Data Valuation in Machine Learning: "Ingredients", Strategies, and Open Challenges

Rachael Hwee Ling Sim[*], **Xinyi Xu**[*] and **Bryan Kian Hsiang Low**
Department of Computer Science, National University of Singapore
{rachaels, xuxinyi, lowkh}@comp.nus.edu.sg

Abstract

Data valuation in *machine learning* (ML) is an emerging research area that studies the worth of data in ML. Data valuation is used in collaborative ML to determine a fair compensation for every data owner and in interpretable ML to identify the most responsible, noisy, or misleading training examples. This paper presents a comprehensive technical survey that provides a new formal study of data valuation in ML through its "ingredients" and the corresponding properties, grounds the discussion of common desiderata satisfied by existing data valuation strategies on our proposed ingredients, and identifies open research challenges for designing new ingredients, data valuation strategies, and cost reduction techniques.

1 Introduction

As *machine learning* (ML) and data sharing become more ubiquitous, there is greater interest by data owners in how much their data is worth and by model owners in how to explain model predictions and what data to add or remove to improve model performance. The former is evidenced by the emergence of data marketplaces (e.g., Ocean Protocol) for users to buy and sell data and initiatives such as the Data Dividend Project where consumers are banding together to demand property rights to their data under the California Consumer Privacy Act and receive payments from big tech companies [Rozemberczki and Sarkar, 2021]. The latter is evidenced by the popularity of interpretable ML methods in various fields, including healthcare.

The above concerns can be addressed via data valuation. In data valuation, the value of data contributed by a data owner is influenced by the data from other owners. Data valuation can be used in *collaborative ML* (CML) to determine a *fair* compensation (or share of a fixed monetary sum) for every data owner to incentivize them to share data [Sim *et al.*, 2020; Tay *et al.*, 2022]. The fair compensation is usually some affine or monotonic transformation of the contributed data value [Jia *et al.*, 2019a; Sim *et al.*, 2020]. Data valuation can also be used in *interpretable ML* to attribute the model predictions and accuracy to the most responsible training examples from the data. Often, only the ranking of data values (instead of their absolute values) is needed. Model owners can improve their models by actively collecting data similar to the high-value training examples or removing noisy and misleading training examples with low values [Ghorbani and Zou, 2019; Yoon *et al.*, 2020]. Lastly, data valuation is useful for domain adaptation by assigning higher value to data with an underlying distribution similar to the target one [Yoon *et al.*, 2020].

It is thus timely to introduce a technical survey that (a) provides a new formal study of data valuation in ML through its "ingredients" and the corresponding properties (e.g., performance metric and its design principles) (Sec. 2), (b) grounds the discussion of common desiderata satisfied by existing data valuation strategies (e.g., Shapley value) on our proposed ingredients (Sec. 3), and (c) identifies open research challenges for designing new ingredients, data valuation strategies, and cost reduction techniques with specific examples (Sec. 4). All these distinguish the novel contributions of our survey from that of Cong *et al.* [2021] and Pei [2020]. Other surveys have either focused on the economics perspective (e.g., nature of data as a merchandise, costs for data acquisition, government regulation) [Fricker and Maksimov, 2017; Liang *et al.*, 2018; Muschalle *et al.*, 2013; Raskar *et al.*, 2019; Zhang and Beltrán, 2020] or the *cooperative game theory* (CGT) based valuation strategies [Bax, 2019].

2 "Ingredients" of Data Valuation

Suppose that there is a set of n data owners denoted as $N \triangleq \{1, \ldots, n\}$. The objective of data valuation is to determine the value (or contribution/payoff) $\{\phi_k\}_{k\in N}$ of dataset D_k for data owner k. The valuation function ϕ is a function of several *"ingredients"*: the family of datasets $\mathcal{D} \triangleq \{D_k\}_{k\in N}$ from all data owners, the learning algorithm $\mathcal{A}$, and the performance metric U. Formally, $\phi_k \triangleq \phi(k, \mathcal{D}, \mathcal{A}, U)$, as shown in Fig. 1. The learning algorithm $\mathcal{A}$ is a function that takes in an arbitrary training dataset D and outputs a model. The performance metric U takes in any model $\mathcal{A}(D)$ or dataset D to output a real-valued performance score. For any subset $C \subseteq N$ of data owners, $D_C \triangleq \bigcup_{k\in C} D_k$, so, D_N denotes the aggregated dataset of all owners. Before giving examples of the data valuation strategy ϕ, we will discuss how the choices of the ingredients can affect the choice and computation of ϕ.

[*]Equal contribution.

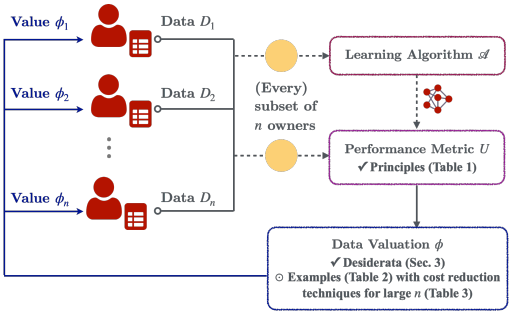

Figure 1: Overview of the data valuation problem with n data owners. The learning algorithm $\mathcal{A}$ (e.g., logistic regression) takes in data and produces a model. The performance metric U (e.g., validation accuracy) used to evaluate the value of data/model is selected to follow certain principles. The data valuation strategy ϕ (e.g., Shapley value [Sec. 3.2]) is selected to satisfy certain desiderata. We may consider the performance metric for multiple subsets of data owners and techniques to reduce the computational cost.

2.1 Data, D

Both the number n of data owners and the maximum dataset size $\max_{k \in N} |D_k|$ affect the cost of computing or approximating ϕ. The presence of noisy data affects the choice of the performance metric U [Ghorbani and Zou, 2019]. Lastly, whether D_k is a fixed dataset or a random/biased sample from an underlying distribution may also affect the choice of ϕ [Abay *et al.*, 2019; Koenecke and Varian, 2020].

2.2 Learning Algorithm, $\mathcal{A}$

The value of data may vary across learning algorithm $\mathcal{A}$, and the choice of $\mathcal{A}$ affects the choice of performance metric U. When $\mathcal{A}$ is undecided, U would solely be designed based on data-driven principles. As another example, when the model has no parameters or the parameters are not directly comparable across models (e.g., neural networks), the performance metric cannot be directly based on the parameters. Furthermore, the choice of $\mathcal{A}$ (and U) affects the cost of computing ϕ and its approximations. We will outline several learning algorithms that are important to our discussion of data valuation.

The *k-nearest neighbours* (k-NN) is a commonly used learning algorithm that predicts the label for a datum based on the known values of several similar data/nearest neighbours [Jia *et al.*, 2019a]. Another class of learning algorithms involves *empirical risk minimization* (ERM) using gradient-based techniques such as logistic regression and *deep neural networks* (DNNs) [Ghorbani and Zou, 2019]. Bayesian models such as Bayesian linear regression offer the distinct advantage of providing an uncertainty measure in the predictions; intuitively, a lower uncertainty is more desirable [Sim *et al.*, 2020]. Generative models [Goodfellow *et al.*, 2014; Kingma and Welling, 2014] are typically used to learn an effective representation of the data and can be used to generate synthetic data that can be operationally easier to share than real data due to privacy regulation [Tay *et al.*, 2022].

2.3 Performance Metric, U

The performance score $U_{\mathcal{A}}(D) \triangleq U(\mathcal{A}(D))$ or $U(D)$ for more desirable models or datasets should be higher. We have identified a few design principles which are model- or data-driven. Table 1 summarizes how different performance metrics in existing work satisfy multiple design principles.

Model-driven Principles

P1 **Validation performance (VP).** Intuitively, a model with better performance on a validation set is more valuable. Possible choices of VP include accuracy [Ghorbani and Zou, 2019; Jia *et al.*, 2019b] and negated loss functions. VP also depends on the task, such as whether we are predicting the likelihood vs. severity of illness (in a medical use case). VP can distinguish between data that improve model performance from data that worsen model performance (e.g., noisy/adversarial data or from a different domain). Thus, VP is suitable for use cases that involve model interpretability, domain adaptation, and noisy/adversarial data. However, it may be practically difficult to obtain a validation set that all data owners agree on [Sim *et al.*, 2020].

P2 **Uncertainty.** A model with a lower uncertainty in its model parameters and predictions is more valuable. This principle is theoretically ensured by using *information gain* (IG) on the model parameters/function from observing the data outputs as the performance metric [Sim *et al.*, 2020]. IG measures the expected reduction in entropy capturing uncertainty across all possible data outputs. Moreover, IG can be used as a surrogate measure of the predictive accuracy of a trained model [Krause and Guestrin, 2007; Kruschke, 2008] when the test queries are not known *a priori*. Hence, it is especially useful when a validation set is not available.

P3 **Order of data arrival.** For certain $\mathcal{A}$ and optimization techniques (e.g., *deep neural networks* (DNNs) and gradient-based optimization), sharing the same data earlier is more valuable. The rationale is that model performance is less likely to be improved or influenced by new data if the model (e.g., DNN) is closer to convergence at the later stages in a gradient-based optimization. Thus, data owners should be more highly valued for sharing data in the early stage of such optimization, like in federated learning [Song *et al.*, 2019].

Data-driven Principles

These principles are defined based on D instead of the model (i.e., $U(D)$ directly instead of $U(\mathcal{A}(D))$).

P4 **Monotonicity.** Having more data is more (or at least equally) valuable [Sim *et al.*, 2020; Tay *et al.*, 2022; Xu *et al.*, 2021b].

P5 **Submodularity.** The same data is less valuable when added to a larger dataset (i.e., law of diminishing returns). The IG on Bayesian model parameters as U satisfies submodularity [Sim *et al.*, 2020].

P6 **Similarity to reference distribution.** Dataset D is more valuable when it is similar and representative of the aggregated data D_N (or the target task's distribution). Tay *et al.* [2022] measured similarity using the translated negative *maximum mean discrepancy* (MMD) between the underlying distributions $\mathbb{P}$ and $\mathbb{P}^*$ for the respective datasets

D and D_N, i.e., $-\text{MMD}(\mathbb{P}, \mathbb{P}^*)$. As this quantity is intractable, Tay *et al.* [2022] have resorted to its (biased) estimate $-\widehat{\text{MMD}}_b(D, D_N \cup \mathcal{G})$, where $\mathcal{G}$ is a synthetic dataset generated by a generative model trained on D_N. Separately, Xu *et al.* [2021a] have considered the cosine similarity (vector alignment) between (i) the model parameter update from using data from each owner k only and (ii) the aggregated model parameter update across all owners (which uses data from every owner).

P7 **Intra-diversity.** Dataset D_k is more valuable and will improve the model's predictive performance if it contains more diverse data points covering a larger region of the input space. Xu *et al.* [2021b] have measured diversity using the volume (i.e., square root of the determinant) of the left $m \times m$ Gram matrix as U (of data with m features). The volume is theoretically shown to positively correlate with the regression performance: A high volume suggests a low mean squared error.

P8 **Replication robustness.** Replicated data should not be highly valuable as adding the same data repeatedly to an ML model does not improve its performance significantly. Xu *et al.* [2021b] have provided an upper bound on their diversity-based performance metric from direct (up to infinite) replication, which does not increase the diversity in data. This principle prevents dishonest owners from gaming the data valuation [Ohrimenko *et al.*, 2019].

P9 **Cost.** Data that is costly (to obtain/maintain) should be more valuable to incentivize the data owners' participation [Heckman *et al.*, 2015]. The costs include that of data acquisition, storage, and transport [Agarwal *et al.*, 2019; Devereaux *et al.*, 2016].

P10 **Timeliness.** Data that is timely (i.e., up-to-date) is more valuable [Stahl and Vossen, 2016]. Service providers can adjust their market operation strategies using real-time data to provide customized services, for example, to users of the bicycle-sharing services [Zhao *et al.*, 2020]. This strategy is ineffective with outdated data.

3 Data Valuation Strategies, ϕ

We will discuss various data valuation strategies ϕ. ϕ is a more general concept than the performance metric U as ϕ can additionally depend on the datasets D_C of others. In other words, $\phi_k \triangleq U(D_k)$ or $U_{\mathcal{A}}(D_k)$ is only one possible specification of ϕ. We will describe several desiderata from existing works and *cooperative game theory* (CGT) literature for designing ϕ. Importantly, the suitable desiderata to a certain application scenario of data valuation (and the properties of its ingredients) can guide the specific choice of ϕ. For example, fairness is related to how well each owner's value ϕ_k reflects his contribution and is captured by D4-D6 [Chalkiadakis *et al.*, 2011]. To ensure fairness, we look for a ϕ that satisfies D4-D6 in Table 2. We adopt the notation convention from CGT and let v be the characteristic function that maps each subset of owners/coalition to a value, i.e., $v : 2^N \mapsto \mathbb{R}$. For data valuation, we define $v(C) \triangleq U_{\mathcal{A}}(D_C)$ or $U(D_C)$.

Principles	Performance Metric $U(\cdot)$
Model-driven	
Validation performance (P1)	• Classification accuracy [Jia *et al.*, 2019a; Jia *et al.*, 2019b; Jia *et al.*, 2019c; Ghorbani and Zou, 2019; Ghorbani *et al.*, 2020; Yoon *et al.*, 2020]
Uncertainty (P2)	• Information gain on Bayesian models' parameters [Sim *et al.*, 2020] • *Uncertainty in predictions*
Order of data arrival (P3)	• Higher weight in earlier iterations of federated learning [Song *et al.*, 2019] • *Other ML applications such as reinforcement learning*
Data-driven	
Monotonicity (P4)	• [Sim *et al.*, 2020]
Submodularity (P5)	• Volume of the left Gram matrix [Xu *et al.*, 2021b]
Similarity (P6)	• Translated negated MMD divergence between data and generative models' learned data distribution [Tay *et al.*, 2022][1] • Gradient alignment [Xu *et al.*, 2021a]
Diversity (P7), Replication robustness (P8)	• [Xu *et al.*, 2021b]
Cost (P9)	• [Heckman *et al.*, 2015]
Timeliness (P10)	• [Stahl and Vossen, 2016]

Table 1: Summary of design principles of performance metrics. *Italics* indicates work yet to be done.

D1 **Low computational cost.** The total cost of computing ϕ, which depends on the number of evaluations of $\mathcal{A}$ and the cost of each evaluation, should be low enough to be tractable/efficient. The number of evaluations may depend on the number n of data owners. When $\mathcal{A}$ is expensive and n is large, we will prefer ϕ that can be computed *without* or with fewer (e.g., $\mathcal{O}(\log n)$) evaluations of $\mathcal{A}$. As an example, when millions of data owners [Singhal *et al.*, 2021] contribute their data for the training of a DNN that minimises the empirical risk, we want to avoid retraining on multiple data subsets from scratch.

D2 **Dependence on $\mathcal{D}$.** The value ϕ_k of D_k should depend on other datasets in $\mathcal{D}$ [Agarwal *et al.*, 2019]. Formally, owner k's relative valuations to j, (i) $\phi_k - \phi_j$ and (ii) ϕ_k/ϕ_j cannot be simplified into a function $g(D_k, D_j, \mathcal{A}, U)$ which is independent of the others' datasets. To illustrate, suppose that the dataset D_i of owner $i \neq k, j$ is updated to include a duplicate of D_j. We expect owner k's relative valuations to increase as j's data is made redundant by D_i. (i) may affect the ranking of owners k and j in interpretable ML while (ii) will alter their compensation in CML [Sim *et al.*, 2020; Xu *et al.*, 2021a]. One possible dependency is that more unique data should be more valuable.[2] Formally, if D_j is similar to others' data and D_k is exclusive to owner k, it is possible for $\phi_k > \phi_j$ even if $U(D_k) \leq U(D_j)$.

D3 **Clone robustness.** If an owner k additionally joins the collaboration as a clone k' (i.e., $D_{k'} = D_k$), then the value assigned to k (and k') should not increase. Formally, let ϕ'_k denote the value of owner k in the clone collaboration setting. We require $\phi_k > \phi'_k + \phi'_{k'}$. This will ensure that the owners cannot unfairly obtain a higher

[1] A possible use case is when the data owners cannot agree on the downstream model/task.

[2] Technical whitepaper: Guide To Data Valuation For Data Sharing (link).

valuation [Ohrimenko *et al.*, 2019].[3]

D4 **Uselessness.** If including D_k does not improve the performance score for any coalition C with aggregated data D_C, then owner k should be valueless: For all $k \in N$, $(\forall C \subseteq N \setminus \{k\}\ \ v(C \cup \{k\}) = v(C)) \Rightarrow \phi_k = 0$.

D5 **Symmetry.** If including D_k yields the same improvement as including D_j in the performance score for any coalition C (e.g., $D_k = D_j$), then their values are equal: For all $k, j \in N$ s.t. $k \neq j$, $(\forall C \subseteq N \setminus \{k, j\}, v(C \cup \{k\}) = v(C \cup \{j\})) \Rightarrow \phi_k = \phi_j$.

D6 **Strict Desirability.**[4] Ensuring fairness may extend beyond symmetry to require owner k to have a higher value than owner j if including D_k yields a greater improvement than including D_j to some coalition, but the reverse is not true. For all $k, j \in N$ s.t. $k \neq j$,

$$(\exists B \subseteq N \setminus \{k, j\}\ \ v(B \cup \{k\}) > v(B \cup \{j\})) \wedge (\forall C \subseteq N \setminus \{k, j\}\ \ v(C \cup \{k\}) \geq v(C \cup \{j\})) \Rightarrow \phi_k > \phi_j.$$

For *weak desirability*, $>$ is replaced by $\geq$. Intuitively, strict and weak desirability enforces whom a data owner k must outvalue or cannot be outvalued.

D7 **Group Rationality.**[5] The value $v(N)$ of the aggregated dataset equals the total value distributed to the N data owners, $\phi_N = \sum_{k \in N} \phi_k$, i.e, $v(N) = \phi_N$. This is important when the absolute value of ϕ_k is used, e.g., $v(N)$ denotes the revenue (from the collaboration) to be distributed among the data owners according to ϕ_k [Richardson *et al.*, 2020]. Note $v(N) > \phi_N$ implies wastage while $v(N) < \phi_N$ is not feasible without external resources. On the other hand, this desideratum may be unnecessary when only the ranking of data points is used (e.g., in interpretable ML applications) or when ϕ_k is rescaled (e.g., in CML applications where owners share a fixed sum).

D8 **Additivity.** The data value computed using the performance metric $U = U_1 + U_2$ is the sum of the data value computed using each performance metric independently. Formally, for all k, $\phi(k, \mathcal{D}, \mathcal{A}, U) = \phi(k, \mathcal{D}, \mathcal{A}, U_1) + \phi(k, \mathcal{D}, \mathcal{A}, U_2)$. As an example, U_1 and U_2 can be the validation accuracy on two separate validation sets stored on separate servers. Thus, decentralized valuation becomes feasible. This desideratum may be unnecessary when centralized valuation is feasible.

D9 **Lipschitzness.** The value of ϕ_k should not change significantly when D_k changes slightly (e.g., slight perturbation to data point) or when $\mathcal{A}$ changes slightly (e.g., change in neural network architecture).

We classify the data valuation strategies ϕ into three categories: *leave-one-out* (LOO), CGT, and desiderata-based according to (the motivation of) the desiderata that ϕ satisfies.

[3]D3 addresses cloning of an owner k and his dataset while P8 addresses replication of data points within any dataset, including D_k.

[4][Maschler and Peleg, 1966].

[5][Jia *et al.*, 2019b]. If $U(\emptyset) = 0$, then this definition is the efficiency assumption in CGT [Chalkiadakis *et al.*, 2011].

3.1 Leave-one-out (LOO) Based

The LOO based strategy evaluates the value of k's dataset D_k as the change in the performance metric of the output model (or dataset) after leaving out D_k from the overall aggregated data D_N. Formally, $\phi_k \triangleq v(N) - v(N \setminus \{k\})$. LOO satisfies D2–D5 and can be adapted to satisfy D1 and D7. To compute the LOO value [Cook and Weisberg, 1980] for every owner $k \in N$, we need to evaluate $v(N \setminus \{k\})$ an additional n times. This is potentially (computationally) expensive due to the training of $\mathcal{A}(D_N \setminus D_k)$ in each $v(N \setminus \{k\})$.

Two fundamental drawbacks of LOO based strategies are that ϕ_k may be inaccurate and the violation of D6, especially when multiple owners have similar datasets. To illustrate, consider the k-nearest neighbours algorithm as $\mathcal{A}$ and the validation accuracy as U. If owner k has a clone j s.t. $D_k = D_j$, their LOO values $\phi_k = \phi_j = 0$ regardless of the validation set as the predictions do not change [Ghorbani and Zou, 2019]. As another example, consider Bayesian linear regression as $\mathcal{A}$ and IG (on the regression parameters) as U. For any owner k, when $|D_N \setminus D_k|$ is large, $\phi_k \approx 0$ as IG is submodular and the increase in IG due to D_k diminishes with more data [Sim *et al.*, 2020]. These drawbacks of LOO prompt the investigation of more principled ways to leave out $m > 1$ owners' datasets from D_N instead, as in CGT based strategies.

3.2 Cooperative Game Theory (CGT) Based

Several data valuation works adopt well-known CGT solution concepts including (but not limited to) *Shapley value* (SV) [Shapley, 1953], least core, and *Banzhaf index* (BI) as ϕ_k. In these solution concepts, ϕ_k explicitly depends on k's *marginal contribution* (MC) to every coalition (subset of owners) $C \subseteq N \setminus \{k\}$ or the value of the coalition $C \cup \{k\}$. The MC of k to C is $\Delta_k^C \triangleq v(C \cup \{k\}) - v(C)$ and the value of a coalition including k is $v(C \cup \{k\})$.

Both SV and BI set $\phi_k = \sum_{C \subseteq N \setminus \{k\}} w_C \Delta_k^C$. In BI, every coalition $C \subseteq N \setminus \{k\}$ has an equal weight w_C while in SV, the weight w_C is the lowest when $|C| = n/2$ and increases as $|C|$ tends to 1 or n. SV is the *unique* solution concept to satisfy D4, D5, D7, and D8 simultaneously. This has led to its popularity in existing works (e.g., [Ghorbani and Zou, 2019; Jia *et al.*, 2019b]). SV and BI assume all owners are committed to forming the grand coalition N for collaboration.

In contrast, *(egalitarian) least core* sets $(\phi_k)_{k \in N}$ to the vector with the least l_2-norm that minimises the greatest deficit between the characteristic function value and the total payoff to any coalition C, i.e, $\max_C v(C) - \sum_{k \in C} \phi_k$ [Chalkiadakis *et al.*, 2011]. It is a group notion of fairness and ensures that each C gets its dues and hence would not deviate from the grand coalition [Yan and Procaccia, 2021].

3.3 Desiderata Based

We will group a few Shapley variants to discuss the desiderata they address, and then describe two other strategies with different motivations.

Data Shapley [Ghorbani and Zou, 2019] values are sensitive to the exact choice of D_N and do not satisfy stability

[6]The techniques that target Δ_k^C can be used for BI and LOO based strategies.

Choice of ϕ	Satisfied desiderata
LOO	D1*, D2–D5, and D7 (after normalization)
Shapley value (SV)	D1*, D2, and D4–D8
Banzhaf index (BI)	D1*, D2–D6, D7 (after normalization), and D8 (without normalization)
Egalitarian least core	D1*, D2, D4, D5, D6 (weak desirability), and D7
Desiderata based	• D9 by $\mathfrak{D}$-Shapley [Ghorbani *et al.*, 2020] • D2, D3, D4, D5, D8 by Robust Shapley [Ohrimenko *et al.*, 2019] • D1*, D2, and D4–D6, D8 by Beta Shapley [Kwon and Zou, 2021] • D1*, D2, D4, and D5 by Variational Value [Bian *et al.*, 2021] • D1 and D2 by data value from DVE [Yoon *et al.*, 2020]

Table 2: Summary of the satisfied desiderata of different valuation strategies ϕ. Note that * means that approximation is required, while D means that the strategy is designed to satisfy the desideratum.

Quantity of interest	Assumptions
Application-agnostic	
Shapley value (SV)	• Bounded v: Permutation Sampling [Maleki *et al.*, 2013] and Group Testing [Jia *et al.*, 2019b] • Monotonic v and sparse $(\phi_k)_{k\in N}$ [Jia *et al.*, 2019b] • Weighted majority game [Fatima *et al.*, 2008]
Least core	• δ-probable least core and (δ, ϵ)-probably approximate least core [Yan and Procaccia, 2021]
Data-valuation specific with Shapley value as ϕ_k [6]	
ϕ_k	• k-NN algorithms [Jia *et al.*, 2019a]
Approximate $\widehat{\phi}_k$	• Stable $\mathcal{A}$ [Jia *et al.*, 2019b]
Δ_k^C	• Incrementally trainable $\mathcal{A}$ [Jia *et al.*, 2019b]
Approximate $\widehat{\Delta}_k^C$ (hence approximate $\widehat{\phi}_k$)	• DNN training can be approximated with only one *epoch*: Gradient Shapley [Ghorbani and Zou, 2019] • Value $\Delta_k^C \approx 0$ when $U_{\mathcal{A}}(D_C) \approx U_{\mathcal{A}}(D_N)$: TMC-Shapley [Ghorbani and Zou, 2019] • Ensemble games [Rozemberczki and Sarkar, 2021] • $\lvert D_k\rvert = 1$, $\mathcal{A}$ outputs a model that minimizes a strictly convex loss with computable gradients and Hessian: Influence functions [Jia *et al.*, 2019b]

Table 3: Summary of cost reduction techniques. Gradient Shapley, TMC-Shapley, and influence function do not have theoretical guarantees on their approximation quality.

(D9). When a data owner is added or removed, all the data Shapley values must be recomputed. To satisfy D9, Ghorbani *et al.* [2020] have proposed Distributional Shapley ($\mathfrak{D}$-Shapley) which depends on the underlying distribution instead of the fixed dataset D_N. $\mathfrak{D}$-Shapley values are stable under slight perturbations to the data points themselves and the underlying data distribution. As similar data will have similar $\mathfrak{D}$-Shapley, Ghorbani *et al.* [2020] have proposed learning a separate regression model to interpolate and predict $\mathfrak{D}$-Shapley for unseen data. As a positive result, the cost of predicting the $\mathfrak{D}$-Shapley for unseen data does not depend on $|D_N|$ or the learning algorithm $\mathcal{A}$ but only on the regression model (D1). Moreover, Ohrimenko *et al.* [2019] have argued that SV does not satisfy D3 and proposed Robust Shapley to satisfy it, but loses D7. Interestingly, Kwon and Zou [2021] have suggested that SV is suboptimal for quantifying the impact of individual datum and proposed losing D7 to assign larger weights for MC to smaller coalitions C.[7] The proposed Beta Shapley is shown to be effective for noisy data identification in interpretable ML applications.

Bian *et al.* [2021] have modelled the probability of any coalition C forming out of the power set 2^N as an energy-based model, and defined ϕ_k as the optimal decoupled probability that an owner k is in the formed coalition C. They have then approximated ϕ_k with the m-step Variational Value which satisfies D2, D4, and D5, but not D7 or D8. Interestingly, they have shown that a one-step gradient ascent for maximizing the mean-field objective can recover SV or BI.

Yoon *et al.* [2020] (DVRL) have used reinforcement learning to automatically learn data values ϕ_k and improve $\mathcal{A}$ during training. They have simultaneously learned a regression model/*data valuation estimator* (DVE) to pick the 'valuable' data to train $\mathcal{A}$ (i.e., trained over randomly sampled batches of data iteratively). Consequently, they have viewed the probabilities that the DVE assigned to k-th datum as its value ϕ_k. DVRL is scalable to large datasets and complex models (D1) since the computational cost (for data valuation) does not directly depend on $|D_N|$ or that of $\mathcal{A}$. It also satisfies D2 but is *not* guaranteed to satisfy the other desiderata (D4–D9). A practical limitation is that it is w.r.t. individual datum (or singleton datasets $|D_k| = 1$) and requires n to be large.

4 Open Challenges

4.1 New Ingredients

In Sec. 2, we have seen how the properties of the data and learning algorithm determine the performance metric and steps in the computation of ϕ. Hence, new data and learning algorithm settings should be examined in future work. A few examples are given as open challenges.

Open Challenges. Firstly, if a data owner k contributes noisy and adversarial data, then we may want him to get a lower performance score and data value. However, this new principle may not be satisfied by the current options for U. It can be practically challenging to obtain a clean, noiseless validation set for P1 based metrics. P6 and P7 based performance metrics may contrarily value noisy and adversarial data highly. Future work may propose new performance metrics U that will satisfy the principle and allow CML applications to remove the assumption for high-quality data used in [Tay *et al.*, 2022; Xu *et al.*, 2021b].

Next, data owners (e.g., patients) may require privacy guarantees on their data. If weak guarantees are acceptable, then future work can consider using synthetic data [Tay *et al.*, 2022]. If *differential privacy* (DP) is needed, then there are a few issues and questions to consider. Firstly, the need to compute the value of multiple coalitions $C \subset N$ in ϕ may increase the privacy leakage. This is observed when we directly use DP gradient descent [Abadi *et al.*, 2016] in each training. Next, as DP would reduce the sensitivity of a model's output to data, how would DP affect the MC and its estimation? Finally, what DP learning algorithm $\mathcal{A}_{\text{p}}$ and other performance metrics U can provably guarantee the principle that data/model become less valuable when the privacy requirements are higher?

Future work can also propose performance metrics U for existing principles. For example, for P6 similarity-based metrics, the similarity of model parameters or predictions to a ref-

[7] Beta Shapley satisfies all the other desiderata of Data Shapley if the weight of every MC is positive.

erence model (i.e., trained on everyone's data) can be used. Furthermore, other unexplored learning algorithms $\mathcal{A}$ (e.g., multi-agent reinforcement learning and meta-learning) may need new performance metrics U and principles.

4.2 New Data Valuation Strategies, ϕ

In Sec. 3, we suggest that in data valuation applications, D7 and D8 may be unnecessary, and their removal will make more strategies satisfactory.

Open challenges. For CGT based strategies, future work should further explore the research direction of Kwon and Zou [2021] to generalize and unify existing strategies, including SV and BI. To guide data valuation users on the right ϕ to use, future work can outline the differences, advantages, and disadvantages of using different CGT solution concepts and the impact on the relative ϕ_k values of owners.

Moreover, for lower computational cost (D1) and to satisfy other desiderata, future work can explore the research directions of Ghorbani *et al.* [2020], Wang *et al.* [2021], and Yoon *et al.* [2020] and outline how to learn and predict ϕ on datasets with more than one datum.

4.3 New Cost Reduction Techniques

Achieving low computational cost (satisfying D1) is a practical challenge shared by both the LOO and CGT based strategies. We need to compute $\Delta_k^{N\setminus\{k\}}$ or $v(N \setminus \{k\})$ (Δ_k^C or $v(C)$) for n and 2^n times in total to compute ϕ_k for LOO and CGT based strategies, respectively. We will first outline existing techniques to reduce the computational cost (i.e., summarized in Table 3) and follow up with some open challenges.

Application-agnostic approximations can be used beyond data valuation and often come with theoretical guarantees. SV approximations can reduce the cost of computing ϕ_k from exponential to polynomial time in n [Fatima *et al.*, 2008; Jia *et al.*, 2019b; Maleki *et al.*, 2013; Mitchell *et al.*, 2022]. Least core relaxations can be computed with (sub)linear time in n [Yan and Procaccia, 2021]. However, these improvements may be insufficient due to the costly training of $\mathcal{A}$ such as training DNN on large datasets. Hence, this motivates the need for *data valuation-specific techniques* to exploit properties of $\mathcal{A}$ and U. The techniques may target two different quantities of interest: (i) the value ϕ_k of owner k, or (ii) his marginal contribution Δ_k^C to any coalition C.

For (i), Jia *et al.* [2019a] have shown that for k-NN learning algorithms, every owner k's SV can be computed exactly in polynomial time. Separately, Jia *et al.* [2019b] have proven that for uniformly stable learning algorithms, every datum's SV can be approximated by $v(N)/n$ (with theoretical guarantee) and thus, no retraining is needed. Differently, Ghorbani and Zou [2019] have proven that for *Lipschitz-stable* $U_{\mathcal{A}}$, any datum's $\mathfrak{D}$-Shapley value [Ghorbani *et al.*, 2020] can be estimated using a regression model trained only on a fraction of data and their computed $\mathfrak{D}$-Shapley.[8] Fatima *et al.* [2008] and Rozemberczki and Sarkar [2021] have specifically chosen $v(\cdot)$ to give rise to *weighted voting games* where both ϕ_k and Δ_k^C can be approximated efficiently with theoretical error guarantees. Specifically, they define MC Δ_k^C to be 1 iff k is the *swing player* (i.e., C makes an incorrect classification but $C \cup \{k\}$ makes a correct classification), and 0 otherwise.

On the other hand, for (ii), Jia *et al.* [2019a] have suggested that for *incrementally trainable* $\mathcal{A}$,[9] we can exactly compute an MC for each of the n owners in the time taken for one evaluation of $\mathcal{A}$. Thus, we can shave off a linear factor of n further from the permutation sampling approximation. TMC-Shapley [Ghorbani and Zou, 2019] eliminates the cost of retraining $\mathcal{A}$ in the later steps of every permutation in permutation sampling by approximating MC Δ_k^C with 0 when the performance metric $v(S)$ evaluated on data from $S \subseteq C$ is close to $v(N)$. Jia *et al.* [2019b] have suggested using the influence function heuristic [Koh and Liang, 2017] to approximate Δ_k^C for singleton datasets (i.e., $|D_k| = 1$) and save a factor of n from the avoided retraining. Lastly, Gradient Shapley [Ghorbani and Zou, 2019] approximates the model trained on multiple passes of the training data with another trained on a single pass and different hyperparameters. This replaces $\mathcal{A}$ with another incrementally trainable $\mathcal{A}'$.

Open Challenges. The challenges focus on data valuation-specific techniques. For (i), by studying CGT literature or analysing $U_{\mathcal{A}}$, future work can identify new properties of $\mathcal{A}$ and U that will enable efficient approximation of ϕ_k with theoretical error guarantees. Existing solutions for (ii) may be unsuitable for some learning algorithms and models (e.g., complex DNNs) as they are not incrementally trainable and do not have convex loss functions. Basu *et al.* [2021] have pointed out that the influence function heuristics deteriorate significantly for complex DNNs. Also, Koh *et al.* [2019] have suggested that influence functions are less accurate at measuring group effects, such as of D_k when its size is large. Moreover, TMC-Shapley and Gradient Shapley do not come with theoretical error guarantees. Hence, improvements and new techniques should be proposed, and their approximation errors should be studied theoretically and empirically.

5 Conclusion

This paper presents a technical survey to provide a formal study of data valuation in ML through its ingredients and the corresponding properties (e.g., performance metric and its design principles). We ground the discussion of common desiderata satisfied by existing data valuation strategies on our proposed ingredients and group existing works into LOO, CGT, and desiderata based strategies. Moreover, we outline open research challenges in data valuation. An interesting direction is to use ML to define ϕ and learn and predict the value of data. This survey serves as a technical guide for researchers in this field and a guideline for practitioners of data valuation.

[8] $\mathfrak{D}$-Shapley satisfies stability (D9) due to the formulation of ϕ and may hold for a non-stable learning algorithm $\mathcal{A}$.

[9] $\mathcal{A}$ is incrementally trainable if we can obtain $\mathcal{A}(D \cup D_k)$ from $\mathcal{A}(D)$ with a significantly lower computation cost than retraining from scratch. An example is logistic regression optimized with gradient descent over one *epoch* only [Ghorbani and Zou, 2019].

Ethical Statement

This technical survey primarily focuses on the algorithmic aspect of data valuation in machine learning and not the regulatory considerations governing the usage of data (e.g., data privacy issues). Hence, practitioners and researchers should keep in mind and closely observe such regulations since the protection of data privacy (especially personal data) is crucial. Furthermore, as briefly described in Sec. 4.1, one possibility is to adopt *differential privacy* to provide rigorous privacy guarantees. Therefore, future works are encouraged to incorporate formal treatments of data privacy (not limited to differential privacy) into valuation to ensure the proposed methods can be adopted in practice while observing the regulations.

Acknowledgments

This research is supported by the National Research Foundation, Singapore under its Strategic Capability Research Centres Funding Initiative. Any opinions, findings and conclusions, or recommendations expressed in this material are those of the author(s) and do not reflect the views of the National Research Foundation, Singapore.

References

[Abadi *et al.*, 2016] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proc. ACM CCS*, 2016.

[Abay *et al.*, 2019] Nazmiye Ceren Abay, Yan Zhou, Murat Kantarcioglu, Bhavani Thuraisingham, and Latanya Sweeney. Privacy preserving synthetic data release using deep learning. In *Machine Learning and Knowledge Discovery in Databases*, pages 510–526, 2019.

[Agarwal *et al.*, 2019] Anish Agarwal, Munther Dahleh, and Tuhin Sarkar. A marketplace for data: An algorithmic solution. In *Proc. EC*, pages 701–726, 2019.

[Basu *et al.*, 2021] Samyadeep Basu, Philip Pope, and Soheil Feizi. Influence functions in deep learning are fragile. In *Proc. ICLR*, 2021.

[Bax, 2019] Eric Bax. Computing a data dividend. arXiv:1905.01805, 2019.

[Bian *et al.*, 2021] Yatao Bian, Yu Rong, Tingyang Xu, Jiaxiang Wu, Andreas Krause, and Junzhou Huang. Energy-based learning for cooperative games, with applications to valuation problems in machine learning. arXiv:2106.02938, 2021.

[Chalkiadakis *et al.*, 2011] Georgios Chalkiadakis, Edith Elkind, and Michael Wooldridge. Computational aspects of cooperative game theory. In R. J. Brachman, W. W. Cohen, and T. G. Dietterich, editors, *Synthesis Lectures on Artificial Intelligence and Machine Learning*. Morgan & Claypool Publishers, 2011.

[Cong *et al.*, 2021] Zicun Cong, Xuan Luo, Pei Jian, Feida Zhu, and Yong Zhang. Data pricing in machine learning pipelines. arXiv:2108.07915, 2021.

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

[Devereaux *et al.*, 2016] P.J. Devereaux, Gordon Guyatt, Hertzel Gerstein, Stuart Connolly, and Salim Yusuf. Toward fairness in data sharing. *New England Journal of Medicine*, 375(5):405–407, 2016.

[Fatima *et al.*, 2008] Shaheen S Fatima, Michael Wooldridge, and Nicholas R Jennings. A linear approximation method for the Shapley value. *Artificial Intelligence*, 172(14):1673–1699, 2008.

[Fricker and Maksimov, 2017] Samuel A Fricker and Yuliyan V Maksimov. Pricing of data products in data marketplaces. In *Proc. ICSOB*, pages 49–66, 2017.

[Ghorbani and Zou, 2019] Amirata Ghorbani and James Zou. Data Shapley: Equitable valuation of data for machine learning. In *Proc. ICML*, pages 2242–2251, 2019.

[Ghorbani *et al.*, 2020] Amirata Ghorbani, Michael Kim, and James Zou. A distributional framework for data valuation. In *Proc. ICML*, 2020.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Proc. NeurIPS*, volume 27, 2014.

[Heckman *et al.*, 2015] Judd Randolph Heckman, Erin Laurel Boehmer, Elizabeth Hope Peters, Milad Davaloo, and Nikhil Gopinath Kurup. A pricing model for data markets. In *Proc. iConference*, 2015.

[Jia *et al.*, 2019a] Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nezihe Merve Gurel, Bo Li, Ce Zhang, Costas J Spanos, and Dawn Song. Efficient task-specific data valuation for nearest neighbor algorithms. *Proc. VLDB Endowment*, 12(11):1610–1623, 2019.

[Jia *et al.*, 2019b] Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J Spanos. Towards efficient data valuation based on the Shapley value. In *Proc. AISTATS*, pages 1167–1176, 2019.

[Jia *et al.*, 2019c] Ruoxi Jia, Xuehui Sun, Jiacen Xu, Ce Zhang, Bo Li, and Dawn Song. An empirical and comparative analysis of data valuation with scalable algorithms. arXiv:1911.07128, 2019.

[Kingma and Welling, 2014] Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In *Proc. ICLR*, 2014.

[Koenecke and Varian, 2020] Allison Koenecke and Hal Varian. Synthetic data generation for economists. arXiv:2011.01374, 2020.

[Koh and Liang, 2017] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proc. ICML*, pages 1885–1894, 2017.

[Koh *et al.*, 2019] Pang Wei Koh, Kai-Siang Ang, Hubert Teo, and Percy S Liang. On the accuracy of influence functions for measuring group effects. In *Proc. NeurIPS*, volume 32, 2019.

[Krause and Guestrin, 2007] Andreas Krause and Carlos Guestrin. Nonmyopic active learning of Gaussian processes: an exploration-exploitation approach. In *Proc. ICML*, pages 449–456, 2007.

[Kruschke, 2008] John Kruschke. Bayesian approaches to associative learning: From passive to active learning. *Learning & Behavior*, 36(3):210–226, 2008.

[Kwon and Zou, 2021] Yongchan Kwon and James Zou. Beta Shapley: a unified and noise-reduced data valuation framework for machine learning. In *Proc. AISTATS*, 2021.

[Liang *et al.*, 2018] Fan Liang, Wei Yu, Dou An, Qingyu Yang, Xinwen Fu, and Wei Zhao. A survey on big data market: Pricing, trading and protection. *IEEE Access*, 6:15132–15154, 2018.

[Maleki *et al.*, 2013] Sasan Maleki, Long Tran-Thanh, Greg Hines, Talal Rahwan, and Alex Rogers. Bounding the estimation error of sampling-based Shapley value approximation. arXiv:1306.4265, 2013.

[Maschler and Peleg, 1966] Michael Maschler and Bezalel Peleg. A characterization, existence proof and dimension bounds for the kernel of a game. *Pacific J. Mathematics*, 18(2):289–328, 1966.

[Mitchell *et al.*, 2022] Rory Mitchell, Joshua Cooper, Eibe Frank, and Geoffrey Holmes. Sampling permutations for shapley value estimation. *Journal of Machine Learning Research*, 23(43):1–46, 2022.

[Muschalle *et al.*, 2013] Alexander Muschalle, Florian Stahl, Alexander Löser, and Gottfried Vossen. Pricing approaches for data markets. In Malu Castellanos, Umeshwar Dayal, and Elke A. Rundensteiner, editors, *Enabling Real-Time Business Intelligence*, pages 129–144, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[Ohrimenko *et al.*, 2019] Olga Ohrimenko, Shruti Tople, and Sebastian Tschiatschek. Collaborative machine learning markets with data-replication-robust payments. In *Proc. NeurIPS SGO & ML Workshop*, 2019.

[Pei, 2020] Jian Pei. A survey on data pricing: from economics to data science. *IEEE Transactions on Knowledge & Data Engineering*, 2020.

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

[Richardson *et al.*, 2020] Adam Richardson, Aris Filos-Ratsikas, and Boi Faltings. Budget-bounded incentives for Federated learning. In *Federated Learning*, pages 176–188. Springer, 2020.

[Rozemberczki and Sarkar, 2021] Benedek Rozemberczki and Rik Sarkar. The Shapley value of classifiers in ensemble games. In *ACM Symposium on Neural Gaze Detection*, 2021.

[Shapley, 1953] Lloyd Shapley. A value for n-person games. In H. W. Kuhn and A. W. Tucker, editors, *Contributions to the Theory of Games*, volume 2, pages 307–317. Princeton Univ. Press, 1953.

[Sim *et al.*, 2020] Rachael Hwee Ling Sim, Yehong Zhang, Mun Choon Chan, and Bryan Kian Hsiang Low. Collaborative machine learning with incentive-aware model rewards. In *Proc. ICML*, pages 8927–8936, 2020.

[Singhal *et al.*, 2021] Karan Singhal, Hakim Sidahmed, Zachary Garrett, Shanshan Wu, John Rush, and Sushant Prakash. Federated reconstruction: Partially local federated learning. In *Proc. NeurIPS*, 2021.

[Song *et al.*, 2019] Tianshu Song, Yongxin Tong, and Shuyue Wei. Profit allocation for Federated learning. In *Proc. IEEE Big Data*, pages 2577–2586. IEEE, 2019.

[Stahl and Vossen, 2016] Florian Stahl and Gottfried Vossen. Data quality scores for pricing on data marketplaces. In N.T. Nguyen, G. Trawiński, H. Fujita, and T-P. Hong, editors, *Intelligent Information and Database Systems*, pages 215–224, Berlin, Heidelberg, 2016. Springer Berlin Heidelberg.

[Tay *et al.*, 2022] Sebastian Shenghong Tay, Xinyi Xu, Chuan Sheng Foo, and Bryan Kian Hsiang Low. Incentivizing collaboration in machine learning via synthetic data rewards. In *Proc. AAAI*, 2022.

[Wang *et al.*, 2021] Tianhao Wang, Yu Yang, and Ruoxi Jia. Learnability of learning performance and its application to data valuation. arXiv:2107.06336, 2021.

[Xu *et al.*, 2021a] Xinyi Xu, Lingjuan Lyu, Xingjun Ma, Chenglin Miao, Chuan Sheng Foo, and Bryan Kian Hsiang Low. Gradient driven rewards to guarantee fairness in collaborative machine learning. In *Proc. NeurIPS*, volume 34, 2021.

[Xu *et al.*, 2021b] Xinyi Xu, Zhaoxuan Wu, Chuan Sheng Foo, and Bryan Kian Hsiang Low. Validation free and replication robust volume-based data valuation. In *Proc. NeurIPS*, volume 34, 2021.

[Yan and Procaccia, 2021] Tom Yan and Ariel D Procaccia. If you like Shapley then you'll love the core. In *Proc. AAAI*, pages 5751–5759, 2021.

[Yoon *et al.*, 2020] Jinsung Yoon, Sercan Arik, and Tomas Pfister. Data valuation using reinforcement learning. In *Proc. ICML*, 2020.

[Zhang and Beltrán, 2020] Mengxiao Zhang and Fernando Beltrán. A survey of data pricing methods. SSRN:3609120, 2020.

[Zhao *et al.*, 2020] Yi Zhao, Ke Xu, Feng Yan, Yuchao Zhang, Yao Fu, and Haiyang Wang. Auction-based High Timeliness Data Pricing under Mobile and Wireless Networks. *IEEE International Conference on Communications*, 2020.