PhD THESIS DECLARATION

I, the undersigned

FAMILY NAME	Minervini
NAME	Marco
Student ID no.	1239946

Thesis title:

Essays in Data Governance: Evidence from the Health Sector

PhD in	Public Policy & Administration
Cycle	30
Student's Advisor	Grandori Anna
Calendar year of thesis defence	2020

DECLARE

under my responsibility:

1) that, according to Italian Republic Presidential Decree no. 445, 28th December 2000, mendacious declarations, falsifying records and the use of false records are punishable under the Italian penal code and related special laws. Should any of the above prove true, all benefits included in this declaration and those of the temporary "embargo" are automatically forfeited from the beginning;

2) that the University has the obligation, according to art. 6, par. 11, Ministerial Decree no. 224, 30th April 1999, to keep a copy of the thesis on deposit at the "Bilioteche Nazionali Centrali" (Italian National Libraries) in Rome and Florence, where consultation will be permitted, unless there is a temporary "embargo" protecting the rights of external bodies and the industrial/commercial exploitation of the thesis;

3) that the Bocconi Library will file the thesis in its "Archivio Istituzionale ad Accesso Aperto" (Institutional Registry) which permits online consultation of the complete text (except in cases of temporary "embargo");

4) that, in order to file the thesis at the Bocconi Library, the University requires that the thesis be submitted online by the student in unalterable format to Società NORMADEC (acting on behalf of the University), and that NORMADEC will indicate in each footnote the following information:

- PhD thesis Essays in Data Governance: Evidence from the Health Sector;
- by Minervini Marco ;
- defended at Università Commerciale "Luigi Bocconi" – Milano in the year 2020;
- the thesis is protected by the regulations governing copyright (Italian law no. 633, 22nd April 1941 and subsequent modifications). The exception is the right of Università Commerciale "Luigi Bocconi" to reproduce the same, quoting the source, for research and teaching purposes;

5) that the copy of the thesis submitted online to Normadec is identical to the copies handed in/sent to the members of the Thesis Board and to any other paper or digital copy deposited at the University offices, and, as a consequence, the University is absolved from any responsibility regarding errors, inaccuracy or omissions in the contents of the thesis;

6) that the contents and organization of the thesis is an original work carried out by the undersigned and does not in any way compromise the rights of third parties (Italian law, no. 633, 22nd April 1941 and subsequent integrations and modifications), including those regarding security of personal details; therefore the University is in any case absolved from any responsibility whatsoever, civil,

administrative or penal, and shall be exempt from any requests or claims from third parties;

7a) that the thesis is not subject to “embargo”, i.e. that it is not the result of work included in the regulations governing industrial property; it was not written as part of a project financed by public or private bodies with restrictions on the diffusion of the results; is not subject to patent or protection registrations.

Date __10/12/2019_____

Abstract

The realization of the potential of the data revolution has been mixed and uneven. Except from rare settings characterized by data abundance, several settings run the risk of not being able to take advantage of advanced analytics technologies due to their inability to build sustainable collaborative initiatives capable of combining data of sufficient size and quality, especially where access to data sources is fragmented across organizations.

The present dissertation intends to improve the understanding on how best to combine data from multiple actors and what are the appropriate governance policies, structures and procedures, by enriching knowledge on the phenomenon both from a theoretical and empirical point of view.

The dissertation is composed by three independent but conceptually interconnected chapters.

The first chapter is mostly intended to improve the conceptualization of the "resource" that is object of governance (i.e. data) encompassing three core elements: a microfounded characterization of the stages of the data production process building on philosophy of information; a characterization of data as an economic good; and an exploration of how certain traits of data that are widely discussed in the data literature may affect the governance and organizational design of data pooling initiatives according to existing lenses in organization theory.

The second chapter captures the heterogeneity of governance forms of data pooling initiatives by conceptualizing them as configurations of property rights à la Ostrom. By systematically analysing how contribution, access, extraction, removal and management rights are allocated across actors involved in cases of data pooling initiatives in the healthcare sector, the chapter fills a knowledge gap on how these initiatives are designed, by appreciating elements of variance and consistency across them.

The third chapter develops a club theoretical model that is intended to understand how certain contingent conditions may lead to the non-emergence or limited sustainability of interorganizational data pooling initiatives. The model allows to: 1) organically consider the interaction between factors that may influence the emergence of a data pooling problem; 2) disentangle the data pooling problem and decomposing it in a provision problem, a membership problem and a stag hunt problem; 3) analyse, through a scenario simulation, how different types of data may affect both the emergence of a data pool and the broadness of its membership.

Contents

ACKNOWLEDGEMENTS

Se le occasioni per ringraziare in privato sono frequenti e spesso non còlte, le occasioni di ringraziare in un documento pubblico sono estremamente più rare. Per la prima volta, ho voglia di cogliere questa occasione. Sebbene gli errori contenuti in questo documento sono esclusivamente da attribuire a me, sono molte le persone che hanno contribuito direttamente e indirettamente al completamento di questa tesi e di questo lungo percorso.

Ringrazio Anna, per avermi accompagnato con pazienza lungo questi anni, tollerando repentini cambi di direzione e le più disparate speculazioni teoriche, e per avermi condotto saggiamente nella conoscenza delle teorie organizzative. La tua passione e curiosità intellettuale sarà sempre per me fonte di ispirazione.

Ringrazio Giovanni, senza il quale non sarei qui a scrivere questi ringraziamenti. Se esiste un responsabile di questa scelta di vita sei sicuramente tu. Grazie per aver accompagnato sapientemente un confuso neolaureato alla ricerca della sua dimensione.

Ringraziamenti speciali vanno ad Amelia. Fonte di supporto su tutti i fronti: accademico, umano ed economico. La tua empatia e umanità ti rende una delle poche persone a cui sono riuscito a fare gli stessi ringraziamenti di persona. Spero di riuscire ad essere un domani un Amelia per un giovane dottorando.

Ringrazio Ellie e Michele, compagni di ansie (molte) e di fantastiche mangiate e bevute serali (troppo poche). Siete stati un’ineguagliabile fonte di supporto e compagnia.

Ringrazio Paolo, Gianmario, Gloria, Sahar, Naila, Ramya, Marina, Silvia, Riccardo, Stefano, Marco, Maria, Italo, Simone, Nicoletta, Aleksandra, Giorgio, Francesco e Livio per aver contribuito in forma differente a rendere più piacevoli le mie giornate in ufficio.

Ringrazio la mia famiglia (Mamma, Papà e Lisa) per il loro amore e supporto incondizionato. Ringrazio Papà, per avermi insegnato l'onestà e l'integrità. Ringrazio Mamma, per avermi insegnato la disciplina. Ringrazio Zia Marianna, esempio di generosità verso la quale nutro un debito incalcolabile. Ringrazio Antonella, le quali rassicurazioni in alcuni momenti chiave hanno messo le fondamenta per completare questo percorso ('Tu pensa a studiare').

Nessuna riga di questa tesi sarebbe stata scritta senza il supporto e l'amore incondizionato di Barbara. Non c'è stata riga di questa tesi, e non solo, che non sia passata sotto i tuoi occhi attenti. A qualsiasi ora del giorno e della notte. Mi hai sostenuto nei momenti di fragilità giustificati e mi hai confrontato in quelli ingiustificati. Mi hai tenuto la mano lungo ogni momento di questo percorso, senza mai lasciarla, sopportandone tutte le conseguenze. Con te accanto, le paure diventano piccole ed i sogni enormi. Grazie.

Alle 3 L della mia vita e ad una B che continua a renderla più speciale.

INTRODUCTION

Recent years have been characterized by an unprecedented generation of digital data and experts suggest that this exponential trend will continue in the following years (Reinsel, Gantz, & Rydning, 2017).

This trend is the result of the interaction of three different phenomena. First, the marked diffusion and use of control instruments (e.g. sensors) and the diffusion of smartphones have increased the ability to render into data many aspects of the world ("datafication") (Koutroumpis, Leiponen, & Thomas, 2017b; Mayer-Schönberger & Cukier, 2013a). Second, private and public organizations have undergone several initiatives aimed at collecting or transforming in digital form information that was originally collected in analog form ("digitization"). Further, the evolution of complementary technologies like artificial intelligence, which make data even more appealing, have furthered the incentives to increase the pace of datafication and digitization (Agrawal, Gans, & Goldfarb, 2018).

This unprecedented generation of data has spurred a marked enthusiasm on the ability of the Big Data revolution to generate benefits across industries and for society (Brynjolfsson & McAfee, 2014; Cukier & Mayer-Schoenberger, 2013; Mayer-Schönberger & Cukier, 2013a; McKinsey Global Institute, 2011). However, most of the claimed potential has not fully unleashed (McKinsey Global Institute, 2016), also due to the uneven and heterogenous distribution of data across settings with spots of abundance, scarcity or absence (Borgman, 2015).

The same unfulfilled promise has characterized the health sector. In 2011 The McKinsey Global Institute published a report highlighting the transformational potential of Big Data in health in the US (Groves, Kayyali, Knott, & Van Kuiken, 2013; McKinsey Global Institute, 2011). The

report predicted a reduction in health costs by 300 to 450 billion $, due to: 1) Targeted disease prevention; 2) Alignment around proven pathways and coordinated care among providers; 3) Payment innovation; 4) Accelerated discovery in R&D and improved trials operations. However, only 10% of the benefits predicted by the report has actually manifested and data sharing challenges have been mentioned as one of the main reasons for this failure (McKinsey Global Institute, 2016).

Most of the promises mentioned above are spurred by the impressive confidence that societal actors have placed on the potential of algorithms and analytics. These tools widely rely on statistical principles and, as a consequence, are eager of data and are effective and non-biased only if alimented with enough high-quality data (Redman, 2018). Except from some rare cases where single economic actors[1] are in control of vast amount of data, in order to be large enough and useful, data needs to be pooled at large scale among multiple actors (Mattioli, 2017; Roski, Bo-Linn, & Andrews, 2014). However, existing evidence shows high concern for the long run sustainability of attempts to combine data from multiple contributing actors (Gliklich, Dreyer, & Leavy, 2014; Zaletel & Kralj, 2015) and there is a very limited understanding on how best to combine data from multiple sources and what are the appropriate governance policies, structures and procedures (Bates, Heitmueller, Kakad, & Saria, 2018; Holmes et al., 2014). Relying on the wording of Mattioli (2017), there is broad evidence of the presence of a data pooling problem and scant knowledge to support the proper design of governance arrangements that are able to sustain adequate contribution to data pools.

The present dissertation intends to improve the understanding on how best to combine data from multiple actors and what are the appropriate governance policies, structures and

[1] I am mainly referring to tech giants.

procedures, by enriching knowledge on the phenomenon both from a theoretical and empirical points of view. The theoretical section contributes by: 1) improving the conceptual clarification of the resource that is object of governance (i.e. data); 2) systematizing the theoretical understanding of the many facets of the data pooling problem; 3) advancing a framework to capture the design of data pooling initiatives. The empirical section instead explores how data pools are actually governed, structured and managed and identifies elements of variation and regularities in their design.

The empirical setting chosen in the dissertation is the health sector. The health sector shows an interesting paradox. Despite the diffused concern on limited data sharing and on the sustainability of data pooling initiatives that can be found in popular media and in academic literature, the health sector is instead characterized by a rich amount of highly heterogenous attempts to create, share and integrate data between multiple sources and actors.

The present dissertation intends to take advantage of this contradiction. It intends to develop a mutually beneficial relationship between an empirical setting in dramatic call for a more theoretical understanding of the phenomenon and a set of theoretical perspectives that may markedly benefit from a setting that has been limitedly explored and that shows a rich organizational variance.

The dissertation will be structured in three main chapters, which are structured in a paper form. A brief description of each follows.

The first chapter is mostly intended to improve the conceptualization of the “resource” that is object of governance (i.e. data) and to introduce the key concepts and intuitions that will be employed through the dissertation. First, it captures the biophysical stages that data take through the data production chain, from data collection to the generation of a data derivative (an

algorithm or data intensive research). Second, it identifies a set of characteristics of data that may be relevant for understanding the governance of data poling initiatives (Non subtractability, Excludability, Option Value, Sensitivity, Volume, Velocity, Variety) and discusses what are the type of governance challenges that may emerge in case data have one or more of the identified characteristics.

The second chapter attempts to develop a configurational framework intended to capture the governance forms that initiatives intended to create, share, integrate and use data may take. By conceptualizing governance forms as configurations of property rights à la Ostrom, it relies on this framework to comparatively capture how several data pools in the health sector are actually structured and managed.

The third chapter attempts to capture some of the intuitions of the first chapter by including them within a club theoretical model that is intended to microfound the data pooling problem, i.e. the non-emergence or limited sustainability of interorganizational initiatives that pool data with the purpose to generate an outcome that is fundamentally based on statistical principles (e.g. an algorithm or data intensive research). The model allows to: 1) organically consider the interaction between factors that may influence the emergence of a data pooling problem; 2) disentangle the data pooling problem and decomposing it in a provision problem, a membership problem and a stag hunt problem; 3) analyse how different types of data may affect the emergence of a data pool and the broadness of its membership.

While the chapters are structured as they were independent papers, they are widely interdependent. The first chapter offers the building blocks for the two following ones. The second chapter, on the one side, relies on the rich conceptualization of the resource proposed in the first chapter to refine the bundle of property rights à la Ostrom that is then employed as a

framework for the comparative analysis. On the other side, by exploring a setting where data are characterized by the copresence of sensitivity, option value and non subtractability, it shows how the analysed initiatives are designed to address two of the core questions that emerge from the first chapter: 1) how do pooling initiatives balance the trade-off between expanding access to maximize the benefits of data option value and reducing access to avoid undesired implications on data subject due to data sensitivity? 2) How are data pooling initiatives designed to incentivize contributors given the limited (if not absent) possibility to rely on a market type of exchange?

In a similar vein, the third chapter relies on the intuition generated in the first chapter. The choice of employing a club theoretical model derives from the economic classification of the good performed in the first chapter. In addition, the characterization of the data production chain allows to capture what are the costs and benefits that an organization faces in the process from data collection to the generation of a data derivative and how these costs and benefits may influence the decision to pursue this endeavour and, in case, to pursue it individually or collaboratively.

After the exposition of the three chapters, a final concluding section follows.

1 DIMENSIONALIZING DIGITAL DATA SYSTEMS FOR GOVERNANCE

1.1 INTRODUCTION

The overwhelming phenomenon of datafication, i.e. the ability to render into data many aspects of the world (Mayer-Schonberger & Cukier, 2013), has significantly and increasingly attracted the attention of scholars and practitioners. This attention has frequently been manifested in several attempts to characterize data according to a set of (varying) dimensions and typologies of data from a multiplicity of perspectives.

However, there is a scarcer understanding on the implication of those characteristics on the emergence and persistence of governance forms intended to the creation, sharing, integration and use of data. Further, scholar contributions that have started to explore how economic characteristics may influence the (non) emergence of specific governance forms have been characterized by a tendency to refer to data either as a vague term and as unique monolith or have focused on several different units of analysis, sometimes leading to opposite conclusions.

The present chapter improves the conceptualization of data in many respects. In the first place it better characterizes the resource that is object of governance identifying the key biophysical manifestations of data as resource in the data "production chain". Second, it identifies the set of dimensions that have been employed to characterize data in the existing literature that may be relevant for governance design and explores both what are the implications of data having these characteristics in terms of the governance challenges that certain characteristics generate and what are the possible governance responses to these emerging challenges. Finally, it explores how the copresence of and the interaction between different features of data generate

a set of peculiar governance challenges, which governance responses may need to be better explored.

The chapter clarifies the debate on data as economic good, suggesting that while economic characterization of data varies across the production chain, after its digitalisation we might consider data as a club good more than a common pool resource, as frequently argued. This distinction suggests that: 1) the primary concern for club goods is their provision more than their depletion; 2) excludability allows to govern a governance dimension that is not available in common pool resources, i.e. membership size.

The chapter also identifies how the 3Vs (Volume, Velocity, Variety) that are commonly employed to define Big Data also have other governance implications apart from being sources of technical complexity. Higher velocity may reduce reliance on formal contracts and increase predictive knowledge. Higher variety, under some contingent conditions, generates 'team production' problems and increases the probability of engaging more heterogeneous actors, calling for more hierarchical solutions and for reliance on more articulated collective decision-making solutions. With respect to data volume, differently from how it is commonly considered in organization theory, the chapter suggests that organizational design needs to be information maximizing rather than information minimizing (Galbraith, 1974).

It then explores how the copresence of some of the identified features of data interact among them. The copresence and interaction between data sensitivity, non subtractability and option value suggests that it may be highly complex and inefficient to rely on market exchange solutions to govern the transaction of data. It is instead more plausible to observe associational contracts and or the establishment of independent legal entities, like foundations and associations. Further, the copresence of data option value, sensitivity and non subtractability

generates a trade-off between the potential value that could be generated expanding access to data due to knowledge expansion on the potential uses and the risk that some of the multiple uses may be detrimental to data subjects. Purely open and closed solutions lie on the two sides of the trade-off and may be even unfeasible. The paper suggests that how pooling initiatives are designed to balance this trade-off deserves to be better explored. Further, the impossibility to rely on market exchange models and the complexity in data evaluation exclude the solution of monetary incentive to incentivize contribution of data by data collectors. The type of solutions pursued by pooling initiatives need to be better explored.

The paper is structured as follows. Section 2 discusses the biophysical states that data take through the data production process. Section 3 reviews some of the main sets of economic and non-economic dimensions that have been employed to characterize data and discusses their governance implications. Section 4 discusses how the interaction between some of the identified dimensions generates key trade-offs and governance challenges for pooling initiatives.

1.2 DATA PHYSICAL FORMS

As suggested by Ostrom (2005) in her study of the governance of natural resources, in order to grasp how a resource is governed it is necessary to understand what are the biophysical and material conditions that the resource of interest takes. However, as argued by Rosenberg (2014), "the term data does heavy lifting yet is barely remarked upon". The increasingly extensive use of the term "data" has further expanded the meanings that have been attached to it, thus reducing terminological clarity. By building on the literature in Information Studies and Philosophy of Information, the primary goal of this section is to clarify what are the physical forms that the resource of interest takes through the data production and processing chain.

Scholars of natural resource commons have found helpful to distinguish between *resource systems* and *resource units* (E. Ostrom & Hess, 2007). A resource system (e.g. a lake) is composed by and is the source of the flow of resource units (e.g. certain quantity of water, fishes). In the data production and processing chain, there are three resource units (Signal, Captum and Derivative) and two resource systems (Signal Pool and Capta Pool). The two resource systems are composed by a pool of resource units of one type and are the source of extraction of another type of resource unit.

Signal refers to a change in status that *can be* collected (Kitchin, 2014). Looking at the "data production process", the very origin of what is actually captured into bytes is the physical world. What elements of the physical world are captured and how they are captured is the result of a choice and a paltry subset of the potential elements (signals) in the physical world that could be captured. A medical encounter between a patient and a clinician is an infinite set of signals (a signal pool) that may be captured. In this setting, the clinician, although subject to multiple factors that may influence the choice, decides what are the signals to capture and how to capture them. In a similar vein, also in online settings like a social network, the actions made by a user in a social network may be multiple (writing, clicking, scrolling). The social network has potential access to all these signals and determines which signals to capture.

Captum refers to the *signals* within the *signal pool* that are actually captured into bytes (Kitchin, 2014), i.e. a *captum* is the result of the selection process described in the paragraph above. The result of a medical encounter is the generation of a set of *capta* on a patient *(*e.g. body temperature, symptoms and heart rate). A combination of set of captured signals from a multiplicity of patients is a *capta pool.*

Derivative refers to data generated through the processing of the *capta pool* (Floridi, 2010). The result of a statistical analysis performed on a multiplicity of *capta* (e.g. a mean, a regression coefficient, a trained prediction model) is what is defined as a *derivative*. The primary goal of comparative effectiveness initiatives, data intensive research and initiatives that develop prediction algorithms is to generate high quality *derivatives* through the statistical processing of a *capta pool*. As quality of a data derivative is subject to basic principles of statistics, the *capta* collected needs to have an adequate degree of semantical and technical homogeneity, and the *capta pool* needs to have sufficient size for the purpose and, in many cases, a sufficient diversity of sources to avoid bias.

Figure 1 represents the data production and processing process as described.

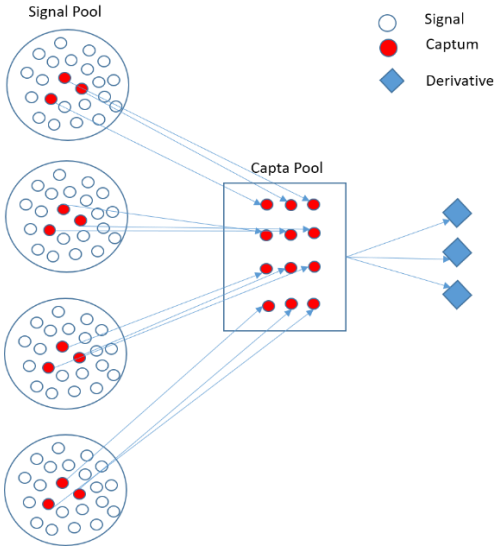

Figure 1 - Data Physical States

Across the data production stages, it is possible to identify different roles and these roles will be employed for clarity across the following sections of the chapter. When referring to personal data, the *signal pool* corresponds to the set of traits and behaviour of a *data subject*. Otherwise, when data are not personal, no *data subject* is involved. *Signals* are then transformed into *capta* by a *data collector*. *Capta* are contributed to a *capta pool* by a *data contributor*. *Derivatives*

are extracted from the pool by a *data user*. Across different cases an actor may cover just one of these roles or cover simultaneously a multiplicity of them.

In addition to capture "physical forms" and the type of actors that may be involved in the data production chain, in order to understand governance challenges and governance solutions in data pooling initiatives, a characterization of the resource is needed. The following section performs a characterization of data identifying the dimension that are mostly relevant in terms of their governance implications.

1.3 CHARACTERIZING DATA: RELEVANT DIMENSIONS FOR GOVERNANCE

Characterization of the resource features is primarily an abstraction effort that allows to locate data goods in a broader family of goods, thus allowing to capture how to theoretically frame the governance and organizational design challenges derived from the type of good and to draw from existing literature to derive design implications. Besides the effect of each dimension considered independently, I will also explore whether and how the combined presence of the analysed features interact to generate peculiar and idiosyncratic challenges.

While the characteristics identified are not fully exhaustive of all the characteristics that have been attached to data, in the present chapter I explored the ones that may be considered more relevant, either because they have clear and strong implications in terms of governance challenges and, consequently, of governance design (Excludability, Subtractability, Option Value, Sensitivity) or because they have been so widely employed to describe data, that it becomes relevant to attempt a translation effort in terms of their governance implications (Volume, Velocity, Variety).

1.3.1 Economic Characteristics (Subtractability and Excludability)

Two attributes have been identified in the political-economy literature to identify the nature of a good (Hess & Ostrom, 2003; V. Ostrom & Ostrom, 1977): its degree of 1) Subtractability and 2) Excludability. Different combinations of degree of subtractability and excludability of the good object of governance differently affect the type of acceptable choice set that rational actors face and, consequently, pose different types of dilemmas and governance challenges. Therefore, for each combination of degrees, different governance solutions may be more or less feasible or preferable.

In the case of data, different perspectives emerged in defining which type of economic good data is. For this reason, in this section, I will first identify how data as an economic good has been characterized in the literature, trying to reconcile differences in perspectives. Then, I will suggest that data may be better characterized as club goods and look at what are the implications of dealing with club goods in terms of governance design.

1.3.1.1 Rivalry / Subtractability[2]

Most of the literature that made attempts to broadly characterize (digital) data have defined it as a non-rival good (Acquisti, Lane, Stodden, Bender, & Nissenbaum, 2010; Duch-brown & Mueller-langer, 2017; Koutroumpis & Leiponen, 2013). A (non) rival good is a good for which the benefits consumed by one actor (do not) subtract from the benefits available to others (Hess & Ostrom, 2003). However, Purtova (2015) argues that data may be also considered as rival. The distinction is mainly driven by the ambiguous decomposition of the data resource used by several authors. In this respect, the distinction between *signal*, *capta* and *derivative* can help in disentangle this apparent discrepancy. Before being transformed in digital forms, i.e. being

[2] (Non) subtractability and (non) rivarly will be used interchangeably through the paper and the dissertation.

transformed from *signal* to *capta*, signals may be frequently seen as a rival resource. The time spent online by users that generates the signal is clearly a rival resource. If a specific social network "consumes" the signal/time of a user, another social network cannot consume it (Purtova, 2015). The same applies in clinical settings. It would be unethical to repeat an intervention for the mere purpose of data (capta) collection. Rivalry of signals also manifests in non-interventional data collections. A patient may reduce its willingness to answer if a request is repeated by a second clinician. This dynamic equally refers to non-personal data. The signal generated in a specific moment of time can be extremely rival. For instance, capta collected through complex experiments. Further, by subtracting the access to a signal, actors might also subtract the opportunity to determine what (elements of the) signal to capture.

In this sense, the idea that "data" are in general nonrival overlooks the rivalry that may be present in several cases at a *signal* capture stage. Not every signal can be accessed by a data collector without reducing the possibility to do the same at equal access cost for another collector. This condition may lead to two commonly observed phenomena: either the concentration of *capta* under a single organization or the strong fragmentation of *capta* across multiple organizations, as repeated collection of the same signal for a second collector becomes markedly more costly or impossible.

While signal may have elements of subtractability when collected in a highly specific moment or condition or when it relies on scarce resources like time and willingness of a data subject, once it is translated into digital form, i.e into *capta* and then eventually pooled into a *capta pool*, it becomes a non-subtractable good. A slightly more complex reasoning applies to derivatives. While in several cases the use of a derivative by one actor does not reduce the stream of benefits for others, it may be that for some derivatives non subtractability is only partial. While a *digital derivative* is physically non subtractable, it may be that the downstream

benefits generated from its use have instead a degree of subtractability. An example is when a data derivative is employed to generate a research publication. The derivative itself (e.g. a statistical estimation of the effect of a specific treatment on the mortality of a patient) is not subtractable. However, one of the mechanisms through which benefits are extracted, i.e. authorship, can instead be characterized by elements of subtractability. For instance, certain authorship positions generate higher benefits than others in terms of career returns. Similarly, a higher number of authors may equally dilute benefits in terms of future career perspectives.

1.3.1.2 Excludability

The excludability of a good is commonly determined by the costs sustained by actors to exclude others from benefiting from it. In other words, a good is non-excludable when it is particularly costly to exclude individuals from using the flow of benefits originated by it either through physical barriers or legal instruments (Hess & Ostrom, 2003). As highlighted by Duch-Brown & Mueller-Langer (2017), different perspectives emerge with respect to the excludability of data on the whole spectrum from non-excludability to excludability.

Some authors claim that the absence of intellectual property rights leads to a substantial non excludability and support their argument by highlighting the empirical reluctance of many actors in engaging in data exchanges. However, this perspective seems to mistakenly not fully distinguish between the nature of the good and the property regime imposed on the good (Hess & Ostrom, 2003). Further, the very fact of being able to determine whether to grant access to/exchange the good or not, seems to go in favor of a substantial ability to exclude another actor from the stream of benefits originated by data.

Differently, Drexl (2016) maintains that data are factually excludable through technical means. In a similar vein, Kerber (2016), while more cautiously affirming that excludability should be

considered as a continuum more than a dichotomy and warning against potential empirical evolutions of it, claims that it is technologically feasible to keep data secret and protect them against copying and leaking to the public. Both authors support their arguments by highlighting the absence of under-provision of data, which might be instead expected in case the good is non-rival and non-excludable (a public good).

Koutroumpis, Leiponen, & Thomas (2018) argue that excludability is only limited to pure secrecy .While it is true that an actor can exclude someone else from having access to data, once access is granted it is complex to exclude the undesired extraction of benefits later. Following Drexl and Kerber, the argument from Koutroumpis et al. related to excludability is highly dependent on the type of technological tools that are employed for granting access. It is increasingly possible to grant controlled access to users through secure environments, i.e. technologies that allow actors to access and extract derivatives from a pool but exclude the possibility that an actor can appropriate and reproduce them. As such, it is technically feasible to exclude noncontributors from accessing and using a capta pool. In light of this, a capta pool can be considered as an excludable good.

1.3.1.3 Implications

Previous sections suggest that capta and especially capta pool are mostly excludable and non-subtractable. As such, they configure themselves as a club good. While Benkler (2014) argues that the frequent confusion with common goods has led to a limited knowledge of what could be the most effective governance arrangements in the management of this kind of goods, club theory (Buchanan, 1965; Cornes & Sandler, 1996) seems instead to provide some theoretical guidance in understanding the implications of this characterization for the governance of the resource.

1.3.1.4 Club Theoretical Approach and Club Theory

Before exploring the propositions originated by club theory, it is important to make a distinction between a "Club Theoretical" approach and "Club Theory". A club theoretical approach is a theoretical approach that frames individual decision making on producing and benefiting from a good as a dual calculative decision: 1) How much of the good to produce? and 2) How many actors to involve in the production of the optimal quantity of the good?. In other words, a club theoretical approach considers each production decision as a decision to generate a club. A club theoretical approach would consider the decision to produce individually as a club of membership size 1.

A club theoretical approach simply needs one assumption to work: benefits derived from the good included in the club can be excludable at the discretion of the contributors. Then, peculiarity of the good and the production function (McGuire, 1972) may play a role in determining what could be the best solution in terms of provision and membership.

1.3.1.5 Club Theory

Compared to a more generalist club theoretical approach, club theory imposes further restrictions on the type of good some of its theoretical predictions apply on.

For club theory, a club good is a good that is (Buchanan, 1965):

- Excludable: Individuals who do not contribute to financing the club can be prevented, at relatively low cost, from gaining access to the benefits of club membership.
- Congestible: Although consumption is not entirely rivalrous, each member of the club imposes a negative externality on his fellows.

- Divisible: Once club membership has reached an optimal size, individuals who want to join but have been excluded can form a new club to produce and consume the same good.

Club Theory is then concerned to establish optimal club size, which is described as a result of a straightforward exercise in equating costs and benefits at the margin. An optimal club optimizes costs and benefits against three conditions (Cornes & Sandler, 1996):

- A Provision Condition. There exists an optimal club size in terms of quantity of a good to be included in a club.
- A Membership Condition. There exists an optimal club size in terms of number of members to be involved in the club.
- Utilization Condition. There exists an optimal utilization level by each group member.

Utilization Condition has limited relevance in data pooling initiatives. Differently from the empirical settings that were employed as example at the base of the theoretical development of club theory (e.g. swimming pools and sport clubs), in data pool the congestion problem typical of club goods does not derive from intensity of use of the capta pool, but from the inclusion of additional members which may generate negative externalities of other nature. As such, for the purpose of this chapter I will focus only on Provision Condition and on Membership Condition, which helps to highlight what are the core distinguishing elements of club goods and what are some of the core governance implications of dealing with this type of good.

1.3.1.5.1 Provision Condition

The primary problem for club goods is their provision, i.e. actors should be primarily willing to contribute to the provision of the club good. In this respect, club goods are different than common pool resources. In common pool resources, the resource is already present, and the

core governance concern is to avoid its depletion. However, for club goods the resource needs to be primarily generated and provided.

This distinction does not mean that once the club good is created, it may not be subject to some risks of depletion. For instance, a source of depletion might be misuse from club members. However, it simply means that before running the risk of depletion, a club good runs the risk of non existence, i.e. of not being provided.

As such, the governance of club goods needs to take into account a two-step process. Before becoming a common (E. Ostrom, 1990), a club good like a capta pool may be an anticommon (Buchanan & Yoon, 2000; Heller, 1998). This makes governance design more complex, and the design of club goods needs to strike an accurate balance in order to ensure the realization of the comedy of the commons (Rose, 1986) and avoiding it turns into a tragedy (Hardin, 1968). As argued by Frost & Morner (2010), institutional design for club goods need to be primarily inclusive before being exclusive.

In fact, existing literature exploring data commons suggests that other types of social dilemmas tend to emerge more markedly than the tragedy of the commons (Strandburg, Frischmann, & Madison, 2017a). Instead, the identified dilemmas are: 1) coordination dilemmas in standard settings; 2) coordination problems in settings where contributions by all actors in the system are fully complementary to make any research possible (like in rare diseases settings (Strandburg, Frischmann, & Cui, 2014); 3) dilemmas in the reallocation of superaddictive returns from the combination of contributions. All dilemmas that refer more to the provision phase than to the depletion after it.

1.3.1.5.2 Membership condition

Two factors interact to make membership relevant when referring to club goods and even more when referring to data pools. On the one side, it is possible for an actor to determine optimal membership as the club good is excludable, a choice possibility that is not present in other types of goods like public good and common pool resource. On the other side, changes in membership generate a complex combination between positive and negative externalities. In data pooling contexts, positive externalities derive from cost sharing in data collection and in superadditivity generated through data combination. Negative externalities derive instead from three core sources of congestion: coordination, moral hazard and rivalry. As such, at certain levels of membership the provision of club goods may be non-sustainable.

Thus, in terms of governance prescription, club theory has implications at two levels: at the planner level (between club perspective or total economy viewpoint) and at the club level (within club perspective).

At the planner level the core argument of club theory is that multiple clubs of more limited size may be welfare maximizing compared to a unique club. Apart from simple maximization, it also suggests that a unique club may even lead to the non provision of the club good.

The within club perspective also follows the same idea. There is an optimal club size in terms of membership that maximizes members benefits. Adding or removing a member to the optimal club size would make everyone worse off. This hypothesis is supported by experimental evidence, which shows both that when actors are included in clubs that correspond to their preferred size they cooperate more and that actors are able to recognize optimal club size quite often (Tutić, 2013).

This is also consistent with some empirical results that show both that collaborative pooling initiatives that started with broad membership then tended to shrink in the longer run, and that smaller initiatives where most effective and more able to manage the emerging complexity (Welch, Loaufi, Fusi, & Manzella, 2016).

1.3.2 Option Value

Mayer-Schonberger et al. (2013) emphasize how data can be considered as a multipurpose good. Their potential uses can be multiple and in many cases data generated for a function may well serve other (unrelated) functions.

Mayer-Schonberger et al. (2013) define it as the option value of data and emphasize the high complexity that actors face when required to evaluate data due to the aforementioned high potential of being used for multiple purposes that are different from the one for which they have been generated.

Apart from complexity in evaluation, two existing theoretical perspectives from organization theory and economics may help in understanding how data option value may affect the governance of the resource: the Penrosian (Penrose, 1959) theory of the firm and transaction cost economics perspective (Williamson, 1981).

The first has implications on the governance of access and use of the capta pool. The second, as the name suggests, has implications on the core transaction that may take place in pooling initiatives, i.e. when capta are contributed to the capta pool.

Implications on access to and use of the capta pool

Penrose proposes a distinction between resources and services. Each resource has infinite services and the actual services employed by the owner of the resource are functions of

knowledge and sectoral constraints. Knowledge expands the potential services of the resource so that increasing knowledge monotonically increases the potential uses.

Thus, adding an additional actor in the use of a resource increases (or at least does not decrease) knowledge, in turn increasing the potential service that the data resource may offer. However, this expansion of potential services to be generated by the resource expands potential beneficial uses as well as potential misuses. In this respect, option value is the source of a very complex trade-off that influences the decision to involve additional actors in the use of a pool. Within this reasoning there is an implicit role of non-rivalry. If a resource was rival, it would have been allocated to (the best) unique use. However, nonrivalry allows to allocate the resource to all uses.

Impact on transaction (from capta to the capta pool)

Instead, from a transaction costs economics perspective, the option value of capta or database may lead to the increase of transaction costs, thus leading actors to not pursue any form of external transaction or to exclude some modes of transaction. The high unpredictability of the set of potential secondary uses is a strong source of uncertainty in the transaction. When a transaction is characterized by a high level of uncertainty, the contractual relation between two actors may be highly complex to be fully specified, resulting as significantly incomplete. Consequently, the contributor may not pursue the transaction being worried about the extensive space of several unpredictable undesirable uses or may prefer alternative solutions rather than a mere exchange transaction.

1.3.3 Sensitivity

Sensitive data are commonly defined through a list of types of data, rather than through a working definition. According to the EU General Data Protection Regulation sensitive data are

data that refers to: “racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership […] genetic data, biometric data, health data”. Similarly to the EU approach, every country employs a specific list of types of data that are considered sensitive to impose more restrictive regulations in terms of how they should be captured and the sanctions that may derive from misuse or violation. Only recent evidence attempts to capture a potential “spectrum of sensitivity” for particular sub categories of data (Rumbold & Pierscionek, 2018), showing that the perceived sensitivity of a particular type of data may vary widely both between societies or ethnic groups and within those groups, especially in light of the implications that the misuse of data has on the data subjects.

Sensitivity is a feature of the resource that has been widely overlooked in organization theory and design. Sensitivity of the resource is an additional factor to the typical opportunistic behaviour that may be envisioned in usual collaborative initiatives. The core concern is not only related to value appropriation between parties in the collaborative, but to a more indirect impact on potential damages to data subjects, as improper uses of data may generate negative externalities on them.

As such, data sensitivity can be considered as an additional determinant of organizational design.

Impact on rights allocation

On the one side, sensitivity of data can be considered as a source of moral hazard toward the data subject. Users in the pool may improperly use the combined set of data in a way that may damage the data subject. Further, existing evidence suggests that even mere access to data by the collecting entity creates negative emotional and cognitive outcomes for the data subject, which in turn translates in lower willingness to share or information misreporting (Martin, Borah, & Palmatier, 2017).

On the other side, sensitivity may be also considered as a source of moral hazard between users of the pool and contributing data collectors. The misuse by a user of the pool may damage the contributing data collector as it may endanger its ability to collect data in the future.

Protection of data subjects takes place through two governance means: 1) endowing data subjects with a set of rights on the resource; 2) imposing restrictions on data users in terms of access and uses of the captured data.

In many cases data subjects are endowed with decision rights directly guaranteed through regulatory prescription. For instance, many national laws require informed consent or limit redisclosure without the consent of a data subject (Miller & Tucker, 2014). The data subject is thus able to determine whether capta collected on her can be contributed to a pool. Further, experimental evidence suggests that concerns on data access may be reduced by allowing data subjects to observe and control data management procedures (Martin et al., 2017).

Other governance solutions refer instead to the minimization of access or usage rights. Regulation may prescribe restrictions on discriminatory usage (Miller & Tucker, 2014) and restrictions on secondary uses. In settings where moral hazard on third parties is generated by combination of capta, solutions like Chinese walls, where data exchange across actors is

forbidden, might be instead implemented (Grandori, 1991). However, in data pooling initiatives, solutions may be more complex as the combination of data is usually a necessary condition to generate any benefit out of the pool. As such, while sharing might take place, strong restrictions might be applied on the number and types of actors that might have access to the pool, with the goal of minimizing it.

Impact on modes of transaction

Aside from access issues, sensitivity also impacts the possible mechanisms that may be used to transfer data across technologically separable interfaces (Williamson, 1981). When data refer to people and especially when they are sensitive, while there might be a transfer of data from one actor to the other, the consequences of misuse are still borne by the data subject. This is a trait of the resource that Leiponen defines as inalienability (Koutroumpis et al., 2018). Inalienability affects both the act of collection between the data subject and the data collector and the possible transfer between the collector and a third party. As mentioned above, data transfer to a third party does not relieve the transferee from bearing the consequences of misuse. If data collector A decides to share sensitive data on patient P1 to data user B, in case data user B misuses the data damaging patient P1, also data collector A suffers a loss, given that data subject P1 may not be willing to allow collection of capta on her anymore. In this respect, the collector may prefer governance solutions that allow to exert control on secondary uses, thus avoiding a full transfer of rights. For instance, a pure sale transaction where a unique buyer acquires data from multiple collectors and where collectors do not exert decision rights on use may be highly implausible to observe.

1.3.4 Physical Characteristics

The implications of physical characteristics on the governance of data have been far less explored. However, in many instances, “Big Data” are usually distinguished from “Small Data” exactly in light of their physical characteristics. More precisely the three mostly used physical characteristics are the 3Vs: Volume, Velocity and Variety (Gandomi & Haider, 2015; Laney, 2001).

Given the fact that many actors in the data ecosystem tend to distinguish data according to these characteristics and that some actors suggest thresholds to determine whether data should be considered big or not, a richer understanding of whether and how this distinction is valuable for understanding the emergence of specific governance forms is needed in order to confirm or not the need to consider these dimensions for both understanding and designing the functioning of governance.

1.3.4.1 Volume

Kitchin & McArdle (2016) define the volume of a database as the product of the volume of records and the amount of records. In other words, volume corresponds to the size of capta pool.

Existing literature in organization theory has looked at information volume as a contingent factor for organizational design. The information processing view of organizational design (Galbraith, 1974) looks at organizational design as an information reducing tool. Optimal organizational design is the one that reduces the set of information to be processed in order to reduce the cognitive limit of the decision maker. If taken superficially this perspective suggests that data volume becomes a core organizational challenge and that organizational design needs to be oriented toward reducing data volume.

However, a more refined assessment suggests that Galbraith perspective cannot be directly applied to the issue of data volume.

First, it is important to draw a conceptual distinction between data as a resource and task-related information. According to Galbraith, information size refers to the set of information that an actor needs to process in order to perform a task (task-related information). In data pooling initiatives information is the core resource on which a set of tasks are eventually performed.

Second, even by relying on the Galbraith perspective, it becomes explicit how recent technological changes may have inverted some of Galbraith prescriptions. Galbraith suggests that organizational interventions should be oriented in two directions: either toward the reduction of information or toward the use of tools that reduce decision makers cognitive burden. Algorithms and statistical tools may be considered as being part of the latter category. However, in order to be effective information reducing tools, algorithms require more rather than less information. The greater the information abundance, the more the information system can be an adequate decision support thus being a credible tool to reduce cognitive burden. As such, when the target is alimenting an algorithm (or any form of statistical analysis), organizational design should be informed by the goal of generating more data rather than reducing them. Thus, in data pooling initiatives, data volume becomes an asset and not a challenge.

1.3.4.2 Velocity

While velocity is usually seen as a source of technical complexity and of a consequent need for costly standardization effort, it also has a bright side in reducing the need for formalization and in increasing cooperation across actors.

Velocity is characterized in many ways. Velocity may refer either to frequency of generation, i.e. frequency of signals, or to frequency of recording, i.e. frequency of captum, or to frequency of transmission of the capta to the capta pool (Kitchin & McArdle, 2016; Koutroumpis & Leiponen, 2013). Brynjolfsson & McAfee (2012) also refer to the increasing opportunity to benefit from frequency of signals and capta by generating frequent derivatives (e.g. real-time estimations of people flow).

When referred to signal and capta, velocity is commonly seen as a technological and organizational challenge as higher frequency usually calls for more complex technical solutions and especially for more accurate standards.

However, data velocity has other governance implications. When considering data velocity as the frequency of contribution of capta to the capta pool, it has governance implications both in terms of: 1) reducing the degree of formalization of data pooling initiatives in non-cooperative settings; and 2) favouring coordination in contribution across contributors in cooperative ones.

In a non-cooperative game theoretical perspective, increasing frequency of contribution of capta implies that contributing actors play more frequent and smaller pay-off games. Thus, the pay-off for an actor to betray (either from violating a mutual agreement to contribute or through the misuse of contributed capta) in each play, is much lower than the one in a unique play in a one-shot large contribution. Thus, ceteris paribus, frequency of contribution can be an explanatory factor for observing relational contracts rather than more formal ones (Baker, Gibbons, & Murphy, 2002).

In a cooperative setting, higher frequency of contribution can increase predictive knowledge (Puranam, Raveendran, & Knudsen, 2012). More frequent contribution in a pooling initiative, instead of a unique contribution of the same entity, increases confidence of other involved

actors that each contributor is making its part, thus in turn reducing uncertainty on whether the joint goal may be achieved.

1.3.4.3 Variety

Variety refers to the degree to which data differ in the way they are captured and stored. In other words, variety refers to the structural heterogeneity in a dataset (Gandomi & Haider, 2015, p. 138). The most common distinction distinguishes between structured, semistructured and unstructured "data" (Kitchin & McArdle, 2016) [3].

While from a technical point of view the challenge is to develop technical tools and algorithms able to univocally retrieve, combine and jointly analyse capta of different nature, variety of capta can also have implications in terms of organizational and governance design.

In order to capture when variety becomes strongly relevant for governance design of a data pooling initiative, it is important to make a distinction between two types of pooling. Employing the analogy with a structured column and row database, it is possible to distinguish between two different kinds of pooling. *Row pooling* and *column pooling*.

[3]Most of the speculations suggested in the current section, similarly apply to contexts where variety is not simply technical. Very similar implications may be observed in settings where different actors are endowed with different variables on the same data subject, even if they fall in one simple category (either structured, unstructured, semistructured).

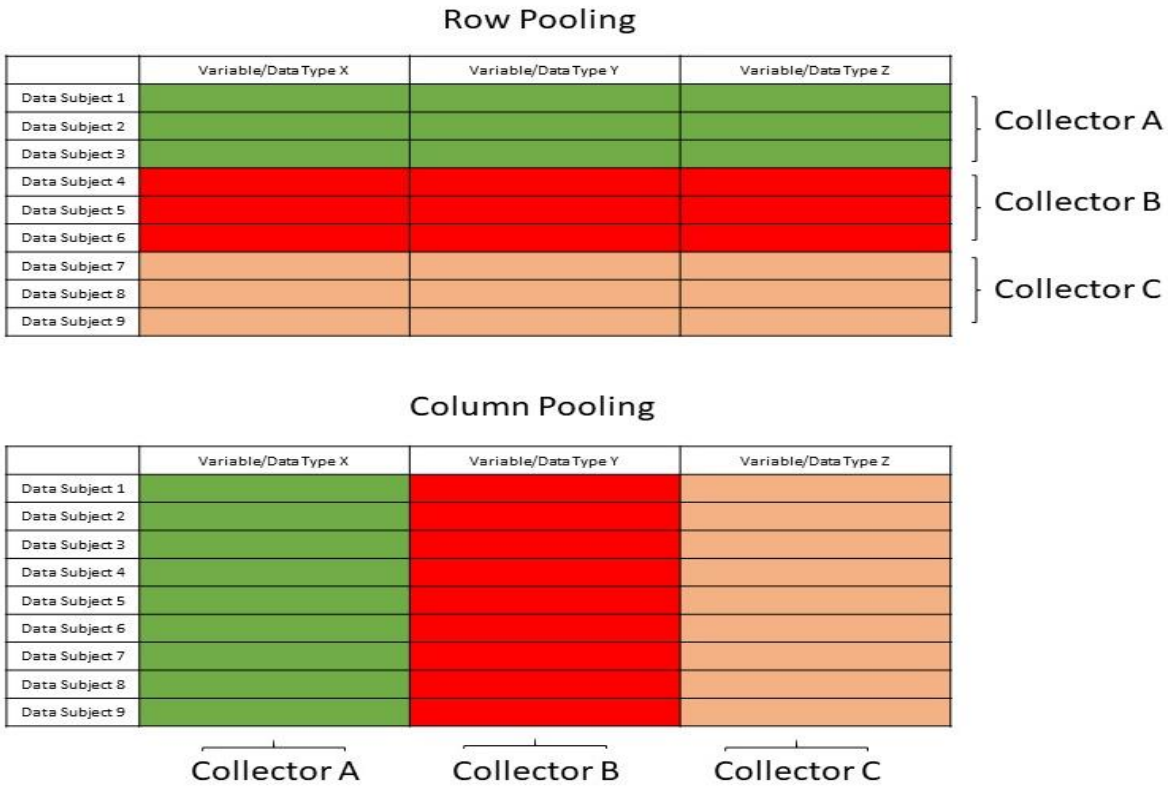

Figure 2 - Row Pooling and Column Pooling

In row pooling a data collector has access to multiple types of signals on the same data subject and consequently captures capta on the same data subject in multiple forms. For instance, in the process of care, a clinician in a hospital collects several types of capta on a single patient: 1) Structured capta, like the ones he uses to fill claims for patient reimbursement; 2) Unstructured capta, like text and images generated through the process of care; 3) Semistructured capta, like the ones originated through sensors. In a row pooling case, clinicians contribute “a patient” (or a set of patients) to the pool. In other words, following the analogy with a purely structured database, the contributing data collector contributes to the whole row. While even in the case of row pooling the whole pool has technical variety in terms of modes of collection of capta, this variety is instead orthogonal with respect to contributing collectors.

Instead, column pooling implies that access to signals of the same data subject is fragmented across multiple data collectors. Following the previous example, a column pool is a pool where each different clinician in different organizations collects different types of capta on the same data subject. Clinician in organization A collects structured data, clinician in organization B collects unstructured ones and clinician in organization C collects semistructured ones. In this type of setting, actors tend to be more heterogeneous in nature. In column pooling data variety is correlated with collectors' heterogeneity rather than being orthogonal.

In column pooling data variety gains more salience in terms of governance challenges for two reasons:

- In column pooling initiatives, actors tend to be more heterogenous in nature.
- In column pooling initiatives, the team production problem, i.e. the complexity to trace the link between contribution and outputs, becomes an extremely salient governance challenge.

Higher actors heterogeneity has several implications on the challenges that collaborative initiatives that pool data face. Primarily there are more marked coordination challenges as heterogenous actors may rely on different standards and practices. As such coordination between actors becomes more computationally complex. Heterogeneity of actors also implies heterogeneity in incentives and motives, in turn implying higher potential for conflict of interest between parties. Higher degree of potential for conflict of interest implies higher formalization and more participative decision making, with the presence of boards and committees, as participants may be concerned that organizational solutions that encompass an unique decision maker may be subject to the risk of prevalence of one of the specific interests (Chompalov, Genuth, & Shrum, 2002). Further, actors heterogeneity makes it extremely more complex for

an individual actor to monitor participants behaviours (Grandori, 2001). Even in this case a monitoring board might be needed, in order to have sufficient variety of actors to accurately appreciate behaviours. In this latter case, as the problem is not representativity of interest but monitoring, an independent board may be preferable.

Moreover, data variety in column pooling makes it more complex to trace the link between contribution and outputs, generating significant "team production" problems (Alchian & Demsetz, 1972).

In row type of pools, the link between contribution to the pool and benefits is more explicit than in column types of pools. In row types of pools, the homogeneity of contribution allows at least to rank order contribution. In column pools, the degree to which adding an additional data type or an additional "variable" in the analysis adds to the quality of the analysis may be far more complex to assess, especially ex ante. Even in a very narrow and single purpose pooling initiative, like a pooling initiative leading to a very simple multivariate regression, the degree to which an additional variable has explanatory/predictive power can only be understood ex post through the analysis itself, after actually pooling the data. This traceability problem becomes far more complex with emerging technologies where the contribution of each individual "column" may be even less intelligible.

1.4 DISCUSSION

In the previous sections I have explored how each characteristic of data may have implications in terms of governance challenges independently one from the other. In this section, I focus on some relevant phases of data pooling to explore how different features interact among them.

One of the core challenges in data pooling is the combination of capta in the capta pool, i.e. in the combination of capta collected by a multiplicity of data collectors in a unique pool that

allows to generate a high-quality derivative. The act of contributing capta into the pool can be considered as a transaction, i.e. as an exchange of good or service between technologically separable units (Williamson, 1981). Previous sections identify three characteristics of data that affect the complexity of transaction: data option value (multipurposedness), data sensitivity and non subtractability. Option value increases transaction uncertainty as potential uses of the transferred resource may not be fully anticipated. In case of sensitive data, the severity of the impact of potential misuses is furthered. In addition, non subtractability of the resource allows users to take advantage of all the potential uses, without having the need to choose one among the multiple purposes. In this respect, even when "misuse" is not the best allocation of the data resource, it is still possible to use data both for the best allocation and for misuse.

In addition to their impact on transaction uncertainty, both data non subtractability and sensitivity have other impacts on the transaction. Non subtractability allows for secondary transfers of data without any loss by the transferee. Assuming a market exchange, there is the practical risk that the buyer may resell data at a lower price as it did not suffer collection costs. As mentioned above, sensitive data also encompass the element of data non alienability. Despite the transfer of data, the consequences of misuse remain on the data subject. Thus, non alienability does not only affect the data subject, but by transitivity it affects the data collector. Consequently, in case of pooling the contributing collector may prefer to exert some control on use as it may also bear the negative externalities of potential misuse.

In light of this, for what concerns the contractual governance of the transaction, it is more plausible to observe associational contracts or the establishment of an independent legal entity such as an association or a foundation rather than exchange contracts.

Another core challenge in data pooling initiatives relates to access to and use of the capta pool, as a core trade-off emerges. While extending access to a multiplicity of actors may increase knowledge on potential uses, this extension may also lead to broader knowledge on potential misuses. As such, especially when dealing with sensitive data, a trade-off emerges between expanding access to realize the data potential and restricting access to protect both data subject from misuse and contributing collectors from opportunistic behaviour that may in turn weaken the ability of data collectors contributing to the pool to further collect data in future times. As such purely closed solutions as well as open ones seem to be not optimal in solving this trade-off. Thus, the different ways through which this trade-off is governed need to be better explored. How are pooling initiatives designed so that they balance the potential of multiplying the benefits that may be generated by the pool while preserving sensitivity? How are access rights and decision rights on access and use allocated in pooling initiatives?

Another core design challenge emerges with respect to incentivizing data collectors in contributing to the pool. As mentioned above, due to a multiplicity of factors, market-like solutions, where full rights on capta are simply transferred to another actor in exchange for a monetary payment, seem to be rarely possible. Further, uncertainty on potential uses and uncertainty on potential benefits make the establishment of any pricing/monetary mechanism to incentivize contribution to the pool an extremely complex endeavour.

As the primary concern for club good like data is ensuring its provision, how does pooling initiatives manage to incentivize contribution, as market contracts and price solutions may be inadequate due to the complex combination of option value, sensitivity and non-rivalry?

1.5 CHAPTER REFERENCES

Alchian, A. A., & Demsetz, H. (1972). Production, Information Costs, and Economic Organization. American Economic Review, 62(5), 777–795.

Buchanan, J. M. (1965). An Economic Theory of Clubs. Economica, 32(125), 1. https://doi.org/10.2307/2552442

Buchanan, J. M., & Yoon, Y. J. (2000). Symmetric Tragedies: Commons and Anticommons. The Journal of Law and Economics, 43(1), 1–14. https://doi.org/10.1086/467445

Chompalov, I., Genuth, J., & Shrum, W. (2002). The organization of scientific collaborations. Research Policy, 31(5), 749–767. https://doi.org/10.1016/S0048-7333(01)00145-7

Cornes, R., & Sandler, T. (1996). The Theory of Externalities, Public Goods and Club Goods.

Floridi, L. (2010). Information. Oxford University Press. https://doi.org/10.1093/actrade/9780199551378.001.0001

Frost, J., & Morner, M. (2010). Overcoming knowledge dilemmas: governing the creation, sharing and use of knowledge resources. International Journal of Strategic Change Management, 2(2/3), 172–199. https://doi.org/10.1504/IJSCM.2010.034413

Galbraith, J. R. (1974). Organization Design: An Information Processing View. Interfaces, 4(3), 28–36. https://doi.org/10.1287/inte.4.3.28

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Grandori, A. (1991). Negotiating efficient organization forms. Journal of Economic Behavior & Organization, 16(3), 319–340. https://doi.org/10.1016/0167-2681(91)90017-R

Grandori, A. (2001). Organization and Economic Behavior. https://doi.org/10.4324/9780203273814

Hardin, G. (1968). The Tragedy of the Commons. Science, 162(3859), 1243–1248. https://doi.org/10.1126/science.162.3859.1243

Heller, M. A. (1998). THE TRAGEDY OF THE ANTICOMMONS: PROPERTY IN THE TRANSITION FROM MARX TO MARKETS. HARVARD LAW REVIEW, 111(3), 621–688. Retrieved from https://www.degruyter.com/view/j/gj.2003.3.1/gj.2003.3.1.1081/gj.2003.3.1.1081.xml

Hess, C., & Ostrom, E. (2003). IDEAS, ARTIFACTS, AND FACILITIES: INFORMATION AS A COMMON-POOL RESOURCE. LAW AND CONTEMPORARY PROBLEMS, 66(1&2), 111–146. Retrieved from http://scholarship.law.duke.edu/lcp/vol66/iss1/5

Kitchin, R., & McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. Big Data & Society, 3(1), 1–10. https://doi.org/10.1177/2053951716631130

Koutroumpis, P., Leiponen, A., & Thomas, L. (2018). Data Strategy. In Academy of Management Global Proceedings. https://doi.org/10.5465/amgblproc.surrey.2018.0085.abs

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. Retrieved from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Martin, K. D., Borah, A., & Palmatier, R. W. (2017). Data Privacy: Effects on Customer and Firm Performance. Journal of Marketing, 81(1), 36–58. https://doi.org/10.1509/jm.15.0497

McGuire, M. (1972). Private Good Clubs and Public Good Clubs: Economic Models of Group Formation. The Swedish Journal of Economics, 74(1), 84. https://doi.org/10.2307/3439011

Miller, A. R., & Tucker, C. (2014). Privacy Protection , Personalized Medicine and Genetic Testing. Working Paper, 1–34.

Ostrom, E. (1990). Governing the Commons: The Evolution of Institutions for Collective Action. (Cambridge University Press, Ed.), Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge.

Ostrom, E. (2005). Understanding Institutional Diversity. Princeton, New Jersey: Princeton University Press.

Ostrom, E., & Hess, C. (2007). Understanding Knowledge as a Commons. Understanding Knowledge as a Commons From Theory to Practice (Vol. 15). https://doi.org/10.1002/asi

Ostrom, V., & Ostrom, E. (1977). Public Goods and Public Choices. Indiana University, Workshop in Political Theory and Policy Analysis. Retrieved from https://books.google.it/books?id=nSypXwAACAAJ

Puranam, P., Raveendran, M., & Knudsen, T. (2012). Organization design: the epistemic interdepencence perspective. Academy of Management Review, 37(3), 419–440.

Rose, C. (1986). The Comedy of the Commons: Custom, Commerce, and Inherently Public Property. The University of Chicago Law Review, 53(3), 711. https://doi.org/10.2307/1599583

Rumbold, J. M. M., & Pierscionek, B. K. (2018). What Are Data? A Categorization of the Data Sensitivity Spectrum. Big Data Research, 12, 49–59. https://doi.org/10.1016/j.bdr.2017.11.001

Strandburg, K. J., Frischmann, B. M., & Cui, C. (2014). The Rare Diseases Clinical Research Network and the Urea Cycle Disorders Consortium as Nested Knowledge Commons, (14), 155–207.

Strandburg, K. J., Frischmann, B. M., & Madison, M. J. (2017). Governing Knowledge Commons: An Appraisal. Governing Medical Knowledge Commons, 421–429.

Tutić, A. (2013). Experimental evidence on the theory of club goods. Rationality and Society, 25(1), 90–120. https://doi.org/10.1177/1043463112463874

Welch, E., Loaufi, S., Fusi, F., & Manzella, D. (2016). Institutional and Organizational Factors for Enabling Data Access, Exchange and Use in Genomics Organizations.

Williamson, O. E. (1981). The Economics of Organization: The Transaction Cost Approach. American Journal of Sociology, 87(3), 548–577. https://doi.org/10.1086/227496

2 UNDERSTANDING THE GOVERNANCE OF DATA POOLS: EVIDENCE FROM THE HEALTH SECTOR

2.1 INTRODUCTION

When referring to data and "big data", the attention from scholars and practitioners has been mainly devoted to the analytical challenges and opportunities related to the use of data and most of the exploration has been done at the intra-organizational level (Thomas & Leiponen, 2016). More limited attention has been given to phenomena of data creation, sharing, integration and use that involve more than one organization or multiple individuals that do not belong to a single organization.

Existing literature, especially in the field of management of information systems, that has focused on interorganizational initiatives intended for the creation, sharing, integration and use of data has limitedly focused on cases where data was the core object of transaction and the core good of interest (Adjerid, Adler-Milstein, & Angst, 2018), rather than simply being a complementary element to a primary transaction of a good (Elgarah et al., 2005) or being instrumental for coordination of action (Markus & Bui, 2012).

The fragmented literature that explores the governance of interorganizational initiatives which have data as a core good of interest tends to be highly descriptive and based on single case studies. The rare comparative efforts either explore very limited traits of governance, mostly forgetting elements of internal governance of the explored initiatives, or they fall short in capturing regularities across their design.

In order to fill this gap, I perform a comparative case study which comparatively assesses elements of consistency and variance in the way 15 cases of interorganizational attempts to

combine data are governed, conceptualizing their governance design as a configuration of property rights à la Ostrom, employing the Ostrom bundle of rights as a framework for the comparative analysis.

In the first place, the analysis of the cases helps in refining the bundle of rights originally conceived for the governance of digital libraries and knowledge commons to the specific setting of interorganizational initiatives that attempt to combine data. Also, thanks to this refinement, it allows to distinguish between three fundamentally different types of initiatives: Data Pools, Exchanges and Repositories.

After a rich identification of elements of consistency and variance in the way property rights (Contribution, Access, Extraction, Removal and Decision rights) are allocated across the analyzed cases, I identify some regularities across them: 1) Combining the way access rights are allocated and the mode through which decisions on extraction rights are taken, a typology of modes of governance of access and extraction emerges; 2) The way contribution rights and extraction rights configure across cases seem to suggest two prevalent models: Multiple Contributors - Internal Access & Single Contributor - External Access; 3) There is a suggestive U shaped relationship between the number of contributors and the probability of observing a polyarchic rather than a committee mode of decision for the allocation of extraction rights.

The paper contributes to the existing literature by increasing knowledge about the internal governance of interorganizational initiatives that attempt to combine data. It also contributes to identify how challenges that may derive from the nature of data involved in the pool (sensitive, non rival, option value) are in fact solved by the observed initiatives. The appreciation of elements of variance across initiatives is also highly generative for new research. Both performance implications of varying governance solutions and the identification of explanatory

antecedents are further research avenues that the comparative appreciation of diversity in design allows.

The paper is structured as follows. After a literature review in section 2, section 3 describes the method and the theoretical framework employed for the analysis and describes the case selection process. Section 4 identifies the elements of consistency and variance in the way property rights are allocated across the analyzed cases. Section 5 explores whether some configurations emerge and section 6 discusses how the analyzed initiatives address the core challenges that derive from the peculiar nature of data. Conclusions follow.

2.2 LITERATURE REVIEW

While the attention from scholars and practitioners has been mainly devoted to the analytical challenges and opportunities related to the use of data, more limited attention has been given to phenomena of data creation, sharing, integration and use, especially when these involve more than one organization or multiple individuals that do not belong to a single organization.

Existing literature, especially in the field of management of information systems, that has focused on the interorganizational design of initiatives intended for the creation, sharing, integration and use of data has given limited attention to cases where data was the core object of transaction and the core good of interest (Adjerid et al., 2018), rather than simply being a complementary element to a primary transaction of a good (Elgarah et al., 2005) or being instrumental for coordination of action between multiple organizations (Markus & Bui, 2012).

However, across different streams of literature there have been some attempts to capture how collaborative initiatives that have data as a core good are governed. While the majority of contributions are highly descriptive single case studies sparse especially across specialistic

health journals, other streams of research attempt to more systematically capture differences between cases.

The recently published book "Governing Medical Knowledge Commons" (Strandburg, Frischmann, & Madison, 2017b) includes a rich set of cases in the health sector that involve the sharing/pooling of data between multiple actors employing an extension of the IAD framework proposed by Ostrom (2005). While it employs a coherent theoretical framework within which cases are analyzed by different authors contributing to the book, the core purpose of the contribution is, in the tradition of the commons literature, to show that self-organized commons are a valid alternative solution to markets and government intervention to solve the Hardin "tragedy of the commons". Its extension in the governance of knowledge production is again intended to challenge the idea that an "optimal amount" of knowledge cannot be generated in absence of either: 1) "markets supported by intellectual property rights directed to exclusivity and ownership"; or 2) governments that "intervene in markets (or avoid markets) in various ways to sponsor and subsidize innovation" (Strandburg et al., 2017b, p. 13). While their contribution consistently shows that a self-organizing alternative is an alternative that is empirically frequent, it does not do much in terms of comparatively capturing how these initiatives are governed appreciating elements of consistency and variance through which this self-organization takes place.

More comparative efforts are similarly highly fragmented across literatures.

Comparative attempts intended to capture differences in the governance of initiatives that attempt to combine data across multiple organizations have been performed within the health informatics literature. Both Vest, Campion, & Kaushal (2013) and Everson (2017) attempt to identify a typology of different forms of health information exchanges. However, the

governance traits actually taken into account to capture variance in governance design have been extremely limited. In both cases, the most relevant governance trait explored is what is the entity that provides the convening role for the health information exchange. Following Vest et al. (2013) , Everson (2017) distinguishes exchanges as Vendor based (the convener is the vendor of the Electronic health record software), Enterprise based (the convener is large healthcare system), Community based (the convener is a neutral third party organization) and argues that different conveners in turn affect the type of members that are willing to take part or are accepted to participate in the HIE. However, except from these two governance traits, no other element has been explored, leaving especially unexplored the internal governance of this type of initiatives.

Susha, Janssen, & Verhulst (2017a) explore instead public-private data collaboratives and advance a rich characterization of the variance of 10 data collaboratives across a multiplicity of sectors (Health, Environment, Education, Infrastructures). However, while they are extremely able and rich in capturing the impressive variance between the different collaboratives, they then fall short in identifying elements of consistency across initiatives.

In this paper I attempt to enrich knowledge on the internal governance of interorganizational attempts to combine data by performing a comparative case study, which apart from capturing the rich variance of the observed initiatives, also explores elements of consistency and especially explores regularities on the way certain modes of allocation of property rights cooccur across them.

2.3 METHOD AND THEORETICAL FRAMEWORK

The method employed in the present paper is a comparative case study. Cases identified have been initially coded against the set of property and decision rights identified in Ostrom & Hess

(2007). Through the coding of initiatives, the initial set of property rights identified in Ostrom & Hess (2007) has been refined to better capture the specificity of data pooling initiatives. Each case analyzed has been then coded against the refined set of property rights. For each property right I have explored what are the elements of consistency and variance across the explored cases. Across elements of variance I also attempt to identify whether some configurations emerge in the way certain rights are allocated.

Case Selection

Being a knowledge and data intensive sector (Oderkirk & Ronchi, 2017), the health sector has repeatedly attempted with varying degrees of success to pool and link data between different actors and organizations in the system (Vest & Gamm, 2010). In addition, the health sector is characterized by the diffuse custom of reporting descriptive narratives of collaborative attempts to combine data in peer reviewed field journals. I relied on this peculiarity in order to identify the cases to be analyzed. I have searched for peer reviewed articles in multiple databases: Scopus, Web of Science, Ebsco and PubMed. The terms employed to search for the different cases were: regist* AND governance; data AND governance AND health; "data governance" AND health; "health information organization" OR "health information exchange"; data pool AND health; data consorti* AND health. All the papers that described a single case of an attempt to pool data in the health sector have been considered as a case to be analyzed. The result of the search gave 1240 papers of which 26 identified single *case studies* describing an initiative that attempts to combine data across multiple actors. Given that some of the paper covered the same *case*, the actual number of *cases* identified is 15. Most of the papers were not exhaustive in describing the governance of the data pool. As such, the information needed to code the cases have also derived from primary sources identified in the “data pool” website (Statute, Frequently Answered Questions, Internal Bylaws).

2.3.1 Refining Property Rights

Property rights have been defined as "an enforceable authority to undertake particular actions in a specific domain. Property rights define actions that individuals may take in relation to other individuals regarding some "thing". If one individual has right, someone else has a commensurate duty to observe that right." (Hess & Ostrom, 2003, p. 124).

In the present section, my goal is to rely on the identified cases to refine the definition of the property rights that are allocated in data pooling initiatives.

The bundle of rights originally identified within the Institutional Analysis and Development (IAD) framework emerged for the analysis of natural resources (Schlager and Ostrom, 1992). As such the rights that have been initially envisioned are the following:

1) Access: The right to enter a defined physical area and enjoy non subtractive benefits (e.g. canoeing in a lake or sitting in the sun)
2) Extraction: The right to obtain resource units of product of a resource system (e.g. catch fishes or divert water)
3) Management: The right to regulate internal use patterns and transform the resource by making improvements
4) Exclusion: The right to determine who will have access rights and withdrawal rights and how those rights may be transferred
5) Alienation: the right to sell or lease management and exclusion rights

In their attempt to extend the bundle of rights to the setting of digital libraries, Ostrom and Hess (2007), envision a slightly different set of property rights:

(1) Access: The right to enter a defined physical area and enjoy non- subtractive benefits
(2) Contribution: The right to contribute to the content
(3) Extraction: The right to obtain resource units or products of a resource system
(4) Removal: The right to remove one's artefact from the resource
(5) Management/Participation: The right to regulate internal use patterns and transform the resource by making improvements
(6) Exclusion: The right to determine who will have access, contribution, extraction, and removal rights and how those rights may be transferred
(7) Alienation: The right to sell or lease extraction, management/participation, and exclusion rights

When considering the governance of a digital library and in general of knowledge commons, Ostrom & Hess (Hess & Ostrom, 2003; E. Ostrom & Hess, 2007) add two rights to the set of rights that apply to natural resources: Contribution rights and Removal rights. In the first case, a core distinction between the natural resource settings previously explored and a digital library is that, as the good object of governance is primarily excludable, there is the need to contribute it in order to generate the pool. For instance, an academic article needs to be primarily contributed before being part of a digital library. The natural resources usually explored in the commons literature are instead already present in nature. Thus, contribution rights were not originally considered, while they become a relevant right in the governance of digital libraries.

The inclusion of removal rights is instead due to non subtractability of the digital good object of governance. When the good object of governance is non subtractable, extraction and removal rights need to be distinguished. If the resource unit was subtractable, the simple act of using it would have removed it from the resource system. However, as digital resource units are non subtractable, their removal should be an intended act that is independent from its use.

Through the analysis of the cases, the set of rights have been further refined to better fit to the specific setting of data pooling initiatives and to better appreciate differences in the way different initiatives are governed. Exclusion and Management rights have been conflated into a unique category of decision rights. Alienation rights have not been considered in this setting, as in no case explored there were provisions on the possibility to sell or lease decision rights (management and exclusion) or extraction rights. Other rights, which have been considered as unique in the Hess & Ostrom framework, have been instead “fragmented” following the fragmentation of rights that is observed in the analysed initiative. For instance, with respect to Access rights, simply looking at a unique right of access would have missed the substantial difference between which are the actors that have access exclusively to the patient records

(*capta*) that they collected and the ones that have access to the whole set of capta collected and contributed to the pool.

Table 1 lists all the rights that have been taken into consideration in the coding of cases.

Type of right	Definition	Object of the right
Contribution	The right to contribute to the content	Contribution of patient records
Access	The right to enter a defined area	Access to own collected database
		Access the whole database (pool)
Extraction	The right to obtain resource units or products of a resource system	Extraction of patient records on a data subject
		Extraction of data derivatives (summary data derived from the pool)
		Recognition of authorship rights
Exclusion and Management (Decision Rights)	The right to determine who will have contribution, access and extraction rights	Determine contribution rights
		Determine access rights
		Determine extraction rights
Removal	The right to remove one's artefact from the resource	Removal of patient records on an individual
		Removal of own collected database

Table 1 - Set of property rights employed in the initial coding of cases

2.3.2 Restricting Cases: Differences between Exchanges, Repositories and Pools

Despite the apparent similarity between interorganizational initiatives that combine data, a fundamental difference emerges across three types of initiatives. After a first analysis of the identified cases, it became explicit that cases where not fully comparable among them as they were pursuing different purposes and, consequently, right allocation differed too markedly across them, as the stream of benefits come from different types of rights on the resource. The core differences identified across the three types of initiatives (Exchanges, Repositories, Pools) are resumed in Table 2.

Exchanges are mostly designed for sequential coordination purposes. As patients may receive care across different organizations, health information exchanges have the purpose of combining different information so that they could be available at the point of care. The core goal of exchanges is to grant to the caregiver access to the best amount of information on the individual patient she is taking care of. Four of the identified cases fall under this category (Health Records Bank Startup; Nehen Portal; EPACCS; Connect Virginia).

Repositories are instead mostly designed for reuse purposes and to reduce search and transaction costs for upstream users. In the set of cases identified, WWARN is a representative case of a repository. Researchers that study malaria can contribute some databases from clinical trials they performed on malaria. Then, after the contribution of researchers' data in the repository, other actors may be allowed (through different mechanisms of authorization) to reuse the capta contributed. In repositories, the contributor does not have to adapt its mode of collection with respect to other contributors. The possibility for the user to employ data coming from different contributors is due to "randomness" in compatibility or to ex-post homogenization efforts, when possible.

Data pools have instead the purpose to jointly generate a data derivative. A data derivative is not a mere combination of data, but it is data that are generated from the processing of other data. For this reason, homogeneity of data to be analysed need to be also semantic other than technical. Cases like MSBase are initiatives where multiple clinical centers contribute data to then generate different types of statistics, in the forms of benchmark or in the forms of statistical analysis for research purposes.

	Exchange	Repository	Pool
Purpose	Availability of a single patient record at the point of care	Data Reuse; Reduce transaction and search costs for users	Generate one or more data derivatives
Standardization needed	Technical	None (Compatibility is accidental, not designed ex ante)	Technical and semantical
Type of Interdependence	Sequential	Absent	Pooled
Right that generates stream of benefits	Access to capta on a single patient	Access on multiplicity of patient records and extraction rights	Extraction rights and access to derivative
Cases	Health Records Bank Startup; Nehen Portal; Epaccs; Connect Virginia	WWARN	VCOR; South Minnesota Beacon CDR; PCCR; EMBT; BigMouth; QPR; Kadoore Biobank; MSBAse; PORTAL; UK Biobank
Existing Literature on Governance	Health Information Exchanges and Interorganizational Coordination Hub	Data Collaboratives (Patent Pools?)	Absent

Table 2 - Types of interorganizational initiatives that combine data

This distinction also allows to reframe existing comparative contributions that analysed the governance of effort to combine data across multiple actors.

As obviously suggested by the name, the comparative efforts on Health Information Exchanges (Everson, 2017; Vest et al., 2013) capture the governance of Exchanges. De facto, given that the purpose is coordination of care, they are not much different from the already explored interorganizational coordination hubs (Markus & Bui, 2012). Instead, through a deeper exploration of the cases explored in their paper, the data collaboratives analysed by Susha et al. (2017b; 2017a), belong to the category of repositories. In none of the cases they explored there was interdependence and coordination in data generation. As the core purpose of repositories

is the reduction of search and transaction costs for upstream users, they resemble the purpose of Patent Pools.

To my knowledge, data pools have instead received far less attention. Also for this reason, the following sections will focus on the 10 cases of data pools.

2.4 Elements of Variance and Stability in Property Rights Allocation

2.4.1 Contribution Rights

Object of the right	Entity (potentially) entitled of the right	Stable features among pools	Varying features among pools
Contribution of Patient Records	• Data Subject • Surveyor • Team (represented by a Principal Investigator) • Organization (Hospital or health system)	Closed contribution Contribution conditional on consent	«Level» of allocation of contribution rights and type of contributing actor Number of contributors

A feature that is stable across all initiatives is that contribution rights are closed. In all the explored cases, it would be impossible to contribute patient records to the pool without an authorization. In all initiatives, contributors need to undergo an approval process in order to be able to contribute. Closure of contribution is intended to avoid pollution, i.e. to avoid improper contribution that might in turn affect the global quality of the pool and of the derived derivatives. Data pools may be subject to contamination derived from unintentional pollution, for instance through mistakes in collection by an unexperienced contributor, or from deliberate pollution (Mindel, Mathiassen, & Rai, 2018).

A second element that is consistent across initiatives is that, when contribution does not come directly from the data subject (i.e. the patient), the possibility to contribute is conditional on informed consent by the data subject. All data pools explored are pools that contain health data,

a type of highly sensitive data. As sensitivity encompasses inalienability (Koutroumpis et al., 2018), i.e. the consequence of use of patients records are still borne by the data subject, consent on contribution from the data subject becomes fundamental and ethical and it is also usually required by law.

Strong variety is instead observed with respect to which and how many actors enjoy contribution rights.

The type of actor that is entitled to contribute the pool differs markedly across different initiatives. Right to contribute patient records is granted to:

- The data subject herself, as in the case of SAPCON (myapnea.org). In this initiative, patients affected by sleep apnea directly answer to an online survey, built ex ante by a group of researchers.
- A team of ad hoc data collectors that are under an employment relation with a single organization that generates the pool. For instance, both the Kadoore Biobank and UK Biobank employ “survey teams” that perform contribution toward the Biobank, by performing ad hoc visits and taking biological samples from patients.
- A group of organizations that collect data during the care process. Within this group there is a more subtle distinction between “Principal Investigator-Centered” initiatives and “Organization-Centered” ones. In the former type, authorization to contribute is given to an individual, the principal investigator, who then asks his own organization the authorization to contribute (as in MSBase Registry). In the latter, the authorization to contribute is granted to the organization legal person and then delegated to a representative individual (like in Big Mouth).

In other cases contribution rights are more hybridlike distributed. For instance, in the Quebec Pain Registry contribution is instead combined between patients and hospital nurses delegated by each participating hospital.

Number of contributing organizations also markedly varies. When contribution comes from different organizations, across the identified cases, the number of contributing organizations range from 5 (in the Quebec Pain Registry) to the more than 500 in the EBMT registry.

2.4.2 Access Rights

Object of the right	Entity (potentially) entitled of the right	Stable features among pools	Varying features among pools
Access own collected database	• Contributor	Each contributor can have access to its own collected database	

Access the whole database (pool)	• Data Steward:
	  • Direct employees of the pool
	  • External providers
	• No one
		No contributor has access to the whole database	Federated vs Centralized
			
			Exclusive access to contributors or external access

Allocation of access rights show both elements of stability and variance.

Across all initiatives, each contributor has access to its own set of contributed patient record.

However, a markedly different scenario emerges with respect to access rights to the whole pool. Across all cases that have multiple contributing organizations, none of them is entitled with the right of having access to the full pool. In some initiatives, like the iCare2 PCCR and the PORTAL network, access to the whole pool is never granted to any actor (federated model). In the other cases, only a "data steward" is entitled to access the whole pool. If the pooling initiative has an internal staff, the role of data steward is granted to members of the internal staff, as in the case of EBMT and the Quebec Pain Registry. Otherwise, in absence of internal

employees, access is instead granted to an external organization which has the only exclusive role of being a data steward, as in the case of the South Minnesota Beacon Community Clinical Data Repository. They decided to rely on the Regenstrief Institute, which is the only actor who has the right to access the whole pool and perform analysis on it.

For all actors that are not the designated data steward, the possibility to access to the pool is contingent on an approval process and usually linked to extraction rights, i.e. to the possibility to perform an analysis on the pool. Further, even after approval, the full access to the pool is rarely granted, i.e. the approval process also encompasses the determination of which subsection of the pool can be accessed by the authorized actor.

2.4.3 The Inverted Relationship between Access vs Extraction Rights

In data pooling initiatives, access and extraction rights are not necessarily jointly assigned. Further, while in the governance of natural resources the number of actors that enjoy access rights is usually greater than the number of actors who are granted extraction rights, due to differences in the nature of the governed resource, in data pooling initiatives this relationship is inverted.

In the governance of natural resources access rights are defined as the non-subtractive use of the resource system. For instance, in the governance of a lake, a group of tourists simply accessing a lake without fishing, are considered actors that enjoy access rights to the resource system (the lake) but are not endowed with extraction rights on the resource units contained and derived from it (the fishes). However, the lake example (and several other natural resources) has two core differences with respect to the data pools I am exploring in this paper. First, both the lake and the fishes do not have any trait of sensitivity. Second, fishes (the resource units) are rival goods. Use by one actor prevent any other actor from both fishing it and especially

eating it or gaining money through the sale of it. As such, when a natural resource is object of governance, the allocation of access rights is broader than extraction rights, i.e. the number of actors that enjoy access rights tend to be broader than the number of actors that enjoy extraction rights. In data pool, sensitivity and non-rivalry of the resource object of governance may invert this relationship. Mere access to the pool may violate privacy of data subject. As such, access rights to the pool tend to be minimized, i.e. restricted to the minimum set of actors. However, due to non-rivalry of digital patient records and of data derivatives, extraction rights may be instead more broadly assigned and especially access to derivatives can be granted to a broader number of actors than the ones that are granted access to the pool.

2.4.4 Extraction Rights

Object of the right	Entity (potentially) entitled of the right	Stable features among pools	Varying features among pools

Extraction of data derivatives	• Contributors
	• Data Stewards
	• External actors:
	  • Private
	  • Public
		Extraction rights to contributors are always granted after an approval process	Extraction granted to external actors vs only internal
			
			Extraction rights linked with access right vs non linked

Extraction rights are considered as the right to generate data derivatives from the pool.

Except for some cases where some extraction rights are allocated in advance to the actors in charge to perform some predetermined analyses on the pool, the possibility to extract a derivative tends to be allocated ad hoc in all cases observed. The possibility to extract a derivative from the pool is decided through alternative decision-making mechanisms, either through committee or polyarchy (a distinction that will be explored in the next subsection on decision rights).

Pooling initiatives differ with respect to whether extraction rights are linked to access rights. In the majority of the cases explored, extraction rights are linked to (partial) access rights**.** The

actor who is granted extraction rights is also granted access rights to the section of the pool needed to perform the authorized analysis. Instead, other cases separate access and extraction rights**.** Two different solutions seem to emerge, an organizational and a technical one: either through delegation of analysis to a data steward (as in VCOR) or through the development of technical means that disentangle access and extraction rights (as in PORTAL). In the VCOR case, access to the pool, i.e. to patient level micro-data is never granted, even when extraction rights are granted to a registry participant. The actor endowed with the right of extraction is either granted access to already aggregated data (which already configure as data derivatives) or is allowed to determine which analysis shall be performed by the internal staff of VCOR. A technological solution can be instead found in PORTAL. POPMEDNET, the technology that is employed in the PORTAL network and that is employed across all the pooling initiatives under the broader PCORI initiative, allows actors to simply send the code for the analysis to each member/contributor of interest for the analysis. The analysis is then separately performed on the set of capta of each contributor and then integrated into a unique analytical result. In this case the possibility to extract a derivative becomes independent from access, through a technological mean. An intermediate solution can be found in the iCare2 PCCR initiative. In PCCR, the access committee determines whether extraction rights are linked to access rights or extraction is performed by internal staff.

Initiatives also differ with respect to the allocation of access and extraction rights to non-contributors.

In cases like MSBase and in the EMBT registry, partial access and the possibility to extract derivatives from the pool is exclusively granted to contributors. Having contributed patient records to the pool is a necessary condition for enjoying access and extraction rights. Exclusivity can be interpreted as an incentive mechanism through which contribution is

incentivized in contexts where there are no alternative mechanisms (for lack of resources or for complexity in structuring alternative types of incentive mechanism, like prices) and collection is fragmented.

Extraction rights to external actors is instead granted by both biobanks and by the Quebec Pain Registry. The possibility to get extraction rights from the UK Biobank is open to any bonafide researcher (against a service fee). As the UK Biobank does not have any internal researcher, opening up access to external researchers is the only mechanism through which it can ensure the use of the pool and to partially ensure the sustainability of its activities[4].

The other initiatives that grant extraction rights to non-contributors, while granting access and extraction rights to both contributors and non-contributors, discriminate between the two categories. The Quebec Pain Registry has lower application and access fees for contributors rather than for noncontributors[5]. For the Kadoore Biobank, instead, time discrimination applies. The Kadoore Biobank, being born as a research project, has internal researchers. Thus, internal researchers of the Kadoore Biobank have exclusive access to the pool for a period of time, and only afterward they are open to external actors, but only through collaboration.

2.4.5 Decision Rights on Extraction and Access

Object of the right	Entity (potentially) entitled of the right	Stable features among pools	Varying features among pools
Determine allocation of access and extraction right	• Contributors • External experts	Decision to allocate access and extraction rights are always taken collectively	Decision by Committee vs Decision by Polyarchy

[4] Except from access fees, the UK Biobank has been substantially funded for a total of 244 million £ mostly by the Medical Research Council, the Wellcome Trust, The UK Department of Health.
[5] The Quebec Pain Registry also applies a discrimination between Academic and Industry non-contributors.

Decision rights on access and extraction also show elements of stability and variance across the analysed cases.

Primarily, it is interesting to note that determination of access and extraction rights is never centralized in a unique individual, even in the cases where collection takes place within a unique organization. Across all the explored cases, determination of access and extraction rights always take place through collective decision making.

Cases explored show two different modes of collective decision making in granting access and extraction rights to the pool: by committee or by polyarchy (Sah & Stiglitz, 1998).

By committee means that whoever wants to access the pool (or part of it except from own contribution) needs to get the approval of a committee which determines whether and how to grant access and extraction rights to the requestor.

In a polyarchic solution, the decision to grant access and extraction rights is not delegated through a committee but is left to each collector, which directly determines whether to grant access to each requestor.

2.4.6 Removal Rights

The way removal rights are allocated across actors is instead more homogenous across cases.

Two types of removal rights are allocated to the different actors. Data subject enjoys removal rights with respect to all the capta referring to her, while data contributor enjoys removal rights on all the capta the data contributor contributed to the pool.

Data subjects enjoy removal rights with no penalty across all cases. The possibility for a data subject to enjoy removal rights may be explained by the fact that data subjects need protection against data option value. Even though a data subject gave initial consent to the contribution of

capta on her to the pool, the possibility to remove it from the pool afterward is a form of protection of the data subject in case new undesired potential uses of the pool emerge. In absence of removal rights, a data subject may be less willing to give consent to contribution in the first place. The same principle applies for data contributors. Contributing organizations are always allowed to remove the patient records they contributed.

2.4.7 Modes of Governance for Access and Extraction

		Decision rights on extraction rights allocation	
Access rights allocation		Polyarchic	By Commitee
	Centralized	Facilitating Pools (BigMouth; PCCR) or Substudy pools (EMBT; MSBase; SAPCON)	Commitee Application pools (UK Biobank, Kadoore Biobank, VCOR, Southesteearn Minnesota Beacon; QPR)
	Federated	Query based pools (PORTAL)	

An interesting combination emerges with respect to how access rights are initially allocated and how decisions on granting access and extraction rights are taken.

As mentioned above, access rights may be centralized (meaning that there is at least an actor that has full access rights to the pool) or federated (meaning that no actor has full access to the pool). In the federated case the pool can be considered as “potential”. Every actor involved in a federated pool has homogeneously collected patient records so that they can be easily combined, but they are technically not located in the same place or joined. Decisions to grant access and extraction rights are instead either taken by committee or by polyarchy.

Combining the two dimensions, different modes of governance emerge.

2.4.7.1 Centralized and Polyarchic: Facilitating Pools and Sub-Study Pools

Under the combination of centralized access and polyarchic decision rights on access, two different forms are identified: facilitating pools and sub-study based pools.

2.4.7.1.1 Facilitating Pools (Centralized, Polyarchic)

In facilitating pools data are centralized to serve the core purpose of facilitating a first evaluation of feasibility of certain analyses and to perform extremely simple count or statistical summaries on the pool. The pooled data are centralized and used exclusively to evaluate the feasibility of certain studies. Once feasibility is established, access and extraction rights are then negotiated across all the members. This solution can be defined as a facilitating pool. The purpose is to reduce the cost of evaluating the basic feasibility of an analysis.

An example of a facilitating pool is Big Mouth. In Big Mouth, the data pool contains all the capta captured by all the 8 participating institutions on their dental patients. However, the pool is employed to estimate the feasibility (especially in terms of numerosity of cases) of future studies and to perform basic predetermined analysis. The tool to perform counts and analysis is available to every contributing organization. However, if a contributing organization wants to perform a more complex analysis on the capta contained in the pool, it needs to negotiate access with each of the other contributors, including the terms of sharing and the potential ways through which sharing is recognized (for instance by establishing modes of authorship allocation).

2.4.7.1.2 Sub-Study Based Pools (Centralized, Polyarchic)

In sub-study pools data are centralized to ensure the performance of a set of core analyses performed by the internal staff of the organization to which the data are contributed. Except from the core analysis, each actor who contributed capta is allowed to propose a certain use to

all the other contributors, which, for each proposal, determine whether to grant access and extraction rights to the proposer.

An example of a sub-study based pool is MSBase. In MSBase, all collecting contributors contribute patient records to the centralized pool, which is employed to perform the "core analyses" (mostly a main research publication and a set of reports for Pharmaceutical companies) which are performed by the internal staff of MSBase under the guidance of the Scientific Leadership Group. However, each contributor is allowed to propose a sub-study to other selected contributors, which in turn can decide to join the sub-study or not. The actor that proposed the sub-study is allowed to access and extract a derivative only from patient records collected by the ones that accepted to join the sub-study.

2.4.7.2 Query Based Pools (Federated, Polyarchic)

In query-based pools the whole capta pool is never centralized under the control of any single actor. In this respect the pool is simply an abstract concept. It is a "potential" pool since collectors have decided to homogeneously capture capta. Each contributing actor can send a query (i.e. a code that runs on all the other contributor datasets), to other contributors, which are free to accept or reject that the query is performed on their own capta.

An empirical example of this solution is the PORTAL (Patient Outcomes Research To Advance Learning) initiative, a clinical research network under PCORnet, a network or research network sponsored by the PCORI (Patient-Centered Outcomes Research Institute). All initiatives within PCORnet rely on the POPMEDNET technology, which embodies the mechanisms mentioned above. After a steering committee approval, each contributor to the PORTAL initiatives can send to the other contributors its query of interest and each member is free to accept or decline that the query is performed on its own data.

2.4.7.3 Committee Application Pools (Centralized, Committee Based)

The broadest set of cases fall instead in the category where access rights are centralized to at least one unique actor and determination of access and extraction rights on a subset of the pool for the others is determined by a Committee. In these cases, the committee is delegated to determine who can extract derivatives from the pool after evaluating the quality and appropriateness of the proposed use.

Several analysed initiatives correspond to this category: the VCOR, the SE Minnesota Beacon CDR and the two analysed Biobanks. Interestingly, the Kadoore Biobank has two access committees. One access committee is composed by internal members of the Kadoore Biobank. The second access committee, composed by external experts, is instead involved in deciding to grant extraction rights only in the case the first committee refuses it for an unjustified reason.

2.5 DO CONFIGURATIONS EMERGE?

When analysing the allocation of rights jointly and not individually, two types of consistencies in organizational design seem to emerge:

1) There are two modes through which contribution rights and extraction rights are combined:
    a. Multiple Contributors - Internal Access *(MC-IA)*
    b. Single Contributors - External Access *(SC-EA)*.
2) In Multiple Contributors initiatives there is a U-shaped relationship between the numbers of contributors and the probability to observe a polyarchic mode of decision making to allocate extraction rights.

2.5.1 Configurations of Contribution and Extraction Rights

For what concerns the relationship between the diffusion of contribution rights and the diffusion of extraction rights, in cases where contributors are a multiplicity of organizations, the possibility to extract a derivative is rarely granted to external actors[6]. Contrarily, when collection is performed within a single organization (by surveyor under an employment contract with the collecting organization), the possibility to extract a derivative is also granted to external actors (usually against the payment of a (two tier) fee[7]).

Two complementary explanations apply to the emergence of these types of configurations. On the one side, in MC-IA initiatives, leaving the exclusive possibility to extract derivatives to contributors acts as an incentive mechanism in compensation of the costs of data collection and contribution that contributors have suffered. On the other side, a multiplicity of contributors that are also users ensure enough variety of knowledge allowing the realization of data option value.

Instead, for SC-EA initiatives, which have performed collection through employment contracts, exclusivity may not be sufficient or needed for incentive purposes and granting access to external actors against a fee may also help covering the initial costs of collection and maintenance costs. In addition, SC-EA might not be "endowed" with a sufficient internal user base. Thus, they might need to rely on a broader external user base, and on a consequent broader knowledge base, in order to maximize data option value[8] (Mayer-Schönberger & Cukier, 2013b). The UK Biobank is emblematic of the latter mechanism. Not having internal researchers, it needs external users to realize data value. The Kadoore Biobank has instead

[6] The only exception is the Quebec Pain Registry, which is unusual due its very limited number of contributing organizations.
[7] Fees are usually established as follows, a fixed application fee and a cost recovery fee.
[8] It should also be noted that the interest on maximization of option value may come from funding actors, as in the case of Kadoore Biobank, where openness after a certain embargo period is required by the funders.

internal researchers. However, variety of knowledge may be non-sufficient. In fact, after an exclusivity period where internal researchers have exclusive extraction rights, as knowledge variety may result insufficient, the Kadoore Biobank also opens up to external actors. However, internal researchers still keep parts of the stream of academic benefits when opening access, by granting access to external actors only contingent on having an academic collaboration.

2.5.2 Relationship between Number of Contributors and Modes of Decision Making

Within initiatives with multiple contributors, another interesting relation refers to the number of contributors involved and the mode through which extraction rights are granted. Polyarchic mode seems to be employed both to initiatives with a small set of contributors and to the ones with a large number of contributors. However, representative mechanisms like data access committees are employed in midsize initiatives. To explain this combination, it is important to capture what is the rationale that may guide the use of a data access committee or a polyarchic solution.

The core purpose of imposing these modes of decision making is to guarantee to contributing actors that contributed patient records will not be used for undesired purposes. Thus, a core trade-off that emerges is to find a mechanism that combines efficiency with the ability to control secondary uses by each actor.

In the case of very small initiatives setting up a representative body may be inefficient. It is more efficient for an actor to negotiate ad hoc with every other contributor that is part of the pool. Thus, in very small-scale initiatives there is no value in simplifying the process of granting extraction rights through a committee.

In midsize initiatives, the mechanism mentioned above would be too demanding for the requestor. As such, it becomes more reasonable to concentrate the decision process within a unique body, which is usually designed to include a representative for each contributor.

In large scale initiatives, instead, a representative committee runs the risk of not being representative enough, so that a contributor may comfortably rely on its decisions to ensure that no secondary undesired use is made. In other words, either the decision-making body becomes too large to the point of becoming ineffective or alternative modes are employed to ensure that actors are not concerned about secondary use. Both MSBase and the EBMT registry employ in fact a two layered mechanism. A data access committee makes the first discrimination on who can enjoy extraction rights but then each contributor is allowed to confirm or reject the authorization on the set of patient records that it has collected and contributed. This mechanism also avoids the problem of reaching consensus at aggregate level. Even if the majority of contributors is not in favour of combining data, it is still possible to generate a smaller pool.

To sum up, the design of modes of allocation of decision rights in data pooling initiatives tries to balance the interests of the users and of the contributors. Employing a data access committee facilitates the process of requesting extraction rights for the requesting user, having a unique point of access. However, after a certain size in terms of number of contributors, full delegation of decision rights to a centralized committee may instead become problematic for contributors as the access committee may be not representative enough of their interests. Under this condition, a polyarchic mode of deciding to grant extraction rights may be more attractive for contributors and thus more conducive to ensure and sustain wide contribution.

2.6 DISCUSSION: GOVERNING CORE DATA GOVERNANCE CHALLENGES

Digital data in health can be characterised as non-rival, sensitive and full of option value.

The copresence of these features creates two fundamental governance challenges. On the one side, governance design needs to balance the potential value that may be generated by data option value, which is usually enhanced by extending users, with the risk of violation of

sensitive data and their undesirable secondary use, which may markedly endanger the sustainability of these initiatives. On the other side, as the copresence of sensitivity, non-rivalry and option value may limit the set of available solutions for incentivizing contribution, collaborative initiatives need to identify ways to solve this governance challenge.

Building on the analysis performed in the previous sections, in this section I try to identify which solutions are implemented by the observed data pooling initiatives to address the aforementioned challenges.

2.6.1 Governing Trade-off in Access

How data pooling initiatives govern the trade-off in access generated by the copresence of digital data option value, non-rivalry and sensitivity? Across all cases, access is always mediated through an approval process. The possibility to access and to extract a derivative from the pool is never unconditionally open. Further, the allocation access rights to the whole pool is “minimised”. As shown in the previous section, access rights to the pool are either not granted at all or granted to a "data steward" which is independent from the contributors. The role of data steward is usually granted either to the internal staff, if the pool has some, or to an external provider which has the exclusive role of being a data steward.

Instead, when access is granted to users different from the data steward, it is only granted to a subset of the pool and all initiatives require that, before being granted the right of access and extraction, the user needs to explicitly state the intended use. While this solution does not prevent secondary uses, as use may be not fully observable, the explicit statement of intended use at least facilitates monitoring, as any alternative use may be considered as a violation. Absence of it would make monitoring far more complex.

It seems evident that these solutions are strongly directed to ensure data protection due to sensitivity, at the potential detriment of the generation of option value. A core problem observed in data and information intensive initiatives is a problem that reminds the Arrow Information Paradox (Arrow, 1962). In order to evaluate the quality of data and to establish potential uses, the user might need to access the data, thus generating a potential violation that may not be balanced by actual use. In some cases, a potential solution that can be helpful for users is the development of metadata to describe content. However, metadata may be sometimes incomplete and producing them may be an expensive task to be performed by the pooling initiative. In this respect, the correspondence between contributors and users may eliminate the arrow information paradox. Having contributed data, potential users already know the "quality" of the good contributed.

2.6.2 Incentive Mechanisms

How do these initiatives incentivize contribution from different actors, given that market exchange solutions relying on monetary incentives are extremely complex to be implemented?

First, the cases confirm that the market solution seems to be non-available. None of the cases explored has a mechanism where data were contributed against a payment. Except from the two Biobanks, which instead relied on employment relation to collect capta, all other initiatives relied on non monetary instruments to incentivize contribution.

Multiple mechanisms to incentivize contribution have been observed.

Primarily, in several initiatives, contribution is the only way through which it is possible to extract a derivative from the pool. Contribution is a necessary condition for being entitled with extraction rights.

In addition, not only the possibility to extract a derivative but also access to the derivative may be restricted exclusively to contributors. In VCOR, which is a clinical quality registry, benchmarking measures are the type of derivative to which only contributors may have access. Further, and more subtly, in several cases it is possible to compare own performances with a benchmark value only through the homogeneous and complete collection of capta. In this case, the individual incentives for collecting capta in a complete way are also aligned with the collective interest of complete contribution by each contributing actor.[9]

In some other cases, granting the mere ability to extract a derivative conditional on contribution may be not sufficient. On the one side, because it may still leave space for shirking: contributing actors may just contribute the bare minimum in order to then being able to extract a derivative. On the other side, because the derived benefit may be too small compared to the collection effort. As such, different types of graduated mechanisms are employed to discriminate across levels of contribution.

When the pool is intended for research purposes, authorship on the derived paper is a mechanism that is used as a tool for discrimination. To discriminate across degrees of contribution, two mechanisms were identified in the cases: a minimum quantity or a rank ordered mechanism. In the first case, exemplified by the EBMT registry, the user that requests extraction rights should state in advance the minimum quantity of patients that are needed to be included as author in the derived publication. In this case there is a baseline criterion suggested by the pool.[10] Authorship order is then established in terms of "intensity of contribution"[11].

[9] This need for full completeness may also be a form of disincentive as continuity of contribution may be costly to maintain.
[10] A similar criterion is required to be specified for the inclusion in the paper acknowledgments.
[11] With the exception of specific authorship positions which are usually granted to the subject who is given extraction rights.

Alternatively, the number of authors is fixed and authorship slots are allocated through a rank ordered mechanism, i.e. the one who contributed the most are included as authors in the derived paper.

While both mechanisms appear as a valuable option for incentivizing contributors in substitution for monetary incentives, they may have different implications on which actors may self-select in these initiatives, with potential important implications on data representatives and pool sustainability.

To conclude, how do pooling initiatives address the incentive problem and ensure sufficient contribution?

In pools with benchmarking purposes, there seems to be natural alignment of incentives for the ones that decide to contribute. In order to benefit from comparison of own data with a benchmark measure, the contributor needs to be as complete as possible in capta collection.

In pools with research purposes, inclusion as author in a derived paper and authorship order is the incentive mechanism employed by several initiatives to ensure intensity of contribution.

2.7 CONCLUSIONS AND LIMITATIONS

The present paper contributes primarily to fill the knowledge gap about how attempts to pool data between multiple actors and organizations are structured and managed. This comparative effort enriches a very scant literature on the internal governance of data pooling initiatives and more broadly contributes to the literature of the governance of interorganizational information systems.

In order to capture how data pooling initiatives are governed and to capture elements of variance and consistency, I conceptualize their governance design as a configuration of property rights à la Ostrom.

This framework allows to richly capture elements of consistency and variance in the allocation and management of each of the property rights identified (Contribution, Access, Extraction, Removal and Decision rights). Contribution to data pools is always closed and the right to contribute is conditional on approval both from the data subject and from the pool itself. However, variance is instead observed in the number and type of contributors. In some cases, collection takes place within a single organization through surveyor under employment relation, while in other cases the contributors are instead multiple organizations or even directly multiple data subjects. With respect to access rights, the sensitivity of health data leads to a strong minimization of the allocation of access rights to the whole contributed pool. In a case, technological means allow to preclude full access to any actor. In the other cases, access rights to the whole pool are concentrated within a data steward: either an internal member of the pool, if it has any, or an external provider with the only function of data stewardship. Granting of extraction rights is usually mediated through an approval process, but cases differ on whether these grant extraction rights only contingent to contribution or extraction rights may also be granted to non-contributing external actors. The way removal rights are allocated is instead constant, both data subject and contributor can remove their contributed capta. With respect to decision rights on access and extraction, decision rights are always granted collectively, i.e. the decision to grant access and extraction rights is never left to a single individual. However,

variance is observed with respect to whether collective decision making is taken through a committee mode or a polyarchic one.[12]

Relying on this classification, I have identified a set of interesting combinations of property right allocations and regularities across the analyzed cases. The combination of allocation of access rights and modes of allocation of access and extraction rights allow to identify different modes of governance of pooling initiatives (Facilitating Pools, Sub-Study Pools, Query Based Pools, and Committee Application Pools). Further, interesting regularities seem to emerge with respect to how contribution and extraction rights are configured. Across the cases analyzed, two types of combinations seem to prevail: single contributor-external access and multiple contributor-internal access. Other combinations, especially the cooccurrence of the presence of multiple contributors and granting extraction rights to external access, seem less plausible, due to incentive reasons, as exclusivity in use is the most common incentive mechanism employed in multiple contributor pooling initiatives.

Another relationship seems to emerge with respect to the number of contributors and the mode through which allocation of extraction rights takes place. Especially when the number of contributors is particularly large, data pooling initiatives seem to rely on polyarchic modes of decision for the allocation of extraction rights, either in addition to or in substitution of a committee mode of decision. Reliance on a polyarchic solution is intended to reduce the agency problem that may emerge in centralizing decision only to a “representative” committee.

Also due to its prevalent descriptive purpose, the present paper is more generative than conclusive.

[12] Table 4 in appendix resumes these findings.

Primarily, the identified speculative regularities need to be further explored and consolidated relying on a broader sample size.

Then, the variance in governance traits that the employed framework allowed to identify, opens up to several questions with respect to performance implications of the different solutions. Are the identified solutions equifinal or do they generate different result in terms of performances (research production, accuracy of estimation, availability of derivatives at the point of care); sustainability (ability to generate the pool constantly and to be attractive to contributors); and, given the statistical nature of the endeavor, in term of representativity? Several questions follow from knowing what the elements of variance in the internal governance of data pooling initiatives are. Does granting contribution rights to data subjects instead of clinicians or ad hoc surveyors influence performance, sustainability and representativity of data pooling initiatives? Are these effects contingent on data features? Do decisions by polyarchy or by committee affect actors' willingness to participate in the pool? What are the actors that prefer polyarchic methods of decision rather than committees? And especially, how do these differences in preferences affect the representativity and quality of the pool?

2.8 CHAPTER REFERENCES

Adjerid, I., Adler-Milstein, J., & Angst, C. (2018). Reducing Medicare Spending Through Electronic Health Information Exchange: The Role of Incentives and Exchange Maturity. Information Systems Research, (February), isre.2017.0745. https://doi.org/10.1287/isre.2017.0745

Arat, M., Arpaci, F., Ertem, M., & Gürman, G. (2008). Turkish Transplant Registry: A comparative analysis of national activity with the EBMT European Activity Survey. Bone Marrow Transplantation, 42(SUPPL.1), 142–145. https://doi.org/10.1038/bmt.2008.144

Arrow, K. (1962). Economic welfare and the allocation of resources for invention. In National Bureau Committee for Economic Research (Ed.), The Rate and Direction of Inventive Activity: Economic and Social Factors (Vol. I, pp. 609–626). Princeton University Press. https://doi.org/10.1521/ijgp.2006.56.2.191

Butzkueven, H., Chapman, J., Cristiano, E., Grand'Maison, F., Hoffmann, M., Izquierdo, G., … Malkowski, J. P. (2006). MSBase: An international, online registry and platform for collaborative outcomes research in multiple sclerosis. Multiple Sclerosis, 12(6), 769–774. https://doi.org/10.1177/1352458506070775

Chen, H. (2013). Governing International Biobank Collaboration: A Case Study of China Kadoorie Biobank. Science, Technology and Society, 18(3), 321–338. https://doi.org/10.1177/0971721813498497

Choinière, M., Ware, M. A., Pagé, M. G., Lacasse, A., Lanctôt, H., Beaudet, N., … Truchon, R. (2017). Development and Implementation of a Registry of Patients Attending Multidisciplinary Pain Treatment Clinics: The Quebec Pain Registry. Pain Research and Management, 2017, 1–16. https://doi.org/10.1155/2017/8123812

Chute, C. G., Hart, L. A., Alexander, A. K., & Jensen, D. W. (2014). The Southeastern Minnesota Beacon Project for Community-driven Health Information Technology: Origins, Achievements, and Legacy. EGEMs (Generating Evidence & Methods to Improve Patient Outcomes), 2(3), 16. https://doi.org/10.13063/2327-9214.1101

Densen, P. M., Fielding, J. E., Getson, J., & Stone, E. (1980). The Collection of Data on Hospital Patients — The Massachusetts Health Data Consortium Approach. New England Journal of Medicine, 302(3), 171–173. https://doi.org/10.1056/NEJM198001173020311

Elgarah, W., Falaleeva, N., Saunders, C. C., Ilie, V., Shim, J. T., & Courtney, J. F. (2005). Data exchange in interorganizational relationships: review through multiple conceptual lenses. The DATA BASE for Advances in Information Systems, 36(1), 8–29. https://doi.org/10.1145/1047070.1047073

Everson, J. (2017). The implications and impact of 3 approaches to health information exchange: community, enterprise, and vendor-mediated health information exchange. Learning Health Systems, 1(2), e10021. https://doi.org/10.1002/lrh2.10021

Feldman, S. S., Schooley, B. L., & Bhavsar, G. P. (2014). Health information exchange implementation: Lessons learned and critical success factors from a case study. Journal of Medical Internet Research, 16(8), e19. https://doi.org/10.2196/medinform.3455

Hess, C., & Ostrom, E. (2003). IDEAS, ARTIFACTS, AND FACILITIES: INFORMATION AS A COMMON-POOL RESOURCE. LAW AND CONTEMPORARY PROBLEMS, 66(1&2), 111–146. Retrieved from http://scholarship.law.duke.edu/lcp/vol66/iss1/5

Iacobelli, S. (2013). Suggestions on the use of statistical methodologies in studies of the European Group for Blood and Marrow Transplantation. Bone Marrow Transplantation, 48(SUPPL.1), S1–S37. https://doi.org/10.1038/bmt.2012.282

Koutroumpis, P., Leiponen, A., & Thomas, L. (2018). Data Strategy. In Academy of Management Global Proceedings. https://doi.org/10.5465/amgblproc.surrey.2018.0085.abs

Markus, M. L., & Bui, Q. “Neo.” (2012). Going Concerns: The Governance of Interorganizational Coordination Hubs. Journal of Management Information Systems, 28(4), 163–198. https://doi.org/10.2753/MIS0742-1222280407

Mayer-Schönberger, V., & Cukier, K. (2013). Big Data: A Revolution That Will Transform How We Live, Work, and Think. Boston New York: Houghton Mifflin Harcourt. Retrieved from http://www.amazon.com/Big-Data-Revolution-Transform-Think/dp/0544227751

McGlynn, E. A., Lieu, T. A., Durham, M. L., Bauck, A., Laws, R., Go, A. S., … Kahn, M. G. (2014). Developing a data infrastructure for a learning health system: The PORTAL

network. Journal of the American Medical Informatics Association, 21(4), 596–601. https://doi.org/10.1136/amiajnl-2014-002746

Mindel, V., Mathiassen, L., & Rai, A. (2018). The Sustainability of Polycentric Information Commons. MIS Quarterly, 42(2). https://doi.org/10.25300/MISQ/2018/14015

Oderkirk, J., & Ronchi, E. (2017). Governing Data for Better Health and Health Care. OECD Observer, 309(Q1), 19–20.

Ostrom, E. (2005). Understanding Institutional Diversity. Princeton, New Jersey: Princeton University Press.

Ostrom, E., & Hess, C. (2007). Understanding Knowledge as a Commons. Understanding Knowledge as a Commons From Theory to Practice (Vol. 15). https://doi.org/10.1002/asi

Paolino, A. R., McGlynn, E. A., Lieu, T., Nelson, A. F., Prausnitz, S., Horberg, M. A., … Steiner, J. F. (2016). Building a Governance Strategy for CER: The Patient Outcomes Research to Advance Learning (PORTAL) Network Experience. EGEMS (Washington, DC), 4(2), 1216. https://doi.org/10.13063/2327-9214.1216

Petrova, M., Riley, J., Abel, J., & Barclay, S. (2018). Crash course in EPaCCS (Electronic Palliative Care Coordination Systems): 8 years of successes and failures in patient data sharing to learn from. BMJ Supportive and Palliative Care, 8(4), 447–455. https://doi.org/10.1136/bmjspcare-2015-001059

Pisani, E., & Botchway, S. (2017). Sharing individual patient and parasite-level data through the WorldWide Antimalarial Resistance Network platform: A qualitative case study [version 1; referees: 1 approved], 6311(0). https://doi.org/10.12688/wellcomeopenres.12259.1

Redline, S., Baker-Goodwin, S., Bakker, J. P., Epstein, M., Hanes, S., Hanson, M., … Rothstein, N. (2016). Patient partnerships transforming sleep medicine research and clinical care: Perspectives from the Sleep Apnea Patient-Centered Outcomes Network. Journal of Clinical Sleep Medicine, 12(7), 1053–1058. https://doi.org/10.5664/jcsm.5948

Sah, R. K., & Stiglitz, J. E. . (1998). Committees , Hierarchies and Polyarchies. The Economic Journal, 98(391), 451–470. Retrieved from https://www.jstor.org/stable/2233377

Sherman, S., Shats, O., Ketcham, M. A., Anderson, M. A., Whitcomb, D. C., Lynch, H. T., … Brand, R. E. (2011). PCCR: Pancreatic cancer collaborative registry. Cancer Informatics, 10, 83–91. https://doi.org/10.4137/CIN.S6919

Stone, E. M., Bailit, M. H., Greenberg, M. S., & Janes, G. R. (1998). Comprehensive health data systems spanning the public-private divide: The massachusetts experience. American Journal of Preventive Medicine, 14(3 SUPPL.), 40–45. https://doi.org/10.1016/S0749-3797(97)00045-7

Strandburg, K. J., Frischmann, B. M., & Madison, M. J. (2017). Governing Medical Knowledge Commons. (K. J. Strandburg, B. M. Frischmann, & M. J. Madison, Eds.). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316544587

Stub, D., Lefkovits, J., Brennan, A. L., Dinh, D., Brien, R., Duffy, S. J., … Reid, C. M. (2018). The Establishment of the Victorian Cardiac Outcomes Registry (VCOR): Monitoring and Optimising Outcomes for Cardiac Patients in Victoria. Heart Lung and Circulation, 27(4), 451–463. https://doi.org/10.1016/j.hlc.2017.07.013

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., … Collins, R. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Medicine, 12(3), 1–10. https://doi.org/10.1371/journal.pmed.1001779

Susha, I., Janssen, M., & Verhulst, S. (2017a). Data Collaboratives as a New Frontier of Cross-Sector Partnerships in the Age of Open Data: Taxonomy Development. Proceedings of the 50th Hawaii International Conference on System Sciences, 2691–2700. https://doi.org/http://hdl.handle.net/10125/41480

Susha, I., Janssen, M., & Verhulst, S. (2017b). Data collaboratives as “bazaars”? Transforming Government: People, Process and Policy, 11(1), 157–172. https://doi.org/10.1108/TG-01-2017-0007

Thomas, L. D. W., & Leiponen, A. (2016). Big data commercialization. IEEE Engineering Management Review, 44(2), 74–90. https://doi.org/10.1109/EMR.2016.2568798

Vest, J. R., Campion, T. R., & Kaushal, R. (2013). Challenges, alternatives, and paths to sustainability for health information exchange efforts. Journal of Medical Systems, 37(6). https://doi.org/10.1007/s10916-013-9987-7

Vest, J. R., & Gamm, L. D. (2010). Health information exchange: persistent challenges and new strategies. Journal of the American Medical Informatics Association, 17(3), 288–294. https://doi.org/10.1136/jamia.2010.003673

Walji, M. F., Kalenderian, E., Stark, P. C., White, J. M., Kookal, K. K., Phan, D., … Ramoni, R. (2014). BigMouth: A multi-institutional dental data repository. Journal of the American Medical Informatics Association, 21(6), 1136–1140. https://doi.org/10.1136/amiajnl-2013-002230

Yasnoff, W. A., & Shortliffe, E. H. (2014). Lessons learned from a health record bank start-up. Methods of Information in Medicine, 53(2), 66–72. https://doi.org/10.3414/ME13-02-0030

2.9 APPENDIX

Table 3 - Cases Explored (Data Pools are in bold)

Name	Corresponding Paper	Website
Victorian Cardiac Outcome Registry (VCOR)	(Stub et al., 2018)	https://vcor.org.au/
South Minnesota Beacon Community Clinical Data Repository	(Chute, Hart, Alexander, & Jensen, 2014)	https://semnbeacon.wordpress.com/
Pancreatic Cancer Care Registry (then iCare2) (PCCR)	(Sherman et al., 2011)	http://pccrproject.org/
European Bone Marrow Transplant Patient Registry (EBMT)	(Arat, Arpaci, Ertem, & Gürman, 2008; Iacobelli, 2013)	https://www.ebmt.org/ebmt-patient-registry
BigMouth	(Walji et al., 2014)	https://bigmouth.uth.edu/
Quebec Pain Registry (QPR)	(Choinière et al., 2017)	https://quebecpainregistry.com/
Kadoore Biobank	(Chen, 2013)	https://www.ckbiobank.org/site/
MSBase NeuroImmunology Registry	(Butzkueven et al., 2006)	https://www.msbase.org/

Partners Patient Outcomes Research To Advance Learning Network (PORTAL)	(McGlynn et al., 2014; Paolino et al., 2016)	https://www.pcori.org/research-results/2013/kaiser-permanente-strategic-partners-patient-outcomes-research-advance
UK Biobank	(Sudlow et al., 2015)	https://www.ukbiobank.ac.uk/
Sleep Apnea Patient-Centered Outcomes Network (SAPCON)	(Redline et al., 2016)	https://myapnea.org/
Health Records Bank Startup	(Yasnoff & Shortliffe, 2014)	http://www.healthbanking.org/
New England Healthcare Exchange Network (NEHEN)	(Densen, Fielding, Getson, & Stone, 1980; Stone, Bailit, Greenberg, & Janes, 1998)	http://nehenportal.com/
Coordinate my care (Epaccs)	(Petrova, Riley, Abel, & Barclay, 2018)	https://www.coordinatemycare.co.uk
WWARN	(Pisani & Botchway, 2017)	https://www.wwarn.org/
Connectvirginia	(Feldman, Schooley, & Bhavsar, 2014)	https://connectvirginia.org/connectvirginia-hie-inc-created/

Table 4 - Variance and Consistency in Property Rights Allocation

Type of Right	Object of the right	Entity (potentially) entitled of the right	Stable features among pools	Varying features among pools
Contribution	Contribution of Patient Records	• Data Subject • Surveyor • Team (represented by a Principal Investigator) • Organization (Hospital or health system)	Closed contribution Contribution conditional on consent	«Level» of allocation of contribution rights and type of contributing actor Number of contributors
Access	Access own collected database (set of patient records (*capta*))	• Contributor	Each contributor can have access to its own collected database	

	Access the whole database (pool)	• Data Steward:
		  • Direct employees of the pool
		  • External providers
		• No one
			No contributor have access to the whole database	Federated vs Centralized
				
				Exclusive access to contributors or external access

Extraction	Extraction of patient records on a data subject	• No one		

	Extraction of data derivatives	• Contributors
		• Data Stewards
		• External actors:
		  • Private
		  • Public
			Extraction rights to contributors are always granted after an approval process	Extraction granted to external actors vs only internal
				
				Extraction rights linked with access right vs non linked

	Recognition of authorship	• (Top) Contributors		Minimum quantity criterion vs rank order mechanism
Decision	Determine allocation of access and extraction right	• Contributors • External experts	Decision to allocate access and extraction rights are always taken collectively	Decision by Committe vs Decision by Polyarchy
Removal	Removal of patient records on data subject	• Data Subject	Data subject are always allowed to remove patient records	
	Removal of own collected database	• Contributors	Contributors are always allowed to remove patient records	

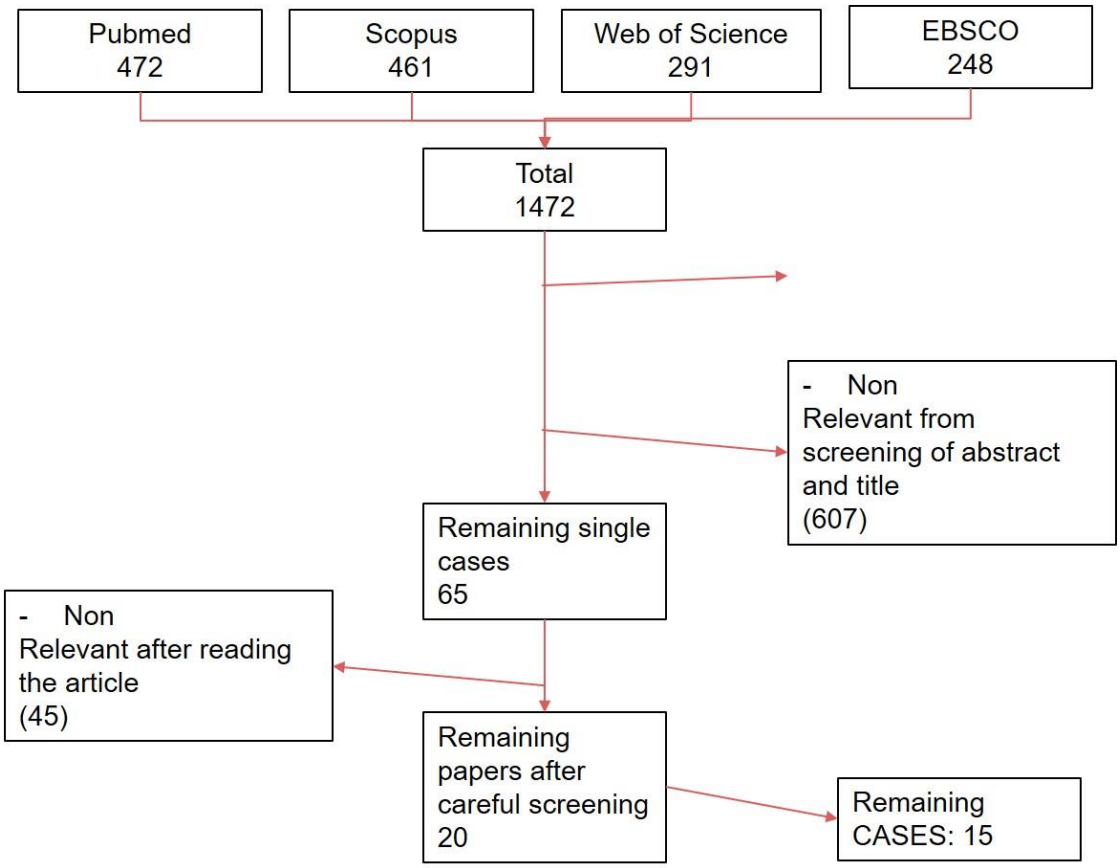

Figure 3 - Case Identification Process

3 MICROFOUNDING THE DATA POOLING PROBLEM: A CLUB THEORETICAL PERSPECTIVE

In 2011, a McKinsey report created immense enthusiasm about the potential of Big Data (McKinsey Global Institute, 2011). However, years later the declared potential has only been partially and unevenly unleashed (McKinsey Global Institute, 2016). While analytic technologies have dramatically advanced (Agrawal et al., 2018), the availability of adequate complementary goods (i.e. sizeable good quality data) has been highly heterogenous across settings, with spots of abundance, scarcity or absence (Borgman, 2015). Mattioli (2017) defines it as the "Data Pooling Problem", arguing that in settings where data are scattered between multiple organizations, innovation may be significantly hindered by the lack of cooperation in pooling industrial, commercial, and scientific data. Further, many initiatives that attempt to combine data across multiple organizations show consistent struggles in ensuring their sustainability (Gliklich et al., 2014; Koutroumpis, Leiponen, & Thomas, 2017a; Zaletel & Kralj, 2015) and there is a very limited understanding of how best to combine data from multiple sources and what are the appropriate governance policies, structures and procedures (Bates et al., 2018; Holmes et al., 2014; Jarvenpaa & Markus, 2018).

This absence of understanding is both related to the relative newness of the phenomenon as well as to an undertheorization of the problem.

Most of the existing studies on the organizational design of Interorganizational Information Systems (IOIS) have focused on initiatives that were supportive of another core interorganizational transaction instead of data being the key object of transaction and a "data

product" the key source of benefits. The electronic data interchanges (Elgarah et al., 2005) and interorganizational coordination hubs (Markus & Bui, 2012) that have been explored in the IOIS literature were mostly intended to support and favour the exchange of a good or service or to support a business process between two or more organizations.

Interorganizational initiatives where data is the key object of transaction, on the other hand, have been analysed only in a limited manner. In the latter setting, the rewards of the organizations involved mostly depend on the benefits generated through the processing of the pooled data, and not on the results of a collateral core business transaction or on the reduction of coordination problems and costs. Very little is known about how this difference impacts incentives and challenges in this type of data intensive collaborations. Therefore, when a "data product" is the key source of benefits for the members of an interorganizational initiative, we have very little knowledge on how to design the appropriate interorganizational governance structures and manage them.

The few existing studies that have explored the design of interorganizational initiatives with data as the core object of transaction are still mostly descriptive in nature (e.g. Mattioli 2017; Susha, Janssen, and Verhulst 2017) and lack in a microfounded theoretical perspective, especially when referring to the key object of governance, i.e. data (Jarvenpaa & Markus, 2018). Further, the rare studies that have attempted to develop a theoretical argument, are mostly focused on settings where data were created within a single organization and are ex post dyadically shared externally (Susha et al., 2017b; van den Broek & van Veenstra, 2015), overlooking the more substantive challenge of a joint and sustainable contribution.

The present paper attempts to fill these gaps by theoretically microfounding the data pooling problem, combining club theory and institutional economics perspectives (Buchanan, 1965; E.

Ostrom, 1990; E. Ostrom & Hess, 2007) with a more refined microfoundation of the resource that is the object of governance, i.e. data, building on the literature on the philosophy of information and information studies (Floridi, 2010; Kitchin, 2014).

Alongside mere conceptual clarity, microfounding data and capturing its idiosyncrasies allows us to better clarify the structure of costs and benefits that actors face through the data production and processing chain.

These identified idiosyncrasies are then included in a club theoretical model that holistically captures how these costs and benefits interact between themselves. The model allows us to disentangle the different contingent conditions under which collaborative initiatives may not emerge or may struggle to be sustainable, due to: mere lack of incentives in generating the data; the presence of incentives that push toward the independent generation and processing of the same data; the presence of a stag hunt problem where, while collaboration may lead to a superior outcome, risks of coordination failure may lead to independently generate a suboptimal "data product"; and, finally, an excessive number of members in the collaborative that may both dilute benefits and unnecessarily increase the risk of moral hazard and costs of data homogenization.

Further, the model helps to capture how different characteristics of data may affect the actors' cost and benefit considerations that influence their willingness to independently or collaboratively produce data intended to generate a piece of data intensive research or an algorithm.

Differently from IOIS where data exchange is collateral to another transaction, we suggest that a reduction in the costs of capturing data generates an enclosure effect rather than favouring the emergence of broad collaborative initiatives and that the emergence of multiple data pooling initiatives, instead of a unique industry wide solution, may be a preferable solution.

Finally, we suggest that features of data that have been associated with the potential success of the "Big Data Revolution" at the intraorganizational level, like data option value and data exhaust (Mayer-Schönberger & Cukier, 2013a), may instead have an hindering effect on the emergence and sustainability of interorganizational data pooling initiatives.

The paper is structured as follows. We first introduce the microfoundation of the resource, clarifying the core concepts that will be used throughout the paper, instead of the vague and confounding term "data", and better conceptualizing the data pooling problem. Then, we introduce the principles of club theory and introduce the cost benefit model that captures the idiosyncrasies of the data production and processing chain. A computational method is employed to identify core model implications. A discussion of the most salient implications and of how contingent features of data affect the probability of the emergence of collaborative data pooling initiatives follows. The remaining sections discuss limitations and conclude.

3.1 BACKGROUND / MOTIVATION

3.1.1 IOIS

Most existing studies on the organizational design of IOIS have focused on initiatives that were supportive of another core interorganizational transaction instead of data being the key object of transaction and a "data product" the key source of benefits.

In electronic data interchanges (Elgarah et al., 2005), the decision to adopt or build an IOIS can be highly influenced by features of the core transaction that the data exchange is supposed to support. Adoption of EDI has been explained by elements of buyer and supplier dependency (Hart & Saunders, 2008), resource criticality and replaceability (Dipanjan Chatterjee & Ravichandran, 2013), efficiency gains to the transaction due to reduction in inventory stocks or cost reduction in transferring collateral documents across organizations (Bakos, 1987). In

contexts where the core object of transaction is data, most of the aforementioned transactional antecedents (D. Chatterjee & Ravichandran, 2004) are not applicable (Adjerid et al., 2018).

Furthermore, even in more data and information intensive interorganizational initiatives, such as Health Information Exchanges (HIE) and Interorganizational Coordination Hubs (ICH), benefits for participants derive from exchange or access to single or combined pieces of information. Instead, in a new algorithmic world, benefits mostly derive from a "data product" generated through the statistical processing of a set of data and not strictly from the mere exchange and access to the underlying data. Thus, benefits derive from a process that is comparable to a production process rather than a communication and exchange one. Two core implications derive from this difference: 1) Actors involved in pooling data to generate a data product cannot simply be considered mere contributors, but are coproducers; 2) The core argument that has been applied to most interorganizational communication and exchange networks , where they assume that these interorganizational initiatives are generally most beneficial when they are used by all (or most) members of the community (Markus & Bui, 2012), deserve to be reconsidered against the "production function" through which data are transformed into a "data product" and in turn into benefits for the participants of the interorganizational initiative.

3.1.2 Data Collaboratives

Although particularly fragmented across disciplinary fields, there exists an emerging literature on interorganizational collaborations where data is the core object of transaction. However, most of these studies are highly descriptive in nature (e.g. Mattioli 2017; Susha, Janssen, and Verhulst 2017), lacking in a microfounded theoretical perspective in both understanding the incentives in the generation and contribution of data to data intensive collaborative initiatives and in understanding how traits and peculiarities of the core boundary resource, i.e. data, do

affect incentives and, in turn, the very emergence and the type of observed collaboration (Jarvenpaa & Markus, 2018). Further, the rare studies that have attempted to develop a theoretical argument, are mostly focused on settings where data were created within a single organization and are ex post dyadically shared externally (Susha et al., 2017b; van den Broek & van Veenstra, 2015), overlooking the more substantive and diffuse challenge of joint and sustainable contribution in settings where access to data sources is fragmented across organizations (Borgman, 2015; Mattioli, 2017).

3.2 MODEL

3.2.1 Microfounding the Data Resource

As suggested by Ostrom (2005) in her study of the governance of natural resources, in order to grasp how a resource is governed it is necessary to understand what are the biophysical and material conditions that the resource of interest takes. However, as argued by Rosenberg (2014), "the term data does heavy lifting yet is barely remarked upon". The increasingly extensive use of the term "data" has further expanded the meanings that have been attached to it, thus reducing terminological clarity. By building on the literature in Information Studies and Philosophy of Information, the primary goal of this section is to clarify what are the physical forms that the resource of interest takes through the data production and processing chain.

Scholars of natural resource commons have found it helpful to distinguish between resource systems and resource units (E. Ostrom & Hess, 2007). A resource system (e.g. a lake) is composed of and is the source for the flow of resource units (e.g. certain quantity of water, fish). In the data production and processing chain, there are three resource units (Signal, Captum and Derivative) and two resource systems (Signal Pool and Capta Pool). The two resource

systems are composed of a pool of one resource unit and are the source for extraction of another resource unit.

Signal refers to a change in status that *can be* collected (Kitchin, 2014). Looking at the "data production process", the very origin of what is actually captured in bytes is the physical world. What elements of the physical world are captured and how they are captured is the result of a choice and a paltry subset of the potential elements (signals) in the physical world that could be captured. A medical encounter between a patient and a clinician consists of an infinite set of signals (a signal pool) that may be captured. In this setting, the clinician, although subject to multiple factors that may influence the choice, decides what are the signals to capture and how to capture them. In a similar vein, also in online settings like social networks, the actions made by a user in a social network may be multiple (writing, clicking, scrolling). The social network potentially has access to all these signals and determines which signals to capture.

Captum refers to the *signals* within the *signal pool* that are actually captured into bytes (Kitchin, 2014), i.e. a *captum* is the result of the selection process described in the paragraph above. The result of a medical encounter is the generation of a set of *capta* on a patient (e.g. body temperature, symptoms and heart rate). A combination of a set of captured signals from a multiplicity of patients is a *capta pool.*

Derivative refers to data generated through the processing of the capta pool (Floridi, 2010). The result of a statistical analysis performed on a multiplicity of capta (a mean, a regression coefficient, a trained prediction model) is what is defined as a derivative. The primary goal of comparative effectiveness initiatives, data intensive research and initiatives that develop prediction algorithms is to generate high quality derivatives through the statistical processing of a capta pool. As the quality of a data derivative is subject to basic principles of statistics, the

capta collected need to have an adequate degree of semantic and technical homogeneity, and the capta pool needs to have a sufficient size for the purpose and, in many cases, a sufficient diversity of sources to avoid bias.

Figure 4 represents the data production and processing process as described.

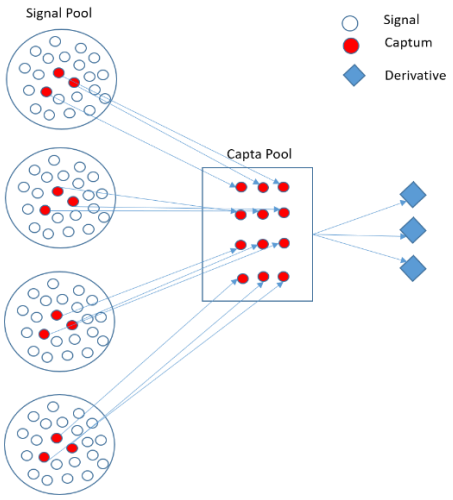

Figure 4 - Data Physical States

3.2.2 The Data Pooling Problem

The challenge for most of the initiatives that process data is to produce the highest quality *derivative* in a sustainable manner, getting from a set of *signal pools* to a *derivative* through the "data production and processing chain" described above.

As the generation of a high-quality *derivative* (whether it be a mean, a regression coefficient or a trained prediction model) is subject to the basic principles of statistics, the *capta pool* collected needs to have an adequate degree of semantic and technical homogeneity and a sufficient size for the purpose.

However, the ways through which the generation of a high-quality derivative takes place differ empirically.

In some well-known cases, the data production and processing chain is mostly integrated within one single organization. An organization that can have access to enough sources of *signals*, is free to independently determine how to capture them and consequently have enough homogeneous *capta* to generate the desired high-quality *derivatives*.

In several other cases, access to *signal pools* is structurally more fragmented. In this context, it may be impossible or extremely costly for a single organization to capture the proper amount and variety of *capta* needed. The alternative solution is instead to collaborate between a multiplicity of organizations, which may, for instance, have easier access to the *signal pools* needed. When collaboration fails or struggles to be sustainable, we are in the circumstance that Mattioli (2017) defines as the Data Pooling Problem.

3.3 MICROFOUNDING THE DATA POOLING PROBLEM: A CLUB THEORETICAL APPROACH

3.3.1 Club Theory

In order to microfound the data pooling problem, I will use a club theoretical approach (Buchanan, 1965). Club theory is a particularly adequate perspective for the problem at stake, as most of the resource systems and resource units involved in the data production and processing chain are club goods.

Capta and *capta* pools, and in most cases *derivatives,* show a high degree of excludability (Drexl, 2016; Kerber, 2016). The degree of excludability of a good is determined by how costly it is to exclude individuals from using the flow of benefits either through technical barriers or legal instruments (E. Ostrom, 1990). In most cases, current technologies allow to exclude other actors from accessing *capta, capta pools* and *derivatives* (Drexl, 2016; Kerber, 2016). The fact

that it is possible to exclude someone else from the stream of benefits derived from a good, does not imply that it will automatically happen. For this reason, it is fundamental to distinguish between the nature of the resource and the governance imposed on the resource.[13] The case of open data is a case in point. While it is technically possible for an organization to keep the data for itself (as many organizations do), it has the possibility to decide to open them to everyone. Thus, excludability broadens the choice set available to a decision maker, by allowing her to determine whether to exclude or not other actors from benefiting from the resource.[14]

A similar reasoning applies to the subtractability dimension. Being digital, *capta* and *capta pools* and in some cases *derivatives,* enjoy a high degree of non-rivalry in consumption (Acquisti, Lane, Stodden, Bender, & Nissenbaum, 2010; Duch-brown, Martens, & Mueller-Langer, 2017; Koutroumpis & Leiponen, 2013; Thomas & Leiponen, 2016). A good is non-rival in consumption when one person's use of the good does not reduce or diminish another person's use. In order to capture its impact on actors' benefits rather than simply focusing on use, more accurately, Ostrom uses the term non-subtractability, in spite of non-rivalry in consumption (E. Ostrom, 1990). A good is non-subtractible when the benefits consumed by one individual do not subtract from the benefits available to others. The degree of subtractability of a good is the degree to which the consumption by one individual subtracts from the benefits available to others. The lower the degree of subtractability, the more freedom a decision maker has to decide to jointly enjoy a resource with another actor (Frischmann, 2012).

[13] This distinction is also relevant when considering the use of the term "public good". While in some cases the term is employed to describe a good that is non-rival and non-excludable in nature, in other settings it is used to describe a type of governance that is imposed on a good that is not a public good in nature.

[14] Non-excludable goods, like common pool resources and public goods are instead characterized by the high cost (to the extreme of "infinite" costs) of excluding individuals from the stream of benefits derived from the good. Clean air and most natural resources are part of this category.

A club theoretical approach[15] is able to capture within its framework the degrees of freedom that a good that shows excludability and a certain degree of non-subtractability grants to the decision maker. Further, it also includes two other valuable elements of flexibility. First, it includes in the framework the costs that one (or more) actor(s) face to generate the club good. Second, it allows for a certain flexibility in how individual costs and benefits vary according to the size of the club good generated and of the number of members involved in the production and enjoyment of it.

Within a club theoretical perspective, for any good, an actor faces two intertwined decisions: what is the optimal amount of good to be produced ("provision decision") and what is the optimal number of actors to involve in the joint production of the good and in the enjoyment of the stream of benefits derived from it (i.e. how many members to involve in the "club", a so-called membership decision). In case the good of interest is purely subtractable (e.g. an apple), the answer to the membership decision becomes trivial. If actor A includes another actor (B) in the "club" to jointly enjoy a purely subtractable good, actor A can only enjoy half of the good, as the other half will be enjoyed by actor B. Even if the two actors split the cost of the apple, they are similarly splitting the benefits derived from it. As such, with a purely subtractable good, the only decision an actor takes is how much to buy/produce for consumption, taking fixed "club" membership at 1.

However, if a good is characterized by a degree of non-subtractability, adding an additional actor to the "club" may reduce costs more markedly than benefits.[16] Thus, optimal club

[15] A club theoretical approach can apply to every type of good that is excludable, so it does not exclusively apply to club goods.

[16] Or may even increase benefits by keeping costs fixed. The key point is that they should not jointly change with the same intensity in the opposite direction.

membership may be higher than 1 as the decision maker may be willing to expand the membership to maximize its utility.

3.3.2 Application to Data Pooling

According to a club theoretical approach, in the setting of the data production and processing chain, if actor *i* aims to generate a *derivative* and aims to maximize its benefits she faces two joint decisions:

- A *provision decision*: What is the optimal quantity Q* to be generated in order to generate the benefit maximizing *derivative*? As the core goal of a data pool is to generate one or more high-quality *derivatives*, actors' benefits are a function of the quality of the *derivative,* which is modelled as function of the quantity of *capta* included in the *capta pool.*
- A *membership decision*: What is the optimal number of members N* to be involved in the joint production and enjoyment of the *capta pool*?

As mentioned above, a club theoretical approach is particularly flexible to be adapted to different circumstances, allowing the researcher to model the cost and benefits that affect the two joint decisions of provision and membership to the setting of interest. The following section introduces the key elements of the model.

3.3.3 Elements of the Model

This section primarily captures a set of cost and benefits that are idiosyncratic to the data production and processing chain and of the data pooling problem, then it exposes through an example and a following more formal version, how these different costs and benefits interact in a coherent model.

Contributors to data pools face three types of costs:

Cost of Data Capture ($C_P(Q)$): Cost of data capture is the marginal cost sustained by an actor to generate a *captum*, i.e. to transform a signal into a *captum*. Cost of data capture may vary substantially between actors and context. At one extreme, production costs may be very close to 0, as in the case of *capta* that are commonly defined as "data exhaust" (Mayer-Schönberger & Cukier, 2013a). This type of *capta* are usually generated for another primary purpose and/or are the by-product of people's actions. An example are *capta* derived from monitoring technologies, like electrocardiograms, which may be used primarily to alert providers in case a patient condition dramatically varies. However, it also collects patients' heartbeat.

On the other extreme, there are *capta* whose primary purpose is exactly their collection. Many surveys fall into this category. In order to capture some individual features, the collector suffers a broad set of costs like identification of subjects, negotiation of access and the cost of a professional who undertakes the job.

Most of the settings fall in between these two extremes. For instance, data captured for observational studies in the health sector are actually collected through the process of care. In this sense, access to a patient is already granted, but there may be some additional costs in term of: time spent to transform into *capta* a set of *signals* that would not have been otherwise captured or more cautiousness in capturing them or getting patient approval for secondary use.

Cost of data capture may be characterized by a threshold. The threshold S_i represents the amount of *capta* that an organization may capture at a relatively low cost. After that threshold, an organization suffers a higher cost to collect *capta* (marginal cost of data capture C_P is multiplied by a multiplicative factor m). An emblematic example is the health sector. A hospital has easy access to the *signals* of its own patients. However, if it needs further *capta* it has two alternatives, either to collaborate with another hospital or to try to collect patient *capta*

independently and in an *ad hoc* fashion. Ad hoc collection, if possible (otherwise m goes to infinite), may be significantly costlier.

Homogenization Costs ($C_A(N)$): Data pools are characterized by a pooled interdependence (Thompson, 1967). In order to create high quality derivatives, the *capta* contributed to the pools need to be homogenous among contributors. In order to homogenise *capta*, actors may suffer homogenization costs, i.e. they need to undergo adaptation costs before or after *capta* collection. Only a high degree of homogeneity, both from a technical and a semantic point of view allows one to perform statistical analysis on all the contributed *capta*. Ceteris paribus, homogenization costs are assumed to increase as the number of members of the pool increases.

Moral Hazard Costs ($\mathbf{C_{M_i}(N)}$): The third family of costs are all the costs derived from the risk of other actors' misbehaviour after contribution takes place. Contribution implies incurring a set of risks: competitive risks (i.e. by having access to the contributed capta, other members of the pool may gain competitive advantage); reputational risks (i.e. contributed capta may disclose undesirable behaviours by the contributor); privacy risks (i.e. contribution to a pool increases the risks of privacy violation in case of personal and sensitive data). Ceteris paribus, moral hazard costs are assumed to increase linearly as the number of members of the pool increases (Clemons & Hitt, 2004)

Benefits

Resource Generated Benefits

Resource generated benefits are benefits that are directly derived by the processing and use of the *capta pool* and not through social interactions between contributing actors. As the core goal of a data pool is to generate one (or more) *derivative*(s), actors' benefits are modelled as a function of the quality of the *derivative*, which in turn is modelled as a function of the quantity

of *capta* included in the *capta pool*. According to the core purpose of the derivative, the benefits derived from it may be more or less diluted by the membership size (N) of the pool.

Production Function and Threshold (α(Q)) The generation of the *derivative* may be interpreted as a production function, where the *capta* within the *capta pool* are the core input to generate the *derivative*. The quality of the generated derivative(s) increase in quantity Q of capta pooled with decreasing returns (Varian, 2018). Further, the production function is considered to be discontinuous at quantity *T*. Once the amount of *capta* collected in the *capta pool* is sufficient to ensure the use of a superior analytical tool or achieve the adequate power to generate an accurate prediction, the quality of the derivative has a marked jump. For, after a threshold T, the production function is multiplied by a factor θ.

Benefits Allocation Function (β(N)) Beta (β(N)) represents the mechanism through which benefits derived from the generation of the *derivative* are allocated. As the generated digital *derivative* shows certain degrees of non-subtractability, the allocation of benefits derived from the processing of the pool of *capta* may not linearly decrease in the number of actors involved (N). However, different contexts and purposes may lead to different modes of allocation of the stream of benefits derived from the *derivative*.

In most settings, the non subtractability of the digital *derivative* allows for an equal and non-subtractive allocation of the aforementioned benefits among all the contributors. For instance, a prediction algorithm that has been trained using the *pooled capta* can be equally employed across the different contributors. In a similar vein, in the case of multicentric benchmarking initiatives, like surgical registries, the access to aggregate measures of performance (e.g. mortality rate after a specific surgical procedure) against which a contributing surgeon can

compare her performance can be equally available to all contributors. In homogeneous actors' settings, β may be represented by a constant that is close to 1.

If instead the *derivative* generated from the *capta pool* is employed for research publication purposes, the benefit allocation function may be declining in the number of contributors. In research publications, the benefits are not simply derived from the *derivative*, which basically are open to all non-contributors, but on the authorship benefits that derive from the publication of it. In this setting, benefits may decline with the increase of membership.

In the model the benefit allocation function is capture by the function $1/N^s$. Where s represents degree of substractability of benefits derived from the *derivative*.

Model Dynamics

A representative example can be helpful in describing how the model works. A clinical unit wants to improve its decision making with regards to how to treat patients with a specific disease. A potential solution to improve its decision making is to rely on a statistical tool that helps in identifying on which patients alternative treatments are more effective. If the clinical unit wants to rely on a statistical tool, it needs to collect data on patients with high consistency (suffering data capture costs C_P for each patient). The more patient data it will collect to more accurate the prediction will be and the better the decision making, but with increasing but marginally decreasing benefits from adding additional data. If it collects more than a certain quantity T of data, the clinical unit may rely on more data eager but more sophisticated data analytics tools and get significantly more accurate predictions. However, the clinical unit encounters a quantity of patients with a specific disease that is lower than the quantity T that would allow it to rely on a data eager superior analytical tool. If it wishes to collect more data than its available quantity, it has two alternatives: 1) trying to recruit patients that are external

to the clinical unit, suffering a higher marginal cost of collection; 2) trying to collaborate with clinical units from other hospitals pooling data together and then benefiting from the joint analysis of them, getting either benefits through achieving a quantity T or higher that allows one to rely on a data eager analytical tool or by sharing the costs of data capture to achieve the desired optimal quantity. However, by collaborating, benefits may be diluted and the clinical unit will suffer both cost of data homogenization and costs of moral hazard, which increase as the number of other clinician units involved.

A formal and general version of this example follows. Let *i* be a utility maximizing organization endowed with quantity signals S_i that can capture at marginal cost C_P . It wants to generate a derivative. Benefits generated from the generated derivative depends on its quality. Quality of the data derivative increases in Q at decreasing marginal rate (represented by the function Q^r with r ∈ (0, 1)). After threshold T, benefits are multiplied by the factor θ. Organization i may decide to either produce quantity Q individually or to produce quantity Q collaborating with other actors. If organization i decides to produce individually, it will face marginal cost C_P to produce quantity $Q \leq S_i$ and cost mC_P to produce any capta outside its own endowment. If organization *i* decides to produce through a collaboration, it will face production costs $C_P Q/N$ and alignment and moral hazard costs $C_A + C_M$ for each additional organization it will involve in the collaborative endeavor. Benefits generated through the collaborative are divided according to degree of rivalry of the derivative, represented by $1/N^s$, with s ∈ [0, 1].

To summarize, organization *i* faces the following combined utility function:

	Q>T	Q<T
Joint Production	$U(Q,N) = \frac{\theta Q^r}{N^s} - c_P\left(\frac{Q}{N}\right) - (c_A + c_M)(N-1)$ $subject\ to\ \frac{Q}{N} \leq S_i$	$U(Q,N) = \frac{Q^r}{N^s} - c_P\left(\frac{Q}{N}\right) - (c_A + c_M)(N-1)$ $subject\ to\ \frac{Q}{N} \leq S_i$
Individual Production	$U(Q) = \theta Q^r - c_P(Q) - c_P(m-1)\ (Q - S_i)$	$U(Q) = \theta Q^r - c_P(Q) - c_P(m-1)\ (Q - S_i)$

3.4 METHOD

To derive our core results, we employ a simulation method. The use of simulation method allows us to add discipline to the process of disciplined imagination that should characterize theory building (Weick, 1989), while allowing theorists to make more realistic assumptions rather than to compromise with analytically convenient ones (Harrison, Lin, Carroll, & Carley, 2007).

	Parameter	Range
Costs of data capture	C_P	[0.1 to 0.9]
Multiplying factor for "external" capture	m	[1 to 9]
Endowment of signals of each organization	S_i	[100 to 10000]
Homogeneization & Moral Hazard costs	$(C_A + C_M) = a$	[1 to 99]
Exponent of the "production function"	r	[0.1 to 0.9]
Benefit multiplier above threshold	θ	[1 to 4]
Threshold level	T	[1000 to 150000]
Degree of rivalry	s	[0 to 1]

Table 5 - Simulation parameters

We computationally simulate the model with 100000 randomly combined realizations of the parameters listed in Table 5.

Optimal Club Size For each realization of the parameters, we identify the utility maximizing combination of Q and N for the representative organization.

$$arg\max_{Q,N}\begin{cases}\frac{\theta Q^r}{N^s} - c_P\left(\frac{Q}{N}\right) - (c_A + c_M)(\mathrm{N}-1) \text{ for } \mathrm{Q} > \mathrm{T} \text{ and } \mathrm{N} > 1 \text{ subject to } \frac{Q}{N} \leq S_i \\ \frac{Q^r}{N^s} - c_P\left(\frac{Q}{N}\right) - (c_A + c_M)(\mathrm{N}-1) \text{ for } \mathrm{Q} < \mathrm{T} \text{ and } \mathrm{N} > 1 \text{ subject to } \frac{Q}{N} \leq S_i \\ \theta Q^r - c_P(Q) - c_P(m-1)\,(Q - S_i)\; for\; Q\; >\; T\; and\; N = 1 \\ Q^r - c_P(Q) - c_P(m-1)\,(Q - S_i)\; for\; Q\; <T\; and\; N = 1\end{cases}$$

Values of Q* and N*, represent the optimal Provision condition (Q*, optimal club size in terms of capta included in the pool) and Membership condition (N*, optimal club size in terms of number of organizations involved in the data pooling club). Each organization establishes what is the optimal quantity Q* to be captured and pooled in order to generate the benefit maximizing derivative and what is the optimal number of members N* to be involved in the joint production and enjoyment of the capta pool. In other words, an organization would be worse off with any other combination of N and Q, different from the optimal combination (Q*,N*). If organizations are utility maximising, we should expect to observe data pooling initiatives of size Q* and N*.

Stag Hunt Further, we collect the utility maximizing Q if the representative actor decides to produce independently:

$$arg\max_{Q}\begin{cases}\theta Q^r - c_P(Q) - c_P(m-1)\,(Q - S_i)\; for\; Q\; >\; T\; and\; N = 1 \\ Q^r - c_P(Q) - c_P(m-1)\,(Q - S_i)\; for\; Q\; <T\; and\; N = 1\end{cases}$$

Club theory-based provision and membership conditions rely on the assumption that once the incentives are in place, no coordination problem may emerge among actors. According to a non-game theoretical view of club theory, if there is an N* where per capita utility is maximized, actors would automatically choose to collaborate to reach it. However, a game theoretical perspective may suggest that this implication is not obvious. Organization *i* may prefer to produce independently a “safe” but suboptimal quantity Q<Q* instead of jointly

collaborate to produce quantity Q*. As collaborative data collection encompasses the effort to coordinate and adapt modes of collection, a possible failure in coordinating may lead all the organizations to be worse off than simply acting individually and independently in capturing data. While a stag hunt game emerges also when organizations have no incentive to produce any quantity of Q individually, existing laboratory evidence shows that the more attractive the safe option (producing individually) compared to the risky one (producing collectively), the more an organization would tend to avoid collaboration (Devetag & Ortmann, 2007), thus increasing the risk of emergence of a stag hunt problem.

In this paper we define as a *Stag Hunt* condition all the situations where:

$$Q^* > \mathrm{T} \text{ and } \mathrm{N}^* > 1 \; and \; \exists\,(Q,\, N = 1)\; s.t.\; U(Q, N = 1) > 0$$

meaning that the optimal solution is a collaborative solution with quantity Q^* above the threshold, but there is a suboptimal positive quantity Q that an organization may find convenient to produce individually.

Feasible Membership Size at Threshold Further, for each realization we identify the feasible set of membership size (N_T) when Q=T

$$N_T = \left\{ \forall\, n : \frac{\theta T^r}{n^s} - c_P\left(\frac{T}{n}\right) - (c_A + c_M)(\mathrm{n} - 1) > 0 \right\}$$

and we identify the maximum and the minimum value of the feasible set at threshold, i.e. we identify the minimum and the maximum value of N at Q=T at which a satisficing representative organization would be willing to participate to the pool:

$$n_{max} = max(N_T)$$

$$n_{min} = min(N_T)$$

These two values have two meanings. On the one side, they can be interpreted as a measure of sensitivity. As actors may fail to coordinate around an optimal N, the broader the distance between N* and n_{max} and n_{min}, the more plausible it is that a pool would emerge nonetheless. On the other side, n_{max} can be considered as a measure of how big an initiative could be, before diluting incentives to contribute below 0. In this paper, n_{max} is employed as a measure to understand under what conditions an industry-wide data pooling solution can be feasible.

Both visual representation and standardized regression analysis is employed to understand how variation in simulation parameters do affect the optimal Q* and N*, and consequently whether we might expect any of the outcomes interest in Table 6. Outcomes of interests are classified into 6 macro categories.

ALGORITHMIC PROVISION (Q*>T)	NON ALGORITHMIC PROVISION (Q*<T)
Collaborative "Algorithmic" Provision (CAP) $Q^* > T \ \& \ N^* > 1$ and $\nexists \ (Q, \ N = 1) \ s.t. \ U(Q, N = 1) > 0$	Collaborative Provision below T (CP) $Q^* < T \ \& \ N^* > 1$
Collaborative "Algorithmic" Provision with Stag Hunt (CAPSH) $Q^* > T \ \& \ N^* > 1$ and $\exists \ (Q, \ N = 1) \ s.t. \ U(Q, N = 1) > 0$	Individual Provision below T (IP) $Q^* < T \ \& \ N^* = 1$
Individual (Non Collaborative) "Algorithmic" Provision (IAP) $Q^* > T \ \& \ N^* = 1$	Full non Provision (NP) $Q^* < T \ \& \ N^* = 1$

Table 6 - Classification of outcomes of interest

The core distinction across outcomes of interest in Table 6 relates to algorithmic or non-algorithmic provision, i.e. whether it would be optimal for the representative organization to (individually or jointly) capture a quantity Q* of signals that allows it to take advantage of a

data eager analytical technology. Within the cases where algorithmic provision takes place, three different groups of outcomes may emerge:

- A pure "collaborative algorithmic provision" (CAP), where the optimal solution is to collaboratively capture an amount of capta that allow to take advantage of a data eager analytic technology and where no actor has any incentive in producing any quantity Q independently as it would be non-convenient.
- A "collaborative algorithmic provision with Stag Hunt" (CAPSH). The optimal solution is to collaboratively capture an amount of capta that allows to take advantage of a data eager analytic technology. However, there exists a suboptimal positive quantity Q that an organization may find convenient to produce individually.
- A "individual algorithmic provision" (IAP). The optimal solution for each organization is to individually and independently capture an amount of capta that allows it to take advantage of a data eager analytic technology, even though this implies to capture signals that are outside its own "endowment", suffering higher costs of data capture.

Cases where algorithmic provision does not take place can be distinguished into three categories:

- "Non Algorithmic Collaborative provision" (CP), where the optimal solution is to collectively capture a quantity Q of capta that is smaller than the threshold that would allow to take advantage of a data eager analytic technology.
- "Non Algorithmic Independent provision" (IP), where the optimal solution is to capture a quantity Q of capta that is smaller than the threshold that would allow to take advantage of a data eager analytic technology.

- "Full Non provision" (NP), where the optimal solution is to not capture any quantity Q of capta.

3.5 MODEL IMPLICATIONS

3.5.1 The Enclosure Effect of Low Costs of Data Capture

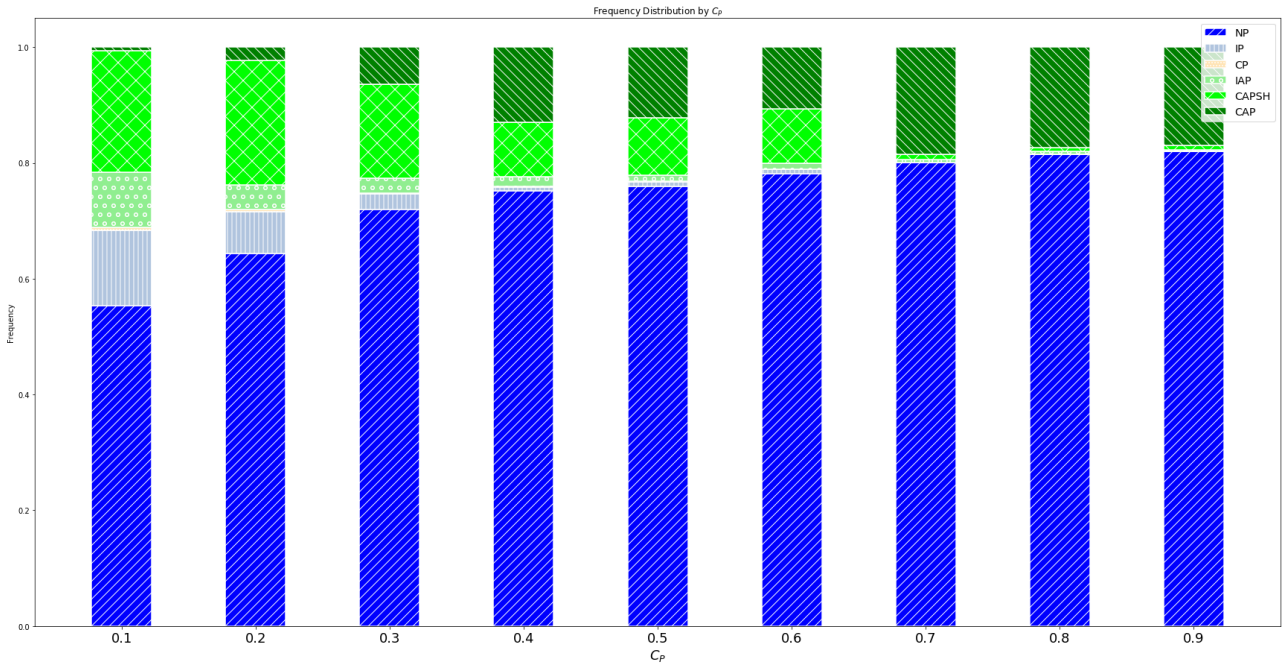

Figure 5 - Frequency of outcomes of interest for different levels of costs of data capture

As shown in Figure 5, a reduction in costs of capturing data (C_P) increases provision in its different forms. As expected, the less costly it is to capture data the more plausible it is that a *capta pool* of size greater than 0 is generated. However, Figure 1 shows that the type of provision observed at different levels of data capture costs varies according to the level of parameter C_P.

Especially at low levels of costs of data capture, i.e. when data are particularly cheap to be collected, we observe the emergence of individual algorithmic provision. When costs of collection are particularly low it may be optimal for each actor to independently collect *capta*,

even accepting to bear higher costs (mC_P) for collecting *capta* outside its own endowment (S_i)[17].

Further, lower costs of data capture also increase the probability of observing a stag hunt condition. Lower costs of data capture increase the feasibility for each organization to capture at least a suboptimal quantity of capta Q. Even if the utility maximizing solution is still to capture a quantity Q* collaboratively, an organization may be more tempted to conservatively produce just its suboptimal quantity Q of capta, potentially reducing the probability to observe algorithmic provision.

Proposition 1*: Lower costs of capture decrease the probability of observing full non provision*

Proposition 2: *Lower costs of capture increase the probability of individual algorithmic provision*

Proposition 3: *Lower costs of capture increase the probability of the emergence of a stag hunt problem, potentially reducing the probability of algorithmic provision*

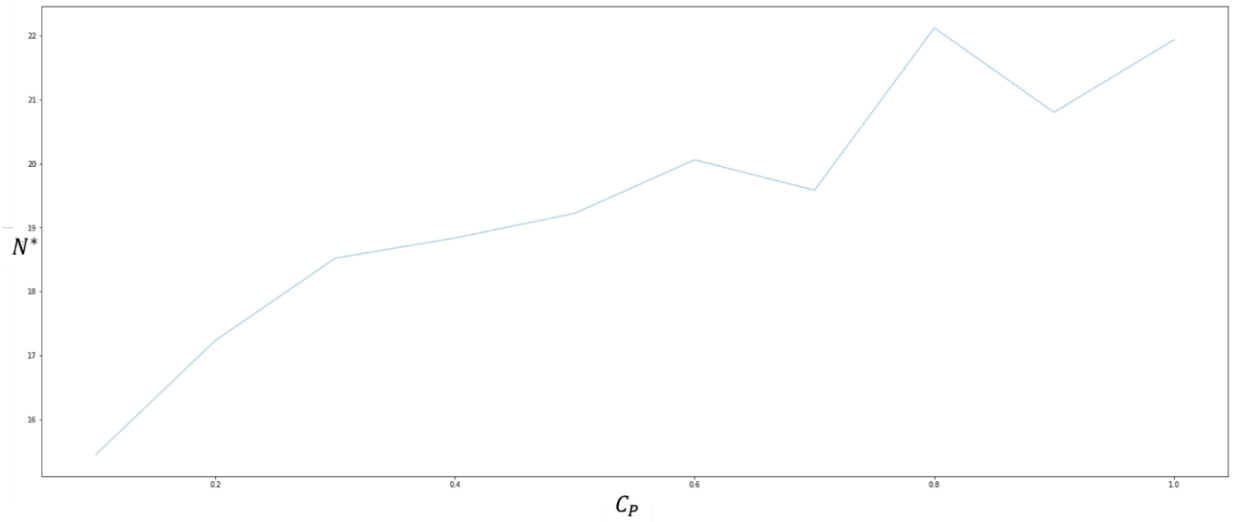

Figure 6 - Average size of optimal club for different level of costs of data capture

[17] In other terms this may imply that actors in the system would collect the same *capta* several times. This being still the optimal solution for each of them.

Figure 6 further shows that in case of provision, optimal club size declines as costs of collection decline.

Proposition 4*:* *Lower costs of capture reduce optimal club size*

These results shed light on why it is relevant to distinguish between IOIS where data is the core element of production and transaction and IOIS where data collection is a collateral effort to coordinate another primary production and transaction process, given that simply translating existing propositions from existing literature in IOIS to the data pooling problem may be misleading.

Reduction in costs of data capture in a IOIS that is supportive of another primary transaction is seen as a reduction in "unit costs of coordination" (Malone, Yates, & Benjamin, 1987), undoubtedly pushing toward higher participation to the IOIS and thus toward broader collaborative initiatives, ceteris paribus. Instead, reduction in costs of data capture when the core goal of a (potential) IOIS is the production of a data product refers to a reduction in unit costs of production, thus having a more complex and potentially opposite effect. In fact, reduction in costs of data capture may lead to smaller collaborative initiatives in terms of membership size and increases the probability, through multiple mechanisms, that an independent initiative of a single organization may emerge rather than a collaborative one.

3.5.2 Club Membership Size

Figure 3 shows the distribution of optimal club size (N*) across the simulated realizations where provision (Q*>0) is observed.

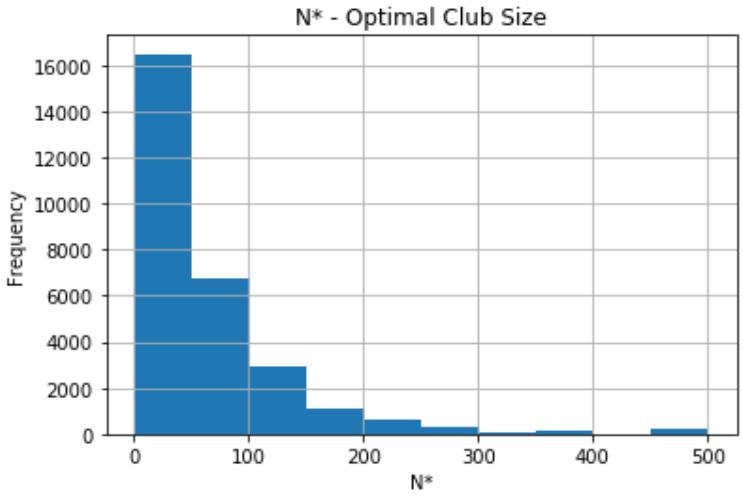

Figure 7 - Distribution of optimal N in case of Q>0*

Across the 100000 combined simulations, rarely the optimal size of N* corresponds to the whole population size (which in the simulation accounts for 500 organizations).

More accurately, only the 0.28% of cases where at least 100 capta (Q) are generated encompass an optimal solution where N* corresponds to the whole population[18]. In most other cases, the optimal membership size is decidedly lower.

Proposition 5: *In data pooling initiatives, optimal club size rarely corresponds to the whole population of organizations in a field*

[18] Even imposing costs of homogeneization and moral hazard and degree of subtractability to their minimum values ($a = 1$ and s=0) only in 1.32% of cases involving the whole population is the optimal membership solution (N*).

Even by relaxing the maximization assumption and assuming that an organization would participate to the club whenever it has positive utility, in only 17% of cases where algorithmic provision takes place the solution where the whole population is involved (n_{max} = 500) in the club is a feasible solution.

s	% of feasible cases of algorithmic provision with full participation
0	42.02%
0.1	28.69%
0.2	19.69%
0.3	11.94%
0.4	8.62%
0.5	6.40%
0.6	3.85%
0.7	1.93%
0.8	0.91%
0.9	0.11%
1	0.00%

Table 7 - Percentage of feasible cases of algorithmic provision with full participation by degree of subtractability

Decomposing this percentage in terms of degrees of non-subtractability (Table 7), even in cases of full non subtractability ($s = 0$), still in 58% of the cases where an algorithm can be provided, involving the whole population in a single data pooling club would drive individual utility to jointly contribute Q=T below 0. More problematically, in contexts where the derivative is intended for research purposes, where degree of subtractability has been estimated to be 0.5 (Bikard, Murray, & Gans, 2015), only in 6% of the cases where we might observe algorithmic provision, the involvement of the whole population can be considered a feasible solution. The core implication of these results is in line with the core gist of club theory. When excludability is possible, for the provision of goods that show a certain degree of non-substractability but are subject to crowding, the optimal solution to maximize benefits for the whole population may

be to fragment the population into smaller clubs, each in charge of generating its own high-quality *data product.* Attempts to generate a unique data product for the whole population, may instead lead to the non-emergence of any data product. This is a risk that is not considered in most of the IOIS cases that have been explored in the existing IS literature, where the desirability of industry-wide IOIS is driven by the principle that optimal membership solution almost always relies on the full participation of industry members in a single IOIS (Markus & Bui, 2012). In a certain way, the presence of data silos of the emergence of data pools that are limitedly permeable to additional membership and external data sharing (Mattioli, 2017), may be seen as an optimal club solution instead of undesirable outcomes.

3.6 Data Characteristics and Implications on Algorithmic Provision and Membership

The present section explores how some of the characteristics that have been employed to characterize data in the recent literature covering the topic of Big Data have implications on the emergence of high-quality derivatives, on the optimal membership of data pools and on the probability of observing a stag hunt problem. As the explored factors affect two parameters of the model and as these parameters may affect the outcome of interest in different directions, the magnitude of the effect of each parameter is evaluated through a regression with standardized coefficients (Table 8).

	(1) P(Algorithmic Provision) beta/_star	(2) N* beta/_star	(3) P(Stag Hunt) beta/_star
$(C_A + C_M)$	-.0760228***	-.1377899***	-.015198**
C_P	-.0859258***	.0216291***	-.5868177***
r	.6260388***	.328729***	.3321766***
m	-.0118107***	.0119087***	.1199033***
θ	.1303677***	.0486531***	-.0432602***
T	-.0300856***	.0045354	-.0252168***
s	-.2307455***	-.251416***	-.1045375***
S_i	.0734652***	-.1077544***	.061768***
N	100000	100000	19761

Table 8 - Effects of simulation parameters on core outcomes of interest - regression results with standardized coefficients (p<0.05 **p<0.01 ***p<0.001)*

3.6.1 Data Exhaust

A relevant feature that has been employed to characterize the Big Data Revolution is the increasing abundance of data exhaust. Data Exhaust are defined as data that have been generated for another primary function (Mayer-Schönberger & Cukier, 2013a; McKinsey Global Institute, 2011; Thomas & Leiponen, 2016). Most of the literature focuses on the competitive advantage that individual firms may gain by having access to data exhaust (Mayer-Schönberger & Cukier, 2013a). However, less attention has been devoted to what are the implications of data exhausts on the emergence and structuring of data pools.

Implications on the Model Components

Data exhausts are characterized by a low opportunity cost of capture, as costs of capture have been already undergone when they were collected for their primary purpose. However, as *capta* have been collected for other purposes, the cost of adapting them between multiple

organizations may be even more marked. As such, for data exhaust the slope of C_A increases. Adding an additional actor involves a costly adaptation process, either after collection (if possible) or before collection (adapting modes of collection ex ante to the purpose of joint contribution increase capture costs).

Implications on the Probability of Generating a High-Quality Derivative (Algorithmic Provision)

A low cost of capture implies an increase in the probability of the existence of at least a combination of values of N and Q such that a high-quality data *derivative* is produced. However, as C_A increases, the probability of the existence of at least a condition where the high-quality data *derivative* is generated decreases.

Despite the lower costs of capture of data exhausts, the generation of a high-quality *derivative* may fail, due to higher costs of aligning data that were generated for another primary purpose. In fact, regression results suggest that the two factors tend to mostly balance each other (i.e. a standard deviation change in both factors affect probability of algorithmic provision with the same magnitude).

However, the risk of non-provision of high-quality data *derivative* is amplified by the stag hunt problem. As suggested in proposition 3, lower costs of capture increase the probability of the emergence of a Stag Hunt problem, potentially reducing the probability of algorithmic provision. Thus, in settings characterized by fragmentation of data access, the impact of the increasing abundance of data exhaust on the probability of emergence of a high-quality *derivative* are uncertain and deserve further empirical exploration.

Implications on Optimal Data Pool Membership

Both low costs of data capture (C_P) and higher costs of homogenization (C_A) negatively affect optimal membership size. As such, they have an enclosure effect.

Proposition 6: *The presence of data exhausts is negatively associated with optimal membership size*

3.6.2 Sensitive Data

Sensitive data are commonly defined through a list of types of data, rather than through a working definition. According to the EU General Data Protection Regulation, sensitive data are data that refer to: "racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership […] genetic data, biometric data, health data". Similarly to the EU approach, every country employs a specific list of types of data that are considered sensitive to impose more restrictive regulations in terms of how they should be captured and the sanctions that may derive from misuse or violation. Only recent evidence attempts to capture a potential "spectrum of sensitivity" for particular sub categories of data (Rumbold & Pierscionek, 2018), showing that the perceived sensitivity of a particular type of data may vary widely both between societies or ethnic groups and within those groups, especially in light of the implications that the misuse of data has on the data subjects.

Implications on the Model Components

The degree of sensitivity of data affects some of the costs that actors in the data production and processing chain face. More sensitive data require higher costs of data capture (C_P) as the transformation of a *signal* into *capta* is commonly more strictly regulated and requires more demanding procedures. Apart from regulatory requirements, negotiating the possibility to capture a sensitive *signal* may be more complex and demanding as data subjects may be less

willing to allow to capture them or at least to reuse them for a secondary purpose (Bell, Ohno-Machado, & Grando, 2014; Whiddett, Hunter, Engelbrecht, & Handy, 2006).

The degree of sensitivity also increases the potential impact of other actors' misbehaviour. The improper access to or use of *capta* by another actor has a markedly higher impact in cases where it refers to sensitive *capta*, both in terms of loss of trust by the sources of the *capta* and even in terms of sanctions to which an actor may be subject. For this reason, "Moral Hazard Costs" (C_M) increase with the sensitivity of *capta* involved in the pool.

Implications on the Probability of Generating a High-Quality Derivative (Algorithmic Provision)

Sensitivity increases costs of data capture and costs of moral hazard. As such, it decreases the probability of generation of a high-quality derivative, i.e. that there exists at least a combination of N and Q such that Q>T.

Proposition 7: *Data sensitivity is negatively associated with the probability of observing the provision of high-quality derivative*

Implications on Optimal Data Pool Membership

Ceteris paribus, the impact of sensitivity on the optimal club membership is uncertain, i.e. it depends on the intensity to which sensitivity differently affects costs of production and moral hazard costs. Increases in costs of production may lead to an increase in N*, thus leading to a higher probability that a collaboration emerges, as it would be preferable to share the costs of collecting the desired amount of *capta* among multiple actors. However, increases in moral hazard costs push toward a reduction in N* as any further actor involved in the pool increases the risk of an undesirable violation. Regression results in Table 4 suggest that a standard

deviation change in moral hazard costs is markedly more influential on the optimal number of members in the pool than a standard deviation change in costs of capture.

Thus, we suggest that in contexts where data are more sensitive, we might observe initiatives with smaller (or even individual) membership.

Proposition 8: *Data sensitivity is negatively associated with optimal membership size*

3.6.3 Data Option Value

The "option value" of data is one of the features that Mayer-Schönberger & Cukier (2013a) consider pivotal for the "Big Data Revolution". Offering a broad set of examples, they argue that "data's true value is like an iceberg floating in the ocean. Only a tiny part of it is visible at first sight, while much of it is hidden beneath the surface". As such, their potential uses can be multiple, and, in many cases, data generated for a function may well serve another (unrelated) function. For this reason, we introduced in the simulation a further parameter *p*, i.e. a benefit multiplier that hypothesizes that the same *capta pool* may generate from 1 to 4 different *derivatives* ($p \in [1,4]$). Regression results with the inclusion of *p* are represented in Table 9.

	(1) P(Algorithmic Provision) beta/_star	(2) N* beta/_star	(3) P(Stag Hunt) beta/_star
$(C_A + C_M)$	-.0676118 ***	-.1327998 ***	.0216324 ***
C_P	-.0766808 ***	.0411474 ***	-.108897 ***
r	.6899419 ***	.3142003 ***	.057302 ***
m	-.0113682 ***	.0329118 ***	.2274372 ***
θ	.1323282 ***	.0369436 ***	-.1289452 ***
T	-.0427686 ***	.0056748 *	-.0073225
s	-.2206281 ***	-.266949 ***	-.1450441 ***
S_i	.0709693 ***	-.1386118 ***	.0613002 ***
p	.0999147 ***	.029705 ***	.1988982 ***
N	100000	100000	26368

Table 9 - Effects of parameters on core outcomes of interest with parameter p- regression results with standardized coefficients (p<0.05 **p<0.01 ***p<0.001)*

Implications on the Model Components

The option value of *capta* and *capta pools* has two implications on the model. The presence of an option value and the possibility of using the same *capta pool* for multiple purposes may positively affect the amount of benefits that a *capta pool* may generate. However, at the same time, the possibility that *capta* or the *capta pool* may be used for purposes different to those for which they were collected and contributed as well as may be potentially used in unexpected ways, increase marginal Moral Hazard Costs (C_M).

Implications on the Probability of Generating a High-Quality Derivative (Algorithmic Provision)

The increase of the potential benefits that may be generated through the same *capta pool* increases the probability of the existence of at least a condition where the high-quality data

derivative is generated. However, the increase in the risk of potential misuse may be a disincentive for contribution. Regression results suggest that a standard deviation increase in the potential uses of data positively affects the probability of emergence of a high-quality derivative slightly more than a standard deviation change in costs of moral hazard.

However, as option value acts at all levels of contribution, the multiplication of potential purposes of data increases the benefits derived from the safe solution of individual provision of a suboptimal quantity increasing the probability of stag hunt.

Thus, while data option value seems to have weak positive effects on the probability of generating a high-quality derivative, the increase in the risk of stag hunt suggests that we cannot derive a clear proposition on the direction to which a higher degree of data option value affect the probability of provision of a high-quality derivative. Thus, this is a trade-off that deserves further empirical evaluation.

Implications on Optimal Data Pool Membership

Ceteris paribus, option value may tend to decrease the number of members of the pool. The increase in benefits afforded by data option value weakly affects the decision on the optimal number of actors to involve in the data pooling club. However, as data option value increases the breadth of potential uses, they consequently increase the risk of potential misuses. For instance, an actor may prefer to reduce the number of members involved in pooling or even to self-produce and individually take advantage of the *capta pool* in order to reduce the risk of an undesired use.

Proposition 9: *Data option value is negatively associated with optimal membership size*

Data Characteristics	Elements of the Model Affected		P(Provision)	Membership: Impact on optimal N
Data Option Value (Mayer-Schönberger & Cukier, 2013a)	p	+	Uncertain	-
	C_M	+		
Data Exhaust (Mayer-Schönberger & Cukier, 2013a)	C_P	-	Uncertain	-
	C_A	+		
Sensitivity (Rumbold & Pierscionek, 2018)	C_P	+	-	-
	C_M	+		

Table 10 - Data characteristics and implications

3.7 LIMITATIONS AND FURTHER DEVELOPMENTS

The model focuses only on the hard "technological" aspects of the data pooling problem, as it considers benefits as exclusively deriving from the pooling and processing of the resource. Other sources of softer benefits, like skills development in data collection for analytical purposes and professional and relational benefits merely derived from participation and independently from the quality of the 'data product' generated, have not been included in the model. A valuable extension of the model would be to include these elements.

The paper inherits some of the limitations of club theory. Club theory is particularly limited in considering how actors coordinate after incentives to collaborate are in place. It assumes that if incentives are in place, actors will find a way to coordinate. However, as coordination problems may emerge even when incentives are in place, a richer consideration on how to incorporate coordination problems and consequent coordination mechanisms in the microfoundation of the data pooling problem is needed.

3.8 CONCLUSIONS

Microfounding data and characterizing them as a production factor allows us to understand and capture in a more complete way the idiosyncrasies that characterize the data production and

processing chain. In this way, it grants the possibility to comprehend the type of incentives that actors face in this context and to rigorously derive how contingent factors may affect the emergence, sustainability and success of initiatives that attempt to combine data between multiple actors to generate a data *derivative,* i.e. a statistical product derived from the analysis of other data.

For instance, differently from what we might expect from IOIS that are collateral to another transaction and/or are based on data exchange, by treating the data pooling challenge as a production process we suggest that low costs of capturing data may have an "enclosure effect" rather than broadening the number of participants to the IOIS. Settings where data can be captured at a low cost may be characterized by collaborative initiatives of a smaller size or, in cases where costs of capture are particularly low, by the independent generation of data by each individual organization. In addition, low costs of data capture may lead to a stag hunt problem, where organizations may prefer to produce a 'safe' small quantity of data instead of engaging in a pooling initiative to achieve a sufficient size to take advantage of data eager technologies.

By exploring the impact of data characteristics that have been identified as the main source of intraorganizational Big Data success, such as data exhaust and data option value (Mayer-Schönberger & Cukier, 2013a), we show that they might have different implications at the interorganizational level. Both data option value and data exhaust have uncertain effects on the degree to which they favour interorganizational collaboration. An uncertainty that deserves further exploration.

Further, by distinguishing between costs and benefits that are inherently related to the production and transformation of data and the costs and benefits that are inherently related to the number of actors involved in the production and transformation of them, the present paper

disentangles the problem of lack of collaboration in data pooling initiatives in more refined subproblems and identifies under which contingent conditions different subproblems may emerge. Distinguishing the data pooling problem from: 1) the mere absence of incentives to generate the needed data; 2) the presence of incentives subject to the broadness of the collaboration; 3) or the presence of coordination (stag hunt) problems, allows us to direct governance and organizational design efforts in substantially different directions. Subsidies for incentivizing data capture may be highly beneficial in case we observe a provision problem. However, they may be detrimental if used in response to a stag hunt condition where they might further increase incentives to choose a safer but suboptimal individual solution. Similarly, relying on third party data custodians (Thomas & Leiponen, 2016) to process data, thus avoiding the risk of data misuse by members of the pool, positively affect the probability of provision by reducing the risk of moral hazard by pool members, still is a very weak solution when a stag hunt problem emerges. Increasing the observability of other actors choices seems instead to be a proper response to solve stag hunt problem (Devetag & Ortmann, 2007; Puranam et al., 2012).

A club theoretical perspective allows to go beyond the pure dichotomy between open and closed data, by allowing to appreciate a more continuous spectrum of choice that organizations and individuals face in collaborating in data intensive initiatives. It suggests that in collaborative initiatives where the outcome of interest is the result of a statistical analysis, thus showing diminishing returns to the quantity of data included, adding additional contributors to the optimal quantity needed may be detrimental rather than beneficial and may eventually lead to the non-emergence or non-sustainability of the collaborations. In other words, instead of simply focusing on clubs composed by 1 (closed data) or by everyone (open data), a club theoretical approach suggests that the optimal solution can frequently be an intermediate one and allows

both the researcher and the organizational designer to capture what may be the optimal number of members to be involved to achieve the desired outcome of producing a high quality data derivative, seeing how it varies according to contingent conditions.

Finally, we introduce club theory to the Management of Information Systems literature. As most of the goods involved in IS show some degrees of non-rivalry in consumption and an increasing level of excludability, the use of this theoretical perspective may benefit more widely other explorations within the field, having an especially interesting application in the whole field of Interorganizational Information Systems.

3.9 CHAPTER REFERENCES

Acquisti, A., Lane, J., Stodden, V., Bender, S., & Nissenbaum, H. (2010). The Economics and Behavioral Economics of Privacy. In Privacy, Big Data, and the Public Good (pp. 76–95). New York: Cambridge University Press. https://doi.org/10.1017/CBO9781107590205.005

Adjerid, I., Adler-Milstein, J., & Angst, C. (2018). Reducing Medicare Spending Through Electronic Health Information Exchange: The Role of Incentives and Exchange Maturity. Information Systems Research, (February), isre.2017.0745. https://doi.org/10.1287/isre.2017.0745

Agrawal, A., Gans, J., & Goldfarb, A. (2018). Prediction machines : the simple economics of artificial intelligence. Harvard Business Review Press.

Bakos, J. Y. (1987). Interorganizational Information Systems: Strategic Implications for Competition and Cooperation. Massachusetts Institute of Technology.

Bates, D. W., Heitmueller, A., Kakad, M., & Saria, S. (2018). Why policymakers should care about "big data" in healthcare. Health Policy and Technology, 7(2), 211–216. https://doi.org/10.1016/j.hlpt.2018.04.006

Bell, E. A., Ohno-Machado, L., & Grando, M. A. (2014). Sharing my health data: a survey of data sharing preferences of healthy individuals. AMIA Symposium, 2014, 1699–1708. Retrieved from http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed13&NEWS=N&AN=25954442

Bikard, M., Murray, F., & Gans, J. S. (2015). Exploring Trade-offs in the Organization of Scientific Work: Collaboration and Scientific Reward. Management Science, 61(7), 1473–1495. https://doi.org/10.1287/mnsc.2014.2052

Borgman, C. L. (2015). Big Data, Little Data, No Data. Cambridge, MA: The MIT Press.

Buchanan, J. M. (1965). An Economic Theory of Clubs. Economica, 32(125), 1. https://doi.org/10.2307/2552442

Chatterjee, D., & Ravichandran, T. (2004). Inter-organizational information systems research: a critical review and an integrative framework. In 37th Annual Hawaii International

Conference on System Sciences, 2004. Proceedings of the (Vol. 00, p. 10 pp.). IEEE. https://doi.org/10.1109/HICSS.2004.1265398

Chatterjee, Dipanjan, & Ravichandran, T. (2013). Governance of interorganizational information systems: A resource dependence perspective. Information Systems Research, 24(2), 261–278. https://doi.org/10.1287/isre.1120.0432

Clemons, E. K., & Hitt, L. M. (2004). Poaching and the Misappropriation of Information: Transaction Risks of Information Exchange. Journal of Management Information Systems, 21(2), 87–107. https://doi.org/10.1080/07421222.2004.11045802

Devetag, G., & Ortmann, A. (2007). When and why? A critical survey on coordination failure in the laboratory. Experimental Economics, 10(3), 331–344. https://doi.org/10.1007/s10683-007-9178-9

Drexl, J. (2016). Designing competitive markets for industrial data - Between Propertisation and Access.

Duch-brown, N., Martens, B., & Mueller-Langer, F. (2017). The economics of ownership , access and trade in digital data (JRC Digital Economy Working Paper 2017-01).

Elgarah, W., Falaleeva, N., Saunders, C. C., Ilie, V., Shim, J. T., & Courtney, J. F. (2005). Data exchange in interorganizational relationships: review through multiple conceptual lenses. The DATA BASE for Advances in Information Systems, 36(1), 8–29. https://doi.org/10.1145/1047070.1047073

Floridi, L. (2010). Information. Oxford University Press. https://doi.org/10.1093/actrade/9780199551378.001.0001

Frischmann, B. M. (2012). Infrastructure: the social value of shared resources. New York: Oxford University Press.

Gliklich, R., Dreyer, N., & Leavy, M. (2014). Registries for Evaluating Patient Outcomes: A User’s Guide. Rockville (MD).

Harrison, J. R., Lin, Z., Carroll, G. R., & Carley, K. M. (2007). Simulation modeling in organizational and management research. Academy of Management Review, 32(4), 1229–1245. https://doi.org/10.5465/amr.2007.26586485

Hart, P., & Saunders, C. (2008). Power and Trust: Critical Factors in the Adoption and Use of Electronic Data Interchange. Organization Science, 8(1), 23–42. https://doi.org/10.1287/orsc.8.1.23

Holmes, J. H., Elliott, T. E., Brown, J. S., Raebel, M. A., Davidson, A., Nelson, A. F., … Steiner, J. F. (2014). Clinical research data warehouse governance for distributed research networks in the USA: a systematic review of the literature. Journal of the American Medical Informatics Association, 21(4), 730–736. https://doi.org/10.1136/amiajnl-2013-002370

Jarvenpaa, S. L., & Markus, M. L. (2018). Data Perspective in Digital Platforms: Three Tales of Genetic Platforms. Proceedings of the 51st Hawaii International Conference on System Sciences, 4574–4583. Retrieved from https://scholarspace.manoa.hawaii.edu/bitstream/10125/50466/1/paper0579.pdf

Kerber, W. (2016). A New (Intellectual) Property Right for Non-Personal Data? An Economic Analysis.

Kitchin, R. (2014). The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences. 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom: SAGE Publications Ltd. https://doi.org/10.4135/9781473909472

Koutroumpis, P., & Leiponen, A. (2013). Understanding the value of (big) data. In 2013 IEEE International Conference on Big Data (pp. 38–42). IEEE. https://doi.org/10.1109/BigData.2013.6691691

Koutroumpis, P., Leiponen, A., & Thomas, L. (2017). The (Unfulfilled) Potential of Data Marketplaces (ETLA Working Papers No. 53). ETLA Working Papers. Retrieved from http://pub.etla.fi/ETLA-Working-Papers-53.pd

Malone, T. W., Yates, J., & Benjamin, R. I. (1987). Electronic markets and electronic hierarchies. Communications of the ACM, 30(6), 484–497. https://doi.org/10.1145/214762.214766

Markus, M. L., & Bui, Q. "Neo." (2012). Going Concerns: The Governance of Interorganizational Coordination Hubs. Journal of Management Information Systems, 28(4), 163–198. https://doi.org/10.2753/MIS0742-1222280407

Mattioli, M. (2017). The Data-Pooling Problem. Berkeley Technology Law Journal, 32(1).

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

McKinsey Global Institute. (2011). Big data: The next frontier for innovation, competition, and productivity.

McKinsey Global Institute. (2016). The Age of Analytics: Competing in a Data-Driven World.

Ostrom, E. (1990). Governing the Commons: The Evolution of Institutions for Collective Action. (Cambridge University Press, Ed.), Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge.

Ostrom, E. (2005). Understanding Institutional Diversity. Princeton, New Jersey: Princeton University Press.

Ostrom, E., & Hess, C. (2007). Understanding Knowledge as a Commons. Understanding Knowledge as a Commons From Theory to Practice (Vol. 15). https://doi.org/10.1002/asi

Puranam, P., Raveendran, M., & Knudsen, T. (2012). Organization design: the epistemic interdepencence perspective. Academy of Management Review, 37(3), 419–440.

Rumbold, J. M. M., & Pierscionek, B. K. (2018). What Are Data? A Categorization of the Data Sensitivity Spectrum. Big Data Research, 12, 49–59. https://doi.org/10.1016/j.bdr.2017.11.001

Susha, I., Janssen, M., & Verhulst, S. (2017a). Data Collaboratives as a New Frontier of Cross-Sector Partnerships in the Age of Open Data: Taxonomy Development. Proceedings of the 50th Hawaii International Conference on System Sciences, 2691–2700. https://doi.org/http://hdl.handle.net/10125/41480

Susha, I., Janssen, M., & Verhulst, S. (2017b). Data collaboratives as "bazaars"? Transforming Government: People, Process and Policy, 11(1), 157–172. https://doi.org/10.1108/TG-01-2017-0007

Thomas, L. D. W., & Leiponen, A. (2016). Big data commercialization. IEEE Engineering Management Review, 44(2), 74–90. https://doi.org/10.1109/EMR.2016.2568798

Thompson, J. D. (1967). Organizations in Action: Social Science Bases of Administrative Theory. New York: McGraw-Hill.

van den Broek, T. van den, & van Veenstra, A.-F. van. (2015). Modes of governance in inter-organisational data collaborations : Complete Research. Twenty-Third European Conference on Information Systems (ECIS), Münster, Germany, 2015, 0–12.

Varian, H. (2018). Artificial Intelligence, Economics, and Industrial Organization. Cambridge, MA. https://doi.org/10.3386/w24839

Weick, K. E. (1989). Theory Construction as Disciplined Imagination. The Academy of Management Review, 14(4), 516. https://doi.org/10.2307/258556

Whiddett, R., Hunter, I., Engelbrecht, J., & Handy, J. (2006). Patients' attitudes towards sharing their health information. International Journal of Medical Informatics, 75(7), 530–541. https://doi.org/10.1016/j.ijmedinf.2005.08.009

Zaletel, M., & Kralj, M. (2015). Methodological guidelines and recommendations for efficient and rational governance of patient registries. Ljubljana.

CONCLUSIONS

The present dissertation lays the foundation for getting closer to knowing how to best design data pooling initiatives. Although chapters are conceptually linked across them, each of them puts a brick of different nature to the foundations.

Chapter 1 performs a marked conceptualization effort, in an attempt both to better understand the resource object of governance and to build the conceptual infrastructure of the dissertation. It identifies the different biophysical states of data in the data production chain and performs a characterisation of data intended to appreciate which characteristics of data may influence the governance of data pools and how. From a conceptual point of view: 1) it confirms that there is value in not treating data as a monolith and that a richer distinction of the physical states that data takes helps in reconciling contradictory classifications of the resource that may have opposite governance implications; 2) it argues that it is conceptually more appropriate to categorize a data pool as a club good (Buchanan, 1965) rather than as common pool resource (E. Ostrom, 1990); 3) it suggests that the dimensions employed to characterize data from a technical point of view (Volume, Velocity, Variety (Gandomi & Haider, 2015)) may have different governance implications than simply the ones derived from the technical complexity of processing 'Big' Data. In addition, it suggests that, when data are non-rival, sensitive and have option value, some governance solutions (especially market based solutions) may be unfeasible or limitedly effective. Consequently, it raises the question on how data pooling initiatives manage to incentivize contribution. Similarly, as data option value, non rivalry and sensitivity also generate a trade-off in governing access to data, it suggests that purely open or closed solutions may be ineffective, but questions how intermediate solutions to govern data access are designed. Questions that are left to the following empirical chapter.

In Chapter 2, I perform a comparative study of several data pooling initiatives in the health sector, conceptualising their governance design as a configuration of property rights à la Ostrom. It is the first study that captures, in such a rich way, elements of consistency and regularities in the design of the internal governance of data pooling initiatives, filling a vast gap in term of knowledge and systematisation of how data pooling initiatives are designed, contributing both to the literature of management of information systems and to the literature in health policy. The comparative approach employed also allowed to identify and clarify an often-overlooked difference between Repositories, Exchanges and Data Pools. In addition, the study contributes to the literature on meta-organizations and new organizational forms, advancing a new refined framework to capture meta-organizational design (Gulati, Puranam, & Tushman, 2012) in a richer way but with a proper balance in terms of parsimony. It enriches the dichotomy between open and close boundaries, suggesting that rights in the same initiatives may differ with respect to their openness or closure. It enriches the distinction between heterarchical modes of decision making, suggesting the presence of two alternative modes of heterarchical decision making: polyarchy and committee (Grandori, 2013; Sah & Stiglitz, 1998).

In Chapter 3, I develop a club theoretical model that comprehensively captures how different idiosyncrasies of data interact between them in influencing organization's decision to generate data for analytical purposes individually or collaboratively and appreciates, through a computational method, how variance in certain characteristics of data may affect the emergence and the breadth of data pooling initiatives.

Primarily, it shows that differently from the typical Interorganizational Information Systems where data exchange is collateral to another transaction, a reduction in the costs of capturing data generates an enclosure effect rather than favouring the emergence of broad collaborative

initiatives and that the emergence of multiple data pooling initiatives, instead of a unique industry wide solution, may be a preferable governance solution. Then, it suggests that features of data that have been associated with the potential success of the "Big Data Revolution" at the intraorganizational level, like data option value and data exhaust (Mayer-Schönberger & Cukier, 2013a), may instead have an hindering effect on the emergence and sustainability of interorganizational data pooling initiatives.

In addition, it suggests that there may be problems of different nature (lack of incentives (provision problem); excess of members (membership problem); and stag hunt problem) behind the non-emergence of an optimal data pool. Acknowledging the existence of this difference and recognizing the contingent conditions that lead to the emergence of these different problems, allows to better chose appropriate solutions to reduce them and to avoid polices that may be counterproductive. For instance, incentives for reduction in data collection costs may be beneficial in case non emergence of a data pool is due to an incentive problem, while they may be detrimental in case non emergence is due to a stag hunt problem.

To my knowledge, it is also the first attempt to employ computational methods to the application of club theory. Club theory has historically struggled in implementing its core theoretical intuition of simultaneity in determining the quantity of the club good to be generated and the quantity of actors to involve in the joint provision and use of the good, due to mathematical tractability problems (Sandler & Tschirhart, 1997). The present work shows the feasibility and the potential of simulation methods in being able to fully embody club theoretical intuitions and eventually in allowing to further expand their application without the limits of mathematical tractability.

The present dissertation is more generative than conclusive. Some of the intuitions presented in the dissertation need more solid confirmatory evidence. In other cases, some conceptualizations (especially derived from Chapter 1) have not been fully developed through the dissertation but may deserve a dedicated exploration outside of it. Further, chapter findings and limitations are generative of additional research.

Chapter 1 suggests that a market type solution for incentivizing the contribution to pools may be limitedly feasible. This hypothesis seems to be confirmed by the fact that among the observed cases in Chapter 2 none of them relied on a market mechanism to generate a pool. However, this speculation needs to be verified on a larger set of observations and in different empirical settings from the health sector and health data. For instance, the exploration in different settings may help in disentangle whether in absence of some of the features of data that are present in the health sector (sensitivity, option value, non subtractability) a market type of contracting may be instead feasible. In this respect, Chapter 2 also shows that the incentive problem is frequently addressed in health data pools by linking contribution with the recognition of authorship of papers derived from analysis on the pool. What happens in settings where this type of collateral mechanism of recognition of contribution is absent?

Chapter 1 introduces the distinction between *row pooling* and *column pooling*, suggesting that the latter type of pooling may need more complex governance solutions and that it may be less plausible for column pooling initiatives to emerge, due to higher complexity to trace the link between contribution and outputs, thus generating a team production problem (Alchian & Demsetz, 1972). De facto, in Chapter 2, only one case may be characterized as a column pooling initiative, while all the others were row pooling ones. However, the core focus of the analysis was not put on this distinction. As such, it may be interesting both to confirm this regularity

with a broader sample and to explore comparatively how the governance design of row and column pooling differ.

The whole dissertation assumes that data intensive research, benchmarking initiatives and the production of algorithms in the form of a product may be considered similar as they all rely on statistical principles for their generation. Consequently, it argues that most of the implications found in the dissertation can be identically applied across the three settings. However, while it is still complex to trace a net line between the three categories, most of the cases explored can be categorized under the first two (data intensive research and benchmarking initiatives). In this respect, a broader coverage of pooling initiatives that have the production of an algorithm in the form of a product as core purpose is needed to confirm the that the type of governance challenges and incentive structures they face is similar to the other more explored types of initiatives.

Chapter 2 shows a marked variance in the way pooling initiatives are designed. While recognition of this variance is fundamental for the exploration of which solution can be more effective, it is clearly nonconclusive for determining which of the solutions are most effective, and under what conditions. As already argued in Chapter 2, several questions follow from knowing what the elements of variance in the internal governance of data pooling initiatives are. Does granting contribution rights to data subjects instead of clinicians or ad hoc surveyors influence performance, sustainability and representativity of data pooling initiatives? Are these effects contingent on data features? Does decision by polyarchy or by committee affect actors' willingness to participate in the pool? What are the actors that prefer polyarchic methods of decision rather than committees? And especially, how do these differences in preferences affect the representativity and quality of the pool?

The simulation model in Chapter 3 employs a set of parameters which are based more on plausibility rather than on actual evidence. There seems to be very scant evidence on: 1) the actual costs that are involved in data production and in coordination among multiple organizations; 2) the most accurate shape of the data production function; 3) the "shape of the function" that maps the number of actors involved in the pool with the perceived risk of data violation (leakage, opportunistic secondary use, privacy violation). A better quantification of these factors may significantly help in calibrating the model so that aside from theoretical reasoning, the model may also have sounder and more precise empirical implications.

In sum, while the road to better understanding the governance of data pooling initiatives is still long, I believe that the present dissertation solidly paved its initial part.

BIBLIOGRAPHY

Acquisti, A., Lane, J., Stodden, V., Bender, S., & Nissenbaum, H. (2010). The Economics and Behavioral Economics of Privacy. In *Privacy, Big Data, and the Public Good* (pp. 76–95). New York: Cambridge University Press. https://doi.org/10.1017/CBO9781107590205.005

Adjerid, I., Adler-Milstein, J., & Angst, C. (2018). Reducing Medicare Spending Through Electronic Health Information Exchange: The Role of Incentives and Exchange Maturity. *Information Systems Research*, (February), isre.2017.0745. https://doi.org/10.1287/isre.2017.0745

Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction machines : the simple economics of artificial intelligence*. Harvard Business Review Press.

Alchian, A. A., & Demsetz, H. (1972). Production, Information Costs, and Economic Organization. *American Economic Review*, *62*(5), 777–795.

Arat, M., Arpaci, F., Ertem, M., & Gürman, G. (2008). Turkish Transplant Registry: A comparative analysis of national activity with the EBMT European Activity Survey. *Bone Marrow Transplantation*, *42*(SUPPL.1), 142–145. https://doi.org/10.1038/bmt.2008.144

Arrow, K. (1962). Economic welfare and the allocation of resources for invention. In National Bureau Committee for Economic Research (Ed.), *The Rate and Direction of Inventive Activity: Economic and Social Factors* (Vol. I, pp. 609–626). Princeton University Press. https://doi.org/10.1521/ijgp.2006.56.2.191

Bakos, J. Y. (1987). *Interorganizational Information Systems: Strategic Implications for Competition and Cooperation*. Massachusetts Institute of Technology.

Bates, D. W., Heitmueller, A., Kakad, M., & Saria, S. (2018). Why policymakers should care about “big data” in healthcare. *Health Policy and Technology*, *7*(2), 211–216. https://doi.org/10.1016/j.hlpt.2018.04.006

Bell, E. A., Ohno-Machado, L., & Grando, M. A. (2014). Sharing my health data: a survey of data sharing preferences of healthy individuals. *AMIA Symposium*, *2014*, 1699–1708. Retrieved from

http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed13&NEWS=N&AN=25954442

Bikard, M., Murray, F., & Gans, J. S. (2015). Exploring Trade-offs in the Organization of Scientific Work: Collaboration and Scientific Reward. *Management Science*, *61*(7), 1473–1495. https://doi.org/10.1287/mnsc.2014.2052

Borgman, C. L. (2015). *Big Data, Little Data, No Data*. Cambridge, MA: The MIT Press.

Brynjolfsson, E., & McAfee, A. (2014). *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*.

Buchanan, J. M. (1965). An Economic Theory of Clubs. *Economica*, *32*(125), 1. https://doi.org/10.2307/2552442

Buchanan, J. M., & Yoon, Y. J. (2000). Symmetric Tragedies: Commons and Anticommons. *The Journal of Law and Economics*, *43*(1), 1–14. https://doi.org/10.1086/467445

Butzkueven, H., Chapman, J., Cristiano, E., Grand'Maison, F., Hoffmann, M., Izquierdo, G., … Malkowski, J. P. (2006). MSBase: An international, online registry and platform for collaborative outcomes research in multiple sclerosis. *Multiple Sclerosis*, *12*(6), 769–774. https://doi.org/10.1177/1352458506070775

Chatterjee, D., & Ravichandran, T. (2004). Inter-organizational information systems research: a critical review and an integrative framework. In *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the* (Vol. 00, p. 10 pp.). IEEE. https://doi.org/10.1109/HICSS.2004.1265398

Chatterjee, Dipanjan, & Ravichandran, T. (2013). Governance of interorganizational information systems: A resource dependence perspective. *Information Systems Research*, *24*(2), 261–278. https://doi.org/10.1287/isre.1120.0432

Chen, H. (2013). Governing International Biobank Collaboration: A Case Study of China Kadoorie Biobank. *Science, Technology and Society*, *18*(3), 321–338. https://doi.org/10.1177/0971721813498497

Choinière, M., Ware, M. A., Pagé, M. G., Lacasse, A., Lanctôt, H., Beaudet, N., … Truchon, R. (2017). Development and Implementation of a Registry of Patients Attending

Multidisciplinary Pain Treatment Clinics: The Quebec Pain Registry. *Pain Research and Management*, *2017*, 1–16. https://doi.org/10.1155/2017/8123812

Chompalov, I., Genuth, J., & Shrum, W. (2002). The organization of scientific collaborations. *Research Policy*, *31*(5), 749–767. https://doi.org/10.1016/S0048-7333(01)00145-7

Chute, C. G., Hart, L. A., Alexander, A. K., & Jensen, D. W. (2014). The Southeastern Minnesota Beacon Project for Community-driven Health Information Technology: Origins, Achievements, and Legacy. *EGEMs (Generating Evidence & Methods to Improve Patient Outcomes)*, *2*(3), 16. https://doi.org/10.13063/2327-9214.1101

Clemons, E. K., & Hitt, L. M. (2004). Poaching and the Misappropriation of Information: Transaction Risks of Information Exchange. *Journal of Management Information Systems*, *21*(2), 87–107. https://doi.org/10.1080/07421222.2004.11045802

Cornes, R., & Sandler, T. (1996). *The Theory of Externalities, Public Goods and Club Goods*.

Cukier, K., & Mayer-Schoenberger, V. (2013). The Rise of Big Data. *Foreign Affairs*, *92*(3), 27–40.

Densen, P. M., Fielding, J. E., Getson, J., & Stone, E. (1980). The Collection of Data on Hospital Patients — The Massachusetts Health Data Consortium Approach. *New England Journal of Medicine*, *302*(3), 171–173. https://doi.org/10.1056/NEJM198001173020311

Devetag, G., & Ortmann, A. (2007). When and why? A critical survey on coordination failure in the laboratory. *Experimental Economics*, *10*(3), 331–344. https://doi.org/10.1007/s10683-007-9178-9

Drexl, J. (2016). *Designing competitive markets for industrial data - Between Propertisation and Access*.

Duch-brown, N., Martens, B., & Mueller-Langer, F. (2017). *The economics of ownership , access and trade in digital data* (JRC Digital Economy Working Paper 2017-01).

Elgarah, W., Falaleeva, N., Saunders, C. C., Ilie, V., Shim, J. T., & Courtney, J. F. (2005). Data exchange in interorganizational relationships: review through multiple conceptual lenses. *The DATA BASE for Advances in Information Systems*, *36*(1), 8–29. https://doi.org/10.1145/1047070.1047073

Everson, J. (2017). The implications and impact of 3 approaches to health information exchange: community, enterprise, and vendor-mediated health information exchange. *Learning Health Systems*, *1*(2), e10021. https://doi.org/10.1002/lrh2.10021

Feldman, S. S., Schooley, B. L., & Bhavsar, G. P. (2014). Health information exchange implementation: Lessons learned and critical success factors from a case study. *Journal of Medical Internet Research*, *16*(8), e19. https://doi.org/10.2196/medinform.3455

Floridi, L. (2010). *Information*. Oxford University Press. https://doi.org/10.1093/actrade/9780199551378.001.0001

Frischmann, B. M. (2012). *Infrastructure: the social value of shared resources*. New York: Oxford University Press.

Frost, J., & Morner, M. (2010). Overcoming knowledge dilemmas: governing the creation, sharing and use of knowledge resources. *International Journal of Strategic Change Management*, *2*(2/3), 172–199. https://doi.org/10.1504/IJSCM.2010.034413

Galbraith, J. R. (1974). Organization Design: An Information Processing View. *Interfaces*, *4*(3), 28–36. https://doi.org/10.1287/inte.4.3.28

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Gliklich, R., Dreyer, N., & Leavy, M. (2014). *Registries for Evaluating Patient Outcomes: A User's Guide*. Rockville (MD).

Grandori, A. (1991). Negotiating efficient organization forms. *Journal of Economic Behavior & Organization*, *16*(3), 319–340. https://doi.org/10.1016/0167-2681(91)90017-R

Grandori, A. (2001). *Organization and Economic Behavior*. https://doi.org/10.4324/9780203273814

Grandori, A. (2013). Epistemic economics and organization: Forms of rationality and governance for a wiser economy. *Epistemic Economics and Organization: Forms of Rationality and Governance for a Wiser Economy*, 1–156. https://doi.org/10.4324/9780203786772

Groves, P., Kayyali, B., Knott, D., & Van Kuiken, S. (2013). The “big data” revolution in healthcare: accelerating value and innovation. *McKinsey Global Institute*, (January), 1–22. https://doi.org/10.1145/2537052.2537073

Gulati, R., Puranam, P., & Tushman, M. (2012). Meta-organization design: Rethinking design in interorganizational and community contexts. *Strategic Management Journal*, *33*(6), 571–586. https://doi.org/10.1002/smj.1975

Hardin, G. (1968). The Tragedy of the Commons. *Science*, *162*(3859), 1243–1248. https://doi.org/10.1126/science.162.3859.1243

Harrison, J. R., Lin, Z., Carroll, G. R., & Carley, K. M. (2007). Simulation modeling in organizational and management research. *Academy of Management Review*, *32*(4), 1229–1245. https://doi.org/10.5465/amr.2007.26586485

Hart, P., & Saunders, C. (2008). Power and Trust: Critical Factors in the Adoption and Use of Electronic Data Interchange. *Organization Science*, *8*(1), 23–42. https://doi.org/10.1287/orsc.8.1.23

Heller, M. A. (1998). THE TRAGEDY OF THE ANTICOMMONS: PROPERTY IN THE TRANSITION FROM MARX TO MARKETS. *HARVARD LAW REVIEW*, *111*(3), 621–688. Retrieved from https://www.degruyter.com/view/j/gj.2003.3.1/gj.2003.3.1.1081/gj.2003.3.1.1081.xml

Hess, C., & Ostrom, E. (2003). IDEAS, ARTIFACTS, AND FACILITIES: INFORMATION AS A COMMON-POOL RESOURCE. *LAW AND CONTEMPORARY PROBLEMS*, *66*(1&2), 111–146. Retrieved from http://scholarship.law.duke.edu/lcp/vol66/iss1/5

Holmes, J. H., Elliott, T. E., Brown, J. S., Raebel, M. A., Davidson, A., Nelson, A. F., … Steiner, J. F. (2014). Clinical research data warehouse governance for distributed research networks in the USA: a systematic review of the literature. *Journal of the American Medical Informatics Association*, *21*(4), 730–736. https://doi.org/10.1136/amiajnl-2013-002370

Iacobelli, S. (2013). Suggestions on the use of statistical methodologies in studies of the European Group for Blood and Marrow Transplantation. *Bone Marrow Transplantation*, *48*(SUPPL.1), S1–S37. https://doi.org/10.1038/bmt.2012.282

Jarvenpaa, S. L., & Markus, M. L. (2018). Data Perspective in Digital Platforms: Three Tales of Genetic Platforms. *Proceedings of the 51st Hawaii International Conference on System Sciences*, 4574–4583. Retrieved from https://scholarspace.manoa.hawaii.edu/bitstream/10125/50466/1/paper0579.pdf

Kerber, W. (2016). A New (Intellectual) Property Right for Non-Personal Data? An Economic Analysis.

Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*. 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom: SAGE Publications Ltd. https://doi.org/10.4135/9781473909472

Kitchin, R., & McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*, *3*(1), 1–10. https://doi.org/10.1177/2053951716631130

Koutroumpis, P., & Leiponen, A. (2013). Understanding the value of (big) data. In *2013 IEEE International Conference on Big Data* (pp. 38–42). IEEE. https://doi.org/10.1109/BigData.2013.6691691

Koutroumpis, P., Leiponen, A., & Thomas, L. (2017a). *The (Unfulfilled) Potential of Data Marketplaces* (ETLA Working Papers No. 53). *ETLA Working Papers*. Retrieved from http://pub.etla.fi/ETLA-Working-Papers-53.pd

Koutroumpis, P., Leiponen, A., & Thomas, L. (2018). Data Strategy. In *Academy of Management Global Proceedings*. https://doi.org/10.5465/amgblproc.surrey.2018.0085.abs

Koutroumpis, P., Leiponen, A., & Thomas, L. D. W. (2017b). *How Control Instruments and Information Technologies Drove* (ETLA Working Papers).

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. Retrieved from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Malone, T. W., Yates, J., & Benjamin, R. I. (1987). Electronic markets and electronic hierarchies. *Communications of the ACM*, *30*(6), 484–497. https://doi.org/10.1145/214762.214766

Markus, M. L., & Bui, Q. "Neo." (2012). Going Concerns: The Governance of Interorganizational Coordination Hubs. *Journal of Management Information Systems*, *28*(4), 163–198. https://doi.org/10.2753/MIS0742-1222280407

Martin, K. D., Borah, A., & Palmatier, R. W. (2017). Data Privacy: Effects on Customer and Firm Performance. *Journal of Marketing*, *81*(1), 36–58. https://doi.org/10.1509/jm.15.0497

Mattioli, M. (2017). The Data-Pooling Problem. *Berkeley Technology Law Journal*, *32*(1).

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

Mayer-Schönberger, V., & Cukier, K. (2013b). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston New York: Houghton Mifflin Harcourt. Retrieved from http://www.amazon.com/Big-Data-Revolution-Transform-Think/dp/0544227751

McGlynn, E. A., Lieu, T. A., Durham, M. L., Bauck, A., Laws, R., Go, A. S., … Kahn, M. G. (2014). Developing a data infrastructure for a learning health system: The PORTAL network. *Journal of the American Medical Informatics Association*, *21*(4), 596–601. https://doi.org/10.1136/amiajnl-2014-002746

McGuire, M. (1972). Private Good Clubs and Public Good Clubs: Economic Models of Group Formation. *The Swedish Journal of Economics*, *74*(1), 84. https://doi.org/10.2307/3439011

McKinsey Global Institute. (2011). *Big data: The next frontier for innovation, competition, and productivity*.

McKinsey Global Institute. (2016). *The Age of Analytics: Competing in a Data-Driven World*.

Miller, A. R., & Tucker, C. (2014). Privacy Protection , Personalized Medicine and Genetic Testing. *Working Paper*, 1–34.

Mindel, V., Mathiassen, L., & Rai, A. (2018). The Sustainability of Polycentric Information Commons. *MIS Quarterly*, *42*(2). https://doi.org/10.25300/MISQ/2018/14015

Oderkirk, J., & Ronchi, E. (2017). Governing Data for Better Health and Health Care. *OECD Observer*, *309*(Q1), 19–20.

Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. (Cambridge University Press, Ed.), *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge.

Ostrom, E. (2005). *Understanding Institutional Diversity*. Princeton, New Jersey: Princeton University Press.

Ostrom, E., & Hess, C. (2007). *Understanding Knowledge as a Commons*. *Understanding Knowledge as a Commons From Theory to Practice* (Vol. 15). https://doi.org/10.1002/asi

Ostrom, V., & Ostrom, E. (1977). *Public Goods and Public Choices*. Indiana University, Workshop in Political Theory and Policy Analysis. Retrieved from https://books.google.it/books?id=nSypXwAACAAJ

Paolino, A. R., McGlynn, E. A., Lieu, T., Nelson, A. F., Prausnitz, S., Horberg, M. A., … Steiner, J. F. (2016). Building a Governance Strategy for CER: The Patient Outcomes Research to Advance Learning (PORTAL) Network Experience. *EGEMS (Washington, DC)*, *4*(2), 1216. https://doi.org/10.13063/2327-9214.1216

Petrova, M., Riley, J., Abel, J., & Barclay, S. (2018). Crash course in EPaCCS (Electronic Palliative Care Coordination Systems): 8 years of successes and failures in patient data sharing to learn from. *BMJ Supportive and Palliative Care*, *8*(4), 447–455. https://doi.org/10.1136/bmjspcare-2015-001059

Pisani, E., & Botchway, S. (2017). Sharing individual patient and parasite-level data through the WorldWide Antimalarial Resistance Network platform: A qualitative case study [version 1; referees: 1 approved], *6311*(0). https://doi.org/10.12688/wellcomeopenres.12259.1

Puranam, P., Raveendran, M., & Knudsen, T. (2012). Organization design: the epistemic interdepencence perspective. *Academy of Management Review*, *37*(3), 419–440.

Redline, S., Baker-Goodwin, S., Bakker, J. P., Epstein, M., Hanes, S., Hanson, M., … Rothstein, N. (2016). Patient partnerships transforming sleep medicine research and clinical care: Perspectives from the Sleep Apnea Patient-Centered Outcomes Network. *Journal of Clinical Sleep Medicine*, *12*(7), 1053–1058. https://doi.org/10.5664/jcsm.5948

Redman, T. C. (2018). If Your Data Is Bad, Your Machine Learning Tools Are Useless.

Harvard Business Review, 1–5. Retrieved from https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless

Reinsel, D., Gantz, J., & Rydning, J. (2017). *Data Age 2025: The Evolution of Data to Life-Critical*. *IDC White Paper*.

Rose, C. (1986). The Comedy of the Commons: Custom, Commerce, and Inherently Public Property. *The University of Chicago Law Review*, *53*(3), 711. https://doi.org/10.2307/1599583

Roski, J., Bo-Linn, G. W., & Andrews, T. A. (2014). Creating Value In Health Care Through Big Data: Opportunities And Policy Implications. *Health Affairs*, *33*(7), 1115–1122. https://doi.org/10.1377/hlthaff.2014.0147

Rumbold, J. M. M., & Pierscionek, B. K. (2018). What Are Data? A Categorization of the Data Sensitivity Spectrum. *Big Data Research*, *12*, 49–59. https://doi.org/10.1016/j.bdr.2017.11.001

Sah, R. K., & Stiglitz, J. E. . (1998). Committees , Hierarchies and Polyarchies. *The Economic Journal*, *98*(391), 451–470. Retrieved from https://www.jstor.org/stable/2233377

Sandler, T., & Tschirhart, J. (1997). Club theory: Thirty years later. *Public Choice*, *93*(3–4), 335–355. https://doi.org/10.1023/A:1017952723093

Sherman, S., Shats, O., Ketcham, M. A., Anderson, M. A., Whitcomb, D. C., Lynch, H. T., … Brand, R. E. (2011). PCCR: Pancreatic cancer collaborative registry. *Cancer Informatics*, *10*, 83–91. https://doi.org/10.4137/CIN.S6919

Stone, E. M., Bailit, M. H., Greenberg, M. S., & Janes, G. R. (1998). Comprehensive health data systems spanning the public-private divide: The massachusetts experience. *American Journal of Preventive Medicine*, *14*(3 SUPPL.), 40–45. https://doi.org/10.1016/S0749-3797(97)00045-7

Strandburg, K. J., Frischmann, B. M., & Cui, C. (2014). The Rare Diseases Clinical Research Network and the Urea Cycle Disorders Consortium as Nested Knowledge Commons, (14), 155–207.

Strandburg, K. J., Frischmann, B. M., & Madison, M. J. (2017a). Governing Knowledge

Commons: An Appraisal. *Governing Medical Knowledge Commons*, 421–429.

Strandburg, K. J., Frischmann, B. M., & Madison, M. J. (2017b). *Governing Medical Knowledge Commons*. (K. J. Strandburg, B. M. Frischmann, & M. J. Madison, Eds.). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316544587

Stub, D., Lefkovits, J., Brennan, A. L., Dinh, D., Brien, R., Duffy, S. J., … Reid, C. M. (2018). The Establishment of the Victorian Cardiac Outcomes Registry (VCOR): Monitoring and Optimising Outcomes for Cardiac Patients in Victoria. *Heart Lung and Circulation*, *27*(4), 451–463. https://doi.org/10.1016/j.hlc.2017.07.013

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., … Collins, R. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Medicine*, *12*(3), 1–10. https://doi.org/10.1371/journal.pmed.1001779

Susha, I., Janssen, M., & Verhulst, S. (2017a). Data Collaboratives as a New Frontier of Cross-Sector Partnerships in the Age of Open Data: Taxonomy Development. *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2691–2700. https://doi.org/http://hdl.handle.net/10125/41480

Susha, I., Janssen, M., & Verhulst, S. (2017b). Data collaboratives as "bazaars"? *Transforming Government: People, Process and Policy*, *11*(1), 157–172. https://doi.org/10.1108/TG-01-2017-0007

Thomas, L. D. W., & Leiponen, A. (2016). Big data commercialization. *IEEE Engineering Management Review*, *44*(2), 74–90. https://doi.org/10.1109/EMR.2016.2568798

Thompson, J. D. (1967). *Organizations in Action: Social Science Bases of Administrative Theory*. New York: McGraw-Hill.

Tutić, A. (2013). Experimental evidence on the theory of club goods. *Rationality and Society*, *25*(1), 90–120. https://doi.org/10.1177/1043463112463874

van den Broek, T. van den, & van Veenstra, A.-F. van. (2015). Modes of governance in inter-organisational data collaborations : Complete Research. *Twenty-Third European Conference on Information Systems (ECIS), Münster, Germany, 2015*, 0–12.

Varian, H. (2018). *Artificial Intelligence, Economics, and Industrial Organization*. Cambridge, MA. https://doi.org/10.3386/w24839

Vest, J. R., Campion, T. R., & Kaushal, R. (2013). Challenges, alternatives, and paths to sustainability for health information exchange efforts. *Journal of Medical Systems*, *37*(6). https://doi.org/10.1007/s10916-013-9987-7

Vest, J. R., & Gamm, L. D. (2010). Health information exchange: persistent challenges and new strategies. *Journal of the American Medical Informatics Association*, *17*(3), 288–294. https://doi.org/10.1136/jamia.2010.003673

Walji, M. F., Kalenderian, E., Stark, P. C., White, J. M., Kookal, K. K., Phan, D., … Ramoni, R. (2014). BigMouth: A multi-institutional dental data repository. *Journal of the American Medical Informatics Association*, *21*(6), 1136–1140. https://doi.org/10.1136/amiajnl-2013-002230

Weick, K. E. (1989). Theory Construction as Disciplined Imagination. *The Academy of Management Review*, *14*(4), 516. https://doi.org/10.2307/258556

Welch, E., Loaufi, S., Fusi, F., & Manzella, D. (2016). *Institutional and Organizational Factors for Enabling Data Access, Exchange and Use in Genomics Organizations*.

Whiddett, R., Hunter, I., Engelbrecht, J., & Handy, J. (2006). Patients' attitudes towards sharing their health information. *International Journal of Medical Informatics*, *75*(7), 530–541. https://doi.org/10.1016/j.ijmedinf.2005.08.009

Williamson, O. E. (1981). The Economics of Organization: The Transaction Cost Approach. *American Journal of Sociology*, *87*(3), 548–577. https://doi.org/10.1086/227496

Yasnoff, W. A., & Shortliffe, E. H. (2014). Lessons learned from a health record bank start-up. *Methods of Information in Medicine*, *53*(2), 66–72. https://doi.org/10.3414/ME13-02-0030

Zaletel, M., & Kralj, M. (2015). *Methodological guidelines and recommendations for efficient and rational governance of patient registries*. Ljubljana.