Research Statement - Adam Dziedzic

I investigate machine learning, databases, security, and privacy. The overarching goal of my research endeavors is to cater to the needs of platforms that facilitate collaboration between its participants to help them analyze data across databases and train machine learning (ML) models on multiple datasets. In many contexts, such as healthcare and finance, separate parties may wish to collaborate and learn from each other's data but are prevented from doing so due to privacy regulations, such as Canadian PIPEDA and European Union's GDPR, or system constraints, where each party may use different model architecture or data storage engine. The core components of collaborative platforms are federated databases for data analysis and processing, and federated learning for training and serving of ML models while preserving data privacy and confidentiality.

The first step in ML pipelines is **data processing**. My work enables efficient data analysis, transformation, and migration [3, 6, 7, 8, 9]. I was the main contributor to the BigDAWG federated polystore engine to enable collaboration within an organization by processing data across diverse database systems [14, 18, 19]. BigDAWG executes queries using different data models. For example, it enables simultaneous access to time-series data stored in an array database, text data contained in a key-value store, and tabular data stored in a relational database. I also extended relational operators for graceful degradation and the avoidance of performance cliffs in F1 Query, which is a highly parallel and scalable federated query processing platform at Google that serves the needs of a large number of users and systems by processing 40 billion queries daily. Next, I helped to build a recommendation system for hybrid physical designs that caters to mixed of transactional and analytical workloads [10] in the SQL Server database. The goal behind these efforts is to provide unified, robust, self-tuning, and self-administering database services.

The next step in ML pipelines, after the data is prepared, is **training and inference** of ML models. To increase their performance and enable better resource management in the shared cloud environment [11, 17] as well as decrease vulnerabilities inherent in ML platforms [5], I built band-limited convolutional neural networks (CNNs). Convolutions are fundamental signal processing operations that amplify certain frequencies of the input and attenuate others. My results suggest that neural networks exhibit a spectral bias and ultimately learn filters with a strong inclination towards lower frequencies. From the systems perspective, this gives us a knob to tune the resource utilization, namely, GPU memory and computation, as a function of how much of the high-frequency spectrum we choose to represent.

In many settings, such as healthcare or finance, sensitive data are stored in databases and used to train machine learning models. These use cases require **data confidentiality and privacy**. To this end, my research focused on new privacy-preserving mechanisms for multi-label classification, which is a prevalent task in the healthcare setting but understudied in the ML literature. As an example, an X-ray image can be labeled with one or more pathologies whereas the standard classification allows us to assign only a single label per input [22]. The new primitive, private multi-winner voting, allows revealing multi-label outputs while satisfying bounded differential privacy guarantees, which is the gold standard for rigorous privacy guarantees.

Finally, my research on the **CaPC - Confidential and Private Collaborative learning framework** [2] integrates these three building blocks. CaPC targets medical and finance domains with hospitals or banks acting as collaborating parties. CaPC is the first method that provides confidentiality, by operating on encrypted data during the inference time, as well as differential privacy of the training data for all the collaborating parties. In contrast, standard federated learning (FL) supports only confidentiality since the shared gradients, parameters, or other forms of model updates, still, leak private information. In fact, recent work shows how an active adversary can easily reconstruct private data points of individual users during training in FL [1].

The highlights of my research include building a polystore system to enable data analytics over diverse data models, analysis of neural network models in the Fourier domain that resulted in acceleration of their training and inference, as well as creating the first collaborative learning framework that provides strong privacy guarantees. In the future, I would like to continue this line of work and solve problems that emerge in the collaborative platforms, specifically creating defenses against model extraction attacks, providing robustness to out-of-distribution examples, and enabling privacy protection for various ML tasks, such as multilabel classification.

Data Transformation, Migration, and Indexing

Highlights. *I was the main contributor to the BigDAWG polystore system, which enables users to analyze data across diverse databases. This is especially useful for large organizations that manage many data engines and want to analyze their data holistically. The main challenge was to design transformations between databases that support diverse data models and processing types (streaming, transactional, and batch-oriented). Furthermore, large amounts of data require indexing, which is a complex task. I helped to build a solution that automatically recommends physical designs and targets cloud-based databases.*

Summary. A plethora of new database systems have been created in recent years. This allows customers to adjust the choice of their database to a specific need. However, many companies end up with data siloed in different systems. My goal was to address this issue and construct a data-processing engine that could span many databases and enable data analysis across different data silos. I designed new tools to enable and accelerate data migration between diverse databases such as: relational to store structure data, array-based for scientists to analyze numerical data in the form of multi-dimensional arrays, time-series for detection of temporal trends, and columnar databases dedicated to large-scale data analysis. I built a fast and reliable data exchange hub that became a backbone of the BigDAWG polystore system [20].

Once the data is loaded to a given database, users want to access it easily and in a timely manner, which is achieved using common indexing techniques such as B+trees or Columnstores [16]. The primary goal is to automatically recommend indexes that could cater to diverse workloads and allow reducing the total cost of ownership so that no human intervention is required to maintain a database. To this end, I carried out a detailed analysis of many index structures and proposed how to apply them to specific workloads. My study was focused on Hybrid physical designs that combine B+trees and Columnstore indexes. I precisely characterized the index structures to optimize their selection for a given workload and showed that Hybrid indexes can provide synergy and orders of magnitude benefits for mixed workloads that contain many write-intensive transactions as well as read-heavy decision support queries.

Impact. BigDAWG became a reference architecture in the database community for a polystore system. The task of selecting indexes is a multi-dimensional optimization problem that is daunting even for an expert DBA (Data Base Administrator). I helped to extend Database Engine Tuning Advisor (DTA), a physical design tuning tool for SQL Server, to analyze and recommend both B+ tree and Columnstore indexes when suitable for a given workload. The work was released as part of the Community Technology Preview (CTP) release of Microsoft SQL Server 2017, which is ranked as the 3rd most popular database system according to the DB-Engines Ranking.

Band-limited Training and Inference

Highlights. *In general, a convolutional filter applies to the entire frequency spectrum of the input data. However, natural data, such as images or time series, contain the most information in the lower frequencies. Based on this insight, I created a new form of a neural network called a band-limited convolutional neural network [4, 11] that artificially constrains the frequency spectra of convolutional filters and data during training and inference. The band-limited models provide an important gain in compression (e.g., 50%) with very little compromise in the model performance (up to 1.5% drop in accuracy), and allow us to effectively control the resource usage (GPU and memory). This method requires no modification to existing training algorithms or neural network architectures, unlike other compression schemes.*

Summary. I leveraged the Fast Fourier Transformation (FFT) for Convolutional Neural Networks (CNNs). The Fourier domain allows us to analyze data in the frequency spectrum and FFT enables fast transformation from the time or spatial domain to the frequency domain. The main insight was that neural networks act similarly to humans with respect to frequency analysis. The human eye is not sensitive to high frequencies that usually represent noise, so we can remove them to denoise data and reduce the size of the representations.

The design of band-limited models started from the theoretical foundations in the Fourier domain and led to the improvement of the convolution algorithms as well as their efficient implementation on GPUs using CUDA. I designed and implemented a new FFT-based convolutional operation that selectively constrains the Fourier spectrum utilized during both forward and backward passes in the neural network. For the compression, I leveraged new insights about exploiting the conjugate symmetry of 2D FFT and applied many techniques to improve the efficiency of FFT-based convolution: fast multiplication of complex numbers, reuse of the frequency maps (computed for the forward pass and used in the backward pass), operator fusion, an extension of input and filter sizes for faster FFT transformations.

The experimental results over CNN training for time-series and image classification tasks led to several interesting findings. First, band-limited models lead to more explainable AI. I showed that during training, neural networks start learning from the low-frequency coefficients, and as the training progresses higher frequency coefficients are leveraged. Second, the amount of compression used during training should match the amount of compression used during inference to avoid significant losses in accuracy. This provides practitioners with guidance in how to substantially reduce model memory usage and ensure high-quality outputs. Third, coefficient-based compression schemes (that discard a fixed number of Fourier coefficients) are more effective than ones that adaptively prune the frequency spectra (discard a fixed fraction of Fourier-domain mass). Finally, the test accuracy of the band-limited models gracefully degrades as a function of the compression rate.

On the systems level, my band-limited method with 50% compression in the frequency domain results in only a 1.5% drop in accuracy for the ResNet-18 model trained on CIFAR-10 data while reducing the GPU memory usage by 40% and the computation time by 30% in comparison to the full-spectra counterparts. Additionally, band-limiting provides a new perspective on adversarial robustness. Attacks on neural networks tend to involve high-frequency perturbations of input data. My experiments show that band-limited training produces models that can successfully remove high-frequency noise produced by some adversaries.

Impact. Overall, band-limiting provides two main benefits: (1) explainable AI through the analysis of neural networks in the frequency domain, and (2) accelerated execution and reduction in storage requirements of the convolution operation. The concept was further developed in terms of theoretical analysis of neural networks from the Fourier perspective, especially in terms of robustness to adversarial examples. The band-limited models were applied to one of the 5G problems that involve a fair spectrum sharing [12, 13]. Currently, I work on an extension of the approach to handle more use cases of WiFi setups.

Private Collaborative Learning

Highlights. *Collaborative Machine Learning enables multiple parties to learn from each other by exchanging predictions between their machine learning models. In practice, such a collaboration operates on sensitive data, where the inputs for predictions have to be kept confidential while the outputs should not breach the privacy of data used to train the models. In this context, I led a multi-institutional effort that introduced the first generic framework for collaborative learning which preserves privacy and confidentiality. Unlike prior work, the framework is agnostic with respect to the learning technique and model architecture.*

Summary. Machine learning benefits from large training datasets, which may not always be possible to collect by any single entity, especially when using privacy-sensitive data. In many contexts, such as healthcare and finance, separate parties may wish to collaborate and learn from each other's data but are prevented from doing so due to privacy regulations. Some regulations prevent explicit sharing of data between parties by joining datasets in a central location (confidentiality). Others also limit implicit sharing of data, e.g., through model predictions (privacy). Before this work, there was no method that enabled machine learning in such a setting, where both confidentiality and privacy are preserved, to prevent both explicit and implicit sharing of data. Other approaches to collaboration via ML, for example, Federated Learning (FL) only provides confidentiality, not privacy, since shared model updates contain private information. On the other hand, differentially private learning protects the privacy of training data but assumes unreasonably large datasets. Furthermore, both of these learning paradigms produce a central model whose architecture was previously agreed upon by all parties rather than enabling collaborative learning where each party learns and improves their own local model.

I am leading the project on CaPC Learning: Confidential and Private Collaborative Learning [2], which is the first method provably achieving both confidentiality of model inputs and privacy of model outputs in a collaborative setting. CaPC leverages secure multi-party computation (MPC) and homomorphic encryption (HE) in combination with differentially private aggregation. The

CaPC protocol allows participants to collaborate without having to explicitly join their training sets or train a central model. Each party can improve the accuracy and fairness of their model, even in settings when datasets are non-IID and architectures of ML models in different parties are heterogeneous. In CaPC, collaborating parties exchange machine learning model predictions instead of parameters or gradients, which dramatically reduces both the communication overhead and the attack space.

The differentially private aggregation had previously only been studied in the single-label setting. However, applications of machine learning to domains like healthcare often involve multi-label and multi-site data: a patient may have multiple diagnoses made from data located at different institutions. This requires multi-label inference that is both privacy-preserving and collaborative (i.e., supports inference across multiple sites). CaPC introduces the first differentially private approach to multi-label classification in this setting by providing three new privacy-preserving mechanisms for voting: Binary, Tau, and Powerset. Binary voting operates independently per label through composition. Tau voting bounds votes optimally in their Euclidean norm. Powerset voting operates over the entire binary vector by viewing the possible outcomes as a power set. These mechanisms enable privacy-preserving multi-label learning by extending the canonical single-label technique: PATE. The theoretical analysis shows tradeoffs between the methods, for instance, the Powerset voting requires strong correlations between labels to outperform Binary voting. The empirical comparison of the techniques with DPSGD on large real-world healthcare data and standard multi-label benchmarks demonstrate that the new techniques outperform all others in the centralized setting. Finally, the multi-label CaPC with the new privacy-preserving mechanisms can be used to collaboratively improve models in a multi-site (distributed) setting.

Impact. CaPC was a proof-of-concept that opened new possibilities of collaboration via Machine Learning while not sacrificing data privacy. The method sparked an interest in the industry and currently, I collaborate with Intel to apply CaPC to real-world problems. Finally, I co-authored blog posts (e.g., hosted on www.CleverHans.io) that clearly formulate open problems and disseminate research ideas in a format that engages a large audience to encourage contributions towards collaborative ML.

Future Work

I am especially interested in how to build robust ML systems [21] for collaboration. On the single model level, the lack of well-calibrated confidence estimates makes neural networks inadequate in safety-critical domains such as autonomous driving or healthcare. The pre-trained transformers detect out-of-distribution (OOD) examples surprisingly well in comparison to previous methods, such as LSTMs [15]. However, there remains room for future research and one direction is to propose new self-supervised objectives that can enhance model robustness. In the vision domain, my goal is to investigate new approaches to increase the OOD robustness, for example, by analyzing the internal representations in convolutional layers and harnessing statistical tests to provide a well-founded estimation of uncertainty.

From the perspective of a participant in collaborative learning protocols such as CaPC, model stealing is one of the major concerns since many organizations use proprietary algorithms to solve their problems. Differential privacy is used to protect training data in CaPC but it does not prevent model extraction, where an adversary with black-box access but no prior knowledge of an ML model aims at duplicating the functionality of the victim model. Current active defenses against model extraction perturb model outputs, which lowers the quality for legitimate users. Passive defenses try to detect an attack but do not work for attackers with in-distribution data. The reactive defenses against model extraction attacks, such as watermarking, address model extraction after an attack has been completed. Instead, I propose to design pro-active defenses that prevent model stealing *before* it succeeds. One of the methods requires users to complete a proof-of-work (POW) or proof-of-elapsed-time (POET) before they can read the model's predictions. This increases the running time of model extraction without lowering the quality of model outputs. Intuitively, the method shifts the trade-off between the quality of answers and robustness, that was introduced by previous defenses, to a trade-off between the computational cost or elapsed time and robustness to model stealing. The preliminary results show that this pro-active defense can deter attackers by greatly increasing (even up to 100x) the time needed to leverage query access for model extraction. These efforts will lead to creating collaborative platforms while protecting the privacy of data and models.

References

[1] Franziska Boenisch, Adam Dziedzic, Roei Schuster, Ali Shahin Shamsabadi, Ilia Shumailov, and Nicolas Papernot. When the Curious Abandon Honesty: Federated Learning Is Not Private. *arXiv preprint 2112.02918*, 2021.

[2] Christopher A. Choquette-Choo, Natalie Dullerud, Adam Dziedzic, Yunxiang Zhang, Somesh Jha, Nicolas Papernot, and Xiao Wang. CaPC Learning: Confidential and Private Collaborative Learning. *International Conference on Learning Representations (ICLR)*, 2021.

[3] Adam Dziedzic. Data Loading, Transformation and Migration for Database Management Systems. *Master's Thesis, University of Chicago*, 2017.

[4] Adam Dziedzic. Input and Model Compression for Adaptive and Robust Neural Networks. *PhD Thesis, University of Chicago*, 2017.

[5] Adam Dziedzic and Sanjay Krishnan. Analysis of Random Perturbations for Robust Convolutional Neural Networks. *arXiv preprint 2002.03080*, 2020.

[6] Adam Dziedzic and Jan Mulawka. Analysis and comparison of NoSQL databases with an introduction to consistent references in Big Data storage systems. In *Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2014*, volume 9290, page 92902V. International Society for Optics and Photonics, 2014.

[7] Adam Dziedzic, Jennie Duggan, Aaron J. Elmore, Vijay Gadepally, and Michael Stonebraker. BigDAWG: a Polystore for Diverse Interactive Applications. *IEEE Viz Data Systems for Interactive Analysis (DSIA)*, 2015.

[8] Adam Dziedzic, Aaron Elmore, and Michael Stonebraker. Data Transformation and Migration in Polystores. *IEEE High Performance Extreme Computing (HPEC)*, 2016.

[9] Adam Dziedzic, Manos Karpathiotakis, Ioannis Alagiannis, Raja Appuswamy, and Anastasia Ailamaki. DBMS Data Loading: An Analysis on Modern Hardware. *Accelerating Analytics and Data Management Systems (ADMS)*, 2016.

[10] Adam Dziedzic, Jingjing Wang, Sudipto Das, Bolin Ding, Vivek R Narasayya, and Manoj Syamala. Columnstore and B+ tree – Are Hybrid Physical Designs Important? *ACM International Conference on Management of Data (SIGMOD)*, 2018.

[11] Adam Dziedzic, John Paparrizos, Sanjay Krishnan, Aaron Elmore, and Michael Franklin. Band-limited Training and Inference for Convolutional Neural Networks. *International Conference on Machine Learning (ICML)*, 2019.

[12] Adam Dziedzic, Vanlin Sathya, Monisha Ghosh, and Sanjay Krishnan. Machine Learning Based Detection of Multiple Wi-Fi BSSs for LTE-U CSAT. *International Conference on Computing, Networking and Communications (ICNC)*, 2020.

[13] Adam Dziedzic, Vanlin Sathya, Muhammad Rochman, Monisha Ghosh, and Sanjay Krishnan. Machine Learning enabled Spectrum Sharing in Dense LTE-U/Wi-Fi Coexistence Scenarios. *IEEE Open Journal of Vehicular Technology (OJVT)*, 2020.

[14] Vijay Gadepally, Kyle O'Brien, Adam Dziedzic, Aaron Elmore, Jeremy Kepner, Samuel Madden, Tim Mattson, Jennie Rogers, Zuohao She, and Michael Stonebraker. BigDAWG version 0.1. *IEEE High Performance Extreme Computing Conference (HPEC)*, 2017.

[15] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. Pretrained Transformers Improve Out-of-Distribution Robustness. *Association for Computational Linguistics (ACL)*, 2020.

[16] Sanjay Krishnan, Adam Dziedzic, and Aaron Elmore. Deeplens: Towards a Visual Data Management System. *Conference on Innovative Data Systems Research (CIDR)*, 2018.

[17] Sanjay Krishnan, Aaron Elmore, Michael Franklin, John Paparrizos, Zechao Shang, Adam Dziedzic, and Rui Liu. Artificial Intelligence in Resource-Constrained and Shared Environments. *ACM Special Interest Group on Operating Systems (SIGOPS)*, 2019.

[18] Tim Mattson, Vijay Gadepally, Zuohao She, Adam Dziedzic, and Jeff Parkhurst. Demonstrating the BigDAWG Polystore System for Ocean Metagenomics Analysis. *Conference on Innovative Data Systems Research (CIDR)*, 2017.

[19] John Meehan, Stan Zdonik, Shaobo Tian, Yulong Tian, Nesime Tatbul, Adam Dziedzic, and Aaron Elmore. Integrating Real-Time and Batch Processing in a Polystore. *IEEE High Performance Extreme Computing (HPEC)*, 2016.

[20] Kyle O'Brien, Vijay Gadepally, Jennie Duggan, Adam Dziedzic, Aaron Elmore, Jeremy Kepner, Samuel Madden, Tim Mattson, Zuohao She, and Michael Stonebraker. BigDAWG Polystore Release and Demonstration. *arXiv preprint 1701.05799*, 2017.

[21] Adelin Travers, Lorna Licollari, Guanghan Wang, Varun Chandrasekaran, Adam Dziedzic, David Lie, and Nicolas Papernot. On the Exploitability of Audio Machine Learning Pipelines to Surreptitious Adversarial Examples. *arXiv preprint 2108.02010*, 2021.

[22] Arnold Y. L. Wong, Garrett Harada, Remy Lee, Sapan D. Gandhi, Adam Dziedzic, Alejandro Espinoza-Orias, Mohamad Parnianpour, Philip K. Louie, Bryce Basques, Howard An, and Dino Samartzis. Preoperative paraspinal neck muscle characteristics predict early onset adjacent segment degeneration in anterior cervical fusion patients: A machine-learning modeling analysis. *Journal of Orthopaedic Research (JOR)*, 2021.