Yufei Song

(412)-450-7330 | yufeison@andrew.cmu.edu | Pittsburgh, PA | GitHub | LinkedIn

EDUCATION

Carnegie Mellon University *M.S. in Information Systems Management* *GPA: 3.74* Pittsburgh, PA | May 2026

- Coursework: Statistics, Distributed Systems, Cloud Computing, Machine Learning in Production, Deep Learning, Database Management, Time Series Forecasting in Python, Business Analysis, GenAI in Software Development

University of Surrey *B.S. in Accounting and Finance* *GPA: 3.76* Jun. 2024

SKILLS

Programming: Python, SQL, R, Java
Data Analysis: Power BI, Tableau, Excel, SAP, Pandas, NumPy, Matplotlib, Seaborn, scikit-learn, TensorFlow, NLP
Data Systems & Cloud: PostgreSQL, SQLite, MongoDB, Hadoop (MapReduce), Spark, AWS, Azure, GCP, Snowflake, Git, Docker

EXPERIENCE

Research Assistant (Data Analytics & NLP) | School of Computer Science, CMU Pittsburgh, PA | Nov. 2025 - Present

- Built a scalable Python data pipeline to extract, preprocess, and validate 12M+ JSTOR records, implementing schema normalization and automated data quality checks; optimized LLM token usage by ~32% for large-scale text processing (9.6K+ documents)
- Fine-tuned a BERT-based model for text classification on 2.3M+ Twitter bios using LLM-assisted labeling; built end-to-end preprocessing, validation, and evaluation pipelines; trained on AWS, achieving strong precision, recall, and F1-score
- Developed interactive network analysis visualizations to support downstream research workflows and large-scale network analysis

Analytics Engineer | USA Today · Gannett Media Pittsburgh, PA | Aug. 2025 - Dec. 2025

- Built scalable Python-based ingestion pipelines to crawl, clean, and apply NLP processing on news articles, enabling detection of suspicious articles for analytics and content protection
- Designed analytics-ready data models in MongoDB Atlas from BigQuery sources, powering text vectorization and similarity scoring pipelines across historical and incoming articles, reducing false positives in suspicious content detection by ~24% and lowering editorial review workload
- Built operational Streamlit BI dashboards enabling monitoring of content risk signals, duplication rates, and investigation outcomes; improved decision efficiency and consistency for editorial teams
- Productionized analytics pipelines with Docker, supporting consistent deployments across development and analytics environments

Data Strategy Intern | Publicis Groupe Shanghai, China | Apr. 2024 - Jul. 2024

- Built a cross-platform metrics framework across TikTok, RedNote, and Weibo by modeling search volume as demand, user-generated content volume as supply, and engagement metrics as performance, enabling analytics-ready insights, reducing ad-hoc analysis requests by ~40%
- Developed Python-based analyses and bubble-chart visualizations to map high-demand low-supply opportunities, supporting competitor benchmarking, trend tracking, and clear identification of white-space content and market opportunities
- Translated BI insights into actionable marketing strategies, guiding keyword prioritization and content investment decisions; contributed to a ~7% client performance lift within one quarter through data-driven recommendations and stakeholder workshops

Business Analytics Intern | Waterdrop Beijing, China | Jul. 2023 - Aug. 2023

- Engineered dynamic KPI reporting dashboards and automated SQL reporting pipelines for clinical and business performance, cutting weekly reporting turnaround time by ~50%; conducted funnel analysis to identify conversion drop-offs, driving targeted operational changes that improved conversion efficiency
- Built financial projections to support short-term revenue planning, operational decisions, and staffing adjustments, helping diagnose and recover from a ~21% revenue drop within two weeks
- Led onboarding workshops on analytics workflows, reducing new intern ramp-up time and recurring reporting errors

PROJECTS

Real-Time User Behavior Analytics Pipeline | *Python, FastAPI, Kafka, Spark, MongoDB, Streamlit, Docker* Dec. 2025

- Built an end-to-end real-time data pipeline to process 2.8M+ event-level user behavior records using Kafka for streaming ingestion and Spark Structured Streaming for real-time transformation, with MongoDB serving as the downstream analytics store
- Analyzed real-time user behavior and conversion funnels by building a streaming analytics pipeline to process event-level data
- Implemented FastAPI-based ingestion services and containerized the ingestion, processing, and analytics layers with Docker, validating data flows via Postman to ensure reproducible environments and production-style service integration

E-Commerce Transaction Analytics Platform | *AWS, API Gateway, Lambda, Kinesis, S3, Redshift, Power BI* Jan. 2026

- Built an AWS data pipeline and ETL workflows processing 541K+ e-commerce transactions, transforming data into analytics-ready schemas in Amazon Redshift using API Gateway, Lambda, Kinesis, S3, DynamoDB, Python, and SQL
- Enabled Power BI dashboards on Redshift to analyze product, revenue, and geographic performance insights

ACTIVITIES & AWARDS

Workshop Director, CMU Data Science Club Spring 2025
3rd Place (Team Lead), 2024 CMU IronViz Data Visualization Challenge Fall 2024