A Case for Data Valuation Transparency via DValCards

Keziah Naggita[1], **Julienne LaChance**[2]

[1]Toyota Technological Institute at Chicago
[2]SONY AI
knaggita@ttic.edu, Julienne.LaChance@sony.com

Abstract

Following the rise in popularity of data-centric machine learning (ML), various data valuation methods have been proposed to quantify the contribution of each datapoint to desired ML model performance metrics (e.g., accuracy). Beyond the technical applications of data valuation methods (e.g., data cleaning, data acquisition, etc.), it has been suggested that within the context of data markets, data buyers might utilize such methods to fairly compensate data owners. Here we demonstrate that data valuation metrics are inherently biased and unstable under simple algorithmic design choices, resulting in both technical and ethical implications. By analyzing 9 tabular classification datasets and 6 data valuation methods, we illustrate how (1) common and inexpensive data preprocessing techniques can drastically alter estimated data values; (2) subsampling via data valuation metrics may increase class imbalance; and (3) data valuation metrics may undervalue underrepresented group data. Consequently, we argue in favor of increased transparency associated with data valuation in-the-wild and introduce the novel Data Valuation Cards (DValCards) framework towards this aim. The proliferation of DValCards will reduce misuse of data valuation metrics, including in data pricing, and build trust in responsible ML systems.

1 Introduction

Recently, focus has shifted away from model-centric machine learning (ML) in favor of data-centric ML, whereby increased emphasis is placed on the importance of meaningful, high-quality data to a desired ML output (Singh 2023). Within this paradigm, data valuation methods quantify the contribution of each datapoint (i.e., datum) to a given ML model performance metric (e.g., accuracy, loss, or a fairness measure such as equalized odds) (Ghorbani and Zou 2019; Cook and Weisberg 1980; Arnaiz-Rodriguez and Oliver 2023; Pang et al. 2024; Wang, Wu, and He 2024). Increasingly, data valuation metrics as *influence functions* are utilized for various technical applications (Hammoudeh and Lowd 2024; Sim, Xu, and Low 2022; Fleckenstein, Obaidi, and Tryfona 2023), including data cleaning and subsampling (Yoon, Arik, and Pfister 2020; Ghorbani and Zou 2019; Koh and Liang 2017; Kwon and Zou 2021; Tang et al. 2021), data acquisition (Ghorbani and Zou 2019; Kwon and Zou 2021; Jia et al. 2021), feature attribution (Chen et al. 2023; Zhao et al. 2024), and active learning (Ghorbani, Zou, and Esteva 2022), with the specific application scenario influencing the choice of valuation function (Sim, Xu, and Low 2022). Additionally, data valuation techniques have been reappropriated to measure or modify the algorithmic fairness of ML systems (Black and Fredrikson 2021; Arnaiz-Rodriguez and Oliver 2023; Pang et al. 2024; Wang, Wu, and He 2024). Within the context of data markets[1], it has been proposed that data buyers utilize data valuation methods for data pricing estimation in order to fairly compensate data owners according to their individual impact on model performance (Laoutaris 2019; Paraschiv and Laoutaris 2019; Jia et al. 2019b,a). However, the practical limitations of in-the-wild data valuation are not yet well exposed.

Here, we highlight the inherent properties of data valuation metrics, notably bias and instability under simple algorithmic design choices, by examining diverse case studies. These experiments aim to address pragmatic questions: (1) Do standard data preprocessing techniques predictably alter data values?; (2) What are the technical side-effects of modifications to an ML system via data valuation? For instance: can data cleaning augment class imbalance?; and (3) What are the ethical side-effects of such modifications? Namely: are members of underrepresented groups more likely to yield undervalued data? Taken together, the context-dependent implications of these results underscore the need for increased transparency regarding data valuation in-the-wild. Alternatively, the properties of data valuation metrics may limit their applicability to specific tasks entirely, as we argue in the case of data market pricing. Ultimately, we address the transparency gap by proposing a framework that we call DValCards, which accompany applications of data values and report the intended use, design choices, performance, and other critical information. We hope that the use of DValCards facilitates communication between creators, users, and affected parties of data valuation metrics, thereby encouraging appropriate use of the technology.

Code — https://github.com/knaggita/DValCards-Case

[1]A platform or mechanism that facilitates the exchange of data between buyers and sellers (Tian et al. 2022).

1.1 Related Work

Data valuation. The data valuation metrics we consider (leave one out (LOO), Truncated monte-carlo Shapley (TMC-Shapley), gradient Shapley (G-Shapley), etc.) were contextualized into an influence function taxonomy by Hammoudeh and Lowd (2024) and are introduced here Section 2. Prior works have analyzed known limitations of data valuation methods with some proposing novel variants which attempt to address them. Zhou et al. (2023) find that Shapley estimators do not necessarily satisfy the fairness properties of true Shapley values. Schoch, Xu, and Ji (2022) develop a Shapley-based metric which better discriminates between in- and out-of-class contributions; here, we further analyze their method according to its impact on class imbalance. Ghorbani, Kim, and Zou (2020) propose a distributional Shapley framework to augment stability of data values under perturbations. Wang et al. (2024) show that when applied to data selection, Data Shapley may perform no better than random selection without specific constraints on utility functions: for instance, when applied to homogeneous data. Wang and Jia (2023) discuss the instability of data value rankings across different model runs and propose a more robust data valuation metric; however, we demonstrate that their method (Banzhaf) still exhibits rank instability across algorithmic design choices.

More generally, we focus specifically on LOO, Banzhaf values and Shapley-based values due to their popularity in real-world applications. Modeling choices have been found to result in varied feature attributions, with the specific task better informing the choice of Shapley-based approach (Chen et al. 2020). More efficient Shapley value estimation methods have been proposed, e.g. (Covert and Lee 2021; Chen et al. 2018; Kwon, Rivas, and Zou 2021; Jethani et al. 2021). Yona, Ghorbani, and Zou (2021) propose an extended Shapley method addressing joint credit assignment, and data valuation metrics have been extended to the federated learning setting, e.g. (Wang et al. 2020; Liu et al. 2022; Song, Tong, and Wei 2019; Jiang et al. 2023).

AI/ML transparency frameworks. Modern ML transparency documentation frameworks are largely inspired by early documentation strategies including Data statements for natural language processing (Bender and Friedman 2018), Datasheets for datasets (Gebru et al. 2021), and Model cards for model reporting (Mitchell et al. 2019). Existing frameworks are designed to enable users to comprehensively report essential characteristics of ML data, models, methods, or systems, and often cite similarities to nutrition labels or engineering datasheets (Chmielinski et al. 2022; Krasin et al. 2017; Arnold et al. 2019). Frameworks may be contextualized for specific domains or applications, such as Healthsheets for healthcare applications (Rostamzadeh et al. 2022), Reward reports for reinforcement learning (Gilbert et al. 2023), or the Foundation Model Transparency Index (Bommasani et al. 2023). Human-centric elements may be included for data reporting, such as the annotator demographic information recommended by Díaz et al. (2022). Data values are distinct from prior subjects of transparency documentation for a number of reasons, making existing frameworks inadequate for data value reporting; this is discussed in more detail in Section 4.

2 Methodology

Experimental overview. In this paper, we restrict our attention to the task of supervised classification. Let $\mathcal{D} = \{z_i = (\mathbf{x}_i, y_i)\}_{i=1}^n$ denote the training data, where $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d$ are the features and $y_i \in \mathcal{Y}$ is the target class of the datum, z_i. Assume the model, $\mathcal{A}$, is trained on a subset of the data, $\mathcal{S} \subseteq \mathcal{D}$, to optimize the selected utility function, $\mathcal{V}(\mathcal{S}, \mathcal{A}) : 2^n \rightarrow \mathbb{R}$, where 2^n is the collection of all subsets of $\mathcal{D}$, including the empty set. To simplify notation, let $\mathcal{V}(\mathcal{S})$ denote $\mathcal{V}(\mathcal{S}, \mathcal{A})$. Throughout the paper, $\mathcal{V}(\mathcal{S})$ denotes the accuracy of the model on the validation (test) set, when trained on $\mathcal{S}$.

We utilize three diverse experiments as illustrative case studies, specifically:

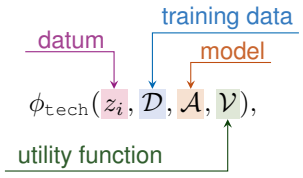
1) **Metric instability:** 12 data imputation methods are applied as preprocessing techniques to 9 tabular datasets which are then used to train supervised classification models. The corresponding data values for each condition are reported using 4 data valuation metrics.
2) **Class imbalance:** 4 data valuation metrics are used to subsample data from 9 tabular datasets. The class imbalance is reported before and after subsampling using the balance estimates described in Appendix Section D.2 in the arXiv version of the paper (Naggita and LaChance 2025) (hereafter, we will only use Appendix X).
3) **Underrepresented group bias:** 4 tabular datasets were analyzed to identify the prevalence of underrepresented attribute groups and their impact on 4 data valuation methods. Group and attribute representation is reported before and after subsampling using the balance estimates described in Appendix Section D.4.

Datasets. We selected 9 real-world, permissively licensed (CC BY), tabular classification datasets from the OpenML-CC18 benchmark (Creative Commons License; Bischl et al. 2019). Dataset selection criteria are detailed in Appendix Section C.1. The datasets are reported by OpenML-CC18 labels: **18** (Mfeat-morphological), **23** (Contraceptive method choice), **31** (German credit), **37** (Pima Indians diabetes database), **54** (Vehicle silhouette), **1063** (KC2 Software defect prediction), **1068** (PC1 Software defect prediction), **1480** (Indian liver patient) and **40994** (climate-model-simulation-crashes). Basic dataset characteristics are listed in Appendix Table 1.

Data preprocessing. To test the impact of data imputation methods on data valuation metrics, we utilize tabular datasets with no missing values and induce missingness according to three percentages ($1\%, 10\%$ and 30%) and three patterns (missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR)), as defined in Appendix Section A.1. Then we perform data imputation using 12 methods: row removal (i.e., discard all rows with any missing data values), column removal (i.e., remove attribute with missing data values), mean (i.e., replace a missing value with the mean of that attribute), mode (i.e.,

replace a missing value with the most frequent values within the attribute), k-nearest neighbor (KNN) (Murti et al. 2019), optimal transport (OT) (Muzellec et al. 2020), random sampling (i.e., randomly select samples from the attribute to fill the missing value), multivariate imputation by chained equations (MICE) (Van Buuren and Groothuis-Oudshoorn 2011), linear interpolation (Huang 2021), linear round robin (LRR) (Muzellec et al. 2020), MLP round robin (MLP RR) (Muzellec et al. 2020), and random forest (RF) (Hong and Lynn 2020). We include supplemental details in Appendix Section C.2.

Data valuation. The objective of the data valuation approach is to compute the datum value that reflects the marginal contribution of the datum to $\mathcal{V}$. Let the value of the datum z_i to $\mathcal{V}$ be given by:



(1)

where $tech$ is the datum valuation approach used to compute value of the datum. For simplicity, we use $\phi_{\texttt{tech}}(z_i)$ to denote $\phi_{\texttt{tech}}(z_i, \mathcal{D}, \mathcal{A}, \mathcal{V})$. In general, $\mathcal{V}(\mathcal{S}) - \mathcal{V}(\mathcal{S} \setminus \{z_i\})$ is defined as the marginal contribution of a datum to the utility function, $\mathcal{V}$. Different valuation approaches have different variants of this formulation.

For all experiments, we evaluate the data valuation approaches: truncated Monte Carlo Shapley (TMC-Shapley) (Ghorbani and Zou 2019), gradient Shapley (G-Shapley) (Ghorbani and Zou 2019), and leave one out (LOO) (Cook and Weisberg 1980). See Appendix Sections A.2 for method descriptions and C.3 for learning algorithm and additional details. Additionally, we analyze Banzhaf (Wang and Jia 2023) with respect to metric instability, class-wise Shapley (CS-Shapley) (Schoch, Xu, and Ji 2022) for class imbalance analysis, and FairShap (Arnaiz-Rodriguez and Oliver 2023) fairness-based metrics exclusively for the fairness experiment, beyond the standard metrics.

3 Results

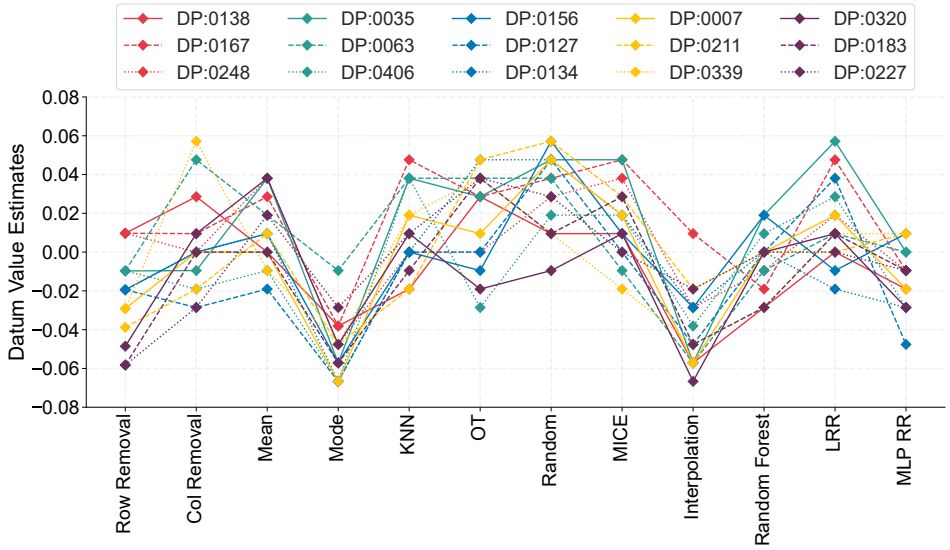
3.1 Metric Instability Case Study: Data Imputation

We find that varying the applied data imputation method results in appreciable variation of data values, with all other experimental conditions held constant (see Figure 1). Notably, the data rank order change is statistically significant when cross-comparing data values corresponding to any two differing imputation methods, according to Kendall's τ coefficient: $\tau < 1$ and $p < 0.05$ (Kendall 1938). This trend holds across all the data valuation methods considered (TMC-Shapley, G-Shapley, LOO and Banzhaf); see Appendix Figure 10.

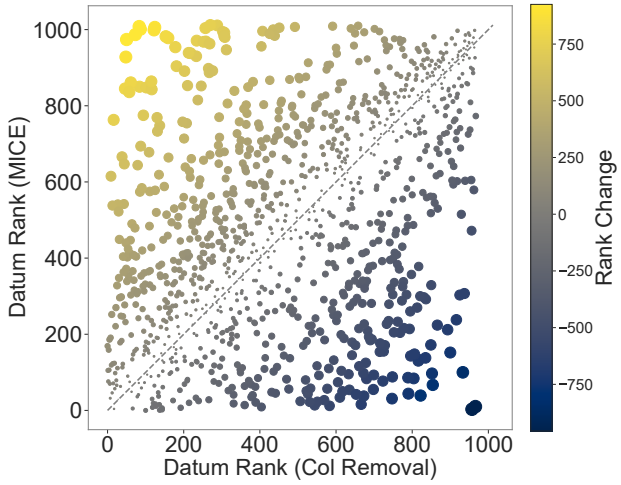
Variance in data values. Figure 1a shows a snapshot of 15 fixed data points in dataset 1063 (KC2 Software defect prediction, with MNAR-1) and the variance of leave-one-out (LOO) data values after various imputation methods were performed. The points were selected as non-imputed values spanning the quartiles of the initial data value set condition (MNAR, row removal), with three points per bin. This snapshot demonstrates a crucial point: that data values can vary significantly according to imputation method applied; not only in an absolute sense (which may be meaningless to compare cross-system), but even causing drastic relative changes in score for non-imputed data points. For completeness, mean data values are reported systematically across all imputation methods and data valuation metrics considered in Appendix Figure 11. In general, utilizing data imputation methods tends to increase the average data value (see Appendix Table 2a and Figure 13) and, in some cases, the maximum data value (see Appendix Table 2b). This further shows that common and inexpensive data pre-processing techniques can drastically alter estimated data values.

Variance in data rank. To illustrate datum rank changes across all data points for a single dataset following common and low-cost preprocessing, we select two imputation methods (column removal and MICE) and cross-compare how datum rank is impacted for all data values in dataset 23 (Contraceptive method choice, with TMC and MAR-30). These results are shown in Figure 1b; we would expect a stable valuation metric to reasonably maintain consistent rank scores, and display a trend along the diagonal (shown as the grey dotted line). The wide variability of rank scores in this case study suggests that data value instability may not uniquely impact high or low data values. To better systematically assess rank order changes, we report the Kendall's τ coefficient across each pair of imputation methods acting on dataset 37 (Pima Indians Diabetes Database, with MNAR-10) for all data valuation metrics considered in Appendix Figure 10. We find that the imputation method of row removal and the data valuation metric LOO are associated with significant rank changes in comparative analyses across imputation strategy.

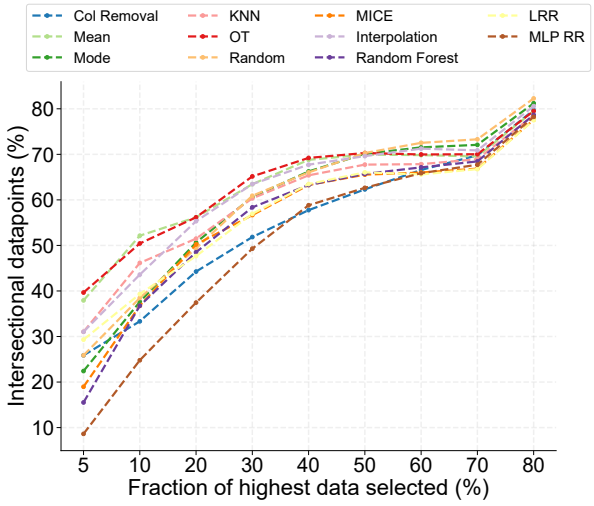
Implications to data subsampling. Indeed, for many practical applications of data valuation metrics, the data values are used to select a subset of the initial dataset according to highest or lowest data values, such as in data cleaning. Thus, we ask: are data valuation metrics capturing the same points as the sub-selected data fraction varies, if only the imputation method is modified? We show that the same points are not necessarily captured in Figure 1c; in this, we present the percentage of data points captured by TMC values following applications of different imputation methods when compared to the baseline method, row removal (dataset 23, MCAR-30). Analogous plots with both the highest- and lowest-valued data fractions are shown for each of the metrics considered in Appendix Figure 14. Moreover, data values assessed prior to imputation could lead to the premature disposal of otherwise high-valued data as assessed post-imputation. In the following section, we explore class-based implications of value-based data subsampling.



(a) Variation in data values for fixed data points by imputation method



(b) Cross-comparison of datum rank values by imputation method



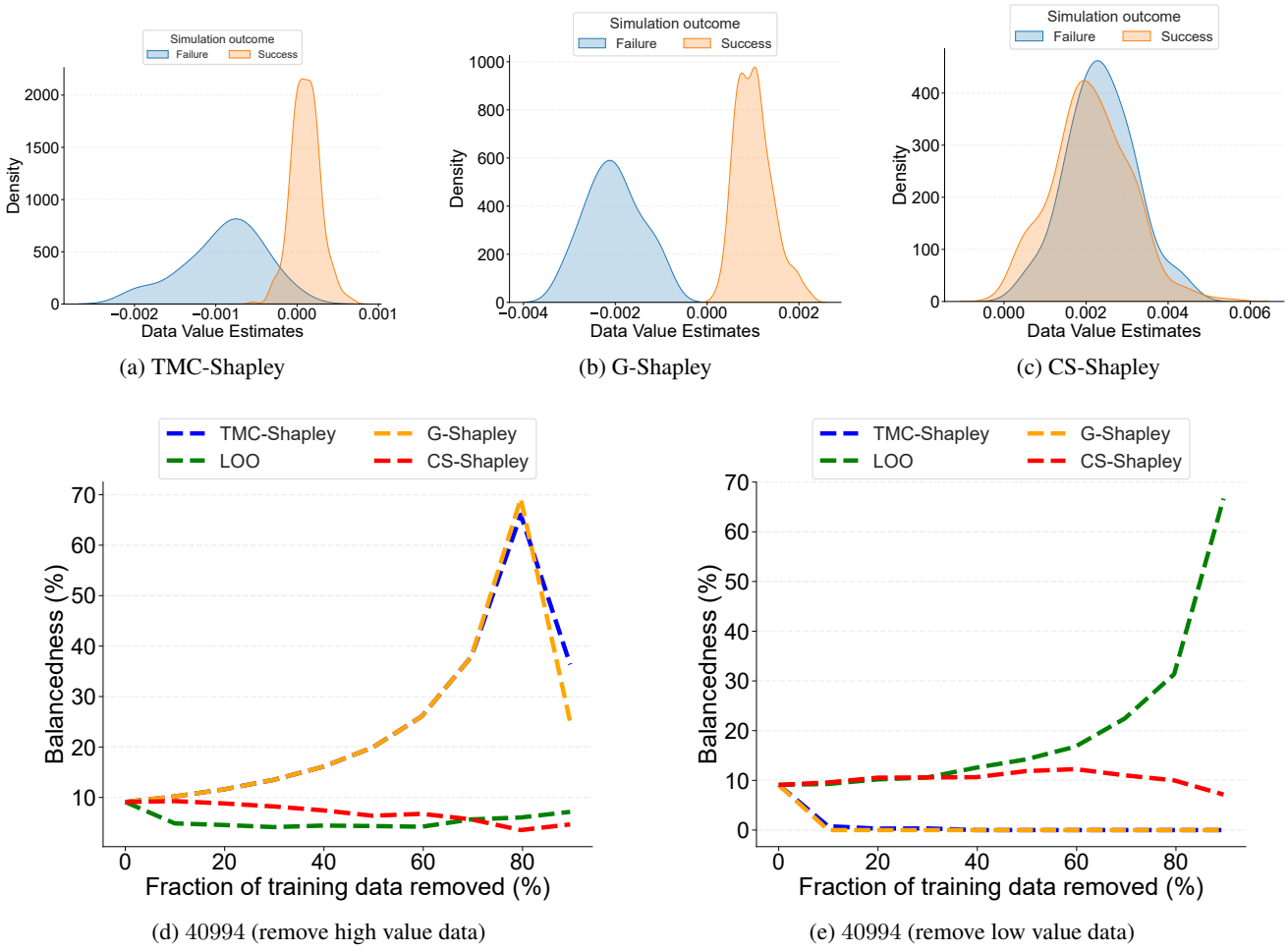
(c) Variations in value-selected data subsets by imputation method

Figure 1: Data values are unstable to choice of data preprocessing method. (**a**) Leave-one-out (LOO) value estimates vary as a function of imputation method; data points are selected to span 5 quintiles of data value scores from the row removal results (grouped by color). By cross-comparing value estimates by imputation method, it is clear that value rank order varies in addition to raw values. (**b**) TMC-Shapley value-based data ranks are compared across two different imputation methods (column removal and MICE) to assess agreement. The Kendall $\tau = 0.3214$, p-value < 0.05, indicating a statistically significant positive correlation but not agreement, as can be observed by the scattering of points away from the diagonal (grey dashed line). Point size indicates scale of rank change; see also Appendix Figure 12 for changes in value and rank across all points. (**c**) The percentage of shared points between high data value sets as a function of various imputation methods and row removal as the baseline method. Analogous plots including low data value selection are provided in Appendix Figure 14.

3.2 Technical Impact Case Study: Class Imbalance

We find that the distribution of data values can vary greatly as a function of class membership and data valuation metric. As a result, data value-based subsampling may increase class imbalance.

Data value distributions may be class-dependent. We observe that most standard data valuation metrics exhibit class-based bias, with sample results shown in Figures 2a and 2b for TMC-Shapley and G-Shapley. All results in Figure 2 are shown for dataset 40994 (climate-model-simulation-crashes) under MCAR-10 and random imputation. Notably, the associated binary classes are imbal-



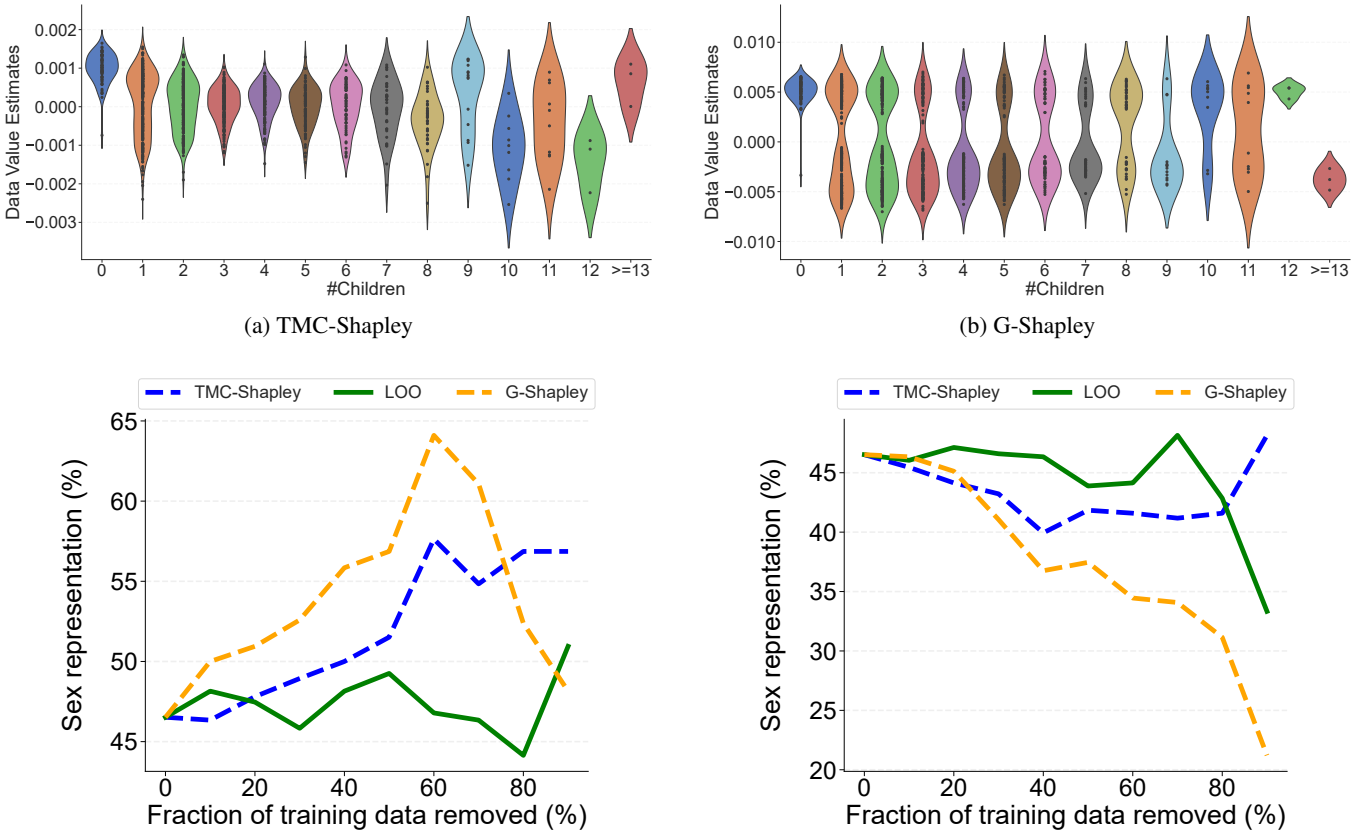
(a) TMC-Shapley (b) G-Shapley (c) CS-Shapley

(d) 40994 (remove high value data) (e) 40994 (remove low value data)

Figure 2: Data values and class imbalance. (**a**, **b**, and **c**) Data value distributions according to three valuation metrics (TMC-Shapley, G-Shapley, and CS-Shapley, respectively), for a binary classifier with 91.3% simulation-outcome *success* (dataset 40994, MCAR-10). We observe marked class-based differences in data value distributions for TMC-Shapley and G-Shapley; by contrast, class-wise Shapley (CS-Shapley) improves consistency between classes. (**d**, **e**) Class balance (as defined in Appendix D.2 as b) versus percentage of data removed, as a function of four data valuation metrics (TMC-Shapley, G-Shapley, LOO and CS-Shapley). Value-based data subsampling may impact class imbalance. In this example, TMC-Shapley and G-Shapley increase class balance with removal of high-value data and decrease balance with removal of low-valued data.

anced, with the larger class ("simulation success") comprising 91.3% of the data. In Figures 2a and 2b, the Shapley-based data valuation metrics can be seen to produce lower data values for the less frequent class ("simulation failure", blue) than the more frequent class ("simulation success", orange). By contrast, the application of the class-wise Shapley (CS-Shapley) metric reduces the class-based bias on the same data: see Figure 2c, in which the distinct classes correspond to similar data value distributions. This trend is unsurprising, as CS-Shapley was developed to better discriminate between training instances' in-class and out-of-class contributions to a classifier. However, the differences observed across data valuation metrics indicate the utility of clear transparency documentation, especially given the impact of the choice of data valuation method on other performance metrics. Additional class-based value distribution plots are shown in Appendix Figure 15 for diverse datasets and metrics.

Value-based subsampling may impact class balance. To illustrate how class imbalance can change as a function of value-based data subsampling, we show class balancedness (defined in Appendix D.2 as $b \in [0, 1]$ where $b = 0$ indicates at least one class is entirely absent from the training set, and $b = 1$ denotes perfect class balance) as a function of percentage removed data for each metric, e.g. in Figures 2d and 2e. Given the same initial dataset with imbalanced classes, we observe that TMC-Shapley and G-Shapley result in reduced class balance as low-valued data is removed; this is indicative of data removal from the lower-valued, smaller class ("simulation failure") corresponding to the value distributions shown in Figures 2a and 2b. The opposite trend



(a) TMC-Shapley

(b) G-Shapley

(c) 31 LRR (sex) remove high value data

(d) 31 LRR (sex) remove low value data

Figure 3: Data values and attribute group. (**a**, **b**) Data value distributions according to two valuation metrics (TMC-Shapley and G-Shapley, respectively), by attribute group "number of children ever born" (dataset 23, MAR-10). We observe marked attribute-based differences in data value distributions with variance across valuation metric. In (**a**), removal of low-valued data may disproportionately remove data from underrepresented attribute groups (i.e, greater "number of children ever born").
(**c**, **d**) Percentage binary sex representation (as defined in Appendix D.4 as g) versus percentage of data removed, as a function of three data valuation metrics (TMC-Shapley, LOO, and G-Shapley). Value-based data subsampling may impact attribute imbalance. In this example, TMC-Shapley and G-Shapley tend to increase percentage sex representation with removal of high-value data and decrease representation with removal of low-valued data.

holds as high-valued data is removed, indicative of data pulled from the majority class, until an inflection point is reached. By contrast, CS-Shapley results in a relatively consistent class balance when either high- or low-valued data is removed. LOO results in reduced class balance as high-valued data is removed and increased class balance as low-valued data is removed. Analogous plots for diverse datasets and imputation methods may be found in Appendix Figure 16. We systematically review all datasets, imputation methods, missingness conditions and value metrics according to their impact on class balance following subsampling in Appendix Tables 3 and 4, corresponding to removal of low- or high-valued data, respectively. Results are reported according to absolute class balance scores (i.e., balancedness < 0.25) and to relative class balance with respect to the original dataset. We find that when 20% of low-valued data is removed, the absolute and relative class balance worsens for most datasets; the removal of high-valued data does not generally reduce class balance. Finally, the choice of metric may have diverse effects on downstream performance metrics, such as accuracy (see Appendix Figure 25) or attribute balance (see Section 3.3).

3.3 Ethical Impact Case Study: Potential Undervaluation of Marginalized Groups

We find that data valuation metrics may exhibit attribute-based bias as a function of dataset and preprocessing conditions. As a result, the choice of metric in the context of specific downstream applications, like data subsampling, may impact attribute balance in an unpredictable manner. When an attribute (e.g. skin tone, gender) denotes a sensitive characteristic associated with protected or marginalized groups of people, the potential for selective removal of these infrequent data samples is ethically (and possibly legally) problematic. Similar implications apply to other value-based applications including pricing in data markets.

Data values by attribute group. Our experiments show that data valuation metrics manifest distinct and potentially biased distributions across attribute groups. Two examples are provided in Figures 3a and 3b, which display the variance in TMC-Shapley and G-Shapley values according to attribute for dataset 23 (Contraceptive method choice, with MAR-10); here, the attribute of interest is "number of children ever born". In these, we observe distinct distributions for data value across the various attribute classes. Lower data values in TMC-Shapley were often associated with underrepresented groups (Figure 3a), i.e., with greater "number of children ever born"; thus downstream subsampling based on low-value data removal may pull relatively more data from these underrepresented groups. Heterogeneity of distributions according to attribute group may be observed under a multitude of experimental conditions: additional analogous plots to Figures 3a and 3b are shown in Appendix Figure 18. These show that CS-Shapley may also produce distinct distribution clusters for specific attributes, and thus a method chosen to protect class balance may still result in the selective removal of data from underrepresented attribute groups, or other issues caused by data undervaluation.

Subsampling and attribute balance. Sample plots in Figures 3c and 3d illustrate how attribute balance may be impacted by value-based subsampling. For dataset 31 (German credit, LRR, MCAR-30), we see that as an increasing fraction of low-valued data is removed, TMC- and G-Shapley tend to result in worsening female-to-male binary sex representation, with generally smaller effects resulting from LOO. Analogous plots to Figures 3c and 3d for diverse imputation methods can be found in Appendix Figure 20, and imputation method is systematically assessed for its impact on attribute balance and equalized-odds difference (EOD, a fairness metric) for age and sex in Appendix Figures 21 and 22, respectively. We cross-compare model accuracy with EOD for both binary sex and age in Appendix Figure 25 for dataset 31 (German credit, mean, MAR-30), as increasing fractions of data are removed, demonstrating that EOD is not necessarily correlated with changes in predictive accuracy.

For all missingness conditions, imputation methods and standard value metrics we present results in which subsampling improves EOD fairness for sex (see Appendix Table 5) and age (see Appendix Table 6) on dataset 31 (German credit, mean, MAR-30). From this systematic analysis we find that across all conditions, subsampling typically does not improve EOD fairness. Similarly, we assess the impact on attribute representation balance for all conditions, for sex (see Appendix Table 7) and age (see Appendix Table 8). The results are found to vary more widely for attribute representation balance, and this may be impacted by the initial attribute representation balance from the original dataset.

For comparison, we additionally show the distribution of data values by attribute group and class according to accuracy and three fairness metrics (equalized odds "Odds", average absolute equalized odds "Odds2", and equal opportunity "EOp") using the protocol described in Arnaiz-Rodriguez and Oliver (2023) on select datasets (see Appendix Figure 19 and Equation 7). As expected, the distribution of accuracy- and fairness-based values display distinct characteristics, as the removal of points of low influence to accuracy may negatively impact fairness outcomes; this is assessed systematically in Appendix Figures 23d and 24, across binary sex and age.

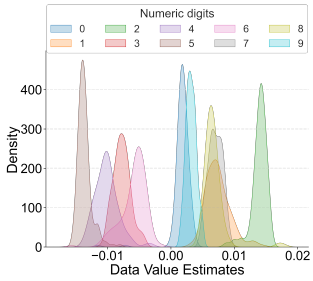
3.4 Fair Compensation

We briefly comment on the oft-cited recommendation that data valuation metrics be utilized as, or a major constituent of, a data pricing scheme (Sim, Xu, and Low 2022; Tian et al. 2022; Jia et al. 2019a; Agarwal, Dahleh, and Sarkar 2019; Azcoitia and Laoutaris 2022). Our results indicate that a naïve utilization of the LOO and Shapley-based metrics is unsuitable for establishing equitable compensation. In Section 3.1, we illustrate the instability of LOO, TMC-Shapley and G-Shapley to 12 common data preprocessing (imputation) methods. Such instability induces no confidence in data metrics as a pricing scheme; that is, it is unclear to data market participants how minor algorithmic design choices may impact data costs. Likewise, control over algorithmic design may provide data buyers with a mechanism by which to artificially adjust data prices to the detriment of data owners. In Section 3.3, we demonstrate the potential for attribute group bias in data values; as a data pricing scheme, this puts data buyers at risk of explicitly undervaluing data offered by members of marginalized groups or other "outlier" types. (Interestingly, such an effect could make homogeneous data more expensive from a buy-side perspective.) Notably, data valuation metrics are *unfair* by design, as evidenced by their utility for data outlier removal and cleaning.

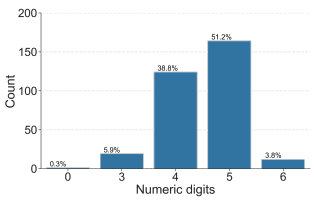
Furthermore, we argue that data valuation metrics lack properties of an effective economic pricing strategy: for instance, an inherent asymmetry is given to the seller, as data owners must submit their data in order to receive an assigned price. Prior works have highlighted this and a number of other practical challenges with the use of data valuation metrics as a pricing scheme, which include computational expense (Hammoudeh and Lowd 2024), the handling of replicated data (Xu et al. 2021; Agarwal, Dahleh, and Sarkar 2019; Wang and Jia 2023; Ohrimenko, Tople, and Tschiatschek 2019), the translation to a monetary value (Coyle and Manley 2023), asymmetry in data marketplace design (Azcoitia and Laoutaris 2022; Agarwal, Dahleh, and Sarkar 2019; Han et al. 2023), privacy leakage (Tian et al. 2022; Wang et al. 2023; Kang, Pedarsani, and Ramchandran 2024) and protections against strategic sellers (Castro Fernandez 2022; Agarwal, Dahleh, and Sarkar 2019). In many practical contexts, fair and consistent compensation may more readily be obtained by assigning data values *apriori* and decoupling values from learning algorithms and performance metrics.

4 DValCards for Data Valuation Transparency

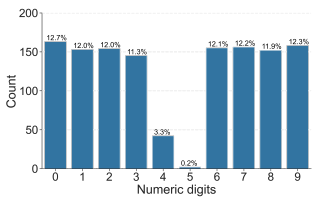
Given the limitations of data valuation metrics explored in previous sections, we propose a transparency framework to promote confidence in, and appropriate use of, such metrics.



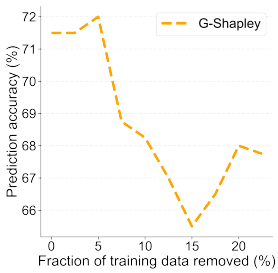
(a) Distribution of data values



(b) The distribution of numeric digits for the **excluded** instances



(c) The distribution of numeric digits for the **included** instances



(d) Low value data removal

Figure 4: Distribution of G-Shapley values before inclusion/exclusion (**a**), after removing the bottom 20% (**b**), and after selecting the top 80% (**c**). Lastly, (**d**) highlights the negative and unexpected impact of removing low-value datapoints from the Dataset 18 (Mfeat-morphological) training set. Results are used in DValCard (see Figure 5) sections descriptions.

DValCard: G-Shapley for Numeric-digits Prediction

Introduction. Developed by the paper authors in May 2024 for demonstration purposes, this DValCard presents results based on dataset **18** (Mfeat-morphological). Date (revised v2): May 2025.

System Flowchart

- **DVal in the life cycle context**. Data valuation is conducted during the data preprocessing stage of model training (see Figure 6). The DVal candidate data is a subset of, or equal to, the model training data.

DVal Candidate Data

- **Data information**. The DVal candidate data is drawn from the Multiple Features Morphological dataset (Duin 1998), which contains six morphological features representing handwritten digits (0–9) extracted from Dutch utility maps. The dataset includes approximately 10% of each digit class, totaling 1600 instances. While the original dataset contains no missing values, we introduced 1% random missingness pattern for experimental purposes.
- **Data preprocessing**. Interpolation imputation method (Huang 2021) is used to impute missing data.

DVal Method

- **DVal technique**. G-Shapley (Ghorbani and Zou 2019).
- **Learning algorithm**. Logistic regression model, with the linear *solver* and *max_iter* 5000 for data valuation and classification.
- **Performance metric**. Learning algorithm accuracy score: the sum of true positives and true negatives out of all algorithmic predictions.
- **Evaluation data**. The evaluation data, comprising 400 numeric digit instances, is drawn from the same distribution as the DVal candidate data. It represents a 20% split of the original dataset, while the DVal candidate data corresponds to the remaining 80%.

DVal Report

- **Data values**. The data values range from a minimum of -0.0156906 to a maximum of 0.0178875. The distribution of data values shows an over-representation of the numeric digits 3, 4, 5, and 6 in the discarded data (refer to Figures 4a and 4b).
- **Chosen/included instances**. Numeric digits 5 and 4 are underrepresented in the top 80% highest-value instances, comprising only 0.2% and 3.3% of the selected data, respectively (Figure 4c).

Ethical Statement and Recommendations

- **Intended users, and in/out-of-scope use cases**. The DvalCard is demonstrative and meant for researchers, engineers and all those interested in data valuation.
- **Potential ethical issues to consider**. The chosen data valuation method (technique) does not perform well on the intended task, as seen in Figure 4d. We caution against using it for data subsampling in a digit classification task on dataset **18** (Mfeat-morphological).
- **Legal considerations**. The DValCard and experimentation details are provided under the MIT license (MIT License).
- **Environmental considerations**. Computations were performed on a laptop CPU; full hardware specifications are in Appendix C.3. Data value computation took approximately 24 hours.
- **Recommendations**. Due to accuracy decreases resulting from application of value-based subsampling, we recommend the consideration of alternative data valuation methods, e.g. CS-Shapley to better handle imbalanced classes (Schoch, Xu, and Ji 2022).

Figure 5: Sample DValCard illustrating G-Shapley-based data valuation for sampling training data in a digit classification task.

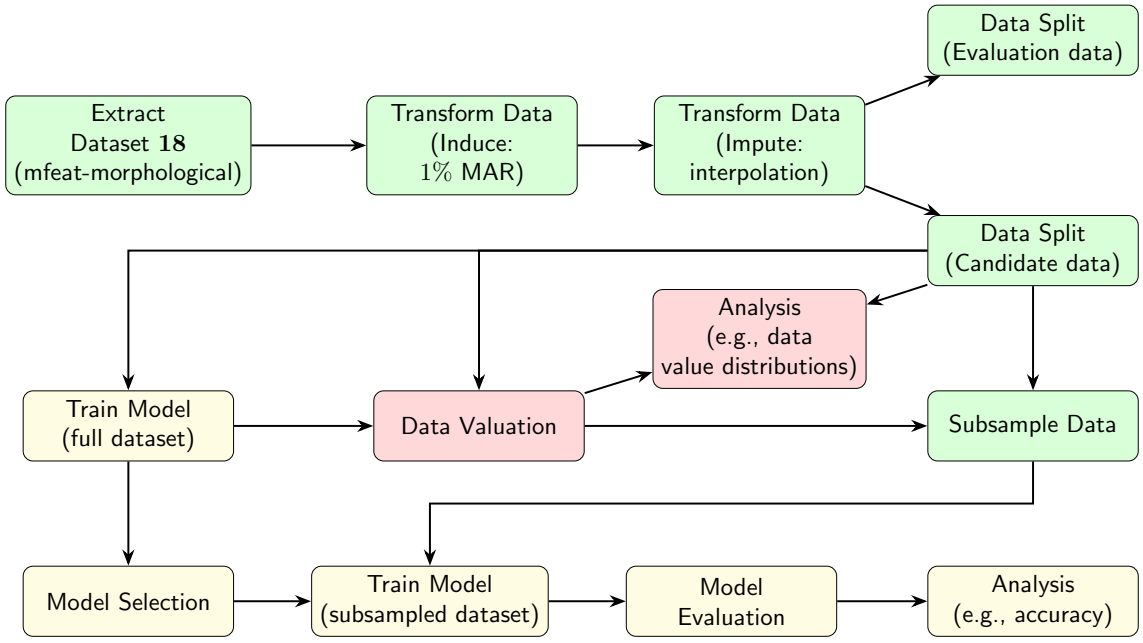


Figure 6: The system flowchart, a part of the DValCard (Figure 5), illustrates where in the model lifecycle data valuation could be applied in practice. Processes most closely associated with data in green; data valuation and analysis of data values in red; and the ML model in yellow. Evaluation data was used in both analysis processes (omitted arrows indicating this for clarity) to assess the contribution of DVal Candidate data to the trained model and to evaluate ML models trained on subsampled datasets.

There exist key differences between data valuation methods and the subjects of existing transparency documents: in particular, data values can (1) form part of the *data life cycle*; (2) form part of the *model life cycle*; or (3) be utilized as standalone measures. Within a data life cycle, data values may be used for dataset curation, e.g. in explanations of data diversity, density or association (Mitchell et al. 2023) or instance removal (Gebru et al. 2021). Within a model or system life cycle, data values are used for model training, e.g. for data weighting, selection, cleaning and preprocessing (Arnaiz-Rodriguez and Oliver 2023; Yoon, Arik, and Pfister 2020; Koh and Liang 2017; Kwon and Zou 2021; Tang et al. 2021; Ghorbani and Zou 2019; Kwon and Zou 2021). Furthermore, data values may be independently used for tasks including data pricing.

Consequently, existing transparency documents do not well capture the flexibility required for data valuation reporting: system cards (Alsallakh et al. 2022) assume the existence of ML models contained within a broader pipeline, while datasheets (Gebru et al. 2021) exclude models entirely, as examples.

Another key feature of data values is that accurate reporting of *when* values are computed is essential, with respect to other ML system components; in Section 3.1 we illustrate the impact of simple preprocessing choices on data values. This motivates our recommendation that DValCard authors include ML system flowcharts (e.g., in Figure 6) to clearly detail the order of operations. Correspondingly, certain performance measures, such as attribute balance, may change as the result of data value-based processes, such as value-based subsampling (see Section 3.3). Thus, we encourage reporting performance before and after applying data values.

Figure 5 illustrates an example of the proposed DValCard framework consisting of six sections highlighted in blue: "*Introduction*", "*System Flowchart*" (Figure 6), "*DVal Candidate Data*", "*DVal Method*", "*DVal Report*", "*Ethical Statement and Recommendations*". Appendix H presents the DValCard template, outlining its proposed general sections designed to flexibly incorporate the key components of the data valuation method(s), while enhancing transparency in the valuation process and accountability of the performance within the context of the intended application.

5 Limitations and Ethical Considerations

Our experiments primarily centered on a small set of data valuation metrics: TMC-Shapley, G-Shapley, and LOO. We selected these methods based on three criteria: they are the most frequently cited in the literature, serve as a foundation for many modern methods that often refine or address the limitations of these fundamental approaches (e.g., CS-Shapley), and are widely applied in data pricing and data markets, with Shapley values being particularly prominent. While alternative metrics may exist that better address some of the technical and ethical challenges we examine, transparency remains essential to foster clear communication between stakeholders in practice.

Moreover, our choice to highlight practical case studies is inherently restrictive; for example, we do not extend beyond the tabular supervised classification domain nor explore preprocessing methods beyond imputation. Additionally, the OpenML-CC18 benchmarking datasets we utilize do not have comprehensive associated transparency documentation (e.g., datasheets). Thus, in some settings, the exact provenance of the original data and the use of ethical curation practices remain unclear. To the best of our knowledge, we are the first to empirically study the practical limitations of data valuation in real-world use cases and propose a specific framework for data valuation transparency. We hope that future researchers can test the framework in practical applications.

Lastly, challenges may arise in enforcing the DValCards standard and incentivizing researchers and practitioners to adopt and implement the documentation effectively. The current proposed DValCard template aims to initiate a discussion and encourage practitioners and researchers to modify it to ensure accurate and comprehensive documentation of the data valuation process. With agreement on the standard, practitioners and researchers can integrate the DValCard into their documentation. We believe we can successfully follow a similar route taken by other documentation and transparency methods to incentivize researchers and practitioners to incorporate DvalCards into existing documentation frameworks.

6 Conclusion

We introduce the DValCards framework to support decision-making and promote the appropriate use of data valuation methods. Through three case studies, we demonstrate notable disparities of data valuation in practice: the variability in data values caused by common data preprocessing techniques, the influence of data values on class imbalances, and the disparate valuation of underrepresented attribute groups. We argue that comprehensive and transparent documentation covering appropriate use, implementation specifics, performance metrics, and fairness and ethical considerations of data valuation methods will significantly improve usage.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback, which significantly improved our paper. This work was supported in part by the National Science Foundation under grants CCF-2212968 and ECCS-2216899, and by the Simons Foundation under the Simons Collaboration on the Theory of Algorithmic Fairness.

References

Agarwal, A.; Dahleh, M.; and Sarkar, T. 2019. A Marketplace for Data: An Algorithmic Solution. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, EC '19, 701–726. New York, NY, USA: Association for Computing Machinery.

Alsallakh, B.; Cheema, A.; Procope, C.; Adkins, D.; McReynolds, E.; Wang, E.; Pehl, G.; Green, N.; and Zvyagina, P. 2022. System-Level Transparency of Machine Learning. Technical report, Meta AI.

Arnaiz-Rodriguez, A.; and Oliver, N. 2023. FairShap: A Data Re-weighting Approach for Algorithmic Fairness based on Shapley Values. arXiv:2303.01928.

Arnold, M.; Bellamy, R. K.; Hind, M.; Houde, S.; Mehta, S.; Mojsilović, A.; Nair, R.; Ramamurthy, K. N.; Olteanu, A.; Piorkowski, D.; et al. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 63(4/5): 6–1.

Azcoitia, S. A.; and Laoutaris, N. 2022. Try before you buy: a practical data purchasing algorithm for real-world data marketplaces. In *Proceedings of the 1st International Workshop on Data Economy*, DE '22, 27–33. New York, NY, USA: Association for Computing Machinery.

Bender, E. M.; and Friedman, B. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6: 587–604.

Bischl, B.; Casalicchio, G.; Feurer, M.; Hutter, F.; Lang, M.; Mantovani, R. G.; Rijn, J. N. V.; and Vanschoren, J. 2019. OpenML Benchmarking Suites. *arXiv:1708.03731v2 [stat.ML]*.

Black, E.; and Fredrikson, M. 2021. Leave-one-out Unfairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 285–295. New York, NY, USA: Association for Computing Machinery.

Bommasani, R.; Klyman, K.; Longpre, S.; Kapoor, S.; Maslej, N.; Xiong, B.; Zhang, D.; and Liang, P. 2023. The foundation model transparency index. *arXiv preprint arXiv:2310.12941*.

Castro Fernandez, R. 2022. Protecting Data Markets from Strategic Buyers. In *Proceedings of the 2022 International Conference on Management of Data*, SIGMOD '22, 1755–1769. New York, NY, USA: Association for Computing Machinery.

Chen, H.; Covert, I. C.; Lundberg, S. M.; and Lee, S.-I. 2023. Algorithms to estimate Shapley value feature attributions. *Nature Machine Intelligence*, 5(6): 590–601.

Chen, H.; Janizek, J. D.; Lundberg, S.; and Lee, S.-I. 2020. True to the model or true to the data? *arXiv preprint arXiv:2006.16234*.

Chen, J.; Song, L.; Wainwright, M. J.; and Jordan, M. I. 2018. L-Shapley and C-Shapley: Efficient model interpretation for structured data. *arXiv preprint arXiv:1808.02610*.

Chmielinski, K. S.; Newman, S.; Taylor, M.; Joseph, J.; Thomas, K.; Yurkofsky, J.; and Qiu, Y. C. 2022. The dataset nutrition label (2nd gen): Leveraging context to mitigate harms in artificial intelligence. *arXiv preprint arXiv:2201.03954*.

Cook, R.; and Weisberg, S. 1980. Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics*, 22(4): 495–508.

Covert, I.; and Lee, S.-I. 2021. Improving kernelshap: Practical shapley value estimation using linear regression. In *International Conference on Artificial Intelligence and Statistics*, 3457–3465. PMLR.

Coyle, D.; and Manley, A.-M. 2023. What is the value of data? A review of empirical methods. *Journal of Economic Surveys*.

Creative Commons License. 2013. Creative Commons CC BY License Description. https://creativecommons.org/licenses/by/4.0/. Accessed: 2024-04-29.

Díaz, M.; Kivlichan, I.; Rosen, R.; Baker, D.; Amironesei, R.; Prabhakaran, V.; and Denton, E. 2022. Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2342–2351.

Duin, R. 1998. Multiple Features. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5HC70.

Fleckenstein, M.; Obaidi, A.; and Tryfona, N. 2023. Data Valuation: Use Cases, Desiderata, and Approaches. In *Proceedings of the Second ACM Data Economy Workshop*, DEC '23, 48–52. New York, NY, USA: Association for Computing Machinery.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Ghorbani, A.; Kim, M.; and Zou, J. 2020. A distributional framework for data valuation. In *International Conference on Machine Learning*, 3535–3544. PMLR.

Ghorbani, A.; and Zou, J. 2019. Data Shapley: Equitable Valuation of Data for Machine Learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 2242–2251. PMLR.

Ghorbani, A.; Zou, J.; and Esteva, A. 2022. Data shapley valuation for efficient batch active learning. In *2022 56th Asilomar Conference on Signals, Systems, and Computers*, 1456–1462. IEEE.

Gilbert, T. K.; Lambert, N.; Dean, S.; Zick, T.; Snoswell, A.; and Mehta, S. 2023. Reward reports for reinforcement learning. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 84–130.

Hammoudeh, Z.; and Lowd, D. 2024. Training data influence analysis and estimation: a survey. *Machine Learning*.

Han, M.; Light, J.; Xia, S.; Galhotra, S.; Fernandez, R. C.; and Xu, H. 2023. A Data-Centric Online Market for Machine Learning: From Discovery to Pricing. *ArXiv*, abs/2310.17843.

Hong, S.; and Lynn, H. S. 2020. Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. *BMC Medical Research Methodology*, 20(1): 199.

Huang, G. 2021. Missing data filling method based on linear interpolation and lightgbm. In *Journal of Physics Conference Series*, volume 1754 of *Journal of Physics Conference Series*, 012187.

Jethani, N.; Sudarshan, M.; Covert, I. C.; Lee, S.-I.; and Ranganath, R. 2021. Fastshap: Real-time shapley value estimation. In *International Conference on Learning Representations*.

Jia, R.; Dao, D.; Wang, B.; Hubis, F. A.; Gurel, N. M.; Li, B.; Zhang, C.; Spanos, C.; and Song, D. 2019a. Efficient task-specific data valuation for nearest neighbor algorithms. *Proceedings of the VLDB Endowment*, 12(11): 1610–1623.

Jia, R.; Dao, D.; Wang, B.; Hubis, F. A.; Hynes, N.; Gürel, N. M.; Li, B.; Zhang, C.; Song, D.; and Spanos, C. J. 2019b. Towards Efficient Data Valuation Based on the Shapley Value. *CoRR*, abs/1902.10275.

Jia, R.; Wu, F.; Sun, X.; Xu, J.; Dao, D.; Kailkhura, B.; Zhang, C.; Li, B.; and Song, D. 2021. Scalability vs. Utility: Do We Have To Sacrifice One for the Other in Data Importance Quantification? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8239–8247.

Jiang, M.; Roth, H. R.; Li, W.; Yang, D.; Zhao, C.; Nath, V.; Xu, D.; Dou, Q.; and Xu, Z. 2023. Fair federated medical image segmentation via client contribution estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16302–16311.

Kang, J. S.; Pedarsani, R.; and Ramchandran, K. 2024. The Fair Value of Data Under Heterogeneous Privacy Constraints in Federated Learning. *Transactions on Machine Learning Research*.

Kendall, M. G. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2): 81–93.

Koh, P. W.; and Liang, P. 2017. Understanding Black-box Predictions via Influence Functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 1885–1894. PMLR.

Krasin, I.; Duerig, T.; Alldrin, N.; Ferrari, V.; Abu-El-Haija, S.; Kuznetsova, A.; Rom, H.; Uijlings, J.; Popov, S.; Veit, A.; et al. 2017. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github. com/openimages*, 2(3): 18.

Kwon, Y.; Rivas, M. A.; and Zou, J. 2021. Efficient computation and analysis of distributional shapley values. In *International Conference on Artificial Intelligence and Statistics*, 793–801. PMLR.

Kwon, Y.; and Zou, J. Y. 2021. Beta Shapley: a Unified and Noise-reduced Data Valuation Framework for Machine Learning. In *International Conference on Artificial Intelligence and Statistics*.

Laoutaris, N. 2019. Why online services should pay you for your data? The arguments for a human-centric data economy. *IEEE Internet Computing*, 23(5): 29–35.

Liu, Z.; Chen, Y.; Yu, H.; Liu, Y.; and Cui, L. 2022. GTG-shapley: Efficient and accurate participant contribution evaluation in federated learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(4): 1–21.

MIT License. 1980. MIT License. https://en.wikipedia.org/wiki/MIT_License. Last accessed 2024-05-10.

Mitchell, M.; Luccioni, A. S.; Lambert, N.; Gerchick, M.; McMillan-Major, A.; Ozoani, E.; Rajani, N.; Thrush, T.; Jernite, Y.; and Kiela, D. 2023. Measuring Data. arXiv:2212.05129.

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, 220–229.

Murti, D. M. P.; Pujianto, U.; Wibawa, A. P.; and Akbar, M. I. 2019. K-Nearest Neighbor (K-NN) based Missing Data Imputation. In *2019 5th International Conference on Science in Information Technology (ICSITech)*, 83–88.

Muzellec, B.; Josse, J.; Boyer, C.; and Cuturi, M. 2020. Missing Data Imputation using Optimal Transport. In *International Conference on Machine Learning*, 7130–7140. PMLR.

Naggita, K.; and LaChance, J. 2025. A case for data valuation transparency via DValCards. *CoRR*, abs/2506.23349.

Ohrimenko, O.; Tople, S.; and Tschiatschek, S. 2019. Collaborative Machine Learning Markets with Data-Replication-Robust Payments. *CoRR*, abs/1911.09052.

Pang, J.; Wang, J.; Zhu, Z.; Yao, Y.; Qian, C.; and Liu, Y. 2024. Fair Classifiers Without Fair Training: An Influence-Guided Data Sampling Approach. *arXiv preprint arXiv:2402.12789*.

Paraschiv, M.; and Laoutaris, N. 2019. Valuating user data in a human-centric data economy. *arXiv preprint arXiv:1909.01137*.

Rostamzadeh, N.; Mincu, D.; Roy, S.; Smart, A.; Wilcox, L.; Pushkarna, M.; Schrouff, J.; Amironesei, R.; Moorosi, N.; and Heller, K. 2022. Healthsheet: development of a transparency artifact for health datasets. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1943–1961.

Schoch, S.; Xu, H.; and Ji, Y. 2022. CS-Shapley: class-wise Shapley values for data valuation in classification. *Advances in Neural Information Processing Systems*, 35: 34574–34585.

Sim, R. H. L.; Xu, X.; and Low, B. K. H. 2022. Data Valuation in Machine Learning: "Ingredients", Strategies, and Open Challenges. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 5607–5614. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Singh, P. 2023. Systematic review of data-centric approaches in artificial intelligence and machine learning. *Data Science and Management*.

Song, T.; Tong, Y.; and Wei, S. 2019. Profit allocation for federated learning. In *2019 IEEE International Conference on Big Data (Big Data)*, 2577–2586. IEEE.

Tang, S.; Ghorbani, A.; Yamashita, R.; Rehman, S.; Dunnmon, J. A.; Zou, J.; and Rubin, D. L. 2021. Data valuation for medical imaging using Shapley value and application to a large-scale chest X-ray dataset. *Scientific Reports*, 11(1): 8366.

Tian, Z.; Liu, J.; Li, J.; Cao, X.; Jia, R.; and Ren, K. 2022. Private Data Valuation and Fair Payment in Data Marketplaces. *CoRR*, abs/2210.08723.

Van Buuren, S.; and Groothuis-Oudshoorn, K. 2011. MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3): 1–67.

Wang, H.; Wu, Z.; and He, J. 2024. FairIF: Boosting Fairness in Deep Learning via Influence Functions with Validation Set Sensitive Attributes. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 721–730.

Wang, J. T.; and Jia, R. 2023. Data Banzhaf: A Robust Data Valuation Framework for Machine Learning. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, 6388–6421. PMLR.

Wang, J. T.; Yang, T.; Zou, J.; Kwon, Y.; and Jia, R. 2024. Rethinking data shapley for data selection tasks: misleads and merits. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Wang, J. T.; Zhu, Y.; Wang, Y.-X.; Jia, R.; and Mittal, P. 2023. Threshold KNN-shapley: a linear-time and privacy-friendly approach to data valuation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Red Hook, NY, USA: Curran Associates Inc.

Wang, T.; Rausch, J.; Zhang, C.; Jia, R.; and Song, D. 2020. A principled approach to data valuation for federated learning. *Federated Learning: Privacy and Incentive*, 153–167.

Xu, X.; Wu, Z.; Foo, C. S.; and Low, B. K. H. 2021. Validation Free and Replication Robust Volume-based Data Valuation. In *Advances in Neural Information Processing Systems*, volume 34, 10837–10848. Curran Associates, Inc.

Yona, G.; Ghorbani, A.; and Zou, J. 2021. Who's Responsible? Jointly Quantifying the Contribution of the Learning Algorithm and Data. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, 1034–1041. New York, NY, USA: Association for Computing Machinery.

Yoon, J.; Arik, S.; and Pfister, T. 2020. Data Valuation using Reinforcement Learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 10842–10851. PMLR.

Zhao, H.; Chen, H.; Yang, F.; Liu, N.; Deng, H.; Cai, H.; Wang, S.; Yin, D.; and Du, M. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2): 1–38.

Zhou, Z.; Xu, X.; Sim, R. H. L.; Foo, C. S.; and Low, B. K. H. 2023. Probably approximate Shapley fairness with applications in machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 5910–5918.