

Dong He

donghe@cs.washington.edu | dongheuw.github.io | linkedin.com/in/dongheuw | +1 (206) 295-6340

Education

University of Washington

PhD in Computer Science, advised by [Prof. Magdalena Balazinska](#)

Sep 2019 – Aug 2024 (expected)

Seattle, WA

Fudan University

BSc in Computer Science (Honors), advised by [Prof. Yang Chen](#)

Sep 2015 – Jul 2019

Shanghai

- Graduated with Highest Distinction (6 out of 118)

Work Experience

Snowflake

PhD Intern, SQL Team

Oct 2023 – present, remote

- Research collaboration between UW and Snowflake on query workload analysis and optimization using LLMs.

Software Engineer Intern, Machine Learning Platform Team

Jun 2023 – Sep 2023, Bellevue, WA

- **Created Snowflake's Distributed Model Trainer (SLT)**: Led the design and implementation (in Python) of the first distributed PyTorch training solution within Snowflake. My work automates the model training process and creates a user-friendly interface that hides underlying complexities, enabling users to train deep models with their data inside Snowflake effortlessly.
- **Multi-Modal Support**: MLP, ResNet, and LLM training support for tabular, image, and text data in Snowflake with SLT.

Microsoft

Jun 2021 – Sep 2021

Research Intern, [Microsoft Jim Gray Systems Lab](#) led by [Prof. Raghuram Ramakrishnan](#)

Remote

- **Created Tensor Query Processor (TQP)**: Led the design and implementation (in Python & C++) of the industry's first query processor that compiles SQL queries into PyTorch programs and executes them on various hardware (CPUs, GPUs, TPUs, etc.).
- **Full TPC-H Support & Enhanced Performance**: Enabled full TPC-H benchmark support with TQP (until 2021 no GPU database was able to support full TPC-H benchmark), improving query execution time by 10x over specialized CPU and GPU systems (DuckDB, HeavyDB, BlazingSQL) and providing acceleration for a 9x speedup for queries containing ML inference.
- **Impact**: First-authored a VLDB paper on TQP and won the **Best Demo Award** at VLDB 2022 (top academic conference in databases). TQP was also **featured by Azure CTO** at Microsoft Build 2023.

Goldman Sachs

Jul 2018 – Sep 2018

Summer Analyst, Engineering

Hong Kong

- **Global Engineering Challenge Champion**: Clinched the Global Winner title in the Intern Engineering Challenge.
- **Revamped PnL Reconciliation Process**: Redesigned and re-implemented the logic (in Java) for the true-up job which reconciles the estimated vs. actual profit and loss. Deployed enhancements led to a 50% reduction in memory usage, significantly minimizing the risk of job failures.

Tencent

Jan 2018 – Feb 2018

Research Intern, YouTu X-Lab led by [Prof. Jiaya Jia](#) and [Prof. Yu-Wing Tai](#)

Shenzhen

- **Optimized Neural Network Inference Efficiency**: Analyzed node liveness and dependencies (in C++) in production-level neural networks, achieving up to 30% reduction in memory consumption through memory sharing.
- **Enhanced Data Collection and Annotation Process**: Created tools (in Python) for gathering and annotating large-scale image data, streamlining the training process for image classification models.

Selected Publications

MaskSearch: Querying Image Masks at Scale. [Preprint][Website][Code] **D. He**, J. Zhang, M. Daum, A. Ratner, M. Balazinska.

Query Processing on Tensor Computation Runtimes. [Paper][Talk][Keynote by Microsoft Azure CTO][MarkTechPost][SyncedReview] **D. He**, S. Nakandala, D. Banda, R. Sen, K. Saur, K. Park, C. Curino, J. Camacho-Rodríguez, K. Karanasos, M. Interlandi. VLDB 2022.

Share the Tensor Tea: How Databases can Leverage the Machine Learning Ecosystem. [Paper] Y. Asada*, V. Fu*, A. Gandhi*, A. Gemawat*, L. Zhang*, **D. He**, V. Gupta, E. Nosakhare, D. Banda, R. Sen, M. Interlandi. VLDB 2022. **Best Demo Award** (Demonstration of *Query Processing on Tensor Computation Runtimes*).

DeepEverest: Accelerating Declarative Top-K Queries for Deep Neural Network Interpretation. [Paper][Website][Code][Talk] **D. He**, M. Daum, W. Cai, M. Balazinska. VLDB 2022.

VOCAL: Video Organization and Interactive Compositional AnaLytics. [Paper][Website][Talk] M. Daum*, E. Zhang*, D. He, M. Balazinska, B. Haynes, R. Krishna, A. Craig, A. Wirsing. CIDR 2022.

VSS: A Storage System for Video Analytics. [Paper][Code][Talk] B. Haynes, M. Daum, D. He, A. Mazumdar, M. Balazinska, A. Cheung, L. Ceze. SIGMOD 2021.

TASM: A Tile-Based Storage Manager for Video Analytics. [Paper][Code][Talk] M. Daum, B. Haynes, D. He, A. Mazumdar, M. Balazinska. ICDE 2021.

Accelerating Mobile Applications at the Network Edge with Software-Programmable FPGAs. [Paper] S. Jiang, D. He, C. Yang, C. Xu, G. Luo, Y. Chen, Y. Liu, J. Jiang. INFOCOM 2018.

For a complete list of my publications, please visit [my website](#) and [my Google Scholar profile](#).

Selected Projects

Tensor Query Processing

May 2021 – Jun 2022

Project owner & lead

Microsoft, University of Washington

- **Created Tensor Query Processor (TQP):** Led the design and implementation (in Python & C++) of the world's first query processor, TQP, that compiles SQL queries into tensor programs and executes them on various hardware (CPUs, GPUs, TPUs).
- **Full TPC-H Support & Hardware Adaptability:** Enabled TQP to support the full TPC-H benchmark on various hardware with reduced development effort, demonstrating the tensor abstraction's capability to relational SQL queries.
- **Significant Speedups:** Improved execution time for TPC-H queries by 10x over specialized CPU and GPU systems such as DuckDB and HeavyDB, and achieved a 9x speedup over CPU baselines when ML inference is used with SQL queries.
- **Impact:** The demo of TQP won the **Best Demo Award** at VLDB 2022. TQP was also **featured by Azure CTO** at Microsoft Build 2023.

Data Management Systems for Model Explanation and Exploration [Website]

Oct 2019 – present

Project owner & lead

University of Washington

• MaskSearch: Querying Image Masks at Scale

- **Created MaskSearch:** Led the design and implementation (in Python) of a system, MaskSearch, that accelerates image retrieval queries based on properties of mask annotations of the images by designing a novel indexing technique and an efficient filter-verification query execution framework. These queries are essential for a variety of applications such as identifying spurious correlations learned by ML models and detecting maliciously manipulated images.
- **Achieved Excellent Empirical Performance:** MaskSearch accelerates individual queries by up to 100x, using indexes only 5% the size of the original data, and consistently outperforms existing methods in various multi-query workloads.
- **DeepEverest: Accelerating Interpretation by Example Queries [Website]**
 - **Created DeepEverest:** Designed and implemented a system (in C++ & Python), DeepEverest, for efficiently executing *interpretation by example* queries that identify examples based on activation patterns in neural networks, by designing an efficient indexing technique and a theoretically-proven instance-optimal query execution algorithm with empirical optimizations.
 - **Optimized Storage and Query Performance:** DeepEverest accelerates individual *interpretation by example* queries by up to 63x while reducing storage to less than 20% of full materialization and consistently outperforms baselines in various multi-query workloads that simulate model explanation processes.

VisualWorld Video Data Management Systems [Website]

Apr 2020 – Sep 2023

Project contributor

University of Washington

- **VOCAL:** a set of video data management systems that support efficient data cleaning, exploration, and organization for large-scale video data, as well as processing complex compositional queries, even when no pretrained model exists.
- **TASM:** a video storage manager which enables spatial random access to encoded videos. TASM speeds up content retrieval queries by up to 94% and improves the throughput of the full scan phase of object detection queries by up to 2x.
- **VSS:** a system that arranges video data on disk in an efficient, granular format, caches frequently retrieved regions in the most useful formats, and eliminates redundancies found in videos captured from multiple cameras with overlapping fields of view.

a video storage system that decouples video operations such as machine learning from the low-level details required to store and efficiently retrieve video data. VSS improves read performance by up to 54% and reduces storage costs by up to 45%.

FPGA-Based Edge Computing for Accelerating Mobile Applications

Jul 2017 – Aug 2017

Project contributor

Peking University

- **Developed FPGA-Based Edge Computing Model:** Engineered a prototype (in C++ & Python) that minimizes response time and energy consumption for interactive mobile applications by offloading computation to an FPGA-based edge.

- **Performance Improvements:** Achieved up to 3x/15x faster response times over CPU-based edge/cloud offloading and enhanced energy efficiency by up to 29.5%.

Selected Awards

- [Best Demo Award](#), 48th International Conference on Very Large Databases (VLDB) 2022
- [Madrona Prize](#), for the most commercializable research project, Madrona Venture & UW [[GeekWire](#)] [[BusinessWire](#)] 2022
- Paul G. Allen Fellowship, University of Washington 2019 – 2020
- Wangdao Scholar, Undergraduate Research Opportunities Program, Fudan University 2018
- Silver Medal, ACM International Collegiate Programming Contest (ACM-ICPC), Asia Regional 2015 – 2016
- Silver Medal (Top 100 in China), National Olympiad in Informatics (NOI), China National Finals 2014
- First Prizes, National Olympiad in Informatics in Provinces (NOIP) 2009 – 2014

Invited Talks & Presentations

- Snowflake, Query Processing on PyTorch Jul 2023
- UW Madison, Data Management for Model Explanation and Exploration Apr 2023
- Huawei Cloud, Query Processing on Tensor Computation Runtimes Feb 2023
- UW CSE Affiliates Day, Data Management for Model Exploration and Debugging Nov 2022
- VLDB 2022, Accelerating Declarative Top-K Queries for Deep Neural Network Interpretation [[Video](#)] Sep 2022
- VLDB 2022, Query Processing on Tensor Computation Runtimes [[Video](#)] Sep 2022
- VLDB 2022, How Databases can Leverage the Machine Learning Ecosystem Sep 2022
- RelationalAI, Query Processing on Tensor Computation Runtimes Jun 2022
- Microsoft Jim Gray Systems Lab, Query Processing on Tensor Computation Runtimes Sep 2021

Teaching & Service

- Teaching Assistant, UW CSEP 590A: Machine Learning for Big Data Spring 2022
- Head Teaching Assistant, UW CSED 516: Scalable Data Systems and Algorithms Fall 2021
- Student Volunteer, VLDB 2020 Sep 2020

Mentoring Experience

- Master / Undergraduate Students: Jason Li (2022-2023), Mona Gandhi (2022–2023), Tim Li (2022).
- Highschool Students: Parie Kumar (2022).

Professional Skills

- **Programming Languages:** C/C++, Python, Java, Pascal, Javascript, Matlab, ...
- **Machine Learning:** PyTorch, TensorFlow, PyTorch Lightning, PEFT, Keras, Scikit-Learn, ...
- **Technical:** Database Systems (Query Optimization & Execution, Indexing Techniques), Algorithms & Data Structures, Machine Learning Systems (Software & Hardware), Video Analytics, Computer Vision, Natural Language Processing, ...
- **Other Tools:** MySQL, PostgreSQL, DuckDB, Spark, Hadoop, Hive, Google Cloud, AWS, Snowflake, Databricks, Docker, Selenium, LaTeX, Git, SVN, ...