Edward Gan

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

Researcher and engineer interested in developing novel algorithms and systems for ML data management.

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 Platform**: I led development of the statistical analyses in model monitoring including data quality, data drift, and model accuracy measures. This required designing a novel staged architecture for allowing both optimized computation and user customization.
- **Data Profiling**: I worked on the Spark backend for the Databricks notebook data exploration tool, which includes a custom aggregator for high-cardinality categorical data, and worked with design and security to launch this across all notebooks.
- **ML Platform Usability**: I led an effort to improve ML platform usability by implementing improved navigation flows, an experiments browser, and driving a documentation overhaul.

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

- **Data summaries for scalable analytics**: I proposed a novel system architecture and statistical methods for analytics on small pre-computed summaries. These methods allow for interactive queries on large, high-cardinality datasets inspired by cloud monitoring needs.
- **ML Data Management**: I developed hyperparameter tuning methods for transfer learning across different domains, as well as sampling techniques for labeling expensive video data. These allow model training to work efficiently with very limited training data.
- **MacroBase**: MacroBase is a system for explaining shifts in data streams. I contributed optimized feature selection routines and deployed the system with collaborators at Microsoft.

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

- **Tensorflow Extended (TFX)**: I implemented C++ streaming operators to speed up end to end processing of data validation by 10% and evaluated methods for automatic feature engineering.

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

- **ML Price Recommendation**: I developed the backend for an e-mail upsell campaign by extracting calibrated scores from the booking prediction model, outperforming existing similar campaigns.

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

– **Data Pipelines**: I developed Python APIs, scheduling logic, and UX to allow users to customize and execute ETL backfills on-demand, enabling more iterative workflows.

Selected Publications

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

– Algorithms for pre-computing efficient 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 model calibration in 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 a general method of 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