Zhengjie Miao

Ph.D. Candidate in Computer Science

D339 LSRC Building, 308 Research Drive
Durham, NC 27708
☎ *+1 (718) 916 6276*
✉ *zjmiao@cs.duke.edu*
⌂ *www.cs.duke.edu/~zjmiao*

Education

2017–present **Ph.D. in Computer Science**, *Duke University*, Durham NC, *GPA: 3.9/4.0.*
- Dissertation: Explanations in the Data Science Pipeline
- Advisor: Prof. Sudeepa Roy
- Committee: Prof. Ashwin Machanavajjhala, Prof. Aditya Parameswaran (UC Berkeley), Prof. Kristin Stephens-Martinez, and Prof. Jun Yang

2015–2016 **M.S. in Computer Science**, *Columbia University*, New York NY, *GPA: 4.0/4.0.*

2011–2015 **B.S. in Computer Science and Technology**, *Peking University*, Beijing - China, *GPA: 3.5/4.0.*

Research Interests

I broadly work in data management, natural language processing, and visual analytics. Currently my research focuses on providing *explanations* to help people write analytical queries and understand query results, and augmenting their data for data integration and data mining tasks.

Awards

2019 Microsoft Research PhD Fellowship Finalist

2019 Outstanding Ph.D. Research Initiation Project Award, *Computer Science Department, Duke University*

2019 VLDB Travel Grant

2018, 2019 ACM SIGMOD Travel Award

2015 7th Place in ACM/ICPC Greater New York Regional

2014 Award for Excellent Detailed Analysis, *IEEE Visual Analytics Science and Technology (VAST) Challenge Mini-Challenge 1*

2013 The May Fourth Scholarship, *Peking University*

2012 Silver medal in ACM/ICPC Asia Regional Contest in Tianjin

2009 First Prize in National Olympiad in Informatics in Hunan Province

Peer Reviewed Full Research Papers

SIGMOD 21 **Rotom: A Meta-Learned Data Augmentation Framework for Entity Matching, Data Cleaning, Text Classification, and Beyond**, *Zhengjie Miao, Yuliang Li, and Xiaolan Wang.*
ACM SIGMOD International Conference on Management of Data, June 2021

SIGMOD 21 **Putting Things into Context: Rich Explanations for Query Answers using Join Graphs**, *Chenjie Li, Zhengjie Miao, Qitian Zeng, Boris Glavic, and Sudeepa Roy.*
ACM SIGMOD International Conference on Management of Data, June 2021

WWW 20 **Snippext: Semi-supervised Opinion Mining with Augmented Data**, *Zhengjie Miao, Yuliang Li, Xiaolan Wang, and Wang-Chiew Tan*, Link to 📄.
The Web Conference (WWW) 2020, April 2020

SIGMOD 19 **Explaining Wrong Queries Using Small Examples**, *Zhengjie Miao, Sudeepa Roy, and Jun Yang*, Link to 📄.
ACM SIGMOD International Conference on Management of Data, June 2019

SIGMOD 19 **Going Beyond Provenance: Explaining Query Answers with Pattern-based Counterbalances**, *Zhengjie Miao**, *Qitian Zeng**, *Boris Glavic, and Sudeepa Roy*, * denotes equal contribution, Link to 📄.
ACM SIGMOD International Conference on Management of Data, June 2019

CIDR 17 **Combining Design and Performance in a Data Visualization Management System**, *Eugene Wu, Fotis Psallidas, Zhengjie Miao, Haoci Zhang, Laura Rettig, Yifan Wu, Thibault Sellam*, Link to 📄.
Conference on Innovative Data Systems Research, Jan 2017

Peer Reviewed Demonstration Papers

VLDB 20 **I-Rex: An Interactive Relational Query Explainer for SQL**, *Zhengjie Miao, Tiangang Chen, Alexander Bendeck, Kevin Day, Sudeepa Roy, and Jun Yang*, Link to 📄.
Proceedings of the VLDB Endowment (PVLDB), Vol. 13, No. 12

VLDB 19 **CAPE: Explaining Outliers by Counterbalancing**, *Zhengjie Miao**, *Qitian Zeng**, *Chenjie Li, Boris Glavic, Oliver Kennedy, and Sudeepa Roy*, * denotes equal contribution, Link to 📄.
Proceedings of the VLDB Endowment (PVLDB), Vol. 12, No. 12

VLDB 19 **LensXPlain: Visualizing and Explaining Contributing Subsets for Aggregate Query Answers**, *Zhengjie Miao, Andrew Lee, and Sudeepa Roy*, Link to 📄.
Proceedings of the VLDB Endowment (PVLDB), Vol. 13, No. 12

SIGMOD 19 **RATest: Explaining Wrong Relational Queries Using Small Examples**, *Zhengjie Miao, Sudeepa Roy, and Jun Yang*, Link to 📄.
ACM SIGMOD International Conference on Management of Data, June 2019

Research Experience

2017–present **Database Research Group**, *Duke University*.
Research Assistant, advised by Prof. Sudeepa Roy and Prof. Jun Yang.

- **Helping Novices Learn and Debug Relational Queries** ↗[Project website]
  - Designed and implemented web-based debugging tools for Relational Algebra and SQL, which find a small counterexample for two input queries where the input queries return different results, and allow syntax-consistent tracing for the query execution.
  - Designed and implemented algorithms to find the smallest counterexample using data provenance and SMT solvers
  - Designed algorithms for generating generalized explanations on the semantic differences between queries
  - Mentored a group of graduate and undergraduate students on designing and implementing features of our tools

- **Explaining Surprising Query Answers Using Patterns**
  - Designed the framework that provides explanations for surprising outcomes of an aggregate query by finding patterns and outliers in the data
  - Formalized the concept of aggregate regression patterns and the definition of counterbalancing explanations using aggregate regression patterns
  - Designed and implemented the explanation generating algorithm

Summer 2021 **Microsoft Research**.
Research Intern, supervised by Dr. Yeye He.

- **Automatic next step suggestion for data preparation**
  - Designed and implemented a learning-based algorithm to suggest Pandas operators for Jupyter notebooks

Summer 2020 **Megagon Labs**.
Research Intern, supervised by Dr. Yuliang Li and Dr. Wang-Chiew Tan.

- **Automatic discovery of data augmentation policies for DB and NLP tasks**
  - Designed and implemented a meta-learned data augmentation framework for sequence classification tasks (text classification, entity matching, error detection, etc.) based on pre-trained language models
  - Proposed the optimization that enables the model to learn how to choose and combine augmented data

Summer 2019 **Megagon Labs**.
Research Intern, supervised by Dr. Yuliang Li and Dr. Wang-Chiew Tan.

- **Opinion extraction for building subjective databases**
  - Studied problems on data augmentation and semi-supervised learning for aspect-based sentiment analysis
  - Designed and implemented MixDA, a novel data augmentation technique by interpolating the representations of text sequences

Services

Journal Reviewer: ACM Transactions on Database Systems (TODS)

Reviewer: ICDT (2021)

Committee Member: PVLDB Reproducibility (2019)

Student Mentor: CS+: CompSci Projects Beyond the Classroom, Duke University (2020)

Student Volunteer: ACM SIGMOD (2020)

Teaching Experience

VLDB 2021 Tutorial **Data Augmentation for ML-driven Data Preparation and Integration**, *Yuliang Li, Xiaolan Wang, Zhengjie Miao, and Wang-Chiew Tan*.

Spring 2019 Teaching Assistant, Introduction to Database Systems (Duke CompSci 316)

- Assisted in writing and grading the assignments and projects; deployed the RATest debugging tool for Relational Algebra Queries.

Spring 2018 Teaching Assistant, Everything Data (Duke CompSci 216)

- Assisted in writing and grading the assignments, labs, and projects.