Big Data Visualization and Analytics: Future Research Challenges and Emerging Applications

Gennady Andrienko, Natalia Andrienko, Steven M. Drucker, Jean-Daniel Fekete, Danyel Fisher, Stavros Idreos, Tim Kraska, Guoliang Li, Kwan-Liu Ma, Jock D Mackinlay, et al.

Big Data Visualization and Analytics: Future Research Challenges and Emerging Applications

Gennady Andrienko[1] Natalia Andrienko[1] Steven Drucker[2] Jean-Daniel Fekete[3] Danyel Fisher[4]
Stavros Idreos[5] Tim Kraska[6] Guoliang Li[7] Kwan-Liu Ma[8] Jock D. Mackinlay[9] Antti Oulasvirta[10]
Tobias Schreck[11] Heidrun Schmann[12] Michael Stonebraker[6]
David Auber[13] Nikos Bikakis[14] Panos K. Chrysanthis[15] George Papastefanatos[16] Mohamed A. Sharaf[17]

[1]Fraunhofer, Institute IAIS, Germany [2]Microsoft Research [3]Inria, France [4]Honeycomb.io [5]Harvard, USA
[6]MIT, USA [7]Tsinghua University, China [8]University of California-Davis, USA [9]Tableau
[10]Aalto University, Finland [11]Graz University of Technology, Austria [12]University of Rostock, Germany
[13]University Bordeaux, France [14]University of Ioannina, Greece [15]University of Pittsburgh, USA
[16]ATHENA Research Center, Greece [17]United Arab Emirates University, UAE

ABSTRACT

In the context of data visualization and analytics, this report outlines some of the challenges and emerging applications that arise in the Big Data era. In particularly, fourteen distinguished scientists from academia and industry, and diverse related communities, i.e., Information Visualization, Human-Computer Interaction, Machine Learning, Data management & Mining, and Computer Graphics have been invited to express their opinions.

Keywords

Big Data Challenges, Future Directions, Research Opportunities, Information Visualization, HCI, Machine Learning, Data Management & Mining, Computer Graphics, Visual Analytics

1. INTRODUCTION

Data visualization and analytics are nowadays one of the cornerstones of Data Science, turning the abundance of Big Data being produced through modern systems into actionable knowledge. Indeed, the Big Data era has realized the availability of voluminous datasets that are dynamic, noisy and heterogeneous in nature. Transforming a data-curious user into someone who can access and analyze that data is even more burdensome now for a great number of users with little or no support and expertise on the data processing part. Thus, the area of data visualization and analysis has gained great attention recently, calling for joint action from different research areas from the *Information Visualization*, *Human-Computer Interaction*, *Machine Learning*, *Data management* & *Mining*, and *Computer Graphics*.

Several traditional problems from those communities, such as efficient data storage, querying and indexing for enabling visual analytics, ways for visual presentation of massive data, efficient interaction and personalization techniques that can fit to different user needs, are revisited with Big Data in mind. This is to enable modern visualization systems that offer scalable techniques to efficiently handle billion objects datasets, while limiting the visual response to a few milliseconds [38][7][6][35][19][8].

In this report, in the context of the 3rd International Workshop on Big Data Visual Exploration and Analytics (BigVis)[1], the organizing committee invited *fourteen distinguished scientists*, from different communities to provide their insights regarding the challenges and the applications they find more interesting in coming years, related to *Big data visualization and analytics*[2].

Particularly, each scientist summarizes his thoughts regarding the following two aspects:

- ***the top future research challenges in Big Data visualization and analytics***
- ***the top emerging applications in the context of Big Data visualization and analytics***

We present their responses in the following sections, while the challenges are summarized as follows.

First, a challenge related to the *Machine Learning* (ML) field is enabling *interactive machine learning.* The challenge is to build *interactive tools and visualizations for machine learning* that can provide user recommendations and support user-driven ML applications. Also, it involves interactive *methods for interpretation, debugging, and comparisons of ML models.*

Another challenge is related to the *scalability* and *efficiency* of data visualization and interaction operations. In that direction, the challenges involve how to build tools that can *perform interactive operations and complex analytics over massive sets of data*. In that respect, there is the need for novel approaches (e.g., *progressive data processing*) that can efficiently handle large streaming, sampled, uncertain, high-dimensional, noisy data.

[1] *https://bigvis.imsi.athenarc.gr/bigvis2020*

[2] This report is also appearing at ACM SIGMOD Blog: *https://wp.sigmod.org*

Bringing closer data management, storage and processing infrastructure with the user visual interaction reveals an interesting opportunity *to re-define and re-evaluate an algebra in the context of visual applications*. Such an algebra can help database engineers to choose the best storage and indexing schemes, design special operators for visualizations and support optimizations for multiple operators, support collaborative visualizations by multiple users on smart devices and finally accelerate the performance of visual interaction with new hardware.

Interactive data visualization should become *more automated and understand what the users need to solve their problem*. For example, the tools should recommend what queries or views of the data the users might want to inspect, learn what sorts of data will be useful and what level of precision they need or choose design parameters based on users' needs. In general, the tools should provide enough guidance in the visual analytics process, supporting adaptive and personalized visual analysis systems and comprehensive visual mappings of data.

In essence, as important changes in the world can often be in data not seen by humans, data visualization systems *should provide sustainable insights and insights recommendations*. Thus, we need to develop tools, *which help non-data scientists to discover insights and continuously monitor the data for changes*. Tools that *automatically produce data stories and explanations* for humans that are not data scientists or analysts explaining the relevant statistics and machine learning to support the human deep-dives when the computer's advice.

Along the same direction, data science applications should become more *user-oriented*. Data analytics should become accessible to a broader range of users and stakeholders who can take advantage of the data. *Novel interfaces* and *sophisticate design of the user interactions* will assist users to understand data types, properties of the data, and their complexity in a more natural, more intuitive manner, exploiting the complementary skills of humans and computers. *Computer architectures and data systems* (which are primarily built for computational performance) *need to support the partnership with humans*, e.g., humans using their rich understanding of the world will supervise the autonomous machine-learning modules.

Following the challenge of human-computer partnership in data analytics, the combination *of human vision research and modern computational sciences* offers a new foundation to visualization research. Human factors, related to human vision and perception, could be directly integrated throughout the visualization pipeline, for supervising or providing feedback to computers.

Besides, like other areas in computational sciences, there is the challenging need for *developing benchmarks and benchmarking techniques* for assessing data visualization and visual analytic tools, in terms of performance, usability, recommendation accuracy, etc.

Finally, the support of *federated visualization* over different data sources may be extremely valuable and, in such cases, several challenges will raise, related to *addressing data privacy and ethical issues*.

The following sections present the view of each scientist.

2. Gennady & Natalia Andrienko

Visual Analytics for Data Science: A Critical View. Visual representations are often used in data analysis. In traditional data mining approaches, visualizations appear at the very end of analytical workflows, aiming at interpretation of identified patterns and their communication to various recipients, e.g., other analysts, decision makers or general public. According to the visual analytics philosophy [1][5], while human efforts must be reduced as much as possible by computational processing, visualization needs to be employed throughout the entire analytical workflow whenever an analyst is supposed to take informed decisions concerning further steps. The role of visualization is to convey the necessary information to the human in a form enabling effective perception and cognition. Hence, the task of visual analytics is to develop analytical workflows in which human cognition is effectively supported by visualizations and computational processing. Visual analytics should also be involved in development of new algorithms and software tools for automated analysis and modelling (e.g., machine learning methods) for checking whether and how well the methods are doing what they are intended to do.

Future Challenges. Our major research topic is *visually-driven analysis of spatio-temporal data*. A representative example is movement data consisting of sequences of time-referenced positions. A variety of methods and tools exist already (see theoretical foundations in the book [2]); new methods are developing actively (this is confirmed by a large number of accepted papers on this topic at IEEE VIS and other highly-selective conferences) and applied successfully in such domains as transportation [3] and sport analytics [4]. While the following thoughts have arisen from our experiences with spatio-temporal data, they are applicable to other types of complex data.

Emerging Applications. Several key recent developments create great opportunities for empowering data science by visual analytics. The first one is the appearance and wide spread of data science-oriented languages such as Python and R. These languages enable step-by-step data analysis and support integration of visualizations in analytical workflows. Analytical notebooks based on these languages help to document analysis, enable reproducibility and help to share results. These technologies are so easy to use that nowadays everyone can become an analyst. There exist textbooks explaining the basics on simple examples, and many pieces of code are available online for use and adaptation. This phenomenon has its back side: self-made analysts often lack fundamental knowledge of the overall analysis process, and miss understanding of why, when, and how visualizations need to be used in analysis. The Internet is overfilled by visualizations, often looking very impressive and fancy, that communicate spurious patterns in inadequate ways. Unfortunately, most of the available code examples and the majority of text books don't go beyond applying basic graphics to simple data and do not demonstrate the analytical value of the graphics. *Therefore, we see the major challenge in educating data scientists on how to use visualizations correctly and effectively within non-trivial analytical workflows, understanding and taking into account the data types, properties of the data, and their complexity.*

3. Steven Drucker

Future Challenges & Emerging Applications.

Systems for Machine Learning (ML) and ML for Systems. While perhaps overhyped, the huge amount of attention directed towards Machine Learning is apparent throughout research and industry. Papers are appearing covering numerous aspects, from fundamental theoretical advances like causal reasoning and general intelligence to applications of machine learning in systems and knowledge work. This explosion in machine learning is enabled primarily by the vast amounts of data available and systems that allow the training for models using those large data sets. This in turn enables clever applications of these techniques above and beyond the basic frontiers of ML (classification, clustering, and regression). Given this explosion, we need far better techniques for working with the data and the models for ML. This includes helping troubleshoot models, understanding where models work and don't work, comparing models with each other, and giving understandable explanations for model behavior. Since visualization is fundamentally about helping humans interpret and interact with data, *interactive tools and visualizations for machine learning is a one of the top challenges for visualization in the coming decade*. *At the same time, using the output of models to help build better visualizations (whether it's for recommending a single or sequence of visualizations) or to interact at a higher level with data by helping find optimal ways of leveraging human intuition and knowledge while exploiting more powerful computation is a key component of new applications*.

To concretize these areas, here are some of the recent research papers that presage some of the emerging applications in this area.

- ***Methods for interpretation of ML models, both specific types of models (such as Additive Models) which both perform well and are somewhat interpretable, and more general techniques for interacting with arbitrary models.*** Closely related to the above, as a requirement of the European laws for General Data Protection Regulation (GDPR), any decision made algorithmically must be explainable and whenever data needs to be explained, visualization is an important component. Recent work on this includes the research of Hohman et al. [21] from Rich Caruana's GAM models [28], Wattenberg & Viegas [53], and others on creating more interpretable models or LIME [35] and Lundberg [30].

- ***Systems for troubleshooting and debugging models and the spaces where they are effective as well as comparing models with each other.*** Recent work includes the research of Saleema Amershi in Model Tracker [1] and Besmira Nushi on Error Terrain Analysis[3].

- ***Generating recommendations for visualizations based on models of user behavior.*** Work such as Moritz's Draco system [33] and Kim's GraphScape system [24].

- ***Interpreting visualization for subsequent reuse.*** Work by Poco & Heer [35] and Agrawala et al. [40] [4].

- ***Creating higher level interactions with data through NLP and other modalities*** such as the VODER work of Srinivasan et al. [44].

[3] *https://slideslive.com/38915701/error-terrain-analysis-for-machine-learning-tool-and-visualizations*

4. Jean-Daniel Fekete

Future Challenges. To be effective, visualization and visual analytics should be interactive, meaning that computing visual representations should happen in a few seconds, interacting on them should be responsive, and analytics should also be done in accordance with the acceptable limits of human latency as described in the literature [24]. *Building systems that remain interactive at scale and using complex analytics is a major challenge for the visualization field, which may become irrelevant if it does not address the scalability challenge properly*.

Emerging Applications. To address this scalability challenge, my new focus of research is "*Progressive Data Analysis and Visualization*", *a new paradigm of computation that, instead of performing computations in one step that can take an arbitrarily long time to complete, splits them in a series of short chunks of approximate computations that improve with time*. Therefore, instead of waiting for an unbounded amount of time the results of computations for visualization and analytics, analysts can see the results unfolding progressively. They can, therefore, maintain their attention and start making some decisions earlier than if they had to wait for the whole computations to finish.

Meanwhile, while the results are being computed, analysts can also interact with the ongoing computation, changing computation parameters and sometimes steering the algorithms.

This new paradigm is just starting to emerge and will require more time to become mainstream, as explained in the report we published after a Dagstuhl seminar conducted in 2018 [15]. However, *I am confident that Progressive Data Analysis will allow visualization and analytics to become more scalable while remaining interactive to facilitate the exploration of the wealth of data that the world is gathering, coupled with new powerful methods to analyze it developed in machine learning in particular* [43].

This new paradigm is not only important for visualization and visual analytics but also requires strong collaborations with researchers in Databases and Machine Learning who recognize that progressive data analysis will lead to more scalable exploratory systems.

5. Danyel Fisher

Future Challenges & Emerging Applications. We are learning to ask new things of our data. It's increasingly practical to interactively explore Big Data, asking novel questions to discover unexpected phenomena. The lines between different forms of analysis -from relational queries, to unstructured data, to rich media- are blurring. *I'm looking forward to our improving all steps of the process: to learning how to best express questions; how to get interactive-speed responses to those questions; and to iterate on those insights to ask the next round of questions*.

These steps are interconnected and interdependent. To get interactive responses, for example, we might use *progressive computation*; e.g., [17][16]. That technique requires us to think about *communicating uncertainty* (e.g., [53][20]) and giving the analyst a way to record how much that uncertainty affects their analysis process.

One way to help all these stages is to focus on particular problem domains. I've spent my last few years working on analysis tools for *sampled, uncertain, high-dimensional, streaming data*. As a

[4] *http://graphics.stanford.edu/projects/dataExtract*

domain as a whole, that's huge and intimidating. Fortunately, I can target my work on Honeycomb[5] toward their real problems, and so can take advantage of the constraints of their particular context. Honeycomb is an APM (Application Performance Monitoring) tool for debugging distributed systems. Our users are DevOps, who are responsible for deploying new code -- and recovering when it fails. Tools like BubbleUp[6], a histogram comparison tool, help users isolate specific classes of failures rapidly. Our underlying data structure is similar to Facebook's SCUBA [28].

I believe that focusing relentlessly on specific use cases will make otherwise broad questions simpler. If we can really understand what the user needs to solve their problem, we can learn what sorts of data they will ingest, what queries they might want to ask, what performance characteristics they expect, and what level of precision they need.

6. Stratos Idreos

Future Challenges & Emerging Applications. Visualizing data is one of the best ways to find patterns and information in Big Data. The reason why data visualization is interesting for the data management community is that this is an inherently data-intensive problem. In addition, data scientists may pose arbitrary queries as they create new visualizations or interact with existing ones. This means that such systems get: (1) diverse queries, (2) sequences of queries where each query may depend on the previous one, (3) queries that may be OK to abort, and (4) workloads which need rapid response times to remain interactive even if correctness is not immediately at 100%. Due to this mismatch with typical database applications, there are several long-term and exciting challenges that represent wonderful opportunities for data management researchers given the rich history of the field in data-intensive algorithms and systems. I highlight two of those opportunities as they arise from recent work in the area.

- ***First, what is the equivalent of the relational algebra for visual analytics*?** It might seem daunting to condense the vast space of possible actions a data scientist may perform into a small set of operations, but this is exactly what the original relational algebra achieved. And then, more complex operations can be synthesized from primitive algebra operations. *On top of that algebra, we can then build systems that rely on a common interface and a small set of operators, allowing the community to collectively attack this problem by considering alternative designs and implementations that respect the same model and API as it happened with relational operators*. This abstraction is one of the secrets both for the adoption of the relational model across diverse applications and for the ability to relatively easily experiment with alternative implementations.
- ***Second, what is the equivalent of the b-tree, and the row-oriented and column-oriented storage schemes*?** While these are by no means the only indexing and storage options, knowing the extreme designs or some of the most versatile designs, and then heavily focusing on them, allowed relational databases to mature in terms of both speed and robustness. Data visualization and visual analytics is (typically) a data-intensive as opposed to a compute-intensive problem. This means that the way we store and move data is the bottleneck. *Supporting alternative storage schemes and choosing the right one for the right queries is key*. Thus, studying the design of data structures that can absorb the new and diverse access patterns as well as the interactive response times required by visual analytics is a massive opportunity for the data management community.

7. Tim Kraska

Future Challenges.

- ***Interactive data exploration for large data and more complex operations*.** Tools like Tableau, PowerBi and co praise themselves as interactive data exploration tools. Yet, for larger datasets they rely on pre-computed data cubes, materialized views, and similar techniques to stay interactive. Unfortunately, these techniques severely restrict what the user is able to do. For example, to create a data cube one needs to know upfront, what type of questions the data cube is supposed to answer. This essentially prevents interactive responses for completely new questions. *A key research challenge is, how we can build systems which guarantee interactive response times regardless of the question and data size*. As part of Northstar, we started to explore how we can leverage progressive computation and sampling to achieve this.
- ***Sustainable insights and insight recommendation.*** *We need to make finding insights easier*. *Thus, we need to develop tools which help non-Data Scientists to discover insights and continuously monitor the data for changes*. For example, a system should automatically recommend interesting insights and visualization about things that might interest the user. SeeDB or VizML are systems, which started to explore that. *At the same time, those insights should also be sustainable*. For example, current insight recommendation systems largely ignore the risk of finding spurious insights by testing too many hypotheses.
- ***Novel interfaces*.** *We should make data analytics more accessible to a broader range of users*. This requires to fundamentally rethink the user interface. *We put everything on the table HCI has to offer; from novel visualizations, interaction patterns, touch screens, up to natural language interfaces*. Interestingly, changing the user interface often also has severe implications on how the backend has to be developed. We (the SIGMOD community) tend to first develop the system and then add the user interface as an afterthought, often leading to clunky old-style interaction. I think it should be the other way around. *Design the user interactions first and then figure out the system, which can actually support them*.

Emerging Applications. I believe, there exists not a single area which is not already impacted by analytics. Thus, it is close to impossible to find a new emerging application for analytics. However, *I am a strong believer that we need to broaden the scope of users, who are able to take advantage of the data*. Current tools are mainly designed for experts or significantly restrict what a user can do. For example, there is not a single tool out there, which makes it easy for a coffee shop owner to analyze his customer base and make predictions about future sales. At the same time, those people could also tremendously benefit from their data. This requires to rethink the way users interact with data.

[5] *http://honeycomb.io*

[6] *https://docs.honeycomb.io/working-with-your-data/bubbleup*

8. Guoliang Li

Future Challenges.

- ***Automatic visualization*.** Existing data visualization methods require uses to write visualization queries. However, in many cases, users cannot precisely write queries, because it is hard to understand all the underlying data, e.g., in data lake. Thus, *it is important to automatically visualize the data, including visualization-aware data discovery, data integration, and data cleaning*.
- ***Federate visualization*.** Most of data visualization systems do not consider data privacy. *It is challenging to support privacy-preserving visualization. Furthermore, it is more promising to support federate visualizations on data across different sources.* Considering two companies A and B, A requires to use the data of both A and B to do visualizations, but B cannot release the data to A and can only release some privacy-preserved data.
- ***Visualization benchmark*.** Like ImageNet or the classic TPC benchmarks, *it is important to develop benchmarks for performance and recommendation. The benchmarks should be faithful to the visual analysis tasks, provide reusable traces and data, and in the case of recommendation, have high coverage and quality of its labels*.
- ***Visualization databases*.** Databases are widely used and deployed in many real applications. It calls for a visualization database that (1) *designs special operators for visualizations and supports optimizations for multiple operators*; (2) *efficiently visualizes a large-scale data*; (3) *supports interactive visualizations with users*; (4) *supports collaborative visualizations by multiple users on smart devices*; (5) *accelerates the performance with new hardware*.

Emerging Applications.

- ***Visualization on clouds*.** Most of existing visualization systems are on premise or on smart devices. However, they cannot handle Big Data and thus it is promising to utilize client-cloud collaboration techniques to support data visualizations, where the data and visualization results are automatically transmitted between clients and clouds. *It requires to address the challenges of data consistency, data security and data sampling*.
- ***Visualize the time-series data for 5G and IOT*.** The 5G age is coming and there are generating more and more high-*volume and high-speed time-series data*. *It is promising to visualize the time-series data and help users find insights from them instantly*.
- ***Visualization for debugging*.** *It is promising to utilize visualizations for debugging, including both bug debugging and data debugging*. The former utilizes visualizations to find bugs in a workflow, e.g., finding the root causes why a SQL query is slow. The latter aims to find the reasons why data visualization result is wrong, e.g., wrong data sources, wrong parameters, wrong visualization queries. *There are various applications in healthcare and education*.

9. Kwan-Liu Ma

Future Challenges. Many Big Data problems find machine learning a promising solution. *One existing challenge is how to effectively assist machine learning with visualization, which includes both interpreting and optimizing machine learning* [11][18]. The inverse problem probably interests visualization researchers more. It has been shown machine learning can assist visualization design and generation [27][40][26]. *Usability of visualization must be validated and enhanced* before it can make a true impact to Big Data applications. A promising approach to enhancing the usability of visualization is exactly to add intelligence to the overall process of selecting data, feature extraction, visual encoding, layout generation, annotation, rendering, and response to interaction, allowing the user to focus on perceiving and interpreting information. This marriage makes computer and human each does its best. Finally, e*thic including privacy* is another area of increasing importance that must be addressed when designing visualization solutions for exploring Big Data [12][48][49].

Emerging Applications. When involving Big Data, there are tremendous opportunities to employ visualization and analytics in a wide range of applications from *health care, manufacturing, IoT, cybersecurity, to learning and design*. Progress in each area is being made but the deployment, adoption, and thus benefits of visualization remain to be seen. *One area of my personal interest is the analysis of agricultural and environmental data collected with all means of monitoring*. The high dimensional and heterogeneous nature of the collected data presents tremendous opportunities for visualization research and innovations, from designing data aggregation and visual composition strategies, resolving visual scalability for adapting to different display and interaction types, enhancing computing scalability to meet the required level of interactivity, and addressing uncertainty due to data generation as well as data and visual transformations [14][51][36][47].

10. Jock D. Mackinlay

Emerging Applications. *One of the biggest opportunities for Big Data innovation is a more natural, more intuitive integration of the complementary skills of humans and computers*.

Humans have incredibly rich understandings of the world that empowers their work with data that describes the world. Computers have superficial understandings of the world but can process Big Data 24×7×365 with sophisticated computations based on statistics and machine learning.

Imagine the potential impact of a better partnership between humans and computers, one in which computers monitor the data coming from the world and advise humans where to visually explore and analyze data.

User interfaces that support effective visual analysis will allow humans to quickly determine when computers have generated false positives. Humans will also be able to dive in deeply, especially with their colleagues, to explore something that deserves human attention and decision-making. Human deep dives can lead to significant value to organizations. Additionally, false positives can be reduced over time by using telemetry from human visual analysis to update the machine learning that decides when to advise humans.

Future Challenges. This vision of a *monitoring partnership* involves a broad range of interesting research challenges, including the following:

- ***Scale*** is a traditional research challenge for the database community, including the processing of federated queries that combine long and/or wide data tables. *Today, the scale of monitoring is bottlenecked by data scientists and analysts who are using their rich understanding of the world to tell computers where to monitor. However, important changes in the world can often be in data not seen by humans*. Imagine scaling our computer monitoring to ALL the data *streaming* from the world using machine-learning algorithms trained on human data work, including the directives from data scientists and analysts, to identify important changes in the world.
- ***Speed*** to process streaming data is important in scenarios that lead to human analysis like intrusion detection and when machine-learning modules are autonomously deployed. *Computer architectures built for speed also need to support the partnership with humans, including humans using their rich understanding of the world to supervise the autonomous machine-learning modules*.
- ***Automatic data stories*** are key to fostering the monitoring partnership. Computers need to explain why they are advising humans to engage in visual analysis. *A key research challenge is explaining the relevant statistics and machine learning to support the human deep-dives when the computer's advice is not a false positive, particularly explanations for humans that are not data scientists or analysts*.

This small list is intended to suggest the range of interesting research challenges associated with the monitoring partnership vision. Toward the human end of the research range, my area of expertise, it is important to embrace the richness of the world. Even unicorn data scientists typically need to collaborate with other people in their organizations to be effective. Each organization (and often parts of organizations) have unique data they need to monitor. Most people are engaged in the mission of their organization, which means they have latent interest in the data describing their organizations. Therefore, *almost everyone in an organization will benefit from partnering with computers monitoring changes in the world*.

11. Antti Oulasvirta

Future Challenges. The greatest challenge is, still, the human. You cannot do a visualization without understanding human perception. But when it comes to that, visualization as a field relies excessively on trial and error and empirical testing. The ideal should be a field that has human factors integrated directly throughout the visualization pipeline, akin to mature engineering disciplines where theories and models (e.g., from physics) are used to derive optimal solutions. In the case of visualization, that necessarily means models of human perception, attention, and cognition -instead of physics- as these determine success/failure across visual analytics tasks. What does this mean in practice? One recent example is our work with Luana Micallef, who sadly passed away recently, and Tino Weinkauf, on perceptual optimization of scatterplots [30]. We used models from human vision research as objective functions and generated visualizations computationally, such that the underlying structure of the data could be better perceived by human observers [30]. This becomes more and more important as the complexity of the dataset grows. But we can go further than generative design! We can better ***explain*** data. Using methods like Bayesian optimization, we can now much better fit models of human behavior directly to observational data, such as clickstream data, in order to make better sense of them [21]. We can go way beyond variables and actually start explaining the processes that generated the observed data! We can also use models from psychology to ***adapt*** visualizations to individuals and their tasks and preferences. Models can be used to compute visualizations that are strike optimal trade-offs among tasks and user groups. I firmly believe that the combination of human vision research and modern computational sciences offers a new foundation to visualization research that is less reliant on heuristics, manual tuning, and empirical testing, and takes a step closer to mature engineering disciplines.

Emerging Applications. I don't think there is a need to open new applications for Big Data visualization and analytics, rather to solve the core problem really well, and in a principled manner. ***Why*** *should a particular visualization favored over another one in some context*? ***Why*** *should one choose particular design parameters over other ones*? *What are the* ***limits*** *of a particular type of visualization, what can it do and -more importantly- what can it* ***not*** *do*? Core problems like these need principled answers, and the answers will not come from computer science only but will need to seriously engage with vision science. *The main application of Big Data visualization and analytics will be just doing it 10× better than now: better efficacy and higher efficiency*.

12. Tobias Schreck

Future Challenges. Visual analytics approaches integrate interactive data visualization with automatic data analysis methods, supporting expert-in-the-loop exploration of large data sets for potentially interesting and actionable endings. *As data sets grow larger, and become more heterogeneous, there is increased risk that users may have difficulties to effectively navigate the data and take longer to arrive at in-sights.* For example, as dimensionality and size of data sets grow, there is an exploding amount of possible data subsets to select and visualize, in which users may get lost navigating.

Approaches to user guidance, adaption and personalization in the visual data exploration process may help this situation. Recent works [9][10] have identified requirements and desiderata for guidance-based visual analytics systems. For example, different types of guidance for orienting, directing or even prescribing visual exploration paths can be distinguished. *While the need and potential of guidance in the visual analytics process is eminent, how to design, integrate and evaluate actual guidance approaches in the visual analytics process, supporting specific tasks of specific users on specific data, is less clear to date*. There exists a large and exciting theoretical and applied research space for guidance approaches in many areas, and we can build on a number of promising starting points in this direction. For example, mining user interaction patterns can help to classify and predict user interest and search strategy, and help to guide the exploration. For example, recent research has mined user interaction input during search tasks [9], or proposed to derive user interest models from eye-tracking the user [43]. Principles from information retrieval like relevance feed-back and recommending [34] can be adapted for the visual exploration process, promising to arrive more efficiently at relevant findings. *Besides mining user interaction, also approaches for estimation the visual quality and information content in visualizations have been proposed, for*

different types of visualizations [25], *and can be used to rank and suggest candidate views for inspection by users. As a third line of research, approaches for description of design rules and learning-based approaches may allow to automatize the otherwise interactive and open-ended search for appropriate visual mappings of data, and support visualization automation* [33][40].

To summarize, *I see promising research directions in implementing guidance for the visual analytics process, supporting more effective and efficient, adaptive and personalized visual analysis systems*.

Emerging Applications. There are certainly very many promising application areas for Big Data visualization and analytics. Progress has been made already in a number of domains including biomedical applications, financial data visualization, social media and text visualization, just to name a few. *We recently observe strongly growing interest in visual analysis of data in industrial contexts, driven by efforts to collect data from industrial production processes by sensor equipment becoming available, and introduction of data-driven approaches*. In [50], the authors give a survey of relevant use cases and first solutions for industrial data visualization, including internal and external equipment environment visualization, and for purpose of creation, including design, production, testing, and service.

Visual analysis of data in industrial contexts presents challenging problems, due to large amounts of data produced by increasing numbers of sensors applied to the whole production pipeline, and delivering data at high frequency. The data is often of heterogeneous nature, comprising different units of measurement which interact with each other. The temporal alignment (normalization) of data along possibly long-running production phases poses a challenge with respect to data accuracy and alignment. The filtering and selection of relevant data is an-other challenge, which often needs to be found by trial-and-error in lack of best practices and experience. There exist bodies of data analysis and visualization methods applicable in principle for industrial data, including real-time monitoring, prediction, identification of influence factors (correlations), and anomaly detection. Each of these methods can be very useful to solve problems of resource planning, quality control, or equipment condition monitoring. *How to adequately select, filter, integrate and transform data, and which analysis and visualization techniques to combine, poses an interesting and rewarding application challenge. In addition, a research problem we are currently pursuing involves the unification of domain process knowledge with process data patterns in an integrated production information system, supporting planning and monitoring* [44]. While first concepts exist, much needs to be done in the future.

13. Heidrun Schumann

Future Challenges. Big Data visualization and analytics require a combination of visual, interactive and automatic analysis methods. However, each of these aspects covers quite a lot challenging topics, which are communicated in separate journals, and discussed at separate conferences and workshops. My question is: *How can this diversity be used in order to provide effective and efficient analysis tools that offer this wide range of functions*?

For specific applications, methods from different fields can certainly be inspected and selected on demand. Then, tailored analysis systems can be developed. But, given the large number of applications: *Will we manage to develop independent analysis systems for each application*?

I think, we should also pursue a more generic approach. However, if we take a generic approach, then the question arises: *How to combine different methods from different fields in order to create a unified application*? I don't believe that it makes sense (or is even possible) to integrate all required or desired functionalities into one and the same easy-to-use analysis tool. *So how can different tools and methods be combined to visualize and analyze large data in a uniform way*?

My key message is therefore: *Not only the large volume of Big Data is a challenge, but also the large volume of methods and tools needed to process it.*

Emerging Applications. I identified *climate change and health care as emerging application examples because of their high social relevance*. Although visual analysis methods are already used in both applications, various challenges require the development of new strategies and solutions. These challenges are for example:

- ***Related to the data.*** The data are heterogeneous and complex, come from different sources, are calculated or measured, are subject to uncertainties, and particularly, they are available on different scales.
- ***Related to the users.*** Scientists from several domains need to discuss and analyze the data. For example, in the case of studying the impact of climate change, scientists from ecology, physics, biology, geology, and mathematics compare their data and projections based on different models and parameterizations to predict the impact of climate change on different aspects.

14. Michael Stonebraker

Future Challenges & Emerging Applications. One of the consequences of having "big data" is that there is a lot of it. The issue of scale breaks many of our cherished assumptions. First the data does not fit in main memory, and any platform that assumes main memory will fail. Also, the data is inevitably skewed – i.e., 90% of the data is in 10% of the available real estate. Hence, the issue of clutter is omnipresent. At scale, it is very costly to prebuild visualization structures, and most data sets are multi-user, leading to different access patterns and use cases. Multi-user data sets are often updated, and assuming the data is static is dangerous. Lastly, terascale and petascale data will always come out of a backing database system, and our community should always deal with server-side DBMSs. As a result, *I think the biggest challenge our community needs to solve is scalability*.

The days of rendering teapots are long gone. Data scientists in industry have to support user communities of business analysts, physical scientists and decision makers. Their objective is to gain actionable insights from very large data sets. If the domain is simple, they will go straight to analytics. In complex domains (where it is not clear what question to even ask), scalable visualization will be the "go to" technology. *Hence, supporting data scientists is likely to be the "sweet spot" for Big Data visualization technology*.

15. Conclusions

In this report, we presented a list of major challenges, which have been provided by fourteen distinguished scientists who took part in a "virtual" panel as part of the BigVis 2020 Workshop. The report aimed at providing insights, new directions and opportunities for research in the field of Big Data visualization and analytics.

REFERENCES

[1] S. Amershi, M. Chickering, et al.: ModelTracker: Redesigning Performance Analysis Tools for Machine Learning. CHI 2015

[2] G. Andrienko, N. Andrienko, P. Bak, D. Keim, S. Wrobel: Visual Analytics of Movement. Springer 2013

[3] G. Andrienko, N. Andrienko, et al.: Visual Analytics of Mobility and Transportation: State of the Art and Further Research Directions. TITS 18(11), 2017

[4] G. Andrienko, N. Andrienko, et al.: Constructing Spaces and Times for Tactical Analysis in Football, TVCG, 2019

[5] N. Andrienko, T. Lammarsch, G. Andrienko, et al.: Viewing Visual Analytics as Model Building. CGF 2018

[6] M. Behrisch, D. Streeb, F. Stoffel, D. Seebacher, et al.: Commercial Visual Analytics Systems-advances in the Big Data Analytics Field, TVCG 25(10), 2019

[7] N. Bikakis: Big Data Visualization Tools Survey, Encyclopedia of Big Data Technologies, Springer 2019

[8] N. Bikakis, T. Sellis: Exploration and Visualization in the Web of Big Linked Data: A Survey of the State of the Art, LWDM Workshop 2016

[9] E.T. Brown, A. Ottley, et al.: Finding Waldo: Learning About Users from their Interaction TVCG 20(12), 2014

[10] D. Ceneda, T. Gschwandtner, et al.: Characterizing Guidance in Visual Analytics. TVCG 23(1), 2017

[11] J. Choo, S. Liu: Visual Analytics for Explainable Deep Learning. IEEE CGA 38(4), 2018

[12] J.-K. Chou, Y. Wang, K.-L. Ma: Privacy Preserving Visualization: A Study on Event Sequence Data. Comput. Graph. Forum 38(1), 2019

[13] C. Collins, N. Andrienko, et al.: Guidance in the Human Machine Analytics Process. Visual Informatics 2(3), 2018

[14] C. D. Correa, Y.-H. Chan, K.-L. Ma: A framework for uncertainty-aware visual analytics. VAST 2009

[15] J.D. Fekete, D. Fisher, A. Nandi, M. Sedlmair: Progressive Data Analysis and Visualization. Dagstuhl Seminar, 2018

[16] J.D. Fekete, R. Primet. Progressive Analytics: A Computation Paradigm for Exploratory Data Analysis. CoRR 2016

[17] D. Fisher, et al.: Trust me, I'm Partially Right: Incremental Visualization lets Analysts Explore Large Datasets Faster CHI 2012

[18] T. Fujiwara, O.-H. Kwon, K.-L. Ma: Supporting Analysis of Dimensionality Reduction Results with Contrastive Learning. TVCG 26(1), 2020

[19] P. Godfrey, J. Gryz, P. Lasek: Interactive Visualization of Large Data Sets. TKDE 28(8), 2016

[20] J. Hullman: Why Authors Don't Visualize Uncertainty. TVCG 26(1), 2019

[21] F. Hohman, A. Head, R. Caruana, R. DeLine, S. Drucker: Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. CHI 2019.

[22] A. Kangasrääsiö, et al.: Parameter Inference for Computational Cognitive Models with Approximate Bayesian Computation. Cognitive Science, 2019

[23] D. Keim, J. Kohlhammer (eds.): Mastering the Information Age: Solving Problems with Visual Analytics. Eurographics 2010

[24] Y. Kim, et al.: GraphScape: A Model for Automated Reasoning about Visualization Similarity and Sequencing. CHI 2017

[25] N.W. Kim, L. Shao, M. El-Assady, et al.: Quality Metrics for Information Visualization. CGF 37(3), 2018

[26] O.-H. Kwon, T. Crnovrsanin, K.-L. Ma: What Would a Graph Look Like in this Layout? A Machine Learning Approach to Large Graph Visualization. TVCG 24(1), 2018

[27] O.-H. Kwon, K.-L. Ma: A Deep Generative Model for Graph Layout. TVCG 26(1), 2020

[28] A. Lior, J. Allen, O. Barykin, V. Borkar, B.Chopra, et al.: SCUBA: diving into Data at Facebook. PVLDB 6(11), 2013

[29] Y. Lou, R. Caruana, J. Gehrke, G. Hooker: Accurate Intelligible Models with Pairwise Interactions. KDD 2013

[30] S. Lundberg, S. Lee. A Unified Approach to Interpreting Model Predictions. NIPS 2017

[31] L. Micallef, et al.: Towards Perceptual Optimization of the Visual Design of Scatterplots. TVCG 23(6), 2017

[32] L. Micallef, G. Palmas, et al.: Towards Perceptual Optimization of the Visual Design of Scatterplots. TVCG 23(6), 2017

[33] D. Moritz, et al.: Formalizing Visualization Design Knowledge as Constraints: Actionable and Extensible Models in Draco. TVCG 25(1), 2019

[34] B. Mutlu, E.E. Veas, C. Trattner VizRec: Recommending Personalized Visualizations. TiiS 6(4), 2016

[35] L. Po, N. Bikakis, F. Desimoni, G. Papastefanatos: Linked Data Visualization: Techniques, Tools and Big Data. Morgan & Claypool, 2020

[36] A. Preston, M. Gomov, K.-L. Ma: Uncertainty-Aware Visualization for Analyzing Heterogeneous Wildfire Detections. IEEE CGA 39(5),2019

[37] J. Poco, J. Heer: Reverse-Engineering Visualizations: Recovering Visual Encodings from Chart Images. CGF 36(3), 2017

[38] X. Qin, Y. Luo, N. Tang, G. Li: Making Data Visualization more Efficient and Effective: A survey. VLDBJ 2020

[39] M.T. Ribeiro, S. Singh, C. Guestrin: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. KDD 2016

[40] B. Saket, D. Moritz, H. Lin, V. Dibia, C. Demiralp, J. Heer: Beyond Heuristics: Learning Visualization Design. CoRR 2018

[41] M. Savva, N. Kong, A. Chhajta, L. Fei-Fei, M. Agrawala, J. Heer: ReVision: Automated Classification, Analysis and Redesign of Chart Images. UIST 2011

[42] B. Shneiderman: Response Time and Display Rate in Human Performance with Computers. ACM Comput. Surv. 16(3), 1984

[43] N. Silva, et al.: Eye Tracking Support for Visual Analytics Systems: Foundations, Current Applications and Research Challenges. ETRA 2019

[44] A. Srinivasan, S.M. Drucker, A. Endert, J. Stasko: VODER: Augmenting Visualizations with Interactive Data Facts to Facilitate Interpretation and Communication. TVCG 25(1), 2019

[45] S. Thalmann., J, Mangler, et al.: Data Analytics for Industrial Process Improvement. CBI 2018

[46] C. Turkay, N. Pezzotti, et al.: Progressive Data Science: Potential and Challenges. CoRR 2019

[47] Y. Wang, K.-L. Ma: Revealing the fog-of-war: A visualization-directed, uncertainty-aware approach for exploring high-dimensional data. IEEE BigData 2015

[48] X.-M. Wang, W. Chen, J.-K. Chou, C. Bryan, H. Guan, W. Chen, R. Pan, K.-L. Ma: GraphProtector: A Visual Interface for Employing and Assessing Multiple Privacy Preserving Graph Algorithms. TVCG 25(1), 2019

[49] X.-M. Wang, J.-K. Chou, W. Chen, H. Guan, W. Chen, T. Lao, K.-L. Ma: A Utility-Aware Visual Approach for Anonymizing Multi-Attribute Tabular Data. TVCG 24(1), 2018

[50] M. Wattenberg, F. Viegas: Visualization: The Secret Weapon of Machine Learning. EuroVis 2017 (Keynote)

[51] Y. Wu, G.-X. Yuan, K.-L. Ma: Visualizing Flow of Uncertainty through Analytical Processes. TVCG 18(12), 2012

[52] F. Zhou., X. Lin, et al.: A Survey of Visualization for Smart Manufacturing. Journal of Visualization 22 (2), 2019

[53] T. Zuk, S. Carpendale: Theoretical Analysis of Uncertainty Visualizations. Visualization and Data Analysis, 2006