Jinjin Zhao

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

Interests

I am interested in systems for data governance, particularly for data science and machine learning. In my research, I have worked on problems in Jupyter Notebook tracking, data lineage and provenance, and machine learning error classification. Currently, I am excited about applying database principles to semi-structured data management and building data systems that enhance model development.

Education

PhD **University of Chicago**, Computer Science — Sept. 2019 to present

- 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

- Graduated Summa Cum Laude
- Minor in Statistics and Machine Learning

Projects

Understanding Datasets for Model Traning in Open Repositories (on-going**)**

- Exploring quantifiable factors to capture patterns in how datasets in open data repositories, like HuggingFace, are used
- Creating a system that links results from secondary artifacts, such as research papers, to initial datasets

Tracing Variation in Data Science Workflows with Jupyter Notebook Logging

- Developed a custom tool for Jupyter Notebooks and Python to log execution traces of 93 data science assignments at the University of Chicago
- Analysed the traces to capture user variation trends in data science usage (e.g., most errors are resolved within 1-2 code executions)
- Validated some common conceptions in data science (e.g., data cleaning takes about 80% of the work)

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

- Fine-grained array lineage is defined as tracking contributions from initial array cells to final array cells after some transformations
- Introduced new compression and query algorithms that improve storage space and query time by up to 2000x and 1500x, respectively.

Improving Triplet Labeling for Image Error Classification with Low-Dimension Features (on-going**)**

- Appling triplet labeling (i.e., given three images, chose the odd-one-out) to the problem of classifying machine learning errors in images
- Combining human feedback (RLHF) with small machine learning models to reduce the human cost of triplet labeling

Experience

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

- Worked closely with a 6-person startup team to design an initial product for Airflow pipeline reproducibility.
- Implemented a core feature that captured data lineage between Airflow tasks.

- Prototyped an internal large language model (LLM) tuning framework for natural language.

Princeton Plasma Physics Lab, Research Intern

NJ, USA
June 2018 to July 2018

- Generated a database from two years of experiments on the DIII-D fusion reactor using the MDSplus interface, focusing specifically on predictors for pedestal features.
- Trained a fully connected neural network architecture to predict parameters that influence fusion production capabilities.

Meta, Software Engineer Intern

WA, USA
June 2017 to Aug. 2017

- Designed an API in PHP/Hack that stores and downloads files (e.g., logs and builds) during code execution.
- Combined the API with a new MySql metadata database that linked to Facebook's internal search framework.
- Used to upload over 200 million log 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 that generated playlists based on nearby concerts.
- Implemented a music player in a separate service environment that linked to the Spotify API.

Publications

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

Jinjin Zhao, Avigdor 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, Avigdor Gal, Sanjay Krishnan

HILDA@SIGMOD Paper

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

Shinan Liu, Tarun Manla, 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

Experimental Based Pedestal Prediction using Machine Learning. 2018

Jinjin Zhao, Egemen Kolemen, Xiaoyan Li, Florian Laggner

APS Division of Plasma Physics Poster

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 C.S. Graduate Student Ministry
- 2020 Lab Coordinator, CDAC (summer research program for high schoolers)
- OSDI'2020 Diversity Grant, USENIX Association
- 2019 Neubauer Graduate Scholarship, University of Chicago
- 2016 YHacks 1&1 Prize Winner

Skills

Languages: Python, SQL, C

Tools: Airflow, Amazon Web Services, Google Cloud Platform