

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://cse.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 machine learning (an area popularly known as advanced analytics or data science), especially devising data management-inspired abstractions, systems, frameworks, and algorithms to make the end-to-end process of building and using machine learning algorithms for data analytics 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 systems, algorithm design, theoretical analysis, empirical analysis, and working with practitioners to deploy my research.

SELECTED HONORS ACM SIGMOD Distinguished PC Member 2017
Google Faculty Research Award 2017
Invited Keynote at 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 *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

RESEARCH PROJECTS

HAMLET: Analyzed the accuracy effects of joins on machine learning and feature selection, and devised methods to avoid joins safely when learning over normalized data in order to improve performance and usability for analysts.

Project HAMLET website: <http://cse.ucsd.edu/~arunkk/hamlet>

Ideas applied in production internally by Facebook, LogicBlox, and MakeMyTrip.

SANTOKU: Built a toolkit to exploit database dependencies during learning, scoring, and feature selection for some popular ML techniques in the R environment in order to improve performance and usability for analysts.

Project SANTOKU website: <http://cse.ucsd.edu/~arunkk/santoku>

ORION: Developed new optimization techniques to push machine learning computations down through relational joins, instead of materializing the joins, in order to improve usability for analysts and improve performance at scale.

Project ORION website: <http://cse.ucsd.edu/~arunkk/orion>

Ideas applied in production internally by LogicBlox and Microsoft.

COLUMBUS: Built a declarative system for the task of feature selection over structured data in order to improve usability for analysts, and developed a new cost-based optimizer to improve performance.

Project COLUMBUS website: i.stanford.edu/hazy/victor/columbus

BISMARCK: Built a system that provides a unified architecture for in-RDBMS implementations of machine learning models to improve developability for software engineers, and developed techniques to improve performance at scale.

Project BISMARCK website: <http://i.stanford.edu/hazy/victor/bismarck>
Code contributed to the open-source library MADlib (Website: madlib.net). Code or ideas incorporated into products by Cloudera, EMC, and Oracle.

STACCATO: Built a system for managing OCR data in an RDBMS to improve usability for database administrators, and applied probabilistic models to improve accuracy, while mitigating the resultant performance issues.

Project STACCATO website: hazy.cs.wisc.edu/hazy/staccato

INDUSTRIAL EXPERIENCE

Research Intern, Microsoft Jim Gray Systems Lab Jun – Aug 2014
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
Built a tool for large-scale principal component analysis on the distributed computation platform REEF. My work has been used by teams inside Microsoft.

Research Assistant, Oracle Labs/Advanced Analytics Jun – Aug 2012
Built a tool for large-scale matrix factorization on MapReduce/Hadoop. My work has been incorporated into the product Oracle R for Enterprise. Invited article on Oracle's product blog: blogs.oracle.com/R/entry/low_rank_matrix_factorization_in.

Research Intern, IBM Research Almaden Jun – Aug 2011
Incorporated scalable ensemble learning and other meta-learning techniques into SystemML. My work has been incorporated into the product IBM BigInsights.

Intern, NetApp India May – July 2008
Prototyped a new Web Services framework for platform-independent communication with NetApp storage devices. My ideas were incorporated into the NetApp filer SDK.

TEACHING

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 / MENTORING

Lingjiao Chen, PhD at UW-Madison. Fall 2015–

Vraj Shah, MS at UCSD. Fall 2016–

Anthony Thomas, MS at UCSD. Winter 2016–

Mingyang Wang, MS at UCSD. Spring 2016

Fengyan Li, MS at UW-Madison. First employment: Google. 2015–16

<i>Fujie Zhan</i> , BS at UW-Madison. First employment: Epic Systems.	2015–16
<i>Zhiwei Fan</i> , BS at UW-Madison. Onward to MS at UW-Madison.	2015–16
<i>Boqun Yan</i> , BS at UW-Madison. First employment: Google.	2014–15
<i>Mona Jalal</i> , MS at UW-Madison.	2014–15
<i>Pradap Konda</i> , PhD at UW-Madison.	2012–13

**THESIS
COMMITTEE**

<i>Chunbin Lin</i> , PhD at UCSD. “Accelerating Query Processing on Compressed Data”	Spring 2017
<i>Nishant Agarwal</i> , MS at UCSD. “A Real-Time Temporal Clustering Algorithm for Short Text, and its Applications”	Spring 2017
<i>Sumedha Kattar</i> , MS at UCSD. “Finding the burnability index of a point on a map using the historical fire data”	Spring 2017

SERVICE

Program Committee:

ACM SIGMOD 2018
VLDB 2018
ACM SIGMOD 2017
ACM SIGMOD 2017 Demonstrations
ACM SIGMOD 2017 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:

ACM SIGMOD 2017 Student Research Competition Judge
IEEE ICDE 2017 PhD Symposium Panelist
IEEE ICDE 2017 Demonstrations Judge

Outreach/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

<i>Democratizing Advanced Analytics Beyond Just Plumbing</i> ACM SIGMOD DEEM Workshop (Invited Academic Keynote)	May 2017
<i>Democratizing Feature Engineering and Model Selection in Advanced Analytics</i> Opera Solutions, San Diego, CA (Invited)	May 2017
<i>Democratizing Distributed Advanced Analytics</i>	

UCSD Center for Networked Systems Lecture	Apr 2017
<i>CEREBRO: A System to Manage Deep Learning for Relational Data Analytics</i> CIDR “Gong Show”	Jan 2017
<i>Accelerating Advanced Analytics</i> Google, Mountain View, CA (Invited)	Dec 2016
<i>The Data Strikes Back! Research Challenges in Advanced Analytics</i> UCSD AI Seminar	Oct 2016
<i>Exploiting Database Dependencies to Accelerate Advanced Analytics</i> UCSD Database Seminar	Oct 2016
<i>Model-based Pricing of Relational Data in the Cloud</i> UCSD Database Seminar	Oct 2016
<i>Accelerating Advanced Analytics (Invited)</i>	Jan-Mar 2016
New York University Microsoft Research, Redmond, WA University of Illinois at Urbana-Champaign Cornell University University of California, San Diego 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)	
<i>Machine Learning over Joins of Multiple Tables</i> Wisconsin Institutes of Discovery Seminar	2015
<i>Learning Generalized Linear Models over Normalized Data</i> ACM SIGMOD	2015
<i>Stop that Join! Optimizing Feature Selection over Normalized Data for Naive Bayes</i> Wisconsin Database Group Seminar	2015
<i>On Learning Generalized Linear Models over Joins</i> Wisconsin Database Group Seminar	2014
<i>Usability and Developability Challenges in Advanced Analytics</i> Indian Institute of Technology, Madras (Invited)	2014
<i>On Learning over Joins</i> Microsoft Big Data Security Symposium (Invited) Microsoft Jim Gray Systems Lab	2014 2014
<i>On Integrating Advanced Analytics with Scalable Structured Data Management</i> Wisconsin CS Preliminary Exam	2014
<i>Scalable and Distributed PCA on REEF</i> Microsoft Cloud and Information Systems Lab	2013
<i>Commoditizing Large-Scale Analytics for the Enterprise around R</i> Microsoft Jim Gray Systems Lab (Invited)	2013
<i>Columbus: Feature Selection on Data Analytics Systems</i> 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

REFERENCES

Jeffrey Naughton
 Principal Scientist, Google
 Professor, Emeritus, University of Wisconsin-Madison
naughton@google.com

Jignesh M. Patel
 Professor, University of Wisconsin-Madison
jignesh@cs.wisc.edu

Christopher Ré
 Associate Professor, Stanford University
chrismre@cs.stanford.edu

David J. DeWitt
 Adjunct Professor, Massachusetts Institute of Technology
 John P. Morgridge Professor, Emeritus, University of Wisconsin-Madison
david.dewitt@outlook.com

Xiaojin Zhu
 Sheldon & Marianne Lubar Professor, University of Wisconsin-Madison
jerryzhu@cs.wisc.edu