The value of data: an analysis of closed-urban-data-based and open-data-based business models

Bruno Carballa Smichowski

▶ **To cite this version:**

Bruno Carballa Smichowski. The value of data: an analysis of closed-urban-data-based and open-data-based business models. 2018. hal-01736484

HAL Id: hal-01736484
https://hal.science/hal-01736484v1

Preprint submitted on 17 Mar 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - ShareAlike 4.0 International License

SciencesPo
CITIES AND DIGITAL TECHNOLOGY CHAIR

Working Paper

> N°01/2018

The value of data
An analysis of closed-urban-data-based and open-data-based business models

Bruno Carballa Smichowski

SciencesPo

CITIES AND DIGITAL TECHNOLOGY CHAIR

The “Cities and Digital Technology” Chair of Sciences Po’s Urban School has been launched in March 2017 to better grasp the impact of digital technologies on urban governance. Funded by four sponsoring firms (Cisco, La Poste, RTE, Caisse des Dépôts), the Chair aims to create new research fields exploring the interaction between digital technology and cities in an empirical and comparative perspective.

The value of data: an analysis of closed-urban-data-based and open-data-based business models

Bruno Carballa Smichowski (Chronos ; CEPN - Université Paris Nord)

bruno.carballa@groupechronos.org

Abstract

The aim of this article is to shed some light on what makes data valuable and analyze the business models that are based on urban data and on open data. The article is structured as follows. Section 1 shows what makes data valuable, namely size, quality and scope. After pointing out that a property regime for data does not exist, Section 2 explains the two main legal strategies that firms employ to appropriate data in order to build business models around it: intellectual-property-based strategies (copyright, patents and sui generis database right) and the terms of use and trade secret combo. Section 3 defines closed-urban-data-based business models and studies them by focusing on how data is obtained, how value is created for the end user from it, and how value is captured. Four types of these business models are distinguished: aggregated-data-based services providers, individual-data-based service providers, trust-based algorithmic coordination platforms and transactional intermediaries. Section 4 focuses on the business models that build on open data to create value and distinguishes five families: government open data, for-profit private firm standalone open data, nonprofit standalone open data, multi-stakeholder data pooling and common-based open data crowdsourcing.

Keywords: urban data, open data, business model

Introduction

The economic relevance of data has become clear to firms, governments and society in general. Assertions such as “data is the new oil” are now a commonplace. Nevertheless, although some scholars have made contributions to the understanding of what makes data valuable (Chignard & Benyayer, 2015; Mayer-Schönberger & Cukier, 2013), the topic is far from having been exhausted. Moreover, the abuse of buzzword ‘platform’ to analyze firms that differ strongly in their core business and in the role played by data in their business models (search engines, social networks, e-commerce, music streaming, connected devices, etc.) is hampering the progress the understanding of data-based business models.

This article aims at contributing to the understanding of what makes data valuable and the different ways in which organizations can use it to create and capture value from it. Because the scope of the economic uses of data is enormous, we have decided to narrow down the analysis to two types of data that we consider of particular interest. The first one is urban data, a resource which governance is currently at the center of political and economic interests that entangle digital firms, States and civil society (The Economist, 2010). In order to get a clear picture of what is at stake and what solutions to the different issues tied to the governance of urban data (privacy, ownership of the data, city halls’ control over public utilities, etc.) are desirable and practicable, understanding urban-data-based business models is essential. The second one is open data. As concerns about the hoarding of tons of data by a handful of tech giants grow, open data has been put forward as a key element in the building of an alternative to regain control over data (Verdier & Murciano, 2017). But in order for open data to gain terrain, studying the business models that can make open data sustainable is of paramount importance.

The understanding of the business models that exist around these two types of data requires an *ex-ante* comprehension of what makes data valuable and, in the case of closed-urban-data-based business models, the legal mechanisms through which data can be appropriated. Consequently, the article will be structured as follows. Section 1 shows what makes data valuable, namely size, quality and scope. After pointing out that a property regime for data does not exist, Section 2 explains the two main legal strategies that firms employ to appropriate data in order to build business models around it: intellectual-property-based strategies (copyright, patents and sui generis database right) and the terms of use and trade secret combo. Section 3 defines closed-urban-data-based business models and studies them by focusing on how data is obtained, how value is created for the end user from it, and how value is captured. Four types of these business models are distinguished: aggregated-data-based services providers, individual-data-based service providers, trust-based algorithmic coordination platforms and transactional intermediaries. Section 4 focuses on the business models that build on open data to create value and distinguishes five families: government open data, for-profit private firm standalone open data, nonprofit standalone open data, multi-stakeholder data pooling and common-based open data crowdsourcing.

1 The value of data

Data is valuable because of what it allows to do. Benyayer and Chignard (2015) summarize what data allows to do in four verbs: describe, explain, predict and prescribe. Nevertheless, data alone is not enough to conjugate those verbs. Statistical analysis methods and algorithms are applied to datasets to obtain valuable information from them and act in consequence. Data feeds algorithms and statistical analysis software to create "information-age refineries" (Cohen, 2017). The value created from data is therefore the combined result of the dataset and the analysis applied to it. The two are inherently inseparable in the value creation process, just like the movement of a car cannot be fully attributed either to the engine or to fuel. But as in the past years computing and storage capacity became increasingly cheap and fast notably as a consequence of the microprocessors revolution (Cohen, Zysman, & DeLong, 2000) and the emergence of cloud computing (Kushida, Murray, & Zysman, 2015), and as software development and data analysis became increasingly abundant, cheaper and outsourceable (Mayer-Schönberger & Cukier, 2013), the capacity to obtain data is today the key source of value creation in data-based business models.

Nevertheless, not any kind of data is valuable. In order for a dataset to allow for proper descriptions, explanations, predictions and prescriptions it needs to have certain properties, namely size, quality and scope. The better a dataset fulfills these conditions the more valuable it becomes. Having this in mind will be important to understand the design of the data-based business models we will describe in sections 3 and 4.

Before developing on each of these three properties, three important remarks have to be made. First, these properties are not merely technical features but in many respects they are the result of human choices, such as licenses or the way in which data collection is designed. Second, as we will show along the following lines, each of these three properties have a different ponderation in making the data valuable depending on the use intended. The value of data is therefore contextual to its use (OECD, 2015). For example, a few months ago Netflix decided to change its rating system from a 10 stars system to a binary like/don't like system. This choice responds to the fact that many users might consider that a 10 stars rating system, unlike a simple like/don't like one, implies some pondering that takes time and, therefore, they would avoid rating systematically. The new rating system probably increased the size of Netflix's database on users' reviews while it decreased its quality in that each review is less precise. One can imagine that Netflix considers that, for the purposes of developing its service based on the study of users' preferences, the size of its database on users' reviews is more important than its quality.

Third, we will only discuss what makes a certain dataset valuable as a resource once it has been already produced, but we will not address the issue of where that value comes from. Indeed, scholars keen to the concept of digital labor (Casilli, 2015; Fuchs, 2014) argue that (most of) the value of data comes from the labor performed by internet users in different ways (clicks, likes, online reviews, browsing, wearing a connected device, etc.) that create the data[1]. Because our focus is on business models built around data, we will investigate

[1] For a good critique of the overreach of the concept of digital labor see Broca (2017).

the *properties* that make data valuable and avoid discussing the source(s) of that value, a topic that would exceed the limits of this article

1.1 Size

The larger a dataset is the more valuable it becomes. The above-mentioned valuable uses of data (describe, explain, predict and prescribe) rely on extracting insightful patterns using statistical techniques. As the results of the latter are more precise and robust as the dataset increases in size, the more data there is the more solid the conclusions that can be drawn from it are.

Moreover, when data is used to 'train' algorithms (i.e. when data serves to feed machine learning) on which many digital firms and other firms largely rely on, the size of datasets is of paramount importance. Algorithms work by processing large amounts of data and they improve due to it. Once an algorithm (or, more generally, a model) is correctly designed to fit its data (once the right questions are asked in the proper way)[2], the more data that an algorithm can work on, the more likely it will be that it will improve over time. Once the team of developers is good enough, the 'algorithm race' becomes a matter of who has more data. As the famous quote by Google's Chief Scientist Peter Norvig goes, "we don't have better algorithms than anyone else; we just have more data" (Cleland, 2011). As machine learning becomes increasingly relevant (and this trend will only intensify with the rise of artificial intelligence incorporated to urban services, among other uses), size gains importance because it allows for faster and more accurate training of algorithms and, therefore, better predictions and prescriptions.

1.2 Quality

While increasing the size of a dataset generally allows extracting more value from it, this might not happen if we are dealing with low quality data. The quality of data refers to the characteristics of a dataset that make it easier to extract meaningful information from it. The meaning of quality is therefore highly dependent on the use intended, since data becomes information in a certain context (Floridi, 2014). This is one of the reasons why a dataset can be very worthy for a certain use but of little interest for others. For example, a dataset about cellphone charging stations in an airport that contains data about how many times cellphones have been plugged in is valuable for managers who want to decide where to place more charging stations based on the number of people that use the existing ones. But if the dataset does not specify how long each charge lasted, it is of little relevance for companies that develop lithium batteries.

It is difficult to list all the properties that constitute quality. In order to illustrate the multi-dimensional nature the term 'quality' acquires to qualify data, we will retain the following categories of quality employed by Floridi (2014): accuracy, objectivity, accessibility, security, relevancy, timeliness, interpretability and understandability. These dimensions of quality are not meant to be definitive or exhaustive, but rather an indication to the reader of what lies behind the word 'quality'. Having these categories in mind will help understanding the strategic choices organizations make around them to create and capture value from it. Other scholars (Batini & Scannapieco, 2006; Olson, 2003; Wang, 1998) have proposed different

[2] For a more detailed explanation of under which conditions does more data improve a model, see Amatriain (2015).

dimensions of the quality of data[3]. Moreover, as mentioned above, the importance of each of the above-quoted dimensions of quality and the extent to which they are fulfilled will depend on the use intended. In our previous example, the relevancy of the dataset is high for the airport manager and low for the company that develops lithium batteries.

While it might seem obvious why each of these dimensions of quality make data valuable, let us reflect on the role that data's lifespan has in affecting data's value through accuracy or relevancy. Some scholars argue that 'fresh' data is always better than 'old' data because data would lose value over time precisely because many of the above-quoted dimensions of the quality of data (especially relevancy and accuracy) would be lessened over time (Lyon, 2016; Sokol & Comerford, 2016). At first glance this assertion seems evident: new data is always preferable and old data is of little interest to describe, explain, predict and prescribe today. The economic implications of such a view are important. If that is the case, then firms hoarding large datasets for long periods of time do not have a competitive advantage over other firms that might want to contest its market but do not have the 'old' data they need. Nevertheless, the fact that data loses value over time is true only to a certain extent and limited to certain uses of data, while in some other uses, on the contrary, having 'old' data can be a source of value creation. For example, Waze requires real-time data to be able to offer an on-demand service that consists in providing routes to users that minimize the time spent on the road. For the immediate purposes of offering that service data loses value by the minute. Nonetheless, as geolocalized data about drivers' routes has a wide range of possible uses, other uses besides the immediate one Waze recurs to in order to provide its service benefit from the accumulation of historical data. Waze could (and certainly does) use its historical databases to understand better how drivers react to alternative routes propositions and use those insights to optimize its algorithms. Researchers and city halls (some of the latter such as the French city of Lille are currently negotiating deals of data exchange with Waze) could use historical databases to gain knowledge on how traffic jams are created and improve the coordination of the functioning of street lights. Then, when analyzing the impact data's lifespan has on its value, one must always bear in mind that different (sometimes even complementary) uses of data exist even within a single organization. As with the three properties that make data valuable (size, quality and scope), the lifespan of data can be either a source of value or depreciation depending on the intended use.

1.3 Scope

The scope of data refers to two related yet distinct properties. One is the fact that a dataset can be easily linked to others. This is what Marzloff (2013) refers to when he writes that "the value of a datum is proportional to the square of the number of data to which it is associated"[4] and what Mayer-Schönberger and Cukier (2013) call "recombining data". For example, a dataset about energy consumption of households alone is valuable to policymakers working on subsidizing energy or for energy providers that want to improve their pricing model. But if the same dataset can be linked to another dataset about socio-demographic information at the household level it becomes even more valuable for these two parties. Then, data gains value through the enhanced utility that comes from linking datasets (Roché, 2016).

[3] For a good review of the literature on the quality of information see Batini and Scannapieco (2006).

[4] The translation is ours.

The other property that constitutes the scope of data is what Mayer-Schönberger and Cukier (2013) call "option value of data": how many different domains a single dataset can provide information about. Datasets that can create links between seemingly unrelated domains are valuable as they enrich the comprehension of a phenomenon (description and explanation), and hence the possibilities of acting (predicting and prescribing) on it in the 'right' way. For example, Google's search data informs about what people look for in the internet. This data can be exploited in many ways because it provides information about a myriad of things. Google has used it to predict flue epidemics by analyzing searches related to the symptoms or medicines. Spin doctors could use search history about politicians to understand better what makes certain candidates noticeable. Car makers could use Google searches about car models to understand what people look for in a car and improve their marketing strategies… and the list could go on.

As the different uses of a same dataset and its combination with other datasets are rarely done within a single organization, the scope of data usually depends on how interoperable it is. Interoperability, in turn, depends on how data is structured (e.g. in the previous example about energy consumption, whether the level of aggregation chosen is the household, the building or the neighborhood) and on its technical properties such as the format. If, for example, the data is in a pdf format, it is less easily linkable to other datasets than if it is under a standardized machine-readable format. If real time applications want to use it, an API will be needed for this dimension of scope to exist. The easier it is for different agents to use data, the more it will circulate, and, hence, the more possibilities there are of enriching and/or exploiting the dataset. This does not mean that every agent has an interest in making its data circulate. Most private firms have an interest in *not* circulating their data to maintain a competitive advantage (Chignard & Benyayer, 2015), as we will show in Section 3. In those cases, the social value of data is in conflict with its economic value for certain actors. Nevertheless, even in cases where firms have no interest in opening their data, they might share momentarily data with other agents (subcontractors, partners, regulators, etc.) and benefit from interoperability to do so. For this reason, data sharing is not just about making the data legally available, but also capturing and processing data in such a manner that it is useful to the largest possible extent to fulfill the needs of all the parties that can use it. In other words, successful data sharing requires that the dataset is of high quality for all the parties involved.

As we have just shown, these two dimensions of scope (the capacity to link several datasets and the capacity to use a single dataset for several purposes) render it valuable because they increase the possibility of making good descriptions, explanations, predictions and prescriptions. But in addition to increasing the value of data by allowing more value-creating opportunities, being able to use a single dataset for several purposes also generates economies of scope. The latter term refers to the reduction of average cost that occurs in a firm as a consequence of diversifying its production. In the case of data, the same competences and investments needed to collect the data can create more value if different uses of it can be found. Indeed, the life cycle of the data can be described as a seven-stage cycle (Chignard & Benyayer, 2015):

1. Creation and collection
2. Transportation (through networks, captors, software, etc.)
3. Storage and security

4. Preparation and qualification (especially cleaning up a database)
5. Analysis
6. Visualization
7. Destruction

Economies of scope arise because a single investment in phases 1 to 3 (and even phase 4, if the end-uses of the data are similar) can result in a variety of uses in phases 5 and 6. Companies that are aware of this and participate in the first tier of the value chain of data (creation and collection) think strategically in terms of the valorization of data by designing data collection in a manner that allows for multiple uses.

Just as economies of scope in physical capital had been one of the main drivers of productivity that allowed for the rise of the corporation in the United States from the late XIXth century on (Chandler, 1993), as the production of goods and services becomes 'datafied', one might expect enormous productivity gains to be made from data-based economies of scope.

2 Legal strategies for data appropriation

It has become a common place to take as a point of departure in approaching data from an economic perspective the idea that data is a non-rival good (A. Lambrecht & Tucker, 2015; Sokol & Comerford, 2016), which means that its use by an agent does not impede the use of the same data by another. This contrasts with rival goods such as a bike, which can only be used by one agent simultaneously, or a sandwich, which can only be eaten integrally by one person. Although technically true, this assertion can be misleading, as it might suggest that an agent that uses a dataset cannot exclude other agents from using it, just like a person that breaths air cannot exclude another person from breathing at the same time. It has to be kept in mind that appropriation is not only the result of the technical properties of data, but also of the legal framework that determines the different ways in which data can be appropriated.

Many organizations build data-based business models with the intention of having at disposal (either through internal means or by recurring to third parties) large, good quality and largely scoped datasets, which allows for more value creation, and recur to legal tools to capture that value through the appropriation of data. The goal of this section is to briefly describe those legal tools.

2.1 Legal regimes for data: an ongoing debate

The first thing that should be noted is that there is not such a thing as a property regime for data. Data itself is not considered as an object of appropriation by any legislation. Natural persons (i.e. individuals) do not have property rights over the data they produce or that refers to them (personal data). Neither are legal persons automatically granted property rights over the data they produce or collect. In other words, from a legal perspective data is not a good. What legal regimes should apply to data is currently an open debate. Benabou and Rochfeld (2015) identify four streams in this ongoing discussion between scholars and legislators.

The first one is the realistic approach. According to this view, individuals should have property rights over their personal data. This would imply individuals having the legal authority to decide on the use of the data that refers to them (including the possibility of selling it) although limitations in terms of the collection, treatment and trade of the data may be set. The second one considers data as *res nullius*, which means that data would not belong to anybody until it is collected. In this case, contrary to the realistic approach, the link between individuals and data would be legally ruled out. The third approach, similar to the first one, considers individuals in to be owners of the data, but under an intellectual property (IP) regime, since individuals would be considered to be creators of the data. The fourth one considers data as a common. According to this approach, property over data should not be attributed neither to natural persons on individual bases nor to the legal persons that collect or produce the data, but to a community of natural or legal persons that produced it and that would manage it as a shared resource[5]. Each of these approaches has enormous implications. Although addressing them falls outside of the scope of this paper, the reader should bear in mind that, as the legislator catches up with the 'datification' of society, adopting any of these views would profoundly affect data-based business models.

On that note, it is important to point out that in the recent General Data Protection Regulation (GDPR) adopted in April 2016 (and enforceable from 25 May 2018) the European Union has discarded the proprietary approaches to data (first and third approaches mentioned above). Nevertheless, the fact that natural and legal persons cannot own data does not mean they cannot *appropriate* it, which in many cases leads to de facto propertization (Cohen, 2017). In the next section we will briefly describe the legal tools that are commonly employed to appropriate data.

2.2 Legal tools for the appropriation of data

Property can be thought as a 'bundle of rights' (Commons, 1893) that links the holder of rights to other natural or legal persons around a thing. Following Schlager and Ostrom's (1992) categorization, we can decompose the bundle of rights over a thing into five rights: access, withdrawal, management, exclusion (impeding a third-party from accessing or using the thing) and alienation (giving it to a third-party). In that sense, although, as said in the previous-section, there is not a property regime for data, there are nonetheless legal means through which these rights can be exerted and lead to de facto propertization. Two types of legal strategies exist to do it: IP-based ones and those that combine general terms of use with trade secret.

2.2.1 IP-based strategies: copyright, patents and sui generis database right

IP-based appropriation strategies are those that rely on the intellectual property regime to obtain one or many of the above-mentioned rights that constitute property. In most countries of the world, including the United States and European Union member states, databases can be protected through copyright. As copyright is supposed to be a legal tool to protect

[5] For example, Merzeau's (2013) proposition of a system of "identity commons" for personal data and Carballa Smichowski's (2016) idea of a regime of data commons under reciprocity licenses for the data generated in sharing economy platforms. For a broader discussion on possible legal regimes for data (including data commons) see Peugeot (2014), Benabou & Rochfeld (2015) and Anciaux & Farchy (2015).

creative human effort, the latter has to be proved in order to obtain the protection. The way in which the information is arranged in the database can be considered to represent a creative human effort, justifying so a protection that can last up to 70 years from the date of creation or publication. Moreover, substantial investment to constitute the database has to be proven in order to protect it under copyright. In the In the European Union there is another mean of protecting a database: *sui generis* database right. This right, introduced in the Database Directive of 1996, allows protecting non-original databases for 15 years provided that there has been quantitatively and/or qualitatively substantial investment either in obtaining, verifying or presenting the contents of the database. Moreover, if there is a substantial change to the contents of the database the 15 year period starts again. It is very important to point out that, in the case of copyright over databases in the European Union, “full copyright protection does not apply to the contents of the database ("the data") but only to its structure” (Duch-Brown, Martens, & Mueller-Langer, 2017, p. 13).

Another more indirect way through which data can be appropriated by the means of intellectual property law is copyrighting or patenting the software linked to the data. This strategy is mostly used to exclude third parties from accessing data that may or may be not protected when the access requires technical interoperability with the software in question. Even when the data is not protected, a firm can use its copyright or patent over the software that has to be put in touch with third parties vendors (for example, an API) or would-be competitors to legally excluding them from using their software, making it impossible so to access the data through it.

2.2.2 The terms of use and trade secret combo for de facto propertization of data

We have seen that although there is not a property regime for data, databases can be directly protected using copyright and, in the case of the European Union, also by recurring to sui generis database right. Nevertheless, firms with a data-based business model rarely employ this strategy. Instead they tend to appropriate data by recurring to trade secret. The main advantage of trade secret is that, contrary to copyright or patents, it provides a perpetual monopoly as long as a third party does not duplicate it by its own means. That is the reason why Coca Cola never patented the formula of its most famous soda and continues to rely on trade secret after more than a century of commercializing it.

Another advantage of not protecting databases with IP is that, as it has been established in European jurisprudence after the Ryanair vs. PR Aviation case, copyright and sui generis database protection laws include exceptions that allow third parties to obtain part of the database its producer displays online through web scrapping. On the contrary, if the data is not protected the exceptions do not apply and the producer of the database can legally enforce its right of excluding third parties from accessing and using its database by simply specifying this prohibition in the general terms of use of the website (Lambrecht, 2015).

Finally, and perhaps more importantly in the case of data, trade secret allows firms to keep the content of the data they possess as a secret, as one of the conditions for being protected by trade secret is precisely that the information must not be unveiled. The other conditions are that the fact that the data remains secret must give it commercial value and that “it must have been subject to reasonable steps by the rightful holder of the information to keep it secret (e.g., through confidentiality agreements)” (WIPO, 2017).

It is precisely for the sake of reasonableness that the legal strategy of data appropriation based on trade secret is generally combined with general terms of use. The latter give free rein to the data collector to dispose of the personal data generated by the users of a platform. When individuals access webpages or download apps they have to consent to the general terms of use. These terms of use are a contract that typically gives the owner of the webpage or app contentment to use the information generated or its transfer, even if (paradoxically) the question of who is the legal owner of the data has not been answered beforehand (Benabou & Rochfeld, 2015). In the case of apps the terms of use generally include users giving consent to access other information located in the phone. The consent of the person involved included in the terms of use is a key legal tool on which most data-based business models are built because it is both one of the foundations that legally legitimize the treatment of personal data (Rochfeld, 2017)[6] and trade secret (Cohen, 2017; Duch-Brown et al., 2017). As user enrollment is typically near-automatic and consent to the (habitually draconian) terms of use of platforms is mandatory to be able to use them, the terms of use/trade secret combo is the most effective legal strategy for firms to be able to de facto propertize data.

Before closing this section let us point out that although firms can effectively appropriate data by recurring to the legal strategies we have briefly described, this does not mean that they can use it in any imaginable way. Indeed, regardless of the legal strategy employed to appropriate data, data holders need to comply with other rights such as privacy law, right to forget or freedom of expression. The exertion of these rights can limit the possible uses of data and therefore constrain the design of data-based business models. It is interesting to point out that many of these protections rely on the category of "personal data", but "the rapid evolution of data collection and analysis technology may create ambiguous borderline cases in the definition of personal data" (Duch-Brown et al., 2017, p. 16), a concept which boundaries are not clearly defined in legal terms (GIGREF, 2015). For example, it is yet unclear whether data generated by connected devices in a person's house (e.g. energy consumption or temperature readings) or data relating many people such as user reviews in platforms are to be considered personal data. With the increasing importance the internet of things and machine to machine communication will probably have in the future, the jurisprudence to be set around the boundaries of the concept of personal data will certainly have a major impact on the evolution of data-based business models.

We can conclude that, given the wide margin of maneuver the current legal framework gives firms to propertisize data, value creation in closed-data-based business models increasingly depends on the ability of firms to collect data by either attracting users or obtaining data from third parties. In the next section we will describe the different ways in which, relying on the legal strategies presented in this section, organizations with closed-urban-data-based business model obtain data and create and capture value from it.

[6] In the case of the new European GDPR the necessity of treati ng data for the purpose of the legitimate interests of the responsible of the treatment is considered a legitimate reason to treat data. A business model could be invoked as a legitimate interest (Rochfeld, 2017), which opens the door to a stronger appropriation of data by firms in spite of the additional protections the GDPR introduced.

3 A typology of closed-urban-data-based business models

We have seen in Section 1 that the value that can be created from a dataset depends on its size, quality and scope, and that its depreciation over time, although possible, is not a general rule. In Section 2 we have shown that the legal regimes that should be applied to data are being debated and that, regardless of the differences between legal frameworks across countries, firms have a lot of room to appropriate data. In this section we will offer a typology of closed-urban-data-based business models. These business models will illustrate the different ways in which organizations leverage the determinants of the value of data describe in Section 1 and, through the legal tools described in Section 2, capture it.

We will divide the section in two subsections. In the first one we will define the concepts of urban data, business model and data-based business models in order to arrive to a definition of what will be the object of study of the following subsection: closed-urban-data-based business models. The second subsection will offer a typology that will describe the different closed-urban-data-based business models in terms of how data is obtained, how value is created from it and how it is captured.

3.1 Defining urban-data-based business models

The concept of business model, although widespread in the fields of management and business, is relatively new to economics. There is not, however, a canonical definition of it in any of these fields. Nevertheless, some features are recurrent in the literature on business models (Zott, Amit, Massa, 2011). First, it represents a holistic view of the company. Second, although it focuses on a specific firm, it also takes into account its environment (suppliers, clients, regulations, etc.). Third, it incorporates the analysis of key activities and the resources employed to carry them on. We will take Harracá's definition of a business model as a description of "the distinctive and fundamental principles and mechanisms by which an organization deploys a strategy to create, sell, and use values (of use and change), in order to fulfill its primary goals" (Harracá, 2017, p. 9). Let us point out that the primary goal of the organization is not necessarily profit making, even when profit exists and constitutes a necessary part of its strategy.

Having provided a definition of a business model, we need now to narrow down our object of analysis to distinguish data-based business models from other business models that employ data. Indeed, every organization produces and collects data to fulfill its primary goal. Hospitals store clinical data about their patients and create statistics for management purposes. Virtually every firm keeps records of its sales, expenditures and so on. Nevertheless, only some organizations rely primarily on data to create and/or capture value. We will define data-based business models as those business models in which the obtainment (be it from internal or external sources, or both) of large datasets and their exploitation are at the core of their value creation and/or value capture strategy of an organization. Following this definition, a supermarket that collects large datasets about its clients' purchases falls out of the scope of data-based business models. Although many supermarkets do use these datasets to find patterns in consumer behavior and adapt their marketing strategy consequently, it cannot be said this data analysis is at the core of their value creation and/or value capture strategies. Let us point out that the degree to which data acquisition and analysis can be considered to be at the core of the value creation and value capture process of an organization is inevitably subject to interpretation.

Nevertheless, we consider this criterion distinctive enough to separate an array of organizations for which data is at the core of their business model.

The next step in characterizing our object of study is defining what we mean by urban data. Despite the popularity of the term, definitions are lacking. It is not our intention to provide a definition that will contribute to the literature, but rather one that will allow us to draw a general contour of the type of data the organizations whose business model we include in our study rely on to develop their business models. We consider urban data to be the data that fulfills two conditions. First, it provides information about the "political, social, and economic conduits" (Swyngedouw & Swyngedouw, 2004) of a city or metropolis. Second, that information loses explanatory power outside of the scope of the city or metropolis it refers to. Therefore, it is not sufficient for data to have been generated within a city or refer to it to be considered to be urban data; it has to provide information about a city (in the broadest sense) that is *proper to that city*. Consequently, data-based business models such as those of generalist social networks or search engines, although certainly reliant on data to create and capture value, are excluded from our analysis for not relying on what we consider *urban* data.

Having provided definitions of the concepts of urban data and (data-based) business models, we can now narrow down the concept of business model we will employ to the basic elements that relate to the role urban data plays in closed-urban-data-based business models. Harraca's definition of a business model is deliberately comprehensive. When using the concept to understand an organization or a type of organization, choices have to be made in terms of the level of detail and the dimensions of the analysis. Business canvases are a classic way to frame this at the level of a company. Since our intention is to create a typology based on the role played by data in value creation and value capture (and not an in-depth analysis at the firm level), we have decided to frame our description of business models in the following manner. For each type of urban-data-based business model we will focus on three processes:

1. **The obtainment of data**: how the data is produced, collected and/or bought and from what actors

2. **Data-based value creation for the end user**: what competences are employed in conjunction with data to offer a valuable service to end users and what makes it valuable

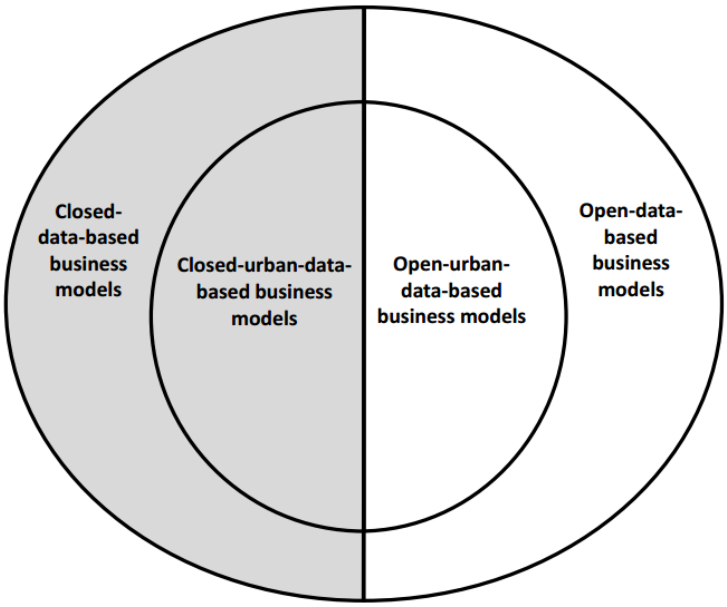
3. **Data-based value capture**: how the organization captures value from the data-based value creation process

For each typology, we will stress the role played by the different determinants of the value of data examined in Section 1 (size, quality and scope) in the value creation and value capture strategies. Although we will not repeat it along this section, the reader should bear in mind that the legal mechanisms that allow for the appropriation of data on which the closed-urban-data-based business models we will describe rely on are those presented in Section 2. As the reader will appreciate, the typology we will offer in this section is not built around fields of activity (e.g. energy, transportation, waste management, etc.) but rather on the role played by data in the business model.

Let us point out that although we will study business model that are centered on urban data, this does not mean that the four families of business models we will describe are the only ways in which urban data can enter a business model. For example, the mobile operator Orange sells anonymized datasets about users' geolocalization although their business model is not centered (yet?) around this urban data but rather on selling subscriptions to telecommunication services. Inversely, the urban-data-based business models we will analyze in this article could be applied to non-urban data. For example, data brokers sell all kinds of non-urban data. In that sense, in this article we will study all the data-based business models that are *compatible* with urban data. Some might be more suited for urban data, while other less.

The reader should also bear in mind that there are many non-urban-data-based business models such as those of social networks that are not analyzed in this article. Urban-data-based business models are to be considered a category within data-based business models. This article focuses precisely on this category. Moreover, both urban and non-urban data-based business models can rely on either open and closed data, as shown in Figure 1.

Figure 1: A classification of data-based business models



3.2 A typology of closed-urban-data-based business models

We have identified four patterns in the way organizations produce and capture value based on the appropriation of urban data, namely aggregated-data-based service providers, individual-data-based service providers, trust-based algorithmic coordinating platforms and transactional intermediaries. Table 1 synthetizes these business models.

3.2.1 Aggregated-data-based service providers

Aggregated-data-based service providers rely on the collection of data originated from multiple sources to offer a service to an end user based on its sorting, categorization and analysis. Examples of these types of service providers include Waze, CityMapper, Yelp, Trip Advisor, Foursquare, Easy Transport, Zen Bus and Mapster.

The data is obtained from two sources. The first one is volunteered (tagging, pointing out a closed route, reviewing a restaurant, etc.) and observed data (geolocalization, time spent walking to a station, etc.) from users. The second one is the integration of third party data, usually from open datasets (e.g. location of bus stops from the transportation authority) or web scrapping (e.g. addresses and contact information of restaurants, rating of a hotel in another website). The user experience of the platform is designed to engage as much as possible users in contributing to enlarging the dataset. This, combined with the data obtained from third parties, is what makes the first brick of value in this urban-data-based business model family: size.

Indeed, the service to be offered depends on having a lot of data because a large enough dataset collected from a multitude of individuals and other sources is crucial for the platform to create value for end users through an accurate description (where are the best Thai restaurants located in my neighborhood, where will traffic jams take place after 5pm) and prescription (which Thai restaurants should I trust based on reviews and my budget, which bus line should I take to get home before 3pm, etc.). Moreover, the data-based value creation comes from size also because of the existence of network effects. Indeed, the more users the platform has, the more valuable it becomes to each individual user, because having more users translates into a better service (more trusty ratings, more information available, etc.) The platform's main competence lies in aggregating, sorting and analyzing datasets to provide useful information to users and answer their inquiries. In that manner, we can say that the value creation process of aggregated-data-based service providers can be summed up as "from large aggregated datasets to individual solutions".

In terms of value capture, the most common strategies of these business models are selling targeted advertisement (including selling more visibility in the platform to companies) and selling datasets to third parties. Because most users are individuals that are not willing to pay much for the service, direct monetization does not usually take place. The data collected not only serves to provide the service that assures in turn the continual collection of user-generated data, but also as a commodity. As we have seen before, the size of these thematic datasets is what makes them valuable. When identifying users across other platforms is possible (in that manner the possibility of logging in through Google, Facebook or Twitter accounts plays an important role), the dataset gains in the scope that comes from combining it with third party data.

3.2.2 Individual-data-based service providers

Individual-data-based service providers offer a service (be it online or offline) to an end user that relies mainly on that user's generated data. Examples include Enevo, Terradona, Prediwaste, Eugène by Uzer, Strava and most of the business models of connected devices.

The sources of data are users' volunteered (e.g. what kind of bike I use, what product I scanned before recycling it) and observed (e.g. what route I took based on GPS location, how many times I scanned milk cartons in a week, how often do I turn on the heating and

at what temperature) data. The data can be complemented with data obtained from third parties (e.g. the map of the city, information about the composition of the product, what is the temperature outside of the house).

The key aspect of the value creation process, which is generally taken into account in the design of the service, is the quality of the data. Because, contrary to what happens in aggregated-data-based services, the service to be provided to a user relies mainly on the data obtained from the user in question and not in having reached a critical mass of data that makes the service attractive enough for each individual user, the key aspect to take into account to create value for the end user is the quality of the data. Service providers design the data collection process in such a manner that they can guarantee that the data is accurate enough, objectively interpretable for the purpose intended, that it arrives within a reasonable delay, etc. Value creation relies on the service provider being able to obtain high quality data about a single user and analyzing it in such a manner that translates into a valuable service for him/her. Those services typically consist in description (historical records of my physical activity, what recyclable products I buy more regularly, at what time of the day I tend to turn on the heating) and prescription (how should I train myself to reach my goals, in which bin should I place carton milks, when should I start heating the house for it to be warm when I get home).

Value capture can take many overlapping forms in individual-data-based services business models. One of them is subscriptions, which are more common when the end user is professional (e.g. Prediwaste) or a government entity (e.g. Terradona). When end users are not professionals, selling premium services is a typical option. Another option is offering data-targeted complementary services. Because the data collected is of high quality and centered on the continuous use of a user, it allows the service provider to have precise insights about his/her habits. The service provider can find behavioral patterns and predict the needs of the user in order to sell complementary services (e.g. Qurrent's e-shop, Amazon's shopping suggestions based on data collected from its smart devices). In this manner, we can say that the logic of the business model of individual-data-based service providers is the inverse of that of aggregated-data-based services, since it can be summarized as "from individual solutions to large aggregated datasets".

3.2.3 Trust-based algorithmic coordination platforms

Trust-based algorithmic coordination (TBAC) platforms, usually referred to as "collaborative economy" or "sharing economy" platforms, are platforms that allow for the production and/or distribution of goods or services (either in exchange for money or not) based on network interactions between mainly private individuals channeled through digital platforms and where most of the users' participation is not driven by wage relationships. TBAC platforms have the following characteristics:

a) They set the conditions of network exchange and/or production (including labor conditions and value distribution) through algorithmic coordination

b) They create a digital support that not only serves as a virtual matchmaking space that allows network interactions to take place offline *stricto sensu*, but also incorporates mechanisms such as reputation systems and third-party identity verification that make these interactions viable.

Examples of these platforms include Uber, Airbnb, TaskRabbit, Deliveroo, BlaBlaCar and La Ruche Qui Dit Oui.

In these platforms the data is mainly originated by users, both in the form of volunteered (name, skills, car brand, etc.) and of observed data (the rate of cancelled trips, when do people travel the most to Portugal).

The data is used in a first stage as the fuel of algorithmic coordination, the latter representing the value added by the platform. Data on inventories, billing, payment, geolocalization, within-platform interaction between individuals and more are collected by the platform. The platform offers an automated response to the situation described by these different pieces of data in order to coordinate production and/or exchange between individuals both in a 'soft' manner (by giving them incentives) and in a 'hard' manner (by giving them instructions). It is important to stress that algorithmic coordination goes beyond matchmaking. In TBAC platforms, algorithms not only put in touch two sides of the market (drivers and riders, hosts and travelers, etc.): they embed the conditions of exchange and/or production between individuals so that a series of tasks can be performed in a particular manner. While matchmaking platforms such as e-commerce websites limit themselves to offering a digital environment that facilitates matchmaking between two sides of a market, TBAC platforms also coordinate the performance of the tasks that take place after the matchmaking and/or direct the matchmaking: Uber sets the route and makes sure drivers stick to it; Deliveroo coordinates the logistics between deliverers and restaurants in real time; Airbnb gives more visibility to listings that comply with certain criteria (instant booking setting, good reputation, activity rate, cancellation rate, etc.) and suggest revenue-maximizing prices to hosts, which exceeds mere matchmaking. Algorithmic coordination of networked interactions is therefore a special type of data-fueled prescription.

Data plays a second role in terms of value creation, as it is a source of improvement of the algorithmic coordination service the platform provides. Indeed, "as a platform gains more users, it can collect more user data, leading to better insights into consumers and their needs, which can be used to improve quality, attracting even more users" (Sokol & Comerford, 2016). As mentioned in subsection 1.1., when machine learning is involved, this process becomes more important because algorithms can 'learn' faster if they are contrasted with more data.

In terms of value capture, TBAC platforms tend to use their gatekeeper position to charge users a commission for each transaction carried on within the platform. Charging a monthly subscription to use the platform is another option, although it is less common. Finally, another value capture strategy deployed by TBAC platforms is using the insights gained through the analysis of the data created by users to expand into neighboring markets. For example, the lodging platform Airbnb started recently an offer called Trips, entering so the business of travel agencies. Similarly, Uber used the knowledge it gained from analyzing user data of its ride-hailing platform to start Uber Eats, a food delivery service.

3.2.4 Transactional intermediaries

Transactional intermediaries are organizations that create and capture value by playing the role of intermediaries in transactions of data. Their intermediation can be distinguished from

that done by the actors that recur to the business models we have analyzed before in that it is neither based on coordinating offline interactions (TBAC platforms) nor on providing a data-based service mediated and supported by the data (aggregated-data-based and individual-data-based service providers). Transactional intermediaries play the role of reselling and/or providing the necessary tools for transactions of data (either monetized or not) to take place. Examples of transactional intermediaries include data brokers such as Acxiom, Nielsen, Experian, Dawex, Equifax and firms such as Enigma or Navitia, which centralize open datasets. Let us point out that although this business model is generally based on closed data, the example of Navitia and Enigma, among others, show that it can also be applied to open data. In that sense, this business model, which can be placed at the center of the inner circle of Figure 1, blurs the open/closed data border we used to structure this article.

Transactional intermediaries obtain data by two means depending on if the data is closed or not. If the data is closed, as in the case of data brokers, they purchase it from other companies, typically digital firms for which the selling of data is part of their value capture strategy. If the data is open, they obtain it through the collection of open data (e.g. Enigma) or accessible data (e.g. web scrapping) or through the pooling of closed data that other organizations decide to open for transactional intermediaries to centralize and manage. In this case, the production of the open pooled dataset follows the same logic that we will describe in subsection 4.4. “Multi-stakeholder open data pooling”. In a nutshell, the rationale behind it is that different organizations holding complementary datasets have an interest in pooling it to increase its size and scope, and therefore its value. The choice of opening it responds to the fact that, by opening the dataset, third parties can enrich it and it can more easily become a standard, which might be important for the firms pooling the data to capture value through other means such as selling services related to the open dataset. The difference between transactional intermediaries using pooled open data and multi-stakeholder open data pooling relies on the fact that the former centralize the management of the dataset, while the latter rely on a multi-stakeholder governance scheme.

Transactional intermediaries create value by leveraging size and scope. Data brokers are a good example. They buy data from different sources, which allows them to both have large datasets about millions of individuals (size) and information about several interconnected aspects of each of them (credit scoring, online shopping, browsing history, etc.), increasing so the scope of the dataset. This, in turn, makes profiling (describing the person) more accurate, and the decisions based on that profiling (prescription and prediction) more solid. Additionally, they sometimes add value also by cleaning and homogenizing the dataset, which increases its quality.

Value capture takes two forms. For closed-data-based transactional intermediaries, value is captured through selling the data. When datasets are sold three pricing methods exist (Chignard & Benyayer, 2015). The first one is the cost of production approach, which consists in setting a price equal to the costs engaged in producing and storing the data (captors, maintaining a network, storage costs, etc.). The second one consists in pricing the data at a price close to what it will make the buyer gain. This is certainly difficult to quantify, but proxy measures such as the number of transactions that took place related to the sold dataset exist. The third method, which takes into account the other two, is market pricing. Buyers and sellers bargain to reach a price located between the value the dataset will generate and its cost of production. The resulting price will depend on the relative power of

the buyer and the seller and on information asymmetries. Indeed, sometimes one of the parties has a better understanding or more information than the other, which helps it reaching a price that benefits it in detriment of the less informed party. For example, when Amazon started it signed a deal with AOL to run the technology behind its e-commerce website that included getting hold of AOL's data. Because AOL failed to understand that that data was very valuable for Amazon to improve the performance of its recommendation engine, it agreed to a price considerably lower than the one it would have charged had it known it. For open-data- based transactional intermediaries, because the openness of the data makes it impossible to sell it, value capture generally happens by selling premium services related to the open datasets. This is the case of Navitia, which offers premium support and analytics.

4 The business models of open data

In Section 1 we have explained what makes data valuable, namely size, quality and scope. In Section 2 we have seen that, although there is not such a thing as a property regime for data, both IP-based and non-IP-based legal strategies to appropriate data exist and their use is widespread. In Section 3 we have shown how organizations create and capture value by recurring to closed-urban-data-based business models that rely on those appropriation strategies. Nevertheless, this does not mean that value can only be created and captured from data if it is appropriated. In this section we will show that there are different business model families based on open data. Let us point out that typology we will offer is not limited to urban data. Urban data can be the object of any of the business models we will present. Moreover, as mentioned before, although all of these business models can be applied to urban data, some are more suited for it than others. Nevertheless, by presenting all of the open-data-based business models (right-sided half of Figure 1), we hope to offer a broader understanding of how relying on open data to develop a business model can shape value creation value capture, and governance choices.

As mentioned above, open-data-based business models imply in some cases a joint governance of the data with other stakeholders and the choice of open licenses. For those reasons, when analyzing the business models of open data we will add to our tryptic reading grid (obtainment of data, data-based value creation and data-based value capture) the governance and licensing dimensions. Table 1 synthetizes these business models.

4.1 Government open data

Government open data is possibly the most well-known business model of open data. The rationale behind this form of open data is that the government decides to open some of the data it holds because it considers it to be a public service and/or a citizen right in terms of access to information and accountability of public institutions.

Government open data is (co)produced by the State bureaucracy (e.g. data on parking tickets, data on the functioning of street lights, etc.) as well as collected by the State from third parties. For example, the State might demand electrical companies to provide information about energy consumption and open that data. In terms of governance, this means that although the State might not be the only producer of the datasets it provides as open data, it centralizes its governance by recurring to its legal authority.

The value creation process is twofold. In a first stage, the value of government open data relies mostly on its quality and size. Because the government has the authority to collect data from some third-parties and can recur to public servants to produce data related to its functions that other actors cannot, it can provide datasets of high quality. Moreover, because the State's function of guarantor of the public interest leads it to obtain information that is comprehensive, government data typically has the size required for its proper exploitation. The second stage of the value creation process around government open data is the most important one. In economic terms, the data opened by the government functions as a public infrastructure: the government finances the production and collection of certain datasets and gives access for free to it to the population. This data is then used for different purposes by all types of actors (private firms, public institutions, civil society, researchers, data journalists, etc.) that will produce value over it. In this sense, we can say along with Benyayer and Chignard (2015) that "data has value only if it circulates". This claim is certainly true in societal terms. As we have seen above, appropriation of urban data is at the core of many business models. It is precisely because the State does not follow a value capture logic and has societal interests in mind that its open-data-based business model is not aimed at capturing value but rather at facilitating value creation to third parties and financing it through taxes.

4.2 For-profit private firm standalone open data

For-profit private firm standalone open data refers to cases in which a single for-profit private firm decides to open a certain dataset it holds. Two complementary goals motivate this decision. The first one is the enlargement (increasing value through size) and/or enrichment (increasing the value through quality) of the opened dataset. The second one is creating business opportunities related to the enlarged and improved resulting dataset.

The original dataset is produced in a first stage through intra-firm data collection (e.g. crowdsourced data about the location of bus stops through a platform owned by the firm) and/or through intra-firm production (a firm that does clinical data and stores the results). Once opened, the dataset can be enlarged and enriched by third parties.

The value of the dataset comes, in a first place, from the quality the firm can bring because of the expertise or specialized knowledge it put into its production. In a second stage the dataset gets more valuable precisely because it has been opened. As said above, the larger the dataset, and the higher its quality, the more it can be learnt from it. Opening data increases the possibility of third parties augmenting its size and quality.

The governance of private firm standalone data is limited to the firm simply opening a certain baseline dataset and letting others enrich it and eventually modify it. No collective governance takes place in this case. Licenses, although having in common being open in the sense of allowing accessing, using and sharing the dataset (Open Data Institute, 2017b), vary regarding the revenue model of the firm that has opened the data.

When the freemium model is chosen, a core dataset is kept open but an extended one (usually useful for commercial purposes) is sold. In these cases, licenses distinguish commercial from non-commercial use of data. Another possible revenue strategy consists in developing data-related services around the open data such as paid data visualization

toolkits to analyze the open dataset. An example of this is HERE's Open Location platform, which provides open cartographical data and offers cartographical licenses for firms that want to use their maps. Finally, a third way of monetizing for-profit private firm standalone open data is creating future business opportunities by gaining knowledge and expertise through the study of the open dataset. This helps companies to develop capabilities they can monetize in the future by creating new products, offering new services or gaining competences to develop new businesses. This is the case of the agribusiness multinational Syngenta, which opens some of its R&D and agriculture data. In words of one of its Science & Technology Fellows, Derek Scuffell, "... if we don't have an open data approach then Syngenta will miss out on opportunities – those opportunities could be in new technologies or new research" (Open Data Institute, 2017a).

4.3 Nonprofit standalone open data

Nonprofit standalone open data refers to open datasets created (as opposed to already existing opened datasets) by a single organization or individual, usually motivated by contributing to a cause and not for commercial purposes. Examples of this are the datasets provided by Wikileaks, the datasets produced by Inside Airbnb through web scrapping of Airbnb's website, the data on different topics published by the famous blogger Nate Silver in his blog FiveThirtyEight or open-data-based journalism.

These datasets are produced by an organization or individual using their expertise in a certain domain to create new open data over already-existing open or accessible datasets. Because a single organization or individual produces it, the governance is centralized: the organization/individual is in charge of the production of the open dataset and of its management. Because these datasets do not have commercial motivations, they usually have permissive open licenses allowing for commercial use and only demanding recognition of authorship. The value created comes mainly from the quality the creator brings with its expertise. As these datasets are generally built on open or accessible data, the value added they contain is rooted in the quality the creator of the database brings to build it.

In terms of revenue model, because these are generally not profit-motivated projects, donations are the most common way of financing the time dedication required to produce these datasets, although the extent of voluntary labor must not be overlooked. Finally, another (sometimes unplanned) revenue strategy consists in gaining notoriety through the production of open data, which might result in future business opportunities. Here again, openness enhances value capture: the more the data can easily circulate and be used, the more notoriety the creator will get, and the more opportunities he/she will have of obtaining business opportunities related to the expertise used to produce that data. The fact that open licenses over these kinds of datasets tend to require the recognition of authorship is consistent with this revenue strategy.

4.4 Multi-stakeholder open data pooling

Multi-stakeholder open data pooling consists in at least two agents of any sort (private firms, governments, NGOs, collectives of citizens, private individuals, etc.) creating a dataset through pooling data they already own or they have created and applying an open license to it. For actors other than profit-oriented firms, the logic of this data pooling consists in being able to fulfill better their noncommercial mission. For example, the regional governments of

Bretagne and Pays de Loire in France contribute to an open data pool about energy called PRIDE to be able to design better policy using the more accurate and exhaustive information that comes from an enlarged and enriched dataset that open data allows for. Another example of this logic is Transdev's Catalogue, a platform of pooled open transportation data. For profit-oriented firms, there are many commercial motivations, notably creating a related business, good publicity (when their opened data helps to tackle a societal issue), gaining expertise and increasing interoperability.

The data is produced, in a first stage, as a result of stakeholders pooling complementary data. In a second stage, some stakeholders might provide their expertise to increase the quality of the data. In a third stage, third parties not belonging to the stakeholders that created the original pooled dataset can contribute to the dataset. Accordingly, governance is multi-partner and shared between the different stakeholders that created the dataset in the first stage, although most of the times there is a leading entity in charge of coordinating the pooling and maintenance of the data. When a public institution is a stakeholder, it usually plays the role of the leader.

In terms of value creation, in the first stage these business models leverage on the openness of data to increase its value mainly through enlarging its scope. By allowing the combination of previously unlinked datasets (which usually implies setting a standardized format) the value of data increases for all the actors involved, including third parties that can access it. If the dataset is large and/or attractive enough, third parties would have incentives to make contributions in that same format and therefore increase its value by augmenting its size. An example of this is the Open, Improved Settlement Data project carried on by CIESIN, Facebook and the World Bank. Facebook has shared commercially-purchased satellite imagery data with CIESIN, which in turn had census data of the places to which the satellite images correspond. In addition, Facebook has shared "state of the art computer visioning techniques" (i.e. an increase in the quality of data) with CIESIN to identify buildings. The pooled dataset, which has been opened, helps understanding how human settlements are distributed across landscape. The resulting scope coming from linking these two datasets makes the resulting pooled dataset valuable in that it allows for many different applications (notably research, humanitarian planning and crisis response) that could be carried on had the datasets remained in silos. For Facebook, this information is valuable because it helps it develop technologies to improve connectivity, a business line in which the company is engaged.

In terms of value capture, several compatible strategies exist. Because this open-data-based business model family relies on stakeholders pooling data, financial contributions from stakeholders is a traditional source of revenue. When the State is involved, public funding, which follows the logic of government open data, generally takes place. While financial contributions might seem like an unjustified expense to profit-oriented firms at first, they make sense commercially for different reasons. Moreover, firms can also provide data-related services using an open dataset that, precisely because of its openness, becomes larger. Firms can also use that more valuable open dataset to gain expertise and knowledge, which can result in positive economic returns to all the actors. For example, Transfermuga, an open dataset about transportation in certain regions of the south of France and the north of Spain, allows incumbent transportation providers to provide a better service by linking their data to other stakeholders'. At the same time, it gives regulators a better picture of

transportation in the region and allows for the creation of a platform that tells users what the best itineraries are. It also creates business opportunities for start-ups. Because the ultimate goal of multi-stakeholder data pooling is to increase the flow and reuse of data by complementary actors, licenses are usually permissive and allow commercial use. The latter is not to be hindered but actually fostered, since it brings value to the open dataset by increasing its use and, eventually, generating contributions to it.

4.5 Commons-based open data crowdsourcing

Commons-based open data crowdsourcing consists in a community crowdsourcing data to create an open data common in order to tackle a societal issue (e.g. crowdsource environmental data) or to provide a dataset that it wants to keep open to the benefit of all. Examples of this include OpenStreet Maps, Open Food Facts, Digital Matatus, Transport For Cairo, Accra Mobile, OpenSideWalks or the Barcelona citizen sensing project Making Sense.

As the name indicates, the data is obtained through crowdsourcing by individuals. Nonetheless, as it is the case with OpenStreet Maps, the dataset is usually enriched by adding other open data from third parties. The main source of value of this type of data is size. Crowdsourcing is a way of producing data that relies on small contributions by a multitude of actors (typically individuals) to obtain a dataset that would have been costly and lengthy to produce by a firm or by the State. Crowdsourcing is also a form of producing data that contributes to its value in terms of quality. Because the data is open and the data is produced through crowdsourcing, individuals can not only contribute by adding data, but also by correcting it. The fact that there is a multitude of individuals with first-hand knowledge about and quick access the data makes the correction and improvement of the dataset more effective than if that task was centralized. For example, people living in a street can signal that it will be closed to transit for a month on OpenStreet Map faster than if an organization intended to do it in a centralized manner.

Commons-based data crowdsourcing datasets are governed following a community governance scheme where contributors have a say in the development of the project. It is for this reason that we can speak of data commons. When these projects attain a certain critical mass, they are generally governed through a foundation that serves as the legal environment to develop community governance as a tool to manage revenue sources.

Regarding value capture, donations are a common source of revenue making. Another common source of financing is public funding, usually in the form of research grants, as in the case of the citizen sensing project Making Sense. Because the primary goal of commons-based open data crowdsourcing is to tackle a societal issue or to provide an open dataset that benefits the general population, it is common and logical for the State to contribute to these projects as they fulfill some of its missions. Another (not so usual) revenue capture strategy is the selling of products related to the common open dataset. For example, Open Street Maps sells merchandising with its logo (shirts, jackets, mugs, etc.). It also gets commissions from the selling of products related to cartographical data collection (GPS, mapping books, batteries, mobile phones, etc.) from certain retailers with which they have passed contracts. This example is interesting not only because of the relevance OpenStreetMap has among common-based data crowdsourcing projects, but also because of what we can learn in terms of the design of business models for data commons from it.

While the purchase of merchandising is closer to voluntary contributions in that what motivates the sale is reciprocity (although the actual purchase of an object with the logo fosters contributions because people feel they are not just giving away money), the purchasing of a GPS, for example, shows that commons-based crowdsourced data commons can create business opportunities for third parties. The logic behind this business opportunity creation is related to the domain to which the project refers to, and it might therefore take place more easily in some projects than in others. Harass Map, “an advocacy, prevention, and response tool that uses crowdsourced data to map incidents of sexual harassment in Egypt” (Young, 2014), for example, has no regular revenue sources. One could imagine that selling products related to sexual harassment (pepper sprays, for example) is more difficult than selling products related to digital cartography (GPS, apps, maps, mobile phones, batteries, etc.).

Table 1: A synthetic overview of closed-urban-data-based and open-data-based business model families

CATEGORY	BUSINESS MODEL	DESCRIPTION	OBTAINMENT OF DATA	DATA-BASED VALUE CREATION FOR THE END USER	VALUE CAPTURE	LEGAL STATUS OF THE DATA	GOVERNANCE OF THE DATA
CLOSED-URBAN-DATA-BASED	**Aggregated-data-based service providers**	Aggregated-data-based service providers rely on the collection of data originated from multiple sources to offer a service to an end user based on its sorting, categorization and analysis **Ex**: Waze, CityMapper, Yelp, Trip Advisor, Easy Transport, Zen Bus and Mapster	Volunteered and observed data from users Third party data to complement (usually open data)	The platform creates value by aggregating, sorting and analyzing data in order to offer users useful information and answer their inquiries. In order for that to happen the size of the database (and therefore of the user base) is crucial. The service offered by the platform becomes increasingly valuable for users when the number of users increases because of the presence of network effects.	Targeted advertisement Sell of data to third parties	Closed data protected by business secret	Centralized by the service provider
	Individual-data-based service providers	Individual-data-based service providers offer a service (be it online or offline) to an end user that relies mainly on that user's generated data **Ex**: Enevo, Terradona, Prediwaste, Eugène by Uzer, Strava and most of the business models of connected devices	Volunteered and observed data from users Third party data to complement	Descriptions and prescriptions that depend mostly on the quality of the data	Subscriptions Premium services Data-targeted complementary services	Closed data protected by business secret	Centralized by the service provider

	Trust-based-algorithmic coordination platforms	Trust-based algorithmic coordination platforms, usually referred to as "collaborative economy" or "sharing economy" platforms, are platforms that allow for the production and/or distribution of goods or services based on network interactions between mainly private individuals channeled through digital platforms **Ex**: Uber, Airbnb, TaskRabbit, Deliveroo, BlaBlaCar and La Ruche Qui Dit Oui	Volunteered and observed data from users	On a first level: data as the fuel of algorithmic coordination, which is the value added the platform creates On a second level: data as a source of improvement of the algorithmic coordination service the platform provides	Transaction fees Monthly subscriptions (less common)	Closed data protected by business secret	Centralized by the service provider
	Transactional intermediaries	Transactional intermediaries intermediate transactions of data by reselling data and/or offering to third parties the necessary tools to complete a transaction of data **Ex**: Acxiom, Nielsen, Experian, Dawex, Equifax, Enigma and Navitia	When based on closed data, the data is generally bought from third parties, typically from digital firms	Value creation depends mainly on the size and the scope that aggregating and centralizing data brings. Sometimes it also relies on increasing the quality of the data though data cleaning and homogenization	When based on closed data, value capture takes place through reselling the data to clients	Closed data protected by business secret	Centralized by the transactional intermediary

OPEN-DATA-BASED			When based on open data, the data is obtained through integrating open data from third parties, scrapping data and the pooling of data by third parties		When based on open data, value capture takes place mainly through selling related data-based services	Permissive open licenses	
	Government open data	Data that the government decides to open some of the data it holds because it considers it a public utility and/or a citizen right in terms of access to information and accountability of public institutions **Ex**: data.gouv.fr, data.gov	(Co)produced by the State bureaucracy The States collects it from third parties using its legal authority	In a first stage: valuable data because of the size and quality the State can assure In a second stage: re-valorization of the data through its use by third-parties	No value capture strategy. The production and procurement of the data is financed through taxation	Permissive open licenses	Centralized by the State
	For-profit private firm standalone open data	For-profit private firm standalone open data refers to cases in which a single private firm decides to open a certain dataset it owns **Ex**: HERE's Open Location platform, Properati's open real-state-related datasets, Syngenta's agricultural and R&D open datasets	In a first stage: intra-firm data collection and/ or intra-firm data production In a second stage: enrichment of the dataset by third parties	In a first stage: the value of the original dataset comes from the quality that the knowledge and expertise of the firm that produced it can bring In a second stage: openness allow thid parties to enlarge the dataset and improve its quality	Freemium model Data-related services around the open dataset Creating business opportunities through the knowledge and	Open licenses with different options in terms of commercial clauses depending on the revenue model	Centralized by the firm although limited to the creation of the original dataset

					expertise that can be built on the open dataset		
	Nonprofit standalone open data	Nonprofit standalone open data refers to open datasets created (as opposed to already existing opened datasets) by a single organization or individual, usually motivated by contributing to a cause and not for commercial purposes **Ex**: Wikileaks, Inside Airbnb, Five Thirty Eight	Produced by an organization or individual that uses its expertise in a certain domain to create new open data over already-existing open or accessible datasets	Rooted in the quality the creator of the database brings to build it	Donations Voluntary labor Creating business opportunities through incrementing the creator's notoriety	Open licenses allowing for commercial use	Centralized by the firm although limited to the creation of the original dataset
	Multi-stakeholder open data pooling	Multi-stakeholder open data pooling consists in at least two agents of any sort (private firms, governments, NGOs, collectives of citizens, private individuals, etc.) creating a dataset through pooling data they already own or they have created and applying an open license to it **Ex:** Open, Improved Settlement Data project, PRIDE (of Bretagne and Pays de Loire's regional governments open energy data pool), Catalogue	In a first stage: stakeholders pool complementary data In a second stage: some stakeholders might use their expertise to improve the quality of the dataset In a third stage: third parties might	In a first stage: the pooling of complementary data creates value through scope In a second stage: contributions by third parties create value by increasing the size of the dataset	Financial contributions by stakeholders Public funding (if a public institution is a stakeholder) Data-based services based on the open dataset Gainng knowledge and expertise using	Open licenses allowing for commercial use	Multi-partner shared governance by the stakeholders with one of them being the coordinator. If the State is a stakeholder it usually plays the role of coordinator

			contribute to the dataset		the open dataset		
	Commons-based open data crowdsourcing	Commons-based open data crowdsourcing consists in a community crowdsourcing data to create an open data common in order to tackle a societal issue (e.g. crowdsource environmental data) or to provide a dataset that it wants to keep open to the benefit of all **Ex**: OpenStreet Maps, Open Food Facts, OpenSideWalks, Making Sense, Digital Matatus, Transport for Cairo, Accra Mobile	Crowdsourced by individuals Third party data to complement (usually open data)	The value created relies on the size and the quality of the dataset that crowdsourcing can in some cases assure better than centralized methods of creating datasets	Donations Public funding (typically research grants) Sell of products related to the open dataset	Open licenses allowing for commercial use	Governed by the community usually through a foundation

Conclusions

Data is valuable in as much as it allows to describe, explain, predict, and prescribe (Chignard & Benyayer, 2015). But in order to do so, value-creating organizations need datasets to be large, linkable to other datasets, reusable for other purposes and/or of high quality. Organizations that build successful business models around data are those that manage to obtain datasets that leverage on these context-defined characteristics in different degrees depending on the intended use to create and capture value. In most cases, this process requires appropriating data.

Although there is no property regime for data, firms have a wide marge of maneuver to appropriate it by recurring to two legal strategies. The first one consists in protecting databases or the software required to access it with intellectual property rights. The second one, which is more common, is based on obtaining users' consent to appropriate the data they generate through the terms of contract and maintaining an exclusive control over the resulting datasets under the umbrella of trade secret.

By recurring to these strategies, many firms build business models based on de facto propertization of data. In this paper we have studied the principles of those business models when they are based on a particular type of data: urban data. We have defined the latter as data that fulfills two conditions. First, it provides information about the "political, social, and economic conduits" (Swyngedouw & Swyngedouw, 2004) of a city or metropolis. Second, that information loses explanatory power outside of the scope of the city or metropolis it refers to. Based on this definition, we have distinguished four types of closed-urban-data-based business models: aggregated-data-based services providers, individual-data-based service providers, trust-based algorithmic coordination platforms and transactional intermediaries.

The first one refers to platforms such as Waze that rely mostly on user-generated data they complement with third party open data to offer users useful information and answer to queries. Because this service requires a critical mass of data, and because of the existence of network effects, the key of value creation relies on the size of the datasets, which in turn depends on the number of users. Revenue comes generally from targeted advertisement and selling data to third parties. Individual-data-based service providers are firms such as Strava or most providers of services based on connected devices. Their value creation relies on offering a service that relies on high quality data about the user to offer him/her accurate descriptions and useful prescriptions. The data collected can be then used to offer targeted complementary services, although value capture relies in most cases mostly in subscriptions to and premium versions of the main data-based service. Trust-based algorithmic coordination platforms such as Uber or Airbnb base their value creation on user-generated data that feeds the algorithms on which their coordination service relies. On a second level, especially through machine learning, the data collected contributes to improving the service. Value capture relies mostly on transaction fees or subscriptions to the platform. Finally, transactional intermediaries are organization such as data brokers that intermediate transactions of data by reselling it and/or by offering to third parties the necessary tools to complete a transaction of data. Their value creation relies on centralizing large datasets (size) of a wide scope, and sometimes on increasing its quality through data cleaning and homogenization. Value capture depends on selling data in the case of closed-

data-based transactional intermediaries and on providing complementary data-based service in the case of open-data-based transactional intermediaries.

While these four business model families are (with the exception of open-data-based transactional intermediaries) based on appropriating data through the legal strategies explained in Section 2, not all data-based business models follow this trend. On the contrary, some business models create value precisely by opening data. Governments create value for the general public by opening data of high quality they can produce and obtain from third parties by recurring to their authority. Some private firms open certain datasets so that third parties will enlarge it and enrich it and then, leveraging on the improved dataset, they capture value through data-related services, freemium models (a dataset is opened, while other complementary closed ones are sold) and by creating business opportunities derived from the knowledge gained from analyzing the improved open dataset. Some organizations and individuals use their expertise to produce high quality datasets they open and obtain revenue from donations and, more indirectly, from business opportunities derived from the notoriety that the openness of high quality data brings about. In other cases, different stakeholders (governments, private firms, NGOs, etc.) pool complementary data to create an open database and create value through scope. The initial value is incremented by third parties that can access and contribute to the database, something stakeholders can benefit from because they can gain knowledge from the more valuable pooled open data and offer better services based on it. Finally, in some cases like OpenStreetMaps individuals crowdsource open data that is valuable because of the superior size and quality that certain datasets can attain when they are crowdsourced. These datasets, which are governed as commons, can be financed through donations, public funding and, less commonly, through the selling of products related to the object of the database.

We can conclude that a variety of stable data-based business models exist. Although in the case of urban data many of them are based on de facto propertization of data, different business models manage to create value based on opening data. The future of all of these business models will be highly dependent on the evolution of the legal regimes applying to data, which is currently under debate among legal scholars and policy makers. In a context in which machine learning keeps gaining relevance, the capacity to appropriate large amounts of data can become a competitive advantage for a few firms. The definition of personal data, a concept which boundaries are yet unclear in the dawn of the era of connected devices (among other things), will be a key factor in determining how value will be created from data and, more importantly, how it will be distributed.

References

Amatriain, X. (2015). In Machine Learning, What is Better: More Data or better Algorithms. Retrieved November 10, 2017, from https://www.kdnuggets.com/2015/06/machine-learning-more-data-better-algorithms.html

Anciaux, A., & Farchy, J. (2015). Données personnelles et droit de propriété: quatre chantiers et un enterrement. *Revue Internationale de Droit Économique*, *29*(3), 307–331.

Batini, C., & Scannapieco, M. (2006). *Data Quality Concepts, Methodologies and Techniques. 2006*. Springer-Verlag.

Benabou, V.-L., & Rochfeld, J. (2015). À qui profite le clic: le partage de la valeur à lère numérique. *Paris: Odile-Jacob*, 42–43.

Broca, S. (2017). Le digital labour, extension infinie ou fin du travail? *Tracés. Revue de Sciences Humaines*, (32), 133–144.

Carballa Smichowski, B. (2016). *Data as a common in the sharing economy: a general policy proposal. Document de travail du CEPN, Paris: Centre d'economie de l'Université Paris Nord*.

Casilli, A. (2015). *Digital Labor: travail, technologies et conflictualités*. Editions de l'INA.

Chandler Jr, A. D. (1993). *The visible hand: The managerial revolution in American business*. Harvard University Press.

Chignard, S., & Benyayer, L.-D. (2015). *Datanomics. Les nouveaux business models des données*. FYP editions.

Cleland, S. (2011, October 3). Google's "Infringenovation" Secrets. Retrieved October 31, 2017, from https://www.forbes.com/sites/scottcleland/2011/10/03/googles-infringenovation-secrets/#137c528930a6

Cohen, J. E. (2017). Law for the platform economy.

Cohen, S. S., Zysman, J., & DeLong, B. J. (2000). Tools for Thought: What is New and Important about the" E-conomy"? *Berkeley Roundtable on the International Economy*.

Commons, J. R. (1893). *The distribution of wealth*. Macmillan and Company.

Duch-Brown, N., Martens, B., & Mueller-Langer, F. (2017). The economics of ownership, access and trade in digital data.

Floridi, L. (2014). Big Data and information quality. In *The philosophy of information quality* (pp. 303–315). Springer.

Fuchs, C. (2014). *Digital Labour and Karl Marx*. Routledge.

GIGREF. (2015, October). Economie des données personnelles. Les enjeux d'un business éthique. Retrieved from http://www.cigref.fr/wp/wp-

content/uploads/2015/11/CIGREF-Economie-donnees-perso-Enjeux-business-ethique-2015.pdf

Harracá, M. (2017, Spring). *Business models and organizational forms: searching the edge of innovation in Google and Amazon*. Université Paris XIII, Paris.

Kushida, K. E., Murray, J., & Zysman, J. (2015). Cloud computing: from scarcity to abundance. *Journal of Industry, Competition and Trade*, *15*(1), 5–19.

Lambrecht, A., & Tucker, C. E. (2015). Can Big Data Protect a Firm from Competition?

Lambrecht, M. (2015). Ryanair c. PR Aviation, Note sous CJUE (2e ch.) C-30/14.

Lyon, L. (2016, May 16). The End of Big Data. Retrieved November 6, 2017, from https://www.databasejournal.com/features/db2/the-end-of-big-data.html

Marzloff, B. (2013). *Sans bureau fixe: transitions du travail, transitions des mobilités* (Vol. 5). FYP éditions.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston: Eamon Dolan / Houghton Mifflin Harcourt.

Merzeau, L. (2013). L'intelligence des traces. *Intellectica-La Revue de l'Association Pour La Recherche Sur Les Sciences de La Cognition (ARCo)*, *1*(59), 115–135.

OECD. (2015). *Data-Driven Innovation: Big Data for Growth and Well-Being*. OECD Publishing.

Olson, J. E. (2003). *Data quality: the accuracy dimension*. Morgan Kaufmann.

Open Data Institute. (2017a). Open enterprise case study: Syngenta. Retrieved November 21, 2017, from https://theodi.org/open-enterprise-big-business-case-study-syngenta

Open Data Institute. (2017b). What is open data? Retrieved November 7, 2017, from http://theodi.org/what-is-open-data

Peugeot, V. (2014, April 13). Données personnelles: sortir des injonctions contradictoires. VECAM. Retrieved from https://vecam.org/Donnees-personnelles-sortir-des-injonctions-contradictoires

Roché, É. (2016). Open data et business models. *LEGICOM*, (56), 121–127.

Rochfeld, J. (2017). Données personnelles: quels nouveaux droits? *Statistique et Société*, *5*(1), 45– 51.

Schlager, E., & Ostrom, E. (1992). Property-rights regimes and natural resources: a conceptual analysis. *Land Economics*, 249–262.

Sokol, D. D., & Comerford, R. (2016). Antitrust and Regulating Big Data. *Geo. Mason L. Rev.*, *23*, 1129.

Swyngedouw, E., & Swyngedouw, E. (2004). *Social power and the urbanization of water: flows of power*. Oxford University Press Oxford.

The Economist. (2010, February 27). Data, data everywhere.

Verdier, H., & Murciano, C. (2017). Les communs numériques, socle d'une nouvelle économie politique. *Esprit*, (5), 132–145.

Wang, R. Y. (1998). A product perspective on total data quality management. *Communications of the ACM*, *41*(2), 58–65.

WIPO. (2017). How are Trade Secrets Protected?? Retrieved October 31, 2017, from /sme/en/ip_business/trade_secrets/protection.htm

Young, C. (2014). HarassMap: using crowdsourced data to map sexual harassment in Egypt. *Technology Innovation Management Review*, *4*(3), 7.

Zott, C., Amit, R., & Massa, L. (2011). The business model: recent developments and future research. *Journal of Management*, *37*(4), 1019–1042.