David Liu

Senior Data Scientist · 14 years building production ML systems across biotech, e-commerce, healthcare, recruiting, and fintech.

SUMMARY

I build production machine learning systems end-to-end — from data ingestion and feature engineering through model training, serving, and measurement. My focus is what actually ships: defining the right metric, getting the right data, and ensuring the model holds up in front of real users. Currently leading personalization at Shipt for millions of grocery shoppers; previously a core ML engineer at Freenome (cancer-detection blood test) and Change Healthcare. On the side I run AutoTrader, a fully autonomous market-prediction system, and publish independent research at the boundary of multi-agent AI, distribution-shift model selection, and statistical genomics.

EXPERIENCE

Senior Data Scientist — Shipt (Personalization Team)

SEP 2024 — PRESENT

Lead data scientist on Shipt's Personalization team; recommender systems behind every personalized shelf on the platform. Mentored 4 data scientists including 2 direct reports.

- Delivered up to **23% Personalized GMV lift** across A/B-tested shelves (Trending Items, Similar Items, Deals For You, Complementary Items) over 90 days.
- Designed and built a two-tower deep-retrieval model replacing the legacy ALS recommender; **3–15% lift** on Deals shelves and **11% higher CTR**; offline +406% NDCG, +87% novelty, +19% long-tail coverage vs. baseline.
- Designed user/product embeddings going beyond text matching (descriptions, categories, pricing, retailer info, dietary preferences, behavioral signals); FAISS ANN infrastructure on GCS for low-latency retrieval at scale.
- Created the Personalization Interaction Score — a composite funnel metric (views → clicks → ATC → purchases, weighted by depth) that replaced single-metric optimization. Adopted by teammates for next-gen real-time recommenders.
- Developed a price-weighted ATC approach for coldstart users; **47% engagement lift** and **8% more first-time orders** over 90 days; technique adopted across other shelves.
- Built an automated Shelf Attribution pipeline using fuzzy-string matching; identified previously reported 5% attribution figures were irreproducible (actual: 1–4%); transparent reporting catalyzed a shift to defensible success metrics.
- Designed and built a Customer Intelligence Platform (CIP) and an "Essentials" recommender shelf usable as both a standalone shelf and a relevance filter for others.
- Partnered with Engineering to replace brittle CSV-based recommendation delivery with Kafka-based pipelines across all legacy recommenders.
- Designed a Retrieval-Augmented Generation system over Shipt's internal retailer catalogs and built the business case that secured agentic-AI investment.
- Served as the team's sole data scientist for its first four months — maintained 16 Discovery Science repos while designing next-generation infrastructure.

Machine Learning Research Engineer — Freenome

NOV 2020 — JUN 2024

Core ML engineer at a genomics company developing a blood test for early-stage cancer detection. Worked at the intersection of infrastructure and research.

- Key contributor to Freenome's core multiomics cancer-detection model that predicts cancer stage (1–4) from blood-draw data; built data abstractions for petabyte-scale genomic datasets that unblocked cross-analyte feature development and accelerated training/evaluation cycles.
- Built a model-comparison system tracking research-vs-production model performance side-by-side — required for FDA audit compliance, gave the team confidence that production models stayed aligned with research intent.
- Designed and built large portions of Freenome's distributed ML training and serving platform used daily by 30+ scientists; scaled CPU-bound training across O(100) machines, supported leave-one-out / K-fold evaluation, and built reproducible model-artifact storage.
- Led adoption of PyTorch, MLFlow, and RayTune across the ML team — replaced legacy tooling to unlock GPU acceleration, experiment tracking, and hyperparameter tuning at scale.
- Built a cloud-cost monitoring system surfacing the biggest GCP storage and compute expenses; visibility alone drove optimizations saving the company **over $10M annually** — one of the highest-ROI projects I've worked on.
- Recognized with a Servant Leadership Award, elected by managers and peers across the engineering organization.

Sr. Machine Learning Engineer — Change Healthcare

JAN 2020 — NOV 2020

ML systems for health-insurance claims processing — accuracy translates directly into operational cost savings.

- Designed a ranking model matching human workers to claims tasks based on skill, history, and complexity; **$7M annual value** by reducing manual task assignment and the need for additional hires.
- Built a classification model partitioning sensitive patient documents (image + text) to route claims to the correct processing workflow.
- Developed internal AWS tooling and production API infrastructure for the ML team's deployment pipeline.
- Led a cross-functional tiger team prototyping a conversational chatbot (Rasa + HuggingFace NLP) for internal claims-inquiry workflows.

Data Scientist — Riviera Partners

JAN 2019 — DEC 2019

ML models for an executive-recruiting firm; full pipeline from data collection to model serving.

- Developed a model suite: a job-departure-likelihood classifier, a regression model predicting team sizes from resume features, and a candidate-ranking model using a custom NDCG listwise loss.
- Built an end-to-end framework for rapid model prototyping, training, evaluation, and serving — enabled the team to iterate on new models without re-engineering infrastructure each time.
- Wrote scrapers harvesting structured candidate data from public sites and APIs.

Undergraduate Researcher — UC Berkeley

JAN 2017 — DEC 2018

- **California Institute for Energy and Environment (CIEE)** — built a recurrent neural network for predicting building energy usage, exploring how temporal consumption patterns can inform smarter grid management.
- **Bengson Research Lab, Sonoma State** — applied ML to EEG data to predict individualized occipital lobe activation patterns; demonstrated early feasibility for brain-computer interface applications.

Earlier Roles

2015 — 2017

- **Data Science Intern, Castlight Health** (2017) — entity matching and deduplication pipeline using gradient-boosted classifiers with hard-negative mining; 85–95% precision/recall across hospital, facility, and practitioner entities.
- **Data Science Contractor, Riviera Partners** (2016) — team-size prediction model from public data; Python wrapper for survival-model time-series analysis; Flask model-serving infrastructure.
- **URAP, Berkeley Institute of Data Science** (2016) — mapped UC Berkeley course progression across majors via class-taxonomy organization and deduplication.
- **Data Science Intern, Doximity** (2015) — gradient-boosted classifier for malformed scraped articles; reverse geocoding + fuzzy string matching to link doctors in news articles to facility profiles.

SELECTED PROJECTS

Meta Council — Multi-Expert AI Decision Support Platform 2025 — PRESENT · RESEARCH & PRODUCT

Multi-agent LLM framework where N expert agents (each with a unique persona and analytical framework) analyze queries in parallel, then a weighted synthesis step produces structured decision documents with confidence scores, dissent preservation, and risk matrices. Evaluated across 750+ benchmark runs spanning 6 domains and 5 models (3B to frontier-class): synthesis outperforms single-best by 29–58% ($p<0.0001$, $d=2.16$); the optimal aggregation method is domain-dependent; synthesis amplifies model quality non-linearly. Published as an independent research paper.

AutoTrader — ML-Powered Stock Prediction System 2024 — PRESENT · PERSONAL

Fully autonomous market-prediction system: collects nightly market data for 600+ tickers, engineers 500+ features across 8 source families, trains 1,800+ dual models (classifier for direction, regressor for magnitude), and delivers confidence-ranked predictions before market open every weekday. Built every piece — data ingestion, custom feature store, dual-model training framework with walk-forward validation and Optuna, FAISS-powered similarity search, tiered email subscription system with Stripe billing, multi-cloud GCP+Azure infrastructure — running autonomously for ~$235/month.

Genomics Methods Note (chr 6 MHC dominance) 2026 · RESEARCH

Single-subject methods paper documenting that the chr 6 MHC dominance in consumer-chip psychiatric polygenic risk scoring is European-cohort-specific and absent from East-Asian-cohort GWAS. Companion artifacts: an interactive d3 chromosome browser of EUR vs EAS GWS hits, and a browser-only client-side 23andMe v5 chip-PRS explorer. Five rounds of multi-lens subagent review pre-publication.

XGBoost Visual Guide 2026 · OPEN-SOURCE

Interactive visual textbook explaining XGBoost and gradient boosting from first principles — 10 sections covering decision trees, ensemble methods, animated step-by-step gradient boosting, learning rate effects, XGBoost-specific innovations (histogram splits, sparsity handling), early stopping, feature importance, and a hyperparameter cheat sheet. D3.js + Chart.js.

Sentic — Multi-dimensional sentiment analysis 2017 · OPEN-SOURCE

Python library for multi-dimensional sentiment analysis going beyond positive/negative polarity (mood, attention, sensitivity, aptitude, pleasantness) across 20+ languages, built on the SenticNet4 knowledge base. Available on PyPI.

PUBLICATIONS

Ancestry-matched GWAS removes the chromosome 6 MHC dominance in consumer-chip psychiatric polygenic risk scoring: a single-subject 23andMe v5 case analysis in an East Asian (Han Chinese) subject
2026 · Independent Research · CC BY 4.0 · Pharmacogenomics, polygenic risk score, schizophrenia, East Asian, single-subject, CPIC

Weighted Multi-Expert Synthesis for High-Stakes Decision Support: A Multi-Agent LLM Framework with Dissent Preservation
2026 · Independent Research · CC BY 4.0 · Multi-agent systems, LLM, decision support, weighted synthesis, dissent preservation, confidence calibration

Stability Bonus Regularization for Model Selection Under Positive-Class Distribution Shift
2026 · Independent Research · CC BY 4.0 · Model selection, distribution shift, cross-validation, class imbalance, regularization

EDUCATION

University of California, Berkeley — BS, Computer Science and Data Science (Dual Degree), Class of 2018. Berkeley Institute of Data Science Undergraduate Research Apprenticeship (2015).

TECHNICAL SKILLS

LANGUAGES	Python · SQL · Bash · Java · C/C++
ML & DATA	PyTorch · XGBoost · Scikit-learn · Pandas · MLFlow · FAISS · Gensim · NLTK · SpaCy
CLOUD	GCP · AWS · Azure
DATA INFRASTRUCTURE	PostgreSQL · Snowflake · MySQL · Spark · Kafka
ORCHESTRATION	Airflow · Flyte · Metaflow · GitHub Actions
INFRASTRUCTURE	Docker · Kubernetes · Git · CI/CD