JUNJIE XING

☎ 734-834-8572 ✉ junjiexing@microsoft.com

Education

University of Michigan, Ann Arbor **Sep. 2020 – May. 2025**
Ph.D. in Computer Science and Engineering *Ann Arbor, USA*
Advisor: Prof. H. V. Jagadish

University of Michigan, Ann Arbor **Sep. 2018 – Apr. 2020**
M.S.E. in Computer Science and Engineering *Ann Arbor, USA*

Shanghai Jiao Tong University **Sep. 2014 – Jun. 2018**
B.Eng in Information Security *Shanghai, China*

Work Experience

Microsoft Research **May 2025 – Present**
Senior Researcher *Manager: Surajit Chaudhuri*

- Conduct research in data integration, data exploration, and data management systems, focusing on making large-scale datasets / database more accessible and useful.

Publication

MMTU: A Massive Multi-Task Table Understanding and Reasoning Benchmark
Junjie Xing, Yeye He, Mengyu Zhou, Haoyu Dong, Shi Han, Lingjiao Chen, Dongmei Zhang, Surajit Chaudhuri, H. V. Jagadish
Accepted by NeurIPS 2025.

Table-LLM-Specialist: Self-trained Language Model Specialist for Tables using Iterative Generator-Validator Fine-tuning
Junjie Xing, Yeye He, Mengyu Zhou, Haoyu Dong, Shi Han, Dongmei Zhang, Surajit Chaudhuri
Accepted by EMNLP 2025.

OpenForge: Probabilistic Metadata Integration
Tianji Cong, Fatemeh Nargesian, Junjie Xing, H. V. Jagadish
In Proceedings of PVLDB Vol.18, 2025.

Data-Driven Insight Synthesis for Multi-Dimensional Data
Junjie Xing, Xinyu Wang, H. V. Jagadish
In Proceedings of PVLDB Vol.17, 2024. Nomination for ***Best Research Paper Award****.*

ARTS: A System for Aggregate Related Table Search
Junjie Xing, H. V. Jagadish
In Proceedings of ICDE Demo 2024.

Research Projects

Finding Aggregate Related Tables **Aug. 2022 – Jul. 2023**
Database Group, CSE department, University of Michigan *Advisor: Prof. H. V. Jagadish*

- Defined a new table relatedness measure, aggregate relatedness, which requires a holistic understanding of column semantics on both textual and numerical columns.
- Proposed a novel column semantics understanding technique, with pre-trained large language models (LLMs).
- Annotated a new benchmark for the task, over a large table repository built with Data.Gov.
- Implemented a system that can effectively and efficiently identify aggregate related tables, and evaluated on the annotated benchmark.

Data-Driven Insight Synthesis for Multi-Dimensional Data **Sep. 2020 – Jun. 2022**
Database Group, CSE department, University of Michigan *Advisor: Prof. H. V. Jagadish, Prof. Xinyu Wang*

- Proposed to learn a data-driven interestingness measure from user annotation.
- Created an annotation algorithm that integrates clustering and QuickSort and reduced the annotation cost heavily.
- Developed a multi-round annotation system that interacts with Amazon Mechanical Turk, and automatically generates and posts human intelligent tasks for each round.
- Developed an efficient insights synthesis algorithm using Markov Chain Monte Carlo.
- Implemented all the ideas in a system and evaluated it on real-word datasets.

Cross-disciplinary Projects

Social Media Archive **June. 2022 – Dec. 2022**

Inter-university Consortium for Political and Social Research, University of Michigan *PI: Prof. Libby Hemphill*

- Goal: to build a social media archive that stores social media data from facebook, twitter, reddit, .etc, and provides easy access to social scientists with little SQL/coding background.
- Designed database schema for twitter data on both traditional database management systems and NoSQL databases.
- Developed a social media query benchmark for both PostgreSQL and Elasticsearch.
- Evaluated the system performance against the query benchmark and different system settings.

MIBIOS: A Database Framework for Microbiome Datasets **Sep. 2020 – May. 2021**

Medical School, University of Michigan *PI: Prof. Thomas Schmidt*

- Developed a database back end for microbiome data.
- Developed a form-based web interface that provides access to the microbiome data and functions that selects desired data with different attributes and exports to different formats.
- Further, developed a data analysis interface that provides widely used data visualizations and statistical analysis methods in the microbiome research community.

OKN: Open Knowledge Network **Jan. 2020 – Dec. 2020**

Transportation Research Institute, University of Michigan *PI: Prof. Robert Hampshire*

- Developed a dataset warehouse and knowledge network for transportation related datasets and research works.
- Developed a web interface for dataset search, based on Elasticsearch and Neo4j.
- Developed a map-view data visualization interface for transportation datasets with Mapbox.

Internship Experience

Adobe **May 2024 – Aug 2024**

Machine Learning Engineer Intern *Supervisor: Dr. Yunyao Li*

- Hybrid Entity Linking for NL2SQL
- Developed a hybrid approach for entity linking that combines string and semantic similarity measures.

Microsoft Research **Feb 2024 – May 2024**

Research Intern *Supervisor: Dr. Yeye He*

- Table-Specialist: Self-trained Language Model Specialist for Tables using Iterative Fine-tuning
- Developed a self-trained framework for LLM fine-tuning for both generative and classification table tasks.

Teaching Experience

EECS484: Database Management Systems **Jan 2022 – Apr 2022**

Instructor, together with Prof. Barzan Mozafari *EECS, University of Michigan*

- Gave lectures for the first-half semester, including topics of database modeling, entity-relationship diagram, relational algebra and basic SQL queries.
- Other services: office hour, managing teaching assistants, and arranging exams.
- I had the privilege of serving as an instructor at the University of Michigan, where I led a lecture series attended by over 350 students. My commitment exceeded 25 hours per week, encompassing a broad range of responsibilities. This role significantly enhanced my skills in student engagement, simplifying complex concepts and algorithms for diverse audiences, and effectively managing a team of 10 teaching assistants.

Technical Skills

Programming Languages: Python, C++, SQL, HTML/CSS, JavaScript
Tools & Frameworks: PyTorch, Tensorflow, PostgreSQL, Elastcsearch, ReactJS