Yao LU

Co-founder & CTO @ A stealth-mode startup

i @ yao.lu

luyao@comp.nus.edu.sg

I am deeply passionate about contributing to research and production at the intersection of machine learning and systems. Throughout my academic and industry journey, I have engaged with various tiers of the technology landscape and acquired hands-on experience in the data layer, core systems, advanced cloud infrastructures, and a diverse range of AI applications.

Working History

2023.9 – **Co-founder & CTO @ a stealth-mode startup**

A US-based tech startup offering efficient large-language-model-as-a-service platforms and solutions.

2018.11 – 2023.8 **Researcher @ Microsoft Research**, Redmond, WA, USA

Data Systems Group, manager: Vivek Narasayya and Surajit Chaudhuri
Research at the intersection of data systems and machine learning. Topics include:

1. *Improving data systems for machine learning*

I worked on improving the system efficiency using (1) black-box optimizations, e.g., UDF re-ordering, auto-parallelism, auto-scheduling, multi-query optimization [T.1,W1], reinforcement learning [P.3], on data center and heterogeneous infrastructures (IoT) [P.1], and (2) gray-box optimization, e.g., plan rewrite using proxy models [P.5,P.8,P.9,A.2,T.1].

2. *Improving data systems using machine learning*

I worked on improving different components of existing data systems using ML models built on query history, structured tables, execution plans and telemetries. Relevant projects include pre-training summarization models of structured datasets for cardinality estimation [P.,W.12,W.23], partition selection in answering big-data queries [P.7], and efficiently adapting ML models to data and workload drifts [P.2,A.1].

Education

2013 – 2018 **PhD in Computer Science and Engineering**

University of Washington, Seattle, WA

Research area: Data systems for ML

Advisor: Linda Shapiro

Committee member: Magdalena Balazinska, Srikanth Kandula

2010 – 2013 **MSc in Computer Science**

Fudan University, Shanghai, China

Research area: Computer vision and ML

2006 – 2010 **BEng in Computer Science**

Tongji University, Shanghai, China

Publications

Peer-Reviewed Conference Publications

2023 P.1 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.

2022 P.2 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.

P.3 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.

P.4 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.

P.5 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.

P.6 **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.

2020 P.7 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.

2018 P.8 **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.**

P.9 **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**.

P.10 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.

2017 P.11 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.

P.12 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**.

P.13 **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.

2016 P.14 **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.**

P.15 **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.

2012 P.16 **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.

P.17 **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.

2011 P.18 Wei Zhang, **Yao Lu**, Xiangyang Xue, Jianping Fan. Automatic Image Annotation with Weakly Labeled Datasets. ACM Multimedia. Scottsdale, AZ, USA. 2011.

P.19 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.

P.20 **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.

Patents

2021 A.1 **Yao Lu**, Srikanth Kandula. Adapting Learned Cardinality Estimators to Data and Workload Drifts. US Patent App. #17/566,996.

2019 A.2 Surajit Chaudhuri, Srikanth Kandula, **Yao Lu**. Accelerating Machine Learning Inference with Probabilistic Predicates. US Patent App. #16/003,495.

2017 A.3 Xue Bai, Jue Wang, **Yao Lu**. Flexible Video Object Boundary Tracking. US Patent #9,569,866.

Doctoral Thesis

2018 T.1 **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

2023 W.1 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.

2021 W.2 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.

2017 W.3 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.

2016 W.4 **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

2018 M.1 Li Wang, Weiyuan Shao, **Yao Lu**, Hao Ye, Jian Pu, Yingbin Zheng. Crowd Counting with Density Adaption Networks. arXiv preprint 2018. arXiv:1806:10040.

Prior Working Experiences

2017 **Research Intern @ Microsoft Research**, Redmond, WA, USA
DMX Group, mentored by Srikanth Kandula and Christian Konig
Worked on ML-based cardinality estimation.

2016 **Research Intern @ Microsoft Research Asia**, Beijing, China
Systems and Networking Group, worked on ML workload optimization. Project led to

a best SIGMOD demo award and production impact in Azure Cosmos DB.

2016 **Research Intern @ Microsoft Research**, Redmond, WA, USA
Mobility and Networking Group, worked on object tracking algorithms in videos.

2015 **Research Intern @ Microsoft Research**, Redmond, WA, USA
Mobility and Networking Group, worked on systems for ML. Project led to publications and production impact in Azure Cosmos DB.

2014 **Research Intern @ Adobe Research**, Seattle WA, USA
Creative Technology Lab, worked on video object segmentation. Project shipped to Adobe After Effects as the rigid mask tracker and face tracker.

2010-2015 **Research Assistant @ Fudan University** Media Lab, Shanghai, China *w/ Xiangyang Xue*

Research Assistant @ University of Washington, Seattle WA, USA *w/ Linda Shapiro*
Worked on ML algorithms and applications in computer vision. Topics include image and video segmentation, object detection, image labeling, and action detection in videos.

2009 **Software Development Engineer Intern @ Microsoft MSN China**, Shanghai, China

Selected Awards

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 and 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 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

Mentoring Experiences

Intern Mentoring

2022 **Weiyuan Wu** (PhD student at Simon Fraser University)

Microsoft Research Intern: ML for query optimization

2022 **Md Mahmudulla Hassan** (PhD student at UTexas at El Paso)
Microsoft Bing Intern: ML for anomaly detection

2021-now **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)
Microsoft Research Intern: ML for CE and workload modeling

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

2019 **Xiao Huang** (PhD student at Texas A&M University)
Microsoft Research Intern: ML for cardinality estimation

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

2019-now **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

Artificial Intelligence / Computer Vision:

Program Committee Member: IEEE MIPR 2018 – 2023, AAAI 2019 – 2024, IEEE/CVF CVPR 2019 – 2023, IEEE ICCV 2019, ACM Multimedia Asia 2019, 2021, IEEE WACV 2020 – 2024, ACCV 2020, 2022, IEEE ECCV 2020, 2022

Journal Reviewing: Neurocomputing 2017-now, The Visual Computer 2017-2021, Pattern Recognition 2018-2021, Computer Vision and Image Understanding 2023

Databases / Systems:

Program Committee Member: SMDB Workshop 2020-2021, AIDB Workshop 2020–2023, VLDB 2023, 2024

Journal Reviewing: The VLDB Journal 2022-now.