

Biographic Summary

Joseph Gonzalez is an Associate Professor at UC Berkeley and co-director of the Berkeley Sky Computing Lab and the LMSYS Org, where he studies the design of next-generation cloud and AI systems for high-performance machine learning. Gonzalez's research spans AI systems, large language model (LLM) inference and serving, agentic AI, neural network design, and distributed systems for large-scale machine learning, with applications in natural language processing, computer vision, robotics, and autonomous driving. Gonzalez has published over 100 papers in top-tier venues spanning machine learning (NeurIPS, ICML, ICLR, AISTATS, UAI, CVPR, ECCV, ICCV) and computer systems (VLDB, OSDI, NSDI, EuroSys, MLSys, SoCC), which have been cited more than 80,000 times with an h-index of 102 (Google Scholar, 2026). He has made significant contributions to LLM serving systems (vLLM, SGLang), open LLM evaluation (LMSYS Chatbot Arena), the design of efficient neural networks for meta-learning, and prediction serving systems. Gonzalez co-developed and teaches the large upper-division data science class (Data100) and co-teaches Introduction to Machine Learning (CS189/289A) at UC Berkeley. Prior to joining Berkeley, Gonzalez co-founded Turi Inc. (formerly GraphLab) based on his thesis work and created the GraphX project (now part of Apache Spark). Gonzalez's innovative work has earned him significant recognition, including the Sigmod Systems Award, the NSF Early CAREER Award, and the Okawa Research Grant.

For more information visit <http://eecs.berkeley.edu/~jegonzal>.

Employment and Professional Experience

Nov 2020 — Present	Co-founder, RunLLM (formerly Aqueduct)
July 2021 — Present	Associate Professor in the EECS Department
Jan 2016 — July 2021	Assistant Professor in the EECS Department
May 2013 — July 2016	Co-founder at Turi Inc., Acquired by Apple Inc.
Jan 2013 — July 2015	Post-doc in the UC Berkeley AMPLab
June 2011 — Aug 2011	Yahoo! Research Intern
June 2007 — Aug 2007	AT&T Labs Intern

Education

Ph.D., Machine Learning: December 2012. Machine Learning Department, School of Computer Science at Carnegie Mellon University.

Title: *“Parallel and Distributed Algorithms and Systems for Probabilistic Reasoning.”*

Thesis Advisor: Carlos Guestrin

M.S., Machine Learning: December 2009. Machine Learning Department, School of Computer Science at Carnegie Mellon University.

B.S. with Honors, Computer Science: June 2006. California Institute of Technology.

Awards

- **Diane S. McEntyre Award for Excellence in Teaching [2023]:** Award for excellence in teaching in the Department of Electrical Engineering and Computer Sciences at UC Berkeley.
- **Sigmod Systems Award [2022]:** Award for work on Apache Spark: an innovative, widely-used, open-source, unified data processing system encompassing relational, streaming, and machine-learning workloads.
- **NSF Early CAREER Award [2019]:** Funding awarded for research into the design of models and system for efficient prediction. Details: "*CAREER: Towards the Design of Models and Systems for Efficient Prediction*" (Award number 1846431).
- **Okawa Foundation Research Grant Award [2018]:** Funding award for research into the design of systems for machine learning life-cycle management.
- **Nominated for ACM Dissertation Award [2013]:** My thesis was nominated by CMU for the ACM Dissertation Award.
- **AT&T Labs Fellowship (2007):** Graduate research stipend for academic achievement as an under-represented minority.
- **NSF Graduate Research Fellowship (2007):** Graduate research stipend for 3 years.
- **NASA Space Act Award (2005):** Awarded for a sizeable contribution to space exploration.
- **NASA Inventions and Contributions Board Award (2005):** Awarded for the development of an innovative new technology that has made a contribution to space exploration.
- **Caltech Presidential Award (2002-2006):** I was awarded tuition for research and academic achievements.

Peer Reviewed Publications

- [1] Parth Asawa, Alan Zhu, Abigail O'Neill, Matei Zaharia, Alexandros G. Dimakis, and Joseph E. Gonzalez. "How to Train Your Advisor: Steering Black-Box LLMs with Advisor Models". In: *International Conference on Machine Learning (ICML 2026)*. 2026. URL: <https://arxiv.org/abs/2510.02453>.
- [2] Aditya Desai, Kunal Kumar Agrawal, Sehoon Yang, Alejandro Cuadron, Luis Gaspar Schroeder, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. "vAttention: Verified Sparse Attention via Sampling". In: *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR 2026)*. 2026. URL: <https://arxiv.org/abs/2510.05688>.
- [3] Shu Liu, Soujanya Ponnappalli, Shreya Shankar, Sepanta Zeighami, Alan Zhu, Shubham Agarwal, Ruiqi Chen, Samion Suwito, Shuo Yuan, Ion Stoica, Matei Zaharia, Alvin Cheung, Natacha Crooks, Joseph E. Gonzalez, and Aditya G. Parameswaran. "Supporting Our AI Overlords: Redesigning Data Systems to Be Agent-First". In: *Conference on Innovative Data Systems Research (CIDR)*. 2026. eprint: 2509.00997. URL: <https://arxiv.org/abs/2509.00997>.
- [4] Michael Luo, Xiaoxiang Shi, Colin Cai, Tianjun Zhang, Justin Wong, Yichuan Wang, Chi Wang, Yanping Huang, Zhifeng Chen, Joseph E. Gonzalez, and Ion Stoica. "Autellix: An Efficient Serving Engine for LLM Agents as General Programs". In: *USENIX Symposium on Networked Systems Design and Implementation (NSDI)*. 2026. eprint: 2502.13965. URL: <https://arxiv.org/abs/2502.13965>.

- [5] Qiuyang Mang, Wenhao Chai, Zhifei Li, Huanzhi Mao, Shang Zhou, Alexander Du, Hanchen Li, Shu Liu, Edwin Chen, Yichuan Wang, Xieting Chu, Zerui Cheng, Yuan Xu, Tian Xia, Zirui Wang, Tianneng Shi, Jianzhu Yao, Yilong Zhao, Qizheng Zhang, Charlie Ruan, Zeyu Shen, Kaiyuan Liu, Zhaoyang Hong, Alex Gu, Ziyi Zhang, Runyuan He, Dong Xing, Zerui Li, Zirong Zeng, Yige Jiang, Lufeng Cheng, Ziyi Zhao, Youran Sun, Suyang Zhong, Junpeng Wang, Donglin Li, Wenyan Huang, Jialiang Gu, Wesley Zheng, Wangmeiyu Zhang, Ruyi Ji, Xuechang Tu, Zihan Zheng, Zhaozi Wang, Zexing Chen, Jingbang Chen, Jialu Zhang, Aleksandra Korolova, Peter Henderson, Pramod Viswanath, Vijay Ganesh, Saining Xie, Zhuang Liu, Dawn Song, Sewon Min, Ion Stoica, Joseph E. Gonzalez, Jingbo Shang, and Alvin Cheung. “FrontierCS: Evolving Challenges for Evolving Intelligence”. In: *International Conference on Machine Learning (ICML 2026)*. 2026. URL: <https://arxiv.org/abs/2512.15699>.
- [6] Huanzhi Mao, Aditya Ghai, Imra Dawoodani, Tony Ginart, Shishir G. Patil, John Emmons, and Joseph E. Gonzalez. “MFCL Audio: An Audio Function Calling Evaluation for Large Language Models”. In: *International Conference on Machine Learning (ICML 2026)*. 2026. URL: <https://icml.cc/virtual/2026/poster/61489>.
- [7] Mihran Miroyan, Tsung-Han Wu, Logan King, Tianle Li, Jiayi Pan, Xinyan Hu, Wei-Lin Chiang, Anastasios N. Angelopoulos, Trevor Darrell, Narges Norouzi, and Joseph E. Gonzalez. “Search Arena: Analyzing Search-Augmented LLMs”. In: *International Conference on Learning Representations (ICLR 2026)*. 2026. URL: <https://arxiv.org/abs/2506.05334>.
- [8] Melissa Z. Pan, Negar Arabzadeh, Riccardo Cogo, Yuxuan Zhu, Alexander Xiong, Lakshya A. Agrawal, Huanzhi Mao, Emma Shen, Sid Pallerla, Liana Patel, Shu Liu, Tianneng Shi, Xiaoyuan Liu, Jared Quincy Davis, Emmanuele Lacavalla, Alessandro Basile, Shuyi Yang, Paul Castro, Daniel Kang, Koushik Sen, Dawn Song, Joseph E. Gonzalez, Ion Stoica, Matei Zaharia, and Marquita Ellis. “Characterizing Agents in Production”. In: *International Conference on Machine Learning (ICML 2026)*. 2026. URL: <https://arxiv.org/abs/2512.04123>.
- [9] Shishir G. Patil, Saurav Kumar, Prabal Dutta, and Joseph E. Gonzalez. “Efficient ML Model Updates for Deeply Embedded Microcontrollers”. In: *Proceedings of the 21st European Conference on Computer Systems (EuroSys 2026)*. 2026.
- [10] Yifan Qiao, Shan Yu, Shu Anzai, Haoran Ma, Shuo Yang, Yang Wang, Miryung Kim, Yongji Wu, Yang Zhou, Jiarong Xing, Joseph E. Gonzalez, Ion Stoica, and Harry Xu. “ConServe: Fine-Grained GPU Harvesting for LLM Online and Offline Co-Serving”. In: *International Conference on Machine Learning (ICML 2026)*. 2026. URL: <https://arxiv.org/abs/2410.01228>.
- [11] Luis Gaspar Schroeder, Aditya Desai, Alejandro Cuadron, Kyle Chu, Shu Liu, Mark Zhao, Stephan Krusche, Alfons Kemper, Matei Zaharia, and Joseph E. Gonzalez. “vCache: Verified Semantic Prompt Caching”. In: *International Conference on Learning Representations (ICLR 2026)*. 2026. URL: <https://arxiv.org/abs/2502.03771>.
- [12] Tsung-Han Wu, Mihran Miroyan, David M. Chan, Trevor Darrell, Narges Norouzi, and Joseph E. Gonzalez. “Are Large Reasoning Models Interruptible?”. In: *International Conference on Machine Learning (ICML 2026)*. 2026. URL: <https://arxiv.org/abs/2510.11713>.
- [13] Jintao Zhang, Haoxu Wang, Kai Jiang, Shuo Yang, Kaiwen Zheng, Haocheng Xi, Ziteng Wang, Hongzhou Zhu, Min Zhao, Ion Stoica, Joseph E. Gonzalez, Jianfei Chen, and Jun Zhu. “SLA: Beyond Sparsity in Diffusion Transformers via Fine-Tunable Sparse-Linear Attention”. In: *International Conference on Learning Representations (ICLR 2026)*. 2026. URL: <https://arxiv.org/abs/2509.24006>.
- [14] Asim Biswal, Liana Patel, Siddharth Jha, Amog Kamsetty, Shu Liu, Joseph E. Gonzalez, Carlos Guestrin, and Matei Zaharia. “Text2SQL Is Not Enough: Unifying AI and Databases with TAG”. In: *Conference on Innovative Data Systems Research (CIDR)*. 2025. eprint: 2408.14717. URL: <https://arxiv.org/abs/2408.14717>.

- [15] Shiyi Cao, Shu Liu, Tyler Griggs, Peter Schafhalter, Xiaoxuan Liu, Ying Sheng, Joseph E. Gonzalez, Matei Zaharia, and Ion Stoica. “MoE-Lightning: High-Throughput MoE Inference on Memory-Constrained GPUs”. In: *Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2025)*. 2025. URL: <https://arxiv.org/abs/2411.11217>.
- [16] Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. “Why Do Multi-Agent LLM Systems Fail?” In: *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*. 2025. URL: <https://arxiv.org/abs/2503.13657>.
- [17] Christopher Chou, Lisa Dunlap, Koki Mashita, Krishna Mandal, Trevor Darrell, Ion Stoica, Joseph E. Gonzalez, and Wei-Lin Chiang. “VisionArena: 230K Real-World User-VLM Conversations with Preference Labels”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2025)*. 2025. URL: <https://arxiv.org/abs/2412.08687>.
- [18] Aditya Desai, Shuo Yang, Alejandro Cuadron, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. “HashAttention: Semantic Sparsity for Faster Inference”. In: *International Conference on Machine Learning (ICML)*. 2025. eprint: 2412.14468. URL: <https://arxiv.org/abs/2412.14468>.
- [19] Lisa Dunlap, Joseph E. Gonzalez, Trevor Darrell, Fabian Caba Heilbron, Josef Sivic, and Bryan Russell. “Discovering Divergent Representations between Text-to-Image Models”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2025)*. 2025. URL: <https://arxiv.org/abs/2509.08940>.
- [20] Lisa Dunlap, Krishna Mandal, Trevor Darrell, Jacob Steinhardt, and Joseph E. Gonzalez. “VibeCheck: Discover and Quantify Qualitative Differences in Large Language Models”. In: *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR 2025)*. 2025. URL: <https://arxiv.org/abs/2410.12851>.
- [21] Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. “How to Evaluate Reward Models for RLHF”. In: *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR 2025)*. 2025. URL: <https://arxiv.org/abs/2410.14872>.
- [22] Dacheng Li, Shiyi Cao, Chengkun Cao, Xiuyu Li, Shangyin Tan, Kurt Keutzer, Jiarong Xing, Joseph E. Gonzalez, and Ion Stoica. “S*: Test Time Scaling for Code Generation”. In: *Findings of the Association for Computational Linguistics: EMNLP 2025*. 2025. eprint: 2502.14382. URL: <https://arxiv.org/abs/2502.14382>.
- [23] Dacheng Li, Shiyi Cao, Tyler Griggs, Shu Liu, Xiangxi Mo, Eric Tang, Sumanth Hegde, Kourosh Hakhmaneshi, Shishir G. Patil, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. “Language Models Can Easily Learn to Reason from Demonstrations: Structure, Not Content, Is What Matters!” In: *Findings of the Association for Computational Linguistics: EMNLP 2025*. 2025. eprint: 2502.07374. URL: <https://arxiv.org/abs/2502.07374>.
- [24] Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. “WorldModelBench: Judging Video Generation Models as World Models”. In: *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*. 2025. URL: <https://arxiv.org/abs/2502.20694>.
- [25] Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. “From Crowdsourced Data to High-Quality Benchmarks: Arena-Hard and BenchBuilder Pipeline”. In: *International Conference on Machine Learning (ICML)*. 2025. eprint: 2406.11939. URL: <https://arxiv.org/abs/2406.11939>.
- [26] Shu Liu, Asim Biswal, Amog Kamsetty, Audrey Cheng, Luis Gaspar Schroeder, Liana Patel, Shiyi Cao, Xiangxi Mo, Ion Stoica, Joseph E. Gonzalez, and Matei Zaharia. “Optimizing LLM Queries in

- Relational Data Analytics Workloads”. In: *Proceedings of Machine Learning and Systems 7 (MLSys 2025)*. 2025. URL: <https://arxiv.org/abs/2403.05821>.
- [27] Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. “RouteLLM: Learning to Route LLMs with Preference Data”. In: *International Conference on Learning Representations (ICLR)*. 2025. eprint: 2406.18665. URL: <https://arxiv.org/abs/2406.18665>.
 - [28] Shishir G. Patil, Hanchen Mao, Fanyue Yan, Chung-Cheng Jack Ji, Vivek Suresh, Ion Stoica, and Joseph E. Gonzalez. “The Berkeley Function Calling Leaderboard (BFCL): From Tool Use to Agentic Evaluation of Large Language Models”. In: *Proceedings of the 42nd International Conference on Machine Learning (ICML 2025)*. 2025. URL: https://gorilla.cs.berkeley.edu/blogs/8_berkeley_function_calling_leaderboard.html.
 - [29] Justin Wong, Yury Orlovskiy, Michael Luo, Sanjit A. Seshia, and Joseph E. Gonzalez. “SimpleStrat: Diversifying Language Model Generation with Stratification”. In: *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*. 2025. URL: <https://arxiv.org/abs/2410.09038>.
 - [30] Tsung-Han Wu, Giscard Biamby, Jerome Quenum, Ritwik Gupta, Joseph E. Gonzalez, Trevor Darrell, and David M. Chan. “Visual Haystacks: A Vision-Centric Needle-in-a-Haystack Benchmark”. In: *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR 2025)*. 2025. URL: <https://arxiv.org/abs/2407.13766>.
 - [31] Tsung-Han Wu, Joseph E. Gonzalez, Trevor Darrell, and David M. Chan. “CLAIRA: Leveraging Large Language Models to Judge Audio Captions”. In: *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2025)*. 2025. URL: <https://arxiv.org/abs/2409.12962>.
 - [32] Tsung-Han Wu, Heekyung Lee, Jiaxin Ge, Joseph E. Gonzalez, Trevor Darrell, and David M. Chan. “Generate, but Verify: Reducing Hallucination in Vision-Language Models with Retrospective Resampling”. In: *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*. 2025. URL: <https://arxiv.org/abs/2504.13169>.
 - [33] Ling Yang, Zhaochen Yu, Tianjun Zhang, Minkai Xu, Joseph E. Gonzalez, Bin Cui, and Shuicheng Yan. “SuperCorrect: Advancing Small LLM Reasoning with Thought Template Distillation and Self-Correction”. In: *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR 2025)*. 2025. URL: <https://arxiv.org/abs/2410.09008>.
 - [34] Chen Zhang, Kuntai Du, Shu Liu, Woosuk Kwon, Xiangxi Mo, Yufeng Wang, Xiaoxuan Liu, Kaichao You, Zhuohan Li, Mingsheng Long, Jidong Zhai, Joseph E. Gonzalez, and Ion Stoica. “JENGA: Effective Memory Management for Serving LLM with Heterogeneity”. In: *Proceedings of the ACM SIGOPS 31st Symposium on Operating Systems Principles (SOSP 2025)*. 2025. URL: <https://arxiv.org/abs/2503.18292>.
 - [35] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. “Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference”. In: *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*. 2024. URL: <https://arxiv.org/abs/2403.04132>.
 - [36] Lisa Dunlap, Yuhui Zhang, Xiaohan Wang, Ruiqi Zhong, Trevor Darrell, Jacob Steinhardt, Joseph E. Gonzalez, and Serena Yeung-Levy. “Describing Differences in Image Sets with Natural Language”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024)*. 2024. URL: <https://arxiv.org/abs/2312.02974>.
 - [37] Naman Jain, Tianjun Zhang, Wei-Lin Chiang, Joseph E. Gonzalez, Koushik Sen, and Ion Stoica. “LLM-Assisted Code Cleaning for Training Accurate Code Generators”. In: *Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024)*. 2024. URL: <https://arxiv.org/abs/2311.14904>.

- [38] Dacheng Li, Rulin Shao, Anze Xie, Eric P. Xing, Xuezhe Ma, Ion Stoica, Joseph E. Gonzalez, and Hao Zhang. “DistFlashAttn: Distributed Memory-Efficient Attention for Long-Context LLMs Training”. In: *Conference on Language Modeling (COLM)*. 2024. eprint: 2310.03294. URL: <https://arxiv.org/abs/2310.03294>.
- [39] Michael Luo, Justin Wong, Brandon Trabucco, Yanping Huang, Joseph E. Gonzalez, Zhifeng Chen, Ruslan Salakhutdinov, and Ion Stoica. “Stylus: Automatic Adapter Selection for Diffusion Models”. In: *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*. 2024. URL: <https://arxiv.org/abs/2404.18928>.
- [40] Federico Mora, Justin Wong, Haley Lepe, Sahil Bhatia, Karim Elmaaroufi, George Varghese, Joseph E. Gonzalez, Elizabeth Polgreen, and Sanjit A. Seshia. “Synthetic Programming Elicitation for Text-to-Code in Very Low-Resource Programming and Formal Languages”. In: *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*. 2024. URL: <https://arxiv.org/abs/2406.03636>.
- [41] Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. “Gorilla: Large Language Model Connected with Massive APIs”. In: *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*. 2024. eprint: 2305.15334. URL: <https://arxiv.org/abs/2305.15334>.
- [42] Suzanne Petryk, David M. Chan, Anish Kachinthaya, Haodi Zou, John Canny, Joseph E. Gonzalez, and Trevor Darrell. “ALOHA: A New Measure for Hallucination in Captioning Models”. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2024)*. 2024. URL: <https://arxiv.org/abs/2404.02904>.
- [43] Suzanne Petryk, Spencer Whitehead, Joseph E. Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. “Simple Token-Level Confidence Improves Caption Correctness”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2024. eprint: 2305.07021. URL: <https://arxiv.org/abs/2305.07021>.
- [44] Daniel Rothchild, Andrew S. Rosen, Eric Taw, Connie Robinson, Joseph E. Gonzalez, and Aditi S. Krishnapriyan. “Investigating the Behavior of Diffusion Models for Accelerating Electronic Structure Calculations”. In: *Chemical Science* (2024). URL: <https://arxiv.org/abs/2311.01491>.
- [45] Sheng Shen, Shijia Yang, Tianjun Zhang, Bohan Zhai, Joseph E. Gonzalez, Kurt Keutzer, and Trevor Darrell. “Multitask Vision-Language Prompt Tuning”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2024)*. 2024. URL: <https://arxiv.org/abs/2211.11720>.
- [46] Ying Sheng, Shiyi Cao, Dacheng Li, Coleman Hooper, Nicholas Lee, Shuo Yang, Christopher Chou, Banghua Zhu, Lianmin Zheng, Kurt Keutzer, Joseph E. Gonzalez, and Ion Stoica. “SLoRA: Scalable Serving of Thousands of LoRA Adapters”. In: *Proceedings of Machine Learning and Systems 6 (MLSys 2024)*. 2024. URL: <https://arxiv.org/abs/2311.03285>.
- [47] Ying Sheng, Shiyi Cao, Dacheng Li, Banghua Zhu, Zhuohan Li, Danyang Zhuo, Joseph E. Gonzalez, and Ion Stoica. “Fairness in Serving Large Language Models”. In: *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 2024)*. 2024. URL: <https://arxiv.org/abs/2401.00588>.
- [48] Sijun Tan, Xiuyu Li, Shishir G. Patil, Ziyang Wu, Tianjun Zhang, Kurt Keutzer, Joseph E. Gonzalez, and Raluca Ada Popa. “LLoCO: Learning Long Contexts Offline”. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*. 2024. URL: <https://arxiv.org/abs/2404.07979>.
- [49] Sarah Wooders, Shishir Liu, Paras Jain, Xiangxi Mo, Joseph E. Gonzalez, Vincent Liu, and Ion Stoica. “CloudCast: High-Throughput, Cost-Aware Overlay Multicast in the Cloud”. In: *21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 2024)*. 2024. URL: <https://www.usenix.org/conference/nsdi24/presentation/wooders>.

- [50] Tsung-Han Wu, Giscard Biamby, David Chan, Lisa Dunlap, Ritwik Gupta, Xudong Wang, Joseph E. Gonzalez, and Trevor Darrell. “See, Say, and Segment: Teaching LMMs to Overcome False Premises”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024)*. 2024. URL: <https://arxiv.org/abs/2312.08366>.
- [51] Tsung-Han Wu, Long Lian, Joseph E. Gonzalez, Boyi Li, and Trevor Darrell. “Self-Correcting LLM-Controlled Diffusion Models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024)*. 2024. URL: <https://arxiv.org/abs/2311.16090>.
- [52] Jiezhi Yang, Khushi Desai, Charles Packer, Harshil Bhatia, Nicholas Rhinehart, Rowan McAllister, and Joseph E. Gonzalez. “CARFF: Conditional Auto-Encoded Radiance Field for 3D Scene Forecasting”. In: *Proceedings of the European Conference on Computer Vision (ECCV 2024)*. 2024. URL: <https://arxiv.org/abs/2401.18075>.
- [53] Ling Yang, Zhaochen Yu, Tianjun Zhang, Shiyi Cao, Minkai Xu, Wentao Zhang, Joseph E. Gonzalez, and Bin Cui. “Buffer of Thoughts: Thought-Augmented Reasoning with Large Language Models”. In: *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*. 2024. URL: <https://arxiv.org/abs/2406.04271>.
- [54] Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. “RAFT: Adapting Language Model to Domain Specific RAG”. In: *Conference on Language Modeling (COLM)*. 2024. eprint: 2403.10131. URL: <https://arxiv.org/abs/2403.10131>.
- [55] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. “LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset”. In: *Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024)*. 2024. URL: <https://arxiv.org/abs/2309.11998>.
- [56] Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. “SGLang: Efficient Execution of Structured Language Model Programs”. In: *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*. 2024. URL: <https://arxiv.org/abs/2312.07104>.
- [57] Robert Avram, Jeffrey E. Olgin, Zeeshan Ahmed, Louis Verreault-Julien, Alvin Wan, Julia Barrios, Justin Wan, Joseph E. Gonzalez, Derek So, Krishan Soni, and Geoffrey H. Tison. “CathAI: Fully Automated Coronary Angiography Interpretation and Stenosis Estimation”. In: *npj Digital Medicine*. 2023. URL: <https://www.nature.com/articles/s41746-023-00880-1>.
- [58] David M. Chan, Suzanne Petryk, Joseph E. Gonzalez, Trevor Darrell, and John Canny. “CLAIR: Evaluating Image Captions with Large Language Models”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*. 2023. URL: <https://arxiv.org/abs/2310.12971>.
- [59] Lisa Dunlap, Alyssa Umno, Han Zhang, Jiezhi Yang, Joseph E. Gonzalez, and Trevor Darrell. “Diversify Your Vision Datasets with Automatic Diffusion-Based Augmentation”. In: *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*. 2023. eprint: 2305.16289. URL: <https://arxiv.org/abs/2305.16289>.
- [60] Joseph E. Gonzalez and Yucheng Low. “The Story of GraphLab—From Scaling Machine Learning to Shaping Graph Systems Research (VLDB 2023 Test-of-Time Award Talk)”. In: *Proceedings of the VLDB Endowment*. 2023. URL: <https://dl.acm.org/doi/10.14778/3611479.3611538>.
- [61] Kirthivasan Kandasamy, Joseph E. Gonzalez, Michael I. Jordan, and Ion Stoica. “VCG Mechanism Design with Unknown Agent Values under Stochastic Bandit Feedback”. In: *Journal of Machine Learning Research* 24 (2023). URL: <https://arxiv.org/abs/2004.08924>.
- [62] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. “Efficient Memory Management for Large Language Model

- Serving with PagedAttention”. In: *Proceedings of the 29th ACM Symposium on Operating Systems Principles (SOSP 2023)*. 2023. URL: <https://arxiv.org/abs/2309.06180>.
- [63] Zhuohan Li, Lianmin Zheng, Yinmin Zhong, Vincent Liu, Ying Sheng, Xin Jin, Yanping Huang, Zhifeng Chen, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. “AlpaServe: Statistical Multiplexing with Model Parallelism for Deep Learning Serving”. In: *USENIX Symposium on Operating Systems Design and Implementation (OSDI)*. 2023. URL: <https://www.usenix.org/conference/osdi23/presentation/li-zhouhan>.
 - [64] Kevin Lin, Kyle Lo, Joseph E. Gonzalez, and Dan Klein. “Decomposing Complex Queries for Tip-of-the-Tongue Retrieval”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. 2023. eprint: 2305.15053. URL: <https://arxiv.org/abs/2305.15053>.
 - [65] Peter Schafhalter, Sukrit Kalra, Le Xu, Joseph E. Gonzalez, and Ion Stoica. “Leveraging Cloud Computing to Make Autonomous Vehicles Safer”. In: *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2023)*. 2023. URL: <https://arxiv.org/abs/2308.03204>.
 - [66] Ying Sheng, Lianmin Zheng, Binhang Yuan, Zhuohan Li, Max Ryabinin, Daniel Y. Fu, Zhiqiang Xie, Beidi Chen, Clark Barrett, Joseph E. Gonzalez, Percy Liang, Christopher Ré, Ion Stoica, and Ce Zhang. “FlexGen: High-Throughput Generative Inference of Large Language Models with a Single GPU”. In: *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*. 2023. URL: <https://arxiv.org/abs/2303.06865>.
 - [67] Sarah Wooders, Xiangxi Mo, Amanda Narang, Kevin Lin, Ion Stoica, Joseph M. Hellerstein, Natacha Crooks, and Joseph E. Gonzalez. “RALF: Accuracy-Aware Scheduling for Feature Store Maintenance”. In: *Proceedings of the VLDB Endowment* 17.3 (2023), pp. 563–576. DOI: 10.14778/3632093.3632116. URL: <https://doi.org/10.14778/3632093.3632116>.
 - [68] Yaoqing Yang, Ryan Theisen, Liam Hodgkinson, Joseph E. Gonzalez, Kannan Ramchandran, Charles H. Martin, and Michael W. Mahoney. “Test Accuracy vs. Generalization Gap: Model Selection in NLP Without Accessing Training or Testing Data”. In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2023)*. 2023. URL: <https://arxiv.org/abs/2202.02842>.
 - [69] Tianjun Zhang, Fangchen Liu, Justin Wong, Pieter Abbeel, and Joseph E. Gonzalez. “The Wisdom of Hindsight Makes Language Models Better Instruction Followers”. In: *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*. 2023. URL: <https://arxiv.org/abs/2302.05206>.
 - [70] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. “Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena”. In: *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*. 2023. eprint: 2306.05685. URL: <https://arxiv.org/abs/2306.05685>.
 - [71] Yonghao Zhuang, Lianmin Zheng, Zhuohan Li, Eric Xing, Qirong Ho, Joseph E. Gonzalez, Ion Stoica, Hao Zhang, and Hexu Zhao. “On Optimizing the Communication of Model Parallelism”. In: *Proceedings of Machine Learning and Systems*. 2023. URL: https://proceedings.mlsys.org/paper_files/paper/2023/hash/d0b9a3081f811b2a307c38ad457a487c-Abstract-mlsys2023.html.
 - [72] Ionel Gog, Sukrit Kalra, Peter Schafhalter, Joseph E. Gonzalez, and Ion Stoica. “D3: A Dynamic Deadline-Driven Approach for Building Autonomous Vehicles”. In: *Proceedings of the Seventeenth European Conference on Computer Systems. EuroSys ’22*. Rennes, France: Association for Computing Machinery, 2022, pp. 453–471. ISBN: 9781450391627. DOI: 10.1145/3492321.3519576. URL: <https://doi.org/10.1145/3492321.3519576>.
 - [73] Wenshuo Guo, Kirthivasan Kandasamy, Joseph Gonzalez, Michael Jordan, and Ion Stoica. “Learning Competitive Equilibria in Exchange Economies with Bandit Feedback”. In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. Ed. by Gustau Camps-Valls,

- Francisco J. R. Ruiz, and Isabel Valera. Vol. 151. Proceedings of Machine Learning Research. PMLR, Mar. 2022, pp. 6200–6224. URL: <https://proceedings.mlr.press/v151/guo22a.html>.
- [74] Paras Jain, Safeen Huda, Martin Maas, Joseph E. Gonzalez, Ion Stoical, and Azalia Mirhoseini. “Learning to Design Accurate Deep Learning Accelerators with Inaccurate Multipliers”. In: *2022 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. 2022, pp. 184–189. DOI: 10.23919/DATE54114.2022.9774607. URL: https://www.parasjain.com/projects/21autoapprox_date/paper.pdf.
- [75] Xiaoxuan Liu, Lianmin Zheng, Dequan Wang, Yukuo Cen, Weize Chen, Xu Han, Jianfei Chen, Zhiyuan Liu, Jie Tang, Joseph E. Gonzalez, Michael Mahoney, and Alvin Cheung. “GACT: Activation Compressed Training for Generic Network Architectures”. In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato. Vol. 162. Proceedings of Machine Learning Research. PMLR, July 2022, pp. 14139–14152. URL: <https://proceedings.mlr.press/v162/liu22v.html>.
- [76] Shishir G. Patil, Paras Jain, Prabal Dutta, Ion Stoica, and Joseph Gonzalez. “POET: Training Neural Networks on Tiny Devices with Integrated Rematerialization and Paging”. In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato. Vol. 162. Proceedings of Machine Learning Research. PMLR, July 2022, pp. 17573–17583. URL: <https://proceedings.mlr.press/v162/patil22b.html>.
- [77] Suzanne Petryk, Lisa Dunlap, Keyan Nasseri, Joseph Gonzalez, Trevor Darrell, and Anna Rohrbach. “On Guiding Visual Attention With Language Specification”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 18092–18102.
- [78] Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. “Reliable Visual Question Answering: Abstain Rather Than Answer Incorrectly”. In: *European Conference on Computer Vision (ECCV)*. arXiv, 2022. DOI: 10.48550/ARXIV.2204.13631. URL: <https://arxiv.org/abs/2204.13631>.
- [79] Albert Wilcox, Ashwin Balakrishna, Brijen Thananjeyan, Joseph E. Gonzalez, and Ken Goldberg. “LS3: Latent Space Safe Sets for Long-Horizon Visuomotor Control of Sparse Reward Iterative Tasks”. In: *Proceedings of the 5th Conference on Robot Learning*. Ed. by Aleksandra Faust, David Hsu, and Gerhard Neumann. Vol. 164. Proceedings of Machine Learning Research. PMLR, Nov. 2022, pp. 959–969. URL: <https://proceedings.mlr.press/v164/wilcox22a.html>.
- [80] Bichen Wu, Ruizhe Cheng, Peizhao Zhang, Tianren Gao, Joseph E. Gonzalez, and Peter Vajda. “Data Efficient Language-Supervised Zero-Shot Recognition with Optimal Transport Distillation”. In: *International Conference on Learning Representations*. 2022. URL: <https://openreview.net/forum?id=G89-1yZLFHk>.
- [81] Tianjun Zhang, Benjamin Eysenbach, Ruslan Salakhutdinov, Sergey Levine, and Joseph E. Gonzalez. “C-Planning: An Automatic Curriculum for Learning Goal-Reaching Tasks”. In: *International Conference on Learning Representations (ICLR)*. arXiv, 2022. DOI: 10.48550/ARXIV.2110.12080. URL: <https://arxiv.org/abs/2110.12080>.
- [82] Tianjun Zhang, Tongzheng Ren, Mengjiao Yang, Joseph Gonzalez, Dale Schuurmans, and Bo Dai. “Making Linear MDPs Practical via Contrastive Representation Learning”. In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato. Vol. 162. Proceedings of Machine Learning Research. PMLR, July 2022, pp. 26447–26466. URL: <https://proceedings.mlr.press/v162/zhang22x.html>.
- [83] Zhengming Zhang, Ashwinee Panda, Linyue Song, Yaoqing Yang, Michael Mahoney, Prateek Mittal, Ramchandran Kannan, and Joseph Gonzalez. “Neurotoxin: Durable Backdoors in Federated Learning”. In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato.

- Vol. 162. Proceedings of Machine Learning Research. PMLR, July 2022, pp. 26429–26446. URL: <https://proceedings.mlr.press/v162/zhang22w.html>.
- [84] Lianmin Zheng, Zhuohan Li, Hao Zhang, Yonghao Zhuang, Zhifeng Chen, Yanping Huang, Yida Wang, Yuanzhong Xu, Danyang Zhuo, Eric P. Xing, Joseph E. Gonzalez, and Ion Stoica. “Alpa: Automating Inter- and Intra-Operator Parallelism for Distributed Deep Learning”. In: *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22)*. Carlsbad, CA: USENIX Association, July 2022, pp. 559–578. ISBN: 978-1-939133-28-1. URL: <https://www.usenix.org/conference/osdi22/presentation/zheng-lianmin>.
 - [85] Raghav Anand, Jeffrey Ichnowski, Chenggang Wu, Joseph M. Hellerstein, Joseph E. Gonzalez, and Ken Goldberg. “Serverless Multi-Query Motion Planning for Fog Robotics”. In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. May 2021, pp. 7457–7463. DOI: 10.1109/ICRA48506.2021.9561571. URL: <https://goldberg.berkeley.edu/pubs/ICRA21-ichnowski-serverless-motion-planning-submitted.pdf>.
 - [86] Robert Avram, Jeffrey Olgin, Alvin Wan, Zeeshan Ahmed, Louis Verreault-Julien, Sean Abreau, Derek Wan, Joseph E. Gonzalez, Derek So, Krishan Soni, and Geoffrey Tison. “CATHAI: FULLY AUTOMATED CORONARY ANGIOGRAPHY INTERPRETATION AND STENOSIS DETECTION USING A DEEP LEARNING-BASED ALGORITHMIC PIPELINE”. In: *Journal of the American College of Cardiology* 77.18_Supplement_1 (2021), pp. 3244–3244. DOI: 10.1016/S0735-1097(21)04598-8. eprint: <https://www.jacc.org/doi/pdf/10.1016/S0735-1097%2821%2904598-8>. URL: <https://www.jacc.org/doi/abs/10.1016/S0735-1097%2821%2904598-8>.
 - [87] Jianfei Chen, Lianmin Zheng, Zhewei Yao, Dequan Wang, Ion Stoica, Michael Mahoney, and Joseph Gonzalez. “ActNN: Reducing Training Memory Footprint via 2-Bit Activation Compressed Training”. In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, July 2021, pp. 1803–1813. URL: <https://proceedings.mlr.press/v139/chen21z.html>.
 - [88] Kaiyuan Eric Chen, Yafei Liang, Nikhil Jha, Jeffrey Ichnowski, Michael Danielczuk, Joseph Gonzalez, John Kubiawicz, and Ken Goldberg. “FogROS: An Adaptive Framework for Automating Fog Robotics Deployment”. In: *2021 IEEE 17th International Conference on Automation Science and Engineering (CASE)*. Aug. 2021, pp. 2035–2042. DOI: 10.1109/CASE49439.2021.9551628. URL: <https://arxiv.org/abs/2108.11355>.
 - [89] Ruizhe Cheng, Bichen Wu, Peizhao Zhang, Peter Vajda, and Joseph E. Gonzalez. “Data-Efficient Language-Supervised Zero-Shot Learning With Self-Distillation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2021, pp. 3119–3124.
 - [90] Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Bichen Wu, Zijian He, Zhen Wei, Kan Chen, Yuandong Tian, Matthew Yu, Peter Vajda, and Joseph E. Gonzalez. “FBNetV3: Joint Architecture-Recipe Search Using Predictor Pretraining”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2021, pp. 16276–16285.
 - [91] Lisa Dunlap, Kirthivasan Kandasamy, Ujval Misra, Richard Liaw, Michael Jordan, Ion Stoica, and Joseph E. Gonzalez. “Elastic Hyperparameter Tuning on the Cloud”. In: *Proceedings of the ACM Symposium on Cloud Computing*. SoCC ’21. Seattle, WA, USA: Association for Computing Machinery, 2021, pp. 33–46. ISBN: 9781450386388. DOI: 10.1145/3472883.3486989. URL: <https://doi.org/10.1145/3472883.3486989>.
 - [92] Sayna Ebrahimi, Suzanne Petryk, Akash Gokul, William Gan, Joseph E. Gonzalez, Marcus Rohrbach, and Trevor Darrell. “Remembering for the right reasons: Explanations reduce catastrophic forgetting”. In: *Applied AI Letters* 2.4 (2021), e44. DOI: <https://doi.org/10.1002/ail2.44>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/ail2.44>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/ail2.44>.

- [93] Yu Gai, Paras Jain, Wendi Zhang, Joseph E. Gonzalez, Dawn Song, and Ion Stoica. “Grounded Graph Decoding Improves Compositional Generalization in Question Answering”. In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. arXiv, 2021. DOI: 10.48550/ARXIV.2111.03642. URL: <https://arxiv.org/abs/2111.03642>.
- [94] Aditya Ganapathi, Priya Sundareshan, Brijen Thananjeyan, Ashwin Balakrishna, Daniel Seita, Jennifer Grannen, Minh Hwang, Ryan Hoque, Joseph E. Gonzalez, Nawid Jamali, Katsu Yamane, Soshi Iba, and Ken Goldberg. “Learning Dense Visual Correspondences in Simulation to Smooth and Fold Real Fabrics”. In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. May 2021, pp. 11515–11522. DOI: 10.1109/ICRA48506.2021.9561980. URL: <https://arxiv.org/abs/2003.12698>.
- [95] Ionel Gog, Sukrit Kalra, Peter Schafhalter, Matthew A. Wright, Joseph E. Gonzalez, and Ion Stoica. “Pylot: A Modular Platform for Exploring Latency-Accuracy Tradeoffs in Autonomous Vehicles”. In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. May 2021, pp. 8806–8813. DOI: 10.1109/ICRA48506.2021.9561747. URL: <https://arxiv.org/abs/2104.07830>.
- [96] J. Weston Hughes, Jeffrey E. Olgin, Robert Avram, Sean A. Abreau, Taylor Sittler, Kaahan Radia, Henry Hsia, Tomos Walters, Byron Lee, Joseph E. Gonzalez, and Geoffrey H. Tison. “Performance of a Convolutional Neural Network and Explainability Technique for 12-Lead Electrocardiogram Interpretation”. In: *JAMA Cardiology* 6.11 (Nov. 2021), pp. 1285–1295. ISSN: 2380-6583. DOI: 10.1001/jamacardio.2021.2746. eprint: https://jamanetwork.com/journals/jamacardiology/articlepdf/2782549/jamacardiology_hughes_2021_oi_210051_1635348688.2775.pdf. URL: <https://doi.org/10.1001/jamacardio.2021.2746>.
- [97] Jeffrey Ichnowski, Paras Jain, Bartolomeo Stellato, Goran Banjac, Michael Luo, Francesco Borrelli, Joseph E Gonzalez, Ion Stoica, and Ken Goldberg. “Accelerating Quadratic Optimization with Reinforcement Learning”. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 21043–21055. URL: <https://proceedings.neurips.cc/paper/2021/file/afdec7005cc9f14302cd0474fd0f3c96-Paper.pdf>.
- [98] Anand Padmanabha Iyer, Qifan Pu, Kishan Patel, Joseph E. Gonzalez, and Ion Stoica. “TEGRA: Efficient Ad-Hoc Analytics on Evolving Graphs”. In: *18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21)*. USENIX Association, Apr. 2021, pp. 337–355. ISBN: 978-1-939133-21-2. URL: <https://www.usenix.org/conference/nsdi21/presentation/iyer>.
- [99] Paras Jain, Ajay Jain, Tianjun Zhang, Pieter Abbeel, Joseph E. Gonzalez, and Ion Stoica. “Contrastive Code Representation Learning”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2021, pp. 5954–5971. DOI: 10.18653/v1/2021.emnlp-main.482. URL: <https://arxiv.org/abs/2007.04973>.
- [100] Eric Liang, Zhanghao Wu, Michael Luo, Sven Mika, Joseph E Gonzalez, and Ion Stoica. “RLlib Flow: Distributed Reinforcement Learning is a Dataflow Problem”. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 5506–5517. URL: <https://proceedings.neurips.cc/paper/2021/file/2bce32ed409f5ebcee2a7b417ad9beed-Paper.pdf>.
- [101] Ujval Misra, Richard Liaw, Lisa Dunlap, Romil Bhardwaj, Kirthevasan Kandasamy, Joseph E. Gonzalez, Ion Stoica, and Alexey Tumanov. “RubberBand: Cloud-Based Hyperparameter Tuning”. In: *Proceedings of the Sixteenth European Conference on Computer Systems. EuroSys ’21*. Online Event, United Kingdom: Association for Computing Machinery, 2021, pp. 327–342. ISBN: 9781450383349. DOI: 10.1145/3447786.3456245. URL: <https://doi.org/10.1145/3447786.3456245>.
- [102] Charles Packer, Pieter Abbeel, and Joseph E Gonzalez. “Hindsight Task Relabelling: Experience Replay for Sparse Reward Meta-RL”. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran

Associates, Inc., 2021, pp. 2466–2477. URL: <https://proceedings.neurips.cc/paper/2021/file/1454ca2270599546dfcd2a3700e4d2f1-Paper.pdf>.

- [103] Samuel Paradis, Minho Hwang, Brijen Thananjeyan, Jeffrey Ichnowski, Daniel Seita, Danyal Fer, Thomas Low, Joseph E. Gonzalez, and Ken Goldberg. “Intermittent Visual Servoing: Efficiently Learning Policies Robust to Instrument Changes for High-precision Surgical Manipulation”. In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. May 2021, pp. 7166–7173. DOI: 10.1109/ICRA48506.2021.9561070. URL: <https://arxiv.org/abs/2011.06163>.
- [104] Nathan Pemberton, Johann Schleier-Smith, and Joseph E. Gonzalez. “The RESTless Cloud”. In: *Proceedings of the Workshop on Hot Topics in Operating Systems*. HotOS ’21. Ann Arbor, Michigan: Association for Computing Machinery, 2021, pp. 49–57. ISBN: 9781450384384. DOI: 10.1145/3458336.3465280. URL: <https://doi.org/10.1145/3458336.3465280>.
- [105] Devin Petersohn, Dixin Tang, Rehan Durrani, Areg Melik-Adamyian, Joseph E. Gonzalez, Anthony D. Joseph, and Aditya G. Parameswaran. “Flexible Rule-Based Decomposition and Metadata Independence in Modin: A Parallel Dataframe System”. In: *Proc. VLDB Endow.* 15.3 (Nov. 2021), pp. 739–751. ISSN: 2150-8097. DOI: 10.14778/3494124.3494152. URL: <https://doi.org/10.14778/3494124.3494152>.
- [106] Nicholas Rhinehart, Jeff He, Charles Packer, Matthew A. Wright, Rowan McAllister, Joseph E. Gonzalez, and Sergey Levine. “Contingencies from Observations: Tractable Contingency Planning with Learned Behavior Models”. In: *IEEE International Conference on Robotics and Automation (ICRA)*. arXiv, 2021. DOI: 10.48550/ARXIV.2104.10558. URL: <https://arxiv.org/abs/2104.10558>.
- [107] Johann Schleier-Smith, Vikram Sreekanti, Anurag Khandelwal, Joao Carreira, Neeraja J. Yadwadkar, Raluca Ada Popa, Joseph E. Gonzalez, Ion Stoica, and David A. Patterson. “What Serverless Computing is and Should Become: The next Phase of Cloud Computing”. In: *Commun. ACM* 64.5 (Apr. 2021), pp. 76–84. ISSN: 0001-0782. DOI: 10.1145/3406011. URL: <https://doi.org/10.1145/3406011>.
- [108] Priya Sundaresan, Jennifer Grannen, Brijen Thananjeyan, Ashwin Balakrishna, Jeffrey Ichnowski, Ellen Novoseller, Minho Hwang, Michael Laskey, Joseph E. Gonzalez, and Ken Goldberg. “Untangling Dense Non-Planar Knots by Learning Manipulation Features and Recovery Policies”. In: *Robotics: Science and Systems (RSS)*. arXiv, 2021. DOI: 10.48550/ARXIV.2107.08942. URL: <https://arxiv.org/abs/2107.08942>.
- [109] Brijen Thananjeyan, Ashwin Balakrishna, Suraj Nair, Michael Luo, Krishnan Srinivasan, Minho Hwang, Joseph E. Gonzalez, Julian Ibarz, Chelsea Finn, and Ken Goldberg. “Recovery RL: Safe Reinforcement Learning With Learned Recovery Zones”. In: *IEEE Robotics and Automation Letters* 6.3 (July 2021), pp. 4915–4922. ISSN: 2377-3766. DOI: 10.1109/LRA.2021.3070252. URL: <https://arxiv.org/abs/2010.15920>.
- [110] Brijen Thananjeyan, Kirthevasan Kandasamy, Ion Stoica, Michael Jordan, Ken Goldberg, and Joseph Gonzalez. “Resource Allocation in Multi-armed Bandit Exploration: Overcoming Sublinear Scaling with Adaptive Parallelism”. In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, July 2021, pp. 10236–10246. URL: <https://proceedings.mlr.press/v139/thananjeyan21a.html>.
- [111] Vainavi Viswanath, Jennifer Grannen, Priya Sundaresan, Brijen Thananjeyan, Ashwin Balakrishna, Ellen Novoseller, Jeffrey Ichnowski, Michael Laskey, Joseph E. Gonzalez, and Ken Goldberg. “Disentangling Dense Multi-Cable Knots”. In: *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2021, pp. 3731–3738. DOI: 10.1109/IROS51168.2021.9636397. URL: <https://arxiv.org/abs/2106.02252>.

- [112] Alvin Wan, Lisa Dunlap, Daniel Ho, Jihan Yin, Scott Lee, Suzanne Petryk, Sarah Adel Bargal, and Joseph E. Gonzalez. “NBDT: Neural-Backed Decision Tree”. In: *International Conference on Learning Representations*. 2021. URL: <https://openreview.net/forum?id=mCLVeEpp1NE>.
- [113] Guanhua Wang, Zhuang Liu, Brandon Hsieh, Siyuan Zhuang, Joseph Gonzalez, Trevor Darrell, and Ion Stoica. “sensAI: ConvNets Decomposition via Class Parallelism for Fast Inference on Live Data”. In: *Proceedings of Machine Learning and Systems*. Ed. by A. Smola, A. Dimakis, and I. Stoica. Vol. 3. 2021, pp. 664–679. URL: <https://proceedings.mlsys.org/paper/2021/file/c4ca4238a0b923820dcc509a6f75849b-Paper.pdf>.
- [114] Xin Wang, Thomas E. Huang, Benlin Liu, Fisher Yu, Xiaolong Wang, Joseph E. Gonzalez, and Trevor Darrell. “Robust Object Detection via Instance-Level Temporal Cycle Confusion”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2021, pp. 9143–9152.
- [115] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph E. Gonzalez, Kurt Keutzer, and Peter Vajda. “Visual Transformers: Where Do Transformers Really Belong in Vision Models?”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2021, pp. 599–609.
- [116] Zhanghao Wu, Paras Jain, Matthew Wright, Azalia Mirhoseini, Joseph E Gonzalez, and Ion Stoica. “Representing Long-Range Context for Graph Neural Networks with Global Attention”. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 13266–13279. URL: <https://proceedings.neurips.cc/paper/2021/file/6e67691b60ed3e4a55935261314dd534-Paper.pdf>.
- [117] Kevin Yang, Tianjun Zhang, Chris Cummins, Brandon Cui, Benoit Steiner, Linnan Wang, Joseph E Gonzalez, Dan Klein, and Yuandong Tian. “Learning Space Partitions for Path Planning”. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 378–391. URL: <https://proceedings.neurips.cc/paper/2021/file/03a3655fff3e9bdea48de9f49e938e32-Paper.pdf>.
- [118] Yaoqing Yang, Liam Hodgkinson, Ryan Theisen, Joe Zou, Joseph E Gonzalez, Kannan Ramchandran, and Michael W Mahoney. “Taxonomizing local versus global structure in neural network loss landscapes”. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 18722–18733. URL: <https://proceedings.neurips.cc/paper/2021/file/9b72e31dac81715466cd580a448cf823-Paper.pdf>.
- [119] Tianjun Zhang, Paria Rashidinejad, Jiantao Jiao, Yuandong Tian, Joseph E Gonzalez, and Stuart Russell. “MADE: Exploration via Maximizing Deviation from Explored Regions”. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 9663–9680. URL: <https://proceedings.neurips.cc/paper/2021/file/5011bf6d8a37692913fce3a15a51f070-Paper.pdf>.
- [120] Tianjun Zhang, Huazhe Xu, Xiaolong Wang, Yi Wu, Kurt Keutzer, Joseph E Gonzalez, and Yuandong Tian. “NovelD: A Simple yet Effective Exploration Criterion”. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 25217–25230. URL: <https://proceedings.neurips.cc/paper/2021/file/d428d070622e0f4363fcea11f4a3576-Paper.pdf>.
- [121] Zhengming Zhang, Yaoqing Yang, Zhewei Yao, Yujun Yan, Joseph E. Gonzalez, Kannan Ramchandran, and Michael W. Mahoney. “Improving Semi-supervised Federated Learning by Reducing the Gradient Diversity of Models”. In: *2021 IEEE International Conference on Big Data (Big Data)*. Dec. 2021, pp. 1214–1225. DOI: 10.1109/BigData52589.2021.9671693.

- [122] Lianmin Zheng, Ruochen Liu, Junru Shao, Tianqi Chen, Joseph E. Gonzalez, Ion Stoica, and Ameer Haj Ali. “TenSet: A Large-scale Program Performance Dataset for Learned Tensor Compilers”. In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*. 2021. URL: <https://openreview.net/forum?id=aIfp8kLuv9>.
- [123] Ashwin Balakrishna, Brijen Thananjeyan, Jonathan Lee, Felix Li, Arsh Zahed, Joseph E. Gonzalez, and Ken Goldberg. “On-Policy Robot Imitation Learning from a Converging Supervisor”. In: *Proceedings of the Conference on Robot Learning*. Ed. by Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura. Vol. 100. Proceedings of Machine Learning Research. PMLR, Oct. 2020, pp. 24–41. URL: <http://proceedings.mlr.press/v100/balakrishna20a.html>.
- [124] Jianfei Chen, Yu Gai, Zhewei Yao, Michael W. Mahoney, and Joseph E. Gonzalez. “A Statistical Framework for Low-bitwidth Training of Deep Neural Networks”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 883–894. URL: <https://proceedings.neurips.cc/paper/2020/file/099fe6b0b444c23836c4a5d07346082b-Paper.pdf>.
- [125] Daniel Crankshaw, Gur-Eyal Sela, Corey Zumar, Xiangxi Mo, Joseph E. Gonzalez, Ion Stoica, and Alexey Tumanov. “InferLine: ML Inference Pipeline Composition Framework”. In: *Proceedings of the ACM Symposium on Cloud Computing*. SoCC ’20. Association for Computing Machinery, Nov. 2020. URL: <http://arxiv.org/abs/1812.01776>.
- [126] Ankur Dave, Chester Leung, Raluca Ada Popa, Joseph E. Gonzalez, and Ion Stoica. “Oblivious Cooperative Analytics Using Hardware Enclaves”. In: *Proceedings of the Fifteenth European Conference on Computer Systems (EuroSys)*. EuroSys ’20. Heraklion, Greece: Association for Computing Machinery, 2020. ISBN: 9781450368827. URL: <https://doi.org/10.1145/3342195.3387552>.
- [127] Rolando Garcia, Eric Liu, Vikram Sreekanti, Bobby Yan, Anusha Dandamudi, Joseph E. Gonzalez, Joseph M. Hellerstein, and Koushik Sen. “Hindsight Logging for Model Training”. In: *Proc. VLDB Endow.* 14.4 (Dec. 2020), pp. 682–693. ISSN: 2150-8097. DOI: 10.14778/3436905.3436925. URL: <https://doi.org/10.14778/3436905.3436925>.
- [128] Jeffrey Ichnowski, William Lee, Victor Murta, Samuel Paradis, Ron Alterovitz, Joseph E. Gonzalez, Ion Stoica, and Ken Goldberg. “Fog Robotics Algorithms for Distributed Motion Planning Using Lambda Serverless Computing”. In: *2020 IEEE International Conference on Robotics and Automation, ICRA 2020, Paris, France, May 31 - August 31, 2020*. IEEE, 2020, pp. 4232–4238. DOI: 10.1109/ICRA40945.2020.9196651. URL: <https://doi.org/10.1109/ICRA40945.2020.9196651>.
- [129] Paras Jain, Ajay Jain, Aniruddha Nrusimha, Amir Gholami, Pieter Abbeel, Kurt Keutzer, Ion Stoica, and Joseph E. Gonzalez. “Breaking the Memory Wall with Optimal Tensor Rematerialization”. In: *Proceedings of Machine Learning and Systems*. 2020, pp. 497–511. URL: <https://arxiv.org/abs/1910.02653>.
- [130] Samvit Jain, Xun Zhang, Yuhao Zhou, Ganesh Ananthanarayanan, Junchen Jiang, Yuanhao Shu, Paramvir Bahl, and Joseph Gonzalez. “Spatula: Efficient cross-camera video analytics on large camera networks”. In: *5th IEEE/ACM Symposium on Edge Computing, SEC 2020, San Jose, CA, USA, November 12-14, 2020*. IEEE, 2020, pp. 110–124. DOI: 10.1109/SEC50012.2020.00016. URL: <https://doi.org/10.1109/SEC50012.2020.00016>.
- [131] Zhuohan Li, Eric Wallace, Sheng Shen, Kevin Lin, Kurt Keutzer, Dan Klein, and Joseph E. Gonzalez. “Train Big, Then Compress: Rethinking Model Size for Efficient Training and Inference of Transformers”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. ICML’20. July 2020. URL: <https://proceedings.icml.cc/static/paper%5C%5Ffiles/icml/2020/6626-Paper.pdf>.
- [132] Xiayue Charles Lin, Joseph E. Gonzalez, and Joseph M. Hellerstein. “Serverless Boom or Bust? An Analysis of Economic Incentives”. In: *12th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud)*. USENIX Association, July 2020. URL: <https://www.usenix.org/conference/hotcloud20/presentation/lin>.

- [133] Mong H. Ng, Kaahan Radia, Jianfei Chen, Dequan Wang, Ionel Gog, and Joseph E. Gonzalez. “BEV-Seg: Bird’s Eye View Semantic Segmentation Using Geometry and Semantic Point Cloud”. In: *Proceedings of the Workshop in Scalability for Autonomous Driving at CVPR’20*. 2020. arXiv: 2006.11436. URL: <https://arxiv.org/abs/2006.11436>.
- [134] Devin Petersohn, William W. Ma, Doris Jung Lin Lee, Stephen Macke, Doris Xin, Xiangxi Mo, Joseph E. Gonzalez, Joseph M. Hellerstein, Anthony D. Joseph, and Aditya G. Parameswaran. “Towards Scalable Dataframe Systems”. In: *Proceedings of Very Large Data Bases (PVLDB)*. Vol. 13. 2020. URL: <http://www.vldb.org/pvldb/vol13/p2033-petersohn.pdf>.
- [135] Daniel Rothchild, Ashwinee Panda, Enayat Ullah, Nikita Ivkin, Ion Stoica, Vladimir Braverman, Joseph E. Gonzalez, and Raman Arora. “FetchSGD: Communication-Efficient Federated Learning with Sketching”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. ICML’20. July 2020. URL: <https://proceedings.icml.cc/static/paper%5C%5Ffiles/icml/2020/5927-Paper.pdf>.
- [136] Vidit Saxena, Joakim Jalden, and Joseph E. Gonzalez. “Thompson Sampling for Linearly Constrained Bandits”. In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (AISTats)*. Ed. by Silvia Chiappa and Roberto Calandra. Vol. 108. Proceedings of Machine Learning Research. Online: PMLR, Aug. 2020, pp. 1999–2009. URL: <http://proceedings.mlr.press/v108/saxena20a.html>.
- [137] Vikram Sreekanti, Chenggang Wu, Saurav Chhatrapati, Joseph E. Gonzalez, Joseph M. Hellerstein, and Jose M. Faleiro. “A Fault-Tolerance Shim for Serverless Computing”. In: *Proceedings of the Fifteenth European Conference on Computer Systems (EuroSys)*. EuroSys ’20. Heraklion, Greece: Association for Computing Machinery, 2020. ISBN: 9781450368827. URL: <https://doi.org/10.1145/3342195.3387535>.
- [138] Vikram Sreekanti, Chenggang Wu, Xiayue Charles Lin, Johann Schleier-Smith, Jose M. Faleiro, Joseph E. Gonzalez, Joseph M. Hellerstein, and Alexey Tumanov. “Cloudburst: Stateful Functions-as-a-Service”. In: *Proceedings of Very Large Data Bases (PVLDB)*. Vol. 13. 2020. URL: <http://www.vldb.org/pvldb/vol13/p2438-sreekanti.pdf>.
- [139] Priya Sundaresan, Jennifer Grannen, Brijen Thananjeyan, Ashwin Balakrishna, Michael Laskey, Kevin Stone, Joseph E. Gonzalez, and Ken Goldberg. “Learning Rope Manipulation Policies Using Dense Object Descriptors Trained on Synthetic Depth Data”. In: *2020 IEEE International Conference on Robotics and Automation, ICRA 2020, Paris, France, May 31 - August 31, 2020*. IEEE, 2020, pp. 9411–9418. DOI: 10.1109/ICRA40945.2020.9197121. URL: <https://doi.org/10.1109/ICRA40945.2020.9197121>.
- [140] Ajay Kumar Tanwani, Raghav Anand, Joseph E. Gonzalez, and Ken Goldberg. “RILaaS: Robot Inference and Learning as a Service”. In: *IEEE Robotics and Automation Letters* 5.3 (2020), pp. 4423–4430. URL: <https://ieeexplore.ieee.org/document/9103220>.
- [141] Brijen Thananjeyan, Ashwin Balakrishna, Ugo Rosolia, Joseph E. Gonzalez, Aaron D. Ames, and Ken Goldberg. “ABC-LMPC: Safe Sample-Based Learning MPC for Stochastic Nonlinear Dynamical Systems with Adjustable Boundary Conditions”. In: *Proceedings of the Int. Workshop on the Algorithmic Foundations of Robotics (WAFR)*. 2020. URL: <https://arxiv.org/abs/2003.01410>.
- [142] Brijen Thananjeyan, Ashwin Balakrishna, Ugo Rosolia, Felix Li, Rowan McAllister, Joseph E. Gonzalez, Sergey Levine, Francesco Borrelli, and Ken Goldberg. “Safety Augmented Value Estimation From Demonstrations (SAVED): Safe Deep Model-Based RL for Sparse Cost Robotic Tasks”. In: *IEEE Robotics Autom. Lett.* 5.2 (2020), pp. 3612–3619. URL: <https://arxiv.org/abs/1905.13402>.
- [143] Alvin Wan, Xiaoliang Dai, Peizhao Zhang, Zijian He, Yuandong Tian, Saining Xie, Bichen Wu, Matthew Yu, Tao Xu, Kan Chen, Peter Vajda, and Joseph E. Gonzalez. “FBNetV2: Differentiable Neural Architecture Search for Spatial and Channel Dimensions”. In: *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2020. URL: <https://arxiv.org/abs/2004.05565>.

- [144] Xin Wang, Thomas E. Huang, Trevor Darrell, Joseph E. Gonzalez, and Fisher Yu. “Frustratingly Simple Few-Shot Object Detection”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. ICML’20. July 2020. URL: <https://proceedings.icml.cc/static/paper%5C%5Ffiles/icml/2020/2957-Paper.pdf>.
- [145] Yaoqing Yang, Rajiv Khanna, Yaodong Yu, Amir Gholami, Kurt Keutzer, Joseph E. Gonzalez, Kannan Ramchandran, and Michael W. Mahoney. “Boundary thickness and robustness in learning models”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 6223–6234. URL: <https://proceedings.neurips.cc/paper/2020/file/44e76e99b5e194377e955b13fb12f630-Paper.pdf>.
- [146] Lianmin Zheng, Chengfan Jia, Minmin Sun, Zhao Wu, Cody Hao Yu, Ameer Haj-Ali, Yida Wang, Jun Yang, Danyang Zhuo, Koushik Sen, Joseph E. Gonzalez, and Ion Stoica. “Ansor: Generating High-Performance Tensor Programs for Deep Learning”. In: *14th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2020, Virtual Event, November 4-6, 2020*. USENIX Association, 2020, pp. 863–879. URL: <https://www.usenix.org/conference/osdi20/presentation/zheng>.
- [147] Joseph M. Hellerstein, Jose M. Faleiro, Joseph E. Gonzalez, Johann Schleier-Smith, Vikram Sreekanti, Alexey Tumanov, and Chenggang Wu. “Serverless Computing: One Step Forward, Two Steps Back”. In: *Conference on Innovative Data Systems Research (CIDR ’19)*. Jan. 2019. URL: <https://arxiv.org/abs/1812.03651>.
- [148] Samvit Jain, Ganesh Ananthanarayanan, Junchen Jiang, Yuanchao Shu, and Joseph E. Gonzalez. “Scaling Video Analytics Systems to Large Camera Deployments”. In: *HotMobile ’19, Proceedings of the 20th International Workshop on Mobile Computing Systems and Applications*. Feb. 2019. URL: <https://arxiv.org/abs/1809.02318>.
- [149] Samvit Jain, Xin Wang, and Joseph Gonzalez. “Accel: A Corrective Fusion Network for Efficient Semantic Segmentation on Video”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019. URL: <http://arxiv.org/abs/1807.06667>.
- [150] Richard Liaw, Romil Bhardwaj, Lisa Dunlap, Yitian Zou, Joseph E. Gonzalez, Ion Stoica, and Alexey Tumanov. “HyperSched: Dynamic Resource Reallocation for Model Development on a Deadline”. In: *Proceedings of the ACM Symposium on Cloud Computing*. SoCC ’19. Santa Cruz, CA, USA: Association for Computing Machinery, 2019, pp. 61–73. ISBN: 9781450369732. URL: <https://doi.org/10.1145/3357223.3362719>.
- [151] Vidit Saxena, Joakim Jaldén, Joseph E. Gonzalez, Mats Bengtsson, Hugo M. Tullberg, and Ion Stoica. “Contextual Multi-Armed Bandits for Link Adaptation in Cellular Networks”. In: *Proceedings of the Workshop on Network Meets AI (NetAI) at SIGCOMM*. 2019, pp. 44–49. URL: <https://doi.org/10.1145/3341216.3342212>.
- [152] Ajay Kumar Tanwani, Nitesh Mor, John Kubiawicz, Joseph E. Gonzalez, and Ken Goldberg. “A Fog Robotics Approach to Deep Robot Learning: Application to Object Recognition and Grasp Planning in Surface Decluttering”. In: *International Conference on Robotics and Automation, ICRA 2019, Montreal, QC, Canada, May 20-24, 2019*. 2019, pp. 4559–4566. URL: <https://doi.org/10.1109/ICRA.2019.8793690>.
- [153] Xin Wang, Fisher Yu, Lisa Dunlap, Yi-An Ma, Ruth Wang, Azalia Mirhoseini, Trevor Darrell, and Joseph E. Gonzalez. “Deep Mixture of Experts via Shallow Embedding”. In: *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019*. 2019, p. 192. URL: <http://auai.org/uai2019/proceedings/papers/192.pdf>.
- [154] Xin Wang, Fisher Yu, Ruth Wang, Trevor Darrell, and Joseph E. Gonzalez. “TAFE-Net: Task-Aware Feature Embeddings for Low Shot Learning”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019. URL: <https://arxiv.org/abs/1904.05967>.

- [155] Zuxuan Wu, Xin Wang, Joseph E. Gonzalez, Tom Goldstein, and Larry S. Davis. “ACE: Adapting to Changing Environments for Semantic Segmentation”. In: *International Conference in Computer Vision (ICCV)*. Oct. 2019. URL: <http://arxiv.org/abs/1904.06268>.
- [156] Tianjun Zhang, Zhewei Yao, Amir Gholami, Kurt Keutzer, Joseph E. Gonzalez, George Biros, and Michael W. Mahoney. “ANODEV2: A Coupled Neural ODE Evolution Framework”. In: *Neural Information Processing Systems (NeurIPS)*. 2019. URL: <https://arxiv.org/abs/1906.04596>.
- [157] Wenting Zheng, Raluca Ada Popa, Joseph E. Gonzalez, and Ion Stoica. “Helen: Maliciously Secure Coopetitive Learning for Linear Models”. In: *IEEE Symposium on Security and Privacy (Oakland)*. IEEE Computer Society, 2019. URL: <https://people.eecs.berkeley.edu/~wzheng/helen%5C%5FfIEEEsp.pdf>.
- [158] Rolando Garcia, Vikram Sreekanti, Neeraja Yadwadkar, Daniel Crankshaw, Joseph E. Gonzalez, and Joseph M. Hellerstein. “Context: The Missing Piece in the Machine Learning Lifecycle”. In: *Proceedings of the KDD Workshop on Common Model Infrastructure (CMI)*. Aug. 2018. URL: <http://www.vikrams.io/papers/flor-cmi18.pdf>.
- [159] Samvit Jain and Joseph E. Gonzalez. “Fast Semantic Segmentation on Video Using Block Motion-Based Feature Interpolation”. In: *The Third International Workshop on Video Segmentation (IWVS)*. Mar. 2018. URL: <https://arxiv.org/abs/1803.07742>.
- [160] Eric Liang, Richard Liaw, Robert Nishihara, Philipp Moritz, Roy Fox, Joseph Gonzalez, Ken Goldberg, and Ion Stoica. “Ray RLLib: A Composable and Scalable Reinforcement Learning Library”. In: *Proceedings of the 35th International Conference on Machine Learning. ICML ’18*. ACM, July 2018. URL: <https://arxiv.org/abs/1712.09381>.
- [161] Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E. Gonzalez, and Ion Stoica. “Tune: A Research Platform for Distributed Model Selection and Training”. In: *Proceedings of the ICML Workshop on AutoML*. 2018. URL: <https://arxiv.org/abs/1807.05118>.
- [162] Xiangxi Mo, Paras Jain, Ajay Jain, Alexey Tumanov, Joseph E. Gonzalez, and Ion Stoica. “A Case for Dynamic GPU Inference Multitenancy and Scheduling”. In: *Proceedings of the Learning Systems Workshop at NIPS 2018*. Dec. 2018. URL: [http://learningsys.org/nips18/assets/papers/102CameraReadySubmissionGPU%5C%5FVirtualization%5C%20\(8\).pdf](http://learningsys.org/nips18/assets/papers/102CameraReadySubmissionGPU%5C%5FVirtualization%5C%20(8).pdf).
- [163] Xin Wang, Yujia Luo, Dan Crankshaw, Alexey Tumanov, Fisher Yu, and Joseph E. Gonzalez. “IDK Cascades: Fast Deep Learning by Learning not to Overthink”. In: *Conference on Uncertainty in Artificial Intelligence (UAI)*. July 2018. URL: <https://arxiv.org/abs/1706.00885>.
- [164] Xin Wang, Fisher Yu, Zi-Yi Dou, and Joseph E. Gonzalez. “SkipNet: Learning Dynamic Routing in Convolutional Networks”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. July 2018. URL: <https://arxiv.org/abs/1711.09485>.
- [165] Bichen Wu, Alvin Wan, Xiangyu Yue, Peter Jin, Sicheng Zhao, Noah Golmant, Amir Gholaminejad, Joseph E. Gonzalez, and Kurt Keutzer. “Shift: A Zero FLOP, Zero Parameter Alternative to Spatial Convolutions”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018. URL: <https://arxiv.org/abs/1711.08141>.
- [166] Francois W. Belletti, Evan R. Sparks, Michael J. Franklin, Alexandre M. Bayen, and Joseph E. Gonzalez. “Random Projection Design for Scalable Implicit Smoothing of Randomly Observed Stochastic Processes”. In: *Artificial Intelligence and Statistics (AISTats ’17)*. July 2017. URL: <http://proceedings.mlr.press/v54/belletti17a/belletti17a.pdf>.
- [167] Daniel Crankshaw, Xin Wang, Guilio Zhou, Michael J. Franklin, Joseph E. Gonzalez, and Ion Stoica. “Clipper: A Low-Latency Online Prediction Serving System”. In: *14th USENIX Symposium on Networked Systems Design and Implementation (NSDI 17)*. Boston, MA: USENIX Association, 2017, pp. 613–627. ISBN: 978-1-931971-37-9. URL: <https://www.usenix.org/conference/nsdi17/technical-sessions/presentation/crankshaw>.

- [168] Joseph M. Hellerstein, Vikram Sreekanti, Joseph E. Gonzalez, Sudhanshu Arora, Arka Bhattacharyya, Shirshanka Das, Akon Dey, Mark Donsky, Gabriel Fierro, Sreyashi Nag, Krishna Ramachandran, Chang She, Eric Sun, Carl Steinbach, and Venkat Subramanian. “Establishing Common Ground with Data Context”. In: *Conference on Innovative Data Systems Research (CIDR '17)*. 2017.
- [169] Neeraja J. Yadwadkar, Bharath Hariharan, Joseph E. Gonzalez, Burton Smith, and Randy H. Katz. “Selecting the Best VM Across Multiple Public Clouds: A Data-driven Performance Modeling Approach”. In: *Proceedings of the 2017 Symposium on Cloud Computing*. SoCC '17. Santa Clara, California: ACM, Sept. 2017, pp. 452–465. ISBN: 978-1-4503-5028-0. URL: <https://doi.acm.org/10.1145/3127479.3131614>.
- [170] Wenting Zheng, Ankur Dave, Jethro G. Beekman, Raluca Ada Popa, Joseph E. Gonzalez, and Ion Stoica. “Opaque: An Oblivious and Encrypted Distributed Analytics Platform”. In: *14th USENIX Symposium on Networked Systems Design and Implementation (NSDI 17)*. Boston, MA: USENIX Association, 2017, pp. 283–298. ISBN: 978-1-931971-37-9. URL: <https://www.usenix.org/conference/nsdi17/technical-sessions/presentation/zheng>.
- [171] Ankur Dave, Alekh Jindal, Li Erran Li, Reynold Xin, Joseph E. Gonzalez, and Matei Zaharia. “GraphFrames: An Integrated API for Mixing Graph and Relational Queries.” In: *SIGMOD Grades Workshop*. 2016.
- [172] Neeraja J. Yadwadkar, Bharath Hariharan, Joseph E. Gonzalez, and Randy Katz. “Multi-Task Learning for Straggler Avoiding Predictive Job Scheduling”. In: *Journal of Machine Learning Research (JMLR '16)*. 2016.
- [173] Daniel Crankshaw, Peter Bailis, Joseph E. Gonzalez, Haoyuan Li, Zhao Zhang, Michael J. Franklin, Ali Ghodsi, and Michael I. Jordan. “The Missing Piece in Complex Analytics: Low Latency, Scalable Model Management and Serving with Velox”. In: *Conference on Innovative Data Systems Research (CIDR '15)*. 2015.
- [174] Daniel Crankshaw, Xin Wang, Joseph E. Gonzalez, and Michael J. Franklin. “Scalable Training and Serving of Personalized Models”. In: *Proceedings of the Learning Systems Workshop at NIPS 2015*. 2015.
- [175] Veronika Strnadova-Neeley, Aydin Buluc, Jarrod Chapman, John Gilbert, Joseph E. Gonzalez, and Leonid Oliner. “Efficient Data Reduction for Large-Scale Genetic Mapping”. In: *ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB '15)*. 2015.
- [176] Neeraja J. Yadwadkar, Bharath Hariharan, Joseph E. Gonzalez, and Randy Katz. “Faster Jobs in Distributed Data Processing using Multi-Task Learning”. In: *SIAM International Conference on Data Mining (SDM '15)*. 2015.
- [177] David Bader, Aydin Buluc, John Gilbert, Joseph E. Gonzalez, Jeremy Kepner, and Timothy Mattson. “The Graph BLAS effort and its implications for Exascale”. In: *SIAM Workshop on Exascale Applied Mathematics Challenges and Opportunities (EX14)*. 2014.
- [178] Joseph E. Gonzalez, Reynold S. Xin, Ankur Dave, Daniel Crankshaw, Michael J. Franklin, and Ion Stoica. “GraphX: Graph Processing in a Distributed Dataflow Framework”. In: *11th USENIX Symposium on Operating Systems Design and Implementation (OSDI 14)*. 2014, pp. 599–613.
- [179] Xinghao Pan, Stefanie Jegelka, Joseph E. Gonzalez, Joseph K. Bradley, and Michael I. Jordan. “Parallel Double Greedy Submodular Maximization”. In: *Neural Information Processing Systems (NIPS '14)*. 2014.
- [180] Veronika Strnadova, Aydin Buluc, Leonid Oliner, Joseph E. Gonzalez, Stefanie Jegelka, Jarrod Chapman, and John Gilbert. “Fast Clustering Methods for Genetic Mapping in Plants”. In: *16th SIAM Conference on Parallel Processing for Scientific Computing*. 2014.

- [181] T. Mattson, D. Bader, J. Berry, A. Buluc, J. Dongarra, C. Faloutsos, J. Feo, J. Gilbert, J. Gonzalez, B. Hendrickson, J. Kepner, C. Leiserson, A. Lumsdaine, D. Padua, S. Poole, S. Reinhardt, M. Stonebraker, S. Wallach, and A. Yoo. “Standards for graph algorithm primitives”. In: *2013 IEEE High Performance Extreme Computing Conference (HPEC)*. Sept. 2013, pp. 1–2. URL: <https://doi.org/10.1109/HPEC.2013.6670338>.
- [182] Xinghao Pan, Joseph E. Gonzalez, Stefanie Jegelka, Tamara Broderick, and Michael I. Jordan. “Optimistic Concurrency Control for Distributed Unsupervised Learning”. In: *NIPS ’13*. 2013. URL: <https://arxiv.org/abs/1307.8049>.
- [183] Evan Sparks, Ameet Talwalkar, Virginia Smith, Xinghao Pan, Joseph E. Gonzalez, Tim Kraska, Michael I. Jordan, and Michael J. Franklin. “MLI: An API for Distributed Machine Learning”. In: *International Conference on Data Mining (ICDM)*. IEEE. Dec. 2013. URL: <https://ieeexplore.ieee.org/abstract/document/6729619>.
- [184] Reynold Xin, Joseph E. Gonzalez, Michael Franklin, and Ion Stoica. “GraphX: A Resilient Distributed Graph System on Spark”. In: *SIGMOD Grades Workshop*. 2013. URL: <https://dl.acm.org/citation.cfm?id=2484427>.
- [185] Amr Ahmed, Mohamed Aly, Joseph Gonzalez, Shravan Narayanamurthy, and Alex Smola. “Scalable Inference in Latent Variable Models”. In: *Conference on Web Search and Data Mining (WSDM)*. 2012. URL: <http://www.cs.cmu.edu/~jegonzal/papers/ahmed%5C%5Fscalable%5C%5Finference%5C%5Fin%5C%5Flatent%5C%5Fvariable%5C%5Fmodels.pdf>.
- [186] Joseph E. Gonzalez, Yucheng Low, Haijie Gu, Danny Bickson, and Carlos Guestrin. “PowerGraph: Distributed Graph-Parallel Computation on Natural Graphs”. In: *OSDI ’12*. 2012. URL: <https://www.usenix.org/system/files/conference/osdi12/osdi12-final-167.pdf>.
- [187] Yucheng Low, Joseph E. Gonzalez, Aapo Kyrola, Danny Bickson, Carlos Guestrin, and Joseph M. Hellerstein. “Distributed GraphLab: A Framework for Machine Learning and Data Mining in the Cloud.” In: *Proceedings of Very Large Data Bases (PVLDB)*. Aug. 2012. URL: <https://arxiv.org/abs/1204.6078>.
- [188] Joseph E. Gonzalez, Yucheng Low, Arthur Gretton, and Carlos Guestrin. “Parallel Gibbs Sampling: From Colored Fields to Thin Junction Trees”. In: *Artificial Intelligence and Statistics (AISTATS)*. May 2011. URL: <http://proceedings.mlr.press/v15/gonzalez11a.html>.
- [189] Yucheng Low, Joseph E. Gonzalez, Aapo Kyrola, Daniel Bickson, Carlos Guestrin, and Joseph M. Hellerstein. “GraphLab: A New Parallel Framework for Machine Learning”. In: *Conference on Uncertainty in Artificial Intelligence (UAI)*. 2010. URL: <https://arxiv.org/abs/1006.4990>.
- [190] Joseph E. Gonzalez, Yucheng Low, and Carlos Guestrin. “Residual Splash for Optimally Parallelizing Belief Propagation”. In: *Artificial Intelligence and Statistics (AISTATS)*. Apr. 2009. URL: <http://proceedings.mlr.press/v5/gonzalez09a.html>.
- [191] Joseph E. Gonzalez, Yucheng Low, Carlos Guestrin, and David O’Hallaron. “Distributed Parallel Inference on Large Factor Graphs”. In: *Conference on Uncertainty in Artificial Intelligence (UAI)*. July 2009. URL: <https://arxiv.org/pdf/1205.2645.pdf>.

Preprint arXiv Technical Reports

- [1] Shiyi Cao, Ziming Mao, Joseph E. Gonzalez, and Ion Stoica. *K-Search: LLM Kernel Generation via Co-Evolving Intrinsic World Model*. 2026. eprint: 2602.19128. URL: <https://arxiv.org/abs/2602.19128>.
- [2] Mayank Mishra, Shawn Tan, Ion Stoica, Joseph E. Gonzalez, and Tri Dao. *MRNN: Non-Linear RNNs with Matrix-Valued States for Scalable Language Modeling*. 2026. eprint: 2603.14360. URL: <https://arxiv.org/abs/2603.14360>.

- [3] Abigail O'Neill, Alan Zhu, Mihran Miroyan, Narges Norouzi, and Joseph E. Gonzalez. *Cooperate to Compete: Strategic Coordination in Multi-Agent Conquest*. 2026. eprint: 2604.25088. URL: <https://arxiv.org/abs/2604.25088>.
- [4] Shiyi Cao, Yichuan Wang, Ziming Mao, Pin-Lun Hsu, Liangsheng Yin, Tian Xia, Dacheng Li, Shu Liu, Yineng Zhang, Yang Zhou, Ying Sheng, Joseph E. Gonzalez, and Ion Stoica. *Locality-Aware Fair Scheduling in LLM Serving*. 2025. eprint: 2501.14312. URL: <https://arxiv.org/abs/2501.14312>.
- [5] Alejandro Cuadron, Dacheng Li, Wenjie Ma, Xingyao Wang, Yichuan Wang, Siyuan Zhuang, Shu Liu, Luis Gaspar Schroeder, Tian Xia, Huanzhi Mao, Nicholas Thumiger, Aditya Desai, Ion Stoica, Ana Klimovic, Graham Neubig, and Joseph E. Gonzalez. *The Danger of Overthinking: Examining the Reasoning-Action Dilemma in Agentic Tasks*. 2025. eprint: 2502.08235. URL: <https://arxiv.org/abs/2502.08235>.
- [6] Hanchen Li, Runyuan He, Qiuyang Mang, Qizheng Zhang, Huanzhi Mao, Xiaokun Chen, Hangrui Zhou, Alvin Cheung, Joseph E. Gonzalez, and Ion Stoica. *Continuum: Efficient and Robust Multi-Turn LLM Agent Scheduling with KV Cache Time-to-Live*. 2025. eprint: 2511.02230. URL: <https://arxiv.org/abs/2511.02230>.
- [7] Kevin Lin, Charlie Snell, Yu Wang, Charles Packer, Sarah Wooders, Ion Stoica, and Joseph E. Gonzalez. *Sleep-Time Compute: Beyond Inference Scaling at Test-Time*. 2025. eprint: 2504.13171. URL: <https://arxiv.org/abs/2504.13171>.
- [8] Jintao Zhang, Kaiwen Zheng, Kai Jiang, Haoxu Wang, Ion Stoica, Joseph E. Gonzalez, Jianfei Chen, and Jun Zhu. *TurboDiffusion: Accelerating Video Diffusion Models by 100-200x*. 2025. eprint: 2512.16093. URL: <https://arxiv.org/abs/2512.16093>.
- [9] Alexander Krentsel, Peter Schafhalter, Joseph E. Gonzalez, Sylvia Ratnasamy, Scott Shenker, and Ion Stoica. *Managing Bandwidth: The Key to Cloud-Assisted Autonomous Driving*. 2024. eprint: 2410.16227. URL: <https://arxiv.org/abs/2410.16227>.
- [10] Shishir G. Patil, Tianjun Zhang, Vivian Fang, Noppapon C., Roy Huang, Aaron Hao, Martin Casado, Joseph E. Gonzalez, Raluca Ada Popa, and Ion Stoica. *GoEX: Perspectives and Designs towards a Runtime for Autonomous LLM Applications*. 2024. eprint: 2404.06921. URL: <https://arxiv.org/abs/2404.06921>.
- [11] Ion Stoica, Matei Zaharia, Joseph E. Gonzalez, Ken Goldberg, Koushik Sen, Hao Zhang, Anastasios Angelopoulos, Shishir G. Patil, Lingjiao Chen, Wei-Lin Chiang, and Jared Q. Davis. *Specifications: The Missing Link to Making the Development of LLM Systems an Engineering Discipline*. 2024. eprint: 2412.05299. URL: <https://arxiv.org/abs/2412.05299>.
- [12] Shuo Yang, Ying Sheng, Joseph E. Gonzalez, Ion Stoica, and Lianmin Zheng. *Post-Training Sparse Attention with Double Sparsity*. 2024. eprint: 2408.07092. URL: <https://arxiv.org/abs/2408.07092>.
- [13] Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. *MemGPT: Towards LLMs as Operating Systems*. 2023. eprint: 2310.08560. URL: <https://arxiv.org/abs/2310.08560>.
- [14] Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E. Gonzalez, and Ion Stoica. *Rethinking Benchmark and Contamination for Language Models with Rephrased Samples*. 2023. eprint: 2311.04850. URL: <https://arxiv.org/abs/2311.04850>.
- [15] Sarah Chasins, Alvin Cheung, Natacha Crooks, Ali Ghodsi, Ken Goldberg, Joseph E. Gonzalez, Joseph M. Hellerstein, Michael I. Jordan, Anthony D. Joseph, Michael W. Mahoney, Aditya Parameswaran, David Patterson, Raluca Ada Popa, Koushik Sen, Scott Shenker, Dawn Song, and Ion Stoica. *The Sky Above The Clouds*. 2022. DOI: 10.48550/ARXIV.2205.07147. URL: <https://arxiv.org/abs/2205.07147>.

- [16] Jeffrey Ichnowski, Kaiyuan Chen, Karthik Dharmarajan, Simeon Adebola, Michael Danielczuk, Victor Mayoral-Vilches, Hugo Zhan, Derek Xu, Ramtin Ghassemi, John Kubiatoewicz, Ion Stoica, Joseph Gonzalez, and Ken Goldberg. *FogROS 2: An Adaptive and Extensible Platform for Cloud and Fog Robotics Using ROS 2*. 2022. DOI: 10.48550/ARXIV.2205.09778. URL: <https://arxiv.org/abs/2205.09778>.
- [17] Brijen Thananjeyan, Justin Kerr, Huang Huang, Joseph E. Gonzalez, and Ken Goldberg. *All You Need is LUV: Unsupervised Collection of Labeled Images using Invisible UV Fluorescent Indicators*. 2022. DOI: 10.48550/ARXIV.2203.04566. URL: <https://sites.google.com/berkeley.edu/luv>.
- [18] Alan Pham, Eunice Chan, Vikranth Srivatsa, Dhruva Ghosh, Yaoqing Yang, Yaodong Yu, Ruiqi Zhong, Joseph E. Gonzalez, and Jacob Steinhardt. *The Effect of Model Size on Worst-Group Generalization*. 2021. DOI: 10.48550/ARXIV.2112.04094. URL: <https://arxiv.org/abs/2112.04094>.
- [19] Daniel Rothchild, Alex Tamkin, Julie Yu, Ujval Misra, and Joseph Gonzalez. *C5T5: Controllable Generation of Organic Molecules with Transformers*. 2021. DOI: 10.48550/ARXIV.2108.10307. URL: <https://arxiv.org/abs/2108.10307>.
- [20] Brijen Thananjeyan, Kirthivasan Kandasamy, Ion Stoica, Michael I. Jordan, Ken Goldberg, and Joseph E. Gonzalez. *PAC Best Arm Identification Under a Deadline*. 2021. DOI: 10.48550/ARXIV.2106.03221. URL: <https://arxiv.org/abs/2106.03221>.
- [21] Matthew A. Wright and Joseph E. Gonzalez. *Transformers are Deep Infinite-Dimensional Non-Mercer Binary Kernel Machines*. 2021. DOI: 10.48550/ARXIV.2106.01506. URL: <https://arxiv.org/abs/2106.01506>.
- [22] Doris Xin, Devin Petersohn, Dixin Tang, Yifan Wu, Joseph E. Gonzalez, Joseph M. Hellerstein, Anthony D. Joseph, and Aditya G. Parameswaran. *Enhancing the Interactivity of Dataframe Queries by Leveraging Think Time*. 2021. DOI: 10.48550/ARXIV.2103.02145. URL: <https://arxiv.org/abs/2103.02145>.
- [23] Vikram Sreekanti, Harikaran Subbaraj, Chenggang Wu, Joseph E. Gonzalez, and Joseph M. Hellerstein. “Optimizing Prediction Serving on Low-Latency Serverless Dataflow”. In: *CoRR* abs/2007.05832 (2020). arXiv: 2007.05832. URL: <https://arxiv.org/abs/2007.05832>.
- [24] Alvin Wan, Daniel Ho, Younjin Song, Henk Tillman, Sarah Adel Bargal, and Joseph E. Gonzalez. “SegNBDT: Visual Decision Rules for Segmentation”. In: *CoRR* (2020). arXiv: 2006.06868 [cs.CV]. URL: <https://arxiv.org/abs/2006.06868>.
- [25] Bohan Zhai, Tianren Gao, Flora Xue, Daniel Rothchild, Bichen Wu, Joseph E. Gonzalez, and Kurt Keutzer. “SqueezeWave: Extremely Lightweight Vocoders for On-device Speech Synthesis”. In: *CoRR* abs/2001.05685 (2020). arXiv: 2001.05685. URL: <https://arxiv.org/abs/2001.05685>.
- [26] Tianjun Zhang, Huazhe Xu, Xiaolong Wang, Yi Wu, Kurt Keutzer, Joseph E. Gonzalez, and Yuan-dong Tian. *BeBold: Exploration Beyond the Boundary of Explored Regions*. 2020. DOI: 10.48550/ARXIV.2012.08621. URL: <https://arxiv.org/abs/2012.08621>.
- [27] Xin Wang, Fisher Yu, Trevor Darrell, and Joseph E. Gonzalez. “Task-Aware Feature Generation for Zero-Shot Compositional Learning”. In: *CoRR* (2019). arXiv: 1906.04854 [cs.CV]. URL: <https://arxiv.org/abs/1906.04854>.
- [28] Vladimir Feinberg, Alvin Wan, Ion Stoica, Michael I. Jordan, Joseph E. Gonzalez, and Sergey Levine. “Model-Based Value Estimation for Efficient Model-Free Reinforcement Learning”. In: *CoRR* abs/1803.00101 (Feb. 2018). arXiv: 1803.00101. URL: <http://arxiv.org/abs/1803.00101>.
- [29] Noah Golmant, Nikita Vemuri, Zhewei Yao, Vladimir Feinberg, Amir Gholami, Kai Rothauge, Michael W. Mahoney, and Joseph Gonzalez. “On the Computational Inefficiency of Large Batch Sizes for Stochastic Gradient Descent”. In: *CoRR* abs/1811.12941 (Nov. 2018). arXiv: 1811.12941. URL: <http://arxiv.org/abs/1811.12941>.

- [30] J. Weston Hughes, Taylor Sittler, Anthony D. Joseph, Jeffrey E. Olgin, Joseph E. Gonzalez, and Geoffrey H. Tison. “Using Multitask Learning to Improve 12-Lead Electrocardiogram Classification”. In: *CoRR* abs/1812.00497 (Dec. 2018). arXiv: 1812.00497. URL: <http://arxiv.org/abs/1812.00497>.
- [31] Sicheng Zhao, Bichen Wu, Joseph Gonzalez, Sanjit A. Seshia, and Kurt Keutzer. “Unsupervised Domain Adaptation: from Simulation Engine to the RealWorld”. In: *CoRR* abs/1803.09180 (Mar. 2018). arXiv: 1803.09180. URL: <http://arxiv.org/abs/1803.09180>.
- [32] Richard Liaw, Sanjay Krishnan, Animesh Garg, Daniel Crankshaw, Joseph E. Gonzalez, and Ken Goldberg. “Composing Meta-Policies for Autonomous Driving Using Hierarchical Deep Reinforcement Learning”. In: *CoRR* abs/1711.01503 (Nov. 2017). arXiv: 1711.01503. URL: <http://arxiv.org/abs/1711.01503>.
- [33] Francois W. Belletti, Evan R. Sparks, Michael J. Franklin, Alexandre M. Bayen, and Joseph E. Gonzalez. “Scalable Linear Causal Inference for Irregularly Sampled Time Series with Long Range Dependencies”. In: *CoRR* abs/1603.03336 (2016). arXiv: 1603.03336. URL: <http://arxiv.org/abs/1603.03336>.
- [34] Joseph E. Gonzalez, Peter Bailis, Michael I. Jordan, Michael J. Franklin, Joseph M. Hellerstein, Ali Ghodsi, and Ion Stoica. “Asynchronous Complex Analytics in a Distributed Dataflow Architecture”. In: *CoRR* abs/1510.07092 (2015). arXiv: 1510.07092. URL: <http://arxiv.org/abs/1510.07092>.

Other Technical Reports and Invited Publications

- [1] David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluís-Miquel Munguia, Daniel Rothchild, David R. So, Maud Texier, and Jeff Dean. “The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink”. In: *Computer* 55.7 (2022), pp. 18–28. DOI: 10.1109/MC.2022.3148714. URL: <https://arxiv.org/abs/2204.05149>.
- [2] David E. Culler, Prabal Dutta, Gabe Fierro, Joseph E. Gonzalez, Nathan Pemberton, Johann Schleier-Smith, Kalyanaraman Shankari, Alvin Wan, and Thomas Zachariah. “CoVista: A Unified View on Privacy Sensitive Mobile Contact Tracing”. In: *IEEE Data Eng. Bull.* 43.2 (2020), pp. 83–94. URL: <http://sites.computer.org/debull/A20june/p83.pdf>.
- [3] Eric Jonas, Johann Schleier-Smith, Vikram Sreekanti, Chia-Che Tsai, Anurag Khandelwal, Qifan Pu, Vaishaal Shankar, Joao Menezes Carreira, Karl Krauth, Neeraja Yadwadkar, Joseph E. Gonzalez, Raluca Ada Popa, Ion Stoica, and David A. Patterson. *Cloud Programming Simplified: A Berkeley View on Serverless Computing*. Tech. rep. UCB/EECS-2019-3. EECS Department, University of California, Berkeley, Feb. 2019. URL: <http://www2.eecs.berkeley.edu/Pubs/TechRpts/2019/EECS-2019-3.html>.
- [4] Dan Crankshaw, Joseph E. Gonzalez, and Peter Bailis. “Research for Practice: Prediction-Serving Systems”. In: *Commun. ACM* 61.8 (July 2018), pp. 45–49. ISSN: 0001-0782. URL: <http://doi.acm.org/10.1145/3190574>.
- [5] Ion Stoica, Dawn Song, Raluca Ada Popa, David A. Patterson, Michael W. Mahoney, Randy H. Katz, Anthony D. Joseph, Michael Jordan, Joseph M. Hellerstein, Joseph E. Gonzalez, Ken Goldberg, Ali Ghodsi, David E. Culler, and Pieter Abbeel. *A Berkeley View of Systems Challenges for AI*. Tech. rep. UCB/EECS-2017-159. EECS Department, University of California, Berkeley, Sept. 2017. URL: <http://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-159.html>.
- [6] Rong Gu, Qianhao Dong, Haoyuan Li, Joseph E. Gonzalez, Zhao Zhang, Shuai Wang, Yihua Huang, Scott Shenker, Ion Stoica, and Patrick P. C. Lee. *DFS-Perf: A Scalable and Unified Benchmarking Framework for Distributed File Systems*. Tech. rep. UCB/EECS-2016-133. EECS Department, University of California, Berkeley, July 2016. URL: <http://www2.eecs.berkeley.edu/Pubs/TechRpts/2016/EECS-2016-133.html>.

- [7] Matei Zaharia, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J. Franklin, Ali Ghodsi, Joseph E. Gonzalez, Scott Shenker, and Ion Stoica. “Apache Spark: A Unified Engine for Big Data Processing”. In: *Commun. ACM* 59.11 (Sept. 2016), pp. 56–65. ISSN: 0001-0782. URL: <https://doi.acm.org/10.1145/2934664>.
- [8] Joseph E. Gonzalez. “From Graphs to Tables the Design of Scalable Systems for Graph Analytics”. In: *Proceedings of the 23rd International Conference on World Wide Web. WWW '14 Companion*. Seoul, Korea: ACM, 2014, pp. 1149–1150. ISBN: 978-1-4503-2745-9. URL: <https://doi.acm.org/10.1145/2567948.2580059>.

Journal Editor

- [2020] **Editor of the IEEE Data Engineering Bulletin Special Issue: Data Technologies Behind Digital Contact Tracing**
- [2018] **Editor of the IEEE Data Engineering Bulletin Special Issue: Machine Learning Life-cycle Management**

Selected Invited Talks

- [2026] **Invited Keynote at the Data Discovery Symposium** (UC Berkeley, May 2026) *Invited keynote at this interdisciplinary symposium on data-driven discovery.*
- [2026] **Invited Talk at Samsung Research America AI Summit** (San Jose, May 2026) *Finding a New North Star for AI Research in the Post AGI Era: Examining how scaling systems and humans enables capable AI agents.*
- [2026] **Invited Faculty Talk at the Amazon AI PhD Fellowship Summit** (Seattle, April 2026) *An Algorithm for Innovation: Invited faculty speaker at the annual Amazon AI Fellowship gathering of PhD fellows and faculty advisors.*
- [2025] **Invited Keynote at the Samsung AI Forum** (Online, September 2025) *Invited keynote talk on advances in AI systems research for Samsung’s annual AI research forum.*
- [2022] **Invited Speaker at the Cloud Intelligence and AIOps Workshop** *The Evolution of Model Deployment from the Cloud to the Sky*
- [2022] **Invited Speaker at the China Database Committee Conference** *The Co-Evolution of Data and ML Systems from the Cloud to the Sky*
- [2021] **UC Berkeley Office of the CTO Cloud Computing Meetup** *To the Cloud and Beyond A Brief View of Cloud Computing Research at UC Berkeley*
- [2020] **Speaker in the Colloquium Lecture Series hosted by the Department of Computer and Information Science at the University of Pennsylvania** *Advances in Machine Learning Systems Research*
- [2020] **Keynote Speaker at the 14th Annual Machine Learning Symposium at the New York Academy of Sciences** *Advances in AI-Systems Research*
- [2020] **Interviewed on This Week in Machine Learning Podcast**
- [2020] **Keynote at the International Solid-State Circuits Conference Forum on Machine Learning Processors** *Hardware Opportunities in the Machine Learning Lifecycle*

- [2020] **Keynote at the Workshop on MLOps (co-located and MLSys'20):** *Advances in Prediction Serving Systems*
- [2019] **Spoke in a Panel for Hispanic Engineers and Scientists (HES):** *Discussed career planning and my research trajectory to undergrads at UC Berkeley.*
- [2018] **Data Eng Conference Keynote** *Managing the Machine Learning Lifecycle*
- [2018] **ISAT TAMALE Workshop Presentation:** *This invitation only workshop covered the Tools and Algorithms for Machine learning. Gonzalez presented his work on prediction serving and experiment management systems.*
- [2017] **SOSP Workshop Keynote** *Research at the Intersection of AI and Systems*
- [2017] **AAAI Systems Workshop Keynote** *Rise to the Challenges of AI Systems*
- [2016] **ICML Workshop Keynote** *Prediction Serving: What happens after learning?*
- [2015] **NIPS Workshop Keynote** *Intelligent Services: Serving Machine Learning Predictions*
- [2015] **ODSC** *Intelligent Services: Serving Machine Learning Predictions*
- [2015] **CIDR Talk** *The Missing Piece in Complex Analytics: Low Latency, Scalable Model Management and Serving with Velox*
- [2014] **OSDI Conference** *GraphX: Graph Processing in a Distributed Dataflow Framework*
- [2014] **Annual meeting of the International Society for Bayesian Analysis (ISBA)** *Concurrency Control For Scalable Bayesian Inference.*
- [2014] **Tutorial at the International Conference for Machine Learning (ICML)** *Emerging Systems for Large-Scale Machine Learning.*
- [2014] **Session on Graph Algorithms Building Blocks at the International Parallel and Distributed Processing Systems (IPDPS)** *GraphX: Unifying Table and Graph Analytics.*
- [2014] **NetApp ATG University Day** *Large Scale Graph Analytics: Applications and Systems*
- [2014] **Keynote Speaker: Workshop on Big Graph Mining at the International World Wide Web Conference (WWW)** *From Graphs to Tables: The Design of Scalable Systems for Graph Analytics.*
- [2013] **SIAM CSE'13 Minisymposium Frontiers in Large-Scale Graph Analysis** *Large-Scale Graph-Structured Machine Learning: GraphLab in the Cloud and GraphChi in your PC*
- [2012] **OSDI Conference** *PowerGraph: Distributed Graph-Parallel Computation on Natural Graphs*
- [2011] **IDGA Data Center Consolidation Summit** *GraphLab: A New Parallel Framework for Machine Learning*
- [2010] **DARPA Future Ideas Symposium** *Invited speaker at the DARPA future ideas symposium.*
- [2009] **UAI Conference.** *Distributed Parallel Inference on Large Factor Graphs*

Advising

Postdocs

1. Jianfei Chen [2019-2021] *Tsinghua University*
2. Aditya Desai [2024-2026] *IIT Bombay (Assistant Professor)*
3. Ionel Gog [2018-2021] *Google Brain*
4. Kirthivasan Kandasamy [2019-2021] *UW-Madison*
5. Yifan Qiao [2024-2025] *Inferact (Staff Engineer)*
6. Alexey Tumanov [2016-2019] *Georgia Tech.*
7. Matthew Wright [2019-2021]
8. Yaoqing Yang [2019-2022] *Dartmouth*

PhD Candidates

1. Francois Belletti (PhD) [2015-2018]: **Thesis:** *Alternate Representations for Scalable Analysis and Control of Heterogeneous Time Series*
2. Asim Biswal [2023-Present]: *Studying LLM-powered query processing and optimization for relational data analytics.*
3. Shiyi Cao [2023-Present]: *Studying efficient LLM inference and serving systems, including speculative decoding and mixture-of-experts optimization.*
4. Daniel Crankshaw (PhD) [2015-2019]: **Thesis:** *The Design and Implementation of Low-Latency Prediction Serving Systems Google (Staff Software Engineer).*
5. Lisa Dunlap (PhD) [2021-2026]: **Thesis:** *So What's the Vibe? Data-Driven Diagnosis of Model Behavior Anthropic.*
6. Vlad Feinberg [2017-2018]: *Departed without degree; Distinguished Engineer at Google DeepMind and Gemini pre-training lead.*
7. Guangyu Feng [2025-Present]: *Studying asynchronous patterns for agentic tool use.*
8. Yu Gai [2019-Present]: *Studying graph neural networks for question answering.*
9. Paras Jain (PhD) [2018-2023]: **Thesis:** *Towards Robust and Scalable Large Language Models Genmo (CEO & co-founder).*
10. Dacheng Li [2023-Present]: *Studying AI systems for large-scale LLM serving and evaluation, with contributions to the LMSYS ecosystem and world model benchmarking.*
11. Hanchen Li [2025-Present]: *Studying agent-first data and serving systems, including KV-cache-aware scheduling for multi-turn LLM agents.*
12. Richard Liaw [2017-2020]: *Studying the design of systems for hyperparameter search.*
13. Charles Lin [2019-2020]: *Studying economics of serverless computing.*
14. Kevin Lin (PhD) [2019-2024]: **Thesis:** *Structured Contexts For Large Language Models Letta.*
15. Charles Packer (PhD) [2019-2024]: **Thesis:** *Building Agentic Systems in an Era of Large Language Models Letta (CEO & co-founder).*
16. Shishir Patil (PhD) [2019-2024]: **Thesis:** *Teaching Large Language Models to Use Tools at Scale Meta.*
17. Nathan Pemberton [2020-2022]: **Thesis:** *The Serverless Datacenter: Hardware and Software Techniques for Resource Disaggregation Amazon (Applied Scientist).*

18. Devin Petersohn [2015-2021]: **Thesis:** *Dataframe Systems: Theory, Architecture, and Implementation Snowflake.*
19. Suzanne Petryk (PhD) [2019-2024]: **Thesis:** *Reliable Multimodal Models Apple.*
20. Daniel Rothchild (PhD) [2018-2023]: **Thesis:** *Accelerating Electronic Structure Calculations with Machine Learning Genentech.*
21. Peter Schafhalter (PhD) [2019-2025]: **Thesis:** *Towards Cloud-Assisted Autonomous Driving Microsoft AI (Technical Staff).*
22. Eyal Sela (PhD) [2018-Present]: *Studying algorithmic methods for efficient deep learning inference.*
23. Vikram Sreekanti [2016-2020]: **Thesis:** *The Design of Stateful Serverless Infrastructure Herald.ai (CEO).*
24. Brijen Thananjeyan [2018-2022]: **Thesis:** *Safe Reinforcement Learning Using Learned Safe Sets Google DeepMind (Senior Research Scientist).*
25. Alvin Wan (PhD) [2018-2022]: **Thesis:** *Efficiently Designing Efficient Deep Neural Networks OpenAI.*
26. Xin Wang (PhD) [2016-2020]: **Thesis:** *The Design of Dynamic Neural Networks for Efficient Learning and Inference OpenAI (Technical Staff).*
27. Colin Wang [2025-Present]: *Studying multimodal AI systems and polyolithic agent architectures.*
28. Justin Wong (PhD) [2020-2026]: **Thesis:** *Mitigating Post-Training Effects on Generative Diversity in Language Models Amazon.*
29. Yichuan Wang [2024-Present]: *Studying storage-efficient vector indexing and RAG systems for personal devices. Learning based approaches to accelerate combinatorial optimization problems.*
30. Sarah Wooders (PhD) [2020-2024]: **Thesis:** *Ensuring Data Freshness Across Clouds for Model Serving Letta (CTO & co-founder).*
31. Tsung-Han (Patrick) Wu [2023-Present]: *Studying multimodal models and their evaluation, including hallucination reduction, false-premise correction, and vision-centric benchmarks.*
32. Parth Asawa [2024-Present]: *Studying continual learning and sample-efficient model adaptation.*
33. Huanzhi Mao [2025-Present]: *Studying LLM tool use and function calling; leads the Berkeley Function Calling Leaderboard (BFCL).*
34. Mihran Miroyan [2024-Present]: *Studying evaluation and benchmarking of search-augmented large language models.*
35. Mayank Mishra [2025-Present]: *Studying efficient LLM pretraining and novel model architectures for scalable language modeling.*
36. Xiangxi (Simon) Mo (PhD) [2023-2026]: **Thesis:** *Building Open Source Inference Serving Systems Inferact (CEO & co-founder).*
37. Abby O'Neill [2025-Present]: *Studying AI-AI interactions and cooperative multi-agent system design.*
38. Neeraja Yadwadkar (PhD) [2015-2018]: **Thesis:** *Machine Learning for Automatic Resource Management in the Datacenter and the Cloud UT Austin (Assistant Professor).*

39. Tianjun Zhang (PhD) [2018-2024]: **Thesis:** *System Architecture for Agentic Large Language Models Meta GenAI.*
40. Lianmin Zheng (PhD) [2019-2024]: **Thesis:** *Scalable and Efficient Systems for Large Deep Learning Models UCLA (Assistant Professor); xAI.*
41. Alan Zhu [2024-Present]: *Studying LLM personalization and agent adaptability in complex multi-agent environments.*

Masters Students

- | | | |
|---------------------------------|------------------------------------|-------------------------------|
| 1. Raghav Anand [2019-2020] | 15. Alexander Kristoffersen [2023] | 29. Chirag Sharma [2023] |
| 2. Alex Cao [2026] | 16. Alexander Ku [2019-2020] | 30. Linyue Song [2022] |
| 3. Anusha Dandamudi [2021] | 17. Raguvir Kunani [2022] | 31. Vikranth Srivatsa [2023] |
| 4. James DeLoye [2025] | 18. Scott Lee [2019-2020] | 32. Hari Subbaraj [2019-2020] |
| 5. Aman Dhar [2019-2020] | 19. Eric Leong [2022] | 33. Victor Sun [2019-2020] |
| 6. Mohamed Elgharbawy [2023] | 20. Debbie Liang [2022] | 34. Nikita Vemuri [2019-2020] |
| 7. Edward Fang [2019-2020] | 21. Eric Liu [2019-2020] | 35. Derek Wan [2021] |
| 8. Tianren Gao [2021] | 22. Yika Yujia Luo [2018-2019] | 36. Rachel Xin [2025] |
| 9. Daniel Ho [2019-2020] | 23. Kevin Miao [2022] | 37. Flora Xue [2019-2020] |
| 10. Roy Huang [2025] | 24. Ashwinee Panda [2019-2020] | 38. Edward Yam [2022] |
| 11. Samvit Jain [2017-2019] | 25. Alan Pham [2022] | 39. Andy Yan [2019-2020] |
| 12. Sona Jeswani [2019-2020] | 26. Aditya Ramkumar [2022] | 40. Fanjia Yan [2025] |
| 13. Alvin Kao [2019-2020] | 27. Vedaad Shakib [2022] | 41. Bohan Zhai [2021] |
| 14. Shreyas Krishnaswamy [2023] | 28. Jiarui Shan [2023] | 42. Corey Zumar [2017-2018] |

Undergraduate Research Assistants

- | | | |
|--------------------------|------------------------|----------------------|
| 1. Daiyaan Arfeen | 13. Weston Hughes | 25. Younjin Song |
| 2. Avinash Arjavalingam | 14. Yuxin Ji | 26. Ishaan Srivasta |
| 3. Michael Chau | 15. Henry Jin | 27. Priya Sundaresan |
| 4. Saurav Chhatrapati | 16. Sun Ah Lee | 28. Andrew Tan |
| 5. Harrison Constantinio | 17. Cathy Li | 29. Julie Yu |
| 6. Anusha Dandamudi | 18. Michael Luo | 30. Anthony Yu |
| 7. Ziyi Dou | 19. William Ma | 31. Bohan Zhai |
| 8. Rehan Durrani | 20. Sean Meng | 32. Ziyun Zhao |
| 9. Tianren Gao | 21. Mong Ng | 33. Kevin Zhu |
| 10. Noah Golmant | 22. Aniruddha Nrusimha | 34. Tyler Zhu |
| 11. Jennifer Grannen | 23. Aditya Ramkumar | 35. Yitian Zou |
| 12. Yijie Huang | 24. Allen Shen | |

Teaching

- **Introduction to Machine Learning [CS189/289A]:** Joseph co-teaches this large upper-division and graduate machine learning course at UC Berkeley covering foundational and modern ML topics including probability, regression, neural networks, optimization, transformers, and generative AI. Most recently taught Fall 2025 (co-taught with Narges Norouzi). For more information visit <https://eecs189.org>.
- **Principles and Techniques of Data Science [Data100]:** Joseph developed and teaches the large upper division data science class at UC Berkeley. This class covers topics ranging from data cleaning and visualization to machine learning and hypothesis testing with a focus on real-world data and problems. Students from across campus take this class which routinely has more than 1000 students a semester. Most recently taught Spring 2024, Fall 2024, Spring 2025, and Fall 2025. For more information visit <http://ds100.org>.
- **Conversations with Thought Leaders in Technology (CS198-100):** Joseph developed and co-hosted this class with Vik Singh external industry leader. This interactive zoom seminar class created in the middle of the COVID-19 pandemic leveraged the virtual format to invite world class technology leaders to speak with students about their careers and the lessons they learned. In this class, we heavily leveraged slack and zoom to have student write and vote on questions that were then used to drive the interactive interview. Speakers in this series included Henry Schuck, Sanjay Poonen, Mamoon Hamid, Archana Agrawal, Hilary Mason, and Nick Caldwell.
- **AI Systems Graduate Seminar [CS294-162]:** Joseph developed and teaches this graduate seminar covering key topics in machine learning and AI systems research, combining paper reading with structured discussions and class projects. Taught Spring 2019 (as CS294-159), Fall 2019, Spring 2022, and Fall 2023 (LLM Edition). For more information visit <https://ucbrise.github.io/cs294-ai-sys-sp22/>.
- **Foundations of Data Science [Data 8]:** Joseph co-taught this large introductory data science course at UC Berkeley. Taught Spring 2023.
- **Introduction to Databases [CS186]:** Joseph co-taught this 500+ student class with Joseph Hellerstein and helped to update content and create a series of new lectures covering concepts at the intersection of databases and data science. Taught Spring 2016.

Workshops

- **[2019] ACM-IMS Interdisciplinary Summit on the Foundations of Data Science:** I co-organized this event and moderated the panel: Deep Learning, Reinforcement Learning, and Role of Methods in Data Science.
- **[2019] SOSP ML Systems Workshop** I co-organized the SOSP learning systems workshop. (<http://learningsys.org/sosp19/index.html>)
- **[2019] NeurIPS ML Systems Workshop** I co-organized the NeurIPS learning systems workshop. (<http://learningsys.org/neurips19/>)
- **[2019] Workshop on Human In the Loop Learning (HILL)** I co-organized the first workshop on Human in the Loop Learning co-located with ICML
- **[2018] NIPS ML Systems Workshop:** I co-organized the NIPS learning systems workshop. (<http://learningsys.org/nips18/>)
- **[2018] NextProf Nexus Workshop** I co-organized this three day workshop intended to strengthen and diversify the next generation of academic leaders in engineering. (<http://nextprof2018.engin.umich.edu/nexus/>)

- **[2017] SOSP ML Systems Workshop I** launched and co-organized the first SOSP learning systems workshop. (<http://learningsys.org/sosp17/>)
- **[2017] NIPS ML Systems Workshop I** co-organized the NIPS learning systems workshop. (<http://learningsys.org/nips17/>)
- **[2016] ICML ML Systems Workshop I** co-organized the ICML learning systems workshop. (<https://sites.google.com/site/mlsys2016/>)
- **[2015] NIPS Learning Systems Workshop I** help co-organize the NIPS learning systems workshop. (<http://learningsys.org>)
- **[2014] DIMACS Workshop Organizer I** organized the DIMACS workshop on the Systems and Analytics of Big Data (<http://dimacs.rutgers.edu/Workshops/Analytics/>)
- **[2013] NIPS Workshop Organizer I** helped organize the third annual NIPS “Big Learning: Algorithms, Systems, and Tools” workshop. (<http://biglearn.org>)
- **[2012] NIPS Workshop Organizer I** helped organize the second annual NIPS “Big Learning: Algorithms, Systems, and Tools” workshop. (<http://biglearn.org>)
- **[2011] NIPS Workshop Organizer I** organized and led the workshop entitled “Big Learning: Algorithms, Systems, and Tools for Learning at Scale” For more information visit the workshop website <http://biglearn.org>
- **[2009] NIPS Workshop Organizer I** organized and led the first NIPS BigLearn workshop entitled “Large-Scale Machine Learning: Parallelism and Massive Datasets.” For more information visit the workshop website <http://www.select.cs.cmu.edu/meetings/biglearn09>

Funding Activities

- **[2019]** NSF National Science Foundation, CAREER: Towards the Design of Models and Systems for Efficient Prediction. UCB Award ID: 047303-001. PI: Joseph Gonzalez (UC Berkeley). Total: \$450,648.
- **[2018]** NSF National Science Foundation, ECDI: Secure Fog Robotics Using the Global Data Plane. UCB Award ID: 045561-001. Co-PI: Joseph Gonzalez (UC Berkeley). Total: \$2,000,000.
- **[2018]** NSF National Science Foundation, Expedition Award: Secure, Real-Time Decisions on Live Data. UCB Award ID: 044393-001. Co-PI: Joseph Gonzalez (UC Berkeley). Total: \$9,999,999. Project length: 5 years.
- **[2017]** DOD Advanced Research Projects Agency, Hierarchical Learning-Guided Automatic Model Discovery. UCB Award ID: 042431-001. Co-PI: Joseph Gonzalez (UC Berkeley). Total: \$3,811,944. Project length: 4 years.
- **[2017-present]** Raised gift funding for RISELab from Alibaba, Amazon, Ant Financial ARM, Ant, Capital One, Ericsson, Facebook, Futurewei, GE, Google, Huawei, IBM, Intel, Microsoft, Nvidia, Scotiabank, Splunk, and VMware. Roughly \$5,000,000 annually.
- **[2013]** Raised 6.75M in series A funding for the GraphLab Inc. startup.
- **[2010]** Applied for and was awarded a grant to have early access to the Intel Single-chip Cloud Computer (SCC) as part of the Many-core Applications Research Community.
- **[2008 - 2009] Helped Lead a DARPA Interdisciplinary Sciences and Technology Study (ISAT) Group** to investigate the future of parallel machine learning from an interdisciplinary perspective I also participated in the final Woodshole annual ISAT meeting to prepare a proposal for the DARPA director.
- **[2008]** Applied for and was awarded funding for (BAA 08-34) “Machine Learning and AI in the context of Multicore and Cluster Computing.”

University Service

- **[2025-2026] CS Division Vice Chair for Graduate Matters, UC Berkeley EECS:** Responsible for overseeing graduate student affairs in the Computer Science division, advising on graduate student petitions, academic standing, and policy matters.
- **[2025-2026] EECS Graduate Matters Committee:** Co-chair of the committee overseeing graduate curriculum, pedagogy, and student affairs for the EECS department.
- **[2025-2026] DIVCO (Divisional Council) Member — Elected (UC Berkeley Academic Senate, Berkeley Division):** Elected by the faculty to serve on the Divisional Council, the executive committee of the Berkeley Division of the Academic Senate, participating in campus-wide academic policy. Ran for this seat to help the campus address the looming challenges and opportunities of AI in higher education.
- **[2023-2024] EECS Undergraduate Study Committee (CS co-Vice Chair):** Co-chair of the EECS Undergraduate Study Committee, responsible for overseeing undergraduate curriculum and academic policy matters for the Computer Science division.
- **[2024-2025] EECS Undergraduate Study Committee (Member):** Continued service on the EECS Undergraduate Study Committee.

Reviewing

- [2027] MLSys General Chair
- [2026] NeurIPS Area Chair
- [2026] ICLR Area Chair
- [2026] ICML Program Committee
- [2026] ECCV Reviewer
- [2026] CVPR Reviewer
- [2026] ACL Rolling Review (ARR) Reviewer
- [2025] ICLR Reviewer
- [2024] NeurIPS Datasets and Benchmarks Track Area Chair
- [2024] ACL Rolling Review (ARR) Reviewer
- [2024] COLM Reviewer
- [2023] ICLR 2024 Area Chair
- [2023] NeurIPS 2023 Reviewer
- [2022] ICLR 2023 Area Chair
- [2021] OSDI Program Committee
- [2021] SysML Program Committee
- [2020] SysML Conference Workshop Committee
- [2020] EuroSys Program Committee
- [2019] SOSP Program Committee
- [2019] MLSys (formerly SysML) Program Committee
- [2018] ICML Area Chair
- [2018] LXAI Workshop Program Committee
- [2018] MLSys (formerly SysML) Program Committee
- [2018] Sigmod Program Committee
- [2017] AISTats Senior Program Committee
- [2016] VLDB PC Member (Best Reviewer Award)
- [2017] EuroSys PC Member
- [2016] ICML Reviewer
- [2016] AISTATS Reviewer
- [2014] OSDI External Review Committee
- [2014] HotCloud Review Committee
- [2014] ICML Reviewer
- [2013] Transactions on Pattern Analysis and Machine Intelligence
- [2013] NIPS Reviewer
- [2013] Super Computing Reviewer
- [2012] Parallel Computing Reviewer
- [2010] ICML Reviewer
- [2009] JMLR Reviewer
- [2007] JMLR Reviewer
- [2007] IPSN Reviewer

Commercial Involvement

- **Advisor to Inferact (2026-Present):** I serve as an advisor to Inferact, a startup commercializing vLLM that launched in January 2026 with \$150M in seed funding.
- **Advisor to Letta Inc. (2024-Present):** I serve as an advisor to Letta, a UC Berkeley Sky Lab spinout founded by Charles Packer and Sarah Wooders to build stateful LLM agent infrastructure (based on the MemGPT research project).
- **Advisor to Genmo Inc. (2023-Present):** I serve as an advisor to Genmo, a generative video AI company co-founded by my former student Paras Jain.
- **Co-founder of RunLLM (formerly Aqueduct, 2020-Present):** I co-founded this company with former students and I am head of product.
- **Advisory Board for Anyscale Inc. (2019-Present):** I am an advisor for Anyscale.
- **Advisory Board for DeepScale Inc. (2017-2019):** I was an advisor for DeepScale and was involved in technical discussions around vision and systems for autonomous driving.
- **Turi Inc. (2013-2016):** I co-founded Turi Inc. (formerly GraphLab Inc.) to commercialize my research and help lead the initial technology roadmap. When Turi was acquired by Apple Inc. in July of 2016 it had grown to over 50 employees and had raised nearly 25M in venture capital.
- **Yahoo! Research (2011):** Developed the next generation of the GraphLab abstraction to enable large-scale machine learning on natural graphs derived from social media and web-content. (*Alex Smola: smola@yahoo-inc.com*)
- **AT&T Labs Research (2007):** Developed models for statistically assessing DSL quality from limited noisy data. (*Steven Phillips: phillips@research.att.com*)
- **Intern at ADAPT (2006)** Worked on an automated AdWords auction agent. I developed and implemented models for assessing word value. (*Alex Bäcker: alex@caltech.edu*)
- **Microsoft Developer Internship (2005):** Worked with MSN Search team developing techniques to use behavioral information to identify search spam. (*Greg Hullender: greghull@windows.microsoft.com*)
- **Caltech Research Fellowship (2004):** Developed a new query-less search technology that uses prior reading interests to identify novel documents. (*Alex Bäcker: alex@caltech.edu*)
- **NASA Jet Propulsion Labs Fellowship (2003):** Developed a new algorithm for efficiently evaluating line-of-sight on digital elevation maps at JPL. (*Robert Chamberlain: rgc@jpl.nasa.gov*)

Publicly Released Software

- **vLLM:** High-throughput and memory-efficient LLM inference and serving engine; one of the most widely used open-source LLM serving systems. <https://github.com/vllm-project/vllm>
- **SGLang:** A structured generation language and high-performance serving system for LLMs and vision-language models. <https://github.com/sgl-project/sglang>
- **Arena (formerly LMSYS Chatbot Arena):** An open platform for evaluating LLMs through crowdsourced human preference judgments; the leading open benchmark for LLM quality. Originally launched as Chatbot Arena under the LMSYS research organization at UC Berkeley, spun off into Arena Intelligence Inc. in April 2025 (\$600M valuation), raised a \$150M Series A in January 2026 (valuation \$1.7B), and rebranded to Arena. <https://arena.ai>
- **RouteLLM:** A framework for cost-effective LLM routing, dynamically selecting between strong and weak models based on query difficulty. <https://github.com/lm-sys/RouteLLM>
- **Gorilla / Berkeley Function Calling Leaderboard (BFCL):** Gorilla is an LLM trained to generate accurate API calls; the associated BFCL is the leading benchmark for evaluating LLM tool-use and function-calling ability, now in V4 with agentic evaluation. <https://gorilla.cs.berkeley.edu>

- **MemGPT / Letta:** A framework for building stateful LLM agents with persistent, self-managed memory; originally published as MemGPT and rebranded to Letta, which was spun off into Letta Inc. (2024). <https://github.com/letta-ai/letta>
- **LEANN:** A low-storage vector index enabling RAG on personal devices with 97% storage savings over standard indexes while maintaining high recall; enables private, on-device semantic search over millions of documents. (MLSys 2026) <https://github.com/yichuan-w/LEANN>
- **Alpa:** A compiler for distributed deep learning. <https://github.com/alpa-projects/alpa>
- **SkyPlane:** Fast bulk data transfers across regions and clouds. <https://skyplane.org>
- **RLlib:** A distributed reinforcement learning system. <https://docs.ray.io/en/master/rllib.html>
- **Tune:** A system for distributed hyperparameter search and model development. <https://docs.ray.io/en/latest/tune.html>
- **Modin:** A system for scalable execution of Pandas dataframes. <https://modin.readthedocs.io/>
- **Clipper Prediction Serving System** A system for serving predictions across a wide range of machine learning frameworks. <https://clipper.ai>
- **Opaque:** Secure analytics on top of Apache Spark. <https://github.com/mc2-project/opaque>
- **GraphX:** A graph computation framework built into the widely adopted Apache Spark open-source project. <https://spark.apache.org/graphx/>
- **GraphLab/PowerGraph** A sophisticated API for building parallel and distributed machine learning algorithms on top of multicore and cloud architectures. GraphLab generalizes the MapReduce abstraction to support iterative asynchronous computation on graph structured dependent data. <http://graphlab.org>
- **Distributed SplashBP** This library implements the SplashBP algorithm for factor graph inference in the distributed setting using MPI. http://www.select.cs.cmu.edu/code/mpi_splash.tar.gz