

Edward Gan

Email : edgan8@gmail.com

Web : edgan8.github.io

Software engineer with a research background passionate about data management and monitoring for machine learning.

EXPERIENCE

Waymo

New York, NY

Senior Software Engineer

11/2022 – Present

- **Agent Prediction Model Evaluation:** I launched a quality evaluation system for agent prediction models. Using tailored metrics over data-mined scenarios, we were able to catch regressions and enable the rollout of causal agent predictions.
- **Data Scaling:** I stabilized the training data distribution for our agent prediction models to improve cache re-use and cut storage costs.

Databricks

San Francisco, CA

Senior Software Engineer

06/2020 – 08/2022

- **Model Monitoring:** I led development of the metrics platform for our model monitoring product, working with customers to ensure the public API was both efficient and customizable.
- **ML Platform Usability:** I grew usage of our ML platform by streamlining navigation UIs, enabling search over experiments, and adding in-notebook data visualization.

Stanford Computer Science, Future Data Systems Group

Stanford, CA

PhD Research

09/2015 – 06/2020

- **High-performance Analytics:** I developed data aggregation systems for high-cardinality monitoring workloads. These were integrated into Apache Druid and internal dashboards at Microsoft.
- **ML Data Management:** I developed methods for optimizing ML compute and data resources, including hyperparameter tuning for transfer learning and sample selection for labeling.

Google Brain

Mountain View, CA

Research Intern

06/2019 – 09/2019

- **Tensorflow Data Validation:** I sped up production Ads model validation 10% by introducing streaming approximations for core statistical primitives.

Facebook

Menlo Park, CA

Software Engineer

08/2013 – 07/2015

- **Data Pipelines:** I developed the company-wide execution engine and UI/API for ETL backfills.

EDUCATION

Stanford University

Stanford, CA

PhD in Computer Science, advised by Peter Bailis

Sep 2015 – June 2020

- **Thesis:** Data summaries for scalable, high-cardinality analytics

Harvard University

Cambridge, MA

A.B. Summa Cum Laude in Computer Science and Mathematics

May 2013

SKILLS

- **Proficient:** Python, Java, C++, SQL, Spark. **Familiar:** Javascript, PyTorch, React

SELECTED PUBLICATIONS

- CoopStore: Optimizing Precomputed Summaries for Aggregation** VLDB
Edward Gan, Peter Bailis, Moses Charikar 2020
- System for optimizing data summaries in high cardinality query engines.
- Approximate Selection with Guarantees using Proxies** VLDB
Daniel Kang, Edward Gan*, Peter Bailis, Tatsunori Hashimoto, Matei Zaharia* 2020
- Sample-efficient methods for calibrating models used for text/video retrieval.
- CrossTrainer: Practical Domain Adaptation with Loss Reweighting** DEEM
Justin Chen, Edward Gan, Kexin Rong, Sahaana Suri, Peter Bailis 2019
- Automatic hyperparameter tuning for transfer learning across datasets.
- DIFF: A Relational Interface for Large-Scale Data Explanation** VLDB
Firas Abuzaïd, Peter Kraft, Sahaana Suri, Edward Gan, . . . , Peter Bailis, Matei Zaharia 2019
- Semantics for a SQL operator to explain differences between datasets.
- Moment-Based Quantile Sketches for . . . Aggregation Queries** MLSys, VLDB
Edward Gan, Jialin Ding, Kai Sheng Tai, Vatsal Sharan, Peter Bailis 2018
- Memory-efficient algorithms for estimating quantiles in distributed systems.
- Scalable Kernel Density Classification via Threshold-Based Pruning** SIGMOD
Edward Gan, Peter Bailis 2017
- Efficient unsupervised, non-parametric outlier classification.
- MacroBase: Prioritizing Attention in Fast Data** SIGMOD
P. Bailis, E. Gan, S. Madden, D. Narayanan, K. Rong, S. Suri 2017
- End-to-end system design for anomaly explanation on multi-dimensional event log data.