Mutian He

Data Engineer / Data Science Intern

Email: mhe3@scu.edu | **Phone**: 408-548-1060 | San Jose, CA
LinkedIn: www.linkedin.com/in/mutian-he **Github:** github.com/austin10231 **Portfolio**: www.mutianhe.com

SUMMARY

Data Engineering and Machine Learning graduate student with experience building end-to-end data pipelines and ML-driven systems. Skilled in designing data workflows and developing systems that support data-driven decision making.

EDUCATION

Santa Clara University, Leavey School of Business Santa Clara, US
Master of Science in Information Systems Dec 2026

- Data Analytics with Python, Cloud Computing, Database Systems, Machine Learning, Deep Learning, NLP

University of Glasgow, School of Computing Science Glasgow, UK
Bachelor of Science in Computer Science (BSc CS) Jun 2024

- Data Structures & Algorithms, Database Systems, Software Engineering, Human-Computer Interaction

PROJECTS

Image Search System (Multimodal Machine Learning) — 2026 | PyTorch, CLIP, FAISS | Live Demo | GitHub

- Built a **CLIP-based multimodal retrieval system** enabling text-to-image search by projecting images and queries into a shared embedding space.
- Designed an end-to-end retrieval pipeline including **data preprocessing, batch image embedding generation, vector indexing, and Top-K similarity search**.
- Leveraged **FAISS for efficient approximate nearest neighbor (ANN) search**, significantly improving scalability and latency over brute-force retrieval.
- Implemented **evaluation metrics (Recall@1/5/10 and Median Rank)** to quantitatively assess retrieval performance and validate model effectiveness.

Cloud-Based Stock Data Pipeline –Data Engineering (2026) | AWS (S3, Glue, Athena), SQL, Kafka, ETL | Github

- Built a real-time stock data pipeline to enable **low-latency market data ingestion** and analysis, supporting **near real-time trading insights**.
- Designed a scalable AWS data lake (S3 + Glue) to **centralize fragmented data sources** and improve **data accessibility**.
- Automated streaming ingestion workflows, reducing **manual data processing** and enabling **continuous data availability**.
- Enabled serverless querying with Amazon Athena, allowing **faster exploratory analysis** and reducing **infrastructure overhead.**

FinSight AI – Cloud Engineering (2026) | (Python, AWS, LLM/NLP) | Live Demo | Github

- Built an end-to-end system to **automate extraction and analysis of SEC 10-K risk disclosures (Item 1A)** from unstructured filings.
- Developed an **AWS-based pipeline (S3, Textract, Bedrock)** to convert raw documents into structured, analyzable risk data.
- Implemented **cross-year risk comparison** to identify newly added and removed risk factors across filings.
- Designed an **LLM-powered agent** to prioritize risks (impact, likelihood, urgency) and generate structured reports.

Bank Marketing Subscription Prediction – Machine Learning (2026) | Pandas, Scikit-Learn, SMOTE, SHAP | Github

- Built a machine learning pipeline to **predict customer subscription likelihood**, enabling **more effective targeting in marketing campaigns.**
- Addressed **severe class imbalance (1:9)** using SMOTE and threshold tuning, improving **minority-class recall and campaign reach**.
- Optimized decision threshold to **prioritize high-probability customers**, supporting **more efficient allocation of marketing resources**.
- Applied SHAP to **identify key drivers of customer conversion**, informing **data-driven marketing strategy** (ROC-AUC: 0.80).

SKILLS

Programming: Python, SQL (MySQL, SQLite, PostgreSQL, MongoDB), Java

Data Engineering: Apache Kafka, Apache Airflow, ETL/ELT Pipelines, Streaming Architecture, Data Modeling

Machine Learning / Deep Learning: Scikit-learn, Feature Engineering, Cross-Validation (K-Fold), Hyperparameter Tuning, Model Evaluation (ROC-AUC, F1, Recall@K), Supervised/ Unsupervised Learning (Random Forest, XGBoost, K-Means, PCA), Multimodal Learning (CLIP, Embeddings, Transformer)

Cloud & Platforms: AWS (EC2, S3, Glue, Athena, RDS, Lambda), Data Lake Architecture

Tools: Git, Docker, Linux, Jenkins, Streamlit, Flask, VS Code, Jupyter Notebook