

Supun Nakandala

Phone: (+1) 812-558-6888
Email: snakanda@eng.ucsd.edu
Web: scnakandala.github.io

3232 EBU3B CSE
9500 Gilman Drive
La Jolla, CA 92093

Research Interests My research interest lies broadly in the intersection of Systems and Machine Learning, an emerging area which is increasingly referred to as *Systems for ML*. In this space I operate as a data management researcher. Taking inspirations from classical data management techniques, I build new abstractions, algorithms, and systems to improve efficiency, scalability, and usability of machine learning workloads.

Education **University of California**, San Diego, CA Sept 2017 - Present
PhD, Computer Science. GPA: 3.96/4.00
Thesis Advisor: Prof. Arun Kumar
Courses: Database System Implementation, Advanced Data Analytics Systems, Data Models in Big Data Era, Advanced Compiler Design, Machine Learning, Recommender Systems and Web Mining, Algorithm Design and Analysis

University of Moratuwa, Sri Lanka Aug 2010 - April 2015
Bachelor of the Science of Engineering, Computer Science & Engineering.
GPA: 4.11/4.20.
Department Topper and Gold Medalist

Publications *Incremental and Approximate Inference for Faster Occlusion-based Deep CNN Explanations*
Supun Nakandala, Arun Kumar, and Yannis Papakonstantinou
ACM SIGMOD 2019 (**Honorable Mention for Best Paper Award**)

Cerebro: Efficient and Reproducible Model Selection on Deep Learning Systems
Supun Nakandala, Yuhao Zhang, and Arun Kumar
ACM SIGMOD 2019 DEEM Workshop

Demonstration of Krypton: Optimized CNN Inference for Occlusion-based Deep CNN Explanations
Allen Ordookhanians, Xin Li, Supun Nakandala, and Arun Kumar
VLDB 2019 Demo | SysML 2019 Demo

Materialization Trade-offs for Feature Transfer from Deep CNNs for Multimodal Data Analytics
Supun Nakandala, Arun Kumar
SysML 2018 Short paper

Gendered Conversation in a Social Game-Streaming Platform
Supun Nakandala, Giovanni Cimpaglia, Norma Su, and Yong-Yeol Ahn
AAAI ICWSM 2017

Apache Airavata Security Manager: Authentication and Authorization Implementations for a Multi-Tenant eScience Framework
Supun Nakandala, Hasini Gunasinghe, Suresh Marru, and Marlon Pierce
IEEE e-Science 2016

Anatomy of the SEAGrid Science Gateway
Supun Nakandala, Sudhakar Pamidigantam, Suresh Marru, Marlon Pierce
NSF XSEDE 2016

Pre-Prints	<p><i>Vista: Declarative Feature Transfer from Deep CNNs at Scale</i> Supun Nakandala, Arun Kumar https://adalabucsd.github.io/papers/TR_2019_Vista.pdf</p> <p><i>Resource-Efficient and Reproducible Model Selection on Deep Learning Systems</i> Supun Nakandala, Yuhao Zhang, and Arun Kumar https://adalabucsd.github.io/papers/TR_2019_Cerebro.pdf</p> <p>Compiling Classical ML Pipelines into Tensor Computations for One-size-fits-all Prediction Serving Supun Nakandala, Gyeong-In Yu, Matteo Interlandi, and Markus Weimer https://adalabucsd.github.io/papers/TR_2019_Hummingbird.pdf</p>
Ongoing Projects	<div data-bbox="438 632 1435 978"> <p>Project Hummingbird Started June 2019</p> <p>In the past few years several optimized systems have been developed for accelerating deep learning prediction serving. However, in many domains, classical ML methods are still widely used. In this project we try to answer the following question: <i>Can we represent classical ML pipelines using tensor computations to reuse deep net prediction serving systems for classical ML prediction serving?</i> To realize this goal, project HUMMINGBIRD takes inspiration from query processing/optimization techniques in RDBMS and compiles classical ML pipelines into tensor computations. Experiments on real-world use cases show that HUMMINGBIRD enables significant speedups (even up to 10X) and seamless hardware acceleration for classical ML prediction serving compared to other existing systems.</p> </div> <div data-bbox="438 1010 1435 1388"> <p>Project Cerebro Started September 2018</p> <p>Deep neural networks are revolutionizing many ML applications. But there is a major bottleneck to wider adoption: the pain of <i>model selection</i>. This empirical process involves exploring the deep net architecture and hyper-parameters, often requiring hundreds of trials. Alas, most ML systems focus on training one model at a time, reducing throughput and raising costs; some also sacrifice reproducibility. We are developing Cerebro, which is a system to raise deep net model selection throughput at scale and ensure reproducibility. CEREBRO uses a novel parallel execution strategy we call <i>model hopper parallelism</i> which is inspired by the multi-query optimization technique. Experiments on <i>Criteo</i> and <i>ImageNet</i> datasets show CEREBRO offers up to 10X speedups and improves resource efficiency significantly compared to existing systems like Parameter Server, Horovod, and task-parallel tools.</p> </div>
Research Impact	<p>Ideas from project CEREBRO integrated into MADlib/Greenplum by Pivotal 2019</p> <p>CEREBRO system is being used by behavioral science researchers at UC San Diego medical school 2019</p> <p>“Gendered Conversation in a Social Game-Streaming Platform” paper gains lot of media attention and creates awareness about the bleak issue of sexism in online game streaming platforms 2017</p> <p>APACHE AIRAVATA science gateways middleware and the SEAGRID science gateway are widely used by computational science researchers to execute and manage computational jobs on university clusters and national supercomputing infrastructure 2017</p>
Work Experience	<p>Research Intern June 2019 - Sept 2019</p> <p><i>Microsoft Cloud Information Services Lab</i> Mentor: Matteo Interlandi, Markus Weimer Translating classical machine learning pipelines into tensor computations for unified and faster scoring of machine learning models.</p>

	Research Software Developer <i>Science Gateways Research Center - Indiana University</i> Manager: Marlon Pierce Developed APACHE AIRAVATA, which is a software framework to compose, manage, execute, and monitor large scale applications and workflows on distributed computing resources such as local clusters, supercomputers, computational grids, and computing clouds.	Oct 2015 - Aug 2017
Teaching Experience	Teaching Assistant - Advanced Data Analytics Systems	UCSD - Spring 2019
Scholarships and Awards	NSF travel award to attend ACM SIGMOD 2019 Gold Medal for the Best Academic Performance Travel award to attend 4 th South Asia Workshop on Research Frontiers in Computing Mahapola Higher Education Merit Scholarship	NSF - 2019 University of Moratuwa - 2015 National University of Singapore - 2014 Govt. of Sri Lanka - 2010
Technical Talks	<i>Resource-Efficient and Reproducible Model Selection on Deep Learning Systems</i> CNS Research Review 2019 <i>Compiling Classical ML Pipelines into Tensor Computations for One-size-fits-all Prediction Serving</i> <i>Incremental and Approximate Inference for Faster Occlusion-based Deep CNN Explanations</i> <i>Cerebro: A System for Efficient and Reproducible Model Selection on Deep Learning Systems</i> <i>Materialization Trade-offs for Feature Transfer from Deep CNNs for Multimodal Data Analytics</i>	UCSD Microsoft Cloud Information Services Lab 2019 ACM SIGMOD 2019 ACM SIGMOD 2019 UCSD CNS Research Review 2018