# Mingming Cai

206-790-3693 mingmingcai306@gmail.com Mingming's LinkedIn Mingming's Homepage

## Biography

- Ph.D. candidate in Transportation Economics and human-centered AI, with a concurrent M.S. in Statistics from UW.
- Academic journey includes applications of machine learning and deep learning methods in demand forecasting, statistical modeling of human behavior and preference, and big data mining and spatio-temporal network analysis.
- Proficient in Python (5+ years), R (5+ years), and SQL (3+ years), with dedicated experience in software system design using C++ and C# (1+ years), distributed computing using MapReduce and Spark (2+ years), Git (2+ years), and AWS (1+ year).
- Passionate about state-of-art and responsible AI techniques (Interpretability, Fairness) to promote human-centered AI products.
- A collaborative person with effective communication and contributions to diverse projects within interdisciplinary teams.

## Education

**University of Washington** September 2020 - May 2025 (Expected)
*Ph.D. in Urban Planning and Economics (GPA: 3.9 / 4.0) - Minor: Machine Learning* *Seattle, WA*
- Relevant Coursework: Machine Learning for Big Data (Recommendation System, NLP, Social-Network Graph Analysis), Statistical Modeling & Prediction & Computing (TensorFlow, PyTorch), Time Series Analysis, Statistical Analysis of Social Networks

**University of Washington** September 2023 - May 2025 (Expected)
*M.S. in Statistics (GPA: 3.8 / 4.0) - Minor: Data Science* *Seattle, WA*
- Relevant Coursework: Causal Modeling, Bayesian Statistics, Design and Analysis of Experiments

**Wuhan University** September 2017 - June 2020
*M.S. in Land Economics and Resources Management (GPA: 3.9 / 4.0) - Minor: Transportation Engineering* *Wuhan, China*
- Relevant Coursework: Python Programming, Database System (SQL), Data Visualization (D3.js), Operational Research

**Wuhan University** September 2013 - June 2017
*B.E. in Environmental Science & Land Resources Management (GPA: 3.8 / 4.0) - Minor: Computer Science* *Wuhan, China*
- Relevant Coursework: Advanced Mathematics, Linear Algebra, Probability Theory and Statistics, Database Principles (SQL, Oracle), Programming Language Design (C), Software System Design (C, C++), Microeconomics, Macroeconomics, Data Structures and Algorithms, Systems Programming (C, C++)

## Research Projects

**Equity Implications of US Suburban and Rural On-Demand Mobility Services Using Explainable ML** | *Leading Researcher*
*Funded by Teaching Old Models New Tricks (TOMNET) Transportation Center, U.S. Department of Transportation, ongoing*
- Employed ML/DL models (XGBoost, LightGBM, Bi-RNNs) to uncover the nonlinear effects of sociodemographics, trip attributes, and the built environment characteristics on the usage of on-demand mobility services.
- Used SHapley Additive exPlanation (SHAP) values to reveal the feature importance for predictions.
- Introduced bias-mitigation regularization into the models, which reduced the prediction error gap between income groups while making trade-offs in prediction accuracy.

**Airbnb Recommender System: NLP-Driven Semantic Information Mining for User Preferences** | *Leading Researcher*
*Funded by Paul G. Allen School of Computer Science & Engineering, University of Washington, 2023*
- Implemented end-to-end sentiment analysis (fine-tuned BERT), topic modeling (developed and fine-tuned BERTopic, LDA), and multi-task learning (using PyTorch) to extract semantic information from millions of user review comments and develop personalized recommendations, countering popularity bias and cold-start problem.
- Efficiently preprocessed large textual data (tokenization, stopwords removal, lemmatization, etc.) by utilizing NLTK & scikit-learn.
- Tuned hyperparameters via coherence analysis and utilized Pyspark and Google Cloud to accelerate the model training process for the large dataset.

**Time-series Forecasting of Real Estate Taxes using Deep Learning Techniques** | *Leading Researcher*
*Funded by National Natural Science Foundation of China, 2017*
- Developed time-series prediction models using ARIMA, exponential smoothing models, and recurrent neural networks to forecast real estate tax revenue for Shanghai pilot projects, achieving 72.9% accuracy with RNN and 5%-12% lower with the other models.
- Enhanced data pipeline and feature engineering via variable transformation, missing data imputation, and input standardization.
- Led the research and published research findings in an academic journal.

## Work Experience

**Data Scientist Intern** September 2024 - December 2024
*Amazon* *Sunnyvale, CA*

- Defined the research problems and designed the scientific solutions for the performance improvement on RAG system in enhancing LLM capabilities of responding to user queries to robotic assistants.
- Applied and extended state-of-art information retrieval techniques to improve RAG system in guiding robot responsive behavior, achieving an increase in Recall@K for context retrieval and a reduction in latency.
- Converted the methodological innovation design into code in Python, tested it independently, and delivered it in collaboration with software engineers.

**Data Analyst - Modeling Intern** June 2023 - August 2023
*Chicago Metropolitan Agency for Planning* *Chicago, IL*

- Developed Python scripts for efficient measurement of transit accessibility index, integrating multi-source data inputs.
- Automated model deployment for computing new index values with updated datasets.
- Optimized the EMME auto-routing algorithm and cross-validated it with Google Maps routing, reducing the percentage of highly mismatched routes by 70% and moderately mismatched routes by 45%.
- Visualized the transit accessibility index using R and Tableau, deploying interactive online maps.

**Research Scientist Intern - ML Modeling** June 2020 - August 2020
*Wuhan Transportation Development Strategy Institute* *Wuhan, China*

- Implemented Conv-LSTM neural networks to predict citywide dockless bike sharing demand, capturing spatio-temporal dependencies in the demand and improving overall prediction accuracy (MAPE) by 5%-15% compared to ARIMA, LSTM, GRU.
- Developed performance-based stacked tree-based models (XGBoost, LightGBM, CatBoost) to predict ride-hailing travel time using city-scale mobility data, achieving a MAPE of 3%-18% for off-peak and peak predictions.
- Collected open data using web scraping techniques in Python (Beautiful Soup, Scrapy), acquiring real-time data of transit services.

**Data Science and Big Data Analytics Intern** June 2018 - August 2018
*Wuhan Geomatics Institute* *Wuhan, China*

- Developed spatio-temporal data mining algorithms to identify short-term and long-term citywide commute behavior patterns using cell phone signaling data, informing government decision-making on transportation investments.
- Enhanced data visualization using Vega and D3.js. Published in the 2019 annual report of the Wuhan Municipal Government.

## Certificates & Awards

- 2024 Outstanding PhD Student Award for Academic Achievement and Leadership, University of Washington
- 2022 Special Student Service and Contribution, University of Washington
- 2019 Grand Prize, Geographic Information Science and Technology Innovation Contest, Ministry of Natural Resources of China
- 2018 First-level Scholarships, Wuhan University (Top 5 %)
- 2017 Outstanding Undergraduate Thesis, Department of Education of Hubei Province, China
- 2016 Honorable Mention, Interdisciplinary Contest in Modeling , Consortium for Mathematics and its Applications (COMAP)
- 2015 National Scholarship, Ministry of Education of China (Top 1 %)
- 2014 National Scholarship, Ministry of Education of China (Top 1 %)

## Technical Skills

**Programming Languages**: Python, R, SQL, JavaScript, Html
**Python Libraries for ML/DL**: PyTorch, TensorFlow, Keras, Scikit-learn, SciPy, Numpy, Pandas, Geopandas
**NLP**: Scikit-learn, Genism, spaCy, Pyspark MLlib
**Statistical Analysis**: Time Series Analysis, Causal Inference, Bayesian Inference, Network Analysis, Experiment Design and Analysis
**Distributed Computing**: Spark, Pyspark, MapReduce, Hive, Yarn
**Spatial Analysis & Simulation**: ArcGIS, QGIS, GeoDa, SUMO, MATSim
**Data Visualization**: Tableau, Power BI, Vega, D3.js, CSS