The Future of Machine Learning Data Practices and Repositories

Workshop summary

Datasets are a central pillar of machine learning (ML) research—from pretraining to evaluation and benchmarking. However, a growing body of work highlights serious issues throughout the ML data ecosystem, including the under-valuing of data work [1, 2], ethical issues in datasets that go undiscovered [3, 4], a lack of standardized dataset deprecation procedures [5, 6], the (mis)use of datasets out-of-context [7], an overemphasis on single metrics rather than holistic model evaluation [8-11], and the overuse of the same few benchmark datasets [7, 9, 12]. Thus, developing guidelines, goals, and standards for data practices is critical; beyond this, many researchers have pointed to a need for a more fundamental culture shift surrounding data and benchmarking in ML [1, 13-17]. At present it is not clear how to mobilize the ML community for such a transformation. In this workshop, we aim to explore this question, including by examining the role of data repositories in the ML data landscape. These repositories have received relatively little attention in this context, despite their key role in the storage, documentation, and sharing of ML datasets. We envision that these repositories, as central purveyors of ML datasets, have the potential to instigate far-reaching changes to ML data and benchmarking culture via the features they implement and the standards they enforce (e.g., minting DOIs, requiring licenses, facilitating the provision of structured metadata).

To address these goals, our proposed workshop aims to facilitate a broad conversation about the impact of machine learning datasets on research, practice, and education—working to identify current issues, propose new techniques, and establish best practices throughout the ML dataset lifecycle. Administrators of machine learning data repositories, including OpenML, HuggingFace Datasets, and the UCI ML Repository, will contribute their perspective on how ML datasets are created, documented, and used and discuss the practical challenges of implementing and enforcing best practices on their platforms. By involving representatives from three major ML repositories and influential researchers from ML, law, governance, and the social sciences, our intent is that this workshop can serve as a catalyst for real positive changes to the ML data ecosystem.

More specifically, topics of interest include (but are not limited to):

- Data repository design and challenges, particularly those specific to ML
- Dataset publication and citation
- FAIR and AI-ready datasets
- Licensing for ML datasets
- ML dataset search and discovery
- Comprehensive data documentation
- Data documentation methods for foundation models

- Data curation and quality assurance
- Best practices for revising and deprecating datasets
- Dataset usability
- Dataset reproducibility
- FAIR ML models
- Benchmark reproducibility
- Holistic and contextualized benchmarking
- Benchmarking and leaderboard ranking techniques
- Overfitting and overuse of benchmark datasets
- Non-traditional/alternative benchmarking paradigms

Invited speakers/panelists

We have recruited a set of speakers with a variety of disciplinary expertise (e.g., machine learning, information sciences, education, law and policy, library and archival sciences) and from a variety of professional avenues (e.g., academia, industry, government, national labs). In particular, we plan to include presentations from the following speakers (10 confirmed interest, 2 invited/pending response):

Block 1: ML Benchmarking & Evaluation

- **Jennifer Wortman Vaughan**, Senior Principal Researcher at Microsoft Research — model evaluation *[confirmed]*
- **Isabelle Guyon,** Director, Research Scientist at Google DeepMind; Professor of Artificial Intelligence at Université Paris-Saclay — AI competitions and benchmarks *[confirmed]*
- **Ce Zhang**, Associate Professor at the University of Chicago — data-centric benchmarking *[confirmed]*

Block 2: Data Curation & Provenance

- **Razvan Amironesei**, Social Scientist at NIST — data genealogy *[confirmed]*
- **Jerone Andrews**, AI Research Scientist at Sony AI, London — responsible data curation *[confirmed]*
- **Shayne Longpre**, Ph.D. candidate at MIT; Lead of the Data Provenance Initiative — data provenance *[confirmed]*

Block 3: Law, Governance, & Policy

- **Jason Schultz**, Professor of Clinical Law at NYU; director of NYU's Technology Law & Policy Clinic — legal considerations for ML, dataset deprecation *[confirmed]*
- **Irene Pasquetto**, Assistant Professor of Information Sciences at the University of Maryland; Senior Research Fellow and Senior Editor at the Shorenstein Center on Media, Politics, and Public Policy — data governance and archival science for ML *[confirmed]*

- **Abeba Birhane**, Senior Fellow in Trustworthy AI at Mozilla Foundation; Assistant Professor at Trinity College Dublin — data auditing *[invited]*

Block 4: Repositories

- **Pieter Gijsbers**, AI Engineer for OpenML; Eindhoven University of Technology — metadata standards *[confirmed]*
- **Lucie Aimee-Kaffee**, Applied Policy Researcher at HuggingFace **—** AI ethics and policy *[confirmed]*
- **Rorie Edmunds**, Samples Community Manager at DataCite Tokyo — standards for trustworthy repositories *[invited]*

Tentative schedule

We intend to provide titles for all talks prior to the event. Each block of invited talks will include three 20-minute presentations (one from each speaker listed in the previous section).

08:45am - 09:00am	Opening Remarks
09:00am - 10:00am	Invited Talks, Block 1: ML Benchmarking & Evaluation
10:00am - 10:15am	*Break*
10:15am - 11:15am	Invited Talks, Block 2: Data Curation & Provenance
11:15am - 12:15pm	Poster Clusters
12:15pm - 01:15pm	*Lunch*
01:15pm - 02:15pm	Invited Talks, Block 3: Law, Governance, & Policy
02:15pm - 02:45pm	Paper Awards & Spotlight Presentations
02:45pm - 03:00pm	*Break*
03:00pm - 04:00pm	Invited Talks, Block 4: Repositories
04:00pm - 05:00pm	Roundtable & Panel Discussion
05:00pm - 05:15pm	Closing Remarks

Poster Clusters

This segment will highlight accepted submissions, providing authors with an interactive platform to showcase their work. To better facilitate discussion, we will form “clusters” of posters based on topic. This clustered format will both help attendees find posters that are most interesting to

them and enable poster presenters to connect with other researchers working on similar topics. Further, each cluster will be assigned an overarching question to prompt big-picture conversations and help identify ideas and questions to consider during the afternoon session presentations and panel discussion.

Paper Awards & Spotlight Presentations

Three submissions will be selected to receive an award (e.g., best paper). In this segment, authors from these three submissions will be presented with award certificates and give 10-minute spotlight talks.

Roundtable & Panel Discussion

We will close our workshop with an interactive roundtable and panel discussion segment. This segment is intended to encourage participants to synthesize insights from the various perspectives highlighted throughout the workshop and to spark discussion about potential next steps. During the first half of this segment, we will form small groups of participants, ensuring that a variety of disciplines and backgrounds are represented in each group. These groups will work to summarize their takeaways and identify questions and discussion points for the panel. To guide the conversation, we will provide discussion prompts to each group, centered on the overall goal of improving the ML data and benchmarking ecosystem. In the second half, we will moderate a panel discussion, involving one representative (i.e., one of the invited speakers) from each block.

Diversity commitment

Data practices and repositories are broadly relevant to individuals working throughout ML and beyond; our intent is to promote this workshop to a broad audience and include participants from a wide range of fields, backgrounds, and seniority levels.

Motivated by the need for a range of viewpoints in these conversations, we aimed for diversity across several dimensions in recruiting committee members and speakers. Both our set of invited speakers and our committee span various seniority levels (junior/mid-level/senior researchers), disciplines (including machine learning, social science, law, policy, philosophy, and information science), and types of affiliations (universities, repositories, and industry/government labs). Our speakers and committee members also bring diverse perspectives in terms of gender, geographic location, and race and ethnicity.

Workshop promotion

We plan to promote our workshop through a variety of channels, leveraging the diverse networks of our organizing committee members. We will publish a workshop website, which we

will link in promotional materials and disseminate via email announcements, relevant mailing lists, personal websites, social media posts, and individual invitations. We will also advertise to the community at upcoming conferences, (e.g., NeurIPS Datasets and Benchmarks 2024) and via relevant interest groups (e.g., FAIR4ML).

Previous related workshops

Navigating and Addressing Data Problems for Foundation Models (DPFM) at ICLR 2024 considers several similar issues (data curation and quality, governance, ethical ML) in the context of training data for foundation models. In addition, previous workshops on data-centric AI, including Data-Centric AI at NeurIPS 2021, DataPerf at ICML 2022, Data-centric Machine Learning Research (DMLR) at ICML 2023, and DMLR at ICLR 2024, have initiated important conversations about the critical role of data in AI, including topics such as data governance, AI alignment, and benchmarking techniques. Our proposed workshop carries forward the momentum of these previous workshops, expanding on these conversations by developing connections with data repository administrators and experts—which we believe can and will play a critical role in enforcing best practices and addressing current challenges.

Anticipated audience size

Based on the attendance at similar workshops (listed in the previous section) in recent years, we expect over 100 in-person attendees.

Virtual access to workshop materials and outcome

This will be an in-person workshop, but we are committed to ensuring that all participants (across time zones, limited by visa issues, etc.) are able to access the contents of our workshop. We will record all of the events throughout the day and post them to the website. Papers and posters will also be made available on the website following the workshop. For exceptional circumstances in which speakers or poster presenters cannot be present in person, we will provide accommodations, such as streaming their talk on Zoom or displaying their poster on their behalf.

Submissions and reviewing

Paper submissions may be of any length and any format. Accepted papers will be non-archival and shared on the workshop website. We will follow ICLR's recommended timeline: paper

submissions will be due February 3, 2025 and decisions will be released on or before March 5, 2025.

The reviewing process will strictly adhere to the ICLR workshop guidelines. To mitigate potential biases and conflicts of interest, we will conduct a double-blind review process for submitted papers and avoid matching reviewers and submissions with shared affiliations. In making decisions, we will strive for a wide representation of topics and research areas, helping to recruit a diverse set of workshop participants. We will prioritize submissions of ongoing or novel work over those that have previously been published.

Program committee members will be recruited from a variety of disciplines and backgrounds. We aim to provide each submission with at least two high-quality reviews.

Organizers and biographies

Markelle Kelly (co-organizer, University of California Irvine)

- **Email**: kmarke@uci.edu
- **Webpage**: https://markellekelly.com
- **Organizational experience**: first-time workshop organizer
- **Biography**: Markelle Kelly is a computer science Ph.D. candidate in the DataLab group at the University of California, Irvine, advised by Dr. Padhraic Smyth. Her research aims to understand and improve human-AI interaction, as well as to develop tools for more interpretable, accountable and data-centric machine learning. Markelle is a member of the Steckler Center for Responsible, Ethical, and Accessible Technology, the Irvine Initiative in AI, Law, and Society, and the HPI Research Center in Machine Learning and Data Science at UCI. She also serves as a curator and librarian for the UCI Machine Learning Repository. Previously, she has interned with eBay and Apple, working broadly on tools and processes for ML model evaluation, and at Project Jupyter as a software engineer. She received her Bachelor's degree in statistics from California Polytechnic State University, San Luis Obispo in 2020 and her Master's degree in computer science from UCI in 2022.

Rachel Longjohn (co-organizer, University of California Irvine)

- **Email**: rlongjoh@uci.edu
- **Webpage**: https://rlongjohn.github.io/
- **Organizational experience**: first-time workshop organizer
- **Biography**: Rachel Longjohn is a Statistics Ph.D. candidate at the University of California, Irvine (UCI) advised by Dr. Padhraic Smyth. Her research interests include forensic statistics, uncertainty quantification, and algorithmic evaluation, explainability, and fairness in machine learning. She is also interested in supporting efforts in data curation and reproducible research and serves as a curator and librarian for the UCI

Machine Learning Repository. She received her Master's degree in Statistics from UCI in 2021, and her Bachelor's degree in Applied and Computational Mathematics, with a specialization in Computer Programming, from the University of Southern California in 2019.

Meera Desai (co-organizer, University of Michigan)

- **Email**: madesai@umich.edu
- **Webpage**: https://meera-desai.com/
- **Organizational experience**: Co-organizer of Women and Machine Learning Workshop at NeurIPS (2021)
- **Biography**: Meera Desai is a fourth year Ph.D. candidate at the University of Michigan School of Information where she is advised by Abigail Jacobs and Dallas Card. She is an affiliate of the Center for Ethics, Society, and Computing. Prior to University of Michigan, Meera was a research fellow at the American Museum of Natural History, and received a B.A. in physics from Barnard College. Her research focuses on the responsible development and use of natural language processing systems. In particular, she is interested in measurement in data practices and evaluation for large language models.

Shivani Kapania (co-organizer, Carnegie Mellon University)

- **Email**: kapania@cmu.edu
- **Webpage**: https://www.shivanikapania.com/
- **Organizational experience**:Co-organizer of the CRAFT session "User Engagement in Algorithm Testing and Auditing: Exploring Opportunities and Tensions between Practitioners and End Users" at FAccT 2023
- **Biography**: Shivani Kapania is a Ph.D. student at the Human-Computer Interaction Institute at Carnegie Mellon University, School of Computer Science, where she works with Sarah Fox on creating mechanisms for accountability and worker agency in interactions with machine learning systems. Previously, she worked as a Research Intern with the FATE group at Microsoft Research, as a Fellow at Weizenbaum Institute, and as a Pre-Doctoral Researcher with the Google Research India and the Technology, AI, Society and Culture (TASC) team, focusing on mixed-methods research interrogating the practices of dataset production.

Maria Antoniak (co-organizer, University of Copenhagen)

- **Email**: maria.antoniak@colorado.edu
- **Webpage**: https://maria-antoniak.github.io/
- **Organizational experience**: Co-Organizer for tutorials at FAccT and ICWSM, Workshops Co-Chair for ICWSM 2024, Publicity Co-Chair for FAccT 2025, Ethics Co-Chair for NAACL 2024 and 2025.
- **Biography**: Maria Antoniak is a Postdoc at the Pioneer Centre for AI at the University of Copenhagen and an incoming Assistant Professor in Computer Science at the University of Colorado Boulder. Previously, she was a Young Investigator at the Allen Institute for

AI, and she completed her PhD in Information Science at Cornell University and her master's degree in Computational Linguistics at the University of Washington. She has also spent time at places like ETH Zurich, Microsoft Research FATE, Twitter Cortex, and Facebook Core Data Science. Her work focuses on natural language processing and cultural analytics.

Padhraic Smyth (co-organizer, University of California Irvine)

- **Email**: smyth@ics.uci.edu
- **Webpage**: https://ics.uci.edu/~smyth/index.html
- **Organizational experience**: Program Chair, ACM SIGKDD Conference 2011; Program Chair Uncertainty in AI Conference 2013; Associate Program Chair, IJCAI 2022; General Chair, AI and Statistics Conference, 1997; Tutorials Chair, AAAI Conference 1998; Tutorials Chair ACM SIGKDD Conferences 1995 and 1996; co-Organizer for Dagstuhl School on Automating Data Science, 2018; co-Organizer for Workshop on Algorithmic and Statistical Approaches for Large Social Network Data Sets, NeurIPS 2012; co-Organizer for Workshop on Temporal and Spatial Machine Learning, ICML 2001.
- **Biography**: Padhraic Smyth is a Distinguished Professor in the Department of Computer Science at UC Irvine, with joint appointments in the Department of Statistics and in the Department of Education. His research interests include machine learning, artificial intelligence, pattern recognition, and applied statistics and he has published over 200 papers on these topics. He is an ACM Fellow, IEEE Fellow, AAAI Fellow and AAAS Fellow, and was a recipient of the ACM SIGKDD Innovation Award.He served as program chair and in various senior/area chair positions for a variety of conferences in machine learning and AI. He has also served in editorial and advisory positions for journals such as the Journal of Machine Learning Research, the Journal of the American Statistical Association, and the IEEE Transactions on Knowledge and Data Engineering.

Sameer Singh (advisor, University of California Irvine)

- **Email**: sameer@uci.edu
- **Webpage**: https://sameersingh.org/
- **Organizational experience**: ACL Workshop on Knowledge Graphs and Large Language Models (KaLLM), 2024; Deep Learning Day, Knowledge Discovery and Data Mining (KDD), 2020; NeurIPS Workshop on Knowledge Representation & Reasoning Meets Machine Learning (KR2ML), 2019; 1st AKBC Workshop on Knowledge Bases and Multiple Modalities (KBMM), 2019; NAACL Workshop on "Automated Knowledge Base Construction", 2016; AAAI Workshop on "Declarative Learning Based Programming", 2016; NeurIPS Workshop on "Machine Learning Systems (LearningSys)", 2015; NeurIPS Workshop on "Automated Knowledge Base Construction", 2014; CIKM Workshop on "Automated Knowledge Base Construction", 2013; NeurIPS Workshop on "Big Learning", 2011, 2012, 2013; ICML Workshop on "Inferning": Interactions between Inference and Learning, 2012, 2013.

- **Biography**: Sameer Singh is a Professor of Computer Science at the University of California, Irvine (UCI). He is working primarily on robustness and interpretability of machine learning algorithms, along with models that reason with text and structure for natural language processing. Sameer was a postdoctoral researcher at the University of Washington and received his PhD from the University of Massachusetts, Amherst, during which he interned at Microsoft Research, Google Research, and Yahoo! Labs. He has received the NSF CAREER award, selected as a DARPA Riser, UCI ICS Mid-Career Excellence in research award, and the Hellman and the Noyce Faculty Fellowships. His group has received funding from Allen Institute for AI, Amazon, NSF, DARPA, Adobe Research, Hasso Plattner Institute, NEC, Base 11, and FICO. Sameer has published extensively at machine learning and natural language processing venues, including paper awards at KDD 2016, ACL 2018, EMNLP 2019, AKBC 2020, and ACL 2020.

Joaquin Vanschoren (advisor, Eindhoven University of Technology)

- **Email**: j.vanschoren@tue.nl
- **Webpage**: https://joaquinvanschoren.github.io/home/
- **Organizational experience**: Joaquin has served as a workshop chair for NeurIPS Workshop on Meta-Learning 2018-2021, NeurIPS Workshop on Data-Centric AI 2021, AAAI Workshop on Meta-Learning 2021, ICML Workshop on Automatic Machine Learning 2016-2021, DALI Workshop on The Data Science Process 2017, ECMLPKDD Workshop on Automatic Machine Learning 2017, ECMLPKDD Workshop on Meta-Learning and Algorithm Selection 2015, ECAI Workshop on Meta-Learning and Algorithm Selection 2014, ECMLPKDD Workshop on Learning from Unexpected Results 2012, and ECAI Workshop on Planning to Learn 2012. He also chaired the NeurIPS datasets and benchmarks track from 2021 to 2023.
- **Biography**: Joaquin Vanschoren is an associate professor at TU Eindhoven and lead of the Automated Machine Learning lab. He aims to deeply understand, explain, and democratize AI for the benefit of all humanity. He and his team build AI systems that learn continually and automatically assemble themselves to learn faster and better, often inspired by the human brain. His work can be found in top AI conferences and journals. He founded OpenML, a leading open science platform for machine learning, to streamline and accelerate reproducible AI research. He was the inaugural chair of the NeurIPS Datasets and Benchmarks track and is editor-in-chief of the DMLR journal, to incentivize and reward good data and evaluation practices in AI. He also chairs the MLCommons AI Safety working group to help make AI models safer through science, am one of the founders of the Croissant standard for sharing AI resources, and contribute to data-centric AI. He is a founding member of the European AI societies ELLIS and CLAIRE, authored the first book on AutoML, gave tutorials at NeurIPS and AAAI, won several awards (including the Dutch Data Prize and Amazon Research Award), and he has been interviewed for news articles in Nature, Science, and podcasts.

Amy Winecoff (advisor, Center for Democracy & Technology)

- **Email**: awinecoff@cdt.org

- **Webpage**: https://cdt.org/staff/amy-winecoff/
- **Organizational experience**: Amy recently organized a covening of industry, government, and civil society stakeholders to discuss the challenges and opportunities of AI documentation for robust governance.
- **Biography**: Amy Winecoff serves as a Senior Technologist in the AI Governance Lab at the Center for Democracy & Technology (CDT), where her work centers on creating technically-informed, practical strategies to enhance AI regulation and governance, aimed at protecting the interests of individuals impacted by AI systems. Her research emphasizes building the foundations for robust governance, particularly in the areas of AI documentation and measurement. Her work has been featured in academic venues like RecSys, CHI, AIES, and First Monday, as well as in policy-focused publications through CDT and Tech Policy Press. She has also served as a responsible technology advisor for startup accelerators, and in her prior roles as a data scientist, she developed and deployed production recommendation systems for e-commerce companies. She has also spent time as a research fellow at Princeton University's Center for Information Technology Policy (CITP) and as an assistant professor at Bard College. Amy received her PhD in Psychology and Neuroscience from Duke University, which allows her to bring a social science lens to her work on AI governance.

Daniel S. Katz (advisor, University of Illinois Urbana-Champaign)

- **Email**: dskatz@illinois.edu
- **Webpage**: https://danielskatz.org/
- **Organizational experience**: Dan has been involved in the organization of almost 400 conferences and workshops in positions from general chair and program chair to member of the program committee.
- **Biography**: Dan is Chief Scientist at the National Center for Supercomputing Applications (NCSA), Research Associate Professor in the Siebel School of Computing and Data Science, and Research Associate Professor in the School of Information Sciences (iSchool) at the University of Illinois Urbana-Champaign. He is also a Better Scientific Software (BSSw) Fellow. He was previously a Senior Fellow in the Computation Institute (CI) at the University of Chicago and Argonne National Laboratory and Guest Faculty at Argonne, a Program Director in the Division of Advanced Cyberinfrastructure at the National Science Foundation, Director for Cyberinfrastructure Development at the Center for Computation & Technology (CCT) at Louisiana State University (LSU), Principal Member of the Information Systems and Computer Science Staff and Supervisor of the Parallel Applications Technologies group at JPL, and Computational Scientist at Cray Research, Inc.

References

[1] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades

in High-Stakes AI. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). https://doi.org/10.1145/3411764.3445518

[2] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). https://doi.org/10.1145/3442188.3445918

[3] Kenneth L. Peng, Arunesh Mathur, and Arvind Narayanan. 2021. Mitigating dataset harms requires stewardship: Lessons from 1000 papers. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2). https://doi.org/10.48550/arXiv.2108.02922

[4] Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision?. In 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). https://doi.org/10.1109/WACV48630.2021.00158

[5] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. In Patterns, 2(11). https://doi.org/10.1016/j.patter.2021.100336

[6] Alexandra Sasha Luccioni, Frances Corry, Hamsini Sridharan, Mike Ananny, Jason Schultz, and Kate Crawford. 2022. A Framework for Deprecating Datasets: Standardizing Documentation, Identification, and Communication. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). https://doi.org/10.1145/3531146.3533086

[7] Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the everything in the whole wide world benchmark. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2). https://doi.org/10.48550/arXiv.2111.15366

[8] Rachel L. Thomas and David Uminsky. 2022. Reliance on metrics is a fundamental challenge for AI. In Patterns, 3(5). https://doi.org/10.48550/arXiv.2002.08512

[9] Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. 2021. The benchmark lottery. arXiv preprint. https://doi.org/10.48550/arXiv.2107.07002

[10] Ryan Burnell, Wout Schellaert, John Burden, Tomer D. Ullman, Fernando Martinez-Plumed, Joshua B. Tenenbaum, Danaja Rutar, Lucy G. Cheke, Jascha Sohl-Dickstein, Melanie Mitchell, Douwe Kiela, Murray Shanahan, Ellen M. Voorhees, Anthony G. Cohn, Joel Z. Leibo, and Jose Hernandez-Orallo. 2023. Rethink reporting of evaluation results in AI. In Science 380(6641). https://doi.org/10.1126/science.adf6369

[11] Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the Eye of the User: A Critique of NLP Leaderboards. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). https://doi.org/10.18653/v1/2020.emnlp-main.393

[12] Bernard Koch, Emily Denton, Alex Hanna, and Jacob G. Foster. 2021. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1. https://doi.org/10.48550/arXiv.2112.01716

[13] Katy Ilonka Gero, Payel Das, Pierre Dognin, Inkit Padhi, Prasanna Sattigeri, and Kush R. Varshney. 2023. The incentive gap in data work in the era of large models. In Nature Machine Intelligence 5(6). https://doi.org/10.1038/s42256-023-00673-x

[14] Meera A. Desai, Irene V. Pasquetto, Abigail Z. Jacobs, and Dallas Card. 2024. An archival perspective on pretraining data. In Patterns, 5(4). https://doi.org/10.1016/j.patter.2024.100966

[15] Amy K. Heger, Liz B. Marquis, Mihaela Vorvoreanu, Hanna Wallach, and Jennifer Wortman Vaughan. 2022. Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata. https://doi.org/10.1145/3555760

[16] Jerone Andrews, Dora Zhao, William Thong, Apostolos Modas, Orestis Papakyriakopoulos, and Alice Xiang. 2024. Ethical considerations for responsible data curation. In Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS '23). https://doi.org/10.48550/arXiv.2302.03629

[17] Stella Biderman and Walter J. Scheirer. 2020. Pitfalls in Machine Learning Research: Reexamining the Development Cycle. In Proceedings of Machine Learning Research. https://doi.org/10.48550/arXiv.2011.02832