MO' DATA, MO' PROBLEMS: HOW DATA COMPOSITION COMPROMISES DATA SCALING PROPERTIES IN MACHINE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

The accumulation of data in the machine learning setting is often presented as a panacea to address its many modeling problems—including issues with correctness, robustness, and bias. But when does adding more data help, and when does it hinder progress on desired model outcomes? We model data accumulation from multiple sources and present analysis of two practical strategies that result the addition of more data degrading overall model performance. We then demonstrate empirically on three real-world datasets that adding training data can result in reduced overall accuracy and reduced worst-subgroup performance while introducing further accuracy disparities between subgroups. We use a simple heuristic for determining when the accumulation of more data may worsen the issues the additional data is meant to solve. We conclude with a discussion on considerations for data collection and suggestions for studying data composition in the age of increasingly large models.

1 INTRODUCTION

The accumulation of data (labeled or unlabeled) in machine learning is often touted as the reliable solution to many of its modeling problems. The benefits of more data on performance have been observed across many domains including tabular (Chen et al., 2018), language (Brown et al., 2020), vision (Chen et al., 2020), and multi-modal data (Wang et al., 2021). Beyond accuracy, increasing dataset size has also been shown to improve adversarial robustness (Carmon et al., 2019) and robustness against distribution shift (Miller et al., 2021). Furthermore, when adding more data also improves subgroup representation, group-level disparities in classification can also be reduced (Rolf et al., 2021). However, practically acquiring more data for training involves much more than a naive increase in the number of training samples. In this work, we define this goal of acquiring more training data examples as *data accumulation*. In practical situations, those engineering the system scale sources to supplement an existing training set – thus data accumulation must not only consider dataset size but also how the composition of the accumulated data changes with scale. Even though such considerations are common knowledge in practical machine learning engineering (Shankar et al., 2022), these challenges still remain relatively under-explored theoretically and empirically by the machine learning research community.

Moreover, dataset qualities do not exist in isolation; algorithmic techniques have been developed for addressing dataset limitations in the distribution shift and supervised domain adaptation literature (Kouw & Loog, 2018). However, current works in these areas narrowly focus on model-based interventions to improve model performance and are thus inadequate for precisely characterizing the effects of a broader set of dataset properties on model outcomes. Acknowledging the reality that there is often agency in the design and composition of the training dataset is an opportunity to design downstream model properties through decision-making about the data. In fact, our paper joins a growing line of work focused on data interventions (Gadre et al., 2023; Marion et al., 2023; Compton et al., 2023). Data accumulation is thus a data-oriented alternative perspective to complement current work which remains narrowly anchored to model improvements.

In this paper, we take a pragmatic approach to data accumulation and construct scenarios that more explicitly factor in corresponding changes to data composition (Figure 1). Motivated by statistics

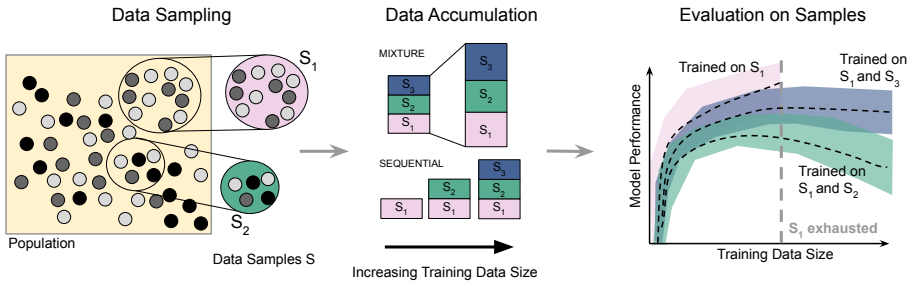

Figure 1: Illustrative pipeline of how we consider the effects of data composition on data scaling properties. We hope to understand cases where the addition of more data in model training leads to a degradation in overall model performance.

literature on the performance penalties incurred by scaling data under sampling bias (Meng, 2018), we explore how this phenomenon may arise in the machine learning setting. We take a principled approach by first formalizing models of data accumulation which gives intuition for why increasing the dataset size may not be sufficient to guarantee better performance. We then test our theoretical models of data composition by examining the effects of increasing training dataset size on real-world tabular datasets for predicting census income, restaurant review sentiment, and medical outcomes.

Our contributions are as follows. In this work, we:

1. **Model realistic case studies of data composition changes in data accumulation:** We present models for data composition changes that occur due to common strategies at unilateral increases to dataset size (ie. scaling up training set size). We motivate and formalize cases of data accumulation from a single-source and multi-source setting.
2. **Analyze data accumulation impacts on downstream performance:** We theoretically demonstrate how data scaling can lead to worse model outcomes and present a simple heuristic to determine when to add more data. We show that under differing sampling regimes, there are scenarios where data accumulation can worsen model performance.
3. **Demonstrate empirical results on the trade-offs between scale and data composition:** We discuss the performance impact of two different practical strategies for data accumulation in the multi-source setting — a sequential data addition case and that of scaling up a mixture of data sources. We illustrate on 3 real-world datasets how the mechanism through which data accumulation occurs impacts model properties.

Most importantly, we hope for this work to be a critical starting point in formalizing the complex dynamics underlying data decision-making as part of the machine learning process. The details of data practices are often overlooked by the machine learning research community altogether, despite its key role in determining the nature of model outcomes (Paullada et al., 2021). We hope this work can be a strong starting point for a deeper investigation by the machine learning community into more principle-based foundations of meaningful data practices.

2 Related Work

Data scaling laws influence model outcomes "Scaling laws" more broadly refer to how increases in model "size" lead to improved performance. Typically, model size is described in terms of the number of model parameters, compute, and other measured factors characterizing the model (Kaplan et al., 2020; Bahri et al., 2021). Data "scaling laws" (Zhang et al., 2020; Bansal et al., 2022; Zhai et al.) specifically reveal the way in which training on larger and larger datasets yields improved performance. Furthermore, adding more data has been suggested as a way to improve the fairness of a model (Chen et al., 2018).

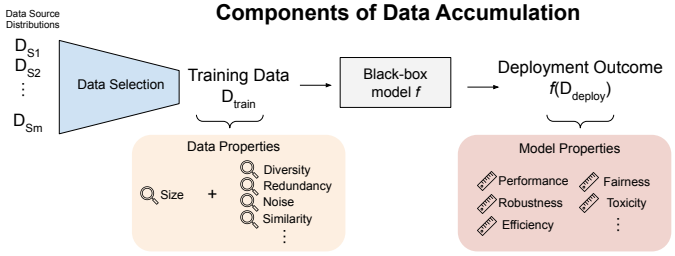

Figure 2: Our work on data accumulation characterizes training data properties interact with scale and how they impacted model outcomes. This study is data-centric and is orthogonal but complimentary to model-based interventions provided in adjacent directions such as domain adaptation.

Data composition influences model outcomes However, the composition of training data, at any size, has been shown to influence model outcomes. Data properties such as data diversity, redundancy, and noise can all contribute to model performance, robustness, fairness, and efficiency Mitchell et al. (2022). These data properties are typically determined by how the data is collected. For example, Rolf et al. (2021) suggests sampling directly from group-specific distributions in order to improve model performance on certain under-represented subgroups.

Prior work in domain adaptation and distribution shift has also considered data composition independent of size (Kouw & Loog, 2018; Gulrajani & Lopez-Paz, 2020; Yang et al., 2023). However, in these works, training data is often considered to be set in stone and imposed as a pre-existing and static constraint. Thus, many of the challenges incurred by data composition are typically addressed with algorithmic interventions rather than data-centric decision-making.

Realistic data scaling impacts data composition When realistically increasing the size of a training dataset, compositional changes in the data can be introduced. Thus, a more pragmatic perspective to data scaling that factors in changes to overall data composition is required – we call this process *data accumulation* (Figure 2). In surveys, sampling bias exacerbates mis-estimation error as the sample size increases, as observed in settings estimating vaccine uptake (Bradley et al., 2021) and election polling (Meng, 2018).

Thus far, relatively few works in machine learning have critically examined the effect of increasing training data size while factoring in the potential changes to data induced by scaling. In the image classification setting, recent work found that performance heavily depends on the pre-training source data (Nguyen et al., 2022) and spurious correlations may be introduced when combining data sources (Compton et al., 2023). Using a theoretical model, Hashimoto (2021) looks at data as a fixed mixture of different sources (e.g., different categories of Amazon reviews) to characterize excess loss as dataset size increases.

3 Two Models for Data Scaling

Much of the past work on the impact of data scaling on model performance assumes a *single-source setting*, where data from a single data sampling process is scaled up, and the distribution of the dataset remains fixed. It is under this setting that many claims on data scaling are typically considered. However, in most practical settings, data accumulation occurs in a *multi-source setting*, where the final training dataset is pieced together from multiple distinct data sampling processes. In order to practically increase the size of the training set, data from multiple sources are collected and combined. Unlike the single source setting, there is not one static scaled-up data collection process – instead the resultant larger dataset is a mixture of multiple data sources, and thus presents more complex data composition changes as the dataset size increases.

In this paper, we specifically consider two practical approaches of this multi-source setting: the SEQUENTIAL case and the MIXTURE case. We consider the following scenarios: 1) enlarging the dataset by sampling from a mixture of fixed sources and 2) sequentially adding data across different sources for model training. The key observation we make in this paper is how overall and subgroup performance can vary as we increase the sample size n in both of these scenarios.

Preliminaries Let $\{x, y\}^n \sim \mathcal{D}$ be data generated from some underlying distribution. Let $D_{S_1}, ..., D_{S_m}$ be a series of empirical distributions sampled with varying types of sampling bias from the underlying distribution $\mathcal{D}$; these empirical source distributions are finite and are different fixed sizes $n_{s_1}, ..., n_{s_m}$. The training distribution of sized n, $D_{train,n}$, is composed of these m sources. The deployment distribution D_{test}, where we hope to achieve good performance, is sampled directly from $\mathcal{D}$ without any sampling bias. We define $\delta : \mathcal{P}(\mathcal{X}, \mathcal{Y}) \times \mathcal{P}(\mathcal{X}, \mathcal{Y}) \to \mathbb{R}_{\geq 0}$ as a divergence between distributions.

Mixture Case (MIXTURE) In the mixture case, $D_{train,n}$ comes from a mixture of sources: $D_{train,n} = \sum_{i=1}^{m} \alpha_i D_{S_i}$. Here, the coefficients $\alpha_i \in [0, 1]$ ($\sum_i \alpha_i = 1$) specify what proportion of data points are sampled from each source. Given a fixed vector α and a dataset size n, $n \times \alpha_i$ data points are included from each distribution D_{s_i}; the ratio of sources is independent of n. Since the ratio of sources is fixed, we also expect the divergence between train and test distributions $\delta(D_{train,n}, D_{test})$ to remain constant as n increases.

Sequential Case (SEQUENTIAL) In the sequential case, the training data is a strictly increasing collection of the underlying sources: $\hat{D}_{train,n} = (\bigcup_{i=1}^{k-1} \hat{D}_{S_i}) \cup \frac{n - \sum_{i=1}^{k-1} n_{s_i}}{n_{S_k}} \hat{D}_{S_k}$[1]. Here, k is set to the source index such that $\sum_{i=1}^{k-1} n_{s_i} < n \leq \sum_{i=1}^{k} n_{s_i}$. In other words, for a desired dataset size n, we start by adding data from the first source and continue to add data from sources sequentially until we reach n. This addresses the common scenario where acquiring more data incurs additional cost; all data from existing sources are used before a new source is introduced. The resulting distribution can also be viewed as a mixture of source distributions where α depends on n: $D_{train,n} = \sum_{i=1}^{k} \alpha_i D_{S_i}$ where $\alpha_i = \frac{n_{s_i}}{n}$ for $i < k$ and $\alpha_m = \frac{n - \sum_{i=1}^{k-1} n_{s_i}}{n_{s_m}}$.

A key observation we make is that $\delta(D_{train,n}, D_{test})$ in the sequential case will actually depend on the number of samples.

Example 3.1. *Consider a training set of two sources: D_{S_1} a small high-quality dataset, D_{S_2} a large lower-quality dataset. We can model the divergence between train and test distributions as follows **if** δ composed linearly:*

$$\delta(D_{train,n}, D_{test}) = \begin{cases} \delta(D_{S_1}, D_{test}) \textit{ if } n \leq |D_{S_1}| \\ \frac{|D_{S_1}|}{n}\delta(D_{S_1}, D_{test}) + (1 - \frac{|D_{S_1}|}{n})\delta(D_{S_2}, D_{test}) \textit{ otherwise} \end{cases}$$

While we cannot assume that divergences compose linearly, we can limit our scope to $f-$divergences and use the convexity of this class of divergences to show that in the SEQUENTIAL case, $\delta(D_{train,n}, D_{test})$ might increase with n.

Lemma 3.2. *Let $D_{train,n}$ be constructed in the* SEQUENTIAL *case from k sources: $D_{S_1}, ..., D_{S_k}$, then if $\delta(D_{S_k}, D_{test}) - \frac{cn}{n_{s_k}} \geq \delta(D_{train,n}, D_{test})$:*

$$\delta(D_{train,n}, D_{test}) \geq \delta(D_{train,n-n_{s_k}}, D_{test})$$

where δ belongs to the family of f-divergences and c is a divergence-dependent constant where $\delta(D_{train,n}, D_{test}) + c = \sum_{i=1}^{m} \frac{n_{s_i}}{n} \delta(D_{S_i}, D_{test})$.

Lemma 3.2 gives a relationship between the new data source and the test set that would cause the divergence between train and test distributions can increase with n in the SEQUENTIAL case[2]. In a fixed dataset size setting, Acuna et al. (2021) relates increased divergence to empirical risk for $f-$divergences in particular by giving a generalization bound. Prior works have also considered different discrepancy measures including L_1 distance (Ben-David et al., 2006), $\mathcal{H}\Delta\mathcal{H}$ divergence (Ben-David et al., 2010), margin disparity discrepancy (Zhang et al., 2019). In conjunction with Lemma 3.2, these results from prior works give an intuition for a larger empirical risk upper bound when the divergence between train and test distributions increases. However, what remains unanswered is whether this bound remains while the training set size increases.

[1] $\hat{D}$ denotes the *set* of examples or data points in the distribution D

[2] See proof in the appendix

DATASET	NUMBER OF ROWS	OUTCOME	SOURCE	SUBGROUP
Folktables (Ding et al., 2021)	1,664,500	Binary Income Level	State	Race
Yelp (Yelp, 2023)	6,990,280	Multi-Class Review Stars	State	Restaurant Category
MIMIC-IV (Johnson et al., 2020)	197,756	Binary Readmission	Admission Type	Race

Table 1: Dataset overview for experiments.

Using the two cases of data accumulation, we can reason about what we might observe empirically even if we do not have an oracle divergence metric. As the training dataset size increases in the MIXTURE case, we would expect $\delta(D_{train,n}, D_{test})$ to remain constant and the training loss to decrease. Thus, we expect the upper bound of the test loss to become tighter as the dataset size grows. However, in the SEQUENTIAL case, if there is a combination of sources that causes $\delta(D_{train,n}, D_{test})$ to grow faster than the decrease in the training set risk, this upper bound becomes looser and we may see an increased test set risk.

3.1 CRITERIA FOR REJECTING MORE DATA

We consider the SEQUENTIAL case where a practitioner may have trained on some data D_{train} to obtain some predictor $f_{D_{train}}$, and encounters another data source D_i. The question is whether it would be beneficial to enlarge the dataset and train on $D_{next} = D_{train} \cup D_i$ in order to best perform on the test distribution that we care about D_{test}. More formally, we are concerned with the excess risk under some proper loss function l from adding more data:

$$L(f_{D_{train}}, f_{D_{next}}, D_{test}) = \mathbb{E}_{(x,y)\sim D_{test}} \left[l(f_{D_{train}}(x), y) - l(f_{D_{next}}(x), y) \right]$$

If $L(f_{D_{train}}, f_{D_{next}}, D_{test}) > 0$, we would want to incorporate the additional data to achieve a lower risk on the test distribution. Otherwise, we would reject the additional data in order to not increase risk on the test distribution or to avoid incurring extra costs for data access and model training while not improving the risk. Practically, finding $f_{D_{next}}$ already requires training a model based on an additional dataset. To avoid this, we suggest a rejection criterion based on access to the existing data mode D_{train}, the additional data D_{next}, and the test distribution D_{test}. While in reality the test set cannot be accessed, we assume we can use some part of the training set (e.g., the D_{s_1}) that is similar to the test distribution). Prior works have suggested that measuring excess Kullbeck-Leibler (KL) divergence between train and test distributions correlates with the resulting loss (Xie et al., 2023b). We also use the excess KL as a heuristic:

$$\Delta_{KL}(D_{train}, D_{next}, D_{test}) = \delta_{KL}(D_{next}||D_{test}) - \delta_{KL}(D_{train}||D_{test}) \quad (1)$$

4 EXPERIMENT SETUP

Datasets For our investigation, we study three real-world tasks and datasets, chosen because of their rich feature sets and the open-access availability. Dataset details can be found in Table 1. See appendix for code, additional experiments and details, and synthetic data experiments.

Models and Evaluation We consider the scenario where the initial dataset of interest comes from a single data source, i.e., Source A, with limited training examples (e.g., South Dakota (SD) in Folktables). We also consider at least one second available data source, i.e., Source B, from which additional training examples can be drawn. The experimental goal is to investigate the effects of manipulating training data composition on model outcomes as measured on a test set sampled exclusively from the initial dataset Source A (e.g., SD) – which we call the *reference test set*. Further experiments are also evaluated on a *generalized test set*, which is randomly sampled from a mixture of all available data sources.

To observe data accumulation in the MIXTURE case, Source A and Source B are sampled at a fixed ratio from a combined dataset of Source A and Source B in order to increase the training set size.

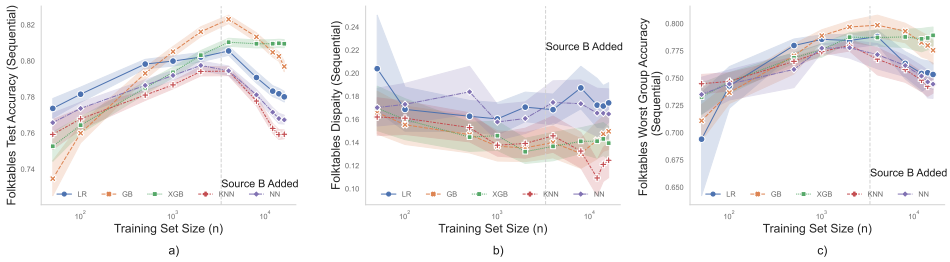

Figure 3: Folktables results on **Source A reference test set** in the SEQUENTIAL case over 5 trials on **(a)** accuracy, **(b)** accuracy disparity, and **(c)** worst subgroup accuracy. Source A from South Dakota and Source B from California is added once South Dakota data has been exhausted.

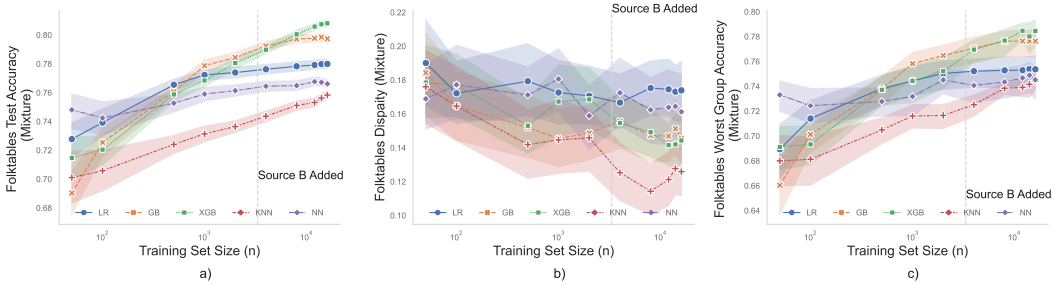

Figure 4: Folktables results on **Source A reference test set** in the MIXTURE case over 5 trials for **(a)** accuracy, **(b)**accuracy disparity, and **(c)** worst subgroup accuracy. The mixture is the same ratio as the final dataset for the SEQUENTIAL case in Figure 3) (75% CA and 25% SD)

In the SEQUENTIAL case, we start by adding training data from Source A and then from Source B when all points from Source A have been included.

We consider a variety of different models: logistic regression (LR), gradient boosting (GB) (Friedman, 2001), k-Nearest Neighbors (kNN), XGBoost (XGB) (Chen & Guestrin, 2016), and MLP Neural Networks (NN). Let f denote the model we are evaluating and let g be a group function that maps each data point to a subgroup, we evaluate the following metrics over D_{test}: **Accuracy**: ($\mathbb{E}_{(x,y)\sim D_{test}}[f(x) = y]$), **Disparity**: The difference between the best and worst-performing subgroups ($\max_{g'} \mathbb{E}_{(x,y)\sim D_{test}}[f(x) = y|g(x) = g'] - \min_{g'} \mathbb{E}_{(x,y)\sim D_{test}}[f(x) = y|g(x) = g']$), and **Worst group accuracy**: The accuracy on the worst-performing subgroup and the metric of interest in studying subpopulation shifts in the distribution shift literature (Koh et al., 2021) ($\min_{g'} \mathbb{E}_{(x,y)\sim D_{test}}[f(x) = y|g(x) = g']$).

5 RESULTS AND ANALYSIS

Single Source Datasets Benefit from Data Scaling Properties We consider the initial stage of the SEQUENTIAL case—prior to sampling from any additional data sources—to be equivalent to a single-source data setting. We find that increasing the dataset size in this single source setting yields improved performance (Figure 3a). Consistently, maximum accuracy is achieved when the most data points in a single source are used. Single source data increases also improve worst-subgroup performance, as demonstrated in prior literature (Sagawa et al., 2019). For some models, increases in data slightly improve disparity (e.g., XGB disparity drops from 17.0% to 13.6%) while for other models even within the single source setting do little to minimize the differences between subgroup disparity[3].

[3]Results with confidence intervals and additional results in larger n for single source scaling are available in the appendix

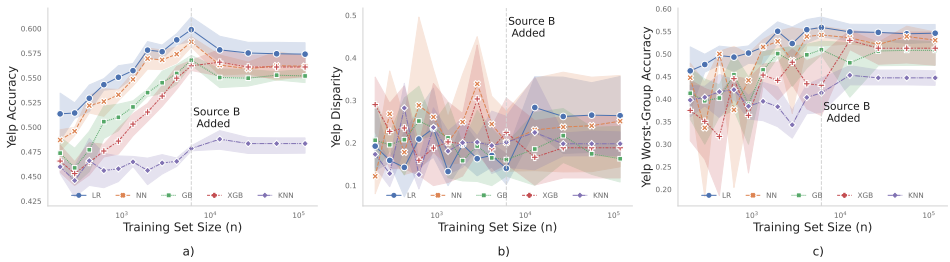

Figure 5: Yelp results on **Source A reference test set** in the SEQUENTIAL case over 5 trials on **(a)** accuracy, **(b)** accuracy disparity, and **(c)** worst subgroup accuracy. Source A is from New Jersey and Source B is from Pennsylvania.

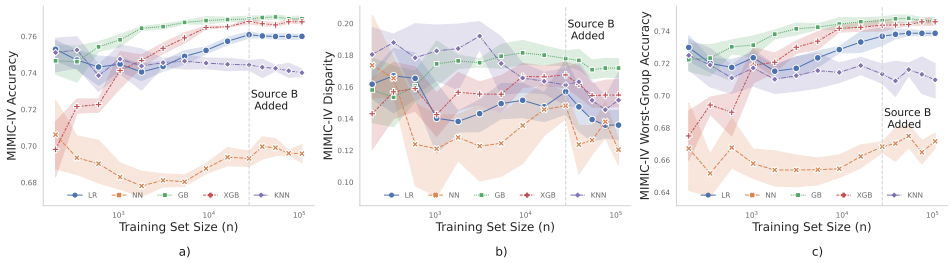

Figure 6: MIMIC-IV results on **Source A reference test set** in the SEQUENTIAL case over 5 trials on **(a)** accuracy, **(b)** accuracy disparity, and **(c)** worst subgroup accuracy. Source A is from admission type URGENT and Source B is from EW. EMER.

Multi-source Dataset Scaling Can Lead to Worse Outcomes on a Reference Test Set In the SEQUENTIAL case (Figure 3a, 5a, 6a), we observe that adding additional data from a separate source, thereby quadrupling the size of the training set, leads to a dip in performance on a reference test set of interest, in addition to worse fairness metrics and worse robustness. For the Folktables dataset (Figure 3a), this dip is observed empirically as a statistically significant reduction of test accuracy for all models except XGB (e.g., LR: -2.5%; GB: -2.6%; kNN: -3.5%, NN: -2.7%). This reduction in performance occurs when additional data is added from source B (n_B = 12000) once source A (n_A = 4000) is exhausted; the training data size has tripled. Decreases to worst subgroup performance were also significant and observed in all models except XGB (e.g., LR: -3.5%; GB: -2.3%; kNN: -2.3%, NN: -2.7%). We did not observe a significant decrease in disparity with the addition of more data.

In the MIXTURE case (Figure 4), we find that scaling a fixed mixture of sources yields monotonically improved performance on the reference test set—i.e., there is no observed dip in performance, and the increase of the dataset size is correlated with increasing test accuracy, and better worst subgroup performance. From n = 4000 to 16000, we see test accuracy consistently improve for all models (e.g., XGB:+1.8%; kNN:+1.5%) and we see worst subgroup performance also consistently improve across all models. As in the SEQUANTIAL case, we do not see a significant impact of data scaling on subgroup disparities due to the high variance in worst and best subgroup size in the test set. However, comparing the MIXTURE case to the SEQUENTIAL over the same range of n before all data is added, the accuracy achieved on the reference dataset for the best n in the SEQUENTIAL case remains consistently above the maximum accuracy in the MIXTURE case. At n_{max} for SEQUENTIAL case, test accuracy is higher (LR: 78.0%, GB: 79.8%, XGB: 80.8%, kNN: 75.8%, NN:76.7%) than n_{max} for the MIXTURE case (LR: 80.6%, GB:82.3%, XGB:81.0%, kNN:79.4%, NN: 79.6%).

Multi-source Dataset Scaling Can Lead to Better Generalization In (Figure 5b, 6b, 7a), we see that increases to the training dataset, even from sources of different distribution, still yield improvements on a generalized test set, sampled from all available sources (e.g., For N = 50/4000/16000, LR: 72.7% / 76.8% / 78.4%; GB: 70.4% / 78.9% / 80.2%; XGB: 72.1% / 79.5% / 80.5%; NN: 70.8% / 77.4% / 77.9%). This indicates that increasing dataset size across sources yields improvements to

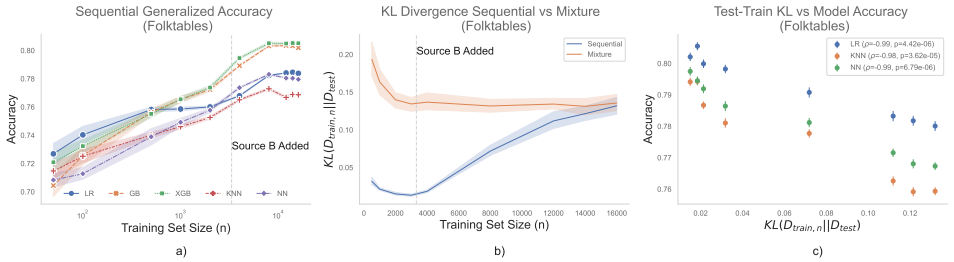

Figure 7: **(a)** Folktables accuracy in the SEQUENTIAL case changes on the **generalized test set**, **(b)** Comparison of Train-Test KL: SEQUENTIAL setting vs MIXTURE setting, **(c)** Relationship between Train-Test KL and Accuracy in the SEQUENTIAL setting for different classifiers

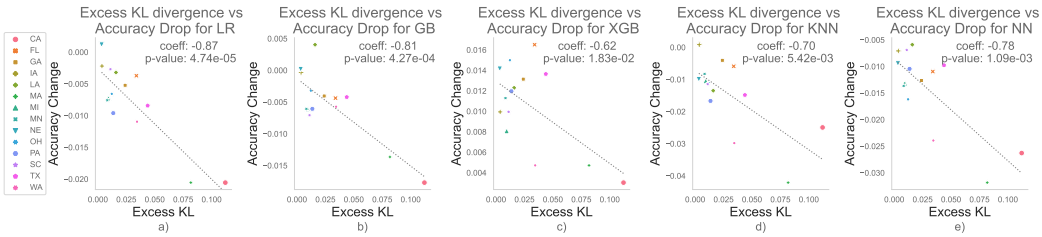

Figure 8: The relationship between Excess KL and the resulting accuracy drop from increasing for **a)** logistic regression (LR), **b)** gradient boosting (GB), **c)** XGBoost (XGB), **d)** K-Nearest Neighbor (kNN) and **e)** Neural Network (NN). We observe a statistically significant correlation between Excess KL (Equation 1) and accuracy drop across all 5 models where we observe significant decreases in model performance.

model generalization, even if this may not translate to improved performance on the reference test set of interest.

5.1 PERFORMANCE VIA THE LENS OF A PRACTICAL DIVERGENCE

In Section 3, we presented two models: SEQUENTIAL and MIXTURE. We will now connect our empirical results to our proposed theoretical models of data accumulations.

Divergence comparison: SEQUENTIAL VS MIXTURE The first step is to empirically validate that our specific choice of divergence, KL divergence, increases in the SEQUENTIAL setting as the training set size grows. We approximate densities through kernel density estimation with a Gaussian kernel on scaled PCA projections (3 components). Figure 7b compares the KL divergence between the training set and test set at different values of n (data size) for the Folktables Income dataset. In the SEQUENTIAL case, scaling up n results in an increase in train-test divergence while in the MIXTURE case, this divergence remains static.

Translating Divergence to Accuracy The next step is to validate that increased train-test divergence translates into a reduction in accuracy. We find a significant negative correlation between the KL divergence between the train and test dataset with the resulting model accuracy for 3 out of 5 models; as train-test divergence increases, test accuracy decreases. Figure 7c shows this correlation for the 3 algorithms where we observe a significant correlation. There was also a negative correlation between for Gradient Boosted Trees (GB). We did not observe a decrease in performance for XGBoost (XGB), thus such a correlation for the XGB model is not expected.

Excess KL and Rejecting More Data Finally, we validate our proposed heuristic of excess KL (Δ_{KL}) for deciding when to include more data (Section 3.1). If Δ_{KL} is larger than 0 there is a significant distribution shift in the larger dataset and there is thus likely to be an

increase or flat-lining of loss. We consider a large set of states as additional data sources: some are closer to South Dakota (e.g., Minnesota) while others are very distant (e.g., Florida). When comparing excess KL between the new bigger training distribution and the original dataset relative to the reference dataset, we find a significant negative correlation between accuracy change and excess KL across different states for all the classifiers. These results show that excess KL is indeed a reasonable heuristic for estimating the accuracy drop induced by additional data. Furthermore, the relative ordering of source states in terms of accuracy drop remains consistent across classifiers. However, the scale of accuracy drop for XGB is an order of magnitude smaller than other classifiers. Our results suggest that more data, albeit from a different distribution, does not affect XGB adversely to the same degree.

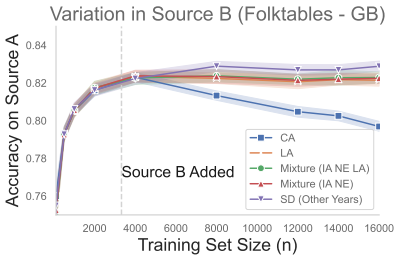

Figure 9: Accuracy on source A when different additional sources are used.

If we replace California with a different state to be the additional data source based on excess KL, we observe better performance as more data (Figure 9). Surprisingly, using Louisiana alone is as helpful as using a mixture of states near South Dakota (e.g., Nebraska and Iowa). Ultimately, the best improvement in performance comes from using South Dakota data from future years (2015-2018) but this source is only slightly better than using a state across the country (e.g. Louisiana) from the same year [4]

6 DISCUSSION

We present two practical cases of data accumulation from multiple sources. We observe a decline or flat-lining in overall accuracy and worst subgroup accuracy when we add data in the SEQUENTIAL case in 3 real world datasets. Since the SEQUENTIAL case can be widespread, we urge caution when enlarging training datasets in real-world data collection. We use excess KL divergence as a heuristic for estimating when adding more data might be undesirable. Ultimately, we expect the trade-off between a smaller high-quality data set and a larger low-quality dataset to be model-dependent. Nevertheless, our results motivate practitioners, particularly in high-stakes applications, to carefully consider the costs and benefits of adding more data.

Limitations & Future Work For future work, there are many opportunities to investigate data-driven performance trade-offs under more complex data accumulation and data curation schemes, including cases involving data pruning (Sorscher et al., 2022; Hooker et al., 2020) and synthetic data (Nikolenko, 2019) or settings that are a combination of the MIXTURE and SEQUENTIAL settings we present. Our work focuses on the tablular data setting – many of the claims made around data scaling laws relate to the context of large language models (Kaplan et al., 2020; Bansal et al., 2022), computer vision models (Zhai et al.) and image-text models (Nguyen et al., 2022) in an over-parameterized regime (Nakkiran et al., 2021), involving much more complex and expressive algorithms (ie. transformers, convolution neural nets). The data scaling claims in the tabular and small-scale setting we investigate may be sufficient illustrations of possible data dynamics but fall short of painting a complete picture of a general law for data accumulation, if one were to exist.

Conclusion Data decision-making is a critical factor in the effective execution of machine learning – yet little is understood of how practical data curation and collection strategies ultimately impact model outcomes. This work is an initial inquiry into what is often an overlooked aspect of the machine-learning process. We challenge long-held assumptions around data scaling, revealing the complexity in enlarging dataset size in the practical setting and discussing how this complexity could yield scenarios in which this assumption, that more data is all you need, no longer holds. We hope this can be a vehicle towards more thorough future modeling and investigations into the practical data decision-making process that underscores much of the model behavior in deployed systems.

[4]Our metric can also be applied to mixtures of sources added sequentially. We add a discussion of this in the appendix.

REFERENCES

David Acuna, Guojun Zhang, Marc T Law, and Sanja Fidler. f-domain adversarial learning: Theory and algorithms. In *International Conference on Machine Learning*, pp. 66–75. PMLR, 2021.

Martin Arjovsky, Kamalika Chaudhuri, and David Lopez-Paz. Throwing away data improves worst-class error in imbalanced classification. *arXiv preprint arXiv:2205.11672*, 2022.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *arXiv preprint arXiv:2102.06701*, 2021.

Yamini Bansal, Behrooz Ghorbani, Ankush Garg, Biao Zhang, Colin Cherry, Behnam Neyshabur, and Orhan Firat. Data scaling laws in nmt: The effect of noise and architecture. In *International Conference on Machine Learning*, pp. 1466–1482. PMLR, 2022.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.

Valerie C Bradley, Shiro Kuriwaki, Michael Isakov, Dino Sejdinovic, Xiao-Li Meng, and Seth Flaxman. Unrepresentative big surveys significantly overestimated us vaccine uptake. *Nature*, 600 (7890):695–700, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. *Advances in Neural Information Processing Systems*, 32, 2019.

Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? *Advances in neural information processing systems*, 31, 2018.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL http://doi.acm.org/10.1145/2939672.2939785.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Rhys Compton, Lily Zhang, Aahlad Puli, and Rajesh Ranganath. When more is less: Incorporating additional datasets can hurt performance by introducing spurious correlations. *arXiv preprint arXiv:2308.04431*, 2023.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490, 2021.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp. 1189–1232, 2001.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Tatsunori Hashimoto. Model performance scaling with multiple data sources. In *International Conference on Machine Learning*, pp. 4107–4116. PMLR, 2021.

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058*, 2020.

Zachary Izzo, Ruishan Liu, and James Zou. Data-driven subgroup identification for linear regression. *arXiv preprint arXiv:2305.00195*, 2023.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv. *PhysioNet. Available online at: https://physionet. org/content/mimiciv/1.0/(accessed August 23, 2021)*, 2020.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Wouter M Kouw and Marco Loog. An introduction to domain adaptation and transfer learning. *arXiv preprint arXiv:1812.11806*, 2018.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. When less is more: Investigating data pruning for pretraining llms at scale. *arXiv preprint arXiv:2309.04564*, 2023.

Xiao-Li Meng. Statistical paradises and paradoxes in big data (i) law of large populations, big data paradox, and the 2016 us presidential election. *The Annals of Applied Statistics*, 12(2):685–726, 2018.

John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, pp. 7721–7735. PMLR, 2021.

Margaret Mitchell, Alexandra Sasha Luccioni, Nathan Lambert, Marissa Gerchick, Angelina McMillan-Major, Ezinwanne Ozoani, Nazneen Rajani, Tristan Thrush, Yacine Jernite, and Douwe Kiela. Measuring data. *arXiv preprint arXiv:2212.05129*, 2022.

Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5715–5725, 2017.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. Quality not quantity: On the interaction between dataset design and robustness of clip. *arXiv preprint arXiv:2208.05516*, 2022.

Sergey I Nikolenko. Synthetic data for deep learning. *arXiv preprint arXiv:1909.11512*, 2019.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021.

Esther Rolf, Theodora T Worledge, Benjamin Recht, and Michael Jordan. Representation matters: Assessing the importance of subgroup allocations in training data. In *International Conference on Machine Learning*, pp. 9040–9051. PMLR, 2021.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Shreya Shankar, Rolando Garcia, Joseph M Hellerstein, and Aditya G Parameswaran. Operationalizing machine learning: An interview study. *arXiv preprint arXiv:2209.09125*, 2022.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.

Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre Alvise-Rebuffi, Ira Ktena, Krishnamurthy Dvijotham, and Taylan Cemgil. A fine-grained analysis on distribution shift. *arXiv preprint arXiv:2110.11328*, 2021.

Garrett Wilson and Diane J Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46, 2020.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. *arXiv preprint arXiv:2305.10429*, 2023a.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. Data selection for language models via importance resampling. *arXiv preprint arXiv:2302.03169*, 2023b.

Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. Change is hard: A closer look at subpopulation shift. *arXiv preprint arXiv:2302.12254*, 2023.

Yelp. Yelp dataset challenge. *Yelp Blog*, 2023. URL `https://www.yelp.com/dataset`.

X Zhai, A Kolesnikov, N Houlsby, and L Beyer. Scaling vision transformers. arxiv 2021. *arXiv preprint arXiv:2106.04560*.

Yian Zhang, Alex Warstadt, Haau-Sing Li, and Samuel R Bowman. When do you need billions of words of pretraining data? *arXiv preprint arXiv:2011.04946*, 2020.

Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *International conference on machine learning*, pp. 7404–7413. PMLR, 2019.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1): 43–76, 2020.

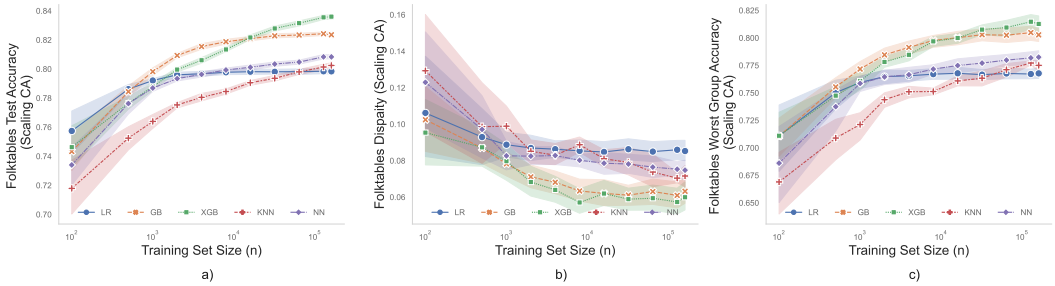

Figure 10: The power of more data: scaling properties for single source South Dakota Folktables data

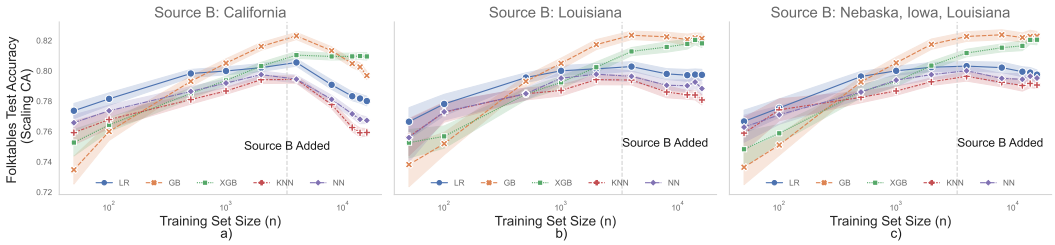

Figure 11: Different choices of source B based on excess KL

A Additional Experiments and Results

A.1 Scaling Folktables Data

Since much of prior work is in deep learning models such as language models and vision pre-training, we first establish the potential of more data. Looking at different training set sizes from a single source of Califonia in Figure 10, we see that the addition of more data improves overall model performance, reduces disparities between groups, and improves worst group accuracy.

Similarly, we test the effect of scaling by expanding South Dakota data by adding data from more years. In Figure 12, we augment data found in South Dakota by adding data from previous years and observe the same phenomenon of more data improving overall model accuracy and worst group accuracy.

We also include the effect of adding more data from different possible source B datasets. We consider the states which had low empirical excesses KL (Figure 11

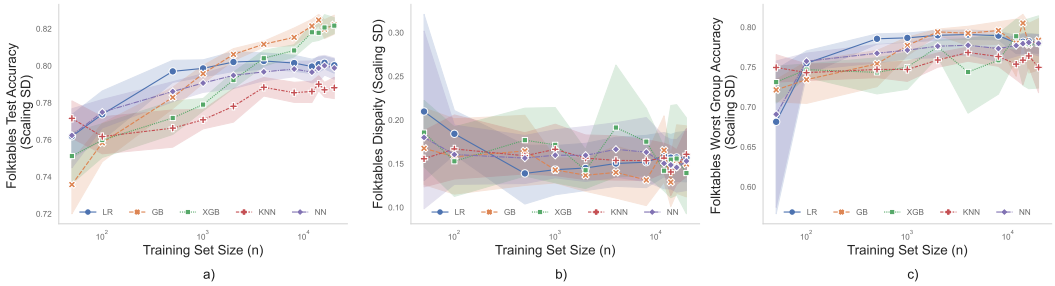

Figure 12: The power of more data: scaling properties for single source California Folktables data

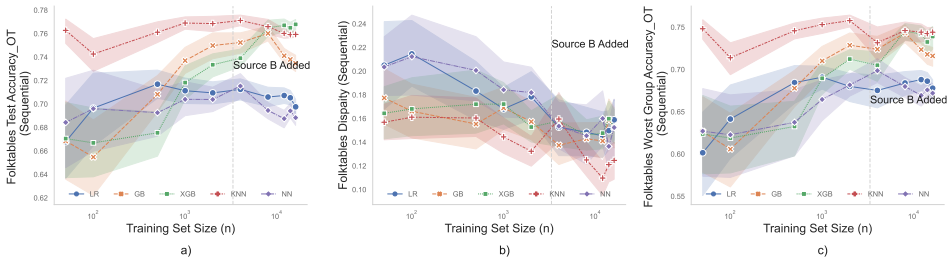

Figure 13: AUC for Folktables results on **Source A reference test set** in the SEQUENTIAL case over 50 trials for **(a)** changes in AUC with increasing data, **(b)** changes in AUC disparity across subgroups, and **(c)** worst case subgroup AUC. source A from South Dakota and Source B from California is added once South Dakota data has been exhausted.

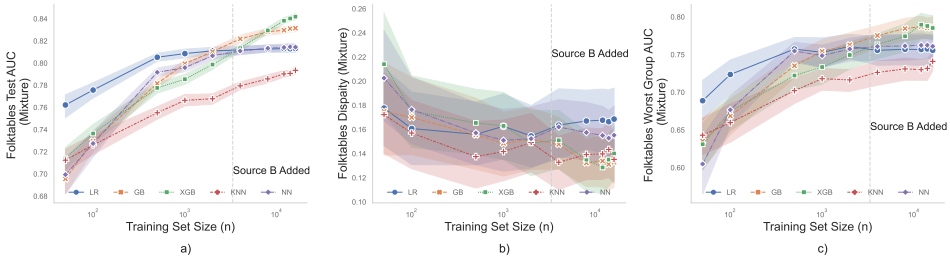

Figure 14: AUC for Folktables results on **Source A reference test set** in the MIXTURE case over 30 trials for **(a)** changes in AUC with increasing data, **(b)** changes in AUC disparity between subgroups, and **(c)** worst subgroup AUC.

A.2 AUC RESULTS AND THRESHOLD OPTIMIZED ACCURACY

In our main results, we study accuracy with the threshold at 0.5. We conduct additional experiments on AUC for both the SEQUENTIAL (Figure 13, 15, 16) and MIXTURE (Figure 14) cases. In the MIXTURE case, AUC climbs steadily as the data is added. In contrast, AUC stagnates and decreases (with the exception of XGB) in the SEQUENTIAL case. When looking at subgroup disparity, the curves are a lot noisier and difficult to compare.

In addition to AUC, we also compare accuracy with dynamic threshold selection (i.e. using Youden Index). One threshold was selected using a held out validation set for the entire population. In this setting, we again observe statistically significant decreases in accuracy for models in the SEQUENTIAL setting.

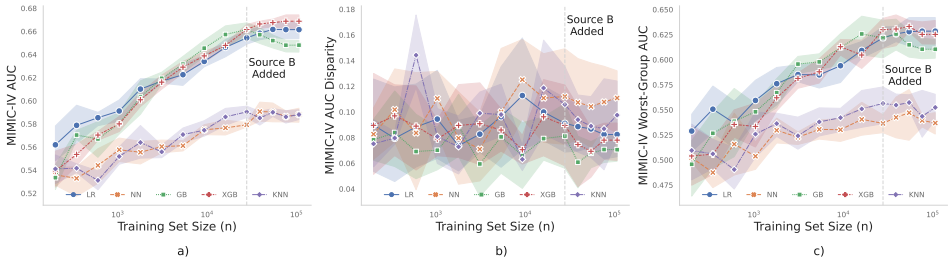

Figure 15: AUC for MIMIC-IV results on **Source A reference test set** in the SEQUENTIAL case over 50 trials for **(a)** changes in AUC with increasing data, **(b)** changes in AUC disparity across subgroups, and **(c)** worst case subgroup AUC.

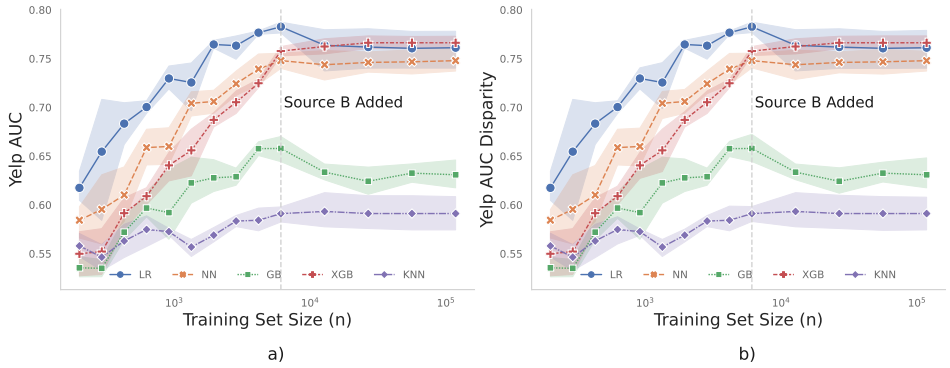

Figure 16: AUC for MIMIC-IV results on **Source A reference test set** in the SEQUENTIAL case over 50 trials for **(a)** changes in AUC with increasing data, **(b)** changes in AUC disparity across subgroups, and **(c)** worst case subgroup AUC.

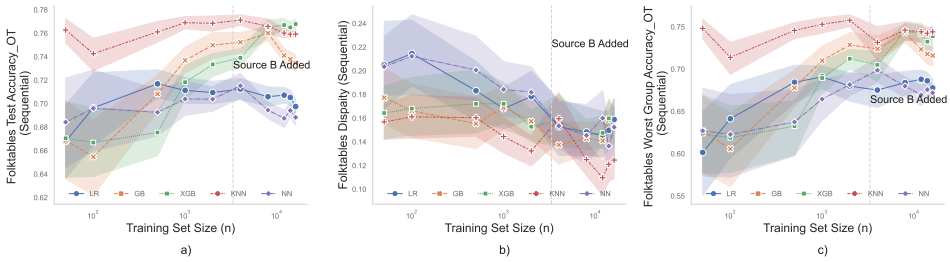

Figure 17: Accuracy (with threshold selection) for Folktables results on **Source A reference test set** in the SEQUENTIAL case over 50 trials for **(a)** changes in accuracy (with threshold selection) with increasing data, **(b)** changes in accuracy (with threshold selection) disparity across subgroups, and **(c)** worst case subgroup accuracy (with threshold selection). source A from South Dakota and Source B from California is added once South Dakota data has been exhausted.

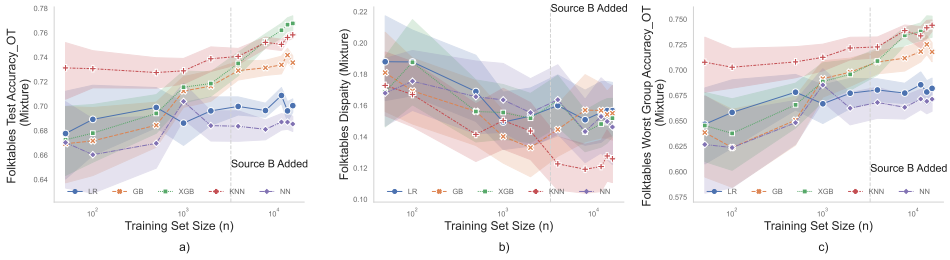

Figure 18: Accuracy (with threshold selection) for Folktables results on **Source A reference test set** in the MIXTURE case over 30 trials for **(a)** changes in accuracy (with threshold selection) with increasing data, **(b)** changes in accuracy (with threshold selection) disparity between subgroups, and **(c)** worst subgroup accuracy (with threshold selection).

A.3 Additional Sequential Experiments

Sequentially Adding Data Across Years We also examine the sequential setting with respect to the year of data collection. This scenario might arise when an organization may want to train on the current year's data but the dataset size is insufficient. Data from previous years might then be added.

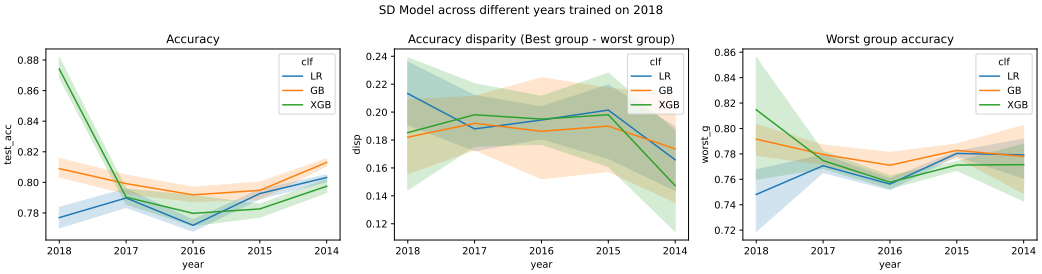

Figure 19: Test time overall accuracy, disparity, and worst group accuracy of a model trained on 2018 South Dakota Income dataset evaluated on 2014-2017 South Dakota Income dataset.

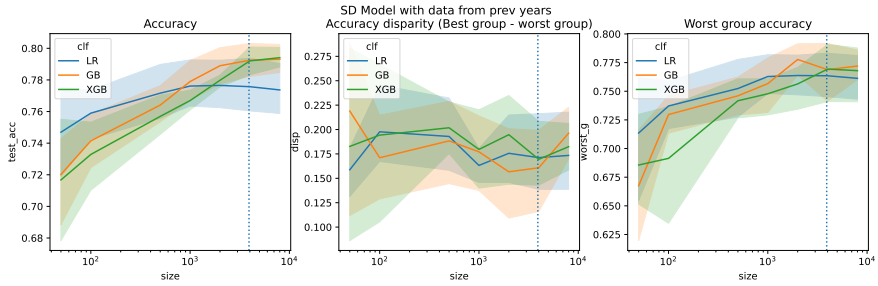

Figure 20: Folktables results on **Source A reference test set** in the SEQUENTIAL case over 5 trials for **(a)** changes in accuracy with increasing data, **(b)** changes in accuracy disparity across subgroups, and **(c)** worst case subgroup performance. Source A from 2018 SD and Source B from 2016 SD is added once 2018 data has been exhausted.

A.4 Detailed Main Paper Results

Please see Section 5 of the main paper for a summary of main results and analysis.

One observation to elaborate on is that it requires much more data from the MIXTURE case to match the performance of a model trained on less data that more closely matches the distribution of the test set—a result we expect from our theoretical model. For instance, for LR trained on Folktables, at N = 4000, which is the maximum size for the single-source setting, accuracy is 80.6% compared to a 75.2% test accuracy for a similar model in the MIXTURE case. Following the SEQUENTIAL strategy though, at N = 14000, that test accuracy drops to 79.1% with the addition of data points from another Source B, while the MIXTURE case test accuracy remains fairly stagnant with the scaling of the mixture, to 74.9%. These effects are influenced by modeling decisions - for example, for the GB and XGB models, there is a larger observed performance drop from scaling in the SEQUENTIAL case (GB: 82.6% to 80.9%; XGB: 81.7% to 80.7%) and a larger performance improvement from scaling in the MIXTURE case (GB: 75.5% to 76.7%; XGB: 74.8% to 77.3%), though overall absolute best performance is still observed in the sequential case.

A.5 Toy Experiments

Before running real-world experiments, we first tested our concept on synthetic data. We consider two source A and source B where $y_A(x) = \sin(x)$ and $y_B(x) = -\sin(x)$. $\hat{D}_1$ is sized $n_A = 10$ comes from source A and $\hat{D}_2$ is sized $n_B = 90$ comes from source B. The training set $\hat{D}_{train} = \hat{D}_1 \cup \hat{D}_2$ while the test set comes from just source A.

Folktables SEQUENTIAL Setting Results N_{Train}	50 (95% CI)	4000 (95% CI)	16000 (95% CI)
Test Accuracy			
LR	0.774 ± 0.005	$\mathbf{0.806 \pm 0.003}$	0.780 ± 0.003
GB	0.735 ± 0.010	$\mathbf{0.823 \pm 0.002}$	0.797 ± 0.003
XGB	0.753 ± 0.008	$\mathbf{0.810 \pm 0.002}$	0.810 ± 0.002
NN	0.766 ± 0.005	$\mathbf{0.794 \pm 0.003}$	0.767 ± 0.002
KNN	0.759 ± 0.008	$\mathbf{0.794 \pm 0.002}$	0.759 ± 0.002
Maximum Subgroup Disparity			
LR	0.204 ± 0.043	0.168 ± 0.013	0.174 ± 0.018
GB	0.166 ± 0.019	0.140 ± 0.014	0.150 ± 0.016
XGB	0.170 ± 0.019	0.137 ± 0.012	0.139 ± 0.013
NN	0.170 ± 0.017	0.175 ± 0.018	0.165 ± 0.022
KNN	0.162 ± 0.017	0.146 ± 0.016	0.125 ± 0.015
Worst-Subgroup Accuracy			
LR	0.694 ± 0.047	0.788 ± 0.007	0.753 ± 0.012
GB	0.711 ± 0.016	0.799 ± 0.010	0.776 ± 0.011
XGB	0.733 ± 0.014	0.787 ± 0.008	0.789 ± 0.009
NN	0.735 ± 0.017	0.772 ± 0.012	0.745 ± 0.012
KNN	0.745 ± 0.009	0.767 ± 0.013	0.744 ± 0.007

Folktables MIXTURE Setting Results N_{Train}	50 (95% CI)	4000 (95% CI)	16000 (95% CI)
Test Accuracy			
LR	0.728 ± 0.012	0.776 ± 0.003	$\mathbf{0.780 \pm 0.003}$
GB	0.690 ± 0.015	0.792 ± 0.003	$\mathbf{0.797 \pm 0.003}$
XGB	0.715 ± 0.015	0.790 ± 0.003	$\mathbf{0.808 \pm 0.002}$
NN	0.748 ± 0.012	0.764 ± 0.003	$\mathbf{0.766 \pm 0.003}$
KNN	0.701 ± 0.018	0.744 ± 0.004	$\mathbf{0.758 \pm 0.003}$
Maximum Subgroup Disparity			
LR	0.190 ± 0.026	0.167 ± 0.018	0.174 ± 0.018
GB	0.184 ± 0.024	0.156 ± 0.016	0.146 ± 0.016
XGB	0.179 ± 0.022	0.155 ± 0.014	0.144 ± 0.013
NN	0.169 ± 0.019	0.172 ± 0.020	0.161 ± 0.023
KNN	0.176 ± 0.022	0.125 ± 0.018	0.126 ± 0.015
Worst-Subgroup Accuracy			
LR	0.689 ± 0.022	0.752 ± 0.012	0.754 ± 0.012
GB	0.660 ± 0.023	0.771 ± 0.011	0.776 ± 0.011
XGB	0.691 ± 0.018	0.769 ± 0.009	0.785 ± 0.010
NN	0.733 ± 0.013	0.741 ± 0.012	0.745 ± 0.014
KNN	0.680 ± 0.020	0.725 ± 0.009	0.744 ± 0.006

Table 2: Summary of model results at key checkpoints of training set size N_{Train}. We consider three models: logistic regression (LR), gradient boosting (GB), and XG-Boost (XGB). These results are for models trained on the Folktables dataset. At $N_{Train} = 4000$, in the SEQUENTIAL setting, data from an additional Source B is added to the samples from Source A. For all N_{Train} in the MIXTURE setting, Source A and B are sampled at a 75:25 ratio. All performance numbers are reported on a test set from Source A.

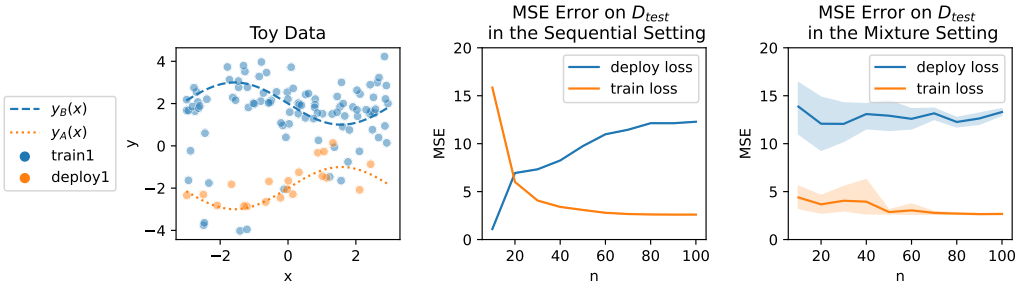

Figure 21: Toy experiments where in the sequential case, we see error on the test set getting worst as the size of the training examples in creases in the Sequential Setting

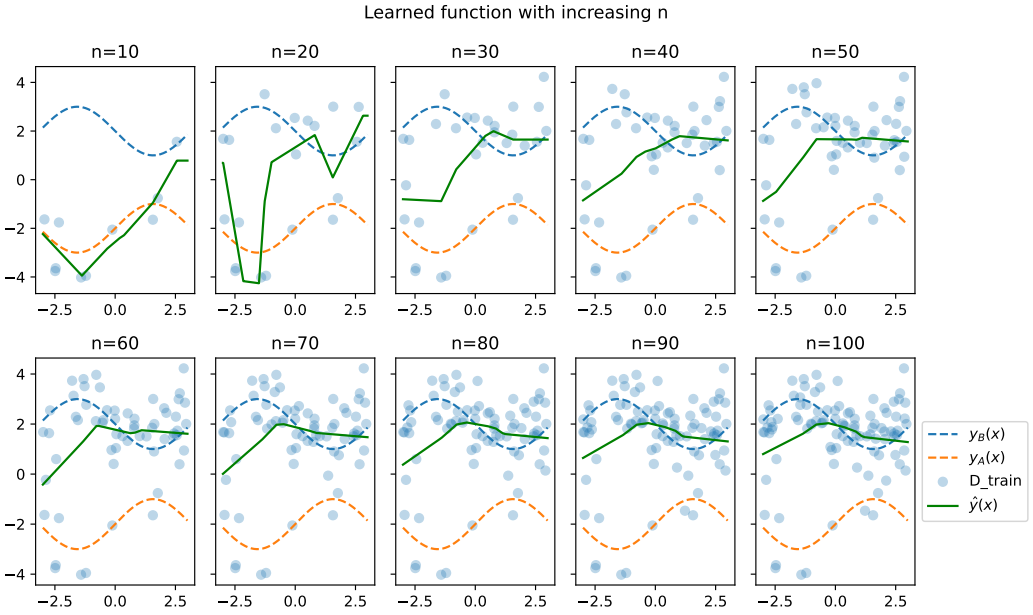

Figure 22: Visualization of learn function as the number of training data points increases. We see that $\hat{y}$ becomes closer to $y_B(x)$ as more data points are added from $\hat{D}_2$ in the sequential Setting.

B Extensions to Notions of Data Quality

The canonical model of data in machine learning assumes that empirical data are sampled iid from an underlying distribution. There are several consequences of these given conditions, most notably the following assumptions:

- Test and training set data distributions are identical.
- Training data distribution remains consistent as the dataset size increases.

However, something that is not as often considered is how the training set distribution *changes* as dataset size increases, at times becoming increasingly divergent from a fixed test distribution. In this work, we specifically investigate this phenomenon.

B.1 Sources of Sampling Bias

Not all cheap data is equally bad and not all distributions are equally shifted. In this paper, we discuss sampling bias issues incurred via data accumulation. Here, we awknowledge some common

sources of sampling bias that may arise when the following types of data are favored in the data sampling process:

1. Easy-to-access data: Machine learning datasets are often amassed using data that is readily available on the internet or even data that is collected without consent. In the healthcare domain, publicly available datasets from a few select hospitals may dominate the training set for a specific task.
2. Complete data: Parts of datasets with missing features may be discarded. Individuals may be included in surveys only if a survey is filled out in its entirety. Differential knowledge of family medical history may impact which health records are used to identify and develop certain diagnostic tools.
3. Unambiguous data: Data points with multiple or conflicting labels may be discarded. Outlier or ambiguous images or text data may be discarded to bolder the statistical robustness of the resultant classifier.

C Detailed Related Work

C.1 Data Scaling

"Scaling laws" more broadly refer to how increases in model "size" lead to improved performance. Typically, model size is described in terms of the number of model parameters, compute and other measured factors characterizing the model (Kaplan et al., 2020; Bahri et al., 2021). Data "scaling laws" (Zhang et al., 2020; Bansal et al., 2022; Zhai et al.) specifically reveal the way in which training on larger and larger datasets yield improved performance. For instance, in (Touvron et al., 2023), large language models trained on datasets with at least 1T tokens beats out models with an order of magnitude more parameters.

Looking at scaling laws for multiple sources in particular, (Hashimoto, 2021) model data as coming from k sources $q_1, ..., q_k$ where $p = \sum_{k\in[k]} q_k p_k$ is the training data; these sources could be different categories of amazon reviews. They measure the resulting excess loss on some test distribution that arises from training with a specific data mixture of n datapoints $p_{n,q}$:

$$L(n, q) = \mathbb{E}l(\hat{\theta}(p_{n,q}, x, y)) - \inf_{\theta} \mathbb{E}l(\theta; x, y)$$

In their experiments, they find that $L(n, q)$ only decreases as n increases; even when q does not match the test distribution well. But they propose an estimation technique based on optimal experimental design which estimates the excess loss based on dataset size n and composition q.

C.2 The Limitations of More Data

Unfortunately, more data can also lead us astray. (Meng, 2018) present a model which decomposes estimation error into three components: data quality, data quantity, and problem difficulty:

$$\hat{\theta} - \theta = \rho_{R,\theta} \times \sqrt{\frac{1-f}{f}} \times \sigma_\theta$$

where R is a binary random variable representing whether an individual responded, θ is the quantity of interest we are trying to estimate, and f captures how much of the underlying population (i.e. $f = 1$ corresponds to the entire population while $f = 0$ corresponds to no data). $\rho_{R,\theta}$ represents data quality; if R corresponded to a perfected random sample, this correlation between θ and R should be zero. $\sqrt{\frac{1-f}{f}}$ represents error arising from data quantity; if we survey the entire population $f = 1$, the error would be zero. σ_θ captures problem difficulty and would be zero if θ is a constant. Furthermore, (Meng, 2018) shows that data quality cannot be compensated by more data when sampling is biased and not truly probabilistic, estimation error scales according to $\sqrt{N}$ the population size and not the sample size. Looking at a $Z-$score for the estimation of θ:

$$Z_{n,N} = \frac{\bar{\theta}_n - \bar{\theta}_N}{\sqrt{V_{SRS}(\bar{\theta}_n)}} = \sqrt{N-1}\rho_{R,\theta}$$

Here we see that while the Z-score should go to zero under random sampling ($\rho_{R,\theta} \approx 0$), the sampling error can scale according to the population size $\sqrt{N}$ when there is sampling bias. If we rewrite the above equations with respect to the effective sample size by setting the mean squared error of SRS (simple random sampling) estimator equal to the mean squared error of our biased sampling:

$$n_{eff} \leq \frac{f}{1-f}\frac{1}{\mathbb{E}\rho_{R,\theta}^2 R} = \frac{n}{1-f}\frac{1}{N\mathbb{E}\rho_{R,\theta}^2 R}$$

When $\mathbb{E}\rho_{R,\theta}^2 R$ decreases at a rate of $O(1)$, the effective sample size n_{eff} decreases rapidly. In other words, the actual effective sample size depends crucially on data quality.

C.3 Representation matters (Rolf et al, 2021)

Does collecting more data from underrepresented groups help? In Rolf et al, the key mechanism of change is the allocation of the proportion of groups in the dataset, which the model practitioner can either choose directly or learn for the optimal loss. This assumes the ability to sample directly from the group-specific distribution. The work also maps the approach to importance weighting (reweight training samples with respect to group distributions) and group distributional robust optimization (minimizes the maximum empirical risk over all groups).

We can formulate this as a function of the group proportions of the training data. For dataset $S = \{x_i, y_i, g_i\}_{i=1}^n$ where features x_i, label y_i, and discrete group g_i for groups $G - \{1, \ldots, |G|\}$ are measured. The population prevalence $\gamma_g = P_{(X,Y,G)\sim D}[G = g]$ is related to the ability to empirical allocation of groups in the data $\alpha_g = \frac{1}{n}\sum_{i=1}^n \mathbb{I}[g_i = g]$.

Sampling from allocation α is defined as independently sampling of $|G|$ disjoint datasets S_g and concatenating according to $S(\alpha, n) = \bigcup_{g\in G} S_g$ with

$$S_g = \{x_i, y_i, g\}_{i=1}^{n_g}, \quad (x_i, y_i) \sim_{iid} D_g$$

C.4 More data can be helpful for fairness sometimes (Chen et al, 2018)

The paper focuses on many different ways to improve fairness of a model, one of which may be adding more training data. As relevant to this group, error due to variance (as opposed to statistical bias and statistical noise) can be estimated via a distribution learning curve. An assumption of the model is therefore that any new data will be from the same distribution as the training data, on which the learning curve is estimated.

Unfairness can be defined as $\Gamma(\hat{Y}, n) := |\gamma_0(\hat{Y}, n) - \gamma_1(\hat{Y}, n)|$ for predictions $\hat{Y}$, sample size n, and group-specific unfairness γ. Based on prior empirical studies, these type II learning curves can be approximated as asymptotic inverse power-law $\gamma_a(\hat{Y}, n_a) = \alpha_a n_a^{-\beta_a} + \delta_a$.

However, when subgroups are difficult to identify, the role of data is less known. Recent work by Izzo et al. (2023) presents data-driven strategies for finding subpopulations based on groups where a linear relationship exists between features and the label. While this approach does not explicitly add more data, the composition of data here is imperative for understanding downstream outcomes.

Given an unbalanced dataset, subsampling (reducing the dataset size as a result) has been shown to achieve better worst-group performance than empirical risk minimization on the entire dataset (Arjovsky et al., 2022).

C.5 Relationship to Domain Adaptation, Distribution Shift, and Transfer Learning

Domain Adaptation The field of domain adaption provides another lens to view the problem of overcoming data quality challenges. Domain Adaption is concerned with the problem of training models with source data and performing well on a target domain (Kouw & Loog, 2018; Zhuang et al., 2020). In Unsupervised Domain Adaptation, the majority of the work in the area, labeled source domain data, and unlabeled target domain data are available for training(Wilson & Cook, 2020). In Supervised Domain Adaptation, only a few or scarce samples from the target dataset and

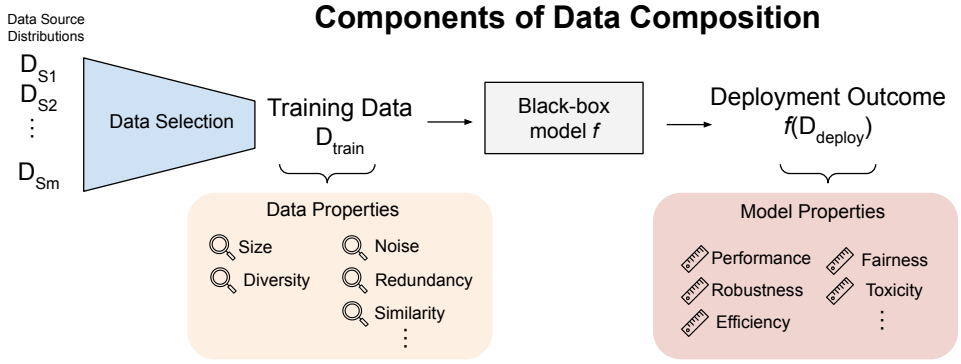

Figure 23: Overview of components of data composition. The domain adaptation literature considers a fixed $D_{train} = \{D_{Source}, D_{Target}\}$ and develops model-based techniques to achieve good *performance* on $D_{deploy} = D_{Target}$ regardless of data properties (Kouw & Loog, 2018). Works in the area of distribution shift assume some difference in *similarity* between D_{train} and D_{deploy} measure drop in *performance* and worst-group performance (a subset of *fairness* metrics (Koh et al., 2021).

the source dataset are fully labeled(Motiian et al., 2017). In this setting, the target domain can be thought of as high-quality data, and the source domain can be thought of as low-quality data (Kouw & Loog, 2018).

A related area of literature concerned with shifts in data quality is distributional robustness and distribution shift. Worst subgroup performance (subpopulation shift) and overall performance (domain generalization) on an out-of-distribution dataset are metrics presented in the domain adaptation literature aimed at measuring distribution shift (Koh et al., 2021). Methods for domain generalization have been critiqued as no better than empirical risk minimization (Gulrajani & Lopez-Paz, 2020). Spurious correlation and unseen data shift observed by Wiles et al. (2021) can be thought of as a fine-grained analysis of barriers for domain generalization. Subpopulation shift, the type of distribution shift, describes a difference in subgroup prevalence between training and unseen datasets that might result in compromised worst group performance (Yang et al., 2023). Works in this area consider a fixed dataset for training(e.g. In- & Out-of-Distribution) to develop algorithmic interventions and give theoretical guarantees based on a uniform sample size or asymptotic results. Little is known about how sample size and data composition contribute to distributional differences themselves.

Data Composition A recent line of work has looked beyond shifts in train and test distributions and instead asked the question of how properties of datasets such as size, diversity, noise, redundancy, and similarity affect model outcomes. *Data Composition* is a broader framework for how data impacts outcomes rather than algorithmic interventions to achieve certain model outcomes (Figure 23). For example, Nguyen et al. (2022) study vision-language models and how combining datasets and increasing dataset size impacts overall performance and distributional robustness. In fact, Gadre et al. (2023) introduces a common task where given a fixed set of models, the goal is to find the best subset of training data. Table 3 summarizes existing work in data composition in terms of dataset domain, data properties measured and model outcomes of interest.

D Dataset Details

D.1 Folktables ACSIncome

Folktables ACSIncome dataset (Ding et al., 2021) is a binary classification task to predict the income of an American adult using 10 census features. Our data sources include data from 2014 12 states: Nebraska (NE), Iowa (Iowa), Minnesota (MN), Ohio (OH), Pennsylvania (PN), Michigan (MI), Texas (TX), Louisiana (LA), Georgia (GA), Florida (FL), California (CA), South Carolina (SC), Washington (WA), Massachusetts (MA); chosen for a mixture of demographic composition and population. We consider subgroups defined by race.

	Domain	Data Properties	Model Outcomes
Nguyen et al. (2022)	Vision-Language	Noise, Size	Robustness
Xie et al. (2023a)	Language	Diversity	Performance, Efficiency
Marion et al. (2023)	Language	Redundancy	Performance
?	Language	Redundancy	Performance
Gadre et al. (2023)	Multi-modal	Size, Similarity, Redundancy	Performance
Our Work	Tabular Data	Similarity, Size	Performance, Fairness (Disparity), Robustness (Worst Group)

Table 3: Summary of Existing Work in Data Composition

D.2 YELP REVIEWS

The Yelp dataset contains crowd-sourced reviews of businesses from states across Canada and the US. The data sources are defined by state, and subgroups are defined by restaurant category. We predict the number of review stars as a multi-class prediction problem using business features review text, totaling 134,092 features.

D.3 MIMIC-IV CLINICAL RECORDS

The Medical Information Mart for Intensive Care (MIMIC)-IV database contains intensive care unit (ICU) patient data between 2008-2019 from the Beth Israel Deaconess Medical Center (BIDMC). We predict binary patient readmission using diagnosis codes extracted from the patient record. There are 9 possible admission types, including urgent care, surgical same-day, and emergency ambulatory observation. Subgroups are patient self-reported race. We predict 15-day patient readmission, and we use 49,469 diagnosis codes recorded in the clinical record as features. The disparity subgroup is the self-reported race group at the time of admission.

E EXPERIMENTAL DETAILS

The investigated scenario is one in which a data modeler has data for one constrained setting of limited training examples, and relies on an external data source or set of sources to supplement the training dataset with additional examples.

Dataset sources are defined by a single sampling process from a general available population of data points. Each distinct sampling process is considered a separate *data source*.

We refer to *data accumulation* as the process of increasing the size of a dataset, often for model training. Note that this constitutes a slightly different context from simpler notions of *data scaling*, which typically involves collecting more samples from a single data sampling process, i.e. scaling up data collection from a single source. Data accumulation includes data scaling but could also involve other approaches to increase dataset size – it thus represents a more generalized notion of increasing dataset size via a variety of available strategies. In this work, we investigate the performance of model performance in two cases of data accumulation (i.e. single-source case and multi-source case) and for two settings of the multi-source case (i.e. SEQUENTIAL and MIXTURE setting).

In the *single-source* setting, training dataset size is collected from a single sampling process from the general population. In order to increase dataset size, the amount of data collected via this process is increased - this is effectively equivalent to much of the prior work on *data scaling*. In the *multi-source* setting, training data is collected via a combination of samples sourced from distinct sampling processes. This means the final training set is effectively an amalgamation of data from more than one dataset source.

Generally, there are at least two settings for the multi-source case, as described in Figure. The SEQUENTIAL setting involves the addition of data from new sources as the overall dataset size increases. The MIXTURE setting includes scaling up a fixed ratio of a mixture of data from more

than one source. One can also imagine a combination of both settings where the ratio of the mixture setting changes as the dataset size increases.

F PROOFS

Proof for Lemma 3.1

Lemma F.1. *Let $D_{train,n}$ be constructed in the* SEQUENTIAL *case from k sources: $D_{S_1}, ..., D_{S_k}$, then if $\delta(D_{S_k}, D_{test}) - \frac{cn}{n_{s_k}} \geq \delta(D_{train,n}, D_{test})$*

$$\delta(D_{train,n}, D_{test}) \geq \delta(D_{train,n-n_{s_k}}, D_{test})$$

where δ belongs to the family of f-divergences and c is a divergence-dependent constant where $\delta(D_{train,n}, D_{test}) + c = \sum_{i=1}^{m} \frac{n_{s_i}}{n} \delta(D_{S_i}, D_{test})$.

Proof. Without loss of generality, the last source in the composition of $D_{train,n}$ is partially used, we define the size of the last source as simply n_k and forget that there is unused data in the last source. Thus we can simply the overall training distribution as $D_{train,n} = \sum_i^k \alpha_i D_{S_i}$ where $\alpha_i = \frac{n_i}{n}$. Furthermore, let n be the total number of examples with k sources and let $n' = n - n_k$ be the total number of examples with $k-1$ sources.

By convexity of δ (Jenson's):

$$\begin{aligned}
&\delta(D_{train,n'}, D_{test}) \\
&\leq \sum_i^{k-1} \frac{n_i}{n'} \delta(D_{S_i}, D_{test}) \\
&= \sum_i^{k-1} \frac{n_i}{n'} \frac{n}{n} \delta(D_{S_i}, D_{test}) \\
&= \frac{n}{n'} \sum_i^{k-1} \frac{n_i}{n} \delta(D_{S_i}, D_{test}) \\
&= \frac{n}{n'} \big(\sum_i^{k} \frac{n_i}{n} \delta(D_{S_i}, D_{test}) - \frac{n_k}{n} \delta(D_{S_k}, D_{test}) \big)
\end{aligned}$$

Since $\delta(D_{train,n}, D_{test}) \leq \sum_i^k \frac{n_i}{n} \delta(D_{S_i}, D_{test})$ then for some constant c:

$$= \frac{n}{n'} \big(\delta(D_{train,n}, D_{test}) + c - \frac{n_k}{n} \delta(D_{S_k}, D_{test}) \big)$$

Thus we need the following condition to be true:

$$\begin{aligned}
\frac{n}{n'}\big(\delta(D_{train,n}, D_{test}) + \\
c - \frac{n_k}{n}\delta(D_{S_k}, D_{test})\big) &\leq \delta(D_{train,n}, D_{test}) \\
\frac{n}{n'}\big(c - \frac{n_k}{n}\delta(D_{S_k}, D_{test})\big) &\leq (1 - \frac{n}{n'})\delta(D_{train,n}, D_{test}) \\
\frac{n}{n'}\big(c - \frac{n_k}{n}\delta(D_{S_k}, D_{test})\big) &\leq (\frac{n' - n}{n'})\delta(D_{train,n}, D_{test}) \\
n\big(c - \frac{n_k}{n}\delta(D_{S_k}, D_{test})\big) &\leq (n' - n)\delta(D_{train,n}, D_{test}) \\
\frac{n}{n' - n}\big(c - \frac{n_k}{n}\delta(D_{S_k}, D_{test})\big) &\geq \delta(D_{train,n}, D_{test}) \\
\frac{n}{n - n'}\big(\frac{n_k}{n}\delta(D_{S_k}, D_{test}) - c\big) &\geq \delta(D_{train,n}, D_{test}) \\
\frac{n_k}{n - n'}\delta(D_{S_k}, D_{test}) - \frac{cn}{n - n'} &\geq \delta(D_{train,n}, D_{test})
\end{aligned}$$

since $n_k = n - n'$

$$\delta(D_{S_k}, D_{test}) - \frac{cn}{n_k} \geq \delta(D_{train,n}, D_{test})$$

This final condition gives tells us that $\delta(D_{S_k}, D_{test})$ must be at least larger than $\delta(D_{train,n}, D_{test}$ in order for the resulting divergence after adding source k to be larger. How much larger depends on c which we can think of as the convexity constant of the divergence for f-divergences. □

Theorem F.2. *(Acuna et al., 2021) For $l : \mathcal{Y} \times \mathcal{Y} \to [0, 1] \in dom\phi*$, every h in some hypothesis class $\mathcal{H}$:*

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_{test}}\left[l(h(x), y)\right] \leq \mathbb{E}_{(x,y)\sim\mathcal{D}_{train,n}}\left[l(h(x), y)\right] + \delta^{\phi}_{h,\mathcal{H}}(\mathcal{D}_{train,n}, \mathcal{D}_{test}) + \lambda$$

where ϕ^* is the Fenchel conjugate of a convex, lower semi-continuous function ϕ that satisfies $\phi(1) = 0$, $\delta^{\phi}_{h,\mathcal{H}}$ is a discrepancy upper bounded by a corresponding f-Divergence, and λ is the sum of risk from the ideal joint hypothesis h^* over the train and test distributions (i.e. $\lambda = \mathbb{E}_{(x,y)\sim\mathcal{D}_{test}}\left[l(h^*(x), y)\right] + \mathbb{E}_{(x,y)\sim\mathcal{D}_{train,n}}\left[l(h^*(x), y)\right]$)