Weiwei Qi

Tel: 646-683-6801 | E-mail: wq2151@cumc.columbia.edu | Website: https://wq1701.github.io/

EDUCATION

Columbia University in the City of New York **New York, US**

Master of Science in Biostatistics, Theory and Methods track | GPA 3.79/4 Expected 2021.5

- **Course**: Data Science I & II, Biostatistical Methods I & II, Probability, Statistical Inference, Epidemiology, Longitudinal Analysis, Deep Learning in BMEN, ML w/ Probabilistic Prog, NLP for social science, Design of Medical Experiments

China Pharmaceutical University **Nanjing, China**

Bachelor of Business Management, Pharmaceutical track | GPA 3.34/4 2015.9–2019.6

- **Course**: Management, Marketing, Linear Algebra, Micro & Macro Economics, Operational Research, Securities Investments, Pharmaceutics, Pharmacoeconomics, Statistics Principles, Mathematical Statistics

SKILLS

- Programming: Python, PyTorch, Pyro, Jupyter Lab, R, SQL, SAS, MATLAB, Excel, GCP, Git[wq1701], Shell scripting, Linux
- Field: Data Analysis, Data Visualization, Statistical Models & Machine Learning, Deep Learning & CNN

EXPERIENCE

Research Foundation for Mental Hygiene (NYSPI/Columbia University) **New York, US**

Research Technician/Statistical Analyst Since 2020.6

- Write, debug, and submit Shell/R/MATLAB scripts to HPC cluster (Linux environment) for neuroimaging processing and large-scale computation under Linux environment; report and present results weekly to supervisor for next-step analysis;
- Use R to generate over 1000 shell & MATLAB scripts for computation; Developed functions based on **GenLouvain dynamic community detection** algorithm to compute neural flexibility in 268 nodes of brain and clean the data in R for statistical analysis;
- Run moderation analysis to test hypothesis of interaction between age and regional neural flexibility; Validated that the neural flexibility of some brain network has significant moderator effect to memory and reasoning; Concluded that age affects specific cognitive performances significantly through correlation analysis.
- Use R and Python to visualize and report significant results as informative figures and tables for publication.

US-China Health Summit, Covid-19 News Report Group **New York, US**

R developer 2020.3- 2020.5

- Developed R functions to download / update / merge data; Rewrite the R script with Tidyverse pipeline to optimize workflow.
- Provided visualization support for News Group; Developed functions to translating inputs keys for reader-based localization;
- Responsible for program maintenance and functional update.

CUG Golden Shield Environmental Technology **Wuhan, China**

Assistant to Chief Executive Officer 2019.2-2019.6

- Assisted the CEO with daily administrative duties; Translate literatures for leadership decision-making and competence analysis;
- **Excel** and **Word** documents processing, such as **PivotTable** and **Formula Functions** to provide information support;
- Performed **desk research** regarding competitors, advanced soil remediation technology methods and drafted reports to executives.

PROJECT

Recommendation System based on Probabilistic Matrix Factorization 2020.10-2020.12

- Defined Gaussian and Poisson distribution as prior for item rating based on descriptive statistics, and transform original data into matrix form. Use the product of movie and user latent feature as the prediction for calculating loss and optimize latent variable distribution. Use **Pyro** as the main package for coding the model and optimize KL divergence and ELBO for variational inference;
- Validate the optimal dimension for latent features is 30 and improved model performance with MAE=0.9; One of the few Pyro-based large-scale Matrix-factorization examples on GitHub.

Natural Language Processing on Medical Transcript

- Performed data cleaning and text-mining on 4000 medical transcripts with regex/stemming/vectorization/TF-IDF. Used PCA for dimension-reduction on transcript data to improve model training speed;
- Trained multiple supervised machine learning models to classify the transcript by medical specialty and keywords; Used Resampling to improve certain model performance due to unbalanced labels. The best model has 0.6 accuracy and 0.64 F1-score. Adopted Latent Dirichlet Allocation as unsupervised model to extract word bags and keywords.

Data Augmentation for Brain-Tumor Segmentation, Columbia Engineering 2020.3-2020.5

- Write python code to build, train, validate, and test the Neural Network (U-Net) for brain-tumor segmentation using **PyTorch** and **TorchIO** on Google Cloud Platform with total dataset size of 335; Achieved a Dice Score of 0.81 on the test dataset;
- Implement traditional data augmentation methods such as flip, random bias field, and random noise to increase the dataset size to avoid the network from overfitting / memorizing the training set. Finally, the model dice score improved 6.1% with augmented data;
- Adopt state-of-the-art Generative Adversarial Network (GAN) architecture as a new data augmentation method and apply it into generating new high-quality data to further grow the dataset, and to generalize the model performance in tumor segmentation.

Data Analysis and Visualization for Drug Usage in Connecticut 2019.10-2019.12

- Discovered the Number 1 drug of causing death (Fentanyl) and the association between age and illicit drug use by visualizing and analyzing the dataset. Built a project website with the team for final presentation;
- Matched the original county/city column with new geo-coordinates data downloaded from Google to expand the original dataset for visualizing the distribution of drug illicit use, age, race, location on state map. Built and deployed the interactive map to the project website with **ggplot2, Dashboard, Leaflet** and **R Shiny app**;
- Built and validated regression models to predict the death rate associated with predictors such as age, drug types, race, and location.