Toby Liu

Santa Barbara, CA | 713-294-5792 | tobyliu2004@gmail.com | Github | LinkedIn | Portfolio

EDUCATION

UC Santa Barbara, B.S. in Statistics & Data Science, B.A in Economics, L&S Honors, Expected June 2026
Relevant Coursework: Data Structures and Algorithms, Statistical Machine Learning(Graduate),Probability & Statistics, Linear Algebra, Regression Analysis, Design of Experiments, Time Series(Graduate), Bayesian Statistics

TECHNICAL SKILLS

Languages: Python, JavaScript, SQL, R, Java, C++, TypeScript
Web: React, FastAPI, HTML/CSS, REST APIs
Tools: Git/GitHub, Docker, CI/CD, Pandas, NumPy, Jupyter, Streamlit, pytest, A/B Testing
Databases: PostgreSQL, MongoDB, SQL(MySQL, SQLite), Redis
Cloud & Data Engineering: AWS(S3, EC2), ETL, Data Pipelines
AI/ML: PyTorch, TensorFlow, Scikit-learn, XGBoost, Deep Learning, LLMs

EXPERIENCE

Houston Methodist (Medical Artificial Intelligence & Innovation Lab) Houston, TX, **Software Engineer Intern**, May 2025 – August 2025

- Engineered multi modal ML pipeline integrating 4 genomic datasets using statistical design principles to predict bladder cancer patient response to chemotherapy treatment, achieving 76.59% AUC, outperforming published baseline by 2.6%
- Architected modular ETL pipeline processing 500,000+ features from 285 TCGA patient records, implementing variance-based feature selection algorithm that reduced dimensionality by 98%
- Built ensemble machine learning framework comparing 5 algorithms (XGBoost, Random Forest, Logistic Regression, ElasticNet, MLP) with 5-fold cross-validation, reducing prediction variance from 15% to 5.2%

Boston Children's (Computational Biology Department), Boston, MA, **Data Science Intern**, June 2024 – August 2024

- Processed ChIP-seq and ATAC-seq genomic datasets through DANPOS3 pipeline for 5 rare pediatric disorders, analyzing chromatin accessibility patterns across 100+ patient samples
- Optimized R visualization workflow for peak calling results, reducing manual plotting time from 2 hours to 20 minutes per dataset
- Generated 200+ programmatic data visualizations (heatmaps and coverage plots) identifying differential chromatin regions between patient and control groups

PROJECTS

Competitive Intelligence Platform | Python, HDBSCAN, SQLite, FastAPI, NLP | October 2025 (Unwrapathon)

- Designed end-to-end data pipeline with Reddit API integration processing 4,662 comments; implemented .npz caching reducing embedding generation from 30s to <1s on subsequent runs
- Architected modular backend with 13 Python modules integrating sentence-transformers and HDBSCAN, designed SQLite schema with efficient indexing for 2.5MB database
- Developed automated comparison engine testing 28 clustering configurations across 32 feature categories with statistical testing module for programmatic gap identification

Research Paper RAG Assistant | Python, PyTorch, PostgreSQL, FastAPI, React | August 2025

- Architected full-stack retrieval-augmented QA system with React frontend and FastAPI backend enabling search across 3,000 ArXiv AI/ML papers
- Engineered hybrid search combining BM25 keyword algorithm with sentence-transformers embeddings in PostgreSQL/pgvector with optimized indexing
- Designed RESTful API with cosine similarity ranking returning top 5 passages per query; built web interface with source highlighting and BibTeX export