香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

数据科学与分析学域
DATA SCIENCE AND ANALYTICS THRUST
信息枢纽
INFORMATION HUB

DSA Research Experiences for Undergraduates

Research Project

Section1: Faculty Information

Full Name	Wei WANG	Tel	13003187004
Thrust/Hub	DSA/INF	Office	W4 515
Email	weiwcs@ust.hk		

Section2: Research Project Proposal

Project Title	Generative Models for Relational Datasets
Project Description (max 800 words)	1. Overview Relational datasets, characterized by structured tables and inter-table dependencies, underpin critical applications in database management, analytics, and privacy-aware systems. This project investigates advanced generative models—including Probabilistic Circuits (PCs), Tensor Decomposition, and Renormalization Group (RG) methods—to capture the joint distributions of relational data while enabling efficient exact/approximate inference. By enhancing these models' capacity to represent complex relational structures, the project aims to address challenges in database benchmarking, privacy-preserving data publishing, selectivity estimation, and approximate query processing. 2. Research Objectives • Model Adaptation for Relational Data Develop techniques to adapt PCs, tensor models, and RG methods to relational tables and databases, ensuring they capture intra-table correlations (e.g., column dependencies) and inter-table relationships (e.g., foreign keys, joins). Improve expressiveness via hierarchical architectures or attention mechanisms to model high-dimensional, sparse relational data. • Selectivity Estimation for Complex Queries Design inference algorithms to estimate the selectivity of single-table queries and ad hoc multi-table joins (e.g., SQL WHERE clauses with predicates on multiple columns). Enable approximate query processing by sampling synthetic data from the learned generative model while preserving statistical fidelity. • Incremental Model Maintenance Create update mechanisms to efficiently adjust models as underlying data evolves (e.g., row insertions, schema changes), minimizing full retraining costs.

	Explore dynamic tensor factorization or online PC learning to maintain accuracy under streaming data scenarios. 3. Methodology Model Architecture Design: Extend PCs with relational-aware sum-product networks (SPNs) and tensor models with database schema embeddings. Query-Aware Training: Incorporate query workloads into model training to prioritize fidelity for frequent or critical query patterns. Benchmarking: Evaluate models on standard relational datasets (e.g., TPC-H, IMDb) using metrics like query error rate, sample quality (KL divergence), and update latency. 4. Expected Contributions Theoretical: A unified framework for generative modeling of relational data, bridging probabilistic inference and database theory. Technical: Open-source implementations of scalable models for selectivity estimation and data synthesis, compatible with SQL engines. Practical: Tools for privacy-safe data sharing (via synthetic data generation) and improved query optimizers in database systems. 5. Applications Database Benchmarking: Generate synthetic databases with realistic constraints for stress-testing systems. Privacy Preservation: Publish statistically accurate, non-reversible synthetic data for secure analytics. Query Optimization: Accelerate query planning via fast selectivity estimates for complex predicates. 6. Broader Impact The project's models and algorithms will advance scalable AI-driven database tools, enabling efficient analytics on large-scale relational data while addressing privacy concerns. The methodologies developed could

	also inform generative modeling in graph-structured data and time-series databases.
Proposed Research Duration	Start Date: __2025___ / MAR__ / _02____ End Date: __ 2026___ / MAR__ / _01____
Student/Researcher Duties	-
Technical Skills Required	☑ Python ☐ Machine Learning ☐ Big Data ☐ R ☐ Deep Learning ☑ SQL ☐ C/C++ ☑ Other: Interest in cognitive aspect of learning is a plus
Preferred Student/Researcher Background	*List preferred coursework, experience, or skills (e.g., statistics, programming, AI).*
Maximum Number of Students/Researchers	☐ 1 ☑ 2

Section3: Pre-Application Research Exposure Meeting

Faculty members are encouraged to schedule a Research Exposure Meeting to introduce students to their projects.

Preferred Date	Mon / Wed / Fri morning
Preferred Time	n/a
Meeting Mode	☑ In-Person ☑ Online
Venue (if in-person)	
Meeting Link (if online)	On request