Challenges in Evaluating Interactive Visual Machine Learning Systems

Nadia Boukhelifa, Anastacia Bezerianos, Remco Chang, Christopher Collins, Steven Drucker, Alex Endert, Jessica Hullman, Chris North, Michael Sedlmair

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 interactive visual machine learning 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 them using DBSCAN, or to learn a mathematical fitness or objective function), mediated by an interactive visual interface (e.g., [3,4,7,15]).

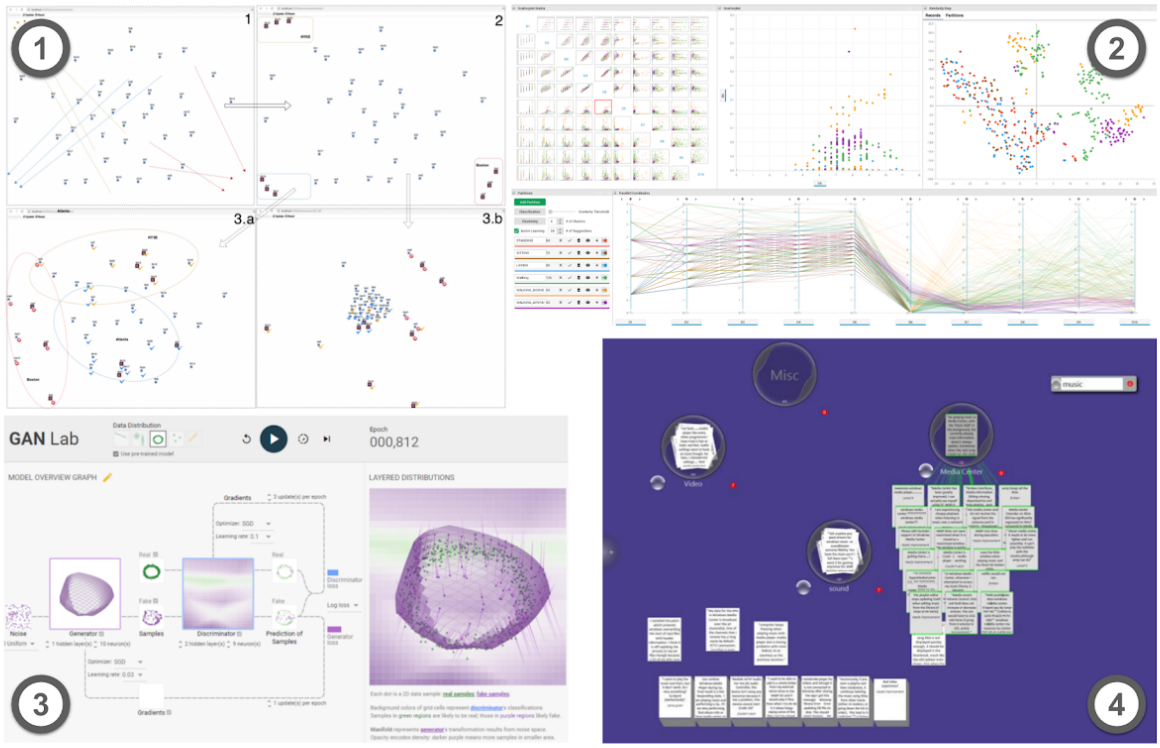

Figure 1. Examples of IVML systems where the analyst interacts with the visualization, and the system updates the machine learning model and adapts the visualization accordingly. (1) Semantic Interaction [6] enables analysts to modify a suggested projection of documents (top), that updates the underlying weighted distance of two alternative projection models (bottom). (2) mVis [11] 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 Lab [24] helps analysts learn about GAN models by visualizing their structure and enabling users to interactively train and experiment with the models. (4) iCluster [14] 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.

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 interactive visual machine learning 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., [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 act on their reactions to these models, such as by 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 a 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 & 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 theory [10]).

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 **trade-offs**; (3) assessing the effects of different sources of **uncertainty** on the use of IVML systems; and (4) providing practical evaluation **guidelines & metrics** for IVML systems.

#1 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 co-adaptation 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 it is the ML component that learns and evolves, 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.

Considering the evaluation 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 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); (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 co-evolution 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 co-evolution) 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 tests [4]). 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 [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 de-couple 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 where the different components can be more easily isolated.

#2 MANY **TRADE-OFF** 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 (**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.

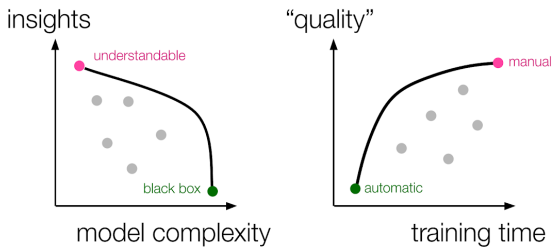

Figure 2. IVML evaluations can be modeled as a set of optimal evaluation studies along a Pareto front [27]. Multiple Pareto fronts can exist when it comes to the evaluation goal, such as between (Left) measuring insights and model complexity, where simplified models may be easier to understand, while 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.

Trade-off considerations for IVML evaluation are not limited to those three dimensions (i.e., accuracy, complexity, and interpretability), and can be found in any visual analysis system that includes an automated component. But other 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, however, e.g., 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.

#3 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 assessed 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 decisions [18] 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, ...); 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, impacts 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 re-sampling, adjusting the decision

threshold, or combining the results of many classifiers [21]). 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 work [33] 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.

#4 - EVALUATION **GUIDELINES & 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 user engagement level. A high level of engagement (e.g., through increased interaction and inference tasks [8]) 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 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

Interactive visual machine learning (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 & metrics** to evaluate the different components of interactive visual machine learning systems.

ACKNOWLEDGMENT

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 over 80 participants, to discuss current challenges and a research agenda for IVML evaluation. Thanks to Mohammad Chegini and Minsuk Kahng and for providing images 1.2 and 1.3.

REFERENCES

1. Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P.N., Inkpen, K. and Teevan, J., 2019, May. Guidelines for human-ai interaction. In Proc. of the 2019 CHI Conf. Hum. Factors Comput. Syst. (pp. 1-13).
2. Bansal, G., Nushi, B., Kamar, E., Lasecki, W.S., Weld, D.S. and Horvitz, E., 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. In Proc. of the AAAI Conf. on Human Comp. and Crowdsourcing (Vol. 7, No. 1, pp. 2-11).
3. Bertini, E. and Lalanne, D., 2009, June. Surveying the complementary role of automatic data analysis and visualization in knowledge discovery. In Proc. of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration (pp. 12-20).
4. Boukhelifa, N., Bezerianos, A., Cancino, W. and Lutton, E., 2017. Evolutionary visual exploration: Evaluation of an IEC framework for guided visual search. Evolutionary computation, 25(1), pp.55-86.
5. Boukhelifa, N., Bezerianos, A. and Lutton, E., 2018. Evaluation of interactive machine learning systems. In Human and Machine Learning. (pp. 341-360). Springer, Cham.
6. Bian, Y., Dowling, M. and North, C., 2019. Evaluating Semantic Interaction on Word Embeddings via Simulation. In Proc. IEEE VIS workshop EVIVA-ML, 2019.
7. Brown, E.T., Liu, J., Brodley, C.E. and Chang, R., 2012, October. Dis-function: Learning distance functions interactively. In 2012 IEEE Conference VAST (pp. 83-92). IEEE.
8. Cashman, D., Wu, Y., Chang, R., Ottley, A. Inferential tasks as a data-rich evaluation method for visualization. In interactive visual machine learning. In Proc. IEEE VIS workshop EVIVA-ML, 2019.
9. Chabin, T., Barnabé, M., Boukhelifa, N., Fonseca, F., Tonda, A., Velly, H., Lemaitre, B., Perrot, N. and Lutton, E., 2017, October. LIDeOGraM: an interactive evolutionary modelling tool. In International Conference on Artificial Evolution (Evolution Artificielle) (pp. 189-201). Springer, Cham.
10. Charmaz, K., 2006. *Constructing grounded theory: A practical guide through qualitative analysis*. sage.
11. Chegini, M., Bernard, J., Shao, L., Sourin, A., Andrews, K. and Schreck, T., 2019. mVis in the Wild: Pre-Study of an Interactive Visual Machine Learning System for Labelling. In Proc. IEEE VIS workshop EVIVA-ML, 2019.
12. Collins, C., Andrienko, N., Schreck, T., Yang, J., Choo, J., Engelke, U., Jena, A. and Dwyer, T., 2018. Guidance in the human–machine analytics process. Visual Informatics, 2(3), pp.166-180.
13. Dimara, E., Franconeri, S., Plaisant, C., Bezerianos, A. and Dragicevic, P., 2018. A task-based taxonomy of cognitive biases for information visualization. IEEE Trans. Vis. Comput. Graph.

14. Drucker, S.M., Fisher, D. and Basu, S., 2011, September. Helping users sort faster with adaptive machine learning recommendations. In IFIP Conf. on HCI (pp. 187-203). Springer, Berlin, Heidelberg.
15. El-Assady, M., Sevastjanova, R., Sperrle, F., Keim, D. and Collins, C., 2017. Progressive learning of topic modeling parameters: A visual analytics framework. IEEE Trans. Vis. Comput. Graph., 24(1), pp.382-391.
16. Endert, A., Hossain, M.S., Ramakrishnan, N., North, C., Fiaux, P. and Andrews, C., 2014. The human is the loop: new directions for visual analytics. Journal of intelligent information systems, 43(3), pp.411-435.
17. Endert, A., Ribarsky, W., Turkay, C., Wong, B.W., Nabney, I., Blanco, I.D. and Rossi, F., 2017. The state of the art in integrating machine learning into visual analytics. In Cmp. Gra. F (Vol. 36, No. 8, pp. 458-486).
18. Fernandes, M., Walls, L., Munson, S., Hullman, J. and Kay, M., 2018, April. Uncertainty displays using quantile dotplots or cdfs improve transit decision-making. In Proc. of the 2018 CHI Conf. Hum. Factors Comput. Syst. (pp. 1-12).
19. Gunning, D., 2017. Explainable artificial intelligence (xai). Defense Advanced Research Projects Agency (DARPA), nd Web, 2.
20. Hong, S., Hullman, J., and Bertini, E. Human factors in model interpretability: Industry practices, challenges, and needs, 2020. Proc. of ACM Conf. on Computer Supported Cooperative Work (CSCW).
21. Howard, A., Zhang, C. and Horvitz, E., 2017, March. Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems. In 2017 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO) (pp. 1-7). IEEE.
22. Hullman, J., Qiao, X., Correll, M., Kale, A. and Kay, M., 2018. In pursuit of error: A survey of uncertainty visualization evaluation. IEEE Trans. Vis. Comput. Graph., 25(1), pp.903-913.
23. Jiang, L., Liu, S. and Chen, C., 2019. Recent research advances on interactive machine learning. Journal of Visualization, 22(2), pp.401-417.
24. Kahng, M. and Chau, D.H., 2019. How does visualization help people learn deep learning? evaluation of GAN Lab. In Proc. IEEE VIS workshop EVIVA-ML, 2019.
25. Kim, Y.S., Reinecke, K. and Hullman, J., 2017, May. Explaining the gap: Visualizing one's predictions improves recall and comprehension of data. In Proc. of the 2017 CHI Conf. Hum. Factors Comput. Syst. (pp. 1375-1386).
26. Kim, Y.S., Walls, L.A., Krafft, P. and Hullman, J., 2019, May. A bayesian cognition approach to improve data visualization. In Proc. of the 2019 CHI Conf. Hum. Factors Comput. Syst. (pp. 1-14).
27. Kung, H.T., Luccio, F. and Preparata, F.P., 1975. On finding the maxima of a set of vectors. *Journal of the ACM (JACM)*, *22*(4), pp.469-476.
28. Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S. and Doshi-Velez, F., 2019. An evaluation of the human-interpretability of explanation. arXiv preprint arXiv:1902.00006.
29. Mackay, W.E., 1990. Users and customizable software: A co-adaptive phenomenon (Doctoral dissertation, MIT, Sloan School of Management).
30. Micallef, L., Schulz, H.J., Angelini, M., Aupetit, M., Chang, R., Kohlhammer, J., Perer, A. and Santucci, G., 2019. The Human user in progressive visual analytics. In 21st EG/VGTC Conference on Visualization.
31. Sacha, D., Sedlmair, M., Zhang, L., Lee, J.A., Weiskopf, D., North, S. and Keim, D., 2016, August. Human-centered machine learning through interactive visualization. ESANN.
32. Valdez, A.C., Ziefle, M. and Sedlmair, M., 2017. Priming and anchoring effects in visualization. IEEE Trans. Vis. Comput. Graph., 24(1), pp.584-594.
33. Wall, E., Blaha, L.M., Franklin, L. and Endert, A., 2017, October. Warning, bias may occur: A proposed approach to detecting cognitive bias in interactive visual analytics. In 2017 IEEE Conference VAST (pp. 104-115). IEEE.
34. Zhang, Y., Coecke, B., and Chen, M. 2019. On the cost of interactions in interactive visual machine learning. in Proc. EVIVA-ML, 2019.

N. Boukhelifa is Research Scientist in visualization and interactive modelling for agronomy at INRAE, France. Contact her at nadia.boukhelifa@inrae.fr.

A. Bezerianos is an Associate Professor at Université Paris-Saclay. She works at the intersection of Human-Computer Interaction and Information Visualization. Contact her at anab@lri.fr.

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

C. Collins is an Associate Professor of Computer Science and Canada Research Chair in Linguistic Information Visualization at Ontario Tech University. Contact him at christopher.collins@ontariotechu.ca.

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

A. Endert is an Associate Professor in the School of Interactive Computing at Georgia Tech. 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 an Associate Professor of Computer Science at Northwestern University. Her work addresses visualization, uncertainty communication, and behavioral information design. Contact her at jhullman@northwestern.edu.

C. North is a Professor of Computer Science at Virginia Tech. He is Associate Director of the Discovery Analytics Center, and leads the Visual Analytics research group. Contact him at north@vt.edu.

M. Sedlmair is a junior professor at the University of Stuttgart, where he works at the intersection of Human-Computer Interaction, Visualization, and Data Science. Contact him at michael.sedlmair@visus.uni-stuttgart.de