Jinjin Zhao

✉ j2zhao@uchicago.edu ☎ +1 312 358 4946 🔗 jinjinz.com j2zhao

About

Research Interests: I have been interested in creating data systems to enable emerging machine learning applications. In particular, my current research seeks to answer: what are the system needs for agents and models to interact with complex external data resources?

Languages: Python, SQL, C

Skills: PyTorch, scikit-learn, Pandas, Airflow, Git, Unix, Amazon Web Services, Google Cloud Platform

Education

PhD **University of Chicago**, Computer Science — Sept. 2019 to 2025 (est.)

- Advisor: Sanjay Krishnan (ChiData Database Group)
- Completed M.S. degree as a part of Ph.D. program

BSE **Princeton University**, Computer Science — Sept. 2015 to May 2019

- Summa Cum Laude
- Minor in Statistics and Machine Learning

Selected Projects

TableVault: Contextual End-to-End Management of LLM Data Pipelines (Website**)** (On-going**)**

- A file-based open-source project to manage data tables and artifacts in complex and dynamic data workflows.
- Extend database transaction techniques, such as two-phase locking and write ahead logs, to enable reverts, restarts, and pauses on all write executions.
- Track data dependencies between tables by enforcing YAML configurations for Python executions.
- Enable governed agentic behavior with executions recursively spawning new data tables.

DSLog: A Compressed Query and Storage Framework for Fine-Grained Array Lineage

- Augment the Numpy library to record cell-to-cell operational lineage efficiently with memory optimizations.
- Introduce a new range-based compression algorithm that improves storage space and query time of the resulting lineage graph by up to 2000x and 20x respectively.

Tracing Variation in Data Science Workflows with Jupyter Notebook Logging

- Develop a tool for Jupyter Notebooks and Python to log execution traces of data science assignments at University of Chicago.
- Analyze the traces to capture user variation trends in data science usage (e.g. most errors are resolved within 1-2 code excutions).
- Validate some common conceptions in data science (e.g. data cleaning takes about 80% of the work).

Experience

Linea Labs, Research Intern — CA, USA
June 2023 to Sept. 2023

- Worked closely within a 6-person startup team out of Berkeley EPIC lab to design and implement an initial MVP for Airflow pipeline reproducibility, with ownership of core lineage-tracking and LLM code-analysis features.

Princeton Plasma Physics Lab, Research Intern — NJ, USA
June 2018 to July 2018

- Compiled data on two years of DIII-D tokamak experiments and trained machine learning models to predict pedestal features driving fusion output.

Meta, Software Engineer Intern WA, USA
June 2017 to Aug. 2017

- Built a full-stack Hack (PHP) and MySQL solution that stored and retrieved test artifacts (e.g. logs and build files) during internal pre-commit automated runs.
- Used to store over 200 million files per week, touching on most internal code development.

Meta, Facebook University Intern CA, USA
June 2016 to Aug. 2016

- Designed and built an independent Android app with an internal music player that generated Spotify playlists based on nearby concerts.

Publications

Fast Capture of Cell-Level Provenance in Numpy 2025

Jinjin Zhao, Sanjay Krishnan

ProvenanceWeek@SIGMOD Paper

TableVault: Managing Dynamic Data Collections for LLM-Augmented Workflows 2025

Jinjin Zhao, Sanjay Krishnan

NOVAS@SIGMOD Paper

Learning Lineage Constraints for Data Science Operations 2025

Jinjin Zhao

arXiv Paper

Quantifying Variation in Data Science Workflows with Fine-Grained Procedural Logging. 2024

Jinjin Zhao, Avidgor Gal, Sanjay Krishnan

Under Submission Paper

Compression and In-Situ Query Processing for Fine-Grained Array Lineage. 2024

Jinjin Zhao, Sanjay Krishnan

ICDE Paper

Data Makes Better Data Scientists. 2023

Jinjin Zhao, Avidgor Gal, Sanjay Krishnan

HILDA@SIGMOD Paper

AMIR: Active Multimodal Interaction Recognition from Video and Network Traffic in Connected Environments. 2023

Shinan Liu, Tarun Mangla, Ted Shaowang, **Jinjin Zhao**, Sanjay Krishnan, Nick Feamster
UbiComp/IMWUT Paper"

Data Station: Delegated, Trustworthy, and Auditable Computation to Enable Data-Sharing Consortia with a Data Escrow. 2022

Siyuan Xia, Zhiru Zhu, Chris Zhu, **Jinjin Zhao**, Kyle Chard, Aaron J. Elmore, Ian Foster, Michael Franklin, Sanjay Krishnan, Raul Castro Fernandez

VLDB Paper

Towards Causal Query Answering for Debugging Video Analytics Systems. 2022

Ted Shaowang*, **Jinjin Zhao***, Stavos Sintos, Sanjay Krishnan

HILDA@SIGMOD Paper

Prediction of DIII-D Pedestal Structure From Externally Controllable Parameters. 2021

Emi Zeger, Florian Laggner, Alessandro Bortolon, Cristina Rea, Orso Meneghini, Samuli Saarelma, Brian Sammuli, Sterling Smith, **Jinjin Zhao**

IEEE Transactions on Plasma Science Paper

Activities And Awards

- **2018 - 2024 Teaching Assistant**: COS 397/497 Fall 2018 *(Princeton University)*, CMSC 16100 Autumn 2019 *(University of Chicago)*, CMSC 21800 Autumn 2020/2023 *(University of Chicago)*, DATA 13600 Spring 2024 *(University of Chicago)*
- ICDE'2024 Travel Award, NSF
- 2022 University Unrestricted Fellowship, University of Chicago
- 2020 - 2022 Curriculum and Social Minister, UChicago CS Graduate Student Ministry
- 2020 Lab Coordinator, CDAC (high school data science research summer program)
- OSDI'2020 Diversity Grant, USENIX Association
- 2019 Neubauer Graduate Scholarship, University of Chicago
- 2016 YHacks 1&1 Prize Winner