Evaluation Metrics for Data Valuation Methods

Zhiyu Chen and Jicong Fan*

School of Data Science
The Chinese University of Hong Kong, Shenzhen, China
zhiyuchen2@link.cuhk.edu.cn fanjicong@cuhk.edu.cn

Abstract. Data Valuation has emerged as a critical research topic at the intersection of artificial intelligence and economics, enabling principled quantification of individual data importance. Despite many proposed algorithms, a central challenge remains: how to evaluate their quality, reliability, and fairness. This paper presents the first systematic review of evaluation metrics for data valuation, synthesising perspectives from data science and economics. We categorise existing metrics into four data science-oriented classes and complement them with economic principles and practical guidance for metric selection. This cross-disciplinary framework promotes more rigorous and standardised evaluation protocols in future data valuation research.

Keywords: Data Valuation · Evaluation Metrics · Economics · Data Science

1 Introduction

The global explosion of data that began in the early 2000s has ushered in another wave of the Industrial Revolution. With data as the blood and algorithms as the skeleton, artificial intelligence (AI) has been developed to mimic human behaviour and assist in nearly every aspect of daily life [20]. After decades of rapid advancement, AI can now not only imitate humans but also far exceed them in both speed and accuracy. This progress can be attributed to several factors, such as the invention of more powerful chips, the expansion of memory capacity, and the continual refinement of algorithms. Yet, one of the most crucial drivers is AI's unprecedented access to vast and diverse datasets. As in human learning, the capacity of AI systems to generalize and improve is deeply tied to the quantity and quality of data from which they learn.

However, simply increasing data volume does not guarantee better AI performance. Redundant, low-quality, or incomplete data may hinder training efficiency or degrade model generalization [12]. Mislabelled, contaminated, or erroneous samples [8, 16], in particular, can cause negative effects. Manual data curation remains labour-intensive, underscoring the need for principled strategies to quantify the contribution of each datum. Since 2010, a growing body of research on data valuation has emerged [10]. These studies propose algorithms that assign a numerical value to each data instance, quantifying its relative importance or contribution to the model's performance [10, 13, 19].

Despite the diversity of proposed valuation methods, a fundamental challenge remains: how to evaluate and compare the effectiveness of these data valuation methods.

* Corresponding author

In contrast to model evaluation, where metrics such as accuracy, RMSE, AUROC, or ARI are well established, there is still no consensus on what constitutes a "good" data valuation method. Existing works employ a diverse range of ad-hoc evaluation criteria, such as data removal plots [20] and correlation with ground-truth Shapley [25]. However, these metrics provide only partial insights and may yield conflicting conclusions across different datasets or training paradigms. As highlighted by Jiang et al.(2023), the lack of standardized, comprehensive evaluation metrics impedes methodological progress, making it difficult to determine which valuation methods are reliable [26].

This paper addresses this gap by offering the first unified review of evaluation metrics for data valuation methods from a cross-domain perspective. While most surveys focus either on the machine learning viewpoint [9, 10, 40] or on economic theories of data pricing [15, 35, 36], our approach integrates both domains. Jiang et al. (2023) introduced *OpenDataVal*, a unified benchmark for data valuation methods [26], which includes a number of experimental evaluations. However, their discussion of evaluation metrics remains limited in scope and is confined to the data science perspective. Unlike Bendechache et al. (2023) which synthesizes what "data value" is via models, dimensions, and applications [3], our paper centres on how to evaluate data-valuation methods, providing a systematic, cross-domain taxonomy of evaluation metrics (performance, consistency, interpretability, efficiency, and economic rationality/cost/societal impact) and concrete procedures for benchmarking their validity and fairness.

Our main contributions are summarized as follows:

- We present a unified review of data valuation methods rooted in both data-science and economic theories, clarifying their motivations and theoretical underpinnings.
- We propose a comprehensive taxonomy of evaluation metrics that integrates both data-science and economic perspectives.
- We provide a practical guidance framework that assists practitioners in applying and interpreting evaluation metrics, emphasizing context-specific caveats.

The remainder of the paper is structured as follows. Section 2 reviews core data valuation approaches from both perspectives. Section 3 motivates the need for principled evaluation. Section 4 presents the taxonomy of data-science-oriented evaluation metrics. Section 5 discusses complementary economic evaluation criteria. Section 6 provides practical guidance for applying evaluation metrics. Section 7 concludes with a synthesis of cross-domain insights.

2 Overview of Data Valuation Methods

Data valuation is an inherently cross-domain problem studied from both economic and data-scientific perspectives. The former views data as an intangible but tradable asset, focusing on market mechanisms and pricing fairness; the latter treats it as a utility contributor, quantifying each sample's influence on model performance. We outline both perspectives and their intersection below.

In this context, the notion of "data" adopted throughout this paper is not restricted to tabular feature vectors. A datum may represent a numerical vector, an image, a video or audio sequence, a text document or prompt, a node within a graph, or even an entire graph [11]. Because our evaluation framework is formulated in terms of marginal utility

contribution with respect to a learning objective, its applicability depends only on the existence of a well-defined learning algorithm and measurable utility function, rather than on the structural form of the data itself.

2.1 Data Science Perspectives: Influence and Utility Contribution of Data

From the data science perspective, data valuation focuses on quantifying the influence or utility of each training example with respect to a given model and learning objective. This line of research originates in statistical diagnostics and has since evolved into general frameworks for attributing model performance to individual samples.

Classical Regression Diagnostics Early work in linear regression introduced quantitative measures of how individual observations influence model estimates. Given a dataset $D = \{(x_i, y_i)\}_{i=1}^n$, let $X \in \mathbb{R}^{n \times p}$ denote the design matrix whose rows are feature vectors x_i^T, and $y \in \mathbb{R}^n$ the response vector. The ordinary least squares (OLS) estimator is $\hat{\beta} = (X^T X)^{-1} X^T y$ and $\hat{\beta}_{(-i)}$ denotes the estimated coefficients after removing the i^{th} data point.

- **Cook's Distance** [5]: It measures how much the entire fitted model changes if observation i is removed.

$$D_i = \frac{(\hat{\beta}_{(-i)} - \hat{\beta})^\top \left(X^\top X\right) (\hat{\beta}_{(-i)} - \hat{\beta})}{p\hat{\sigma}^2} \tag{1}$$

  where p is the number of predictors and $\hat{\sigma}^2$ is the residual variance.

- **DFBETAS** [2]: It assesses how each coefficient β_j shifts when datum i is removed.

$$\text{DFBETAS}_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_{j(-i)}}{s_{(-i)}\sqrt{c_{jj}}} \tag{2}$$

  where c_{jj} is the j^{th} diagonal element of $(X^T X)^{-1}$ and $s_{(-i)}$ is the residual standard error excluding observation i.

Both diagnostics serve as early quantitative data valuation scores, identifying outliers and influential points that disproportionately affect model performance.

Influence Functions in Modern Machine Learning In large-scale learning problems, retraining after removing each point is infeasible. Influence functions generalize these classical ideas of Leave-One-Out [6] by approximating the effect of small perturbations on the loss function through first-order analysis [28].

Given a loss function $\ell(z, \theta)$ defined on each training example $z_i = (x_i, y_i)$, the empirical risk is $L(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(z_i, \theta)$. The optimal model parameters $\hat{\theta}$ are obtained by minimising this risk, satisfying the first order condition $\nabla_\theta L(\hat{\theta}) = 0$. When the i^{th} sample is up-weighted by a small ε, the new optimum $\hat{\theta}_{\varepsilon,i}$ satisfies:

$$\left.\frac{d\hat{\theta}_{\varepsilon,i}}{d\varepsilon}\right|_{\varepsilon=0} = -H_{\hat{\theta}}^{-1} \nabla_\theta \ell\left(z_i, \hat{\theta}\right) \tag{3}$$

where $H_{\hat{\theta}} = \nabla_\theta^2 L(\hat{\theta})$ is the Hessian at the optimum. The influence of training point z_i on the test loss $\ell(z_{\text{test}}, \hat{\theta})$ is then approximated by:

$$I_{\text{up,loss}}(z_i, z_{\text{test}}) = -\nabla_\theta \ell\left(z_{\text{test}}, \hat{\theta}\right)^\top H_{\hat{\theta}}^{-1} \nabla_\theta \ell\left(z_i, \hat{\theta}\right) \tag{4}$$

This expression quantifies how much removing or perturbing z_i changes the model's loss on a given test example, allowing practitioners to trace predictions back to the most influential training data points without exhaustive retraining.

Although influence-based methods provide valuable insights into individual data effects, they generally assume independence among samples and fail to capture higher-order interactions. This limitation has motivated the adoption of cooperative game-theoretic frameworks, such as the Shapley Value.

2.2 Economic Perspectives: Pricing and Market Design for Data

From the economic perspective, data valuation is grounded in market design theory and information economics. The goal is to design pricing and allocation schemes that are fair, incentive-compatible, and arbitrage-free, taking into account the strategic behaviour of data owners, intermediaries, and buyers.

A foundational line of research emphasizing the interests of data sellers is the **Query-based Data Pricing framework** [29]. In this model, data is conceptualised as a collection of database queries Q. A pricing function $p(Q)$ assigns a monetary value to each query such that arbitrage-freeness holds:

$$p(Q_1 \cup Q_2) \leq p(Q_1) + p(Q_2) \tag{5}$$

This prevents a buyer from combining cheaper queries Q_1, Q_2 to reconstruct a more expensive query and pay less. The model emphasises the informativeness of queries and the assumption of rational buyers/sellers in a data market.

Focusing on the interests of data contributors, Li et al. (2014) proposed a **Privacy-based Data Pricing Model** [34]. In this formulation, a market-maker sells query results perturbed to satisfy differential privacy and compensates data owners for the resulting privacy loss. The key idea links price to accuracy: buyers pay more for more accurate results, while sellers receive compensation proportional to their data disclosure. This approach extends the arbitrage-free pricing model to privacy-aware markets, establishing a quantitative relationship between economic value and information leakage.

Integrating the role of intermediaries and market institutions, Hao (2023) conceptualised data pricing as part of a broader **digital-market ecosystem** [21]. This work analyses how different organizational structures, including centralised data brokers, platform-mediated exchanges, and federated markets, affect pricing formation and data ownership. It highlights the complex interactions among data suppliers, intermediaries and consumers, and explains how market power, network effects and regulatory interventions shape the economic value of data in practice.

It is worth noting that methods proposed from the economic perspective seldom provide explicit mechanisms for valuing individual data points. Instead, they focus on the composition, exchange and redistribution of data value within the market ecosystem.

These models provide macro-level insights into transaction incentives and equilibria, complementing micro-level utility-based approaches.

2.3 Where the Two Meet: Data Shapley and Its Accelerations

Data Shapley applies the Shapley Value from cooperative game theory to supervised learning, providing a principled way to quantify each sample's marginal contribution to model performance [20].

Given a dataset $D = \{(x_i, y_i)\}_{i=1}^n$, a learning algorithm $\mathcal{A}$ and a utility function $V(S)$, the Shapley value for point i is defined as:

$$\phi_i = C \sum_{S \subseteq D \backslash \{i\}} \frac{V(S \cup \{i\}) - V(S)}{\binom{n-1}{|S|}} \tag{6}$$

where $S \subseteq D\backslash\{i\}$ represents a subset of data excluding point i, and C is a scaling constant that does not affect the ranking of values. This formulation satisfies the null player, symmetry, and linearity axioms, ensuring a fair and interpretable allocation of value across data instances.

Despite its theoretical appeal, exact computation of ϕ_i is exponential in n, requiring evaluation over all 2^n subsets. To mitigate this, subsequent research introduces various approximation and acceleration strategies. The Truncated Monte Carlo(TMC) Shapley methods estimate the expected marginal contributions by sampling random data permutations [20]. Jia et al. (2019) [25] proposed a suite of algorithms for efficient Shapley value estimation. For instance, their Group-Testing approach reduces the exponential computation cost to $O(\sqrt{n}\log^2 n)$. Beta Shapley [31] further introduces cardinality-dependent coalition weights, providing a noise-robust generalisation that improves mislabelled-data detection and subsampling performance.

Domain-specific extensions further adapt the Data Shapley valuation to distributed and large-scale learning. Federated Shapley [43] aligns valuation with federated training protocols, capturing client participation effects while minimising communication overheads. Data Shapley in One Run [42] integrates value estimation into a single training process, avoiding repeated retraining and improving scalability for deep models.

In summary, Data Shapley bridges economic fairness and computational efficiency. It inherits equitable principles from cooperative game theory while incorporating machine learning optimisation for practical feasibility.

3 Why Evaluate Data Valuation Methods?

Despite advances in data valuation, a key question remains: how to evaluate the reliability of these methods. Unlike model assessment with established metrics such as accuracy or AUROC, there is no consensus on measuring the quality of a valuation algorithm. The difficulty arises from its cross-domain nature: economics stresses fairness and incentive compatibility, whereas data science emphasises utility and robustness.

From the data-science view, evaluation verifies whether estimated values truly reflect each sample's contribution to model performance. Effective methods should reward informative samples and penalise noisy or mislabelled ones. Empirical criteria

include data removal or addition plots and correlation analyses comparing estimated and reference values. By contrast, economic metrics assess whether a valuation mechanism adheres to key theoretical principles, with a focus on conceptual soundness and incentive alignment rather than numerical performance indicators.

The absence of standardised evaluation protocols hinders comparability and reproducibility, as studies often rely on ad-hoc datasets and metrics [26]. Establishing unified evaluation criteria is therefore essential to link theoretical fairness with empirical validity and to enable systematic progress in data valuation research.

4 Evaluation Metrics in the View of Data Science

4.1 Taxonomy

From the data-science viewpoint, evaluating a data-valuation method requires quantitative metrics that measure how well an algorithm reflects the true contribution of each data instance to model performance. Because the goals of data valuation vary, no single metric can fully characterise quality across all dimensions.

To address this, we categorise existing evaluation metrics into four groups according to their primary evaluation objective:

- Performance-based,
- Consistency-based,
- Interpretability-based,
- Efficiency-based.

This taxonomy provides a unified lens for comparing diverse evaluation strategies used in recent literature. A summary of the representative evaluation method corresponding to each category is presented in **Table 1**, which serves as a roadmap for the detailed discussions in the following subsections.

To support the evaluation framework proposed in this review, we provide an open-source repository: https://github.com/Zhiyu2723/Evaluation_Metrics_DV.git. It includes modular Python implementations of the seven data-science metrics (A1-A7). The economic criteria (B1-B3), which rely mainly on conceptual assessment, are not included.

4.2 Performance-Based Metrics

Performance-based metrics provide the most direct and widely adopted approach to evaluating data valuation methods. They examine whether the assigned values accurately reflect each data point's marginal contribution to model performance. In other words, they assess whether data points deemed "valuable" truly enhance predictive accuracy, robustness, or generalisation when included. These metrics align closely with the core objective of machine learning: leveraging the most informative data to maximise predictive effectiveness.

Perspective	Metric	Brief Description
Data Science Performance-Based	Data removal/addition plots	Model performance change versus proportion of samples removed or added by value ranking.
	Noisy label detection	Low-value samples correspond to mislabelled data.
Data Science Consistency-Based	Approximation fidelity to ground-truth Shapley	MAE or rank correlation between estimated and exact Shapley values.
	Stability and reproducibility analysis	Agreement of valuations across seeds, model types, or data splits.
Data Science Interpretability-Based	Human evaluation of value rankings	Expert inspection of top/bottom-valued samples for semantic consistency.
Data Science Efficiency-Based	Runtime and memory analysis	Measure scalability of computational resources.
	Sample efficiency analysis	Convergence of valuation estimates with sampling iterations or subset size.
Economics	Rationality and fair compensation analysis	Compliance with efficiency, monotonicity, subadditivity, and fair-share principles.
	Cost-efficiency and risk-adjusted value analysis	Ratios of value to cost or risk-adjusted valuations under uncertainty.
	Multi-dimensional and societal value assessment	Weighted-attribute or net-social-value frameworks.

Table 1: Taxonomy of evaluation metrics for data valuation methods.

A1. Data Removal/Addition Plots Given a dataset $D = \{(x_i, y_i)\}_{i=1}^n = \{z_i\}_{i=1}^n$ and valuation scores ϕ_i, let $\phi_{i_1} \geq \phi_{i_2} \geq \ldots \phi_{i_n}$ be the ranking from highest to lowest estimated value. Let $V(S)$ denote the utility of a model trained on subset $S \subseteq D$. To evaluate whether this ranking reflects true marginal utility, data removal and addition plots measure model performance as data are incrementally removed or added in fixed fractions rather than point by point.

Let $a \in (0, 1)$ be the removal/addition step size. Define $k_t = \lfloor t \cdot a \cdot n \rfloor$, $t = 0, 1, \ldots, T$, $T = \left\lfloor \frac{1}{a} \right\rfloor$. At step t, the high-value and low-value removal subsets are:

$$S_t^{\text{rem, high}} = D \setminus \left\{ z_{i_1}, z_{i_2}, \ldots, z_{i_{k_t}} \right\} \text{ and } S_t^{\text{rem, low}} = D \setminus \left\{ z_{i_n}, z_{i_{n-1}}, \ldots, z_{i_{n-k_t+1}} \right\} \quad (7)$$

The corresponding performance curves are $R_{\text{high}}(t) = V(S_t^{\text{rem, high}})$, $R_{\text{low}}(t) = V(S_t^{\text{rem, low}})$.

Analogously, for addition plots, we construct training subsets by adding the top-valued or bottom-valued k_t samples at step t. The high-value and low-value addition sets are

$$S_t^{\text{add, high}} = \{ z_{i_1}, z_{i_2}, \ldots, z_{i_{k_t}} \} \text{ and } S_t^{\text{add, low}} = \{ z_{i_n}, z_{i_{n-1}}, \ldots, z_{i_{n-k_t+1}} \} \quad (8)$$

with corresponding performance curves: $A_{\text{high}}(t) = V\left(S_t^{\text{add, high}}\right)$ and $A_{\text{low}}(t) = V\left(S_t^{\text{add, low}}\right)$.

To interpret these curves, one typically compares them against a random-ranking baseline. Let $R_{\text{rand}}(t)$ and $A_{\text{rand}}(t)$ denote the removal and addition curves obtained when the scores $\{\phi_i\}$ are replaced by random values. A valuation method is considered effective if, for most t,

$R_{\text{high}}(t) < R_{\text{rand}}(t)$ and $A_{\text{high}}(t) > A_{\text{rand}}(t)$, meaning that removing high-valued points degrades performance faster than random removal, while adding high-valued points accelerates performance improvement relative to random addition. The low-valued curves $R_{\text{low}}(t)$ and $A_{\text{low}}(t)$ are often reported as an additional sanity check, but the primary comparison is against random or ground-truth rankings rather than between high-valued and low-valued orderings alone. Figure 1 illustrates these behaviours on a dataset generated by a subset of the MNIST dataset [33].

This metric was first introduced by Ghorbani and Zou (2019) [20] and has since been widely adopted in the evaluation of Shapley-based and non-Shapley data valuation methods, including Jia et al. (2019) [25], Kwon and Zou (2021) [31], and Wang et al. (2023) [26].

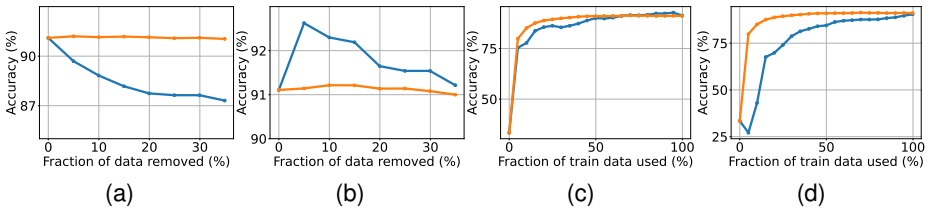

Fig. 1. Performance-based evaluation using data removal and addition plots on MNIST (subsets generated by classes 4, 5, 9). (a) Removal of highest-valued samples; (b) removal of lowest-valued samples; (c) addition of highest-valued samples; (d) addition of lowest-valued samples.

A2. Noisy Label Detection Another important form of performance-based evaluation examines whether a data valuation method can reliably identify mislabelled or corrupted samples. Define the set of truly corrupted labels as $\mathcal{M} = \{i : y_i \text{ is incorrect }\}$. After computing and ranking $\{\phi_i\}_{i=1}^n$ in ascending order. The noisy-label detection is then defined as:

$$\text{Det}(k) = \frac{\left|\mathcal{M} \cap \left\{i_1, \ldots, i_{\lfloor k \cdot n \rfloor}\right\}\right|}{|\mathcal{M}|}, \quad k \in [0, 1/2], \quad \text{Det}(k) \in [0, 1]. \tag{9}$$

which measures the proportion of true errors discovered within the bottom k-fraction. An effective valuation method has large fractions of low-valued points correspond to large $\text{Det}(k)$.

Figure 2a illustrates this metric on the Iris [14] dataset with injected 20% label noise. The left panels shows that Permutation Shapley achieves a steep rise in $\text{Det}(k)$, indicating that low-valued points strongly overlap with corrupted labels.

This evaluation strategy originates from [20], where Ghorbani and Zou(2019) first demonstrated that low Shapley-valued samples in Fashion-MNIST [46] often correspond to ambiguous or mislabelled images. The same principle has since been adopted in subsequent works, including Efficient Data Shapley [25], Beta Shapley [31], and Data Banzhaf [41], each of which reported a strong correlation between low-valued data and label noise or annotation inconsistency.

4.3 Consistency-Based Metrics

Consistency-based metrics complement performance-based evaluation by assessing the approximation fidelity and stability of data values. A reliable valuation method should yield consistent scores under changes in random initialisation, training seeds, or model architecture. High consistency suggests that the valuation captures intrinsic data importance rather than artifacts from stochastic optimisation or model-specific noise. Metrics in this class are useful for ensuring reliability in production pipelines and benchmarking reproducibility.

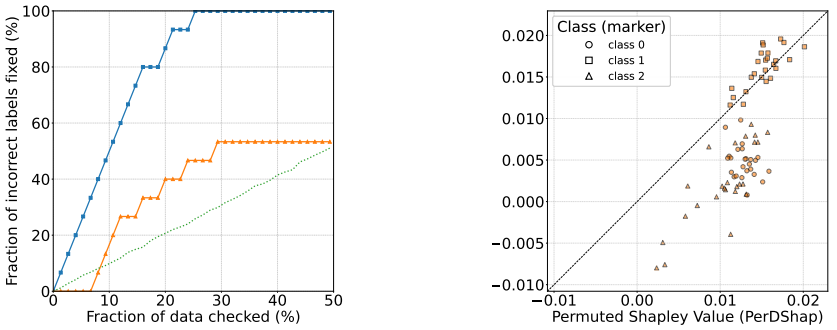

Fig. 2. a) Noisy label detection curve for Iris with 20% injected noise. b) Linear comparison between Permutation-based Data Shapley values and TMC Data Shapley values.

A3. Approximation Fidelity to Ground-Truth Data Shapley Another common quantitative criterion, especially in works proposing faster Data Shapley value approximations, is approximation fidelity. Here, the goal is to measure how closely an estimated valuation reproduces the true Data Shapley values, which is intractable for large datasets. For datasets with greater than 30 points, Permutation-Based Data Shapley is often served as a substitution. To measure fidelity, researchers compute both ground-truth value ϕ_i and candidate approximation $\hat{\phi}_i$ on the same dataset and compare them via pointwise deviation:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{\phi}_i - \phi_i\right| \in [0,\infty) \quad \text{or} \quad \text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\phi}_i - \phi_i\right)^2 \in [0,\infty) \tag{10}$$

or rank-based similarity:

$$\rho = \text{Spearman}(\phi, \hat{\phi}) \in [-1,1] \quad \text{or} \quad \tau = \text{Kendall}(\phi, \widehat{\phi}) \in [-1,1] \tag{11}$$

which quantify ordinal agreement. These comparisons are often visualised as scatter plots $(\phi_i, \hat{\phi}_i)$, as shown in Figure 2b. In this figure, TMC-Shapley values align closely with Permutation-based Data Shapley values, producing a near-diagonal trend and indicating high approximation fidelity.

This metric was first introduced by Jia et al. (2019) in Efficient Data Shapley [25], where fidelity measures were used to validate their sublinear-time approximation method. Kwon and Zou (2021) [31] further applied the same evaluation to demonstrate that Beta Shapley yields lower variance and stronger rank correlation with true Shapley values than Monte Carlo estimators. More recently, Wu et al. (2024) [45] employed these fidelity measures to benchmark their MLPbV framework against existing approximation algorithms, while OpenDataVal [26] incorporated correlation-based fidelity as one of its standardised evaluation dimensions.

A4. Stability and Reproducibility Analysis The goal of this metric is to evaluate the robustness and reproducibility of a data valuation method under varying experimental conditions. Let $\phi^{(r)} = (\phi_1^{(r)}, \ldots, \phi_n^{(r)})$ denote the valuation scores produced on run r, where runs differ in random seed, mini-batch order, or model hyper-parameters, etc. A stable method should satisfy:

$$\text{Corr}\left(\phi^{(r_i)}, \phi^{(r_j)}\right) \approx 1 \quad \forall i \neq j,\ \text{Corr} \in [-1,1] \tag{12}$$

where Corr is a rank-based similarity measure such as Spearman's ρ or Kendall's τ. High agreement indicates that the valuation reflects intrinsic data importance, rather than fluctuations arising from stochastic training dynamics.

This evaluation framework was formalized in OpenDataVal by Wang et al. (2023), which benchmarked both intra-method and inter-model stability across diverse datasets and valuation algorithms [26]. Similarly, Data Banzhaf (Wang and Jia, 2022) analyzed the robustness of the Banzhaf-value estimator under stochastic gradient noise, highlighting the need for variance reduction in data valuation [41]. These studies underscore that for Shapley-value–based methods—especially those employing deep neural networks—the stability analysis is not merely optional but essential for ensuring the credibility of the estimated data contributions.

Stability analysis reveals whether an algorithm reliably identifies influential samples across perturbations and is crucial for reproducibility in production settings. However, strong stability does not imply correctness, and low stability may reflect architectural differences rather than methodological flaws. Thus, stability should be interpreted as a complementary indicator, used alongside performance-based evaluations when assessing valuation quality.

4.4 Interpretability-Based Metrics

Interpretability-based metrics examine whether the numerical values assigned to by a valuation algorithm are semantically meaningful. Interpretability is especially critical in domains where human oversight and accountability are required, such as healthcare, finance and autonomous decision systems. Rather than emphasising predictive performance or consistency, these metrics evaluate the transparency of valuation outcomes. In other words, whether humans can make sense of why particular data points are deemed important is the main focus of these metrics. One representative evaluation strategy in this category is human evaluation of value rankings.

A5. Human-Evaluation of Value Rankings Human evaluation connects quantitative data valuations to human judgment and domain knowledge. While quantitative measures capture algorithmic consistency or performance impact, human evaluation examines whether the assigned values are semantically meaningful—that is, whether high-value samples correspond to representative, informative, or clean data, and low-value samples correspond to mislabelled, noisy, or ambiguous instances. This form of qualitative validation provides an interpretability check that bridges algorithmic reasoning and human intuition.

To conduct this qualitative evaluation, data points are first ranked according to their estimated values, after which the highest- and lowest-ranked samples are visually examined by human experts. Unlike quantitative metrics, this approach assesses whether the valuation method aligns with human intuition regarding sample quality and representativeness. High-value examples are typically expected to be clear, canonical, and semantically consistent instances of their class, whereas low-value examples are frequently mislabelled, corrupted, or visually ambiguous.

Figure 3 illustrates this process for an MNIST subset. The lowest-valued samples (Low-1 to Low-3) exhibit ambiguous digit form, such as distorted shapes and irregular stroke patterns. In contrast, the next three samples (High-1 to High-3) correspond to the highest-valued points and appear visually representative of corresponding classes. Such alignment with human intuition provides qualitative evidence that the method captures meaningful notions of data quality.

This evaluation strategy was first popularized in early works on data valuation such as Influence Functions (Koh and Liang, 2017) [28], Data Shapley (Ghorbani and Zou, 2019) [20], and Beta Shapley (Kwon and Zou, 2021) [31], each of which visually demonstrated that low-valued samples tend to contain mislabelled or atypical content, whereas high-valued samples are more prototypical. These studies established human evaluation as a persuasive means of verifying that data valuation algorithms capture intuitive notions of data quality and importance.

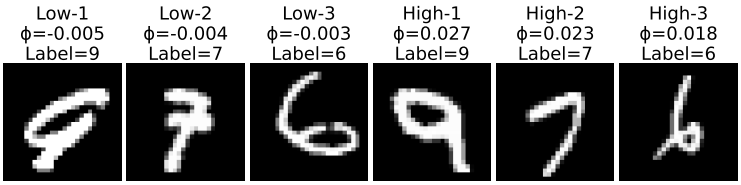

Fig. 3. Human inspection of MNIST samples ranked by Permutation-based Data Shapley values.

4.5 Efficiency-Based Metrics

Efficiency-based metrics evaluate the computational feasibility and scalability of data valuation algorithms. Since many valuation methods, particularly those based on cooperative-game theory, require repeated model retraining or sampling over data subsets, their computational cost can be prohibitive for large-scale machine learning applications. Efficiency evaluation thus focuses on quantifying how computation time, memory usage, and estimation accuracy scale with dataset size and model complexity. Two commonly used metrics in this category are runtime and memory and sample efficiency analysis, both of which capture different aspects of algorithmic practicality.

A6. Runtime and Memory Analysis Runtime and memory scaling analysis aims to assess the scalability and efficiency of a valuation algorithm, especially for Shapley-based methods that require repeated model training or performance evaluation over data subsets. This evaluation thus measures the algorithm's feasibility under realistic resource constraints, emphasizing its trade-off between computational efficiency and estimation accuracy.

To conduct this analysis, researchers systematically record the computational resources consumed during valuation, including wall-clock runtime, GPU or CPU utilization, number of model retraining or evaluations, and peak memory usage. The cost is measured as a function of dataset size n, model dimension d, or the number of sampled permutations k. In practice, results are reported either as growth rates (e.g., linear, sub-linear, or quadratic scaling) or through log–log performance plots illustrating the runtime–accuracy trade-off. This allows direct comparison of algorithmic efficiency across different valuation approaches.

This metric is widely known in other domains. It was first formalized to be utilized in Data Valuation by Jia et al. (2019) in Efficient Data Shapley [25], which demonstrated that their sublinear-time approximation method dramatically reduced computation compared to standard Monte Carlo Shapley estimation, achieving orders-of-magnitude acceleration while maintaining fidelity. The OpenDataVal benchmark by Wang et al. (2023) [26] further institutionalized this evaluation by providing standardized runtime and resource comparison protocols across diverse datasets and data valuation algorithms, establishing it as a core evaluation dimension.

A7. Sample Efficiency Analysis Sample efficiency analysis evaluates how quickly a sampling-based valuation method converges to stable and reliable estimates as the number of sampled permutations or subsets increases. For stochastic Shapley estimators, such as Permutation-based Data Shapley or TMC Shapley, let $\hat{\phi}^{(k)}$ denote the estimated values after k samples. Whereas approximation fidelity measures how accurately $\hat{\phi}^{(k)}$ matches the ground truth Shapley values at convergence, sample efficiency examines the rate at which $\hat{\phi}^{(k)} \to \hat{\phi}^{(\infty)}$ with increasing k.

To perform this evaluation, researchers track intermediate valuation estimates across sampling iterations and compute convergence indicators such as Spearman's ρ or Kendall's τ between intermediate and final rankings, along with MAE or MSE trends. Alternatively, performance-based diagnostics plot downstream performance (e.g., test accuracy or AUROC) against the number of iterations to visualise convergence speed. Faster stabilisation or steeper curves indicate higher sample efficiency.

This evaluation was first introduced by Ghorbani and Zou (2019) in Data Shapley [20] and later refined in Efficient Data Shapley (Jia et al., 2019) [25], which demonstrated that sublinear-time approximations could achieve convergence with significantly fewer samples. Subsequent works, including Beta Shapley (Kwon and Zou, 2021) [31] and Data Banzhaf (Wang and Jia, 2022) [41], also reported convergence analyses to verify the stability and efficiency of their stochastic estimators. The OpenDataVal benchmark [26] further standardized this evaluation, adopting sample efficiency curves as a key diagnostic for assessing the trade-off between accuracy and computational cost across methods.

5 Evaluation Metrics in the View of Economics

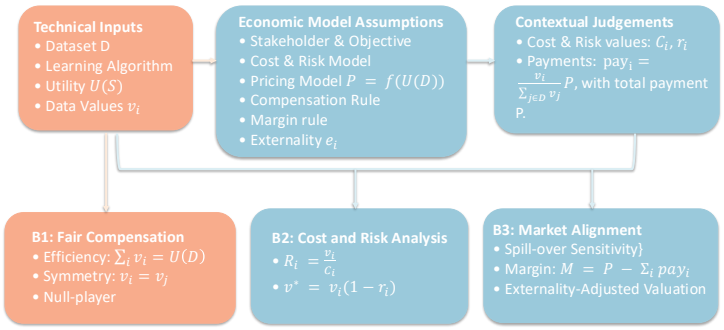

Fig. 4. Illustration of implementing B1–B3. Technical inputs support B1, while B2 and B3 additionally require economic model assumptions and contextual judgments.

The two perspectives in evaluation metrics are inherently complementary rather than competing. While evaluation in data science emphasises empirical validity, the economic perspective focuses on the theoretical soundness, fairness, and institutional consistency of valuation mechanisms. From this viewpoint, data function as economic goods exchanged among multiple stakeholders, including data owners, buyers, and intermediaries. In economics, the assessment of valuation mechanisms is conducted by checking whether the mechanism adheres to normative criteria drawn from cooperative game theory, pricing theory, and welfare economics.

In the following subsections, we synthesise key economic evaluation criteria, including Fair Compensation Analysis, Cost Efficiency, and Risk-Adjusted value analysis, Multi-Dimensional and Societal Value Assessment. These criteria evaluate the outputs of a valuation mechanism, rather than empirical model performance, and assess whether the assigned values align with economic principles. Then we discuss how they can be operationalised and aligned with data-science-oriented evaluations to form a unified assessment framework.

B1. Fair Compensation Analysis This metric evaluates whether a data-valuation mechanism allocates value to contributors in accordance with foundational fairness axioms of cooperative game theory. A method that achieves fair compensation rewards contributors proportionally to their marginal impact on utility, while assigning negligible value to those who do not contribute.

Let $U(D)$ denote the total utility generated by dataset D, and let v_i denote the value assigned to contributor i. Following [20], a fair valuation should satisfy:

– **Efficiency** The total value allocated to contributors equals the total utility of the dataset:

$$\sum_i v_i = U(D) \tag{13}$$

– **Symmetry** Contributors with identical marginal effects receive identical valuations:

$$v_i = v_j \tag{14}$$

– **Null-Player** The contributor whose inclusion never changes utility receives zero value:

$$v_i = 0 \tag{15}$$

To apply this criterion, one first computes the valuation v_i and examines: i) the efficiency gap $|\sum_i v_i - U(D)|$; ii) the consistency of values among contributors known to have equivalent marginal influence; iii) whether contributors with $v_i \approx 0$ indeed exhibit negligible utility change. These checks assess how closely the valuation method adheres to the fairness principles.

B2. Cost-Efficiency and Risk-Adjusted Value Analysis Whereas fair compensation concerns equity among contributors, cost-efficiency analysis examines whether a valuation method reflects the economic costs and risks of acquiring or managing data. A well-calibrated method should assign higher value to data that deliver strong utility relative to their cost and discount data that incur significant financial, operational, or regulatory risk. This aligns data valuation with principles of rational resource allocation and business decision-making.

Cost and risk considerations are central to the economic frameworks of Fleckenstein et al. (2023), who propose a composite model incorporating acquisition cost, ownership, privacy exposure, and maintenance burden as weighted dimensions of data value [15]. Coyle et al. (2024) further discuss risk-adjusted valuation using real-options theory to quantify uncertainty and optionality in data-driven investments [7]. Together, these works motivate evaluating valuation algorithms not only by utility contribution but also by economic feasibility.

Let C_i denote the estimated cost of obtaining or maintaining datum i, and let r_i denote its normalised risk (e.g., privacy exposure or regulatory vulnerability). A valuation method can be evaluated using cost-efficiency and risk-adjusted indicators, such as:

$$R_i = v_i / C_i \in (0, \infty) \quad and \quad v^* = v_i(1 - r_i) \in (-\infty, \infty) \tag{16}$$

A method is considered economically sound if high-utility data generally yield high R_i and non-inflated v_i^*, indicating alignment between estimated value and economic feasibility.

B3. Intermediary and Regulatory Alignment Analysis This metric evaluates whether a data-valuation mechanism aligns with the structural interests of intermediaries and regulatory objectives. Under this perspective, data are treated as economic goods transacted via intermediaries. Therefore, valuation methods should reflect not only contributor and buyer interests but also the wider market architecture and governance concerns.

Studies of data intermediaries, such as Schweihoff et al. (2024), examine how intermediary platforms organise data sharing, governance, and value-creating in the data economy [23]. Concurrently, research on externalities in data markets, such as Hossain and Chen (2023), models how buyer and seller interactions and spill-over effects shape welfare outcomes [39]. Taken together, these studies provide the theoretical foundation for evaluating valuation methods through an intermediary or regulatory lens, examining not just direct contributor or buyer interests, but how data governance, market structure, and externalities influence value allocation.

Let v_i denote the value assigned by a valuation method to datum i. A mechanism aligned with intermediary/regulatory concerns should satisfy or exhibit the following features:

	Data-Science							Economics		
	A1	A2	A3	A4	A5	A6	A7	B1	B2	B3
CG Influence [28]	✓	✓	✓	×	✓	✓	×	×	×	×
LiSSA [1]	✓	×	×	×	✓	×	✓	×	×	×
Arnoldi Influence [37]	×	✓	✓	✓	✓	×	✓	×	×	×
EKFAC Influence [18]	×	×	×	✓	✓	×	×	×	×	×
Datainf [30]	×	✓	✓	×	✓	×	✓	×	×	×
Nyström Influence [22]	×	×	×	×	✓	✓	✓	×	×	×
Data Shapley [20]	✓	✓	×	×	✓	×	×	×	×	×
Data Banzhaf [41]	✓	✓	✓	×	✓	✓	✓	×	×	×
CS Shapley [38]	✓	✓	✓	✓	×	✓	✓	✓	×	×
Beta Shapley [31]	✓	✓	✓	✓	×	✓	✓	✓	×	×
Delta Shapley [44]	✓	×	✓	×	×	✓	✓	✓	×	×
KNN Shapley [24]	✓	×	✓	✓	✓	✓	✓	✓	×	×
GT Shapley [25]	✓	✓	✓	×	×	✓	✓	✓	✓	×
Owen Shapley [4]	×	×	✓	✓	×	✓	✓	✓	×	×
Least Core [47]	✓	✓	✓	✓	×	✓	✓	✓	×	✓
Data OOB [32]	✓	✓	×	✓	×	✓	✓	×	×	×
LAVA [27]	✓	✓	✓	✓	×	✓	✓	✓	×	×
DVRL [48]	✓	✓	×	✓	×	✓	✓	×	×	×
Query-Based DP [29]	×	×	×	×	×	✓	×	✓	×	×
Option Approach [7]	×	×	×	×	×	×	×	✓	✓	×
Scoring-Based DV [15]	×	×	×	×	×	×	×	✓	✓	✓
Records DV [17]	×	×	×	×	×	×	×	✓	✓	×
Privacy DP [34]	×	×	×	×	×	×	×	✓	✓	✓
Equitable DM [23]	×	×	×	×	×	×	×	×	×	✓
Data Intermediaries [39]	×	×	×	×	×	×	×	×	×	✓

Table 2. Mapping of existing data valuation method to the proposed evaluation taxonomy. Each row corresponds to a representative valuation method from the literature, while each column denotes one of the ten evaluation metrics introduced in this review.

- **Spill-over Sensitivity**. Data pooled or brokered via intermediaries often creates positive network benefits or externalities. An evaluation investigates whether v_i increases with an externality factor e_i (for instance, pool size or number of downstream users).
- **Brokerage Margin Transparency**. In a transparent intermediation model, the owners' total allocated value $V = \sum_i v_i$ plus intermediary surplus M should approximate the total market value U: a small gap $\big|U - (V + M)\big|$ signals alignment with fair intermediation.
- **Externality-Adjusted Valuation**. Regulators also care about both positive and negative externalities (e.g., innovation spill-overs, privacy risks). Given a normalised externality score x_i, the evaluation checks whether valuations v_i are positively or negatively correlated accordingly (e.g., high v_i for high positive x_i, and low v_i for high negative x_i).

6 Discussion: Practical Guidance for Applying Evaluation Metrics

Although our taxonomy covers many evaluation metrics, their application requires careful judgment. Different metrics serve distinct purposes, and misuse may yield misleading conclusions.

Performance-based metrics, such as removal and addition plots (A1), are appropriate for assessing whether a method truly reflects marginal utility. However, they may be unreliable under high stochasticity, small datasets, or unstable retraining. Noisy label detection (A2) faces similar limitations and may fail when ambiguity or systematic noise drives valuation outcomes.

Consistency-based metrics (A3, A4) assess reliability and reproducibility, particularly for approximation methods. Yet stability does not imply correctness: methods may be consistently biased, and instability may arise from model randomness. These metrics should therefore complement performance-based evaluation.

Interpretability-based evaluation, such as human inspection of extreme-ranked samples, is compelling for visual or semantically meaningful data but remains subjective and less suited to high-dimensional settings.

Efficiency-based metrics, including runtime and sample-efficiency, are critical for large-scale applications, though comparisons are sensitive to implementation and hardware differences.

Economic-oriented criteria concern fairness, cost-efficiency, and regulatory alignment. They are crucial in governance contexts but often depend on context-specific assumptions.

The framework also extends naturally to large language models (LLMs). Because the metrics are model-agnostic and utility-driven, they apply whenever a measurable objective is defined. In LLMs, utility can be defined through standard language modelling objectives, most notably perplexity. As demonstrated by Wang et al. [42], Shapley-based data valuation can be efficiently integrated into large-scale neural training processes. In such settings, model performance $V(S)$ can be operationalised as validation perplexity. Consequently, performance-based evaluation metrics such as data removal or addition plots (A1) can measure how perplexity changes when high- or low-valued text samples are removed or added, while noisy-label detection (A2) can be adapted to identify corrupted, low-quality, or adversarial text data. However, despite these natural extensions, data valuation for LLMs remains an emerging research frontier. Therefore, while the proposed evaluation metrics are conceptually applicable to LLMs, the development of robust and scalable data valuation methods tailored to large-scale generative models remains an open and promising direction for future research.

Overall, metric selection should reflect the purpose of the study: empirical machine learning tasks benefit primarily from performance-, consistency-, and efficiency-based criteria, whereas applications involving data sharing, pricing, or governance require economic-oriented assessment. A robust evaluation framework should combine insights across these perspectives, balancing empirical validity with economic soundness.

7 Conclusion

Data valuation lies at the intersection of artificial intelligence and economics, shaping how data is measured, traded, and governed. Despite rapid methodological progress, the evaluation of data valuation methods remains fragmented and lacks standardisation. This paper addresses this gap by presenting the first unified, cross-domain review of evaluation metrics, combining empirical evaluation practices from data science with institutional and welfare-oriented principles from economics. From the data science perspective, we organised existing evaluation metrics into four categories–performance, consistency, interpretability, and efficiency–capturing a distinct aspect of algorithmic behaviour. From the economic perspective, we introduced complementary criteria on fair compensation, cost-efficiency, risk-adjusted value, and regulatory alignment. Together, these perspectives connect micro-level reliability with the macro-level institutional soundness. Bridging these viewpoints is essential. By framing evaluation as a cross-disciplinary challenge and promoting transparent, reproducible benchmarks, this study advances more rigorous and equitable data valuation research.

Acknowledgments. The work was partially supported by the National Natural Science Foundation of China under Grant No.62376236 and the Shenzhen Stability Science Program 2023.

Disclosure of Interests. The authors have no competing interests.

References

1. Agarwal, N., Bullins, B., Hazan, E.: Second-order stochastic optimization for machine learning in linear time. Journal of Machine Learning Research **18**(116), 1–40 (2017)
2. Belsley, D.A., Kuh, E., Welsch, R.E.: Regression diagnostics: Identifying influential data and sources of collinearity. John Wiley & Sons (2005)
3. Bendechache, M., Attard, J., Ebiele, M., Brennan, R.: A systematic survey of data value: models, metrics, applications and research challenges. IEEE Access **11**, 104966–104983 (2023)
4. Castro, J., Gómez, D., Tejada, J.: Polynomial calculation of the shapley value based on sampling. Computers & operations research **36**(5), 1726–1730 (2009)
5. Cook, R.D.: Detection of influential observation in linear regression. Technometrics **19**(1), 15–18 (1977)
6. Cook, R.D., Weisberg, S.: Residuals and influence in regression. New York: Chapman and Hall (1982)
7. Coyle, D., Gamberi, L.: A real options approach to data valuation. Business Economics **59**(4), 227–234 (2024)
8. Dai, W., Fan, J.: AutoUAD: Hyper-parameter optimization for unsupervised anomaly detection. In: The Thirteenth International Conference on Learning Representations (2025)
9. Deng, J., Li, T.W., Zhang, S., Liu, S., Pan, Y., Huang, H., Wang, X., Hu, P., Zhang, X.: `dattri`: A library for efficient data attribution. Advances in Neural Information Processing Systems **37**, 136763–136781 (2024)
10. Ebiele, M., Bendechache, M., Brennan, R.: Quantitative data valuation methods: A systematic review and taxonomy. ACM Journal of Data and Information Quality **17**(2), 1–39 (2025)
11. Fan, J.: Graph minimum factor distance and its application to large-scale graph data clustering. In: Forty-second International Conference on Machine Learning (2025)
12. Fan, J.: An interdisciplinary and cross-task review on missing data imputation. arXiv preprint arXiv:2511.01196 (2025)
13. Fan, Z., Fang, H., Zhou, Z., Pei, J., Friedlander, M.P., Liu, C., Zhang, Y.: Improving fairness for data valuation in horizontal federated learning. In: 2022 IEEE 38th International Conference on Data Engineering (ICDE). pp. 2440–2453. IEEE (2022)
14. Fisher, R.A.: Iris. UCI Machine Learning Repository (1936). https://doi.org/https://doi.org/10.24432/C56C76
15. Fleckenstein, M., Obaidi, A., Tryfona, N.: A review of data valuation approaches and building and scoring a data valuation model. Harvard Data Science Review **5**(1) (2023)
16. Fu, D., Fan, J.: Uniod: A universal model for outlier detection across diverse domains. arXiv preprint arXiv:2507.06624 (2025)
17. Galperti, S., Levkun, A., Perego, J.: The value of data records. Review of Economic Studies **91**(2), 1007–1038 (2024)
18. George, T., Laurent, C., Bouthillier, X., Ballas, N., Vincent, P.: Fast approximate natural gradient descent in a kronecker factored eigenbasis. Advances in neural information processing systems **31** (2018)
19. Ghorbani, A., Kim, M., Zou, J.: A distributional framework for data valuation. In: International Conference on Machine Learning. pp. 3535–3544. PMLR (2020)

20. Ghorbani, A., Zou, J.: Data shapley: Equitable valuation of data for machine learning. In: International conference on machine learning. pp. 2242–2251. PMLR (2019)
21. Hao, J., Deng, Z., Li, J.: The evolution of data pricing: From economics to computational intelligence. Heliyon **9**(9) (2023)
22. Hataya, R., Yamada, M.: Nyström method for accurate and scalable implicit differentiation. In: International Conference on Artificial Intelligence and Statistics. pp. 4643–4654. PMLR (2023)
23. Hossain, S., Chen, Y.: Equilibrium of data markets with externality. Proceedings of the 41st International Conference on Machine Learning (2024)
24. Jia, R., Dao, D., Wang, B., Hubis, F.A., Gurel, N.M., Li, B., Zhang, C., Spanos, C.J., Song, D.: Efficient task-specific data valuation for nearest neighbor algorithms. Proc. VLDB Endow. **12**(11) (2019)
25. Jia, R., Dao, D., Wang, B., Hubis, F.A., Hynes, N., Gürel, N.M., Li, B., Zhang, C., Song, D., Spanos, C.J.: Towards efficient data valuation based on the shapley value. In: The 22nd International Conference on Artificial Intelligence and Statistics. pp. 1167–1176. PMLR (2019)
26. Jiang, K., Liang, W., Zou, J.Y., Kwon, Y.: Opendataval: a unified benchmark for data valuation. Advances in Neural Information Processing Systems **36**, 28624–28647 (2023)
27. Just, H.A., Kang, F., Wang, J.T., Zeng, Y., Ko, M., Jin, M., Jia, R.: Lava: Data valuation without pre-specified learning algorithms. The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=JJuP86nBl4q
28. Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions. In: International conference on machine learning. pp. 1885–1894. PMLR (2017)
29. Koutris, P., Upadhyaya, P., Balazinska, M., Howe, B., Suciu, D.: Query-based data pricing. Journal of the ACM (JACM) **62**(5), 1–44 (2015)
30. Kwon, Y., Wu, E., Wu, K., Zou, J.: Datainf: Efficiently estimating data influence in lora-tuned llms and diffusion models. The Twelfth International Conference on Learning Representations (2024)
31. Kwon, Y., Zou, J.: Beta shapley: a unified and noise-reduced data valuation framework for machine learning. arXiv preprint arXiv:2110.14049 (2021)
32. Kwon, Y., Zou, J.: Data-oob: Out-of-bag estimate as a simple and efficient data value. In: International conference on machine learning. pp. 18135–18152. PMLR (2023)
33. LeCun, Y., Cortes, C., Burges, C.: Mnist handwritten digit database. ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist **2** (2010)
34. Li, C., Li, D.Y., Miklau, G., Suciu, D.: A theory of pricing private data. ACM Transactions on Database Systems (TODS) **39**(4), 1–28 (2014)
35. Mohan, S.K., Bharathy, G., Jalan, A.: Enterprise data valuation—a targeted literature review. Journal of Economic Surveys (2025)
36. Pohl, M., Haertel, C., Staegemann, D., Turowski, K.: Data valuation methods-a literature review. The annual Americas Conference on Information Systems (AMCIS) (2023)
37. Schioppa, A., Zablotskaia, P., Vilar, D., Sokolov, A.: Scaling up influence functions. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 8179–8186 (2022)
38. Schoch, S., Xu, H., Ji, Y.: Cs-shapley: class-wise shapley values for data valuation in classification. Advances in Neural Information Processing Systems **35**, 34574–34585 (2022)
39. Schweihoff, J., Lipovetskaja, A., Jussen-Lengersdorf, I., Möller, F.: Stuck in the middle with you: Conceptualizing data intermediaries and data intermediation services. Electronic Markets **34**(1), 48 (2024)
40. Sim, R.H.L., Xu, X., Low, B.K.H.: Data valuation in machine learning:" ingredients", strategies, and open challenges. In: IJCAI. pp. 5607–5614 (2022)
41. Wang, J.T., Jia, R.: Data banzhaf: A robust data valuation framework for machine learning. In: International Conference on Artificial Intelligence and Statistics. pp. 6388–6421. PMLR (2023)

42. Wang, J.T., Mittal, P., Song, D., Jia, R.: Data shapley in one training run. In: Yue, Y., Garg, A., Peng, N., Sha, F., Yu, R. (eds.) International Conference on Learning Representations. vol. 2025, pp. 12358–12395 (2025), https://proceedings.iclr.cc/paper_files/paper/2025/file/20fdaf67581e6d7157376d1ed584040a-Paper-Conference.pdf
43. Wang, T., Rausch, J., Zhang, C., Jia, R., Song, D.: A principled approach to data valuation for federated learning, pp. 153–167. Springer (2020)
44. Watson, L., Kujawa, Z., Andreeva, R., Yang, H.T., Elahi, T., Sarkar, R.: Accelerated shapley value approximation for data evaluation. arXiv preprint arXiv:2311.05346 (2023)
45. Wu, O., Zhu, W., Li, M.: Is data valuation learnable and interpretable? arXiv preprint arXiv:2406.02612 (2024)
46. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. CoRR **abs/1708.07747** (2017), http://arxiv.org/abs/1708.07747
47. Yan, T., Procaccia, A.D.: If you like shapley then you'll love the core. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 5751–5759 (2021)
48. Yoon, J., Arik, S., Pfister, T.: Data valuation using reinforcement learning. In: International Conference on Machine Learning. pp. 10842–10851. PMLR (2020)