# TOWARDS A HUMAN-CENTRIC DATA ECONOMY

by

SANTIAGO ANDRÉS AZCOITIA

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in

Ph.D. Program in Telematics Engineering

Universidad Carlos III de Madrid

Advisor:

Nikolaos Laoutaris

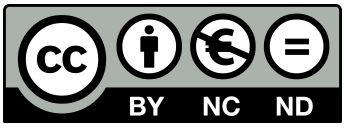
Tutor:

Nikolaos Laoutaris

April 2023

This work has been supported by IMDEA Networks Institute.



# Acknowledgements

Enrolling in a Computer Science PhD means riding a roller coaster of hard work, disappointment and joy, of illusion and delusion, of designing, coding, analysing, writing, presenting... Doing it after more than 20 years' experience in the industry may be described by many people as a pure act of insanity. Starting it three months before the COVID-19 lockdown, a hard test of mental stability and tenacity. But no human or artificial intelligence was able to predict that.

Under these circumstances, I successfully managed to publish all my work in great conferences and journals. And here I am ending my dissertation with these grateful opening words.

It is probably my closest family that has the clearest idea about the amount of work and energy behind this dissertation. It is Irene, my wife, and my children, Lucía, Jaime and Javier, that have suffered my sudden mood swings as reviews, rejections and acceptances have followed one after the other over the last three and a half years. So big thanks to all of you for being there and for not kicking a sometimes nasty PhD candidate out of your life!

I would like to thank Nikos Laoutaris, my supervisor, for coming up with this crazy PhD, for betting for a seasoned consultant like me to execute it, and for his advice, mind-blowing brainstorming sessions and all the support he provided during the process. Thanks also to my co-authors Costas Iordanou and Marius Paraschiv for their help in writing our joint papers and their valuable advice. And thanks to my teammates in the Data Transparency Group, and to the rest of the crew at IMDEA for bearing with me the last years.

Some other people have indirectly helped me with this PhD. On the personal side, my family (parents and sisters, blood and -in-law) and beloved lifelong friends for supporting cutting-edge research with love, beer, talks, walks, barbecues, wine and laughs. Intangible yet invaluable tools to stabilise the mind and accelerate the production of code and revolutionary ideas. Many thanks to you all, always.

On the professional side, thanks to my ex-colleagues in Axon Consulting, Deloitte and Telefónica, with some of whom I still keep close contact. Together with my teachers and professors in the far old college and school years, they have shaped me and taught me almost everything I used to succeed in this ambitious undertaking.

Although it is only one person that signs a doctoral dissertation, many more people behind the scenes have contributed to that piece of research. That is why hereinafter I intentionally avoid using the first person singular.

# Published and submitted content

During the course of this PhD, the author submitted, co-authored and published a number of papers that are included in this thesis. The following list summarises those publications, where and to what extent they are included in this thesis, and the role of the author in them:

**(1) Santiago Andrés Azcoitia** and Nikolaos Laoutaris, A Survey of Data Marketplaces and their Business Models. *ACM SIGMOD Record, 51(3), (Sep 2022)*, ACM, New York, NY, USA.
URL: https://sigmodrecord.org/publications/sigmodRecord/2209/pdfs/04_Surveys_Azcoitia.pdf.
(pre-print) **(Santiago Andrés Azcoitia** and Nikolaos Laoutaris. A Survey of Data Marketplaces and their Business Models. (Jan 2022). ArXiv.
URL: https://doi.org/10.48550/arXiv.2201.04561.

- This work is fully included in this thesis in chapters 1, 3, and 8.
- The material from this source included in this thesis is not singled out with typographic means and references.
- The author helped in designing the methodology, located target data marketplaces, carried out the survey, and developed all the materials in this paper.

**(2) Santiago Andrés Azcoitia**, Costas Iordanou, and Nikolaos Laoutaris. Measuring the Price of Data in Commercial Data Marketplaces. *First ACM Data Economy Workshop in CoNEXT'22 (2022)*. Association for Computing Machinery, New York, NY, USA.
URL: https://doi.org/10.1145/3565011.3569053

- This work is fully included in this thesis in chapters 4 and 8
- The material from this source included in this thesis is not singled out with typographic means and references.
- The author helped in designing the methodology, developed data marketplace crawlers and parsers, and contributed to writing the paper.

**(3) Santiago Andrés Azcoitia**, Costas Iordanou, and Nikolaos Laoutaris. Understanding the Price of Data in Commercial Data Marketplaces. Accepted for publication in the 39th IEEE *International Conference on Data Engineering (ICDE 2023)*.

(pre-print) **Santiago Andrés Azcoitia**, Costas Iordanou, and Nikolaos Laoutaris. What Is the Price of Data? A Measurement Study of Commercial Data Marketplaces. (Oct 2021). ArXiv. URL: https://doi.org/10.48550/arXiv.2111.04427

(Poster) **Santiago Andrés Azcoitia**, Costas Iordanou, and Nikolaos Laoutaris. Poster: What Is the Price of Data? A Measurement Study of Commercial Data Marketplaces. *ACM Internet Measurement Conference (2021)*. ACM, New York, NY, USA. URL: https://conferences.sigcomm.org/imc/2021/pdf/8.pdf

- This work is fully included in this thesis in chapter 4
- The material from this source included in this thesis is not singled out with typographic means and references.
- The author helped in designing the methodology, developed data marketplace crawlers and parsers, implemented classifiers and regression models, analysed and interpreted the data, and contributed to writing the paper.
- The author prepared a poster and presented it in the corresponding session of IMC'21.

**(4) Santiago Andrés Azcoitia** and Nikolaos Laoutaris. Try Before You Buy: A practical data purchasing algorithm for real-world data marketplaces. *First ACM Data Economy Workshop in CoNEXT'22 (2022)*. Association for Computing Machinery, New York, NY, USA. URL: https://doi.org/10.1145/3565011.3569054

(pre-print) **Santiago Andrés Azcoitia** and Nikolaos Laoutaris. Try Before You Buy: A practical data purchasing algorithm for real-world data marketplaces. (Dec 2021). ArXiv. URL: https://doi.org/10.48550/arXiv.2012.08874

- This work is fully included in this thesis in Chapters 5 and 6.
- The material from this source included in this thesis is not singled out with typographic means and references.
- The author helped in designing the methodology, developed data marketplace simulators and implemented the corresponding purchasing algorithms, carried out the different tests, analysed and interpreted the data, and contributed to writing the paper.

**(5) Santiago Andrés Azcoitia**, Marius Paraschiv, and Nikolaos Laoutaris. Computing the relative value of spatio-temporal data in data marketplaces. *In Proceedings of the 30th International Conference on Advances in Geographic Information Systems (SIGSPATIAL 2022)* Association for Computing Machinery, New York, NY, USA, Article 23, 1–11.
URL: https://doi.org/10.1145/3557915.3561470

(pre-print) **Santiago Andrés Azcoitia**, Marius Paraschiv, and Nikolaos Laoutaris. Computing the relative value of spatio-temporal data in wholesale and retail data marketplaces. (Feb 2020). ArXiv.
URL: https://doi.org/10.48550/arXiv.2002.11193

- This work is fully included in this thesis in Chapters 5 and 7.
- The material from this source included in this thesis is not singled out with typographic means and references.
- The author helped in developing and testing different calculation and approximation algorithms to the Shapley value, tested and analysed them, interpreted their results, and contributed to writing the paper.

# Other Research Merits

Other published works and public activity related to this PhD include:

(6) Nikhil Jha, Martino Trevisan, Luca Vassio, Marco Mellia, Stefano Traverso, Alvaro Garcia-Recuero, Nikolaos Laoutaris, Amir Mehrjoo, **Santiago Andrés Azcoitia**, Ruben Cuevas Rumin, Kleomenis Katevas, Panagiotis Papadopoulos, Nicolas Kourtellis, Roberto Gonzalez, Xavi Olivares, George-Marios Kalatzantonakis-Jullien, A PIMS Development Kit for New Personal Data Platforms, in IEEE Internet Computing, vol. 26, no. 3, pp. 79-84, 1 (May-June 2022).

- The author helped with the survey of PIMS and the characterisation of their business models, and wrote the corresponding section of this paper.

(7) (Talk) **Santiago Andrés Azcoitia**. What's the price of data? A Measurement Study of Commercial Data Marketplaces. *IMDEA Networks Institute (Jun 2021)*

- The author prepared the material and gave the talk at IMDEA's auditorium.

(8) (Talk) **Santiago Andrés Azcoitia**. Towards a human-centric data economy. *Universidad Politécnica de Madrid, TryIT Conference (Mar 2022).*

- The author prepared the material and gave the talk.

(9) (Poster) **Santiago Andrés Azcoitia**, Costas Iordanou, and Nikolaos Laoutaris. Poster: What Is the Price of Data? *12th IMDEA Networks Annual International Workshop* (Jun 2022).

- The author prepared the poster and presented it in the workshop.

(10) (Talk) **Santiago Andrés Azcoitia**. Towards a human-centric data economy. *IMDEA Networks Institute (Jun 2022).*

- The author prepared the material and gave the talk at IMDEA's auditorium.

# Abstract

Spurred by widespread adoption of artificial intelligence and machine learning, "data" is becoming a key production factor, comparable in importance to capital, land, or labour in an increasingly digital economy. In spite of an ever-growing demand for third-party data in the B2B market, firms are generally reluctant to share their information. This is due to the unique characteristics of "data" as an economic good (a freely replicable, non-depletable asset holding a highly combinatorial and context-specific value), which moves digital companies to hoard and protect their "valuable" data assets, and to integrate across the whole value chain seeking to monopolise the provision of innovative services built upon them. As a result, most of those valuable assets still remain unexploited in corporate *silos* nowadays.

This situation is shaping the so-called data economy around a number of *champions*, and it is hampering the benefits of a global data exchange on a large scale. Some analysts have estimated the potential value of the data economy in US$2.5 trillion globally by 2025. Not surprisingly, unlocking the value of data has become a central policy of the European Union, which also estimated the size of the data economy in 827€ billion for the EU27 in the same period. Within the scope of the *European Data Strategy*, the European Commission is also steering relevant initiatives aimed to identify relevant cross-industry use cases involving different verticals, and to enable sovereign data exchanges to realise them.

Among individuals, the massive collection and exploitation of personal data by digital firms in exchange of services, often with little or no consent, has raised a general concern about privacy and data protection. Apart from spurring recent legislative developments in this direction, this concern has raised some voices warning against the unsustainability of the existing digital economics (few digital champions, potential negative impact on employment, growing inequality), some of which propose that people are paid for their data in a sort of worldwide data labour market as a potential solution to this dilemma [114, 115, 155].

From a technical perspective, we are far from having the required technology and algorithms that will enable such a human-centric data economy. Even its scope is still blurry, and the question about the value of data, at least, controversial. Research works from different disciplines have studied the data value chain, different approaches to the value of data, how to price data assets, and novel data marketplace designs. At the same time, complex legal and ethical issues with respect to the data economy have risen around privacy, data protection, and ethical AI practices.

In this dissertation, we start by exploring the data value chain and how entities trade data assets over the Internet. We carry out what is, to the best of our understanding, *the most thorough survey of commercial data marketplaces*. In this work, we have catalogued and characterised ten different business models, including those of personal information management systems, companies born in the wake of recent data protection regulations and aiming at empowering end users to take control of their data. We have also identified the challenges faced by different types of entities, and what kind of solutions and technology they are using to provide their services.

Then we present *a first of its kind measurement study that sheds light on the prices of data in the market* using a novel methodology. We study how ten commercial data marketplaces categorise and classify data assets, and which categories of data command higher prices. We also develop classifiers for comparing data products across different marketplaces, and we study the characteristics of the most valuable data assets and the features that specific vendors use to set the price of their data products. Based on this information and adding data products offered by other 33 data providers, we develop a regression analysis for revealing features that correlate with prices of data products. As a result, we also implement the basic building blocks of a novel data pricing tool capable of providing a hint of the market price of a new data product using as inputs just its metadata. This tool would provide more transparency on the prices of data products in the market, which will help in pricing data assets and in avoiding the inherent price fluctuation of nascent markets.

Next we turn to topics related to data marketplace design. Particularly, we study how buyers can select and purchase suitable data for their tasks without requiring *a priori* access to such data in order to make a purchase decision, and how marketplaces can distribute payoffs for a data transaction combining data of different sources among the corresponding providers, be they individuals or firms. The difficulty of both problems is further exacerbated in a human-centric data economy where buyers have to choose among data of thousands of individuals, and where marketplaces have to distribute payoffs to thousands of people contributing personal data to a specific transaction.

Regarding the selection process, we compare different purchase strategies depending on the level of information available to data buyers at the time of making decisions. *A first methodological contribution of our work is proposing a data evaluation stage prior to datasets being selected and purchased by buyers in a marketplace*. We show that buyers can significantly improve the performance of the purchasing process just by being provided with a measurement of the performance of their models when trained by the marketplace with *individual* eligible datasets. *We design purchase strategies that exploit such functionality and we call the resulting algorithm Try Before You Buy*, and our work demonstrates over synthetic and real datasets that it can lead to near-optimal data purchasing with only $O(N)$ instead of the exponential execution time - $O(2^N)$ - needed to calculate the optimal purchase.

With regards to the payoff distribution problem, we focus on computing the relative value of spatio-temporal datasets combined in marketplaces for predicting transportation demand and travel time in metropolitan areas. Using large datasets of taxi rides from Chicago, Porto and New York *we show that the value of data is different for each individual, and cannot be approximated by its volume. Our results reveal that even more complex approaches based on the "leave-one-out" value, are inaccurate. Instead, more complex and acknowledged notions of value from economics and game theory, such as the Shapley value, need to be employed if one wishes to capture the complex effects of mixing different datasets on the accuracy of forecasting algorithms.* However, the Shapley value entails serious computational challenges. Its exact calculation requires repetitively training and evaluating every combination of data sources and hence $O(N!)$ or $O(2^N)$ computational time, which is unfeasible for complex models or thousands of individuals. Moreover, our work paves the way to new methods of measuring the value of spatio-temporal data. *We identify heuristics such as entropy or similarity to the average that show a significant correlation with the Shapley value and therefore can be used to overcome the significant computational challenges posed by Shapley approximation algorithms in this specific context.*

We conclude with a number of open issues and propose further research directions that leverage the contributions and findings of this dissertation. These include monitoring data transactions to better measure data markets, and complementing market data with actual transaction prices to build a more accurate data pricing tool. A human-centric data economy would also require that the contributions of thousands of individuals to machine learning tasks are calculated daily. For that to be feasible, we need to further optimise the efficiency of data purchasing and payoff calculation processes in data marketplaces. In that direction, we also point to some alternatives to repetitively training and evaluating a model to select data based on *Try Before You Buy* and approximate the Shapley value. Finally, we discuss the challenges and potential technologies that help with building a federation of standardised data marketplaces.

The data economy will develop fast in the upcoming years, and researchers from different disciplines will work together to unlock the value of data and make the most out of it. Maybe the proposal of getting paid for our data and our contribution to the data economy finally flies, or maybe it is other proposals such as the robot tax that are finally used to balance the power between individuals and tech firms in the digital economy. Still, we hope our work sheds light on the value of data, and contributes to making the price of data more transparent and, eventually, to moving towards a human-centric data economy.

# Glossary

| Acronym | Meaning |
|---|---|
| **AAPE** | Average Average Percentage Error |
| **AASTD** | Average Average Standard Deviation |
| **ACM** | Association for Computing Machinery |
| **AI** | Artificial Intelligence |
| **AMO** | Refers to AMO market and cryptocurrency [66] |
| **API** | Application Programming Interface |
| **A-TBYB** | Assisted Try-Before-You-Buy |
| **AWS** | Amazon Web Services |
| **B2B** | Business to Business |
| **BPMN** | Business Process Model and Notation [144] |
| **CAM** | Continuous Automated Monitoring (in Gaia-X) |
| **CCPA** | California Consumer Privacy Act |
| **CDF** | Cumulative Distribution Function |
| **CEO** | Chief Executive Officer |
| **CORE** | Consensus Revenue Estimate |
| **CosSim** | Cosine Similarity |
| **COVID** | COronaVIrus Disease |
| **CSV** | Comma Separated Values |
| **DE** | Data Economy |
| **DI** | Data Interchangeability |
| **DIH** | Data Intelligence Hub |
| **DIN** | Standard drawn up at the German Institute for Standardisation (DIN) |
| **DLT** | Distributed Ledger Technologies |
| **DM** | Data Marketplace |
| **DME** | Data Marketplace Enabler |
| **DMP** | Data Management Platform |
| **DNN** | Deep Neural Network |
| **DP** | Data Provider |
| **DTE** | Data Trading Entities |

| Acronym | Meaning |
|---|---|
| **DTW** | Dynamic Time Warp |
| **ESRI** | Environmental Systems Research Institute, Inc. |
| **EU** | European Union |
| **FAIR** | Findable, Accessible, Interoperable, and Reusable |
| **FL** | Federated (Machine) Learning |
| **GBR** | Gradient Boosting Regressor |
| **GDPR** | General Data Protection Regulation |
| **GIS** | Geographic Information System |
| **HCDE** | Human-Centric Data Economy |
| **HERE** | Refers to HERE Technologies [177] |
| **IaaS** | Infrastructure-as-a-Service |
| **IDC** | International Data Corporation |
| **IDS** | International Data Spaces |
| **IDSA** | International Data Spaces Association |
| **IEC** | International Engineering Consortium |
| **IMDEA** | Instituto Madrileño de Estudios Avanzados |
| **IoT** | Internet of Things |
| **IOTA** | Internet of Things Application |
| **JSON** | JavaScript Object Notation |
| **JSON-LD** | JavaScript Object Notation for Linked Data |
| **kNN** | k-Nearest Neighbours algorithm |
| **LOO** | Leave-One-Out |
| **M2M** | Machine to Machine |
| **MAE** | Mean Absolute Error |
| **MC** | Monte Carlo |
| **ML** | Machine Learning |
| **MM** | Millions |
| **MRE** | Mean Relative Error |
| **MSE** | Mean Squared Error |
| **MUP** | Marginal Utility Profile |
| **NB** | Naive Bayes |
| **NFT** | Non-Fungible Token |
| **NGO** | Non-Governmental Organisation |
| **NLP** | Natural Language Processing |
| **NumSim** | Numeric Similarity |
| **NYC** | New York City |
| **N/A** | Not available |
| **OECD** | The Organisation for Economic Cooperation and Development |

| Acronym | Meaning |
|---|---|
| **PaaS** | Platform-as-a-Service |
| **PDK** | Platform Development Kit |
| **PET** | Privacy-Enhancing Technology |
| **PIMS** | Personal Information Management System |
| **PMP** | Private Marketplaces |
| **PoI** | Point of Interest |
| **QoS** | Quality of Service |
| **RAM** | Reference Architecture Model (of IDS standard) |
| **RDF** | Resource Description Framework |
| **RDTW** | Relative Dynamic Time Warp |
| **RF** | Random Forest algorithm |
| **RMSE** | Relative Mean Squared Error |
| **RS** | Random Sampling |
| **SARIMA** | Seasonal Auto-Regressive Integrated Moving Average |
| **SDL** | Secure Development Language |
| **SEC** | American Securities and Exchange Commission |
| **SP** | Service Provider |
| **SS** | Structured Sampling |
| **S-TBYB** | Standalone Try-Before-You-Buy |
| **STD** | Standard Deviation |
| **SV** | Shapley Value |
| **TBYB** | Try-Before-You-Buy |
| **TCOD** | Total Cost of Data |
| **TEE** | Trusted Execution Environment |
| **TLS** | Transport Layer Security |
| **TMC** | Truncated Monte Carlo |
| **TRS** | Truncated Random Sampling |
| **TSS** | Truncated Structured Sampling |
| **UML** | Unified Modelling Language |
| **UNCTAD** | United Nations Conference on Trade And Development |
| **URL** | Uniform Resource Locator |
| **US** | United States (of America) |
| **USA** | United States of America |
| **UTM** | Universal Travel Mercator |
| **XLS** | Microsoft Excel spreadsheet |
| **XLSX** | Microsoft Excel spreadsheet files |

# Table of Contents

# List of Tables

# List of Figures

# Part I

# Background

# Chapter 1

# Introduction

Paying for information is not a new idea: insiders have been hired and spies have been trained to achieve a competitive advantage while doing business or fighting wars since ancient times. Such primitive information exchanges exclusively involved humans, yet they were often decisive and undeniably influenced the course of history (e.g., Ephialtes' betrayal proved decisive in the Battle of Thermopilae).

With the advent of telecommunications, information was no longer transmitted by people but by electromagnetic signals, and the exchange of information became almost instantaneous. Later still, computing, electronics and digital communications gave birth to a new generation of sensors and increasingly automated data collection. However, the majority of information flows and services in the web 2.0 are intended to be consumed by humans.

An even more revolutionary twist will likely drive the future growth of the so-called knowledge economy thanks to the internet of things (IoT), artificial intelligence (AI), and ubiquitous communication systems such as 5G. According to IDC, 30% of data will be generated by sensors in real time by 2025 [159]. In the current context of the major digitalisation of the economy, a myriad of applications and data-hungry machine learning (ML) models are - to give a couple of meaningful examples - helping companies and public institutions improve their efficiency, as well as assisting individuals in health issues. This means that machines may replace humans as the main data consumers. In some settings, such M2M data exchanges will be required to happen in real time, too.

In the first generation of AI/ML models based on statistical or machine learning, they were the programmers that introduced the intelligence in the code of expert systems. The use of data was limited and there was no inferred knowledge base. On the contrary, second-generation AI/ML models allowed for statistical learning. It is the model itself that learns from training data and adapts to produce outputs according to a certain objective. Hence, the quality of data for training such models is key, and data has consequently become a very relevant production factor, comparable in importance to land, capital, labour, or infrastructure.

As digitalisation progresses, machines are increasingly playing a leading role in the data value chain, starting from its collection through probes and sensors and ending in digital services provided to end users that leverage ML models that "*consume*" such data after one or more intermediate processing steps. AI and ML are increasing the demand for data and the availability of more quality data is enabling new models and use cases in a virtuous circle. In fact, the global amount of data created annually is expected to grow by 530% from 2018 to 2025 [159]. This influx of new and existing data is thus accelerating the data economy, and data is undoubtedly becoming a cornerstone of modern economic systems.

Estimating the value of data and setting the price of a dataset become harder tasks due to the elusive nature of the traded "*commodity*". Unlike oil, to which it is often compared [30], data can be copied, transmitted, and processed with close to zero cost. This follows directly from its digital nature. Moreover, its value depends heavily on the context and the purpose for which it is used, and can be very different for each "*consumer*". Some authors compared data to labour, too [11]. However, unlike labour, data is a non-depletable and non-rivalrous good meaning that its supply is not affected by its consumption.

The structure of the data economy is heavily influenced by the unique characteristics of "*data*" as an economic good [97, 164]. Data is massively collected from users in exchange of services over the Internet, and it often remains stuck unexploited in *silos*. Companies in the data economy span across the whole value chain, from the very first collection of data to the provision of final services to end users, in order to protect their data assets from being exposed to third parties at the risk of losing a competitive advantage. As a result, the data economy is led by a handful horizontally integrated global oligopolies (Apple, Microsoft, Amazon, Alphabet, Facebook, Alibaba, or Tencent), it shows a blatant geographical imbalance, with two countries championing digital services (US and China), and it inherits a still huge digital divide between developed and developing countries that also extends to the creation of value from data [145].

From the perspective of individuals, personal data is increasingly being collected by a myriad of digital service providers for different purposes, and it is expected to play a crucial role in the ongoing digitalisation process within the so-called fourth industrial revolution. This situation is raising a growing concern about online privacy. There are still significant economic and technical challenges at the time of protecting privacy of end users, tracking the use of personal data and calculating a fair value of data in the market. In spite of new legislation enforcing that users grant "*data processors*" their consent to use their personal data, they are rarely offered any reward other than digital services for such usage. At the same time, some companies are making huge benefits by sharing personal anonymised data, inferred or built after processing personal data of a mass of individuals, and usually by providing (often online) services that leverage such information.

From an economic perspective, Jaron Lanier questions whether the current situation would be optimal or even sustainable in the long term in "Who Owns the Future?" [114], and he proposes that digital companies pay people for their data according to its "utility". Weyl and Posner also introduce a radical market around paying individuals for their data. Not only would this create a

new income source for families, but it would also change the role of individuals in the economy radically, from passive data subjects to active *prosumers* [155]. Some voices are claiming against data privacy violations and free collaboration with the industry, which might eventually challenge IoT development and adoption (note that IoT heavily relies on automatically processing data collected through sensors, often capturing personal data attributable to end users). Furthermore, they request the empowerment of users to track the use of shared personal data. Getting a fair reward for the data they generate to digital service providers would also stop companies abusing indiscriminate data collection over the Internet. Even though it may seem that users are just trying to get part of the cake, it is widely accepted that a fair reward would eventually improve the quality and availability of such data: as the willingness to collaborate with the industry grows, so do available resources to train AI models and to maximise the outcome of massive digitalisation.

Aided by recent legislative developments in data protection, including the General Data Protection Regulation in the EU or the California Consumer Privacy Act [140, 182], *Personal Information Management Systems (PIMS)* have appeared with the purpose of empowering individuals to take back control of their PI currently being collected by Internet service providers with little or no consent. They let users collect their personal information from Internet data and service providers; exercise erasure and modification rights over this data; manage cookie, privacy, and access permissions settings of their devices in a friendly way; and grant or withdraw their consent for the platform to share their personal data with third parties for different purposes [94].

In the business to business (B2B) market, entities with different business models are responding to the demand of data and looking at ways to operationalise data trading. First-generation *general-purpose DMs* are being complemented by *niche DMs* that target specific industries (e.g., Caruso for the connected car, Veracity for energy and transportation), and cover data sourcing for specific innovative purposes, such as feeding ML algorithms (e.g., Mechanical Turk, DefinedCrowd), or trading IoT real-time sensor data (e.g., IOTA, Terbine). Additionally, some leading *data-management systems* (e.g., Snowflake, Cognite) and *niche* digital solutions (e.g., Carto, Openprise, LiveRamp) are integrating secure data exchange features and capabilities to their existing products with the aim of breaking data silos [64]. Some PIMS have also implemented marketplaces for helping users monetise their personal data [169].

Still, entities trading data over the internet face lots of challenges related to protecting ownership, fighting piracy and theft, pricing of data, assessing suitable data for their purpose, fairly distributing payoffs to data owners and across the chain, etc. Moreover, the value of personal data is far from clear. The market capitalisation of data-driven market champions was US$9,271 billions[1] in the first quarter of 2022. Dividing it by the global population roughly gives US$1,000 per individual. This assumes that they do business out of data only, whereas they sell devices or software, as well. Weyl and Posner dare estimate the transfer of a 9% of the data economy from data-driven companies to data owners thanks to their radical market of data as labour, meaning US$20k for a family of four, while increasing the overall size of the economy by 3%.

[1] See https://en.wikipedia.org/wiki/Big_Tech

Even though most studies agree that AI and data will have a heavy overall impact on the economy, they publish different estimates. For example, the data economy was estimated to reach US$2.5 trillion globally by 2025 [87], whereas a recent market study within the scope of the European Data Strategy estimates a size of up to 827€ billion for the EU27 [37]. A more recent study estimated the potential of AI to deliver additional global economic activity of around US$13 trillion by 2030 [28]. Measuring the data economy has been appointed also as a relevant challenge in the field of economic [12].

There is also a significant multiplier, $\times 5$ to $\times 10$ depending on the study, between the size of data markets and the impact of data on the economy. According to the EU [62], the size of data markets, meaning "*the marketplace where digital data is exchanged as 'products' or 'services' as a result of the elaboration of raw data*" was 80 billion € in 2020, making an impact of 444 € billion. According to UNCTAD, the size of data markets in Japan by the same year was 40 € billion, and 220 € billion in the US [145].

Unlocking the value of data is central to the European Strategy for Data [186], aimed at creating a digital single market around data in the EU that allows to unlock the enormous potential and opportunities of data for European citizens. The digital single market for data must ensure data sovereignty by individuals and companies, while it contributes to improving the digital competitiveness of Europe. As part of the strategy, the European Commission is adopting legislative measures [182, 183, 185, 189], and fostering other policy initiatives to promote open data, to enable the reuse of public information [184], and to issue guidelines on data sharing [188], among others. Finally, the EU is fostering a couple of standardisation initiatives closely related to this dissertation, namely the International Data Spaces (IDS) standard, and the Gaia-X project.

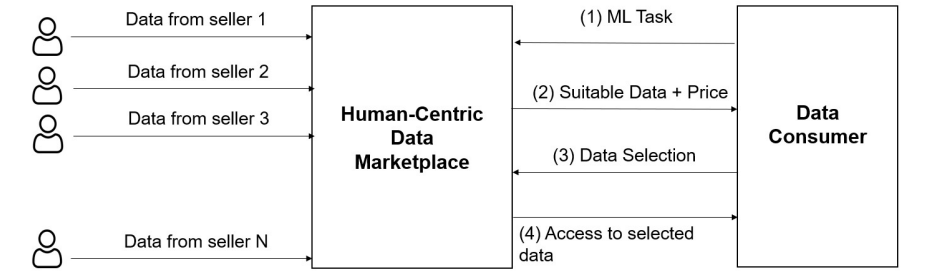
## 1.1. Objectives

Within this context, IMDEA's initiative towards a Human-Centric Data Economy (HCDE) believes in a fair and functioning data economy in which data is controlled and used fairly and ethically in a human-oriented manner. HCDE mission is to empower individuals by improving their right to self-determination regarding their personal data. Our mission inside HCDE is to enable users to understand the value of their contribution to the data economy and the usage of their personal data that players of this ecosystem do. A HCDE must track the activity of individuals in the data ecosystem and eventually derive fair algorithms to incentivise their contributions just as wages reward work done by workers.

From a technical perspective, we are far from reaching the required technology and algorithms that will enable the business and economic models we are conceiving as fairest and most sustainable in the long term. However, some steps can already be taken to better understand the value of data and how it is traded in the market.

This dissertation addresses, among others, the following research questions in part II:

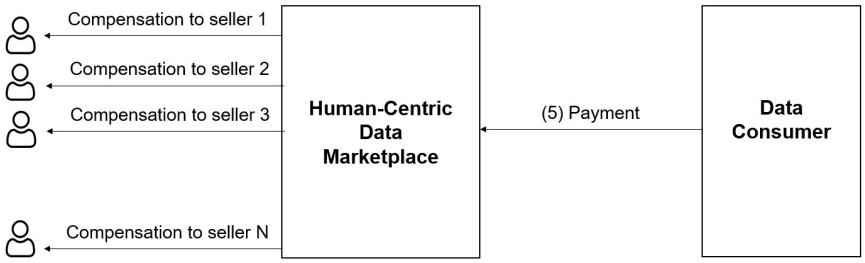
- How are entities trading data doing business over the Internet?
- What kind of data products are being traded? How is data being traded in the market?
- What is the price of data products being sold in commercial data marketplaces?
- Which features of data products are driving their prices in the market?




**How do I set a price to my data?**
*S. Andrés Azcoitia, C. Iordanou, and N. Laoutaris. Measuring the price of data in commercial data marketplaces. In Proc. 1st ACM Workshop on Data Economy (DE '22)*

**How do I select data that suits my task?**
*S. Andrés Azcoitia and N. Laoutaris. 2022. Try before you buy: a practical data purchasing algorithm for real-world data marketplaces. In Proc. 1st ACM Workshop on Data Economy (DE '22)*

(a) Data pricing and selection problems




**How do I distribute payoffs among users fairly?**
*S. Andrés Azcoitia, M. Paraschiv, and N. Laoutaris. Computing the relative value of spatio-temporal data in data marketplaces. In Proc. ACM SIGSPATIAL 2022*

(b) Payoff distribution problem

Figure 1.1: Challenges of a human-centric data marketplace

Were data consumers forced to pay people for their data, they would be required to select among a huge number of data sources of thousands of individuals to feed data to ML tasks. This brings two fundamental problems to the table, shown in Fig. 1.1a. First, a price for the data must be agreed by the parties involved in the exchange, therein lies the data pricing problem. Second, data consumers would look forward to minimising the number of individuals whose information they purchase by carefully selecting those that suits their needs, as opposed to the current practice of indiscriminately hoarding vast amounts of data.

Ultimately, data marketplaces would also be required to calculate the contribution of individual users to each specific data transaction. This might be straightforward if an individual price is set by each seller, but it is far from trivial in the most realistic case that the marketplace defines a price for datasets combining data from different sources. Furthermore, data exchange and trading processes must be efficient to avoid hampering the benefits of a human-centric data economy.

This dissertation defines the former problems in the context of human-centric data marketplaces, and addresses the following research questions in part III:

- How can data consumers select only suitable data for their task from different eligible sources in an efficient way?
- What is the relative value of data from different individuals for A ML task?
- How can we efficiently reward users based on the value of their data?

Moreover, this dissertation addresses the pricing problem from a market perspective by defining a methodology and developing technical components to measure the prices of data products in commercial marketplaces. In part IV, we leverage this study and some of its technical components to present a high-level design of a data pricing tool to answer the question "How do I set a price to my data?".

## 1.2. Contributions

This PhD has contributed to the development of essential knowledge and technical components for the realisation of a human-centric data economy. Specifically, we have delivered methodologies and technical contributions for:

1. Developing natural language processing (NLP) classifiers to compare data products across data marketplaces and addressing the challenges posed by incompatible/non-standard data categories and criteria used by different platforms in order to classify their data products.

2. Designing and implementing a novel feature importance analysis to identify what are the key features, as extracted from metadata scraped from real commercial data marketplaces, that drive the market prices of data products. This study implements basic building blocks of a data pricing tool based on market prices and transaction data, whose design was proposed as part of this work.

3. Proposing a new data evaluation phase prior to buyers purchasing and accessing the data, and defining a new family of "Try-Before-You-Buy" purchasing strategies that allow an efficient near optimal selection of data, and that optimise the amount of data acquired to feed a ML model, hence reducing data leakages and contributing to preserving privacy.

4. Identifying context-specific heuristics to avoid the repetitive processing of complex model to calculate the value of data for a specific ML task, paving the way to more efficient data marketplace designs in the future.

5. Proposing a data pricing tool based on the prices observed in the market. Such a tool leverages key technical components developed in the course of this PhD, and can also benefit from the collaboration of commercial marketplaces.

In addition, we have contributed significant findings related to:

1. Understanding and modelling the data value chain, its stakeholders, their business models and identifying key challenges to help with bootstraping a human-centric data economy.

2. Studying the business models of Personal Information Management Systems,

3. Studying the price of data products in commercial data marketplaces, including the range of data prices, an analysis of the categories that command the highest prices, and a list of relevant features that are driving the prices of data in the market.

4. Understanding the different relative value of spatio-temporal data from companies and individuals in predicting demand or travel time in large metropolitan areas, which is not necessarily proportional to the volume of their data.

5. Testing approximations to the Shapley value as a general baseline metric of the utility of a data source in ML tasks, nd finding an algorithm capable of doing so in polynomial time - $O(N)$ to $O(N^2)$ -.

6. Identifying some heuristics - measuring the amount of information carried by a data source and its averageness - that show significant correlation with the value of data in these use cases and hence can significantly speed up the valuation of data sources.

## 1.3. Structure

This thesis is structured as follows. Chapter 2 summarises and discusses related works, and the reasons why our contributions in this PhD are novel.

Parts II and III address the key research questions of this thesis. Part II introduces the reader to the data economy and presents our works contributing to its measurement. Chapter 3 presents a comprehensive survey of entities trading data over the Internet and their business models, summarises relevant European initiatives such as IDSA and Gaia-X, and frames the scope of our contributions. Chapter 4 presents an innovative study of data products and prices in 10 commercial data marketplaces, applies a novel ML methodology to compare across data marketplaces, and uses regression models and feature importance analysis to find out relevant characteristics of data that are determining the prices of financial, marketing and healthcare products in the market.

Part III presents a feasible human-centric data marketplace aligned with the current state of the art, and deals with issues related to the value of data. Chapter 5 describes the process of buying and selling data in commercial marketplaces. Chapter 6 introduces a data appraisal stage in the purchasing process, and presents the Try-Before-You-Buy family of algorithms, which allow data buyers to acquire and gain access only to the data they need and best fit their purposes. Chapter 7 deals with the relative value of data of companies and individuals, focusing on spatio-temporal data for prediction tasks, and discusses how DMs can distribute payoffs for data among data sources more efficiently.

Finally, part IV discusses the findings and concludes the thesis. Chapter 8 emphasises the scientific value of our contributions, and presents future research lines and works that derive from them. Finally, chapter 9 summarises the key takeaways from our work and closes this dissertation.

Table 1.1 summarises the research questions, the chapter dealing with each of them, and any related publication as indexed in the chapter "*published works*" in the preamble.

Table 1.1: Structure of the thesis

| **Research question** | **Corresponding Chapter** | **Related Publications** |
|---|---|---|
| How are entities trading data doing business over the Internet? | Chapter 3 A data economy primer | (1) S. Andrés Azcoitia and N. Laoutaris, A Survey of Data Marketplaces and their Business Models. ACM SIGMOD Record 2022 |
| How is data being traded in the market? | | |
| What kind of data products are being traded? | Chapter 4 Measuring the price of data in commercial data marketplaces | (2) S. Andrés Azcoitia, C. Iordanou, and N. Laoutaris. Measuring the Price of Data in Commercial Data Marketplaces. 1st ACM DE Workshop (2022) |
| What is the price of data products being sold in commercial data marketplaces? | | (3) S. Andrés Azcoitia, C. Iordanou, and N. Laoutaris. Understanding the Price of Data in Commercial Data Marketplaces. (Accepted) 39th IEEE ICDE (2023) |
| Which features of data products are driving their prices in the market? | | (3) S. Andrés Azcoitia, C. Iordanou, and N. Laoutaris. What Is the Price of Data? A Measurement Study of Commercial Data Marketplaces. ArXiv (2021) |
| How can data consumers select only suitable data for their task from different eligible sources in an efficient way? | Chapter 6 Try-Before-You-Buy - a novel data purchasing strategy | (4) S. Andrés Azcoitia and N. Laoutaris. Try Before You Buy: A practical data purchasing algorithm for real-world data marketplaces. 1st ACM DE Workshop (2022) |
| What is the relative value of data from different individuals for A ML task? | Chapter 7 - Splitting the value of combined data among multiple contributors | (5) S. Andrés Azcoitia, M. Paraschiv, and N. Laoutaris. Computing the relative value of spatio-temporal data in data marketplaces. In Proc. of SIGSPATIAL (2022) |
| How can we efficiently reward users based on the value of their data? | | |
| How do I set a price to my data? | Chapter 8 - Discussion | (2) S. Andrés Azcoitia, C. Iordanou, and N. Laoutaris. Measuring the Price of Data in Commercial Data Marketplaces. 1st ACM DE Workshop (2022) |

# Chapter 2

# Related Works

This chapter discusses and summarises works related to this dissertation, frames the scope of our research and discusses the novelty of our contributions in the following fields:

1. surveys related to entities trading data over the Internet (Sect.2.1),
2. value of data (Sect.2.2),
3. data pricing (Sect.2.3), and
4. data marketplace design (Sect.2.4).

## 2.1. Existing surveys of data marketplaces and entities trading data

Due to the increasing importance of data in the economy, a number of different works have also published the results of surveys about entities trading data over the Internet in the last years. The first systematic survey about data marketplaces was published in 2013. It already defined data marketplaces as platforms dealing both with buyers and sellers and allowing for selling, buying, or trading data [163], and identified categories of 55 data marketplaces and vendors. Yearly updates of this survey were published [146, 169–171] until 2017, expanding the sample to up to 72 entities, and identifying trends in this period.

A more recent work emphasised the increasing relevance of data trading in the B2B market, and developed a morphological box to clearly display different dimensions of analysis [168]. Given the increasing importance of ML and AI, some vision papers have focused on marketplaces aimed at trading data for ML and AI models. A recent paper included a survey of 24 of them in different stages of development [112]. Finally, some other papers have complemented these surveys by interviewing data marketplaces and have pointed out key challenges of data trading over the internet [64, 101].

Our survey presented in chapter 3 is more up-to-date (half of the surveyed entities were founded in 2016 or later), it is broader in scope, and provides an in-depth analysis of up to three

times more entities than previous surveys. It concludes by explaining some key challenges found during the market research, and it extends the morphological box to also identify and characterise up to ten different business models that data trading entities are adopting. Further - and following our study of 20 of them - our work is also, to the best of our knowledge, the first to address the business models and challenges of PIMS.

Furthermore, none of these works dealt with commercial products being sold by commercial data marketplaces. To the best of our knowledge, our work presented in chapter 4 is the first empirical measurement study that deals with commercial prices of data. Our study on this subject matter is also timely: the lack of empirical studies about the prices of data was considered as a key challenge in a recent comprehensive survey about data pricing [153], a topic we also address in this thesis.

## 2.2. The value of data

"*Data*" has become a valuable production factor comparable in importance to capital, land, or technology. An increasing number of business models rely on data to provide services to end users or to improve decision-making, and hence estimating the value of data has become a central problem that has attracted the attention of practitioners from different disciplines. From accounting and estimating the value of data assets of a firm in a due diligence process [190] to estimating the value of data samples to train specific ML models [75], the range of data valuation tasks and their targets is wide and covers very different goals. Due to the elusive nature of "*data*" as an economic good, it is far from obvious how to estimate its value, and its unique characteristics dramatically condition how it is exploited and traded in the economy.

Before summarising the literature dealing with this topic, Sect. 2.2.1 discusses some metaphors comparing data to other well-known economic goods in an attempt to explain its peculiar economics. Then we review the distinctive attributes of data as an economic good in Sect. 2.2.2. Finally, Sect. 2.2.3 provides an overview of the state of the art and highlights relevant publications dealing with the value of data. The existing research around this topic is immense and comprises works from very different disciplines. It is not the objective of this section to provide a thorough list of publications dealing with the topic, but to point the reader to valuable survey works and to frame the scope of our contributions.

### 2.2.1. "Data is like..." – Metaphors about the value of data

It is precisely the elusive nature of data as an economic good and an asset that has inspired a number of metaphors by the industry in the last years. Perhaps the most famous (and abused) quotation is that "*Data is the new oil*", initially attributable to Clive Humby's keynote in the Senior Marketer's Summit at Kellogg School of the American Advertisers Association back in 2006 [30]. He emphasises that data is easily available to marketers, and its huge value if adequately "refined", and transformed to relevant information for making decisions and taking actions. He used the term

commodity, which is a gross oversimplification of what data is. Unlike oil, data can be copied / transmitted / processed with close to zero cost. Notice that whereas two litres of gasoline yield a similar mileage on two similar cars under similar driving styles, nothing of this sort applies to data since 1) two datasets of equal volume may carry vastly different amounts of usable information, 2) the same information may have tremendously different value for Service A than for Service B, and 3) even if the per usage value of two services is the same, Service A may use the data 1,000 times more intensely than Service B leading to extremely different produced benefits.

Ed King. CEO of OpenPrise compared data to water [103] in the sense that data is everywhere, companies live on data and can be "drowned" in useless data, too. Data, like water, need to be cleansed and processed for consumption. Like in selling water, packaging is very important. He also touches upon the value of data for specific purposes and states that it is not necessarily tied to its price. For example, free open data can be used to enrich and significantly enhance expensive datasets. However, unlike water or oil, data is non-depletable meaning a data source cannot be exhausted by repetitive and continuous consumption and usage.

Some authors dare compare data to a currency, in the sense that it has economic value and can be purchased, sold and traded, as long as ownership is clearly established and protected [73]. Other works question and review the role of governments as a data producer, consumer and facilitator of this data market that will enable the new currency [61]. The idea of providing *micropayments* to users for their personal data as a way to overcome the abuse of privacy on the Internet has received a lot of public attention [114, 155]. More recent work describes fundamental technological challenges that need to be addressed for the above vision to be fulfilled [115].

In that scenario, individuals would be able to have an income out of their personal data and out of their contribution to the digital economy. This transforms data in a production factor comparable to labour [11]. Should this possibility materialise, users may need to associate to data unions that defend their rights on the exploitation of their personal data. Some PIMS are already positioning themselves as a data union of "users practising their data rights" [174]. Association to such unions should probably be recognised as a digital right, and supporting them may require regulatory intervention upon large digital firms, similar to obligations regarding labour unions that have been imposed on large traditional enterprises. Data unions would probably be required to act as non-for-profit organisations (i.e., as fiduciaries) for passive data users to trust them [54]. However, data as labour is not an accurate metaphor for data in the economy, either. Unlike labour, data is a non-rivalrous good meaning that its supply is not affected by its consumption, and thus selling data to a consumer A does not prevent a data provider from selling (a copy of) the same data to a consumer B.

More recently, Tim O'Reilly compared data to sand, thereby highlighting the huge effort required to develop digital products and services from raw data and sending a radically different message regarding the distribution of the value of data across the chain. Raw data from an individual is worthless, unless combined with that of other million individuals and adequately treated through complex processing and data pipelines to feed an innovative use and business case [180].

In the end, data is a digital sub-product, and hence likely to benefit from the recommendations on the monetisation of this kind of economic goods [164]. Still, some authors have pointed at some specific characteristics of data that are not strictly applicable to other digital goods, such as the uncertainty as to the units of data to be traded (or rather their dependency on the type of data and the context in which it is used), its inherently combinatorial and aggregated value, and the fact that data often is consumed by machines, whereas digital products and services target people and enterprises. Finally, data and digital goods are sold in different ways. Unlike digital goods, data is often an intermediate production factor to be combined and processed with other data to build digital services [153].

In conclusion, the economics of "data" are peculiar, and no single metaphor or comparison is able to accurately capture its specific characteristics and define its behaviour in the economy. Rather than being accurate in all situations, the intention of such metaphors has been to highlight specific features of "data" in keynote speeches, position papers, reports, or scientific papers.

### 2.2.2. Data as an economic good

Many works discuss on the characteristics of data as an economic good. In this section, we point to some of them and summarise key features of data, which will serve as a preliminary basic background to understand the challenges of marketplaces and entities dealing with data and trading it in the market [47, 113]. As an economic good, data is:

- *Freely replicable*, meaning it can be copied and transmitted at close to zero cost.
- *Non-depletable* or *Non-perishable*, meaning that it is not exhausted by a repetitive and continuous consumption.
- *Reusable* for different purposes and clients.
- *Non-rivalrous*, that is, selling data to a consumer A does not prevent from selling (a copy) of the same data to another consumer B.
- *Permanent*, unlike information that is perishable and depreciates over time.

Unsurprisingly, these peculiar characteristics determine the behaviour of data in the economy. Next we summarise some considerations found in the literature about the value of data:

- Since the marginal cost of processing data is reduced, *its value increases with its use*.
- It has an *inherently combinatorial value*, the more data is combined and used, the more value it usually generates.
- However, its *value depends on the context* and on the existing information: more data does not necessarily mean better data.

- *Its value is affected by quality*, but features affecting that quality (e.g., completeness, accuracy, timeliness) also depend on the context.

- Uniqueness also affects the value of data: *the more you share it, the lower its value becomes*.

- *Packaging and delivery methods* (linkability, interoperability) are important.

- *Its value is affected by externalities*, such as the implications of data sharing in terms of privacy.

In a nutshell, data is a key factor for producing digital services and goods, and the goal of data pipelines and value chains is to transform data into information and then into more valuable digital services and goods that benefit from the increasing value of combining different pieces of data. Hence, enabling a market around data means trading intermediate inputs along these pipelines and allowing third parties to combine them into more complex use cases aimed to improve decision making, to increase the efficiency of productive and operational processes, and to create new innovative services for end users.

Furthermore, the characteristics of data and the former considerations regarding its value affect how it is traded in the market [97, 164] and shape the data economy and the nascent data markets. Since it is easily replicable and its value reduces as it is shared in the market, most companies are often reluctant to share data and give up monetising it because they are scared of losing competitive advantages in doing so. This is having severe consequences and undermining the potential benefits of data in the market:

1. Firms tend to hoard huge amounts of data, avoiding any sharing with third parties. As a result most data being collected by companies remains in silos nowadays.

2. Since it is often difficult to find data in the market, most companies look for innovative ways to collect those data. As a result, data collection is inefficient.

3. The abuse of personal data collection with little or no consent and the amount of personal data stored by different parties are raising a sentiment of lack of privacy on the Internet.

Regardless its nature, data is becoming a cornerstone in the digital economy, and it is now considered a key asset of data-driven companies. Therefore, it is necessary to measure its value to understand the importance of firms to the economy, or just to monitor the development of the data economy. The practice of data governance intends to improve the management of data, make an inventory of data assets in firms, and apply asset valuation methodologies the measuring and understanding the value of data. In the next section, we provide an overview about these valuation methodologies and we frame the scope of the work and contributions of this thesis in this field.

### 2.2.3. Measuring the value of data

For most people, the 'value of data' has been linked to that of personal data, and more specifically to its application in marketing and advertising. As we will show in chapters 3 and 4, this view is very restrictive, and there are many different types of data being traded in data marketplace that can be used in very different use cases other than targeting users to improve the performance of online advertising campaigns [87]. Due to the wide spectrum of data and use cases that can be found in the market [118], many different methodologies and works attempt to estimate its value often resulting in apparently contradicting estimations.

From a macroeconomics perspective, the OECD published an interesting survey summarising different approaches to determine the monetary value of personal data [139], including very heterogeneous methods such as examining market capitalisation, revenues or net income of data-driven firms per individual, analysing revenues or net income per record/user, or assessing the cost of data breaches, which in turn assumes personal data as a liability. Another common methodology to approach this problem is through economic experiments and surveys to users' willingness to pay to protect their data [34]. A more recent literature survey works adds impact-based valuation that also considers the social and economic outcomes of data use cases [57].

In chapter 4, we address the value of data from a market perspective, which assumes the value is related to the prices of data in data markets, and the volume of money traded in such markets. Previous papers have presented similar market-based solutions close to the one we followed, but focused on prices observed in online advertising [100]. Some tools like this were implemented later on and are able to estimate the relative value of different user profiles and the revenue users generate for social networks like Facebook [31].

Other works have assessed the value of data for specific tasks from a perspective in the border between microeconomics and computer science. As opposed to macroeconomic approaches, these works provide a detailed valuation of data for specific purposes and contexts. Finding scalable and fair ways to compute value-based contributions to a ML problem is key in calculating the contribution of individuals to the data economy, and moving towards a human-centric data economy will require to find fair scalable ways to do so.

In particular, in this thesis we have focused on the value of spatio-temporal data for ML prediction tasks. The use of spatio-temporal data in transportation and smart city applications has attracted much attention from the research community. Different works look at how knowledge extraction from spatio-temporal data can significantly improve the effectiveness of transportation [198], mobility prediction [10], or last mile delivery [46], among others.

In this context, some authors have also studied the *intrinsic* value of spatio-temporal information, and calculate it as the reduction of the uncertainty about the position of an individual [136]. Similar to ours, other works have dealt with the *extrinsic* value of spatio-temporal data for a specific problem in a specific context [8, 9], while others have introduced the notion of privacy in pricing spatio-temporal data [135]. These valuable works differ from ours in 1) they adapt different notions of value instead of using the more generic Shapley value used in our work, 2) they

look at different applications domains (e.g, location-based marketing, traffic prediction) than the ones we study, and 3) they deal with different problems, such as identifying the best moment for bidding or acquiring a new data point, but they do not address that of rewarding data sources contributing to a single dataset based on their value. To the best of our knowledge, our work presented in chapter 7 is the first to apply Shapley values to ML tasks using spatio-temporal data.

As a consequence of the complexity in understanding the value of data, setting the price of a dataset is often a complex challenge for data sellers. In the next section, we review the literature related to data pricing.

## 2.3. Data pricing

Data pricing is a thriving research field gathering researchers from very different disciplines. Several authors agree that pricing techniques are critical to realise the value of data [119], and some valuable works have published comprehensive surveys about the pricing of data and digital products [153]. Some lessons learnt while pricing digital goods may also be applicable to data pricing [164]. For example, *versioning* and *personalisation* (i.e., tailoring data to each buyer) using different levers such as the age of information, its resolution, format or features are being used by data sellers to structure their offer. Other common practices in pricing digital goods are being applied to the pricing of data, such as *segmentation* and *group pricing* (e.g., designing specific products targeting researchers), *bundling* of complementary data products, or the provision of free samples or *freemium* pricing models. Other authors claim that some techniques to set the prices of cloud services [196] are also applicable to data products.

Different pricing strategies may be considered in a data marketplace [134], including usage-based, subscription-based, package-pricing (e.g., a number of API calls for a fixed price), flat rates, two-part tariffs. In reviewing the existing literature, we have identified six different "*schools*" when it comes to approaching data pricing:

- Data **auctions** are initially complex to design to ensure truthfulness for non-rival products with unlimited supply. Some authors proposed auction designs to set the prices of digital goods and data products and showed that non-deterministic auctions can be competitive. Later it was shown that this does not hold for asymmetric auctions (i.e., those whose outcome depends on the order of arrival of the bids) [4]. Randomisation techniques have been used, as well. Consensus estimates (CORE) auctions randomise the auction mechanism [79], and other authors have made bidders compete by randomly grouping them [80].

- Novel **AI/ML** data marketplace architectures have been proposed under the concept of value-based pricing [40–42]. Some authors have also combined value-based pricing with auctions to make bidders compete for the quality of data, too [3]. Mechanism design plays an important role in designing incentives for players to participate in federated ML models [175]. Other collaborative marketplaces, very close to federated learning, have

designed value-based incentives to compensate each player depending on how much its data contributes to improving a joint model [142].

- **Query pricing** looks into the problem of pricing views or queries as versions of a database [20, 107]. Arbitrage opportunities have been extensively studied [21, 121] and pricing functions have been defined for brokers to maximise their revenues, including empirical evaluation on real-world instances [39]. Empirical tools and frameworks like Qirana [53] have been proposed to establish query markets [108].

- Some authors claim that **privacy** should be considered in the pricing of personal data, and have defined pricing strategies and marketplaces based on differential privacy loss of consumers [76] and also query pricing [117]. Some authors have developed privacy-preserving data marketplaces [137], which propose compensating sellers regardless their data is included in the final set of traded data.

- **Quality-based pricing** prices different versions of data by evaluating and assigning weights to certain quality features [86]. Some works have defined data pricing strategies for sellers based on this idea [201].

- **Dynamic pricing** introduces the temporal dimension in data pricing. Some query pricing mechanisms support history-aware pricing (depending on the buyers' purchase history) [53, 107]. Dynamic prices have also been applied to spatio-temporal data [204], to maximise revenues considering different arrival times for buyers [138], and for disincentivising undesirable behaviours of buyers and sellers [132].

The pricing of personal data has received a lot attention from the privacy and measurement communities. There are measurement studies based on prices carried over the Real Time Bidding protocol [143, 150]. These works report auction prices for advertising spaces in web pages and for trying to attract the attention of the right individuals and, therefore, have nothing to do with the price of B2B datasets traded in modern marketplaces for different purposes that we discuss in chapter 4.

"*Versioning*", that is producing and offering different versions of a data product, with different utility and price, is actively used by different pricing mechanisms. Freshness, history, features, scope, volume, format, resolution or accuracy of data are being used to offer different versions of a data product. The injection of noise is also used by data marketplaces (e.g., to location or to numerical values resulting from a query) to generate *a priori* less valuable and more privacy-preserving products.

Finally, some authors have studied the pricing of "*bundles*" of data products, which makes sense as long as bundling contributes to increasing the willingness to pay for information goods [164]. There is a framework to study the conditions under which bundling produces more revenues [29], and when pure bundling is optimal [85].

After reviewing the different works in the literature, we conclude that quality-based pricing is the closest approach to our work on data pricing, which provides a reasonable price for data products based on the characteristics of similar products in the market. However, we are not aware of any previous study that has been able to derive weights for metadata features from real market data as we do in chapter 4 of this thesis.

In conclusion, all these works are complementary to our proposal of helping data sellers set a price for their data products by building a data pricing tool based on the metadata of products already available in the market. This approach assumes that data is an asset whose value can be approximated by that of similar products in the market. Moreover, this somehow assumes that data, at least that within a certain category and having specific characteristics, is a commodity, as well. Hence, this approach might not be very fully functional for quoting unique datasets, or in early stages of the data economy when the amount of information to feed the tool is still scarce. However, it will be increasingly useful in time as more data is unlocked, data trading thrives, and new products are offered in the market and added to the information base.

## 2.4. Data marketplaces in the research community

The terms '*data market*' and '*data marketplaces*' are trending topics in different fields of scientific research [60]. Recent vision papers state the challenges of building a data marketplace from a broader perspective that the scope of this dissertation. Not only do they focus on data transactions and value, but they do also emphasise challenges related to discoverability, integration, and transparency, and deal with the systems perspective [64, 101].

A recent comprehensive literature review has revealed 82 publications related to data markets or marketplaces in the period 2016-2021 and has pointed to the wide range of relevant topics, challenges and solutions in the field [60]. In this thesis, we focus on the design of ML-oriented data marketplaces, meaning those designed to trade data for training ML models. Recent theoretical work on the intersection between computer science and economics has looked into designing such marketplaces, and has proposed solution concepts and algorithms which we summarise in this subsection.

Most ML-oriented theoretical DM proposals leverage a data valuation framework similar to the one we propose in this thesis [3, 41]. Different value-based marketplaces have been designed based on this framework, that measures the value of data as the gain in some accuracy metric brought to the buyer's specific ML task. In this context, there is an ongoing debate regarding whether DMs should return trained models, insights or bulk data [50].

Some designs prepare versions of trained ML models by adding noise to the weights of the model or to the data used to train it, and they offer buyers an arbitrage-free mix of them with different value and price [41]. Other designs add a bidding process in which buyers can pay for higher accuracy, and deal with aggregating data from different sellers and with fairly compensating sellers based on the utility that their data brings to the ML task [3].

In such settings, calculating payoffs according to the accuracy (value) that data brings to a model is a complex problem, often dealt with by approximating the Shapley values of data [35, 75, 95, 96, 151, 160, 181, 191], some of which we test and apply in Chapter 7. We provide more details about those algorithms in Sect. 5.3.3. Also related to the computation of the relative value of data for a specific ML task, a family of works have studied the calculation of Shapley values for specific problems such as those related to characteristics of points in the plane, online advertising, or traffic load in telecommunication networks [32, 172, 203]. None of the above works has looked at valuation issues relating to spatio-temporal data as we do in chapter 7.

Some privacy-preserving data marketplace designs introduce the ability of data owners to manage their privacy preferences by using the concept of differential privacy, and proposes a way to calculate its impact on their compensations [122]. Privacy enhancing technologies (PETs) will definitely play a role in data marketplaces [77], and some designs already propose to use technologies such as multi-party computation to let buyers process third party data using brokers as intermediaries. Brokers can ensure data policies are followed but cannot access the data nor the results of the process [158]. Others tackle the problem of fair payment distribution, as well [178], and some broker prototypes, designed as smart contracts, use cryptography techniques to train - still simple - ML models while protecting data privacy [109].

Distributed ledgers, and particularly blockchains are oftentimes proposed as an immutable information registry about data transactions, policies and data usage in data marketplaces. Some prototypes look to integrate IoT data flows in ML models [207]. The trend is towards federated and decentralised marketplace platforms [77]. Some commercial DMs (IOTA for IoT data, Swash for personal data, or MedicalChain for health-related data) are implementing blockchain-based data exchange platforms, often settling payments for transactions using their own cryptocurrency, supported by their own public technical whitepapers [7, 67, 154].

Federated machine learning, or simply federated learning (FL), has also worked on proposing incentives for workers to contribute to training a shared model [99], and some works talk about Federated Learning as a Service (FLaaS) [106]. Some collaborative DMs propose a framework close to FL for different parties (acting in this case as both data providers and consumers) to train a common model by contributing their own data and calculating payoffs depending on how much its model improves with data of other parties, and how much its data contributes to improving the model of other participants [142].

Other works look forward to increasing the awareness of the costs of training models collaboratively using FL [33]. Inspired by our work, some authors have also proposed adding a data appraisal layer to select private training data from a ML marketplace [197], and we have found similar proposals for data marketplaces based on FL [33]. Similar to this dissertation, other more recent works do also state that the appraisal of AI/ML data prior to purchasing it will have a significant impact in making data sharing more efficient in terms of computing cost [84].

# Part II

# Understanding and measuring the data economy

# Chapter 3

# A data economy primer

This chapter contextualises our work within the so-called data economy, defined by the European Commission as "*an ecosystem of different types of market players [..,] collaborating to ensure that data is accessible and usable. This enables the market players to extract value from this data, by creating a variety of applications with a great potential to improve daily life*" [44]. Building a thriving data economy has been a key policy objective of the EU, and it is closely related to other relevant policies such as the Cloud Strategy, the Advanced Computing and the AI strategy within the so-called "*Shaping Europe's Digital Future*" programme.

The complex ecosystem of entities trading data on the Internet combined with the inherently complex economics of data, limits our ability to understand this evolving ecosystem and to contribute, from the research point of view, to its evolution. In this chapter we strive to present a comprehensive survey of commercial data trading entities, covering a wide spectrum of target data types, business models, and technologies they use. We develop a taxonomy system and identify ten business models by studying 104 entities in depth. Moreover, this analysis points to a number of open challenges that we believe should attract the attention of the research community, including issues of pricing, federation of different marketplaces, and ownership protection.

First, we introduce the data value chain in Sect. 3.1, and then present the results of the broadest survey of entities trading data over the internet in Sect. 3.2. This is also the first survey to analyse personal information management systems (PIMS) whose business model we describe in more detail. In Sect. 3.3 we provide an overview of two relevant initiatives aimed at building a trustful sovereign data trading ecosystem in Europe: the International Data Spaces (IDS) and Gaia-X projects. Finally, Sect. 3.4 concludes with the key takeaways from this comprehensive study, and frames the scope of rest of the dissertation.

As a starting point, we provide some background, and we introduce some of the terms and definitions we will use throughout the chapter, along with our view about the data value chain.

## 3.1. Understanding the data trading value chain

In the context of data trading, *actors* in the value chain are legal entities or individuals playing an effective role in producing any data-driven service or data product, be it intermediate or final, that is offered and eventually acquired or exchanged in the market. We will generally refer to them as data trading entities (DTE). Our survey aims to understand what the roles of such DTEs are, how they interact between them, how they do business, and what mechanisms they use to set prices for data. We encapsulate all this information in the concept of a *business model*, a term that has been defined in various ways in the literature [147]. For the purpose of this thesis, we will refer to a DTE's *business model* as the description of its value proposition within the chain, the processes or activities it covers, the inputs it requires, and the outputs it provides the market with, as well as the relationship the entity maintains with other actors in the ecosytem [43].

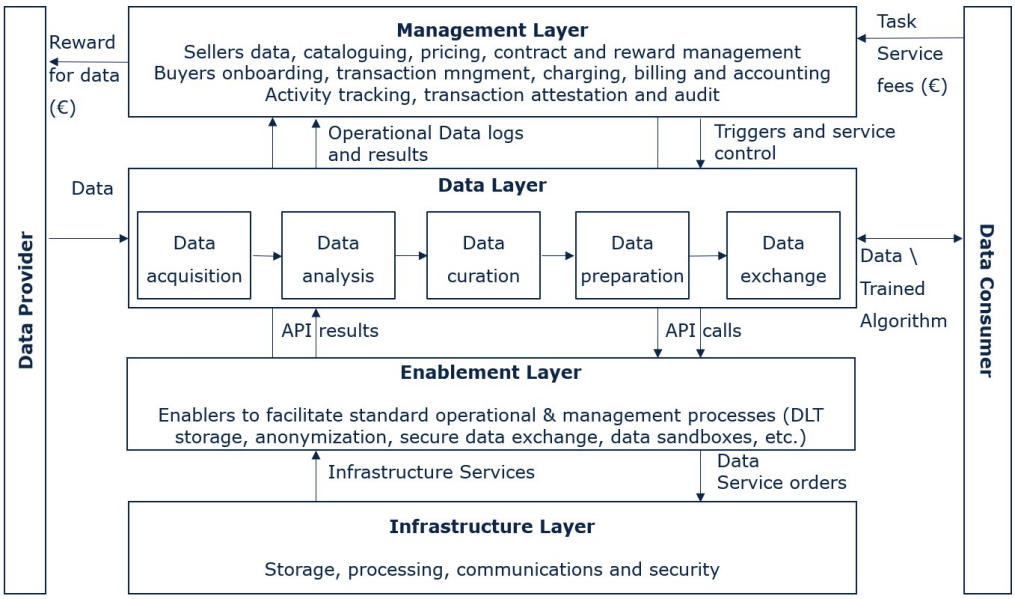



Figure 3.1: A layered approach to data trading

Understanding the data value chain is a key first step in order to identify relevant business models. Previous studies have already explained the data value chain in general [49, 82], and in specific contexts [127, 128]. From a broad data trading perspective, Fig. 3.1 shows a diagram of four stacked functional layers that allow sellers and buyers to connect.

At the bottom, the *infrastructure layer* provides the basic processing, secure storage and communication functions to the upper layers in the stack.

On top of such infrastructure, the *enablement layer* provides generic application programming interfaces (APIs) and functions to DTEs. Some solutions and PDKs in the market do not intend to directly provide services to the end users, but rather to provide a platform with common useful functions that *enable* other DTEs to carry out a controlled data exchange which optionally may involve an economic transaction.

In the next level, a more technical and operational *data layer* deals with data processing itself and responds to the effective delivery of data or data-driven services to customers, be they consumers, or other DTEs. Reaching from data collection or extraction to its final delivery to the end consumer, this process usually requires intermediate preprocessing, curation and data enrichment steps. In addition, it may involve third parties whose data is acquired and combined, and therefore other secure data exchanges.

Finally, the top *management layer* deals with data discovery, coordinates transactions, keeps track of contracts and service level agreements, and ensures the accountability and transparency of all the operations and processes in the data layer. In contrast to the operational *data layer*, it works with metadata and transactional data. Other functions of the management layer include helping data owners catalogue, structure and price their data offer, governing transactions (e.g., through contract management, charging, billing and accounting processes), and increasing the overall transparency of data trading. In the case of transactions involving data from multiple sellers, it is also in charge of distributing the resulting payments among them.

Note that our definition allows for cascading transactions, which is oftentimes the case before sufficiently processed data is transformed to a data-driven service to end-consumers. For example, a model that outputs consumer segmentation data at postcode level requires at least the following steps: i) gathering anonymised segmentation data (often from disparate sources), ii) combining such information with geo-located identity data into a single coherent dataset, and iii) aggregating this output into individual postcodes by processing it together with postcode border shapefiles (often obtained from a third party, too) in a geographical information system. Each of this steps requires (part of) the layers depicted in Fig. 3.1.

In the next section, we present the most comprehensive survey of entities trading data over the Internet. We identify ten different business models that cover part of the former layers and roles in the value chain, and we extract valuable information on how the different entities exchange data and do business.

## 3.2. A survey of entities trading data and their business models

A number of open directories already exist on the web for listing data trading entities [45, 52, 157]. Such directories loosely tag as '*data marketplace*' heterogeneous entities with hugely different objectives, focus, customers, and business models, etc. For example, traditional *data providers* have been collecting, enriching, and curating public and private information from different sources and silos for years, building successful business models mainly in the areas of marketing (Acxiom, Experian, etc), financial and business intelligence (Bloomberg, Thomson Reuters, etc.). More recently, *data marketplaces* (DMs), i.e., two-sided platforms for matching data sellers and data buyers and mediating in data exchanges and transactions, have also arrived on the scene [163].

First-generation *general-purpose DMs* trading *any* kind of data are being complemented by *niche DMs* that target specific industries (e.g., Caruso for the connected car, Veracity for energy and transportation), and cover data sourcing for specific innovative purposes, such as feeding ML algorithms (e.g., Mechanical Turk, DefinedCrowd), or trading IoT real-time sensor data (e.g., IOTA, Terbine). Additionally, some leading *data-management systems* (e.g., Snowflake, Cognite) and *niche* digital solutions (e.g., Carto, Openprise, LiveRamp) are integrating secure data exchange features and capabilities to their platforms with the aim of breaking data silos [64].

Aided by recent legislative developments, including the General Data Protection Regulation (GDPR) in the EU or the California Consumer Privacy Act (CCPA) [140, 182], *Personal Information Management Systems (PIMS)* have appeared with the purpose of empowering individuals to take back control of their personal information currently being collected by Internet service providers, with little or no consent. Some of them have also implemented marketplaces for helping users monetise their personal data [169].

This chapter presents the most comprehensive survey of entities trading data over the Internet, and a taxonomy of ten different business models that we identified and characterised while gathering information for this survey. Other survey papers have been published regarding DMs [146, 163, 168, 169]. Ours is more up-to-date (half of the surveyed entities were founded in 2016 or later), broader in scope, and provides an in-depth analysis of three times more entities than previous works, as shown in Fig. 3.2. Further - and following our study of 20 of them - this work is also, to the best of our knowledge, the first to address the business models and challenges of personal information management systems (*PIMSs*). The remainder of the section is structured as follows:

- Subsection 3.2.1 presents the methodology we followed to carry out the survey.
- Subsection 3.2.2 presents the scope of the survey.
- Subsection 3.2.3 characterises the business models compiled during our study.
- Subsection 3.2.4 provides quantitative results on how different entities trade and exchange data over the Internet.

### 3.2.1. Methodology of the survey

The methodology used to carry out the survey consisted of the following steps:

1. **Identifying of target companies** trading or making business by delivering data. Companies were identified by either searching the web with relevant key words, or by browsing through relevant articles and papers available on the internet.
2. **Making a quick first assessment** and classifying companies according to the type of data they are trading, the industry they target, their type of clients and business model.

3. **Formulating a comprehensive set of questions** covering all the aspects we want to answer in this study, and defining a preliminary set of possible answers to each of them. This was further refined during the research process to generate a taxonomy for presenting the results of the survey.

4. **Carrying out a desktop research** to dive deeper into each specific company, answering to the survey questions in a data sheet, and generating a detailed information dossier about the company for consultation purposes in a latter stage as needed.

5. **Building the data taxonomy by homogenising the answers** and refining them to allows for comparing between them.

6. **Analysing the results** of this study, both from a technical and a business perspective, characterising key business models, and listing entities that have adopted each of them.

In the next sections, we list the questions we set out to answer while studying the different entities trading data on the Internet in Sect. 3.2.1.1, and we discuss some limitations of our study in Sect. 3.2.1.2.

#### 3.2.1.1. Survey questions

Table 3.1 summarises the questions considered in the survey and the different possible answers we found when studying the different entities. It also refers to the subsection where we present the results and discuss each topic.

Apart from answering the former questions, we gathered some general data to classify each entity, understand its maturity, and measure its popularity. These KPIs include the foundation year, country of origin, companies backing the project, the number of employees, how much money they raised, its AlexaRank and its trend in the last months.

#### 3.2.1.2. Data collection approach and limitations

Data acquisition was the result of a desktop research based on secondary information available on the Internet. As a consequence, the survey relies on information that the target entities are directly publishing on their websites, as well as any related material, such as whitepapers, public videos, product brochures and presentations.

Whenever an answer was not found for any question in the case of a specific entity, ”N/A” (meaning *not available*) labels were used. In general, this situation is due to either a lack of information when analysing such entities, or due to insufficient detail of such information to answer the question. We report the percentage of entities for which we have information in each subsection presenting the results of the survey.

Table 3.1: Survey questions and taxonomy of the results produced in the survey

| Field | Question | Values | Sect. |
|---|---|---|---|
| Type of data | Which kind of data is the entity trading? | IoT Sensor Data; Personal Data; Geo-located data; Contact data; Marketing; Corporate data; AI / ML models; Human-generated data; Multimedia; Industry; Trading Data; Web data; Automotive-related; Identity data; Healthcare data; Genetic data; Any | 3.2.4.1 |
| Providers | Who are the data subjects? | Individuals; Businesses; Sensors; Any source | 3.2.4.2 |
| Consumers | Who are the data consumers? | B2C; B2B; Any | |
| Targets | Who is the target of the entity? In case of B2B exchanges, which department or specific industry? | Digital Service Providers; Marketing; Market research; Financial; Automotive; Individuals; Energy, Logistics, Oil & Gas; Healthcare; Retailers; Any | |
| Pricing mechanisms | Which data pricing mechanisms are available? | Fixed subscription; Bid by Buyer; Fixed by seller; Auction; Customised; Free; Revenue Sharing; CPM; CPC; %Gross Media spent; Volume-based; Open; N/A | 3.2.4.3 |
| Actor(s) setting prices of datasets | Who sets the price of traded datasets? | Providers; Platform; Subjects; Buyers; Open | 3.2.4.4 |
| Payment redistribution | How does the platform redistribute payments to data subjects / sellers? | One-to-one; Contribution-reputation-based; N/A | |
| Data transaction payment | Which payment method and/or currency is used in data transactions? | Fiat currency; Token; Internal credits; N/A | 3.2.4.5 |
| Platform pricing policy towards data subjects | How are data subjects charged for accessing the platform? | Free; Connection fee; Time subscription; IaaS platform charges; Shipping fees; Freemium; Open; N/A | 3.2.4.6 |
| Platform pricing policy towards data buyers | How are data buyers charged for accessing the platform? | Free; Connection fee; Subscription; Revenue sharing; IaaS platform charges; Shipping fees; Customised; N/A | |
| Platform pricing policy towards data sellers | How are data providers / sellers charged for accessing the platform? | Free; Connection fee; Time subscription; Revenue sharing; Freemium; IaaS platform charges; Shipping fees; Partnership; One-off fee; Sales commission; N/A | |
| Access for providers | How do data providers get access to the platform? | API for data providers; Web-services; Mobile App; compatible DPs' systems; N/A | 3.2.4.7 |
| Access for buyers | How do data buyers get access to the platform? | API for data buyers, Web-services, Proprietary mobile app, compatible data buyers' systems, direct contact, N/A | |
| Data sources | Where is data coming from? | internet; Self-generated; Sellers; Data Providers; Users; IoT devices | |
| Data delivery | How is data delivered? | Access to encrypted data by sending a revokable PK, through an API, through a specific software or platform, by training models with the data; dataset bulk download; Web services | |
| Data storage | Where is data to be traded stored? | Centralised public cloud backend, Decentralised private clouds, Centralised private cloud, Data subject's device, Distributed depending on data provider, Centralised backend, DLT, Centralised backend or Public cloud, Decentralised public cloud, Data subject's device and Centralized servers, N/A | 3.2.4.8 |
| Transaction / Management data storage | Where is the information about data transactions stored? | DLT, Public cloud backend, DLT or centralised management, Centralised backend, Distributed depending on data provider, N/A | |
| Structure of data | Who determines the structure of data to be stored? | Data owner, Application, Data sellers, Data providers and the platform, Platform, Data providers, N/A | |
| Data preview | How can the data buyer see or test the data before it is transacted? | No way, Free sample Data, Free access, Demo, Trial period, Reputation mechanism, Testing sandbox | 3.2.4.9 |
| Data security measures | How do entities prevent unauthorised access to data while stored? And while it is being moved? | Encryption and SSL, DLT decryption key distribution, Distribution of PK, Special additional measures for PI, Secure storage for PK, User authentication, DLT data replication and immutability protection, Distributed secure data storage | 3.2.4.10 |

### 3.2.2. Scope of the Survey

Several iterations were required in order to come up with a comprehensive set of data trading entities, and fully understand the current market situation. We initially checked more than 190 companies offering data products on the Internet.

After a first quick assessment, we discarded some entities for their subsequent in-depth study and documentation. In particular, we rejected online advertising platforms not offering a private marketplace, concept projects either lacking information or discontinued in time, entities no longer providing service nor providing any data exchange or data-driven service as such.

Finally, we filtered out some entities whose business model was already well represented in the survey. For example, we found many data providers with similar business models, but we only included 38 of them in the survey. We prioritised those providing clear pricing information. As an exception, we did include every active PIMS we found to provide the reader with a thorough overview of this brand-new business model.

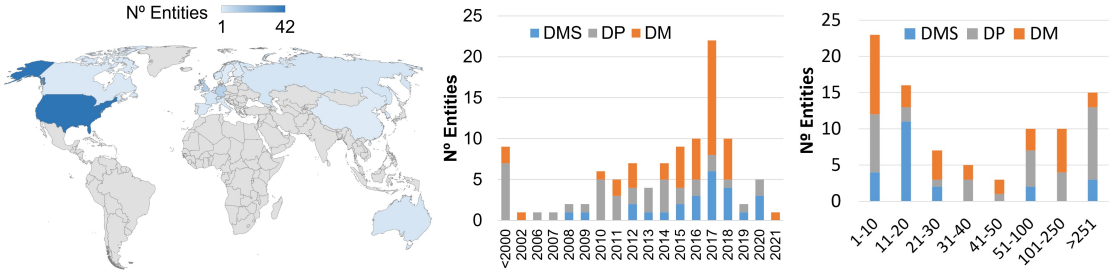
As a result, we discarded the following entities for the in-depth study: AAAChain, Acxiom, Adcolony, Adelphic, Adform, Adition, admaxim, Adobe Advertising Cloud, Adot, AdSquare, adsWizz, adXperience, Algorithmia, Amaxon Mechanical Turk, Amobee, Apervita, Axonix, Bidtheatre, BigChain, BigToken, Bottos, Bluetalon, CentroBasis, Clearview.ai, Cogito, Complementics, CoverUS, CXSense, Datacoup, Dataguru, DataHub, Datax.io, DataXpand, dbc, Evotegra, Experian, Eyeota, Fyber, Hu-manity, ifeelgoods, IBM, , iExec, Imbrex, ImproveDigital, Informatica Data Exchange, InMobi, LiveIntent, LUCA, Magnite, Mediarithmics, Microbilt, MyHealthMyData, Nielsen, OpenPDS, Opiria Blockchain, Optum Data Exchange, Orderly, OwnData, OwnYourInfo, PickcioChain, PlaceIQ, Pubmatic, Qlik Datamarket, Relevant Audience, Reply.io, Reveal Mobile, ROKU (Oneview), Rubicon project, RythmOne, Smaato, Smartclip, StreetCred, Synchronicity-IoT, Tabmo - HAWK, Taboola, TapTap, The DX network, Tremorvideo, Trufactor, Wove, Xandr, xDayta.

After this initial filter, we selected 104 of them listed in Table 3.2 to be analysed in detail. The final set includes companies of different sizes from 23 countries, as Fig. 3.2 shows. We collected information published by these companies on their web-sites to better understand their business models. For example, we investigated the data that they trade, how they collect and manage it, whom they sell it to, exactly what they provide to customers, and how they deliver and price their services. Furthermore, we collected information about when these companies were founded (half of them in 2016 or later) and how many employees they have (40% with less than 20 employees).

Most companies in our sample are either *scaling* their customer base (29) or are in *commercial* development stage (61). In addition, we have included *developing* companies working in new innovative concepts around IoT, personal and ML data, or integrating blockchain in decentralised architectures (e.g., DataBroker/Settlemint and Dataeum). Finally, we chose not to include any *open data* providers and repositories, but instead focus only on those offering paid data products.

Table 3.2: List of entities selected for in-depth study (links accessed: Dec'22)

| | | | |
|---|---|---|---|
| 1DMC | Data Republic | HERE | Qiy Foundation |
| Advaneo | DataPace | HxGn Content | Quexopa |
| Airbloc | Datarade | Intrinio | Refinitiv |
| Aircloak | DataScouts | IOTA | Salesforce |
| AMO | Datasift | Knoema | S&P Global DM |
| ArcGIS DM | Datavant | Kochava | SAP data marketplace |
| Atoka | DataWallet | LemoChain | SayMine |
| AWS Marketplace | Datum | LiveRamp | Skychain |
| Azure | Dawex | LonGenesis | Snowflake |
| BattleFin | Decentr | Lotame | Streamr |
| Benzinga | DefinedCrowd | Madana | Swash |
| Bloomberg EAP | Demyst | Meeco | TelephoneLits.Biz |
| BookYourData | dHealth Network | MedicalChain | Terbine |
| BronId | Digi.me | Mobility Data Marketplace | The Adex |
| BurstIQ | Enigma | Multimedia Lists | TheTradeDesk |
| Carto | ErnieApp | mydex | Sales Leads |
| Caruso Dataplace | Factset | Narrative | TAUS data marketplace |
| citizenme | Factual | Nokia Data Marketplace | v10 data |
| Cognite | Fysical | Ocean Protocol | Veracity |
| Convex | GeoDB | OpenCorporates | Vetri |
| Crunchbase | Google Cloud Marketplace | Openprise | Vinchain |
| Cybernetica | GXChain | Oracle DMP | Webhose.io |
| datablockchain.io | Handshakes | OSA Decentralized | Weople |
| Databroker | HAT | Otonomo | Wibson |
| Dataeum | Health Verity | People.io | Xignite |
| Data Intelligence Hub | HealthWizz | Quandl | Zenome |




(a) Entities by country (b) Entities by year of foundation (c) Entities by nº employees

Figure 3.2: Summary of entities included in the survey

### 3.2.3. Data trading business models

First, we found that the business models of DTEs heavily depend on who they consider their customers to be, which in turn depends on which side of the chain they approach data trading from. *Data management systems* (DMS) focus on managing the information an enterprise or individual owns. Conversely, traditional *data providers* (DP) focus on data consumers, and conceal data owners and often even their partners when selling their products. Whereas the former approached data trading in order to allow secure data exchanges within an organisation or to authorise third parties, the latter implemented data trading platforms to complement their existing products or services with those of third parties. In addition, *data marketplaces* (DM) were conceived from the beginning as two-sided platforms dealing both with buyers and sellers.

Table 3.3: Taxonomy of data trading business models

| | **Data Providers (DP)** | **Data Marketplaces (DM)** | **Data Management Systems (DMS)** |
|---|---|---|---|
| End-to-end DTEs | Service Providers (SP) | General-purpose DM | Embedded DM |
| | Data Providers (DP) | Niche-DM | PIMS |
| | Private marketplaces (PMP) | | Survey PIMS |
| Enablers | | DM enablers (DME) | PIMS-enabler (PIMS-E) |

Within the scope of the survey, we included 41 DMs and 25 DMSs. As far as DPs are concerned, they often provide their products in commercial DMs, and we managed to list 2,015 of them selling their products in a sample of nine public or semi-private DMs. Hence, they are by far the most frequent business model we found. Since the way they operate is often similar, we took a diversified sample of 38 to understand how they deliver data and how they price their products. We dug deeper into the characteristics of the entities belonging in each category, and identified up to ten different business models. which we summarise in Table 3.3.

Interestingly, we found that some PIMSs and DMs only implement partial data trading functionality. Such *enablers* provide a range of solutions that includes, for example, anonymising personal information (AirCloak), providing an homogeneous anonymised identity to buyers (Datavant), facilitating secure exchanges (Cybernetica), or empowering individuals to exert their rights on the information that data providers hold about them (Saymine). When it comes to charging and billing, enablers usually charge for transactions (e.g., calls to the API, volume of data processed, etc.). Even though some enablers focus on specific types of data (e.g., IoT-related, data for ML models, personal data), or industries (e.g., health), we do spot some *general-purpose* enablers as well (e.g., those providing secure data exchange of distributed data).

With regards to entities providing full-fledged seller-to-buyer solutions, Table 3.4 summarises their characteristics (in rows), the section the topic is dealt with, and the differences between their business models (in columns), which we further explain in the next sub-sections.

Table 3.5 summarises the list of entities trading data that were analysed in depth in the survey, including their business model. For bigger companies such as SAP or Oracle, the business model reflects the role of their data trading solutions.

Table 3.4: Summary of business models

| | **Data Trading Entities (DTE)** | | | | | |
|---|---|---|---|---|---|---|
| | **Providers** | | **Data Marketplaces (DM)** | | **Data Management Systems (DMSs)** | |
| **Concept (Sect.)** | **DP/SP** | **PMP** | ***General-purpose* DM** | ***Niche* DM** | ***Embedded* DM** | **PIMS** |
| **Data exchange (3.2.3)** | Public, semi-private, private | Private | Public / Semi-private | | Private | Public / Semi-private |
| **Scope (3.2.4.1)** | Focused | | Diversified | Focused | | |
| **Type of data (3.2.4.1)** | Any | Specific data to be used within their service / platform | Any | Industry or type-specific | Data exchanged within the system | Personal data |
| **Roles / Players interacting (3.2.3)** | Partners, Customers | | Sellers, buyers | | Owner, requester | Users, data Providers, buyers |
| **Gets data from (3.2.4.7)** | Internet, self-generated, partners, users | Partners, Data providers | Data providers | Data providers, self-enriched | Data providers | Users, Data providers |
| **Provides buyers with (3.2.4.7)** | API, datasets | API, access to data through the system | API, datasets | | API, Access to data through the system | API, Key to decrypt data |
| **Owners get access through (3.2.3)** | Partnership | Partnership, the service platform | Web-services | | Data Management platform | Mobile App Web services |
| **Buyers get data through (3.2.4.7)** | Web-services, APIs | Web-service, the service platform | Web-services | Web-services, APIs | Data Management platform | Web-services, APIs, compatible systems |
| **Access pricing for buyers (3.2.4.6)** | 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 (3.2.4.6)** | Partnership (when applicable) | Partnership, time subscription | Predominantly free. Some freemium subscription, and revenue-share charges | | Subscription to the platform | Free |
| **Data pricing schemes (3.2.4.3)** | Fixed one-off, subscription, customised, volume-based | Subscription, domain-specific (e.g., cost per click, cost per 1,000 impressions) | Fixed one-off, subscription and customised | Customised, volume/usage-based, fixed one-off | Open | Open, bid by buyer |
| **Data price set by (3.2.4.4)** | Platform | Platform, buyers | Platform, providers | | Open | Users, Platform |
| **Payment (3.2.4.5)** | Fiat currency | | | Fiat currency, token | Open | Token, fiat currency |
| **Platform type (3.2.4.8)** | Centralised | | Centralised or decentralised | | Centralised | Decentralised |

Table 3.5: List of entities included in the survey and their business model (links accessed: Feb'23)

| Entity | Bss. Model | Entity | Bss. model | Entity | Bss. Model | Entity | Bss. model |
|---|---|---|---|---|---|---|---|
| 1DMC | DM | Data Republic | Emb. DM | HERE | PMP | Qiy Foundation | DME |
| Advaneo | DM | DataPace | DM | HxGn Content | DP | Quexopa | DP |
| Airbloc | PIMS-E | Datarade | DM | Intrinio | DP | Refinitiv | PMP |
| Aircloak | DME | DataScouts | DP | IOTA | DM+DME | Salesforce | DM |
| AMO | DM | Datasift (Fairhair) | SP | Knoema | DM | S&P Global DM | PMP |
| ArcGIS DM | PMP | Datavant | DME | Kochava | PMP | SAP DM | Emd. DM |
| Atoka | DP | DataWallet | PIMS+DM | LemoChain | DME | SayMine | PIMS |
| AWS | DM | Datum | PIMS+DM | LiveRamp | PMP | Skychain | DM |
| Azure | DM | Dawex | DM | LonGenesis | DM | Snowflake | Emd. DM |
| BattleFin | DM | Decentr | PIMS+DM | Lotame | PMP | Streamr | DM |
| Benzinga | DP | Defined Crowd | DP | Madana | DM | Swash | PIMS |
| Bloomberg EAP | DP | Demyst | DM | Meeco | PIMS-E | TelephoneLists | DP |
| BookYourData | DP | dHealth | DME | MedicalChain | PIMS-E | Terbine | DM |
| BronId | SP | Digi.me | PIMS-E | Mobility DM | DM | The Adex | PMP |
| BurstIQ | DM | Enigma | DP | MMedia Lists | DP | TheTradeDesk | PMP |
| Carto | PMP | ErnieApp | Surv. PIMS | mydex | PIMS+DM | Sales Leads | DP |
| Caruso | DM | Factset | PMP | Narrative | DM | TAUS DM | DM |
| Citizenme | Surv. PIMS | Factual | SP | Nokia DM | DME | v10 data | DP |
| Cognite | Emd. DM | Fysical | DP | Ocean Protocol | DME | Veracity | DM |
| Convex | DM | GeoDB | PIMS+DM | Open Corporates | DP | Vetri | PIMS+DM |
| Crunchbase | PMP | Google Cloud | DM | Openprise | PMP | Vinchain | SP |
| Cybernetica | DME | GXChain | DME | Oracle DMP | Emd. DM | Webhose.io | DP |
| Datablockchain | DME | Handshakes | DP | OSA Decentr. | SP | Weople | PIMS+DM |
| Databroker / Settlemint | DM | HAT | PIMS-E | Otonomo | DM | Wibson | PIMS+DM |
| Dataeum | PIMS+DM | Health Verity | DM | People.io | Surv. PIMS | Xignite | DP |
| DIH | DM | HealthWizz | PIMS | Quandl | DM | Zenome | DM |

#### 3.2.3.1. Providers

We consider two types. *Data Providers* (DPs, aka vendors [163]) are entities that provide *data* as a product, be they raw or enriched data, access to information through a graphical user interface, or information contained in reports to third parties. They usually combine data from different sources (e.g., from the public Internet, from partners, or from other providers) to enrich their products and add value to their offer.

*Service Providers* (SPs) are entities providing digital services to end users, be they individuals or enterprises, based on data they own, or on that which they collect from the Internet, or acquire from third parties. Examples of them are Clearview.ai, a company that provides identity data based on pictures of people publicly available on the Internet, or Factual, which offers marketing insights based on the movement of people. The boundaries between data and service providers are often blurry: are not personal identifications provided by Clearview.ai or insights by Factual data in the end?

From our point of view, supply side platforms and demand side platforms are SPs in the online marketing industry. The former allow publishers and digital media owners to manage and sell their ad spaces, whereas the latter allow advertisers to buy such advertising space, often by means of real-time automated auctions. Also in online marketing, data management platforms (DMPs) refer to audience data management systems that allow advertisers to enrich their audience data with that provided by the DMP. Some marketing-related SPs (Liveramp, Lotame, Openprise, among others) are integrating *private marketplaces* into their platforms to allow secure exchanges, monetisation, trading and integration of audience data from trusted partners (among them the so-called *data brokers*) within the platform. Such marketplaces are frequently an add-on to DMP subscriptions, and therefore can only be accessed by their users.

Despite the fact that the term *PMP* often refers to data trading platforms operated by marketing-related service providers, similar business models have also flourished in trading geo-located data (Carto, Here), business technographic data (Crunchbase), and financial data (Factset, Quandl, Refinitive). They all provide their users with a marketplace to enrich their data in the platform with relevant second-party and third-party data. As opposed to public or semi-private DMs, data exchange in PMPs is a *private functionality* of data and service providers that complements their main value proposition, and hence is only accessible by their customers on the buy side, or authorised data partners on the sell side.

Interestingly, as well as directly commercialising their services through their websites, DPs and SPs also use intermediaries to advertise their services, provide access to free samples of data, or offer specific data products. We found that 45% of data brokers (like Experian, Acxiom or Gravy Analytics) that offer their products through marketing-related PMPs (e.g., Liveramp, TheTradeDesk or LOTAME) commercialise their products in other DMs such as AWS or Data-Rade, too. This is also the case with providers such as RepRisk, Equifax or Arabesque S-Ray, which make use of financial PMPs.

#### 3.2.3.2. Data Marketplaces

DMs are mediation platforms that put providers in touch with potential buyers, and manage data exchanges between them. Such exchanges usually involve some kind of economic transaction, as well, either through payments in fiat currency or in a cryptocurrency often created and controlled by the platform. DMs are either public - i.e., open to any data seller or buyer - or semi-private, meaning any seller or buyer is subject to the approval of the platform in order to be allowed to trade data. Furthermore, DMs often deal with data categorisation, curation and management of metadata to help buyers discover relevant data products.

Whereas *general-purpose* DMs like AWS, Advaneo or DataRade trade any type of data, *niche* DMs are focused on certain industries (martech, automotive, energy) and on certain types of data (spatio-temporal data, or that coming from IoT sensors). By analyzing the date of foundation, we spotted a clear trend towards real-time data streaming marketplaces to harness the potential of IoT (e.g., IOTA, Terbine), and those specialised in training ML models (e.g., Skychain, Ocean Protocol), both very active lines of scientific research.

The large number of identified marketplaces, each one having proprietary on-boarding processes, access protocols/APIs and user-interfaces, makes it challenging for data providers to establish presence in all of them and thus reach the widest possible audience. The fragmentation of the DM ecosystem calls for establishing inter-operability standards that will allow different marketplaces to federate (*Challenge 1*). Sellers and buyers are often invited to subscribe for free to the platform. However, some platforms charge for freemium subscriptions or charge IaaS-like fees for delivering data. A few of them opt for charging sellers according to the money they make through the platform, either through commissions or revenue sharing.

In addition, buyers oftentimes pay marketplaces for data. Both the data seller and the platform are in charge of setting the prices for data products - in most cases one-off charges for downloading or gaining access to datasets, or periodic subscriptions to data feeds in *general-purpose* DMs. Conversely, *niche* DMs more frequently resort to volume or usage-based charging for APIs, and price customisation depending on who the data buyer is and on what the purpose of purchasing the data is.

Some DMs build on top of *data marketplace enablers* (DMEs). For example, Ocean Protocol provides marketplace functionality for ML data trading, whereas GeoDB and Decentr are DMs that use Ocean Protocol.

#### 3.2.3.3. Data Management Systems

On the one hand, enterprise DMSs are increasingly offering add-ons to carry out secure data exchanges within an organisation, and to enrich its corporate information base by acquiring data from second or third-party providers. Such *embedded DMs* - meaning they are built in an already existing DMS - rarely include full marketplace functionality, but rather restrict themselves to securing data exchanges, and to controlling the delivery and access to data assets within the

walled-garden of information under their control. Some of them charge IaaS-like fees for delivering data, and a recurring subscription fee to authorised sellers.

On the other hand, *PIMSs* look to empower individuals to take control of their personal data, and act as a single point of control to manage them. They leverage recent data protection laws so as to let users collect personal information controlled by digital service providers, exercise their erasure or modification rights as granted by law, manage permissions of mobile apps to give away their data, manage cookie settings, etc.

In addition, some of them seek their users' consent to share their personal information with third parties through the platform in exchange for a reward. Almost half of the surveyed *PIMSs* include marketplace functionalities, and focus on trading personal data for marketing purposes such as user profiling and ad targeting. Therefore, they leave data subjects (as the owners) and data providers to negotiate fees for consenting to get access to their data. This way they become personal data brokers, letting users monetise their data, and controlling who has access to it and for what purpose.

Recently, health-related *PIMSs* (Longenesis, HealthWizz, MedicalChain) specialise in managing healthcare-related information of their users. We found that health-related *PIMSs* often resort to blockchains to provide additional security to such sensitive data, and comply with a strong sectorial regulation. However, it is unclear whether and how their solutions protect against data replication and distribution off the chain.

Finally, *survey PIMSs* (e.g., Citizen.me, ErnieApp or People.io) aim to facilitate targeted marketing surveys among their users, leveraging information about their profile to offer an accurately targeted audience, and rewarding users for participating in the processes.

As opposed to enterprise DMSs, *PIMSs* are more decentralised platforms that often leverage the users' devices to store information, and they are always offered for free to individuals. Some charge one-off fees, subscription, or data delivery fees to potential data buyers. Given the relevance of PIMS in a human-centric data economy, we explain them in more detail in Sect. 3.2.4.7.

### 3.2.4. Results of the survey

Having characterised qualitatively the business models of DTEs, this section takes a closer quantitative look into crucial aspects of data trading. In the following sections, we tackle questions related to:

- the kind of data being traded (Sect. 3.2.4.1),
- the source and the target of DTEs (Sect. 3.2.4.2),
- pricing schemes (Sect. 3.2.4.3) and
- responsible parties to set the prices (Sect. 3.2.4.4),
- payment methods (Sect. 3.2.4.5),

- billable concepts and charging (Sect. 3.2.4.6),
- business models used to trade data (Sect. 3.2.4.7),
- type of storage and architecture (Sect. 3.2.4.8),
- how buyers can test data (Sect. 3.2.4.9), and
- general data security issues (Sect. 3.2.4.10)

#### 3.2.4.1. What kind of data is being traded?

Very different kinds of data are being traded in the market. In fact, DTEs are often classified based on the kind of data they trade. For example, we will talk about *marketing DPs* or *marketing PIMS*, meaning DTEs specialised in providing data or managing and trading personal information for marketing-related purposes. We will also discuss the aforementioned *general-purpose* DTEs aiming to trade *any* kind of data.

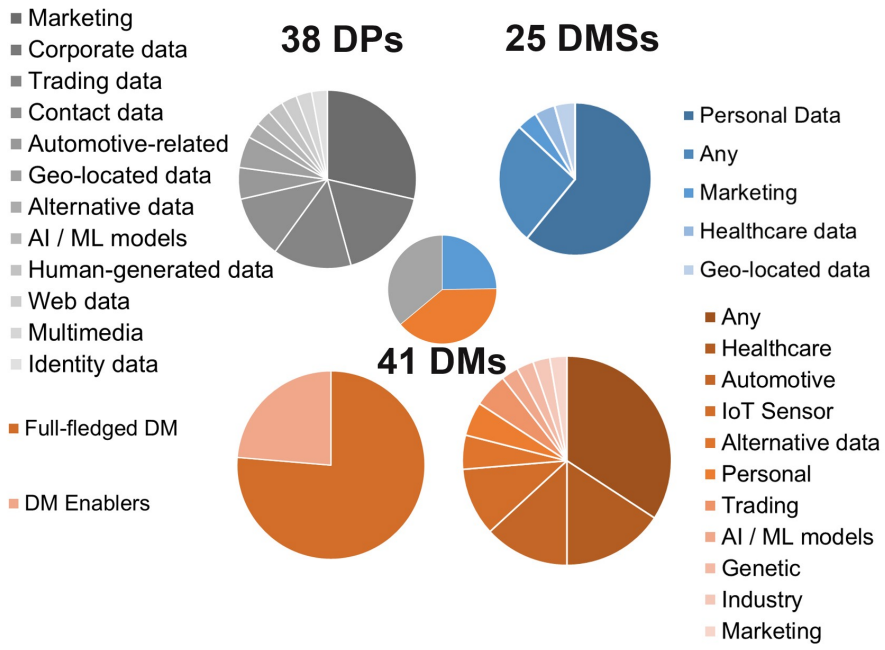



Figure 3.3: Data trading entities and the kind of data they trade

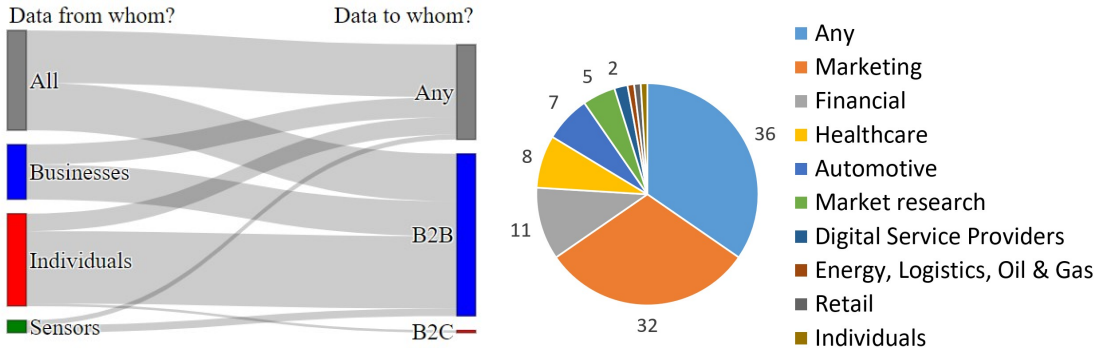
Figure 3.3 shows a breakdown of the kind of data traded by DMSs (in blue), DMs (in orange) and DPs (in grey). There are notable differences in what kind of data entities do trade depending on their business model.

- DPs specialise in a market *niche*, either a specific type of data or a customer segment. Only one of them (Quexopa) is publicly focusing on data from a certain region (Latin America). Even though the range of data DPs deal with is diverse, it turns out that most providers in our sample are related to marketing, corporate, contact or financial data.

- Within DMSs, *PIMSs* focus on personal and healthcare-related data, whereas business-oriented DMSs are usually designed to trade different types of corporate data.

- With regards to DMs, at least 14 of them are *general-purpose* and trade *any* kind of data, whereas *niche* DMs deal with healthcare, automotive, IoT-related, trading or alternative investment data.

#### 3.2.4.2. Data from whom? Targeting whom?




(a) Flows of data in commercial marketplaces

(b) Target of data trading entities

Figure 3.4a shows where or from whom data trading entities get their data from and to whom they intend to sell their data and their services. Data may come from different sources, such as PI owned by individuals, data related to companies, industries, measurements from sensors, etc. However, most data trading entities are designed to trade data from *all* these sources. Even though most entities are clearly oriented to the business market, and we can state that most data nowadays flows to enterprises, no restriction seems to prevent individuals from also acquiring data. DTEs usually target specific industries, and often specific departments within their business customers, which we summarise in Fig. 3.4b. Unsurprisingly, it is marketing and financial departments that DTEs' data and services are most often targeting, with more and more marketplaces oriented to healthcare and to specific industries lately.

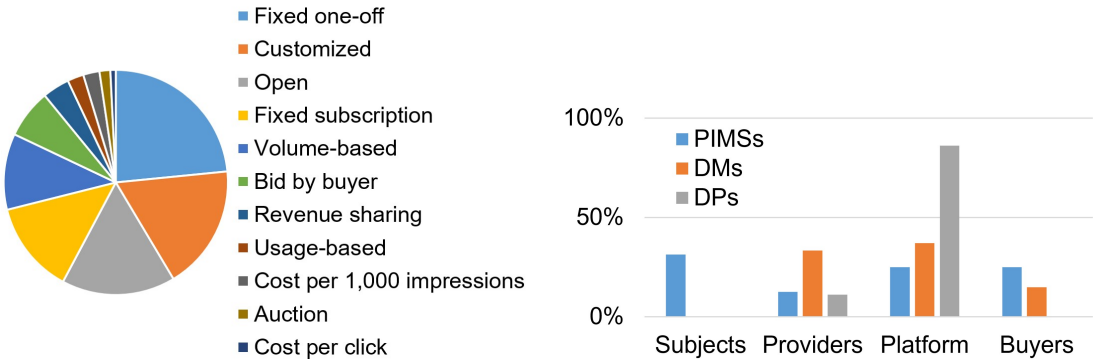
In the next two sections, we discuss about the pricing schemes data trading entities are using for pricing data products, and who is in charge of setting the prices.

#### 3.2.4.3. How is data being priced?

Figure 3.5a provides a summary of the most widely adopted pricing mechanisms. Our conclusions are in line with the current state-of-the-art [153]:

- Most buyers pay a lump-sum as a **fixed one-off** for a dataset, or a **fixed subscription** charge for accessing a stream or a service for a period of time.

- **Customised**. Price (and the product) is personalised depending on who the buyer is and what the data is intended to be used for.




(a) Data pricing mechanisms

(b) Stakeholders setting the prices of data

Figure 3.5: Data transaction pricing

- **Open**. The platform lets buyers and sellers agree on the prices, and eventually consent to the data exchange. This is the preferred approach of *enablers* that focus on facilitating data exchanges but do not involve themselves in setting their economic terms.

- **Volume-based**. Price directly depends on the volume of information that is downloaded or accessed (e.g., contacts, images), often with volume discounts.

- **Bid by buyer**. Buyers place bids (e.g., in a *PIMS*) that sellers must accept for the transaction to take place. This avoids sellers deciding on an upfront asking price.

- **Usage-based** pricing is frequently used for API calls and offered in tiers. Charges include volume discounts and depend on the number and type of calls.

In addition, we identify some other interesting mechanisms being used in specific contexts.

First, MyDex used to charge transactions using **revenue sharing**: when a buyer purchased the rights to access a user's personal data, the platform claimed its rights to 4% of the revenues that such a buyer made on the platform from that individual. Digi.me, a *PIMS enabler*, also mentions this pricing scheme. Although innovative in terms of pricing data, its feasibility is still to be proven: would a *PIMS* be able to control how much money buyers are making from each user and charge accordingly? None of them are using this scheme now.

**Cost per 1,000 impressions**, **cost per click** and **percentage of gross media expenses** are specific to PMPs of online advertising platforms (e.g., LiveRamp, Oracle or Kochava), thanks to their end-to-end control of online ad campaigns.

Finally, **auctions** are very popular price setting mechanisms in other fields, and they are widely used in online advertising where advertisers bid in real time to show their ads to a user browsing a certain web page [150]. Nonetheless, they are not so common when selling data, due to its non-rivalrous nature. Even though some works of research have already defined a whole family of auctions that artificially creates competition among interested buyers [79,81], we found only one enabler (Ocean Protocol) that mentions auctions as a potential mechanism to set the prices of data products.

Other interesting takeaways from this analysis are that most data products in *general-purpose* DMs are made available for free, and that some of them such as DIH, Advaneo, and Google Cloud Marketplace lack any significant offer of paid products. We observe that these free data products are either open data from public repositories, or samples uploaded by data providers.

Surprising though it may seem in the case of entities whose aim is to make profit, marketplaces like DIH or Advaneo collect and link open data made available by authorities or public institutions. They give up on generating revenues from reselling paid data, and they monetise the effort to organise and facilitate the exploitation of free open data assets in other different ways. For example, some DMs offer free datasets as part of a subscription to the platform (e.g., Carto), whereas others charge for processing and integrating data within the platform (e.g., Advaneo, and those managed by cloud service providers). Such a vast amount of data may also serve as a '*hook*' for sellers and buyers, and as a complement to third-party paid data products.

Moreover, we find that some providers are making use of public marketplaces to upload outdated samples of their products so that buyers can manipulate them and get to know how useful the whole data product would be for their purposes. This practice would indeed be interesting for marketplaces, provided it was they who eventually sold the corresponding paid product after the trial. However, data providers usually refer interested buyers to their own commercial channels, and the host marketplace merely acts as a showcase for their products.

#### 3.2.4.4. Who sets the price of data products?

The answer to this question again depends on the business model, as Fig, 3.5b shows. Whereas providers tightly control the price of their data or services, *PIMSs* give more control to their individual users (the actual data subjects), and usually let them agree with buyers on transaction prices. Although DMs usually play an active role in setting the prices for data products on their platform, they always do it in conjunction with providers. In fact, some of them (Dawex, Battlefin) charge for advisory services in setting the prices for their data products. Such advisory services for pricing are empirically provided since developing a more rigorous methodology for data pricing remains an open challenge (*Challenge 2*).

#### 3.2.4.5. How do entities deal with payments?

Whereas data providers have traditionally been charged for their services in fiat money (dollars, euros, etc.), 55% of surveyed *PIMSs* and almost 40% of marketplaces are using cryptocurrencies instead. The benefits they seek by using this alternative include an increased speed of transfers, a higher availability if compared to going through banks or establishments, and a greater liquidity. Real-time data exchanges like the ones trading with IoT sensor data are broadly opting for cryptocurrencies when it comes to settle payments.

#### 3.2.4.6. How do data trading platforms charge users for accessing their services?

DTEs that operate as platforms do not only charge users for the data they consume, but for other concepts such as delivering data, or even just for gaining access to their services. Again, such additional platform charges vary greatly between business models.

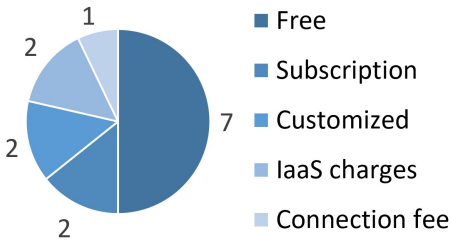
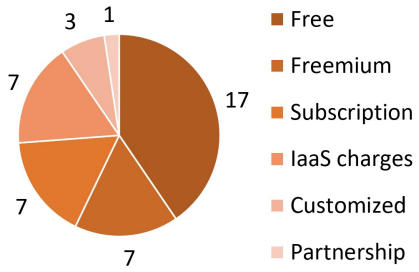
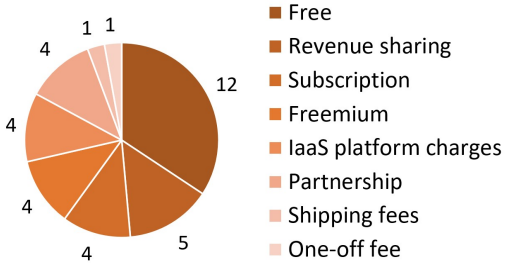



Figure 3.6: Charges to buyers accessing PIMS

In general, PIMS are free for data subjects. On the one hand, this makes sense since they provide the platform with PI to work with, and also make the promise of increased privacy and data protection more appealing. On the other hand, it raises concern over the profitability of users who are unwilling to share their data and are using PIMS for such purposes. Data buyers, who are also usually welcome and free to join the platform, often just pay for the data they acquire. In some cases, potential buyers are asked for a one-off connection fee or charged a periodic subscription (see Figure 3.6) to get access to the platform. Finally, some PIMS demand details about the buyer signing up to the platform to customise access charges.




(a) Charges to buyers




(b) Charges to sellers

Figure 3.7: Charges and pricing in data marketplaces

Conversely, charging buyers and especially data sellers for access is more usual in the case of DMs (see Figs. 3.7a and 3.7b), either through:

- time-based subscriptions, often using a freemium model;
- revenue sharing, where the platform keeps a percentage of the total sales;
- one-off fees to connect to the system.

A few *niche DMs* (Otonomo) and most PMPs offer partnership models to big data sellers, an *ad hoc* agreement to share data frequently used by DPs. Interestingly, a niche DM (Caruso) requires a partnership agreement to be signed by buyers, which requires their participation as shareholders if they are willing to use the platform.

Regarding PIMS and DM-enablers, they welcome full-fledged PIMS and DMs to use their technology and usually charge pay-as-you-go IaaS/PaaS-like fees based on the number of API calls or the volume of data they deliver. Some DMs (e.g., AWS or Snowflake) do charge data shipping fees to both parties, too.

#### 3.2.4.7. How do entities trade data?

Some specific characteristics of data, in particular its zero-cost replicability and its inherently combinatorial value, make this attractive asset considerably more difficult to be priced and safely traded [153, 164]. In economics, a good or service is called *excludable* if it is possible to prevent consumers who have not paid for it from having access to it. In addition, data is non-depletable and hence a *non-rivalrous* good: selling data to a consumer A does not prevent the owner from selling it again to another consumer B.

Entities trading data aim at somehow making it *excludable* and therefore a club good. Indeed, this is a key challenge in building a flourishing economy around data. This section provides some additional insights about how entities in our survey are attempting to achieve this goal. In the following subsections, we answer questions regarding where entities take data from, what they provide buyers with, and how users - both from the buy and sell sides - gain access to data.

Since the conclusions are very different for DPs, DMs and PIMS, we present them separately. Given their novelty and the importance of the latter in a human-centric data economy, we will pay more attention to these platforms empowering individuals to take control of their data.

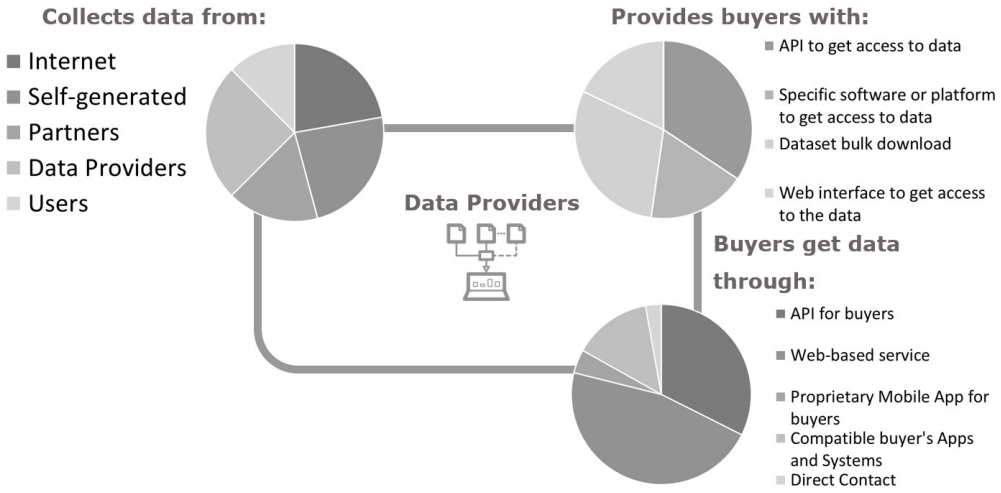



Figure 3.8: How do data providers work?

**Data Providers**. As Fig. 3.8 shows, DPs leverage the internet and access to exclusive self-enriched data sources to provide buyers with access to data either through APIs or bulk downloads, and preferably through web-services or specific applications. Note that they are not meant to be two-sided platforms, but players oriented to provide their data or their data-driven services to their customers. Should they require proprietary information from third parties, they establish partnerships or bilateral agreements to access such exclusive information, which they eventually enrich and resell. Therefore, DPs control the whole go-to-market process, and conceal the identity of their partners and the sources of their information, unless disclosing them adds any value (e.g., credibility) to their business and hence helps with sales.

As an exception, PMPs integrated in data-driven services (e.g., spatio-temporal data marketplaces integrated in GIS cloud SPs) allow third-party DPs to sell data within their platform. Unlike DMs, PMPs carefully select authorised DPs, who often sign private partnership agreements with them. Moreover, they deliver data to be used within their system or services, and only to their users, which is why such marketplaces qualify as *private*.

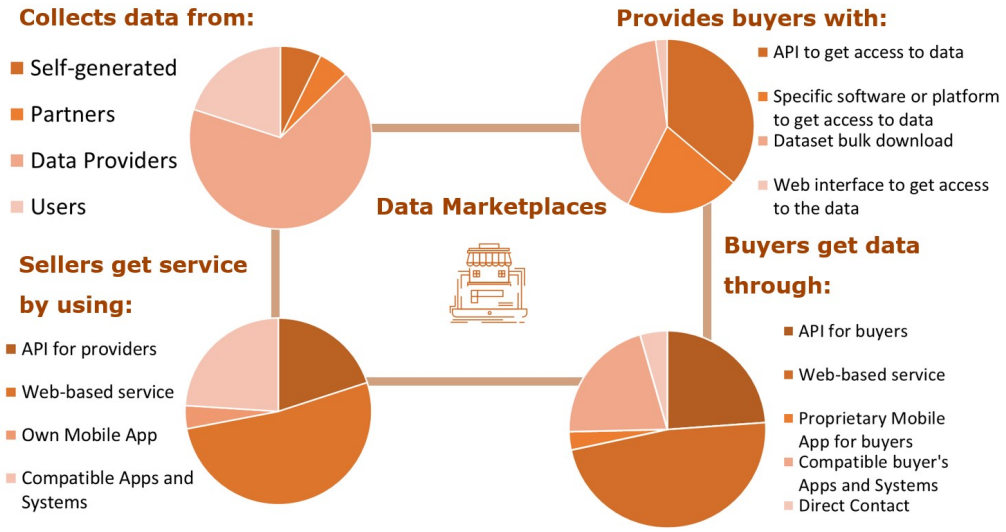



Figure 3.9: How do surveyed data marketplaces work?

**Data Marketplaces**. Figure 3.9 shows that DMs collect and sometimes enrich or combine data from different DPs (sellers), who have signed the DM's public terms of use. Similar to DPs, data is often delivered to buyers as a bulk download or through APIs. Although some of them restrict delivery methods to get access to data through their platforms (e.g., AWS marketplace offers access to data stored in Amazon S3 services). they often resort to APIs and web services for users to manage their transactions and data within the system.

**Personal Information Management Systems (PIMS)**. Figure 3.10 shows a diagram of how PIMS work. They deal with three types of users. First, *data subjects* are the individuals whose personal data is managed by the PIMS, which also help them exert their rights to erase, modify or retrieve their personal data collected by third parties like social networks or e-mail services. PIMS

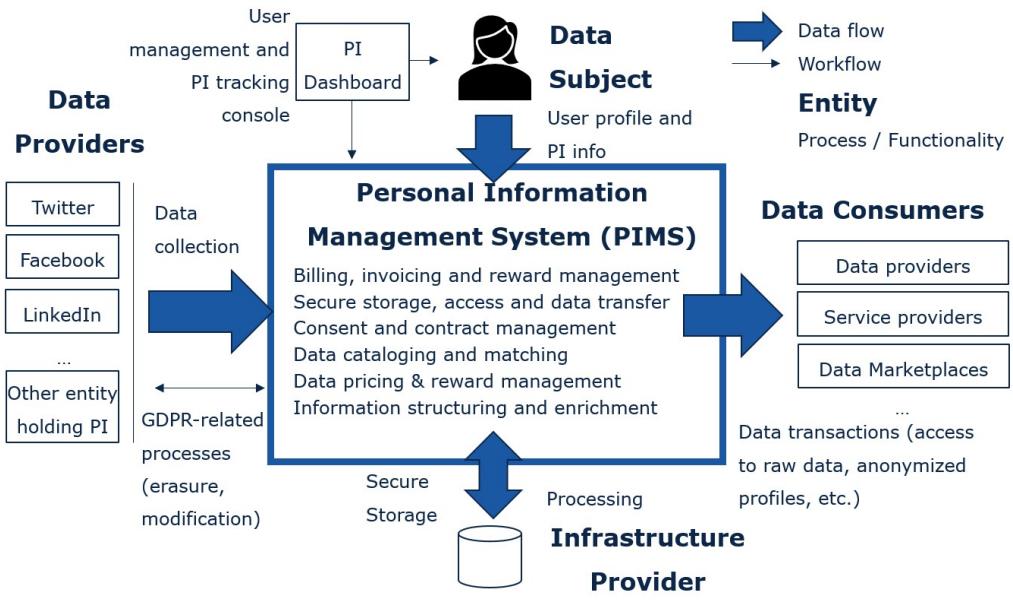



Figure 3.10: Diagram and functions in a PIMS

usually call such third parties their *data providers*, and refer to data subject as *users*. Individual users can share their personal information with the platform, as well, and all their personal data is usually stored in their devices. Finally, PIMS may share and some of them even sell personal data of individuals to potential *data consumers* upon the owners' consent. Thanks to their knowledge of data protection laws, some PIMS (e.g., Wibson) offer data providers and consumers, in most of the cases enterprises, advisory on the compliance with this legislation.

Therefore, PIMS are complex platforms that involve a number of processes, including secure storage, access and data transfer, execution of data protection rights with different data providers, cataloguing and structuring personal data, which they also process, enrich and aggregate. They manage the consent of users to share or sell their data with third parties for different purposes, and they may implement marketplace functions (data cataloguing, pricing, transaction management, billing, invoicing, and managing rewards to end users).

Most frequently, PIMS users get access to the platform through a mobile application, which also allows them to manage their consent to share their PI and monitor data transactions, as Fig. 3.11 shows. Most of them provide buyers with APIs or web services to gain access to data. As opposed to DPs and DMs, some PIMS ask entities willing to gain access to acquired data to integrate their apps and systems (MyDex, GeoDB, DataWallet).

PIMS deliver data in technologically innovative ways. For example, some of them provide access to encrypted data streams by sending temporary keys that are revoked once the subscription expires. Some of them resort to hashed temporary URLs to provide buyers with access to data for a certain period. Some PIMS still do not provide any automated platform for buyers to get data, but instead they negotiate directly with them, and generate the data to be shared case by case.

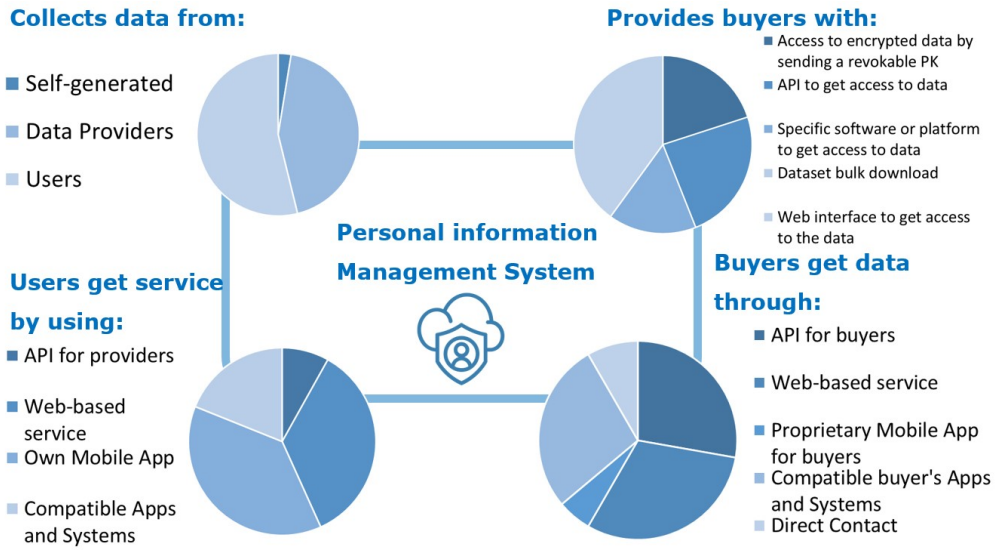



Figure 3.11: How do PIMS work?

### 3.2.4.8. How do entities store data products?

*PIMSs* usually opt for a more decentralised architecture by leveraging data subjects or providers to store and process users' data. With some exceptions, they avoid making copies of personal information, which is retrieved from the users' personal data storage, instead. On the contrary, DPs and DMs have traditionally preferred a centralised architecture.

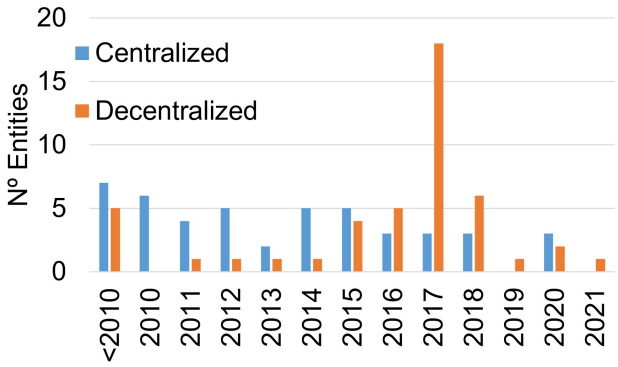



Figure 3.12: Data management architecture in time

Figure 3.12 shows a trend towards decentralisation of the storage of both data products and transactional data. In fact, distributed ledger technologies (DLTs) are increasingly being used to store transactional or management data related to data trading. Due to the high cost of storing data in a DLT, it is not yet being considered as a alternative for storing data for sale, except for specific concept models and developing prototypes related to healthcare (BurstIQ, MedicalChain), marketing (Datum, already closed) and automotive (AMO), whose feasibility is yet to be proven.

#### 3.2.4.9. Can buyers "try" before buying?

Buyers cannot be sure about the true value of a dataset before they get access to it and they demand ways to do that, e.g., by training a ML algorithm and test its resulting accuracy/performance. This chicken-and-egg problem, known as Arrow's information (or disclosure) paradox, often deters potential buyers from actually purchasing data. A big challenge for the data economy is thus to come up with ways to reduce the uncertainty for buyers (*Challenge 3*) [169]. Next we look at how surveyed entities approach this open challenge.

We found that 69% of entities answer this question on their websites, which reflects this is indeed an important issue for them. They claim to be using one or more of the following mechanisms: publishing or sending in advance **free samples** of data to potential buyers; allowing **free access** to part of the data (e.g., some fields of a structured data base); offering a **trial period** in which to have access to a data feed or subscription-based service; providing buyers with a **sandbox** (Battlefin, Otonomo) that lets them experiment with real data before bidding for it or making a purchase decision; offering a live **demo** of their services and their data products; or hosting a **reputation mechanism** for buyers to rank data and providers.

#### 3.2.4.10. What security measures are taken?

*PIMSs* and DMs often publish high-level information related to security of data and data exchanges to gain the trust of potential users, be they buyers or sellers. Unsurprisingly, it is *PIMSs* that express the greatest concern about it: most of them include a section dedicated to data security, and address users' and buyers' frequent concerns in this respect. On the contrary, DPs are generally reluctant to give away any information about security, which is considered an internal policy.

Some of the measures taken to secure data include: user authentication and identification; TLS encryption; Anonymisation or de-identification of personal information; delivery through revokable DLT decryption keys or public-key cryptography allowing buyers to decrypt an encrypted data stream or dataset; use of temporary URL to deliver data; secure data connectors; tamper-proof data delivery through data signatures and message chaining, which sometimes make use of a blockchain to ensure immutability; or use of specific service and software that certifies the origin of data.

However, DMs still fail to provide a fully effective solution to avoid unauthorised data replication, and protecting data ownership was found to be a key challenge (*Challenge 4*). Moving from providing data to providing services has traditionally been the most commonly accepted recipe to mitigate this risk [164]. For example, ML *niche* DMs look to sell model training services [50], rather than bulk data to train models as *general-purpose DMs* do.

DMSs and PMPs sell data to be used within their systems and services, and heavily restrict outgoing data flows. Extending the scope of controlled environments like *embedded* DMs might be a means to impose severe barriers to data replication and enhance the control of the access

to data. Still it needs to be proven whether an open version of such a controlled trust model can be scaled and bootstrapped to the entire Internet as standards like the International Data Spaces [18] and initiatives like Gaia-X [72] are aiming to on a smaller scale. In the next section, we provide more details on these two initiatives. Related to these, the University of Berkeley has also started the sky computing initiative aimed to standardise the interconnection between existing cloud computing platforms, but it does not deal with transparent data exchanges [38].

## 3.3. Standardisation efforts: IDSA and Gaia-X

Both Gaia-X and International Data Spaces (IDS) architectures look forward to shifting from a centralised hub to an open-network or federated data sharing ecosystem for two parties to exchange data online [51] (see Fig. 3.13). They follow the global trend towards distributed data marketplaces we have spotted in our survey, and they advocate consumers accessing data stored by data providers and they aim to build a federation of stakeholders playing different roles according to well-defined rules in a data ecosystem based on trust.

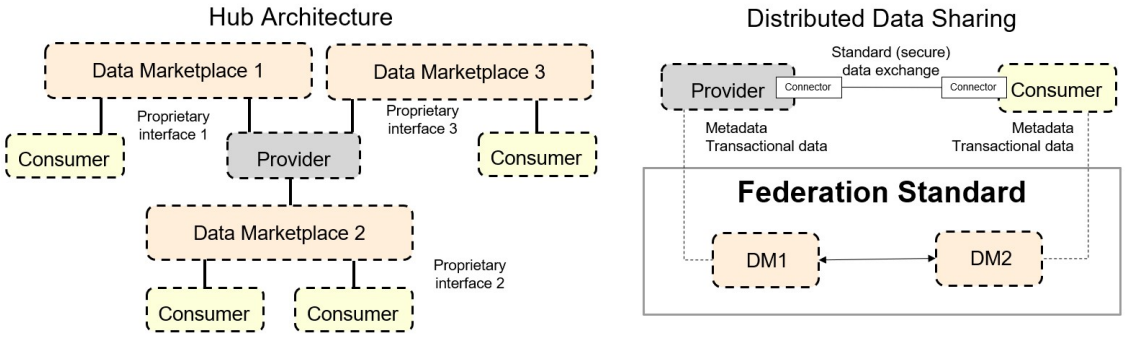



Figure 3.13: Centralized hub vs. distributed data sharing

Federated data exchanges overcome an important drawback of communities of hub or centralised marketplaces, namely the need for data providers to deal with different governance models and interfaces, which requires repeating integration efforts. In fact, this was found to be a challenge for both data providers and consumers in Sect. 3.2.4.1, due to the large number of different existing commercial data marketplaces. On the contrary, a federated model sets up some rules, protocols and standards so that any provider or consumer willing to join the federation is able to do so by following the governance rules and connecting through a single interface (usually known as “connector”) that adheres to the standard, and manages exchanges within the ecosystem.

For federated data exchanges to take-off, agreeing on a common data exchange standard is crucial to avoid ending up in different federated platforms and maximise the efficiency advantage of an open-network model for data trading entities. However, defining a standard that fits the wide range of data types, use cases, and industry verticals of the data economy is definitely a humongous task. The objective of this section is to provide an overview of two important

European initiatives in this direction involving significant industry players worldwide: the International Data Spaces (IDS) [18] and the Gaia-X [72] project. Following the notation used in both standards, we will capitalise and write in *italics* any concept and name defined by them.

### 3.3.1. International Data Spaces (IDS)

The International Data Spaces is a global standard defined by the International Data Spaces Association (IDSA), which groups more than 130 member organisations from 22 countries as of December 2022. All members "*share a vision of companies able to self-determine usage rules and realise the full value of their data in secure, trusted, equal partnerships*". They include research institutions, solution developers, data providers, service providers, and data consumers from dozens of industry verticals.

IDS reference architecture model provides an overview of the standard and high-level specifications for its different components. It defines IDS as "*a virtual data space leveraging existing standards and technologies, as well as governance models well-accepted in the data economy, to facilitate secure and standardised data exchange and data linkage in a trusted business ecosystem. It thereby provides a basis for creating smart-service scenarios and facilitating innovative cross-company business processes, while at the same time guaranteeing data sovereignty for data owners*" [15]. Essentially, data spaces define a standardised interoperable ecosystem of data exchanges based on trust in certified companies that are able to share their data while keeping full control of their assets. This concept of keeping data assets under control is referred to as *data sovereignty* in the standard.

Data *Connectors* are key basic components enabling secure exchange of data within the ecosystem. They allow for interoperability through Application Programming Interfaces (APIs), for ensuring the application of data exchange and usage policies, for implementing security measures, and for increasing trust through identification and certification of participants. IDSA offers an open source implementation of a basic *Connector*, and a number of additional implementations are being commercialised or openly shared by members of IDSA [17].

The standard is structured in five layers, and comprises three vertical perspectives across them, as shown in Fig. 3.14. It provides a standardised architecture and useful middleware to enable secure exchanges of data between certified entities within the ecosystem. It leverages an underlying cloud / edge computing and networking infrastructure, out of the standardisation scope. Next, we summarise each of these layers and relate them to our work and contributions.

The **business layer** defines the participants in the IDS, their roles and the basic tasks they are assuming in the ecosystem. Business models defined in 3.2.3 may play one or more roles in the IDS ecosystem. For example, full-fledged data marketplaces would play the role of *Data Broker* and *Clearing House*, whereas an "enabler" developing reusable components and services to exchange data according to the standard would be a *Service Provider*.

The **functional layer** defines the requirements to achieve the strategic objectives of IDS standard, and desirable features of its components. *Trust* is achieved by controlling the responsibilities

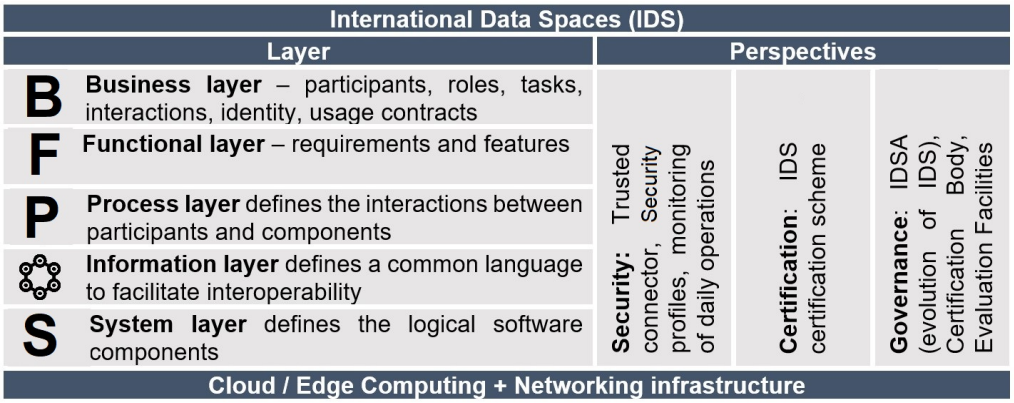



Figure 3.14: IDS layers and perspectives (based on IDS RAM v3.0 [15])

assigned to the different roles, by certifying participants and software components used in data exchanges, and by managing and ensuring the identity of participants. In particular:

- Data sovereignty is supported by authorisations of X.509 certificates attached to *Connectors* in the ecosystem, by data usage policies and their enforcement, by technical certification of components and by an underlying trustworthy communication to secure data on-the-move.

- The ecosystem of data is based on data descriptions, open brokering – meaning participants are able to subscribe or search for data in multiple data broker systems -, and the governance of standard vocabularies used in descriptions and during data discovery.

- Standardised interoperability is the basis of IDS, and it is ensured by certified *Connectors* running on top of the participants' IT environments, which are able to receive data from the enterprise back-end either through pull or push mechanisms, and to write it into the back-end of *Data Consumers*.

- *Connectors* are provided with *Value Added Apps* distributed by open publicly accessible *App Stores*. Such applications support data processing, transformation and enrichment workflows. The development and publication of *Data Apps* is open to developers adhering to the standard.

- *Usage Restrictions* and pricing can be defined by *Data Owners*, and IDS provides typical standard contracts to help in transactions and that can be automatically negotiated and enforced.

The **process layer** provides high-level diagrams for a number of interactions that can happen between components in IDS and follows the Business Process Model and Notation (BPMN) [144]. Such diagrams specify the different roles of participants in a process, the interactions between them, the inputs and outputs they provide to each other, and the general process workflow. The specification in IDS' *Reference Architecture Model* (RAM) 3.0 was done at a very high

level (Level 0, 1 and 2), and lacks details or instructions about how the different tasks are carried out. Regarding its scope, the standard specifies processes related to the onboarding of companies willing to join the ecosystem (identity acquisition, connector configuration and provisioning, security and availability setup), to the exchange of data (finding data providers and invoking data operations), and to certifying, publishing and using *Data Apps*.

The **information layer** defines the *Information Model*, a domain-agnostic common language designed to facilitate the interoperability in a data space. It allows to describe, publish, provide and identify data products and reusable *Data Apps*, often referred to as *Digital Resources*, and describes participants and components in the ecosystem. It is defined at three increasingly specific levels: conceptual, declarative and programmatic representations. Only the first one is defined in the IDS Reference Architecture Model. The IDS Onthology Draft and the IDS Information Model Library provide details on the declarative and the programmatic information model, respectively [13, 16].

The information model of IDS allows to thoroughly describe *Digital Resources* so that they can be automatically discovered, traded and used within the ecosystem. The concern hexagon defines key "*aspects of concern*" that must be taken into consideration and defined for this to happen. Those aspects are the features of metadata that a *Provider* must specify for its *Digital Resources* to be discoverable and traded in the IDS. For example, the *Content* concern deals with the nature of a *Resource*, which includes its concept, described independently of its physical or logical instantiation, its representation and its context. The *Commodity* concern helps assess the adequacy of a *Data Resource*, and provides information about its lineage, its quality, usage policy, and pricing. Different instances of the same data with different quality and pricing in the platform may be offered by the same *Provider* in the platform.

The Conceptual Information Model provides a description of each aspect of concern, and the different sub-aspects they cover, together with UML diagrams that explain the relationship between concepts in the description of *Data Resources*. IDSA is in charge of evolving the standard and its RDF representation.

The **System Layer** is the technical core of IDS and specifies the data and service architecture responding to the requirements of the *Functional Layer*. The three basic technical components are the *Connectors*, the *App Store* and the *Broker*, and they are supported by four other components: the *Identity Provider*, the *Vocabulary Hub*, the *Update Repository* (source of updates to any deployed *Connector*) and the *Trust Repository* (that carries out remote attestation checks and that is the source of trustworthy software stacks and fingerprints). The *System Layer* provides a reference *Connector* architecture using application containers for individual data services.

The **Security Perspective** is a central requirement of the IDS architecture, and critical to ensuring trust in the ecosystem. It includes functionality across the five layers and specifies a *Trusted Connector* that extends the *Base Connector*, and implements and monitors security in everyday operations. Four levels of trust in *Connectors* are defined. Even though they are interoperable, a *Connector* may refuse connecting to peers granting lower security standards.

Finally, IDSA has issued the DIN 27070 standard through the German Institute for Standardisation. As of April 2022, the standard is officially accessible in German only, and it provides the "*requirements and reference architecture of a security gateway for the exchange of industry data and services*", which corresponds to a *Connector* in the *IDS RAM*. The standard specifies the different security levels of *Connectors* (or *Security Gateways*), and the requirements of the architecture for complying with a *Secure Development Cycle* (SDL) as specified in DIN EN IEC 62443-4-2, and particularly for supporting the following groups of requirements of the ecosystem: communication integrity, data usage control, common information model, identity and access management, connection to broker services, operating system integrity, apps and app store connection, and transparency of data usage.

### 3.3.2. The Gaia-X project

The Gaia-X initiative was originally a German proposal launched in the Digital Summit on October 2019 [72]. The initiative intends to develop an open, transparent and secure digital standard ecosystem where data can be shared and services can be provided in an environment of *trust*. The Gaia-X Ecosystem is structured in three levels:

- Gaia-X Association,
- Gaia-X National Hubs, and
- Gaia-X open-source Community.

Gaia-X European Association for Data and Cloud AISBL (or Gaia-X Association), based in Brussels, is the core of the organisation, whose goals are developing the standard and operating Gaia-X Federation services. Established in September 2020 thanks to the joint effort of 22 companies and organisations (11 from Germany and 11 from France), the Association, based in Brussels, included 364 member companies as of December 2022,[1].

*National Hubs* [2] are the central points for interested parties in each country. They connect potential stakeholders to Gaia-X, scale up and spread use cases, coordinate and collaborate with the Gaia-X Association to develop standards, consolidate initiatives in countries, and ensure efficient cooperation within a particular industry. As of December 2022, there are 20 National Hubs of Gaia-X, most of them in Europe but also in other countries such as the US, Korea, or Japan.

The open-source *Gaia-X Community* provides a way for users and providers to collaborate with the Association on specific work packages. It is open for everyone who wants to join, and share their knowledge in a collaboration platform or in webinars. A *Technical Committee* supervises, steers work packages and ensures that the outputs are in line with the standard.

In the next sections, we summarise the architecture of Gaia-X, its operating model, and its data and infrastructure ecosystems.

[1] See Gaia-X members, last accessed Feb'23.
[2] see Gaia-X Hubs, last accessed Feb'23.

#### 3.3.2.1. Gaia-X Architecture

The Gaia-X architecture document provides a high-level description of the different concepts and components of this initiative. Similar to IDS, the high-level requirements for Gaia-X Architecture deal with interoperability and portability of data and services, sovereignty over data, security and trust. For that purpose, the architecture follows three key design principles: federation, decentralisation and openness. Also in line with IDS, a significant effort has been made to develop and catalogue relevant use cases for different industry verticals, involving key market players. Complementing IDSA's, Gaia-X architecture deals with infrastructure services, as well, and defines an infrastructure ecosystem that supports the data ecosystem with processing or interconnection services.

The Gaia-X architecture document summarises the key concepts and stakeholders in the Gaia-X Ecosystem [70]. Participants in the Federation are *Providers*, *Federators* and *Consumers*.

*Providers* provide *Resources* (be they *Data*, *Software*, *Nodes* for processing, or *Interconnection* - i.e., communication facilities - between *Nodes*) according to a *Service Offering*, that defines the terms and conditions of the services in *Resource Templates*, and make them available for ordering. *Providers* operate *Service Instances* that realise services of the offering and include *Self-Descriptions*.

*Federators* are in charge of providing *Federation Services*, which belong in four groups: *Identity and Trust*, *Federated Catalogue*, *Sovereign Data Exchange*, and *Compliance*. *Federation services* work as a connection layer between the *Infrastructure Ecosystem*, that focuses on exchanging processing and communication services, and the *Data Ecosystem*, that deals with data exchange and the provision of data-driven services.

*Consumers* search for *Service Offerings* that are relevant for their purpose, and consume *Service Instances* to enable their digital offerings for end users. The Gaia-X Association defines a common *Contract* model between participants in the ecosystem, including computable and non-computable elements.

The term *Policy* is widely used within Gaia-X architecture referring in general to statement of objectives, rules, or assertions that define and determine the behaviour of an entity in the ecosystem. There are *General Policies* defined by the Gaia-X Association (see the Policy Rules Document [71]) for all *Providers*, *Service Offerings* and *Contracts*, but the architecture supports particularisation, as well, including:

- *Provider* or *Usage Policies* (e.g., data can only be used for x days or y times), and
- *Consumer* or *Search Policies* (e.g., providers must comply with rule x or come from jurisdiction y).

#### 3.3.2.2. Gaia-X Architecture in practice

Let us make an example about how Gaia-X architecture would work in the context of a human-centric data economy. We will assume that Alice is working with a PIMS to manage her personal data within the Gaia-X Ecosystem and we will explain the complex relationship between players in this use case using the language of Gaia-X. We will also assume that all commercial platforms participating in the use case adhere to the standard. Figure 3.15 shows a diagram summarising the parties involved, their role, the service provisioning and data flows in this use case.

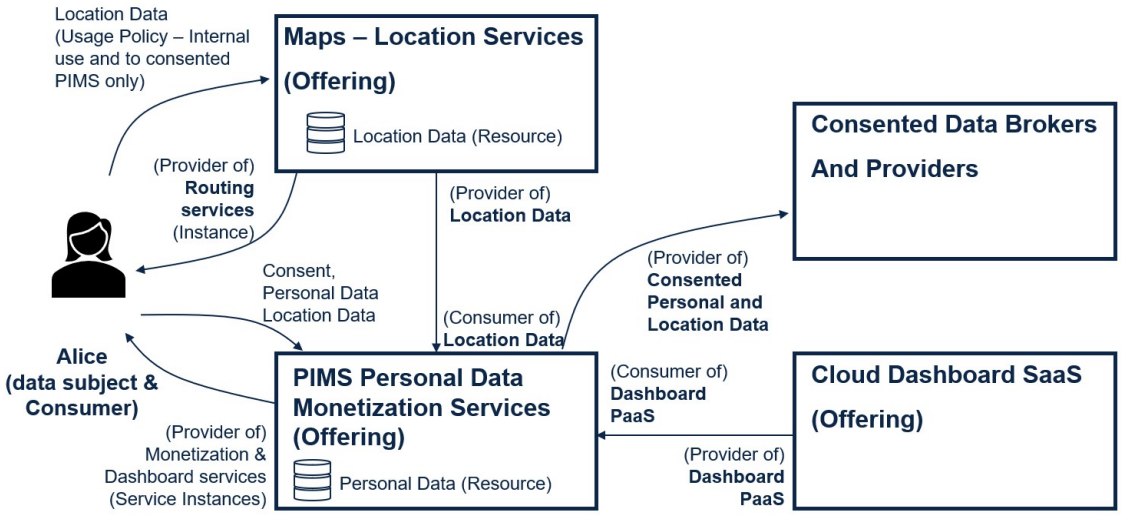



Figure 3.15: An example of Gaia-X architecture in practice

Alice, the data subject according to the GDPR, has an account in a geo-location application that stores and manage her information. This application is the *Resource Owner* for such geo-location information, and a *Provider* of geo-location services (through its *Service Offering*) to Alice (the *Consumer*), and it operates the different *Service Instances* (e.g., a routing service to provide the best route to a certain point of interest based on Alice's position, or a dashboard showing the collection of places Alice has visited in the last year).

At the same time, Alice is the user of a PIMS that allows her to control and monetise her personal data. For that purpose, the PIMS ask the geo-location application storing Alice's personal information for access to such data. Such access must be granted by the *Contracts* and *Policies* of the *Resource Policies* and according to the rights included in the GDPR. In this case, the geo-location application is a data *Provider* to the PIMS. And the PIMS also provides services to Alice (e.g., a personal location dashboard, a management console) and to third parties whom Alice consented the PIMS to exchange her data with, hence requiring and activating the corresponding *Usage Policies*. The PIMS is both a *Provider* and a *Consumer* of Alice's personal data. At the same time, the PIMS is also a *Consumer* of PaaS services from a PaaS cloud service provider within Gaia-X infrastructure ecosystem, which the PIMS uses to generate the dashboard of Alice's management console.

#### 3.3.2.3. Gaia-X Operating model

The Gaia-X Federated Ecosystem allows *Participants* to integrate already-existing ecosystems (e.g., Catena-X for the automotive industry), or create new ones for particular purposes. It imposes some constraints to external ecosystems adhering to Gaia-X in order to be able provide or consume *Service Instances* or *Resources* of Gaia-X federation. For example, even when external ecosystems may use their own *Trust Anchors*, meaning entities trusted by every participant in there, they must use Gaia-X *Trust Anchors* to allow interoperability with other *Participants*.

*Trusted Anchors* are nominated by *Participants* and used to operate the ecosystem. They include *Label* issuers, *Trust / Identity Service Providers* and entities that validate *Self-Descriptions*. Even though the Gaia-X Association will initially play that role, the idea is that other companies or bodies help with undertaking these tasks and act as *Trust Anchors* in the long term, as well.

Gaia-X *Compliance* is a process of validating some simple essential rules in *Self-Descriptions* that allows to include them in FAIR catalogues (of *Data Resources*, of *Providers*, etc.) [195] and be made discoverable by other *Participants* in the federation. *Compliance* includes basic automated tests regarding the format and syntax of *Self-Descriptions* (JSON-LD delivering a RDF graph in line with Gaia-X definitions), a validation of the crypto signature (identity checked by Gaia-X *Trust Service Providers*, and signed by a *Trust Anchor*), consistency checks and verification on attribute values.

Gaia-X *Labels* are a way to "qualify" and label *Resources* and *Participants* issued by specific governance *Trust Anchors* acting as label issuers. Gaia-X *Self-Descriptions* are machine-readable immutable cryptographically signed sets of statements (a RDF Graph of Claims serialised in JSON-LD) that describe *Participant Roles*, *Resources* and *Service Offerings*. Since they are initially provided by the interested party, they must be validated through the *Compliance* process before being included in the federated *Catalogs*. They contain statements with *Claims*, and may contain credentials signed by 3rd parties endorsing them.

Should not they agree with *Claims* in *Self-Descriptions*, or should they spot any error, *Participants* can also initiate a process to revoke them. Whereas the *Compliance* process validates whether a description is formally valid, it is the Gaia-X Community and its users that validate whether the contents reflect the reality of the service or the entity a *Self-Description* describes. As a result, *Self-Description* endorsement and revocation can be used to build a reputation system within the Gaia-X Ecosystem.

According to the Architecture document, "*the Gaia-X Registry is a public distributed, non-reputable, immutable, permissionless database with a decentralised infrastructure and the capacity to automate code execution*" [70]. It facilitates the provision of a DLT with smart contracts functionality, and stores key governance information such as identity and the result of processes related to the validation of *Trust Anchors*, links to schemas and *Self-Descriptions* of *Catalogs* and other core elements of the ecosystem.

#### 3.3.2.4. Gaia-X Federation Services

There are five main federated services defined in Gaia-X: *Inter-catalogue synchronization*, *Identity and Access Management*, *Data Exchange Services*, *Gaia-X Trust Framework*, and Gaia-X *Portals and APIs*. Next, we summarise these federated services and their relationship with roles and concepts of the ecosystem.

**Inter-catalogue synchronisation** or **Federated Catalog** is an indexed repository of *Self-Descriptions* that allow *Consumers* to discover potential *Providers* and *Service Offerings* in the ecosystem, and monitor potential changes to such offerings. *Ecosystem Catalogs* can be linked to the global *Federated Catalog*. Cross references are based on unique identifiers of *Participants* and *Resources*. To comply with the standard, *Catalogs* cannot rank results, and must allow private *Self-Descriptions* that *Providers* can choose to share in private. Gaia-X Association provides a *Self-Description* browser and an extensible set of *Schemas* that must be supported by any *Catalog* and that *Self-Descriptions* must follow.

**Identity and Access Management** covers the identification, authentication and authorisation of *Participants*, identification and credential management and the verification of analogue credentials. Unique identifiers and a minimum list of attributes are assigned to *Participants* and *Resources*. A federated *Trust Framework* guarantees proof of identity. Only *Participants* following the policies, technical specifications and interoperability criteria set up by the Gaia-X Association can be awarded the Gaia-X *Labels*. *Self-Descriptions* and reputation of *Participants* contribute to establishing trust, too.

**Data Exchange Services** enable *Participants* to exchange data in the ecosystem. This requires a *Data Agreement Service* to manage the data exchange agreements, and a *Data Logging Service* to monitor and enforce *Usage Policies* in data flows after the signature of the corresponding *Contract*. They are complemented by *Resource* discovery services using *Federated Catalogs*.

**Trust Framework** includes mechanisms to make sure that *Participants* adhere to the Policy Rules during onboarding and service delivery [69]. Participants agree to a code of conduct, accept specific terms and conditions (e.g., data encryption and protection) and agree to third-party attestations during onboarding. Compliance services validate that *Self-Descriptions* are formally valid through a so-called *Continuous Automated Monitoring* (CAM) service.

**Gaia-X Portals and APIs** support the process of onboarding and accreditation of *Participants*, and allow them to carry out basic operations and services, such as managing *Self-Descriptions*, search for *Service Offerings* and for other *Participants*, etc.

According to the Gaia-X Architecture document, the former federated services contribute to Gaia-X architectural requirements of interoperability, portability, sovereignty, security and trust, and support the principles of decentralisation, distribution, federation and sharing.

#### 3.3.2.5. Gaia-X Infrastructure Ecosystem

Gaia-X defines an infrastructure ecosystem that focuses on storage and computing *Nodes* that execute *Software Resources* to process data, and *Interconnection Services* to ensure secure data exchange between *Nodes*. Gaia-X assumes that the best-effort Open Internet do not suffice to provide the quality of service required for all services. Therefore, a framework was defined to address the requirements of *Interconnection Services*, based on: i) *Self-Descriptions* of *Interconnection Resources*, ii) Quality of service (QoS) assessment by means of inter node measurements, for example, and iii) interoperable networking services including, but not limited to the Internet, dedicated point to point connectivity, etc.

Regarding the *Self-Descriptions* of *Interconnection Services*, the standard defines mandatory attributes for services offering a differentiated QoS, such as minimum requirements for latency, availability or throughput. The standard talks about a *Service Composition Framework* that allows to combine L1-L4 Interconnection Services provided by different *Participants* in more complex *Service Instances*. The Architecture document provides a first classification of the type of *Interconnection Services*, that includes networks, routes, connections, physical media and orchestration resources. They also discuss about the need to allow for *Interconnection Platforms* that avoid the need to build $N \cdot (N-1)$ communication facilities between $N$ Participants.

### 3.3.3. Gaia-X vs IDS

Gaia-X and IDS have some overlapping definitions and they are closely related. In fact, IDSA is an active member of Gaia-X from the beginning of this initiative, and its standard a central element of Gaia-X initiative, and there is an ongoing collaboration between Gaia-X and IDS.

Comparing the conceptual models, the underlying principles and requirements, and some concepts of both standards are similar to those of IDS. To facilitate this comparison, IDSA released a position paper that explains the relationship with Gaia-X [14]. Table 3.6 summarises the similarities and differences between both standards.

Gaia-X has adopted IDS *Connectors* for ensuring secure data exchanges. Both initiatives are actively promoting the collaboration of companies in use cases using the joint ecosystem, a key task to foster the exchange of data and to start bringing value to companies joining those Associations. Both initiatives restrict data exchanges to a limited trusted ecosystem of certified (and monitored) companies. Yet it should be proven that this standard can be somehow extended to the whole Internet ecosystem and it is ultimately supported by major private data holders (big Internet and tech companies).

Table 3.6: Gaia-X vs. IDS

| **Gaia-X** | **IDS** |
|---|---|
| Requirements: interoperability and portability of data and services, sovereignty over data, security and trust | Requirements: trust, data sovereignty, ecosystem of data, Standardised value adding apps and enablement of data markets |
| Architectural principles: Federation and decentralisation | Architectural principles: Federation and decentralisation |
| Federation defined around *Federated Services* | Federation defined around *Connectors* and *Certification* |
| *Federated Catalogue* / *Inter-Catalogue Synchronisation* | IDS *Broker*, *Vocabulary Provider* and *Information Model* |
| Sovereign data exchange | IDS *Usage Control* and *Clearing House* |
| Identity and trust | IDS *Identity Provider* and *Dynamic Attribute Provisioning* service |
| Nodes | Connectors |
| Open *Self-Descriptions* and *Schemas* | Structured but flexible *Information Model* |
| Data Ecosystem *Participants* (*Providers* and *Consumers*) | *Data Provider* and *Data Consumer* |
| Infrastructure Ecosystem and data exchange | Data exchange only |
| Services and *Service Instances* | *App Stores*, *App Provider* |
| Ongoing implementations: architecture and components under development | Already existing implementations by industry members |

## 3.4. Key Takeaways

We have catalogued ten different business models, based on a comprehensive survey that analysed 104 entities trading data on the Internet. Through this extensive study, it has become clear to us that most of the challenges these entities face have to do with *trust*. On the one hand, sellers express an ambition for absolute control of their data, and demand strong commitment from marketplaces to avoid unauthorised replication, resale or use of their data assets. On the other hand, potential buyers would benefit from testing data and knowing its value before closing a transaction, and from certifying that information comes from trustful data sources.

Unsurprisingly, the most successful market players nowadays are horizontally integrated service providers that protect (rather than share) their most valuable data assets. Traditional providers are being challenged by marketplace platforms that mediate between data sellers and buyers and facilitate data transactions. Nowadays, *niche* marketplaces coexist with *general-purpose* platforms. In both cases, their business model is yet to prove viable, and it is unclear whether specialisation is convenient or rather a winning one-fits-all solution will arise. On the one hand, *niche* DMs have clear advantages over *general-purpose* ones. First, because focusing on certain data space and leveraging their specific expertise let them provide value-added services both to buyers and sellers along with data sharing. Second, because their platform is adapted to the kind of data they trade, and they concentrate their commercial efforts on attracting a specific buyer segment. On the other hand, *niche* DMs target a much smaller market, and the concept of a one-stop-shop for any kind of data is arguably attractive.

Unlike public DMs, *embedded* DMs and PMPs consider data trading more as an add-on to the services they already provide. This *commodification* of data trading has two important competitive advantages. First, they leverage an existing potential customer base on the buy side, which lets them concentrate on finding the right data partners to attract their captive demand. Second, they sell data to be used within a specific environment, which significantly reduces the risk of replication and lets them provide more focused, processed, and thereby more valuable data.

Fighting against the data-for-services dynamics of the Internet becomes the main challenge of *PIMSs*, provided the rights of new data protection legislation are enforced by competent authorities. They are focusing on gaining the *trust* of users to build a minimum viable base of millions of them, yet their feasibility is still to be proven. Consequently, they are struggling to make themselves known, leveraging an increasing concern around privacy on the Internet. Conversely, the variety of existing isolated platforms may undermine their trustworthiness. A future consolidation may help them acquire users, though it may well turn the odds against them unless they differentiate themselves from the '*big data lords*'- why trusting your data to a *PIMS* instead to Internet service providers? Adopting data *trust* models might be a way to overcome this challenge.

Bottom line, the data economy is a developing and oftentimes unproven paradigm and ecosystem still undergoing major development. A huge corporate, entrepreneurial and research effort aims to *de-silo* data and enable a healthy trading of such an important asset. In this study we have revealed significant differences between what is working in the market right now and what the market is developing. Through commodifying and specialising data trading, the market is moving away from horizontally integrated monolithic siloed data providers, and towards distributed '*niche*' exchange platforms.

In the next chapters we will address and provide solutions for some of the challenges we have identified during this survey. Moreover, the outcomes of this PhD can help in addressing some challenges faced by IDS and they can also contribute to Gaia-X federated services and to data marketplaces implemented over these standards.

First, chapter 4 addresses the problem of data pricing (*Challenge 2*), and presents a first of its kind measurement study of the growing DM ecosystem. We identify a set of data features closely related to the price of data products in the market, hence this work is closely related to IDS *Information Model* and to Gaia-X *Self-Descriptions* of data products. The analysis carried out in this section would be useful for IDS in two ways:

1. Features related to prices of data products in the market should be included in the *Information Model* to capture all the inputs required to estimate the value of *Digital Resources*.

2. A future data pricing tool whose design we present in chapter 8 could make use of this work, metadata stored according to the *Information Model* and the transactions of *Data Brokers* and *Clearing Houses* to provide a hint to participants in the ecosystem about the value of their data.

Chapter 6 addresses the problem of buyers selecting suitable data for their purpose and specific task (*Challenge 3*). Should Gaia-X and IDS be successful and widely adopted, they will bring to market a number of different data assets applicable to the same problem, and thus buyers will be in need of selecting among them. This is exactly the topic we are addressing in chapter 6. To help with this problem, *Clearing Houses* and *Data Brokers* may implement mechanisms that allow *Consumers* to test pieces of data on their models before they sign a contract to buy or let access to this data.

Finally, chapter 7 deals with computing the relative value of pieces of data for a specific task. Not only would this be valuable for selecting among eligible data sources and cleaning data samples [75], but also for data marketplaces to distribute payoffs of a data transaction covering data from different *Providers* according to the value each of them brings to the consumer. Within Gaia-X, both processes would be valuable *Service Instances* if implemented by data marketplaces in this ecosystem. Within IDS, they could be again valuable functions of *Clearing Houses* and *Data Brokers*.

# Chapter 4

# Measuring the price of data in commercial data marketplaces

An issue of paramount importance is that of *data pricing*. In chapter 3, we strove to identify different schemes data marketplaces apply when setting the prices of data products, and we made some interesting findings. For example, some marketplaces leave it to sellers to set a price for their data products. Moreover, 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 step.

We have also discussed in Sect. 2.2.2 the peculiar characteristics of data as an economic good. Due to the elusive nature of the traded "*commodity*" and due to the complex business models under which it is made available, pricing is a very complex matter, even more complex than pricing material goods [153]. The research community at the intersection between computer science and economics has studied several aspects of data pricing (see Sect. 2.3 for more details).

Ultimately it is the market that decides and sets prices via complex mechanisms and feedback loops that are hard to capture. Despite some other works trying to measure the price of personal data of individuals [34, 143, 150], there is no systematic measurement study about the price of data products traded in commercial data marketplaces.

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

To answer such questions we followed our own novel methodology summarised in Fig. 4.1. First, we checked data marketplaces in our survey and we selected 10 of them that fulfil necessary criteria for a measurement study. For these ones we developed custom crawlers for retrieving information about the products they trade. Using these crawlers, and adding the portfolio of

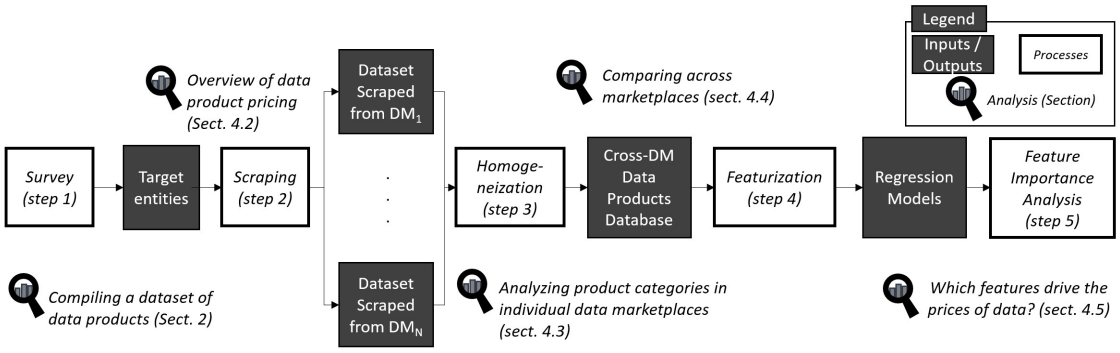



Figure 4.1: Summary of our methodology

another 30 data providers, we obtained information for more than 210,000 data products and a catalogue of more than 2,100 distinct sellers. We also developed data product category classifiers, meaning ML models for identifying products of similar categories or responding to the same use case across marketplaces, and executed 9 different regression models to understand which features are driving their prices.

The remainder of the chapter is structured as follows. First, we frame the scope of our analysis and show some initial outcomes of our measurement study in Sect. 4.1. In Sect. 4.2, we present an analysis on data product pricing in commercial marketplaces. Furthermore, Sect. 4.3 dives deeper into analysing the products in individual marketplaces, particularly in AWS' DM and DataRade, which account for the largest number of price references in our sample. We then address the challenges of comparing across data marketplaces and develop tools for enriching our sample by labelling data products homogeneously in Sect. 4.4. Finally, in Sect. 4.5, we apply several methodologies for analysing the importance of different metadata features in determining the price of commercial data products.

## 4.1. Compiling a dataset of data products

In this section we explain the scope and methodology of our study. Section 4.1.1 lists the data marketplaces and providers whose products were included in our work. Section 4.1.2 provides some insights into the web scraping exercise and the software modules we developed for this purpose. Finally, Sect. 4.1.3 summarises some statistics about the information we managed to collect from these entities.

### 4.1.1. Scope of the measurement study

Existing works and surveys on commercial data marketplaces [146, 163, 168, 169], our own exercise presented in chapter 3, an extensive web search and a consultation with experts in the area allowed us to compile a list of data marketplaces. From our analysis, we also identified a

subset of marketplaces that fulfilled the criteria for using them as sources of data for a reproducible measurement study. Such criteria include that they grant access to their product catalogue without requiring an account, or through an account but without a vetting process or upfront paid registration, that they have a reasonably large catalogue that includes sufficient descriptions of their data products, and that they include a clear description of their pricing policy.

Out of the 180 initial DMs, only 10 companies fulfilled all of the above criteria. Most of them did not make it to the list simply because they do not allow non-paying users to browse their catalogues. For example, marketing-related private marketplaces such as *Liveramp*, *LOTAME* or *TheTradeDesk* neither provide public per-product information nor any price references. However, they do provide information about their data partners. By analysing the presence of these partners in other marketplaces, we did find that 45% of them in those private marketplaces sell through general-purpose public ones, such as *AWS* or *DataRade*, as well, and hence we have been able to include products of theirs in this study.

We also discarded several otherwise *scrapable* general-purpose DMs such as *Data Intelligence Hub* (DIH), *Google Cloud DM* because they included only free data products. We chose to scrape the largest of these free open data marketplaces, *Advaneo*, to help in training our data product category classifiers.

Table 4.1: Summary of scraped DMs

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

Table 4.1 lists the 10 DMs that we use as data sources in our study. Overall, we include 6 general-purpose and 4 niche DMs, as well as 30 data providers[1] that, in addition to commercialising their own 777 data products through DMs, provide valuable pricing information on their own websites, which they also use to advertise and deliver their products.

[1] 42matters, Airbtics, Apptopia, Benzinga, Bizprospex, BoldData, BookYourData, bronID, BuiltWith, DataScouts, Demografy, ebCard, Enigma, ESGAnalytics, HGXN, IFDAQ, ipinfo.io, MultimediaLists, MyDex, OikoLab Weather, Onclusive. Open Corporate, PanXchange, Pipecandy, Shutterstock, Storm Glass, TelephoneListsBiz, Unwrangle, USASalesLeads, and Walklists.

### 4.1.2. Scraping data products and data providers over the Internet

We developed our own web crawler to render and download web pages, and specialised parsers for extracting metadata from data products in data marketplaces and structuring it in tables that we can process with relational databases. Figure 4.2 shows a functional diagram of such scraping tool. It requires the following inputs shown on the left:

- A list of URLs to be downloaded and parsed,
- A list of proxy servers to be (optionally) used in the process of downloading web pages from the Internet, and
- A list of header configuration to be (optionally) used to disguise the downloading tool and make it look like different browsers when requesting web pages.

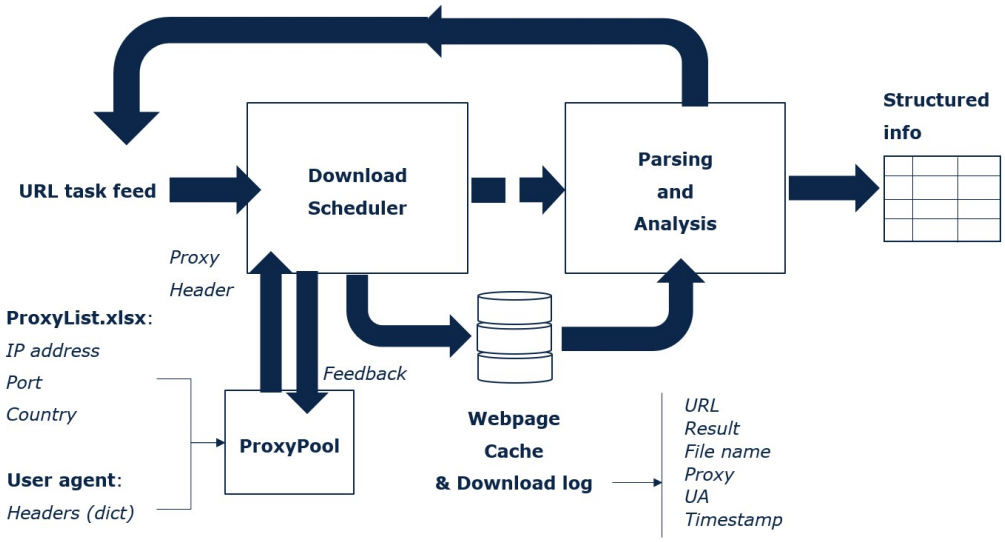



Figure 4.2: Functional diagram of the scraping tool

The crawling process takes two steps. First, the *scheduler* deals with downloading and rendering web pages, which are stored in a common *web page cache*. In this process, it also generates a download log, and manages the list of proxy servers (e.g., demoting those that do not respond). We follow common crawling good practices [88]. For example, we avoid 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.

Second, *ad hoc* parsers specialised in each data marketplace read web pages stored in the *Web page cache*, detect whether they correspond to a data product, extract all the metadata available from them (e.g., title, description, provider, related products, etc.) and generate a structured output writing metadata information to a *csv* text file. Some products include URLs to other related products within the same marketplace. Parsers use this information to update the URL task feed, if they detect new products.

Since every data marketplace defines its own format and publishes data products using different structures, specific parsers had to be developed for each platform and the format of the structured output differs between them. Section 4.5.3 explains how we managed to homogenise those outputs and to assemble a single data set encompassing products of all the marketplaces.

### 4.1.3. Results of the scraping activity

We collected information related to 215,075 products from 2,115 distinct sellers in total. We analysed the information at the level of entities trading data on the Internet and at the level of the data products they offer in the market.

Regarding companies trading data, 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. Dealing with several marketplaces with different interfaces and APIs multiplies their effort to commercialise their data products. This is especially relevant for small market players and for niche providers (58% of them) focusing on one product only in our sample. As a result, most data providers (81%) work with only one DM in addition to selling their products through their own web site.

Still, bigger data providers do actually work with several data marketplaces. We have already referred to 45% of providers in financial and marketing-related marketplaces that sell through general-purpose DMs, such as *AWS* or *DataRade*, as well. Moreover, we also spotted DMs advertising and offering their products in other DMs (e.g., *Battlefin* or *CARTO* through *AWS*), making the relationships between players in the ecosystem even more complex.

At the level of data products, we scraped all available metadata such as the product id, title, description, source, seller and, when available, its geographic scope, volume, category, use cases, update rate, historical 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.

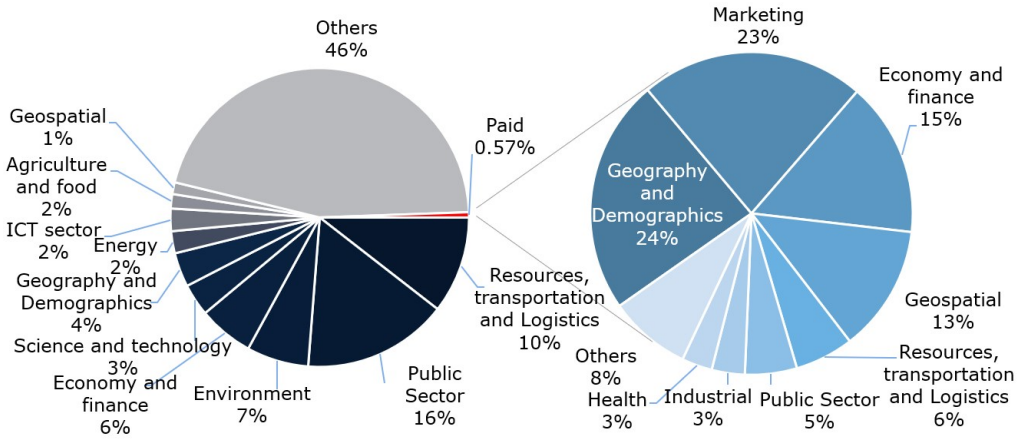



Figure 4.3: Breakdown of products in *general-purpose* DMs

By matching the different categories used by marketplaces, we arrived at Figure 4.3, that presents the most frequent data categories of data products in *general-purpose* DMs. The pie chart on the left includes free and paid data products, whereas the one on the right includes only those that are paid (10,860 products). '*Marketing*' and '*Economy and Finance*' fall among the most popular categories for paid data products. Moreover, the presence of '*Geography and Demographics*' and '*Geospatial*' data emphasises the importance of geo-located data in the sample, as well. As we will discuss later in Sect. 4.4, comparing across DMs entails complex challenges, and this methodology is not fully accurate.

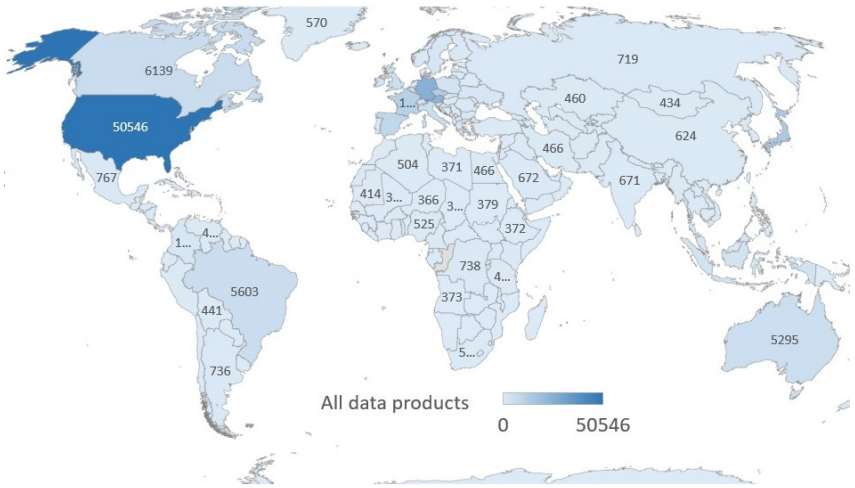



Figure 4.4: Data products by country

Regarding the geographical scope of data products, we found that marketplaces 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 4.4 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.

## 4.2. Overview of data product pricing

It may appear initially surprising that, despite being commercial entities in the B2B space, most of the surveyed and some of the scraped DMs offer predominately free (most of the time open) data. Again we point to the fact that these are privately held companies [2,90] and not open data NGOs or government initiatives. Our conjecture is that since DMs are two-sided platforms, pre-populating them with free data is a very reasonable bootstrapping strategy, since it can attract the initial "buyers", which in turn will attract commercial sellers and thus help the marketplace grow its revenue.

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

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 [123, 153] (see Sect. 3.2.4.3). Predominant among the 11,823 non-free data products are the *subscription-based* model (i.e., buyer paying for a subscription to get access to data for a period of time), and the *one-off* model (i.e., lump sum payment for data), seller-led in both cases. The first one is used mostly for "live" data usually accessed via an API (e.g., IoT sensor data), whereas the second is used for more static data, which are usually downloaded as one or more files.

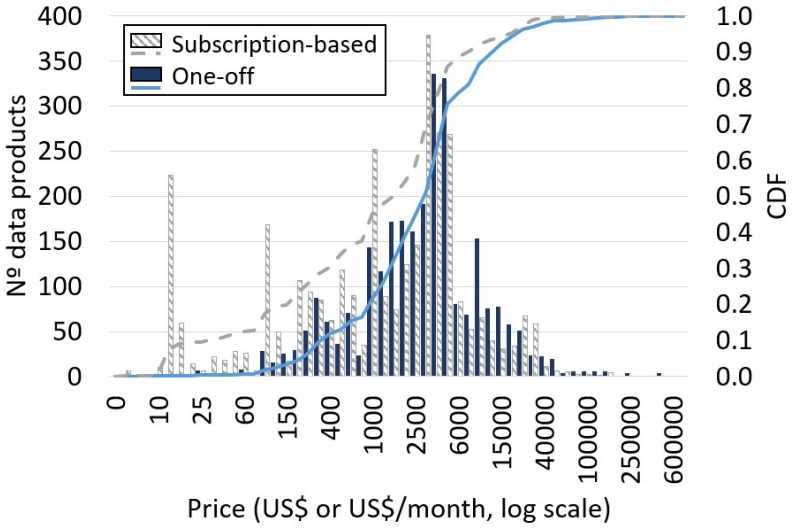



Figure 4.5: Histogram and CDF of data products

4,162 products from 443 distinct providers provided clear information about their prices. Figure 4.5 shows a histogram and the corresponding cumulative distribution function (CDF) of monthly prices for data products. Regarding those offered under a *subscription model*, we see prices across a wide range up to US$150,000 per month. Cheap products below US$100 per month are often curated and cleaner versions of open data. For example, a seller offers a historical compilation of quarterly reports submitted to the US Securities and Exchange Commission (SEC), also downloadable from their websites. They also include low-cost "promotion samples" of more expensive products from well-known sellers, such as GIS data and supporting metadata for a small area of some US cities. The median price is US$1,417 per month. Almost one-third of all products, including targeted market data for example, are sold for US$2-5k monthly.

Comparing to products sold under a *one-off model*, (1) the latter tend to be more expensive: median price US$2,176 vs. US$1,417 per month for *subscription-based* products; maximum price US$500,000, more than 3 times higher than the maximum in *subscription-based* access, and (2) *one-off* products have a price histogram more normally distributed around its median at US$2,176. Among 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. 4.5. We will come back to them later.

## 4.3. Analysing product categories in individual data marketplaces

To get a more in-depth understanding of data pricing, we analyse the catalogue of individual data marketplaces. They may offer data products with different pricing schemes or currency units. We present the methodology we used to convert to a common comparable standard unit of price measurements in Sect. 4.3.1. Then, we analyse the data products in AWS' DM, the one with the largest base of paid products with prices, all of them subscription-based at the time of crawling this data marketplace. In Sect. 4.3.2, we set out to study what is the range of prices of data products and which categories of data command the highest prices. Finally, we repeat the exercise with DataRade and we present its results in Sect. 4.3.3.

### 4.3.1. Challenges and proposed solutions

Comparing the prices of products in a single data marketplace is not necessarily straightforward, since most of them allow sellers to set prices using different currencies and to use different pricing schemes. Next, we explain how we dealt with these issues in analysing the prices of individual data marketplaces.

Some data marketplaces work with several currencies and prices in different currencies are not directly comparable. Therefore, we need to convert them to a common currency. To avoid the volatility of exchange rates, we use the average rates in the last 6 months to convert to US dollars. Hereinafter, we assume all the prices are converted to US dollars using this methodology.

Comparing between different pricing schemes usually requires converting to a standard unit of measurement [65, 124]. In the case of DataRade, you can find both one-off and subscription-based products. One-off prices are set as a lump-sum ($p_{off}$) that the buyer pays for downloading a snapshot of their data, whereas in the case of subscription-based products buyers pay a fee ($p_{sub}$) to have access to the corresponding data, including any update, for a certain period of time, say $T_{sub}$. Almost all the products do also specify their update rate, meaning the inverse to the period of time $T_u$ the data provider takes to generate a new updated version of the data.

To compare one-off to subscription-based products we convert all prices to the cost per month for the buyer to get access to up-to-date version that data. In the case of subscription-based products, this is directly the price $p$ divided by the subscription period in months $T_p$. Should the buyer provide different prices for different subscription periods, we will take the cheapest option

for the buyer, meaning the one with the lowest $p_{sub}/T_{sub}$ US$/month, which is usually the one with longer $T_{sub}$. One-off products usually have longer update periods ($T_u >= 1$ month), that often correspond to months, quarters, or even a year. Therefore, users willing to purchase up-to-date data pay a price $p_{off}$ every time the dataset is updated, hence they pay $p_{off}/T_u$ US$/month on average, assuming $T_u$ is measured in months.

Next, we apply this methodology to study prices in AWS Marketplace and DataRade.

### 4.3.2. Amazon Web Services Marketplace

AWS classifies data products in its marketplace 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*'.

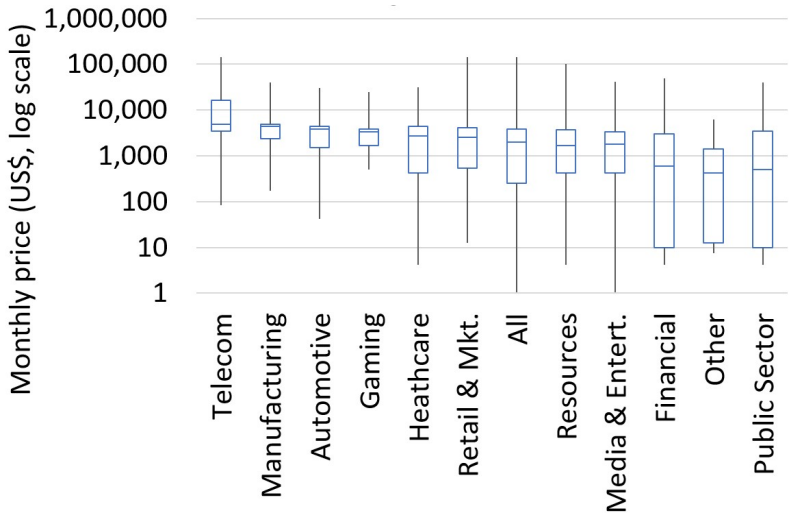



Figure 4.6: Subscription prices by industry in AWS.

Figure 4.6 shows a box plot of products by category in AWS. The X-axis shows the different categories sorted by 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 ($\times 2.6$, $\times 2.3$ and $\times 2$, respectively). Most low-value products belong to the '*Public Sector*', '*Financial*' (e.g., stock price feeds), and ' *Other*' categories, whereas the most expensive ones relate to '*Telecom*' and '*Retail, Location and Marketing*'.

### 4.3.3. DataRade

DataRade classifies products in a hierarchy of more than 300 categories of data. Figure 4.7 shows a box-plot of the prices of data for the first level categories in DataRade. In this case, it is especially '*Credit Rating*', but also '*Mobile App*' and '*Healthcare*' data that show the highest

median prices ($\times 8.3$, $\times 3.7$ and $\times 2.5$ above the overall median, respectively). Again, the most expensive products are related to marketing, in this case to the categories of '*Commerce*', '*B2B*', and '*Consumer*' data.

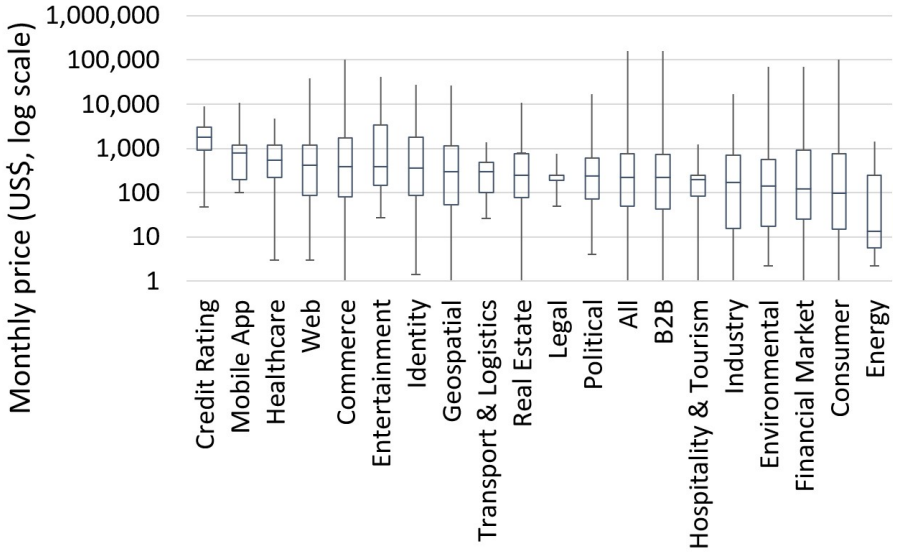



Figure 4.7: Subscription prices by category in DataRade.

## 4.4. Comparing across marketplaces

### 4.4.1. Challenges and solutions proposed

Comparing information about data products from different marketplaces is an even more difficult task since we need to deal with different information models. This involves additional challenges to the ones that we spotted when analysing products in individual data marketplaces, First, each marketplace provides metadata of different granularity and level of detail. Second, each platform uses different categorisation to describe their products. To overcome these challenges, we developed a methodology to homogenise the categorisation of data collected in order to be able to compare similar products across marketplaces.

#### 4.4.1.1. Dealing with different levels of detail

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

We normalised and stored in this cross-DM database all the information from the scraped datasets. We managed to fully automate the extraction of most of the fields (18 out of 27), which were directly scraped from the web pages of the different DMs. This extraction was semi-automated for 5 fields, meaning that they were automatically extracted for certain marketplaces, or retrieved from product descriptions for others, in a process that required a manual check afterwards. For example, *update rate* of data is usually included in the general description of a data product, but the presence of the word '*monthly*' may not necessarily point to a monthly update rate. Information about data volume or data subject units was automatically extracted only for DataRade and BookYourData, and required computer-aided manual typing in the rest of the DMs (we highlight and extract numbers and their context from data descriptions). Manual checks were performed by three different experts. Any ambiguities and disagreements were resolved by majority voting.

#### 4.4.1.2. Dealing with different categorisation systems

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

We addressed this issue by developing a series of natural language processing (NLP) naïve Bayes (NB) classifiers [55, 59, 111]. In our first attempt, we wanted to identify similar data products – those that belong to the same category – between two different (source and destination) DMs. As a result, we trained both multinomial and complement versions of NB classifiers to detect data products from the source DM that belong in a certain category by using feature vectors based on the information provided by the data product description from the source DM. We used bag of words [105] and data pre-processing steps such as removing stop words and words with numbers, using stemming and TF-IDF transformation [130, 162]. Then we validated the resulting classifier against a manually labelled sample from the destination DM. Manual labelling was performed by three different experts. Any ambiguities and disagreements were resolved again by majority voting.

### 4.4.2. Comparing DataRade to AWS Marketplace

We utilised the above methodology to build different classifiers that help us compare data products between the two DMs including more price references, namely DataRade (destination DM) and AWS (source DM). We generated our feature vectors based on AWS data product de-

scriptions (source DM) and applied the resulting classifiers to DataRade data products (destination DM). We were interested in finding out: (1) what percentage of products from those categories we could identify in DataRade, (2) whether categorisation and pricing were coherent between them, and (3) whether we could enrich our metadata by adding AWS's inferred categories.

Using our cross-DM database, we generated train/test datasets 80/20 splits in order to train and test the corresponding classifiers. We observed that multinomial classifiers outperformed the complement NB for this task so we proceeded with the former ones. The resulting classifiers yield an acceptable $F_1$ score above 0.85 (average for 50 executions with different random 80/20 train/test splits). In fact, they identified meaningful and reasonable stems when tagging products related to each category. For example, for the two categories including more data products:
**Financial:** 'system', 'sec', 'exchang', 'type', 'file', 'form', 'edgar', 'secur', 'act', and 'compani'.
**Retail, Location and Marketing:** 'locat', 'topic', 'b2b', 'score', 'echo', 'trial', 'compani', 'visit', 'intent', 'consum'.

Then we validated the models against a manually labelled sample from DataRade. Manual labelling was performed by three different experts. Any ambiguities and disagreements were resolved again by majority voting. The validation set included 745 manually pre-labelled data products with both '*Financial*' and '*Retail, Location and Marketing*' tags. The models trained only with data from AWS did not perform so well on the validation set ($F_1$ scores of 0.73 and 0.43 for '*Financial*' and '*Retail, Location and Marketing*' data). To further generalise our methodology and to improve its accuracy, we added training data from other marketplaces. In particular:
(1) The ***Financial*** classifier was trained with 95,208 labelled descriptions of products from 4 different entities (Advaneo, Carto, AWS, and Refinitiv), and 45,298 financial products.
(2) The ***Retail, Location and Marketing*** classifier was trained with 3,828 descriptions from 3 entities (AWS, BookYourData and TelephoneLists), including 1,614 marketing products.

By adding products belonging to the same category from other DMs we observed better balance between precision and recall and an overall improvement of model generalisation. We also observed an increase of the $F_1$ score in the test set. Particularly, adding information from Refinitiv improves the $F_1$ score from 0.73 to 0.79. In the case of '*Retail, Location and Marketing*', adding information from specialised marketing DMs (e.g., BookYourData), drastically improves the $F_1$ score from 0.43 to 0.74. We tested multiple classifiers, with and without stemming, and we found that using word-based instead of stem-based features led in general to more accurate results in both cases (+5% $F_1$ score). Table 4.2 shows the accuracy obtained by both classifiers.

Table 4.2: Score of data product classifiers

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

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

### 4.4.3. Comparing AWS Marketplace to DataRade

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

### 4.4.4. Labelling data products across marketplaces homogeneously

We repeated the process described before to attach AWS data category labels to all the products in our database. As a result, we managed to enrich our sample by homogeneously labelling products based on their descriptions according to the categories and criteria used in AWS DM.

Figure 4.8 shows a box-plot of the prices of all the products consistently classified by AWS' industry. The four categories with highest median prices are the same as in AWS, but in a different order. Similarly, most products belong to '*Financial*' and '*Retail, Location and Marketing*', and the most expensive products do also belong to the latter category.

## 4.5. Which features drive the prices of data?

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 in Sect. 4.5.1, and ii) see how particular sellers price their data products in Sect. 4.5.2. Then we explain the resulting list of relevant features in Sect. 4.5.3. Finally, Sect. 4.5.4 uses this information to train regression models predicting the prices of data products based on their metadata and on the prices of other products in the market. We do not intend to build state-of-the-art price predic-

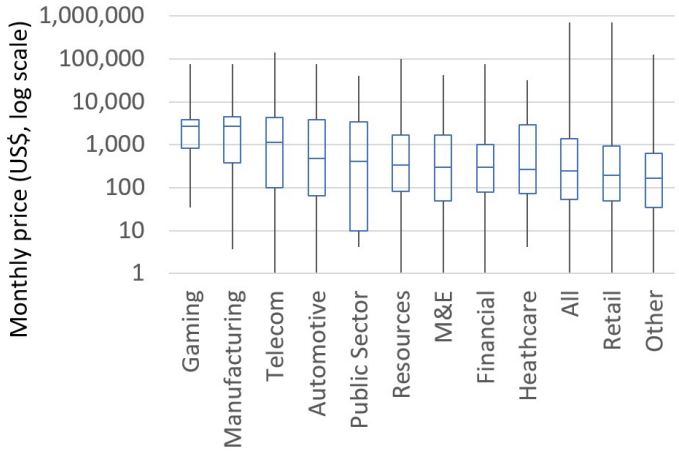



Figure 4.8: Monthly costs of all products by AWS' industry.

tors, but rather to understand which features are driving the price of data. Therefore, we conduct feature importance analysis on the resulting regression models and we find out which features have the highest impact on the observed prices for the different data products in our corpus.

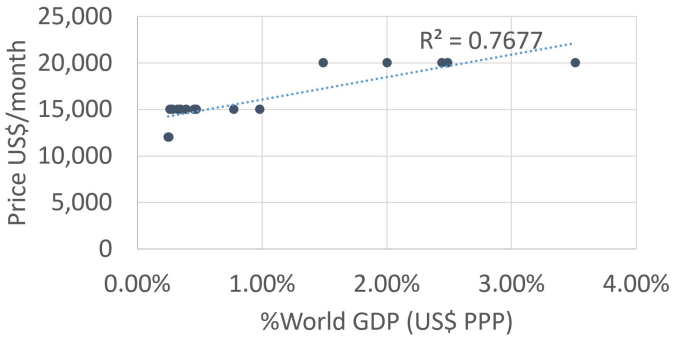
### 4.5.1. Key features of expensive data products

Section 4.2 pointed to a long tail of 33 data products worth more than US$30,000 per month. We found that:

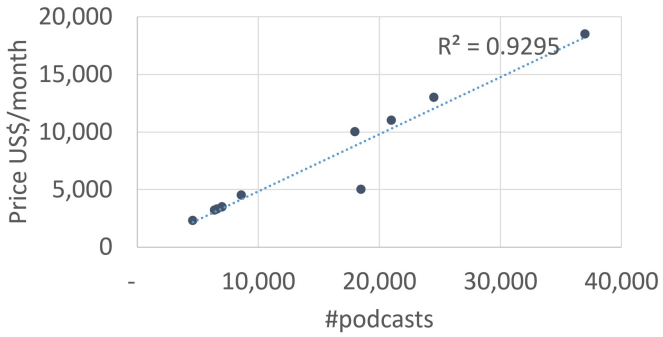
- All of them include huge amounts of data from millions of people, tens of thousands of locations, thousands of enterprises, etc.
- 20 (61%) of them offer daily updates.
- 11 (33%) of them do not provide any past data, and only 4 (12%) include over 2 years historical data.
- 22 (67%) of them are US-focused, 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).

### 4.5.2. Seller-specific pricing strategies

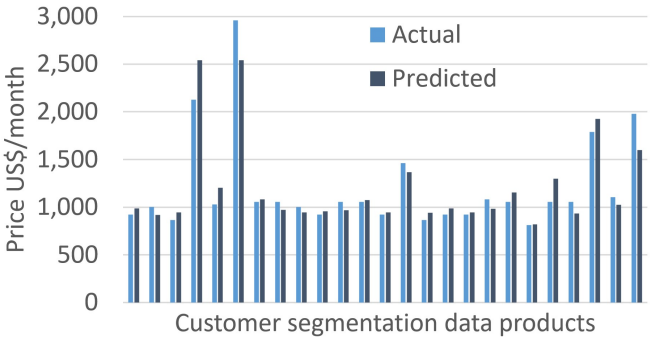
We also looked at how *specific* sellers set the prices of data. We found that surprisingly simple regression models relying on specific metadata features were able to accurately predict those prices. Understanding sellers' pricing strategies proved to be valuable in order to find features related to asking prices observed in commercial DMs.




(a) Mobile coverage




(b) Podcast metadata information




(c) Consumer segmentation

Figure 4.9: Pricing regression examples from specific sellers

Figure 4.9 depicts three such examples for telecom, recommender systems, and consumer segmentation data. We observe that:

- A seller offering mobile network infrastructure and coverage data by country groups products in a few price tiers that depend on their gross domestic product (see Fig. 4.9a).

- A seller offering podcast metadata uses language to segment its products, and prices them proportionally to the nº of podcasts they include (US$0.5 per podcast, see Fig. 4.9b).

- 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. 4.9c, $R^2$ = 0.88, MAE = 10% of the average, MRE = 9%).

### 4.5.3. Building a feature matrix to feed regression models

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

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

Before that, an additional pre-processing step is needed in order to transform the fields of our cross-DM database into a set of valuable features that can be ingested by ML regression algorithms. This process uses the NLTK [24] and Scikit-learn [152] Python libraries and includes mainly the following steps:

1. Extraction of '*word*' features from the title and the textual description of each data product. We use bag of words [105] and data pre-processing steps such as removing stop words and words with numbers, TF-IDF transformation [162], and stemming [130]. In addition, we have sellers' names removed from the vocabulary, so as to avoid bias introduced by knowing their identity. Finally, we prepare matrices for different vocabulary lengths and optimise each algorithm for this parameter.

2. Breakdown of volume-related fields in 13 different groups depending on their nature. For example, we separate data products targeting 'entities' or 'companies', from those whose subjects are 'individuals' in different features. The resulting comparable units are in turn normalised, and a new overarching feature ('units') measuring the percentage of units covered is added to compare products across groups of units.

3. Calculation of country-level binary features to indicate whether a certain country is covered.

4. Homogenisation of the units measuring the time scope, what we will call *history*.

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

As a result of this *featurisation* process, we reduce each sample product to a feature vector and produce a feature matrix to train our regression models. Table 4.3 lists feature groups and some examples of their individual features. We organise features in 10 disjoint sets according to their nature and the basic questions they answer about data products.

Table 4.3: List of feature groups

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

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

### 4.5.4. Analysing feature importance

Regression models can be used for feature importance analysis. After explaining how we managed to optimise regression models to fit the market prices of data, we use a range of such techniques to understand which features have the higher impact on the prices of data products.

#### 4.5.4.1. Optimising Regression models

Owing to their stochastic nature, training several regression algorithms and comparing their outcomes is key to obtaining robust conclusions. Consequently, we have tested variations of 9 different regressors with different values for their main parameters (e.g., number of estimators, depth, etc.) as included in the Scikit-learn [152] Python library, and inputs of different vocabulary lengths. Such models work with the log instead of the absolute value of product prices as the dependent variable so as to normalise the distribution of prices and avoid negative price predictions. We were hoping to find at least 3 models that produce sufficiently accurate price predictions, measured as the $R^2 score$ of their output w.r.t. actual prices.

To reduce the complexity of each model, we removed low-value features, i.e., those that had a negative leave-one-out (LOO) value, provided the accuracy of the model was not negatively affected. A feature having negative LOO value means that the model improved its average accuracy

Table 4.4: 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 |

in 10 random executions for different train and test data splits when such feature was removed from the input matrix. Finally, we performed a cross-validation to check the variance of the model accuracy when training and testing in 5-folds, and 20-random training-test splits of the input data.

We found that three target models worked reasonably well (i.e., they yield an $R^2$ score greater or equal to 0.70), namely Random Forest [25], k-Nearest Neighbours [110], and Gradient Boosting [68, 129] regression models. On the contrary, we discarded linear, Elastic-Net [206], Ridge [89], Bayesian Ridge [125], and Lasso [179] regressions even though they worked well in specific simulations.

In addition, we also tested a Deep Neural Network regressor using the TensorFlow [1] and Keras [102] libraries. We followed all common good practices recommended for such activity by first standardising the input data. We tested RELU/Leaky RELU activation functions for all hidden layers, and a linear activation function for the output layer. As loss function we used the mean absolute error (MAE). To avoid overfitting we randomly applied Drop-out between training epochs and to avoid dying/exploding neurons we also applied Batch normalisation between all layers. We used the Adam optimiser [104] with a tuned learning rate decay to train the model faster at the beginning and then decrease the learning rate with further epochs to make training more precise. Finally, we used Callbacks to stop the training at the optimal epoch.

Table 4.4 presents a summary of the accuracy obtained by the different regressors by category of data products, including the $R^2$ score, the MAE and the mean squared error (MSE) with regards to the actual log prices. For the sake of robustness, our results were consistent across subsequent 5-fold and 20 random train/test splits: $R^2$ score showed a standard deviation below $4\%$ of the average in each round. Note that due to the total (low) number of observations that we have in our datasets, DNN models are not recommended, nevertheless, we wanted to explore them since we believe that they will further improve our results as soon as we manage to increase the overall size of our datasets. Consequently, we avoided carrying out any the feature importance analysis on DNN models.

#### 4.5.4.2. Analysing the importance of individual features

We carried out this process for financial, marketing, healthcare and all data products in our sample. Financial and marketing data were the most popular data categories in our sample, whereas healthcare data was chosen as a relevant disjoint category of less though increasingly

Table 4.5: Top 10 most relevant features not related to volume by category and regression model

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

popular products showing a different behaviour in terms of prices. As a result, we obtained at least one model that achieves a $R^2$ score of 0.78 by category and accurately fits the prices of data products (see Table 4.4). We ran two different individual feature importance analysis:

1. measuring the accuracy lost by randomly shuffling the values of a certain feature among samples (permutation importance analysis [173]), and

2. measuring the prediction accuracy lost when one individual feature is removed from the inputs (leave-one-out or LOO value)

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

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

Table 4.5 lists other features not related to data volume in descending order of importance. Next we provide some details about the most important features of each specific category:

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

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

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

#### 4.5.4.3. Analysing the importance of groups of features

Since LOO is often negligible for individual features, we have repeated this analysis for groups of features answering to the same question regarding the data product (see Table 4.3). In this case, we have used the following two methods:

1. Measuring the prediction accuracy lost when a group of features is removed from the input dataset (leave-one-out - LOO).

2. Measuring the average (in 20 random train/test split executions) Shapley value of each group of features.

The Shapley value is defined as the average $R^2$ score added by combining the information of a certain group of features with every possible mix of the rest of groups. This is a well-known and widely-used concept in game theory, economics and ML [75, 165], and it is considered a 'fair' method to distribute the gains obtained by cooperation. In our case, we applied the Shapley value to distribute the gains in accuracy of our regression models among the groups of features that contributed to achieving such an accuracy. Furthermore, we ran 5-fold feature importance analysis in the case of LOO, in a similar way as we did for individual features, and 20 calculations of the Shapley values for random 80/20 train/test splits of our input data.

Table 4.6: Feature analysis by feature group

(a) LOO values by feature group

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

(b) Shapley values by feature group

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

Whereas LOO measures gains or loses in accuracy of a model when features belonging in a group are removed from the input matrix, Shapley values better capture the complementarity among groups and take into consideration their individual predictive power, as well. Table 4.6a and Table 4.6b list the LOO and the Shapley values by group of features in descending order of importance. The standard deviation of Shapley values across executions is acceptable (average below 0.029 for financial and marketing datasets, 0.057 for healthcare-related data, and below 0.017 for all the data), and the ranking of relevant feature groups remains stable.

Figure 4.10 plots the percentage of the sum of Shapley and LOO values that each feature group represents, what we call their *predictive power*, and illustrates how important each group is for determining the prices of each category of products. We have piled together and coloured in gradients groups responding to the same question about data products.

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

By looking at Fig. 4.10, we can confirm that features related to '**volume**' and '**descriptions**' are the most relevant groups driving data prices: at least half of the predictive power owes to those two groups of features according to their Shapley values. While '*volume*' is clearly the

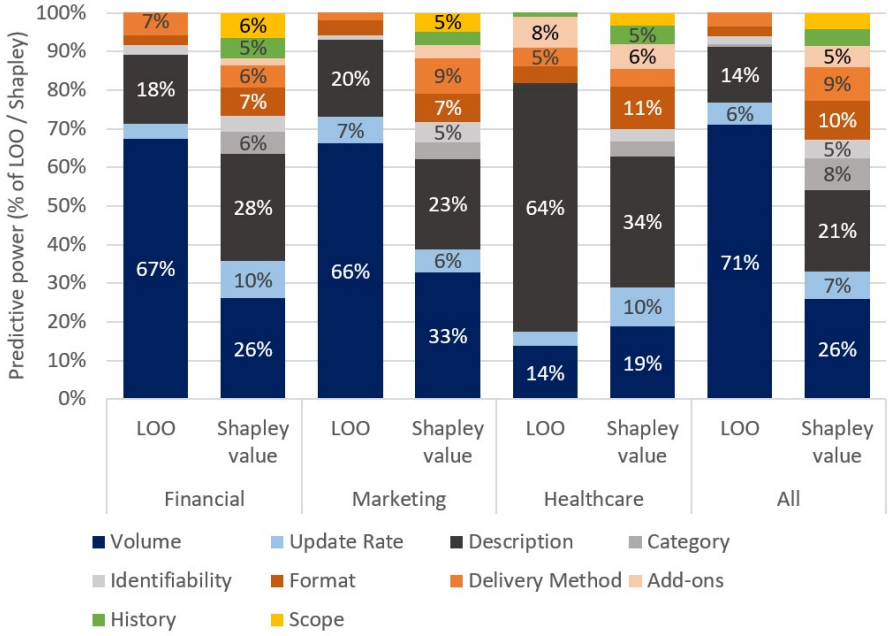



Figure 4.10: Predicting power of feature groups

most relevant group for marketing data products, it is not so relevant for healthcare-related data due to the heterogeneity of products belonging in this category and due to the lower sensitivity of their prices to volume.

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

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

## 4.6. Key Takeaways

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

Matching categories of data products in different data marketplaces we found that those related to "*Resources, transportation and logistic*", "*Public Sector*", "*Environment*" and "*Economy and Finance*" were the most frequent. Among paid products (0.57% of them), it is "*Geography and Demographics*", "*Marketing*", "*Economy and Finance*" and "*Public Sector*" the categories of data including more products.

Still, we learnt that comparing across marketplace is far from simple. Not only do they use different categorisation hierarchies, but they apply different criteria to label product categories, as well. Therefore, matching and directly comparing categories of different marketplaces may end up comparing pears to apples, even if the categories being compared are apparently the same. We circumvented this problem by using ML classifiers that are able to learn the criteria that DMs follow to label datasets belonging in a certain category, and apply the same criteria to label datasets in other data marketplaces. By using these classifiers, we enriched our sample by consistently labelling products according to AWS's categories.

Focusing on the products that carried a price, some 4,200 of them, we observed that:

- Prices vary in a wide range from few, to several hundreds of thousands of US dollars. The median price for data products sold under a *subscription* model is US$1,400 per month, and US$2,200 for those sold as an *one-off* purchase.

- Those related to "*Telecom*", "*Manufacturing*", "*Automotive*" and "*Gaming*" command the highest median prices, and that the most expensive ones consistently relate to *Retail and Marketing* across marketplaces.

- Using regression models, it is possible to fit the prices of commercial products from their features with $R^2$ above 0.84.

- Due to the heterogeneity of the sample there is no single feature that drives the prices, but instead we spotted meaningful features that drive the prices of specific categories of

data. For example, data update rate is a key price driver for *financial* and *healthcare*-related products, whereas geo-spatial localisation and the possibility of connecting data points from the same owner are for *marketing* data.

- Overall our models use features related to the category and description of the nature of the different data products (i.e., 'Financial', 'Retail', 'stock', 'contact', 'list', etc.), features related to the data products volume and units, as well as singular characteristics extracted from the data products description (i.e, words like 'custom', 'accuracy', 'quality', etc.) to forecast their market price. Groups of features related to '*what*' and '*how much*' data a product contains are driving 66% of its price.

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

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. We are continuously monitoring commercial DMs to see how they evolve, to enrich our database and to find out more about the price and the value of data. Using as building blocks the models and algorithms we present in this chapter, we are building a quotation tool for data products using real market data, which we describe as future work later in Sect. 8.3.

# Part III

# Buying and selling data

# Chapter 5

# Background and definitions

Designing and building a successful data marketplace calls for addressing a plethora of technology, business, and economics challenges [64]. According to our findings in chapter 3, DMs spend effort and money to attract enough sellers, and then try to convince buyers to purchase data through the platform. Therein lie three fundamental problems:

1. the *pricing problem* for sellers, which we already discussed in chapter 4,
2. the *purchasing problem* for buyers to select suitable data for their ML task, which we discuss in chapter 6, and
3. the problem of *distributing payoffs* to sellers according to their contribution in transactions combining data from different providers, which we address in the context of spatio-temporal information in chapter 7.

Before entering into the details of the solutions we propose to the last two problems, we introduce our data marketplace model and some key concepts we will use in the next chapters.

## 5.1. Data marketplace model and general definitions

Despite the wide literature related to data marketplaces in the research community (see Sect. 2.4), most theoretical concepts in research papers regarding pricing, privacy-preserving techniques and value-based payment distribution are still under development. To the best of our knowledge, there is no practical nor commercial implementation of these novel innovative AI/ML marketplace designs yet.

Therefore, we seek to define a data marketplace model closer to the real world platforms we found in our survey paper and that actually can be implemented in practice. Figure 5.1 shows the reference DM model used throughout this third part of the thesis. It works as follows:
(1) We will assume that the DM provides buyers with a "sandbox" for them to submit their ML task, including the model $M$, the accuracy function they seek to optimise, and the test set they want the DM to use in order to evaluate such accuracy.

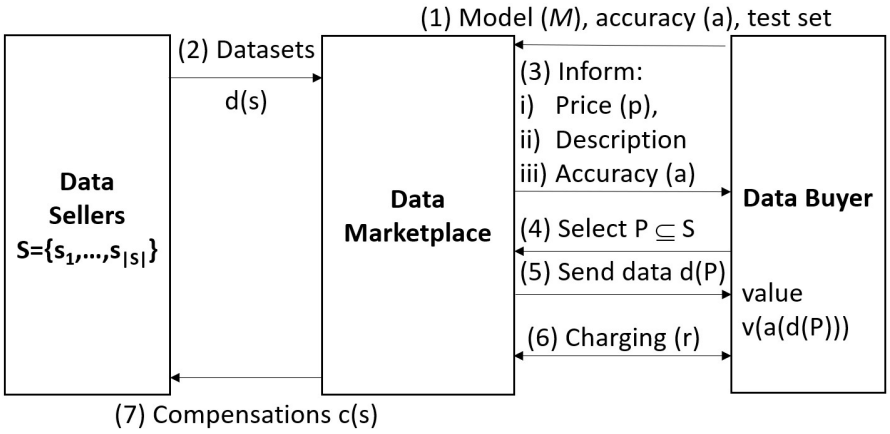



Figure 5.1: Reference data marketplace model

(2) The DM will identify suitable data for the task and obtain it from the corresponding sellers.
(3) Existing DMs typically list the datasets they offer and provide buyers with a description and a price for each dataset (or combinations of them). The "sandbox" allows buyers to test their specific model on each eligible dataset, and to get an *exact* answer in terms of the achieved accuracy, without being able to see or copy the raw data and before committing to purchase it.
(4) With that information, the buyer will make a decision on which datasets (P) to purchase and place an order to obtain them.
(5) The DM will send the corresponding data, which the buyer will use to train its model, obtaining a value $v$ that directly depends on the accuracy achieved.
(6) The DM will charge for its services and receive the corresponding payment from the buyer.
(7) Finally, the DM will share part of this payment as a compensation to data sellers, which in some cases may be done according to their contribution to the value $v$ (*Challenge 5*).

We will denote by $\mathcal{S}$ the set of suitable combinations of sellers in $S = \{s_1, ..., s_{|S|}\}$ for the ML task of a particular buyer. We will denote by $d(s)$ the dataset offered by seller $s \in S$, by $p(s)$ its *price*, and by $a(d(s)) \in [0, 1]$ the *accuracy* that the buyer's ML task can achieve if trained by $d(s)$. Similarly, for a subset of the sellers $K \in \mathcal{S}$, we will denote by $d(K)$ their aggregated dataset, and by $a(d(K))$ the maximum accuracy that can be achieved using all or a subset of the data in $d(K)$. We will also introduce the *value function* $v(a)$, that tells the value for a buyer whose ML model achieves an accuracy of $a$.

Buyers will always look forward to optimising their profit, so they will decide to purchase a dataset $d(s)$, if its (expected) marginal value exceeds its cost, i.e. $v(a(d(s|P))) \geq p(s)$. To simplify the analysis, we will assume that $v(a) = a$, i.e., that the accuracy achieved directly reflects the economic value for buyers.

When datasets are bought sequentially, we will append a subscript to identify the round to the notation we already defined. Hence, $S_n \subset S$ will refer to the set of eligible datasets in round $n$. We will denote by $P_n$ the set of data already under the buyer's control in round $n$, allowing to achieve an accuracy $a_n = max_{S' \in 2^{P_n}} a(S')$ and a value $v_n = v(a_n)$.

## 5.2. The buyers' problem

Chapter 5 introduced some technical problems data marketplaces face from the sellers' and buyers' perspective. In this chapter, we will focus on the problem of selecting which datasets to buy, given the prices set by sellers, which can be further broken down into two interrelated sub-problems: (a) compute how useful these datasets will be to their ML algorithms, something that can be captured by various accuracy metrics, and (b) compute how such accuracy can be converted into monetary gains (via improved sales, acquisition of new customers, retention of existing ones, etc.). Whereas sub-problem (b) seems to be easier for buyers [27], sub-problem (a) is inherently more challenging. Buyers need to have access to the data before they can compute its value for their ML task, but such access is only granted *after* purchasing the data – this is known as Arrow's information paradox. Sub-problem (a) is further exacerbated when the buyer can choose among $2^N$ combinations of N eligible datasets.

In theoretical works, the value of any subset of datasets for a data buyer is considered as known *a priori* [166]. In reality, however, things are completely different [149], and the most common practice is that service providers collect as much data as possible from different sources and from end users, which leads to enormous implications in privacy. Selecting suitable data from a number of different sources is already a problem in emerging niche-DMs focused on specific AI/ML tasks, such as image recognition [167] or those related to natural language processing [176]. Buyers will also need to select suitable personal data from individuals using PIMS or geo-location DMs [74]. In most DMs, data sellers provide only a description of their datasets, a price, sometimes an outdated sample, and buyers have to make purchase decisions with that information only. Some of them offer "sandboxed" environments that allow data buyers to experiment versions of the data without being able to copy or extract them [2, 22, 148]. Few of these DM also allow buyers to make offers (bids) for data when sellers do not indicate a fixed price, or if they are willing to pay something below the asking price. This case suffers as well from the fundamental problem of not knowing the value of a dataset before purchasing it.

## 5.3. Distributing payoffs among data sellers

Once a transaction is closed, the data marketplace receives a payment from the buyer and hence it must distribute (a part of) it among sellers contributing data to the transacted dataset. If the DM chooses to build its total price as the sum of the prices of individual datasets, then it may use the latter to distribute payoffs according to how much sellers asked for their data. However, some platforms set the price of a combined dataset in different ways, such as depending on the accuracy achieved in the corresponding model [3, 41]. In those cases, distributing payoffs among sellers becomes a non-trivial technical problem to solve.

Let $P$ denote a set of data sources, each one contributing a dataset $d(s_n), s_n \in P$ that has been purchased by a buyer to feed their model ($\mathcal{M}$) and optimise its accuracy $a$ over a test set.

Such accuracy is gauged by similarity metrics that define the notion of *value* of a dataset $S_K$, where $K \subseteq P$, which we denote as $v(S_K)$. Thus, $v(S_K)$ represents the accuracy of the model, according to the chosen metric, when training is performed on the data from all sources $k \in K$, and prediction is performed on the fixed test set. Let us assume that the marketplaces charges an amount $r$ to the buyer as a result of such a transaction.

Our objective is to find a value assignment method, $\mu(s_i)$, that captures the relative importance of the data originating from source $s_i \in P$ to the predictions of the model. The value assignment method $\mu(s_i)$ would thus be a function of $\mathcal{M}$, the dataset of source $s_i$, the datasets of sources in P other than $s_i$, and $v$.

$$\mu(s_i) = f(S_{s_i}, \{S_j\}, \mathcal{M}, v), j \in P - \{s_i\} \tag{5.1}$$

In this setting, compensations would be $c(s_i) \propto \mu(s_i)$, and the sum of them a fraction of the transaction charge $r$, excluding a percentage the marketplace keeps to cover the processing involved in the purchasing process. Usually, data marketplaces may think of forcing $c(s_i) >= 0, \forall s_i \in P$ to avoid making sellers "pay" for selling their data, which could be difficult to be explained and may undermine the trust in the marketplace.

The Shapley value is widely acknowledged as a "fair" method to distribute the value of a game between the players of a coalition in the game theory and ML literature. Even though other techniques and notions of value have been explored to establish the value of data in ML settings [193, 199, 200], we focus on the Shapley value, since it is by far the most widely used by the research community. For that reason, we introduce this concept in Sect. 5.3.1, and we present an illustrative toy example of how it works in Sect. 5.3.2. However, the exact calculation of the Shapley value is computationally challenging. Section 5.3.3 refers to works of the research community related to the approximation of Shapley values in different settings.

### 5.3.1. Using the Shapley value to compute the relative value of data

Establishing individual contributions of players to a collaborative game has long been a central problem of cooperative game theory. To this end, Shapley proposed that a player's value should be proportional to their average marginal contribution to any coalition they may join [165].

Let $S$ be a set of sources and $d(S)$ their aggregate data, with value $v(S)$. The *Shapley value* is a uniquely determined vector of the form $(\phi(s_1), ..., \phi(s_{|S|}))$, with $s_1, ..., S_{|S|} \in S$, where

$$\phi(s_i) = \sum_{K \subseteq S \setminus \{s_i\}} \frac{|K|!(|S| - |K| - 1)!}{|S|!} [v(S_K \cup S_{s_i}) - v(S_K)], \tag{5.2}$$

$K \subseteq S \setminus \{s_i\}$ takes the value of all possible coalitions of sources, excluding $s_i$, the expression $v(S_K$ represents the value of the combined data from the $K$ sources, and the expression $v(S_K \cup S_{s_i})$ represents the value of the combined data from the $K$ sources and source $s_i$. The Shapley value is acknowledged as a fair method of assigning value to players in cooperative game theory

because it is the only method mapping values of a game to payoffs to players that meets the following salient properties:

• **Efficiency**: The sum of Shapley values is the value of the grand coalition:

$$\sum_{s_i \in S} = v(S). \tag{5.3}$$

• **Symmetry**: Equivalent players get the same Shapley values:

$$\forall K \subseteq S, v(K \cup s_i) = v(K \cup s_j) \Rightarrow \phi(s_i) = \phi(s_j). \tag{5.4}$$

• **Linearity**: If two ML tasks with value functions v and w are combined, the Shapley values of the combination correspond to the sum of Shapley values of the individual tasks:

$$\phi^{v+w}(s_i) = \phi^v(s_i) + \phi^w(s_i). \tag{5.5}$$

• **Null-player**: the Shapley value of a source whose contribution is zero (null-player) is zero:

$$\forall K \subseteq S, v(K \cup s) = v(K) \Rightarrow \phi(s) = 0. \tag{5.6}$$

• **Strict desirability** [126]: The Shapley value of a source $i$ whose contributions to the accuracy of the task are better than another source $j$, is higher than the Shapley value of $j$.

$$\begin{gathered}\forall s_i, s_j \in S, (\exists B \subseteq S \setminus \{s_i, s_j\}, v(B \cup \{s_i\}) > v(B \cup \{s_j\})) \wedge \\ (\forall C \subseteq S \setminus \{s_i, s_j\}, v(C \cup \{s_i\}) \geq v(C \cup \{s_j\})) \Rightarrow \phi(s_i) > \phi(s_j).\end{gathered} \tag{5.7}$$

We can use the Shapley values, according to Eq. 5.1, as a credit assignment method, where $\mu(s_i) = \phi(s_i)$. Should the marketplace be willing to avoid negative compensations to sellers, there is an additional consideration to take into account. Shapley values may be negative if the average marginal contribution of a data source is negative on average, even when it adds net value to the coalition of the rest of data sources. To avoid negative Shapley values propagating to negative compensations to sellers, the platform will need to take care of those negatives before calculating the compensations. At the cost of breaking the efficiency property, a straightforward solution would be considering only positive Shapley values in calculating those compensations:

$$\mu(n_i) = \begin{cases} 0 & \phi(n_i) <= 0 \\ \phi(n_i) & \phi(n_i) > 0 \end{cases} \tag{5.8}$$

### 5.3.2. A toy model

Let us make an example to understand how Shapley values work in practice. Let us assume a simple model to estimate the average age of a population $S = \{Alice, Bob, Carlos\}$, whose ages are $d(S) = \{20, 10, 60\}$, 30 years on average. We will use an accuracy function based on the relative error of the estimation compared to a target average age (e.g., an average age of a test sample, a guess of the buyer, etc.) $\tau$,

$$v(K) = \begin{cases} 1 - \frac{|\overline{d(K)} - \tau|}{\tau} & |\overline{d(K)} - \tau| <= \tau \\ 0 & |\overline{d(K)} - \tau| > \tau \end{cases} \tag{5.9}$$

which means a correct estimation would yield 1 value and an estimation of an average age of 10 years would yield 0.33 value. In the base scenario, we will also assume that the average test value is that of the test set, i.e., $\tau = \overline{d(S)}$.

First, to compute the Shapley value, we need to calculate the value of any possible permutation of the sources, which means $|S|!$ training and evaluation cycles. Since the value of a subset of players K ($v(K), K \subseteq S$) just depends on the elements in K and not in the order of the elements in this case, $v(S)$ needs to be calculated for the $2^{|S|}$ possible combinations of the data sources only. Table 5.1a summarises the value of any combination of data sources. To simplify the notation, we have referred to the combination $K = (Alice, Bob, Carlos)$ as "$ABC$".

Table 5.1: Calculation of the Shapley Value

(a) v(S) for any combination of data sources

| $K \subseteq S$ | Guess | Error | $v(K)$ |
|---|---|---|---|
| A | 20 | 10 | 0.67 |
| B | 10 | 20 | 0.33 |
| C | 60 | 30 | 0 |
| AB | 15 | 15 | 0.5 |
| AC | 40 | 10 | 0.67 |
| BC | 35 | 5 | 0.83 |
| ABC | 30 | 0 | 1 |

(b) Calculation of Shapley value of Alice

| Permutation | Marginal value | $v(S_K \cup S_{s_i})$ |
|---|---|---|
| ABC | $v(A)$ | 0.67 |
| ACB | $v(A)$ | 0.67 |
| BAC | $v(BA) - v(B)$ | 0.17 |
| BCA | $v(BCA) - v(BC)$ | 0.17 |
| CAB | $v(CA) - v(C)$ | 0.67 |
| CBA | $v(CBA) - v(CB)$ | 0.17 |
| Average marginal value | | 0.42 |

Once we know the value of any possible combination of data sources, we can directly apply eq. 5.2 to calculate the exact Shapley values of each element in $S$. Table 5.1b shows how the Shapley value has been calculated for *Alice*. The first column lists all the possible permutations of the data sources, the column "Marginal value" shows how to calculate the marginal value that *Alice* is bringing to previous "players" in each permutation, and the third column shows marginal values. The Shapley value at the bottom is the average marginal contribution[1].

As a result, we find that *Alice*'s Shapley value is 0.42 in this case. If we repeat this exercise for *Bob* and *Carlos* we will find that their Shapley values are 0.33 and 0.25 respectively. The closer the player is to the target average age we are trying to guess, the higher its Shapley value. In this case, it is *Alice*'s data who proved to be more valuable for this specific task.

Generally, we say that Shapley values provide a value-based distribution of payoffs because they are tied to the value that data of each seller brings to the ML task submitted by the buyer. Therefore, Shapley values will strongly depend on the model, the test set, the valuation function submitted by "buyers", and the data already available to them. Next we repeat the former exercise varying each of these factors in the baseline scenario (*use case 1*) one at a time.

[1]For the sake of clarity, we have respected the order of the elements of the permutation in the "Marginal value" column. When reproducing the calculation, it must be taken into account that the value of a permutation does not depend on the order of its elements, and hence $v(BA) = v(AB)$.

***Use case 2 - changing the model***. Let us assume we are trying to find the oldest person in a given population instead of its average age. This means that we are using a different model in this case, one that returns the maximum age instead of the average of people whose data we have purchased. If we repeat the former exercise, we will find that *Alice*'s data is no longer the most valuable, but it is *Carlos*', the oldest person in the sample. In this example, the Shapley values of *Alice*, *Bob* and *Carlos* are 0.14, 0.06, and 0.8 respectively.

***Use case 3 - using a biased test set***. Unlike what happens in the base case scenario, the test set used by buyers as ground truth shows biases, or it simply does not reflect the universal truth in real settings. As a result, the target $\tau$ sought by buyers is not necessarily aligned with that achieved by using all the data available. In the base case scenario, such target is aligned to the average age of the population in $S$, i.e., $\tau = \overline{d(S)}$. As a result, the model achieves the maximum accuracy when we buy all the data in S, and thus Shapley values sum to one. However, this is not usually the case in reality. Since the Shapley value adapts to the objectives pursued by buyers in their ML task, using a different test set can have a significant impact on the Shapley values.

Let us suppose that the buyer, based on beliefs or estimates, is looking for people that are on average 33 years old (*use case 3*). In this scenario, the Shapley values of *Alice*, *Bob* and *Carlos* are 0.32, 0.24, and 0.35 respectively. *Carlos* becomes more valuable because now the buyer is looking for older people on average. The Shapley values now sum to 0.91 which is the accuracy achieved when using all the available information and comparing $\overline{d(S)}$ to $\tau$ in this second case.

***Use case 4 - using RMSE as the valuation function***. Changing the valuation function also affects the results of the model. Let us assume that we use the relative mean square error instead the relative absolute error in the valuation function of *use case 4*, namely:

$$v'(K) = \begin{cases} 1 - \frac{|\overline{d(K)} - \tau|^2}{\tau^2} & |\overline{d(K)} - \tau| <= \tau \\ 0 & |\overline{d(K)} - \tau > \tau \end{cases} \tag{5.10}$$

In this case, the differences between the value of *Alice*'s and *Bob*'s data to the value of *Carlos*' increases, due to the additional penalty that RMSE imposes to larger errors. In this example, the differences between the Shapley values of *Alice*, and *Carlos* grows from 0.17 to 0.33, meaning that Carlos would receive a much lower payback for his data if the buyer and the marketplace opt for RMSE as the valuation function.

***Use case 5 - using already existing data***. Finally, the Shapley values also depend on the data already available to the buyer at the beginning of the purchasing process. Imagine the buyer already has purchased data from *Dennis* and *Eric* whose age is 40. This means that the average of all the population will be 34 in this case, and that the buyer already achieves a 0.82 accuracy even without purchasing any data. Notice that in this case, negative marginal contributions start to appear. For example, using only *Carlos*' data misleads the model and yields a -0.2 marginal contribution. As a result, if we carefully repeat the exercise we find that the Shapley values of

*Alice*, *Bob* and *Carlos* are 0.11, 0.09, and -0.02, respectively. Remarkably, Shapley values sum to 0.18 in this case, which is the maximum marginal value that sellers can bring to the buyer in this case due to the data already available in the beginning of the purchasing process.

Table 5.2 shows how Shapley values change for slight variations of this base use case. It shows that standalone small variations in the model, the valuation function, the test set or the initial data have a dramatic impact on the value that data of different players bring to the buyer.

Table 5.2: Shapley values for slight variations of our toy example

| 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 |
| 5 | Existing Data | 0.11 | 0.09 | -0.02 | 0.18 |

Despite the simplicity of this toy model, the variations illustrated in these use cases reflect real issues that happen when the Shapley value is used to approximate the relative value of data in real settings working with more complex ML models. The resulting values will heavily depend on the model and the valuation function chosen by the buyer, and they will be affected by any existing data and by any bias in the test set.

Unfortunately, the exact calculation of Shapley values is only feasible when the number of data sources is low, and it suffers from severe scalability problems when that number increases. Even calculating the Shapley values of a few data sources for heavy ML models that take time to be trained can be challenging, as we will show in chapter 7. This is why researchers resort to different techniques to *approximate*, rather than exactly *calculate*, Shapley values. Next, we summarise such techniques, and we end the chapter with a discussion about Shapley "fairness".

### 5.3.3. Approximating Shapley values

Previous works have already used the Shapley value to measure the utility of training data for different ML problems [160]. Unfortunately, calculating the Shapley values has also been proven to be NP-hard for many domains [19]. Since it takes into account all possible coalitions, for each source, the number of terms scales with $2^{|S|}$, where $|S|$ represents the number of sources, therefore it quickly becomes computationally unfeasible.

Several works have looked at computational aspects of Shapley value and have developed efficient approximation algorithms. Such approximations come also at a cost. Even though the definition of Shapley is general in scope, its approximations must often be tailored to a particular type of problem. An algorithm that works well for certain problems is not necessarily applicable or simply does not work so well for others. Some authors warn that some approximations may also undermine the "fairness" of exact calculations of Shapley values [205].

We have identified three different families of approximation techniques, allowing for orthogonal improvements in the process, namely: i) reducing the number of coalitions to evaluate, ii) reducing the dimensionality, and iii) improving the efficiency of the calculation.

***Reducing the number of evaluated coalitions***

One family of techniques approximate Shapley by calculating the marginal contributions of only a fraction of all possible permutations of the data sources. For example, they resort to random sampling methods [35], whose error is bound by the size of the sample and the variance of the values, or they use Monte Carlo algorithms that define conditions to stop evaluating random permutations once they are met by the interim approximate Shapley values (e.g., their relative variation being below a certain threshold after evaluating a minimum number of r permutations) [75]. Bayesian inference has been proposed to accelerate the execution of Monte Carlo approximations to Shapley, taking advantage of the interdependence of Shapley values of the different players [181]. Other works have shown that structured sampling helps in dramatically improving the efficiency of Shapley approximations in certain problems [63, 191].

***Reducing the dimensionality of the problem***

A different family of approximations reduce the complexity of Shapley value calculation by decreasing the number of sources $|S|$ for which the algorithm is trained and evaluated. They apply clustering and dimensionality reduction [151], and algorithms such as k-nearest neighbors [95] and approximate Shapley value in close to linear time.

***Improving the efficiency of the calculation of the marginal value***

Recent works propose using appraisal functions adapted to the current model, and they also introduce multi-party computation to evaluate and select private training data from a marketplace [197]. Working with such heuristics, for example using ML influence functions [193], the data marketplace can circumvent the need to retrain the model for each combination of data sources, and thus can remarkably accelerate the process of distributing payoffs compared to exactly calculating Shapley value or even approximating Shapley [96].

### 5.3.4. Shapley fairness in the context of a human-centric data economy

Selecting the most suitable data for a ML task and fairly rewarding users contributing data to a transaction are key processes to bootstrap a successful data marketplace. Both processes are somehow intertwined, and the platform can use the input of the evaluation of sellers' data for a buyer's ML task during data selection process to calculate the contribution of individual sellers later on. Most data valuation algorithms cited in this chapter allow to discriminate and filter data or data sources according to the value they bring to a particular task fed into the system by the buyer. This is particularly interesting for buyers to make the most out of the data stored in the marketplace, but it might not always be desirable when distributing payoffs for data among users contributing data to the platform in a human-centric data economy.

Distributing payoffs according to the value the data of a person brings may motivate users in providing more valuable data to the platform, as long as they are given a precise idea of what is meant by "valuable data". Explainable AI and friendly user interfaces and dashboards will be key to achieve this objective. So will be transparency and accountability regarding data usage and the contributions made by each user.

Even if explainability, transparency and accountability are thoroughly achieved, a value-based distribution of payoffs might not be always desirable. In some settings, users have no control over the "quality" of their data (see the toy example of estimating the average age of a population, we cannot decide on our age!), and providing such incentives will turn out to be useless, and may even frustrate less valuable end users in these situations. For example, we may think of people contributing their personal health data to the healthcare system, as opposed to other use cases such as people annotating training data for image classifiers, or taxi drivers contributing their location data to the city hall to improve mobility. Patients cannot generally choose to report valuable health-related data to fight the effects of a disease unless they suffer from it, whereas taxi drivers may choose to drive around certain areas or within specific time frames, stop and wait for new customers, or seek for them in the city while they produce more mobility data. Similarly, people annotating images may pay more attention and improve the quality of the training data they produce if it is found to be faulty. Whereas distributing payoffs according to value might be beneficial to improve the amount and quality of the data reported by users in the last two cases, it is not so clear that it is a convenient policy to incentivise the sharing of healthcare data.

Acknowledging this, the data marketplace architecture defined in this chapter allows for the necessary flexibility for buyers and data marketplaces to freely decide how to select and how to distribute payoffs for data in a human-centric data economy. It allows but not obliges to distribute payoffs to sellers based on the value they bring to the specific task. Nor does the design stick to using the actual buyer's ML task to select data and to distribute payoffs. As an alternative, buyers could use simpler versions of their models, or models that detect and discard outliers and distribute rewards uniformly among the rest of data samples. Whatever model the buyer and the marketplace choose for selecting data, coherency calls for using the same model to reward data sources based on their value. By doing this, the processing and information generated during data selection can be reused to accelerate any approximation to Shapley when distributing payoffs. Still, our data marketplace design allows different models for these two tasks, if deemed appropriate.

In the next chapters, we explain our contributions to the problem of buyers selecting suitable data, and to the problem of data marketplaces distributing payoffs based on the value of data. We apply them to the context of spatio-temporal data being reported to feed ML prediction models.

# Chapter 6

# Try-Before-You-Buy - a novel data purchasing strategy

In previous chapters we have identified that buyers oftentimes face problems when sourcing suitable data for their specific ML tasks. This is due to the special characteristics of "data" as a tradable asset: data is an experience good whose value heavily depends on the purpose for which it is eventually used and its context, hence similar data valuable for a task A is not necessarily valuable for another task B. Because of this, letting buyers know the particular value of a piece of data for them before acquiring it would definitely be a valuable feature of data marketplaces. However, granting access to such data before closing a transaction, even if it is just for assessing its value, may kill the business if a malicious buyer is able to copy it.

In this chapter, we propose a method for optimising data purchasing decisions by allowing buyers to try data on their own models and evaluate how suitable it is for their specific task, that without gaining access to such data. In practice, this means adding a data appraisal step to the purchasing process before consumers make a decision on which/whose data to acquire. We show that if a marketplace provides potential buyers with a measurement of the performance of their models on *individual* datasets, then they can select which of them to buy with an efficacy that approximates that of knowing the performance of each possible combination of datasets offered by the DM. We call the resulting algorithm *Try Before You Buy* (TBYB).

Overall, TBYB aims to increase the efficiency of buying datasets online and to reduce the risk of sourcing processes. We believe that this is key to allowing both DMs and global data supply to grow. Existing commercial marketplaces such as Advaneo, Otonomo or Battlefin already offer "sandboxes" for potential buyers to play with outdated samples of data before acquiring it (see Sect. 3.2.4.9 for more information). We believe that such functionality could be easily extended to implement TBYB in real data marketplaces.

Furthermore, anticipating the value of data is even more critical in a human-centric data economy that empowers users to take control of their data. In this setting, data consumers would have to choose (and eventually pay for) suitable data among that of thousands or millions of individ-

uals. However difficult implementing this vision may be, paying for data would certainly help address the existing sort of tragedy of commons around privacy online due to the current practice of digital service providers amassing as much information as possible about end users. Having to pay for it would automatically limit the appetite for data, which would then seek to limit the amount of data they acquire to the minimum necessary for their task.

The rest of the chapter is structured as follows. First, Sect. 6.1 defines the different purchase strategies we will be comparing TBYB to, ranging from the optimal purchase to more naïve strategies lacking any information about the value of datasets for the particular task, and hence much closer to the situation that buyers face nowadays. Section 6.2 describes the methodology and the synthetic model that we design and implement to compare the outcome of those strategies and their results under different conditions. Then we validate those results using real spatio-temporal datasets from companies and anonymised individuals in use cases related to demand and travel time prediction in metropolitan areas in Sect. 6.3. Finally, we present our conclusions and key takeaways from this analysis in Sect. 6.4.

## 6.1. Data Purchase Strategies

Throughout this chapter, we will assume the data marketplace model and the buyers' problem previously described in chapter 5. In this section, we present a series of purchase strategies that buyers may follow to acquire data in such a setting, including our proposed algorithm called Try-Before-You-Buy (TBYB).

### 6.1.1. Optimal Purchase

The *optimal purchase* assumes that the buyer knows $a(d(S))$ for any subset $S \in \mathcal{S}$. This allows for an optimal purchase $\mathcal{S}^\star$ that maximises the profit, i.e., the difference between the value that the buyer extracts from the data, and the cost paid to purchase them:

$$\mathcal{S}^\star = \arg\max_{S \in \mathcal{S}} \left( v(a(d(S))) - \sum_{s \in S} p(s) \right), \tag{6.1}$$

subject to $v(a(d(\mathcal{S}^\star))) \geq \sum_{s \in \mathcal{S}^\star} p(s)$.

Such a full information scenario is optimal from a buyer's perspective, but not scalable nor practical: a DM would need to compute the accuracy of each ML algorithm over $2^{|S|}$ combinations of eligible datasets.

### 6.1.2. Try Before You Buy

We propose that the DM provides buyers with the accuracy of their models trained on individual eligible datasets, but not on combinations of them. The algorithm is sequential and greedy in nature, and can run for up to $|S|$ iterations. We will consider two versions.

#### 6.1.2.1. Stand-alone version - S-TBYB

*Stand-alone Try Before You Buy (S-TBYB)* assumes that the marketplace provides the accuracy of individual datasets on their model, that is, the buyer knows $a(d(s))$ on all $s \in S$. Then S-TBYB starts buying datasets in descending order of *expected profit* until a stop condition is reached.

For the first dataset, the profit is not expected but exact, so the best dataset $s^\star \in S$ is bought provided $v(a(d(s^\star))) - p(s^\star) \geq -\lambda \cdot v(a^\star)$, where:

- $a^\star \leq 1$ is the best accuracy that can be delivered, as informed by the data marketplace, and

- $\lambda$ is the risk parameter and models the maximum admissible loss for the buyer, relative to the potential value of the sourcing operation ($v(a^\star) - v_n$ in round n). For example, $\lambda = 0.1$ means that buyers will buy $s^\star$ if its price is lower than the marginal value they expect to get plus $10\%$ of the maximum value that they could add by buying new data.

Assuming some risk is sometimes necessary. For example, in some sourcing problems, the marginal value of new data increases as more information is bought. In such a setting, buyers may be required to assume some temporary losses when acquiring the first datasets, in the hope that they provide additional accuracy, and become profitable when fused together with other data.

**n-th iteration**. The buyer will proceed as follows:

1. Identify the best possible dataset $s_n^\star \in S_n$ such that $s_n^\star = \arg\max_{s \in S_n} \left(v(a(d(s))) - p(s)\right)$
2. Purchase $s_n^\star$ if its estimated marginal value exceeds its price, and a risk threshold that depends on the remaining expected value to be obtained out of the operation, i.e., if $v(E\{a(d(s_n^\star \cup P_n)\}) - v_n - p(s_n^\star) \geq -\lambda \cdot (v(a^\star) - v_n)$
3. If the buy condition is met then, $s_n^\star$ is added to the purchased dataset, $P_{n+1} = P_n \cup d(s^\star)$, and the next round starts,
4. Else if no dataset in $S$ meets this requirement, then the process stops

The estimation $E\{a(s^\star \cup P_n)\}$ is specific to each problem. We estimate the relative added accuracy of $s^\star$ by multiplying its individual accuracy $a(s^\star)$, and the ratio of the marginal contribution and the individual accuracy of the last purchased dataset:

$$E\{a(s_n^\star \cup S_n)\} = \frac{a_n - a_{n-1}}{a(P_n - P_{n-1})} \cdot a(s_n^\star) \cdot (a^\star - a_n). \tag{6.2}$$

#### 6.1.2.2. Assisted version - A-TBYB

*Assisted Try Before You Buy (A-TBYB)* assumes the buyer is allowed to ask the marketplace *every round* for the marginal accuracy of any still eligible datasets.

**n-th iteration**. The purchase process will be the following:

1. Ask the marketplace for complementary datasets $S_n \in \mathcal{S}$, and $a(d(s)|P_n), \forall s \in S_n$ given the task $(\mathcal{M}, a)$ and $P_n$

2. If there are eligible datasets, i.e., $S_n \neq \emptyset$:

   - Identify the best possible eligible dataset $s_n^\star \in S_n$ such that $s_n^\star = \arg\max_{s \in S_n} (v(a(d(s \cup P_n))) - p(s))$
   - Buy provided $v(a(d(P_n \cup d(s_n^\star)))) - v_n - p(s_n^\star) \geq -\lambda \cdot (v(a^\star) - v(a(P_n)))$
   - If the buy condition is met then, $s_n^\star$ is added to the set of controlled datasets: $P_{n+1} = P_n \cup d(s_n^\star)$ and the next round starts
   - Else if the buy condition is not met, then the process stops

As a result, the model will be processed a maximum of $\sum_{i=0}^{r-1} |S| - i$ times for $r$ rounds, and no estimation of the marginal value is needed. To prevent abuses, a marketplace implementing this solution may charge (part of) the processing costs of TBYB to buyers, or may set up a maximum limit of trials for a certain task, that can be updated as the buyer purchases data.

### 6.1.3. Buying without trying

Most commercial marketplaces provide buyers with a description of datasets, their metadata, source, procedure used to collect them, etc. Buyers must make a purchase decision based on this information, the reputation of the sellers and the asking prices. In this section, we describe these value-unaware strategies to which we will compare the performance of TBYB later on.

#### 6.1.3.1. Volume-based purchasing

Most data marketplaces provide buyers with a description of datasets including their volume (e.g., nº samples), which is often used to choose among the different offers. Let $vol(s)$ denote the volume of dataset $s$, used as merit figure by the following volume-based purchasing heuristic:

**n-th iteration**. We will assume that a greedy buyer would select the dataset $s_n^\star \in S_n$ with the highest $vol(s)/p(s)$ ratio every round, provided:

$$p(s_n^\star) \leq -\lambda \cdot (v(a^\star) - v_n), \tag{6.3}$$

which assumes that even in the worst case the maximum relative admissible loss is not exceeded in the operation.

#### 6.1.3.2. Price-based purchasing

It may happen that the marketplace just publishes the list of suitable datasets $S$, and their prices. This setting resembles real situations where information about data offer is insufficient or misleading to the buyer's purposes. We assume such a buyer would randomly select among datasets whose price is lower than their maximum relative admissible loss. *Price-based purchasing* assumes buyers randomly select among datasets whose price is lower than their maximum relative admissible loss.

**n-th iteration**. The buyer will randomly select one of the datasets in the set of eligible datasets that avoid exceeding the maximum loss if any of them is bought, that is, the set $S_n \subseteq \mathcal{S}$ such that, $\forall s \in S_n, p(s) \leq -\lambda \cdot (v(a^\star) - v_n)$. If $S_n = \emptyset$ then the process stops.

## 6.2. Performance evaluation with synthetic data

We will use synthetic data to evaluate the performance of the different purchase strategies presented in Sect. 6.1 across a wide range of parameters. Our synthetic model is easy to reproduce, captures a wide range of parameters, and allows us to extract useful insights about the relative performance of different data purchase strategies. As we will show later in Sect. 6.3, our conclusions from this section are also validated by results with real data and use cases.

### 6.2.1. Synthetic model description

We will denote as $TCOD$, which stands for *total cost of data*, the cost of buying all the available datasets in $S$, i.e., $TCOD = \sum_{s \in S} p(s)$. Hence, if $TCOD < a(d(\mathcal{S}^\star))$ the buyer is guaranteed to make a profit. In the more interesting case of $TCOD \gg a(d(\mathcal{S}^\star))$, the buyer needs to carefully select which datasets to buy in order to avoid ending up with a loss. Our synthetic model assumes that:

$$\forall K \subseteq S, v(a(d(K))) = a(d(K)) = \left( \frac{\sum_{s_i \in K} DI^i}{\sum_{s_i \in S} DI^i} \right)^{MUP}, \tag{6.4}$$

where:

- **MUP is the Marginal Utility Profile** parameter that controls the concaveness/convexity of $v(\cdot)$ from 0 to 1 as more data is bought. $MUP < 1$ means that buying additional datasets will have a decreasing marginal utility for the buyer. $MUP > 1$ means that the marginal contribution of new data sources will increase as more datasets are bought. Finally, $MUP = 1$ means that all datasets yield the same marginal accuracy regardless the purchase sequence.

- **DI is the Data Interchangeability parameter** that controls the relative value of different datasets in $S$. $DI = 1$ means making all datasets fully interchangeable, that is, it only matters how many datasets are bought, but not which ones. For $DI > 1$ and $MUP = 1$, dataset $s_i$ is DI times more valuable than dataset $s_{i-1}$, $1 \leq i \leq |S|$. It is only if $DI \neq 1$ when not only how many datasets are bought, but also which ones matters.

With regards to pricing, we will consider the following schemes:

P1) All datasets having the **same price** ($p = TCOD/|S|$).

P2) Datasets with **uniformly distributed random prices**.

P3) Dataset prices **based on their Shapley values**, as previously introduced in Sect. 5.3.1.

Figure 6.1 shows the methodology we used to test the different purchase strategies. We developed a simulator to run purchase processes for different values of the parameters of our synthetic model like MUP, DI, TCOD and for different pricing options. As outputs, the simulator returns the datasets purchased, their cost, the accuracy and the profit obtained by the buyer,

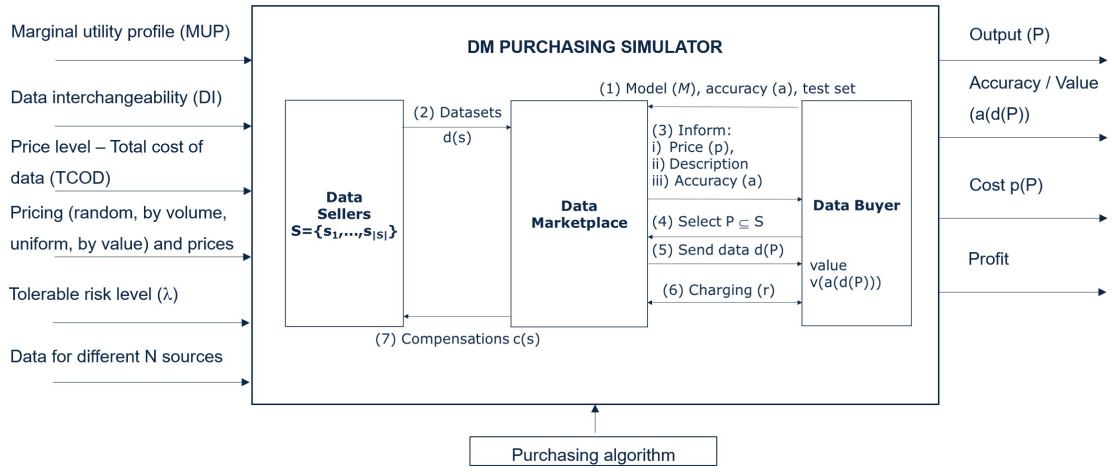



Figure 6.1: Data marketplace purchase simulator

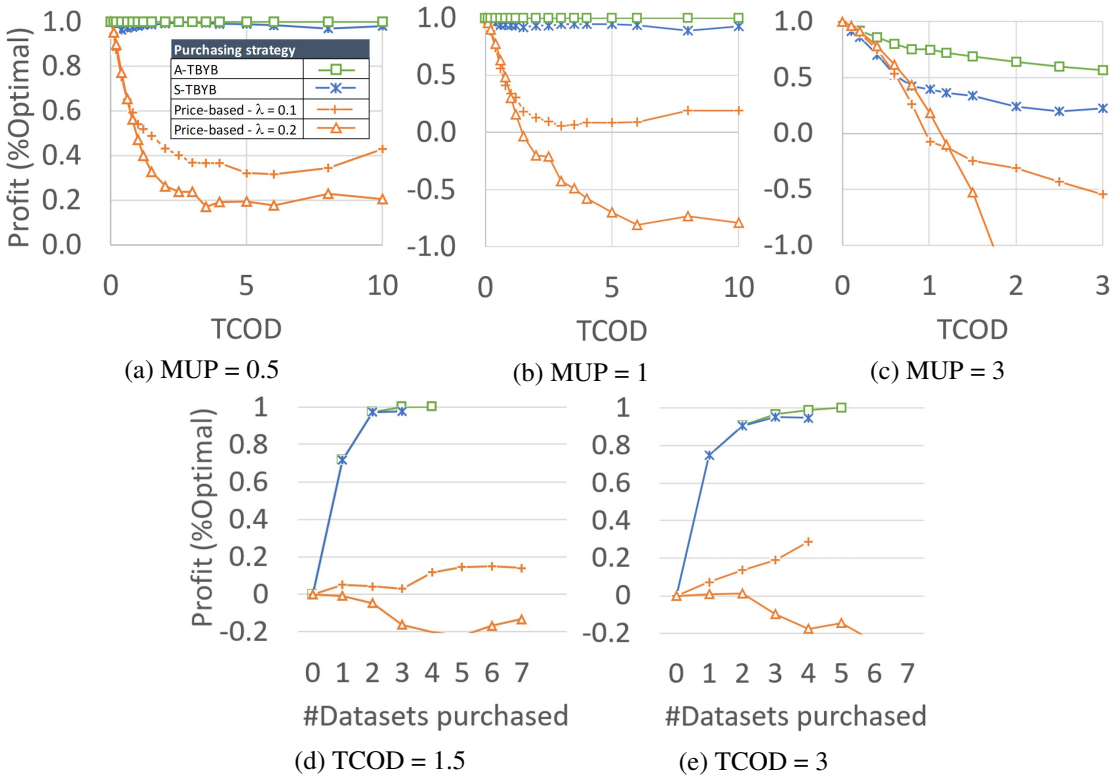
Next we will compare the profit of TBYB to that of the optimal purchase, and to that of other value-agnostic heuristics. Whenever randomisation is used, e.g., in pricing datasets or in executing value-unaware strategies, we report average values of 50 executions.

### 6.2.2. Results for different utility profiles

Assuming that some datasets are more important than others ($DI = 2$), Fig. 6.2 (a-c) shows the average outcome of the purchasing processes following different strategies for different MUP. In the plots, we compare their outcomes to the optimal profit for the buyer.

A-TBYB matches the optimal purchasing for both concave ($MUP = 0.5$, subplot 6.2a) and linear ($MUP = 1$, subplot 6.2b) value profiles, across the entire range of TCOD values. Similarly, the simpler S-TBYB, yields above 90% of the optimal profit for concave and linear MUPs, even when the performance of value-agnostic heuristics drops below 50% and even leads to losses when $MUP \geq 1$. Under convex value profiles ($MUP = 3$, subplot 6.2c), all strategies underperform because reaching a higher accuracy (and therefore a higher value for buyers) requires buying more datasets, which, in turn, eats away the profit margins for buyers. Even in these cases, TBYB yields a profit and avoids losses if $TCOD \gg 1$. A risk-prone buyer i.e., assuming a higher $\lambda$, purchasing data based on prices will generally do better if prices for data are low (TCOD less or close to 1), but will face more loses if prices are higher.

To explain why TBYB outperforms the value-unaware strategies, we plot in Fig. 6.2 (d-e) average "*purchase sequences*", showing the buyers' average profit as they purchase datasets using different algorithms. TBYB algorithms buy both the most valuable datasets (they achieve higher accuracy from the first round), and the right number of them (they stop buying before profits decrease). On the contrary, value-unaware heuristics overbuy, and randomly select affordable datasets respecting the buyer's risk appetite, which generally leads to lower profits, or even loses for risk-prone buyers.



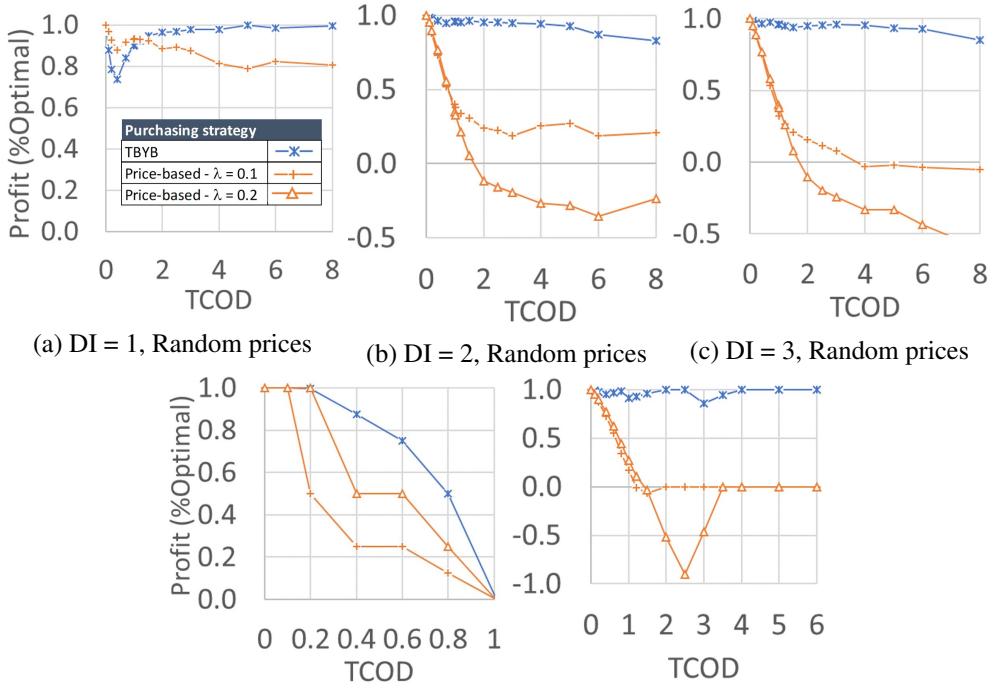

(a) MUP = 0.5 (b) MUP = 1 (c) MUP = 3

(d) TCOD = 1.5 (e) TCOD = 3

Figure 6.2: Profit vs. TCOD for different MUP and value-unrelated prices (a-c). Purchase sequences for MUP = 1 (d-e))

### 6.2.3. The effect of data interchangeability

To find out how TBYB is affected by the interchangeability of datasets, we have run a set of simulations for different values of the parameter DI. Figures 6.3 (a-c) show the relative profit of different purchase algorithms for different DI values assuming MUP=1. A-TBYB always matches the optimal purchase.




(a) DI = 1, Random prices (b) DI = 2, Random prices (c) DI = 3, Random prices




(d) DI = 2, value-based prices (e) DI = 2, Equitable prices

Figure 6.3: Profit (MUP = 1) for different DI and pricing schemes

The subplot on the left depicts results for perfectly interchangeable datasets (DI=1), whereas the next two show cases of datasets that are increasingly less interchangeable (DI=2 and DI=3). These plots show that the performance benefits of TBYB over the heuristics increase when eligible datasets have different value, alone and combined with other datasets. On the contrary, knowing the value of data beforehand is less advantageous when they are almost interchangeable.

### 6.2.4. The effect of data pricing

So far we have used prices that do not reflect the value of different datasets. In this section, we look at the effect of dataset pricing on the performance of TBYB, especially when prices are set proportionally to their value for an ML algorithm.

Subfigures 6.3 (b, d and e) show the performance of our purchasing algorithms for different pricing models. A-TBYB always matches the optimal purchase in these cases, too. Pricing based on value significantly reduces the gap between S-TBYB vs. price-based purchasing strategies, although S-TBYB still outperforms price-based purchasing for the same level of risk. Notice,

however, that pricing data in accordance to their actual value for buyers is not realistic, since such value depends on the task and the buyer, who has no incentives to disclose it to sellers.

### 6.2.5. Summary

Table 6.1 summarises the impact of changing the parameters of our synthetic model in the performance gap between TBYB and price-based purchasing. Bottomline, TBYB, even in its simplest, stand-alone version, always outperforms value-unaware heuristics, especially in the most realistic scenarios involving high TCOD, concave value functions, non-interchangeable datasets, and value-unaware pricing. In most cases, TBYB is very close to the performance of an optimal purchase using full information.

Table 6.1: Impact of parameters on the gap between TBYB and price-based purchasing

| Param. | Impact | Explanation |
|---|---|---|
| `TCOD` | The higher TCOD, the more valuable TBYB | More difficult for other strategies to find the right datasets in terms of price - value to buy |
| `MUP` | The higher MUP, the more difficult to find the optimal. TBYB loses effectiveness but still outperforms other algorithms | TBYB buys more valuable datasets, minimises temporary losses and limits the risk of buyers, since it allows for a better estimation of expected marginal value of datasets |
| `DI` | The less interchangeable datasets are, the more advantage of using TBYB | With perfectly interchangeable datasets, TBYB only improves the estimation of the marginal utility as information increases |
| `Pricing` | TBYB gap with price-based purchasing narrows for value-unaware prices | Price-based purchasing works better if value is embedded in price |

## 6.3. Validation with real data

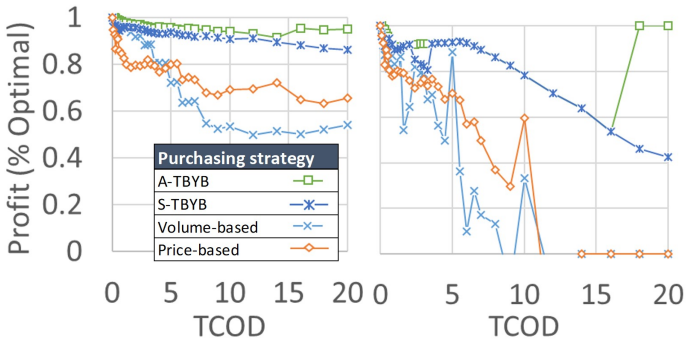
In reality, however, different datasets may mix in much more complex ways, that cannot be represented by any parameter setting of the above model. For example, $s_i$ can be very useful if combined with $s_j$, but not so useful if combined with others that individually yield the same accuracy as $s_j$. To validate our conclusions from the previous section, we tested the performance of different data purchase strategies using real spatio-temporal data reported by taxi companies and individual drivers, which we use to forecast demand and predict travel time within a city. For that purpose, we adapted the data marketplace simulator described in Fig. 6.1 to work with generic ML models, valuation functions, training and test data.

Furthermore, we introduce volume-based data pricing, which is quite common in practice as we discussed in chapter 4. *Volume-based pricing* assumes the price of a dataset is proportional to its volume (drivers or rides). We also test volume-based purchasing.

### 6.3.1. Demand prediction based on B2B data

We will assume that data buyers are looking to predict vehicle-for-hire demand in a city based on historical data, for which they can acquire data from a number of taxi companies (data sellers) through a DM. For this purpose, we have used real data reported by taxi companies to the regulatory authority of Chicago [141] containing 11.1 MM rides corresponding to the first 8 months of 2019 for $N = 16$ companies - the 15 largest ones servicing 94% of the total demand, plus a hypothetical 16th company aggregating the rest of the data.

We computed the exact Shapley value of the data of each company for the task of predicting demand using a multiseasonal SARIMA model with hourly and daily sub-components. The model was trained to predict taxi demand in the second half of March (control period) at the level of districts, for which we trained it with taxi rides from the previous six weeks (observation period). We compared the prediction to the ground truth given by the real data provided by all the 16 companies in the control period. We deliberately chose to predict the taxi demand of a medium size district (community area 11, Jefferson Park), where companies need to combine data to achieve a good prediction accuracy, and their Shapley values are very different (standard deviation equal to 76% of the average). The maximum accuracy the DM was able to give using all the information was $a^\star = 0.8963$ in this case.



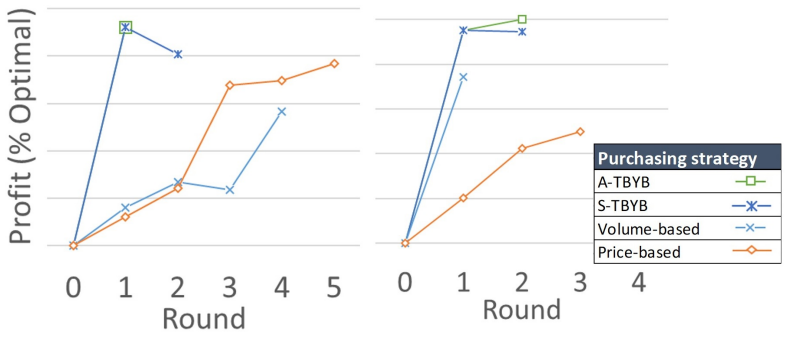

(a) Value-unrelated price (b) Volume-based price

Figure 6.4: Profit for data buyers in predicting demand using B2B data

We have simulated all purchase algorithms for different TCOD, pricing models, and $\lambda$ parameters. Figure 6.4a shows that both A-TBYB and S-TBYB achieve above 90% of the optimal buyer's profit under value-unaware prices. A-TBYB outperformed S-TBYB because it addresses the complexity of combining datasets at the (extra computational) cost of recursively asking for accuracy estimations every purchasing round. The results are in line with the ones we obtained using synthetic data.

Volume-based purchasing proved even less effective than price-based purchasing. This is because value and volume are not significantly correlated in this particular case ($R^2 = 0.54397$), hence buying depending on the nº drivers does not necessarily lead to higher accuracy.

Figure 6.4b shows the results corresponding to volume-based prices. Profit reduces faster as TCOD grows in this case than in that of value-unaware prices. This is because more valuable and larger datasets command higher prices. Still TBYB selects cheaper more valuable data.




(a) TCOD = 3 (b) TCOD = 8

Figure 6.5: Purchase sequences for volume-based prices when predicting demand using B2B data

Figures 6.5a and 6.5b show purchase sequences that illustrate why TBYB works for volume-related prices. TBYB improves price-based purchasing both by selecting the best datasets, and by stopping before profit reduces. This feature is especially relevant when data prices grow (TCOD $>> 1$). Picking datasets based on volume does not improve price-based purchasing this time.

As in the theoretical use case, the gap between TBYB and price-based purchasing narrows when prices as set proportionally to the Shapley value of datasets, for the same reasons. In particular, price-based purchasing was able to reach a moderate accuracy, far from the optimal but profitable for the buyer, by combining several of the cheapest datasets.

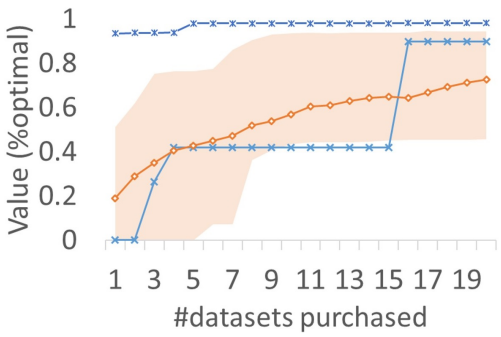
### 6.3.2. Prediction based on individuals data

Some authors have proposed that individuals should get paid for their data [114, 155]. PIMS have appeared to empower users to take control of their data and to decide who they share their data with (and at what price). Current PIMS act as a sort of data unions and usually consider data from every user equally valuable (same price for every user) or valuable according to their volume. PIMS do not disclose the identity of individuals but at most anonymised ids and some metrics (e.g., the volume of their data), and buyers must select whose data to purchase.

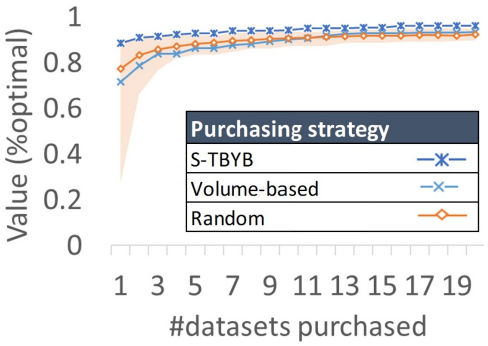
In this case, buyers' profits depend on the number of people whose data they purchased and on the accuracy achieved. Consequently, an algorithm that is able to make better predictions by picking less drivers, or less volume of data if prices are based on volume, will generally be more profitable. In this section, we will show that TBYB allows buyers to better filter individuals whose data suits their task.

For this purpose, we consider data of 2.845 drivers in the same setting we used for demand prediction of B2B data. At city level, volume-based pricing strategy matches TBYB and is able to reconstruct the demand with data of 5 drivers only. However, the situation is different at the level

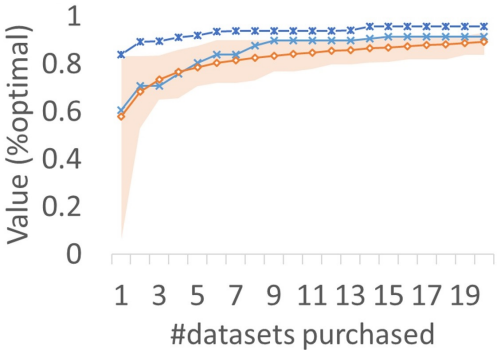
of districts. Figure 6.6a shows the purchase sequences of the different algorithms in district 56 of Chicago (Garfield Ridge), the shaded region delimiting percentile 0.1 and 0.9 of 50 executions picking drivers randomly. Whereas TBYB is able to reconstruct the demand with only 5 drivers, picking them by volume requires 49, a similar amount to the average drivers required if picked randomly. Therefore, regardless the price, TBYB always achieves higher accuracy with less drivers that the other algorithms. If prices are set by volume, TBYB leads to even bigger profits, because volume-based purchasing heads buyers for the most expensive drivers in this case.




(a) Demand prediction in Chicago district 56




(b) Travel Time prediction to Porto Airport




(c) Travel Time prediction to Porto São Bento

Figure 6.6: Value vs. nº of people whose data is purchased

We also validated our findings in a completely different setting, a random forest model that was able to predict car travel time within the metropolitan area of Porto trained on a dataset detailing 1.71 million routes of 448 taxis from Jul'13 to Jun'14 [98]. We randomly split input data between train (80%) set and test (20%) set, and trained the model to predict car travel time based on the UTM coordinates of the start and end points, the time of the year (month, weekday, hour), the type of day (workday, holidays or day before holidays) and the travel distance. The model was able to provide reasonably good accuracy ($R^2$ score above 0.8) over the test set.

In this use case, we obtained the value delivered by the different purchasing strategies when predicting car travel time to the airport (see Fig. 6.6b), and to São Bento station (see Fig. 6.6c). Similar to the previous case, TBYB outperforms volume-based and price-based purchasing strategies, by selecting the most suitable drivers. In predicting travel time to the airport, TBYB achieves more than 0.95 of the maximum accuracy with barely 5 drivers, whereas picking drivers randomly requires 25 drivers, and volume-based purchasing requires 30 drivers to get to this value.

Our empirical results validate the conclusions of the previous simulations based on synthetic data. Not only is TBYB able to provide buyers with more value, but it also reduces the number of individuals whose data is shared with third parties, hence reducing unnecessary data leakages.

## 6.4. Key takeaways

In this chapter, we have proposed a preliminary "data evaluation" phase prior to buyers selecting which datasets to buy and, in this context, we have introduced a family of dataset purchase algorithms that we call "Try Before You Buy" (or TBYB) that allow data buyers to identify the best datasets to buy with only $O(|S|)$ information about the accuracy of ML algorithms on individual datasets, instead of $O(2^{|S|})$ information required by an optimal strategy using full information. Effectively, TBYB needs to know only the accuracy of a ML model on *individual* datasets, and with this information it can approximate the optimal *combination* of datasets that maximises the profit of the buyer, i.e., the difference between the value extracted from the datasets minus the cost of purchasing them. The accuracy of individual datasets can either be pre-computed by the DM or the data sellers for some common algorithms, and be made available as part of the dataset description. Another alternative is for the DM to use recently developed "sandboxed" environments that allow data buyers to experiment versions of the data without being able to copy or extract them (hence the "Try" part on the algorithm's name; Otonomo, Swash, Advaneo, Caruso or Battlefin are examples of commercial marketplaces that implement such functionality).

We have compared the performance of TBYB against several generic heuristics that do not use information about the value of a dataset for the particular AI/ML task at hand, as well as against an optimal solution that uses full information. We have started with a synthetic evaluation and then we have validated our conclusions using real-world spatio-temporal data and a use case in predicting demand for taxi rides and travel time in metropolitan areas.

**Our findings** are as follows:

- TBYB remains close to the optimal in most scenarios, and its benefit increases with the catalogue size ($|S|$).

- TBYB is almost optimal when buying more data yields a progressively diminishing return in value for the buyer. Otherwise, TBYB finds it more difficult to match the optimal performance, although it still outperforms the heuristics.

- The benefit of TBYB becomes maximal when prices of datasets do not correlate with their actual value for the buyer. When pricing reflects such value, the performance of TBYB is still superior but the gap with value-unaware strategies becomes smaller.

- When dealing with personal data, TBYB can significantly reduce the number of individuals whose information is disclosed to buyers, hence helping in preserving privacy.

Overall, our work demonstrates that near optimal dataset purchasing is realistic in practice. Not only does TBYB improve the profit for buyers, but acquiring less data to reach to similar accuracy also reduces data leakages, and hence contributes to preserving privacy. We believe that this is key for allowing both DMs and the data offer side to grow, as well. Interestingly, this approach could be implemented by real-world data marketplaces using the already existing "sandbox" functions to try data on the models of potential buyers.

# Chapter 7

# Splitting the value of combined data among multiple contributors

Data-driven decision-making is bringing significant improvements to many sectors of the economy, including in several applications related to ubiquitous computing in the areas of transportation, mobility, and crowd-sensing. A solid body of research has studied matters of route optimisation and city infrastructure planning [10, 46, 198], whereas companies are increasingly deploying and operating sophisticated systems for optimising their operations using live data. Such models and algorithms often require combining data from different sources and domains.

Data is by now considered a key production factor, comparable in importance to labour, capital, and infrastructure. Companies often need data from third parties, and for this they resort to the different types of data marketplaces that we discussed in chapter 3. PIMS like Digi.me, Swash, or Meeco allow individuals to sell their personal data, including their location, whereas general-purpose (AWS) and domain-specific marketplaces, some of them integrated in already-existing services (HERE, Carto, ESRI), allow companies to sell spatio-temporal data as real-time streams (Streamr, IOTA, GeoDB), as downloadable datasets, or accessible through APIs.

In almost all commercial marketplaces, pricing is left to sellers and buyers to agree. Sellers may set a fixed price, or let buyers bid for data [131], or even do a combination of the two. Such empirical pricing operates with minimal information, namely a high-level description of the dataset, including the number of data points it contains. The research community has already proposed different marketplace architectures to deal with AI/ML tasks, and industry-led initiatives aim to design trustworthy data spaces to share data [18, 72]. In a nutshell, such marketplaces are able to train ML models [3, 41] or run code [2, 156] from potential buyers on data provided by sellers. They also ensure that data is accessed according to the terms agreed by both parties, that no data is leaked or replicated, that the intellectual property of buyers is protected, and that transactions and data usage are tracked and accounted. Once a transaction is closed, the marketplace needs to distribute payoffs to sellers contributing to that transaction. In some settings, doing so according to the value of their data may incentivise sellers to provide better information.

This chapter looks at this open problem for the case of spatio-temporal data. In particular, we study how to compute the relative value of different spatio-temporal datasets used in i) forecasting future demand for a service across space and time, and ii) forecasting the travel time between two points A and B in a metropolitan area. Companies already offering service in overlapping areas can, for example, pool together their data to increase the accuracy of forecasting and its coverage. Improved forecasting can be used by the same companies to improve operations, such as dispatching vehicles, providing consumers with better information, or provisioning service points. It can also be sold to third parties often after bidding and bargaining to agree on the price. In the latter case, computing the relative value of each contributing source helps buyers select the most suitable data sources, and provides a fair way for marketplaces to split revenue among them.

For the purpose of our work, we concentrate on vehicle-for-hire demand prediction in Chicago and New York, and car travel time forecasting in Porto. While our examples and findings are specific to these particular urban mobility use cases, the methods that we apply for assigning value to spatio-temporal data are more general in scope, and can thus be used in other use cases beyond transportation, such as tourism, health services, entertainment, energy or telecommunications. We answer questions such as "Does combining multiple datasets of past taxi rides always benefit the forecasting accuracy of future services?". Also, when it does, "How should we attribute the improved forecasting precision to the individual datasets used to produce it?".

To do so, we use the Shapley value from collaborative game theory as a baseline metric for establishing the importance of each *player* (be they taxi companies or individual drivers) in the context of a *coalition* of data providers. The Shapley value is widely accepted for this purpose due to its salient properties we introduced in Sect. 5.3.1. But at the same time it entails serious computational challenges, since its direct calculation in a coalition of $N$ players requires enumerating and calculating the value of $O(2^N)$ sub-coalitions. This may be possible for a few tens of data providers, which is the case of companies in wholesale markets, but becomes impossible when considering hundreds or thousands of them in a retail data market setting.

Furthermore, we look at the trade-off between fairness and scalability/practicality by studying and comparing against simpler heuristics used to estimate the value of data, based on:

1. *Data volume*, in our case taxi rides, which has been used by marketplaces trading marketing or user profiling data [131], for example. While certainly more practical, this assumes that every ride has equal value for predicting demand or travel time.

2. *Leave-one-out (LOO)*, which has been used for "denoising" datasets, by omitting data points that reduce the accuracy of a ML algorithm [75]. Unlike the Shapley value, LOO examines only a single sub-coalition per source.

3. Measures of the amount of information held in data such as Shannon's *entropy* [136].

4. *Similarity metrics* that compare inputs to the aggregate dataset of the whole coalition, often used in detection of outliers [83].

## 7.1. Methodology

This chapter follows the notation and definitions previously introduced in Sects. 5.1 and 5.3. We particularise the definitions and the problem statement to our specific spatio-temporal forecasting tasks in Sect. 7.1.1, which include:

- a spatio-temporal demand forecasting model, and
- a travel time prediction model.

Section 7.1.2 presents the different scenarios we have evaluated, referring to the corresponding section of this chapter. Once we have introduced the target ML tasks of our study, we elaborate on how to compute the value of spatio-temporal data based on the Shapley value, and we discuss the Shapley approximation algorithms we tested in Sect. 7.1.3. Section 7.1.4 provides an intuition about the value of spatio-temporal time series for predicting demand, and points at the possibility of using heuristics tailored to approximate the value of data for these particular tasks. Finally, Sect. 7.1.5 presents some heuristics we will compare to the Shapley value in computing the relative value of spatio-temporal data.

### 7.1.1. Definitions and problem statement

Let $S$ denote a set of data sources, each one contributing a dataset $d(s_n), s_n \in S$. A dataset is a set of spatio-temporal observations $(x, t)$ denoting the spatial ($x$) and temporal ($t$) coordinates over a common period and geographical area describing the trajectory of taxi rides.

Such dataset is then split into training and test sets for experimentation purposes. We then train predictive algorithms ($\mathcal{M}$) on subsets of the complete training set (containing a part of the total number of data sources) and perform predictions on the test set. The accuracy of the trained model is gauged by similarity metrics that define the notion of *value* of a dataset $d(K)$, where $K \subseteq S$, which we denote as $v(d(K))$. Thus, $v(d(K))$ represents the accuracy of the predictive model, according to the chosen metric, when training is performed on the data from all sources $k \in K$, and prediction is performed on the fixed test set.

Our objective is to find a value assignment method, $\mu(s_i)$, that captures the relative importance of the data originating from source $s_i$ to the predictions of the following ML tasks:

1. Forecasting transportation demand in a metropolitan area based on historical observations.
2. Forecasting car travel time between two points of a metropolitan area based on historical spatio-temporal data of vehicles driving in the city.

As we described in Sect. 5.1, data buyers would provide the model, the accuracy metric and the test set in a real setting, whereas the marketplace would help with finding a combination of suitable data from their sellers to improve the model accuracy, and would reward sellers proportionally to the value of their data. Next we provide more details on how this general prediction

framework is adapted for the former two ML tasks, and particularly on the models, the valuation functions and the three large real datasets we have used, which we summarise in table 7.1.

Table 7.1: Datasets used in computing the relative value of data

| **Id** | ***Dataset1*** | | ***Dataset2*** | ***Dataset3*** |
|---|---|---|---|---|
| **City** | **Chicago** | | **New York** | **Porto** |
| **Time period** | 01-01-2013 - 09-01-2019 | 01-01-2019 - 09-01-2019 | 04-01-2019 - 05-31-2019 | 07-01-2013 - 07-01-2014 |
| **Rides** | 94 MM | 11.1 MM | 65 MM | 1.71 MM |
| **Companies** | 160 (101 individual licenses) | 58. 94% rides by top 15 companies | 33 | N/A |
| **Districts** | 77 (administrative communities) | | 261 | 100 (cluster regions after pre-processing) |
| **Taxi Ids** | 19,014. 55% of them from 5 companies | 6,469 | N/A | 448 |

#### 7.1.1.1. Spatio-temporal demand forecasting

For the purpose of this use case, we will focus on metropolitan vehicle-for-hire markets and we will assume that i) service demand observations will be taxi rides reported in a certain spatial coordinates at a certain time, and ii) data sources will be the databases of taxi companies that contain a log of such taxi rides. Our objective will be to forecast the aggregated demand in a control period ($T_c$) taking as an input the demand reported in an observation period ($T_o$).

Increasing the accuracy of such a prediction model is important both for operational needs (e.g., knowing where to dispatch drivers in anticipation of demand) and planning issues (e.g., deciding where to place taxi service points). Hence, it would make sense for companies to collaborate in case their prediction accuracy could be significantly improved by pooling their data.

*Datasets*: In order to compute results for real scenarios, we will make use of two datasets.

(*Dataset 1*) We use a public dataset of taxi rides from the city of Chicago,[1] which is a log of taxi rides that licensed companies report to local regulatory bodies. It contains more than 94 million rides from 160 companies, spanning from 2013 to 2019. For the analysis, we will filter data for the first half of 2019. We will consider the demand for the main 15 taxi companies in that city, plus an additional hypothetical 16th company that aggregates the information from the rest of companies and accounts for less than 5% of the total demand.

(*Dataset 2*) We validate the results against a dataset of taxi rides in New York City from April to May 2019.[2]. It includes more than 65 million rides from 33 companies in 261 districts spanning from April to May 2019. Unfortunately, it does not include data at individual taxi level.

*Model*: Figure 7.1 shows a block diagram that describes the general demand prediction model. For predicting demand in the control period $T_c$, we use a multi-seasonal SARIMA algorithm with hourly, daily and weekly sub-components, trained with data over $T_o$, and producing a forecast

[1] see data.cityofchicago.org, last accessed Feb'23.
[2] see nyc.gov, last accessed Feb'23.

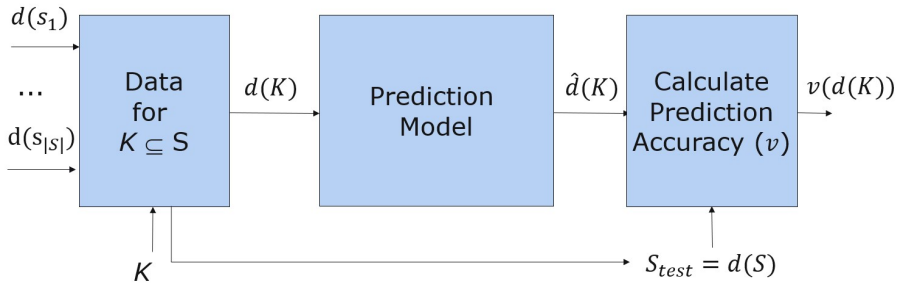



Figure 7.1: Demand prediction framework

($\hat{d}(K)$) over $T_c$ which is compared to the ground truth $d(S)$. The best parameters of the SARIMA model are obtained by grid analysis, and selecting the set of parameters that returns the best prediction using all historical data available.

*Value function*: The model can easily accommodate any similarity metric as the value function $v$. Throughout this chapter, we will use cosine similarity (hereinafter, CosSim) to compare transportation demand predictions to the test set $d(S)[t]$. *CosSim* assumes that the value of the data produced by the source subset $K$ can be defined as:

$$v_1(S_K) = \text{CosSim}(d(S)[t], \hat{d}(K)[t]) = \frac{\sum_{t \in T_c} d(S)[t] \cdot \hat{d}(K)[t]}{\sqrt{\sum_{t \in T_c} (d(S)[t])^2} \cdot \sqrt{\sum_{t \in T_c} (\hat{d}(K)[t])^2}}, \quad t \in T_c, \tag{7.1}$$

Moreover, we have validated the Shapley values obtained for CosSim against two other metrics: numerical similarity (hereinafter, NumSim) and relative dynamic time warp (hereinafter, RDTW).

The value of data of a subset $K$ of data sources using *Numsim* is defined as:

$$v_2(S_K) = \text{NumSim}(d(S)[t], \hat{d}(K)[t]) = 1 - \frac{1}{n} \cdot \sum_{t \in T_c} \frac{|\overline{d(S)}[t] - \overline{\hat{d}(K)}[t]|}{\overline{d(S)}[t] + \overline{\hat{d}(K)}[t]}, \quad t \in T_c \tag{7.2}$$

and works with normalised output ($\overline{d(K)}$) and is thus module-independent. The averaging performed does help in reducing inconsistencies introduced by the element-wise comparison. Finally, the value of data using *RDTW* is defined as

$$v_3(S_K) = \text{RDTW}(d(S)[t], \hat{d}(K)[t]) = 1 - \frac{\text{DTW}(\overline{d(S)}[t], \overline{\hat{d}(K)}[t])}{\text{DTW}(\overline{d(S)}[t], 0)} \quad t \in T_c, \tag{7.3}$$

RDTW is often used in signal processing and automatic speech recognition, where robust comparison of time signals is required.

#### 7.1.1.2. Travel time prediction

For the purpose of this task, we use periodic polls to GPS location of taxis in metropolitan areas and we assume that i) observations are trajectories reported between an origin location X and a destination Y at a certain time, ii) data sources are individual taxis sending a log of their rides to the company, and iii) data sources send their information frequently, so that we can easily reconstruct their route between X and Y, calculate their average and instant speed, distance travelled, etc. Our objective is forecasting car travel time between two points within a metropolitan area based on historical information.

*Dataset*: In this use case, we make use of a dataset[3] that provides the routes of 448 taxis within the city of Porto from Jul-13 to Jun-14 *Dataset 3*. Taxis mount mobile data terminals that are polled every 15 seconds and inform their position to the central office dispatching cars and drivers.

We run two experiments to predict travel time to Porto Airport and to São Bento train station as key points of interest (PoI) in the city. We filter taxi rides ending close to those two points, and we randomly split input data between train (80%) and test (20%) sets. We pre-process taxi rides in the training set by extracting the time it takes to reach the target PoIs from any intermediate points informed for each route, and by calculating intermediate timestamps, instant and average speeds, total travel distance, etc. We also divide the city in 100 clusters by applying k-means clustering algorithm to the set of possible locations, and add binary features informing the clusters traversed in each ride.

*Model*: To predict travel time between two points in a city, we use a Random Forest model [25] trained with the following input features: the UTM coordinates of the start and end points, the time of the year (month, weekday, hour, quarter), the type of day (i.e., workday, holidays or day before holidays) and the travel distance. The model is trained with the actual ride time in this training data set, and tested by making travel time predictions over the inputs of the test set, which we compare to the ground truth (the actual travel time of rides in the test set) to evaluate the performance of the model. We optimise the number of trees using grid analysis and we select the minimum number that yields an accuracy close (less than 1% difference) to the maximum reported in the analysis.

*Value function*: To compare the predictions of the model to the real travel time informed in the test set, we resort to $R^2$ score. Using all the information available in the training set, the models provide reasonable results (an average $R^2$ score above 0.8 in both PoIs).

### 7.1.2. Scenarios

We study the problem of computing the relative value of data of individual cars and whole company fleets in predicting transportation demand and travel time in different cities. Table 7.2

[3] see Kaggle ECML/PKDD 15: Taxi Trajectory Prediction, last accessed Feb'23

summarises those scenarios and use cases, the dataset used, the granularity level of the analysis, its scope, and the section we deal with it within this chapter.

Table 7.2: Scenarios for computing the relative value of data.

| # | Task | City | Dataset | Level | Scope | Sect. |
|---|---|---|---|---|---|---|
| 1.1 | Demand prediction | Chicago | *Dataset1* | Company | City-wide | 7.2.1 |
| 1.2 | | | | | District | 7.2.2 |
| 1.3 | | | | Taxi | City-wide | 7.3.1 |
| 1.4 | | | | | District | 7.3.2 |
| 2 | | New York City | *Dataset2* | Company | District | 7.4 |
| 3 | Travel time prediction | Porto | *Dataset3* | Taxi | PoI | 7.5 |

To calculate the relative values of data sources for these prediction task, we resort to the Shapley value as a baseline metric (see Sect. 5.3.1). However, computing the Shapley value for thousands of individual taxis is complex and requires efficient approximation algorithms. Before presenting the results of our exercise, we discuss about the methodology we used to approximate Shapley values and the three algorithms we tested in the following subsection. All of them resort to reducing the complexity by evaluating the model only for a subset of all possible combinations of data sources.

### 7.1.3. Computing the Shapley value for spatio-temporal forecasting

We test Monte Carlo (MC) [75], Random Sampling (RS) [35], and Structured Sampling (SS) algorithms to approximate Shapley values, and we compare the approximations to the exact values calculated using Eq. 5.2. When structuring samples of permutations in SS algorithms, we have tailored a solution to address problems where the position of *players* in a permutation strongly determines their marginal contribution [63, 191]. Whereas RS and MC randomly select the permutations to evaluate, SS ensures that all companies or individuals appear the same number of times in each position of the sample permutations. This we achieve by defining and processing the average marginal value brought by each data source for samples of $r$ blocks of Latin squares each of them containing $|S|$ permutations.

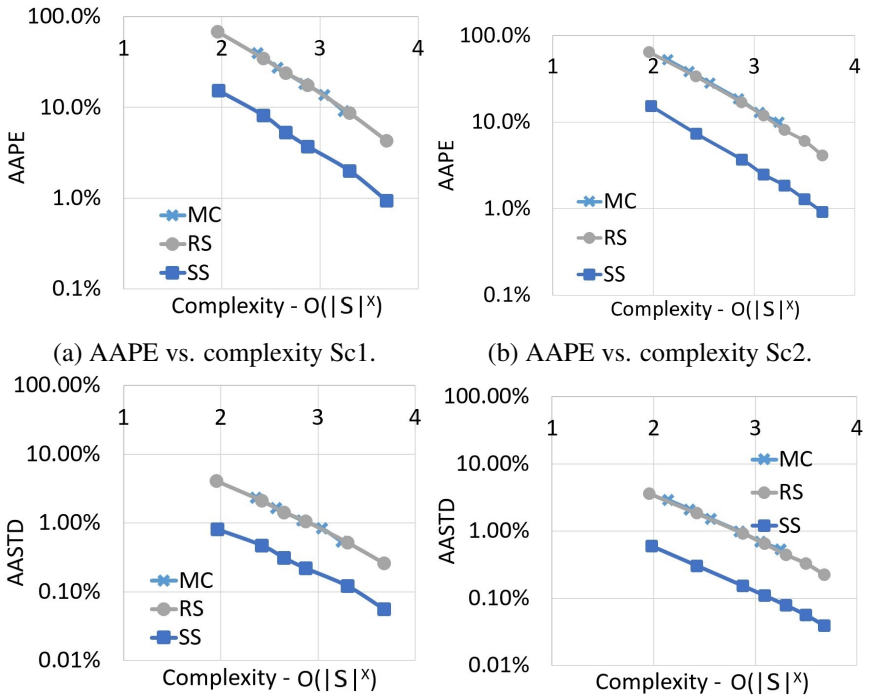
First, we carry out a test for approximating the value of companies in two different demand prediction scenarios (city-wide and medium-size district 35) of Chicago (*Scenario 1* or *Sc1*). Then, we repeat the test for approximating the value of individuals in 10 different tuples of 6 to 16 drivers yielding a good accuracy (close to 0.8) in predicting travel time to Porto's airport (*Scenario 2* or *Sc2*). Our cases then cover individuals and companies whose Shapley values are similar, and situations in which the value of companies and individual drivers differ significantly (up to more than $\times 5$). Given their stochastic nature, we test truncated and non-truncated versions of MC, RS and SS algorithms in rounds of 50 executions for each combination of input parameters and tuple of data sources.

We measure the performance of the algorithms in terms of:

- *Accuracy*, through the average average percentage error (henceforth, AAPE)[4] compared to the real Shapley values.

- *Robustness*, measured as the average average standard deviation (hereinafter, AASTD)[5] of their outputs in each round.

- *Complexity*, measured in terms of the actual number of training-prediction cycles computed in each case. A complexity of 2 means that the algorithm required $O(|S|^2)$ such cycles.




(a) AAPE vs. complexity Sc1.

(b) AAPE vs. complexity Sc2.

(c) AASTD vs. complexity Sc1.

(d) AASTD vs. complexity Sc2.

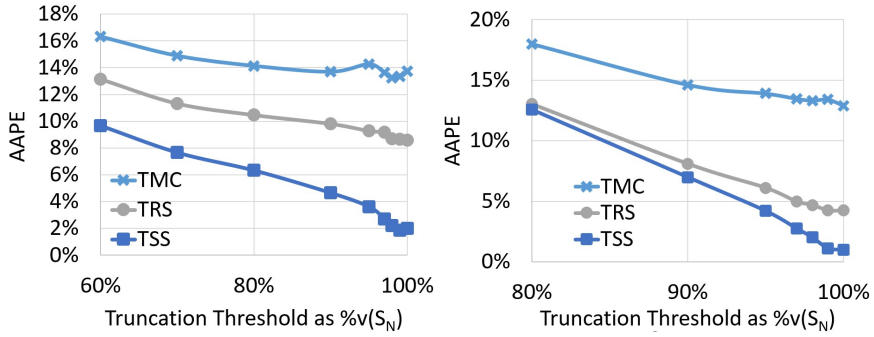
Figure 7.2: Accuracy and robustness vs. complexity of the algorithms

Figure 7.2 shows a comparison of MC, RS and SS in terms of accuracy and robustness for different levels of complexity (X-axis). We introduce complexity by reducing the maximum variation of Shapley values for the MC model to converge, and by increasing the number of permutations evaluated in RS and SS approximations. In general, the more combinations evaluated, the more accurate and, especially, the more robust the approximations are. As shown in the figure, SS clearly outperforms both RS and MC, and delivers more consistent outputs across executions, which are also much closer to the exact Shapley values. This is thanks to the planning of the sample permutations, which helps in reducing the randomness of the sample, and hence increases the accuracy of the approximation.

Figure 7.3 shows the effect of truncation on the accuracy for the three algorithms. According

[4]First average error across companies for each test, then average the average error across all executions

[5]First average standard deviation of the approximate Shapley value across all executions, then average the average standard deviation across companies




(a) AAPE vs. truncation Sc1. (b) AAPE vs. truncation Sc2.

Figure 7.3: Accuracy vs. truncation threshold

to our results, Truncated SS (TSS) is slightly more sensitive to truncation, but it always outperforms Truncated RS (TRS) and Truncated MC (TMC). Moreover, it is possible to easily control the trade-off between accuracy and execution time by tuning $r$ and the truncation threshold. We chose to use a *truncation threshold* of $0.95 \cdot v(S_N)$ since it divides the execution time by 4, while maintaining the AAPE below 5%.

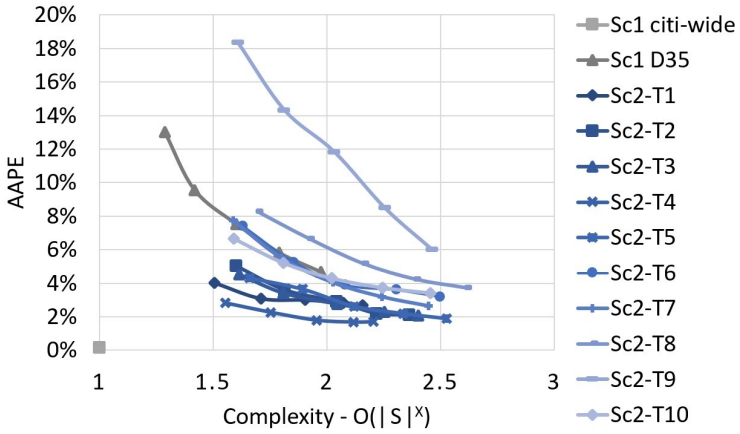



Figure 7.4: TSS AAPE vs complexity for r = 1, 2, 4, 8, 16.

To select $r$, we tested the TSS algorithm for increasingly complex ($r = 1, 2, 4, 8, 16$) approximations. Figure 7.4 shows the AAPE for the different test tuples of drivers and companies in both scenarios. Overall, we observe that for $r = 2$ and a complexity between $O(|N|)$ to $O(|N|^2)$ the AAPE is 5.2% on average and below 15% in all the cases, which we consider reasonable for distributing payoffs.

In conclusion, having evaluated the above approximation algorithms extensively in terms of precision and robustness vs. computational time, we have selected an *ad hoc* tweaked version of the TSS algorithm since it clearly outperformed the rest of them, and achieved the best trade-off on all the datasets we used for testing: it is able to approximate payoff distributions based on Shapley values with an error of less than 10%, which we consider sufficient for this purpose, in $O(|N|)$ to $O(|N|^2)$ computation time.

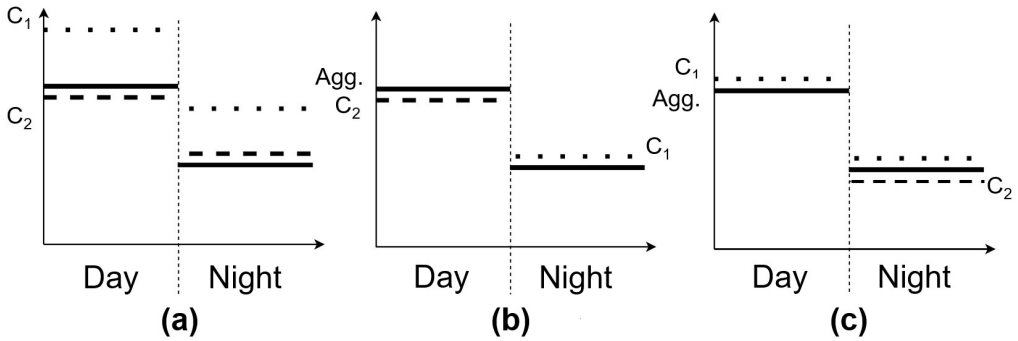



Figure 7.5: Data aggregation can influence its value in nontrivial ways.

Based on the lessons learnt throughout these tests, next we provide an intuition about how Shapley values work in demand prediction tasks. It will prove helpful in interpreting the results of real use case scenarios.

### 7.1.4. Some intuition regarding Shapley values in spatio-temporal forecasting

Consider a group of taxi companies agreeing to pool together their spatio-temporal data, containing demand for taxi rides within a city. One method to determine the value of a company is to observe how well the company is able to reconstruct the total aggregate, that is, the data coming from all companies, by solely using its own. As such, the data of one single company, or a group thereof, is used to train a predictive model, and the reconstruction error, between the prediction and an actual ground truth, is measured. This error, or rather its opposite, the reconstruction accuracy, represents the value of that company or group.

Aggregation leads to a highly non-trivial behaviour of the value function, which we illustrate with the toy example depicted in Fig. 7.5. A number of companies combine their data, to produce a spatio-temporal output or signal (continuous line), representing the total aggregate demand. For simplicity, the time scale is that of a single day, split into day-time and night-time, and also all signals are drawn as constant. Companies whose overall behaviour is closer to the *average* may be able to predict the complete aggregate signal by themselves, without a need to form coalitions with other companies. As such, their value will be ranked high by our algorithm. In Fig. 7.5(a), company $C_1$ is less valuable than $C_2$, as the signal of $C_2$ better emulates the total aggregate.

In the same setting, we also discuss the problem of complementarity, depicted in Fig. 7.5(b). Company $C_1$ is only offering its transport services during the night, while company $C_2$ is active solely during the day. Taken individually, the data of neither of these two is able to reconstruct the complete aggregate. However, they gain tremendous value as a coalition, since the resulting signal covers the entire time-span of the aggregate.

Data aggregation, however, does not always lead to an increase in value. A simple example is presented in Fig. 7.5(c), one company, $C_1$ provides data spanning the entire day (both day-time and night-time), and is also close to the total aggregate, while the other, $C_2$, only provides

data during the night. The predictive accuracy of both datasets combined is lower than that of $C_1$, because the absence of reports for day from $C_2$ will make most estimators believe that the demand gap between day and night is smaller than the real one.

It is thus clear that, depending on the particular characteristics of different datasets, mixing data may or may not be beneficial. Moreover, we have pointed to specific characteristics of data relevant in producing reliable demand forecasts, in this case the time span or, more generally, the time distribution of the observations. The following section presents a series of such simpler heuristics, some of them general in scope, while others are tailored to our specific problems.

### 7.1.5. Simpler heuristics for value estimation

Even though the Shapley value is widely acknowledged as a fair method to distribute payoffs among sources contributing data to a data transaction, it is very complex to calculate (see Sect. 5.3.1). Finding simpler heuristics to approximate the value of data as given by Shapley is desirable, since it would allow data marketplaces to select suitable data and reward data sellers much more efficiently. In this section, we present some candidate heuristics, which we will later on test against Shapley values to check whether they can be used to approximate the value of spatio-temporal data in ML prediction tasks.

One might initially think that the value of the data coming from a provider $s_i$ is given by its volume. In fact, data marketplaces often establish the price of their datasets proportionally to this figure of merit. Hence, we will also consider value distribution based on data volume, which results in the value assignment metric $\mu(s_i) = |d(s_i)|$, where $|d(s_i)|$ stands for the data volume of source $s_i$, or the number of data points originating from this particular source.

The *Leave One Out* (LOO) method, widely used in ML, considers that the value of a source $n_i$ is the difference in performance when the data corresponding to that particular source is removed from the training set. Hence, we define the LOO value of source $n_i$ as $LOO(s_i) = v(S_N) - v(S_{N-\{s_i\}})$. LOO can be computed in $O(|N|)$ time and has proven to be valuable for optimising the outcome of an algorithm by trimming data with negative $LOO$ values. In accordance with Eq. 5.1, the value assignment method in this case, is provided by $\mu(s_i) = LOO(s_i)$.

We also apply specific heuristics tailored to each use case. For predicting demand, we compute the similarity of an input $d(S_K)$ to the aggregate input $d(S)$ using the similarity metric $v$. This tests how close the shape of a each dataset is to the aggregate demand, and is often used in direct detection of time series outliers [83].

For predicting travel time, we apply Shannon's concept of entropy to the histogram of values a dataset provides for relevant spatio temporal features of our model. We define the entropy of a feature $j$ of dataset $S_K$ as:

$$H_j(S_K) = -\sum_{x \in \mathcal{X}_j} p(x) \cdot log(p(x)). \tag{7.4}$$

Entropy measures the amount of information or surprise in the values of feature $j$ of the dataset, and requires $\mathcal{X}_j$ to be a discrete set of values. In particular, we calculate three temporal entropy values by discretising timestamps to their month, weekday and hour, and a spatial entropy by clustering observations and grouping those that are close in space.

We will address the question of how appropriate these heuristics are for estimating the value of data in the next sections presenting the results of our exercise.

## 7.2. Computing the value of company data in predicting demand in Chicago

We start with the case that different taxi companies pool together their data to improve their demand forecasts, either for their own use, or to sell it to an external buyer. In both cases, it is relevant to know how important the contribution of each company is.

We first check the cases that make collaboration between companies meaningful. For those cases, we will then compute a fair measure of the importance of each individual company based on the quality of the data it offers. We will look at those two matters at both city level, as well as independently for each of the 77 different administrative areas (hereinafter, districts) in which Chicago is divided.

### 7.2.1. Demand forecasting at city level

Figure 7.6 shows a prediction sample for a control period between Apr 15th and Apr 28th 2019 based on the observations of the previous weeks. It compares the real observed demand to the predicted demand using information from *all* companies and only from the company labelled as *C0*. Similar plots are obtained for the rest of the companies.

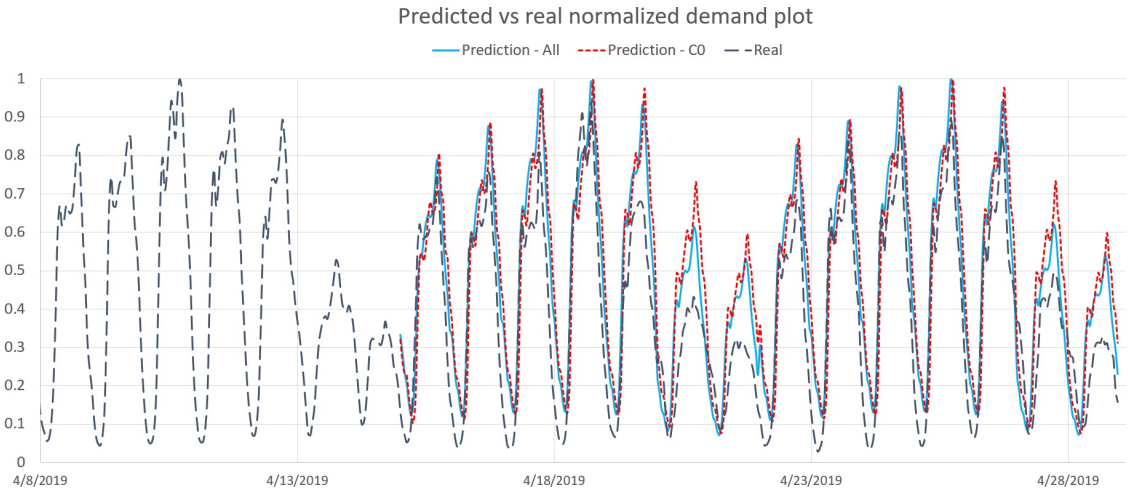



Figure 7.6: Example plot of real city-wide demand vs SARIMA model fit i) using the information from all companies and ii) using only that of company C0.

The demand prediction that each company is able to produce on its own yields, in general, an accuracy above 96% at city level as shown in Table 7.3. This means that all companies have enough data to independently predict the future demand with at most a 4% maximum error. Granted that all companies have sufficient data to perform demand prediction accurately on their own, the incentives for collaboration via pooling their data together are very small.

Table 7.3: City-wide accuracy by company.

| Co | Accuracy | Co | Accuracy |
|---|---|---|---|
| All | 0.9833 | C8 | 0.9797 |
| C0 | 0.9686 | C9 | 0.9861 |
| C1 | 0.9835 | C10 | 0.9829 |
| C2 | 0.9794 | C11 | 0.9659 |
| C3 | 0.9737 | C12 | 0.9845 |
| C4 | 0.9801 | C13 | 0.9725 |
| C5 | 0.9736 | C14 | 0.9767 |
| C6 | 0.9800 | C15 | 0.9724 |
| C7 | 0.9804 | | |

### 7.2.2. Demand forecasting at district level

We carried out a similar analysis by isolating the rides of each of the 77 districts of Chicago. Estimating the future demand in this case becomes more challenging and, as we will show soon, often requires collaboration between different companies.

Figure 7.7 shows the relationship between the forecast accuracy and the number of rides reported within a district. Not surprisingly, we see that the accuracy is higher in districts with a higher number of reported rides. Predictions at district level are more susceptible to irregular local events than city-wide predictions. For instance, despite being one of the districts with the highest number of reported rides, district number 7 (Lincoln Park), appears to be an outlier in Fig. 7.7. While analysing manually the dataset we found out that a large number of the reported rides were due to a one time event – a James Bay concert at the Riviera Theater, on March 19th. The resulting irregular spike that evening largely explains why the forecasting accuracy remains lower than other districts with smaller volume of demand but more regular patterns.

Another interesting case is district 33 (Near South Side), where the NFL Stadium, McCormick Place and different Museums and city attractions are located. Even though it is reporting a reasonably high number of rides (70k, ranked the fifth district in the city in terms of number of rides), the forecasting algorithm is unable to produce a prediction of high accuracy (goes up to 66% accuracy even with all the available information used). This is due to the event-driven nature of demand in this area, which is not captured by the assumed SARIMA algorithm.[6]

[6]Areas like this may be amenable to a better prediction accuracy by more complex forecasting algorithms using contextual information but this goes outside the scope of this dissertation since our focusing is on judging the importance of different datasets for a (reasonable) predictor as opposed to designing the best predictor possible.

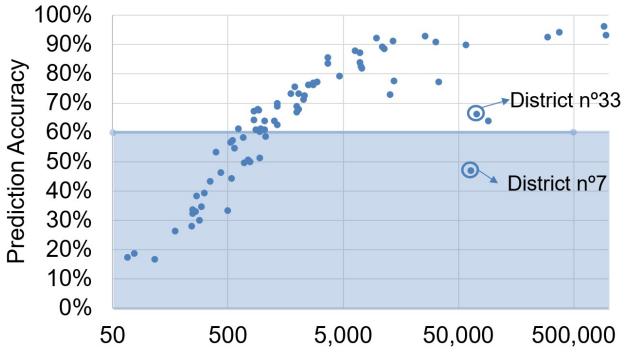



Figure 7.7: Prediction accuracy at district level.

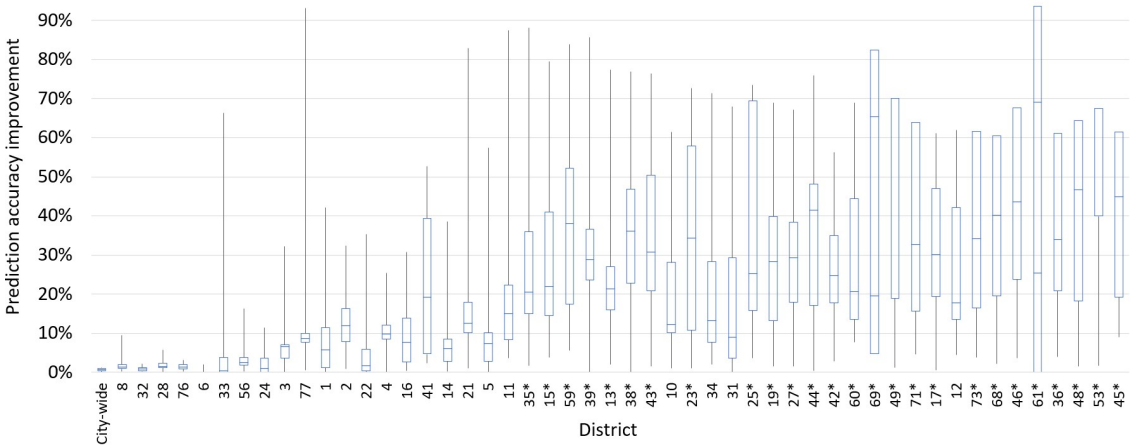



Figure 7.8: Potential prediction accuracy improvement by cooperation at district level.

Out of the 77 districts, the forecasting algorithm is able to achieve an accuracy above 60% for 50 of them (those above the shaded region in Fig. 7.7). This means that even by aggregating all the information available, the particular forecasting algorithm would not be able to predict the future demand with sufficient accuracy for 27 districts.

In order to check whether our findings at city level still hold at district level, we execute the forecasting algorithm in each of these 50 districts for all 16 companies. Figure 7.8 depicts box-plots (over companies) of the forecasting accuracy improvement from collaboration (Y-axis) in each district (X-axis). Districts are sorted in descending order with respect to the total number of reported rides. We also include city-wide results at the leftmost point of the plot. We find that there is always at least one company that is able to build a forecast model on its own which is very close to the one built by using all the data available. It is not necessarily always the same company across all districts, neither always the biggest one. Smaller companies tend to benefit more from cooperation, which is also less likely to be beneficial in the most popular districts, meaning those that report a large number of rides. However, as we move to smaller districts, the benefits of collaboration start increasing. It is at such areas where it makes sense for different taxi companies to pool their data together to achieve a more accurate forecast.

Table 7.4: Shapley value, LOO and nº rides (Rd%) for three districts.

| | 15 | | | 17 | | | 19 | | |
|---|---|---|---|---|---|---|---|---|---|
| Co | SV | LOO | Rd(%) | SV | LOO | Rd(%) | SV | LOO | Rd(%) |
| 1 | 11.2 | 0.5 | 2.5 | 14.0 | 0.6 | 8.3 | 2.0 | 0.0 | 3.4 |
| 2 | 1.8 | -0.1 | 0.8 | 0.0 | -0.1 | 0.5 | 1.5 | 0.0 | 0.5 |
| 3 | 1.0 | 0.0 | 0.3 | 0.2 | 0.0 | 0.5 | 0.3 | 0.0 | 0.0 |
| 4 | 0.4 | -0.1 | 0.2 | 0.2 | 0.0 | 0.0 | 0.4 | 0.0 | 0.1 |
| 5 | 2.3 | -0.1 | 0.9 | 0.4 | 0.0 | 0.5 | 0.7 | 0.0 | 0.8 |
| 6 | 16.4 | -1.2 | 37.9 | 28.0 | 8.7 | 56.2 | 24.1 | 3.3 | 38.6 |
| 7 | 1.1 | -0.3 | 0.4 | 0.2 | 0.0 | 0.4 | 0.2 | 0.2 | 0.5 |
| 8 | 1.1 | -0.1 | 0.8 | 0.3 | 0.4 | 1.4 | 1.5 | 0.2 | 0.5 |
| 9 | -0.3 | 0.0 | 0.2 | 0.0 | 0.0 | 0.2 | -0.6 | -0.1 | 0.3 |
| 10 | 2.3 | 0.4 | 1.4 | 0.2 | -0.2 | 0.8 | 0.9 | 0.1 | 0.7 |
| 11 | 0.6 | 0.0 | 0.3 | 0.2 | 0.1 | 0.5 | 1.4 | 0.0 | 0.5 |
| 12 | 4.4 | -0.1 | 1.9 | 0.4 | 0.1 | 0.9 | 2.4 | 0.1 | 1.9 |
| 13 | 17.9 | 0.8 | 18.1 | 0.3 | -0.2 | 1.3 | 4.3 | 0.0 | 1.3 |
| 14 | 16.7 | -0.9 | 34.0 | 17.2 | 0.0 | 27.6 | 26.4 | 1.9 | 50.4 |
| 15 | 0.4 | 0.0 | 0.1 | 0.8 | 0.0 | 0.3 | 0.4 | 0.0 | 0.1 |
| 16 | 0.2 | -0.1 | 0.2 | 0.0 | 0.0 | 0.8 | 2.4 | 0.1 | 0.5 |

Focusing on the districts where collaboration makes most sense, we will now compute the relative importance of the data that each company brings, via the notion of the Shapley value.

### 7.2.3. Computing the value of information at district level

For the 26 districts marked with an asterisk in Fig. 7.8, taxi companies would benefit from an increase in forecasting accuracy by combining their data. For each one of them we have computed the Shapley value of the 16 companies using the Shapley formula from Eq. 5.2. To do that we used cosine similarity as the value function, a test dataset obtained by combining the taxi ride data of all companies active in each district, and the output of the SARIMA model, once trained on the taxi ride data from a particular coalition, as the prediction.

Table 7.4 summarises the Shapley value, the LOO value and the percentage of rides reported by each company in the first 3 districts. Figure 7.9 shows the relationship between the number of rides and the Shapley value for our forecast at district level, which represents the average marginal contribution of its data to the obtained forecast accuracy for that district. Each point in the plot represents a company in one of the 26 districts.

Observing Table 7.4 one may see that different taxi companies can have Shapley values that differ by several orders of magnitude within the same district. Also, the Shapley value of a given company may vary from district to district by a factor of more than $\times 10$ in some cases (see, for instance, companies 1 and 13). Some companies have negative Shapley values in certain districts, meaning they bring *on average* a negative contribution (i.e., they reduce the forecast accuracy) to the coalitions they join.

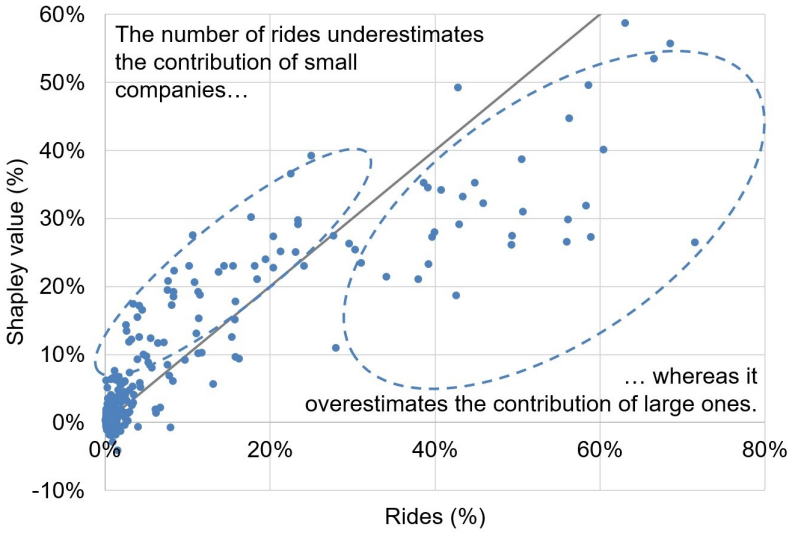



Figure 7.9: Shapley value vs. nº rides reported by companies for a sample of districts. Each point in the plot represents a company in a district.

From Fig. 7.9 we see that the Shapley values of companies do not correlate well with their number of rides. In fact, the Shapley value for small companies tends to be higher than their corresponding percentage of rides, whereas it is the opposite for large companies. In other words, if we approximated the importance of companies just by the volume of data (rides) they contribute, we would be rewarding large companies, at the expense of smaller ones.

Similarly, LOO values are weakly correlated ($R^2 = 0.38$) with Shapley values at the level of districts. Notice that even some companies reporting many rides have negative LOO (e.g., company 6 in district 15). As with the case of number of rides, were someone to split an accepted bid based on LOO values, the allocation of payments would deviate significantly from what a splitting based on Shapley would produce.

### 7.2.4. Validating the results using other metrics

We have compared the Shapley values using the different metrics introduced in Sect. 7.1.1.1 in a sample of districts with high (8, 28), medium (6 and 56) and small (11) demand by using data from $|N| = 16$ companies. We show the results for district 11 in table 7.5, compared to those calculated by using CosSim as the value function. We deliberately chose district 11 since it shows the highest dispersion of Shapley values. As it can be seen, the values using different metrics are highly correlated, and the top 4 companies are the same in the three cases. In fact, it turns out that the fraction ($\frac{\phi_i}{\sum_{j \in N} \phi_j}$) is very similar for the three value functions ($R^2 = 0.92$ in the case of CosSim vs. NumSim, $R^2 = 0.87$ in the case of CosSim vs. RDTW, for the 5 districts), meaning that they would produce similar distribution of payoffs among data sources.

Table 7.5: Shapley value of $|S| = 16$ companies for different value functions $v$ in district 11

| Co | CosSim | NumSim | RDTW | Co | CosSim | NumSim | RDTW |
|---|---|---|---|---|---|---|---|
| 1 | 0.02 | 0.01 | 0.02 | 10 | 0.11 | 0.08 | 0.06 |
| 2 | 0.03 | 0.03 | 0.02 | 11 | 0.03 | 0.02 | 0.02 |
| 3 | 0.02 | 0.02 | 0.00 | 12 | 0.10 | 0.07 | 0.08 |
| 4 | 0.01 | 0.01 | 0.01 | 13 | 0.06 | 0.04 | 0.02 |
| 5 | 0.04 | 0.03 | 0.01 | 14 | 0.12 | 0.11 | 0.10 |
| 6 | 0.12 | 0.11 | 0.09 | 15 | 0.00 | 0.00 | 0.00 |
| 7 | 0.08 | 0.07 | 0.06 | 16 | 0.06 | 0.04 | 0.03 |
| 8 | 0.03 | 0.03 | 0.03 | All | 0.86 | 0.69 | 0.56 |
| 9 | 0.01 | 0.01 | 0.01 | | | | |

### 7.2.5. Summary

Predicting demand at city level does not require collaboration between taxi companies since each one can independently estimate city-wide demand. However, when attempting to estimate demand at district level, different companies need to combine their data if they are to achieve a high prediction accuracy. In these cases, neither the data volume a company is providing nor its LOO value reflect accurately its contribution to achieving a better forecast of future demand as given by its corresponding Shapley value. Finally, these observations hold for the three value functions tested, which do result in similar Shapley values and payoff distributions across a number of representative districts.

## 7.3. Computing the value of individuals' data in predicting demand in Chicago

In the previous section we applied methods for estimating the value of aggregate data held by taxi companies. In this section we will go a step further, and apply approximations to Shapley value for estimating the value of data held by individual drivers. Recently, some PIMS have appeared that allow for "retail" data marketplaces in which individuals are able to sell their data (see the discussion about PIMS in chapter 3). They will face additional challenges in terms of scalability to compute the Shapley value over hundreds or thousands of taxi drivers, and distribute payoffs for their data to individual drivers according to their value.

The idea of providing *micropayments* to users for their personal data has received a lot of public attention [114, 155]. Finding fair ways to compute value-based contributions to a ML problem is key in calculating the contribution of individuals to the data economy, and will also require to find fair scalable ways to do it. More recent work describes fundamental technological challenges that need to be addressed for the above vision to be fulfilled [115].

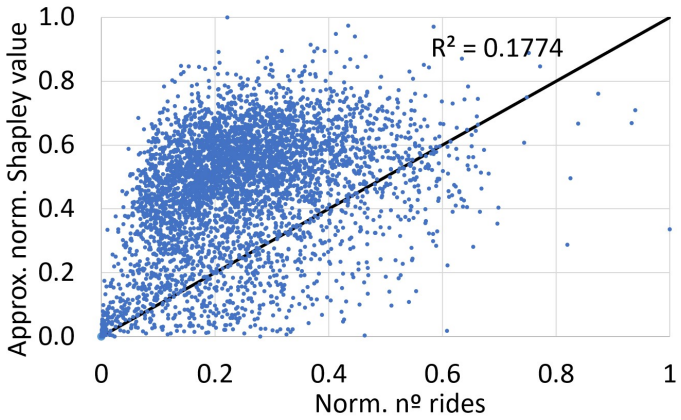



Figure 7.10: Approx. Shapley value vs. nº of rides across drivers.

### 7.3.1. City-wide results

We have computed a TSS Shapley value approximation for a set of $|N| = 4968$ taxi drivers that provided service in Chicago during March and April 2019. In this way we computed the contribution of each driver's data to the forecasting accuracy achieved by the model in predicting the demand in the second half of April using taxi rides from the previous six weeks for training.

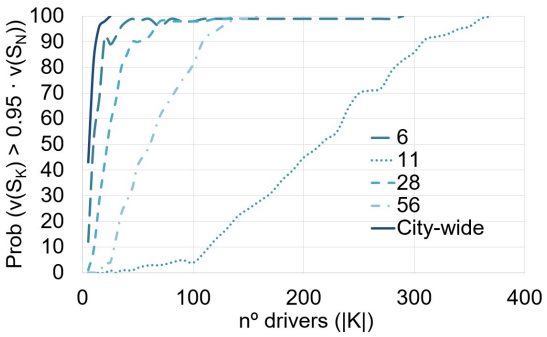
In the same way that we proceeded in the wholesale use case, we compared the Shapley value with the number of rides reported by each driver. Figure 7.10 shows a plot of these two metrics across all drivers. We see that there is no clear relationship between them ($R^2 = 0.1774$), and there seems to be other factors affecting Shapley value. For example, the value of a tuple S of drivers ($v(S)$) seems to be more correlated to the similarity of the input signal to the average ($R^2 = 0.6736$) than to the number of rides that drivers in S are reporting, as suggested by our toy model in Sect. 7.5. Another interesting finding is that it takes a very small number of drivers to estimate the city-wide aggregate demand. With 7 randomly selected drivers, on average, we can reconstruct the shape of the demand at city level with a 95% accuracy.

### 7.3.2. Results at district level

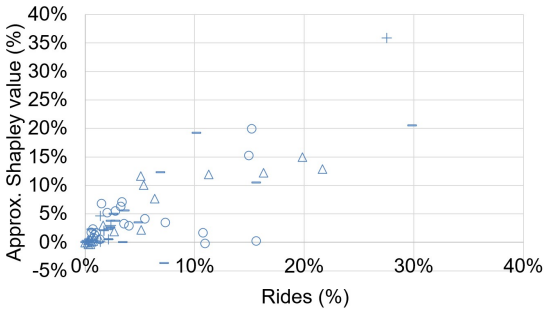
As we just saw, it is possible to build accurate demand forecasts at city level using only a very small number of drivers. We will check whether this also holds for demand forecasts at the district level.

For that purpose, we will first quantify the number of necessary drivers, and then proceed to compute the relative value of each driver's data. Figure 7.11a shows the probability that using a number of drivers indicated in the X-axis one can achieve a prediction accuracy at least 95% of that achieved when using information from all the drivers. Different lines correspond to districts with high (28), medium (6 and 56) and small (11) demand for taxi rides.

The plot shows that whereas for forecasting city-wide demand, or demand in large districts, few drivers suffice, forecasting the demand of medium-sized and smaller districts requires infor-




(a) Prob $(v(S_K) > 95\% \cdot v(S_N)$ vs nº drivers in $K$.




(b) Approx. SV vs. nº of rides for tuples K of $|K| = 20$ drivers with $v(S_K) > 95\% \cdot v(S_N)$.

Figure 7.11: Results at district level for individual taxi drivers.

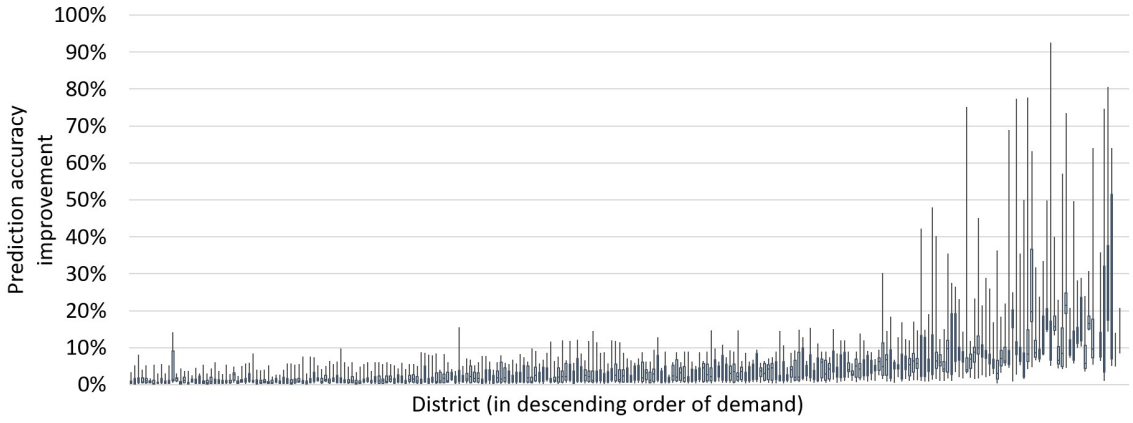
mation from many more drivers. This can be understood by noting that in large districts, the aggregate demand is much more predictable since it is the result of the aggregation of large numbers of independent variables (people that may need a taxi ride). Such demands are known to be easier to forecast, but achieving high forecasting accuracy in medium and small districts requires using the data from tens if not hundreds of drivers. Computing the actual Shapley value is impractical for such numbers of players, but it can be approximated by using the truncated structured sampling approximation discussed earlier in Sect. 7.1.3.

We computed the Shapley values for smaller sets of drivers whose data achieve an accuracy very close to $v(S_N)$ when combined. Figure 7.11b shows a scatter plot of the approximate Shapley value (Y-axis) vs. the percentage of reported rides (X-axis) for a number of such sets of drivers in district 28. Each point represents a driver, and drivers from the same set are represented with the same marker. As observed earlier at city level, the real value of a driver may be very different from that predicted by its number of rides.

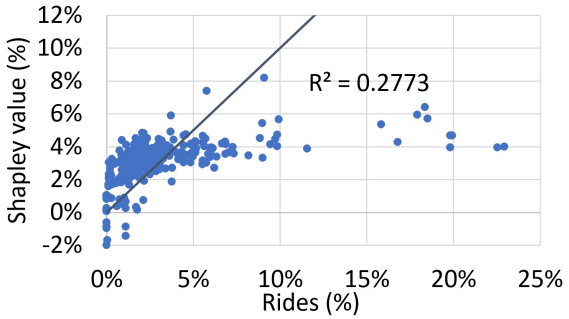
## 7.4. Computing the value of data in predicting demand in NYC

We have validated the results by repeating the analysis using a large dataset of taxi rides in NYC. The conclusions from NYC are similar to the ones we drew in detail for Chicago.

Figure 7.12a shows a box-plot (over companies) of the prediction accuracy improvement (Y-axis) by district (X-axis) in the case of NYC. More than 75% of taxi companies are able to predict demand with an accuracy of above 0.9 in 219 districts. Cooperation can improve by at least 10% the accuracy of individual forecasting for more than 50% of the companies in 20 of the smallest districts. There are 4 districts with very few rides in which the forecasting algorithm cannot achieve a high accuracy even combining the data from all companies. In the districts where cooperation between companies made sense, the number of rides reported by each company is again weakly correlated with its Shapley values ($R^2$ ranging from 17% to 40%, 27% on average, see Fig. 7.12b), and the same holds for the LOO value. In summary, repeating the analysis for a second large dataset verified the conclusions of our analysis based on the Chicago dataset.




(a) Box-plot (over companies) of potential prediction accuracy.




(b) Shapley value vs. nº rides in small districts.

Figure 7.12: Area-level results by company for demand prediction in NYC.

## 7.5. Computing the value of data in predicting travel time in Porto

As a third and completely different use case, we measured the relative value of the information provided by ubiquitous sensors placed in cars for predicting the travel time between two points of a city, using trajectory data from taxis in Porto. A first analysis reveals substantial differences in the amount of data provided by each individual car.

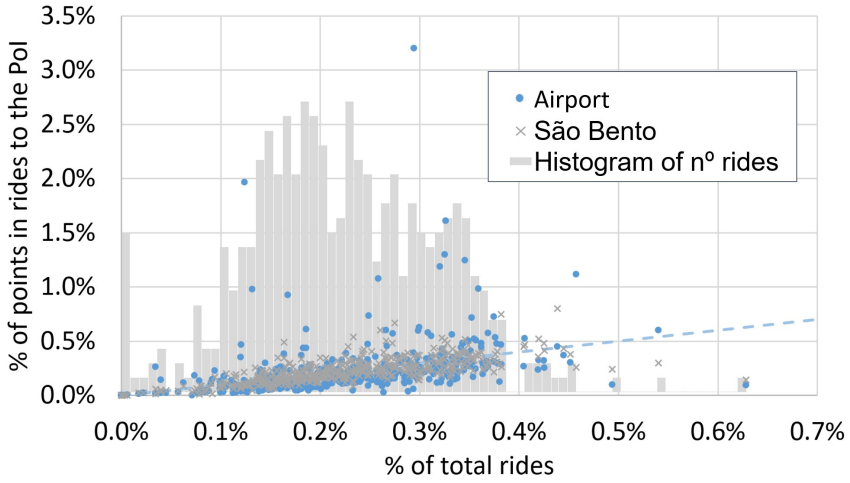
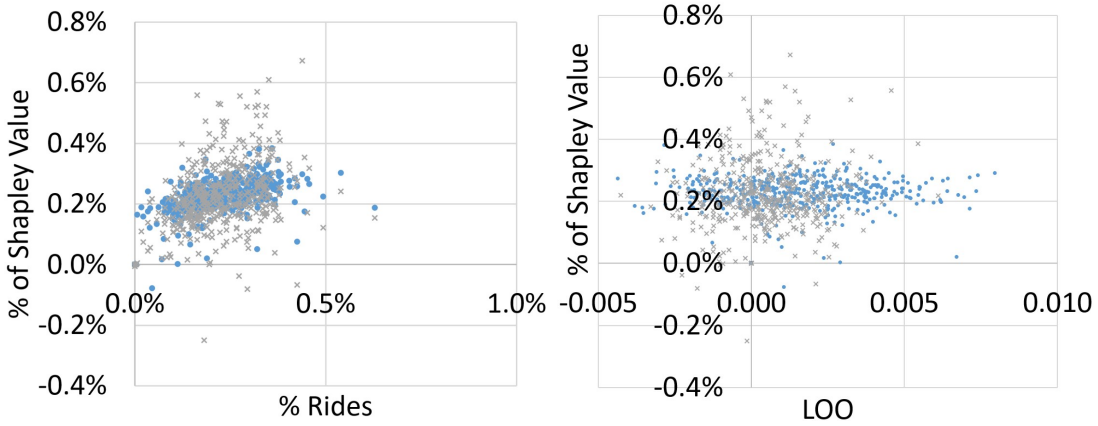



Figure 7.13: Total vs nº rides to Porto's airport and to São Bento St.

Figure 7.13 shows a histogram of the number of rides reported by each taxi (in grey bars), and a scatter plot comparing the percentage they represent in the whole dataset (X-axis), and in the subset of rides ending in the corresponding point of interest (Y-Axis). Not only do vehicles provide different volume of information, but they also exhibit different behaviour. For example, the one reporting the largest number of rides, hence more to the right, turns out to be reporting only a few rides to the airport and the train station. On the contrary, some drivers that appear well above the dotted line travel much more frequently to the airport than the average.

Even though PIMS usually reward data of individuals equitatively, our experiments show that each taxi brings very different value to the model. Whereas the Shapley value of some vehicles is close to twice the average, others account for a negligible or even negative contribution. Some of them are valuable for predicting the travel time to the airport, but not so valuable for predicting the travel time to the São Bento train station.

It is also quite common to set the price of data based on its volume (see Sect. 3.2.4.3), assuming that the more points reported, the more valuable a piece of data would be to the model. However, as the correlation plot in Fig. 7.14a shows, this is not exactly the case. The Pearson correlation (R) of Shapley values for predicting travel time with the number of rides reported is 0.56 and 0.39 respectively. Such results evidence that the more rides a driver reports the more likely it is that its data will improve the accuracy of the model. However, two individuals reporting the same number of rides may bring a very different value to the model, which supports the hypothesis that other complex dependencies and factors must be considered in this calculation.




(a) $R_1 = 0.56$, $R_2 = 0.39$ (b) LOO vs Shapley values.

Figure 7.14: Pearson correlation of Shapley values with the volume of data and LOO values.

As Fig. 7.14b shows, LOO values are small ($|LOO| < 0.01$) and uncorrelated with the Shapley value or with the number of rides. As a result, any reward attribution or ranking of data sources based on LOO would be close to random.

We also considered the spatio-temporal *diversity* of data and its relationship to data value. We want to check whether wider time span and more dispersion across space lead to higher value.

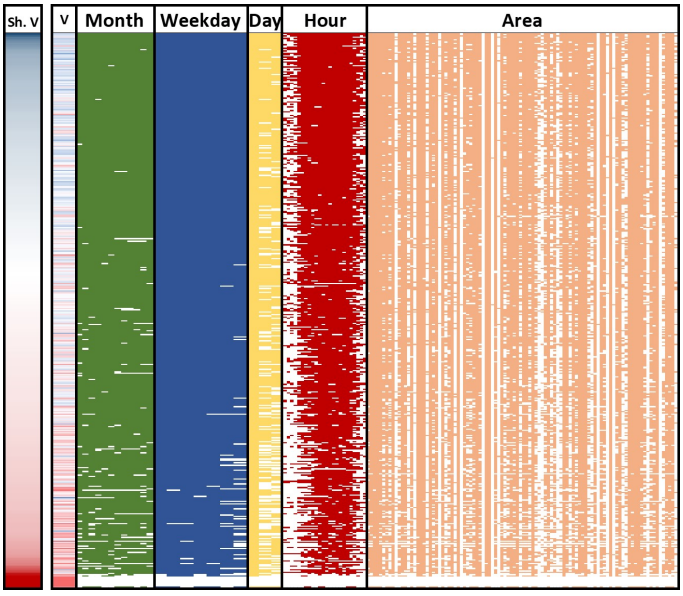
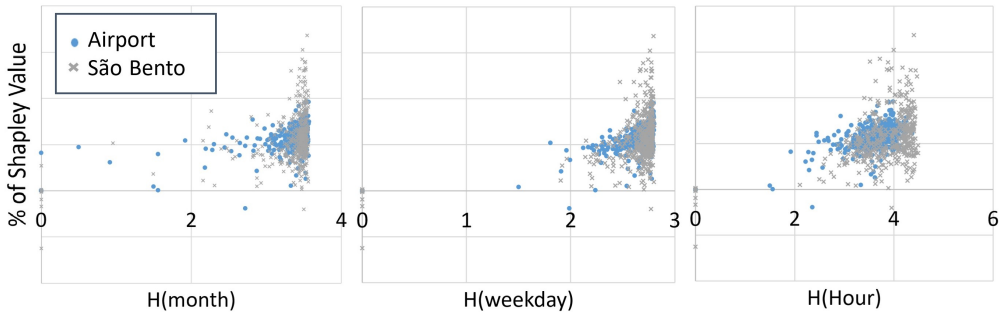



Figure 7.15: Shapley value vs representativeness of data in predicting time to the airport.

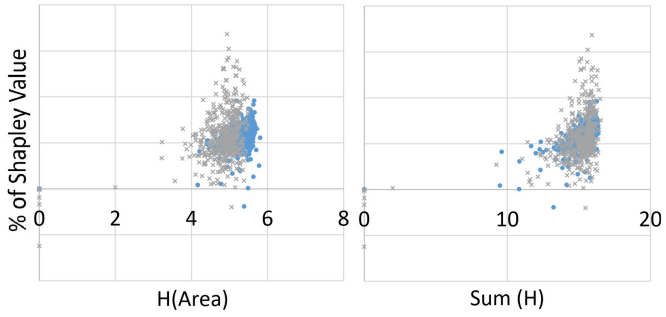
Figure 7.15 shows how representative data from different sources is for the different features considered in the model. Rows represent individual taxis sorted in descending order of their Shapley values, shown in the first column. Blue indicates a Shapley value above the average, whereas red cars as we move to the bottom of the table showed to less useful for improving the accuracy of the model. The columns on the right refer to the different family of features used by

the model. The column labelled as '$V$' refers to the volume of rides reported. The darker blue the more rides than the average a taxi reports. On the contrary, red colour indicates that only a few rides were reported. As the figure shows, there is apparently no significant correlation between the Shapley value and the volume of rides.

The rest of the columns on the right are coloured depending on whether each taxi reports rides for a specific month, weekday, type of day, hour or area in the city (out of 100 different areas obtained by applying k-means clustering to all the spatial data points). The analysis shows that taxis close to the top provide more comprehensive data that covers almost all the features that the model is using. As we move down in Fig.7.15, more blank cells start to appear, meaning that less valuable riders fail to provide data for specific time frames, days, or areas of the city. Drivers that only report rides at certain hours or in specific areas, or those that stopped their activity for several months apparently mislead the prediction algorithm and, hence, are less valuable than those that provide more *diverse* data.




(a) $R_1 = 0.63$, $R_2 = 0.48$ (b) $R_1 = 0.67$, $R_2 = 0.47$ (c) $R_1 = 0.69$, $R_2 = 0.57$




(d) $R_1 = 0.62$, $R_2 = 0.45$ (e) $R_1 = 0.69$, $R_2 = 0.60$

Figure 7.16: Pearson correlation of Shapley values with entropy features related to the diversity of data for predicting travel time to Porto's airport ($R_1$) and São Bento Station ($R_2$).

We have measured this *diversity* through the concept of information entropy. It turns out that the entropy rates of spatial and temporal features show significant correlation with Shapley values (R ranging from 0.45 to 0.69), as Figs. 7.16(a-e) show. Therefore the higher the entropy of information of an individual taxi, the more *diverse* in terms of time and space, and hence the more valuable for the prediction algorithm it will be. On the contrary, taxis failing to provide data

for specific days of the week, those that only report rides at certain hours or in specific areas, or those that stopped their activity for several months apparently mislead the prediction algorithm and, hence, are less valuable. This result points to another potential approach for approximating the value of data without having to compute the Shapley value, but instead relying on the above mentioned diversity feature.

## 7.6. Key Takeaways

In this work we have looked at the problem of how to compute the relative importance of different spatio-temporal datasets that are combined in order to improve the accuracy of demand and travel time forecasting for taxi rides in large cities such as Chicago, Porto and New York.

We first studied the value of data fusion at the granularity of companies. Since the number of such companies covering the same geographical area is typically small, the relative value of their data can be computed directly from the definition of the Shapley value. This, however, becomes infeasible at the level of individual taxi drivers, since the latter may amount to several thousands for large metropolitan areas. To address this issue, we compare different approximation techniques, and conclude that an *ad hoc* version of structured sampling [63] performs much better than other more popular approaches such as Monte Carlo [75] and random sampling.

By applying our model and valuation algorithms to taxi-ride data from Chicago, Porto and New York, we find that sufficiently large companies hold enough information to independently predict the overall demand, at city level, or in large districts, with over 96% accuracy. This effectively means that inter-company collaboration does not make much sense in such cases. On the contrary, companies have to combine their data in order to achieve a sufficient forecasting accuracy in smaller districts. We compute the relative value of different contributions in such cases by computing the Shapley value for each taxi company. We find that the values differ by several orders of magnitude, and that the importance of the data of a given company can vary as much as $\times 10$ across districts. More interestingly, the Shapley value of a company's dataset does not correlate with its volume, i.e., some companies that report relatively few rides have a larger impact on the forecasting accuracy than companies that report many more rides. The LOO heuristic also fails to approximate the per company value as given by Shapley.

Similar phenomena are observed at the finer level of individual drivers. We show that by combining data from relatively few drivers one can easily detect peak hours at city level. At district level, however, more data needs to be combined, and this requires making use of our fastest approximations of the Shapley value based on structured sampling. Moreover, using trajectory data from taxis in Porto, we observe again, this time for estimating the travel time within a city, that the value of information contributed by each driver may vary wildly, and that it cannot be approximated based on the volume of rides they report nor via the LOO heuristic.

*Overall, using multiple datasets, different forecasting objectives, and at different granularities, our work shows that computing, even approximately, the Shapley value seems to be a "nec-*

*essary evil" if one wants to split fairly the value of a combined spatio-temporal dataset. Simple heuristics based on volume and LOO fail to approximate the results produced via the Shapley value. Other heuristics tailored to each problem, such as the similarity to the aggregate when predicting demand, or spatio-temporal Shannon entropy when predicting travel time in a city, seem to be doing a better job at approximating Shapley.* We consider this to be the most important learning stemming from our work which has looked in detail at multiple real-world spatio-temporal datasets and problems, across different granularities, and under different objective functions. We believe that the fast-growing ecosystem of data marketplaces and PIMS surveyed in chapter 3 can greatly benefit from this finding as it transits from very basic towards more elaborate and *fairer* pricing schemes.

Our main result has been that the importance of each dataset differs and cannot be deduced via simple heuristics based on data volume or leave-one-out methods, but instead one needs to look deeper and consider the complex ways in which different datasets complement one another, which is what the Shapley value does. This applies when combining data from entire companies, but as well for data from individual drivers.

However, the Shapley value has inherent scalability issues and, even using efficient approximation algorithms, requires a $O(N^2)$ computation time which is still insufficient for thousands of taxis or complex prediction models. We have also noticed the correlation between data volume and its Shannon information entropy with the value of spatio-temporal data. Were Shapley values based on functions combining volume and entropy able to approximate those based on repetitively training a ML model, they would arguably be much more efficient in terms of computing time.

# Part IV

# Conclusion

# Chapter 8

# Discussion

In this chapter, we discuss some technical challenges of a future human-centric data economy, and we point to potential solutions and future work needed to address them. First, we briefly summarise the challenges of data trading identified throughout this dissertation. Next we present some future research directions. Section 8.2 discusses how our measurement study could be extended to gain more insights into the data economy, to provide transparency and to manage this thriving ecosystem more effectively. Then Sect. 8.3 points at building a data quotation tool leveraging our measurement study, and aimed to serve as guidance about market prices of data. Finally, we discuss about other open issues, valuable technology and future works relating to federating human-centric data marketplaces in Sect. 8.4.

## 8.1. Open Challenges of data trading

Despite its remarkable potential and observed initial growth (50% yearly increase of the number of products offered in AWS and DataRade.ai in 2021), the market for data is still at its nascent phase. Like all nascent economies, the data economy faces a yet uncertain future. Regardless of which companies and business models finally succeed, we identified some key, intertwined, open challenges related to increasing the *practicality* and *trust* of the ecosystem:
(*Challenge 1*) The *current fragmentation of data markets*, as reflected by the ever-growing large number of companies in market surveys and the variety of data being traded (see Sect. 3.2.4.1), makes us think that a consolidation could take place in the future and a new single monopoly or 'niche' data trading champions may arise. Instead of waiting for such champions to solve it for everyone, and given the importance and complexity of the task, solutions must be sought that respect transparent Internet and web governance and expansion principles, including openness, standardisation, and layering [116]. We point to this challenge as future works later in Sect. 8.4.
(*Challenge 2*) Regarding data economics, there are open problems and unexplored questions related to pricing, as we pointed out in Sects. 2.4, 3.2.4.3, and 3.2.4.4. Moreover, a healthy data market requires knowledgeable neutral references (like web services suggesting a range of prices

to second-hand car sellers based on the model, year, nº km, etc.) to avoid ending up in a radical and sustained price fluctuation of data products. To help with this, we have contributed a measurement study about prices of data in commercial data marketplaces in chapter 4.
(*Challenge 3*) Due to the fact that 'data' is an experience good, it is far from obvious for potential buyers to anticipate the value of data in certain settings such as ML tasks [75]. Hence, developing new solutions like the ones implemented in chapter 6 allowing buyers to first locate [26] and then select better data, which improve the - still primitive - ones presented in Sect. 3.2.4.9, is important.
(*Challenge 4*) Dealing with *ownership* and fighting against piracy and theft of data is of uttermost importance to ensure trustworthiness in data trading. This task is even more arduous when malicious players are able to copy and transmit data at zero cost, and the market lacks a sound notion of authorship, as we pointed out in Sect. 3.2.4.10.
(*Challenge 5*) Related to *data provenance*, computing *fair compensations* for providers at scale is an additional challenge for DMs, and especially for PIMSs dealing with individuals (see Sect. 2.4). Such compensations must ideally be based on the value they bring to a specific task or buyer, and DMs must be accountable and transparent about the process [64].

Finally, significant regulatory challenges related to data trading lie ahead, both for competition authorities and *ex ante* regulatory bodies. Due to their market power, tech companies are increasingly under the scrutiny of regulators both in the US and the EU. Policymakers are currently evaluating the imposition of some degree of data sharing to dominant tech firms in their effort to balance its market power [23]. However, designing such a policy is complex in the case of data assets due to its potential harm to privacy and security.

Within the realm of *personal data*, *protecting privacy* was the main purpose of recent lawmaking in the EU and the US. New legislative proposals in the EU [185, 189] aim to foster data sharing and '*offer an alternative to data handling practice of major tech platforms*' [187]. Assuming regulatory bodies are able to enforce such regulations, some authors have proposed that individuals are compensated for their personal data [114, 155], while others suggest that entities collecting personal data must be required to act as fiduciaries [54], or even that the mass collection and sharing of sensitive personal data must be banned and prosecuted [192].

In the next sections we propose some extensions of our work that may help with addressing some of the challenges above.

## 8.2. Measuring a human-centric data economy in collaboration with real world data marketplaces

As Peter Drucker stated, "*If you can't measure it, you can't manage it*". And you manage what you measure, so you had better count on reliable accurate right measurements about the data economy to support strategic management and policy decision-making.

First, we think that extending our measurement study and involving data marketplace platforms can render better knowledge about the volume of data traded in the market and the eco-

nomic value unleashed by data exchanges. In part II of this dissertation we have carried out the most complete survey of entities trading data on the Internet up to data, and a first of its kind measurement study about the prices of data in commercial data marketplaces. Even though we managed to create a data base of more than two thousand data providers, and a graph showing their relationship with target marketplaces, it restricts to data product offerings and hence is unaware of actual data transactions happening in this ecosystem. Since an increasing number of data marketplaces are relying on blockchains to register such operations, including lighthouse initiatives such as Gaia-X and IDSA, we believe that further valuable insights might be obtained by tapping into the information of those registries, such as who sells what, to whom and at what price. Furthermore, extending this approach in time will definitely give a more accurate notion about the evolution of the data trading ecosystem and where it is heading. As opposed to our work, which scraped only publicly available information on the Internet, this initiative may likely require the collaboration of data marketplace platforms, and it will also need to be carefully designed to preserve the privacy of all the parties involved in data transactions. Still, we believe it could definitely be a valuable add-on to existing data marketplaces and to services of standardisation initiatives such as IDS and Gaia-X.

Second, we think that more insights into how data is being traded can be acquired by automatically processing and analysing the terms and conditions of use of commercial data marketplaces. As part of our survey, we located and downloaded a number of those documents. Most data trading platforms make available public versions of them in their websites. These documents could be analysed to get more systematic information about how data is being traded and how PIMS deal with their users. One could use text mining and NLP techniques to analyse those documents, build a catalogue of "standard" common terms and conditions and find out which of them are applied by each platform, how they differ between the different types of data, etc. Similar works have been carried out in the field of automated public policy analysis [48]. Moreover, this work would also prove valuable in enforcing the rights of users granted by law, such as the ones imposed on data intermediaries (e.g., PIMS) by the current proposal of a "Data Act" by the EU [189].

## 8.3. Building a data pricing tool

As we have shown in part II, data sellers find it challenging to decide on prices. 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 8.1 depicts the main building blocks of such a data quotation tool. Moreover, we have shown the feasibility of such a tool by providing a first implementation of some of these modules in chapter 4.

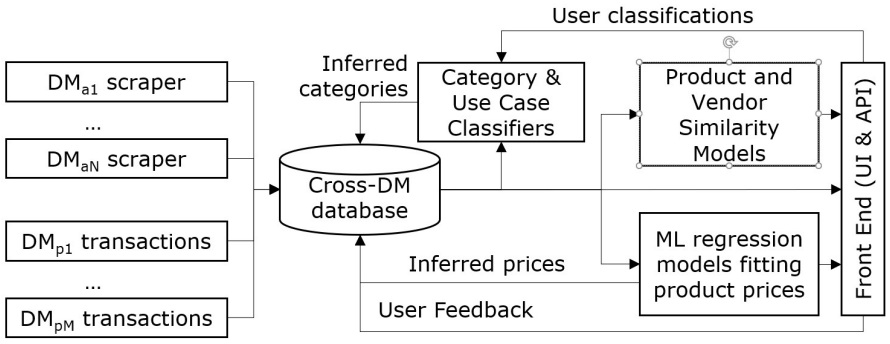



Figure 8.1: Block diagram of a data quotation tool

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. 4.1, 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 standardised and structured data about data products to the cross-DM database, the centralised data warehouse of the tool. Table 4.3 summarised some fields describing data products that are driving the prices of data, and hence must be extracted and loaded into the cross-DM database 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 standardised data categories and use cases to data product. Such models 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. We showed that this is feasible in Sect. 4.4.

Similarly, ML regression models will be trained to fit prices found in commercial DMs taking as input all the metadata features presented in Table 4.3. Regression models like the ones we tested in Sect.4.5 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. Such models will be used by the tool to provide information about similar products or vendors when the user is accessing the contents of the cross-DM database.

The front-end will provide at least the following functions for the end 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 refine its results, such as the competition or the price trend of data observed in the segment of the market the data product is aimed to. Both factors require continuous harvesting and analysis of market data across time, and will provide valuable information for stakeholders to make decisions on pricing or purchasing products in the market.

## 8.4. Federating human-centric data marketplaces

To solve the current fragmentation observed in the data ecosystem, there is an urgent need to improve the interoperability of data markets. Figure 8.2 depicts a reference architecture of a federated data marketplace ecosystem.

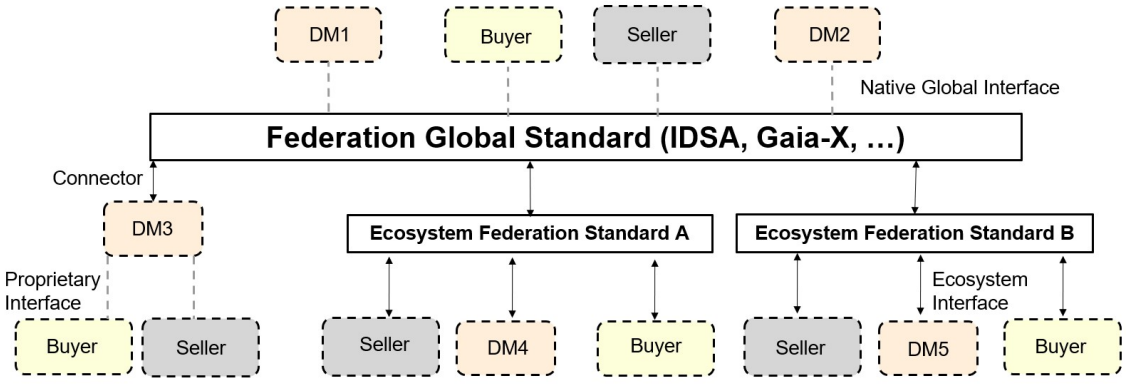



Figure 8.2: Reference architecture of a federation of data marketplaces

The architecture includes DMs of different types. There are standalone marketplaces with proprietary interfaces used by their buyers and sellers (e.g., Snowflake or AWS marketplace). There are also groups of marketplaces federated over an ecosystem standard such as Ocean Protocol, with a population of buyers and sellers. Finally, global federation standards aim to interconnect different ecosystem standards as well as DMs, native buyers and sellers originally built to support

them. In the figure, DM1 and DM2 are data marketplaces that support an emerging Federation Global Standard like Gaia-X or IDS, and can thus be accessed by any buyer and seller that also supports them natively, as well as federate with other DMs that follow these standards. On the other extreme, DM3 is a standalone data marketplace that is accessed by its own buyers and sellers through its proprietary interface. DM3 can become part of the global federation through a compatible certified connector. DM4 and DM5 participate in different ecosystems and thus can only be accessed by other entities federated within each one of these distinct ecosystems. However, as individual standalone data marketplaces, federated ecosystems can communicate with the global federated standard via the corresponding data connectors.

Designing a successful data marketplace calls for solving a plethora of challenges beyond the ones we have addressed in this dissertation and making complex decisions about how the platform would operate. Some key open issues that require further work are the following:

- Ensuring data sovereignty and protecting data ownership and the rights of data providers against piracy and theft (in and outside the platform).
- Making data easily and fully discoverable and portable.
- Accounting for and explaining payoffs resulting from data transactions.
- Designing a robust charging scheme that prevents strategic behaviours of buyers (e.g., arbitrage), or sellers (e.g., benefiting from selling different copies of the same data).
- Ultimately ensuring a fair allocation of economic incentives between all the parties involved in a data transaction, while making the platform economically sustainable.

Fortunately, some existing cutting-edge technology will decisively help in overcoming these daunting challenges, namely:

**Effective data provenance**, even spanning outside of trust ecosystems, may be built upon advances on watermarking [5, 42, 58, 93, 120], hashing [78, 202], trusted execution environments [161], privacy-preservation techniques [158], and network tomography [36].

**Effective data protection**. By combining a blockchain with hardware-based Trusted Execution Environments (TEE), it is possible to create computing platforms that cannot, by design, decrypt and process data in a way that violates the agreements that are recorded in smart contracts [133].

**Usage-based economics** may be built upon crowdsourcing of the market [91, 92], tracking cookies when accessing data services, cryptography [56], and binding between file-level ownership and Non Fungible Tokens (NFTs) [194] and using blockchain-based smart contracts will not only increase ownership, but also enable other usage-tracking technologies.

**Information Centric Network** principles [6] at its data layer provide a base to handle data naming, routing, and in-network storage and replication.

With regards to efficiency and sustainability, we believe that providing scalable and explainable alternatives to ML data marketplaces that avoids buyers sharing confidential intellectual property with DMs, and that makes the data purchasing process more efficient is an open issue. In Chapter 7 we pointed to some heuristics that are correlated with the value of data for spatio-temporal tasks. They are both more efficient to be computed and easier to understand by end users. Some authors have also identified and used influence functions to evaluate the impact of data in a ML model [193], and even outlier detection metrics can also be used to evaluate data more efficiently without disclosing the model to the marketplace. Another way would be to use simpler or different faster models to evaluate data, instead of sharing a working or master version of the ML model the buyer wants to optimise. A recent paper is also proposing to use reinforcement learning to select data for a specific ML task, by attaching weights to training samples [200]. Working with such weights, one could also calculate a distribution of payoffs to data providers according to the possibility that their samples are selected.

Using these faster alternative data valuation methods allows for accelerating both data selection and payment division processes. However, this time saving comes at a cost, as well. First, the buyer would need to train the end model after the sourcing process. Second, since heuristics, influence functions or simpler models just approximate the behaviour of the master model, the selection of data will be sub-optimal, and the corresponding Shapley value calculation will be an approximation to the real Shapley value, just as data Shapley or other popular approximation methods [75]. Measuring the trade-off between between efficiency and accuracy must also be part of the objectives of works evaluating the feasibility of using these alternative methods.

# Chapter 9

# Conclusion

With the advent of AI and ML, data is becoming a cornerstone of an increasingly digitalised and data-driven economy. In this context, the widespread collection of personal data is raising privacy and data protection concerns among people, let alone an increasing fear of being displaced from the labour market by AI, ML models and robots. Some visionary authors have proposed that people take control of their data, and get paid for them in accordance to their contribution to a data driven economy. We have also witnessed public and private initiatives, and rule-making heading towards this direction in the last years.

In this thesis, we have systematically studied key aspects of the vast ecosystem of entities trading data on the Internet, and carried out a first of its kind measurement study of the prices of data in commercial data marketplaces. We have also studied the business models and challenges of Personal Information Management Systems (PIMS) as a sort of primitive data union of users striving to exert their rights relating to the protection of their personal data. Building on this idea of moving towards a human-centric data economy that pays people back for the data they supply, we have contributed to solving the problems of buyers selecting and consuming only suitable data for their tasks (as opposed to the current common practice of digital firms amassing as much data as possible), and platforms deciding on a fair retribution to users contributing data to a specific data transaction.

Chapter 3 presented what is, to the best of our knowledge, the most comprehensive survey of entities trading data over the Internet. We checked and learnt about more than 190 entities, and identified ten different business models they are adopting in the data economy. We responded to key questions regarding how they share, sell or exchange data, and we introduced relevant ongoing standardisation efforts looking forward to federating the fragmented data ecosystem. This introductory chapter allowed us to frame the scope of our work within these initiatives.

Data pricing is an open hot topic in research, and also a challenge for entities trying to monetise their data assets in the market. Chapter 4 presented a novel measurement study of the prices of data products in commercial data marketplaces. We designed and developed a novel methodology to gather information about real data products being traded in commercial data marketplaces.

As a result, we showed that data products listed in commercial DMs may cost from few to hundreds of thousands of US dollars, and that products about telecommunications, manufacturing, automotive, and gaming command the highest prices nowadays. We also developed classifiers for comparing categories of data products across different DMs, as well as a regression analysis that revealed features that correlate with prices of data products belonging in specific categories, such as update rate or history for financial data, and volume and geographical scope for marketing data.

Chapter 5 introduced a generic process for acquiring data in a data marketplace offering a number of suitable datasets from different vendors, and it defined two key problems for these platforms: buyers selecting the most valuable datasets for their task, and marketplaces distributing payoffs to individual vendors contributing data to a single transaction according to the value their data brings to the buyer. We dealt with these two problems in chapters 6 and 7. Interestingly, the complexity of both problems is further exacerbated by an increasing supply of data by many individuals or data providers as one may expect of a thriving human-centric data economy.

With regards to the data selection process, we proposed adding a preliminary appraisal step before sharing any data, and presented a new method for optimising data purchasing decisions. We showed that if a marketplace provides potential buyers with a measure of the performance of their models on *individual* datasets, then they can select which of them to buy with an efficacy that approximates the optimal purchase, i.e., the combination of datasets yielding the maximum profit for the buyer and the specific task, which can only be determined by knowing the performance of each possible combination of them. We called the resulting algorithm *Try Before You Buy* (TBYB) and demonstrated over synthetic and real datasets that TBYB can lead to near optimal data purchasing with only $O(N)$ instead of $O(2^N)$ information and execution time.

With regards to the payoff distribution problem, we studied the problem of computing the relative value of spatio-temporal datasets combined in marketplaces for predicting transportation demand and travel time in metropolitan areas. Using large datasets of taxi rides from Chicago, Porto and New York we showed that simplistic but popular approaches for estimating the relative value of data, such as splitting it equally among the data sources, more complex ones based on volume or the "leave-one-out" heuristic, are inaccurate. Instead, more complex notions of value from economics and game-theory, such as the Shapley value, need to be employed if one wishes to capture the complex effects of mixing different datasets on the accuracy of forecasting algorithms. However, the Shapley value entails serious computational challenges. Its exact calculation requires repetitively training and evaluating combinations of data in $O(N!)$ or $O(2^N)$ computational time, which is unfeasible for complex models or thousands of individual contributors. To help with this challenge, we identified some heuristics, such as entropy or similarity to the average, that showed a significant correlation with the Shapley value, and therefore could be used to overcome the significant computational challenges posed by Shapley approximation algorithms. Therefore, our work has also paved the way to new methodology to measure the value of spatio-temporal data in prediction tasks.

Furthermore, we have also pointed at open challenges and future related research directions in chapter 8. The data economy is still at its nascent phase, and hence there are lots of hot research topics related to data marketplaces such as ensuring data discoverability or protecting data ownership, apart from the ones we tackle in this dissertation. Moreover, we have also proposed specific future works leveraging the contributions of this dissertation.

First, we think that our work on understanding and measuring the data economy can be further extended in time and improved by enriching our survey and price measurement study with data from actual transactions, and by automatically processing and analysing the terms of use of commercial marketplaces.

Second, we propose a high-level design of a data pricing tool aimed to provide users with a hint of the market price of data, based on historical prices of similar datasets in commercial marketplaces. Chapter 4 shows the feasibility of building such a tool and shares some finding sand insights resulting from the implementation of some of its key modules.

Third, we identify specific challenges related to standardizing a federation of data marketplaces, and we point at specific technologies that can help with the problems of protecting data ownership. tracking data provenance and establishing a sustainable and efficient data economy based on data usage.

In conclusion, we believe the data economy will develop fast in the upcoming years, and that researchers from different disciplines will work together to unlock the value of data and make the most out of it. Maybe the revolutionary proposal of getting paid for our data according to the value it generates in the economy finally materialises, or maybe it is other proposals such as the robot tax that are finally used to balance the power between individuals and tech firms in the economy. Still, we hope the work and findings of this dissertation will contribute to shedding light, and increasing transparency about the data value chain, the value of data, and its pricing in existing commercial data markets on the Internet.

# References


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

[2] Advaneo. Data marketplace. access to the world of data. Accessed: Feb'23.

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

[4] G. Aggarwal, A. Fiat, A. V. Goldberg, J. D. Hartline, N. Immorlica, and M. Sudan. Derandomization of auctions. In *Proc. of STOC*. ACM, 2005.

[5] R. Agrawal and J. Kiernan. Watermarking relational databases. In *Proc. of VLDB*, 2002.

[6] B. Ahlgren, C. Dannewitz, C. Imbrenda, D. Kutscher, and B. Ohlman. A survey of information-centric networking. *IEEE Comm. Magazine*, 2012.

[7] Airbloc. Airbloc protocol Technical Whitepaper. 2.0, 2018. Accessed: Feb'23.

[8] H. Aly, J. Krumm, G. Ranade, and E. Horvitz. On the value of spatiotemporal information: principles and scenarios. In *Proc. of ACM SIGSPATIAL*, 2018.

[9] H. Aly, J. Krumm, G. Ranade, and E. Horvitz. To buy or not to buy: Computing value of spatiotemporal information. *ACM Transactions on Spatial Algorithms and Systems*, 2019.

[10] L. Amichi, A. C. Viana, M. Crovella, and A.F. Loureiro. From movement purpose to perceptive spatial mobility prediction. In *Proc. of SIGSPATIAL*. ACM, 2021.

[11] I. Arrieta-Ibarra, L. Goff, D. Jiménez-Hernández, J. Lanier, and E. G. Weyl. Should we treat data as labor? moving beyond "free". *AEA Papers and Proceedings*, 108, 2018.

[12] C. Artigas. Keynote at the 1st acm workshop on the data economy, 2022.

[13] International Data Spaces Association. International Data Spaces Information Model: Ontology draft, 2021. Accessed: Feb'23.

[14] International Data Spaces Association. Gaia-X and International Data Spaces, 2021. Accessed: Feb'23.

[15] International Data Spaces Association. International Data Spaces Reference Architecture Model v4.0, 2022. Accessed: Feb'23.

[16] International Data Spaces Association. International Data Spaces Information Model, 2022. Accessed: Feb'23.

[17] International Data Spaces Association. Data Connector Report, 2022. Accessed: Feb'23.

[18] International Data Spaces Association. Web page, 2023. Accessed: Feb'23.

[19] Y. Bachrach, E. Elkind, R. Meir, D. Pasechnik, M. Zuckerman, J. Rothe, and J. Rosenschein. The cost of stability in coalitional games. In *Symposium of Algorithmic Game Theory, SAGT*, 2009.

[20] M. Balazinska, B. Howe, and D. Suciu. Data markets in the cloud: An opportunity for the database community. In *Proc. VLDB*, 2011.

[21] M. Balazinska, Bill Howe, P. Koutris, D. Suciu, and P. Upadhyaya. A discussion on pricing relational data. In *Search of Elegance in the Theory and Practice of Computation*, 2013.

[22] Battlefin. Better Your Investments Using Alternative Data. Accessed: Feb'23.

[23] C. Biancotti and P. Ciocca. Opening internet monopolies to competition with data sharing mandates. *Peterson Institute for International Economics. Policy Brief*, 2019.

[24] Edward L. Bird, S. and Ewan K. *Natural Language Processing with Python*. O'Reilly Media Inc, 2009.

[25] L. Breiman. Random forests. *Machine Learning*, 45(1), 2001.

[26] D. Brickley, M. Burgess, and N. Noy. Google dataset search: Building a search engine for datasets in an open web ecosystem. In *Proc. of WWW*, 2019.

[27] Y. M. Brovman, M. Jacob, N. Srinivasan, S. Neola, D. Galron, R. Snyder, and P. Wang. Optimizing similar item recommendations in a semi-structured marketplace to maximize conversion. In *Proc. ACM RecSys*. ACM, 2016.

[28] J. Bughin, J. Seong, J. Manyika, M. Chui, and R. Joshi. Notes from the ai frontier: Modeling the impact of ai on the world economy. *McKinsey Global Institute*, 2018.

[29] Daskalakis C., Deckelbaum A., and Tzamos C. Strong duality for a multiple-good monopolist. *Econometrica*, 85(3), 2017.

[30] C. Humby. Data is the new oil! *Keynote at ANA Senior Marketer's Summit, Kellogg School*, 2006. Accessed: Feb'23.

[31] J. Cabañas, Á. Cuevas, and R. Cuevas. Fdvt: Data valuation tool for facebook users. In *Proc. of CHI Conf.*, 2017.

[32] S. Cabello and T. M. Chan. Computing Shapley values in the plane. *Discrete & Computational Geometry*, 2022.

[33] T. Cao, H. Truong, T. Truong-Huu, and M. Nguyen. Enabling awareness of quality of training and costs in federated machine learning marketplaces. In *Proc. of IEEE/ACM International Conference on Utility and Cloud Computing*, 11 2022.

[34] 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. WWW*, 2013.

[35] J. Castro, D. Gomez, and J. Tejada. Polynomial calculation of the shapley value based on sampling. *Computers and Operations Research*, 36, 2009.

[36] R. Castro, M. Coates, G. Liang, R. Nowak, and B. Yu. Network Tomography: Recent Developments. *Stat. Sci.*, 19(3), 2004.

[37] G. Cattaneo, G. Micheletti, M. Glennon, C. La Croce, and C. Mitta. The european data market monitoring tool. d2.9 final study report. *European Commission*, 2020.

[38] S. Chasins, A. Cheung, N. Crooks, A. Ghodsi, K. Goldberg, J. E. Gonzalez, J. M. Hellerstein, M. I. Jordan, A. D. Joseph, M. W. Mahoney, A. Parameswaran, D. Patterson, R. A. Popa, K. Sen, S. Shenker, D. Song, and I. Stoica. The sky above the clouds, 2022.

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

[40] L. Chen, P. Koutris, and A. Kumar. Model-based pricing: Do not pay for more than what you learn! In *Proc. of DEEM'17*, 2017.

[41] L. Chen, P. Koutris, and A. Kumar. Towards model-based pricing for machine learning in a data marketplace. In *Proc. of SIGMOD'19*. ACM, 2019.

[42] Z. Chen, Z. Wang, and C. Jia. Semantic-integrated software watermarking with tamperproofing. *Multimedia Tools Appl.*, 77(9), 2018.

[43] H. Chesbrough and R. Rosenbloom. The role of the business model in capturing value from innovation: Evidence from xerox corporation's technology spin-off companies. *Industrial and Corporate Change*, 11, 2002.

[44] European Commission. Communication on Building a European Data Economy, 2017. Accessed: Feb'23.

[45] European Commission and IDC. EU Data Landscape, 2021. Accessed: Feb'23.

[46] C. F. Costa and M. A. Nascimento. Last mile delivery considering time-dependent locations. In *Proc. of SIGSPATIAL*. ACM, 2021.

[47] D. Coyle, S. Diepeveen, J. Wdowin, J. Tennison, and L. Kay. The value of data – policy implications. *Bennett Institute for Public Policy, Cambridge*, 2020.

[48] H. Cui, R. Trimananda, A. Markopoulou, and S. Jordan. Poligraph: Automated privacy policy analysis using knowledge graphs, 2022.

[49] E. Curry. *New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe*, chapter The Big Data Value Chain: Definitions, Concepts, and Theoretical Approaches. Springer International Publishing, 2016.

[50] M. Dahleh. Why the Data Marketplaces of the Future Will Sell Insights, Not Data, 2018.

[51] S. Dalmolen, H. J.M. Bastiaansen, M. Kollenstart, and M. Punter. Infrastructural sovereignty over agreement and transaction data ('metadata') in an open network-model for multilateral sharing of sensitive data. In *Proc. of ICIS 2019*. AIS, 2020.

[52] DataRade.ai. Platforms, 2022. Accessed: Feb'23.

[53] S. Deep and P. Koutris. The design of arbitrage-free data pricing schemes. In *ICDT*, 2017.

[54] S. Delacroix and N. D. Lawrence. Bottom-up data Trusts: disturbing the 'one size fits all' approach to data governance. *International Data Privacy Law*, 9(4), 2019.

[55] L. Denoyer and P. Gallinari. Bayesian network model for semi-structured document classification. *Inf. Process. Manage.*, 40(5), 2004.

[56] D. Derler and D. Slamanig. Highly-efficient fully-anonymous dynamic group signatures. In *Proc. of ASIACCS*, 2018.

[57] S. Diepeveen and J. Wdowin. The value of data policy implications report – accompanying literature review. *Bennett Institute for Public Policy, Cambridge*, 2020.

[58] G. J. Doërr and J.L. Dugelay. A guide tour of video watermarking. *Signal Processing: Image Communication*, 18, 2003.

[59] P. Domingos and M. Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29(2–3), 1997.

[60] S. W. Driessen, G. Monsieur, and W. J. Van Den Heuvel. Data market design: A systematic literature review. *IEEE Access*, 10, 2022.

[61] W. D. Eggers, R. Hamill, and A. Ali. Data as the new currency Government's role in facilitating the exchange. *Deloitte Review*, 2013. Accessed: Feb'23.

[62] Open Evidence, The Lisbon Council, and IDC. The European Data Market Monitoring Tool, 2020. Accessed: Feb'23.

[63] S. S. Fatima, M. Wooldridge, and N. R. Jennings. A linear approximation method for the shapley value. *Artificial Intelligence*, 2008.

[64] R. C. Fernandez, P. Subramaniam, and M. J. Franklin. Data market platforms: Trading data assets to solve data problems. In *Proc. of VLDB Endow.*, volume 13, 2020.

[65] Organisation for Economic Co-operation and Development. Benchmarking broadband prices in the oecd, 2004.

[66] AMO Foundation. Fulfilling driving experiences. Driving data platform for innovative mobility. Accessed: Feb'23.

[67] Ocean Protocol Foundation and BigchainDB GmbH. Ocean Protocol: Tools for the web3 data economy. whitepaper, 2022. Accessed: Feb'23.

[68] J. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29, 11 2000.

[69] Gaia-X. Gaia-X Trust Framework, 2022. Accessed: Feb'23.

[70] Gaia-X. Gaia-X Architecture Document, 2022. Accessed: Feb'23.

[71] Gaia-X. Gaia-X Policy Rules Document, 2022. Accessed: Feb'23.

[72] Gaia-X. Web page, 2023. Accessed: Feb'23.

[73] C. Gates and P. Matthews. Data is the new currency. In *Proc. of NSPW*. ACM, 2014.

[74] GeoDB. Decentralized peer-to-peer big data sharing ecosystem. Accessed: Feb'23.

[75] A. Ghorbani and J. Zou. Data shapley: Equitable valuation of data for machine learning. In *Proc. of ICML*, 2019.

[76] A. Ghosh and A. Roth. Selling privacy at auction. In *Proc. of EC*. ACM, 2011.

[77] L. Giaretta, T. Marchioro, E. Markatos, and Š. Girdzijauskas. Towards a decentralized infrastructure for data marketplaces: Narrowing the gap between academia and industry. In *Proc. Workshop on the Data Economy*, DE '22. ACM, 2022.

[78] A. Gionis, P. Indyk, and R. Motwani. Similarity search in high dimensions via hashing. In *Proc. of VLDB*, 1999.

[79] A. V. Goldberg and J. D. Hartline. Competitiveness via consensus. In *Proc. of ACM-SIAM SODA*, 2003.

[80] A. V. Goldberg, J. D. Hartline, A. R. Karlin, M. Saks, and A. Wright. Competitive auctions. *Games and Economic Behavior*, 55(2), 2006. Special Issue: Electronic Market Design.

[81] A. V. Goldberg, J. D. Hartline, and A. Wright. Competitive auctions and digital goods. In *Proc. of ACM-SIAM SODA*, 2001.

[82] GSMA and AT Kearney. The Data Value Chain, 2018. Accessed: Feb'23.

[83] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han. Outlier detection for temporal data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 26(9), 2014.

[84] N. Gupta, H. Patel, S. Afzal, N. Panwar, R. S. Mittal, S. Guttula, A. Jain, L. Nagalapatti, S. Mehta, S. Hans, P. Lohia, A. Aggarwal, and D. Saha. Data quality toolkit: Automatic assessment of data quality and remediation for machine learning datasets, 2021.

[85] N. Haghpanah and J. Hartline. When Is Pure Bundling Optimal? *The Review of Economic Studies*, 88(3), 08 2020.

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

[87] N. Henke, J. Bughin, M. Chui, J. Manyika, T. Saleh, and B. Wiseman. The age of analytics: Competing in a data-driven world. *McKinsey Global Institute*, 2016.

[88] M. Hils, D. W. Woods, and R. Böhme. Measuring the emergence of consent management on the web. In *Proc. of the ACM IMC*. ACM, 2020.

[89] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 42(1), 2000.

[90] Data Intelligence Hub. (DIH) extract value from data securely. Accessed: Feb'23.

[91] C. Iordanou, N. Kourtellis, J. M. Carrascosa, C. Soriente, R. Cuevas, and N. Laoutaris. Beyond content analysis: Detecting targeted ads via distributed counting. In *Proc. of CoNEXT*. ACM, 2019.

[92] C. Iordanou, C. Soriente, M. Sirivianos, and N. Laoutaris. Who is fiddling with prices? building and deploying a watchdog service for e-commerce. In *Proc. of SIGCOMM*. ACM, 2017.

[93] D. Isler, E. Cabana, and N. Laoutaris. Freqywm: Frequency watermarking for the new data economy, 2022.

[94] N. Jha, M. Trevisan, L. Vassio, M. Mellia, S. Traverso, A. Garcia-Recuero, N. Laoutaris, A. Mehrjoo, S. Andrés Azcoitia, R. Cuevas Rumin, K. Katevas, P. Papadopoulos, N. Kourtellis, R. Gonzalez, X. Olivares, and G. M. Kalatzantonakis-Jullien. A pims development kit for new personal data platforms. *IEEE Internet Computing*, 26(3), 2022.

[95] R. Jia, D. Dao, B. Wang, F. Hubis, N. Gurel, C. Zhang, C. Spanos, and D. Song. Efficient task-specific data valuation for nearest neighbor algorithms. *Proc. of VLDB*, 12, 2019.

[96] R. Jia, D. Dao, B. Wang, F. A. Hubis, N. Hynes, N. M. Gürel, B. Li, C. Zhang, D. Song, and C. J. Spanos. Towards efficient data valuation based on the shapley value. In *Proc. of ML Research*, 2019.

[97] C. I. Jones and C. Tonetti. Nonrivalry and the economics of data. *American Economic Review*, 110(9), 2020.

[98] Kaggle. Ecml/pkdd 15: Taxi Trajectory Prediction, 2015. (Accessed: Feb'23).

[99] J. Kang, Z. Xiong, D. Niyato, S. Xie, and J. Zhang. Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory. *IEEE IoT Journal*, 6(6), 2019.

[100] Y. Kassa, J. Gonzalez, Á. Cuevas, R. Cuevas, M. Marciel, and R. Gonzalez. Your data in the eyes of the beholders: Design of a unified data valuation portal to estimate value of personal information from market perspective. In *Proc. of ARES*, 2016.

[101] J. Kennedy, P. Subramaniam, S. Galhotra, and R. Castro Fernandez. Revisiting online data markets in 2022: A seller and buyer perspective. *SIGMOD Record*, 51(3), 2022.

[102] Keras. Keras - Simple. Flexible. Powerful. Accessed: Feb'23.

[103] E. King. Data is like water. Accessed: Feb'23.

[104] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proc. of ICLR*, 2015.

[105] Y. Ko. A study of term weighting schemes using class information for text classification. In *Proc. of SIGIR*. ACM, 2012.

[106] N. Kourtellis, K. Katevas, and D. Perino. Flaas: Federated learning as a service. In *Proc. of Workshop on Distributed Machine Learning*, 2020.

[107] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu. Querymarket demonstration: Pricing for online data markets. *Proc. of VLDB*, 5, 2012.

[108] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu. Query-based data pricing. *Journal of ACM*, 62(5), November 2015.

[109] V. Koutsos, D. Papadopoulos, D. Chatzopoulos, S. Tarkoma, and P. Hui. Agora: A privacy-aware data marketplace. In *IEEE ICDCS*, 2020.

[110] O. Kramer. Unsupervised k-nearest neighbor regression, 2011.

[111] G. Krishnaveni and T. Sudha. Naïve bayes text classification ś a comparison of event models. *Imperial Journal of Interdisciplinary Research*, 3, 2016.

[112] A. Kumar, B. Finley, T. Braud, S. Tarkoma, and P. Hui. Marketplace for ai models, 2020.

[113] Moody D. L. and Walsh P. Measuring the value of information - an asset valuation approach. In *European Conference on Information Systems ERCIS*, 1999.

[114] J. Lanier. *Who Owns the Future?* Simon & Schuster, 2013.

[115] N. Laoutaris. Why online services should pay you for your data? the arguments for a human-centric data economy. *IEEE Internet Computing*, 2019.

[116] N. Laoutaris and C. Iordanou. What do information centric networks, trusted execution environments, and digital watermarking have to do with privacy, the data economy, and their future? *SIGCOMM Computing Comm. Rev.*, 2021.

[117] C. Li, D. Y. Li, G. Miklau, and D. Suciu. A theory of pricing private data. *ACM Trans. Database Syst.*, 2015.

[118] Wendy C.Y. LI, NIREI Makoto, and YAMANA Kazufumi. Value of Data: There's No Such Thing as a Free Lunch in the Digital Economy. Discussion papers 19022, Research Institute of Economy, Trade and Industry (RIETI), 2019.

[119] F. Liang, W. Yu, D. An, Q. Yang, X Fu, and W. Zhao. A survey on big data market: Pricing, trading and protection. *IEEE Access*, 6:15132–15154, 2018.

[120] X. Liang and S. Xiang. Robust reversible audio watermarking based on high-order difference statistics. *Signal Processing*, 173, 2020.

[121] B. R. Lin and D. Kifer. On arbitrage-free pricing for general data queries. *Proc. VLDB Endow.*, 7(9), 2014.

[122] J. Liu, J. Lou, J. Liu, L. Xiong, J. Pei, and J. Sun. Dealer: An end-to-end model marketplace with differential privacy. *Proc. of VLDB*, 14(6), 2021.

[123] A. Löser, F. Stahl, A. Muschalle, and G. Vossen. Pricing approaches for data markets. In *Proc. of the BIRTE*, 08 2012.

[124] O'Rourke M. and Rogerson D. A practical guide on benchmarking telecommunication prices, 2014.

[125] David J. C. MacKay. Bayesian interpolation. *Neural Comput.*, 4(3), 1992.

[126] M. Maschler and B. Peleg. A characterization, existence proof and dimension bounds for the kernel of a game. *Pacific Journal of Mathematics*, 18, 1966.

[127] Analysis Mason. Online data economy value chain, 2014. Accessed: Feb'23.

[128] Analysis Mason. What is the IoT value chain and why is it important?, 2020. Accessed: Feb'23.

[129] L. Mason, J. Baxter, P. Bartlett, and M. Frean. Boosting algorithms as gradient descent. In *Proc. of NIPS*. MIT Press, 1999.

[130] S. Matic, C. Iordanou, G. Smaragdakis, and N. Laoutaris. Identifying sensitive urls at web-scale. In *Proc. of IMC*. ACM, 2020.

[131] S. Mehta, M. Dawande, G. Janakiraman, and V. Mookerjee. How to sell a dataset?: Pricing policies for data monetization. In *Information Systems Research, Forthcoming*, 2019.

[132] D. Moor. Data markets with dynamic arrival of buyers and sellers. In *Proc. of NetEcon*. ACM, 2019.

[133] M. Müller, A. Simonet-Boulogne, S. Sengupta, and O. Beige. Process mining in trusted execution environments: Towards hardware guarantees for trust-aware inter-organizational process analysis. In *Process Mining Workshops*. Springer International Publishing, 2022.

[134] A. Muschalle, F. Stahl, A. Löser, and G. Vossen. Pricing approaches for data markets. In *Enabling Real-Time Business Intelligence*. Springer, 2013.

[135] K. Nguyen, J. Krumm, and C. Shahabi. Spatial privacy pricing: The interplay between privacy, utility and price in geo-marketplaces. In *Proc. SIGSPATIAL*, 2020.

[136] K. Nguyen, J. Krumm, and C. Shahabi. Quantifying intrinsic value of information of trajectories. In *Proc. of SIGSPATIAL*. ACM, 2021.

[137] C. Niu, Z. Zheng, F. Wu, X. Gao, and G. Chen. Achieving data truthfulness and privacy preservation in data markets. *IEEE Trans. KDE*, 31, 2019.

[138] C. Niu, Z. Zheng, F. Wu, S. Tang, and G. Chen. Online pricing with reserve price constraint for personal data markets. In *2020 IEEE ICDE*, 2020.

[139] OECD. Exploring the economics of personal data: A survey of methodologies for measuring monetary value. *OECD Digital Economy Papers*, 2013.

[140] State of California. California consumer privacy act, 2018.

[141] City of Chicago. Taxi Trips, 2023. Accessed: Feb'23.

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

[143] L. Olejnik, M. Tran, and C. Castelluccia. Selling off privacy at auction. In *Proc. of NDSS*, 2014.

[144] Object Management Group (OMG). Business process model and notation (BPMN) 2.0., 2011. Accessed: Feb'23.

[145] United Nations Conference on Trade and Development UNCTAD. Digital economy report. cross-border data flows and development: For whom the data flow, 2021.

[146] W. Org, J. Becker, K. Backhaus, H. Grob, B. Hellingrath, T. Hoeren, S. Klein, H. Kuchen, U. Müller-Funk, U. Thonemann, G. Vossen, F. Stahl, and F. Schomm. *The Data Marketplace Survey Revisited*. Westf. Wilhelms-Univ., ERCIS, 2014.

[147] A. Osterwalder. The business model ontology. a proposition in a design science approach. In *PhD Thesis*, 2004.

[148] Otonomo. One-Stop Shop for Vehicle Data. Accessed: Feb'23.

[149] Hubert P. and Ricco G. Imperfect information in macroeconomics. *Sciences Po publications*, 2018.

[150] P. Papadopoulos, N. Kourtellis, P. 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. ACM IMC*, 2017.

[151] M. Paraschiv and N. Laoutaris. Valuating User Data in a Human-Centric Data Economy. *arXiv e-prints*, page arXiv:1909.01137, 2019.

[152] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85), 2011.

[153] J. Pei. Data pricing – from economics to data science. In *Proc. of SIGKDD*. ACM, 2020.

[154] Serguei Yu. Popov. The tangle, 2015.

[155] E. Posner and G. Weyl. *Radical Markets. Uprooting Capitalism and Democracy for a Just Society*. Princeton Univ. Press, 2018.

[156] Ocean Protocol. https://oceanprotocol.com/Tools for the Web3 Data Economy, 2023. Accessed: Feb'23.

[157] Daria R. The Future of Data Marketplaces, 2019. Accessed: Feb'23.

[158] K. Koch; S. Krenn; T. Marc; S. More; S. Ramacher. Kraken: A privacy-preserving data market for authentic data. In *Proc. of 1st ACM Workshop on the Data Economy*, 2022.

[159] D. Reinsel, J. Gantz, and J. Rydning. The digitization of the world - from edge to core. *Data Age 2025*, 2018.

[160] B. Rozemberczki, L. Watson, P. Bayer, H. Yang, O. Kiss, S. Nilsson, and R. Sarkar. The shapley value in machine learning. In *Proc. of the IJCAI*, 2022.

[161] M. Sabt, M. Achemlal, and A. Bouabdallah. Trusted execution environment: What it is, and what it is not. In *IEEE Trustcom/BigDataSE/ISPA*, 2015.

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

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

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

[165] Lloyd S. Shapley. *A Value for n-Person Games*. RAND Corporation, Santa Monica, California, 1952.

[166] Y. Shen, B. Guo, Y. Shen, X. Duan, X. Dong, and H. Zhang. A pricing model for big personal data. *Tsinghua Science and Technology*, 2016.

[167] Shutterstock. Web page, 2022. Accessed: Feb'23.

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

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

[170] F. Stahl, F. Schomm, and G. Vossen. Data marketplaces: An emerging species. *Frontiers in Artificial Intelligence and Applications*, 2014.

[171] F. Stahl, F. Schomm, G. Vossen, and L. Vomfell. A classification framework for data marketplaces. *Vietnam Journal of Computer Science*, 3, 03 2016.

[172] R. Stanojevic, N. Laoutaris, and P. Rodriguez. On economic heavy hitters: Shapley value analysis of 95th-percentile pricing. In *Proc. of ACM IMC*, 2010.

[173] C. Strobl, A.L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. Conditional variable importance for random forests. *BMC bioinformatics*, 9:307, 08 2008.

[174] Swash. Reimagining data ownership. Accessed: Feb'23.

[175] Z. Tang, Z. Lv, and C. Wu. A brief survey of data pricing for machine learning. In *Proc. of International Conference on Signal, Image Processing and Pattern Recognition (SIPP)*, 2020.

[176] TAUS. Data Marketplace, 2022. Accessed: Feb'23.

[177] HERE Technologies. Marketplace. Accessed: Feb'23.

[178] Z. Tian, J. Liu, J. Li, X. Cao, R. Jia, and K. Ren. Private data valuation and fair payment in data marketplaces. *arXiv*, 2022.

[179] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58, 1996.

[180] Tim O'Reilly. Data Is the New Sand. *The Information*, 2021. Accessed: Feb'23.

[181] S. Touati, M. S. Radjef, and L. Sais. A bayesian monte carlo method for computing the shapley value: Application to weighted voting and bin packing games. *Computers and Operations Research*, 2021.

[182] European Union. General data protection regulation, April 2016.

[183] European Union. Regulation (eu) 2018/1807 on a framework for the free flow of non-personal data in the european union, 2018.

[184] European Union. Directive (eu) 2019/1024 on open data and the re-use of public sector information (recast), 2019.

[185] European Union. Data governance act, 2020.

[186] European Union. A European strategy for data, 2020. Accessed: Feb'23.

[187] European Union. Press release: Commission proposes measures to boost data sharing and support european data spaces, November 2020. Accessed: Feb'23.

[188] European Union. Guidance on private sector data sharing, 2022. Accessed: Feb'23.

[189] European Union. Proposal for a regulation of the european parliament and of the council on harmonised rules on fair access to and use of data (data act), 2022.

[190] Collins V. and Lanz J. Managing data as an asset. *Certified Public Accountants (CPA) Journal*, June 2019.

[191] T. van Campen, H. Hamers, B. Husslage, and R. Lindelauf. A new approximation method for the shapley value applied to the wtc 9/11 terrorist attack. *Social Network Analysis and Mining*, 8, 2017.

[192] C. Veliz. *Privacy is Power: Why and How You Should Take Back Control of Your Data*. Bantam Press, 2021.

[193] Koh P. W. and Liang P. Understanding black-box predictions via influence functions. In Doina Precup and Yee Whye Teh, editors, *Proc. of ICML*, volume 70. PMLR, 2017.

[194] Q. Wang, R. Li, Q. Wang, and S. Chen. Non-fungible token (nft): Overview, evaluation, opportunities and challenges. *arXiv*, 2021.

[195] M. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. W. Boiten, L. O. o Bonino da Silva Santos, P. Bourne, J. Bouwman, A. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. Evelo, R. Finkers, and B. Mons. The fair guiding principles for scientific data management and stewardship. *Nature Scientific Data*, 3, 2016.

[196] C. Wu, R. Buyya, and K. Ramamohanarao. Cloud pricing models: Taxonomy, survey, and interdisciplinary challenges. *ACM Computing Surveys*, 52, 2019.

[197] X. Xu, A. Hannun, and L. Van Der Maaten. Data appraisal without data sharing. In *Proc. of Machine Learning Research*, 2022.

[198] Li Yan, Haiying Shen, Zhuozhao Li, Ankur Sarker, John A. Stankovic, Chenxi Qiu, Juanjuan Zhao, and Chengzhong Xu. Employing opportunistic charging for electric taxicabs to reduce idle time. *Proc. ACM Interact. Mob. Wearable Ubiquitous Tech.*, 2, 2018.

[199] T. Yan and A. D. Procaccia. If you like shapley then you'll love the core. *Proc. of the AAAI Conf.*, 35(6), 2021.

[200] J. Yoon, S. Arik, and T. Pfister. Data valuation using reinforcement learning. In Hal Daumé III and Aarti Singh, editors, *Proc. of the ICML*, volume 119. PMLR, 2020.

[201] H. Yu and M. Zhang. Data pricing strategy based on data quality. *Computers and Industrial Engineering*, 112, 2017.

[202] K. Zhao, H. Lu, and J. Mei. Locality preserving hashing. *Proc. of the AAAI Conference*, 28(1), 2014.

[203] K. Zhao, S. H. Mahboobi, and S. Bagheri. Shapley value methods for attribution modeling in online advertising. *ArXiv*, 2018.

[204] Z. Zheng, Y. Peng, F. Wu, S. Tang, and G. Chen. An online pricing mechanism for mobile crowdsensing data markets. In *Proc. of Mobihoc*. ACM, 2017.

[205] Z. Zhou, X. Xu, R. H. L. Sim, C. S. Foo, and K. H. Low. Probably approximate shapley fairness with applications in machine learning. *Proc. AAAI Conf.*, 2023.

[206] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B. (Statistical Methodology)*, 67, 2005.

[207] K. R. Özyilmaz, M. Doğan, and A. Yurdakul. Idmob: Iot data marketplace on blockchain. In *Crypto Valley Conference on Blockchain Technology (CVCBT)*, 2018.