Solving Problems with Visual Analytics:
Challenges and Applications

Daniel Keim
University of Konstanz, Germany
Keim@uni-konstanz.de

Leishi Zhang
University of Konstanz, Germany
Leishi.Zhang@uni-konstanz.de

ABSTRACT
Never before in history data has been generated and collected in such high volumes as it is today. Keeping up to date with the flood of data, using standard tools for data analysis and exploration, is fraught with difficulty. Visual analytics seeks to provide people with better and more effective ways to understand and analyze large datasets, while also enabling them to act upon their findings immediately. The field integrates the analytic capabilities of the computer and the abilities of the human analyst, allowing novel discoveries and empowering individuals to take control of the analytical process. In this paper we present the challenges of visual analytics and exemplify them with a couple of application examples that illustrate the existing potential of current visual analysis techniques but also their limitations.

Categories and Subject Descriptors
H.5.2 **[Information Interfaces and Presentation**]: User Interfaces -*Theory and methods*

General Terms
Human Factors

Keywords
Visual Analytics

1. INTRODUCTION
One of the greatest challenges in the 21st century is to handle massive data collections. The fast development of data storage devices and means to collect data enables people in science and business domains to gather vast amounts of information from different sources. For instance, by 2003 information on 20,000-25,000 genes in the human DNA and the sequences of the 3,000,000,000 chemical base pairs that make up the human DNA had been collected by the Human Genome Project to help understanding how the human body functions [1]. In a different application, every day about 100,000 articles from over 2,500 news sources in 43 languages are collected by the Europe Media Monitor (EMM) to aggregate worldwide news and issue alerts [2]. These data are rich information sources that can help to support scientific discovery and decision making. However, extracting meaningful knowledge from such data is challenging. People are often confronted by disparate, conflicting, and dynamic information from multiple heterogeneous sources and get lost due to the lack of ability to analyze it. This is the well-known *information overload problem*.

As a young upspring science field that aims at tackling the information overload problem, *visual analytics* "combines *automated data analysis* with *interactive visualizations* for an effective understanding, reasoning, and decision making on the basis of very large and complex datasets" [3]. The field utilizes both the computational power of computers and the innate human ability to visually perceive patterns and trends to help people extract meaningful patterns from data that are too large or too complex to be handled by automated data analysis methods or visualization alone. The essential idea is to develop computer based intelligent systems that allow human analysts to examine the massive information stream at the right level of abstraction through appropriate visual representations and to take effective actions in real-time.

Visual analytics is essential in many application areas where large information spaces have to be processed and analyzed. However, challenges exist in many aspects of the field alongside the opportunities. In this paper, we discuss these challenges and exemplify them with some real world applications. In section 2, we briefly introduce the process and paradigm of visual analytics. In section 3, we detail the challenges of visual analytics in the technical aspects of the field. In section 4, we use two example applications to illustrate the potential of current visual analytics techniques, and in section 5 we provide some concluding remarks.

2. VISUAL ANALYTICS
Visual analytics is a multidisciplinary approach which takes advantage of various related research areas such as *visualization, data mining, data management, data fusion, statistics,* and *cognition science*.

2.1 Process
Figure 1 illustrates the process of visual analytics. First of all, heterogeneous data sources need to be processed and integrated. Automated analysis techniques can then be applied to generate models of the original data. These models can be visualized for evaluation and refinement. In addition to checking the models, visual representations can be abstracted from the data using a variety of interactive visualization techniques that are best suited for the specific data type, structure, and dimensionality. In the visual analytics process, knowledge can be gained from visualization, automatic analysis, as well as the interactions between visualization, models and the human analysts.

The feedback loop stores this knowledge of insightful analysis in the system and assists the analyst in drawing faster and better conclusions in the future. An important aspect of visual analytics is user interaction: analysts should be able to query and explore, often multiple abstract views of the data, with rapid feedback, and in the same way steer the analysis process by modifying parameters or choosing alternative analysis methods.
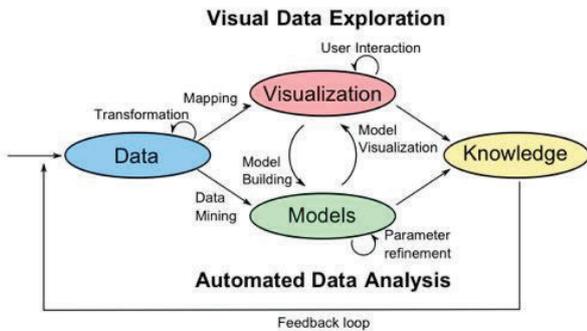

Figure 1. Visual Analytics Process [3]

2.2 Paradigm

The main objective of visual data analysis is to present, navigate, aggregate, and see the details of the data such that complex questions can be answered. Shneiderman [4] proposed the well-known information seeking paradigm "*Overview First, Zoom and Filter, Details on Demand*". The "golden rule" starts with an overview for a better orientation, lets the user decides how to filter interesting data, and shows details only on demand for selected data. It has successfully guided the design of many visualization systems and techniques. For visual analytics, we extend Shneiderman's paradigm: "*Analyze First; Show the Important; Zoom, Filter and Analyze Further; Details on Demand*" [5]. The paradigm brings more analytical skills into the loop before and after the generation of interactive visual representations to help extract abstract models from data sets that are too large or too complex to be analyzed in a straightforward manner.

3. CHALLENGES

Visual analytics is a growing field with many open problems and challenges. Many challenges originate from the specific applications of visual analytics. Each individual application has its own practical requirements in its particular problem domain. However, some challenges are common to more than one domain and application:

Scalability: It is difficult to come up with scalable visual analytics solutions with regard to both visual representations and automatic analysis. The solution needs to scale in size, dimensionality, data types, and levels of quality. Effective methods are needed to deal with noisy high-resolution input data as well as continuous input data streams of high bandwidth. The relevant data patterns and relationships need to be visualized on different levels of details, and with appropriate levels of data and visual abstraction.

Uncertainty: Dealing with uncertainty in visual analytics is nontrivial because of the large amount of noise and missing values originating from heterogeneous data sources and bias introduced by automatic analysis methods as well as human perception. To face this problem, the notion of data quality and the confidence of the analysis algorithm need to be appropriately represented. The analysts need to be aware of the uncertainty and be able to analyze quality properties at any stage of the data analysis process.

Hardware: More efficient computational methods and powerful hardware are needed to support near real time data processing and visualization for large data streams. In addition to high-resolution desktop displays, advanced display devices such as large-scale power walls and small portable personal assistants 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.

Interaction: Novel interaction techniques are needed to fully support the seamless intuitive visual communication with the system. The analyst should be able to fully focus on the task at hand and not be distracted by overly technical or complex user interfaces and interactions. User feedback should be taken as intelligently as possible, requiring as little user input as possible. Such interactions guarantee the full support of the user in navigating and analyzing the data, memorizing insights and making informed decisions.

Evaluation: Due to the interdisciplinary nature and complex visual analytics process, it is hard to assess the quality of visual analytics solutions. A theoretically founded evaluation framework needs to be developed to assess the effectiveness, efficiency and user acceptance of new visual analytics techniques, methods, and models. Such a framework will lead to a better understanding of the field and more successful and efficient development of innovative methods and techniques.

Infrastructure: So far, most current visual analytics solutions develop their own infrastructures for solving their specific problems. Although some systems can be connected through various communication mechanisms such as direct library linking and web services, there is still a mismatch between the level of service provided and the real need for visual analytics in terms of fast and precise answers with progressive refinement, incremental re-computation, and steering the computation towards data regions that are of higher interest to the user. More research is needed to develop a high level infrastructure to bind together all the processes, functions, and services supplied by various disciplines. There is also a need to build repositories of available visual analytics solutions to ensure that common components are reusable.

4. APPLICATION

Visual analytics is a highly application oriented discipline driven by practical requirements. In this section, we use two visual analytics applications to exemplify the possibilities of combining human analytical skills and computer intelligence to solve complex problems.
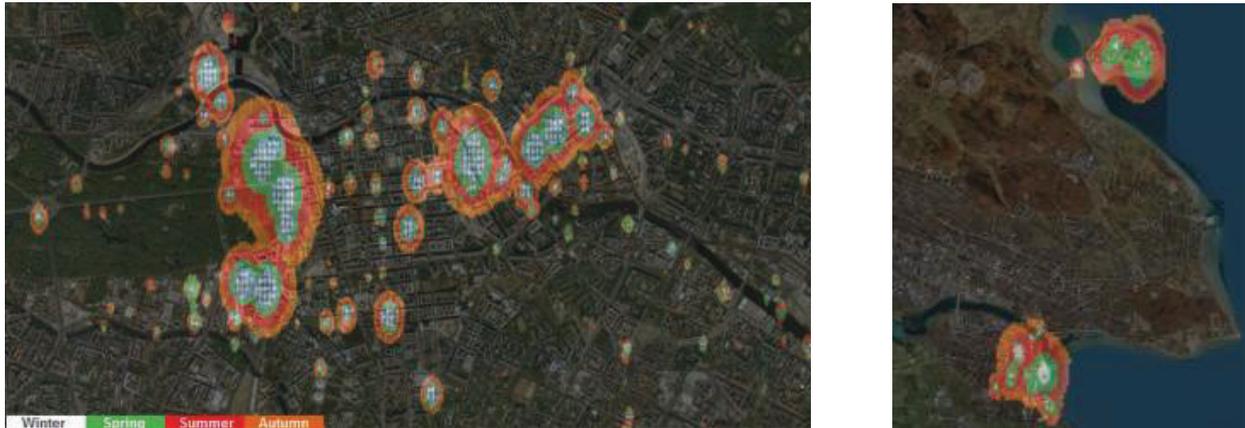
Figure 2. Location and seasonal differences of photos taken in Berlin (left) and Konstanz (right): more places of interest are identified in Berlin, and the small size of the white region in Konstanz indicates that this region is mostly visited during the warm seasons, whereas Berlin is visited all year round [6].

4.1 Spatial-temporal Data Analysis

Spatial-temporal data analysis works on data with contain both space and time dimensions, for example, telecommunication traffic data or satellite image data. Spatial data contain measurements together with links to geographic location. Finding spatial relationships and patterns within the data is of particular interest, requiring the development of appropriate management, representation and analysis functions. Temporal data on the other hand records measurements taken over time. Important tasks here include identification of patterns, trends, and correlations of the data items over time. The analysis of data with both spatial and temporal dimensions adds complexities of scale and uncertainty.

Our first application analyzes concentrations and movements of tourists in a city, based on GPS-referenced photos taken at different geographical locations and published online by the tourists. Such data are available from free online photo album services such as Flickr (www.flickr.com) or Panoramio (www.panoramio.com). The Growth Ring Map [6] is a newly developed technique that supports the exploration of both the frequencies and temporal patterns of events occurring at close-by places. The method defines significant places based on the density of a spatial clustering of the photos. To visualize all photos taken at one significant place, all photos taken at the place are first sorted by their time stamps and then projected on top of the map. Each photo is represented as a colored pixel around the central point (where the photo was taken) in an orbital layout, the older the photo, the closer it is to the central point. Color-hue is used to map semantic properties of the photo, time or place. For example, seasonal differences in visiting the places may be investigated by mapping the seasons to four distinct colors (winter-white, spring-green, summer-red, and autumn-blue). The resulting map shows both the intensity of photos taken at different locations and the seasonal differences (see Figure 2 for an example).

4.2 News Stream Data Analysis

The analysis of large quantities of news is an interesting application of visual analytics. A large number of news sources

publish thousands of news articles on world events every day and a large number of qualitative and quantitative measures can be calculated from each news article. Making sense of this data is becoming increasingly complex due to the rate of the incoming news, as well as the inherent complexity of analyzing large quantities of evolving unstructured text corpora. Therefore, automated text analysis techniques have to be coupled with user interaction and human knowledge feedback.

Analyzing important people and organizations and their relations to world events is of great interest to users in different application areas. Tracking the appearance of key public figures in the news creates a new and rich information space, in which the temporal context plays an important role. In such cases, it is often necessary to provide an overview of the temporal development of parallel news stories, while allowing the analyst to focus on each individual news article that is a part of an important story. Our tool employs the novel Cloudline [7] technique, which efficiently utilizes limited screen space to present multiple time series within a focus-plus-context display. In Figure 3, each row represents a politician appearing frequently in the news during February 2011. Each news article is represented by a circle, whose size and opacity are determined by its local temporal density. By using kernel density estimation methods on the news time-stamped data, important clusters with high density can be easily identified. The benefits of the method of bypassing time series aggregation are two-fold: first, fine-grained structure of each time-series is preserved and, second, direct access to the individual items is possible. The tool allows the user to logarithmically distort the timeline and put emphasis on the more recent events. Using magnification lenses with different parameters, a time window of any range can be explored in detail, providing atomic access to each individual news article even within a logarithmically distorted time line. The tool provides flexibility in terms of adapting the technique to different amounts of data, time ranges and tasks. By combining interaction and visual mapping of automatically processed text content, the user is supported by the system in analyzing the evolving corpora and putting each news article into the context of global news trends.
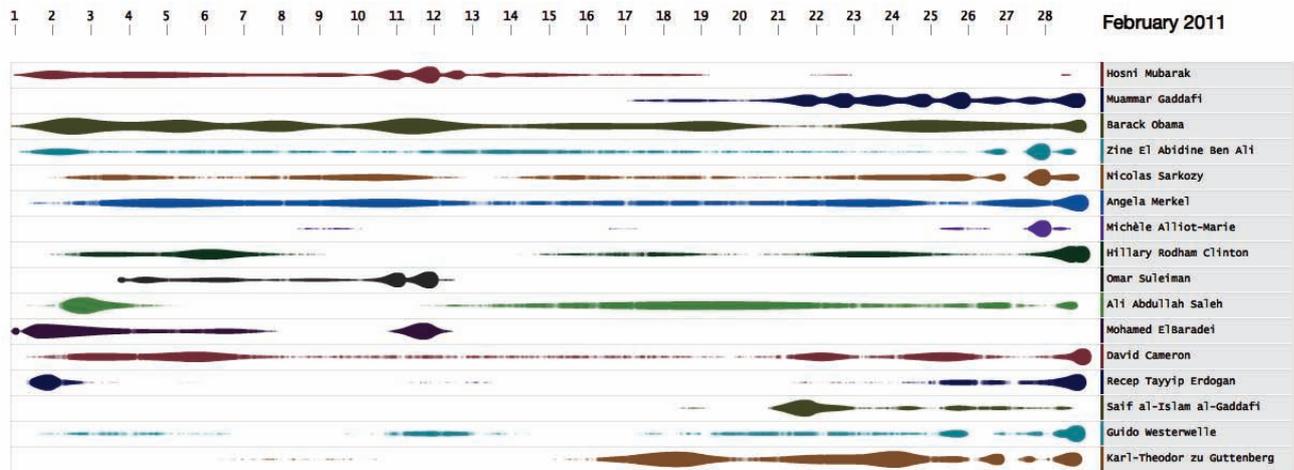
Figure 3. Cloudline: Compact Display of Event Episodes in Multiple Time-Series. The visualization shows multiple politicians appearing frequently in the news during February 2011.

5. CONCLUSIONS

Nearly all grand challenge problems of the 21st century, such as climate change, the energy crisis, the financial crisis, the health crisis and the security crisis, require the analysis of very large and complex datasets, which can be done neither by the computer nor the human alone. Visual analytics is a young active science field that comes with a mission of empowering people to find solutions for complex problems from large complex datasets. By tightly integrating human intelligence and intuition with the storage and processing power of computers, many recently developed visual analytics solutions successfully help people in analyzing large complex datasets in different application domains. However, challenges remain in many disciplines of visual analytics. In this paper, we highlight some of the most important challenges and use two real world applications to show the potential of applying visual analytics techniques to help people synthesize information from heterogeneous sources and derive insight from large data sets.

6. REFERENCES

[1] The Human Genome Project: http://www.genome.gov/.

[2] EMM: http://emm.newsbrief.eu/overview.html. in the same way steer the analysis process by modifying parameters or choosing alternative analysis methods.

[3] D. A. Keim, J. Kohlhammer, et al. (Editors), *Mastering the Information Age - Solving Problems with Visual Analytics*, Eurographics, 2010.

[4] B. Shneiderman, *The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations*, in Proceedings of the IEEE Symposium on Visual Languages, pp. 336-343, Washington. IEEE Computer Society Press, 1996.

[5] D. A. Keim, G. Andrienko, et al., *Visual analytics: Definition, process, and challenges*, Information Visualization, Springer, pp. 154-175, 2008.

[6] G. Andrienko, N. Andrienko, et al., *Analysis of community-contributed space- and time referenced data (by example of Panoramio photos)*, in Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 540--541, 2009.

[7] M. Krstajic, E. Bertini, D. A. Keim, *CloudLines: Compact Display of Event Episodes in Multiple Time-Series*, InfoVis 2011, to appear.