

Christopher M. Jermaine

<http://www.cs.rice.edu/~cmj4>

cmj4@cs.rice.edu

Research Focus **Relational database systems and applied statistical methods.** I focus on developing systems that facilitate the application of methods from probability and statistics to perform analytics over very large data sets. Within this context, my research covers distributed database systems, database query languages and query optimization, applied machine learning and data mining. In addition to developing general-purpose systems software, I work in application areas that include biomedical informatics, genetics, and text analytics.

Education

Georgia Institute of Technology, Atlanta **June 1998-Dec 2002**
College of Computing, Doctor of Philosophy
Advisor: Edward Omiecinski

University of Arizona, Tucson **August 1997-May 1998**
Dept. of Computer Science, PhD student
Advisor: Richard Snodgrass

The Ohio State University, Columbus **September 1995-May 1997**
Dept. of Comp. & Info. Sci., Master of Science
Advisor: Renee Miller

The University of California, San Diego **December 1993**
Dept. of Math., Bachelor of Arts, Math/Comp. Sci.

Academic Positions Held

University of Florida, Gainesville **Fall 2002-December, 2008**
Assistant Professor, CISE Department

Rice University **January 2009-Present**
Associate Professor, Computer Science Department

Notable Systems Built

The SimSQL System. 120,000 lines of source code. SimSQL is a distributed relational database system augmented to run large-scale recursive machine learning computations (see [9] under Refereed Conference Publications). Web page: <http://www.rice.edu/SimSQL/SimSQL.html>.

The DataPath System. 70,000 lines of source code. DataPath is a high-performance parallel database engine based upon the “data centric” processing paradigm (see [15] under Refereed Conference Publications). This technology has been commercialized by collaborator Alin Dobra (Tera Insights, Inc.)

Awards	<p>Alfred P. Sloan Foundation Research Fellowship, 2008-2010.</p> <p>ACM KDD Conference Best Paper Runner Up, 2009. For our paper on the spatial anomaly detection (see [17] under Refereed Conference Publications).</p> <p>IBM Almaden Pat Goldberg Best Paper Award, 2009. For the best paper-published by IBM in the fields of math, computer science, electrical engineering.</p> <p>ACM SIGMOD Conference Best Paper Award, 2007. For our paper on the DBO database engine (see [26] under Refereed Conference Publications).</p> <p>National Science Foundation CAREER Award, 2004.</p>
---------------	---

Conference and Workshop Organization	<p>General Chair. 2018 ACM SIGMOD Conference.</p> <p>Program Co-Chair. 2014 ACM SIGMOD Tutorials Track.</p> <p>Program Co-Chair. 2013 IEEE ICDE Conference Technical Program.</p> <p>Program Chair. 2013 VLDB PhD Workshop.</p> <p>Program Chair. 2011 ACM SIGMOD Demonstrations Track.</p> <p>Program Co-Chair. 2011 VLDB Workshop Program.</p>
---	--

Editorial Positions Held	<p>Associate Editor. <i>ACM Transactions on Database Systems</i>. 2010 to present.</p> <p>Associate Editor. <i>Very Large Database Journal</i>. 2007 to 2013.</p> <p>Associate Editor. <i>IEEE Transactions on Knowledge and Data Engineering</i>. 2008 to 2013.</p> <p>Associate Editor. <i>IEEE Comp. Soc. Data Engineering Bulletin</i>, 2014-present.</p>
---------------------------------	---

Selected Conference Technical Program Committee Service	<p>Very Large Database Conference (VLDB): 2004 to 2012, 2015, 2016.</p> <p>Conference of the ACM Special Interest Group on the Management of Data (SIGMOD): 2006, 2009, 2010, 2012, 2013, 2015, 2016.</p> <p>IEEE International Conference on Data Engineering (ICDE): 2009, 2010, 2011, 2014, 2015.</p> <p>Conference of the ACM Special Interest Group on Knowledge Discovery in Data (KDD): 2004 to 2009, 2012, 2013.</p> <p>Many others, including: SIAM Data Mining (SDM), IEEE International Conference on Data Mining (ICDM), ACM Conference on Knowledge Management (CIKM).</p>
--	--

Recent Internal Rice Service	<p>Chair, Tenure-Track Rice CS Faculty Search Committee, 2015.</p>
-------------------------------------	---

Current PhD Students	<ol style="list-style-type: none">1. Risa Myers. 2010-present (PhD expected 2016). Research topic: Data mining and machine learning for biomedical informatics.2. Jacob Gao. 2011-present (PhD expected 2017). Research topic: Flexible and efficient user-defined functions for parallel database systems.3. Shangyu Luo. 2012-present (PhD expected 2018). Research topic: Adding linear algebra support to a relational database system.4. Letao Qi. 2013-present (PhD expected 2019). Research topic: Natural language processing for biomedical informatics.5. Arkabandhu Chowdhury. 2013-present (PhD expected 2019). Research topic: Natural language processing for biomedical informatics.6. Rohan Mukherjee. 2014-present (PhD expected 2020). Research topic: Data mining of large software repositories.7. Sourav Sikdar. 2014-present (PhD expected 2020). Research topic. Parallel and distributed machine learning and data mining.
-----------------------------	---

Current Post-docs and Research Staff	<ol style="list-style-type: none">1. Michael Gubanov. Postdoctoral scholar, 2015-present. Research topic: Adding linear algebra support to a relational database system.2. Kia Teymourian. Postdoctoral scholar, 2015-present. Research topic: Building distributed object stores.3. Carlos Monroy, PhD. Research staff. 2015-present. Research topic: Building distributed object stores.
---	---

PhD Theses Supervised	<ol style="list-style-type: none">1. Abhijit Pol: <i>Maintaining very large samples using the geometric file</i>, University of Florida, 2007. First Employment: Yahoo! Inc.2. Shantunu Joshi: <i>Sampling-based randomization techniques for approximate query processing</i>. University of Florida, 2007. First employment: Oracle, Inc.3. Jayendra G. Venkateswaran: <i>Indexing techniques for metric databases with costly searches</i>, University of Florida, 2008 (advised jointly with Tamer Kahveci). First employment: Oracle, Inc.4. Xiuyao Song: <i>Novel change detection techniques in multidimensional data mining</i>, University of Florida, 2008 (advised jointly with Sanjay Ranka). First employment: Yahoo! Inc.5. Subramanian Arumugam: <i>Efficient algorithms for spatiotemporal data management</i>, University of Florida, 2008. First employment: Greenplum, Inc.6. Mingxi Wu: <i>Statistical methods for fast anomaly detection</i>, University of Florida, 2008. First employment: Oracle, Inc.
------------------------------	---

7. **Manas H. Somaiya**: *Novel mixture models to learn complex and evolving patterns in high dimensional data*, University of Florida, 2009 (advised jointly with Sanjay Ranka). First employment: Amazon.com.
8. **Fei Xu**: *Correlation-aware statistical methods for sampling-based group by estimates*, University of Florida, 2009. First employment: Microsoft.
9. **Anna Drummond**: *Statistical machine learning for text mining with Markov chain Monte Carlo inference*, Rice University, 2014. First employment: Houston startup.
10. **Zhuhua Cai**: *Very large scale machine learning*, Rice University, 2014. First employment: Google, Inc.
11. **Luis Perez**: *Query processing and optimization for stochastic analytics*, Rice University, 2014. First employment: Self-employed consultant.
12. **Niketani Pansare**: *Large-scale online aggregation via distributed systems*, Rice University, 2014. First employment: IBM Almaden.

Current and Recent Research Grants

1. **National Science Foundation**, “Data Mining and Cleaning for Medical Data Warehouses.” 9/2010-9/2015, \$1,200,000. Sole PI at Rice (\$600K to UT-HS).
Project goal: This project focuses on statistical models and learning algorithms for quantifying and correcting errors in clinical data warehouse records.
2. **National Science Foundation**, “Design and Implementation of the DBO Database System.” 10/2009-10/2013, \$750,000. Sole PI.
Project goal: The project is concerned with the design and development of a unique system called DBO. Like traditional relational database systems, DBO can run database queries from start to finish and produce exact answers over very large archives. However, unlike any existing research or production system, DBO uses sampling algorithms to produce a statistical estimate for the final query answer at all times throughout query execution.
3. **National Science Foundation**, “The MCDB Database System for Managing and Modeling Uncertainty.” 9/2009-9/2015, \$500,000. Sole PI.
Project goal: The project is concerned with the design and implementation of a prototype database system called MCDB that allows an expert-level analyst or statistician to attach arbitrary stochastic models to the database data in order to simulate values for unknown or inaccurate data.
4. **Department of Energy**, “The MCDB System for Management and Analysis of Petabyte-Scale Uncertain Data.” 9/2009-9/2013, \$600,000. Sole PI.
Project goal: This project is concerned with scaling up MCDB for use in a large-scale distributed environment.
5. **National Science Foundation**, “III: Medium: SimSQL: A Database System Supporting Distributed Execution of Machine Learning Codes.” 9/2014-9/2017, \$1,200,000. Sole PI.
Project goal: This project will perform the fundamental research necessary to make machine learning in-a-database a mature technology.

6. **National Science Foundation**, “ABI Innovation: Algorithms and Models for Distributed Computation of Bayesian Phylogenetics.” 8/2014-8/2017, \$1,150,888. PI (co PI: Luay Nakhleh; 50% to Jermaine).

Project goal: The project’s aim is to develop parallel algorithms for Bayesian phylogenetic inference that are suitable for use in a modern cluster compute environment.

7. **DARPA**, “PLINY: Mining and Understanding Software Enclaves.” Fall 2014 to Fall 2018, approximately \$11,000,000. Co-PI (many other co-PIs; approximately \$3,000,000 to Jermaine).

Project goal: This project seeks to develop database, data mining, machine learning, and programming languages technologies that can mine millions of lines of open-source computer code to build models that can help programmers complete programming tasks.

Courses Taught

Applications of Discrete Structures, University of Florida. Summer 2003, Summer 2004, Summer 2005, Summer 2006.

Information and Database Systems, University of Florida. Fall 2002, Fall 2003, Fall 2004, Fall 2005, Fall 2006, Fall 2008.

Indexing Large Databases, University of Florida, Fall 2007.

Database System Implementation, University of Florida. Spring 2003, Spring 2004, Spring 2005, Spring 2006, Spring 2007, Spring 2008.

Indexing Large Databases, Rice University, Spring 2009.

Design and Analysis of Algorithms, Rice University. Fall 2009, Fall 2010.

Introduction to Program Design, Rice University. Fall 2011, Fall 2012, Fall 2013, Fall 2014.

Tools and Models for Data Science, Rice University. Fall 2015.

Introduction to Database Systems, Rice University. Spring 2010, Spring 2011, Spring 2012, Spring 2013, Spring 2014, Spring 2015.

Invited Presentations

University of California, Irvine (2015), *Large-scale machine learning with the SimSQL system.*

Carnegie Mellon University, Pittsburgh (2015), *Large-scale machine learning with the SimSQL system.*

Technical University of Berlin, Berlin, Germany (2015), *Large-scale machine learning with the SimSQL system.*

IBM Zurich, Zurich, Switzerland (2015), *Large-scale machine learning with the SimSQL system.*

Microsoft Research, Redmond (2014), *Large-scale machine learning with the SimSQL system.*

University of Michigan, Ann Arbor (2014), *Large-scale machine learning with the SimSQL system.*

University of Wisconsin, Madison (2014), *Large-scale machine learning with the SimSQL system.*

Google, Mountain View (2014), *Large-scale machine learning with the SimSQL system.*

University of Texas, Austin (2014), *Large-scale machine learning with the SimSQL system.*

Purdue University, West Lafayette (2014), *Large-scale machine learning with the SimSQL system.*

University of California, Santa Barbara (2014), *Large-scale machine learning with the SimSQL system.*

University of Edinburgh, Edinburgh, UK (2013), *Large-scale machine learning with the SimSQL system.*

EPFL, Lausanne, Switzerland (2013), *Large-scale machine learning with the SimSQL system.*

ETH Zurich, Zurich, Switzerland (2013), *Large-scale machine learning with the SimSQL system.*

NASA Langley Research Center, National Institute of Aerospace Lecture Series (2012), *Stochastic Analytics.*

National e-Science Institute Workshop on Statistical and Probabilistic Databases (at Johns Hopkins, 2011), *Stochastic Analytics.*

Emory University, Atlanta (2011), *The Monte Carlo Database System.*

Fudan University, China (2010), *The Monte Carlo Database System.*

Duke University, Durham (2010), *The Monte Carlo Database System.*

New England Database Society, Brown/Brandeis/MIT (2010), *The DataPath Database System.*

The University of California, Berkeley (2009), *The Monte Carlo Database System.*

IBM Almaden Pat Goldberg Award Lecture (joint with Peter Haas, 2009), *The Monte Carlo Database System.*

SIGKDD Workshop on Knowledge Discovery from Uncertain Data (2009), *Managing and Mining Uncertain Data: What Might We Do Better?*

Publications: Summary

Total of nine research papers in SIGMOD, ten in VLDB, six in ICDE. These are generally regarded as the top data management publication venues. Since 2009: four papers in SIGMOD, four in VLDB, two in ICDE.

Refereed Conference Publications

(Student/postdoc authors supervised by Jermaine are noted with an asterisk)

- [1] Risa Myers^{*}, John Frenzel, Joseph Ruiz, Christopher M. Jermaine: Correlating Surgical Vital Sign Quality with 30-Day Outcomes Using Regression on Time-Series Segment Features. In *Proc. SDM 2015*, 11 pages.
- [2] Yanxin Lu, Joe Warren, Christopher M. Jermaine, Swarat Chaudhuri, Scott Rixner: Grading the Graders: Motivating Peer Graders in a MOOC. In *Proc. WWW 2015*, 11 pages.
- [3] Zhuhua Cai^{*}, Luis Leopoldo Perez^{*}, Shangyu Luo^{*}, Jacob Gao^{*}, Zografoula Vagena^{*}, Christopher M. Jermaine: An Experimental Comparison of Platforms for Implementing and Executing Large-Scale Machine Learning Codes. In *Proc. ACM SIGMOD 2014*, 12 pages.
- [4] Luis Leopoldo Perez^{*}, Christopher M. Jermaine: History-Driven Query Optimization with Materialized Intermediate Views. In *Proc. IEEE ICDE, 2014*, 12 pages.
- [5] Anna Drummond^{*}, Christopher M. Jermaine: Senders, Receivers, and Authors in Document Classification. In *Proc. ICDM 2014*: 7 pages.
- [6] Anna Drummond^{*}, Yanxin Lu, Swarat Chaudhuri, Christopher M. Jermaine, Joe Warren, Scott Rixner: Learning to Grade Student Programs in a Massive Open Online Course. In *Proc. ICDM 2014*: 7 pages.
- [7] Zhuhua Cai^{*}, Christopher M. Jermaine, Zografoula Vagena^{*}, Dionysios Logothetis, Luis Leopoldo Perez^{*}: The Pairwise Gaussian Random Field for High-Dimensional Data Imputation. In *Proc. ICDM 2013*: 61, 10 pages.
- [8] Anna Drummond^{*}, Chris Jermaine, Zografoula Vagena^{*}: Topic Models For Feature Selection in Document Clustering. In *Proc. SDM 2013*: 521, 9 pages.
- [9] Zhuhua Cai^{*}, Zografoula Vagena^{*}, Luis Leopoldo Perez^{*}, Subramanian Arumugam^{*}, Peter J. Haas, Christopher M. Jermaine: Simulation of Database-valued Markov Chains Using SimSQL. In *Proc. ACM SIGMOD 2013*: 637, 12 pages.
- [10] B. Bue and C. Jermaine: Multiclass Domain Adaptation with Iterative Manifold Alignment. In *Proc. IEEE WHISPERS 2013*, 4 pages.
- [11] Zhuhua Cai^{*}, Chris Jermaine: The Latent Community Model for Detecting Sybils in Social Networks. In *Proc. NDSS 2012*, 10 pages.
- [12] Niketan Pansare^{*}, Chris Jermaine, Peter J. Haas, Nitendra Rajput: Topic Models over Spoken Language. In *Proc. ICDM 2012*: 1062-1067.
- [13] Manas Somaiya^{*}, Christopher Jermaine, Sanjay Ranka: Mixture Models for Learning Low-dimensional Roles in High-dimensional Data. In *Proc. ACM KDD 2010*: p 909, 9 pages.
- [14] Fei Xu^{*}, Ravi Jampani^{*}, Mingxi Wu^{*}, Chris Jermaine, Tamer Kahveci: Surrogate ranking for very expensive similarity queries. In *Proc. IEEE ICDE, 2010*: p 848, 12 pages.
- [15] Subi Arumugam^{*}, Alin Dobra, Christopher M. Jermaine, Niketan Pansare^{*}, Luis Leopoldo Perez^{*}: The DataPath system: a data-centric analytic processing engine for large data warehouses. In *Proc. ACM SIGMOD, 2010*: p 519, 12 pages.

- [16] Luis Leopoldo Perez*, Subi Arumugam*, Christopher M. Jermaine: Evaluation of probabilistic threshold queries in MCDB. In *Proc. ACM SIGMOD*, 2010: p 687, 12 pages.
- [17] Mingxi Wu*, Xiuyao Song*, Chris Jermaine, Sanjay Ranka, John Gums: A LRT Framework for Fast Spatial Anomaly Detection. *Proc. ACM KDD* 2009: p 887, 9 pages. **Runner-up Best Paper Award.**
- [18] Alin Dobra, Chris Jermaine, Florin Rusu, Fei Xu: Turbo-Charging Estimate Convergence in DBO. In *Proc. VLDB*, 2010 (paper published as *PVLDB* 2(1): 419-430 (2009)).
- [19] Xiuyao Song*, Chris Jermaine, Sanjay Ranka, John Gums: A Bayesian Mixture Model with Linear Regression Mixing Proportions. *Proc. ACM KDD* 2008: p 659, 10 pages.
- [20] Ravi Jampani*, Fei Xu*, Mingxi Wu*, Luis Leopoldo Perez*, Christopher M. Jermaine, Peter J. Haas: MCDB: a Monte Carlo Approach to Managing Uncertain Data. In *Proc. SIGMOD, 2008*: p 687, 14 pages.
- [21] Florin Rusu, Fei Xu*, Luis Leopoldo Perez*, Mingxi Wu*, Ravi Jampani*, Chris Jermaine, Alin Dobra: The DBO Database System. In *Proc. SIGMOD, 2008*: p 1223, 4 pages (software system demonstration).
- [22] Shantanu Joshi*, Chris Jermaine: Robust Stratified Sampling Plans for Low Selectivity Queries. In *Proc. IEEE ICDE, 2008*: p 199, 12 pages.
- [23] Mingxi Wu*, Chris Jermaine: A Bayesian Method for Guessing the Extreme Values in a Data Set. In *Proc. VLDB, 2007*: p 471, 12 pages.
- [24] Fei Xu*, Chris Jermaine: Randomized Algorithms for Data Reconciliation in Wide Area Aggregate Query Processing. In *Proc. VLDB, 2007*: p 639, 12 pages.
- [25] Xiuyao Song*, Mingxi Wu*, Chris Jermaine, Sanjay Ranka: Statistical Change Detection for Multidimensional Data. In *Proc. ACM KDD, 2007*: p 667, 10 pages.
- [26] Chris Jermaine, Subramanian Arumugam*, Abhijit Pol*, Alin Dobra: Scalable Approximate Query Processing with the DBO Engine. In *Proc. ACM SIGMOD, 2007*: p 725, 12 pages. **Best Paper Award.**
- [27] Ruoming Jin, Leonid Glimcher, Chris Jermaine, Gagan Agrawal: New Sampling-Based Estimators for OLAP Queries. In *Proc. IEEE ICDE, 2006*: p 18, 10 pages.
- [28] Subramanian Arumugam*, Chris Jermaine: Closest-Point-of-Approach Join for Moving Object Histories. In *Proc. IEEE ICDE, 2006*: p 86, 10 pages.
- [29] Mingxi Wu*, Chris Jermaine: Outlier Detection by Sampling with Accuracy Guarantees. In *Proc. ACM KDD, 2006*: p 767, 6 pages.
- [30] Jayendra Venkateswaran*, Deepak Lachwani, Tamer Kahveci, Chris Jermaine: Reference-based Indexing of Sequence Databases. In *Proc. VLDB, 2006*: p 906, 12 pages.
- [31] Chris Jermaine, Alin Dobra, Subramanian Arumugam*, Shantanu Joshi*, Abhijit Pol*: A Disk-Based Join With Probabilistic Guarantees. In *Proc. ACM SIGMOD, 2005*: p 563, 12 pages.
- [32] Abhijit Pol*, Chris Jermaine: Relational Confidence Bounds Are Easy With The Bootstrap. In *Proc. ACM SIGMOD, 2005*: p 587, 12 pages.
- [33] Chris Jermaine, Alin Dobra, Abhijit Pol*, Shantanu Joshi*: Online Estimation For Subset-Based SQL Queries. In *Proc. VLDB, 2005*: p 745, 12 pages.

- [34] Chris Jermaine, Abhijit Pol*, Subramanian Arumugam*: Online Maintenance of Very Large Random Samples. In *Proc. ACM SIGMOD, 2004*: p 299, 12 pages.
- [35] Chris Jermaine: Robust Estimation With Sampling and Approximate Pre-Aggregation. In *Proc. VLDB, 2003*: p 235, 12 pages.
- [36] Chris Jermaine: Playing Hide-And-Seek With Correlations. In *Proc. ACM KDD, 2003*: p 242, 6 pages.
- [37] Wai Gen Yee, Shamkant B. Navathe, Edward Omiecinski, Chris Jermaine: Bridging Response Time and Energy-Efficiency in Broadcast Schedule Design. In *Proc. EDBT, 2002*: p 572, 18 pages.
- [38] Chris Jermaine, Edward Omiecinski: Lossy Reduction for Very High Dimensional Data. In *Proc. IEEE ICDE, 2002*: p 663, 10 pages.
- [39] Chris Jermaine: The Computational Complexity of High-Dimensional Correlation Search. In *Proc. IEEE ICDM, 2001*: p 249, 7 pages.
- [40] Chris Jermaine, Edward Omiecinski, Wai Gen Yee: Maintaining a Large Spatial Database with T2SM. In *Proc. ACM GIS, 2001*: p 100, 6 pages.
- [41] Chris Jermaine: Computing Program Modularizations with the k -Cut Method. In *Proc. WCRE, 1999*: p 224, 11 pages.
- [42] Chris Jermaine, Anindya Datta, Edward Omiecinski: A Novel Index Supporting High Volume Data Warehouse Insertion. In *Proc. VLDB, 1999*: p 235, 11 pages.

Refereed Journal Publications

- [43] Florin Rusu, Zixuan Zhuang, Mingxi Wu, Christopher M. Jermaine: Workload-Driven Antjoin Cardinality Estimation. *ACM Trans. Database Syst.* 40(3): 16-46
- [44] Supriya Nirkhiwale, Alin Dobra, Christopher M. Jermaine: A Sampling Algebra for Aggregate Estimation. *PVLDB* 6(14): 1798-1809 (2013)
- [45] Graham Cormode, Minos N. Garofalakis, Peter J. Haas, Chris Jermaine: Synopses for Massive Data: Samples, Histograms, Wavelets, Sketches. *Foundations and Trends in Databases* 4(1-3): 1-294 (2012)
- [46] Niketan Pansare*, Vinayak R. Borkar, Chris Jermaine, Tyson Condie: Online Aggregation for Large MapReduce Jobs. *PVLDB* 4(11): 1135-1145 (2011)
- [47] Jayendra Venkateswaran*, Bin Song, Tamer Kahveci, Chris Jermaine: TRIAL: A Tool for Finding Distant Structural Similarities. *IEEE/ACM Trans. Comput. Biology Bioinform.* 8(3): 819-831 (2011)
- [48] Ravi Jampani*, Fei Xu*, Mingxi Wu*, Luis Leopoldo Perez*, Chris Jermaine, Peter J. Haas: The Monte Carlo Database System: Stochastic Analysis Close to the Data. *ACM Trans. Database Syst.* 36(3): 18-63 + 15 page appendix (2011)
- [49] Peter J. Haas, Christopher M. Jermaine, Subi Arumugam*, Fei Xu*, Luis Leopoldo Perez*, Ravi Jampani*: MCDB-R: Risk Analysis in the Database. *PVLDB* 3(1): 782-793 (2010)

- [50] Mingxi Wu^{*}, Chris Jermaine, Xiuyao Song, Sanjay Ranka: A Model-Agnostic Framework for Fast Spatial Anomaly Detection. *TKDD* 4(4): 20 (2010)
- [51] Shantanu Joshi^{*}, Christopher M. Jermaine: Sampling-based estimators for subset-based queries. *VLDB J.* 18(1): 181-202 (2009)
- [52] Mingxi Wu^{*}, Chris Jermaine: Guessing the extreme values in a data set: a Bayesian method and its applications. *VLDB J.* 18(2): 571-597 (2009)
- [53] Fei Xu^{*}, Christopher M. Jermaine, Alin Dobra: Confidence bounds for sampling-based group by estimates. *ACM Trans. Database Syst.* 33(3): (2008)
- [54] Chris Jermaine, Subramanian Arumugam^{*}, Abhijit Pol^{*}, Alin Dobra: Scalable approximate query processing with the DBO engine. *ACM Trans. Database Syst.* 33(4): (2008)
- [55] Shantanu Joshi^{*}, Christopher M. Jermaine: Materialized Sample Views for Database Approximation. *IEEE Trans. Knowl. Data Eng.* 20(3): 337-351 (2008)
- [56] Manas Somaiya^{*}, Christopher M. Jermaine, Sanjay Ranka: Learning correlations using the mixture-of-subsets model. *ACM Trans. Know. Disc. and Data Mining* 1(4): (2008)
- [57] Jayendra Venkateswaran^{*}, Tamer Kahveci, Christopher M. Jermaine, Deepak Lachwani: Reference-based indexing for metric spaces with costly distance measures. *VLDB J.* 17(5): 1231-1251 (2008)
- [58] Abhijit Pol^{*}, Christopher M. Jermaine, Subramanian Arumugam^{*}: Maintaining very large random samples using the geometric file. *VLDB J.* 17(5): 997-1018 (2008)
- [59] Chris Jermaine, Edward Omiecinski, Wai Gen Yee: The Partitioned Exponential File for Database Storage Management. *VLDB Journal* 16(4): p 417, 21 pages (2007).
- [60] Chris Jermaine: Online Random Shuffling of Large Database Tables. *IEEE Trans. Knowl. Data Eng.* 19(1): p 73, 12 pages + 3 page appendix (2007).
- [61] Xiuyao Song^{*}, Mingxi Wu^{*}, Christopher M. Jermaine, Sanjay Ranka: Conditional Anomaly Detection. *IEEE Trans. Knowl. Data Eng.* 19(5): p 631, 15 pages (2007).
- [62] Chris Jermaine, Alin Dobra, Subramanian Arumugam^{*}, Shantanu Joshi^{*}, Abhijit Pol^{*}: The Sort-Merge-Shrink join. *ACM Trans. Database Syst.* 31(4): p 1382, 35 pages + 10 page appendix (2006).
- [63] Chris Jermaine: Finding the Most Interesting Correlations in a Database: How Hard Can It Be? *Information Systems* 30(1): p 21, 25 pages (2005).
- [64] Wai Gen Yee, Shamkant B. Navathe, Edward Omiecinski, Chris Jermaine: Efficient Data Allocation over Multiple Channels at Broadcast Servers. *IEEE Trans. Computers* 51(10): p 1231, 6 pages (2002).