FANGZHU SHEN

fangzhu.shen@duke.edu https://fangzhushen.github.io
Durham, North Carolina, USA

RESEARCH INTERESTS

I am broadly interested in database management, data analysis and LLM agents, focusing on the end-to-end data lifecycle. I have been woroking on problems in data cleaning, influence attribution in social networks, and selectivity estimation in online learning setting. Currently, I am exploring how LLM agents are transforming data management, particularly in data cleaning and query verification with formal verification methods.

EDUCATION

Duke University Durham, NC, USA
Ph.D. in Computer Science *January 2023 - Present*

- Advisor: Prof. Sudeepa Roy
- GPA: 3.98/4.0
- Relevant Coursework: Natural Language Processing, Causal Inference in Data Analysis - Fairness & Explanations, Data Science, Design & Analysis Algorithms

Duke University Durham, NC, USA
M.S. in Economics and Computation *August 2021 - December 2022*

- Transitioned to Ph.D. program
- Relevant Coursework: Database Systems, Algorithms, Theory & Algorithm Machine Learning, Advanced Computer Networks, Computational Microeconomics

Central University of Finance and Economics Beijing, China
B.A. in Finance *September 2016 - July 2020*

PUBLICATIONS

(* = equal contribution)

- **The Cost of Representation by Subset Repairs** [pdf] (Proceedings of the VLDB Endowment (PVLDB) 2024, Vol. 18, No. 2)
  - Yuxi Liu*, **Fangzhu Shen***, Kushagra Ghosh, Amir Gilad, Benny Kimelfeld, and Sudeepa Roy.
- **Causal What-If and How-To Analysis Using HypeR** [pdf] (ICDE 2023, Demonstration Paper)
  - **Fangzhu Shen***, Kayvon Heravi*, Oscar Gomez, Sainyam Galhotra, Amir Gilad, Sudeepa Roy, Babak Salimi.

RESEARCH EXPERIENCES

Data Cleaning

- My work *"The Cost of Representation by Subset Repairs"* (PVLDB 2025) addressed the challenge of preserving data distribution while minimizing repair costs. (1) We introduced Representative Subset Repair to maintain proportional representation of sensitive sub-populations while achieving data consistency, which is critical for unbiased analytics and downstream ML tasks. (2) We studied the complexity and devised dynamic programming-based algorithms for special cases and efficient optimization algorithms for general cases using Linear Programming. (3) We evaluated our approaches experimentally and showed their effectiveness in large-scale real-world datasets.

Causal Inference

- I have built and presented a demonstration titled *'Causal What-If and How-To Analysis Using HypeR'* (ICDE 2023 Demonstration) created a SQL-like interface integrating causal inference for answering complex hypothetical queries. We delivered interactive visual modules to guide interpretation of what-if and how-to queries.

INDUSTRY EXPERIENCES

PhD Software Engineering Intern, Uber *June 2025 - August 2025*
Coordinated Structural Pricing Team *New York, NY, USA*
Project: Dynamic Structural Estimation for Pricing Model: An End-to-End Learning Framework

- To address the challenge of forecasting hidden marketplace parameters, developed a dynamic learning framework that integrates deep learning with economic-driven structural pricing model.
- Designed a novel residual-meta learning architecture to handle high variance and multi-target predictions, compared multiple ML models (MLP, LSTM, Transformer), resulting in a significant improvement in accuracy (11% reduction in overall loss compared to baseline).
- Engineered a differentiable pricing model by re-implementing and approximating the equilibrium solver in PyTorch, enabling its integration into end-to-end training framework.
- Conducted large-scale experiments across 100+ cities and built a reproducible ETL and training pipeline.

Software Engineering Intern, Carl Zeiss Meditec *June 2022 - August 2022*
AI Solution & Service Team *Raleigh, NC, USA*

- Implemented and evaluated multiple ML algorithms including decision trees, Bayesian ridge regression, and neural networks with Scikit-learn for file conversion time prediction using Python and SQL.
- Designed and built feature engineering pipelines to process and analyze data from Azure Cloud SQL databases
- Deployed the prediction model as a real-time service using FastAPI, enabling seamless integration with production systems.

TEACHING EXPERIENCE

- **Teaching Assistant**, Introduction to Database Systems, Duke University *Fall 2025*
- **Teaching Assistant**, Causal Inference, Fairness, and Explanations in Data Analysis *Spring 2025*
- **Teaching Assistant**, Introduction to Databases, Duke University *Fall 2024*
- **Teaching Assistant**, Introduction to Database Systems, Duke University *Fall 2023*

AWARDS

- VLDB 2025 Travel Award
- ICDE 2023 Travel Award
- Duke Scholar Award (25% Tuition Waiver) *August 2021 - December 2022*
- Academic Scholarships of Central University of Finance and Economics *2017, 2018, 2019*

SERVICE

- Shadow PC member: International Conference on Very Large Databases (VLDB) 2026

SKILLS

- Programming: Python, SQL, C, Rust
- Database Systems: PostgreSQL, MySQL, MongoDB
- Development: PyTorch, Scikit-learn, Hugging Face, Docker, Git, Linux