Luchao Qi

☎ (443)839-9129
✉ lqi9@jhu.edu
📍 3111 N Charles Street 4C
Baltimore, MD 21218

🔗 https://luchaoqi.github.io/
🔗 https://github.com/LuchaoQi
🔗 https://www.linkedin.com/in/LuchaoQi/

EDUCATION

The Johns Hopkins University May 2020

M.Sc.Eng. Biomedical Engineering

Northeastern University May 2018

B.S. Biomedical Engineering

SKILLS

Programming: Python, R, SQL, Shell Scripting

Packages: NumPy, Pandas, Scikit-Learn, NLTK, dplyr, tidyverse, Keras

Machine Learning: GLM, Random Forest, SVM, KNN, K-Means, PCA

Data Visualization: Tableau, Matplotlib, Seaborn, ggplot2, plotly

Data Science: A/B testing, NLP, Hadoop, Spark, HDFS

WORKING EXPERIENCE

Research Assistant, The Johns Hopkins Data Science Lab

Baltimore, MD | Sep 2019 - Present

Association analysis between lifestyle patterns and body mass index (BMI) via generalized linear model

- Wrangle time-series data of 32971 subjects and build pipeline to front-end dashboard using **SQL**
- Transform features using **normalization** to enhance machine learning pipelines
- Train a generalized linear model (**GLM**) to predict user BMI with 46.07 mean squared error (**MSE**)
- Reduce prediction error by 13% using feature selection method (**hypothesis testing, Random Forest**)
- Tested associations between BMI and physical activity with age, race and gender
- Identify statistically significant (p-value < 0.5) impact of lifestyle patterns on BMI to encourage the performance of multiple good health behaviors

Data Analyst Intern, The Johns Hopkins Bloomberg School of Public Health

Baltimore, MD | May 2019 – Aug 2019

Survival analysis in time-series data using Python, R

- Cleaned National Health and Nutrition Examination Survey (NHANES) data using **dplyr**, **tidyverse**
- Reduced dimensionality of data using **PCA** to capture essence of the data
- Selected features using **tree-based model**, **AIC/BIC** to achieve better predictive performance of model
- Constructed a neural network on 3000 patients using **Keras** to predict patient mortality with 71% accuracy for the purpose of benchmarking and performance evaluation of daily activities
- Improved classification accuracy to 86.45% using **regularized logistic regression**
- Hosted R shiny website comparing **PCA**, **k-means**, **UMAP**, **t-SNE** and visualizing clustering results using **ggplot2**, **plotly** (demo: https://luchaoqi.github.io/Shiny_clustering/#1)

Visiting Student Researcher, Paul C. Lauterbur Lab at SIAT

Shenzhen, CN | Nov 2016 - Jan 2017

EMG signal pattern recognition for hand gestures using spectral analysis

- Designed, constructed and assembled EMG data acquisition system for arm activities recognition
- Converted time-domain data of 200 gestures into frequency domain using **fast fourier transform** to denoise signal
- Classified different hand movements using support vector machines (**SVMs**) with 82% accuracy
- Improved accuracy by 3% training a **neural network** providing insight for medical rehabilitation system

SELECTED PROJECTS

Amazon Product Review Rating Prediction

June 2019 – Aug 2019

Detection of suspicious or fake Amazon product reviews using machine learning in Python

Demo: https://github.com/LuchaoQi/my-python/blob/master/amazon_project.ipynb

- Extracted Amazon Food Reviews data from Kaggle and cleaned data using **pandas**, **numpy** and **dfply**
- Tokenized unstructured text of user reviews using **scikit-learn** and **nltk** for feature construction
- Predicted customer rating categories using **logistic regression** with 0.94 AUC
- Reduced prediction error by 3% using **random forest** to better detect suspicious or fake online reviews

RNA-Seq - Next Generation Sequencing (NGS)

Nov 2018 - Jan 2019

Differential gene expression (DGE) analysis & Gene set enrichment analysis (GSEA) of RNA-Seq data

Demo: https://github.com/LuchaoQi/NGS

- Created tools (**Shell script, R, Python**) that can be used to perform one-stop analysis from downloading the raw gene data from Sequence Read Archive (**SRA**) to investigating the differentially expressed gene matrix
- Performed gene set enrichment analysis (**GSEA**) of **RNA-Seq** profiles obtained from Gene Expression Omnibus (**GEO**)
- Identified top 50 genes that are over-represented that may have an association with disease phenotypes

PUBLICATIONS

1. **Luchao Qi**, Brian Caffo, et al. Associations between Body Mass Index (BMI) and Physical Activity: National Health and Nutritional Examination Survey (NHANES) 2005-2006. Submitted to Am J Epidemiol 2019.
2. **Qi L**, Zhang Q, Tan Y, et al. Non-contact High-frequency Ultrasound Microbeam Stimulation: A Novel Finding and Potential Causes of Cell Responses. *IEEE Trans Biomed Eng* 2019.
3. **Qi L**, Zhang Q, Lam KH, et al. Calcium fluorescence response of human breast cancer cells by 50-MHz ultrasound microbeam stimulation. Presented at 2017 IEEE International Ultrasonics Symposium (IUS), 6-9 Sept. 2017 2017.