THE SURPRISING AMOUNT OF ARBITRARINESS IN SHAPLEY-VALUE DATA VALUATION

Hannah Diehl & Ashia Wilson
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
77 Massachusetts Ave, MA 02139, USA
{hdiehl,ashia07}@mit.edu

ABSTRACT

The growing economic importance of data has generated interest in principled methods for data valuation. Particular attention has been given to the Shapley value, a result from cooperative game theory that defines the unique distribution of a game's rewards to contributors subject to specified fairness axioms. By casting a machine learning task as a cooperative game, Shapley-based data valuation purports to equitably attribute model performance to individuals. However, the practical operationalization of this process depends on a wide array of practitioner decisions. Many of these decisions lie outside of the scope of the underlying machine learning task, introducing a potential for arbitrary decision making. The sensitivity of valuation outcomes to these intermediate decisions threatens the desired fairness properties. In light of these surfaced concerns, we evaluate the face-value equitability of Shapley for data valuation.

1 INTRODUCTION

As data is increasingly recognized as a valuable asset, there are growing calls for ethical and legal frameworks governing data ownership. While existing legislation, such as the European Union's *General Data Protection Regulation* and the *California Consumer Privacy Act*, focuses primarily on privacy and security, much of the broader landscape of data-related policy remains unaddressed. At the same time, the rapid rise of generative AI has intensified concerns about intellectual property infringement, emphasizing the urgent need to trace model behavior to specific training data points.

In this context, the technical question of data valuation—the process of assigning a numerical score to each data contributor in a dataset—has become highly relevant Zhang et al. (2024). Data valuation can shape the distribution of monetary rewards and determine how individuals and communities access and participate in the data economy. Furthermore, it has the potential to influence which creative inputs are acknowledged and credited, shaping perceptions of intellectual property and contributions in data-driven innovations Worledge et al. (2023). The social implications of these decisions are profound, requiring a focus on ensuring fair outcomes to maintain trustworthiness.

This paper spotlights a new challenge in the operationalization of Shapley-based data valuation: valuation sensitivity to the choices made in mapping machine learning tasks to a cooperative game. Using a vignette based on linear regression, we illustrate the surprising complexity of the decision space even in seemingly straightforward contexts. By highlighting these nuances, we aim to contribute towards bridging the gap between theory and the socially responsible deployment of data valuation methodologies.

2 RELATED WORK

The Economic Value of Data A growing body of literature explores the economic value of data from a variety of perspectives. The difficulty of handling the value of data as an asset in accounting settings is explored in(Moody & Walsh). (Stein & Maass, 2022) provides a review of methods within business and accounting for valuing data on the basis of costs, income-potential, and sale value on a market. Others have assessed the intrinsic value of data by appealing to privacy considerations

(Kleinberg et al., 2001), property law(Jurcys et al., 2020), and investment value(Cheong et al., 2023). A significant amount of recent literature has sought to describe functional characteristics of data marketplaces (Agarwal et al., 2019; Tian et al., 2023; Zhang et al., 2024). Additionally, many works have sought to quantify intrinsic quality of data and its instrumental value in machine learning applications (Sim et al., 2022; Schoch et al., 2023). When learned models have economic value through revenue-generation or cost-savings (Agarwal et al., 2019), data valuation seeks to attribute that economic performance to individual contributors to the training data.

The Shapley value (**?**), a key concept in cooperative game theory, has been widely applied to data valuation (**??**). Other cooperative game-based valuation methods have since been proposed (Kwon & Zou, 2022; Wang & Jia, 2023). For a comprehensive review of semivalue-based data valuation, including the Shapley value, see Zhang et al. (2024, Sec. 8.2). Other credit allocation methods from cooperative game theory, notably *the core*, have also been explored for data valuation(Yan & Procaccia, 2021). Some task-agnostic methods have been proposed but have not seen widespread uptake(Just et al., 2022; Ki et al., 2023).

Data Valuation Use Cases The data valuation literature has proposed a wide array of practical uses of values, including incentivizing data contribution from high-quality contributors(Zhu et al., 2019), attributing credit within multi-institute collaborations(Kumar et al., 2022), allocating payouts (Han et al., 2023), informing data acquisition(Ghorbani et al., 2020), and cleaning datasets of 'noisy' data (Namba et al., 2024; **?**; Zhou et al., 2022). Despite the data Shapley value and related notions being defined only within the context of a fixed learning application, some have taken these values to possibly correspond to some intrinsic quality of the data, leading to the Shapley value being used to imply value for another task (Schoch et al., 2023) or inform data pricing and market behavior (Tian et al., 2022; 2023).

Several studies have explored the use of semivalues in domain-specific applications. For instance, royalty-sharing frameworks (Wang et al., 2024) and revenue distribution models (Zhu et al., 2019; Ma et al., 2021) seek to leverage the Shapley value to fairly allocate payments, whether to copyright holders in content generation or contributors in medical data collection. In federated learning, the Shapley value has been proposed for fair credit attribution across institutions (Wang et al., 2020; Kumar et al., 2022).

Fairness and Arbitrariness The axiomatic fairness guarantees provided by the *Shapley value* and related notions merely constrain payouts to contributors within a defined game. Determining if outcomes derived from these valuations are 'fair' or 'equitable' likely requires consideration of context, similar to what has been shown for fairness claims for other machine learning tasks (Dwork et al., 2020). This assessment should be informed by broader work on algorithmic fairness in machine learning(Barocas et al.). The equitability of a single component of the data valuation pipeline - the fair allocation of rewards of a specified game - does not imply fairness of the process as a whole (Bower et al., 2017; Dwork & Ilvento, 2019). The precise definition of what makes policy outcomes dependent on data valuation fair is likely to be challenging to define (Binns, 2021),(Bothmann et al., 2024).

Given this, the practical fairness of data valuation and its consequences is dependent on implementation choices made by practitioners. This latitude introduces further ethical considerations when they constitute 'arbitrary' decisions. The moral implications of 'arbitrariness' are not straightforward and have recently begun to be studied in other ML fairness contexts (Black et al., 2022; Ganesh et al., 2025; Creel & Hellman, 2022). (Creel & Hellman, 2022) proposes that arbitrariness itself does not pose a fundamental moral issue, but rather the " but the systematicity of their arbitrariness" can give rise to ethical concerns. This calls for further attention to data valuation pipelines to assess whether analyst decisions are arbitrary in a way that promote systematic unfairness before claims of equity can be made.

3 Practical Specification of Shapley-Based Data Valuation

Data Shapley Let $\mathcal{D}$ denote a set of players of a cooperative game defined by a utility function U that maps subsets of $\mathcal{D}$ to the score attained when that coalition plays the game. The Shapley (1952) value of a contributor i denotes the reward that should be allocated to i in order to distribute the team's collective score, $U(\mathcal{D})$, in a way that satisfies the a desirable set of fairness axioms. These fairness axioms and the form of the Shapley value are provided in full in Appendix A. The *data Shapley value* extends this notion of value to data valuation in machine learning by casting supervised learning problems as cooperative games (Ghorbani & Zou, 2019; Jia et al., 2019). This notion of value is widely recognized to be defined with respect to a specific learning task. Per (Ghorbani & Zou, 2019), the data Shapley value of an observation may change due to changes to the learning architecture, and should change under change of task, with the provided example "regressing to the age of heart disease onset instead of heart disease incidence". Furthering this, we seek to highlight the sensitivity of the *data Shapley value* to the specification of the utility function even within the same substantive machine learning task.

Practitioner Choices: Defining the Game The utility function specified for data Shapley valuation depends on all choices entailed in specifying the learning task, including the learning algorithm and the model performance metric. Some of these choices are made implicitly and without careful scrutiny, especially those that do not influence the training behavior on the entire dataset and thus have no influence on the underlying ML task. We call particular attention to the role of specifying the behavior of the learning algorithm on small coalitions and monotonic transformations of the model performance function.

4 Illustrative Example

4.1 Scenario Description

Consider a scenario in which a regional health agency is using data-driven decision-making to optimize healthcare and improve resource allocations. As a model task, we consider hospital length of stay predictions — a key public health application (Stone et al., 2022). Suppose the agency acquires data from regional hospitals, implements a policy informed by a learned model, and subsequently observes a related decrease in costs. The agency seeks to allocate a portion of the cost savings to compensate data contributors, aiming to reward contributions fairly while avoiding reinforcing inequities.

In the underlying machine learning task, the agency models the length of hospitalization using a small set of real-valued features ($d = 4$) describing patient admission states using regularised linear least squares regression[1]. The regularization hyperparameter is optimized using leave-one-out cross validation (LOOCV). The dataset, $\mathcal{D}$, contains $n = 25$ electronic health records from two hospitals: a large, wealthy hospital (A) and a smaller hospital serving an economically disadvantaged community (B). The distinction between data from hospitals (A) and (B) is encoded in the distribution of admission states (see Appendix B for additional details on data generation).

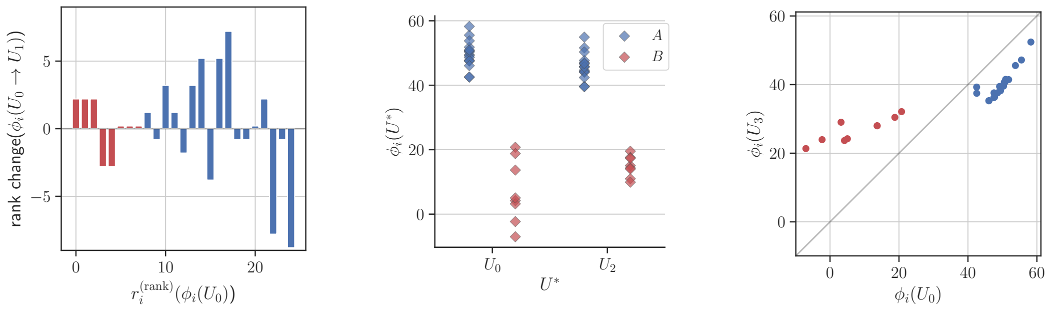
In the following experiments, we identify several junctures where an analyst may face ambiguity in making a decision in defining the learning algorithm or valuation metric. We propose a justifiable alternative utility function, selecting one among many possible variations. We then present experimental results on its impact in our outlined scenario and discuss broader implications for real-world applications.

4.2 Multiple Utility Specifications

Baseline Valuation For the baseline valuation, we derive one justifiable specification of the learning algorithm and model performance based on the underlying task. The assumptions implicit in this specification are assessed through the exploration of variations in subsequent sections.

The learning algorithm, $\mathcal{A}_{\lambda_0}$, is taken to be regularized linear least squares regression using the regularization hyperparameter determined by LOOCV during training in the underlying ML task. In practice, the computational complexity of Shapley value estimation may make retuning hyperparameters infeasible. Further, hyperparameter tuning may not be possible to reasonably define on small training coalitions. As a result, many applications will see at least some hyperparameter or model selection that is not incorporated into the learning algorithm, used for Shapley valuation.

[1] We assume the response (hospitalization duration) is truly a linear function of the covariates (admission states). While model misspecification may affect valuations, we reserve this exploration for future work.



(a) Variant 1

(b) Variant 2

(c) Variant 3

Figure 1: Variability of Shapley valuation outcomes under alternative utility specifications for the same learning task. Blue marks denote individuals in majority population(A), red marks denote individuals from minority population(B). (a) rank changes arising from re-specifying learning behavior on small coalitions; (b) valuation distributions without and with hyperparameter tuning into algorithm; (c) valuation outcomes when scaled RMSE is reported.

The model performance is evaluated on an held-out test set $\mathcal{T}$. While elements of $\mathcal{T}$ should ideally be assigned a value for its use, we only assign elements of $\mathcal{D}$ a value, and leave equitable value assignment strategies to members of $\mathcal{T}$ for future work. Performance score, V, is reported as the mean square error (MSE) of the model on this test set. This yields baseline utility $U_0 = V \circ \mathcal{A}_{\lambda_0}$.

Variant 1: Defining Behavior on 'Small' S Coalitions In general, machine learning algorithms $\mathcal{A}$ are designed for large datasets, and may not have appropriate behavior on small coalitions.

Within the example scenario, the practitioner may recognize that fitting a linear model with d free parameters to a training coalition with fewer observations does not correspond to an analysis the agency would have carried out in a counterfactual setting where only S was available as a training set. The analyst could incorporate this process assumption to capture an observation's actual marginal improvement to practical outcomes by defining an alternative utility where the learning algorithm returns a naive 'untrained' model when $|S| < d$. This results in alternative utility: $U_1(S) = U_0(S)\mathbb{I}\{|S| \geq d\} + U_0(\emptyset)\mathbb{I}\{|S| < d\}$.

This adjustment excludes marginal improvements from small coalitions in the value computation. Fig. 1a illustrates the resulting rank shifts, highlighting reward sensitivity across all outcome policies.

Han et al. (2023) note that Shapley valuations tend to be dominated by marginal contributions from small coalitions. As a result, valuation outcomes are highly sensitive to the practitioner's decisions in defining behavior on a small dataset. As shown experimentally, this particular variation to U systematically benefits individuals in population B As it does not change $\mathcal{A}(\mathcal{D})$, this specification is not intrinsic to the learning task and instead is up to the practitioner's discretion, giving rise to an ethics concern that arbitrariness is causing systematic impacts.

Variant 2: Defining Role of Hyperparameters in Algorithmic Counterfactual Hyperparameter tuning, such as cross-validation, is commonly treated as part of the learning algorithm. It may seem natural to tune hyperparameters for counterfactual models as well, but this is often avoided due to computational costs. However, whether or not hyperparameter tuning is performed can significantly impact data valuations. As cross-validation worsens the small coalition issue raised in variant 1, we consider optimization on a representative held-out validation set.

We simulate this behavior by defining the variant utility function $U_2 = \max_{\lambda \in L}(V_0 \circ \mathcal{A}_\lambda)$ where L denotes the domain of the LOOCV optimization in the underlying task.

Figure 1b illustrates how incorporating regularization hyperpameter tuning into the algorithm affects valuation distributions across hospital populations. We see that this choice raises valuations for Hospital B while reducing within-group variability, whereas Hospital A sees a slight decline. Without regularization, the model overfits to Hospital B — the minority in the test set — leading to large negative marginal effects when individuals from Hospital B are added to the empty coalition. Regularization mitigates this impact, which ultimately benefits Hospital B.

Variant 3: Alternative Reporting of Performance Metric Model scores in machine learning often undergo monotonic transformations without affecting real-world decisions. For instance, while models are commonly trained using mean square error (MSE), performance can equivalently be reported with root mean square error (RMSE) without loss of information or change in relative performances. A practitioner might assume that similar transformations in data valuation are benign. However, with semivalues, monotonic transformations can significantly alter valuations, even leading to rank reversals—an unexpected consequence given that such transformations typically preserve model rankings. To study this effect, we introduce $U_4 = K\sqrt{U_0}$, a variation of the baseline valuation method U_0 that replaces MSE with RMSE with scaling to preserve the sum of values. Valuation outcomes under this variant are depicted in Fig. 1c.

Results All three variants give rise to systematic changes to Shapley valuation outcomes despite corresponding to the same substantive learning task. Valuation changes corresponding to rank changes and wealth transfer from one demographic group to another are possible. These represent changes to the exact closed-form shapley values associated with each game.

5 Discussion

While the use of the Shapley value offer a principled framework for attributing value to data contributors, the equitability promise for the data valuation application depends on the appropriate specification of the underlying game. As shown, there is ambiguity in which game should be used. Alternative specifications of the underlying game can correspond to the same essential machine learning task, however lead to significantly different valuations for the same points. This highlights a concern that the Shapley value may not be corresponding to a contributors' 'true' value to a given learning task. This calls for significantly more attention to this requirement for practical data valuation.

Challenges arise from the imperfect mapping between machine learning tasks and cooperative games in the classical sense. While the composition of any learning algorithm and any model evaluation metric can constitute a utility function that defines a game, the subtle specification of the learning algorithm and metric can map the same machine learning task to different games with dramatically different valuations. Further complexities arise from the use of approximation methods and the need to translate computed valuations into concrete outcomes. Since raw valuation scores do not always translate directly to actionable insights, practitioners often employ normalization, thresholding, or ranking mechanisms. Careful evaluation of these choices is critical to ensuring fair and meaningful applications of data valuation.

While, the purpose of this paper is to highlight the impact of subtle decisions in data valuation, the development of prescriptive norms for making these decisions remains an area for future work. Establishing principled best practices—tailored to different applications and learning tasks—would provide much-needed guidance to practitioners. A formal framework for aligning data valuation choices with practical objectives would enhance the reliability and interpretability of semivalue-based data valuation.

REFERENCES

Anish Agarwal, Munther Dahleh, and Tuhin Sarkar. A Marketplace for Data: An Algorithmic Solution, May 2019. URL http://arxiv.org/abs/1805.08125. arXiv:1805.08125 [cs].

Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness and Machine Learning.

Reuben Binns. Fairness in Machine Learning: Lessons from Political Philosophy, March 2021. URL http://arxiv.org/abs/1712.03586. arXiv:1712.03586 [cs].

Emily Black, Manish Raghavan, and Solon Barocas. Model Multiplicity: Opportunities, Concerns, and Solutions. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 850–863, Seoul Republic of Korea, June 2022. ACM. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533149. URL https://dl.acm.org/doi/10.1145/3531146.3533149.

Ludwig Bothmann, Kristina Peters, and Bernd Bischl. What Is Fairness? On the Role of Protected Attributes and Fictitious Worlds, November 2024. URL http://arxiv.org/abs/2205.09622. arXiv:2205.09622 [cs].

Amanda Bower, Sarah N. Kitchen, Laura Niss, Martin J. Strauss, Alexander Vargas, and Suresh Venkatasubramanian. Fair Pipelines, July 2017. URL http://arxiv.org/abs/1707.00391. arXiv:1707.00391 [cs].

Hyongmook Cheong, Boyoung Kim, and Ivan Ureta Vaquero. A Data Valuation Model to Estimate the Investment Value of Platform Companies: Based on Discounted Cash Flow. *Journal of Risk and Financial Management*, 16(6):293, June 2023. ISSN 1911-8074. doi: 10.3390/jrfm16060293. URL https://www.mdpi.com/1911-8074/16/6/293. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute.

Kathleen Creel and Deborah Hellman. The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision-Making Systems. *Canadian Journal of Philosophy*, 52(1):26–43, January 2022. ISSN 0045-5091, 1911-0820. doi: 10.1017/can.2022.3. URL https://www.cambridge.org/core/product/identifier/S0045509122000030/type/journal_article.

Cynthia Dwork and Christina Ilvento. Fairness Under Composition. *LIPIcs, Volume 124, ITCS 2019*, 124:33:1–33:20, 2019. ISSN 1868-8969. doi: 10.4230/LIPIcs.ITCS.2019.33. URL http://arxiv.org/abs/1806.06122. arXiv:1806.06122 [cs].

Cynthia Dwork, Christina Ilvento, Guy N. Rothblum, and Pragya Sur. Abstracting Fairness: Oracles, Metrics, and Interpretability, April 2020. URL http://arxiv.org/abs/2004.01840. arXiv:2004.01840 [cs].

Prakhar Ganesh, Afaf Taik, and Golnoosh Farnadi. The Curious Case of Arbitrariness in Machine Learning, January 2025. URL http://arxiv.org/abs/2501.14959. arXiv:2501.14959 [cs] version: 1.

Amirata Ghorbani and James Zou. Data Shapley: Equitable Valuation of Data for Machine Learning, June 2019. URL http://arxiv.org/abs/1904.02868. arXiv:1904.02868 [cs, stat].

Amirata Ghorbani, Michael P. Kim, and James Zou. A Distributional Framework for Data Valuation, February 2020. URL https://arxiv.org/abs/2002.12334v1.

Dongge Han, Michael Wooldridge, Alex Rogers, Olga Ohrimenko, and Sebastian Tschiatschek. Replication-Robust Payoff-Allocation for Machine Learning Data Markets. *IEEE Transactions on Artificial Intelligence*, 4(5):1114–1128, October 2023. ISSN 2691-4581. doi: 10.1109/TAI.2022.3195686. URL http://arxiv.org/abs/2006.14583. arXiv:2006.14583 [cs].

Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gurel, Bo Li, Ce Zhang, Dawn Song, and Costas Spanos. Towards Efficient Data Valuation Based on the Shapley Value, February 2019. URL https://arxiv.org/abs/1902.10275v4.

Paulius Jurcys, Christopher Donewald, Mark Fenwick, Markus Lampinen, and Andrius Smaliukas. Ownership of User-Held Data: Why Property Law Is the Right Approach. *SSRN Electronic Journal*, 2020. ISSN 1556-5068. doi: 10.2139/ssrn.3711017. URL https://www.ssrn.com/abstract=3711017.

Hoang Anh Just, Feiyang Kang, Tianhao Wang, Yi Zeng, Myeongseob Ko, Ming Jin, and Ruoxi Jia. LAVA: Data Valuation without Pre-Specified Learning Algorithms. September 2022. URL https://openreview.net/forum?id=JJuP86nBl4q.

Nohyun Ki, Hoyong Choi, and Hye Won Chung. Data Valuation Without Training of a Model, March 2023. URL http://arxiv.org/abs/2301.00930. arXiv:2301.00930 [cs].

Jon Kleinberg, Christos H. Papadimitriou, and Prabhakar Raghavan. On the value of private information. In *Proceedings of the 8th conference on Theoretical aspects of rationality and knowledge*, TARK '01, pp. 249–257, San Francisco, CA, USA, July 2001. Morgan Kaufmann Publishers Inc. ISBN 978-1-55860-791-0.

Sourav Kumar, A. Lakshminarayanan, Ken Chang, Feri Guretno, Ivan Ho Mien, Jayashree Kalpathy-Cramer, Pavitra Krishnaswamy, and Praveer Singh. Towards More Efficient Data Valuation in Healthcare Federated Learning using Ensembling, September 2022. URL http://arxiv.org/abs/2209.05424. arXiv:2209.05424 [cs].

Yongchan Kwon and James Zou. Beta Shapley: a Unified and Noise-reduced Data Valuation Framework for Machine Learning, January 2022. URL http://arxiv.org/abs/2110.14049. arXiv:2110.14049 [cs, stat].

Shuaicheng Ma, Yang Cao, and Li Xiong. Transparent Contribution Evaluation for Secure Federated Learning on Blockchain. In *2021 IEEE 37th International Conference on Data Engineering Workshops (ICDEW)*, pp. 88–91, April 2021. doi: 10.1109/ICDEW53142.2021.00023. URL http://arxiv.org/abs/2101.10572. arXiv:2101.10572 [cs].

Daniel Moody and Peter Walsh. Measuring The Value Of Information: An Asset Valuation Approach.

Hiroyuki Namba, Shota Horiguchi, Masaki Hamamoto, and Masashi Egi. Thresholding data shapley for data cleansing using multi-armed bandits, 2024. URL https://arxiv.org/abs/2402.08209.

Stephanie Schoch, Ritwick Mishra, and Yangfeng Ji. Data Selection for Fine-tuning Large Language Models Using Transferred Shapley Values, June 2023. URL http://arxiv.org/abs/2306.10165. arXiv:2306.10165 [cs].

Lloyd S. Shapley. A Value for N-Person Games. Technical report, RAND Corporation, March 1952. URL https://www.rand.org/pubs/papers/P295.html.

Rachael Hwee Ling Sim, Xinyi Xu, and Bryan Kian Hsiang Low. Data Valuation in Machine Learning: "Ingredients", Strategies, and Open Challenges. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 5607–5614, Vienna, Austria, July 2022. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-1-956792-00-3. doi: 10.24963/ijcai.2022/782. URL https://www.ijcai.org/proceedings/2022/782.

Hannah Stein and Wolfgang Maass. *Requirements for Data Valuation Methods*. January 2022. ISBN 978-0-9981331-5-7. URL http://hdl.handle.net/10125/80086.

Kieran Stone, Reyer Zwiggelaar, Phil Jones, and Neil Mac Parthaláin. A systematic review of the prediction of hospital length of stay: Towards a unified framework. *PLOS Digital Health*, 1(4): e0000017, April 2022. ISSN 2767-3170. doi: 10.1371/journal.pdig.0000017. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9931263/.

Yingjie Tian, Yurong Ding, Saiji Fu, and Dalian Liu. Data Boundary and Data Pricing Based on the Shapley Value. *IEEE Access*, 10:14288–14300, 2022. ISSN 2169-3536. doi: 10.1109/ACCESS.2022.3147799. Conference Name: IEEE Access.

Zhihua Tian, Jian Liu, Jingyu Li, Xinle Cao, Ruoxi Jia, Jun Kong, Mengdi Liu, and Kui Ren. Private Data Valuation and Fair Payment in Data Marketplaces, February 2023. URL http://arxiv.org/abs/2210.08723. arXiv:2210.08723 [cs].

Jiachen T. Wang and Ruoxi Jia. Data Banzhaf: A Robust Data Valuation Framework for Machine Learning, March 2023. URL http://arxiv.org/abs/2205.15466. arXiv:2205.15466 [cs, stat].

Jiachen T. Wang, Zhun Deng, Hiroaki Chiba-Okabe, Boaz Barak, and Weijie J. Su. An Economic Solution to Copyright Challenges of Generative AI, September 2024. URL http://arxiv.org/abs/2404.13964. arXiv:2404.13964 [cs].

Tianhao Wang, Johannes Rausch, Ce Zhang, Ruoxi Jia, and Dawn Song. A Principled Approach to Data Valuation for Federated Learning, September 2020. URL http://arxiv.org/abs/2009.06192. arXiv:2009.06192 [cs, stat].

Theodora Worledge, Judy Hanwen Shen, Nicole Meister, Caleb Winston, and Carlos Guestrin. Unifying Corroborative and Contributive Attributions in Large Language Models, November 2023. URL http://arxiv.org/abs/2311.12233. arXiv:2311.12233 [cs].

Tom Yan and Ariel D. Procaccia. If You Like Shapley Then You'll Love the Core. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(6):5751–5759, May 2021. ISSN 2374-3468. doi: 10.1609/aaai.v35i6.16721. URL https://ojs.aaai.org/index.php/AAAI/article/view/16721. Number: 6.

Jiayao Zhang, Yuran Bi, Mengye Cheng, Jinfei Liu, Kui Ren, Qiheng Sun, Yihang Wu, Yang Cao, Raul Castro Fernandez, Haifeng Xu, Ruoxi Jia, Yongchan Kwon, Jian Pei, Jiachen T. Wang, Haocheng Xia, Li Xiong, Xiaohui Yu, and James Zou. A Survey on Data Markets, November 2024. URL http://arxiv.org/abs/2411.07267. arXiv:2411.07267 [cs].

Zijian Zhou, Xinyi Xu, Rachael Hwee Ling Sim, Chuan Sheng Foo, and Kian Hsiang Low. Probably Approximate Shapley Fairness with Applications in Machine Learning, December 2022. URL http://arxiv.org/abs/2212.00630. arXiv:2212.00630 [cs].

Liehuang Zhu, Hui Dong, Meng Shen, and Keke Gai. An Incentive Mechanism Using Shapley Value for Blockchain-Based Medical Data Sharing. In *2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*, pp. 113–118, May 2019. doi: 10.1109/BigDataSecurity-HPSC-IDS.2019.00030.

A SHAPLEY VALUE

Consider a cooperative game where a coalition of players, S, can attain a score according to function $U(S)$. **?** defines the *Shapley value*, ϕ_i, of player i on team $\mathcal{D}$ participating in this game as:

$$\phi_i = \frac{1}{|\mathcal{D}|} \sum_{S \subseteq \mathcal{D} \setminus \{i\}} \frac{U(S \cup \{i\}) - U(S)}{\binom{|\mathcal{D}|-1}{|S|}}, \tag{1}$$

1. *Null Player Axiom:* If the addition of player i to any subset of the team results in no change to that coalition's reward, contributor i should be given value 0.
2. *Symmetry Axiom:* If players i and j have equivalent contributions when added to any subset S, they should be given the same value.
3. *Linearity Axiom:* If U is sum of sub-scores $U_1, U_2, ...U_n$, ϕ_i should be the sum of Shapley values for each sub-score.
4. *Group Rationality Axiom:* The full value of the team's collective reward should be distributed to players: $\sum_{i \in \mathcal{D}} \phi_i = U(\mathcal{D}) - U(\emptyset)$.

B Experiment Details

Parameters		Value
Total contributors	n	25
from Hospital A	n_A	17
from Hospital B	n_B	8
Observed features	d	4
Model Parameters	m	4
Covariate Means		
Hospital A	$\bar{x}_A$	[2, 10, 1, 5]
Hospital B	$\bar{x}_B$	[12, 2, 5, 1]
Covariate Variance		
Hospital A	$\Sigma_{x,A}$	1.5 I_d
Hospital B	$\Sigma_{x,B}$	1.0 I_d
Response Noise Scale		
Hospital A	$\sigma_{y,A}$	0.5
Hospital B	$\sigma_{y,B}$	1.1
True Model Parameters	β	[3, -3, 1, -1]

Table 1: Parameters for experiment

Data Generative Model Covariate data is generated according to a two component multivariate gaussian mixture model, with each component corresponding to one hospital population. Data from Hospital H (which we take to be representative of some Population H) is drawn according to

$$x_i \sim \mathcal{N}(\bar{x}_H, \Sigma_x^2)$$
$$y_i = x_i^\top \beta^\star + \epsilon_i; \quad \epsilon_i \overset{\text{iid.}}{\sim} \mathcal{N}(0, \sigma_y^2)$$

with the numerical parameters specified in Table 1.

Utility Specification

$$\mathcal{A}_\lambda(S) = \arg\min_\beta \textstyle\sum_{i \in S} (y_i - x_i^\top \beta)^2 + \lambda \cdot \|\beta\|^2 \,,$$
$$V(\beta) = -\textstyle\sum_{(\overline{x},\overline{y}) \in \mathcal{T}} (\overline{y} - \overline{x}^\top \beta)^2 \,.$$

Exact Semivalue Computation The score for a coalition fit with regularizer λ on a test set with Hospitals A and B represented in w_A and w_B fractions of the test set respectively is computed by computing the fit

$$\hat{\beta}(X, y; \lambda) = (X^T X + \lambda)^{-1}(X^T y)$$

This learned model is scored using negative MSE on each subpopulation. Given the gaussian distribution for each subpopulation and linear response, this can be done analytically allowing approximation of the score on a large test set[2].

$$\begin{aligned} V_H(S;\lambda) &= - \mathop{\mathbb{E}}_{x,y \sim H}[(\hat{\beta}(S;\lambda)x - y)^2] \\ &= -((\hat{\beta} - \beta^*) \cdot x_h)^2 - \sigma_x^2 ||(\hat{\beta} - \beta^*)||^2 + \sigma_y^2 \end{aligned}$$

The utility evaluated on a test set with f_A representation from population A and f_B representation from population B is:

$$U(S;\lambda) = \sum_{H \in \{A,B\}} f_H V_H(S;\lambda)$$

Experimental results are computed for a test set with 75% representation of Population A and 25% representation of Population B.

[2] In the retrospective payout setting, the test set can be collected after model deployment, so the constraint on the size of $\mathcal{D}$ would not necessarily imply small $\mathcal{T}$

The Shapley value - possible to compute exactly due to the small dataset size - is comptued as:

$$\phi_i(U) = \sum_{S \in Pow(\mathcal{D})} w^{(i)}(S)U(S)$$

Where

$$w^{(i)}(S) = \begin{cases} \binom{n-1}{|S|-1}^{-1} & i \in S \\ \binom{n-1}{|S|}^{-1} & i \notin S \end{cases}$$

Utility Variation Experiment Denote the optimal regularization from leave-one-out cross validation as λ_0. For the drawn dataset, among candidate set $L = \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$, $\lambda_0 = 10^{-2}$. Using the primitives described in 'Exact Semivalue Computation', the baseline valuation method and variations are formalized as follows:

- **Baseline Valuation**: $U_0(S) = U(S; \lambda_0)$
- **Small S Variation**: $U_1(S) = U(S; \lambda_0)\mathbb{I}\{|S| \geq 4\} + U(\emptyset)\mathbb{I}\{|S| < 4\}$
- **Hyperparameter Retuning Setting**: $U_2(S) = \max_{\lambda \in L} U(S, \lambda)$
- **Score Transformation**: $U_3(S) = \sqrt{U(S; \lambda_0)}$