

# Zhengjie Miao

*Ph.D. Candidate in Computer Science*

D339 LSRC Building, 308 Research Drive  
Durham, NC 27708  
☎ +1 (919) 660-6594  
✉ [zjmiao@cs.duke.edu](mailto:zjmiao@cs.duke.edu)  
🌐 [www.cs.duke.edu/~zjmiao](http://www.cs.duke.edu/~zjmiao)  
📀 QWv8VwYAAAAJ


## Research Interests

I broadly work in data management, machine learning, and visual analytics. Currently my research focuses on helping non-expert users (1) write analytical queries and understand query results; and (2) integrate and prepare their data efficiently using machine learning techniques.

## 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: Sudeepa Roy
  - Committee: Ashwin Machanavajjhala, Aditya Parameswaran (UC Berkeley), Kristin Stephens-Martinez, and Jun Yang
- 2015–2016 **M.S. in Computer Science**, *Columbia University*, New York NY, *GPA: 4.0/4.0*.
- Advisor: Eugene Wu
- 2011–2015 **B.S. in Computer Science and Technology**, *Peking University*, Beijing - China, *GPA: 3.5/4.0*.
- Advisor: Xiaoru Yuan

## Research Experience

- 2017–present **Research Assistant**, *Duke Database Research Group*, supervised by Prof. Sudeepa Roy and Prof. Jun Yang.
- **HNRQ: Helping Novices Learn and Debug Relational Queries**  [Project website]
    - Built tools helping people understand and debug Relational Algebra and SQL queries. For two input queries, HNRQ tools find a small counterexample where the input queries return different results, and allow syntax-consistent tracing for the query execution.
    - Mentored a group of graduate and undergraduate students on building the tools.
  - **Explaining Surprising Query Answers**: Built frameworks that provide explanations for surprising outcomes of an aggregate query by finding patterns and outliers in the data, or by augmenting the provenance with automatic join path discovery.
- Summer 2021 **Research Intern**, *Microsoft Research (DMX Group)*, supervised by Dr. Yeye He.
- **Automatic next step suggestion for data preparation**: Designed and implemented learning-based algorithms to recommend table-manipulation operators (e.g. Pandas DataFrame operators) for data preparation pipelines.
- Summer 2020 **Research Intern**, *Megagon Labs*, supervised by Dr. Yuliang Li and Dr. Wang-Chiew Tan.
- **Automatic discovery of data augmentation policies for DB and NLP tasks**: Built a meta-learned data augmentation framework for sequence classification tasks (text classification, entity matching, error detection, etc.) using pre-trained language models.
- Summer 2019 **Research Intern**, *Megagon Labs*, supervised by Dr. Yuliang Li and Dr. Wang-Chiew Tan.
- **Opinion extraction for building subjective databases**: Designed and implemented Snippet, an label-efficient opinion mining pipeline using novel data augmentation techniques.

---

## Peer Reviewed Full 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, Link to [📄](#).  
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, Link to [📄](#).  
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 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<sup>\*</sup>, Qitian Zeng<sup>\*</sup>, Boris Glavic, and Sudeepa Roy, <sup>\*</sup> 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 Short/Demonstration Papers

- VLDB 21 **Data Augmentation for ML-driven Data Preparation and Integration**, Yuliang Li, Xiaolan Wang, Tutorial Zhengjie Miao, and Wang-Chiew Tan, Link to [📄](#).  
Proceedings of the VLDB Endowment (PVLDB), Vol. 14 No. 12
- 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<sup>\*</sup>, Qitian Zeng<sup>\*</sup>, Chenjie Li, Boris Glavic, Oliver Kennedy, and Sudeepa Roy, <sup>\*</sup> 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. 12, 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

---

## Awards

- 2019 Microsoft Research PhD Fellowship Finalist
- 2019 Outstanding Ph.D. Research Initiation Project Award, *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

---

## Professional Services

- 2022 **External Reviewer:** International Conference on Data Engineering
- 2021 **External Reviewer:** International Conference on Database Theory
- 2020 **Journal Reviewer:** ACM Transactions on Database Systems
- 2020 **Student Volunteer:** ACM SIGMOD International Conference on Management of Data
- 2019 **Committee Member:** Proceedings of the VLDB Endowment Reproducibility

---

## Teaching & Mentoring Experience

- Summer 2020 Undergrad Student Mentor, CS+: CompSci Projects Beyond the Classroom, Duke University
- Spring 2019 Teaching Assistant, Introduction to Database Systems (Duke CompSci 316)
- Spring 2018 Teaching Assistant, Everything Data (Duke CompSci 216)

---

## References

**Sudeepa Roy (sudeepa@cs.duke.edu)**, Assistant Professor of Computer Science, Duke University

**Jun Yang (junyang@cs.duke.edu)**, Bishop-MacDermott Family Professor of Computer Science, Duke University

**Wang-Chiew Tan (wangchiew@fb.com)**, Research Scientist, Facebook AI

**Yeye He (yeyehe@microsoft.com)**, Principal Researcher, Microsoft Research