齐维为

Tel: 15951070153 | E-mail: wq2151@cumc.columbia.edu | Website: https://wq1701.github.io/

教育背景

哥伦比亚大学 **美国纽约**

生物统计学，理学硕士-理论与方法方向 | GPA 3.79/4 2019.9-2021.5

- **相关课程：**数据科学, 生物统计方法, 概率论, 统计推断, 流行病学, 纵向数据分析, 深度学习, 贝叶斯机器学习, 自然语言处理, 临床实验设计

中国药科大学 **中国南京**

工商管理，管理学学士-医药方向 | GPA 3.34/4 2015.9–2019.6

- **相关课程：**管理学, 市场营销, 线性代数, 经济学, 运筹学, 投资学, 药学, 分析化学，药物经济学, 统计学原理, 数理统计

技能

- 编程语言/软件: Python, R, SQL, Shell Script, MATLAB, linux, SAS, MS Office, Tableau, GCP, Git[wq1701], LaTex
- 专业知识: 回归分析, 可视化, 机器学习, 神经网络, 假设检验
- 综合素质: 英语（托福 108）, 逻辑与批判性思维, 团队合作与沟通, 数据导向思维

工作经验

纽约长老会医院/哥伦比亚大学 （Prof. Seonjoo Lee, 哥伦比亚大学生物统计系副教授） **美国纽约**

研究技术员/统计分析师 2020.6-2021.4

- 编写并调试 Shell/MATLAB 脚本并提交至高性能计算集群(HPC, linux 环境)进行神经图像的并行处理以及大规模计算；用 R 实现自动化批生成超过 1000 个实验对象的运算脚本；
- 在 MATLAB 中开发基于 GenLouvain 广义动态社区检测算法的程序用于计算大脑中 268 个节点神经灵活性，并在 R 中进行数据清洗和统计分析；对年龄和区域性神经流动性进行调节分析并假设检验其交互作用，验证部分网络的神经灵活性对记忆与推理能力存在显著调节作用。通过相关性分析得出年龄对成年人的部分认知指标表现存在显著影响；
- 将静息态与作业性核磁扫描数据按脑功能分组计算区域性神经灵活性，并分析拒绝关于成年人神经活动性在静息态下存在显著灵活性的假设，进一步提出其大脑活动表现在静息态下偏向随机性流动而非系统分配的假设。

中美健康峰会, 新冠报道组 **美国纽约**

R 程序开发组 2020.3-2020.5

- 开发 R 程序实现后台数据自动整合与输入翻译可控自动化；以 tidyverse 为主优化 R 脚本的工作流；
- 为新冠早报的编写提供实时可视化支持以及程序除错和优化。

中地金盾环境科技有限公司 **中国武汉**

CEO 助理 2019.2-2019.6

- 协助上级完成日常管理任务；按需求翻译文献及资料用于后续决策分析；进行案头研究并收集数据，对市场及潜在竞争对手进行初步分析及撰写报告；
- MS office 文档处理，用 PivotTable 和 excel 函数等工具进行信息整合和展示。

项目经验

基于概率矩阵分解的电影推荐系统 （课程：Machine Learning w/ Probabilistic Programming） 2020.10-2020.12

- 为超过两万部电影评分设置正态与泊松先验分布，将原始评分列表数据转化为矩阵，并通过电影与用户潜变量的乘积计算损失与优化潜变量分布。以 pyro 为主要库编写概率模型并优化 KL 散度与变分推断的证据下界(ELBO)；
- 验证潜变量的最佳参数为 30 且模型在测试集的表现接近业界前沿标准（MAE=0.9）；为 GitHub 上为数不多的 pyro 矩阵分解实例。

电子病历的自然语言处理

- 对 4000 名个体的电子病历进行数据清理和文档挖掘，例如正则化表达、提取词根、向量化及计算词频-逆文档频率；通过主成分分析(PCA)对冗杂的电子病历数据进行降维以提高后续模型学习速度。
- 通过训练多种监督式机器学习模型来对电子病历进行分类，对不平衡的标签进行重采样以提高特定模型的表现-最好模型准确度达到 0.6，F1 分数为 0.64；用隐含狄利克雷分布进行无监督机器学习提取电子病历关键词与主题。

脑肿瘤分割的神经网络数据增强-哥伦比亚大学工程学院（Prof. Jia Guo） 2020.3-2020.5

- 利用 PyTorch 以及 TorchIO 搭建用于大脑肿瘤分割的深度 U 型神经网络，依靠 Google 云平台进行模型训练加速；
- 采用多种传统数据增强方法来实时提高神经网络的分割表现，例如模糊，噪点，对称旋转。确认拉伸等增强方法将破坏神经网络的学习效果。其余方法将模型表现（Dice score）提高 6.1%，最终 Dice score 为 0.81；
- 尝试前沿的数据增强方法，例如 GAN（生成对抗网络）生成高质量脑图，来提高分割模型的泛化能力。

康州药物使用分析与可视化 2019.10-2019.12

- 通过描述性分析及可视化得出误用致死率最高的药物(Fentanyl)以及年龄和误用药物之间的关联；用 R markdown 与 GitHub 搭建项目网站用于展示与汇报；
- 将原数据中的地名与谷歌的地理坐标信息进行匹配并清洗、拓展数据，对误用药物病患的人数、年龄、种族以及误用药物的详细地理坐标进行可交互与可视化处理；编写 ggplot/Rshiny/Leaflet 代码搭建可交互地图及部署至网站；
- 构建线性回归模型得出致死率与病患年龄/药物/种族以及地理位置等变量的关系

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/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 regional neural flexibility in 268 nodes in 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.

US-China Health Summit, Covid-19 News Report Group **New York, US**
R developer 2020.3- 2020.5

- Develop R functions to clean, merge and translate input such as row and columns names to produce Chinese version chart output;
- Match countries with different colors to create tidy and consistent plots; Rewrite the R script with pipeline to optimize workflow.

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 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.