

## Experience

### IBM Watson Health – Data Scientist

Cambridge, MA Apr. 2017 -

† Trained word vectors using medical corpus which feed into recurrent networks to tag research papers with different genetic mutation effects. The system was deployed to preprocessing documents later be analyzed by biologists.

† Designed Avro data schemas for claim data for ingesting and publishing. Implemented ETL pipeline to transfer claim data from data lake to data reservoir for downstream application.

† Built knowledge graphs and implemented graph traversal strategies to facilitate validation and inference on the entity relationships extracted from natural language models.

### Fidelity Investments – Data Science Intern

Smithfield, RI Jun. 2016 - Sept. 2016

† Worked on projects for the AI lab including face recognition, voice assistant and text classification with neural networks.

### Zhenjiang Library – Database Engineer

Zhenjiang, China Nov. 2010 - Feb. 2015

† Developed and maintained PL/SQL package for analytic use cases and report generation.

## Competitions

† KDD Cup 2018 – Top 30

Forecasting hourly air quality measures for various locations in Beijing and London over the next 48 hours using neural networks.

† Kaggle Expert, currently 350 in competition rank.

† TensorFlow Speech Recognition Challenge – Top 4%

Google Brain's competition on small foot-print keyword spotting. Ensemble a collection of deep learning models includes ResNet, VGG, LSTM recurrent networks, convolutional RNN and depth separable CNN based on recent research papers. Beat Google's benchmark model by 11% in accuracy.

† Jigsaw Toxic Comment Classification - Top 4%

Classifying Wikipedia comments into sentiment classes using fastText word embedding with recent published deep learning architectures like VDCNN, DPCNN and LSTM networks with max k pooling.

† Safe Driver Prediction – Top 2%

Competition on ranking insurance claim holder on their risk of filing a claim in the next year. Stacking an army of gradient boosting machines, regularized greedy forests, factorization machines and neural networks with entity embedding.

## Research

† Zhou L., Zhang R., Chakraborty P., Farooq F., Hensley Alford S., *Comparative Effectiveness of Edoxaban and Warfarin in Prevention of Stroke and Systemic Embolism in Non-valvular Atrial Fibrillation Using Observational Healthcare Data*, in ISPOR 2018

† Babaian T., Zhang R., Lucas W. *DTMi – a New Interface for Informed Navigation*, in proceedings HCI International 2017

## Education

### Bentley University – Master of Science in Business Analytics / Data Science 4.0

Sept 2015 - Dec 2016

† Coursework: Programming, Databases, Hadoop, Time Series, Optimization, Statistical Learning, Data Mining

† Research Assistant: Worked with Prof. Babaian on developing D3 interactive data visualization for graph datasets.

### Udacity – Data Analyst Nano Degree

Nov 2015 – Feb 2016

† Series of projects on statistical testing, data munging, visualization, machine learning and A/B testing.

## Skills

† Deep Learning, Machine Learning, Statistic Analysis, Optimization, Data Analysis, Software Development, Algorithms

† Languages: Python, R, Javascript, Java, Scala

† Libraries: Pandas, Numpy, Scipy, Sklearn, Tensorflow, Keras, PyTorch, NLTK, Gensim, Seaborn, Bokeh, Angular, D3

† Systems/Tools: Linux, Hadoop, Spark, Map-Reduce, Hive, Pig, MongoDB, SQL, Shell scripts, Git, Jenkins, Docker