Edward Gan

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

Software engineer and researcher working at the intersection of data processing and machine learning. Experienced with developing both algorithms and systems to address data science needs and then translating them into product impact.

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**: The model monitoring product tracks model predictions and model quality for customers. I was responsible for design and implementation of the statistical analyses 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 in Notebooks provides users an integrated interface for exploring BI and ML datasets. I developed much of the core statistics logic including a custom optimized Spark operator for aggregating 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 with quick start guides 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 for data analytics using on pre-aggregated summaries that targets use cases at Microsoft and Imply. This required new statistical methods for high-cardinality roll-ups over data summaries.
- **ML Data Management**: A recurring challenge in ML is limited training data. I developed hyperparameter tuning methods for incorporating data from different domains, as well as data sampling techniques for labeling expensive video or text examples.
- **MacroBase**: MacroBase is a system for explaining shifts in data streams. I contributed optimized feature selection routines and integrated 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 replaced Python aggregations in the data validation phase with C++ streaming operations to speed up end to end processing by 10%. I also evaluated methods for automatic feature engineering through conditional numeric binning.

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

– **ML Price Recommendation**: The price suggestion model captures opportunities for host revenue, but was only used in one part of the product. I refactored the 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

CoopStore: Optimizing Precomputed Summaries for Aggregation VLDB
***Edward Gan**, Peter Bailis, Moses Charikar* *2020*

– Algorithms for efficiently pre-aggregating summaries in high cardinality query engines.

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.

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

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

DIFF: A Relational Interface for Large-Scale Data Explanation VLDB
*Firas Abuzaid, 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*

– Distributed quantile estimation using a maximum entropy model.

Scalable Kernel Density Classification via Threshold-Based Pruning SIGMOD
***Edward Gan**, Peter Bailis* *2017*

– Unsupervised, non-parametric outlier classification, outperforming scikit-learn.

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 and Awards

- **Languages**: Proficient with Python, Java, SQL, Spark, PyTorch. Familiar with Javascript, Scala, C++.
- **Awards**: NSF Graduate Research Fellowship 2015-2020