Wei Dai (David)

CONTACT INFORMATION

Email: david@davidwd.org
Phone: (203) 300-908
Homepage: `http://www.davidwd.org`

RESEARCH INTERESTS

Large-scale machine learning system and algorithm; End-to-end ML platform; Automatic feature engineering; Deep learning and medical imaging

EDUCATION

Carnegie Mellon University, Pittsburgh, PA

Ph.D. in Machine Learning **2012 – 2018**

- Research Advisor: Eric P. Xing

M.Sc. in Machine Learning **2012 – 2016**

California Institute of Technology, Pasadena, CA

B.Sc. with Honor in Computer Science **2010 – 2012**

- Research Advisor: Andreas Krause

Wesleyan University, Middletown, CT

B.A. with High Honor in Physics and Mathematics **2007 – 2010**

- Research Advisor: Francis W. Starr

EMPLOYEMENT

Apple Inc, Seattle, WA

Machine Learning Engineer **Feb 2019 – Present**

- Support the deployment of machine learning models at Apple.

Petuum Inc, Pittsburgh, PA

Senior Director of Engineering **Sept 2017 – Feb 2019**
Director of Product Development **May 2017 – Sept 2017**
Senior Manager **Jul 2016 – May 2017**

- As part of the founding team, I help grow the engineering team from <10 in 2016 to over 60 full-time engineers in 2Q18. Both a trench worker and a manager, I lead the product and engineering teams to deliver quarterly platform milestones for 3Q17 – 1Q18. Currently I lead engineering at Petuum, marshaling the work of 8 teams and 4 front-line managers.
- In my tech lead role, I lead the backend system design of Petuum's AI platform that combines all stages of machine learning—data ingestion and cleaning, feature engineering, embedding, training, data and model visualization, pipeline deployment and serving, monitoring—into one unified framework. Comparable systems include AzureML, TFX, FBLearner Flow, and Michelangelo.
- I led the former medical imaging team to develop best in class deep learning methods for chest x-ray to assist radiologists in diagnosis, resulting in Petuum's first contract.
- Proposed and implemented org adjustment for platform teams with broad feedback and support, resulting in a simpler org structure that empowers engineering managers and tech leads.
- Fostering engineering culture by hosting bi-weekly tech talks, performing 1-on-1, planning and executing performance review and promotions for all engineering. Support long-term efforts in formalizing engineering and management ladders.

Bosch Research, Pittsburgh, PA

Research Intern **May 2016 – Aug 2016**

- Developed state-of-the-art convolutional neural networks for environmental sound analysis using TensorFlow. Published two papers in ICASSP 2017.

Facebook, Menlo Park, CA

Software Engineering Intern **May 2015 – Aug 2015**

- Part of the FBLearn team (now Applied ML), I developed a distributed large-scale logistic regression using LBFGS and Petuum parameter server.
- Benchmarked the implementation against Facebook's internal system and Vowpal Wabbit; showed that the implementation achieves high system throughput and produces comparable to better models.

Google, Pittsburgh, PA

Software Engineering Intern **May 2013 – Aug 2013**

- Contributed to the Ad Quality backend; developed a hyperparameter tuning framework to optimize SmartAds training system with convex and non-convex optimization methods.
- The system was later published as Google Vizier.

LinkedIn, Mountain View, CA

Software Developer Intern **Jun 2012 – Aug 2012**

- Implemented a number of background tasks in the payment backend using Java, Oracle SQL, Python, and Spring Framework.

OpenX, Pasadena, CA

Software Developer Intern **Apr 2012 – Jun 2012**

- Simulated a large number of users to load-test several internal servers using Erlang and Tsung; developed Tsung modules to enable Thrift protocols.

Caltech Computer Science Department, Pasadena, CA

Research Assistant **Jun 2011 – Sept 2011**

- Contributed to the Community Seismic Network project which applies machine learning to detect earthquakes using smartphones.
- Applied *coreset* to training Gaussian mixture model using smartphone acceleration sensor data.

Selected Publications

W. Dai, Y. Zhou, N. Dong, H. Zhang, E. P. Xing
"Toward Understanding the Impact of Staleness in Distributed Machine Learning"
o appear in International Conference on Learning Representations (ICLR), 2019.

Y. Zhou, Y. Yu, **W. Dai**, Y. Liang, E. P. Xing
"Distributed Proximal Gradient Algorithm for Partially Asynchronous Computer Clusters"
Journal of Machine Learning Research (JMLR), 2018.

N. Dong, M. Kampffmeyer, X. Liang, Z. Wang, **W. Dai**, E. P. Xing
"Unsupervised Domain Adaptation for Automatic Estimation of Cardiothoracic Ratio"
Medical Image Computing and Computer Assisted Intervention (MICCAI), 2018.

Z. Wang, N. Dong, **W. Dai**, S. D. Rosario, E. P. Xing
"Classification of Breast Cancer Histopathological Images using Convolutional Neural Networks with Hierarchical Loss and Global Pooling"
International Conference on Image Analysis and Recognition (ICIAR), 2018. [Oral Presentation]

H. Zhang, S. Xu, G. Neubig, **W. Dai**, Q. Ho, G. Yang, E. P. Xing
"Cavs: A Vertex-centric Programming Interface for Dynamic Neural Networks"
Annual Technical Conference (ATC), 2018. [Oral Presentation]

H. Zhang, Z. Zheng, S. Xu, **W. Dai**, Q. Ho, X. Liang, Z. Hu, J. Wei, P. Xie, E. P. Xing
"Poseidon: An Efficient Communication Architecture for Distributed Deep Learning on GPU Clus-

ters"
Annual Technical Conference (ATC), 2017. [Oral Presentation]

X. Liang, L. Lee, **W. Dai**, E. P. Xing
"Dual Motion GAN for Future-Flow Embedded Video Prediction"
International Conference on Computer Vision (ICCV), 2017.

I. E.H. Yen, X. Huang, **W. Dai**, P. Ravikumar, I. Dhillon, E. P. Xing
"PPDSparse: A Parallel Primal-Dual Sparse Method for Extreme Classification"
Knowledge Discovery and Data Mining (KDD), 2017.

Y. Zhou, Y. Yu, **W. Dai**, Y. Liang, E. P. Xing
"Distributed Proximal Gradient Algorithm for Partially Asynchronous Computer Clusters"
arXiv:1704.03540, 2017.

W. Dai, J. Doyle, X. Liang, H. Zhang, N. Dong, Y. Li, E. P. Xing
"SCAN: Structure Correcting Adversarial Network for Organ Segmentation in Chest X-rays"
arXiv:1703.08770, 2017.

W. Dai*, C. Dai*, S. Qu, J. Li, S. Das
"Very Deep Convolutional Neural Networks for Raw Waveforms"
International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017.

J. Li, **W. Dai**, F. Metze, S. Qu, S. Das
"A Comparison of Deep Learning Methods for Environmental Sound Detection"
International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017.

A. Harlap, H. Cui, **W. Dai**, J. Wei, G. R. Ganger, P. B. Gibbons, G. A. Gibson, E. P. Xing
"Addressing the Straggler Problem for Iterative Convergent Parallel ML"
ACM Symposium on Cloud Computing (SoCC), 2016.

Y. Wang, V. Sadhanala, **W. Dai**, W. Neiswanger, S. Sra, E. P. Xing
"Parallel and Distributed Block-Coordinate Frank-Wolfe Algorithms"
International Conference of Machine Learning (ICML), 2016.

J. K. Kim, Q. Ho, S. Lee, X. Zheng, **W. Dai**, G. Gibson, E. P. Xing.
"STRADS: A Distributed Framework for Scheduled Model Parallel Machine Learning"
European Conference on Computer Systems (EuroSys), 2016.

Y. Zhou, Y. Yu, **W. Dai**, Y. Liang, E. P. Xing
"On Convergence of Model Parallel Proximal Gradient Algorithm for Stale Synchronous Parallel System"
Artificial Intelligence and Statistics (AISTATS), 2016.

E P. Xing, Q. Ho, P. Xie, **W. Dai**
"Strategies and Principles of Distributed Machine Learning on Big Data"
Engineering, Volume:2, pp. 179 - 95, 2016.

J. Wei, **W. Dai**, A. Qiao, H. Cui, Q. Ho, G. R. Ganger, P. B. Gibbons, G. A. Gibson, E. P. Xing
"Managed Communication and Consistency for Fast Data-Parallel Iterative Analytics"
ACM Symposium on Cloud Computing (SoCC), 2015. **[Best Paper]**

E. P. Xing, Q. Ho, **W. Dai**, J. K. Kim, J. Wei, S. Lee, X. Zheng, P. Xie, A. Kumar, Y. Yu
"Petuum: A New Platform for Distributed Machine Learning on Big Data"
Knowledge Discovery and Data Mining (KDD), 2015. [Oral Presentation]

J. Yuan, F. Gao, Q. Ho, **W. Dai**, J. Wei, X. Zheng, E. P. Xing, T. Liu, and W. Ma
"LightLDA: Big Topic Models on Modest Compute Cluster"
International World Wide Web Conference (WWW), 2015. [Oral Presentation]

W. Dai, A. Kumar, J. Wei, Q. Ho, G. Gibson, E. P. Xing
"Analysis of High-Performance Distributed ML at Scale through Parameter Server Consistency Models"
AAAI Conference on Artificial Intelligence (AAAI), 2015. [Oral Presentation]

H. Cui, A. Tumanov, J. Wei, L. Xu , **W. Dai**, J. Haber-Kucharsky, Q. Ho, G. R Ganger, P. B. Gibbons, G. A. Gibson, E. P. Xing
"Exploiting Iterative-ness for Parallel ML Computations"
Symposium on Cloud Computing (SoCC), 2014.

H. Cui, J. Cipar, Q. Ho, J. K. Kim, S. Lee, A. Kumar, J. Wei, **W. Dai**, G. R. Ganger, P. B. Gibbons, G. A. Gibson, E. P. Xing
"Exploiting Bounded Staleness to Speed Up Big Data Analytics"
Annual Technical Conference (ATC), 2014.

W. Dai, J. Wei, X. Zheng, J. K. Kim, S. Lee, J. Yin, Q. Ho, E. P. Xing
"Petuum: A Framework for Iterative-Convergent Distributed ML"
NIPS, Big Learning Workshop, 2013.

W. Dai, S. K. Kumar, F. W. Starr
"Universal two-step crystallization of DNA-Functionalized Nanoparticles"
Soft Matter, Vol. 6, pp. 6130-6135, 2010.

W. Dai, C. W. Hsu, F. Sciortino, F. W. Starr
"Valency Dependence of Polymorphism and Polyamorphism in DNA-Functionalized Nanoparticles"
Langmuir, Vol. 26, pp. 3601-3608, 2010.

W. Dai
"Effect of Valency on the Dynamics and Thermodynamics of DNA-linked Nanoparticles Materials"
Bachelor of Arts Honor Thesis: Wesleyan University, 2010.

Invited and Contributed Talks

AAAI Conference on Artificial Intelligence, January 2015. Title: Analysis of High-Performance Distributed ML at Scale through Parameter Server Consistency Models

Carnegie Mellon Univ., 2015 Spring: Invited guest lecturer for 10-605 Machine Learning with Large Datasets on Parameter Server. Title: Parameter Server and Stuff that Makes Large-scale Machine Learning Work.

California Institute of Technology: Summer Undergraduate Research Fellowship Seminar Day, October 2011. Title: A Smartphone that Learns: Toward Adaptive Earthquake Detection on Smartphones. (Advanced to final round in Perpall Speaking Competition.)

American Physical Society, March 2010 in Seattle, USA. Title: Phase Behavior of DNA-Functionalized Nanoparticles: Dependence on Number and Orientation of Attached DNA strands.

Awards and Honors

30 under 30 by the Pittsburgh Business Times, 2018.

Best Paper Award, ACM Symposium on Cloud Computing (SoCC), 2015.

High Honors from Wesleyan University Physics Department: Awarded for my undergraduate honor thesis work, 2010.

Freeman Asian Scholarship: A four-year full scholarship awarded to two students per country from eleven Asian countries for outstanding scholastic and leadership achievements, 2007.

Professional Service

Reviewer for IEEE Transaction on Medical Imaging, 2017, 2018.

Reviewer for Science Advance, 2017

Reviewer for PLOS ONE, 2017, 2018.

Reviewer for IEEE Transactions on Big Data, 2016.

Reviewer for AAAI Conference on Artificial Intelligence, 2016.

Technical Skills

Machine Learning: Parameter Server, Distributed Optimization, Deep Learning, Computer Vision, Medical Imaging

Technologies: TensorFlow, Spark

Languages

Python, C/C++