How do Recent Machine Learning Advances Impact the Data Visualization Research Agenda?

Timo Ropinski (Organizer)* Ulm University
Daniel Archambault† Swansea University
Min Chen‡ Oxford University
Ross Maciejewski§ Arizona State University
Klaus Mueller¶ Stony Brook University
Alexandru Telea‖ University of Groningen
Martin Wattenberg** Google Inc.

ABSTRACT

Nowadays, machine learning approaches have revolutionized many domains, by enabling machines to solve problems which could before solely be solved when involving humans. As this pushes the human out of the loop, the human-in-the-loop paradigm, which is one of the main pillars of data visualization research, might be endangered. Thus, we would like to investigate, which old visualization challenges are rendered obsolete, and which new visualization challenges arise from the recent advances in machine learning. Along these lines, we will - among other aspects - investigate the role of visualization when training networks, but also in how to make machine-made decisions more transparent to humans.

Index Terms: I.2.6 [Computing Methodologies]: Learning; I.3.8 [Computer Graphics]: Applications

1 INTRODUCTION

In recent years machine learning research has made tremendous advances, finally allowing computers to solve problems which before seemed to be unsolvable by computational processes alone. While the recent and widely acclaimed triumph of AlphaGo over Go champion Lee Sedol in March last year [17] was only the capstone on top of a period of successes, machine learning has helped to solve many computational science problems in areas such as medicine, biology, astronomy and meteorology - just to name a few. Many of these problems have before been considered only solvable by bringing the human into the loop, a paradigm often exploited by visualization researchers.

Within this panel we would like to investigate how the latest advances in machine learning affect the data visualization research agenda. Therefore, we would like to discuss both sides of the medal. On the one side, we will discuss the implications of the machine learning advances with respect to visualization researchers' exploitation of the human-in-the-loop paradigm. The key question in this context would be: Could these advances ultimately render visualization research obsolete? On the other side, we would like to investigate which new opportunities the shift towards learning systems brings to the visualization community. With the increasing importance of learning systems, it becomes mandatory to train systems effectively, while at the same time it is required to understand the functionality of a trained system, in order to enable predictions. First visualization approaches for improving training [13], and understanding convolutional neural networks [18, 16, 7, 19] have already been proposed. But also data visualizations guided by machine learning [10], and combinations of data visualization and machine learning [12] have been investigated. However, the key question in this context remains: How can visualization be used to support training and to communicate a neuronal network's behavior effectively? Especially in a world, where the power of algorithms is increasing in a variety of areas, such as medical diagnosis, insurance policies, or financial rating, it is essential to be able to understand the ratio behind these machine-made decisions.

These rather new applications of visualization might have the potential to define a new visualization research agenda, which deeply roots visualization in a technical world, which is currently transiting from programmable to learning systems. In such a world, visualization might be essential to support the training process of these learning systems, but also to understand their behavior. Thus, the user's main purpose might shift from making decisions towards analyzing decisions, which have been made by machines. With such a shift, the human-in-the-loop paradigm and the role of visualization must be reevaluated. Thus, while visualization is still often quoted as a key technology enabling our information-based society to cope with the big data challenge arising from inexpensive data acquisition and storage, this panel will investigate if this statement still holds. While a 2012 MIT sloan report even rated data visualization as the most valuable technique to deal with this large data challenge [9], the question is how would they rate the role of visualization today?

These and other questions shall be addressed within this panel, where we would like to reevaluate the role of data visualization in a world where the need for the human-in-the-loop paradigm seems to vanish, or might at least shift. The goal of such a reevaluation is to (re)position the area of data visualization, such that we as a research community can still justify the meaningfulness of our work, and can develop a robust research agenda which can stand independent of the advances the machine learning community might make in the next couple of years.

2 PANELIST STATEMENTS

As preparation for the planned panel, each panelist has been asked to write a short position statement, which is included in this outline together with the panelists' biographies. The position statement should reflect on the impact of recent machine learning advances on the visualization research agenda, and ideally also be addressing the following questions:

- How far do you see the application of the human-in-the-loop paradigm within data visualization research threatened by the recent machine learning advances?
- Are the application areas of data visualization shrinking due to recent advances in machine learning?
- How can visualization researchers capitalize from the shift from programmable systems to learning systems?
- How do fundamental visualization concepts change when enabling the human to judge machine-made decisions, rather than making man-made decisions?
- When perceptual problems are solvable, what does this mean for more cognitive problems, where the user has to do some planning? Do we need to distinguish between recognition and planning problems?

*e-mail: timo.ropinski@uni-ulm.de
†e-mail: d.w.archambault@swansea.ac.uk
‡e-mail: min.chen@oerc.ox.ac.uk
§e-mail: rmacieje@asu.edu
¶e-mail: mueller@cs.stonybrook.edu
‖e-mail: a.c.telea@rug.nl
**e-mail: wattenberg@google.com

- Where do you see new synergies emerging in visualization and machine learning?

Daniel Archambault - *Towards a Tighter Integration of Machine Learning and Information Visualisation*

Position statement. Machine learning and information visualisation have similar goals but have different means to achieve them. One common goal is to find interesting features in the data. Information visualisation researchers try to find ways of representing the information visually, leveraging user creativity and human expertise to make these discoveries. Machine learning research creates automatic algorithms to detect these features automatically.

Attacking this problem from a machine learning perspective has the advantage of scalability. Machine learning approaches are extremely scalable, and can process far more raw data than a human. However, computers are not all that creative and the features automatically found in the data are constrained by the selection of machine learning method and its application to the data. As information visualisation is able to use user expertise, our methods can get around a number of these limitations through soft knowledge but is bound by the technique and the visual processing and memory abilities of the human. In many ways, the techniques and algorithms developed by our fields have complementary benefits and limitations. We can use machine learning to process large volumes of data and then use custom visual interfaces to analyse the results of the machine learning. This perspective would be a *visualisation as output* way of thinking.

In my experience working with researchers in the machine learning community, the easiest way for us to work together is through visualisation as output. There is nothing wrong with a visualisation as output strategy – actually there are truly many benefits in taking on this perspective. Machine learning algorithms often produce large volumes of output and it is often a challenge to make sense of it. However, probably the more challenging problem is a tighter integration of machine learning techniques where information visualisation methods are used during the mining process.

Accomplishing this goal is extremely difficult. For many machine learning techniques, it is often really hard to understand how they achieve their highly generalisable results. A tighter integration of machine learning and information visualisation would require a way to understand the inner workings of the machine learning algorithms in order to interact with them. However, progress in this direction would be extremely beneficial. Simple immediate benefits in terms of execution time as less relevant data can be undersampled. More interesting and wide-reaching benefits would be we could maybe begin to understand how the machine learning approach works and potentially where missclassifications could occur.
Short biography. Daniel Archambault is a Senior Lecturer at Swansea University in the United Kingdom. His introduction to bridging the divide between machine learning and visualisation began during his post-doctoral studies at University College Dublin. He was one of two information visualisation scientists in a lab full of data mining researchers. During this time, he primarily worked on problems involving graph and text mining and how information visualisation techniques could be applied in these areas. This led him to run two editions of the SocMedVis workshop at AAAI ICSWM which explored the integration of visualisation techniques in social media analysis. He still contributes to the AAAI ICWSM conference and was on of the Senior IPC members of the conference for 2017. In the visualisation community, he organised the machine learning methods tutorial at EuroVis 2016 and 2017 with Ian Nabney and Jaakko Peltonen. He is interested in exploring the integration of graph mining with information visualisation methods and the integration of machine learning methods with information visualisation in general.

Min Chen - *The Space of Machine Learning*

Position statement. Consider all possible programs that we can create or use. Whether we like it or not, some of these programs are being created using *machine learning* (ML). This is not because ML can develop better algorithms or more reliable systems than trained

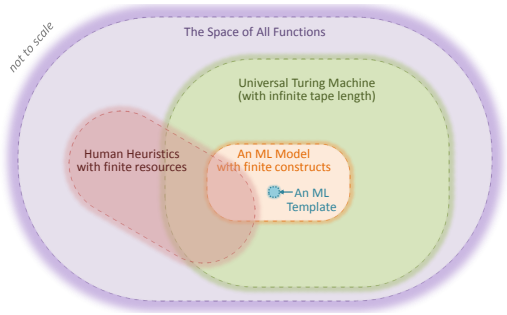

Figure 1: The space of functions, and its subspaces.

computer scientists and software engineers. ML is merely a computational tool that helps us write an approximate software function, for which we do not quite know the exact algorithm. Today this tool is becoming more and more powerful and useful because of the increasing availability of training data and high performance computing for optimization. With such a tool, we can explore new areas in the *space of functions*, which is programmable on a stored program computer and where conventional algorithms are not yet found or effective.

In theoretical computer science, a programming language is said to be *Turing complete* if it can describe any single-taped Turing machine. We can consider an underlying model used by ML (e.g., neural networks, decision trees, Bayesian networks, support vector machines, genetic algorithms, etc.) as a programming language. In such a language, there is a very limited set of constructs (e.g., node, edge, weight). For any practical ML applications, humans usually do the clever part of the programming by defining a *template*, such as determining the candidature variables for a decision tree, the order of nodes in a neural network, or the possible connectivity of a Bayesian network. The ML tool then does the tedious and repetitive part of the programming for fine-tuning the template using training data to make a function. It is known that most commonly used ML models, such as forward neural networks, decision trees, random forests, and static Bayesian networks are not Turing complete. Some others are Turing complete (e.g., recurrent neural networks and dynamic Bayesian networks), but their demand for training data is usually exorbitant. For any ML model, once a template is determined by the humans, the *parameter space* that the ML tool can explore is certainly not Turing complete as illustrated in Fig. 1.

Although the parameter space that an ML tool can explore is tiny in comparison with the space normally available to a human programmer, the ML can meticulously comb every bit of the space. The number of "locations" that an ML tool can visit is exponentially related to the number of constructs used in the template designed by the human programmer. For example, there are 2^n candidature functions for a given neural network template with a maximum of n edges, and there are k^n functions for a given static Bayesian network template that has n edges and each edge may take any of the k values $\in [0,1]$. The number of these "locations" has a direct impact on the amount of training data required, and the amount of computation time for training.

The above theoretical discourse suggests that humans have many significant roles to play in ML, and visualization can enable humans to play such roles more effectively and efficiently:

- Most complex data intelligence workflows require some functions with known algorithms, some without; some theoretically within the spaces of ML models, and some outside; some practically within the reach of ML, and some beyond (e.g., lack of training data, rapid context change). We should always take a multifaceted approach. (i) When a required function has known algorithms, we should always use an algorithmic function written by humans. ML is only an intermedi-

ate solution, which is not a replacement for gaining better understanding and discovering new algorithms. (ii) When a required function does not have known algorithms, and is within the theoretical space and practical reach of an ML model, we should attempt an ML solution. Because such a function is approximated, humans must always in the loop in its development and deployment. (iii) When a required function does not have known algorithms, and is beyond the theoretical space and practical reach of any ML model, it should be performed directly by humans with the aid of statistical analysis and interactive visualization. This is in fact the essence of *visual analytics* [15].

- When an ML tool has some difficulties to explore a sufficient number of "locations" in its parameter space defined by a template (e.g., because of a sparse training dataset), humans can use their soft knowledge to assist the ML tool with the aid of model-developmental visualization (e.g., [14]).
- When a software engineer develops an algorithm or a system, there are many different forms of quality control. Visualization can assist humans to control the quality of any software function created using ML throughout the lifecycle of an ML function. For example, visualization can help observe the training process, depict evolution of the connectivity and parameters within a template, analyze sensitivity to training data, testing data and real-world data, and monitor its performance after its deployment and in different contexts (e.g., [1]).
- It is also necessary to remind us that ML is particularly useful for approximating functions for which we do not have an exact algorithm. In fact, most perceptual and cognitive functions that we use during visualization fall into such a category. Hence, we can potentially use ML tools to create such functions for simulating human behaviors during visualization. This will no doubt contribute significantly to the development of theoretical models for visualization [3]).

Short biography. Min Chen developed his academic career in Wales between 1984 and 2011. He is currently the professor of scientific visualization at Oxford University and a fellow of Pembroke College. His research interests include visualization, computer graphics and human-computer interaction. He has co-authored some 200 publications, including his contributions in areas of volume graphics, video visualization, face modelling, automated visualization and theory of visualization. He has been awarded over 11M research grants from EPSRC, JISC (AHRC), TSB (NERC), Royal Academy, Welsh Assembly Government, HEFCW, Industry, and several UK and US Government Agencies. He is currently leading visualization activities at Oxford e-Research Centre, working on a broad spectrum of interdisciplinary research topics, ranging from the sciences to sports, and from digital humanities to cybersecurity. His services to the research community include papers co-chair of IEEE Visualization 2007 and 2008, Eurographics 2011, IEEE VAST 2014 and 2015; co-chair of Volume Graphics 1999 and 2006, EuroVis 2014; associate editor-in-chief of IEEE Transactions on Visualization and Computer Graphics, editor-in-chief of Wiley Computer Graphics Forum, and co-director of Wales Research Institute of Visual Computing. He is a fellow of British Computer Society, European Computer Graphics Association, and Learned Society of Wales. URL: https://sites.google.com/site/drminchen/

Ross Maciejewski - *Does Visual Analytics Contribute to Algorithm Aversion?*

Position statement. As the amount of data available for analysis has increased, leaps in machine learning and data mining techniques have occurred, enabling large-scale modeling of all sorts of phenomena. Such modeling is often performed offline in a relatively black-box manner where results are presented to be used (or ignored) by the domain experts. Here, the visual analytics community postulates that the integration of domain knowledge into an interactive sense-making loop will improve modeling results from machine learning claiming that experts have some inherent knowledge that cannot be easily encapsulated by the machine learning. Anecdotal evidence from the visualization community has suggested that the direct integration of domain knowledge does improve the overall model efficacy. However, research from the management science community has found mixed results of human-in-the-loop with evidence indicating that in forecasting tasks, machine predictions consistently outperform human forecasters [4, 5, 8]. In fact, work by Akes, Dawes, and Christensen [2] found that domain expertise diminished people's reliance on algorithmic forecasts, creating an algorithmic aversion [6] which leads to a worse performance in prediction tasks. Specifically, humans quickly lose confidence in algorithmic forecasts after seeing algorithmic mistakes. However, one underlying premise of visualization is that users can explore model results, adjust for errors, and improve the model overall. This means that visualization is intentionally showing users the algorithmic mistakes under the guise of domain knowledge injection and explainability. As such, the visualization of machine learning results may be directly at odds with convincing users to adopt a given algorithm, and visualization could potentially contribute to algorithmic aversion during forecasting tasks and lead to reduced performance. Furthermore, humans come with a large variety of biases, in particular, confirmation bias- seeking out information to confirm decisions, overconfidence bias- being too confident in abilities which leads to taking risks, and anchoring- overreliance on the first piece of information found, are specific human biases that are known to affect decision-making. Thus, by giving users the option to integrate their domain knowledge, we are also enabling them to inject bias into the model. What is the point of using technology to learn something new when you are bending it to fit your pre-existing notions? Visual analytics focuses on detect the expected, and discovering the unexpected, but what if the unexpected fails to fit a users preconceived data view? Does visualization encourage users to reconsider, or encourage algorithm aversion? As such, how much (if any) human should be included as part of machine learning?

Short biography. Ross Maciejewski is an Associate Professor at Arizona State University in the School of Computing, Informatics Decision Systems Engineering. His primary research interests are in the areas of geographical visualization, predictive analysis, and visual analytics focusing on public health, dietary analysis, social media, criminal incident reports, and the food-energy-water nexus. He is a recipient of an NSF CAREER Award (2014) and was recently named a Fulton Faculty Exemplar and Global Security Fellow at Arizona State University.

Klaus Mueller - *The Case for Intelligent Visual Analytics that builds Trust*

Position statement. In my graduate visualization course this year I began the first midterm exam with the following somewhat provocative question: With all the advances that are made in machine learning these days, is interactive visualization needed at all for data analysis? Cant we just automate it? Here I was reminded of a similar question by the professor teaching my first AI course in the 1990s which stated Can computers think? I put this question in the midterm so the graduate students 134 of them would start to think about this issue which I believe is quite existential to our research domain and the jobs the students take after graduation. My AI professor two decades ago probably thought similarly.

Here is what my answer key for this question was: Visual analytics enables human intelligence, commonsense, imagination, creativity, intuition, and domain knowledge (1 point for each for a max of 4 points) to play a part in the data analysis and decision making, and as a consequence get better results (2 points), break ties (2 points), and be more confident (2 points). Did all students max out for 10 points? Not really, but they did much better in the final exam when I ran the question again.

Clearly, there will be many people that will just trust a machine-derived result. It will work when an approximate result is good enough. Its a matter of trust. But there are even very mundane situations where trust is challenged. For example, I might stand on

a scale every morning to measure my weight. When my weight keeps going down I wonder if the scale is right. When my weight keeps going up, I wonder the same. I want to know how the scale arrived at the result it gave me.

This example might have been metaphorical but trust is a key issue, and its importance scales with the stakes at hand legal, finance, business, medicine, and so on. Likewise, true innovation is made when the innovator goes beyond the capabilities of the available tools. Enabling innovation, whatever the scale, with intelligent visual analytics tools is a definite goal.

Short biography. Klaus Mueller received a PhD in computer science from the Ohio State University. He is currently a professor in the Computer Science Department at Stony Brook University and is also an adjunct scientist in the Computational Science Initiative at Brookhaven National Labs. His current research interests are visualization, visual analytics, data science, medical imaging, and high-performance computing, He won the US National Science Foundation CAREER award in 2001 and the SUNY Chancellor Award in 2011. Mueller has authored more than 170 peer-reviewed journal and conference papers, which have been cited more than 7,500 times. He is a frequent speaker at international conferences, has participated in numerous tutorials on various topics, and was until recently the chair of the IEEE Technical Committee on Visualization and Computer Graphics. He serves as an Associate Editor-in-Chief of IEEE Transactions on Visualization and Computer Graphics and is a senior member of the IEEE. For more information, please see http://www.cs.sunysb.edu/ mueller

Alexandru Telea - *Harnessing the Power of Machine Learning needs Visualization*

Position statement. Machine learning techniques and tools have become dramatically more mature, scalable, available, precise, and generically applicable to a wide range of problems in the last few years. As such, various tasks such as data segmentation, classification, pattern recognition, model inference, and pattern synthesis have become much easier to address in the context of real-world applications. Many of these tasks relied on human input, either in terms of parameter tuning or in terms of visual annotation or exploration of the involved data in the past. Machine learning has largely automated many such tasks, reducing the users burden. In this context, several voices have questioned the future need of a human-in-the loop approach and, thus, of the value of data visualization.

Contrary to these views, we argue that visualization has become only more useful in the context of machine learning. To be efficient and effective, machine learning requires a far more complex set-up than classical data analysis techniques: Examples are architecture and design of a (deep) neural network; training and parameterization of the learning; detailed examination of the test and validation results; and repeating the loop to increase the accuracy, robustness, and overall quality of the modeled process.

The power of machine learning techniques comes with a high price, expressed in terms of the complexity of design, operation, and fine-tuning of the used techniques. The most successful methods in the area (random forest classifiers, support vector machines, and above all deep neural networks) are also the hardest to understand, even for specialists. For typical users, they operate pretty much as ‘black magic or black boxes. When a machine learning system is not performing as desired, the question of how to ‘fix the system is largely an open one. Several voices from the machine learning community have recently expressed the strong need for means to open up such black boxes to comprehend their operation, understand where and why they fail, and iteratively guide users in an intuitive, efficient, and effective way towards achieving their goals.

We argue that the human-in-the-loop and his (visual) tools of trade are not dissappearing, but are just being shifted on a higher level: Rather than understand raw data and simple processes, we now have to understand how a complex (learning) system (mis)interprets complex data; how we can guide a system to understand our view on the data; and when and why does our view on the data and its underlying semantics start diverging from what the machine ‘thinks. We see a strong similarity between these questions with those asked in classical software engineering (expression of requirements, program comprehension, software debugging). As such, we argue that visualization techniques for large, high-dimensional, relational, uncertain, hybrid, and time-dependent data, well proven and developed in software visualization, can be leveraged and extended to support the user in the design-training-testing loop in machine learning.

Short biography. Alexandru Telea received his PhD (2000) in Computer Science from the Eindhoven University of Technology, the Netherlands. Until 2007, he was assistant professor in visualization and computer graphics at the same university. Since 2007, he is professor of computer science at the University of Groningen, the Netherlands. His interests include multiscale visual analytics, software and graph visualization, and methods at the crossroads of scientific and information visualization. He has published over 200 internationally peer-reviewed papers in this fields, and is the author of the textbook Data Visualization: Principles and Practice (CRC Press, 2008/2014).

Martin Wattenberg - *Machine Learning & Visualization*

Position statement. At first glance, machine learning and visualization might seem like opposites: in one case, the computer makes decisions automatically; in the other, humans are given the power to analyze data and decide themselves. However, these two fields have many close connections.

One key link is the issue of interpretability. When a machine learning system takes and action, it is often important to understand its reasons, if the machine can in fact be said to have reasons at all. Unfortunately, some of the most successful recent machine learning models have millions of parameters, with no clear modular organization. The challenge of making sense of model decisions is critical–and a natural place where visualization can help. Fortunately, there is some hope: in some recent intriguing work on systems for vision and for sequence processing, we have seen how visualization can sometimes act like a kind of microscope for meaning, finding interpretations even in individual model parameters. An important next step will be to create visualizations that highlight meaningful ensembles of parameters, as a way of extracting additional structure from model decisions.

Visualizations of how models work have so far been aimed at researchers and engineers, but a critical extension is to find ways of explaining model decisions that lay users can understand. In many cases, a simple and comprehensible summary of the "why" of a decision may greatly improve the user experience of an AI system. It seems likely that visualization can play a key role in providing such summaries, making this a promising area for research.

So far, we have discussed how visualization can help machine learning. But can machine learning help visualization? In particular, it is natural to ask whether machine learning can be used to create better visualization systems, perhaps automatically choosing the best visual marks and encodings. A key question in this area is to define an evaluation function: how do we tell the machine what makes a "good" visualization? This is a subtle area with many pitfalls. For example, if we teach machines to imitate what humans we do, we may find they inherit bad habits (over-recommending pie charts, to name one example). On the other hand, it is also possible that with the right perceptual evaluation function, machines may even create effective visualizations no human has yet considered.

Short biography. Martin Wattenberg is a computer scientist and artist. He is a co-leader, with Fernanda Vigas, of Google's "Big Picture" data visualization group, part of the Google Brain team. Before joining Google, he and Vigas led IBMs Visual Communication Lab, where they created the ground-breaking public visualization platform Many Eyes. Wattenberg is known for his visualization-based artwork, which has been exhibited in venues worldwide. He holds a Ph.D. in mathematics from U.C. Berkeley.