Samuel Randall
Email, Website

Applied Mathematician & ML Systems Engineer
Helping companies reduce GPU inference costs by 20-50% and improve ML system reliability at scale.

Education

Stanford University, Masters — August 2020 — December 2022
Computational & Mathematical Engineering
DGSAC Exceptional Master's Student Award, GPA: 3.86

Johns Hopkins University, Bachelors — September 2015 — June 2019
Majors: Applied Mathematics and Public Health.
Minors: Computer Science and Environmental Science.

Experience

moco, Machine Learning Researcher + Consultant — February 2025 — Present
Focus: reducing compute cost in large-scale ML inference.

- Developed pipeline consisting of an analysis step and then an optimization step to optimize ML classifiers.
  - Analyze the geometry of input data, latent representations, and model weights
  - Introduce early-exit and computation-skipping modifications based on the analysis.
- Reduced inference FLOPs by 20–50% and latency by 20–30%, <0.5% accuracy loss (in all cases) across TinyBERT sentiment classification, ResNet-18 image/audio models, and XGBoost fraud detection model.

Prime Health, Technical Lead — November 2025 — Present

- Co-architected and built the backend for an AI agent for personal health coaching. Designed the system's core logic and data layer; enabling accurate, deterministic, debuggable and auditable recommendations and actions.

BlueLightAI, Principal Applied Scientist — January 2023 - January 2025

- Promoted from SWE → Lead SWE → Principal Applied Scientist
- Improved accuracy of financial intent recognition ML model for refund-related queries from ~50% to 92% with clustering-classification pipeline to select corrective training data from a data lake.
- Presented results to Global 500 financial institution, securing a pilot engagement.
- Built and benchmarked clustering algorithms to debug ML models: detecting drifted, hard-to-classify, and mislabeled data.

Athena Security, Founding iOS Engineer — June 2020 — February 2022

- Built the iPad client for a computer vision system to detect fevers and weapons.

Skills

Systems & Optimization: Model Optimization, Machine Learning, AI Agents
Languages: Python, Swift, C++
Frameworks and Libraries: PyTorch, CVXPY, JAX, NumPy, PyGSP, NetworkX, FastAPI
Math: Graph Theory, Convex Optimization, Graph Algorithms, Geometric Algorithms.