Visual Analytics for the Big Data Era – A Comparative Review of State-of-the-Art Commercial Systems

Leishi Zhang* University of Konstanz, Germany
Andreas Stoffel† University of Konstanz, Germany
Michael Behrisch‡ University of Konstanz, Germany
Sebastian Mittelstädt§ University of Konstanz, Germany
Tobias Schreck¶ University of Konstanz, Germany
René Pompl‖ Siemens AG
Stefan Weber** Siemens AG
Holger Last†† Siemens AG
Daniel Keim‡‡ University of Konstanz, Germany

ABSTRACT

Visual analytics (VA) system development started in academic research institutions where novel visualization techniques and open source toolkits were developed. Simultaneously, small software companies, sometimes spin-offs from academic research institutions, built solutions for specific application domains. In recent years we observed the following trend: some small VA companies grew exponentially; at the same time some big software vendors such as IBM and SAP started to acquire successful VA companies and integrated the acquired VA components into their existing frameworks. Generally the application domains of VA systems have broadened substantially. This phenomenon is driven by the generation of more and more data of high volume and complexity, which leads to an increasing demand for VA solutions from many application domains. In this paper we survey a selection of state-of-the-art commercial VA frameworks, complementary to an existing survey on open source VA tools. From the survey results we identify several improvement opportunities as future research directions.

Index Terms: H.4 [Information Systems]: INFORMATION SYSTEMS APPLICATIONS, K.1 [Computing Milieux]: THE COMPUTER INDUSTRY—Markets

1 INTRODUCTION

We are at the beginning of a big data era when data is generated at an incredible speed everywhere — from satellite images to social media posts, from online transaction records to high-throughput biological experiment results, and from mobile phone GPS signals to digital pictures and videos posted online [3]. According to IBM [9] 2.5 quintillion bytes of data are generated every day. Thus, 90% of todays data has been created in the last two years alone. This phenomenon leads to an increasing interest and effort from both academia and industry towards developing VA solutions with improved performance. On the academic side, a number of advanced VA techniques and open source toolkits have been developed [21]. On the industrial side, a large variety of companies, ranging from specialized data discovery vendors such as *Tableau, QlikTech,* and *TIBCO*, to multinational corporations such as *IBM, Microsoft, Oracle* and *SAP*, have all devoted much effort to develop their own commercial products for analyzing data of increasing volume and variety that arrives ever quicker.

Stakeholders from both academia and industry are well-aware of the importance of gaining an overview of the state-of-the-art solutions to stimulate innovative ideas and avoid redundant effort. Such overview enables people to understand limitations of existing solutions and thus to identify space for improvement. In the last couple of years, effort has been made to survey and compare the functionality of existing open-source VA toolkits [21] as well as commercial Business Intelligence (BI) applications [19, 28]. Such studies are important to assess what tools are available, what techniques they implement, and how good they are with respect to certain application tasks. However, a thorough survey of specific visual analysis functionality of existing commercial VA tools is still lacking, given that the range of tools in existing surveys is restricted to BI applications and focuses on the usability aspects of a product. Towards this end, we conducted a survey on a wider range of commercial VA tools including not only BI VA products but also a number of general purpose VA tools, and put our focus on evaluating their capability of handling data of large volume and variety efficiently. While existing surveys are largely based on user surveys, we devote much effort to evaluate the system performance and functionality by installing the software and testing with reference datasets.

We conducted our survey by first building an encompassing list of 15 relevant commercial systems. The choice is made by investigating current market share. A wide range of systems were selected, covering software that falls into different categories, for example, data discovery and visualization software, enterprise BI systems, network analysis toolkits, innovative and niche products; some products fall into more than one category. We assigned each system a priority level to make sure that we can focus on a smaller number of "core" systems without losing the whole picture. In the second phase, a structured questionnaire was designed for evaluating the functionality of each product from different perspectives, including *data management*, *visualization*, *automatic analysis*, and *system and performance*. We then contacted all vendors to get their answers to our questionnaire. Although many vendors responded with detailed answers, we did not manage to get responses from all of them.

In this paper we report the results for those ten systems whose vendors answered our questionnaire, including *Tableau* [14], *Spotfire* [4], *QlikView* [13], *JMP (SAS)* [11], *Jaspersoft* [10], *ADVIZOR Solutions* [6], *Board* [7], *Centrifuge* [8], *Visual Analytics* [15], and *Visual Mining* [16]. For the remaining systems in our initial list, some of which are regarded as key products in the market (*Cognos (IBM)*, *SQL Server BI (Microsoft)*, *Business Objects (SAP)*, *Teradata*, and *PowerPivot (Microsoft)*, we managed to find many answers to the questionnaire by ourselves, which allows us to gain a better understanding and overview of state-of-the-art VA systems.

*e-mail:leishi.zhang@uni-konstanz.de
†e-mail:andreas.stoffel@uni-konstanz.de
‡e-mail:michael.behrisch@uni-konstanz.de
§e-mail:sebastian.mittelstädt@uni-konstanz.de
¶e-mail:tobias.schreck@uni-konstanz.de
‖e-mail:rene.pompl.ext@siemens.com
**e-mail:stefan_hagen.weber@siemens.com
††e-mail:holger.last@siemens.com
‡‡e-mail:keim@uni-konstanz.de

IEEE Conference on Visual Analytics Science and Technology 2012
October 14 - 19, Seattle, WA, USA
978-1-4673-4753-2/12/$31.00 ©2012 IEEE

But to provide a fair comparison we do not include our findings about those five tools in the survey. This means unfortunately all the systems that support linguistic analysis on text documents (*Business Objects*, *Cognos* and *Teradata*) fall out of the comparison tables. However some of the relevant findings are used to support the analysis and discussion in this paper. To provide further references, we also investigated a number of analytical tools that are known for their text analysis functionality, including *nSpace (Oculus)* [12], *Palentir* [2], and *In-Spire (PNNL)* [1] and integrate some of our findings in the discussion.

In the last phase, further evaluation was carried out on the systems in the top priority list. After installing all the systems on the same machine under the same configuration, we performed a series of loading stress test to check the scalability of each system. The analytical and visualization capability of the selected systems is further tested using two benchmark dataset provided by different research communities representing real-world data analysis challenges.

The main contributions of this paper are: (1) we complement the existing survey of open-source toolkits [21] and user surveys of BI tools [19, 28] by conducting an encompassing survey of commercial VA tools; (2) we structure a comparison of the tools along a harmonized schema; and (3) we draw some careful conclusion and give recommendations to potential users on which tools are applicable for what types of applications. (4) We identify future directions for developing VA systems. The remainder of this paper is organized as follows: In the next section, we discuss related work. In Section 3, we analyze the functionality of each product. In Section 4, we show the result of our data evaluation. We summarize our key findings in Section 5, before drawing conclusion and discussing space for improvement in current commercial products and identifying interesting future directions in Section 6.

2 Related work

In this section, we review work on the definition of VA, existing VA systems and surveys on the market for commercial products.

Visual Analytics Methodology. The VA methodology is based on combining data visualization, data analytics, and human-computer interaction to solve application problems. Its general approach, application examples, and research challenges are detailed in [27, 26]. Recently, the infrastructure working group within the EU VisMaster project [5] identified a number of shortcomings of the current state of application of VA technology in practice [26] (Chapter 6). The lack of standardization in software components, functionality and interfaces was regarded as a major problem, leading to a loss in efficiency and scalability due to massive re-implementation of software components. Hence, standardization was proposed as the key approach to enable a market for software components which eventually should lead to streamlined production of application-oriented VA systems.

Open Source Toolkits. A number of open-source VA toolkits exist; each covers a specific set of functionalities for visualization, analysis and interaction. For example, *InfoVis Toolkit* [18], *Prefuse* [23], *Improvise* [29], and *JUNG* [24]. Using existing toolkits for required functionality instead of implementing from scratch provides much efficiency while developing new VA solutions, although the level of maintenance, development and user community support of open source toolkits can vary drastically. Besides, a relatively high amount of programming expertise and effort is often required to integrate these components into a new system. In [21], a survey of 21 existing open source toolkits is presented. The functionality of these toolkits is compared along three criteria: (1) visualization functions, (2) analysis capabilities, and (3) supported development environment. The aim of the survey is to provide a reference to developers for choosing a base framework for a given problem.

Commercial VA Systems. An alternative is to resort to software suites which integrate required functionality in software systems which work either standalone, or integrate, more or less seamlessly, into an existing information infrastructure. Example systems include *Tableau* [14], *Spotfire* [4], and *QlikView* [13]. Commercial toolkits typically require no or only limited configurations or program adjustments, to become operational. They may provide, subject to the business policy of the vendor, specific levels of maintenance, development and user support. As part of the software market for (corporate) information systems, the BI market segment provides commercial tools for analyzing business data. The BI software market consists of long-standing software suites, which have developed out of core database or statistical data analysis suites. Other products are developed and marketed as standalone tools or add-ons to existing information systems. Common tasks of BI systems include *reporting* of historic and current data, *analysis* (intelligence) of data, and *prediction* including what-if-analysis.

BI System User Surveys. *Gartner Research* surveys the BI software market annually and publish their result online [19]. They maintain a set of 14 functional requirements that BI tools aim at, structured along three categories: (1) integration into existing environments, (2) information delivery and (3) information analysis functionality. A set of 21 products is included in the 2012 survey which outlines the strengths and possible risks of each selected product, relative to the market and product history. A characterization of the 21 products as challengers (2 products), market leaders (8 products), niche solutions (11 products), and visionaries (0) is provided.

In another report, a detailed survey of 16 current BI products is provided by *Passionned Group* [28]. Eight evaluation criteria are defined by the study, ranging from software architecture, functionality, to usability and analytic capabilities. The products are categorized into (1) standalone enterprise-level solutions, (2) BI products which come integrated with database systems software, (3) data discovery and visualization tools, and (4) innovative and niche products. A scoring scheme is defined to compare product along these criteria individually. Also, an all-against-all comparison along aggregated scores is provided.

Open Source and Commercial Tool Landscape. There is a wide spectrum of tools from which VA applications can be built. In general, the open source domain provides state-of-the-art functionality, which may include early and sometimes prototypical techniques. Often a library has to be embedded into a front-end and connected to a back-end data infrastructure, to obtain an end-user application. However we also see exceptions. For example, Gephi, an open source graph visualization tool, also features a rich user front-end interface. Open source tools are mainly developed and maintained on a voluntary basis.

On the other hand, in the commercial sector, we see more conservative visualization techniques, which in most instances are already integrated with user front ends and data back end infrastructure. Whereas in the open source market, development takes place in an open, sometimes unpredictable manner, development in the commercial area takes place under competition, in a closed way, often involving pilot users. Intermediate results are not discussed with the larger public.

Open source tools are freely available, whereas commercial products generally require costly licensing. Licensing fees vary drastically. For an industrial investment decision, the *total cost of ownership* is relevant, which includes roll out, development and adaption, life cycle management, and user training, among other factors. It depends also on the environment in which the tools are deployed. The discussion of this is beyond the scope of this work. To determine the total costs a consultancy process is required, involving users, vendors, and business process specialists.

Table 1: Data Handling Functionality

	Usability		Preprocessing and Data Handling			
	Import User Guidance	Collaborative Working/ No. of Users	Preprocessing (Semi-/Automatic/ Expression Lang.)	Column calculations/ Column and Row Combinations	Joints/ Joints on Filtered Tables	Querying functions (Group-by, Sum, Average, Count, Ordering)
Tableau	DnD, Wizards, Previews, Type Guessing	✓, Unlimited	✓, - ,✓	✓, ✓	(✓, ✓)*1	✓, ✓, ✓, ✓, ✓
QlikView	DnD, Wizards, Previews, Type Guessing	✓, Unlimited	✓, - ,✓	✓, ✓	✓, ✓	✓, ✓, ✓, ✓, ✓
Spotfire	Wizards	✓, Unlimited	✓, - ,✓	✓, ✓	(✓, ✓)*1	✓, ✓, ✓, ✓, ✓
JMP	DnD, Wizards, Previews, Type Guessing	-	✓, ✓,✓	✓, ✓	✓, ✓	✓, ✓, ✓, ✓, ✓

DnD: Drag-and-Drop File Import *(...)*1: With Exporting Intermediate Step*

In this paper we concentrate on a functional comparison of a selected number of tools. We relate our work with the existing surveys as follows. Gartner reports and Passionned survey aim at providing an overview of functionality of major BI products as a reference to potential customers and market analysts. The result is largely based on feedback from current users, although the vendors are contacted to supply additional information (business strategy, vision, etc.). We take a rather different perspective and approach - we survey the identified vendors with a structured questionnaire consisting of questions covering different aspect of system performance and functionality, and test-driving the selected toolkits in a standardized environment and on benchmark datasets. We also extend the scope of the tool selection by including a number of characteristic VA tools which provide solutions to specific problem domains that are not included in BI tools. The main objective of our survey is to provide an comparative review of the state-of-the-art VA systems and highlight possible technical advances for future research and development.

3 Functional Comparison

Typically, there are three main actions in a VA system work flow, *data management*, *data modeling* and *visualization* [26]. 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 VA process, knowledge can be gained from visualization, automatic analysis, as well as the interactions between visualization, models and the human analysts.

Based on the evaluation strategy described in section 1, a structured questionnaire consisting of 52 questions was designed to evaluate the functionality of each system (see Appendix 1). Questions are categorized into 4 classes in order to cover the three main actions in a system work flow as well as the system performance: *data management*, *automatic analysis*, *visualization*, and *system and performance*. The questionnaire was sent to 15 different vendors and 10 answers were received.

Among the 10 systems, 4 fall into the top priority list: *Tableau*, *Spotfire*, *QlikView*, and *JMP*. We managed to acquire academic or evaluation licenses from each vendor and evaluated the functionality and performance of the four systems further by installing each system and testing with real data. In addition, we verified the information provided by vendors wherever possible. Next we detail our results.

3.1 Data Management

Following the Knowledge Discovery in Databases pipeline defined by Fayyad et al. [17], the primary steps for VA tools are data loading, integration, preprocessing, transformation, data mining, and data interpretation. In a data management related functional comparison of commercial VA tools one can subsume all data loading, integration, and exporting options under *data management functionality*. Operational steps, such as data preprocessing or transformation, as well as their relation to usability aspects can be classified as *data handling functionality*.

Regarding data management, all VA systems allow connecting to relational database systems, such as SQL, PostgreSQL, and Oracle. But only a few tools allow access to vertically scalable storage system, such as Hadoop, Vertica (Column-oriented), and MongoDB (Document-oriented), or web-based on-demand database systems, such as Amazon S3 and Salesforce Database System (None-SQL, Object-oriented).

The import of raw (structured or unstructured) data files was assessed too. The most prominent data file formats, which are Microsoft Excel and plain text file (CSV), are supported by all assessed tools. Yet, only a few tools import dedicated geo-related files, such as ESRI or Google's KML, or allow to process the content of Adobe PDF or Microsoft Word files.

Another data management aspect is related to the simultaneous access to multiple data sources. In a data warehouse scenario, the analyst often needs to access various distributed databases. In most systems, multiple data connections can be maintained. However, to use some of the dashboarding facilities, a data unification batch needs to be processed to consolidate the data sources.

The data/result exporting is the final step in the data analysis pipeline. It serves the purpose of presenting results to a broader audience or save intermediate results. In the latter case, it is often necessary to write results back into the databases. Yet, this data handling mechanism is rarely implemented. Only *Tableau, JMP*, and *Visual Analytics* support a direct database write-back. The obvious standard way to present results is via (interactive) dashboards either hosted on-premise (on a company's secured local server) or on the VA producer's public gallery, via HTML or Adobe Flash Websites.

Mobility is one of the hot topics for commercial VA systems. *Tableau*, *Spotfire*, *QlikView*, and *JMP* take advantage of their underlying presentation platform and offer Apple iPad apps for accessing interactive dashboards in meetings, at customer sites and at operation centers. Another approach towards mobility is the presentation through HTML5-capable browser engines (e.g. Android/BlackBerry/Nokia built-in browsers support HTML5).

The next functional comparison is related to all mandatory data handling steps during data transformation. Table 1 emphasizes two aspects. First, it depicts a use case oriented data handling comparison of the four tools that fall into our top priority list (*Tableau, QlikView, Spotfire,* and *JMP*). And second, it gives an insight into the data handling usability and feature richness.

After the loading procedure, a data cleaning and transformation step is often needed. For example, handling missing/null values and

Table 2: Automatic Analysis Methods

	Statistics			Data Modelling				Data Projection			Visual Query Analysis
	Univariate	Bivariate	Multivariate	Clustering	Classification	Network Modelling	Predictive Analysis	PCA	MDS	SOM	
Tableau	✓	✓	-	-	-	-	-	-	-	-	-
QlikView	✓	✓	(✓)*	(✓)*	(✓)*	-	(✓)*	(✓)*	(✓)*	(✓)*	-
Spotfire	✓	✓	(✓)*	P / H	(DT, NB, ANN)*	-	*(AR,HW)**	(✓)*	(✓)*	-	(✓)*
JMP	✓	✓	✓	P / H	DT, ANN	-	✓	✓	-	✓	✓
Jaspersoft	✓	✓	-	-	-	-	-	-	-	-	-
ADVIZOR	✓	✓	✓	-	SVM		MVLR	-	-	-	✓
Visual Analytics	✓	✓	-	P / H	-	✓	-	-	-	-	✓
Centrifuge	✓	✓	-	P / H	-	✓	-	-	-	-	✓
Visual Mining	✓	✓	-	-	-	-	-	-	-	-	-
Board	✓	✓	-	-	-	-	-	-	-	-	-

(...): only with additional upgrades*
DT, NB, ANN: decision tree, naive bayes, artificial neural network
SVM: support vector machine
P / H: partitioned based clustering / hierarchical clustering
AR, HW: ARIMA, Holt-Winters
MVLR: multivariate linear regression

normalizing data over one or more dimensions. Most commercial VA systems provide the user the option of manipulating data with a proprietary expression language. For example, Tableau patented in 2003 *VizQL* [20], a structured, declarative query language that translates user-actions into database queries and handles the mapping of the results to their visual representations.

Since data preprocessing can range from data sampling or filtering, to more sophisticated approaches such as binning or outlier detection, we decided to derive different data handling tasks that occur in most data analytics tasks. The first one, called *Column calculations*, describes a batch modification of every row record in a selected column, for example, string to date conversion or numerical columns scaling. *Combining columns or rows*, into a single column/row, is another required data handling step. More related to the analytical part of data analysis is the task *Joins/Joins on Filtered Tables*. Most of the commercial VA systems have difficulties in combining tables that are filtered according to the user's needs. Accordingly, the user has to overcome these problems by exporting the filtered table, reloading it from file, and doing the join operation as a distinctive intermediate step.

3.2 Automatic Analysis Methods

Various techniques for automatic analysis of data exist, ranging from simple aggregation to advanced data modeling algorithms. In our survey, we divide automated analysis functions that are implemented by the investigated systems into four categories: *statistics*, *data modeling*, *dimensionality reduction*, and *visual query analysis*.

The first category includes statistics functions for: 1) *univariate analysis* that operates on one dimensional data, for example the calculation of the *mean*, *minimum* and *maximum*, and *standard deviation*; 2) *bivariate analysis* that reveals interrelations of two dimensions, for example, *Pearson correlation* and *Spearman's rank correlation coefficient*; and 3) *multivariate analysis* that models the relations over multiple dimensions, for example, *discriminant analysis* and *variance analysis*. These functions provide different levels of statistical analysis and allow the user to explore the data and relations from different perspectives. As shown in Table 2, all the systems provide some simple statistics methods for univariate and bivariate analysis, but multivariate analysis is only supported by *Spotfire*, *JMP* and *ADVIZOR*.

Methods in the second category allow the user to model the data and find patterns using various data mining algorithms. Most commonly implemented algorithms include: 1) *clustering* algorithms that group data items based on their similarities; 2) *classification* algorithms that assign data items into different classes based on training data with class labels for each data item; 3) *network modeling* techniques that model the relationships between data items as a network (graph), where nodes represent entities (e.g. persons, organizations) and links represent relationships (e.g. co-authors, friends); 4) *predictive modeling* techniques that analyze current and historical facts to make predictions about future events. Note that with *Spotfire* some of the automatic analysis methods are only available with additional upgrades.

The third category describes dimension reduction techniques that can be applied to transform high dimensional data into lower dimensional space. Such transformation leverages the dimensionality problem by reducing the number of dimensions prior to analysis or visualization while keeping the essence of the data intact. The result is often used to generate 2D or 3D projections (typically scatter plots) of the data. The commonly used dimension reduction techniques are *Principle Component Analysis (PCA)*, *Multidimensional Scaling (MDS)* and *Self Organizing Map (SOM)*.

Among all the systems, *Visual Analytics* and *Centrifuge* are the only two that support network modeling. Both systems also support cluster analysis on the networks. *JMP* and *Spotfire* appear to cover all the other data modeling functionalities. They are also the only two systems that implement dimension reduction techniques for handling high-dimensional data.

Another useful feature for automatic data analysis is pattern search. Given a target pattern, an automatic searching mechanism can be designed to look for similar patterns in the data. Some systems enable the user to define a target pattern with the help of the graphical user interface. Once a pattern is defined, the system will automatically search for similar patterns and visualize the results accordingly. We call such functionality *visual query analysis* and use it as the fourth category. Such functionality is favorable to many users as it provides a fast and intuitive means of pattern analysis. Surprisingly only half of the system we surveyed support the visual query analysis (see Table 2).

3.3 Visualization Techniques

To analyze the visualization functionality of each system, we divide visualization techniques into *graphical representations* of data and *interaction techniques*. The former refer to the visual form in which the data or model is displayed, for example, a bar chart or a line chart. Graphical representations are often also called "visualizations" by the tools, and often refer to the static graphical models representing the data. Interaction techniques describe how the user can interact with the graphical models, for example, zooming or panning, and has to be implemented on top of one or more graphical representation to provide users with more freedom and flexibility while exploring graphical representations of the data. In this section we analyze which of these two types of visualization techniques are supported by each surveyed product and detail our findings.

On a high level, we classify the visualization techniques by the

Table 3: Visualization techniques

	Numerical Data					Geo-related Data	Network Data	
	Bar- line-pie- Chart Histogram	Scatterplot	Heatmaps	Parallel Coordinates	Scatterplot Matrix	Projection on Map	Treemap	Other Graphs
Tableau *	✓	✓	✓	(✓)	✓	✓	-	-
QlikView	✓	✓	(✓)	(✓)	✓	(✓)	✓	-
Spotfire	✓	✓	✓	✓	✓	✓	✓	✓
JMP *	✓	✓	✓	✓	✓	✓	✓	-
Jaspersoft	✓	✓	✓	-	-	✓	-	-
ADVIZOR	✓	✓	✓	✓	✓	✓	-	-
Visual Analytics	✓	✓	✓	-	-	✓	-	-
Centrifuge	✓	✓	✓	-	-	✓	-	✓
Visual Mining	✓	✓	✓	(✓)	-	✓	-	-
Board	✓	✓	✓	(✓)	-	✓	-	-

(✓): *not available as default, user interaction required (eg., transform line charts to parallel coordinates)*
* *tool that suggests appropiate visualizations to the user*

type of visualized data: 1) *numerical data*; 2) *text/web*; 3) *geo-related data*; and 4) *network data (graph)*. On a lower level, we investigate individual graphical representations implemented by the surveyed systems to visualize different types of data. For example, for visualizing numerical data, a large number of techniques exist, from *bar chart*, *line chart*, *pie chart* and *scatter plots*, which are often used to visualize numerical data with few dimensions, to *parallel coordinates*, *heatmaps*, and *scatter plot matrix*, which are used for displaying data with higher dimensionality.

Text/web data visualization is a relatively new field, with techniques such as *word cloud* [25] and *theme river* [22] having been developed in recent years. The generation of more and more geo-tagged data increases the demand for geo-spatial data visualization. Often the analyst wants to see geo-related information projected on a conventional 2D map or 3D globe.

Another important branch are graph visualizations, which are widely used for displaying relationships in data and which are applied in emerging fields such as social network analysis and biological regulatory network analysis. Depending on whether there is a hierarchical relation in the graph data, the field can be further divided into hierarchical and non-hierarchical graph visualization. While many force-directed placement techniques can be applied to visualize graphs in general, a number of techniques exist for visualizing graphs with a hierarchical structure, for example, the treemap and the hyperbolic view.

Surprisingly, the number of visualization techniques that are implemented by the surveyed VA systems is rather small compared to the number of techniques that are available from research. Table 3 shows the main visualization techniques that are implemented by (at least one of) the products we surveyed.

As we can see from the result, all products implement standard visualization techniques such as line charts, bar charts, pie charts and histograms. These techniques are commonly used to analyze data with very few dimensions. Scatterplot, scatterplot matrices and heatmaps can be found in most of the tools for analyzing data with higher dimensionality. But to our surprise only few products implement the parallel coordinates visualization, which is considered to be effective for visualizing high dimensional data. Also none of the systems provide functionality for textual data visualization (therefore we removed the column from the comparison table).

In terms of network analysis, only *QlikView*, *Spotfire*, *JMP*, *Visual Analytics* and *Centrifuge* provide functionality for visualizing network data. In addition, functionality for visualizing geo-related data is rather limited in many systems, although most of them do allow the user to project data on top of a static map.

Both *Tableau* and *JMP* implement recommendation facilities which suggest suitable visualizations for the input data. This is very helpful in the initial analysis, especially for people who are not familiar with visualization techniques or the data. These products are marked with * in Table 3.

For most visual analytics tasks it is essential to interact with the data and visualization models. For example, to *filter* the data, to *drill down* to a subset of the dimensions or data items, to *zoom and pan* the view to see the visualization model at different levels of detail, to interactively change the focus of the view without losing the whole picture (*focus+context* distortion techniques), and to *link and brush* different views to see the data from different perspectives.

Most of the tools we surveyed support interactive filtering and zooming as well as the distortion of views (e.g. logarithmic scale). Providing multiple views simultaneously connected by linking-and-brushing functionality is one of the most effective approaches and a major strength of some tools.

3.4 System and Architecture

In addition to the functional characteristics of the VA tools, several non-functional features determine its usability. For example, platform, scalability and architecture. Another important non-functional characteristic is security with respect to data transmission, collaborative working environment, anonymization and role-based content access. Table 4 depicts the system, architecture and security features of the surveyed systems.

According to our findings, VA systems can be subdivided into stand-alone desktop programs and server-sided dashboarding tools. However, the architecture has direct impact on the scalability and performance. In case of client-server architectures, dedicated computing server machines can be added to scale to the given processing needs. *Tableau, QlikView* and *Spotfire* support this so-called *vertical scalability*. Of all tools, only *QlikView* and *Jaspersoft's* cloud-based Platform-as-a-Service (PaaS) offering adapts flexibly to the task's processing needs.

The deployment platform is another aspect to consider, especially for medium and large-sized organizations. Most tools support on the client-side Microsoft Windows XP, Vista, and 7. On the server side Microsoft Windows Server 2003/2008 dominate the platform installation environment. Only a few tools allow an installation on Apple MacOS, Linux distributions or are JVM-based (Java Virtual Machine) applications.

As external viewers, browser-based access to HTML5 or Flash-based dashboards are popular. *Tableau*, *Spotfire*, *QlikView*, *JMP* and *Board* go even one step further and offer a dedicated iPad app to take advantage of the underlying mobile platform.

The memory concept also plays an important role for the perfor-

Table 4: Scalability and Performance Functionality

	Usability						Security		
	Stand Alone/ Client-Server/ Cloud	Platforms	External Viewers	Scalability	Memory concept	BI Infrastucture Integration	Role-based content access	Transmission encryption	Anonymization concept
Tableau	✓, ✓, -	Windows XP, Vista, 7, Server 2003/2008	Browser, Apple iPad	✓	In-Memory Engine, Dedicated Comp. Server	SAP	✓	HTTPS, SSL, LDAP, MS Auth.	-
QlikView	✓, ✓, -	Windows XP, Vista, 7, Server 2008	Browser	✓	In-Memory Engine, Dedicated Comp. Server	SAP, Third party	✓	HTTPS, SSL, LDAP, MS Auth.	✓
Spotfire	✓, ✓, -	Windows XP, Vista, 7, Server 2008	Browser, Apple iPad	✓	In-Memory Engine, Dedicated Comp. Server	SAP, Oracle eBusiness, Siebel, Salesforce.com	✓	HTTPS, SSL, LDAP, MS Auth.	-
JMP	✓, ✓, -	Windows XP, Vista, 7, MacOS X	Browser, Apple iPad	-	RAM	SAS BI Interface	✓	HTTPS, SSL SAS On Demand	✓
Jaspersoft	✓, ✓, ✓	Windows XP, Vista, 7, Server 2003/2008, Linux, MacOS X	Browser	✓	In-Memory Engine	Web-Service, REST-API	✓	HTTPS, SSL, LDAP, MS Auth.	-
ADVIZOR	✓, ✓, -	Windows XP, Vista, 7, Server 2003/2008	Browser	-	In-Memory Engine	DB Interface	✓	HTTPS, SSL, MS Auth.	-
Visual Analytics	✓, ✓, -	Java	Browser	-	RAM	DB Interface	✓	AES-256 or SHA-256 Client/Server,	-
Centrifuge	-, ✓, -	Windows XP, Vista, 7, Linux	Browser	✓	In-Memory Engine	DB Interface	✓	-	✓
Visual Mining	-, ✓, -	Windows XP, Vista, 7, Server 2003/2008	Browser	-	RAM	DB Interface	✓	HTTPS, SSL	-
Board	✓, ✓, -	Windows XP, Vista, 7, Server 2003/2008	Browser, Apple iPad, Office Add-In	✓	In-Memory Engine, Dedicated Comp. Server	DB Interface	✓	HTTPS, SSL	-

mance and scalability in terms of processable data size. Nearly all vendors acknowledge this fact and come up with a proprietary in-memory data engine. For example, *QlikView's* patented in-memory data analysis engine assumes a star schema in the data and thus associates fields with the same name in a global and fast array-like data structure. The indexes are determined by parallelized scans, taking advantage of todays multi-core processors. Moreover, it handles caching and query prediction intelligently by taking the cost of a query reconstruction into account, too. Other vendors, such as *Tableau*, *Spotfire*, *Jaspersoft*, *Board* and *ADVIZOR* have their own approaches to the topic. However, their common point is the capability of handling big amounts of data. Despite the great advances in this field one has to acknowledge the fact that sophisticated calculations, especially with a lot of data joins, are still limited by the RAM size and lead to paging.

Security considerations have also to be taken into account. Security is not only regarded as plain transmission security, but also content-wise access security. Role-based content access, which restricts or permits well-defined data views, is implemented in all systems. If the data needs to be published openly, automatic anonymization features are required. In our test it was therefore not assessed, whether the systems allow to modify one or more name columns (e.g. by a hashing algorithm) manually and then create a new anonymized view (file), but rather if this publishing functionality is supported by a built-in export functionality.

4 Benchmarking the System Performance

In addition to surveying the vendors, we further evaluated the functionality and performance of the four systems in our top priority list, *Tableau*, *QlikView*, *Spotfire*, and *JMP*. First we installed the four systems on our local computer under the same system configuration. A use case study is then carried out on the systems using two benchmark datasets 1) the "Practice Fusion Medical Research Data" provided by Microsoft Azure Marketplace representing real-world challenge in health data analysis, and 2) the "Geospatial and Microblogging Data" provided by VAST challenge 2011 representing challenges in spatial-temporal data analysis. The essential idea is to test the analytical and visualization capability of each system. Besides, a series of loading stress tests are applied to test the scalability of each system. Next we detail our findings.

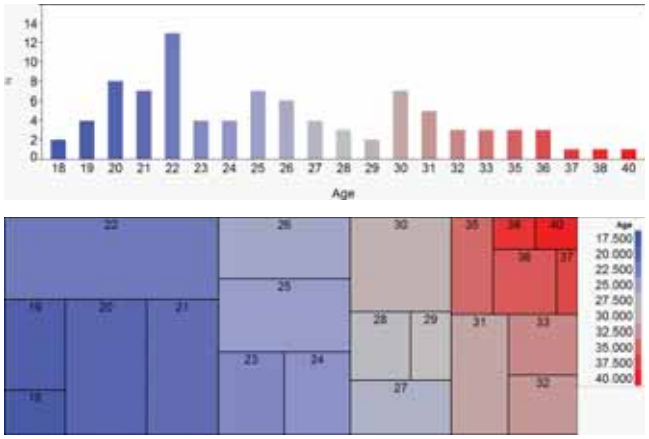


Figure 1: Histogram and treemap visualizations of the pregnancy diagnosis in the "Practice Fusion Medical Research Data" in JMP. The age is mapped into colors in both visualizations. The shares of pregnancy are mapped to the height of bars in the histogram, and the area in the treemap.

4.1 Use Case Study

Practice Fusion Medical Research Data contains a sample of 15,000 de-identified health records stored in 7 different tables, recording information about patients, diagnosis, medications, prescriptions, allergies, immunizations, and vitals respectively. All the tables share a common field *PatientGuide*, which means information in different tables can be linked and aggregated across the different tables.

In our study, we use the data to test the data handling capability of each system, as well as some basic analytical capability with respect to answering simple analysis questions and visualize related information. To achieve this we started from a simple question "What is the distribution of pregnancy age?" and try to find out how easy it is to get the answer using different systems and what type of visualization each system provides.

To answer the question, the data set has to be preprocessed before further analysis. First of all, tables containing patient and di-

agnosis records have to be joined. Next the age of the patient at the moment of the diagnosis need to be calculated based on the year of the diagnosis and the patient's birth year. The last step is to filter out non pregnancy related diagnosis and patients with invalid ages. We had no problem with all the systems during the preprocessing stage. After the filtering, 91 pregnancy diagnoses with a valid age were found among the 77,400 diagnoses in the data.

Using the pregnancy diagnoses records, we tried the basic visualization functionality of each system. First we try to see if we could generate a histogram from the data to show the age distribution over pregnancy. While all the systems were able to render histograms from the data with absolution values (number of pregnancies), creating histograms with percentage values seemed to be more challenging in *Spotfire* and *QlikView* - both systems require additional effort to convert absolute values to percentage before rendering. *Tableau* offers a wizard for creating calculated columns in the visualization, and *JMP* includes a similar aggregation function in the visualization wizard. It is not difficult to find out the answer to our question in the result histograms - the pregnancy age ranges between 18 and 44, and 22 is the peak age that has the highest pregnancy rate.

We further checked the flexibility of customizing visualizations by trying to assign data values to different visual parameters (e.g. color, size) in each system. We tested the possibility of double coding the data values to both height and color of the bars in the histogram. Although this is possible with all the systems, it is relatively easier in *Tableau* and *Spotfire* because the user can change the settings directly on top of the interactive visualization or via menu functions. With *JMP* is less easy, because the system tends to automatically assign the colors of the data column to the corresponding bars in the histogram, and once a visualization is created, it is not possible to change the color encodings unless the user resets the colors in the data column and generates a new histogram. With *QlikView* the user has to define customized functions for assigning colors to bars. This is undesirable to non-programmers, but for users with more programming experience, the system provides much freedom to customize their visualizations. For example, a user defined bi-polar colormap can be generated using some functions in the program library. One slight disadvantage with the current implementation is the fact that the colormap cannot to saved.

Last we try to see the possibility of generating a slightly more "advanced" visualization technique - Treemap with the systems. Except *Tableau*, all the other systems support treemap visualization. The implementation in both *Spotfire* and *QlikView* orders the rectangle in lexical order of the visualized data columns by default. The configuration of the treemap visualization in all cases are similar to the corresponding histogram visualization: while the visualizations in *Tableau* and *Spotfire* are easily configurable, *QlikView* provides less flexibility, although the system does allow the user to write their own functions for changing configurations. With *JMP*, once a visualization is created, modification is restricted. For instance, it is not possible to change the mapping of dimensions in *X* and *Y* axes, however it is easy to create the same visualization with different settings. Figure 1 shows a histogram and a treemap visualization generated by *JMP* as example outputs.

Geospatial and Microblogging Data encodes the characterization of an epidemic spread. Two datasets are included, the first one contains geo-tagged microblogging messages with time stamps, the second one contains map information for the artificial "Vastopolis" metropolitan area. We use the data in our second use case to see how geo-temporal data can be analyzed and visualized in different systems.

As a preprocessing step we transformed each of the 1,023,057 messages into a tabular form containing the timestamp, x-geolocation, y-geolocation and the message text. We store this data in a CSV file for further analysis. In all tools, the overarching analysis goal is to visualize the geo-referenced disease outbreaks over the given time span.

Importing the 185 MB CSV file into the tools worked without any problem. However, only *Tableau* and *Spotfire* recognized the standard date format correctly. *QlikView* and *JMP* required us to define a conversion to their proprietary date format. After loading, the data extraction step requires the calculation of two specific columns: (1) the inversion of the y-coordinate (due to the different notions of origin in the image and the standard Cartesian coordinate system) and (2) the extraction of interesting disease keywords, including "breath", "chest", "diarrhea", "cough", "fever", "flu", "pneumonia", and "sick", in the text. All the tools were able to extract the disease indicators with an if-then-else statement that checks whether the keywords are present or not. However, more sophisticated text analysis/mining features, such as sentiment analysis, stemming or stop word removal, are not present in our packaged versions of the Visual Analytics tools.

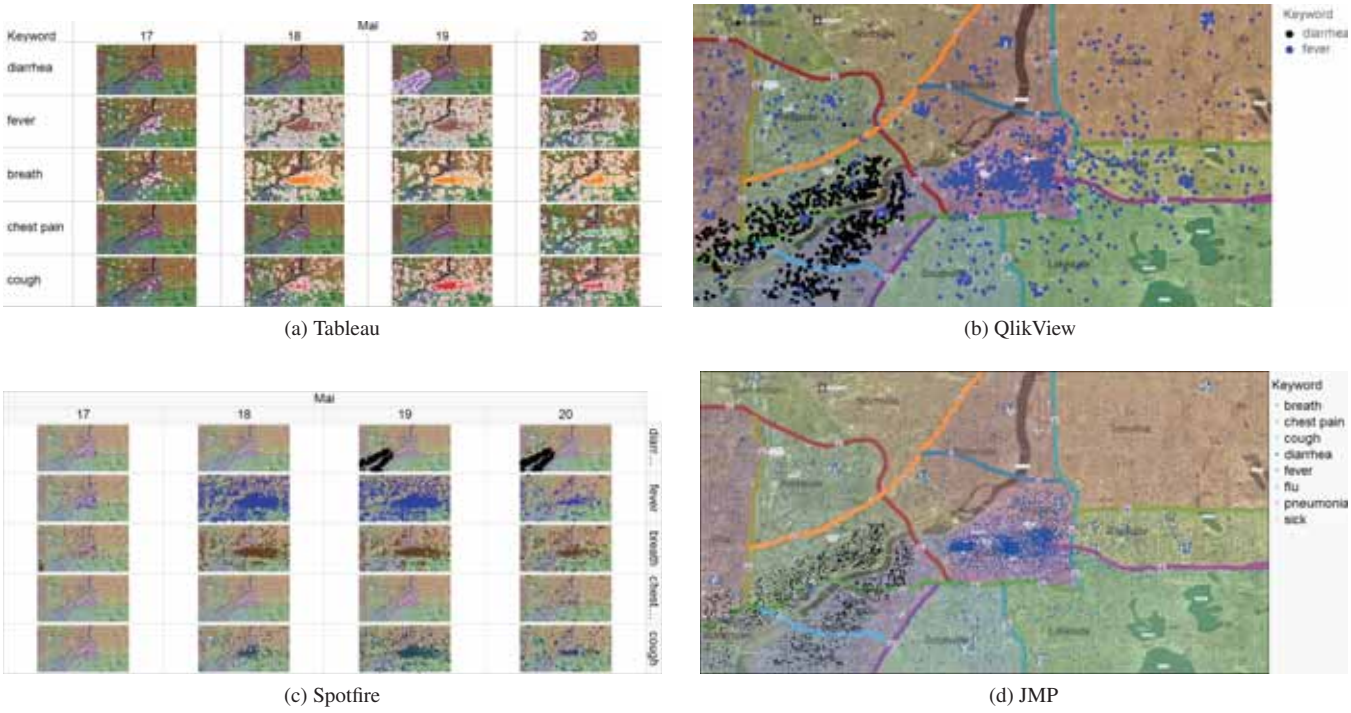
In order to visualize the results, we decided on a small multiple map presentation that takes the geo-spatial, as well as the temporal information into account. Each line in the small multiple view should represent the development of one disease indicator over time. As Figure 2 shows, all tools allowed us to load the data into a 2D scatterplot and set a user-determined background image (the Vastopolis map). Furthermore, none of the tools showed problems with the image space geo-location parameters given in the data set. While the standard interaction paradigm for exploring the data is an on-demand time interval filtering, only *Spotfire* and *Tableau* have a built-in functionality to visualize a series of small multiples with different filtering parameters each. *JMP* and *QlikView*, on the other hand, let the user explore the content differences on a single screen. From a visualization perspective, a small multiple view is one of the best solutions to get an all-embracing overview of the data. However, the high number of interactive screens has an impact on the system's performance. *Spotfire* renders the small multiple screens fast and allows sufficiently fast brushing and linking. *JMP* and *QlikView* also render the single screen fast, but vary greatly in the time needed by brushing and linking.

Some of the known VAST Challenge 2011 findings can be easily retrieved from the map visualizations. For example, in Figure 2 all tools clearly showed the uncorrelatedness of the disease indicators "diarrhea" and "fever", thus leading to the hypothesis of two disease patterns. However, while the small multiple views (a) and (c) give the user the ability to perceive the delayed outbreaks of the two diseases on one screen, (b) and (d) leave the user with the problem of choosing the correct filter predicate to make this observation. Another example: Figure 2 (a) and (c) let the user hypothesize that the wind direction is from west to east, which can be seen in the "breath" outbreak occurrences. Also, Figure 2 (a) and (c) let the user hypothesize about the location of the hospitals in Cornertown, Suburbia, Southville and Lakeside.

4.2 System Performance

Scalability with respect to the size of the analyzed data sets is an important aspect of a system's performance. In practice, big data files are often held on sophisticated database storage systems, which themselves can manage operations such as filtering and grouping. Many VA systems can work with DBMSs and it was not our goal to test the capacity and connection speed for any particular DBMS. Instead we experimentally tested the upper boundary of data load that a VA system can handle on its own.

We generated a series of test data sets of increasing size. Our test data are uniformly generated records of 50 dimensions, containing 3 categorical and 47 numerical values. We provided our test data as CSV files of 100 MB (204,683 records), 200 MB (409,358 records), 500 MB (1,023,348 records), 1 GB (2,095,847 records), 10 GB (20,957,918 records), 20 GB (41,915,609 records),



(a) Tableau

(b) QlikView

(c) Spotfire

(d) JMP

Figure 2: Visualization of Spatial-temporal Data in *Tableau, QlikView, Spotfire* and *JMP*

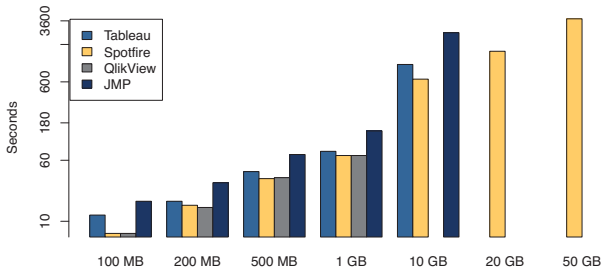


Figure 3: Loading Stress Test

and 50 GB (104,789,361). The evaluation was conducted on a workstation with an Intel Core i7-2600 CPU, and 16 GB of main memory. The operating system and the tools are installed on a 128 GB SSD drive. In addition, the workstation has a 1 TB HDD storage for user data, which we used for storing the workbooks created with the tools.

For each system we measured the time required for loading the data set into a project and displaying the data table. Figure 3 shows for each VA system the time to load the data. Only *Spotfire* was able to handle a data size of 50 GB. *QlikView* failed to load the 10 GB file on our test system. *Tableau* and *JMP* reached their limits at 20 GB. At 10 GB Tableau was not able to display the data table anymore. In all other cases the times taken for displaying the data table was negligible. *Spotfire* was even able to show the data table instantly for the 50 GB test after the data was loaded.

5 Summary of Key Findings

Generally speaking, the tasks supported by all investigated VA systems fall into four categories: exploration, dashboards, reporting, and alerting. Exploration allows users to generate and verify hypotheses. The advantage is the ability to easily create and modify visualizations and statistical models. The result of the exploration is usually additional knowledge or statistical models. In contrast, dashboards are either used to communicate findings or to provide standardized interfaces for regularly occurring analysis problems. Usually a dashboard consists of a fixed set of visualizations and controls, allowing interactions such as selection, filtering, and drilling down. The reporting task generates a static summary of information from the data sources. Reports are either generated on demand or on a regular basis. The representation of the information in the reports is standardized, allowing easy comparison of different reports. The alerting task provides automatic notification when the data sources reach predefined states. These states are typically thresholds or indicators, but more complex ones may incorporate evaluations of statistical models. Alerts are used to inform users about unusual events that need attention.

Among all the systems we surveyed, a number have roots back in academic research, for example *Tableau* from Stanford University, *Spotfire* from University of Maryland, and *ADVIZOR* from Bell Labs. These vendors appear to be leaders in interactive visualization and automatic analysis, and put effort in integrating innovative visualization techniques. For example, *Tableau* benefits from its unique visual query language, VizQL, that translates user actions into a database query and then expresses the response graphically. *Spotfire* provides powerful automatic analysis functionality and is regarded as a pioneer in predictive analysis. *ADVIZOR* implements different types of interactive charts, some of which are not included in many other VA systems.

Tableau is still expanding its statistics and automatic analysis functionality over the latest releases. *Spotfire* already has advanced its functionality in all aspects we investigated - from automatic analytics, to interactive visualization, from system architecture to data management. However, some advanced data analysis components are only available with additional upgrades and cost. (see Table 2).

QlikView appears advanced regarding data compression and memory optimization. It has strong interactive drill-down capabil-

ities and fast response time because of its in-memory architecture. The system accesses information from standard database applications and displays data associatively using highlighting colors. But not many statistics and automatic analyses are included in the system.

Several other systems, such as *JMP* and *Cognos* (which is not included in our study) also provide strong analytical capabilities by integrating their own VA components. For example, *JMP* integrates SAS, and *Cognos* integrates SPSS. In particular, the integration of interactive visualization with automatic analysis functionalities makes *JMP* an advanced data discovery system for data modeling and predictive analysis.

Systems more oriented towards BI, such as *Centrifuge*, *Board*, *Visual Mining* and *Jaspersoft* put much focus on presentation-oriented features (e.g. dashboards, reports), which allow the user to generate in a straightforward way graphical representation of standard data. Among those, *Jaspersoft* is one of the least costly BI products on the market, although it appears to be a little behind other BI systems in terms of functionality and infrastructure. *BOARD* earns the name of an innovative product by integrating BI and Corporate Performance Management (called Management Intelligence by the tool's advocates). One issue we noticed is that the interactivity of most of the dashboard facilities is rather limited.

While network analysis is still not a fully developed functionality in many VA systems, *Centrifuge* and *Visual Analytics* put much focus on applying interactive network visualizations and automatic analysis methods to help understanding hidden relations in data. Visual Analytics is widely used in financial transaction data analysis and fraud detection. A range of reactive and proactive analyses is supported, including entity extraction, social network analysis, geo-spatial analysis, etc.

Linguistic analysis on text documents is not supported by many VA systems, despite the increasing amount of text documents generated on- and off-line and need to analyze them. To our knowledge, only three systems in our initial list (*Business Objects*, *Cognos* and *Teradata*) have text mining functionality. However, for more specific text mining tasks, *Oculus* provides a nice open source toolkit *nSpace* [12] which includes a number of useful functions including faceted search, faceted trends, and evidence marshalling. Besides, *Palentir* [2], and *In-Spire* [1] are also known for their text analysis capabilities.

6 Concluding Remarks

VA system development is a fast moving field with effort been made by multiple disciplines including statistics, machine learning, information visualization, human computer interaction, data management, and memory optimization. Besides open source toolkits, a large number of commercial products were developed, marketed, and employed, relying in practice on corporate IT as well as IT consulting services. In the past ten years, on the one hand some existing VA software companies expanded rapidly (e.g. *Tableau Software*, *QlikTech (QlikView)*) due to the growing market. On the other hand, big software vendors such as *IBM*, *Oracle* and *Microsoft* started to either acquire successful VA software companies and integrate acquired VA components into their own framework (e.g. *IBM* bought *Cognos*, *Oracle* acquired *Siebel* and *Hyperion*, *SAP* purchased *Business Objects*, and *TIBCO* acquired *Spotfire*) or to develop their own VA components (e.g. *SAS* developed *JMP*, *Microsoft* developed *Sharepoint* and *PowerPivot*). Such phenomena are not surprising in a dynamic market where the trend is led by the practical need in application domains. The trend is most likely going to continue if we look at the increasing volume, velocity and variety of data that are generated in different application domains nowadays.

In this paper, we report our survey on a selection of state-of-the-art VA systems as a basis for analyzing current market and trend, discussing space for improvement and identifying future research directions. We evaluate the functionality and performance of each system by surveying the vendor with a structured questionnaire as well as testing with real world data. We detail our findings and outline the main characteristic of each system. Our survey provides a comparative review of ten products on the market. We also investigate a larger number of systems, including *Cognos*, *SQL Server BI*, *Business Objects*, *Teradata*, *PowerPivot*, *Panopticon*, *KNIME*, *Oculus*, *Palentir* and *in-Spire* to gain a better overview of the VA software market. Future work will include harmonizing findings of the latter tools, which are still being collected, with the presented systems.

Through our study, we identify a number of challenges which may lead to possible future directions:

Semi- and Unstructured Data. The increasing speed of data generation brings both opportunity and challenge. In particular, more and more semi- or unstructured data are generated on- or off-line. A large number of data analysis and visualization techniques are available for analyzing structured data, but methods for modeling and visualizing semi- or unstructured data are still underrepresented. An effective VA system often needs to be able to handle both, and ideally integrate the analysis of both types of data for supporting decision making.

Advanced Visualization. Compared to open source VA systems, it seems that commercial products take longer time to integrate innovative visualization techniques. In particular, some big software vendors tend to focus on only a small number of "standard" visualization techniques such as line charts, bar charts and scatter plots, which have limited capability in handling large complex data. The success of *Tableau*, *Spotfire* and *ADVIZOR* demonstrate the possibility and benefit of transferring technical advances developed by academic research into industrial products.

Customizable Visualization. One useful feature which is often ignored in visualization function design is customizable visualization. Given the same data and visualization technique, different parameter settings may lead to totally different visual representations and give people different visual impressions. Designing customizable visualization functions leaves the user the freedom of changing visual parameter setting and more opportunity to gain insight from the visualization.

Real Time Analysis. More and more data are generated in real-time on the Internet (e.g. online news streams, twitter streams, weblogs) or by modern equipment or devices (e.g. sensors, GPS, satellite cameras). If analysis is applied appropriately, these data provide rich information resources to many tasks. Therefore, improving analytical capability to handle such data is a development opportunity in current commercial products. We expect to see more functionality in this respect in the future.

Predictive Analysis. The demand of predictive modeling is increasing, especially in the business domain, but only very few systems support predictive analysis. Even with those systems that support predictive analysis, not many predictive modeling methods are implemented.

Acknowledgements

This work was partially funded by the German Research Foundation (DFG) under grant GK-1042 "Explorative Analysis and Visualization of Large Information Spaces" and by the European Commission (FP7) under the grant "Modeling and Simulation of the Impact of Public Policies on SMEs (MOSIPS)". The authors wish to thank Christine Jacob for her work on testing the different applications.

REFERENCES

[1] http://in-spire.pnnl.gov/.

[2] http://palantir.com/.

[3] http://radar.oreilly.com/2012/01/what-is-big-data.html.

[4] http://spotfire.tibco.com/.

[5] http://vismaster.eu/.

[6] http://www.advizorsolutions.com/.

[7] http://www.board.com/.

[8] http://www.centrifugesystems.com/.

[9] http://www.ibm.com/software/data/bigdata/.

[10] http://www.jaspersoft.com/.

[11] http://www.jmp.com/.

[12] http://www.oculusinfo.com/nspace/.

[13] http://www.qlikview.com/.

[14] http://www.tableausoftware.com/.

[15] http://www.visualanalytics.com/.

[16] http://www.visualmining.com/.

[17] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery: an overview. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in knowledge discovery and data mining*, chapter From data mining to knowledge discovery: an overview, pages 1–34. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996.

[18] J.-D. Fekete. The infovis toolkit. In *INFOVIS*, pages 167–174, 2004.

[19] J. Hagerty, R. Sallam, and J. Richardson. Magic quadrant for business intelligence platforms. Technical report, Gartner Technology Research, 2012.

[20] P. Hanrahan. Vizql: a language for query, analysis and visualization. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, SIGMOD '06, pages 721–721, New York, NY, USA, 2006. ACM.

[21] J. R. Harger and P. J. Crossno. Comparison of open-source visual analytics toolkits. In *Proceedings of the SPIE Conference on Visualization and Data Analysis*, 2012.

[22] S. Havre, B. Hetzler, and L. Nowell. Themeriver: Visualizing theme changes over time. In *Proc. IEEE Symposium on Information Visualization*, pages 115–123, 2000.

[23] J. Heer, S. K. Card, and J. A. Landay. prefuse: a toolkit for interactive information visualization. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '05, pages 421–430, New York, NY, USA, 2005. ACM.

[24] Java Universal Network/Graph Framework. http://jung.sourceforge.net/, 2012.

[25] O. Kaser and D. Lemire. Tag-cloud drawing: Algorithms for cloud visualization. *CoRR*, abs/cs/0703109, 2007.

[26] D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, editors. *Mastering The Information Age - Solving Problems with Visual Analytics*. Eurographics, 2010.

[27] J. Thomas and K. Cook. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE Computer Society, 2005.

[28] D. van Beek and N. Manley. The business intelligence product survey. Technical report, Passionned Group, 2012.

[29] C. Weaver. Building highly-coordinated visualizations in improvise. In *INFOVIS*, pages 159–166, 2004.