This forum provides a space to engage with the challenges of designing for intelligent algorithmic experiences. We invite articles that tackle the tensions between research and practice when integrating AI and UX design. We welcome interdisciplinary debate, artful critique, forward-looking research, case studies of AI in practice, and speculative design explorations. — Juho Kim and Henriette Cramer, Editors

Data Excellence for AI: Why Should You Care?

Lora Aroyo, Google Research, Matthew Lease, University of Texas at Austin, Praveen Paritosh, Google Research, Mike Schaeckermann, Amazon

The efficacy of machine learning (ML) models depends on both algorithms and data. Training data defines what we want our models to learn, and testing data provides the means by which their empirical progress is measured. Benchmark datasets define the entire world within which models exist and operate, yet research continues to focus on critiquing and improving the algorithmic aspect of the models rather than critiquing and improving the data with which our models operate. If “data is the new oil,” we are still missing work on the refineries by which the data itself could be optimized for more effective use.

Data is potentially the most undervalued and deglamorized aspect of today’s AI ecosystem. Data issues are often perceived and characterized as mundane “preprocessing” that has to be done before getting to the real work (i.e., modeling). ML practitioners often view data wrangling as tedious and time-consuming. In contrast, Sambasivan et al. [1] provide examples of how data quality is crucial to ensure that AI systems can accurately represent and predict the phenomena they are claiming to measure. Amershi et al. [2] identify three aspects of the AI domain that make it fundamentally different from prior software application domains. One of the challenges identified focuses on processes for data discovery, management, and versioning, which are much

more complex and difficult than in traditional software engineering.

Real-world datasets are often “dirty” with various data-quality problems and present the risk of “garbage in = garbage out” in terms of the downstream AI systems we train and test on such data. This has inspired a steadily growing body of work on understanding and improving data quality. It also highlights the importance of rigorously managing data quality using mechanisms specific to data validation, instead of relying on model performance as a proxy for data quality. Just as we rigorously test our code for software defects before deployment, we need to test for data defects with the same degree of rigor so that we can detect or prevent weaknesses in ML models caused by underlying issues in data quality.

Benchmark datasets are often missing much of the natural ambiguity of the

real world, as data instances with annotator disagreement are often aggregated to eliminate disagreement (obscuring uncertainty) or filtered out of datasets entirely. Successful benchmark models fail to generalize real data, and inflated benchmark results mislead our assessment of state-of-the-art capabilities. Thus, ML models become prone to develop “weak spots,” classes of examples that are difficult or impossible for a model to accurately evaluate because they are missing from the evaluation set [3].

Measuring data quality is challenging, nebulous, and often circularly defined with “ground truth” on which models are tested. When dataset quality is considered, the ways in which it is measured in practice are often poorly understood, and sometimes simply wrong. Challenges identified include fairness and bias issues in labeled datasets, quality issues in datasets, limitations of benchmarks, reproducibility concerns in machine learning research, and lack of documentation and replication of data.

Measurement of AI success today is often metrics driven, with emphasis on rigorous model measurement and A/B testing. Measuring the goodness of the model’s fit to the dataset, however, ignores any consideration of how well the dataset fits the real-world problem. Goodness-of-fit metrics, such as F1, Accuracy, and AUC, do not tell us much about data *fidelity* (i.e., how well the dataset represents reality) and *validity* (i.e., how well the data explains things related to the phenomena captured by the data). No

Insights

- Measurement of AI success is often metrics driven and does not tell us much about data fidelity and validity. No standardized metrics exist yet for characterizing the goodness-of-data.
- If “data is the new oil,” our use of data remains crude, and we are missing work on the refineries by which the data itself could be optimized.
- Real-world datasets are often “dirty,” with various data-quality problems and the risk of “garbage in = garbage out.”

standardized metrics exist today for characterizing the goodness of data.

HOW CAN LESSONS FROM EXCELLENCE IN SOFTWARE ENGINEERING HELP DATA EXCELLENCE?

Decades of catastrophic failures within high-stakes software projects have helped establish the crucial importance of upfront investments in *software engineering excellence*. It was through careful post hoc analysis of these kinds of disasters that software engineering has matured as a field and achieved a robust understanding of the costs and benefits: *processes* like systematic code reviews, *standards* like coding conventions and design patterns, and *infrastructure* for debugging and experimentation, as well as *incentive structures* that prioritize careful quality control over hasty rollouts. An analogous framework for *data excellence* does not yet exist, bearing the risk of similarly disastrous catastrophes to arise from the use of datasets of unknown or poor quality in AI systems. Consider some *key properties that can pave the way for data excellence in analogy with software engineering excellence*:

Maintainability: Maintaining data at scale has similar challenges as maintaining software at scale. We can follow the lessons learned from software engineering and apply them to data and its maintenance.

Validity: Validating data reveals how well the data helps us explain things related to the phenomena captured by the data, for example, via correlation between the data and external measures. Education research tries to explore whether grades are valid by studying their correlations to external indicators of student outcome. For datasets to have operational validity, we need to know whether they account for the potential complexity, subjectivity, multivalence, or ambiguity of the intended construct; whether they can predict features of the represented construct; and whether the data-collection procedure allows for generalizations within the intended phenomena.

Reliability: This captures internal aspects of data validity, such as

Figure 1. Visual summary of Data Excellence Workshop.

the consistency, replicability, and reproducibility of data. Irreproducible data allows us to draw whatever conclusions we want to draw, while giving us the facade of being data driven, when it is actually dangerously hunch driven. We need reliable mechanisms to account for the human aspects of data collection.

Fidelity: Users of data often assume that the dataset accurately and comprehensively represents the phenomenon, which is almost never the case. Various types of sampling from larger corpora can affect the fidelity of a dataset. For example, temporal splitting can introduce bias if not done right in cases such as signal processing or sequential learning; and user-based splitting not keeping the data of users separated is a potential bias source (e.g., if data from the same user is in test and train sets).

REPORT FROM THE 1ST DATA EXCELLENCE WORKSHOP AT HCOMP 2020

Researchers in human computation (HCOMP) and ML-related fields have had a longstanding interest in human-annotated data for model training and testing. A series of workshops—Meta-

Eval2020 at AAAI, REAIS2019 at HCOMP, SAD2019 at TheWebConf, and SAD2018 at HCOMP—raised awareness about the issues of data quality for ML evaluation and provided a venue for related scholarship. As human-annotated data represents the compass that the entire ML community relies on, data-focused research can potentially have a multiplicative effect on accelerating progress in ML broadly.

The 1st Data Excellence Workshop at HCOMP 2020 provided a platform for an interdisciplinary group of more than 100 industrial and academic researchers from 60 international institutions to discuss and inspire a framework for data excellence in AI systems. At the outset of this workshop, we proposed that *data is “excellent” when it accurately represents a phenomenon, and is: a) collected, stored, and used responsibly, b) maintainable over time, c) reusable across application, and d) exhibits empirical and explanatory power*. Leveraging lessons learned from decades of software engineering inspired an analogous framework for data excellence in AI systems with respect to:

- Identifying properties and defining metrics for data excellence
- Gathering case studies with examples of data excellence and data catastrophes
- Cataloging best practices and incentive structures for data excellence
- Discussing the cost-benefit trade-

We need reliable mechanisms to account for the human aspects of data collection.

Figure 2. Visual overviews of Data Excellence Workshop papers.

off for investments in data excellence

- Cataloging methodologies for the reliability, validity, maintainability, and fidelity of data.

Seven invited talks and three contributed papers covered research and practical experiences from a wide range of industry, academia, and government organizations. Participants contributed case studies of both data catastrophes and excellence, including empirical and theoretical methodologies for the reliability, validity, maintainability, and fidelity of data (Figures 1 and 2).

Invited talks. *Emily Dinan* discussed data challenges for neural dialogue models that may learn and mimic patterns of offensive or otherwise toxic behavior when trained on large unlabeled corpora of human interactions. *Aleksander Mądry* presented results in support of the notion that ML models pick up biases of the real world and our data pipelines. *Quang Duong* introduced the complex task of medical data labeling and highlighted several differences from traditional crowdsourced labeling, including increased task complexity and duration, the requirement of medical expertise in human

graders, and additional personae such as workforce managers, who not only coordinate expert labelers and oversee data quality but also contribute to the development of complex grading guidelines and training of graders. *Andrea Olgiati* discussed the evolution of best practices in software engineering and drew parallels between it and the process of dataset creation for AI applications; for example, unit tests for dataset creation ought to verify data labels' syntax (Are labels in the right format?) and semantics (Do labels have the right meaning and magnitude?).

Ian Soboroff provided a historical overview of datasets and data evaluation methods produced by NIST's Text Retrieval Conference (TREC) and called out the need for "relentless introspection" (i.e., building and scrutinizing a dataset within the community of use). *Peter*

Data intersects software engineering and scientific inquiry.

Hallinan defined data excellence as an optimal trade-off between dataset quality, cost, and speed, given the fixed constraints of privacy and product requirements. He further suggested that dataset quality, cost, and speed are system properties resulting from design levers that can be controlled, including portfolio-level and dataset-level control levers. *Ben Hutchinson* emphasized the need to view datasets as "engineered infrastructure." He introduced a rigorous framework for dataset development transparency that supports decision making and accountability.

Papers. Han et al. [4] define "annotation artifacts" as a type of dataset bias in which annotations capture workers' idiosyncrasies that are irrelevant to the task itself. Their work provided empirical results to address the questions of what factors affect the generation of annotation artifacts and how to reduce annotation artifacts by adjusting workflow design. *Johanne Christensen* and *Benjamin Watson* [5] advocated for new methods of producing labeled training data for machine learning that can discover and support diverse opinions and flexible problem solving

to serve all users equitably. Kapania et al. [6] reported on data practices in high-stakes AI, from interviews with 53 AI practitioners in India, East and West African countries, and the U.S. They also reported on the challenges faced by practitioners with two essential desiderata within high-stakes domains—data reliability and fidelity—and discussed strategies for their improvement and for avoiding data disasters, resulting in safer and robust systems for all.

DISCUSSION

Data as science versus data as engineering. Data intersects software engineering and scientific inquiry. Both interpretations, of data as engineering and as science, emerged as productive metaphors that can teach us about the central questions of data excellence, but possibly point to different directions: engineering is goal focused; science is curiosity focused. Engineering is more strongly represented in industry, science more strongly in academia. Engineering has less theory and more practice; science has more practice and less theory. Discussions highlighted open questions for both perspectives:

Data as science

- How can we as a community address the reproducibility crisis in data collection for AI systems that has implications throughout the life cycle of AI systems, including model design, development, evaluation, and deployment?

- How can we make progress on measuring AI model performance on the underlying task we care about, as opposed to the benchmark itself?

Data as engineering

- How can we best leverage synergies between human judgment and algorithmic approaches like low-shot and unsupervised learning in the process of annotating data?

- How can we formalize data requirements specifications so that they are consistent, repeatable, and well-defined?

- Which stakeholders are most concerned with data fidelity?

Best practices. The workshop discussion also addressed questions of

establishing, communicating, and incentivizing best practices toward data excellence for novices and experienced practitioners alike, and highlighted open questions:

Datasets as living artifacts

- What infrastructure, tools, and practices can facilitate maintaining datasets that are alive (nonstatic) and grow over time?

- For continuously evolving benchmark datasets that are collected with models-in-the-loop, how can we prevent the dataset from drifting away from a reasonable distribution for the original task?

Documentation

- How much dataset documentation is enough, and how is dataset documentation justified when other priorities are competing for attention on a pressing deadline?

- What are best practices in dataset design documents for discussing the biases your dataset might have and how others might help address those biases?

CONCLUSION

Optimizing the cost, size, and speed of data collection has attracted significant attention in the *first-to-market* rush with data for AI.

However, important aspects of the maintainability, reliability, validity, and fidelity of datasets have often been neglected. We argue that we have reached an inflection point in the field of ML in which attention to neglected data quality is poised to significantly accelerate progress. We advocate for research defining and creating processes to achieve *data excellence* and showcase examples, case studies, and methodologies to enable a shift in our research culture to value excellence in data practices in order *to enable the next generation of breakthroughs in ML and AI*.

ENDNOTES

1. Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., and Aroyo, L. “Everyone wants to do the model work, not the data work”: Data cascades in high-stakes AI. *Proc. of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, 2021, Article 39, 1–15.
2. Amershi, S. et al. Software engineering

for machine learning: a case study. *Proc. of the 41st International Conference on Software Engineering: Software Engineering in Practice*. IEEE Press, 2019, 291–300.

3. Aroyo, L. and Paritosh, P. Uncovering unknown unknowns in machine learning. Google AI blog. Feb. 11, 2021; <https://ai.googleblog.com/2021/02/uncovering-unknown-unknowns-in-machine.html>
4. Reducing annotation artifacts in crowdsourcing datasets for natural language processing. Han, D., Kim, J., and Oh, A. 1st Data Excellence Workshop at HCOMP 2020.
5. Christensen, J. and Watson, B. Machine learning training to support diversity of opinion. 1st Data Excellence Workshop at HCOMP 2020.
6. Kapania, S., Sambasivan, N., Olson, K., Highfill, H., Akrong, D., Paritosh, P., and Aroyo, L. Data desiderata: Reliability and fidelity in high-stakes AI. 1st Data Excellence Workshop at HCOMP 2020.

👤 **Lora Aroyo** is a research scientist at Google Research’s New York office currently working on human-labeled data quality. She is a four-time winner of the IBM Faculty Award for her work on CrowdTruth used in adapting the IBM Watson system to the medical domain and in capturing ambiguity in understanding misinformation. Prior to joining Google, she worked at VU Amsterdam as a full professor in computer science and was chief scientist at NY-based start-up Tagasauris.
→ l.m.aroyo@gmail.com

👤 **Matthew Lease** is an associate professor in the School of Information at the University of Texas at Austin. He is also a faculty leader of Good Systems (<http://goodsystems.utexas.edu/>), a UT Austin Grand Challenge to design responsible AI technologies. His research spans information retrieval, crowdsourcing, and natural language processing.
→ ml@utexas.edu

👤 **Praveen Paritosh** is a senior research scientist at Google, leading research on data excellence and evaluation for AI systems. He is also the designer of large-scale human crowdsourcing systems for Freebase and the Google Knowledge Graph.
→ pkp@google.com

👤 **Mike Schaeckermann** is a human-computer information researcher working at the intersection of machine learning, data annotation, and high-stakes domains like medicine. He is currently an applied scientist at Amazon AI. His Ph.D. research at the University of Waterloo was recognized with a Google Ph.D. Fellowship and the Canadian Computer Science Distinguished Dissertation Award.
→ mikeschaeckermann@gmail.com