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) Jul 2023 - present

Data Narratives from Data Analysis to Dashboard

- Lead statistical analysis of 10+ internal data modalities, identifying baseline distributions and quantifying variability
- Published a company-wide reference dashboard of "normal" ranges for all data modalities, enabling faster decision-making (Plotly and Streamlit)
- Identified behavioral deviations in data to uncover confounding variables affecting data distributions and established normalization procedures for consistent downstream analysis
- Contributed data analysis to : Toxicity of Antisense Oligonucleotides is Determined by the Synergistic Interplay of Chemical Modifications and Nucleotide Sequences, Not by Either Factor Alone

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, GCP and parallelization
- Optimized BigQuery tables using clustering/partition indexes, reducing query costs by ~70% per query

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
- Built a modular Python library fully compatible with the scikit-learn API, accelerating experimentation workflows
- Achieved 15% improvement in prediction accuracy through data cleaning and feature engineering

Deep Learning for RNA Biology

- Enabled downstream latent space exploration via a fine-tuned SpliceBERT featurizer
- Leading an ongoing PoC to establish the utility of Attention-based models for enhancing model explainability

MLOps and CI/CD

- Architected a Python-based ML model library using Pydantic 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
- 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)
- Implemented distributed 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)

- Developed a scalable ETL framework for multidimensional data to populate a data warehouse
- Automated the microbatch 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

Languages & Databases: Python, SQL, DB2, Sybase, Greenplum, Snowflake, Google BigQuery (GCP)

Frameworks: PyTorch, PyTorch Lightning, DataIKU, WandB, HuggingFace, Spark

DevOps & Other Tools: Agile, Docker, Kubernetes, JIRA, Git, Jenkins(CI/CD), Linux, Bash, Tableau, Streamlit

Packages & Utilities: NumPy, Pandas, scikit-learn, SciPy, Matplotlib, Seaborn, PySpark, PyMC, Pydantic, Plotly

EDUCATION

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