Steven Liu — San Francisco, CA | GitHub | Portfolio | LinkedIn | steventheliu@gmail.com |

WORK EXPERIENCE

Software Engineer / Data Scientist — AIMdyn — Santa Barbara, CA (Remote) — (2021 - 2024)

- Developed predictive and machine learning models for dynamic simulations, predictive analysis, and optimization, tackling challenges like ship refueling pathing, network traffic forecasting, and COVID-19 predictions.
- **COVID-19 Forecasting Model:** Ranked in the top 5% for predicting deaths, hospitalizations, and infections, surpassing competitors like Johns Hopkins and Microsoft.
- **Ship Refueling Optimization:** Improved fuel efficiency by 20%, driving cost savings and environmental benefits through modeling.
- Automated data pipelines resulting in reduced ML model training time by 30%.
- Provided actionable insights using **Tableau**, **Matplotlib**, and **Seaborn**, informing data-driven business strategies.
- Conducted robust model testing with **Unittest** and containerized applications with **Docker** for scalability and reliability.
- Collaborated with cross-functional teams via **Jira** and **Confluence** to align goals and ensure timely communication.

EDUCATION

University of California, San Diego — *B.S. in Data Science* — *3.72 GPA* — (2018 - 2021)

- **Data Structures & Algorithms** - Recursion, Higher-Order, OOP, Complexity, and Data Types
- **Application of Data Science** - Statistics, Machine Learning Algorithms, A/B Testing, Web Scraping and Data Systems
- **Database Management** - Relational Database, Schema Design, Query Language and Optimization
- **Scalable Analytics Systems** - Big Data, Memory Hierarchy, Distributed Systems, Model Selection, ETL, Deployment at Scale
- **Modeling & Machine Learning** -Natural Language Processing, Supervised/Unsupervised, Robotics, Deep Learning

PERSONAL PROJECTS

Transaction Data Pipeline - Set up a data pipeline for batch and real-time processing of transaction data

- **Batch Processing:** Built **Spark** jobs to process and transform historical transaction data, including ingestion from **PostgreSQL** and file-based sources.
- Validated data quality through schema checks, missing data handling and deduplication.
- **Real-Time Streaming:** Configured **Kafka** producers and consumers to handle synthetic transaction data streams.
- **Data Storage:** Optimized **PostgreSQL** and **MongoDB** with indexing, partitioning, and retention policies.
- **Orchestration:** Automated workflows with **Airflow**, leveraging **Docker Compose** for setup and task management.
- **Monitoring:** Integrated logging for DAG execution in Airflow and verified **Kafka** consumer performance metrics.

Ticketmaster Replica - Set up web application to view and book tickets for events with distributed microservices

- Optimized for searching events with search based database such as **ElasticSearch**
- Prevents issues such as double-booking with **Optimistic Locking** with **Redis** cache
- Ensure data consistency among data storages **Redis** and **PostgreSQL** with **CDC** (Change Data Capture) and Kafka
- Improved user experience by updating available tickets for events in real-time and ticket confirmation via **SSE**
- Distributed seamless access to events with **CDN** (**CloudFront**) deliver static content globally
- Improved user experience during surges from popular events with random waiting queue for booking
- Services are implemented in **Docker** containers to be scalable

CERTIFICATIONS

- **AWS Certified Cloud Practitioner** - 11/2024

TECHNICAL SKILLS

- **Languages**: Python, CSS, HTML, JavaScript, SQL, C++, Java, MATLAB, R
- **Packages**: Flask, BeautifulSoup, Unittest, Coverage, Tkinter, Selenium
- **Data Science:** Pandas, NumPy, SciPy, Scikit-Learn, TensorFlow, PyTorch, Spark, Dask, Hadoop, Matplotlib
- **Tools & Platforms**: Git, GitHub, Docker, CI/CD, Airflow, Grafana, Prometheus
- **System Design:** Flask, Django, PostgreSQL, MongoDB, Cassandra, ElasticSearch, Redis
- **Cloud:** AWS (S3, RDS, Lambda, CloudFront)