VRAJ SHAH

Phone: (+1) 858-291-2697
Email: vps002@eng.ucsd.edu
Web: pvn25.github.io
LinkedIn: linkedin.com/vraj

EDUCATION

University of California, San Diego, CA
Sept 2016 - June 2022
PhD and MS, Computer Science & Engineering
Thesis Advisor: Prof. Arun Kumar

Indian Institute of Technology, Indore, India
Aug 2012 - June 2016
Bachelor of Technology, Computer Science & Engineering

INDUSTRY EXPERIENCE

IBM Research
Staff Research Scientist
May 2023 - Present
Research Scientist
August 2022 - April 2023

- Developing novel AI models and human-in-the-loop tools for natural language processing tasks, while reducing the barrier of entry for domain experts in adopting AI.

Microsoft
June 2018 - Sept 2018
Research Intern

- Implemented computational graph-level optimizations and analytical cost model for ML operators inside Microsoft's deep learning system for inference.

Infor Corporation
June 2017 - Sept 2017
Research Intern

- Integrated ML algorithms inside LogicBlox forecasting engine to scale training with data parallelization strategy.

RESEARCH EXPERIENCE

University of California, San Diego
Sept 2016 - June 2022
Graduate Student Researcher

- **CategDups.** Presents novel data artifacts, benchmarks, and empirical analyses to help ML practitioners prioritise their effort in cleaning Categorical duplicates and Automated machine learning(AutoML) developers to build better deduplication workflows.
- **SortingHat.** Created the first benchmark for ML feature type inference by leveraging database schema semantics to objective quantify and substantially improve the accuracy of the task. This helps to objectively validate and improve AutoML platforms.
- **ML Data Prep Zoo.** Vision of how we leverage ML for systematically standardizing and automating data preparation for ML with a zoo of labeled dataset and ML models.
- **SpeakQL.** Developed a system for making spoken SQL querying effective and efficient. The speech-driven interface allows the users to query in any domain with infinite vocabulary using interactive query correction.
- **Hamlet.** Analyzed the accuracy effects of joins on high-capacity ML algorithms, when learning over normalized data to reduce the data sourcing burden for ML.

University of Alberta, Canada
May 2015 - July 2015
Research Assistant

- Developed a Big Data Adaptor as Eclipse plugin which efficiently handles GitHub's data dump. The tool saves developers' time and effort in data prep for GitHub analysis.

PUBLICATIONS

An Empirical Study on (Non-)Importance of Cleaning Categorical Duplicates before ML.
Vraj Shah, Thomas Parashos, and Arun Kumar.
Under Submission | Paper.

Towards Benchmarking Feature Type Inference for AutoML Platforms.
Vraj Shah, Jonathan Lacanlale, Premanand Kumar, Kevin Yang, Arun Kumar.

SIGMOD 2021 | Paper.

SpeakQL: Towards Speech-driven Multimodal Querying of Structured Data.
Vraj Shah, Side Li, Arun Kumar, Lawrence Saul.
SIGMOD 2020 | Paper.

Demonstration of SpeakQL: Speech-driven Multimodal Querying of Structured Data.
Vraj Shah, Side Li, Kevin Yang, Arun Kumar, Lawrence Saul.
SIGMOD 2019 (*Demo track*) | Paper.

The ML Data Prep Zoo: Towards Semi-Automatic Data Preparation for ML.
Vraj Shah, Arun Kumar.
DEEM Workshop, SIGMOD 2019 | Paper.

SpeakQL: Towards Speech-driven Multi-modal Querying.
Vraj Shah.
SIGMOD 2019 (*Student Research Competition*) | ***Awarded Second Runner-up*** | Paper.

Are Key-Foreign Key Joins Safe to Avoid when Learning High Capacity Classifiers?
Vraj Shah, Arun Kumar, Xiaojin Zhu.
VLDB 2018 | Paper.

SpeakQL: Towards Speech-driven Multi-modal Querying.
Dharmil Chandarana, **Vraj Shah**, Arun Kumar, Lawrence Saul.
HILDA Workshop, SIGMOD 2017 | Paper.

GitHub's Big Data Adaptor: An Eclipse Plugin.
Ali Sajedi, **Vraj Shah**, Eleni Stroulia.
IBM CASCON 2015 | Paper.

RESEARCH IMPACT

- *Improving Feature Type Inference Accuracy of TFDV with SortingHat*
  Vraj Shah, Kevin Yang, and Arun Kumar | Technical Report.
  Models from project SortingHat explored for production use by TensorFlow Data Validation in collaboration with Google.
- Ongoing Collaboration with AWS and OpenML to leverage our data and ML models for ML feature type inference for deployment use.

PATENTS

Speech Based Structured Querying
Arun Kumar, **Vraj Shah**, Dharmil Chandarana

AWARDS

Second Runner-up, ACM SIGMOD Student Research Competition	2019
NSF Travel Award to attend ACM SIGMOD	2019
NSF Travel Award to attend VLDB	2018
Microsoft Travel Award to attend ACM SIGMOD	2017
Research Experience program honor for poster at *University of Alberta International Research Symposium*	2015
MITACS Globalink Research Award	2015
DAAD WISE Fellowship	2015

SERVICE

Program Committe: VLDB 2024, SIGMOD 2024, CODS-COMAD 2024, NeurIPS Table Representation Learning 2022, SIGMOD DEEM 2022, VLDB 2022

External Reviewer: VLDB 2019, VLDB Demo 2018

TEACHING EXPERIENCE

- Teaching Assistant - DSC 102: Advanced Data Analytics Systems Winter 2020
  Co-created the first edition of the course programming assignments (PAs), which includes data exploration with AWS and Dask, and feature engineering and model selec-

tion with Spark. The PAs have been used by 450+ UCSD students so far and are now used in every DSC102 course offering.

- Teaching Assistant - CSE 132C: Database System Implementation Spring 2020

SKILLS

Languages: C, C++, Python, Java, SQL, R.

Web Development: HTML, CSS, JavaScript, PHP.

Tools & Libraries: Scikit-learn, Dask, Keras, Tensorflow, Matlab, AWS EC2/S3.

RELEVANT COURSEWORK

Probabilistic Reasoning and Learning, Machine Learning, Recommender System and Web Mining, Advanced Compiler Design, Principles of Database Systems, Algorithms.

MENTORSHIP EXPERIENCE

Francisco Cornejo-Garcia, BS, Cypress College	*Summer 2020*
Kevin Yang, BS, UCSD.	*Fall 2019 - Spring 2020*
Jonathan Lacanlale, BS, California State University, Northridge	*Summer 2019*
Thomas Parashos, BS, California State University, Northridge	*Summer 2019*