Navigating and Addressing **D**ata **P**roblems for **F**oundation **M**odels (DPFM), an ICLR 2024 Workshop

Abstract

Foundation Models (FMs, e.g., GPT-3/4, LLaMA, DALL-E, Stable Diffusion, etc.) have demonstrated unprecedented performance across a wide range of downstream tasks. Following the rapid evolution, as researchers strive to keep up with the understanding of the capabilities and limitations of FMs as well as their implications, attention is now shifting to the emerging notion of data-centric AI. The curation of training data has been shown to be crucially important for the performance and reliability of FMs and a wealth of recent works demonstrate that data-perspective research sheds light on a promising direction toward critical issues such as safety, alignment, efficiency, security, privacy, interpretability, etc. Our technical agenda is composed of four modules with 12 **confirmed** speakers:

A. Data Quality, Dataset Curation, and Data Generation–Recent Achievements and Current Efforts
B. A Data Perspective to Efficiency, Interpretability, and Alignment–Latest Advancement and Breakthroughs
C. A Data Perspective to Safety and Ethics–Risks, Limitations, and Opportunities
D. Copyright, Legal Issues, and Data Economy–A Broader Landscape

Workshop Summary

Foundation Models (FMs, e.g., GPT-3/4, LLaMA, DALL-E, Stable Diffusion, etc.) have been achieving sweeping success on a wide range of tasks. Pre-trained on a massive amount of data in a self-supervised manner, these versatile models are able to adapt to target tasks fairly easily and be deployed in various downstream scenarios, yielding exciting results that attract the interests of broad communities. As researchers strive to keep up with the understanding of the capabilities and limitations of FMs as well as their implications following the rapid evolution, looking forward, the attention is now shifting to the emerging notion of data-centric AI.

The curation of training data has been shown to be crucially important for the performance and reliability of FMs and a wealth of recent works (e.g., LAION / Datacomp, GPT-3 paper, GLaM paper, RedPajama, RLHF/LlaMa 2, Safety/Helpfulness datasets, etc.) demonstrate that data-perspective research sheds light on a promising direction toward critical issues such as safety, alignment, efficiency, security, privacy, interpretability, etc.

The recent year has seen a spur of individual works exploring many frontiers related to this topic, providing now an excellent opportunity to bring together brilliant minds to search for a systematic framework and roadmap for research. To move forward, this workshop aims to

discuss and explore a better understanding of the new paradigm for research on data problems for foundation models.

Our technical agenda is composed of four modules:

- **A. Data Quality, Dataset Curation, and Data Generation–Recent Achievements and Current Efforts**
    - *How to quantify data quality in the context of FMs or select a good subset?*
        - What are the aspects to consider and what are the quantitative metrics?
        - How this can be performed considering the scale or nature of data for FMs?
    - How to model the influence of data throughout the lifecycle of FMs (pre-training, fine-tuning, deployment, etc.)? What's the impact of each part and how does data influence interaction?
        - What are the data-perspective efforts in adapting FMs to target tasks/scenarios for deployment? A particular interest is data curation/labeling for fine-tuning, the role of in-context learning, and adaptation in dynamic environments.
    - The capability of foundation models brings it to a new level for generating data. How to control the generation process to produce high-quality or task-relevant data and what can it be used for?
        - Is it good for directly being used to train a model in the same way as natural data? What problem this may cause?
        - How can it help with alignment, improving fairness/safety, or/and adaptation in low-resource scenarios where labeled data is scarce?
    - For FMs pre-trained on massive and broad data, how to consider data acquisition/composition/quality/resource efficiency at scale?
        - Resources requirements for training and deployment of FMs may be out of reach for many, what does this mean for researchers and practitioners working on data problems and how to adapt to the new norm?
        - Scaling laws help in a lot of scenarios and enable research on large models with a small budget, but with complications that some unique capabilities for very large models such as chain-of-thoughts only emerge after a sufficiently large scale. What complications will this cause?

- **B. A Data Perspective to Efficiency, Interpretability, and Alignment–Latest Advancement and Breakthroughs**
    - Other than the impressive capability and generalizability, foundation models have reached unprecedented scales in terms of model size and training data, pronouncing the efficiency problems from all aspects.
        - This includes data efficiency in model training such as the typically resource-intensive pre-training, or fine-tuning at deployment which often has limited labeled data,

        - and also the efficiency of inference methods for interpretability, explainability, fact tracing, etc. How do existing approaches scale up to foundation models and what are the new solutions to these problems?
    - Alignment is one of the most active topics for FMs. How can data-perspective research best contribute to important issues such as data attribution/interpretability, harmlessness/truthfulness, AI safety(fake/harmful contents), etc?

- **C. A Data Perspective to Safety and Ethics–Risks, Limitations, and Opportunities**
    - The unprecedented size and capability of FMs pose unprecedented challenges for safety/trustworthy issues (e.g., jailbreaking, security loopholes, harmful contents, misuse, privacy violations, etc.). <u>What are the risks and current limitations?</u>
        - How data-perspective research can help and which issues benefit most from data-perspective research?
    - How does data-perspective research contribute to the evaluation of FMs (e.g., fairness/ethics, defects of FMs/failure cases)?
        - How to improve these issues from the data side?

D. Copyright, Legal Issues, and Data Economy–A Broader Landscape

- <u>Copyright issues and privacy concerns</u> are the sword of Damocles for the deployment of FMs. What are the current risks and limitations?
    - How data-perspective research can contribute technical, economic, and governance solutions to this topic?
- <u>What is the perspective of data economy?</u> What are the potential market solutions for the acquisition of data?
    - What are the research opportunities for data problems associated with it? How to quantify the value of data, schemes for data exchange, etc.

This workshop aims to bring together pioneers and practitioners on research with data problems for foundation models to discuss its new paradigm and search for a roadmap, facilitating the understanding of emerging research problems that attracted broad interests and exchanging insights on directions forward. We strive to build a community behind this essential topic and provide the platform to connect, share ideas, explore for consensus, and create collaboration opportunities. It is worth mentioning that the current data practices of foundation models are largely opaque[1]. One mission of this workshop is to create a community effort on open source data efforts at the pretraining stage itself. Subsequent efforts include creating datasets, benchmarks (e.g., MLCommons and DataPerf), and dedicated venues (e.g., DMLR) to promote research on data problems for foundation models and ultimately facilitate the widespread deployment of FMs in a sociotechnical-friendly way that provides benefit at large.

Examples of our target communities include researchers on data problems (e.g., data-centric AI, dataset/data curation, data market) and foundation models (alignment, safety/trustworthiness,

[1] Ref: The Foundation Model Transparency Index, https://crfm.stanford.edu/fmti/fmti.pdf

fairness/ethics), practitioners of downstream applications, tech companies providing innovative solutions and beyond.

Modality

The workshop is intended to be a **hybrid** event and live-streamed throughout the day via **Zoom**. Both speakers and audiences will be able to attend either in-person or online and our moderators will help take questions from virtual participants.

Speakers and Panelists:

***ALL** the speakers listed below **have confirmed** their interest in giving a talk at the prospective workshop.

Module A, Data Quality, Dataset Curation, and Data Generation

- - Prof. Luke Zettlemoyer (U Washington/site lead- Meta FAIR Seattle, data generation)
- - Prof. Ludwig Schmidt (U Washington, dataset curation)
- - Prof. Hannaneh Hajishirzi (U Washington/AI2, data quality)
- - Prof. Swabha Swayamdipta (USC, dataset quality)

Module B, Efficiency, Interpretability, Alignment

- - Dr. Mike Lewis (Meta, RoBERTa/LIMA, efficiency/alignment, etc.)
- - Dr. Ari Morcos(Stealth Startup/formerly Meta, data efficiency)
- - Dr. Sara Hooker (Cohere/formerly Google Brain, interpretability)

Module C, Safety, Sociotechnical, and Ethics

- - Prof. Nicolas Papernot (U Toronto, data and security)
- - Dr. Remi Denton (Google, sociotechnical and ethics)
- - Pamela Mishkin (OpenAI, Safety and Fairness)

Module D, Copyright, Legal Issues and Data Economy:

- - Prof. Pamela Samuelson (Berkeley Law, data copyright and legal issues, remote)
- - Prof. Haifeng Xu (U Chicago, data economics)

Biographies:
Module A

	Luke Zettlemoyer is a Professor in the Paul G. Allen School of Computer Science & Engineering at the University of Washington, and a Research Director at Meta. His research focuses on empirical methods for natural language semantics, and involves designing machine learning algorithms, introducing new tasks and datasets, and, most recently, studying how to best develop self-supervision signals for pre-training. His honors include being named an ACL Fellow as well as winning a PECASE award, an Allen Distinguished Investigator award, and multiple best paper awards. Luke received his PhD from MIT and was a postdoc at the University of Edinburgh.
	Ludwig Schmidt is an assistant professor in computer science at the University of Washington. His research interests revolve around the foundations of machine learning, often with a focus on datasets, evaluation, reliable generalization, and large models. He is also a research scientist in the AllenNLP team at AI2 and a member of LAION. Before joining UW, he was a postdoc at UC Berkeley where his mentors were Moritz Hardt and Ben Recht. He received my PhD from MIT advised by Piotr Indyk. His group also contributes to machine learning via code repositories and datasets, e.g., OpenCLIP, OpenFlamingo, and LAION-5B.
	Hanna Hajishirzi is a Torode Family Associate Professor at the University of Washington and a Senior Director of NLP at AI2. Her research spans different areas in NLP and AI, more recently on the science of language models and language models for science. Honors include an NSF CAREER, Sloan Fellowship, Allen Distinguished Investigator Award, Intel rising star award, UIUC alumni award. She has received a best paper and several honorable mention paper awards.
	Swabha Swayamdipta is an Assistant Professor of Computer Science and a Gabilan Assistant Professor at the University of Southern California. Her research interests are in natural language processing and machine learning, with a primary interest in the estimation of dataset quality, understanding and evaluation of generative models of language, and using language technologies to understand social behavior. At USC, Swabha leads the Data, Interpretability, Language and Learning (DILL) Lab. She received her PhD from Carnegie Mellon University, followed by a postdoc at the Allen Institute for AI. Her work has received outstanding paper awards at ICML 2022, NeurIPS 2021 and an honorable mention for the best paper at ACL 2020. Her research is supported by awards from the Allen Institute for AI and Intel Labs.

Module B

	Mike Lewis is a research scientist at Meta AI working on open foundation models. Prior projects include the Cicero Diplomacy agent, and the Bart and Roberta pretrained language models. Previously he was a postdoc at the University of Washington (working with Luke Zettlemoyer), and has a PhD from the University of Edinburgh (advised by Mark Steedman). He received a Best Paper Award at EMNLP 2016, Best Resource Paper at ACL 2017, and Best Paper Honourable Mention at ACL 2018. His work has been extensively covered in the media, with varying levels of accuracy.
	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.
	Sara Hooker is a Director at Cohere and she leads Cohere For AI, a research lab that seeks to solve complex machine learning problems. She led a team of researchers and engineers working on making large language models more efficient, safe and grounded. Prior to Cohere, she was a research scientist at Google Brain doing work on training models that go beyond test-set accuracy to fulfill multiple desired criteria -- interpretable, compact, fair and robust.

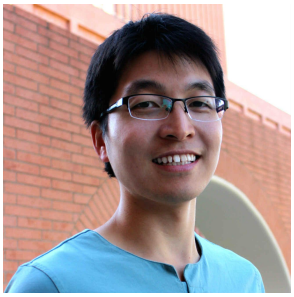
Module C

	Nicolas Papernot is an Assistant Professor at the University of Toronto, in the Department of Electrical and Computer Engineering and the Department of Computer Science. He is also a faculty member at the Vector Institute where he hold a Canada CIFAR AI Chair, and a faculty affiliate at the Schwartz Reisman Institute. He was named an Alfred P. Sloan Research Fellow in Computer Science in 2022 and a Member of the Royal Society of Canada College in 2023. His research interests are at the intersection of security, privacy, and machine learning. His research has been cited in the press, including the BBC, New York Times, Popular Science, The Atlantic, the Wall Street Journal and Wired. He currently serve as a Program Committee Chair of the IEEE Conference on Secure and Trustworthy Machine Learning (SaTML), which he co-founded in

	2023. He earned my Ph.D. in Computer Science and Engineering at the Pennsylvania State University, working with Prof. Patrick McDaniel and supported by a Google PhD Fellowship. Upon graduating, he joined Google Brain for a year; He continue to spend time at Google DeepMind.
	Dr. Remi Denton (they/them) is a Staff Research Scientist at Google, within the Technology, AI, Society, and Culture team, where they study the sociocultural impacts of AI technologies and conditions of AI development. Their recent research centers on emerging text- and image-based generative AI, with a focus on data considerations and representational harms. Prior to joining Google, Emily received their PhD in Computer Science from the Courant Institute of Mathematical Sciences at New York University, where they focused on unsupervised learning and generative modeling of images and video. Prior to that, they received their B.S. in Computer Science and Cognitive Science at the University of Toronto. Though trained formally as a computer scientist, Emily draws ideas and methods from multiple disciplines and is drawn towards highly interdisciplinary collaborations, in order to examine AI systems from a sociotechnical perspective. They've published in multiple top-tier venues spanning social science and computing disciplines, including Big Data & Society, CSCW, FAccT, and NeurIPS.
	Pamela Mishkin is a researcher at OpenAI. She is interested in how to make language models safe and fair, from a technical and policy perspective. She previously led product management at The Whistle, a small start-up building tech tools for international human rights groups. Before that, she researched economic policy at the Federal Reserve Bank of New York and worked with the Department of Digital Culture, Media and Sport in the UK on online advertising policy. She has a BA in Computer Science and Math from Williams College and an MPhil in Technology Policy from the University of Cambridge (Herchel-Smith Fellow).

Module D

	Pamela Samuelson is the Richard M. Sherman Distinguished Professor of Law and Information at the University of California, Berkeley. She is recognized as a pioneer in digital copyright law, intellectual property, cyberlaw and information policy. Since 1996, she has held a joint appointment at Berkeley Law School and UC Berkeley's School of Information. Samuelson is a director of the internationally-renowned Berkeley Center for Law & Technology. She is co-founder and chair of the board of Authors Alliance, a nonprofit organization that promotes the public interest in access to knowledge. She also serves on the board of directors of the Electronic Frontier Foundation, as well as on the advisory boards for the Electronic Privacy Information Center, the Center for Democracy & Technology, and Public Knowledge. Samuelson has written and

	published extensively in the areas of copyright, software protection and cyberlaw. She is also a Fellow of the Association of Computing Machinery.
	**Haifeng Xu** is an assistant professor in computer science at the University of Chicago, and directs the Strategic Intelligence for Machine Agents (SIGMA) lab which focuses on designing intelligent AI systems that can effectively learn and act in informationally complex multi-agent setups. His recent research focuses on the economics of machine learning, including designing markets for data and MLaaS, and designing ML algorithms in economic contexts. Haifeng has been recognized by a few awards, including Early Career Spotlight of the International Joint Conference on Artificial Intelligence (IJCAI), a Google Faculty Research Award, and honorable mention for both ACM SIGecom Dissertation Award and IFAAMAS Distinguished Dissertation Award. Prior to UChicago, he was the Alan Batson Assistant Professor at the University of Virginia, and even before that a postdoc at Harvard.

Tentative schedule

We aim to bring together researchers, practitioners, and broader audiences alike to navigate and explore data problems with foundation models and search for a roadmap forward. A core part of the workshop is to provide a platform for exchanging insights, sharing ideas, and promoting fruitful collaborations. Thus, we intertwine the sessions with talks by people on the front lines and frequent intermissions for discussion to best facilitate the exchange of ideas between participants and also with the speakers. This also allows personnel exchange between different workshops and connects to broader audiences. We offer presentation opportunities for the high-quality works submitted to the workshop to showcase versatile progress currently being made. We will take questions from online participants through moderated discussions. Awards for accepted papers will be selected based on nominations of reviewers and the consensus of the organizing committee (excluding conflicts of interest) before the day of the event.

Keynote talks are allocated 35 minutes each (25 minutes talk + 10 minutes discussion); subsequent invited talks are allocated 30 minutes each (25 minutes talk + 5 minutes discussion).

We provide a tentative schedule as the following *(specific timing such as for coffee breaks will be adjusted to keep in line with the official program for the conference)*

Time	**Session**	**Topic**
8:00-8:10	Introduction. Opening remarks.	-

8:10-8:45	Keynote talk #1–Module A.	**Data Quality**
8:45-9:20	Keynote talk #2–Module A.	**Dataset Curation**
9:20-9:50	Invited talk #1–Module A.	**Data Generation**
9:50-10:05	*Coffee break #1/discussion/poster*	
10:05-10:40	Keynote talk #3–Module D.	**Copyright and Legal Issues**
10:40-11:10	Invited talk #2–Module A.	**Data Curation**
11:10-11:40	Invited talk #3–Module D.	**Data Economy**
11:40-12:10	Best paper presentations. (10min*3)	-
12:10-1:10	*Lunch break/discussion/poster*	
1:10-1:45	Keynote talk #4–Module B.	**Efficiency, Interpretability, Alignment**
1:45-2:15	Invited talk #4–Module B.	**Efficiency, Interpretability, Alignment**
2:15-2:50	Keynote talk #5–Module C.	**Safety, Sociotechnical, and Ethics**
2:50-3:05	*Coffee break #2/discussion/poster*	
3:05-3:35	Invited talk #5–Module C.	**Safety, Sociotechnical, and Ethics**
3:35-4:05	Invited talk #6–Module B.	**Efficiency, Interpretability, Alignment**
4:05-4:35	Invited talk #7–Module C.	**Safety, Sociotechnical, and Ethics**
4:35-5:00	Best paper presentations. (5min*5)	-
5:00-5:10	Announcement of awards. Conclusion. Closing Remarks.	
5:10-6:00	*Discussion/poster*	
6:00	End	-

We plan to publish a position paper together at the end of this workshop. A tentative venue is the new Journal of Data-centric Machine Learning Research (DMLR), which has recently started to accept submissions.

Organizers and Biographies

Members of the organizing team:

- Prof. Ruoxi Jia (Assistant professor, Virginia Tech)
- Prof. Tatsunori Hashimoto (Assistant professor, Stanford University)
- Prof. Pang Wei Koh (Assistant professor, University of Washington)
- Dr. Jerone Andrews (Research Scientist, Sony AI)
- Sang Michael Xie (PhD student, Stanford University)
- Lingjiao Chen (PhD student, Stanford University)
- Feiyang Kang (PhD student, Virginia Tech)

The organizing team is composed of a balanced proportion of experienced researchers with abundant past organizing of professional events and PhD students who are first-time workshop organizers. **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** co-organized an AI and Future Crime Workshop in 2019, drawing attendees from the UK government and the financial sector. **Sang Michael Xie** co-organized the ME-FoMo Workshop at ICLR 2023. **Lingjiao Chen and Feiyang Kang are first-time workshop organizers.**

Biographies:

	Ruoxi Jia is an assistant professor in the the Bradley Department of Electrical and Computer Engineering at Virginia Tech. She earned her PhD in the EECS Department from UC Berkeley and a B.S. from Peking University. Jia's recent work focuses on data-centric and trustworthy machine learning. Ruoxi is the recipient of the NSF CAREER Award, the Chiang Fellowship for Graduate Scholars in Manufacturing and Engineering, the 8108 Alumni Fellowship, and the Okamatsu Fellowship, Virginia's Commonwealth Cyber Initiative award, Cisco Research Awards, and Amazon-VT Initiative Research Awards. She was selected for the Rising Stars in EECS in 2017. Ruoxi's work has been featured in multiple media outlets such as MIT Technology Review, New York Times, IEEE Spectrum, and Wired. Her work has been adopted in the financial sector and tech companies. **Previous organizational experience**: She 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).

	Tatsunori Hashimoto 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.
	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. **Previous organizational experience:** He organized the Workshop on Distribution Shifts at NeurIPS 2021–2023.
	Jerone Andrews 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). **Previous organizational experience:** In 2019, he co-organized an AI and Future Crime Workshop, drawing attendees from the UK government and the financial sector.

	Sang Michael Xie is a sixth year PhD student at Stanford University advised by Percy Liang and Tengyu Ma. He has worked on data-centric methods for language model pretraining, theory and methods for understanding how pretraining improves transfer to downstream tasks, and pretraining and self-training on unlabeled data for reliable machine learning. He is a recipient of the NDSEG Fellowship. Outside of Stanford, he has interned at Google Brain. **Previous organizational experience:** He co-organized the ME-FoMo Workshop at ICLR 2023.
	Lingjiao Chen is a PhD candidate at Stanford University, advised by Prof. Matei Zaharia and Prof. James Zou. His research interest lies broadly at the intersection of data systems and machine learning, with a recent focus on the generative AI service market. His work has been supported by a Google Fellowship and covered by mainstream media including Fortune, The Wall Street Journal, and Scientific American. **First-time workshop organizer.**
	Feiyang Kang is a PhD student in the Bradley Department of Electrical and Computer Engineering at Virginia Tech, advised by Prof. Ruoxi Jia. His research interests lie in data-centric methods for trustworthy machine learning. His recent projects involve data valuation, data selection, data influence with applications in large language models and computer vision. His work has been adopted into production pipelines by leading tech companies. **First-time workshop organizer.**

Diversity Commitment

We strive to achieve balance and parity across multiple levels–from seniority levels, field/community, position/location, to personal backgrounds for invited speakers, organizing team, covered topics, and targeted audiences.

Our invited speakers are diverse in terms of seniority levels (established/junior researchers), positions (academic/major tech corporations/startups), gender (45% female, 45% male, 10% non-binary), and race/ethnicity (multiple). Our organizing team is made up of 43% academic researchers, 14% industrial researchers, and 43% PhD students from different personal backgrounds and geographic locations. The workshop agenda has a comprehensive agenda for

different types of data problems and dedicated no less than half the sessions to broader considerations including safety/fairness, ethics/alignment, legal issues, etc.

With the widespread interest and usage of foundation models, this workshop is dedicated to attracting audiences at large rather than from specific sub-fields. The organizing team plans to spend substantial effort in promoting the workshop to participants from diverse fields and backgrounds.

Access

We will maintain a dedicated website for the workshop and promote it to audiences. For invited talks and panel discussions, we will post the abstract of the talk and biography of the speakers as soon as they are confirmed (we anticipate this to be months ahead of the workshop). We will make our best effort to make the slides and recordings of talks available online after the event. The workshop is intended to be a hybrid event and live streamed throughout the day of workshop. Both speakers and audiences will be able to attend either in-person or online and we will take questions from virtual participants. Accepted papers will be posted on the website and will also be provided the opportunity to be published on arXiv with notes on them being part of the workshop.

Funding

Upon acceptance, the organizing team of the workshop will solicit sponsorship from companies including Amazon, Sony, and federal agencies. The funding will be used for accessibility (tech support for live streaming and maintenance of the website) and for creating awards for best long/short papers and best talks. The remaining funding will be awarded to student participants for travel grants. The exact amount for each award will depend on the availability of funding.

Anticipated audience size

Based on successful outcomes of previous workshops on related topics, the relevance of the proposed agenda to audiences, and the attractiveness of the speakers, we expect the workshop to host 100-200 people in the room at all times and cover 200-500 audiences in total throughout the event.

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

To ensure the success of the workshop and maximize the coverage to potential audiences, the organizing team plans to spend substantial effort in promoting this event. We consider the following approaches: 1. advertise the proposed workshop in the community of upcoming conferences (e.g., NeurIPS 2023) and related workshops; 2. send out workshop flyers and

call-for-papers in related-community/institutions/companies; 3. promote the workshop on social media platforms (e.g., Twitter) /workshop websites/personal websites.

Previous related workshops

This is the first workshop dedicated to the emerging research on data problems for foundation models. The topic is still at a fairly early stage and we aim to bring together front-line researchers to discuss what's the new research paradigm and search for a roadmap forward. This workshop differentiates from existing workshops/series. A series of workshops on data-centric AI emerged in recent years and has been hosted successfully and attracting growing attention, including Data-centric Machine Learning Research (DMLR) Workshop at ICML'23, DataPerf: Benchmarking Data for Data-Centric AI at ICML'22, Data-Centric AI Workshop at NeurIPS'21, Economics of Privacy and Data Labor at ICML'20. The emergence of foundation models and their success brings up unique challenges and opportunities that hold the potential to substantially reshape the current research landscape. Also, the past two years have seen a wealth of workshops on foundation models, including ES-FoMo: Efficient Systems for Foundation Models at ICML'23, Foundation Models for Decision Making at NeurIPS'23 & NeurIPS'22, Mathematical and Empirical Understanding of Foundation Models (ME-FoMo) Workshop at ICLR'23, Trustworthy and Reliable Large-Scale Machine Learning Models Workshop at ICLR'23. Research from data perspectives holds a promising position in contributing to some of the most interesting and important research questions related to foundation models and their deployment. We possess a keen interest in facilitating the success of research on data problems in the era of foundation models and continuing this workshop as a series as we anticipate the interests and usages continue to grow in the foreseeable future.

Program Committee and Reviewing

The Reviewing process will be conducted strictly adhering to the workshop guidelines and the organizing team plan will make efforts to ensure conflicts of interest are avoided to the maximum extent possible. Specifically, organizers will limit their extent of involvement in the reviewing process and at a minimum, will not handle submitted works from the same organization/known people/listed conflict of interests. Initial review of submitted works will be performed by the Program Committee in a double-blind manner. Each submitted work will be assigned 3-4 reviewers and each reviewer will be assigned 3-4 papers. The organizing team will serve as meta-reviewers to the extent allowed by the guidelines. At the end of the reviewing period, meta-reviewers will work to make sure each work receives 3 reviews where at least 2 are high-quality. Decisions for acceptance will be made by meta-reviewers based on the initial reviews. Reviewers will also be asked whether to nominate the work for awards. The final decision on awards will be made by the meta-reviewers unanimously while avoiding conflicts of interest. Call-for-paper will solicit novel and original works and explicitly discourage submissions that have already been published and presented at other conferences.

Based on the attractiveness and popularity of the scope of the workshop and the broad span of the 4 modules, we anticipate receiving 100~200 submissions. Thus, we will work to ensure the same number of reviewers for the Program Committee. We will ask each submission to nominate one person to serve on the Program Committee. We list below a pool of candidate reviewers with the expertise needed for the Program Committee (e.g., data problems for foundation models, research on foundation models, data-centric AI, etc.). Upon acceptance of this proposal, we will start to solicit people to serve on the Program Committee.

Pool of Candidate Reviewers

Aakanksha Chowdhery chowdhery@google.com	Nikunj Saunshi nsaunshi@google.com
Abulhair Saparov as17582@nyu.edu	Patrick Fernandes pattuga@gmail.com
Adam Fisch fisch@mit.edu	Pradeep Dasigi pradeepd@allenai.org
Ahana Gangopadhyay ahana@wustl.edu	Raj Ratn Pranesh raj.ratn18@gmail.com
Aishwarya Balwani abalwani6@gatech.edu	Rishi Bommasani nlprishi@stanford.edu
Akhilesh Deepak Gotmare dg.akhilesh@gmail.com	Rohan Taori rtaori@cs.stanford.edu
Albert Gu albertfgu@gmail.com	Rohith Kuditipudi rohithk@stanford.edu
Alexander Wettig awettig@cs.princeton.edu	Ruibo Liu ruibo.liu.gr@dartmouth.edu
Aman Madaan amadaan@cs.cmu.edu	Ruoqi Shen shenr3@cs.washington.edu
Amir Bar amirb4r@gmail.com	Saachi Jain saachij@mit.edu
Asher Trockman ashert@cs.cmu.edu	Sachin Goyal sachingo@andrew.cmu.edu
Ashish Hooda ashish1995hooda@gmail.com	Sadhika Malladi SadhikaMalladi@princeton.edu
Ashwini Pokle apokle@andrew.cmu.edu	Sameera Horawalavithana yasanka.horawalavithana@pnnl.gov
Avinash Madasu avinashmadasu17@gmail.com	Sanjay Subramanian sanjayss@berkeley.edu
Benjamin L. Edelman bedelman@g.harvard.edu	Sarah Masud Sarah.Masud.Preum@dartmouth.edu
Benjamin Newman bnewman@ucr.edu	Saurabh Garg sgarg2@andrew.cmu.edu
Bilal Alsallakh bilal@voxelai.com	Sewon Min sewon@cs.washington.edu
Bingchen Zhao zhaobc.gm@gmail.com	Shashank Shekhar shekhar@emory.edu
Charlie Victor Snell csnell22@berkeley.edu	Sheng Shen sheng.s@berkeley.edu
Christina Baek ke.baek@berkeley.edu	Shivam Garg shivamgarg@stanford.edu
Collin Burns collin.burns@columbia.edu	Shulei Wang shuleiw@illinois.edu
Daniel Y Fu danfu@stanford.edu	Shuyan Zhou shuyanzh@cs.cmu.edu
Dingli Yu dingliy@cs.princeton.edu	Simon Kornblith simon@simonster.com
Ellyn Ayton ellyn.ayton@pnnl.gov	Simran Arora simarora@stanford.edu
Erik Jones erjones@berkeley.edu	Tejasri Nampally ai19resch11002@iith.ac.in
Hao Tan sameidasabove@adobe.com	Thao Nguyen thaottn@cs.washington.edu
Haokun Liu haokunl@cs.unc.edu	Tianle Cai tianle.cai@princeton.edu
Haonan Duan haonand@cs.toronto.edu	Tianwei Yue tyue@andrew.cmu.edu
Hong Liu hongliu9903@gmail.com	Tianyu Gao tianyug@princeton.edu
Iz Beltagy beltagy@allenai.org	Tim Dettmers Dettmers@cs.washington.edu
Jacob M. Springer jspringer@cmu.edu	Tomasz Korbak tomasz.korbak@gmail.com
Jeff Z. HaoChen jhaochen@cs.stanford.edu	Udita Patel upatel22@umd.edu
Jiacheng Liu liujc@cs.washington.edu	Vaishnavh Nagarajan vaishnavh@google.com
Jianan Zhou JIANAN004@e.ntu.edu.sg	Vamsi Aribandi aribandiv@gmail.com
Jingling Li jingling@cs.umd.edu	Vanya Bannihatti Kumar

Jinqi Luo jinqiluo@upenn.edu Kamilė Lukošiūtė KamileLukosiute@gmail.com Kefan Dong kefandong@stanford.edu Kevin Miao kevinmiao@cs.berkeley.edu Liunian Harold Li liunian.harold.li@cs.ucla.edu Lucio M. Dery ldery@andrew.cmu.edu Mayee F Chen mfchen@stanford.edu Mengzhou Xia mengzhou@princeton.edu Michael Zhang mzhang@cs.stanford.edu Mirac Suzgun msuzgun@stanford.edu Nikhil Vyas nikhil@g.harvard.edu Pratyush Maini pratyushmaini@cmu.edu Preetum Nakkiran preetum@nakkiran.org Qian Huang qhwang@cs.stanford.edu Adams Wei Yu <adamsyuwei@gmail.com> Dmitris Tsipras <tsipras@stanford.edu> David Adelani <d.adelani@ucl.ac.uk> Nikhil Kandpal <nkandpa2@cs.unc.edu> Philip Keung <keung@amazon.com> Stanislav Fort <stan@anthropic.com> Mayee Chen <mfchen@stanford.edu> Yi Tay <ytay017@gmail.com> Lisa Dunlap <lisabdunlap@berkeley.edu> Suchin Gururangan <sg01@cs.washington.edu> Hadi Salman <hady@mit.edu> Ellen Wu <zeqiuwu1@uw.edu> Steve Mussman <mussmann@cs.washington.edu> Kawin Ethayarajh <kawin@stanford.edu> Gabriel Ilharco <gamaga@cs.washington.edu> Andrew Ilyas <ailyas@mit.edu> Kenton Lee <kentonl@google.com> Ilija Radosavovic <ilija@berkeley.edu> Yash Savani <ysavani@cs.cmu.edu>	vbanniha@andrew.cmu.edu Vishakh Padmakumar vishakh@nyu.edu Wei Hu vvh@umich.edu Xiang Lisa Li xlisali@stanford.edu Xiang Wang xiangwang1223@gmail.com Xiangyu Yue xyyue@ie.cuhk.edu.hk Xindi Wu xindiw@princeton.edu Xuechen Li lxuechen@cs.stanford.edu Yanda Chen yc3384@columbia.edu Yangjun Ruan yjruan@cs.toronto.edu Yann Dubois yanndubs@stanford.edu Yara Rizk yara.rizk@ibm.com Yining Chen cynnjjs@cs.stanford.edu Yiyuan Li yiyuanli@cs.unc.edu Yoonho Lee yoonholee95@gmail.com Yossi Gandelsman yossi@gandelsman.com Zhiyuan Li zhiyuanli@stanford.edu Zi-Yi Dou zdou@cs.ucla.edu Yiyuan Li <bill.lyy.nisioptimum@gmail.com> Aakanksha Chowdhery <chowdhery@google.com> Karan Goel <kgoel@cs.stanford.edu> Michi Yasanuga <myasu@cs.stanford.edu> Mattia Rigotti <mrg@zurich.ibm.vom> Alaa Khaddaja <alaakh@mit.edu> Zhuohan Li <zhuohan@cs.berkeley.edu> Tim Brooks <tim@timothybrooks.com> Nitish Joshi <nitish@nyu.edu> Vaishaal Shankar <vs@vaishaal.com> Jesse Dodge <jessed@allenai.org> Victor Zhong <victor@victorzhong.com> Derek Tam <dt.derek.tam@gmail.com> Yizhong Wang <yizhongw@cs.washington.edu> Lavinia F. Pieptea <laviniaflorentinapieptea@my.unt.edu> Kai Xiao <kaix@mit.edu> Yu Sun <yusun@berkeley.edu>

Extended List:

Kevin Duh (Johns Hopkins University) Bang Liu (University of Montreal (UdM)) Hamidreza Mahyar (McMaster University) Wenhu Chen (University of Waterloo) Yue Dong (University of California) Lili Mou (University of Alberta) Peyman Passban (BenchSci) Aref Jafari (University of Waterloo) Vasileios Lioutas (University of British	Dan Alistarh (Institute of Science and Technology Austria) Bang Liu (University of Montreal (UdM)) Hassan Sajjad (Dalhousie University) Tiago Falk(INRS University) Yu Cheng (Microsoft) Anderson R. Avila (INRS University) Peyman Passban (BenchSci) Rasoul Kaljahi (Oracle)

Colombia (UBC)) Malik H. Altakrori (McGill University & MILA) Ali Vahdat (Thomson Reuters) Prasanna Parthasarathi (McGill University & MILA) Shohreh Shaghaghian (Thomson Reuters) Ehsan Kamalloo (University of Alberta) Ali Saheb Pasand (University of Waterloo) Soheila Samiee (BASF) Mohammed Senoussaoui (INRS) Flávio Ávila (Amazon) Peng Lu (Huawei Noah's Ark Lab) Joao Monteiro (Service Now) Xiaoguang Li (Huawei Noah's Ark Lab) Can Liu (Amazon Alexa AI) Amina Shabbeer (Amazon) M. Skylar Versage (Amazon) Tanya Roosta (Amazon) Prashanth Rao (Royal Bank of Canada) Ovidiu Serban (Imperial College London) Tony Tong (Royal Bank of Canada) Jiahao Sun (Royal Bank of Canada) Ryan Ong (Imperial College London) Weihang Zhang (Imperial College London) Manying Zhang (Institut National des Langues et Civilisations Orientales) Lianlong Wu (Oxford University) Mojtaba Valipour (University of Waterloo) Chandra Bhagavatula (Allen Institute for AI) Jinming Zhao (Monash University) Khalil Slimi (ServiceNow) Mohammadreza Tayaranian (Huawei Noah's Ark Lab) Ning Shi (University of Alberta) Weiyi Lu (Amazon)	Joao Monteiro (Service Now) Shahab Jalalvand (Interactions) Aref Jafari (University of Waterloo) Jad Kabbara (MIT) Ahmad Rashid (University of Waterloo) Ehsan Kamalloo (University of Alberta) Soheila Samiee (BASF) Hamidreza Mahyar (McMaster University) Flávio Ávila (Verisk Analytics) Peng Lu (UdeM) Mauajama Firdaus (University of Alberta) Tanya Roosta (Amazon) Tianyu Jiang (University of Utah, US) Juncheng Yin (Western University, Canada) Jingjing Li (Alibaba) Meng Cao (McGill and Mila, Canada) Wen Xiao (UBC, Canada) Chenyang Huang (University of Alberta) Mojtaba Valipour (University of Waterloo) Lili Mou (University of Alberta) Yue Dong (University of California, Riverside) Makesh Sreedhar (NVIDIA) Hossein Rajabzadeh (University of Waterloo) Mohammadreza Tayaranian (McGill University) Suyuchen Wang (MILA) Crystina Zhang (University of Waterloo) Parsa Kavehzadeh (York University) Nandan Thakur (University of Waterloo) Heitor Guimarães (INRS University) Amirhossein Kazemnejad (McGill University/MILA) Hamidreza Saghir (Microsoft) Yuqiao Wen (University of Alberta) Arthur Pimentel (INRS University)