# Kai Zhang

kaz4006@med.cornell.edu/ (+1) 6465772755

## EDUCATION BACKGROUND

**Cornell University** 9/2019–12/2020

M.S. in Biostatistics and Data Science | Overall GPA: 3.98 / 4.0

**Selected Courses:** Statistical Learning, Generalized Linear Model, Survival Analysis, Pharmaceutical Statistics, Design and Analysis of Biomedical Studies, Hierarchical Modeling and Longitudinal Data Analysis, Categorical Data Analysis, Causal Inference

**The Chinese University of Hong Kong (Shenzhen)** 9/2015–7/2019

B.S. in Statistics with the 1st Honor Degree | Overall GPA: 3.40 / 4.0 **(Ranked 7/71)** | Major GPA: 3.66/4.0 (**Ranked 3/71**)

**Selected Courses:** Machine Learning, Data Mining, Data Management, Statistical Software, Regression Analysis, Time Series, Survival Analysis, Stochastic Process, Optimization, Statistical Inference, Nonparametric Statistics, Survey Sampling, Statistical Topic in Epidemiology, Statistical Models in Financial Market, Financial Data Analysis

**Online Courses:** Applied Machine Learning in Python (Coursera), Extreme Gradient Boosting with XGBoost (DataCamp), SQL Fundamentals (DataCamp), Neural Network and Deep Learning (Coursera), Deep Learning with Keras (DataCamp)

## RESEARCH EXPERIENCES

***Single-cell RNA Data Mining on Mouse Retina (Supervisor: Prof. Kathy Zhou)***

**Weill Cornell Medicine, New York, USA** 1/2020–

- Pre-processed raw droplet-based single-cell data following *Scater* workflow, including distinguish cells from empty droplets using Dirichlet-multinomial model-based method, performing quality control by outlier detection, removing non-biological difference across cells by deconvolution normalization, discovering highly variable genes/features by Poisson distribution-based method, and reducing dimensionality using Principle Component Analysis (PCA).
- Performed and compared multiple clustering algorithms on cells, including K-means, hierarchical clustering, DBSCAN, SNN graph-based clustering, combined K-means and graph-based clustering, and Resampling-based Sequential Ensemble Clustering (RSEC) algorithm.
- Detected marker genes for clusters using pairwise t-test with Benjamini-Hochberg (BH) procedure for multiple comparisons.
- Working on cell type annotation using reference-based analysis and classification methods.

***Causal Mediation Analysis of Factors of Frailty in Older Adults (Supervisor: Arindam RoyChoudhury)***

**Weill Cornell Medicine, New York, USA** 5/2020–

- Performed and compared the traditional Baron-Kenny’s four-steps procedure with Sobel test and simulation-based causal inference method, to detect the mediation effect of body composition on relationship between physical performance and strength in old people.

***Campus-wide Trajectory Network Analysis (Supervisor: Prof. Jianmin Jia, Dr. Jianjun Zhou)***

**Shenzhen Research Institute of Big Data, Shenzhen, China** 1/2019-6/2019

- Designed a processing pipeline of extracting CUHKSZ students’ moving trajectory based on the raw wifi connection data from 1710 students’ devices including time and location details.
- Modeled and visualized student’s behavior as a novel network of school facilities to perform social network analysis using NetworkX toolkit, and extracted SNA statistics (e.g. degree centrality, eigenvector centrality, entropy) during various time periods, especially for exam weeks.
- Run the Lasso regression and XGBOOST on network measures against GPA, to find the dominant factors of individuals’ academic performance.

***Prediction of Poor Academic Performance Using Behavior Data (Supervisor: Dr. Jianjun Zhou)***

**Shenzhen Research Institute of Big Data, Shenzhen, China** 9/2018-1/2019

- Extracted behavioral features of 1710 CUHKSZ students from raw wifi connection data, including estimation of students’ monthly spending time in various facilities (e.g. library, dorm, teaching building, sports hall, etc.), and detection of potential friendship based on spatiotemporal overlaps.
- Predicted students’ potential decline in academic performance using Random Forest and XGBoost, based on behavioral data.

## WORK EXPERIENCE

**Credit Rating Intern, Changjiang Securities, Wuhan, China** 7/2018–8/2018

- Preprocessed economic data of listed manufacturing companies during 2014-2017, including missing data imputation, discretization of continuous variables based on conditional inference, variables selection using RF algorithm, dimension reduction using PCA method.
- Built risk rating card using logistic regression model on WOE transformed data.

## AWARD

| | |
|---|---|
| Dean’s List | 2017-2018 |
| Master's List Scholarship of Shaw College (Top 10% Academic Scholarship) | 2018-2019 |

## LANGUAGES AND SKILLS

**GRE General:** V: 154 / Q: 170 / AW: 3.5 / Total: 324

**Programming Languages / Software:** R (tidyverse, ggplot2, R shiny), Python (pandas, scikit-learn, Keras), SQL