Yajing Gao

Data Scientist

SUMMARY

Data Scientist with a Ph.D. in Biomedical Engineering who is interested in integrating knowledge from science, math, and engineering to solve problem in novel and efficient ways. Experienced in general programming and algorithms, statistics and analytics, data science models and visualization.

EDUCATION

DUKE UNIVERSITY
Ph.D. in Biomedical Engineering
2016 | Durham, NC

DUKE KUNSHAN UNIVERSITY
Duke-Tsinghua Machine Learning Summer School
2016 | Kunshan, China

DUKE UNIVERSITY
B.S.E. in Biomedical Engineering, Mechanical Engineering, Chemistry, and Mathematics
2009 | Durham, NC

CONTACT

Email: yajing@gmail.com
Phone: 919.339.1119
Location: San Francisco, CA
Website: yajingg.github.io
Github: yajingg
LinkedIn: yajingg

SKILLS

COMPUTER SCIENCE
Java • C++ • Python • R
SQL • HIVE • MATLAB • HTML

DATA SCIENCE
Data Analysis • Data Visualization
Machine Learning • Model Building

SCIENCE
Statistics • Differential Equations
Numerical Analysis • Drug Delivery
Fluid Mechanics • Thermodynamics

EXPERIENCE

METIS | Data Science Fellow
Jan 2017 to Present | San Francisco, CA

- Developed several Data Science projects over the course of a 12-week immersive bootcamp, incorporating skills such as data wrangling, web scraping, regression, database, and deep learning.

DUKE UNIVERSITY | Graduate Research Assistant
2009 to 2016 | Durham, NC

- Thesis Topic: Mechanistic Models of Anti-HIV Drug Delivery.
- Used models based on physical processes to simulate drug delivery and drug effectiveness of HIV prevention products.
- Collaborated with non-profits in healthcare to test performance with clinical data, and deliver results as prototype in the pipeline for drug design.

DUKE UNIVERSITY | Teaching Assistant
2007 to 2011 | Durham, NC

- Assisted with Statistics, Mathematics, and Biomedical Engineering courses.
- Communicated solutions to students verbally during office hours and graded written homework.

PROJECTS

VISUALIZING AND DIFFERENTIATING NY SUBWAY STATIONS

- Worked with team to filter raw data from MTA website with Pandas, analyzed with Fast Fourier Transform to give phase and amplitude for rider frequencies.
- Visualized station ridership by animated dots over shape file that changes size and color to match the log of total riders.

ANALYZING TRENDS IN DOMESTIC BOX OFFICE

- Scraped raw data with Selenium from websites for movie and economic data. Blockbuster gross can be estimated by annual data on top 100 movies.
- Presented the optimization results which showed factors such as the economy and top blockbusters do not significantly contribute to decreased normalized movie sales.

CLASSIFYING AT FAULT DRIVERS IN FATAL ACCIDENTS

- Queried data with a structured relational database schema for all fatal crashes in 2015 from the National Highway Traffic Safety Administration.
- Used supervised learning to differentiate between at fault and not at fault drivers by age, gender, car make, model, and year. Showing male drivers 20 to 30 are more likely to be at fault in a fatal accident.

EMOJI ANALYSIS OF TWEETS

- Exploratory insights into emoji use by mining big data from twitter using Amazon Web Services (AWS) and MongoDB.
- Natural Language Processing (NLP) and unsupervised learning techniques from scikit-learn used to get quantitative features of emoji tweets. Visualized real time tweets with D3.