Arun Kumar

3218 EBU3B (CSE building)
9500 Gilman Drive, Mail Code 0404
La Jolla, CA 92093

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

EMPLOYMENT

University of California, San Diego
Assistant Professor
Department of Computer Science and Engineering — From 2016
Halıcıoğlu Data Science Institute — From 2019

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

Databases and data management/systems for machine learning/artificial intelligence-based data analytics, with a focus on problems related to usability, developability, performance, and scalability. I enjoy working on problems that are motivated by real applications and are formally grounded. My work spans the whole gamut of building systems, algorithm design, theoretical analysis, empirical analysis, and working with practitioners (data scientists and ML/software engineers) to deploy my research.
Research Webpage: https://adalabucsd.github.io/

AWARDS AND HONORS

NSF CAREER Award	2020
VMware Early Career Faculty Grant	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 by the Indian gov.	2003–08

MAJOR ONGOING PROJECTS

Project Triptych: — From 2016
The goal of this umbrella project is to build end-to-end *model selection management systems* to simplify and accelerate the processes of preparing data/features and building ML models. We draw upon the classical lessons of declarative specification and automated query optimization from the database systems world to lay a principled foundation for the design and implementation of next generation ML systems, including new AutoML frameworks. We exploit the semantics of the data and the ML task to reduce grunt work for users and remove various performance and scalability

bottlenecks, which in turn reduces costs and helps democratize ML analytics. This project improves how both classical statistical models and deep learning models are built. It has the following active components: CEREBRO for optimized deep learning model selection, MORPHEUS to optimize statistical ML model building, and SORTINGHAT on benchmarking ML data preparation, and others on AutoML frameworks. Past components inlude HAMLET to exploit database schema in ML analytics, SLAB to benchmark linear algebra-based ML systems, and NIMBUS to price ML models in data marketplaces.
Project Webpage: `https://adalabucsd.github.io/triptych.html`

Project Genisys: From 2017
The goal of this umbrella project is make it easier to deploy deep learning models to see, hear, and understand unstructured data and query sources such as speech, images, video, time series, and text–a vision for ML systems I call *database perception*. Once again, we draw upon the classical lessons of higher-level query specification and automated query optimization to enable seamless type-agnostic data analytics systems. It has the following active components: SPEAKQL to enable data systems to hear speech-based queries and data, VISTA and KRYPTON to see image data for inference optimizations, and PANORAMA to see video data for practical querying.
Project Webpage: `https://adalabucsd.github.io/genisys.html`

SELECTED RESEARCH IMPACT

Ideas from project CEREBRO integrated into MADlib and Greenplum in collaboration with Pivotal/VMWare. 2019–20
Models/ideas from project SORTINGHAT explored for integration with TensorFlow Data Validation in collaboration with Google 2019–20
CEREBRO being used by UCSD public health researchers 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 UWashington 2013
Code/ideas from project BISMARCK shipped as part of analytics products by Oracle, EMC, 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, etc.): 21
Other peer-reviewed conference and journal papers: 9
Peer-reviewed workshop and demonstration papers: 12
Papers under submission: 3
Number of citations: 1568 and h-index: 15 (as per Google Scholar in August 2020)
Full list of publications: `https://adalabucsd.github.io/publications.html`

CONFERENCE PUBLICATIONS

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

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

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

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 Highlight 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 R, 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

BOOKS AND JOURNAL PUBLICATIONS

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

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

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

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 R, 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

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

Towards A Polyglot Framework for Factorized ML
David Justo, Lukas Stadler, Nadia Polikarpova, and Arun Kumar
Under submission

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

Application of Convolutional Neural Network Algorithms for Advancing Sedentary and Activity Bout Classification
Supun Nakandala et al. (about 11 authors)
Under submission

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 (Go to 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

TEACHING

CSE 291/234: Data Systems for Machine Learning. UCSD.	Fall 2020
CSE 132C: Database System Implementation. UCSD.	Spring 2020
DSC 102: Systems for Scalable Analytics. UCSD.	Winter 2020
CSE 232A: Graduate Database Systems. UCSD.	Fall 2019
CSE 190D: Topics in Database System Implementation. UCSD.	Spring 2019
CSE 291F: Advanced Data Analytics and ML Systems. UCSD.	Winter 2019
CSE 232A: Graduate Database Systems. UCSD.	Fall 2018
CSE 290D: Seminar on Integrative AI Engineering. UCSD.	Fall 2018
CSE 190A: Topics in Database System Implementation. UCSD.	Spring 2018
CSE 291A: Advanced Data Analytics and ML Systems. UCSD.	Winter 2018
CSE 290A: Seminar on Advanced Data Science. UCSD.	Fall 2018
CSE 190D: Topics in Database System Implementation. UCSD.	Spring 2017
CSE 290B: Seminar on Advanced Data Science. UCSD.	Spring 2017
CSE 291G: Topics in Advanced Analytics. UCSD.	Winter 2017
CS 564: DBMS: Design and Implementation. UW-Madison.	Fall 2015

ADVISING (CURRENT)

Side Li, PhD, CSE, UCSD.	Fall 2019–
Tara Mirmira, PhD, CSE, UCSD.	Fall 2019–
Supun Nakandala, PhD, CSE, UCSD.	Fall 2017–
Vraj Shah, MS & PhD, CSE, UCSD.	Fall 2016–
Yuhao Zhang, MS & PhD, CSE, UCSD.	Fall 2018–
Advitya Gemawat, BS, HDSI, UCSD.	Winter 2019–
Kabir Nagrecha, BS, CSE, UCSD.	Fall 2019–

ADVISING (ALUMNI)

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

STUDENT AWARDS

Kabir Nagrecha received an Honorable Mention for UCSD CSE BS Student Research Award. 2020
Advitya Gemawat awarded an HDSI Undergraduate Scholarship. 2019
Side Li awarded a JSOE PhD and an HDSI PhD Fellowships. 2019
Tara Mirmira awarded an HDSI PhD Fellowship. 2019
Vraj Shah is second runner-up at SIGMOD Student Research Competition. 2019
Anthony Thomas received Best Poster Award at UCSD JSOE Research Expo. 2019
Sothyrak Srey received the UCSD CSE MS Student Contributions to Diversity Award upon my nomination. 2019
Digvijay Karamchandani received 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

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

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

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

SERVICE

Organization:
Associate Editor, Scalable Data Science Category, VLDB 2021
Co-Chair, Diversity and Inclusion, ACM SIGMOD 2021
Lead Organizer, SoCal DB Day 2018
Co-Chair, ACM SIGMOD Workshop on Data Management for End-to-End Machine Learning (DEEM) 2018

Organizing Committee, ACM SIGKDD Workshop on Common Model Infrastructure (CMI) 2018
Organizing Committee, Extremely Large Databases (XLDB) Conference 2018

Program Committee:
CIDR: 2021
ACM SIGMOD: 2020, 2019, 2018, 2017
VLDB: 2020, 2019, 2018
ACM SIGMOD Workshop on Data Management for End-to-End ML (DEEM): 2020, 2019, 2017
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:
VLDB 2017, ACM SIGMOD 2013, IEEE ICDE 2013
IEEE INFOCOM 2010, IEEE GLOBECOM 2009, IEEE SECON 2009

Proposal Reviewer or Panelist:
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:
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

Outreach and Contributions to Diversity, Equity, and Inclusion:
June 2020–: Co-chair for Diversity and Inclusion for ACM SIGMOD 2021.
June 2020: Gave talks on careers in data science to high school students in QI's Big Data Summer Camp.
December 2019–: Co-founded the HDSI Diversity, Equity, and Inclusion Committee and became a member.
November 2019: Official representative of CSE at the NSF-sponsored Workshop on Departmental BPC Plans; co-authored CSE's Departmental BPC plan.
August 2019: Panelist for a discussion on impostor syndrome in CSE's SPIS summer program for freshmen.
August 2019: Gave a talk on careers and research in data science to freshmen in CSE's SPIS summer program.
July 2019: Gave talks on careers in data science to high school students in SDSC's REHS summer program and QI's Big Data Summer Camp.
Apr 2019: Organized a panel discussion on resources and community for LGBTQ+ people at UCSD on CSE Celebration of Diversity Day

Mar 2019: 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
Nov 2018: Represented CSE and UCSD at oSTEM annual conference as official sponsor with official desk presence
Nov 2018: Panelist for a Q & A event organized by oSTEM UCSD chapter for out LGBTQ+ students in STEM
Nov 2018: Hosted a research group open house for oSTEM UCSD chapter students
Jun 2018: Spoke and gave out certificates at the UCSD Rainbow Graduation
Spring 2018: Member, UCSD LGBTQIA+ Undergraduate Scholarships Committee
Fall 2017–: Active member of CSE Diversity, Equity, and Inclusion Committee
Nov 2017: Attended the annual conference of oSTEM representing CSE and UCSD
Nov 2017: Panelist for a Q & A event organized by oSTEM UCSD chapter for out LGBTQ+ students in STEM
Nov 2017: Hosted a research group open house for oSTEM UCSD chapter students
Oct 2017: Blogged publicly about my coming out experience. Go to webpage
Oct 2017: Co-proposed new CSE PhD scholarship for contributions to diversity
Apr 2017: Spoke about my coming out experience in graduate school as a panelist at the IEEE ICDE 2017 PhD Symposium
Apr 2017: Part of the faculty group on diversity issues during CSE external review
Fall 2016–: Listed on the UCSD LGBT Resource Center "Out List" of faculty mentors for LGBTQ+ students

Department/University Level:
2019–20: HDSI Faculty Recruiting Committee
2019–20: CSE Bylaws Committee
2017–20: CSE MS Committee
December 2019–: Co-founded the HDSI Diversity, Equity, and Inclusion Committee and became a member.
November 2019: Official representative of CSE at the NSF-sponsored Workshop on Departmental BPC Plans; co-authored CSE's Departmental BPC plan.
Winter 2019: Co-created a new CSE MS depth area for data science
2017–: Active founding member of CSE Diversity, Equity, and Inclusion Committee
2017: UCSD SDSC Sustainability Committee
2016–17: CSE PhD Admissions Committee

INTERVIEWS AND MEDIA

Interviewed by UCSD's Data Science Student Society in 2020 on my research, the data science field, and career advice: Go to webpage

Interviewed by Software Engineering Daily at Strata Data Conference 2019 for an article on systems for ML: Go to webpage

Article on ACM SIGARCH Blog in 2019 about a SIGMOD 2019 research paper of mine: Go to webpage

Interviewed by ACM SIGMOD 2018 WebDB Workshop for an ACM SIGMOD Blog article in 2018 on the intersection of ML and data systems: Go to webpage

EXTRAMURAL FUNDING

Grants:

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

National Science Foundation 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

National Institutes of Health grant titled "Diet and Physical Activity Assessment Methodology." Co-PI (Lead PI: Loki Natarajan, UCSD). 2018–22

National Science Foundation CISE-IIS-III Small grant titled "Towards Speech-Driven Multimodal Querying." Lead PI. 2018–21

Gifts:

VMWare Early Career Faculty Grant. Sole PI. 2020
Google Faculty Research Award. Sole PI. 2019–20
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. 2016–17

TALKS

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

Multi-Query Optimization for Deep Learning Systems
Microsoft Jim Gray Systems Lab, Online (Invited) June 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 August 2019

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

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

Multi-Query Optimization 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: `https://goo.gl/raJFpu`)
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

TECHNICAL SKILLS

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

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

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