Visual Analytics:
Definition, Process, and Challenges

Daniel Keim[1], Gennady Andrienko[2], Jean-Daniel Fekete[3], Carsten Görg[4],
Jörn Kohlhammer[5], and Guy Melançon[6]

[1] Department of Computer and Information Science, University of Konstanz,
78457 Konstanz, Germany,
keim@informatik.uni-konstanz.de

[2] Fraunhofer Institute for Intelligent Analysis and Information Systems(IAIS),
Schloss Birlinghoven 53754 Sankt Augustin, Germany,
gennady.andrienko@iais.fraunhofer.de

[3] Université Paris-Sud, INRIA, Bât 490,
F-91405 Orsay Cedex, France,
Jean-Daniel.Fekete@inria.fr

[4] School of Interactive Computing & GVU Center, Georgia Institute of Technology,
85 5th St., NW, Atlanta, GA 30332-0760, USA,
goerg@cc.gatech.edu

[5] Fraunhofer Institute for Computer Graphics Research,
Fraunhoferstraße 5, D-64283 Darmstadt, Germany,
joern.kohlhammer@igd.fraunhofer.de

[6] INRIA Bordeaux – Sud-Ouest, CNRS UMR 5800 LaBRI,
Campus Université Bordeaux I,
351 Cours de la libération, 33405 Talence Cedex, France,
Guy.Melancon@labri.fr

1   Introduction and 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 our
way of dealing 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 generate 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.

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
- 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 massive amounts of disparate, conflicting and dynamic information, which are available from multiple heterogeneous sources. We want to simply and effectively exploit and use the hidden opportunities and knowledge resting in unexplored data sources.

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 driving problem: It is the ability to identify methods and models, which can turn the data into reliable and provable knowledge. Any technology, that claims to overcome the information overload problem, has to provide answers for the following problems:

- Who or what defines the "relevance of information" for a given task?
- How can appropriate procedures in a complex decision making process be identified?
- How can the resulting information be presented in a decision- or task-oriented way?
- What kinds of interaction can facilitate problem solving and decision making?

With every new "real-life" application, procedures are put to the test possibly under circumstances completely different from the ones under which they have been established. The awareness of the problem how to understand and analyse our data has been greatly increased in the last decade. Even as 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 quite complex. Even as fully-automated data processing methods represent the knowledge of their creators, 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 examine the procedures.

The overarching driving vision of **visual analytics** is to turn the information overload into an opportunity: Just as *information visualization* 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 visualization of these processes will provide the means of communicating about them, instead of being left with 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 (see Figure 1).

On a grand scale, visual analytics solutions provide technology that combines the strengths of human and electronic data processing. Visualization becomes the medium of a semi-automated analytical process, where humans and machines cooperate using their respective distinct capabilities for the most effective results.
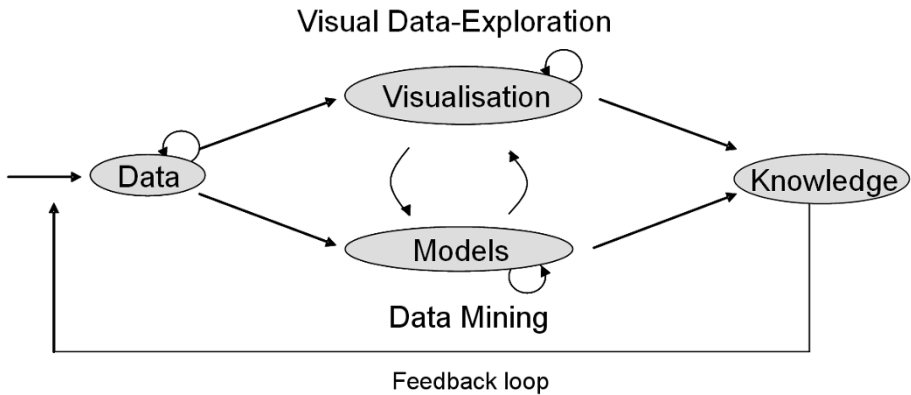
Visual Data-Exploration

Fig. 1. Tight integration of visual and automatic data analysis methods with database technology for a scalable interactive decision support.

The user has to be the ultimate authority in giving the direction of the analysis along his or her specific task. At the same time, the system has to provide effective means of interaction to concentrate on this specific task. On top of that, in many applications different people work along the path from data to decision. A visual representation will sketch this path and provide a reference for their collaboration across different tasks and abstraction levels.

The diversity of these tasks can not be tackled with a single theory. Visual analytics research is highly interdisciplinary and combines various related research areas such as visualization, data mining, data management, data fusion, statistics and cognition science (among others). Visualization has to continuously challenge the perception by many of the applying sciences that visualization is not a scientific discipline in its own right. Even if the awareness exists, that scientific analysis and results must be visualized in one way or the other, this often results in ad hoc solutions by application scientists, which rarely match the state of the art in interactive visualization science, much less the full complexity of the problems. In fact, all related research areas in the context of visual analytics research conduct rigorous, serious science each in a vibrant research community. To increase the awareness of their work and their implications for visual analytics research clearly emerges as one main goal of the international visual analytics community (see Figure 2).

Because visual analytics research can be regarded as an integrating discipline, application specific research areas should contribute with their existing procedures and models. Emerging from highly application-oriented research, dispersed research communities worked on specific solutions using the repertoire and standards of their specific fields. The requirements of visual analytics introduce new dependencies between these fields.
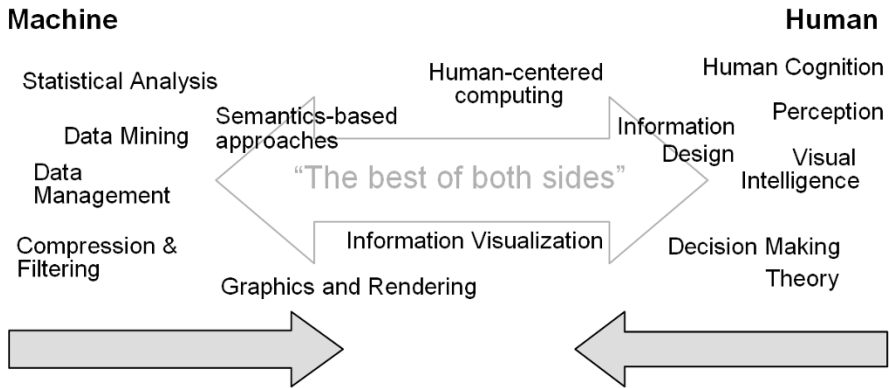
Machine   **Human**

Statistical Analysis   Human-centered   Human Cognition
                                     computing
                    Semantics-based   Perception
      Data Mining   approaches   Information
                                          Design   Visual
Data   "The best of both sides"   Intelligence
Management
                              Information Visualization
Compression &   Decision Making
Filtering   Theory
              Graphics and Rendering

Fig. 2. Visual analytics integrates scientific disciplines to improve the division of labor between human and machine.

2   Definition of Visual Analytics

In "Illuminating the Path" [39], Thomas and Cook define visual analytics as the science of analytical reasoning facilitated by interactive visual interfaces. In this paper, however, we would like to give a more specific definition:

Visual analytics combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex data sets.

The goal of visual analytics is the creation of tools and techniques to enable people to:

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

By integrating selected science and technology from the above discussed disciplines and as illustrated in Figure 2, there is the promising opportunity to form the unique and productive field of visual analytics. Work in each of the participating areas focuses on different theoretical and practical aspects of users solving real-world problems using Information Technology in an effective and efficient way. These areas have in common similar scientific challenges and significant scientific added-value from establishing close collaboration can be identified. Benefit of collaboration between the fields is identified to be two-fold:

- Jointly tackling common problems will arrive at better results on the local level of each discipline, in a more efficient way.
- Integrating appropriate results from each of the disciplines will lay the fundament for significantly improved solutions in many important data analysis applications.

Visual Analytics versus Information Visualization

Many people are confused by the new term visual analytics and do not see a difference between the two areas. While there is certainly some overlay and some of the information visualization work is certainly highly related to visual analytics, traditional visualization work does not necessarily deal with an analysis tasks nor does it always also use advanced data analysis algorithms.

Visual analytics is more than just visualization. It can rather be seen as an integral approach to decision-making, combining visualization, human factors and data analysis. The challenge is to identify the best automated algorithm for the analysis task at hand, identify its limits which can not be further automated, and then develop a tightly integrated solution with adequately integrates the best automated analysis algorithms with appropriate visualization and interaction techniques.

While some of such research has been done within the visualization community in the past, the degree to which advanced knowledge discovery algorithms have been employed is quite limited. The idea of visual analytics is to fundamentally change that. This will help to focus on the right part of the problem, i.e. the parts that can not be solved automatically, and will provide solutions to problems that we were not able to solve before.

One important remark should be made here. Most research efforts in Information Visualization have concentrated on the process of producing views and creating valuable interaction techniques for a given class of data (social network, multi-dimensional data, etc.). However, much less has been suggested as to how user interactions on the data can be turned into intelligence to tune underlying analytical processes. A system might for instance observe that most of the user's attention concern only a subpart of an ontology (through queries or by repeated direct manipulations of the same graphical elements, for instance). Why not then use this knowledge about the user's interest and update various parameters by the system (trying to systematically place elements or components of interest in center view, even taking this fact into account when driving a clustering algorithm with a modularity quality criteria, for instance).

This is one place where Visual Analytics maybe differs most from Information Visualization, giving higher priority to data analytics from the start and through all iterations of the sense making loop. Creativity is then needed to understand how perception issues can help bring more intelligence into the analytical process by "learning" from users' behavior and effective use of the visualization.

3   Areas Related to Visual Analytics

Visual analytics builds on a variety of related scientific fields. At its heart, Visual Analytics integrates Information and Scientific Visualization with Data Management and Data Analysis Technology, as well as Human Perception and Cognition research. For effective research, Visual Analytics also requires an appropriate Infrastructure in terms of software and data sets and related analytical problems repositories, and to develop reliable Evaluation methodology (see Figure 3).
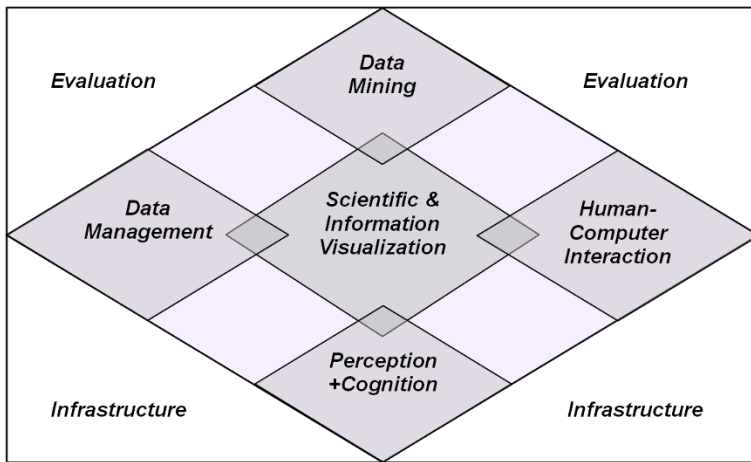
Fig. 3. Visual Analytics integrates Scientific and Information Visualization with core adjacent disciplines: Data management and analysis, spatio-temporal data, and human perception and cognition. Successful Visual Analytics research also depends on the availability of appropriate infrastructure and evaluation facilities.

An example for a common problem in several of the disciplines is that of scalability with data size. The larger the data set to be handled gets, the more difficult it gets to manage, analyze, and visualize these data effectively. Researching appropriate forms to represent large data volumes by smaller volumes containing the most relevant information benefits each of the data management, analysis, and visualization fields. On top of these individual progresses, a synergetic collaboration of all these fields may lead to significantly improved processing results. Consider a very large data stream. Appropriate data management technology gives efficient access to the stream, which is intelligently processed and abstracted by an automatic analysis algorithm which has an interface to the data management layer. On top of the analysis output, an interactive visualization which is optimized for efficient human perception of the relevant information allows the analyst to consume the analysis results, and adapt relevant parameters of the data aggregation an analysis engines as appropriate. The combination of the individual data handling steps into a Visual Analytics pipeline leads to improved results and makes data domains accessible which are not effectively accessible by any of the individual data handling disciplines. Similar argumentations apply to other related fields and disciplines. In many fields, visualization is already used and developed independently as a means for analyzing the problems at hand. However, a unified, interdisciplinary perspective on using visualization for analytical problem-solving will show beneficial for all involved disciplines. As common principles, best practices, and theories will be developed, these will become usable in the individual disciplines and application domains, providing economies of scale, avoiding replication of work or application of only sub-optimal techniques.

3.1   Visualization

Visualization has emerged as a new research discipline during the last two decades. It can be broadly classified into *Scientific* and *Information* Visualization.

In Scientific Visualization, the data entities to be visualized are typically 3D geometries or can be understood as scalar, vectorial, or tensorial fields with explicit references to time and space. A survey of current visualization techniques can be found in [22,35,23]. Often, 3D scalar fields are visualized by isosurfaces or semi-transparent point clouds (direct volume rendering) [15]. To this end, methods based on optical emission- or absorption models are used which visualize the volume by ray-tracing or projection. Also, in the recent years significant work focused on the visualization of complex 3-dimensional flow data relevant e.g., in aerospace engineering [40]. While current research has focused mainly on efficiency of the visualization techniques to enable interactive exploration, more and more methods to automatically derive relevant visualization parameters come into focus of research. Also, interaction techniques such as focus&context [28] gain importance in scientific visualization.

Information Visualization during the last decade has developed methods for the visualization of abstract data where no explicit spatial references are given [38,8,24,41]. Typical examples include business data, demographics data, network graphs and scientific data from e.g., molecular biology. The data considered often comprises hundreds of dimensions and does not have a natural mapping to display space, and renders standard visualization techniques such as (x, y) plots, line- and bar-charts ineffective. Therefore, novel visualization techniques are being developed by employing e.g., Parallel Coordinates and their numerous extensions [20], Treemaps [36], and Glyph [17]- and Pixel-based [25] visual data representations. Data with inherent network structure may be visualized using graph-based approaches. In many Visualization application areas, the typically huge volumes of data require the appropriate usage of automatic data analysis techniques such as clustering or classification as preprocessing prior to visualization. Research in this direction is just emerging.

3.2   Data Management

An efficient management of data of various types and qualities is a key component of Visual Analytics as this technology typically provides the input of the data which are to be analyzed. Generally, a necessary precondition to perform any kind of data analysis is an integrated and consistent data basis [18,19]. Database research has until the last decade focused mainly on aspects of efficiency and scalability of exact queries on homogeneous, structured data. With the advent of the Internet and the easy access it provides to all kinds of heterogeneous data sources, the database research focus has shifted toward integration of heterogeneous data. Finding integrated representation of 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 mapping of database schemata but includes also the cleaning and harmonization of uncertainty and missing data in the volumes of heterogeneous data. Modern applications require such intelligent data fusion to be feasible in near real-time and as automatically as possible [32]. New forms of information sources such as data streams [11], sensor networks [30] or automatic extraction of information from large document collections (e.g., text, HTML) result in a difficult data analysis problem which to support is currently in the focus of database research [43]. The relationship between Data Management, Data Analysis and Visualization is characterized such that Data Management techniques developed increasingly rely on intelligent data analysis techniques, and also interaction and visualization to arrive at optimal results. On the other hand, modern database systems provide the input data sources which are to be visually analyzed.

3.3 Data Analysis

Data Analysis (also known as Data Mining or Knowledge Discovery) researches methods to automatically extract valuable information from raw data by means of automatic analysis algorithms [29,16,31]. Approaches developed in this area can be best described by the addressed analysis tasks. A prominent such task is supervised learning from examples: 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 [13]. A huge number of algorithms have been developed to this end such as Decision Trees, Support Vector Machines, Neuronal Networks, and so on. A second prominent analysis task is that of cluster analysis [18,19], which aims to extract structure from data without prior knowledge being available. Solutions 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 for subsequent analysis steps. Further data analysis tasks include tasks such as association rule mining (analysis of co-occurrence of data items) and dimensionality reduction. While data analysis initially was developed for structured data, recent research aims at analyzing also semi-structured and complex data types such as web documents or multimedia data [34].

It has recently been recognized that visualization and interaction are highly beneficial in arriving at optimal analysis results [9]. 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. Visualization is also a suitable means for appropriately communicating the results of the automatic analysis, which often is given in abstract representation, e.g., a decision tree. Visual Data Mining methods [24] try to achieve exactly this.

3.4 Perception and Cognition

Effective utilization of the powerful human perception system for visual analysis tasks requires the careful design of appropriate human-computer interfaces. Psychology, Sociology, Neurosciences and Design each contribute valuable results to

the implementation of effective visual information systems. Research in this area focuses on user-centered analysis and modeling (Requirement Engineering), the development of principles, methods and tools for design of perception-driven, multimodal interaction techniques for visualization and exploration of large information spaces, as well as usability evaluation of such systems [21,12]. On the technical side, research in this area is influenced by two main factors: (1.) The availability of improved display resources (hardware), and (2.) Development of novel interaction algorithms incorporating machine recognition of the actual user intent and appropriate adaptation of main display parameters such as the level of detail, data selection and aggregation, etc. by which the data is presented[44]. Important problems addressed in this area include the research of perceptual, cognitive and graphical principles which in combination lead to improved visual communication of data and analysis results; The development of perception-theory-based solutions for the graphical representation of static and dynamic structures; And development of visual representation of information at several levels of abstraction, and optimization of existing focus-and-context techniques.

3.5   Human-Computer Interaction

Human-computer interaction is the research area that studies the interaction between people and computers. It involves the design, implementation and evaluation of interactive systems in the context of the user's task and work [12]. Like visual analytics itself, human-computer interaction is a multi-disciplinary research area that draws on many other disciplines: computer science, system design, and behavioral science are some of them. The basic underlying research goal is to improve the interaction between users and computers: how to make computers more receptive to the users' intentions and needs. Thus, the research areas discussed in the previous section about perception and cognition are also much related to human-computer interaction [21].

As pointed out in the introduction, visual analytics aims to combine and integrate the strengths of computers and humans into an interactive process to extract knowledge from data. To effectively switch back and forth between tasks for the computer and tasks for the human it is crucial to develop an effective user interface that minimizes the barrier between the human's cognitive model of what they want to accomplish and the computer's understanding of the human's task. The design of user interfaces focuses on human factors of interactive software, methods to develop and assess interfaces, interaction styles, and design considerations such as effective messages and appropriate color choice [37].

3.6   Infrastructure and Evaluation

The above described research disciplines require cross-discipline support regarding the evaluation of the found solutions, and need certain infrastructure and standardization grounding to build on effectively. In the field of information visualization, standardization and evaluation came into the focus of research only recently. It has been realized that a general understanding of the taxonomies

regarding the main data types and user tasks [2] to be supported are highly desirable for shaping visual analytics research. A common understanding of data and problem dimensions and structure, and acceptance of evaluation standards will make research results better comparable, optimizing research productivity. Also, there is an obvious need to build repositories of available analysis and visualization algorithms, which researchers can build upon in their work, without having to re-implement already proven solutions.

How to assess the value of visualization is a topic of lively debate [42,33]. A common ground that can be used to position and compare future developments in the field of data analysis is needed. The current diversification and dispersion of visual analytics research and development resulted from its focus onto specific application areas. While this approach may suit the requirements of each of these applications, a more rigorous and overall scientific perspective will lead to a better understanding of the field and a more effective and efficient development of innovative methods and techniques.

3.7   Sub-communities

Spatio-Temporal Data: While many different data types exist, one of the most prominent and ubiquitous data types is data with references to time and space. The importance of this data type has been recognized by a research community which formed around spatio-temporal data management and analysis [14]. In geospatial data research, data with references in the real world coming from e.g., geographic measurements, GPS position data, remote sensing applications, and so on is considered. Finding spatial relationships and patterns among this data is of special interest, requiring the development of appropriate management, representation and analysis functions. E.g., developing efficient data structures or defining distance and similarity functions is in the focus of research. Visualization often plays a key role in the successful analysis of geospatial data [6,26].

In temporal data, the data elements can be regarded as a function of time. Important analysis tasks here include the identification of patterns (either linear or periodical), trends and correlations of the data elements over time, and application-dependent analysis functions and similarity metrics have been proposed in fields such as finance, science, engineering, etc. Again, visualization of time-related data is important to arrive at good analysis results [1].

The analysis of data with references both in space and in time is a challenging research topic. Major research challenges include [4]: scale, as it is often necessary to consider spatio-temporal data at different spatio-temporal scales; the uncertainty of the data as data are often incomplete, interpolated, collected at different times, or based upon different assumptions; complexity of geographical space and time, since in addition to metric properties of space and time and topological/temporal relations between objects, it is necessary to take into account the heterogeneity of the space and structure of time; and complexity of spatial decision making processes, because a decision process may involve hetero-

geneous actors with different roles, interests, levels of knowledge of the problem domain and the territory.

Network and Graph Data: Graphs appear as flexible and powerful mathematical tools to model real-life situations. They naturally map to transportation networks, electric power grids, and they are also used as artifacts to study complex data such as observed interactions between people, or induced interactions between various biological entities. Graphs are successful at turning semantic proximity into topological connectivity, making it possible to address issues based on algorithmics and combinatorial analysis.

Graphs appear as essential modeling and analytical objects, and as effective visual analytics paradigms. Major research challenges are to produce scalable analytical methods to identify key components both structurally and visually. Efforts are needed to design process capable of dealing with large datasets while producing readable and usable graphical representations, allowing proper user interaction. Special efforts are required to deal with dynamically changing networks, in order to assess of structural changes at various scales.

4   The Visual Analytics Process

A number of systems for information visualization, as well as specific visualization techniques, motivate their design choice from Shneiderman's celebrated mantra "Overview first, Filter and zoom, Details on demand". As is, the mantra clearly emphasizes the role of visualization in the knowledge discovery process. Recently, Keim adjusted the mantra to bring its focus toward Visual Analytics: "Analyze first, Show the Important, Zoom, filter and analyze further, Details on demand". In other words, this mantra is calling for astute combinations of analytical approaches together with advanced visualization techniques.

The computation of any visual representation and/or geometrical embedding of large and complex datasets requires some analysis to start with. Many scalable graph drawing algorithms try to take advantage of any knowledge on topology to optimize the drawing in terms of readability. Other approaches offer representations composed of visual abstractions of clusters to improve readability. The challenge then is to try to come up with a representation that is as faithful as possible to avoid introducing uncertainty. We must not fall into the naïve assumption that visualization can offer a virgin view on the data: any representation will inevitably favor an interpretation over all possible ones. The solution offered by Visual Analytics is then to let the user enter into a loop where data can be interactively manipulated to help gain insight both on the data and the representation itself.

The sense-making loop structures the whole knowledge discovery process supported through Visual Analytics. A generic scenario can be given following a schema developed by van Wijk [42], which furthermore admits to be evaluated and measured in terms of efficiency or knowledge gained. A choice for an initial representation and adequate interactions can be made after applying different

statistical and mathematical techniques, such as spatio-temporal data analysis or link mining depending on the nature of the dataset under study. The process then enters a loop where the user can gain knowledge on the data, ideally driving the system toward more focused and more adequate analytical techniques. Dually, interacting on the visual representation, the user will gain a better understanding of the visualization itself commanding for different views helping him or her to go beyond the visual and ultimately confirm hypotheses built from previous iterations (see Figure 4).
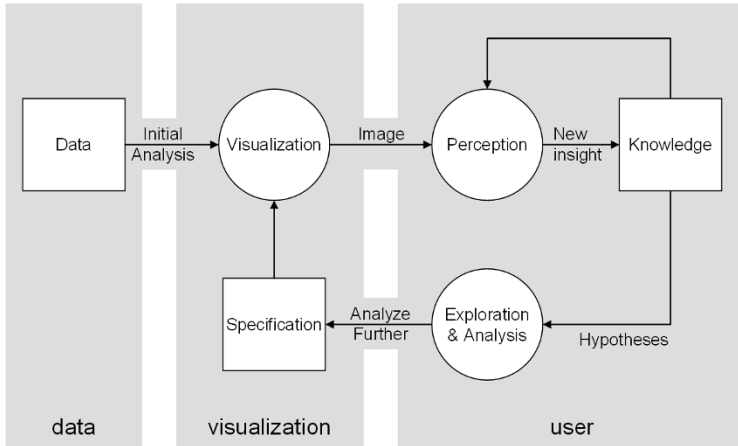

Fig. 4. The sense-making loop for Visual Analytics based on the simple model of visualization by Wijk [42].

5   Application Challenges

Visual Analytics is a highly application oriented discipline driven by practical requirements in important domains. Without attempting a complete survey over all possible application areas, we sketch the potential applicability of Visual Analytics technology in a few key domains.

In the **Engineering** domain, Visual Analytics can contribute to speed-up development time for products, materials, tools and production methods by offering more effective, intelligent access to the wealth of complex information resulting from prototype development, experimental test series, customers' feedback, and many other performance metrics. One key goal of applied Visual Analytics in the engineering domain will be the analysis of the complexity of the production systems in correlation with the achieved output, for an efficient and effective improvement of the production environments.

Financial Analysis is a prototypical promising application area for Visual Analytics. Analysts in this domain are confronted with streams of heterogeneous information from different sources available at high update rates, and of varying

reliability. Arriving at a unifying, task-centered view on diverse streams of data is a central goal in financial information systems. Integrated analysis and visualization of heterogeneous data types such as news feeds, real-time trading data, and fundamental economic indicators poses a challenge for developing advanced analysis solutions in this area. Research based on results from Information Visualization is regarded as promising in this case.

Socio-economic considerations often form the basis of political decision processes. A modern society can be regarded as a complex system of interrelationships between political decisions and economic, cultural and demographic effects. Analysis and Visualization of these interrelationships is promising in developing a better understanding of these phenomena, and to arrive at better decisions. Successful Visual Analytics applications in this domain could start being developed based on currently existing Geo-Spatial analysis frameworks.

Public **Safety & Security** is another important application area where Visual Analytics may contribute with advanced solutions. Analysts need to constantly monitor huge amounts of heterogeneous information streams, correlating information of varying degrees of abstraction and reliability, assessing the current level of public safety, triggering alert in case of alarming situations being detected. Data integration and correlation combined with appropriate analysis and interactive visualization is promising to develop more efficient tools for the analysis in this area.

The study of **Environment and Climate change** often requires the examination of long term weather records and logs of various sensors, in a search for patterns that can be related to observations such as changes in animal populations, or in meteorological and climatic processes for instance. These requirements call for the development of systems allowing visual and graphical access to historical monitoring data and predictions from various models in search for or in order to validate patterns building over time.

These diverse fields of applications share many problems on an abstract level, most of which are addressed by Visual Analytics. The actual (software) solution must be adapted to the specific needs and terminologies of the application area and consequently, many researchers currently focus on a specific customer segment. Much can be achieved, if the European research infrastructure in this field becomes strong enough to encourage the exchange of ideas on a broad scale, to foster development of solutions applicable to multiple domains, achieving synergy effects.

6   Technical Challenges

The primary goal of Visual Analytics is the analysis of vast amounts of data to identify and visually distill the most valuable and relevant information content. The visual representation should reveal structural patterns and relevant data properties for easy perception by the analyst. A number of key requirements need to be addressed by advanced Visual Analytics solutions. We next outline important scientific challenges in this context.

Scalability with Data Volumes and Data Dimensionality: Visual Analytics techniques need to be able to scale with the size and dimensionality of the input data space. Techniques need to accommodate and graphically represent high-resolution input data as well as continuous input data streams of high bandwidth. In many applications, data from multiple, heterogeneous sources need to be integrated and processed jointly. In these cases, the methods need to be able to scale with a range of different data types, data sources, and levels of quality. The visual representation algorithms need to be efficient enough for implementation in interactive systems.

Quality of Data and Graphical Representation: A central issue in Visual Analytics is the avoidance of misinterpretations by the analyst. This may result due to uncertainty and errors in the input data, or limitations of the chosen analysis algorithm, and may produce misleading analysis results. To face this problem, the notion of data quality, and the confidence of the analysis algorithm needs to be appropriately represented in the Visual Analytics solutions. The user needs to be aware of these data and analysis quality properties at any stage in the data analysis process.

Visual Representation and Level of Detail: To accommodate vast streams of data, appropriate solutions need to intelligently combine visualizations of selected analysis details on the one hand, and a global overview on the other hand. The relevant data patterns and relationships need to be visualized on several levels of detail, and with appropriate levels of data and visual abstraction.

User Interfaces, and Interaction Styles and Metaphors: Visual Analytics systems need to be easily used and interacted with by the analyst. The analyst needs to be able to fully focus on the task at hand, not on overly technical or complex user interfaces, which potentially distract. To this end, novel interaction techniques need to be developed which fully support the seamless, intuitive visual communication with the system. User feedback should be taken as intelligently as possible, requiring as little manual user input as possible, which guarantees the full support of the user in navigating and analyzing the data, memorizing insights and making informed decisions.

Display Devices: In addition to high-resolution desktop displays, advanced display devices such as large-scale power walls and small portable personal assistant, graphically-enabled devices need to be supported. Visual Analytics systems should adapt to the characteristics of the available output devices, supporting the Visual Analytics workflow on all levels of operation.

Evaluation: Due to the complex and heterogeneous problem domains addressed by Visual Analytics, so far it has been difficult to perform encompassing evaluation work. A theoretically founded evaluation framework needs to be developed which allows assessing the contribution of any Visual Analytics system toward the level of effectiveness and efficiency achieved regarding their requirements.

Infrastructure: Managing large amounts of data for visualization or analysis requires special data structures and mechanisms, both in memory and disks. Achieving interactivity means refreshing the display in 100ms at worst whereas analyzing data with standard techniques such as clustering can take hours to complete. Achieving the smooth interaction required by the analysts to perform their tasks while providing high-quality analytical algorithms need the combination of asynchronous computation with hybrid analytical algorithms that can trade time with quality. Moreover, to fully support the analytical process, the history of the analysis should also be recorded and interactively edited and annotated. Altogether, these requirements call for a novel software infrastructure, built upon well understood technologies such as databases, software components and visualization but augmented with asynchronous processing, history managements and annotations.

7   Examples for Visual Analytics Applications

7.1   Visual Analytics Tools for Analysis of Movement Data

With widespread availability of low cost GPS devices, it is becoming possible to record data about the movement of people and objects at a large scale. While these data hide important knowledge for the optimization of location and mobility oriented infrastructures and services, by themselves they lack the necessary semantic embedding which would make fully automatic algorithmic analysis possible. At the same time, making the semantic link is easy for humans who however cannot deal well with massive amounts of data. In [5] we argue that by using the right visual analytics tools for the analysis of massive collections of movement data, it is possible to effectively support human analysts in understanding movement behaviors and mobility patterns.

Figure 5 shows a subset of raw GPS measurements presented in so-called space-time cube. The large amount of position records referring to the same territory over a long time period makes it virtually impossible to do the analysis by purely visual methods.

The paper [5] proposes a framework where interactive visual interfaces are synergistically combined with database operations and computational processing. The generic database techniques are used for basic data processing and extraction of relevant objects and features. The computational techniques, which are specially devised for movement data, aggregate and summarize these objects and features and thereby enable the visualization of large amounts of information. The visualization enables human cognition and reasoning, which, in turn, direct and control the further analysis by means of the database, computational, and visual techniques. Interactive visual interfaces embrace all the tools.

Thus, in order to detect and interpret significant places visited by the moving entities, the positions of stops are extracted from the data by means of appropriate database queries. Then, clustering methods are applied to detect frequently visited places. Interactive visual displays put the results in the spatial and temporal contexts. The spatial positions of the stops can be observed on
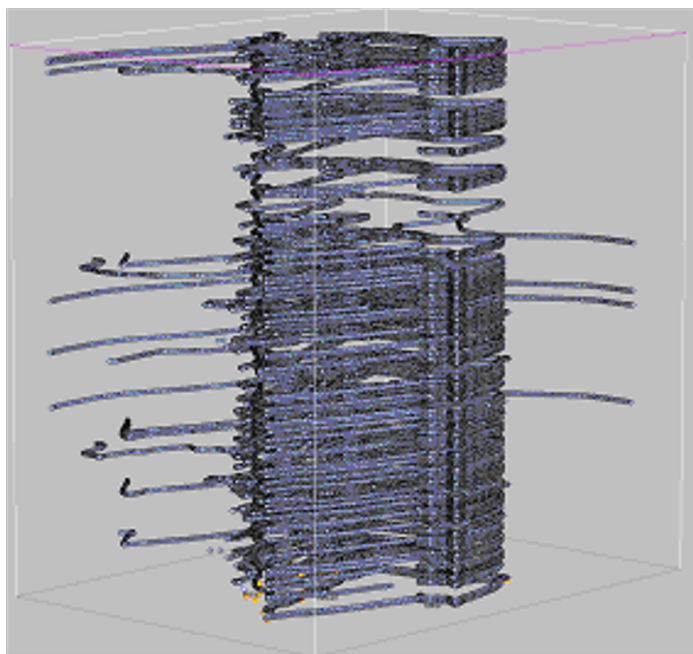
Fig. 5. A visual display of a large amount of position records is unreadable and not suitable for analysis.
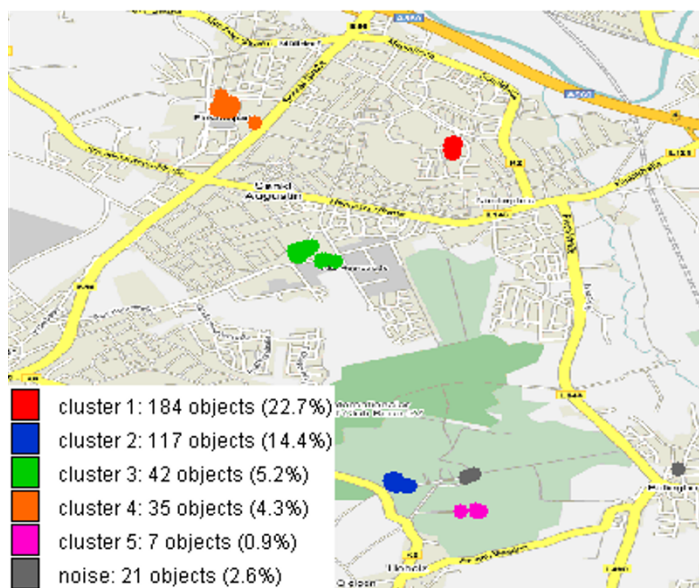

cluster 1: 184 objects (22.7%)
cluster 2: 117 objects (14.4%)
cluster 3: 42 objects (5.2%)
cluster 4: 35 objects (4.3%)
cluster 5: 7 objects (0.9%)
noise: 21 objects (2.6%)

Fig. 6. Positions of stops have been extracted from the database. By means of clustering, frequently visited places have been detected.
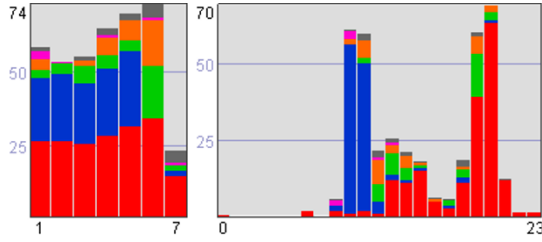
Fig. 7. The temporal histograms show the distribution of the stops in the frequently visited places (Figure 6) with respect to the weekly (left) and daily (right) cycles.

a map (Figure 6) or 3D spatial view. Temporal histograms (Figure 7) are used to explore the temporal distribution of the stops throughout the time period and within various temporal cycles (daily, weekly, etc.). These complementary views allow a human analyst to understand the meanings or roles of the frequently visited places.

In order to detect and interpret typical routes of the movement between the significant places, the analyst first applies a database query to extract sequences of position records between the stops, from which trajectories (time-referenced lines) are constructed. Then, clustering is applied with the use of specially devised similarity measures. The results are computationally generalized and summarized and displayed in the spatial context (Figure 8).

7.2   Multilevel Visualization of the Worldwide Air Transportation Network

The air transportation network has now become more dense and more complex at all geographical levels. Its dynamic no more rests on simple territorial logics. The challenge is to gain insightful understandings on how the routes carrying the densest traffic organize themselves and impact the organization of the network into sub-communities at lower levels. At the same time, subnetworks grow on their own logic, involving tourism, economy or territorial control, and influence or fight against each other. Because of the network size and complexity, its study can no more rely on traditional world map and requires novel visualization. A careful analysis of the network structural properties, requiring recent results on small world phenomenon, reveals its multilevel community structure.

The original network is organized into a top level network of communities (Figure 9(a)). Each component can then be further decomposed into sub-communities. Capitals such as New York, Chicago, Paris or London (Figure 9(b)) clearly attract most of the international traffic and impose routes to fly the world around because of airline partnerships (economical logic). Asia (Figure 9(c)) clearly stands apart from these core hubs because of strong territorial ties endorsed by national Asian airline companies (territorial logic). Visualization of social networks such as the worldwide air transportation is challenged by the necessity to scale with the growing size of network data while being able to offer
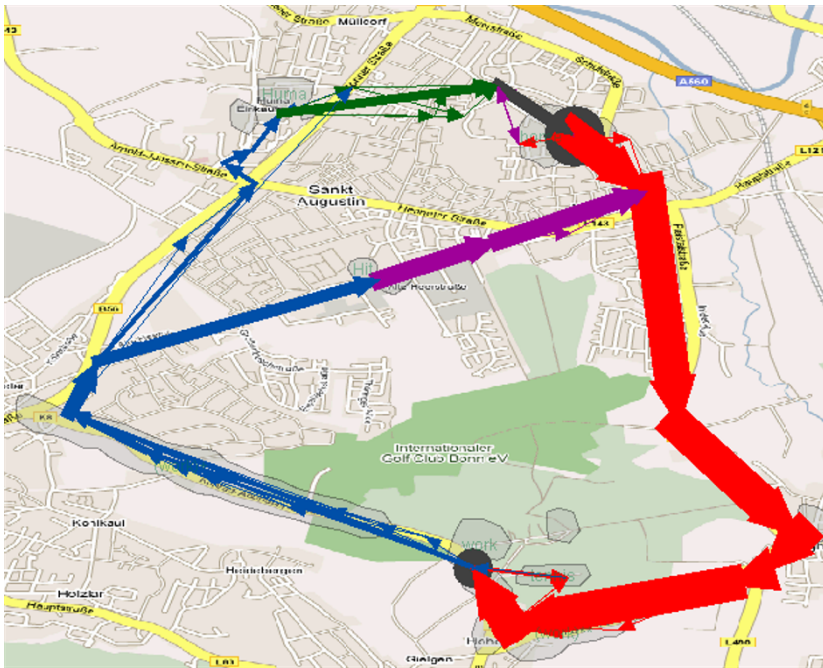
Fig. 8. A result of clustering and summarization of movement data: the routes between the significant places.

readable visual representations and fluid interaction. Visualization today brings the field of social sciences close to the study of complex systems and promises to deliver new knowledge across these disciplines [7,3,10].

8   Conclusions

The problems addressed by Visual Analytics are generic. Virtually all sciences and many industries rely on the ability to identify methods and models, which can turn data into reliable and provable knowledge. Ever since the dawn of modern science, researchers needed to find methodologies to create new hypotheses, to compare them with alternative hypotheses, and to validate their results. In a collaborative environment this process includes a large number of specialized people each having a different educational background. The ability to communicate results to peers will become crucial for scientific discourse.

Currently, no technological approach can claim to give answers to all three key questions that have been outlined in the first section, regarding the

- relevance of a specific information
- adequacy of data processing methods and validity of results
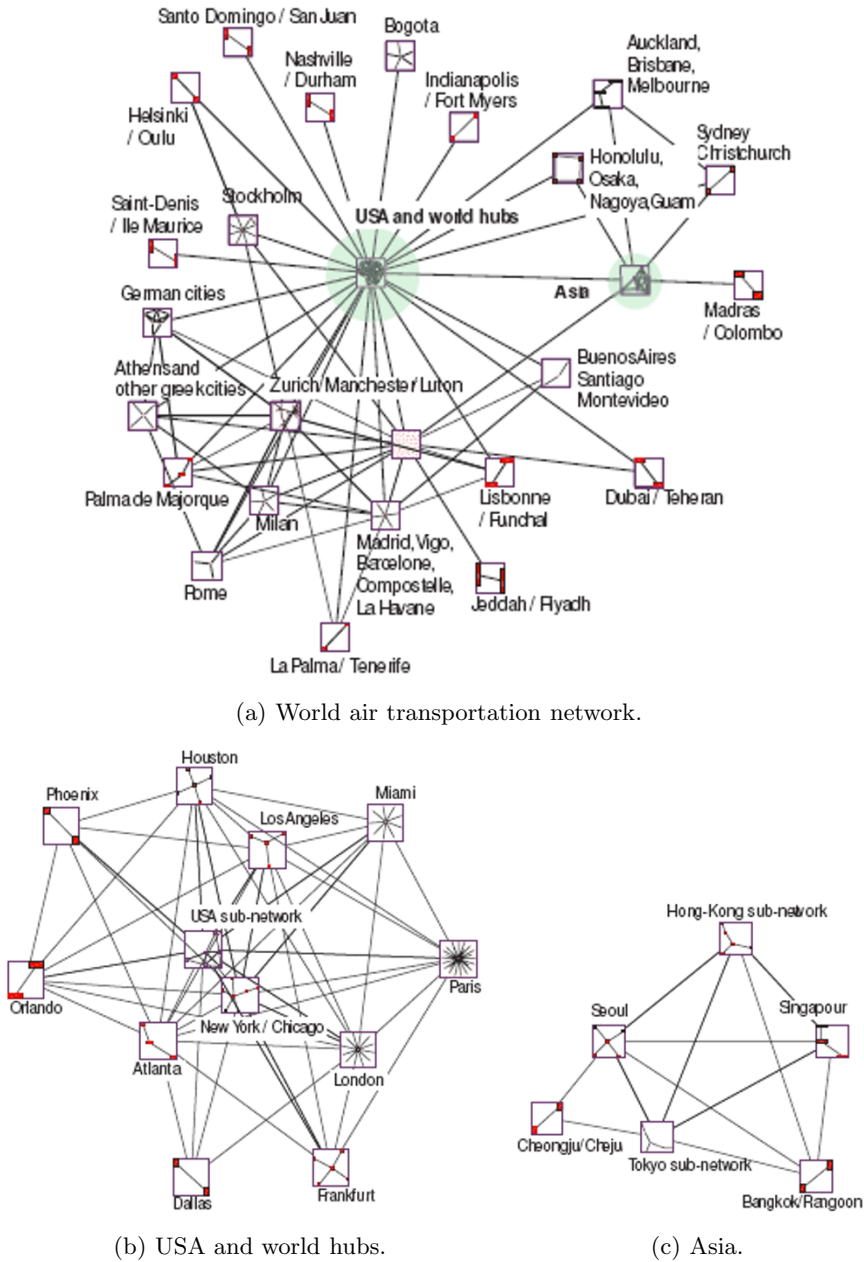- acceptability of the presentation of results for a given task

(a) World air transportation network.

(b) USA and world hubs.

(c) Asia.

Fig. 9. Multilevel Visualization of the Worldwide Air Transportation Network

Visual Analytics research does not focus on specific methods to address these questions in a single "best-practice". Each specific domain contributes a repertoire of approaches to initiate an interdisciplinary creation of solutions.

Visual Analytics literally maps the connection between different alternative solutions, leaving the opportunity for the human user to view these options in the context of the complete knowledge generation process and to discuss these options with peers on common ground.

References

1. Aigner, W., Miksch, S., Müller, W., Schumann, H., Tominski, C.: Visual methods for analyzing time-oriented data. IEEE Transactions on Visualization and Computer Graphics 14(1), 47–60 (2008)
2. Amar, R.A., Eagan, J., Stasko, J.T.: Low-level components of analytic activity in information visualization. In: INFOVIS, p. 15 (2005)
3. Amiel, M., Melançon, G., Rozenblat, C.: Réseaux multi-niveaux: l'exemple des échanges aériens mondiaux. M@ppemonde 79(3) (2005)
4. Andrienko, G., Andrienko, N., Jankowski, P., Keim, D., Kraak, M.-J., MacEachren, A., Wrobel, S.: Geovisual analytics for spatial decision support: Setting the research agenda. Special issue of the International Journal of Geographical Information Science 21(8), 839–857 (2007)
5. Andrienko, G., Andrienko, N., Wrobel, S.: Visual analytics tools for analysis of movement data. ACM SIGKDD Explorations 9(2) (2007)
6. Andrienko, N., Andrienko, G.: Exploratory Analysis of Spatial and Temporal Data. Springer, Heidelberg (2005)
7. Auber, D., Chiricota, Y., Jourdan, F., Melançon, G.: Multiscale visualization of small world networks. In: INFOVIS (2003)
8. Card, S.K., Mackinlay, J., Shneiderman, B.: Readings in Information Visualization: Using Vision to Think. Morgan Kaufmann, San Francisco (1999)
9. Ceglar, A., Roddick, J.F., Calder, P.: Guiding knowledge discovery through interactive data mining, pp. 45–87. IGI Publishing, Hershey (2003)
10. Chiricota, Y., Melançon, G.: Visually mining relational data. International Review on Computers and Software (2005)
11. Das, A.: Semantic approximation of data stream joins. IEEE Transactions on Knowledge and Data Engineering 17(1), 44–59 (2005), Member-Johannes Gehrke and Member-Mirek Riedewald
12. Dix, A., Finlay, J.E., Abowd, G.D., Beale, R.: Human-Computer Interaction (.), 3rd edn. Prentice-Hall, Inc., Upper Saddle River (2003)
13. Duda, R., Hart, P., Stock, D.: Pattern Classification. John Wiley and Sons Inc., Chichester (2000)
14. Dykes, J., MacEachren, A., Kraak, M.-J.: Exploring geovisualization. Elsevier Science, Amsterdam (2005)
15. Engel, K., Hadwiger, M., Kniss, J.M., Rezk-salama, C., Weiskopf, D.: Real-time Volume Graphics. A. K. Peters, Ltd., Natick (2006)
16. Ester, M., Sander, J.: Knowledge Discovery in Databases - Techniken und Anwendungen. Springer, Heidelberg (2000)
17. Forsell, C., Seipel, S., Lind, M.: Simple 3d glyphs for spatial multivariate data. In: INFOVIS, p. 16 (2005)

18. Han, J., Kamber, M. (eds.): Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco (2000)
19. Hand, D., Mannila, H., Smyth, P. (eds.): Principles of Data Mining. MIT Press, Cambridge (2001)
20. Inselberg, A., Dimsdale, B.: Parallel Coordinates: A Tool for Visualizing Multivariate Relations (chapter 9), pp. 199–233. Plenum Publishing Corporation, New York (1991)
21. Jacko, J.A., Sears, A.: The Handbook for Human Computer Interaction. Lawrence Erlbaum & Associates, Mahwah (2003)
22. Johnson, C., Hanson, C. (eds.): Visualization Handbook. Kolam Publishing (2004)
23. Keim, D., Ertl, T.: Scientific visualization (in german). Information Technology 46(3), 148–153 (2004)
24. Keim, D., Ward, M.: Visual Data Mining Techniques (chapter 11). Springer, New York (2003)
25. Keim, D.A., Ankerst, M., Kriegel, H.-P.: Recursive pattern: A technique for visualizing very large amounts of data. In: VIS '95: Proceedings of the 6th conference on Visualization '95, Washington, DC, USA, p. 279. IEEE Computer Society Press, Los Alamitos (1995)
26. Keim, D.A., Panse, C., Sips, M., North, S.C.: Pixel based visual data mining of geo-spatial data. Computers &Graphics 28(3), 327–344 (2004)
27. Kerren, A., Stasko, J.T., Fekete, J.-D., North, C.J. (eds.): Information Visualization. LNCS, vol. 4950. Springer, Heidelberg (2008)
28. Krúger, J., Schneider, J., Westermann, R.: Clearview: An interactive context preserving hotspot visualization technique. IEEE Transactions on Visualization and Computer Graphics 12(5), 941–948 (2006)
29. Maimon, O., Rokach, L. (eds.): The Data Mining and Knowledge Discovery Handbook. Springer, Heidelberg (2005)
30. Meliou, A., Chu, D., Guestrin, C., Hellerstein, J., Hong, W.: Data gathering tours in sensor networks. In: IPSN (2006)
31. Mitchell, T.M.: Machine Learning. McGraw-Hill, Berkeley (1997)
32. Naumann, F., Bilke, A., Bleiholder, J., Weis, M.: Data fusion in three steps: Resolving schema, tuple, and value inconsistencies. IEEE Data Eng. Bull. 29(2), 21–31 (2006)
33. North, C.: Toward measuring visualization insight. IEEE Comput. Graph. Appl. 26(3), 6–9 (2006)
34. Perner, P. (ed.): Data Mining on Multimedia Data. LNCS, vol. 2558. Springer, Heidelberg (2002)
35. Schumann, H., Müller, W.: Visualisierung - Grundlagen und allgemeine Methoden. Springer, Heidelberg (2000)
36. Shneiderman, B.: Tree visualization with tree-maps: 2-d space-filling approach. ACM Trans. Graph. 11(1), 92–99 (1992)
37. Shneiderman, B., Plaisant, C.: Designing the User Interface. Addison-Wesley, Reading (2004)
38. Spence, R.: Information Visualization. ACM Press, New York (2001)
39. Thomas, J.J., Cook, K.A.: Illuminating the Path. IEEE Computer Society Press, Los Alamitos (2005)
40. Tricoche, X., Scheuermann, G., Hagen, H.: Tensor topology tracking: A visualization method for time-dependent 2d symmetric tensor fields. Comput. Graph. Forum 20(3) (2001)
41. Unwin, A., Theus, M., Hofmann, H.: Graphics of Large Datasets: Visualizing a Million (Statistics and Computing). Springer, New York (2006)

42. van Wijk, J.J.: The value of visualization. In: IEEE Visualization, p. 11 (2005)
43. Widom, J.: Trio: A system for integrated management of data, accuracy, and lineage. In: CIDR, pp. 262–276 (2005)
44. Yi, J.S., Kang, Y.a., Stasko, J.T., Jacko, J.A.: Toward a deeper understanding of the role of interaction in information visualization. IEEE Trans. Vis. Comput. Graph. 13(6), 1224–1231 (2007)