Edward Gan

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

Software engineer with a research background working at the intersection of data and ML.

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**: I was responsible for design and implementation of statistical analyses in our model monitoring product including: model and data drift, model accuracy measures, slicing and grouping, custom user-defined metrics, etc. This required understanding user workflows to set up a layered architecture and planning development work for the team.
- **Data Profiling**: Data profiling provides users an integrated interface for exploring datasets. I implemented the backend, which includes a custom Spark aggregator for categorical data, and worked cross-functionally with design and security to launch this across all Notebooks at Databricks.
- **ML Studio**: As part of the release of ML Studio, I led the engineering work to make ML features more accessible. I developed improved navigation flows, a new experiments browser, and drove a documentation overhaul for the ML platform.

Stanford Computer Science, DAWN Lab Stanford, CA
PhD Research *Sep 2015 – June 2020*

- **Data summaries for scalable analytics**: I proposed a system architecture and new statistical methods for analytics on pre-aggregated summaries, targeting use cases at Microsoft and Imply.
- **ML Data Management**: I developed hyperparameter tuning methods for incorporating data from different domains, as well as sampling techniques for labeling expensive video datasets.
- **MacroBase**: MacroBase is a system for explaining shifts in data streams. I contributed optimized feature selection routines and deployed the system with internal cloud monitoring at Microsoft.

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

- **Tensorflow Extended (TFX)**: TFX is a platform for training and deploying ML models. I implemented C++ streaming operators 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**: I refactored an existing price suggestion model to output calibrated scores for marketing up-sells, improving our e-mail conversion rate.

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

- **Data Pipelines**: I developed Python APIs, scheduling logic, and UX to improve the usability of ad-hoc ETL backfills on the Airflow-like company data workflow platform.

Selected Publications

Approximate Selection with Guarantees using Proxies VLDB

Daniel Kang, **Edward Gan****, *Peter Bailis, Tatsunori Hashimoto, Matei Zaharia* *2020*

– Statistically-efficient methods for data labeling in ML models used for text/video retrieval.

Moment-Based Quantile Sketches for . . . Aggregation Queries MLSys, VLDB

Edward Gan, *Jialin Ding, Kai Sheng Tai, Vatsal Sharan, Peter Bailis* *2018*

– Distributed quantile estimation using a maximum entropy model.

MacroBase: Prioritizing Attention in Fast Data SIGMOD

P. Bailis, ***E. Gan****, S. Madden, D. Narayanan, K. Rong, S. Suri* *2017*

– Anomaly detection and feature selection on multi-dimensional event log data.

Skills

Proficient with Python, Java, SQL, Spark, PyTorch. Familiar with Javascript, Scala, C++.