

Department: Visualization Viewpoints

Editor: Theresa-Marie Rhyne, theresamarierhyne@gmail.com

Challenges in Evaluating Interactive Visual Machine Learning Systems

N. Boukhelifa

INRAE, Université Paris-Saclay

A. Bezerianos

Université Paris-Saclay, CNRS, Inria

R. Chang

Tufts University

C. Collins

Ontario Tech University

S. Drucker

Microsoft Research

A. Endert

Georgia Tech

J. Hullman

Northwestern University

C. North

Virginia Tech

M. Sedlmair

University of Stuttgart

Abstract—In interactive visual machine learning (IVML), humans and machine learning algorithms collaborate to achieve tasks mediated by interactive visual interfaces. This human-in-the-loop approach to machine learning brings forth not only numerous intelligibility, trust, and usability issues, but also many open questions with respect to the evaluation of the IVML system, both as separate components, and as a holistic entity that includes both human and machine intelligence. This article describes the challenges and research gaps identified in an IEEE VIS workshop on the evaluation of IVML systems.

■ **RECENT ADVANCES** in machine learning saw the rise of powerful automatic methods to build

robust predictive models from data. In an attempt to enhance understanding and improve performance, researchers have pursued human-centered approaches. For instance, in interactive visual machine learning (IVML), a human operator and a machine collaborate to achieve a task (e.g., to classify points using GAN, to cluster

Digital Object Identifier 10.1109/MCG.2020.3017064

Date of current version 23 October 2020.

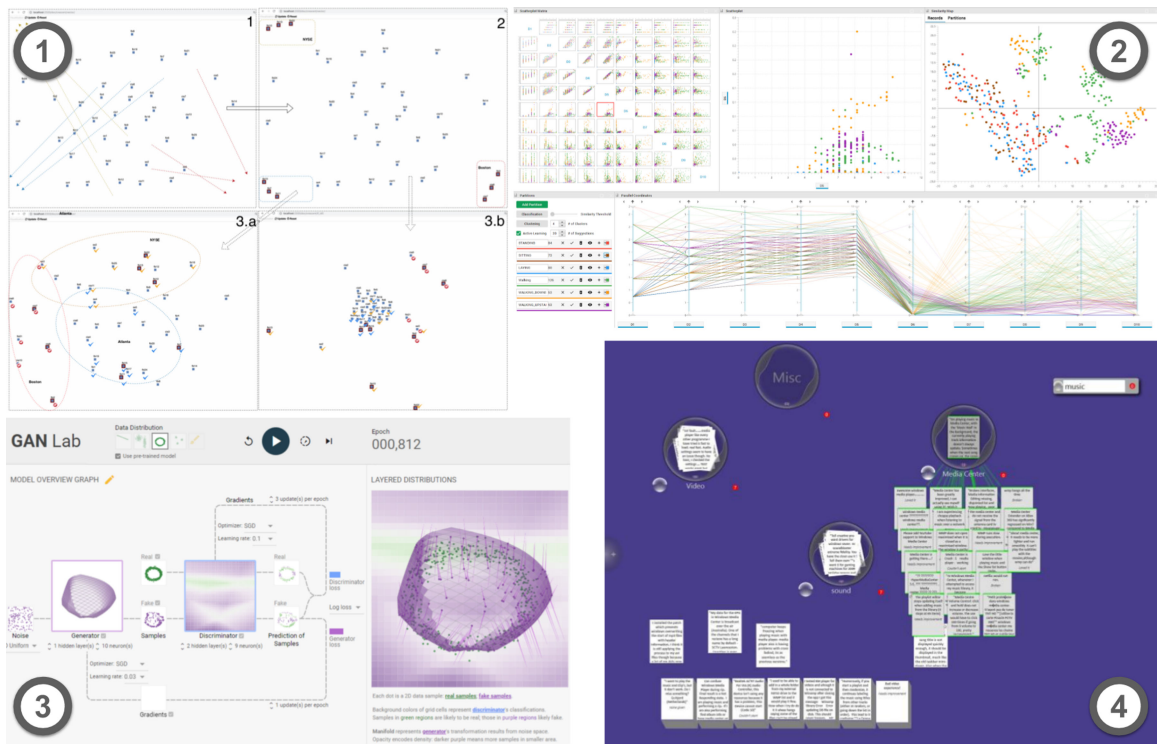


Figure 1. Examples of IVML systems where the analyst interacts with the visualization, and the system updates the machine learning model and adapts the visualization accordingly. (1) Semantic Interaction⁶ enables analysts to modify a suggested projection of documents (top) that updates the underlying weighted distance of two alternative projection models (bottom). (2) mVis¹¹ helps the analyst interactively partition an unlabeled dataset or explore and verify the partitions of a labeled one. The machine learning labeling model gets updated accordingly. (3) GAN Lab²⁴ helps analysts learn about GAN models by visualizing their structure and enabling users to interactively train and experiment with the models. (4) iCluster¹⁴ helps users sort and cluster large numbers of documents. The user adds items to clusters, and the system learns from those interactions and recommends new items and clusters.

them using DBSCAN, or to learn a mathematical fitness or objective function), mediated by an interactive visual interface (e.g., the papers by Bertini *et al.*,³ Boukhelifa *et al.*,⁴ Brown *et al.*,⁷ and El-Assady *et al.*¹⁵).

Typically, an IVML system comprises an automated service, a user interface, and a learning component. Often the goal of IVML (examples in Figure 1) is to help the human and the machine intelligence work together more efficiently and effectively than they would individually.

The interactive and visual approach to machine learning (ML) is appealing for many reasons. For example, it allows for the integration of valuable expert knowledge, guidance, and model steering.¹² It can also aid analysts in reasoning about uncertainty in machine learning output and makes output

more interpretable and likely to increase user trust.

Recent work in IVML has focused more on developing working prototypes, and less on methods to evaluate IVML systems and their various components. Most of the work that exists in the evaluation arena focuses on explainable machine learning (e.g., the paper by Lage *et al.*²⁸). This body of work aims to test whether users appropriately trust a machine learning algorithm when it is more likely to be correct than they are, and whether they understand what the artificial intelligence (AI) is “learning” (while it is being trained) or “thinking” (when it makes or suggests decisions).

Human-in-the-loop approaches to machine learning extend the role of the human beyond interpreting and understanding the underlying models or decisions. Humans can also react to

or act directly on these models, for example by accepting suggestions or altering model parameters. This brings forth not only numerous intelligibility and usability issues, but also many open questions with respect to the evaluation of the various facets of the IVML system, both as separate components, and as a holistic entity that includes both human and machine intelligence.^{5,31} For example, IVML tools need to be assessed more generally on their ability to increase task efficiency and correctness, as well as other possible metrics.

We believe that the evaluation of IVML systems is harder than the evaluation of their individual components in isolation (i.e., the automated service, the visualization and user interface, and the learning component). In what follows, we describe four important challenges to consider when evaluating IVML systems, identified in an IEEE VIS 2019 workshop on the evaluation of interactive visual machine learning systems (EVIVA-ML, <https://eviva-ml.github.io/>). The workshop brought together visualization researchers and practitioners to discuss experiences and viewpoints on how to effectively evaluate IVML systems. We first transcribed the workshop, including the keynote, paper presentations, and panel discussion. Then two co-authors performed open coding to identify distinct topics discussed in the workshop, and to group them in bigger themes (following grounded theory¹⁰).

Four major challenges and associated research opportunities emerged from our analysis. We highlight challenges in (1) identifying the human and AI *roles* within an IVML system partnership, such that it is clear what each contributes to the analysis; (2) defining the success criteria of the partnership, taking into account multiple possible *tradeoffs*; (3) assessing the effects of different sources of *uncertainty* on the use of IVML systems; and (4) providing practical evaluation *guidelines and metrics* for IVML systems.

IDENTIFICATION OF ROLES WITHIN THE PARTNERSHIP

IVML systems are particularly complex because they integrate multiple components that are themselves complex, such as large datasets and potentially uncertain ground truth, probabilistic and black-box machine learning models, function-rich interfaces and visualizations, and last but not

least, human analysts that may have biases, or imprecise or hard to express goals. There is also the complex interplay and tight coupling between those components. Take, for example, the coadaptation phenomenon between the user and the machine learning component.²⁹ As the user sees suggestions from the system, for example, when working with clustering, they may adjust or refine their original clustering criteria influenced by the system recommendations. Or if they use the system over time, their analysis goals may evolve as their understanding of the data improves. Thus, not only does the ML component learn and evolve, but so does the human operating it. Evaluating such complex dynamic systems can be challenging, as the desired outcome of the IVML process, from the perspective of the human, may continuously evolve and both partners (humans and ML) will adapt to it.

We consider next the methodologies currently adopted. On the one hand holistic evaluations that take an IVML system as an integrated entity suffer from the *attribution problem*. Observed results are loosely attributed to the system as a whole, but without accurate explanations or insights as to what component or components played a bigger role to achieve those results. For example, was there an improvement in the results (compared to not using an IVML) due to a more robust machine learning algorithm, to users' expertise and their pertinent feedback to the AI, or to the iterative tuning of results going back and forth between the human and the AI? Such integrated evaluations are perhaps easier to run, but they struggle to tease out where success or failure occurs, since IVML elements are intertwined.

On the other hand, reductionist evaluation approaches break down the IVML into multiple components and study different variants of the system. This type of evaluation requires many considerations including how to: (a) break down the system; (b) identify IVML configurations to compare (accounting for their unique properties and potential interaction, e.g., VIS alone versus ML alone versus VIS+ML); and (c) decide what tasks and evaluation method to use to test the different combinations (e.g., quantitative, qualitative, insight, or simulation-based).³⁴

The coupling of elements and their inherent coevolution makes it hard to isolate any one of

those components at one time (more so than other non-ML visual analytics systems), leading to complicated study designs. Even when a viable slicing and dicing of the IVML is identified, it may only be appropriate in specific study designs. For example, isolating interaction from the ML may be appropriate for simulation-based studies, but could create potential study confounds if the number of insights is selected as an evaluation metric. Since it is often the continuous dialogue between the human and the AI (and their coevolution) that can lead to insights, removing or reducing interaction can hamper insight generation.

Multiple kinds of evaluations are adopted to evaluate a single IVML system, combining user-centered and algorithm-centered evaluations (e.g., insight evaluation and algorithmic convergence tests⁴). Apart from being time-consuming and complicated to run, those studies have the additional challenge of having to “stitch” the results back together to a set of unified and meaningful results, that can inform future research and provide insights useful to the community that goes beyond the usefulness of the specific IVML system.

Research Opportunities

Moving forward, we need to identify the role of each component in IVML and create a taxonomy of the different types of partnerships and how to evaluate them. These roles can be identified based on, for example, a level of abstraction, such as the high-level and low-level roles proposed in the paper by Endert *et al.*¹⁶ Here, the analyst is focused on the high-level ideas and the big picture, and the AI algorithm is learning and filling in the details.

To evaluate the different types of roles and partnerships, we need methods that are able to tell us what the human versus what the ML contributed to the analysis. Depending on the target use case, it may make sense to decouple and compare components (e.g., run a head-to-head study between human + AI against AI, against human). These roles can also guide us in identifying appropriate metrics for evaluating each of the different IVML components in our studies. Given these observations, IVML system designers need to consider designing IVML systems

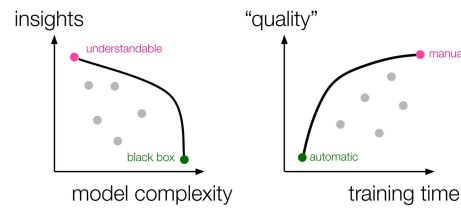


Figure 2. IVML evaluations can be modeled as a set of optimal evaluation studies along a Pareto front.²⁷ Multiple Pareto fronts can exist when it comes to the evaluation goal, such as between (Left) measuring insights and model complexity, where simplified models may be easier to understand, while complex black-box models, may fit the data better, but can be hard to understand; or between (Right) measuring training time and model quality, where automatic solutions optimize time at the expense of quality, and manual solutions improve quality but increase time.

where the different components can be more easily isolated.

MANY TRADEOFF CONSIDERATIONS

An important consideration when evaluating any interactive system is to define success criteria. That is to say, how do we know that the system in question helps people achieve their goals? For example, are we trying to build systems that help people get insights? Or is our goal to hit the *export model* button at the end and get a really good model?

For IVML, multiple conflicting success criteria may be in play, such as accuracy, complexity, and interpretability (see Figure 2). While a sophisticated AI model can capture some relationships in the data accurately and can find unexpected groupings, the model and results may be too complex for a human to understand, reducing trust. A more human-readable and perhaps simpler model, on the other hand, may not be able to capture important relationships in the data and can yield poorer results despite greater interpretability.

Tradeoff considerations for IVML evaluation are not limited to those three dimensions (i.e., accuracy, complexity, and interpretability) that can be found in any visual analysis system that includes an automated component. Additional factors can be in conflict in IVML. For instance,

an interpretable IVML may be aimed at optimizing for analysts' insight.²⁰ Analysts in such settings might end up abusing the model (by extensively experimenting with it, affecting the learning process, or by actively simplifying it to increase understandability), making it worse in terms of performance accuracy. For example, when using an interactive evolutionary algorithm to iteratively model biological processes in an IVML system (essentially building mathematical functions), agronomists evolved both functions that had high fitness (but also high mathematical complexity), but also ones that sacrificed accuracy for the sake of simpler and more interpretable models.⁹ Or if we want to optimize time, we automate more tasks (possibly sacrificing accuracy along the way) and reduce the amount of interaction in the machine learning process. If insight is a process, then by automating away pieces of the "reasoning," we may also be reducing the total or deeper insight.

Research Opportunities

We can think of the IVML evaluation *as a set of optimal evaluation studies* along a Pareto front, which represents a diverse set of compromise points between conflicting objectives.²⁷ The dimensions of this Pareto front can be decided together with experts in the application domain, visualization, and machine learning. Such dimensions could include: time of executing the task, quality of the model, quality of insights, difficulty level, and cognitive load, as well as the cost of interaction.³⁴ We need to parameterize our evaluation space and identify the objectives we want to optimize. Some of those objectives may be subjective in nature, which means we may need to involve the end user in setting those parameters.

Furthermore, when it comes to IVML evaluation, important metrics should not be evaluated independently, say between accuracy and interpretability. Recent work shows, for example, that a more accurate AI can lead to worse decisions by a human-AI pair when the newly updated model conflicts with the human's expectations about how the model works.² Rather, a range of possible optimal IVML configurations (i.e., trade-offs) need to be considered and compared, much like in a Pareto front optimization. For example, a

non-optimal solution according to say model accuracy and time might turn out to be the best solution, because it leads to more data insights and understanding, or simply because it is preferred.

MULTIPLE SOURCES OF UNCERTAINTY

The evaluation of IVML systems needs to account for multiple sources of uncertainty; pertaining to the predictions of the IVML system itself as well as the context in which IVML is operated. The first source of uncertainty concerns the fact that ML predictions are typically probabilistic, producing outputs that can be accessed via various error typologies such as signal detection methods capturing false positives and false negatives.¹ The degree to which the model architecture, or specification, truly captures the generating process simulated by the ML contributes persistent, unquantifiable uncertainty. This is a concern in any system that includes a ML component.

Beyond these sources, in IVML, the human component contributes further uncertainty into how human+ML decisions will be made. The analysts trying to build intuitions (e.g., mental models) and reason with the algorithms may themselves be inconsistent in how they respond to information such as model confidence in a classification. In this context, the evaluation is not simply about assessing how accurate the analysts' reading of a dataset is, but also about how effectively the IVML system supports reasoning and decision making under uncertainty. While analysts might properly interpret a visualization of predictions, they may not necessarily make optimal decisions¹⁸ or update their beliefs about the outcome in ways that align with statistical mechanisms.²⁶ They may, thus, provide inappropriate feedback to the ML component (resulting in poor training).

Evaluating reactions to uncertain predictions is a challenging task given that people's intuitions about statistical processes are often biased but may appear correct under certain evaluation approaches.²² Evaluations of human-in-the-loop approaches may not capture different types of cognitive biases, including priming and anchoring effects,^{13,30,33} which may appear in some contexts but not others. Even without

uncertainty, the same person presented with the same visualization, say a scatterplot, might reach a different conclusion based on what they have seen before.³²

Another evaluation challenge is accounting for uncertainty due to circumstantiality and variability in the IVML context of usage (e.g., user tasks, dataset characteristics, etc.); in IVML systems, it is their end-users that train the model. There are various contexts in which machine learning models are created, and evaluators may not be able to foresee the possible use cases to evaluate (e.g., validate what is going on in the machine learning under the hood, as every user may utilize it very differently). A particular example of circumstantiality is when people have private data (e.g., health data) and so they want to be able to build a model themselves.¹⁴ In such IVML systems, evaluators may not have access to raw private data, and thus, they study the system with alternative datasets at the expense of having results that are perhaps less certain, or less pertinent to the target audience. Thus, the circumstances of evaluating this particular IVML are unique, as there are privacy requirements and the problems attacked may vary from one user to another.

Research Opportunities

We need more research on how to evaluate the effects of uncertainty in the use of IVML systems. In particular, it is useful to explore how communicating various forms of uncertainty, which characterize model development and usage, impact how humans draw conclusions from data-driven estimates or models, how they affect trust and confidence in model predictions, and how in turn these affect the feedback humans give to the ML component. Recent applications of Bayesian inference to visualization evaluation may offer one promising avenue for evaluating IVML.²⁶

Bias from the ML side alone (e.g., due to class-imbalance in the datasets used to train these machine learning algorithms) is a well-known problem and research suggests how to tackle this issue before reaching the stage of evaluation with humans (e.g., using different forms of resampling, adjusting the decision threshold, or combining the results of many

classifiers²¹). While cognitive biases (from the human side) are difficult to eliminate and may require implementing specific exploration strategies.¹³ IVML evaluations should consider biases during evaluations and strive to highlight their possible sources and calibrate the results accordingly. We can be inspired by work³³ that has started to look at what can be measured and communicated to IVML users in real-time in an attempt to mitigate such biases.

Finally, given that the ML component of IVML systems is trained by the end user, we need to consider their evaluation as circumstantial, i.e., prone to variations in IVML use context, user tasks, privacy concerns, and dataset characteristics.

EVALUATION GUIDELINES AND METRICS

There is a lack of guidelines to design IVML systems and taxonomies to characterize those systems and their associated tasks (with a few recent exceptions^{1,23}). Even when it comes to existing usability guidelines, IVML systems can breach established guidelines such as consistency due to changes as a result of learning over time or failing to prevent errors due to some poorly understood probabilistic behaviors.¹ As such, IVML tool builders struggle to select the most appropriate evaluation methods. As different types of IVML systems exist,^{3,17} there are no guidelines to help determine what degree of integration is needed to support what analysis tasks, and the corresponding pertinent evaluation method.

In terms of metrics used to evaluate IVML systems, besides performance metrics such as accuracy of model predictions, existing evaluations measure interpretability, trust, and user confidence in ML results. However, most evaluations do not explain why the collaboration between the AI and the human was successful (or a failure). They are not able to explain whether or how the human was able to use the model predictions, and whether this information resulted in changes in how they are looking at the ML problem to be solved. Some recent work tries to elicit people's predictions (that express the mental model they have developed), as a way to gauge how well they are interpreting a model.²⁵ But it is still difficult to

prescribe how people's mental model of the underlying ML model should be developing when interacting with model predictions that change over time.

Research Opportunities

A number of evaluation metrics have been recently proposed such as intelligibility, quality of explanation,²⁸ and appropriate trust. There is a need to investigate how reliable these metrics are. For example, when it comes to trust as a metric, it is important to create IVML systems that foster "appropriate" user trust.¹⁹ Thus, IVML systems should maximize not only cases where the machine learning is correct and the user trusts and accepts the ML recommendation, but also cases where the ML is incorrect and the user rejects the system's advice. In cases where confidence in model prediction is low, we need to investigate smooth hand-off mechanisms between the AI and the user.

Another metric is a user engagement level. A high level of engagement (e.g., through increased interaction and inference tasks⁸) could indicate that users enjoy the tool, and that they are likely to learn more through the usage. However, soliciting frequent user feedback and interaction may be counterproductive due to user fatigue. This opens avenues for research on how to measure engagement itself, directly (through user self-reporting) or indirectly (through logging of interactions). This question of how to elicit feedback from the analyst does not only pertain to measuring success. IVML systems can be trained by user interactions that are implicit or explicit,⁵ and it is possible we need to adapt our metrics accordingly.

We can also be inspired by other fields in learning and behavioral sciences to help us better understand how people respond to information and learn from it, which will be very useful in helping us evaluate people's trust in and the use of ML and human-in-the-loop AI.

Finally, although this challenge is more general and broad than the previous three, addressing the first three challenges could go a long way to providing starting guidelines and metrics for evaluating IVML systems.

CONCLUSION

IVML systems combine human and machine intelligence to collaboratively achieve a task. Human analysts and ML are partners solving problems as a unit: analysts do not merely interpret the decisions and ML models, but they actively act on, and react to, these models. By acting on their interpretation of model decisions, the role of humans in IVML goes beyond just understanding the underlying model and predictions. Therefore, tackling some of the aforementioned challenges in evaluating IVML systems would result in IVML systems that could also serve as effective explainable ML tools.

We have identified unique challenges with respect to the evaluation of the various facets of IVML systems, as well as research opportunities associated with them. We highlight the roles of people and machine learning and how they can inform future evaluations, consider the different tradeoffs related to the objectives of these systems, discuss the effect of different types of uncertainty and context of use on the decisions reached when using IVML, and highlight the subtleties of selecting appropriate guidelines and metrics to evaluate the different components of IVML systems.

ACKNOWLEDGMENTS

This viewpoint comes from discussions of six paper presentations, a keynote, and a panel held at IEEE VIS 2019 EVIVA-ML workshop (eviva-ml.github.io), which brought together more than 80 participants to discuss current challenges and a research agenda for IVML evaluation. Thanks to M. Chegini and M. Kahng for providing images 1.2 and 1.3.

REFERENCES

1. S. Amershi *et al.*, "Guidelines for human-AI interaction," in *Proc. CHI Conf. Human Factors Comput. Syst.*, May 2019, pp. 1–13.
2. G. Bansal, B. Nushi, E. Kamar, W. S. Lasecki, D. S. Weld, and E. Horvitz, "Beyond accuracy: The role of mental models in human-AI team performance," in *Proc. AAAI Conf. Human Comput. Crowdsourcing*, 2019, vol. 7, no. 1, pp. 2–11.

3. E. Bertini and D. Lalanne, "Surveying the complementary role of automatic data analysis and visualization in knowledge discovery," in *Proc. ACM SIGKDD Workshop Vis. Anal. Knowl. Discovery: Integrating Autom. Anal. Interactive Exploration*, Jun. 2009, pp. 12–20.
4. N. Boukhelifa, A. Bezerianos, W. Cancino, and E. Lutton, "Evolutionary visual exploration: Evaluation of an IEC framework for guided visual search," *Evol. Comput.*, vol. 25, no. 1, pp. 55–86, 2017.
5. N. Boukhelifa, A. Bezerianos, and E. Lutton, "Evaluation of interactive machine learning systems," in *Human and Machine Learning*. Berlin, Germany: Springer, 2018, pp. 341–360.
6. Y. Bian, M. Dowling, and C. North, "Evaluating semantic interaction on word embeddings via simulation," in *Proc. IEEE Eval. Interactive Visual Mach. Learn. Syst.*, 2019, pp. 1–5.
7. E. T. Brown, J. Liu, C. E. Brodley, and R. Chang, "Dis-function: Learning distance functions interactively," in *Proc. IEEE Conf. Visual Analytics Sci. Technol.*, Oct. 2012, pp. 83–92.
8. D. Cashman, Y. Wu, R. Chang, and A. Ottley, "Inferential tasks as a data-rich evaluation method for visualization," in *Proc. IEEE VIS Workshop Eval. Interactive Visual Mach. Learn. Syst.*, 2019, pp. 1–5.
9. T. Chabin *et al.*, "LIDeOGraM: An interactive evolutionary modelling tool," in *Proc. Int. Conf. Artif. Evol. (Evol. Artificielle)*, Oct. 2017, pp. 189–201.
10. K. Charmaz, *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis*. Newbury Park, CA, USA: Sage, 2006.
11. M. Chegini, J. Bernard, L. Shao, A. Sourin, K. Andrews, and T. Schreck, "mVis in the wild: Pre-study of an interactive visual machine learning system for labelling," in *Proc. IEEE VIS workshop Workshop Eval. Interactive Visual Mach. Learn. Syst.*, 2019, pp. 1–4.
12. C. Collins *et al.*, "Guidance in the human-machine analytics process," *Vis. Informat.*, vol. 2, no. 3, pp. 166–180, 2018.
13. E. Dimara, S. Franconeri, C. Plaisant, A. Bezerianos, and P. Dragicevic, "A task-based taxonomy of cognitive biases for information visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 2, pp. 1413–1432, Feb. 2020.
14. S. M. Drucker, D. Fisher, and S. Basu, "Helping users sort faster with adaptive machine learning recommendations," in *Proc. 13th IFIP TC Int. Conf. Human-Comput. Interact.*, Sep. 2011, pp. 187–203.
15. M. El-Assady, R. Sevastjanova, F. Sperrle, D. Keim, and C. Collins, "Progressive learning of topic modeling parameters: A visual analytics framework," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 1, pp. 382–391, Jan. 2018.
16. A. Endert, M. S. Hossain, N. Ramakrishnan, C. North, P. Fiaux, and C. Andrews, "The human is the loop: New directions for visual analytics," *J. Intell. Inf. Syst.*, vol. 43, no. 3, pp. 411–435, 2014.
17. A. Endert *et al.*, "The state of the art in integrating machine learning into visual analytics," *Comput. Graph. Forum*, vol. 36, no. 8, pp. 458–486, 2017.
18. M. Fernandes, L. Walls, S. Munson, J. Hullman, and M. Kay, "Uncertainty displays using quantile dotplots or CDFs improve transit decision-making," in *Proc. CHI Conf. Human Factors Comput. Syst.*, Apr. 2018, pp. 1–12.
19. D. Gunning, Explainable artificial intelligence (XAI), Defense Advanced Research Projects Agency (DARPA), Arlington County, VA, USA, 2017.
20. S. Hong, J. Hullman, and E. Bertini, "Human factors in model interpretability: Industry practices, challenges, and needs," in *Proc. ACM Conf. Comput. Supported Cooperative Work*, 2020, pp. 1–26.
21. A. Howard, C. Zhang, and E. Horvitz, "Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems," in *Proc. IEEE Workshop Adv. Robot. Social Impacts*, 2017, pp. 1–7.
22. J. Hullman, X. Qiao, M. Correll, A. Kale, and M. Kay, "In pursuit of error: A survey of uncertainty visualization evaluation," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 1, pp. 903–913, Jan. 2019.
23. L. Jiang, S. Liu, and C. Chen, "Recent research advances on interactive machine learning," *J. Visualization*, vol. 22, no. 2, pp. 401–417, 2019.
24. M. Kahng and D. H. Chau, "How does visualization help people learn deep learning?" in *Proc. IEEE VIS workshop EVIVA-ML*, 2019, pp. 1–3.
25. Y. S. Kim, K. Reinecke, and J. Hullman, "Explaining the gap: Visualizing one's predictions improves recall and comprehension of data," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2017, pp. 1375–1386.
26. Y. S. Kim, L. A. Walls, P. Krafft, and J. Hullman, "A Bayesian cognition approach to improve data visualization," in *Proc. CHI Conf. Human Factors Comput. Syst.*, May 2019, pp. 1–14.
27. H. T. Kung, F. Luccio, and F. P. Preparata, "On finding the maxima of a set of vectors," *J. ACM*, vol. 22, no. 4, pp. 469–476, 1975.

28. I. Lage *et al.*, "Human evaluation of models built for interpretability," in *Proc. AAAI Conf. Human Comput. Crowdsourcing*, vol. 7, no. 1, 2019, pp. 59–67.
29. W. E. Mackay, "Users and customizable software: A co-adaptive phenomenon," Doctoral dissertation, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, USA, 1990.
30. L. Micallef *et al.*, "The Human user in progressive visual analytics," in *Proc. 21st EG/VGTC Conf. Visualization*, 2019, pp. 19–23.
31. D. Sacha *et al.*, "Human-centered machine learning through interactive visualization," in *Proc. 24th Eur. Symp. Artif. Neural Netw., Comput. Intell. Mach. Learn.*, Aug. 2016, pp. 641–646.
32. A. C. Valdez, M. Ziefle, and M. Sedlmair, "Priming and anchoring effects in visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 1, pp. 584–594, Jan. 2018.
33. E. Wall, L. M. Blaha, L. Franklin, and A. Endert, "Warning, bias may occur: A proposed approach to detecting cognitive bias in interactive visual analytics," in *Proc. IEEE Conf. Visual Analytics Sci. Technol.*, 2017, pp. 104–115.
34. Y. Zhang, B. Coecke, and M. Chen, "On the cost of interactions in interactive visual machine learning," in *Proc. IEEE VIS Workshop Eval. Interactive Visual Mach. Learn. Syst.*, 2019, pp. 1–5.

N. Boukhelifa is currently a Research Scientist in visualization and interactive modeling for agronomy with INRAE, Paris, France. She is the corresponding author of this article. Contact her at nadia.boukhelifa@inrae.fr.

A. Bezerianos is currently an Associate Professor with Université Paris-Saclay, Essonne, France. She works at the intersection of human–computer interaction and information visualization. Contact her at anab@lri.fr.

R. Chang is currently an Associate Professor of computer science with Tufts University, Medford, MA, USA. His current research interests include visual analytics, information visualization, human–computer interaction and databases. Contact him at remco@cs.tufts.edu.

C. Collins is currently an Associate Professor of computer science and Canada Research Chair in linguistic information visualization with Ontario Tech University, Oshawa, ON, Canada. Contact him at christopher.collins@ontariotechu.ca.

S. Drucker is currently a Partner & Research Manager with Microsoft Research, Redmond, WA, USA, heading the Visualization in Data and Analytics (VIDA) Research group. Contact him at sdrucker@microsoft.com.

A. Endert is currently an Associate Professor with the School of Interactive Computing, Georgia Tech, Atlanta, GA, USA. His work explores new methods for user interaction with visual analytic tools to help people make sense of data. Contact him at endert@gatech.edu.

J. Hullman is currently an Associate Professor of computer science with Northwestern University, Evanston, IL, USA. Her work addresses visualization, uncertainty communication, and behavioral information design. Contact her at jhullman@northwestern.edu.

C. North is currently a Professor of computer science with Virginia Tech, Blacksburg, VA, USA. He is an Associate Director of the Discovery Analytics Center, and leads the Visual Analytics research group. Contact him at north@vt.edu.

M. Sedlmair is currently a Junior Professor with the University of Stuttgart, Stuttgart, Germany, where he works at the intersection of human–computer interaction, visualization, and data science. Contact him at michael.sedlmair@visus.uni-stuttgart.de.

Contact department editor Theresa-Marie Rhyne at theresamarieryhne@gmail.com.