Wenjia Wang

5231 5th Ave, Pittsburgh, PA 15232 | +1 (530) 5649145 | wew89@pitt.edu | https://wenjiaking.github.io/

EDUCATION BACKGROUND

University of Pittsburgh **08/2020-present**

Ph.D. in Biostatistics (Advisor: George C. Tseng) ***Overall GPA:* 3.98/4**

University of California, Davis **09/2018-12/2019**

M.S. in Statistics ***Overall GPA:* 3.91/4**

East China Normal University **09/2013-07/2017**

B.S. in Mathematics and Applied Math ***Major GPA*: 3.85/4**

Statistics Courses: Longitudinal Data Analysis, Baysian Analysis, Stochastic Processes (Markov Chain, Poisson Process, etc.), Computational Statistics (optimization, EM algorithm, MCMC, etc.), Omics Data Analysis, Survival Analysis, Machine Learning (Carnegie-Mellon University), Advanced Deep Learning (CMU), etc.

Advanced Math Courses: Abstract Algebra, Real Analysis, Functional Analysis, Numerical Analysis, ODE, Topology, etc.

Public Health Courses: Epidemiology, Essential of Public Health, Molecular Basis of Human Inherited Deseases.

HONORS & SCHOLARSHIPS

Excellent Graduate of East China Normal University:	07/2017
First-class & Second-class National Scholarship: for top 3% and 10% students respectively.	2014 & 2015
University-level First Prize of Daxia Fund Project	2015

RESEARCH EXPERIENCE

Fast P-value Calculation for Higher Criticism (HC) and Berk-Jones (BJ) Tests in Meta Analysis **11/2020-Present**

- Thesis (with manuscripts); Advisor: George C. Tseng, professor in biostatistics department, University of Pittsburgh
- Propose the cross-entropy based importance sampling method with appropriate importance density to efficiently compute the null distribution for HC and BJ. Modify the existing analytic methods to improve the numerical stability and flexibility
- Construct the quantile-probability curves for HC and BJ of various domains and dimensions for fast vectorizing computation

Develop the pipeline for long-read sequencing (generated by PacBio, Oxford Nanopore, LoopSeq, etc.) **03/2021-Present**

- Collaborator: Silvia Liu, assistant professor in pathology department, University of Pittsburgh
- The pipeline with raw FASTQ input includes quality control, alignment (combining multiple aligners e.g. STARLong, Minimap2, etc.), fusion calling, filtering, isoform identification, ORF prediction and functional analysis.
- Conduct simulations by PBSIM, analyze anchor length, sensitivity and precificity, finally apply it our ten real cancer data sets

Research of Neurons Extraction from in-vivo Calcium Imaging Data (Neuron Science) **08/2019-02/2020**

- Independent study; Advisor: Shizhe Chen, assistant professor in statistics department, University of California, Davis
- Explore the structures of various imaging data, theoretically analyze the efficiency and drawbacks of unsupervised basis learning approaches in extracting neurons from different calcium imaging data, and write summaries and a manuscript review
- Improve methods and do simulations to remove neuropil/out-of-focus contamination and extract subcellular compartments

PROJECTS

Explore in utero intestinal metabolome and where the in utero microbiome come from **12/2020-Present**

- Collaborator: Liza Konnikova, assistant professor in Yale University School of Medicine.
- Preprocess the data of metabolites expression in various tissues, conduct DE analysis, pathway enrichment analysis, ANOVA
- Assess gene expression of bile acid and SCFA receptors and carrier proteins based on 11 single cell data sets

Patient Self-evaluation System for Medicine Recommendation (Python, ML) **05/2019-06/2019**

- Apply non-linear binary classifiers (SVM, random forest, Adaboost, neural network) to a realistic data of patients situations from a pharmacy company, evaluate models based on F1 score, ROC, PR curves, and compare space and time complexity

Analysis of House Rentals Posts from Craigslist Website (R, data mining) **11/2018-12/2018**

- Scrape 45847 posts about house rentals from Craigslist in 4 areas, clean the messy data and extract features from texts
- Regress to rental prices based on factor analysis, visualize results and provide the best choice based on custom's demand

Numerical Algorithms of Solving Linear Least Squares Problem (MATLAB, optimization algorithms) **12/2016-04/2017**

- Research in algorithms of tensor decomposition, matrix factorization involving SVD and Cholesky decomposition
- Compare algorithms SOR, Krylov subspace method (CG) and preconditioning methods under varied ill conditions

PROFESSIONAL SKILLS

Programming Language: proficient in R (developing R packages and R Shiny), MATLAB, Python (scikit-learn, TensorFlow, PyTorch), SAS (SAS Certified Specialist: Base Programming Using SAS 9.4), STATA, Bash, SQL

Machine Learning: CNN, RNN, Hidden Markov Models, Bayesian Networks, Reinforcement Learning, Recommender Systems

Computer Vision: Image processing, Feature extraction, Segmentation and Classification