

Christopher M. Jermaine

150 Sabine St, #231

Houston, TX, 77007

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

cmj4@cs.rice.edu

<b>Education</b>	<b>Georgia Institute of Technology, Atlanta</b> College of Computing, Doctor of Philosophy Advisor: Edward Omiecinski	<b>June 1998-Dec 2002</b>
	<b>University of Arizona, Tucson</b> Dept. of Computer Science, PhD student	<b>August 1997-May 1998</b>
	<b>The Ohio State University, Columbus</b> Dept. of Comp. & Info. Sci., Master of Science	<b>September 1995-May 1997</b>
	<b>The University of California, San Diego</b> Dept. of Math., Bachelor of Arts, Math/Comp. Sci.	<b>December 1993</b>
<b>Work Experience</b>	<b>University of Florida, Gainesville</b> Assistant Professor, CISE Department Associate Professor, CISE Department	<b>Fall 2002-Summer, 2009</b>
	<b>Rice University</b> Associate Professor, Computer Science Department	<b>Jan 2009-Present</b>
<b>Research Focus</b>	<b>Relational database systems</b> and applied statistical methods. I focus on using tools from probability and statistics within a database to permit flexible, scalable analysis of database data, with an emphasis on medical applications.	
<b>Awards</b>	<b>Alfred P. Sloan Foundation Research Fellowship</b> , 2008-2010 <b>ACM SIGMOD Conference Best Paper Award</b> , 2007 For our paper on the DBO database engine (see [8] under Refereed Conference Publications). The award citation describes the paper as a “technical tour-de-force” solving a key open problem in analytic query processing. <b>National Science Foundation CAREER Award</b> , 2004 For work on approximate query processing.	
<b>Current Research Grants</b>	<b>National Science Foundation</b> , “Data Mining and Cleaning for Medical Data Warehouses.” 9/2010-9/2014, \$1.2M. Sole PI at Rice (\$600K to UT-HS). <b>National Science Foundation</b> , “Design and Implementation of the DBO Database System.” 10/2009-10/2013, \$750,000. Sole PI. <b>National Science Foundation</b> , “The MCDB Database System for Managing and Modeling Uncertainty.” 9/2009-9/2014, \$500,000. Sole PI. <b>Department of Energy</b> , “The MCDB System for Management and Analysis of Petabyte-Scale Uncertain Data.” 9/2009-9/2013, \$600,000. Sole PI.	

<b>Selected Recent Service</b>	<p><b>Associate Editor.</b> <i>ACM Transactions on Database Systems</i>.</p> <p><b>Associate Editor.</b> <i>Very Large Database Journal</i>. (Ending 2014)</p> <p><b>Associate Editor.</b> <i>IEEE Trans. on Know. and Data Eng.</i> (Ended 2013)</p> <p><b>Program Co-Chair.</b> 2014 ACM SIGMOD Tutorials Track.</p> <p><b>Program Co-Chair.</b> 2013 IEEE ICDE Conference Technical Program.</p> <p><b>Program Chair.</b> 2013 VLDB PhD Workshop.</p> <p><b>Program Chair.</b> 2011 ACM SIGMOD Demonstrations Track.</p> <p><b>Program Co-Chair.</b> 2011 VLDB Workshop Program.</p> <p><b>Program Committee Member.</b> Many conferences, including PC member for technical program of VLDB (2004 to 2012, 2015), ACM SIGMOD (2006, 2009, 2010, 2012, 2013), and ACM KDD (2004 to 2009, 2012, 2013), and others (in 2013: SIGMOD 2013, IEEE ICDE 2013, ACM SoCC 2013, ACM KDD 2013, SIAM SDM 2013).</p>
<b>Current Advisees</b>	<p><b>PhD Students:</b> Anna Gutowska, Risa Myers, Niketan Pansare, Luiz Perez, Cai Zhuhua, Qi Letao, Shangyu Luo, Jacob Gao, Arkabandhu Chowdhury</p>
<b>Graduated PhD Students</b>	<p><b>Abhijit Pol.</b> First employment: Yahoo!, Inc.</p> <p><b>Shantanu Joshi.</b> First employment: Oracle Corp.</p> <p><b>Mingxi Wu.</b> First employment: Oracle Corp.</p> <p><b>Subramanian Arumugam.</b> First Employment: Greenplum, Inc.</p> <p><b>Fei Xu.</b> First employment: Microsoft Corp.</p> <p><b>Xiuyao Song.</b> First Employment: Yahoo!, Inc. (50% advised by S. Ranka)</p>
<b>Recent Presentations</b>	<p><b>EPFL</b>, Switzerland (2013); <b>ETH Zurich</b>, Switzerland (2013); <b>University of Edinburgh</b>, UK (2013): “Large Scale Bayesian Machine Learning with the SimSQL System”</p> <p><b>Microsoft</b>, Redmond (2014); <b>University of Michigan</b> (2014); <b>University of Wisconsin</b>, Madison (2014): “Large Scale Machine Learning with the SimSQL System”</p>

### Refereed Conference Publications

- [1] Zhuhua Cai<sup>\*</sup>, Luis Leopoldo Perez<sup>\*</sup>, Shangyu Luo<sup>\*</sup>, Jacob Gao<sup>\*</sup>, Zografoula Vagena<sup>\*</sup>, Christopher M. Jermaine: An Experimental Comparison of Platforms for Implementing and Executing Large-Scale Machine Learning Codes. To appear, *Proc. ACM SIGMOD* 2014, 12 pages.
- [2] Luis Leopoldo Perez<sup>\*</sup>, Christopher M. Jermaine: History-Driven Query Optimizaton with Materialized Intermediate Vews. To appear, *Proc. IEEE ICDE*, 2014, 12 pages.
- [3] Zhuhua Cai<sup>\*</sup>, Christopher M. Jermaine, Zografoula Vagena<sup>\*</sup>, Dionysios Logothetis, Luis Leopoldo Perez<sup>\*</sup>: The Pairwise Gaussian Random Field for High-Dimensional Data Imputation. In *Proc. ICDM* 2013: 61, 10 pages.
- [4] Anna Drummond<sup>\*</sup>, Chris Jermaine, Zografoula Vagena<sup>\*</sup>: Topic Models For Feature Selection in Document Clustering. In *Proc. SDM* 2013: 521, 9 pages.

- [5] 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.
- [6] B. Bue and C. Jermaine: Multiclass Domain Adaptation with Iterative Manifold Alignment. In *Proc. IEEE WHISPERS* 2013, 4 pages.
- [7] Zhuhua Cai\*, Chris Jermaine: The Latent Community Model for Detecting Sybils in Social Networks. In *Proc. NDSS* 2012, 10 pages.
- [8] Niketan Pansare\*, Chris Jermaine, Peter J. Haas, Nitendra Rajput: Topic Models over Spoken Language. In *Proc. ICDM* 2012: 1062-1067.
- [9] 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.
- [10] 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.
- [11] 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.
- [12] Luis Leopoldo Perez\*, Subi Arumugam\*, Christopher M. Jermaine: Evaluation of probabilistic threshold queries in MCDB. In *Proc. ACM SIGMOD*, 2010: p 687, 12 pages.
- [13] 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.**
- [14] 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)).
- [15] 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.
- [16] 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.
- [17] 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).
- [18] Shantanu Joshi\*, Chris Jermaine: Robust Stratified Sampling Plans for Low Selectivity Queries. In *Proc. IEEE ICDE*, 2008: p 199, 12 pages.
- [19] Mingxi Wu\*, Chris Jermaine: A Bayesian Method for Guessing the Extreme Values in a Data Set. In *Proc. VLDB*, 2007: p 471, 12 pages.
- [20] Fei Xu\*, Chris Jermaine: Randomized Algorithms for Data Reconciliation in Wide Area Aggregate Query Processing. In *Proc. VLDB*, 2007: p 639, 12 pages.

- [21] Xiuyao Song<sup>\*</sup>, Mingxi Wu<sup>\*</sup>, Chris Jermaine, Sanjay Ranka: Statistical Change Detection for Multidimensional Data. In *Proc. ACM KDD, 2007*: p 667, 10 pages.
- [22] Chris Jermaine, Subramanian Arumugam<sup>\*</sup>, Abhijit Pol<sup>\*</sup>, Alin Dobra: Scalable Approximate Query Processing with the DBO Engine. In *Proc. ACM SIGMOD, 2007*: p 725, 12 pages. **Best Paper Award.**
- [23] Ruoming Jin, Leonid Glimcher, Chris Jermaine, Gagan Agrawal: New Sampling-Based Estimators for OLAP Queries. In *Proc. IEEE ICDE, 2006*: p 18, 10 pages.
- [24] Subramanian Arumugam<sup>\*</sup>, Chris Jermaine: Closest-Point-of-Approach Join for Moving Object Histories. In *Proc. IEEE ICDE, 2006*: p 86, 10 pages.
- [25] Mingxi Wu<sup>\*</sup>, Chris Jermaine: Outlier Detection by Sampling with Accuracy Guarantees. In *Proc. ACM KDD, 2006*: p 767, 6 pages.
- [26] Jayendra Venkateswaran<sup>\*</sup>, Deepak Lachwani, Tamer Kahveci, Chris Jermaine: Reference-based Indexing of Sequence Databases. In *Proc. VLDB, 2006*: p 906, 12 pages.
- [27] Chris Jermaine, Alin Dobra, Subramanian Arumugam<sup>\*</sup>, Shantanu Joshi<sup>\*</sup>, Abhijit Pol<sup>\*</sup>: A Disk-Based Join With Probabilistic Guarantees. In *Proc. ACM SIGMOD, 2005*: p 563, 12 pages.
- [28] Abhijit Pol<sup>\*</sup>, Chris Jermaine: Relational Confidence Bounds Are Easy With The Bootstrap. In *Proc. ACM SIGMOD, 2005*: p 587, 12 pages.
- [29] Chris Jermaine, Alin Dobra, Abhijit Pol<sup>\*</sup>, Shantanu Joshi<sup>\*</sup>: Online Estimation For Subset-Based SQL Queries. In *Proc. VLDB, 2005*: p 745, 12 pages.
- [30] Chris Jermaine, Abhijit Pol<sup>\*</sup>, Subramanian Arumugam<sup>\*</sup>: Online Maintenance of Very Large Random Samples. In *Proc. ACM SIGMOD, 2004*: p 299, 12 pages.
- [31] Chris Jermaine: Robust Estimation With Sampling and Approximate Pre-Aggregation. In *Proc. VLDB, 2003*: p 235, 12 pages.
- [32] Chris Jermaine: Playing Hide-And-Seek With Correlations. In *Proc. ACM KDD, 2003*: p 242, 6 pages.
- [33] 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.
- [34] Chris Jermaine, Edward Omiecinski: Lossy Reduction for Very High Dimensional Data. In *Proc. IEEE ICDE, 2002*: p 663, 10 pages.
- [35] Chris Jermaine: The Computational Complexity of High-Dimensional Correlation Search. In *Proc. IEEE ICDM, 2001*: p 249, 7 pages.
- [36] Chris Jermaine, Edward Omiecinski, Wai Gen Yee: Maintaining a Large Spatial Database with T2SM. In *Proc. ACM GIS, 2001*: p 100, 6 pages.
- [37] Chris Jermaine: Computing Program Modularizations with the  $k$ -Cut Method. In *Proc. WCRE, 1999*: p 224, 11 pages.
- [38] Chris Jermaine, Anindya Datta, Edward Omiecinski: A Novel Index Supporting High Volume Data Warehouse Insertion. In *Proc. VLDB, 1999*: p 235, 11 pages.

\* Indicates co-author who was a PhD student or postdoc majority-supervised by Dr. Jermaine.

## Refereed Journal Publications

- [39] Supriya Nirkhiwale, Alin Dobra, Christopher M. Jermaine: A Sampling Algebra for Aggregate Estimation. *PVLDB* 6(14): 1798-1809 (2013)
- [40] 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)
- [41] Niketan Pansare\*, Vinayak R. Borkar, Chris Jermaine, Tyson Condie: Online Aggregation for Large MapReduce Jobs. *PVLDB* 4(11): 1135-1145 (2011)
- [42] 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)
- [43] 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)
- [44] 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)
- [45] Mingxi Wu\*, Chris Jermaine, Xiuyao Song, Sanjay Ranka: A Model-Agnostic Framework for Fast Spatial Anomaly Detection. A Model-Agnostic Framework for Fast Spatial Anomaly Detection. *TKDD* 4(4): 20 (2010)
- [46] Shantanu Joshi\*, Christopher M. Jermaine: Sampling-based estimators for subset-based queries. *VLDB J.* 18(1): 181-202 (2009)
- [47] 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)
- [48] Fei Xu\*, Christopher M. Jermaine, Alin Dobra: Confidence bounds for sampling-based group by estimates. *ACM Trans. Database Syst.* 33(3): (2008)
- [49] Chris Jermaine, Subramanian Arumugam\*, Abhijit Pol\*, Alin Dobra: Scalable approximate query processing with the DBO engine. *ACM Trans. Database Syst.* 33(4): (2008)
- [50] Shantanu Joshi\*, Christopher M. Jermaine: Materialized Sample Views for Database Approximation. *IEEE Trans. Knowl. Data Eng.* 20(3): 337-351 (2008)
- [51] 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)
- [52] 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)
- [53] Abhijit Pol\*, Christopher M. Jermaine, Subramanian Arumugam\*: Maintaining very large random samples using the geometric file. *VLDB J.* 17(5): 997-1018 (2008)
- [54] Chris Jermaine, Edward Omiecinski, Wai Gen Yee: The Partitioned Exponential File for Database Storage Management. *VLDB Journal* 16(4): p 417, 21 pages (2007).
- [55] Chris Jermaine: Online Random Shuffling of Large Database Tables. *IEEE Trans. Knowl. Data Eng.* 19(1): p 73, 12 pages + 3 page appendix (2007).

- [56] Xiuyao Song\*, Mingxi Wu\*, Christopher M. Jermaine, Sanjay Ranka: Conditional Anomaly Detection. *IEEE Trans. Knowl. Data Eng.* 19(5): p 631, 15 pages (2007).
- [57] 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).
- [58] Chris Jermaine: Finding the Most Interesting Correlations in a Database: How Hard Can It Be? *Information Systems* 30(1): p 21, 25 pages (2005).
- [59] 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).