Yanlei Diao

Office: Laboratoire d'informatique (LIX)
1 rue Honoré d'Estienne d'Orves
Campus de l'École Polytechnique
91120 Palaiseau, France

Phone: +33 1 77 57 80 13
Email: yanlei.diao@polytechnique.edu
URL: http://www.lix.polytechnique.fr/~yanlei.diao/

Research Interests

Big Data Analytics, Data Management Systems, Scalable Intelligent Information Systems
big and fast data analytics, multi-objective optimization of cloud analytics, performance modelling through large-scale machine learning, explainable anomaly detection, data stream processing, interactive data exploration, insight discovery from RDF graphs, uncertain data management, high-performance genomic data processing, RFID/sensor data management, flash memory databases

Education

University of California, Berkeley — Berkeley, CA
Ph.D., Electrical Engineering and Computer Science — August 2005
Advisor: Professor Michael J. Franklin
Thesis: "Query Processing for Large-Scale XML Message Brokering"

Hong Kong University of Science and Technology — Hong Kong, China
M.S., Computer Science — June 2000
Advisor: Professor Hongjun Lu
Thesis: "Learning-based Web Query Processing"

Fudan University — Shanghai, China
B.S., Computer Science — July 1998

Academic Positions

Professor — 09/2015 – present
Laboratoire d'Informatique
Ecole Polytechnique, France

Professor
College of Information and Computer Sciences — 09/2019 – present
University of Massachusetts Amherst

Associate Professor
College of Information and Computer Sciences — 09/2011 – 08/2019
University of Massachusetts Amherst

Assistant Professor — 09/2005 – 08/2011

Department of Computer Science
University of Massachusetts Amherst

Visiting Professor 07/2011 – 12/2011
Department of Computer Science
Brown University

Visiting Professor 06/2008 – 08/2008
Computer Science and Artificial Intelligence Laboratory 06/2010 – 08/2010
Massachusetts Institute of Technology (MIT)

HONORS AND AWARDS

- **European Research Council (ERC) Consolidator Award**, €2.5M, 2017-2022
- **Keynote**, 15th ACM International Conference on distributed and event-based systems (DEBS), 2021; 2nd Int'l Workshop on Exploratory Search in Databases and the Wed (ExploreDB), co-located with SIGMOD 2015
- **Distinguished Lecture Series**, Max Planck Institut (MPI) Informatik, December 2021; IBM Almaden Research Center, April 2021; Naver Research Labs, January 2020; Northeastern University, 2022; Technische Universitaet Darmstadt, November 2017; University of Texas at Austin, 2005
- **CRA-W Borg Early Career Award**, 2013 (*one female scientist selected each year across all fields of Computer Science and Engineering for significant contributions in research and outreach*)
- **Distinguished Alumni**, School of Computer Science, Fudan University, Shanghai China, 2013
- **IBM Innovation Award on Scalable Data Analytics**, 2010
- **National Science Foundation CAREER Award**, 2008
- **Finalist for Microsoft Research New Faculty Fellowship**, 2008
- **Nomination for Distinguished Teaching Award**, University of Massachusetts Amherst, Fall 2007 and Spring 2008
- **ACM-SIGMOD Dissertation Award Honorable Mention**, 2005
- First Class Honor, Fudan University, 1998
- Intel Fellowship, Fudan University, 1996-1998
- GE Fellowship, Fudan University, 1995-1996

PROFESSIONAL SERVICE

- ### ADVISORY BOARDS AND AWARD COMMITTEES

  Chair, ACM SIGMOD Award Committee, 2022
  Member, ACM SIGMOD Award Committee, 2020-2023
  Chair, ACM SIGMOD Research Highlight Award, 2015-2019

Member, ACM SIGMOD Executive Committee,[1] April 2014 – August 2019
Member, VLDB Endowment, August 2016 - present
Member, ACM SIGMOD Software Systems Award Committee, 2014 – 2019

- **JOURNALS**

Editor-in-Chief, ACM SIGMOD Record, April 2014 – August 2019
Associate Editor, ACM Transactions on Database Systems (TODS),[2] 2014 – 2020
Associate Editor, Proceedings of VLDB (PVLDB), 2012 – 2013, 2015 – 2016

Reviewer, ACM Transactions on Database Systems (TODS)
Reviewer, International Journal on Very Large Data Bases (VLDB Journal)
Reviewer, ACM Transactions on Information Systems (TOIS)
Reviewer, ACM Transactions on Internet Technology (TOIT)
Reviewer, IEEE Transactions on Knowledge & Data Engineering (TKDE)
Reviewer, Journal of Computer Science and Technology (JCST)

- **CONFERENCES AND WORKSHOPS**

PC Co-Chair, IEEE International Conference on Data Engineering (ICDE), 2017
PC Co-Chair, ACM Symposium on Cloud Computing (SoCC), 2016
PC Chair, Paris Big Data Summit, May 2016
Area Chair, ACM SIGMOD Conference, 2014 – 2015
Area Chair and Organizing Committee Member, the Abstract track and Gong Show at the Conference on Innovative Data Systems Research (CIDR), 2011 – 2015
Chair, New England Database (NEDB) Summit, 2011
Co-Chair, New Researcher Symposium, ACM SIGMOD Conference, 2010 – 2011
Co-Chair, International Workshop on Data Management for Sensor Networks, 2008 – 2009
Steering Committee, International Workshop on Data Management for Sensor Networks, 2010
Publicity Chair, ACM SIGMOD Conference on Management of Data (SIGMOD), 2009
Co-organizer, New England Database Society Seminar Series, 2009-present

Committee Member, ACM International Conference on Management of Data (SIGMOD), 2011, 2014
Committee Member, International Conference on Very Large Data Base (VLDB), 2007, 2008, 2009, 2010, 2013, 2014
Committee Member, IEEE International Conference on Data Engineering (ICDE), 2007, 2008, 2009, 2010
Committee Member, ACM International Conference on Information and Knowledge Management (CIKM), 2010
Committee Member, Conference on Innovative Data Systems Research (CIDR), 2009, 2011, 2013
Committee Member, International Workshop on Networking Meets Databases (NetDB), 2008
Committee Member, International Workshop on RFID Data Management (RFDM), 2008
Committee Member, International Workshop on Data Management for Sensor Networks, 2007-2010
Committee Member, International Workshop on Scalable Stream Processing, 2007
Committee Member, International XML Database Symposium (XSym), 2006

- **PANELS**

Panelist, International Workshop on Data Management for Sensor Networks (DMSN), 2010

[1] A board of 9 members, making major decisions for the entire ACM Management of Data (SIGMOD) community.
[2] An ACM flagship journal, ranked as A* by CORE 2014 ranking.

Panelist, International Workshop on RFID Data Management (RFDM), 2008
Panelist, National Science Foundation, 2007

- **DIVERSITY EFFORTS**

Mentor and Speaker, CRA-W Graduate Cohort, 2011, 2013, 2014
Mentor and Participant, CRA-W Distributed Mentor Project (DMP), 2010, 2012, 2013
Massachusetts Aspirations in Computing Affiliate Award (MACAA), Cambridge MA, 2010 – 2014
Dot Diva Launch Event, Cambridge MA, September 2010
DataBase Mentoring (DB Me) Workshop, ACM SIGMOD 2010
Speaker, Women in Science Club at UMass Boston, April 2010

PUBLICATIONS

I. BIG DATA: SYSTEMS & OPTIMIZATION

[1] Chenghao Lyu, Qi Fan, Fei Song, Arnab Sinha, Yanlei Diao, Wei Chen, Li Ma, Yihui Feng, Yaliang Li, Kai Zeng, Jingren Zhou. "Fine-Grained Modeling and Optimization for Intelligent Resource Management in Big Data Processing." Proceedings of VLDB Endowment (PVLDB), August 2022.

[2] Fei Song, Khaled Zaouk, Chenghao Lyu, Arnab Sinha, Qi Fan, Yanlei Diao, Prashant Shenoy. "Spark-based Cloud Data Analytics using Multi-Objective Optimization." Proc. of the IEEE International Conference on Data Engineering (ICDE), pp. 396-407, 2021.

[3] Khaled Zaouk, Fei Song, Chenghao Lyu, and Yanlei Diao. "Neural-based Modeling for Performance Tuning of Spark Data Analytics." CoRR abs/2101.08167 (2021). https://arxiv.org/abs/2101.08167

[4] Khaled Zaouk, Fei Song, Chenghao Lyu, Arnab Sinha, Yanlei Diao, Prashant J. Shenoy. "UDAO: A Next-Generation Unified Data Analytics Optimizer." PVLDB 12(12): 1934-1937 (2019)

[5] Boduo Li, Yunmeng Ban, Yanlei Diao, and Prashant Shenoy. "Supporting Scalable Analytics with Latency Constraints." PVLDB 8(11): 1166-1177, July 2015. (21% acceptance rate)

[6] Buduo Li, Ed Mazur, Yanlei Diao, Andrew McGregor, and Prashant Shenoy. "Scalla: A Platform for Scalable One-pass Analytics using MapReduce". ACM Transactions on Database Systems (TODS), 37(4):27-64, December 2012. (*Special Issue on Best Papers of SIGMOD 2011*)

[7] Buduo Li, Ed Mazur, Yanlei Diao, Andrew McGregor, and Prashant Shenoy. "A Platform for Scalable On-pass Analytics using MapReduce". In Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD), 985-996, June 2011. (*One of the Best Papers*)

[8] Devesh Agrawal, Deepak Ganesan, Ramesh Sitaraman, Yanlei Diao, and Shashi Singh. "Lazy-Adaptive Tree: An Optimized Index Structure for Flash Devices." In Proceedings of the 35th International Conference on Very Large Data Bases (VLDB), 361-372, August 2009. (16.7% acceptance rate)

[9] Ed Mazur, Boduo Li, Yanlei Diao, and Prashant Shenoy. "Towards Scalable One-Pass Analytics using MapReduce." In Proceedings of the *1st International Workshop on Data Intensive Computing in the Clouds,* May 2011. 10 pages.

II. BIG DATA: DATA ANALYTICS AND STREAM PROCESSING

Explainable Anomaly Detection on Data Streams

[10] Vincent Jacob, Fei Song, Arnaud Stiegler, Yanlei Diao, Nesime Tatbul. "Exathlon: An Open Benchmark for Explainable Anomaly Detection." Proc. VLDB Endow. 14(11): 2613-2626 (2021). Also available online at http://arxiv.org/abs/2010.05073

[11] Vincent Jacob, Fei Song, Yanlei Diao, and Nesime Tatbul. "A Demonstration of the Exathlon Benchmarking Platform for Explainable Anomaly Detection." Proc. VLDB Endow. 14(12): 2827-2830 (2021).

[12] Bijan Rad, Fei Song, Vincent Jacob and Yanlei Diao. "Explainable Anomaly Detection on High-Dimensional Time Series Data." Proc. of ACM International Conference on Distributed and Event-based Systems (DEBS), pp. 2-14, 2021.

[13] Fei Song, Yanlei Diao, Jesse Read, Arnaud Stiegler, Albert Bifet: EXAD: A System for Explainable Anomaly Detection on Big Data Traces. ICDM Workshops 2018: 1435-1440

[14] Fei Song, Boyao Zhou, Quan Sun, Wang Sun, Shiwen Xia, and Yanlei Diao. "Anomaly Detection and Explanation Discovery on Event Streams." In Proceedings of the International Workshop on Real-time Business Intelligence and Analytics (BIRTE), VLDB 2018. Position paper, 5 pages.

[15] Haopeng Zhang, Yanlei Diao, Alexandra Meliou: EXstream: Explaining Anomalies in Event Stream Monitoring. In Proceedings of the 20th International Conference on Extending Database Technology (EDBT), pp. 156-167, 2017. (21.8% acceptance rate)

Complex Event Processing on Data Streams

[16] Haopeng Zhang, Yanlei Diao, and Neil Immerman. "On Complexity and Optimization of Expensive Queries in Complex Event Processing." In Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD), 217-228, 2014. (25% acceptance rate)

[17] Haopeng Zhang, Yanlei Diao, and Neil Immerman. "Recognizing Patterns in Streams with Imprecise Timestamps." Information Systems, *Elsevier*, 38(3), 1187-1211, November 2013.

[18] Haopeng Zhang, Yanlei Diao, and Neil Immerman. "Recognizing Patterns in Streams with Imprecise Timestamps." Journal "Proceedings of the VLDB Endowment" (PVLDB), 3(1): 244-255, 2010.

[19] Jagrati Agrawal, Yanlei Diao, Daniel Gyllstrom, and Neil Immerman. "Efficient Pattern Matching over Event Streams". In Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD), pages 147-160, June 2008. (18% acceptance rate)

[20] Eugene Wu, Yanlei Diao, and Shariq Rizvi. "High-Performance Complex Event Processing over Streams." In Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD), pages 407-418, June 2006. (13% acceptance rate). *1148 citations as of 06/2021. Google Scholar listed it a top cited paper published in 2006 and named it a classic in Databases & Information Systems in 2017.*

[21] Daniel Gyllstrom, Jagrati Agrawal, Yanlei Diao, and Neil Immerman. "On Supporting Kleene Closure over Event Streams." In Proceedings of the 24th International Conference on Data Engineering (ICDE), pages 1391-1393, April 2008. (31% acceptance rate)

[22] Daniel Gyllstrom, Eugene Wu, Hee-Jin Chae, Yanlei Diao, Patrick Stahlberg, and Gordon Anderson. "SASE: Complex Event Processing over Streams." System Demo. In Proceedings of the 3rd Biennial Conference on Innovative Data Systems Research (CIDR), pages 407-411, January 2007.

Interactive Data Exploration and Insight Discovery

[23] Efficient Exploration of Interesting Aggregates in RDF Graphs. Yanlei Diao, Paweł Guzewicz, Ioana Manolescu, Mirjana Mazuran. In Proceedings of ACM Conference on Management of Data (SIGMOD), 2021.

[24] Efficient Version Space Algorithms for" Human-in-the-Loop" Model Development.. Luciano Palma, Yanlei Diao, Anna Liu. Technical report, https://hal.inria.fr/hal-03064769/, 2020.

[25] Enhui Huang, Luciano Di Palma, Laurent Cetinsoy, Yanlei Diao, and Anna Liu: AIDEme: An active learning based system for interactive exploration of large datasets. NeurIPS 2019, demonstration track.

[26] Yanlei Diao, Pawel Guzewicz, Ioana Manolescu, Mirjana Mazuran: Spade: A Modular Framework for Analytical Exploration of RDF Graphs. Proc. VLDB Endow. 12(12): 1926-1929 (2019).

[27] Luciano Di Palma, Yanlei Diao, Anna Liu: A Factorized Version Space Algorithm for "Human-In-the-Loop" Data Exploration. ICDM 2019: 1018-1023.

[28] Enhui Huang, Liping Peng, Luciano Di Palma, Ahmed Abdelkafi, Anna Liu, and Yanlei Diao: Optimization for Active Learning-based Interactive Database Exploration. Proc. of the VLDB Endowment (PVLDB), 12(1): 71-84, September 2018.

[29] Wenzhao Liu, Yanlei Diao, and Anna Liu. "An Analysis of Query-Agnostic Sampling for Interactive Data Exploration." Communications in Statistics – Theory and Methods. November 2017.

[30] Olga Papaemmanouil, Yanlei Diao, Kyriaki Dimitriadou, Liping Peng. "Interactive Data Exploration via Machine Learning Models." IEEE Data Eng. Bull., 39(4): 38-49, 2016.

[31] Kyriaki Dimitriadou, Olga Papaemmanouil, and Yanlei Diao. AIDE. "An Active Learning-based Approach for Interactive Data Exploration." IEEE Transactions on Knowledge and Data Engineering (TKDE), 28(11): 2842-2856, 2016.

[32] Yanlei Diao, Kyriaki Dimitriadou, Zhan Li, Wenzhao Liu, Olga Papaemmanouil, Kemi Peng, Liping Peng. "AIDE: An Automatic User Navigation System for Interactive Data Exploration." Proceedings of Very Large Databases (PVLDB), 8(12): 1964-1967, August 2015. (System Demo)

[33] Kyriaki Dimitriadou, Olga Papaemmanouil, and Yanlei Diao. "Explore-by-Example: An Automatic Query Steering Framework for Interactive Data Exploration." In Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD), 517-528, 2014. (25% acceptance rate)

[34] Ugur Cetintemel, Mitch Cherniack, Justin DeBrabant, Yanlei Diao, Kyriaki Dimitriadou, Alex Kalinin, Olga Papaemmanouil, Stan Zdonik. "Query Steering for Interactive Data Exploration." In the 6th Biennial Conference on Innovative Data Systems Research (CIDR), 2013.

[35] Kyriaki Dimitriadou, Olga Papaemmanouil and Yanlei Diao. "Interactive data exploration based on user relevance feedback." In Proceedings of the 30th International Conference on Data Engineering Workshops, ICDE, March 31 - April 4, 2014.

Uncertain Data Management

[36] Liping Peng and Yanlei Diao. "Supporting Data Uncertainty in Array Databases." Proceedings of the ACM International Conference on Management of Data (SIGMOD), June 2015. (26% acceptance rate)

[37] Thanh T.L. Tran, Yanlei Diao, Charles Sutton, and Anna Liu. "Supporting User-Defined Functions on Uncertain Data". Journal "Proceedings of the VLDB Endowment" (PVLDB), 6(6), 469-480, August 2013. (22.7% acceptance rate)

[38] Thanh T. L. Tran, Liping Peng, Yanlei Diao, Andrew McGregor, and Anna Liu. "CLARO: Modeling and Processing Uncertain Data Streams." VLDB Journal, 21(5):651-676, November 2012.

[39] Liping Peng, Yanlei Diao, and Anna Liu. "Optimizing Probabilistic Query Processing on Continuous Uncertain Data". Journal "Proceedings of the VLDB Endowment" (PVLDB), 4(11): 1169-1180, 2011.

[40] Thanh Tran, Andrew McGregor, Yanlei Diao, Liping Peng, and Anna Liu. "Conditioning and Aggregating Uncertain Data Streams: Going Beyond Expectations." Journal "Proceedings of the VLDB Endowment" (PVLDB), 3(1): 1302-1313, 2010.

[41] Thanh Tran, Liping Peng, Boduo Li, Yanlei Diao, and Anna Liu. "PODS: Modeling and Processing of High-Volume Uncertain Data Streams." In Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD), 159-170, June 2010. (20.8% acceptance rate)

[42] Yanlei Diao, Boduo Li, Anna Liu, Liping Peng, Charles Sutton, Thanh Tran, and Michael Zink. "Capturing Uncertainty in High-Volume Stream Processing." In Proceedings of the 4th Biennial Conference on Innovative Data Systems Research (CIDR), January 2009. 11 pages. (27.1% acceptance rate)

Large-Scale XML Message Brokering

[43] Yanlei Diao and Michael J. Franklin. "High-Performance XML Message Brokering." Data Stream Management: Processing High-Speed Data Streams (Garofalakis, Gehrke, and Rastogi, eds.). Springer Data-Centric Systems and Applications Series, November 2010.

[44] Yanlei Diao and Michael J. Franklin. "Publish/Subscribe over Streams." Encyclopaedia of Database Systems. Springer, 2009. 6 pages.

[45] Yanlei Diao and Michael J. Franklin. "XML Publish/Subscribe." Encyclopaedia of Database Systems. Springer, 2009. 6 pages.

[46] Yanlei Diao. "Query Processing for Large-Scale XML Message Brokering." PhD Dissertation, University of California, Berkeley, August 2005. *ACM-SIGMOD Dissertation Award Honorable Mention.*

[47] Yanlei Diao, Shariq Rizvi, and Michael J. Franklin. "Towards an Internet-Scale XML Dissemination Service." In Proceedings of the 30th International Conference on Very Large Data Bases (VLDB), pages 612-623, August 2004. (16% acceptance rate)

[48] Yanlei Diao, Mehmet Altinel, Michael J. Franklin, Hao Zhang, and Peter Fischer. "Path Sharing and Predicate Evaluation for High-Performance XML Filtering." ACM Transactions on Database Systems (TODS) 28(4): 467-516, December 2003. ACM Press, New York, NY.

[49] Yanlei Diao, Daniela Florescu, Donald Kossmann, Michael J. Carey, and Michael J. Franklin. "Implementing Memoization in a Streaming XQuery Processor." In Proceedings of the 2nd International XML Database Symposium (XSym), pages 35-50, August 2004. (25% acceptance rate)

[50] Yanlei Diao and Michael J. Franklin. "Query Processing for High-Volume XML Message Brokering." In Proceedings of the 29th International Conference on Very Large Data Bases (VLDB), pages 261-272, September 2003. (14% acceptance rate)

[51] Yanlei Diao, Peter Fischer, Michael J. Franklin, and Raymond To. "YFilter: Efficient and Scalable Filtering of XML Documents." System Demo. In Proceedings of the 18th International Conference on Data Engineering (ICDE), pages 341-342, February 2002.

[52] Yanlei Diao and Michael J. Franklin. "High-Performance XML Filtering: An Overview of YFilter." IEEE Data Engineering Bulletin 26(1): 41-48, March, 2003.

III. Big Data: Application Domains

RFID/Sensor Networks

[53] Yanming Nie, Richard Cocci, Zhao Cao, Yanlei Diao, and Prashant Shenoy. "SPIRE: Efficient Inference and Compression over RFID Streams." IEEE Transactions on Knowledge and Data Engineering (TKDE), 24(1): 141-155, January 2012.

[54] Zhao Cao, Charles Sutton, Yanlei Diao, and Prashant Shenoy. "Distributed Inference and Query Processing for RFID Tracking and Monitoring." Journal "Proceedings of the VLDB Endowment" (PVLDB), 4(5): 326-337, 2011.

[55] Devesh Agrawal, Boduo Li, Zhao Cao, Deepak Ganesan, Yanlei Diao, Prashant Shenoy. "Exploiting the Interplay Between Memory and Flash Storage In Embedded Sensor Devices." In Proceedings of the IEEE International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA), August 2010. 10 pages. (23% acceptance rate)

[56] Thanh Tran, Charles Sutton, Richard Cocci, Yanming Nie, Yanlei Diao, and Prashant Shenoy. "Probabilistic Inference over RFID Streams in Mobile Environments." In Proceedings of the 25th International Conference on Data Engineering (ICDE), pages 1096-1107, March 2009. (17% acceptance rate)

[57] Yanlei Diao, Deepak Ganesan, Gaurav Mathur, and Prashant Shenoy. "Re-thinking Data Management for Storage-centric Sensor Networks." In Proceedings of the 3rd Biennial Conference on Innovative Data Systems Research (CIDR), pages 22-31, January 2007. (44% acceptance rate)

[58] Zhao Cao, Yanlei Diao, and Prashant Shenoy. "Architectural Considerations for Distributed RFID Tracking and Monitoring." In Proceedings of the 5th International Workshop on Networking Meets Databases (NetDB), October 2009. 6 pages.

[59] Richard Cocci, Thanh Tran, Yanlei Diao, and Prashant Shenoy. "Efficient Data Interpretation and Compression over RFID Streams." In Proceedings of the 24th International Conference on Data Engineering (ICDE), pages 1445-1447, April 2008. (31% acceptance rate)

[60] Richard Cocci, Yanlei Diao, and Prashant Shenoy. "SPIRE: Scalable Processing of RFID Event Streams." In Proceedings of the 5th RFID Academic Convocation, April 2007. 6 pages.

High-Performance Genomic Data Processing

[61] Abhishek Roy, Yanlei Diao, Uday Evani, Avinash Abhyankar, Clinton Howarth, Rémi Le Priol, Toby Bloom. "Massively Parallel Processing of Whole Genome Sequence Data: An In-Depth Performance Study." In Proceedings of the ACM SIGMOD Conference, 187-202, 2017. (19% acceptance rate)

[62] Abhishek Roy, Yanlei Diao, and Toby Bloom. "Building Highly-Optimized, Low-Latency Pipelines for Genomic Data Analysis." In Proceedings of the 7th Biennial Conference on Innovative Data Systems Research (CIDR), January 2015. 12 pages. (33% acceptance rate)

[63] Abhishek Roy, Yanlei Diao, Evan Mauceli, Yiping Shen, Bai-Lin Wu. "Massive Genomic Data Processing and Deep Analysis." Proceedings of the VLDB Endowment (PVLDB), 5(12): 1906-1909, 2012. (System Demo.)

IV. Computer Networks

[64] Fang Yu, Yanlei Diao, Randy Katz, and T. V. Lakshman. "Fast Packet Pattern Matching Algorithms." Algorithms for Next Generation Network Architecture (Graham Cormode and Marina Thottan, eds.). Springer Computer Communications and Networks Series, October 2009.

[65] Fang Yu, Zhifeng Chen, Yanlei Diao, T.V. Lakshman, and Randy H. Katz. "Fast and Memory-Efficient Regular Expression Matching." ACM/IEEE Symposium on Architectures for Networking and Communications Systems (ANCS), pages 93-102, December 2006. (19% acceptance rate)

V. Information Retrieval and Data Mining

[66] Michael Bendersky, W. Bruce Croft, and Yanlei Diao. "Quality-Biased Ranking of Web Documents". In Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM), 10 pages, 2011. (8.3% for oral presentation)

[67] Desislava Petkova, W. Bruce Croft, and Yanlei Diao. "Refining Keyword Queries for XML Retrieval by Combining Content and Structure." In Proceedings of 31st European Conference on Information Retrieval (ECIR), pages 662-669, April 2009. (31.9% acceptance rate)

[68] Yanlei Diao, Hongjun Lu, Songting Chen, and Zengping Tian. "Toward Learning Based Web Query Processing." In Proceedings of the 26th International Conference on Very Large Data Bases (VLDB), pages 317-328, September 2000. (15% acceptance rate)

[69] Yanlei Diao, Hongjun Lu, and Dekai Wu. "A Comparative Study of Classification Based Personal E-mail Filtering." In Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), pages 408-419, April 2000. (29% acceptance rate)

[70] Songting Chen, Yanlei Diao, Hongjun Lu and Zengping Tian. "FACT: A Learning Based Web Query Processing System." System Demo. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data (SIGMOD), page 587, May 2000.

VI. Software Engineering

[71] Kaituo Li, Christoph Reichenbach, Yannis Smaragdakis, Yanlei Diao, and Christoph Csallner. "SEDGE: Symbolic Example Data Generation for Dataflow Programs." In the Proceedings of the 28th IEEE/ACM International Conference on Automated Software Engineering (ASE), 235-245, November 2013. (13.6% acceptance rate)

Research Grants

European Funding

- Yanlei Diao (PI). "Charting a New Horizon of Big and Fast Data Analysis through Integrated Algorithm Design," **European Research Council (ERC) Consolidator Award**, €2.5 million, 09/2017-10/2023.

  (The process included two written proposals, two phases of peer reviews, and an onsite interview with a panel of 19 senior computer scientists. *13.8% acceptance rate, ranked #1 by the panel.*)

- Yanlei Diao (PI). "A New Database Service for Interactive Exploration on Big Data," **Agence Nationale de la Recherche (ANR)**, €300,000 10/2016-09/2020.

- Yanlei Diao (PI). Université de Paris Saclay, Chaire IDEX Big Data pour l'innovation, €505,000, 09/2015-08/2021.

US Federal, State, and University Funding

- Yanlei Diao (PI) and Toby Bloom. "Collaborative Research: ABI Development: A New Platform for Highly-Optimized, Low-Latency Pipelines for Genomic Data Analysis." **National Science Foundation**, DBI-1356486, $865,382.00, 09/15/2014-08/31/2019.

- Christopher Hill (PI), Prashant Shenoy, Yanlei Diao, with other Co-PIs. "Commonwealth Computing Cloud for Data Driven Biology (C3DDB)." Mass Life Sciences Center, $4.5 million, for purchasing compute clusters at MGHPCC for data-intensive computing in life sciences.

- Prashant Shenoy (PI) and Yanlei Diao. "Integrating Inference and Complex Event Processing for RFID Logistics Management." **National Aeronautics and Space Administration (NASA)**, $225,000, 02/2015-01/2017.

- Yanlei Diao (PI) and Anna Liu. "III: Small: High-Performance Complex Processing of Continuous Uncertain Data." **National Science Foundation**, IIS-1218524, $495,961, 09/01/2012-08/31/2017.
  - REU Supplement, $16,000, 06/2013-08/2015.

- Yanlei Diao (PI). "CAREER: Efficient, Robust RFID Stream Processing for Tracking and Monitoring." **National Science Foundation** IIS-0746939, $588,879, 06/2008-05/2014.
  - REU Supplement, $16,000, 07/02/2009-06/30/2010.

- Kevin Fu (PI), Yanlei Diao (co-PI), Charles Ross, Deepak Ganesan, Wayne Burleson. "MRI: Acquisition of an RFID Testbed Using Renewable Energy for Object Identification and Habitat Monitoring." **National Science Foundation** CNS-0923313, $450,010, 10/2009-09/2012.
- Yanlei Diao (PI) and Anna Liu. "III-COR-small: Capturing Data Uncertainty in High-Volume Stream Processing." **National Science Foundation** IIS-0812347, $441,594, 09/2008-08/2012.
    - REU Supplement, $16,000, 06/2010-05/2011.
- Deepak Ganesan (PI), Yanlei Diao (co-PI), Prashant Shenoy. "NeTS-NOSS: STONES: Storage-centric Networked Embedded Systems." National Science Foundation CNS-0626873, $500,000, 09/2006-08/2010.
- Yanlei Diao (PI) and Prashant Shenoy. "Big Data Informatics Initiative: Platforms, Algorithms, and Applications for Real-time Big Data Analytics." **UMass President's Science & Technology Initiatives Fund Award**, $167,500.
- Yanlei Diao (PI), Li-Jun Ma, Samuel Madden, and Yiping Shen. "A Cloud Service for Massive Scale Genomic Data Processing and Deep Analysis." **Massachusetts Green High Performance Computing Center (MGHPCC) Seed Fund**, $144,000 and $94,000 for UMass Amherst.

INDUSTRY FUNDING AND OTHERS

- Yanlei Diao. "Explainable Anomaly Detection on Streams of SWIFT Messages." **SWIFT Research Award**, $50,000, University of Massachusetts Amherst.
- Yanlei Diao and Peter Haas. "Strategic Maintenance of Machine Learning Models Under Data Drift." **Adobe Research Award**, $45,000, University of Massachusetts Amherst.
- Yanlei Diao (PI). "Intelligent Resource Management via Deep Learning." **Alibaba Innovative Research** (AIR), $100,000, 2019.
- Yanlei Diao (PI). "Boosting Big Data Analysis with New Modeling and Optimization Methods." **Huawei HIRP Open 2018**, $70,000, Ecole Polytechnique.
- Yanlei Diao (PI). "Deep Learning based Anomaly Detection and Explanation Discovery." **Huawei HIRP Open 2017**, $67,000, Ecole Polytechnique.
- Huan Li (PI) and Yanlei Diao. "Flash-based energy efficient storage and real- time processing in WSN." **National Natural Science Foundation of China** (NSFC), Grant No: 61170293. $94,486.00. 1/1/2012 - 12/31/2015.
- Gerome Miklau (PI), Yanlei Diao, and Evimaria Terzi. "A Platform for Data-Intensive Cybersecurity Monitoring." **Advanced Cybersecurity Center**, $75,000 in total and $50,000 for UMass Amherst.
- Yanlei Diao (PI). "Streaming Analytics in Large-Scale Monitoring Applications." **IBM Scalable Data Analytics for A Smarter Planet Innovation Award**, $20,000.
- Yanlei Diao (PI) and Andrew McGregor. "Data Analytics in the Cloud: Exact Answers Fast, Approximate Answers Faster." **NEC Labs Research Award**, $40,000.
- Yanlei Diao (PI) and Prashant Shenoy. "Streamlining Large-Scale Data Analysis: Exact Answers Fast, Approximate Answers Faster." **Google Research Award**, $50,000.
- Yanlei Diao (PI). "In-Network Event Processing over Distributed Streams." **Cisco Research Gift**, $89,916.

SOFTWARE

MASS_SCALLA 0.1 (http://scalla.cs.umass.edu/): A scalable, low-latency analytics platform that fundamentally transforms the batch-oriented MapReduce cluster computing paradigm into an incremental parallel processing paradigm, and further extends to near real-time analytics.

SASE 1.0 (http://sase.cs.umass.edu/): A stream processing system developed at the University of Massachusetts Amherst that provides fast pattern matching over event streams. The release of SASE 1.0 contains the core stream processing engine of the SASE system.

YFilter 1.0 (http://yfilter.cs.umass.edu/): A high-performance filtering system that allows users or applications to submit queries to be continuously executed against streaming XML messages. This release supports queries written in a subset of XPath 1.0, and has received over 1000 downloads.

INVITED TALKS AND KEYNOTES

- UDAO: A Multi-Objective Optimizer for Cloud Data Analytics via Large-Scale Machine Learning

  Distinguished Lecture, DATA Lab, Northeastern University, April 27, 2022
  Distinguished Lecture, Max Planck Institut (MPI) Informatik, December 16, 2021

- Proactive Monitoring on High-Volume Event Streams through Large-Scale Machine Learning

  Keynote, 15th ACM International Conference on distributed and event-based systems (DEBS), 2021

- Model Learning and Explanation Discovery for Exploring Large Datasets

  Distinguished Lecture, IBM Almaden Research Center, San Jose, April 30, 2021
  Distinguished Lecture, Naver Research Labs, Grenoble, France, January 13, 2020

- Towards a Unified Data Analytics Optimizer

  Spark Summit, London, October 4, 2018

- Boosting Big Data Analytics with New Modeling and Optimization Methods

  MIT, April 24, 2018
  DataBricks, San Francesco, April 25, 2018
  Huawei American Research Center, April 26, 2018

- Big and Fast Data Analytics through Integrated Algorithm Design

  Huawei European Research Symposium, January 20, 2018
  Distinguished Lecture, Technische Universitaet Darmstadt, November 30, 2017
  University of Grenoble, September 29, 2017

- "Interactive Data Exploration via ML/DB Co-Design"

Keynote, 2nd International Workshop on Exploratory Search in Databases and the Wed (ExploreDB), co-located with SIGMOD 2015

- "Scalable, Low-Latency Data Analytics and its Applications"

  Telecom ParisTech, January 2015;
  ETH Zurich, October 2014; Macau University, October 2014; Google, April 11 2014;
  Google September 2013; China East Normal, June 2013; Beihuang University, January 2013;
  INRIA Research Center, December 2012;
  Hong Kong University of Science and Technology, Hong Kong Baptist University, January 2012;
  Fudan University, October 2011; Facebook October 2010

- "Addressing New Challenges in Data Stream Processing"

  University of Wisconsin Madison, September 1, 2010;
  AT&T Research Center, August 3 2010;
  IBM Watson Research Center, August 2 2010;
  Microsoft Research Silicon Valley, May 11 2010;
  Yahoo! Research, May 10 2010;
  Portland State University, May 7 2010;
  Massachusetts Institute of Technology, April 29 2010;
  University of Pittsburgh and Carnegie Mellon University, April 19 2010;
  University of California San Diego, April 14 2010;
  University of California Los Angeles, November 17 2009;
  University of California Irvine, November 16 2009;
  University of California Berkeley, November 13 2009;
  IBM Almaden Research Center, July 28 2009;
  HP Labs, July 27 2009

- "An Optimized Index Structure for Flash Devices"

  Non-volatile Memories Workshop, April 13 2010

- "SASE+: Expressing and Processing Complex Event Patterns over Streams"

  Worcester Polytechnic Institute, Department Colloquium, April 25, 2008;
  New England Database Day, Cambridge, MA, February 4, 2008;
  Renming University and Beihuang University, Beijing, China, January 4, 2008;
  Microsoft Research Center, Redmond, WA, October 12, 2007;
  StreamBase Systems, Lexington, Massachusetts, February 16, 2007

- "Query Processing for Large-Scale XML Message Brokering"

  Tsinghua University and Peking University, September 5-6, 2006;
  Distinguished Faculty Lecture, University of Texas at Austin, December 12-14, 2005;
  AT&T Research Lab, December 15, 2005

- "XML Filtering, Transformation, and Routing with YFilter"

  Cisco Systems, January 5, 2007;
  IBM Almaden Research Center, September 23, 2004;
  Oracle Corporation, December 9, 2003;
  BEA Systems, May 22, 2003

TEACHING

Instructor **University of Massachusetts Amherst**

Information Systems (CMPSCI 445): F2007, S2008, F2009, F2012, S2013, S2015
Database Design and Implementation (CMPSCI 645): S2006, S2007, S2009, S2011, S2018, S2020, S2021
Advanced Topics in Database Systems (CMPSCI 745): F2008, F2010, F2013
Hot Topics in Databases (CMPSCI 691TD): S2009, S2010
Advanced Database Systems (CMPSCI 691LL): F2006

Instructor **Ecole Polytechnique**

Systems for Big Data (M1), Spring 2017, Winter 2018
Systems for Big Data Analytics (M2), Spring 2016, Fall 2016, Fall 2017

UNIVERSITY SERVICE

- **Member**
  Faculty Recruiting committee, 2008-2009, 2010-2011, 2011-2012, 2013-2014, 2014-2015, 2019-2020
  Personnel Committee, 2006-2007, 2010-2011
  Undergraduate Recruiting Committee, 2009-2010, 2011-2012
  Diversity/Outreach Committee, 2005-2008, 2010-2011, 2012-2013, 2013-2014, 2014-2015
  Graduate Program Committee, 2007-2008, 2009-2010, 2012-2013
  Graduate Admissions Committee, 2005-2006, 2011-2012, 2020-2021

- **Speaker**
  Professionalism Seminar, Fall 2005, Spring 2018
  Lab Description Seminar, Fall 2005, Fall 2006, Fall 2007, Fall 2009
  CS Women Luncheon Meeting, Fall 2005

ADVISING

- **Current PhD Students and Postdocs**

Sein Minn	(Postdoc, 02/2022-, Ecole Polytechnique)
Fei Song	(Postdoc, 01/2017-, Ecole Polytechnique)
Abhishek Roy	(PhD, 09/2013-, UMass Amherst)
Chenghao Lyu	(PhD, 09/2018-, UMass Amherst)
Qi Fan	(PhD, 12/2019-, Ecole Polytechnique)
Vincent Jacob	(PhD, 12/2019-, Ecole Polytechnique)

- **Past Students**

Pawel Guzewicz	PhD, 10/2021, Ecole Polytechnique / Inria
PhD Thesis:	*EXPRALYTICS: Expressive and Efficient Analytics for RDF Graphs*

Luciano Di Palma PhD Thesis:	PhD, July 2021, Ecole Polytechnique *New Algorithms and Optimizations for Human-in-the-Loop Model Development*
Enhui Huang PhD Thesis:	PhD, July 2021, Ecole Polytechnique *Active Learning Methods for Interactive Exploration on Large Databases*
Khaled Zaouk PhD Thesis:	PhD, 02/2021, Ecole Polytechnique *Neural-Based Modeling for Performance Tuning of Cloud Data Analytics*
Liping Peng PhD Thesis:	PhD, December 2017, UMass Amherst. Netflix *Supporting Scientific Analytics under Data Uncertainty and Query Uncertainty.*
Haopeng Zhao PhD Thesis	PhD, May 2017, UMass Amherst. Google *High-Performance Complex Event Processing for Decision Analytics.*
Boduo Li PhD Thesis:	PhD, May 2015, UMass Amherst. Research Scientist, Facebook *A Platform for Scalable Low-Latency Analytics using MapReduce*. *2011 Facebook Fellowship Runner-up*
Thanh Tran PhD Thesis: MS Project:	PhD, May 2013, UMass Amherst. Data Scientist, Twitter *High-Performance Processing of Continuous Uncertain Data*. *Probabilistic inference for RFID-based object tracking and monitoring.*
Zhao Cao PhD Thesis:	PhD, December 2011. Director, Huawei *Distributed Inference and Query Processing for RFID Tracking and Monitoring*
Wenzhao Liu MS Project:	MS, 2016, UMass Amherst. Microsoft *An Analysis of Query-Agnostic Sampling for Interactive Data Exploration*
Yunmeng Ban	MS, 2014, UMass Amherst. Facebook *2014 Microsoft Research Graduate Women's Scholarship*
Edward Mazur MS Project:	MS, 2011, UMass Amherst. Google Inc. *Towards Scalable One-Pass Analytics using MapReduce*.
Richard Cocci MS Project:	MS, 2008, UMass Amherst. Harvard Law School *Efficient Data Interpretation and Compression over RFID Streams*.
Ravishankar Rajamony MS Project:	MS, 2008, UMass Amherst. Goldman Sachs *Improved Memory Management in XML Data Stream Processing*.
Daniel Gyllstrom MS Project:	MS, 2007, UMass Amherst *On Supporting Kleene Closure over Event Streams*.
Francesco Pierri	MS, 2018, Ecole Polytechnique
Alexander Sevin	MS, 2017, Ecole Polytechnique
Shu Shang	MS, 2017, Ecole Polytechnique and Inria
Zheng Zhang	MS, 2017, Ecole Polytechnique and Inria
Arnaud Stiegler	Summer intern, 2018, Ecole Polytechnique and UMass Amherst
Magdalena Matczak	Summer intern, 2017, Ecole Polytechnique and Columbia University
Remi le Priol	Summer intern, 2017, Ecole Polytechnique and New York Genome Center

- **Visiting Students**
  Zhao Cao (2008-2010), IBM Research, China
  Yanming Nie (2007-2008), Northwestern Polytechnical University, China

- **External PhD Dissertation Committees**
  Petro Holanda, Universiteit Leiden
  Julien Pilourdault (September 2017), Université de Grenoble
  Alper Okcan (2014), Northeastern University
  Asterios Katsifodimos (2013), INRIA, France
  Xiaoyong Liu (2006), Jiwoon Jeon (2006), University of Massachusetts Amherst
  Mo Liu (2009-2010), Worcester Polytechnic Institute

- **Graduate Student Researchers**
  Ramya Sarma (Summer 2018), Shubham Mukherjee (Fall 2017-Spring 2018), Junghee Jo (Fall 2008-Spring 2009), Jagrati Agrawal (Spring 2007-Spring 2008), Hee-Jin Chae (Summer 2006)

- **Undergraduate Student Researchers**
  Michael Rabie (Spring 2015), Kevin Gurney (Spring 2015), Thai Nguyen (Spring 2015), Brian Staplegon (Spring 2014), Timothy Allman (Spring 2014), Albert Williams (Spring 2014), Longhao Piao (Spring 2014), Rebecca Bryan (2013), Said Mastawi (2013), Yang Tang (2013), Sofya Vorotnikova (2012), DiHuynh (2012), Andrew Huang (Spring 2008), Jacob Mitchell (Fall 2009), Alex Jackson (Spring 2010), Eugene Wu (Fall 2006)