

Bridging Information Visualization with Machine Learning

Edited by

Daniel A. Keim¹, Tamara Munzner², Fabrice Rossi³, and Michel Verleysen⁴

1 Universität Konstanz, DE, daniel.keim@uni-konstanz.de

2 University of British Columbia – Vancouver, CA, tmm@cs.ubc.ca

3 Université Paris I, FR, Fabrice.Rossi@univ-paris1.fr

4 Université Catholique de Louvain, BE, michel.verleysen@uclouvain.be

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 15101 “Bridging Information Visualization with Machine Learning”. This seminar is a successor to Dagstuhl seminar 12081 “Information Visualization, Visual Data Mining and Machine Learning” held in 2012. The main goal of this second seminar was to identify important challenges to overcome in order to build systems that integrate machine learning and information visualization.

Seminar March 1–6, 2015 – <http://www.dagstuhl.de/15101>

1998 ACM Subject Classification H.5 Information Interfaces and Presentations, I.3 Computer Graphics, I.5.4 Computer Vision, H.2.8 Database Applications, H.3.3 Information Search and Retrieval, I.2.6 Learning

Keywords and phrases Information visualization, Machine learning, Visual data mining, Exploratory data analysis

Digital Object Identifier 10.4230/DagRep.5.3.1


1 Executive Summary

Daniel A. Keim

Tamara Munzner

Fabrice Rossi

Michel Verleysen

License  Creative Commons BY 3.0 Unported license
© Daniel A. Keim, Tamara Munzner, Fabrice Rossi, and Michel Verleysen

Motivations and context of the seminar

Following the success of Dagstuhl seminar 12081 “Information Visualization, Visual Data Mining and Machine Learning” [1, 2], which provided to the participants from the IV and ML communities the ground for understanding each other, this Dagstuhl seminar aimed at bringing once again the visualization and machine learning communities together.

Information visualization and visual data mining leverage the human visual system to provide insight and understanding of unorganized data. Visualizing data in a way that is appropriate for the user’s needs proves essential in a number of situations: getting insights about data before a further more quantitative analysis (e.g., for expert selection of a number of clusters in a data set), presenting data to a user through well-chosen table, graph or other structured representations, relying on the cognitive skills of humans to show them extended information in a compact way, etc.



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 3.0 Unported license

Bridging Information Visualization with Machine Learning, *Dagstuhl Reports*, Vol. 5, Issue 3, pp. 1–27

Editors: Daniel A. Keim, Tamara Munzner, Fabrice Rossi, and Michel Verleysen



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

The scalability of visualization methods is an issue: human vision is intrinsically limited to between two and three dimensions, and the human preattentive system cannot handle more than a few combined features. In addition the computational burden of many visualization methods is too large for real time interactive use with large datasets. In order to address these scalability issues and to enable visual data mining of massive sets of high dimensional data (or so-called “big data”), simplification methods are needed, so as to select and/or summarize important dimensions and/or objects.

Traditionally, two scientific communities developed tools to address these problems: the machine learning (ML) and information visualization (IV) communities. On the one hand, ML provides a collection of automated data summarizing/compression solutions. Clustering algorithms summarize a set of objects with a smaller set of prototypes, while projection algorithms reduce the dimensionality of objects described by high-dimensional vectors. On the other hand, the IV community has developed user-centric and interactive methods to handle the human vision scalability issue.

Building upon seminar 12081, the present seminar aimed at understanding key challenges such as interactivity, quality assessment, platforms and software, and others.

Organization

The seminar was organized in order to maximize discussion time and in a way that avoided a conference like program with classical scheduled talks. After some lightning introduction by each participant, the seminar began with two tutorial talks one about machine learning (focused on visualization related topics) followed by another one about information visualization. Indeed, while some attendants of the present seminar participated to seminar 12081, most of the participants did not. The tutorials helped establishing some common vocabulary and giving an idea of ongoing research in ML and IV.

After those talks, the seminar was organized in parallel working groups with periodic plenary meeting and discussions, as described below.

Topics and groups

After the two tutorials, the participants spend some time identifying topics they would like to discuss during the seminar. Twenty one emerged:

1. Definition and analysis of quantitative evaluation measures for dimensionality reduction (DR) methods (and for other methods);
2. In the context of dimensionality reduction: visualization of quality measures and of the sensitivity of some results to user inputs;
3. What IV tasks (in addition to DR related tasks) could benefit from ML? What ML tasks could benefit from IV?
4. Reproducible/stable methods and the link of those aspects to sensitivity and consensus results;
5. Understanding the role of the user in mixed systems (which include both a ML and an IV component);
6. Interactive steerable ML methods (relation to intermediate results);
7. Methods from both fields for dynamic multivariate networks;
8. ML methods that can scale up to IV demands (especially in terms of interactivity);

9. Interpretable/transparent decisions;
10. Uncertainty;
11. Matching vocabularies/taxonomies between ML and IV;
12. Limits to ML;
13. Causality;
14. User guidance: precalculating results, understanding user intentions;
15. Mixing user and data driven evaluation (leveraging a ROC curve, for instance);
16. Privacy;
17. Applications and use cases;
18. Prior knowledge integration;
19. Formalizing task definition;
20. Usability;
21. Larger scope ML.

After some clustering and voting those topics were merged into six popular broader subjects which were discussed in working groups through the rest of the week:

1. Dynamic networks
2. Quality
3. Emerging tasks
4. Role of the user
5. Reproducibility and interpretability
6. New techniques for Big Data

The rest of the seminar was organized as a series of meeting in working groups interleaved with plenary meetings which allowed working groups to report on their joint work, to steer the global process, etc.

Conclusion

As reported in the rest of this document, the working groups were very productive as was the whole week. In particular, the participants have identified a number of issues that mostly revolve around complex systems that are being built for visual analytics. Those systems need to be scalable, they need to support rich interaction, steering, objective evaluation, etc. The results must be stable and interpretable, but the system must also be able to include uncertainty into the process (in addition to prior knowledge). Position papers and roadmaps have been written as a concrete output of the discussions on those complex visual analytics systems.

The productivity of the week has confirmed that researchers from information visualization and from machine learning share some common medium to long term research goals. It appeared also clearly that there is still a strong need for a better understanding between the two communities. As such, it was decided to work on joint tutorial proposals for upcoming IV and ML conferences. In order to facilitate the exchange between the communities outside of the perfect conditions provided by Dagstuhl, the blog “Visualization meets Machine Learning¹” was initiated.

It should be noted finally that the seminar was very appreciated by the participants as reported by the survey. Because of the practical organization of the seminar, participants did not know each other fields very well and it might have been better to allow slightly more

¹ <http://vismeetsml.b.uib.no/>

time for personal introduction. Some open research questions from each field that seems interesting to the other fields could also have been presented. But the positive consequences of avoiding a conference like schedule was very appreciated. The participants were pleased by the ample time for discussions, the balance between the two communities and the quality of the discussions. Those aspects are quite unique to Dagstuhl.

References

- 1 Daniel A. Keim, Fabrice Rossi, Thomas Seidl, Michel Verleysen, and Stefan Wrobel. Dagstuhl Manifesto: Information Visualization, Visual Data Mining and Machine Learning (Dagstuhl Seminar 12081). *Informatik-Spektrum*, 35:58–83, 8 2012.
- 2 Daniel A. Keim, Fabrice Rossi, Thomas Seidl, Michel Verleysen, and Stefan Wrobel, (editors). *Information Visualization, Visual Data Mining and Machine Learning (Dagstuhl Seminar 12081)*, Dagstuhl Reports, 2(2):58–83, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2012. <http://dx.doi.org/10.4230/DagRep.2.2.58>

2 Table of Contents

Executive Summary

Daniel A. Keim, Tamara Munzner, Fabrice Rossi, and Michel Verleysen 1

Overview of Tutorial Talks

Machine Learning and Visualisation
Ian Nabney 6

Visualization Analysis and Design
Tamara Munzner 6

Working Groups

Dynamic Networks
Tamara Munzner, Stephen North, Eli Parviainen, Daniel Weiskopf, and Jarke van Wijk 6

Machine Learning Meets Visualization: A Roadmap for Scalable Data Analytics
Daniel Archambault, Kerstin Bunte, Miguel Á. Carreira-Perpiñán, David Ebert, Thomas Ertl, and Blaz Zupan 7

User and Machine Learning Dialogue for Visual Analytics
Francois Blayo, Ignacio Díaz Blanco, Alex Endert, Ian Nabney, William Ribarsky, Fabrice Rossi, Cagatay Turkay, and B. L. William Wong 12

Bridging the Analytics Gap: Human-centered Machine Learning
Michael Sedlmair, Leishi Zhang, Dominik Sacha, John Aldo Lee, Daniel Weiskopf, Bassam Mokbel, Stephen North, Thomas Villmann, and Daniel Keim 13

Emerging tasks at the crossing of machine learning and information visualisation
Barbara Hammer, Stephen Ingram, Samuel Kaski, Eli Parviainen, Jaakko Peltonen, Jing Yang, and Leishi Zhang 16


Reproducibility and interpretability
Helwig Hauser, Bongshin Lee, Torsten Möller, Tamara Munzner, Fernando Paulovich, Frank-Michael Schleich, and Michel Verleysen 24

Participants 27

3 Overview of Tutorial Talks

3.1 Machine Learning and Visualisation

Ian Nabney (Aston University – Birmingham, GB)

License  Creative Commons BY 3.0 Unported license
© Ian Nabney

This talk describes two principal modes of data projection (or dimensionality reduction): topographic mappings and latent variable models. Principal Component Analysis is defined and it shown how it can be generalised to a non-linear projection based on distance preservation (topographic mapping exemplified by Neuroscale) or as a density model for the data (latent variable model exemplified by Generative Topographic Mapping – GTM). We then discuss how GTM can be extended to deal with missing values, discrete and mixed data types, hierarchies and feature selection. Illustrations from real applications are provided throughout.

3.2 Visualization Analysis and Design

Tamara Munzner (University of British Columbia – Vancouver, CA)

License  Creative Commons BY 3.0 Unported license
© Tamara Munzner

Computer-based visualization (vis) systems provide visual representations of datasets designed to help people carry out tasks more effectively. Visualization is suitable when there is a need to augment human capabilities rather than replace people with computational decision-making methods. The design space of possible vis idioms is huge, and includes the considerations of both how to create and how to interact with visual representations. Vis design is full of trade-offs, and most possibilities in the design space are ineffective for a particular task, so validating the effectiveness of a design is both necessary and difficult. Vis designers must take into account three very different kinds of resource limitations: those of computers, of humans, and of displays. Vis usage can be analyzed in terms of why the user needs it, what data is shown, and how the idiom is designed. I will discuss this framework for analyzing the design of visualization systems.

4 Working Groups

4.1 Dynamic Networks

Tamara Munzner (University of British Columbia – Vancouver, CA), Stephen North (Info-visible – Oldwick, US), Eli Parviainen (Aalto University, FI), Daniel Weiskopf (Universität Stuttgart, DE), and Jarke van Wijk (TU Eindhoven, NL)

License  Creative Commons BY 3.0 Unported license
© Tamara Munzner, Stephen North, Eli Parviainen, Daniel Weiskopf, and Jarke van Wijk

Networks are ubiquitous. Telecom networks, biological networks, software call graphs, citation graphs, sensor networks, financial transactions, social networks are some examples. In all

these cases, it is not only the network structure that is relevant. Nodes and edges have associated multivariate data, and also, they are often dynamic. Attributes change, and also, in many cases networks are derived from streams of events (messages, communications, transactions), where each event has at least a time stamp, and two nodes as associated data.

Such large and complex networks are notoriously hard to visualize and understand. Just showing the structure of networks with a few hundred nodes already gives rise to the so-called hairball images, dynamics and associated data are yet another dimension of complexity. In the visualization community, novel representations and interaction techniques are proposed, but the problem is far from solved. Hence, the generic question is what machine learning can offer to provide more insight in such networks. Typical tasks are the identification of outliers and anomalous behavior, partitioning a sequence of time steps into clusters, identification of trends and discontinuities, and finding dynamic clusters of nodes.

A lively discussion gave rise to three possible approaches. As model for the data we used a simple sequence of networks $G_i, i = 1, \dots, N$. The first approach concerns the use of a predictive model. Given such a model, one can predict for each time step a graph G'_i , given the other graphs $G_j, j \neq i$. Next, the difference between prediction and actual data can be shown, to reveal how the given data differs from expectation. A second idea is to use dynamic clustering: derive clusters across multiple graphs, such that emerging and disappearing clusters can be shown. Finally, one approach could be to translate each network into some feature vector, and next apply machine learning on these feature vectors.

Conceptually, all these approaches seem plausible and promising, however, also many questions remain. First, all these require models and metrics, for instance to make predictions, to cluster, and to select features; second, a question is if one should strive for generic solutions, or that questions on network data are strongly application dependent and require custom solutions.

The participants of the workshop were excited about the topic and the possible approaches. However, the group lacked expertise to make further steps. Therefore, we decided not to continue and join other working groups.

4.2 Machine Learning Meets Visualization: A Roadmap for Scalable Data Analytics

Daniel Archambault (Swansea University, GB), Kerstin Bunte (UC Louvain, BE), Miguel Á. Carreira-Perpiñán (University of California – Merced, US), David Ebert (Purdue University – West Lafayette, US), Thomas Ertl (Universität Stuttgart, DE), and Blaz Zupan (University of Ljubljana, SI)

License © Creative Commons BY 3.0 Unported license

© Daniel Archambault, Kerstin Bunte, Miguel Á. Carreira-Perpiñán, David Ebert, Thomas Ertl, and Blaz Zupan

4.2.1 Introduction

The big data problem requires the development of novel analytic tools for knowledge discovery and data interpretation (for example [1, 2]). The fields of visualization and machine learning have been addressing this problem from different perspectives and advances in both communities need to be leveraged in order to make progress. Machine learning has proposed algorithms that can address and represent large volumes of data enabling visualizations to scale. Conversely, visualization provides can leverage the human perceptual system to interpret and uncover hidden patterns in these data sets.

In this short report we identify areas where machine learning can assist the process of data visualization and areas where visualization can drive machine learning processes. These areas are summarized in Figure 1.

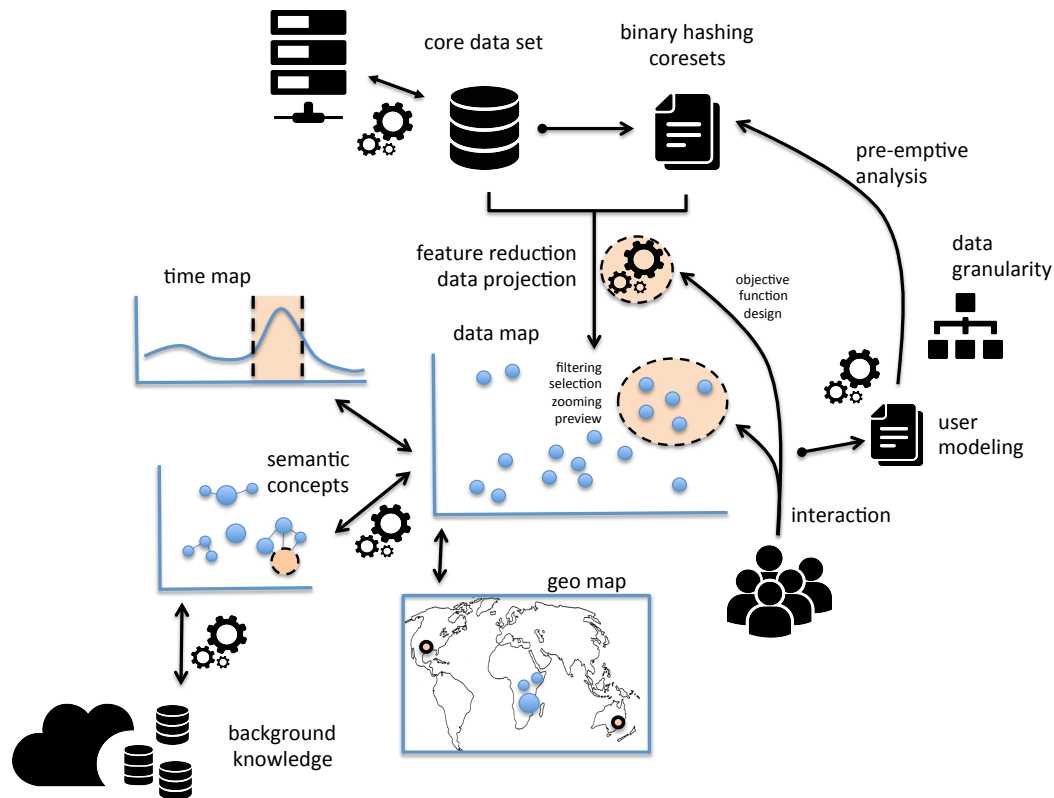
4.2.2 Visualization benefits from Machine Learning

Traditional uses of machine learning for visualization have included exploratory procedures such as feature selection, dimensionality reduction and clustering. Here we describe additional machine learning concepts that may be of benefit for visualization research.

Binary hashing. Binary hashing has emerged in recent years as an efficient way to speed up information retrieval of high-dimensional data, such as images or documents. Given, say, a query image, searching in a large database of images for the nearest images to the query is a high-dimensional nearest neighbor finding problem whose exact solution is computationally very expensive. For example, representing each image with a 300-dimensional vector of SIFT features would take over one terabyte for one billion images. In binary hashing, one maps every image to a compact binary vector so that Hamming distances in binary space approximately preserve distances in image space. Searching for neighbors in binary space is much faster because 1) the dimensionality of the binary vector is much smaller than the dimensionality of the image, 2) Hamming distance computations can be done very efficiently with hardware support for binary arithmetic, and 3) the size of the binary-vector database is small enough that it can even fit in RAM memory rather than disk. In the earlier example, using 32 bits per image the database would take 4 GB. The success of binary hashing depends on being able to learn a good hash function, which maps images to bit vectors so that distances are approximately preserved. Initial algorithms learned a dimensionality reduction mapping and simply truncated it to output binary values [3], while recent efforts try to optimize the function directly respecting the binary nature of its outputs [4, 5].

As an example application, consider visualizing a stream of Tweets. Given a new Tweet, we can turn it into a high-dimensional vector using a bag-of-words representation and then map it to binary space using the binary hash function. Searching in a binary database of Tweets quickly retrieves a selection of approximate neighboring Tweets, which can be refined to keep only true neighbors by computing distances between the retrieved bag-of-words vectors and the query.

Coresets. Besides the dimensionality which leads to high computational costs and memory requirements also the number of samples influences the efficiency of many applications whenever very large amounts are collected as in Astronomy, Photography, streaming and so on. Random sampling, feature extraction and ϵ -samples are often used strategies to deal with this problem. This leads to a general concept combining these ideas referred to as coresets [6, 7]. The aim is to find a small (weighted) subset of the data, which guarantees, that a training procedure based on this subset provides comparable good results also for the original set. The effectiveness has been shown for several objectives, ranging from for example dimension reduction, clustering and Gaussian Mixture Models and surprisingly also coresets with size independent from the size of the data set have been proven. Moreover, efficient parallel and distributed strategies to find coresets are proposed, which makes them perfectly suitable for big data analysis and streaming settings. Information visualization can directly benefit from this concept, since it usually depends on pairwise similarities or distances resulting in quadratic complexity with respect to the number of data items.



■ **Figure 1** Possible interplay between machine learning and data visualization. The core data set (top), possibly storing the information from the data stream, is preprocessed for binary hashing and coresets discovery. Preprocessing enables index-based data retrieval, selection of the representative data instances, and fast distance computation. Multi-view visualization initially displays data in the coreset, but also supports user in digging deeper and retrieving data from neighborhood, time, location or concept-specific spaces. Data-related semantic concepts are retrieved from related data bases and organized in ontology or network. Visualizations are interlinked: any change in selection in one view updates the information in all other views. Machine learning algorithms for clustering, assessment of concept enrichment, outlier detection and classification of uncharacterized data instances are triggered on the fly. User's interactions are recorded and modeled, and provide means of predicting them and executing the most likely data-intensive operations that the user can trigger in the future before they are actually needed. User can change the attributes or position of data instances in any visualization, thus visually changing the objective function that is optimized in the visualizations. Change of objective function is followed by repositioning of data elements in the visualizations.

Inclusion of background knowledge. Besides the core data which we are trying to analyze, there may be additional information available that may shed light on the interaction between data entities, or additionally explain the discovered data patterns. Background knowledge may be incorporated at various stages of data analysis. For machine learning, it may serve as a prior that constrains the hypothesis space and steers the optimization towards models that are consistent both with data and additional information. For visualization, background knowledge may provide information that support interpretation. What characteristics outside of the data space are common to a set of co-clustered data points? What is the match between the visualized data and the concepts that are related to the problem investigated but were not included in the original data set? Crucial to exploration of the interplay between the data any additional information are graphical user interfaces to access and explore such interaction, and quantification of relations between data instances and concepts to draw statistically founded conclusions. An example of the later are enrichment analysis techniques from bioinformatics [8], which ranks the data annotation terms according to their association with a selected group of data entities.

Visualization of classifiers. Recent approaches accommodate for the growing demand of interpretable models, which lead to visualizations, not only showing the data, but also an inferred classification model [9, 10]. This enables the use of the human perceptual qualities to detect: 1) potential mis-labelings which might emerge as outliers, 2) noisy regions which are difficult to classify, 3) the modality of each class and 4) overfitting effects of the model for example.

Visualization of machine learning processes. Recent work in both the machine learning and human computer interaction communities has focused on how to use visualization in order to improve how we tune machine learning approaches. Specifically, the approaches have been applied to the problem of network alarm triage [11] and optimizing machine learning approaches for given performance constraints [12, 13, 14]. This work provides a way to optimize machine learning processes for given tasks, instead of treating the approach as a black box.

Steerability, semantic zoom and user constraints. One of the most promising applications for information visualization to machine learning is steerability. Steerable approaches in the field of visualization allow for the user to interactively guide large computations towards areas of interest. Such approaches, when applied in conjunction with machine learning can be very powerful, allowing heavy weight computations to be targeted to areas of interest in a very large data set. Steerable approaches first emerged in the field of scientific visualization [15] and have been subsequently been applied to the process of visualizing graphs [16, 17].

Moreover, user constraints, like for example walls in maps or must-link and cannot-link constraints for clustering, can be accumulated by interactions with a display. Any machine learning algorithms suitable for constraint-based optimization as satisfiability optimization can benefit from such interactive solutions. First steps to directly incorporate user constraints into the optimization process of visualizations has been taken for example in [18]. These constraints can be imposed through user interaction and the resultant computation could be used in conjunction with a semantic zoom.

4.2.3 Way Forward

Visual design of objective functions. Recently, some methods have been proposed to make model parametrization and data exploration more intuitive without requiring deep methodological knowledge of the data expert. Those approaches provide for example a

simplex where a point in the area corresponds to a parametrization of the underlying model comparable to multidimensional sliders. Other tools facilitate an interactive data exploration, by visually combining modules implementing different data processing steps, which could be combined by the user. However, the parametrization is a very high level design mechanism and limited in its impact on the final model. To change the fundamental design and assumptions of the model one would need to interact on much lower levels such as the mathematical formulation. It would be interesting when a user could visually combine mathematical atoms to form new objectives as for instance using graphical models in Bayesian formulations, which are inferred automatically.

Modelling of user interactions. Machine learning should not only be used for summarizing data. One approach is to use machine learning to learn user actions and predict the likely future ones. The area of adaptive user interfaces and intelligent user interfaces could be applied to the field of information visualization to determine likely future interactions with the system to give it a *head start* on heavyweight computational processes in a steerable environment.

Data fusion. In making quality decisions, us, humans, tend to use all available information that is directly or only indirectly related to the problem. In machine learning, the notion of wide-range data integration has been explored by data methods of fusion. So far, data fusion has primarily focused on development of predictive models by combining different data sources through, say, through kernel-based methods [19] or collective matrix factorization [20]. The research in this field is important to big data, as it addresses the variety and span of data sources. To bring the resulting models to the data analyst, however, data fusion would need to be combined with data visualization using the approaches that have yet to be conceptualized and developed.

References

- 1 Junghoon Chae, Dennis Thom, Harald Bosch, Yun Jang, Ross Maciejewski, David S Ebert, and Thomas Ertl. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 143–152. IEEE, 2012.
- 2 Harald Bosch, Dennis Thom, Florian Heimerl, Edwin Puttmann, Steffen Koch, Robert Kruger, Michael Worner, and Thomas Ertl. Scatterblogs2: Real-time monitoring of microblog messages through user-guided filtering. *IEEE Trans. on Visualization and Computer Graphics*, 19(12):2022–2031, 2013.
- 3 Yair Weiss, Antonio Torralba, and Rob Fergus. Spectral hashing. In Daphne Koller, Yoshua Bengio, Dale Schuurmans, Leon Bottou, and Aron Culotta, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 21, pages 1753–1760, 2009.
- 4 Miguel Á. Carreira-Perpiñán and Ramin Raziperchikolaei. Hashing with binary autoencoders. In *Proc. of the 2015 IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, June 7–12 2015.
- 5 Ramin Raziperchikolaei and Miguel Á. Carreira-Perpiñán. Learning hashing with affinity-based loss functions using auxiliary coordinates. arXiv:1501.05352, January 21 2015.
- 6 Pankaj K. Agarwal, Sarel Har-Peled, and Kasturi R. Varadarajan. Geometric approximation via coresets. In *Combinatorial and Computational Geometry, MSRI*, pages 1–30. University Press, 2005.
- 7 Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *Proceedings of the Forty-third Annual ACM Symposium on Theory of Computing, STOC'11*, pages 569–578, New York, NY, USA, 2011. ACM.

- 8 Jui-Hung Hung, Tun-Hsiang Yang, Zhenjun Hu, Zhiping Weng, and Charles DeLisi. Gene set enrichment analysis: performance evaluation and usage guidelines. *Briefings in Bioinformatics*, 13(3):281–291, 2012.
- 9 A. Schulz, A. Gisbrecht, K. Bunte, and B. Hammer. How to visualize a classifier? In B. Hammer and T. Villmann, editors, *New Challenges in Neural Computation (NC2)*, ser. *Workshop of the GI-Fachgruppe Neuronale Netze and the German Neural Networks Society in connection to DAGM 2012*, Graz, Austria, August 2012. LNCS.
- 10 Alexander Schulz, Andrej Gisbrecht, and Barbara Hammer. Using nonlinear dimensionality reduction to visualize classifiers. volume 7902 of *IWANN(1)*, pages 59–68. Springer, 2013.
- 11 S. Amershi, B. Lee, A Kapoor, R. Mahajan, and B. Christian. Cuet: Human-guided fast and accurate network alarm triage. In *Proc. CHI 2011*, pages 157–166, 2011.
- 12 A. Kapoor, B. Lee, D. Tan, and E. Horvitz. Interactive optimization for steering machine classification. In *Proc. of CHI 2010*, pages 1343–1352, 2010.
- 13 A. Kapoor, B. Lee, D. Tan, and E. Horvitz. Performance and preferences: Interactive refinement of machine learning procedures. In *Proc. of AAAI 2012*, 2012.
- 14 S. Amershi, M. Chickering, S. M. Drucker, B. Lee, P. Simard, and Jina Suh. Modeltracker: Redesigning performance analysis tools for machine learning. In *Proc. CHI 2015*, 2015.
- 15 S.G. Parker and C.R. Johnson. SCIron: A scientific programming environment for computational steering. In *Proc. of Supercomputing*, 1995.
- 16 D. Archambault, T. Munzner, and D. Auber. GrouseFlocks: Steerable exploration of graph hierarchy space. *IEEE Trans. on Visualization and Computer Graphics*, 14(4):900–913, 2008.
- 17 D. Archambault, H. C. Purchase, and B. Pinaud. The readability of path-preserving clusterings of graphs. *Computer Graphics Forum*, 29(3):1173–1182, 2010.
- 18 Kerstin Bunte, Matti Järvisalo, Jeremias Berg, Petri Myllymäki, Jaakko Peltonen, and Samuel Kaski. Optimal neighborhood preserving visualization by maximum satisfiability. In *Proceedings of AAAI-14, The Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.
- 19 Shi Yu, Léon-Charles Tranchevent, Bart De Moor, and Yves Moreau. *Kernel-based Data Fusion for Machine Learning*. Springer-Verlag, Berlin, Heidelberg, 2011.
- 20 Marinka Zitnik and Blaz Zupan. Data fusion by matrix factorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37:41–53, 2014.

4.3 User and Machine Learning Dialogue for Visual Analytics

Francois Blayo (Ipseite SA – Lausanne, CH), Ignacio Díaz Blanco (University of Oviedo, ES), Alex Endert (Georgia Institute of Technology, US), Ian Nabney (Aston University – Birmingham, GB), William Ribarsky (University of North Carolina – Charlotte, US), Fabrice Rossi (Université Paris I, FR), Cagatay Turkay (City University – London, GB), and B. L. William Wong (Middlesex University, GB)

License © Creative Commons BY 3.0 Unported license
 © Francois Blayo, Ignacio Díaz Blanco, Alex Endert, Ian Nabney, William Ribarsky, Fabrice Rossi, Cagatay Turkay, and B.L. William Wong

Thomas and Cook (2005) presented the visual analytics community with a challenge to create visualization technologies that work interactively and smoothly with computational algorithms. They describe such a dialog as analytic discourse. They described this as “...visually-based methods to support the entire analytic reasoning process”, including the analysis of data as well as structured reasoning techniques such as the construction

of arguments, convergent- divergent investigation, and evaluation of alternatives. These methods must support not only the analytical process itself but also the progress tracking and analytical review processes.

The merger of machine learning and visual analytics presents many potential opportunities for visual data analysis. Visual analytics leverages the cognitive and perceptual abilities of humans to enable them to explore, reason, and discover data features visually. Machine learning leverages computational abilities of computers to perform complex data-intensive calculations to produce results for specific questions or tasks. Currently, visual analytic techniques exist that make use of select machine learning models or algorithms (often, dimension reduction techniques). However, there are additional techniques that can apply to the broader visual data analysis process. Doing so reveals opportunities for how to couple user tasks and activities with such models.

The discussion at this Dagstuhl seminar focuses on the role of the user in this process of integrating machine learning into visual analytics. We discussed challenges and difficulties of designing a system that would enable analytic discourse. How should specific machine learning techniques be incorporated into the visual data exploration process? We present a discussion of the role of user interaction in such a dialog between machine learning techniques, interactive visualisation and cognitive processes, and provide a scenario to illustrate these concepts. What would be or should be the role of the user when we combine machine learning with interactive visualization in ways that would enable users to steer and drive the computational algorithms?

We claim that user interactions are an important aspect of such a combination. In visual analytics, user interactions have been designed and implemented as mechanisms by which users can augment the visualization parameters, filter data, and other direct changes to the application. In machine learning, user interaction has been used as directed feedback on results of computation (e.g., classification models, predictive models, etc.). However, we challenge these two communities to consider an additional lens through which user interaction can be viewed. We posit that every user interaction encodes some (potentially small) part of analytical reasoning and insight. The challenge posed to the community is how to adequately leverage these bits of analytical reasoning and integrate them into the holistic visual analytics system.

4.4 Bridging the Analytics Gap: Human-centered Machine Learning

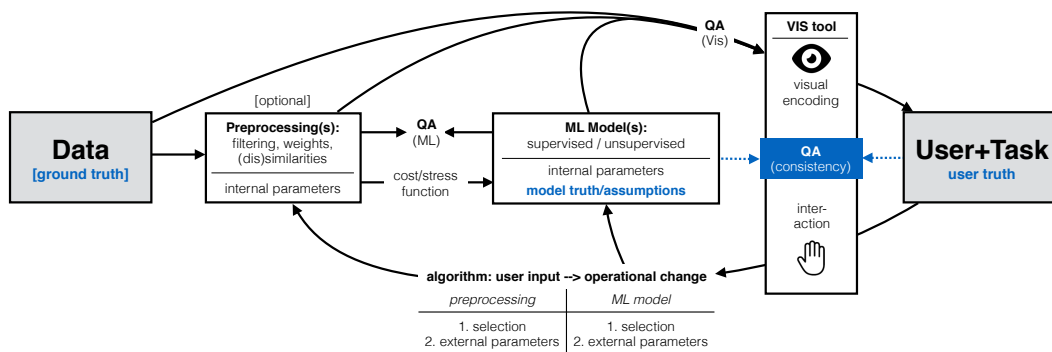
Michael Sedlmair (Universität Wien, AT), Leishi Zhang (Middlesex University, GB), Dominik Sacha (Universität Konstanz, DE), John Aldo Lee (UC Louvain, BE), Daniel Weiskopf (Universität Stuttgart, DE), Bassam Mokbel (Universität Bielefeld, DE), Stephen North (Infovisible – Oldwick, US), Thomas Villmann (Hochschule Mittweida, DE), and Daniel Keim (Universität Konstanz, DE)

License © Creative Commons BY 3.0 Unported license

© Michael Sedlmair, Leishi Zhang, Dominik Sacha, John Aldo Lee, Daniel Weiskopf, Bassam Mokbel, Stephen North, Thomas Villmann, and Daniel Keim

The goal of visual analytics systems is to solve complex problems by integrating automated data analysis methods with interactive visualizations. While numerous visual analytics systems have been developed for specific application problems, a general understanding of how this integration can be realized is still largely missing.

Towards the goal of better understanding this interplay, our working group developed a



■ **Figure 2** Our conceptual framework. The main components of the pipeline are shown in the center. Visual encoding of results at different stages of the pipeline are indicated via arrows at the top, which point through the VIS tool to the User. User interactions are indicated through the arrows at the bottom, again via the VIS tool. Different versions of “truth” are highlighted in blue, together with a quality assurance (QA) component that helps ensuring consistency between the ML components and the user.

framework that conceptualizes how integration of machine learning methods and interactive visualizations can be implemented (see Figure 2). We identified aspects of machine learning methods, which are amenable to be controlled interactively by the user, such as the choice and parameterization of machine learning models. While some of these aspects can be automatically optimized by pre-defined cost functions, in many applications it is crucial to allow the user to control them interactively. Our framework makes the crucial interplay between automated algorithms and interactive visualizations more concrete. To show its utility, we used it to analyze several existing visual analytics systems against the framework, demonstrating that it provides useful understanding as well as guidelines when developing and evaluating them.


Based on our framework, we finally characterized a set of 11 open challenges:

1. Mapping user input to ML model adaptation – At the core of our conceptual framework lies the idea that external parameters of an ML model or preprocessor can be adapted via iterative, and direct user interactions. Some simple examples exist, such as updating a parameter of a dimension reception model based on how a user moves around points in a scatterplot. However, mapping user inputs to more complex actions, such as switching between different model types, remains an open challenge.
2. Discontinuous changes – Implementing such more complex interactions, may sometimes cause major, abrupt changes in the underlying ML components. A major challenge is how to communicate such abrupt changes in a perceptually understandable way to the user, in order to keep her in the loop.
3. Effective learning from small data – An algorithmic challenge that our envisioned human-in-the-loop scenario poses is learning from a small number of user interactions, likely in the single or low double digits. While these interactions will be used as stimuli for training the ML model’s internal parameters, most ML methods require a larger set of input data to train the model, typically hundreds or thousands of input stimuli.
4. Interactive and Scalable Algorithms – Another technical challenge is that the user should not be disrupted by long response times occurring during adaptation of ML models. Therefore, training procedures must be efficient in terms of computation time. In this regard, new approximation approaches, and methods for including intermediate results will be needed.

5. Balance between Model and Visual Quality – A major challenge in a rich human-in-the-loop analysis process is assuring both ML model quality and visualization quality. However, the two types of quality preservation do not always align. For example, a visual embedding that preserves the input data structure well may not have good readability due to high dimensionality. Data sparsity and noise can cause clutter and poor group separation. While some techniques exist, the challenge is to provide a clear indication of both quality measures to the user and help them to find the right balance between the two, so meaningful analysis can be carried out.
6. Consistency between Model and Human – In current visual analytics systems, checking consistency between model and user is often done manually. The user must evaluate the model and provide feedback to the system. When a conflict between the two arises, the outcome can be biased. Such problems can be alleviated by developing automatic methods that check consistency quality, highlight inconsistency, and recommend appropriate actions. Note that, though consistency between human and machine is desirable, it does not guarantee correctness per se.
7. User Guidance – Current general purpose systems, such as R, offer multiple choices of preprocessors and ML models that can be applied to analyze data. Application users who are not ML experts, however, often find it difficult to know which choices are most suitable for the data and task at hand, and to find good parameter settings for the selected ML components. Assistance and guidance from the system is therefore of utmost importance.
8. Better Perceptual Quality Measures – While numerous quality measures have been designed for algorithmic purposes, we find few measures that have a truly perceptual motivation. Current visualization measures do not cover any complex approaches from perceptual psychology to accurately capture mechanism of human visual perception. Such models are especially difficult if they want to include human-computer interaction and data dependency. Having an accurate model of human perception would not only be helpful for guiding users through the space of visualization design choices, but also for ensuring consistency between the user and ML models as discussed above.
9. Uncertainty Description, Quantification, and Propagation – Another challenge is that we need to describe and compute uncertainty introduced by the various pieces within the pipeline of using visualization and machine learning together. Describing and quantifying uncertainty in the interplay between user, task, and ML model is a non-trivial endeavor.
10. Visualization of Uncertainty – Once we have a quantification of uncertainty, what shall we do with it? One research question deals with the visualization of such uncertainty. There is much previous work on visualization techniques to display data uncertainty of spatial data, such as volume or flow visualization. We find much less work on uncertainty visualization of abstract data, such as high-dimensional data visualization, common in ML applications.
11. Uncertainty Reduction – A related challenge is how we can reduce the amount of uncertainty. One possibility is to steer the visual analytics process toward a “sweet spot” where the process becomes less sensitive to the influence of uncertainty. Here, sensitivity analysis or similar approaches might be adapted. Another approach to reduce the uncertainty from user input (such as inaccuracies introduced by annotation uncertainty) could be automatic checks for consistency with the machine-learning model. This idea is tightly linked to having an appropriate quality assessment for consistency between model and user.

4.5 Emerging tasks at the crossing of machine learning and information visualisation

Barbara Hammer (Bielefeld University, DE), Stephen Ingram (UBC, Vancouver, CA), Samuel Kaski (Aalto University, Helsinki, FI), Eli Parviainen (Aalto University, Helsinki, FI), Jaakko Peltonen (Aalto University / University of Tampere, FI), Jing Yang (University of North Carolina, Charlotte, US), and Leishi Zhang (Middlesex University, London, UK)

License  Creative Commons BY 3.0 Unported license
© Barbara Hammer, Stephen Ingram, Samuel Kaski, Eli Parviainen, Jaakko Peltonen, Jing Yang, and Leishi Zhang

4.5.1 Introduction

An ever increasing number of domains is accompanied by digital fingerprints: industry 4.0 with heterogeneous sensor streams monitoring and controlling industrial processes; smart sensor signals of everyday life which become ubiquitous in the context of smart phones, wearable devices, and digitalisation of the automotive sector; highly sensitive medical diagnostics based on a variety of different biotechnologies leading to individualised -omics sources; the financial market which is characterised by a multitude of digitally stored indicators; social life which is tightly mirrored in social media and social networks; or even politics which, increasingly, makes use of digital information and the underlying ways of decision making [1]. This digital data revolution places new challenges towards computer scientists: they are not only developing new technologies for efficient data measurement, pervasive data storage, privacy preservation, etc, but they also face the challenge to enable humans to cope with the information buried in these data and take according action. This has been identified as one of the major questions when it comes to big data, and the term ‘big data analytics’ has been coined as a key capacity of modern society [2].

Machine learning (ML) and information visualisation (InfoVis) constitute two pivotal disciplines which enable humans to unravel the information hidden in digital data. Albeit these two disciplines address similar questions and challenges, their underlying technologies and theoretical background often differ. Research directions such as the developments put under the overarching umbrella of ‘scalable visual analytics’ constitute promising attempts to bridge this gap [3], and there do exist formalisms and tools which successfully rely on aspects of both worlds [4]. The goal of this contribution is to discuss such links by zooming onto the tasks and questions which are shared by ML and InfoVis, and their respective approaches to tackle these tasks. Thereby, we do not cover the full spectrum. Rather we put spotlights on interesting aspects at two different levels of scientific granularity: differences and shared technology, respectively, as concerns central paradigms of the data processing pipeline in InfoVis and ML, on the one hand; and topics which we regard as emerging topics in the domains of ML and InfoVis, which share a common research question but which are looked at from two different points of view in the two disciplines. We discuss each of these spotlights separately in a short paragraph in the sequel.

4.5.2 Classical dimensionality reduction

Often, data are vectorial, but high dimensional, such that its direct inspection as points in the plane is impossible. Their intuitive visual access constitutes one of the classical tasks which are addressed by both, InfoVis and ML – but technologies differ [5]. InfoVis provides a number of different techniques to display such data, such as scatter plots, parallel coordinates, heat maps, glyphs, or Chernoff faces, as well as interactive exploration e.g.

based on four methods. In this context, a major question which is investigated, is how these visualisation technologies align with human perception [6]. Conversely, ML almost solely relies on a static display of high dimensional data as a scatter plot, but it explores a variety of different approaches to learn suitable two-dimensional coordinates from the given data which preserve as much structure of the original data as possible [7]. The focus is on the different mathematical ways to formalise the concept of structure preservation, and its efficient computational modelling.

These foci constitute two different views on the problem: InfoVis concentrates on human perception and puts the user into the centre, while ML focusses on (often nonlinear) aspects in the given data and their mathematical formalisation. These different views of the same problem open the way towards new paradigms, which combine rich visual display technologies and interaction methods as offered by InfoVis with highly flexible data driven structure preservation as provided by ML technology. Such enriched data displays have a great potential for emerging areas such as biomedical data analysis where, often, heterogeneous information or additional structures have to be taken into account [8, 9, 10, 11]

4.5.3 Modelling

Both, ML and InfoVis essentially model observed data in such a way that the information buried in the data can easily be accessed by humans. Thereby, a crucial part is to identify general paradigms and workflows which allow researchers to access the given data in a principled and scientifically valid way.

For InfoVis, general workflows such as the InfoVis mantra ‘overview first, zoom and filter, then details on demand’ and clear relations of the technology to be used for display and the type of data to be displayed are well established [12, 13]. These modelling paradigms offer guidelines for the ‘scientific language’ which can be used for data visualisation and the realisation of the dynamics of such display. These principles are usually not tailored to the exact values of the data to be displayed.

For ML, the key aim is to model the given, observed data, and one overarching paradigm underlying modelling in ML is the language of probability theory and statistics: often, learning is phrased as probabilistic modelling of the given data points which are regarded as samples of an underlying data distribution; modelling refers to the inference of the latter, i.e. estimating generative probabilistic models from a finite number of given observations [14]. Thereby, computational learning theory provides a mathematical justification that this principle is valid. Such modelling is data centred, in the sense that different models result from different measurements, and the influence of the observed data on the final model can be quantified by the deviation of the resulting distribution and the prior.

In principle, probabilistic modelling is universal, being capable of modelling every possible underlying regularity – in practice, assumptions have to be made to avoid overfitting, and regularisation which is based on prior knowledge or universal priors (such as sparsity) has to be used. A good choice of priors remains a challenge in particular for sparse measurements and heterogeneous data sources. Here human intuition could help to regularise accordingly, opening up an interesting support line from the InfoVis field.

Interestingly, a Bayesian view on InfoVis, which treats data and also user interactions as observations, opens the ground towards an automation of display selection and adaptation of the views according to the data. Recently, some promising research along this line has been proposed, see e.g. [15].

4.5.4 Quantitative evaluation

Both, InfoVis and ML face the challenge to quantitatively and qualitatively evaluate their techniques. The used methods differ fundamentally, a fact which closely mirrors the user centred versus data centred view of the two disciplines.

For InfoVis, the evaluation of a system usually takes place in the form of user studies or user feedback, such as expert evaluation, lab studies, or field studies [16]. These enable a formal evaluation of important aspects of InfoVis systems such as their functionality, effectiveness, efficiency, or usability. Such evaluations are often time consuming, and they require a clear study design. Notably, these techniques do not make explicit assumptions about human perception, since humans are directly evaluating the models using their cognitive capabilities. Interesting attempts try to match human perception and formal mathematical measurements, which could result in a speed up of the design process due to the availability of computable measures mirroring human perception [17].

For ML, evaluation is almost solely data centred, and evaluation measures have its roots in statistics. Since the majority of ML technologies can be found in the field of so-called supervised learning, classical evaluation measures for ML technology refer to cost measures such as the classification error or regression error as evaluated in a cross-validation. It has been a long debate how to evaluate unsupervised methods such as clustering or dimensionality reduction for data visualisation, and widely accepted quantitative measures for the latter just emerged recently [18, 19]. One main problem in this context consists in the fact that data visualisation and unsupervised data analysis is a mathematically ill-posed problem, and it depends on the setting at hand, which aspects of the data are of interest for the user. It is often not clear how to formalise these fuzzy goals in terms of mathematical cost functions and model priors. In this respect, ML can benefit from the insights and evaluation technology which is common in InfoVis, since it enables to take the user expectation into account without the necessity to express the latter within mathematical terms.

Conversely, by focussing on an underlying data distribution and the generalisation ability of a model to new data, ML can rely on strong techniques offered by statistics. A general technology which allows to evaluate the generalisation ability and robustness of a model, for example, is provided by sampling methods such as bootstrap statistics or cross-validation [20]. Hence it is easily possible to automatically evaluate a given algorithm or model as concerns its statistical robustness – a prerequisite which is independent from the overarching goal of modelling.

4.5.5 Big and streaming data

Albeit ML and InfoVis constitute two key technologies when it comes to big data, both techniques also face a number of new challenges in this context [2]. Both disciplines have to cope with the increasing computational and memory demands when it comes to big data. Hadoop's map-reduce, as an example, constitutes a widely used technology in both domains [21].

Besides these grounds, both domains develop new data structures and algorithms to speedup computational costs for core methods such as spatio-temporal data representation or dimensionality reduction. Interesting recent proposals, for example, rely on an intricate hierarchical representation of data and a suitable summary of the information content at each hierarchical level: within InfoVis, so-called nanocubes enable to deal with tens of billions of data points efficiently [22]. In ML, a similar concept which has its roots in statistical physics has recently been proposed to speed up dimensionality reduction techniques from quadratic to only log-linear complexity [23, 24].

Often, data are not only big but arrive continuously over time. In such cases, the challenge is to face the specific data characteristics caused by its dynamic arrival. In InfoVis, streaming data visualisation deals with the problem to take user expectation and perception of temporal changes into account. As an example, dynamic graph drawing tries to optimally balance dynamic changes and constant characteristics within a visual display of dynamic graphs [25]. Besides human perception, enriched mathematical concepts such as parameterised lines can open the way to novel, efficient dynamic displays [26].

For ML, one of the core problems of streaming data analysis consists in the fact that a crucial assumption underlying classical ML is violated: data are usually no longer independently and identically distributed, rather trends occur. Thus, ML methods have to cope with the challenge of data trend, emerging and vanishing concepts, and intricate data dependencies over time, with quite a few novel approaches and theoretical models popping up to deal with these problems [27].

4.5.6 Few data

In the context of heterogeneous data and user interaction, a phenomenon, which lies on the opposite side, takes place: methods face the challenge to learn from few data only.

One example instantiation of this challenge is the detection of rare events within large data sets, popular applications being e.g. network intrusion detection, rare event detection, customer preference learning, crime detection, or change point detection [28, 29]. Here, specific ML and InfoVis techniques have to be used which are capable of dealing with highly imbalanced data sets and putting its focus on the few observed anomalies in the data, since the majority of observations belong to the class of ‘normal’ events in such settings.

Another application area deals with very few labeled events only, such as instantaneous learning from few examples. This becomes possible provided auxiliary information is taken into account, such as strong priors in Bayesian modelling of visual categories [30], or the wisdom of the crowd which manifests itself in social media [31].

One domain where learning from few examples would be very useful is the automated annotation of given data. Typically, interactive systems are offered by InfoVis technology which enable experts to annotate such events; still, this is usually too time consuming for the full data. Here automation as offered by ML would help. Currently, most automated annotation systems are specialised to the respective domain, covering e.g. genomic data annotation, texts, images, or specific events in time series data [32, 33]. An interplay of ML and InfoVis techniques could help to generalise these approaches towards a domain independent technique.

4.5.7 Causality

Interactive data analysis is concerned with insights into the given information such as characteristic patterns, summaries, or typical cases. Often, the causality of observations constitutes a key question humans are interested in: which measurements and observations are relevant for a certain effect and how do they relate to each other? What is the cause of a particularly interesting / annoying / relevant observation, and how can this effect be changed? While correlations of events can easily be determined based on classical statistics, the notion of causality – which event is the cause of which other event – requires a more in depth analysis. Typically, it does not suffice to analyze available observations only, rather it demands for a mediated probability or expert insight.

Interestingly, in recent years, the automated inference of causality from measured data

has become more and more relevant in different areas of ML, caused by increasing data sets e.g. in neurobiology (such as action potentials of neurons, based on which neural connectivity should be predicted) [34, 35]. There do exist possibilities to infer causality in some settings, provided suitable prior assumptions are integrated into the models. One example is offered by independent component analysis, which is capable of unraveling mixed sources based on the notion of statistical independence only, and which can also be used for causality detection for linear relations. Naturally, human interaction can also help to clarify unclear cases which can occur due to highly nonlinear effects or sparse sampling; here an interactive analysis where ML and InfoVis provide different insights can be beneficial.

Having identified causal relationships, it remains a challenge to present these insights in such a way that the user can use this information for decision making in complex settings. A challenge is given by the fact that data are high dimensional, and causality is usually not only spotting relationships between simple measurements, rather it relates to significant macro-properties of the system, such as traffic jams and road network design in interactive traffic analysis. InfoVis provides a few technologies how to display such information efficiently and effectively in different contexts [36, 37].

4.5.8 Computational creativity

Automatic storytelling has been dubbed as one emerging area in InfoVis which goes beyond the mere display of data; rather it enables to build a whole story and line of argumentation around given data, supporting the arguments by suitable visualisations where appropriate [38]. Besides novel InfoVis tools, this task faces the challenge to infer a reasonable storyline automatically or with the help of the user from the given data; hence there is the need for fundamental arguing principles and inference mechanisms, typically techniques from ML and AI. Further, stories are often built around interesting exceptional events, hence rare meaningful events have to be detected automatically, as already discussed in section 4.5.6.

In ML, this question also touches on what is referred to as ‘computational creativity’: where are the relevant novel uncommon insights buried in the data? This imprecise notion can be partially matched with mathematical measures such as the entropy, which measures the amount of surprise in a data set, and successful technical systems which make use of these principles e.g. for efficient reinforcement learning have been proposed [39].

4.5.9 Collaborative work

Web and social media, among other aspects, enable an ever increasing availability of collaborative sources for data analysis: they provide basic data sources and background information based on which data analysis can be enriched, popular examples being e.g. collaborative filtering [40]; automated annotation and the wisdom of the crowd enables to rely on label information which, due to the sheer size of the participants, can be statistically very reliable; further, the web provides an environment where humans can increasingly work together and collaborate, making according platforms mandatory, examples are MOOCs or shared bioinformatics data bases.

These developments provide new possibilities but also new challenges for InfoVis and ML, such as the following: how to visualise and analyse data which comes from different sources, how to align the usually slightly different data representations and persistently store the involved information? One crucial aspect is, for example, a common data space or language shared by the collaborators, a question which is tackled under the umbrella of transfer learning in ML [41], and addressed in first systems in the InfoVis field [42].

4.5.10 Discussion

We have discussed some tasks and questions shared by InfoVis and ML, pointing out the different view of the two disciplines due to their user centred versus data centred view. This difference often results in different technologies, which can be combined to open up revenues for new, even more powerful technologies. With the advent of big data and distributed sensors, data sets and analysis tasks become ever more complex: data sources are heterogeneous, data are distributed, and massive volumes have to be addressed. At the same time the tasks, which can be tackled, are no longer restricted to simple correlations, but complex questions which relate to planning and decision making are investigated. This calls for a combination of the two fields, such that it becomes possible to address these challenged with integrated methods which can automate inference wherever possible, but which can use interactive analysis wherever expert feedback is mandatory.

References

- 1 Tom Khalil. Big data is a big deal. White House, Sep 2012.
- 2 Committee on the Analysis of Massive Data, Committee on Applied and Theoretical Statistics, Board on Mathematical Sciences and Their Applications, Division on Engineering and Physical Sciences, and National Research Council. *Frontiers in Massive Data Analysis*. National Academic Press, 2013.
- 3 Daniel A. Keim. Solving problems with visual analytics: The role of visualization and analytics in exploring big data. In *Datenbanksysteme für Business, Technologie und Web (BTW), 15. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS), 11.-15.3.2013 in Magdeburg, Germany. Proceedings*, pages 17–18, 2013.
- 4 Tuukka Ruotsalo, Giulio Jacucci, Petri Myllymäki, and Samuel Kaski. Interactive intent modeling: information discovery beyond search. *Commun. ACM*, 58(1):86–92, 2015.
- 5 Stephen Ingram, Tamara Munzner, Veronika Irvine, Melanie Tory, Steven Bergner, and Torsten Möller. Dimstiller: Workflows for dimensional analysis and reduction. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology, IEEE VAST 2010, Salt Lake City, Utah, USA, 24-29 October 2010, part of VisWeek 2010*, pages 3–10, 2010.
- 6 Michael Friendly. Milestones in the history of thematic cartography, statistical graphics, and data visualization. In *13th International Conference on Database and Expert Systems Applications (DEXA 2002), Aix en Provence*, pages 59–66. Press, 1995.
- 7 Andrej Gisbrecht and Barbara Hammer. Data visualization by nonlinear dimensionality reduction. *Wiley Interdisc. Rev.: Data Mining and Knowledge Discovery*, 5(2):51–73, 2015.
- 8 Matthew Brehmer, Stephen Ingram, Jonathan Stray, and Tamara Munzner. Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists. *IEEE Trans. Vis. Comput. Graph.*, 20(12):2271–2280, 2014.
- 9 Jing Yang, Yujie Liu, Xin Zhang, Xiaoru Yuan, Ye Zhao, Scott Barlowe, and Shixia Liu. PIWI: visually exploring graphs based on their community structure. *IEEE Trans. Vis. Comput. Graph.*, 19(6):1034–1047, 2013.
- 10 Jaakko Peltonen and Ziyuan Lin. Information retrieval approach to meta-visualization. *Machine Learning*, 99(2):189–229, 2015.
- 11 Sana Malik, Fan Du, Megan Monroe, Eberechukwu Onukwugha, Catherine Plaisant, and Ben Shneiderman. Cohort comparison of event sequences with balanced integration of visual analytics and statistics. In *Proceedings of the 20th International Conference on Intelligent User Interfaces, IUI 2015, Atlanta, GA, USA, March 29 to April 01, 2015*, pages 38–49, 2015.
- 12 Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *VL*, pages 336–343, 1996.

- 13 Melanie Tory and Torsten Möller. Rethinking visualization: A high-level taxonomy. In *10th IEEE Symposium on Information Visualization (InfoVis 2004), 10-12 October 2004, Austin, TX, USA*, pages 151–158, 2004.
- 14 Christopher M. Bishop. A new framework for machine learning. In *Computational Intelligence: Research Frontiers, IEEE World Congress on Computational Intelligence, WCCI 2008, Hong Kong, China, June 1-6, 2008, Plenary/Invited Lectures*, pages 1–24, 2008.
- 15 Leanna House, Scotland Leman, and Chao Han. Bayesian visual analytics: Bava. *Statistical Analysis and Data Mining*, 8(1):1–13, 2015.
- 16 Catherine Plaisant. The challenge of information visualization evaluation. In *Proceedings of the Working Conference on Advanced Visual Interfaces, AVI'04*, pages 109–116, New York, NY, USA, 2004. ACM.
- 17 D. J. Lehmann, S. Hundt, and H. Theisel. A study on quality metrics vs. human perception: Can visual measures help us to filter visualizations of interest? *it – Information Technology*, 57, 2015 2015.
- 18 John Aldo Lee and Michel Verleysen. Scale-independent quality criteria for dimensionality reduction. *Pattern Recognition Letters*, 31(14):2248–2257, 2010.
- 19 John Aldo Lee and Michel Verleysen. Two key properties of dimensionality reduction methods. In *2014 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2014, Orlando, FL, USA, December 9-12, 2014*, pages 163–170, 2014.
- 20 Bradley Efron. *The Jackknife, the bootstrap and other resampling plans*. CBMS-NSF Reg. Conf. Ser. Appl. Math. SIAM, Philadelphia, PA, 1982. Lectures given at Bowling Green State Univ., June 1980.
- 21 Byron Ellis. *Real-Time Analytics: Techniques to Analyze and Visualize Streaming Data*. Wiley Publishing, 1st edition, 2014.
- 22 Lauro Lins, James T. Klosowski, and Carlos Scheidegger. Nanocubes for real-time exploration of spatiotemporal datasets. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2456–2465, 2013.
- 23 Zhirong Yang, Jaakko Peltonen, and Samuel Kaski. Scalable optimization of neighbor embedding for visualization. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 127–135, 2013.
- 24 Laurens van der Maaten. Accelerating t-sne using tree-based algorithms. *Journal of Machine Learning Research*, 15(1):3221–3245, 2014.
- 25 J. Ellson, E.R. Gansner, E. Koutsofios, S.C. North, and G. Woodhull. Graphviz and dynagraph – static and dynamic graph drawing tools. In M. Junger and P. Mutzel, editors, *Graph Drawing Software, Mathematics and Visualization*, pages 127–148. Springer-Verlag, Berlin/Heidelberg, 2004.
- 26 O.D. Lampe and H. Hauser. Interactive visualization of streaming data with kernel density estimation. In *Pacific Visualization Symposium (PacificVis), 2011 IEEE*, pages 171–178, March 2011.
- 27 Robi Polikar and Cesare Alippi. Guest editorial learning in nonstationary and evolving environments. *IEEE Trans. Neural Netw. Learning Syst.*, 25(1):9–11, 2014.
- 28 Joseph F. Murray, Gordon F. Hughes, and Dale Schuurmans. Machine learning methods for predicting failures in hard drives: A multiple-instance application. *Journal of Machine Learning research*, 6:816, 2005.
- 29 Ping Chen, Jing Yang, and Linyuan Li. Synthetic detection of change point and outliers in bilinear time series models. *Int. J. Systems Science*, 46(2):284–293, 2015.
- 30 Lei Le, Emilio Ferrara, and Alessandro Flammini. On predictability of rare events leveraging social media: a machine learning perspective. *CoRR*, abs/1502.05886, 2015.
- 31 Li Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Computer*

- Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on*, pages 178–178, June 2004.
- 32 Ivo Pedruzzi, Catherine Rivoire, Andrea H. Auchincloss, Elisabeth Coudert, Guillaume Keller, Edouard De Castro, Delphine Baratin, Béatrice A. Cuche, Lydie Bougueleret, Sylvain Poux, Nicole Redaschi, Ioannis Xenarios, and Alan Bridge. Hamap in 2013, new developments in the protein family classification and annotation system. *Nucleic Acids Research*, 41(Database-Issue):584–589, 2013.
 - 33 Dengsheng Zhang, Md. Monirul Islam, and Guojun Lu. A review on automatic image annotation techniques. *Pattern Recognition*, 45(1):346–362, 2012.
 - 34 Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *arXiv.org preprint*, arXiv:1412.3773 [cs.LG], December 2014. Submitted to Journal of Machine Learning Research.
 - 35 Tatsuya Tashiro, Shohei Shimizu, Aapo Hyvärinen, and Takashi Washio. Parcelingam: A causal ordering method robust against latent confounders. *Neural Computation*, 26(1):57–83, 2014.
 - 36 Hao Zhang, Maoyuan Sun, Danfeng (Daphne) Yao, and Chris North. Visualizing traffic causality for analyzing network anomalies. In *Proceedings of the 2015 ACM International Workshop on International Workshop on Security and Privacy Analytics, IWSPA'15*, pages 37–42, New York, NY, USA, 2015. ACM.
 - 37 Nivedita R. Kadaba, Student Member, Pourang P. Irani, and Jason Leboe. Visualizing causal semantics using animations. In *IEEE Transactions on Visualization and Computer Graphics*, 2007.
 - 38 Robert Kosara and Jock D. Mackinlay. Storytelling: The next step for visualization. *IEEE Computer*, 46(5):44–50, 2013.
 - 39 Jürgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990-2010). *IEEE T. Autonomous Mental Development*, 2(3):230–247, 2010.
 - 40 Thomas Hofmann and Justin Basilico. Collaborative machine learning. In *From Integrated Publication and Information Systems to Virtual Information and Knowledge Environments, Essays Dedicated to Erich J. Neuhold on the Occasion of His 65th Birthday*, pages 173–182, 2005.
 - 41 Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Trans. on Knowl. and Data Eng.*, 22(10):1345–1359, October 2010.
 - 42 Kristi Morton, Magdalena Balazinska, Dan Grossman, Robert Kosara, and Jock D. Mackinlay. Public data and visualizations: How are many eyes and tableau public used for collaborative analytics? *SIGMOD Record*, 43(2):17–22, 2014.

4.6 Reproducibility and interpretability

Helwig Hauser (University of Bergen, NO), Bongshin Lee (Microsoft Research – Redmond, US), Torsten Möller (Universität Wien, AT), Tamara Munzner (University of British Columbia – Vancouver, CA), Fernando Paulovich (University of Sao Paulo, BR), Frank-Michael Schleif (University of Birmingham, GB), and Michel Verleysen (Université Catholique de Louvain, BE)

License © Creative Commons BY 3.0 Unported license
© Helwig Hauser, Bongshin Lee, Torsten Möller, Tamara Munzner, Fernando Paulovich, Frank-Michael Schleif, and Michel Verleysen

Evaluating visualizations and visualization tools is a broad area and is at the heart of visualization research. Especially considering that a visualization requires a human to be understood and used, the focus has always been on how to evaluate the user experience. However, visualization research uses more and more sophisticated algorithms including some developed in the field of machine learning. Most of these algorithms have a stochastic nature, which makes that their result (or output) may depend on various settings, such as the small variations in the data, some random initialization or random step in an optimization procedure, etc. Therefore human evaluation of visualizations include various elements related on one side to the human nature of evaluations, and on the other side to the stochastic nature of the methods. The discussion in the group during the Dagstuhl seminar has concentrated on 1) how to distinguish these two aspects, and 2) what are really the different effects that have to be measured, in terms of robustness, generalizability, stability, etc.

Evaluation of visual data analysis tools can thus be viewed under a holistic perspective. Let us consider the the process of (visual) data analysis as a special type of algorithm. It takes inputs just like any other algorithm in form of data and/or parameters. Its output is some type of number or other complex entity (as is common for any algorithm). Sometimes this output will be some kind of decision made by the user, and hence it could be seen as a classification (into 0 or 1 or any other class of possible decisions). The only difference would be that while a traditional algorithm would simply be a structured sequence of computer code, the new holistic way of algorithms could include components that are determined by the so-called user-in-the-loop. In order to better distinguish this holistic view from the traditional view, we call these hal-gorithms.

This is akin to the Turing Test. The purpose of the Turing Test is simply to find out whether the algorithm one interacts is purely a machine or has components that can only be performed by a “real” human.

Under this holistic view of an algorithm it makes sense to ask on how to evaluate the quality of this hal-gorithms. With other words, we are considering the question on how different algorithmic performance test would extend to a scenario where the human is an integral part of the hal-gorithms. Again the evaluation of the quality necessitates to distinguish between performances (or differences of performances) that result from the algorithm itself, or from the user-in-the-loop supplementary layer.

The discussion during the Dagstuhl seminar has also covered terminology. Words such as robustness, stability, generalizability and sensitivity are sometimes used without having in mind a clear definition of their respective meaning and differences. Some can cover various situations too. The following is a first attempt to clarify both the terminology and its use in the holistic context.

4.6.1 Robustness of algorithms

The term *robustness* with respect to algorithms refers to the ability of an algorithm to gracefully handle any type of input. For instance the robustness of an algorithm with regards to outliers is of great concern. The concept of robustness is not far from the concept of stability (described below),

4.6.2 Robustness of hal-gorithms

Transferring the concept of robustness to hal-gorithms can have different meanings. For example if the target visual data analysis tool was created for a specific user group ('experts'), will it handled users that are not part of this group gracefully?

4.6.3 Generalizability of algorithms

The concept of *generalizability* of an algorithm is a contribution of the machine learning community. The idea is that train the algorithm (i.e. estimate optimal parameter settings) on a small subset of the known data. The performance of the algorithm is the evaluated on a hold-out set, which allows estimating how the algorithm would *generalize* to a greater set of possible (unknown) data. Estimating how an algorithm generalizes gives some indication on how to choose between several algorithms or settings.

4.6.4 Generalizability of hal-gorithms

The concept of generalizability is not new to the visualization community. The "User Performance" and "User Experience" evaluation methods speak exactly to aspects of understanding visual encoding principles by a larger set of users. However, there is a difficulty of properly testing relatively complex (visual analysis) tools. Often times there are too many confounding factors to consider. On the other hand, many tools are created for specific applications and particular experts. Having access to a larger number of these specific users is often not possible. Further, it is often not feasible to create multiple tools for different subsets of these users (the 'training' user set). Hence, during the design of a visual analysis tool (often referred to as a Design Study) the algorithm / tool is iterated upon and refined with a set of particular users one is working with. Hence, the generalizability of these tools is not tested.

4.6.5 Stability analysis of algorithms

Stability analysis is a term that often refers to the numerical stability of algorithms or discretization schemes. It is tied to the analysis of errors in the numerical computation. Hence, it is tied to the propagation of errors over several iterations. If the errors increase, the algorithm is numerically unstable. If the errors decrease, the algorithm is stable and often an analysis of the speed of convergence is followed. Even without 'errors' stability issues may be encountered due to the stochastic nature of data. On the other hand if 'errors' also refer to possible small variations in the data, the stability concept is not far from the robustness concept, and from the sensitivity one. Stability can also be related to the objective function: does the results of an algorithm change significantly if the objective function (the criterion that is optimized by the algorithm) is slightly modified?

4.6.6 Stability analysis of hal-gorithms

In cases where the generalizability of hal-gorithms can not be tested, perhaps a restricted view can be taken and a stability analysis can be performed. I.e. perhaps it can be well defined under what conditions and circumstances our hal-gorithms can be guaranteed to perform well. Further, one can ask whether several users working together come to an answer faster or to a better answer.

4.6.7 Sensitivity analysis of algorithms

Last but not least, *sensitivity analysis* is a branch of statistics that considers the change of outputs with respect to the inputs. Here, one distinguishes between global sensitivity analysis and local sensitivity analysis. Global sensitivity analysis is considering the possible change in outputs over all possible input variables by constraining just one input. On the other hand, local sensitivity analysis constraints all inputs to a specific value and analysis the change of output with respect to a small change in input of one of the inputs.

4.6.8 Sensitivity analysis of hal-gorithms

Sensitivity analysis is perhaps the most interesting and neglected aspect of hal-gorithms. How does the result change if the particular user using the visual analysis system changes?

Participants

- Daniel Archambault
Swansea University, GB
- Francois Blayo
Ipseite SA – Lausanne, CH
- Kerstin Bunte
UC Louvain-la-Neuve, BE
- Miguel Á. Carreira-Perpiñán
Univ. of California – Merced, US
- Ignacio Díaz Blanco
University of Oviedo, ES
- David S. Ebert
Purdue University – West
Lafayette, US
- Alex Endert
Georgia Inst. of Technology, US
- Thomas Ertl
Universität Stuttgart, DE
- Barbara Hammer
Universität Bielefeld, DE
- Helwig Hauser
University of Bergen, NO
- Stephen Ingram
University of British Columbia –
Vancouver, CA
- Samuel Kaski
Aalto University, FI
- Daniel A. Keim
Universität Konstanz, DE
- Bongshin Lee
Microsoft Res. – Redmond, US
- John A. Lee
UC Louvain-la-Neuve, BE
- Torsten Möller
Universität Wien, AT
- Bassam Mokbel
Universität Bielefeld, DE
- Tamara Munzner
University of British Columbia –
Vancouver, CA
- Ian Nabney
Aston Univ. – Birmingham, GB
- Stephen North
Infovisible – Oldwick, US
- Eli Parviainen
Aalto University, FI
- Fernando Paulovich
University of Sao Paulo, BR
- Jaakko Peltonen
Aalto University / University of
Tampere, FI
- William Ribarsky
University of North Carolina –
Charlotte, US
- Fabrice Rossi
University of Paris I, FR
- Frank-Michael Schleif
University of Birmingham, GB
- Michael Sedlmair
Universität Wien, AT
- Cagatay Turkyay
City University – London, GB
- Jarke J. van Wijk
TU Eindhoven, NL
- Michel Verleysen
University of Louvain, BE
- Thomas Villmann
Hochschule Mittweida, DE
- Daniel Weiskopf
Universität Stuttgart, DE
- William Wong
Middlesex University, GB
- Jing Yang
University of North Carolina –
Charlotte, US
- Leishi Zhang
Middlesex University, GB
- Blaz Zupan
University of Ljubljana, SI

