

# An Overview of Data Contribution Evaluation Methods for Federated Learning

Chenxi Ding, Mingchuan Zhang

Henan University of Science and Technology, Luoyang 471000, China

---

**Abstract:** With the rapid development of artificial intelligence, especially machine learning technology, the demand for data is also increasing. As a distributed learning method, federated learning (FL) allows total multiple parties to co-train models locally while protecting data privacy. Due to heterogeneity in data, devices, and other factors, the contribution of participants to model quality also varies. If these differences are ignored, it poses a challenge to achieving a fair FL environment. Therefore, fairly assessing the contribution levels of each data provider becomes a critical issue in FL. This plays a significant role in fairly motivating participants and promoting the sustainable development of FL. This paper reviews existing evaluation methods from the perspective of data contribution assessment in FL, discusses ongoing challenges in this area, and explores future research directions.

**Keywords:** Data Evaluation; Fair Incentive; Federated Learning; Machine Learning.

---

## 1. Introduction

In recent years, the rapid advancement of artificial intelligence [1-2] has propelled the data economy to new heights. Among them, the continuous stream of data resources serves as the driving force behind this technological progress. However, from the perspective of data security and privacy protection, countries around the world have introduced various policies and regulations to regulate and limit the use of data [3]. These policies have imposed stricter controls on sensitive data within certain institutions or organizations, such as banking secrets and medical patient information, preventing the centralized collection and processing of data. On the one hand, these restrictions safeguard individual privacy and enhance the data security. On the other hand, they significantly disrupt traditional centralized learning methods, especially in scenarios requiring cross-institutional sharing or processing of large-scale data. This has led to the phenomenon known as "data island," [4] which in turn hinders the realization of the potential value of the data and the contribution of knowledge.

In response to this challenge, while safeguarding data privacy and considering the interests of all parties involved, FL [5-6] has been proposed. This approach allows multiple participants to collaboratively train a high-quality machine learning model without sharing their raw data directly. Meanwhile, they train models locally and exchange only model updates, such as parameters and gradients, to achieve this collaboration. This approach not only effectively prevents the direct exchange of sensitive information and reduces the risk of data leakage, but also enables data collaboration across organizational boundaries. While safeguarding individual privacy and data security, FL maximizes the potential of data held by various stakeholders. It has been widely applied in areas such as edge computing [7], healthcare [8], and smart home [9].

The success of FL relies heavily on the active participation of participants in collaboration using high-quality data. Given

the variation in data quality, computational capabilities, and resources among participants, their contributions to the improvement of model performance are also different. Obviously, if the server overlooks the aforementioned facts, it will inevitably lead to unfairness, thereby diminishing the enthusiasm for participation in collaboration [10-11]. In this scenario, fairly evaluating the data contribution of each participant and then giving them timely incentives is a key issue in FL research.

Additionally, the integrity of the data set may be subject to various restrictions, such as confidentiality concerns and interests of different parties, which making it challenging for data owners to provide complete information [12]. In settings where data lacks completeness, the existing data contribution assessment methods might not be applicable for cases with absent attributes or constrained datasets. Hence, developing customized evaluation schemes for data contributions, suited to varying data scenarios, has emerged as a pivotal issue in the study of FL. This approach guarantees that the input and endeavors of all contributors receive a just evaluation and appropriate rewards, motivating a broader array of data proprietors to engage and contribute actively to the collective endeavors of FL.

## 2. Background Knowledge

### 2.1. Federated Learning

Depending on the distinct characteristics and distributions of data, FL can be categorized into three types: horizontal federated learning (HFL), vertical federated learning (VFL), and federated transfer learning (FTL) [13]. In HFL, different parties have datasets that share the same features but originate from different sources. VFL indicates that the participants' data sources are the same but have different feature spaces, and it focuses on collaboration through the lens of feature integration. FTL addresses situations where there's little commonality in the data features or sources among the participants. The basic process of FL as shown in Fig. 1.

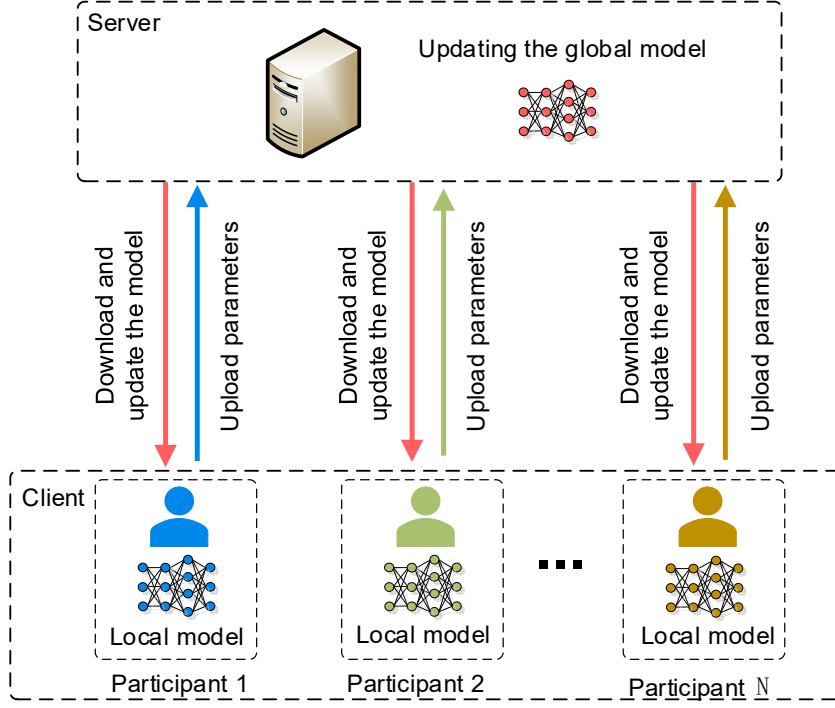


Fig 1. The framework of federated learning with  $N$  participants

The FL framework mainly consists of a central server and many participants. Taking HFL as an example, this optimization problem can be expressed as follows:

$$\min F(w) := \sum_{i=1}^N p_i F_i(w) \quad (1)$$

where  $w$  represents the model parameters,  $N$  denotes the total number of participants,  $F(\cdot)$  stands for the global objective function, and  $p_i$  is the weight of each participant in the  $t^{\text{th}}$  round, with  $p_i \geq 0$ .

$$F_i(w) := \ell(w, D_i) \quad (2)$$

where  $\ell$  is loss function. To optimize the problem in equation (1), McMahan et al. proposed FedAvg algorithm [5]. Specifically, in each round, participants primarily engage in local model training and updating. Here, participants perform local updates using stochastic gradient descent (SGD) [14], resulting in the following update formula:

$$w_i^{t+1} = w_i^t - \eta^t \nabla F_i(w_i^t) \quad (3)$$

where  $\eta^t$  is the learning rate.

On the server side, it mainly collects and updates the model parameters. For example, FedAvg collects model information based on weighted aggregation of the amount of data of each party, then

$$w^{t+1} = \sum_{i=1}^N \frac{|D_i|}{|D_N|} w_i^{t+1} \quad (4)$$

where  $|D_i|$  represents the dataset size of the participant and  $|D_N|$  is the total data volume.

## 2.2. Contribution Evaluation Metrics

The motivation of data owners in FL collaborations hinges on the fairness of the cooperation, making it crucial to fairly and effectively assess the contribution of data. Viewing from the lens of data privacy and access, datasets can be categorized into complete and incomplete types, with the latter primarily referring to situations where test or validation sets are missing. Depending on the completeness of datasets in federated settings, methods for assessing contributions can be divided into those that rely on complete datasets and those adapted for scenarios with missing datasets.

### 2.2.1. Evaluation Metrics under the Complete Dataset

In general, FL tasks have complete dataset information by default. The most common way to evaluate data contribution metrics is to test the accuracy of the trained model based on it [15]. For example, in image classification tasks, it is common to measure the contribution of the data in terms of the error value. The higher model accuracy indicates greater value and contribution of a data point. The data contribution estimation methods described in subsequent chapters are applicable to the full dataset. Additionally, the credibility of the data, as well as the quality and quantity of the data used for training, can also serve as metrics for evaluating the contribution of data to FL. In practical scenarios, it is necessary to comprehensively analyze the value of data from multiple perspectives.

### 2.2.2. Evaluation Metrics under the Incomplete Dataset

Given considerations such as data privacy protection, legal regulations, and constraints of the learning task, it becomes challenging for a consortium to access complete dataset information. Therefore, it's essential to further investigate methods for assessing data contributions under conditions of incomplete dataset information. As illustrated below, this paper outlines several common evaluation metrics.

- **Statistical Methods.** This approach evaluates the value of data using various statistical methods, such as those based on data distribution. For instance, the gram determinant [16] can quantify the volume of data in Euclidean space; The smaller the volume, the more concentrated the data distribution, indicating a relatively lower contribution rate. This can be formally expressed as

$$U(S) = (R_S^T R_S)^{\frac{1}{2}} \quad (5)$$

where  $R_S$  represents the aligned subset of the data. For example, Xu et al. [17] proposed a volumetric metric to evaluate the value of individual data points based on the above strategy.

- **Model Similarity.** In an alliance, the value of data can be measured by the similarity of the trained data model to the global model. For example, cosine similarity uses the cosine

value between the model gradient updates from a local data combination and the global gradient updates to measure the contribution of that combination. Among them, Jiang et al. [18] constructed a data space and a gradient space, and introduced cosine similarity to measure the contribution of each participant.

### 3. Data Contribution Evaluation Methodology

Due to the actual situation, the subjectivity of participants in federated settings, among other factors, the methods for evaluating contribution are often complex and costly. Therefore, developing a sound and reliable methodology is crucial for determining the value of data. This chapter will discuss several prominent evaluation approaches.

#### 3.1. Individual Method

From an individual perspective, the impact of a single participant's data on model accuracy can be assessed using a value function, thereby defining its contribution value [19]. Moreover, the metrics for evaluating individual value can take various forms. For example, the individual reputation method [20] assesses a participant's value based on their historical activity level, reflecting their credibility. Meanwhile, the mutual information method [21], based on feature selection, indicates that a higher mutual information value suggests a stronger dependency between features and labels, thus implying a higher credibility of the participant's data information. From an individual perspective, Lv et al. [22] proposed the pairwise correlated agreement (PCA) method, which uses the statistical correlation to evaluate their contributions.

Additionally, to evaluate the value of individual data points, can introduce influence functions [23] that measure the degree to which a sample impacts model parameters. For instance, [24] have designed payment schemes based on the impact of data points on the model's loss function, aiming to incentivize participants to contribute high-quality data for training purposes. Shyn et al. [25] proposed FedCCEA and created an accuracy approximation model (AAM). Estimate simulation test accuracy by sample data size to evaluate the quality and contribution of customer data. Without the assistance of the validation set, Xu et al. [26] proposed a cosine gradient sharpley value (CGSV) method, which evaluates the contributions of the participants through the model information uploaded by them.

#### 3.2. Leave-One-Out (LOO)

As a cross-validation technique, LOO [27] was introduced early to assess the value of data. This approach takes the model's accuracy of portfolio value of a participant after removing the participant to be assessed as the participant's contribution to the alliance, which can be expressed as

$$U_i = \Gamma(N) - \Gamma(N \setminus \{i\}) \quad (6)$$

where  $\Gamma$  is the utility function. Both the Individual method and LOO are simple and efficient, which are often used as benchmark indicators for evaluation results. However, LOO also suffers from high computational cost, and in [28], Koh et al. reduced the computational complexity by approximating the influence function.

#### 3.3. Shapley Value (SV)

The SV [29] is primarily utilized in game theory to address

the issue of distributing benefits in multi-party cooperative scenarios. It quantifies the marginal contribution of each participant to the collective alliance by assigning their respective SVs, making it a critical tool in research across various fields. Having a utility function  $\Gamma(\cdot)$ , the SV for participant  $k$  can be interpreted as their marginal contribution to  $\Gamma(S)$ , where  $S$  is a subset of participants from  $N: S \subseteq \{1, \dots, N\}$ . Mathematically, the  $U_i(N, \Gamma)$  for a participant  $i$  within  $N$  is expressed as

$$U_i(N, \Gamma) = E_{\pi \in \Pi(D)} [\Gamma(S_i^\pi \cup \{i\}) - \Gamma(S_i^\pi)] \quad (7)$$

where  $\pi \in \Pi(D)$  represents a certain ordering of all participants, and  $S_i^\pi$  refers to the combination of participants that are before participant  $i$  in the sequence  $\pi$ . In practical applications, utility functions  $\Gamma(\cdot)$  can be created using evaluation metrics such as test accuracy and model loss. The SV of participant  $i$  can also be simplified as:

$$U_i(N, \Gamma) = \frac{1}{N} \sum_{S \subseteq N \setminus i} \frac{(\Gamma(S \cup i) - \Gamma(S))}{\binom{|N|-1}{|S|}} \quad (8)$$

where  $\frac{1}{N}$  represents the probability of  $i$  being in any position within the sequence. Moreover, the SV provides solutions for various cooperative scenarios. Among them, Sim et al. [30] measure the contributions of each participant based on the SV and information gain, thereby allocating corresponding model rewards to them. From perspectives of fairness and efficiency, Fan et al. [31-32] explore the issue of evaluating participant contributions in both HFL and VFL.

Although SV provides a comprehensive and reasonable way, it is also plagued by high computational costs [33]. Recently, there has been significant research aimed at enhancing computational efficiency. Specifically, considering the high cost of communication and the impact of participant order on contribution assessment, Wang et al. [34] proposed the federated SV, effectively reflecting the true value of data sources without incurring additional costs. Kwon et al. [35] introduced a distributed SV framework and developed analytical expressions in contexts such as linear regression and binary classification, greatly improving computational efficiency.

#### 3.4. Least Core

Also as a scheme for the distribution of benefits in game theory, the nuclear method has attracted attention. Yan et al. [36] have noted that the SV, when assessing contributions, exhibits limitations in terms of fairness. They introduced the least core method to evaluate the contributions of participating entities in various combinations. This evaluation scheme can be formalized as

$$\min c = \begin{cases} \sum_{i \in N} U_i = \Gamma(N) \\ \sum_{i \in S} U_i + c \geq \Gamma(S), \forall S \subseteq N \end{cases} \quad (9)$$

where  $c$  represents the cost function. However, this method may lead to non-unique evaluation results, so its effectiveness still needs to be verified in subsequent work.

#### 3.5. Optimal Transport

In 1781, French mathematician Gaspard Monge introduced the concept of optimal transport [37], also known as the Monge problem. The problem primarily seeks the method of relocating a quantity of sand of a certain shape to another specified shape at the minimum cost and the Fig. 2 illustrates this problem.

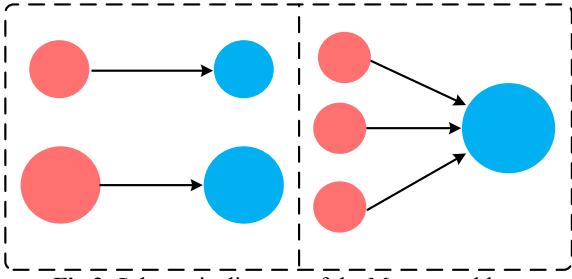


Fig 2. Schematic diagram of the Monge problem

Due to the constraint of the forward measure, the original cargo mass cannot be split and assigned to different destinations during transportation. Therefore, a feasible or optimal solution does not necessarily exist for the Monge problem. In response to this problem, Kantorovich relaxed the restrictions and raised the Kantorovich problem [37]. Specifically, given  $P \in \mathcal{M}(X)$ ,  $Q \in \mathcal{M}(Y)$ , and a cost function  $c: X \times Y \rightarrow \mathbb{R}^+$ , the formulation can be expressed as follows

$$\inf\{K(\gamma) := \int_{X \times Y} c(x, y) d\gamma(x, y) : \gamma \in \Pi(P, Q)\} \quad (10)$$

where the joint probability measure  $\gamma$  is the space of transportation schemes, then

$$\Pi(P, Q) := \{\gamma \in \mathcal{M}(X \times Y) : (\pi_x)_\# \gamma = P, (\pi_y)_\# \gamma = Q\} \quad (11)$$

This problem introduces the idea of mass splitting, which allows data points to be split and transported to different places, and also became the definition in common use today. The schematic of this problem is shown in Fig.3.

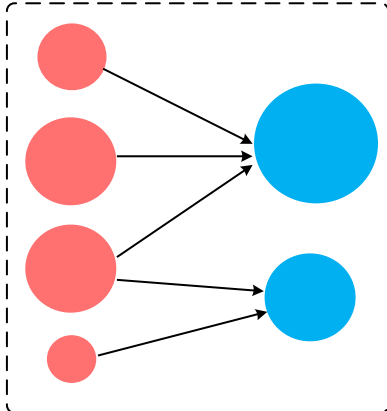


Fig 3. Schematic diagram of the Kantorovich problem

In FL setting, suppose the distribution of data point  $i$  is  $p_i$ , and let  $P, Q, \nu$  be defined in the probability distribution space  $\mathcal{M}(X)$ . The Wasserstein distance satisfies

$$W_p(P, Q) \leq W_p(P, \nu) + W_p(\nu, Q) \quad (12)$$

This indicates that in the constructed probabilistic metric space, there exists a geodesic such that any point on this path can serve as an interpolation for measuring the distance between the two endpoints [38]. In the FL framework, by exploring the relationship between the Wasserstein distance and the federated model, the interpolation metric can be introduced to construct the connection between the data point and the server, which can be translated into the evaluation of the contribution value that data points bring to the alliance.

Without relying on learning algorithms, Just et al. [39] innovatively employed the category Wasserstein distance from optimal transport theory as a proxy, effectively utilizing this distance to identify valuable data points. This paper [40] have successfully integrated the Wasserstein distance into FL

framework, laying the groundwork for subsequent applications of this technique in assessing the data contributions of FL participants.

## 4. Summary and Future Outlook

### 4.1. Summary

This article delves into how to fairly and accurately assess the value of data involved in training within FL. From the perspective of data information integrity, it comprehensively examines the advancements in evaluating data contributions in FL and provides an in-depth introduction and analysis of mainstream evaluation schemes. Moreover, this paper offers clear insights into understanding contribution assessment issues in various scenarios.

### 4.2. Future Outlook

While significant progress has been made in existing research, there remain areas for improvement in effective value measurement, reasonable evaluation schemes, and the efficacy of incentive mechanisms. The following analysis addresses potential directions for future research.

- Most of the existing studies communicate under the rational nature of both parties. However, in practice, there can be issues like malicious users and information tampering during transmission, presenting security risks. Future work should consider optimizing communication transmissions, establishing enhanced privacy protection frameworks, and developing proactive defense technologies to adapt to more complex distributed scenarios.
- FL also deals with a significant amount of dynamic data, which may lead to catastrophic forgetting phenomena and a lack of sufficient information for accurate model parameter estimation, challenging the assessment of contributions. Future research should focus on developing evaluation strategies that are adaptable to dynamic data environments.

## Acknowledgments

This work was supported in part by the Key Technologies R & D Program of Henan Province under Grant No. 22210 2210049, 232102211008 and 242102210102.

## References

- [1] Huynh-The T, Pham Q V, Pham X Q, et al. Artificial intelligence for the metaverse: A survey[J]. Engineering Applications of Artificial Intelligence, 2023, vol. 117(105581).
- [2] Nilsson N J. The quest for artificial intelligence[M]. Cambridge, UK, Cambridge University Press, 2009.
- [3] Dilmaghani S, Brust M R, Danoy G, et al. Privacy and security of big data in AI systems: A research and standards perspective[C]. 2019 IEEE international conference on big data. Los Angeles, CA, USA, 2019, pp. 5737-5743.
- [4] Yang Q, Liu Y, Chen T, et al. Federated machine learning: Concept and applications[J]. ACM Transactions on Intelligent Systems and Technology, 2019, vol. 10(2), pp. 1-19.
- [5] McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data[C]. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, USA, 2017, vol. 54, pp. 1273-1282.
- [6] Kairouz P, McMahan H B, Avent B, et al. Advances and open problems in federated learning[J]. Foundations and trends® in machine learning, 2021, vol. 14(1-2), pp. 1-210.

- [7] Duan Q, Huang J, Hu S, et al. Combining federated learning and edge computing toward ubiquitous intelligence in 6G network: Challenges, recent advances, and future directions[J]. *IEEE Communications Surveys & Tutorials*, 2023.
- [8] Rahman A, Hossain M S, Muhammad G, et al. Federated learning-based AI approaches in smart healthcare: concepts, taxonomies, challenges and open issues[J]. *Cluster computing*, 2023, vol. 26(4), pp. 2271-2311.
- [9] Nguyen D C, Pham Q V, Pathirana P N, et al. Federated learning for smart healthcare: A survey[J]. *ACM Computing Surveys (Csur)*, 2022, vol. 55(3), pp. 1-37.
- [10] T. Li, M. Sanjabi, and V. Smith, Fair resource allocation in federated learning[C]. In *Proceedings of the 8th International Conference on Learning Representations*. Addis Ababa, Ethiopia, 2020.
- [11] Ezzeldin Y H, Yan S, He C, et al. Fairfed: Enabling group fairness in federated learning[C]. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. Washington, DC, USA, 2023, vol. 37(6), pp. 494-7502.
- [12] Sim R H L, Xu X, Low B K H. Data Valuation in Machine Learning: "Ingredients", Strategies, and Open Challenges[C]. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence*. Vienna, Austria, 2022, pp. 5607-5614.
- [13] Zhang T, Mao S. An introduction to the federated learning standard [J]. *GetMobile: Mobile Computing and Communications*, 2022, vol. 25(3), pp.18-22.
- [14] Bottou L. Large-scale machine learning with stochastic gradient descent[C]. In *Proceedings of the 19th International Conference on Computational Statistics*. Paris, France, 2010, pp. 177-186.
- [15] Chen Y, Yang X, Qin X, et al. Dealing with label quality disparity in federated learning[J]. *Federated Learning: Privacy and Incentive*, 2020, pp. 108-121.
- [16] Banica T, Curran S. Decomposition results for Gram matrix determinants[J]. *Journal of mathematical physics*, 2010, vol. 51(11), p. 113503.
- [17] Xu X, Wu Z, Foo C S, et al. Validation free and replication robust volume-based data valuation[C]. In *Proceedings of 35th International Conference on Neural Information Processing Systems*. 2021, vol. 34, pp. 10837–10848.
- [18] Jiang M, Roth H R, Li W, et al. Fair federated medical image segmentation via client contribution estimation[C]. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 16302-16311.
- [19] Wang Y, Li G L, and Li K Y. Survey on contribution evaluation for federated learning[J]. *Ruan Jian Xue Bao*, 2022, vol. 34(3), pp.1168–1192.
- [20] Kang J, Xiong Z, Niyato D, et al. Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory[J]. *IEEE Internet of Things Journal*, 2019, vol. 6(6), pp.10700-10714.
- [21] Miller N, Resnick P, Zeckhauser R. Eliciting informative feedback: The peer-prediction method[J]. *Management Science*, 2005, vol. 51(9), pp. 1359-1373.
- [22] Lv H, Zheng Z, Luo T, et al. Data-free evaluation of user contributions in federated learning[C] In *Proceedings of 19th International Symposium on Modeling and Optimization in Mobile, Ad hoc, and Wireless Networks*. 2021, pp. 81-88.
- [23] Cook R D, Weisberg S. Characterizations of an empirical influence function for detecting influential cases in regression[J]. *Technometrics*, 1980, vol. 22(4), pp. 495-508.
- [24] Richardson A, Filos-Ratsikas A, Faltings B. Rewarding high-quality data via influence functions[J]. *arXiv preprint arXiv: 1908.11598*, 2019.
- [25] Shyn S K, Kim D, Kim K. Fedceca: A practical approach of client contribution evaluation for federated learning[J]. *arXiv preprint arXiv: 2106.02310*, 2021.
- [26] Xu X, Lyu L, Ma X, et al. Gradient driven rewards to guarantee fairness in collaborative machine learning[J]. In *Proceedings of 35th International Conference on Neural Information Processing Systems*. 2021, pp. 16104–16117.
- [27] Cawley G C, Talbot N L C. Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers[J]. *Pattern recognition*, 2003, vol. 36(11), pp. 2585-2592.
- [28] Koh P W, Liang P. Understanding black-box predictions via influence functions[C]. In *Proceedings of the 34th International Conference on Machine Learning*. Sydney, NSW, Australia, 2017, vol. 70, pp. 1885–189.
- [29] Shapley L S. A value for n-person games. In: *Contributions to the Theory of Games [M]*. Princeton, New Jersey, Princeton University Press, 1953, vol. 2(28), pp. 307–318.
- [30] Sim R H L, Zhang Y, Chan M C, et al. Collaborative machine learning with incentive-aware model rewards[C]. In *Proceedings of the 37th International Conference on Machine Learning*. 2020, pp. 8927–8936.
- [31] Fan Z, Fang H, Zhou Z, et al. Improving fairness for data valuation in horizontal federated learning[C]. In *Proceedings of the 38th International Conference on Data Engineering*. Kuala Lumpur, Malaysia, 2022, pp. 2440-2453.
- [32] Fan Z, Fang H, Zhou Z, et al. Fair and efficient contribution valuation for vertical federated learning[J]. *arXiv preprint arXiv: 2201.02658*, 2022.
- [33] Jia R, Dao D, Wang B, et al. Towards efficient data valuation based on the shapley value[C]. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*. Naha, Okinawa, Japan, 2019, pp. 1167-1176.
- [34] Wang T, Rausch J, Zhang C, et al. A principled approach to data valuation for federated learning[J]. *Federated Learning: Privacy and Incentive*, 2020, pp. 153-167.
- [35] Kwon Y, Rivas M A, Zou J. Efficient computation and analysis of distributional shapley values[C]. In *Proceedings of 24th International Conference on Artificial Intelligence and Statistics*. 2021, vol. 130, pp. 793-801.
- [36] Yan T, Procaccia A D. If you like shapley then you'll love the core[C]. In *Proceedings of the 33th AAAI Conference on Artificial Intelligence*. 2021, vol. 35(6), pp. 5751–5759.
- [37] Villani C. *Optimal transport: old and new[M]*. Berlin: springer, 2009, vol. 338.
- [38] Peyré G, Cuturi M. Computational optimal transport: With applications to data science[J]. *Foundations and Trends® in Machine Learning*, 2019, vol. 11(5-6), pp. 355-607.
- [39] Just H A, Kang F, Wang J T, et al. Lava: Data valuation without pre-specified learning algorithms[C]. In *Proceedings of the 7th International Conference on Learning Representations*. Kigali, Rwanda, 2023.
- [40] Rakotomamonjy A, Nadjahi K, Ralaivola L. Federated Wasserstein Distance [J]. *arXiv preprint arXiv: 2310.01973*, 2023.