

**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

### NEXTBEE MEDIA

Data Scientist

San Mateo  
Sept. 2019 to Current

Led development of the Lighthouse App from inception to deployment.

Available at: <https://lighthouse.nextbee.com/>

- Defined the **MySQL database** schema and wrote the **SQL queries** to collect aggregate customer data from 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 the **binary classification** techniques of **logistic regression** and **random forest** to predict the likelihood of customers making another purchase.
- Developed a **Heroku Flask app** to share mock-ups of **interactive visualizations** made using **Plotly**.
- Made **user-interface** mock-ups using the InVision App.
- Used **Git** to collaborate with other data scientists and front-end developers.
- **Helped non-technical staff** and clients understand data analytics figures.

### METIS

Data Scientist

San Francisco  
Apr. 2019 to June 2019

Completed multiple **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.

### 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 myself. 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 proceed with in expensive experiments.
- Handled all molecular cloning for South SF site - **maintained database** of sequence data for over 200 DNA constructs.
- Wrote a **Python script** to **automate design** of short DNA oligos, which is over 200 times faster than manual design.
- **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

- Generated the input material for a high-throughput sequencing process, which were analyzed using **machine learning** techniques.
- **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.

## DATA PROJECTS

### YELP REVIEW CLASSIFIER AND TOPIC MODELING

- Built a **web scraper** to collect Yelp data on California climbing gyms.
- 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.
- Used **natural language processing (NLP)** techniques to **model topics** for 1-star and 5-star reviews.

### PREDICTING POPULARITY OF ROCK CLIMBS

Available at: <https://harrisonized.github.io/2019/05/08/mountain-project-recommender.html>

- Used a **Postgres SQL database** to minimize disk storage and memory usage.
- Used **generalized linear models** to predict the number of users who have a rock climb on their to-do-list on Mountain Project. **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.

### MEDICAL NOTES CLASSIFICATION

Available at: <https://www.github.com/harrisonized/medical-notes-kaggle>

- Created a **Python script** to **ingest data** from unstructured text files into a **MongoDB NoSQL database**.
- **Classified text** from medical notes into four clinical domains, achieving an accuracy score of 0.880 on the out-of-sample test set.

## EDUCATION

University of California, San Diego  
Double Major: B.S. Physics, B.S. Physiology & Neuroscience