Understanding the Price of Data in Commercial Data Marketplaces

1nd Santiago Andrés Azcoitia
IMDEA Networks Institute
Universidad Carlos III de Madrid
Leganés, Spain
santiago.azcoitia@imdea.org

2rd Costas Iordanou
Cyprus University of Technology
Limassol, Cyprus
kostas.iordanou@cut.ac.cy

3th Nikolaos Laoutaris
IMDEA Networks Institute
Leganés, Spain
nikolaos.laoutaris@imdea.org

***Abstract*—A large number of Data Marketplaces (DMs) have appeared in the last few years to help owners monetize their data, and data buyers optimize their marketing campaigns, train their ML models, and facilitate other data-driven decision processes. In this paper, we present a first of its kind measurement study of the growing DM ecosystem, focused on understanding which features of data are actually driving their prices in the market. We show that data products listed in commercial DMs may cost from few to hundreds of thousands of US dollars. We analyze the prices of different categories of data and show that products about telecommunications, manufacturing, automotive, and gaming command the highest prices. We also develop classifiers for comparing data products across different DMs, as well as a regression analysis for revealing features that correlate with data product prices of specific categories, such as update rate or history for financial data, and volume and geographical scope for marketing data.**

***Index Terms*—Data economy, data marketplaces, measurement, data pricing**

I. Introduction

Data-driven decision making powered by Machine Learning (ML) algorithms is changing how the society and the economy work and is having a profound positive impact on our daily life. A McKinsey report predicted that data-driven decision-making could reach US$2.5 trillion globally by 2025 [30], whereas a recent market study within the scope of the European Data Strategy estimates a size of 827 billion euro for the EU27 [14]. ML is driving up the demand for data in what has been called the fourth industrial revolution.

To satisfy this demand, several data marketplaces (DMs) have appeared in the last few years. DMs are mediation platforms that aim to connect data providers (acting as sellers) to data consumers (acting as potential buyers), and to manage data transactions between them. This ecosystem includes open data repositories [28], [33], general-purpose [2], [7], [18], [19], [21], and specialized or niche DMs targeting specific industries, such as automotive [13], [50], financial [8], [55], marketing [41], [42], and logistics [65], to name a few.

An issue of paramount importance is that of *data pricing*. Some marketplaces leave it to sellers to set a price for their data products. Many of them do not list prices of their products, but leave it to buyers and sellers to agree on a price after a negotiation. Due to the elusive nature of the traded "commodity", pricing is a very complex matter, even more than in the case of material goods [53]. Unlike oil, to which it is often compared [17], data can be copied / transmitted / processed with close to zero cost. Even the use of the term commodity is a gross oversimplification of what data is. Notice that whereas two liters of gasoline yield a similar mileage on two similar cars under similar driving styles, nothing of this sort applies to data since 1) two datasets of equal volume may carry vastly different amounts of usable information, 2) the same information may have tremendously different value for Service A than for Service B, and 3) even if the per usage value of two services is the same, Service A may use the data 1,000 times more intensely than Service B leading to extremely different produced benefits. Some authors compared data to labor, too [6]. However, unlike labor, data is non-rivalrous meaning that its supply is not affected by its consumption, and thus selling data for a Service A does not prevent a provider from selling (a copy of) the same data for a Service B.

The research community at the intersection between computer science and economics has studied several aspects of data pricing. Still its elusive nature, and the complex business models under which it is made available makes it very hard to prescribe a price for data. Ultimately it is the market that decides and sets prices via complex mechanisms and feedback loops that are hard to capture. Despite some other works trying to measure the price of personal data of individuals [12], [43], [51], there is no systematic measurement study about the price of data products traded in commercial data marketplaces.

Our Contributions: In this paper we present what is, to the best of our knowledge, the first systematic measurement study of marketplaces for B2B data products. This ecosystem, despite being quite vibrant commercially, remains completely unknown to the scientific community. Very basic questions such as "*What is the range of prices of data traded in modern DMs?*", "*Which categories and types of data products command the highest prices?*", "*Which are the features, if any, that correlate with the most expensive data products?*" appear to have no answer and evade most meaningful speculations.

Our research has been supported by MLEDGE project (REGAGE22e00052829516), funded by the Ministry of Economic Affairs and Digital Transformation and the European Union-NextGenerationEU/PRTR, and by the European Union's HORIZON project DataBri-X (101070069).

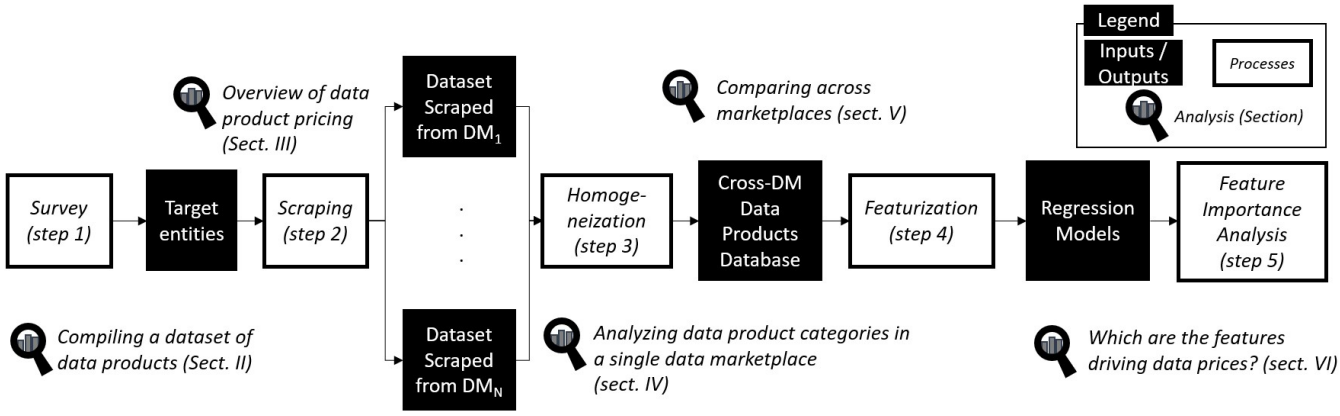

Fig. 1: Summary of our methodology

To answer such questions we followed the methodology summarized in Fig. 1. First, we checked existing surveys for compiling a list of data marketplaces [4], [57], [60]–[62]. We then selected 10 of them that fulfill necessary criteria for a measurement study. For these ones we developed custom crawlers for retrieving information about the products they trade. Using these crawlers, and adding the portfolio of another 30 data providers, we obtained information for more than 210,000 data products and a catalog of more than 2,100 distinct sellers[1]. We also developed data product category classifiers, meaning ML models for identifying products of similar categories across marketplaces, and executed 9 different regression models to understand which features are actually driving their prices.

Our Findings: Analyzing the collected data we observed that the majority of data products were either given for free, or did not carry a fixed price, but rather were up for direct negotiation between the seller and interested buyers. Focusing on the ones that carried a price, some 4,200 of them, we observed that:

- Prices vary in a wide range from few, to several hundreds of thousands of US dollars. The median price for data products sold under a *subscription* model is US$1,400 per month, and US$2,200 for those sold as an *one-off* purchase.
- Using classifiers, we enriched our sample by consistently labeling products according to AWS's categories.
- We found that those related to *telecoms*, *manufacturing*, *automotive* and *gaming* command the highest median prices, and that the most expensive ones relate to *retail and marketing*.
- Using regression models, we managed to fit the prices of commercial products from their features with R^2 above 0.84.
- Due to the heterogeneity of the sample there is no single feature that drives the prices, but instead we spotted meaningful features that drive the prices of specific categories of data. For example, data update rate is a key price driver for *financial* and *healthcare*-related products, whereas geo-spatial localization and the possibility of connecting data points from the same owner are for *marketing* data.
- Overall our models use features related to the category and description of the different data products (i.e., 'Financial', 'Retail', 'stock', 'contact', 'list', etc.), features related to the data products volume and units, as well as singular characteristics extracted from the data products description (i.e., words like 'custom', 'accuracy', 'quality', etc.) to forecast the data product price. Features related to '*what*' and '*how much*' data a product contains are driving 66% of its price.

Like in all measurement studies of Internet-scale phenomena, we will refrain from claiming that any of our findings are "typical" or "representative". What we do claim, however, is that to the best of our knowledge, our measurement study is the first one that attempts to characterize the DM sector, and our above mentioned quantitative results were previously totally unknown. Also, as it will become evident from our methodology later, and to the best of our knowledge, we collected all publicly available pricing information that was accessible during the time of our study.

The remainder of the paper is structured as Fig. 1 shows. First, we frame the scope of our analysis and show some initial outcomes of our measurement study in Sect. II. In Sect. III, we present an analysis on data product pricing in commercial marketplaces. Furthermore, Sect. IV dives deeper into analyzing AWS' DM and Datarade, which account for the largest number of price references in our sample. We then develop tools for enriching our sample and compare across DMs in Sect. V. Finally, in Sect. VI, we apply several methodologies for analyzing the importance of different metadata features in determining the price of commercial data products.

II. Compiling a dataset of data products

Existing works and surveys on commercial data marketplaces [4], [57], [60]–[62], an extensive web search and a consultation with experts in the area allowed us to compile a list of data marketplaces and understand the different business models they use to compete in this ecosystem. From our analysis, we identified a subset of DMs that fulfilled the criteria for using them as sources of data for a reproducible measurement study. Such criteria include that they grant access to their product catalog without requiring an account, or

[1]Please, find datasets generated during our research, and code to reproduce our experiments at https://gitlab.com/sandresazcoitia1/data-pricing-tool.

through an account but without a vetting process or upfront paid registration, that they have a reasonably large catalog that includes sufficient descriptions of their data products, and that they include a clear description of their pricing policy. Out of the 180 initial DMs, only 10 companies fulfilled all of the above criteria. Most of them did not make it to the list simply because they do not allow non-paying users to browse their catalogs. For example, marketing-related private marketplaces such as *Liveramp*, *LOTAME* or *TheTradeDesk* neither provide public per-product information nor any price references. However, they do provide information about their data partners. By analyzing this information, we did find that 45% of providers in those private marketplaces sell through general-purpose public ones, such as *AWS* or *DataRade*, as well, and hence we have included their products in this study. We also discarded several otherwise *scrapable* general-purpose DMs such as *Data Intelligence Hub* (DIH), *Google Cloud DM* because they included only free data products. We chose to scrape the largest of these free open data marketplaces, *Advaneo*, to help in training our data product category classifiers.

TABLE I: Summary of scraped DMs

Marketplace	#Products	#Paid prod.	#Sellers
Advaneo	198,743	1	N/A
AWS	4,263	2,674	262
DataRade	1,592	1,592	1,262
Snowflake	889	889	200
Knoema	158	158	142
DAWEX	160	160	79
Carto	8,182	5,283	42
Crunchbase	9	9	15
Veracity	115	95	38
Refinitiv	187	187	76
Other providers	777	775	30

Table I lists the 10 DMs that we use as data sources in our study. Overall, we include 6 general-purpose and 4 niche DMs, as well as 30 data providers[2] that, in addition to commercializing their own 777 data products through DMs, provide valuable pricing information on their own websites.

We developed our own web crawler to render and download web pages, and specialized parsers for extracting metadata. We followed common crawling good practices [31]. For example, we avoided visiting several times the same product page in each scraping round and we set up a random wait time from 1 to 2 minutes after requesting a web page in order to avoid flooding the target servers with requests.

We collected information related to 215,075 products from 2,115 distinct sellers in total. We noticed the huge market fragmentation with lots of data providers working with a large number of marketplace platforms. This is natural in a cross-industry nascent market, though hard for data providers to manage. In fact, most data providers (81%) work with only one DM in addition to selling their products through their own web site. 45% of providers in niche financial and marketing-related marketplaces sell through general-purpose DMs, such as *AWS* or *DataRade*, as well. We also spotted DMs advertising and offering their products in other DMs (e.g., *Battlefin* or *CARTO* through *AWS*). Finally, small and niche providers (58% of them) are focusing on one product only.

We scraped all available metadata for data products such as the product id, title, description, source, seller and, when available, its geographic scope, volume, category, use cases, update rate, historic time span, format, etc. We searched for and eliminated duplicates from a single seller within the same DM. We paid special attention to information related to pricing and actual prices of data products.

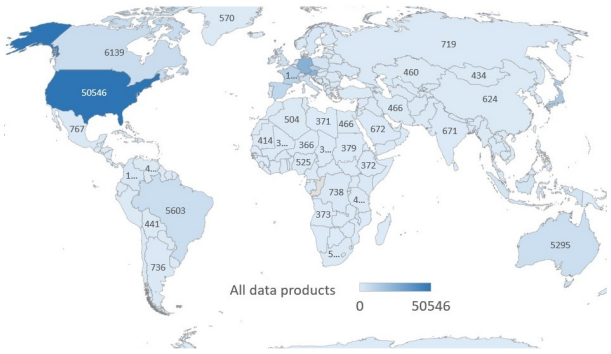

Fig. 2: Data products by country

Regarding the geographical scope of data products, we found that DMs aggregate information from different countries. 14,472 (7%) of the products did not inform about their scope, and 1,177 (around 10% out of the 11,823 paid products) claimed to be global. Figure 2 shows the number of data products covering each country. Regarding the number of *paid* data products, US leads this ranking: around 30% of paid products cover this country. Canada (9.3%), UK (9.2%), Germany (7.6%), France (7.4%), and Spain (7.1%) follow the US in the ranking of countries by number of *paid* products.

III. Overview of Data Product Pricing

It may appear initially surprising that, despite being commercial entities in the B2B space, most of the surveyed and some of the scraped DMs offer predominately free (most of the time open) data. Again we point to the fact that these are privately held companies [2], [21] and not open data NGOs or government initiatives. Our conjecture is that since DMs are two-sided platforms, pre-populating them with free data is a very reasonable bootstrapping strategy, since it can attract the initial "buyers", which in turn will attract commercial sellers and thus help the marketplace grow its revenue.

Next, we focus on the 11,823 paid data products, for which we managed to extract information about their pricing, and whose price is higher than zero. Despite being few compared to the free ones, this sample provides valuable insights about the current status of commercial DMs, as well as to where this segment of the economy is heading to, and how.

[2]42matters, Airbtics, Apptopia, Benzinga, Bizprospex, BoldData, BookYourData, bronID, BuiltWith, DataScouts, Demografy, ebCard, Enigma, ESGAnalytics, HGXN, IFDAQ, ipinfo.io, MultimediaLists, MyDex, OikoLab Weather, Onclusive. Open Corporate, PanXchange, Pipecandy, Shutterstock, Storm Glass, TelephoneListsBiz, Unwrangle, USASalesLeads, and Walklists.

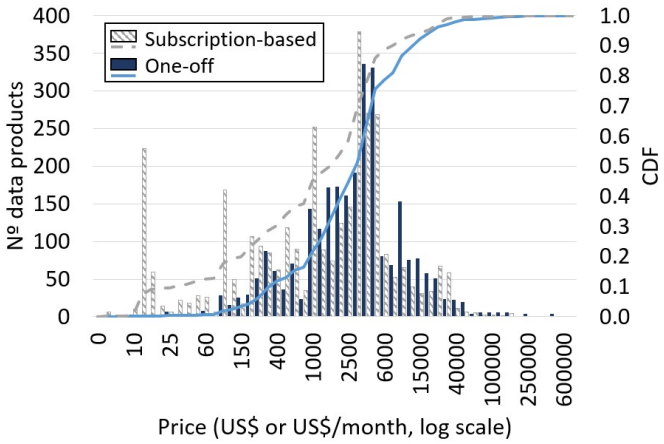

Fig. 3: Histogram and CDF of data products

There is a great magnitude of pricing schemes for data products, such as seller-led, buyer-led (bidding), revenue-sharing, tiered-pricing, subject to negotiation, usage-based, etc [4], [44], [53]. Predominant among the 11,823 non-free data products are the *subscription-based* model (i.e., buyer paying for a subscription to get access to data for a period of time), and the *one-off* model (i.e., lump sum payment for data), seller-led in both cases. The first one is used mostly for "live" data usually accessed via an API (e.g., IoT sensor data), whereas the second is used for more static data, which are usually downloaded as one or more files.

4,162 products from 443 distinct providers provided clear information about their prices. Figure 3 shows a histogram and the corresponding CDF of monthly prices for data products. Regarding those offered under a *subscription model*, we see prices across a wide range up to US$150,000 per month. Cheap products below US$100 per month are often curated and cleaner versions of open data. For example, a seller offers a historical compilation of quarterly reports submitted to the US Securities and Exchange Commission (SEC), also downloadable from their websites. They also include low-cost "promotion samples" of more expensive products from well-known sellers, such as GIS data and supporting metadata for a small area of some US cities. The median price is US$1,417 per month. Almost one-third of all products, including targeted market data for example, are sold for US$2-5k monthly.

Comparing to products sold under a *one-off model*, (1) the latter tend to be more expensive: median price US$2,176 vs. US$1,417 per month for *subscription-based* products; maximum price US$500,000, more than 3 times higher than the maximum in *subscription-based* access, and (2) *one-off* products have a price histogram more normally distributed around its median at US$2,176. Within the heterogeneous set of products within the US$1,000-4,000 interval, we found a large group of voluminous targeted contact data products. Interestingly, we observe a long tail of valuable data products in Fig. 3. We will come back to them later.

IV. Analyzing data product categories in a single data marketplace

To get a more in-depth understanding of data pricing, we analyzed the catalog of AWS' DM, the one with the largest base of paid products with prices. AWS classifies data products by *category*. Specifically, a product can belong to none, one, or several categories corresponding to industries or sectors of the economy. For instance, credit cards transaction data products are classified both as '*Financial*' and '*Retail, Location and Marketing*', whereas weather related ones are not labeled. We mark such unclassified products as '*Other*'.

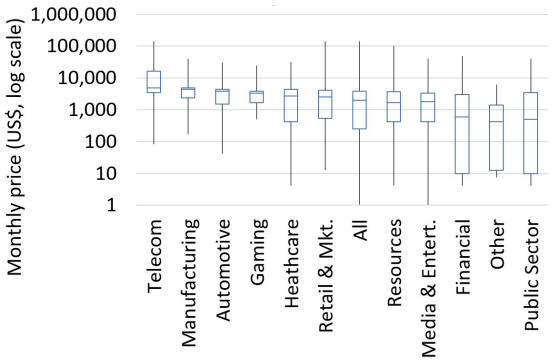

Fig. 4: Subscription prices by industry in AWS.

V. Comparing across marketplaces

Comparing information about data products from different marketplaces is not a straightforward task since i) they provide metadata of different granularity and level of detail, and ii) they use different categorization to describe their products. To overcome these challenges, we developed a methodology to homogenize the categorization of data collected in order to be able to compare similar products across marketplaces.

A. Dealing with different levels of detail

Some marketplaces provide more information than others about their offers. To sort this out, we built a common cross-DM database utilizing a superset of all the different description fields found in different data marketplaces. Apart from their category and text descriptive fields, data product records include the time scope, the volume and units, any potential limitations (e.g., maximum number of users), add-ons, granularity of the information, geo-scope at country level, data delivery methods, update frequency and data format.

We normalized and stored in this cross-DM database all the information from the scraped datasets. We managed to fully automate the extraction of most of the fields (18 out of 27), which were directly scraped from the web pages of the different DMs. This extraction was semi-automated for 5 fields, meaning that they were automatically extracted for certain marketplaces, or retrieved from product descriptions for others, in a process that required a manual check afterwards. For example, *update rate* of data is usually included in the general description of a data product, but the presence of the

word '*monthly*' may not necessarily point to a monthly update rate. Information about data volume or data subject units was automatically extracted only for DataRade and BookYourData, and required computer-aided manual typing in the rest of the DMs (we highlight and extract numbers and their context from data descriptions). Manual checks were performed by three different experts. Any ambiguities and disagreements were resolved by majority voting.

B. Dealing with different categorization systems

Every marketplace has its own way to classify data. In this case, AWS tags data products in 10 different categories, whereas DataRade allows data products to be positioned in a hierarchy with more than 300 categories and more than one (out of 150) use cases. Furthermore, boundaries between tags are often blurry, and the criteria followed by different DMs to label a data product with a certain category tag are not necessarily coherent. For example, only certain marketplaces mark '*credit card transaction*' data products as '*financial*', whereas all DMs label them as related to '*marketing*'. Thus, even if we find apparently comparable categories across different marketplaces, we may miss relevant data products due to inconsistencies in their categorization processes.

We addressed this issue by developing a series of natural language processing naïve Bayes (NB) classifiers [20], [22], [39]. In our first attempt, we wanted to identify similar data products – those that belong to the same category – between two different (source and destination) DMs. As a result, we trained both multinomial and complement versions of NB classifiers to detect data products from the source DM that belong in a certain category by using feature vectors based on the information provided by the data product description from the source DM. We used bag of words [36] and data preprocessing steps such as removing stop words and words with numbers, using stemming and TF-IDF transformation [47], [56]. Then we validated the resulting classifier against a manually labeled sample from the destination DM. Manual labeling was performed by three different experts. Any ambiguities and disagreements were resolved again by majority voting.

We utilized the above methodology to build different classifiers to help us compare data products between the two DMs including more price references, namely DataRade (destination DM) and AWS (source DM). We generated our feature vectors based on AWS data product descriptions (source DM) and applied the resulting classifiers to DataRade data products (destination DM). We were interested in finding out: (1) what percentage of products from those categories could we identify in DataRade, (2) whether categorization and pricing were coherent between them, and (3) whether we could enrich our metadata by adding AWS's inferred categories to all products.

We utilized our cross-DM database to generate the train/test datasets at 80/20 split in order to train and test the corresponding classifiers. We observed that multinomial classifiers outperformed the complement NB for this task so we proceeded with the former ones. The resulting classifiers yield an acceptable F_1 score above 0.85 (average for 50 executions with different random 80/20 train/test splits). In fact, they identified meaningful and reasonable stems when tagging products related to each category. For example, for the two categories including more data products:

Financial: 'system', 'sec', 'exchang', 'type', 'file', 'form', 'edgar', 'secur', 'act', and 'compani'.

Retail, Location and Marketing: 'locat', 'topic', 'b2b', 'score', 'echo', 'trial', 'compani', 'visit', 'intent', 'consum'.

We then validated the models against a manually labeled sample from DataRade. Manual labeling was performed by three different experts. Any ambiguities and disagreements were resolved again by majority voting. The validation set included 745 manually pre-labeled with both '*Financial*' and '*Retail, Location and Marketing*' tags. The models trained only with data from AWS did not perform so well on the validation set (F_1 scores of 0.73 and 0.43 for '*Financial*' and '*Retail, Location and Marketing*' data). To generalize further our methodology and improve its accuracy, we enriched the train data with information from other DMs. In particular:

(1) The ***Financial*** classifier was trained with 95,208 labeled descriptions of products from 4 different entities (Advaneo, Carto, AWS, and Refinitiv), and 45,298 financial products.

(2) The ***Retail, Location and Marketing*** classifier was trained with 3,828 descriptions from 3 entities (AWS, BookYourData and TelephoneLists), including 1,614 marketing products.

By adding products belonging to the same category from other DMs we observed better balance between precision and recall and an overall improvement of model generalization. We also observed an increase of the F_1 score in the test set. Particularly, adding information from Refinitiv improves the F_1 score from 0.73 to 0.79. In the case of '*Retail, Location and Marketing*', adding information from specialized marketing DMs (e.g., BookYourData), drastically improves the F_1 score from 0.43 to 0.74. We tested multiple classifiers, with and without stemming, and we found that using word-based instead of stem-based features led in general to more accurate results in both cases (+5% F_1 score). Table II shows the accuracy obtained by both classifiers.

TABLE II: Score of data product classifiers

	Accuracy	**Precision**	**Recall**	F_1 **Score**
Test - Financial	0.93	0.97	0.81	0.88
Test - Retail	0.95	0.96	0.88	0.91
Val. - Financial	0.89	0.72	0.88	0.79
Val. - Retail	0.78	0.81	0.68	0.74

We used them to label data products in DataRade. As a result, we located 619 and 701 '*Financial*' and '*Retail, Location and Marketing*' data products in this DM, which represent 39% and 44% of the total sample, respectively. As happened in AWS, not only do those categories contain the largest number of products in this marketplace, but the most expensive ones are tagged as '*Retail, Location and Marketing*', as well. We repeated the process for the rest of the 11 AWS data categories, and thus we managed to enrich our sample by homogeneously labeling products based on their descriptions.

Does this methodology work if we switch source and destination DMs? In order to answer this question, we trained NB classifiers to detect products in AWS related to relevant use cases and categories in DataRade. In this case, DataRade acted as the source DM, i.e., it provided descriptions and tagging information to train the classifiers, whereas AWS' role was the destination DM, whose products we labeled with some of DataRade's tags and driven by the criteria we learned from the source DM. In particular, we focused on products belonging to the '*B2B Marketing*', '*Audience Targeting*' and '*Risk Management*' use cases in DataRade, some 46, 48 and 30 products out of 745 respectively. Since the training set is imbalanced and the number of samples is low, complement NB outperformed multinomial NB in this case. We trained the classifiers and obtained the log-probability of belonging in each category for all the data products in AWS. As a result, at least 16 out of the top 20 data products showing the highest log-probability turned out to be useful for those specific use cases, according to the assessment of three different experts.

VI. Which are the features driving data prices?

So far we have seen an overview of data pricing, looked at the prices of particular categories, developed and applied a methodology to homogeneously label products across marketplaces in our sample. Our final goal is to understand the prices of data in commercial data marketplaces.

For that purpose, we first extract features to train regression models for predicting the prices of real commercial data products. We do not intend to build state-of-the-art price predictors, but rather to understand which features are driving the price of data. Therefore, we conduct feature importance analysis on the resulting regression models and we find out which features have the highest impact on the observed prices for the different data products in our corpus.

A. Building a feature matrix to feed regression models

An additional preprocessing step is needed in order to transform the fields of our cross-DM database into a set of valuable features that can be ingested by ML regression algorithms. This process uses the NLTK [9] and Scikit-learn [52] Python libraries and includes mainly the following steps:

1) Extraction of '*word*' features from the title and the textual description of each data product. We use bag of words [36] and data preprocessing steps such as removing stop words and words with numbers, TF-IDF transformation [56], and stemming [47]. In addition, we have sellers' names removed from the vocabulary, so as to avoid bias introduced by knowing their identity. Finally, we prepare matrices for different vocabulary lengths and optimize each algorithm for this parameter.
2) Breakdown of volume-related fields in 13 different groups depending on their nature. For example, we separate data products targeting 'entities' or 'companies', from those whose subjects are 'individuals' in different features. The resulting comparable units are in turn normalized, and a new overarching feature ('units') measuring the percentage of units covered is added to compare products across groups of units.
3) Calculation of country-level binary features to indicate whether a certain country is covered by a data product.
4) Homogenization of the units of time when measuring the time scope of the products, what we will call *history*.

Before feeding the models, we reduce the number of input features by discarding those that have a unique value, which may appear when filtering the complete dataset by *category*. Next, we unify groups of features showing a high cross-correlation among them, i.e., $R^2 \geq 0.9$.

As a result of this *featurization* process, we reduce each sample product to a feature vector and produce a feature matrix to train our regression models. Table III lists feature groups and some examples of their individual features. We organize features in 10 disjoint sets according to their nature and the basic questions they answer about data products.

We evaluated the linear correlation of individual features with respect to data product prices. Not surprisingly, it turns out that none of them is linearly correlated to price, as opposed to what we found for specific sellers. Our challenge now is measuring which features and groups of features are more significant in determining the price of data products in commercial marketplaces.

B. Analyzing feature importance

Regression models can be used for feature importance analysis. Next we use a range of such techniques to understand which features have the higher impact on data product prices.

1) Optimizing Regression models: Owing to their stochastic nature, training several regression algorithms and comparing their outcomes is key to obtaining robust conclusions. Consequently, we have tested variations of 9 different regressors with different values for their main parameters (e.g., num. of estimators, depth, etc.) as included in the Scikit-learn [52] Python library, and inputs of different vocabulary lengths. Such models work with the log instead of the absolute value of product prices as the dependent variable so as to normalize the distribution of prices and avoid negative price predictions. We were hoping to find at least 3 models that produce sufficiently accurate price predictions, measured as the $R^2 score$ of their output w.r.t. actual prices.

To reduce the complexity of each model, we removed low-value features, i.e., those that had a negative leave-one-out (LOO) value, provided the accuracy of the model was not negatively affected. A feature having negative LOO value means that the model improved its average accuracy in 10 random executions for different train and test data splits when such feature was removed from the input matrix. Finally, we performed a cross-validation to check the variance of the accuracy of the model when training and testing in 5-folds, and 20-random training-test splits of the input data.

We found that three target models worked reasonably well (i.e., they yield an R^2 score greater or equal to 0.70), namely Random Forest [10], k-Nearest Neighbours [38], and Gradient Boosting [23], [46] regression models. On the contrary,

TABLE III: List of feature groups

Question	Group	Definition	Nº features	Example of features
What?	**Category**	Labels attached to the product that define the type of data it contains	11	'Weather', 'Gaming', 'Financial'
	Description	Stem-like features obtained from data product descriptions	up to 2000	'wordmarket', 'wordidentifi', 'wordlist'
	Identifiability	Tells whether the product allows the buyer to recognize the activity of individuals or to identify specific companies	2	'idSessions', 'IdCompanies'
How much?	**Volume**	Normalized nº units covered broken down by the nature of such units	14	'units', 'people', 'entities'
	Update rate	Defines the frequency between data updates as announced by the seller	11	'real time', 'monthly', 'hourly'
How?	**Delivery method**	Defines how the buyer can have access to data	8	'S3Bucket', 'Download', 'FeedAPI'
	Format	Defines the way in which data is arranged	17	'txt', 'shapefile', 'xls'
	Add-ons	Tells whether the product attaches any add-on or has any limitations	2	'ProfServices', 'Limitations'
When?	**History**	Time scope included	1	'History'
Where?	**Geo scope**	Metrics about countries included in the data product	up to 249	'Nº Countries', 'USA', 'Canada'

TABLE IV: Accuracy achieved by regression models

Model	Financial			Marketing			Healthcare			All		
	R^2	MAE	MSE	R^2	MAE	MSE	R^2	MAE	MSE	R^2	MAE	MSE
RF	0.85	0.2	0.14	0.86	0.21	0.13	0.78	0.25	0.15	0.84	0.23	0.16
kN	0.78	0.31	0.26	0.74	0.33	0.24	0.77	0.26	0.17	0.69	0.37	0.31
GB	0.82	0.23	0.16	0.8	0.28	0.19	0.73	0.27	0.19	0.79	0.3	0.22
DNN	0.73	0.33	0.35	0.77	0.30	0.22	0.68	0.26	0.18	0.72	0.33	0.28

TABLE V: Top 10 most relevant features not related to volume by category and regression model

Financial			Marketing			Healthcare		
RF	**kNN**	**GBR**	**RF**	**kNN**	**GBR**	**RF**	**kNN**	**GBR**
S3Bucket	Email	S3Bucket	IdSessions	History	csv	wordhealth	csv	wordlist
wordsubmit	Download	wordmonthli	Download	USA	yearly	wordtrend	daily	Del. Methods
Download	daily	wordstock	REST API	IdSessions	REST API	wordmedic	wordmarket	wordhospit
txt	IdCompanies	worddeliv	wordcustom	Nº Countries	wordqualiti	wordglobal	wordgo	wordidentifi
wordedgar	USA	Del. Methods	USA	Financial	wordaccur	csv	Limitations	wordamerica
wordcustom	wordmarket	txt	yearly	Others	wordidentifi	Del. Methods	location data	wordhealth
wordlist	Retail	wordneed	monthly	wordcontact	wordwebsit	wordinsight	wordpopul	wordreport
wordcontact	wordcontact	wordsubmit	IdCompanies	Email	UI Export	wordreport	wordprofil	wordstudi
wordsystem	real time	wordreport	wordname	UI Export	wordcover	wordregion	wordinsight	wordupdat
wordcompar	wordprice	wordcontact	location data	Download	wordfield	wordlist	Download	wordcontact

we discarded linear, Elastic-Net [68], Ridge [32], Bayesian Ridge [45], and Lasso [64] regressions even though they worked well in specific simulations.

In addition, we also tested a Deep Neural Network regressor using the TensorFlow [1] and Keras [34] libraries. We followed all common good practices recommended for such activity by first standardizing the input data. We tested RELU/Leaky RELU activation functions for all hidden layers, and a linear activation function for the output layer. As loss function we used the mean absolute error (MAE). To avoid overfitting we randomly applied Drop-out between training epochs and to avoid dying/exploding neurons we also applied Batch normalization between all layers. We used the Adam optimizer [35] with a tuned learning rate decay to train the model faster at the beginning and then decrease the learning rate with further epochs to make training more precise. Finally, we used Callbacks to stop the training at the optimal epoch.

Table IV presents a summary of the accuracy obtained by regressor and category of data products, including the R^2 score, the MAE and the mean squared error (MSE) with regards to the actual log prices. For the sake of robustness, our results were consistent across subsequent 5-fold and 20 random train/test splits: R^2 score showed a standard deviation below 4% of the average in each round. Note that due to the total (low) number of observations that we have in our datasets, DNN models are not recommended, nevertheless, we wanted to explore them since we believe that they will further improve our results as soon as we manage to increase the overall size of our datasets. Consequently, we avoided using any DNN model in the feature importance analysis.

2) Analyzing the importance of individual features: We carried out this process for financial, marketing, healthcare and all data products in our sample. Financial and marketing data were the most popular data categories, whereas healthcare data was chosen as a relevant disjoint category of less though increasingly popular products showing a different behavior in terms of prices. As a result, we obtained at least one model that achieves a R^2 score of 0.78 by category and accurately fits the prices of data products (see Tab. IV). We ran two different individual feature importance analysis:

1) measuring the accuracy lost by randomly shuffling the values of a certain feature among samples (permutation importance analysis [63]), and
2) measuring the prediction accuracy lost when one individual feature is removed from the inputs (leave-one-out or LOO value)

We have found that 50% of the positive LOO and 67% of the ΔR^2 score by shuffling values owe to the top 10 most relevant features on average for specific categories of data. Note that we would need more than 25 features to achieve equivalent scores if we include all the products. Whereas features related to units and the volume of data clearly lead the ranking for financial and marketing data products, they are less important for healthcare-related ones.

We cross-validated our results in 5-fold executions of both methods and took averages in order to disregard features that showed to be important only in specific tests. As regards robustness, we compared the top-20 ranking of every individual test to the top-20 average ranking of that algorithm and category. It turns out that both rankings have at least 5 features in common in 95% of the cases, and a median of 13 common individual features.

Table V lists other features not related to data volume in descending order of importance. Next we provide some details about the most important features of each specific category:

Financial: Not only do volume-related features such as '*units*' and '*entities*' rank number one, but they are on average four times more important than the second feature in the ranking. Other features relate to specific characteristics of financial data products and help models identify data products either by their category (e.g., '*Retail*') or their description. For instance, RF relies on the word '*edgar*', which stands for SEC's Electronic Data Gathering, Analysis, and Retrieval System, all algorithms identify business '*contact*' lists, a family of financial products, and they also use '*stock*' and '*market*'. The word '*custom*' helps identify information about *customers*, but also refers to the valuable possibility of personalizing data products (e.g., select which companies we want financial data from). Features related to delivery methods (e.g., '*S3bucket*' or '*Download*') and update rate (e.g., '*real time*' or '*daily*') stand out in terms of relevance, as well.

Marketing: With regards to marketing data products, features related to volume, such as '*units*' and '*entities*' lead the ranking, as well. Again categories (e.g., '*Financial*', '*Others*') and specific words pointing to relevant characteristics of data play a relevant role, too. For example, words like '*contact*' are used to locate contact lists, a family of marketing products, the stems '*qualiti*' and '*accur*' refer to the high-quality and accuracy of data, as advertised by sellers. A number of features, such as the stem '*identifi*', emphasize the value of identification for marketing data. In addition, the presence of '*IdSessions*' and '*IdCompanies*' features indicates that being able to reconstruct sessions of anonymized individuals and being able to identify merchants are price drivers for marketing products. Unlike financial data, the fact that a dataset includes *'location data'* is also used to set prices of marketing data. Finally, the scope of data is important, as suggested by features like '*USA*' and '*Nº Countries*' ranking high in the results of RF and kNN models.

Healthcare: The '*what*' is more important than the '*how much*' when fitting the observed prices of healthcare products. This is due to the heterogeneity of data products belonging in this category, ranging from contact lists of healthcare practitioners and hospitals to data about clinical trials or specific medications. Therefore, stems like '*trial*', '*hospit*' or '*studies*' help in identifying what a dataset is about. The stem '*go*' refers to an official check-in and rating system that was used to limit the spread of COVID in the US. Features related to the update rate, data format ('*csv*'), the number of available delivery options ('*Del. Methods*') and the presence of '*Limitations*' (e.g., limited number of reports, or limited data exports included) determine product prices, too.

3) Analyzing the importance of groups of features: Since LOO is often negligible for individual features, we have repeated this analysis for groups of features answering to the same question regarding the data product (see Tab. III). In this case, we have used the following two methods:

1) Measuring the prediction accuracy lost when a group of features is removed from the input dataset (LOO).
2) Measuring the average (in 20 random train/test split executions) Shapley value of each group of features.

The Shapley value is defined as the average R^2 score added by combining the information of a certain group of features with every possible mix of the rest of groups. This is a well-known and widely-used concept in game theory, economics and ML [24], [59], and it is considered a 'fair' method to distribute the gains obtained by cooperation. In our case, we applied the Shapley value to distribute the gains in accuracy of our regression models among the groups of features that contributed to achieving such an accuracy. Furthermore, we ran 5-fold feature importance analysis in the case of LOO, in a similar way as we did for individual features, and 20 calculations of the Shapley values for random 80/20 train/test splits of our input data.

Whereas LOO measures gains or loses in accuracy of a model when features belonging in a group are removed from the input matrix, Shapley values better capture the complementarity among groups and take into consideration their individual predictive power, as well. Table VI and Table VII list the LOO and the Shapley values by group of features in descending order of importance. The standard deviation of Shapley values across executions is acceptable (average below 0.029 for financial and marketing datasets, 0.057 for healthcare-related data, and below 0.017 for all the data), and the ranking of relevant feature groups remains stable.

Figure 5 plots the percentage of the sum of Shapley and LOO values that each feature group represents, what we call their *predictive power*, and illustrates how important each group is for determining the prices of each category of products. We have piled together and colored in gradients groups responding to the same question about data products.

Note that the algorithms, in the absence of certain features, try to replace or infer them through other features in order to come up with the best estimation possible. We have observed that this happens with '**category**' labels or '**add-ons**', and it is also the reason why LOO values are generally smaller than the corresponding Shapley values.

TABLE VI: LOO values by feature group

Group	Financial			Marketing			Healthcare			All		
	RF	kNN	GBR	RF	kNN	GBR	RF	kNN	GBR	RF	kNN	GBR
Description	0.027	0.025	0.066	0.021	0.034	0.098	0.054	0.425	0.052	0.023	-0.020	0.079
Volume	0.092	0.182	0.167	0.171	0.138	0.199	0.048	0.014	0.052	0.138	0.123	0.142
Geo Scope	-0.005	-0.007	-0.001	-0.003	-0.006	0.000	0.015	0.000	-0.011	-0.003	-0.002	0.000
Del. Method	0.005	0.032	0.011	0.000	0.018	0.008	0.019	0.017	0.003	0.002	0.010	0.008
Format	0.002	0.004	0.010	0.007	0.001	0.023	0.007	0.030	0.000	0.002	0.007	0.006
Category	-0.002	0.001	0.001	-0.001	-0.003	0.001	0.013	-0.033	-0.006	0.001	0.000	0.003
Add-ons	-0.001	0.007	-0.001	-0.001	0.000	0.001	0.000	0.022	0.000	0.001	0.001	0.000
Identifiability	-0.002	0.016	0.002	-0.001	0.006	0.004	0.010	0.000	-0.009	0.000	0.008	0.000
History	-0.001	0.000	0.000	-0.003	0.004	0.000	0.009	0.000	0.000	0.002	0.000	-0.001
Update Rate	0.001	0.023	0.001	0.036	0.000	0.016	0.010	0.021	0.000	0.021	-0.002	0.014

TABLE VII: Shapley values by feature group

Group	Financial			Marketing			Healthcare			All		
	RF	kNN	GBR	RF	kNN	GBR	RF	kNN	GBR	RF	kNN	GBR
Description	0.155	0.266	0.222	0.247	0.153	0.152	0.232	0.290	0.236	0.113	0.176	0.187
Volume	0.211	0.216	0.184	0.290	0.241	0.241	0.168	0.125	0.131	0.211	0.210	0.174
Format	0.087	0.006	0.086	0.027	0.046	0.094	0.090	0.077	0.082	0.072	0.087	0.071
History	0.072	0.000	0.059	0.009	0.037	0.036	0.063	0.001	0.046	0.058	0.010	0.037
Update Rate	0.088	0.056	0.084	0.060	0.032	0.050	0.046	0.145	0.041	0.067	0.034	0.067
Del. Method	0.036	0.054	0.044	0.093	0.075	0.049	0.030	0.040	0.035	0.062	0.062	0.074
Identifiability	0.034	0.038	0.028	0.052	0.027	0.048	0.040	0.001	0.031	0.056	0.022	0.039
Geo Scope	0.056	0.046	0.050	0.032	0.044	0.036	0.030	0.001	0.040	0.061	0.015	0.024
Category	0.071	0.021	0.044	0.018	0.043	0.037	0.017	0.031	0.039	0.070	0.063	0.055
Add-ons	0.021	0.003	0.021	0.012	0.028	0.038	0.048	0.053	0.041	0.055	0.026	0.045

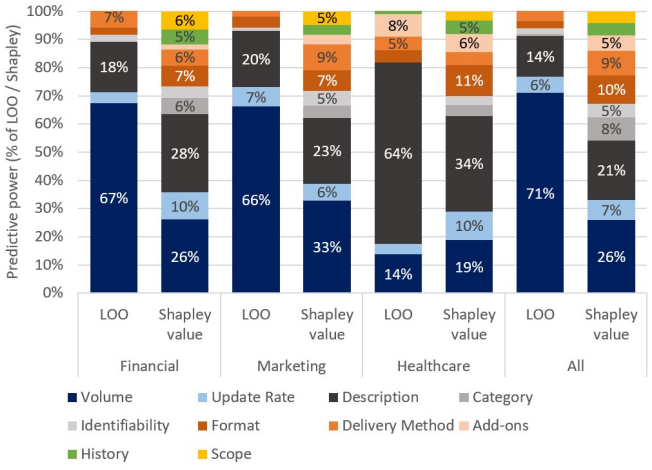

Fig. 5: Predicting power of feature groups

By looking at Fig. 5, we can confirm that features related to '**volume**' and '**descriptions**' are the most relevant groups driving data prices: at least half of the predictive power owes to those two groups of features according to their Shapley values. While '*volume*' is clearly the most relevant group for marketing data products, it is not so relevant for healthcare-related data due to the heterogeneity of products belonging in this category and due to their lower price-sensitivity to volume.

Data '**update rate**' and its '**format**' are consistently relevant across all data categories, but to a lesser extent (6-11% of the prediction score), whereas the Shapley values of the other groups differ across categories: '**history**' (meaning the time span of data delivered) is more relevant for financial and healthcare-related data, '**delivery methods**' are more relevant for marketing data, and ´**identifiability**' is important in general, but especially for marketing products. These results are in line with our discussion based on the relevance of individual features in the previous section.

In summary, it is mostly '*what*', as captured in product description and categories, and '*how much*' data is being traded that determine the price of a product. Since relevant descriptive features are diverse and strongly differ across data categories, we failed to find a single feature other than '*units*' that, with some aforementioned exceptions, consistently shows a significant *predictive power*. However, we did find interesting features driving the prices of specific categories of data, such as update rate for financial products, and the ability to provide exact locations and those related to identifiability for marketing data. '*How*' data is delivered to buyers proved to be important too, and accounts for 15-24% of *predictive power* according to Shapley. Finally, historical time span ('*when*') and geographical scope ('*where*') of data products, whose score oscillates around 5% for every data category, are less relevant in driving their prices.

VII. Related works

Even though several surveys related to data marketplaces have been recently published [4], [57], [60]–[62], our work is, to the best of our knowledge, the first empirical measurement study that deals with the prices of data products sold in commercial data marketplaces.

In fact, the lack of empirical data around dataset prices is considered as a key challenge in data pricing research [53]. According to some authors, some techniques to set the prices of digital products [58] or cloud services [66] are applicable to data products, as well. Some authors proposed auction designs to set the prices of digital goods and data products [26], [27]. Novel AI/ML data marketplace architectures have been proposed under the concept of value-based pricing [3], [16], [49] and the value of privacy [48]. Moreover, some authors defined pricing strategies and marketplaces based on differential privacy [25], [40] or queries to a database [15], [37]. All of them work on analyzing the theoretical properties for fair, arbitrage-free pricing, but leave the responsibility of actually

defining absolute prices to both buyers and sellers. Quality-based pricing [29] is the one closest to our approach. According to it, the value of data must be assessed by evaluating and assigning weights to certain quality features. Even though some additional works have provided data pricing strategies for sellers based on this idea [67], we are not aware of any measurement study that has been able to derive weights for such features from real market data.

The pricing of personal data of individuals has received attention from the privacy and measurement communities. There are measurement studies based on prices carried over the Real Time Bidding protocol [43], [51] as well as more traditional survey-based studies [12]. These works report prices for the data and the attention of individuals and, therefore, have nothing to do with B2B datasets traded in modern DMs.

Cross-marketplace analysis and discoverability of data has been pointed out as a significant challenge by data marketplace vision papers [54]. Google Dataset Search has proposed a standard for providing metadata for their crawlers [11]. Discoverability is the *leit motiv* of DMs and data aggregators, such as DataRade, but do not touch upon pricing questions.

Finally, part of this work explaining the challenges in scraping and comparing across data marketplaces and outlining the design of a data quotation tool was published as a workshop paper [5]. This paper is adding substantially new material, such as the procedures we used to populate a cross-DM database, the development and test of classifiers to compare across DMs (see Sect. V), and the training of regression models to fit data prices and carry out feature importance analysis (see Sect. VI).

VIII. Conclusions and future work

Our work has provided a first glimpse into the growing market for B2B data. Despite having worked in a range of pricing topics in the past, prior to conducting this study, we did not have the slightest idea even for fundamental questions such as "What are typical prices for data products sold online?", or "What types of data command higher prices?". Our work has produced answers to those and many other questions. We have seen that while the median price for data is few thousands, there exist data products that sell for hundreds of thousands of dollars. We have also looked at the categories of data and the specific per-category features that have the highest impact on prices. Having scraped metadata for hundreds of thousands of data products listed by 10 real-world data marketplaces and other 30 data providers we found fewer than ten thousand that were non-free and included prices. We believe that this is due to prices being often left to direct negotiation between buyers and sellers, and also because most marketplaces use free data to bootstrap their marketplace and attract the first "buyers" and then commercial sellers.

Moreover, the paper represents a first step towards developing a price recommendation tool for new data products [5], and has even provided a first implementation of some of its key components, namely i) the metadata and taxonomy required to describe data products, ii) crawlers and parsers to automate the collection of such information from key leading marketplaces, iii) classifiers to compare across them, and iv) regression models to understand which are the most relevant features driving product prices. The significant monthly growth rate we have seen at AWS and other marketplaces makes us believe that in the future the paid catalog of data marketplaces is bound to grow and therefore, we will continue monitoring them to see how they evolve.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Tal- war, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale Machine Learning on Heterogeneous Systems, 2015. Software available from tensorflow.org.

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

[3] A. Agarwal, M. Dahleh, and T. Sarkar. A Marketplace for Data: An Algorithmic Solution. In Proc. of ACM EC, 2019.

[4] S. Andrés Azcoitia and N. Laoutaris. A Survey of Data Marketplaces and their Business Models. SIGMOD Record, 2022.

[5] S. Andrés Azcoitia, C. Iordanou, N. Laoutaris, "Measuring the Price of Data in Commercial Data Marketplaces," ACM Data Economy Workshop, 2022.

[6] I. Arrieta-Ibarra, L. Goff, D. Jiménez-Hernández, J. Lanier, and E. G. Weyl. Should we Treat Data as Labor? Moving Beyond "Free". AEA Papers and Proceedings, 108:38–42, 2018

[7] AWS. Amazon Web Services Marketplace. https://aws.amazon.com/marketplace. Last accessed: Oct'22

[8] Battlefin. Better your investments using alternative data. https://www.battlefin.com/. Last accessed: Oct'22.

[9] E. L. Bird, Steven and E. Klein. Natural Language Processing with Python. O'Reilly Media Inc, 2009.

[10] Breiman. Random forests. Mach. Learn., 45(1):5–32, Oct. 2001.

[11] D. Brickley, M. Burgess, and N. Noy. Google Dataset Search: Building a Search Engine for Datasets in an Open Web Ecosystem. In Proc. of ACM WWW conf., 2019.

[12] J. P. Carrascal, C. Riederer, V. Erramilli, M. Cherubini, and R. de Oliveira. Your Browsing Behavior for a Big Mac: Economics of Personal Information Online. In Proc. of ACM WWW Conf., 2013

[13] Caruso. Your solution. one platform. multibrand in-vehicle data. https://www.caruso-dataplace.com/. Last accessed: Oct'22.

[14] G. Cattaneo, G. Micheletti, and al. The European Data Market Monitoring Tool. Key Facts and Figures, First Policy Conclusions, Data Landscape and Quantified Stories. Final Study Report. European Commission, 2020.

[15] S. Chawla, S. Deep, P. Koutris, and Y. Teng. Revenue Maximization for Query Pricing. Proc. of the VLDB Endow., 13, 2019.

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

[17] Clive Humby. Data is the New Oil! Keynote at ANA Senior Marketer's Summit, Kellogg School, 2006.

[18] DataRade. Datarade. choose the right data with confidence. https://datarade.ai/. Last accessed: Oct'22.

[19] Dawex. DAWEX Data Exchange, unleash the value of your data. https://www.dawex.com/. Last accessed: Oct'22.

[20] L. Denoyer and P. Gallinari. Bayesian Network Model for Semi-structured Document Classification. Inf. Process. Manage., 40(5), 2004.

[21] DIH. Data intelligence hub. extract value from data securely. https://dih.telekom.net/. Last accessed: Oct'22.

[22] P. Domingos and M. Pazzani. On the optimality of the Simple Bayesian Classifier under Zero-one Loss. Mach. Learn., 1997

[23] J. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics, 29, 2000

[24] A. Ghorbani and J. Zou. Data shapley: Equitable Valuation of Data for Machine Learning. Proc. of the ICML, 2019.

[25] A. Ghosh and A. Roth. Selling privacy at auction. In Proc. of the ACM EC '11, 2011.

[26] A. V. Goldberg and J. D. Hartline. Competitiveness via Consensus. In Proc. of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '03, page 215–222, USA, 2003. Society for Industrial and Applied Mathematics.

[27] A. V. Goldberg, J. D. Hartline, and A. Wright. Competitive Auctions and Digital Goods. In Proc. of the ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics, 2001.

[28] Harvard. Dataverse. https://dataverse.harvard.edu/. Accessed: Oct'22.

[29] J. R. Heckman, E. Boehmer, E. H. Peters, M. Davaloo, and N. G. Kurup. A Pricing Model for Data Markets. In Proc. iConference 2015.

[30] N. Henke, J. Bughin, and al. The age of analytics: Competing in a Data-driven World. McKinsey Global Institute, 2016.

[31] M. Hils, D. W. Woods, and R. Böhme. Measuring the Emergence of Consent Management on the Web. In Proc. of the ACM IMC'20, page 317–332. 2020.

[32] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased Estimation for Nonorthogonal Problems. Technometrics, 42(1):80–86, Feb. 2000.

[33] Kaggle. Datasets. https://www.kaggle.com/datasets. Accessed: Oct'22.

[34] Keras. Simple. flexible. powerful. https://keras.io/. Accessed: Jun '22.

[35] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, Proc of ICLR '15, 2015.

[36] Y. Ko. A Study of Term Weighting Schemes using Class Information for Text Classification. In Proc. of ACM SIGIR 2012.

[37] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu. Querymarket Demonstration: Pricing for Online Data Markets. Proc. of the VLDB Endow., 5, 2012.

[38] O. Kramer. Unsupervised k-Nearest Neighbor regression. 2011.

[39] G. Krishnaveni and T. Sudha. Naïve Bayes Text Classification - a Comparison of Event Models. Imperial Journal of Interdisciplinary Research, 3, 2016.

[40] C. Li, D. Y. Li, G. Miklau, and D. Suciu. A theory of pricing private data. ACM Transactions on Database Systems 39(4), 2015.

[41] LiveRamp. Data marketplace. https://liveramp.com/our-platform/data-marketplace/. Last accessed: Oct'22.

[42] LOTAME. Private data exchange (pdx). trusted data relationships made easy. https://www.lotame.com/pdx/. Last accessed: Oct'22.

[43] C. C. Lukasz Olejnik, Minh-Dung Tran. Selling off privacy at auction. In Proc. of the NDSS Symposium, 2014.

[44] A. Löser, F. Stahl, A. Muschalle, and G. Vossen. Pricing Approaches for Data Markets. In Proc. of the International Workshop on Business Intelligence for the Real-Time Enterprise, 2012.

[45] D. J. C. MacKay. Bayesian interpolation. Neural Comput., 4(3), 1992.

[46] L. Mason, J. Baxter, P. Bartlett, and M. Frean. Boosting Algorithms as Gradient Descent. In Proc. of the International Conference on Neural Information Processing Systems, 1999.

[47] S. Matic, C. Iordanou, G. Smaragdakis, and N. Laoutaris. Identifying Sensitive Urls at Web-scale. In Proceedings of the ACM IMC, 2020.

[48] C. Niu, Z. Zheng, F. Wu, S. Tang, X. Gao, and G. Chen. Unlocking the Value of Privacy: Trading Aggregate Statistics over Private Correlated Data. In Proc. of ACM SIGKDD, 2018.

[49] O. Ohrimenko, S. Tople, and S. Tschiatschek. Collaborative Machine Learning Markets with Data-replication-robust Payments. CoRR, 2019.

[50] Otonomo. One-stop shop for vehicle data. https://otonomo.io/. Last accessed: Oct'22.

[51] P. Papadopoulos, N. Kourtellis, P. R. Rodriguez, and N. Laoutaris. If you are not paying for it, you are the product: How much do advertisers pay to reach you? In Proc. of the ACM IMC, 2017.

[52] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. Journal of Machine Learning Research.

[53] J. Pei. Data Pricing – from Economics to Data Science. In Proc. of the ACM SIGKDD, page 3553–3554, 2020.

[54] M. F. Raul Castro Fernandez, Pranav Subramaniam. Data Market Platforms: Trading Data Assets to Solve Data Problems. In Proc. of the VLDB Endow., 2020.

[55] Refinitiv. Data catalog. our data, your way. https://www.refinitiv.com/en/financial-data. Last accessed: Oct'22.

[56] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. Info. Processing and Management, 1988.

[57] F. Schomm, F. Stahl, and G. Vossen. Marketplaces for data: An initial survey. ACM SIGMOD Record, 2013.

[58] C. Shapiro and H. R. Varian. Information Rules: A Strategic Guide to the Network Economy. Harvard Business School Press, 2000.

[59] L. S. Shapley. A Value for n-Person Games. RAND Corporation, 1952.

[60] M. Spiekermann. Data marketplaces: Trends and monetisation of data goods. Intereconomics, 2019.

[61] F. Stahl, F. Schomm, L. Vomfell, and G. Vossen. Marketplaces for digital data: Quo vadis? Computer and Information Science, 10, 2017.

[62] F. Stahl, F. Schomm, and G. Vossen. The Data Marketplace Survey Revisited. Westf. Wilhelms-Univ., ERCIS, 2014.

[63] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. Conditional Variable Importance for Random Forests. BMC Bioinformatics, 2008.

[64] R. Tibshirani. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society (Series B), 1996.

[65] Veracity. Veracity by DNV GL. Find the Right Tools for your Industry Needs. https://store.veracity.com/. Last accessed: Oct'22.

[66] C. Wu, R. Buyya, and K. Ramamohanarao. Cloud Pricing Models: Taxonomy, Survey, and Interdisciplinary Challenges. ACM Computing Surveys, 2019.

[67] H. Yu and M. Zhang. Data Pricing Strategy based on Data Quality. Computers and Industrial Engineering, 112:1–10, 2017.

[68] H. Zou and T. Hastie. Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society. Series B. (Statistical Methodology), 2005.