# Ho Wong

(857) 265-5870 | howong112@outlook.com | www.linkedin.com/in/hwong511 | hwong511.github.io

## EDUCATION

**University of Wisconsin-Madison**, Madison, WI August 2024 - December 2025
*M.S. Data Science in Human Behavior*
**Cumulative GPA:** 4.00/4.00
**Certificate:** *Google Advanced Data Analytics* – Coursera, 2025

## EXPERIENCE

**Capstone Project (Industry Partnership)** | Carnegie Learning, Remote May 2025 - December 2025

- Built an XGBoost model automating student engagement detection from IMS Caliper Analytics logs (0.639 AUC) and achieved minimal train-test gap through rigorous temporal feature engineering and student-level cross-validation
- Engineered temporal and behavioral features from 19,000+ multimodal records, integrating restricted-use observation data with event-level telemetry via time-based joins
- Communicated model insights and feature importance to stakeholders through interactive Tableau dashboards and a live presentation at DataJam

**Capstone Project (Industry Partnership)** | American Family Insurance, Remote May 2025 - December 2025

- Architected a multi-source data pipeline integrating restricted-use demographic, insurance, survey, and climate datasets across 16 Florida counties using KNN matching and temporal joins
- Engineered ensemble machine learning models (XGBoost, Random Forest, LMEM) that extracted unbiased feature importance rankings from unsampled data for GenAgent calibration, revealing homeownership as the dominant migration predictor over hurricane severity metrics
- Collaborated with team to design self-supervised prompt optimization pipeline using DSPy and Pydantic that iteratively refined LLM instructions across 200+ optimization cycles

**Clinical Research Coordinator** | Massachusetts General Hospital, Boston, MA February 2023 - February 2024

- Collaborated with clinical teams to collect and organize data from multiple sources including participant assessments, tracking systems, and follow-up records across intervention studies
- Performed data cleaning and quality control on clinical datasets, identifying inconsistencies and ensuring data integrity for research analysis and reporting
- Managed participant tracking workflows and conducted follow-up with hard-to-reach populations, maintaining organized records to support timely project completion

**Research Assistant** | Boston College Mind Perception Lab, Boston, MA January 2020 - December 2022

- Designed independent research study collecting primary data across multiple countries, developing data collection methodology to analyze behavioral patterns during COVID-19 pandemic
- Performed statistical analysis in R examining relationships between variables and real-world outcomes, identifying significant patterns through data visualization and structured reports
- Cleaned and analyzed datasets to identify cultural differences in perception patterns, presenting findings to academic stakeholders through reports and visualizations

## SKILLS

Python, R, SQL | Statistical Analysis & Experimental Design | Regression & Mixed Models | Machine Learning | Spark, AWS/GCP | Tableau

## PROJECTS

**Car Insurance Fraud Detection |** *Python, PyTorch, CNN, Vision Transformers, Hugging Face, Kaggle*
Objective: Detect fraudulent car insurance claims from damage photos using deep learning, addressing the critical business challenge of identifying fraud in highly imbalanced real-world datasets.

- Built and productionized deep learning models using PyTorch and Hugging Face Vision Transformers achieving 98% accuracy and 0.95 AUC-ROC, demonstrating end-to-end capability from exploration to deployment
- Developed comprehensive data preprocessing and augmentation pipeline handling extreme class imbalance, optimizing model performance through systematic hyperparameter tuning and cross-validation
- Maintained model reliability through rigorous evaluation frameworks monitoring precision and recall metrics critical for production deployment

**Airline Satisfaction Prediction |** *Python, RStudio, Random Forest, scikit-learn, tidymodels*
Objective: Predict passenger satisfaction to enable proactive service improvements and identify key drivers of customer retention in the competitive airline market.

- Built end-to-end ML pipeline using Python scikit-learn and R tidymodels to process 100,000+ records, deploying Random Forest ensemble models achieving 96%+ accuracy with robust cross-validation
- Identified key satisfaction drivers through feature importance analysis and experimentation, providing data-driven recommendations to improve user experience and retention
- Implemented reproducible preprocessing and feature engineering workflows establishing pipeline architecture suitable for production deployment and continuous iteration

**Big Data Processing with Spark and Hive |** *Python, PySpark, Hive, HDFS, Docker*
Objective: Build scalable distributed data infrastructure to analyze competitive programming datasets, demonstrating enterprise-level big data engineering capabilities for processing and modeling large-scale datasets.

- Contributed to scalable data pipeline infrastructure processing millions of records through optimized ETL workflows using PySpark, SQL, and distributed processing, reducing query execution time by 60%
- Built production-grade environment using Docker containerization and multi-node Spark clusters, establishing reproducible infrastructure supporting analytics workflows and model deployment
- Developed end-to-end ML workflow on distributed infrastructure demonstrating capability to work across data engineering and machine learning platforms

**Real-Time Weather Data System |** *Python, Cassandra, Spark, gRPC, Docker*
Objective: Design fault-tolerant distributed system for real-time weather data ingestion and querying, demonstrating expertise in building scalable streaming data architectures with enterprise reliability requirements.

- Built real-time data ingestion pipeline using Python and gRPC enabling streaming data collection across distributed Cassandra cluster with millisecond-latency queries
- Designed and implemented database schemas with optimized partition and clustering keys, writing infrastructure code for fault-tolerant data storage and retrieval
- Developed containerized deployment using Docker and cloud infrastructure fundamentals, maintaining documentation for system architecture and data observability practices