Commercial Visual Analytics Systems – Advances in the Big Data Analytics Field

Michael Behrisch, Dirk Streeb, Florian Stoffel, Daniel Seebacher, Brian Matejek
Stefan Hagen Weber, Sebastian Mittelstädt, Hanspeter Pfister, Daniel Keim

Abstract—Five years after the first state-of-the-art report on Commercial Visual Analytics Systems we present a reevaluation of the Big Data Analytics field. We build on the success of the 2012 survey, which was influential even beyond the boundaries of the InfoVis and Visual Analytics (VA) community. While the field has matured significantly since the original survey, we find that innovation and research-driven development are increasingly sacrificed to satisfy a wide range of user groups. We evaluate new product versions on established evaluation criteria, such as available features, performance, and usability, to extend on and assure comparability with the previous survey. We also investigate previously unavailable products to paint a more complete picture of the commercial VA landscape. Furthermore, we introduce novel measures, like suitability for specific user groups and the ability to handle complex data types, and undertake a new case study to highlight innovative features. We explore the achievements in the commercial sector in addressing VA challenges and propose novel developments that should be on systems' roadmaps in the coming years.

Index Terms—System Comparison, Commercial Landscape, Visual Analytics Research, Advances, Development Roadmap.

1 Introduction

In 1890, Herman Hollerith revolutionized the world of data analysis with a creative and innovative idea: he used punch cards to collect and analyze the US census data. Using punch cards saved two years and five million dollars over the manual tabulation techniques used in the previous census while enabling more thorough analysis of the data [1]. We currently face an analogous development in the Big Data Analysis field, where commercial Visual Analytics (VA) systems allow a faceted confirmatory or a data-driven exploratory analysis of large amounts of data in significantly less time than years ago. Today, the success of many businesses relies on efficient and effective analysis of massive quantities of data.

Bertin [2] and Tukey [3] consider the possible levels of data, information, and analysis. They summarize data analysis into three levels: presentation, confirmatory, and exploratory analysis. Over the last decade, a significant amount of research explores presentation and confirmatory analysis in the commercial VA field. Specifically, dashboarding systems enable users to gain quick insights with faceted filtering functionality. Confirmatory analysis scenarios are supported either by focusing on simple visual interactive Overview + Detail displays or by incorporating increasingly automatic analysis techniques into coordinated view systems. We claim that although current commercial VA systems have been developed with the aim to support the exploration of large quantities of data, they currently do not sufficiently support exploratory analysis scenarios. In particular, we see a scarcity of supportive environments where the domain expert and the machine work in an interplay towards formulating and validating hypotheses. This dearth is due to several reasons: (1) often users are left alone in finding a starting point in their analysis; (2) the communication of non-trivial hypotheses is challenging; (3) automatic algorithms for validating interesting findings are not scalable or even implemented in the systems. This survey counterbalances the efforts of the community against the needs and requirements imposed by the Big Data Era. Further, we ask which steps should be taken in the future by examining past directions to allow for exploratory data analysis in Big Data scenarios.

We revise and update the 2012 state-of-the-art report on commercial VA systems following the original methodology and rationale of Zhang et al. [4]. We build our comparative market overview on an encompassing list of 46 relevant commercial VA systems.[1] These chosen systems reflect current market shares [5]–[7] and encompass the broad product categories within the field: e.g., data discovery, visual software, Business Intelligence (BI), innovative, and niche products.

Our survey is structured along two primary dimensions. In a *user/task* oriented view we claim that three user groups with potentially overlapping skill sets are interested in commercial VA systems: (1) Upper management, e.g., CEOs, who make critical business decisions based on prepared presentations; (2) Domain experts who have extensive domain knowledge and can formulate hypotheses; (3) Data analysts and engineers who do not necessarily know the data in advance but have the challenge of finding a needle in a large amount of complex data, potentially at high velocity.

The second dimension structuring this survey relates to the *functional capabilities* of commercial VA products. Therefore, we approached all 41 vendors (five offered two candidate products)

- *M. Behrisch, B. Matejek, H. Pfister are with Harvard University, Cambridge, USA*
  E-mail: {behrisch, bmatejek, pfister}@g.harvard.edu
- *D. Streeb, F. Stoffel, D. Seebacher, D. Keim are with University of Konstanz, Germany.*
  E-mail: {dirk.streeb, florian.stoffel, daniel.seebacher, daniel.keim}@uni-konstanz.de.
- *S. H. Weber, S. Mittelstädt are with Siemens AG, Corporate Research Germany.*
  E-mail: {stefan_hagen.weber, sebastian.mittelstaedt}@siemens.com.

Manuscript received August 01, 2017; revised August 01, 2017.

1. The complete list is on our website http://commercialtools.dbvis.de/

to get responses to a structured questionnaire[2] targeting the following feature sets: Data Handling and Management; Automatic Analysis; Complex Data Types, Visualization; and User-Guidance, Perception, Cognition, and Infrastructure. This thorough insight and overview builds the basis for our feature richness and degree of innovation comparison scheme. We intentionally complement the objective feature assessment with our expertise derived from applying the systems. Lastly, we venture a glimpse into the future by contrasting recent advances in the sector with interesting developments from the VA research community.

We claim the following contributions for this paper: (1) we update and increment the 2012 survey of commercial VA systems from Zhang et al. [4] and complement current user surveys of BI tools [5]–[7] by conducting a faceted evaluation of commercial VA systems; (2) we introduce an elaborate evaluation scheme for judging the feature richness and degree of innovation for the diverse commercial VA systems; (3) we present a detailed quantitative performance evaluation that incorporates measures for each of the main steps in a generic analysis-workflow: (4) we contrast the current developments in the field with a selection of trending topics from the VA research community and identify future directions for developing VA systems; (5) we give practical guidelines and recommendations to potential users on which systems are applicable to what types of applications. Our paper is structured as follows: At first, we revisit previous work in Section 2. Then, Section 3 describes the selection and evaluation criteria for surveying the commercial VA field. Section 4 gives an overview of the commercial VA landscape and introduces a user-dependent requirement analysis. In Section 5, we present the results based on our online questionnaire and our own experiences with the systems, followed by a performance and case study evaluation in Section 6, respectively Section 7. Subsequently, Section 8 combines our key findings and discusses the major insights, trends, and open challenges. We conclude in Section 10 with a summary and an outlook on the next years.

2 Reflections and Related Work

In 2012, Zhang et al. [4] published a survey on the state of commercially available Visual Analytics systems. They compare available features (e.g., visualization, automatic analysis techniques, and usability), perform a case study based on the 2011 VAST Challenge data set, and evaluate performance with a loading stress test on a selection of ten systems. One of their main findings is that vendors with an academic background are market leaders and have gradually increased the number of visualization and automatic analysis techniques in their products [4, p. 180]. However, they note that commercial products lag behind open-source VA systems when it comes to inclusion of novel visualizations [4, p. 181]. They conclude that the open challenges for most systems are semistructured and unstructured data, advanced and customizable visualization, and real-time and predictive analysis. Overall, their survey has had an impact within and, more interestingly, outside the Visual Analytics community.

With this survey we not only want to build directly upon the work of Zhang et al. from 2012 [4] but also reflect on how their survey was received across the research community boundaries: we reviewed in total 66 publications ranging from conference papers to Master's and Ph.D. theses to even books. All these references build on Zhang et al.'s work [4]. Although the VA community (19 references) and the visualization community (12) can be seen as the target audience for this survey, around half of the 66 publications are not focused on VA or visualization. Application-driven publications (25) contribute the most to this (external) group.

These publications cover a diverse range of topics. Applications frequently focus on Big Data (8), e.g., most recently Akoka et al. [8], Business Intelligence/Processes (6), e.g., Aggarwal et al. [9] give an introduction to the analytics process, and geo-related use cases (5), e.g., Fernandez et al. [10] show an application to seaport monitoring. Most interestingly, Zhang et al.'s survey not only reached researchers in other fields but also industry, e.g., Zillner et al.'s work [11] resonates on industrial Big Data efforts. Numerous authors make use of the summary of the state-of-the-art and the review of systems. Others use the reference to underline the importance of their research gap or question.

In recent years various surveys on VA systems have been published and are continuously updated. On the commercial side, Forrester Research [6], Gartner [5], and Business Application Research Center (BARC) [7] offer surveys on commercial BI systems with data visualization. Their evaluation criteria reflect the business perspective on the market and they primarily focus on cost, rollout, and market presence, but also consider content creation and user satisfaction. Their findings are mostly based on customer and vendor surveys. Gartner [5], for example, ranks vendors in quadrants based on subjective ratings along "Ability to Execute" and "Completeness of Vision" dimensions. BARC [7] presents a survey of more than 2,000 users, an approach we do not replicate.

Openly available comparisons from blogs are more informal and feature open-source software more prominently. To give one example, Rost [12] recreates a bubble chart in 24 ways using available systems as well as charting libraries. Lastly, academic surveys tend to build on smaller samples of systems and include more formal evaluations. Umaquinga et al. [13] compare both commercial and open-source systems from the visualization and data analysis domains. Nair et al. [14] compare visualization capabilities of D3.js [15] to *Tableau* [16].

Researchers have numerous diverse expectations towards future developments in the field. Lemieux et al. [17] expect closer collaboration with record managers. Li et al. [18] mention data streams, scalability, and uncertainty as major challenges in geo data analysis and call for more sophisticated analysis methods [18, p. 128]. Piovano et al. [19] mention that both interfaces for non-technical users and mobile analysis scenarios should be improved. Nocke et al. [20] see an increasing demand for the analysis of network data. Interestingly, the commercial sector resonates some mentioned expectations. *Tableau* [16] expects the top developmental trends for the future to include more non-technical users, mobile analysis, and cloud integration.

We design our survey to facilitate a scientific perspective on available commercial systems with a strong focus on Visual Analytics. We deliberately do not evaluate pricing and license models since realistic models are only available after lengthy negotiation rounds and the gained information is mostly confidential. Overall, we follow a comparative evaluation scheme that combines several facets into an all-embracing picture: (1) an online questionnaire for the product's feature base, (2) a performance evaluation, which

2. The complete question catalog can be found online: http://commercialtools.dbvis.de/questions

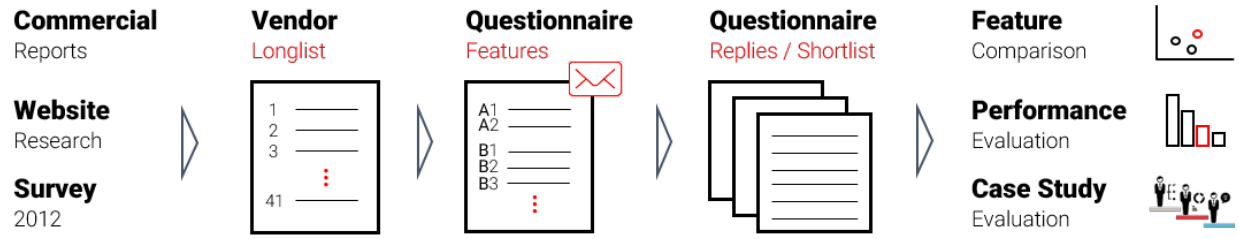


Fig. 1. We compare a representative set of ten commercial VA systems to gain an overview of the current state-of-the-art in the field and to derive practical guidelines for which user groups can specifically benefit from which systems. Therefore, we extract a longlist of 41 vendors—offering in total 46 commercial VA products—and approach each vendor with a structured questionnaire of 50 questions. We derive a shortlist of ten products and thoroughly test their feature richness and degree of innovation (Section 5), their performance (Section 6), and their usability (Section 7).

encounters all necessary steps of a VA-inspired workflow, and (3) a use case evaluation on an established and challenging data set (VAST challenge 2015 [21]) to prove each system's usefulness for a specific target group.

3 Methodology

In the following section, we summarize our selection procedure for deriving a representative set of systems for a commercial VA landscape overview. Generally, our approach is comparable to the evaluation strategy presented by Zhang et al. in 2012 [4] and is based on the questionnaire responses from 41 VA system vendors.

3.1 Longlist Selection Criteria

The commercial VA landscape provides an abundance of products targeted at partially overlapping usage scenarios—such as exploratory analysis, confirmatory analysis, and result presentation—or niche products for specific data types. Our goal is to make an informed selection of candidates for a detailed investigation. To that end, we started with an initial set of 46 relevant commercial VA systems on our *product longlist*, which we collected from the main commercial business reports from BARC [7], Forrester Research [6], and Gartner [5]. We consolidate our product longlist with Zhang et. al's 2012 survey to ensure validity and comparability, and expand/validate our vendor longlist with queries using prevalent keywords. Our full longlist of commercial VA systems can be accessed online at http://commercialtools.dbvis.de/systems. At this point, we explicitly exclude the even more dynamic open-source developments and decide to contrast commercial and research developments in the VA community. One of the motivations for this survey is to examine the knowledge transfer from academia to the market, which updates and extends the work of Zhang et al. [4]. Moreover, we focus on integrated VA systems and have to exclude platform solutions. These solutions typically comprise a set of stand-alone though integrated tools for business intelligence functionality. The obvious example of this category is the *Oracle BI platform*, which consists of an entire set of distinct products for data collection and storage, data and access management, reporting, and analysis. These products are targeted at medium-to-large scale companies with a respective network setup and cannot be simulated and compared in our survey setup.

3.2 Questionnaire Design

In early 2017, we approached each vendor on our longlist with a feature-assessment questionnaire consisting of 50 multiple-choice and free-text questions split into twelve question groups. The questions cover system architecture, data import and preparation, automatic and visual analysis, presentation of results, working with sensitive data, and collaboration, and follow the analytics process sketched in [22] and [23]. Most questions (17) target the central visual and automatic analysis process. However, we also value import and preparation of data (8) as they are crucial to practical success and attract increasing interest in the research community. Further questions target, amongst others, presentation of results (3), extensibility (2), user support (4), working with sensitive data (3), and collaborative features (2). We also ask the vendors to position their product in the landscape and predict future challenges.

3.3 Shortlist Selection Criteria

During 2017, we received eight responses to our online questionnaire for Advizor Solutions *Advizor*, IBM *Cognos*, SAS *JMP*, SAS *Visual Analytics*, Tableau Software *Tableau*, Tibco *Jaspersoft*, Tibco *Spotfire*, and Microsoft *PowerBI* (Desktop). The average time to fill out our survey was 4.3 hours (263 min. 28 sec.) and the median almost 6 hours (357 min. 41 sec.). The survey sessions could be paused, and the vendors were explicitly allowed to distribute the survey within the company, such that the department with the most expertise could fill in the respective answers.

We include all systems for which we received a complete answer to our *product shortlist*. Additionally, we add Qliktech *QlikView* and SAP *Lumira*, respectively a major competitor and a prominent newcomer in the field. For these systems, we managed to find many answers to our questionnaire by ourselves. This results in a shortlist of ten systems, as shown in Figure 2. In contrast to Zhang et al.'s 2012 survey, we do not include *Board*, *Palantir*, *Centrifuge*, and *Visual Mining* as they did not answer our questionnaire. Further, they cannot be regarded as key players in the field anymore since they represent rather niche products tailored to the analysis of specific data types.

3.4 Evaluation Facets

Taking the product shortlist as a representative selection of systems from the commercial VA landscape, we thoroughly assess the products on their feature richness and degree of innovation (Section 5), their performance (Section 6), and their usability (Section 7). We outline the evaluation criteria for each separate evaluation facet in the corresponding sections.

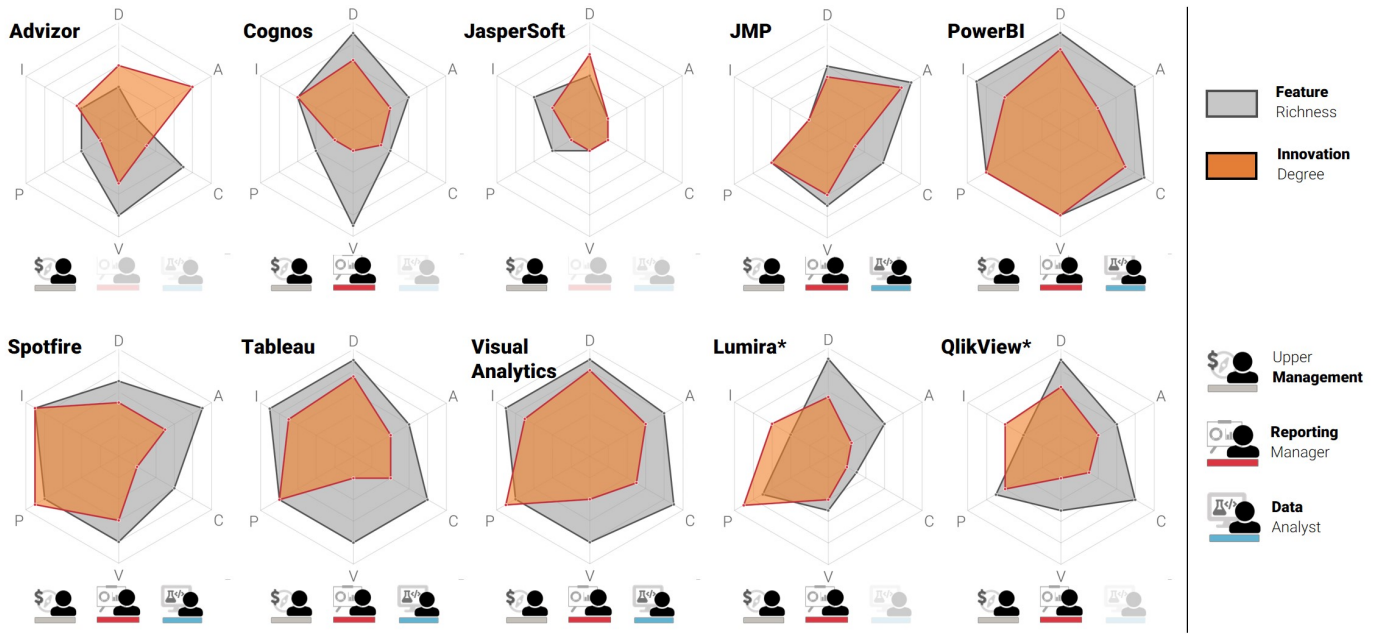


Fig. 2. Overview of the ten evaluated systems in this survey. Dimensions in the radar chart are: **D**=Data Handling and Management, **A**=Automatic Analysis, **C**=Complex Data Types, **V**=Visualization, **P**=User-Guidance, Perception, Cognition and **I**=Infrastructure. Commercial VA systems are designed for specific user groups with varying and overlapping skill sets and requirements w.r.t. data handling, analysis, and reporting.

4 Commercial VA System Landscape

Historically, the commercial VA systems landscape emerged from a partial intersection of distinct fields due to new demands posed by more complex analysis tasks and larger data sets. The early static visualization systems were designed as a supplemental facet to data warehousing approaches, such as *Extract, Transform, and Load (ETL)* and *Online Analytical Processing (OLAP)* software. Over the course of the last 20 years, the landscape matured to enterprise solutions helping even small businesses process large volumes of data and visualize insights.

A number of commercial VA systems trace their roots back to academic research. For example, *Spotfire* was founded in 1997 as a spin-off from the University of Maryland's Human-Computer Interaction Lab. In 2003, *Tableau* was started as a Stanford University spin-off capitalizing on the Polaris research system [24]. In the same year, *Advizor* spun-off successfully from Bell Labs.

All research spin-offs have approached the emerging market with distinct skill-sets manifesting the diverse approaches in the commercial VA sector today. For example, *Tableau*'s core architecture resides on *VizQL* [25], a declarative query definition language that translates user actions into database queries and the respective data responses back into graphical representations. Similarly, *Spotfire*'s architecture builds on top of *IVEE: An information visualization & exploration environment* [24], a research prototype for the dynamic queries idea in which the database query process is translated into visual metaphors. The unifying concept that made these products successful is that they tightly integrate the user into the analysis workflow by allowing an incremental and, most importantly, interactive exploration of the data set. On the other hand, *Spotfire* recognized early on that automatic (statistical) analysis functionality would play a key role in the next round of commercial VA systems. Orthogonally, *Advizor* followed the credo that different types of interactive and coordinated visualization displays allow for a multi-faceted data analysis—a concept known as coordinated views. We can still retrieve the main distinct tendencies of today's commercial VA system landscape from the historical examples above:

Data Representation: Over the course of the last five to ten years, a significant amount of work explores the interactive presentation of data with coordinated views. The main established category description is "dashboarding" tools. In these nowadays mostly web-based systems, users can gain quick insights into their data with faceted filtering functionality.

Confirmatory Analysis: Two diverging trends became apparent five years ago for confirmatory analysis scenarios [4]. First, most visual interactive systems are based on simple Overview+Detail visualizations with rudimentary Detail-on-Demand functionality. In contrast, data, statistics, and algorithm-driven approaches enable users to communicate and validate their hypotheses.

Exploratory Analysis: While today's commercial VA landscape is dominated by confirmatory analysis systems and data presentation tools, the exploration of large quantities of (high-dimensional) data imposes new challenges on the analysts. Consequently, the third group of systems in the commercial VA landscape comprises fully-fledged exploratory analysis systems, which tightly integrate the domain expert with the analysis techniques. These systems make the user aware of potential data uncertainties, handle missing data aspects, or suggest the next step along the exploration path.

Developments

In comparison to our report from 2012, we see several trends in the commercial VA landscape: first, traditional statistics systems, like *JMP*, focus on adding more interactive visualization capabilities; second, systems formerly focused purely on data visualization, like *Tableau*, integrate more and more automatic analysis features; third, some BI/dashboarding systems, including *Cognos*, adopt more interactive workflows.

Needless to say the commercial VA market itself is evolving. New products designed from scratch, such as SAP *Lumira* or SAS *Visual Analytics*, have entered the market. Other companies

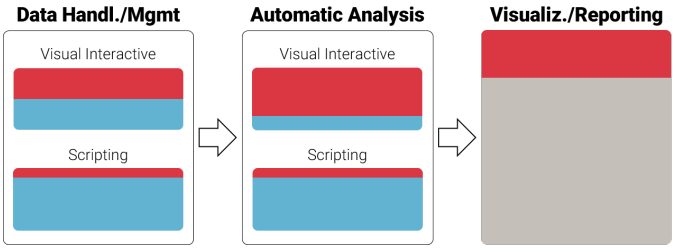


Fig. 3. Basic analysis workflow: Upper management (gray) is primarily interested in presenting prepared results. Reporting managers (red) rely on interactive interfaces all along the workflow. Data analysts (blue) need flexible systems to adapt workflows to deal with their challenging and ill-defined problems.

have been acquired, e.g., TIBCO acquired Jaspersoft in 2014. In this survey we do not outline all market developments, but defer the interested readers to the business reports of Gartner [5], Forrester [6], and BARC [7] for more detailed information on these topics. Overall, we can state that while the market has been diverse and dynamic in recent years, *QlikView*, *Spotfire*, *PowerBI*, and *Tableau* are the established key players in the field.

On the other hand, many specialized systems occupying their economic niche have also evolved. The most prominent examples are Geographic Information Systems, such as *Esri ArcGIS* or *GeoTime*. Other examples, such as *Palantir* or *Centrifuge*, provide specialized solutions for the security, financial, or health sectors and present network analysis and visualization systems. However, we also see innovative VA products for the general public, like *Visual Analytics*, that are challenging the proven systems.

4.1 Requirement and User-dependent Task Analysis

As a result of this diverse field, it is challenging to compare all systems without considering the requirements imposed by different use cases, investigated by the various user roles within a company.

Several scientific works focus on the question of which user roles and tasks exist in enterprise data analysis environments. Kandel et al. [26] describe, similarly to our user categorization, three analyst archetypes (Application User, Scripters, and Hackers) that differ in terms of skill set and typical workflows. Kandogan et al. [27] claim that the data analytics user landscape is even more nuanced and should rather be categorized into the types of analytical work. In their framework, they distinguish three work types with respect to its scope, duration, nature, and target users: (1) mid-level business people tend to carry out *tactical work* that requires quantitative analysis and numerical data processing, dealing with short analysis durations (i.e., weeks); (2) high-level executives focus on *strategic work* that is longer-term and forward-looking with a strong focus on predictive analysis; (3) line-of-business managers consider *operational work* that involves timely analysis of transactional data [27].

While Kandogan et al.'s and Kandel et al.'s work outline a nuanced user landscape, our user-dependent task categorization has a similar but far broader scope and aims to reflect the central visualization goals (presentation, confirmatory, and exploratory analysis) already outlined in Section 1 and discussed in Section 4. As a direct mapping of the visualization goals into the commercial VA landscape we can safely distinguish three broad main user categories for the commercial VA sector, which we briefly describe in this section and for which we highlight their primary needs.

Upper Management: For users with this role, it is most important to present information convincingly and consistently. Typical application scenarios are, for instance, board or shareholder meetings, where interactive presentations are usually avoided. Therefore, a clear presentation of facts, as well as storytelling capabilities, are of utmost importance.

Reporting Manager: Users of this group are often tasked with confirmatory or hypothesis-driven analysis, such as finding the best selling items or checking perceived trends in sales records. For them, it is important that systems offer a broad range of interactive analysis and visualization techniques.

Data Analyst: Users of this group are mainly interested in exploratory analysis to find new and valuable knowledge in given data. For users with this role, extensibility, interactivity, and data handling are highly important. Although they prefer interactive workflows, they are capable of extending systems, e.g., with R scripts [28], to enable analyses not feasible using out-of-the-box systems.

Figure 3 contrasts the described user groups regarding their modes of use along the analysis workflow. While the upper management (gray) focuses on the *reporting/presentation* of prepared results and thus relies on the other user groups, reporting managers (red) need interactive interfaces all along the workflow to formulate and verify their hypothesis (*confirmatory analysis*). Data analysts (blue), on the other hand, require adaptive systems that can reflect their non-standard data science workflows.

5 Feature Comparison

Fig. 4. **Feature Evaluation:** We evaluate the qualitative questionnaire responses based on the structural content analysis of Mayring [29] in which they derive a structured overview of the systems' capability per feature group/criterion. Subsequently, we rank the systems relative to each other on a Likert-scale between 0 (not supported) to 5 (fully supported) for feature richness and 0 (no focus) to 5 (incorporates the latest research) for the degree of innovation.

For our functional comparison, we investigate state-of-the-art commercial VA systems with respect to the following feature groups: Data Handling and Management; Automatic Analysis; Complex Data Types, Visualization; User-Guidance, Perception, Cognition; and Infrastructure. While a myriad of interesting subcategorizations might lead to meaningful feature comparisons, we stay close to the standard data analysis and visualization pipelines, such as the *Information Visualization Reference Model* of Card et al. [22] or the *Knowledge Discovery Pipeline* of Fayyad [23]. Accordingly, we incorporate must-have categories, such as Data Handling and Management (see: Section 5.1), Automatic Analysis (see: Section 5.2), and Visualization (see: Section 5.4) to reflect the primary analysis step. However, we also enlarge the scope with a more fine-granular investigation of meta-categories, such as support for Complex Data Types (see: Section 5.3), User-Guidance, Perception, Cognition (see: Section 5.5), and Infrastructure considerations (see: Section 5.6).

Feature Comparison Methodology: With our feature comparison scheme, we hope to derive a comparative overview of each system's state-of-the-art feature and topic groups. As Figure 4 depicts, the main data set for deriving a comparative rating of a system's (a) feature richness and (b) degree of innovation is the qualitative questionnaire feedback. Following the structured content analysis theory of Mayring [29], we consolidated, summarized, and aggregated the questionnaire responses into findings and rankings.

First, in an initial data preparation phase, we selected all answers that contribute to each of our topic groups. To ensure the response validity, we did a manual sanity check if answers appear to be unclear.

Second, three of our authors went collaboratively through the responses to describe feature sets and map the systems supporting these features (*coding phase*). A fourth author played the devil's advocate role at this stage by enforcing the consistency, understandability, and validity of the decisions. Whenever decisions were unclear, the coders either had to produce a good argument or suggest a recoding. The process of (re-)coding and allocating vendor responses was repeated several times until a consensus was reached among the authors. A coding/allocation round took between 0.5 and 3 hours depending on the feature group, the number of questions/responses, and the number of multiple-choice questions in the respective topic group.

Third, after getting a structured overview about the systems' distinguishing feature sets in the coding phase, all involved authors discussed the findings and ranked the systems relative to each other on a Likert-scale between 0 (not supported) to 5 (fully supported) scale for feature richness and 0 (no focus) to 5 (incorporates the latest research) for the degree of innovation. Our argumentation for evaluating the feature richness is the following: A well-designed system with many analytical features can and should satisfy diverse task needs. In contrast, pure feature bloat makes the user experience negatively more complex.This rationale leads to the inclusion of the User-Guidance, Perception, Cognition section (see: Section 5.5), which emphasizes this balance.

Five years of new application tasks since the original 2012 survey [4] have led to improved VA systems and new competitors. At the same time research has progressed tremendously in many varied directions. One of the motivations of this survey is to evaluate knowledge transfer from academia to the market, which updates and extends on the work of Zhang et al. [4]. With this focus in mind, we discuss at the end of each section the gap between commercial and research fields dealing with data handling, automatic analysis, complex data types, infrastructure, and usability. Therefore, we use a representative selection of current research topics and visions that could become of interest for the commercial VA sector in the next years. Our selection of trending topics is not exhaustive and is the result of a subjective selection process. Nevertheless, while other topics could be discussed in their respective sections, we claim that an investigation of the missing pieces will help to develop a roadmap for the development of better visual interactive exploratory analysis interfaces.

5.1 Data Handling and Management

All primary data analysis and visualization pipelines, such as the *Information Visualization Reference Model* of Card et al. [22] or the *Knowledge Discovery Pipeline* of Fayyad [23] begin with a form of raw data that is loaded, integrated, and (pre-)processed so that it can be analyzed visually or automatically in subsequent steps.

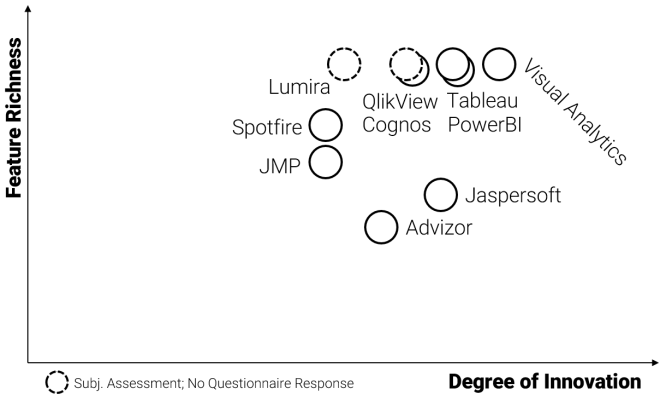


Fig. 5. **Data handling and -management:** All key players in the commercial VA field deliver a similar set of data handling and management features. *Visual Analytics* shows an innovative handling of noisy and raw data (esp., imputation of missing values). Detailed feature overviews can be found in the appendix tables http://commercialtools.dbvis.de/appendix.

Various descriptive names have been coined for this highly time-consuming step, such as *ETL (Extract, Transform, and Load)* in the BI domain or *Data Wrangling* in the data analysis domain [30].

Our primary focus in this feature comparison category is to assess the *data handling and management* functionality based on the following feature sets: (1) import/export, (2) integration of heterogeneous sources, (3) *ETL* features and (4) (semi-)automatic data preprocessing support.

As the overview in Figure 5 depicts, most systems offer a similar feature richness in terms of data handling. Almost all systems support standard import/export formats, such as CSV, Excel/Google Spreadsheets, JSON, XML and relational databases. Also, the support for NoSQL databases is gaining increasing importance. Interestingly, *Tableau*, *JMP*, *QlikView*, and *Spotfire* allow a write-back of modified data entries into the databases.

The integration of various (heterogeneous) data sources has strongly improved over the last five years. Most vendors guide users with visual query interfaces or wizards through the data integration process.

In terms of *ETL* features, almost all systems deliver a wide range of functionalities out-of-the-box, e.g., type guessing, defining derived attributes, or joining tables over multiple attributes, and give technically advanced users the option to define complex queries through scripting, query, or fully-fledged programming languages, such as SQL, R, *MatLab*, Python, or Java. Some systems, e.g., *Advizor* or *Cognos*, even offer bridges to specialized *ETL* in-house solutions.

For (semi-)automatic data preprocessing we see more variation. While filtering by values or the imputation of values (min/max/avg) is supported by all systems in a purely interactive fashion, the *automatic* suggestion of possible data transformations (e.g., a meaningful dimension normalization) is only supported by *Visual Analytics*, *Spotfire* and *Advizor*. Statistically inspired systems, such as *JMP* and *Visual Analytics*—and even *Advizor*, *PowerBI*, and *Tableau* to a lesser extent—go one step further and implement (sub-) sampling algorithms, which are critical for the analysis of Big Data sources. *JMP* even allows analyzing missing value distributions. Anonymization concepts are implemented rudimentarily via calculated fields (*Tableau*, *PowerBI*) or through database views (*Spotfire*, *Visual Analytics*).

Research Discussion

Over the last five years, commercial VA system vendors have put great effort into the ease-of-use and integration of heterogeneous data sources, which is an extensive endeavor implementation-wise. On the other hand, the research community raises the bar in this category and shows visual analytics approaches for a (semi-) automatic data integration.

Data Uncertainty and Trust: Basing the analysis on uncertain data is dangerous. While the research community has put extensive efforts into data uncertainty and trust, the commercial VA system field to date largely ignores this facet. At the same time, workshops such as the IEEE Vis 2014 *Workshop on Provenance for Sensemaking* reflect the importance of this research topic. Sacha et al. reason on the interplay between uncertainty and trust during the knowledge generation process within VA [31]. Bors et al. present a review of uncertainty and provenance methods [32]. On the visualization side, Correa et al. [33] show how different visual mappings support the understanding of uncertainty in data projections.

Data (Pre-)Processing: According to informal inquiries the typical data scientist spends 50% to 80% of the time on data collection and preparation before it can be explored [34]. As a result, an entire commercial landscape focuses only on this topic. Data Wrangler [30] and its corresponding commercial spin-off *Trifacta* most prominently represent the research community. These systems learn from user-guided transformation steps with the aim to automatically infer potentially useful data cleaning procedures and even sequences thereof [35]. Similarly, Wu et al. use an example-based learning algorithm to derive a grammar of potentially useful text editing operations [36]. Heer et al. present a system that even predicts a sequence of future interaction steps [35].

5.2 Automatic Analysis

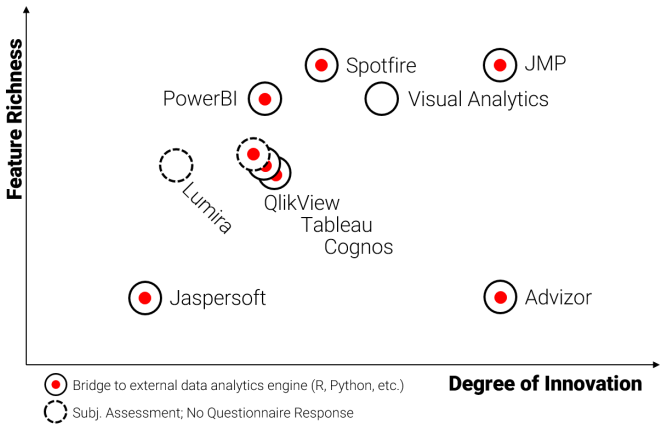


Fig. 6. **Automatic Analysis:** Positioning along the feature richness dimension represents the systems' "out-of-the-box" functionality. Most systems (marked with a red dot) can be enhanced by calling external data analytics tools (e.g., R, Python, or SPSS). Detailed feature overviews can be found in the appendix tables http://commercialtools.dbvis.de/appendix.

This feature comparison category focuses on automatic analysis techniques for statistical, predictive, comparative or high-dimensional data analysis. While previously many systems purely focused on an interactive visual analysis of data, we recognize a paradigm shift towards a full-fledged integration of purely automatic techniques. Unlike in 2012, we can see today that almost all systems offer an option, with variable convenience, to call external data analytics engines, such as R, Python, or SPSS.

To provide a fair comparison, we assess the systems based on their integrated data analysis functionality, as depicted in Figure 6. All systems with an external analysis engine bridge are marked with a red dot. In terms of standard data analysis techniques, e.g., clustering and outlier detection, classification, and regression models, all systems except *Jaspersoft* deliver at least basic functionalities. k-means is the clustering algorithm of choice, decision trees are mostly implemented for classification tasks, and least-square linear and logistic regression models are standard. Functionality-wise *JMP* is outstanding. Their analysis suite offers the most state-of-the-art analysis functionality including self-organizing maps, PCA, k-nearest neighbor classifiers, and decision tree ensembles. Since version 10, *Tableau* includes automatic data analysis functionality out-of-the-box such as regression analysis, outlier detection, and clustering.

Interestingly, prediction functionality is implemented in several systems for instants-based regular 1D time-series (*JMP*, *Tableau*, *Cognos*, *Spotfire*, *Visual Analytics*, *PowerBI*). In contrast, none of the systems support a purely automatic sensitivity or what-if analysis but rather emphasize visual interactive techniques for these use cases. Other loosely related automatic analysis functionalities comprise the support for alert mechanisms or automatic refreshing of data sources, which are supported by *PowerBI*, *Advizor*, *Visual Analytics* and *Tableau*.

One Big Data analysis feature of *Advizor* and *Visual Analytics* is specifically innovative: Their systems are able to calculate (intermediate) visualization results based on either sampling-based calculations, prediction methods, or incremental updates for long-running tasks (*Visual Analytics* supports the latter).

Research Discussion

Five years ago there was an established understanding of Big Data. Today, however, the algorithmic implications become more and more obvious. First, BI decisions should be based on the comparative analysis of potential scenarios. Hence, relying on just one algorithm with potentially hidden parameter settings can be risky. Second, the prevalent credo to store as much information as possible and derive meaning out of it in a fully detached working step leads to complex and potentially high-dimensional data sets. Third, a data analyst does not want to wait hours for simple calculations (further discussed in Section 5.6). A range of VA challenges can be derived from this fact.

Model and Parameter Space Exploration: *JMP* reflects the importance of model comparison and exploration with a statistically inspired listing of *global* error scores. However, Mühlbacher and Piringer show evidence that (algorithmic) models can fit well globally but may be locally inaccurate [37]. Similarly and impressively, the work of Matejka and Fitzmaurice shows how visually distinct scatter plots with identical summary statistics (mean, std. deviation, and Pearson's correlation) can be generated iteratively [38]. This generalization of Anscombe's Quartet emphasizes that VA approaches need to combine automated quality metrics and a respective visualization. VA approaches for comparing and refining the behavior of classification models are presented in [39] and [40]. Cao et al. present a treemap-like glyph for a comparative analysis of multidimensional cluster results [41]. Some recent works focus on cluster comparison and ensemble building [42], [43]. More generally, Sacha et al. claim that a (visual) comparison

of data models can help in the knowledge generation process [44]. Sedlmair et al. present a conceptual framework for a VA-driven parameter space analysis [45].

High-Dimensional Data Analysis: Only elementary high-dimensional (HD) data analysis techniques can be found in the commercial sector. On the other hand, data sets are becoming increasingly complex, particularly in terms of dimensionality, and thus require more sophisticated HD data analysis techniques. While feature selection methods remove irrelevant and redundant dimensions by analyzing the entire data set with global metrics [46], subspace analysis tries to overcome the curse of dimensionality by effectively selecting subsets of dimensions (i.e., subspaces) to allow for descriptive and informative data interpretation [47]. One prominent example are subspace clustering approaches, such as CLIQUE [48] or PROCLUS [49], which consider that many dimensions are irrelevant and can even mask existing clusters. However, while subspace clustering techniques can deliver meaningful results, the number of reported clusters is typically large and may contain substantial redundancy [50].

Often HD analysis techniques depend on pairwise similarities or distances resulting in quadratic runtime and memory complexity with respect to the number of data items. However, recent advances leverage the idea of coresets [51], [52], such as t-SNE [53], or approach with multi-scale analysis, such as Hierarchical Stochastic Neighbor Embedding [54]. These advancements make projection techniques available even for Big Data sources.

Feature Encoding and Learning: For an analysis of non-numeric data sets, such as in image, audio, 3D, or text databases, a descriptive feature encoding is key. For engineered feature descriptors, a large research corpus for the various data domains exists (see also the Research Discussion in Section 5.3). An all-embracing enumeration of the central approaches is out of the scope of this survey. As an alternative to engineered feature extraction methods, supervised algorithmic approaches can be applied to learn generative models that represent the dominant data features. These deep learning approaches, such as described in the form of convolutional neural networks (CNNs) for image classification [55], [56] or in the form of recurrent neural networks (RNNs) for text classification or machine translation [57], [58] require labeled data for training.

While these deep learning algorithms achieve state-of-the-art performance results, finding a suitable neural network configuration (i.e., number of hidden layers and neurons per hidden layer) is often a time-consuming trial-and-error task. In recent years, an increasing amount of work focuses on visualizing neural networks with respect to the different configurations. For example, Abadi et al. show a web-based system for exploring different hidden layer and neuron configurations [59] and Strobelt et al. show an RNN visualization that helps to explore hidden state representations over time [60].

5.3 Complex Data Types

Big Data is defined by the four big V's: Volume (amount of data), Veracity (data uncertainty), Variety (data types), and Velocity (data streams). In terms of variety, data usually does not come as numeric values in well-defined tables. This poses new challenges on data handling (see Section 5.1), but also demands more analysis and visualization capabilities during the entire analysis workflow. (1) During the data loading process, an automatic typing is significantly more challenging. (2) During the analysis, only sophisticated similarity measures can adequately compare complex data types. (3) At the same time, non-standard visualizations are needed to represent essential aspects of complex data types, e.g., trajectories of high-dimensional vectors.

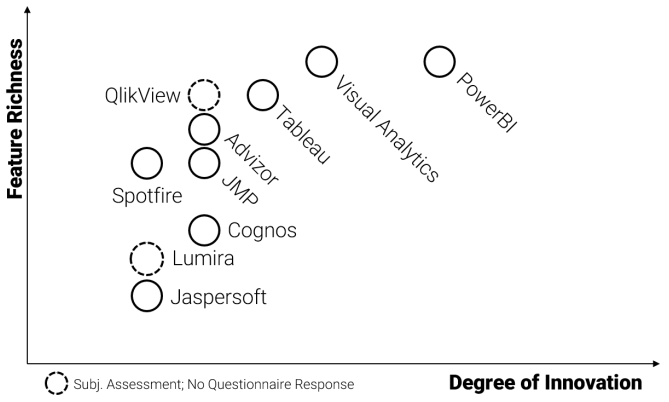


Fig. 7. **Complex Data Types:** All systems support numeric, text, and time-series data. Rankings are based on advanced functionality that comprises support for multimedia and streaming data. Detailed feature overviews can be found in the appendix tables http://commercialtools.dbvis.de/appendix

We find that the support for complex data types has improved over the last five years. In particular, basic text analysis functionality and the support for geographic positions are commonplace today. However, as shown in Figure 8 even more data types are available. Data type support varies greatly and reflects the focus of each distinct system. For example, image analysis has gotten more and more common. *PowerBI* and *Visual Analytics* support most complex data types, closely followed by *Advizor* and *Tableau*. Nonetheless, three dimensional objects, which will gain additional importance with 3D-printing, are not supported by any system. *Cognos* is the only system that supports video analysis. Interestingly, fewer than half of the systems support relational data and networks despite the growing research interest. Some systems offer separate extensions to support more data types, e.g., *Spotfire* for networks.

Big Data sources vary significantly with respect to quality. Dealing with uncertainty in gathered data has become more critical as, for example, sensor networks value throughput over accuracy. Commercial VA systems have begun to adapt to this new demand by increasing support for numeric and error intervals.

High velocity, the third characteristic of Big Data, especially drives our innovativeness assessment. Real-time analysis of data streams adds new applications of data analysis, such as production chain monitoring. Microsoft *PowerBI* sets new standards and allows users to deal with streaming data as it approaches (single item or batch mode) by supporting real-time APIs, like PubNub, and its proprietary real-time APIs MS Flow and MS Azure. *Visual Analytics* natively supports streaming data as well. Other systems offer updates at fixed time intervals or manual one-click updates.

Research Discussion

Initial approaches towards the *Variety* aspect of Big Data have been conducted over the last few years. As more specific forms of Big Data evolve, more targeted commercial products will fill these specific gaps (see: Discussion in Section 8).

Complex Data Type Analysis: The analysis of complex data types inevitably demands a feature extraction step. If the analysis becomes even more sophisticated, then similarity functions to

	Supported Data Types	Advizor	Cognos	Jaspersoft	JMP	PowerBI	Spotfire	Tableau	Visual Analytics	Lumira*	QlikView*
Data Types	Numeric with error				✓	✓	✓	✓	✓		
	Numeric intervals	✓			✓	✓	✓		✓		
	Complex time series	✓			✓	✓		✓	✓		✓
	Item Sets	✓	✓			✓		✓	✓		✓
	Relational/Networks	✓				✓		✓	✓		✓
	Geographic positions	✓	✓	✓	✓	✓	✓	✓	✓		✓
	Geographic traces					✓	✓	✓			✓
	Geographic areas	✓	✓			✓	✓	✓	✓	✓	✓
	Text	✓	✓	✓	✓	✓	✓	✓	✓		
	String-like (e.g. DNA)	✓				✓					
	Image		✓	✓	✓	✓	✓		✓		
	Video		✓								

Fig. 8. **Supported data types:** All commercial VA systems support standard data types including geographic positions and text. None supports audio, three-dimensional objects, or images with medical domain specific meta-data (e.g., DICOM). Blue marks were manually added as *Lumira* and *QlikView* did not answer our questionnaire.

compare the feature vectors will play a role (e.g., for clustering or filtering-by-content). The choice of feature vectors and similarity functions is a central research challenge; it often requires knowledge of the application context, and sometimes even the user. To date, a significant number of feature extraction methods have been proposed for different types of structured data [61], [62]. However, descriptors are often defined in a heuristic way and yield rather abstract information, which is difficult to interpret and leverage by non-expert users. Thus, it remains difficult to decide which descriptor to choose for the retrieval or analysis problem at hand.

Quality Metrics: Recently, feature-based approaches have been introduced to assess the visual quality of data representations in order to guide users in the exploration process [63]–[65]. The idea is to search automatically for an improved or alternative view of interest to the user. Doing so describes the data on a space that is different from the data itself. In these approaches, the feature space is based on the characteristics of visual patterns [63], [64], [66] and has the advantage that images are closely related to what the user inspects, namely, a visualization of the considered data.

Streaming Data: Velocity in Big Data, a substantial research topic of interest, is only roughly reflected by the commercial VA sector. For example, Vehlow present a VA approach for showing dynamic aspects in social networks [67]. Fischer et al. focus on real-time analysis and visualization of network traffic data [68], [69]. Liu et al. [70] and Keim et al. [71] explore text and topic developments over time. More generally, Wanner et al. survey in [72] the state-of-the-art for event detection in text data streams, such as in the analysis of social media sources.

5.4 Visualization

The combination of graphical interfaces and interaction techniques, such as brushing and linking, faceted filtering, or Focus+Context, enables a visual analysis of the underlying data set. Similar to the automatic analysis (see: Section 5.2), we see that the integrated feature richness has stagnated, while extensibility has increased drastically. Positive counter-examples for a consequent extension of integrated visualization features are *Spotfire*, *Visual Analytics*, and *Advizor*. Since 2012, now *Spotfire* allows rendering of pixel bar charts, *Advizor* offers extended parallel coordinate plots and SAS *Visual Analytics*ntegrates icicle plots.

Almost all systems offer some directly accessible bridges to visualization libraries. For example, *Tableau* offers an entire SDK

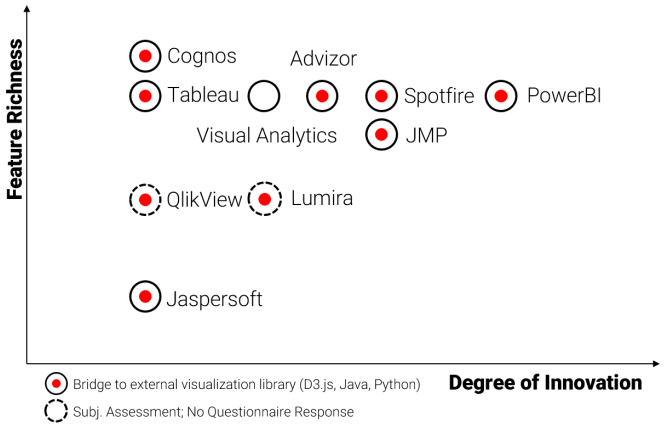


Fig. 9. **Visualization:** All systems support standard visualizations, such as bar, pie, and line charts, box, and scatter plots, treemaps and geographical maps. Advanced functionality comprises support for, amongst others, node-link diagrams, heatmaps, dense-pixel displays, glyphs, icicle plots or horizon charts. Detailed feature overviews can be found in the appendix tables http://commercialtools.dbvis.de/appendix

for writing custom C/C++, Java, or Python code. Users can extend *Spotfire*, *JMP*, or *Cognos* with custom visualizations through either proprietary scripting languages or Javascript-based visualization libraries (e.g., D3.js). *PowerBI* pursues an interesting expandability approach: developers can provide visualizations as free or paid add-ins through a dedicated app store, thus giving users without programming experience the possibility to employ sophisticated visualizations.

In the commercial sector, the definition of glyph design as one generic solution to visualize HD datasets is restricted to changing the shape appearance and color-mapping. Glyph-based visualizations, such as combining glyph designs and 2D layouts, are not yet standard. The visualization layout for coordinated views is mostly restricted to grid-like arrangements where a potential time-dependency is reflected by juxtapositioning or animation (*PowerBI*, *Tableau*, *Lumira*, *JMP*). Superpositioning/layering of visualizations is mostly available in combination with geographical maps; *Advizor* offers a rudimentary space-time cube metaphor.

Beyond WordClouds, text visualizations are not in the focus of the commercial field. We even notice that the visual and automatic support for (un-)structured text is limited in most systems.

The structured presentation of (visualization) insights also falls into this category. *Advizor*, *JMP*, *PowerBI*, *Spotfire*, and *Visual Analytics* offer a journal or linear history metaphor for this purpose. More advanced provenance features, such as tree-like histories (*JMP*, *Spotfire*) and data-flow graphs (*Spotfire*, *Advizor*), are rarely seen in the commercial VA landscape. *Spotfire*, *PowerBI*, *JMP*, and *Advizor* are specifically innovative for their tracking and saving of user interactions and even data selections. An automatic and on-the-fly analysis of these user logs will open up new possibilities for visual analytics, where the system systematically guides the user to new findings.

All product vendors state that their visualization engines are capable of supporting an entire variety of output devices, such as Large-scale/Powerwall-sized displays, (multi-display) desktop environments, tablet, phone or even simultaneous combinations of the above. *PowerBI* is particularly innovative with its support for HoloLens augmented reality headsets and Apple's smartwatch. *PowerBI*, *Jaspersoft*, and *Tableau* even support customized visualization presentations per specific device type.

Research Discussion

We claim that there is a vast gap between the current state-of-the-art in research and the commercial system landscape. Although it is known that it takes up to ten years in the software industry to transition research results into commercial systems, we see that advanced visualizations (i.e., beyond line, pie, and bar charts, and scatter plots) as well as advanced interaction metaphors remain in the research community.

Interaction Design: We see a consequently adapted brushing-and-linking and interactive filtering functionality in all systems. However, most VA research prototypes rely on advanced drill-down functionality [73]–[75]. Multi-scale analysis, such as presented for geo-related data [76] or text data [43], view distortion and adoption techniques, such as presented for Treemaps by Tu and Shen [77], as well as novel navigation concepts, such as link-sliding [78], are only found in research prototypes.

Adaptive/Scalable Visualizations: The commercial VA sector works primarily with interactive, non-adaptive InfoVis-like charts. Neither advanced concepts, such as semantic zoom functionality [79], [80], which adapts the visual metaphors based on the current information-aggregation level, nor off-screen visualization approaches [81] are employed to overcome the typical Big Data information overload. As mentioned above, all systems rely on a grid-like layout of charts. Innovative analytic provenance techniques, such as presented by Heer et al. [82], in VisTrails [83], or in Small Multiples & Large Singles [84], could be generic alternatives.

Visualization Grammar: In recent years considerable efforts try to define and improve declarative languages for visualization design [15], [85], [86] on which, e.g., *Tableau* is built (cf. Polaris [24]). Recently, Vega-Lite has been introduced by Satyanarayan et al. [87] as a high-level grammar for specifying complex interaction concepts on top of the declarative visualization language Vega [88]. Additionally, recent work such as Lyra [89] or iVisDesigner [90] present authoring tools, which allow composing visualizations with multiple layers and annotations.

5.5 User-Guidance, Perception, Cognition

This section summarizes a broad scope spanning from aspects of perception, such as the pre-attentiveness of some visual artifacts via cognitive features (e.g., trust and cognitive biases), all the way to decision support in overwhelming or confusing situations. Systems with a focus on these considerations present users with reasonable defaults and guide them through the analysis process.

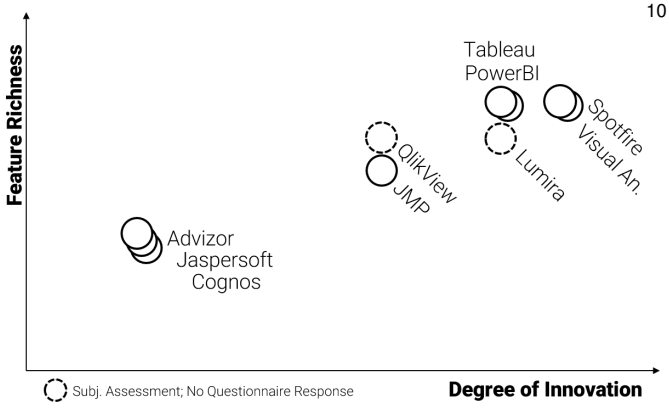


Fig. 10. **User-Guidance and Perception, Cognition** With respect to human-centered analysis, prior leaders *Spotfire* and *Tableau* are challenged by *Visual Analytics*, *PowerBI*, and *Lumira*. Detailed feature overviews can be found in the appendix tables http://commercialtools.dbvis.de/appendix

While the specification granularity and feature richness is drastically improved in today's VA systems, mainly due to the integration of analytic and visualization bridges to external interfaces, we see that more elaborate guidance through the visualization design process needs to be established. To reduce the cognitive overload, four systems suggest potentially useful visualizations, such as with "Show Me" buttons (*Tableau*, *JMP*, *Visual Analytics*). *Lumira* and *Spotfire* go one step further and analyze not only the data type but also value distributions for their suggestions. *Lumira* presents a ranked list of "Related Visualizations" to guide users to different data facets for the selected chart option.

Visual query interfaces, especially for the data integration/preprocessing step, are available in all but one system. Basic combinations of data sources, like inner, outer, and cross joins, are hence accessible to users with little pre-knowledge. Most systems even offer data previews to anticipate the effect of queries and data transformations, like changing a data column type. At the same time, almost all systems offer guidance for analysis and visualization tasks via some form of wizard. *PowerBI* distinguishes from the other systems by providing a natural language interface that translates into visual queries and parameter settings.

Sophisticated analysis and visualization operations require advanced scripting capabilities, which restricts the user group for these functionalities. As an alternative, two systems (*Tableau* and *Spotfire*) entirely rely on drag-and-drop interfaces throughout their data analysis. Scripting is an interesting feature when it comes to automatization of reoccurring tasks, too. Five systems (*Advizor*, *JMP*, *PowerBI*, *Spotfire*, *Visual Analytics*) offer this functionality. *Spotfire* and *Advizor* even support building macros using wizards. *JMP* uniquely supports recording and rerunning series of procedures.

On the perception side, many systems demand the construction of custom color maps for color-blind people. None of the systems promote pre-attentive combinations of visual variables actively.

Research Discussion

From a research perspective, we see progress in the area of general usability. On the other hand, many opportunities exist to

make the workflow guidance even better. As an example, recent advances in color perception research are not yet reflected [91], [92]. These research results show that an appropriate selection of the used color scheme will mitigate contrast biases or improve the visualization readability and understandability. For example, systems could suggest better color maps tailored to the current user tasks [93].

Guided Exploration: In general the workflow model implemented by the systems is rather idealistic and did not enter the Big Data era. While the system infrastructure was enhanced and is now more capable of dealing with large amounts of data (see Section 5.6), users are still left alone in an overwhelming visualization space too large to explore manually. While initial approaches towards guided exploration (e.g., *Show Me* visualization recommendations) have been made, significantly more effort could be devoted to the question "What comes next in my data exploration?" A general purpose framework for a user-guided, interest-driven exploration of high-dimensional datasets is presented by Behrisch et al. [94]. The importance and facets of guidance during the VA process are emphasized by Ceneda et al. [95]. Tang et al. [96] suggest top insights from multi-dimensional data. Showing useful subsets of possible visualizations [97]–[99] followed by feedback-driven exploration [100] and human-centered machine learning [101] could be a suitable workflow for tackling overwhelming search spaces. Measures of both interestingness [102] and trust are necessary to guide users with diverse levels of expertise.

Quality Assessment: On the other hand, more open and user-centered workflows have the potential to increase problems of biases. Researchers put significant effort in cognitive aspects regarding uncertainty and trust [31]. Quality metrics for visualizations are one research approach in this direction, represented by an entire range of **-gnostics* papers [65], [66], [103] as already mentioned in the Discussion in Section 5.3. These papers present a space of potentially interesting and interpretable visual patterns to the user. Chen proposes a promising approach for in-situ quality assessment of automatic analysis and algorithm choices [104]. Further, statistical evaluation should be included *automatically*. Systems should notify and guide the user in case of data quality issues or a wrong choice of an analysis approach (i.e., using the wrong statistical test).

Analytic Provenance: Data manipulation and algorithmic transformations can significantly change the data interpretability. Consequently, exploratory findings should be justifiable from analytic provenance systems. VisTrails [83], Knime [105], the graphical histories of Heer et al. [82], and Small Multiples & Large Singles [84] are prominent examples for the usefulness of this approach.

5.6 Infrastructure

First, almost all systems are available in a client-server structure, on-premise or off-site in the cloud. hic stand alone software is gone We claim that the time of monolithic stand-alone data analysis applications is over. This claim manifests Nowadays, a wide range of presentation device types are supported, such as Powerwall-sized displays, smartphones, and even smartwatches (*PowerBI*). Secondly, many companies offer a rich set of loosely coupled cohesive modules. For example, *Cognos* integrates with *SPSS Predictive Analysis* and Tibco offers *Jaspersoft ETL* and *Advanced Data Services* as on-demand features. Thirdly, systems integrate third-party and open-source software. In their 2012 survey, Zhang

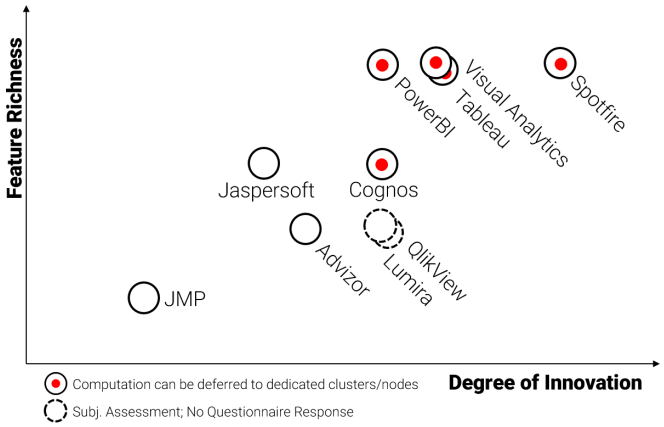


Fig. 11. **Infrastructure** Diverse architectures lead to a wide-spread field. In general, systems benefit from scalable, modular architecture designs. Practically, specific use case demands will define which architecture best meets these demands.

et al. [4] emphatically requested the integration of commonly used scripting languages, like R or Python. Today, most systems provide this functionality at several stages in the analysis process, e.g., data import, calculation of derived dimensions, or automatic analysis. Hence, integrating (intermediate) results into existing workflows becomes more important than offering a complete solution.

Lastly, demand has grown beyond simple desktop machines with the increasing diversity of data hardware. While multi-core support is standard, exploitation of one or more GPUs is only supported by some products (*Spotfire*, *Visual Analytics*). Distributing workload to dedicated compute clusters/nodes is more common and logical in combination with a client-server architecture.

The most prevalent memory concept is in-memory. However, this is limited by available hardware and hybrid storage concepts are necessary for very large datasets. Loading data selectively on-demand can have significant benefits when combining distinct data sources. Many systems introduce an additional layer, which evaluates queries and defers work to data sources when beneficial. For example, *Jaspersoft*, *PowerBI*, and *Cognos* use a query "proxy," which translates, distributes, and joins queries for each database technology. Depending on the type of query and aggregation, some complexity will be pushed down to the databases, to avoid pulling all the data in-memory.

Another aspect of diverse data sources emerges when dealing with sensitive data. Systems need to take care of multiple privilege structures. All systems externalize this problem to the underlying databases. Similarly, anonymization of sensitive data still remains a manual task. While desktop appliances are designed for one user per installation, almost all client-server designs include multi-tenancy capabilities.

We judge the innovation degree in this feature category by the systems' support for handling and effectively distributing long-running tasks, among other tests. For example, *Spotfire* allows sampling-based approaches that can be calculated either on a GPU or deferred to dedicated compute cluster nodes. *Visual Analytics* and *Tableau* support incremental calculations.

Research Discussion

Current businesses, research communities, and governments continue to be overwhelmed with Big Data. Overall, the commercial

Infrastructure			Advizor	Cognos	Jaspersoft	JMP	PowerBI	Spotfire	Tableau	Visual Anal.	Lumira*	QlikView*
Architecture		Stand alone	✓		✓	✓	✓	✓	✓			✓
		Client/Server	✓	✓	✓			✓		✓	✓	
		Cloud on premise	✓	✓	✓		✓	✓	✓	✓	✓	✓
		Cloud at product company		✓			✓	✓	✓	✓	✓	✓
		Cloud in Internet	✓	✓	✓		✓	✓	✓	✓	✓	✓
Performance	PC	Multi-CPU support	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		Single-GPU support						✓	✓	✓		
		Multi-GPU support						✓		✓		
	Cloud	Deferrable dedicated compute clusters/nodes		✓			✓	✓	✓	✓	✓	✓

Fig. 12. **Infrastructure:** Almost all commercial VA systems provide a client-server or software-as-a-service infrastructure. *Spotfire* and *Visual Analytics* allow GPU-accelerated computations. Blue marks were manually added as *Lumira* and *QlikView* did not answer our questionnaire.

VA field is diverse as depicted in Figure 11. While these systems profit from the cloud and client-server concepts already, we see that the research community puts an even broader focus on these ideas.

Progressive Analytics: We claim that although progress has been made in terms of scalability, none of the commercial VA systems deliver *exploratory data analysis*. A core challenge is to give the user a feedback-loop with a latency under ∼10 seconds to maintain their efficiency during exploration tasks. A novel analytics paradigm called *progressive data analysis* is a potential solution: results of long-running computations are delivered not in one long pass but in multiple foreseeable steps. Thus, the result quality starts with estimates and improves progressively. Hellerstein presents initial thoughts along these lines with progressive aggregated queries [106]. VA approaches are presented by Williams and Munzner for projections [107] or by Fekete and Primet with the focus on a progressive infrastructures [108] and more generally from a workflow- and trust-perspective by Stolper et al. [109], Fisher et al. [110], and Zgraggen et al. [111].

Dealing with larger datasets: In the geo-spatial visualization domain, we note that current VA prototypes can handle billions of data points. For example, Liu et al. present in imMens an approach for binned aggregations that can also be applied to other domains [112]. Further RAM efficiency considerations are presented in Nanocubes by Lins et al. [113]. More generally, efficient ways to speed up especially the k-nearest neighbor search in large high-dimensional datasets (e.g., as in image or document databases) have emerged in recent years. For example, in binary hashing, data items are mapped to a compact binary vector so that Hamming distances in binary space approximately preserves distances in the original data space [114], [115]. These binary-vector databases are typically an order of two magnitudes smaller than standard feature descriptor databases.

6 Performance Evaluation

Essential challenges of Big Data are the scalability of analytical systems, their capability to analyze large masses of data, and the data representation [116]. To test the capabilities of the VA systems regarding these three significant challenges we carry out a performance evaluation. In practice, many of the VA systems are designed to run on large distributed systems connected to sophisticated Big Data-capable storage solutions, e.g., Google's distributed database Spanner [117].

A complete and all-embracing performance evaluation of all 46 available systems would exceed the scope of this paper. Instead, we focused on evaluating the system performance for the ten systems on our short list in a structured experiment. We conducted nine performance tests in February 2017 and one in March 2018[3]. *Jaspersoft* and *Cognos* were excluded, since we could not ensure the same testing environment. Thus, our final selection of BI systems for the performance evaluation were: *Advizor*, *QlikView*, *Lumira*, *JMP*, *Visual Analytics*, *Tableau*, *Spotfire*, and *PowerBI*.

6.1 Performance Evaluation Methodology

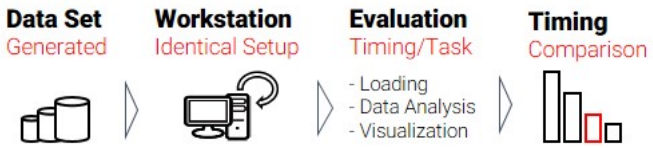


Fig. 13. **Feature Evaluation:** We evaluate the performance of commercial VA systems with three distinct performance tests reflecting the main data analysis steps: (1) Data Loading, (2) Computational Data Analysis, (3) Visualization performance. All performance trials are executed on the same workstation setup with generated, complex data sets.

Data Set: To ensure a fair performance comparison we created random data using the Java-based jFairy[4] fake data generator. Our goal was to create real-world inspired datasets suitable for all later analysis and visualization steps, which includes various complex data types such as dates, geospatial data, and full-text to test the data handling capabilities. We created data sets of sizes 1 GB (1,705,423 rows), 5 GB (8,520,659 rows), 10 GB (17,030,776 rows), 50 GB (85,073,441 rows), 100 GB (170,021,191 rows), and 500 GB (849,248,817 rows) and stored them in a PostgreSQL database.

Workstation: All performance experiments were performed on the same workstation powered by an Intel® Core™ i7-4770 (3.40 GHz) with 32 GB RAM. The OS (Windows 10 Enterprise 64x) and all VA systems were installed on a 240 GB SSD. For storage and temporary data, a 2 TB conventional hard drive was used. After assessing the performance for each of the VA systems, we reset the system to the same clean-installation state. *The tests of Visual Analytics have been performed with an instance running on the SAS Cloud, and are not comparable with the desktop systems. Therefore, the results are marked with a star (*) and hatched in the respective performance bar charts.*

Study Design: Three major challenges of Big Data are scalability, analysis, and representation [116], which we test for in three consecutive tasks.

First, we perform a *loading stress test* to evaluate data handling and scalability in which we measure the time required to load the generated data sets. Some systems, such as *Spotfire*, allow for in-database data analysis with some functional limitations. All data was completely imported from the database to ensure comparability and that all functionality is available. *Results for Advizor and PowerBI are included in the chart, but are not comparable as the data was loaded from a different data source due to compatibility*

3. MS PowerBI responded in March 2018; Product versions can be found online: http://commercialtools.dbvis.de/performance
4. https://github.com/Codearte/jfairy

issues. Therefore, the results are marked with a star (*) and hatched in the respective performance bar charts.

Second, we perform a standard *data analysis task*. We calculate Spearman's rank correlation coefficient for two dimensions. We chose this task since it reflects a common data analysis scenario but is not overbearingly computationally expensive but still contains summarization, multiplication, and division components. Many systems supported this measure out-of-the-box, one exception being *Tableau*, for which we could rely on its R integration.

Third, we carry out a simple *visualization task*. We create a scatter plot visualization from two dimensions on all data set items, ranging from 1,705,423 to 849,248,817 items to test the limits of each system.

6.2 Performance Results

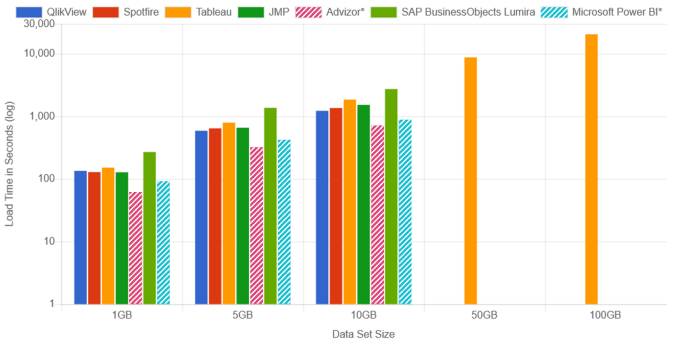


Fig. 14. **Loading Performance**. After *Lumira*, *Tableau* requires the most time to load the data. However, it is the only system capable of loading data sets exceeding the machine's memory. *Advizor* and *PowerBI* are not able to load the data from our standardized PostgreSQL database, but can import from CSV files. Non-existing bars represent the inability to load the data set.

Although the performance evaluation does reflect an imaginary usage scenario of VA systems, we can still observe two categories of systems in use and some common limits: First, in-memory data systems, and second, systems which cache data on the hard drive, such as *Tableau*. As expected, this reflects a trade-off between speed and size. In Figure 14 we see that *Tableau* takes more time to load the data sets compared to the in-memory systems. However, we also note that *Tableau* is the only system capable of loading data sets bigger than 10 GB (not limited by the 32 GB RAM). Nonetheless, no system was able to load the 500 GB data set. A hybrid approach, such as implemented in *KNIME* [105], could provide a solution to partially omit the trade-off between in-memory performance and on disk capacity. For the sake of completeness, we include the results for *Advizor* and *PowerBI*, although we had to load the data from a CSV file.

The performance difference between in-memory and hard-drive cache systems is also visible in the data analysis stress test. In our specific case we calculate the Spearman's rank correlation coefficient for two data dimensions (see Figure 15). We see minor differences between the in-memory systems. *Tableau* needs more time to calculate the correlation coefficient, which could be due to the (de-)serialization required for invoking the R bridge. The results for *Visual Analytics* are not directly comparable since the correlation was measured on the SAS Cloud. However, we included the measurements since it is interesting to see that regardless of the data set size, the analysis performance remains almost constant. For *Advizor*, we could not find a way to calculate the rank correlation coefficient. Similar to *Tableau*, *PowerBI* provides an R bridge,

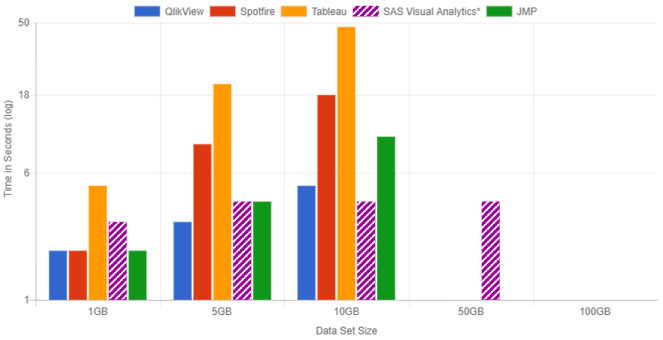


Fig. 15. **Analysis Performance**. We found minor performance differences between the in-memory systems (*JMP*, *QlikView*, *Spotfire*). *Tableau* needs more time to calculate the correlation coefficient, which could be caused by the R integration. *Advizor* and *PowerBI* were excluded from this test, since we could not find a way to calculate the rank correlation coefficient. The results for *Visual Analytics* are not directly comparable since the tests were performed on the SAS Cloud. Non-existing bars represent the inability to perform the analysis task.

allowing us to calculate the correlation coefficient. However, the R bridge of *PowerBI* is limited to 150,000 rows, which did not allow us to perform performance tests on every test data set.

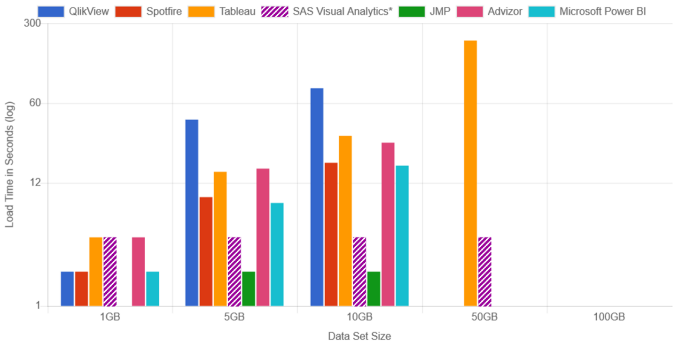


Fig. 16. **Visualization Performance**. *JMP* achieves remarkable results, needing only two seconds to create a scatter plot for a 10 GB data set. *Spotfire*, *Tableau*, and *Advizor* achieve similar performance. One outlier is *QlikView*, which although having the best analysis performance, shows the slowest visualization performance for data sets up to 10GB. The results for *Visual Analytics* are not directly comparable since the tests were performed on the SAS Cloud. Non-existing bars show the systems inability to visualize the data set.

The evaluation of visualization performance shows interesting outcomes, as Figure 16 depicts. *JMP* is the fastest when tasked with displaying large amounts of data. *Tableau*, *Spotfire*, *Advizor*, and *PowerBI* achieve similar performance, with minor differences. One outlier is *QlikView*, which needs more than twice as long as any other system, standing in contrast to the best analysis performance. Again, the *Visual Analytics* results are not directly comparable (computed on the SAS Cloud), but should emphasize the advantages distributed VA infrastructures.

7 Case Study Evaluation

In this section, we report on an informal case study performed with *JMP*, *Lumira*, *QlikView*, *Spotfire*, *Tableau*, *Visual Analytics*, and *PowerBI*, which we conducted in Feb. 2017 and April 2018.[5] We

5. MS PowerBI responded in March 2018; Product versions can be found online: http://commercialtools.dbvis.de/performance

test the systems to derive an understanding of which user groups benefit from their analytical, visualization, and user-guidance features most. Thus, the insights from working with the systems complement the results in Section 5, "Feature Comparison."

7.1 Case Study Evaluation Methodology

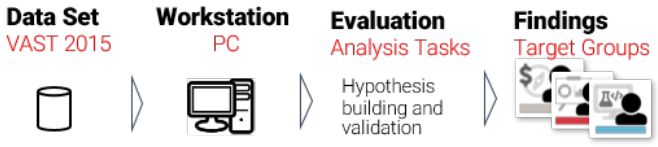


Fig. 17. **Case Study Evaluation:** To derive an understanding of which user groups benefit most from the systems' analytical, visualization and user-guidance features, we conduct an informal case study in which we resemble known data insights from the VAST Challenge 2015 data set.

Data Set: We use data from the 2015 VAST Challenge [21], specifically mini challenges MC1 and MC2, to examine the visualization and analysis capabilities of each system. Both data sets describe three days of visitor movement and communication in a fictitious amusement park and contain around 30,000,000 rows of data with a total size of 1.3 GB. The data contains different attribute types such as strings, dates, timestamps, and geo-coordinates.

Workstation: Technically, except for *Visual Analytics* (hosted on SAS premises), we used the available trial versions and installed these on an Intel® Core™ i5-4590 (3.30 GHz) with 16 GB RAM, SSD, Windows 10 Enterprise x64 (reset after each experiment).

Study Design: Based on the ground truth given in the VAST Challenge 2015 Reviewer Guide [118], we attempt to resemble findings for different visitor groups (MC1) and find patterns of nonoperating attractions (MC2). All experiments were performed by the same person, who participated in the VAST Challenge in 2015. Before trying to mimic the ground truth for each system, we invested at least two days in getting to know the user interface, integrated data analysis capabilities, and how features such as brushing and linking are implemented.

Evaluation Criteria: The following criteria serve as the framework for our experiments: (1) initial visualization using a map or a map-like technique, (2) time-based exploration, e.g., the combination of a visualization with time filters, (3) exploration process with interactive filter and drill-down capabilities, (4) (semi-)automated analysis support to identify groups of individuals or outliers in the data. For the park movement data (MC1) we specifically try to reflect: (1) the visitor distribution per day; opening and closing hours, (2) different visitor groups (people that visit every/only kids-friendly attractions), (3) attractions that were closed during the opening hours. For the network-related communication data (MC2) we attempt to mirror the following findings: (1) identify visitors with a high volume of communication, (2) validate the communication increase on Sunday starting at 11:30AM, (3) recognize devices that sent batches of messages.

7.2 Case Study Results

Next, we summarize our findings, report on outstanding performances, and give general impressions from our experiments.

Data Handling, Selection and Filtering: Each system is able to load the data set from text files (CSV), allows users to customize field types, and provides data previews. As Figure 18 depicts, *JMP* allows specifying (simple) data sampling strategies for loading even large data sets. For quick inspection of data attributes, *JMP*, *Lumira*, and *Spotfire* provide dedicated views that visualize value distribution, and, depending on the attribute type, further meta data (e.g., *Spotfire* in Figure 18). *PowerBI* provides raw data previews of 200 rows by default. Similar insights can be created manually with dedicated views in most of the other systems. All systems support interactive data selection and filter creation besides their manual specification and offer different views based on the current selection. Some applications, e.g., *Visual Analytics* and *Lumira*, maintain a set of global (data set level) and local (current visualization or analysis) filters, which allows quick hypothesis testing after a finding has been made. *PowerBI* has a strong focus on reporting on a number of different pages and therefore provides page-level filtering. Besides *QlikView*, which requires manual scripting for data not in Microsoft Excel format, all systems provide intuitive ways to connect to a variety of different data sources. Notably, *Lumira* provides an advanced data processing stage with instant previews and a large number of operations. We found *PowerBI* convenient to use, as the user interface of the *Power Query Editor* is similar to MS Excel.

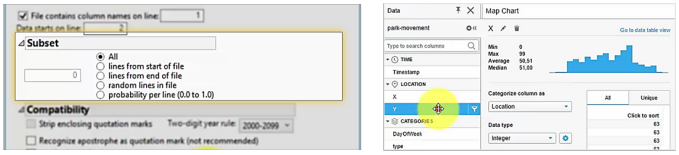


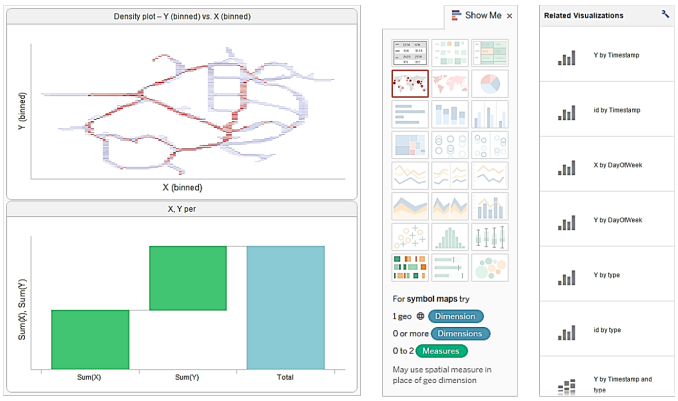
Fig. 18. Data sampling methods for importing and (simple) overview visualizations (e.g., via value distribution histograms) are often secondary analysis features and rarely as good integrated into the analytics workflow as in *JMP* or *Spotfire*.

Visualization: Except for *QlikView*, all tested systems provide a graphical drag-and-drop environment for visualization design. All systems provide facilities to support brushing and linking which allows users to combine different chart types during data exploration. Custom coloring, element size, and interactive selection of parts of the view, as well as the contained data records, are standard. The creation of maps with artificial coordinates is straight forward in some systems, such as *Tableau* and *Spotfire*. Others, however, do not allow non-geographic locations. Similarly, an upper limit of records in parallel coordinate plots, e.g., by *Lumira* or *PowerBI*, limits the exploration capabilities of some systems, as it requires the user to either apply sampling or data filtering beforehand. For time-related visualizations, such as grouped bar charts or multiple line charts, all systems provide good support that allow us to reflect time-related findings easily.

User-Guidance: Some systems provide recommendations for what to visualize given a current data set. Figure 19 shows some examples of this. However, the type of selected data attributes seems to determine these suggestions. For example, the *Show Me* feature of *Tableau* (Figure 19, middle) explicitly states the attribute types required for a visualization.

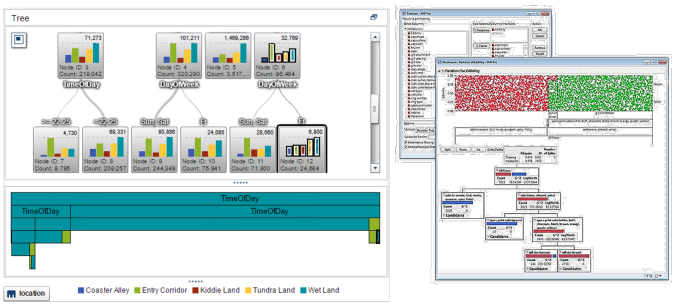
During our experiments, we found that only a minority of systems, such as *Spotfire*, provides previews of recommendations (Figure 19, left). Most systems provide generic previews/symbols.

(Semi-)Automated Support: Automated data analysis that also includes meaningful and sensible default settings for clustering (k-means) or classification (decision trees) is only supported to a limited extent. In our case study, none of the systems was able to produce interesting or striking results without manual intervention.



Spotfire Tableau Lumira

Fig. 19. Related or recommended visualization dialogs of three different systems: *Spotfire* shows data-based previews (left), *Tableau* gives generic previews (middle), and *Lumira* provides textual descriptions (right).



Visual Analytics JMP

Fig. 20. Decision tree construction and exploration in *Visual Analytics* (left) and *JMP* (right).

General Findings: All of the tested systems are able to identify findings from the ground truth. The subjective time to findings varies greatly, as loading the data set in *QlikView* requires scripting, while all other systems are able to load the data by drag and drop or graphical dialogs. Similarly, the time required to create a visualization is higher when using *QlikView*, as the provided wizard and the general visualization customization options exhibit more details compared to the others. For experienced data analysts this is a huge benefit, as almost every aspect of created objects can be customized. In general, we found the recommendations to visualize parts of the data sets not useful, as they mostly make these recommendations based on the type of the data fields. That being said, it is challenging to provide useful suggestions without any prior knowledge of the user's task. All systems have hard limits with respect to what is shown in a visualization display. Various strategies to cope with that limit include automated sampling or error messages stating that there are too many items to show. *Tableau* seems to filter for unique data values and shows them in the visualization, while almost all other systems stop with error messages stating that there are *too many items to show* (*Visual Analytics*), an upper limit of records (10,000) would be exceeded (*Lumira*), or simply that there are *too many values* (*PowerBI*).

8 Discussion and Key Findings

Our discussion focuses on the two primary evaluation criteria of this survey: (1) We reflect our findings from the feature-centered and case-study centered evaluation in a consolidated form and reason for which user groups certain systems might be beneficial. (2) We report our high-level findings and discuss the implications for the current and future commercial and research VA landscape.

8.1 Selecting a Commercial VA System

While ideally one would provide specific guidance on which VA system to select, there remains too many open research questions to provide formal and robust guidelines at this point. Instead, we provide several insights on which specific VA system might be effective, based on our observations in Section 5, and our own empirical knowledge gathered by applying the systems in Section 7. We summarize our findings into an overview table in Figure 21, which highlights the systems' strength with respect to the described data analysis levels on the left and the systems' feature richness scores for the main categories on the right (Data handling/Mgmt., Automatic, and Visualization capabilities). Next, we bridge the gap between our feature-centric and the user-centric considerations outlined in Section 4.1 and report on our insights collected throughout the survey:

The role of the *upper management* challenges the commercial VA sector by requiring not only expressive data visualizations, but also ways of externalizing how these insights where generated (i.e., storytelling capabilities). This has an impact on the expected feature-base of a commercial VA system: (1) Broad range of data visualizations should emphasize distinct—and sometimes even orthogonal—aspects of the underlying insights; (2) Exporting and sharing functionalities are essential to communicate the findings to a broader audience; (3) A non-trivial implication is that the interfaces must allow the upper management themselves to understand and investigate the underlying data. For the upper management, all systems provide reporting or presentation functionalities, e.g., in form of images, PDF, or Office documents. Systems that provide extensive reporting and storytelling facilities, such as *Lumira*, *QlikView*, *PowerBI*, and *Tableau*, give adequate support for this user group and combine (static) presentation functionality, as well as dashboard-like interactivity for selecting (e.g., filters on demand). We must mention *Cognos*, *Advizor*, and *Jaspersoft* specifically in this category for their strong—and advertised—focus on dashboard-driven reporting and analysis functionality. Nearly all vendors provide cloud-based sharing of reports with strongly varying support for interactivity. On the analysis side, drag-and-drop interfaces as in *PowerBI*, *Tableau*, *Spotfire*, or *Visual Analytics* allow even users without programming skills to conduct faceted data research.

The role of a *reporting manager* is characterized by a strong domain knowledge, which allows this group to formulate, reject, and confirm hypotheses. From this standpoint we can consequently derive certain expected features: (1) A quick and understandable formulation of data hypothesis is the core requirement for this user group; (2) Data manipulation operations have to be integrated seamlessly into the systems and should not hinder a fluent exploration process; (3) Expressive power to emphasize aspects of their findings. For reporting managers, *Cognos*, *Spotfire*, *Lumira*, *QlikView*, *Tableau*, and *Visual Analytics* provide an easy to use interface for data analysis and visualization, support brushing and linking, as well as interactive and visual definition of data filters. In particular, the interfaces of *Cognos*, *PowerBI*, *Spotfire*, *Lumira*, and *Tableau* allow the user to change quickly between a large amount of potentially useful visualizations and modify their data mappings. The definition of derived attributes is specifically well integrated in *Cognos* and *Lumira* where even derived dimensions/attributes can be specified without scripting knowledge. On the automatic

Fig. 21. Commercial VA Systems are designed with specific target user groups and usage scenarios in mind: While systems like *Advizor*, *QlikView* or *Jaspersoft* especially target the dashboarding/result presentation market, systems like *Cognos* or *Lumira* are designed with hypothesis definition/validation in mind. The main leaders in the field, *JMP*, *Spotfire*, *Tableau*, *Visual Analytics* and *PowerBI* try to satisfy all user groups.

analysis side *JMP*, followed by *Spotfire*, give the reporting manager a full range of algorithmic data analysis support for clustering, classification, and regression analysis tasks.

The *data analyst* poses the highest requirements on commercial VA systems in terms of functionality. While their typical workflow makes extensively use of the integrated, interactive data analysis features, their typical analysis questions often require the full data mining toolbox. Hence, this user group expects the following features: (1) An easy-to-use and fully functional extension capability to integrate a multitude of data and visualization libraries; (2) A well-thought integration of these bridges into the overall analysis workflow within the commercial VA system; (3) Big data capabilities focusing not only on Volume (amount of data) but also on Veracity (data uncertainty), Variety (data types), and Velocity (data streams). For the data analyst, *JMP*, *Tableau*, *Visual Analytics*, *Spotfire*, and esp. *PowerBI* provide methods to apply deep methodological knowledge and extensive customization possibilities. *JMP* especially offers a wide variety of integrated data analysis algorithms. *Spotfire* and *Tableau*, on the other hand, put a strong emphasis on hiding the complexity of these algorithms. As stated before, the extension capability of commercial VA systems has increased tremendously and is built on integrating third-party libraries or proprietary analysis products, such as in the SAS VIYA product suite. On the data analysis side, the extensions are mostly integrated into the analysis workflow by generating derived or calculated columns and on the visualization side, mostly rely on JavaScript APIs, such as in *Tableau* or *Spotfire*. *Spotfire* and *Visual Analytics* handle the Big Data capabilities of variety and velocity particularly well. For example, *Spotfire* allows users to analyze time series with even sophisticated analysis means (clustering, prediction) and *Visual Analytics* mentions real-time image processing capabilities. The notion of veracity (data uncertainty), on the other hand, has scarcely found its way into the commercial VA sector.

8.2 Summary of Key Findings

Although we have to recognize that the commercial VA landscape has matured and enters a consolidation phase, we also see developments and market insights with respect to the following topics:

Landscape Evolution: Over the last five years we witnessed that the core set of players remained stable, while other system vendors diversified to adapt to specific task requirements. Similar to the results presented in our commercial counterpart "Gartner – Magic Quadrant for Business Intelligence and Analytics Platforms" [119], we see that the systems with a root back in academic research, for example, *Tableau* from Stanford University, *Spotfire* from University of Maryland, and *Advizor* from Bell Labs, remain innovation leaders with respect to interactive visualization and automatic analysis. On the other hand, novel players, such *Visual Analytics* or *Lumira*, backed up by large companies, show feature-rich Visual Analytics suites.

Diversification: Another significant trend is the diversification of the system landscape. While several years ago the great vision was to present an all-embracing VA system, today the systems offer a rich set of secondary software components and bridges. Tibco, SAS, and Qliktech even offer two separate VA-related products to account for the diverse skill-sets of their users (see also: Requirement and User-dependent Task Analysis Section 4.1). As a great success, we value that today Zhang et al's [4] central feature-request is a standard feature: systems do not try to deliver every possible analysis and visualization feature out-of-the-box anymore but offer extensibility (i.e., software bridges) to various specialized commercial or open-source software. However, to date, this functionality is only accessible to users with advanced programming skills. This will hopefully change in the future.

Architecture Design: Another impactful change relates to the "backend evolution." Years ago the accepted architecture was a software monolith with rich data integration functionality. Today, most systems offer at least a cloud service for hosting reports. *Visual Analytics* even follows the fundamentally different approach to center their architecture around a high-performance (off-premise) server instance. This "Lazr" server distributes compute-intensive data analysis and visualization rendering tasks into the SAS cloud. If this architecture becomes predominant, it has critical implications on the declarative specification of visualization and data analysis tasks, the handling and specification of interactivity, and the incremental/progressive result presentation.

Cloud Services: The last critical finding relates to the accessibility of standard VA functionality. We see that an increasing number of vendors offer not only mere data reporting services in the cloud, but also "VA in the Browser". Accordingly, both educated data analysts and less experienced users can explore their data sets. Consequently, novel exploration/representation guidance

functionality needs to be established. To give an example, these systems have to communicate the data/model uncertainty while a regression model is built, instead of requiring users to invoke a "model evaluation" manually. A consequent next step would be that the system itself would have to suggest more suitable models (with better accuracies). On the visualization side, the VA research already shows what an innovative "Show Me" (i.e., visualization recommendation) can look like.

Vendor Self-Assessment: Parts of our questionnaire asked for a subjective assessment of the challenges the vendors foresee for the next five years. Interestingly, two main topics prevailed. First, all vendors acknowledged that the growing data volume and variety demands a new level of "augmented analytics capabilities" to derive results more effectively and efficiently. Future noted cornerstones of successful VA systems included automating time-consuming tasks, such as connecting, cleaning, and mashing up data, injecting more AI capabilities when extracing insights from data models, e.g., through recommendations, forecasts, and proactive alerts, and offering data lineage and control options. While this answer could be expected, the second topic sparked our special attention: three vendors mentioned that, although the systems' analytic capabilities increase by leap and bounds, the "data literacy" within organizations is not growing at the same rate. One participant even mentioned that in many companies the top management still does not trust data analytics. Consequently, the commercial VA field will only maintain its upward trend if data analytics and presentation progress together. Automated and pro-active intelligence will save time in the analysis, but advanced analytics workflows also need to be transparent and accessible to justify decision-making processes.

9 Limitations and Future Work

In this paper, we report on a selection of ten state-of-the-art commercial VA systems. We thoroughly analyze this selection for feature, performance, and workflow-related aspects. Although more commercial VA systems exists (our long list of 46 can be found under http://commercialtools.dbvis.de/systems), our informed selection of VA systems can be regarded as a representative basis for analyzing the current VA system market. Future work should aim to expand this selection, which is necessarily dependent on the vendor's willingness to invest human resources into answering online questionnaires (the average time to finish our online questionnaire was 4h 23min 28sec).

Our selection of questions for the online survey (see also: http://commercialtools.dbvis.de/questions) was intended to examine the systems concerning several broad categories (cf. Section 3.2). During the questionnaire design, we put a specific focus on assessing the Visual Analytics capabilities of the systems. With the development of the VA landscape different question sets may arise that will still fit our assessment categories. This allows for comparability to future results.

One of the key points of the paper is to contrast the developments in the commercial sector with the advances in the VA research community. Therefore, we decided on a selection of trending topics, which could become of interest for the commercial VA sector. This subjective selection process represents a daring view into the future and is not intended to serve as an exhaustive enumeration of indispensable development steps for the sector. Nevertheless, we claim that a critical analysis of development potentials will help to outline a roadmap for improvements in the commercial VA landscape.

Lastly, while we give insights on how to select an appropriate commercial VA system, a systematic matching of feature sets, tasks, skill sets, requirements and user preferences is extremely challenging. We decided against attempting to describe this matching formally as there are still many unknowns and doing so would require a substantial amount of future work. In particular, we do not think this is possible without additionally considering the circumstances of the potential customer, such as, e.g., the existing infrastructure, the amount and diversity of employees, and the complexity of the analytical questions.

10 Conclusion

This survey represents an unprejudiced view onto the commercial Visual Analytics landscape, which is structured along the following evaluation criteria. First, we review the feature-richness and degree of innovation for each of the products' feature groups (Data Handling and Management; Automatic Analysis; Complex Data Types; Visualization; User-Guidance, Perception, Cognition; Infrastructure) and contrast the commercial developments with a selection of recent advances in the research community. With this approach, we hope to establish a "What-can-come-next" view onto respective topics. We complement this feature-driven evaluation by a user-centered view in which we give practical guidance on which systems might be suitable for specific target groups. To be of further practical use, we conduct a system evaluation based on loading, analysis and visualization performance to understand how well "Big Data" requirements are met. Lastly, we apply the systems to the established VAST Challenge dataset to further test usability.

Overall, data analysts will be challenged with a surplus of data and advanced analytics requirements in the future. They will need to build automated mechanisms to create visualizations and predictions. They will also have to make these available to a wide range of business users beyond data scientists. Likewise, entry barriers must be reduced to make analytical tools accessible to the broad mass and to improve data and visualization literacy.

References

[1] IBM Archives, "Herman Hollerith," Online, Feb. 2017, accessed Feb, 15 2017. [Online]. Available: https://www-03.ibm.com/ibm/history/exhibits/builders/builders_hollerith.html

[2] J. Bertin, *La Graphique et le traitement graphique de l'information*, ser. Nouvelle Bibliothèque Scientifique. Flammarion, 1977.

[3] J. W. Tukey, *Exploratory Data Analysis*. Addison-Wesley, 1977.

[4] L. Zhang, A. Stoffel, M. Behrisch, S. Mittelstädt, T. Schreck, R. Pompl, S. Weber, H. Last, and D. A. Keim, "Visual analytics for the big data era - a comparative review of state-of-the-art commercial systems," in *Proc. IEEE Conference on Visual Analytics Science and Technology*, 2012.

[5] C. Howson, A. Woodward, C. J. Idoine, J. L. Richardson, J. Tapadinhas, and R. L. Sallam, "Magic quadrant for analytics and business intelligence platforms," Gartner, Inc., Tech. Rep., 2 2018.

[6] B. Hopkins, S. Sridharan, E. Cullen, and J. Lee, "The forrester wave: Enterprise insight platform suites, q4 2016," Forrester Research, Inc., Cambridge, Tech. Rep., 12 2016.

[7] "The bi survey 18," BARC, Tech. Rep., 2018. [Online]. Available: http://bi-survey.com/

[8] J. Akoka, I. Comyn-Wattiau, and N. Laoufi, "Research on big data a systematic mapping study," *Computer Standards & Interfaces*, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0920548917300211

[9] N. Aggarwal, B. Berk, G. Goldin, M. Holzapfel, and E. Knudsen, *Getting Analytics Right*, 1st ed., S. Cutt, Ed. Beijing, Boston: O'Reilly, 2016.

[10] P. Fernndez, J. M. Santana, S. Ortega, A. Trujillo, J. P. Surez, C. Domnguez, J. Santana, and A. Snchez, "Smartport: A platform for sensor data monitoring in a seaport based on fiware," *Sensors*, vol. 16, no. 3, p. 417, 2016. [Online]. Available: http://www.mdpi.com/1424-8220/16/3/417

[11] S. Zillner, S. Rusitschka, and M. Skubacz, "Big data story: Demystifying big data with special focus on and examples from industrial sectors," EU BIG European Big Data Public Private Forum, White paper, 3 2014.

[12] L. C. Rost, "What i learned recreating one chart using 24 tools," 2016, accessed: 03/15/2017. [Online]. Available: https://source.opennews.org/articles/what-i-learned-recreating-one-chart-using-24-tools/

[13] A. C. Umaquinga C., D. H. Peluffo O., J. C. Alvarado P., and M. V. Cabrera A., *Memorias de las I Jornadas Internacionales de Investigacin Cientfica UTN*. Ibarra, Ecuador: La Universidad Tcnica del Norte, 2016, ch. Estudio descriptivo de tcnicas aplicadas en herramientas Open Source y comerciales para visualizacin de informacin de Big Data.

[14] L. Nair, S. Shetty, and S. Shetty, "Interactive visual analytics on big data: Tableau vs d3.js," *Journal of e-Learning and Knowledge Society*, vol. 12, no. 4, 2016. [Online]. Available: http://je-lks.org/ojs/index.php/Je-LKS_EN/article/view/1128

[15] M. Bostock, V. Ogievetsky, and J. Heer, "D^3 data-driven documents," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 12, pp. 2301–2309, 2011. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2011.185

[16] Tableau Germany, "Top 10 trends der business intelligence für 2016," online Whitepaper, 2016, accessed on 08.12.2016. [Online]. Available: https://www.tableau.com/de-de/learn/whitepapers/top-10-business-intelligence-trends-2016

[17] V. Louise Lemieux, B. Gormly, and L. Rowledge, "Meeting big data challenges with visual analytics: The role of records management," *Records Management Journal*, vol. 24, no. 2, pp. 122–141, 2014. [Online]. Available: http://dx.doi.org/10.1108/RMJ-01-2014-0009

[18] S. Li, S. Dragicevic, F. A. Castro, M. Sester, S. Winter, A. Coltekin, C. Pettit, B. Jiang, J. Haworth, A. Stein, and T. Cheng, "Geospatial big data handling theory and methods: A review and research challenges," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 115, pp. 119–133, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0924271615002439

[19] L. Piovano, D. Garrido, R. Silva, and I. Galloso, "What (smart) data visualizations can offer to smart city science," *Communications and Strategies*, vol. 96, no. 4, pp. 89–112, 2014. [Online]. Available: https://ssrn.com/abstract=2636382

[20] T. Nocke, S. Buschmann, J. F. Donges, N. Marwan, H. J. Schulz, and C. Tominski, "Review: visual analytics of climate networks," *Nonlinear Processes in Geophysics*, vol. 22, pp. 545–570, sep 2015.

[21] "Vast challenge 2015," http://hcil2.cs.umd.edu/newvarepository/VAST%20Challenge%202015/challenges/Mini-Challenge%201/, last Accessed: 20-03-2017.

[22] S. K. Card, J. D. Mackinlay, and B. Shneiderman, *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999.

[23] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD Process for Extracting Useful Knowledge from Volumes of Data," *Communications of the ACM*, vol. 39, no. 11, pp. 27–34, Nov. 1996.

[24] C. Stolte, D. Tang, and P. Hanrahan, "Polaris: A system for query, analysis, and visualization of multidimensional relational databases," *IEEE Trans. Vis. Comput. Graph.*, vol. 8, no. 1, pp. 52–65, 2002. [Online]. Available: http://dx.doi.org/10.1109/2945.981851

[25] P. Hanrahan, "Vizql: a language for query, analysis and visualization," in *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*. ACM, 2006, pp. 721–721.

[26] S. Kandel, A. Paepcke, J. M. Hellerstein, and J. Heer, "Enterprise data analysis and visualization: An interview study," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2917–2926, 2012.

[27] E. Kandogan, A. Balakrishnan, E. M. Haber, and J. S. Pierce, "From data to insight: Work practices of analysts in the enterprise," *IEEE Computer Graphics and Applications*, vol. 34, no. 5, pp. 42–50, 2014.

[28] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org

[29] P. Mayring, "Qualitative content analysis," *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, vol. 1, no. 2, 2000. [Online]. Available: http://www.qualitative-research.net/index.php/fqs/article/view/1089

[30] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer, "Wrangler: Interactive visual specification of data transformation scripts," in *ACM Human Factors in Computing Systems (CHI)*, 2011. [Online]. Available: http://vis.stanford.edu/papers/wrangler

[31] D. Sacha, H. Senaratne, B. C. Kwon, G. P. Ellis, and D. A. Keim, "The role of uncertainty, awareness, and trust in visual analytics," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 1, pp. 240–249, 2016. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2015.2467591

[32] C. Bors, T. Gschwandtner, S. Miksch, and J. Gärtner, "Qualitytrails: Data quality provenance as a basis for sensemaking," in *Proceedings of IEEE VIS Internation Workshop Analytic Provenance for Sensemaking*, 2014.

[33] C. D. Correa, Y. Chan, and K. Ma, "A framework for uncertainty-aware visual analytics," in *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology, IEEE VAST 2009, Atlantic City, New Jersey, USA, 11-16 October 2009, part of VisWeek 2009*, 2009, pp. 51–58. [Online]. Available: http://dx.doi.org/10.1109/VAST.2009.5332611

[34] "Trifacta: Data wrangling tools and software," https://www.trifacta.com/, accessed: 2018-06-21.

[35] J. Heer, J. M. Hellerstein, and S. Kandel, "Predictive interaction for data transformation," in *CIDR 2015, Seventh Biennial Conference on Innovative Data Systems Research, 2015*, 2015. [Online]. Available: http://www.cidrdb.org/cidr2015/Papers/CIDR15_Paper27.pdf

[36] B. Wu, P. Szekely, and C. A. Knoblock, "Learning data transformation rules through examples: Preliminary results," in *Proceedings of the Ninth International Workshop on Information Integration on the Web*, ser. IIWeb '12. New York, NY, USA: ACM, 2012, pp. 8:1–8:6. [Online]. Available: http://doi.acm.org/10.1145/2331801.2331809

[37] T. Mühlbacher and H. Piringer, "A partition-based framework for building and validating regression models," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 12, pp. 1962–1971, 2013. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2013.125

[38] J. Matejka and G. Fitzmaurice, "Same stats, different graphs: Generating datasets with varied appearance and identical statistics through simulated annealing," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2017, pp. 1290–1294.

[39] A. Kapoor, B. Lee, D. S. Tan, and E. Horvitz, "Interactive optimization for steering machine classification," in *Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010, Atlanta, Georgia, USA, April 10-15, 2010*, 2010, pp. 1343–1352. [Online]. Available: http://doi.acm.org/10.1145/1753326.1753529

[40] J. Talbot, B. Lee, A. Kapoor, and D. Tan, "Ensemblematrix: Interactive visualization to support machine learning with multiple classifiers," in *ACM Human Factors in Computing Systems (CHI)*, 2009. [Online]. Available: http://vis.stanford.edu/papers/ensemblematrix

[41] N. Cao, D. Gotz, J. Sun, and H. Qu, "Dicon: Interactive visual analysis of multidimensional clusters," *IEEE transactions on visualization and computer graphics*, vol. 17, no. 12, pp. 2581–2590, 2011.

[42] P. Kothur, M. Sips, H. Dobslaw, and D. Dransch, "Visual analytics for comparison of ocean model output with reference data: Detecting and analyzing geophysical processes using clustering ensembles," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1893–1902, 2014.

[43] Y. Zhang, W. Luo, E. A. Mack, and R. Maciejewski, "Visualizing the impact of geographical variations on multivariate clustering," *Comput. Graph. Forum*, vol. 35, no. 3, pp. 101–110, 2016. [Online]. Available: http://dx.doi.org/10.1111/cgf.12886

[44] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. P. Ellis, and D. A. Keim, "Knowledge generation model for visual analytics," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 1604–1613, 2014. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2014.2346481

[45] M. Sedlmair, C. Heinzl, S. Bruckner, H. Piringer, and T. Möller, "Visual parameter space analysis: A conceptual framework," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 2161–2170, 2014. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2014.2346321

[46] H. Liu and H. Motoda, *Computational methods of feature selection*. CRC Press, 2007.

[47] H.-P. Kriegel, P. Kröger, and A. Zimek, "Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 3, no. 1, p. 1, 2009.

[48] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, *Automatic subspace clustering of high dimensional data for data mining applications*. ACM, 1998, vol. 27, no. 2.

[49] C. C. Aggarwal, J. L. Wolf, P. S. Yu, C. Procopiuc, and J. S. Park, "Fast algorithms for projected clustering," in *ACM SIGMoD Record*, vol. 28, no. 2. ACM, 1999, pp. 61–72.

[50] M. Hund, I. Färber, M. Behrisch, A. Tatu, T. Schreck, D. A. Keim, and T. Seidl, "Visual quality assessment of subspace clusterings," in *KDD Workshop – Interactive Data Exploration and Analytics (IDEA)*, 2016.

[51] P. K. Agarwal, S. Har-Peled, and K. R. Varadarajan, "Geometric approximation via coresets," *Combinatorial and computational geometry*, vol. 52, pp. 1–30, 2005.

[52] D. Feldman and M. Langberg, "A unified framework for approximating and clustering data," in *Proceedings of the forty-third annual ACM symposium on Theory of computing*. ACM, 2011, pp. 569–578.

[53] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[54] N. Pezzotti, T. Höllt, B. P. F. Lelieveldt, E. Eisemann, and A. Vilanova, "Hierarchical stochastic neighbor embedding," *Comput. Graph. Forum*, vol. 35, no. 3, pp. 21–30, 2016. [Online]. Available: http://dx.doi.org/10.1111/cgf.12878

[55] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[56] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[57] J. L. Elman, "Finding structure in time," *Cognitive science*, vol. 14, no. 2, pp. 179–211, 1990.

[58] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[59] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[60] H. Strobelt, S. Gehrmann, B. Huber, H. Pfister, and A. M. Rush, "Visual analysis of hidden state dynamics in recurrent neural networks," *arXiv preprint arXiv:1606.07461*, 2016.

[61] Y. Rui, T. S. Huang, and S. Chang, "Image retrieval: Current techniques, promising directions, and open issues," *J. Visual Communication and Image Representation*, vol. 10, no. 1, pp. 39–62, 1999. [Online]. Available: https://doi.org/10.1006/jvci.1999.0413

[62] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, 2000.

[63] A. Dasgupta and R. Kosara, "Pargnostics: Screen-space metrics for parallel coordinates," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 1017–1026, Nov 2010. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2010.184

[64] D. J. Lehmann, F. Kemmler, T. Zhyhalava, M. Kirschke, and H. Theisel, "Visualnostics: Visual guidance pictograms for analyzing projections of high-dimensional data," *Computer Graphics Forum*, vol. 34, no. 3, pp. 291–300, 2015. [Online]. Available: http://dx.doi.org/10.1111/cgf.12641

[65] M. Behrisch, B. Bach, M. Hund, M. Delz, L. von Rüden, J.-D. Fekete, and T. Scheck, "Magnostics: Image-based Search of Interesting Matrix Views for Guided Network Exploration," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 31–40, Oct. 2017. [Online]. Available: magnostics.dbvis.de

[66] L. Wilkinson, A. Anand, and R. L. Grossman, "Graph-theoretic scagnostics," in *IEEE Symp. on Information Visualization (InfoVis 2005)*, 2005, p. 21. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/INFOVIS.2005.14

[67] C. Vehlow, F. Beck, and D. Weiskopf, "Visualizing dynamic hierarchies in graph sequences," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 10, pp. 2343–2357, 2016. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2015.2507595

[68] F. Fischer, F. Mansmann, and D. A. Keim, "Real-time visual analytics for event data streams," in *Proceedings of the 27th Annual ACM Symposium on Applied Computing*. ACM, 2012, pp. 801–806.

[69] F. Fischer and D. A. Keim, "Nstreamaware: Real-time visual analytics for data streams to enhance situational awareness," in *Proc. of the 11th Workshop on Visualization for Cyber Security*. ACM, 2014, pp. 65–72.

[70] S. Liu, J. Yin, X. Wang, W. Cui, K. Cao, and J. Pei, "Online visual analytics of text streams," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 11, pp. 2451–2466, 2016. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2015.2509990

[71] D. A. Keim, M. Krstajic, C. Rohrdantz, and T. Schreck, "Real-time visual analytics for text streams," *IEEE Computer*, vol. 46, no. 7, pp. 47–55, 2013. [Online]. Available: http://dx.doi.org/10.1109/MC.2013.152

[72] F. Wanner, A. Stoffel, D. Jäckle, B. C. Kwon, A. Weiler, and D. A. Keim, "State-of-the-Art Report of Visual Analysis for Event Detection in Text Data Streams," in *EuroVis - STARs*, R. Borgo, R. Maciejewski, and I. Viola, Eds. Eurographics Association, 2014, pp. 125–139.

[73] B. Bach, N. Henry-Riche, T. Dwyer, T. Madhyastha, J.-D. Fekete, and T. Grabowski, "Small MultiPiles: Piling Time to Explore Temporal Patterns in Dynamic Networks," *Computer Graphics Forum*, 2015. [Online]. Available: https://hal.inria.fr/hal-01158987

[74] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister, "Upset: Visualization of intersecting sets," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 1983–1992, 2014. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2014.2346248

[75] S. Gratzl, A. Lex, N. Gehlenborg, H. Pfister, and M. Streit, "Lineup: Visual analysis of multi-attribute rankings," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 12, pp. 2277–2286, 2013. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2013.173

[76] J. Wang, X. Liu, H.-W. Shen, and G. Lin, "Multi-resolution climate ensemble parameter analysis with nested parallel coordinates plots," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 81–90, 2017.

[77] Y. Tu and H. Shen, "Balloon focus: a seamless multi-focus+context method for treemaps," *IEEE Trans. Vis. Comput. Graph.*, vol. 14, no. 6, pp. 1157–1164, 2008. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2008.114

[78] T. Moscovich, F. Chevalier, N. Henry, E. Pietriga, and J. Fekete, "Topology-aware navigation in large networks," in *Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI 2009, Boston, MA, USA, April 4-9, 2009*, 2009, pp. 2319–2328. [Online]. Available: http://doi.acm.org/10.1145/1518701.1519056

[79] J. Abello and F. van Ham, "Matrix zoom: A visual interface to semi-external graphs," in *10th IEEE Symposium on Information Visualization (InfoVis 2004), 10-12 October 2004*, 2004.

[80] N. Elmqvist, T.-N. Do, H. Goodell, N. Henry, and J.-D. Fekete, "ZAME: Interactive Large-Scale Graph Visualization," *IEEE Pacific Visualization Symposium*, pp. 215–222, Mar. 2008. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4475479

[81] F. Perteneder, E.-M. B. Grossauer, J. Leong, W. Stuerzlinger, and M. Haller, "Glowworms and fireflies: Ambient light on large interactive surfaces," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ser. CHI '16. New York, NY, USA: ACM, 2016, pp. 5849–5861. [Online]. Available: http://doi.acm.org/10.1145/2858036.2858524

[82] J. Heer, J. D. Mackinlay, C. Stolte, and M. Agrawala, "Graphical histories for visualization: Supporting analysis, communication, and evaluation," in *IEEE Information Visualization (InfoVis)*, 2008.

[83] S. P. Callahan, J. Freire, E. Santos, C. E. Scheidegger, C. T. Silva, and H. T. Vo, "Vistrails: visualization meets data management," in *Proceedings of the ACM SIGMOD International Conference on Management of Data, Chicago, Illinois, USA, June 27-29, 2006*, 2006, pp. 745–747. [Online]. Available: http://doi.acm.org/10.1145/1142473.1142574

[84] S. van den Elzen and J. J. van Wijk, "Small multiples, large singles: A new approach for visual data exploration," *Comput. Graph. Forum*, vol. 32, no. 3, pp. 191–200, 2013. [Online]. Available: http://dx.doi.org/10.1111/cgf.12106

[85] H. Wickham, *ggplot2 - Elegant Graphics for Data Analysis*, ser. Use R. Springer, 2009. [Online]. Available: http://dx.doi.org/10.1007/978-0-387-98141-3

[86] J. Heer and M. Bostock, "Declarative language design for interactive visualization," *IEEE Trans. Vis. Comput. Graph.*, vol. 16, no. 6, pp. 1149–1156, 2010. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2010.144

[87] A. Satyanarayan, D. Moritz, K. Wongsuphasawat, and J. Heer, "Vega-lite: A grammar of interactive graphics," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 1, pp. 341–350, 2017. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2016.2599030

[88] "Vega: A visualization grammar," Online, March 2017. [Online]. Available: http://trifacta.github.io/vega

[89] A. Satyanarayan and J. Heer, "Lyra: An interactive visualization design environment," *Comput. Graph. Forum*, vol. 33, no. 3, pp. 351–360, 2014. [Online]. Available: http://dx.doi.org/10.1111/cgf.12391

[90] D. Ren, T. Höllerer, and X. Yuan, "ivisdesigner: Expressive interactive design of information visualizations," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 2092–2101, 2014. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2014.2346291

[91] S. Mittelstädt, D. Jäckle, F. Stoffel, and D. A. Keim, "ColorCAT: Guided Design of Colormaps for Combined Analysis Tasks," in *Eurographics Conference on Visualization (EuroVis) - Short Papers*, 2015.

[92] C. Gramazio, D. H. Laidlaw, and K. B. Schloss, "Colorgorical: Creating discriminable and preferable color palettes for information visualization," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 1, pp. 521–530, 2017. [Online]. Available: https://doi.org/10.1109/TVCG.2016.2598918

[93] C. Tominski, G. Fuchs, and H. Schumann, "Task-driven color coding," in *2008 12th International Conference Information Visualisation*, July 2008, pp. 373–380.

[94] M. Behrisch, F. Korkmaz, L. Shao, and T. Schreck, "Feedback-Driven Interactive Exploration of Large Multidimensional Data Supported by

Visual Classifier," in *IEEE Conference on Visual Analytics Science and Technology*. IEEE CS Press, Oct. 2014, pp. 43–52.
[95] D. Ceneda, T. Gschwandtner, T. May, S. Miksch, H. J. Schulz, M. Streit, and C. Tominski, "Characterizing guidance in visual analytics," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 1, pp. 111–120, Jan 2017.
[96] B. Tang, S. Han, M. L. Yiu, R. Ding, and D. Zhang, "Extracting top-k insights from multi-dimensional data," in *Proceedings of the 2017 ACM International Conference on Management of Data*, ser. SIGMOD '17. New York, NY, USA: ACM, 2017, pp. 1509–1524. [Online]. Available: http://doi.acm.org/10.1145/3035918.3035922
[97] D. J. Lehmann and H. Theisel, "Optimal sets of projections of high-dimensional data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 609–618, Jan 2016.
[98] Y. Kim, K. Wongsuphasawat, J. Hullman, and J. Heer, "Graphscape: A model for automated reasoning about visualization similarity and sequencing," in *Proc. Conference on Human Factors in Computing Systems CHI*, 2017, pp. 2628–2638. [Online]. Available: http://doi.acm.org/10.1145/3025453.3025866
[99] K. Wongsuphasawat, Z. Qu, D. Moritz, R. Chang, F. Ouk, A. Anand, J. Mackinlay, B. Howe, and J. Heer, "Voyager 2: Augmenting visual analysis with partial view specifications," in *Human Factors in Computing Systems (CHI)*. ACM, 2017.
[100] L. Shao, D. Sacha, B. Neldner, M. Stein, and T. Schreck, "Visual-Interactive Search for Soccer Trajectories to Identify Interesting Game Situations," in *IS&T Elec. Imag. Conf. on Vis. and Data Analysis*, 2016.
[101] D. Sacha, M. Sedlmair, L. Zhang, J. A. Lee, D. Weiskopf, S. C. North, and D. A. Keim, "Human-Centered Machine Learning Through Interactive Visualization: Review and Open Challenges," in *Proceedings of the 24th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges, Belgium*, Apr. 2016.
[102] L. Shao, T. Schleicher, M. Behrisch, T. Schreck, I. Sipiran, and D. A. Keim, "Guiding the exploration of scatter plot data using motif-based interest measures," *J. Vis. Lang. Comput.*, vol. 36, pp. 1–12, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.jvlc.2016.07.003
[103] A. Dasgupta and R. Kosara, "Pargnostics: Screen-space metrics for parallel coordinates," *IEEE Trans. Vis. Comput. Graph.*, vol. 16, no. 6, pp. 1017–1026, 2010. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2010.184
[104] M. Chen and A. Golan, "What may visualization processes optimize?" *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 12, pp. 2619–2632, Dec 2016.
[105] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel, "KNIME: the konstanz information miner," in *Proceedings of 31st Conf. on Data Analysis, Machine Learning and Applications*, 2007, pp. 319–326. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-78246-9_38
[106] J. M. Hellerstein, P. J. Haas, and H. J. Wang, "Online aggregation," in *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, May 13-15, 1997, Tucson, Arizona, USA.*, 1997, pp. 171–182. [Online]. Available: http://doi.acm.org/10.1145/253260.253291
[107] M. Williams and T. Munzner, "Steerable, progressive multidimensional scaling," in *10th IEEE Symposium on Information Visualization (InfoVis 2004), 10-12 October 2004, Austin, TX, USA*, 2004, pp. 57–64. [Online]. Available: http://dx.doi.org/10.1109/INFVIS.2004.60
[108] J. Fekete and R. Primet, "Progressive analytics: A computation paradigm for exploratory data analysis," *CoRR*, vol. abs/1607.05162, 2016. [Online]. Available: http://arxiv.org/abs/1607.05162
[109] C. D. Stolper, A. Perer, and D. Gotz, "Progressive visual analytics: User-driven visual exploration of in-progress analytics," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 1653–1662, 2014. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2014.2346574
[110] D. Fisher, I. Popov, S. Drucker *et al.*, "Trust me, i'm partially right: incremental visualization lets analysts explore large datasets faster," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012, pp. 1673–1682.
[111] E. Zgraggen, A. Galakatos, A. Crotty, J. Fekete, and T. Kraska, "How progressive visualizations affect exploratory analysis," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 8, pp. 1977–1987, 2017. [Online]. Available: https://doi.org/10.1109/TVCG.2016.2607714
[112] Z. Liu, B. Jiang, and J. Heer, "*imMens*: Real-time visual querying of big data," *Comput. Graph. Forum*, vol. 32, no. 3, pp. 421–430, 2013. [Online]. Available: http://dx.doi.org/10.1111/cgf.12129
[113] L. D. Lins, J. T. Klosowski, and C. E. Scheidegger, "Nanocubes for real-time exploration of spatiotemporal datasets," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 12, pp. 2456–2465, 2013. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2013.179
[114] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Advances in neural information processing systems*, 2009, pp. 1753–1760.
[115] M. A. Carreira-Perpinán and R. Raziperchikolaei, "Hashing with binary autoencoders," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 557–566.
[116] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171–209, 2014.
[117] J. C. Corbett, J. Dean, M. Epstein, A. Fikes, C. Frost, J. J. Furman, S. Ghemawat, A. Gubarev, C. Heiser, P. Hochschild *et al.*, "Spanner: Googles globally distributed database," *ACM Transactions on Computer Systems (TOCS)*, vol. 31, no. 3, p. 8, 2013.
[118] "Vast challenge 2015 reviewer guide," http://hcil2.cs.umd.edu/newvarepository/VAST%20Challenge%202015/challenges/Mini-Challenge%201/solution/VAST%20Challenge%202015%20Solution%20Guide.zip, last Accessed: 22-03-2017.
[119] R. L. Sallam, C. Howson, C. J. Idoine, T. W. Oestreich, J. L. Richardson, and J. Tapadinhas, "Magic quadrant for business intelligence and analytics platforms," Gartner Inc., Tech. Rep., 2 2017.

Michael Behrisch is a Postdoctoral Researcher with the Visual Computing Group at the Harvard University Cambridge, USA. His research interest include the visualization of relational data, pattern analysis in visualizations, and user-centric exploration approaches for large data spaces. He received his PhD. in Computer Science from the University of Konstanz in 2017.

Dirk Streeb is a PhD student at the Graduate School of Decision Sciences and associated to the Data Analysis and Visualization group of Daniel Keim at the University of Konstanz, Germany, since 2016. He received his MSc. in Social and Economic Data Analysis at the University of Konstanz in the same year. His research interest focuses on analyst behavior in data analysis using Visual Analytics systems, including adaptive visualization and algorithmic analysis.

Florian Stoffel is a research associate and PhD Student at the Data Analysis and Visualization group of Prof. Daniel A. Keim at the University of Konstanz since 2013. His research interests include text and crime data analysis and visualization, as well as visual analytics and machine learning.

Daniel Seebacher is a resarch associate and PhD student at the Data Analysis and Visualization group of Daniel Keim at the University of Konstanz since 2016. His research interests include similarity search in heterogeneous high-dimensional data. He received a MSc. in Computer Science from the University of Konstanz in 2015.

Brian Matejek Brian Matejek is a PhD student with the Visual Computing Group of Hanspeter Pfister at Harvard University. He received his B.S.E. and M.S.E. from Princeton University in 2014 and 2016, respectively. His research interests include the applications of computer vision and compression techniques to neuroscience problems.

Stefan Hagen Weber is Senior Key Expert for Visual Analytics at Siemens AG, Corporate Technology and has more than 15 years experience in interactive data mining, visualization and machine learning.

Sebastian Mittelstädt received his PhD in Computer Science at the University of Kostanz in 2015 and joined afterwards Siemens AG, Corporate Technology where he is working in the field of Visual Analytics and Interactive Exploration of large Data Spaces

Hanspeter Pfister is the An Wang Professor of Computer Science at the Harvard John A. Paulson School of Engineering and Applied Sciences and an affiliate faculty member of the Center for Brain Science. His research in visual computing lies at the intersection of visualization, computer graphics, and computer vision and spans a wide range of topics, including biomedical visualization, image and video analysis, 3D fabrication, and visual analytics in data science. Pfister has a PhD in computer science from the State University of New York at Stony Brook and an MS in electrical engineering from ETH Zurich, Switzerland. Pfister was elected as chair and is currently a director of the IEEE Visualization and Graphics Technical Committee.

Daniel Keim is a Full Professor and the Head of the Information Visualization and Data Analysis Research Group inthe University of Konstanzs Computer Science Department. He received his Ph.D. and habilitation degrees in computer science from the University of Munich. He has been the Program Co-chair of the IEEE Information Visualization Conference, the IEEE Conference on Visual Analytics Science and Technology (VAST), and the ACM SIGKDD Conference on Knowledge Discovery and Data Mining.