Aaron Beppu

Applied ML engineer with experience bringing mission-critical ML systems to production.

- Broad impact, leading cross-functional projects spanning infrastructure, backend, ML, and UX
- Deep experience in real-time prediction, distributed systems, and integrating ML into business-critical systems

Experience

2021 - 2025: Staff ML Engineer, Hyperscience (Remote from San Francisco, CA)

Hyperscience applies ML to automate business processes involving human-readable documents, emphasizing customized, task-specific models.

Vision-Language Models (VLMs):

- Designed and implemented a benchmarking platform for evaluating models and prompts across diverse document tasks.
- Fine-tuned and evaluated OSS models; achieved production accuracy and automation targets with as few as 20 annotated examples.

Document Intelligence (non-generative ML):

- Cut training time for key models by 50% through pipeline profiling and optimization.
- Improved generalization to new document types with explicit “out-distribution” detection and altered confidence handling.
- Built annotation-assist tools (e.g. anomaly detection) yielding faster and more consistent labeled data.
- Delivered production models for Named Entity Recognition and PII redaction; redaction later became its own offering.
- Developed clustering/analysis tools for unlabeled data, prioritizing examples to maximize annotation efficiency.

2013 - 2021: Sr → Principal Software Engineer, Sift Science (San Francisco, CA)

Sift predicts fraud and abuse from fine-grained event data. Contributed to nearly every component of data ingestion, model training, serving, and automation.

ML / Data Science:

- Developed methodology to measure bias in ML models without access to sensitive attributes, proving bias effects were causal rather than correlational.
- Designed and deployed an ML calibration system, enabling seamless hot-swapping of models while preserving both accuracy metrics and score distributions.
- Spearheaded optimizations of ML training pipelines, achieving up to 3× speedups.
- Led redesign of a monolithic prediction service into multi-service architecture, supporting live routing for A/B and blue/green deployments.

Product / Process:

- Built a flexible workflow automation product, allowing customers to define custom rule-based actions around ML predictions.
- Delivered a real-time reporting system, providing up-to-the-moment aggregates on customer usage.
- Led engineering contributions to compliance efforts: full-fleet encryption-at-rest and GDPR-compliant data deletion architecture.

Organizational Initiatives:

- Created and organized a rotation system for paying down technical debt.
- Piloted a project-based team structure for cross-disciplinary initiatives.
- Founding member of the internal Ethics Committee.

2013: Software Engineer, Prismatic (San Francisco, CA)

- Built and improved backend services including topic modeling, document lifecycle, and social media integrations.

2011 - 2013: Software Engineer, Etsy (New York, NY)

- Improved product search ranking via query-specific ML models trained on click and purchase data; introduced weighted stratification to enhance result diversity.
- Built distributed data processing jobs for query/clickstream data; led major migration away from EMR.

2008 - 2010: Software Engineer, A9 (Palo Alto, CA)

- Conducted clickstream analysis and search analytics using Hadoop.
- Developed Bayesian/graphical models of user attention in search sessions.

Education

2005 - 2008 **BA, Cognitive Science**, UC Berkeley
Graduated with Honors, Departmental Citation

Technologies

In production, I have used (descending order of proficiency):

- **Languages:** Java, Python, Scala, Clojure, JavaScript, PHP
- **Deep Learning:** PyTorch, HuggingFace Transformers, PEFT
- **Distributed Processing:** Hadoop, Spark, Beam/Dataflow
- **Databases:** HBase, Mongo, Postgres, BigTable, Athena, BigQuery
- **Cloud Platforms:** AWS, GCP (broad service experience)
- **Infrastructure / Orchestration:** Kubernetes, Terraform, Salt
- **Other Tools:** ClearML, Kafka, SQS, Protobuf, Avro, ElasticSearch, Airflow

Patents & Publications

- Patent: **US10339472B2**
- Conference Paper: **Iterated Learning and the Cultural Ratchet**. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pp. 2089–2094.