

# Wenyi (Wesley) Tao

(201) 201-9455 \* [wt2271@columbia.edu](mailto:wt2271@columbia.edu) \* <https://wesleytao.github.io/>

## EDUCATION

**Columbia University** *M.A. in Statistics, New York* GPA: 4.00/4.00 Expected Dec 2018  
Courses: Reinforcement Learning, Machine Learning, Analysis of Algorithm, Bayesian Statistics, Advanced Data Analysis

**Fudan University** *B.A. in Economics, Shanghai* GPA: 3.53/4.00 (Top 1 final year) Sept.2012 - July.2017  
Courses: Applied Statistical Tools, Econometrics, Advanced Mathematics, Data Structures, Intro to Database

## SKILLS

Model: Lasso Ridge, SVM, AdaBoost, Bagging, K-means, Logistic, Hierarchy Bayesian, Latent Dirichlet Allocation  
Tools: MySQL, Python, R, Git, OpenCV, Scikit-learn, Spacy, Pytorch, PySpark, Tableau

## MACHINE LEARNING EXPERIENCE AND INDEPENDENT PROJECT

- 09.2018-present **Machine Learning Engineer Intern**, Pactera OneConnect AI Lab New York
- Collaborated with other engineers to build a data-driven chatbot powered by machine learning models
  - Implemented LSTM, SVM, Naïve Bayes models to upgrade hard-coded dialogue management and improved the product by making the chatbot memorize the context of dialogue history
  - Replaced the low-level DBAPI with ORM mapper and made the code base easier to maintain and managed
- 06.2018-09.2018 **Machine Learning Engineer Intern**, Adatos A.I., New York
- Decreased the mean absolute error (in around 300 trees) from 8 to 3 by replacing existing Gaussian blob detection with deep-learning powered tree counting model and speeded up the tree counting process from 1.83s per image to 0.1s per image by designing and fine-tuning an Unet Model
  - Created well-organized and well-documented code for the end-to-end automatic tree counting pipeline
  - Solved the poor performance on the boundary problem of semantic segmentation models (FCN and UNet)
- 04.2018 **Independent Researcher**, Collaborative Filtering with EM Clustering New York
- Built from scratches a recommender system with Bayesian Clustering Algorithm without using any ML framework
  - Tried multiple parametrized distribution within the conjugate family, evaluated the different models based on speed and performance
  - Use cluster structure robustness, perplexity to tune the hyperparameter and use utility score to evaluate the performance
- 05.2018-09.2018 **Research Assistant**, Harvard T. H. Chan School of Public Health New York
- Assisted a Ph.D. candidate on a public health project by scrapping and transforming data from twitter and google trend
  - Discovered that people of Marseille Longchamp have symptoms of asthma 2 hours after local PM2.5 hit 35  $\mu\text{g}/\text{m}^3$
- 08.2017 **Research Assistant**, Fudan Institute of Data Science Shanghai
- Found people under the non-flat-rate policy use 12 percent more electricity after 21:00 compared to the control group
  - Quantified the price elasticity between the two group users, proposed a new pricing policy for the Shanghai Electric Power Company
- 04.2018 **Group Project Leader** Tweets Sentiment Analysis on Sino-US Trade War New York
- Implemented Latent Dirichlet Allocation to perform topic modeling for millions of tweets
  - Use interactive D3 to visualize the retweet relation networks and found those influential opinion leaders on this topic
  - Visualize majority opinion's shift across a series of events after performing sentiment analysis on all the comments

## COMPETITION EXPERIENCE

- 04.2018 **1 out of 12 teams:** Indeed Data Mining for Target Marketing
- Discovered the dominant features like license-required that can boost around 5%-12% of average clicks per day
  - Found job posted before November statistically have more clicks than job posted after November excluded other effects
- 01.2018 **Top 5%:** Alibaba Cloud Algorithm Competition- Helping Balloons Navigate the Weather
- Planned safe and fastest flight routes for ten unmanned balloons to deliver packages to their destinations
  - Creatively implemented the A-Star algorithm in 3-Dimension space to solve the obstacle changing problem

## INTEREST

- Swimming (2000m nonstop with medley)