Lawrence Du

Machine Learning • Data Science • Molecular Biology

✉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, GANs, GraphNN, Attention), Louvain/Leiden,*k*-fold cross-validation, SVMs, PCA, *k*-means clustering, decision trees

Tools
Tensorflow, Jax, Numpy, Numba, Pandas, Sklearn, Conda, Flask, AWS (EC2, S3, CloudFormation, Batch, Step Functions), Jenkins, Docker, Metaflow

Programming
Python, SQL, C#, Java, Dart, Bash, MATLAB/Octave, HTML/CSS, C/C++, Perl

Languages
Mandarin Chinese and some Spanish

Experience

Software Engineer II - Machine Learning Engineering • 23andMe

Apr 2020 - Present (Sunnyvale, CA)

- Worked on improvements in polygenic risk score modeling through model ensembling.
- Engineered a large-scale pipeline for calculating quality metrics of imputed single-nucleotide polymorphisms (~10 trillion datapoints) across an distributed cloud cluster (using AWS Batch, Metaflow, AWS Glue, and AWS Athena)

Data Scientist - Ancestry Product • 23andMe

Nov 2018 - Apr 2020 (Sunnyvale, CA)

- Developed current version of Recent Ancestor Locations (RAL) - a machine-learning based country matching algorithm which serves >10 million customers worldwide.
- Built Recent Ancestor Locations to run as a microservice on AWS backed hardware using MLflow for model artifact tracking.
- Improved graph-based techniques for unsupervised identification of populations by genetic relationships.

Bioinformatician IV • Scripps Research

May 2018 - Oct 2018 (San Diego, CA)

- Wrote robust automated sequencing pipelines for Oxford Nanopore data using Common Workflow Language (CWL) for realtime microbial diagnostics, *de novo* genome assembly, and variant calling.

Independent Consultant • Juno Diagnostics (startup - $25 million Series A closed in May 2021)

Sept 2017 - Feb 2018 (San Diego, CA)

- Developed patent – EP3773534A1 - Deep learning-based methods, devices, and systems for prenatal testingalong with a Tensorflow based deep learning software package for detecting prenatal genetic abnormalities.
- Created a genetic abnormality simulator derived from statistical analysis of human high throughput sequencing data.

Data Science Fellow • Insight

Jan 2017 - Apr 2017 (Remote Session - San Diego, CA)

- Wrote DeepPixelMonster and created an interactive Python Flask web application hosted on Amazon AWS integrated with a Tensorflow back-end for GAN based art generation.

PhD Student Biology • UC San Diego • Scott A. Rifkin Lab

Aug 2010 - May 2017 (La Jolla, CA)

- Performed research on RNA expression noise during animal development by imaging single molecule RNA expression data >5,000 embryos and analyzing data using self-written MATLAB tools for image segmentation, fluorescence quantification, and image deconvolution.
- Wrote DeepNuc - a CNN model for classifying over 500,000 transcriptional start site (TSS) flanking sequences from humans, mice, fruit flies, and nematodes and over 60,000 microRNA target sequences (from publically available CLEAR-CLIP data).
- Transgenically modified over 70 nematode and Drosophila strains using techniques such as GIBSON assembly, CRISPR/Cas9, MosSCI, and*PhiC31* integrase.

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

Hobbies	I enjoy running, painting. During the pandemic,I developed a VR space sim game
Extracurricular Activities	23andMe Spitballers Ultimate Frisbee • Machine Learning Society of San Diego • UCSD GSA Lobby Corps •BioEASI Art and Science Board • Genetics Training Program Grant • Hughes Scholar (2009) •Cornell Undergraduate Research Board (Vice President) • Friends of Farmworkers (tutor)