularly suitable for information visualisation

In her overview on evaluation methodologies for information visualisation[26], Carpendale carefully discusses the various approaches in quantitative and qualitative evaluation, following the terminology of McGrath: field study, field experiment, laboratory experiment, experimental simulation, judgement study, sample survey, formal theory, and computer simulation. In particular, she emphasises the importance of qualitative approaches as a valid group of evaluation techniques. Plaisant[86] discusses the challenges of evaluation, also in the context of information visualisation. In addition to controlled experiments, the need for longitudinal studies is stressed. Recommended steps to improve evaluation and facilitate adoption are: repositories (data and tasks), collecting case studies and success stories, and strengthening the role of toolkits. Chen and Yu[28] report on a meta-analysis of empirical studies in information visualisation. They included only studies on tree or network visualisations and restricted themselves to studies with information retrieval tasks. Due to the very strict requirements, of the original 35 studies selected only 6 remained for the final analysis. They found that due to the diversity of studies it is very difficult to apply meta-analysis methods. They conclude that the measurement of higher cognitive abilities is especially hard and more task standardisation in cognitive ability testing is required.

Repositories of data, tasks, case studies are useful

Task based evaluation is generally not suitable to measure insight

Zhu[129] focuses on the definition of effectiveness of visualisation and how to measure it. Current definitions of effectiveness are reviewed and a more comprehensive definition of effectiveness is introduced, which comprises accuracy, utility, and efficiency. As one aspect of efficiency of evaluation, North in his Visualisation Viewpoints paper[82] focuses on the question of how to measure insight, the ultimate goal of all of visualisation. One of his main observations is that task-based evaluation is too narrow. What works well for a given task might not work at all for tasks not studied. Generalisation from simple to difficult tasks is hard. Either more complex tasks are needed, or one may eliminate benchmark tasks completely and put the emphasis on more qualitative insights. Involving learning processes as in traditional education will be helpful. Finally, Munzner[80] presents a nested model for visualisation design and evaluation. She subdivides the process of creating visualisations into four nested levels: domain problem characterisation, data/operation abstraction design, encoding/interaction technique design, and algorithm design. She then argues that distinct evaluation methodologies should be used for each of these levels because each of the levels has different threats to its validity.

Generalisation is difficult

8.2.2 Examples of Evaluation

We now discuss a number of specific approaches to evaluation of visual analytics that are used in practice.

Program understanding and software visualisation. In the field of program understanding and software visualisation, evaluation of combined analysis and visualisation techniques and tools already has a rich history.

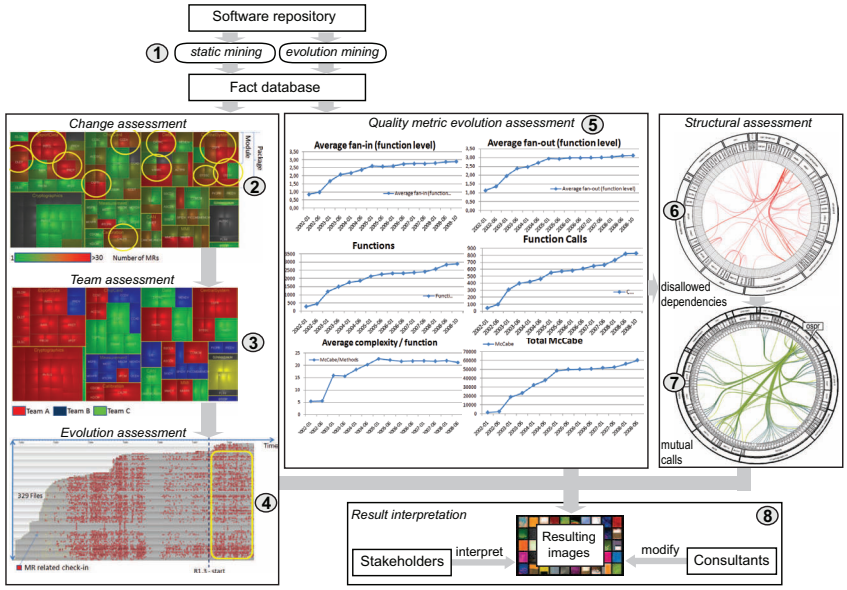

Figure 8.3: Visual analytics for software product and process assessment. A large software repository is mined (1), a variety of aspects are visualised and assessed (2)-(7), findings are discussed with stakeholders and consultants (8)

Several particular difficulties for evaluation exist in this field. The diversity of tasks and questions in software analysis requires analysts to easily combine and customise their analysis and visualisation tools in ways, which often go beyond what these tools were designed to support. Many such tasks require high precision analyses, such as extracting exact call, control flow, and data flow graphs in static analysis or the exact detection of structural or execution patterns in program verification. This is a challenge both for the analysis tools, but also for the creation of visualisations able to convey exact insights at fine-grained levels. Scale is also a problem. Modern software projects have millions of lines of code, structured in tens of thousands of entities, which are modified continuously by hundreds of developers over many years. Although most software systems are hierarchically organised, which enables level-of-detail and aggregation techniques to simplify visualisations, many analyses such as impact and correctness assessment involve concerns, which cut across boundaries and make localised examination difficult.

Scale is a problem

Shneiderman and Plaisant[101] argue for the need for longitudinal studies as an effective means for understanding the issues involved in tool adoption. A useful taxonomy of evaluation procedures is given by Kraemeret al.[84], who considers surveys, interviews, observational studies and think-aloud studies. These studies form valuable input in organising effective evaluations in the larger context of visual analytics.

There are many types of evaluation including, longitudinal studies, surveys, interviews, observation and think-aloud studies

Several evaluations of the effectiveness of visual analytics solutions in software

A large study of visual analytics in software maintenance suggest a tight integration of analysis and visualisation tools

maintenance are presented by Voinea et al.[120]. They used software visualisation, repository data mining, and static analysis tools to support assessments of software product quality attributes such as maintainability, modularity and complexity, and to elicit process patterns such as implicit developer networks and workflow structure. These techniques have been used on a variety of repositories such as open-source code bases with millions of lines of code, but also on commercial projects. Insight was collected from a variety of users including over sixty master students and over twenty professional software developers in the industry, over a period of several years. Usage patterns involved using tools during education, open-source development, longitudinal product development, and short assessment sessions. Evaluation results pointed strongly to the need of using simple visualisation metaphors, tight integration of analysis and visualisation tools within the accepted workflow by the targeted stakeholder group, and visualisation design, which closely reflects the concepts and values seen as important by the users.

The usage of visual analytics in software maintenance is a good illustration of the particular mix of heterogeneous data sources, combined analysis and visualisation, and the overall hypothesis forming, refinement, validation, and presentation of results for decision making support.

Illustrative example. An example of visual analytics for software product and process assessment (Voinea et al.[121]) is shown in Figure 8.3. In this application, an industrial automotive company developed a large software stack over a period of eight years. Towards the end of the process, it was seen that the product could not be completed within the time, and an assessment of the causes of the problem was required within one week in order to decide on the future of the project. Following static code and software evolution data mining (1), several hypotheses were generated for the problem causes, and several visualisations such as tree-maps, metric charts, compound graphs, and time-lines were used to assess the team and code evolution (2–4), quality evolution (5), and system architecture (6–7). The findings were combined and discussed with the stakeholders (i. e., project managers and team leaders), who obtained extremely valuable insights to guide their decision making.

The whole evaluation process needs to be considered from hypothesis forming to presentation of the results

This example outlines several interesting aspects related to visual analytics evaluation. The tools and techniques involved in analysis were used by a user group (the consultants), which was separate from the actual stakeholders (the project owners). Images and insights were presented by the tool users to the stakeholders, but result interpretation and decision making was left entirely to the latter group. The usage of simple business graphics, as opposed to more sophisticated visualisations, was seen as crucial for acceptance. As such, what was evaluated as successful was the entire process of hypothesis forming, refinement, validation, and presentation, rather than specific tool usability.

Collaborative visual analytics adds a further level of difficulty to its evaluation

Collaborative visual analytics. Visual analytics often involves a highly collaborative analysis process. Hence, evaluation plays an important role to determine how successful collaborative visual analytics systems can support the

reasoning processes in teams, something that is often difficult to evaluate in a controlled manner. Only few approaches have addressed this issue specifically, one of them presented by Isenberg and Fisher[63]. In their paper, the authors describe their Cambiera system that has dedicated support for awareness for collaborative visual analytics. In their informal evaluation of Cambiera, the authors presented two pairs of researchers with a data set in which they asked the participants to identify a story line. In particular, the authors paid attention to the use of awareness features provided by their tool and found that these were used by the participants in a variety of different ways.

8.2.3 Contests

One aim of evaluation is to find out what is the best approach to solve a certain problem. An interesting alternative way of evaluation is to compete: present a problem to the community and challenge researchers and developers to show that their solution is best. Competitions and contests have shown their value for advancing a field quickly.

Information retrieval has a long history of providing test collections

Early developments. In some research communities large scale, competitive evaluation efforts have a long history and developed into a central focus. For example, in text retrieval, large test collections are provided, aiming to encourage research and communication among industry, academia and governments. Different research areas are addressed with different tracks, which study issues such as text retrieval in blogs, legal documents, cross-language documents, Web pages etc.

Besides the development of text source material, it led to the development of standards for evaluation, e.g., the adoption of metrics like precision and recall.

An advantage in these cases is that the 'ground truth', i. e., what constitute good results, can be established objectively, even though generation of this ground truth often requires human experts. For exploratory data analysis and visualisation this is much harder to establish. A common format for contest in these communities is therefore to visualise a given data set and to report on findings.

The first graph drawing contest was in 1994

Graph drawing community. The graph drawing community focuses on the development of methods and techniques for producing diagrams of graphs that are aesthetically pleasing and follow layout conventions of an application domain. Annual contests have been held in conjunction with the Symposium of Graph Drawing since 1994. The categories have varied over the years, including free-style (all type of drawings for arbitrary datasets, judged on artistic merit and relevance), evolving graphs, interactive graph analysis, and social networks. Particularly interesting and exciting for participants is the on-site challenge format, where teams are presented a collection of graph data and have approximately one hour to submit their best drawings.

Information visualisation community. The information visualisation community started in 2003 with a contest at the yearly IEEE InfoVis Conference.

Information visualisation contests have run successfully since 2003

Catherine Plaisant, Jean-Daniel Fekete, and Georges Grinstein have been involved in the first three contests, and have given a thorough report on their experiences and lessons learned[87]. Participants were provided with a large dataset, and had typically four months to prepare their submissions, in the form of a two page summary, a video, and a Web page. Examples of datasets used are tree-structured data (from various sources), citation networks, multivariate data of technology companies, and data on Hollywood movies. Results were judged for the quality of the data analysis (what interesting insights have been found); quality of the techniques used (visual representation, interaction, flexibility); and quality of the written case study. Judges reviewed submissions, and they reported that this was difficult and time-consuming, as there was no ground truth available and because processes and results were difficult to compare. Participants in the contest worked hard and were motivated. Students could use the contest to test their PhD research and small companies reported that they appreciated the exposure.

Contest datasets are a valuable resource and should be made available in repositories

Some other recommendations given to organisers of contests are to facilitate student participation, to provide examples, and to use datasets with established ground truth. They emphasise that the contest is only a first step, and that the prepared datasets and the submissions are valuable material. They argue that an infrastructure is needed to facilitate the use of the datasets, leading to a repository of benchmarks. Also, given that the efforts required are above what can be expected from volunteers, they argue for support by funding agencies to plan and build up long term coordinated evaluation programs and infrastructures, to increase the impact of such contests.

Software visualisation community. Challenges involving evaluation of combined visualisation and analysis tools and methods are well established in software visualisation and several conferences have organised challenge tracks, specifically for software visualisation techniques.

Datasets are prepared and offered to participants for investigation several months in advance, with a number of questions being asked. Questions and tasks range from generic software understanding, such as investigating the evolution of large-scale software repositories in order to identify trends of interest for typical software maintenance tasks, up to precisely defined tasks such as assessing the modularity or presence of certain design and coding patterns for a given software system. Participants typically use visualisation and analysis tools of their own design to answer these questions. Contest entries are typically published as short papers in the conference proceedings.

Apart from challenges focusing on pre-selected datasets, software visualisation conferences also encourage the submission of short tool demonstration papers. In contrast to challenges, which focus on the insight gained when analysing a given data set, tool demo papers focus on a more general set of aspects that make a tool efficient and effective, such as scalability, genericity, integration with accepted workflows, and ease of use.

Challenges and tool demo contests in software visualisation share a number of particular aspects. Several de facto standard datasets have emerged from the

research community, such as the Mozilla Firefox, KDE, and ArgoUML code bases. Compared to some other sub-domains of visual analytics, generation and acquisition of realistic, challenging, data is not seen as a problem in software visualisation. Many open source repositories exist, which contain large and complex systems. These repositories cover a wide range of aspects, such as long-term evolving code, multiple designs, architecture, programming languages and patterns, and access to specific questions and challenges of the developers, present in design documents and commit logs. Furthermore, tools, technologies and data interchange formats are relatively well standardised across the field.

Many open source repositories exist within the software visualisation domain

Visual analytics community. In 2006 a highly relevant contest emerged: the VAST Contest[89], renamed to VAST Challenge in 2008, held in conjunction with the Visual Analytics Software and Technology symposium.

VAST Challenge

In several respects, the challenges in visual analytics contests are close to perfect. The data provided are large. Each year a new challenge is addressed, typically with a security or intelligence aspect. Several different datasets are provided for a challenge, each giving different cues and different aspects. For instance, the 2008 challenge scenario concerned a fictitious, controversial socio-political movement; and the data consisted of cell phone records, a chronicle of boat journeys with passenger lists, a catalogue of Wiki edits, and geo-spatial data of an evacuation after a bomb attack. The datasets are carefully generated by the National Visualisation and Analytics Center (NVAC) Threat Stream Generator project team at PNNL, and a ground truth (as well as false trails) is hidden in these.

High quality, complex data is at the heart of the successful VAST Challenge

In many respects, the VAST Challenge is highly successful. It encourages and stimulates researchers and students, and it has led to a repository of large heterogeneous datasets with ground truths. These datasets are used now by researchers to test new methods, but also in education. A large part of its success can be attributed to the high quality and high motivation of the organisers. Another important ingredient is the support by government agencies (NIST and PNNL), especially for constructing the datasets and for judging the results.

8.3 Next Steps

Overall, visual analytics is a highly promising concept for increasing the effectiveness of obtaining new insights. It helps solving actual data-related problems in many application fields where multivariate, highly dimensional, and complex datasets are involved. However, there are several challenges to the adoption of visual analytics in actual application areas, and evaluation is a crucial one of these. In the preceding sections, we have discussed why evaluation is highly important, why evaluation is hard in visual analytics, and that despite the efforts so far, there is a lack of solid findings. We now put forward recommendations to improve this situation, which we hope

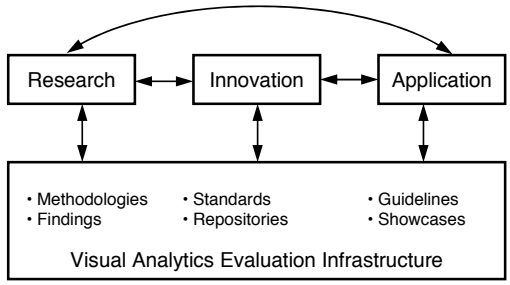

Figure 8.4: Overview of stakeholders and recommendations

will ultimately lead to successful adoption of visual analytics on a large scale.

We formulate our recommendations by urging the need for a solid evaluation infrastructure for visual analytics, consisting of a number of components. An overview of this evaluation infrastructure is shown in Figure 8.4. The main categories of stakeholders are researchers, innovators (translating new ideas into useful tools), and, most importantly, the users of these tools. To enable the latter to take advantage of the opportunities offered by visual analytics, various aspects of the innovation process have to be strengthened, leading to the following set of recommendations.

Stimulate research on evaluation methodologies for visual analytics. In the preceding sections, we have seen that visual analytics methods can be evaluated in a variety of ways, ranging from informal user studies to studies of the adoption in practise. All these have their own strengths and weaknesses. Given the complexity of the topic, it is highly unlikely that the current approaches used are already fully developed, and we are convinced that there is much room for improvement. A promising area is, for instance, the use of eye-tracking techniques and measurement of physical signals in general, as these give a wealth of information on what the user is actually doing, although the interpretation of this information is very hard. Instrumentation of software gives another detailed view on the actions of users, but here again correct interpretation is difficult. Experiments with professional, overburdened users is notoriously difficult, and often students are taken as replacements. The extent to which the results obtained through the use of non-experts can be translated to professional use is unclear. In short, we expect that methodologies for evaluation can be improved significantly, and that this will provide means to obtain more insight with less costs into the quality of visual analytics artefacts.

Stimulate evaluation of visual analytics methods and techniques. Despite enthusiastic efforts of the research community in visual analytics, there is still a great lack of solid results on the quality and scope of visual analytics artefacts. Given the novelty, size and complexity of the field, this is very understandable, but significant effort is required to improve this situation. There is a strong awareness now that evaluation is important, and in many research projects much

effort is already expended on evaluation. To stimulate this further, new research programs should emphasise and encourage evaluation, including efforts aimed at evaluating existing methods.

A particularly important driver for evaluation is setting the right frame of reference. In many cases, visualisation tools and techniques, which have successfully passed evaluations related to ease of use, interaction, and perception, are still not adopted by practitioners in the field. One major reason for this is that such tools are not seen as bringing essential value for their targeted users. This situation is even more important for visual analytics technologies which, by definition, combine many aspects, data sources and tools. As such, one promising direction is to evaluate the effectiveness of an entire visual analytics solution, or workflow, rather than to focus only on its separate components. However, to summarise the quality in a simple measure (e.g., time on task) is too simplistic, and leads to a 'credit assignment' problem (e.g., which features of the systems helped or not). This is crucial for formative evaluation. Ideally we should evaluate a complete workflow, but with carefully designed measurements that are able to tap into parts of the process, including interpretative feedback from users, maybe after the event to prevent interference effects.

Stimulate standardisation. Standardisation is a vital aspect in building up a body of knowledge. Visual analytics subsumes a wide variety of types of data, techniques, applications and users. In order to make evaluation results comparable and retrievable, standard definitions and taxonomies of all these aspects are important. In other fields, such as statistics, techniques can be accurately classified based on the characteristics of the data analysed, and the results have a clear and precise meaning. In visual analytics, such a precision cannot be attained, almost by definition. A fundamental assumption is that the data to be analysed is so complex that the human in the loop, with their own great strengths but also their complexity and variety, is essential. However, this does not mean that we should not try to reach higher levels. For example, the use of standard measures and tests to assess the perceptual and cognitive skills of participants would be a good step forward. In general, standardisation enables researchers, innovators, and users to exchange results, where the meaning of the various aspects is as clear and unambiguous as possible. We think that efforts in this direction are important and should be stimulated, as it helps to build a foundation and infrastructure that many will benefit from.

Stimulate repositories. Standardisation is one aspect to enable and stimulate exchange of results; the development of repositories is its natural complement. Evaluation of visual analytics can be performed much more effectively and efficiently if central repositories are set up and maintained that provide relevant material. Such repositories could provide:

- Datasets for a variety of applications and at various levels of detail. Preferably they should also include information on results to be found, that can act as a ground truth.

- Data generation tools to generate benchmark datasets, again of many different types and ranges of complexity, where the information to be found can be inserted on request.
- Analysis tools and libraries to perform automated analysis and evaluation, whenever possible.
- Standardised questionnaires to assess users experience of the artefacts tested;
- Detailed results of previous evaluation studies, to enable further analysis and comparison.

In certain fields, such as software visualisation, the emergence of lively open-source communities provides a good, low-cost, solution. Datasets such as software repositories are the prime vehicle of information interchange in such fields, and are open to everyone for examination. Given the focus on software technologies, this field also sees a strong development and sharing of analysis and visualisation tools, and strong interaction between researchers, industry practitioners and individual developers.

Collect showcases. For the adoption of visual analytics technology, the outside world has to become more aware of its possibilities and advantages. Potential users, include software and system developers, which could take advantage of integration of visual analytics technology in their products, as well as end-users. Collection and dissemination of showcases, including successful evaluation approaches, is important in this respect. Such showcases provide an overview of the possibilities, and should clearly show the benefits, in terms of novel and valuable insights obtained as well as reduced costs for data analysis. Here, we can exploit the particularities of each field to stimulate dissemination and create awareness.

Stimulate development of guidelines. Potential users need guidance on what technology to adopt, how to apply it in order to solve their problems, and how to evaluate the effectiveness. Development of guidelines, tutorials, handbooks deserves attention. These should be useful and understandable for the target audience, and be grounded in results from the scientific community as well as real world practise and experience.

9 Recommendations

Governments, businesses, research institutes, and people themselves are collecting and hoarding large amounts of data, which are presently often not utilised in the best possible way for solving the world's pressing problems. We need better and more usable solutions to extract information and ultimately knowledge from these rich data resources. Our ultimate goal as a research community is to provide visual analytics methodologies, tools, and infrastructure that will benefit society in general. The international research initiatives in the area of visual analytics including the European VisMaster project have acted as a catalyst in instigating better collaboration between leading institutes and universities working on various aspects of visual analytics. The successful collaboration and interaction of different communities enables us for the first time to identify common challenges and problems across many disciplines. This book is a stepping stone towards our goal and lays down the path to a shared solution.

Each of the chapters has described the specific challenges and opportunities within the specific domain. This chapter summarises the challenges according to what we believe to be the main entities of visual analytics and then consolidates the recommendations presented in the individual chapters into higher level recommendations for enabling successful visual analytics research. The challenges and recommendations highlight the interdisciplinary nature of visual analytics and the importance of working together. While many initiatives have been started in different countries including many EU member states, only under a worldwide and EU-wide umbrella can significant overlap be avoided, and continued strong collaboration between the research groups be fostered.

9.1 The Challenges

Visual analytics is concerned with data, users, and designing a technology that enables the user to make sense of the data in order to extract information and augment their knowledge. Each of the chapters in this book has identified challenges associated with visual analytics with respect to its particular domain, but many challenges are common to more than one domain.

This section presents a summary of the challenges, organised into four entities:

- **Data:** the challenge of dealing with very large, diverse, variable quality datasets.
- **Users:** the challenge of meeting the needs of the users.

- **Design:** the challenge of assisting designers of visual analytic systems.
- **Technology:** the challenge of providing the necessary infrastructure.

9.1.1 Data

An obvious challenge is dealing with very large datasets, whether this is in terms of storage, retrieval, transmission (as with distributed databases or Cloud storage), algorithm processing time, and scalability of visualisations. It is also apparent that many analytic applications use in-memory storage rather than a database approach, as traditional databases cannot meet the challenging functionality required by visual analytics.

Data is often heterogeneous and can be of poor quality with, missing, incomplete, or erroneous values. This adds to the complexity of integrating data from many sources. In addition, data often requires transformation of some sort (e.g., scaling and mapping) or requires specialised data types, which are seldom provided by current database systems.

Streaming data presents many challenges – coping with very large amounts of data arriving in bursts or continuously (as with analysing financial transactions or Internet traffic), tackling the difficulties of indexing and aggregation in real-time, identifying trends and detecting unexpected behaviour when the dataset is changing dynamically.

Semantic management (managing metadata) is currently not well catered for, which is surprising, given the wealth of information contained in rich metadata. In addition, we can also add further meaningful information gathered during the analysis and visualisation phases.

9.1.2 Users

There are many challenges related to system usability and process understanding. For users to have confidence in the data they should be aware (or be able to discover) where the data comes from, and also what transformations have been applied on its way through the process pipeline (e.g., data cleansing, analysis and visualisation). Furthermore, a clear understanding of the uncertainties in the data and results of the analysis can help minimise cognitive and perceptual biases, which without attention can significantly affect the interpretation of the results.

Another challenging aspect is using visual analytics to simplify the models and patterns extracted by advanced data mining techniques, so called 'visual data mining'. Existing methods are largely non-intuitive and require significant expertise. Similar efforts are required to assist users in understanding visualisation models, such as the level of abstraction (actual data or aggregated view) and visual metaphors (how the data is represented on the screen). Expert analysts require this flexibility and so do the more naive users, who in addition, require guidance in, for instance, choosing appropriate analysis tools and visualisation methods for the task at hand. Users often wish or need to collaborate in order to

share, or work cooperatively on, the data, results of analysis, visualisations and perhaps workflows. Providing the necessary distribution infrastructure as well as the user interface is a challenging task.

The degree of interactivity is important for all users. Rapid feedback is critical in visual interfaces and this presents challenges to many of the domains associated with visual analytics. Evaluating visual analytics applications is particularly difficult due to the complexity of human interaction with multiple processes (e.g., analysis and visualisation). The question of how to classify success or decide what is a good solution is problematic when dealing with exploratory tasks, which are typically ill-defined or open-ended.

9.1.3 Design

One of main challenges is to utilise our existing theoretical and practical knowledge by making it readily available to designers of visual analytics systems, possibly in the form of design guidelines. For instance, there is a wealth of experimental results in the field of visual perception and cognition that would be of considerable benefit to interaction designers, if it were organised appropriately. In general, we have a host of technology, but for a given task, the challenge is to provide guidance on what to use (e.g., method of analysis, type of visualisation), how to use it and how to decide if it was a good choice. We need to find ways of making appropriate test datasets, tools, and results of evaluation studies available to the community.

Designing and implementing visual analytics applications would be faster and potentially more reliable and flexible, if a unified architectural model was used. Designing a suitable component-based framework is certainly a challenge.

9.1.4 Technology

Various challenges have been identified regarding more technical aspects of visual analytics. One is in relation to the duration of the analysis phase, which tends to be much longer than traditional transactions dealt with by a standard database management system. Therefore, methods are required not only to support long commit phases, but also to furnish partial results from the analysis. Providing this 'progressive analysis' would give the analyst a rapid overview and hence, a basis for steering the analysis in a particular direction, from which details could be sought. This interactive functionality requires notification services – current database management systems utilise a trigger mechanism that is not suitable for visual analytics, especially when the trigger for recalculation comes from the visualisation sub-system rather than from the analysis.

Providing multi-scale analysis is a particular challenge identified by the geospatial visualisation community. For example, the analyst may wish to look for

patterns over a long period of time, months or even years, as well as patterns in daily activity.

Finally, we need to devise a methodology for providing basic visualisation functions, such as linking and brushing, which can be used to coordinate different views of the data. These connection mechanisms ought to be incorporated in the next generation of lightweight web-based visualisation tools, in order to facilitate the analysis of data to a wider audience.

9.2 Meeting the Challenges

Our recommendations consolidate the recommendations from the individual chapters (under the 'Next Steps' section) and put forward higher level recommendations for enabling successful visual analytics research under the following areas:

- International, European, and national funding agencies
- Visual analytics research community
- Broader research community
- Industry and other potential users

9.2.1 International, European, and National Funding Agencies

These recommendations are aimed at the funding agencies throughout the world, including the US and Canada as well as the EC and the national funding agencies. Some EU countries have already started initiatives in this direction.

Recommendation 1: Need for Continuing Support

National and international agencies should endeavour to continue support for visual analytics

VisMaster considers it of great importance that a parallel effort, both at the international and national level, is extended. The national funding agencies have the opportunity to step in earlier (while more time is needed at the international level) and provide additional support for visual analytics research. The umbrella, under which the various efforts in the EU can be integrated, has to be further entrenched by the EC. Within the FET area of ICT, this could be established through a FET Proactive Topic that funds several projects in visual analytics, implementing the vision that this research roadmap lays down. Such open basic research schemes or comparable national instruments will support pilot projects in visual analytics as a next step.

Recommendation 2: Appreciate Visual Analytics in Related Research Areas

The European Commission, as well as the main national funding agencies, should assess their existing programs and identify possibilities to acknowledge the consideration of visual analytics questions in related interdisciplinary research projects. Visual analytics is not just a 'user interface' for research prototypes; it is a different approach all together, fully integrating the human expertise in the human-machine dialogue. We recommend that funding agencies and their units start to think in new ways about an interdisciplinary integration of visual analytics components in the context of existing programmes. A first example can be seen in the Objective ICT-2011.4.4 Intelligent Information Management, target outcome b).

Consider integrating visual analytics into existing research projects

Recommendation 3: Foster Interdisciplinary Research Projects

The fastest path to new solutions to highly complex problems is the promotion of interdisciplinary research projects where heterogeneous teams address real-world problems and develop visual analytics methods and tools that help in investigating and solving these problems.

Promote interdisciplinary research projects, addressing real-world problems

Interdisciplinary PhD programs and summer schools that support young researchers in travelling to other universities and research institutions are a perfect way of exchanging knowledge about complementary disciplines and doing transdisciplinary research.

Recommendation 4: Avoid Redundant Development

Visual analytics, not only its visualisation part, is relevant to many areas of current research projects. The development and sharing of open-source analysis and visualisation tools should be fostered by and for the visual analytics community. Building and publicising repositories of visual analytics techniques will enable the rapid implementation of effective and efficient visual analytic solutions and help to avoid costly redundant development effort.

Develop repositories of visual analytic, tools, techniques, guidelines and datasets to accelerate the implementation of appropriate applications

Potential users and other research areas need guidelines on what visual analytics technology to adopt, how to apply it in order to solve their problems, and how to evaluate their effectiveness. Guidelines, tutorials, handbooks, etc., which are useful and understandable have to be developed for the targeted audience, and be grounded in results from the scientific community as well as real world practice and experience.

The evaluation of visual analytic techniques, including software tools, models and theories is very important for the continued growth of the discipline. This can be performed much more effectively and efficiently if central repositories are set up and maintained to provide relevant material. Such repositories

Promote the evaluation of all aspects of visual analytic solutions

could provide datasets at various levels of detail for a variety of applications, together with data generation tools, analysis tools, results of evaluations and libraries.

9.2.2 Visual Analytics Community

The following recommendations address the visual analytics research community. Most, if not all of the research institutions active in visual analytics within the EU have been involved in VisMaster. Through this collaboration, the project identified recommendations for each institute that should help the community to strengthen visual analytics research throughout and outside the EU.

Recommendation 5: Spread the Word Within Academia

Continue and expand upon interdisciplinary workshops and support visual analytics within education

Visual analytics technologies should be presented at data management, data mining and human-computer interaction conferences. This can be achieved by organising special sessions, panel discussions and dedicated workshops or by publishing related success stories. In addition, educational curricula ought to be updated to reflect the increasing prominence of visual analytics, and educators should engage in related interdisciplinary teaching efforts.

Interdisciplinary scientific workshops like VAKD (in conjunction with ICDM), INTERACT-VA, or GeoVA(T) should be continued as they give researchers from different disciplines, the opportunity to discuss the research problems in visual analytics and establish contacts for further interdisciplinary cooperation in solving these problems. The working groups of VisMaster have been successful in organising such workshops, drawing together different communities.

Support the annual EuroVA Symposium

VisMaster successfully organised the EuroVAST symposium on visual analytics in June 2010 in Bordeaux, France, which attracted over 70 scientists and researchers from all over the world. EuroVAST 2010 will not be a singular event; it will continue as the EuroVA Symposium in cooperation with EuroVis, with the next taking place in 2011 in Bergen, Norway. This symposium should be established as the European event where researchers can submit their visual analytics research work, in addition to IEEE VAST in the US.

Collaboration should be continued with agencies outside the EU

Further collaboration should be continued on an international level beyond the EU with NVAC, the Purdue Center of Excellence and the DHS (all in the US), as well as other initiatives in Canada, Australia and elsewhere.

Recommendation 6: Spread the Word Within Industry and Governmental Agencies

Publicise work of the visual analytics community in dealing with real-world problems

The VisMaster Industry Day has shown the existing interest in visual analytics from areas such as business intelligence, finance, security, and media technologies. The visual analytics community should publicise their work within both their industry and application domains, and show the hands-on benefits of using visual analytics technologies to solve some of the most pressing problems facing industry.

The effective use of visual analytics in exploring government data should be encouraged

Politicians and policy makers should use visual analytics to explore data collected by their statistical departments in order to increase their understanding of the needs and desires of their constituencies. In addition, their constituents could evaluate their policy decisions, thus helping to strengthen and legitimise democratic institutions.

Demonstrate the possibilities and benefits of visual analytics to a wide audience

The outside world has to be made more aware of the possibilities and advantages of visual analytics. The collection and dissemination of showcases, including successful evaluation approaches, is important in demonstrating the possibilities and the benefits of visual analytics. Potential users and decision makers must easily recognise the novel and valuable insights, which only visual analytics can enable, as well as the reduced costs of obtaining the insights.

Recommendation 7: Build a Visual Analytics Infrastructure

Stimulate the design of conceptual and software architectures for visual analytics

Researchers in databases, analytics, visualisation and communication should be given incentives to work together to iterate on the design of a conceptual architecture for visual analytics applications. This will avoid an explosion of partial solutions to the overall infrastructure problem with issues of interoperability and also situations in which specialists in one domain implement another domain's modules. Applied-research projects are required to design and experiment with software architectures, again to promote interoperability and compatibility across domains, and thought needs to be given to managing and promoting the specifications of the resulting architectures at an international level.

Recommendation 8: Understand and Reach Out to Users

Encourage research on evaluation methods for visual analytics

Research on evaluation methodologies for visual analytics should be encouraged. We expect that methodologies for evaluation can be improved significantly, and this will provide a means to obtain more insight with visual analytics technologies. In addition, scientifically verified best practices will reduce the cost of implementing high-quality tools and provide convincing arguments for their adoption.

The understanding of human perceptual and cognitive processes in dealing with spatial and temporal information should be improved as well as the

Develop design guidelines based on visual perception and cognition research

understanding of visual displays of and interaction with such information. On this basis, appropriate design rules and guidelines should be developed for interactive displays of information, with particular focus on non-expert users of visual analytics systems.

Support the visual analytics process for personal as well as professional analysts

Furthermore, a new generation of lightweight accessible dynamic visual analytics tools should be developed to support a range of personal and professional analysts in the best possible way. Effective solutions for training both specialist and non-specialist users interested in adopting these visual analytics tools are also necessary.

9.2.3 Broader Research Community

Visual analytics is a highly interdisciplinary topic. While VisMaster approached specific related research communities to start or enhance collaboration, there are more topics and fields of research that relate to visual analytics. Certain communities feel that there should be an integration of research efforts, rather than just an exchange of views.

Recommendation 9: Integrate Visual Analytics and Related Research Areas

Build a new integrated research community

Our investigation and analysis show that there is a widely recognised need for an integration of visual analytics and related research areas for building a new integrated research community. This is especially true for the area of knowledge discovery and data mining (KDD). For many years, KDD and visualisation research have been complementary. With the rise of the field of visual analytics, the need to integrate these two disciplines becomes more apparent. The experience from the IEEE VAST conference and the VAKD workshop held at SIGKDD suggest that we should bring the integration of these topics to a new level to enable its full potential.

Recommendation 10: Organise Interdisciplinary Events

Organise interdisciplinary events

The VisMaster project has demonstrated the success of jointly organising and implementing interdisciplinary events in enabling researchers to discuss common problems and establish future contacts. Such events are more beneficial if one finds mediators that can translate between communities, as achieved at the INTERACT workshop by inviting experts from areas of cognitive science, HCI, and visualisation. Invited lectures and seminars by prominent researchers in universities and research organisations are also ideal for the dissemination of knowledge and experience.

9.2.4 Industry and Potential Users of Visual Analytics Technology

The following recommendation targets the industry users and all other potential users of visual analytics technology. Through many research projects and presentations, the VisMaster partners saw that industry, policy makers, governments, social scientists, and individuals from many other areas are in need of visual analytics tools. Many of them highly appreciate visual analytics solutions since they recognise the limits of purely analytical or purely visual solutions. We invite our potential users to be open to new approaches and to help us understand their needs.

Recommendation 11: Evaluate and Express the Need for Visual Analytics

Industry users should explore possible uses of visual analytics

Drawing on the examples of successful prototypes, industrial users should evaluate their potential need for visual analytics technology, possibly with the assistance of knowledgeable researchers and developers, and communicate this need through the appropriate channels. It is advisable to focus on small, incremental improvements and use case studies of successful pilot projects to generate broader support for and acceptance of visual analytics in the organisation.

9.3 Future Directions

An interesting observation is that all grand challenge problems of the 21th century, such as the climate change, energy, financial, health or security crisis, require the exploration and analysis of very large and complex data sets which can neither be done by the computer nor the human alone. Making scientific discoveries and solving complex problems require a tight integration of human intelligence and intuition with the storage and processing power of today's computers. Visual analytics will therefore likely develop into a general science of problem solving and interactive discovery. It will change the way we approach large complex datasets and the unsolved grand challenge problems associated with them. This research roadmap is designed to pave the way for visual analytics to become such a tool of scientific discovery. It is also designed to help the readers to recognise the potential of visual analytics and encourage them to engage in the next steps towards this goal.

Bibliography

[1] J. Abello, F. van Ham, and N. Krishnan. ASK-GraphView: A large scale graph visualization system. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):669–676, 2006.

[2] C Ahlberg, C Williamson, and B Shneiderman. Dynamic queries for information exploration: an implementation and evaluation. In *CHI '92: Proceedings of the SIGCHI conference on human factors in computing systems*, pages 619–626. ACM Press, 1992.

[3] W. Aigner, S. Miksch, W. Müller, H. Schumann, and C. Tominski. Visual methods for analyzing time-oriented data. *IEEE Transactions on Visualization and Computer Graphics*, 14(1):47–60, 2008.

[4] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005*, pages 111–117, 2005.

[5] G. Andrienko, N. Andrienko, P. Jankowski, D. A. Keim, M. J. Kraak, A. MacEachren, and S. Wrobel. Geovisual analytics for spatial decision support: Setting the research agenda. *International Journal of Geographical Information Science*, 21(8):839–858, 2007.

[6] G. Andrienko, N. Andrienko, S. Rinzivillo, M. Nanni, D. Pedreschi, and F. Giannotti. Interactive visual clustering of large collections of trajectories. In *IEEE Symposium on Visual Analytics Science and Technology*, pages 3–10, 2009.

[7] N. Andrienko and G. Andrienko. *Exploratory Analysis of Spatial and Temporal Data: a Systematic Approach*. Springer, 2006.

[8] L. Anselin. What is special about spatial data? alternative perspectives on spatial data analysis. Technical Report 89-4, National Center for Geographic Information and Analysis, 1989.

[9] A. Aris and B. Schneiderman. A node aggregation strategy to reduce complexity of network visualization using semantic substrates. Technical Report HCIL-2008-10, Human-Computer Interaction Lab, University of Maryland, 2008.

[10] P. Bak, I. Omer, and T. Schreck. Visual analytics of urban environments using high-resolution geographic data. In M. Painho, M.Y. Santos, and H. Pundt, editors, *Geospatial Thinking, Lecture Notes in Geoinformation and Cartography*. Springer, 2010.

[11] D. Barbará, W. DuMouchel, C. Faloutsos, P. J. Haas, J. M. Hellerstein, Y. E. Ioannidis, H. V. Jagadish, T. Johnson, R. T. Ng, V. Poosala, K. A. Ross, and K. C. Sevcik. The New Jersey data reduction report. *IEEE Data Eng. Bull.*, 20(4):3–45, 1997.

[12] J. C. Barrett, B. Fry, J. Maller, and M. J. Daly. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, 21(2):263–265, 2005.

[13] BELIV. Beyond Time and Errors: Novel Evaluation Methods for Information Visualization, ACM Press.

[14] E. Bertini and D. Lalanne. Investigating and reflecting on the integration of automatic data analysis and visualization in knowledge discovery. *SIGKDD Explorations*, 11(2):9–18, May 2010.

[15] E. Bertini and G. Santucci. Is it darker? improving density representation in 2D scatter plots through a user study. In *Conference on Visualization and Data Analysis*, volume 5669, pages 158–167, 2005.

[16] E. Bertini and G. Santucci. Give chance a chance: modeling density to enhance scatter plot quality through random data sampling. *Information Visualization*, 5(2):95–110, 2006.

[17] E. Bertini and G. Santucci. Visual quality metrics. In *BELIV '06: Proceedings of the 2006 AVI workshop on beyond time and errors*, pages 1–5. ACM, 2006.

[18] P. A. Boncz, M. L. Kersten, and S. Manegold. Breaking the memory wall in MonetDB. *Commun. ACM*, 51(12):77–85, 2008.

[19] J. Boulos, N. Dalvi, B. Mandhani, S. Mathur, C. Re, and D. Suciu. MYSTIQ: a system for finding more answers by using probabilities. In *SIGMOD '05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 891–893. ACM, 2005.

[20] P. Buono, A. Aris, C. Plaisant, A. Khella, and B. Shneiderman. Interactive pattern search in time series. In *Proceedings of SPIE*, volume 5669, pages 175–186, 2005.

[21] C. M. Burns and J. R. Hajdukiewicz. *Ecological interface design*. CRC, 2004.

[22] T. Butkiewicz, W. Dou, Z. Wartell, W. Ribarsky, and R. Chang. Multi-focused geospatial analysis using probes. *IEEE Transactions on Visualization and Computer Graphics*, 14:1165–1172, 2008.

[23] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, and R. Rosati. Conceptual modeling for data integration. In *Conceptual Modeling: Foundations and Applications*, pages 173–197, 2009.

[24] S. K. Card and J. Mackinlay. The structure of the information visualization design space. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis '97)*, pages 92–99, 1997.

[25] S. K. Card, J. D. Mackinlay, and B. Shneiderman. *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999.

[26] S. Carpendale. Evaluating information visualization. In A. Kerren, J. Stasko, J. Fekete, and C. North, editors, *Information Visualization: Human-Centered Issues and Perspectives*, pages 19–45. Springer, 2008.

[27] C. Chen. *Information Visualization - Beyond the Horizon*. Springer, 2004.

[28] C. Chen and Y. Yu. Empirical studies of information visualization: a meta-analysis. *Int. J. Hum.-Comput. Stud.*, 53(5):851–866, 2000.

[29] E. H. Chi and J. Riedl. An operator interaction framework for visualization systems. In *INFOVIS '98: Proceedings of the 1998 IEEE Symposium on Information Visualization*, pages 63–70, 1998.

[30] L. Chittaro, C. Combi, and G. Trapasso. Data mining on temporal data: a visual approach and its clinical application to hemodialysis. *Journal of Visual Languages & Computing*, 14(6):591–620, 2003.

[31] E. F. Codd, S. B. Codd, and C. T. Salley. Providing OLAP (On-Line Analytical Processing) to user-analysis: An IT mandate, 1993.

[32] L. Colgan, R. Spence, and P. Rankin. The cockpit metaphor. *Behaviour & Information Technology*, 14(4):251–263, 1995.

[33] V. Coltheart. *Fleeting memories: Cognition of brief visual stimuli*. MIT Press, 1999.

[34] O. De Bruijn and R. Spence. A new framework for theory-based interaction design applied to serendipitous information retrieval. *ACM Trans. Comput.-Hum. Interact.*, 15(1):1–38, 2008.

[35] M. C. Ferreira de Oliveira and H. Levkowitz. From visual data exploration to visual data mining: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 9(3):378–394, 2003.

[36] A. Dix, J. Finlay, and G. D. Abowd. *Human-computer interaction*. Prentice Hall, 3rd edition, 2004.

[37] S. R. dos Santos and K. W. Brodlie. Gaining understanding of multivariate and multidimensional data through visualization. *Computers & Graphics*, 23(1):311–325, June 2004.

[38] J. Dykes, A. M. MacEachren, and M-J. Kraak, editors. *Exploring Geovisualization*. Elsevier, 2005.

[39] S. G. Eick. Visualizing multi-dimensional data. *SIGGRAPH Comput. Graph.*, 34(1):61–67, 2000.

[40] S. G. Eick, J. Mauger, and A. Ratner. Visualizing the performance of computational linguistics algorithms. In *2006 IEEE Symposium On Visual Analytics Science And Technology*, pages 151–157, 2006.

[41] G. Ellis and A. Dix. A taxonomy of clutter reduction for information visualisation. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1216–1223, 2007.

[42] G. Ellis, J. Finlay, and A. Pollitt. HIBROWSE for hotels: bridging the gap between user and system views of a database. In *IDS'94 Workshop on User Interfaces to Databases*, pages 45–58. Springer, 1994.

[43] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios. Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1):1–16, January 2007.

[44] R. Elmasri and S. B. Navathe. *Fundamentals of Database Systems*. Addison-Wesley, 2007.

[45] W. Fikkert, M. D'Ambros, T. Bierz, and T. Jankun-Kelly. Interacting with visualizations. *Human-Centered Visualization Environments*, pages 77–162, 2007.

[46] B. R. Gaines. Modeling and forecasting the information sciences. *Inf. Sci.*, 57-58:3–22, 1991.

[47] F. Giannotti and D. Pedreschi. *Mobility, Data Mining and Privacy: Geographic Knowledge Discovery*. Springer, 2008.

[48] J. Gibson. *The Ecological Approach to Visual Perception*. Lawrence Erlbaum Associates, 2nd (reprint) edition, 1986.

[49] E. W. Gilbert. Pioneer maps of health and disease in England. *Geographical Journal*, 124(2):172–183, 1958.

[50] G. Grinstein, C. Plaisant, S. Laskowski, T. O'Connell, J. Scholtz, and M. Whiting. VAST 2008 challenge: Introducing mini-challenges. In *IEEE Symposium on Visual Analytics Science and Technology (VAST '08)*, pages 195–196, 2008.

[51] D. Guo, J. Chen, A. M. MacEachren, and K. Liao. A visualization system for space-time and multivariate patterns (vis-stamp). *IEEE transactions on visualization and computer graphics*, pages 1461–1474, 2006.

[52] R. B. Haber and D. A. McNabb. Visualization idioms: A conceptual model for scientific visualization systems. In B. Shriver, G. M. Nielson, and L. J. Rosenblum, editors, *Visualization in Scientific Computing*, pages 74–93. IEEE, 1990.

[53] T. Hägerstrand. What about people in regional science? *Papers of the Regional Science Association*, 24:7–21, 1970.

[54] J. Han and M. Kamber. *Data mining: concepts and techniques*. Morgan Kaufmann, 2006.

[55] D. J. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. The MIT Press, 2001.

[56] C. D. Hansen and C. R. Johnson. *The visualization handbook*. Academic Press, 2005.

[57] J. Heer, S. K. Card, and J. A. Landay. Prefuse: a toolkit for interactive information visualization. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 421–430. ACM, 2005.

[58] J. Hollan, E. Hutchins, and D. Kirsh. Distributed cognition: Toward a new foundation for human-computer interaction research. *ACM Trans. Comput.-Hum. Interact.*, 7(2):174–196, 2000.

[59] K. J. Holyoak and P. Thagard. *Mental leaps: Analogy in creative thought*. MIT Press, 1996.

[60] G. J. Hunter and M. F. Goodchild. Managing uncertainty in spatial databases: Putting theory into practice. *Journal of Urban and Regional Information Systems Association*, 5(2):55–62, 1993.

[61] E. Hutchins. *Cognition in the Wild*. MIT Press, 1994.

[62] W. H. Inmon, editor. *Building the Data Warehouse*. Wiley, 2002.

[63] P. Isenberg and D. Fisher. Collaborative brushing and linking for co-located visual analytics of document collections. *Computer Graphics Forum*, 28(3):1031–1038, June 2009.

[64] D. A. Keim. Visual exploration of large data sets. *Communications of the ACM (CACM)*, 44(8):38–44, 2001.

[65] D. A. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler. Challenges in visual data analysis. In *Information Visualization (IV 2006), Invited Paper, July 5-7, London, United Kingdom*. IEEE Press, 2006.

[66] D. A. Keim, F. Mansmann, and J. Thomas. Visual analytics: How much visualization and how much analytics? *SIGKDD Explorations*, 11(2):5–8, December 2009.

[67] D. A. Keim and J. Thomas. Scope and challenges of visual analytics, 2007. Tutorial at IEEE Visualization, http://infovis.uni-konstanz.de/tutorials/.

[68] G. Klein. *Sources of Power: How People Make Decisions*. MIT Press, Feb 1999.

[69] M. J. Kraak and F. Ormeling. *Cartography: visualization of geospatial data*. Pearson Education, 2003.

[70] J. Kruger, J. Schneider, and R. Westermann. Clearview: An interactive context preserving hotspot visualization technique. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):941–948, 2006.

[71] J. Lin, E. Keogh, S. Lonardi, J. P. Lankford, and D. M. Nystrom. VizTree: a tool for visually mining and monitoring massive time series databases. In *VLDB '04: Proceedings of the Thirtieth international conference on Very large data bases*, pages 1269–1272. VLDB Endowment, 2004.

[72] Z. Liu, N. Nersessian, and J. Stasko. Distributed cognition as a theoretical framework for information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1173–1180, 2008.

[73] A. M. MacEachren. *How maps work*. Guilford Press New York, 1995.

[74] A. M. MacEachren and I. Brewer. Developing a conceptual framework for visually-enabled geocollaboration. *International Journal of Geographical Information Science*, 18(1):1–34, 2004.

[75] O. Z. Maimon and L. Rokach. *Data mining and knowledge discovery handbook*. Springer New York, Inc., 2005.

[76] F. Mansmann, F. Fischer, S. C. North, and D. A. Keim. Visual support for analyzing network traffic and intrusion detection events using treemap and graph representations. In *CHiMiT '09: Proceedings of the Symposium on Computer Human Interaction for the Management of Information Technology*, pages 19–28. ACM, 2009.

[77] R. E. Mayer. The search for insight: Grappling with Gestalt psychology's unanswered questions. *The nature of insight*, pages 3–32, 1995.

[78] J. E. McGrath. Methodology matters: Doing research in the social and behavioural sciences. In *Readings in Human-Computer Interaction: Toward the Year 2000*. Morgan Kaufmann, 1995.

[79] H. J. Miller and J. Han. *Geographic data mining and knowledge discovery*. CRC Press, 2001.

[80] T. Munzner. A nested process model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928, November/December 2009.

[81] U. Neisser. *Cognition and Reality*. W.H. Freeman, San Francisco, 1976.

[82] C. North. Toward measuring visualization insight. *IEEE Computer Graphics and Applications*, 26(3):6–9, 2006.

[83] C. O'Malley and S. Draper. Representation and interaction: Are mental models all in the mind. *Models in the Mind: Theory, Perspective & Application*, pages 73–91, 1992.

[84] M. Di Penta, R. Stirewalt, and E. Kraemer. Designing your next empirical study on program comprehension. In *Proc. IEEE Intl. Conf. on Program Comprehension (ICPC)*, pages 281–285, 2007.

[85] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of International Conference on Intelligence Analysis*, volume 2005, pages 2–4, 2005.

[86] C. Plaisant. The challenge of information visualization evaluation. In *Proceedings of the working conference on Advanced Visual Interfaces*, pages 109–116. ACM, 2004.

[87] C. Plaisant, J-D. Fekete, and G. G. Grinstein. Promoting insight-based evaluation of visualizations: From contest to benchmark repository. *IEEE Trans. Vis. Comput. Graph.*, 14(1):120–134, 2008.

[88] C. Plaisant, G. Grinstein, and J. Scholtz. Guest editors' introduction: Visual-analytics evaluation. *IEEE Computer Graphics and Applications*, 29(3):16–17, May/June 2009.

[89] VAST Challenge Portal. http://vac.nist.gov.

[90] M. C. Potter. Very short-term conceptual memory. *Memory & Cognition*, 21(2):156–161, 1993.

[91] K. Puolamäki and A. Bertone. Introduction to the special issue on visual analytics and knowledge discovery. *SIGKDD Explorations*, 11(2):3–4, December 2009.

[92] G. Ross and M. Chalmers. A visual workspace for constructing hybrid multidimensional scaling algorithms and coordinating multiple views. *Information Visualization*, 2(4):247–257, 2003.

[93] D. Rusu, B. Fortuna, D. Mladenić, M. Grobelnik, and R. Sipoš. Visual analysis of documents with semantic graphs. In *Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery*, pages 66–73, 2009.

[94] J. D. Saffer, V. L. Burnett, G. Chen, and P. van der Spek. Visual analytics in the pharmaceutical industry. *IEEE Computer Graphics and Applications*, 24(5):10–15, 2004.

[95] P. Saraiya, C. North, and K. Duca. An insight-based methodology for evaluating bioinformatics visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 11(4):443–456, 2005.

[96] J. Seo and B. Shneiderman. Interactively exploring hierarchical clustering results. *Computer*, 35(7):80–86, 2002.

[97] J. Seo and B. Shneiderman. *From Integrated Publication and Information Systems to Information and Knowledge Environments*, volume 3379 of *Lecture Notes in Computer Science*, chapter A Knowledge Integration Framework for Information Visualization, pages 207–220. Springer, 2005.

[98] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *IEEE Symposium on Visual Languages*, pages 336–343, 1996.

[99] B. Shneiderman. Extreme visualization: squeezing a billion records into a million pixels. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 3–12, 2008.

[100] B. Shneiderman and C. Plaisant. *Designing the user interface: strategies for effective human-computer interaction*. Addison-Wesley, 4th edition, 2004.

[101] B. Shneiderman and C. Plaisant. Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. In *Proc. AVI workshop on Novel Evaluation Methods for Information Visualization*, pages 1–7. ACM, 2006.

[102] D. J. Simons and R. A. Rensink. Change blindness: Past, present, and future. *Trends in Cognitive Sciences*, 9(1):16–20, 2005.

[103] T. A. Slocum, R. B. Mcmaster, F. C. Kessler, and H. H. Howard. Thematic cartography and geovisualization (Prentice Hall Series in Geographic Information Science). 2008.

[104] R. Spence. *Information Visualization - Design for Interaction*. Pearson Education Limited, 2nd edition, 2007.

[105] R. Spence. The broker. In *Human Aspects of Visualization*. Springer LNCS, 2010.

[106] R.J. Sternberg and J.E. Davidson. *The nature of insight*. MIT Press Cambridge, MA, 1995.

[107] C. Stolte, D. Tang, and P. Hanrahan. Polaris: A system for query, analysis, and visualization of multidimensional relational databases. *IEEE Transactions on Visualization and Computer Graphics*, 08(1):52–65, 2002.

[108] P-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison Wesley, 2005.

[109] R. Theron. Visual analytics of paleoceanographic conditions. In *Proceedings of the IEEE Symposium on Visual Analytics Science & Technology*, pages 19–26, 2006.

[110] J. Thomas. Taxonomy for visual analytics: Seeking feedback. *VAC Views*, May 2009.

[111] J. Thomas and K. Cook, editors. *Illuminating the Path: Research and Development Agenda for Visual Analytics*. IEEE Press, 2005.

[112] W. Tobler. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(2):234–240, 1970.

[113] C. Tominski, J. Abello, and H. Schumann. CGV–An interactive graph visualization system. *Computers & Graphics*, 33(6):660–678, 2009.

[114] X. Tricoche, G. Scheuermann, and H. Hagen. Tensor topology tracking: A visualization method for time-dependent 2D symmetric tensor fields. *Computer Graphics Forum*, 20(3):461–470, 2001.

[115] E. R. Tufte. *The visual display of quantitative information*. Graphics Press Cheshire, CT, 1983.

[116] J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, Reading MA, 1977.

[117] L. Tweedie, R. Spence, H. Dawkes, and H. Su. Externalising abstract mathematical models. In *Proceedings of the SIGCHI conference on human factors in computing systems: common ground*, page 406. ACM, 1996.

[118] L. Tweedie, R. Spence, D. Williams, and R. Bhogal. The Attribute Explorer. In *CHI '94: Conference companion on human factors in computing systems*, pages 435–436. ACM, 1994.

[119] J. J. van Wijk and E. R. van Selow. Cluster and calendar based visualization of time series data. In *INFOVIS*, pages 4–9, 1999.

[120] L. Voinea, J. J. Lukkien, and A. Telea. Visual assessment of software evolution. *Science of Computer Programming*, 65(3):222–248, 2007.

[121] L. Voinea and A. Telea. Case study: Visual analytics in software product assessments. In *In Proceedings of IEEE VISSOFT*, pages 65–72, 2009.

[122] C. Ware. *Information visualization: perception for design*. Morgan Kaufmann, 2004.

[123] C. Ware. *Visual thinking for design*. Morgan Kaufmann, 2008.

[124] J. Widom. Trio: A system for integrated management of data, accuracy, and lineage. Citeseer, 2005.

[125] P. C. Wong and J. Thomas. Visual analytics. *IEEE Computer Graphics and Applications*, 24(5):20–21, 2004.

[126] J. S. Yi, Y. Kang, J. Stasko, and J. Jacko. Toward a deeper understanding of the role of interaction in information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1224–1231, 2007.

[127] J. S. Yi, Y. Kang, J. T. Stasko, and J. Jacko. Understanding and characterizing insights: how do people gain insights using information visualization? In *BELIV '08: Proceedings of the 2008 conference on BEyond time and errors*, pages 1–6. ACM, 2008.

[128] J. Zhang and D. A. Norman. Representations in distributed cognitive tasks. *Cognitive science*, 18(1):87–122, 1994.

[129] Y. Zhu. Measuring effective data visualization. *Advances in Visual Computing*, pages 652–661, 2007.

List of Figures

Glossary of Terms

ACID	Atomicity, Consistency, Isolation, and Durability (database transaction properties)
API	Application Program Interface
BRETAM	Breakthrough, Replication, Empiricism, Theory, Automation, Maturity (learning model)
CAD	Computer-Aided Design
Cloud	Internet-based computing
DBMS	Database Management System
DHS	Department of Homeland Security
DMX	Data Mining Extensions
GATE	General Architecture for Text Engineering
GIS	Geographic Information System
GPS	Global Positioning System
GPU	Graphics Processing Unit
HCI	Human Computer Interaction
HDF	Hierarchical Data Format
HTML	HyperText Markup Language
ICA	International Cartographic Association
IEEE	Institute of Electrical and Electronics Engineers
INSPIRE	Infrastructure for Spatial Information in Europe
KDD	Knowledge Discovery and Data Mining
KDE	Kernel Density Estimation
MDS	Multi-Dimensional Scaling
NVAC	National Visualization and Analytics Center
OECD	Organisation for Economic Co-operation and Development
OLAP	Online Analytical Processing
PDA	Personal Digital Assistant
PMML	Predictive Model Markup Language
PNNL	Pacific Northwest National Laboratory
RDBMS	Relational Database Management System
RDF	Resource Description Framework
RSS	Really Simple Syndication
SMS	Short Message Service
SOM	Self Organising Map
SPSS	Statistical Package for the Social Sciences (SPSS Inc.)
SQL	Structured Query Language

`UIMA`	Unstructured Information Management Architecture
`UML`	Unified Modelling Language
`VTK`	Visualisation ToolKit
`Web`	World Wide Web (Internet)
`WEKA`	Waikato Environment for Knowledge Analysis
`XML`	Extensible Markup Language

visual analytics

Today, in many spheres of human activity, massive amounts of data are collected and stored. As the volumes of data available to lawmakers, civil servants, business people and scientists increase, their effective use becomes more challenging. Simply keeping up to date with the flood of data, using standard tools for data management and analysis, is fraught with difficulty. Not only is the task of managing this complex, heterogeneous data daunting; extracting information and knowledge from it is a serious challenge. The field of visual analytics seeks to provide people with better and more effective means to understand and analyse very large datasets, while at the same time enabling them to act upon their findings without delay, in real-time. Visual analytics integrates the analytic capabilities of the computer and the abilities of the human analyst. In so doing, visual analytics allows novel discoveries, unexpected insights and empowers individuals to take control of the analytical process, thus leading to beneficial and profitable innovation.

This book is the result of the VisMaster project, a coordination action funded by the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission to form a strong visual analytics research community.

ISBN 978-3-905673-77-7

VisMaster
Visual Analytics - Mastering the Information Age