Data-IQ: Characterizing subgroups with heterogeneous outcomes in tabular data

Anonymous Author(s)
Affiliation
Address
email

Abstract

High model performance, on average, can hide that models may systematically underperform on subgroups of the data. We consider the tabular setting, which surfaces the unique issue of *outcome heterogeneity* - this is prevalent in areas such as healthcare, where patients with *similar features* can have *different outcomes*, thus making reliable predictions challenging. To tackle this, we propose *Data-IQ*, a framework to systematically stratify examples into subgroups with respect to their outcomes. We do this by analyzing the behavior of individual examples during training, based on their predictive confidence and, importantly, the aleatoric (data) uncertainty. Capturing the aleatoric uncertainty permits a principled characterization and then subsequent stratification of data examples into three distinct subgroups (Easy, Ambiguous, Hard). We experimentally demonstrate the benefits of Data-IQ on four real-world medical datasets. We show that Data-IQ's characterization of examples is most robust to variation across similarly performant (yet different) models, compared to baselines. Since Data-IQ can be used with *any* ML model (including neural networks, gradient boosting etc.), this property ensures consistency of data characterization, while allowing flexible model selection. Taking this a step further, we demonstrate that the subgroups enable us to construct new approaches to both feature acquisition and dataset selection. Furthermore, we highlight how the subgroups can inform reliable model usage, noting the significant impact of the Ambiguous subgroup on model generalization.

1 Introduction

Most machine learning models are optimized using empirical risk minimization (ERM), to maximize average performance during training [1]. However, in real-world settings, while models may perform well on average, they might underperform on specific subgroups of data [2–4]. Most of the current literature has focused on this problem in computer vision, where the underperforming subgroups are typically associated with data examples that have spurious correlations [1, 5] or mislabelling [6].

In this paper, we focus on tabular data, the most ubiquitous format in medicine and finance, where data is based on relational databases [7, 8]. Specific to the tabular setting, we formalize an understudied source of underperformance, namely *heterogeneity of outcomes*. This phenomenon is vital in healthcare, where patients with *similar features* can have *different outcomes* [9, 10]. For example, [11] showed that prognostic models for risk prediction perform well on average, but underperform on specific cancer types due to heterogeneity of risk (outcome). Prior works have audited subgroups belonging to sensitive attributes (e.g. demographics, race or gender), as it is well-known that ML models generally underperform on these subgroups [12, 13]. However, this approach is limiting, as it needs the sensitive attributes to be specified, and it also does not capture the case where complex feature interactions may lead to underperformance.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

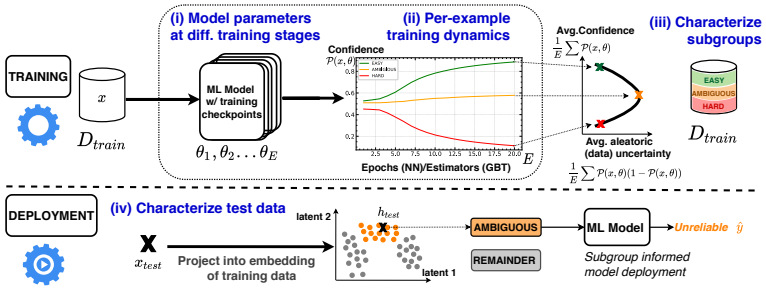


Figure 1: Data-IQ systematically characterizes data into subgroups using (i) any ML model, trained in stages (epochs/iterations) & can be checkpointed. At training time, Data-IQ (ii) leverages the model's checkpoints to analyze the training behavior of individual examples and (iii) characterizes each example based on its aleatoric (data) uncertainty & prediction confidence. (iv) At deployment, the subgroups inform model usage by embedding the training set in a representation space and characterizing the new data points in this representation space.

We take a different approach to automatically stratify data into subgroups, usable with *any ML model trained in stages (epochs/iterations)*; e.g., neural networks, gradient boosting etc. Specifically, we study the behavior of individual examples during training, called *training dynamics*. This allows us to formalize that examples can lie on the spectrum from easy to hard to predict. More concretely, let's consider the task of patient mortality prediction. Based on their features, sicker patients more often have a mortality event. Thus, they are easy to learn for any model and will be predicted *correctly* with *high confidence* (Easy). However, a subgroup might have a heterogeneous outcome: survival despite their poor prognosis. This heterogeneity could result from randomness, making it practically impossible for a model to learn. These examples will be predicted *incorrectly* yet with *high confidence* (or equivalently have low confidence for the correct class) (Hard). In tabular data, there are also examples with inherent ambiguity where the predicted probability for the correct class remains low. They appear where the current features are insufficient to distinguish the example correctly, regardless of the model used [14, 15] (Ambiguous). These subgroups naturally arise in real data; see Fig.1 (ii).

Identifying these subgroups is practically valuable, as improving accuracy and robustness often depends on the data's characteristics and quality [16–19]. As mentioned in [19, 20], the "data" work is often undervalued as merely operational, yet failing to account for it can have immense practical harm [11, 19]. Consequently, our goal is to build a systematic framework with the following desired properties (**P1-P4**), motivated by the considerations of practitioners at various stages of the ML pipeline. In satisfying **P1-P4**, we seek to address the "dire need for an ML-aware data quality that is not only principled, but also practical for a larger collection (. . .) of ML models" [18]:
(P1) Robust data characterization: the characterization of data examples should be robust, such that it is consistent across similar performing models, that have different architectures/parametrizations.
(P2) Principled data collection: the characterization should be informative and actionable, providing practitioners insights that enable both quantitative feature collection and selection between datasets.
(P3) Reliable model deployment: the characterization should enable reliable model usage, both by unmasking unreliable subgroups or using the subgroups to tailor the data for better performance.
(P4) Plug & play: the characterization should be applicable to a variety of ML models widely used on tabular data, including neural networks, gradient boosting (and variants) etc.

To fulfill *P1-P4*, we propose **Data-IQ**, a systematic framework that characterizes examples based on the inherent qualities (IQ) of the data; at both training and deployment time. As outlined in Fig.1, Data-IQ leverages confidence and in the *"data-centric AI"* spirit focused on the data: aleatoric uncertainty (i.e. uncertainty inherent to the data). This permits Data-IQ to provide ML-aware data quality that is principled and practical for a variety of ML models, making the following contributions:

Contributions: ① Data-IQ models the aleatoric (data) uncertainty, which permits subgroup identification that is most robust to variation across different but similar performing models/parameterizations, compared to baselines i.e. *P1* (Sec.4.1). ② Data-IQ aids with principled data collection *P2* in two ways: Firstly, it permits to quantify the value of an acquired feature by measuring how the feature reduces an example's aleatoric uncertainty. This information enables a more principled approach to feature acquisition. Secondly, it permits to compare datasets based on the proportion of ambiguous examples. We demonstrate that the proportions link to how well a model trained with the dataset generalizes (Sec.4.2). ③ Experimentally, the subgroups identified by Data-IQ can inform reliable model deployment, i.e. P3. We highlight cases, where evaluation on average might mask unreliable performance, including data sculpting, model robustness, and uncertainty estimation methods (Sec.4.3). ④ Data-IQ by construction is "plug-and-play" i.e. *P4* with *any* ML model that can be checkpointed, granting practitioners flexibility to use Data-IQ with their model of choice (Sec.3.4).

2 Related work

This paper primarily engages with the literature on data characterization and contributes to the nascent area of data-centric AI [21, 22]. An extended discussion of related work is found in Appendix A.1.

Data characterization. The literature to characterize data samples has used a myriad of different metrics. However, their goals have typically been different, such as spurious correlation or mislabelling, compared to Data-IQ, whose goal is to characterize subgroups with respect to the outcome predictions. Furthermore, none of these methods completely addresses all the desired properties (P1-P4). The closest to our work on data quality is Data Maps [23]. A key contrast to Data-IQ is that Data Maps use confidence and prediction variability to flag instances. In Sec. 3, we show that this prediction variability corresponds to the model uncertainty (i.e. epistemic uncertainty). Alternatively, Data-IQ takes a different and more principled approach, capturing the inherent data uncertainty (known as *aleatoric uncertainty*) [24]. Epistemic uncertainty is reducible by collecting more data. In comparison, aleatoric uncertainty is irreducible even with more samples. This is due to the fact that it captures properties inherent to the data [24–26]; only better features can reduce the aleatoric uncertainty [26]. Later in Fig. 4, we show on real data that capturing the aleatoric uncertainty allows Data-IQ to be more robust to variation across different models, compared to Data Maps (*P1*). This allows practitioners to characterize their data in such a way that the insights are more consistent. We further show theoretically in Sec. 3.3, why the characterization by Data-IQ indeed provides a more principled definition for *Ambiguous* examples, compared to Data Maps.

Besides Data Maps, other related methods address specific computer vision problems: identifying mislabelled images using area under the margin (AUM) [6], gradient norm to identify "important examples" to aid pruning during training [27], or underperformance due to spurious image correlations [1]. The tabular setting considered in this paper requires new methods, due to the specific problem of heterogeneous outcomes for examples with similar features (i.e. "feature collision"). The ambiguity in the tabular, "feature collision" sense, is different or non-existent in modalities such as images.

Data-Centric AI. The assessment of data quality is a critical but often overlooked problem in ML [19]. While the focus in ML is typically on optimizing models, the task of ensuring high quality data (or even improving one's data) can be equally valuable to improving performance [16, 19]. Even when it is considered, the process of assessing datasets is adhoc or artisinal [19, 21, 28]. The recent growth of the data-centric AI space aims to build systematic tools for "data collection, labeling, and quality monitoring processes for datasets to be used in machine learning" [22]. Data-IQ contributes to this nascent body of work, specifically around ML-aware data quality monitoring [18].

3 Formulation

This section gives a detailed formulation of Data-IQ and motivates our proposed example stratification that uses aleatoric uncertainty and confidence. We then describe how Data-IQ stratifies examples into subgroups at both training and testing time. Finally, we show Data-IQ's formulation permits usage with *any* ML model trained in stages, e.g. neural networks, GBDTs etc, unlike other approaches.

3.1 Preliminaries

We consider the typical supervised learning setting, where the aim is to assign an input $x \in \mathcal{X} \subseteq \mathbb{R}^{d_X}$ to a class $y \in \mathcal{Y} \subset \mathbb{N}$. We have a dataset $\mathcal{D}$ with $N \in \mathbb{N}^*$ examples, i.e. $\mathcal{D} = \{(x^n, y^n) \mid n \in [N]\}$ drawn IID from an unknown distribution. Our goal is then to learn a model $f_\theta : \mathcal{X} \to \mathcal{Y}$, parameterized by $\theta \in \Theta$. Typically, the parameters θ are learned to minimize empirical risk, by minimizing the average training loss , i.e. $\text{ERM}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(x_i, y_i; \theta)$, with a loss function $\ell : \mathcal{X} \times \mathcal{Y} \times \Theta \to \mathbb{R}^+$.

This brings us to the essence of the problem: "not all examples are created equally". e.g. patients with similar features might have heterogeneous outcomes, reflected in their labels y being different. These correspond to subgroups within $\mathcal{D}_{\text{train}}$ on which a predictive model might systematically underperform. We formalize this concept of hidden heterogeneous subgroups by assigning to each example x^n a hidden *subgroup* label $g^n \in \mathcal{G}$, where $\mathcal{G} = \{\text{Easy}, \text{Ambiguous}, \text{Hard}\}$. Before giving a precise description of how those group labels are assigned, it is useful to detail the context. Several works have established that the training dynamics of a model, contains signal about the quality of the data itself [29–31]. For instance, it takes more epochs/iterations for a model to assign the correct

label to noisier/more difficult training examples. With Data-IQ, we build on those observations and assign a label g^n to each example x^n by studying its training dynamic, which is then used to estimate the aleatoric uncertainty and predictive confidence of each example. The following sections detail how this is done and how this contrasts with existing approaches.

3.2 Uncertainty decomposition during training

Recall that practitioners desire flexibility in the choice of the model. Hence, we focus on *any* ML model that is trained in stages and can be checkpointed during training, $f_\theta : \mathcal{X} \to \mathcal{Y}$ parameterized by θ and on a given example *from the training set* $(x, y) \in \mathcal{D}_{\text{train}}$. Assume that the model f_θ corresponds to a conditional categorical distribution, assigning a probability to each class given the input x: $f_\theta(x) = P(Y \mid X = x, \vartheta = \theta)$. During iterative training, the model parameters θ vary, where over $E \in \mathbb{N}^*$ epochs/iterations, these parameters take E different values at each checkpoint, i.e. $\theta_1, \theta_2, \ldots, \theta_E$. Since our analysis relies on the model's training dynamics, we want to take those different parameters into account. For the sake of notation, we introduce a random variable ϑ that has an empirical distribution over this set of parameters captured through the training process $\vartheta \sim P_{\text{emp}}(\{\theta_e \mid e \in [E]\})$. The variability of the model's parameters at training time is then reflected by the variance $\mathbb{V}_\vartheta[\cdot]$.

The uncertainty we model is based on the random variable $Y \mid X = x$ that represents the possible labels given the input x. Since the ground-truth label y is available for training examples, we would like to distinguish between 2 cases: ① the predicted label corresponds to the ground-truth label $Y = y$ and ② the predicted label is different from the ground-truth label $Y \neq y$. To this end, we introduce a binary random variable $\tilde{Y}$ that is set to one when the predicted label equals the ground-truth label ($\tilde{Y} = 1$ if $Y = y$) and that is zero otherwise ($\tilde{Y} = 0$ if $Y \neq y$). As discussed earlier, we are interested in the uncertainty on the predictive random variable $\tilde{Y} \mid X = x$. This uncertainty is modeled by the variance $v(x) = \mathbb{V}_{\tilde{Y}|X}\left[\tilde{Y} \mid X = x\right]$. We will now show that this quantity can be evaluated with the model predictions.

We start by noting that the definition of $\tilde{Y}$ implies that $\tilde{Y} \mid X = x, \vartheta = \theta$ is a Bernoulli random variable with parameter[1] $\mathcal{P}(x, \theta) = P(Y = y \mid X = x, \vartheta = \theta) = [f_\theta(x)]_y$. From this observation, we can decompose $v(x)$ with the law of total variance and make each term explicit:

$$\begin{aligned} v(x) &= \underbrace{\mathbb{V}_\vartheta\left[\mathbb{E}_{\tilde{Y}|X,\vartheta}\left[\tilde{Y}|X = x, \vartheta\right]\right]}_{\text{Bernoulli Mean: } \mathcal{P}(x,\vartheta)} + \underbrace{\mathbb{E}_\vartheta\left[\mathbb{V}_{\tilde{Y}|X,\vartheta}\left[\tilde{Y}|X = x, \vartheta\right]\right]}_{\text{Bernoulli Var: } \mathcal{P}(x,\vartheta)(1-\mathcal{P}(x,\vartheta))} \\ &= \underbrace{\mathbb{V}_\vartheta\left[\mathcal{P}(x,\vartheta)\right]}_{\text{Epistemic uncertainty: } v_{\text{ep}}(x)} + \underbrace{\mathbb{E}_\vartheta\left[\mathcal{P}(x,\vartheta)(1-\mathcal{P}(x,\vartheta))\right]}_{\text{Aleatoric uncertainty: } v_{\text{al}}(x)}. \end{aligned} \tag{1}$$

In Eq. (1), we have split the overall uncertainty into two components: *epistemic* and *aleatoric uncertainty*. This type of decomposition is similar to those in the context of Bayesian neural networks [32]. To understand the distinction between uncertainties, it is useful to closely examine the variances in the *first* equality of Eq. (1). For *epistemic uncertainty* v_{ep}, variance is evaluated on the *model parameters* ϑ. Hence, epistemic uncertainty originates from the fact that a model's predictions oscillate when we change its parameters. For the *aleatoric uncertainty* v_{al}, the variance is evaluated on the *predicted label* $\tilde{Y} \mid X, \vartheta$. Hence, the variability originates from the inability to predict the correct label with high confidence. While existing works use epistemic uncertainty to stratify examples, we argue that aleatoric uncertainty is a better principled choice to capture the inherent data uncertainty.

3.3 Stratification based on data uncertainty

We now explain how the above notion of uncertainty permits to assign a group label $g \in \mathcal{G}$ to each training example x. First, we use the empirical distribution $\vartheta \sim P_{\text{emp}}(\{\theta_e \mid e \in [E]\})$ to explicitly write the two types of uncertainties in Eq. (1),where $\bar{\mathcal{P}}(x) = {}^1\!/_E \sum_{e=1}^{E} \mathcal{P}(x, \theta_e)$:

$$v_{\text{ep}}(x) = \frac{1}{E}\sum_{e=1}^{E}\left[\mathcal{P}(x,\theta_e) - \bar{\mathcal{P}}(x)\right]^2 \qquad v_{\text{al}}(x) = \frac{1}{E}\sum_{e=1}^{E}\mathcal{P}(x,\theta_e)(1-\mathcal{P}(x,\theta_e)), \tag{2}$$

[1] In this case, $[f_\theta(x)]_y$ denotes the component y of the probability vector $f_\theta(x)$.

Stratification at training time. Before giving a precise definition of the group labels, let us give an intuitive definition for each group. ① Easy: examples that have low data uncertainty that the model can correctly predict with high confidence, ② Ambiguous: examples that have high data uncertainty, hence the model is unable to predict with confidence and ③ Hard: examples that have low data uncertainty that the model is unable to predict (i.e. predicted incorrectly yet with high confidence or equivalently have low confidence for the correct class). We note that we need the model's prediction for the ground-truth class to delineate Easy and Hard examples. In practice, we use the model's average confidence for the ground-truth class $\bar{\mathcal{P}}(x)$ defined previously for this purpose. We make use of this concept to detail how labels are assigned to training examples $(x, y) \in \mathcal{D}_{\text{train}}$:

$$g(x, \mathcal{D}_{\text{train}}) = \begin{cases} \text{Easy} & \text{if } \bar{\mathcal{P}}(x) \geq C_{\text{up}} \ \wedge \ v_{\text{al}}(x) < P_{50}\left[v_{\text{al}}(\mathcal{D}_{\text{train}})\right] \\ \text{Hard} & \text{if } \bar{\mathcal{P}}(x) \leq C_{\text{low}} \ \wedge \ v_{\text{al}}(x) < P_{50}\left[v_{\text{al}}(\mathcal{D}_{\text{train}})\right] \\ \text{Ambiguous} & \text{otherwise} \end{cases} \tag{3}$$

where C_{up} and C_{low} are upper and lower confidence threshold resp. and P_n the n-th percentile. We provide a practical method to set C_{up} and C_{low}, applicable to any dataset in Appendix A.

In contrast to Data-IQ which uses Aleatoric uncertainty $v_{\text{al}}(x)$; Data Maps [23] identifies ambiguous training examples $(x, y) \in \mathcal{D}_{\text{train}}$ as those with high epistemic uncertainty $v_{\text{ep}}(x)$. We consider a typical scenario to see how this characterization might cause problems – illustrated in Fig. 2. Consider an example x in which the model cannot classify confidently during the entire training $\mathcal{P}(x, \theta_e) = 0.5 \ \forall e \in [E]$. In this case, the epistemic uncertainty $v_{\text{ep}}(x)$ vanishes, as the prediction is consistently unconfident (i.e. low variability of the model predictions). This implies that Data Maps would consider this example as non-ambiguous, despite the ambiguous model prediction for this example. This problem can be traced back to the definition of epistemic uncertainty, which measures the sensitivity of a model prediction with respect to the model's parameters.

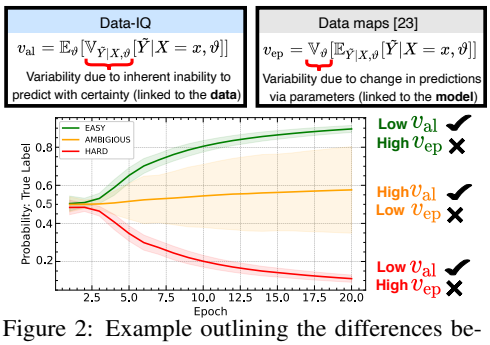


Figure 2: Example outlining the differences between v_{al} (Data-IQ) and v_{ep} (Data Maps) & showing the type of uncertainty matters.

A more principled definition for ambiguous examples should capture examples for which the model cannot predict the appropriate label with high confidence (i.e. data uncertainty). This is precisely what the aleatoric uncertainty $v_{\text{al}}(x)$ captures (Data-IQ). Furthermore, it is easy to verify that the previous example $\mathcal{P}(x, \theta_e) = 0.5 \ \forall e \in [E]$ maximizes the aleatoric uncertainty (see Fig. 2). Since high aleatoric uncertainty captures ambiguous examples for various values of the model's parameters, we believe that it better reflects the inherent quality of the data. In that sense, we expect this quantity to be more stable and robust to variation for different ML model parameters/architecture changes (*P1*). We experimentally validate the consistency in Sec. 4.

Stratification at inference time. Most previous methods are only applicable at training time. To address this limitation and improve the practical utility of our method, we also stratify examples into subgroups at deployment time. However, if we try to apply the above stratification for incoming data at deployment time, we face a problem: $\bar{\mathcal{P}}(x)$ requires the ground-truth class y.

For this reason, we follow an alternative approach based on representation learning that does not require access to ground-truth labels. The idea is the following: we construct a low-dimensional UMAP embedding [33] $h : \mathcal{X} \rightarrow \mathcal{H}$ of the training set's examples $x_{\text{train}} \in \mathcal{D}_{\text{train}}$. In doing this, we note two things (see Appendix C): ① Ambiguous examples have distinctive features and are clustered in embedding space. Thus, it is possible to distinguish the Ambiguous examples using the embedding. ② It is not possible to reliably distinguish Easy examples from Hard examples based on the embedding, because Hard examples are a minority with outcome randomness that have similar features, as the Easy examples. Combining these observations, we note it is possible to identify Ambiguous test examples. This label is assigned by computing the related embedding $h(x_{\text{test}})$ and comparing this embedding to the nearest neighbor embedding from the training set, i.e. $d[h(x_{\text{test}}), h(x_{\text{train}})] \forall x_{\text{train}} \in \mathcal{D}_{\text{train}}$. For models like neural networks with an implicit representation space, the same analysis can be done using the model's representation space.

3.4 Using Data-IQ with a variety of models, beyond Neural Networks (P4)

The baseline methods discussed are primarily applicable only to neural networks. However, practically in tabular settings (e.g. healthcare/finance etc), practitioners often use other highly performant iterative learning algorithms such as Gradient Boost Decision Trees (GBDTs) or variants [7, 34]. Data-IQ's formulation by construction is naturally adaptable to *any* ML model trained in stages, that can be checkpointed. This satisfies *P4*, which allows practitioners the flexibility to use Data-IQ with their application-specific model of choice. Appendix A provides guidelines, space and time considerations, as well as, discussing the specifics of how Data-IQ is easily adapted, for example to GBDTs.

4 Experiments

This section presents a detailed empirical evaluation demonstrating that Data-IQ satisfies (**P1**) Robust data characterization, (**P2**) Principled data collection and (**P3**) Reliable Model Deployment, introduced in Sec.1. Recall that (**P4**) Plug and play is satisfied by construction of Data-IQ.

Datasets. We conduct experiments on four real-world medical datasets, with diverse characteristics (different sizes, binary/multiclass, varying degrees of task difficulty etc) and highlight real-world applicability with heterogeneous patient outcomes: (1) Covid-19 dataset of Brazilian patients [35], (2) Prostate cancer datasets from both the US [36] and UK [37], (3) Support dataset of seriously ill hospitalized adults [38], (4) Fetal state dataset of cardiotocography [39]. We describe the datasets in greater detail in Appendix B, along with further experimental details. *We observe similar performance across different datasets, but given the space limitations, we typically show pertinent results for a single dataset, and include results for the other datasets in Appendix C.*

4.1 (P1) Robust data characterization

Robustness to variation. As per *P1*, we desire that Data-IQ identifies subgroups in a manner robust to variation across different models. This would allow a practitioner to obtain consistent insights about their data even when using different model architectures/parameterizations. When comparing the different methods from Sec. 2, we note that each method has its own specific metric used to characterize examples. To assess robustness to variation, we compare the consistency of the different characterization metrics, evaluated on models with different architectures (number of layers or hidden units) and parameterizations (optimizers). Specifics of the differences are outlined in Appendix B. Note that all models achieve similar training performance and are trained to convergence, with early stopping on a validation set. The main paper does not explicitly compare across model types (neural networks vs gradient boosting). The reason is that besides Data-IQ, the majority of methods are constrained to neural networks. For this comparison, see Appendix C.

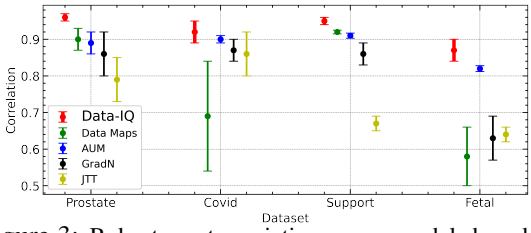
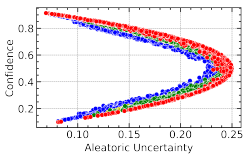


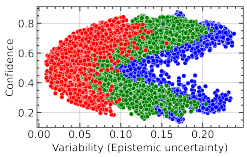
Figure 3: Robustness to variation across models based on Spearman correlation, where Data-IQ has the highest correlation (i.e. consistency) across all datasets.

Quantitatively, we compute the Spearman rank correlation between all model combinations, see Fig. 3. We observe that Data-IQ is the most consistent and robust to variation across different models, having the highest score on all datasets, satisfying *P1*. Further, the baseline methods themselves are also not consistent in performance ordering across datasets, which is undesirable. Ultimately, the robustness means practitioners can feel confident in the consistency of data insights, derived using Data-IQ.

To further compare Data-IQ and Data Maps[23], we examine 3 distinct models that achieve similar performance on the Covid-19 [35] tabular dataset, and we produce a characterization of the training set using each model in Fig. 4. We note that Data Maps groups can be recovered from (5) by replacing the aleatoric uncertainty v_{al} from Data-IQ with its epistemic counterpart v_{ep}. The y-axis is the same for both methods and corresponds to $\bar{\mathcal{P}}(x)$. The x-axis corresponds to $v_{\text{al}}(x)$ for Data-IQ and to $v_{\text{ep}}(x)$ for Data Maps. Each model is assigned a color in Fig. 4. We note three things



(a) Data-IQ



(b) Data Maps [23]

Figure 4: Data-IQ's robustness to variation across diff. models (i.e.colors)

① Data-IQ's characterization of the data is significantly more stable across models. ② Linked to the points in Sec.3, Data Map's high and low confidence examples in fact have high epistemic uncertainty v_{ep}, which can lead to incorrect conclusions when attempting to use Data Maps to characterize data. ③ Data-IQ always distributes the data around a bell shape, which standardizes its interpretation. We provide a theoretical analysis to explain this bell shape observation in Appendix A.

Data insights from subgroups. Given the distinct differences between the subgroups, we seek to understand what factors make these subgroups different and how they can provide insight into the dataset. Such insights are especially useful in clinical settings. Results for the prostate cancer dataset are illustrated in Fig. 5 (with other datasets in Appendix C). To visualize the different groups of patients within each subgroup, we cluster each subgroup (Easy, Ambiguous and Hard) using a Gaussian Mixture Model (GMM) similar to [5], selecting the optimal clusters based on the Silhouette score. We assess cluster quality vs alternatives in Appendix C.

In general, across datasets, the subgroups are: (1) Easy: Severe patients with a death outcome, and less severe patients with a survival outcome. (2) Ambiguous: Patients with similar features, but different outcomes. This could suggest that the features, we have at hand are insufficient to separate the differences in outcomes. (3) Hard: Severe patients with a survival outcome, and less severe patients with a death outcome. i.e. opposite outcomes as expected due to randomness in the outcomes.

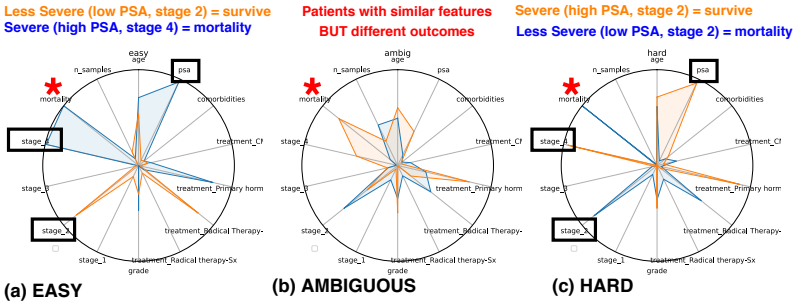
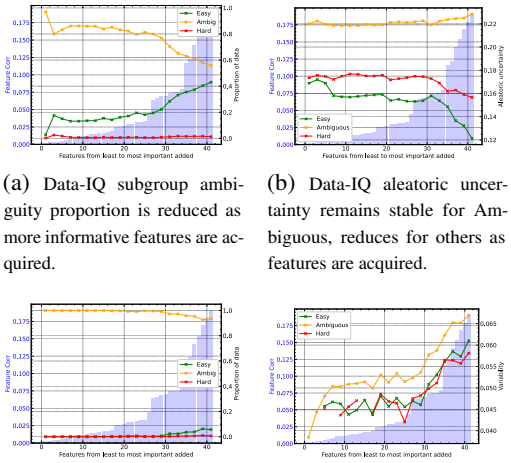


Figure 5: Comparing subgroups identified by Data-IQ (descriptions above). Colors represent the GMM clusters.

4.2 (P2) Principled data collection

Principled feature acquisition. As shown in Sec.4.1, the Ambiguous subgroup has examples with similar features, yet different outcomes. Recall that this case of ambiguity in the tabular setting is very different from ambiguity in other modalities such as images. Rather, in this setting, the ambiguity is due to insufficient features to adequately separate the examples. We link this to the concept that the Ambiguous subgroup has a high aleatoric uncertainty that is irreducible, even if we collect more data examples. Rather, aleatoric uncertainty can only be reduced by acquiring better features [26]. We leverage this idea and show that Data-IQ's example characterization provides a principled approach to assessing the benefit of acquiring a specific feature. This is a distinctly different task from feature selection, where all features are present and we select the most "important feature". Additionally, this is different from active learning which quantifies the value of acquiring examples, rather than features.



(a) Data-IQ subgroup ambiguity proportion is reduced as more informative features are acquired.

(b) Data-IQ aleatoric uncertainty remains stable for Ambiguous, reduces for others as features are acquired.

(c) Data Maps subgroup proportions largely unaffected as features acquired.

(d) Data Maps variability increases across subgroups as features acquired.

Figure 6: Quantifying the value of feature acquisition based on change in ambiguity. Only Data-IQ captures this relationship.

With the above in mind, a valuable feature should decrease the example ambiguity (i.e. aleatoric uncertainty). Hence, a decrease in the proportion of Ambiguous examples can serve as a proxy for the feature's potential value to the dataset. Understanding the value of features is useful in settings such as healthcare, where feature acquisition comes at a cost. To showcase the potential, we construct a semi-synthetic experiment, where we rank sort the features based on correlation with the target.

We then train different models, where we sequentially "acquire" features of increasing value (based on correlation). Fig. 30. shows results for the Support dataset. For Data-IQ, Fig 30 (a) shows that as we acquire "valuable" features, the proportion of the Ambiguous subgroup drops, whilst the Easy subgroup increases, with significant changes for the important features. This shows that Data-IQ's subgroup characterization can be used to quantify a feature's value, by its ability to decrease ambiguity. In contrast, Data Maps, Fig 30 (c), shows minimal response to feature acquisition, suggesting it may not be sensitive enough to capture the feature's value.

Further, for Data-IQ we see that the examples that remain as Ambiguous, after features are collected, maintain a consistent aleatoric uncertainty. This is desired as it demonstrates for those examples which remain Ambiguous, that indeed the features collected are not informative enough to reduce their inherent (aleatoric) uncertainty, i.e. those remaining still need better features (see Fig 30 (b)). While for Data Maps the added features, in fact increase the variability for *all* subgroups making it harder to stratify (see Fig 30 (d)). This links to the fact that Data Maps subgroups can't capture the value of the acquired features. We further show experimentally in Appendix C that for Ambiguous examples, it is not simply a case of increasing the size of the dataset (i.e. more examples). In fact, this can increase the proportion of Ambiguous examples due to the increased probability of feature collisions as the dataset size increases. Ultimately, this motivates the usefulness of principled feature acquisition (which Data-IQ can guide), as a way to decrease dataset ambiguity.

Principled dataset comparison. Extending beyond feature-level, an understudied scenario involves systematically selecting between entire datasets in two cases: (1) purchasing data from data markets [40–42] and (2) organizations where the data is siloed, with lengthy access processes [43, 44]. In both cases, synthetic versions of the real dataset has begun to be used [43]. For now, we ignore privacy concerns, and focus on data fidelity and quality, which compares the real and synthetic datasets using statistical measures. However, as per [45], the conclusions can vary across different metrics. In practice, competing "synthetic" datasets can be generated by different ML models or vendors. Thus, while they model the same underlying distribution, depending on the process used, one version might be superior. We now ask whether Data-IQ could permit us to systematically select between synthetic datasets? We consider the setting where the real data is *not accessible*. Hence, we can't use existing evaluation metrics, yet still wish to compare the synthetic datasets (e.g. comparing vendors).

We simulate this scenario by generating synthetic data using 2 different models, representing 2 synthetic data vendors: (V1) CTGAN [46] and (V2) Gaussian Copula [47]. We then characterize the dataset subgroup proportions using Data-IQ. We hypothesize that datasets with greater proportions of Easy examples generalize better. As is common, we validate fidelity by training with synthetic data and testing with real data [48, 49], where the best fidelity data produces the best model performance on real data (test set). Table 1 shows that datasets with the highest quality as measured by Data-IQ indeed produces the best performance on real test data (i.e. Rank 1). Further, it shows that the same "vendor" does not always produce the best dataset, highlighting the value of comparative assessment. Ultimately, these two aspects demonstrate that when the real data is unavailable, Data-IQ is a useful tool in the hands of practitioners wishing to assess data quality, especially when selecting between different datasets.

Table 1: Comparison of accuracy performance rank and (dataset quality). Synthetic dataset w/ better quality (↑ easy)produces the best real data test performance.

Dataset	(V1) CTGAN	(V2) Gaussian Copula
Prostate	**Rank 1 (67% Easy)**	Rank 2 (37% Easy)
Covid	**Rank 1 (59% Easy)**	Rank 2 (38% Easy)
Support	**Rank 1 (71% Easy)**	Rank 2 (65% Easy)
Fetal	Rank 2 (67% Easy)	**Rank 1 (95% Easy)**

4.3 (P3) Reliable Model Deployment

Less is more: data sculpting based on subgroups. What role do the data subgroups, specifically Ambiguous examples, play in ensuring model generalization? Using multi-country prostate cancer data, we train a baseline model on US data (SEER) and assess generalization when deployed on patients in the UK (CUTRACT) and vice versa. In Fig. 7, we see that test time generalization performance, monotonically increases as we decrease the proportion of Ambiguous training data (see Appendix C.13 for absolute numbers). Ultimately, this illustrates the value of sculpting the training dataset, by removing ambiguous examples, as a way to improve the reliability of a deployed model.

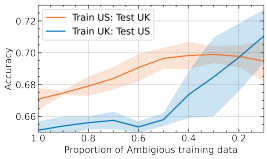


Figure 7: Performance improvement via data sculpting based on subgroups

Group-DRO: Not a silver bullet in tabular settings. Once subgroups of underperformance are identified, it is generally assumed that methods such as Group Distributionally Robust Optimization

(DRO) [50] can be applied to improve model performance and robustness. We compare Group-DRO with groups identified by Data-IQ, and as baselines: George [5] and Just-Train-Twice (JTT) [1]. As per the previous experiments, the largest underperforming group (in proportion) is the Ambiguous examples. Similar to the literature, we evaluate the performance change from a baseline model, after Group-DRO is applied. The results in Table 2 show that Group-DRO using Data-IQ's groups both improves overall performance and improves performance on the Ambiguous group. Whilst, for the other baselines, the performance actually degrades.

Nevertheless, while using Data-IQ can boost performance, it is evident that simply applying Group-DRO is not a silver bullet to equalize subgroup performance, given the sometimes small improvement. The rationale is our tabular setting is different, from the spurious correlation setting in computer vision where Group-DRO typically shines. Ultimately, we believe, based on the feature acquisition results, that in tabular settings, practitioners would be well served acquiring better features to improve performance and reduce ambiguity.

Table 2: Comparison of different model improvement/robustness techniques

Dataset	Group	Baseline	Group-DRO (**Data-IQ**)	Group-DRO (George)	JTT
Prostate	Overall	0.837	**0.838**↑	0.781 ↓	0.834
	Ambiguous	0.743	**0.744** ↑	0.678 ↓	0.726 ↓
	The Rest	0.934	0.934	0.930 ↓	0.934
Covid	Overall	0.719	**0.723**↑	0.647 ↓	0.719
	Ambiguous	0.708	**0.714**↑	0.652 ↓	0.707 ↓
	The Rest	0.785	0.785	0.634 ↓	0.785
Support	Overall	0.736	**0.737**↑	0.678 ↓	0.735 ↓
	Ambiguous	0.630	**0.633**↑	0.602 ↓	**0.633**↑
	The Rest	0.847	0.847	0.763 ↓	0.843
Fetal	Overall	0.766	**0.823**↑	**0.823**↑	0.766
	Ambiguous	0.765	**0.833**↑	0.827 ↑	0.765
	The Rest	0.866	0.879↑	**0.897**↑	0.866

Subgroup-informed usage of uncertainty estimation. Uncertainty estimation methods are essential in safety-critical areas such as healthcare, yet are typically assessed on average. We ask the question: since subgroups have different performance properties, are uncertainty estimates equally reliable for each subgroup? As done in the literature [51, 52], if an uncertainty estimate is reliable and informative of predictive performance, it can be used to defer "uncertain examples". This is done by rank sorting examples based on uncertainty and thresholding proportions of tolerated uncertainty [51–54]. Ideally, as the threshold proportion of examples increases (i.e. inclusion of more uncertain examples), we should see a monotonic decrease of accuracy. We assess this by training a Bayesian Neural Network (BNN)[55] to obtain uncertainty estimates. We then compute the performance across different threshold proportions $\tau \in \{0.1, 0.2, \ldots, 1\}$ for the Ambiguous subgroup specifically, and as is commonly done, across the entire dataset (i.e. *average*).

Fig. 8 shows an example in which the *average* examples exhibit the monotonic decreasing relationship, as expected. However, the Ambiguous examples categorized by Data-IQ , contrary to expectations, show an increase in accuracy as more uncertain examples are included. This suggests that the uncertainty estimates are not as informative of predictive performance for the Ambiguous examples. Data-IQ could thus, be leveraged by practitioners at deployment time to understand which examples require auditing before deferring, as the "average" monotonic decreasing behavior, based on the uncertainty estimates might not always hold. Ultimately, the result further highlights how subgroup characterization via Data-IQ could assist practitioners in unmasking unreliable performance, which is not evident on average.

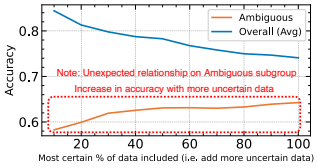


Figure 8: Data-IQ subgroups can unmask unreliable prediction deferral by uncertainty methods

5 Discussion

In this paper, we introduce Data-IQ, a systematic framework that can be used with *any* ML model with checkpoints, to characterize examples into subgroups with respect to the outcome. Through several experiments, we demonstrate that the usage of aleatoric uncertainty, which captures properties more inherent to the data, is indeed more principled, being more robust to variation across models and/or parameterizations. Data-IQ's consistency is unmatched by any compared baseline. Data-IQ should not automate and replace the intuition of a data scientist. Rather, as we have demonstrated, Data-IQ should serve as a systematic "data-centric ML" tool that assists and empowers data scientists with the "data" work at training time, whilst also guiding reliable model usage at deployment time.

Finally, we outline limitations and future opportunities. ① While Data-IQ characterizes examples; the current formulation does not allow us to understand which attributes are responsible for the characterization per example. This would be an interesting extension around dataset explainability, allowing practitioners to better probe their data. ② In high-stakes settings such as healthcare, to mitigate possible adverse effects (e.g. difficulty of Easy vs Hard), Data-IQ should be used with a "human-in-the-loop", allowing experts to complement and validate findings with domain knowledge.

References

[1] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021.

[2] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019.

[3] John C Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally robust losses against mixture covariate shifts.

[4] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018.

[5] Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33:19339–19352, 2020.

[6] Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, and Kilian Q Weinberger. Identifying mislabeled data using the area under the margin ranking. *Advances in Neural Information Processing Systems*, 33:17044–17056, 2020.

[7] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *arXiv preprint arXiv:2110.01889*, 2021.

[8] Jinsung Yoon, Yao Zhang, James Jordon, and Mihaela van der Schaar. Vime: Extending the success of self-and semi-supervised learning to tabular domain. *Advances in Neural Information Processing Systems*, 33, 2020.

[9] Zepeng Huo, Lida Zhang, Rohan Khera, Shuai Huang, Xiaoning Qian, Zhangyang Wang, and Bobak J Mortazavi. Sparse gated mixture-of-experts to separate and interpret patient heterogeneity in ehr data. In *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–4. IEEE, 2021.

[10] Gerald Gartlehner, Suzanne L West, Alyssa J Mansfield, Charles Poole, Elizabeth Tant, Linda J Lux, and Kathleen N Lohr. Clinical heterogeneity in systematic reviews and health technology assessments: synthesis of guidance documents and the literature. *International Journal of Technology Assessment in Health Care*, 28(1):36–43, 2012.

[11] Eun Jeong Oh, Ravi B Parikh, Corey Chivers, and Jinbo Chen. Two-stage approaches to accounting for patient heterogeneity in machine learning risk prediction models in oncology. *JCO Clinical Cancer Informatics*, 5:1015–1023, 2021.

[12] Weijie Jiang and Zachary A Pardos. Towards equity and algorithmic fairness in student grade prediction. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 608–617, 2021.

[13] Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. Ethical machine learning in healthcare. *Annual Review of Biomedical Data Science*, 4:123–144, 2021.

[14] Tin Kam Ho and Mitra Basu. Complexity measures of supervised classification problems. *IEEE transactions on pattern analysis and machine intelligence*, 24(3):289–300, 2002.

[15] Ana C. Lorena, Luís P. F. Garcia, Jens Lehmann, Marcilio C. P. Souto, and Tin Kam Ho. How complex is your classification problem? a survey on measuring classification complexity. *ACM Comput. Surv.*, 52(5), sep 2019.

[16] Abhinav Jain, Hima Patel, Lokesh Nagalapatti, Nitin Gupta, Sameep Mehta, Shanmukha Guttula, Shashank Mujumdar, Shazia Afzal, Ruhi Sharma Mittal, and Vitobha Munigala. Overview and importance of data quality for machine learning tasks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3561–3562, 2020.

[17] Nitin Gupta, Hima Patel, Shazia Afzal, Naveen Panwar, Ruhi Sharma Mittal, Shanmukha Guttula, Abhinav Jain, Lokesh Nagalapatti, Sameep Mehta, Sandeep Hans, et al. Data quality toolkit: Automatic assessment of data quality and remediation for machine learning datasets. *arXiv preprint arXiv:2108.05935*, 2021.

[18] Cedric Renggli, Luka Rimanic, Nezihe Merve Gürel, Bojan Karlas, Wentao Wu, and Ce Zhang. A data quality-driven view of mlops. *IEEE Data Engineering Bulletin*, 2021.

[19] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

[20] Kiri Wagstaff. Machine learning that matters. *arXiv preprint arXiv:1206.4656*, 2012.

[21] Andrew Ng, Lora Aroyo, Cody Coleman, Greg Diamos, Vijay Janapa Reddi, Joaquin Vanschoren, Carole-Jean Wu, and Sharon Zhou. Neurips data-centric ai workshop, 2021.

[22] Neoklis Polyzotis and Matei Zaharia. What can data-centric ai learn from data and ml engineering? *arXiv preprint arXiv:2112.06439*, 2021.

[23] Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, 2020.

[24] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

[25] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009.

[26] Efstratios Gavves. Bayesian deep learning, lecture 10, university of amsterdam, November 2019.

[27] Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in Neural Information Processing Systems*, 34, 2021.

[28] Sezal Chug, Priya Kaushal, Ponnurangam Kumaraguru, and Tavpritesh Sethi. Statistical learning to operationalize a domain agnostic data quality scoring. *arXiv preprint arXiv:2108.08905*, 2021.

[29] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017.

[30] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019.

[31] Yuanzhi Li, Tengyu Ma, and Hongyang R Zhang. Learning over-parametrized two-layer neural networks beyond ntk. In *Conference on learning theory*, pages 2613–2682. PMLR, 2020.

[32] Yongchan Kwon, Joong-Ho Won, Beom Joon Kim, and Myunghee Cho Paik. Uncertainty quantification using bayesian neural networks in classification: Application to biomedical image segmentation. *Computational Statistics & Data Analysis*, 142:106816, 2020.

[33] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018.

[34] Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.

[35] Pedro Baqui, Ioana Bica, Valerio Marra, Ari Ercole, and Mihaela van Der Schaar. Ethnic and regional variations in hospital mortality from covid-19 in brazil: a cross-sectional observational study. *The Lancet Global Health*, 8(8):e1018–e1026, 2020.

[36] Máire A Duggan, William F Anderson, Sean Altekruse, Lynne Penberthy, and Mark E Sherman. The surveillance, epidemiology and end results (seer) program and pathology: towards

strengthening the critical relationship. *The American journal of surgical pathology*, 40(12):e94, 2016.

[37] CUTRACT Prostate Cancer UK. Prostate cancer uk.

[38] William A Knaus, Frank E Harrell, Joanne Lynn, Lee Goldman, Russell S Phillips, Alfred F Connors, Neal V Dawson, William J Fulkerson, Robert M Califf, Norman Desbiens, et al. The support prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of internal medicine*, 122(3):191–203, 1995.

[39] Diogo Ayres-de Campos, Joao Bernardes, Antonio Garrido, Joaquim Marques-de Sa, and Luis Pereira-Leite. Sisporto 2.0: a program for automated analysis of cardiotocograms. *Journal of Maternal-Fetal Medicine*, 9(5):311–318, 2000.

[40] Pantelis Koutroumpis, Aija Leiponen, and Llewellyn DW Thomas. Markets for data. *Industrial and Corporate Change*, 29(3):645–660, 2020.

[41] Ramesh Raskar, Praneeth Vepakomma, Tristan Swedish, and Aalekh Sharan. Data markets to support ai for all: Pricing, valuation and governance. *arXiv preprint arXiv:1905.06462*, 2019.

[42] Mohammad Rasouli and Michael I Jordan. Data sharing markets. *arXiv preprint arXiv:2107.08630*, 2021.

[43] Khaled El Emam, Lucy Mosquera, and Richard Hoptroff. *Practical synthetic data generation: balancing privacy and the broad availability of data*. O'Reilly Media, 2020.

[44] Andre Goncalves, Priyadip Ray, Braden Soper, Jennifer Stevens, Linda Coyle, and Ana Paula Sales. Generation and evaluation of synthetic patient data. *BMC medical research methodology*, 20(1):1–40, 2020.

[45] Ahmed Alaa, Boris van Breugel, Evgeny Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models, 2022.

[46] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in Neural Information Processing Systems*, 32, 2019.

[47] Zheng Li, Yue Zhao, and Jialin Fu. Sync: A copula based framework for generating synthetic data from aggregated sources. In *2020 International Conference on Data Mining Workshops (ICDMW)*, pages 571–578. IEEE, 2020.

[48] Cristóbal Esteban, Stephanie L Hyland, and Gunnar Rätsch. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633*, 2017.

[49] Michael Platzer and Thomas Reutterer. Holdout-based empirical assessment of mixed-type synthetic data. *Frontiers in big Data*, page 43, 2021.

[50] 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.

[51] Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific reports*, 7(1):1–14, 2017.

[52] Tong Xia, Jing Han, Lorena Qendro, Ting Dang, and Cecilia Mascolo. Uncertainty-aware covid-19 detection from imbalanced sound data. *arXiv preprint arXiv:2104.02005*, 2021.

[53] Ranganath Krishnan and Omesh Tickoo. Improving model calibration with accuracy versus uncertainty optimization. *Advances in Neural Information Processing Systems*, 33:18237–18248, 2020.

[54] Jishnu Mukhoti and Yarin Gal. Evaluating bayesian deep learning methods for semantic segmentation. *arXiv preprint arXiv:1811.12709*, 2018.

[55] Soumya Ghosh, Jiayu Yao, and Finale Doshi-Velez. Structured variational learning of bayesian neural networks with horseshoe priors. In *International Conference on Machine Learning*, pages 1744–1753. PMLR, 2018.

[56] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.

[57] Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning*, pages 2242–2251. PMLR, 2019.

[58] Gaojie Jin, Xinping Yi, Liang Zhang, Lijun Zhang, Sven Schewe, and Xiaowei Huang. How does weight correlation affect generalisation ability of deep neural networks? *Advances in Neural Information Processing Systems*, 33:21346–21356, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] Please refer to our "contributions" paragraph in Section 1. Furthermore, all claims made about Data-IQ are verified in Section 4 and in the supplementary material.
   (b) Did you describe the limitations of your work? [Yes] See Section 3.3 and Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Data-IQ provides practioners with a tool to help them with their "data" work and to better understand the characteristics of their data. In fact, Data-IQ may help reveal the negative impacts of ML models. However, as discussed in Section 5, expert knowledge should still be used to validate our characterization.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We have carefully read the ethics review guideline and confirm that our paper respects the guidelines.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 3 and Appendix A of the supplementary material
   (b) Did you include complete proofs of all theoretical results? [Yes] See Section 3 and Appendix A of the supplementary material
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] Will be released upon acceptance
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix B, detailing all relevant information
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Included as relevant to the experiments in Section 4 and Appendix C
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix B
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See citations [35], [36]. [37], [38],[39]. We cite these in the main paper and supplementary material
   (b) Did you mention the license of the assets? [Yes] See Appendix B of the supplementary material.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A] We only use existing assets
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] It is discussed in detail in the publications by the dataset creators, which we have cited
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Appendix B of the supplementary material.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]