

# Wangqian Ju

LinkedIn: [linkedin.com/in/wangqian-will-ju/](https://www.linkedin.com/in/wangqian-will-ju/)

Github: [github.com/willju-wangqian](https://github.com/willju-wangqian)

Email: [wju@iastate.edu](mailto:wju@iastate.edu)

Mobile: +1(515)708-2829

## EDUCATION

---

### Iowa State University

Ames, IA

- *Ph.D. Candidate in Statistics; Advised by Dr. Heike Hofmann*

Aug. 2019 - Expected May. 2024

3.94 GPA; Kempthorne Award in Statistics;

- *Master of Science in Statistics*

Aug. 2019 - June. 2021

### University of California - Davis

Davis, CA

- *Bachelor of Science in Statistics; Minor in Computer Science;*

Aug. 2015 - June. 2019

3.92 GPA; UC Davis Statistics Outstanding Performance Citation; Graduation with Honors

## SKILLS

---

- **Statistical Analysis:** Generalized Linear Model, ANOVA, Bayesian, Time Series, Statistical Inference
- **Coding:** R, Shiny App, Python, C, Matlab, SQL, Bash, Git, JavaScript
- **Frameworks:** Scikit, TensorFlow, Keras

## RESEARCH EXPERIENCE

---

### Open-source Implementation of the CMPS Algorithm - R and C

May. 2020 - Present

*Center for Statistics and Applications in Forensic Evidence - Research Assistant*

- Implemented the Congruent Matching Profile Segment (CMPS) algorithm conceptually described by the National Institute of Standards and Technology for objective comparison of striated tool marks in **R**, and used the similarity score computed by the CMPS algorithm for bullet land engravings comparison
- Modified the basic CMPS algorithm to grant users more flexibility in parameters selection
- Processed 420 (3.4 GB) 3D topographical images of bullet land engravings stored in x3p (XML 3-D Surface Profile) format into bullet signatures, numeric sequences that describe striations in bullet land engravings
- Applied the implemented CMPS algorithm to the processed bullet signature data, achieved 100% classification rate, and obtained reproducible results
- Built an **R** package `cmprR` based on the implementation of the CMPS algorithm with **C** source code that improves the overall run-time efficiency by 200 times and visualization tools for the output of the algorithm
- Defined new metrics for summarizing CMPS scores from land-by-land level to bullet-by-bullet level and measuring the effects of various sets of parameters
- Built an automatic pipeline for parameter selection based on the new metrics

### Data Processing and Visualization tool - R and Shinyapp

Jan. 2021 - Present

*Center for Statistics and Applications in Forensic Evidence - Research Assistant*

- Built an **R** Shiny app named `bulletInspectR` that allows users to upload x3p files, perform data processing, identify and mark damaged data or outliers, visualize possible results, and download results in csv or rds format
- Implemented a pipeline that performs quick and automatic data preparation
- Modularized the Shiny app so that users can write and insert modules freely, implemented modules for x3p data processing, and processed 600+ data using the Shiny app

### Functional Data Analysis with New York Taxi Data - R

Jan. 2018 - June. 2019

*Statistics Research Training Group; Advised by Dr. Hans-Georg Mueller*

- Cleaned and processed (25+ GB) New York taxi (yellow) trip data for year 2016 and 2017 in **R**
- Introduced new covariates by collecting and processing New York temperature and weather data (humidity, precipitation, etc) and applied the functional principal component analysis techniques to the weather data
- Fitted the global Frechet regression model with respect to the Wasserstein distance to predict temporal pick-up/drop-off density of the Manhattan taxi data

### Performance testing for BugSwarm with Cobertura - Python

Jan. 2018 - June. 2018

*Davis Excellent/Eclectic/Extreme Computational Analytics Lab*

- Used Cobertura to calculate code coverage and other metrics of tests of selected Java Projects, and built a pipeline that collects data from multiple Cobertura reports
- Modified a pipeline that generates code coverage reports for hundreds of projects in a multi-threaded way with Cobertura and BugSwarm

## DATA ANALYSIS PROJECTS

---

### Green Beans Daily Sales Prediction - R

Jan. 2021 - Apr. 2021

- Conducted exploratory data analysis on data of individual stores and fitted primary models to guide further analysis
- Fitted a statistical model with a two-stage prior that allows individual stores to investigate the relationship between sales and prices through the posterior estimations of the model parameters
- Used **JAGS** (Just Another Gibbs Sampler) to obtain credible intervals, posterior distributions, posterior predictive distributions, and other estimations of the model parameters

### New York City Real Estate Price Prediction - R and Python

Jan. 2020 - Apr. 2020

- Investigated and processed 50,000+ real estate data in New York City and performed exploratory data analysis
- Built a random forest model to predict the real estate price using model prediction results (Ordinary Least Squares, K-Nearest Neighbors, Principal Component Regression, Elastic Net, etc) as features and achieved a loss score of 0.35
- Built an artificial neural network model in **Python** with Keras and compared its result with that of the random forest model

### Dog Breeds Scanner - Python

Sep. 2018 - Dec. 2018

- Utilized transfer learning techniques to built and trained a convolutional neural network model
- Classifies 10 most popular dog breeds in the US with over 95% accuracy
- Deployed the model online as an iOS mobile application

### Reddit Comments Analysis - Python

Sep. 2018 - Dec. 2018

- Conducted exploratory data analysis on Reddit comments and collected, cleaned, and processed over 800,000 Reddit comments for the time period of 2016 presidential election
- Trained a Word2Vec model, applied K-Means clustering to the obtained word vectors, and used the clusters for embedding sentences and obtaining sentence vectors
- Clustered the Reddit comments, conducted sentimental analysis to identify people's attitudes toward political issues, and visualized results with word cloud

### Network Data Visualization with Facebook Data - Python

Dec. 2016 - May. 2017

- Implemented the Shi and Malik's Normalized Cut algorithm, applied it to Facebook network data, and built visualization tools for the results
- Built a **Python** package that organizes and collects the implemented functions and methods

## PUBLICATIONS

---

- **“An Open-Source Implementation of the CMPS Algorithm for Assessing Similarity of Bullets,”** The R Journal, In Preparation

## TEACHING

---

### Teaching Assistant

Iowa State University

- Fall 2019 - Spring 2020: Introduction to Statistics
- Fall 2019 - Spring 2020: Engineering Statistics

### Academic Assistance and Tutoring Centers

UC Davis

- Spring 2018: Individual Tutor, Fundamentals of Statistical Data Science
- Spring 2018: Individual Tutor, Elementary Statistics
- Fall 2018: Drop-in Tutor, Calculus

## SELECTED COURSEWORK

---

- **Theory:** Theory of Probability and Statistics; Statistical Inference; Regression Analysis; Analysis of Variance
- **Methods:** Statistical Methods; Statistical Data Science; Applied Modern Multivariate Statistical Learning
- **Coding:** Object-Oriented Programming and Data Structures; Algorithms; Machine Learning and Artificial Intelligence