HARRISON WANG
DATA SCIENTIST

harrison.c.wang@gmail.com
https://harrisonized.github.io/about
https://www.linkedin.com/in/harrisonized/

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

DATA PROJECTS

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.

CLIMBING TRACKER WEB APPLICATION

Available at: https://harrisonized-climbing-app.herokuapp.com

- Deployed a **Heroku Flask App** to track climbing progress.
- Executes **SQL queries** and generates **visualizations** all on the server side.

PREDICTING POPULARITY OF ROCK CLIMBS

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

- Created 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.

SF RENT PETITION TIME-SERIES ANALYSIS

- Used the **time-series analysis** method of SARIMA on counts of rent petitions in SF to forecast customer demand for up to five years.
- Used unemployment rate as a feature in **linear regression** to improve training and prediction speed while achieving similar accuracy.
- **Visualized** the distribution of rent petitions using **Tableau**.

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.

EXPERIENCE

METIS
San Francisco
Data Scientist
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
South San Francisco
Research Associate 2
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 our **executives** and upper management.

GENE YEO LAB, UCSD
La Jolla
Staff Research Associate 1
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.

EDUCATION

University of California, San Diego — 2015
Double Major: B.S. Physics, B.S. Physiology & Neuroscience