Jaspreet Kaur Bhamra

jbhamra24@gmail.com | +1 (858) 241-1769 | linkedin.com/in/jaspreet-kaur-bhamra | github.com/jaspreetbhamra

PROFESSIONAL EXPERIENCE

CREYON BIO (Data Scientist - ML) — Jul 2023 - Aug 2025

Data Narratives from Data Analysis to Dashboard

- Identified behavioral deviations in data to uncover confounding variables affecting data distributions and established normalization procedures for consistent downstream analysis

Publication: Toxicity of Antisense Oligonucleotides is Determined by the Synergistic Interplay…

Statistical Modeling and Software

- Statistical and biophysical modeling of diverse data types (transcriptomic) using Monte Carlo and Bayesian inference
- Optimized workflow by 83% (compute time - 4hrs to 40mins) using Snakemake, Kubernetes, Docker and parallelization
- Optimized BigQuery tables using clustering/partition indexes, reducing query costs by ~70% per query (GCP)

Interpretable/Explainable Models and Machine Learning

- Engineered a custom suite of Generalized Additive Models (GAMs), developed for enhanced interpretability
- Integrated Monte Carlo sampling to generate probabilistic ensembles, to quantify prediction uncertainty
- Achieved 15% improvement in prediction accuracy through data cleaning and feature engineering (XGBoost, LGBM)

Deep Learning for RNA Biology

- Enabled downstream latent space exploration via a fine-tuned SpliceBERT as a featurizer
- Lead a PoC to establish the utility of Attention-based models for enhancing model explainability (PyTorch)

MLOps and CI/CD - Model Deployment, Monitoring, and Iteration

- Architected an ML model library using Python, Pydantic (compatible with the Scikit-learn API) for robust model and data provenance, reducing model training setup time from 2 hours to under 15 minutes

SAN DIEGO SUPERCOMPUTER CENTER (Machine Learning Engineer) — Apr 2022 - Jul 2023

Multimodal Deep Learning Model (SmokeyNet)

- Deep Learning Model to detect wildfire smoke using statistical and unstructured data (images) via PyTorch Lightning
- Designed a multimodal architecture leading to a 22% reduction in average smoke detection time (RNN, LSTM, ViT, VViT)
- Built ensemble architectures to also use multimodal time series data (statistical analysis, feature engineering, ML)
- Automated pipeline, integrating data from multiple sources based on fire alerts to source new sequences for training

Model Logging, Tracking, Optimization, Version Control

- Integrated model pipeline with WandB for MLOps (model tracking, logging, monitoring)
- Implemented Distributed CUDA Training in PyTorch Lightning to make use of multiple GPUs

Publications: Multimodal Wildland Fire Smoke Detection, **Workshop @ NeurIPS22** (arxiv.org/abs/2212.14143)
MDPI (mdpi.com/2072-4292/15/11/2790)

MORGAN STANLEY (Data Engineer) — Aug 2018 - Aug 2021

Micro-Batching Ingestion Framework (Data Warehouses)

- Automated the ETL setup process leading to time savings of around 70% per job
- Enabled cost savings of 1000x USD by automating migration of 60+ TB of production unstructured LOB data

DB Monitor: Using Predictive Modeling to Predict Outages

- Implemented statistical models using logs for anomaly detection, using data analysis and quantitative research
- Enabled early detection of issues helping to reduce database outages by 67% on average

SKILLS

ML & Deep Learning: PyTorch, PyTorch Lightning, HuggingFace Transformers, PEFT, scikit-learn, Constitutional AI, RLHF

LLM & Agents: LangChain, LangGraph, LangSmith, RAG, MCP, Vector DBs (FAISS, Chroma), Neo4j

Statistics & Analysis: Bayesian Inference, Monte Carlo Sampling, GAMs, PyMC, SciPy, NumPy, Pandas, Scikit-learn

Backend and Data Engineering: Python, SQL, Bash, Java, Snowflake, BigQuery, Greenplum, PySpark, ETL pipelines

MLOps & Infrastructure: Docker, Kubernetes, W&B, Jenkins (CI/CD), Git, Snakemake, GCP

Visualization & Deployment: Streamlit, Plotly, Tableau, Seaborn, Matplotlib

PROJECTS (TL;DR) - All Projects on GitHub

- Developing solutions using Claude Code/Codex, for faster iteration, using skill, commands and subagents
- Experimented with RAG workflows, multi-agent orchestration pipelines using LangGraph and ReAct-style reflection loops
- Research to better understand LLMs eg Mechanistic Interpretability of LLMs

EDUCATION

UNIVERSITY OF CALIFORNIA - SAN DIEGO — Jun 2023
Masters, Computer Science (Machine Learning) — CGPA: 3.76 / 4