# Private Data Measurements for Decentralized Data Markets

**Charles Lu** LUCHAR@MIT.EDU
*Department of Media Arts & Sciences*
*Massachusetts Institute of Technology*
*Cambridge, MA 02139, USA*

**Mohammad Mohammadi Amiri** MAMIRI@RPI.EDU
*Department of Computer Science*
*Rensselaer Polytechnic Institute*
*Troy, NY 12180, USA*

**Ramesh Raskar** RASKAR@MIT.EDU
*Department of Media Arts & Sciences*
*Massachusetts Institute of Technology*
*Cambridge, MA 02139, USA*

**Reviewed on OpenReview:** *https://openreview.net/forum?id=YdI2TULi8E*

## Abstract

As training data is fundamental to current machine learning, incentivizing data access will be crucial in data-limited application areas such as healthcare. Data markets have been proposed to incentivize greater data access. However, information asymmetry about data value between data owner and data consumer can impede otherwise beneficial transactions from taking place. In this paper, we study data measurements of *relevance* and *diversity* to resolve this information asymmetry. Unlike previous work in data valuation, our heuristic-based approach is cheap to compute, task-agnostic, and does not require centralized data access — properties that are well-suited for a decentralized marketplace setting. We evaluate our approach on several medical imaging datasets and find that relevance measurements are effective at discriminating between data domains, while diversity measures are more useful in selecting sellers that have similar distributions. Code for our experiments is available at `https://github.com/clu5/data-valuation`.

**Keywords:** Data Markets, Data Valuation, Data Measurements

## 1 Introduction

Access to massive amounts of training data has proved foundational to many artificial intelligence (AI) breakthroughs, from earlier deep learning models in computer vision to the current paradigm of large language models (Sun et al., 2017; Kaplan et al., 2020). However, data access is limited in many important application areas, such as healthcare. Furthermore, AI companies face increased scrutiny for their large-scale data collection practices, leading to public backlash and litigation from software developers to journalists. [1] As AI continues to be rapidly developed and widely deployed, more equitable and transparent practices of

1. See `https://stablediffusionlitigation.com`, `https://githubcopilotlitigation.com`, and `https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html` for more context on lawsuits filed against Stable Diffusion, GitHub, OpenAI respectively.

data acquisition need to be developed (Posner and Weyl, 2018; Delacroix and Lawrence, 2019).

Recently, data markets have been promoted to provide compensation frameworks for data contributors, learning to greater data sharing and access (Castro Fernandez, 2023; Agarwal et al., 2019). As the ethical and legal risks of data acquisition become an ever more pressing issue, techniques and platforms to value data assets need to be developed and implemented. The problem of data valuation has already attracted study in areas such as federated learning, information economics, collaborative science (Wang et al., 2020; Sim et al., 2020; Fleckenstein et al., 2023; Belter, 2014).

In this work, we motivate the need for federated data valuation and evaluate one approach using heuristic measures of relevance and diversity. We envision that our approach to data valuation could enable seller selection in situations when third-party brokers cannot be trusted, the modeling task is unknown, or when model training is impractical at valuation time.

## 2 Decentralized Data Markets

Decentralized data markets can address problems with current centralized settings by providing a more equitable and efficient exchange of data resources. Additionally, fully decentralized data markets could realize an alternative model of collective data governance without comprising the rights of individual data owners — redistributing the economic benefits from AI technology to those whose data enables AI research and development (Posner and Weyl, 2018; Duncan, 2023).

### 2.1 Brokerless Data Markets

Current data brokers are highly centralized and have aggregated vast amounts of data, often without a user's knowledge, consent, or compensation (Roderick, 2014; Crain, 2018). This massive centralization of data has led to increased data breaches, privacy erosion, and data misuse. For example, the 2017 Equifax data breach exposed the private records of more than 150 million people, and Google's Project Nightingale allowed Google employees access to non-anonymous medical records of 50 million people without their consent (Zou et al., 2018; Schneble et al., 2020).

In contrast, decentralized data markets may be a more robust, equitable, and efficient approach to data acquisition (Posner and Weyl, 2018; Raskar et al., 2019; Kennedy et al., 2022). In a decentralized marketplace, buyers can transact directly with sellers, bypassing intermediate brokers. This bypassing of data brokers results in lower transaction costs and greater market efficiency by allowing data owners to capture more of the revenue generated from their data. Additionally, compensating data owners may incentivize greater data access from a wider and more diverse range of data producers, in turn attracting more market participants. A greater number of participants in the market would increase price transparency and internalize externalities such as privacy risks and data breaches (Posner and Weyl, 2018).

### 2.2 Federated Data Valuation

A survey of data market participants found that current estimating data value is prohibitively time-consuming and one of the biggest sources of friction (Kennedy et al., 2022). Additionally, most current work in data valuation, such as Data Shapley (Ghorbani and Zou, 2019), assumes a centralized setting where all data is fully accessible to train AI models. Besides being computationally expensive, these data valuation methods also assume a particular modeling objective, which may be challenging when specifying apriori for each data buyer.

In decentralized settings, a data seller would not permit buyers' data access to estimate its value before payment since data is easily copied. However, a buyer would be reluctant to pay a fair price for data if they cannot be assured of its value. Therefore, a fundamental information asymmetry arises between data buyers and sellers, closely related to Arrow's Information Paradox (Arrow, 1972), that results in increased search costs and fewer transactions taking place. Thus, new methods of data appraisal for the decentralized setting with only limited data access need to be developed (Chen et al., 2023).

To allow a buyer to search for the most promising sellers in a decentralized marketplace, we evaluate heuristic data measurements, which have the advantage of being computationally cheap to compute, task-agnostic, and only require indirect data access. Many different data measurements have been developed to quantify intrinsic, task-agnostic characteristics (Mitchell et al., 2022; Lai et al., 2020; Lee et al., 2006). Data measurements can be general-purpose, such as central tendency (e.g., mean, median) and "distance" (e.g., Euclidean distance, KL divergence) or modality-specific, such as Fréchet Inception Distance (Alfarra et al., 2022) and lexical diversity (Jarvis, 2013). Recently, Amiri et al. (2023) proposed to use conditional diversity and relevance measurements to value data without requiring model training or validation data evaluation. We extend their work by evaluating several other definitions of diversity and relevance in the context of private and federated data valuation on medical imaging datasets.

## 3 Private Data Measurements

We adopt several proposed definitions of diversity and relevance measures in our decentralized data marketplace setting. Figure 1 shows an abstract representation of the data measurement from the buyer's perspective, which we describe more formally below.

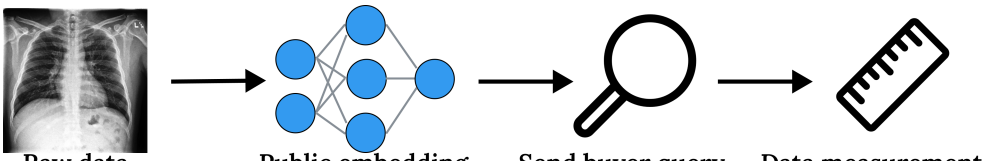



Figure 1: **Overview of our data measurements pipeline.** A buyer embeds their data through some embedding model and sends a private query of matrix projections to each seller. Each seller responds with data measurements that allow the buyer to compare and transact with sellers that have the most relevant data.

### 3.1 Marketplace Description

Suppose that each data buyer has a utility function that comes from a set of specified utility functions that depend on the seller's data and test data, $u : \mathcal{X}^{\text{seller}} \times \mathcal{X}^{\text{test}} \to \mathbb{R}^+,\ u \in \mathcal{U}$. For example, some buyers may care about empirical risk, $u_{\text{risk}} \triangleq \mathbb{E}\left[\ell(\hat{h}(\mathbf{X}^{\text{test}}, \mathbf{Y}^{\text{test}})\right]$, where $\hat{h} : \mathcal{X} \to \mathcal{Y}$ is a model trained on the seller's data and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is some loss function. However, other buyers may have other utility functions, such as finding the most similar data point to the mean test data $u_{\text{sim}} \triangleq \min_{X_i \in \mathbf{X}^{\text{seller}}} d(\bar{X}^{\text{test}}, X_i)$, where $d$ is some distance metric $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Therefore, the data value $V : \mathcal{X} \to \mathbb{R}^+$ of a particular seller's dataset $\mathbf{X}_j^{\text{seller}}$ should correlate to high utility across multiple utility functions under consideration for all potential buyers:

$$V(\mathbf{X}_j^{\text{seller}}) = \mathbb{E}_{u \sim \mathcal{U}}\left[u\left(\mathbf{X}_j^{\text{seller}}, \mathbf{X}^{\text{test}}\right)\right]. \tag{1}$$

However, explicitly calculating Eq. 1 is computationally intractable for large datasets or many utility functions, even in centralized settings.

In the decentralized setting, we cannot directly access $\mathbf{X}^{\text{seller}}$ or $\mathbf{X}^{\text{test}}$ for privacy (see Section 2.1). Instead, we hope to approximate Eq. 1 with private data measurement heuristics $\mu$ that should correlate with the buyer's utility function. Optionally, the seller can preprocess their data to reduce dimensionality $\widetilde{\mathbf{X}}^{\text{seller}} = f(\mathbf{X}^{\text{seller}})$ where $f : \mathcal{X} \to \widetilde{\mathcal{X}}$ is some embedding function. For example, if $X$ is a h × w image, $X \in \mathcal{X} = \mathbb{R}^{\text{h}\times\text{w}}$, then $\widetilde{X} \in \mathbb{R}^{\text{d}}$ may be a d-dimensional embedding of $X$.

Additionally, we assume that the buyer has a small number of IID data $\mathbf{X}^{\text{buyer}} \in \mathcal{X}^{\text{n}}$ that can be used to form a k-dimensional data query $\mathbf{Q}$, that can be communicated to each seller. Then, the data measurements heuristics, $\mu \in \mathbb{R}^2$, diversity and relevance, can be computed on the $j$-seller for the $i$-th buyer by:

$$\mu_{ij} = \left\{\text{Rel}\left(\mathbf{Q}_i^{\text{buyer}}, \widetilde{\mathbf{X}}_j^{\text{seller}}\right),\ \text{Div}\left(\mathbf{Q}_i^{\text{buyer}}, \widetilde{\mathbf{X}}_j^{\text{seller}}\right)\right\}, \tag{2}$$

where several possible definitions of Rel and Div are considered in Section 3.3.

We choose $g$ to be k principal directions obtained from PCA decomposition on $\mathbf{X}^{\text{buyer}}$ embedded through the same pretrained model as the seller's data $f : \mathbb{R}^{\text{n}\times\text{d}'} \to \mathbb{R}^{\text{n}\times\text{d}},\ d' \gg d$:

$$\mathbf{Q}^{\text{buyer}} = \pi_k\left(f\left(X^{\text{buyer}}\right)\right), \tag{3}$$

where $\pi_k : \mathbb{R}^{n\times d} \to \mathbb{R}^{k\times d}$ computes the $k$ principal directions.

Then, we choose the seller preprocessing function $f$ to be

$$\widetilde{\mathbf{X}}^{\text{seller}} = f\left(\mathbf{X}^{\text{seller}}\right)\mathbf{Q}^\top, \tag{4}$$

Additionally, differential privacy could be incorporated with the Gaussian mechanism $M$ to provide $(\epsilon, \delta)$-privacy Dwork et al. (2014) by adding noise generated from a Normal distribution $\mathcal{N}(0, \sigma^2 I)$, where the variance, $\sigma^2$, is chosen to be $\frac{\Delta^2 \cdot 2\ln(1.25/\delta)}{\epsilon^2}$ where $\Delta$ is the L2 sensitivity of the query[2]. This protects the privacy of each seller sample against membership inference and reconstruction attacks at the cost of noisier data measurements (Rahman et al., 2018; Dwork et al., 2017).

2. In differential privacy, $\epsilon$ is the desired privacy budget and $\delta$ is the probability that the privacy budget is violated. These parameters should be chosen with respect to the privacy-utility trade-off in context.

### 3.2 Advantages of Federated Data Measurements

In Section 3.1, we described the decentralized data measurements. We now make some remarks on measurement misreporting, privacy, and embeddings before describing specific definitions of diversity and relevance measurements.

**Preventing Data Measurement Misreporting.** One potential issue with using federated data measurements is malicious misreporting of relevance and diversity by the sellers. A strategic seller could artificially inflate the data measurement value to appear more relevant and diverse to a buyer. To counteract this, a buyer could send multiple queries containing "dummy" directions that may be random or adversarial chosen in addition to the true principal directions. Thus, the buyer would penalize sellers with large data measurements in these "fake" directions while upweighting sellers with high value in the "real" directions. This will incentivize the sellers to honestly report their true data measurements as they do not know which directions are real or fake. Sending additional queries will increase communication overhead but this may be tolerable since each query is cheap — being only a $k \times d$ matrix, where $k \ll n$; each of our queries is $10 \times 512$ floats in our experiments. Each seller's response contains only a single scalar for diversity measurements and a n-dimensional vector for relevance.

**Privacy-preserving Communication:** In most cases, a seller only communicates a scalar response, which protects against data copying while still allowing the buyer to compare sellers based on unique information. The seller can also protect against membership set attacks using differential privacy. On the other hand, the buyer is protected against data reconstruction attacks since only the principal directions are communicated to the seller. Without the magnitude component of the variance (principal components), the seller cannot accurately reconstruct the buyer's data.

**Linear representation space.** Using the embedding function, $f$, allows high-dimensional, multi-modal data to be represented in a shared, lower-dimensional space, which is more amendable to linear separable techniques such as PCA (Jacot et al., 2018). For many practical applications, instead of directly measuring the raw data space (e.g. text data), the buyer and seller can both embed their data into the same representation space (e.g. word embeddings) (Pennington et al., 2014; Mikolov et al., 2015). This embedding step can be precomputed to save computation for each buyer and seller.

### 3.3 Diversity and Relevance

For our approach, we focus on relevance and diversity as two fundamental heuristic measures of data value. We consider several existing definitions of relevance and diversity for our market setting.

### 3.4 Relevance Measures

Relevance should capture some relative notion of similarity between the buyer and seller. For example, if the buyer has chest X-ray (CXR) images of African-American patients, then a seller with CXR images of the same demographic would be more relevant CXR from a different demographic. Likewise, CXR data would be more relevant than MRI data or photography images.

One ubiquitous notion of distance is **L2 distance**. Specifically, we consider negative L2 of the mean projected vectors between buyer and seller: $\left\|\bar{X}^{\text{buyer}} - \bar{X}^{\text{seller}}\right\|_2$, where $\bar{X} \triangleq \frac{1}{k}\sum_{i=1}^{k} \pi_k X_i$ is the mean vector and $\pi$ is the buyer's projection operator. Another common measure of relevance is **cosine similarity**, which we again compute between mean projected vectors: $\frac{\bar{X}^{\text{buyer}} \cdot \bar{X}^{\text{seller}}}{\left\|\bar{X}^{\text{buyer}}\right\|_2 \left\|\bar{X}^{\text{seller}}\right\|_2}$. Lastly, we consider similarity measure proposed by Amiri et al. (2023), defined by the geometric mean of the overlap between the buyer's and seller's principal components: $\sqrt[k]{\prod_{i=1}^{k} \min\left(\lambda_i^{\text{buyer}}, \lambda_i^{\text{seller}}\right) / \max\left(\lambda_i^{\text{buyer}}, \lambda_i^{\text{seller}}\right)}$, where $\lambda$ is the magnitude of the principal components of the projected covariance matrix $\pi(X^{\text{seller}})^\top \pi(X^{\text{seller}})$. One downside to this method is that both the eigenvalues and eigenvectors must be communicated, which could allow approximate reconstructions of the buyer data.

### 3.5 Diversity Measures

Diversity measures should capture the inherent heterogeneity or redundancy within a dataset. Intuitively, data with greater diversity should correspond to better generalization and robustness during test time prediction. If there is insufficient diversity in the training set, the model may overfit and have poor generalization on unseen test data (Xu et al., 2021; Friedman and Dieng, 2022). For example, a seller with X-ray images from 100 unique patients would typically be considered more valuable than 100 X-rays from a single patient. Xu et al. (2021) proposed to measure data diversity in a validation-free manner using the **volume of the gram matrix**. We consider the projected version of volume: $\sqrt{\det\left(\pi(X^{\text{seller}})^\top \pi(X^{\text{seller}})\right)}$.

In the context of natural language processing, Lai et al. (2020) considered data diversity as the **dispersion of the features** of the data, defined as the geometric sum of the standard deviation of each feature: $\left(\prod_{i=1}^{d} \sigma_i\right)^{\frac{1}{d}}$, where $\sigma_i$ is the standard deviation of feature $i$ of the projected seller data $\pi(X^{\text{seller}})$.

A third definition of diversity is the **Vendi score** (Friedman and Dieng, 2022) defined as $\exp\left(-\sum_{c=1}^{C} \lambda_c \log \lambda_c\right)$, where $\lambda$ is eigenvalue of the projected covariance matrix $\frac{1}{m^{\text{seller}}} \pi(X^{\text{seller}})^\top \pi(X^{\text{seller}})$ and $C$ is the number of directions.

## 4 Experiments

We evaluate our approach on the MedMNIST medical imaging benchmark (Yang et al., 2023). We precompute embeddings with CLIP ViT-B/16 to embed each image into a 512-dimensional vector (Radford et al., 2021). For the buyer query, we project the seller's data onto the first 10 principal directions of the buyer's data. See Appendix A for more details on the experimental setup.

**Qualitative Evaluation:** In Figure 2, we compare combinations of diversity and relevance measurements. The buyer used 100 embedded images from the BloodMNIST dataset to create the query, and we compared sellers with data from four other domains: MNIST, MedMNIST, CIFAR-10, and a noisy version of BloodMNIST (see Figure 4). For each of the four domains, we start with 10,000 in-domain (ID) BloodMNIST data points and gradually

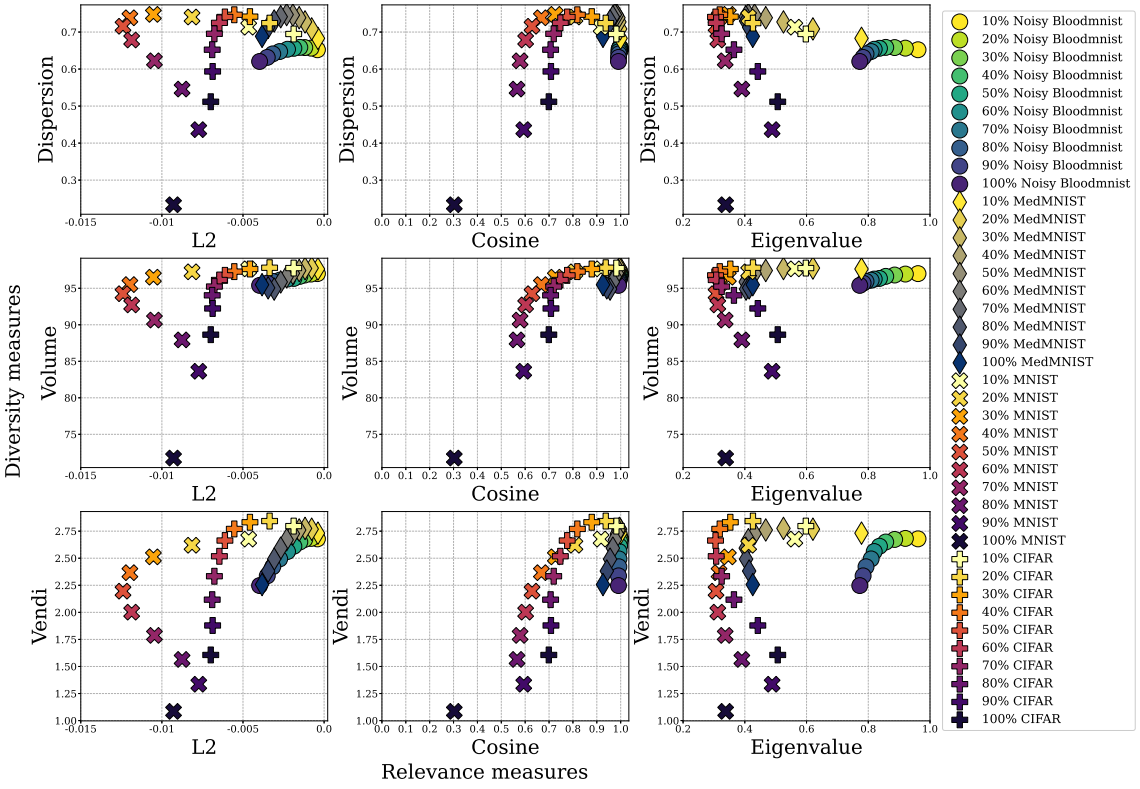



Figure 2: **Relevance measurements discriminate between in-domain and out-of-domain data.** We plot relevance and diversity when sellers with 10,000 in-domain BloodMNIST data have increasing proportion replaced with out-of-domain data (Noisy Bloodmnist, MedMNIST, MNIST, CIFAR). We see that relevance measures such as cosine similarity discriminate out-of-domain data better than diversity measures.

replace an increasing proportion with out-of-domain (OD) data starting from 10% OD / 90% ID until 100% OD data.

Ideally, sellers with more ID data and less OD data should have higher data measurement values. Additionally, sellers with more relevant OD data (noisy BloodMNIST) should have higher values than irrelevant OD data (CIFAR and MNIST). We observe that diversity measures perform poorly at discriminating between ID and OD sellers. In contrast, relevance measures perform better at discriminating between domains. In particular, we find that cosine similarity assigns sellers with mostly ID data high relevance values while assigning sellers with mostly OD data much lower values.

See Appendix B.1 for results on other MedMNIST datasets. In the Appendix, we evaluate the effect of other data characteristics, such as the number of unique classes (Figure B.2) and noise corruptions ( Figure B.4).

**Quantitative Evaluation:** Since our data valuation framework is task agnostic, we evaluate the correlation between data measurements and test accuracy for three prediction

Table 1: **Kendall rank correlation between relevance and diversity measurements and average test accuracy.** In general, we find position correlations between diversity measurements and test prediction performance across MedMNIST datasets for binary classification, multiclass classification, and clustering prediction tasks.

| Data Measurement | | Blood | Derma | Retina | Path | Tissue | Organ | Average |
|---|---|---|---|---|---|---|---|---|
| Relevance | L2 Dist. | 0.04 | 0.24 | 0.24 | 0.06 | 0.17 | -0.03 | 0.12 |
| | Cosine Sim. | 0.00 | 0.26 | 0.25 | 0.05 | 0.11 | 0.09 | 0.13 |
| | Eigenvalue | 0.01 | 0.33 | 0.24 | 0.03 | 0.20 | -0.04 | 0.13 |
| Diversity | Dispersion | 0.09 | 0.32 | 0.25 | 0.12 | 0.32 | 0.17 | 0.21 |
| | Volume | 0.08 | **0.34** | **0.29** | **0.22** | **0.33** | **0.23** | **0.25** |
| | Vendi Score | **0.12** | 0.33 | 0.24 | 0.19 | 0.28 | 0.23 | 0.23 |

tasks: binary classification, multiclass classification, and clustering. For each dataset, we sample data from a subset of classes for the buyer's data query and a held-out test set. For each of the 500 sellers, we introduce class heterogeneity by sampling classes from a Dirichlet distribution as typically done in non-IID federated learning experiments Li et al. (2021). We train a model using the seller's data, evaluate performance on the buyer's test set, and calculate the rank correlation of each data measurement and test set performance. We report correlations averaged over prediction tasks; this can be thought of as a crude approximation of Eq. 1.

Intuitively, we expect that sellers with more data from the same classes as the buyer will result in better-performing models and should have a higher data measurement value. Indeed, we find a moderate to strong positive correlation between relevance and diversity measurements and test accuracy for several datasets in Figure 1. Interestingly, we find that diversity measures, such as volume, tend to be more correlated with test accuracy than relevance measures (strongest correlations shown in Figure 11). See Table 2 for an expanded table across prediction tasks.

## 5 Conclusion

In this paper, we motivated the need for new data valuation approaches for decentralized data marketplaces. Importantly, these techniques should be federated and privacy-preserving for domains such as healthcare, where medical data is sensitive, and data value must be estimated without direct access to the seller's data. In this work, we evaluated several definitions of relevance and diversity on multiple benchmark medical imaging datasets. We found that relevance is more useful in differentiating between different domains (medical vs. non-medical data). In contrast, diversity measures are more correlated with prediction accuracy and may be more suited when sellers have more data from similar distributions. Developing robust and generalizable data measurements will be important in resolving information asymmetries between data buyers and sellers to facilitate decentralized data transactions. However, further work is needed to validate our approach to data valuation and seller selection in the context of data markets on other data modalities and domains.

## Broader Impact Statement

We believe that AI developers must reconcile important ethical questions regarding data acquisition in current AI development. Class-action lawsuits have been filed against several AI companies for their data collection practices, raising questions about data compensation and consent from data owners. Current data acquisition norms may actively discourage further data sharing, which can hamper the progress and impact of AI, especially in data-limited domains such as healthcare.

Our work advances technical challenges in operationalizing data marketplaces, which promise an alternative model of data acquisition, compensation, and ownership. In addition, our approach does not require centralized access to seller data, which contrasts with previous work in data valuation and, therefore, better protects the privacy of data producers during the data valuation stage. Centralizing all data with the broker can result in undesirable privacy and security risks, such as data breaches. In contrast, decentralized data marketplaces may be more robust and transparent. Bypassing intermediate data brokers will enhance privacy and increase market efficiency. Transaction costs can be reduced, and revenue can be directly captured by data producers. This will enable a greater number of data transactions and sustain more types of markets.

## References

Anish Agarwal, Munther Dahleh, and Tuhin Sarkar. A marketplace for data: An algorithmic solution. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 701–726, 2019.

Motasem Alfarra, Juan C Pérez, Anna Frühstück, Philip HS Torr, Peter Wonka, and Bernard Ghanem. On the robustness of quality measures for gans. In *European Conference on Computer Vision*, pages 18–33. Springer, 2022.

Mohammad Mohammadi Amiri, Frederic Berdoz, and Ramesh Raskar. Fundamentals of task-agnostic data valuation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9226–9234, 2023.

Kenneth Joseph Arrow. *Economic welfare and the allocation of resources for invention*. Springer, 1972.

Christopher W Belter. Measuring the value of research data: A citation analysis of oceanographic data sets. *PLoS One*, 9(3):e92590, 2014.

Raul Castro Fernandez. Data-sharing markets: Model, protocol, and algorithms to incentivize the formation of data-sharing consortia. *Proceedings of the ACM on Management of Data*, 1(2):1–25, 2023.

Lingjiao Chen, Bilge Acun, Newsha Ardalani, Yifan Sun, Feiyang Kang, Hanrui Lyu, Yongchan Kwon, Ruoxi Jia, Carole-Jean Wu, Matei Zaharia, et al. Data acquisition: A new frontier in data-centric ai. *arXiv preprint arXiv:2311.13712*, 2023.

Matthew Crain. The limits of transparency: Data brokers and commodification. *New Media & Society*, 20(1):88–104, 2018. doi: 10.1177/1461444816657096. URL `https://doi.org/10.1177/1461444816657096`.

Sylvie Delacroix and Neil 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.

Jamie Duncan. Data protection beyond data rights: Governing data production through collective intermediaries. *Internet Policy Review*, 12(3):1–22, 2023.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Cynthia Dwork, Adam Smith, Thomas Steinke, and Jonathan Ullman. Exposed! a survey of attacks on private data. *Annual Review of Statistics and Its Application*, 4:61–84, 2017.

Mike Fleckenstein, Ali Obaidi, and Nektaria Tryfona. A review of data valuation approaches and building and scoring a data valuation model. 2023.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022.

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

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Scott Jarvis. Capturing the diversity in lexical diversity. *Language Learning*, 63:87–106, 2013.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Javen Kennedy, Pranav Subramaniam, Sainyam Galhotra, and Raul Castro Fernandez. Revisiting online data markets in 2022: A seller and buyer perspective. *ACM SIGMOD Record*, 51(3):30–37, 2022.

Yi-An Lai, Xuan Zhu, Yi Zhang, and Mona Diab. Diversity, density, and homogeneity: Quantitative characteristic metrics for text collections. *arXiv preprint arXiv:2003.08529*, 2020.

Yang W Lee, Leo L Pipino, James D Funk, and Richard Y Wang. *Journey to data quality*. The MIT Press, 2006.

Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study, 2021.

Tomas Mikolov, Kai Chen, Gregory S Corrado, and Jeffrey A Dean. Computing numeric representations of words in a high-dimensional space, May 19 2015. US Patent 9,037,464.

Margaret Mitchell, Alexandra Sasha Luccioni, Nathan Lambert, Marissa Gerchick, Angelina McMillan-Major, Ezinwanne Ozoani, Nazneen Rajani, Tristan Thrush, Yacine Jernite, and Douwe Kiela. Measuring data. *arXiv preprint arXiv:2212.05129*, 2022.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Eric Posner and Eric Weyl. *Radical markets: Uprooting capitalism and democracy for a just society*. Princeton University Press, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Md Atiqur Rahman, Tanzila Rahman, Robert Laganière, Noman Mohammed, and Yang Wang. Membership inference attack against differentially private deep learning model. *Trans. Data Priv.*, 11(1):61–79, 2018.

Ramesh Raskar, Praneeth Vepakomma, Tristan Swedish, and Aalekh Sharan. Data markets to support ai for all: Pricing, valuation and governance. *arXiv preprint arXiv:1905.06462*, 2019.

Leanne Roderick. Discipline and power in the digital age: The case of the us consumer data broker industry. *Critical Sociology*, 40(5):729–746, 2014. doi: 10.1177/0896920513501350. URL `https://doi.org/10.1177/0896920513501350`.

Christophe Olivier Schneble, Bernice Simone Elger, and David Martin Shaw. Google's project nightingale highlights the necessity of data science ethics review. *EMBO molecular medicine*, 12(3):e12053, 2020.

Rachael Hwee Ling Sim, Yehong Zhang, Mun Choon Chan, and Bryan Kian Hsiang Low. Collaborative machine learning with incentive-aware model rewards. In *International conference on machine learning*, pages 8927–8936. PMLR, 2020.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Tianhao Wang, Johannes Rausch, Ce Zhang, Ruoxi Jia, and Dawn Song. A principled approach to data valuation for federated learning. *Federated Learning: Privacy and Incentive*, pages 153–167, 2020.

Xinyi Xu, Zhaoxuan Wu, Chuan Sheng Foo, and Bryan Kian Hsiang Low. Validation free and replication robust volume-based data valuation. *Advances in Neural Information Processing Systems*, 34:10837–10848, 2021.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Hangyu Zhu, Jinjin Xu, Shiqing Liu, and Yaochu Jin. Federated learning on non-iid data: A survey. *Neurocomputing*, 465:371–390, 2021.

Yixin Zou, Abraham H Mhaidli, Austin McCall, and Florian Schaub. " i've got nothing to lose": Consumers' risk perceptions and protective actions after the equifax data breach. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 197–216, 2018.

## Appendix A. Experimental Setup

We use the buyer's data to determine the principal directions to send as the query to each of the sellers. The number of directions was determined by checking a Scree plot of the magnitude of the buyer's eigenvalues.

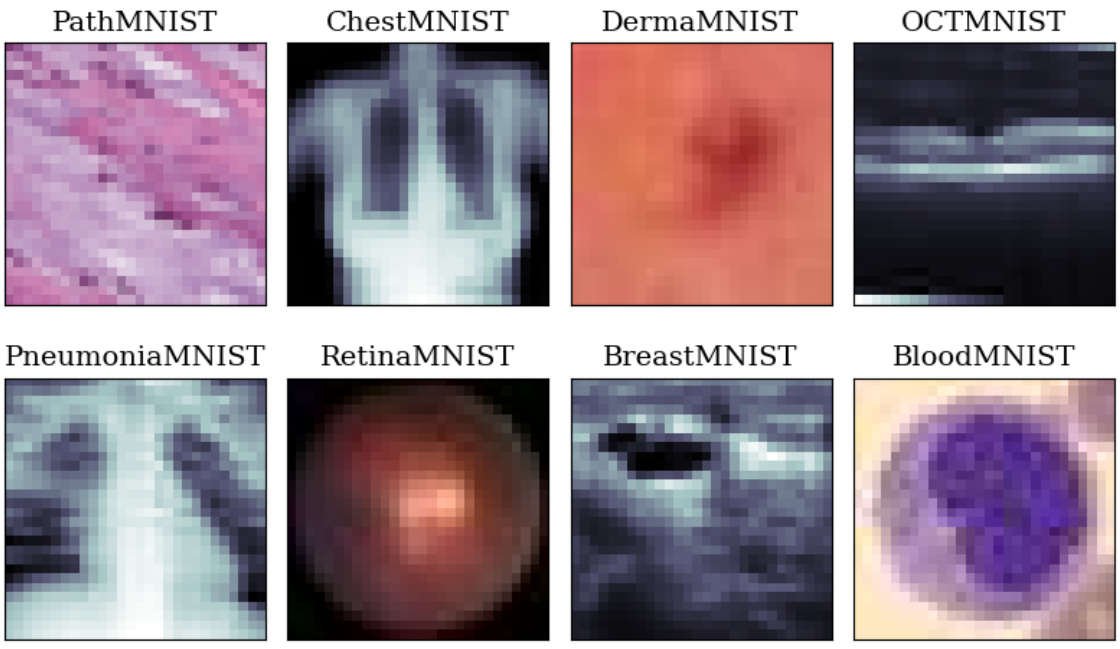



Figure 3: **Example images from datasets in the MedMNIST benchmark.** See `medmnist.com` for more information.

To create the noisy version of the MedMNIST dataset used in the qualitative experiments, we use the following PyTorch code:

```
    f = transforms.Compose([
    transforms.ToTensor(),
    transforms.GaussianBlur(5),
    transforms.RandomResizedCrop(size=(28, 28)),
])
```

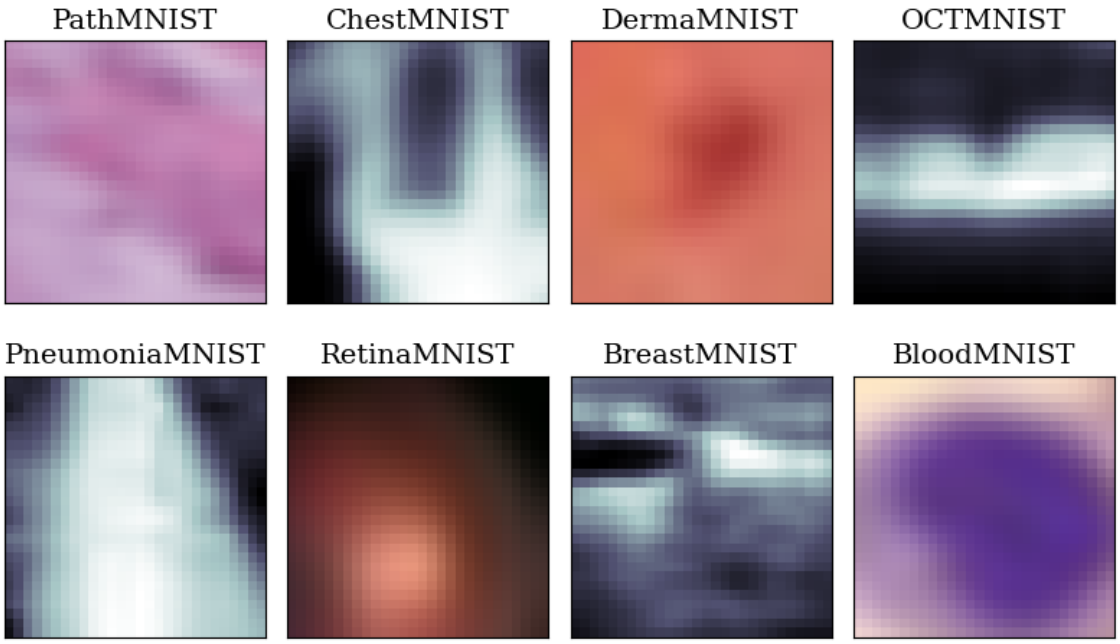



Figure 4: **Noisy images from datasets in the MedMNIST benchmark.** See `medmnist.com` for more information.

In Table 1, we show how correlated each data measure of relevance and diversity is with prediction accuracy in a binary classification task using a subset of the original classes from each MedMNIST dataset. We select the following classes for each dataset class to use as the subset for the buyer and test set in the quantitative experiments, which are treated as the "positive" classes in binary classification and the only classes evaluated in multi-class classification and clustering:

- PathMNIST: 'colorectal adenocarcinoma epithelium', 'cancer-associated stroma', 'normal colon mucosa'
- DERMAMNIST: 'melanoma', 'basal cell carcinoma', 'actinic keratoses and intraepithelial carcinoma'
- RetinaMNIST: '1', '2', '3'
- BloodMNIST: 'neutrophil', 'monocyte', 'lymphocyte'

- TissueMNIST: 'Glomerular endothelial cells', 'Podocytes', 'Proximal Tubule Segments'
- OrganAMNIST: 'lung-left', 'lung-right', 'heart', 'liver'

To introduce heterogeneity between the data sellers, we sample class weights from a Dirichlet probability distribution (initialized from the initial class proportions in each dataset) as typically done in Federated Learning experiments with non-IID data distributions Zhu et al. (2021).

In addition, we use the following number of samples for the buyer and each seller:

- PathMNIST: 100 samples in buyer, 1500 samples per seller
- DERMAMNIST: 100 samples in buyer, 1500 samples per seller
- RetinaMNIST: 100 samples in buyer, 500 samples per seller
- BloodMNIST: 500 samples in buyer, 5000 samples per seller
- TissueMNIST: 1500 samples in buyer, 25000 samples per seller
- OrganAMNIST: 500 samples in buyer, 5000 samples per seller

For each data seller, we train a simple Logistic Regression model on its corresponding dataset of labeled embeddings and evaluate the held-out test set for the task of binary classification of the above-mentioned positive classes for each dataset. For multiclass classification, we train a Random Forest model with 10 trees and a max depth of 5 and evaluate F1 score. For clustering, we fit a K-Means model with the same number of clusters as classes in the buyer and evaluate homogeneity score. Finally, we compute relevance and diversity measures between each seller and the buyer and calculate the Kendall rank correlation between prediction accuracy for the model trained using that seller's dataset and its data measurements.

## Appendix B. Additional Experiments

### B.1 Additional Qualitative Results

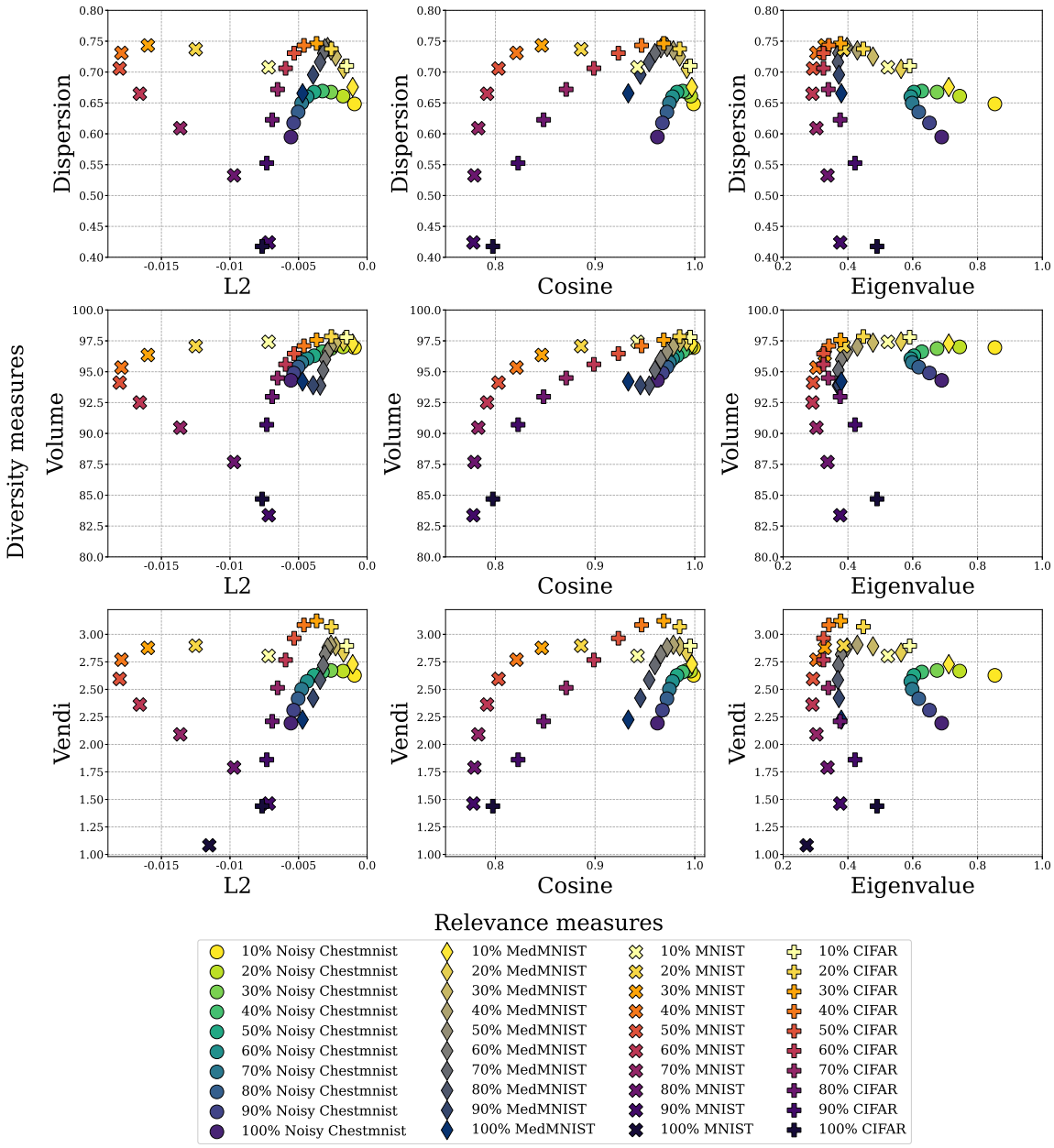



Figure 5: Qualitative assessment of different relevance and diversity measures when the in-domain distribution is the ChestMNIST dataset.

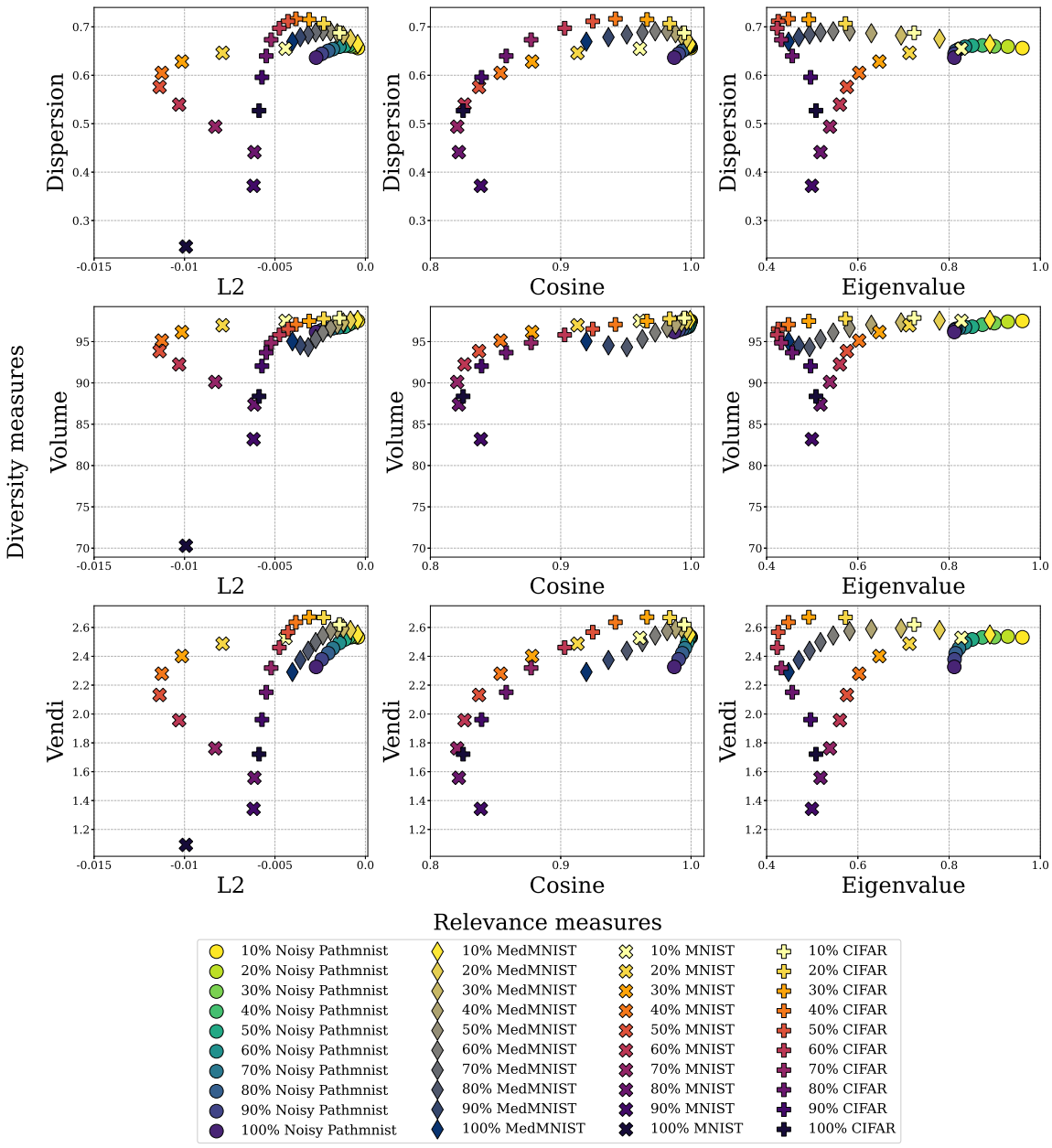



Figure 6: Qualitative assessment of different relevance and diversity measures when the in-domain distribution is the PathMNIST dataset.

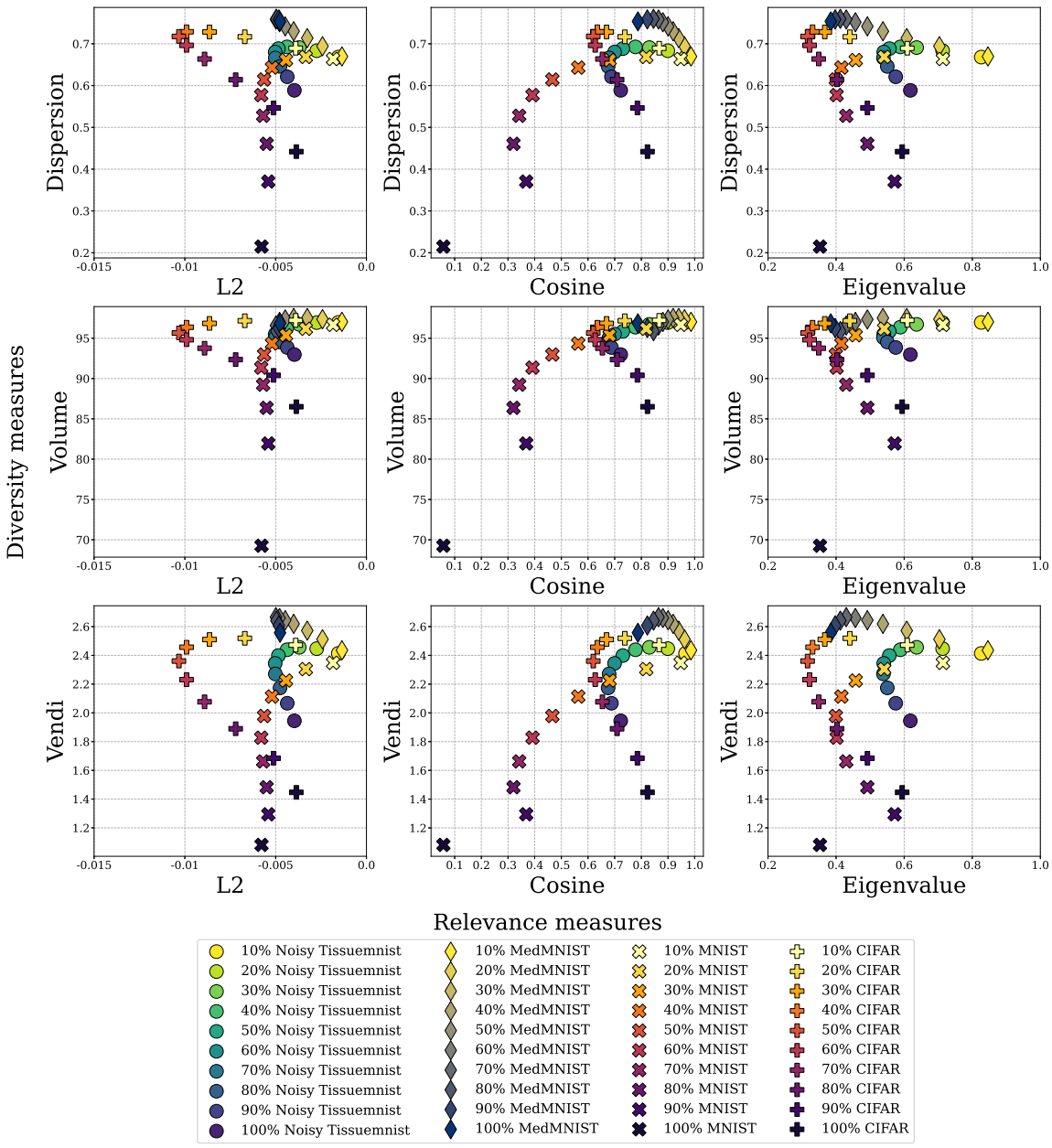
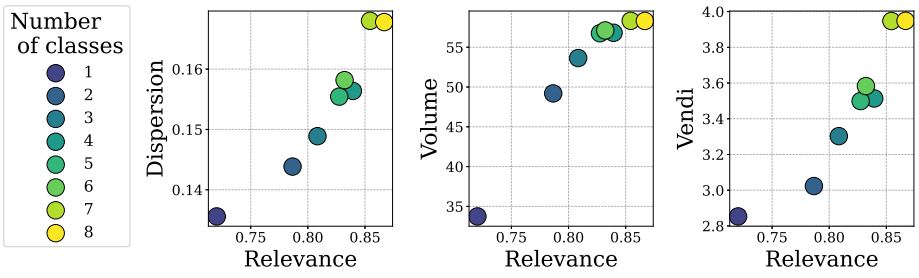



Figure 7: Qualitative assessment of different relevance and diversity measures when the in-domain distribution is the TissueMNIST dataset.

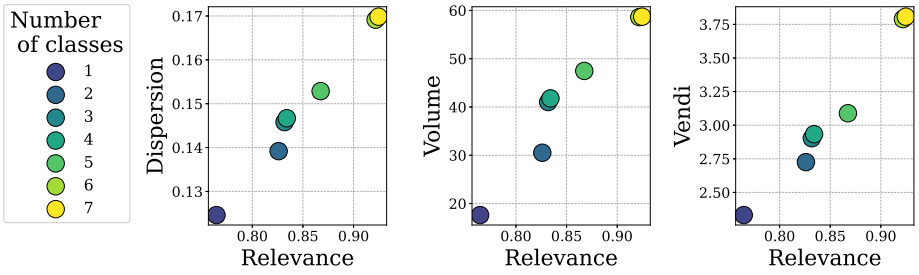
### B.2 Effect of Number of Unique Classes on Diversity

We plot diversity for sellers with data from varying numbers of unique classes. Intuitively, we expect that as the number of classes increases, the diversity also increases, which is reflected across all three diversity measures in the BloodMNIST and DermaMNIST datasets. However, in the PathMNIST and OrganAMNIST, diversity decreases with both the disper-

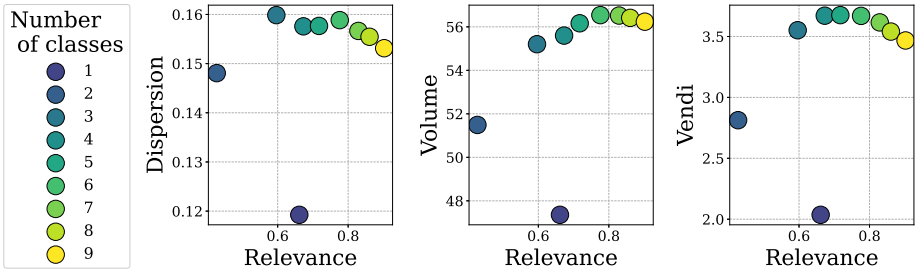
sion and Vendi Score measures for sellers with the highest number of unique classes (this behavior may be related to both the frequency of each class as some MedMNIST datasets can be highly imbalanced). The Volume-based definition of diversity more closely follows this desirable property (middle column) across all datasets.



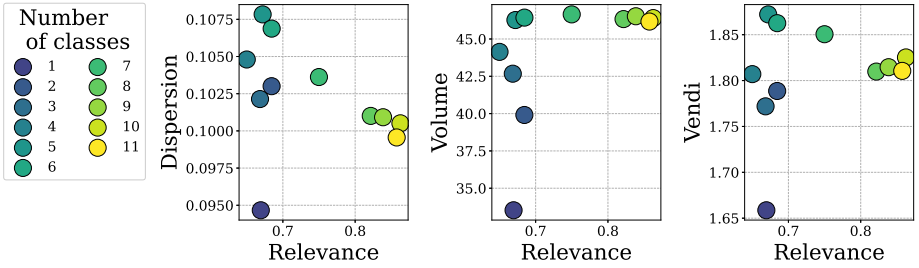

(a) BloodMNIST




(b) DermaMNIST




(c) PathMNIST




(d) OrganAMNIST

Figure 8: **Comparing different diversity measures with varying numbers of unique classes.** The buyer has data from all classes.

### B.3 Evaluating Absolute Diversity.

First, we compare the diversity between MedMNIST datasets and non-medical datasets (CIFAR, MNIST, FashionMNIST) without any buyer projection in Figure 9. All three diversity methods show that CIFAR has the highest absolute diversity, followed by MedMNIST, a combination of 8 different MedMNIST datasets. Intuitively, CIFAR is composed of natural images of animals and vehicles, which contain more variation both on a pixel and semantic level than medical datasets. Interestingly, even FashionMNIST, which consists of grayscale images of clothes, contains slightly more diversity than some of the individual MedMNIST datasets. This may be reflected in the CLIP model's bias, as most of the training data consisted of text-image pairs publicly available on the Internet.

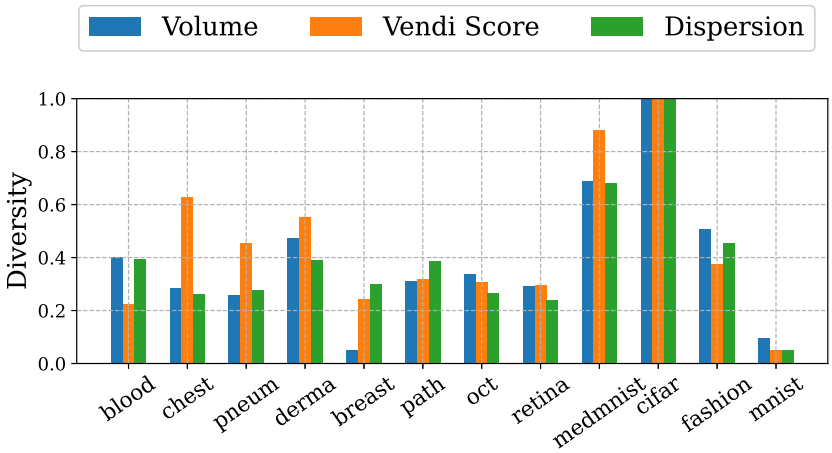
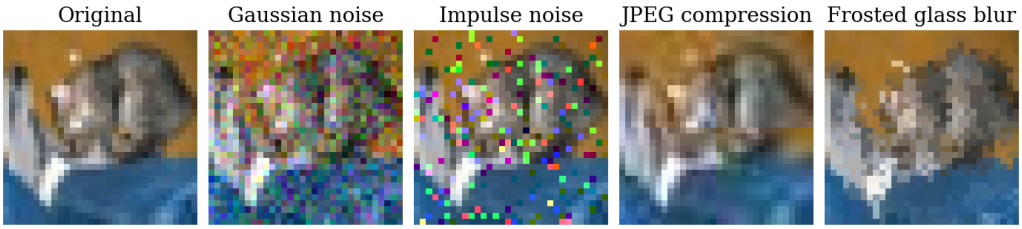
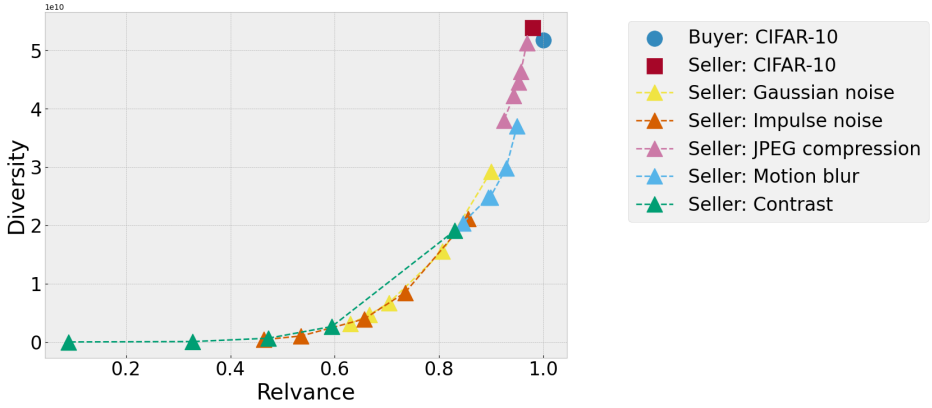



Figure 9: Diversity of 1000 CLIP-embedded samples for each dataset. MedMNIST is a combination of 8 individual medical datasets. Across all three diversity measures, CIFAR has the highest diversity, while MNIST has the lowest diversity. Each diversity measure is min-max scaled between datasets to lie within 0 and 1.

### B.4 Effect of Noise Corruptions on Relevance and Diversity




(a) Different corruptions from CIFAR-C dataset.



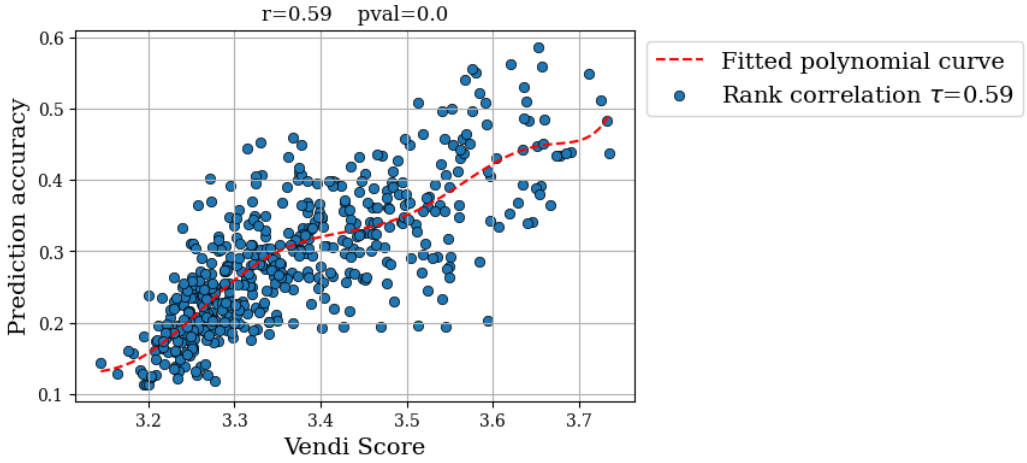

(b)

Figure 10: Effect of Noise Corruptions on Relevance and Diversity.

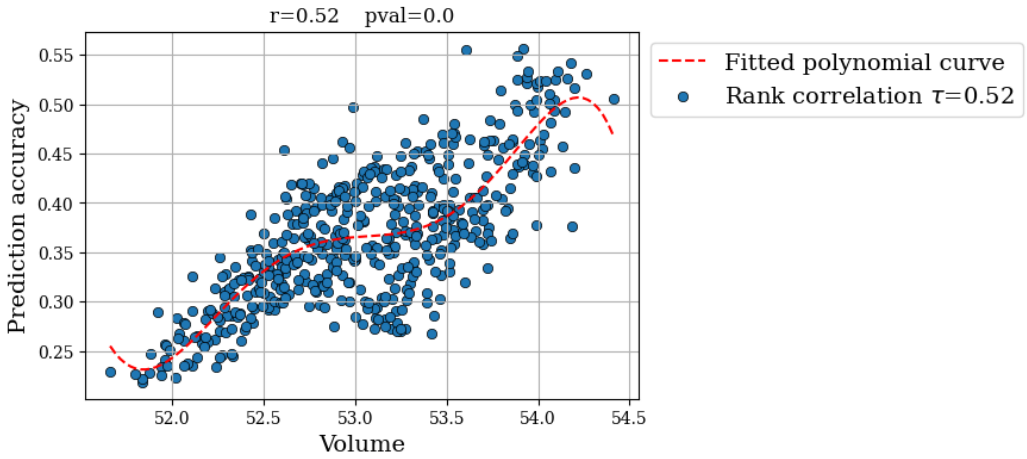
### B.5 Additional Quantitative Results

Table 2: **Correlation between relevance and diversity measurements and test accuracy.** For each seller, we compute data measurements and train an ML model on their data to predict a held-out test set. We report the rank correlation between each data measurement and test accuracy on three prediction tasks. Underlined values denote the highest correlation with prediction accuracy for either the category of relevance or diversity, while bolded values indicate the highest correlated data measure overall. In general, we find that diversity measures are more correlated with prediction performance than measures of relevance.

| Task | Measurement | | Blood | Derma | Retina | Path | Tissue | Organ | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| Binary | Relevance | L2 | 0.23 | 0.21 | 0.26 | 0.02 | 0.06 | 0.00 | 0.13 |
| | | Cosine | 0.19 | 0.22 | 0.28 | 0.02 | -0.02 | 0.02 | 0.12 |
| | | Eigen. | 0.21 | **0.34** | **0.31** | 0.01 | 0.21 | 0.00 | 0.18 |
| | Diversity | Disp. | 0.12 | 0.29 | 0.12 | 0.14 | 0.23 | 0.12 | 0.17 |
| | | Vol. | **0.26** | 0.33 | 0.19 | **0.23** | **0.28** | **0.16** | **0.24** |
| | | Vendi | 0.17 | 0.29 | 0.09 | 0.12 | 0.26 | 0.13 | 0.18 |
| Multiclass | Relevance | L2 | 0.11 | 0.43 | 0.35 | -0.01 | 0.07 | -0.05 | 0.15 |
| | | Cosine | 0.03 | 0.44 | 0.26 | -0.01 | 0.06 | 0.00 | 0.13 |
| | | Eigen. | 0.09 | 0.55 | 0.35 | -0.02 | 0.06 | -0.04 | 0.17 |
| | Diversity | Disp. | 0.08 | 0.57 | 0.48 | 0.17 | **0.29** | **0.08** | 0.28 |
| | | Vol. | **0.11** | 0.57 | **0.52** | **0.24** | 0.24 | 0.07 | **0.29** |
| | | Vendi | 0.03 | **0.59** | 0.47 | 0.17 | 0.15 | 0.06 | 0.25 |
| Clustering | Relevance | L2 | -0.22 | 0.08 | 0.10 | 0.16 | 0.37 | -0.03 | 0.08 |
| | | Cosine | -0.22 | 0.11 | **0.21** | 0.15 | 0.29 | 0.24 | 0.13 |
| | | Eigen. | -0.28 | 0.11 | 0.07 | 0.10 | 0.33 | -0.09 | 0.04 |
| | Diversity | Disp. | 0.07 | 0.10 | 0.16 | 0.06 | 0.44 | 0.30 | 0.19 |
| | | Vol. | -0.12 | **0.12** | 0.17 | 0.18 | **0.46** | 0.47 | 0.21 |
| | | Vendi | **0.17** | 0.11 | 0.15 | **0.27** | 0.42 | **0.51** | **0.27** |




(a) Correlating Vendi score measurements and F1 score for multiclass classification on the DermaMNIST dataset.




(b) Correlating Volume and F1 score for multiclass classification on the RetinaMNIST dataset.

Figure 11: Correlating between data measurements of diversity and prediction accuracy on MedMNIST datasets.