First publ. in: Information Visualization (IV 2006), Invited Paper, July 5-7, London: IEEE Pr., 2006, pp. 9-16

Challenges in Visual Data Analysis*

Daniel A. Keim, Florian Mansmann, Jörn Schneidewind, and Hartmut Ziegler
University of Konstanz, Germany
{keim, mansmann, schneide, ziegler}@inf.uni-konstanz.de

Abstract

In today's applications data is produced at unprecedented rates. While the capacity to collect and store new data grows rapidly, the ability to analyze these data volumes increases at much lower pace. This gap leads to new challenges in the analysis process, since analysts, decision makers, engineers, or emergency response teams depend on information "concealed" in the data. The emerging field of visual analytics focuses on handling massive, heterogenous, and dynamic volumes of information through integration of human judgement by means of visual representations and interaction techniques in the analysis process. Furthermore, it is the combination of related research areas including visualization, data mining, and statistics that turns visual analytics into a promising field of research. This paper aims at providing an overview of visual analytics, its scope and concepts, and details the most important technical research challenges in the field.

1. Introduction

Information overload is a well-known phenomenon of the information age, since due to the progress in computer power and storage capacity over the last decades, data is produced at an incredible rate. Meanwhile, our ability to collect and store data is growing faster than our ability to analyse it. However, the analysis of these massive, typically messy and inconsistent, volumes of data is crucial in many application domains. For decision makers, analysts or emergency response teams it is an essential task to rapidly extract relevant information from the flood of data. Today, a selected number of software tools is employed to help analysts to organize their data, generate overviews and explore the information space in order to extract potentially useful information. Most of such data analysis systems still rely on interaction metaphors developed over a decade ago and it is questionable whether they are able to meet the demands of the information age. In fact, huge investments in terms of time and money are often wasted, because the possibilities to properly interact with the databases are still too limited. Visual analytics aims at bridging this gap by employing more intelligent means in the analysis process. The basic idea of visual analytics is to visually represent information, allowing the human to directly interact with it, to gain insight, to draw conclusions, and to ultimately make better decisions. Visual representation of the information reduces complex cognitive work needed to perform certain tasks. People may use visual analytics tools and techniques to synthesize information and derive insight from massive, dynamic, and often conflicting data by providing timely, defensible, and understandable assessments. [10]

The goal of visual analytics research is to turn the information overload into an opportunity. Decision-makers should be enabled to examine massive, multi-dimensional, multi-source, time-varying information stream to make effective decisions in time-critical situations. For informed decisions, it is indispensable to include humans into the data analysis process to combine flexibility, creativity, and background knowledge with the enormous storage capacity and the computational power of today's computers. The specific advantage of visual analytics is that decision makers may focus their full cognitive and perceptual capabilities on the analytical process, while allowing them to apply advanced computational capabilities to augment the discovery process. This paper gives an overview on visual analytics and discusses the most important research challenges in this field. Technical research challenges are detailed to show how visual analytics can help to turn information overload as generated by today's applications into a useful asset.

The rest of the paper is organized as follows: section 2 defines visual analytics and discusses its scope. Technical challenges are described in Section 3. Finally, section 4 presents the visual analytics mantra as a solution.

*Part of this work has been presented at Dagstuhl Seminar 05231 "Scientific Visualization: Challenges for the Future" [3]

2. Scope of Visual Analytics

In general, *visual analytics* can be described as "the science of analytical reasoning facilitated by interactive visual interfaces" [10]. To be more precise, visual analytics is an iterative process that involves collecting information, data preprocessing, knowledge representation, interaction, and decision making. The ultimate goal is to gain insight into the problem at hand which is described by vast amounts of scientific, forensic or business data from heterogeneous sources. To achieve this goal, visual analytics combines the advantages of machines with strenghts of humans. While methods from knowledge discovery in databases (KDD), statistics and mathematics are the driving force on the automatic analysis side, capabilities to perceive, relate and conclude turn visual analytics into a very promising field of research.

Historically, visual analytics has evolved out of the fields of information and scientific visualization. According to Colin Ware, the term visualization is meanwhile understood as "a graphical representation of data or concepts" [14], while the term was formerly applied to form a mental image. Nowadays fast computers and sophisticated output devices create meaningful visualizations and allow us not only to mentally visualize data and concepts, but to actually see and explore the representation of the data under consideration on a computer screen. However, transformation of data into meaningful visualizations is a non-trivial task that can not be automatically improved through steadily growing computational resources. Very often, there are many different ways to represent the data and it is unclear which representation is the best one. State-of-the-art concepts of representation, perception, interaction and decision-making need to be applied and extended to be suitable for visual data analysis.

The fields of information and scientific visualization deal with visual representations of data. Scientific visualization examines potentially huge amounts of scientific data obtained from sensors, simulations or laboratory tests with typical applications being flow visualization, volume rendering, and slicing techniques for medical illustrations. In most cases, some aspects of the data can be directly mapped onto geographic coordinates or into virtual 3D environments. We define information visualization more generally as the communication of abstract data relevant in terms of action through the use of interactive visual interfaces. There are three major goals of visualization, namely a) presentation, b) confirmatory analysis, and c) exploratory analysis. For presentation purposes, the facts to be presented are fixed a priori, and the choice of the appropriate presentation technique depends largely on the user. The aim is to efficiently and effectively communicate the results of an analysis. For confirmatory analysis, one or more hypotheses about the data serve as a starting point. The process can be described as a goal-oriented examination of these hypotheses. As a result, visualization either confirms these hypotheses or rejects them. *Exploratory data analysis* as the process of searching and analyzing databases to find implicit but potentially useful information, is a difficult task. At the beginning, the analyst has no hypothesis about the data. According to John Tuckey, tools as well as understanding are needed [12] for the interactive and usually undirected search for structures and trends.

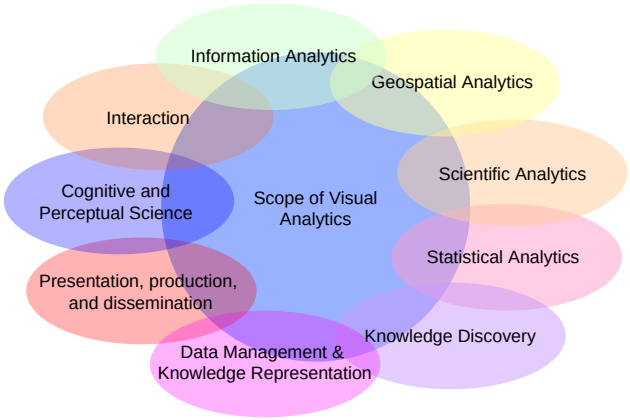

Figure 1. Visual analytics as a highly interdisciplinary field of research.

Visual analytics is more than just visualization and can rather be seen as an integrated approach combining visualization, human factors and data analysis. Figure 1 illustrates the detailed scope of visual analytics. With respect to the field of visualization, visual analytics integrates methodology from information analytics, geospatial analytics, and scientific analytics. Especially human factors (e.g., interaction, cognition, perception, collaboration, presentation, and dissemination) play a key role in the communication between human and computer, as well as in the decision-making process. In this context, *production* is defined as the creation of materials that summarize the results of an analytical effort, *presentation* as the packaging of those materials in a way that helps the audience understand the analytical results using terms that are meaningful to them, and *dissemination* as the process of sharing that information with the intended audience [11]. In matters of data analysis, visual analytics further benefits from the methodologies developed in the fields of data management & knowledge representation, knowledge discovery, and statistical analytics. Note that visual analytics, is not likely to become a separate field of study [15], but its influence will spread over the research areas it comprises.

According to Jarke J. van Wijk, "visualization is not

’good’ by definition, developers of new methods have to make clear why the information sought cannot be extracted automatically” [13]. From this statement, one immediately recognizes the need for the visual analytics approach using automatic methods from statistics, mathematics and knowledge discovery in databases (KDD) wherever they are applicable. Visualization is used as a means to efficiently communicate and explore the information space when automatic methods fail. In this context, human background knowledge, intuition, and decision-making either cannot be automated or serve as input for the future development of automated processes.

Overlooking a large information space is a typical visual analytics problem. In many cases, the information at hand is conflicting and needs to be integrated from heterogeneous data sources. Moreover, the system lacks the knowledge human experts posses. By applying analytical reasoning, hypotheses about the data can be either affirmed or discarded, eventually leading to a better understanding of the data, thus supporting the analyst in his task to gain insight. Contrary to that, a well-defined problem where the optimum or a good estimation can be calculated by non-interactive analytical means should not be classified as a visual analytics problem. In such a scenario, non-interactive analysis should be clearly preferred due to efficiency reasons. Likewise, visualization problems not involving methods for automatic data analysis do not fall into the category of visual analytics problems.

The fields of visualization and visual analytics both rely on methods from scientific, geospatial, and information analytics. They both benefit from the knowledge out of the field of interaction as well as of cognitive and perceptual science. They do differ in that visual analytics also integrates the methodology from the fields of statistical analytics, knowledge discovery, data management & knowledge representation and presentation, production & dissemination.

3. Technical Challenges

With information technology becoming a standard in most areas in the past years, from medicine to science, from education to business, from astronomy to networking (Internet), more and more digital information is generated and collected. Today, we continuously face rapidly increasing amounts of data: New sensors, faster recording methods and decreasing prices for storage capacities in the previous years allow storing huge amounts of data that used to be unimaginable a decade ago. Applications like flow simulations, molecular dynamics, nuclear science, computer tomography or astronomy generate amounts of data that can easily reach terabytes; the Large Hadron Collider (LHC) at CERN, for example, generates a volume of 1 petabyte of

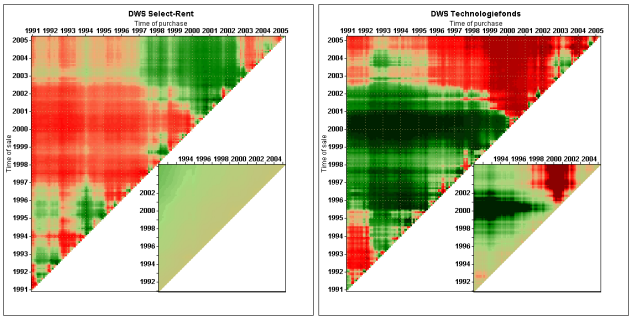

Figure 2. Visual Analysis of Financial Data: The Growth Matrix

data per year. Parallel to the growth of datasets, the computational power of the computer systems for processing these amounts of data also evolved: Faster processors, more main memory, faster networks, parallel and distributed computing and larger storage capacities increase the data throughput every year. Having no possibility of adequately exploring the large amounts of data which have been collected due to their potential usefulness, the data becomes useless and the databases become data “dumps” [2].

Commonly, the amount of data (often terabytes) to be visualized exceeds the limited amount of pixels of a display by several orders of magnitude. Filtering, aggregation, compression, principle component analysis or other data reduction techniques are then needed to reduce the amount of data as only a small portion of it can be displayed. Visual scalability is defined as the capability of visualization tools to effectively display large data sets in terms of either the number or the dimension of individual data elements [1]. We thus not only have to compare the absolute data growth and hardware performance in order to cope with a problem, but also the software and the algorithms to bring this data onto the screen in an appropriate way. *Scalability* in general is a key challenge of visual analytics as it determines the ability to process large datasets by means of computational overhead as well as appropriate rendering techniques. In the last decade, information visualization has developed numerous techniques to visualize datasets, but only some of them are scalable to the huge data sets used in visual analytics. As the amount of data is continuously growing and the amount of pixels on the display remains rather constant, the rate of compression to visualize the dataset on the display is continuously increasing, and therefore, more and more details are lost. It is the task of visual analytics to create a higher-level view of the dataset to gain insight, while maximizing the amount of details at the same time.

An example of a highly scalable visual analytics technique in the field of financial data analysis is the Growth Matrix [4]. It visualizes all possible time intervals of a fund

over a time span of 14 years (about 11.000 intervals), and simultaneously compares the performance of each interval with 14.000 funds that are in the database in just one image (see Fig. 2). To generate each image, 154 million values have to be calculated. Using a large display wall, it is possible to compare several hundreds of these Growth Matrices.

Dynamic processes, arising in business, network or telecommunications generate tremendous streams of time related or real time data. Examples are sensor logs, web statistics, network traffic logs or atmospheric and meteorological applications. Analysis of such *data streams* is an important challenge, since it plays an essential role in many areas of science and technology. As the sheer amount of data often does not allow to record all the data at full detail, effective compression and feature extraction methods are needed to manage the data. Furthermore, it is vital to provide analysis techniques and metaphors that are capable of analyzing large real time data streams in time, and to present the results in a meaningful and intuitive way. This enables quick identification of important information and timely reaction on critical process states or alarming incidents.

Real-world applications often access information from a number of heterogeneous information sources and thus require *synthesis of heterogeneous types of data* in order to perform an effective analysis. These heterogeneous data sources may include collections of vector data, strings and text documents, graphs or sets of objects. A typical application domain is computational biology, where the human genome is accompanied by real-valued gene expression data, functional annotation of genes, genotyping information, a graph of interacting proteins, equations describing the dynamics of a system, localization of proteins in a cell, and natural language text in the form of papers describing experiments, partial models, and numerous other data sources. The problem of integrating these data sources touches upon many fundamental problems in decision theory, information theory, statistics, and machine learning evidently posing a challenge for visual analytics, too. The focus on scalable and robust methods for fusing complex and heterogeneous data sources is thus a key to a more effective analysis process.

Interpretability or the ability to recognize and understand the data is one of the biggest challenges in visual analytics. Generating a visually correct output from raw data and drawing the right conclusions largely depends on the quality of the used data and methods. Many possible quality problems (e.g., data capture errors, noise, outliers, low precision, missing values, coverage errors, double counts) can already be contained in the raw data. Furthermore, preprocessing of data in order to use it for visual analysis bears many potential quality problems (i.e., data migration and parsing, data cleaning, data reduction, data enrichment, up-/ down-sampling, rounding and weighting, aggregation and combining). Data can be inherently incomplete or simply out of date. The challenges are on the one hand to determine and to minimize these errors on the pre-processing side, and to provide a flexible yet stable design of the visual analytics application to cope with data quality problems on the other hand. From the technical point of view such application can be either designed to be insensitive to data quality issues through employment of data cleaning methods or to explicitly visualize errors and uncertainty in the application to make the analyst aware of the problem. Many domains ranging from terrorism informatics to natural sciences and business intelligence would potentially benefit from the methods for enhancing data quality. Homeland Security applications in general have to deal with many missing values and uncertainty. For instance, consider a screening program in the context of Homeland Security in an airport. The system should identify potential terrorists, but also try to minimize false positives in order to avoid incorrectly targeting innocent travelers. A falsely inserted data record should not influence the principle way in which the system observes and analyses other persons. Moreover, updated data of a potential terrorist might not be available in the database for many weeks, but the visual monitoring and analysis of patterns should still work even though the records in the database are widely incomplete.

The field of *problem solving, decision science, and human information discourse* constitutes a further visual analytics challenge. Many psychological studies about the process of problem solving have been conducted. In a usual test setup the subjects have to solve a well-defined problem where the optimal solution is known to the researchers. However, real-world problems are manifold. In many cases these problems are intransparent, consist of conflicting goals, and are complex in terms of large numbers of items, interrelations, and decisions involved. In addition to these aspects, information exhibits its own dynamics by changing over time and has thus a strong impact on the optimal solution. To date, decision making is aided by so-called Decision Support Systems which try to represent expert knowledge through rules in order to reduce the risk of human errors when too much interrelated information is involved. Decision making in groups makes the human decision-making process even more complicated. As for human information discourse, it is estimated that human subjects can reliably distinguish between approximately seven categories in the process of absolute judgement [6]. This number has therefore direct consequences for the design of effective user interfaces. To assess further limitations of human capabilities for more complex tasks is challenging as subjects react differently to screen designs depending on their personality and their cultural, educational, and professional background. Summing up,

the process of decision support for problem solving requires understanding of technology on the one hand, but also comprehension of typical human capabilities such as logic, reasoning, and common sense on the other hand. Intuitive displays and interaction devices should be constructed to steer analysis and to communicate analytical results through meaningful visualizations and clear representations.

Another challenge in the context of visual analytics is to provide *semantics* for future analysis tasks and decision-centered visualization. Semantic meta data extracted from heterogeneous sources may capture associations and complex relationships. Therefore, providing techniques to analyse and detect this information is crucial to visual analytics applications. Ontology-driven techniques and systems have already started to enable new semantic applications in a wide span of areas such as bioinformatics, financial services, web services , business intelligence, and national security. However, further research is necessary in order to increase capabilities for creating and maintaining large domain ontologies and automatic extraction of semantic meta data, since the integration process between different ontologies to link various datasets is hardly automated yet. In order to perform a more powerful analysis of heterogeneous data sources, more advanced methods for the extraction of semantics from heterogenous data are a key requirement. Thereby, research challenges arise from the size of ontologies, content diversity, heterogeneity as well as from computation of complex queries and link analysis over ontology instances and meta data. New techniques are necessary to resolve semantic heterogeneities in order to discover complex relationships.

User acceptability is a further challenge; many novel visualization techniques have been presented, yet their widespread deployment has not taken place, primarily due to the users' refusal to change their working routines. Therefore, the advantages of visual analytics tools need to be communicated to the audience of future users to overcome usage barriers, and to eventually tap the full potential of the visual analytics approach. One example is the IBM Remail project [8] which tries to enhance human capabilities to cope with email overload. Concepts such as "Thread Arcs", "Correspondents Map", and "Message Map" support the user in efficiently analysing his personal email communication. MIT's project Oxygen [7] even goes one step further, by addressing the challenges of new systems to be pervasive, embedded, nomadic, adaptable, powerful, intentional and eternal.

Another challenge for visual analytics is the integration of the visualization techniques into other applications and systems. Visual analytics tools and techniques should not stand alone, but should integrate seamlessly into the applications of diverse domains, and allow interaction with other already existing systems. Although many visual analytics tools are very specific (like in astronomy or nuclear science) and therefore rather unique, in many domains (like business applications or network security) integration into existing systems would make sense. This requires interactive capabilities and input/output software interfaces for communication with other applications. Real-time visual monitoring and analysis require a connection to databases and frequent updates, and can also be used for automated analysis. In addition to that, visual analytics applications can be integrated or connected to new innovative interaction and visualization devices for improved performance.

Evaluation as a systematic determination of merit, worth, and significance of a system is crucial to its success. When evaluating a system, different aspects can be considered such as functional testing, performance benchmarks, measurement of the effectiveness of the display, economic success, user studies, assessment of its impact on decision-making to name just a few. Not all of these aspects are orthogonal, they rather commonly represent comparisons with previous systems to assess the novel system's adequacy. In this context, objective rules of thumbs to facilitate design decisions would be a great contribution to the community.

4 Solutions

Visual analytics combines strengths from information analytics, geospatial analytics, scientific analytics, statistical analytics, knowledge discovery, data management & knowledge representation, presentation, production & dissemination, cognition, perception, and interaction. It is a goal-oriented process to gain insight into heterogeneous, contradictory and incomplete data through the combination of automatic analysis methods with human background knowledge and intuition.

The example shown in Figure 3 helps to illustrate the ratio behind visual analytics. The figure shows analysis of stock market data using the CircleView [5] approach, visualizing prices of 240 stocks from the S&P 500 over 6 months, starting from January 2004 (periphery of the circle) to June 2004 (center of the circle). Instead of just visualizing the data, it is analysed by applying a relevance function in the first step. Since from the analyst's point of view it may be more interesting to analyse actual stock prizes rather than historic ones, the basic idea is to show actual stock prizes in full detail and to present historic values as aggregated high level views. That is, the relevance value of each data point is determined by its time stamp and the data is presented at different levels of details (i.e., day, week, month). This helps to reduce the amount of data and to provide overviews even on large data sets. In the next step the stocks are clustered to identify groups of stocks with similar performance over time. If the analyst then identifies

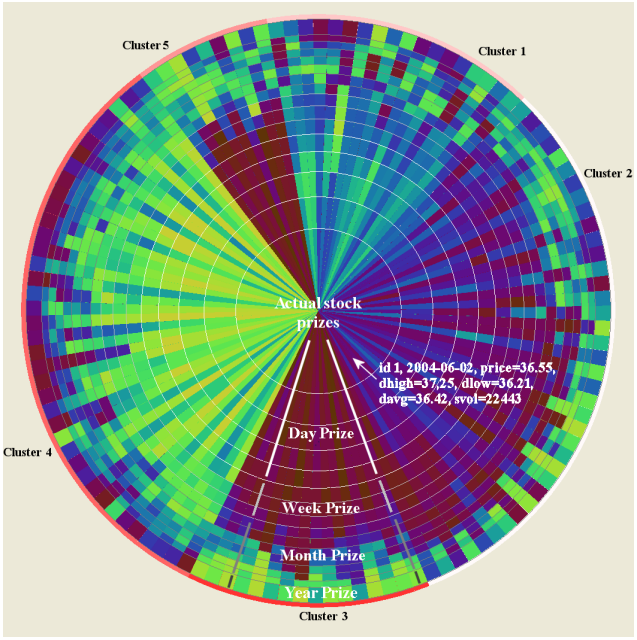

Figure 3. CircleView technique

stocks of interest or relevant groups of stocks, he can proceed by selecting relevant subsets of data and performing a detailed analysis using drill down operations on items with lower resolution or select interesting parts of the data and investigate them in more detail using chart techniques.

Unlike described in the information seeking mantra ("overview first, zoom/filter, details on demand") [9], the visual analytics process comprises the application of automatic analysis methods before and after the interactive visual representation is used. This is primarily due to the fact that current and especially future data sets are complex on the one hand and too large to be visualized in a straightforward manner on the other hand. Therefore, we present the visual analytics mantra:

"Analyse First -
Show the Important -
Zoom, Filter and Analyse Further -
Details on Demand"

The visual analytics mantra could be exemplarily applied in the context of data analysis for network security. Visualizing the raw data is unfeasible and rarely reveals any insight. Therefore, the data is first analysed (i.e., compute changes, intrusion detection analysis, etc.) and then displayed. The analyst proceeds by choosing a small suspicious subset of the recorded intrusion incidents by applying filters and zoom operations. Finally, this subset is used for a more careful analysis. Insight is gained in the course of the whole visual analytics process.

Acknowledgements

This work has been funded by the German Research Society (DFG) under the grant GK-1042, Explorative Analysis and Visualization of Large Information Spaces, Konstanz.

References

[1] S. G. Eick. Visual scalability. *Journal of Computational & Graphical Statistics*, March 2002.

[2] D. A. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 8(1):1–8, January–March 2002.

[3] D. A. Keim. Visual Exploration of Large Geo-Spatial Data Sets. In T. Ertl, E. Gröller, K. Joy, and G. Nielson, editors, *Dagstuhl Seminar 05231: Scientific Visualization*, June 2005. http://www.dagstuhl.de/05231/.

[4] D. A. Keim, T. Nietzschmann, N. Schelwies, J. Schneidewind, T. Schreck, and H. Ziegler. A spectral visualization system for analyzing financial time series data. In *EuroVis 2006: Eurographics/IEEE-VGTC Symposium on Visualization, Lisbon, Portugal, 8-10 May*, 2006.

[5] D. A. Keim, J. Schneidewind, and M. Sips. Circleview - a new approach for visualizing time related multidimensional data sets. In *ACM Advanced Visual Interfaces (AVI)*. Association for Computing Machinery (ACM), ACM Press, 2004.

[6] G. A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63:81–97, 1956.

[7] MIT Project Oxygen. http://oxygen.lcs.mit.edu/.

[8] S. L. Rohall, D. Gruen, P. Moody, M. Wattenberg, M. Stern, B. Kerr, B. Stachel, K. Dave, R. Armes, and E. Wilcox. Remail: a reinvented email prototype. In *Extended abstracts of the 2004 Conference on Human Factors in Computing Systems, CHI 2004, Vienna, Austria, April 24 - 29, 2004*, pages 791–792, 2004.

[9] 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.

[10] J. Thomas and K. Cook. *Illuminating the Path: Research and Development Agenda for Visual Analytics*. IEEE-Press, 2005.

[11] J. J. Thomas and K. A. Cook. A Visual Analytics Agenda. *IEEE Transactions on Computer Graphics and Applications*, 26(1):12–19, January/February 2006.

[12] J. W. Tuckey. *Exploratory Data Analysis*. Addison-Wesley, Reading MA, 1977.

[13] J. J. van Wijk. The value of visualization. In *IEEE Visualization*, 2005.

[14] C. Ware. *Information Visualization - Perception for Design*. Morgan Kaufmann Publishers, 1st edition, 2000.

[15] P. C. Wong and J. Thomas. Visual analytics - guest editors' introduction. *IEEE Transactions on Computer Graphics and Applications*, September/October 2004.