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

Kevin Yang

Data Scientist / Engineer

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

Skills

Tools and Languages	Python, R, SQL, Bash, Linux, Git, Jupyter Notebook, Visual Studio, RStudio, Vim, SLURM, HPC, Snowflake, Jenkins, MLflow, Airflow, snakemake, pytest, tox, Agile, JIRA, CI/CD, TDD, Quarto
Packages	PyTorch, NumPy, PySpark, pandas, polars, scikit-learn, SciPy, duckdb, matplotlib, seaborn, NumPyro
Machine Learning	Neural Networks/Deep Learning, Bayesian Networks, NLP, Regression, Clustering and Classification, Decision Trees, XGBoost, Time Series, SVM, K-means, Hypothesis A/B Testing, ANOVA

Technical Experience

Consultant w/ Amaris Consulting

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

- Leverage ontologies and NLP methods to standardize >20 million annotations across >5000 experiments.
- Spearhead development of a data wrangling interface, improving the efficiency of data manipulation by 80%.
- Reduce 1-2 hours of manual labor per day by constructing a pipeline to automatically trigger database updates.
- Organize metadata from >2000 poorly annotated single-cell datasets (>180,000 cells) for effective retrieval.

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*

- Augmented a clustering algorithm on high-dimensional pan-cancer data over 250GB in size.
- Cut memory usage by 75% by creating an R package to process and wrangle high-dimensional datasets.
- Exploited numba and snakemake to parallelize data analysis pipelines, gaining 10x speed increases.

Projects (see GitHub)

SummarizedExperiment
- Contribute methods to combine advanced data structures optimized for processing datasets exceeding 500GB.
- Develop 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
- Reduce error by 75% over state-of-the-art methods by designing machine learning models to infer cancer population structure.
- Exploit HPC clusters to gain 10x speed up on pipeline executions.
- Create publication quality data visualizations to compare L1 losses for over 5 experiments.

3D Image Reconstruction of Cancer Tumours
- Decrease 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.
- Accelerate 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*