David Yang

davidheweiyang@gmail.com | (587)-889-2618 | Calgary, Canada

github.com/davidhy8 | linkedin.com/in/david-h-yang/ | davidhyang.com

TECHINAL SKILLS

Languages & Technologies: Python, R, Java, SQL, Tableau, MS Excel, Git, Minitab, UNIX, HTML, CSS, LaTeX

Frameworks & Libraries: Data wrangling and analysis (Pandas, NumPy, dplyr), Data Visualization (Matplotlib, ggplot2), Machine Learning (scikit-learn, Tensorflow), Webscraping (BeautifulSoup), Automation Testing (Selenium), Web development (Flask, R Shiny)

EDUCATION

M.Sc. Mathematics and Statistics – Specialization: Statistics — Sep 2021 – Mar 2024 (Expected)
University of Calgary | GPA 3.7/4.0 | Thesis project: Parallelization of MCMC Phylogenetic Analyses | TA: Calculus I
Coursework: Deep Learning, Generalized Linear Models, Statistical Inference, Bayesian Statistics, Theory of Probability

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 SARS-CoV-2 Analysis
Coursework: Computer science I & II, Calculus I & II & III (AU), Linear Methods I & II (AU), Special Topics in Computer Science

EXPERIENCE

Graduate Researcher — Sep 2021 – Present
University of Calgary — Calgary, Canada

- Pinpointed ~50 out of >30,000 significant 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 methods) on noisy biological datasets with high dimensionality and multi-collinearity.
- Generated scientific figures using **data visualization** libraries in **R** which elucidated key research findings to external institutions leading to the receival of monetary grants valuing greater than $50,000.
- 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%.
- Implemented **time-series analysis** in **R** on human blood biomarker data to identify key components related to cancer metastasis.

Web Automation Developer – Part-time — Apr 2023 – Present
ADM Lucid Solutions Inc. — Calgary, Canada

- Developed automation test scripts with **Selenium** and **Java** to validate the integrity of web applications (cucumber, POM, JMeter).
- Produced video tutorials discussing **automation testing frameworks** like Lighthouse, Netbeans, Docker, and Cucumber.

Undergraduate Researcher — May 2018 – Sep 2021
University of Calgary — Calgary, Canada

- Identified sampling bias in SARS-CoV-2 sequence collection by **analyzing** and **visualizing** COVID-19 data via **Python & R Shiny.**
- Devised a novel representative **sampling strategy** based on scientific deductions of COVID-19 and implemented a **software pipeline** involving **Python** and **Perl** to drastically reduce sampling bias during SARS-CoV-2 sequence selection.

Chief Information Officer, Co-Founder — Jun 2018 – Aug 2021
Canadian Organization for Undergraduate Health Research — Calgary, Canada

- Designed the framework for an Android mobile health tracking application (*palz*) with **Android SDK** in Android Studio (**Java**).
- Leveraged **data analytics** from social media platforms and website traffic to guide internal recruitment of five regional teams and various national committees which resulted in the employment of almost 100 individuals.

PROJECTS

NBA prediction web application: Python Flask web application that **scrapes** the internet for NBA data using **BeautifulSoup** and trains a neural network with hyperparameter tuning (**Tensorflow)** to predict NBA win-loss.

Image Classification with deep learning: Developed and deployed a **convolutional neural network** with **Tensorflow** that performs repurposed image classification tasks by building upon a model pretrained on the ImageNet dataset via **transfer learning**.

Predictive modelling for heart disease: Performed **logistic** and **lasso regression** analysis (i.e. feature selection & model evaluation) in **R** for a clinical dataset to select the model with the greatest prediction accuracy during the implementation of **cross-validation**.

Bayesian Inference of Zero-Inflated Dataset: Programmed custom Bayesian statistical models in **R** using **OpenBUGS** to model zero-inflated datasets with **Gibbs sampling** to obtain Bayesian credible intervals.