Fundamentals of Task-Agnostic Data Valuation

Mohammad Mohammadi Amiri[1], **Frédéric Berdoz**[2], **Ramesh Raskar**[1]

[1]MIT, Media Lab, 75 Amherst St, Cambridge, MA 02139, USA
[2]EPFL, Lausanne, Switzerland
mamiri@mit.edu, frederic.berdoz@epfl.ch, raskar@mit.edu

Abstract

We study valuing the data of a data owner/seller for a data seeker/buyer. Data valuation is often carried out for a specific task assuming a particular utility metric, such as test accuracy on a validation set, that may not exist in practice. In this work, we focus on task-agnostic data valuation without any validation requirements. The data buyer has access to a limited amount of data (which could be publicly available) and seeks more data samples from a data seller. We formulate the problem as estimating the differences in the statistical properties of the data at the seller with respect to the baseline data available at the buyer. We capture these statistical differences through second moment by measuring *diversity* and *relevance* of the seller's data for the buyer; we estimate these measures through queries to the seller without requesting the raw data. We design the queries with the proposed approach so that the seller is blind to the buyer's raw data and has no knowledge to fabricate responses to the queries to obtain a desired outcome of the diversity and relevance trade-off. We will show through extensive experiments on real tabular and image datasets that the proposed estimates capture the diversity and relevance of the seller's data for the buyer.

Introduction

Data is the main fuel of the modern world enabling artificial intelligence and driving innovation and technological growth. The demand for data has grown substantially, and it is extremely valuable for sectors to acquire high quality data to discover knowledge and improve their products and services. As the demands for data have grown substantially, data products have become valuable assets to purchase and sale. This calls for establishing a data marketplace that connects different parties and facilitates trading data.

A data marketplace mainly includes three components, data sellers, broker, and data buyers; data sellers own the data and share it with the broker in exchange for rewards; data buyers want to acquire data, and broker facilitates trading data. As a valuable resource, it is important to establish a principled method to quantify the worth of the sellers' data and its value for the buyers. This is addressed via data valuation which is the essential component for realization of a fair marketplace for sellers and buyers. Data valuation arises in various applications such as collaborative machine learning (Sim et al.; Tay et al.), federated learning (Song et al.; Richardson et al.), data marketing (Schomm et al.; Muschalle et al.), advertisement (Bergemann et al.; Zheng et al.), recommendation systems (Immorlica et al.; Che et al.), and data sharing (Rasouli et al.; Gradwohl et al.).

Data valuation is carried out either based on "intrinsic" or "extrinsic" factors; intrinsic data valuation is data-driven and based on the quality of dataset (Niu et al.; Raskar et al.), while extrinsic data valuation considers demand-supply and game-theoretic mechanisms (Luong et al.; Zhang et al.). It is a common practice to couple intrinsic data valuation with a utility metric for validation (Ghorbani et al.; Jia et al.), or with a specific machine learning (ML) task (Agarwal et al.; Chen et al.). In particular, for ML applications, data is often valued assuming existence of a validation set using validation accuracy as a metric (Wang et al.; Yan et al.). Also, ML models trained with a target task are used to estimate the value of the data used for training the models (Pei; Liu et al.). On the other hand, extrinsic data valuation techniques consider external factors such as competition and demands (Agarwal et al.; Bimpikis et al.), which requires estimating costumers' demands for products and competitors price levels to price a product (Toni et al.; Cong et al.).

Enforcing a close coupling between intrinsic data valuation and existence of a validation set may not be practical since a validation set that all the parties agree on may not exist, and a particular validation set may not sufficiently represent the data distribution for a learning task (Xu et al.). Furthermore, having a validation set may provide the chance to malicious sellers to modify their datasets to overfit on the validation set. Also, considering a specific ML model/task for data valuation may not be aligned with the interests of all the parties. We instead take a step back and consider an intrinsic data valuation without any validation requirements and before performing any tasks such as training a ML model. We take a step towards addressing the challenge of formulating a model- and task-agnostic intrinsic valuation of data at a seller for a buyer. The authors in (Xu et al.) develop a technique independent of validation based solely on the diversity of seller's data, which captures the variation/dissimilarity across data samples; this provides the same value of data at a seller for all the buyers. However, we believe that diversity of data alone may not be sufficient for data valuation for two reasons. First, performing data valuation independent of the buyers makes it hard to realize the relevance of the seller's data for the buyer. Consider the case when a buyer is interested in health data, such as chest X-ray images, while a seller has a very diverse set of images of animals. Thus, the inherently diverse dataset at the seller is irrelevant for the buyer, and this needs to be captured by data valuation. Additionally, a seller can fabricate data to increase its diversity through, for example, adding random noise.

We focus on an intrinsic task-agnostic data valuation considering the fact that data at each seller has a distinct value for each buyer (Raskar et al.). We measure the value of data at a seller in dependence with the already available data at the buyer, some of which may be publicly available, which stays local at the buyer and is not shared with any parties. This provides a unique valuation of a seller's data for each buyer. We aim to value the data through comparing the statistical properties of the two datasets and formulate the problem as estimating **diversity** and **relevance** of the seller's data for the buyer. We then estimate diversity and relevance by measuring the differences and similarities in the statistical properties of the two datasets through second moment. This is carried out through queries from the buyer to the seller designed such that it is infeasible for the seller to fabricate responses to the queries and manipulate the data to achieve a desired outcome of diversity and relevance pair.

Notations: A multi-variate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is denoted by $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$; $\mathbf{0}_n$ represents an all-zero vector of dimension n; l^2 norm of a vector is denoted by $\|\cdot\|$; $\text{Diag}(\lambda_1, ..., \lambda_d)$ returns a $d \times d$ diagonal matrix with diagonal entries $\lambda_1, ..., \lambda_d$. Cardinality of a set is shown by $|\cdot|$.

Problem Motivation and Formulation

We consider a data marketplace with an arbitrary number of buyers and sellers, assuming that each buyer (she) has access to some data samples[1] and wants to buy extra data from one or multiple sellers. The goal is to measure the value of the data at a seller (he) for a buyer without focusing on a specific task for which the buyer is buying data. Data as a random variable is entirely defined by its distribution, and data distribution contains all the statistical information about the data. As a result, comparing data distributions at the buyer and seller could provide a comprehensive means for data valuation. However, in practice (as for ML applications), the data distribution is unknown, and it is often computationally impossible to approximate it using only a limited number of samples. Hence, we may instead directly use the data samples as the realizations of their distribution to capture their statistical properties. We further argue that the differences and similarities in the statistical properties of the data at different parties are reflected by two metrics, **diversity** and **relevance**. Accordingly, we aim to estimate these two metrics using the data at the seller and the buyer for data valuation.

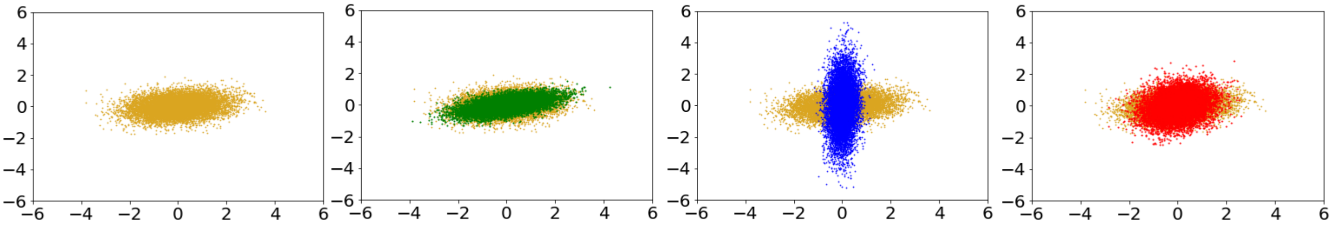
Diversity measures how much of different statistical properties the seller's data adds to the buyer's data, where we note that her data is limited to capture all the statistical properties of the original distribution. Whereas, relevance captures the similarity in the statistical properties of the two datasets. Consider a buyer with some cat and dog images, both with only black color. Intuitively, images of colorful cats and dogs seem to be a perfect addition to the buyer's data, where it provides some statistical properties that the buyer's data has not seen (because of the difference in colors), and some similarity in the statistical properties (having the same animals). Other images (except cats and dogs) could provide highly diverse data for the buyer; however, the relevance may be very limited, which prevents the buyer's data to capture the entire distribution. On the contrary, a dataset with black cats and dogs is highly relevant to the buyer's data, while it may not add any new statistics to it.

Another example is that of sample complexity in ML, which, given a data distribution, is defined as the minimum number of independent and identically distributed samples required for the ML model to generalize to that distribution without overfitting. Adding diverse data to the buyer's data helps the ML model to cover a wider range of statistical properties; however, this will require a larger sample size to guarantee that the model can generalize well to the new (statistically more diverse) dataset. While, after receiving a more relevant dataset at the buyer, it is likely that the ML model generalizes well (satisfy the sample complexity requirements) to the (limited) data statistics. As a result, there is a trade-off between the amount of diversity and relevance that the buyer is willing to receive and the performance (in this case whether the model generalizes well to capture all the statistical properties of the data).

Our goal is to develop a task-agnostic data valuation through measuring diversity and relevance between two datasets. We consider buyer's data as the baseline dataset and measure the diversity and relevance of a seller's data with respect to this baseline. Let us denote the buyer's and seller's data with matrices $\boldsymbol{B} \in \mathbb{R}^{n_b \times d}$ and $\boldsymbol{S} \in \mathbb{R}^{n_s \times d}$, respectively. The underlying assumption is that the datasets at the buyer and seller have the same feature space and as in ML applications have been zero-centered and normalized (this will guarantee that the datasets have the same support set). Data valuation is defined assuming that data could be readily used at the buyer without any computationally heavy post-processing (except zero-centering and normalization).

Considering the buyer's dataset $\boldsymbol{B}$ as the baseline, we denote the diversity and relevance of another dataset with respect to the baseline dataset by $D_{\boldsymbol{B}}$ and $R_{\boldsymbol{B}}$, respectively, such that $D_{\boldsymbol{B}} : \mathbb{R}^{n_s \times d} \to [0, 1]$ and $R_{\boldsymbol{B}} : \mathbb{R}^{n_s \times d} \to [0, 1]$. Accordingly, both diversity and relevance accept a dataset $\boldsymbol{S}$ (seller's data matrix) as input and map it to a real number in the interval $[0, 1]$. According to the above definition, the output of $D_{\boldsymbol{B}}$ and $R_{\boldsymbol{B}}$ is general enough, since any bounded interval can be normalized to the interval $[0, 1]$. A larger $D_{\boldsymbol{B}}$ ($R_{\boldsymbol{B}}$) indicates a larger diversity (relevance) of a dataset with respect to $\boldsymbol{B}$. As a result, for any specific realization of the measures $D_{\boldsymbol{B}}$ and $R_{\boldsymbol{B}}$ defined above, being close to 0 indicates the minimum diversity (relevance), while a measure close to 1 translates into the maximum diversity (relevance) of a dataset compared to the baseline dataset $\boldsymbol{B}$.

[1] This could be private data at each buyer or a publicly available dataset or a combination of both.



(a) Buyer

(b) Seller 1 vs. buyer

(c) Seller 2 vs. buyer

(d) Seller 3 vs. buyer

Figure 1: Data scatters illustration in 2-D for buyer and sellers 1 to 3's data with covariance matrices $[[1, 0.1], [0.1, 0.25]]$, $[[0.9, 0.2], [0.2, 0.15]]$, $[[0.1, 0.05], [0.05, 2]]$, and $[[0.5, 0.1], [0.1, 0.5]]$, respectively.

Any realization of $D_{\boldsymbol{B}}$ and $R_{\boldsymbol{B}}$ should satisfy the following two intuitive cases:

- **Case 1:** $D_{\boldsymbol{B}}(\boldsymbol{B}) = 0$; $R_{\boldsymbol{B}}(\boldsymbol{B}) = 1$; that is, the same dataset has no diversity and maximum relevance.
- **Case 2:** Using any distance measure, if the distance between the distributions of $\boldsymbol{B} \in \mathbb{R}^{n_b \times d}$ and $\boldsymbol{S} \in \mathbb{R}^{n_s \times d}$ is unbounded[2], we have $D_{\boldsymbol{B}}(\boldsymbol{S}) = 1$; $R_{\boldsymbol{B}}(\boldsymbol{S}) = 0$; that is, for distributions P_b and P_s of data at the buyer and seller, respectively, with their distance denoted by $\ell(P_b, P_s)$, $\lim_{\ell(P_b,P_s)\to\infty} D_{\boldsymbol{B}}(\boldsymbol{S}) = 1$ and $\lim_{\ell(P_b,P_s)\to\infty} R_{\boldsymbol{B}}(\boldsymbol{S}) = 0$.

Unlike the above task-agnostic data valuation formulation, a task-dependent data valuation may be a function of a learning algorithm, which takes as input a training dataset and outputs a ML model; also, it may depend on a utility function which takes as input the output of the learning algorithm (ML model) and/or a dataset and outputs a real value score (Sim et al.). Next we motivate our approach to measure diversity and relevance through a simple example.

Motivating Example

Here we focus on a 2-D feature space, i.e., $d = 2$, where $\boldsymbol{B} \in \mathbb{R}^{n_b \times 2}$ and $\boldsymbol{S} \in \mathbb{R}^{n_s \times 2}$. We consider the case where entries of data matrices $\boldsymbol{B}$ and $\boldsymbol{S}$ are distributed according to $\mathcal{N}(\mathbf{0}_2, \boldsymbol{\Sigma}_b)$ and $\mathcal{N}(\mathbf{0}_2, \boldsymbol{\Sigma}_s)$, respectively, where we note that data distribution is unknown to the nodes. For simplicity, we assume that the number of data samples at the buyer and each seller is 10^4, i.e., $n_b = n_s = 10^4$. We aim to measure the diversity and relevance of various datasets with respect to the baseline dataset (buyer's data) with a covariance matrix $\boldsymbol{\Sigma}_b = [[1, 0.1], [0.1, 0.25]]$. Fig. 1a illustrates the scatters of the buyer's data in 2-D. We observe that the buyer's data is scattered mostly across the first dimension.

We consider five sellers with various datasets. The datasets in the first three sellers have covariance matrices $\boldsymbol{\Sigma}_{s_1} = [[0.9, 0.2], [0.2, 0.15]]$, $\boldsymbol{\Sigma}_{s_2} = [[0.1, 0.05], [0.05, 2]]$, and $\boldsymbol{\Sigma}_{s_3} = [[0.5, 0.1], [0.1, 0.5]]$, respectively. Figs. 1b, 1c, and 1d demonstrate the scatters of the first three sellers' data compared to the buyer's data in 2-D. Accordingly, it is intuitive to conclude that, for the buyer, seller 1's data is the most similar compared to the data of the other two sellers, while seller 2's data has the least similarity among the three sellers. We expect seller 3's data to have some level of similarity and some level of difference compared to the buyer's data. We further consider sellers 4 and 5 with datasets with covariance matrices $\boldsymbol{\Sigma}_{s_4} = [[1, 0.1], [0.1, 0.25]]$ and $\boldsymbol{\Sigma}_{s_5} = [[50, 0], [0, 50]]$, respectively. Given the covariance matrix of the buyer's data, it is expected that seller 4's data should result in Case 1, i.e., minimum diversity and maximum relevance, while seller 5's data should lead to Case 2, i.e., maximum diversity and minimum relevance.

We need a metric to capture the differences in the statistical properties of various datasets compared to the buyer's data and reflect it in diversity and relevance. Our approach focuses on the second moment to capture the variations in distribution. In particular, we consider principal component analysis (PCA) applied to the covariance matrix of data at different nodes, where it measures the variance of data in directions corresponding to the principal components. We first find the principal components, together with their corresponding variance values, of the covariance matrix at the buyer. Then, the principal components of the buyer's data are shared with each seller, and he reports the variance of his covariance matrix in those directions. We then use the volume corresponding to the difference and intersection of the variances in the principal components directions to estimate diversity and relevance of the seller's data for the buyer's data, respectively. This is demonstrated in Fig. 2.

To be precise, the buyer first applies eigendecomposition to the covariance matrix $\frac{1}{n_b}\boldsymbol{B}^T\boldsymbol{B}$, which results in $\frac{1}{n_b}\boldsymbol{B}^T\boldsymbol{B} = [\boldsymbol{u}_1 \quad \boldsymbol{u}_2]\,\mathrm{Diag}(\lambda_1, \lambda_2)\,[\boldsymbol{u}_1 \quad \boldsymbol{u}_2]^T$, where $\boldsymbol{u}_1 = [0.99 \quad 0.13]^T$ and $\boldsymbol{u}_2 = [-0.13 \quad 0.99]^T$ are the eigenvectors (principal components), and $\lambda_1 = 1.01$ and $\lambda_2 = 0.23$ are the eigenvalues (variance in the direction of their corresponding eigenvectors). Next, the seller aims to find the variance of his data in both directions $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$, the eigenvectors of buyer's data. Having vectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ shared with a seller, he estimates the variance of his data in these directions by first computing the covariance matrix $\frac{1}{n_s}\boldsymbol{S}^T\boldsymbol{S}$, then the l^2-norm of this matrix projected onto the directions as follows $\hat{\lambda}_1 = \|\frac{1}{n_s}\boldsymbol{S}^T\boldsymbol{S}\boldsymbol{u}_1\|$,

[2] This is for any reasonable distance metric between two datasets, for instance Kullback–Leibler divergence, or Rényi divergence. For distance metrics with upper bound, such as Jensen–Shannon divergence, this could be rewritten as the maximum distance between the two datasets.

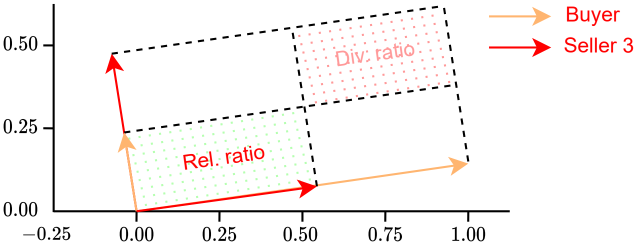


Figure 2: Variance of the buyer's and seller 3's data in the directions of the principal components of buyer's data.

$\hat{\lambda}_2 = \|\frac{1}{n_s}\boldsymbol{S}^T\boldsymbol{S}\boldsymbol{u}_2\|$. We note that, if $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are the eigenvectors of $\frac{1}{n_s}\boldsymbol{S}^T\boldsymbol{S}$, then $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are the eigenvalues of this matrix. Accordingly, at seller 3 with data generated with $\mathcal{N}(\mathbf{0}_2, [[0.5, 0.2], [0.2, 0.5]])$, after receiving vectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ from the buyer, we have $\hat{\lambda}_1 = 0.53$ and $\hat{\lambda}_2 = 0.47$.

Fig. 2 illustrates vectors $\lambda_1\boldsymbol{u}_1$ and $\lambda_2\boldsymbol{u}_2$, the principal components of the buyer's data, as well as $\hat{\lambda}_1\boldsymbol{u}_1$ and $\hat{\lambda}_2\boldsymbol{u}_2$, the variance estimate of seller 3's data in the directions of $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$. The goal is to estimate diversity and relevance of seller 3's data for the buyer based on the knowledge of λ_1, λ_2 and $\hat{\lambda}_1$, $\hat{\lambda}_2$. We argue that the volume measuring the difference (shown by red dots) represents diversity, while the intersection volume (shown by green dots) represents the relevance that seller 3's data has for the buyer. The rationale behind these choices is that the volume capturing the difference ($|\lambda_1 - \hat{\lambda}_1| \times |\lambda_2 - \hat{\lambda}_2|$) represents the dissimilarity between the two distributions measured through the second moment on principal components of the buyer's data, which is translated into the diversity of seller 3's data for the buyer. Whereas, the intersection volume ($\min\{\lambda_1, \hat{\lambda}_1\} \times \min\{\lambda_2, \hat{\lambda}_2\}$) measures the similarity between the two distributions through second moment, which is translated to the relevance of seller 3's data for the buyer.

In order to limit diversity and relevance values to the interval $[0, 1]$, we divide each of the difference and intersection volumes by the whole volume, i.e., $\max\{\lambda_1, \hat{\lambda}_1\} \times \max\{\lambda_2, \hat{\lambda}_2\}$. Furthermore, we take the square root of the result to account for the geometric mean; that is, we estimate diversity and relevance, respectively, as follows:

$$\text{Div.} = \left(\frac{|\lambda_1 - \hat{\lambda}_1| \times |\lambda_2 - \hat{\lambda}_2|}{\max\{\lambda_1, \hat{\lambda}_1\} \times \max\{\lambda_2, \hat{\lambda}_2\}}\right)^{1/2}, \tag{1a}$$

$$\text{Rel.} = \left(\frac{\min\{\lambda_1, \hat{\lambda}_1\} \times \min\{\lambda_2, \hat{\lambda}_2\}}{\max\{\lambda_1, \hat{\lambda}_1\} \times \max\{\lambda_2, \hat{\lambda}_2\}}\right)^{1/2}. \tag{1b}$$

We will show in the Appendix that, with the above estimates of diversity and relevance, we have Div. + Rel. ≤ 1. With the proposed measures, in general the desired output of (Div., Rel.) pair may be around $(0.5, 0.5)$, i.e., the buyer may desire moderate levels of diversity and relevance jointly instead of sacrificing one for the other, although this depends on the buyer's desire which may vary across the buyers.

Fig. 3 shows the diversity and relevance of various sellers' data for the buyer estimated based on our approach given that the buyer has Gaussian samples with covariance matrix $\boldsymbol{\Sigma}_b = [[1, 0.1], [0.1, 0.25]]$. As expected intuitively, data at seller 1 with covariance matrix $\boldsymbol{\Sigma}_{s_1} = [[0.9, 0.2], [0.2, 0.15]]$ resembles buyer's data and adds little diversity to it. Whereas, seller 2 with data with covariance matrix $\boldsymbol{\Sigma}_{s_2} = [[0.1, 0.05], [0.05, 2]]$ provides the buyer with a diverse data with little relevance. Unlike these two sellers, seller 3's data with covariance matrix $\boldsymbol{\Sigma}_{s_3} = [[0.5, 0.1], [0.1, 0.5]]$ has a moderate level of diversity and relevance for buyer with the pair very close to $(0.5, 0.5)$. These results indicate that the proposed estimates of diversity and relevance corroborate our intuition. Also, the proposed approach returns the expected results for the scenarios in Case 1 and Case 2 given the estimated diversity-relevance pair for seller 4's and seller 5's data, respectively.

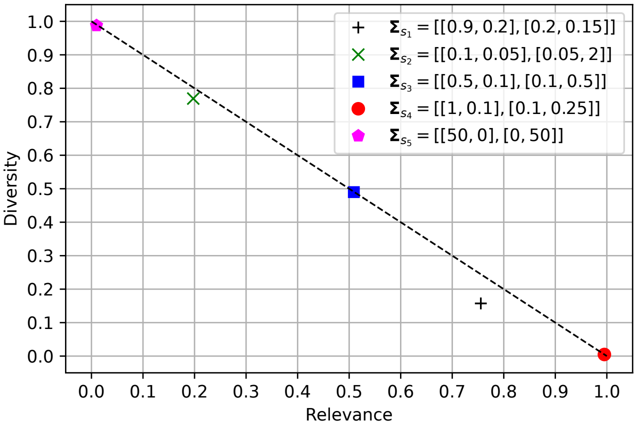


Figure 3: Diversity versus relevance of various datasets with 2-D zero-mean Gaussian distributions with various covariance matrices $\boldsymbol{\Sigma}_s$ with respect to the baseline dataset with covariance matrix $\boldsymbol{\Sigma}_b = [[1, 0.1], [0.1, 0.25]]$.

Diversity and Relevance Estimation

In this section, we present our approach in estimating diversity and relevance of dataset $\boldsymbol{S} \in \mathbb{R}^{n_s \times d}$ (seller's data) compared to the baseline dataset $\boldsymbol{B} \in \mathbb{R}^{n_b \times d}$ (buyer's data). This is carried out by comparing the statistical properties of the two datasets through second moment.

The buyer employs eigendecomposition to the covariance matrix $\frac{1}{n_b}\boldsymbol{B}^T\boldsymbol{B}$; i.e., $\frac{1}{n_b}\boldsymbol{B}^T\boldsymbol{B} = \boldsymbol{U}\ \text{Diag}(\lambda_1, ..., \lambda_d)\ \boldsymbol{U}^T$, where λ_i is the i-th largest eigenvalue, and $\boldsymbol{U} = [\boldsymbol{u}_1 \cdots \boldsymbol{u}_d]$ with $\boldsymbol{u}_i \in \mathbb{R}^d$ denoting the eigenvector corresponding to the i-th eigenvalue. We note that $\lambda_i \geq 0$ since $\frac{1}{n_b}\boldsymbol{B}^T\boldsymbol{B}$ is positive semi-definite. The buyer shares the principal components $\boldsymbol{u}_1, ..., \boldsymbol{u}_d$ with the seller, while $\lambda_1, ..., \lambda_d$ stay local at the buyer. The seller estimates the variance of its covariance matrix $\frac{1}{n_s}\boldsymbol{S}^T\boldsymbol{S}$ along $\boldsymbol{u}_1, ..., \boldsymbol{u}_d$. This is carried out as

$$\hat{\lambda}_i = \Big\|\frac{1}{n_s}\boldsymbol{S}^T\boldsymbol{S}\boldsymbol{u}_i\Big\|, \quad i = 1, ..., d, \tag{2}$$

where the covariance matrix $\frac{1}{n_s}\boldsymbol{S}^T\boldsymbol{S}$ is first projected into $\boldsymbol{u}_i$ and then l^2-norm of the resultant vector provides the estimate of the variance (the data matrices are zero-centered).

We note that if $\boldsymbol{u}_i$ is an eigenvector of $\frac{1}{n_s}\boldsymbol{S}^T\boldsymbol{S}$, then $\hat{\lambda}_i$ is its corresponding eigenvalue. Next, seller and buyer share $\hat{\lambda}_i$ and λ_i, for $i = 1, ..., d$, respectively, with the broker. The broker then tries to estimate the diversity and relevance of the seller's data for the buyer according to $\hat{\lambda}_i$ and λ_i.

We estimate diversity and relevance based on the volume of the space specified by the coordinates corresponding to the principal components (eigenvectors) of the covariance matrix of buyer's data. We have λ_i and $\hat{\lambda}_i$ as the value of buyer's and seller's data on the i-th coordinate, respectively. We estimate the relevance through the volume occupied by both buyer's and seller's data in these coordinates; that is, $\prod_{i=1}^{d} \min\{\lambda_i, \hat{\lambda}_i\}$. Our justification is that this volume captures the similarity in the statistical properties of the two datasets since this is a space occupied by both datasets. On the other hand, diversity is estimated through the volume of the difference between the variance of the buyer's and seller's data in each coordinate; that is, $\prod_{i=1}^{d} |\lambda_i - \hat{\lambda}_i|$. We argue that this volume captures the amount of dissimilarity in the statistical properties of the two datasets. We normalize these estimates through dividing it by the entire volume, i.e., $\prod_{i=1}^{d} \max\{\lambda_i, \hat{\lambda}_i\}$. This yields the following estimates for diversity and relevance, respectively, $\prod_{i=1}^{d} \left(\frac{|\lambda_i - \hat{\lambda}_i|}{\max\{\lambda_i, \hat{\lambda}_i\}}\right)$, $\prod_{i=1}^{d} \left(\frac{\min\{\lambda_i, \hat{\lambda}_i\}}{\max\{\lambda_i, \hat{\lambda}_i\}}\right)$, where each is the product of d terms each ≤ 1; so for large enough d, these estimates may be very close to 0. To address this issue, we take the geometric mean and estimate diversity and relevance of seller's data $\boldsymbol{S}$ for buyer with data $\boldsymbol{B}$, respectively, as follows:

$$D_{\boldsymbol{B}}(\boldsymbol{S}) = \prod_{i=1}^{d} \left(\frac{|\lambda_i - \hat{\lambda}_i|}{\max\{\lambda_i, \hat{\lambda}_i\}}\right)^{1/d}, \tag{3a}$$

$$R_{\boldsymbol{B}}(\boldsymbol{S}) = \prod_{i=1}^{d} \left(\frac{\min\{\lambda_i, \hat{\lambda}_i\}}{\max\{\lambda_i, \hat{\lambda}_i\}}\right)^{1/d}. \tag{3b}$$

Fig. 4 shows the proposed approach with the interactions between different parties to value a seller's data for a buyer.

It is easy to verify that the proposed diversity and relevance estimates satisfy the conditions in Case 1 and Case 2. We will show in the Appendix that the proposed estimates validate additional intuitive properties. Generally speaking, with the proposed approach a safe default target is to have a diversity-relevance pair close to $(0.5, 0.5)$. However, this may change depending on a buyer's desire. For instance for a buyer with a relatively small amount of data, acquiring a highly diverse dataset may not be desirable since it is likely that the sample complexity requirements will not be satisfied given her own limited data samples (in other words, adding diverse data samples complexifies the underlying distribution, which increases the already excessive sample complexity). While, a buyer with a relatively large amount of data may prefer acquiring a more diverse data since most likely she already has enough samples to generalize to her own data distribution (she has enough room to diversify her underlying distribution and increase her sample complexity).

Partial Components. We remark that the proposed approach can be readily extended to the scenario where diversity and relevance could be estimated using the variance in the directions of only partial main components rather than all the d directions; that is, assuming a subset $\mathcal{D} \subset \{1, ..., d\}$,

$$D_{\boldsymbol{B}}(\boldsymbol{S}) = \prod_{i\in\mathcal{D}} \left(\frac{|\lambda_i - \hat{\lambda}_i|}{\max\{\lambda_i, \hat{\lambda}_i\}}\right)^{1/|\mathcal{D}|}, \tag{4a}$$

$$R_{\boldsymbol{B}}(\boldsymbol{S}) = \prod_{i\in\mathcal{D}} \left(\frac{\min\{\lambda_i, \hat{\lambda}_i\}}{\max\{\lambda_i, \hat{\lambda}_i\}}\right)^{1/|\mathcal{D}|}. \tag{4b}$$

This is valid with high dimensional data since not all the principal components carry significant information (Shlens).

Representations. Assuming that different parties have access to the same publicly available pre-trained model (such as VGG16 trained on the ImageNet dataset (Deng et al.)), the proposed estimates to measure diversity and relevance can be employed to the representations of data, instead of raw data. To be precise, different parties can forward propagate the data to the (same) pre-trained model and obtain the activations of the last hidden layer of the model, and then apply the proposed algorithm to estimate diversity and relevance between different datasets. We highlight that the last hidden layer output provides a compact representations of data by capturing its most significant attributes (Zhang et al.).

Experiments

We evaluate our estimates for diversity and relevance using real datasets, namely Adult (Kohavi), MNIST (LeCun et al.), fashion-MNIST (Xiao et al.), Cifar-10 (Krizhevsky) and FairFace (Karkkainen et al.). For the experiments, we estimate diversity and relevance through the partial components analysis, given in (4), where only the principal components of buyer's data with corresponding eigenvalues more than 10^{-2} are chosen. We will observe in the experiments that the proposed approach can capture the increase in diversity (relevance) when reducing (enhancing) the overlaps in demographics or labels between seller's and buyer's datasets.

We first consider the Adult dataset which has various features such as education level, age, occupation, etc., predicting whether an individual's annual income is over 50k or not. We consider a buyer with a dataset of individuals with doctorate degree and annual salary over 50k. We consider various sellers with datasets of individuals with: i) no more than 20 years old; ii) annual salary over 50k and working for less than 40 hours per week; iii) 60 years of age and older; iv) annual salary more than 50k; v) annual salary more than 50k and education level of at least bachelors, i.e., bachelors, masters, prof-school, or doctorate; vi) ducation level of prof-school or doctorate; vii) education level of doctorate.

Fig. 5 illustrates the diversity-relevance pair of each seller's data for the buyer using the proposed estimates. As expected, we observe that the seller with data only from the individuals not older than 20 years provides the most diverse and least relevant data; it is highly likely that none of these individuals hold a doctorate degree and earn at least 50k per year. Also, it is likely that most of those with doctorate degree and annual salary of at least 50k work for more than 40 hours per week; that is why the seller with information about individuals working less than 40 hours per week while

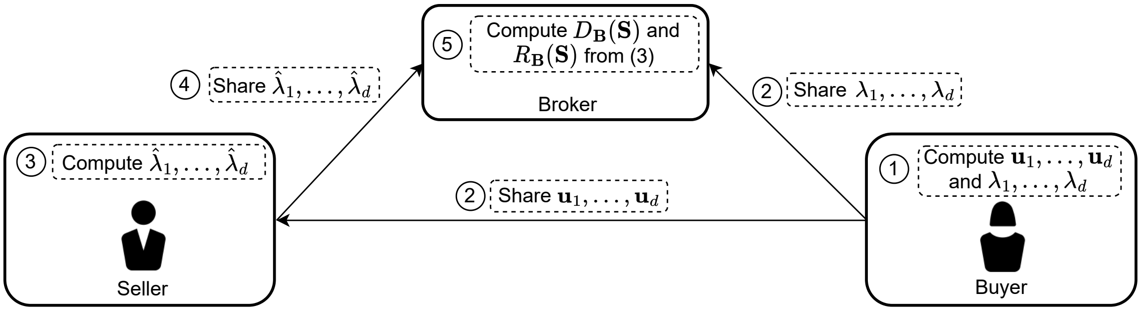

5
Compute $D_{\mathbf{B}}(\mathbf{S})$ and $R_{\mathbf{B}}(\mathbf{S})$ from (3)
Broker
4
Share $\hat{\lambda}_1, \dots, \hat{\lambda}_d$
2
Share $\lambda_1, \dots, \lambda_d$
3
Compute $\hat{\lambda}_1, \dots, \hat{\lambda}_d$
Seller
2
Share $\mathbf{u}_1, \dots, \mathbf{u}_d$
1
Compute $\mathbf{u}_1, \dots, \mathbf{u}_d$ and $\lambda_1, \dots, \lambda_d$
Buyer

Figure 4: The proposed interaction between different parties to estimate diversity and relevance of seller's data for buyer.

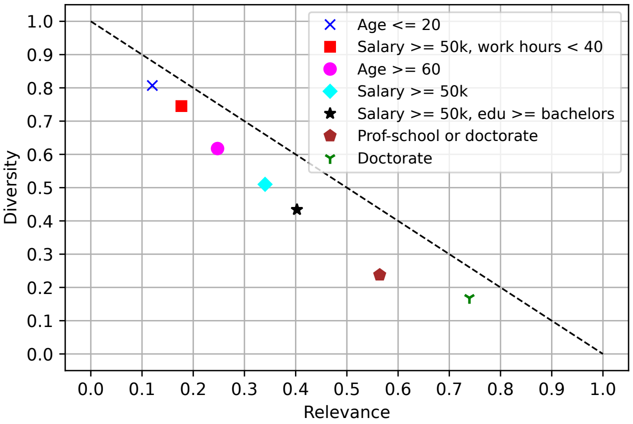


Figure 5: Diversity versus relevance of datasets at various sellers compared to buyer's dataset considering the Adult dataset, where the buyer has data of individuals with doctorate degree earning an annual salary of at least 50k.

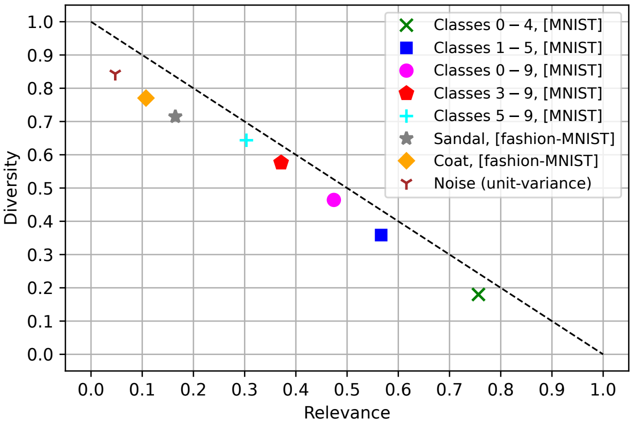


Figure 6: Diversity versus relevance of data at various sellers with data from MNIST, fashion-MNIST and noisy images compared to buyer's data from MNIST with classes 0 to 4.

earning more than 50k annually has more diversity than relevance for the buyer. We also expect that most people over the age of 60 years do not hold doctorate degree and/or earn more than 50k per year. However, as we expect, for the buyer the relevance of data at the 3rd seller is more than that at the first seller which is also reflected in our estimates.

Considering a seller with data of individuals earning at least 50k annually, this provides a slightly more diversity than relevance for the buyer, since most of these individuals may not hold a doctorate degree; the diversity however is smaller than the data of individuals older than 60 years which includes people with more diverse education levels. Limiting the dataset of high income individuals to those holding bachelors degree or higher academic degree (masters, prof-school, doctorate) reduces the diversity and increases the relevance to the buyer's data. On the other hand, the seller with data of individuals having doctorate degree provides a relatively high relevance for the buyer (since most of those probably earn more than 50k annually), while the relevance reduces considering a seller with information of people holding degree from a professional school or doctorate. We observe that our estimates capture these differences.

Fig. 6 evaluate our estimates of diversity and relevance using MNIST and fashion-MNIST datasets. We assume that the buyer has $\sim$ 6000 images and sellers have $\sim$ 10000 images, and each party has access to distinct images. We consider buyer with images of only classes 0 to 4 from MNIST, and five sellers with images from MNIST with different classes, precisely classes 0 to 4 (same as the buyer), 1 to 5, 0 to 9, 3 to 9, and 5 to 9. It is evident that the diversity/relevance should increase/decrease from seller 1 to seller 5, where the proposed estimates illustrate these changes. It is interesting to note that the seller with data from all the classes 0 to 9 provides a diversity-relevance pair close to point $(0.5, 0.5)$. To further evaluate the proposed estimates, we consider sellers with images of a different dataset than the images at the buyer. In particular, we consider two sellers with images of Sandal and Coat classes, respectively, from fashion-MNIST dataset. As expected, we observe that these two sellers provide a more diverse data for the buyer compared to the sellers having images from MNIST. Furthermore, we consider a seller with only noisy images where each pixel (in a 28×28 image in this case) is assumed to have a zero-mean unit-variance Gaussian distribution. Observe that this seller provides the largest diversity and smallest relevance to the buyer compared to other sellers. We note that by increasing the noise variance, the diversity-relevance pair will get closer to $(1, 0)$, i.e., corresponding to Case 2.

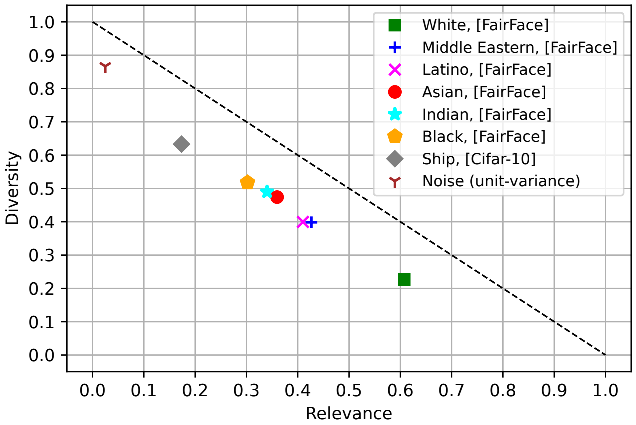


Figure 7: Diversity versus relevance of data at various sellers from Cifar-10, FairFace and noisy images compared to buyer's data from FairFace with images of class White.

Next we consider a buyer with images from the FairFace dataset, which contains face images of different race groups, namely White, Black, Indian, Middle Eastern, Asian (combining East Asian and Southeast Asian). We assume that the buyer and all the sellers have access to the publicly available VGG16 model pre-trained on the ImageNet dataset, and apply the proposed scheme to estimate diversity and relevance to the output of the last hidden layer of this pre-trained model (feature representations of data). We assume that the buyer has images of only White group, and six sellers each has images of one group (the images of White group at a seller are different than the ones at the buyer). We observe in Fig. 7 that the images of Black group provides the most diversity for the buyer, and images of Indian and Asian (East Asian + Southeast Asian) have respectively more diverse than relevant data for the buyer. Whereas, sellers with images of Middle Eastern and Latino groups are more relevant for the buyer compared to the other three groups, and, as expected, the seller with White group images has the least diverse data for the buyer. Overall, human face images follow a particular pattern and does not vary significantly across individuals. To further capture the differences, we consider a seller with Ship images from Cifar-10 passed through the pre-trained VGG16 model on ImageNet. We observe that this seller provides a relatively small relevant data for the buyer. Also, observe that data of the seller with only noisy images is highly diverse and not relevant for the buyer.

Discussions

Here we discuss about various aspects of the proposed approach ranging from its properties to possible extensions.

Privacy Enhancing. With the proposed approach, the only information about buyer's data that is shared with the seller is all/partial principal components of its covariance matrix. This could be easily modified to enhance the privacy of buyer's data by sharing extra directions, in addition to the principal components, with the seller. This can hide the principal components of buyer's data, which are the only directions used to estimate diversity and relevance at the broker.

Robustness to Malicious Seller(s). With the proposed approach, the seller does not have any information about the variance of buyer's data in different directions. This provides a robust mechanism to malicious sellers who try to fabricate their data resulting in a particular diversity-relevance pair and/or add noise to their data; that is, the seller does not have enough information about the buyer's data to manipulate the algorithm to output a specific diversity-relevance pair. Also, adding random noise to the data at the seller may increase the diversity for the buyer, however it reduces the relevance. Thus, the trade-off between diversity and relevance controls a malicious seller in adding noise to his data.

Number of Data Samples. We assume that sellers have large enough data compared to the buyer, and we do not incorporate the size of data at the sellers into the proposed data valuation formulation. This can be extended by considering the size of data at the sellers as an additional metric for data valuation. We can further extend the approach by considering the mean of data samples at different parties.

Weighted Averaging. The proposed diversity and relevance estimates could be extended by weighting the ratios at various principal components of the buyer's data differently. For weights $\omega_1, ..., \omega_d$ with $0 \leq \omega_i \leq 1$ and $\sum_{i=1}^{d} \omega_i = 1$, we can estimate diversity as $D_{\boldsymbol{B}}(\boldsymbol{S}) = \prod_{i=1}^{d} \left(\frac{|\lambda_i - \hat{\lambda}_i|}{\max\{\lambda_i, \hat{\lambda}_i\}} \right)^{\omega_i}$ and relevance as $R_{\boldsymbol{B}}(\boldsymbol{S}) = \prod_{i=1}^{d} \left(\frac{\min\{\lambda_i, \hat{\lambda}_i\}}{\max\{\lambda_i, \hat{\lambda}_i\}} \right)^{\omega_i}$. This is useful when the variance in one particular direction may be of more importance for the buyer. Furthermore, we can output a single value as the data value computed as the combination of diversity and relevance specified by the buyer. For example, if the buyer prefers to have ratio α diverse and $(1-\alpha)$ relevant data, for $0 \leq \alpha \leq 1$, data value of a seller for the buyer can be computed as $\alpha D_{\boldsymbol{B}}(\boldsymbol{S}) + (1-\alpha) R_{\boldsymbol{B}}(\boldsymbol{S})$.

Maximum Diversity Maximum Relevance. With the proposed approach, the diversity and relevance estimates are not independent since their sum can not exceed 1. Although ideally maximum diversity and maximum relevance are desired, we argue that this may not be feasible. Maximum diversity encompasses all the randomness that could be added to the data for more diversity, i.e., when adding more randomness does not increase the diversity. Having maximum relevance in this case may not be feasible.

Conclusions

We studied task-agnostic valuation of data at a seller for a buyer. This is specifically relevant when the buyer has access to data samples apriori which could be used to measure usefulness of seller's data for the buyer. We formulated the problem as the diversity and relevance of the seller's data for the buyer in the efforts to compare the statistical properties of the two datasets. We then provided estimates for the diversity and relevance by measuring the difference and similarity volumes using the space of the principal components of the buyer's data as the baseline; this technique focuses on the second moment analysis by comparing the variance of each dataset on these components. We show that the proposed estimates are successful in capturing the diversity and relevance of two datasets using various real datasets.

References

Agarwal, A.; Dahleh, M.; Horel, T.; and Rui, M. 2021. Towards data auctions with externalities. https://arxiv.org/abs/2003.08345. Accessed: 2021-09-04.

Agarwal et al. 2019. A marketplace for data: an algorithmic solution. In *Proc. ACM Conference on Economics and Computation*, 701–726.

Bergemann et al. 2022. Data, Competition, and Digital Platforms. https://www.mit.edu/~bonatti/dcdp.pdf. Accessed: 2022-08-03.

Bimpikis et al. 2019. Information sale and competition. *Management Science*, 65(6): 2646–2664.

Che et al. 2018. Recommender systems as mechanisms for social learning. *Quarterly Journal of Economics*, 133(2): 871–925.

Chen et al. 2019. Towards model-based pricing for machine learning in a data marketplace. In *Proc. International Conference on Management of Data (SIGMOD)*, 1535–1552.

Cong, Z.; Luo, X.; Pei, J.; Zhu, F.; and Zhang, Y. 2022. Data pricing in machine learning pipelines. *Knowledge and Information Systems*, 64: 1417–1455.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Ghorbani et al. 2019. Data Shapley: equitable valuation of data for machine learning. In *Proc. International Conference on Machine Learning (ICML)*, 2242–2251.

Gradwohl et al. 2022. Pareto-improving data-sharing. https://arxiv.org/abs/2205.11295. Accessed: 2022-05-23.

Immorlica, N.; Mao, J.; Slivkins, A.; and Wu, Z. S. 2020. Incentivizing exploration with selective data disclosure. In *Proc. ACM Conference on Economics and Computation*, 647–648.

Jia, R.; Dao, D.; Wang, B.; Hubis, F. A.; Hynes, N.; Gurel, N. M.; Li, B.; Zhang, C.; Song, D.; and Spanos, C. 2019. Towards efficient data valuation based on the Shapley value. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, 1167–1176.

Karkkainen et al. 2021. FairFace: face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision*, 1548–1558.

Kohavi, R. 1996. Scaling up the accuracy of naive-Bayes classifiers: a decision-tree hybrid. In *Proc. International Conference on Knowledge Discovery and Data Mining*. Portland, Oregon.

Krizhevsky, A. 2009. Learning multiple layers of features from tiny images. *Technical Report, University of Toronto*, 2–3.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. The MNIST database of handwritten digits. *Proceedings of the IEEE*, 86(11): 2278–2324.

Liu, J.; Lou, J.; Liu, J.; Xiong, L.; Pei, J.; and Sun, J. 2021. Dealer: an end-to-end model marketplace with differential privacy. *VLDB Endowment*, 14(6): 957–969.

Luong, N. C.; Hoang, D. T.; Wang, P.; Niyato, D.; Kim, D. I.; and Han, Z. 2016. Data collection and wireless communication in internet of things (IoT) using economic analysis and pricing models: a survey. *IEEE Communications Surveys and Tutorials*, 18(4): 2546–2590.

Muschalle et al. 2012. Pricing approaches for data markets. In *Proc. International workshop on business intelligence for the real-time enterprise*, 129–144.

Niu, C.; Zheng, Z.; Wu, F.; Tang, S.; Gao, X.; and Chen, G. 2018. Unlocking the value of privacy: trading aggregate statistics over private correlated data. In *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2031–2040.

Pei, J. 2020. A survey on data pricing: from economics to data science. *IEEE Transactions on Knowledge and Data Engineering*, 1–1.

Raskar, R.; Vepakomma, P.; Swedish, T.; and Sharan, A. 2019. Data markets to support AI for all: pricing, valuation and governance. https://arxiv.org/abs/1905.06462. Accessed: 2019-05-14.

Rasouli et al. 2021. Data sharing markets. https://arxiv.org/abs/2107.08630. Accessed: 2021-07-20.

Richardson et al. 2020. Budget-bounded incentives for federated learning. *Springer*, 176–188.

Schomm et al. 2013. Marketplaces for data: an initial survey. *ACM SIGMOD Record*, 42(1): 15–26.

Shlens, J. 2014. A tutorial on principal component analysis. https://arxiv.org/abs/1404.1100. Accessed: 2014-04-03.

Sim, R. H. L.; Zhang, Y.; Chan, M. C.; and Low, B. K. H. 2020. Collaborative machine learning with incentive-aware model rewards. In *Proc. International Conference on Machine Learning (ICML)*, 8927–8936.

Sim et al. 2022. Data valuation in machine learning: "ingredients", strategies, and open challenges. In *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*.

Song et al. 2019. Profit allocation for federated learning. In *Proc. IEEE International Conference on Big Data (Big Data)*, 2577–2586.

Tay, S. S.; Xu, X.; Foo, C. S.; ; and Low, B. K. H. 2022. Incentivizing collaboration in machine learning via synthetic data rewards. In *Proc. AAAI Conference on Artificial Intelligence*.

Toni, D. D.; Milan, G. S.; Saciloto, E. B.; and Larentis, F. 2017. Pricing strategies and levels and their impact on corporate profitability. *Revista de Administracao*, 52(2): 120–133.

Wang, T.; Rausch, J.; Zhang, C.; Jia, R.; and Song, D. 2020. A principled approach to data valuation for federated learning. *Lecture Notes in Computer Science, Springer*, 12500: 153–167.

Xiao et al. 2017. Fashion-MNIST: a novel image dataset for benchmarking machine Learning algorithms. https://arxiv.org/abs/1708.07747. Accessed: 2017-09-15.

Xu, X.; Wu, Z.; Foo, C. S.; and Low, B. K. H. 2021. Validation free and replication robust volume-based data valuation.

In *Proc. Conference on Neural Information Processing Systems (NeurIPS)*.

Yan et al. 2021. If you like Shapley then you'll love the core. In *Proc. AAAI Conference on Artificial Intelligence*, 5751–5759.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 586–595.

Zhang et al. 2020. A survey of data pricing methods. *SSRN*, 1–25.

Zheng et al. 2021. Optimal advertising for information products. In *Proc. ACM Conference on Economics and Computation*, 888–906.