Aaron Beppu

I'm an experienced engineer with interests in machine
learning, data mining, and real-time systems.
If you're interested in employing me **tell me about your open position**.

Experience

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

I have been a key contributor within engineering during a period of continued and transformative growth. During that period I have driven many of the projects underpinning that transformation. Here's a sample:

Data hacking:

- Measuring and analysing ML bias without user-level ground-truth about sensitive group membership.
- A model calibration system which supports switching between ML models without disturbing the behavior of downstream systems which consume predictions. Simultaneously preserves specific model accuracy metrics while also matching marginal score distributions.
- Online accuracy metrics for payment fraud. Designed analysis and implemented pipeline and reporting for tracking accuracy, in the presence of incomplete and delayed ground truth data.

ML tooling and platform:

- Led efforts to speed up our production training pipelines while preserving semantics. Achieved up to 3x speedup.
- Redesign of monolithic scoring service, to multiple services supporting different model versions. Supports "instant" changes to model serving based on live configuration and routing system.

Product:

- A flexible workflow automation product. Customers can describe "if X then Y else Z" conditions and automations which have access to our ML features and predictions.
- A real-time reporting product. Customers can see up-to-the-moment reports and aggregates describing their use of our product.

Platform/Infra:

- Several large, cross-cutting migrations: Zero-downtime migration between incompatible versions of HBase. Migration between distributed queue systems (SQS to Kafka). Encrypted all data at rest; involved replacing every instance in our fleet. Key contributor in inter-cloud migration (AWS to GCP)

Org Hacking:

- Led a rotation-based system to pay down technical debt.
- Organized and launched a trial of “project-based teams” – temporary, inter-disciplinary teams scoped to key projects.
- Ethics Committee founding member.

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

Built and improved a range of backend services, including topic modeling, document life-cycle, and social media integrations.

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

- Data-mining system to improve product search ranking: Query-specific models based on click and purchase data. Weighted stratification of results from models to provide result-page diversity. (e.g. see my **Hadoop World 2011 presentation**)
- Big data tools and infrastructure:
  - Migrating from EMR to a Hadoop cluster on our own hardware, including a large codebase of legacy jobs
  - Breaking performance bottlenecks in our ongoing processing
  - Tooling for creating, scheduling and running workflows of jobs
  - Tooling for monitoring and error-recovery

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

- Clickstream analysis and search analytics, using Hadoop.
- Bayesian/graphical model of user attention during search sessions.

Education

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

Public communications

Most of my work is not for public audiences. However, here are some published artifacts:

Patent **US10339472B2**

Technical blog **Models in Disguise: How Sift Science Ships Non-Disruptive Model Changes**

Public speaking **Non-disruptive Model Changes**

Conference Paper **Iterated learning and the cultural ratchet.** In N. A. Taatgen & H. van Rijn (eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. pp. 2089–2094.

Technologies

Languages I have used in production: (descending order of proficiency) Java, Scala, Python, Clojure, JavaScript, PHP

Libraries/frameworks/services I have used in production: - distributed processing: Hadoop, Spark, Beam/Dataflow - DBs: Bigtable, HBase, Mongo, Postgres, Athena, BigQuery - cloud: many

AWS and GCP services - infrastructure config: Terraform, Salt - other: Kafka, SQS, Protobuf, Avro, Elastic Search, Airflow

Non-production technical interests: disciplined convex programming, logical/relational programming, programming language theory