Lawrence Du

Machine Learning • Biotechnology • Cloud

✉ larrydu88@gmail.com | ☎ 626-808-7096 | ⬤ github.com/LarsDu | ▦ linkedin.com/in/LarsDu

2911 McKinley Dr. Santa Clara, CA 95051

Skills

Techniques

Neural networks (CNNs, GraphNN, Transformers), Louvain/Leiden, k-fold cross-validation, SVMs, PCA, decision trees

Tools

Tensorflow, Pytorch, Jax, Numpy, Numba, Pandas, Sklearn, Conda, Flask, AWS, Google Cloud, Jenkins, Terraform, Docker, Metaflow, Kubernetes

Languages

Python, Rust, SQL, C#, Java, Dart, Bash, C/C++, Matlab, HTML/CSS, Some Mandarin and Spanish

Experience

Senior Machine Learning Platform Engineer • Freenome
Aug 2022 - Present (South San Francisco, CA)

- Built end-to-end scalable distributed machine learning pipelines using PyTorch, Ray, and Kubernetes for cancer detection from multi-omics data (deep sequencing cell-free methylated DNA, protein, and more).
- Built out a greenfield research engineering platform/library to support both distributed data parallel and model parallel training strategies for large-scale cancer detection models.
- Developed and refined methylated cell-free DNA based models for Colorectal Cancer (CRC) detection.
- Deployed and managed an organization-wide MLFlow based model tracking system for reproducibility, monitoring, and automated model performance reporting using Terraform, Pulumi, and Google Cloud.

Software Engineer - Machine Learning Engineering • 23andMe
Apr 2020 - Aug 2022 (Sunnyvale, CA)

- Built a large-scale feature engineering ETL pipeline for imputed SNPs (~10 million samples x ~1 million SNPs) using AWS Batch, Metaflow, AWS Glue, and AWS Athena used to feed downstream GWAS and Polygenic Risk Score (PRS) ML models.
- Improved PRS model AUCs and auPRCs through model stacking approaches.

Data Scientist - Ancestry Product • 23andMe
Nov 2018 - Apr 2020 (Sunnyvale, CA)

- Developed Recent Ancestor Locations (RAL) - a high precision, high recall country matching algorithm which serves >10 million customers worldwide.
- Deployed RAL using MLFlow and AWS Fargate
- Improved graph-based techniques for unsupervised identification of populations by genetically based identity-by-descent (IBD) family relationships.

Bioinformatician IV • Scripps Research
May 2018 - Oct 2018 (San Diego, CA)

- Developed a classifier for organ transplant rejection using RNA data.
- Wrote pipelines for Nanopore long-read sequencers using Common Workflow Language.

Independent Consultant • Juno Diagnostics
Sept 2017 - Feb 2018 (San Diego, CA)

- Developed patent – US20210020314A1 - Deep learning-based methods, devices, and systems for prenatal testing along with a Tensorflow based classifier for detecting prenatal genetic abnormalities from high throughput sequencing data.

Data Science Fellow • Insight
Jan 2017 - Apr 2017 (Remote Session - San Diego, CA)

- Built and deployed (as a Flask app on AWS EC2) DeepPixelMonster - a Tensorflow based GAN for creating pixel art, back when GANs were still relatively state-of-the art.

PhD Student Biology • UC San Diego • Scott A. Rifkin Lab
Aug 2010 - May 2017 (La Jolla, CA)

- Wrote DeepNuc - a CNN model for classifying over 500,000 transcriptional start site (TSS) flanking sequences from humans, mice, fruit flies, and nematodes as well as for over 60,000 microRNA target sequences.
- Researched the role of RNA expression noise during animal development by imaging single molecule RNA expression data in >5,000 embryos and analyzing data using self-written MATLAB tools for image segmentation, fluorescence quantification, and image deconvolution.

Education

Ph.D Biology UC San Diego, 2010 - 2017

B.A. Biological Sciences *Genetics and Development, Magna Cum Laude* Cornell University, 2006 - 2010

Activities and interests

- Developing the VR game Rogue Stargun (https://roguestargun.com)
- 3D modeling with Blender3D

- Painting
- DuBlog (https://dublog.net) - Self deployed using Google Cloud, Hugo, and Github Actions