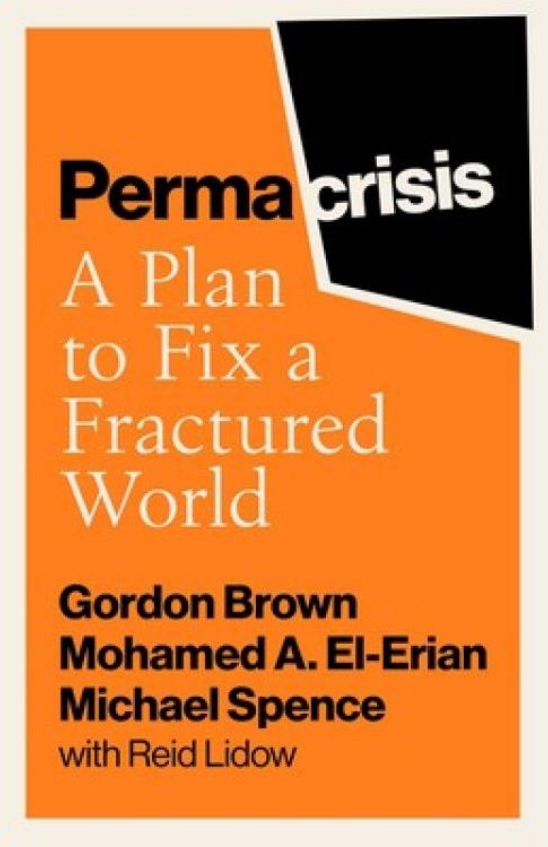
Towards a Human-Centric Data Economy

Santiago Andrés Azcoitia

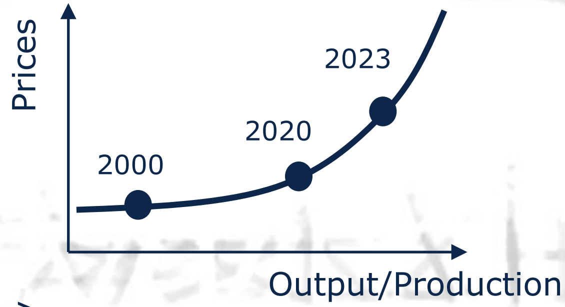
IMDEA Networks Institute

Developing the Science of Networks

AI was recently referred to as one of the tailwinds to propel economic growth in the next decades, …

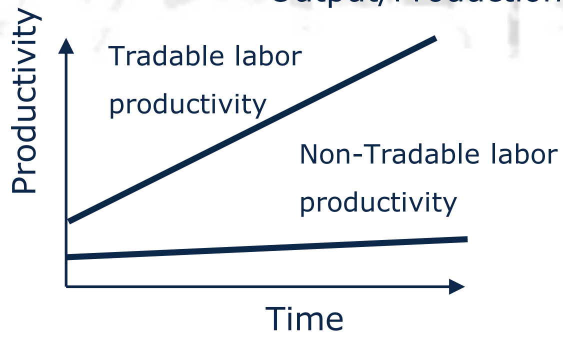


1



Inflation in the last year is a consequence of a **supply-constrained economy** that will last for some time (ageing, protectionism, etc.)

2



Asymmetry in productivity increase:

Even though tradable labour productivity (e-g., manufacturing goods) has significantly increased in the last years, non-tradable labour productivity (e.g., haircuts, waiters, telecom engineers, or travels) has not.

3

Their point is that **AI has the power of dramatically increasing non-tradable labour productivity**

… and AI/ML algorithms require data, thereby we need a global ecosystem to gather, organize and exchange data to create economic value

institute imdea networks
Data Transparency Group
POLITÉCNICA
UNIVERSIDAD POLITÉCNICA DE MADRID

? What makes data a special economic/tradable good?

Data is a peculiar ‘tradable good’…

- Available
- Costly
- Freely replicable
- Non-depletable
- Reusable
- Non-rivalrous

… whose value shows a special behaviour

- Context-specific
- Inherently combinatorial
- Increases with use
- Quality-driven
- Dependent on packaging
- Uniqueness & exclusivity

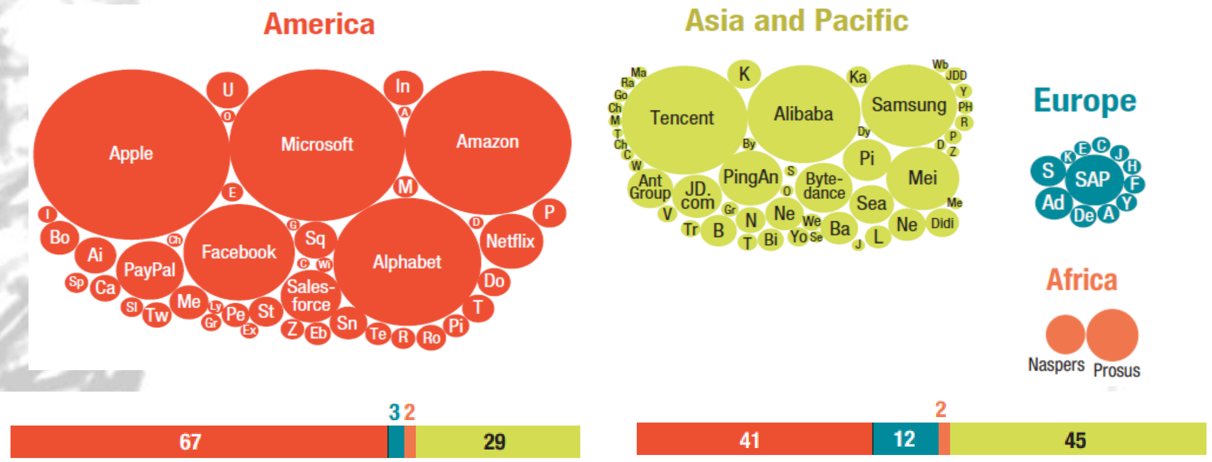
The nascent data economy is hindered by these particularities and, in spite of its huge potential, most data remains in corporate silos nowadays

Current data economy

1) Led by horizontally-integrated oligopolies

2) Geographically imbalanced

Top 100 Global Digital Platforms by market capitalization (2021)



Share in value by Region (%)

Nº top 100 platforms

3) Heavy overall impact

Data economy size and impact

Up to 827 bn€ in 2025 within EU27+UK (EC)

Data-driven decision-making to reach US$2.5 trillion globally by 2025 (McK)

AI to potentially deliver additional global economic activity of $13 trillion by 2030 (McK)

EU Strategy for Data focuses on putting people first in developing technology, and promoting European values and rights in the digital world

- "Building a European Data Economy" and the "European Strategy for data" as a key pillar of the "Shaping Europe's digital future" strategy

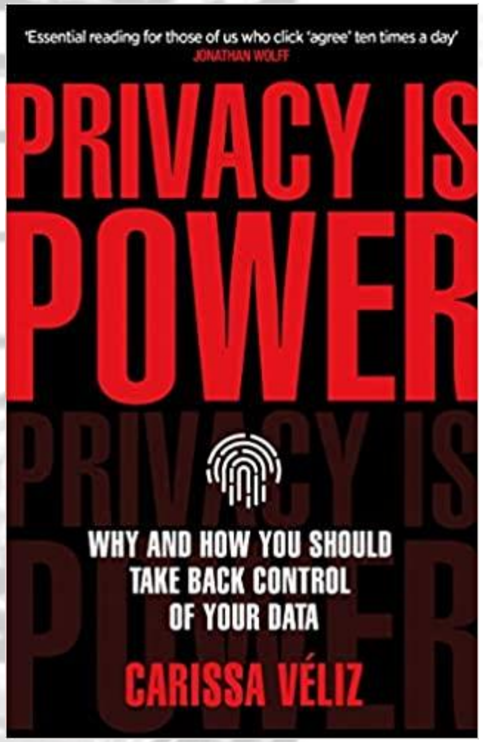
- Related policies about "Artificial Intelligence" with the strategy for "Artificial Intelligence for Europe", and about ensuring EU autonomy with European cloud services.

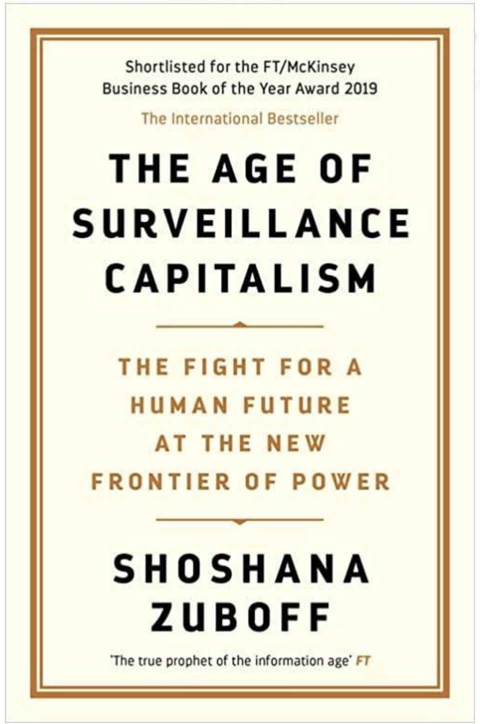
- New regulations:
  1) General Data Protection Regulation
  2) Regulation for the Free Flow of non-Personal Data and guidelines
  3) Data Governance Act
  4) Data Act
  5) AI Act

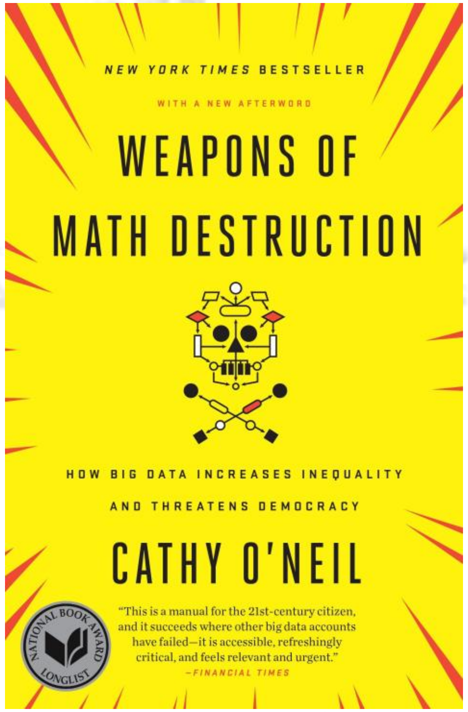
- Other policy initiatives to create a common data space in the EU:
  1) Promote open data initiatives to enable the reuse of public information
  2) Recommendation on access to and preservation of scientific information
  3) Guidelines to private data sharing

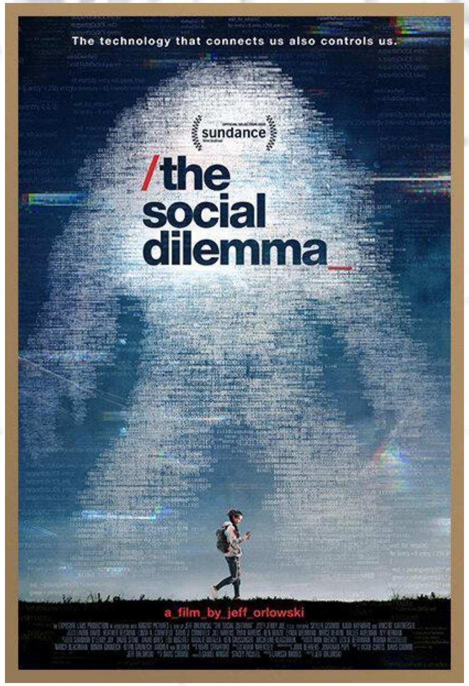
- Initiatives and projects towards sovereign, secure, trusted data exchange standards: International Data Spaces and Gaia-X

The massive collection and exploitation of personal data in exchange of services has raised a general concern about privacy and AI ethics…

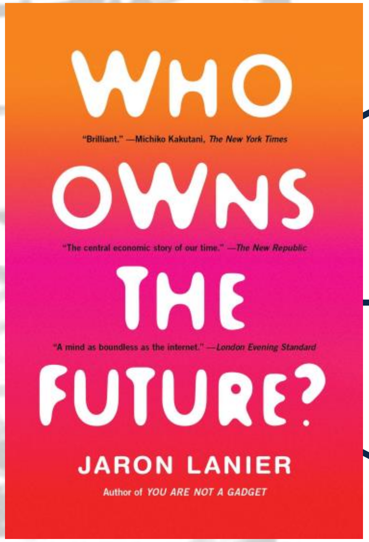


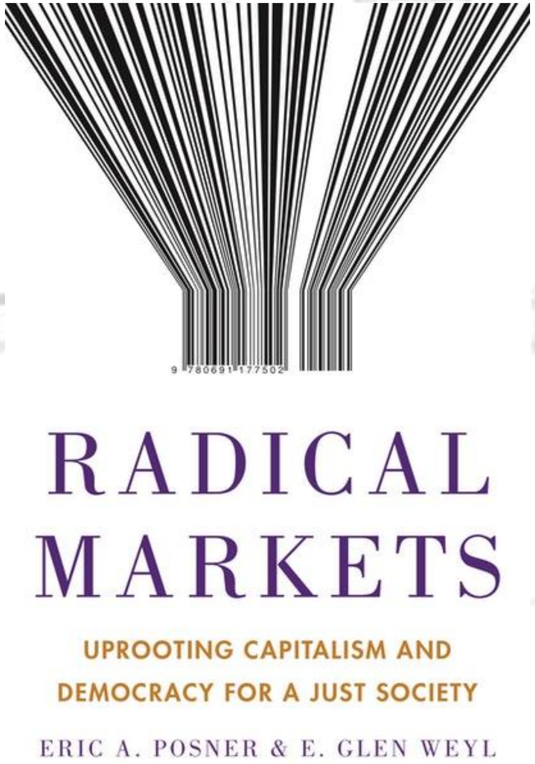






… and remarkable voices have warned against unsustainable digital economics, and proposed to retribute people for their data as a solution





Te deben 18.490€ al año por tus datos: una revolucionaria teoría sacude el capitalismo. El Confidencial 20 Feb 2020.
¿Acabaremos cobrando por ceder nuestros datos? ABC. 26 Feb 2020.
El investigador que propone recibir un salario a cambio de nuestros datos. El País 10 mar 2020.

Some dare estimate a transfer of 9% of the data economy from companies to owners, meaning +US$20k yearly income for a family of 4 in the US

The ‘data dividend’ in California, the ‘data tax’ in NYC, or digital service taxes (DSTs) in Europe may require to put a financial value on data

Los Angeles Times

POLITICS

Newsom wants companies collecting personal data to share the wealth with Californians

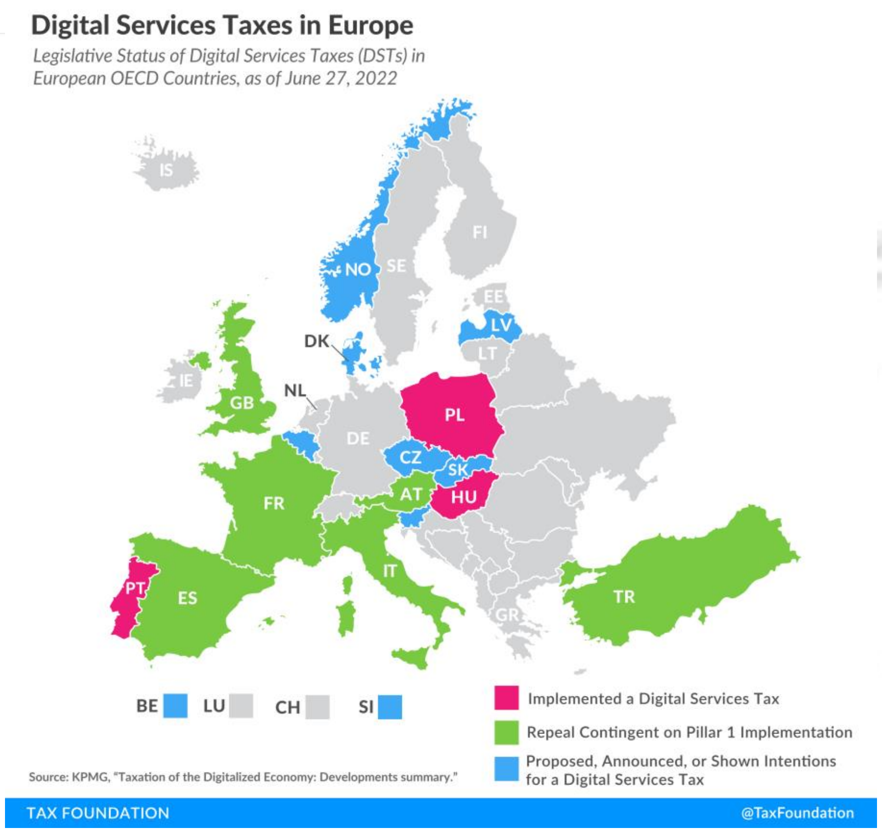
https://www.latimes.com/politics/la-pol-ca-gavin-newsom-california-data-dividend-20190505-story.html

OPINION | COMMENTARY

A Tax on Data Could Fix New York’s Budget

New revenue from information brokers to plug the Covid hole.

https://taxfoundation.org/new-york-data-tax-proposal/



https://taxfoundation.org/digital-tax-europe-2022/

Unlocking the value of data and ensuring data markets is key to harness the potential of AI in the economy

Understanding and Measuring the Data Economy

Addressing Technical Challenges

Regulating the data economy

Most of the material in this presentation is part of my PhD thesis “Towards a Human-Centric Data Economy”

Understanding and Measuring the Data Economy

Addressing Technical Challenges

Regulating the data economy

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

S. Andrés Azcoitia and N. Laoutaris, A Survey of Data Marketplaces and their Business Models. ACM SIGMOD Record Sept. 2022

At a high-level, we spotted 3 main families of business models depending on whom companies target their services:

Owners ⟷ Buyers

Data management Systems (DMS)

mydex, ErnieApp, wecple, digi.me, Swash, geodb, AIRBLOC, meeco, VETRI, COGNITE, DATA REPUBLIC, Mine

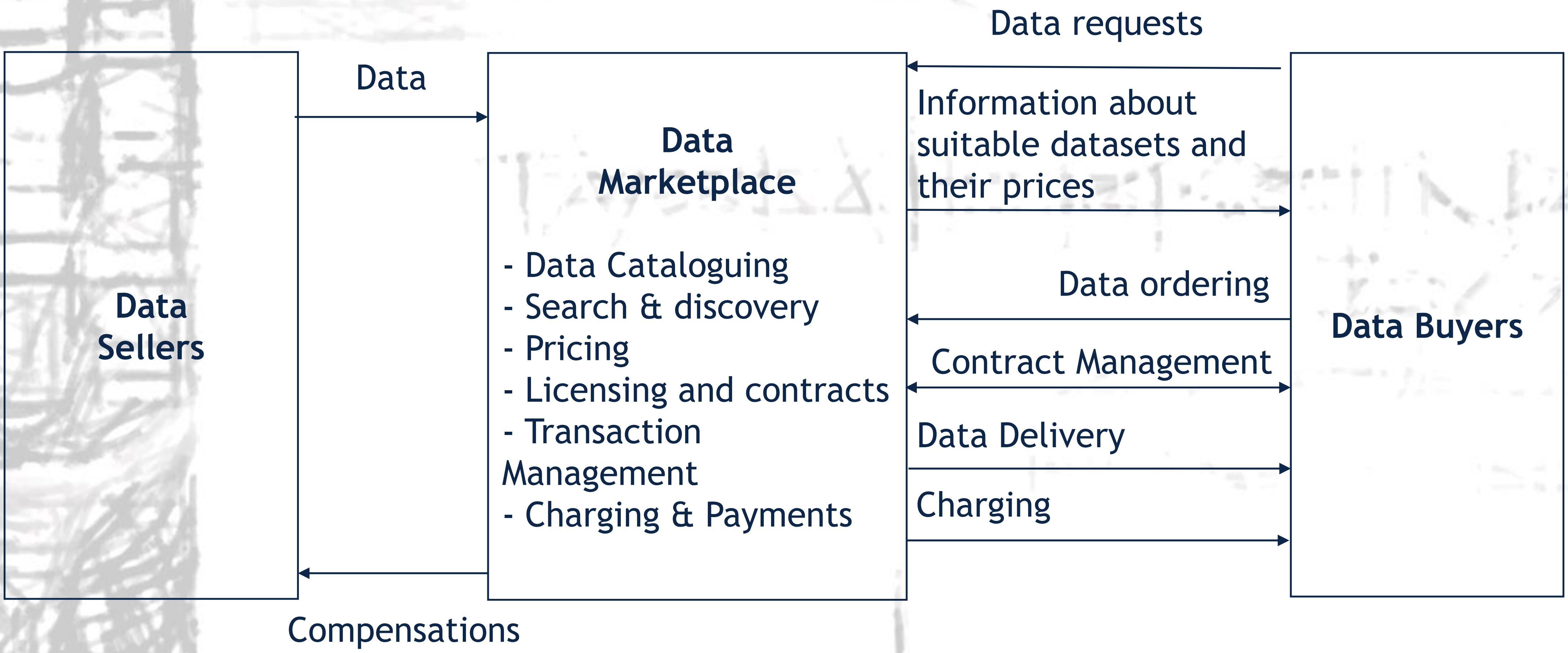
Data Marketplaces (DM)

databroker (the marketplace for data), otonomo, Datarade, TERBINE, VERACITY by DNV GL, Demyst, BattleFin, ADVANEO (data marketplace), DATA INTELLIGENCE HUB, knoema, CARUSO, data marketplace, NOKIA, DAWEX, ocean, IOTA, dHealth Network, BIGCHAIN DB

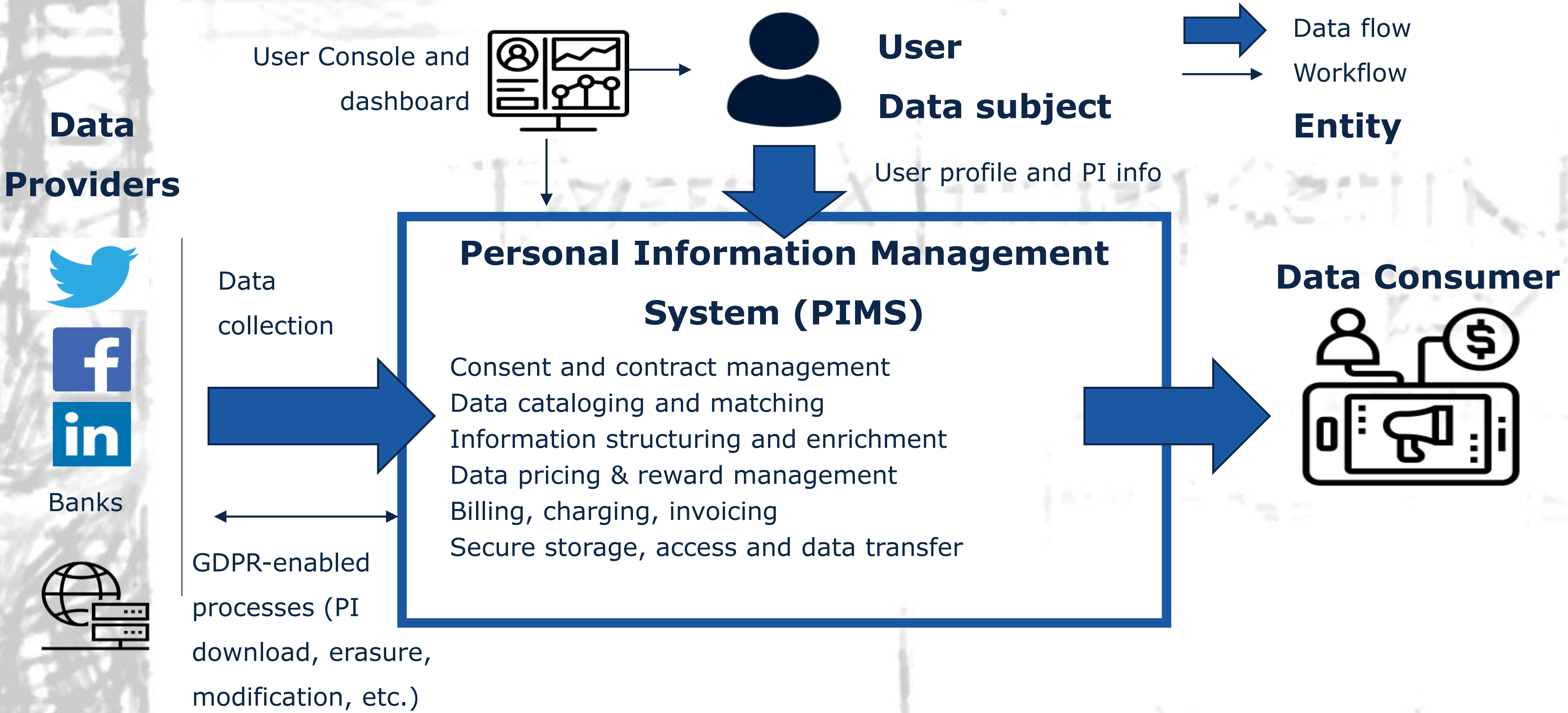
Data Providers (DP)

acxiom, opencorporates, DATA SCOUTS, REFINITIV, experian, Clearview.ai, FYSICAL, BOOK YOUR DATA, bronID, EVOTEGRA, ArcGIS Marketplace, CARTO, LOTAME, theTradeDesk

Data Marketplaces



Personal Information Management Systems (PIMS)



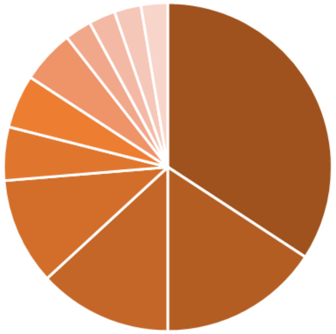
S. Andrés Azcoitia and N. Laoutaris, A Survey of Data Marketplaces and their Business Models. ACM SIGMOD Record Sept. 2022

We can classify entities based on the kind of data they trade, which also depends on their business models

Owners ⟷ Buyers

Data management Systems (DMS) | **Data Marketplaces (DM)** | **Data Providers (DP)**

- Personal Data
- Any
- Marketing
- Healthcare data
- Geo-located data



- Any
- Healthcare
- Automotive
- IoT Sensor
- Alternative data
- Personal
- Trading
- AI / ML models
- Genetic
- Industry
- Marketing

- Marketing
- Corporate data
- Trading data
- Contact data
- Automotive-related
- Geo-located data
- Alternative data
- AI / ML models
- Human-generated data
- Web data
- Multimedia
- Identity data

We characterized up to 10 different business models based on different dimensions of analysis

Concept	DP/SP	PMP	General-purpose	Niche DMs	Embedded DM	PIMS
Data exchange	Public, semi-private. private	Private	Public / semi-private		Private	Public / semi-private
Scope	Focused	Focused	Diversified	Focused	Focused	
Type of data	Any	Specific data to be used within their service / platform	Any	Industry-, or type-specific	Data to be exchanged within the system	Personal data
Roles / Players interacting	Partners, Customers		Sellers, Buyers		Owner, Requester	Users, Data Providers, Buyers
Gets data from	Internet, self-generated, partners, users	Partners, Data providers	Data providers	Data providers, self-enriched	Data providers	Users, Data providers
Provides buyers with	API, Datasets	API, Access to data through the system	API, Datasets		API, Access to data through the system	API, Key to decrypt data
Owners access through	Partnership	Partnership & the platform	Web-services		Data Management platform	Mobile App Web services
Buyers get data through	Web-services, APIs	Web-service, the platform	Web-services	Web-services, APIs	Data Management platform	Web-services, APIs, compatible systems
Type of platform	Centralised		Centralized or Decentralised		Centralised	Decentralised
Access Pricing for buyers	Subscription Pay for data	Included in the main platform	Predominantly free. Some freemium, subscription, and data delivery charges		Add-on to the data management platform	Pay for data
Access Pricing for sellers	Partnership (when applicable)	Partnership Subscription	Predominantly free, freemium, subscription, and revenue-share charges		Subscrition to the platform	Free
Prices set by	Platform	Platform, Buyers	Platform, Providers	Platform, Providers	Open	Users, Platform
Pricing schemes	Fixed one-off, subscription, customized, volume-based	Subscription, domain-specific (CPC, CPM,…)	Fixed one-off, subscription and customised	Customised, volume/usage-based, fixed one-off	Open	Open, Bid by buyer
Payment method	Fiat currency			Fiat currency, token	Open	Token, fiat currency

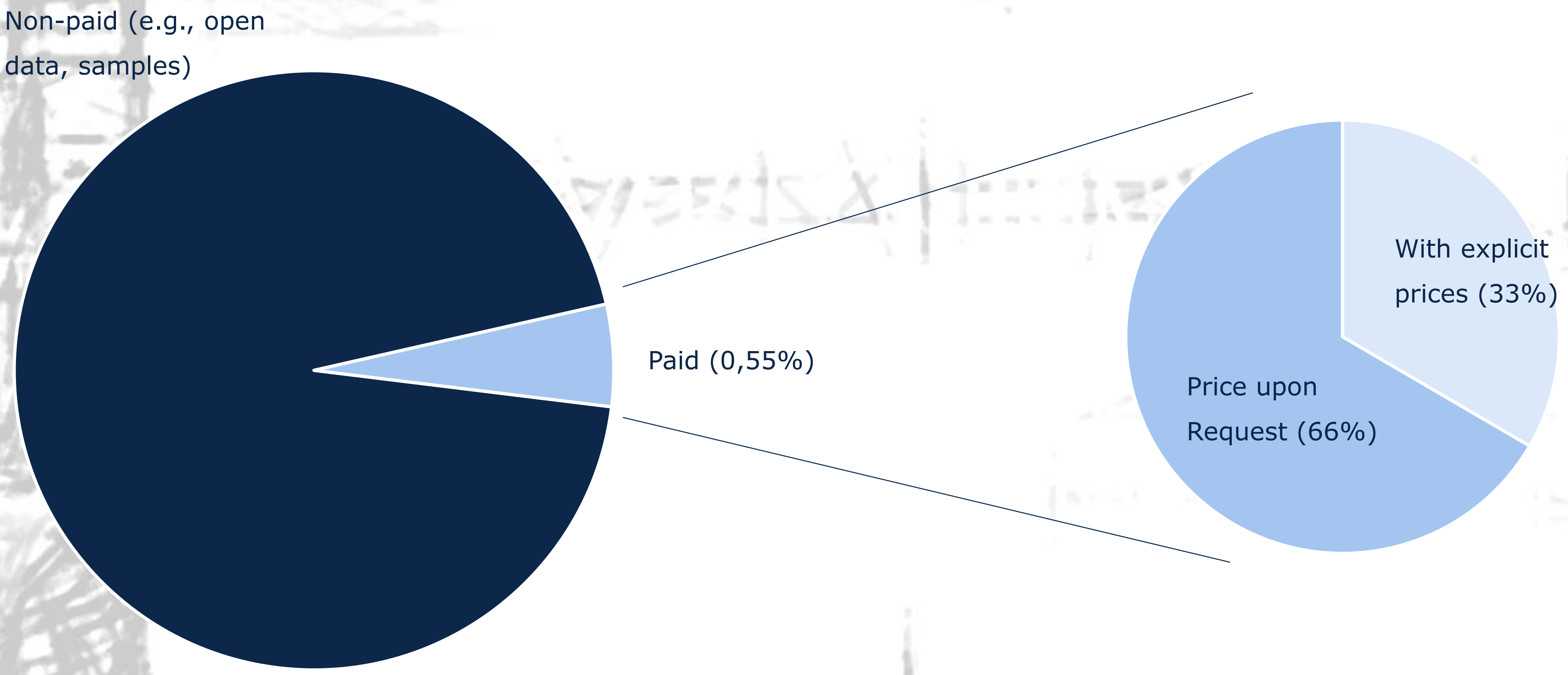
institute imdea networks

Data Transparency Group

POLITÉCNICA UNIVERSIDAD POLITÉCNICA DE MADRID

S. Andrés Azcoitia and N. Laoutaris, A Survey of Data Marketplaces and their Business Models. ACM SIGMOD Record Sept. 2022

We went further and, in another recent market study, we scraped metadata of +210k products from 10 DMs, +2k DPs



S. Andrés Azcoitia, C. Iordanou, and N. Laoutaris. Measuring the Price of Data in Commercial Data Marketplaces. In Proc. of 1st ACM DE Workshop (2022)

What's the price of data? The problem

How valuable is this?

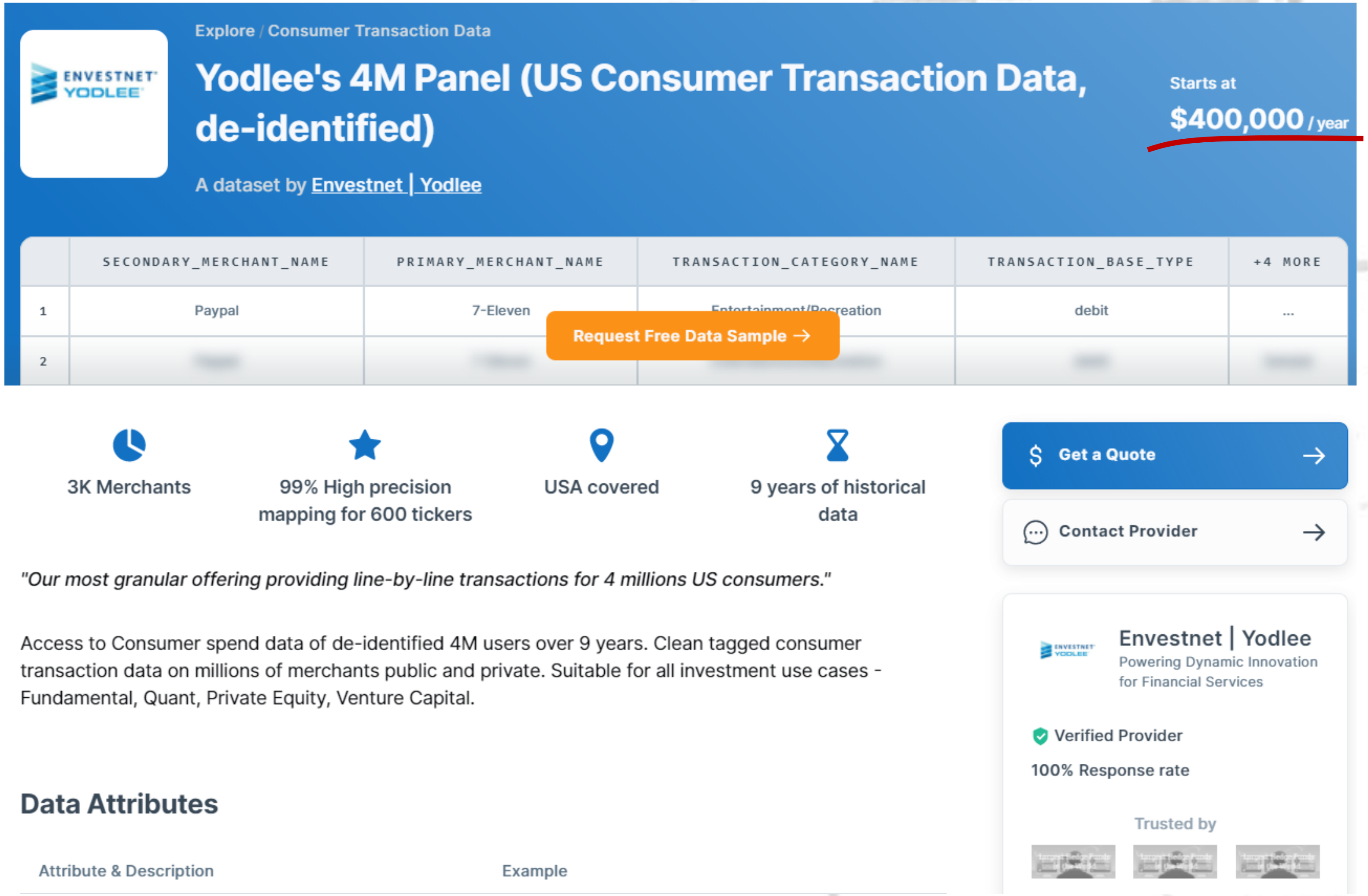
How about this?

? What’s the price of data? The problem

And 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

How does a data product look like in a data marketplace?



S. Andrés Azcoitia, C. Iordanou, and N. Laoutaris. Measuring the Price of Data in Commercial Data Marketplaces. In Proc. of 1st ACM DE Workshop (2022)

How does a data product look like in a data marketplace?

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

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)

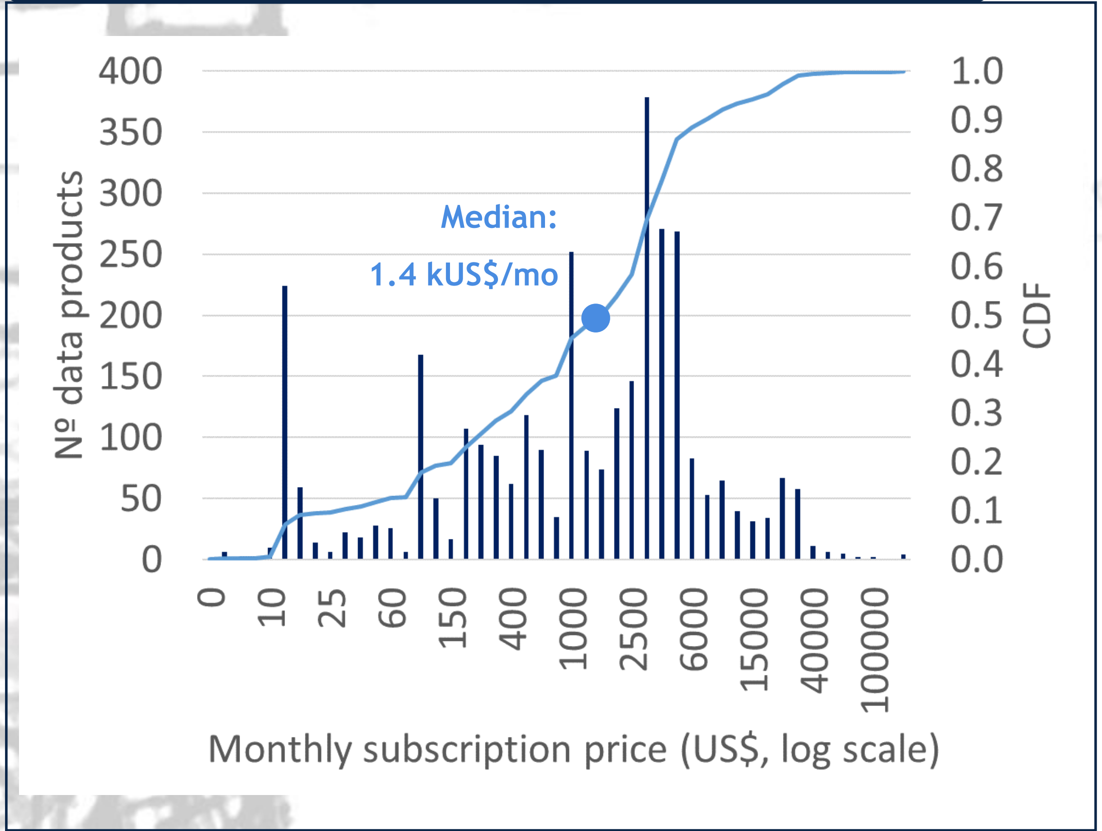
Overview one sheeter: https://info.alliantinsight.com/hubfs/Downloadable%20Content%20/Alliant%20AWS_Geo_Performance.pdf

Provided By

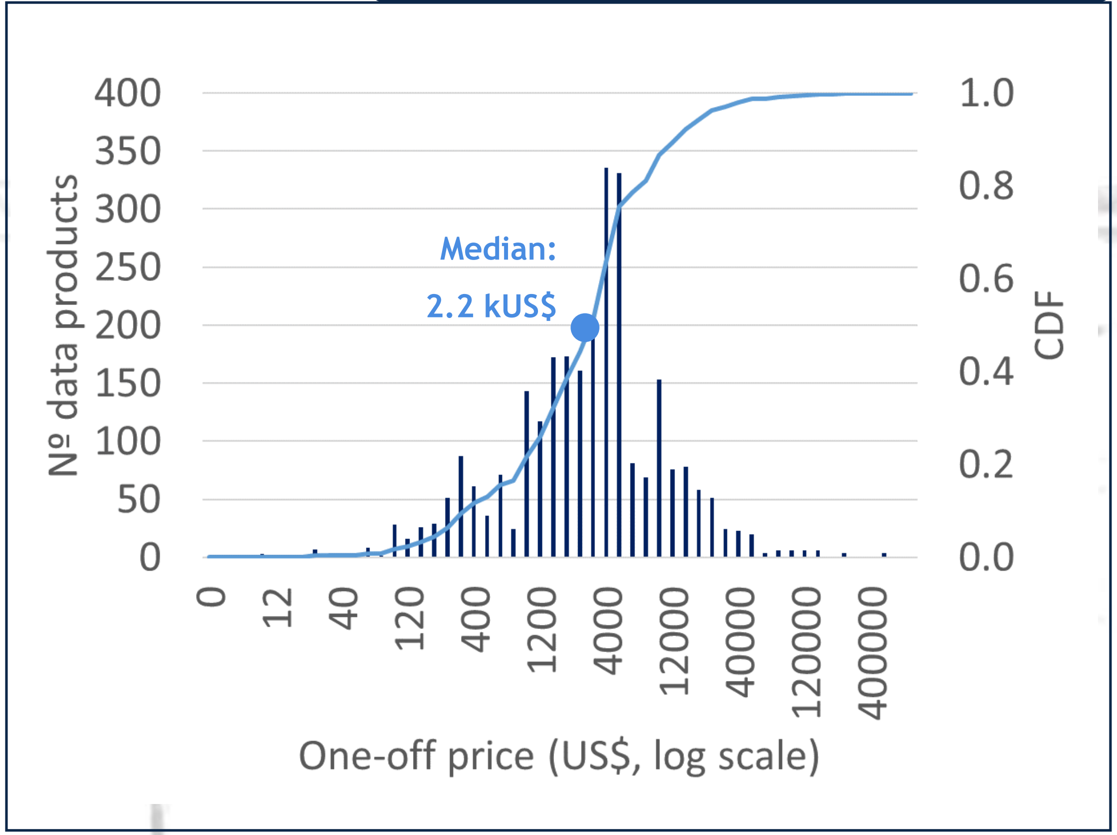
Alliant

We found very heterogeneous data that sells at an immensely wide range of prices up to US$800k or US$150k per month, …

Subscription-based data product prices



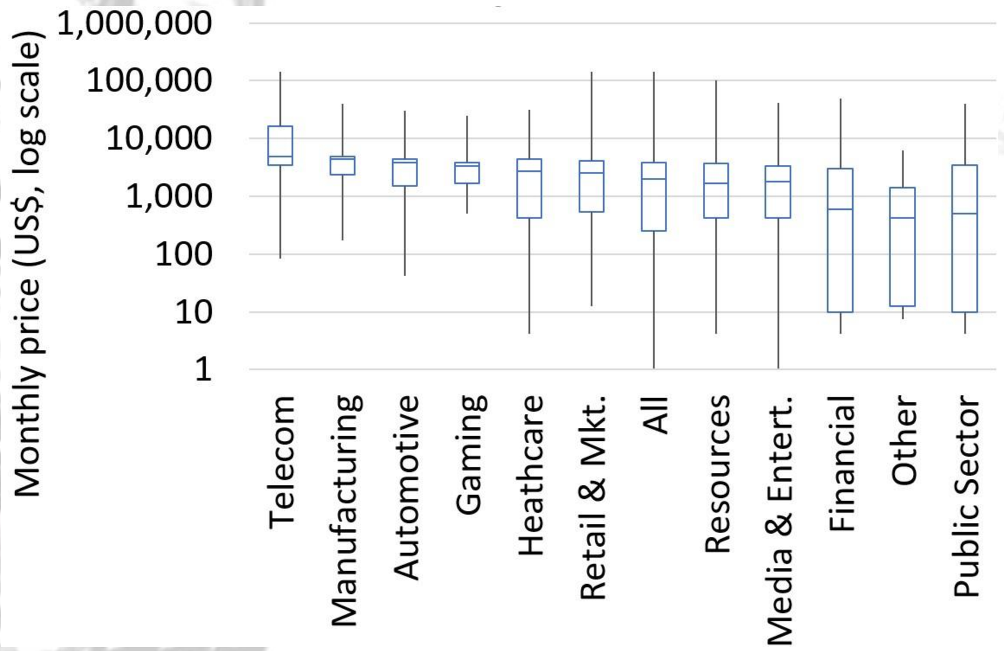
One-off data product prices



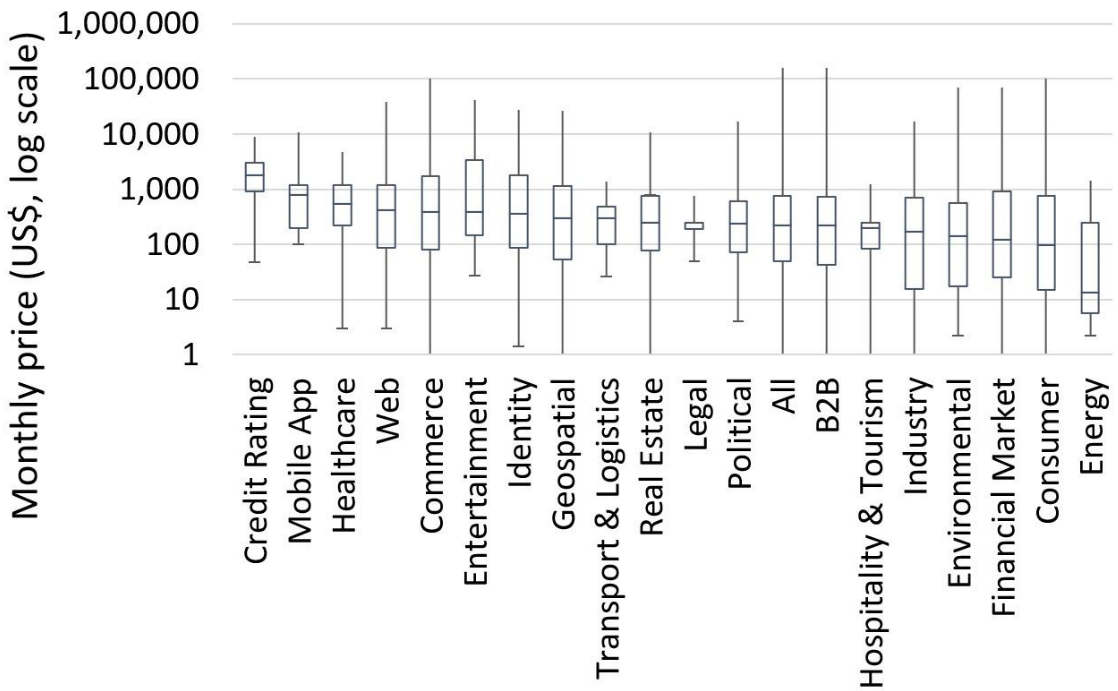
S. Andrés Azcoitia, C. Iordanou, and N. Laoutaris. Measuring the Price of Data in Commercial Data Marketplaces. In Proc. of 1st ACM DE Workshop (2022)

… which depend on the category of data products

Data product prices by category AWS



Data product prices by category DataRade



S. Andrés Azcoitia, C. Iordanou, and N. Laoutaris. Measuring the Price of Data in Commercial Data Marketplaces. In Proc. of 1st ACM DE Workshop (2022)

We built a cross-DM database of metadata of products offered in different DMs

- 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 the price of financial, marketing and health-related data

R^2 score by model and category

Model\Cat.	Financial	Marketing	Healthcare	All
RF	0.85	0.86	0.78	0.84
kN	0.78	0.74	0.77	0.69
GB	0.82	0.80	0.73	0.79
DNN	0.73	0.77	0.68	0.72

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

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 *perhaps* units that relates to the price of data across categories

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

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

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

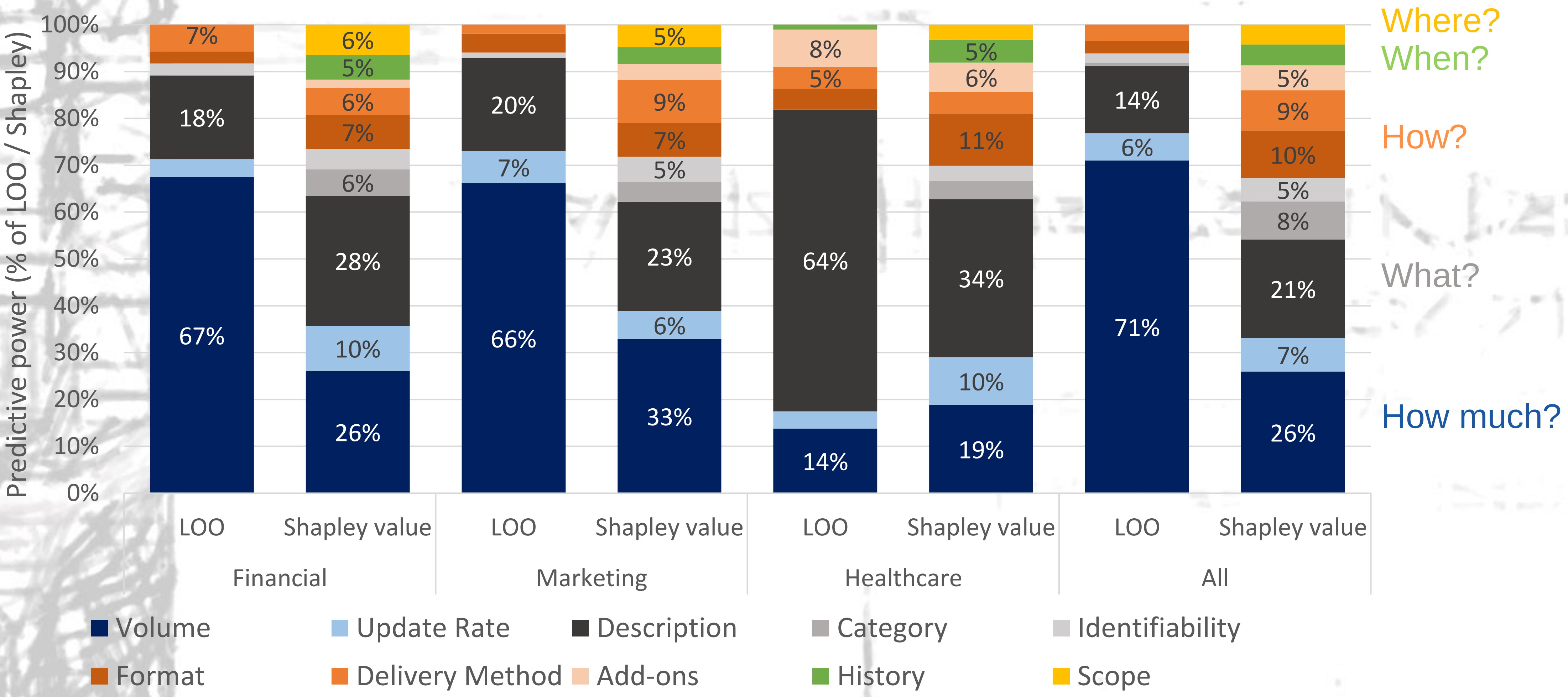
Delivery methods and update rate seem somewhat important for the prices of financial and marketing data

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

Geo-spatial localization and scope and the possibility of connecting data points from the same owner are also present especially in marketing 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

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



S. Andrés Azcoitia, C. Iordanou, and N. Laoutaris. Understanding the Price of Data in Commercial Data Marketplaces. In Proc. of 39th IEEE ICDE (2023)

Understanding and Measuring the Data Economy

Addressing Technical Challenges

Regulating the data economy

During our survey and research of SOTA we identified a number of challenges data markets are facing:

MARKET CHALLENGES

Protecting ownership & earning trust

Federating and standardizing data sharing to deal with the current fragmentation of data markets

Setting up knowledgeable neutral price references

Anticipating the value of data for a specific task

Computing fair compensations for data providers and owners at scale

MARKET CHALLENGES

Protecting ownership & earning trust

Federating and standardizing data sharing to deal with the current fragmentation of data markets

Setting up knowledgeable neutral price references

Anticipating the value of data for a specific task

Computing fair compensations for data providers and owners at scale

Several “schools” of researchers are dealing with data pricing problems with very different approaches:

AUCTION

- Are they useful when pricing data?
- Random auctions [Goldberg01]
- CORE auctions [Goldberg03]
- They artificially create competition between Bidders

AI / ML

- Model-based pricing [Chen18]
- Utility & quality-based [Agarwal19]
- Collaborative ML markets [Ohrimenko19]

QUERY DM

- Query determinacy [Koutris12]
- Arbitrage freeness [Balazinska13]
- Revenue maximization [Chawla19]

PRIVACY DM

- Selling privacy at auction [Ghosh11]
- **Privacy preserving** for buyers (e.g., info of their purchases), sellers (sensitive, PI or info about sales), and third parties (e.g. PI of individuals)

QUALITY-BASED

- Asseses value of data depending on quality features [Heckman15]
- Monopolistic quality-based pricing [Yu17]

DYNAMIC PRICING

- **Pricing dynamic data**: e.g. history-aware pricing (API, Query)
- **Dynamic data pricing**: [Niu19] maximize cumulative revenue in time

Some tools widely used when pricing digital products may be useful in pricing data, as well

BUNDLING

- Data (service) providers price together the access to data products (e.g., data for a platform)
- When is it convenient? In general, it is convenient when price-sensitive buyers consider products as complementary
- There is a framework to study the conditions under which bundling produces more revenues [Daskalakis17]
- Pure bundling is optimal if consumers with higher values for the grand bundle have comparatively higher relative values for smaller bundles [Haghpanah20]
- In general, Both papers assume a multi-product monopolist.

VERSIONING

- Refers to selling different versions of a data product, with different utility and price
- *Freshness, history, features, scope, volume, format, resolution or accuracy* of data are being used to offer different versions of a data product
- **AI / ML**: noise injection to data or models
- **Query DM**: Noise injection to data
- **Location-based**: precision of data location
- **Privacy DM**: noise injection to increase differential privacy ε
- **Quality-based**: different versions of data with different mix of quality features

MARKET CHALLENGES

Protecting ownership & earning trust

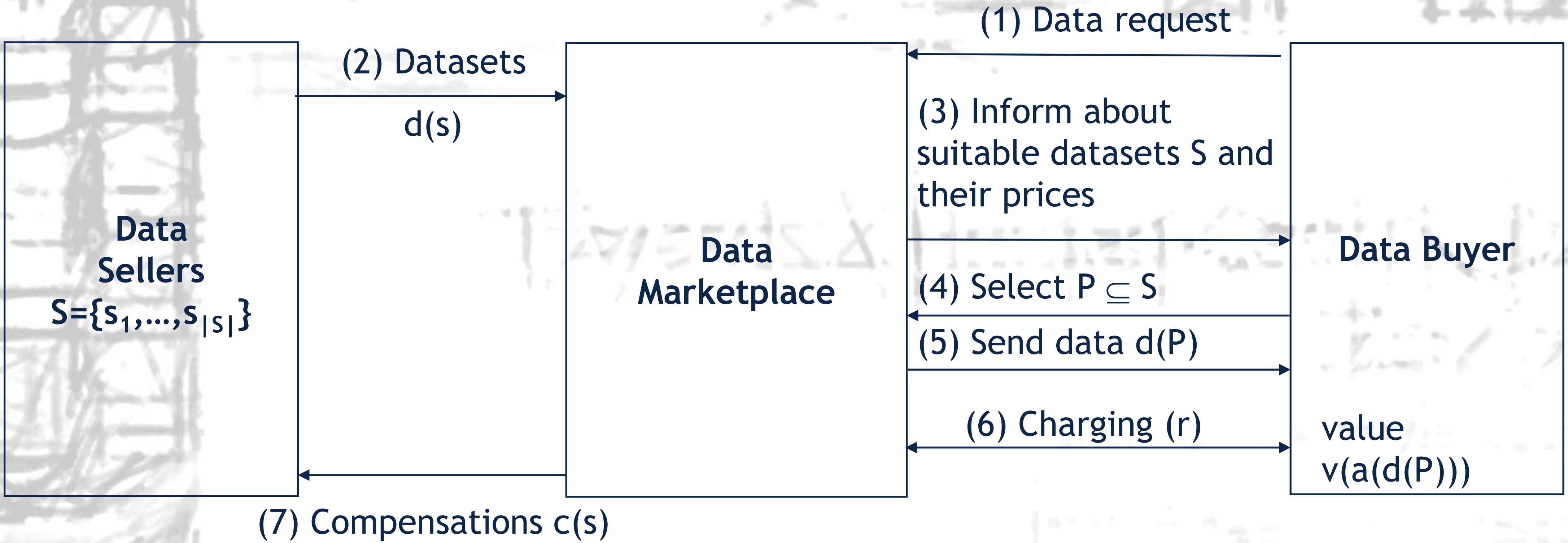
Federating and standardizing data sharing to deal with the current fragmentation of data markets

Setting up knowledgeable neutral price references

Anticipating the value of data for a specific task

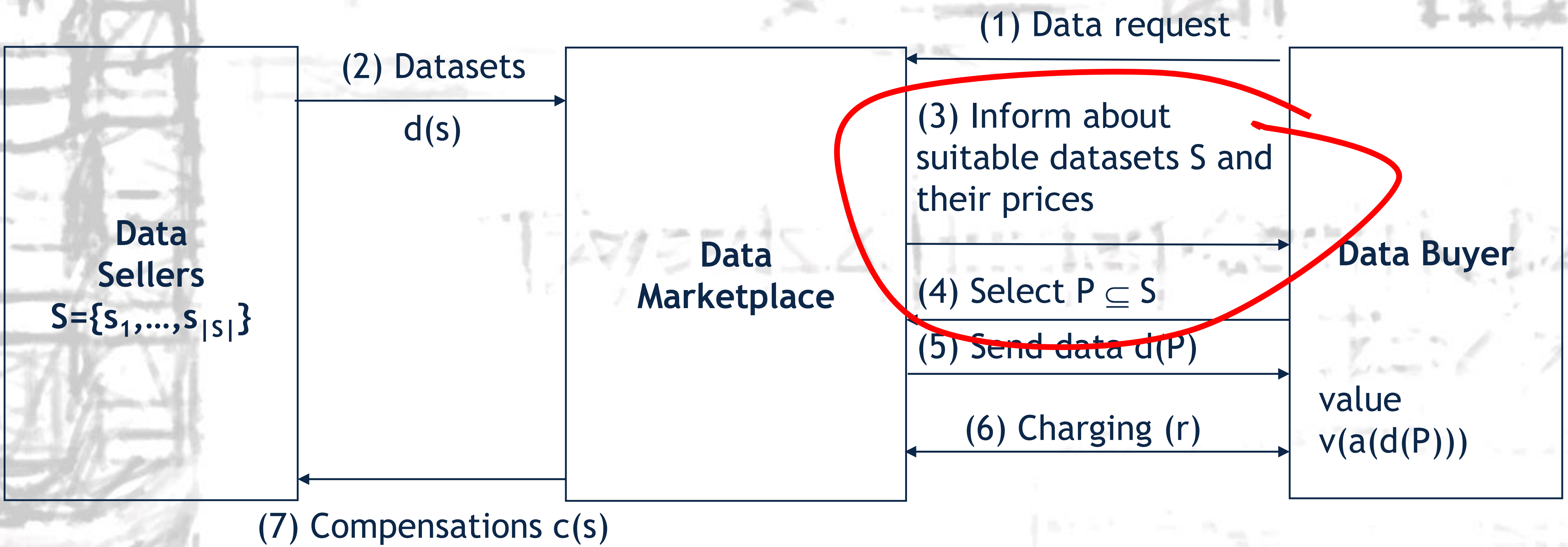
Computing fair compensations for data providers and owners at scale

A data marketplace model



S. Andrés Azcoitia and N. Laoutaris. Try Before You Buy: A practical data purchasing algorithm for real-world data marketplaces. In Proc. of 1st ACM DE Workshop (2022)

P1: How do buyers select data that suit their tasks?



S1: Buy the most valuable combination of datasets

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

S. Andrés Azcoitia and N. Laoutaris. Try Before You Buy: A practical data purchasing algorithm for real-world data marketplaces. In Proc. of 1st ACM DE Workshop (2022)

We proposed a preliminary "evaluation" phase prior to buyers selecting which data to acquire and a family of algorithms (Try-Before-You-Buy)…

Data Sellers
$S=\{s_1,\ldots,s_{|S|}\}$

(2) Datasets
d(s)

Data Marketplace

(1) Model (*M*), accuracy (a), test set

(3) Inform:
i) Price (p),
ii) Description
iii) Accuracy (a)

(4) Select $P \subseteq S$

(5) Send data d(P)

Data Buyer

value
v(a(d(P)))

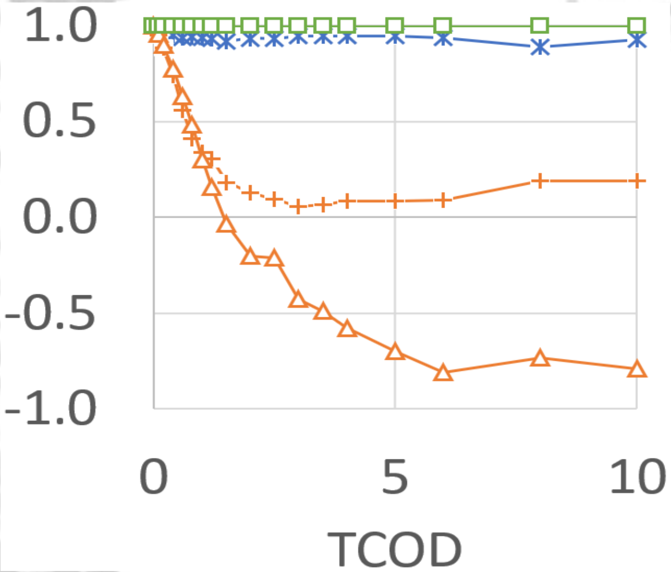
… which is "easily" implementable using "sandboxes" of some commercial DM:

BattleFin otonomo Swash ADVANEO data marketplace CARUSO

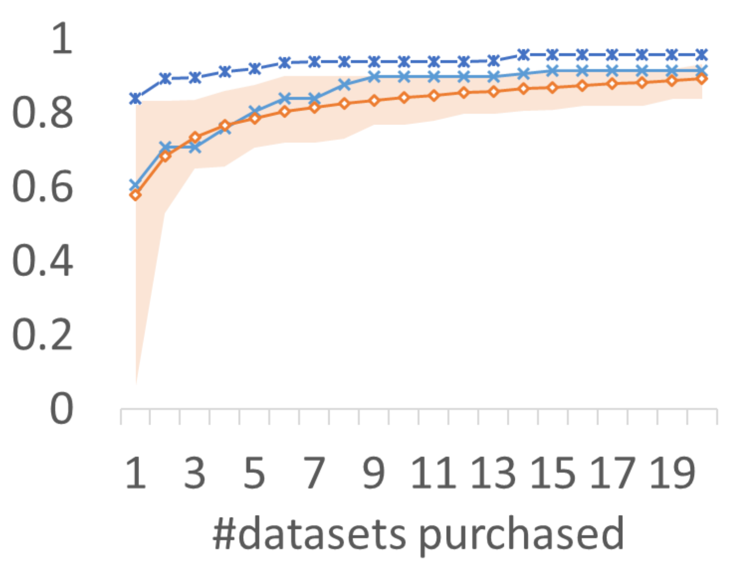
S. Andrés Azcoitia and N. Laoutaris. Try Before You Buy: A practical data purchasing algorithm for real-world data marketplaces. In Proc. of 1st ACM DE Workshop (2022)

We proposed a preliminary “*evaluation*” phase prior to buyers selecting which data to acquire and a family of algorithms (Try-Before-You-Buy)

1 TBYB was shown to yield near-optimal profits to buyers under a wide range of parameters and data in O(N) – O(N^2) execution time



2 TBYB allows buyers to filter individuals whose data is more suitable for a certain task, reducing the amount of information exchanged and hence the privacy leakage



TBYB algorithms select the best datasets and stop purchasing in the right time

S. Andrés Azcoitia and N. Laoutaris. Try Before You Buy: A practical data purchasing algorithm for real-world data marketplaces. In Proc. of 1st ACM DE Workshop (2022)

MARKET CHALLENGES

Protecting ownership & earning trust

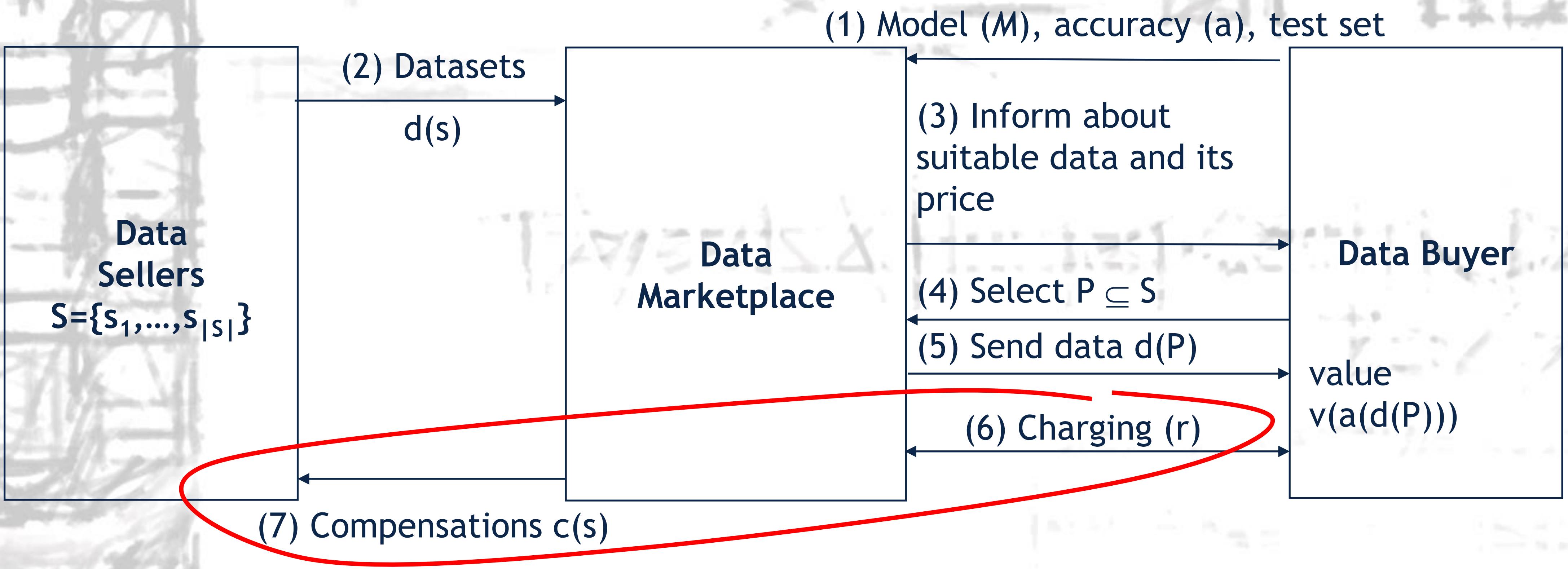
Federating and standardizing data sharing to deal with the current fragmentation of data markets

Setting up knowledgeable neutral price references

Anticipating the value of data for a specific task

Computing fair compensations for data providers and owners at scale

P2: What is the relative value of data from different data sources?



P2: *How do DMs distribute payoffs fairly?*

$$\mu(s_i) = f(S_{s_i}, \{S_j\}, \mathcal{M}, v), j \in P - \{s_i\}$$

$$c(s_i) \; \propto \; \mu(s_i)$$

Can you think of ways to reward data sellers/owners for their data?

Most research works resort to the Shapley value, which is the average marginal contribution of a data source to every possible combination of the rest of them

Nº years old
20
10
60
Average (target) 30

Set	Guess	Error	Score
	20	10	0.67
	10	20	0.33
	60	30	0
	15	15	0.5
	40	10	0.67
	35	5	0.83
	30	0	1

= 0.67 x 2

- = 0.17

- = 0.67

- = 0.17 x 2

Average (SV) 0.42

0.33

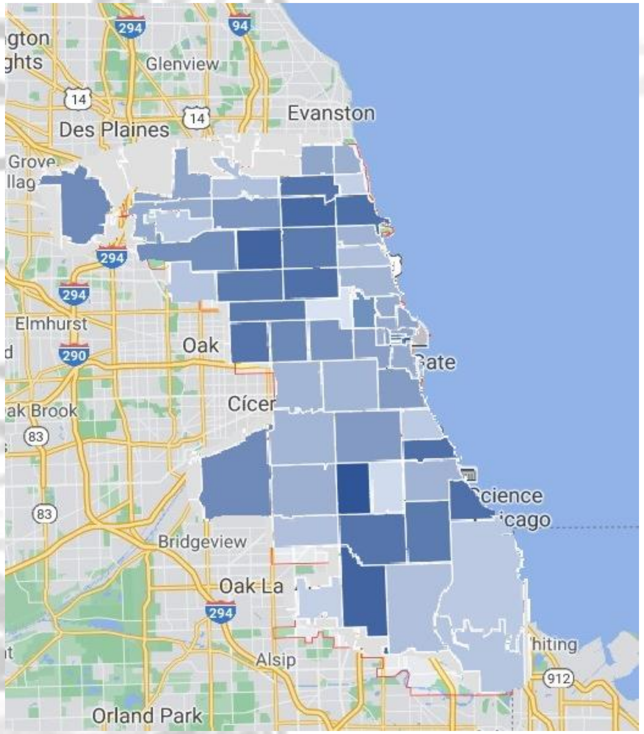
0.25

However, slight variations of the model, the valuation function, the test set or the initial data have a dramatic impact on the value of different players, …

Use case		Alice	Bob	Carlos	Sum
1	Base case	0.42	0.33	0.25	1
2	Max	0.14	0.06	0.8	1
3	Biased test set	0.32	0.24	0.35	0.91
4	Using RMSE	0.49	0.36	0.15	1

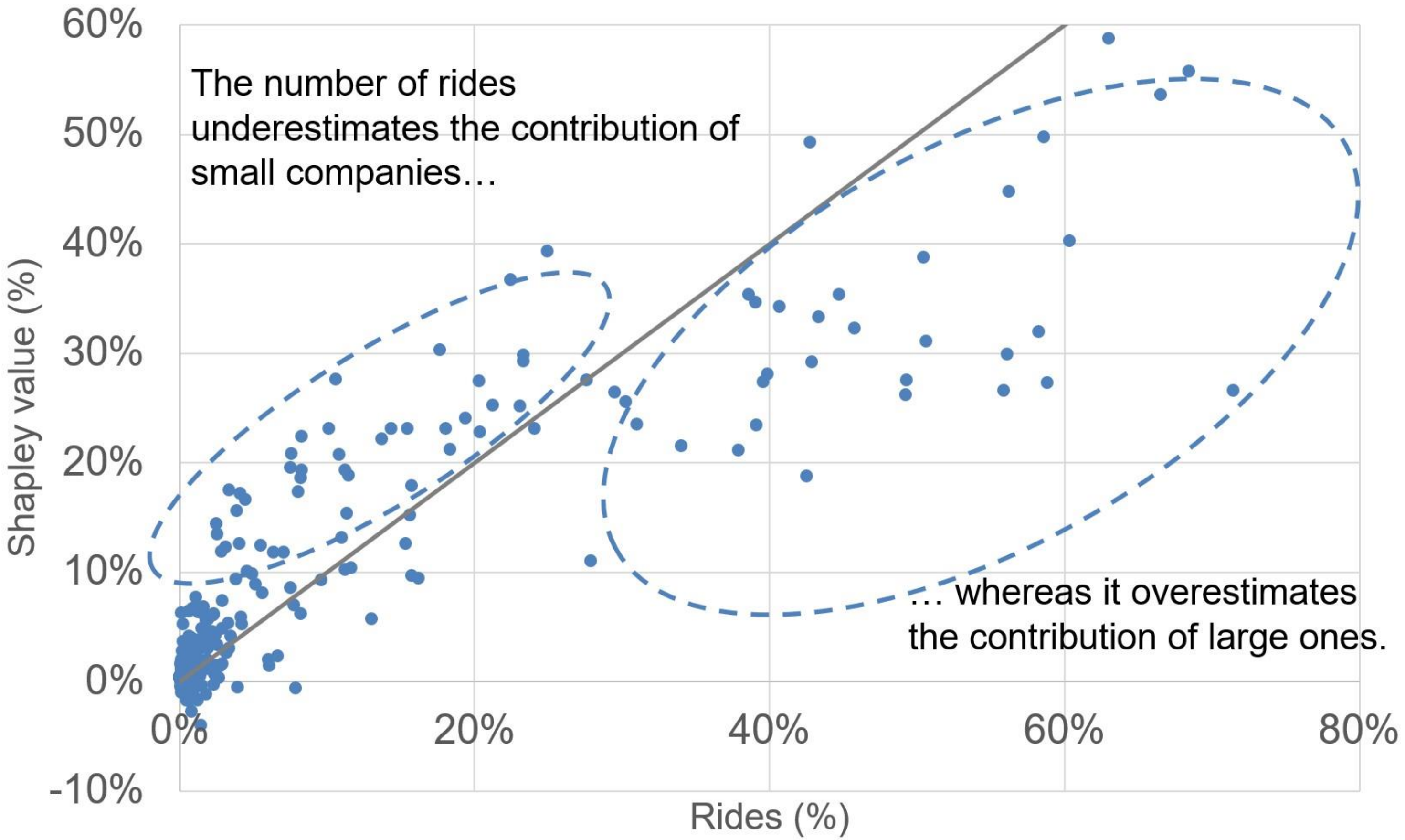
… let alone distributing rewards based on value can be arguable and difficult to explain to end users.

We found that the number of rides does not necessarily reflect the value that data from a taxi company adds to predicting future transportation demand…



Data from a taxi company can be very useful to predict vehicle-for-hire demand in a certain district of the city, but not in others

SV vs. nº rides by company in small districts of Chicago

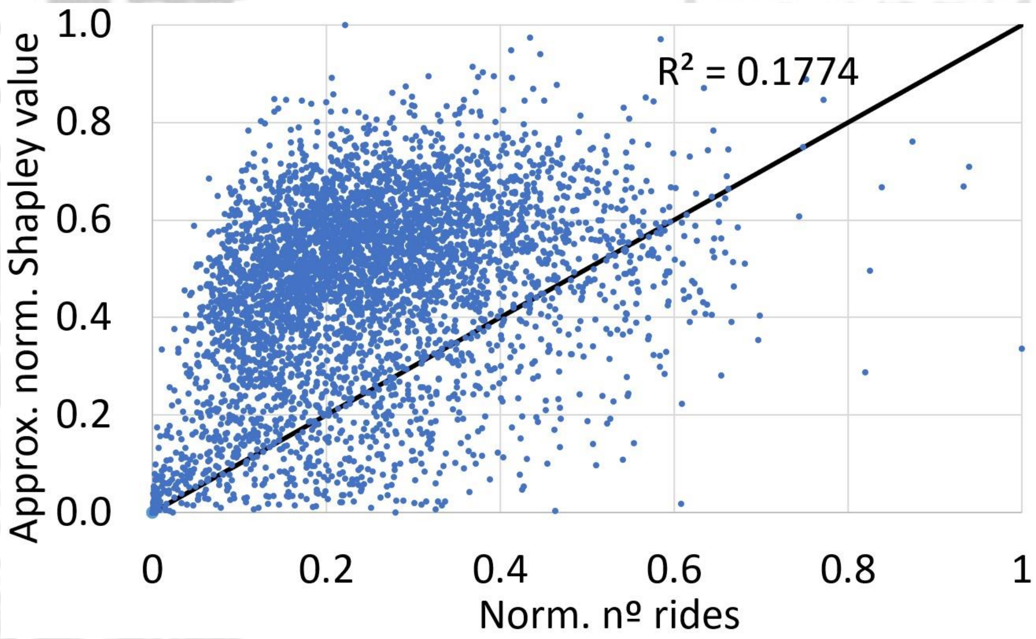


S. Andrés Azcoitia, M. Paraschiv, and N. Laoutaris. Computing the Relative Value of Spatio-Temporal Data in Data Marketplaces. In Proc. of ACM SIGSPATIAL (2022)

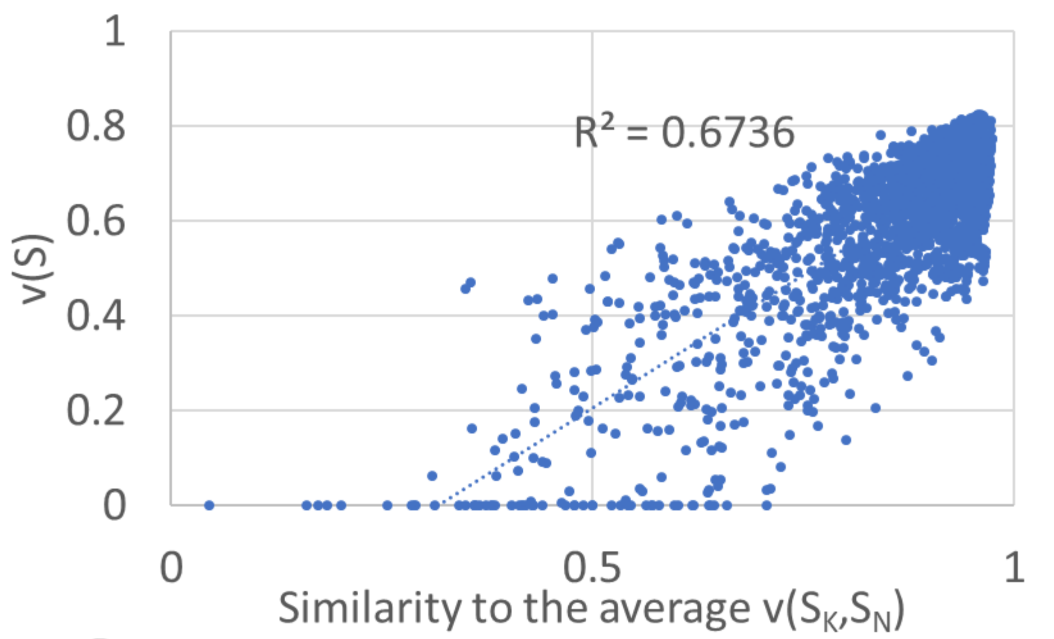
imdea networks · Data Transparency Group · UNIVERSIDAD POLITÉCNICA DE MADRID

… nor does it reflect the value of data from individual taxies at city level, that shows more correlation with the averageness of its data instead (R^2 = 0.67)

Shapley value vs. nº rides by driver at city level



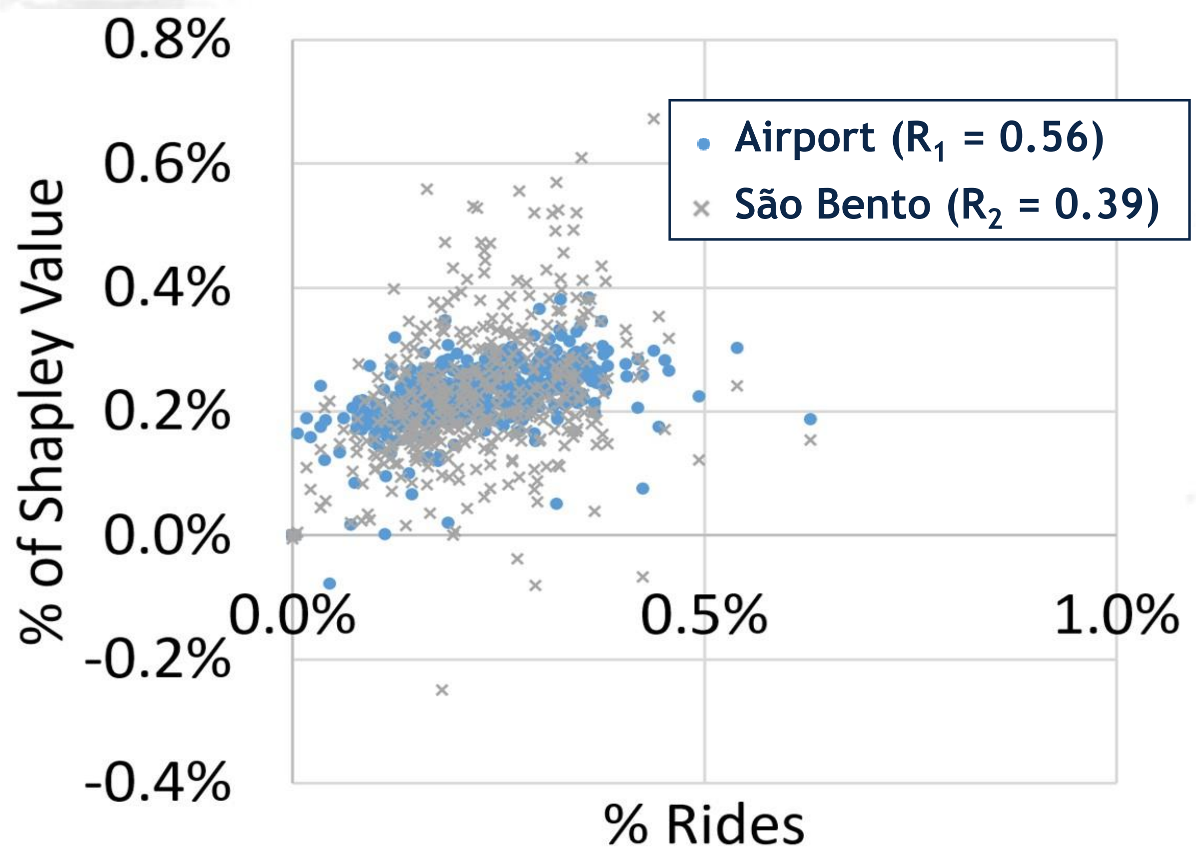
Shapley value vs. averageness at city level



S. Andrés Azcoitia, M. Paraschiv, and N. Laoutaris. Computing the Relative Value of Spatio-Temporal Data in Data Marketplaces. In Proc. of ACM SIGSPATIAL (2022)

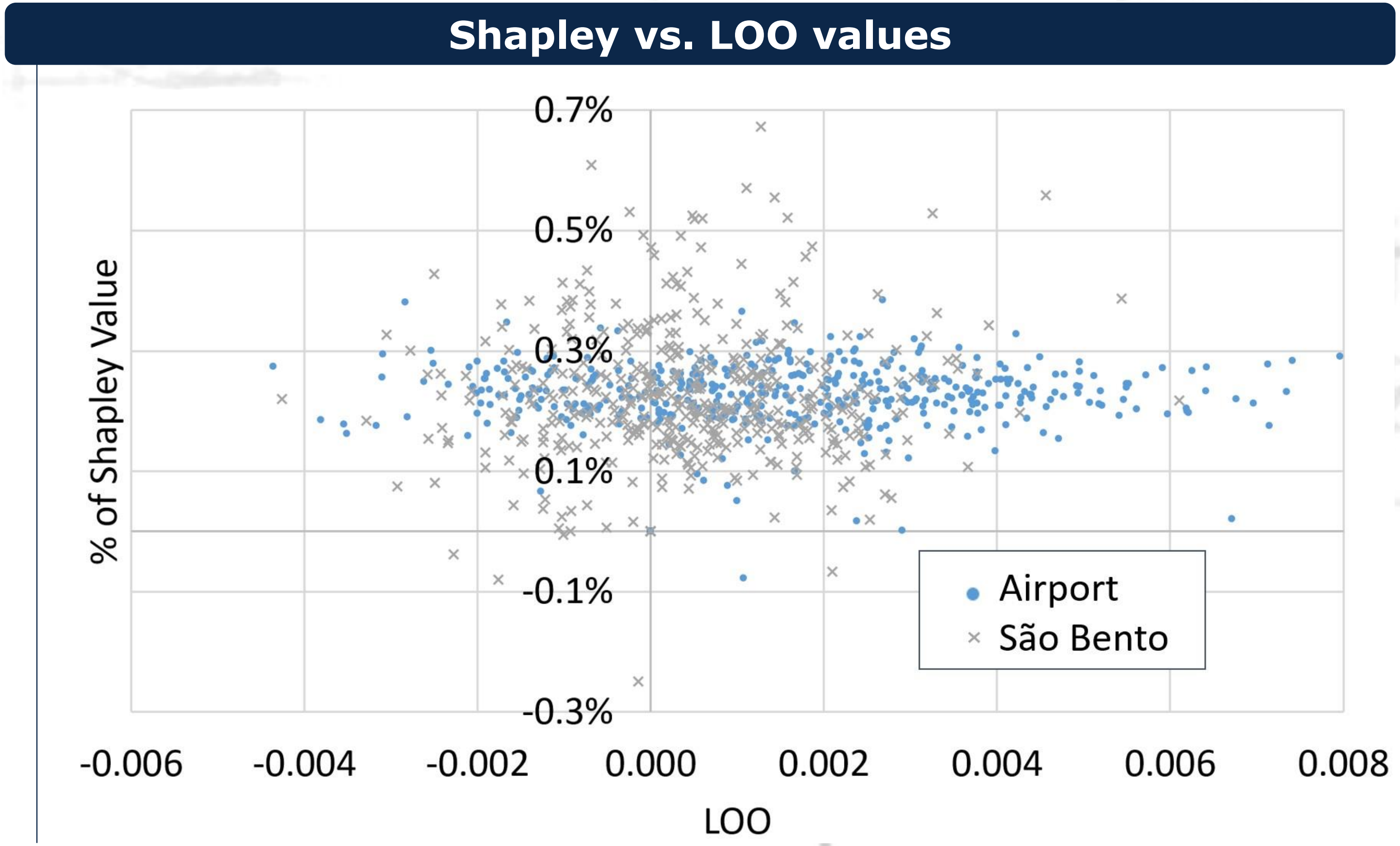
The Shapley value for estimating transportation time in Porto is different for each driver, and weakly correlated with the nº rides reported …

Shapley value vs. % rides reported by each taxi



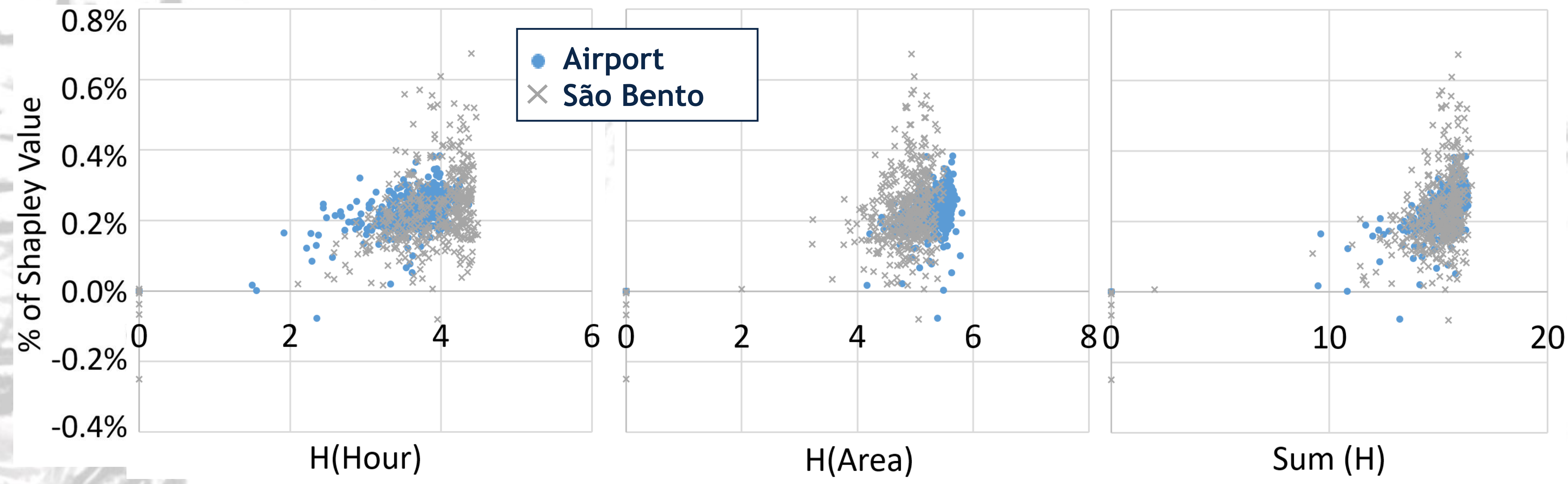
S. Andrés Azcoitia, M. Paraschiv, and N. Laoutaris. Computing the Relative Value of Spatio-Temporal Data in Data Marketplaces. In Proc. of ACM SIGSPATIAL (2022)

… or with their LOO-values.



Interestingly, the diversity of data reported, measured as Shannon's entropy (H) of key spatio-temporal features, showed a stronger correlation in this case

Pearson correlation of Shapley values with data features



$R_1 = 0.69$, $R_2 = 0.57$ $R_1 = 0.62$, $R_2 = 0.58$ $R_1 = 0.69$, $R_2 = 0.60$

imdea networks institute · Data Transparency Group · POLITÉCNICA · UNIVERSIDAD POLITÉCNICA DE MADRID

S. Andrés Azcoitia, M. Paraschiv, and N. Laoutaris. Computing the Relative Value of Spatio-Temporal Data in Data Marketplaces. In Proc. of ACM SIGSPATIAL (2022)

Understanding and Measuring the Data Economy

Addressing Technical Challenges

Regulating the data economy

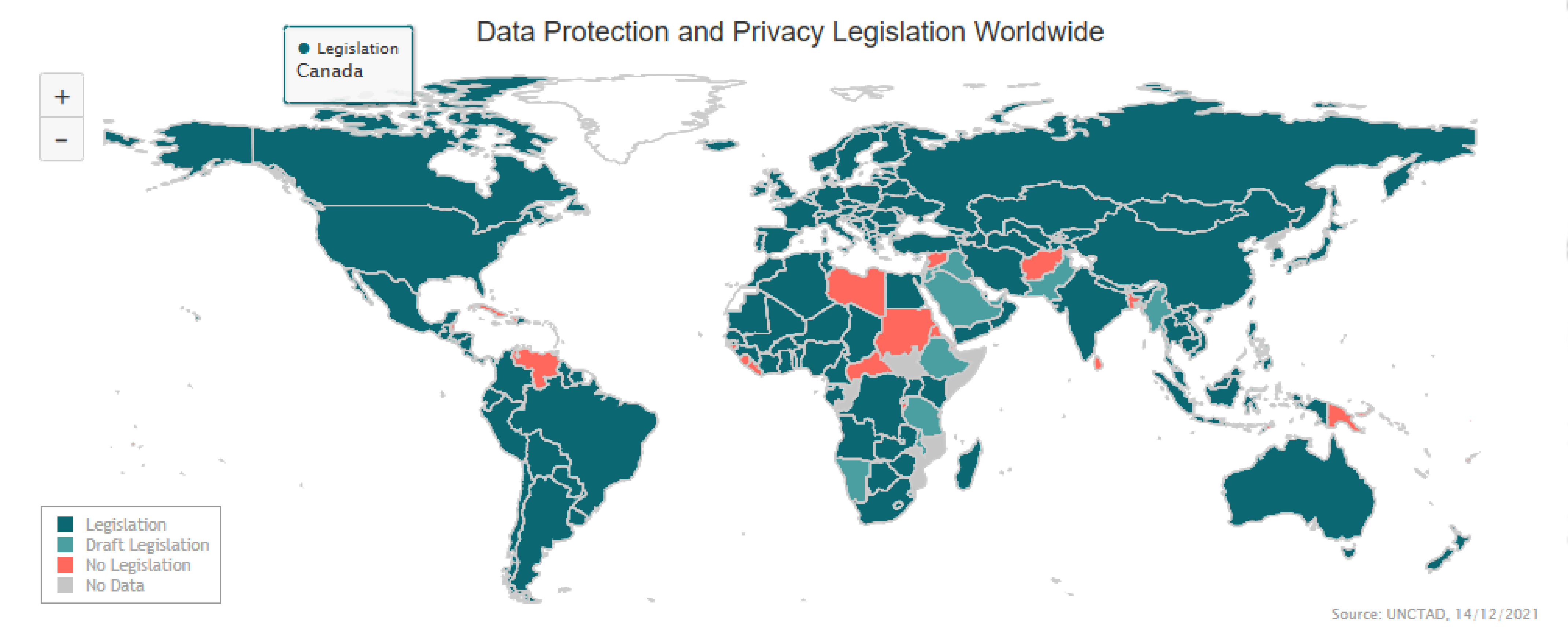
Data markets and data-related regulation respond to different strategies and objectives in the EU

	Shaping Europe's digital future	Data Strategy
Legal basis	Article 114 TFEU	Article 114 TFEU
Objectives	Protection of data subject/end users/business users' rights (fairness)	Reconcile economic goals in realising full potential of data
Targets of regulation	Big Tech – market power dynamics	Shift to other types of operators (alternative)
Form of obligations	Prescriptive and proscriptive	Alternative means – siloed-approach via limitations
Business models	Discontinuation of existing market power	New opportunities for new businesses

- Digital Markets Act
- Digital Services Act
- AI Act

- Data Governance Act
- Data Act

The EU is looking forward to pioneering the regulation of data markets and AI, as it happened with data protection & GDPR back in 2016



However, recent legislations will create frictions with the industry and will make enforcement very challenging

https://unctad.org/page/data-protection-and-privacy-legislation-worldwide

Some data marketplaces may "somehow" be tied to existing complementary platforms

Examples of **private data marketplaces** embedded in data-driven services or management systems:

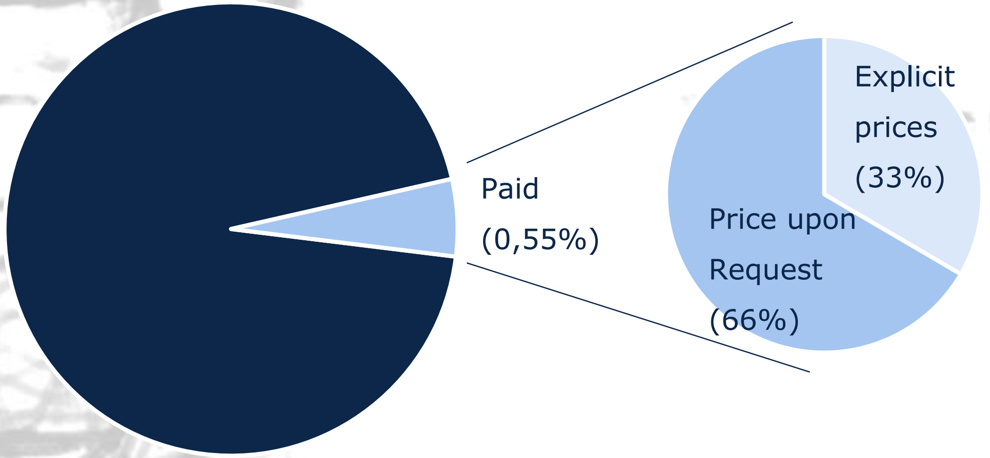
GIS	Marketing / Audience Data
ArcGIS Marketplace	LOTAME
CARTO	theTradeDesk
here	LiveRamp

General purpose DMs
aws marketplace

Key characteristics	
1	Accessible only by users of host systems / platforms
2	Data to be used primarily within the platform
3	Easier to bootstrap the DM targets already-existing platform users

How to comply with the principles of neutrality – 12(a) - and independency – 12(b) of DGA?

Data providers usually tailor prices (and products/services) to users

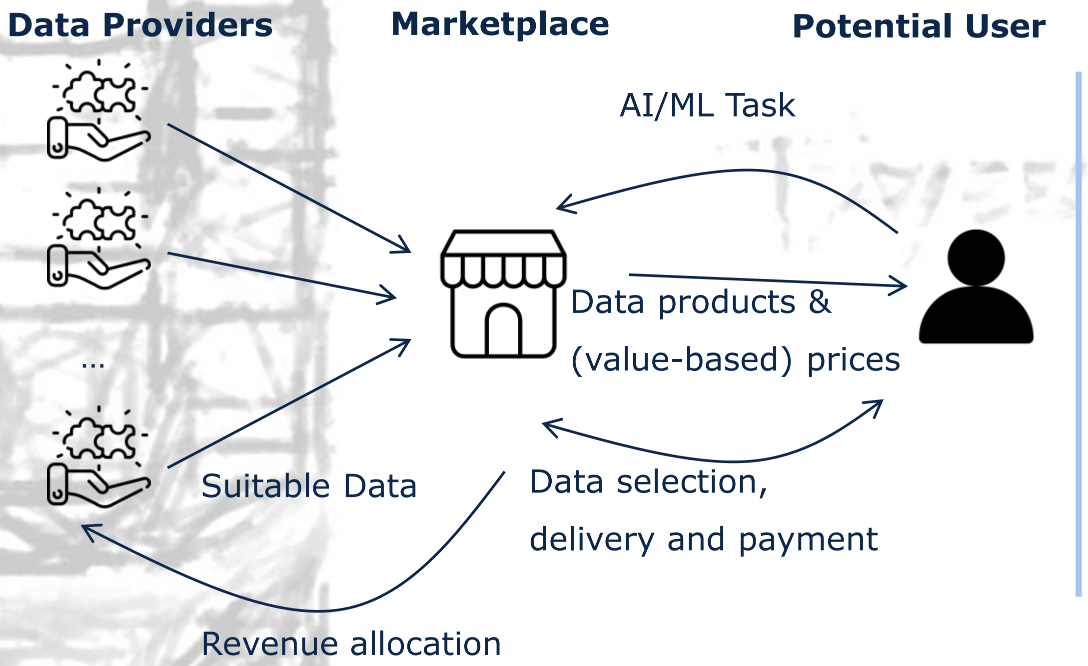


Key characteristics

1. DPs request information about potential users – identity, purpose of using the data
2. The price (and the product) are tailored to their needs
3. Potentially infinite data products can be around (versioning), with different prices and characteristics

How to ensure compliance with non-discrimination principles stated in DGA Art. 12(f)?

Model-based DM and federated leaning architectures tend to value (and price) data based on its contribution to a task



Key characteristics	
1	DMs or data holders are able to train or evaluate the utility of a dataset for a particular AI/ML task
2	Some studies propose to price data / rewards DPs based on the utility it brings to the task
3	DMs may deliver the data, whereas FL hides data and only delivers trained models

How to ensure compliance with non-discrimination principles stated in DGA Art. 12(f)?

Conclusion

Unlocking data silos by solving the challenges of data markets is key to realise the immense potential of AI in the economy, but will require work on:

1. Continuing to develop AI/ML use cases capable of delivering true value to the industry, to Governments, and to end users
2. Streamlining DMs by fighting against fragmentation and piracy, standardizing data sharing, setting knowledgeable price references, and improving the experience of buying and selling
3. Involving end users: protecting ownership and privacy, increasing trust, making the data economy explainable, and rewarding them fairly for their contributions
4. Increasing the information and transparency of data markets, and measuring the true value of data in the economy
5. Reshaping existing policies and regulations, and not only those related to data/AI (Data labor unions? Intellectual property? Robot-tax?)

Thanks for listening and participating!

Now Q&A time!

Santiago Andrés Azcoitia
santiago.azcoitia@imdea.org

References

Andrés Azcoitia, Santiago. Towards a Human-Centric Data Economy. PhD Thesis. UC3M.

Parts	Research Questions	Publications
Part I. Understanding and Measuring the Data Economy	*How are entities trading data doing business?*	*S. Andrés Azcoitia and N. Laoutaris, A Survey of Data Marketplaces and their Business Models. ACM SIGMOD Record Sept. 2022*
	How is data being traded in the market?	
	What kind of data products are being traded?	*S. Andrés Azcoitia, C. Iordanou, and N. Laoutaris. Measuring the Price of Data in Commercial Data Marketplaces. In Proc. of 1st ACM DE Workshop (2022)*
	What is the price of data products in commercial marketplaces?	
	Which features are driving the price of data in the market?	*S. Andrés Azcoitia, C. Iordanou, and N. Laoutaris. Understanding the Price of Data in Commercial Data Marketplaces. In Proc. of 39th IEEE ICDE (2023)*
Part II. Buying and Selling Data	*How can data consumers select suitable data for their tasks?*	*S. Andrés Azcoitia and N. Laoutaris. Try Before You Buy: A practical data purchasing algorithm for real-world data marketplaces. In Proc. of 1st ACM DE Workshop (2022)*
	What is the relative value of data from different individuals for A ML task?	*S. Andrés Azcoitia, M. Paraschiv, and N. Laoutaris. Computing the relative value of spatio-temporal data in data marketplaces. In Proc. of ACM SIGSPATIAL (2022)*
	How can we efficiently reward users based on the value of their data?	

References

- [Agarwal19] - A Marketplace for Data - An Algorithmic Solution
- [Aggarwal08] – Derandomization of auctions
- [Balazinska13] – A discussion on pricing relational data
- [ChanKim04] – Blue Ocean Strategy
- [Chawla19] – Revenue maximization for query pricing
- [Chen18] – Model-based pricing
- [Daskalakis17] - Strong duality for a multiple-good monopolist
- [Deep17] - QIRANA: a Framework for Scalable Query Pricing
- [Fernandez20] - Data Market Platforms: Trading Data Assets to Solve Data Problems
- [Ghosh11] – Selling privacy at auction
- [Goldberg01] – Competitive auctions and digital goods
- [Goldberg03] – Competitiveness via consensus
- [Goldberg08] – Competitive auctions
- [

References

- [Goldfarb19] - Digital Economics
- [Haghpanah20] – When is pure bundling optimal?
- [Kang19] - Incentive Mechanism for Reliable Federated Learning: A Joint Optimization Approach to
- Combining Reputation and Contract Theory
- [Koutris12] – Query-based data pricing
- [Koutris12_2] – Query market demostration
- [Li15] – A theory of Pricing Private Data
- [Liang18] – A survey of Big Data Market: Pricing, Trading and Protection
- [Lin14] - On arbitrage-free pricing for general data queries
- [Moor19] - Data Markets with Dynamic Arrival of Buyers and Sellers
- [Muschalle13] – Pricing approaches for data markets
- [Niu19] - Online pricing with reserve price constraint for personal data markets
- [Noy19] - Google Dataset Search: Building a search engine for datasets in an openWeb ecosystem

References

- [Ohrimenko19] – Collaborative Machine Learning Markets with Data-Replication Robust Payments
- [Ostrom90] - Governing the commons the evolution of institutions for collective action
- [Pantelis13] – Undestanding the value of Big Data
- [Pei20] - A Survey on Data Pricing: from Economics to Data Science
- [Shapiro98] - Information Rules - A Strategic Guide to the Network Economy
- [Yan20] – If you like Shapley you'll love the core
- [Yu17] – Data pricing strategy based on data quality
- [Wu10] – Cloud pricing models: Taxonomy, survey and interdisciplinary challenges
- [Wu10_2] – Best pricing strategy for information services
- [Zheng17] - An Online Pricing Mechanism for Mobile Crowdsensing Data Markets