Yao LU

yao@nus.edu.sg

My passion lies at the intersection of AI, systems and science. Throughout my academic and industry journey, I have engaged with various tiers of the technology landscape and acquired hands-on experience in data science, core systems, advanced cloud infrastructures, and a diverse range of AI applications.

Working History

2024.8 – **Assistant Professor @ School of Computing, National University of Singapore**
2023.9 – 2024.7 Adjunct Assistant Professor
Research topics: AI + data and cloud systems

2014 – 2024 **Researcher @ Microsoft Research**, Redmond, WA, USA.
Full time ← research contractor ← multiple internships in Data Systems Group and Mobile and Networking Group. Research topics: *Improving data systems for AI* and *Improving data systems using machine learning*

Education

2013 – 2018 **PhD in Computer Science and Engineering**
University of Washington, Seattle, WA
Research area: Data systems for AI
Advisor: Linda Shapiro
Committee member: Magdalena Balazinska, Srikanth Kandula

2010 – 2013 **MSc in Computer Science**
Fudan University, Shanghai, China
Research area: Computer vision and AI

2006 – 2010 **BEng in Computer Science**
Tongji University, Shanghai, China

Publications

Peer-Reviewed Conference Publications

- Xinle Wu, **Yao Lu**. Reward Model Routing in Alignment. International Conference on Learning Representations (ICLR). Rio de Janeiro, Brazil. 2026.
- Yuhao Shen, Junyi Shen, Quan Kong, Tianyu Liu, **Yao Lu**, Cong Wang. Speculative Decoding via Hybrid Drafting and Rollback-Aware Branch Parallelism. International Conference on Learning Representations (ICLR). Rio de Janeiro, Brazil. 2026.
- Yongjun He, Haofeng Yang, **Yao Lu**, Ana Klimovic and Gustavo Alonso Resource Multiplexing in Tuning and Serving Large Language Models. USENIX Annual Technical Conference (ATC). Boston, MA, USA. 2025.

- Yulong Hui, **Yao Lu**, Huanchen Zhang. OkraLong: A Flexible Retrieval-Augmented Framework for Long-Text Question Answering. Conference on Empirical Methods in Natural Language Processing (EMNLP) Findings. Suzhou, China. 2025.
- Chenxia Han, Chaokun Chang, Srijan Srivastava, **Yao Lu**, Eric Lo. Scalable Complex Event Processing on Video Streams. ACM International Conference on Management of Data (SIGMOD). Berlin, Germany. 2025.
- Hui Yulong, **Yao Lu**, Huanchen Zhang. UDA: A Benchmark Suite for Retrieval Augmented Generation in Real-world Document Analysis. In Proceeding of the Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS), 2024.
- Yongji Wu, Matt Lentz, Danyang Zhuo, **Yao Lu**. Serving and Optimizing Machine Learning Workflows on Heterogeneous Infrastructures. International Conference on Very Large Data Bases (VLDB). Vancouver, BC, Canada. 2023.
- Beibin Li, **Yao Lu**, Srikanth Kandula. Warper: Efficiently Adapting Learned Cardinality Estimation Models to Data and Workload Drifts. ACM International Conference on Management of Data (SIGMOD). Philadelphia, PA, USA. 2022.
- Pramod Chunduri, Jaeho Bang, **Yao Lu**, Joy Arulraj. Zeus: Efficiently Localizing Actions in Videos using Reinforcement Learning. ACM International Conference on Management of Data (SIGMOD). Philadelphia, PA, USA. 2022.
- Zhihui Yang, Zuozhi Wang, Yicong Huang, Feng Gao, **Yao Lu**, Chen Li, X. Sean Wang. Demonstration of Accelerating Machine Learning Inference Queries with Correlative Proxy Models. International Conference on Very Large Data Bases (VLDB) Demo. Sydney, Australia. 2022.
- Zhihui Yang, Zuozhi Wang, Yicong Huang, **Yao Lu**, Chen Li, X. Sean Wang. Correlative Cascades for Machine Learning Inference. International Conference on Very Large Data Bases (VLDB). Sydney, Australia. 2022.
- **Yao Lu**, Srikanth Kandula, Arnd Christian Konig, Surajit Chaudhuri. Pre-training Summarization Models of Structured Datasets for Cardinality Estimation. International Conference on Very Large Data Bases (VLDB). Sydney, Australia. 2022.
- Kexin Rong, **Yao Lu**, Peter Bailis, Srikanth Kandula, Philip Levis. Approximate Partition Selection for Big-Data Workloads using Summary Statistics. International Conference on Very Large Data Bases (VLDB). Tokyo, Japan. 2020.
- **Yao Lu**, Aakanksha Chowdhery, Srikanth Kandula and Surajit Chaudhuri. Accelerating Machine Learning Inference with Probabilistic Predicates. ACM International Conference on Management of Data (SIGMOD). Houston, TX, USA. 2018. *Course Material in GT8803@GaTech, CS839@UW-Madison, CMPT8343@SFU.*
- **Yao Lu**, Srikanth Kandula and Surajit Chaudhuri. Interactive Demonstration of Probabilistic Predicates. ACM International Conference on Management of Data (SIGMOD) Demo. Houston, TX, USA. 2018. **Best Demonstration Award**.
- Haonan Qiu, Yingbin Zheng, Hao Ye, **Yao Lu**, Feng Wang, Liang He. Precise Temporal Action Localization by Evolving Temporal Proposals. ACM International Conference on Multimedia Retrieval (ICMR). Yokohama, Japan. 2018.
- Siwei Lyu and **Yao Lu** et al. UA-DETRAC 2017: Report of AVSS2017 & IWT4S Challenge on

Advanced Traffic Monitoring. IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). Lecce, Italy. 2017.

- Li Wang, **Yao Lu**, Hong Wang, Yingbin Zheng, Hao Ye and Xiangyang Xue. Evolving Boxes for Fast Vehicle Detection. IEEE International Conference on Multimedia and Expo (ICME). Hongkong, China. 2017. **Platinum Best Paper Award**.
- **Yao Lu** and Linda Shapiro. Closing the Loop for Object Proposals and Edge Detection. The Thirty-First AAAI Conference on Artificial Intelligence (AAAI). San Francisco, CA, USA. 2017.
- **Yao Lu**, Xue Bai, Linda Shapiro, Jue Wang. Coherent Parametric Contours for Interactive Video Object Segmentation. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA. 2016. *Shipped to Adobe After Effects*.
- **Yao Lu**, Aakanksha Chowdhery, Srikanth Kandula. Optasia: A Relational Platform for Efficient Large-Scale Video Analytics. ACM Symposium on Cloud Computing (SoCC). Santa Clara, CA, USA. 2016.
- **Yao Lu**, Wei Zhang, Ke Zhang, Xiangyang Xue. Semantic Context Learning with Large-Scale Weakly-Labeled Image Set. ACM Conference on Information and Knowledge Management (CIKM). Hawaii, HI, USA, 2012.
- **Yao Lu**, Wei Zhang, Chen Jin, Xiangyang Xue. Learning Attention Map from Images. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR). Providence, RI, USA. 2012.
- Wei Zhang, **Yao Lu**, Xiangyang Xue, Jianping Fan. Automatic Image Annotation with Weakly Labeled Datasets. ACM Multimedia. Scottsdale, AZ, USA. 2011.
- Xiangyang Xue, Wei Zhang, Jie Zhang, Bin Wu, Jianping Fan, **Yao Lu**. Correlative Multi-Label Multi-Instance Image Annotation. 13th International Conference on Computer Vision (ICCV). Barcelona, Spain. 2011.
- **Yao Lu**, Wei Zhang, Hong Lu, Xiangyang Xue. Salient Object Detection using Concavity Context. 13th IEEE International Conference on Computer Vision (ICCV). Barcelona, Spain. 2011.

Vision paper

- Lu Y. et al. Computing in the Era of Large Generative Models: From Cloud-Native to AI-Native. arXiv preprint arXiv:2401.12230. 2024.

Patents

- **Yao Lu**, Srikanth Kandula. Adapting Learned Cardinality Estimators to Data and Workload Drifts. US Patent App. #17/566,996.
- Surajit Chaudhuri, Srikanth Kandula, **Yao Lu**. Accelerating Machine Learning Inference with Probabilistic Predicates. US Patent App. #16/003,495.
- Xue Bai, Jue Wang, **Yao Lu**. Flexible Video Object Boundary Tracking. US Patent #9,569,866.

Doctoral Thesis

- Yao Lu. Building and Accelerating a Declarative Platform for Machine Learning Model Serving. Doctoral Dissertation. University of Washington. 2018.

Posters, Workshop Papers and Technical Reports

- Yongjun He, **Yao Lu**, Gustavo Alonso. Deferred continuous batching in resource-efficient large language model serving. In Proceeding of the 4th Workshop on Machine Learning and Systems. 2024.
- Gaurav. Kakkar, et al. EVA: An End-to-End Exploratory Video Analytics System. Proceedings of the 7th Workshop on Data Management for End-to-End Machine Learning. (DEEM). 2023.
- Beibin Li, **Yao Lu**, Chi Wang, Srikanth Kandula. Q-error Bounds of Random Uniform Sampling for Cardinality Estimation. MSR Technical report MSR-TR-2021-29.
- Yao Peng, Hao Ye, Yining Lin, Yixin Bao, Zhijian Zhao, Haonan Qiu, **Yao Lu**, Li Wang, Yingbin Zheng. Large-Scale Video Classification with Elastic Streaming Sequential Data Processing System. ACM Multimedia Workshop on Large-Scale Video Classification Challenge (LSVC). Mountain View, USA. 2017.
- **Yao Lu**, Aakanksha Chowdhery, and Srikanth Kandula, VisFlow: A Declarative Platform for Parallelizing Large-Scale Vision Programs. The 4th International Workshop on Large Scale Visual Recognition and Retrieval (CVPR Workshop), Las Vegas, USA, 2016.

Manuscripts and Pre-prints

- Junyi Shen, Noppanat Wadlom, **Yao Lu**. Batch Query Processing and Optimization for Agentic Workflows. arXiv preprint 2025. arxiv:2509.02121.
- Junyi Shen, Noppanat Wadlom, Lingfeng Zhou, Dequan Wang, Xu Miao, Lei Fang, **Yao Lu**. FlowMesh: A Service Fabric for Composable LLM Workflows. arXiv preprint 2025. arxiv:2510.26913.
- Pramod Chunduri, **Yao Lu**, Joy Arulraj. Tracer: Efficient object re-identification in networked cameras through adaptive query processing. arXiv preprint 2025. arxiv:2507.09448.
- Li Wang, Weiyuan Shao, **Yao Lu**, Hao Ye, Jian Pu, Yingbin Zheng. Crowd Counting with Density Adaption Networks. arXiv preprint 2018. arXiv:1806:10040.

Selected Awards

2025 Google Research Award
2023 VLDB Distinguished Reviewer
2018 ACM SIGMOD Best Demonstration Award
2017 IEEE ICME Platinum Best Paper Award
2014 University of Washington Royalty Research Fund Scholarship
2012 Chinese National Graduate Scholarship
2012 Google Innovation Scholarship
2011 Tencent Scholarship

Invited Talks

2023 **Towards Intelligent Data Systems**
Colloquium talk at Princeton University. Host: Kai Li

University of Sydney. Host: Joachim Gudmundsson
National University of Singapore. Host: Xiaokui Xiao

2022 **Pre-trained Models in Databases**
Database seminar talk at UC Berkeley SkyLab. Hosts: Tiemo Bang and Joeseph Hellerstein
Systems seminar talk at Stanford University. Hosts: Johann Hauswald, Christos Kozyrakis
Systems & database seminar talk at Duke University. Hosts: Danyang Zhuo and Jun Yang
Database seminar talk at Georgia Tech. Hosts: Joy Arulraj and Sham Navathe

2019 **Cardinality Estimation: Is Machine Learning a Silver Bullet?**
AIDB workshop talk @ VLDB

2018 **Machine Learning on Big-Data Systems**
Alibaba Research. Hosts: Bolin Ding and Jingren Zhou
IBM Research Almaden. Hosts: Berthold Reinwald and Fatma Ozcan
Google Research. Host: Cong Yu
Salesforce Research. Hosts: Caiming Xiong
Microsoft Research. Hosts: Yinan Li and Christian Konig

Teaching Experiences

Teaching Instructor

2026 Sem2 CS4262/5462 **Machine Learning Systems**, National University of Singapore (Creator)
2025 Sem1 CS6216 **Advanced Topics in Machine Learning (Systems)**, National University Singapore
2024 Sem2 CS4221 **Database Applications Design and Tuning**, National University Singapore
2024 Sem2 CS4221 **Database Applications Design and Tuning**, National University Singapore
2024 Sem1 CS6216 **Advanced Topics in Machine Learning (Systems)**, National University Singapore

Teaching Assistant

2018 Sum **CSE344 Introduction to Data Management**, University of Washington
Undergraduate course. Instructor: Kevin Zatloukal

2018 Win **CSE515 Statistical Methods in Computer Science**, University of Washington
Graduate course. Instructor: Pedro Domingos

2018 Spr **CSE455 Computer Vision**, University of Washington
Undergraduate course. Instructor: Linda Shapiro

2017 Win **CSE455 Computer Vision**, University of Washington
Undergraduate course. Instructor: Linda Shapiro

2017 Aut **CSE546 Machine Learning**, University of Washington
Graduate course. Instructor: Kevin Jamieson

2017 Spr **CSE576 Computer Vision**, University of Washington
Undergraduate course. Instructor: Linda Shapiro

2016 Spr **UW CSE547 Machine Learning and Big Data**, University of Washington
Graduate course. Instructor: Sham Kakade

2015 Win **CSE455 Computer Vision**, University of Washington
Undergraduate course. Instructor: Linda Shapiro

2014 Spr **CSE415 Introduction to AI**, University of Washington
Graduate course. Instructor: Linda Shapiro

2011 Spr **COMP120004 Linear Algebra**, Fudan University
Undergraduate course. Instructor: Wei Zhang

Students

Current PhDs

Junyi Shen	on ML systems
Noppanat Wadlom	on ML systems
Channe Chwa	on LLM posttraining
Zhengyuan Su	on ML systems
Yuncong Liu	on AI for finance
Bowen Qin	on LLM posttrainin

Current Master by thesis

Jiang Zhou	on Finance AI
Ningxian Jin	on Finance AI
Yongzhi Wang	on Finance AI
Rui Zeng	on LLM posttraining
Lingxiang Zhou	on Finance AI
Mustafa Hussain	on LLM posttraining
Yuan Wan	on Finance AI
Sarthak Kumar	on Finance AI
Nan Xiao	on Finance AI
Jiajia Xiang	on Finance AI
Yifan Wen	on Finance AI
You Li	on Finance AI

Alumni

Prerana Chakraborty (MSc)	on ML systems
Yulin Huang (MSc)	on Finance AI

Undergrads (NUS Final Year Project & Undergrad Research Opportunities Program)

Finance AI: Muchen Liu, Taanish Bhardwaj, Tze Kin Tai, Yi Xian Tan, Yuyang Huang, Zhehao Xu, Ho Yin Kiat, Si Rui Tan, Jeffrey Jian Yu Jie, Sultania Nikhil, Chen Yihao, Niu Wanjia, Billy Ho Cheng En, Shen James, Vanya Priscillia Bendatu, Lian Zhi Xuan, Lae Zong Hon Justyn, Chen Yi Xun, Daron Oh, Casper Eg E En, Sean Wong Sheng Hui, Feilin Liangga Putri, Zhu Rong, Hao Yushun, Jeffrey Liu Weixuan, Daniel Aloysius Png Yong Hui

Interns Mentored

2022 **Weiyuan Wu** (PhD student at Simon Fraser University) MSR: ML for DB

2022 **Md Mahmudulla Hassan** (PhD student at UTexas at El Paso) MS: anomaly detection

2021-2023 **Yongji Wu** (PhD student at Duke University) Co-advised with Danyang Zhuo and Matthew Lentz: Systems for ML

2020 **Beibin Li** (PhD student at University of Washington) MSR: ML for DB

2019 **Kexin Rong** (PhD student at Stanford University) MSR: co-mentored with Srikanth Kandla: ML for AQP

2019 **Xiao Huang** (PhD student at Texas A&M University) MSR: ML for DB

2019-2022 **Zhihui Yang** (PhD student at Fudan and UC Irvine) Co-advised with Chen Li and X. Sean Wang: ML workload optimization

2019-2023 **Pramod Chunduri** (PhD student at Georgia Tech) Co-advised with Joy Arulraj: Video data management systems

Doctoral Thesis Committee Member

2022 **Beibin Li,** Computer Science and Engineering, University of Washington

Professional Services

Program Committee Member: IEEE MIPR 2018 – 2023, AAAI 2019 – 2024, IEEE/CVF CVPR 2019 – 2023, 2026, IEEE ICCV 2019, ACM Multimedia Asia 2019, 2021, IEEE WACV 2020 – 2024, ACCV 2020, 2022, IEEE ECCV 2020, 2022, SMDB Workshop 2020-2021, AIDB Workshop 2020–2023, 2025, VLDB 2023, 2024, 2026, ICDE 2025, MLSys 2025, APSys 2025

Area Chair: MLSys 2025

General Chair: SoCC 2026

Journal Reviewer: Neurocomputing 2017-now, The Visual Computer 2017-2021, Pattern Recognition 2018-2021, Computer Vision and Image Understanding 2023, The VLDB Journal 2022-now.