Visual Analytics

Daniel A. Keim, Florian Mansmann, Andreas Stoffel, Hartmut Ziegler
University of Konstanz, Germany
http://infovis.uni-konstanz.de

SYNONYMS

Visual Analysis; Visual Data Analysis; Visual Data Mining

DEFINITION

Visual analytics is the science of analytical reasoning supported by interactive visual interfaces according to [6]. Over the last decades data was produced at an incredible rate. However, the ability to collect and store this data is increasing at a faster rate than the ability to analyze it. While purely automatic or purely visual analysis methods were developed in the last decades, the complex nature of many problems makes it indispensable to include humans at an early stage in the data analysis process. Visual analytics methods allow decision makers to combine their flexibility, creativity, and background knowledge with the enormous storage and processing capacities of today's computers to gain insight into complex problems. The goal of visual analytics research is thus to turn the information overload into an opportunity by enabling decision-makers to examine this massive information stream to take effective actions in real-time situations.

HISTORICAL BACKGROUND

Automatic analysis techniques such as statistics and data mining developed independently from visualization 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 to exploratory data analysis, which was first stated in the statistics research community by John W. Tukey in his book "Exploratory data analysis" [7].

Later with the availability of graphical user interfaces with proper interaction devices, a whole research community devoted their efforts to information visualization [1, 2, 5, 8]. At some stage, this community recognized the potential of integrating the user in the KDD process through effective and efficient visualization techniques, interaction capabilities and knowledge transfer leading to *visual data exploration* or *visual data mining* [3]. This integration considerably widened the scope of both the information visualization and the data mining fields, resulting in new techniques and plenty of interesting and important research opportunities.

The term *visual analytics* was coined by Jim Thomas in the research and development agenda "Illuminating the Path" [6], which had a strong focus on Homeland Security in the United States. Meanwhile, the term is used in a wider context, describing a new multidisciplinary field that combines various research areas including visualization, human-computer interaction, data analysis, data management, geo-spatial and temporal data processing and statistics [4].

SCIENTIFIC FUNDAMENTALS

Visual analytics evolved from information visualization and automatic data analysis. It combines both former independent fields and strongly encourages human interaction in the analysis process as illustrated in Figure 1. The focus of this section is to differentiate between visualization and visual analytics and thereby motivating its necessity. Thereafter the visual analytics process is described and technical as well as social challenges of visual analytics are discussed.

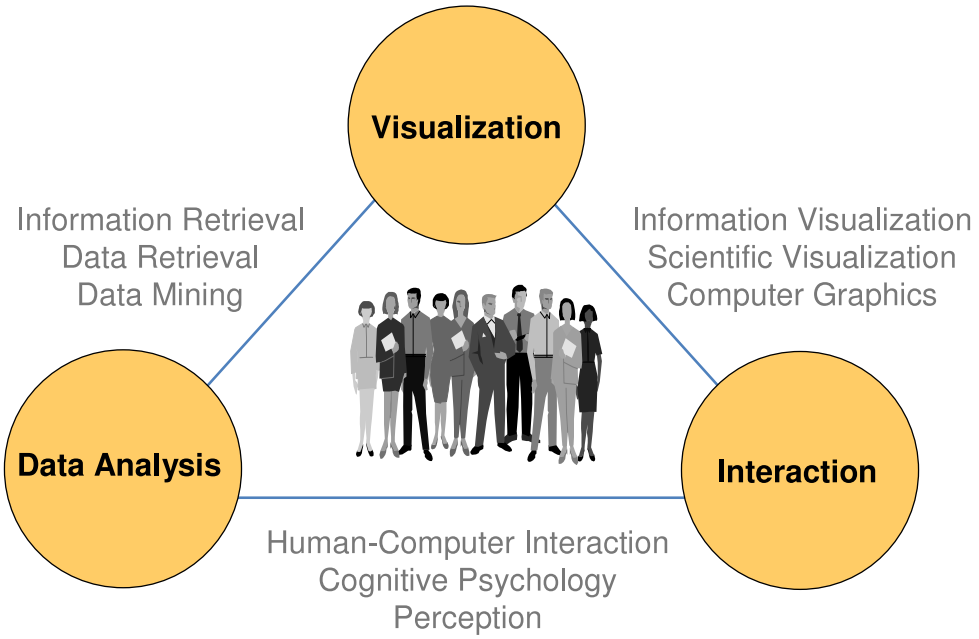

Figure 1: Visual analytics as the interplay between data analysis, visualization, and interaction methods.

Visualization is the communication of data through the use of interactive interfaces and has three major goals: a) presentation to efficiently and effectively communicate the results of an analysis, b) confirmatory analysis as a goal-oriented examination of hypotheses, and c) exploratory data analysis as an interactive and usually undirected search for structures and trends.

Visual analytics is more than only visualization. It can rather be seen as an integral approach combining visualization, human factors, and data analysis. Visualization and visual analytics both integrate 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 matters of data analysis, visual analytics furthermore profits from methodologies developed in the fields of statistical analytics, data management, knowledge representation, and knowledge discovery. Note that visual analytics is not likely to become a separate field of study, but its influence will spread over the research areas it comprises [9].

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. Often the computer system lacks knowledge that is still hidden in the expert's mind. By applying analytical reasoning, hypotheses about the data can be either affirmed or discarded and eventually lead to a better understanding of the data. Visualization is used to explore the information space when automatic methods fail and to efficiently communicate results. Thereby human background knowledge, intuition, and decision-making either cannot be automated or serve as input for the future development of automated processes. In contrast to this, a well-defined problem where the optimum or a good estimation can be calculated by non-interactive analytical means would rather not be described as a visual analytics problem. In such a scenario, the 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 field of visual analytics.

Visual Analytics Process

The visual analytics process is a combination of automatic and visual analysis methods with a tight coupling through human interaction in order to gain knowledge from data. Figure 2 shows an abstract overview of the different stages (represented through ovals) and their transitions (arrows) in the visual analytics process.

In many visual analytics 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 in order to extract meaningful units of data for further processing. Typical preprocessing tasks are data cleaning,

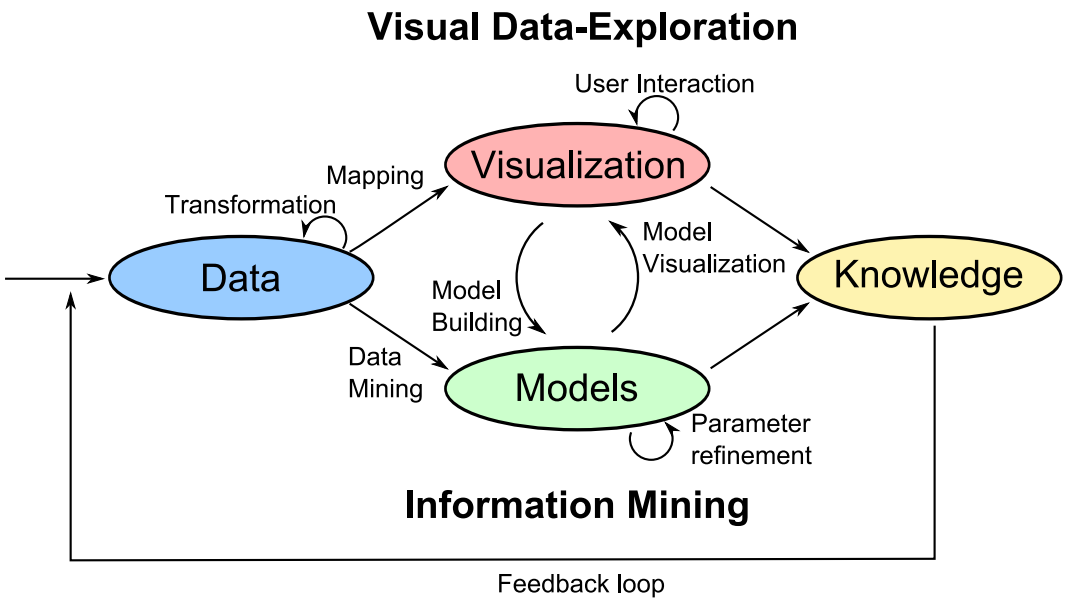

Figure 2: The *Visual Analytics Process* is characterized through interaction between data, visualizations, models about the data, and the users in order to discover knowledge.

normalization, grouping, or integration of heterogeneous data into a common schema.
Continuing with this meaningful data, the analyst can select between visual or automatic analysis methods. After mapping the data the analyst may obtain the desired knowledge directly, but more likely is the case that an initial visualization is not sufficient for the analysis. User interaction with the visualization is needed to reveal insightful information, for instance by zooming in different data areas or by considering different visual views on the data. In contrast to traditional information visualization, findings from the visualization can be reused to build a model for automatic analysis. As a matter of course these models can also be built from the original data using data mining methods. Once a model is created the analyst has the ability to interact with the automatic methods by modifying parameters or selecting other types of analysis algorithms. Model visualization can then be used to verify the findings of these 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, which leads to more confidence in the final results.
In the visual analytics process, knowledge can be gained from visualization, automatic analysis, as well as the preceding interactions between visualizations, models, and the human analysts. The feedback loop stores this knowledge of insightful analyses in the system and contributes to enable the analyst to draw faster and better conclusions in the future.

Technical and Social Challenges

While visual analytics profits from the increasing computational power of computer systems, faster networks, high-resolution displays, as well as novel interaction devices, it must be kept in mind that new technologies are always accompanied with a variety of technical challenges that have to be solved.
Dynamic processes in scientific or business applications often generate large streams of real-time data, such as sensor logs, web statistics, network traffic logs, or atmospheric and meteorological data. The analysis of such *large data streams* which can consist of terabytes or petabytes of data is one of the technical challenges since advances in many areas of science and technology are dependent upon the capability to analyze these data streams. As the sheer amount of data does often not allow to store all data at full detail, effective compression and feature extraction methods are needed to manage the data. Visual analytics aims at providing techniques that make humans capable of analyzing real time data streams by presenting results in a meaningful and intuitive way while allowing to interact with the data. These techniques enable quick identification of important information and timely reaction on critical process states or alarming incidents.

Synthesis of heterogeneous data sources is another challenge that is closely related to data streams because real-world applications often access information from a large number of different information sources including collections of vector data, strings and text documents, graphs or sets of objects. Integrating these data sources includes many fundamental problems in statistics, machine learning, decision theory, and information theory. Therefore, the focus on scalable and robust methods for fusing complex and heterogeneous data sources is key to a more effective analysis process.

One step further in the analysis process, *interpretability and trustworthiness* or the ability to recognize and understand the data can be seen as one of the biggest challenges in visual analytics. Generating a visually correct output from raw data and drawing the right conclusions largely depends upon the quality of the used data and methods. A lot of 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, pre-processing 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 combination). The concrete challenges are on the one hand to determine and to minimize these errors on the pre-processing side, and a flexible yet stable design of the visual analytics applications to cope with data quality problems on the other hand. From a technical point of view the design of such applications should either be insensitive to data quality issues through usage of data cleaning methods or explicitly visualize errors and uncertainty to raise awareness for data quality issues.

In many scenarios, interpreting the raw data only makes little or no sense at all if it cannot be embedded in context. Research on *semantics* may derive this context from meta data by capturing associations and complex relationships. Ontology-driven techniques and systems have already started to enable new semantic applications in a wide span of fields such as bioinformatics, web services, financial services, business intelligence, and national security. Research challenges thereby 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.

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. Often, the huge amount of data that has to be visualized exceeds the limited amount of pixels of a display by several orders of magnitude. In order to cope with such a problem not only the absolute data growth and hardware performance have to be compared, but also the software and the algorithms to bring this data in an appropriate way onto the screen. As the amount of data is continuously growing and the amount of pixels on the display remains rather constant, the rate of compression on the display is steadily increasing. Therefore, more and more details get lost. It is an essential task of visual analytics to create a higher-level view onto the dataset, while maximizing the amount of details at the same time.

The field of *problem solving, decision science, and human information discourse* constitutes a further visual analytics challenge since it not only requires understanding of technology, but also comprehension of typical human capabilities such as logic, reasoning, and common sense. 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. The dynamics of information that changes over time should not be underestimated since it might have a strong impact on the right decision. Furthermore, social aspects such as decision making in groups make the process even more delicate.

While many novel visualization techniques have been proposed, their wide-spread usage has not taken place primarily due to the users' refusal to change their daily working routines. *User acceptance* is therefore a further visual analytics challenge since the advantages of developed tools need to be properly communicated to the audience of future users to overcome usage barriers and to tap the full potential of the visual analytics approach.

Visual analytics tools and techniques should not stand alone, but should *integrate* seamlessly into the applications of diverse domains, and allow interaction with existing systems. Although many visual analytics tools are very specific (i.e., in astronomy or nuclear science) and therefore rather unique, in many domains (e.g., business or network security applications) integration into existing systems would significantly promote their usage by a wider community.

Finally, *evaluation* as a systematic analysis of usability, worth, and significance of a system is crucial to the success of visual analytics science and technology. During the evaluation of a system, different aspects can be

considered such as functional testing, performance benchmarks, measurement of the effectiveness of the display in user studies, assessment of its impact on decision-making, or economic success to name just a few. Development of abstract design guidelines for visual analytics applications would constitute a great contribution.

KEY APPLICATIONS*

Visual analytics is essential in application areas where large information spaces have to be processed and analyzed. Major application fields are *physics* and *astronomy*. Especially the field 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, form continuous streams of terabytes of data that can be recorded and analyzed. With common data analysis techniques, astronomers can separate relevant data from noise, analyze 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 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 visualize the global warming, melting of the poles, the stratospheric ozone depletion, as well as hurricane and tsunami warnings.

In the domain of *emergency management*, visual analytics can help determining the on-going progress of an emergency and identifying 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. bird flu), but also human-made technological catastrophes like industrial accidents, transport accidents or pollution.

Visual analytics for *security* and *geographics* 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 large data sets, like flight customer data, GPS tracking or IP traffic data.

In *biology* and *medicine*, computer tomography, and ultrasound imaging for 3-dimensional digital reconstruction and visualization produce gigabytes of medical data and have been widely used for years. The application area of bio-informatics uses visual analytics techniques to analyze 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 even enlarge the amount 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 areas that are not promising.

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 every day, which accumulate to extremely high data volumes throughout the years. The main challenge in this area is to analyze 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, detection of money laundering, or the analysis of customer data, insurance data, social data, and health care services.

CROSS REFERENCE*

Cluster Vizualization
Comparative Visualization
Data Mining
Data Visualization
Human-Computer Interaction
Multidimensional Visualization Methods

Multivariate Visualization Methods
Parallel Visualization
Scientific Visualization
Visual Classification
Visual Clustering
Visual Content Analysis
Visual Data Mining
Visual Metaphor
Visual On-line Analytical Processing (OLAP)
Visualization for Information Retrieval

RECOMMENDED READING
Between 5 and 15 citations to important literature, e.g., in journals, conference proceedings, and websites.

[1] S. W. Card, J. D. Mackinlay, and Ben Shneiderman, editors. *Readings in information visualization: using vision to think*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999.

[2] Chaomei Chen. *Information Visualization - Beyond the Horizon*. Springer, 2nd edition, 2004.

[3] D. A. Keim. Visual exploration of large data sets. *Communications of the ACM (CACM)*, 44(8):38–44, 2001.

[4] D. A. Keim and Jim Thomas. Scope and challenges of visual analytics, 2007. Tutorial at IEEE Visualization, http://infovis.uni-konstanz.de/tutorials/.

[5] Robert Spence. *Information Visualization - Design for Interaction*. Pearson Education Limited, 2nd edition, 2006.

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

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

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

[9] Pak Chung Wong and Jim Thomas. Visual analytics - guest editors' introduction. *IEEE Transactions on Computer Graphics and Applications*, 24(5):20–21, September/October 2004.