David Yang

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

SKILLS

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

Frameworks & Libraries: Data cleaning and Data analysis (Pandas, NumPy, dplyr), Data Visualization (Matplotlib, ggplot2), Machine Learning (scikit-learn, TensorFlow), Data mining (BeautifulSoup), Automation Testing (Selenium), Web development (Flask, 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, Linear Algebra I/II, Special Topics in Computer Science, Mathematical Statistics

EXPERIENCE

Machine Learning Researcher Sep 2021 – Present
University of Calgary Calgary, Canada

- 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 greater than $100,000.
- 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% (~84 days).
- Identified ~10 key components related to cancer metastasis via **time-series** & **statistical analysis** in R on human blood protein data.

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** (i.e. Lighthouse, NetBeans, Docker) reaching ~40,000 people.

Data Science 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 **data pipeline** involving Python and Perl which reduced sampling bias during SARS-CoV-2 sequence selection (n = >2 million) by around 100%.

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

- Designed the software framework for an Android mobile health tracking application (*palz*) with Android SDK and Java.
- Leveraged **data analytics** from social media platforms and website traffic to guide recruitment of five regional and national teams and advertisement of research program resulting in the employment of 100 individuals and more than 100 applicants for the program.

PROJECTS

NBA prediction web application: Python Flask web application that **web-scrapes** and **preprocesses** >8000 games of NBA data using BeautifulSoup and trains a **neural network** (scikit-learn, TensorFlow**)** to predict NBA win-loss with ~60% accuracy.

Image Classification with deep learning: Developed and deployed a **convolutional neural network** with TensorFlow that performs repurposed image classification by building upon a model pretrained on the ImageNet dataset via **transfer learning** with 98% accuracy.

Predictive modelling for heart disease: Engineered **logistic** and **lasso regression** (i.e. data preprocessing, data visualization, feature selection & model evaluation) **predictive models** in R for a clinical dataset with ~90% accuracy when evaluated with **cross-validation**.

Bayesian Inference of Zero-Inflated Dataset: Programmed custom Bayesian statistical models in R using OpenBUGS to **statistically model** zero-inflated datasets by approximation with the **Gibbs sampling** algorithm implemented from scratch**.**