Lei Cao

lcao@csail.mit.edu
http://people.csail.mit.edu/lcao/
32 Vassar Street, G886, Cambridge, MA 02139, USA

Research Interests

Data Systems: Privacy-preserving Data Systems, Machine Learning for Systems, Cloud Database, Streaming Database, Data Integration and Cleaning, Distributed OLTP, Query Optimization
Data Science: Anomaly Detection, Intelligent Data Systems, Big Data Analytics, Interpretable Machine Learning

Education

Massachusetts Institute of Technology (MIT) Cambridge, MA
Postdoc Associate in Computer Science Nov. 2016 – Jan. 2021

- Research Direction: Data Systems/Data Science
- Advisor: Prof. Samuel Madden

Worcester Polytechnic Institute (WPI) Worcester, MA
Ph.D. in Computer Science Sep. 2010 – Mar. 2016

- Research Direction: Data Management/Big Data Analytics
- Thesis: Outlier Detection in Big Data
- Advisor: Prof. Elke Rundensteiner

Employment

Massachusetts Institute of Technology (MIT) Cambridge, MA
Research Scientist. Jan. 2021 – Now
Postdoc Associate. Supervisors: Prof. Sam Madden and Prof. Mike Stonebraker Nov. 2016 – Jan. 2021

IBM T.J. Watson Research Center Yorktown Heights, NY
Research Staff Member. Oct. 2015 – Nov. 2016

Research Experience

Data Systems Group, MIT Cambridge, MA
Research Scientist. Jan. 2021 – Now
Postdoc Associate. Supervisors: Prof. Sam Madden and Prof. Mike Stonebraker Nov. 2016 – Jan. 2021

- Making database differentially private and faster with accuracy guarantee
- Worked on a scalable distributed OLTP database
- Developed an end-to-end anomaly detection system; used by **Facebook**
- Designed a system supporting the analytics of IoT sequence data; used by **Philips Lighting**
- Designed a system to support machine learning on big EEG data; used by **Mass General Hospital**
- Proposed an image classification model to effectively reject out-of-distribution objects at inference
- Proposed a continuous similarity search paradigm adaptive to human feedback
- Proposed a deep context-aware model enforcing the semantics context constraints in object detection

Database System Research Group, WPI Worcester, MA
Graduate Research Assistant. Supervisor: Prof. Elke Rundensteiner Sep. 2010 – Mar. 2016

- Proposed a scalable streaming anomaly detection framework
- Proposed new semantics and scalable algorithms for detecting anomalies from trajectory data
- Designed a distributed processing paradigm scaling anomaly detection algorithms to big data
- Designed an online system to solve the parameter tuning problem in unsupervised machine learning
- Developed a high-performance stream query engine
- Proposed scalable techniques to support aggregation in complex event processing (CEP)

IBM Research AI, Blockchain, and Quantum Solutions Yorktown Heights, NY

Research Staff Member. Supervisor: Dr. Xuan Liu Oct. 2015– Nov. 2016

- Cognitive supplier chain: used machine learning to optimize shipping and inventory management
- Data science for social good: used data science to measure economic competitiveness

IBM Research AI, Blockchain, and Quantum Solutions Yorktown Heights, NY

Research Intern. Supervisor: Dr. Chandrasekhar Narayanaswami May. 2014 – May. 2015

- Studied how local events influence the sales of grocery stores

Honors and Awards

SIGMOD 2016 Student Travel Award Jun. 2016

Sharing-Aware Outlier Analytics over High-Volume Data Streams

KDD 2015 Student Travel Award Aug. 2015

Online Outlier Exploration Over Large Datasets

SIGMOD 2014 Student Travel Award Jun. 2014

Complex Event Analytics: Online Aggregation of Stream Sequence Patterns

VLDB 2014 Student Travel Fund Aug. 2014

High Performance Stream Query Processing With Correlation-Aware Partitioning

ICDE 2014 Student Travel Scholarship Apr. 2014

Distance-Based Outlier Detection over High-Volume Data Streams

Teaching and Mentoring

CS3431 Database Systems, WPI Jan. – Mar. 2011

Teaching Assistant

- Had office hours, held lab sessions, graded homework, projects, and exams.
- Rating: Student rating: excellent, faculty rating: excellent.

CS4516 Advanced Computer Networks, WPI Oct. – Dec. 2010

Teaching Assistant

- Had office hours, held lab sessions, graded homework, projects, and exams.
- Rating: Student rating: excellent, faculty rating: excellent.

CS3013 Operating Systems, WPI Sep. – Oct. 2010

Teaching Assistant

- Had office hours, held lab sessions, graded homework, projects, and exams.
- Rating: Student rating: excellent, faculty rating: excellent.

WPI MQP Program Sep. – Dec. 2013

Research Mentor

Supervised five undergraduate students

- Developed an infection control system used by UMASS Memorial Hospital.

Invited Talks

SAUL: Towards Effective Data Science

- University of Michigan, April 2021
- University of Maryland, March 2021
- National University of Singapore, February 2021

Toward An End-to-End Anomaly Detection System

- Google Research, July 2020
- UC Irvine, April 2020
- Purdue University, April 2020
- Georgia Institute of Technology, April 2020
- UCLA, April 2020
- Northwestern University, April 2020

- University of Maryland, March 2020
- University of Arizona, March 2020
- CSAIL-MSR Trustworthy AI collaboration, MIT, February 2019
- FinTech@CSAIL, MIT, August 2018
- Signify Research Cambridge, July 2018
- CSAIL Alliances Annual meeting, June 2018
- North East Database Day (NEDB), MIT, January 2018

Taming the Ictal-interictal-injury Continuum - Visualizing & Labeling 30TB of EEG
- Massachusetts General Hospital (MGH)/Harvard Medical School, January 2019
- Google Cambridge, August 2018

Detecting Anomalies from IoT Sequence Data
- Signify Research Cambridge, July 2017
- Stanford University, January 2017
- North East Database Day (NEDB), MIT, January 2017

Outlier Detection in Big Data
- Brown University, June 2016
- Alibaba Seattle, August 2015
- IBM T.J. Watson Research Center, Yorktown Heights, February 2015
- Alibaba Hangzhou, August 2014

Professional Service

Program Committee:

- Information System Area Editor	2021–2024
- SIGMOD Proceeding Chair	2021
- VLDB	2023, 2021, 2020 (Session Chair)
- SIGMOD	2019
- SIGKDD	2022, 2021, 2020, 2019
- ICDE	2023, 2022, 2020, 2019, 2018, 2017
- EDBT	2023
- CIKM	2022, 2021, 2019, 2018
- DASFAA	2022, 2021, 2020, 2019
- VLDB Demo	2019
- IEEE Big data	2022, 2021, 2020, 2019, 2018
- WSDM	2022
- SDM	2022

Reviewer for:

- TODS	2019
- TKDE	2022, 2020, 2019, 2018, 2017, 2016, 2015
- VLDBJ	2022, 2020, 2019, 2018
- Artificial Intelligence	2019
- TKDD	2019, 2018
- SIGMOD	2017, 2016, 2015, 2013, 2012
- VLDB	2017, 2016, 2015, 2013, 2012
- ICDE	2014
- EDBT	2014, 2013

Grant Writing

NSF CSSI (Award#2103832)
- Title: A Self-tuning Anomaly Detection Service
- PIs: Samuel Madden, Elke Rundensteiner
- My Contributions: the content is based on my research; responsible for 90% of the writing
- Result: granted $590,000 for 2021 – 2024

NSF IIS (Award#1910880)
- Title: Outlier Discovery Paradigm
- PI: Elke Rundensteiner
- My Contributions: the content is based on my research; responsible for 90% of the proposal
- Result: granted $499,558 for 2019 – 2022

NSF IIS (Award#1815866)
- Title: Scalable Event Trend Analytics For Data Stream Inquiry
- PI: Elke Rundensteiner
- My Contributions: drafting, editing, and reviewing the proposal
- Result: granted $515,753 for 2018 – 2021

Papers in Preparation

I7. Binwei Yan, **Lei Cao**, Nan Tang and Samuel Madden *The Revisit of Data Cleaning on Machine Learning*, In preparation.

I6. **Lei Cao**, Nan Tang, and Samuel Madden *Query in the Wild: NLP on Data Lake*, In preparation.

I5. Ruoshan Lan, **Lei Cao** and Samuel Madden *The Civilization of IoT Sequence Data*, In preparation.

I4. **Lei Cao**, Haibo Xiu and Samuel Madden *Clustering High Dimensional Data via Graph Embedding*, In preparation.

I3. Haibo Xiu, Jiachen Liu, **Lei Cao** and Samuel Madden *Making Product Quantization Work in Dynamic Data*, In preparation.

I2. Christos Chachamis, **Lei Cao** and Samuel Madden *Learning a High Dimensional Index*, In preparation.

I1. Yizhou Yan*, **Lei Cao***, Samuel Madden, and Elke Rundensteiner *Context-Aware Object Detection With Convolutional Neural Networks*, In preparation (*Equal Contribution).

Papers under Review

U5. Yu Wang, **Lei Cao** and Samuel Madden *Interpretable Outlier Summarization*, Submitted to **SIGMOD2023**.

U4. Jiaming Liang, **Lei Cao** and Samuel Madden *RITA: Group Attention is All You Need*, Submitted to **SIGMOD2023**.

U3. **Lei Cao**, Yizhou Yan, Harihar Subramanyam, Samuel Madden, and Elke Rundensteiner *An End-to-end Anomaly Discovery System*, Submitted to **VLDB2023**.

U2. **Lei Cao**, Yizhou Yan, Samuel Madden, and Elke Rundensteiner *AutoOD: Automatic Outlier Detection*, **SIGMOD 2023**, under revision.

U1. **Lei Cao**, Yizhou Yan, Samuel Madden, and Elke Rundensteiner *ASSET: A System for Exploring Sequential Patterns*, Submitted to **VLDB2023**.

Journal Publications

J3. Caitlin Kuhlman, Karthikeyan Natesan Ramamurthy, Prasanna Sattigeri, Aurlie C Lozano, **Lei Cao**, Chandra Reddy, Aleksandra Mojsilović, Kush R Varshney, *How to Foster Innovation: a Data-driven Approach to Measuring Economic Competitiveness*, IBM Journal of Research and Development, Volume 61, Iss. 6, November 2017.

J2. Yanwei Yu, **Lei Cao***, Elke A Rundensteiner, Qin Wang, *Outlier Detection over Massive-scale Trajectory Streams*, ACM Transactions on Database Systems (***TODS***), Volume 42, Iss. 2, June 2017 (*Corresponding Author).

J1. Elke A Rundensteiner, Olga Poppe, Chuan Lei, Medhabi Ray, **Lei Cao**, Yingmei Qi, Mo Liu, Di Wang, *Exploiting Sharing Opportunities for Real-time Complex Event Analytics*, IEEE Data Engineering Bulletin, Volume 38, Iss. 4, June 2017.

Conference Publications

C32. Dennis Hofmann, Peter Van Nostrand, **Lei Cao**, Samuel Madden, and Elke Rundensteiner *A Demonstration of AutoOD: A Self-Tuning Anomaly Detection System*, ***VLDB*** 2022.

C31. Zhongqiang Gao, Chuanqi Cheng, Yanwei, Yu, **Lei Cao**, Chao Huang, and Junyu Dong *ATLANTIC: Making Database Differentially Private and Faster with Accuracy Guarantee*, ***ICDE*** 2022.

C30. **Lei Cao**, Dongqing Xiao, Yizhou Yan, Samuel Madden, and Guoliang Li *ATLANTIC: Making Database Differentially Private and Faster with Accuracy Guarantee*, ***VLDB*** 2021.

C29. Huayi Zhang, **Lei Cao**, Samuel Madden, and Elke Rundensteiner *ELITE: Robust Deep Anomaly Detection with Meta Gradient*, ***KDD*** 2021.

C28. Huayi Zhang, **Lei Cao**, Elke Rundensteiner, and Samuel Madden *LANCET: Labeling Complex Data at Scale*, ***VLDB*** 2021.

C27. Guoliang Li, Xuanhe Zhou, and **Lei Cao** *AI Meets Database: AI4DB and DB4AI*, ***SIGMOD*** 2021.

C26. Yi Lu, Xiangyao Yu, **Lei Cao**, and Samuel Madden *Epoch-based Commit and Replication in Distributed OLTP Databases*, ***VLDB*** 2021

C25. Yi Lu, Xiangyao Yu, **Lei Cao**, and Samuel Madden *Aria: A Fast and Practical Deterministic OLTP Database*, Proceedings of the ***VLDB*** Endowment, Vol. 13, Iss. 11, August 2020.

C24. Chengliang Chai, **Lei Cao**, Guoliang Li, Jian Li, Yuyu Luo and Samuel Madden *Human-in-the-loop Outlier Detection*, Proceedings of ***SIGMOD***, June 2020.

C23. **Lei Cao**, Huayi Zhang, Yizhou Yan, Elke Rundensteiner, and Samuel Madden *Continuously Adaptive Similarity Search*, Proceedings of ***SIGMOD***, June 2020.

C22. El Kindi Rezig, **Lei Cao**, Giovanni Simonini, Maxime Schoemans, Samuel Madden, Mourad Ouzzani, Nan Tang, and Michael Stonebraker, *Dagger: A Data (not code) Debugger*, Proceeding of the Conference on Innovative Data Systems Research (***CIDR***) 2020.

C21. El Kindi Rezig, **Lei Cao**, Michael Stonebraker, Giovanni Simonini, Wenbo Tao, Samuel Madden, Mourad Ouzzani, Nan Tang, Ahmed K Elmagarmid, *Data Civilizer 2.0: a Holistic Framework for Data Preparation and Analytics*, Proceedings of the ***VLDB*** Endowment, Vol. 12, Iss. 12, August 2019.

C20. **Lei Cao**, Wenbo Tao, Sungtae An, Jing Jin, Yizhou Yan, Xiaoyu Liu, Wendong Ge, Adam Sah, Leilani Battle, Jimeng Sun, Remco Chang, Brandon Westover, Samuel Madden, Michael Stonebraker, *Smile: a System to Support Machine Learning on EEG Data at Scale*, Proceedings of the ***VLDB*** Endowment, Vol. 12, Iss. 12, August 2019.

C19. **Lei Cao**, Yizhou Yan, Samuel Madden, and Elke Rundensteiner, *Efficient discovery of sequence outlier patterns*, Proceedings of the ***VLDB*** Endowment, Vol. 12, Iss. 8, April 2019.

C18. Xiao Qin, **Lei Cao**, Elke Rundensteiner, and Samuel Madden, *Scalable Kernel Density Estimation-based Local Outlier Detection over Large Data Streams*, Processing of ***EDBT***, March 2019.

C17. Yizhou Yan*, **Lei Cao***, Caitlin Kulhman, and Elke Rundensteiner, *SWIFT: Mining Representative Patterns from Large Event Streams*, Proceedings of the ***VLDB*** Endowment, Vol. 12, Iss. 3, November 2018 (*Equal Contribution).

C16. Yizhou Yan, **Lei Cao**, and Elke Rundensteiner, *Distributed Top-N local outlier detection in big data*, Proceedings of ***IEEE Big Data***, December 2017.

C15. Mingrui Wei, **Lei Cao**, Chris Cormier, Hui Zheng, Elke Rundensteiner, *Interactive Analytics System for Exploring Outliers*, Proceedings of ***CIKM***, November 2017.

C14. Caitlin Kulhman, Yizhou Yan, **Lei Cao**, and Elke Rundensteiner, *Pivot-based Distributed K-Nearest Neighbor Mining*, Proceedings of ***ECML PKDD***, September 2017.

C13. Yizhou Yan*, **Lei Cao***, Caitlin Kulhman, and Elke Rundensteiner, *Distributed Local Outlier Detection in Big Data*, Proceedings of ***SIGKDD***, August 2017 (*Equal Contribution).

C12. Yizhou Yan*, **Lei Cao***, and Elke Rundensteiner, *Scalable Top-n Local Outlier Detection*, Proceedings of ***SIGKDD***, August 2017 (*Equal Contribution).

C11. Xiao Qin, Tabassum Kakar, Susmitha Wunnava, Elke A Rundensteiner, and **Lei Cao**, *Maras: Signaling Multi-drug Adverse Reactions*, Proceedings of ***SIGKDD***, August 2017.

C10. Ruoshan Lan, Yanwei Yu, **Lei Cao**, Peng Song, and Yingjie Wang, *Discovering Evolving Moving Object Groups from Massive-scale Trajectory Streams*, Proceedings of ***MDM***, May 2017.

C9. **Lei Cao**, Yizhou Yan, Caitlin Kulhman, Qingyang Wang, and Elke Rundensteiner, *Multi-tactic Distance-based Outlier Detection*, Proceedings of ***ICDE***, April 2017.

C8. **Lei Cao**, Jiayuan Wang, and Elke Rundensteiner, *Sharing-aware Outlier Analytics over High-volume Data Streams*, Proceedings of ***SIGMOD***, June 2016.

C7. **Lei Cao**, Jiayuan Wang, and Elke Rundensteiner, *Multi-query Outlier Detection over Data Streams*, Proceedings of ***DEBS***, June 2016.

C6. **Lei Cao**, Mingrui Wei, Di Yang, and Elke Rundensteiner, *Online Outlier Exploration over Large Datasets*, Proceedings of ***SIGKDD***, August 2015.

C5. Yanwei Yu*, **Lei Cao***, Elke Rundensteiner, and Qin Wang, *Detecting Moving Object Outliers in Massive-scale Trajectory Streams*, Proceedings of ***SIGKDD***, August 2014 (*Equal Contribution).

C4. **Lei Cao**, Qingyang Wang, and Elke Rundensteiner, *Interactive Outlier Exploration in Big Data Streams*, Proceedings of the ***VLDB*** Endowment, Vol. 7, Iss. 13, August 2014.

C3. Yingmei Qi, **Lei Cao**, Medhabi Ray, and Elke A Rundensteiner, *Complex Event Analytics: Online Aggregation of Stream Sequence Patterns*, Proceedings of ***SIGMOD***, June 2014.

C2. **Lei Cao**, Di Yang, Qingyang Wang, Yanwei Yu, Jiayuan Wang, and Elke A Rundensteiner, *Scalable distance-based outlier detection over high-volume data streams*, Proceedings of ***ICDE***, April 2014.

C1. **Lei Cao** and Elke Rundensteiner, *High Performance Stream Query Processing with Correlation-aware Partitioning*, Proceedings of the ***VLDB*** Endowment, Vol. 7, Iss. 4, December 2013.

Reference

Samuel Madden
Professor of EECS
Massachusetts Institute of Technology
madden@csail.mit.edu

Elke Rundensteiner
Professor of Computer Science
Worcester Polytechnic Institute
rundenst@wpi.edu

Michael Stonebraker
Adjunct Professor of EECS
Massachusetts Institute of Technology
stonebraker@csail.mit.edu