Daniel Kang

https://ddkang.github.io   ddkang@stanford.edu

Education

Stanford University: 2016 – 2021
- PhD candidate, Computer Science
- Advisors: Peter Bailis and Matei Zaharia

University of Cambridge: *2015 – 2016*
- Master of Advanced Studies, Mathematics

Massachusetts Institute of Technology (MIT): *2011 – 2015*, GPA: 5.0 out of 5.0
- Bachelor of Science, Computer Science and Mathematics (double major), GPA: 5.0/5.0
- Master of Engineering, Computer Science, GPA 5.0/5.0

Research

Stanford University, efficient and reliable query processing using machine learning (2016 – present)
- Built **NoScope**, a system for accelerating machine learning-based queries that use binary predicates.
- Built **BlazeIt**, a video analytics query engine that introduced FrameQL, a method for querying video datasets for spatiotemporal information of objects present in the video, and two novel optimizations for aggregation and limit queries.
- Developed **SUPG**, query semantics and algorithms for approximate selection queries with statistical guarantees on the recall of the query result, i.e., set of records (these guarantees are required for scientific analysis).
- Built **Smol**, a preprocessing-aware runtime environment for fast end-to-end DNN inference that alleviates the bottleneck of preprocessing in modern visual analytics tasks.
- Built **TASTI**, a method for indexing unstructured data (e.g., text and images) by cheaply clustering together similar records.
- Developed **Model Assertions**, a method of using assertions for monitoring ML models and use these errors as a form of weak supervision and active learning to improve model performance.
- Developed **Learned Observation Assertions**, a DSL and system for finding errors in human-generated and model-generated labels for ML pipelines. We are working to deploy learned observation assertions at an autonomous vehicle company.
- Developed **UAR**, a method for assessing the robustness of ML models to *unforeseen* adversaries, adversaries that are not known at train time. We used UAR to show that existing methods for assessing robustness can be misleading.
- Developed **DAWNBench** and **MLPerf**, a competition and benchmark that standardized "time-to-accuracy" as the metric of choice for machine learning performance.
- Developed **LIT**, a model compression technique that takes advantage of the repetitive structure in deep models for up to 5.5x compression.

MIT, computational biology and mathematics (2011 – 2015)
- Developed machine learning algorithms to study cell state (epigenetics).
- Investigated connections between classical and quantum mechanics by analyzing the Schrödinger operator over modified hyperbolic manifolds

---

Conference and Journal Publications
- **Daniel Kang**, Nikos Arechiga, Sudeep Pillai, Peter Bailis, and Matei Zaharia. "Finding Label and Model Errors in Perception Data with Learned Observation Assertions." In submission to *MLSys* (2021).
- **Daniel Kang***, John Guibas*, Peter Bailis, Tatsunori Hashimoto, and Matei Zaharia. "Task-agnostic Indexes for Deep Learning-based Queries over Unstructured Data." In submission to *MLSys* (2021).

- **Daniel Kang***, Yi Sun*, Dan Hendrycks, Tom Brown, and Jacob Steinhardt. "Testing Robustness Against Unforeseen Adversaries." In submission to *ICLR* (2021).
- **Daniel Kang**, Ankit Mathur, Teja Veeramacheneni, Peter Bailis, and Matei Zaharia. "Jointly Optimizing Preprocessing and Inference for DNN-based Visual Analytics." In *Proceedings of the VLDB Endowment* (2021).
- **Daniel Kang***, Edward Gan*, Peter Bailis, Tatsunori Hashimoto, and Matei Zaharia. "Approximate Selection with Guarantees using Proxies." In *Proceedings of the VLDB Endowment* (2020).
- **Daniel Kang**, Peter Bailis, and Matei Zaharia. "BlazeIt: Fast Exploratory Video Queries using Neural Networks." In *Proceedings of the VLDB Endowment* (2020).
- **Daniel Kang**, Ankit Mathur, Teja Veeramacheneni, Peter Bailis, and Matei Zaharia. "Jointly Optimizing Preprocessing and Inference for DNN-Based Visual Analytics." In *Proceedings of the VLDB Endowment* (2020).
- Peter Kraft, **Daniel Kang**, Deepak Narayanan, Shoumik Palkar, Peter Bailis, and Matei Zaharia. "A Demonstration of Willump: A Statistically-Aware End-to-end Optimizer for Machine Learning Inference." In *VLDB demonstration* (2020).
- **Daniel Kang** and Tatsunori Hashimoto. "Improved Natural Language Generation via Loss Truncation." In *Association for Computational Linguistics* (2020).
- **Daniel Kang***, Deepti Ragahavan*, Peter Bailis, Matei Zaharia. "Model Assertions for Monitoring and Improving ML models." *MLSys* (2020).
- Peter Kraft, **Daniel Kang**, Deepak Narayanan, Shoumik Palkar, Peter Bailis, and Matei Zaharia. "Willump: A Statistically-Aware End-to-end Optimizer for Machine Learning Inference." *MLSys* (2020).
- 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, Atsushi Ike, Bill Jia, **Daniel Kang**, David Kanter, Naveen Kumar, Jeffery Liao, Guokai Ma, Deepak Narayanan, Tayo Oguntebi, Gennady Pekhimenko, Lillian Pentecost, Vijay Janapa Reddi, Taylor Robie, Tom St. John, Tsuguchika Tabaru, Carole-Jean Wu, Lingjie Xu, Masafumi Yamazaki, Cliff Young, and Matei Zaharia. "MLPerf Training Benchmark." In *MLSys* (2020).
- **Daniel Kang**, Peter Bailis, and Matei Zaharia. "Challenges and Opportunities in DNN-Based Video Analytics: A Demonstration of the BlazeIt Video Query Engine." In *CIDR* (2019).
- Animesh Koratana*, **Daniel Kang***, Peter Bailis, and Matei Zaharia. "LIT: Block-wise Intermediate Representation Training for Model Compression." *ICML* (2019).
- Cody Coleman*, **Daniel Kang***, Deepak Narayanan*, Tian Zhao, Jian Zhang, Luigi Nardi, Peter Bailis, Kunle Olukotun, Chris Ré, and Matei Zaharia. "Analysis of DAWNBench, A Time-to-Accuracy Machine Learning Performance Benchmark." *ACM SIGOPS Operating Systems Review* (2019).
- Sandeep Chinchali, Apoorva Sharma, James Harrison, Amine Elhafsi, **Daniel Kang**, Evgenya Pergament, Eyal Cidon, Sachin Katti, and Marco Pavone. "Network offloading policies for cloud robotics: a learning-based approach." *Robotics: Science and Systems* (2019). **Finalist for Best Systems Paper and Best Student Paper.**
- **Daniel Kang**, John Emmons, Firas Abuzaid, Peter Bailis, and Matei Zaharia. "NoScope: optimizing neural network queries over video at scale." *Proceedings of the VLDB Endowment* (2017).
- 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." *NIPS MLSys Workshop* (2017): 102.
- **Daniel Kang**, Richard Sherwood, Amira Barkal, Tatsunori Hashimoto, Logan Engstrom, and David Gifford. "DNase-capture reveals differential transcription factor binding modalities." *PloS one* 12, no. 12 (2017): e0187046.

- Tatsunori Hashimoto*, Richard I. Sherwood*, **Daniel Kang***, Nisha Rajagopa, Amira A. Barkal, Haoyang Zeng, Bart J. M. Emons, Sharanya Srinivasan, Tommi Jaakkola, David K. Gifford. "A Synergistic DNA Logic Predicts Genome-wide Chromatin Accessibility". *Genome Research* (2016).
- Haoyang Zeng, Tatsunori Hashimoto, **Daniel Kang**, and David K. Gifford. "GERV: A Statistical Method for Generative Evaluation Of Regulatory Variants For Transcription Factor Binding". *Bioinformatics* (2015): btv565.
- Kiril R. Datchev, **Daniel Kang**, and Andre P. Kessler. "Non-Trapping Surfaces Of Revolution With Long-Living Resonances". *Mathematical Research Letters* 22.1 (2015): 23-42.
- Ira S. Moskowitz, Paul Cotae, and **Daniel Kang**. "Channel capacity behavior for simple models of optical fiber communication". 8th International Conference on Communications (COMMS 2010).

Workshop Publications

- **Daniel Kang***, Deepti Ragahavan*, Peter Bailis, Matei Zaharia. "Model Assertions for Debugging Machine Learning." *ICLR DebugML Workshop* (2019). **Contributed talk, best student paper**
- **Daniel Kang***, Yi Sun*, Tom Brown, Dan Hendrycks, and Jacob Steinhardt. "Transfer of Adversarial Robustness Between Perturbation Types." *ICML Uncertainty and Robustness in ML workshop* (2019).
- Animesh Koratana*, **Daniel Kang***, Peter Bailis, and Matei Zaharia. "Block-wise Intermediate Representation Training for Model Compression." *NeurIPS CDNNRIA Workshop* (2018).
- **Daniel Kang**, Peter Bailis, and Matei Zaharia. "Blazeit: An optimizing query engine for video at scale." *SysML* (2018).

Awards, Grants

- Co-authored and awarded a $75,000 grant from Stanford HAI for early wildfire detection (2020)
- Co-authored and awarded a $700,000 grant from Toyota to support safe ML (2018)
- Co-authored and awarded a $200,000 grant from Cisco to support ML benchmarking efforts (2018)
- Co-authored and awarded a $100,000 grant from Facebook to support ML benchmarking efforts (2017)
- National Science Foundation Graduate Research Fellowship (2016 – present)
- Churchill scholar (2015-16)
- Phi Beta Kappa (2015)
- Goldwater scholar (2014)
- MIT EECS SuperUROP Research and Innovation Scholar (2014)
- Grand Prize Winner of the Google Code-In (GCI) contest (2011)

Industry Experience

Toyota Research Institute Research Intern: *Winter 2020*
- Developed a new DSL and system, Learned Observation Assertions (LOA), for probabilistically finding human-generated and model-generated errors in ML pipelines.
- Currently working on deploying LOA in production to find errors in Toyota's labeling pipelines.

Dropbox Software Engineering Intern: *Summer 2016*
- Researched improvements to lossless compression of H264 video
- Researched various generic compression methods for Dropbox data
- Improved Brotli compression speeds by 40% or more

Google Software Engineering Intern: *Summer 2013*
- Expanded Google search's question answering system using natural language processing techniques, resulting in higher coverage of answerable questions

Google Software Engineering Intern: *Summer 2012*
- Parallelized FFmpeg's VP8 decoder to support sliced threading resulting in speed gains of up to 30% with two threads and 50% with four threads

- Researched and implemented new features (16x16 transform and selectable transform size) in Google's experimental video codec, resulting in improved compression
- Wrote unit tests with Google's testing framework for Google's experimental video decoder
- Refactored and rewrote the VP9 rate-distortion search and token parser for simplicity and speed
- Filed a patent for increased transform sizes in the VP9 codec

Emergent Analytics Intern: *January 2012*
- Investigated and implemented solutions in Java to stream live market data for the startup hedge fund, Emergent Analytics

Naval Research Laboratory Research Intern: *Summer 2009, 2010, 2011*
- Researched optical fiber communications and methods to remove digital image steganography, with Matlab, under the mentorship of Dr. Ira Moskowitz in 2009 – 10
- Investigated Android security and development in 2011

Video Encoding and Decoding Program Developer: *2010 – 2013*
- Optimized open source projects dealing with video encoding and decoding (x264 and FFmpeg/Libav). Sped up functions up to 40 times
- Implemented the majority of the Intel Sandy Bridge AVX support for x264
- Implemented the majority of the high bit-depth x86/x64 assembly for x264, with overall speedups of up to 15%
- Commissioned to optimize 10-bit H.264 decoding in FFmpeg/Libav for Intel x86/x64 SIMD. Decreased overall decoding times in half or more

CoreCodec Engineer: *2011*
- Optimized CoreCodec products (CoreAVC, CoreMVC) designed to decode MPEG-4 AVC video streams for Intel x86/x64 architecture (SIMD)
- Wrote unit tests in C to verify the correctness of optimized functions

---

Teaching and Mentorship

Research mentor, undergraduate John Guibas: spring 2020 – present
- Supervised John Guibas on TASTI, an indexing method for unstructured data
- Co-first authored a paper in submission to MLSys

Teaching assistant for CS197 (Stanford class to teach research to CS undergraduates):
- Supervised 12 undergraduate students to do introductory research
- Several groups went on to publish workshop papers as a result of the class

Research mentor, master's student Ankit Mathur: fall 2018 – spring 2019
- Supervised Ankit Mathur on a project for fast inference of CNN models
- Co-authored a submission to SIGMOD 2020

Research mentor, undergraduate Animesh Koratana: spring 2018 – spring 2019
- Supervised Animesh Koratana for LIT, a novel method for deep model compression
- Co-first authored a paper accepted to ICML
- Animesh won the Stanford CURIS best poster award based on LIT

Instructor for AddisCoder: summer 2018
- Taught introductory computer science to over 180 underprivileged students in Ethiopia
- Gave lectures and helped design the curriculum and assignments
- Lead a team of over 15 TAs, directed the creation of assignments

Teaching Assistant for AddisCoder: summer 2016
- Assisted teaching students in lab at a program for underprivileged students in Ethiopia
- Helped develop assignments for the program
- Students went out to enroll in colleges including MIT

Research mentor, undergraduate Logan Engstrom: summer 2015
- Supervised Logan Engstrom for DNase-capture, a technique to query the epigenetic state of a cell
- Co-authored a publication to PLoS ONE

Leadership, Service, and Reviewing
- Reviewer for NeurIPS (2020)
- Reviewer for ICML, top 33% (2020)
- Reviewer for NeurIPS (2019)
- Reviewer for ICML (2019)
- Reviewer for the NeurIPS Relational Representation Learning workshop (2018)
- Stanford PhD admissions committee member (2020)
- Stanford PhD admissions committee member (2019)
- Organized *InfoLunch*, a weekly discussion seminar for Stanford PhD students and faculty
- TechX's *TechTalks director* (2013 – 2014)
- *HackMIT organizer* (2013), a 1,000 person hackathon hosted at MIT
- *TechFair organizer* (2013), a tech expo held yearly at MIT, on the corporate relations committee