Edward Gan

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

Software engineer and researcher working at the intersection of data processing and machine learning.

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*

Experience

Databricks — San Francisco, CA
Senior Software Engineer — *June 2020 – Present*

- **Model Monitoring**: Design and implementation of the statistical analyses (model drift, custom model quality metrics, etc..) in our ML model-monitoring platform. Collaborated with internal users to form requirements and led development for 3 engineers.
- **Data Profiling**: Developed optimized data summarization routines in Spark and integrated them with the Databricks Notebook, enabling in-UI exploration for thousands of users weekly.

Google Brain — Mountain View, CA
Research Intern — *June 2019 – September 2019*

- **Tensorflow Extended (TFX) Data Validation**: Added streaming data processing algorithms to the TFX data validation pipeline to speed up end to end processing by 10%, and evaluated methods for automatic feature engineering.

Airbnb — San Francisco, CA
Engineering Intern — *June 2016 – September 2016*

- **ML Price Recommendation**: Refactored price suggestion model to output calibrated scores for marketing up-sells, achieving a higher conversion rate than existing marketing e-mails.

Facebook — Menlo Park, CA
Software Engineer — *Aug 2013 – July 2015*

- **Data Pipelines**: Developed Python API, scheduling logic, and UX for backfilling ETL pipelines on-demand across the company.

Selected Publications

Approximate Selection with Guarantees using Proxies — VLDB
Daniel Kang, **Edward Gan***, Peter Bailis, Tatsunori Hashimoto, Matei Zaharia* — *2020*

- Statistically-efficient algorithms for data labeling when using ML models for text/video retrieval.

CrossTrainer: Practical Domain Adaptation with Loss Reweighting — SIGMOD DEEM
*Justin Chen, **Edward Gan**, Kexin Rong, Sahaana Suri, Peter Bailis* — *2019*

- Robust & efficient techniques for automatic transfer learning across datasets.

Moment-Based Quantile Sketches for ... Aggregation Queries — SysML, VLDB
***Edward Gan**, Jialin Ding, Kai Sheng Tai, Vatsal Sharan, Peter Bailis* — *2018*

- Distributed quantile estimation using a maximum entropy model, incorporated into Apache Druid.

Skills and Awards

- **Languages**: Python, Java, SQL, Spark, PyTorch
- NSF Graduate Research Fellowship 2015-2020