# Try Before You Buy: a Practical Data Purchasing Algorithm for Real-World Data Marketplaces

Santiago Andrés Azcoitia
santiago.azcoitia@imdea.org
IMDEA Networks Institute
Universidad Carlos III Madrid
Leganés, Madrid, Spain

Nikolaos Laoutaris
nikolaos.laoutaris@imdea.org
IMDEA Networks Institute
Leganés, Madrid, Spain

## ABSTRACT

Data trading is becoming increasingly popular, as evident by the appearance of scores of data marketplaces (DMs) in the last few years satisfying the demand for third-party data. For buyers, however, deciding whether paying the requested price makes sense can only be done after having tested the data on their ML model. In this paper, we propose a method for optimizing data purchasing decisions. We show that if a marketplace provides to potential buyers a measure of the performance of their models on *individual* datasets, then they can select which of them to buy with an efficacy that approximates that of knowing the performance of each possible combination of datasets offered by the DM. We call the resulting algorithm *Try Before You Buy* (TBYB) and demonstrate over synthetic and real-world datasets how TBYB can lead to near optimal data purchasing with only $O(N)$ instead of $O(2^N)$ information and execution time.

## CCS CONCEPTS

• **Information systems** → **Data management systems**; • **Computing methodologies** → *Machine learning*; • **Theory of computation** → **Data exchange**.

## KEYWORDS

data economy, value of data, data purchasing, data marketplaces

**ACM Reference Format:**
Santiago Andrés Azcoitia and Nikolaos Laoutaris. 2022. Try Before You Buy: a Practical Data Purchasing Algorithm for Real-World Data Marketplaces. In *Data Economy (DE '22), December 9, 2022, Roma, Italy.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3565011.3569054

## 1 INTRODUCTION

Data is considered a key production factor, comparable in importance to labour, capital, and infrastructure. Companies often need data they do not possess, or cannot collect directly. Therefore, general-purpose [1, 35] and domain-specific [6, 19, 29, 33] DMs have appeared, and leading data management platforms are adding marketplace functionality into their products [8, 18, 39] with the purpose of building a business of mediation between data selling and buying *companies*. Finally, personal information management systems (PIMS) have answered the call of recent legislative developments in personal data protection [15, 26] by offering data control, portability, and monetization services for *individuals*.

Designing and building a successful data marketplace calls for addressing a plethora of technology, business, and economics challenges [4, 11]. According to our survey [3], DMs spend effort and money to attract enough sellers, and then try to convince buyers to purchase data through the platform. Therein lie two fundamental problems: (1) the *pricing problem* for sellers, and (2) the *purchasing problem* for buyers.

Recent theoretical work on the intersection between computer science and economics has looked into those problems, and has proposed solution concepts and algorithms for them [2, 12, 13, 23, 28]. Other works look at the complex problem of data pricing [31] from different perspectives. In this paper we will focus on the problem of selecting which datasets to buy (problem (2)), given the prices set by sellers (problem (1)), which can be further broken down into two interrelated subproblems: (2.a) compute how useful these datasets will be to their ML algorithms, something that can be captured by various accuracy metrics, and (2.b) compute how such accuracy can be converted into monetary gains (via improved sales, acquisition of new customers, retention of existing ones, etc.). Whereas sub-problem (2.b) seems to be easier for buyers [7], subproblem (2.a) is inherently more challenging. Buyers need to have access to the data before they can compute its value for their ML task, but such access is only granted *after* purchasing the data – this is known as Arrow's information paradox. (2.a) is further exacerbated when the buyer can choose among $2^N$ combinations of N datasets.

In theoretical works, the value of any subset of datasets for a data buyer is considered as known *a priori* [37]. In reality, however, things are completely different [30], and the most common practice from data consumers and service providers is collecting as much data as possible from different sources and from end users, leading to enormous implications in privacy. Selecting suitable data from a number of different sources is already a problem in emerging niche-DMs focused on specific AI/ML tasks, such as image recognition [38] or those related to natural language processing [41]. Buyers will also need to select suitable personal data from individuals using PIMS or geo-location DMs [16]. In most DMs, data sellers provide only a description of their datasets, a price, sometimes an outdated sample, and buyers have to make purchase decision with that information alone. Some of them offer "sandboxed" environments that allow data buyers to experiment versions of the data

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
*DE '22, December 9, 2022, Roma, Italy*
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9923-4/22/12…$15.00
https://doi.org/10.1145/3565011.3569054

without being able to copy or extract them [1, 6, 29]. Few of these DM also allow buyers to make offers (bids) for data when sellers do not indicate a fixed price, or if they are willing to pay something below the asking price. This case suffers as well from the fundamental problem (2.a) of not knowing the value of a dataset before purchasing it.

**Our contribution:** Our main contribution is to propose a preliminary "data evaluation" phase prior to buyers selecting which datasets to buy and, in this context, to introduce a family of dataset purchase algorithms that we call "Try Before You Buy" (or TBYB) that allow data buyers to identify the best datasets to buy with only $O(N)$ information about the accuracy of ML algorithms on individual datasets, instead of $O(2^N)$ information required by an optimal strategy using full information. Effectively, TBYB needs to know only the accuracy of a ML model on *individual* datasets, and with this information it can approximate the optimal *combination* of datasets that maximizes the profit of the buyer.

We compare the performance of TBYB against several heuristics lacking information about the value of datasets for the particular task, and against the optimal purchase. We start with a synthetic evaluation and then validate our conclusions using real spatio-temporal datasets from companies and anonymized individuals in use cases related to demand and travel time prediction in metropolitan areas.

**Our findings** are as follows:

- TBYB remains close to the optimal in most scenarios, and its benefit increases with the catalog size (N).
- TBYB is almost optimal when buying more data yields a progressively diminishing return in value for the buyer. Otherwise, TBYB finds it more difficult to match the optimal performance, although it still outperforms the heuristics.
- The benefit of TBYB becomes maximal when prices of datasets do not correlate with their actual value for the buyer.
- When dealing with personal data, TBYB can significantly reduce the number of individuals whose information is disclosed to buyers, hence helping in preserving privacy, too.

Overall, our work demonstrates that TBYB allows for near optimal dataset purchasing that could be easily implemented by real-world DMs. We believe that this is a key functionality for allowing both marketplaces and the data offer to grow.

## 2 MARKETPLACE MODEL & DEFINITIONS

Figure 1 shows the reference DM model used throughout the paper. It works as follows:

(1) We will assume that the DM provides buyers with a "sandbox" for them to submit their ML task, including the model $M$, the accuracy function they seek to optimize, and the test set they want the DM to use in order to evaluate such accuracy.

(2) The DM (or alternatively, the buyer) will identify suitable data for the task and get this data from data sellers.

(3) Existing DMs typically list the datasets they offer and provide for each one a description and a price. The sandbox allows buyers to additionally get an *exact* answer in terms of the achieved accuracy over each candidate dataset for the particular task of the buyer, without being able to see or copy the raw data.

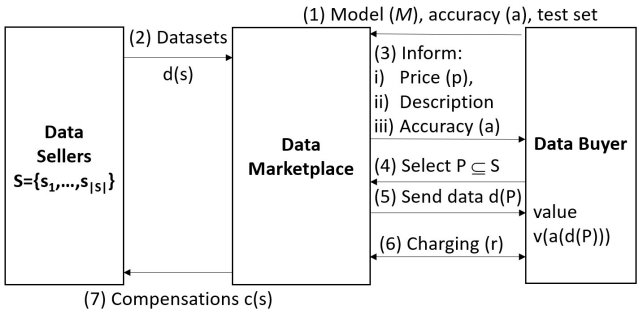



**Figure 1: Reference data marketplace model**

(4) With that information, the buyer will make a decision on which datasets to purchase (P) and orders them,

(5) the DM will send the corresponding data, which the buyer will use to train its model, obtaining a value that depends on the accuracy

(6) The DM will then charge for its services and receive the corresponding payment from the data buyer, and

(7) The DM will share part of this payment as a compensation to the data seller.

We will denote by $\mathcal{S}$ the set of suitable sellers for the ML task of a particular buyer. We will denote by $d(s)$ the dataset offered by seller $s \in \mathcal{S}$, by $p(s)$ its *price*, and by $a(d(s)) \in [0, 1]$ the *accuracy* that the buyer's ML task can achieve if trained by $d(s)$. Similarly, for a subset of the sellers $S \in \mathcal{S}$, we will denote by $d(S)$ their aggregated dataset, and by $a(d(S))$ the maximum accuracy that can be achieved using all or a subset of the data in $d(S)$. We will also introduce the *value function* $v(a)$, that tells the value for the buyer if his ML model gets to an accuracy of $a$.

When datasets are bought sequentially, we will append a subscript to identify the round to the notation we already defined. Hence, $S_n \subset \mathcal{S}$ will refer to the set of eligible datasets in round $n$. We will denote by $P_n$ the set of data already under the buyer's control in round $n$, allowing to achieve an accuracy $a_n = max_{S' \in 2^{P_n}} a(S')$ and a value $v_n = v(a_n)$.

Buyers will always look forward to optimizing their profit, so they will decide to purchase a dataset $d(s)$, if its (expected) marginal value exceeds its cost, i.e. $v(a(d(s|P))) \geq p(s)$. To simplify the analysis, we will assume that $v(a) = a$.

## 3 DATA PURCHASE STRATEGIES

In this section, we will present a series of data purchase strategies including our proposed algorithm called TBYB.

***Optimal purchase*** assumes that the buyer knows $a(d(S))$ for any subset $S \subseteq 2^S$. This allows for an optimal purchase $\mathcal{S}^\star$ that maximizes the profit, i.e., the difference between the value that the buyer extracts from the data, and the cost paid to purchase them:

$$\mathcal{S}^\star = \arg\max_{S \in 2^S} \left( v(a(d(S))) - \sum_{s \in S} p(s) \right), \quad (1)$$

subject to $v(a(d(\mathcal{S}^\star))) \geq \sum_{s \in \mathcal{S}^\star} p(s)$.

***Stand-alone Try Before You Buy (S-TBYB)*** assumes that the DMs provide buyers with the accuracy of their models on individual datasets, but not on combinations of them.

The marketplace provides $a(d(s))$ on all $s \in \mathcal{S}$. Then S-TBYB starts buying datasets in descending order of *expected profit* until a stop condition is reached.

For the first dataset, the profit is not expected but exact, so the best dataset is bought provided $v(a(d(s))) - p(s) \geq -\lambda \cdot v(a^\star)$, where:

- $a^\star \leq 1$ is the best accuracy that can be delivered, as informed by the DM or as estimated by the buyer, and

- $\lambda$ is the risk parameter and models the maximum admissible loss for the buyer, relative to the potential value of the sourcing operation ($v(a^\star) - v_n$ in round n). For example, $\lambda = 0.1$ means that buyers will buy $s$ if its price is lower than the marginal value they expect to get plus 10% of the maximum value that they could add by buying new data.

*n-th iteration.* The buyer will proceed as follows:

(1) Identify the best possible dataset $s^\star \in S_n$ such that $s^\star = \arg\max_{s \in S_n} (v(a(d(s))) - p(s))$
(2) Purchase $s^\star$ if its estimated marginal value exceeds its price, and a risk threshold that depends on the remaining value she expects to get out of the operation, i.e., if $v(E\{a(s^\star \cup S_n)\}) - v_n - p(s^\star) \geq -\lambda \cdot (v(a^\star) - v_n)$
(3) If the buy condition is met then, $s^\star$ is added to $P_{n+1} = P_n \cup d(s^\star)$ and the next round starts,
(4) else if no dataset in $S$ meets this requirement, then the process stops

The estimation $E\{a(s^\star \cup S_n)\}$ is specific to each problem. We estimate the relative added accuracy of $s^\star$ by multiplying its individual accuracy $a(s^\star)$, and the ratio of the marginal contribution and the individual accuracy of the last purchased dataset:

$$E\{a(s^\star \cup S_n)\} = \frac{a_n - a_{n-1}}{a(P_n - P_{n-1})} \cdot a(s^\star) \cdot (a^\star - a_n). \quad (2)$$

***Assisted Try Before You Buy (A-TBYB)*** assumes the buyer is allowed to ask the marketplace *every round* for the marginal accuracy of any still eligible datasets. As a result, the model will be processed a maximum if $\sum_{i=0}^{r-1} |S| - i$ times for $r$ rounds, and no estimation of the marginal value is needed.

***Volume-based purchasing***. Most DMs provide buyers with a description of datasets including their volume (e.g., n° samples), which is often used to choose among the different offers. Let $vol(s)$ denote the volume of dataset $s$, used as merit figure by the following volume-based purchasing heuristic:

*n-th iteration.* We will assume that a greedy buyer would select the dataset $s^\star \in S_n$ with the highest $vol(s)/p(s)$ ratio every round, provided:

$$p(s^*) \leq -\lambda \cdot (v(a^\star) - v_n), \quad (3)$$

which assumes that even in the worst case the maximum relative admissible loss is not exceeded in the operation.

***Price-based purchasing*** assumes buyers randomly select among datasets whose price is lower than their maximum relative admissible loss.

# 4 PERFORMANCE EVALUATION WITH SYNTHETIC DATA

We will use an easily reproducible synthetic model to evaluate the performance of the purchase strategies.

## 4.1 Synthetic model description

We will denote as total cost of data (TCOD), the cost of buying all the available datasets in $\mathcal{S}$, i.e., $TCOD = \sum_{s \in \mathcal{S}} p(s)$. Hence, if $TCOD < a(d(\mathcal{S}^\star))$ the buyer is guaranteed to make a profit. In the more interesting case of $TCOD \gg 1$, the buyer needs to select which datasets to buy carefully, to avoid ending up with a loss. Our synthetic model assumes that:

$$v(a(d(S))) = a(d(S)) = \left( \frac{\sum_{s_i \in S} DI^i}{\sum_{s_i \in \mathcal{S}} DI^i} \right)^{MUP}, \quad (4)$$

where:

- **MUP is the Marginal Utility Profile** parameter that controls the concaveness/convexity of $v(\cdot)$ from 0 to 1 as more data is bought. $MUP < 1$ means that buying additional datasets will have a decreasing marginal utility for the buyer. $MUP > 1$ means that the marginal contribution of new data sources will increase as more datasets are bought. Finally, $MUP = 1$ means that all datasets yield the same marginal accuracy regardless the purchase sequence.

- **DI is the Data Interchangeability parameter** that controls the relative value of different datasets in $\mathcal{S}$. $DI = 1$ means making all datasets fully interchangeable. For $DI > 1$ and $MUP = 1$, dataset $s_i$ is DI times more valuable than dataset $s_{i-1}$, $1 \leq i \leq |\mathcal{S}|$. It is only if $DI \neq 1$ when not only how many datasets are bought, but also which ones matters.

Regarding pricing, we will consider the following schemes:
P1 - All datasets having the **same price** ($p = TCOD/|\mathcal{S}|$).
P2 - Datasets with **uniformly distributed random prices**.
P3 - Dataset prices **based on their Shapley values** [5, 36]

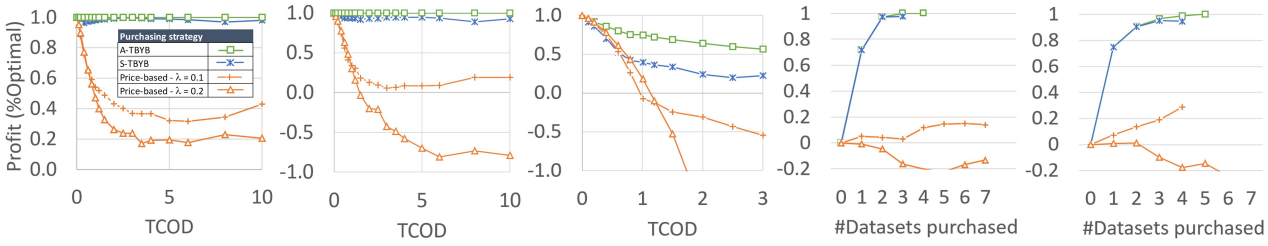
The Shapley value is a well-known concept in cooperative game theory, which we apply here to coalitions of data sources contributing value to a ML task. Shapley stated that the value of a player $\phi(s_i)$ in a coalition $\mathcal{S}$, with $s_i \in \mathcal{S}$ should be proportional to its average marginal contribution to any coalition of the other players it may join [36], i.e.,

$$\phi(s_i) = \sum_{K \subseteq \mathcal{S} \setminus \{s_i\}} \frac{|K|!(|\mathcal{S}| - |K| - 1)!}{|\mathcal{S}|!} [v(d(S_K) \cup d(s_i)) - v(d(S_K))], \quad (5)$$

where $K \subseteq \mathcal{S} \setminus \{s_i\}$ takes the value of all possible coalitions of sources, excluding $s_i$.

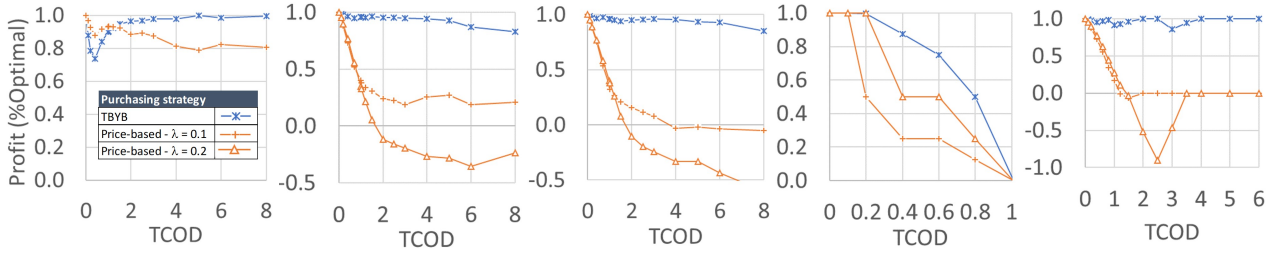
The Shapley value is widely used in ML [34] due to its salient properties (efficiency, symmetry, linearity, null player, etc.). However, it entails serious computational challenges since it requires to repetitively train and evaluate the model for any possible coalition of data sources in $\mathcal{S}$. Some context-specific approximation algorithms have managed to reduce this complexity to $O(N^2)$.

Next we will compare TBYB to the optimal purchase, and to other value-agnostic heuristics, and measure the buyers profit for different values of parameters like MUP, DI, TCOD and the relative price of datasets. Whenever randomization is used, e.g., in pricing datasets or in some of the value-agnostic heuristics, we report average values of 50 executions.




(a) MUP = 0.5 (concave) (b) MUP = 1 (linear) (c) MUP = 3 (convex) (d) TCOD = 1.5) (e) TCOD = 3)

**Figure 2: Profit vs. TCOD for different MUP and value-unrelated prices (a-c). Purchase sequences for MUP = 1 (d-e))**




(a) DI = 1, Random (b) DI = 2, Random (c) DI = 3, Random (d) DI = 2, value-based (e) DI = 2, Equitable

**Figure 3: Profit (MUP = 1) for different DI and pricing schemes**

## 4.2 Results for different utility profiles

Assuming that some datasets are more important than others ($DI$ = 2), Fig. 2 (a-c) shows the outcome of purchasing process following different strategies compared to the optimal for different MUP. A-TBYB matches the optimal purchasing for both concave (MUP=0.5, left subplot) and linear (MUP=1, middle subplot) value profiles, across the entire range of TCOD values. Similarly, the simpler S-TBYB, stays above 90% of the optimal profit for concave and linear MUPs, even when the performance of value-agnostic heuristics drops below 50% and even leads to losses when $MUP \geq 1$. Under convex value profiles (MUP=3), all strategies perform worse, since reaching a higher accuracy (and therefore buyer value) requires buying more datasets, which, in turn, eats away the profit margins for buyers. Even in these cases, TBYB yields a profit and avoids losses if $TCOD \gg 1$.

To explain why TBYB outperforms the value unaware heuristics, we plot in Fig. 2 (d-e) average "*purchase sequences*", showing the buyers' average profit as they purchase datasets using different algorithms. TBYB algorithms buy both the most valuable datasets (they achieve higher accuracy from the first round), and the right number of them (they stop buying before profits decrease). On the contrary, value-unaware heuristics overbuy, and randomly select affordable datasets respecting the buyer's risk appetite, which generally leads to lower profits, or even loses for risk-prone buyers.

## 4.3 The effect of data interchangeability

Figure 3 (a-c) shows the relative profit of different purchase algorithms for different DI values assuming MUP=1. The subplot on the left depicts results for perfectly interchangeable datasets (DI = 1), whereas the next two show cases of datasets that are increasingly less interchangeable (DI=2 and DI=3). These plots show that the performance benefits of TBYB over the heuristics increase when eligible datasets have different value, alone and combined with other datasets. On the contrary, knowing the value of data beforehand is less advantageous when they are almost interchangeable.

## 4.4 The effect of data pricing

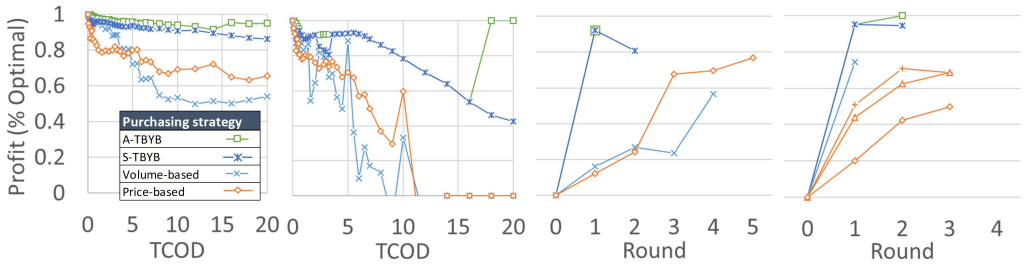
Subfigures 3 (b, d and e) show the performance of our purchasing algorithms for different pricing models. A-TBYB always matches the optimal purchase. Pricing based on value significantly reduces the gap between S-TBYB vs. price-based purchasing strategies, although S-TBYB still outperforms price-based purchasing for the same level of risk. Notice, however, that pricing data in accordance to their actual value for buyers is not realistic, since such depends on the task and the buyer, who has no incentives to disclose it to sellers.

## 4.5 Summary

TBYB, even in its simplest, stand-alone version, always outperforms the value-unaware heuristics, especially in the most realistic scenarios involving high TCOD, concave value functions, non-interchangeable datasets, and value-unaware pricing. In most cases, TBYB is very close to the performance of an optimal purchase using full information.

# 5 VALIDATION WITH REAL DATA

In reality, however, different datasets may mix in much more complex ways, that cannot be represented by any parameter setting of the above model. For example, $d_i$ can be very useful if combined with $d_j$, but not so useful if combined with others that individually yield the same accuracy as $d_j$. To validate our conclusions from




(a) Value-unrelated pric. (b) Volume-based pric. (c) TCOD = 3 (d) TCOD = 8

**Figure 4: Profit (a,b) and purchase sequences for volume-based prices (c,d) when predicting demand using B2B data**

the previous section, we tested the performance of different data purchase strategies using real spatio-temporal data reported by taxi companies and individual drivers, which we use to forecast demand and predict travel time within a city.

Furthermore, we introduce volume-based data pricing, which is quite common in practice [3]. *Volume-based pricing* assumes the price of a dataset is proportional to its volume (drivers or rides). We also test volume-based purchasing.

## 5.1 Demand prediction based on B2B data

We will assume that data buyers are looking to predict vehicle-for-hire demand in a city based on historical data, for which they can acquire data from a number of taxi companies (data sellers) through a DM. For this purpose, we have used real data reported by taxi companies to the regulatory authority of Chicago [27] containing 11.1 MM rides corresponding to the first 8 months of 2019 for $N = 16$ companies - the 15 largest ones servicing 94% of the total demand, plus a hypothetical 16th company aggregating the rest of the data.

We computed the exact Shapley value of the data of each company for the task of predicting demand using a multiseasonal SARIMA model with hourly and daily sub-components. The model was trained to predict taxi demand in the second half of March (control period) at the level of districts, for which we trained it with taxi rides from the previous six weeks (observation period). We compared the prediction to the ground truth given by the real data provided by all the 16 companies in the control period. We deliberately chose to predict the taxi demand of a medium size district (community area 11, Jefferson Park), where companies need to combine data to achieve a good prediction accuracy, and their Shapley values are very different (std. dev. = 76% of the average). The maximum accuracy the DM was able to give using all the information was $a^\star = 0.8963$ in this case.

We have simulated all purchase algorithms for different TCOD, pricing models, and $\lambda$ parameters. Figure 4a shows that both A-TBYB and S-TBYB achieve above 90% of the optimal buyer's profit under value-unaware prices. A-TBYB outperformed S-TBYB because it addresses the complexity of combining datasets at the (extra computational) cost of recursively asking for accuracy estimations every purchasing round. The results are in line with the ones we obtained using synthetic data.

Regarding volume-based purchasing, it proved even less effective than price-based purchasing. This is because value and volume are not significantly correlated in this particular case ($R^2 = 0.54397$), hence buying depending on the nº drivers does not necessarily lead to higher accuracy.

Figure 4b shows the results corresponding to volume-based prices. Profit reduces faster as TCOD grows in this case than in the case of value-unaware prices. This is because more valuable and larger datasets command higher prices. Still TBYB selects cheaper more valuable data.

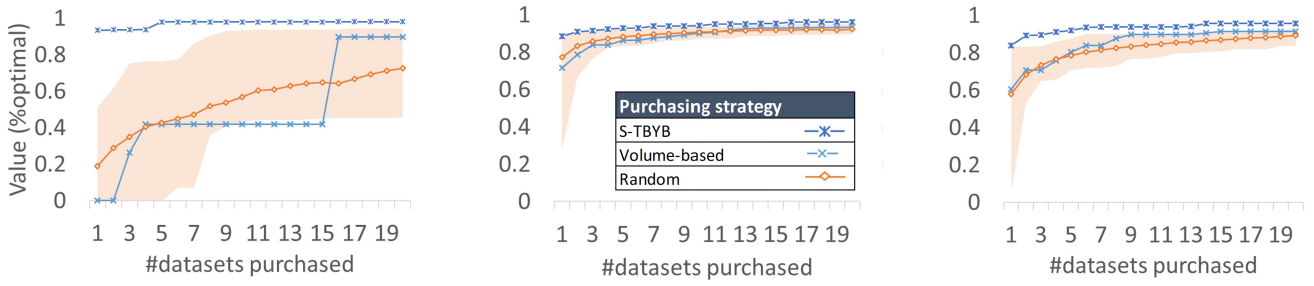
Figures 4c and 4d show purchase sequences that illustrate why TBYB works for volume-related prices. TBYB improves price-based purchasing both by selecting the best datasets, and by stopping before profit reduces. This feature is especially relevant when data prices grow (TCOD >> 1). Picking datasets based on volume does not improve price-based purchasing in this specific example.

## 5.2 Prediction based on individuals data

Some authors have proposed that individuals should get paid for their data [24, 32]. PIMS have appeared to empower users get their data under control and allow them to decide who they share their data with (and at what price). Current PIMS act as data unions and either value similar data equitably (same price for every user) or based on their data volume. PIMS do not disclose the identity of individuals but at most anonymized ids and some metrics (e.g., the volume of their data), and buyers must select whose data to purchase.

In this case, buyers' profits depend on the number of people whose data they purchased and on the accuracy achieved. Consequently, an algorithm that is able to make better predictions by picking less drivers, or less volume of data if prices are based on volume, will generally be more profitable. In this section, we will show that TBYB allows buyers to better filter individual whose data suits their task.

For this purpose, we consider data of 2.845 drivers in the same setting we used for demand prediction of B2B data. At city level, volume-based pricing strategy matches TBYB and is able to reconstruct the demand from data of 5 drivers. However, the situation is different at the level of districts. Figure 5a shows the purchase sequences of the different algorithms in district 56 of Chicago (Garfield Ridge), the shaded region delimiting percentile 0.1 and 0.9 of 50 executions picking drivers randomly. Whereas TBYB is able to reconstruct the demand with only 5 drivers, picking them by volume requires 49, a similar amount to the average drivers required if picked randomly. Therefore, regardless the price, TBYB always achieves higher accuracy with less drivers that the




**(a) Demand pred. Chicago** **(b) Travel Time pred. Porto Airport** **(c) Travel Time pred. Porto São Bento**

**Figure 5: Value vs. nº of people whose data is purchased**

other algorithms. If prices are set by volume, TBYB leads to even bigger profits, because volume-based purchasing heads buyers for the most expensive drivers in this case.

We also validated our findings in a completely different setting, a random forest model that is able to predict car travel time within the metropolitan area of Porto trained on a dataset detailing the 1.71 million routes of 448 taxis from Jul'13 to Jun'14 [21]. We randomly split input data between train (80%) set and test (20%) set, and trained the model to predict car travel time based on the UTM coordinates of the start and end points, the time of the year (month, weekday, hour), the type of day (i.e., workday, holidays or day before holidays) and the travel distance. The model was able to provide reasonably good accuracy ($R^2$ score above 0.8) when we evaluated predictions using the test set.

Similar to the previous example, we obtained the value delivered by the different purchasing strategies when predicting car travel time to the airport (see Fig. 5b), and to São Bento station (see Fig. 5c). Similar to the previous case, TBYB outperforms volume-based and price-based purchasing strategies, by selecting the most suitable drivers. In predicting travel time to the airport, TBYB achieves more than 0.95 of the maximum accuracy with barely 5 drivers, whereas picking drivers randomly requires 25 drivers, and volume-based purchasing requires 30 drivers to get to this value.

Our empirical results validate the conclusions of the previous simulations based on synthetic data. Not only is TBYB able to provide buyers with more value, but it also reduces the number of individuals whose data is shared with third parties, hence reducing unnecessary data leakages.

## 6 RELATED WORK

Different efforts have been done by the research community in order to identify the challenges of existing data marketplaces and define new business models [3, 11]. Most ML-oriented theoretical DM proposals leverage a data valuation framework similar to what we propose in this paper [2, 13]. In general, such marketplaces train buyers' models in a neutral platform by feeding them with their data, and ask for a price depending on the accuracy (and, thus the value) they provide. Some of them assume collaboration between the parties involved in training the algorithm, and approach the field of federated learning [28]. To the best of our knowledge, there is no practical nor commercial implementation of those designs yet. There is an ongoing debate regarding whether DMs should return trained models, insights or bulk data [14]. Some companies, such as SigOpt or Comet.ml, provide algorithm optimization services, federated learning is already available in key cloud platforms, and some works talk about Federated Learning as a Service (FLaaS) [22].

Furthermore, some researchers have studied the dynamics of DMs and proposed mechanisms to prevent strategic behaviour of sellers and buyers [10, 25]. Still, lots of challenges remain open around data trading and pricing [31, 37, 42].

Finally, we have used the Shapley value and its approximation algorithms to price datasets according to their value [5, 9, 17, 20, 34]. Shapley values are often proposed as a fair distribution of payoffs of data transactions among the sources that contributed data. As an alternative, some works propose to use the core [40].

## 7 CONCLUSION AND FUTURE WORK

TBYB was shown to yield near optimal profits to buyers under a wide range of parameters and data. Used with off-the-shelf ML algorithms, or more complex ones in a sandboxed "try before you buy" infrastructure, can make TBYB a practical, high performance alternative to value-unaware purchasing, which is currently the normal in real-world DMs. All of this in the hope that helping buyers select suitable data and achieve higher profits would thus help in bootstrapping and growing the currently nascent data economy.

We are currently working on developing fully functional design of a data marketplace based on TBYB to be implemented by real DMs. This involves additional aspects not covered here. For example, TBYB involves an implicit computation cost for which someone has to pay. On the one hand, charging buyers for it avoids abuse of TBYB functionality and promotes efficient valuation tasks on the buy side, at the cost of reducing the overall benefit of buyers using TBYB. On the other hand, charging sellers part of this processing protects against bad-quality useless datasets. Other relevant additional aspects are dynamic data pricing, protection against arbitrage, price discrimination, data engineering, scalability, and security challenges that will be the focus of forthcoming works. Moreover, we are testing data valuation functions that can evaluate the quality of data in different settings, and guide the purchasing process without the buyers revealing the model to the data marketplace.

## 8 ACKNOWLEDGEMENTS

This paper has been funded by the European Union's Horizon 2020 Research and Innovation program under the DataBri-X (Project No. 101070069) project.

# REFERENCES


[1] Advaneo. 2022. Advaneo. Access to the world of data. https://www.advaneo-datamarketplace.de/. Last accessed: Sep'22.

[2] A. Agarwal, M. Dahleh, and T. Sarkar. 2019. A Marketplace for Data: An Algorithmic Solution. In *Proc. of ACM EC'19*.

[3] S. Andrés and N. Laoutaris. 2022. A Survey of Data Marketplaces and Their Business Models. *ACM SIGMOD Record* 51, 3 (2022).

[4] S. Andrés Azcoitia, C. Iordanou, and N. Laoutaris. 2021. What Is the Price of Data? A Measurement Study of Commercial Data Marketplaces. (2021). arXiv:2111.04427

[5] S. Andrés Azcoitia, M. Paraschiv, and N. Laoutaris. 2022. Computing the Relative Value of Spatio-Temporal Data in Data Marketplaces. *SIGSPATIAL'22* (2022).

[6] Battlefin. 2022. Better Your Investments Using Alternative Data. https://www.battlefin.com/. Last accessed: Sep'22.

[7] Y. M. Brovman, M. Jacob, N. Srinivasan, S. Neola, D. Galron, R. Snyder, and P. Wang. 2016. Optimizing Similar Item Recommendations in a Semi-Structured Marketplace to Maximize Conversion. In *Proc. of RecSys'16*.

[8] CARTO. 2022. Marketplace. https://marketplace.carto.com/me. Last accessed: Sep'22.

[9] J. Castro, D. Gomez, and J. Tejada. 2009. Polynomial calculation of the Shapley value based on sampling. *Computers and Operations Research* 36 (05 2009).

[10] R. Castro Fernandez. 2022. Protecting Data Markets from Strategic Buyers. In *Proceedings of SIGMOD'22*.

[11] R. Castro Fernandez, P. Subramaniam, and M. J. Franklin. 2020. Data Market Platforms: Trading Data Assets to Solve Data Problems. *Proc. VLDB Endow.* 13, 12 (2020).

[12] S. Chawla, S. Deep, P. Koutris, and Y. Teng. 2019. Revenue maximization for query pricing. *Proc. VLDB Endow.* 13 (09 2019).

[13] L. Chen, P. Koutris, and A. Kumar. 2019. Towards Model-Based Pricing for Machine Learning in a Data Marketplace. In *Proc. of SIGMOD'19*. ACM.

[14] M. Dahleh. 2018. Why the Data Marketplaces of the Future Will Sell Insights, Not Data.

[15] EU. 2016. General Data Protection Regulation (GDPR).

[16] GeoDB. 2022. A Decentralized Big Data Ecosystem That Rewards You For The Data You Generate. https://geodb.com/. Last accessed: Sep'22.

[17] A. Ghorbani and J. Zou. 2019. Data Shapley: Equitable Valuation of Data for Machine Learning. (04 2019).

[18] S&P Global. 2022. Marketplace. https://www.marketplace.spglobal.com/en/. Last accessed: Sep'22.

[19] IOTA. 2022. IOTA data marketplace. https://data.iota.org/. Last accessed: Oct'22.

[20] R. Jia, D. Dao, B. Wang, F. A. Hubis, N. Hynes, N. M. Gürel, B. Li, C. Zhang, D. Song, and C. J. Spanos. 2019. Towards Efficient Data Valuation Based on the Shapley Value *(Proc. of ML Research, Vol. 89)*.

[21] Kaggle. 2015 (accessed Sep'22). ECML/PKDD 15: Taxi Trajectory Prediction. https://www.kaggle.com/c/pkdd-15-predict-taxi-service-trajectory-i/data

[22] N. Kourtellis, K. Katevas, and D. Perino. 2020. FLaaS: Federated Learning as a Service. In *Proc. of Workshop on Distributed Machine Learning*.

[23] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu. 2015. Query-Based Data Pricing. *J. ACM* 62, 5 (2015).

[24] J. Lanier. 2013. *Who Owns the Future?* SIMON and SCHUSTER.

[25] D. Moor. 2019. Data Markets with Dynamic Arrival of Buyers and Sellers. In *Proc. of NetEcon '19*.

[26] State of California. 2018. California Consumer Privacy Act (CCPA).

[27] City of Chicago. 2019 (accessed Sep'22). Taxi Trips. https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew

[28] O. Ohrimenko, S. Tople, and S. Tschiatschek. 2019. Collaborative Machine Learning Markets with Data-Replication-Robust Payments. *ArXiv* (2019). arXiv:1911.09052

[29] Otonomo. 2022. One-Stop Shop for Vehicle Data. https://otonomo.io/. Last accessed: Sep'22.

[30] Hubert P. and Ricco G. 2018. Imperfect information in macroeconomics. *Sciences Po publications* (2018).

[31] J. Pei. 2020. Data Pricing – From Economics to Data Science. In *Proc. of SIGKDD'20*. ACM.

[32] E. Posner and G. Weyl. 2018. *Radical Markets. Uprooting Capitalism and Democracy for a Just Society.* Princeton Univ. Press.

[33] Refinitiv. 2022. Data Catalog. https://www.refinitiv.com/en/financial-data. Last accessed: Sep'22.

[34] B. Rozemberczki, L. Watson, P. Bayer, H. Yang, O. Kiss, S. Nilsson, and R. Sarkar. 2022. The Shapley Value in Machine Learning. arXiv:2202.05594

[35] Amazon Web Services. 2022. AWS Marketplace. https://aws.amazon.com/marketplace. Last accessed: Sep'22.

[36] Lloyd S. Shapley. 1952. A Value for n-Person Games. (1952). https://www.rand.org/pubs/papers/P0295.html

[37] Y. Shen, B. Guo, Y. Shen, X. Duan, X. Dong, and H. Zhang. 2016. A pricing model for Big Personal Data. *Tsinghua Science and Technology* 21 (10 2016), 482–490.

[38] Shutterstock. 2022. Shutterstock. https://www.shutterstock.com/. Last accessed: Oct'22.

[39] Snowflake. 2022. Marketplace. https://www.snowflake.com/marketplace/. Last accessed: Sep'22.

[40] Yan T. and Procaccia A. 2020. If You Like Shapley Then You'll Love the Core.

[41] TAUS. 2022. Data Marketplace. https://datamarketplace.taus.net/. Last accessed: Oct'22.

[42] J. Yang, C. Zhao, and C. Xing. 2019. Big Data Market Optimization Pricing Model Based on Data Quality. *Complexity* 2019 (04 2019).