Yinjun Wu, Ph.D

wuyinjun@seas.upenn.edu URL: http://wuyinjun-1993.github.io

Research Interests

Leveraging database techniques to address interpretability, efficiency and debuggability issues in today's data science applications in various high-stake applications such as database systems, health care and self-driving cars.

Education and Training

2021-09 – Present **University of Pennsylvania, PA**
Postdoctoral Researcher in Computer and Information Science
Advisor: **Mayur Naik**

2016-08 – 2021-08 **University of Pennsylvania, PA**
Ph.D. in Computer and Information Science
Advisor: **Susan B. Davidson**
Thesis title: Towards the efficient use of fine-grained provenance in the data science applications
Chair of Dissertation Committee: **Zachary Ives**

2012-08 – 2016-07 **Tsinghua University, Beijing, China**
B.A. in Automation

Industry Research Experience

2020-05 – 2020-08 **NEC Labs American**, Princeton, NJ.
Research Intern

2019-06 – 2019-08 **Microsoft Gray System Lab,** Madison, MI**.**
Research Intern

Teaching Experience

Fall 2017 – Spring 2018 **Teaching Assistant**
CIS 550: Databases and Information System
University of Pennsylvania

Professional Service

2022 **Award coordinator** in *Special Interest Group on Management of Data* (SIGMOD)
2022 **Tutorial session chair** in *Special Interest Group on Management of Data* (SIGMOD)

Reviewing

2024 *International Conference on Learning Representations* (ICLR)

2024 *IEEE International Conference on Data Engineering* (ICDE)

2023 *The Conference on Neural Information Processing Systems* (Neurips)

2022 – 2023 *International Conference on Extending Database Technology* (EDBT)

2022 – 2023 *The International Journal on Very Large Data Bases* (VLDBJ)

2022 *Special Interest Group on Management of Data* (SIGMOD)

2022 *AAAI conference on Artificial Intelligence* (AAAI)

2021 *AAAI conference on Artificial Intelligence* (AAAI)

2021 *International conference on Web Search and Data Mining* (WSDM)

Peer-reviewed conference papers

OOPSLA 2024 **TorchQL: A Programming Framework for Integrity Constraints in Machine Learning**
Aaditya Naik, Adam Stein, **Yinjun Wu**, Mayur Naik, and Eric Wong

ICML 2023 **Do Machine Learning Models Learn Statistical Rules Inferred from Data?**
Aaditya Naik, **Yinjun Wu**, Mayur Naik, and Eric Wong

AAAI 2023 **Learning to Select Pivotal samples for Meta Re-weighting**
Yinjun Wu, Adam Stein, Jacob Gardner, and Mayur Naik,
AAAI conference on Artificial Intelligence (AAAI), 2023

VLDB 2021 **Chef: a cheap and fast pipeline for iteratively cleaning label uncertainties**
Yinjun Wu, James Weimer, and Susan B. Davidson
Proceedings of the VLDB Endowment (VLDB), 2021

AAAI 2021 **Dynamic Gaussian Mixture based Deep Generative Model For Robust Forecasting on Sparse Multivariate Time Series**
Yinjun Wu, Jingchao Ni, Wei Cheng, Bo Zong, Dongjin Song, Zhengzhang Chen, Yanchi Liu, Xuchao Zhang, Haifeng Chen, and Susan Davidson
Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2021

ICML 2020 **Deltagrad: Rapid retraining of machine learning models**
Yinjun Wu, Edgar Dobriban, and Susan Davidson
International Conference on Machine Learning (ICML), 2020

DaMoN 2020 **Lessons learned from the early performance evaluation of Intel Optane DC Persistent Memory in DBMS**
Yinjun Wu, Kwanghyun Park, Rathijit Sen, Brian Kroth, and Jaeyoung Do
Proceedings of the 16th International Workshop on Data Management on New Hardware (DaMoN), 2020

SIGMOD 2020 **PrIU: A provenance-based approach for incrementally updating regression models**
Yinjun Wu, Val Tannen, and Susan B Davidson
Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD), 2020

VLDB 2019 **ProvCite: A Provenance-based Citation System**
Yinjun Wu, Abdussalam Alawini, Daniel Deutch, Tova Milo, and Susan Davidson
Proceedings of the VLDB Endowment (VLDB), 2019

SIGMOD 2018 **Data Citation: Giving Credit where Credit is Due**
Yinjun Wu, Abdussalam Alawini, Susan B Davidson, and Gianmaria Silvello
International Conference on Management of Data (SIGMOD), 2018

VLDB 2017 **Automating data citation in CiteDB**
Abdussalam Alawini, Susan B Davidson, Wei Hu, and **Yinjun Wu**
Proceedings of the VLDB Endowment (VLDB), 2017

Invited talks

2022 **Columbia University**, "Can data provenance assist data-centric AI?"

2021 **Stanford University**, "Toward efficient provenance-aware reasoning in data science applications"

2020 **Northeastern University, "**Provenance-aided approaches to incrementally updating machine learning models"

2020 **Relational AI,** "Provenance-aided approaches to incrementally updating machine learning models"

Awards and Honors

2022 **Morris and Dorothy Rubinoff dissertation Award** in the department of computer and information science at University of Pennsylvania

2014 **Technology Innovation Award** in Tsinghua University for excellent undergraduate research

Research experience

University of Pennsylvania, Department of Computer and Information Science **Philadelphia, PA.**

Postdoctoral researcher **Sept. 2021-Present**

- Collaboratively worked on a project which is on building a programming framework for Pytorch-based machine learning pipeline to facilitate debugging machine learning models in a more scalable and interactive manner; Collaboratively implemented this system and conducted comprehensive experiments over various benchmarks. This work has been submitted to **OOSPLA 2024.**
- Cooperated on a project which is on describing how the distribution shift happens by generating explanations that are more robust and respect inherent structures in data with respect to the state of the art; Collaboratively formalized the algorithm and provided strong empirical evidence to show the benefits of the proposed algorithm. This work has been submitted to **ICLR 2024.**
- Leaded a project on developing a general self-interpretable framework for effectively and accurately explaining how an intervention can influence the output of general machine learning models; Collaboratively implemented the algorithm and empirically demonstrated the effectiveness of the algorithm on various critical domains, in particular, health care. This work has been submitted to **ICLR 2024.**
- Co-leaded a project on learning common sense from data in the format of statistical rules and incorporating the learned common sense into machine learning models to fix model prediction errors during test time in an unsupervised manner; Collaboratively developed the overall algorithms and conducted experiments on multiple datasets across multiple modalities. This work has been published in **ICML 2023.**
- Leaded a project on proposing and developing a labeling-efficient algorithm for label cleaning in the training sample reweighting setting; Independently provided theoretical analysis on the proposed algorithm and collaboratively performed extensive experiments on various benchmarks. This work is published in **AAAI 2023** as one oral presentation.

University of Pennsylvania, Department of Computer and Information Science **Philadelphia, PA.**

Graduate Research Assistant **Aug. 2016-Aug. 2021**

- Developed an interactive, cost-effective and scalable solution, CHEF, to improve the ML model quality by instructing the human annotators to clean the dirty labels of the most influential training samples and simultaneously reducing the time overhead and cost in the overall pipeline; independently providing comprehensive theoretical analysis for this solution and providing an efficient implementation. This work has been published in **VLDB 2021** as a first-author paper.
- Developed an efficient algorithm, DeltaGrad, for rapidly forcing machine learning models to incrementally forget sensitive training samples, which can benefit multiple emerging data science applications including data cleaning and debugging in today's machine learning pipeline; collaboratively provided rigorous theoretical analysis on the correctness of DeltaGrad and independently provided an efficient implementation of this algorithm. This work has been published in **ICML 2020** as a first-author paper.
- Proposed and implemented a provenance-based solution to incrementally update regression models after sensitive training samples are removed. Independently provided a theoretical guarantee on the correctness of the proposed algorithm. This work has been published as a first-author paper in **SIGMOD 2020**.
- Collaboratively developed a provenance-based framework to automatically track the contributions of data curators to open-sourced scientific databases and distribute proper credits to them; independently implemented variants of query optimization techniques to reduce the computational overhead, which can achieve up to an order of magnitude speed-up. This work has ended up with a series of papers, including first-author papers in **VLDB 2019** and **SIGMOD 2018.**

NEC Labs American **Princeton, NJ.**

Research Intern **May. 2020 - Aug. 2020**

- Designed a deep generative model for dynamically modeling the latent clustering structures in time series data with a substantial number of missing values; Conducted rigorous derivations to provide theoretical guarantees on the training process of the proposed model; Independently implemented the proposed model and demonstrated the effectiveness in modeling highly sparse medical time series in comparison to the state of the art. This work has been published in **AAAI 2021**

Microsoft Gray System Lab **Madison, MI.**

Research Intern **Jun. 2019 - Aug. 2019**

- Extensively measured the performance characteristics over a new device called Intel Optane DC persistent memory in various hardware and system configurations by running various microbenchmark workloads; Developed a prototype to demonstrate the potential of this new device in some well-known in-disk data structures to show its I/O performance gains; Experimentally verified the performance benefits of this device in SQL server. This work has been published in **DaMoN 2020.**