# Lawrence Du

**Machine Learning** • **Biotechnology**

mail[larrydu88@gmail.com](mailto:larrydu88@gmail.com) | phone626-808-7096 | github[github.com/LarsDu](http://github.com/LarsDu) | linkedin[linkedin.com/in/LarsDu](http://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, decision trees

#### Tools

Tensorflow, Pytorch, Jax, Numpy, Numba, Pandas, Sklearn, Conda, Flask, AWS (EC2, S3, Athena, CloudFormation), Jenkins, Docker, Metaflow

#### Languages

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

## Experience

### *Software Engineer - Machine Learning Engineering •* [*23andMe*](https://www.linkedin.com/company/23andme/mycompany/)

##### Apr 2020 - Present (Sunnyvale, CA)

* Built a large-scale ETL pipeline for calculating quality statistics of genotyped and imputed single-nucleotide polymorphisms (~10 trillion datapoints) across an distributed cloud cluster (using AWS Batch, Metaflow, AWS Glue) for populating an AWS Athena based feature datastore used to improve polygenic risk score models.
* Developed performance improvements for polygenic risk score models through model stacking.

### *Data Scientist - Ancestry Product •* [*23andMe*](https://www.linkedin.com/company/23andme/mycompany/)

##### Nov 2018 - Apr 2020 (Sunnyvale, CA)

* Developed current version of Recent Ancestor Locations (RAL) - a high precision, high recall country matching algorithm which serves >10 million customers worldwide.
* Built RAL 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 identity-by-descent (IBD) relationships.

### *Bioinformatician IV •* [*Scripps Research*](https://www.scripps.edu/science-and-medicine/cores-and-services/bioinformatics-core/index.html)

##### May 2018 - Oct 2018 (San Diego, CA)

* Wrote a classifier for organ transplant rejection from RNA sequencing data for a research grant.
* 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)*](https://www.linkedin.com/company/juno-diagnostics/)

##### Sept 2017 - Feb 2018 (San Diego, CA)

* Developed patent – [EP3773534A1 - Deep learning-based methods, devices, and systems for prenatal testing](https://patents.google.com/patent/EP3773534A1/en) along 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*](https://insightfellows.com/data-science)

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

* Wrote [DeepPixelMonster](https://github.com/LarsDu/DeepPixelMonster) - a Deep Convolutional Generative Adverserial Network (DCGAN) for creating pixel art as a week-long project. Also deployed aforementioned service as a Flask app on AWS.

### *PhD Student Biology • UC San Diego •* [*Scott A. Rifkin Lab*](http://labs.biology.ucsd.edu/rifkin/)

##### Aug 2010 - May 2017 (La Jolla, CA)

* Wrote [DeepNuc](https://github.com/LarsDu/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).
* 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.
* 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, and [3D modeling](https://sketchfab.com/3d-models/walken-23f4b6f0dcea4bf3afe714a59a9473e9). During the pandemic, [I developed a VR space game in my spare time](https://www.youtube.com/watch?v=HoC-xfeNZ84&t=107s)

**Extracurricular Activities**

23andMe Spitballers Ultimate Frisbee • Machine Learning Society of San Diego • UCSD GSA Lobby Corps • [BioEASI Art and Science Board](https://bioeasi.ucsd.edu/) • Genetics Training Program Grant • Hughes Scholar (2009) • [Cornell Undergraduate Research Board (Vice President)](https://www.cornellcurb.com/) • Friends of Farmworkers (tutor)