Communicating Uncertainty in Machine Learning Systems

Exploring the visualization implications of different *stakeholders* and *sources of uncertainty*

ASPEN HOPKINS and HARINI SURESH

Additional Key Words and Phrases: Machine Learning, Uncertainty, Data Visualization

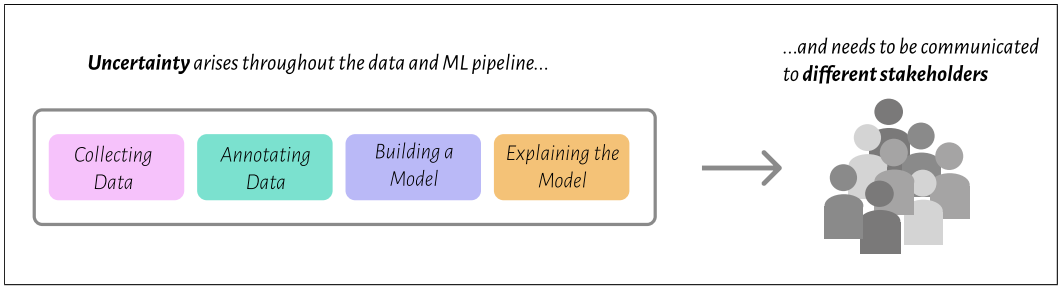

Fig. 1. Multiple Sources of Uncertainty in the Data and ML Pipeline

Machine learning (ML) systems are developed and deployed for a broad range of tasks, informing decisions in domains from healthcare [7] to hiring [21]. These systems aren't used in isolation; they are part of larger social systems and are subject to interpretation by a wide range of stakeholders [22]. In these complex socio-technical settings, presenting model outputs without appropriate contextualization can impart undue authority onto the model. Indeed, recent work has shown that simply receiving model recommendations on their own [10, 11, 16]—or even with generically visualized metadata and explanations [2, 6, 20, 25]—can lead to over-reliance, wherein users come to rely on the decisions of automated systems over their own expertise. This can be quite harmful—models may suffer from underspecification [8], overconfidence [12], or discriminatory behavior [3], requiring careful consideration rather than dependency. More broadly, simply deploying black box models can discourage affected stakeholders to question or push back on model decisions.

ML systems are the product of complex processes, from data collection and annotation to model building and deployment—each of which can introduce forms of uncertainty into the final output. Communicating this context is key to combating the false sense of authority afforded to ML systems. We posit that comprehensive, intuitive communication of a model's uncertainty is a critical tool both for *building effective trust* (i.e., helping users know when to trust the model and when to rely more on their own judgement) and for *enabling contestability* (i.e., by making model outputs less authoritative and empowering affected stakeholders to push back on them). Work in ML has included developing models that better quantify their own uncertainty [15, 19] and has framed uncertainty as an important facet of model transparency [4]. Still, there is a large gap between *computing* uncertainty and *communicating* it in a way that is relevant and understandable to users. Our work is aimed at this gap, exploring *how* and *what* uncertainty measures should be presented to different stakeholders.

We draw from other domains to explore these questions—most relevantly, uncertainty communication has a rich history in the data visualization research community. But the scope and nature of this work is still limited to a particular framing of uncertainty; in seeking to effectively communicate

Authors' address: Aspen Hopkins, dataspen@mit.edu; Harini Suresh, hsuresh@mit.edu.

and understand uncertainty to reduce ambiguity, some information about its underlying nature is lost. When uncertainty is presented visually, it is often through a singular, cumulative encoding, such as in the height of an error bar or boxplot [17], an animation [1, 9], the width of a line chart's ribbon [18], or text and glyphs on or near a graph [5]. The nature of these visual artifacts hints at the balancing act visualization authors face in communicating uncertainty—they must weigh the relative trade-offs of a graph's comprehensibility with the desire to communicate nuance. These binned encodings are visually simple not just for readability, but also due to the uncertainty we as a community have focused on: total uncertainty, a quantified, numerical value representing an apparently complete measure of data's ambiguity. Yet as Kale et al. [14] describe, the *"drive to reduce uncertainty can lead to unwarranted expressions of certainty, which has consequences for decision-making individually and at an organizational level."*

While its communication is a place of ongoing research in both ML and in data visualization, uncertainty is complex, multifaceted, and not always quantifiable [23], an issue further complicated when the uncertainty is heterogeneous—with multiple sources of ambiguity contributing to its composite. Yet heterogenous uncertainty is common, and may warrant unique treatment based on stakeholders' needs [13, 24]. Here we begin to explore, through illustrative examples, how 1) different sources of uncertainty and 2) different stakeholders of machine learning systems require different visual representations of uncertainty.

Disentangling sources of uncertainty: Scientific Analysis of Ocean Core Data. To explore the implications of different sources of uncertainty, we consider heterogeneous uncertainties introduced during collection and analysis of a large, collaborative geological core dataset. The dataset is composed of various modalities of data; we focus on the microspectral images of core rock samples. These microspectral images, and their subsequent use, present many places where uncertainty is introduced, including: several forms of collection bias; variable data quality in physical samples, images, and various experimental results; uncertainty from statistical modeling and interpolation methods; and even invisible minerals such as quartz, which are known to be present but are not caught through hyperspectral analysis. The implications of these forms of uncertainty vary, and often require distinctly different responses. For example, uncertainty that arises due to unfamiliar textures present in the rock may warrant physical inspection and might be a source of great research interest. In contrast, uncertainty due to random noise can be automatically accounted for. These *sources* of uncertainty matter and may require disentanglement when visualizing model outputs.

Designing for different stakeholders: Diagnostic Decision Aids. To then explore how uncertainty visualization needs might differ across stakeholders, we consider ML systems in healthcare contexts. In recent work, we studied how ML recommendations were used in medical decision-making. When presented with diagnostic recommendations for a series of chest x-rays, we found that both radiologists and internal/emergency medicine (IM/EM) physicians were susceptible to incorrect advice—and that this over-reliance effect was more pronounced for IM/EM physicians with less expertise in reading chest x-rays [10]. Visualizing uncertainty along with the model's prediction could be a critical tool for combatting over-reliance in this context. However, behavior differences amongst physicians with different levels of task expertise suggests that there is not a "one-size-fits-all" approach. In other prior work, we describe cases where dental practitioners were shown varying levels of model sensitivities based on experience [13]. There are several competing tensions when thinking about stakeholder needs—for example, we might want to emphasize uncertainty to stakeholders with less domain expertise who are more susceptible to accepting incorrect advice; at the same time, these users may also be more likely to experience information overload if the visualization is *too* detailed or comprehensive. In our work, we begin to enumerate and tease apart these tensions.

Fundamentally, our work asks *how* and *what* uncertainty measures should be visualized in ML systems. We use a series of examples to explore how to visualize heterogeneous uncertainty to different stakeholders, with the long-term aim of enabling appropriate trust and contestability in these systems.

REFERENCES

[1] Gregor Aisch. 2019. Why we used jittery gauges in our live election forecast. https://www.vis4.net/blog/2016/11/jittery-gauges-election-forecast/

[2] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. https://doi.org/10.1145/3411764.3445717

[3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2017. Fairness in machine learning. *Neurips tutorial* 1 (2017), 2.

[4] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Gauthier Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, et al. 2020. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. *arXiv preprint arXiv:2011.07586* (2020).

[5] Georges-Pierre Bonneau, Hans-Christian Hege, Chris R. Johnson, Manuel M. Oliveira, Kristin Potter, Penny Rheingans, and Thomas Schultz. 2014. *Overview and State-of-the-Art of Uncertainty Visualization.* Springer London, London, 3–27. https://doi.org/10.1007/978-1-4471-6497-5_1

[6] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. https://doi.org/10.1109/ICHI.2015.26

[7] Alison Callahan and Nigam H Shah. 2017. Machine learning in healthcare. In *Key Advances in Clinical Informatics*. Elsevier, 279–291.

[8] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. 2020. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395* (2020).

[9] Charles R. Ehlschlaeger, Ashton M. Shortridge, and Michael F. Goodchild. 1997. Visualizing spatial data uncertainty using animation. *Computers Geosciences* 23, 4 (1997), 387–395. https://doi.org/10.1016/S0098-3004(97)00005-8 Exploratory Cartograpic Visualisation.

[10] Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J Berkowitz, Eva Lermer, Joseph F Coughlin, John V Guttag, Errol Colak, and Marzyeh Ghassemi. 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine* 4, 1 (2021), 1–8.

[11] Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 90–99.

[12] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*. PMLR, 1321–1330.

[13] Aspen Hopkins and Serena Booth. 2021. Machine Learning Practices Outside Big Tech: How Resource Constraints Hinder Responsible Development.. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*.

[14] Alex Kale, Matthew Kay, and Jessica Hullman. 2019. Decision-making under uncertainty in research synthesis: Designing for the garden of forking paths. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[15] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6405–6416.

[16] Jennifer M. Logg, Julia A. Minson, and Don A. Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

[17] Chris Olston and Jock D Mackinlay. 2002. Visualizing data with bounded uncertainty. In *IEEE Symposium on Information Visualization, 2002. INFOVIS 2002.* IEEE, 37–40.

[18] Lace Padilla, Matthew Kay, and Jessica Hullman. [n.d.]. Uncertainty visualization. ([n. d.]).

[19] Nicolas Papernot and Patrick McDaniel. 2018. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv:1803.04765* (2018).

[20] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 237, 52 pages. https://doi.org/10.1145/3411764.3445315

[21] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 469–481.

[22] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 59–68. https://doi.org/10.1145/3287560.3287598

[23] Meredith Skeels, Bongshin Lee, Greg Smith, and George G Robertson. 2010. Revealing uncertainty for information visualization. *Information Visualization* 9, 1 (2010), 70–81.

[24] Harini Suresh, Steven R Gomez, Kevin K Nam, and Arvind Satyanarayan. 2021. Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and their Needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[25] Harini Suresh, Natalie Lao, and Ilaria Liccardi. 2020. Misplaced Trust: Measuring the Interference of Machine Learning in Human Decision-Making. In *12th ACM Conference on Web Science*. 315–324.