Kevin Yang

Data Scientist / Engineer

(604) 710-7454
Vancouver, BC
kevinyang10@gmail.com

keviny2.github.io/
github.com/keviny2
linkedin.com/in/keyang2

Skills

Tools and Languages	Python, R, PostgreSQL, Bash, AWS, snakemake, GATK, HPC (SLURM, LSF), Snowflake, Jenkins, MLflow, Airflow, Docker, tox, Git, Jupyter Notebook, Visual Studio, RStudio, pgAdmin, Vim, JIRA, CI/CD
Packages	PyTorch, NumPy, pandas, polars, huggingface, pysam, scikit-learn, statsmodels, SciPy, duckdb, psycopg, matplotlib, seaborn, altair, NumPyro, pytest, ggplot, dplyr, samtools
Machine Learning	Neural Networks/Deep Learning, Natural Language Processing (NLP), Transformers, Bayesian Networks, Computer Vision, XGBoost, Regression, Clustering and Classification, Decision Trees, Time Series, SVM

Technical Experience

Consultant @ Amaris Consulting

Senior Data Engineer **05/2023 — Present**
Genentech - A Member of the Roche Group *San Francisco, CA (Remote)*

- Generated a training set of over 35 million natural language sentences from single-cell metadata with Large Language Models (LLMs) for a multi-modal neural network model blending scRNA-seq data and structured metadata to characterize cells.
- Enhanced the performance of Large Language Models (LLMs) inference by over 1000 fold through strategic optimization leveraging NVIDIA's GPU Cloud (NGC), model quantization techniques, and efficient batch inference processes.
- Spearheaded the development of a software tool for accessing, querying, and analyzing a corpus of single-cell datasets, spanning 250 million cells with individual datasets exceeding 300GB, accelerating research by reducing metadata harmonization needs.
- Implemented an ETL workflow to migrate a 1TB single-cell metadata corpus from the cloud into an analytical database, resulting in a resource cost reduction of 70% and improved query speeds of 300%.
- Applied bioinformatics-specific Large Language Models to standardize over 18 million annotation terms, using biological ontologies as references, which encompassed over 20% of the entire single-cell database.

Data Scientist Intern **01/2023 — 04/2023**
Intact Financial Corporation *Vancouver, BC*

- Leveraged machine learning algorithms to optimize premium pricing, increasing profits by over 50%.
- Reduced feature engineering time by 70% using cluster computing and parallelization tools.
- Cut manual labor by 90% by developing automated pipelines for data ingestion, analysis, and model training.

Machine Learning Research Assistant **05/2020-12/2020**
BC Cancer Research Centre *Vancouver, BC*

- Implemented a clustering algorithm combining DNA-methylation and RNA-expression on pan-cancer datasets over 250GB in size.
- Refactored a pipeline to analyze high-dimensional DNA-methylation data, resulting in a >75% reduction in memory requirements through efficient data analysis and wrangling.

Projects (see GitHub)

SummarizedExperiment

- Contributed methods to combine advanced data structures optimized for processing datasets exceeding 500GB.
- Developed flexible subsetting methods for data analysis, increasing developer capacity by 20%.
- Authored comprehensive unit tests, achieving a code coverage of >90%.

Bayesian Probabilistic Graphical Model for Early Cancer Detection

- Integrated genomic variant data from both liquid and solid tissue biopsies into a machine learning model, successfully reducing error by 75% compared to state-of-the-art methods for inferring cancer population and subpopulation structure.
- Developed data simulation pipelines, which downsample Whole Genome Sequencing (WGS) datasets, to generate realistic high-dimensional biological data from liquid and tissue biopsies for benchmarking of machine learning model performance.

3D Image Reconstruction of Cancer Tumours

- Decreased validation error by 60% for spatial tissue analysis by implementing two cutting-edge deep learning models.
- Mitigated overfitting through weight decay and optimized loss functions, resulting in a 10% reduction in validation error.
- Accelerated neural network training over 100x using GPUs with CUDA in PyTorch.

Education

MSc. Bioinformatics - *GPA: 4.0*, *The University of British Columbia* 2021-2022

- *Canada Graduate Scholarships - Master's (CIHR) ($17 500 over 1 year)*

BSc. Computer Science & Statistics - *GPA: 3.9*, *The University of British Columbia* 2015-2020

- *Dean's Honour List*