VITA: PETER J. HAAS

College of Information and Computer Sciences
University of Massachusetts Amherst
140 Governors Drive
Amherst, MA 01003-9264
(413) 545-3140
phaas@cs.umass.edu

Research Interests

Application of techniques from applied probability and statistics to the design, performance analysis, and control of systems for information management, mining, integration, exploration, decision support, and learning. Statistical and machine-learning techniques for modeling, simulation, design, and control of complex systems, especially discrete-event stochastic systems, with applications to healthcare, manufacturing, computer, telecommunication, work-flow, and transportation systems.

Education

Ph.D. (Operations Research) 1986, Stanford University.
M.S. (Statistics) 1984, Stanford University.
M.S. (Environmental Engineering) 1979, Stanford University.
S.B. *Magna cum Laude* (Engineering and Applied Sciences) 1978, Harvard University.

Experience

College of Information and Computer Sciences and Deprtment of Mechanical and Industrial Engineering, University of Massachusetts Amherst. Professor, 2017–present

Taught graduate course on computer simulation, undergraduate courses on discrete mathematics and on probability. Pursuing research on stochastic optimization in database systems, automatic document summarization, generative neural networks for creation, efficient deployment, and accelerated execution of simulation models, modeling of multiple chronic diseases for decision support, online maintenance of machine-learning models, time-biased sampling for online analytics and machine learning, and reducing training bias in machine learning models.

IBM Almaden Research Center, San Jose, CA. Research Staff Member, 1987–2014; Principal Research Staff Member, 2014–2017 (designation was created by IBM in 2014).

Analytics over massive data With the SystemML team, developed novel compressed linear algebra methods for scalable machine learning. With the Watson division, developed novel methods for principled computation of confidence values for machine-generated hypotheses. Also led an effort to develop technologies for managing uncertain data at scale, including novel Monte-Carlo-based query processing and machine learning techniques in traditional relational database systems and modern map-reduce settings; the resulting SimSQL system, developed jointly with Rice University, has recently been open sourced. Developed best-of-breed parallel and distributed Big Data algorithms for tasks including matrix completion as is used in recommender systems; optimization via gradient descent for machine learning, statistics, and decision support; analysis of dynamic interaction graphs such as Twitter mention-activity graphs; and efficient execution of "groupwise set-valued analytics" such as stratified sampling. Worked on sampling-based methods for visual analytics for model management in machine learning. Also investigated use of BluSpark platform for Internet-of-Things applications in healthcare

Simulation As part of IBM Splash project, played a leading role in developing a platform for combining heterogeneous datasets and simulation, statistical, and optimization models to support collaborative modeling, simulation, and analytics. Conducted basic research on modeling, stability analysis, and simulation of complex discrete-event stochastic systems and demonstrated applicability of theory and methods to local-area network, database, and manufacturing models. Co-developed the first non-Markovian stochastic Petri net model. This work has resulted

in an award-winning monograph and over 25 journal and conference publications.

Query optimization and processing Developed sampling-based algorithms for estimating the size and processing cost of select-join queries in a relational database system. Developed new estimators for "column cardinality" and other statistics used by database query optimizers; several of these estimators, along with related algorithms, have been incorporated into IBM's DB2 database products and have yielded a number of patents. Recent algorithms for distributed estimation of column cardinality have been recognized in *CACM* Research Highlights. Developed indexing techniques for speeding up analytical queries in Hadoop. Developed scan-sharing technique for multi-core main-memory database systems.

Advanced database functionality and analytics Helped develop, code, and direct the implementation of algorithms for correlation and regression analysis in DB2 and in IBM's Visual Warehouse product. Collaborated with J. M. Hellerstein (UC Berkeley) on developing an "online aggregation interface" for relational database systems, including invention of the "ripple join" algorithm, and led effort to develop a prototype interface for DB2. Played key role both in developing the ISO proposed standard for specifying sampling in SQL queries and in providing this sampling functionality within DB2 UDB. Conducted research on novel technologies for exploiting and extending database sampling capabilities in the DB2 product, as well as extending query-optimization technology to deal with sampling. Gave numerous seminars and webcasts to familiarize consultants and IBM customers with DB2's sampling and high-level analytics capabilities. With IBM LEO project, also developed technology to support "autonomic" data management systems that require minimal human intervention and automatically improve their performance by learning from past experience; some of this technology has been incorporated into the DB2 product and prototyped for IBM Informix Dynamic Server. As part of the IBM Infosphere project, developed a "synopsis warehouse" architecture for flexible and scalable data analysis, along with hashing-based and sampling-based algorithms for discovering fuzzy undeclared rules, functional dependencies, keys, similarities, and correlations in relational and XML data.

Other research Other activities have included developing methods for probabilistic information extraction from text, developing, for IBM's Tivoli division, novel methods for real-time detection and prediction of anomalous behaviors in complex software systems, developing stochastic models of workload-balancing strategies in parallel database systems, developing query-optimization techniques for XML data based on statistical learning methods, and developing a method for "watermarking" relational data to combat piracy.

Stanford University. Lecturer, 1998–2002; Consulting Associate Professor, 2003–2010; Consulting Professor, 2011–2017.

Taught annual graduate-level course on computer simulation. Pursued joint research with faculty.

Center for the Mathematical Sciences, U. Wisconsin, Madison, WI. Honorary Fellow, 1992–1993.

Lectured on simulation methods for generalized semi-Markov processes and stochastic Petri nets. With Prof. Jeffrey Naughton, developed sampling-based selectivity estimation methods for database systems.

Department of Decision and Information Sciences, Santa Clara University, Santa Clara, CA. Assistant Professor, 1985–1987.

Taught intro courses in probability and statistics; pursued research on discrete-event stochastic systems.

Stanford University, Stanford, CA. Research & teaching assistant, Department of Operations Research, 1981–1985.

Radian Corporation, Austin Texas. Staff Scientist, 1979–1981.

Performed air-quality modeling studies for EPA, Texas Air Control Board, and corporate clients. Also participated in a state-of-the-art study for the Bureau of Land Management of the effect of a proposed coal mining/power plant complex on atmospheric visibility in several adjacent national parks. Extended several existing computer models of atmospheric dispersion to predict visibility effects, and designed and implemented several new visibility models. Developed a program to model atmospheric dispersion of heavier-than-air toxic gases, as part of a proposed automated emergency evacuation system.

Awards and Recognition

SIGMOD Distinguished PC Member 2021
Invited speaker, INFORMS Simulation Society Research Workshop 2021
Member, Sigma Chi Honor Society, 2020–present
Winter Simulation Conference Best Contributed Theoretical Paper Finalist, 2020
VLDB Best Demonstration Award, 2020
VLDB Best Demonstration Runner-Up, 2020
SIGMOD Research Highlight Award, 2019
Research Highlights recognition in *Commun. ACM*, 2019
Best Paper, EDBT 2018
Distinguished speaker, EDBT 2018
IEEE Computing Edge Recognition, 2017
IBM Research 2016 Pat Goldberg Memorial Best Paper Award
SIGMOD Research Highlight Award, 2016
INFORMS Fellow, 2016
Best Paper, *VLDB*, 2016
Keynote speaker, *Spring Simulation Multi-Conference*, 2016
IBM Outstanding Innovation Award, 2015
IBM Principal Research Staff Member, 2014
PODS Invited Tutorial, 2014
ACM Fellow, 2013
IBM Master Inventor, 2012
IBM Research 2012 Pat Goldberg Memorial Best Paper Award
Best Paper, *NIPS Big Learning Workshop*, 2011
Best Paper Honorable Mention, VLDB Challenges and Visions Track, 2011
Best Paper, *SBP*, 2010
Keynote Speaker, *VLDB Workshop on Management of Uncertain Data*, 2010
IBM High-Value Patent Application Award, 2009
Research Highlights recognition in *Commun. ACM*, 2009
IBM Research 2008 Pat Goldberg Memorial Best Paper Award
IBM Supplemental Patent Award, 2008 (for distinguished patents)
ACM SIGMOD 2007 Test-of-Time Award (10 year best paper)
IBM Research 2006 Pat Goldberg Memorial Best Paper Award
IBM Research Division Award, 2005
IBM Invention Achievement Plateau Awards, 2002, 2004, 2005, 2006, 2007, 2008, 2009, 2013, 2014, 2016
IBM Research 2003 Pat Goldberg Memorial Best Paper Award
INFORMS College on Simulation 2003 Outstanding Publication Award
Meritorious Service Award, *Operations Research*, 1996, 2003
IBM Outstanding Technical Achievement Award, 2003
Thirty IBM Invention Achievement Awards for Patents Filed and/or Issued
Keynote Speaker, *11th Intl. Conf. Scientific and Statistical Database Management*
ACM SIGMOD 1999 Best Paper Honorable Mention
IBM Research 1999 Computer Science Best Paper Award
Leavey Fellow, Santa Clara University
Stanford University Fellowship
Harvard: Blumberg Creative Science Award, Harvard College Honorary Scholarship

Professional Service

VLDB Awards Committee, 2020, 2021, 2022
ACM SIGMOD Best Paper Committee, 2020
Winter Simulation Conference Board Member, 2020–present
Program Chair, *Winter Simulation Conference*, 2019

Co-Editor, *ACM TOMACS*, Special Issue on Model-Data Ecosystems, 2020
Co-Chair, Fifth INFORMS Simulation Research Workshop, 2017
INFORMS Simulation Society Elections Committee 2017
Co-Editor, *ACM TOMACS*, Special Issue in Honor of Donald Iglehart, 2015
ICDE PhD Colloquium Committee, 2015
Reviewer for NSF CAREER Grant proposals, 2015
Invited reviewer, MacArthur Foundation Genius Grants, 2014, 2015
Invited reviewer, Sloan Foundation, 2015
INFORMS Simulation Society Distinguished Service Award Committee, 2014–2016
NSF panelist (Computer Science), 2014
Sponsored session organizer, INFORMS National Meeting, 2014
Chair, INFORMS Simulation Society Elections Committee, 2013–2014
President, INFORMS Simulation Society, 2010–2012
Co-Editor, *ACM TOMACS*, Special Issue on Simulation of Complex Service Systems, 2012
Area/Associate Editor, *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 2004–present
Associate Editor, *ACM Trans. Database Systems*, 2015–present
Associate Editor, *Operations Research*, 1995–2018
Associate Editor, *VLDB Journal*, 2007–2013
Invited-Session Organizer, Winter Simulation Conference, 2012; INFORMS National Meeting, 2014
Co-Chair, Third INFORMS Simulation Research Workshop, 2011
Vice President, INFORMS Simulation Society, 2008–2010
Selection committee for Editor-in-Chief of *ACM Trans. Modeling Computer Simulation*, 2009
Co-Editor, *VLDB Journal*, Special Issue on Uncertain and Probabilistic Databases, 2008–2009
Member, *INFORMS*, 1984–present
Member, *ACM SIGMOD*, 2000–present
Program Committee, *4th Intl. Workshop, Petri Nets and Performance Models*
Program Committee, *11th Intl. Conf. Scientific and Statistical Database Management*
Program Committee, *ACM SIGMOD Intl. Conf. Management of Data*, 2002, 2005, 2007, 2021
Program Committee, *10th Intl. Workshop, Petri Nets and Performance Models*
Program Committee, *Intl. Conf. Very Large Data Bases (VLDB)*, 2004, 2006
Program Committee, *10th ACM SIGKDD Intl. Conf. Knowledge Discovery Data Mining*, 2004
Program Committee, *ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 2011
Outstanding Publication Committee, INFORMS Simulation Society, 2004–2007
Dantzig-Lieberman Memorial Fellowship Committee (Stanford University Dept. of MS&E)
Individual reviewer for over 120 papers in seven conferences and seventeen journals, 1987–present
Tenure and promotion, approx. 60 letters, 1987-present

Service Activities at IBM, Stanford, and UMass

Doctoral Program Director, CICS, UMass, 2022–
Promotion and Tenure Committee, CICS, UMass, 2018–19, 2019–20 (Co-Chair), 2020–21 (Co-Chair), 2021-2022 (Co-Chair)
Ethics Education Committee, CICS, UMass, 2018–19, 2019–20, 2020–21, 2021-2022
PhD Admissions Committee, CICS, UMass, 2017–18
Graduate Program Committee, CICS, UMass, 2017–18
Dissertation Committee (Chair), Matteo Brucato 2021
Dissertation Committee, Anna Fariha, UMass 2020–21
Dissertation Committee, Ryan McKenna 2020–21
Dissertation Committee, Luciano DiPalma, Ecole Polytechnique, 2020–21
Dissertation Committee, Yeounoh Chung, Brown U., 2018–19
Dissertation Committee, Abhishek Roy, UMass, 2018–19
Dissertation Committee, Moojoong Ra, Management Science and Engineering, Stanford, 2017
Dissertation Committee, Jihee Kim, Management Science and Engineering, Stanford, 2013
Dissertation Committee, Dmitry Smelov, Management Science and Engineering, Stanford, 2013

Dissertation Committee, Chang-Han Rhee, Management Science and Engineering, Stanford, 2013
Dissertation Committee, Parag Agrawal, Computer Science, Stanford, 2011
Dissertation Committee, Rainer Gemulla, Computer Science, TU Dresden, 2008
IBM Database Department recruiting at U. Michigan, Harvard, Brown, MIT, Stanford, U. Waterloo
IBM Invention Day, 2013
IBM Research quarterly patent open house mentoring events, 2013–2017
IBM Silicon Valley Development Lab Patent Pipeline event, 2014
IBM hiring review committees for roughly 180 applicants over thirty years
Reviewed roughly 50 IBM Invention Disclosures over thirty years
IBM Almaden Services Research, Best Paper Committee, 2010
IBM Employee Charitable Contribution Campaign (ECCC) Day of Caring volunteer, 2005–2017
IBM Black Family Night volunteer, 2010
IBM ECCC canvasser (approx. 60 people, 100% response rate) 2004, 2007
IBM Almaden Lab External Recognition Committee, 2013
National Engineering Week Volunteer (STEM presentations at low-income middle and high schools) , 2008–2017

Patents Granted

US6732110: Estimation of column cardinality in a partitioned relational database
US6778976: Selectivity estimation for processing SQL queries containing HAVING clauses
US6993516: Efficient sampling of a relational database
US7124146: Incremental cardinality estimation for a set of data values
US7277873: Method for discovering undeclared and fuzzy rules in databases
US7363324: Method, system and program for prioritizing maintenance of database tables
US7406200: Method and system for finding structures in multi-dimensional spaces using image-guided clustering
US7412429: Method and system for data classification by kernel density shape interpolation of clusters
US7512629: Consistent and unbiased cardinality estimation for complex queries with conjuncts of predicates
US7512574: Consistent histogram maintenance using query feedback
US7536403: Method for maintaining a sample synopsis under arbitrary insertions and deletions
US7543006: Flexible, efficient and scalable sampling
US7636735: Method for estimating the cost of query processing
US7647293: Detecting correlation from data
US7792856: Entity-based business intelligence
US7831592: System and method for updating database statistics according to query feedback
US7836356: Method for monitoring dependent metric streams to detect anomalies
US7987177: Method for estimating the number of distinct values in a partitioned dataset
US8140466: System and method for maintaining and utilizing Bernoulli samples over evolving multisets
US8234295: Managing uncertain data using Monte Carlo techniques
US8341180: Risk analysis for data-intensive stochastic models
US8352945: System, method, and apparatus for scan-sharing for business intelligence queries in an in-memory database
US8838648: Efficient discovery of keys in a database
US8903748: Systems and methods for large-scale randomized optimization for problems with decomposable loss functions
US9201989: Interpolation techniques used for time alignment of multiple simulation models
US9524326: Synchronization of time between different simulation models
US9697277: Stratified sampling using adaptive parallel data processing
US9805143: Composite simulation modeling and analysis
US9824167: Result caching for improving statistical efficiency of composite simulation models
US9910860: Split elimination in MapReduce systems
US10013782: Dynamic interaction graphs with probabilistic edge decay
US10635063: Systems and methods for highly parallel processing of parameterized simulations
US10685150: System for design and execution of numerical experiments on a composite simulation model

Books

[B1] *Synopses for Massive Data: Samples, Histograms, Wavelets, Sketches.* G. Cormode, M. Garofalakis, P. J. Haas, and C. Jermaine. (Published in *Foundations and Trends in Databases*, **4**, 2011, 1–294.)

[B2] *Stochastic Petri Nets: Modelling, Stability, Simulation*. P. J. Haas. Springer-Verlag, New York, 2002. **INFORMS College on Simulation 2003 Outstanding Publication Award**.

[B3] *DB2 UDB's High Function Business Intelligence in e-Business*. N. R. Alur, P. J. Haas, D. Momiroska, P. Read, N. H. Summers, V. Totanes, C. Zuzarte. IBM Redbook Series, 2002. ISBN 0-7384-2460-9.

Book Chapters

[C1] Monte Carlo methods for uncertain data. P. J. Haas. *Encyclopedia of Database Systems*, 2nd Ed., Springer, August, 2017.

[C2] Karp-Luby sampling. P. J. Haas. *Encyclopedia of Database Systems*, 2nd Ed., Springer, January, 2017.

[C3] Data-stream sampling: basic techniques and results. P. J. Haas. In *Data Stream Management: Processing High Speed Data Streams*. M. Garofalakis, J. Gehrke, R. Rastogi (eds). Springer-Verlag, 2016.

[C4] Regenerative simulation. P. J. Haas. *Encyclopedia of Operations Research and Management Science*, 3rd Ed., Springer, 2013.

[C5] Toward automated large scale information integration and discovery. P. Brown, P. J. Haas, J. Myllymaki, H. Pirahesh, B. Reinwald, and Y. Sismanis. In *Data Management in a Connected World*, T. Härder and W. Lehner, eds. Springer-Verlag, 2005.

[C6] Efficient data reduction methods for on-line association rule discovery. H. Brönnimann, B. Chen, M. Dash, P. J. Haas, Y. Qiao, and P. Scheuermann. In *Data Mining: Next Generation Challenges and Future Directions*. AAAI Press, 2004, 125–146.

Journal Papers

[J1] NIM: Generative neural networks for automated model and generation of simulation inputs. W. Cen and P.J. Haas. *ACM Trans. Modeling Comput. Simul.*, 2023, to appear.

[J2] Exact PPS sampling with bounded sample size. B. Hentschel, P.J. Haas, Y. Tian. *Information Processing Letters*, **182**, 2023, 106382.

[J3] A new mixed agent-based network and compartmental simulation framework for joint modeling of related infectious diseases– application to sexually transmitted diseases. C. Gopalappa, H. Balasubramanian, P.J. Haas. *Infectious Disease Modeling*, **8**, 2023, 84–100.

[J4] In-database decision support: Opportunities and challenges. A. Abouzied, P. J. Haas, A. Meliou. *IEEE Data Engrg. Bull.*, **45(3)**, 2022, 102–115

[J5] Introduction to the Special Issue for Towards an Ecosystem of Simulation Models and Data. P. J. Haas, and G. Theodoropoulos. *ACM Trans. Modeling Comput. Simul.*, **30(4)**, 2020, 1–3.

[J6] sPaQLTools: A stochastic package query interface for scalable constrained optimization (demo). M. Brucato, M. Mannino, A. Abouzzied, P. J. Haas, and A. Meliou. *PVLDB*, **13(12)**, 2020, 2881–2884. **Best Demonstration Award**.

[J7] SuDocu: Summarizing documents by example. A. Fariha, M. Brucato, P. J. Haas, and A. Meliou. *PVLDB* **13(12)**, 2020, 2861–2864. **Best Demonstration Runner-Up**.

[J8] General temporally-biased sampling schemes for online model management. B. Hentschel, P. J. Haas, and Y. Tian. *ACM Trans. Database Sys.*, **44(4)**, 2019, 14:1–14:45. Invited extended version of [P10].

[J9] Online model management via temporally-biased sampling. B. Hentschel, P. J. Haas, and Y. Tian. *SIGMOD Record*, **48(1)**, 2019, 69–76. **Invited SIGMOD Research Highlights paper**.

[J10] Compressed linear algebra for declarative large-scale machine learning. A. Elgohary, M. Boehm, P. J. Haas, F. R. Reiss, and B. Reinwald. *Commun. ACM*, **62**, 2019, 83–91. **Research Highlights section**.

[J11] Compressed linear algebra for large-scale machine learning. A. Elgohary, M. Boehm, P. J. Haas, F. R. Reiss, and B. Reinwald. *VLDB J.*, **27**, 2018, 719–744. **Invited extended version of [J16]**.

[J12] Literature-based automated discovery of tumor suppressor p53 phosphorylation and inhibition by NEK2. B.-K. Choi, T. Dayaram, N. Parikh, A. D. Wilkins, M. Nagarajan, I. B. Novikov, B. J. Bachman, P. J. Haas, J. L. Labrie, C. R. Pickering, A. K. Adikesavan, S. Regenbogen, K. Scott, L. Kato, A. Lelescu, C. M. Buchovecky, H Zhang, S. H. Bao, S. Boyer, G. Weber, K. L. Scott, Y. Chen, S. Spangler, L. A. Donehower, and O. Lichtarge. *Proc. Nat. Acad. Sci.*, **115(42)**, 2018, 10666–10671.

[J13] Foresight: Recommending visual insights. Ç. Demiralp, P. J. Haas, S. Parthasarathy, T. Pedapati. *PVLDB*, **10**, 2017, 1937–1940. Also presented at 2017 KDD IDEAS Workshop.

[J14] Scaling machine learning via compressed linear algebra. A. Elgohary, M. Boehm, P. J. Haas, F. R. Reiss, and B. Reinwald. *SIGMOD Record*, **46**, 2017, 42-49. **Invited SIGMOD Research Highlights paper**.

[J15] Sampling for scalable visual analytics. B. C. Kwon, J. Verma, P. J. Haas, and Ç. Demiralp. *IEEE Comput. Graphics Applications*, **37**, 2017, 100–108. **Recognized in** *IEEE Computing Edge***, March, 2017**.

[J16] Compressed linear algebra for large-scale machine learning. A. Elgohary, M. Boehm, P. J. Haas, F. R. Reiss, and B. Reinwald. *PVLDB*, **9**, 2016, 960–971., *VLDB* 2016.) **Best Paper Award**

[J17] Guest editors' introduction to special issue honoring Donald Iglehart. P. W. Glynn and P. J. Haas. *ACM Trans. Modeling Comput. Simul.*, **25**, 2015, 21.

[J18] On transience and recurrence in irreducible finite-state stochastic systems. P. W. Glynn and P. J. Haas. *ACM Trans. Modeling Comput. Simul.*, **25**, 2015, 25.

[J19] Shared-memory and shared-nothing stochastic gradient descent algorithms for matrix completion. F. Makari, C. Teflioudi, R. Gemulla, P. J. Haas, and Y. Sismanis. *Knowledge Inform. Sys.*, **42**, 2015, 493–523.

[J20] Guest editors' introduction to special issue on the third INFORMS Simulation Society Research Workshop. P. J. Haas, S. G. Henderson, and P. L'Ecuyer. *ACM Trans. Modeling Comput. Simul.*, **24**, 2014, 1.

[J21] Non-uniformity issues and workarounds in bounded-size sampling. R. Gemulla and P. J. Haas. *VLDB J.*, **22**, 2013, 753–772.

[J22] Data is dead...without what-if models. P. J. Haas, P. P. Maglio, P. G. Selinger, and W.-C. Tan. *PVLDB*, **4**, 2011, 1486–1489. Best Paper Honorable Mention, Challenges and Visions Track.

[J23] Sketches get sketchier. P. J. Haas. *Commun. ACM*, August, 2011. Invited Technical Perspective.

[J24] The Monte Carlo Database System: Stochastic Analysis Close to the Data. R. Jampani, L. Perez, M. Wu, F. Xu, C. Jermaine, and P. J. Haas. *ACM Trans. Database Sys.*, **36**, 2011, Article 3.

[J25] MCDB-R: Risk analysis in the database. S. Arumugam, R. Jampani, L. Perez, F. Xu, C. Jermaine, and P. J. Haas. *PVLDB*, **3**, 2010, 782–793.

[J26] Foreword to Special Issue on Probabilistic Databases. P. J. Haas and D. Suciu. *VLDB Journal*, 18(5), 2009, 987–988.

[J27] Discovering and exploiting statistical properties for query optimization in relational databases: A survey. P. J. Haas, I. F. Ilyas, G. M. Lohman, and V. Markl. *Statistical Analysis and Data Mining*, **1**, 2009, 223–250.

[J28] Distinct-Value Synopses for Multiset Operations. K. Beyer, R. Gemulla, P. J. Haas, B. Reinwald, Y. Sismanis. **Research Highlights section of** *Commun. ACM*, October, 2009.

[J29] Maintaining bounded-size sample synopses of evolving datasets. R. Gemulla, W. Lehner, and P. J. Haas. *VLDB Journal*, 2008, **17**, 173–202. **Special issue devoted to best papers from** *VLDB* **2006**.

[J30] Main-memory scan sharing for multi-core CPUs. L. Qiao, V. Raman, F. Reiss, P. J. Haas, and G. M. Lohman. *PVLDB*, **1**, 2008, 610–621.

[J31] Consistent selectivity estimation via maximum entropy. V. Markl, P. J. Haas, M. Kutsch, N. Megiddo, and T. M. Tran. *VLDB Journal*, 2007, **16**, 55–76. **Special issue devoted to best papers from** *VLDB* **2005**.

[J32] Laws of large numbers and functional central limit theorems for generalized semi-Markov processes. P. W. Glynn and P. J. Haas. *Commun. Statist. Stochastic Models*, **22**, 2006, 201–231.

[J33] An estimator of the number of species from quadrat sampling. P. J. Haas, Y. Liu, and L. Stokes. *Biometrics*, **62**, 2006, 135–141.

[J34] Making DB2 products self-managing: strategies and experiences. S. Lightstone, G. M. Lohman, P. J. Haas, V. Markl, J. Rao, A. Storm, and D. Zilio. *Data Engrg. Bull.*, 2006, **29**, 16–23.

[J35] On functional central limit theorems for semi-Markov and related processes. P. W. Glynn and P. J. Haas. *Commun. Statist.—Theory Meth.*, **33**, 2004, 487–506. Special issue on semi-Markov processes.

[J36] Watermarking relational data: framework, algorithms, and analysis. R. Agrawal, P. J. Haas, and J. Kiernan. *VLDB Journal*, **12**, 2003, 157–169. **Special issue devoted to best papers of VLDB 2002**.

[J37] Estimation methods for delays in non-regenerative discrete-event systems. P. J. Haas. *Commun. Statist. Stochastic Models*, **19**, 2003, 1–35.

[J38] The need for speed: speeding up DB2 using sampling. P. J. Haas. *IDUG Solutions Journal*, **10(2)**, 2003, 32–34.

[J39] On the validity of long-run estimation methods for discrete-event systems. P. J. Haas and P. W. Glynn. *Perf. Eval. Rev.*, **30**, 2002, 35–37. Special issue on the 4th Workshop Math. Perform. Modeling and Analysis (MAMA 2002).

[J40] Estimation of delays in non-regenerative discrete-event systems. P. J. Haas. *Perf. Eval. Rev.*, **28**, 2001, 36–38. Special issue on the 2nd Workshop Math. Perform. Modeling and Analysis (MAMA 2000).

[J41] Estimation methods for non-regenerative stochastic Petri nets. P. J. Haas. *IEEE Trans. Software Engrg.*, **25**, 1999, 218–236. **Special section devoted to best papers from PNPM '97.**

[J42] Interactive data analysis: The CONTROL project. J. M. Hellerstein, R. Avnur, A. Chou, C. Hidber, C. Olston, V. Raman, T. Roth, and P. J. Haas. *IEEE Computer*, **32**, August 1999, 51–59. **Cover feature**.

[J43] On simulation output analysis for generalized semi-Markov processes. P. J. Haas. *Commun. Statist. Stochastic Models*, **15**, 1999, 53–80.

[J44] Estimating the number of classes in a finite population. P. J. Haas and L. Stokes. *J. Amer. Statist. Assoc.*, **93(444)**, 1998, 1475–1487.

[J45] The New Jersey data reduction report. D. Barbara, W. DuMouchel, C. Faloutsos, P. J. Haas, J. M. Hellerstein, Y. E. Ioannidis, H. V. Jagadish, T. Johnson, R. T. Ng, V. Poosala, K. A. Ross, K. C. Sevcik. *IEEE Data Engrg. Bull.* **20**, December, 1997, 3–45.

[J46] Selectivity and Cost Estimation for Joins Based on Random Sampling. P. J. Haas, J. F. Naughton, S. Seshadri, and A. N. Swami. *ACM J. Computer Systems Sciences*, **52**, 1996, 550–569. **Special issue devoted to the best papers from PODS '93**.

[J47] Estimation methods for passage times based on one-dependent cycles. P. J. Haas and G. S Shedler. *Discrete Event Dynamic Systems: Theory and Applications* **6**, 1996, 43–72. This material was also presented at *INFORMS 1995 Applied Probability Conf.*, Atlanta, Georgia.

[J48] Passage times in colored stochastic Petri nets. P. J. Haas and G. S. Shedler. *Commun. Statist. Stochastic Models* **9**, 1993, 31–80.

[J49] The maximum and mean of a random length sequence. P. J. Haas. *J. Appl. Probability* **29**, 1992, 460–466.

[J50] Stochastic Petri nets: modelling power and limit theorems. P. J. Haas and G. S. Shedler. *Probab. Engrg. Informational Sci.* **4**, 1991, 477–498.

[J51] Stochastic Petri net representation of discrete event simulations. P. J. Haas and G. S. Shedler. *IEEE Trans. Software Engrg.* **15**, 1989, 381–393. **Special section devoted to best papers from PNPM '87**.

[J52] Stochastic Petri nets with timed and immediate transitions. P. J. Haas and G. S. Shedler. *Comm. Statist. Stochastic Models* **5**, 1989, 563–600. Special Issue Devoted to Computer-Experimental Methods in Probability.

[J53] Modelling power of stochastic Petri nets for simulation. P. J. Haas and G. S. Shedler. *Probab. Engrg. Informational Sci.* **2**, 1988, 435–459.

[J54] Regenerative generalized semi-Markov processes. P. J. Haas and G. S. Shedler. *Commun. Statist. Stochastic Models* **3**, 1987, 409–438.

[J55] Recurrence and regeneration in non-Markovian networks of queues. P. J. Haas and G. S. Shedler. *Commun. Statist. Stochastic Models* **3**, 1987, 29–52.

[J56] Regenerative stochastic Petri nets. P. J. Haas and G. S. Shedler. *Performance Evaluation* **6**, 1986, 189–204.

[J57] Regenerative simulation methods for local area computer networks. P. J. Haas and G. S. Shedler. *IBM J. Res. Develop.* **29**, 1985, 194–205.

[J58] The effects of NO_2-aerosol interaction on indices of perceived visibility impairment. P. J. Haas and A. J. Fabrick. *Atmos. Environ.*, **15**, 1981, 2171–2177.

Refereed Conference Proceedings Papers

[P1] Piloting an interactive ethics learning environment in undergraduate CS courses. F. Castro, S. Raipura, H. Conboy, A. Arroyo, P. J. Haas, L. Osterweil. *Proc. ACM SIGCSE, Vol. 1*, 2023, 659-665.

[P2] Enhanced simulation metamodeling via graph and generative neural networks. W. Cen and P. J. Haas. *Proc. 2022 Winter Simul. Conf.*, 2748-2759.

[P3] Augmenting decision making via interactive what-if analysis. S. Gathani, M. Hulsebos, J. Gale, P. J. Haas, Ç. Demiralp. *CIDR 2022*.

[P4] SubSumE: A dataset for subjective summary extraction from Wikipedia documents. M. Brucato, A. Fariha, O. Youngquist, A. Meliou, P. J. Haas. *Third Workshop on New Frontiers in Summarization (NewSum) – EMNLP 2021*.

[P5] NIM: Modeling and generation of simulation inputs via generative neural networks. W. Cen, E. A. Herbert, P. J. Haas. *Proc. 2020 Winter Simulation Conference*. **Best Contributed Theoretical Paper Finalist.**

[P6] Stochastic package queries in probabilistic databases. M. Brucato, A. Abouzied, P.J. Haas, A. Meliou. *Proc. 2020 ACM SIGMOD Intl. Conf. Management of Data*, 269–283.

[P7] MNC: Structure-exploiting sparsity estimation for matrix expressions. J. Sommer, M. Boehm, A. Evfimievski, B. Reinwald, P. J. Haas. *Proc. 2019 ACM SIGMOD Intl. Conf. Management of Data*, 1607–1623.

[P8] NIM: Generative Neural Networks for Simulation Input Modeling (Extended Abstract and Poster). W. Cen, E. A. Herbert, P.J. Haas. *2019 Winter Simul. Conf.*, National Harbor, MD, December, 2019.

[P9] NIM: Generative Neural Networks for Modeling and Generation of Simulation Inputs. E. A. Herbert, W. Cen, P.J. Haas. *2019 Summer Simul. Conf.*, Berlin, July, 2019.

[P10] Temporally biased sampling for online model management. B. Hentschel, P. J. Haas, Y. Tian. *Proc. 21st Intl. Conf. Extending Database Tech. (EDBT)*, 2018, 109–120. **Best Paper award.**

[P11] Foresight: Rapid data exploration through guideposts. Ç. Demiralp, P. J. Haas, S. Parthasarathy, T. Pedapati. *IEEE VIS DSIA Workshop*, 2017. Available as CoRR abs/1709.10513.

[P12] Predicting future scientific discoveries based on a networked analysis of the past literature. M. Nagarajan, A. D. Wilkins, B. J. Bachman, I. B. Novikov, S. Bao, P. J. Haas, M. E. Terrón-Díaz, S. Bhatia, A. K. Adikesavan, J. J. Labrie, S. Regenbogen, C. M. Buchovecky, C. R. Pickering, L. Kato, A. M. Lisewski, A. Lelescu, H. Zhang, S. Boyer, G. Weber, Y. Chen, L. Donehower, S. Spangler, O. Lichtarge. *Proc. 21st Intl. Conf. Knowledge Discovery and Data Mining (KDD)*, 2015, 2019–2028.

[P13] Dynamic interaction graphs with probabilistic edge decay. W. Xie, Y. Tian, Y. Sismanis, A. Balmin, and P. J. Haas. *Proc. 31st Intl. Conf. Data Engrg. (ICDE)*, 2015, 1143–1154.

[P14] Groupwise analytics via adaptive MapReduce. L. Peng, K. Zheng, A. Balmin, V. Ercegovac, P. J. Haas, and Y. Sismanis. *Proc. 31st Intl. Conf. Data Engrg. (ICDE)*, 2015, 1059–1070.

[P15] Improving the efficiency of stochastic composite simulation models via result caching. P. J. Haas. *Proc. Winter Simulation Conference*, 2014, 817–828.

[P16] Automated hypothesis generation based on mining scientific literature. S. Spangler, A. D. Wilkins, B. J. Bachman, M. Nagarajan, T. Dayaram, P. J. Haas, S. Regenbogen, C .R. Pickering, A. Comer, J. N. Myers, I. Stanoi, L. Kato, A. Lelescu, J. J. Labrie, N. Parikh, A. M. Lisewski, L. Donehower, Y. Chen, and O. Lichtarge. *Proc. 20th Intl. Conf. Knowledge Discovery and Data Mining (KDD)*, 2014, 1877–1886.

[P17] MCDB and SimSQL: Scalable Stochastic Analysis within the Database. P. J. Haas and C. Jermaine. *1st Intl. Workshop on Big Uncertain Data (BUDA)*, 2014.

[P18] Model-data ecosystems: Challenges, tools, and trends. P. J. Haas. *Proc. 34th ACM SIGACT-SIGMOD-SIGART Symp. Principles of Database Systems*, 2014, 76–87. **Invited tutorial paper**.

[P19] Panel: Are we effectively preparing our students to be Certified Analytics Professionals? R. C. H. Cheng, P. J. Haas, S. Robinson, and L. Schruben. *Proc. Winter Simulation Conference*, 2013.

[P20] Exploring large composite simulation models with Splash (poster). P. J. Haas, P. G. Selinger, I. Terrizzano, H. Xue. *XLDB 2013*.

[P21] Simulation of database-valued Markov chains with SimSQL. Z. Cai, Z. Vagena, C. Jermaine, P. J. Haas. *Proc. 2013 ACM SIGMOD Intl. Conf. Management of Data*, 637–648.

[P22] Eagle-eyed elephant: Split-oriented indexing in Hadoop. M. Y. Eltabakh, F. Özcan, Y. Sismanis, P. J. Haas, H. Pirahesh, and J. Vondrak. *EDBT*, 2013, 89–100.

[P23] Topic models over spoken language. N. Pansare, C. Jermaine, P. J. Haas, and N. Rajput. *ICDM*, 2012, 1062–1067.

[P24] On aligning massive time-series data in Splash. P. J. Haas and Y. Sismanis. *BigData 2012*. Also presented at *XLDB 2012*.

[P25] Splash: a platform for analysis and simulation of health. W. C. Tan, P. J. Haas, R. Mak, C. A. Kieliszewski, P. Selinger, P. P. Maglio, S. Glissmann, M. Cefkin, and Y. Li. *Proc. 2nd ACM SIGHIT Intl. Health Informatics Symp.*, 2012, 543–552.

[P26] Splash: Simulation optimization in complex systems of systems. P. J. Haas, N. C. Barberis, P. Phoungphol, I. G. Terrizzano, W.-C. Tan, P. G. Selinger, and P. P. Maglio. *50th Annual Allerton Conference on Communication, Control and Computing*, 2012, 414–425. IEEE. **Invited paper**.

[P27] On Simulation of non-Markovian stochastic Petri Nets with heavy-tailed firing times. P. W. Glynn and P. J. Haas. *Proc. Winter Simulation Conference*, 2012, 301.

[P28] Splash: A progress report on combining simulations for better health policy. M. Cefkin, S. M. Glissmann, P. J. Haas, Y. Li, P. P. Maglio, R. Mak, P. Selinger, W.-C. Tan. *INFORMS Healthcare Conf.*, 2011.

[P29] Very large scale Bayesian inference using MCDB. Z. Cai, Z. Vagena, C. Jermaine, and P. J. Haas. *NIPS Big Learning Workshop*, 2011.

[P30] Large-Scale Matrix Factorization with Distributed Stochastic Gradient Descent. R. Gemulla, P. J. Haas, Y. Sismanis, C. Teflioudi, and F. Makari. *NIPS Big Learning Workshop*, 2011. **Best Paper Award**.

[P31] Large-scale matrix factorization with distributed stochastic gradient descent. R. Gemulla, P. J. Haas, E. Nijkamp, and Y. Sismanis. *Proc. 17th Intl. Conf. Knowledge Discovery and Data Mining (KDD)*, 2011, 69–77. Also presented at *XLDB 2011*.

[P32] Splash: A progress report on building a platform for a 360 degree view of health. M. Cefkin, S. M. Glissmann, P. J. Haas, L. Jalali, P. P. Maglio, P. Selinger, and W.-C. Tan. *Proc. 5th INFORMS Workshop on Data Mining and Health Informatics*, 2010.

[P33] Ricardo: Integrating R and Hadoop. S. Das, R. Gemulla, P. J. Haas, Y. Sismanis. *Proc. 2010 ACM SIGMOD Intl. Conf. Management of Data*, 987–998.

[P34] Social factors in creating an integrated capability for health systems modeling and simulation. P. P. Maglio, M. Cefkin, P. J. Haas, and P. Selinger. *Proc. 2010 Intl. Conf. Social Computing, Behavioral Modeling, Prediction*, vol. 6007, *Lecture Notes in Computer Science*, Springer-Verlag, 2010, 44–51.

[P35] Report of 08421 Working Group: Classification, Representation and Modeling. A. Das Sarma, A. de Keijzer, A. Deshpande, P. J. Haas, I. F. Ilyas, C. Koch, T. Neumann, D. Olteanu, M. Theobald, and V. Vassalos. In *Uncertainty Management in Information Systems*, C. Koch, B. König-Ries, V. Markl, and M. van Keulen, Eds. Dagstuhl Seminar Proceedings No. 08421, 2009. http://drops.dagstuhl.de/opus/volltexte/2009/1941.

[P36] Report of 08421 Working Group: Probabilistic Databases Benchmarking. C. Koch and C. Ré and D. Olteanu and H.-J. Lenz and M. van Keulen and P. J. Haas and J. Z. Pan. In *Uncertainty Management in Information Systems*, C. Koch, B. König-Ries, V. Markl, and M. van Keulen, Eds. Dagstuhl Seminar Proceedings No. 08421, 2009. http://drops.dagstuhl.de/opus/volltexte/2009/1936.

[P37] Database meets simulation: Tools and techniques. P. J. Haas and C. Jermaine. *Proc. 2009 INFORMS Simulation Society Research Workshop*, 119–124.

[P38] $E = MC^3$: Managing uncertain enterprise data in a cluster-computing environment. F. Xu, V. Ercegovac, P. J. Haas, and E. Shekita. *Proc. 2009 ACM SIGMOD Intl. Conf. Management of Data*, 441–454.

[P39] Uncertainty management in rule-based information extraction systems. E. Michelakis, P. J. Haas, R. Krishnamurthy, and S. Vaithyanathan. *Proc. 2009 ACM SIGMOD Intl. Conf. Management of Data*, 101–114.

[P40] Resolution-aware query answering for business intelligence. Y. Sismanis, A. Fuxman, L. Wang, P. J. Haas, and B. Reinwald. *Proc. 25th Intl. Conf. Data Engrg.*, 2009, 976–987.

[P41] MCDB: A Monte Carlo approach to managing uncertain data. R. Jampani, F. Xu, M. Wu, L. Perez, C. Jermaine, and P. J. Haas. *Proc. 2008 ACM SIGMOD Intl. Conf. Management of Data*, 687–700.

[P42] On reservoir sampling with deletions. R. Gemulla, W. Lehner, and P. J. Haas. *Proc. European Workshop on Data Stream Analysis*, Caserta, Italy, 2007.

[P43] Detecting attribute dependencies from query feedback. P. J. Haas, F. Hueske, and V. Markl. *Proc. 33rd Intl. Conf. on Very Large Data Bases*, 2007, 830–841.

[P44] Maintaining Bernoulli samples over evolving multisets. R. Gemulla, W. Lehner, and P. J. Haas. *Proc. 27th ACM SIGACT-SIGMOD-SIGART Symp. Principles of Database Systems*, 2007, 93–102.

[P45] On Synopses for distinct-value estimation under multiset operations. K. Beyer, P. J. Haas, B. Reinwald, Y. Sismanis, and R. Gemulla. *Proc. 2007 ACM SIGMOD Intl. Conf. Management of Data*, 199–210.

[P46] Integrating query-feedback based statistics into Informix Dynamic Server. A. Behm, V. Markl, P. J. Haas, and K. Murthy. *BTW 2007*.

[P47] A dip in the reservoir: maintaining sample synopses of evolving data sets. R. Gemulla, W. Lehner, and P. J. Haas. *Proc. 32nd Intl. Conf. on Very Large Data Bases*, 2006, 595–606.

[P48] GORDIAN: Efficient and scalable discovery of all composite keys. Y. Sismanis, P. J. Haas, and B. Reinwald. *Proc. 32nd Intl. Conf. on Very Large Data Bases*, 2006, 691–702.

[P49] MAXENT: Consistent cardinality estimation in action (demo). V. Markl, M. Kutsch, T. M. Tran, P. J. Haas, and N. Megiddo. *Proc. 2006 ACM SIGMOD Intl. Conf. Management of Data*, 2006, 775–777.

[P50] Techniques for warehousing of sample data. P. G. Brown and P. J. Haas. *Proc. 22nd Intl. Conf. Data Engrg. (ICDE)*, 2006, 6.

[P51] ISOMER: Consistent histogram construction using query feedback. U. Srivastava, P. J. Haas, V. Markl, and N. Megiddo. *Proc. 22nd Intl. Conf. Data Engrg. (ICDE)*, 2006, 39.

[P52] Integrating a maximum-entropy cardinality estimator into DB2 UDB. M. Kutsch, P. J. Haas, V. Markl, N. Megiddo, and T. M. Tran. *Proc. 10th Intl. Conf. Extending Database Technology (EDBT)*, 2006, 1092–1096.

[P53] Statistical learning techniques for costing XML queries. N. Zhang, P. J. Haas, V. Josifovsky, G. M. Lohman, and C. Zhang. *Proc. 31st Intl. Conf. on Very Large Data Bases*, 2005, 289–300.

[P54] Consistently estimating the selectivity of conjuncts of predicates. V. Markl, N. Megiddo, M. Kutsch, T. M. Tran, P. J. Haas, U. Srivastava. *Proc. 31st Intl. Conf. on Very Large Data Bases*, 2005, 373–384.

[P55] Automated statistics collection in action (demo). P. J. Haas, M. Kandil, A. Lerner, V. Markl, I. Popivanov, V. Raman, and D. Zilio. *Proc. 2005 ACM SIGMOD Intl. Conf. Management of Data*, 2005, 933–935.

[P56] CORDS: Automatic generation of correlation statistics in DB2 (demo). I. Ilyas, V. Markl, P. J. Haas, P. G. Brown and A. Aboulnaga. *Proc. 30th Intl. Conf. on Very Large Data Bases*, 2004, 1341–1344.

[P57] Automated statistics collection in DB2 UDB. A. Aboulnaga, P. J. Haas, M. Kandil, S. Lightstone, G. Lohman, V. Markl, I. Popivanov, and V. Raman. *Proc. 30th Intl. Conf. on Very Large Data Bases*, 2004, 1146–1157.

[P58] CORDS: Automatic discovery of correlations and soft functional dependencies. I. Ilyas, V. Markl, P. J. Haas, P. G. Brown, and A. Aboulnaga. *Proc. 2004 ACM SIGMOD Intl. Conf. Management of Data*, 647–658.

[P59] Automatic relationship discovery in self-managing database systems. I. Ilyas, V. Markl, P. J. Haas, P. G. Brown and A. Aboulnaga. *Proc. Intl. Conf. Autonomic Computing (ICAC '04)*, 2004, 340–341.

[P60] A bi-level Bernoulli scheme for database sampling. P. J. Haas and C. König. *Proc. 2004 ACM SIGMOD Intl. Conf. Management of Data*, 2004, 275–286.

[P61] Stochastic Petri nets for modeling and simulation (tutorial). P. J. Haas. *Proc. Winter Simulation Conference*, 2004, 101–112.

[P62] Efficient Data Reduction with EASE. H. Brönnimann, B. Chen, M. Dash, P. J. Haas, P. Scheuermann. *Proc. 9th Intl. Conf. Knowledge Discovery and Data Mining (KDD)*, 2003, 59–68.

[P63] BHUNT: Automatic discovery of fuzzy algebraic constraints in relational data. P. G. Brown and P. J. Haas. *Proc. 29th Intl. Conf. on Very Large Data Bases*, 2003, 668–679.

[P64] A system for watermarking relational databases (demo). R. Agrawal, P. J. Haas, and J. Kiernan. *Proc. 2003 ACM SIGMOD Intl. Conf. Management of Data*, 674.

[P65] A new two-phase sampling based algorithm for discovering association rules. B. Chen, P. J. Haas, and P. Scheuermann. *Proc. 8th Intl. Conf. Knowledge Discovery and Data Mining (KDD)*, 2002, 462–468.

[P66] A scalable hash ripple join algorithm. G. Luo, P. J. Haas, and J. F. Naughton. *Proc. 2002 ACM SIGMOD Intl. Conf. Management of Data*, 2002, 252–262.

[P67] FAST: A new sampling-based algorithm for discovering association rules. B. Chen, P. J. Haas, and P. Scheuermann. *Proc. 18th Intl. Conf. Data Engrg. (ICDE)*, 2002, 263.

[P68] Hoeffding inequalities for join-selectivity estimation and online aggregation. P. J. Haas. *Computing Science and Statistics (Interface 2000)*, **31**, 2000, 74-78.

[P69] Ripple joins for online aggregation. P. J. Haas and J. M. Hellerstein. *Proc. 1999 ACM SIGMOD Intl. Conf. Management of Data*, 1999, 287–298. **Best Paper Honorable Mention**.

[P70] Techniques for online exploration of large object-relational datasets. P. J. Haas. *Proc. 11th Intl. Conf. Scientific and Statistical Database Management*, 1999, 4–12. **Keynote paper**.

[P71] Online aggregation. J. M. Hellerstein, P. J. Haas, and H. J. Wang. *Proc. 1997 ACM SIGMOD Intl. Conf. Management of Data*, 171–182. Reprinted in *Readings in Database Systems*, 3rd ed., Morgan Kaufmann, 1998. **SIGMOD 2007 Test of Time Award (10 year best paper)**.

[P72] Large-sample and deterministic confidence intervals for online aggregation. P. J. Haas. *Proc. Ninth Intl. Conf. Scientific and Statistical Database Management*, 1997, 51–63.

[P73] Estimation methods for stochastic Petri nets based on standardized time series. P. J. Haas. *Proc. Seventh Intl. Workshop Petri Nets and Performance Models*, 1997, 194–204.

[P74] Improved histograms for selectivity estimation of range predicates. V. Poosala, Y. E. Ioannidis, P. J. Haas, E. J. Shekita. *Proc. 1996 ACM SIGMOD Intl. Conf. Management of Data*, 294–305.

[P75] One-dependent cycles and passage times in stochastic Petri nets. P. J. Haas and G. S. Shedler. *Proc. Sixth Intl. Workshop on Petri Nets and Performance Models*, 1995, 191–202.

[P76] Sampling-based estimation of the number of distinct values of an attribute. P. J. Haas, J. F. Naughton, S. Seshadri, L. Stokes. *Proc. 21st Intl. Conf. on Very Large Data Bases*, 1995, 311–322.

[P77] Sampling-based selectivity estimation for joins using augmented frequent value statistics. P. J. Haas and A. N. Swami. *Proc. 11th Intl. Conf. Data Engineering*, 1995, 522–531.

[P78] On the relative cost of sampling for join selectivity estimation. P.J. Haas, J. F. Naughton, and A. N. Swami. *Proc. Thirteenth ACM SIGACT-SIGMOD-SIGART Symp. Principles of Database Systems*, 1993, 14–24.

[P79] Fixed-precision estimation of join selectivity. P. J. Haas, J. F. Naughton, S. Seshadri, and A. N. Swami. *Proc. Twelfth ACM SIGACT-SIGMOD-SIGART Symp. Principles of Database Systems*, 1993, 190–201.

[P80] Sequential sampling procedures for query size estimation. P. J. Haas and A N. Swami. *Proc. 1992 ACM SIGMOD Intl. Conf. Management of Data*, 1992, 1–11.

[P81] Stochastic Petri nets with simultaneous transition firings. P. J. Haas and G. S. Shedler. *Proc. Intl. Workshop Petri Nets and Performance Models*, IEEE Computer Society Press, Washington, D.C., 1987, 24–32.

[P82] Regenerative simulation of stochastic Petri nets. P. J. Haas and G. S. Shedler. *Proc. Intl. Workshop on Timed Petri Nets*, IEEE Computer Society Press, Silver Spring, MD, 1985, 14–21.

[P83] Analysis of Dispersion Models used for Complex Terrain Simulation. A. J. Fabrick and P. J. Haas. *Proc. DOE/NOAA/ORNL Symposium on Intermediate Range Transport Processes and Technology Assessment*, Gatlinburg, TN, 1981, 319–326.

Other Papers and Technical Reports

[M1] Understanding business users' data-driven decision-making: Practices, challenges, and opportunities. S. Gathani, Ç. Demiralp, P. J. Haas, L.Z. Liu. Submitted for publication.

[M2] Estimating the prevalence of multiple chronic diseases via maximum entropy. P. Amaranath, N. Khargonkar, P. Srinivasan, R. Thaikkat, H. Balasubramanian, P. J. Haas. Submitted for publication.

[M3] Exact PPS sampling with bounded sample size. B. Henschel, P. J. Haas, Y. Tian. arXiv:2105.10809.

[M4] Unknown Examples & Machine Learning Model Generalization. Y. Chung, P. J. Haas, T. Kraska, E. Upfal. arXiv:1808.08294.

[M5] Foresight: Rapid data exploration through guideposts. Ç. Demiralp, P. J. Haas, S. Parthasarathy, T. Pedapati. CoRR abs/1709.10513, 2017.

[M6] Distributed gradient descent via online aggregation. N. Pansare, C. Jermaine, and P. J. Haas. Manuscript.

[M7] Large-scale matrix factorization with distributed stochastic gradient descent. R. Gemulla, P. J. Haas, E. Nijkamp, and Y. Sismanis. IBM Research Report RJ 10481, 2011. (Revised Feb., 2013)

[M8] Problem detection via monitoring of metric-stream dependency structure. P. J. Haas, J. M. Lake, G. M Lohman, and T. Syeda-Mahmood. 2008.

[M9] Simulation methods for manufacturing systems using stochastic Petri nets. P. J. Haas and G. S. Shedler. IBM Research Report RJ 8672, 1992.

[M10] Labelled stochastic Petri nets: passage times. P. J. Haas and G. S. Shedler. IBM Research Report RJ 7933, 1990.

[M11] Stochastic Petri nets: simultaneous transition firing. P. J. Haas and G. S. Shedler. IBM Research Report RJ 7338, 1990.

[M12] Workload imbalance and parallel processing efficiency. P. J. Haas. IBM Research Report RJ 6936, 1989.

[M13] Markovian stochastic Petri nets. P. J. Haas and G. S. Shedler. IBM Research Report RJ 6764, 1989.

[M14] Recurrence and regeneration in non-Markovian simulations. P. J. Haas. Ph.D. Dissertation (advisor: D. L. Iglehart), Department of Operations Research, Stanford University, 1986.

[M15] *User Guide to IMPACT: An Integrated Model for Plumes and Atmospheric Chemistry in Complex Terrain.* A. J. Fabrick and P. J. Haas. DCN 80-241-403-01. Radian Corporation, Austin, TX, 1980.

Presentations

[T1] In-database decision support: Opportunities and challenges. Invited talk at Microsoft Gray Systems Lab, 2022.

[T2] Data-centric decision-making over thousands of simulation models. Invited talk at *2021 INFORMS Simulation Society Research Workshop*.

[T3] Making uncertain data management practical. Invited round table, VLDB 2020.

[T4] Approximate query processing: Overview and Research Challenges. Invited plenary talk, EDBT 2018.

[T5] Time-biased sampling for quick and dirty dynamic analytics. Invited talk at Facebook, 2018.

[T6] Some topics in model-data ecosystems. Invited talk at MIT, 2018.

[T7] Foresight: Rapid data exploration through guideposts. Ç. Demiralp, P. J. Haas, S. Parthasarathy, T. Pedapati. *Proc. DSIA Workshop*, October, 2017.

[T8] Simulation of complex systems. DARPA Proposer's Day, 2015.

[T9] Information management and simulation: innovation at the interface. Keynote talk at *Spring Simulation Multiconference*, April, 2016.

[T10] IBM Almaden activities in multi-model analysis. DARPA Invited Workshop on Multi-Modal Analysis, 2015.

[T11] Model-data ecosystems: Challenges, tools, and trends. Invited tutorial at *PODS*, 2014.

[T12] Splash: A computational platform for collaborating to solve complex real-world problems. Seminar, Center for Applied Mathematics Computing, and Statistics, San Jose State University, 2013.

[T13] Insights from Big Data: High-Performance Algorithms and Solutions, With P. G. Selinger and B. Reinwald. Presentation to members of Korea Electronics and Telecommunication Institute, 2013.

[T14] The Monte Carlo Database System: Querying Large-Scale Uncertain Data. DoD AUKS Invited Workshop, 2012.

[T15] Bringing Stochastic Analytics to the Data. EECS Department, UC Merced, 2011.

[T16] Splash: A Platform for Collaborative Modeling and Simulation. School of Engineering, Arizona State University, 2011.

[T17] On Recurrence and Transience in Heavy-Tailed Generalized Semi-Markov Processes. Dept. of Industrial & Systems Engineering, Georgia Tech, 2011

[T18] Composite Simulation Modeling of Complex Service Systems: Example and Research Challenges. Opening plenary talk, *2011 INFORMS Simulation Society Workshop*.

[T19] Splash: Smarter planet platform for analysis and simulation of health. *Brain to Society Diagnostic Project: 2nd International Roadmap Development Workshop*, 2010.

[T20] MCDB-R: Risk Analysis in the Database. *2010 INFORMS National Meeting*.

[T21] From MUD to MIRE: Managing the Inherent Risk in the Enterprise. Keynote talk, *2010 VLDB Workshop on Management of Uncertain Data*.

[T22] A Model Mashup Environment for Healthcare Support *2009 INFORMS National Meeting*.

[T23] On recurrence and transience in heavy-tailed generalized semi-Markov processes. RiskLab, ETH Zürich, 2009.

[T24] A Monte Carlo approach to managing uncertain data. Dagstuhl Seminar on Uncertainty Management in Information Systems, 2008. Technische Universität Dresden, 2009. Technische Universität Berlin, 2009. ETH Zurich, 2009. Universität Stuttgart,, 2009. New England Database Seminar, 2009.

[T25] An introduction to discrete-event simulation. With P. W. Glynn. *IMA Hot Topics Workshop on Stochastic Models for Intracellular Reaction Networks*, Minneapolis, MN, 2008.

[T26] On transience and recurrence in discrete-event simulations. *14th INFORMS Applied Probability Conf.*, Eindhoven, The Netherlands, 2007.

[T27] Online Aggregation at 10: Ongoing Results and Interactions. With J. M. Hellerstein. *Proc. 2007 ACM SIGMOD Intl. Conf. Management of Data*. (Invited talk in conjunction with SIGMOD 2007 Test of Time Award.)

[T28] Stochastic Petri nets for discrete-event simulation. P. J. Haas. Tutorial presented at *28th Intl. Conf. Application Theory Petri Nets and Other Models of Concurrency.* Siedlce, Poland, June, 2007.

[T29] Towards a Synposis Warehouse. Seminar, Beihang University Computer Science, Beijing, 2006. UC Berkeley Database Group Seminar, 2007, Stanford University InfoLab Seminar, 2007.

[T30] On transience and recurrence in irreducible finite-state stochastic systems. *2005 INFORMS National Meeting*.

[T31] BHUNT: Automatic Discovery of Fuzzy Algebraic Constraints in Relational Data. Database group, UC Berkeley, 2003

[T32] Speeding Up DB2 UDB Using Sampling. *IBM Data Management Conf.*, Anaheim, CA, 2002. *IDUG North America*, Las Vegas, NV, 2003, *DB2 BI Technical Conference*, 2005.

[T33] DB2 UDB Advanced Analytics for Business Intelligence. *IBM Data Management Conf.*, Anaheim, CA, 2002. *IDUG North America*, Orlando, FL, 2002 . *DB2 and Business Intelligence Technical Conf.*, Orlando, 2001.

[T34] Online query processing: A tutorial. With J. Hellerstein. *SIGMOD*, 2001.

[T35] Techniques for online exploration of large data sets. U. Toronto Computer Science Colloquium, 2000. UT Austin Data Mining Seminar, 2000.

[T36] Online aggregation for DB2: A next-generation decision-support interface. Demo at *CASCON '99*.

[T37] Database technology for decision support applications. Panel at *CASCON '99*.

[T38] Sampling and estimation methods for object-relational databases. Database colloquium, UC Berkeley, 1999.

[T39] Confidence-interval methodology for online aggregation. Database colloquium, UC Berkeley,1998.

[T40] Some sampling and estimation methods for SQL databases. *Uinveristy of Washington-Microsoft Research Summer Institute on Data Mining*, Seattle, WA, 1997.

[T41] Standardized time series and generalized semi-Markov processes. *1997 Spring INFORMS National Meeting*, San Diego, CA.

[T42] Simulation output analysis and generalized semi-Markov processes. Dept. of Management Science and Information Systems, UT-Austin, June, 1996.

[T43] Passage times in colored stochastic Petri nets. Two invited lectures for Computer Science Performance Seminar, University of Wisconsin-Madison, 1993.

[T44] Stochastic models for load balancing in parallel database systems. *1992 TIMS/ORSA Joint National Meeting*, Orlando, FL. Invited session on database interface and performance modeling.

[T45] Sequential sampling procedures for query size estimation. *1992 TIMS/ORSA Joint National Meeting*, San Francisco, CA. (Invited session on research issues in relational databases.) Also presented at Dept. of Computer Science, Seminar University of Wisconsin-Madison, 1992.

[T46] Labelled stochastic Petri nets. *ORSA/TIMS Special Interest Conf. Appl. Probab. in the Engineering, Informational, and Natural Sciences*, Monterey, CA, 1991.

[T47] Analysis techniques for Generalized Semi-Markov Processes. Three invited lectures for Operations Research seminar at Stanford University, 1989.

[T48] Simulation of stochastic Petri nets. Dept. of IEOR, UC Berkeley, 1989.

[T49] Regeneration and non-Markovian networks of queues. P. J. Haas and G. S. Shedler. *ORSA/TIMS Conf. Queueing Networks and their Applications*, New Brunswick, New Jersey, 1987.