Measuring the Hidden Cost of Data Valuation through Collective Disclosure

Anonymous Author(s)
Affiliation
Address
email

Abstract

Data valuation methods assign marginal utility to each data point that has contributed to the training of a machine learning model. If used directly as a payout mechanism, this creates a *hidden cost of valuation*, in which contributors with near-zero marginal value would receive nothing, even though their data had to be collected and assessed. To better formalize this cost, we introduce a conceptual and game-theoretic model—the *Information Disclosure Game*—between a *Data Union* (sometimes also called a data trust), a member-run agent representing contributors, and a *Data Consumer* (e.g., a platform). After first aggregating members' data, the DU releases information progressively by adding Laplacian noise under a differentially-private mechanism. Through simulations with strategies guided by data Shapley values and multi-armed bandit exploration, we demonstrate on a Yelp review helpfulness prediction task that data valuation inherently incurs an explicit acquisition cost and that the DU's collective disclosure policy changes how this cost is distributed across members.

1 Introduction

The idea of data dividends [Feygin et al., 2021]—distributing a share of value generated from data back to contributors—has drawn attention as a mechanism to make data economies more inclusive. A way to ground such dividends is through data valuation, in which each point's contribution to predictive performance is quantified. Recent methods such as data Shapley value (DSV) [Jia et al., 2019b], which originated from cooperative game theory, have been applied to data summarization and efficient acquisition [Ghorbani and Zou, 2019]. These methods show that only a fraction of the data is needed for high utility, but nonetheless require access to the entire dataset to compute the marginal contributions—meaning that each individual must first contribute their data, even if their eventual payout is negligible. We refer to this as the *hidden cost of data valuation*.

Following calls for data trusts and unions as institutional forms of collective governance [Delacroix and Lawrence, 2019], we frame our approach as a form of algorithmic collective action in which contributors act collectively through a Data Union (DU) that sets group data disclosure policies, shifting how value and costs are distributed. While recent work shows that differentially-private mechanisms Dwork [2006] can diminish the leverage of collective action in gradient-based training [Solanki et al., 2025], we instead use DP positively as a tool for disclosure and value control.

More precisely, we introduce a conceptual model in which contributors coordinate through a DU. Rather than acting individually in a decentralized marketplace, members pool their data and empower the union to negotiate collectively. The DU aggregates the dataset and controls how information is released to a Data Consumer (DC) by progressively disclosing noisy versions of data points using a differentially-private mechanism. In this way, the DU operates as a collective agent: it sets disclosure

Submitted to the NeurIPS 2025 Workshop on Algorithmic Collective Action. Do not distribute.

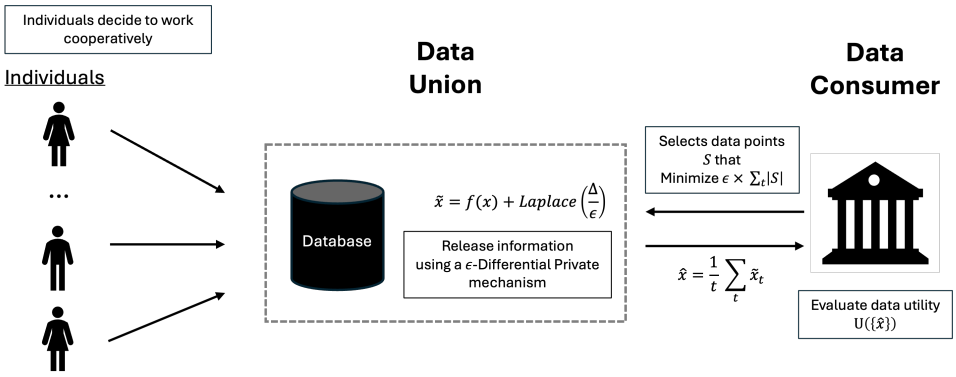

Figure 1: Illustration of the Information Disclosure Game. A Data Union (DU) holds a private dataset and releases information to a Data Consumer (DC) by adding Laplacian noise to data points under ϵ-differential privacy. The DC incrementally acquires these noisy version of the data points, denoises them using an average and train a model to reach a utility target. In this work, we focus on non-parametric k-Nearest Neighbors (kNN) and SBERT embeddings Reimers and Gurevych [2019].

policies that influence which points the DC must acquire to reach its utility target, and thus how dividends are distributed across members. See Figure 1 for an illustration of our approach.

We formalize this interaction as an Information Disclosure Game (IDG), a form of Stackelberg game Simaan and Cruz Jr [1973], Fudenberg and Tirole [1991] in which the DU acts as leader by choosing disclosure rules, and the DC acts as follower by optimizing acquisitions given those rules (Figure 1). This framework allows to make explicit the acquisition cost that valuation imposes on the DC and how collective disclosure policies shape their distribution across members. We instantiate the model by a k-Nearest Neighbors (kNN), in which the relation between individual points and predictive utility is direct. Simulating the DC strategies guided by Shapley values and by multi-armed bandit exploration, we test it on a Yelp review helpfulness prediction task using text embeddings.

Outline. First in Section 2, we review the related work on data valuation and differential privacy before detailing in Section 3 the IDG framework and the objectives of DU and DC. Afterwards in Section 4, we present our empirical evaluation followed by a discussion and conclusion.

2 Background

Data Valuation. In cooperative game theory, an *imputation* is a way to distribute the total value of a game among its players Osborne and Rubinstein [1994]. In the context of data valuation, each data point in a dataset D is treated as a *player* and a *value function* $v : 2^D \to \mathbb{R}$ is used to the utility (*e.g.*, accuracy or loss reduction) of any subset $D' \subseteq D$. The objective is to find an imputation that fairly allocates the total value $v(D)$ among individual data points based on their contributions. Several approaches have been explored for data valuation—each with different fairness, computational and interpretability trade-offs—including *Data Shapley* Ghorbani and Zou [2019], Jia et al. [2019b], *Data Banzhaf* Wang and Jia, *Beta Shapley* Kwon and Zou and the *Core* Yan and Procaccia.

Formally, the Shapley value of a data point $z_i \in D$ is given by:

$$\phi_{z_i}(v) = \mathbb{E}_{D' \sim \mathcal{P}(D \setminus \{z_i\})}\big[v(D' \cup \{z_i\}) - v(D')\big],$$

in which D' is a subset of $D \setminus \{z_i\}$ sampled uniformly at random and $v(D')$ is the value function evaluated on D'. The Shapley value thus quantifies the *expected* incremental contribution of each data point across all possible subsets. The uniqueness of the Shapley value can be derived from satisfying four foundational axioms: *Efficiency* (*i.e.*, the total value is distributed among all players), *Symmetry* (*i.e.*, identical contributions are rewarded equally), *Dummy* (*i.e.*, players that contribute nothing receive zero value) and *Linearity* (*i.e.*, the Shapley values from two games can be combined linearly). One application is *data pricing* in marketplaces Jia et al., Pei, Tian et al., Xia et al., in

which Shapley-based frameworks assign higher prices to points that contribute across many subsets. Shapley values can also inform the *data selection* task, by identifying examples most useful for training smaller high-quality sets. Their effectiveness depends on the structure of the utility function and faces the challenge of high computational cost, with approximations such as G-Shapley Ghorbani and Zou [2019], Yoon et al. [2020] proposed to improve scalability. However, for k-NN classifiers, an exact Shapley formulation exists Jia et al. [2019a] with complexity $O(N \log N)$, making valuation tractable in this setting.

Differential privacy. Differential Privacy (DP) Dwork [2006] provides a mathematically rigorous privacy framework by ensuring that the inclusion or exclusion of any individual in a dataset does not significantly alter the probability distribution of outputs. For instance, the trade-off between privacy and utility can be controlled through noise addition, thus reducing the re-identification risk and providing provable privacy guarantees. Formally, let $D = \{z_1, z_2, \ldots, z_n\}$ be a dataset consisting of n individual data points. A randomized mechanism M satisfies (ϵ, δ)-differential privacy if, for any dataset D, any data point $z_i \in D$, and any subset of possible outputs O, the following holds:

$$\Pr[M(D) \in O] \leq e^{\epsilon} \Pr[M(D \setminus \{z_i\}) \in O] + \delta,$$

in which ϵ represents the privacy loss and δ allows for a small probability that strict ϵ-differential privacy does not hold. Beyond protection, privacy also concerns the incentives individuals face when deciding whether to share their data truthfully. From a mechanism design perspective, the ***Revelation Principle*** Myerson [1983] states that outcomes achievable through indirect strategies can also be implemented by truthful revelation. Differential privacy makes such mechanisms approximately truthful: in McSherry and Talwar [2007], an ε-DP mechanism bounds each agent's influence on the outcome, so the *gain from lying* is at most $O(\varepsilon)$.

3 Information Disclosure Game

We model the interaction between two agents: the DU and the DC. As a member-run collective, the DU manages pooled data with the mission of ensuring fair valuation while accounting for concerns such as privacy. Meanwhile, the DC seeks to optimize utility—often by acquiring data points at minimal cost. These interactions are formalized as a two-phase Stackelberg game, in which the DU acts as a leader by setting how to disclose information and at what cost, and the DC, as a follower, responds through strategic data acquisition.

Complete information disclosure game. A typical formulation is to model the interaction as a pricing game, in which the DU assigns a price p_i to each data point z_i. The DC then solves a knapsack-like problem:

$$\min_{\mathbf{x} \in \{0,1\}^N} \quad \sum_{i=1}^{N} p_i x_i \quad \text{s.t.} \quad U(\mathbf{x}) \geq U_{\text{target}}, \tag{1}$$

in which $U(\mathbf{x})$ denotes the utility obtained from the purchased subset. The DU anticipates this behavior and sets $\mathbf{p}$ with the aim of maximizing the DC's minimized total cost.

This approach provides a straightforward pricing mechanism but has significant limitations. Indeed, because disclosure of a data point is binary—either withheld or complete—many contributors receive no compensation. Incentivizing broader acquisition requires lowering the price of high-value points, which penalizes those who contribute the most. Moreover, releasing data at fixed prices locks in value at a single point in time and weakens privacy guarantees, limiting the DU's ability to balance inclusiveness and protection.

Partial information disclosure game. To address these challenges, we depart from per-point pricing and instead model an iterative disclosure process under DP. At each round t, the DC selects a subset S_t of data points to query. Each query consumes a fixed privacy budget ϵ and returns a noisy version of the data. The DU sets $(\epsilon, T_{\max})$, the per-query budget and the per-point query cap, which together determine a maximum per-point spend $B_{\max} = T_{\max}\epsilon$. The DC's cumulative spend is $B = \epsilon \sum_{t=1}^{T} |S_t|, \quad T \leq T_{\max}$.

The DC's objective is to minimize its cumulative budget spent:

$$\min_{\{S_t\}} \epsilon \sum_{t=1}^{T} |S_t| \quad \text{s.t.} \quad U^{(T)} \geq U_{\text{target}},\ T \leq T_{\max}. \tag{2}$$

Conversely, the DU's objective is to maximize the minimum budget spent by the DC:

$$\max_{(\epsilon, T_{\max})} \min_{\{S_t\}} \epsilon \sum_{t=1}^{T} |S_t| \tag{3}$$

with the goal of spreading spend more evenly across members while managing privacy–utility trade-offs. Compared to the pricing game in Equation 1, partial disclosure replaces per-point ownership costs with iterative access charges. This formulation makes the hidden acquisition cost of valuation explicit in the DC's budget and allows the DU to shape acquisition behaviors without penalizing the most valuable contributors.

4 Experiments on Review Helpfulness Prediction

Review helpfulness prediction provides a natural testbed for data valuation as platforms such as Yelp collect reviews not only for sentiment analysis but also to assess their quality. Indeed, unhelpful reviews may ultimately be discarded, yet they are still necessary to identify and reward helpful ones. This creates a setting in which individual contributions to predictive performance are both explicit and uneven. We use the Yelp dataset Asghar [2016] following the setup of Bilal and Almazroi [2023], who report a k-NN accuracy of 59.6%. Using pretrained text embeddings and k-NN, our best configuration achieves 66.0% test accuracy and 65.2% F1-score, bringing non-parametric performance closer to fine-tuned transformer models while preserving point-level interpretability needed for valuation. As a sanity check, we replicated the acquisition experiment of Jia et al. [2019b] and verified that Shapley-based selection outperforms random sampling. The corresponding curve, along with additional details on the experimental setup, is reported in Appendix D.

To implement a DP iterative release, we add Laplacian noise independently to each feature of every data point (1024-dimensional vectors). For feature x_j, we first project it into a fixed interval $[a_j, b_j]$, in which a_j and b_j are min/max statistics for feature j. The release mechanism is then

$$\tilde{x}_j \;=\; \min\{\max\{x_j,\, a_j\},\, b_j\} \;+\; \text{Laplace}\left(\frac{\Delta_j}{\epsilon_j}\right),$$

in which $\Delta_j = b_j - a_j$ denotes the global sensitivity of feature j after bounding and ϵ_j is the allocated budget for that feature. This design is similar in spirit to a per-feature local DP mechanism Cormode et al. [2018], Wang et al. [2019], in that each coordinate is privatized independently, but here it is applied in a *centralized* setting by the DU. We assume the sensitivity interval $[a_j, b_j]$ is fixed in advance and does not vary significantly across releases, so the noise scale is determined by Δ_j/ϵ_j. On the DC side, each noisy version of a point is averaged to form a denoised estimate. After observing t noisy versions of point i, its center is computed incrementally as $\hat{x}_i^{(t)} = \frac{1}{t}\sum_{k=1}^{t} \tilde{x}_i^{(k)}$. The averaging strategy steadily improves fidelity as noisy samples accumulate, with error decreasing and correlation increasing non-linearly across iterations (see Appendix B for details).

In the remainder of our analysis, we do not attempt to solve the full game between the DU and DC. Instead, we simulate the DC's behavior by implementing and comparing different data selection strategies under the noisy iterative release mechanism. We define the utility target as the validation accuracy (69.6%). For all data selection strategies, we use a fixed privacy budget ϵ per point per query of 1 and assume a constant maximum budget per data point. The next two subsections explore these strategies, including one based on rankings (random and Shapley-based) and another using adaptive n-armed bandit algorithm.

4.1 Data Selection Using Random and Shapley-based Strategies

Random data selection. In this baseline, the DC commits to a random subset of data points for all iterations with the results displayed in Figure 2 being averaged over 10 random permutations. As shown in Figure 2, this approach fails to reach the target utility within 100 iterations unless nearly 100% of the dataset is acquired, which confirms that a purely random acquisition strategy from the DU is not viable under budget constraints.

Data Shapley selection. Afterwards, we assess whether Shapley valuation works over noisy data by the DC selecting all data points for a number of iterations (*i.e.*, bootstrap iterations) and estimating DSVs based on the averaged center points. As illustrated in Figure 3, selecting the top-valued centers based on noisy Shapley estimates allows the DC to reach the target utility. Successful acquisition starts at 10% of the dataset but performance varies depending on the percentage of selected data and the number of iterations. For instance, we observed that in the 60% selection case, data Shapley selection achieves the utility target after roughly 25 bootstrap iterations. Testing other Shapley-based strategies (Appendix D), we found no consistent advantage in committing earlier. Although committing may stabilize performance for a fixed subset, our findings suggest that for Shapley-based strategies, the DC is better off continuing complete iterations to refine noisy point estimates.

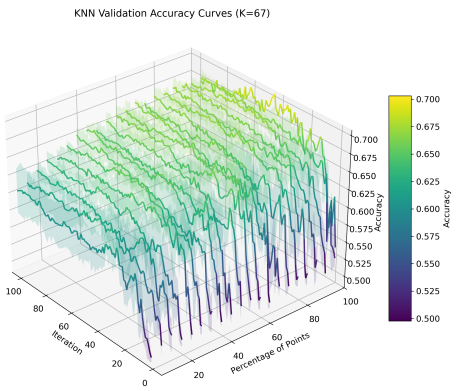

Figure 2: Validation accuracy for random data selection across varying parameters. Unlike Shapley-based methods, random selection fails to consistently reach the utility target (69.6%) within the budgeted iteration range.

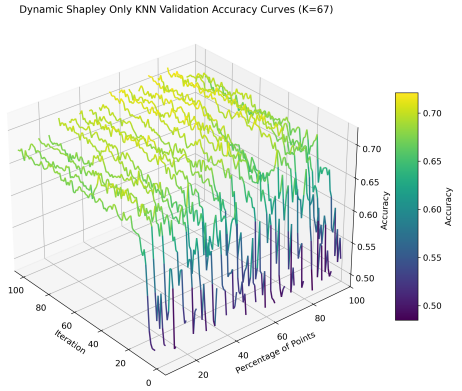

Figure 3: Validation accuracy using estimated data Shapley selection across dataset percentages and iterations. The target accuracy is achieved with as little as 10% of data.

4.2 Data Selection using n-Armed Bandits

As a final strategy, we model data selection as an n-armed bandit problem, in which each action $a \in \{1, \ldots, n\}$ corresponds to selecting a data point z_a. The DC follows an Upper Confidence Bound (UCB) policy under a fixed per-point privacy budget B_{MAX}. Each query incurs a privacy cost ϵ and yields a noisy sample that updates the point's estimated value, measured as incremental utility in a k-NN classifier. Unlike Shapley-based strategies, the DC may exploit the same point repeatedly until its budget is exhausted. This formulation captures the exploration–exploitation trade-off and makes the *cost of valuation* explicit as the exploration cost required to identify valuable points under budget constraints. Formally, each point maintains its estimated value $Q_t(a)$, query count $N_t(a)$, current averaged center center$_a$ and remaining budget B_a^{re}. UCB scores are computed as

$$\text{UCB}_t(a) = \begin{cases} Q_t(a) + c \cdot \sqrt{\dfrac{1}{N_t(a) + \varepsilon}} \cdot \dfrac{B_a^{\text{re}}}{B_a^{\text{MAX}}}, & \text{if } B_a^{\text{re}} > 0 \\ -\infty, & \text{otherwise.} \end{cases}$$

The utility of a queried point is defined as the fraction of neighbors in k-NN that share its label. See Algorithm 1 in Appendix for implementation details.

Hyperparameter search. Identifying an equilibrium can be seen as a hyperparameter search, since the interaction between the DU's release policy and the DC's acquisition strategy does not admit a closed-form solution. To realize this, we conducted a grid search over two key parameters: the *maximum budget per point* (set by the DU) and the *exploration coefficient* c (in UCB). These jointly determine budget usage and the diversity of selected points. Results show that the bandit reliably reaches the utility threshold except when the per-point budget is too small or exploration is disabled; below a budget of 20, success is rare and often due to *lucky point selection*. Gini analysis

of budget allocations confirms that exploration promotes more balanced spending across points and increases the likelihood of success. The correlation between Q values and Shapley values grows with budget but remains moderate overall (Appendix E).Finally, evaluations based on denoised centers yielded test accuracy comparable to or exceeding that obtained from the original data. For instance, a configuration with a per-point budget of 50 reached 69.9% surpassing the 66% benchmark.

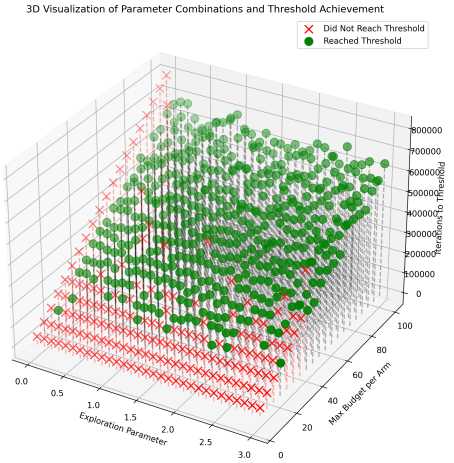

Figure 4: 3D visualization of hyperparameter combinations. Green dots represent successful runs in which the DC reached the utility threshold while red crosses represent failures. Success becomes unlikely with budget-per-arm below 20 or when exploration is zero.

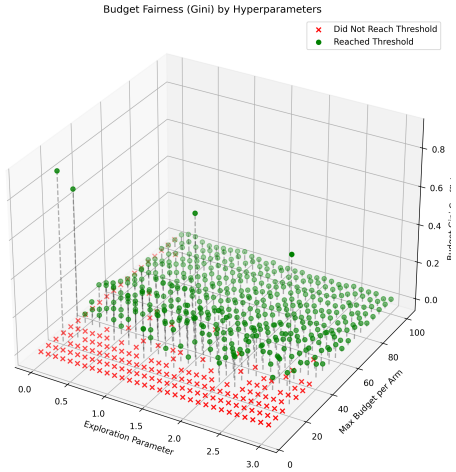

Figure 5: Gini coefficient of budget usage by hyperparameter setting. High values indicate budget concentrated on few data points, typical in "lucky" early selections.

5 Discussion & Conclusion

We proposed a game-theoretic framework for data valuation that integrates privacy and inclusiveness concerns through an information disclosure game between a Data Union (DU) and a data consumer (DC). Rather than assuming direct access to data, our approach models acquisition as an iterative process using a differential private mechanism to enforce a cost on data valuation. On Yelp review helpfulness prediction, Shapley-based strategies remained effective under noise but required a minimal budget—about 10 bootstrap iterations—to reliably identify high-value points. This demonstrates that Shapley valuation entails an inherent cost of exploration in our setting. Similarly, a multi-armed bandit strategy reached the utility target with comparable budgets, reinforcing that valuation itself imposes exploration costs. Gini analysis further showed that as budgets grow, the DC allocates resources more evenly across contributors, increasing inclusiveness.

While our experiments focused on k-NN for computation efficiency, an important direction for future work is extending the framework to differentiable models. Gradient-based methods such as *G-Shapley* Ghorbani and Zou [2019] already show that Shapley values can be approximated from changes in gradients, suggesting ways to scale valuation beyond non-parametric models. In parallel, approaches like *DP-SGD* Abadi et al. [2016] demonstrate that DP can be enforced by adding noise to gradients rather than directly to points. Bridging these lines of work would provide a better privacy-utility trade-off for both the DU and DC. Nonetheless, our results suggest that in a DU setting, reaching target utility requires a minimum level of budget spread. Thus, our framework implies that a *minimum dividend* should be guaranteed to all members, regardless of individual Shapley value. In the context of reviews, this means that every contributor would receive at least some share of value, even if their individual review is ultimately deemed unhelpful, aligning the incentives of all members with the collective outcome. Otherwise, excluding some contributors increases the incentive to form data unions that adopt adversarial disclosure strategies (*e.g.*, prioritizing low-value reviews or injecting excess noise), which may in turn hinder the data consumer's ability to reach its utility target.

References

M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

N. Asghar. Yelp dataset challenge: Review rating prediction. *arXiv preprint arXiv:1605.05362*, 2016.

M. Bilal and A. A. Almazroi. Effectiveness of fine-tuned bert model in classification of helpful and unhelpful online customer reviews. *Electronic Commerce Research*, 23(4):2737–2757, 2023.

G. Cormode, S. Jha, T. Kulkarni, N. Li, D. Srivastava, and T. Wang. Privacy at Scale: Local Differential Privacy in Practice. In *Proceedings of the 2018 International Conference on Management of Data*, SIGMOD '18, pages 1655–1658, New York, NY, USA, May 2018. Association for Computing Machinery. ISBN 978-1-4503-4703-7. doi: 10.1145/3183713.3197390. URL https://doi.org/10.1145/3183713.3197390.

S. Delacroix and N. D. Lawrence. Bottom-up data trusts: Disturbing the 'one size fits all'approach to data governance. *International data privacy law*, 9(4):236–252, 2019.

C. Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer, 2006.

Y. Feygin, H. Li, C. Lala, B. Hecht, N. Vincent, L. Scarcella, and M. Prewitt. A data dividend that works: steps toward building an equitable data economy. 2021.

D. Fudenberg and J. Tirole. *Game theory*. MIT press, 1991.

A. Ghorbani and J. Zou. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, pages 2242–2251. PMLR, 2019.

R. Jia, D. Dao, B. Wang, F. A. Hubis, N. Hynes, N. M. Gürel, B. Li, C. Zhang, D. Song, and C. J. Spanos. Towards Efficient Data Valuation Based on the Shapley Value. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pages 1167–1176. PMLR. URL https://proceedings.mlr.press/v89/jia19a.html.

R. Jia, D. Dao, B. Wang, F. A. Hubis, N. M. Gurel, B. Li, C. Zhang, C. J. Spanos, and D. Song. Efficient task-specific data valuation for nearest neighbor algorithms. *arXiv preprint arXiv:1908.08619*, 2019a.

R. Jia, D. Dao, B. Wang, F. A. Hubis, N. Hynes, N. M. Gürel, B. Li, C. Zhang, D. Song, and C. J. Spanos. Towards efficient data valuation based on the shapley value. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1167–1176. PMLR, 2019b.

Y. Kwon and J. Zou. Beta Shapley: A Unified and Noise-reduced Data Valuation Framework for Machine Learning. URL http://arxiv.org/abs/2110.14049.

F. McSherry and K. Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE, 2007.

R. B. Myerson. Mechanism design by an informed principal. *Econometrica: Journal of the Econometric Society*, pages 1767–1797, 1983.

M. J. Osborne and A. Rubinstein. *A course in game theory*. MIT press, 1994.

J. Pei. A Survey on Data Pricing: From Economics to Data Science. 34(10):4586–4608. ISSN 1041-4347, 1558-2191, 2326-3865. doi: 10.1109/TKDE.2020.3045927. URL https://ieeexplore.ieee.org/document/9300226/.

N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

M. Simaan and J. B. Cruz Jr. On the stackelberg strategy in nonzero-sum games. *Journal of Optimization Theory and Applications*, 11(5):533–555, 1973.

R. Solanki, M. Bhange, U. Aïvodji, and E. Creager. Crowding out the noise: Algorithmic collective action under differential privacy. *arXiv preprint arXiv:2505.05707*, 2025.

Z. Tian, J. Liu, J. Li, X. Cao, R. Jia, and K. Ren. Private Data Valuation and Fair Payment in Data Marketplaces. URL http://arxiv.org/abs/2210.08723.

J. T. Wang and R. Jia. Data Banzhaf: A Robust Data Valuation Framework for Machine Learning. URL http://arxiv.org/abs/2205.15466.

N. Wang, X. Xiao, Y. Yang, J. Zhao, S. C. Hui, H. Shin, J. Shin, and G. Yu. Collecting and analyzing multidimensional data with local differential privacy. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 638–649. IEEE, 2019.

H. Xia, J. Liu, J. Lou, Z. Qin, K. Ren, Y. Cao, and L. Xiong. Equitable Data Valuation Meets the Right to Be Forgotten in Model Markets. 16(11):3349–3362. ISSN 2150-8097. doi: 10.14778/3611479.3611531. URL https://dl.acm.org/doi/10.14778/3611479.3611531.

T. Yan and A. D. Procaccia. If You Like Shapley Then You'll Love the Core. 35(6):5751–5759. ISSN 2374-3468. doi: 10.1609/aaai.v35i6.16721. URL https://ojs.aaai.org/index.php/AAAI/article/view/16721.

J. Yoon, S. Arik, and T. Pfister. Data valuation using reinforcement learning. In *International Conference on Machine Learning*, pages 10842–10851. PMLR, 2020.

A Appendix: MAB Algorithm

Algorithm 1 Budget-Aware UCB for Data Selection

Require:
$\mathcal{D} = \{x_1, \ldots, x_n\}$: Dataset of n data points (arms)
ϵ: Differential privacy budget per query
$T_{\max}$: Maximum number of iterations
U_{target}: Target utility (average positive vote ratio)
c: Exploration coefficient
α: Learning rate
ε: Small constant for numerical stability
$B^{\text{MAX}}(a)$: Maximum privacy budget per point
NOISYRELEASE(x_a, ϵ): Returns a noisy version of x_a
KNNUTILITY($\mathcal{Z}$; $\{\text{center}_a\}, k$): Average positive votes ratio over validation set $\mathcal{Z}$ with current centers

Initialize:
for $a = 1$ **to** n **do**
  $Q(a) \leftarrow 0$, $N(a) \leftarrow 0$, $B^{\text{re}}(a) \leftarrow B^{\text{MAX}}(a)$
  $\text{center}_a \leftarrow 0$
end for
$U \leftarrow \text{KNNUTILITY}(\mathcal{Z}; \{\text{center}_a\}, k)$
$t \leftarrow 1$
while $t \leq T_{\max}$ **and** $U < U_{\text{target}}$ **do**
  for $a = 1$ **to** n **do**
    if $B^{\text{re}}(a) > 0$ **then**
      $\text{UCB}(a) \leftarrow Q(a) + c \cdot \sqrt{\frac{1}{N(a)+\varepsilon}} \cdot \frac{B^{\text{re}}(a)}{B^{\text{MAX}}(a)}$
    else
      $\text{UCB}(a) \leftarrow -\infty$
    end if
  end for
  $A_t \leftarrow \arg\max_a \text{UCB}(a)$
  $\tilde{x}_{A_t} \leftarrow \text{NOISYRELEASE}(x_{A_t}, \epsilon)$
  $\text{center}_{A_t} \leftarrow \frac{\text{center}_{A_t} \cdot N(A_t) + \tilde{x}_{A_t}}{N(A_t)+1}$
  $U_{\text{new}} \leftarrow \text{KNNUTILITY}(\mathcal{Z}; \{\text{center}_a\}, k)$
  $R_t \leftarrow U_{\text{new}} - U$
  $U \leftarrow U_{\text{new}}$
  $Q(A_t) \leftarrow Q(A_t) + \alpha \cdot (R_t - Q(A_t))$
  $N(A_t) \leftarrow N(A_t) + 1$
  $B^{\text{re}}(A_t) \leftarrow B^{\text{re}}(A_t) - \epsilon$
  $t \leftarrow t + 1$
end while
return $\{\text{center}_a\}, U$

B Appendix: The effect of Laplacian noise

We have allocated a privacy budget of $\epsilon = 1$ per feature, resulting in a total budget of 1024 per point. To illustrate how fidelity evolves under iterative disclosure, we report two complementary measures. The first tracks how close denoised representations remain to the original data points, while the second measures the correlation between original Shapley values and those computed on denoised representations. Together, they provide evidence that averaging across noisy samples progressively restores signal.

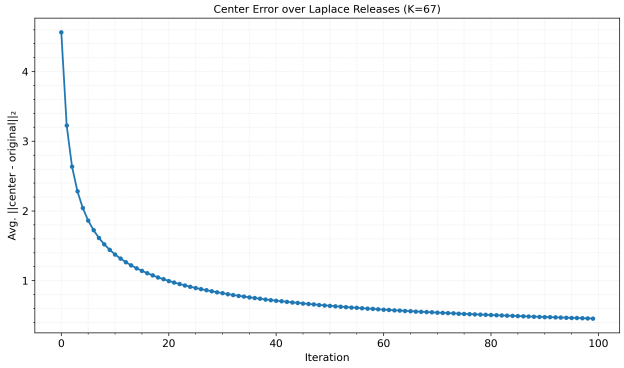

Figure 6: Average ℓ_2 distance between original points and their denoised centers as a function of iterations. Error drops sharply early on with diminishing returns over time.

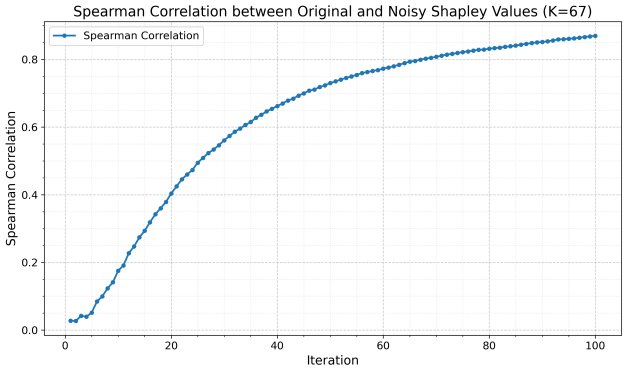

Figure 7: Spearman correlation between original Shapley values and those computed on denoised centers over iterations.

C Appendix: Experimental setup

Dataset and embeddings. We used the Yelp dataset Asghar [2016], which contains text reviews with helpfulness votes. Reviews were encoded with two sentence embedding models: `sentence-transformers/all-mpnet-base-v2` and `intfloat/multilingual-e5-large-instruct`. Both were chosen for their strong performance on semantic similarity tasks and because they were not trained on Yelp, avoiding leakage. A k-Nearest Neighbors classifier was trained with the number of neighbors k tuned on a validation set. The dataset was split into 8000 training, 1000 validation and 1000 test examples.

Computing resources. All experiments were run locally on a MacBook Pro M4 with 24GB of unified memory. Each acquisition experiment completed within a few hours, with bandit simulations being the most computationally intensive.

D Appendix: Data Shapley Selection strategies

Figure 10 illustrates an alternative acquisition strategy in which the DC first estimates Shapley values through a fixed number of bootstrap iterations before committing to acquire the top-ranked points. This procedure is compared against the exact Shapley values provided by an oracle, which serves as a benchmark but is not available in practice. The results highlight that while estimating Shapley values incurs an additional cost, this cost diminishes as more budget is consumed. However, committing after a fixed iteration number does not reduce the total number of iterations required to reach the target utility, as shown in Figure 11, indicating that early commitment provides no fundamental shortcut in convergence.

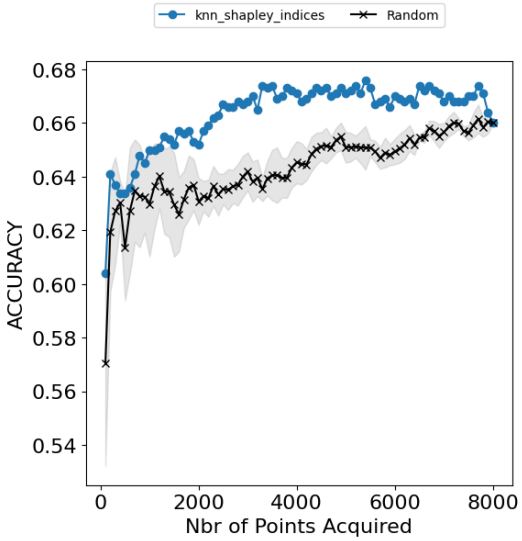

Figure 8: Test accuracy as a function of acquired data points, comparing kNN Shapley-based selection with random selection.

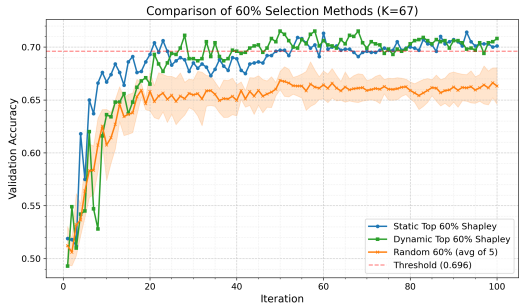

Figure 9: Validation accuracy over iterations for 60% data selection using random, esimated (noisy) data Shapley strategies compared to exact Shapley values given by an oracle. Estimated data Shapley selection reaches the utility target in 25 iterations, significantly outperforming random selection.

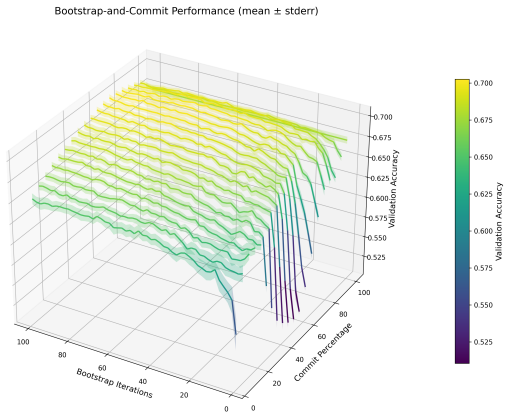

Figure 10: Performance of data Shapley+commitment selection strategy for different combinations of bootstrap and commit parameters. No clear benefit is observed compared to no commitment.

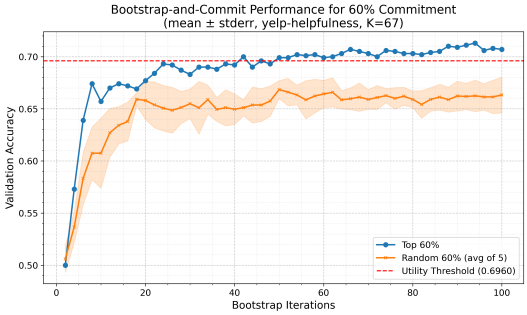

Figure 11: Shapley+commitment strategy on 60% of the data. While utility is eventually reached, performance does not improve over dynamic Shapley.

E Appendix: Correlation between Q-values and Shapley values

We examined the relationship between learned Q-values and Shapley values. In our framework, Q-values are estimates maintained by the multi-armed bandit policy, representing the expected incremental utility of querying a particular data point under the differential privacy budget. Figure 12 shows that the Spearman correlation between these metrics increases with budget but remains moderate overall. This is expected and desired as the objective is for Q-values to be influenced by data utility but not perfectly mimic Shapley values, thereby offering a different valuation with more inclusiveness.

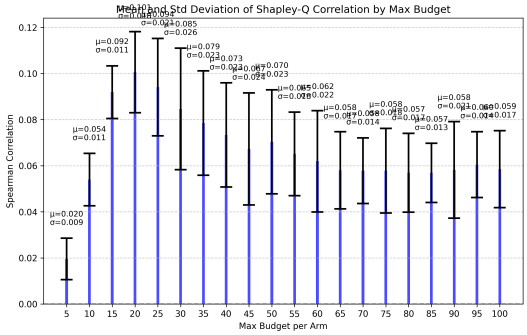

Figure 12: Mean and standard deviation of Spearman correlation between learned Q-values and Shapley values, grouped by maximum budget. Correlation rises with budget but remains modest indicating partially aligned but distinct prioritization.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state that the paper introduces the Information Disclosure Game, a Stackelberg game between a Data Union and Data Consumer, and empirically evaluate disclosure strategies on Yelp review helpfulness prediction. The results and discussion match these claims (see Sections 3 and 4).

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 explicitly mentions that experiments are limited to k-NN for computational efficiency and that extensions to differentiable models is a fundamental future work.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper formalizes the framework (game definition, objectives) but does not present formal theorems or proofs beyond definitions. The contribution is primarily conceptual and empirical.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Appendix C details dataset splits, embedding models and values of k. Algorithm 1 specifies the bandit procedure. Finally, noise mechanisms and parameter settings are described in Section 4 and Appendix B.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The paper uses publicly available Yelp data Asghar [2016], but code has not been released at submission time. Release is planned for the camera-ready version.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Dataset splits, embedding models, k-NN classifier details, noise addition and evaluation metrics are all specified (Section C, Section 4).

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Figures such as 2, 3, and Appendix plots report averages over multiple runs and show standard deviations or shaded areas to indicate variability.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Appendix C specifies that experiments were run on a MacBook Pro M4 with 24GB memory; runtime per experiment is on the order of a few hours.

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The work uses a public dataset (Yelp reviews) with no sensitive personal identifiers, and experiments are consistent with ethical research standards.

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Section 5 discusses inclusiveness and fairer data dividends as positive impacts, while noting challenges in extending to more complex models.

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The work does not release pretrained models or new datasets with misuse risk. It only uses Yelp reviews, which are already public and well studied.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The Yelp dataset is properly cited Asghar [2016], and pretrained embeddings (*e.g.*, all-mpnet-base-v2, multilingual-e5-large-instruct) are credited in Appendix C.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: No new datasets or models are released; only conceptual framework and experiments on existing data.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The work does not involve human subjects or crowdsourcing.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The work does not involve human subjects.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLMs are not part of the core methodology. Only pretrained embeddings (not large models so to speak) are used.