Tianyi Wu

BC, Canada | twa174@sfu.ca | +1 437-984-0331 | LinkedIn Profile | Github Page

Skills

- **Programming Languages:** Python, SQL, R, C/C++, MATLAB, CQL
- **MLOps**: AWS(ECS, EC2, Sagemaker, Cloudwatch, etc.), Git, Github Actions, Docker, Bash Scripting, MLflow, DVC
- **Generative AI**: AWS Bedrock, Pydantic AI, DSPy, Prompt Engineering, OpenAI, Anthropic
- **Machine Learning Algorithms:** Neural Networks, Random Forest, XGBoost, SVM, KNN, K-Means, UMAP, Leiden, Generalized Linear Regression, Stacking
- **Frameworks/Libraries:** LangChain, PyTorch, TensorFlow, PySpark, Spark SQL, MapReduce, Scikit-learn, Pydantic, Numpy, Pandas, Papermill, Markdown, LaTeX, Figma, Tableau, Power BI, Label Studio
- **Quick learner and strong adaptability:** quickly mastered new domain knowledge and technologies within the first week of internship, earning recognition from the supervisor for delivering impactful results
- **Leadership:** showcased effective leadership and teamwork, achieving award-winning results in various competitions.

Education

Simon Fraser University — Master's in Professional Computer Science (**Big Data**) 09/2024 - 04/2026 | Canada

University of Waterloo — Bachelor of Mathematics 09/2021 - 08/2024 | Canada

- Honors **Statistics & Computational Mathematics** Double Major

Experience

Data Scientist | Rio Tinto | *AWS, Docker, Pytorch, Git, LLM, CV* 05/2025 - Present | Vancouver, Canada

- Developed an **agentic LLM pipeline** on **AWS Bedrock** using **Azure OCR** and **Pydantic** for structured data extraction; automated metadata processing for **3,000+** documents, significantly reducing manual data cleansing overhead
- Architected and orchestrated an **end-to-end containerized MLOps pipeline** on AWS **ECS** with parallel processing and full observability via **CloudWatch** and **MLflow**, reducing product delivery time from **3 weeks to less than 1 day**
- Implemented **Human-in-the-Loop (HITL)** workflows by integrating SAM 3 for zero-shot pre-annotation within Label Studio, accelerating CV model training cycles while maintaining high data quality
- **Led a cross-functional project** integrating **1D-CNN** models with hyperspectral data; collaborated directly with geoscientists to co-design label schemas and validation protocols, delivering high-accuracy prediction maps for field operations
- Integrated **GitHub Actions** for **CI/CD** and provided ongoing bug fixes for an external library, ensuring reliable deployments and uninterrupted internal progress

Graduate Teaching Assistant | Simon Fraser University 09/2025 - 12/2025 | Vancouver, Canada

- Assisted graduate students with labs and assignments on big data technologies, including Hadoop, **Spark**, NoSQL, and scalable data pipelines, providing technical guidance and debugging support

Data Analyst | Mashang Consumer Finance| *SQL, Python, Excel* 06/2023 - 07/2023 | China

- Preprocessed and performed quality control on **30+** datasets (**1M+ rows**), transforming raw inputs into structured data using **SQL**
- Analyzed historical customer service data to identify high-risk client groups, developing a **BI dashboard** to monitor and alert key metrics in real-time

Projects

Machine Learning for Portfolio Optimization | *Python, Pandas, Scikit-learn* 04/2025

- **Led a team of 4**, won the **2025 SFU MPCS Innovation Prize Competition** by defining the scope of work, prioritizing technical objectives, and coordinating tasks
- Developed a machine learning driven portfolio optimization tool leveraging **Nested Clustered Optimization** algorithm and Optimal Number of Clustering, improving upon traditional Markowitz Efficient Frontier methods
- Built an interactive website allowing users to input stocks and risk thresholds to generate optimized portfolios dynamically

Second Place Award at 2024 Scotiabank AI-Kathon | *Python, PyTorch, Pandas, NLTK, Transformers, LLM* 01/2024

- **Led a team of 4** and **won second place** at the event, leveraging advanced **LLM** and **NLP** techniques to derive business insights from **9000+** raw customer feedback
- Engineered a filtering pipeline with **LLM prompt engineering** and divide-and-conquer strategies to remove irrelevant reviews and improve data quality
- Integrated **RoBERTa**-based sentiment analysis to identify extreme sentiments and used **BERTopic** for extracting key pain points, uncovering targeted insights, and enabling data-driven product enhancements
- **Utilized** a **guided BERTopic** model with **UMAP** to classify reviews into 20 topics, leveraging preprocessing (**cleaning, tokenization, lemmatization**) for optimized analysis