Ruoxi Yang

(732) 789-8985 | yruoxi0206@gmail.com

Education

New York University   09/2023 – Present
Master of Science in Biostatistics

Rutgers University   09/2019 – 05/2023
Bachelor of Arts in Statistics with Minor in Psychology

Publications

Manuscripts

- Kascelan, J., **Yang, R.**, Shasha D.* *EpiInfer: Forecasting Infection Rates In Epidemics Using Meeting Distributions Across Locales*. Submitted to *Algorithms*, 2024.

Conference Abstracts

- **Yang, R.**, Xiaotao, Z.*, et al. *Oral Microbiome Diversity And Its Association With Liver Fibrosis: Insights From The NHANES 2009-2012 Study*. Digestive Disease Week (DDW) 2025, May 3–6, San Diego, CA.
- Maryfe C., **Yang, R.**, Xiaotao, Z.*, et al. *Bean Consumption Is Associated With Fibrosis Improvement In Liver Stiffness Among Hispanic Patients With MASLD: Pilot Findings From The Mount Sinai MASLD/MASH Registry*. Digestive Disease Week (DDW) 2025, May 3–6, San Diego, CA.

Research Experience

Research Data Analyst   10/2024 – Present
NYU Grossman School of Medicine
Advisor: Dr. Yongzhao Shao

- Conducted high-dimensional data analysis using packages survival, ncvreg, ggm, and data.table to identify significant biomarkers associated with lung cancer survival outcomes.
- Applied advanced statistical techniques, including feature selection methods like sure independence screening (SIS) and minimax concave penalty (MCP), to handle large-scale genomic datasets.
- Utilized Kaplan-Meier survival curves to visualize survival probabilities and uncover patterns across patient subgroups, while generating time-dependent ROC curves to assess the predictive accuracy.
- Expanding this study by incorporating next-generation sequencing datasets, including RNA-Seq gene expression data, to complete the mediation analysis pathway.

Research Data Analyst   09/2024 – Present
Microbiome Epidemiology Lab, Icahn School of Medicine at Mount Sinai
Advisor: Dr. Xiaotao Zhang

- Conducted research on the association between the oral microbiome and liver fibrosis using real medical data collected from Mount Sinai Hospital and NHANES datasets.
- Analyzed microbial diversity metrics, including alpha diversity (using Shannon and Simpson indices) and beta diversity (using Bray-Curtis and Weighted UniFrac distances), to examine their associations with fibrosis.
- Performed relative abundance analysis at family, genus, and species taxa levels, identifying key microbial features linked to liver fibrosis.
- Visualized alpha diversity using boxplots, beta diversity through Principal Coordinates Analysis (PCoA), and identified significant taxa using volcano plots and heatmaps.

Research Assistant   03/2024 – Present
Advisor: Dr. Dennis Shasha

- Collected and analyzed detailed contact tracing data from a small community, creating distributions of daily meetings per person across 21 countries.
- Developed and implemented an SEIR model to estimate key epidemic parameters p_1 (probability of exposure) and p_2 (probability of infection), integrating incubation periods and asymptomatic transmission dynamics.
- Simulated epidemic progression using statistical distributions of daily contact rates to predict newly exposed and infected cases, accounting for symptomatic and asymptomatic carriers.
- Improved predictive performance of SEIR models by leveraging multi-locale inference and probabilistic approaches to better capture real-world epidemic dynamics.

Lab Assistant   03/2024 – Present
The Coruzzi Lab, NYU Center for Genomics & Systems Biology
Advisor: Dr. Tim Jeffers
- Conducted research in plant biology with a focus on gene function and regulation in *Arabidopsis thaliana*, tomatoes, and other crops.
- Screened plasmid samples to ensure successful gene incorporation and utilized techniques like agarose gel electrophoresis for preliminary verification.
- Investigated gene expression patterns and functional impacts on plant development through various molecular biology techniques.

Projects

Natural Language Processing-Based Sentiment Analysis of IMDb Movie Reviews   06/2022 – 09/2022
- Utilized torchtext in Python to preprocess the IMDb dataset by counting word frequency, uniform word case, and removing special symbols.
- Developed proficiency in building deep learning models such as RNNs, CNNs, LSTMs, and Transformers using the torch package.
- Led the group in constructing an RNN model for sentiment analysis on IMDb movie reviews, gradually refining model parameters to improve the performance.

Deep Learning-Based Classification of Pulsar Stars   09/2021 – 12/2021
- Used R to conduct data preprocessing on the HTRU2 dataset by transforming all null values and object data types into numerical values.
- Developed a deep learning model using a simple neural network with three dense layers and implemented ReLU activation for hidden layers and sigmoid activation for the output layer to optimize performance in binary classification.
- Evaluated model performance by plotting accuracy and loss curves for both training and validation data and generated a confusion matrix to assess prediction accuracy.

Data Analysis with SAS Programming   01/2020 – 03/2020
- Employed SAS to import car data comprising 60 observations and 5 variables representing cars sold in the North American market.
- Fitted the five variables into a linear regression model and utilized the "selection" option to check coefficient estimates.
- Identified the optimal regression model and performed model diagnostics and validation checks.

Work Experience

AI/Data Analysis Intern   08/2024 – 10/2024
Circus AI, New York, USA
- Conducted comprehensive research and data collection on apps and e-commerce platforms utilizing AI for automatic content generation, analyzing business strategies to extract actionable insights and assess industry trends.
- Contributed to the development of the Circus AI app by enhancing image and content generation functionalities to expand the user base.

Data Analyst Intern   01/2023 – 06/2023
Sino City USA LLC, Bridgewater, USA
- Analyzed business and contact information from over 200 nail/beauty salons in New Jersey using Microsoft Excel and search engines.
- Constructed the API interface key for Google Map and Yelp software to accurately locate over 60% of small businesses in New Jersey.
- Used Gmail plugins to monitor user engagement post-email receipt and identified target customers among a wide range of merchants.

Skills

Programming Languages: Python, R, SAS, REDCap, Stata, NetLogo, LaTeX
Models & Techniques: predictive modeling, time series analysis, neural networks, natural language processing (NLP), regression, optimization, compartmental epidemic models (SIR, SIRD, SEIR)
Languages: Fluent in English and Mandarin Chinese