# **Research Statement:** Cody Coleman (cody@cs.stanford.edu)

My goal in life, as well as research, is to enable everyone (particularly those with limited resources) to be successful by developing novel computer systems and algorithms that make technology, especially AI, more accessible and useful. As my journey from being born in prison to pursuing a PhD in CS at Stanford demonstrates [13, 21], we all have value, and the secret to maximizing the potential of human civilization is enabling people to find and unlock that value.

Consequently, to maximize the potential of AI, AI must reflect the interests and diversity of the people it impacts. Unfortunately, two trends run contrary to that goal: 1) a small number of well-funded teams increasingly dominate advances in AI, and 2) models have demonstrated negative biases that can adversely affect minority groups. Many of the recent breakthroughs in AI have depended on tremendous amounts of data, computational resources, and expertise that only a small fraction of people and organizations have access to [2, 6, 44]. As a result, advances in AI have become increasingly dominated by the most well-funded teams. At NeurIPS 2016, for example, over 1,000 paper authors were affiliated with American institutions, while 0 papers came from research groups on the African continent [32]. This lack of diversity has led to significant biases and oversights that have negatively impacted disadvantaged and marginalized groups [9, 43]. To increase the probability of positive outcomes, we need more diversity in the people designing applications, developing models, and monitoring their performance.

To this end, my research has focused on lowering the barrier to creating intelligent applications by reducing the cost of training and deploying state-of-the-art machine learning (ML) models. Going forward, I am actively working on novel abstractions that provide a simple but powerful framework for building end-to-end intelligent applications (think the Ruby on Rails of AI) as well as creating practical tools for combating bias in data.

## Research Background

To date, my research can be broadly summarized into three thrusts: machine learning for personalized education, benchmarking and improving machine learning performance, and efficient data selection for deep learning.

**Machine Learning for Personalized Education.** For my Master's thesis [18], I leveraged the wealth of data from Massive Open Online Courses (MOOCs) to understand student behavior and intervene to aid struggling students. Specifically, I found that teachers were heavily enrolled in courses and a largely untapped resource [37], showed that, while controlling for the effect of total time on-site, the number of study sessions was an important predictor of certification rate [31], and developed novel ML approaches to identify struggling students and intervene [19, 47].

**Benchmarking and Improving Machine Learning Performance.** While my Master's showed me the positive impact ML could have on education, and in turn society, I also saw that the computational requirements were a limiting factor. Even though I was fortunate to be apart of a well-funded lab, my experiments were slow and expensive, which was the polar opposite of my image of computer science. The reason I loved computer science was that anyone with a computer and an internet connection could make something that impacted the lives of people around the world. Seeing that ML was powerful but expensive even for well-funded organizations, I shifted toward making AI more accessible during my PhD and joined the Stanford DAWN project [6], which aims to democratize ML.

As a step in this direction, I began my PhD by benchmarking and analyzing ML systems. Despite considerable research on systems, algorithms, and hard to speed up ML workloads, there was no standard means of evaluating end-to-end system performance. Many prior ML performance benchmarks used throughput (either per-kernel or per-iteration) as a proxy for performance [1, 4, 5, 12, 23, 42], which ignores the impact optimizations can have on final model quality. To address this lack of end-to-end evaluation, I introduced training time and cost to a specified accuracy ("time-to-accuracy") as a measure of ML system performance in DAWNBench, the first open benchmark and competition for end-to-end deep learning training and inference [16]. Over the six months of the competition, we had submissions from several teams, including Google, Intel, and fast.ai, and there was a $477\times$ drop in training time and a $23\times$ drop in training cost over our initial multi-GPU seed entries without any loss in accuracy on the ImageNet dataset [15]. In light of DAWNBench's success, I became a founding member of MLPerf to expand our end-to-end approach to performance evaluation to more tasks and provide standard evaluation criteria for ML systems [30]. MLPerf now has over 50 supporting organizations with broad industry and academic involvement. Finally, I analyzed optimized submissions from DAWNBench and MLPerf to understand the impact (if any) these optimizations had on generalization and discover bottlenecks in the best-performing entries. Despite the stochasticity of ML training procedures and the heavy optimizations these submissions received, time-to-accuracy was relatively stable, and models optimized for time-to-accuracy generalized to unseen data nearly as well as standard pre-trained models. However, there was still room for improvement; top entries underutilized hardware by up to $10\times$ and had substantial communication overheads [14].

**Efficient Data Selection for Deep Learning.** In addition to looking for more efficient ways to use hardware, I have also looked for better ways to use data. While there has been work on improving data efficiency by selecting the most informative points to label (active learning) [29, 39, 40] or distilling a dataset down to a representative subset (core-set selection) [10, 11, 24, 27, 38, 45], many existing methods are too computationally expensive for deep learning. These data selection techniques depend on semantically meaningful features or uncertainty estimates from a trained model to identify the most informative examples, and unfortunately, these features are expensive in deep learning. Unlike in other areas of ML, the feature representations that these techniques depend on are learned rather than given, requiring substantial training times. To improve the computational efficiency of data selection in deep learning, I showed that small proxy models could be used to perform data selection (e.g., selecting data points to label for active learning) without significantly impacting final predictive performance. For active learning, this "selection via proxy" approach achieved up to a $41.9\times$ speed-up in data selection (i.e., the time it takes to repeatedly train and select points). For core-set selection, proxies that were over $10\times$ faster to train than their larger, more accurate target models can remove up to 50% of data without harming the final accuracy, making end-to-end training time savings possible [17].

## Ongoing and Future Work

Going forward, I am not only working on reducing the financial barrier to ML, but also lowering the learning curve for curating quality data, building ML models, and creating applications with those models.

**Building Diverse Datasets.** AI is increasingly deployed in real-world situations where there are diverse audiences and real consequences. Unfortunately, the traditional approach of collecting large training datasets and uniformly training over the data works well for the majority but fails for the long-tail. Image recognition systems perform worse for people of color [9] and can develop false associations, labeling black people as the class "basketball" or "Gorilla" and Asian people as the class "ping-pong" [43, 46]. Risk assessment systems have been shown to have significant racial disparities, mislabeling black people as high risk and white people as low risk [3]. Autonomous vehicles still struggle to accurately detect and predict certain vehicles despite having fleets of vehicles able to capture data [28, 41]. While a great deal of prior work on fairness has focused on better optimization procedures [8, 25, 48], this work treats the data as fixed, even though it can be an essential way to improve fairness in practice [26]. Inspired by the computational efficiency improvements from my "selection via proxy" approach above, I am moving beyond data selection as a way to improve coarse-grain metrics like accuracy and develop techniques to select examples that reduce bias, so AI works well for more than just the majority. For example, can we adapt traditional active learning methods like uncertainty sampling that implicitly optimizes for zero-one loss [33] to address worst-case risk [25] or other fairness criteria? Can we uncover false associations (e.g., black people are basketballs) by visualizing small, representative subsets of each class and then correct those associations by collecting counter-examples from large-scale unlabeled data? My prior work makes questions like this tractable for modern deep learning methods.

**Automating AI through High-Level Abstractions for Intelligent Applications.** Finally, even if we make a model less bias on a given task, the tasks we choose to solve will be bias if only a small group can wield AI. However, training, deploying, and monitoring these systems remains challenging even for the most well-equipped companies due to the hidden technical debt of modern ML systems [36]. Many have begun to create systems that simplify this process [7, 22, 34, 35], but most are proprietary and have not been open-sourced. Given my expertise in systems and ML, I want to accelerate the process of creating intelligent applications by developing an open-source framework that provides convenient abstractions like Ruby on Rails did for web applications, or map-reduce did for distributed computing. For example, rather than parsing data into tensors and mapping model outputs, models can be represented with a simple, functional interface that maps between high-level data types (e.g., "`def caption(image: Image): -> List[Character]`" for image captioning or "`def suggest(sentence: List[Character]): -> Tuple[Word, ...]`" for next word prediction) as an extension of Uber's Ludwig system [34]. Labels could come from a mixture of weak supervision labeling functions like Apple's Overton [35] or explicit feedback loops like Berkeley's Velox [20]. Then, a combination of transfer learning and AutoML techniques could turn these function signatures and labels into accurate models without practitioners ever needing to write code in TensorFlow or PyTorch. Unifying these disparate systems under a common and extensible framework that is grounded in the best practices of the field is squarely at the intersection of my expertise in machine learning and systems, and developing such a framework would not only raise many research questions across programming languages, systems, and machine learning, but also enable practitioners and domain experts to address a new wave of real-world problems.

## References


[1] Robert Adolf, Saketh Rama, Brandon Reagen, Gu-Yeon Wei, and David Brooks. Fathom: Reference Workloads for Modern Deep Learning Methods. In *IISWC*, pages 1–10. IEEE, 2016.

[2] Dario Amodei and Danny Hernandez. Ai and compute, 2018. https://blog.openai.com/ai-and-compute/.

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[4] Soheil Bahrampour, Naveen Ramakrishnan, Lukas Schott, and Mohak Shah. Comparative Study of Deep Learning Software Frameworks. *arXiv preprint arXiv:1511.06435*, 2015.

[5] Baidu. DeepBench: Benchmarking Deep Learning Operations on Different Hardware. https://github.com/baidu-research/DeepBench, 2017.

[6] Peter Bailis, Kunle Olukotun, Christopher Ré, and Matei Zaharia. Infrastructure for usable machine learning: The stanford dawn project. *arXiv preprint arXiv:1705.07538*, 2017.

[7] Denis Baylor, Eric Breck, Heng-Tze Cheng, Noah Fiedel, Chuan Yu Foo, Zakaria Haque, Salem Haykal, Mustafa Ispir, Vihan Jain, Levent Koc, et al. Tfx: A tensorflow-based production-scale machine learning platform. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1387–1395. ACM, 2017.

[8] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016.

[9] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA, 23–24 Feb 2018. PMLR.

[10] Trevor Campbell and Tamara Broderick. Automated scalable bayesian inference via hilbert coresets. *arXiv preprint arXiv:1710.05053*, 2017.

[11] Trevor Campbell and Tamara Broderick. Bayesian coreset construction via greedy iterative geodesic ascent. *arXiv preprint arXiv:1802.01737*, 2018.

[12] Soumith Chintala. Convnet-Benchmarks: Easy Benchmarking of All Publicly Accessible Implementations of Convnets. https://github.com/soumith/convnet-benchmarks, September 2017.

[13] Cody Coleman. Digging deeper: How a few extra moments can change lives, 5 2017. https://www.youtube.com/watch?v=stxJMsxxxtA.

[14] Cody Coleman, Daniel Kang, Deepak Narayanan, Luigi Nardi, Tian Zhao, Jian Zhang, Peter Bailis, Kunle Olukotun, Chris Ré, and Matei Zaharia. Analysis of dawnbench, a time-to-accuracy machine learning performance benchmark. *SIGOPS Oper. Syst. Rev.*, 53(1):14–25, 7 2019.

[15] Cody Coleman, Deepak Narayanan, Daniel Kang, Peter Bailis, and Matei Zaharia. Dawnbench v1 deep learning benchmark results, 2018. https://dawn.cs.stanford.edu/2018/04/30/dawnbench-v1-results/.

[16] Cody Coleman, Deepak Narayanan, Daniel Kang, Tian Zhao, Jian Zhang, Luigi Nardi, Peter Bailis, Kunle Olukotun, Chris Ré, and Matei Zaharia. Dawnbench: An end-to-end deep learning benchmark and competition. In *ML System Workshops at NIPS*, 2017. https://dawn.cs.stanford.edu/benchmark/papers/nips17-dawnbench.pdf.

[17] Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. Selection via proxy: Efficient data selection for deep learning. *CoRR*, abs/1906.11829, 6 2019.

[18] Cody A. Coleman. Identifying and characterizing subpopulations in massive open online courses. Master's thesis, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139, 2015.

[19] Cody A. Coleman, Daniel T. Seaton, and Isaac Chuang. Probabilistic use cases: Discovering behavioral patterns for predicting certification. In *Proceedings of the Second ACM conference on Learning@Scale conference*. ACM, 4 2015.

[20] 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. *arXiv preprint arXiv:1409.3809*, 2014.

[21] Angela Duckworth. *Grit: The power of passion and perseverance*, volume 234. Scribner New York, NY, 2016. pages 219 - 222.

[22] Jeffrey Dunn. Introducing fblearner flow: Facebook's ai backbone, 5 2016. https://engineering.fb.com/core-data/introducing-fblearner-flow-facebook-s-ai-backbone/.

[23] Google. TensorFlow Benchmarks. https://www.tensorflow.org/performance/benchmarks, 2017.

[24] Sariel Har-Peled and Akash Kushal. Smaller coresets for k-median and k-means clustering. *Discrete & Computational Geometry*, 37(1):3–19, 2007.

[25] Tatsunori B Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. *arXiv preprint arXiv:1806.08010*, 2018.

[26] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 600. ACM, 2019.

[27] Jonathan Huggins, Trevor Campbell, and Tamara Broderick. Coresets for scalable bayesian logistic regression. In *Advances in Neural Information Processing Systems*, pages 4080–4088, 2016.

[28] Timothy B. Lee. Autopilot was active when a tesla crashed into a truck, killing driver, 5 2019. https://arstechnica.com/cars/2019/05/feds-autopilot-was-active-during-deadly-march-tesla-crash/.

[29] David D Lewis and William A Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12. Springer-Verlag New York, Inc., 1994.

[30] Peter Mattson, Christine Cheng, Cody Coleman, Greg Diamos, Paulius Micikevicius, David Patterson, Hanlin Tang, Gu-Yeon Wei, Peter Bailis, Victor Bittorf, David Brooks, Dehao Chen, Debojyoti Dutta, Udit Gupta, Kim Hazelwood, Andrew Hock, Xinyuan Huang, Bill Jia, Daniel Kang, David Kanter, Naveen Kumar, Jeffery Liao, Deepak Narayanan, Tayo Oguntebi, Gennady Pekhimenko, Lillian Pentecost, Vijay Janapa Reddi, Taylor Robie, Tom St. John, Carole-Jean Wu, Lingjie Xu, Cliff Young, and Matei Zaharia. Mlperf training benchmark. 10 2019.

[31] Yohsuke Miyamoto, Cody A. Coleman, Joseph Williams, Jacob Whitehill, Sergiy Nesterko, and Justin Reich. Beyond time-on-task: The relationship between spaced study and certification in moocs. *Journal of Learning Analytics*, 2(2):47–69, 2015.

[32] Shakir Mohamed, Emily Muller, and Vukosi Marivate. Missing continents: A study using accepted nips papers, 7 2017. http://www.deeplearningindaba.com/blog/missing-continents-a-study-using-accepted-nips-papers.

[33] Stephen Mussmann and Percy S Liang. Uncertainty sampling is preconditioned stochastic gradient descent on zero-one loss. In *Advances in Neural Information Processing Systems*, pages 6955–6964, 2018.

[34] Yaroslav Dudin Piero Molino and Sai Sumanth Miryala. Introducing ludwig, a code-free deep learning toolbox, 2 2019. https://eng.uber.com/introducing-ludwig/.

[35] Christopher Ré, Feng Niu, Pallavi Gudipati, and Charles Srisuwananukorn. Overton: A data system for monitoring and improving machine-learned products, 2019.

[36] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. Hidden technical debt in machine learning systems. In *Advances in neural information processing systems*, pages 2503–2511, 2015.

[37] Daniel T. Seaton, Cody A. Coleman, Jon P. Daries, and Isaac Chuang. Enrollment in mitx moocs: Are we educating educators. *Educause Review (February 2015)*, 2015. http://er.educause.edu/articles/2015/2/enrollment-in-mitx-moocs-are-we-educating-educators.

[38] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

[39] Burr Settles. From theories to queries: Active learning in practice. In Isabelle Guyon, Gavin Cawley, Gideon Dror, Vincent Lemaire, and Alexander Statnikov, editors, *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, volume 16 of *Proceedings of Machine Learning Research*, pages 1–18, Sardinia, Italy, 16 May 2011. PMLR.

[40] Burr Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, 2012.

[41] David Shepardson. Tesla driver in fatal 'autopilot' crash got numerous warnings: U.s. government, 6 2017. https://www.reuters.com/article/us-tesla-crash/tesla-driver-in-fatal-autopilot-crash-got-numerous-warnings-u-s-government-idUSKBN19A2

[42] Shaohuai Shi, Qiang Wang, Pengfei Xu, and Xiaowen Chu. Benchmarking State-of-the-Art Deep Learning Software Tools. In *Cloud Computing and Big Data (CCBD)*. IEEE, 2016.

[43] Pierre Stock and Moustapha Cisse. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 498–512, 2018.

[44] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017.

[45] Ivor W Tsang, James T Kwok, and Pak-Ming Cheung. Core vector machines: Fast svm training on very large data sets. *Journal of Machine Learning Research*, 6(Apr):363–392, 2005.

[46] James Vincent. Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech, 1 2018. https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai.

[47] Jacob Whitehill, Joseph J. Williams, Glenn Lopez, Cody A. Coleman, and Justin Reich. Beyond prediction: First steps toward automatic intervention in mooc student stopout. In *Proceedings of the 8th International Conference on Educational Data Mining*. International Educational Data Mining Society, 2015.

[48] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340. ACM, 2018.