

Hank Sha

San Jose, CA | (510) 320-1116 | hankssha@gmail.com | linkedin.com/in/hank-sha | hanksha.com | Github

EDUCATION

University of California, Santa Barbara

B.S. Statistics and Data Science

Goleta, CA

Sep 2022 – June 2026

- **Relevant Coursework:** Experimental Design, Regression Analysis, Probability & Statistics, Statistical Machine Learning, Bayesian Data Analysis, Time Series Analysis, Applied Stochastic Processes

EXPERIENCE

UC Riverside

Machine Learning Research Assistant

Remote

June 2025 – Mar 2026

- Produced the first systematic characterization of a newly proposed graph-reasoning methodology, measured across 10K+ topological graphs, by converting a single-graph prototype into an unattended evaluation harness and implementing the Graph Neural Network architecture from scratch.
- Delivered 75% runtime reduction, 20% lower GPU memory on local hardware, by profiling the serving path to isolate KV-cache thrash and tokenizer overhead, then applying batched inference, quantization, and CUDA memory tuning.

ABVC BioPharma

Clinical Data Scientist Intern

Remote

Aug 2025 – Nov 2025

- Identified 3 distinct patient response phenotypes, via K-means clustering on longitudinal outcome data, revealing that gradual responders showed the strongest drug response and that 55% of responders plateaued below the 50%-reduction endpoint.
- Isolated 2 comorbidities with statistically significant association to placebo response (chi-squared testing) by modeling placebo effect within a randomized controlled Phase II trial, reconciling ITT and PP via LOCF imputation and benchmarking 4 classifiers to an 83% AUC-ROC XGBoost model.
- Caught 2 data-leakage features before they corrupted a dropout-risk model by building correlation matrix and diagnostic visualizations in Matplotlib and Seaborn, then trained an XGBoost classifier to 83% recall (10-fold cross-validation) across 15+ engineered clinical features using SMOTE resampling against severe class imbalance.

GitRoll

AI/ML Engineer Intern

San Francisco, CA

Sep 2024 – Apr 2025

- Shipped a production Retrieval-Augmented Generation (RAG) pipeline serving 45K+ users, generating GitHub-derived developer metrics as the product's core scoring surface.
- Raised pull-request classification precision to 92% by hand-labeling 216+ issues into a ground-truth set and standing up a human-in-the-loop evaluation framework, shipped as the company's core production metric..

PROJECTS

Project Nexus — Civic engagement track winner, UCSB Data4Good

2026

Python, LangChain, Pinecone, Neo4j

- Delivered live RAG demo that let non-technical judges query 500K+ documents and surface evidence-backed findings in real time via entity extraction and relationship modeling.

Biotech Intelligence Engine — Data platform for biotech-catalyst trading

2026 – present

Python, FastAPI, Celery, TypeScript, Docker

- Unified 5 heterogeneous external sources into a 137K+ entry feed (SEC filings, clinical trials, FDA calendars, news), deployed live through GitHub Actions CI/CD.
- Enforced idempotent ingestion by writing SQL deduplication and cross-table integrity queries against a 137K+ entry Postgres store, eliminating duplicate records introduced by overlapping nightly syncs.

NBA Contract Value — Analytics dashboard for best NBA contracts

2026 – present

Python, BeautifulSoup, pandas, scikit-learn, deck.gl, Typescript

- Built a nightly ETL pipeline ingesting all 334 active NBA contracts via a shared rate-limited HTTP client with a per-host request budget, sustaining unattended scraping without triggering source-side blocking.
- Engineered a composite production metric from true shooting %, usage rate, minutes, and games played, then fit a league-average price-of-output regression on market-priced deals to rank every contract by residual surplus.

TECHNICAL SKILLS

Languages: Python, SQL, R, TypeScript, bash

Statistics & Experimentation: Experimental design, hypothesis testing, regression analysis, Bayesian inference, time series, cross-validation, causal inference (RCT, ITT/PP estimands), class imbalance (SMOTE)

Machine Learning: XGBoost, scikit-learn, PyTorch, clustering, feature engineering, model evaluation (AUC-ROC, precision, recall)

Data & Visualization: pandas, NumPy, Matplotlib, Seaborn, deck.gl, Postgres, MongoDB, ETL, BeautifulSoup

Platform: FastAPI, Docker, AWS, GCP, CI/CD (GitHub Actions), Celery, Redis, Vercel