On the Impact of the Utility in Semivalue-based Data Valuation

Mélissa Tamine, Benjamin Heymann, Patrick Loiseau, Maxime Vono

On the Impact of the Utility in Semivalue-based Data Valuation

Mélissa Tamine [1 2] **Benjamin Heymann** [1] **Patrick Loiseau** [2] **Maxime Vono** [1]

Abstract

Semivalue-based data valuation in machine learning (ML) quantifies the contribution of individual data points to a downstream ML task by leveraging principles from cooperative game theory and the notion of utility. While this framework has been used in practice for assessing data quality, our experiments reveal inconsistent valuation outcomes across different utilities, albeit all related to ML performance. Beyond raising concerns about the reliability of data valuation, this inconsistency is challenging to interpret, as it stems from the complex interaction of the utility with data points and semivalue weights, which has barely been studied in prior work. In this paper, we take a first step toward clarifying the utility impact on semivalue-based data valuation. Specifically, we provide geometric interpretations of this impact for a broad family of classification utilities, which includes the accuracy and the arithmetic mean. We introduce the notion of *spatial signatures*: given a semivalue, data points can be embedded into a two-dimensional space, and utility functions map to the dual of this space. This geometric perspective separates the influence of the dataset and semivalue from that of the utility, providing a theoretical explanation for the experimentally observed sensitivity of valuation outcomes to the utility choice.

1. Introduction

Supervised machine learning (ML) relies on data, but real-world datasets often suffer from noise and biases as they are collected from multiple sources and are subject to measurement and annotation errors (Northcutt et al., 2021). Such variability can impact learning outcomes, highlighting the need for systematic methods to evaluate data quality. In response, *data valuation* has emerged as a growing research field that aims to quantify individual data points' contribution to a learning task, helping to identify informative samples and mitigate the impact of low-quality data. A popular way to tackle the data valuation problem is to adopt a cooperative game-theoretic viewpoint, where each data point is modeled as a player in a coalitional game, and the usefulness of any data subset is measured by a *utility function*. This approach leverages game theory solution concepts called *semivalues* (Dubey et al., 1981), which input data and utility to assign an importance score to each data point (Ghorbani & Zou, 2019; Kwon & Zou, 2022; Wang & Jia, 2023; Jia et al., 2023; 2020). When computing semivalues, the utility function is typically selected as a performance metric, such as the accuracy in classification or the mean squared error in regression. However, this choice is inherently unconstrained—any function mapping data subsets to real values can serve as a utility as long as a higher utility reflects better performance. This flexibility raises a fundamental and legitimate question: *to what extent does the choice of utility impact data valuation outcomes?*

Despite the widespread use of semivalue-based data valuation, there is a limited theoretical understanding of how and why the choice of utility function influences valuation outcomes. In practice, utility functions are often chosen for convenience, typically as standard ML performance metrics (Ghorbani & Zou, 2019), rather than being grounded in theoretical principles. However, Wang et al. (2024) demonstrated that for a particular semivalue-based method, certain utility choices can lead to valuation outcomes no better than random importance assignment when no specific constraints are imposed. This finding underscores a fundamental issue: the flexibility in utility selection introduces variability in data valuation, potentially leading to inconsistent or misleading conclusions. This question's lack of theoretical grounding is particularly concerning in high-stakes decision-making scenarios such as healthcare, where data valuation informs critical tasks (Pandl et al., 2021; Bloch & Friedrich, 2021; Zheng et al., 2024). Practitioners risk making unreliable decisions that undermine model performance and interpretability without a clear understanding of how utility functions shape valuation outcomes. Our study aims to fill in this gap, providing insights to better understand data valuation and its practical applications.

Our contributions can be summarized as follows:

[1]Criteo AI Lab, Paris, France [2]Inria, Fairplay joint team, Palaiseau, France. Correspondence to: Mélissa Tamine <m.tamine@criteo.com>.

1. **Empirical evidence of data valuation outcomes variability across utility functions.** Our experiments reveal that the agreement between two utility functions in assessing data importance varies unpredictably across datasets and semivalues. For a given dataset and semivalue, two utilities may produce similar rankings of data points, while for another pair, they may diverge entirely. This lack of a systematic pattern suggests that a utility's impact on data valuation is not solely determined by its intrinsic properties but rather by its interaction with the dataset and the semivalue.
2. **A geometric interpretation of a utility interaction with data and semivalue.** We propose a geometric framework for a class of binary classification utilities to better understand this interaction. We introduce the concept of *spatial signatures*, which correspond to an embedding of the dataset into a two-dimensional space induced by the semivalue. We show that the utility functions we consider map to the dual of this space, enabling data values to be visualized as projections onto directions defined by the utility. This geometric perspective provides a structured way to understand which datasets and semivalues lead to robust data valuations across utilities and which lead to variable and inconsistent valuations. In particular, it explains why, in our experiments, utility functions influence data valuation inconsistently across different datasets and semivalues.

Related work. Game-theoretic approaches to data valuation have gained traction in recent years due to their formal justification through axioms. The Shapley value (Shapley, 1953; Ghorbani & Zou, 2019), in particular, has been widely adopted as a data valuation method because it uniquely satisfies four key axioms: linearity, dummy player, symmetry, and efficiency. Alternative approaches have emerged by relaxing some of these axioms. By omitting the efficiency requirement, one obtains the semivalue framework (Dubey et al., 1981). Examples of value notions within this class include LOO (Leave-One-Out) (Koh & Liang, 2020), Beta Shapley (Kwon & Zou, 2022), and Data Banzhaf (Wang & Jia, 2023). Furthermore, relaxing the linearity axiom leads to the Least Core, an alternative concept from the cooperative game theory proposed by (Yan & Procaccia, 2021) for data valuation. The Least Core determines an optimal profit allocation where each coalition S receives the minimum required subsidy to prevent any participant from defecting from the grand coalition $\mathcal{D}$. The Distributional Shapley Value (Ghorbani et al., 2020; Kwon et al., 2021) is an extension of Data Shapley designed to assess data contributions based on an underlying data distribution rather than a fixed dataset. Beyond cooperative game theory, several non-game-theoretic data valuation methods have been explored. An overview is provided by (Sim et al., 2022), and some of them are benchmarked by (Jiang et al., 2023).

Notation. We set $\mathbb{N}^* = \mathbb{N} \setminus \{0\}$. For $n \in \mathbb{N}^*$, we denote $[n] := \{1, .., n\}$. For a dataset $\mathcal{D}$, we denote as $|\mathcal{D}|$ its cardinality and as $2^{\mathcal{D}}$ its powerset, i.e., the set of all possible subsets of $\mathcal{D}$, including the empty set $\emptyset$ and $\mathcal{D}$ itself. For $d \in \mathbb{N}^*$, we denote $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}$ an input space and an output space, respectively.

2. Background

2.1. Semivalue-based data valuation set-up

The data valuation problem involves a dataset of interest $\mathcal{D} = \{z_i = (x_i, y_i)\}_{i \in [n]}$, where for any $i \in [n]$ each $x_i \in \mathcal{X}$ is a feature vector and $y_i \in \mathcal{Y}$ is the corresponding label. We focus on binary classification, where $\mathcal{Y} = \{0, 1\}$. Data valuation aims to assign a scalar score to each data point in $\mathcal{D}$, quantifying its contribution to a downstream ML task. These scores will be referred to as *data values*.

Utility functions. Most data valuation methods rely on *utility functions* to compute data values. A utility is a set function $u : 2^{\mathcal{D}} \to \mathbb{R}$ that maps any subset S of the training set $\mathcal{D}$ to a score indicating its usefulness for performing the considered ML task. Formally, this can be expressed as $u(S) = \text{PERF}(\mathcal{A}(S))$, where $\mathcal{A}$ is a learning algorithm that takes a subset S as input and returns a trained model, and PERF is a metric function used to evaluate the model's performance. For classification tasks, PERF can be chosen, for instance, as the accuracy evaluated on a hold-out test set $\mathcal{D}_{\text{test}}$. There are, however, many other choices of performance function, which lead to different utility functions—this is precisely the focus of our study. For convenience, we interchangeably refer to the utility u and the performance metric PERF as u inherently depends on PERF.

Semivalues. The most popular data valuation methods assign a value score to each data point in $\mathcal{D}$ using solution concepts from cooperative game theory, known as semivalues (Dubey et al., 1981). The collection of data valuation methods that fall under this category is referred to as *semivalue-based data valuation*. These methods rely on the notion of *marginal contribution*. Formally, for any $i, j \in [n]$, let $\mathcal{D}_j^{\setminus z_i}$ denote the set of all subsets of $\mathcal{D}$ of size $j-1$ that exclude z_i. Then, the marginal contribution of z_i with respect to other $j-1$ samples is defined as

$$\Delta_j(z_i; u) := \frac{1}{\binom{n-1}{j-1}} \sum_{S \subseteq \mathcal{D}_j^{\setminus z_i}} u\left(S \cup \{z_i\}\right) - u(S) \,.$$

The marginal contribution $\Delta_j(z_i; u)$ considers all possible subsets $S \in \mathcal{D}_j^{\setminus z_i}$ with the same cardinality $j-1$ and measures the average changes of u when datum of interest z_i is removed from $S \cup \{z_i\}$.

Each semivalue-based method is characterized by a weight vector $\omega := (\omega_1, \ldots, \omega_n)$ and assigns a score $\phi(z_i; \omega, u)$ to

each data point $z_i \in \mathcal{D}$ by computing a weighted average of its marginal contributions $\{\Delta_j(z_i;u)\}_{j\in[n]}$. Specifically,

$$\phi(z_i;\omega,u) := \sum_{j=1}^{n} \omega_j \Delta_j(z_i;u)\ . \tag{1}$$

Below, we define the weights for three commonly used semivalue-based methods (Jiang et al., 2023) and illustrate them in Figure 1. Their differences in weighting schemes have geometric implications discussed in Section 4.

Definition 2.1. *Data Shapley* (Ghorbani & Zou, 2019) is derived from the *Shapley value* (Shapley, 1953), a solution concept from cooperative game theory that fairly allocates the total gains generated by a coalition of players based on their contributions. In the context of data valuation, Data Shapley takes a simple average of all the marginal contributions. Its weight vector ω_{shap} is

$$\omega_{\text{shap}} = \left(\frac{1}{n},\dots,\frac{1}{n}\right)$$

Definition 2.2. (α,β)*-Beta Shapley* (Kwon & Zou, 2022) extends Data Shapley by introducing tunable parameters $(\alpha,\beta) \in \mathbb{R}^2$, which controls the emphasis placed on marginal contributions from smaller or larger subsets. The corresponding weight vector $\omega_{\text{beta}} = (\omega_{\text{beta},j})_{j\in[n]}$ has

$$\omega_{\text{beta},j} = \binom{n-1}{j-1} \cdot \frac{\text{Beta}(j+\beta-1, n-j+\alpha)}{\text{Beta}(\alpha,\beta)}$$

for all j, where $\text{Beta}(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$ and Γ is the Gamma function.

Definition 2.3. *Data Banzhaf* (Wang & Jia, 2023) is derived from the *Banzhaf value* (Banzhaf, 1965), a cooperative game theory concept originally introduced to measure a player's influence in weighted voting systems. Data Banzhaf weight's vector $\omega_{\text{banzhaf}} = (\omega_{\text{banzhaf},j})_{j\in[n]}$ is defined as

$$\omega_{\text{banzhaf},j} = \binom{n-1}{j-1} \cdot \frac{1}{2^{n-1}}$$

These semivalue-based methods satisfy fundamental axioms that ensure desirable properties in data valuation. We formally state these axioms in the following. Let $\phi(.,\omega;.)$ be a semivalue-based data valuation method defined by a weight vector ω and let u and v be utility functions. Then, ϕ satisfies the following axioms:

1. Dummy. If $u(S\cup\{z_i\}) = u(S)+c$ for all $S \subseteq \mathcal{D}\backslash\{z_i\}$ and some $c \in \mathbb{R}$, then $\phi(z_i;\omega,u) = c$.
2. Symmetry. If $u(S\cup\{z_j\}) = u(S\cup\{z_j\})$ for all $S \subseteq \mathcal{D}\backslash\{z_i,z_j\}$, then $\phi(z_i;\omega,u) = \phi(z_j;\omega,u)$.
3. Linearity. For any $\alpha_1,\alpha_2 \in \mathbb{R}$, $\phi(z_i;\omega,\alpha_1 u+\alpha_2 v) = \alpha_1\phi(z_i;\omega,u) + \alpha_2\phi(z_i;\omega,v)$.

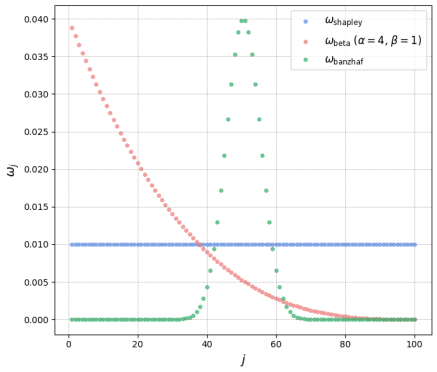


Figure 1. Comparison of Data Shapley, $(4,1)$-Beta Shapley, and Data Banzhaf weighting schemes for $n = 100$.

While all semivalues satisfy the above axioms, Data Shapley uniquely also guarantees *efficiency*: $\sum_{z\in\mathcal{D}} \phi(z,\omega,u) = u(\mathcal{D})$.

2.2. Applications of semivalue-based methods

In practice, semivalue-based methods are mostly applied to perform *data cleaning* or *data subset selection* (Tang et al., 2021; Pandl et al., 2021; Bloch & Friedrich, 2021; Zheng et al., 2024). Both tasks involve ranking data points according to their assigned values.

Data cleaning. Data cleaning aims to improve dataset quality by identifying and removing noisy or low-quality data points. Since semivalue-based methods quantify each point's contribution to a downstream task, low-valued points are natural candidates for removal. Specifically, a common approach is to remove points that fall into the set $\mathcal{N}_\tau$, defined as the subset of data points with the lowest values (Ghorbani & Zou, 2019). Formally, $\mathcal{N}_\tau = \{z_i \in \mathcal{D} \mid \phi(z_i;u,\omega) \le \tau\}$, where τ is a threshold determined through domain knowledge or empirical evaluation.

Data subset selection. Data subset selection involves choosing the optimal training set from available samples to maximize final model performance. Since semivalues measure data quality, prioritizing data points with the highest values is a natural approach. Consequently, a common practice in the literature is selecting, given a size budget k, the subset $\mathcal{S}^{(k)}_{\phi(u,\omega)}$ of data points with top-k data values, i.e., $\mathcal{S}^{(k)}_{\phi(u,\omega)} = \arg\max_{\mathcal{S}\subseteq\mathcal{D},|\mathcal{S}|=k} \sum_{z_i\in\mathcal{S}} \phi(z_i;u,\omega)$ (Wang et al., 2024).

3. Variability of data valuations across utility functions: an experimental investigation

Although utility functions are central in semivalue-based data valuation methods, their impact on data valuation outcomes has received little attention. Wang et al. (2024) is

the only work that theoretically explores this question, focusing specifically on how the choice of utility affects the reliability of Data Shapley for data subset selection. In this section, we broaden this scope by experimentally studying the influence of various utility functions across multiple semivalue-based methods. To conduct this investigation, we propose an application-agnostic metric based on data values *ranking*, enabling a broader perspective beyond one specific application.

3.1. An application-agnostic metric based on rank correlation to compare utility impact

Most data valuation applications depend on the relative ranking of data points based on their assigned values. Data cleaning prioritizes identifying low-ranked points, while data subset selection focuses on points with the highest values. Since the rankings inherently determine the outcome of these applications, if rankings induced by different utility functions are highly similar, it suggests that the utilities are aligned in their ability to prioritize data points for a given application. Therefore, *rank correlation* appears as an intuitive and reasonable metric for evaluating whether utility functions produce consistent data valuation outcomes.

Formally, given a dataset $\mathcal{D}$ and a semivalue characterized by weight vector ω, we compare the impact of two utilities u and v on data valuation outcomes by computing the rank correlation between $\{\phi(z_i; u, \omega)\}_{i\in[n]}$ and $\{\phi(z_i; v, \omega)\}_{i\in[n]}$. Several measures of rank correlation exist to evaluate the similarity between two rankings. One of the most widely used is the *Kendall rank correlation coefficient*,[1] which quantifies the agreement between two rankings by comparing the relative order of all pairs of elements.

Kendall rank correlation coefficient (Kendall, 1938) measures the ordinal association between two sets of values $\{\phi_i^u\}_{i\in[n]}$ and $\{\phi_i^v\}_{i\in[n]}$ assigned to n elements. For any pair of indices (i, j) where $i < j$, the pair is *concordant* if the relative order of ϕ_i^u and ϕ_j^u matches the relative order of ϕ_i^v and ϕ_j^v and is *discordant* otherwise. Let C and D denote the number of concordant and discordant pairs, respectively. The Kendall rank coefficient $\tau(u, v)$ is defined as $\tau(u, v) = \frac{C-D}{\binom{n}{2}}$, and can be equivalently expressed as

$$\tau(u, v) = \frac{1}{\binom{n}{2}} \sum_{i<j} \operatorname{sgn}[(\phi_i^u - \phi_j^u) \cdot (\phi_i^v - \phi_j^v)] \,, \qquad (2)$$

and $\tau(u, v) = 1$ indicates perfect alignment between $\{\phi_i^u\}_{i\in[n]}$ and $\{\phi_i^v\}_{i\in[n]}$ while $\tau(u, v) = -1$ traduces perfect disagreement and $\tau(u, v) = 0$ an absence of correlation.

While similar data value rankings suggest alignment in data valuation applications, the converse is not always true. Low-rank correlation does not necessarily imply misalignment. We derive an analytical example in Appendix B.3, which shows that two utility functions can exhibit low-rank correlation yet consistently separate high-importance from low-importance points, demonstrating aligned behavior despite ranking differences. This is why, in cases of low-rank correlation, we complement the rank correlation measure with an intersection analysis, evaluating the overlap between bottom-ranked subsets of data values to better assess utility alignment in performing data valuation applications.

[1] Along with the *Spearman rank correlation* (Spearman, 1904).

3.2. Experimental evidence of ranking variability

We perform systematic rank correlation computations on various datasets and semivalue-based methods to assess the ranking variability of data values induced by different utility functions.

Experimental setup. Rank correlation computations are performed on several publicly available binary classification datasets widely used in the literature (Ghorbani & Zou, 2019; Wang & Jia, 2023; Kwon & Zou, 2022; Jiang et al., 2023). Table 3 lists these datasets along with their sources. We compute data values using three semivalue-based methods: Data Shapley, $(4, 1)$-Beta Shapley,[2] and Data Banzhaf. Given a dataset $\mathcal{D}$ and a semivalue ω, we evaluate the impact of three commonly used classification utilities: the *accuracy* (ACC), the *recall* (REC), and the *arithmetic mean* (AM). Specifically, for each utility function pair (u, v), we compute the Kendall rank correlation $\tau(u, v)$ to quantify ranking consistency. We extend these experiments to additional classification utilities and reproduce them with the Spearman rank correlation for completeness. The corresponding results are provided in Appendix C.5.

Remark 3.1. Computing $\tau(u, v)$ for each utility pair (u, v) requires obtaining the data values sets $\{\phi(z, \omega, u)\}_{z\in\mathcal{D}}$ and $\{\phi(z, \omega, v)\}_{z\in\mathcal{D}}$. This computation involves a learning algorithm $\mathcal{A}$, and a test dataset $\mathcal{D}_{\text{test}}$ to evaluate the utility on different subsets $S \subseteq \mathcal{D}$, and an approximation method as the exact computation of semivalues is infeasible for large datasets (Ghorbani & Zou, 2019; Jia et al., 2023; Garrido-Lucero et al., 2024). To ensure that any observed differences in both sets' rankings arise solely from the choice of the utility and not from these other sources of variability, we propose a systematic methodology in Appendix C.2 that eliminates extraneous perturbations. The results reveal qualitatively similar insights, reinforcing the observation that the impact of utility functions on data valuation rankings is not solely dictated by their intrinsic properties but also by their interaction with the dataset and the semivalue. Full details are available in Appendices C.5.1 and C.5.2.

[2] The authors in (Kwon & Zou, 2022) identify $(\alpha, \beta) = (16, 1)$ as best suited for data valuation applications, but we use $(\alpha, \beta) = (4, 1)$ as it performs best in the benchmark proposed by (Jiang et al., 2023).

Results analysis. The experimental results in Table 1 reveal that utility pairs (ACC-AM, ACC-REC, and REC-AM) exhibit no systematic agreement or disagreement across datasets and semivalues. In some cases, utility functions produce highly similar rankings, while their rank correlation is markedly low in others.

For example, in the Breast dataset, ACC-AM exhibits strong agreement across all semivalues, with Kendall correlations of 0.98, 0.98, and 0.99 for Shapley, $(4, 1)$-Beta Shapley, and Banzhaf, respectively. However, this same utility pair shows significantly weaker agreement for the CREDIT dataset, with Kendall correlations dropping to 0.51, 0.58, and 0.07, highlighting a sharp dataset-dependent divergence.

Moreover, substantial differences emerge within a dataset depending on the chosen semivalue. For the CPU dataset, the correlation between ACC-REC is relatively high under Shapley and $(4, 1)$-Beta Shapley (Kendall rank correlations of 0.78 and 0.79, respectively), yet it vanishes entirely under Banzhaf (Kendall coefficient $= 0.01$).

In addition, following Remark 3.1, we complement the rank correlation computations with an intersection-based analysis in Appendix C.5.3 for datasets, semivalues, and utility pairs exhibiting low Kendall rank correlation coefficient. This analysis assesses whether differences in value rankings indicate misalignment in performing data valuation applications. The results show no cases where low-rank correlation preserves alignment, suggesting that low-rank correlation effectively reflects misalignment for these datasets, semivalues, and utility pairs.

All those results suggest that a utility function's influence on data valuation rankings is not an intrinsic property of the utility itself but rather emerges from its interaction with both the dataset and the semivalue. The same utility pair can yield highly similar rankings in one context yet diverge entirely in another, indicating that the way a utility assigns value to data points is shaped by how it interacts with the dataset's structure and how the semivalue aggregates marginal contributions. This reinforces the idea that utility-driven data valuation cannot be understood in isolation—its effects are context-dependent, varying with both the dataset characteristics and the weighting mechanism imposed by the semivalue.

This utility-dataset-semivalue dependency remains underexplored in prior work, resulting in a limited understanding of what semivalue-based methods truly capture. This raises concerns, as data valuation methods are intended to enhance the interpretability of dataset quality. Addressing this gap requires a deeper exploration of the interplay between utilities, datasets, and semivalue weights to ensure semivalue-based data valuation delivers its promise.

4. Explaining ranking variability through a geometric interpretation of the utility-dataset-semivalue interplay

Motivated by the experimental results from Section 3, we aim to understand the interplay between utilities, datasets, and semivalues in order to explain ranking variability. We focus on a family of classification utilities that includes accuracy, recall, and arithmetic mean, for which we derive *geometric interpretations* of ranking diversity. This geometric perspective provides a framework to explain the results in Table 1.

4.1. A subclass of *linear fractional performance measures*

This section introduces the specific family of utility functions we consider in our analysis. We build on the framework of *linear fractional performance measures* (Koyejo et al., 2014), which generalizes various classification metrics, including the F-score and misclassification risk.

These measures express classifier performance as a ratio of affine functions of classification probabilities. Formally, given a training dataset $S \in (\mathcal{X} \times \{0,1\})^n$, a test dataset $\mathcal{D}_{\text{test}} = \{(x_j, y_j)\}_{j\in[m]} \in (\mathcal{X} \times \{0,1\})^m$ and a learning algorithm $\mathcal{A}$, we denote $g_S = \mathcal{A}(S)$ a classifier trained on S which maps input features to predicted labels, i.e., $g_S : \mathcal{X} \to \{0,1\}$. A linear fractional performance measure u evaluates the performance of g_S on $\mathcal{D}_{\text{test}}$ as

$$u(S) = \frac{c_0 + c_1\lambda(S, \mathcal{D}_{\text{test}}) + c_2\gamma(S, \mathcal{D}_{\text{test}})}{d_0 + d_1\lambda(S, \mathcal{D}_{\text{test}}) + d_2\gamma(S, \mathcal{D}_{\text{test}})} ,$$

where $(c_0, c_1, c_2, d_0, d_1, d_2) \in \mathbb{R}^6$ determines the structure of u while $\lambda(S, \mathcal{D}_{\text{test}}) = \widehat{\mathbb{P}}_{\mathcal{D}_{\text{test}}}(g_S(x) = 1, y = 1) = \frac{1}{m}\sum_{j=1}^{m} \mathbf{1}[g_S(x_j) = 1, y_j = 1]$ is the empirical probability of true positives and $\gamma(S, \mathcal{D}_{\text{test}}) = \widehat{\mathbb{P}}_{\mathcal{D}_{\text{test}}}(g_S(x) = 1) = \frac{1}{m}\sum_{j=1}^{m} \mathbf{1}[g_S(x_j) = 1]$ is the empirical probability of positive predictions. When the test dataset $\mathcal{D}_{\text{test}}$ is clear from context, we use the shorthand $\lambda(S)$ to denote $\lambda(S, \mathcal{D}_{\text{test}})$ and $\gamma(S)$ to denote $\gamma(S, \mathcal{D}_{\text{test}})$.

Our analysis focuses on a specific subclass of utilities within the framework of linear fractional performance measures. These utilities are characterized by a constant denominator, i.e., $d_1 = d_2 = 0$ and $d_0 \neq 0$, which simplifies their formulation to

$$u(S) = \frac{1}{d_0}\left[c_0 + c_1\lambda(S) + c_2\gamma(S)\right] . \tag{3}$$

We refer to this subclass as the $(\lambda, \gamma)-$*linear utility class*, denoted by $\mathcal{U}_{\lambda,\gamma}$, since each utility function in this class is a linear transformation of the classification statistics λ and γ. Accuracy, recall, and arithmetic mean, used in our experiments (Section 3), belong to this subclass. Their formulation follows Eq. (3), with coefficients (c_0, c_1, c_2, d_0)

DATASET	SHAPLEY			$(4,1)$-BETA SHAPLEY			BANZHAF		
	ACC-REC	ACC-AM	REC-AM	ACC-REC	ACC-AM	REC-AM	ACC-REC	ACC-AM	REC-AM
BREAST	0.93 (0.01)	0.98 (0.01)	0.92 (0.01)	0.94 (0.01)	0.98 (0.01)	0.92 (0.01)	0.82 (0.03)	0.99 (0.01)	0.81 (0.03)
TITANIC	0.42 (0.03)	0.77 (0.01)	0.64 (0.02)	0.46 (0.02)	0.81 (0.01)	0.65 (0.01)	-0.25 (0.04)	0.77 (0.02)	-0.05 (0.05)
CREDIT	0.31 (0.01)	0.51 (0.01)	0.79 (0.01)	0.35 (0.01)	0.58 (0.01)	0.76 (0.01)	-0.31 (0.01)	0.07 (0.01)	0.60 (0.02)
HEART	0.52 (0.02)	0.98 (0.01)	0.50 (0.02)	0.61 (0.01)	0.98 (0.01)	0.59 (0.02)	0.19 (0.02)	0.98 (0.01)	0.18 (0.02)
WIND	0.77 (0.01)	0.98 (0.01)	0.75 (0.01)	0.79 (0.01)	0.98(0.01)	0.79 (0.01)	0.08 (0.03)	0.98 (0.01)	0.07 (0.03)
CPU	0.78 (0.01)	0.92 (0.01)	0.86 (0.01)	0.79 (0.01)	0.93 (0.01)	0.86 (0.01)	0.01 (0.03)	0.75 (0.02)	0.19 (0.04)
2DPLANES	0.31 (0.02)	0.99 (0.01)	0.31 (0.02)	0.33 (0.02)	0.99 (0.01)	0.33 (0.02)	0.37 (0.01)	0.99 (0.01)	0.37 (0.01)
POL	0.56 (0.01)	0.73 (0.01)	0.29 (0.01)	0.56 (0.01)	0.79 (0.01)	0.34 (0.01)	0.67 (0.01)	0.69 (0.01)	0.36 (0.01)

Table 1. Kendall rank correlations between different utility function pairs (Accuracy-Recall, Accuracy-Arithmetic Mean, and Recall-Arithmetic Mean) across multiple datasets and three semivalues: Shapley, $(4,1)$-Beta Shapley, and Banzhaf. Each semivalue is approximated 5 times using Monte Carlo sampling, and Kendall rank correlations are computed for each run. The reported values are the mean across the 5 runs, while the standard errors (in parenthesis) are derived from the standard deviation of these 5 estimates.

UTILITY	(c_0, c_1, c_2)	d_0
ACCURACY (ACC)	$(1-\pi, 2, -1)$	1
RECALL (REC)	$(0, 1, 0)$	π
ARITHMETIC MEAN (AM)	$\left(\frac{1}{2}, \frac{1}{2\pi} + \frac{1}{2(1-\pi)}, \frac{-1}{2(1-\pi)}\right)$	1

Table 2. Examples of utilities which can be represented by Eq. (3). For more examples, see Choi et al. (2009). We set $\pi = \frac{1}{m}\sum_{j=1}^{m} \mathbf{1}[y_j = 1]$, the proportion of positive labels in $\mathcal{D}_{\text{test}}$.

given in Table 2. We chose to restrict our analysis to this class of utilities, which already encompasses popular classification performance metrics, in order to be able to provide theoretical results matching empirical evidence.

4.2. Geometric characterization of dataset, semivalue, and utilty

This section examines the geometric properties of utility functions in $\mathcal{U}_{\lambda,\gamma}$ in relation to the dataset $\mathcal{D}$ and the semivalue ω. A fundamental property of the utility class $\mathcal{U}_{\lambda,\gamma}$ is that, for any $u \in \mathcal{U}_{\lambda,\gamma}$, the data value $\phi(z_i; \omega, u)$ can be represented as a linear combination of $\phi(z_i; \omega, \lambda)$ and $\phi(z_i; \omega, \gamma)$, as stated formally in the following proposition.

Proposition 4.1. *(Linear decomposition of data values for (λ,γ)-linear utilities) Let $\mathcal{D} = \{z_i\}_{i\in[n]}$ be a dataset and ω a semivalue weight vector. For any utility function $u \in \mathcal{U}_{\lambda,\gamma}$ characterized by coefficients $(c_0, c_1, c_2, d_0) \in \mathbb{R}^4$, the data value assigned to z_i can be decomposed as*

$$\phi(z_i; \omega, u) = \frac{c_1}{d_0}\phi(z_i; \omega, \lambda) + \frac{c_2}{d_0}\phi(z_i; \omega, \gamma)\,. \qquad (4)$$

The decomposition in Proposition 4.1 suggests that the influence of the dataset and semivalue is fully captured by the vector $\mathbf{e_i} = (\phi(z_i; \omega, \lambda), \phi(z_i; \omega, \gamma))$ which embeds each data point $z_i \in \mathcal{D}$ into a two-dimensional space $\mathcal{P}_\omega$ induced by the semivalue ω. While the embedding $\mathbf{e_i}$ depends on λ and γ, the spatial structure of $\mathcal{D}$ in $\mathcal{P}_\omega$ is independent of the particular choice of u within $\mathcal{U}_{\lambda,\gamma}$ as this choice is uniquely determined by the coefficients (c_0, c_1, c_2) and d_0.

Definition 4.2. (*Spatial signature*) We define the spatial signature of $\mathcal{D}$ in $\mathcal{P}_\omega$ as the collection of all embedded data points $\mathbf{e} = (\mathbf{e_i})_{i\in[n]} = \big(\phi(z_i; \omega, \lambda), \phi(z_i; \omega, \gamma)\big)_{i\in[n]}$.

Similarly, utility functions in $\mathcal{U}_{\lambda,\gamma}$ can be characterized by the vector $\mathbf{u} = (c_1/d_0, c_2/d_0)$, which belongs to space $\mathcal{U}^* \subset \mathbb{R}^2$ (cf. (3) and (4)). This gives the following result:

Theorem 4.3. *Let $\mathcal{P}_\omega \subset \mathbb{R}^2$ denote the space where each data point $z_i \in \mathcal{D}$ is embedded as $\mathbf{e_i} = \big(\phi(z_i; \omega, \lambda), \phi(z_i; \omega, \gamma)\big)$. Let $\mathcal{U}^* \subset \mathbb{R}^2$ denote the space of linear utilities, where each utility function $\mathbf{u} \in \mathcal{U}^*$ is represented as $\mathbf{u} = \left(\frac{c_1}{d_0}, \frac{c_2}{d_0}\right)$. Then, $\mathcal{U}^*$ is isomorphic to the dual space $\mathcal{P}_\omega^*$ of $\mathcal{P}_\omega$.*

Theorem 4.3 establishes a correspondence between utility functions and linear functionals over the data embedding space. In other words, each data value $\phi(z_i; \omega, u)$ results from applying a linear transformation parameterized by $\mathbf{u}$ to the embedded data representation $\mathbf{e}_i$. This directly leads to the following geometric interpretation.

Corollary 4.4. *The data value $\phi(z_i; \omega, u)$ can be expressed as the scalar product $\phi(z_i; \omega, u) = \mathbf{u} \cdot \mathbf{e}_i = \mathbf{u}(c_1, c_2, d_0) \cdot \mathbf{e}_i(\omega)$. This identifies $\phi(z_i; \omega, u)$ as the projection of $\mathbf{e_i}$ onto the utility direction $\mathbf{u}$, explicitly separating the contribution of u and $\mathcal{D}, \omega$.*

Utility directions and the unit sphere. Building on this geometric framework, we aim to understand how distinct utilities in $\mathcal{U}^*$ induce similar or divergent rankings of data values. Since rankings depend only on relative orderings and not absolute values, the space of all distinct utilities

(in terms of ranking) corresponds to the set of possible directions in $\mathcal{U}^*$, which can be naturally identified with the unit sphere. Proposition 4.5 formalizes this observation.

Proposition 4.5. *(Unit sphere representation of distinct utilities) Let $\mathbf{u} \in \mathcal{U}^*$. The set of utilities that share the same direction as $\mathbf{u}$ is in bijection with the unit sphere $\mathcal{S}^1 \subset \mathcal{U}^*$ and is uniquely represented by $\tilde{\mathbf{u}} = \mathbf{u}/\|\mathbf{u}\| = (c_1, c_2)/\sqrt{c_1^2 + c_2^2}$.*

We visually illustrate the geometric concepts discussed so far in Figure 2 for a specific dataset.

4.3. Insights on ranking diversity for extreme cases of spatial signatures.

The spatial signature of $\mathcal{D}$ in $\mathcal{P}_\omega$ plays a crucial role in shaping the rankings induced by different utility functions. In particular, we analyze two extreme cases: (a) $\mathbf{e}$ is in *general position* (Definition 4.6), (b) $\mathbf{e}$ is *collinear* (Definition 4.7).

Definition 4.6. (*General position*) A spatial signature $\mathbf{e} = (\mathbf{e}_i)_{i\in[n]} \subset \mathcal{P}_\omega$ is in general position if:

1. For all distinct i, j, k, there does not exist a line $L \subset \mathcal{P}_\omega$ such that $\mathbf{e}_i, \mathbf{e}_j, \mathbf{e}_k \in L$.
2. For all distinct i, j, there is no scalar $k > 0$ such that $\mathbf{e}_i = k\mathbf{e}_j$.

Definition 4.7. *(Collinearity)* A spatial signature $\mathbf{e} = (\mathbf{e}_i)_{i\in[n]} \subset \mathcal{P}_\omega$ is collinear if there exists $\mathbf{w} \in \mathcal{P}_\omega$ and scalar $k_i \in \mathbb{R}$ such that $\mathbf{e}_i = k_i\mathbf{w}$ for all $i \in [n]$.

We establish Theorem 4.8, which quantifies the impact of these spatial configurations on ranking diversity.

Theorem 4.8. *Let $\mathbf{e} = (\mathbf{e}_i)_{i\in[n]}$ be a spatial signature in $\mathcal{P}_\omega$. Define the ranking regions as the connected components of the unit sphere $\mathcal{S}^1$ where the linear utilities $\tilde{\mathbf{u}} \in \mathcal{S}^1$ induce identical rankings on $\mathcal{D}$. Then,*

1. *if $\mathbf{e}$ is in general position, the number of distinct ranking regions is maximal and equal to $R_{gen}(n) = 2 \times \binom{n}{2}$.*
2. *if $\mathbf{e}$ is collinear, the number of distinct ranking regions is minimal as it collapses to $R_{col}(n) = 2$.*

To illustrate this result, Figure 3 presents these two extreme cases for a dataset of three points.

For a spatial signature $\mathbf{e}$ that is neither collinear nor in general position, let $\sigma_1 > \sigma_2 > 0$ be the singular values of $\mathbf{e}$ and $\rho = \sigma_1/\sigma_2$. The number of ranking regions $R(n)$ satisfies $2 < R(n) < 2\binom{n}{2}$ and $R(n)$ decreases as ρ increases. As $\rho \to 1^+$ (balanced variance), $R(n) \to 2\binom{n}{2}$. As $\rho \to +\infty$ (collinearity), $R(n) \to 2$. The transition is governed by the alignment of the line crossing the circle at antipodal pairs $\mathcal{H}_{ij} = \{\tilde{\mathbf{u}} \in \mathcal{S}^1 \mid \tilde{\mathbf{u}} \cdot (\mathbf{e}_i - \mathbf{e}_j) = 0\} \subset S^1$ with the principal eigenvector of $\mathbf{e}$'s covariance matrix, which dominates as ρ grows.

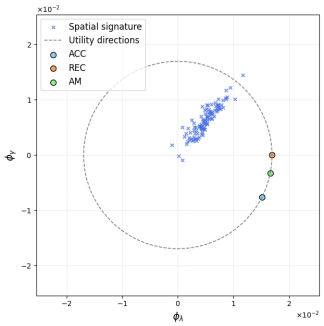
4.4. Explaining experimental results from Section 3

From the spatial signatures of the WIND dataset in Figure 2, we observe an almost collinear structure in $\mathcal{P}_{\omega_{\text{shap}}}$, $\mathcal{P}_{\omega_{\text{beta}}}$, and $\mathcal{P}_{\omega_{\text{banz}}}$. This near-alignment indicates that the embedded points $\mathbf{e}_i$ predominantly lie along the principal eigenvector $\mathbf{w}_1^\omega$ of each spatial signature's covariance matrix. For analytical simplicity, we treat these signatures as fully collinear.

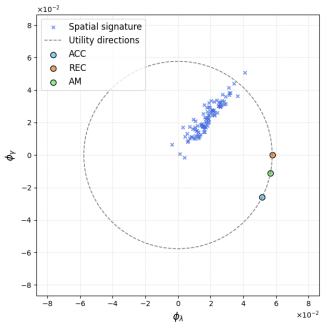
By Theorem 4.8, collinearity collapses the unit circle $\mathcal{S}^1$ into **two ranking regions**, separated by the antipodal points orthogonal to $\mathbf{w}_1^\omega$. These antipodal points, defined as $H_\omega = \{\tilde{\mathbf{u}} \in \mathcal{S}^1 \mid \tilde{\mathbf{u}} \cdot \mathbf{w}_1^\omega = 0\}$, partition $\mathcal{S}^1$ into hemispheres where $\tilde{\mathbf{u}} \cdot \mathbf{w}_1^\omega > 0$ or $\tilde{\mathbf{u}} \cdot \mathbf{w}_1^\omega < 0$. Utilities within the same hemisphere induce identical rankings, yielding perfect correlation. Figure 4 (left) shows that in our experiments: under ω_{shap} and ω_{beta}, the utilities $\tilde{\mathbf{u}}_{\text{acc}}$, $\tilde{\mathbf{u}}_{\text{rec}}$, and $\tilde{\mathbf{u}}_{\text{am}}$ lie in the *same hemisphere*, explaining their strong correlation. Under ω_{banz}, $\tilde{\mathbf{u}}_{\text{rec}}$ aligns *exactly* with the boundary point H_ω, satisfying $\tilde{\mathbf{u}}_{\text{rec}} \cdot \mathbf{e}_i = 0 \quad \forall i$ which orthogonalizes it to the collinear direction $\mathbf{w}_1^\omega$. This degeneracy equalizes utility values across all points, producing tied rankings and the observed weak correlation with $\tilde{\mathbf{u}}_{\text{acc}}$ and $\tilde{\mathbf{u}}_{\text{am}}$. Figures 15, 16, 17, 18, 19, 20 and 21 in Appendix C.6 present similar visualizations for other experimental datasets. We observe that Banzhaf consistently produces nearly collinear spatial signatures across all datasets, leading to reduced ranking diversity as described in Theorem 4.8. In contrast, Shapley and Beta Shapley exhibit more dataset-dependent structures, sometimes aligning with Banzhaf but often showing greater dispersion, resulting in higher ranking variability.

To further illustrate ranking diversity in this setting, particularly transitions across ranking regions, we present Figure 4, which depicts the evolution of the Kendall rank correlation along a convex combination path of utility functions in $\mathcal{S}^1$. This path is specifically designed to cross one of the two antipodals points in H_ω that define the ranking regions, encompassing the three key utilities we analyzed $\tilde{\mathbf{u}}_{\text{acc}}, \tilde{\mathbf{u}}_{\text{rec}}, \tilde{\mathbf{u}}_{\text{am}}$. The resulting visualization captures how rank correlation fluctuates as the utility function moves along this path. To formally derive this figure, we rely on the following proposition.

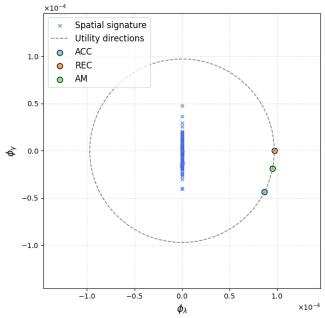
Proposition 4.9. *Let $\mathcal{D}$ be a dataset, ω a semivalue and let $u, v \in \mathcal{U}_{\lambda,\gamma}$ be two utilities respectively characterized by $(c_0, c_1, c_2, d_0) \in \mathbb{R}^4$ and $(c_0', c_1', c_2', d_0') \in \mathbb{R}^4$. The Kendall rank correlation between the data values sets $\{\phi(z, \omega, u)\}_{z\in\mathcal{D}}$ and $\{\phi(z, \omega, v)\}_{z\in\mathcal{D}}$ denoted as $\tau(u, v)$ is defined as*



(a) Spatial signature under ω_{shap}



(b) Spatial signature under ω_{beta}



(c) Spatial signature under ω_{banz}

Figure 2. These three figures illustrate the spatial signatures of the WIND dataset under three semivalues: $\omega_{\text{shap}}, \omega_{\text{beta}}, \omega_{\text{banz}}$. The unit circle represents the set of distinct utilities in terms of ranking, with markers indicating the normalized utility vectors $\tilde{\mathbf{u}}_{\text{acc}}$, $\tilde{\mathbf{u}}_{\text{rec}}$, and $\tilde{\mathbf{u}}_{\text{am}}$ for the three utility functions used in our experiments, namely the accuracy, the recall, and the arithmetic mean that belong to $\mathcal{U}_{\lambda,\gamma}$.

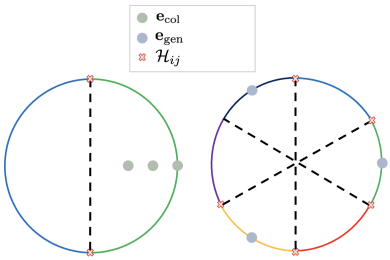


Figure 3. Ranking regions induced by linear utilities on the unit circle $\mathcal{S}^1$ for collinear (left) and general-position (right) spatial signatures of three data points. In the collinear case, only two ranking regions exist—rankings flip when $\tilde{\mathbf{u}}$ crosses the direction orthogonal to $\mathbf{w} = (1, 0)$. In the general-position case, six distinct ranking regions emerge. The red crosses correspond to the sets $\mathcal{H}_{ij} = \{\tilde{\mathbf{u}} \in \mathcal{S}^1 \mid \tilde{\mathbf{u}} \cdot (\mathbf{e}_i - \mathbf{e}_j) = 0\}$, which are the utility points for which two data points $\mathbf{e}_i$ and $\mathbf{e}_j$ have equal utility values, leading to ranking transitions. The black dashed lines partition $\mathcal{S}^1$ into ranking regions by connecting pairs of antipodal points $\mathcal{H}_{ij}$.

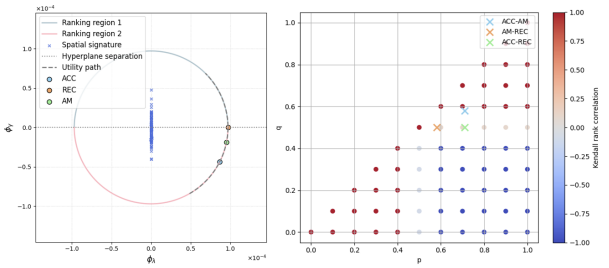


Figure 4. (Left) The spatial signature of the dataset is shown along with the unit circle of utilities $\mathcal{S}^1$, with the two ranking regions induced by different utility functions. The dashed curve represents a *convex path* of utility functions. The points corresponding to accuracy (ACC), recall (REC), and arithmetic mean (AM) utilities are marked. (Right) Kendall rank correlation between rankings induced by pairs of utility functions along the convex path as a function of parameters p and q. The color scale encodes the rank correlation, with blue indicating a negative correlation and red indicating a positive correlation. Pairs ACC-REC, ACC-AM and AM-REC are represented as crosses of different colors.

$$\tau(u, v) = \frac{1}{\binom{n}{2}} \sum_{i<j} \operatorname{sgn}\Big[\frac{1}{d_0 d_0'}[c_1 c_1' D_{ij}^2(\omega, \lambda) + (c_1 c_2' + c_1' c_2) D_{ij}(\omega, \lambda) D_{ij}(\omega, \gamma) + c_2 c_2' D_{ij}^2(\omega, \gamma)]\Big]$$

where $\operatorname{sgn}(x) = 1$ *if* $x > 0$, 0 *if* $x = 0$, *and* -1 *if* $x < 0$ *and for any utility* h, $D_{ij}(\omega, h) = \phi(z_i; \omega, h) - \phi(z_j; \omega, h)$.

5. Conclusion

This work advances the understanding of how utility functions impact semivalue-based data valuation. Our experiments show that the way a utility influences data value rankings is not solely dictated by its intrinsic properties but rather by its specific interaction with the dataset and the semivalue weights. We explain this interaction by introducing a geometric framework that clarifies empirical results.

Limitations. Our study focuses on a specific subclass of classification utility functions, limiting the generality of our theoretical findings. While this restriction enables a deeper analytical understanding, it remains unclear how our results extend to broader classes of utility functions. However, we have extended our experimental analysis to other utility functions beyond this subclass in Appendix C.5, and the empirical conclusions remain consistent. An extension of our theoretical framework to more general utility functions is a key direction for future research.

Broader impact statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Acknowledgments

This work was partially supported by the French National Research Agency (ANR) through grants ANR-20-CE23-0007 and ANR-23-CE23-0002 and through the PEPR IA FOUNDRY project (ANR-23-PEIA-0003).

References

Banzhaf, J. Weighted voting doesn't work: A mathematical analysis. *Rutgers Law Review*, 19(2):317–343, 1965.

Bloch, L. and Friedrich, C. Data analysis with shapley values for automatic subject selection in alzheimer's disease data sets using interpretable machine learning. *Alzheimer's Research & Therapy*, 13, 09 2021. doi: 10.1186/s13195-021-00879-4.

Choi, S., Cha, S.-H., and Tappert, C. A survey of binary similarity and distance measures. *J. Syst. Cybern. Inf.*, 8, 11 2009.

Dubey, P., Neyman, A., and Weber, R. J. Value theory without efficiency. *Mathematics of Operations Research*, 6(1):122–128, 1981. ISSN 0364765X, 15265471. URL http://www.jstor.org/stable/3689271.

Garrido-Lucero, F., Heymann, B., Vono, M., Loiseau, P., and Perchet, V. Du-shapley: A shapley value proxy for efficient dataset valuation, 2024. URL https://arxiv.org/abs/2306.02071.

Ghorbani, A. and Zou, J. Data shapley: Equitable valuation of data for machine learning, 2019. URL https://arxiv.org/abs/1904.02868.

Ghorbani, A., Kim, M. P., and Zou, J. A distributional framework for data valuation, 2020. URL https://arxiv.org/abs/2002.12334.

Jia, R., Dao, D., Wang, B., Hubis, F. A., Gurel, N. M., Li, B., Zhang, C., Spanos, C. J., and Song, D. Efficient task-specific data valuation for nearest neighbor algorithms, 2020. URL https://arxiv.org/abs/1908.08619.

Jia, R., Dao, D., Wang, B., Hubis, F. A., Hynes, N., Gurel, N. M., Li, B., Zhang, C., Song, D., and Spanos, C. Towards efficient data valuation based on the shapley value, 2023. URL https://arxiv.org/abs/1902.10275.

Jiang, K. F., Liang, W., Zou, J., and Kwon, Y. Opendataval: a unified benchmark for data valuation, 2023. URL https://arxiv.org/abs/2306.10577.

Kendall, M. G. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938. ISSN 00063444. URL http://www.jstor.org/stable/2332226.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions, 2020. URL https://arxiv.org/abs/1703.04730.

Koyejo, O. O., Natarajan, N., Ravikumar, P. K., and Dhillon, I. S. Consistent binary classification with generalized performance metrics. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/30c8e1ca872524fbf7ea5c519ca397ee-Paper.pdf.

Kwon, Y. and Zou, J. Beta shapley: a unified and noise-reduced data valuation framework for machine learning, 2022. URL https://arxiv.org/abs/2110.14049.

Kwon, Y., A. Rivas, M., and Zou, J. Efficient computation and analysis of distributional shapley values. In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 793–801. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/kwon21a.html.

Northcutt, C. G., Athalye, A., and Mueller, J. Pervasive label errors in test sets destabilize machine learning benchmarks, 2021. URL https://arxiv.org/abs/2103.14749.

Pandl, K. D., Feiland, F., Thiebes, S., and Sunyaev, A. Trustworthy machine learning for health care: scalable data valuation with the shapley value. In *Proceedings of the Conference on Health, Inference, and Learning*, CHIL '21, pp. 47–57, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383592. doi: 10.1145/3450439.3451861. URL https://doi.org/10.1145/3450439.3451861.

Pozzolo, A. D., Caelen, O., Johnson, R. A., and Bontempi, G. Calibrating probability with undersampling for unbalanced classification. In *2015 IEEE Symposium Series on Computational Intelligence*, pp. 159–166, 2015. doi: 10.1109/SSCI.2015.33.

Shapley, L. S. A value for n-person games. In Kuhn, H. W. and Tucker, A. W. (eds.), *Contributions to the Theory of Games II*, pp. 307–317. Princeton University Press, Princeton, 1953.

Sim, R. H. L., Xu, X., and Low, B. K. H. Data valuation in machine learning: "ingredients", strategies, and open challenges. In Raedt, L. D. (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 5607–5614. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/782. URL https://doi.org/10.24963/ijcai.2022/782. Survey Track.

Spearman, C. The proof and measurement of association between two things. *American Journal of Psychology*, 15: 88–103, 1904.

Tang, S., Ghorbani, A., Yamashita, R., Rehman, S., Dunnmon, J. A., Zou, J., and Rubin, D. L. Data valuation for medical imaging using shapley value and application to a large-scale chest x-ray dataset. *Scientific Reports*, 11(1), April 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-87762-2. URL http://dx.doi.org/10.1038/s41598-021-87762-2.

Vats, D. and Knudson, C. Revisiting the gelman-rubin diagnostic, 2020. URL https://arxiv.org/abs/1812.09384.

Wang, J. T. and Jia, R. Data banzhaf: A robust data valuation framework for machine learning, 2023. URL https://arxiv.org/abs/2205.15466.

Wang, J. T., Yang, T., Zou, J., Kwon, Y., and Jia, R. Rethinking data shapley for data selection tasks: Misleads and merits, 2024. URL https://arxiv.org/abs/2405.03875.

Yan, T. and Procaccia, A. D. If you like shapley then you'll love the core. In *AAAI Conference on Artificial Intelligence*, 2021. URL https://api.semanticscholar.org/CorpusID:211108182.

Zheng, K., Chua, H.-R., Herschel, M., Jagadish, H. V., Ooi, B. C., and Yip, J. W. L. Exploiting negative samples: a catalyst for cohort discovery in healthcare analytics. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

A. Additional definitions

Definition A.1. *(Spearman rank correlation).* The Spearman rank correlation coefficient ρ_S measures the monotonic relationship between two sets of values $\{\phi_i^u\}_{i\in[n]}$ and $\{\phi_i^v\}_{i\in[n]}$ assigned to n elements. Let $\mathrm{rg}(\phi_i^u)$ and $\mathrm{rg}(\phi_i^v)$ denote the ranks of ϕ_i^u and ϕ_i^v within their respective sets. The Spearman coefficient $\rho_S(u,v)$ is defined as the Pearson correlation coefficient between the ranked values:

$$\rho(u,v) = \frac{\mathrm{Cov}\left(\mathrm{rg}(\phi^u), \mathrm{rg}(\phi^v)\right)}{\sigma_{\mathrm{rg}(\phi^u)} \cdot \sigma_{\mathrm{rg}(\phi^v)}}$$

where Cov denotes covariance, $\sigma_{\mathrm{rg}(\phi^u)}$ and $\sigma_{\mathrm{rg}(\phi^v)}$ are the standard deviations of the ranks of $\{\phi_i^u\}$ and $\{\phi_i^v\}$, respectively.

If there are no tied ranks, $\rho_S(u,v)$ simplifies to:

$$\rho_S(u,v) = 1 - \frac{6\sum_{i=1}^n d_i^2}{n(n^2-1)},$$

where $d_i = \mathrm{rg}(\phi_i^u) - \mathrm{rg}(\phi_i^v)$.

The coefficient $\rho_S(u,v)$ satisfies $-1 \leq \rho_S(u,v) \leq 1$, where $\rho_S(u,v) = 1$ indicates perfect monotonic agreement between rankings (ranks increase identically), $\rho_S(u,v) = -1$ indicate perfect monotonic disagreement (ranks are inversely ordered) and $\rho_S(u,v) = 0$ indicates no monotonic association.

We consider a binary classification task, where each sample belongs to one of two classes $\in \{0,1\}$. Given a dataset of size n, we define the following classification utility functions:

Definition A.2. *(Accuracy).* Accuracy measures the proportion of correctly classified instances and is defined as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}},$$

where TP (true positives), TN (true negatives), FP (false positives), and FN (false negatives) denote the number of samples correctly or incorrectly classified with respect to the positive and negative classes.

Definition A.3. *(Recall).* Also known as true positive rate (TPR), recall quantifies the model's ability to correctly identify positive instances. It is given by:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

Definition A.4. *(Arithmetic mean).*

$$\text{AM} = \frac{1}{2}(\text{TPR} + \text{TNR})$$

Definition A.5. *(F1-score).* The F1-score provides a balance between precision and recall by computing their harmonic mean. It is defined as:

$$\text{F1} = 2 \cdot \frac{P \cdot R}{P + R},$$

where precision P is given by:

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Definition A.6. *(Negative log loss).* Negative log loss (NLL) evaluates the confidence of probabilistic predictions. Given a dataset of size n, where each sample has a true label $y_i \in \{0,1\}$ and a predicted probability $p_i \in [0,1]$ for class 1, NLL is defined as

$$\text{NLL} = -\frac{1}{n}\sum_{i=1}^{n} \left[y_i \log p_i + (1-y_i)\log(1-p_i)\right].$$

B. Proofs & complementary analytical results

B.1. Proofs of propositions

Proposition B.1. *(Restate of Proposition 4.1) Let $\mathcal{D} = \{z_i\}_{i\in[n]}$ be a dataset and ω a semivalue weight vector. For any utility function $u \in \mathcal{U}_{\lambda,\gamma}$ characterized by coefficients $(c_0, c_1, c_2, d_0) \in \mathbb{R}^4$, the data value assigned to z_i can be decomposed as*

$$\phi(z_i;\omega,u) = \frac{c_1}{d_0}\phi(z_i;\omega,\lambda) + \frac{c_2}{d_0}\phi(z_i;\omega,\gamma)$$

Proof. By assumption, the utility function u satisfies the decomposition $u(S) = \frac{c_0+c_1\lambda(S)+c_2\gamma(S)}{d_0}, \quad \forall S \subseteq \mathcal{D}$. Thus, for all $z_i \in \mathcal{D}$, $\phi(z_i;\omega,u) = \phi\left(z_i;\omega,\frac{c_0+c_1\lambda+c_2\gamma}{d_0}\right)$. Using the linearity axiom of semivalues defined in 2.1, we get $\phi(z_i;\omega,u) = \frac{1}{d_0}\left(c_0\phi(z_i;\omega,1) + c_1\phi(z_i;\omega,\lambda) + c_2\phi(z_i;\omega,\gamma)\right)$. Since $\phi(z_i;\omega,1) = 0$, we obtain

$$\phi(z_i;\omega,u) = \frac{c_1}{d_0}\phi(z_i;\omega,\lambda) + \frac{c_2}{d_0}\phi(z_i;\omega,\gamma).$$

□

Proposition B.2. *(Restate of Proposition 4.5) Let $\mathbf{u} \in \mathcal{U}^*$. The set of utilities that share the same direction as $\mathbf{u}$ is in bijection with the unit sphere $\mathcal{S}^1 \subset \mathcal{U}^*$ and is uniquely represented by:*

$$\tilde{\mathbf{u}} = \frac{\mathbf{u}}{\|\mathbf{u}\|} = \frac{(c_1,c_2)}{\sqrt{c_1^2+c_2^2}}$$

Proof. We proceed in two steps: first, proving surjectivity, then proving injectivity.

- *Surjectivity.* Every $\mathbf{u} \in \mathcal{U}^*$ can be normalized to $\tilde{\mathbf{u}} = \frac{\mathbf{u}}{\|\mathbf{u}\|} \in \mathcal{S}^1$ with the ranking induced by $\mathbf{u}$ identical to that induced by $\tilde{\mathbf{u}}$.
- *Injectivity.* Suppose $\tilde{\mathbf{u}}_1, \tilde{\mathbf{u}}_2 \in \mathcal{S}^1$ induce the same ranking. Then $\tilde{\mathbf{u}}_1 = \tilde{\mathbf{u}}_2$. In fact, if $\tilde{\mathbf{u}}_1 \neq \tilde{\mathbf{u}}_2$, there exists a point $\mathbf{e}_i$ such that $\tilde{\mathbf{u}}_1 \cdot \mathbf{e}_i > 0$ but $\tilde{\mathbf{u}}_2 \cdot \mathbf{e}_i < 0$, violating identical rankings.

Thus, the map $\mathbf{u} \mapsto \tilde{\mathbf{u}}$ sends each directionally equivalent utility to a unique point on $\mathcal{S}^1$. □

Proposition B.3. *(Restate of Proposition 4.9) Let $\mathcal{D}$ be a dataset, ω a semivalue and let $u, v \in \mathcal{U}_{\lambda,\gamma}$ be two utilities respectively characterized by $(c_0, c_1, c_2, d_0) \in \mathbb{R}^4$ and $(c_0', c_1', c_2', d_0') \in \mathbb{R}^4$. The Kendall rank correlation between the data values sets $\{\phi(z,\omega,u)\}_{z\in\mathcal{D}}$ and $\{\phi(z,\omega,v)\}_{z\in\mathcal{D}}$ denoted as $\tau(u,v)$ is defined as*

$$\tau(u,v) = \frac{1}{\binom{n}{2}}\sum_{i<j}\operatorname{sgn}\left[\frac{1}{d_0 d_0'}[c_1c_1'D_{ij^2}(\omega,\lambda) + (c_1c_2' + c_1'c_2)D_{ij}(\omega,\lambda)D_{ij}(\omega,\gamma) + c_2c_2'D_{ij}^2(\omega,\gamma)]\right]$$

where $\operatorname{sgn}(x) = 1$ *if* $x > 0$, 0 *if* $x = 0$, *and* -1 *if* $x < 0$ *and for any utility* h, $D_{ij}(\omega,h) = \phi(z_i;\omega,h) - \phi(z_j;\omega,h)$.

Proof. From Proposition 4.1, we have for all $i \in [n]$,

$$\phi(z_i;\omega,u) = \frac{c_1}{d_0}\phi(z_i;\omega,\lambda) + \frac{c_2}{d_0}\phi(z_i;\omega,\gamma).$$

$$\phi(z_i;\omega,v) = \frac{c_1'}{d_0'}\phi(z_i;\omega,\lambda) + \frac{c_2'}{d_0'}\phi(z_i;\omega,\gamma).$$

For any pair (i,j),

$$D_{ij}(\omega,u) = \phi(z_i;\omega,u) - \phi(z_j;\omega,u) = \frac{c_1}{d_0}D_{ij}(\omega,\lambda) + \frac{c_2}{d_0}D_{ij}(\omega,\gamma).$$

$$D_{ij}(\omega, v) = \phi(z_i; \omega, v) - \phi(z_j; \omega, v) = \frac{c_1'}{d_0'} D_{ij}(\omega, \lambda) + \frac{c_2'}{d_0'} D_{ij}(\omega, \gamma).$$

By Equation (2),

$$\tau(u, v) = \frac{1}{\binom{n}{2}} \sum_{i<j} \operatorname{sgn}[D_{ij}(\omega, u) D_{ij}(\omega, v)] \tag{5}$$

with $D_{ij}(\omega, u) D_{ij}(\omega, v) = \frac{1}{d_0 d_0'}[c_1 c_1' D_{ij}^2(\omega, \lambda) + (c_1 c_2' + c_1' c_2) D_{ij}(\omega, \lambda) D_{ij}(\omega, \gamma) + c_2 c_2' D_{ij}^2(\omega, \gamma)]$. □

B.2. Proofs of theorems

Theorem B.4. *(Restate of Theorem 4.3) Let $\mathcal{P}_\omega \subset \mathbb{R}^2$ denote the space where each data point $z_i \in \mathcal{D}$ is embedded as $\mathbf{e_i} = \big(\phi(z_i; \omega, \lambda), \phi(z_i; \omega, \gamma)\big)$. Let $\mathcal{U}^* \subset \mathbb{R}^2$ denote the space of linear utilities, where each utility function $\mathbf{u} \in \mathcal{U}^*$ is represented as $\mathbf{u} = \left(\frac{c_1}{d_0}, \frac{c_2}{d_0}\right)$. Then, $\mathcal{U}^*$ is isomorphic to the dual space $\mathcal{P}_\omega^*$ of $\mathcal{P}_\omega$.*

Proof. We define the map $\Psi : \mathcal{U}^* \to \mathcal{P}_\omega^*$ by

$$\Psi(\mathbf{u})(\mathbf{e}_i) = \mathbf{u} \cdot \mathbf{e_i} = \frac{c_1}{d_0} \phi(z_i; \omega, \lambda) + \frac{c_2}{d_0} \phi(z_i; \omega, \gamma)$$

where $\mathbf{u} = \left(\frac{c_1}{d_0}, \frac{c_2}{d_0}\right)$. This map is linear by the linearity of the dot product.

- *Injectivity of* Ψ. Suppose $\Psi(\mathbf{u}) = \Psi(\mathbf{v})$. Then, for all $\mathbf{e_i} \in \mathcal{P}_\omega$,

  $$\mathbf{u} \cdot \mathbf{e}_i = \mathbf{v} \cdot \mathbf{e}_i$$

  Since $\mathcal{P}_\omega$ spans $\mathbb{R}^2$ (λ and γ are linearly independent utilities), this implies $\mathbf{u} = \mathbf{v}$. Thus, Ψ is injective.

- *Surjectivity of* Ψ. Any linear functional $f \in \mathcal{P}^*$ can be written as

  $$f = \alpha \mathbf{e}^\lambda + \beta \mathbf{e}^\gamma \quad \text{for } \alpha, \beta \in \mathbb{R}$$

  where $(\mathbf{e}^\lambda, \mathbf{e}^\gamma)$ is the dual basis. Setting $\alpha = \frac{c_1}{d_0}$ and $\beta = \frac{c_2}{d_0}$, we recover $\mathbf{u} = \left(\frac{c_1}{d_0}, \frac{c_2}{d_0}\right) \in \mathcal{U}^*$ such that $f = \Psi(\mathbf{u})$. Thus, Ψ is subjective.

Since Ψ is a linear bijection (both injective and surjective), it is an isomorphism. □

Theorem B.5. *(Restate of Theorem 4.8) Let $\mathbf{e} = (\mathbf{e}_i)_{i \in [n]}$ be a spatial signature in $\mathcal{P}_\omega$. Define the ranking regions as the connected components of the unit sphere $\mathcal{S}^1$ where the linear utilities $\tilde{\mathbf{u}} \in \mathcal{S}^1$ induce identical rankings on $\mathcal{D}$. Then,*

1. *if $\mathbf{e}$ is in general position, the number of distinct ranking regions is maximal equals to $R_{gen}(n) = 2 \times \binom{n}{2}$.*
2. *if $\mathbf{e}$ is collinear, the number of distinct ranking regions is minimal as it collapses to $R_{col}(n) = 2$.*

Proof. When analyzing the ranking regions induced by linear utilities on $\mathcal{S}^1$, we consider how the set of inequalities $\tilde{\mathbf{u}} \cdot \mathbf{e}_i > \tilde{\mathbf{u}} \cdot \mathbf{e}_j$ partitions $\mathcal{S}^1$. These partitions are determined by each set $H_{ij} = \left\{\tilde{\mathbf{u}} \in \mathcal{S}^1 \mid \tilde{\mathbf{u}} \cdot (\mathbf{e}_i - \mathbf{e}_j) = 0\right\}$, which consists of the points on $\mathcal{S}^1$ where the ranking order of $\mathbf{e}_i$ and $\mathbf{e}_j$ flips.

1. If $\mathbf{e}$ is in general position, then for every pair (i, j), the differences $\mathbf{e}_i - \mathbf{e}_j$ are pairwise non-collinear (since in general position, there is no collinear triples or radial pairs). This implies that for each pair (i, j), the equation $\tilde{\mathbf{u}} \cdot (\mathbf{e}_i - \mathbf{e}_j) = 0$ defines a unique line through the origin in $\mathbb{R}^2$, which intersects $\mathcal{S}^1$ at exactly two distinct antipodal points. These two points partition $\mathcal{S}^1$ into two connected regions, corresponding to the two possible rankings between $\mathbf{e}_i$ and $\mathbf{e}_j$. Since there are $\binom{n}{2}$ such pairs and each contributes two distinct partitioning points, the total number of connected ranking regions is $R_{\text{gen}}(n) = 2 \times \binom{n}{2}$.

2. If $\mathbf{e}$ is collinear, all embedded points satisfy $\mathbf{e}_i = k_i\mathbf{w}$ for $\mathbf{w} \in \mathcal{P}_\omega$ and scalars $k_i \in \mathbb{R}$. The differences satisfy $\mathbf{e}_i - \mathbf{e}_j = (k_i - k_j)\mathbf{w}$, which means that all differences are collinear with $\mathbf{w}$. Consequently, the separating sets H_{ij} are all aligned with the unique perpendicular direction $\mathbf{w}^\perp$, which is a single line through the origin in $\mathbb{R}^2$, intersecting $\mathcal{S}^1$ at exactly two antipodal points. Since all pairwise ranking reversals occur along this single separating direction, all ranking regions collapse into just two distinct regions $R_{\text{col}}(n) = 2$. These two regions correspond to the cases $\tilde{\mathbf{u}} \cdot \mathbf{w} > 0$ and $\tilde{\mathbf{u}} \cdot \mathbf{w} < 0$, meaning that all rankings in each half-space are identical.

□

B.3. Remark on the interpretation of low-rank correlations

Caution is essential when interpreting rank correlations from different utility functions. Poor rank correlations do not necessarily imply inconsistency in the assigned values for a data valuation application; rather, the values generated using different utility functions can still be equally useful for this task. This may occur when data points are clustered into groups of similar values for both utility functions, while the rankings within those groups differ depending on the utility function used. Figure 5 and the analytical example above illustrate this idea.

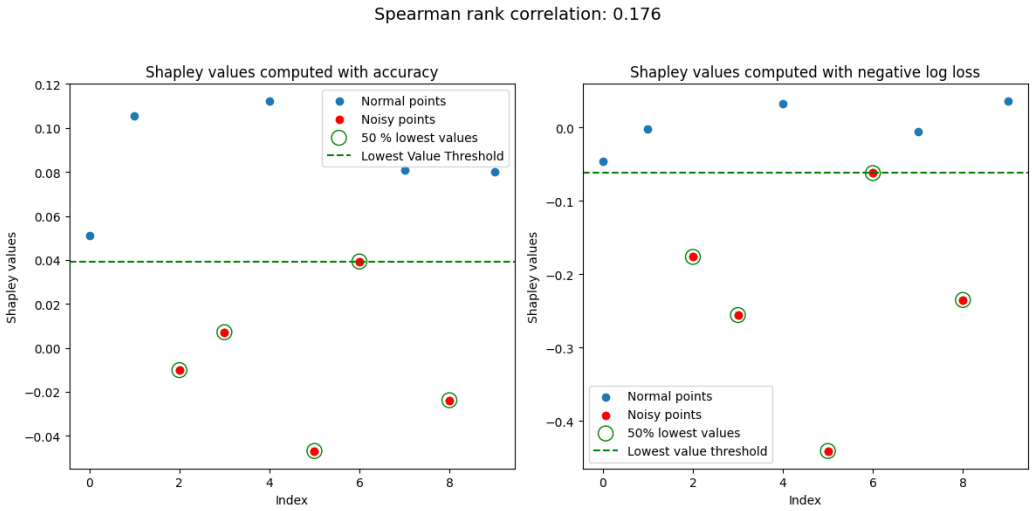


Figure 5. Illustration of Shapley values computed with accuracy and negative log loss as utility functions on the open-source dataset WIND which has been split into a training set with 10 points (50% of which have noisy labels due to label flipping) and a test set with 200 points. The five lowest-valued points (highlighted with green circles) are observed to be associated with noisy points (red) across both utility functions. This figure demonstrates that poor rank correlations (Spearman correlation of 0.176) between different utility functions do not necessarily imply inconsistency in the data valuation task. Despite the low correlation, the five lowest points consistently belong to the same group, indicating that the points are correctly grouped into clusters of similar values.

Analytical illustration of low-rank correlation while clustering effect. Suppose we have a training dataset $\mathcal{D} = \{z_i\}_{i\in[n]}$ partitioned as $\mathcal{D} = \mathcal{D}_{\text{high}} \cup \mathcal{D}_{\text{low}}$ where $\mathcal{D}_{\text{high}}$ contains $n_{\text{high}} = k \cdot m$ points and $\mathcal{D}_{\text{low}}$ contains $n_{\text{high}} = m$ points and. Here, k is a positive constant representing the ratio of the sizes of the two subsets, so $n = (k+1) \cdot m$. Without loss of generality, we assume the indices in $\mathcal{D}$ are structured such that points in $\mathcal{D}_{\text{high}}$ have indices 1 through $k \cdot m$ and those in $\mathcal{D}_{\text{low}}$ have indices $k \cdot (m+1)$ through n. Let $U : 2^{\mathcal{D}} \to \mathbb{R}^+$ be a function defined on any subset $S \subseteq \mathcal{D}$ such as $U(S) = \frac{|\{z_i \in S : z_i \in \mathcal{D}_{\text{high}}\}|}{C_\omega}$ where $C_\omega = \sum_{S\subseteq\mathcal{D}\setminus\{z\}} \omega(S)$, $z \in \mathcal{D}$.[3] We define two utility functions U_+ and U_-, as follows:

$$U_+(S) = U(S) + \sum_{z_i \in S} i \cdot \varepsilon \quad \text{and} \quad U_-(S) = U(S) - \sum_{z_i \in S} i \cdot \varepsilon$$

where $\varepsilon > 0$. We assume ε to be sufficiently small to satisfy $C_\omega \cdot k \cdot m \cdot \varepsilon < 1$. Let $\varphi(z_i; \omega, V)$ denote the semivalue characterized by ω for each point z_i under a utility function V. Then, under this setting, we have:

[3]Since the value of C_ω depends only on the size of $\mathcal{D}\setminus\{z\}$ (which has $n-1$ elements if $\mathcal{D}$ has n elements), and not on the specific point z being excluded, C_ω remains the same regardless of the choice of z.

(i) *Cluster separation*: For any $z_i \in \mathcal{D}_{\text{high}}$ and $z_j \in \mathcal{D}_{\text{low}}$,

$$\varphi(z_i;\omega,U_+) > \varphi(z_j;\omega,U_+) \quad \text{and} \quad \varphi(z_i;\omega,U_-) > \varphi(z_j;\omega,U_-)$$

(ii) *Low Spearman rank correlation*: The Spearman rank correlation between both sets of values $\{\varphi(z_i;\omega,U_+)\}_{z_i\in\mathcal{D}}$ and $\{\varphi(z_i;\omega,U_-)\}_{z_i\in\mathcal{D}}$ denoted as $\rho_s(k,m)$ equals:

$$\rho_s(k,m) = 1 - \frac{6\left(\sum_{j=1}^{k\cdot m}(2j-k\cdot m-1)^2 + \sum_{j=1}^{m}(2j-m-1)^2\right)}{(k+1)\cdot m\left(((k+1)\cdot m)^2-1\right)}$$

and for m fixed, $\exists\, k^*$ such that $\rho_s(k^*,m)\approx 0$.

Proof. (i) *Cluster separation*: For any $z_i \in \mathcal{D}_{\text{high}}$ and $z_j \in \mathcal{D}_{\text{low}}$, we have by definition of U,

$$\varphi(z_i;\omega,U) = \sum_{S\subseteq\mathcal{D}\setminus\{z_i\}} \omega(S)[U(S\cup\{z_i\}) - U(S)] = \sum_{S\subseteq\mathcal{D}\setminus\{z_i\}} \omega(S)\cdot\frac{1}{C_\omega} = 1$$
$$\varphi(z_j;\omega,U) = \sum_{S\subseteq\mathcal{D}\setminus\{z_j\}} \omega(S)[U(S\cup\{z_j\}) - U(S)] = 0.$$

Then, we use the linearity of the semivalue to express the following:

$$\varphi(z_i;\omega,U_+) = \varphi(z_i;\omega,U) + C_\omega\cdot i\cdot\varepsilon = 1 + C_\omega\cdot i\cdot\varepsilon$$
$$\varphi(z_j;\omega,U_+) = \varphi(z_j;\omega,U) + C_\omega\cdot j\cdot\varepsilon = C_\omega\cdot j\cdot\varepsilon$$

Similarly, for U_-, we have:

$$\varphi(z_i;\omega,U_-) = \varphi(z_i;\omega,U) - C_\omega\cdot i\cdot\varepsilon = 1 - C_\omega\cdot i\cdot\varepsilon$$
$$\varphi(z_j;\omega,U_-) = \varphi(z_j;\omega,U) - C_\omega\cdot j\cdot\varepsilon = -C_\omega\cdot j\cdot\varepsilon$$

Since $0 < C_\omega\cdot j\cdot\varepsilon < 1$ and $0 < C_\omega\cdot i\cdot\varepsilon < 1$, we conclude that:

$$\varphi(z_i;\omega,U_+) < \varphi(z_j;\omega,U_+) \quad \text{and} \quad \varphi(z_i;\omega,U_-) < \varphi(z_j;\omega,U_-).$$

(ii) *Low Spearman rank correlation*: The Spearman rank correlation ρ_s is defined as:

$$\rho_s = 1 - \frac{6\sum_{z_i\in\mathcal{D}}^{n} d_i^2}{n(n^2-1)},$$

where $d_i = \text{rank}[\varphi(z_i;\omega,U_+)] - \text{rank}[\varphi(z_i;\omega,U_-)]$ is the difference in ranks for each data point z_i, and $n = (k+1)\cdot m$ is the total number of points in the dataset.

For $\mathcal{D}_{\text{high}}$, based on the values computed in (i), the ranks of $\{\varphi(z_i;\omega,U_+)\}_{z_i\in\mathcal{D}_{\text{high}}}$ are in descending order from $k\cdot m$ to 1, and the ranks of $\{\varphi(z_i;\omega,U_-)\}_{z_i\in\mathcal{D}_{\text{high}}}$ are in ascending order from 1 to $k\cdot m$. Therefore, for each $z_i\in\mathcal{D}_{\text{high}}$, the rank difference d_i is:

$$d_i = \text{rank}[\varphi(z_i;\omega,U_-)] - \text{rank}[\varphi(z_i;\omega,U_+)] = i - (k\cdot m - i + 1) = 2i - k\cdot m - 1.$$

For $\mathcal{D}_{\text{low}}$, based on the values computed in (i), the ranks of $\{\varphi(z_i;\omega,U_+)\}_{z_i\in\mathcal{D}_{\text{low}}}$ are from $k\cdot m+1$ to $(k+1)\cdot m$, and the ranks of $\{\varphi(z_i;\omega,U_-)\}_{z_i\in\mathcal{D}_{\text{low}}}$ are in descending order from $n=(k+1)\cdot m$ to $k\cdot m+1$. Therefore, for each $z_i\in\mathcal{D}_{\text{low}}$, the rank difference d_i is:

$$d_i = \text{rank}[\varphi(z_i;\omega,U_-)] - \text{rank}[\varphi(z_i;\omega,U_+)] = (k\cdot m + i) - ((k+1)\cdot m - i + 1) = 2i - m - 1.$$

Hence, the total sum of squared rank differences is:

$$\sum_{z_i\in\mathcal{D}} d_i^2 = \sum_{z_i\in\mathcal{D}_{\text{high}}} d_i^2 + \sum_{z_i\in\mathcal{D}_{\text{low}}} d_i^2 = \sum_{i=1}^{n_{\text{high}}}(2i-k\cdot m-1)^2 + \sum_{i=1}^{n_{\text{low}}}(2i-m-1)^2 = \sum_{i=1}^{k\cdot m}(2i-k\cdot m-1)^2 + \sum_{i=1}^{m}(2i-m-1)^2.$$

Substituting this into the formula for ρ_s, we obtain the Spearman rank correlation:

$$\rho_s(k,m) = 1 - \frac{6\left(\sum_{i=1}^{k\cdot m}(2i - k\cdot m - 1)^2 + \sum_{i=1}^{m}(2i - m - 1)^2\right)}{(k+1)\cdot m\left(((k+1)\cdot m)^2 - 1\right)}.$$

To illustrate the fact that for m fixed, $\exists\, k^*$ such that $\rho_s(k^*, m) \approx 0$, we calculate $\rho_s(4, 2)$ as an example. For $k = 4$ and $m = 2$, we have $n_{\text{high}} = 4 \cdot 2 = 8$ points in $\mathcal{D}_{\text{high}}$ and $n_{\text{low}} = 2$ points in $\mathcal{D}_{\text{low}}$, with a total dataset size of $n = (4 + 1) \cdot 2 = 10$ points. Substituting these values into the expression for ρ_s, we get:

$$\rho_s(4,2) = 1 - \frac{6\left(\sum_{i=1}^{8}(2i - 8 - 1)^2 + \sum_{i=1}^{2}(2i - 2 - 1)^2\right)}{5\cdot 2\cdot((5\cdot 2)^2 - 1)} \approx -0.03$$

This yields a Spearman rank correlation near zero, supporting our claim of low-rank correlation between the values $\{\varphi(z_i; \omega, U_+)\}_{i\in\mathcal{D}}$ and $\{\varphi(z_i; \omega, U_-)\}_{i\in\mathcal{D}}$ for this particular choice of k and m. Furthermore, to generalize this observation, we plot $\rho_s(k, m)$ for various values of m and k (see Figure 6). As shown in the figure, for each fixed m, choosing $k = 4$ leads to a Spearman rank correlation close to zero.

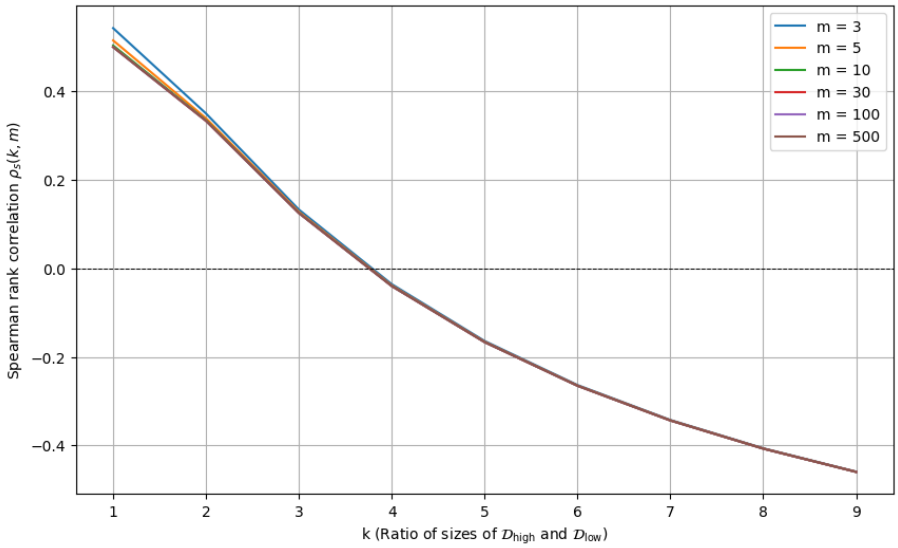


Figure 6. Spearman rank correlation $\rho_s(k, m)$ as a function of k for different values of m.

□

C. Additional settings & experiments

C.1. Datasets

Table 3 provides an overview of the datasets used in Section 3. These datasets are widely utilized in the data valuation literature (Ghorbani & Zou, 2019; Kwon & Zou, 2022; Jia et al., 2023; Wang & Jia, 2023; Jiang et al., 2023). Due to the computational cost associated with repeated model retraining in our experiments, we select a subset of 100 instances for training and 50 instances for testing from each dataset.

Dataset	Source
Breast	https://www.openml.org/d/13
Titanic	https://www.openml.org/d/40945
Credit	(Pozzolo et al., 2015)
Heart	https://www.openml.org/d/43398
Wind	https://www.openml.org/d/847
CPU	https://www.openml.org/d/761
2DPlanes	https://www.openml.org/d/727
Pol	https://www.openml.org/d/722

Table 3. A summary of datasets used in experiments from Section 3.

C.2. Systematic methodology for isolating utility effects in semivalue approximation

For a given dataset $\mathcal{D}$, we propose a systematic methodology to isolate the effect of utility functions on data valuation rankings by controlling for extraneous variability. This is achieved through two key principles:

1. A *fixed learning context* $\mathcal{L} = (\mathcal{A}, \mathcal{D}_{\text{test}})$, ensuring consistency in model training and evaluation.
2. *Consistent semivalue approximations* across utility functions to maintain comparability.

C.2.1. Fixed learning context $\mathcal{L}$

As outlined in Section 2, a utility function u is defined as:

$$u(S) = \text{PERF}(\mathcal{A}(S), \mathcal{D}_{\text{test}}),$$

where $\mathcal{A}$ is a learning algorithm that trains a model on a subset S, and PERF evaluates the model on a test set $\mathcal{D}_{\text{test}}$. The learning algorithm $\mathcal{A}$ specifies the model class, objective function, optimization procedure, and hyperparameters (e.g., learning rate, weight initialization). We define the *learning context* as $\mathcal{L} = (\mathcal{A}, \mathcal{D}_{\text{test}})$.

By fixing $\mathcal{L}$, the only varying factor is the performance metric PERF. Changing PERF (e.g., accuracy vs. recall) induces different utility functions u and v, while all other elements remain unchanged. Consequently, variations in data valuation rankings—quantified via Kendall's $\tau(u, v)$—are exclusively attributable to differences in the chosen performance metrics.

C.2.2. Accounting for approximation variability in semivalue computation

The above reasoning holds if *exact* semivalues are computed. However, in practice, semivalues are estimated using permutation sampling techniques, which introduce stochastic variability. This variability can affect data valuation rankings, making it difficult to attribute observed differences solely to the choice of utility function.

To control for this, we propose *aligned sampling* in addition to fixing the learning context $\mathcal{L}$. The fixed learning context ensures that all model training and evaluation factors remain constant, while aligned sampling guarantees that the same set of permutations is used across utility functions, eliminating variability introduced by the stochastic nature of semivalue approximations.

Fixed set of permutations. Let $\mathcal{P} = \{\pi_1, \pi_2, \ldots, \pi_m\}$ denote a fixed set of m random permutations of the data points in $\mathcal{D}$. The sampler applies this exact set of permutations across multiple utilities $\{u_1, u_2, \ldots, u_q\}$ such that $u_k(\cdot) = \text{PERF}_k[(\mathcal{A})(\cdot), \mathcal{D}_{\text{test}}]$ with fixed $\mathcal{L} = (\mathcal{A}, \mathcal{D}_{\text{test}})$ for all $k \in [q]$.

For a given performance metric PERF_k and the set of permutations $\mathcal{P}$, the sampler estimates the marginal contributions $\{\hat{\Delta}_j(z_i; u_k)\}_{j=1}^n$ for each data point $z_i \in \mathcal{D}$ with respect to the utility u_k such as

$$\hat{\Delta}_j(z_i; u_k) := \frac{1}{m} \sum_{s=1}^{m} \left(u_k\left(S_j^{\pi_s} \cup \{z_i\}\right) - u_k\left(S_j^{\pi_s}\right) \right),$$

where m is the number of permutations used, π_s denotes the s-th permutation and $S_j^{\pi_s}$ represents the subset of data points of size $j-1$ that precedes z_i in the order defined by permutation π_s.

Using the same permutation set $\mathcal{P}$ across performance metrics $u_1, u_2, \ldots, u_q$ ensures consistency in the sampling process. The sampler minimizes random variability by maintaining a fixed order of permutations, allowing each configuration to be evaluated in a stable and controlled framework. This setup provides a uniform basis for data valuation without introducing noise from randomly varying coalition structures.

Determining m: convergence and maximum number of permutations. The number of permutations m used in the marginal contribution estimator is determined based on both a maximum limit and a convergence criterion applied across all configurations. Specifically:

$$m = \max\left(m_{\text{min}}, \min\left(m_{\text{max}}, m_{\text{conv}}\right)\right),$$

where: m_{min} is a predefined minimum number of permutations to avoid starting convergence checks prematurely, m_{max} is a predefined maximum number of permutations set to control computational feasibility, m_{conv} is the smallest number of permutations required for the Gelman-Rubin (GR) (Vats & Knudson, 2020) statistic to converge across all utility functions $u_1, \ldots, u_q$. The use of the Gelman-Rubin statistic as a convergence criterion follows established practices in the literature (Jiang et al., 2023; Kwon & Zou, 2022).

Since m is required to be consistent across all performance metrics, the GR statistic must indicate convergence for each u before the sampling process stops. This means that if the GR statistic has not converged for even one utility, the computation of marginal contributions continues across all utilities.

For each data point z_i, the GR statistic R_i is computed for every 100 permutation across all utilities. The sampling process halts when the maximum GR statistic across all data points and all utilities falls below a threshold (e.g., 1.05), indicating convergence:

$$\max_k \max_{i=1,\ldots,n} R_i^k < \text{threshold}.$$

In this framework, the GR statistic, R_i^k, is used to assess the convergence of marginal contribution estimates for each data point z_i across multiple chains of sampled permutations under each utility u_k. The GR statistic evaluates the agreement between chains by comparing the variability within each chain to the variability across the chains, with convergence indicated when R_i^k approaches 1. Specifically, to compute the GR statistic for each data point z_i under u_k, we determine:

1. The within-chain variance W_i^k which captures the variability of marginal contributions for z_i within each chain. Specifically, if there are c independent chains, W_i^k is calculated as the average of the sample variances within each chain:

$$W_i^k = \frac{1}{C} \sum_{c=1}^{C} s_{i,c}^2,$$

where $s_{i,c}^2$ is the sample variance of marginal contributions for z_i within chain c. This term reflects the dispersion of estimates within each individual chain,

2. And the between-chain variance B_i^k which measures the variability between the mean marginal contributions across the chains. It indicates how much the chain means differ from each other. The between-chain variance is defined as:

$$B_i^k = \frac{n}{C-1} \sum_{c=1}^{C} \left(\bar{\Delta}_c(z_i; u_k) - \bar{\Delta}(z_i; u_k)\right)^2,$$

where $\bar{\Delta}_c(z_i; u_k)$ is the mean marginal contribution for z_i in chain c, and $\bar{\Delta}(z_i; u_k)$ is the overall mean across all chains:

$$\bar{\Delta}(z_i; u_k) = \frac{1}{C} \sum_{c=1}^{C} \bar{\Delta}_c(z_i; u_k).$$

The term B_i^k quantifies the extent of disagreement among the chain means.

Combining both W_i^k and B_i^k, the GR statistic R_i^k for data point z_i under utility u_k is defined as:

$$R_i^k = \sqrt{\frac{(n-1)}{n} + \frac{B_i^k}{W_i^k \cdot n}},$$

where n is the number of samples per chain.

Intra-permutation truncation. Building on existing literature (Ghorbani & Zou, 2019; Jiang et al., 2023), we improve computational efficiency by implementing an intra-permutation truncation criterion that restricts coalition growth once contributions stabilize. Given a random permutation $\pi_s \in \mathcal{P}$, the marginal contribution for each data point $z_{\pi_{s,j}}$ (the j-th point in the permutation π_s is calculated incrementally as the coalition size j increases from 1 up to n. However, instead of expanding the coalition size through all n elements, the algorithm stops increasing j when the marginal contributions become stable based on a relative change threshold.

For each step $l \in [n]$ within a permutation, the relative change V_l^k in the utility u_k is calculated as:

$$V_l^k := \frac{\left| u_k\left(\{z_{\pi_{s,j}}\}_{j=1}^{l} \cup \{z_{\pi_{s,l+1}}\}\right) - u_k\left(\{z_{\pi_{s,j}}\}_{j=1}^{l}\right)\right|}{u_k\left(\{z_{\pi_{s,j}}\}_{j=1}^{l}\right)}.$$

where $\{z_{\pi_{s,j}}\}_{j=1}^{l}$ represents the coalition formed by the first l data points in π_s. This measures the relative change in the utility u_k when adding the next data point to the coalition. The truncation criterion stops increasing the coalition size at the smallest value j satisfying the following condition:

$$j^* = \arg\min\left\{j \in [n] : \left|\{l \leq j : V_l \leq 10^{-8}\}\right| \geq 10\right\}.$$

This means that the coalition size j^* is fixed at the smallest j for which there are at least 10 prior values of V_l (for $l \leq j$ that are smaller than a threshold of 10^{-8}). This condition ensures that the utility u_k has stabilized, indicating convergence within the permutation. This intra-permutation truncation reduces computational overhead by avoiding unnecessary calculations once marginal contributions stabilize, thus improving efficiency.

Aggregating marginal contributions for semivalues estimation. Once the marginal contributions have been estimated consistently across all permutations and configurations, these contributions are aggregated to compute various semi-values, such as the Shapley, Banzhaf, and Beta Shapley values. Each semivalue method applies a specific weighting scheme to the marginal contributions to reflect the intended measure of data point importance.

For a data point z_i under utility u_k, the semivalue $\hat{\varphi}(z_i; \omega, u_k)$ is computed by applying a weighting function ω to the marginal contributions across coalition sizes:

$$\hat{\varphi}(z_i; \omega, u_k) = \sum_{j=1}^{n} \omega_j \, \hat{\Delta}_j(z_i; u_k),$$

where $\hat{\Delta}_j(z_i; u_k)$ is the estimated marginal contribution for coalition size j, and ω_j is the weight assigned to coalition size j in the semivalue calculation.

C.3. Learning algorithm $\mathcal{A}$

This section details the fixed learning algorithm $\mathcal{A}$ used throughout all our experiments.

Specifically, we use logistic regression as the model, trained with the L-BFGS optimization algorithm. The loss function is set to Binary Cross-Entropy (BCE), and ℓ_2-regularization is applied with a $\lambda = 1.0$ coefficient. Model weights are initialized from a fixed normal distribution, and the learning rate is fixed at 1.0.

C.4. Decision threshold calibration

We calibrate the decision threshold based on the proportion of positive labels in the training set. Instead of using a fixed threshold (e.g., 0.5), we adapt the threshold dynamically to match the expected class distribution. The calibration process

works as follows: first, it computes the target proportion of positive labels in the training set. Then, it sorts the predicted probabilities in ascending order and selects the threshold at the position corresponding to the fraction of negative samples. This ensures that the proportion of instances classified as positive matches the empirical distribution observed in the training set, leading to a more dataset-adaptive classification decision.

C.5. Additional experiments for Section 3

C.5.1. Extended experiments with additional classification utilities

To further explore the impact of utility function selection on data valuation rankings, we extend our experiments by replacing recall (REC) and arithmetic mean (AM) with F1-score (F1) and negative log-loss (NLL). While accuracy (ACC) remains a common baseline, these additional utility functions introduce distinct mathematical and statistical properties.

The F1-score is a *linear fractional measure*, as introduced in Section 4. While it shares similarities with linear performance measures such as accuracy, it remains fundamentally different, as accuracy corresponds to the special subset of this class for which $d_1 = d_2 = 0$. On the other hand, negative log-loss relies on *probabilistic predictions* rather than discrete classification decisions, introducing a continuous performance evaluation criterion that accounts for model confidence. We incorporate these utility functions to examine whether their different mathematical structures affect data valuation rankings.

The results in Table 4 show the Kendall rank correlations computed across these alternative utility functions for various datasets and three different semivalue-based data valuation methods: Data Shapley, $(4, 1)$-Beta Shapley, and Data Banzhaf.

DATASET	SHAPLEY			$(4,1)$-BETA SHAPLEY			BANZHAF		
	ACC-F1	ACC-NLL	F1-NLL	ACC-F1	ACC-NLL	F1-NLL	ACC-F1	ACC-NLL	F1-NLL
BREAST	0.98 (0.01)	-0.59 (0.02)	-0.6 (0.02)	0.98 (0.01)	-0.65 (0.01)	-0.66 (0.01)	0.98 (0.01)	0.18 (0.01)	0.18 (0.01)
TITANIC	0.63 (0.01)	-0.53 (0.01)	0.54 (0.01)	0.65 (0.01)	-0.60 (0.01)	-0.61 (0.01)	0.26 (0.02)	0.14 (0.02)	-0.07 (0.01)
CREDIT	0.30 (0.01)	-0.59 (0.02)	-0.43 (0.01)	0.39 (0.01)	-0.66 (0.01)	-0.49 (0.01)	0.05 (0.03)	0.38 (0.01)	0.28 (0.03)
HEART	0.80 (0.01)	-0.04 (0.02)	0.01 (0.02)	0.87 (0.01)	-0.20 (0.02)	-0.17 (0.03)	0.80 (0.01)	-0.07 (0.01)	-0.05 (0.01)
WIND	0.44 (0.01)	0.67 (0.02)	0.69 (0.01)	0.55 (0.02)	0.74 (0.02)	0.73 (0.01)	-0.17 (0.01)	0.26 (0.01)	0.44 (0.01)
CPU	0.84 (0.01)	0.55 (0.01)	0.68 (0.01)	0.87 (0.01)	0.59 (0.01)	0.69 (0.01)	-0.32 (0.01)	-0.53 (0.01)	0.52 (0.01)
2DPLANES	0.23 (0.02)	0.22 (0.02)	0.98 (0.01)	0.41 (0.01)	0.41 (0.01)	0.98 (0.01)	-0.05 (0.05)	-0.03 (0.01)	0.18 (0.01)
POL	0.58 (0.01)	0.58 (0.01)	0.79 (0.01)	0.78 (0.01)	0.74 (0.01)	0.81 (0.01)	-0.29 (0.02)	-0.01 (0.02)	0.13 (0.02)

Table 4. Kendall rank correlations with standard errors in parentheses between different utility function pairs (Accuracy-F1, Accuracy-Negative Log-Loss, and F1-Negative Log-Loss) across multiple datasets and three semivalues: Shapley, $(4, 1)$-Beta Shapley, and Banzhaf. Each semivalue is approximated 5 times using Monte Carlo sampling, and Kendall rank correlations are computed for each run. The reported values correspond to the mean correlation across the 5 runs, while the standard errors are derived from the standard deviation of these 5 estimates.

C.5.2. Spearman rank correlation results

We replicate all rank correlation computations using the Spearman rank correlation coefficient for completeness. The results presented in this section confirm that the trends observed with Kendall rank correlation persist under Spearman rank correlation.

Tables 5 and 6 report the Spearman rank correlation coefficients for the same experimental settings as in Tables 1 and 4, respectively, ensuring consistency in the evaluation across different ranking measures.‘

C.5.3. Intersection analysis

While rank correlations provide a measure of the consistency in the ordering of data values across different utility functions, they may not fully capture the grouping behavior that is often relevant in data valuation applications. As noted in B.3, a low-rank correlation between utility functions does not necessarily imply an inconsistent assigned importance of data points.

To better understand how changes in the utility function impact data point valuations, we conduct an *intersection analysis*. We examine the overlap in the bottom-$n\%$ of data points ranked by different utility functions to identify whether the least valuable points—according to various utility metrics—are consistently grouped together, even if their exact rankings differ. Figures 7, 8, 9, 10, 11, 12, 13, and 14 illustrate the results.

DATASET	SHAPLEY			$(4,1)$-BETA SHAPLEY			BANZHAF		
	ACC-REC	ACC-AM	REC-AM	ACC-REC	ACC-AM	REC-AM	ACC-REC	ACC-AM	REC-AM
BREAST	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)	0.90 (0.02)	0.99 (0.01)	0.89 (0.03)
TITANIC	0.56 (0.03)	0.91 (0.01)	0.82 (0.01)	0.62 (0.03)	0.93 (0.01)	0.84 (0.01)	-0.37 (0.05)	0.91 (0.02)	-0.08 (0.08)
CREDIT	0.46 (0.01)	0.67 (0.01)	0.94 (0.01)	0.50 (0.01)	0.75 (0.01)	0.92 (0.01)	-0.45 (0.01)	0.09 (0.02)	0.79 (0.01)
HEART	0.71 (0.02)	0.99 (0.01)	0.69 (0.02)	0.80 (0.02)	0.99 (0.01)	0.78 (0.02)	0.29 (0.02)	0.99 (0.01) (0.01)	0.27 (0.02)
WIND	0.92 (0.01)	0.99 (0.01)	0.91 (0.01)	0.93 (0.01)	0.99 (0.01)	0.92 (0.01)	0.12 (0.04)	0.99 (0.01)	0.11 (0.04)
CPU	0.93 (0.01)	0.99 (0.01)	0.97 (0.01)	0.93 (0.01)	0.99 (0.01)	0.97 (0.01)	0.01 (0.03)	0.90 (0.01)	0.27 (0.05)
2DPLANES	0.44 (0.03)	0.99 (0.01)	0.44 (0.03)	0.47 (0.03)	0.99 (0.01)	0.47 (0.03)	0.52 (0.01)	0.99 (0.01)	0.52 (0.01)
POL	0.75 (0.01)	0.90 (0.01)	0.42 (0.01)	0.74 (0.01)	0.93 (0.01)	0.48 (0.02)	0.85 (0.01)	0.87 (0.01)	0.52 (0.01)

Table 5. Spearman rank correlations with standard errors in parentheses between different utility function pairs (Accuracy-Recall, Accuracy-Arithmetic Mean, and Recall-Arithmetic Mean) across multiple datasets and three semivalues: Shapley, $(4,1)$-Beta Shapley, and Banzhaf. Each semivalue is approximated 5 times using Monte Carlo sampling, and Spearman rank correlations are computed for each run. The reported values correspond to the mean correlation across the 5 runs, while the standard errors are derived from the standard deviation of these 5 estimates.

DATASET	SHAPLEY			$(4,1)$-BETA SHAPLEY			BANZHAF		
	ACC-F1	ACC-NLL	F1-NLL	ACC-F1	ACC-NLL	F1-NLL	ACC-F1	ACC-NLL	F1-NLL
BREAST	0.99 (0.01)	-0.76 (0.02)	-0.78 (0.02)	0.99 (0.01)	-0.82 (0.01)	-0.83 (0.01)	0.99 (0.01)	0.22 (0.01)	0.23 (0.01)
TITANIC	0.81 (0.01)	-0.71 (0.01)	0.74 (0.01)	0.82 (0.01)	-0.79 (0.01)	-0.80 (0.01)	0.35 (0.02)	0.18 (0.02)	-0.20 (0.01)
CREDIT	0.44 (0.02)	-0.76 (0.02)	-0.61 (0.02)	0.55 (0.02)	-0.83 (0.01)	-0.68 (0.02)	0.06 (0.05)	0.53 (0.01)	0.40 (0.03)
HEART	0.94 (0.01)	-0.04 (0.02)	0.03 (0.03)	0.97 (0.01)	-0.28 (0.04)	-0.23 (0.04)	0.95 (0.01)	-0.10 (0.02)	-0.08 (0.02)
WIND	0.60 (0.02)	0.84 (0.01)	0.86 (0.01)	0.72 (0.02)	0.90 (0.01)	0.89 (0.01)	-0.23 (0.02)	0.34 (0.01)	0.62 (0.01)
CPU	0.95 (0.01)	0.73 (0.01)	0.85 (0.01)	0.97 (0.01)	0.77 (0.01)	0.86 (0.01)	-0.43 (0.01)	-0.71 (0.01)	0.70 (0.01)
2DPLANES	0.33 (0.02)	0.33 (0.02)	0.99 (0.01)	0.58 (0.01)	0.58 (0.01)	0.99 (0.01)	-0.08 (0.07)	-0.04 (0.02)	0.24 (0.05)
POL	0.77 (0.01)	0.77 (0.01)	0.92 (0.01)	0.93 (0.01)	0.90 (0.01)	0.93 (0.01)	-0.41 (0.03)	-0.01 (0.03)	0.21 (0.02)

Table 6. Spearman rank correlations with standard errors in parentheses between different utility function pairs (Accuracy-F1, Accuracy-Negative Log-Loss, and F1-Negative Log-Loss) across multiple datasets and three semivalues: Shapley, $(4,1)$-Beta Shapley, and Banzhaf. Each semivalue is approximated 5 times using Monte Carlo sampling, and Spearman rank correlations are computed for each run. The reported values correspond to the mean correlation across the 5 runs, while the standard errors are derived from the standard deviation of these 5 estimates.

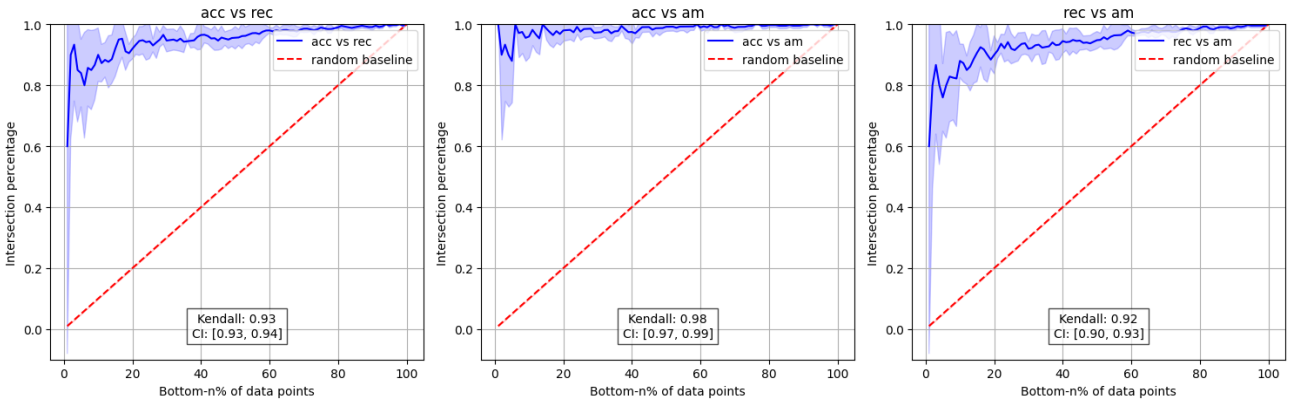


Figure 6. (a) BREAST - Shapley

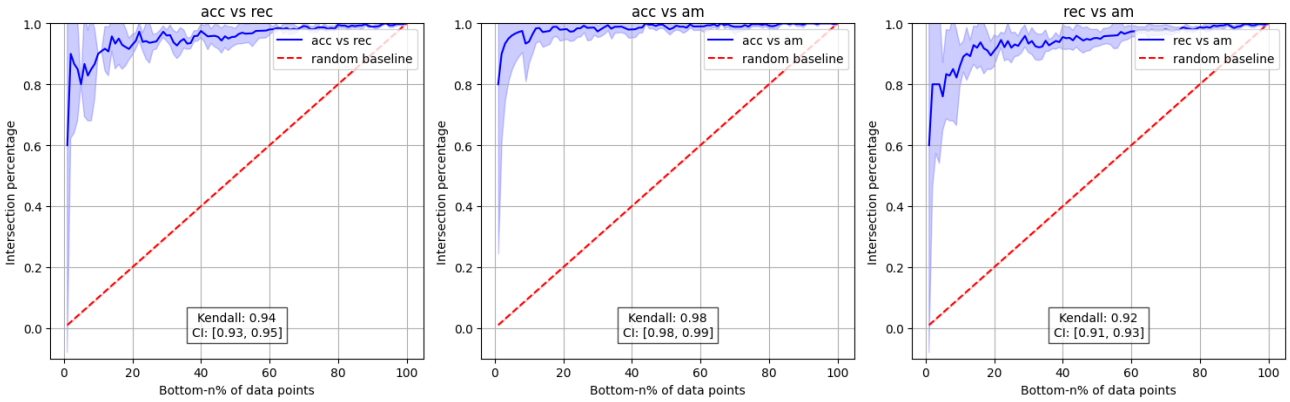


Figure 6. (b) BREAST - (4,1)-Beta Shapley

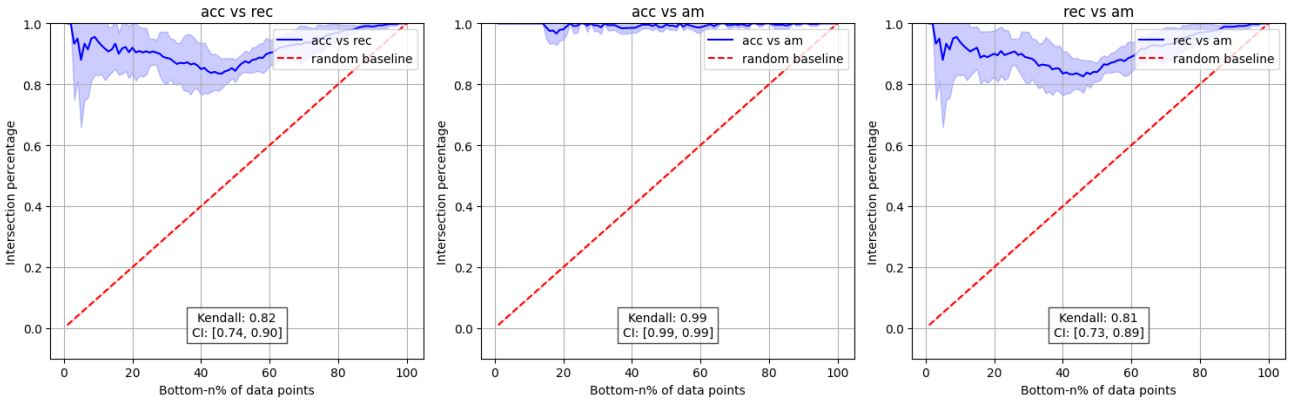


Figure 6. (c) BREST - Banzhaf

Figure 7. Intersection of bottom-$n\%$ ranked data points for the BREAST dataset using different utility functions (accuracy, recall, and arithmetic mean) across semi-value methods. The plot includes a theoretical random baseline, computed using a formula that derives the expected intersection when selecting bottom-$n\%$ points from a random permutation. This provides a reference for evaluating how much the observed intersections deviate from purely random rankings.

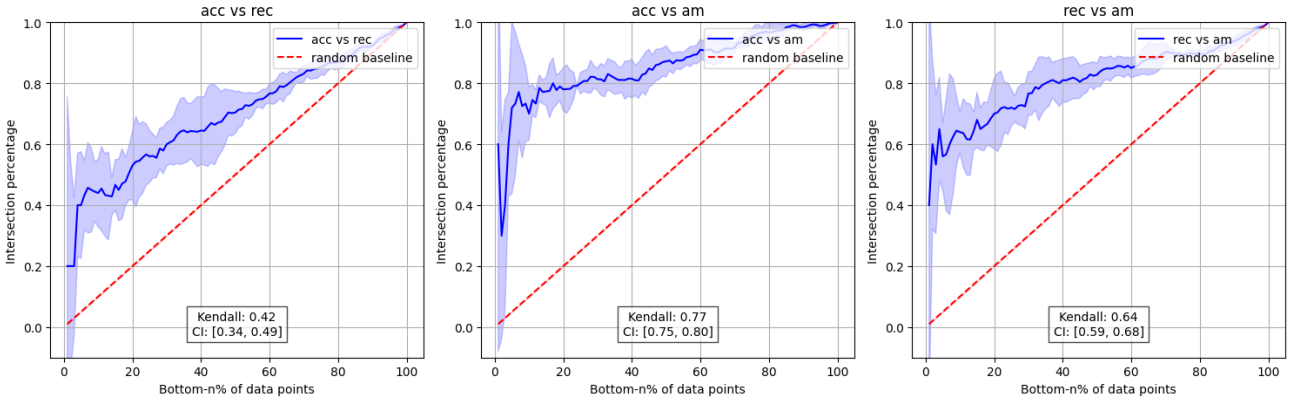


Figure 7. (a) TITANIC- Shapley

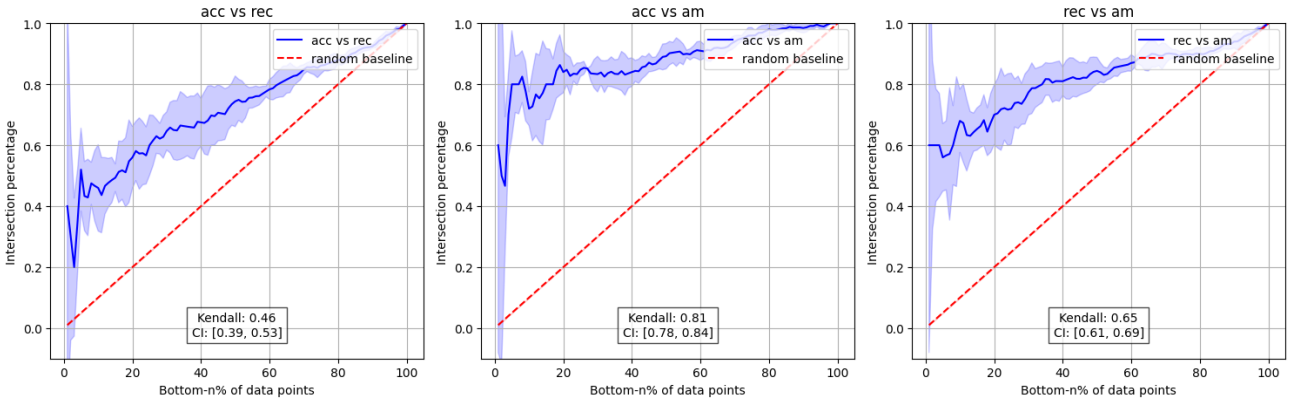


Figure 7. (b) TITANIC- (4,1)-Beta Shapley

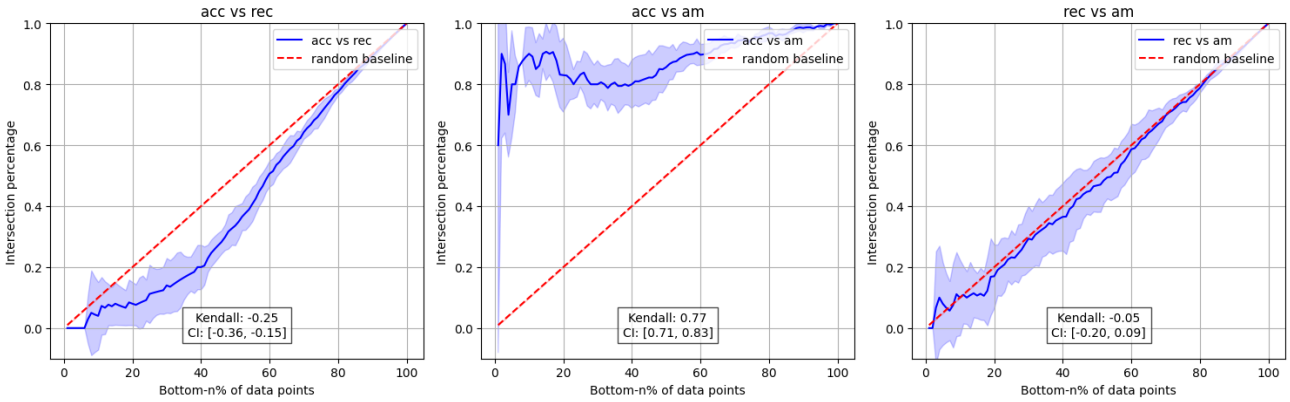


Figure 7. (c) TITANIC - Banzhaf

Figure 8. Intersection of bottom-$n\%$ ranked data points for the TITANIC dataset using different utility functions (accuracy, recall, and arithmetic mean) across semi-value methods. The plot includes a theoretical random baseline, computed using a formula that derives the expected intersection when selecting bottom-$n\%$ points from a random permutation. This provides a reference for evaluating how much the observed intersections deviate from purely random rankings.

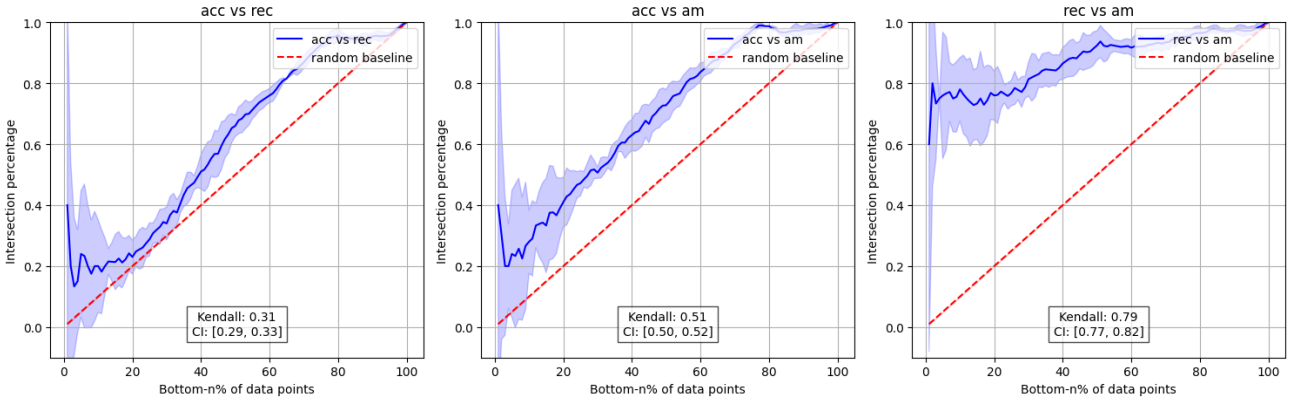


Figure 8. (a) CREDIT - Shapley

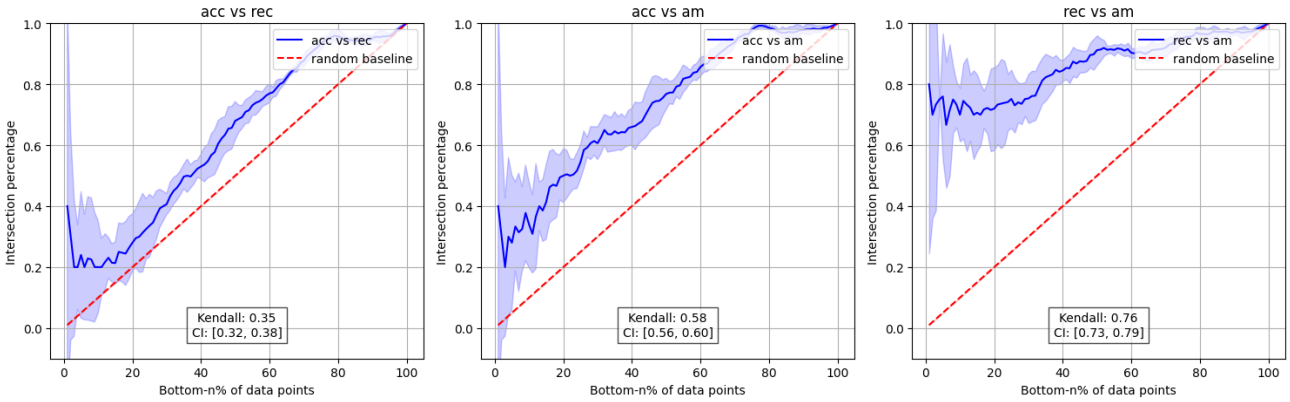


Figure 8. (b) CREDIT - (4,1)-Beta Shapley

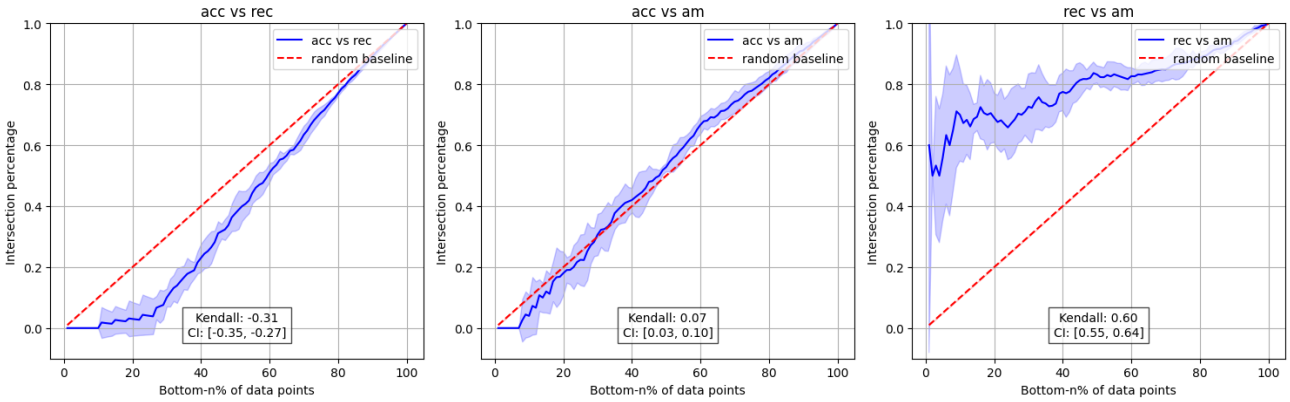


Figure 8. (c) CREDIT - Banzhaf

Figure 9. Intersection of bottom-$n\%$ ranked data points for the CREDIT dataset using different utility functions (accuracy, recall, and arithmetic mean) across semi-value methods. The plot includes a theoretical random baseline, computed using a formula that derives the expected intersection when selecting bottom-$n\%$ points from a random permutation. This provides a reference for evaluating how much the observed intersections deviate from purely random rankings.

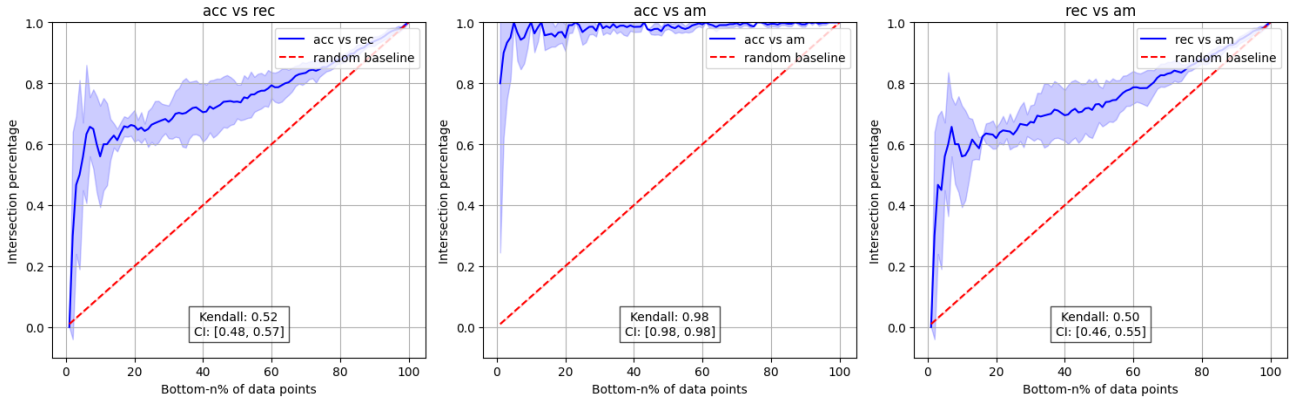


Figure 9. (a) HEART - Shapley

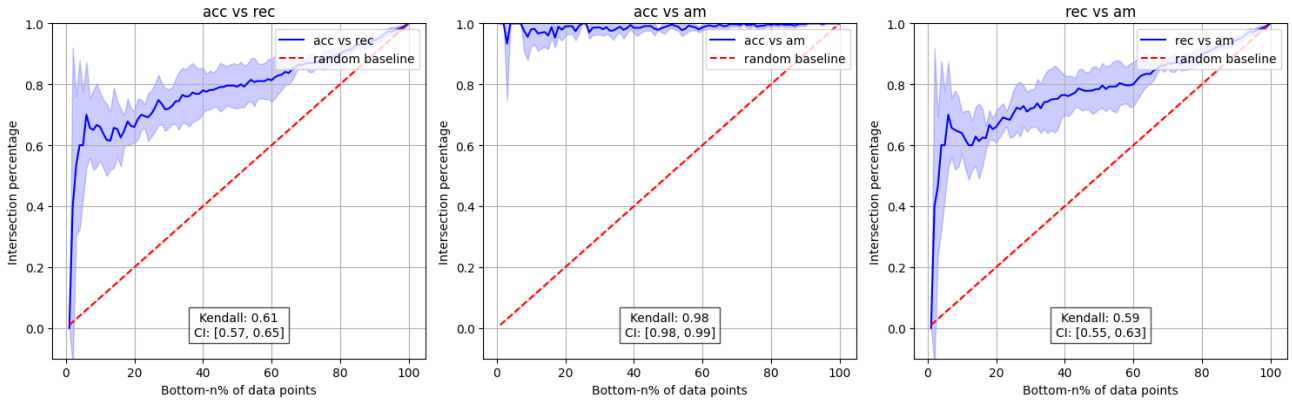


Figure 9. (b) HEART - (4,1)-Beta Shapley

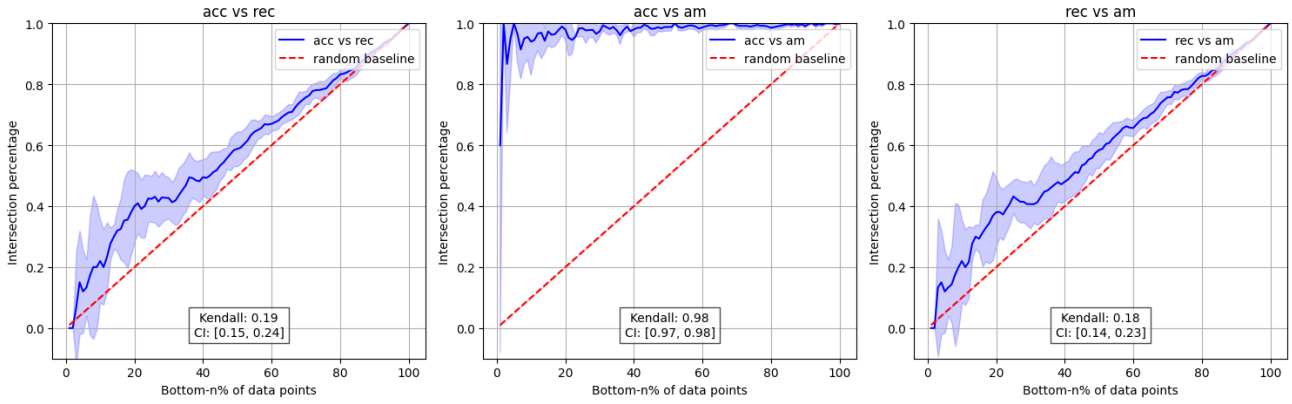


Figure 9. (c) HEART - Banzhaf

Figure 10. Intersection of bottom-$n\%$ ranked data points for the HEART dataset using different utility functions (accuracy, recall, and arithmetic mean) across semi-value methods. The plot includes a theoretical random baseline, computed using a formula that derives the expected intersection when selecting bottom-$n\%$ points from a random permutation. This provides a reference for evaluating how much the observed intersections deviate from purely random rankings.

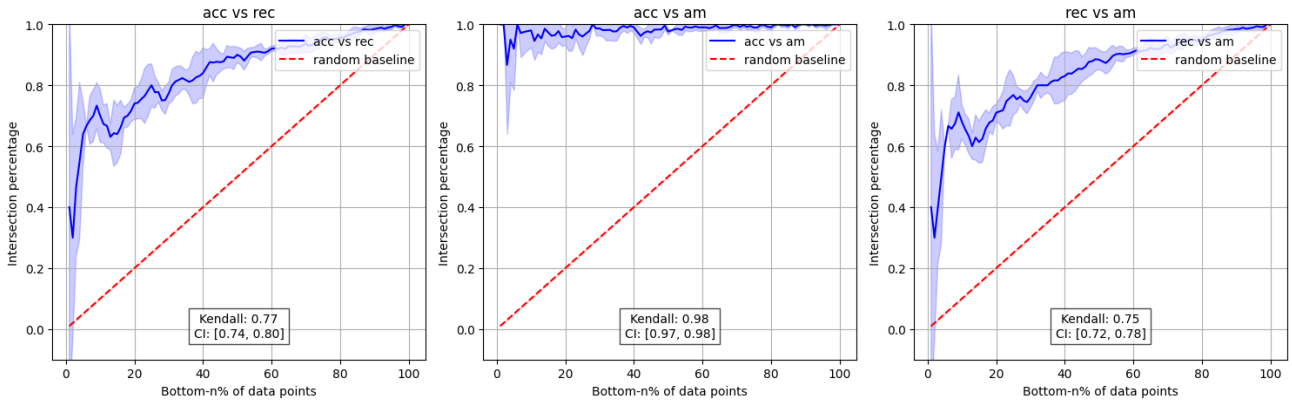


Figure 10. (a) WIND - Shapley

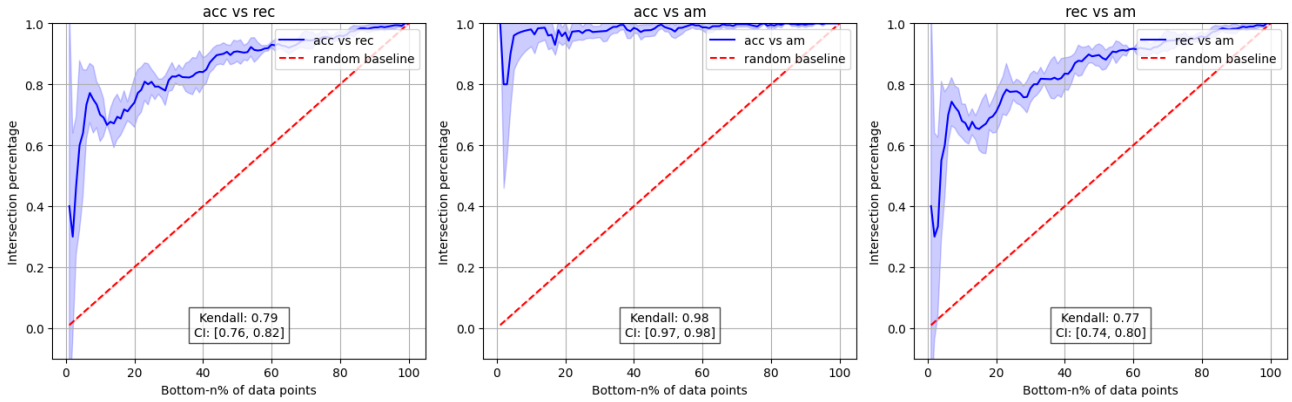


Figure 10. (b) WIND - (4,1)-Beta Shapley

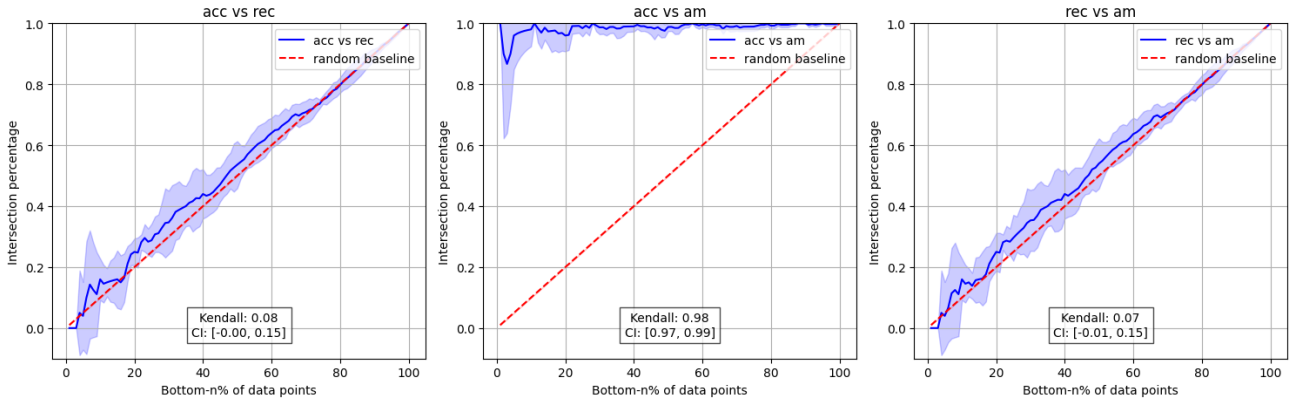


Figure 10. (c) WIND - Banzhaf

Figure 11. Intersection of bottom-$n\%$ ranked data points for the WIND dataset using different utility functions (accuracy, recall, and arithmetic mean) across semi-value methods. The plot includes a theoretical random baseline, computed using a formula that derives the expected intersection when selecting bottom-$n\%$ points from a random permutation. This provides a reference for evaluating how much the observed intersections deviate from purely random rankings.

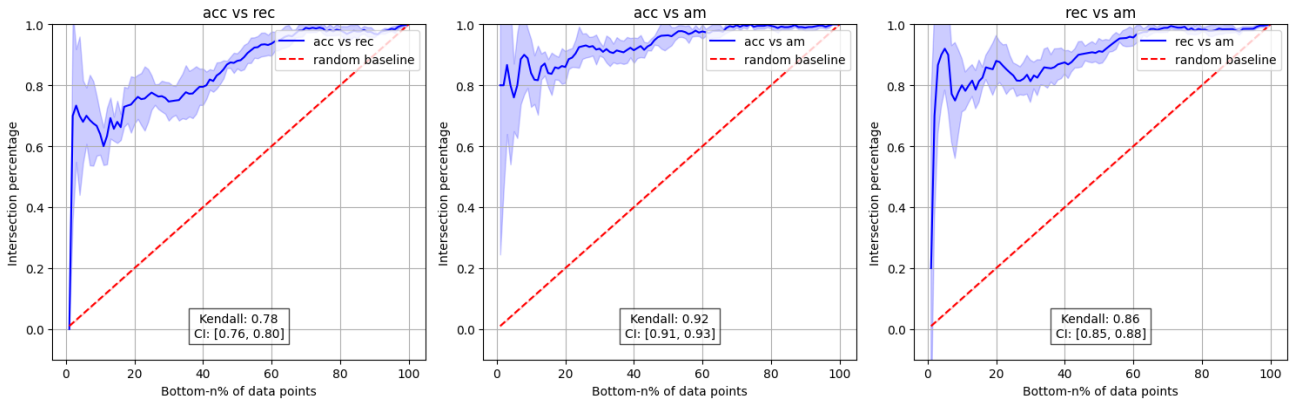


Figure 11. (a) CPU - Shapley

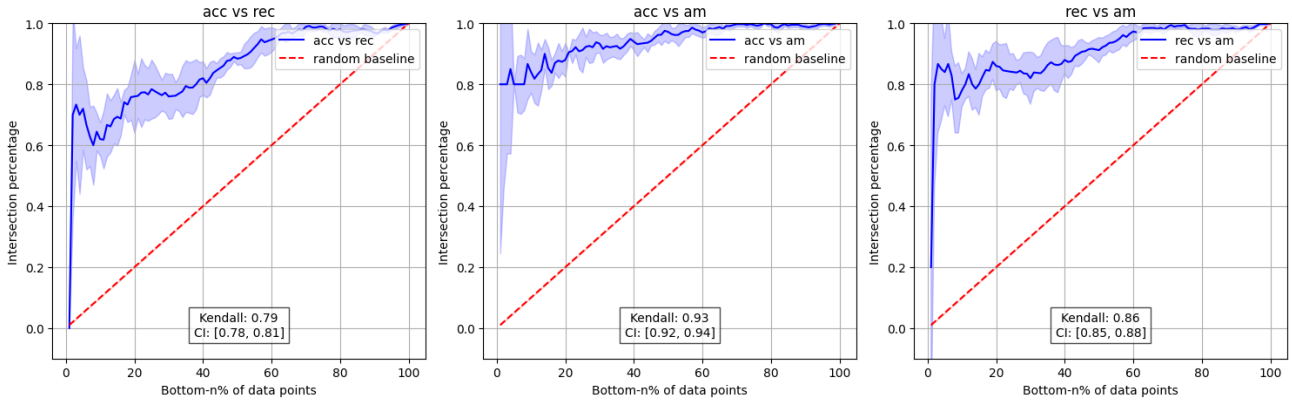


Figure 11. (b) CPU- (4,1)-Beta Shapley

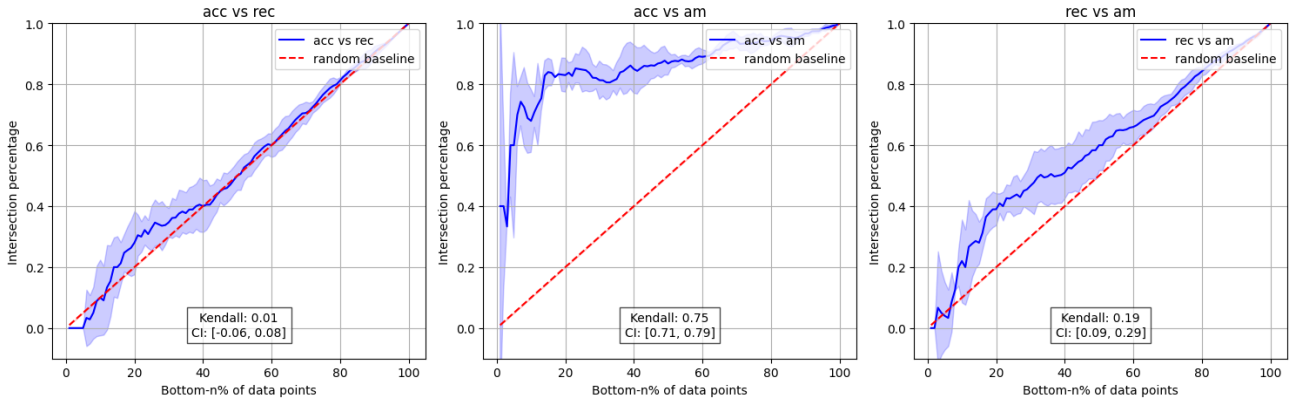


Figure 11. (c) CPU - Banzhaf

Figure 12. Intersection of bottom-$n\%$ ranked data points for the CPU dataset using different utility functions (accuracy, recall, and arithmetic mean) across semi-value methods. The plot includes a theoretical random baseline, computed using a formula that derives the expected intersection when selecting bottom-$n\%$ points from a random permutation. This provides a reference for evaluating how much the observed intersections deviate from purely random rankings.

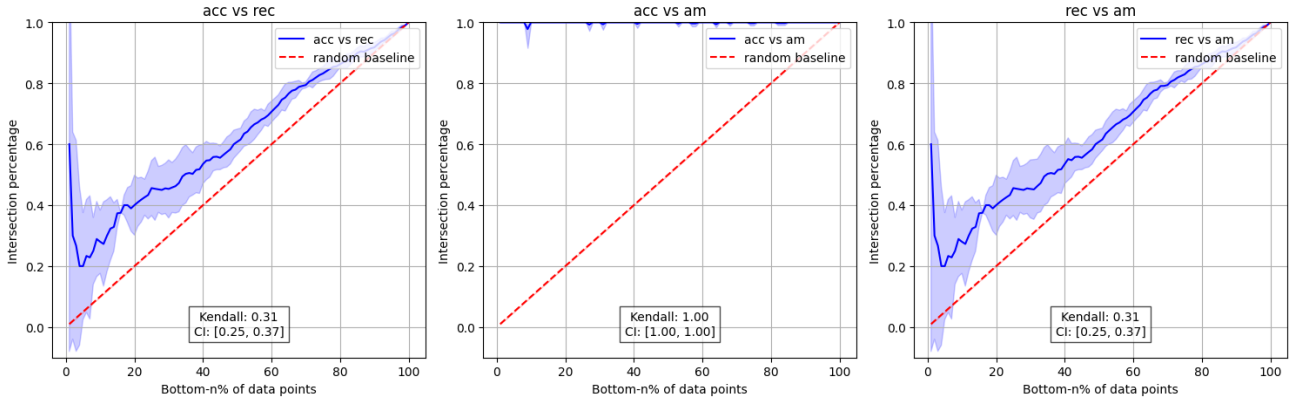


Figure 12. (a) 2DPLANES - Shapley

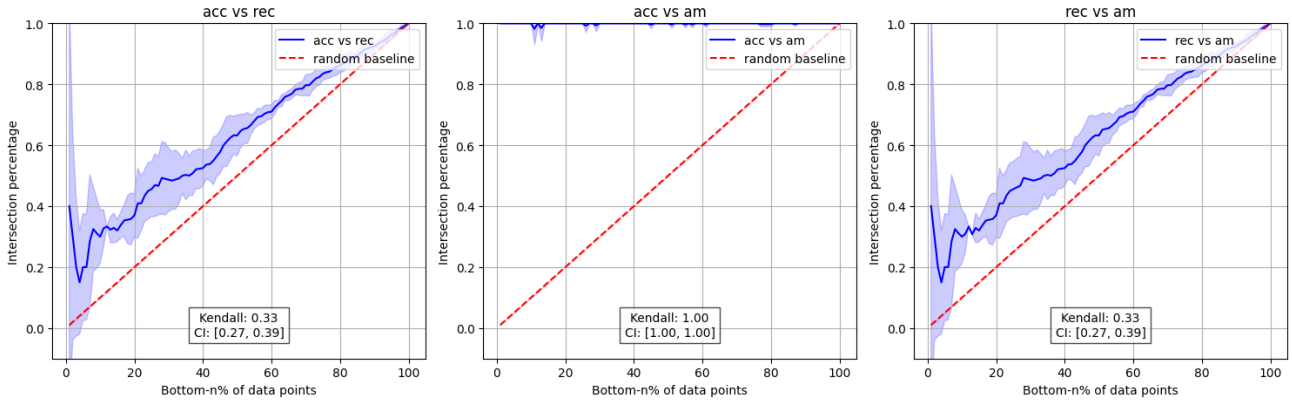


Figure 12. (b) 2DPLANES - (4,1)-Beta Shapley

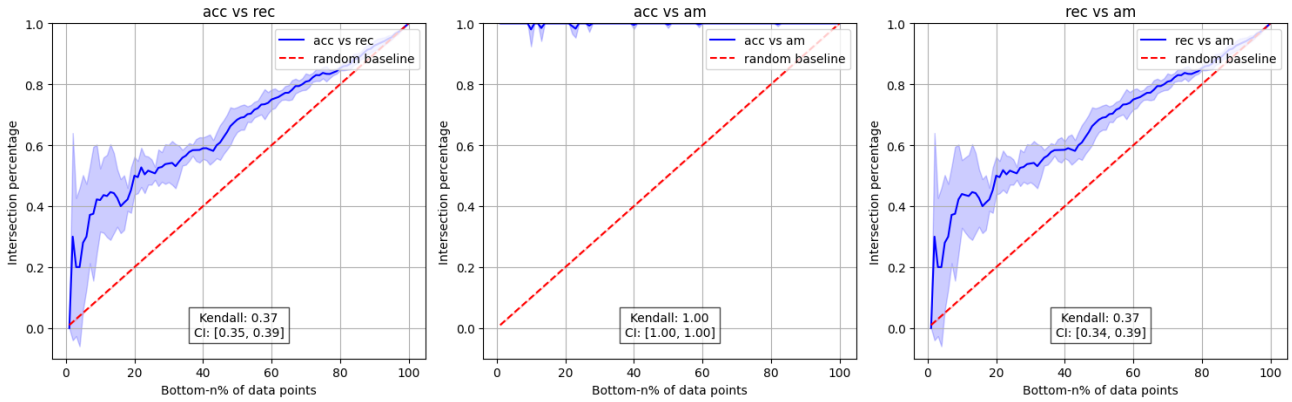


Figure 12. (c) 2DPLANES - Banzhaf

Figure 13. Intersection of bottom-$n\%$ ranked data points for the 2DPLANES dataset using different utility functions (accuracy, recall, and arithmetic mean) across semi-value methods. The plot includes a theoretical random baseline, computed using a formula that derives the expected intersection when selecting bottom-$n\%$ points from a random permutation. This provides a reference for evaluating how much the observed intersections deviate from purely random rankings.

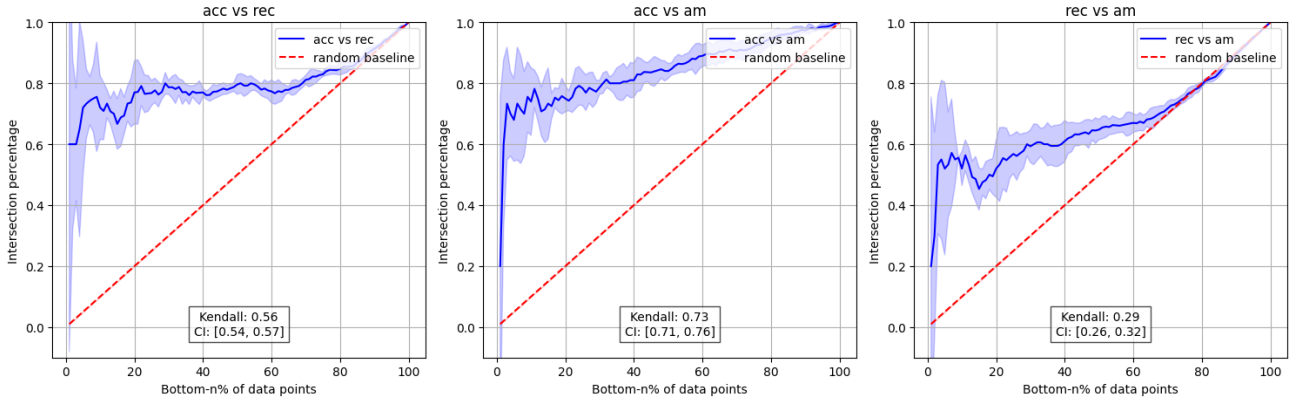


Figure 13. (a) POL - Shapley

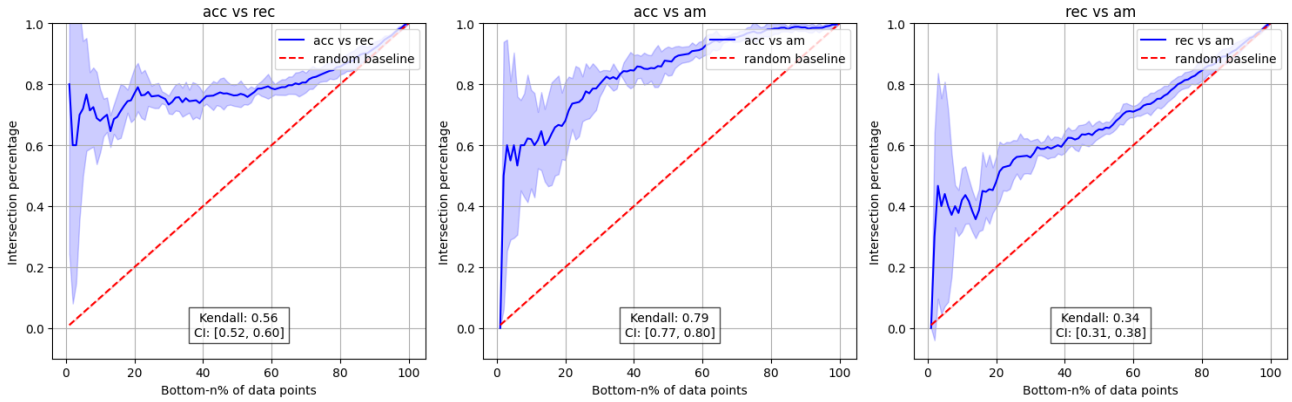


Figure 13. (b) POL - (4,1)-Beta Shapley

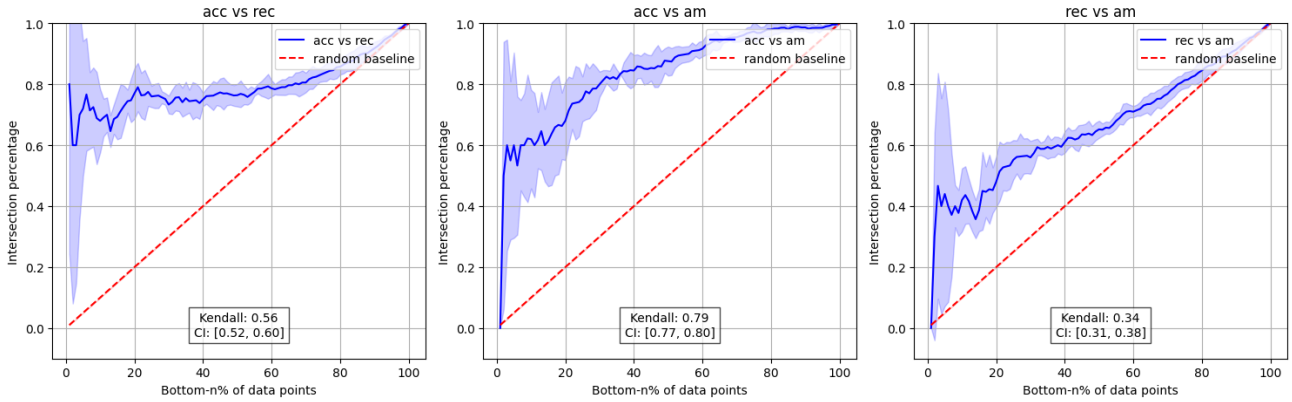
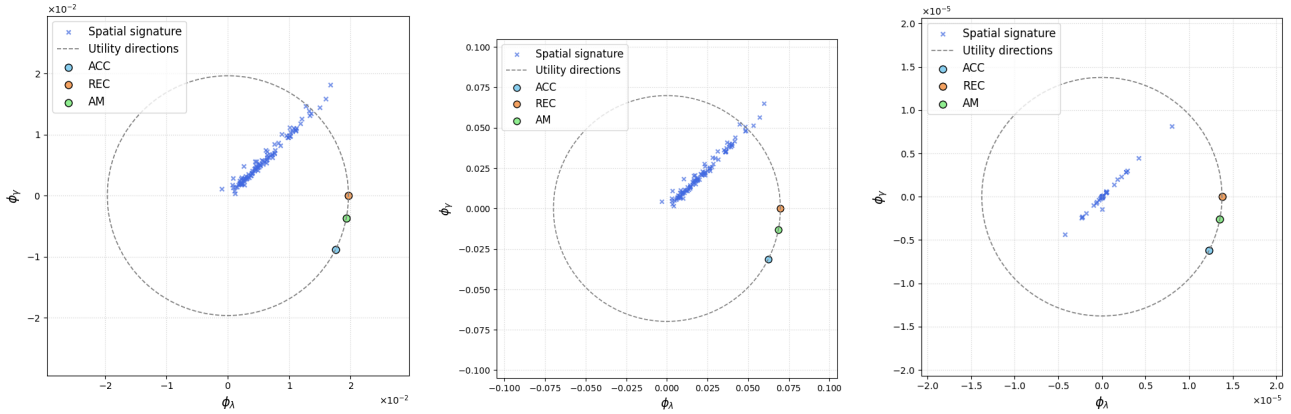


Figure 13. (c) POL - Banzhaf

Figure 14. Intersection of bottom-$n\%$ ranked data points for the POL dataset using different utility functions (accuracy, recall, and arithmetic mean) across semi-value methods. The plot includes a theoretical random baseline, computed using a formula that derives the expected intersection when selecting bottom-$n\%$ points from a random permutation. This provides a reference for evaluating how much the observed intersections deviate from purely random rankings.

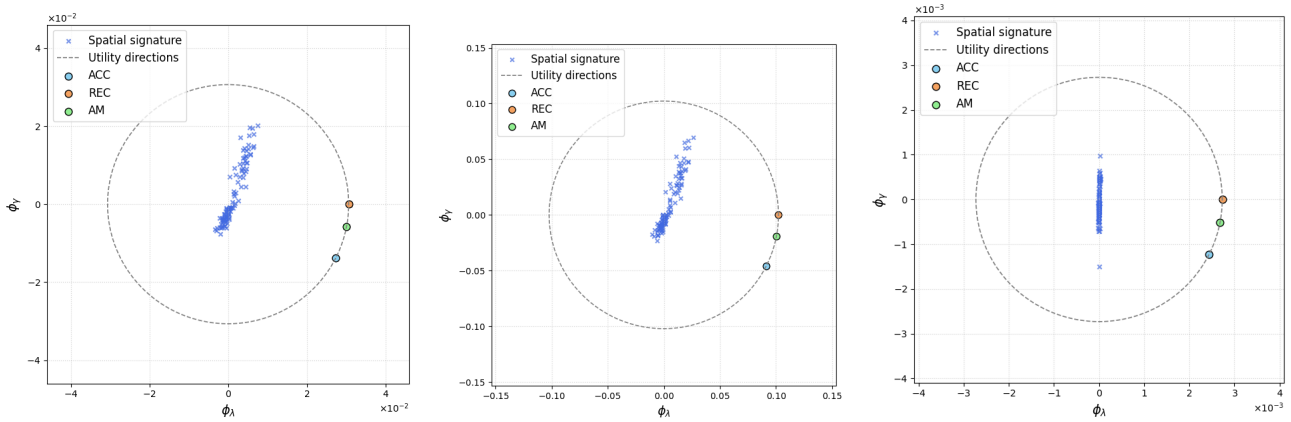
C.6. Additional figures for Section 4

This section contains all figures illustrating the spatial signatures across different datasets and semivalues.



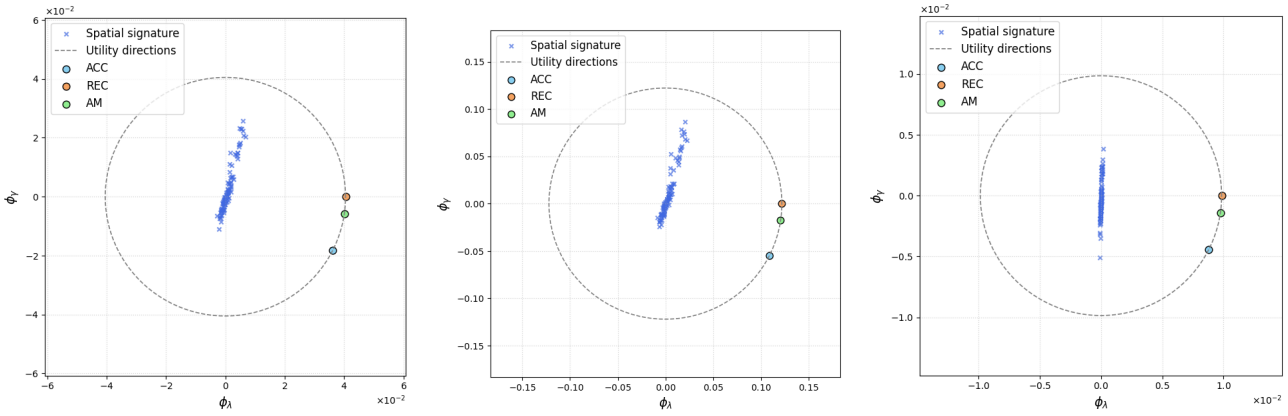
(a) Spatial signature under ω_{shap} (b) Spatial signature under ω_{beta} (c) Spatial signature under ω_{banz}

Figure 15. These three figures illustrate the spatial signatures of the BREAST dataset under three semivalues: $\omega_{\text{shap}}, \omega_{\text{beta}}, \omega_{\text{banz}}$. The unit circle represents the set of distinct utilities in terms of ranking, with markers indicating the normalized utility vectors $\tilde{\mathbf{u}}_{\text{acc}}$, $\tilde{\mathbf{u}}_{\text{rec}}$, and $\tilde{\mathbf{u}}_{\text{am}}$ for the three utility functions used in our experiments, namely the accuracy, the recall, and the arithmetic mean that belong to $\mathcal{U}_{\lambda,\gamma}$.



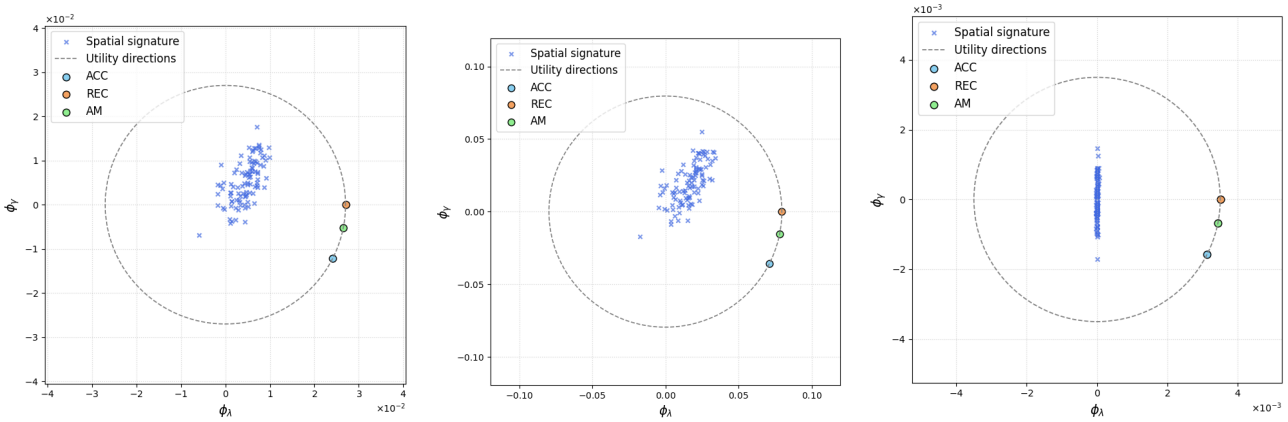
(a) Spatial signature under ω_{shap} (b) Spatial signature under ω_{beta} (c) Spatial signature under ω_{banz}

Figure 16. These three figures illustrate the spatial signatures of the TITANIC dataset under three semivalues: $\omega_{\text{shap}}, \omega_{\text{beta}}, \omega_{\text{banz}}$. The unit circle represents the set of distinct utilities in terms of ranking, with markers indicating the normalized utility vectors $\tilde{\mathbf{u}}_{\text{acc}}$, $\tilde{\mathbf{u}}_{\text{rec}}$, and $\tilde{\mathbf{u}}_{\text{am}}$ for the three utility functions used in our experiments, namely the accuracy, the recall, and the arithmetic mean that belong to $\mathcal{U}_{\lambda,\gamma}$.



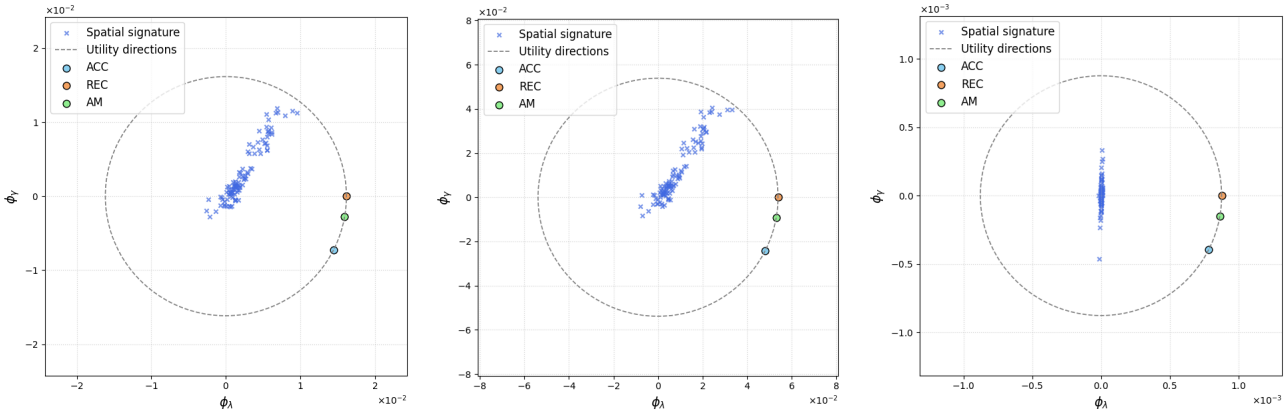
(a) Spatial signature under ω_{shap} (b) Spatial signature under ω_{beta} (c) Spatial signature under ω_{banz}

Figure 17. These three figures illustrate the spatial signatures of the CREDIT dataset under three semivalues: $\omega_{\text{shap}}, \omega_{\text{beta}}, \omega_{\text{banz}}$. The unit circle represents the set of distinct utilities in terms of ranking, with markers indicating the normalized utility vectors $\tilde{\mathbf{u}}_{\text{acc}}$, $\tilde{\mathbf{u}}_{\text{rec}}$, and $\tilde{\mathbf{u}}_{\text{am}}$ for the three utility functions used in our experiments, namely the accuracy, the recall, and the arithmetic mean that belong to $\mathcal{U}_{\lambda,\gamma}$.



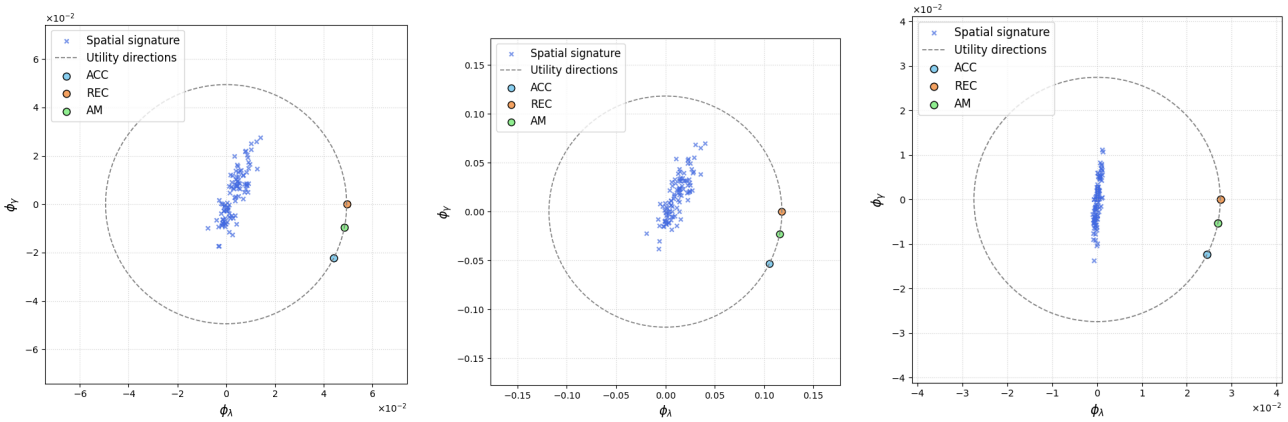
(a) Spatial signature under ω_{shap} (b) Spatial signature under ω_{beta} (c) Spatial signature under ω_{banz}

Figure 18. These three figures illustrate the spatial signatures of the HEART dataset under three semivalues: $\omega_{\text{shap}}, \omega_{\text{beta}}, \omega_{\text{banz}}$. The unit circle represents the set of distinct utilities in terms of ranking, with markers indicating the normalized utility vectors $\tilde{\mathbf{u}}_{\text{acc}}$, $\tilde{\mathbf{u}}_{\text{rec}}$, and $\tilde{\mathbf{u}}_{\text{am}}$ for the three utility functions used in our experiments, namely the accuracy, the recall, and the arithmetic mean that belong to $\mathcal{U}_{\lambda,\gamma}$.



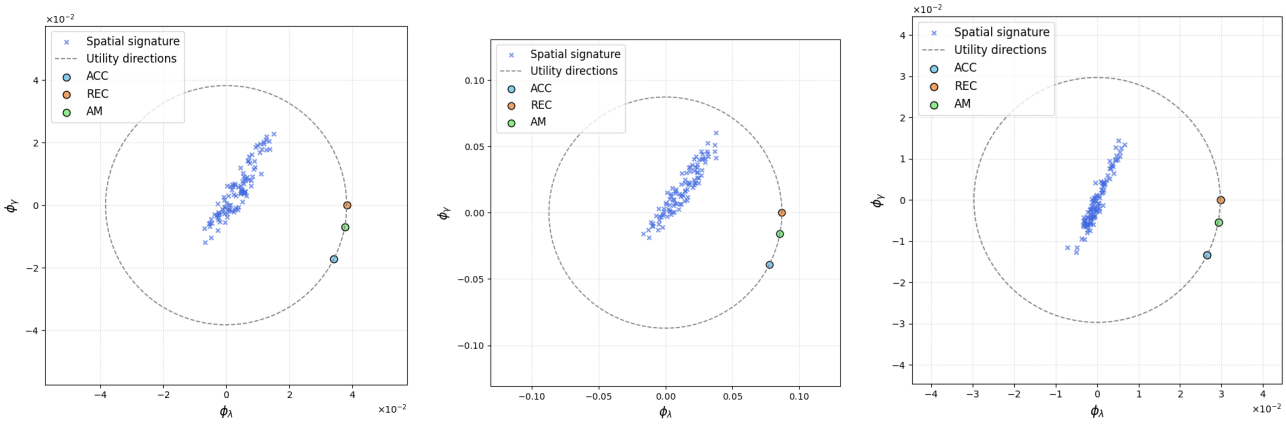
(a) Spatial signature under ω_{shap} (b) Spatial signature under ω_{beta} (c) Spatial signature under ω_{banz}

Figure 19. These three figures illustrate the spatial signatures of the CPU dataset under three semivalues: $\omega_{\text{shap}}, \omega_{\text{beta}}, \omega_{\text{banz}}$. The unit circle represents the set of distinct utilities in terms of ranking, with markers indicating the normalized utility vectors $\tilde{\mathbf{u}}_{\text{acc}}$, $\tilde{\mathbf{u}}_{\text{rec}}$, and $\tilde{\mathbf{u}}_{\text{am}}$ for the three utility functions used in our experiments, namely the accuracy, the recall, and the arithmetic mean that belong to $\mathcal{U}_{\lambda,\gamma}$.



(a) Spatial signature under ω_{shap} (b) Spatial signature under ω_{beta} (c) Spatial signature under ω_{banz}

Figure 20. These three figures illustrate the spatial signatures of the 2DPLANES dataset under three semivalues: $\omega_{\text{shap}}, \omega_{\text{beta}}, \omega_{\text{banz}}$. The unit circle represents the set of distinct utilities in terms of ranking, with markers indicating the normalized utility vectors $\tilde{\mathbf{u}}_{\text{acc}}$, $\tilde{\mathbf{u}}_{\text{rec}}$, and $\tilde{\mathbf{u}}_{\text{am}}$ for the three utility functions used in our experiments, namely the accuracy, the recall, and the arithmetic mean that belong to $\mathcal{U}_{\lambda,\gamma}$.



(a) Spatial signature under ω_{shap} (b) Spatial signature under ω_{beta} (c) Spatial signature under ω_{banz}

Figure 21. These three figures illustrate the spatial signatures of the POL dataset under three semivalues: $\omega_{\text{shap}}, \omega_{\text{beta}}, \omega_{\text{banz}}$. The unit circle represents the set of distinct utilities in terms of ranking, with markers indicating the normalized utility vectors $\tilde{\mathbf{u}}_{\text{acc}}$, $\tilde{\mathbf{u}}_{\text{rec}}$, and $\tilde{\mathbf{u}}_{\text{am}}$ for the three utility functions used in our experiments, namely the accuracy, the recall, and the arithmetic mean that belong to $\mathcal{U}_{\lambda,\gamma}$.