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
Department of Computer Science and Engineering
Assistant Professor — 2016–Now

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, 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 software engineers) to deploy my research.
Research Webpage: https://adalabucsd.github.io/

SELECTED HONORS

Hellman Fellowship	2018
Faculty of the Year from UCSD oSTEM Chapter	2018
ACM SIGMOD Distinguished PC Member	2017
Google Faculty Research Award	2017
Invited Keynote at ACM SIGMOD DEEM Workshop	2017
UW-Madison CS Graduate Student Research Award for best PhD research	2016
Invited Paper at ACM Transactions on Database Systems	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 — Started 2016

The goal of this umbrella project is to build a comprehensive end-to-end *model selection management system* to simplify and accelerate the processes of preparing data/features and building ML models. We draw upon the classical lessons of declarative specification, automated query optimization, and provenance management 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 will 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 has implications for how both classical statistical models and deep learning models are built. This project includes several component projects, including Morpheus to study the interplay between relational and linear algebra, Hamlet to exploit database schema in ML analytics, SLAB to benchmark linear algebra-based ML systems, and others on ML schema

extraction, declarative AutoML, and pricing training data.
Project Webpage: `https://adalabucsd.github.io/triptych.html`

Project Genisys Started 2017
The goal of this umbrella project is make it easier to deploy ML models, especially 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-powered data systems we called *database perception*. Once again, we draw upon the classical lessons of declarative specification and automated query optimization to enable seamless type-agnostic analytics in existing data systems environments. This project also includes several component projects, including SpeakQL to enable data systems to hear speech-based queries and data, Vista to enable data systems to see image data for multimodal analytics, and others on declarative querying of video data and accelerating the explanations of the behavior of deep nets.
Project Webpage: `https://adalabucsd.github.io/genisys.html`

SELECTED RESEARCH IMPACT

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 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
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/news.html`

PUBLICATIONS SUMMARY

Full papers at SIGMOD/VLDB (the top conferences on data systems): 14
Other peer-reviewed conference and journal papers: 6
Peer-reviewed workshop and demonstration papers: 6
Full papers under submission: 4
Number of citations: 1016 and h-index: 13 (as per Google Scholar in Jan 2019)
Full list of publications: `https://adalabucsd.github.io/publications.html`

CONFERENCE PUBLICATIONS

SpeakQL: Towards Speech-driven Multimodal Querying of Structured Data
V. Shah, S. Li, A. Kumar, and L. Saul
Under submission

Materialization Trade-offs for Feature Transfer from Deep CNNs for Multimodal Data Analytics
S. Nakandala and A. Kumar
Under submission

Accelerating Deep CNN Explanations with Incremental and Approximate Inference
S. Nakandala, A. Kumar, and Y. Papakonstantinou
Under submission

Hierarchical and Distributed Machine Learning Inference Beyond the Edge
A. Thomas, Y. Guo, Y. Kim, B. Aksanli, A. Kumar, and T. S. Rosing
Under submission

Enabling and Optimizing Feature Interactions over Normalized Data
S. Li, L. Chen, and A. Kumar
ACM SIGMOD 2019 (To appear)

Model-based Pricing for Machine Learning in a Data Marketplace
L. Chen, P. Koutris, and A. Kumar
ACM SIGMOD 2019 (To appear)

Tuple-Oriented Compression for Large-scale Mini-Batch Gradient Descent
F. Li, L. Chen, Y. Zeng, A. Kumar, J. Naughton, J. M. Patel, and X. Wu
ACM SIGMOD 2019 (To appear)

A Comparative Evaluation of Systems for Scalable Linear Algebra-based Analytics
A. Thomas and A. Kumar
VLDB 2018/2019 (To appear)

In-RDBMS Hardware Acceleration of Advanced Analytics
D. Mahajan, J. K. Kim, J. Sacks, A. Ardalan, A. Kumar, and H. Esmaeilzadeh
VLDB 2018

Are Key-Foreign Key Joins Safe to Avoid when Learning High Capacity Classifiers?
V. Shah, A. Kumar, and X. Zhu
VLDB 2018

Towards Linear Algebra over Normalized Data
L. Chen, A. Kumar, J. Naughton, and J. M. Patel
VLDB 2017

Bolt-on Differential Privacy for Scalable Stochastic Gradient Descent-based Analytics
X. Wu, F. Li, A. Kumar, K. Chaudhuri, S. Jha, and J. Naughton
ACM SIGMOD 2017

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

To Join or Not to Join? Thinking Twice about Joins before Feature Selection
A. Kumar, J. Naughton, J. M. Patel, and X. Zhu
ACM SIGMOD 2016

Learning Generalized Linear Models Over Normalized Data
A. Kumar, J. Naughton, and J. M. Patel
ACM SIGMOD 2015

Materialization Optimizations for Feature Selection Workloads
C. Zhang, A. Kumar, and C. Ré
ACM SIGMOD 2014 (**Best Paper Award**; **Invited to ACM TODS 2016**)

Brainwash: A Data System for Feature Engineering
M. Anderson, D. Antenucci, V. Bittorf, M. Burgess, M. Cafarella, A. Kumar, F. Niu, Y. Park, C. Ré, and C. Zhang
CIDR 2013 (Vision paper)

Probabilistic Management of OCR Data Using an RDBMS
A. Kumar, and C. Ré
VLDB 2012

The MADlib Analytics Library: Or MAD Skills, the SQL
J. Hellerstein, C. Ré, F. Schoppmann, D. Wang, E. Fratkin, A. Gorajek, K. Ng, C. Welton, X. Feng, K. Li, and A. Kumar
VLDB 2012 (Industrial track)

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

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

JOURNAL PUBLICATIONS

Materialization Optimizations for Feature Selection Workloads
C. Zhang, A. Kumar, and C. Ré
ACM TODS 2016 (**Invited paper**)

Model Selection Management Systems: The Next Frontier of Advanced Analytics
A. Kumar, R. McCann, J. Naughton, and J. M. Patel
ACM SIGMOD Record Dec 2015 (Vision paper)

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

WORKSHOPS, DEMOS, AND OTHER PEER-REVIEWED PUBLICATIONS

Demonstration of SpeakQL: Speech-driven Multimodal Querying of Structured Data
V. Shah, S. Li, K. Yang, A. Kumar, and L. Saul
Under submission (Demo)

Demonstration of Nimbus: Model-based Pricing for Machine Learning in a Data Marketplace
L. Chen, H. Wang, L. Chen, P. Koutris, and A. Kumar
Under submission (Demo)

Model-based Pricing: Do Not Pay for More than What You Learn!
L. Chen, P. Koutris, and A. Kumar
ACM SIGMOD 2017 DEEM Workshop

SpeakQL: Towards Speech-driven Multi-modal Querying
D. Chandarana, V. Shah, A. Kumar, and L. Saul
ACM SIGMOD 2017 HILDA Workshop

Demonstration of Santoku: Optimizing Machine Learning over Normalized Data
A. Kumar, M. Jalal, B. Yan, J. Naughton, and J. M. Patel
VLDB 2015 (Demo)

Hazy: Making it Easier to Build and Maintain Big-data Analytics
A. Kumar, F. Niu, and C. Ré
ACM Queue 2013 (**Invited to the Communications of the ACM**)

Distributed and Scalable PCA in the Cloud
A. Kumar, N. Karampatziakis, P. Mineiro, M. Weimer, and V. Narayanan
NIPS BigLearn Workshop 2013

Feature Selection in Enterprise Analytics: A Demonstration using an R-based Data Analytics System
P. Konda, A. Kumar, C. Ré, and V. Sashikanth
VLDB 2013 (Demo)

Flexible Multimedia Content Retrieval Using InfoNames
A. Kumar, A. Anand, A. Balachandran, V. Sekar, A. Akella, S. Seshan
ACM SIGCOMM 2010 (Demo)

TECHNICAL REPORTS, MANUSCRIPTS, AND ARTICLES

ML/AI Systems and Applications: Is the SIGMOD/VLDB Community Losing Relevance?
A. Kumar
Article on ACM SIGMOD Blog (`http://wp.sigmod.org/?p=2454`), 2018

Courting ML: Witnessing the Marriage of Relationa & Web Data Systems to Machine Learning
Interviewed for ACM SIGMOD Blog (`http://wp.sigmod.org/?p=2243`), 2018

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

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

InfoNames: An Information-Based Naming Scheme for Multimedia Content
A. Kumar, A. Anand, A. Balachandran, V. Sekar, A. Akella, S. Seshan
UW-Madison CS Technical Report TR 1677, 2010

TEACHING

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)

Supun Nakandala, PhD at UCSD.	Fall 2017–
Vraj Shah, MS & PhD at UCSD.	Fall 2016–
Yuhao Zhang, MS at UCSD.	Fall 2018–
Kevin Yang, BS at UCSD.	Fall 2018–

ADVISING (ALUMNI)

Side Li, BS at UCSD.	Fall 2017–Spring 2018
Anthony Thomas, MS at UCSD.	Spring 2017–Spring 2018
Mingyang Wang, MS at UCSD.	Spring 2017
Lingjiao Chen, MS at UW-Madison.	Fall 2015–Fall 2018
Fengan Li, BS at UW-Madison.	Fall 2015–Spring 2016
Mona Jalal, MS at UW-Madison.	Fall 2014–Spring 2015
Boqun Yan, BS at UW-Madison.	Fall 2014–Spring 2015

STUDENT AWARDS

Lingjiao Chen awarded a Google PhD Fellowship.	2017–18
Lingjiao Chen is runner-up at SIGMOD 2017 Student Research Competition.	2017

THESIS COMMITTEE

Nikos Koulouris, PhD at UCSD (Advisor: Yannis Papakonstantinou). 2018
Title TBD

Julaiti Alafate, PhD at UCSD (Advisor: Yoav Freund). 2018
Title TBD

Chunbin Lin, PhD at UCSD (Advisor: Yannis Papakonstantinou). 2017
"Accelerating Query Processing on Compressed Data"

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

Sumedha Kattar, MS at 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 at UCSD (Advisor: Alin Deutsch). Spring 2018
"Querying Heterogeneous Data Sources : A Comparative Study"

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

SERVICE

Organization:
Lead Organizer, SoCal DB Day 2018
Co-Chair, ACM SIGMOD 2018 Workshop on Data Management for End-to-End ML (DEEM)
Organizing Committee, ACM SIGKDD 2018 Workshop on Common Model Infrastructure (CMI)
Organizing Committee, Extremely Large Databases (XLDB) Conference 2018

Program Committee:
ACM SIGMOD 2020, 2019, 2018, 2017
VLDB 2019, 2018
SysML 2019
ACM SIGMOD 2017 Demonstrations and Student Research Competition
ACM SIGMOD 2017 Workshop on Data Management for End-to-End ML (DEEM)
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

Other Research-Related:
Speaker at ACM SIGMOD 2018 New Researcher Symposium
Interviewee for ACM SIGMOD 2018 WebDB Workshop Article on "Data meets ML"
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/Contributions to Diversity:
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
Winter–Spring 2018: Member, UCSD LGBTQIA+ Undergraduate Scholarships Committee
Fall 2017–: Member of UCSD 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: 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:
2017–19: CSE MS Committee
2017: UCSD SDSC Sustainability Committee
2016–17: CSE PhD Admissions Committee

EXTRAMURAL FUNDING

Grants:

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

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

Gifts:

Hellman Fellowship. 2018
Opera Solutions Faculty Research Award. 2017
Google Faculty Research Award. Sole PI. 2016–17
NVIDIA GPU Grant. 2017

RESEARCH TALKS

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

Democratizing Machine Learning-based Data Analytics
UCSD CSE Faculty Research Seminar Oct 2018

Multi-Query Optimizations for ML Systems
Microsoft Cloud and Information Services Lab, Mountain View Sep 2018
Google Brain/TFX and DIA, Mountain View Sep 2018

Advice from PhD to Early Career
ACM SIGMOD New Researcher Symposium Oct 2018
Slides: `https://sigmod2018.org/nrs_slides/kumar.pdf`

Morpheus: Factorized Linear Algebra for Scalable Advanced Analytics
Google DIA, 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: C/C++, Java, Python, R, SQL

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

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