# Sunita Sarawagi

Professor
Computer Science and Engineering
IIT Bombay, Powai, Mumbai-400076.
sunita@iitb.ac.in

http://www.cse.iitb.ac.in/ sunita
Fax: +91-22 2572-0022
Phone: +91-22 2576-7904

## Education

**PhD** in Computer Science, University of California at Berkeley
December 1996. Advisor: Michael Stonebraker
Thesis: Query Processing in Tertiary Memory Databases.

**M.S.** in Computer Science, University of California at Berkeley
December 1993. GPA = 3.95/4.0 Advisor: Michael Stonebraker
Thesis: Efficient Organization of Large Multidimensional Arrays.

**B.Tech** in Computer Science, Indian Institute of Technology, kharagpur
May 1991. GPA = 9.93/10.0 Advisors: P.P.Chakrabarti and S. Ghosh
Thesis: Algorithms for Rectangle Cutting and Packing Problems.
Second highest GPA among all graduating B.Tech students.

## Experience

Indian Institute of Technology Bombay March 2020-present
Professor in Charge, Center for Machine Intelligence and Data Science. The Center is the nodal point of contact for all AI/ML related activities of the university and has 70+ associated faculty.

Indian Institute of Technology Bombay March 2014-present
Professor, Computer Science

Google Inc, Mountain View, CA July 2014-July 2016
Visiting Scientist (on sabbatical and leave from IITB) Worked on deep learning models for personalizing and diversifying YouTube and Play recommendations, improving Duo's conversation assistance engine, and extracting attributes of classes from the Knowledge Graph.

Indian Institute of Technology Bombay March 2003-2014
Associate Professor, Computer Science, IIT Bombay

Carnegie Mellon University, Pittsburg Jan 2004-June 2004
Visiting Associate Professor, School of Computer Science. Co-developed the Semi-Markov Conditional Random Field (Semi-CRF) model with William Cohen.

Indian Institute of Technology Bombay Feb 1999-March 2003
Assistant Professor, School of Information Technology.

IBM Almaden Research Center Aug 1996–Feb 1999
Research Staff Member. Working in the data mining group on algorithms for OLAP and Data Mining

University of California at Berkeley May 1992–Aug 1996
Graduate Student Researcher.

## Recent research projects

Neural Models for Sequence Prediction with applications to dialog generation, translation, grammar correction, and time series forecasting. 2015-present

| | |
|---|---|
| Domain Adaptation and Domain Generalization. | 2017-present |
| Continuous, Reusable, Human intervenable and Modular Learning. | 2017-present |
| Machine learning models for reliable aggregate statistics over predicted variables. | 2012-2016 |
| Graphical models for selective node labeling in social networks. | 2011-2013. |
| WWT: Structure extraction from tables and lists on the web. | 2008-2014 |
| Inference algorithms for graphical models in information extraction task. | 2006-2010 |

**Research supervision**

- Doctoral thesis
  1. N. Lokesh (2020–) TBD
  2. Ashish Mittal (2020–) TBD
  3. Vihari Piratla (2017–): Domain Adaptation and Generalization of Machine Learning Models
  4. Abhijeet Awasthi (2017–): Learning with High-Level Supervision
  5. Prathamesh Deshpande (2017–): Accurate Forecasting in Time-series and Temporal Point Processes.
  6. Arun Iyer, 2016: Machine Learning Models for Predicting Aggregate LabelStatistics
  7. Rahul Gupta, 2010. Collective Conditional Random Fields for Information Extraction
  8. Ashish Tendulkar (Co-advised), 2009: Computational Biology using Machine Learning.
  9. Shantanu Godbole 2006: Machine Learning Models for Text categorization

**Sponsored research and Industry interaction**

- Member, Information Technology Sub-Committee of the Central Board of the Bank, RBI, India 2018-
- IBM AI Horizons research project (Co-PI) 2018-
- Google AI/ML India Research Awards for Faculty 2018
  OpenMind: Continuous, Compositional, and Human intervenable Learning
- Flipkart, Academic Collaboration 2017-2018
  Deep learning Models for Demand Forecasting
- Google Research, Faculty Advisor 2016-2017,
  Deep Learning Models for Conversation Assistance
- Yahoo! Research, 2008,2009,2011,2013
  Information Aggregation from tables on the Web
- Member of Scientific Advisory Board, Opera Solutions, CA (2011-2012).
- IBM Faculty Award, IBM Global Services India. 2003,2008
- Microsoft Research, Redmond, USA. 2003,2004,2005,2006,2007
  Exchange of Ideas on Data Mining and Data Cleaning Projects-Microsoft Research.
- Boeing Corporation, 2003
  Consultation on scalable interactive deduplication.
- Spectrum Corporation, 2002
  Licensed software for segmenting Indian Addresses.

**Honors**

- Infosys Science Foundation Award in Computer Science and Engineering, 2019.
- Distinguished Alumni Award, IIT Kharagpur, 2019

- H.H. Mathur Award for Excellence in Applied Sciences, 2019
- H-index 52 and more than 14,000 citations as per Google scholar Jul 2021.
- PAKDD Most Influential Paper Award 2014 for the paper: "Discriminative Methods for Multi-labeled Classification Shantanu Godbole and Sunita Sarawagi in PAKDD 2004".
- Fellow of the Indian National Academy of Engineering (INAE) 2013.
- Runners up for best paper award at IEEE ICDM 2012.
- Honorary mention for Outstanding student paper award in NIPS 2010.
- ACM SIGKDD Service Award 2009.
- IBM Faculty award, 2003, 2008.
- Best paper award at 1998 ACM-SIGMOD International Conference on Management of Data.
- *Eugene C. Gee and Mona Fay Gee Scholarship* for Spring 1995, *Dora Garibaldi Fellowship* for Fall 92 to spring 93, UC Berkeley.
- Institute award for second highest GPA over all departments, IIT Kharagpur, 1991.
- Best undergraduate thesis award, IIT Kharagpur, 1991.
- All India sixteenth rank in IIT Joint Entrance Examination, 1987.

## Professional activities

### Chair/vice chair positions

- VLDB 2011, VLDB 2016, Research track Co-chair.
- PC Co-chair, 2008 ACM-SIGKDD conference, ACM India CoDS-COMAD 2018
- Senior PC, Vice chair: ICML, IJCAI, KDD, NIPS, SIGMOD, ICDE, VLDB conferences

### Board memberships

- Technical Advisory Committee (TAC) Member, Data Science Lab, RBI (2019)
- ACM SIGKDD, member of the Board of directors. 2005-2012.
- Member, Board of trustees, VLDB Foundation (2008–2013)
- Member, SIGMOD Advisory board(2014–present)

### Journal editorship

- Journal of the ACM Transactions on Knowledge Discovery in Databases (TKDD) (2005-2010)
- Journal of the ACM Transactions on Database Systems (TODs) (2004–2007)
- Editor-in-chief, ACM SIGKDD (SIG Knowledge Discovery in Databases) newsletter. May 2003-2005.
- Foundations and Trends in Databases (2020-present)
- Foundations and Trends in Machine Learning (2007-present)
- ACM SIGKDD (SIG Knowledge Discovery in Databases) newsletter. 1999-May 2003.
- Associate editor, IEEE data engineering bulletin 2000-2001

### Award committees

- IEEE John Von Neumann Medal committee (2017-2020)
- ACM SIGMOD Award committee (2017–2021)
- VLDB Early career researcher, Ten year best paper awards (2013-2017)
- ACM SIGKDD 2010,2013,2017,2021: Innovation award and service award committee
- ACM SIGKDD 2001, 2009, 2010, 2014 Best paper award committee

### Program committee member

- ICDE 1997, 2001, 2002,
- SIGMOD 1998, 2002, 2005, 2006
- VLDB 2000, 2002, 2004
- SIGKDD 2001, 2003, 2004, 2005, 2009, 2010, 2017

- ICML 2003, 2011, 2013
- EDBT 2006, 2011
- WWW 2006, 2013
- CIDR 2009, 2010
- Demo committee, SIGMOD 2003
- WSDM 2013

**Selected publications**

- Deep Indexed Active Learning for Matching Heterogeneous Entity Representations. Arjit Jain, Sunita Sarawagi, Prithviraj Sen. In VLDB , 2022.
- Missing Value Imputation on Multidimensional Time Series. Parikshit Bansal, Prathamesh Deshpande, Sunita Sarawagi. VLDB 2021.
- Exploiting Language Relatedness for Low Resource Language Model Adaptation: An Indic Languages Study. Yash Khemchandani, Sarvesh Mehtani, Vaidehi Patil, Abhijeet Awasthi, Partha Talukdar and Sunita Sarawagi. In ACL-IJCNLP 2021 Main Conference.
- Training Data Augmentation for Code-Mixed Translation. Abhirut Gupta, Aditya Vavre and Sunita Sarawagi. In NAACL (Short paper), 2021.
- Error-driven Fixed-Budget ASR Personalization for Accented Speakers. Abhijeet Awasthi, Aman Kansal, Sunita Sarawagi, Preethi Jyothi In ICASSP, 2021.
- Long Horizon Forecasting With Temporal Point Processes Authors:Prathamesh Deshpande, Kamlesh Marathe, Abir De and Sunita Sarawagi In WSDM, 2021.
- NLP Service APIs and Models for Efficient Registration of New Clients Authors:Sahil Shah, Vihari Piratla, Soumen Chakrabarti and Sunita Sarawagi In Findings in EMNLP, 2020.
- Black-box Adaptation of ASR for Accented Speech Authors: Kartik Khandelwal, Preethi Jyothi, Abhijeet Awasthi, Sunita Sarawagi In Interspeech, 2020.
- Efficient Domain Generalization via Common-Specific Low-Rank Decomposition. Vihari Piratla, Praneeth Netrapalli, Sunita Sarawagi In ICML, 2020.
- What's in a Name? Are BERT Named Entity Representations just as Good for any other Name? Sriram Balasubramanian, Naman Jain, Gaurav Jindal, Abhijeet Awasthi and Sunita Sarawagi In Rep4NLP, 2020.
- Learning from Rules Generalizing Labeled Exemplars. Abhijeet Awasthi, Sabyasachi Ghosh, Rasna Goyal, Sunita Sarawagi. In ICLR, 2020.
- Data Programming using Continuous and Quality-Guided Labeling Functions. Oishik Chatterjee, Ganesh Ramakrishnan, Sunita Sarawagi. In AAAI, 2020.
- Parallel Iterative Edit Models for Local Sequence Transduction. Abhijeet Awasthi, Sunita Sarawagi, Rasna Goyal, Sabyasachi Ghosh and Vihari Piratla In EMNLP, 2019.
- Topic-Sensitive Attention on Generic Corpora Corrects Sense Bias in Pretrained Embeddings. Vihari Piratla, Sunita Sarawagi and Soumen Chakrabarti In ACL, 2019.
- Streaming Adaptation of Deep Forecasting Models using Adaptive Recurrent Units. Prathamesh Deshpande Sunita Sarawagi In KDD 2019.
- Posterior Attention Models for Sequence to Sequence Learning Shiv Shankar, Sunita Sarawagi In ICLR, 2019.
- Shiv Shankar, Siddhant Garg, Sunita Sarawagi. Surprisingly Easy Hard-Attention for Sequence to Sequence Learning. In EMNLP (short paper), 2018.
- Aviral Kumar, Sunita Sarawagi, Ujjwal Jain. Trainable Calibration Measures for Neural Networks from Kernel Mean Embeddings, ICML 2018.
- Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Soumen Chakrabarti, Preethi Jyothi, Sunita Sarawagi. Generalizing Across Domains via Cross-Gradient Training In ICLR, 2018.
- Shiv Shankar and Sunita Sarawagi. Labeled Memory Networks for Online Model Adaptation. In AAAI, 2018.

- Pavel Sountsov and Sunita Sarawagi. Length bias in Encoder Decoder Models and a Case for Global Conditioning, EMNLP 2016
- Arun Iyer, Saketh Nath, and Sunita Sarawagi. Privacy-preserving class ratio estimation. In ACM SIGKDD, 2016.
- Alon Y. Halevy, Natalya Fridman Noy, Sunita Sarawagi, Steven Euijong Whang, and Xiao Yu. Discovering structure in the universe of attribute names. In WWW, 2016.
- Aman Madaan, Ashish Mittal, Mausam, Ganesh Ramakrishnan, and Sunita Sarawagi. Numerical relation extraction with minimal supervision. In AAAI Conference on Artificial Intelligence, 2016.
- Immanuel Trummer, Alon Y. Halevy, Hongrae Lee, Sunita Sarawagi, and Rahul Gupta: Mining Subjective Properties on the Web. SIGMOD Conference 2015
- Sunita Sarawagi and Soumen Chakrabarti. Open-domain quantity queries on web tables: Annotation, response, and consensus models. In ACM SIGKDD, 2014.
- Arun Iyer, Saketh Nath, and Sunita Sarawagi. Maximum mean discrepancy for class ratio estimation: convergence bounds and kernel selection. In ICML, 2014.
- Gaurish Chaudhari, Vashist Avadhanula, and Sunita Sarawagi. A few good predictions: Selective node labeling in a social network. In WSDM, 2014.
- Sunita Sarawagi, Namit Katariya, and Arun Iyer. Active evaluation of classifiers on large datasets. In ICDM (**Runners-up for Best paper award**), 2012.
- Rakesh Pimplikar and Sunita Sarawagi. Answering table queries on the web using column keywords. VLDB, 2012.
- Rahul Gupta and Sunita Sarawagi. Joint training for open-domain extraction on the web: Exploiting overlap when supervision is limited. In WSDM, 2011.
- Rahul Gupta, Sunita Sarawagi, and Ajit A. Diwan. Collective inference for extraction mrfs coupled with symmetric clique potentials. JMLR, 11, November 2010.
- Sashank J. Reddi, Sunita Sarawagi, and Sundar Vishwanathan. MAP estimation in binary MRFs via bipartite multi-cuts. In NIPS 2010 (**Oral presentation, Honorary mention for Outstanding student paper award**).
- Girija Limaye, Sunita Sarawagi, and Soumen Chakrabarti. Annotating and searching web tables using entities, types and relationships. In PVLDB Vol 3, No 1. 2010.
- Rahul Gupta and Sunita Sarawagi. Answering table augmentation queries from unstructured lists on the web. In Proc. of the 35th Int'l Conference on Very Large Databases (VLDB), 2009.
- Sunita Sarawagi and Vinay S Deshpande and Sourabh Kasliwal. Efficient Top-K count queries over imprecise duplicates. EDBT 2009.
- Sunita Sarawagi. Information Extraction. FnT Databases Vol 1, No 3, 2008.
- Sunita Sarawagi and Rahul Gupta. Accurate max-margin training for structured output spaces. Proceedings of the 25th International Conference on Machine Learning (ICML), Helsinki, 2008.
- Rahul Gupta, Ajit A. Diwan, and Sunita Sarawagi. Efficient inference with cardinality-based clique potentials. Proceedings of the 24th International Conference on Machine Learning (ICML), USA, 2007.
- Sandeep Satpal and Sunita Sarawagi. Domain adaptation of conditional probability models via feature subsetting. In ECML/PKDD, 2007.
- Rahul Gupta and Sunita Sarawagi. Curating probabilistic databases from information extraction models, In Proc. of the 32nd Int'l Conference on Very Large Databases (VLDB), 2006.
- Sunita Sarawagi. Efficient inference on sequence segmentation models. In Proceedings of the 23rd International Conference on Machine Learning (ICML), Pittsburgh, PA, USA, 2006.
- Imran Mansuri and Sunita Sarawagi. A system for integrating unstructured data into relational databases. In Proc. of the 22nd IEEE Int'l Conference on Data Engineering (ICDE), 2006.
- Amit Chandel, P.C. Nagesh, and Sunita Sarawagi. Efficient batch top-k search for dictionary-based entity recognition. In Proc. of the 22nd IEEE Int'l Conference on Data Engineering (ICDE), 2006.
- Shantanu Godbole, Ganesh Ramakrishnan, and Sunita Sarawagi. Text classification with evolving labelsets. In ICDM, 2005.

- Semi-Markov Conditional Random Fields for Information Extraction, Sunita Sarawagi and William W. Cohen, NIPs 2004.
- Shantanu Godbole and Sunita Sarawagi. Discriminative methods for multi-labeled classification. In PAKDD, 2004.
- Abhay Harpale, Shantanu Godbole, Sunita Sarawagi, and Soumen Chakrabarti. Document classification through interactive supervision of document and term labels. In ECML/PKDD, 2004.
- Exploiting Dictionaries in Named Entity Extraction: Combining Semi-Markov Extraction Processes and Data Integration Methods, William W. Cohen and Sunita Sarawagi, Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, USA, 2004.
- Efficient set joins on similarity predicates, Sunita Sarawagi and Alok Kirpal, Proceedings of the ACM SIGMOD International Conference on Management of Data, 2004.
- Cross training: learning probabilistic mappings between topics, Sunita Sarawagi, Soumen Chakrabarti and Shantanu Godbole, Proc. of the Nineth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-2003), Washington D.C., USA, Aug 2003.
- Factorizing Complex Predicates in Queries to Exploit Indexes, Surajit Chaudhuri and Prasanna Ganesan and Sunita Sarawagi. Proc. ACM SIGMOD International Conf. on Management of Data, San Diego, USA, June 2003
- Interactive deduplication using active learning, Sunita Sarawagi and Anuradha Bhamidipaty, Proc. of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-2002), Edmonton, Canada.
- Scaling multi-class Support Vector Machines using inter-class confusion (poster paper), Shantanu Godbole and Sunita Sarawagi and Soumen Chakrabarti, Proc. of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-2002), Edmonton, Canada.
- ALIAS: An Active Learning led Interactive Deduplication System, Sunita Sarawagi and Anuradha Bhamidipaty and Alok Kirpal and Chandra Mouli, Proc. of the 28th Int'l Conference on Very Large Databases (VLDB) (Demonstration session), Hongkong, August, 2002
- Building Classifiers With Unrepresentative Training Instances: Experiences From The KDD Cup 2001 Competition, B. Anurandha and Anand Janakiraman and Sunita Sarawagi and Jayant Haritsa, Proc. of Workshop on Data Mining Lessons Learnt held in conjunction with the International Conference on Machine Learning, Sydney July 9th-12th, 2002,
- Efficient Evaluation of Queries with Mining Predicates, Surajit Chaudhuri, Vivek Narasayya and Sunita Sarawagi, Proc. of the 18th Int'l Conference on Data Engineering (ICDE), San Jose, USA, 2002.
- Intelligent Rollups in multidimensional OLAP data, G. Sathe and S. Sarawagi Proc. of the 27th Int'l Conference on Very Large Databases (VLDB), Italy, 2001.
- Automatic text segmentation for extracting structured records. V. Borkar, K. Deshmukh, S. Sarawagi, Proc. of the ACM SIGMOD International Conf. on Management of Data, Santa Barbara, USA, 2001.
- User-cognizant multidimensional analysis, S. Sarawagi. VLDB Journal, 10(2-3):224-239, 2001. (Invited)
- Integrating Association Rule Mining with Relational Database Systems: Alternatives and Implications, Sunita Sarawagi, Shiby Thomas, Rakesh Agrawal: Data Mining and Knowledge Discovery journal 4(2/3): 89-125 (2000)
- Informative differences in multidimensional data, Sunita Sarawagi, Data Mining and Knowledge Discovery journal, Data Min. Knowl. Discov. 5(4): 255-276 (2001)
- User Adaptive Exploration of OLAP data cubes S. Sarawagi. Proc. of the 26th Int'l Conference on Very Large Databases (VLDB), 2000.
- $I^3$ : Intelligent, Interactive Investigaton of OLAP data cubes. S. Sarawagi and G. Sathe In Proc. ACM SIGMOD International Conf. on Management of Data (Demonstration section), Dallas USA, May 2000.
- Explaining differences in multidimensional aggregates S. Sarawagi. Proc. of the 25th Int'l Conference on Very Large Databases (VLDB), 1999.
- Mining surprising patterns using temporal description length: S. Chakrabarti, S. Sarawagi and B.Dom, Proc. of the 24th Int'l Conference on Very Large Databases (VLDB), 1998.

- Integrating association rule mining with databases: alternatives and implications , S. Sarawagi, S. Thomas and R. Agrawal. Proc. of the 1998 ACM-SIGMOD International Conference on Management of Data. (winner **Best paper award**)
- Mining generalized association rules and sequential patterns using SQL queries ", S.Thomas and S.Sarawagi, Poster at the 4th International Conference on Knowledge Discovery and Data Mining (KDD-98)
- Discovery-driven exploration of OLAP data cubes, S. Sarawagi, R. Agrawal, N. Megiddo. Proc. of the Sixth Int'l Conference on Extending Database Technology (EDBT), Valencia, Spain, March 1998.
- Modeling Multidimensional Databases, R. Agrawal, A. Gupta, S. Sarawagi. Proc. of the 13th Int'l Conference on Data Engineering (ICDE) , Birmingham, U.K., April 1997.
- On Computing the Data Cube, S. Sarawagi, R. Agrawal and A. Gupta in "On the computation of multidimensional aggregates", Proc. of the 22nd Int'l Conference on Very Large Databases (VLDB), 1996.
- Efficient Organization of Large Multidimensional Arrays, S. Sarawagi and M. Stonebraker. Proc. Tenth International Conference on Data Engineering (ICDE), 1994.
- Benefits of Reordering Execution in Tertiary Memory Datab ases, S. Sarawagi and M. Stonebraker. Proc. of the 22nd Int'l Conference on Very Large Databases (VLDB), 1996.
- Query Processing in Tertiary Memory Databases, Proc. of the Twenty first Int'l conf. on Very Large Databases (VLDB) 1995.
- Database Systems for Efficient Access to Tertiary Memory, Proc. IEEE Mass Storage Symposium, 1995.

**US Patents**

- "Database System and Method Employing Data Cube Operator for Group-By Operations", R. Agrawal, A. Gupta, and S. Sarawagi, U.S. Patent number: AM9-96-003, filed March 1996, granted 1999.
- "Discovery driven exploration of OLAP data cubes", S. Sarawagi and R. Agrawal, U.S. Patent number: AM997036, filed 1997
- "Mining surprising patterns using temporal description length", Patent number: AM9-98-065, filed 1998.
- "Integrating mining with databases: alternatives and implications", S. Sarawagi, S. Thomas and R.Agrawal, U.S. Patent application AM9-98-081, filed 1998.
- System and method for organizing repositories of semi-structured documents such email. US US6592627B1 International Business Machines Corporation Priority 1999-06-10 Filing 1999-06-10 Grant 2003-07-15 Publication 2003-07-15
- Efficient evaluation of queries with mining predicates US US7346601B2 Surajit Chaudhuri, Vivek Narasayya, Sunita Sarawagi. Microsoft Corporation Priority 2002-06-03 Filing 2002-06-03 Grant 2008-03-18 Publication 2008-03-18
- System and method for explaining exceptions in data US US6691098B1 Rakesh Agrawal, Sunita Sarawagi International Business Machines Corporation. Priority 2000-02-08 Filing 2000-02-08 Grant 2004-02-10 Publication 2004-02-10