David Yang

davidheweiyang@gmail.com | (587)-889-2618 | Calgary, Canada
github.com/davidhy8 | linkedin.com/in/david-h-yang/ | davidhyang.com

EXPERIENCE

Data Mining Analyst Apr 2024 – Present
University of Calgary - Centre for Health Genomics and Informatics

- Built an **ETL pipeline** using Python to automatically retrieve, process and transform over 400,000 scientific papers for LLM.
- Engineered a caching object that uses **interval search trees** to efficiently store and manage previously processed data, including timeframes, reducing computational redundancy and improving performance when handling large datasets and repetitive requests.
- Optimized state-of-the-art **transformer architectures** (**BERT**) to classify sub-chunks of long text and integrated **gradient boosting** algorithms to overcome BERT's input limitation, achieving >95% recall in identifying scientific papers discussing viral mutations.
- Implemented **Named Entity Recognition** to extract key information from texts for a Flask web interface, enabling users to view, annotate, and save their annotations in a shared database accessible to all users.
- Developed a **self-training** mechanism that progressively improves **Large Language Models** and **gradient boosting** (LightGBM) frameworks by leveraging previous outputs and user annotations for continuous refinement and enhanced classification accuracy.

Web Automation Developer – Part-time Apr 2023 – Present
ADM Lucid Solutions Inc.

- Created automation test scripts with Selenium to validate web application functionality and data integrity (Cucumber, JMeter).

Machine Learning and Bioinformatics Researcher May 2018 – Mar 2024
University of Calgary

- Pinpointed ~50 out of >30,000 important genomic factors related to Glaucoma disease with R by employing **dimensionality reduction** (regularization, PCA), **data wrangling** (normalization, data imputation), and **statistical testing** techniques (Wald/LRT test, Bootstrapping, Regression) on noisy biological datasets with high dimensionality and multi-collinearity (>30,000 features).
- Generated scientific figures using **data visualization** libraries in R (ggplot2) which elucidated key research findings from **exploratory data analysis** to external institutions leading to the receival of monetary grants valuing $XXX,XXX.
- Created an asynchronous parallelization method for the **Markov chain Monte Carlo** (MCMC) algorithm involved in **Bayesian inference** (evolutionary) which reduced computational run-times by more than 2900% (~84 days).
- Identified selection bias in SARS-CoV-2 sequence collection by **analyzing** and **visualizing** COVID-19 data via Python & Tableau**.**
- Devised a novel representative **sampling strategy** based on scientific deductions of COVID-19 and implemented a **data pipeline** involving Python and Perl which reduced selection bias during SARS-CoV-2 sequence selection (n = >2 million) by around 100%.
- Conducted **benchmarking** and **optimizations** for bioinformatics software that leverages CUDA to parallelize computations.

PROJECTS & NOTABLE ACHIEVEMENTS

Restaurant food waste optimization (Python, scikit-learn, PyTorch, R-Shiny): Implemented ML-driven **forecasting models** for local restaurants to minimize food waste and boost revenue, providing **P&L** and opportunity cost analysis for enhanced procurement decisions.

Gamify life (NextJS, PostgreSQL): Designed and developed a web application that gamifies productivity and goal tracking, featuring user authentication, dynamic daily challenges, and a rewards system with XP and in-game gold.

Co-Founder & Chief Information Officer, *Canadian Organization for Undergraduate Health Research*
Led a not-for-profit organization across **four Canadian cities**, establishing a summer research program that pairs students with professors, design and developed a mobile Alzheimer's tracking app and oversaw student-led research and public education initiatives.

Web automation tutorials: Amassed over **150,000 YouTube views** through creating video tutorials teaching web automation testing.

TECHNICAL SKILLS

Languages & Technologies: Python, R, Java, SQL, Bash, Tableau, MS Excel, Git, Docker, Spark

Frameworks & Libraries: Data cleaning and data analysis (Pandas, NumPy, dplyr), data visualization (Matplotlib, ggplot2), machine learning (scikit-learn, TensorFlow, PyTorch), data mining (BeautifulSoup), automation (Selenium), web development (Flask, React)

EDUCATION

M.Sc. Mathematics and Statistics – Specialization: Statistics Sep 2021 – Mar 2024
University of Calgary | GPA 3.7/4.0 | Thesis project: Parallelization of MCMC Phylogenetic Analyses | Teaching: Calculus

B.Sc. First Class Honours, Cellular, Molecular, and Microbial Biology Sep 2017 - May 2021
University of Calgary | GPA 3.96/4.00 | Honours project: Eliminating Sampling Bias in COVID-19 Evolutionary Analysis