

PROGRAMMING: Python, SQL

DATABASES: PostgreSQL, SQLite3, MongoDB

INFRASTRUCTURES & FRAMEWORKS: AWS, Flask, Heroku

MACHINE LEARNING: Regression, Classification, Natural Language Processing, Clustering, Time Series Analysis

PACKAGES: Pandas, Numpy, Scikit-Learn, StatsModels, BeautifulSoup, Selenium, NLTK, Gensim, Pyspark, Matplotlib, Plotly, Tableau

EXPERIENCE

INVITAE

Dry Lab Operations Data Scientist

San Francisco
Jan. 2019 to Current

- Use **Tableau** to build analytics dashboards that help monitor internal processes
- Perform **root cause analysis** to diagnose customer issues

NEXTBEE MEDIA

Data Scientist

San Mateo
Sept. 2019 to Dec. 2019

Lead development of the Lighthouse App from inception to deployment. Available at: <https://lighthouse.nextbee.com/>

- Defined the **MySQL database** schema and the **SQL queries** used to build customer profiles based on orders data.
- **Segmented customers** into tier groups based on features identified through **domain knowledge** of ecommerce.
- Used the **time-series forecasting** method of SARIMA to predict future revenue and number of new customers.
- Used **logistic regression** and **random forest** to predict the likelihood of each customer making another purchase based on their purchase history.
- Made **interactive analytics dashboards** using **Plotly** and designed the **UI mock-ups** using the InVision App.
- Used **Git** to collaborate with other data scientists and front-end developers, and **helped non-technical staff** and clients understand analytics figures.

METIS

Data Science Fellow

San Francisco
Apr. 2019 to June 2019

- Completed four **business-oriented data science projects** as part of an immersive 12-week program focusing on classical machine learning, database management, deep learning, and project design.
- Wrote **documentation** to help future students get started: <https://harrisonized.github.io/2020/02/02/prepare-laptop-for-ds.html>

BIOVERATIV, FORMERLY TRUE NORTH THERAPEUTICS

Research Associate 2

South San Francisco
Jan. 2017 to Mar. 2019

- **Completed 2 research projects** on the structural biology of our lead drug. Independently designed and optimized experiments to test hypotheses.
- Performed **regression analysis** on protein-engineering data I collected in my experiments. Discovered a log-linear relationship between a physical property of our lead drug and its efficacy at treating disease, making it easy to decide which drug variants to use in downstream experiments.
- Wrote a **Python script** to **automate design** of short DNA oligos, which is over 200 times faster than manual design.
- Handled all molecular cloning for South SF site and **maintained sequence database** for over 200 DNA constructs.
- **Presented** findings weekly at lab meetings to **executives** and upper management.

GENE YEO LAB, UCSD

Staff Research Associate 1

La Jolla
May 2013 to Nov. 2016

- **Co-authored a Cell paper** that included my experiments on the application of a cutting-edge genome-editing technology to tracking RNA in live cells.
- **Co-authored a Neuron paper** that included my experiments on investigating the link between an RNA-binding protein and ALS.

FEATURED DATA PROJECTS

CLIMBING TRACKER WEB APPLICATION

Available at: <https://harrisonized-climbing-gpp.herokuapp.com>

- Built an **analytics dashboard** deployed to **Heroku Flask** to track personal climbing progress.
- Executes **SQL queries** and generates **interactive visualizations** in **Plotly** all on the server side.

YELP REVIEW CLASSIFIER AND TOPIC MODELING

Available at: <https://harrisonized.github.io/2019/06/05/yelp-climbing-gyms.html>

- Built a **web scraper** to collect Yelp data on California climbing gyms with storage in a **MongoDB NoSQL database**.
- Used **multi-class classification** on user reviews to predict the number of stars given by the reviewer. Adjusted the class weights to give minority classes more importance, which improved my out-of-sample accuracy score from 0.635 to 0.867.

PREDICTING POPULARITY OF ROCK CLIMBS

- Used **generalized linear models** to predict the number of users who would have a rock climb on their to-do-list. **Ensembled models** of log-linear and Poisson regression to improve the out-of-sample test score (R^2) from 0.643 of the baseline model to 0.842. The same strategy is used to predict the number of people who would have an item in their online shopping-list.

EDUCATION

University of California, San Diego

Double Major: B.S. Physics, B.S. Physiology & Neuroscience

2015