Weiyuan Wu

youngw@sfu.ca · wooya.me · github.com/dovahcrow

Education

Ph.D. Candidate in Database, Simon Fraser University, Canada Sep. 2019 -
Supervisor: Dr. Jiannan Wang

M.Sc. in Database, Simon Fraser University, Canada May 2017 - Sep. 2019
Thesis: "Enabling SQL-ML Explanation to Debug Training Data"
Supervisor: Dr. Jiannan Wang

B.S. in Computing Science, UESTC, China Sep. 2012 - July 2016

Research Interests

SQL & Machine Learning Debugging, Data Intensive System

Research and Work Experience

Research Assistant Sept. 2017 -
Data Debugging for Machine Learning Pipelines (Supervisor: Dr. Jiannan Wang) *Simon Fraser University*

- Researched related work on SQL debugging, Machine Learning debugging and Federated Learning.
- Conducted experiments using Tensorflow, Python and Scikit-learn.
- Wrote research papers and got them published in top conferences.
- Published papers: Complaint-driven Training Data Debugging for Query 2.0 in SIGMOD 2020, Enabling SQL-based Training Data Debugging for Federated Learning in VLDB 2022, Complaint-Driven Training Data Debugging at Interactive Speeds in SIGMOD 2022.

Project Leader May 2019 -
Data Preparation in Python *dataprep.ai*

- Designed and implemented the core system, including a module for EDA, a module for data collection and a module for data cleaning.
- Managed a team with 20+ members. Established the team processes for communication, code committing, code review, issue triage and release.
- Achieved ~300k downloads and over 1k Github stars within the past two years.
- Published paper DataPrep.EDA: Task-Centric Exploratory Data Analysis for Statistical Modeling in Python in SIGMOD 2021

Project Leader Jan 2021 -
The Fastest Library to Load Data from DB to DataFrames in Rust and Python *github.com/sfu-db/connector-x*

- Conducted related work investigation and performed extensive evaluations.
- Implemented the core pipeline of the system. It accelerates data loading by 13x and reduces the memory footprint by 3x compared to Pandas, the most popular data tool.
- Designed the DSL for easily extending the library. The DSL allows ConnectorX to support 7+ mainstream databases and 4 most widely used dataframes.
- Submitted paper ConnectorX: Accelerating Data Loading From Databases to Dataframes in VLDB 2022

Tech Advisor Mar 2021 -
Pan-European Digital Derivatives Exchange *D2X Group*

- Researched different IPC methods, storage and recovery solutions based on the latency and reliability requirements.
- Designed the system architecture from zero to one, including matching engine, order entry gateway and risk engine.

Database Engineer Intern Sept. 2021 - Dec. 2021
Memory Optimized Distributed Database (Supervisor: Qingqing Zhou) *Tencent*

- Piloted the application of the userpagefault in database page management.
- Addressed the out-of-memory issue by integrating the userpagefault page management using C++.
- Implemented the coroutine support for the page management component.

External Researcher Jan 2018 - Sep. 2019
Vancity

- Performed data augmentation on company's customer data using entity resolution with open data from the web.
- Built ensemble tree-based model for churn prediction.
- Applied sentiment analysis on the customer feedback to continuously monitor the company's performance.

Data Scientist June 2016 - May 2017
Strikingly Inc.

- Built XGBoost based model for churn prediction.
- Built a rule and Machine Learning mixed model for detecting spammer contents.
- Built a data warehouse with ETL pipeline from scratch using Postgres, Amazon Redshift, lambda functions.
- Improved the responsiveness of the analytics dashboard for customers from 3 minutes to 2 seconds by implementing a data cube-based cache layer.

Publications

Xiaoying Wang*, **Weiyuan Wu***, Jinze Wu, Yizhou Chen, Nick Zrymiak, Changbo Qu, Lampros Flokas, George Chow, Jiannan Wang, Tianzheng Wang, Eugene Wu, Qingqing Zhou:
ConnectorX: Accelerating Data Loading From Databases to Dataframes VLDB 2022, Under Review

Lampros Flokas, **Weiyuan Wu**, Yejia Liu, Jiannan Wang, Nakul Verma, Eugene Wu:
Complaint-Driven Training Data Debugging at Interactive Speeds SIGMOD 2022

Yejia Liu, **Weiyuan Wu***, *Lampros Flokas, Jiannan Wang, Eugene Wu:*
Enabling SQL-based Training Data Debugging for Federated Learning VLDB 2022

Brandon Lockhart, Jinglin Peng, ***Weiyuan Wu****, Jiannan Wang, Eugene Wu:*
Explaining Inference Queries with Bayesian Optimization VLDB 2021

Jinglin Peng,* ***Weiyuan Wu****, Brandon Lockhart, Song Bian, Jing Nathan Yan, Linghao Xu, Zhixuan Chi, Jeffrey Rzeszotarski, Jiannan Wang:*
**DataPrep.EDA: Task-Centric Exploratory Data Analysis
for Statistical Modeling in Python** SIGMOD 2021

Xiaoying Wang, Changbo Qu*,* ***Weiyuan Wu****, Jiannan Wang, Qingqing Zhou:*
Are We Ready For Learned Cardinality Estimation? VLDB 2021

Weiyuan Wu*, Lampros Flokas, Eugene Wu, Jiannan Wang:*
Complaint-driven Training Data Debugging for Query 2.0 SIGMOD 2020

Weiyuan Wu*, Lampros Flokas, Eugene Wu, Jiannan Wang:*
Towards Complaint-driven ML Workflow Debugging MLOps 2020, Demo

Skills

Frameworks: Tensorflow, Pandas, Numpy, Scikit-Learn, Dask

Programming Languages: Rust (7y), Python/Cython (4y), C++, Typescript, SQL, Terraform

Platforms: Docker, Kubernetes, AWS, Solana