

# Xiaoxiao (Monica) Yan

917-900-2474 | [monica.yan@columbia.edu](mailto:monica.yan@columbia.edu) | New York, NY, 10025 | <https://github.com/monicayan>

## EDUCATION

**Columbia University**, Fu Foundation School of Engineering and Applied Science

*New York, NY*

M. S. in Data Science

Expected May 2019

*Coursework: Algorithms, Machine Learning, Probability Theory, Statistical Modeling, Computer Systems, Deep learning, Reinforcement Learning*

**Southeast University** (Honor Student of the Year)

*Nanjing, China*

B.S in Materials Science and Engineering (Major GPA: 3.83/4.0)

June 2017

## SKILLS

Machine Learning, Deep Learning, Reinforcement Learning, Algorithms, Data Analysis, Database Management; Python, R, SQL, TensorFlow, Keras, Bash, C++, D3; AWS, Hadoop, Spark, MongoDB, Tableau, LaTeX; Linux, Unix.

## PROJECTS

### Entity Resolution as Service (Capstone Project)

- Set up entire application pipeline for entity resolution problem, including data transfer (Kafka), data storage (MongoDB), domain-independent blocking strategy, and matching algorithms (word embedding and random forest regression).

### Deep Reinforcement Learning Assisted Trading (<https://github.com/monicayan/Deep-Reinforcement-Trading>)

- Extracting financial features from historical data; using deep-Q-network and policy gradient method to guide trading policy;

### MTA Subway System Optimization (<https://github.com/monicayan/MTA-Analysis>)

- Scraping raw data from website; after data cleaning, commit exploration and visualization on passenger flow on each gate of NYC subway system; give possible optimization based on the results.

### Text to Images

- Trying to use stackedGANs structure to generate high-resolution pictures according to user inputs (based on TensorFlow);

### Amazon Online Products Recommendation Analysis

- Analyzed the likelihood of the customers' purchasing different products together, summary stats; currently try to use the data build a product-based recommendation system using logistic regression model.

### Stan Development Team

- Building interface between CmdStan and Python, and PyCmdStan and ArviZ;

### Selected Assignments

- Built word embedding and RNN structures to: 1) classification sentences from different books; 2) generate color and RGB values according to input color names; 3) generate color name and RGB values according to input color. (based on TensorFlow)
- Built a user audio recommendation system using PySpark and MLlib;
- Set up AutoML-style image classifier with CNN structure to classify different flower varieties in the pictures (based on TensorFlow);
- Built various regression models to predict wine quality according to 28 features; compare the differences between each model;
- Analyzed the influence of insurance level on times of medical consultations using generalized linear model (turncPoisson, Exponential).

## EXPERIENCE

### U.S. Science Support Program (a NSF program)

IT Assistant

05/18 - present

*New York, US*

- Wrote Python scraping scripts to get raw data from website, and cleaned the data;
- Built an easy-maintained relational database, set up both back-end and front-end database platform (based on MySQL and Javascript);
- Managing the database and its server as daily routine, developing web-front end as a search engine linked to the database.

### Southeast University

Software Developer

06/14 – 09/14

*Nanjing, China*

- Developed a software for ticketing system based on C++, with selling, searching and returning functions;
- Conducted functional testing, usability testing and performance testing to the software.