

Mukesh Mithrakumar

 [mukeshmithrakumar](#)

 mukesh@mukeshmithrakumar.com

 +1 605-690 7739

 [mukesh-mithrakumar](#)

Professional Experience

LEAD MACHINE LEARNING ENGINEER

84.51°

Oct 2024 – Present
Plymouth Meeting, PA

TECHNICAL LEADERSHIP & ARCHITECTURE:

- Serve as principal technical decision-maker for Kroger search and personalization, owning architecture, design reviews, and multi-quarter platform strategy while providing hands-on technical leadership to 4 ML engineers, 2 feature platform engineers, and 3 data scientists across cross-functional engineering initiatives.
- Drive platform investments and technical roadmaps across product, science, data engineering, QA, front-end, and infrastructure teams, aligning architectural direction with business priorities and quarterly delivery milestones.

SEARCH, PERSONALIZATION & BACKEND SYSTEMS:

- Architected and implemented Kroger's multi-stage search and personalization system, engineering large-scale data pipelines, distributed model training, and semantic embeddings for a 40M+ product catalog; optimized retrieval, ranking, and inference to support 700ms P99.9 end-to-end search latency.
- Built a unified personalization API serving multiple customer experiences, integrating real-time recommendation models and asynchronous workflows on Kubernetes with Helm, PostgreSQL, Redis, and event-driven services, owning API design, data consistency, fault tolerance, and production reliability.
- Designed and productionized a two-stage real-time basket completion recommender integrating purchase history, session context, item co-occurrence, and personalization signals to drive incremental add-to-cart behavior.
- Built low-latency inference infrastructure supporting 10B+ monthly personalized inferences at 1K+ QPS and sub-100ms P99 latency using Vertex AI, GKE, Triton, TensorRT, and Cloud Run.
- Engineered a real-time product availability microservice using Kafka, Beam, Dataflow, and Bloom filters, integrating across search, personalization, and monetization systems to protect \$250M–\$500M in annual revenue.
- Delivered recurring quarter-over-quarter improvements across the search and personalization portfolio, contributing to 10–15% uplift in cart adds and revenue, 2–3% overall sales lift, and ~5% increase in sales per search.

ML PLATFORM & DISTRIBUTED INFRASTRUCTURE:

- Architected and built a feature engineering platform supporting real-time and batch features, enabling personalization using live behavioral signals, improving reliability and resilience, and reducing feature development cycles from 12 weeks to 3 days.
- Architected and built AI-powered productionization tooling, reusable repositories, and automated workflows spanning model optimization, testing, QA, endpoint deployment, observability, and rollout governance, reducing science-to-production delivery time from 16 weeks to 2 weeks.
- Own the MLOps platform architecture spanning distributed training, experimentation, model lifecycle management, and production inference; implemented training and data pipelines processing 1TB+ datasets using PyTorch, Kubeflow, Ray, Horovod, Kafka, Dataflow, Bigtable, and PySpark.

SENIOR MACHINE LEARNING ENGINEER

IQVIA

Aug 2020 – Oct 2024
Plymouth Meeting, PA

AI PRODUCTS & BUSINESS IMPACT:

- Architected and integrated an AI-powered analytics assistant into IQVIA's \$100M+ Orchestrated Analytics product portfolio, delivering near real-time insights to 100K+ users through LLMs, RAG, Text-to-SQL, entity extraction, and knowledge graphs.
- Designed and delivered Next Best Message, Content, and Action recommendation systems and an intelligent widget curation engine across 29 markets, achieving over 90% adoption, 5% click-through rate, and up to 15% uplift.
- Developed a demand forecasting system achieving 95% accuracy on 30-day predictions, improving inventory efficiency by 25%.

ML SYSTEMS & PLATFORM ENGINEERING:

- Developed and productionized distributed deep learning systems spanning recommendation models, CNNs, transformer-based forecasting, and fine-tuned language models using PyTorch, TensorFlow, Horovod, and Dask.
- Architected and built AWS-based MLOps infrastructure with automated CI/CD, reproducible training, and Kubernetes-native orchestration using Kubeflow and Argo, reducing model deployment timelines from months to days.
- Built and operated real-time and batch model serving infrastructure using TensorFlow Serving and TorchServe, supporting 2M weekly recommendations.

TECHNICAL LEADERSHIP & ENGINEERING ENABLEMENT:

- Drove cross-functional technical execution and quarterly and annual roadmaps across product, data science, research, DevOps, and engineering teams, coordinating initiatives involving 50+ contributors and supporting delivery of 15+ ML solutions annually.
- Mentored 10+ ML engineers through technical guidance, design discussions, and workshops, contributing to a 40% improvement in project delivery.

FOUNDER AND CTO ASTRUM AI SOLUTIONS

Aug 2018 - Aug 2020
San Francisco, CA

Built and delivered multiple end-to-end machine learning products for early-stage customers, spanning computer vision, NLP, and generative systems, including cardboard quality detection for recycling, automated check processing for low-resource languages in banking, research paper search and summarization, and generative image tools for an animation studio.

Education

MS. IN COMPUTER SCIENCE

Arizona State University

Tempe, AZ

BS. IN PHYSICS.

South Dakota State University

Brookings, SD

Teaching

INSTRUCTOR

- Associate Instructor, Machine Learning and Machine Learning Operations. (FourthBrain). Summer 2021 & 2022.
- Lead Instructor, Data Science. (General Assembly). Fall 2020.

TEACHING ASSISTANT

- Machine Learning (FourthBrain). Fall 2021.
- Physics (PHY 2048 & PHY 2049). Fall 2015 – Spring 2017

MENTORING

- Data Science (Springboard). 2018-May 2024.
- Data Science (Thinkful). 2018-2020.

Publications

JOURNAL

- Y. Jin, P. Kharel, P. Lukashev, S. Valloppilly, B. Staten, J. Herran, I. Tadic, **M. Mithrakumar**, B. Bhusal, A. O'Connell, K. Yang, Y. Huh, R. Skomski and D. J. Sellmyer. "Magnetism and electronic structure of CoFeCrX (X = Si, Ge) Heusler alloys". Journal of Applied Physics 2016; (Vol.120, Issue 5).

CONFERENCE

- Tong Wu, Mateusz Buda, **Mukesh Mithrakumar**, Yunlong Wang, Srikanth Sankaran Iyer, Tanveer Ahmed Nasir. "Leveraging Language Model for Next Best Action in Promotion Campaigns to Augment HCP Engagement". Pharmaceutical Management Science Association 2023. (Poster).
- **M. Mithrakumar**, B. Bhusal, Y. Huh and P. Kharel. "Tuning Magnetic Properties of Co₂FeGe with Cr Substitution for Fe", South Dakota Academy of Science 2016. (Poster).

WORKSHOP

- **Mukesh Mithrakumar**. "Training a Distributed NER model for drug, disease and condition identification", Healthcare NLP Summit 2021. ([Pre-recorded talk](#)).
- **Mukesh Mithrakumar**. "Synthesis, Structure and Magnetism of Co₂Fe_{1-x}Cr_xGe (x=0,0.25,0.5) alloy", Undergraduate Research, Scholarship and Creative Activity Day 2016. (**Certificate of Excellence for Outstanding participation in Undergraduate Research**).