Data-centric Machine Learning Research (DMLR): Harnessing Momentum for Science

Workshop Summary

For years, curated data has been instrumental in propelling the field of machine learning forward. While the spotlight has often been on the development of machine learning models, there is a growing recognition of the critical role that data quality plays [1, 2]. Building on the momentum of our previous workshops—Data-Centric AI at NeurIPS 2021, DataPerf at ICML 2022, and Data-centric Machine Learning Research (DMLR) at ICML 2023—we aim to continue fostering a collaborative environment for exploring the critical role of data in machine learning. The momentum within the community is evident, with the NeurIPS Datasets and Benchmark track receiving around 1,000 submissions, nearly doubling the total from last year. This surge underscores the growing recognition of the significance of data quantity and quality. We intend to sustain this momentum by discussing the accomplishments within the DMLR ecosystem, encompassing the growing community, infrastructure evolution, and the initiatives surrounding the next generation of public datasets for both research and societal benefits. While maintaining this momentum is our primary goal, the fourth installment in our series of workshops will have a secondary focus on DMLR in the realm of science research and applications.

AI for Science: Unlike general AI, AI for Science uses AI to tackle unique scientific challenges, uncover rare phenomena, deepen our understanding of scientific domains, and accelerate discoveries. The traditional model-centric AI approach primarily focuses on algorithmic improvements and often overlooks the foundational role of data. This is particularly problematic in scientific contexts where there are strong emphases on both prediction from and explanation of data. In science, ML pipelines are often interwoven with the inputs and outputs of theoretical models which depend on robust data and reliable model outputs. In large science projects and missions that produce vast amounts of data (e.g., at CERN [3], NASA), efficient data-centric AI frameworks are essential for maximizing the potential of expensive experiments and missions. In contrast, biomedical research surfaces different data-centric AI issues surrounding sparsity (e.g., to understand OOD cases, estimate causal effects), privacy, and fairness. While there's an increasing body of literature on data-centric AI, this momentum seems slower within the scientific community. The significance of a data-centric AI framework in science is manifold:

- Scientific research is inherently data-driven.
- The integrity of AI systems is intrinsically tied to the quality of their training data. The high stakes in science leave no room for errors due to poor data. For scientists to trust AI systems, data quality, including precise labeling and comprehensive coverage, is vital.
- The vastness of scientific data demands robust data management for consistent future model evaluations.
- Ethical considerations, such as data privacy, biases, and diverse representation, are central to scientific research.
- Science domain experts provide insights that models alone cannot, ensuring data aligns with scientific objectives and bolsters research reliability.
- There is a need for strong norms around rigorous, data-driven machine learning akin to those surrounding mathematical and statistical modeling.

The purpose of this workshop is to create a forum for addressing these crucial topics. We plan to explore a wide range of subjects, from the creation and evaluation of datasets and benchmarks to the development of specialized tools, infrastructure, and governance models [4, 5, 6]. Additionally, we aim to delve into foundational research concerning data quality, data and concept shifts, and acquisition strategies. Our ultimate goal is to cultivate a collaborative environment that brings together a diverse array of researchers, practitioners, domain experts, data and platform providers, and engineers, all of whom are tackling pressing data-related challenges in science research and applications. Topics will include, but are not limited to:

- Data collection and benchmarking techniques
- Data governance frameworks for ML
- Impact of data bias, variance, and drifts
- Role of data in foundation models: pre-training, prompting, fine-tuning
- Optimal data for standard evaluation framework in the context of changing model landscape
- Domain specific data issues
- Data-centric explainable AI

- Data-centric approaches to AI alignment
- Active learning, Data cleaning, acquisition for ML

Modality: hybrid

Session organization: virtual + in-person engagement

We aim at a discussion-centric workshop to allow for in-depth coverage of state-of-art and work-in-progress efforts and panel discussion and poster presentation along the data lifecycle in machine learning research and engineering: creation, quality and processing, governance and management/infrastructure. We will use Slido to gather questions from the virtual audience. Each session will conclude with discussions summarized and published by Scribe. To facilitate participation of people unable to travel, all workshop contents will be available online, and during the workshop day we will support various forms of hybrid presentations and discussions, including organized breakout rooms for virtual attendees during the networking sessions and poster sessions. The workshop will be organized in four components:

- Keynotes and invited talks
- Open panel discussions
- Poster sessions
- Networking sessions

Tentative Schedule:

9:00 - 9:15	Introduction and Opening
9:15 - 10:00	Keynote 1: *James Zou*, Assistant Professor, Stanford University (Confirmed)
10:00 - 10:35	Invited Talk 1: *Marzyeh Ghassemi*, Assistant Professor, MIT (Invited)
10:35 - 11:00	Coffee / networking break
11:00 - 12:00	Panel 1: *Data 2024: What are the important research questions for the DMLR community in light of foundation models?*
12:00 - 1:00	Lunch Break
1:00 - 1:45	Keynote 2: *Baharan Mirzasoleiman*, Assistant Professor, UCLA (Confirmed)
1:45 - 2:20	Invited Talk 2: *Abdulmotaleb El Saddik*, University of Ottawa, (Confirmed)
2:20 - 2:45	Coffee / networking break
2:45 - 3:15	Announcements: • DataPerf + Dynabench announcements • Challenge results • Update on DMLR Journal • Croissant data format
3:15 - 4:15	Panel 2: *Generative AI 2024: What are the power applications of generative AI and what data needs do they have?*
4:15 - 5:15	Poster session
5:15 - 5:30	Concluding remarks
7:00 - 9:00	Post-workshop social event

Invited Speakers

- James Zou, Assistant Professor, Stanford University
- Baharan Mirzasoleiman, Assistant Professor, UCLA
- Marzyeh Ghassemi, Assistant professor, MIT
- Abdulmotaleb El Saddik, University of Ottawa

Panelists

- Feiyang Kang, Assistant Professor, Virginia Tech University
- Ece Kama, Researcher, Microsoft
- Sujit Roy, Researcher, University of Alabama in Huntsville
- Elena Simperl, Professor, King's College London

- Tim Salimans, Researcher, Google
- Anna Khoreva, Research Group Leader, Bosch Center for AI
- Thomas Sutter, Researcher, ETH Zürich
- Tianlong Chen, Assistant Professor, UNC Chapel Hill

Anticipated audience size

We expect 100 to 150 in-person attendees and 30 to 50 virtual attendees. These estimates are based on attendance for prior iterations of the workshop at large international conferences such as ICML and NeurIPS.

Plan to get an audience for a workshop (advertising, reaching out, etc...)

We intend to connect with a more diverse group of people through a variety of channels including social media platforms, academia email lists, professional networks, personalized invitations, Discord servers and Google groups with special interest in machine learning research and development. Part of our target audience is researchers with cross-disciplinary interest intersecting data-centric research with science fields. As DMLR encourages diversity, we also target women-in-science audiences to ensure a more equitable scientific data-centric community. We plan to send our announcements, advertisements and invitations in different language versions and make it more welcoming and inclusive for the diverse audience we are targeting. We will offer incentives to encourage individuals to participate and attend the DMLR workshop. These incentives may include prizes for best paper/poster award, invitation of selected papers to contribute to DMLR journal, exclusive offers/discounts or registration-fee scholarships for participating students to attend. For students interested in attending the DMLR workshop, we will include the information of the student volunteer and D&I subsidies program in our announcements and invitations. We will periodically promote our workshop through media advertisements, we will include engaging visuals to help capture people's attention. The advertisements may include elements like infographics of past DMLR workshops success, appealing landmark scenes of the hosting city, speakers' pictures/biographies, AI-powered science applications published in the past DMLR workshops, and sponsors.

Diversity commitment

During the selection of organizers and speakers, we actively encouraged all forms of diversity. We invited participation for the organizing committee through open forums, meetings and mailing lists. The final selection of organizers and speakers encompasses individuals with diverse gender, ethnicity, affiliations, nationality, career level, and scientific background. Specifically with regard to career level, we have ensured that our structure supports the development and engagement of more junior researchers by including several early career researchers as invited speakers on the panels. Our commitment to diversity and inclusion extends beyond just one event; it's a continuous effort that shapes all our upcoming initiatives, as demonstrated by the ever-evolving roster of organizers, speakers, and sponsored students. In the last iteration of the workshop, we donated three free workshop registration to indigenous Hawaiians and affinity groups.

Access

The workshop will take place in hybrid format, most likely with sessions in Zoom and a poster session in Gather Town. This format allows for the inclusion of remote attendees, enabling them to present their work alongside in-person participants. As with the previous iteration of the workshop, all accepted papers will be published and archived on the workshop webpage located at dmlr.ai/iclr2024. Recordings of the talks will be made accessible either through the SlidesLive platform or alternatively by directly visiting the workshop webpage. Moreover, a selected number of papers will be invited to contribute their research to the newly established DMLR journal.

Previous related workshops

This workshop proposal extends the triumph of prior Data-Centric AI Workshops that we have organized at NeurIPS2021, ICML2022, and ICML2023, aiming to showcase the prevailing state-of-the-art. Each of these prior versions had over 100 participants and 50 submitted presentations. It also carries forward a legacy of successful workshops centered around the significance of data in AI. Some other previous related workshops are:

- Data-Centric AI (DCAI) @ NeurIPS 2021

- Data Excellence (DEW) @ HCOMP 2020
- Evaluating Evaluation of AI Systems (Meta-Eval) @ AAAI 2020
- Rigorous Evaluation of AI Systems (REAIS) @ HCOMP 2020 and 2019

References

[1] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo, “Everyone wants to do the model work, not the data work: Data cascades in high-stakes AI,” in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–15.

[2] E. Strickland, “Andrew Ng, AI minimalist: The machine-learning pioneer says small is the new big,” *IEEE Spectrum*, vol. 59, no. 4, pp. 22–50, 2022.

[3] CERN, “Cern accelerating science,” CERN, (Accessed on 10/19/2023). [Online]. Available: https://home.cern/

[4] L. Aroyo, M. Lease, P. Paritosh, and M. Schaekermann, “Data excellence for AI: why should you care?” *Interactions*, vol. 29, no. 2, pp. 66–69, 2022.

[5] J. Vanschoren, J. N. Van Rijn, B. Bischl, and L. Torgo, “OpenML: networked science in machine learning,” *ACM SIGKDD Explorations Newsletter*, vol. 15, no. 2, pp. 49–60, 2014.

[6] C. Coleman, C. Yeh, S. Mussmann, B. Mirzasoleiman, P. Bailis, P. Liang, J. Leskovec, and M. Zaharia, “Selection via proxy: Efficient data selection for deep learning,” *The International Conference on Learning Representations*, 2020.

Organizers.
This proposal brings together a very diverse group of organizers from several data-related communities including DataPerf, MLCommons, Data-centric AI Workshops, and Dataset 2030.

Contact: Manil Maskey and Lilith Bat-Leah

Manil Maskey (he/him) is a Senior Research Scientist and Project Manager at NASA. He leads the advanced concepts team for the Inter-Agency Implementation and Advanced Concepts Team (IMPACT), where he is developing innovative data systems and data-driven solutions to challenging science problems. He also leads the NASA Science Mission Directorate's Artificial Intelligence Program, where he is transforming NASA's science data into actionable data for machine learning. He has focused on research and application projects in the area of data systems, cloud computing, machine learning, computer vision, and visualization. He is an affiliate faculty at the University of Alabama in Huntsville, a senior member of the Institute of Electrical and Electronics Engineers (IEEE), chair of the IEEE Geoscience and Remote Sensing Society (GRSS) Earth Science Informatics Technical Committee, member of American Geophysical Union (AGU) and AGU Fall Meeting Planning Committee, member of European Geosciences Union (EGU), and member of Association for Advancement of Artificial Intelligence (AAAI). Previously, he has organized several AI workshops on using machine learning for geosciences.

- Email: `manil.maskey@nasa.gov`
- Web page: `https://manilmaskey.github.io`
- Google Scholar: `https://scholar.google.com/citations?user=k4T40hoAAAAJ&hl=en`

Alicia Parrish (she/her) is a research scientist on the Responsible AI team at Google. She received her PhD in linguistics from New York University in 2022, where she worked at the intersection of experimental linguistics, cognitive neuroscience, and NLP. Her research focuses on crowdsourcing methods, adversarial data collection, dataset evaluation, and human disagreements. She served on the program committee for the Linguistics Society of America (LSA) 2019-2022, co-organized New Ways of Analyzing Variation (NWAV) 2018 and the DMLR Workshop at ICML 2023, and she co-organized the Inverse Scaling Prize public competition and Adversarial Nibbler Challenge.

- Email: `aliciaparrish@google.com`
- Web page: `https://aliciaparrish.com/`
- Google Scholar: `https://scholar.google.com/citations?hl=en&user=Kze5eGkAAAAJ`

Chanjun Park (he/him) is a researcher in the field of Natural Language Processing (NLP), with a focus on Data-Centric AI, Machine Translation and Large Language Model (LLM). He is currently working as an Technical Leader (TL) at Upstage LLM Team. In 2023, he received Ph.D. from Korea University under the supervision of Professor Heuiseok Lim for the work on "Data-Centric Neural Machine Translation". From 2018 to 2019, he worked at SYSTRAN as a Research Engineer. Chanjun is the founder and leader of the KU-NMT Group, and has received the Naver Ph.D. Fellowship in 2021. He served as the Virtual Social Chair at COLING 2022, and is currently serving as the Program Chair for the WiNLP Workshop. He has published more than 160 papers in the field of NLP.

- Email: `bcj1210@naver.com`
- Web page: `https://parkchanjun.github.io/`
- Google Scholar: `https://scholar.google.com/citations?user=085jNAMAAAAJ&hl=en`

Xiaozhe Yao (he/him) is a PhD student at ETH Zurich, with research interests spanning from Large-scale machine learning to Data-Centric systems, his research direction is to democratize machine learning and make it accessible to a wider range of audience. Prior to that, he was an innovation fellow at the Library Lab, ETH Zurich where he led projects towards accessible, usable and scalable AI. He co-organized the data cleaning challenge at DataPerf in 2022-2023.

- Email: `xiaozhe.yao@inf.ethz.ch`
- Web page: `https://yao.sh/`
- Google Scholar: `https://scholar.google.com/citations?user=Bhgm1tQAAAAJ&hl=en`

Holger Caesar (he/him) is an Assistant Professor at TU Delft in the Netherlands, working on perception, prediction and planning for autonomous vehicles. He received a PhD from the University of Edinburgh and studied at ETH, EPFL and KIT. Holger created the data team of one of the leading robotaxi companies and his research interests focus on efficient dataset creation and learning from limited data. His datasets COCO-Stuff, nuScenes and nuPlan have had a big impact on their respective fields and have been cited over 4500 times.

- Email: h.caesar@tudelft.nl
- Web page: https://it-caesar.com
- Google Scholar: https://scholar.google.com/citations?user=373LKEYAAAAJ&hl=de

Bernard Koch (he/him) is an Assistant Professor of Sociology at the University of Chicago and Postdoctoral Fellow at the Kellogg School of Management, Northwestern University. His research characterizes how data is created and used across MLR and science, with a particular focus on epistemic and ethical repercussions. His work on increasing concentration on fewer benchmarks across MLR won a best paper award at the NeurIPS 2021 DS&B track. He also has methodological interests in graph learning and deep causal estimation.

- Email: bernard.koch@kellogg.northwestern.edu
- Web page: bernardjkoch.com
- Google Scholar: https://scholar.google.com/citations?user=cBc_sIEAAAAJ&hl=en

Fatimah Alzamzami (she/her) is a PhD candidate at the University of Ottawa with research interests in data-centric AI, NLP and smart cities. Her research focuses on multi-lingual-dialectal online social behavior analysis. She is a member of MCRLAB where she led multiple projects including smart city dataset curation, informal multi-lingual to multi-dialectal Arabic machine translation, data exploration and interpretation pandemic-friendly system, and City Digital Pulse (CDP) system for real-time online affective analysis.

- Email: falza094@uottawa.ca
- Web page: https://www.linkedin.com/in/fatimah-alzamzami-6b11035/
- Google Scholar: https://scholar.google.com/citations?user=RweX7igAAAAJ&hl=en&oi=ao

Zhangyang "Atlas" Wang (he/him) is a tenured Associate Professor and holds the Temple Foundation Endowed Faculty Fellowship #7, in the Department of Electrical and Computer Engineering at The University of Texas at Austin. He is also a faculty member of UT Computer Science (GSC), and the Oden Institute. Prof. Wang has broad research interests spanning from the theory to the application aspects of machine learning (ML). At present, his core research mission is to leverage, understand and expand the role of low-dimensionality, from classical optimization to modern neural networks, whose impacts span over many important topics such as: efficient scaling, training and inference of large language models (LLMs); robustness and trustworthiness; generative AI; and graph learning. Prof. Wang co-founded the new Conference on Parsimony and Learning (CPAL) and serves as its inaugural Program Chair. He is an elected technical committee member of IEEE MLSP and IEEE CI; and regularly serves as area chairs, invited speakers, tutorial/workshop organizers, various panelist positions and reviewers. He is an ACM Distinguished Speaker and an IEEE senior member.

- Email: atlaswang@utexas.edu
- Web page: https://vita-group.github.io/
- Google Scholar: https://scholar.google.com/citations?user=pxFyKAIAAAAJ&hl=en

Jerone Andrews (he/him) is a Research Scientist at Sony AI (Tokyo) within its AI Ethics flagship project. His current research centers on human-centric computer vision, in particular responsible data curation, human-centric representation learning, as well as bias detection and mitigation. Prior to joining Sony, he received an MSci in mathematics from King's College London, which he followed with an EPSRC-funded MRes and Ph.D. in computer science at University College London (UCL). Subsequently, he was awarded a Royal Academy of Engineering Research Fellowship, a British Science Association Media Fellowship with BBC Future, and a Marie Skłodowska-Curie RISE grant. While at UCL, he also spent time as a Visiting Researcher at the National Institute of Informatics (Tokyo) and Telefónica Research (Barcelona).

- Email: jerone.andrews@sony.com
- Web page: https://ai.sony/people/Jerone-Andrews/
- Google Scholar: https://scholar.google.com/citations?user=cEr1ouIAAAAJ&hl=en

Lilith Bat-Leah (she/her) is Vice President, Data Services at Mod Op, responsible for consulting on use cases for data analytics, data science, and machine learning. Lilith has over 11 years of experience managing, delivering, and consulting on identification, preservation, collection, processing, review, annotation, analysis, and production of data in legal proceedings. She also has experience in research and development of machine learning software for eDiscovery. She speaks and writes about various topics in eDiscovery, such as evaluation of machine learning systems, ESI protocols, and discovery of databases. Lilith holds a BSGS in Organization

Behavior from Northwestern University, where she graduated magna cum laude. She is a current co-chair of the DataPerf/DynaBench MLCommons working group and formerly served as co-trustee of the EDRM Analytics and Machine Learning project, as a member of the EDRM Global Advisory Council, as Vice President of the Chicago ACEDS chapter, and as President of the New York Metro ACEDS Chapter.

- Email: `lilith.bat-leah@modop.com`
- Web page: `https://dprism.com/about/lilith-bat-leah/`
- Google Scholar: `https://www.jdsupra.com/authors/lilith-bat-leah/`

Praveen Paritosh (he/him) is a senior research scientist at Google, leading research on data excellence and evaluation for AI systems. He designed the large-scale human curation systems for Freebase and the Google Knowledge Graph. Praveen is the co-chair of the DataPerf working group. He was the co-organizer and chair for the AAAI Rigorous Evaluation workshops, Crowdcamp 2016, SIGIRWebQA 2015 workshop, the Crowdsourcing at Scale 2013, the shared task challenge at HCOMP 2013, and Connecting Online Learning and Work at HCOMP 2014, CSCW 2015, and CHI 2016 toward the goal of galvanizing research at the intersection of crowdsourcing, natural language understanding, knowledge representation, and rigorous evaluations for artificial intelligence.

- Email: `pkparitosh@gmail.com`
- Web page: `https://users.cs.northwestern.edu/~paritosh/`
- Semantic Scholar: `https://www.semanticscholar.org/author/Praveen-K.-Paritosh/2990264`

Paolo Climaco (he/him) is a PhD student at the Bonn Institute for Numerical Simulation, with research interests in the intersection of mathematics and AI. He is a member of the Bonn International Graduate School of Mathematics and is a graduate student member of the Society for Industrial and Applied Mathematics. Paolo's current research focuses on developing mathematical insights into data-centric AI methodologies, particularly data selection approaches, and understanding how such techniques can benefit the domain of science, specifically quantum chemistry.

- Email: `climaco@ins.uni-bonn.de`
- Web page: `https://www.linkedin.com/in/paolo-climaco-5b02871b7/`
- Google Scholar: `https://scholar.google.com/citations?user=bl6LxawAAAAJ&hl=it`

Bolei Ma (he/him) is a PhD student at the Social Data Science and AI Lab of the Department of Statistics, LMU Munich. Previously, he got an M.A. degree in Linguistics and an M.Sc. degree in Computational Linguistics from LMU Munich. His research interest research interest lies in data- and human-centric AI, and social NLP. He is currently a co-organizer of the 29th LIPP Symposium, which focuses on the revitalization of minority languages of the world.

- Email: `bolei.ma@stat.uni-muenchen.de`
- Web page: `https://boleima.github.io/`
- Google Scholar: `https://scholar.google.com/citations?user=9KdJOfAAAAAJ&hl=en&oi=ao`

Steffen Vogler (he/him) is a Senior Imaging Technology Scientist at Bayer, leading research and product development on AI in medical computer vision with special focus on the Radiology domain. His interest is in data-centric AI, ethical AI and health equity. He is a member of the ITU-WHO Focus Group "Artificial Intelligence for Health". Prior to joining Bayer in 2017, he did a PhD in Neurobiology and worked on basic research questions around memory formation in the mammal brain.

- Email: `steffen.vogler@bayer.com`
- Web page: `https://www.linkedin.com/in/steffen-vogler-762783102/`
- Google Scholar: `https://scholar.google.de/citations?user=s6CqZi8AAAAJ`

Danilo Brajovic (he/him) is a Research Associate at the Fraunhofer Institute for Manufacturing Engineering and Automation IPA and a PhD student at the University of Stuttgart. His current research is centered around safe AI in industrial applications, specifically developing methods to ensure the quality of datasets for safety-critical applications. Before joining Fraunhofer, Danilo completed a double degree in cognitive and computer science from Tubingen University. He has also worked on topics related to fair ML at the Max-Planck Institute for Intelligent Systems and on autonomous driving at Bosch Corporate Research and Mercedes-Benz.

- Email: `danilo.brajovic@ipa.fraunhofer.de`
- Web page: `https://www.linkedin.com/in/danilo-brajovic-626438163/`

- ResearchGate: https://www.researchgate.net/scientific-contributions/Danilo-Brajovic-2229753350

Mayee Chen (she/her) is a PhD student in the Computer Science department at Stanford University advised by Professor Christopher Ré. She is interested in understanding and improving how models learn from data. Recently, she has focused on problems in data selection, data labeling, and data representations, especially in the setting where there exist multiple input signals or objectives. Her work on contrastive learning representations and integrating weak supervision with foundation models has been recognized with a best paper award at AAAI AIBSD 2021 and with a best student paper runner up award at UAI 2022, respectively. Previously, she graduated summa cum laude from Princeton University with a concentration in Operations Research and Financial Engineering and a certificate in Applications of Computing, where she worked with Professors Elad Hazan and Miklos Racz.

- Email: mfchen@stanford.edu
- Web page: https://mayeechen.github.io/
- Google Scholar: https://scholar.google.com/citations?user=dhgytncAAAAJ&hl=en

Sang Truong (he/him) is a Ph.D. student at the Stanford AI Lab advised by Professor Sanmi Koyejo, Professor Nick Haber, and Professor Susan Athey. He specializes in developing intelligent agents capable of goal-directed exploration across a range of diverse, uncertain environments inherent in human-centric AI applications like scientific discovery, healthcare, and education. His research interests encompass decision-making under uncertainty, Bayesian optimization, bandit problems, and reinforcement learning.

- Email: sttruong@cs.stanford.edu
- Web page: ai.stanford.edu/~sttruong
- Google Scholar: https://scholar.google.com/citations?user=oXPm0dAAAAAJ&hl=en