Understanding the Price of Data in Commercial Data Marketplaces

*Santiago Andrés Azcoitia**

PhD Student @ IMDEA Networks Institute & Universidad Carlos III de Madrid

Costas Iordanou, Cyprus University of Technology

Nikolaos Laoutaris, Imdea Networks Institute

ICDE 2023

Anaheim, CA April 3 – 7, 2023

Data Transparency Group

institute imdea networks

uc3m

Τεχνολογικό Πανεπιστήμιο Κύπρου

[Developing the Science of Networks]

A quiz

How valuable is this?

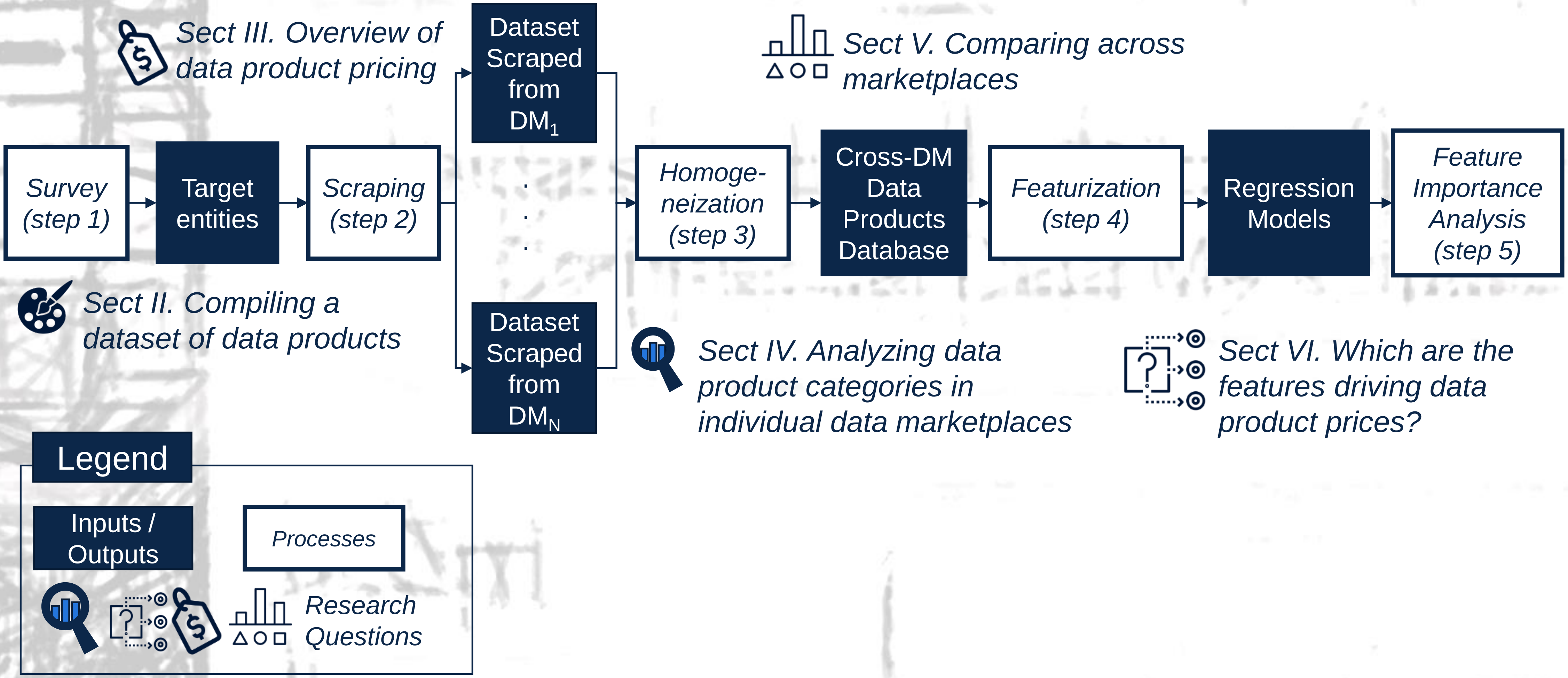
How much is it?

A quiz

And, what’s the price of this?

38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States
53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States
28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba
37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States
49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica
52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States
31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States
42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States
37	Private	280464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States
30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	0	40	India
23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States
32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States
40	Private	121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	0	0	40	?
34	Private	245487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	0	0	45	Mexico
25	Self-emp-not-inc	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States
32	Private	186824	HS-grad	9	Never-married	Machine-op-inspct	Unmarried	White	Male	0	0	40	United-States
38	Private	28887	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States
43	Self-emp-not-inc	292175	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	0	45	United-States
40	Private	193524	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	60	United-States
54	Private	302146	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	0	20	United-States
35	Federal-gov	76845	9th	5	Married-civ-spouse	Farming-fishing	Husband	Black	Male	0	0	40	United-States
43	Private	117037	11th	7	Married-civ-spouse	Transport-moving	Husband	White	Male	0	2042	40	United-States
59	Private	109015	HS-grad	9	Divorced	Tech-support	Unmarried	White	Female	0	0	40	United-States

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



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

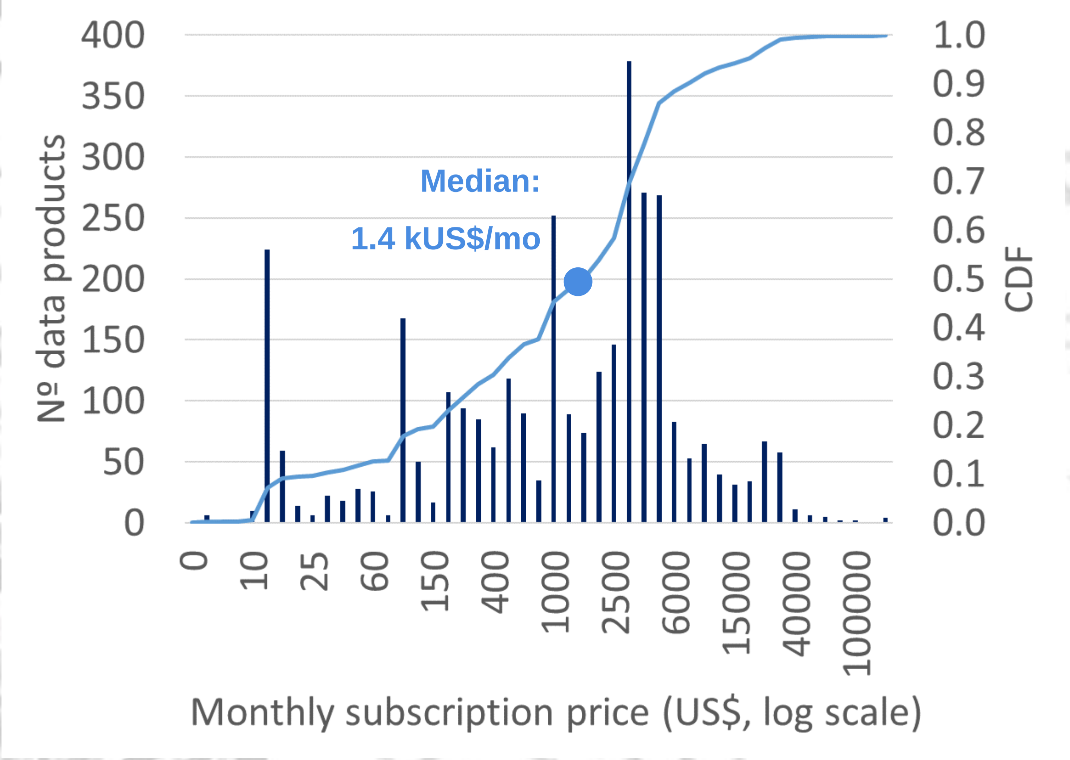
[1] S. Andrés Azcoitia, N. Laoutaris. "*A Survey of Data Marketplaces and Their Business Models*", SIGMOD Record Sept'22

We scraped 10 data marketplaces (DMs) + 30 providers and collected information about 215,075 data products from 2,115 sellers in 2021

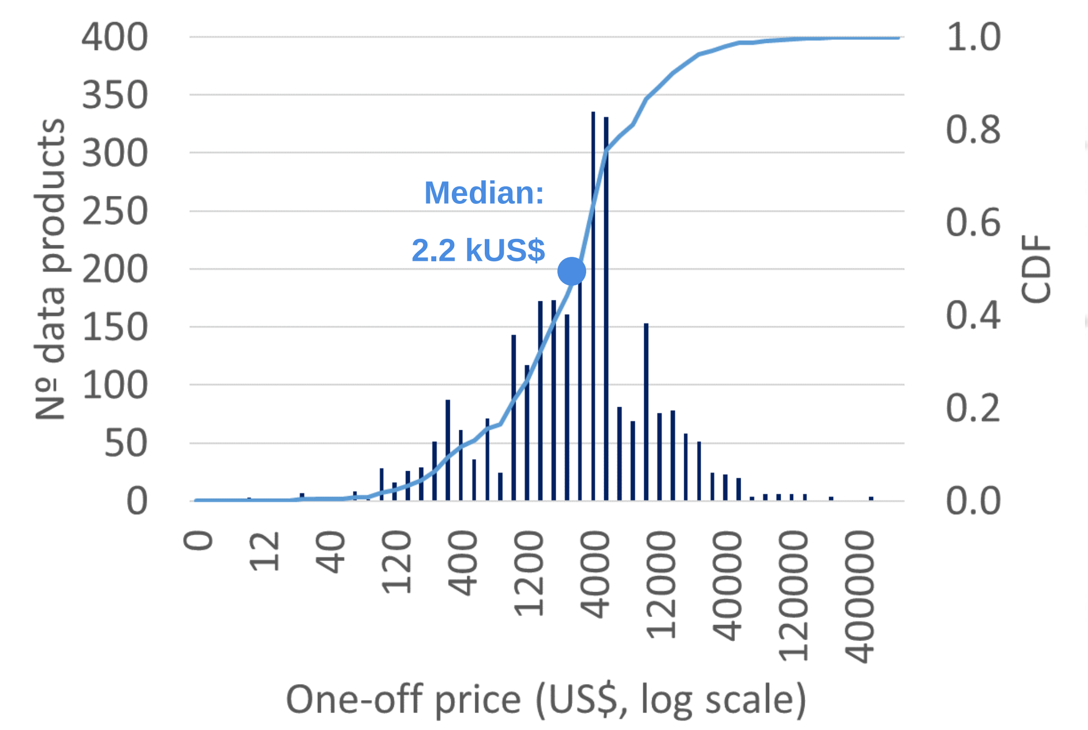
Marketplace	#Products	#Paid prod.	#Sellers
Advaneo	198,743	1	N/A
AWS	4,263	2,674	262
DataRade	1,592	1,592	1,262
Snowflake	889	889	200
Knoema	158	158	142
DAWEX	160	160	79
Carto	8,182	5,283	42
Crunchbase	9	9	15
Veracity	115	95	38
Refinitiv	187	187	76
Other providers	777	775	30

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

Subscription-based data product prices

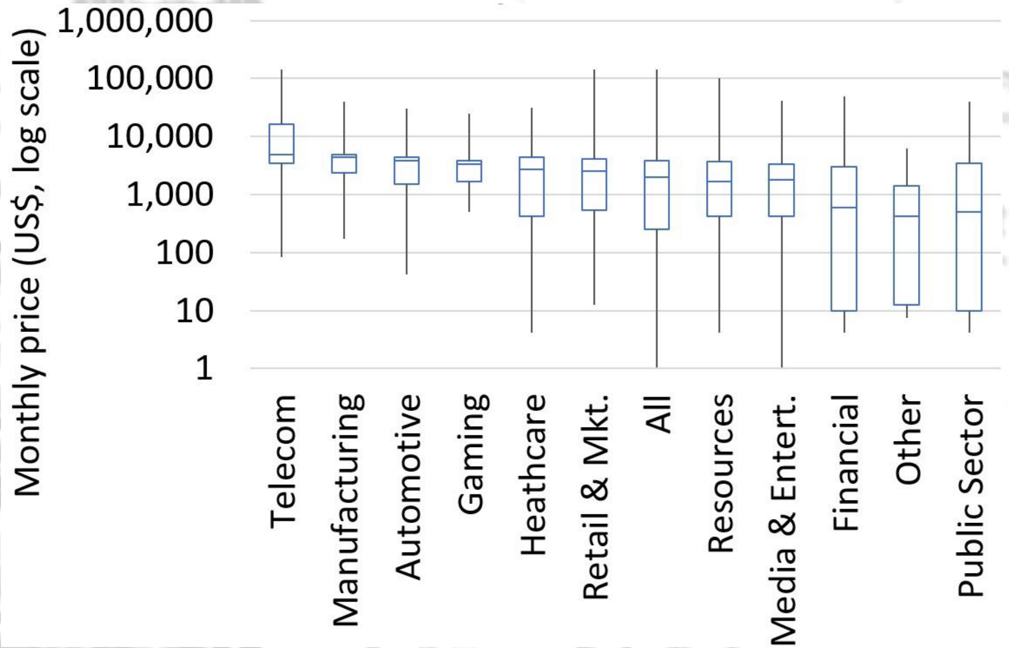


One-off data product prices

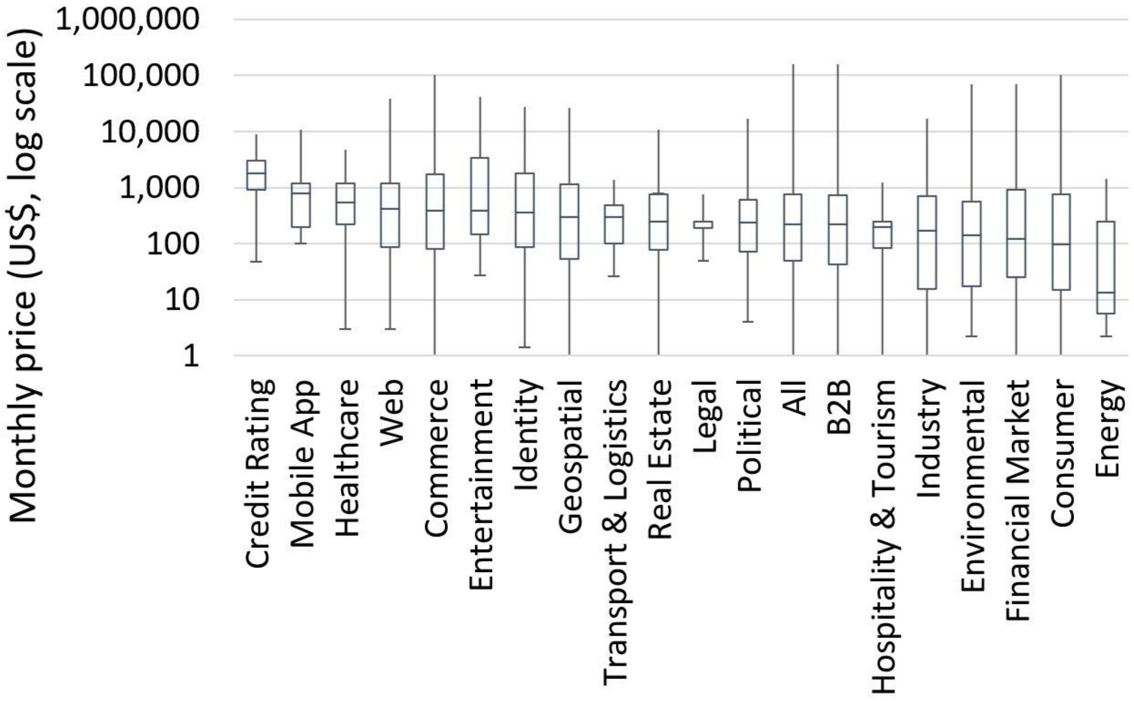


… which depend on the category of data product

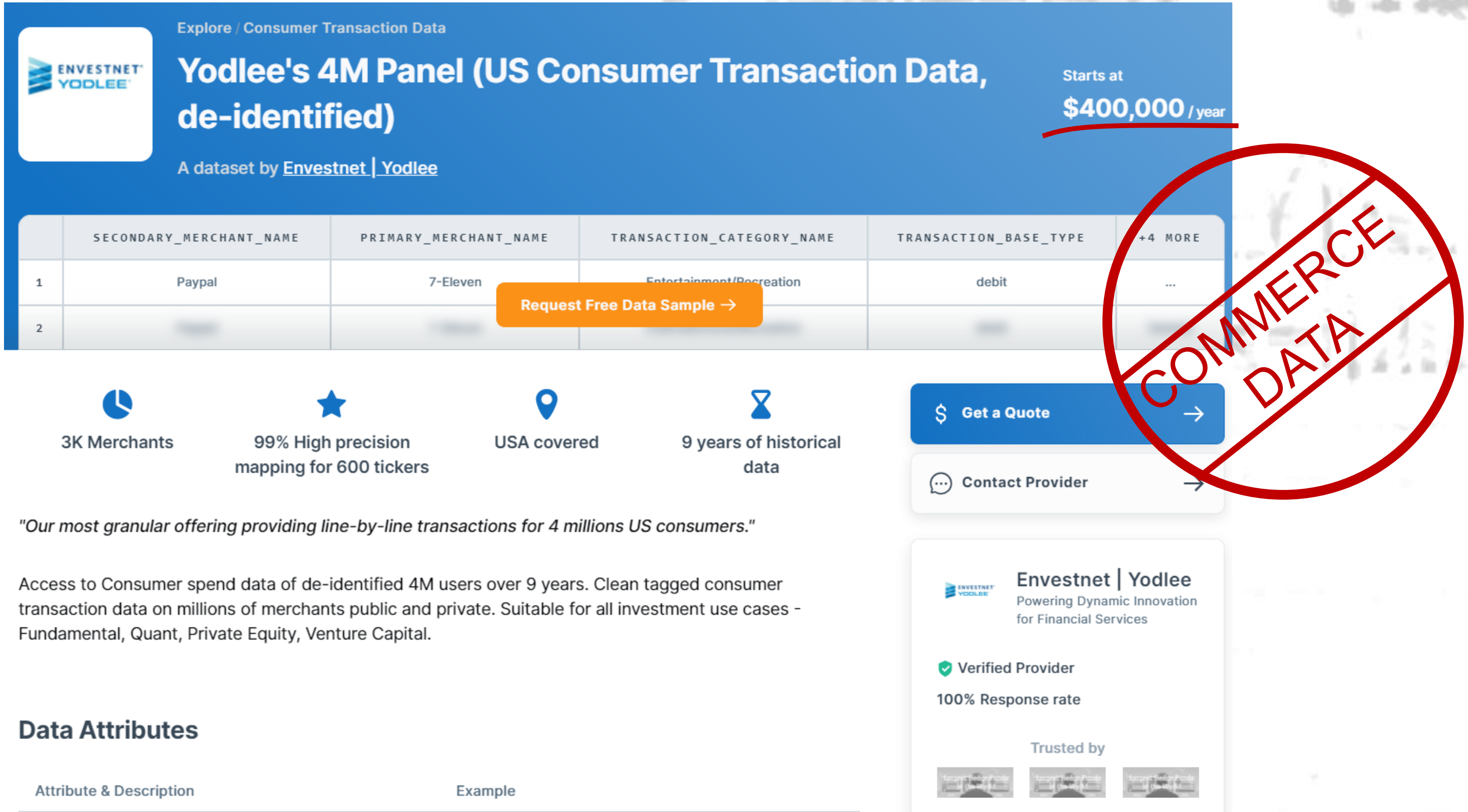
Data product prices by category AWS



Data product prices by 1st level category DataRade



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



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

Alliant
The Audience Company™

Consumer transaction and payment data

Provided By: Alliant

Alliant consumer transaction and payment data, sourced from Alliant's proprietary cooperative database of billions of transactions. Examples: Credit card transactions, dollar and number broken out by block group Alliant's proprietary payment score metric

Continue to subscribe

Product offers | Overview | Usage | Support

Product offers

The following offers are available for this product. Choose an offer to view the pricing and access duration options for the offer. Select an offer and continue to subscribe. Your subscription begins on the date that your request is approved by the provider. Additional taxes or fees might apply.

Public offer

Payment schedule: Upfront payment | Offer auto-renewal: Supported

- $3,500 for 1 month
- $35,000 for 12 months

RETAIL DATA

Overview

Consumer transaction and payment data, aggregated at the geographic block group level. Data is sourced from Alliant's proprietary cooperative database which aggregates hundreds of leading DTC brand's 1st party detailed transactional CRM data. Deterministic view into U.S. geographic block groups transaction and payment detail. Example data points include: -total number and dollar amount of credit card transactions by block group in last 5 years -total number and dollar amount of write offs by block group in last 5 years -Alliant's proprietary payment score metric (grouped 1-20)

Overview one sheeter: https://info.alliantinsight.com/hubfs/Downloadable%20Content%20/Alliant%20AWS_Geo_Performance.pdf

Provided By

Alliant

FINANCIAL DATA

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

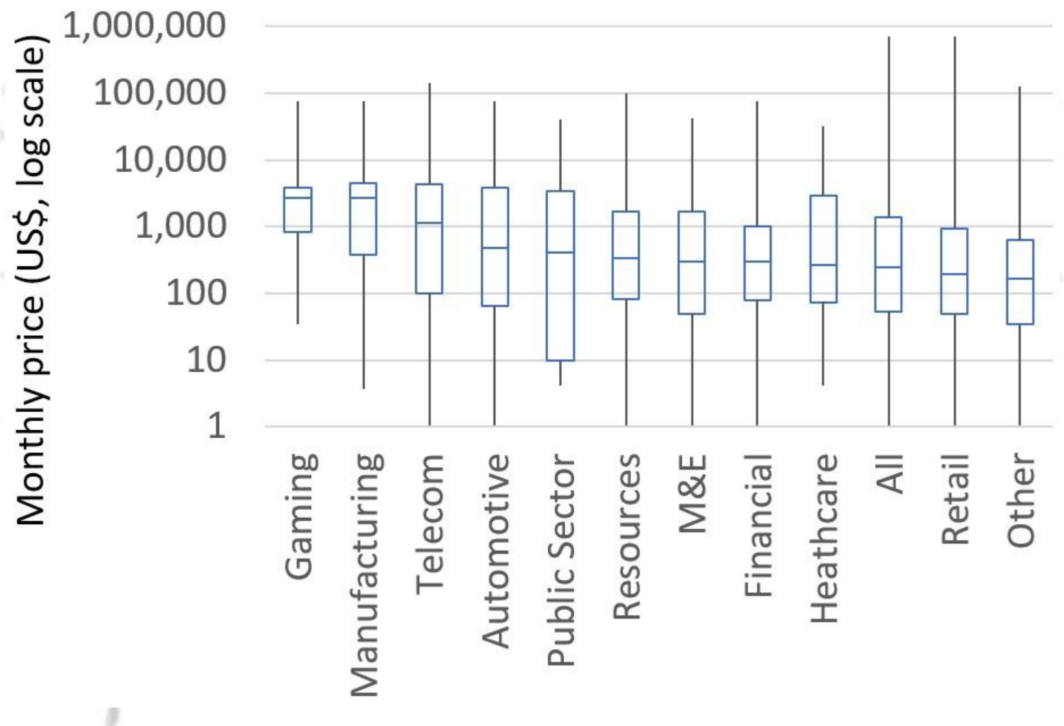
Financial: ‘system’, ‘sec’, ‘exchang’, ‘type’, ‘file’, ‘form’, ‘edgar’, ‘secur’, ‘act’, and ‘compani’.

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

Data Product Price by AWS category (all products)



ICDE 2023
Anaheim, CA April 3 – 7, 2023
institute imdea networks
uc3m
Τεχνολογικό Πανεπιστήμιο Κύπρου

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	IdSessions	USA	yearly	wordhealth	daily	wordhospit
wordsubmit	daily	wordstock	Download	IdSessions	people	wordtrend	wordmarket	wordidentifi
Download	IdCompanies	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	IdCompanies	Email	UIExport	wordreport	wordinsight	wordcontact

The table shows average scores of 5-fold executions of leave-one-out and permutation importance analysis. A median of 13 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	IdSessions	USA	yearly	wordhealth	daily	wordhospit
wordsubmit	daily	wordstock	Download	IdSessions	people	wordtrend	wordmarket	wordidentifi
Download	IdCompanies	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	IdCompanies	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. Even the ‘what’ seems to be more important than the ‘how much’ when pricing healthcare products

Among the rest of 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

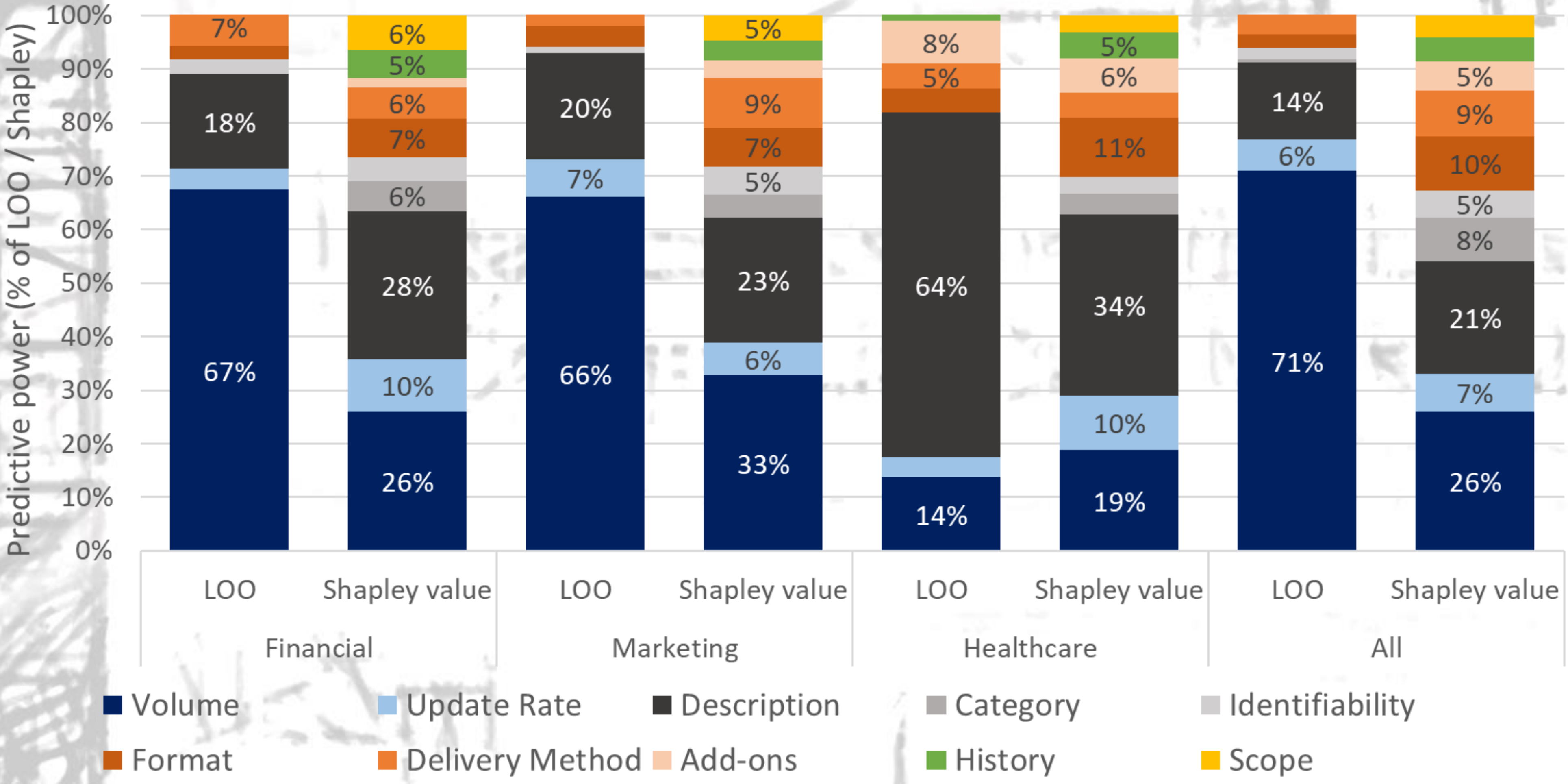
Features relating to delivery methods and update rate seem somewhat important for the prices of financial and marketing data

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

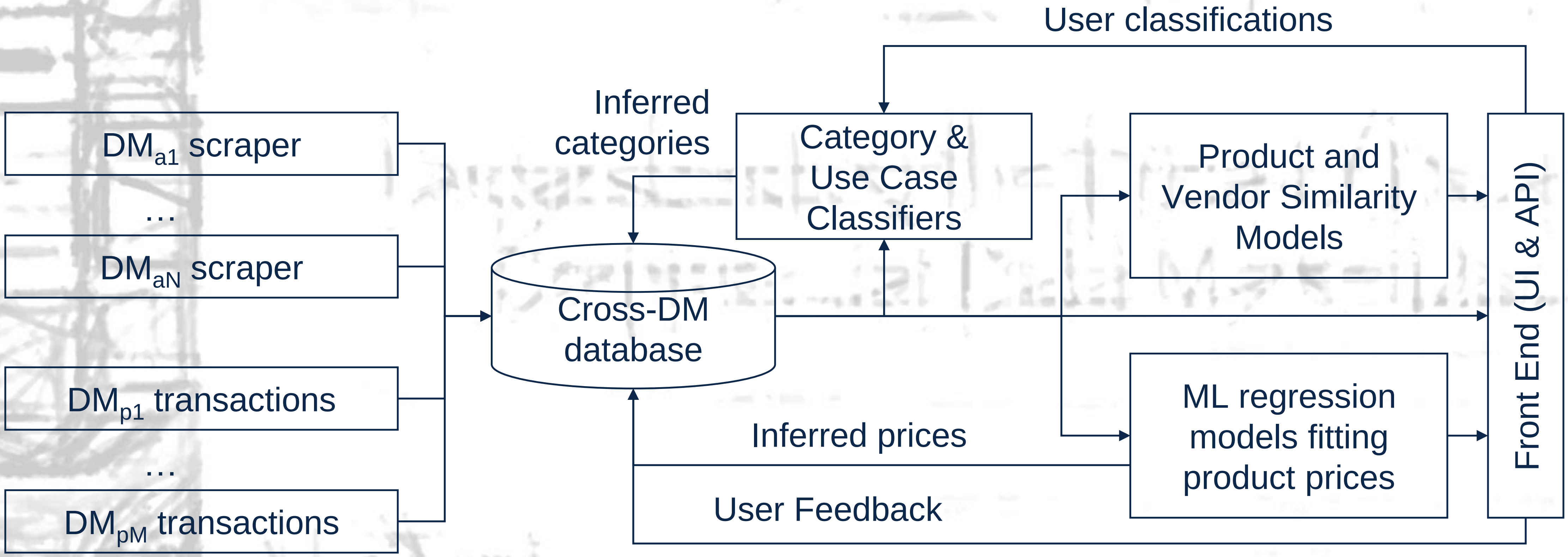
We studied the most influential feature groups, as well, resulting in notorious differences across data categories



In summary, or work is the first paper measuring the price of data in commercial marketplaces. We have 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

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 definitely have limitations, since it neglects: i) the usability for the buyer, ii) the quality of the data, iii) the specific value for a buyer.

Thank you!

Q&A time!

For more information please contact:

Santiago Andrés Azcoitia
santiago.azcoitia@imdea.org

institute imdea networks

Data Transparency Group

uc3m

Τεχνολογικό Πανεπιστήμιο Κύπρου

ICDE 2023
Anaheim, CA April 3 – 7, 2023