Title: Will Synthetic Data Finally Solve the Data Access Problem?

Workshop summary

Accessing large scale and high quality data has been shown to be one of the most important factors to the performance of machine learning models. Recent works show that large (language) models can greatly benefit from training with massive data from diverse (domain specific) sources and aligning with user intention. However, the use of certain data sources can trigger privacy, fairness, copyright, and safety concerns. The impressive performance of generative artificial intelligence popularized the usage of synthetic data, and many recent works suggest (guided) synthesization can be useful for both general purpose and domain specific applications. For example, Yu et al. 2024, Xie et al. 2024, Hou et al. 2024 demonstrate promising preliminary results in synthesizing private-like data, while Wu et al. 2024 highlight existing gaps and challenges. As techniques like self-instruct (Wang et al. 2021) and self-alignment (Li et al. 2024) gain traction, researchers are questioning the implications of synthetic data (Alemohammad et al. 2023, Dohmatob et al. 2024, Shumailov et al. 2024).

Will synthetic data ultimately solve the data access problem for machine learning? This workshop seeks to address this question by highlighting the limitations and opportunities of synthetic data. It aims to bring together researchers working on algorithms and applications of synthetic data, general data access for machine learning, privacy-preserving methods such as federated learning and differential privacy, and large model training experts to discuss lessons learned and chart important future directions.

Topics of interest include, but are not limited to, the following:

- Risks and limitations of synthetic data.
- New algorithms for synthetic data generation.
- New applications of using synthetic data (e.g. in healthcare, finance, gaming and simulation, education, scientific research, or autonomous systems).
- Synthetic data for model training and evaluation.
- Synthetic data for improving specific model capabilities (e.g., reasoning, math, coding).
- Synthetic data to address privacy, fairness, safety and other data concerns.
- Evaluation of synthetic data quality and models trained on synthetic data.
- Conditional and unconditional synthetic data generation.
- Fine-grained control of synthetic data generation.
- Data access with federated learning and privacy-preserving methods.
- New paradigm of accessing data for machine learning.
- Mixing synthetic and natural data.

Tentative schedule

We have planned for invited talks, spotlight talks from contributed papers, poster sessions, and a panel discussion. We highlight the exchange of ideas through question and answer for talks, and active discussion in panel and poster sessions. We have listed a tentative schedule.

8:55-9:05 Opening remarks
9:00-9:30 Invited speaker 1
9:30-10:00 Spotlight paper talks
10-10:30 Break
10:30-11:00 Invited speaker 2
11:00-11:30 Invited speaker 3
11:30-12:30 Poster
12:30-1:30 Lunch break

1:30-2:30 Panel discussion
2:30-3:00 Invited speaker 4
3:00-3:30 Spotlight paper talks
3:30-4:00 Break
4:00-4:30 Invited speaker 5
4:30-5:00 Invited speaker 6
5:00-5:05 Concluding remarks

Invited speakers and panelists

We have invited seven speakers with diverse backgrounds, and five speakers already confirmed.

Confirmed speakers

1. Mary-Anne Hartley (confirmed) is an Assistant Professor at Yale and EPFL. Her recent work on large medical models show the importance of accessing clinical data with the potential to fulfill generative tasks to tackle factuality and trust concerns.
2. Sanmi Koyejo (confirmed) is an Assistant Professor at Stanford working on Trustworthy AI. His award winning research highlights the importance of (data for) evaluating GenAI, and recent papers study the limits of synthetic data.
3. Natalia Ponomareva (confirmed) is a Staff Software Engineer at Google leading a team on private synthetic data. She led the DP-fy ML guide working group and delivered tutorials at ICML and KDD, and she advocated the usage of DP synthetic data in practice.
4. Mihaela van der Schaar (confirmed) is the John Humphrey Plummer Professor of Machine Learning, AI and Medicine at the University of Cambridge. She was the senior organizer of Synthetic Data Workshop NeurIPS'23, and a world-renowned expert on this topic.
5. Eric Xing (confirmed) is the President at Mohamed bin Zayed University of AI, and a Professor at Carnegie Mellon University. His recent work on large models training (LLM360) and evaluation (Chatbot Arena) discuss open-sourcing not only models, but also data.

Pending reply

6. Thomas Scialom (pending) is a Research Scientist at Meta leading Llama 2 and Postraining Llama 3, one of the most well known open-sourced large models. He also worked on CodeLlama, Galactica, Toolformer, Bloom, Nougat, GAIA, etc. Synthetic data is often used in developing large models.
7. Phillip Isola (pending) is an Associate Professor at Massachusetts Institute of Technology. His work pioneered vision generative models, and broadly into generative intelligence and emergent intelligence.

We will also host a panel to discuss the Future of Synthetic Data and Data Access. The potential panelists include Kate Downing (IP attorney), Tom Goldstein (UMD), Julia Kempe (NYU), Arash Vahdat (Nvidia) and Qiang Yang (HKUST).

Organizers

Contact: Peter Kairouz and Zheng Xu {kairouz, xuzheng}@google.com

Herbie Bradley (UK AI Safety Institute) mail@herbiebradley.com Scholar user=oQ0HzPcAAAAJ

- **Bio:** Herbie Bradley is a Research Scientist in the UK AI Safety Institute, working on research to develop more predictive evaluations for advanced AI systems and support AI governance. Herbie is also a fourth-year PhD candidate at the University of Cambridge advised by Adrian Weller, and his PhD research centers on synthetic data generation techniques and open-endedness for large language models.

Rachel Cummings (Columbia University) rac2239@columbia.edu Scholar user=2mYxmokAAAAJ

- **Bio:** Dr. Rachel Cummings is an Associate Professor of Industrial Engineering and Operations Research and (by courtesy) Computer Science at Columbia University. Before joining Columbia, she was an Assistant Professor of Industrial and Systems Engineering and (by courtesy) Computer Science at the Georgia Institute of Technology. Her research interests lie primarily in data privacy, with connections to machine learning, algorithmic economics, optimization, statistics, and public policy. Dr. Cummings is the recipient of numerous awards including an NSF CAREER award, a DARPA Young Faculty Award, a DARPA Director's Fellowship, an Early Career Impact Award, multiple industry research awards, a Provost's Teaching Award, two doctoral dissertation awards, and Best Paper Awards at DISC 2014, CCS 2021, and SaTML 2023. She has previously organized the Workshop on Contextual Integrity for Differential Privacy (2023), the Connections Workshop: Algorithms, Fairness, and Equity (2023), the Workshop on Societal Considerations and Applications (2022), the Workshop on Theory and Practice of Differential Privacy at ICML (2020, 2021), Workshop on Synthetic Data for AI in Finance at ICAIF (2022), Workshop on Synthetic Data Generation: Quality, Privacy, and Bias at ICLR (2021), Workshop on Economic Implications of Data at ICML (2020), Workshop on Privacy-Preserving Artificial Intelligence at AAAI (2020), and the Data Science for Social Good Workshop (2019, 2020).

Giulia Fanti (CMU) gfanti@andrew.cmu.edu Scholar user=Rn_BmTYAAAAJ

- **Bio:** Giulia Fanti is an Assistant Professor of Electrical and Computer Engineering at Carnegie Mellon University. Her research interests span the security, privacy, and efficiency of distributed systems. She is a two-time fellow of the World Economic Forum's Global Future Council on Cybersecurity and a member of NIST's Information Security and Privacy Advisory Board. Her work has been recognized with several awards, including best paper awards, a Sloan Fellowship, an Intel Rising Star Faculty Award, an ACM SIGMETRICS Rising Star Award, an NSF CAREER Award, and an Air Force Young Investigator Award. She has co-organized several events, including ACM SIGMETRICS 2024 (TPC Co-Chair), the CyLab-Africa Summit on Digital Public Goods (Co-Chair), and the 2022-2023 Workshops on Synthetic Data for AI in Finance at ICAIF (Co-Organizer).

Peter Kairouz (Google) kairouz@google.com Scholar user=m8NUgw0AAAAJ

- **Bio:** Peter Kairouz is a research scientist at Google, where he focuses on researching and building federated learning, differential privacy, and other privacy-enhancing technologies. Before joining Google, he was a Postdoctoral Research Fellow at Stanford University. He received his Ph.D. in electrical and computer engineering from the University of Illinois at Urbana-Champaign (UIUC). He is the recipient of the 2012 Roberto Padovani Scholarship from Qualcomm's Research Center, the 2015 ACM SIGMETRICS Best Paper Award, the 2015 Qualcomm Innovation Fellowship Finalist Award, and the 2016 Harold L. Olesen Award for Excellence in Undergraduate Teaching from UIUC, and the 2021 ACM Conference on Computer and Communications Security (CCS) Best Paper Award. Dr. Kairouz has organized several Google-hosted workshops on Private Learning and Analytics (2018, 2019, and 2020), and was a lead organizer of Federated Learning and Analytics in Practice Workshop at ICML 2023. He has given multiple tutorials on the same topic, including one at NeurIPS 2020 and another at AAAI 2021. He has served as the associate editor-in-chief for the IEEE JSAC Series on Machine Learning in Communications and Networks, and as an area chair of AISTATS 2023, 2024, and 2025. He has also led a 58-author effort that produced one of the most comprehensive surveys on advances and open problems in the field, which was published in Foundations and Trends of Machine Learning.

Lipika Ramaswamy (Gretel) lipika@gretel.ai

- **Bio:** Lipika Ramaswamy is a Staff Applied Scientist at Gretel. Her interests lie in practical implementations of privacy research for industry applications. She leads privacy research at Gretel, a multimodal synthetic data platform, delivering product features such as differentially private learning for generative models and empirical privacy measures for synthetic data. Prior to Gretel, Lipika had several years of industry experience building and deploying privacy preserving data analysis frameworks in enterprise software. She holds a Masters in Data Science from Harvard University, where she worked on usability of differential privacy software tools at the Privacy Tools Project. Lipika is an active member of the applied privacy community - she has delivered talks and hosted workshops on topics like adversarial attacks on LLM outputs and mitigation strategies (The Rise of Privacy Tech 2022, Open Data Science Conference 2022, Big Data & AI 2023, Snorkel AI Summit 2023).

Chulin Xie (UIUC) chulinx2@illinois.edu Scholar user=WeJnzAgAAAAJ

- **Bio:** Chulin Xie is a fifth-year Ph.D. candidate in Computer Science at the University of Illinois Urbana-Champaign advised by Bo Li. Her research lies in trustworthy machine learning, with a focus on identifying and mitigating robustness and privacy risks in models such as Federated Learning (FL) and Large Language Models (LLMs). She gained industry experience during research internships at Nvidia, Microsoft, and Google, contributing to privacy-preserving techniques for large-scale machine learning systems. She received an outstanding paper award at NeurIPS 2023 Datasets and Benchmark track for benchmarking the trustworthiness of LLMs. She co-organized several workshops including ICML 2023 workshop on Knowledge and Logical Reasoning in the Era of Data-driven Learning, ACL 2022 workshop on Federated Learning for Natural Language Processing, and CVPR 2021 workshop on Adversarial Machine Learning.

Zheng Xu (Google) xuzheng@google.com Scholar user=TfWlMTYAAAAJ

- **Bio**: Zheng Xu is a staff research scientist working on federated learning and privacy at Google. He earned his Ph.D. in optimization and machine learning from University of Maryland, College Park, in 2019. Before that, he got his master's and bachelor's degree from the University of Science and Technology of China. He has published 30+ papers at peer-reviewed research conferences and journals with 13K+ citations, and received two best student paper awards. He is a co-author of *Advances and Open Problems in Federated Learning*, and a lead author of *A Field Guide to Federated Optimization*, both of which resulted from 20+ collaborators in workshop discussions. He is also a co-author of *How to DP-fy ML: A Practical Guide to Machine Learning with Differential Privacy*, and gives the tutorial on the same topic at ICML 2023 and KDD 2023. He is a co-organizer of the Google Federated Learning and Analytics Workshop 2020, TTIC workshop on New Frontiers in Federated Learning 2023, KDD workshop on Federated Learning for Data Mining and Graph Analytics 2024, and the lead organizer of Federated Learning and Analytics in Practice Workshop at ICML 2023, Privacy Regulation and Protection in Machine Learning Workshop at ICLR 2024.

Anticipated audience size

We are expecting 100+ participants. The estimation is based on our previous experience of FL workshop at ICML'23 in Hawaii and PrivML workshop at ICLR'24 in Vienna, and the observation of other workshops at major machine learning conferences.

Plan to get an audience

We will take the following actions:

- create a workshop webpage with all the information;
- advocate call for papers via sharing it with relevant mailing lists in both academia and industry;
- send email announcements to organizer institutions and the broader research community;
- reach out to traditionally underrepresented institutions;
- and share information on social media such as twitter and linkedin, and advertisement on relevant slack workspaces

Diversity commitment

We encourage diversity in both organizers and speakers, which considers gender, affiliation, location, career stages, knowledge and cultural background. Specifically, the seven invited speakers have four men (two confirmed including one who identifies as Black) and three women (all confirmed); five primarily from academia (Stanford, University of Cambridge, Yale and EPFL, MBZUAI and CMU, MIT) and two primarily from industry (Meta, Google). Geographically, two speakers are affiliated with Europe (Cambridge and EPFL), and one speaker is affiliated with Asia (MBZUAI).

The seven organizers have four women and three men; three primarily from academia (Columbia University, CMU, UIUC), three primarily from a big tech and a startup (Google, Gretel), and one primarily from a government institute (UK AI Safety Institute). The organizers are at different career stages including senior PhD students, assistant and associate professors, and senior industry researchers.

We intentionally invited senior speakers, and will balance it with contributed talks from students. We will seek sponsorship from Google, UK AI Safety Institute, Gretel and other companies to potentially provide financial support for students, and we will consider diversity when providing such support.

Access to workshop materials and outcome

All information will be updated on a webpage on a regular basis. Though the workshop is non-archival, we will use OpenReview for double-blind review process, and host the accepted papers. We are confident to secure funding to provide more virtual access, and financial support for attending the workshop.
Already published work at main machine learning venues (ICLR/ICML/NeurIPS) including papers accepted to the ICLR main conference will be explicitly discouraged in call for papers.

History and related workshops

This inaugural workshop is built upon the success of previous workshops on synthetic data, and many related workshops on data access problems and data centric research at major machine learning conferences in recent years.

Synthetic Data Generation with Generative AI NeurIPS'23 is the closest to ours. The NeurIPS'23 workshop highlights the usage of GenAI for synthetic data to empower trustworthy ML training. The senior organizer of the NeurIPS'23 workshop, Mihaela van der Schaar, is a confirmed speaker at our proposed workshop, who can summarize the discussion in NeurIPS'23 and present recent advances in the well grounded literature. The other organizers and speakers do not overlap with the previous edition. We highlight the remarkable progress in synthetic data development and usage in the GenAI era in the past year, and our speakers can speak for the usage in training powerful GenAI models (e.g., Llama models from Meta), and for privacy and safety (Google). More importantly, our proposal highlights the limits and risks of synthetic data for future data access, with speakers with rich experience of open-sourcing data and model (Eric Xing), working on health domains (Mary-Anne Hartley, Mihaela van der Schaar), and studying the phenomena of synthetic data and trustworthiness (Sanmi Koyejo). The organizers' expertise on both synthetic data and data access with privacy, safety and trustworthy considerations will help to organize a program that fosters a discussion across these communities for the future of synthetic data and data access.

In addition, the Synthetic Data for Computer Vision workshop at CVPR'24 focuses on generating visual data, where our invited speaker Phillip Isola is an organizer. Organizer Rachel Cummings previously helped organize the Synthetic Data Generation workshop at ICLR'21. Data access is an important general problem in machine learning. The Data-centric Machine Learning Research workshop at ICML'24 highlights datasets for foundation models. PrivML workshop at ICLR'24 (Zheng Xu is a lead organizer), FL@FM workshop at NeurIPS'23 (Zheng Xu is a keynote speaker) and FL workshop at ICML'23 (Zheng Xu and Peter Kairouz are lead organizers) highlight the privacy considerations and important techniques for accessing high quality data. We can rely on the community on data problems and techniques in machine learning for the success of the workshop.

Call for papers and review

We will highlight submission instructions and timeline on our webpage, with actions to attract participants as mentioned above. We encourage tiny and short papers in our program following the ICLR guidance. .

In addition to the double-blind reviewing as part of our diversity commitment, we will attract a diverse program committee. Each submission will be reviewed by 3 reviewers. Decisions will be made in a transparent way by the organizers. We will encourage presentation of work with novel perspectives and ambitious goals. We will avoid conflict of interests by explicitly asking the reviewers to indicate any. The final list of accepted papers will be published on the workshop website.