ICLR 2025 2nd Workshop on Navigating and Addressing Data Problems for Foundation Models (DPFM)

Abstract

Foundation models (FMs) have become central to modern machine learning, with data playing a crucial role in their development and sparking increased attention to data-related challenges such as curation and attribution. Adapting traditional data-centric methods to FMs is challenging due to the scale of both data and model architectures, necessitating interdisciplinary collaboration and community efforts. Building on the success of the first Data Problems in Foundation Models (DPFM) workshop at ICLR 2024, the second DPFM workshop will address persistent and emerging data-related challenges in FM deployment. While longstanding issues in data collection, curation, and synthesis remain relevant, new challenges have arisen as FMs are integrated into a growing number of applications and become increasingly multi-modal. Concurrently, the societal impact of AI has intensified, highlighting concerns such as data copyright. These evolving challenges emphasize the need for continued, focused discussions on data-related issues in FM development. Our goals include fostering a comprehensive understanding of these challenges across the entire FM pipeline and creating a platform for interdisciplinary researchers to connect, collaborate, and drive progress. We hope this workshop will serve as a catalyst for innovative solutions to critical data challenges, shaping the future of FMs and their wide-ranging applications.

1 Workshop Summary

Foundation models (FMs) have emerged as a cornerstone of modern machine learning (ML), demonstrating exceptional capabilities across a wide array of tasks by leveraging large-scale datasets and model architectures. Data plays a central role in the development of foundation models. Consequently, research on data-related problems has received significantly increased attention among the ML community. For instance, *data curation*, a long-standing challenge in ML, has garnered renewed interest because it is crucial not only for foundation models' performance [6], but also for reducing computational costs [23] and mitigating undesirable behaviors such as bias and toxicity [20]. Another data-related problem, *training data attribution*, which aims to quantify the contribution of training data for specific model behaviors [18], has become an increasingly active research area due to the need for explainability in FM-based applications [31]. The critical role of data in the era of FMs has also raised new research problems, such as *data copyright protection* [13] and *data pricing/marketplace mechanisms* [2; 4; 3]. Addressing these evolving data challenges necessitates collaborative endeavors across diverse disciplines from traditional core fields in machine learning, such as optimization and statistics, to areas like game theory [14], economics [7], education [22], law [15], and other disciplines, as well as specific application domains with unique data requirements.

Challenges in Conducting Data-related Research for Foundation Models. Adapting existing methods—such as data curation and data attribution approaches originally developed for traditional ML settings—to the context of FMs presents significant challenges due to the unprecedented scale of data volumes and model parameters. Despite the growing interest in this field, several challenges hinder effective progress. Data-related research in machine learning is inherently interdisciplinary, spanning from fundamental theory and algorithms to a wide range of applications. Hence, there remains a critical need for *interdisciplinary collaboration and clear communication*. Additionally, this emerging field grapples with additional challenges, including *poor reproducibility* due to variations in subtle experimental settings and software environments [30], *computational resource disparities* among different research institutions [1], and the need for reliable *standardized benchmarks* [11; 19].

Motivation & Goal of the 2nd DPFM Workshop. The first iteration of the DPFM workshop at ICLR 2024 attracted over 60 paper submissions and had over 300 attendees. Building on this success, we are excited to propose the second iteration of the workshop. *Several existing challenges from last year remain*, including improving systematic data collection, curation, and synthesis for FMs. Understanding how these processes affect training efficiency, model performance, safety, and other model behaviors continues to be crucial. At the same time, *new challenges have emerged* as foundation models are increasingly integrated into various applications through diverse customization methods (e.g., fine-tuning, retrieval-augmented generation (RAG), LLM-enabled agents) and are becoming increasingly multi-modal. These new scenarios prompt us to reexamine the role and impact of data in shaping downstream performance and outcomes. *Several pre-existing issues have also intensified*. For instance, AI's societal impact has become more pronounced, with growing evidence of reduced income and job opportunities for data creators [12]. Additionally, concerns over data copyright and the need for proper data attribution have escalated [8; 24]. *These factors collectively underscore the need to organize the second iteration of the workshop*. In this new iteration, we aim to address both existing and emerging data challenges by bringing together leading researchers from academia and industry to share insights and foster discussions on data-related issues in the context of FMs. By facilitating interdisciplinary dialogue, we will establish a dedicated venue for cross-domain conversations on data issues, bridging diverse research background and fostering collaboration.

Audience and Topics of Interest. Our target audience encompasses a diverse range of stakeholders across the FM ecosystem. This includes researchers tackling data-related challenges (e.g., data curation, attribution, copyright projection), as well as those focusing on different aspects of foundation models (e.g., alignment, safety, fairness/ethics, privacy). Additionally, we aim to engage practitioners for downstream applications, policy makers shaping data regulations, and organizations advocating for open-source and ethical data use. By reaching diverse groups, we seek to foster comprehensive discussion and advance data-related research. The topics of interest include, but are not limited to:

Data Collection and Curation for Foundation Models

- Practical data curation strategies and techniques: What are the current best practices for curating data (e.g., filtering, mixing, repairing) for different stages of FM training? How do existing data curation techniques extend to scenarios such as RAG, LLM agents, and multi-modal settings?
- Theoretical frameworks and scaling laws in data curation: What theoretical frameworks can guide data selection? How do data curation strategies change scaling laws for foundation models?

Data Attribution, Interpretability, and Data Marketplace

- Data attribution and fact tracing in foundation models: How can we efficiently attribute model outputs back to specific training examples? How can we fairly evaluate and compare the effectiveness of different data attribution techniques?
- Economic models and data marketplaces: What are effective economic models for data pricing and creating incentives for data sharing? How can data marketplaces be designed and implemented to facilitate data sharing while ensuring fair compensation?

Law and Technical Solutions for Data Copyright Protection

- Copyright issues and compliance in training data: How to mathematically define the current copyright challenges of FMs' training data? What mitigation strategies can be employed?
- Connections between copyright, privacy, and other fields in machine learning: How do concepts like differential privacy [10] and algorithmic fairness [9] relate to training data copyright issues? Can we adapt existing techniques (e.g., machine unlearning) for copyright projection?

Synthetic Data and Model Collapse

- Synthetic data generation and its impact: How can we generate high-quality synthetic training data for foundation models? What are the impacts of training on synthetic data on model performance, robustness, and other metrics? How can we better understand and mitigate model collapse when training foundation models from both theory and empirical aspects?

Data and Society (Safety, Privacy, Fairness, and Other Social Impacts)

- How can data-centric methods improve AI safety, privacy, and fairness? What are the side effects of data curation on fairness and ethics in foundation models, and how can they be mitigated?

Benchmarks and Evaluations

- How can we design evaluation metrics for data-centric techniques & create reliable dataset benchmarks for FMs? What are the existing/potential pitfalls (e.g., test data contamination [21])?

2 Tentative Schedule & Plans

The workshop will be in-person but *possible for online participation*. We have planned for six 30-min invited talks (22-min talk + 8-min Q&A), four 10-min spotlight talks selected from regular/position papers track submissions (7-min talk + 3-min Q&A), two 5-min spotlight talks selected from Tiny papers track submissions, two 75-min poster sessions, and a 30-min panel discussion.

2.1 Tentative Schedule

We provide a tentative schedule as follows, **with more than half of the time allocated for open discussions or networking** (poster sessions, coffee/lunch breaks, and panel discussions). As a core objective of the workshop is to provide a platform for exchanging insights, sharing ideas, and fostering collaboration, our schedule switches between talks and open discussions (poster sessions, panel discussion, coffee breaks) several times throughout the event. This arrangement allows the audience to *easily connect with speakers following their talks*, encouraging open discussion. It also facilitates *cross-participation between different workshops* and helps connect with a broader audience.

Morning session:

- 9:00–9:05 Opening Remarks
- 9:05–9:35 Invited Talk 1 (Ari Marcos)
- 9:35–10:05 Invited Talk 2 (Baharan Mirzasoleiman)
- 10:05–11:20 Poster Session 1
- 11:20–11:30 Coffee Break
- 11:30–12:00 Invited Talk 3 (Peter Henderson)
- 12:00–12:10 Spotlight Presentation 1
- 12:10–12:20 Spotlight Presentation 2
- 12:20–12:30 Spotlight Presentation (Tiny Papers)
- 12:30–1:30 Lunch Break

Afternoon session:

- 1:30–2:00 Invited Talk 4 (Bryan Low)
- 2:00–2:30 Invited Talk 5 (Danqi Chen)
- 2:30–2:40 Spotlight Presentation 3
- 2:40–2:50 Spotlight Presentation 4
- 2:50–4:05 Poster Session 2
- 4:05–4:35 Invited Talk 6 (Kyle Lo and Luca Soldaini)
- 4:35–5:05 Panel Discussion (Moderator: Tatsunori Hashimoto)
- 5:05–5:10 Closing Remarks

Our workshop will conclude around **5:10pm** to maintain a manageable schedule for attendees. Prior to the workshop, we will put talk and poster titles/abstracts on the workshop website to allow for choice of attendance based on content.

2.2 Highlights of Our Plans

Invited Talks & Spotlight Presentations. (1) Remote participant accommodation: To accommodate online participants, all invited talks and spotlight presentations will be livestreamed via Zoom, with questions from online participants facilitated through moderated discussions. **(2) Extended Q&A session:** If the number of questions exceeds the available time, the organizers will collect them and share them on the Slack channel, allowing presenters to respond later. Based on feedback from the previous DPFM workshop at ICLR 2024, *Q&A sessions for invited talks have been extended to 8-min* to encourage more in-depth interaction between speakers and the audience. **(3) Connecting speakers and audience:** The Slack channel will also serve as a platform for follow-up questions and continued engagement between the audience and presenters. **(4) Selection of spotlight presentations:** We aim to provide presentation opportunities for high-quality submissions to showcase the latest advancements in the field. Spotlight presentations will be selected based on reviewer nominations and discussions within the organizing committee, with conflicts of interest being carefully managed. **(5) Spotlight presentations from Tiny Papers track:** This year, our workshop will introduce a new Tiny Papers Track (see Section 2.3) aimed at increasing the visibility of research from underrepresented, under-resourced, and early-career researchers. We will select two high-quality submissions

from this track for a 5-min oral presentation during the workshop, presented alongside the spotlight presentations from the regular track.

Poster Sessions. (1) Remote participant accommodation: The poster session can be attended either virtually or in person. We include 2 poster sessions so that virtual attendees have the flexibility to choose a session that aligns with their time zone. To enhance the experience for remote attendees, we will setup a dedicated channel on a chat platform like Rocket.Chat to facilitate interactions among workshop participants. If the accepted paper author cannot attend the workshop in-person, the organizers will display posters on behalf of authors. We also encourage the authors to upload an optional video about their works (see Section 2.3). **(2) Extended length**: In response to feedback from the first DPFM workshop at ICLR 2024, we have extended the poster sessions to 75 minutes to foster more in-depth discussions and networking opportunities in this iteration.

2.3 Paper Submission

Regular/Position Papers Track. Our workshop welcomes research and position papers addressing data challenges for foundation models, with options for both long (10-page) and short (4-page) submissions in ICLR template. **Discouraging submissions of published work:** We will explicitly discourage submissions of previously published work at major ML venues (e.g., ICLR, ICML, NeurIPS), including papers accepted to the ICLR main conference.

Tiny Papers Track. This year, our workshop will incorporate a Tiny Papers Track to encourage underrepresented, under-resourced, and early-career researchers who may not yet have the resources to submit full papers. Similar to the Tiny Paper Track in the main conference, our Tiny Papers Track welcomes submissions that are in the early stages of a research project. For example, a modest but self-contained theoretical result, a novel observation from preliminary experiments, or a new perspective on an existing problem. We aim to foster early-stage ideas and provide a platform for researchers to receive feedback and support as they develop their work further.

Optional Videos. We encourage the authors of accepted papers from both tracks to upload a pre-recorded video about their works, enabling remote attendees to access the content flexibly.

Paper Review & Conflict of Interests. The paper review process will follow a double-blind format, ensuring anonymity for both authors and reviewers. We will reach out to diverse reviewers to provide a broad range of expertise. Each submission will be reviewed by at least 3 reviewers, and decisions will be made transparently by the organizing committee. To prevent conflicts of interest, the authors will be asked about potential conflicts via OpenReview. The final list of accepted papers will be published on the workshop website.

Tentative Schedule of Paper Submission. We will follow the suggested dates by ICLR.

- Workshop paper submission deadline: February 3, 2025.
- Workshop paper notification date: March 3, 2025.
- Camera-ready and (optional) posters and video recordings upload: April 5, 2025.

3 Invited Speakers and Panelists

We have invited 7 speakers, listed below with their tentative talk themes, covering existing key data challenges for foundation models and offering diverse perspectives to enrich the discussion. **All speakers have confirmed their availability for the workshop.**

- Ari Marcos (CEO, DatologyAI): Data curation strategies for training foundation models.
- Baharan Mirzasoleiman (Assistant Professor, UCLA): Theory of data-efficient learning.
- Peter Henderson (Assistant Professor, Princeton): Law and technical solutions for data copyright.
- Bryan Low (Associate Professor, NUS): Data attribution and data marketplaces.
- Danqi Chen (Assistant Professor, Princeton): Data selection for instruction tuning.
- Kyle Lo and Luca Soldaini (Lead Scientists, AI2): Open Data for Language Models.

Our invited speakers come from diverse backgrounds, including industry, non-profit research institutions, and academia, offering a comprehensive view of the challenges and opportunities in data-centric foundation model research. The speakers include two from non-profit research institute (Kyle and Luca), one from a startup (Ari), as well as both early-career and tenured faculty members.

Penal Discussion. The workshop will feature a panel discussion focused on the current challenges of data-centric approaches for foundation models. **Tatsunori Hashimoto** (Assistant Professor, Stanford) has confirmed his availability as the panel moderator. All confirmed invited speakers are welcome to join as panelists, with the final list to be determined closer to the workshop date. To ensure a wide range of perspectives from diverse research communities, we may also invite additional experts working at the intersection of computer science and other disciplines, such as **Lindsey Raymond** (Harvard Business School, Economics), **Pamela Samuelson** (UC Berkeley, Law), **Connor T. Jerzak** (UT Austin, Politics), and **Snehalkumar Gaikwad** (UNC Chapel Hill, Ethics). We may also invite experts from specific application domains, such as **Sherrie Wang** (MIT, AI for Earth Science) and **Jimeng Sun** (UIUC, AI for Healthcare), along with leaders from industry and government, such as **Sara Hooker** (Cohere For AI) and **Elena Sizikova** (Food and Drug Administration).

3.1 Biographies of Speakers and Panel Moderator

Ari Morcos is the CEO and co-founder of Datology AI. Previously, he was a senior staff research scientist at Meta AI Research (FAIR Team) in Menlo Park working on understanding the mechanisms underlying neural network computation and function, and using these insights to build machine learning systems more intelligently. Most recently, his work has focused on understanding properties of data and how these properties lead to desirable and useful representations, with a particular emphasis on data curation. Ari's work has been honored with Outstanding Paper awards at both NeurIPS and ICLR. Before joining FAIR, Ari worked at DeepMind in London, and earned his PhD in neuroscience working with Chris Harvey at Harvard University.

Baharan Mirzasoleiman is an Assistant Professor in the Computer Science Department at UCLA, where she leads the BigML research group. Her research aims to address sustainability, reliability, and efficiency of machine learning. She is mainly working on improving the big data quality, by developing theoretically rigorous methods to select the most beneficial data for efficient and robust learning. Besides, she is also interested in improving the models and learning algorithms. The resulting methods are broadly applicable for learning from massive datasets across a wide range of applications, such as medical diagnosis and environment sensing. Baharan received the ETH medal for Outstanding Doctoral Thesis, were recognized as a Rising Star in EECS by MIT, and were awarded the NSF Career Award.

Peter Henderson is an assistant professor at Princeton University with appointments in the Department of Computer Science and the School of Public and International Affairs, as well as the Center for Information Technology Policy. His research focuses on aligning machine learning, law, and policy for responsible real-world deployments. This includes work on AI safety, methods to improve reasoning in foundation models, interdisciplinary methods in law and AI, as well as core work on legal doctrine and policy, particularly around AI governance. He received his J.D. from Stanford Law School and Ph.D. in computer science from Stanford University.

Bryan Low is an Associate Professor of Computer Science at the National University of Singapore, the Director of AI Research at AI Singapore, and the Deputy Director of NUS AI Institute. His research interests include probabilistic & automated machine learning, planning under uncertainty, and multi-agent/robot systems. Dr. Low is the recipient of the (1) Andrew P. Sage Best Transactions Paper Award for the best paper published in all 3 of the IEEE Transactions on Systems, Man, and Cybernetics - Parts A, B, and C in 2006; (2) National University of Singapore Overseas Graduate Scholarship for Ph.D. studies in Carnegie Mellon University (CMU) in 2004-2009; (3) Singapore Computer Society Prize for Best M.Sc. Thesis in School of Computing, National University of Singapore in 2003; and (4) Faculty Teaching Excellence Award in School of Computing, National University of Singapore in 2017-2018.

Danqi Chen is an assistant professor of Computer Science at Princeton University. She co-leads the Princeton NLP Group and is the associate director of Princeton Language and Intelligence (PLI), an initiative that seeks to develop fundamental research of large AI models (e.g., LLMs). Her research interests lie broadly in natural language processing and machine learning. Danqi received her Ph.D. from Stanford University (2018) and B.E. from Tsinghua University (2012), both in Computer Science. Danqi is a recipient of a 2022 Sloan Fellowship, a Lawrence Keyes, Jr./Emerson Electric Co. Faculty Advancement Award, faculty awards from Google, Meta, Amazon, Apple, and Salesforce, and paper awards from ACL 2016, EMNLP 2017, and ACL 2022.

Kyle Lo is the Lead Scientist at Allen Institute for AI on the OLMo and Semantic Scholar projects. Kyle specializes in topics in natural language processing, machine learning and human-AI interaction. Kyle is the core contributor to Dolma, the largest open dataset for language model pretraining to-date, and peS2o, a transformation of S2ORC optimized for pretraining language models of science. Prior to this, Kyle co-led the data curation efforts behind S2ORC, the largest, machine-readable collection of open-access full-text papers to-date, and CORD-19, the most comprehensive, continually-updated set of COVID-19 literature at the time.

Luca Soldaini is a senior research scientist at the Allen Institute for AI in the Semantic Scholar and OLMo teams, and an organizer at Queer In AI. Prior to joining AI2, Luca was a senior applied scientist at Amazon Alexa. Luca completed his Ph.D. in computer science at Georgetown University in 2018 in the Information Retrieval Lab working with Nazli Goharian. Luca's research focuses on best practices for curation and exploration of large corpora, mostly in the context of (Large) Language Model. **Luca and Kyle Lo are co-leading Data Research for OLMo.**

Tatsunori Hashimoto (panel moderator) is an assistant professor at the computer science department at Stanford University. His research uses tools from statistics to make machine learning systems more robust and trustworthy — especially in complex systems such as large language models. Previously, he was a post-doc at Stanford working for John C. Duchi and Percy Liang on tradeoffs between the average and worst-case performance of machine learning models. Before his post-doc, he was a graduate student at MIT co-advised by Tommi Jaakkola and David Gifford and an undergraduate student at Harvard in statistics and math advised by Edoardo Airoldi.

4 Organizers and Biography

Contact: We will set up a dedicated email group and Slack channel to ensure prompt responses to external inquiries.

Members of the organizing team:

- Prof. Ruoxi Jia (Assistant Professor, Virginia Tech)
- Prof. Pang Wei Koh (Assistant Professor, University of Washington)
- Prof. Dawn Song (Professor, UC Berkeley)
- Dr. Jerone Andrews (Research Scientist, Sony AI)
- Feiyang Kang (PhD student, Virginia Tech)
- Hoàng Anh Just (PhD student, Virginia Tech)
- Jiachen T. Wang (PhD student, Princeton University)

Organization experiences. The organizing team consists of both researchers with extensive experience in organizing academic events and Ph.D. students who are first-time workshop organizers. Notably, 3 out of the 7 organizers did not participate in the organization of the first iteration of the DPFM workshop at ICLR 2024. Among them, Hoàng Anh Just and Jiachen T. Wang are organizing a workshop for the first time. Prof. Ruoxi Jia organized AsiaCCS Workshop on Secure and Trustworthy Deep Learning Systems (2022), ICML Workshop on Economics of Privacy and Data Labor (2020), Workshop on AI for Energy-Cyber-Physical Systems (2018), Workshop on Smart Buildings as Enablers for a Smarter Grid (2016). Prof. Pang Wei Koh organized the Workshop on Distribution Shifts at NeurIPS 2021–2023. Dr. Jerone Andrews organized CVPR Workshop on Responsible Data (2024), ICLR Workshop on Data-Centric Machine Learning Research (2024), and Workshop on AI and Future Crime (2019).

Diversity. See Section 7.

4.1 Biographies

Ruoxi Jia is an assistant professor in the Bradley Department of Electrical and Computer Engineering at Virginia Tech. Her research interests span machine learning, security, privacy, and cyber-physical systems, with a recent focus on data-centric and trustworthy AI. Her work has earned her several prestigious awards and fellowships, including the NSF CAREER Award and the Best Social Impact Paper Award at ACL. Her research has been featured in prominent media outlets such as The New York Times, IEEE Spectrum, and MIT Technology Review. Her work on data valuation and selection has been adopted by companies in the financial sector and tech industry.

Pang Wei Koh is an assistant professor in the Allen School of Computer Science and Engineering at the University of Washington. His research interests are in the theory and practice of building reliable and interactive machine learning systems. His research has been published in Nature and Cell, featured in media outlets such as The New York Times and The Washington Post, and recognized by the MIT Technology Review Innovators Under 35 Asia Pacific award and best paper awards at ICML and KDD. He received his PhD and BS in Computer Science from Stanford University. Prior to his PhD, he was the 3rd employee and Director of Partnerships at Coursera.

Dawn Song is a professor in the Department of Electrical Engineering and Computer Science at UC Berkeley. Her research interest lies in AI and deep learning, blockchain/web3, security and privacy. She is the recipient of various awards including the MacArthur Fellowship, the Guggenheim Fellowship, the NSF CAREER Award, the Alfred P. Sloan Research Fellowship, and the MIT Technology Review TR-35 Award. She is an ACM Fellow and an IEEE Fellow. She is ranked the most cited scholar in computer security (AMiner Award).

Jerone Andrews is a Research Scientist at Sony AI (London), where he leads a team focused on responsible data curation, bias detection and mitigation, and representation learning. His work has been recognized with awards such as ICML'24 Best Paper and ECCV'24 Best Paper Candidate. He holds a PhD and MRes in Computer Science from University College London and an MSci in Math from King's College London. His early career included a Royal Academy of Engineering Research Fellowship and a British Science Association Media Fellowship with BBC Future.

Feiyang Kang is a PhD student at Virginia Tech, advised by Prof. Ruoxi Jia. His research interests lie in data-centric AI and its applications in trustworthy ML, from data selection/curation, scaling laws, data valuation to data influence/attribution/interpretability, etc. His recent pursuit includes designing effective schemes to implement data-centric methods for large-scale applications such as foundation models. His work has been adopted into production pipelines by leading tech companies. Previously, he interned at Meta, FAIR team, and NVIDIA AI Research.

Hoàng Anh Just is a PhD student in the Bradley Department of Electrical and Computer Engineering at Virginia Tech, advised by Prof. Ruoxi Jia. His research focuses on data valuation in machine learning, for example, how to fairly valuate the contribution of each datapoint that is used for machine learning training. Additionally, his work also focuses on data curation for improving efficacy and efficiency for training ML models. **Hoàng is a first-time workshop organizer.**

Jiachen Wang is a fourth-year Ph.D. student at Princeton University. His research focuses on data-centric machine learning from a rigorous statistical perspective. Currently, he is developing principled and scalable data attribution techniques for foundation models. Jiachen's works in data attribution, data curation, and privacy-preserving machine learning have been recognized with multiple oral and spotlight presentations at top machine learning conferences [17; 29; 25; 26; 16; 28; 27]. He is the recipient of the 2024 Rising Star in Data Science. **Jiachen is a first-time workshop organizer.**

5 Anticipated Audience Size

Based on the successful outcome of the first edition of DPFM workshop at ICLR 2024, and given the growing research interests in the intersection of foundation models and data-centric machine learning, we expect more than 50 attendees in the room at all times and more than 500 audiences in total throughout the event day.

6 Plan to Get an Audience for the Workshop

To ensure the success of the workshop and maximize engagement with potential audiences, the organizing team plans to dedicate substantial effort to promoting this event. We consider the following approaches:

1. Create a comprehensive workshop webpage with all relevant information, including the schedule, speakers, submission guidelines, and updates.
2. Advertise the workshop in the upcoming machine learning conferences (e.g., NeurIPS 2024) to increase visibility to relevant audiences.

3. Send email announcements and calls for papers to relevant mailing lists in both academia and industry to ensure broad awareness.

4. Share information on social media platforms such as Twitter and LinkedIn, as well as advertise on relevant Slack workspaces, to maximize reach.

5. Pay special attention to **mailing lists related to minority and underrepresented groups in AI**, ensuring inclusivity and broader representation. The organizers' home institutions can help to reach a diverse set of participants and encourage attendance from different disciplines.

6. Encourage accepted paper authors to promote their participation and attract their own networks to the workshop.

7. Prior to the workshop, we will put talk and poster titles up publicly to allow for choice of attendance based on content.

7 Diversity Commitment

We are committed to fostering diversity in both our organizing team and invited speakers, encompassing gender, affiliation, geographic location, career stage, expertise, and cultural background. The current tentative schedule reflects our efforts to balance gender, representation from industry and academia, policy and technical backgrounds, and geographic diversity.

Diversity in the Organizing Team and Speakers. The organizing team and speakers include members from diverse backgrounds, including ethnic backgrounds (White/Asian/Black), historically underrepresented groups, gender identities (men/women), affiliations (acedemia/industry/non-profit organizations), seniority levels (PhD students, junior and senior faculty, industry research scientists), and geographic regions/origins (Western Europe, North America, Southeast Asia, Eastern Europe, ...). Among the organizers, Jiachen Wang and Hoàng Anh Just are first-generation college students. The organizing team also consists of a mix of first-time organizers ($\approx 30\%$) alongside those with previous experience in workshop organization, ensuring a wide array of perspectives and expertise.

Diversity in Discussion Topics. The workshop agenda includes a broad spectrum of data-related topics. We especially encourage participation from researchers across diverse fields to contribute to the panel discussions and enrich the diversity of viewpoints. The workshop agenda contains a broad spectrum of data-related topics with significant implications for diversity-related challenges such as fairness, accountability, and transparency in AI systems. Data practices directly impact how AI models perform across different demographic groups, affecting issues like bias and discrimination. By highlighting these connections, we aim to foster discussions that explore how data problems can be addressed to promote equity and inclusion.

Singapore Representatives. Our organizing team and speakers include representatives from/based in Singapore (Pang Wei Koh and Bryan Low). We are particularly excited to discuss the implications of the recent Model AI Governance Framework [5] introduced by the Singapore government, presenting a unique opportunity to explore data governance practices in an international context.

Registration Fee & Travel Grants for Junior Researchers, Local Singaporeans, and Underrepresented Groups. To make our workshop more accessible, we aim to offer grants for registration fees and travel expenses to participants who may face financial barriers to attending. Priority will be given to junior researchers, local Singaporeans, and historically underrepresented groups. We are seeking sponsorship from companies such as DatologyAI, Cisco, and Amazon to support these initiatives and encourage broader participation.

Hybrid Meeting Format. Recognizing the challenges posed by varying time zones in a hybrid meeting format, we will incorporate a series of actions to ensure wide participation. If the accepted paper author cannot attend the workshop in-person, we will display posters on behalf of authors. We also encourage the authors to upload a pre-recorded video about their works and put them on the workshop website in advance, enabling other attendees to access the content flexibly. Live sessions such as invited talks and panel discussion will be lived streamed through Zoom, and the Q&A session will be facilitated through Slack.

8 Virtual Access to Workshop Materials and Outcomes

All relevant information will be regularly updated on a dedicated webpage.

Accepted Papers, Posters, and Optional Videos. Although the workshop is non-archival, we will use OpenReview for the double-blind review process and will host the accepted papers. The posters and videos will be uploaded to the workshop website (with authors' consent).

Slides and Recordings. With the consent of the speakers, we will provide workshop slides and recordings to all attendees after the event. We will also offer a summarization of relevant code repositories, datasets, and additional reading materials. We will host a YouTube channel to publish the workshop recordings as well as the optional videos submitted by paper authors (with their consent). This will help disseminate the content to a broader audience.

After the workshop, we will write a position paper summarizing the new insights and future directions discussed during the presentations and panel discussions. This paper will serve as a valuable reference for both attendees and the broader research community.

9 Previous Related Workshops

This is the second iteration of the DPFM workshop, following our first event at ICLR 2024. In this new iteration, we focus on both persistent and emerging data challenges during the past year. The data-related research for foundation model is still in its early stages, and the DPFM workshop distinguishes itself from existing workshops and series in several key aspects:

Workshops on Data-Centric ML. Recent years have witnessed the successful launch and growing popularity of workshops in data-centric ML such as DMLR workshop series (ICML'23, ICLR'24, ICML'24), DCAI workshop (NeurIPS'21), and DataPerf workshop (ICML'22). While these workshops focus on the broader field of data-centric ML, the DPFM workshop narrows its focus to the pressing and increasingly important challenges related specifically to data problems in foundation models. Given the unique difficulties and rising significance of foundation model research, a dedicated workshop is essential for reshaping the research landscape and fostering focused discussions.

Workshops on Data Attribution. The workshop on *Attributing Model Behavior at Scale* (NeurIPS'23 and NeurIPS'24) primarily explored efficient data attribution techniques for large-scale models and datasets. While this represents a crucial aspect of data problems, it is intrinsically linked with other areas such as data curation [28] and copyright protection [8; 24]. The DPFM workshop aims to bridge these interconnected topics, bringing together researchers from various subfields to facilitate a more holistic understanding and approach to data challenges in foundation models.

The past two years have seen many workshops on foundation models, such as Multi-modal Foundation Model meets Embodied AI (MFM-EAI) at ICML'24, Long-Context Foundation Models at ICML'24, Reliable and Responsible Foundation Models at ICLR'24, Workshop on Theoretical Foundations of Foundation Models (TF2M) at ICLR'24. Research focusing on data offers promising opportunities to address some of the key challenges in foundation models and their deployment. We are committed to supporting research on data-related challenges in the foundation model era and aim to continue this workshop series, as interest and usage in this area are expected to grow in the coming years.

References

[1] Inequality grows in ai research. https://www.axios.com/2020/10/28/growing-inequality-ai-research, 2020.

[2] Amazon aws data exchange. https://aws.amazon.com/data-exchange, 2023.

[3] Databricks marketplace. https://marketplace.databricks.com, 2023.

[4] Snowflake datamarketplace. https://www.snowflake.com/en/data-cloud/marketplace/, 2023.

[5] https://iapp.org/resources/article/global-ai-governance-singapore/. https://iapp.org/resources/article/global-ai-governance-singapore/, 2024.

[6] A. Albalak, Y. Elazar, S. M. Xie, S. Longpre, N. Lambert, X. Wang, N. Muennighoff, B. Hou, L. Pan, H. Jeong, C. Raffel, S. Chang, T. Hashimoto, and W. Y. Wang. A survey on data selection for language models, 2024.

[7] E. Brynjolfsson, D. Li, and L. R. Raymond. Generative ai at work. Technical report, National Bureau of Economic Research, 2023.

[8] J. Deng and J. Ma. Computational copyright: Towards a royalty model for ai music generation platforms. *arXiv preprint arXiv:2312.06646*, 2023.

[9] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[10] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[11] S. Y. Gadre, G. Ilharco, A. Fang, J. Hayase, G. Smyrnis, T. Nguyen, R. Marten, M. Wortsman, D. Ghosh, J. Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024.

[12] T. Germain. Ai took their jobs. now they get paid to make it sound human. https://www.bbc.com/future/article/20240612-the-people-making-ai-sound-more-human, 2024. Accessed: 2024-10-20.

[13] M. M. Grynbaum and R. Mac. The times sues openai and microsoft. *The New York Times*, pages B1–B1, 2023.

[14] M. Han, J. Light, S. Xia, S. Galhotra, R. C. Fernandez, and H. Xu. A data-centric online market for machine learning: From discovery to pricing. *arXiv preprint arXiv:2310.17843*, 2023.

[15] P. Henderson, X. Li, D. Jurafsky, T. Hashimoto, M. A. Lemley, and P. Liang. Foundation models and fair use. *arXiv preprint arXiv:2303.15715*, 2023.

[16] J. Hong, J. T. Wang, C. Zhang, L. Zhangheng, B. Li, and Z. Wang. Dp-opt: Make large language model your privacy-preserving prompt engineer. In *The Twelfth International Conference on Learning Representations*, 2024.

[17] H. A. Just, F. Kang, T. Wang, Y. Zeng, M. Ko, M. Jin, and R. Jia. Lava: Data valuation without pre-specified learning algorithms. In *The Eleventh International Conference on Learning Representations*, 2022.

[18] P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894. PMLR, 2017.

[19] J. Li, A. Fang, G. Smyrnis, M. Ivgi, M. Jordan, S. Gadre, H. Bansal, E. Guha, S. Keh, K. Arora, et al. Datacomp-lm: In search of the next generation of training sets for language models. *arXiv preprint arXiv:2406.11794*, 2024.

[20] S. Longpre, G. Yauney, E. Reif, K. Lee, A. Roberts, B. Zoph, D. Zhou, J. Wei, K. Robinson, D. Mimno, et al. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. *arXiv preprint arXiv:2305.13169*, 2023.

[21] Y. Oren, N. Meister, N. S. Chatterji, F. Ladhak, and T. Hashimoto. Proving test set contamination in black-box language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[22] N. Prakriya, J.-N. Yen, C.-J. Hsieh, and J. Cong. Accelerating large language model pretraining via lfr pedagogy: Learn, focus, and review. *arXiv preprint arXiv:2409.06131*, 2024.

[23] B. Sorscher, R. Geirhos, S. Shekhar, S. Ganguli, and A. Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, 2022.

[24] J. T. Wang, Z. Deng, H. Chiba-Okabe, B. Barak, and W. J. Su. An economic solution to copyright challenges of generative ai. Technical report, 2024.

[25] J. T. Wang and R. Jia. Data banzhaf: A robust data valuation framework for machine learning. In *International Conference on Artificial Intelligence and Statistics*, pages 6388–6421. PMLR, 2023.

[26] J. T. Wang, P. Mittal, and R. Jia. Efficient data shapley for weighted nearest neighbor algorithms. *arXiv preprint arXiv:2401.11103*, 2024.

[27] J. T. Wang, T. Wu, P. Mittal, D. Song, and R. Jia. Compute-efficient llm training via online batch selection. *Advances in neural information processing systems*, 2024.

[28] J. T. Wang, T. Yang, J. Zou, Y. Kwon, and R. Jia. Rethinking data shapley for data selection tasks: Misleads and merits. In *Forty-first International Conference on Machine Learning*, 2024.

[29] J. T. Wang, Y. Zhu, Y.-X. Wang, R. Jia, and P. Mittal. Threshold knn-shapley: A linear-time and privacy-friendly approach to data valuation. *arXiv preprint arXiv:2308.15709*, 2023.

[30] T. Xia, B. Yu, K. Dang, A. Yang, Y. Wu, Y. Tian, Y. Chang, and J. Lin. Rethinking data selection at scale: Random selection is almost all you need. *arXiv preprint arXiv:2410.09335*, 2024.

[31] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, and M. Du. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38, 2024.