Edward Gan

Email : edgan8@gmail.com
Web : edgan8.github.io

Software engineer working at the intersection of ML platforms and research.

Experience

Scale AI New York, NY
Staff Software Engineer *01/2024 – Present*

- **Training Data Quality Platform**: I developed a platform to scale up data quality experiments by post-training models and running domain-specific LLM benchmarks. We identified multiple strategic data quality issues and the platform was used by dozens of researchers. (Blog post).
- **Public LLM Leaderboard Infra**: I standardized LLM eval execution infra and tracing across the research team, allowing us to keep our public leaderboards up-to-date with day-1 results and analyses.

Waymo New York, NY
Senior Software Engineer *11/2022 – 01/2024*

- **Agent Prediction Model Evaluation**: I launched an eval pipeline for multi-agent interactions on data-mined scenarios. This informed model fixes during the development 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 definition API and compute engine for the model monitoring product.
- **ML Platform Usability**: I grew usage of our ML platform by re-designing the navbar, enabling experiments search, and adding in-notebook data exploration.

Google Brain Mountain View, CA
Research Intern *06/2019 – 09/2019*

- **TFX Data Validation**: Cut the ads data validation runtime by 10% with streaming statistics.

Facebook Menlo Park, CA
Software Engineer *08/2013 – 07/2015*

- **Data Pipelines**: 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, SQL. **Familiar**: Javascript, C++, Spark, PyTorch

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.

DIFF: A Relational Interface for Large-Scale Data Explanation VLDB
*Firas Abuzaid, Peter Kraft, Sahaana Suri, **Edward Gan**, ..., Peter Bailis, Matei Zaharia* *2019*

– SQL operator for explaining 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 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 for explaining anomalies on multi-dimensional event log data.