Yinjun Wu

wuyinjun@seas.upenn.edu URL: http://www.cis.upenn.edu/~wuyinjun
Room 565, 3330 Walnut Street, Philadelphia, Pennsylvania, U.S.

Research Interests

Data science and machine learning problems in the context of today's emerging data-intensive applications, including database systems and medical analysis etc.

Education

Post-Doc. Computer and Information Science, University of Pennsylvania, PA. **2021-2022**
Advisor: Prof. Mayur Naik
Ph.D. Computer and Information Science, University of Pennsylvania, PA. **2016-2021**
Advisor: Prof. Susan B. Davidson
Thesis title: Towards the efficient use of fine-grained provenance in the data science applications
B.A. Automation, Tsinghua University, Beijing, China **2012-2016**
Thesis title: Indexing Technologies in Question-Answering Systems
Relevant courses: Introduction to Artificial Intelligence, Pattern Recognition, Operations Reserach, Stochastic Processes, Probability theory, Introduction to Databases, Data Structures and Algorithms, etc.

Research Experience

University of Pennsylvania, Department of Computer and Information Science **Philadelphia, PA.**
Post-Doc researcher **Sept. 2021-2022**

- Leaded a project on iteratively cleaning the most representative samples to boost the performance of meta-reweighting algorithm; collaboratively worked with other PhD students on the implementation of the system and submitted one paper to Neurips 2022

University of Pennsylvania, Department of Computer and Information Science **Philadelphia, PA.**
Graduate Research Assistant **Aug. 2016-Aug. 2021**

- Developed a research project on "data citation", which automatically tracks the influence of data in various data science applications at various granularities.
- Independently implemented core functions of a data citation system called CiteDB which tracks the influence of data throughout SQL queries using JAVA; collaboratively built an interactive user interface. CiteDB was accepted as a demo paper in **VLDB 2017**.
- Worked with external collaborators on a proof of concept for CiteDB using the IUPHAR Pharmacological Dataset; collaboratively implemented algorithms for CiteDB using JAVA; developed various optimization strategies which achieved orders of magnitude speed-up; conducted extensive experiments using Linux bash scripts and Matlab scripts. This leads to a first-author paper in **SIGMOD 2018** and an invited paper to the IEEE TCDE journal.
- Collaboratively developed a provenance-based citation model; independently implemented an extended version of CiteDB which can handle more general queries, in particular, those involving aggregate and user-defined functions in JAVA; applied clustering algorithms to speed up computations (achieving an order of magnitude speed-up). This work has been accepted as a first-author paper to **VLDB 2019.**
- Developed a provenance-based framework by leveraging properties from the optimization theory and linear algebra; independently implemented algorithms with various optimization strategies by utilizing Pytorch library with provable approximation bound, which provide the power of efficiently updating the regression models (linear regression and logistic regression models) in the case of partial removal over training set and thus can assist ML model debugging, data cleaning, dealing with GPDR laws and other related ML applications. This work has been published as a first-author paper in **SIGMOD 2020**.
- Developed a solution, DeltaGrad, by utilizing well-studied L-BFGS algorithm from the ML domain for the purpose of incrementally updating the general ML models satisfying strong convexity; collaboratively working with some ML experts on providing rigorous theoretical analysis on the correctness of DeltaGrad; independently providing an efficient implementation of this algorithm. This work has been published in **ICML 2020** as a first-author paper.

- Developed an active-learning-like solution, Chef, to improve the ML mode 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 submitted to **VLDB 2021** as a first-author paper.

University of Pennsylvania, Department of Computer and Information science **Philadelphia, PA.**
Teaching Assistant **Sep. 2017 – May. 2018**

- Developed homework questions using Python scripts, the Neo4j database engine, the MongoDB engine and the Tensorflow framework.

Institute of Future Information and Technology, Tsinghua University **Beijing, China**
Undergraduate Research Assistant **Apr. 2014-Jun. 2016**

- Conducted research on developing efficient bitmap indexing technologies and applying those indexing techniques in Question-Answering systems
- Independently designed a probability-based model to analyze the performance of bitmap indexing algorithms from a theoretical perspective, and conducted extensive simulations using Matlab. This leads to a first-author publication in ICCCN 2015.
- Collaboratively designed and implemented a new bitmap indexing algorithm, which achieves 2-3x speed-up compared to previous algorithms. This was published in IEEE Communications Letters as a first-author paper.
- Independently developed a QA prototype system by integrating an existing speech-recognition system and knowledgebase system; improved the search performance in this prototype by implementing our own indexing algorithms; and built up a simplified user interface with Javascript. This work was my undergraduate thesis.

Experience

NEC Labs American **Princeton, NJ.**
Research Intern **May. 2020 - Aug. 2020**

- Designed a deep generative model for dynamically modeling the latent clustering structures in sparse health care time series data
- Conducted rigorous derivations to provide theoretical guarantees on the training process of the proposed model
- Independently implemented the proposed model with Pytorch and demonstrated the effectiveness in forecasting the future values of the health care time series data by comparing it against the state-of-the-art work, which ends up with an accepted paper 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 the Intel Optane DC persistent memory in SQL server by running OLTP/PLAP workloads, which ends up a first-author paper published in **DaMoN 2020**

Professional Activities

- **Reviewer**: WSDM 2021, AAAI 2021, AISTATS 2021, AAAI 2022, EDBT 2022
- **External reviewer**: SIGMOD 2022
- **Conference service:** Awards Coordinator in SIGMOD 2022

Selected Publications

[1] **Y. Wu**, A. Stein, J. R. Gardner, M. Naik "Learning to Select Pivotal samples for Meta Re-weighting", submitted to Neurips 2022

[2] **Y. Wu**, J. Weimer, S. B. Davidson, "Chef: a cheap and fast pipeline for iteratively cleaning label uncertainties", VLDB 2021.

[3] **Y. Wu**, J. Ni, W. Cheng, B. Zhong, D. Song, Z. Chen, Y. Liu, X. Zhang, H. Chen, S. B. Davidson, "Dynamic Gaussian Mixture based Deep Generative Model For Robust Forecasting on Sparse Multivariate Time Series", AAAI 2021.

[4] **Y. Wu**, K. Park, R. Sen, B. Knoth, J. Do, "Lessons learned from the early performance evaluation of Intel Optane DC Persistent Memory in DBMS", In Proceedings of the 16th International Workshop on Data Management on New Hardware (DaMoN 2020) pp. 1-3.

[5] **Y. Wu**, E. Dobriban, S.B. Davidson, "DeltaGrad: Rapid retraining of machine learning models", accepted by ICML 2020.

[6] **Y. Wu**, V. Tannen, S.B. Davidson, "PrIU: A provenance-based approach for incrementally updating regression models" In Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data (SIGMOD 2020) pp. 447-462.

[7] **Y. Wu**, A. Alawini, D. Detuch, T. Milo, S. B. Davidson, "ProvCite: A Provenance-based Citation System", Proceedings of the VLDB Endowment 12 (7), 738-751.

[8] **Y. Wu**, A. Alawini, S. B. Davidson, G. Silvello, "Data Citation: Giving Credit where Credit is Due", Proceedings of the 2018 International Conference on Management of Data (SIGMOD 2018), pp. 99-114. ACM, 2018.

[9] A. Alawini, S. B. Davidson, W. Hu, **Y. Wu**, "Automating data citation in CiteDB", Proceedings of the VLDB Endowment 10 (12), 1881-1884 (authors sorted alphabetically)

Skills

Java, Python, C, C++, Linux bash script, SQL, Pytorch, Tensorflow, Neo4j, MongoDB, Matlab

Conference Presentations

[1] "DeltaGrad: Rapid retraining of machine learning models" Oral presentation at ICML 2020

[2] "PrIU: A provenance-based approach for incrementally updating regression models" Oral presentation at SIGMOD 2020

[3] "Lessons learned from the early performance evaluation of Intel Optane DC Persistent Memory in DBMS" Oral presentation at DaMoN 2020

[4] "ProvCite: A Provenance-based Citation System" Oral presentation at VLDB 2019

[5] "Data Citation: Giving Credit where Credit is Due" Oral presentation at SIGMOD 2018.

[6] "Automating data citation in CiteDB" Demonstration at VLDB 2017.

Awards and Honors

- Co-winner of Morris and Dorothy Rubinoff dissertation Award in the department of computer and information science at University of Pennsylvania (Top 2), 2022
- Scholarship in Tsinghua University for excellent academic performance (1/136), 2013-2015
- First prize in physics contest of college students in Beijing (Top 5%), 2014
- Technology Innovation Award in Tsinghua University for excellent undergraduate research, 2014