Arun Kumar

3218 CSE (EBU3b) and 351 HDSI
9500 Gilman Drive, Mail Code 0404
La Jolla, CA 92093

Email: akk018@ucsd.edu
Web: https://cseweb.ucsd.edu/~arunkk/

EMPLOYMENT

University of California, San Diego

Associate Professor	From 2021
Department of Computer Science and Engineering	
Halıcıoğlu Data Science Institute	
HDSI Faculty Fellow	From 2021
Assistant Professor	
Department of Computer Science and Engineering	2016–2021
Halıcıoğlu Data Science Institute	2019–2021

EDUCATION

University of Wisconsin-Madison

Ph.D. in Computer Sciences.	2011–2016
M.S. in Computer Sciences.	2009–2011

Indian Institute of Technology, Madras

B.Tech. in Computer Science and Engineering.	2005–2009

RESEARCH INTERESTS

Data management and systems for machine learning/artificial intelligence-based data analytics, with a focus on problems related to usability, scalability, resource efficiency, and ease of development/deployment. I enjoy working on problems that are motivated by real applications and are formally grounded. My work spans the whole gamut of building new systems, new algorithmics, new datasets, theoretical analysis, empirical analysis, and working with data/ML/AI practitioners (data scientists, ML/data/software engineers, and domain scientists) to help deploy my research.

Research Group Webpage: https://adalabucsd.github.io/

AWARDS AND HONORS

VLDB Distinguished Associate Editor	2022, 2021
IEEE TCDE Rising Star Award	2021
Amazon Research Award	2021
HDSI Faculty Fellow	2021
Invited Keynote at Brazilian Symposium on Databases	2021
VMware Early Career Faculty Grant	2021, 2020
NSF CAREER Award	2020
ACM SIGMOD Research Highlight Award	2020
Google Faculty Research Award	2020, 2017
Invited Paper at ACM Transactions on Database Systems	2020, 2016
Honorable Mention for Best Paper Award at ACM SIGMOD	2019
ACM SIGMOD Distinguished PC Member	2019, 2017
VLDB Distinguished PC Member	2019
Hellman Fellowship	2018
Faculty of the Year from UCSD oSTEM Chapter	2018
Invited Keynote at ACM SIGMOD DEEM Workshop	2017
UW-Madison CS Graduate Student Research Award for best PhD research	2016
Anthony C. Klug NCR Fellowship in Database Systems	2015

Best Paper Award at ACM SIGMOD 2014
Invited Paper at the Communications of the ACM 2013
National Talent Search Exam (NTSE) Scholarship from Govt. of India 2003–08

MAJOR ONGOING PROJECTS

Project Cerebro From 2018
A layered data platform for simplifying deep learning model building at scale on all axes–model size, dataset size, and tasks–and on all kinds of execution backends.
Project Webpage: `https://adalabucsd.github.io/cerebro.html`

Project SortingHat From 2018
Benchmark tasks, labeled datasets, and empirical evaluation of tools for automated data preparation on ML platforms.
Project Webpage: `https://adalabucsd.github.io/sortinghat.html`

LLMs for NL2SQL From 2023
Benchmark tasks, labeled datasets, and empirical evaluation of data-centric gaps in LLM-based natural language to SQL interfaces.

SELECTED RESEARCH IMPACT

Ideas from SATURN being explored for internal use by Netflix. 2023–24
Ideas from SATURN predecessor HYDRA being used by UCSD computational physics researchers. 2022–23
Ideas from CEREBRO and SATURN being explored for internal use by Meta. 2022–24
CEREBRO being used by UCSD public health researchers. 2019–now
Models/ideas from project SORTINGHAT being explored for adoption by OpenML, AutoGluon, and Amazon Web Services. 2021–22
Models/ideas from project SORTINGHAT explored for integration with TensorFlow Data Validation in collaboration with Google TFX team. 2020–21
Ideas from project CEREBRO integrated into MADlib and Greenplum in collaboration with Pivotal/VMWare. 2019–20
Ideas from project MORPHEUS being integrated into GraalVM by Oracle. 2019–20
Ideas from bolt-on differential privacy paper integrated into TensorFlow Privacy by Georgian Partners and Google. 2019
Code from project MORPHEUS used internally by Avito for e-commerce. 2018
Benchmarking results from project SLAB led to bug fixes and feature earmarks in IBM/Apache SystemML. 2018
Ideas from bolt-on differential privacy paper adopted for various customer use cases by Georgian Partners. 2018
Ideas from project MORPHEUS and ORION explored for internal use by Oracle for banking analytics, Google for ad analytics. 2017
Ideas from project HAMLET used internally by LogicBlox for retail analytics, Facebook for friend recommendations, and MakeMyTrip for customer analytics. 2016
Ideas from project ORION used internally by LogicBlox for retail analytics and Microsoft for Web security analytics. 2015–16
BISMARCK system used for healthcare analytics at U Washington. 2013
Code/ideas from project BISMARCK shipped as part of analytics products by Oracle, Pivotal, and Cloudera. 2011–13
Code from project BISMARCK contributed to the Apache MADlib library. 2011–12
Full list of research impact notes: `https://adalabucsd.github.io/impact.html`

PUBLICATIONS SUMMARY

Full papers at top-tier conferences (SIGMOD/VLDB/CIDR): 30
Other peer-reviewed conference and journal papers: 22
Peer-reviewed workshop and demonstration papers: 15
Papers under submission/preparation: 5

Number of citations: 3248 and h-index: 28 (as per Google Scholar in Apr 2024)
Full list of publications: `https://adalabucsd.github.io/publications.html`

CONFERENCE PUBLICATIONS

Saturn: An Optimized Data System for Multi-Large-Model Deep Learning Workloads
Kabir Nagrecha and Arun Kumar
VLDB 2024

How do Categorical Duplicates Affect ML? A New Benchmark and Empirical Analyses
Vraj Shah, Thomas Parashos, and Arun Kumar
VLDB 2024

Lotan: Bridging the Gap Between GNNs and Scalable Graph Analytics Engines
Yuhao Zhang and Arun Kumar
VLDB 2023

Database-Aware ASR Error Correction for Speech-to-SQL Parsing
Yutong Shao, Arun Kumar, and Ndapandula Nakashole
IEEE ICASSP 2023

Nautilus: An Optimized System for Deep Transfer Learning over Evolving Training Datasets
Supun Nakandala and Arun Kumar
ACM SIGMOD 2022

Distributed Deep Learning on Data Systems: A Comparative Analysis of Approaches
Yuhao Zhang, Frank McQuillan, Nandish Jayaram, Nikhil Kak, Ekta Khanna, Orhan Kislal, Domino Valdano, and Arun Kumar
VLDB 2021

Towards an Optimized GROUP BY Abstraction for Large-Scale Machine Learning
Side Li and Arun Kumar
VLDB 2021

Towards A Polyglot Framework for Factorized ML
David Justo, Lukas Stadler, Nadia Polikarpova, and Arun Kumar
VLDB 2021 Industrial Track

Towards Benchmarking Feature Type Inference for AutoML Platforms
Vraj Shah, Jonathan Lacanlale, Premanand Kumar, Kevin Yang, and Arun Kumar
ACM SIGMOD 2021

Cerebro: A Layered Data Platform for Scalable Deep Learning
Arun Kumar, Supun Nakandala, Yuhao Zhang, Side Li, Advitya Gemawat, and Kabir Nagrecha
CIDR 2021 (Vision paper)

Cerebro: A Data System for Optimized Deep Learning Model Selection
Supun Nakandala, Yuhao Zhang, and Arun Kumar
VLDB 2020

Panorama: A Data System for Unbounded Vocabulary Querying over Video
Yuhao Zhang and Arun Kumar
VLDB 2020

Understanding and Benchmarking the Impact of GDPR on Database Systems
Supreeth Shastri, Vinay Banakar, Melissa Wasserman, Arun Kumar, and Vijay Chidambaram
VLDB 2020

Vista: Declarative Feature Transfer from Deep CNNs at Scale
Supun Nakandala and Arun Kumar
ACM SIGMOD 2020

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

Incremental and Approximate Inference for Faster Occlusion-based Deep CNN Explanations
Supun Nakandala, Arun Kumar, and Yannis Papakonstantinou
ACM SIGMOD 2019
Honorable Mention for Best Paper Award
Invited to ACM TODS 2020
Invited to ACM SIGMOD Research Highlights 2020

Enabling and Optimizing Non-linear Feature Interactions in Factorized Linear Algebra
Side Li, Lingjiao Chen, and Arun Kumar
ACM SIGMOD 2019

Model-based Pricing for Machine Learning in a Data Marketplace
Lingjiao Chen, Paraschos Koutris, and Arun Kumar
ACM SIGMOD 2019

Tuple-oriented Compression for Large-scale Mini-batch Stochastic Gradient Descent
Fengan Li, Lingjiao Chen, Yijing Zeng, Arun Kumar, Jeffrey Naughton, Jignesh M. Patel, and Xi Wu
ACM SIGMOD 2019

Hierarchical and Distributed Machine Learning Inference Beyond the Edge
Anthony Thomas, Yunhui Guo, Yeosong Kim, Baris Aksanli, Arun Kumar, and Tajana S. Rosing
ICNSC 2019

A Comparative Evaluation of Systems for Scalable Linear Algebra-based Analytics
Anthony Thomas and Arun Kumar
VLDB 2018/2019

In-RDBMS Hardware Acceleration of Advanced Analytics
Divya Mahajan, Joon Kyung Kim, Jacob Sacks, Adel Ardalan, Arun Kumar, and Hadi Esmaeilzadeh
VLDB 2018

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

Materialization Trade-offs for Feature Transfer from Deep CNNs for Multimodal Data Analytics
Supun Nakandala and Arun Kumar
SysML 2018 (Short paper)

Towards Linear Algebra over Normalized Data
Lingjiao Chen, Arun Kumar, Jeffrey Naughton, and Jignesh M. Patel
VLDB 2017

Bolt-on Differential Privacy for Scalable Stochastic Gradient Descent-based Analytics
Xi Wu, Fengan Li, Arun Kumar, Kamalika Chaudhuri, Somesh Jha, and Jeffrey Naughton
ACM SIGMOD 2017

CEREBRO*: A System to Manage Deep Learning for Relational Data Analytics*
Arun Kumar
CIDR 2017 (Abstract)

To Join or Not to Join? Thinking Twice about Joins before Feature Selection
Arun Kumar, Jeffrey Naughton, Jignesh M. Patel, and Xiaojin Zhu
ACM SIGMOD 2016

Learning Generalized Linear Models Over Normalized Data
Arun Kumar, Jeffrey Naughton, and Jignesh M. Patel
ACM SIGMOD 2015

Materialization Optimizations for Feature Selection Workloads
Ce Zhang, Arun Kumar, and Christopher Ré
ACM SIGMOD 2014
Best Paper Award
Invited to ACM TODS 2016)

Brainwash: A Data System for Feature Engineering
Michael Anderson, Dolan Antenucci, Victor Bittorf, Matthew Burgess, Michael Cafarella, Arun Kumar, Feng Niu, Yongjoo Park, Christopher Re, and Ce Zhang
CIDR 2013 (Vision paper)

Probabilistic Management of OCR Data Using an RDBMS
Arun Kumar, and Christopher Ré
VLDB 2012

The MADlib Analytics Library: Or MAD Skills, the SQL
Joseph M. Hellerstein, Christopher Ré, Florian Schoppmann, Daisy Zhe Wang, Eugene Fratkin, Aleksander Gorajek, Kee Siong Ng, Caleb Welton, Xixuan Feng, Kun Li, and Arun Kumar
VLDB 2012 (Industrial track)

Towards a Unified Architecture for in-RDBMS Analytics
X. Feng*, Arun Kumar*, B. Recht, and Christopher Ré (*alphabetical order of surnames)
ACM SIGMOD 2012

Mobile Data Collection in WSNs Using Wireless Communication
Arun Kumar and K. M. Sivalingam
IEEE/ACM COMSNETS 2010

JOURNAL PUBLICATIONS AND BOOKS

Low movement, deep-learned sitting patterns, and sedentary behavior in the International Study of Childhood Obesity, Lifestyle, and the Environment (ISCOLE) Paul R. Hibbing, Jordan A. Carlson, Chelsea Steel, Mikael Anne Greenwood-Hickman, Supun Nakandala, Marta M. Jankowska, John Bellettiere, Jingjing Zou, Andrea Z. LaCroix, Arun Kumar, Peter T. Katzmarzyk, and Loki Natarajan
International Journal of Obesity 2023

Database Education at UC San Diego
Arun Kumar, Alin Deutsch, Amarnath Gupta, Yannis Papakonstantinou, Babak Salimi, and Victor Vianu
ACM SIGMOD Record 2022 (Invited)

CHAP-child: An open source method for estimating sit-to-stand transitions and sedentary bout patterns from hip accelerometers among children
Jordan A. Carlson et al. (15 authors)

International Journal of Behavioral Nutrition and Physical Activity 2022

Structured Data Representation in Natural Language Interfaces
Yutong Shao, Arun Kumar, and Ndapandula Nakashole
IEEE Data Engineering Bulletin 2022 (Invited)

CHAP-Adult: A Reliable and Valid Algorithm to Classify Sitting and Measure Sitting Patterns Using Data from Hip- Worn Accelerometers in Adults Aged 35+
John Bellettiere et al. (14 authors)
Journal for the Measurement of Physical Behaviour 2022

VLDB Scalable Data Science Category: The Inaugural Year
Arun Kumar, Alon Halevy, and Nesime Tatbul
ACM SIGMOD Record 2022

Diversity and Inclusion Activities in Database Conferences: A 2021 Report
Sihem Amer-Yahia et al. (33 authors)
ACM SIGMOD Record 2022

VLDB Panel Summary: "The Future of Data(base) Education: Is the Cow Book Dead?"
Zachary Ives, Johannes Gehrke, Jana Giceva, Arun Kumar, and Rachel Pottinger
ACM SIGMOD Record 2021

The CNN Hip Accelerometer Posture (CHAP) Method for Classifying Sitting Patterns from Hip Accelerometers: A Validation Study
Mikael Anne Greenwood-Hickman, Supun Nakandala, Marta M. Jankowska, Fatima Tuz-Zahra, John Bellettiere, Jordan Carlson, Paul R. Hibbing, Jingjing Zou, Andrea Z. LaCroix, Arun Kumar, and Loki Natarajan
Medicine and Science in Sports and Exercise Journal, 2021

Application of Convolutional Neural Network Algorithms for Advancing Sedentary and Activity Bout Classification
Supun Nakandala, Marta M. Jankowska, Fatima Tuz-Zahra, John Bellettiere, Jordan Carlson, Andrea Z. LaCroix, Sheri J. Hartman, Dori Rosenberg, Jingjing Zou, Arun Kumar, and Loki Natarajan
Journal for the Measurement of Physical Behaviour, 2021

Query Optimization for Faster Deep CNN Explanations
Supun Nakandala, Arun Kumar, and Yannis Papakonstantinou
ACM SIGMOD Record 2020
ACM SIGMOD Research Highlight Award

Incremental and Approximate Computations for Accelerating Deep CNN Inference
Supun Nakandala, Kabir Nagrecha, Arun Kumar, and Yannis Papakonstantinou
ACM TODS 2020 (**Invited paper**)

Data Management in Machine Learning Systems
Matthias Boehm, Arun Kumar, and Jun Yang
Synthesis Lectures on Data Management, Morgan & Claypool Publ. (Book), 2019

Materialization Optimizations for Feature Selection Workloads
Ce Zhang, Arun Kumar, and Christopher Ré
ACM TODS 2016 (**Invited paper**)

Model Selection Management Systems: The Next Frontier of Advanced Analytics
AruArun Kumar, Robert McCann, Jeffrey Naughton, and Jignesh M. Patel

ACM SIGMOD Record Dec 2015 (Vision paper)

On Reducing Delay in Mobile Data Collection-Based WSNs
Arun Kumar, Krishna M. Sivalingam, and Adithya Kumar
Springer Wireless Networks 2012

WORKSHOPS, POSTERS, DEMOS, AND OTHER PEER-REVIEWED PUBLICATIONS

Intermittent Human-in-the-Loop Model Selection using Cerebro: A Demonstration
Liangde Li, Supun Nakandala, and Arun Kumar
VLDB 2021 Demo

Predicting Eating Events in Free Living Individuals
Jiayi Wang, Jiue-An Yang, Supun Nakandala, Arun Kumar and Marta M. Jankowska
eScience 2019 Conference (Poster)

Demonstration of Krypton: Optimized CNN Inference for Occlusion-based Deep CNN Explanations
Allen Ordookhanians, Xin Li, Supun Nakandala, and Arun Kumar
VLDB 2019 Demo

The ML Data Prep Zoo: Towards Semi-Automatic Data Preparation for ML
Vraj Shah Shah and Arun Kumar
ACM SIGMOD 2019 DEEM Workshop

Cerebro: Efficient and Reproducible Model Selection on Deep Learning Systems
Supun Nakandala, Yuhao Zhang, and Arun Kumar
ACM SIGMOD 2019 DEEM Workshop

Demonstration of SpeakQL: Speech-driven Multimodal Querying of Structured Data
Vraj Shah, Side Li, K. Yang, Arun Kumar, and Lawrence Saul
ACM SIGMOD 2019 Demo

Demonstration of Nimbus: Model-based Pricing for Machine Learning in a Data Marketplace
Lingjiao Chen, Hongyi Wang, Leshang Chen, Paraschos Koutris, and Arun Kumar
ACM SIGMOD 2019 Demo

Demonstration of Krypton: Incremental and Approximate Inference for Faster Occlusion-based Deep CNN Explanations
Supun Nakandala, Arun Kumar, and Yannis Papakonstantinou
SysML 2019 Demo

Model-based Pricing: Do Not Pay for More than What You Learn!
Lingjiao Chen, Paraschos Koutris, and Arun Kumar
ACM SIGMOD 2017 DEEM Workshop

SpeakQL: Towards Speech-driven Multi-modal Querying
D. Chandarana, Vraj Shah, Arun Kumar, and Lawrence Saul
ACM SIGMOD 2017 HILDA Workshop

Demonstration of Santoku: Optimizing Machine Learning over Normalized Data
Arun Kumar, Mona Jalal, Boqun Yan, Jeffrey Naughton, and Jignesh M. Patel
VLDB 2015 Demo

Hazy: Making it Easier to Build and Maintain Big-data Analytics
Arun Kumar, Feng Niu, and Christopher Ré
ACM Queue 2013
Invited to the Communications of the ACM

Distributed and Scalable PCA in the Cloud

Arun Kumar, Nikos Karampatziakis, Paul Mineiro, Markus Weimer, and Vijay Narayanan
NIPS BigLearn Workshop 2013

Feature Selection in Enterprise Analytics: A Demonstration using an R-based Data Analytics System
Pradap Konda, Arun Kumar, Christopher Ré, and Vaishnavi Sashikanth
VLDB 2013 Demo

Flexible Multimedia Content Retrieval Using InfoNames
Arun Kumar, Ashok Anand, Athula Balachandran, Vyas Sekar, Aditya Akella, Srinivasan Seshan
ACM SIGCOMM 2010 Demo

TECHNICAL REPORTS, MANUSCRIPTS, AND ARTICLES

SNAILS: Schema Naturalness Assessments for Improved LLM Systems
Kyle Luoma and Arun Kumar
Under submission

Routing Over LLMs Using Proxy Metrics for Relative Quality Estimation
Kabir Nagrecha, Arun Kumar, and Hao Zhang
Under submission

Development, Validation, and Transportability of Several Machine-Learned, Non-Exercise Based VO2max Prediction Models for Older Adults
Benjamin T. Schumacher, Michael J. LaMonte, Andrea Z. LaCroix, Eleanor M. Simonsick, Steven P. Hooker, Humberto Parada Jr., John Bellettiere, and Arun Kumar.
Under submission

Contribution to "Reminiscences on Influential Papers"
Arun Kumar
ACM SIGMOD Record 2023

Processing Analytical Queries in the AWESOME Polystore
Xiuwen Zheng, Arun Kumar, and Amarnath Gupta
UCSD Technical Report

Design and Evaluation of An SQL Dialect for Speech-Based Querying
Kyle Luoma and Arun Kumar
UCSD Technical Report

Hydra: An Optimized Data System for Large Multi-Model Deep Learning
Kabir Nagrecha and Arun Kumar
UCSD Technical Report

Some Damaging Delusions of Deep Learning Practice (and How to Avoid Them)
Arun Kumar, Supun Nakandala, and Yuhao Zhang
Extended talk abstract for ACM SIGKDD Deep Learning Day, 2021

Automation of Data Prep, ML, and Data Science: New Cure or Snake Oil?
Arun Kumar
ACM SIGMOD 2021 Panel

Letter from the Rising Star Award Winner
Arun Kumar
IEEE Data Engineering Bulletin, June 2021

VigilaDE: Avoiding False Discoveries with Hierarchical Data in Data Exploration Systems
Nikos Koulouris, Arun Kumar, and Yannis Papakonstantinou
Manuscript, 2020

Scalable Data Science: A New Research Track Category at PVLDB Vol 14 / VLDB 2021
Alon Halevy, Arun Kumar, and Nesime Tatbul
Article on ACM SIGMOD Blog (webpage), 2020

MLSys: The New Frontier of Machine Learning Systems
Alexander Ratner et al. (about 70 authors)
Manuscript on arXiv

ML/AI Systems and Applications: Is the SIGMOD/VLDB Community Losing Relevance?
Arun Kumar
Article on ACM SIGMOD Blog (webpage), 2018

Learning Over Joins
Arun Kumar
UW-Madison CS PhD Dissertation, 2016

A Survey of the Existing Landscape of ML Systems
Arun Kumar, Robert McCann, Jeffrey Naughton, and Jignesh M. Patel
UW-Madison CS Technical Report TR1827, 2015

InfoNames: An Information-Based Naming Scheme for Multimedia Content
Arun Kumar, Ashok Anand, Athula Balachandran, Vyas Sekar, Aditya Akella, Srinivasan Seshan
UW-Madison CS Technical Report TR 1677, 2010

INDUSTRY BLOGS AND TALKS FROM MY GROUP

Saturn: Efficient Multi-Large-Model Deep Learning
Kabir Nagrecha and Arun Kumar
Baylearn Machine Learning Symposium — Oct 2023

Systems for Machine Learning
Kabir Nagrecha
Article on Gradient Pub: webpage. — Aug 2021

Massively Parallel Automated Model Building for Deep Learning
Advitya Gemawat and Frank McQuillan
Article on VMware Blog: webpage. — Apr 2021

Resource-Efficient Deep Learning Model Selection on Apache Spark
Yuhao Zhang and Supun Nakandala
Talk at Spark+AI Summit: webpage. — Jun 2020

Model Selection for Deep Neural Networks on Greenplum Database
Frank McQuillan, Yuhao Zhang, and Arun Kumar
Article on VMware Blog: webpage. — Apr 2020

Faster ML over joins of tables
Arun Kumar
Talk at O'Reilly Strata Data Conferences: webpage. — Mar 2019

Low-Rank Matrix Factorization in Oracle R Advanced Analytics for Hadoop
Arun Kumar
Article on Oracle Blog: webpage. — Feb 2014

TEACHING

CSE 234: Data Systems for Machine Learning. UCSD. Winter '24; Winter '23; Fall '21, '20
CSE 132C: Database System Implementation. UCSD. Spring '23, '22, '21, '20
DSC 208R: Data Management for Analytics. UCSD OMDS. Winter '23

DSC 102: Systems for Scalable Analytics. UCSD. Fall '22; Winter '22, '21, '20
CSE 239: Database Seminar. UCSD Fall '22, '21, '20, '19
CSE 234: Data Systems for Machine Learning. UCSD. Fall '21, '20
CSE 232A: Graduate Database Systems. UCSD. Fall '19, '18
CSE 190: Topics in Database System Implementation. UCSD. Spring '19, '18, '17
CSE 291: Advanced Data Analytics and ML Systems. UCSD. Winter '19, '18, '17
CSE 290: Seminar on Integrative AI Engineering. UCSD. Fall '18
CSE 290: Seminar on Advanced Data Science. UCSD. Fall '17, Spring '17
CS 564: DBMS: Design and Implementation. UW-Madison. Fall '15

ADVISING (CURRENT)

Kyle Luoma, PhD, CSE, UCSD. Co-advisor: Jingbo Shang. Fall 2021–
Xiuwen Zheng, PhD, CSE, UCSD. Co-advisor: Amarnath Gupta. Fall 2020–
Animesh Kumar, MS, CSE, UCSD. Spring 2024–

ADVISING (ALUMNI)

Kabir Nagrecha, PhD, CSE, UCSD (Co-advisor: Hao Zhang). Fall 2021–Spring 2024
First employment: Netflix.

Yuhao Zhang, PhD & MS, CSE, UCSD. Fall 2018–Fall 2023
First employment: Databricks.

Pradyumna Sridhara, MS, CSE, UCSD. Winter 2022–Spring 2023
First employment: UCSD HDSI.

Tanay Karve, MS, CSE, UCSD. Spring 2022–Winter 2023
First employment: Apple.

Vignesh Nanda Kumar, MS, CSE, UCSD. Winter 2022–Winter 2023
First employment: ServiceNow.

Vraj Shah, PhD & MS, CSE, UCSD. Fall 2016–Spring 2022
First employment: IBM Research Almaden.

Liangde Li, MS, CSE, UCSD. Spring 2021–Spring 2022
First employment: TigerGraph.

Supun Nakandala, PhD, CSE, UCSD. Fall 2017–Winter 2022
First employment: Databricks.

Tara Mirmira, MS, CSE, UCSD. Fall 2019–Winter 2022
First employment: PhD at UCSD.

Advitya Gemawat, BS, HDSI, UCSD. Winter 2019–Spring 2021
First employment: Microsoft NERD AI.

Side Li, MS, CSE, UCSD. Fall 2019–Spring 2021
First employment: Google.

Kabir Nagrecha, BS, CSE, UCSD. Fall 2019–Spring 2021
First employment: PhD at UCSD.

Shaoqing Yi, BS, HDSI and Math, UCSD. Fall 2020–Spring 2021
First employment: PhD at UC Berkeley.

Kevin Yang, BS, CSE, UCSD. Fall 2018–Spring 2020
First employment: MS at UPenn.

David Justo, MS, CSE, UCSD (Co-advisor: Nadia Polikarpova). Spring–Fall 2019
First employment: Microsoft.

Lingjiao Chen, MS, CS, UW-Madison. Fall 2015–Fall 2018
First employment: PhD at Stanford.

Side Li, BS, CSE, UCSD. Fall 2017–Spring 2018
First employment: Amazon.

Anthony Thomas, MS, CSE, UCSD. Spring 2017–Spring 2018
First employment: PhD at UCSD

Mingyang Wang, MS, CSE, UCSD. Spring 2017
First employment: Amazon.

Fengan Li, BS, CS, UW-Madison. Fall 2015–Spring 2016
First employment: Google

Mona Jalal, MS, CS, UW-Madison. Fall 2014–Spring 2015

Boqun Yan, BS, CS, UW-Madison. Fall 2014–Spring 2015

UG INTERNS VIA STARS

Francisco Cornejo-Garcia; Cypress College. Summer 2020

Jonathan Lacanlale; California State University, Northridge. Summer 2019

Thomas Parashos; California State University, Northridge. Summer 2019

HIGH SCHOOL MENTEES VIA MAP

Annabel Ng; Westview High School; College: UC Berkeley EECS 2021–22

Jenny Mar; Canyon Crest Academy; College: UC San Diego CSE 2021–22

Nitya Pillai; Del Norte High School; College: UC San Diego CSE 2021–22

James Zhang; Helix Charter High School; College: UC San Diego CSE 2019–20

STUDENT AWARDS AND HONORS

Supun Nakandala wins the highly prestigious ACM SIGMOD Jim Gray Doctoral Dissertation Award. He is the first UCSD student to receive this award and this is the first time this award goes to work in the DB for ML / ML systems space. 2023

Kabir Nagrecha wins a highly competitive Meta PhD Fellowship, the first UCSD student to receive one in this fellowship program's 10 year history. 2022

Kabir Nagrecha wins first place at the UG level of the ACM SIGMOD 2021 Student Research Competition. This is the first time a UCSD student receives this honor. 2021

Side Li wins second place at the graduate level of the ACM SIGMOD 2021 Student Research Competition. This is the first time a UCSD student receives this honor. 2021

Advitya Gemawat awarded the HDSI UG Scholarship Project Award. This is the first

time this award goes to a DB Lab student. 2021

Kabir Nagrecha awarded the CSE UG Award for Excellence in Research. This is the first time this award goes to a DB Lab student. 2021

Kabir Nagrecha awarded both JSOE/CSE PhD and HDSI PhD Fellowships. 2021

Kabir Nagrecha receives an Honorable Mention for the CRA Outstanding UG Researcher Award. This is the first time a DB Lab student receives this honor. 2020

Kabir Nagrecha and *Advitya Gemawat* are both nominated to the CRA Outstanding Undergraduate Research Award by CSE and HDSI, respectively. 2020

Kabir Nagrecha receives an Honorable Mention for UCSD CSE BS Student Research Award. 2020

Advitya Gemawat awarded an HDSI Undergraduate Scholarship. 2019

Side Li awarded both JSOE/CSE PhD and HDSI PhD Fellowships. This is the first time a student receives both of these competitive PhD fellowships. 2019

Tara Mirmira awarded an HDSI PhD Fellowship. 2019

Vraj Shah is second runner-up at the ACM SIGMOD 2019 Student Research Competition. This is the first time this award goes to a UCSD student. 2019

Anthony Thomas receives Best Poster Award at UCSD JSOE Research Expo. 2019

Sothyrak Srey receive the UCSD CSE MS Student Contributions to Diversity Award upon my nomination. 2019

Digvijay Karamchandani receives an Honorable Mention for UCSD CSE MS Student Teaching Award upon my nomination. 2018

Lingjiao Chen awarded a Google PhD Fellowship. 2017–18

Lingjiao Chen is runner-up at SIGMOD Student Research Competition. 2017

THESIS COMMITTEE

Bhanu Gullapalli, PhD, HDSI, UCSD (Advisor: Tauhidur Rahman). 2024
TBD

Kabir Nagrecha, PhD, CSE, UCSD; Chair as Advisor (Co-advisor: Hao Zhang). 2024
"Orchestration Systems to Support Deep Learning at Scale"

Zeyi Wu, MS, CSE, UCSD (Advisor: Lily Weng). 2024
"Mitigating Non-Actionable Recourse"

Yuhao Zhang, PhD, CSE, UCSD; Chair as Advisor. 2023
"High-throughput Data Systems for Deep Learning Workloads"

Ryan Tran, MS, CSE, UCSD (Advisor: Julian McAuley). 2023
"Spoiler Recognition as Semantic Text Matching"

Pradyumna Sridhara, MS, CSE, UCSD; Chair as Advisor. 2023

"Cerebro - An Efficient, End-to-end Platform for Scalable Deep Learning"

Haochen Huang, PhD, CSE, UCSD (Advisor: Yuanyuan Zhou). 2022
"Improving the Dependability of Python-Based Database-Backed Web Applications"

Vraj Shah, PhD, CSE, UCSD; Chair as Advisor. 2022
"Simplifying Data Preparation for Machine Learning on Tabular Data"

Supun Nakandala, PhD, CSE, UCSD; Chair as Advisor. 2022
"Query Optimizations for Deep Learning Systems"

Liangde Li, MS, CSE, UCSD; Chair as Advisor. 2022
"Efficient Systems for Advanced Data Analytics"

Rana Alotaibi, PhD, CSE, UCSD (Advisor: Alin Deutsch). 2022
"Towards Scalable Self-Tuning Polystore Systems"

Benjamin Schumacher, PhD, Epidemiology and Public Health, UCSD (Advisor: Andrea LaCroix). 2022
"Developing, Validating, and Applying Measurements of Relative Intensity Activity from Observational Accelerometry Studies"

Ainur Smagulova, PhD, CSE, UCSD (Advisor: Alin Deutsch). 2021
"Vertex-centric Parallel Computation of SQL Queries"

Abdul Wasay, PhD, CS, Harvard University (Advisor: Stratos Idreos). 2021
"Computation-Cautious Machine Learning Systems"

Julaiti Alafate, PhD, CSE, UCSD (Advisor: Yoav Freund). 2020
"Parallel Boosting and Learning from Diverse Datasets"

Yunhui Guo, PhD, CSE, UCSD (Advisor: Tajana Rosing). 2020
"Efficient Learning across Multiple Domains with Deep Neural Networks"

Nikos Koulouris, PhD, CSE, UCSD (Advisor: Yannis Papakonstantinou). 2020
"Preventing Multiple Comparisons Problems in Data Exploration and Machine Learning"

David Justo, MS, CSE, UCSD (Co-advisor: Nadia Polikarpova). 2019
"Write once, rewrite everywhere: A Unified Framework for Factorized Machine Learning"

Chunbin Lin, PhD, CSE, UCSD (Advisor: Yannis Papakonstantinou). 2018
"Accelerating Analytic Queries on Compressed Data"

Nishant Agarwal, MS, CSE, UCSD (Advisor: Amarnath Gupta). 2017
"A Real-Time Temporal Clustering Algorithm for Short Text, and its Applications"

Sumedha Kattar, MS, CSE, UCSD (Advisor: Ilkay Altintas). 2017
"Finding the burnability index of a point on a map using the historical fire data"

RESEARCH EXAM COMMITTEE

Yutong Huang, PhD, CSE, UCSD (Advisor: Yiying Zhang). 2023
"Efforts in Memory Saving for DNN Training Acceleration"

Narangereit Batsoyol, PhD, CSE, UCSD (Advisor: Steve Swanson). 2023
"P-MASSIVE: a Real-time Search Engine for a Multi-terabyte Mass Spectrometry Database"

Alisha Ukani, PhD, CSE, UCSD (Advisor: Alex Snoeren). 2022
"Characterizing Online Tracking and its Mitigations"

Matthew Parker, MS, CSE, UCSD. 2021
"Investigating the LSM Tree and Applying it to Experimental Results of NVMM-Capable Devices"

Benjamin Schumacher, PhD, Epidemiology and Public Health, UCSD (Advisor: Andrea LaCroix). 2020
"The PAID Study: Physical Activity Intensity and Dementia"

Rana Alotaibi, PhD, CSE, UCSD (Advisor: Alin Deutsch). 2018
"Querying Heterogeneous Data Sources: A Comparative Study"

Nikos Koulouris, PhD, CSE, UCSD (Advisor: Yannis Papakonstantinou). 2018
"Controlling for False Discoveries in Data Exploration Systems"

SERVICE

Organization:
Program Co-Chair (Research Track), ACM IKDD CODS-COMAD 2024
Associate Editor, ACM SIGMOD 2024
Associate Editor, Scalable Data Science Category, VLDB 2022 & 2021 (Inaugural)
Co-Chair, Diversity and Inclusion, ACM SIGMOD 2021 (Inaugural)
(Inaugural) Core Committee member, D & I in DB Initiative (website), Jan–May 2021
(Inaugural) Lead Organizer, SoCal DB Day 2018 (website)
Co-Chair, ACM SIGMOD Workshop on Data Management for End-to-End Machine Learning (DEEM) 2018
(Inaugural) Organizing Committee, ACM SIGKDD Workshop on Common Model Infrastructure (CMI) 2018
Organizing Committee, Extremely Large Databases (XLDB) Conference 2018

Program Committee:
ACM SIGMOD: 2024, 2020, 2019, 2018, 2017
ACM IKDD CODS-COMAD: 2024 IEEE ICDE 2023 Special Track Senior PC
CIDR: 2023, 2022, 2021
ACM SIGMOD Workshop on Data Management for End-to-End ML (DEEM): 2023, 2022, 2021, 2020, 2019, 2017
VLDB: 2022, 2021, 2020, 2019, 2018
ACM SIGMOD Workshop on Human-in-the-loop Data Analytics (HILDA): 2022
MLSys / SysML: 2020, 2019
ACM SIGMOD 2017 Demonstrations; Student Research Competition
IEEE ICDE 2017
USENIX 2016 Workshop on Hot Topics in Cloud Computing (HotCloud)
ACM SIGMOD 2016 Undergraduate Research Poster Competition

Reviewer:
ACM Transactions on Database Systems (TODS) 2017, 2015
IEEE Transactions on Knowledge and Data Engineering (TKDE) 2014

External Reviewer:
ACM SIGMOD 2022, VLDB 2017, ACM SIGMOD 2013, IEEE ICDE 2013

IEEE INFOCOM 2010, IEEE GLOBECOM 2009, IEEE SECON 2009

Proposal Reviewer or Panelist:
Panelist: NSF CISE IIS 2021
Ad Hoc Reviewer: NSF SBIR/STTR Phase II 2020
Reviewer: DOE Solar Energy Technologies Office 2020
Panelist: NSF HDR Data Science Corps 2019

Other Research-Related:
Invited attendee at the Oct 2023 "Boston DB Meeting" at MIT; the youngest attendee of this edition of this prestigious once-in-5-years summit of renowned DB faculty and industry researchers
Co-chair of round table on "Systems for ML" at VLDB 2021
Co-chair of curated session on "Data Management for ML" at ACM SIGMOD 2021
Helped shape the new "Scalable Data Science" research paper category of VLDB 2021
Speaker at ACM SIGMOD 2018 New Researcher Symposium
Co-chair of "Best of ICDE 2017" Selection Committee for TKDE 2018
Judge for ACM SIGMOD 2017 Student Research Competition
Panelist at IEEE ICDE 2017 PhD Symposium
Judge for IEEE ICDE 2017 Demonstrations

Department/University Level:
2024 Jan–2024 Jun: CSE Faculty Recruiting Committee
2021 Oct–: HDSI Data Planet Lead; Fellowships Committee Chair (Inaugural)
2021 Oct–2022 Jun: HDSI Data Infra. & Systems Faculty Recruiting Committee Chair
2021–2023: CSE+HDSI Online Masters in Data Science Committee (Inaugural)
2021–2023: HDSI Industry Liaison Program (ILP) Committee
2019–Now: CSE Bylaws Committee
2017–2023: CSE MS Committee
2021 Mar: Finalized CSE's Departmental BPC plan and got it approved by NSF.
2021 Jan: Revised CSE's Departmental BPC plan and presented at faculty meeting.
2019–20: HDSI Faculty Recruiting Committee
2019–21: HDSI DEI Committee
2019 Dec: Co-founded and joined the HDSI DEI Committee
2019 Nov: Official representative of CSE at the NSF-sponsored Workshop on Departmental BPC Plans; co-authored CSE's Departmental BPC plan.
2019 Winter: Co-created a new CSE MS depth area for data science
2017–21: Active founding member of CSE Diversity, Equity, and Inclusion Committee
2017: UCSD SDSC Sustainability Committee
2016–17: CSE PhD Admissions Committee

Outreach and Contributions to Diversity, Equity, and Inclusion:
2024 Jan: Invited panelist and speaker at the D & I experience sharing session at ACM IKDD CODS-COMAD 2024.
2023 Nov: Invited speaker for SC 2023 Early Career Program Grant Writing Webinar; longer talk on strategies and examples of learning and growing from grant rejections.
2022 Oct: Invited panelist and speaker for SC 2022 Early Career panel discussion; spoke about how to cope with, and grow from, grant rejections as junior faculty.
2022 Jun: Invited panelist for ACM SIGMOD 2022 Diversity & Inclusion panel discussion on "Success and Impact Beyond Traditional Metrics."
2021 Jul–2022 Jun: Mentored 3 high school girls as part of UCSD MAP; all 3 went on to CS majors, 2 at UCSD CSE and 1 at UC Berkeley EECS.
2022 Mar: Talk on MAP and careers in CS and STEM at the UCSD MAP monthly meeting for high school students.

2020–21: Three UCSD students, all international students from Asia, trusted me as the first person they came out to; guided them on LGBTQ+ resources and community on campus and shared about my own coming out experience.
2021 Nov: Founding member of QICSE a group for LGBTQ+ graduate students, postdocs, faculty, and staff at UCSD CSE.
2021 Oct: Talk at oSTEM General Body Meeting on being an out academic in computing/STEM.
2021 Summer: Member, UCSD LGBTQIA+ Undergraduate Scholarships Committee
2021 Jun: Talk on mentoring programs and CS careers at the UCSD ABLE End of Year Celebration for high school girls interested in computing.
2021 May: Talk on MAP and careers in CS and STEM at the UCSD MAP Symposium for high school students and their parents.
2021 May: Blogged publicly about my experiences with rejections in academia to help reduce survivorship bias and impostor syndrome that are pervasive in academia: webpage. Added addendum on rejections to this CV.
2021 Apr: Short talk at CSE faculty meeting inclusion minutes on free speech and inclusivity (video).
2021 Apr: Authored an essay / poem to raise awareness on the abysmal state of inclusion in the CS field (webpage).
2021 Mar: Finalized CSE's Departmental BPC plan and got it approved by NSF.
2021 Feb–: Co-created and maintaining the website of the D & I in DB initiative.
2021 Jan–May: Inaugural Core Committee member of the D & I in DB Initiative for unifying Diversity & Inclusion efforts across database conference venues (website).
2021 Jan: Revised CSE's Departmental BPC plan and presented at faculty meeting.
2020 Nov: Panelist for the UCSD oSTEM Chapter discussion on Impostor Syndrome (webpage).
2020 Nov: Panelist for UCSD OIC Innovation at the Edge webinar on AI ethics and societal impact, especially on legally and societally marginalized groups (video).
2020 Oct: Panelist for the UCSD oSTEM Chapter Out in Engineering event; also spoke at their General Body Meeting.
2020 Jun–2021 Jun: (Inaugural) Co-chair for Diversity and Inclusion for ACM SIGMOD 2021.
2020 Jun: Gave talks on careers in data science to high school students in QI's Big Data Summer Camp.
2020 Feb: Short talk at CSE faculty meeting inclusion minutes on inclusive CS examples (webpage).
2019 Dec–2021 Jun: Co-founded the HDSI Diversity, Equity, and Inclusion Committee and became an official faculty member.
2019 Nov: Official representative of CSE at the NSF-sponsored Workshop on Departmental BPC Plans; co-authored CSE's Departmental BPC plan.
2019 Aug: Panelist for a discussion on impostor syndrome in CSE's SPIS summer program for freshmen.
2019 Aug: Gave a talk on careers and research in data science to freshmen in CSE's SPIS summer program.
2019 Jul: Gave talks on careers in data science to high school students in SDSC's REHS summer program and QI's Big Data Summer Camp.
2019 Apr: Organized a panel discussion on resources and community for LGBTQ+ people at UCSD on CSE Celebration of Diversity Day
2019 Mar: Organized a desk for CSE PhD Visit Day social events with pamphlets and swag collected from diversity-focused and other student resource centers at UCSD
2018 Nov: Represented CSE and UCSD at oSTEM annual conference as official sponsor with official desk presence
2018 Nov: Panelist for a Q & A event organized by oSTEM UCSD chapter for out LGBTQ+ students in STEM

2018 Nov: Hosted a research group open house for oSTEM UCSD chapter students
2018 Jun: Spoke and gave out certificates at the UCSD Rainbow Graduation
2018 Spring: Member, UCSD LGBTQIA+ Undergraduate Scholarships Committee
2017 Fall–2021 Spring: Active member of CSE Diversity, Equity, and Inclusion Committee
2017 Nov: Attended the annual conference of oSTEM representing CSE and UCSD
2017 Nov: Panelist for a Q & A event organized by oSTEM UCSD chapter for out LGBTQ+ students in STEM
2017 Nov: Hosted a research group open house for oSTEM UCSD chapter students
2017 Oct: Blogged publicly about my coming out experience: webpage
2017 Oct: Co-proposed new CSE PhD scholarship for contributions to diversity
2017 Apr: Spoke about my coming out experience in graduate school as a panelist at the IEEE ICDE 2017 PhD Symposium
2017 Apr: Part of the faculty group on diversity issues during CSE external review
2016 Fall–: Listed on the UCSD LGBT Resource Center "Out List" of faculty mentors for LGBTQ+ students

PANELS, INTERVIEWS, MEDIA, AND BLOGS

2024 Jan: Invited panelist and speaker at the D & I experience sharing session at ACM IKDD CODS-COMAD 2024.

2023 Jul: Featured by UCSD Today to celebrate out STEM faculty for San Diego Pride Week; webpage.

2022 Oct: Invited panelist and speaker for ACM-IEEE SC conference's Early Career Program on "Preparing Effective Grant Proposals"; spoke about how to grow from grant rejections: slide deck.

2022 Jun: Invited panelist for ACM SIGMOD 2022 Diversity & Inclusion panel discussion on "Success and Impact Beyond Traditional Metrics."

2021 Dec. Invited panelist for NeurIPS DBAI Workshop panel discussion on "Database and AI"; webpage.

2021 Dec. Interviewed by Stoodnt, a major admissions consulting firm in India, on advice and tips to international students interested in grad school in CS and Data Science: webpage.

2021 Dec. Interviewed by TWIML for a podcast on my research agenda and worldview, as well as projects Cerebro and SortingHat: webpage.

2021 Nov. ACM SIGMOD Blog post on the panel discussion I moderated at ACM SIGMOD 2021 on the status quo and challenges of data preparation in the context of AutoML; webpage

2021 Oct. Talk on being out in CS at the annual General Body Meeting of oSTEM UCSD Chapter; slide deck.

2021 Aug. Invited panelist for ACM KDD 2021 Deep Learning Day panel discussion on "Scaling and Debugging DL Systems."

2021 Aug. Invited panelist for VLDB 2021 panel discussion on "The Future of Data(base) Education"; article.

2021 May. Interviewed by Software Engineering Daily for a podcast on my research

agenda and worldview, as well as projects CEREBRO and SORTINGHAT: webpage

2021 Apr. Short talk at CSE faculty meeting inclusion minutes on free speech and inclusivity; video

2021 Apr. Essay / poem to raise awareness on the abysmal state of inclusion in the CS field; webpage

2021 Feb. Newsletter article by UCSD CNS on my group's research on scalable and optimized systems for deep learning; webpage

2021 Feb. Interview by UCSD oSTEM Chapter on my research and thoughts on the LGBTQ+ community in CS; webpage

2021 Feb. Website of the D & I in DB initiative for unifying Diversity & Inclusion efforts across database conference venues; webpage

2020 Nov. Invited panelist for panel discussion on Impostor Syndrome organized by UCSD oSTEM Chapter; webpage

2020 Nov. Invited panelist for panel discussion on AI ethics organized by UCSD Office of Innovation and Commercialization; video

2020 Feb. ACM SIGMOD Blog post on the new PVLDB research track category Scalable Data Science; webpage

2020 Feb. Interviewed by UCSD's Data Science Student Society on my thoughts on the ethical risks/benefits of AI and other emerging technologies; webpage

2020 Feb. Interviewed by UCSD's Data Science Student Society on my research, the data science field, and career advice; webpage

2020 Feb. Blog post on inclusive CS examples to raise awareness of the importance of inclusion and avoiding exclusionary examples in CS; webpage

2019 Jul. Talk on data science careers to high school students at SDSC's REHS summer workshop and QIś Big Data summer camp at UCSD; slide deck

2019 Jun. Article on ACM SIGARCH Blog about a SIGMOD 2019 research paper of mine; webpage

2019 Apr. Interviewed by Software Engineering Daily at Strata Data Conference 2019 for an article on systems for ML; webpage

2018 Aug. ACM SIGMOD Blog post on the panel discussion I moderated at ACM SIGMOD DEEM Workshop 2019 to raise awareness of the evolving role of the database research community in the ML systems/applications arena; webpage

2018 Apr. Blog post on the culture wars of the database/data management community (c. early 2018) to raise awareness of the inherent intellectual multiculturalism of the research area; webpage

2018 Feb. Interviewed by ACM SIGMOD 2018 WebDB Workshop for an ACM SIGMOD Blog article in 2018 on the intersection of ML and data systems; webpage

2017 Oct. Blog post on my coming out experience from my sociocultural and individual vantage point; webpage

EXTRAMURAL FUNDING

Grants:

NIH R01 grant titled "Leveraging Deep Learning to Classify Sitting Posture and Measure Sedentary Patterns from Accelerometer Data in Diverse Cohorts"
Co-PI. Lead PI: Loki Natarajan, UCSD. 2024–28

NSF CISE-IIS Smart and Connected Health grant titled "MS-ADAPT: Multi-Sensor Adaptive Data Analytics for Physical Therapy"
Co-PI. Lead-PI: Emilia Farcas, UCSD. 2022–26

NIH PAR19-070 Alzheimers grant titled "Identifying Digital Phenotypes of Risk for Alzheimer's Disease and Related Dementias Among Hispanics/Latinos"
Co-PI. Lead PI: Maria Marquine, UCSD. 2021–26

NSF OIA Convergence Accelerator Track D Phase I grant titled "Towards Intelligent Sharing and Search for AI Models and Datasets"
Co-PI. Lead PI: Jingbo Shang, UCSD. 2020–21

NSF CAREER grant titled "Multi-Query Optimizations for Deep Learning Systems"
Sole PI. 2020–25

NSF CISE-IIS-III Small grant titled "Towards Cross-Model Query Optimizations for Multi-model Heterogeneous Data Analytics"
Co-PI. Lead PI: Amarnath Gupta, UCSD. 2019–22

NIH R01 grant titled "Diet and Physical Activity Assessment Methodology.'
Co-PI. Lead PI: Loki Natarajan, UCSD. 2018–22

NSF CISE-IIS-III Small grant titled "Towards Speech-Driven Multimodal Querying"
Lead PI. 2018–22 (with NCE)

Gifts:

Amazon Research Award. Sole PI.	2021
VMWare Early Career Faculty Grant. Sole PI.	2021
VMWare Early Career Faculty Grant. Sole PI.	2020
Google Faculty Research Award. Sole PI.	2020
Oracle Labs Research Award. Sole PI.	2019
Hellman Fellowship. Sole PI.	2018
Opera Solutions Faculty Research Award. Sole PI.	2017
NVIDIA GPU Grant. Sole PI.	2017
Google Faculty Research Award. Sole PI.	2017

TALKS

The New DBfication of ML/AI

Google India (Invited)	Jan 2024
UC Irvine CS Seminar (Invited)	May 2023
NYU Responsible AI Seminar	Mar 2023
ACM CODS-COMAD (Invited keynote)	Jan 2023
IIT Madras, India (Invited)	Dec 2022
Cornell Database Seminar (Invited)	Nov 2022
TU Berlin BIFOLD Colloquium, Germany (Invited)	Aug 2022

The CCF Advanced Disciplines Lectures, China (Invited keynote) Jun 2022
Amazon Web Services (Invited) Mar 2022
NeurIPS DBAI Workshop (Invited) Dec 2021
UCSD DSE Program Guest Speaker Series (Invited) Oct 2021
Google Ads "Let's Talk Tech" (Invited) Oct 2021
Brazilian Symposium on Databases (Invited keynote) Oct 2021
Megagon Labs (Invited) Oct 2021
San Diego ML Meetup-HDSI Data Science Speaker Series (Invited) Oct 2021
University of Melbourne (Invited) Sep 2021
ACM SIGKDD ADS Talk Series (Invited) Aug 2021
ACM SIGMOD (Invited keynote at a curated session) Jun 2021
IEEE ICDE (TCDE Rising Star Award acceptance talk) Apr 2021

How to Transform the Sting of Grant Rejections into Fuel for Growth
ACM-IEEE SC Early Career Program Grant Writing Webinar (Invited) Nov 2023
ACM-IEEE SC Early Career Program (Invited) Oct 2022

Cerebro: A Layered Data Platform for Scalable Deep Learning
Zalando, Germany (Invited) Sep 2022
GuideRails Developer+ Workshop (Invited) May 2022
UCSD HDSI Faculty Seminar Apr 2021
Data Science for Materials Discovery I-AIM Seminar, Online (Invited) Mar 2021
CIDR 2021, Online; Short talk; Video Jan 2021

Factorized Machine Learning: Paths and Roadblocks
Factorized Databases Workshop (Invited) Aug 2022

Some Damaging Delusions of Deep Learning Practice (and How to Avoid Them)
ACM SIGKDD Deep Learning Day; Video Aug 2021

Multi-Query Optimizations for Deep Learning Systems
Dutch Seminar on Data Systems Design (Invited) Jul 2021
HPI + TU Darmstadt (Invited) Feb 2021
UC Berkeley DSF Seminar (Invited) Oct 2020
VMware, Online (Invited) Sep 2020
Microsoft Jim Gray Systems Lab (Invited) Jun 2020

Data and ML for Data Prep for ML: The ML Data Prep Zoo
Amazon Web Services AI and Redshift (Invited) May 2021
Lyon 1 University + CNRS France (Invited) Apr 2021
Google BigQuery ML and Cloud AutoML (Invited) Jan 2021
UW-Madison and Microsoft MLOS Seminar (Invited); Video Oct 2020

On Mentoring Programs and CS Careers
UCSD ABLE End of Year Celebration for high school girls interested in computing careers Jun 2021
UCSD MAP Symposium for high school students and their parents (Under a different title) May 2021

Systems for Data Science: From Research to Practice and Back
UCSD HDSI + Rady Data Science Day (Invited) May 2021

Apache MADlib: Scalable In-RDBMS Machine Learning
Microsoft Azure Data (Invited) Jul 2020

Democratizing Machine Learning-based Data Analytics
USC-MHI Cyber-Physical Systems Seminar (Invited) Sep 2019
UCSD HDSI Faculty Seminar Apr 2019
UCSD CSE Faculty Research Seminar Oct 2018

Data Science: Applications, Careers, and Research
UCSD CSE Summer Program for Incoming Students Aug 2019

Building a Successful Career in Data Science
UCSD SDSC REHS Workshop for high school students Jul 2019
UCSD Qualcomm Institute Big Data Summer Camp Jul 2019

Faster ML over Joins of Tables
O'Reilly Strata Data Conference, San Francisco Mar 2019

Multi-Query Optimizations for ML Systems
University of Washington, Seattle (Invited) Nov 2018
Microsoft Research, Redmond Nov 2018
Microsoft Cloud and Information Services Lab, Mountain View Sep 2018
Google Brain/TFX and DIA, Mountain View Sep 2018

Towards Optimized Distributed ML Systems
UCSD Center for Networked Systems Research Review Oct 2018

Advice from PhD to Early Career
ACM SIGMOD New Researcher Symposium Oct 2018

Morpheus: Factorized Linear Algebra for Scalable Advanced Analytics
Google Data Infrastructure and Analytics, Irvine Aug 2017

Accelerating Model Selection in Advanced Analytics
Teradata, San Diego (Invited) Nov 2017
Opera Solutions Technical Conference, San Diego (Invited) Oct 2017
University of Michigan, Ann Arbor (Invited) Sep 2017

Towards Linear Algebra over Normalized Data
VLDB Aug 2017

Accelerating Advanced Analytics on Multi-table Data
Amazon Machine Learning, Berlin (Invited) Aug 2017

Democratizing Advanced Analytics Beyond Just Plumbing
ACM SIGMOD DEEM Workshop (Invited Academic Keynote) May 2017

Democratizing Feature Engineering and Model Selection in Advanced Analytics
Opera Solutions, San Diego (Invited) May 2017

Democratizing Distributed Advanced Analytics
UCSD Center for Networked Systems Lecture Apr 2017

CEREBRO*: A System to Manage Deep Learning for Relational Data Analytics*
CIDR "Gong Show" Jan 2017

Accelerating Advanced Analytics

Google, Mountain View (Invited) Dec 2016

The Data Strikes Back! Research Challenges in Advanced Analytics
UCSD AI Seminar Oct 2016

Exploiting Database Dependencies to Accelerate Advanced Analytics
UCSD Database Seminar Oct 2016

Model-based Pricing of Relational Data in the Cloud
UCSD Database Seminar Oct 2016

Accelerating Advanced Analytics (Invited) Jan-Mar 2016
New York University
Microsoft Research, Redmond, WA
University of Illinois at Urbana-Champaign
Cornell University
University of California, San Diego; Video
University of Chicago
IBM Research Almaden, CA (under a different title)
University of Maryland, College Park
LogicBlox, Atlanta, GA
Georgia Institute of Technology
Purdue University (under a different title)

Machine Learning over Joins of Multiple Tables
Wisconsin Institutes of Discovery Seminar 2015

Learning Generalized Linear Models over Normalized Data
ACM SIGMOD 2015

Stop that Join! Optimizing Feature Selection over Normalized Data for Naive Bayes
Wisconsin Database Group Seminar 2015

On Learning Generalized Linear Models over Joins
Wisconsin Database Group Seminar 2014

Usability and Developability Challenges in Advanced Analytics
Indian Institute of Technology, Madras (Invited) 2014

On Learning over Joins
Microsoft Big Data Security Symposium (Invited) 2014
Microsoft Jim Gray Systems Lab 2014

On Integrating Advanced Analytics with Scalable Structured Data Management
Wisconsin CS Preliminary Exam 2014

Scalable and Distributed PCA on REEF
Microsoft Cloud and Information Systems Lab 2013

Commoditizing Large-Scale Analytics for the Enterprise around R
Microsoft Jim Gray Systems Lab (Invited) 2013

Columbus: Feature Selection on Data Analytics Systems
Wisconsin Database Group Seminar 2013

Brainwash: A Data System for Feature Engineering
CIDR 2013

Probabilistic Management of OCR Data Using an RDBMS
VLDB 2012
Wisconsin Database Group Seminar 2012

Large-Scale Low-Rank Matrix Factorization using Incremental Gradient Descent
Oracle Labs 2012

Towards a Unified Architecture for in-RDBMS Analytics
ACM SIGMOD 2012

Staccato: Probabilistic Management of OCR Data Using an RDBMS
Wisconsin DB Affiliates Meeting 2011

Scalable Cross-validation and Ensemble Learning in SystemML
IBM Almaden Research Center 2011

Managing Uncertainty in OCR and Speech Data Using an RDBMS
Microsoft Jim Gray Systems Lab 2011

INDUSTRY EXPERIENCE

Consultant, Luth Research LLC Jan–Apr 2019
Data engineering, systems for data analytics and data science

Research Intern, Microsoft Jim Gray Systems Lab Jun–Aug 2014
Manager: Alan Halverson
Performed a theoretical and empirical analysis of the effects of a classical database dependency on the accuracy and performance of some ML algorithms.

Research Intern, Microsoft Cloud and Information Services Lab Jun–Aug 2013
Managers: Markus Weimer and Vijay Narayanan
Built a tool for large-scale principal component analysis on the distributed computation platform REEF. My work was subsequently used by teams inside Microsoft.

Research Assistant, Oracle Labs/Advanced Analytics Jun–Aug 2012
Manager: Vaishnavi Sashikanth
Built a tool for large-scale matrix factorization on MapReduce/Hadoop. My work was incorporated into the product Oracle R for Enterprise. Invited article on Oracle's product blog: webpage.

Research Intern, IBM Research Almaden Jun–Aug 2011
Managers: Berthold Reinwald and Rajasekar Krishnamurthy
Incorporated scalable ensemble learning and other meta-learning techniques into SystemML.My work was incorporated into the product IBM BigInsights.

Intern, NetApp India May–Jul 2008
Worked on investigating and prototyping a new Web Services framework for platform- and programming language-independent communication with NetApp storage systems. Ideas from this work were incorporated into the SDK for NetApp filers.

TECHNICAL SKILLS

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

Data Platforms: PySpark, PostgreSQL, Greenplum, Hadoop, Hive

ML Tools: Keras, TensorFlow, PyTorch, Scikit-learn

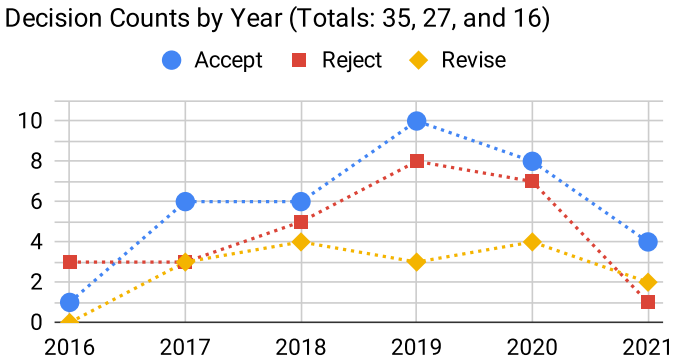
Cloud/Cluster Tools: AWS EC2, S3, and EBS; Kubernetes; Dask; Ray

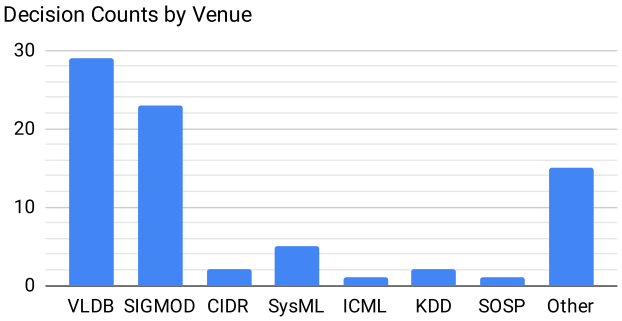
ADDENDUM: REJECTIONS

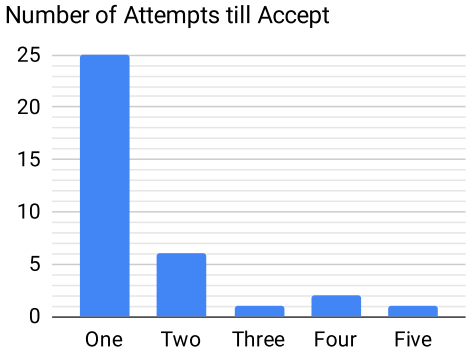
I believe that acceptances and rejections are two sides of the same coin of intellectual progress and that reporting only on the former in an academic CV may perpetuate survivorship bias and/or impostor syndrome in academia. Thus, I share these summary statistics on the rejections I have faced throughout my faculty career.

For simplicity sake, "paper" below includes all forms of peer-reviewed publications listed in the preceding pages. An (eventually) accepted paper or awarded proposal is counted more than once across these statistics if it was not accepted/awarded directly in the first attempt. Furthermore, all but one of my accepted VLDB and SIGMOD papers went through a revision decision. Last update: Summer 2021.

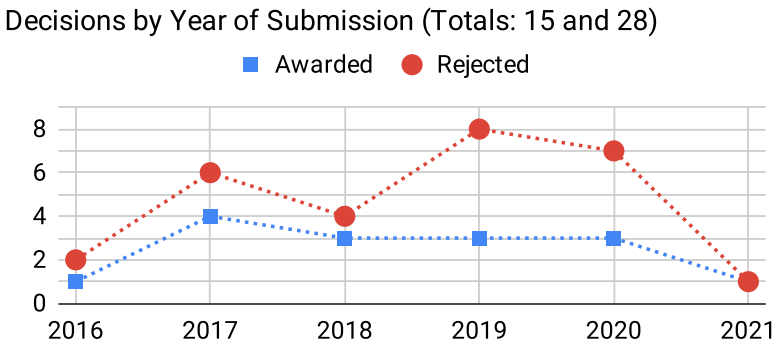
SUMMARY STATISTICS ON PAPER DECISIONS

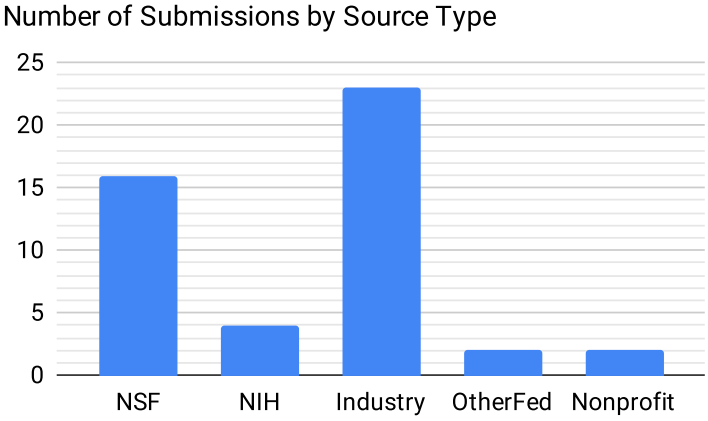






SUMMARY STATISTICS ON PROPOSAL (GRANT/GIFT) DECISIONS





REJECTIONS ON MAJOR COMPETITIVE HONORS OR AWARDS

2021:
Nonprofit: Sloan Research Fellowship
Industry: Google Research Scholar Award, Microsoft Research Faculty Fellowship, VMware Systems Research Award

2020:
Industry: Amazon Research Award, Sony Focused Research Award

2019:
Federal: NSF CAREER Award
Industry: Bloomberg Data Science Research Grant, Google Faculty Research Award, Microsoft Investigator Fellowship, Microsoft Research Faculty Fellowship, NetApp Faculty Fellowship

2018:
Federal: NSF CRII Award
Industry: Amazon Research Award, Bloomberg Data Science Research Grant, Facebook Research Award

2017:
Federal: NSF CAREER Award