Arun Kumar

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

Email: arunkk@eng.ucsd.edu
Phone: (+1) 614-602-9734
Web: http://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
Thesis Co-advisors: Jeffrey Naughton and Jignesh M. Patel

M.S. in Computer Sciences. — 2009–2011
Research Supervisor: Christopher Ré

Indian Institute of Technology, Madras
B.Tech. in Computer Science and Engineering. — 2005–2009

RESEARCH INTERESTS

Data management and its intersection with ML (an area popularly known as advanced data analytics or data science), especially devising data management-inspired abstractions, systems, frameworks, and algorithms to make the end-to-end process of building and deploying ML/AI algorithms for data analytics applications easier (improving the productivity of data scientists and developers) and faster (improving runtime performance and introducing accuracy trade-offs). My work spans the gamut of building new data systems, algorithm design, empirical analysis, theoretical analysis, and working with practitioners to help deploy my research.

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

SELECTED HONORS

ACM SIGMOD Distinguished PC Member	2017
Google Faculty Research Award	2017
Invited Keynote at ACM SIGMOD DEEM Workshop	2017
UW 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 Govt. of India	2003–08

CONFERENCE PUBLICATIONS

A Comparative Evaluation of Systems for Scalable Linear Algebra-based Analytics
A. Thomas and A. Kumar
Under submission

Materialization Trade-offs for Feature Transfer from Deep CNNs for Multimodal Data Analytics

S. Nakandala and A. Kumar
Under submission

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

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

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

OTHER PEER-REVIEWED PUBLICATIONS

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 AND MANUSCRIPTS

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 190A: Topics in Database System Implementation
Instructor. UCSD. Spring 2018

CSE 291A: Advanced Data Analytics and ML Systems
Instructor. UCSD. Winter 2018

CSE 290A: Seminar on Advanced Data Science

Organizer. UCSD. Fall 2018

CSE 190D: Topics in Database System Implementation
Instructor. UCSD. Spring 2017

CSE 290B: Seminar on Advanced Data Science
Organizer. UCSD. Spring 2017

CSE 291G: Topics in Advanced Analytics
Instructor. UCSD. Winter 2017

CS 564: Database Management Systems: Design and Implementation
Instructor. UW-Madison. Fall 2015

CS 764: Topics in Database Management Systems
Guest Lecture (Instructor: Jeffrey Naughton). UW-Madison. Fall 2015

CS 764: Topics in Database Management Systems
Guest Lecture (Instructor: Christopher Ré). UW-Madison. Spring 2013

ADVISING (CURRENT)

Lingjiao Chen, PhD at UW-Madison (Co-advised by Paris Koutris). Fall 2015–
Supun Nakandala, PhD at UCSD. Fall 2017–
Vraj Shah, MS at UCSD. Fall 2016–
Anthony Thomas, MS at UCSD. Winter 2016–
Yaobang Deng, BS at UCSD. Fall 2017–
Side Li, BS at UCSD. Fall 2017–

ADVISING (ALUMNI)

Mingyang Wang, MS at UCSD. Spring 2016
Fengan Li, MS at UW-Madison. First employment: Google. 2015–16
Zhiwei Fan, BS at UW-Madison. Onward to MS at UW-Madison. 2015–16
Fujie Zhan, BS at UW-Madison. First employment: Epic Systems. 2015–16
Boqun Yan, BS at UW-Madison. First employment: Google. 2014–15
Mona Jalal, MS at UW-Madison. 2014–15

THESIS COMMITTEE

Chunbin Lin, PhD at UCSD. Spring 2017
"Accelerating Query Processing on Compressed Data"

Nishant Agarwal, MS at UCSD. Spring 2017
"A Real-Time Temporal Clustering Algorithm for Short Text, and its Applications"

Sumedha Kattar, MS at UCSD. Spring 2017
"Finding the burnability index of a point on a map using the historical fire data"

SERVICE

Organization:
Co-Chair, ACM SIGMOD 2018 Workshop on Data Management for End-to-End ML (DEEM)
Organizing Committee, Extremely Large Databases (XLDB) Conference 2018

Program Committee:
ACM SIGMOD 2019, 2018, 2017
VLDB 2018
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
ACM Transactions on Database Systems (TODS) 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:
Interviewee for ACM SIGMOD 2018 WebDB Workshop Article on "Data meets ML"
Co-chair of "Best of ICDE 2017" Selection Committee for TKDE
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:
2018: Member of UCSD CSE Diversity Committee
Winter 2018: Member, UCSD LGBTQIA+ Undergraduate Scholarship 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
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

TALKS

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, Perl, Python, R, SQL

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