ICDE 2023
Anaheim, CA April 3 – 7, 2023

Data Transparency Group

institute IMDEA networks

Understanding the Price of Data in Commercial Data Marketplaces

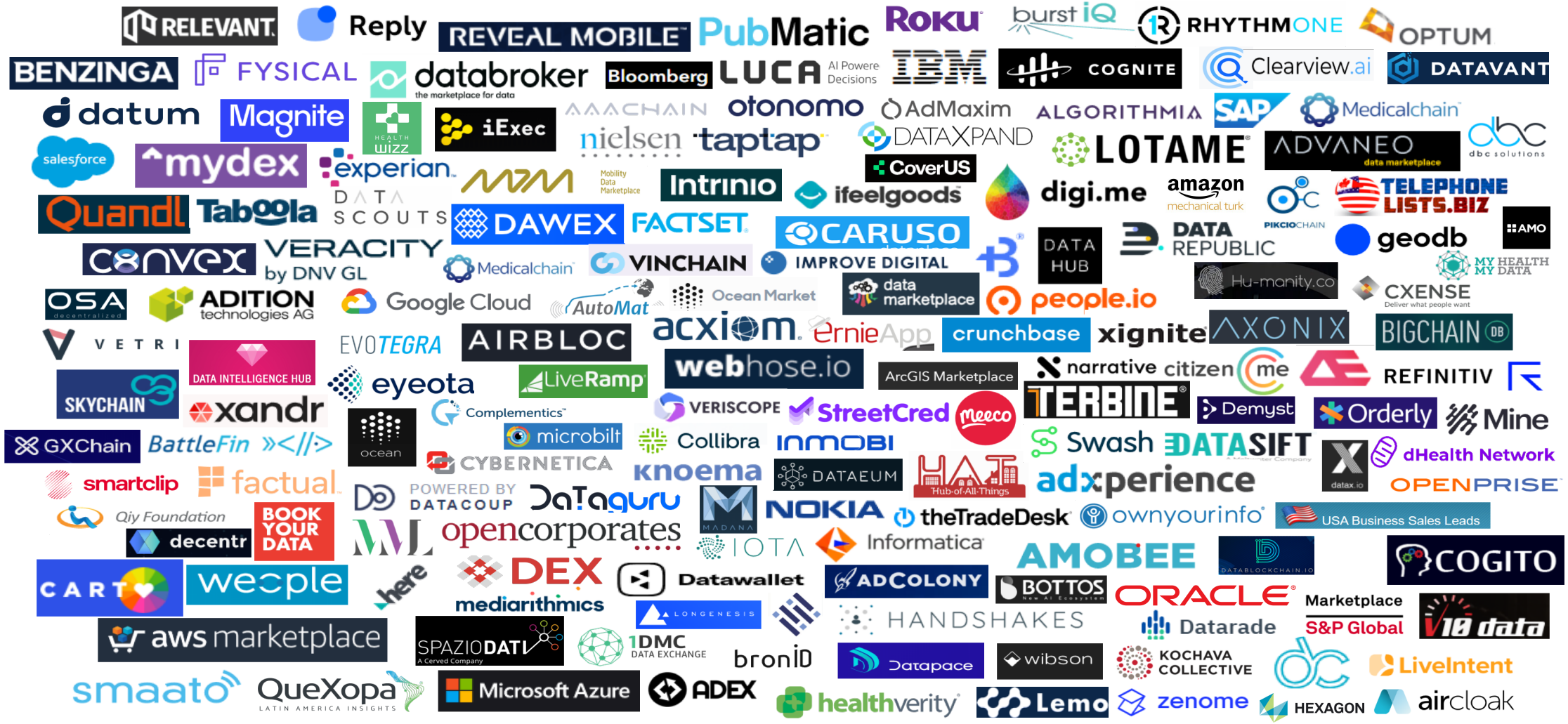
Santiago Andrés Azcoitia, IMDEA Networks Institute

Costas Iordanou, Cyprus University of Technology

Nikolaos Laoutaris, IMDEA Networks Institute

Developing the Science of Networks

We checked more than 190 companies offering data products and services in order to understand how data is traded nowadays[1]



[1] S. Andrés Azcoitia, N. Laoutaris. "*A Survey of Data Marketplaces and Their Business Models*", SIGMOD Record Sept'22

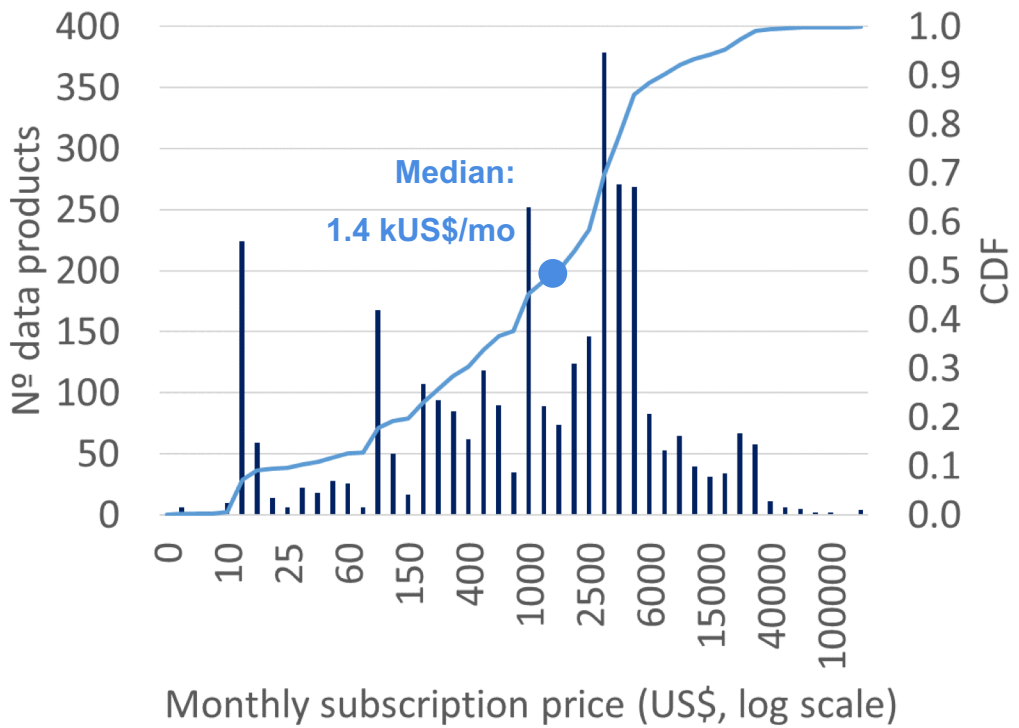
We scraped 10 data marketplaces (DMs) + 30 sellers and collected information about 215,075 data products from 2,115 sellers in total

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

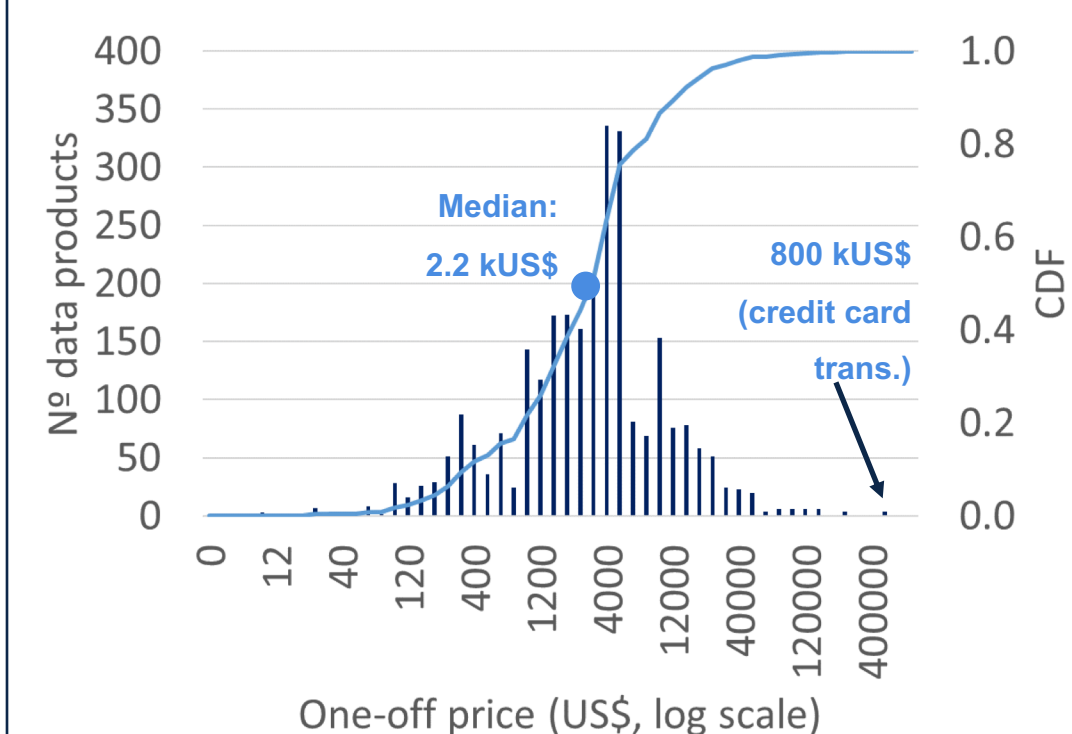
Only 12k data products are "paid" and only 4k from 443 distinct sellers disclose information about their prices!

We found that data sells at an immensely wide range of prices, …

Subscription-based data product prices

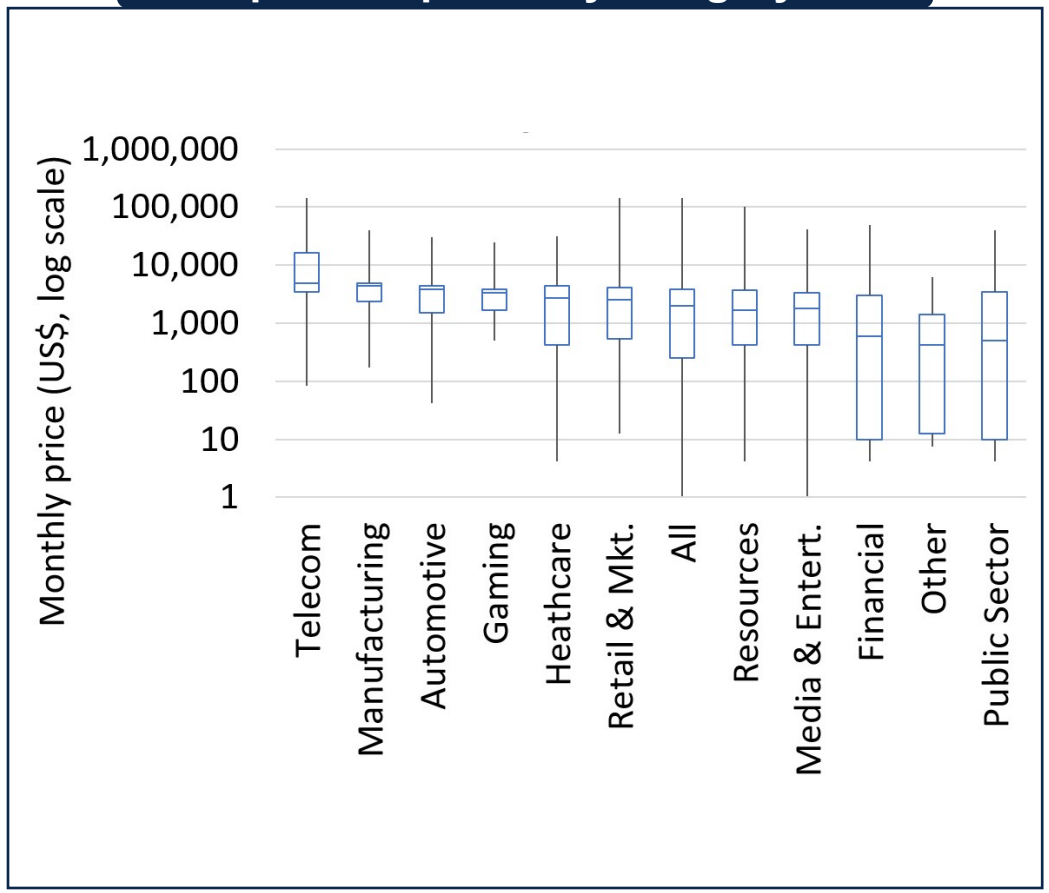


One-off data product prices

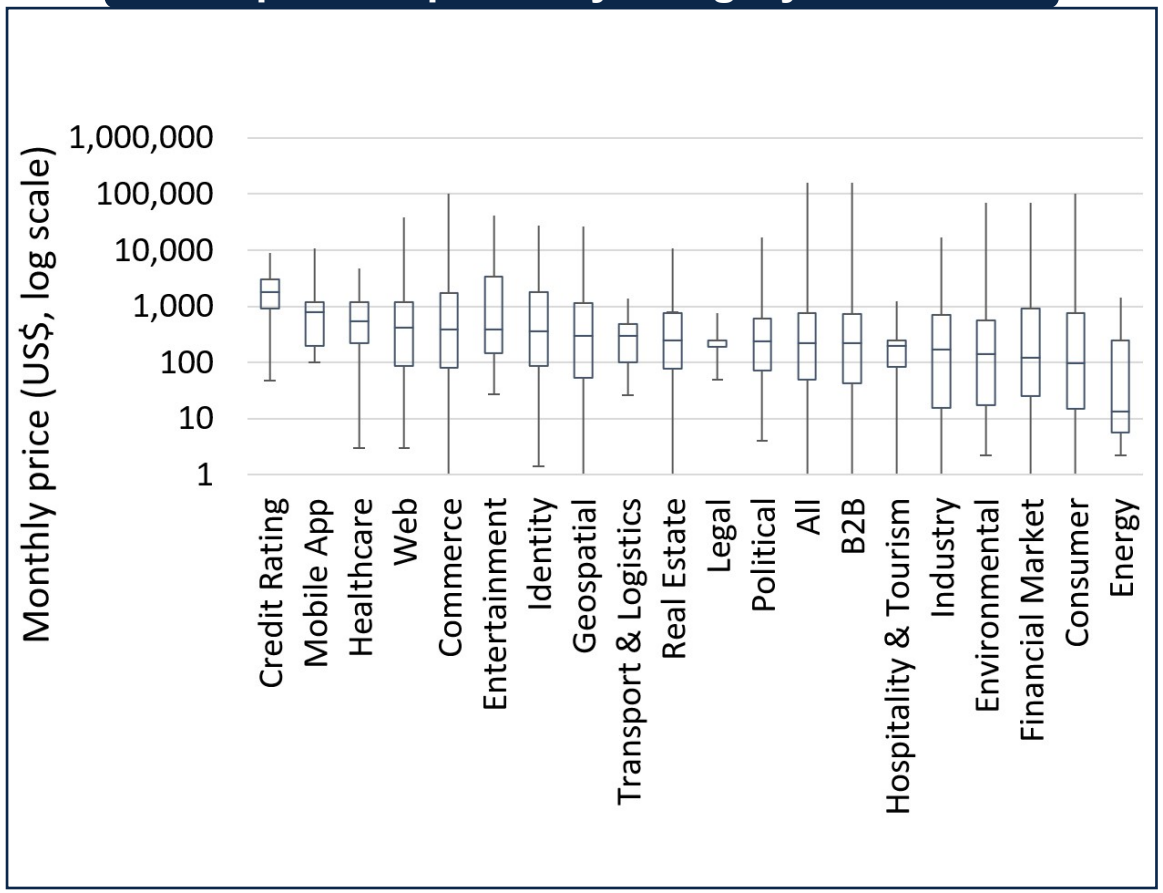


… which depend on the category of data product

Data product prices by category AWS

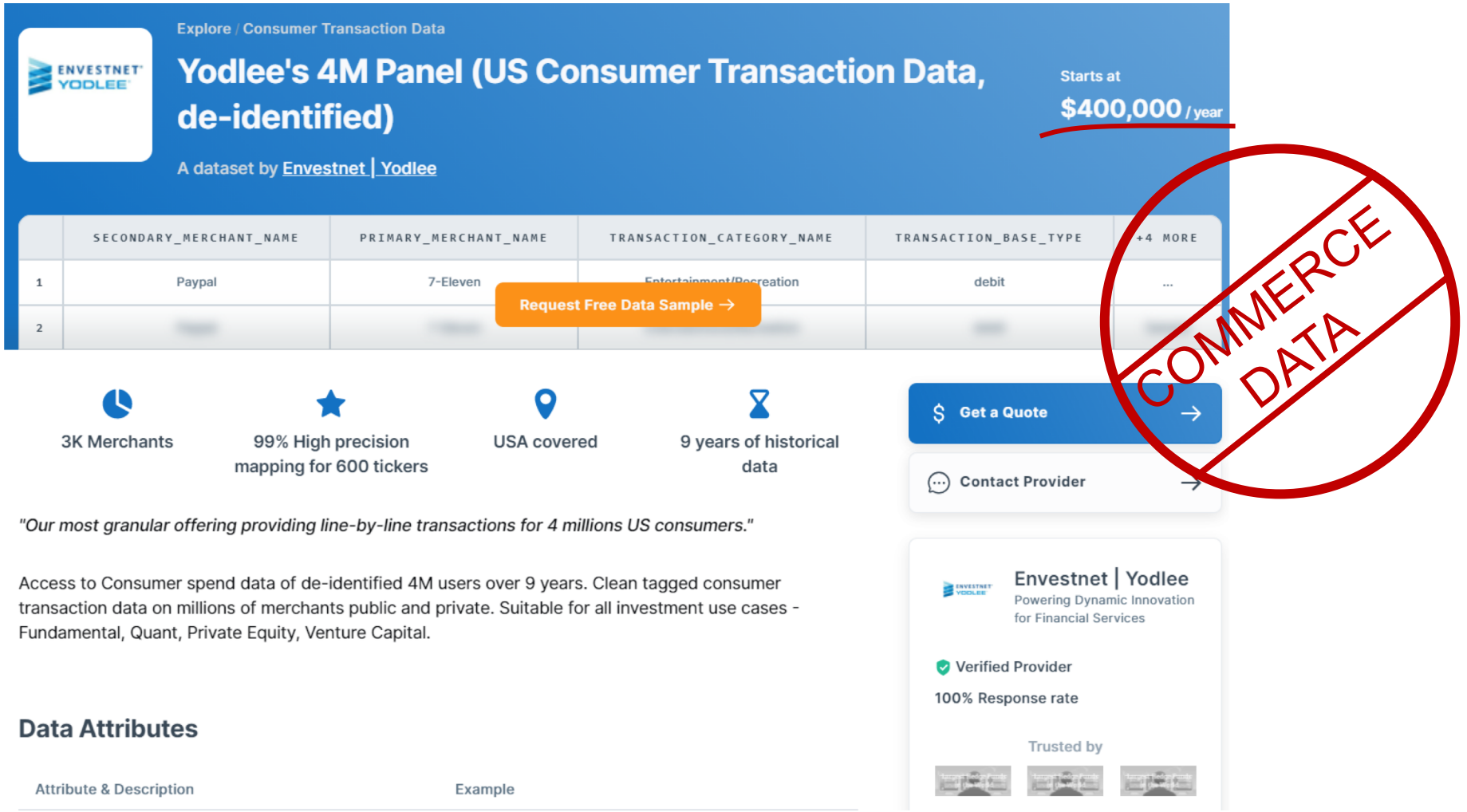


Data product prices by category DataRade

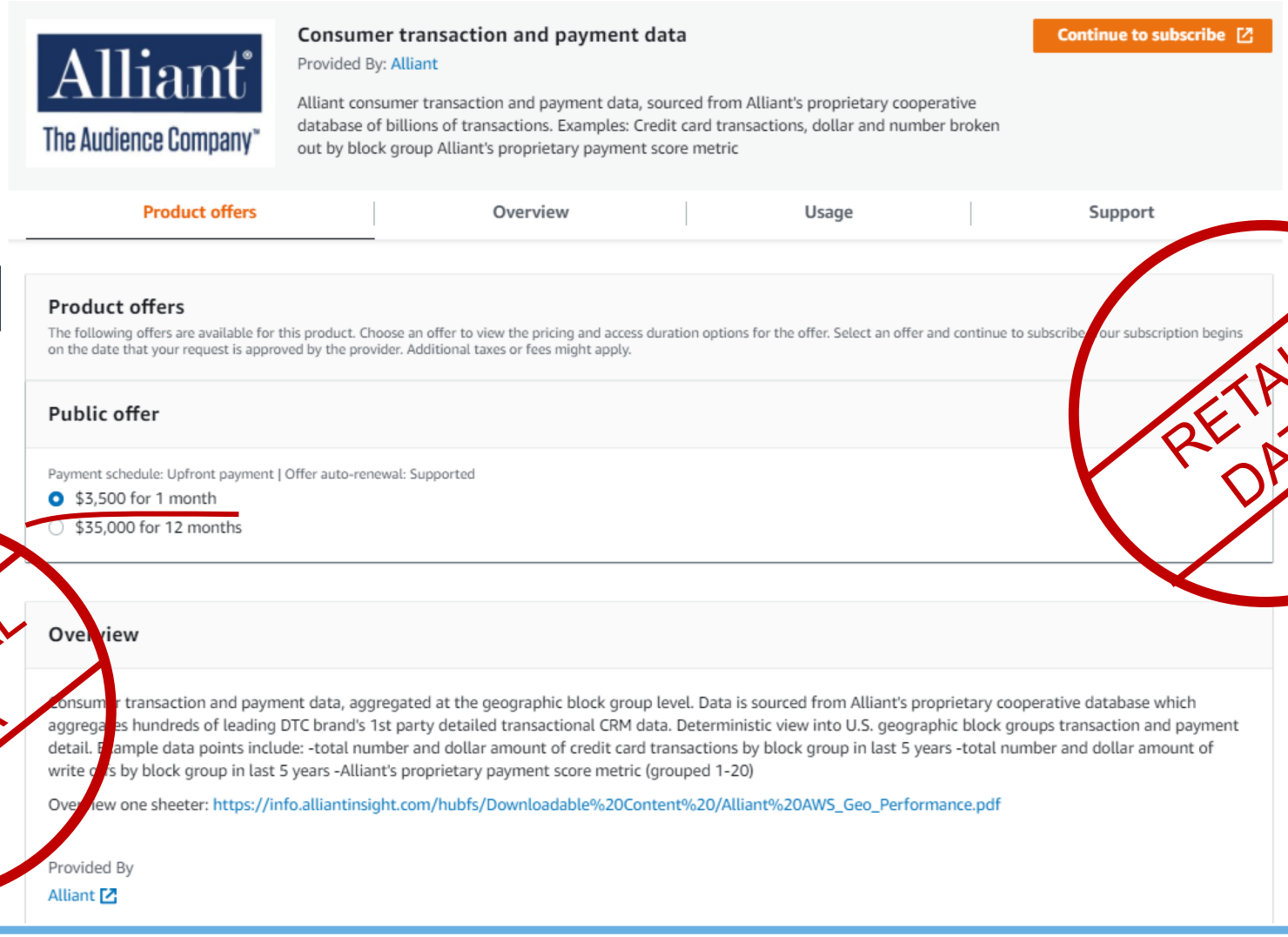


Cross DMs analysis is challenging, since DMs i) provide different detail, and ii) use different categorisation and criteria to assign categories to data products

Datarade



Cross DMs analysis is challenging, since DMs i) provide different detail, and ii) use different categorisation and criteria to assign categories to data products



We trained NLP NB classifiers to learn how a *source* DM labels products that belong in a certain category, and label products in a *destination* DM

Significant stems

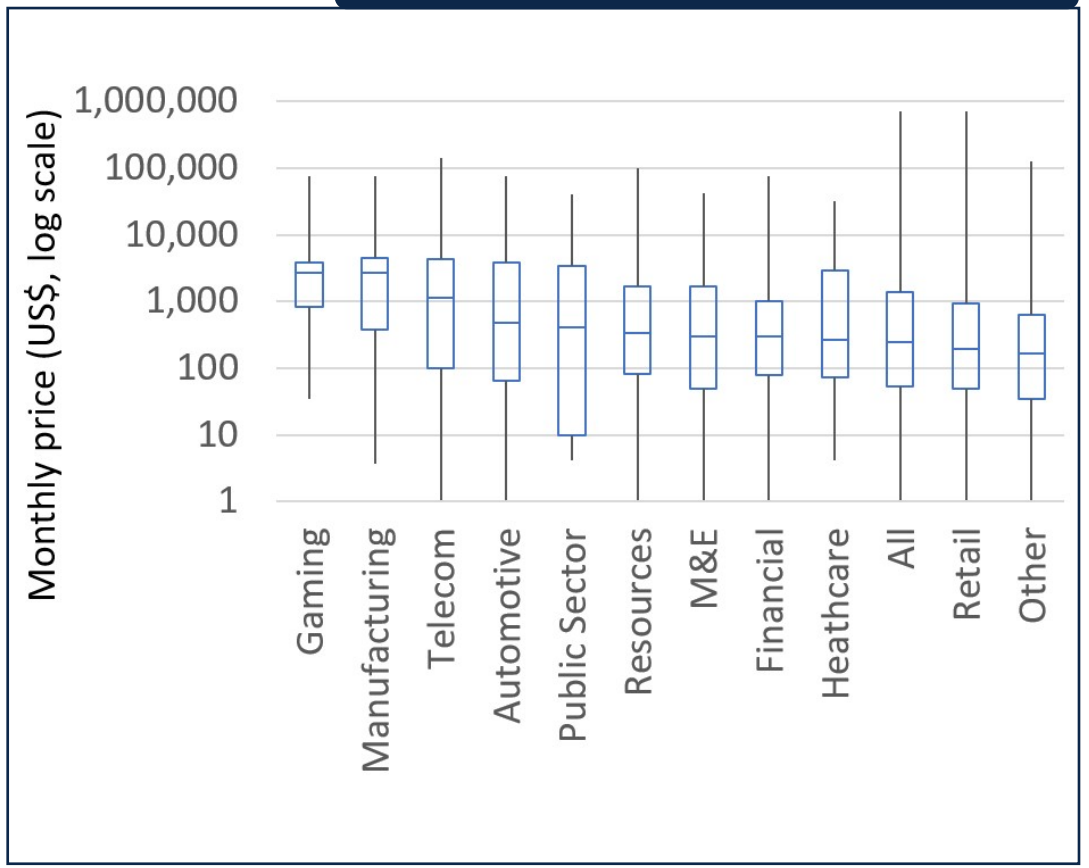
Financial: ‘system’, ‘sec’, ‘exchang’, ‘type’, ‘file’, ‘form’, ‘edgar’, ‘secur’, ‘act’, and ‘compani’.

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

Accuracy score

	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

Boxplot by AWS category (all)



We built a cross-DM database as a superset of metadata fields found in different DMs, and found to be driving the prices of data products

- Id & Description
- Category
- Granularity
- Time scope
- Use cases
- Identifiability
- Volume & units
- Delivery method
- Limitations
- Geo scope
- Update frequency
- Add-ons

So, which are the features actually driving the prices of data products?

We tested 9 regressors and optimized 4 of them. At least one shows $R^2 > 0.78$ for predicting prices of financial, marketing and health-related data

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

Note: MAE and MSE reflect the error in predicting the logarithm of data product prices

We discarded linear, Elastic-Net, Ridge, Bayesian Ridge, and Lasso regressions even though they worked well in specific cases

We studied the most relevant individual features which sellers rely on for pricing financial, marketing and healthcare data

Financial			Marketing			Healthcare		
RF	**kNeigh**	**GB**	**RF**	**kNeigh**	**GB**	**RF**	**kNeigh**	**GB**
units	units	units	units	units	csv	units	csv	wordlist
entities	Email	S3Bucket	entities	History	units	people	units	Del. Methods
S3Bucket	Download	wordmonthli	ldSessions	USA	yearly	wordhealth	daily	wordhospit
wordsubmit	daily	wordstock	Download	ldSessions	people	wordtrend	wordmarket	wordidentifi
Download	ldCompanies	worddeliv	REST API	Nº Countries	REST API	wordmedic	wordgo	wordamerica
people	USA	people	wordcustom	Financial	wordqualiti	wordglobal	Limitations	wordhealth
txt	wordmarket	Del. Methods	USA	Others	wordaccur	csv	location data	wordreport
wordedgar	Retail	txt	yearly	people	wordidentifi	DelMethod	wordpopul	wordstudi
wordcustom	wordcontact	wordneed	monthly	wordcontact	wordwebsit	wordinsight	wordprofil	wordupdat
wordlist	realtime	wordsubmit	ldCompanies	Email	UlExport	wordreport	wordinsight	wordcontact

The table shows average scores of 5-fold executions of leave-one-out and permutation importance analysis. An average of 11 of the top 20 features by category and algorithm appear in every individual test.

Features related to data volume are present in financial and marketing data categories, but seem to be especially relevant for financial data products

Financial			Marketing			Healthcare		
RF	**kNeigh**	**GB**	**RF**	**kNeigh**	**GB**	**RF**	**kNeigh**	**GB**
units	units	units	units	units	csv	units	csv	wordlist
entities	Email	S3Bucket	entities	History	units	people	units	Del. Methods
S3Bucket	Download	wordmonthli	ldSessions	USA	yearly	wordhealth	daily	wordhospit
wordsubmit	daily	wordstock	Download	ldSessions	people	wordtrend	wordmarket	wordidentifi
Download	ldCompanies	worddeliv	REST API	Nº Countries	REST API	wordmedic	wordgo	wordamerica
people	USA	people	wordcustom	Financial	wordqualiti	wordglobal	Limitations	wordhealth
txt	wordmarket	Del. Methods	USA	Others	wordaccur	csv	location data	wordreport
wordedgar	Retail	txt	yearly	people	wordidentifi	DelMethod	wordpopul	wordstudi
wordcustom	wordcontact	wordneed	monthly	wordcontact	wordwebsit	wordinsight	wordprofil	wordupdat
wordlist	realtime	wordsubmit	ldCompanies	Email	UIExport	wordreport	wordinsight	wordcontact

Due to the heterogeneity of the sample, there is no single feature other than units that relates to the price of data in every category. **The 'what' seems to be more important than the 'how much' when pricing healthcare products**

Among the rest of the features, the ones related to 'what' data is offered stand out in terms of importance

Financial			Marketing			Healthcare		
RF	**kNeigh**	**GB**	**RF**	**kNeigh**	**GB**	**RF**	**kNeigh**	**GB**
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

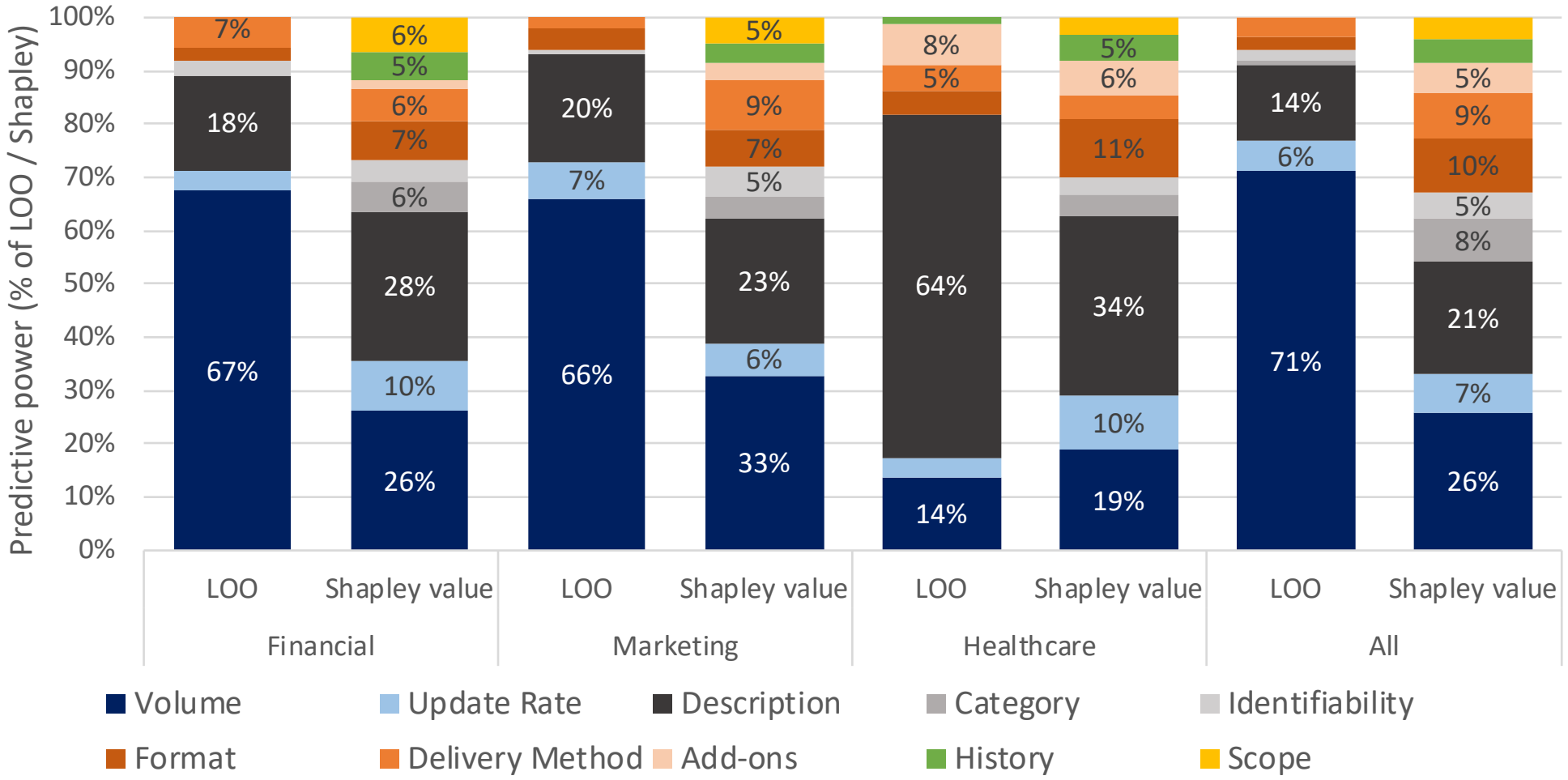
Features relating to delivery methods and update rate seem somewhat important for the prices of financial and marketing data

Financial			Marketing			Healthcare		
RF	**kNeigh**	**GB**	**RF**	**kNeigh**	**GB**	**RF**	**kNeigh**	**GB**
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

Geo-spatial localization and scope and the possibility of connecting data points from the same owner are relevant especially for marketing data.

Financial			Marketing			Healthcare		
RF	**kNeigh**	**GB**	**RF**	**kNeigh**	**GB**	**RF**	**kNeigh**	**GB**
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 studied the most influential feature groups, as well, resulting in notorious differences across data categories



To probe further

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. With the exception of very large companies that have both the data and the skills to develop powerful ML-driven services, the large majority of provably possible ML services, from e-health, to transportation and predictive maintenance, to name just a few, still remain at the idea or prototype level for the simple reason that data, the skills to manipulate them, and the business models to bring them to market, seldom co-exist under the same roof. Data has to somehow meet with the ML and business skills that can unleash its full power for the society and economy.

DATA ECONOMY WORKSHOP

KEYNOTE SPEAKERS

Bruce Pon

Founder of Ocean Protocol

Damian Boeselager

Member of the European Parliament

Carme Artigas

Secretary of State for Digitization and Artificial Intelligence, Government of Spain

DataBri-X

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101070069.

MLEDGE

Cloud and Edge Machine Learning

IMDEA Networks is the beneficiary of this project

More info

This project (REGAGE22e00052829516) has been funded by the Ministry of Economic Affairs and Digital Transformation and the European Union-NextGenerationEU/PRTR.

GOBIERNO DE ESPAÑA
VICEPRESIDENCIA PRIMERA DEL GOBIERNO
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
UNICO I+D Cloud
Financiado por la union europea NextGenerationUE

In summary, this is the first paper measuring and understanding the price of data in commercial marketplaces, we found that:

1. Data products sell at an immensely wide range of prices up to several US$100ks per month

2. We homogenized heterogeneous metadata and classification labels to be able to compare data products across marketplaces

3. Using regression models, we managed to fit the prices of commercial products from their features with R^2 above 0.84.

4. Features related to ‘what’ and ‘how much’ data a product contains are driving 66% of its price, and some other features (geo-scope, history, upate rate) are relevant for specific categories.

5. We’ve made available code and data obtained in this study which you can find in https://gitlab.com/sandresazcoitia1/data-pricing-tool

Thank you!

Q&A time!

For more information please contact:

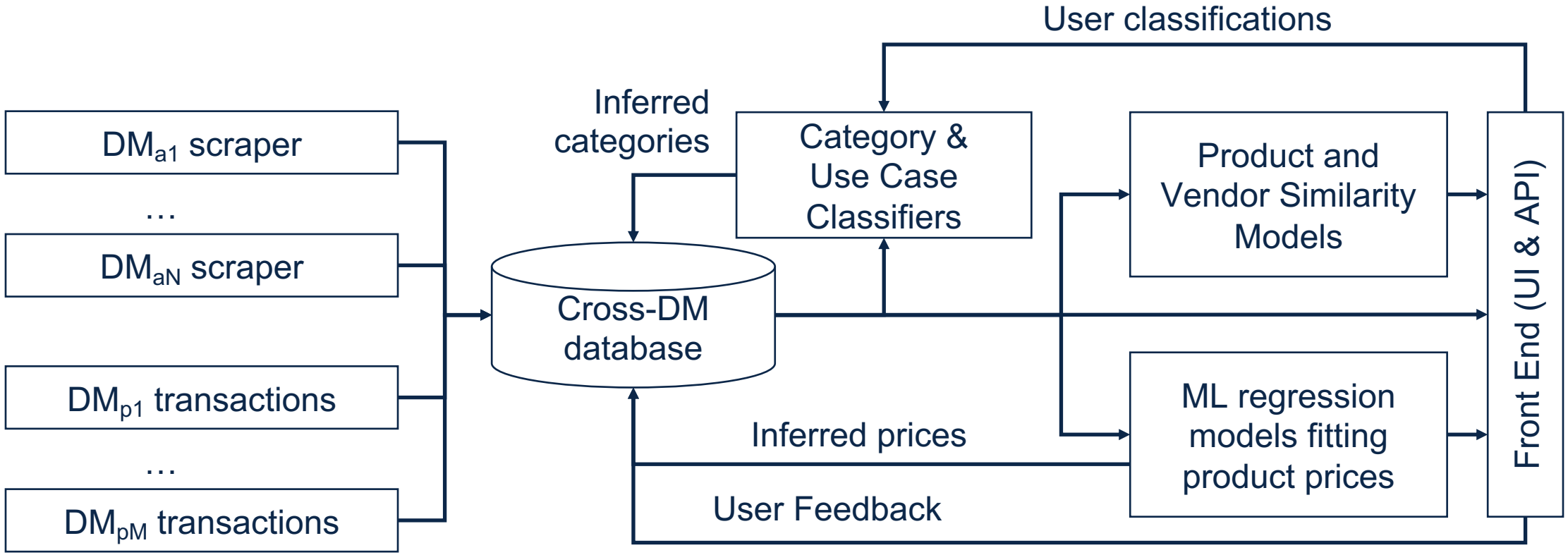
institute imdea networks

Data Transparency Group

Santiago Andrés Azcoitia
santiago.azcoitia@imdea.org

Nikolaos Laoutaris
nikolaos.laoutaris@imdea.org

We are working on a data quotation tool[2] to be able to predict the prices of a data product out of its metadata based on market prices and transactions



Such a tool will have limitations, since it does not consider: i) the usability for the buyer, ii) the quality of the data, iii) the specific value for a buyer.

[2] S. Andrés Azcoitia, C. Iordanou, and N. Laoutaris. Measuring the Price of Data in Commercial Data Marketplaces. 1st ACM Data Economy Workshop at CoNEXT'22

So, what is the price of data in the B2B market? What are the features that are driving the prices of data products?