Measuring the Price of Data in Commercial Data Marketplaces

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

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

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

ABSTRACT

A large number of Data Marketplaces (DMs) have appeared in the last few years to help owners monetise 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, shedding light on several totally unknown facts about it. We show that data products listed in commercial DMs may cost from few to hundreds of thousands of US dollars. We analyse the prices of different categories of data and the challenges of comparing across DMs. We also analise the pricing of specific sellers and products to identify features that apparently correlate with prices, and we point to the need and the challenges of building a quotation tool for data products based on market data.

CCS CONCEPTS

• **Information systems** → **Data extraction and integration**; • **General and reference** → **Measurement**; • **Applied computing** → **Electronic data interchange**.

KEYWORDS

Data economy, data marketplaces, measurement, data pricing

ACM Reference Format:
Santiago Andrés Azcoitia, Costas Iordanou, and Nikolaos Laoutaris. 2022. Measuring the Price of Data in Commercial Data Marketplaces. In *Data Economy (DE '22), December 9, 2022, Roma, Italy.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3565011.3569053

1 INTRODUCTION

Machine Learning (ML) is driving up the demand for data in what has been called the fourth industrial revolution. A report predicted that the market around data-driven decision-making powered by ML algorithms could reach up to US$2.5 trillion globally by 2025 [22], whereas a recent market study within the scope of the European Data Strategy estimates a size of 827€ billion for the EU27 [9].

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/10.1145/3565011.3569053

To satisfy this demand, several data marketplaces (DMs) have appeared in the last few years, acting as mediation platforms between data providers (aka sellers) and potential buyers, and managing the transactions between them. This ecosystem includes open data repositories [19, 25], general-purpose DMs [1, 4, 13–15], and specialised or *niche* DMs for specific industries, such as automotive [8, 34], financial [5, 37], marketing [28, 29], and logistics [43], to name a few.

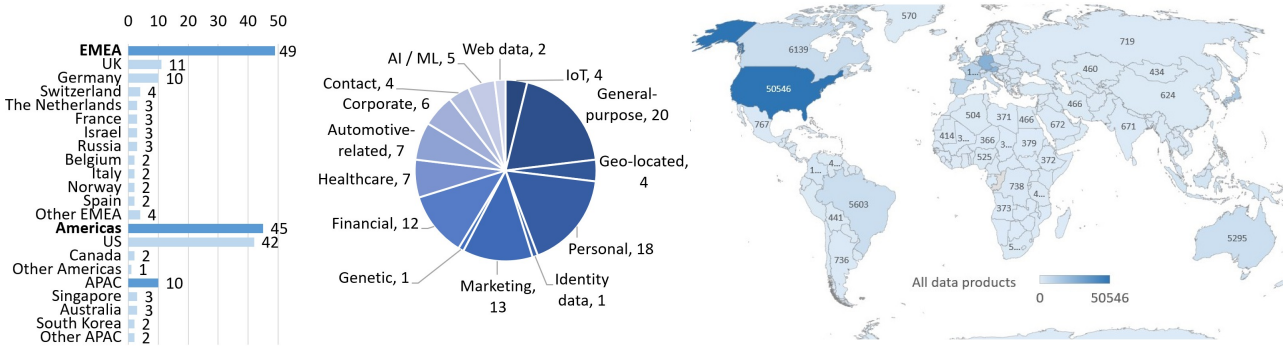
An issue of paramount importance is that of *data pricing*. Some marketplaces leave it to sellers to set a price for their data products, just like in any other marketplace for material goods or services. Many of them do not list prices of their products, but leave it to buyers and sellers to agree on a price following a negotiation process. Pricing data goods is a very complex matter due to their elusive nature. Unlike oil, to which it is often compared [12], data can be copied / transmitted / processed with close to zero cost. It is also a non-rivalrous good, meaning that selling (a copy of) data to a consumer A does not prevent the seller from selling it to another consumer B. Moreover, its value depends heavily on the context and can be very different for each consumer.

The research community at the intersection between computer science and economics has studied several aspects of data pricing [36], and tried to measure the value of personal data [7, 30, 35]. However, there is no systematic measurement study about the actual prices of B2B data products traded in commercial DMs.

Our Contributions: In this paper we present what is, to the best of our knowledge, the first systematic measurement study of DM 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.

To answer such questions we first conducted an extensive survey for compiling a catalogue with more than 190 DMs [3]. We then selected 15 of them that fulfill necessary criteria for a measurement study, for which we crawled information about the providers selling through them and their products. Adding the portfolio of another 30 data providers, we obtained information for more than 210,000 products from more than 2,100 distinct sellers. We studied data pricing by some vendors and the characteristics of the most expensive products to list features that appear to be driving data prices. Finally, we pointed to the challenges of building a quotation tool for products based on market data.

Our Findings: 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



(a) N° entities by Country (b) N° entities by data type (c) Data products by country

Figure 1: Scope of our survey on entities trading with data

buyers. Focusing on the ones that carried a price, some 4,200 of them, we observed the following:

- 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.
- Focusing on Amazon Web Services (AWS) DM we found that data related to *telecoms*, *manufacturing*, *automotive* and *gaming* command the highest median prices.
- We noticed that DMs such as AWS have been growing with a significant 3% monthly rate in 2021.
- We pointed to a number of features like data volume, its category, or its granularity that are driving the prices of data products in commercial data marketplaces.

The remainder of the paper is structured as follows. First, we frame the scope of our analysis in Sect. 2. In Sect. 3, we present a novel analysis on data product pricing in commercial marketplaces. Furthermore, Sect. 4 dives deeper into analysing AWS' DM, which accounts for more price references in our sample. We point to the challenges of comparing across DMs in Sect. 5. Next we identify metadata features either related to the most expensive products or used by sellers to set the price of commercial data products in Sect. 6. Finally, Sect. 7 presents the building blocks of a data quotation tool and introduces some challenges of such a development.

2 COMPILING DATA PRODUCTS

An extensive web search after a consultation with experts in the area allowed us to compile a list of 190 DMs [3]. Previous related surveys mention half the number of DMs [38, 40–42] and, therefore, we believe that we have covered a good part of the market. Out of those 190, we selected 104 for a more in-depth study. We discarded concept projects, online advertising platforms, and service providers not offering data products. Moreover, we balanced our selection in terms of the business models covered, and included entities from 17 countries that trade different types of data as Fig. 1 shows.

From our analysis of the aforementioned marketplaces we identified a subset that fulfilled a number of 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 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. 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* do not provide public per-product information nor any price references. Searching for information about data sellers using commercial DMs, we scraped five more *niche* DMs (these three, and two more *Quandl* and *Factset*, related to financial data). We also discarded several otherwise *scrapable* general-purpose DMs (*Data Intelligence Hub* (DIH) or *Google Cloud DM*) that include only free data products, except for the largest one, *Advaneo*, which we scraped to help in understanding data categorisation.

Table 1 lists the 15 DMs that we use as data sources in our study. Overall, we include 6 general-purpose and 9 niche DMs, and 30 providers commercialising 777 products and publishing their prices on their own websites.

We developed our own web crawler to render and download web pages, and specialised parsers for extracting metadata. We followed common crawling good practices [23]. 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 to avoid flooding 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 from data products, such as the product id, title, description, source, seller and, when available, its geographic scope, volume, category, use cases, update rate, historic

Table 1: Summary of scraped DMs

Marketplace	#Data Products	#Paid products	#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 spatial data catalog	8,182	5,283	42
Crunchbase Marketplace	9	9	15
Veracity	115	95	38
Refinitiv	187	187	76
Liveramp	N/A	N/A	135
LOTAME	16	16	126
TheTradeDesk	N/A	N/A	122
Quandl	N/A	N/A	49
Factset	N/A	N/A	84
Other data providers	777	775	30

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.

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 1c 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.

3 ABOUT DATA PRODUCT PRICING

It may appear initially surprising that, despite being privately-held commercial entities in the B2B space [1, 15], most of the surveyed and some of the scraped DMs offer predominately free (most of the time open) data. 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 DM grow its revenue.

Next, we focus on the 11,823 paid data products, for which we managed to extract some information about their pricing. Despite being few compared to the free ones, this sample provides valuable insights about commercial DMs.

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 [31, 36]. Predominant among the paid data products are the *subscription-based* model (i.e., 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 usually downloaded as one or more files.

4,162 products from 443 distinct providers provided clear information about their prices. Figure 2 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

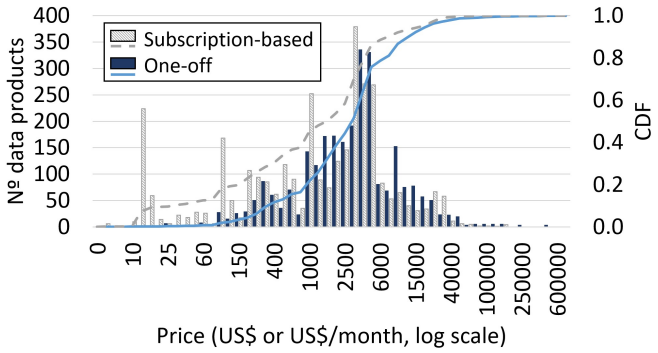


Figure 2: Histogram and CDF of data products

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 and reports for example, are sold for US$2,000-5,000 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. 2. We will come back to them later.

4 ANALYZING CATEGORIES IN AWS

To get a more in-depth understanding of data pricing, we analysed 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 labelled. We mark such unclassified products as '*Other*'.

Figure 3 shows a box plot of products by category in AWS. The X-axis shows the different categories ordered in decreasing median price, whereas the Y-axis represents the monthly price to get access to the data. '*Telecom*', '*Manufacturing*' and '*Automotive*' categories exhibit a median price significantly above the global (×2.6, ×2.3 and ×2, respectively). Most low-value products belong to the '*Public Sector*', '*Financial*' (e.g., stock price feeds), and ' *Other*' categories.

We also conducted a temporal analysis of AWS' DM. Figure 4 shows how the number of data products offered by AWS in each category evolved from Nov'20 to Aug'21. From this figure *we see a*

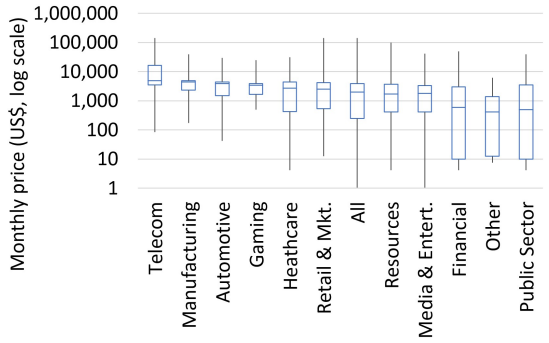


Figure 3: Subscription-based prices by industry in AWS

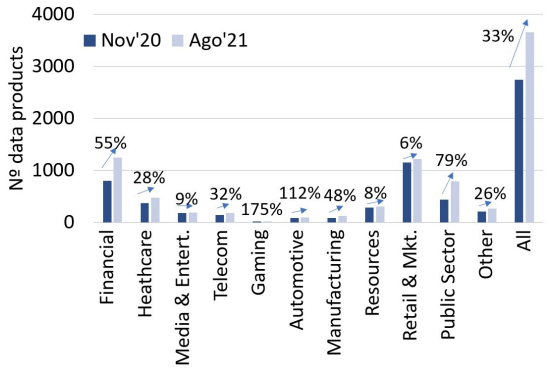


Figure 4: Evolution of AWS' products (Nov'20-Ago'21)

significant growth taking place in AWS DM in terms of data products, with a total of 3% monthly increase. Similarly, we also observe 1% monthly growth of the nº sellers. With regards to categories, '*Gaming*' (175%), '*Automotive*' (112%) '*Public sector*' (79%), '*Financial*' (55%), and '*Manufacturing*' (48%) exhibit the highest relative growth in nº products. Moreover, '*Financial*' and '*Public Sector*' accounted for the highest absolute growth in nº products, and added 444 and 349 respectively in this period.

5 COMPARING ACROSS MARKETPLACES

Comparing across different marketplaces is not a straightforward task since i) they provide metadata of different granularity and level of detail, and ii) they use different categorisation.

Regarding the information describing data products, most DMs offer a different number of fields and leave providers to fill them in. As a result, some products show more information and metadata than others, and often key information appears as text in product descriptions. This makes it difficult to compare and structure the metadata from different marketplaces. To overcome this challenge, an effort must be done to standardizing and structure this metadata, ensuring that it includes all the information required for buyers to understand the value of a product. This can be done by DMs or providers describing products according to an agreed notation [6], by creating a metadata onthology and setting up quality assurance mechanisms [24], or by structuring and homogenizing data collected from different marketplaces.

As regards data categorisation, every marketplace has its own way to classify data. For example, AWS tags data products with one or more out of ten different categories that correspond to industries, Advaneo allows only one category out of a different set of nine other industries, and DataRade labels data products with one or more tags out of a hierarchy with more than 300 categories and more than 150 use cases.

Furthermore, boundaries between tags are often blurry, and the criteria followed by different DMs and providers 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.

The problem has some similarity with classifying web pages on the Internet based on their text, and thus similar solutions might be sought [20]. Although natural language processing may help in labeling products of DM A with categories of DM B, we lack a solid ground truth given the differences in the criteria used to classify products. Moreover, in the case of use cases it is not only description what matters, and some of them require information to be refreshed at a minimum rate, availability of historical data or a minimum granularity for data to be useful.

6 WHICH ARE THE FEATURES DRIVING DATA PRODUCT PRICES?

To list key features determining the price of data products, we first manually inspect our database to i) identify any common distinctive features of top most valuable products, and ii) see how particular sellers price their data products.

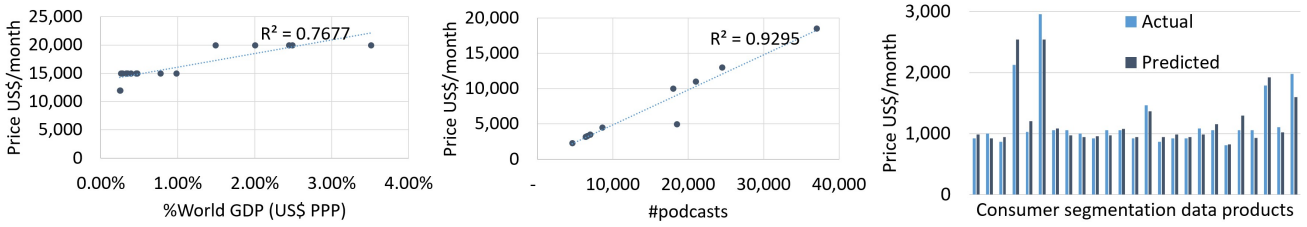
6.1 Key features of expensive data products

Section 3 pointed to a long tail of 33 data products worth more than US$30,000 per month. We found that:

- All of them include huge amounts of data from millions of people, tens of thousands of locations, etc.
- 20 (61%) of them offer daily updates.
- 11 (33%) of them do not provide any past data, and only 4 (12%) over 2 years historical data.
- 22 (67%) of them are US-focused, and 7 (21%) are global.
- 25 (73%) of them relate to *Retail, Location and Marketing*.
- B2B products include precise enterprise and contact data.
- At least 16 (48%) of them enable a granular location-based analysis, and 9 (27%) of them provide geo-located data.
- 7 B2C marketing products (21%) allow for session reconstruction (i.e., connecting data points of individuals/entities).

6.2 Seller-specific pricing strategies

We also looked at how *specific* sellers set the prices of data and we found that surprisingly simple regression models relying on specific metadata features were able to accurately predict their prices. Figure 5 depicts three such examples for telecom, recommender systems, and consumer segmentation data. We observe that:



(a) Mobile coverage **(b) Podcast metadata information** **(c) Consumer segmentation products**

Figure 5: Pricing regression examples from specific sellers

Table 2: List of feature groups

Question	Group	Definition	N° features	Example of features
What?	Category	Labels attached to the product that define the type of data it contains	custom	'Weather', 'Gaming', 'Financial'
	Description	Stem-like features obtained from data product descriptions	custom	'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 n° units covered broken down by the nature of such units	14	'people', 'entities'
	Update rate	Defines the frequency between data updates as announced by the seller	11	'real time', 'monthly', 'hourly'
	Granularity	Defines the resolution and potential aggregations of data	3	'aggLevel','spatialResol', 'timeResol'
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	'N° Countries', 'USA', 'Canada'

- A seller offering mobile network infrastructure and coverage data by country, by grouping products in a few price tiers that depend on their gross domestic product (see Fig. 5a).
- A seller offering metadata about podcasts uses language to segment its products, and prices them proportionally to the n° of podcasts they include (US$0.5 per podcast, see Fig. 5b).
- A well-known leader in consumer segmentation data relies on the population covered, its purchase power, and the granularity of the information provided to set the prices of its country-wide products (see Fig. 5c, R^2 = 0.88, mean absolute error = 10% of the average, mean relative error = 9%).

Understanding pricing strategies of sellers proved valuable to find features related to asking prices observed in commercial DMs.

6.3 Summary

Table 2 summarizes the features we found in different DMs that data prices seem to depend upon. We have grouped them by the aspect of data they describe. For example, product description, categories or use cases capture '*what*' kind of data we are selling, or features related to volume (e.g., how many individuals, enterprises, etc.), data update rate, spatio-temporal resolution or aggregation level determine '*how much*' data a product offers. Information about data products can be retrieved from the asking price published by commercial DMs and, to better capture the real market value, from actual data transactions registered by DMs. Still there are a number of challenges to build such a cross-DM database, including those discussed in Sect. 5, automating the extraction of features from the different DM or the population of unavailable fields, dealing with different languages, etc. Descriptive text-related features can be obtained from product descriptions using natural language processing methods, and categories and use cases be identified by means of classifiers based on the descriptions and the characteristics of data.

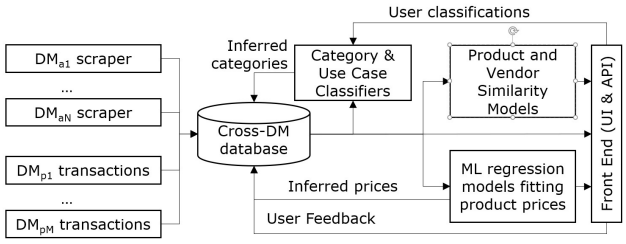


Figure 6: Block diagram of a data quotation tool

On top of this homogenized base of data products, ML regression models may be trained to understand what are the features driving the prices of data, or recommender systems to find similar products in the market and get an approximate quotation for a new product based on similar products in the market, as we do with second-hand houses or cars.

7 BUILDING A DATA QUOTATION TOOL

Data sellers find it challenging to decide a price for their data products [3]. Our success in training simple regression models to understand the prices of data set by specific sellers led to an interesting research question: can we train more complex regression models that are applicable to the whole base of data products? Such models would be able to provide sellers with a hint of the price of a dataset based on real market data about the prices of similar products. Figure 6 depicts the main building blocks of such a data quotation tool.

The blocks on the left are intended to ingest data to the cross-DM database. We envisage two different sources of data: 1) metadata and reviews of data products scraped from N public DMs, using crawlers

and parsers similar to the ones we described in Sect. 2, and 2) more detailed information from M partner DMs about data products, actual data transactions, number of views, reviews of products and vendors. These ingestion modules will provide standardized and structured data about data products to the cross-DM database, the centralized data warehouse of the tool. Table 2 summarized some fields describing data products that are affecting data prices, and hence must be extracted and loaded into the cross-DM to be considered in quoting data prices.

Category and use case ML classifiers will be used to enrich the information in the cross-DM database by attaching standardized data categories and use cases to data product. Such models must consider the challenges stated in Sect. 5, they will take descriptions and metadata of data products as inputs, and they will provide as their output whether data products belong to the specific category or the specific use case the classifier is trained to detect. As ground truth, classifiers will use cleansed information about 1) actual category tags of commercial DMs and 2) classifications informed by the users.

Similarly, ML regression models will be trained to fit prices found in commercial DMs taking as input all the metadata features presented in Tab. 2. Regression models will be able to provide an expected price range for a data product described by the user, or for data products in commercial DMs lacking a price reference. This information will also be used to enrich the cross-DM database.

Similarity models will be able to compute how similar or different two data products or vendors are. They will use similarity functions and the information stored in the cross-DM database. Semantic NLP models will be required to compare descriptions of data products.

The front-end will provide at least the following functions for users of the tool:

- browsing the cross-DM database and looking for data products or data providers;
- finding products or vendors similar to the one specified by the user as an input, or similar to other products or providers in the cross-DM database;
- obtaining price references for a data product based on i) the price of similar products in the market, and ii) on the prediction of the built-in regression models.

The main limitation of a quotation tool such as the one described here is that it would heavily rely on the metadata of commercial products to provide price references for data products. Therefore, it would not take into consideration other relevant factors for pricing a dataset such as i) its adequacy and its usability for the specific task of the potential buyer, ii) the quality of the data provided, iii) the specific value for the buyer, which may substantially differ with the purpose of acquiring a data product.

Finally, the tool may also consider other factors in order to provide more accurate quotations for data products. For example, the competition observed in the segment of the market the data product is aimed to, or the price trend of data for the specific segment over time may also be affecting the price of a data product in the market. The first should be somehow reflected by updated market prices, and the second requires harvesting and analyzing market data across time.

8 RELATED WORKS

Even though several surveys related to data marketplaces have been recently published [3, 38, 40–42], our work is, to the best of our knowledge, the first empirical measurement study that deals with the price data in commercial DMs.

In fact, the lack of empirical data around dataset prices is considered as a key challenge in data pricing research [36]. According to some authors, some techniques to set the prices of digital products [39] or cloud services [44] are applicable to data products, as well. Some authors proposed auction designs to set the prices of digital goods and data products [17, 18]. Novel AI/ML data marketplace architectures have been proposed under the concept of value-based pricing [2, 11, 33] and the value of privacy [32]. Moreover, some authors defined pricing strategies and marketplaces based on differential privacy [16, 27] or for queries to a database [10, 26]. All of them work on analysing 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 [21, 45] is the one closest to our approach, and assesses the value of data by evaluating and assigning weights to quality features like those we listed in Tab. 2.

The pricing of personal data has received attention from the privacy and measurement communities. There are measurement studies based on prices carried over the Real Time Bidding protocol [30, 35] as well as more traditional survey-based studies [7]. 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.

9 CONCLUSIONS AND FUTURE WORK

Our work has provided a first glimpse into the growing market for B2B data. Having scraped metadata for hundreds of thousands of data products listed by 10 real-world data marketplaces and other 30 data providers, 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 features that appear to have impact on prices, and pointed to specific challenges and solutions to compare across marketplaces with different categorization.

Due to the current fragmentation of data markets, there is a need of an overarching solution not only to discover data, but to provide transparency on how much a piece of data might be worth in the market, and why. In that direction, the paper provides a high-level architecture of price recommendation tool for new data products. We are continuously monitoring commercial DMs to see how they evolve and to enrich our database and to find out more about the price and the value of data. We are also working on the different blocks described in Sect. 7 to build a quotation tool for data products using real market data.

10 ACKNOWLEDGEMENTS

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

REFERENCES

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

[2] A. Agarwal, M. Dahleh, and T. Sarkar. A marketplace for data: An algorithmic solution. In *Proc. of ACM EC*, 2019.

[3] S. Andrés and N. Laoutaris. A survey of data marketplaces and their business models. *ACM SIGMOD Record*, 51(3), 2022.

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

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

[6] 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.

[7] 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.

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

[9] 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.

[10] S. Chawla, S. Deep, P. Koutris, and Y. Teng. Revenue maximization for query pricing. *Proc. of the VLDB Endow.*, 13, 2019.

[11] L. Chen, P. Koutris, and A. Kumar. Towards model-based pricing for machine learning in a data marketplace. In *Proc. of ACM SIGMOD*, 2019.

[12] Clive Humby. Data is the new oil! *Keynote at ANA Senior Marketer's Summit, Kellogg School*, 2006.

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

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

[15] DIH. Data intelligence hub. extract value from data securely.

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

[17] 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.

[18] 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.

[19] Harvard. Dataverse. https://dataverse.harvard.edu/. Last accessed: Jun '22.

[20] M. Hashemi. Web page classification: A survey of perspectives, gaps, and future directions. 79(17–18), 2020.

[21] J. R. Heckman, E. Boehmer, E. H. Peters, M. Davaloo, and N. G. Kurup. A pricing model for data markets. In *iConference 2015 Proceedings*, 2015.

[22] N. Henke, J. Bughin, and al. The age of analytics: Competing in a data-driven world. *McKinsey Global Institute*, 2016.

[23] M. Hils, D. W. Woods, and R. Böhme. Measuring the emergence of consent management on the web. In *Proceedings of the ACM Internet Measurement Conference*, IMC '20, page 317–332, New York, NY, USA, 2020. ACM.

[24] IDSA. International data spaces reference architecture model. *International Data Spaces Association*, 2022.

[25] Kaggle. Datasets. https://www.kaggle.com/datasets. Last accessed: Jun '22.

[26] 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.

[27] C. Li, D. Y. Li, G. Miklau, and D. Suciu. A theory of pricing private data. 39(4), 2015.

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

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

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

[31] 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.

[32] 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.

[33] O. Ohrimenko, S. Tople, and S. Tschiatschek. Collaborative machine learning markets with data-replication-robust payments. *CoRR*, 2019.

[34] Otonomo. One-stop shop for vehicle data. https://otonomo.io/. Last accessed: Feb 2022.

[35] 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.

[36] J. Pei. Data pricing – from economics to data science. In *Proc. of the ACM SIGKDD*, page 3553–3554, 2020.

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

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

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

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

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

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

[43] Veracity. Veracity by dnv gl. find the right tools for your industry needs. https://store.veracity.com/. Last accessed: Jun '22.

[44] C. Wu, R. Buyya, and K. Ramamohanarao. Cloud pricing models: Taxonomy, survey, and interdisciplinary challenges. *ACM Comput. Surv.*, 2019.

[45] H. Yu and M. Zhang. Data pricing strategy based on data quality. *Computers and Industrial Engineering*, 112:1–10, 2017.