# Research Statement

Yihan Gao (gaoyihan@gmail.com)

October 28, 2018

My primary research interests are in the field of machine learning, database systems and data mining, and most of my PhD research works focus on using machine learning techniques to solve standard database problems. More specifically, I try to utilize the statistical properties of datasets to achieve variable but generally better performance (depending on the specific dataset). Although algorithms using statistical properties of datasets can be hard to analyze theoretically, they can perform extremely well if the input dataset is well-structured statistically. The following two projects are most representative of my research style:

(a) **Tabular Dataset Compression:** in [4], we studied the problem of compressing tabular datasets using their statistical properties (e.g., highly correlated columns, or skewed attribute value distribution). Generally speaking, datasets with nice statistical properties can be efficiently compressed using tricks such as joint-encoding of multiple columns or non-uniform length coding scheme. Generalizing this basic idea, we developed an end-to-end compression algorithm, using Bayesian Networks as the statistical model and Arithmetic Coding as the encoding scheme. Due to the flexibility of Bayesian Networks, our algorithm can capture a wide variety of statistical structures, and achieved significantly better result compared to existing methods.

(b) **Log Dataset Structure Discovery:** in [1], we studied the problem of converting semi-structured log datasets into a structured relational format without human supervision. The potential high complexity of log datasets makes it very challenging to automatically separate the actual records from noise and identify their formatting. Our algorithm Datamaran, which is based on the Minimum Description Length (MDL) Principle, does not require users to manually identify the record boundaries, which makes it much more applicable to the noisy log files that are ubiquitous in practice.

**Side Project: Crowdsourcing Pricing**

Aside from the works done in traditional database research, I also have some experience in the field of crowdsourcing. In [5], we tackled the problem of dynamically adjusting the task prices in marketplaces (e.g., Amazon Mechanical Turk) to guarantee that the entire batch can be finished on time. We developed a mathematical model for human behaviors in crowdsourcing marketplaces, and designed a pricing algorithm using Markov Decision Process.

## Deep Integration between Machine Learning and Databases

Currently, I'm actively exploring the potential deep integrations between machine learning techniques and database systems. In particular, I'm designing a database system with integrated ML functionalities but hides away the modeling details: similar to how traditional database systems hide away the details of physical storage and indexing from users, the "intelligent" layer of database systems could be designed in such a way that allow users to concisely depict their inference intention without worrying about the details of ML methodology. So far, we have published a few preliminary results that are related to this envisioned system:

(a) **Interpretability of Agnostic ML:** in [3], we investigated the interpretability of conditional probability estimates under the agnostic setting: how do we interpret the output of conditional probability estimators when their model assumptions are not consistent with the input dataset (e.g., logistic regression model on datasets with non-linear distribution). Our results show that even in such cases, the output can still have nice interpretations if they are *calibrated*. This calibration property can be obtained by calibrating these outputs in a post-processing procedure, which then allows us to use these results as if they are real conditional probabilities for most practical purposes.

(b) **Preliminary Investigation of Relational Representation Learning:** in [2], we examined the major factors that affect the performance of representation learning methods in graphs, a simplified setting of our relational learning scenario. Our results show that for these methods, restricting the norm of embedding vector is of crucial importance, while the choice of embedding dimension is less relevant. These results provide us with insight on how to apply similar techniques for relational datasets in our envisioned system.

### Future Research Plan

With the above preliminary investigations done, the next step is to tackle the main challenge of designing automated inference algorithm for relational datasets. Currently, there are three major approaches for handling relational learning tasks: (a) using join operations to concatenate the features of entities (based on their foreign-key references) and use them in standard off-the-shelf ML methods; (b) using probabilistic graphical models to characterize the joint distribution of attributes within the dataset, with template techniques (e.g., Markov Logic Network) to handle the arbitrary connections between entities; (c) using representation learning techniques to learn the embedding vectors of entities, which are then used as input in standard ML methods. Among these approaches, methods using representation learning techniques have reported the most promising results in literature, but a thorough investigation is still needed to understand the pros and cons of all these approaches in our scenario. Additionally, most existing applications require heavy human supervision, and our requirement of fully automated deployment may give rise to new technical challenges.

Once we finish designing the inference algorithm, the next step is to integrate it into standard database systems. Basically, we need to modify existing SQL query language to incorporate the new ML component, and also figure out how existing components (e.g., query optimization, indexing, view materialization) should adapt to the new ML workload. Clearly, there are many new optimization opportunities in such system, but implementing a basic functional prototype system should not be too difficult. From there, we can then start to improve its efficiency/stability or add new functionalities.

## References


[1] Yihan Gao, Silu Huang, and Aditya Parameswaran. "Navigating the Data Lake with Datamaran: Automatically Extracting Structure from Log Datasets". In: *ACM SIGMOD International Conference on Management of Data*. 2018.

[2] Yihan Gao, Chao Zhang, Jian Peng, and Aditya Parameswaran. "The Importance of Norm Regularization in Linear Graph Embedding: Theoretical Analysis and Empirical Demonstration". In: *CoRR* abs/1802.03560 (2018). arXiv: 1802.03560. URL: http://arxiv.org/abs/1802.03560.

[3] Yihan Gao, Aditya Parameswaran, and Jian Peng. "On the interpretability of conditional probability estimates in the agnostic setting". In: *Electron. J. Statist.* 11.2 (2017), pp. 5198–5231. URL: https://doi.org/10.1214/17-EJS1376SI.

[4] Yihan Gao and Aditya Parameswaran. "Squish: Near-Optimal Compression for Archival of Relational Datasets". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2016, pp. 1575–1584.

[5] Yihan Gao and Aditya Parameswaran. "Finish them!: Pricing algorithms for human computation". In: *Proceedings of the VLDB Endowment* 7.14 (2014), pp. 1965–1976.