Amplifying the Value of Data

STRATEGY AND MECHANISMS TO EXCHANGE DATA
BETWEEN COMPANIES IN VALUE NETWORKS

Zur Erlangung des akademischen Grades eines
Doktors der Wirtschaftswissenschaften

(Dr. rer. pol.)

von der KIT-Fakultät für Wirtschaftswissenschaften
des Karlsruher Instituts für Technologie (KIT)

genehmigte
DISSERTATION

von

Wolfgang Badewitz, M. Sc.

Tag der mündlichen Prüfung: 5. Mai 2023
Referent: Prof. Dr. Christof Weinhardt
Korreferent: Prof. Dr. Timm Teubner

Karlsruhe, 2023

Table of Contents

List of Figures

List of Tables

List of Acronyms

AF Arbitrage Freeness.

ALS Alternating Least Squares.

BI Business Intelligence.

C Complexity.

CBF Content-Based Filtering.

CDO Chief Data Officer.

CF Collaborative Filtering.

CIO Chief Information Officer.

CM Capability Matchmaking.

CMft Cloud Manufacturing.

CRISP-DM Cross-Industrial Standard Process Data Mining.

CSR Corporate Social Responsibility.

DASC-PM Data Science Process Model.

DIH Data Intelligence Hub.

DIKW-Pyramid Data-Information-Knowledge-Wisdom-Pyramid.

DL Data Lake.

DLH Data Lakehouse.

DSS Decision Support System.

DWH Data Warehouse.

ERP Enterprise Resource Planning.

ETL Extract - Transform - Load.

FAIR Findable, Accessible, Interoperable, Reusable.

FG Fine Granularity.

FP Fraud Prevention.

GDPR General Data Protection Regulation.

GPN Global Production Network.

HQ Head Quarter.

IDSA International Data Space Association.

IoT Internet of Things.

IT Information Technology.

KBF Knowledge-Based Filtering.

LMF Logistic Matrix Factorization.

ME Market Engineering.

ML Machine Learning.

MRR Mean Reciprocal Rank.

NI Negative Information.

PD Price Differentiation.

PP Provide Privacy.

Q Quality Issues.

RS Recommender System.

RSh Revenue Sharing.

SCM Supply Chain Management.

SHAP SHapley Additive exPlanations.

T Transparency.

URI Uniform Resource Identifier.

WoS Web of Science.

WTP willingness-to-pay.

YODA Yale University Open Data Access.

Chapter 1

Introduction

> There will come a time when our descendants will be amazed
> that we did not know things that are so plain to them.
>
> Lucius Annaeus Seneca

The value of data is on tap but requires its utilization to flow (Hartmann et al., 2016). Fortunately, data persists beyond its use, and its value keeps flowing for other use cases. Data becomes even more valuable the more it is used and intertwined with more data from other sources (Moody and Walsh, 1999). This way, even greater value can be unlocked if the data is exchanged and shared with third parties. Besides a mutual bartering of data, it can also be monetized via direct selling (Wixom, 2014). In both cases, however, the economist must ask what mechanisms should be used in sharing and selling to achieve a desirable outcome for all parties. This is particularly true as companies remain reluctant to share their data, despite the potential benefits (Richter and Slowinski, 2018; Koutroumpis, Leiponen, and Thomas, 2020). To enable free and easy data exchange, researchers, (Abbas et al., 2021), practitioners (Spiekermann, 2019), and policymakers (European Commission, 2020) have directed their attention towards data marketplaces. The hurdles in this regard are many and varied, ranging from technical to economical concerns (Abbas et al., 2021). Of special concern are questions relating to the market design (Fernandez, Subramaniam, and Franklin, 2020), including problems of price determination and revenue allocation.

This dissertation investigates the economic layer of data exchange between companies and in value networks. It proposes a framework to differentiate between collaborative and competitive approaches to data sharing and presents a game-theoretical model to investigate incentive problems within cross-company data applications. On this base, both approaches are more closely analyzed in specific applications from a manufacturing network. This dissertation aims to be a piece of the puzzle for a better economic understanding of data-based collaboration and thereby contributes to the realization of data markets.

1.1 Research Motivation

Over the past decades, data has become an important sometimes even vital business asset across all major industries and domains (Marr, 2021). Such a development was enabled by unprecedented growth in quantity and capability. Data is being generated and collected in ever greater and previously unimaginable amounts. With today's computing power, it has also become possible to effectively process data via machine learning. Whereas previously small samples of data had to be carefully composed and analyzed with great skill and rigor, nowadays, the sheer quantity of data can be exploited – and in a fraction of the time. This development has also allowed working with more uncertainty and noise in the data (Cukier and Mayer-Schoenberger, 2013). The competitive advantage that is achievable through the utilization of big data is undoubted. Big data has long influenced various industries, such as commerce (Akter and Wamba, 2016), manufacturing (O'Donovan et al., 2015), agriculture (Kamilaris, Kartakoullis, and Prenafeta-Boldú, 2017), and supply chains (G. Wang et al., 2016). In figures, the European Data Economy, that aggregates the overall impacts from all steps of the data value chain, had a value of 442,57 Mio. € in 2021 and it is assumed that it will grow by 7.4% annually until 2025 (Glennon et al., 2022).

An awareness has already developed that one should not limit oneself to one's own data but should make greater use of the potential that it holds in combination with further and other data. One of the essential properties of data is that its value grows in combination with other data (Moody and Walsh, 1999). Shared data permits otherwise impossible or hard-to-get insights. This way companies are able to not only preserve their existing business but to create new value propositions (Wixom, Sebastian, and Gregory, 2020). To make an example, data sharing has always been of utmost importance for the effective and efficient operation of supply chains (Stefansson, 2002). In value chains that are highly differentiated, but have to finely intertwine, shared data works as a lubricant. More than ever a digital value chain accompanies the physical value chain (Hofmann and Rüsch, 2017). However, even beyond that, data from suppliers and partners, as well as the data a company collects itself can be utilized in smart services and products (Pflaum and Gölzer, 2018). Thus data sharing is also an important step towards truly smart manufacturing (Mittal et al., 2017). Although data sharing is already taking place today, it could be expanded considerably. Especially two issues are noteworthy. One shortcoming is that many companies only share data if it is indispensable, while generally keeping the bulk of their data confidential from others (Pujol Priego, Osimo, and Wareham, 2019). As a result, the demand for data is far bigger than the supply (Wauters et al., 2018). Secondly, there is a misconception that data sharing implies that all data sets are free and can be used by everyone (Scaria et al., 2018). On the contrary, there are many forms of data sharing besides open data. It is for example a viable approach to sell data as well as to barter it against other data or services (Wixom, 2014). In order to achieve a more intensive exchange of data, data marketplaces have been suggested (Bergemann and Bonatti, 2019; Abbas et al., 2021). A data marketplace combines several functions, but most prominent to the raised concerns it aggregates supply and demand for data, which ultimately allows for clearance. Furthermore, it specifies the conditions of data exchange. Data marketplaces are therefore in a pole position to facilitate the future data economy. In consequence, the attention of politics, business, and academia has started to rise.

The most notable political tailwind came in the form of the European Data Strategy. In it, the European Union envisions a single market for data that is supposed to boost growth and create value by enabling easy and fair access to high-quality data (European Commission, 2020). To this end, the Data Act has been proposed whose primary objectives include "[facilitating] access to and the use of data by consumers and businesses, while preserving incentives to invest in ways of generating value through data." The willingness of policymakers to interfere in data matters has already been documented by the far-reaching reform of the General Data Protection Regulation (GDPR), the Data Governance Act, and the Data Act. Notwithstanding, further regulatory work is needed regarding data marketplaces since not only privacy is affected but a range of economic, political, and social rights and entitlements (Taylor et al., 2020). An important discussion revolves around Data Sovereignty (Hummel, Braun, Tretter, et al., 2021; Braud et al., 2021). A notion that comprises abilities and rights to exercise control over data. In relation to this, there are also open questions regarding the ownership of data (Asswad and Marx Gómez, 2021; Hummel, Braun, and Dabrock, 2021).

Data Markets have also already taken root in practice and application-oriented research, which is presented in a glimpse. Dawex[1] is one of the successful companies in this field. The French start-up was founded in 2015 and has since launched a data marketplace and an exchange platform. While their marketplace works as an open, international, industry-agnostic platform to share data at different terms, their exchange platform allows building own solutions to facilitate data sharing within a network. The International Data Space Association (IDSA)[2] is a non-profit initiative to foster the ideas of secure and sovereign data exchange. IDSA brings together leading companies from various industries and countries, such as Siemens, and IBM, as well as renowned research organizations such as Fraunhofer, and Politecnico Milano. For now, their contribution is the development of a reference architecture (Otto et al., 2019) and the demonstration of use cases (International Data Spaces Association, 2022). Currently, the focus lies on the consolidation of similar projects and initiatives to yield a unified result in the mid-term and reach broad awareness and increased adoption over the course of 10 years (Nagel and Lycklama, 2021).

[1] www.dawex.com/

[2] internationaldataspaces.org

The research performed within such application-oriented projects finds its counterpart in the more theoretical research on data markets, which has also evolved over the course of the last years. The research spans a wide range from service, technology, organization to finance questions (Abbas et al., 2021). Significant numbers of research endeavors have been performed in the pricing of data (Badewitz, Hengesbach, and Weinhardt, 2022; Pei, 2022) since the value of data is not easily determined. The reason that this area still poses major challenges is the intricate nature of data as a commodity. When data is mentioned in these contexts, digital data is meant. The latter brings with it all the characteristics of a digital good. As such, it is infinitely shareable at almost no cost, which poses a challenge in itself as is demonstrated in the development of digital goods auctions two decades ago (Goldberg et al., 2006; Hartline and Karlin, 2007). However, data goods have peculiarities that distinguish them from "ordinary" digital goods such as music or software. Data can be combined with other data to new insights (Moody and Walsh, 1999) and data is only valuable if it provides actionable insights (Rowley, 2007). Once worthless data can become very valuable when new data becomes accessible or when it is made available to parties who can put the data into action. At the same time, data can be highly rivalrous, posing information asymmetries in transactions and a competitive advantage over contestants who lack information. Taken together, this makes the exchange of data between companies a captivating and highly relevant field of research.

1.2 Research Outline

Against this backdrop, this dissertation closer investigates the economics of data sharing. Too often the value of data is curtailed by siloed storage and secrecy. It envisions a world in which it is common to exchange proprietary data to bring them to use and value, and – most important – a world where those who share their data stand to benefit from it. Specifically, it focuses on companies in the secondary sector of the economy, most prominently manufacturing companies. This sector is coined by a high partition of labor and highly differentiated value networks. That form of collaboration has yielded high productivity that by far outweighs the cost of the increased organizational effort. However, it has also precluded the collection of big data across the whole value network for later use. Overcoming that obstacle, for instance with the means of data sharing, promises notable profit to any value network. At the same time, one has to realize that companies from this sector vary a lot regarding their digital capabilities and data literacy; some are even laggards in terms of digitalization. Data is a very sensitive asset. The risks of data sharing seem to be very threatening for small production companies with little expertise in digital matters, while the potential advantages seem theoretical. Thus it does not come as a surprise, that data is often

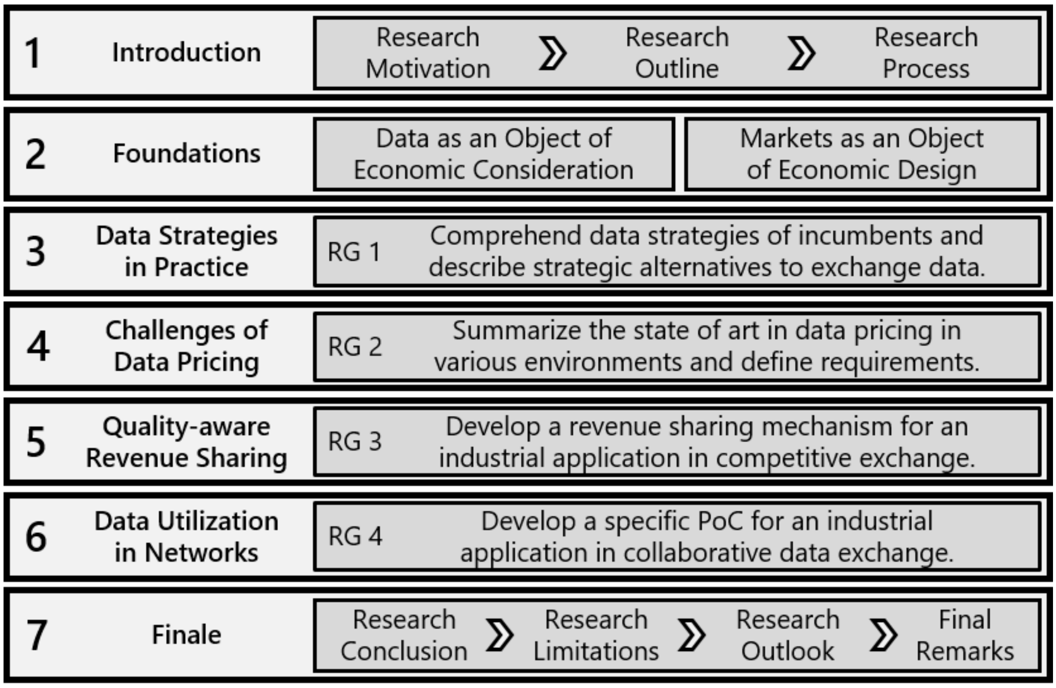

1
Introduction
Research Motivation
Research Outline
Research Process
2
Foundations
Data as an Object of Economic Consideration
Markets as an Object of Economic Design
3
Data Strategies in Practice
RG 1
Comprehend data strategies of incumbents and describe strategic alternatives to exchange data.
4
Challenges of Data Pricing
RG 2
Summarize the state of art in data pricing in various environments and define requirements.
5
Quality-aware Revenue Sharing
RG 3
Develop a revenue sharing mechanism for an industrial application in competitive exchange.
6
Data Utilization in Networks
RG 4
Develop a specific PoC for an industrial application in collaborative data exchange.
7
Finale
Research Conclusion
Research Limitations
Research Outlook
Final Remarks

Figure 1.1: The Dissertation Outline

only shared to a very limited extent. This is where this dissertation contributes. The presented research promotes data exchange between companies or business units in value networks in order to increase its usage and amplify its value for the broader value network. Therefore, four research goals are pursued in this dissertation. Each research goal is covered in a separate chapter, which should be comprehensive and comprehensible in its own right. The covered topics build up on each other and the chapters are ordered from top-level topics and general insights to specific questions, applications, and contributions. In total, this dissertation is divided into seven chapters, including this introduction, a foundation, and a conclusion to the collected work. Figure 1.1 provides an overview of the outline, which is explained in more detail below.

The foundations shed light on the theoretical background of this dissertation from a business and technological perspective. It is subdivided into two parts. First, the relevant research subject of data and information is presented as an economic good. It discusses the characteristics that distinguish data from physical goods and other more traditional digital goods and draws a conclusion about the nature of data as an economic good. Subsequently, principles and frameworks for the processing of data in a business context are presented. Secondly, the concept of markets is described and the art of Market Engineering (ME) is introduced. This has transformed markets from an object of analysis to an object of design. Important economic methodologies, specifically cooperative game theory, are closer discussed to lay the foundation for the mechanisms developed in later aspects of this dissertation.

Research Goal 1 *Comprehend data strategies of incumbents and describe strategic alternatives to exchange data.*

The first research goal is concerned with data strategies in companies from the second sector of the economy, especially manufacturing companies. Based on an extensive review of publicly available incumbent material and qualitative interviews conducted with domain experts the data strategies of manufacturing companies are described and structured. Further, alternative approaches to data sharing between companies are identified and analyzed for their respective risks and benefits. In this framework, a differentiation between competitive and collaborative data exchange, which is of major importance for the second part of this dissertation, will be made.

Research Goal 2 *Summarize the state of the art in data pricing in various environments and define requirements.*

The second research goal focuses on the existing body of knowledge about the pricing of data. Naturally, pricing is an important function in the monetary exchange of data, yet, it poses a major barrier to the commodification of data. That is because the value of data is hardly examined, which is due to its essential properties. Eleven challenges to price data are extracted from the literature and the existing approaches to price data are discussed. It also analyzes the underlying data and market structures in which existing research is thought. Thus the Chapter provides the reader with a comprehensive overview of the state of the art, serves as a toolbox, and gives impulses for further research in this dissertation.

Research Goal 3 *Develop a revenue-sharing mechanism for an industrial application in competitive data exchange.*

The third research goal deals with incentive compatibility issues in the competitive exchange of data. This is inspired by an industry use case in which data from upstream supply chain is used to optimize manufacturing. While the consumer of the data profits significantly from the application. However, the success of the application depends on the quality of data, which is subject to decisions of the providers, who do not benefit from sharing. While it is reasonable from the supply chain perspective to share the necessary data, this constitutes an inherent conflict between the data providers who wish for a high price, but low-cost data product, and the data consumer who wants a low price but high-quality data product. This problem is approached by modeling the data value chain first. The relevant prototype actors are identified and their economic interests lined out. Building on this understanding, the exchange of data between different actors has been modeled as a game in the sense of game theory. This 'Data Provision Game' is the basis for suggesting a pricing mechanism for data. Subsequently, it is evaluated whether that pricing mechanism is able solve the incentive issues and achieve a system-optimal solution.

> **Research Goal 4** *Develop a specific proof-of-concept for an industrial application in collaborative data exchange.*

The fourth research goal addresses collaborative data exchange. Therefore, an Decision Support System (DSS) to solve the specific problem of Capability Matchmaking is proposed. This is the question of whether a certain part can be produced by a certain machine. The DSS solves this task by the use of a Recommender System (RS). The general applicability of the approach is shown by using data from a use case partner. Moreover, the viability as a case of collaborative data exchange is discussed.

The Finale concludes this dissertation. An overall discussion is given on the overarching limitations of this dissertation and further research that is motivated by the results presented and identified limitations. Ultimately, the key results will be summarized, discussed, and reflected in the conclusion.

1.3 Research Process

Research is a fluid process and needs the exchange with other researchers. This includes collaboration with students, student assistants, colleagues, supervisors, and other researchers. Some ideas in this dissertation stem from conversations with fellow researchers and my supervisors, while other ideas arose during the supervision of students and student assistants and thus influenced their work as well. In acknowledgment of the influence others have had on my work, without which it would not have been possible, *'we'* is used throughout this dissertation instead of *'I'*. Recognizing the positive impact of discussion and presentations, many contents have also been published beforehand at scientific conferences. The research presented has been forged into its current form through the preparation of concise scientific articles. It has gained greater clarity through the feedback in multiple rounds of peer review. Given their genesis, the articles have been adopted in a complete form including their introductions and conclusions. However, the chapters were expanded and enriched in the aftermaths of conferences and in the light of new insights won since their first publishing. Chapters, which have been published before, are indicated through a disclaimer at the start of the respective chapter that points toward the publication. Further details on these publications can be found in Appendix A. This includes information on funding, co-authors, and my own contributions to the publication, as well as the added and made changes for this dissertation.

Chapter 2

Foundations

> To be ignorant of what occurred before you were born is to remain always a child. For what is the worth of human life, unless it is woven into the life of our ancestors by the records of history?
>
> Marcus Tullius Cicero

The objectives of this dissertation require knowledge in two fields. First, it is necessary to understand what data is, what properties it has, and how it can be utilized. Second, one must know what markets are, what parts constitute a market, and what steps have to be taken to analyze and design a market. This chapter deals with both, by focusing on introducing the topics and establishing the necessary background to convey the basic idea of the work. There is a conscious decision not to introduce tools and methods as they are presented in the method sections of the respective chapters.

In Section 2.1, we provide a comprehensive overview over important notions onto the concept of data and the most important models regarding its use. We introduce the DIKW-Pyramid (Rowley, 2007), which clarifies the very blurry notions of information and data, improving comprehension of the subtle qualitative differences. Next, laws of information (Moody and Walsh, 1999) and economic properties regarding the fourfold model of goods are examined. We finish with a comprehensive overview of the important FAIR principles (Wilkinson et al., 2016) and the CRISP-DM process model.

In Section 2.2, we will address the design of markets in general. Therefore, the discipline of Market Engineering (ME) is presented in its outline. First, the deliberate design of markets is motivated. Subsequently, the House of Market Engineering discusses the basic features that need to be considered and the implications that arise during the process (Weinhardt, Holtmann, and Neumann, 2003). The basic operating principle of mechanism design is further introduced (Hurwicz and Reiter, 2006) and builds a bridge to game theory, which is its own vast research field.

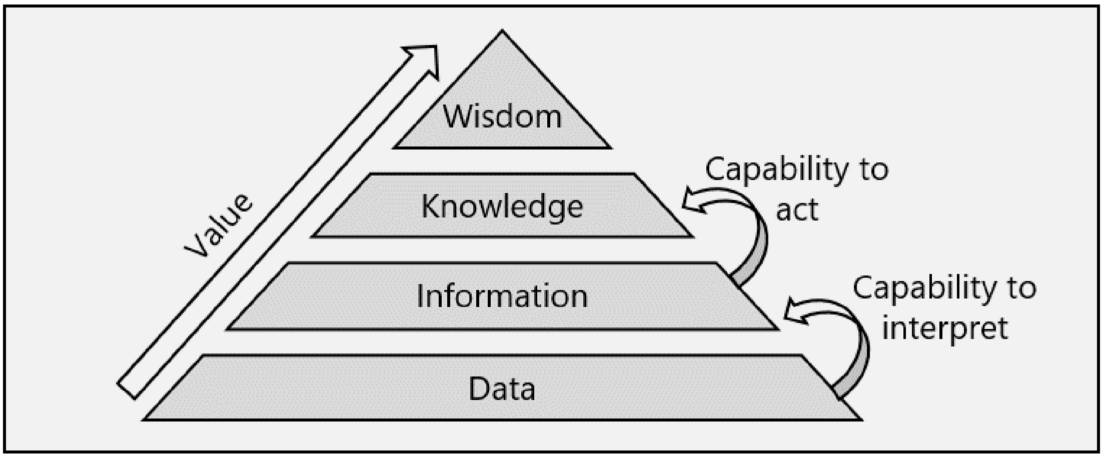


Figure 2.1: The Data-Information-Knowledge-Wisdom-Pyramid illustrates the hierarchy between those terms. Depiction adapted from Rowley (2007)

2.1 Data as an Object of Economic Consideration

Before we take serious steps to study data markets and mechanisms relating to the exchange of data, we need to clarify what is meant by data in the context of this research and what properties data has economically. For starters, we have to concern ourselves with three terms that seem to be clear but become very blurry when looked at closer. Those three terms are data, information, and knowledge. Since the understanding of digital data spans up to digital goods, like music or movies, that is *just data*, it also covers information, as is stored in digital encyclopedias and knowledge written down in digital books. Notwithstanding, data can be used to describe the rawest and uninformative state of any string of binary digits, which the terms information and knowledge cannot. To understand these three concepts and their interrelationships, the Data-Information-Knowledge-Wisdom-Pyramid (DIKW-Pyramid) is a profound guide. It is depicted in Figure 2.1 and here described after Rowley (2007).

The founding layer of this pyramid is data. As mentioned earlier, data in its simplest form is any string of characters. The definitions of data reduce it to this simple form. In order to delimit data from information, most definitions explicitly state what data is lacking: context, organization, interpretation, meaning, and value. However, data is not nothing. It contains discrete and recorded symbols, that bear the potential to be processed and understood. As such data is the elementary building block of the super-ordinate levels. Its value stems from being informative about something. Without data, it would not be possible to ascend the pyramid.

The counterpart to data is information, most often defined as data that has been processed in order to add what data had missing. The fundamental difference between the terms data and information is thus that information realizes some of its potentials. Information has a meaning that is understood and consequently has value for the recipient. The difference is therefore a matter of perspective. One piece of data can be unintelligible for some and informative for others, for instance, a text in a foreign language. Consequently, this text is data for the first group and information for the second. Information adds the capability to interpret data and with capability can be supported by giving the data structure and organization, this is by processing it into a form that can be easily understood by the recipient. If data is pre-processed in order to be more easily understood, this generates value for the recipients.

The definitions of knowledge vary a lot depending on the respective authors. Once again knowledge can be defined as a processed form of information and data. Definitions of knowledge tend to even more explicitly mention people and human contributions. In this understanding, knowledge is accumulated information, experience, or expertise that aids people to act in a certain way given a certain problem. So while information is about understanding data and grasping facts, knowledge allows you to derive action from the information. Knowledge is therefore also associated with skills, experience, and abilities though not all definitions embody that aspects. Knowledge is the ability of people to transform information into action, and thus generates value.

Wisdom is the final layer in the shape of the pyramid reproduced here. Sometimes understanding or enlightenment is also added to the top of the pyramid. It is linked with notions of intuition, foresight, transferability, and judgment. However, all of these concepts stay elusive and abstract to describe a state beyond the scope of information system research and knowledge management.

In this dissertation, the terms data and information are used mostly interchangeably. However, when the context requires a distinction between the terms "data" and "information," the corresponding terms are used. Yet, as has been shown, the transition from one state to the other is fluid, and it is common to use the terms "data" and "information" interchangeably. Moreover, we will always refer to digital data in a strict sense. That means that data is also abstracted from its carrier medium. It does not matter whether the data is stored in a cloud or on a hard drive, written in a book, or set in stone. What does matter is that it prescribes non-arbitrary signs, that encode a meaning (even if the receiver might not understand it), and that that set of signs can be duplicated to any other carrier medium.

2.1.1 Economic Properties of Data

For almost twenty years, big data was a topic in the middle of debate – scholarly as well as industrial. The term big data emphasizes foremost that in our modern times a lot of data is available in digitized forms. But big data is more than that: the main properties attributed to big data are volume, referring to the huge quantity of available data; velocity, referring to the increasing pace at which new and actual data is made available; and variety, referring to the immense scope of formats, which range from rigidly structured to totally unstructured (Sagiroglu and Sinanc, 2013; Kitchin and McArdle, 2016). Clearly, big data had an impact on the way data is thought about and dealt with (Cukier and Mayer-Schoenberger, 2013). However, big data describes primarily the technical attributes of modern data. The economic aspects are not included in these thoughts. Moreover, even in the times of big data, small data has not lost its importance. Two of the first to concern themselves with the economic properties of data were Moody and Walsh (1999). They introduced their seven laws of information, which were groundbreaking for the growing comprehension of what data and information are in an economic sense although only applicable in a general sense. Therefore, they are briefly paraphrased and discussed.

First, they claim that "information is (infinitely) shareable." Information in digital form can be reproduced at almost no cost and without destroying the original copy, which means that in principle anyone could own and use a copy. However, data is protected against free copying in many ways, both by technical means such as licenses and by legal means such as data protection laws. Second, they claim that "the value of information increases with use." Information has no value in itself, but its value stems from the informed action one can take based on it. The more actions are improved by using a piece of information, the more value it generates. However, whether an information is suited to inform an action and thus generates any value depends on context and quality of the information. Some limiting factors such as outdated or inaccurate data, as well as lack in processing capability or understanding are separately described. Third, they claim that "information is perishable." The potential value of data depreciates over time with its ability to inform actions taken now, i.e. they become outdated. The rate with which this happens vary a lot among different data assets. Moreover, the use case of data change over time. While recent data will be used to determine actions to take right now in a direct manner, historic data is of great use for analytic purposes, which influence actions only indirectly. Fourth, they claim that "the value of information increases with accuracy." More accurate information is of greater use to make good decisions. Inaccurate information can even have a negative value as it leads to wrong conclusions and actions. The slope of the relation can be assumed to be convex in most cases, meaning that increases in accuracy will have a larger positive effect on the value of low-accuracy information than on high-accuracy information. The point at which this saturation begins depends on many factors but especially the context; and keep in mind that information can be used in many contexts independently. However, since specific data is used in specific and limited contexts, and accuracy is not for free, this constitutes a tradeoff, which motivates the research work in Chapter 5. Fifth, they claim that "the value of information increases when combined with other information." The ability to interpret information is improved when it is integrated with other information. Combining information from multiple sources can support the generation of a more comprehensive and complete understanding. However, combining information also increases the complexity. Since not every piece of information has a considerable effect, there is the need of an assessment and selection process. Sixth, they claim that “more is not necessarily better.” The amount of data must remain within manageable limits if no negative consequences are to arise. This ics caused by the need to process information before it can be used. Thus, bottlenecks in processing can cause the information to not unveil its full potential. The phenomenon is also known as information overload. Decision-makers tend to seek more information they can process and still feel more confident about their decisions despite them being worse. It expresses the necessity to find a balance between information supply and processing capabilities. Seventh, they claim that “information is not depletable.” Since information stays available after use for further use, it is not scarce per se. Moreover, during usage, more information has been generated as a result of the processing and logging of the results. However, this only applies to information in general. Specific data might be nonetheless hard to collect and obtain. Once collected they do not deplete, but may loose relevance over time or loose their comprehensibility and accessibility due to a inferior curation.

While these laws are a proper starting point in understanding how the value of data behaves under certain circumstances and in correspondence to certain actions, they do not relate data to other economic goods and are in themselves not economic categories. A typical classification of economic goods is the fourfold model of goods (Kolmar, 2021; Vassilakopoulou, Skorve, and Aanestad, 2018; V. Ostrom and E. Ostrom, 2019). It classifies goods along the properties of excludability and subtractability into private, public, common, and club goods (see also Figure 2.2). Excludability is a measure of how well third parties can be denied the use of a good. Subtractability is a measure of how much the use of a good by one party affects the use of that good by others. Strong subtractability implies that a good cannot be used by multiple parties at all, while no subtractability implies that a good can be used by infinite parties without any restrictions. The concept of subtractability is often also represented as rivalry (Kolmar, 2021). Both dimensions are continuous, but four edge cases are typically defined. A private good is a good that is excludable and rival, e.g., an apple, which can be eaten once and effectively protected against others. A public good is quite the opposite, e.g., street lighting, which benefits everyone the same and cannot be withheld from anyone. A club good is excludable and non-rival, e.g., a golf course, which can be used by many but may restrict access. A common good is non-excludable but rival, e.g., fish stocks. Each fish can be caught exactly once, but it is very difficult to stop someone from fishing in the sea. An understanding of the properties of goods in these categories supports the design of an adequate mechanism that reaches efficient use of a resource. (Kolmar, 2021)

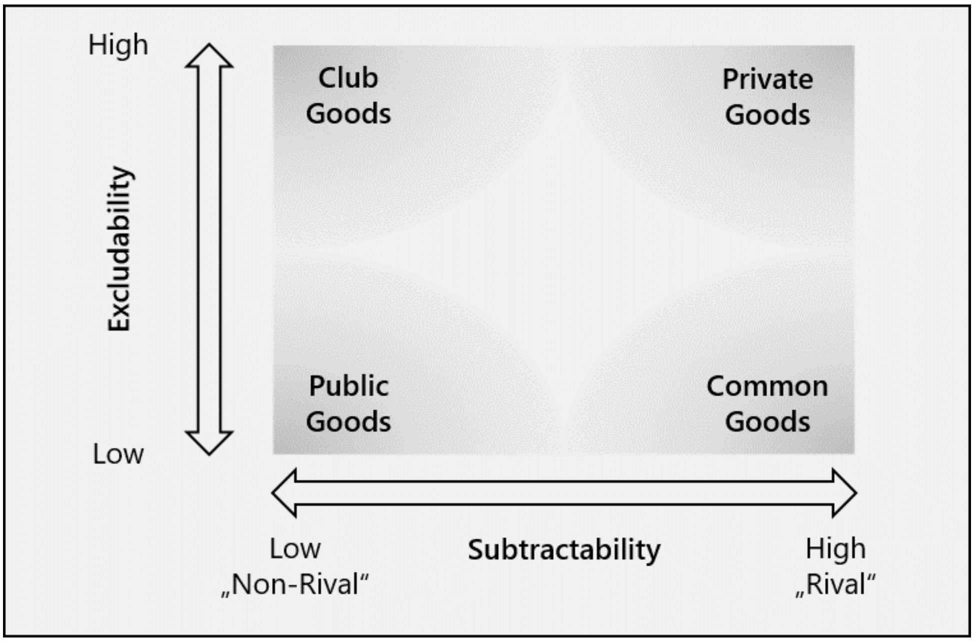


Figure 2.2: The fourfold model of goods illustrates the four edge cases of private, public, common, and club goods within the properties of excludability and rivalry. Depiction adapted from Kolmar (2021) and Vassilakopoulou, Skorve, and Aanestad (2018).

Digital Goods can be described as non-subtractable and partially excludable (Rayna, 2008). They are non-subtractable because they can be copied at insignificant cost. They are partially excludable because anybody owning a copy can themselves start duplicating the good, whereby the original producer loses their power over access control. Via technological, e.g., licenses, and legal barriers, e.g., anti-piracy laws, some power to excludability can be preserved, but the effort for exclusion is considerable. A good example of a digital good are movie files. Data is also a digital good, and in principle, the same arguments apply. Most articles on data sharing are assuming data can be shared in a non-rival manner, i.e., if data is shared the sharing party does neither profit nor lose anything from another one using the data (Badewitz, Hengesbach, and Weinhardt, 2022). For instance, Hess and E. Ostrom (2003) consider information as in most cases non-subtractable but are heavily concerned with the trend to increased excludability of information goods. However, data understood as information, respectively the raw form of information is not that evidently non-rival. If data is processed in order to gain or maintain an information-asymmetric situation (Akerlof, 1978; Stigler, 1961), they are rival in the sense that they are valuable only as long as they are not common knowledge, which is especially true for the initial use. For instance, research findings are only novel once (Vassilakopoulou, Skorve, and Aanestad, 2018). Therefore, research data is precious and not easily shared even though open data policies are common in science (Roche et al., 2014). Also in the business world, being the only one to use certain data can substantially add competitive advantage (Hagiu and Wright, 2020). Therefore, the economic ability to use data might be severely subtracted, though the technical ability to use remains unaffected by someone else's use. This is one important reason companies hold their relevant business data closed (Smichowski, 2016). It also motivates an investigation into the conflict of interest between data buyers and data sellers (H. Cai, Zhu, et al., 2019).

Concluding, data is not easily classified as an economic good. Much depends on the specific context in which data is collected and used. This fact also implies that the properties might subjectively change for one and the same data. Different data can behave more like a private good, e.g., data on individual customer interests on social networks; like a common good, e.g., open research data; like a public good, e.g., open governmental data; or like a club good, e.g., collaboratively curated master data. What situation applies must first be clarified, before the right measures to enable data exchange can be taken. In view of the above consideration, it seems impossible to design a one-size-fits-all market for data and information goods.

2.1.2 Dealing with Data

This section highlights best practices for dealing with data in order to reach its full potential. First, we take a look at the Findable, Accessible, Interoperable, Reusable (FAIR) principles (Wilkinson et al., 2016) that are concerned with leveraging the useful life of data by increasing its re-usability across time and stakeholders. Second, we explain the Cross-Industrial Standard Process Data Mining (CRISP-DM) that clarifies from a technical viewpoint the procedure in generating value from data and highlights the importance of varying contexts.

The FAIR principles stem from a 2016 paper in which more than 50 individuals from more than 40 institutions identified an urgent need to improve scientific data management (Wilkinson et al., 2016). They coined the abbreviation and the according principles as a guideline for data management with the ultimate aim to enhance the re-usability of data. This demand was initially directed at science and is part of the Open Science movement. As such it has also been supported by the G20 in 2016 to foster innovation in science and technology.[1] With the GO FAIR initiative[2] there exists an independent organization which promotes the FAIR data principles and supports the process of FAIR-ification. Meanwhile, companies have recognized the value of the FAIR approach to overcome barriers to collaboration and making links between partners in value networks (Vlijmen et al., 2020). The principles are described in the following based on their representation by the Go-FAIR initiative.

[1] `ec.europa.eu/commission/presscorner/detail/en/statement_16_2967` (accessed 2022-07-12)

[2] `www.go-fair.org/` (accessed 2022-07-12)

Findable In order to find data, data has to be "assigned a globally unique and persistent identifier." This is a key property to ensure that references to data are unequivocally understood. A good example is a Uniform Resource Identifier (URI). Further, data need to be "described with rich metadata." This allows finding data based on its content and context even if the identifier is unknown. However, the identifier has to be "clearly and explicitly included" in the metadata, so the connection between both can be made. Finally, both metadata and data have to be "registered or indexed in a searchable resource". Otherwise, data is only usable for those who are aware of the data resource beforehand.

Accessible Data and metadata should be "retrievable by using a standardized communication protocol." This ensures that there exist low technological barriers to data access. However, this does not rule out organizational barriers. Data is still accessible if it requires authentication and authorization though it must be documented how to authenticate and authorize. This principle applies even to deleted data, as long as the metadata is "accessible even when the data is no longer available." Metadata that indicates the data is no longer maintained is of great support to those trying to access it.

Inter-operable It should be possible to process data with common applications or in combination with other data. That does mean that data should "use a formal, accessible, shared, and broadly applicable language", which is ideally human and machine-readable. For this purpose ontologies are a suitable means. The used vocabulary however should "follow the FAIR principles" as well, meaning it has to guarantee the same ease of use. If multiple data sets are linked to each other, the metadata should "include qualified references" that explain the intent of the reference rather than simply stating the existence.

Re-usable All of the previously discussed principles serve the purpose of improving the re-usability of data. Yet, they are necessary, but not sufficient by themselves to reach a satisfactory level. The re-usability directly scales with the "plurality of accurate and relevant attributes" contained in the metadata. Those do not serve the purpose of finding the data but enable usage of the data by providing metadata generously to cover all possible customer needs.

The CRISP-DM, illustrated in Figure 2.3 is a domain-agnostic, standard procedure to approach data science projects. Although there are alternative models, such as the Data Science Process Model (DASC-PM) (Schulz et al., 2020), it is the de facto standard (Schröer, Kruse, and Gómez, 2021). Its value lies in the fact that it covers the entire process. It contains six phases, which are cycled through. The machine learning tasks in a closer sense are contained in the phases *data preparation* and *modeling*. Even more

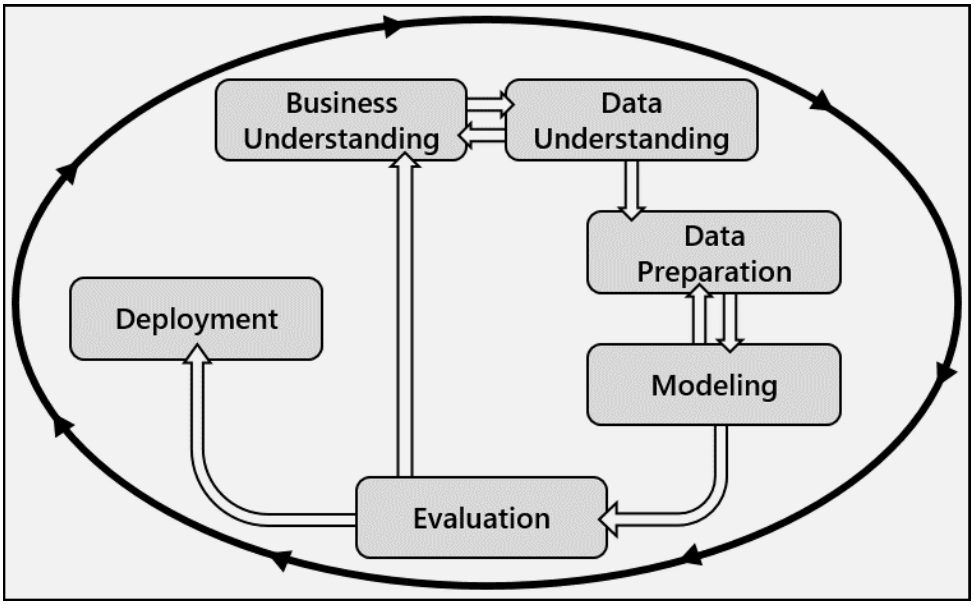


Figure 2.3: The CRISP-DM Cycle illustrates the state-of-the art approach to data projects. Depiction adapted from Shearer (2000).

important, however, is the background and feedback on business aspects. At the beginning of any data project, there is a thorough comprehension of the company's goals and an assessment of the business context (*'business understanding'*) intertwined with a collection, description, and exploration of available data (*'data understanding'*). The latter is very strongly supported by the FAIR data principles. The first point becomes even more important when data is to be shared among multiple parties, as we have already elaborated while examining the economic properties of data. While the *deployment*-phase concerns the successful implementation in operation and organizational learning, *evaluation* ensures a feedback loop to business. It assesses the success in reference to the understanding of the business. In a scenario with many stake-holding parties, we have to ensure in this phase that everyone is to benefit from the solution and therefore takes part in deployment. To guarantee the economic compatibility of the found technical solution, an economic design of the data exchange becomes crucial. This is the topic of the next subsection.

2.2 Markets as an Object of Economic Design

In the liberal line of thought, markets have been described as a result of spontaneous order (Luban, 2019). Patterns of behavior that constitute markets, e.g., rules of property and especially how a property changes hand, can evolve without and even against the will of a designer as the existence of black markets shows. However, spontaneous order merely represents a self-replicating and often stable state, not a superior system in itself (Sugden, 1989). If markets get institutionalized, the rules are either implemented as is – or get deliberately designed. Especially in electronic markets, each mechanism has to be coded. In this, every market owner has the chance to actively design the rules on his marketplace to his will. The art of building marketplaces that deliver quality outcomes is called ME. However, it is not a simple task. ME has to be done in unison with human action. A market that is not embraced by buyers and sellers will fade away. Market design is therefore a difficult matter. As A. E. Roth (2002) notes, "markets don't always grow like weeds—some of them are hothouse orchids". Every action might have unintended consequences (Vernon, 1979), which results in high responsibilities for market engineers as their design choices influence decisions on a broad scale. Complications in ME may arise regarding the strategic environment as well as the real behavior of market participants (A. E. Roth, 2002). Considerable research has been conducted on how to deal with these challenges. This section provides a brief overview of the main aspects and approaches.

2.2.1 Market Engineering

ME is – in analogy to traditional engineering but in contrast to traditional economics – concerned not only with the understanding of economies and economic behavior but also with the construction of viable economic settings. A market engineer strives to build a market. Gimpel et al. (2008) define ME as "the use of legal frameworks, economic mechanisms, management science models, and information and communication technologies for the purpose of (1) designing and constructing places where goods and services can be bought and sold; and (2) providing services associated with buying and selling." This definition shows by itself, that ME is a versatile field that requires interdisciplinary knowledge and approaches. A Market Engineer is typically confronted with a multitude of heterogeneous requirements incorporating technical aspects of the good as well as business expectations of stakeholders, and compliance with legal and societal norms. To think about markets in a structured way, the 'House of Market Engineering' is a valuable framework (see Figure 2.4). It sets the scope of ME and clarifies its components. Furthermore, the visualization already delineates the application fields of ME. In the following, the single aspects are presented with reference to the description from Weinhardt, Holtmann, and Neumann (2003).

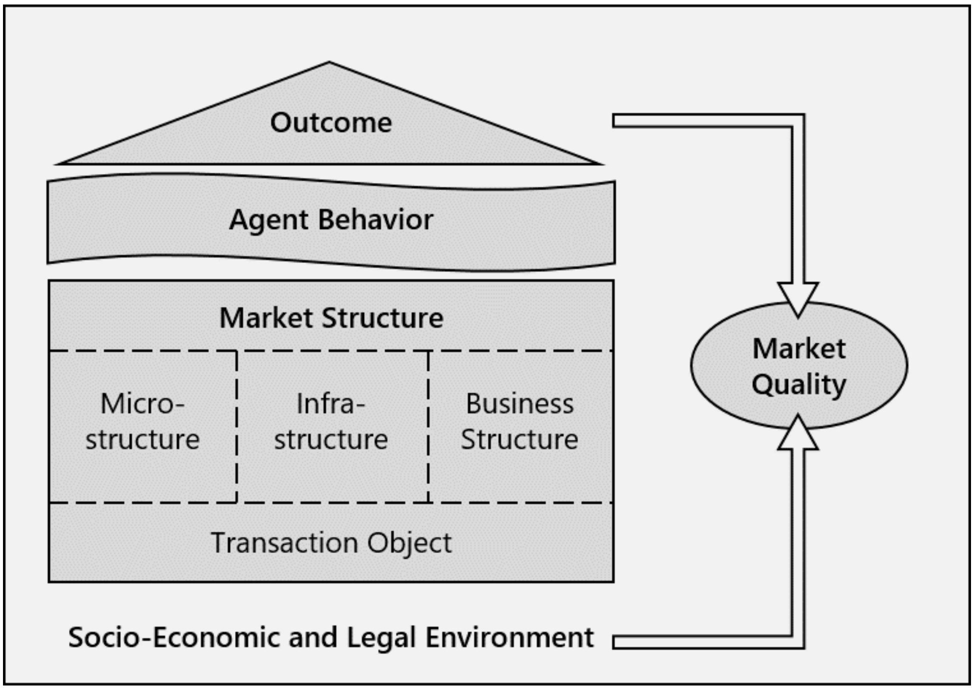


Figure 2.4: The House of Market Engineering illustrates different concepts within ME and their relationships. Depiction adapted from Weinhardt, Holtmann, and Neumann (2003); and Gimpel et al. (2008).

At the top level, markets consist of a market structure, the agents and their behavior, and an outcome. The outcome of a market is the allocation at the end of the trading process, i.e., who gets what at which price. The market is embedded in a socio-economic and legal environment. That environment is given and cannot be influenced by the market engineer directly, though indirect by influencing legislation. Many requirements stem from this environment. The market quality, i.e., whether the market outcome is desirable, can therefore only be assessed against the background of the socio-economic and legal environment. ME can be used for the design of markets as well as for the analysis, and especially to identify whether a market is a suitable means to a specified end in the first place (Gimpel et al., 2008). Thus, market engineers contribute to the body of scientific knowledge and provide rich practical implications for the operation of markets.

The market in a narrow sense is determined by the market structure, which contains the market rules, the technical implementation, the logic of the business, and the model of the trading object. All the adjusting screws available to the market engineer are located here. That begins with the transaction object itself. The product to be traded must first be made tradable, this is the distinctive properties of the good that have to be described. These are also the properties, which are subject to the negotiation and price-finding of participants. As an example, in stockbroking, market engineers have to make the design choice of whether securities are tradable only as whole numbers or in fractions. With reference to data as a transaction object, we have to make the design choice of whether we want to incorporate data quality into the product model and if so how to define it (see chapter 4).

The microstructure describes the abstracted mechanisms of how the outcome is determined. This includes which messages are exchanged, how the good allocation is selected based on them, and what payments have to be made. For instance, the mechanism of an English Auction can be described as follows: Given a starting bid at which the good remains with the seller, each bidder can submit a higher price than the last bid. If nobody is bidding for a certain amount of time, the last bidder gets the good and has to pay his bid, while everyone else pays nothing. In essence, the microstructure is synonymous with the design of the mechanism, which will be examined in more detail in the next subsection.

The infrastructure treats the technical implementation of the market, which means that the mentioned English Auction could be conducted in different infrastructures. A traditional approach would be an auction house with a physical presence. Electronic versions of the English Auction were implemented on peer-to-peer trading platforms, such as eBay. In the future, market engineers might want to implement an English Auction on a blockchain as a self-enforcing, smart contract. Without changing the microstructure, a different infrastructure might have quite an effect on the behavior of market participants.

The business structure clarifies the business logic, that is the business model of the market provider. This includes how to earn money with the market, e.g., with a fee or a commission. Further topics are how to get and keep participants on the market. Microstructure, infrastructure, business structure, and the transaction object can be investigated individually but in a market, they have to be put together and interfere with each other. Some mechanisms are unsuited for certain infrastructures. A commission can distort a carefully designed mechanism. A change in the model of the transaction object, for instance, trading music as digital files instead of discs might bring the need to adapt all three.

The market structure as a whole influences the behavior of market participants, i.e., those who participate in the market. Their behavior in turn influences the outcome of a market. This behavior cannot be directly influenced by the market engineer, but only through the market structure. Mechanisms always include incentives that, e.g., reward certain desired behavior and penalize malicious behavior. It is also possible to restrict certain behaviors by not including the corresponding technical features, e.g., minimizing informal agreements by restraining the possibilities of getting in contact with other participants.

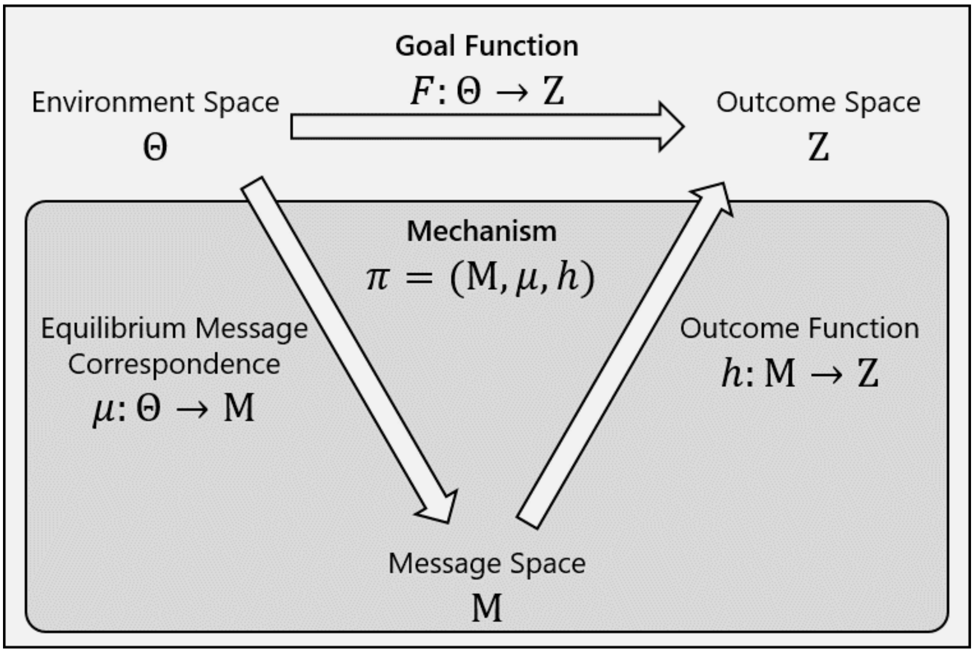


Figure 2.5: The Stanley-Reiter-Diagram above illustrates the mathematical view on mechanism design. Depiction adapted from Hurwicz and Reiter (2006).

A notable development regarding agent behavior is the emergence of behavioral economics. In the past, economic theory has generally assumed that actors act rationally. According to the rational principle, agents maximize their utility, collect and evaluate all information and weigh all their options correctly – also taking probabilities into account. However, this assumption is often inconsistent with findings from practice and from psychology. Therefore more general behavioral principles, which include psychological effects on the reasoning of agents have to be included. The rational principle retains its importance only as a special case and reference point (Camerer, 1999). The market engineer has to deal with these findings. At first, an analytical review of the market based on rational or bounded-rational agents is appropriate. However, testing and evaluation of markets in the field are indispensable before implementing and operating them large-scale.

2.2.2 Mechanism Design

Mechanism Design in a closer sense looks at the microstructure of markets and aims at creating rules that enforce the desired state. This subdomain of ME is coined by a very mathematical and formally abstract view of mechanisms. Figure 2.5 summarizes the single aspects important to the understanding of mechanisms. The following mechanisms are explained with reference to Hurwicz and Reiter (2006). A mechanism tries to implement a goal function that specifies which target states are desirable given specific circumstances. The goal function can also be referred to as social choice (Narahari, 2014) as it is not part of the design, but of the social context, which defines what is desirable. The goal function links the environment space Θ and the outcome space Z. The environment space covers all possible environments. In a single-good auction, that would be all possible vectors of the willingness-to-pay (WTP) of the bidders. The outcome space covers all possible outcomes of the problem. In terms of markets, this is generally the allocation. In our example, this is a vector that answers who gets the good, and who has to pay which price. The goal function then defines that given a specific environment, i.e., the WTP-vector, what allocations, i.e., winning bidder and payments, are desirable. For instance, only those outcomes are desirable in which the bidder with the highest WTP gets the good, he has to pay no more than his WTP, and all losing bidders have to pay nothing. This exemplary goal function incorporates Pareto-efficiency and ex-post rationality. This part of mechanism design is a question of modeling and goal setting. While the environment space is known, the specific environment is not. The design task is to come up with a mechanism, which implements the goal function indirectly without closer ex-ante knowledge about the environment.

Such a mechanism consists of a message space M, an outcome function h, and an equilibrium message. The mechanism is displayed in the lower part of the image in the shaded box. The message space M contains the vector of possible messages agents can send. In the auction example, the message space is classically the bids. However, the message space is part of the design. The message could be a bid that represents the revealed WTP of the agent. However, it could also be just a 'yes' or 'no'-message that represents an interest in the good in the first place. The outcome function determines which outcome is realized based on the specific messages received. On classical bids, one could determine that the highest bid wins and pays its bid. On the 'yes' or 'no' message space, one could determine that a lottery takes place between all 'yes'-votes and pays a previously fixed amount. The outcome function thus is also part of the design. The equilibrium message correspondence μ links mechanism design to game theory. All agent choose their respective message strategically with respect to the strategic choices of all other agents such that the reached outcome will fit their individual desire most. Thus, the equilibrium message correspondence is out of the control of the designer and subject to the strategic actions of the agents in a game. This game, however, is the designed mechanism.

2.3 Data and Markets as Guiding Thoughts

In this Chapter, we laid down the foundations of data markets. Both notions, *data as an object of economic consideration*, as well as *markets as objects of economic design*, accompany us through all chapters. Therefore, it is quite useful to keep the general thoughts of this chapter in mind while reading the following chapters. These serve as a fixed point in the pursuit of research objectives. While we explore specific questions in more detail, we are guided by the concepts and frameworks presented here and we discuss our findings and their limitations in the light of their context.

Data as an asset and how to best deal with them is a super-ordinate topic to the investigation of business data strategies (see Chapter 3). However, the aspects delineated in the House of Market Engineering come also into play when making strategic decisions regarding the internal and external exchange of data. The economic characteristics of data are of particular importance in integrating the challenges of pricing data assets (see Chapter 4) into an overall picture of that asset class. However, a thorough understanding of mechanism design is necessary to grasp the details of the reviewed pricing approaches. The value creation from data to information to knowledge encounters us again in the derivation of a data value chain and the development of the corresponding Data Provision Game (see Chapter 5). Mechanism design is applied in order to solve the raised incentivization problem. The CRISP-DM framework structured the development of the solution of the Capability Matchmaking problem (see Chapter 6). Since the developed solution serves as a proof-of-concept for collaborative data sharing at the same time, the implicit given internal mechanism is also part of the investigation.

Chapter 3

Data Strategies in Practice

> Strategy is making trade-offs in competing.
> The essence of strategy is choosing what not to do.
>
> Michael E. Porter

In light of big tech companies disrupting industries with their digital business models such as equipment-as-a-service, once-untouched industries such as manufacturing have recently come under pressure to digitize their businesses. Here, the right use of data is crucial, at times even vital, to digital transformation. Determining a company's future business depends on its ability to successfully transition to a data-driven enterprise, and this can only be achieved with a strategic approach.

In this Chapter, we shed light on the management of data by incumbents from a strategic perspective. This provides a good understanding of the goals and institutional framework in which the data will be used today and tomorrow. In particular, it includes a view of the alternative strategic approaches to data sharing and carves out the differentiation between competitive and collaborative data exchange.

The research is conducted via a structured search for company resources from a basket of incumbents and an interview study with domain experts. Therefore it is rich in practical examples and of high external validity. Each focus begins with a summary of research results providing the link to academia.

The contents of this chapter are adopted or taken from the working paper: Wolfgang Badewitz, Carl-Philipp Wachter, and Christof Weinhardt (n.d.) "Data Strategies of Industrial Incumbents: Worth a look into practice".
See Appendix A for further details.

3.1 Introduction

The importance of data and data-related capabilities for doing business in the 21st century is commonly understood and agreed upon. Reasonable efforts are undertaken to become and engage in data-driven initiatives. Today, the most valuable companies are big tech companies, e.g., Alphabet or Meta, whose business model is built on data (Marr, 2021). The customers of their products generate almost endless amounts of data. Understanding and analyzing this data, enables those companies to realize lots of use cases and gain high profits in advertisement and else. Traditional companies without such a history have serious trouble engaging in competition with digital-first companies as for example is illustrated by the pressure Tesla has put on the automotive market. To avoid a rude awakening when a digital competitor enters the market and to strengthen their own competitive position, companies need a data strategy that mitigates the risks of improper data handling, such as siloed data, redundant efforts, and information loss to competitors and promotes the benefits of optimized data management, such as leveraged new data sources, improved decision making, and new sources of revenue (Medeiros, Maçada, and Silva Freitas Junior, 2020; DalleMule and Davenport, 2017). However, too often companies and departments adopt digital processes and technologies without understanding their role in the context of the strategic objectives (Marr, 2021). Moreover, the need for a strategic approach to data management is often overlooked for many reasons (Fleckenstein and Fellows, 2018a). To get that link, companies need a data strategy, which we define as follows:

Definition 3.1. *A business data strategy describes an organization's general approach to data assets and all data processing activities, such as collecting, storing, analyzing, and applying data with the ultimate goal of efficient data utilization.*

We thereby illustrate the importance of a coherent, unifying, and integrative pattern of decision-making and long-term purpose (Hax, 1990). As a functional strategy, the main concern is with the use of the respective resource, i.e., data in the company, that has to be aligned with the overarching strategy (Leimeister, 2015). Decisions on data affect various aspects in the data value chain (Curry, 2016), which all have to be covered. In this Chapter, we investigated data strategies in leading traditional, i.e., non-tech companies and took lessons from the practice for the practice. Our guiding research questions were:

RQ 3.1 *What are typical manifestations of a data strategy among incumbent non-tech companies?*

RQ 3.2 *What lessons can be learned for future implementation in non-tech companies in general?*

By answering these two research questions, we provide implications for management for each of the deduced core components that are introduced at the end of the next section. We focused specifically on companies for whom data was not a core business in the past and who are now striving to become data-driven. In this way, we support companies that are also in the midst of the transition from a traditional organization to a data-driven one.

3.2 Method

We applied a mixed research design and approached our research questions from two sides. First, we studied publicly accessible materials from incumbents on their data

Table 3.1: Investigated Companies

HQ	Company	Industry*
US	3M	Conglomerates
DE	Adidas	Household & Personal Products
DE	Allianz	Insurance
US	Amgen	Drugs & Biotechnology
NL	ASML	Semiconductors
DE	BASF	Chemicals
DE	Bayer	Drugs & Biotechnology
US	Boeing	Aerospace & Defense
US	Caterpillar Inc.	Capital Goods
US	Chevron	Oil & Gas Operations
DE	Daimler	Consumer Durables
DE	Deutsche Telekom	Telecommunications Services
US	Dow Inc.	Chemicals
US	Honeywell International	Conglomerates
US	Intel	Semiconductors
US	Johnson & Johnson	Drugs & Biotechnology
IE	Linde	Chemicals
FR	L'Oreal	Household & Personal Products
FR	LVMH	Household & Personal Products
US	Merck & Co	Drugs & Biotechnology
CH	Nestle	Food, Drink & Tobacco
US	Nike	Household & Personal Products
CH	Novartis	Drugs & Biotechnology
DK	Novo Nordisk	Drugs & Biotechnology
US	Procter & Gamble	Household & Personal Products
CH	Roche	Drugs & Biotechnology
DE	Siemens	Conglomerates
US	The Coca Cola Company	Food, Drink & Tobacco

* according to Forbes

strategies. Therefore, we created a company sample and performed a structured review of the implemented data strategies by those companies. By following a structured process in the company study we have a documentation of our search process providing transparency and links to future research. It provides transparency over what companies for what reasons were included in our strategy review. Furthermore, we can be assured of having gathered a relatively complete view and enable other scientists to build up on our findings and expand the review. Since this is common ground for literature reviews (Brocke et al., 2009; Webster and Watson, 2002), we adopt the basic principles of our company strategy review. The foundation of our company corpus is the two stock indices StoxxEurope50 and Dow Jones Industrial. We excluded big tech companies and added further companies from Germany with a higher chance of winning interview partners. Table 3.1 provides an overview of all investigated companies. A central barrier in research was that no company has made its data strategy itself public, instead, we had to refer to manifestations of the strategy in public statements. This includes especially openly accessible materials, such as shareholder information, press releases, interviews with top management, or articles in business newspapers. Our search was conducted on Google Web Search. As search strings, we used *CompanyName AND Keyword* for all companies and keywords. Our keywords were “data strategy”, “data management”, “data asset”, “chief data officer”, “data architecture”, “data lake”, “data warehouse”, “data lakehouse”, “data mesh”, “data capabilities”, “data skill”, “acquisition” as well as “partnership” combined with further keywords referencing to digital and data-related industries, “data usage”, “data sharing”, and “data monetization”. Further, we have looked up more specific sources for topics that came up in doing those searches. The search took place between August 2021 and August 2022. All internet sources in this Chapter were accessed last on November 4th, 2022.

Second, we conducted a qualitative interview study with business professionals in the domain of data strategy, data management, or adjacent fields. This serves to validate our findings and integrate aspects that have been underexposed in the companies’ official announcements. Interviews were conducted using a semi-structured approach for which an interview guideline was developed. Specifically, we followed the method of qualitative expert interviews described by Kaiser (2014). This approach allowed us to balance between rigidness and openness (Myers, 2019). In contrast to structured interviews, which rigidly follow a questionnaire, and unstructured interviews, which are performed with no or only a few prepared questions, a semi-structured interview keeps a balance between those two extremes (Myers, 2019). A questionnaire is pre-defined, but emerging questions can change the direction and focus of interviewees. The results remain comparable, but issues that were not addressed when the questionnaire was created can be considered, and ideas and thoughts that emerge in an ongoing interview can be included. Therefore, a semi-structured interview study is suitable for an exploratory research task such as the present one. We had asked 390 potential experts, which we researched and contacted via LinkedIn, as well as over personal contacts. A total of 13 experts were won over for participation. An overview of the business background of the experts can be found in Table 3.2. In the end, we interviewed thirteen experts from eleven companies. Interview durations were between 27 minutes up to 56 minutes with a mean duration of 42 min. Interviews were transcribed and then read and annotated independently by two researchers.

We deduced five core components of a data strategy and organized our research along with them. The five components were:

1st *objectives of the data strategy*, an examination of the specific purpose, which is the core of any strategy (Hax, 1990);

2nd – 4th *creating responsibilities for data, designing the technical architecture, and gaining data and AI capabilites* resembling an analysis of humans, technology, and organizational aspects (Strohm and Ulich, 1997);

5th and lastly, *approaching data exchange*, referring to the increased importance of cross-organizational data economy (European Commission, 2020).

The results of our research are presented in the following sections along the core components of a data strategy. We start each section with the incorporation of literature on the respective topic and lead via the findings from the expert interviews to practical insights from the company study. We finish with a look at related research work and a conclusion on our own.

Table 3.2: Overview of the experts interviewed in the field of data strategy

Expert	Business Background	Managerial Position	Location (HQ)
E01	Data Science	yes	EU
E02	Data Science	no	EU
E03	Business Intelligence	no	EU
E04	Data Science	yes	EU
E05	Data Strategy	yes	EU
E06	Data Governance	no	EU
E07	Data Management	yes	EU
E08	Private Equity	no	US
E09	Innovation Management	yes	EU
E10	IoT Strategy	yes	EU
E11	Data Strategy	yes	EU
E12	Data Strategy	yes	US
E13	Data Management	yes	US

3.3 Related Work

Fleckenstein and Fellows (2018b) and Marr (2021) have written standard books on data strategy. Those works cover a wide range of topics, that obviously include the topics of this Chapter, esp. the ones about the implementation of a strategy.

Arcondara et al. (2017) has investigated the stock performance of CAC40 companies related to their corporate data involvement. It posed a severe difficulty in their efforts, that there was very little publicly available information about the data involvement, even on big data spending in the financial reports. Their analysis, therefore, was restricted to the existence of a Chief Data Officer (CDO) as a measure and resulted in an arbitrary picture.

Gür et al. (2021) developed a taxonomy of data strategy methodologies in two cycles from empirical to conceptual and from conceptual to empirical. In the first round, they analyzed publicly available whitepapers, insights, and reports on data strategy, which they searched via google. The second round consists of a structured literature review. They found 29 characteristics in 9 dimensions among those were, e.g., data assets and strategy implementation.

DalleMule and Davenport (2017) developed a framework from their experience as practitioners for what they call a robust data strategy. It emphasizes a differentiation between defensive and offensive parts of a data strategy. Defensive measures aim at mitigating risk and ensuring security, privacy, integrity, quality, regulatory compliance, and governance, while offensive measures aim at improving competitive position and profitability. Therefore they require different orientations of management, namely control, and flexibility, respectively, which are highly interwoven with the enabling architecture.

Wilberg et al. (2017) derived a framework, that supports the data strategy development for engineering companies motivated by the fact that specific data initiatives need a viable foundation to build upon which is created by a data strategy. Their framework comprises an analysis stage, which ensures the fit to the overall context of the company, such as business goals and stakeholders, and a concept stage, which evaluates specific data initiatives and finally picks those that should be implemented. The framework was developed based on nine existing process models for data initiatives.

Baecker et al. (2020) research data monetization strategies by analyzing 102 use cases from the industry which were identified from relevant literature retrieved in a structured literature review. With their system bottom-up approach, they claim to turn empirical results into theoretical evidence.

3.4 Objectives of the Data Strategy

The objectives, which were found in the company study and mentioned in the expert interviews can be divided into objectives related to super-ordinate objectives, that relate to the goals of the business strategy itself, and functional objectives in the area of data management. Super-ordinate objectives n the one hand aim at performance and results outside of the data domain and are basically answering the question “Why become data-driven?”. Functional objectives on the other hand stay in the data domain and relate to the processing of data within the company, basically answering the question “How to become data-driven?”. Super-ordinate and functional objectives often go hand in hand. As an example, Dow is striving to become a digital developer of new materials, which aims at innovating their products (super-ordinate level), but they want to do so by ensuring cooperation across the company and overcoming data siloes, which aims at increasing data accessibility throughout the company and creating a data culture to improve the connections between engineering, operations and IT (functional level).[1]

For both functional and super-ordinate objectives being ambidextrous is vital. Ambidexterity is the ability to pursue both efficiency and flexibility while balancing exploitation and exploration (Park, Pavlou, and Saraf, 2020). Exploration strategies target at new opportunities and innovate the existing business product-, process- and market-wise. Exploitation strategies target at improving the design of the current business and efficient management of resources (Sirén, Kohtamäki, and Kuckertz, 2012). Ambidexterity is important to an organization’s success and its survival (Jurksiene and Pundziene, 2016; Clauss et al., 2021). Overemphasizing exploitation can lead to a failure in meeting changing market needs in the long run; overemphasizing exploration may end up in constant new ventures without reaching profitability (Park, Pavlou, and Saraf, 2020). Though the differentiation of ambidexterity on functional and super-ordinate levels can be made, it is important to note that exploratory and exploitative capabilities go hand in hand, and ambidexterity is best achieved cross-functional (O’Cass, Heirati, and Ngo, 2014).

On a super-ordinate layer, ambidexterity poses the question of whether data is used to increase the efficiency of a company’s processes, e.g., increasing productivity, or to innovate the business, e.g., generating new value propositions. Undoubtedly data can be used for both goals, and some experts suggested doing so (E01, E10). Efficiency-targeted projects as well as exploration-targeted projects compete for scarce resources. To explore new opportunities, it is important to let data scientists play with the data and trust in their creativity (E03). Some experts report to focus on one or the other in their company, either on efficiency (E12) or on exploration (E09). Expert 01 recom-

[1] chemicalprocessing.com/articles/2020/digital-transformation-dow-breaks-down-cultural-and-organizational-barriers/

mends enabling exploration projects by defining quality gates and only continuing working on projects that passed. This way exploratory projects can be started without much ado but must prove their worth after a certain time.

An example of an efficiency-driven data strategy is Boeing. They focused on improving productivity on the factory floor and leveraging existing knowledge and capabilities. Using inventory demand signaling as an example, Boeing saved hundreds of millions of dollars.[2]

Adidas use customer data to create new products and follows a strategy of heavily individualizing them. To enable this approach also in the future, they develop apps, which are themselves exploring new opportunities, that serve as a direct touch-point to customers in order to better understand their needs.[3]

On a functional layer being ambidextrous implicates having a data landscape at the company that is efficient to get the data operations up and running, but also being flexible regarding new tools and concepts regarding the management of data. Efficiency in data operations is important in order to ensure data is accessible and usable (E03). For instance, focusing on a tidy data model first might delay some projects, but in the long run, increases the re-usability and yields flexibility for further projects (E12). Thus efficiency in data terms can be interpreted as a hygiene factor. This notion was made before in literature regarding the concept of efficiency as a whole (Clauss et al., 2021). In contrast, Expert 07 reported that they use and try multiple technologies to stay flexible and enable the dynamic development of the data landscape in the company, thus exploration is seen as necessary for future efficiency. Expert 13 points out that tools should serve a single purpose in order to be efficient, but adds that one must remain flexible and observe possible applications that could fulfill the purpose as well, and switch from one tech stack to another as needed. Brought together, efficiency and flexibility play into each other. At each layer, the base has to be efficient to enable flexibility, but flexibility at each layer is needed to ensure long-run efficiency. Finding a balance in this trade-off is a challenge and a sign of good management (E09).

Procter & Gamble has undertaken the act of finding a balance in the development and implementation of their data platform. Before, the different European units had a low degree of standardization. Subsequently, there was a debate between preserving the variability and flexibility and becoming harmonized and standardized across countries and thereby also lifting synergy effects and becoming more efficient. This was resolved in a global core Data Lake (DL) with region-specific data hubs. (Datar, Sarah Mehta, and Hamilton, 2020)

[2] mckinsey.com/business-functions/mckinsey-analytics/our-insights/data-as-jet-fuel-an-interview-with-boeings-cio

[3] cmo.com.au/article/629662/adidas-taps-data-technology-smarts-build-personalised-digital-engagement-consumers/

3.5 Creating Responsibilities for Data

Traditionally, non-tech companies view their data as a description of their relevant business assets, but not as an asset itself (Perrons and Jensen, 2015). As a consequence, data is managed and utilized with far less discipline than traditional assets (Gartner Inc., 2017). Changing the view of data from being an auxiliary resource to being a business asset comes with several advantages. In order to detach data from its original context and lead it to new fields of application it is necessary to consider and manage them independently. Building on this data can develop their full potential and change the value creation of organizations for the better. By turning data into an asset, a company increases the attention on the effective use of data to achieve organizational objectives and creates the need to leverage and manage data effectively (Collins and J. Lanz, 2019). Further, wrapping data into well-defined assets increases the data liquidity, the ease of reusing and recombining data, and thus the utility (Wixom, Piccoli, and Rodriguez, 2021). However, if data is not managed with the same care and rigor as other assets, there exists the danger that a company is swamped in data: Having high cost to maintain databases, but not knowing what data exists, where, and in what quality, whether it is allowed or possible to use it in certain circumstances and multiplying the necessary effort of high-cost and skilled personnel doing the groundwork. Because data has become so critical to the successful operation of a business, the future development and sometimes even the existence of companies are at risk if they do not take an asset view of their data (Marr, 2021).

The first step in moving to an asset view is to inventory existing data (Gartner Inc., 2017; Collins and J. Lanz, 2019). This results in a data catalog, which manages various data assets in a central metadata directory. The catalog has to be kept accurate and curated. Potential use cases shall use the catalog to identify data assets for their purpose (E06, E07). The catalog serves the purpose of increasing data efficiency. Data should be discoverable and easily consumable by potential use cases. A catalog makes this process easier, faster, cheaper, and more transparent (E06). The catalog can also avoid duplicate work, and associated hidden costs, if newly generated data sources during a data science project, e.g., during pre-processing, are imported into the catalog (E06). The invested effort into a data asset has to reflect its value for the company. A coordinated project to ensure equal access and documentation for all data assets is doomed to fail because of the utter amount of data and the low value of most data. A solution is to categorize data assets and differ in the management of data assets of different categories (E04, E06). With this setup, it allows the orchestration of the data landscape (E04) and building bridges between different assets from various parts of a company (E02, E04). A traditional relational catalog can be extended or augmented by a graph database with the advantage to incorporate various relationships between data (E11). E06 points out that an asset view and the catalog support data governance in an organization. Rules and requirements regarding transparency about data can be defined, enforced, and traced in a catalog. For instance, if personal data assets are recorded the risk of violating against privacy legislation is mitigated (E10). The catalog itself has to be filled by the respective departments which are the owners of the data assets. The data governance team creates and monitors compliance (E06).

Though the necessity of a data catalog has become clear, its exact scope is due to a negotiation process between many objectives (E01, E09, E12). Especially a balance between efficiency and granularity has to be found (E12). A too small and coarse catalog might not contain relevant data assets or needed metadata, a too big and detailed catalog poses the risk of outdated metadata and might be too complex for effective use. In both cases, the catalog cannot fulfill its purpose of informing stakeholders about the available data assets and might end up as a dead document. The identification of the most important data sets must be based on the business context since the value of data assets stem from their use case (E01, E13).

Hence a thorough understanding of the domain, as it is part in most data science cycles, e.g., CRISP-DM or DASC-PM, is mandatory. It is advised to think use-case driven and prioritize data assets that target an existing and known business challenge (E01). Thus, the explicit definition and prioritization of business cases support the classification of data assets. Further, it is important to notice that the work on a catalog is never finished (E09). It is a living document in which new data assets are introduced, old ones have to be maintained and updated, and obsolete data assets have to be deleted. With the rise of new technology, e.g., specific ontology software solutions, the ideal scope of a catalog might also become bigger in the future (E09). A broader implication of the asset view is the call for a measurement of the value of data assets (Gartner Inc., 2017). Such a financial evaluation would highlight the economic importance of data assets, increasing the pressure for appropriate management of them. Although it would be principally possible to assign a value to data, accounting struggles to do so (Collins and J. Lanz, 2019). None of the experts knew about an existing or aspired financial assessment of data in their companies. This is mainly for uncertainty about legislative and tax implications (E05, E07, E11). However, a financial evaluation would be valuable for a cost-benefit calculation (E09) and to divide data-related costs between users (E07).

It’s a common goal among organizations to treat data as a strategic asset in the future (Wixom, Piccoli, and Rodriguez, 2021). This is confirmed in our company study. Representatives from L’Oreal, P&G, Honeywell, Siemens Healthineers, and Nestlé, and others have publicly spoken of data as a valuable, strategic asset. Nestlé, for instance, has developed a framework for transforming data to an asset with the goals of improving transparency about the available data assets, standardizing linkage and exchange opportunities for internal and external re-usability, and ensuring data security aspects.[4] All interviewed experts, but one, knew about a data catalog in existence or development. Yet, most catalogs still lack relevant parts of the data landscape in their respective companies (E02, E03, E07, E09, E11) even after considering that it is neither possible nor reasonable to record all data.

After the creation of an according management category, the data asset, a responsible management role, the CDO, should be installed. The prototype CDO is an important senior executive who addresses the appropriate use of data in an organization and is responsible for planning, implementing, operating, and evaluating the data strategy and related policies and projects. Appointing a CDO has gained increased popularity. Based on credible assessments from business insiders 2012 only 12% of surveyed companies had a CDO in 2012, while it were about 70% in 2020 (Treder, 2020). The obvious reason for this development is the increasing relevance of data as a business asset that requires top management attention. Having a CDO supports increasing the visibility and awareness among managers, promotes a cultural change in the company, and improves data governance (Earley, 2017). The detailed tasks of a CDO were investigated by (Nie et al., 2018), who analyzed over 400 job descriptions of CDO's. In their research data analytics and business management are the both top tasks of a CDO, followed by new IT solutions, and enterprise strategy. In accordance with established management theory, it is essential that the CDO has the necessary competencies to fulfill his tasks and the responsibility to do so. This is known as the congruence principle (Bach et al., 2012). A data strategy that set the rules for company-wide data governance needs a CDO, which is capable to enforce it (E04). The same expert reported that the position of the CDO at their company has meanwhile been abolished for its lack in decision power. E01 emphasized that especially in small and medium enterprises this principle is often violated and therefore the company's stance is not pushed forward in practice.

In reality, most CDOs may not be at board level, but at a still senior level. On lower levels, the position might be still titled CDO but a lot of other titles occur, for instance, Head of Data Strategy, Head of Data Management, Head of Data Analytics, etc. Nonetheless, it is desirable to place the CDO as high as possible in the hierarchy (E03, E04, E13). E13 emphasized this as a prerequisite for a well-defined, company-wide operationalization of the data strategy. Also, the tasks of a CDO might be subsumed within the activities of a different or differently titled top-level manager, especially the Chief Information Officer (CIO), but also the Chief Digital Officer (also CDO). In contrast to this practice, E13 argues that the CIO and CDO shall be separated, because traditional IT and data compete about resources while pursuing different goals.

Our company study has revealed that from our company set none has a CDO directly employed in the Executive Leadership. However, seven companies from our set have executive positions that are close to that of a CDO. For instance Boeing, 3M, and

[4] aidataanalytics.network/business-analytics/articles/how-nestle-optimized-the-data-to-ai-pipeline

Honeywell. In 23 companies, including the seven from before, we found some persons in positions that are similar to a CDO position. In case they are below board-level, they are yet naturally restricted in their competencies across the whole group and sometimes focused on specific subtasks, e.g., data architecture or analytics. This complies with the result from the expert study. Experts from five companies reported that the position of a CDO exists, but in no case on top management level. Two more experts reported that central units exist, but no top-level CDO. However, both expressed the wish to install one.

3.6 Designing the Technical Architecture

Besides the organizational responsibilities to deal with data, a company needs a technical architecture for the storage of its data. Data Warehouses (DWHs) and Data Lakes (DLs) have been competing architectures for the past (Shiyal, 2021). Both comprise benefits such as segregation from transactional systems, a single point of contact, and storage of historical data, but they differ in certain aspects. A DWH stores structured, processed data, which makes them a rather inert alternative, but highly efficient for repetitive Business Intelligence (BI) tasks. A DL on the other hand store data in all forms at low cost and get new data quickly at the cost of bypassing the transformation step. A DL is suitable for modern data science but runs the risk of becoming a data swamp as a result of the lack of data and metadata management in their default settings (Brackenbury et al., 2018; Hai, Geisler, and Quix, 2016). Recently, new trends in data architectures have become prominent, namely the Data Lakehouse (DLH) and the data mesh. The DLH is ought to combine the advantages of DLs and DWHs, implementing structure and management features from the latter, but agility and adaptability from the former (Shiyal, 2021). Building architectures that are able to support a single source of truth on a data layer, as well as multiple versions in the management of information, is crucial to balance control and flexibility (DalleMule and Davenport, 2017). In 2017, they suggested using DLs for this, but the lake house might better fit the expressed need. The lakehouse does so by comprising a lake under a common metadata, caching, and indexing layer (Armbrust et al., 2021). It allows to perform a process step on the raw data already loaded into its lake part and store it in a warehouse part, thereby raw as well as processed data are part of it (Oreščanin and Hlupić, 2021). These changes allow for easy, efficient, and reliable data management with one point of control (Shiyal, 2021; Armbrust et al., 2021). Although monolithic, the lakehouse is compatible with the inherently decentralized approach of a data mesh (Armbrust et al., 2021; Oreščanin and Hlupić, 2021). The data mesh is another novel paradigm that emphasizes the domain-focused and decentralized management of arbitrary data within an organization and in between them. It is built upon the four principles of domain ownership, data as a product, self-serve data platform, and federated computational governance (Zhamak, 2022). Data responsibility and storage is intentionally distributed among or left at multiple mesh nodes over which the appropriate data teams have control. This resembles product thinking in the data world. However, a strong federalized computational governance ensures interoperability and prevents chaos (Machado, Costa, and Santos, 2022).

In theory, a centralized solution targets to utilize a bigger share of the company's data by making them findable and accessible. It is a key advantage that formerly siloed data from certain business functions or subdivisions become available for analysis on the top-level (E01, E02, E13). For instance, it can fortify the negotiation power with vendors that serve multiple of a company's subdivisions (E13). A centralized solution reduces the costs of having countless different integration (E12). Centralized solutions avoid redundant storage capacity, as well as redundant labor in data integration. Yet, expert 06 gave to consider that centralized solutions might cause unnecessary Extract - Transform - Load (ETL) processes, that have to be governed if data is needed in one place only. In certain cases, decentralized storage would prove beneficial. Additionally, the risk of having inconsistencies between multiple decentral versions of the same data can be mitigated (E12). Moreover, there still exists the risk that formerly siloed data is not transferred and hosted in shadow IT or get lost. Today, it is possible to integrate and access data in a compatible format from decentral sources as well, enabling said benefits but with novel architectures.

At this point, the expert study indicates that the trend toward centralization has peaked. This is in line with the claims from literature (Armbrust et al., 2021; Machado, Costa, and Santos, 2022). Although almost all experts report that their companies have some centralized databases running, they mainly agree that in the future the management of decentralized solutions will become more relevant (E01, E02, E04, E07, E09). Data is still valuable even if they are not centrally stored (E01). Consequently, it is a learning from the past to leave the data where they are and focus on technologies that enable the extraction of insights from the data in place and only transfer the insights (E04). The benefits of centralization can be achieved through virtual logic at the access and management level, while data storage can remain decentralized (E04, E06, E07, E12). The data mesh that implements this idea is followed with interest by the experts. A data mesh streamlines the process of making changes to data sources while maintaining data governance standards, especially when adding new data sources (E09). The bigger vision of the data strategy and the rules of data governance remain in place, but are adaptable for local sources and allow for necessary differentiation in the handling of data (E04). Consequently, the data mesh increases agility and broadens the covered scope of the enterprise-wide data landscape, as well as prevents the emergence of shadow data. In contrast to central solutions, it simplifies the search for the error and the person responsible. This includes that problems with the data quality can be solved directly at and by the source (E09). Additionally, since a high number of parties have to be involved, successfully implementing a centralized architecture comes with high transaction costs needed to solve issues between them (E11). Therefore a decentralized paradigm is most probable the more time- and cost-efficient alternative (E11, E09). The mesh can serve as the top structure of grown structures and increase the acceptance of decentral data which also drives their utilization (E01). Often this is the only solution to combine and utilize this data at all (E02). The key to productive use of the data mesh is to focus on interoperability (E06), a single point of contact (E12), and a good data dictionary (E02). Since it is most probably highly improvident to implement a completely centralized and holistic storage solution, it is highly unlikely to implement such a solution in any company (E09). Whether a central or decentral solution is implemented depends on the use-case (E01, E13). This complies with what most experts report to be the status quo at their companies: single central DLs and warehouses for specified purposes and endeavors to build a mesh-like platform to access decentral data sources (E01, E02, E04, E06, E08, E09, E10, E13).

The company study revealed that companies had done severe efforts to build and maintain a centralized data architecture in the past. Evidence stems from Nestle, Novartis, Deutsche Telekom, and P&G. Nestlé has built a centralized BI suite around a DWH powered by SAP Hana and Analytical Tools by Microsoft Azure.[5] Novartis works on a giant DL comprising two million patient years of data as well as researched data that sum up to 20 petabytes. This DL is a "crucial asset" for their research endeavors and shall free up their data scientists and enable to find revolutionary new insights.[6] Deutsche Telekom is trying to break up siloed data from different domains, e.g., customer service, finance, or sales, and transfer them into one centralized DL in order to achieve better processes for a cheaper price.[7] P&G also builds a core DL that comprises data from all across the company. Derived on that core DL several regional or functional diverse data hubs are derived and can be supplemented with additional data (Datar, Sarah Mehta, and Hamilton, 2020), marking a consciousness for decentralizing. P&G's CIO Vittorio Cretella explicitly rejected the data mesh concept at least for the moment, saying it was not yet ready for scaled deployment.[8]

In contrast, Roche just opted to deploy a data mesh concept to satisfy its data needs and to modernize its architecture. Moreover, it aligns with the decentralized and empowering corporate culture and supports the steady and rapidly growing number of employees who produce or consume data. According to them, the data mesh is very

[5] `aidataanalytics.network/business-analytics/articles/how-nestle-optimized-the-data-to-ai-pipeline`

[6] `novartis.com/stories/data42-program-shows-novartis-intent-go-big-data-and-digital`

[7] `telcotitans.com/deutsche-telekomwatch/cfo-werner-in-box-seat-on-telekoms-data-strategy-revamp/2420.article`

[8] `techinformed.com/procter-gamble-on-scaling-ai-for-enterprisex/`

much scalable and fits the modern agile product-oriented way of working. However big challenges aside in the congruence of data domains and business functions as well as in the associated corporate change to federated governance.[9] Likewise, Adidas is working on an implementation of the data mesh concept. They refer to scalability as a key advantage, especially because it helps to mitigate bottlenecks in the central team responsible for managing data. Among the issues with the traditional approaches, they report a mismatch to the product-led strategy, inferior discoverability of data assets, data quality issues resulting from the disconnection of a data producer and consumer, and problems resulting from a lack of central domain knowledge. They came across the trade-off between governance rigidity and agility. Loose governance will negatively impact data quality, while rigid rules will slow down data operations. As a middle ground, they suggest scaling quality control in proportion to the potential reusability of a data set. Other challenges involve incentives for data producers, enabling real-time applications, and maintaining interoperability as more nodes are added.[10] Bayer is a company implementing the lakehouse concept. They want to tackle the scalability, connectivity, and flexibility shortcomings of their former applied on-premise solution and circumvent bottlenecks in the central team. They allow for self-service management and assign the responsibility of data to decentral teams. Yet, the lakehouse stays a single platform that provides transparency over who does what and supports governance.[11]

3.7 Gaining Data and AI Capabilities

A very important role in being successful as a data-driven company is to have the right personnel. Data-related skills often belong to the best paid[12] and staff with those skills is hard to hire. This problem is even worse for industries that are not among the major players in the digital economy. Data workers in non-tech companies face different challenges and obstacles than their colleagues in tech companies. This is true in terms of corporate culture, as well as technological advancement. Digital companies often kept a start-up-feeling, with lots of freedom in everyday work life, high autonomy of single workers and flexibility of the organization and operating in a strategic offensive mode, while traditional manufacturing companies have more standardized processes and rules, are slower to change and operate in a defensive mode. Many companies have started upskilling initiatives to tackle the shortage of data scientists in the job market. Chevron can serve as an example. They have launched a six-month data science development program, including a company-wide exchange between ex-

[9] snowflake.com/blog/data-mesh-perspectives-a-qa-with-roche-diagnostics/

[10] medium.com/adidoescode/Introduction-to-data-mesh-adoption-in-adidas-85b1db812fa2

[11] databricks.com/dataaisummit/session/modern-architecture-cloud-enabled-data-and-analytics-platform

[12] indeed.com/career-advice/resumes-cover-letters/skills-employers-look-for

perienced and inexperienced data scientists. Additionally, they run internal data science competitions open for all their employees to attract attention and find hidden talent.[13] A second obstacle in hiring staff is the necessary combination with domain knowledge. Domain knowledge is of uttermost importance to successful data projects. While there already is a lack of skilled data workers, this becomes even worse for data workers with domain knowledge in a certain domain, and even worse if the domain gets more niche. On the other hand, the companies already have subject-matter experts. Ted Colbert, the CEO and then CIO of Boeing, emphasizes that it needs employees who already have necessary in-depth insights into a company to drive success and that it might be easier to equip domain experts with data skills than the other way round. Moreover, huge potential would be awaiting, if the capability is democratized throughout the company.[14] Upskilling can take place in different forms. The simplest form is to grant employees access to online course platforms. A little less lightweight are onsite and offsite training. Multiple experts report this kind of initiative for their companies (E02, E03, E05, E09, E12, E13). The upskilling not only targets technical skills but also knowledge about existing solutions inside the company and shall encourage employees to attain a data-driven mindset. A second notable form are internal communities dedicated to data-related issues. Such communities provide a team spirit to its members and the opportunity to learn from each other as well as a point of contact for other employees, to pose questions and get support in data-driven issues (E03, E12). Further opportunities are development programs, job rotation and collaboration with external partners (E01).

The efforts to become a digital enterprise can be substantially accelerated by forming strategic partnerships with big tech companies. Big tech companies and their product portfolio play a very important role in the digitalization of many companies. Thus it is not surprising that many companies under investigation are customers of and adjust their data efforts to work with those solutions. However, multiple examples could be found, where companies have broken out of a pure customer relationship and formed strategic partnerships with big players. Microsoft and Chevron entered a seven-year partnership in 2017, establishing Microsoft azure as their main provider of cloud services. However, even more than in an ordinary customer relationship, they focus on leveraging collaborative innovation capabilities and learning from each other. While Chevron expects to leverage its capabilities in high-performance computing and IoT, Microsoft wants to strengthen its services to meet the challenges of harsh operating conditions and highly integrated enterprises. For this purpose, employees of both companies will train each other.[15] In 2019, the partnership was further

[13] rigzone.com/news/oil_gas/a/145001/chevron_oil_gas_cos_should_date_not_marry_big_data_tech_vendors/

[14] mckinsey.com/business-functions/mckinsey-analytics/our-insights/data-as-jet-fuel-an-interview-with-boeings-cio

[15] chevron.com/stories/chevron-partners-with-microsoft and news.microsoft.com/

extended to Schlumberger, a technology provider of the oil and gas industry. Together, the three companies are building a customized, open platform to deliver significantly more value from data.[16] LVMH and GoogleCloud have formed a strategic partnership in AI and Machine Learning (ML) in 2021. In addition to utility-oriented goals, the partnership is intended to combine the strengths of both companies in their respective fields and promote co-innovation. A further focus of the agreement is on building data skills and enhancing culture change. Part of the agreement are upskilling and certification programs for LVMH employees and especially launching a Data and AI Academy in Paris, the headquarters of LVMH.[17]

Lastly, strategic acquisitions can build and enhance the capacities of companies. Acquired companies should bring in complementary capabilities, patents, and human resources. They often allow one to leap forward in an area but require a good implementation in existing processes. Acquisitions are not a sure way to success as many examples also are expensive missteps. Intel serves as an example of a company with an extensive acquisition and ventures capital strategy, i.e., Intel Capital. Intel strives for the necessary capacity for the IoT-future, and one of the means is acquiring several companies with complementary expertise in order to fasten its way. In 2016, Intel acquired Itseez, known for its computer vision expertise, further strengthening its portfolio on autonomous driving, digital surveillance, and industrial inspection.[18] This acquisition laid the ground for later endeavors in this area. Subsequently, Nervana Systems was acquired in 2016 supplementing Intel's deep learning soft- and hardware stack.[19] In 2020 another acquisition, Habana, which also is in deep learning chips, replaced the Nervana products.[20] While Intel might have increased its capacities via the first acquisition, later acquisitions for the same capacities have been necessary in the view of its management. However, it is expensive to repeatedly replace old purchases with new ones to keep up with developments in the data field.

3.8 Approaching Data Sharing

Data Sharing is commonly understood as an umbrella term that describes "all possible forms and models underpinning B2B data access or transfer" (European Commission, 2018). External parties are the obvious beneficiaries of sharing, while at the same time it is clear that opening up data inherits risks for the data owner. Thus, it seems advisable to follow a closed approach, marking everything as confidential and only publishing and transmitting data if necessary, e.g., for legal compliance. Indeed most

transform/chevron-fuels-digital-transformation-with-new-microsoft-partnership/

[16] chevron.com/newsroom/2019/q3/chevron-schlumberger-microsoft-announce-collaboration

[17] r.lvmh-static.com/uploads/2021/06/pr_lvmh_google.pdf

[18] newsroom.intel.com/editorials/intel-acquires-computer-vision-for-iot-automotive/

[19] newsroom.intel.com/editorials/foundation-of-artificial-intelligence/

[20] newsroom.intel.com/news-releases/intel-ai-acquisition/

companies follow a closed approach (Pujol Priego, Osimo, and Wareham, 2019). However, companies can profit from a less restrictive approach to data sharing. Sharing data is essential for companies to fully realize the value creation potential digitalization offers (Wixom, Sebastian, and Gregory, 2020). When multiple data streams are combined and interoperability challenges are solved, their value can be leveraged, leading to competitive advantage and additional revenue (Visconti, Larocca, and Marconi, 2017). Supplier, partner, and customer behavior can be influenced by sharing or not sharing data, which can improve operations in many ways. Therefore, data sharing measures are a strategic tool for the efficient management of supply chains and distribution channels. Ultimately, this can lead to a significant new source of revenue (Najjar and Kettinger, 2013).

Companies have shared transactional data that was required to deliver the value proposition of a contract at all times. Wixom, Sebastian, and Gregory (2020) coined the term 'data sharing 1.0' for this kind of sharing. They describe it as minimalistic and reduced to the necessary exchange of data. However, they line out that companies today engage in 'Data Sharing 2.0', which they define as "sharing complementary data assets and capabilities to create new value propositions" (Wixom, Sebastian, and Gregory, 2020). This is characterized by organizations increasingly exchanging data and using shared data that serves to complement existing data that is accessible to the recipient and serves the purpose of filling data gaps. While this differentiation is motivated by the functional role of data, there is not only the decision of whether and which data to share or not, but you also have to think about how you share data.

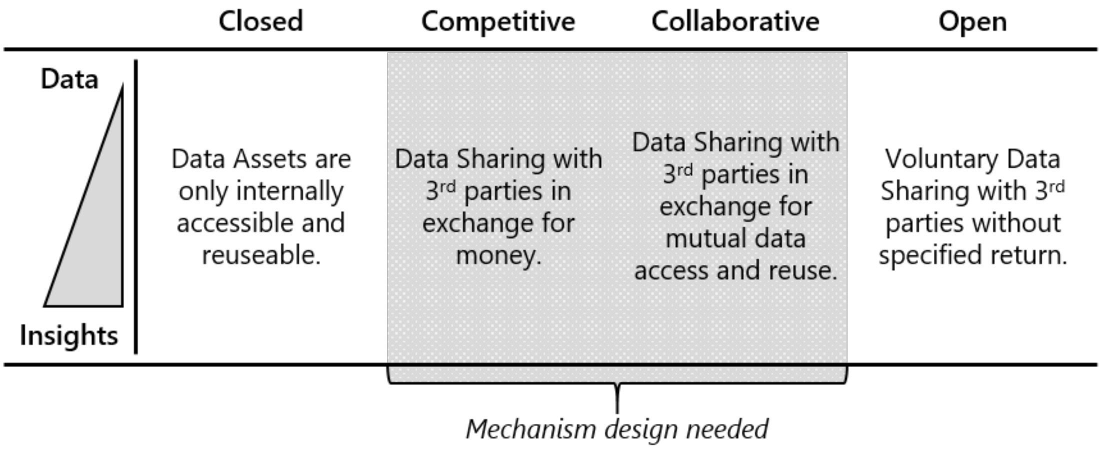
Wixom (2014) differentiate three forms of data monetization, which she defines as 'the act of exchanging information-based products and services for legal tender or something of perceived equivalent value' and thus fit with the concept of data sharing. The 'Wrapping' form comprises opportunities to enrich services and products with data and insights, thereby strengthening the competitive stance of that core offering. The other two forms of data sharing are more explicit about data and not necessarily related to an existing business. Their essential difference is what is received in return. In 'Selling', companies directly exchange data against money. In 'Bartering', companies get tools, services, or special deals in exchange, for instance, that could also be other data. (Wixom, 2014)

While it is easiest to use data only internally and wrapping builds on existing offerings, selling data is the most difficult approach because it requires a new business model (Wixom and Ross, 2017). Later, Woerner and Wixom (2015) supplemented the perspective of 'digital transformation', in which companies are grouped together in business ecosystems with blurring corporate boundaries and are based on information platforms. Most often Data Sharing 2.0 takes place in collaborative ecosystems similar to this (Wixom, Sebastian, and Gregory, 2020).

Walker (2015) describe four strategies to monetize data. First, 'Keep the data proprietary' comprises mostly internal uses of data, e.g., to improve operations or to enter new businesses, but also an external use of data via exclusive licensing in a direct relationship. Second, 'Trade data to business partners for shared benefits' comprises the joint usage of data, especially in supply chains and both upstream as well as downstream. In either case, the goal is to achieve mutual benefits. Third, 'Sell the Data Product (to a host of possible clients)' comprises selling data about assets to the owners of these assets, similar to the wrapping approach, but also other interested partners, more similar to the 'Selling'-Approach. The option for the up-sale of a premium data product flanked by free services is also featured within. Fourth, 'Make the data available (and even free) to many users' means opening up data and data-driven services to many users and monetizing them through advertising strategies, which are discussed based on various characteristics of the underlying data.

Baecker et al. (2020) has identified 12 strategies for monetizing data, five of which involve a data sharing approach. 'Asset Sale' and 'Data Insights Sale' capture the sale of data at varying levels of processing. Strategic Opening of Data' covers the open provision of data, and 'Data Bartering' covers the idea of sharing data for non-monetary benefits, while 'Data Enrichment' refers to the use of shared data within one's own organization. Najjar and Kettinger (2013) describe paths from low data monetization to high data monetization, e.g., in their case study, a retailer first establishes collaborative sharing and proceeds to competitive sharing later. Visconti, Larocca, and Marconi (2017) outline the role of platforms in the sharing of data. This is reflected in our company findings, which most often refer to multi-case platforms instead of single-case applications.

Against the backdrop of our research, we want to motivate a novel categorization of strategic approaches to data sharing. Four alternatives can be identified; closed, collaborative, competitive, and open (see Figure 3.1). All approaches can be applied to the whole range from raw data to insights. This is important to consider in the specific implementation of an approach since challenges differ with the rawness of shared data. Though, the business model might change, e.g., from 'selling' to 'wrapping' (Wixom, 2014). Closed and Open Data Exchange are decisions of the data owner and the data owner alone with an indirect return. Competitive and Collaborative Data Exchange are agreements on data transactions with third parties which include a direct return in form of monetary benefit (competitive) or in form of mutual data access (collaborative). Therefore, in the last two approaches, it is necessary to design mechanisms that lead to the desired outcome by coordinating actions by the various participating companies. This includes for instance mechanisms to solve the principal-agent problem in competitive exchange or the free-rider problem in collaborative exchange. The first is addressed with the design of price incentives in Chapter 5.



Closed	Competitive	Collaborative	Open
Data Assets are only internally accessible and reuseable.	Data Sharing with 3rd parties in exchange for money.	Data Sharing with 3rd parties in exchange for mutual data access and reuse.	Voluntary Data Sharing with 3rd parties without specified return.

Figure 3.1: The Data Sharing Framework Used in this Work

Collaborative data exchange is the shared usage of data for more or less common goals. Data is shared especially with companies to which strong links exist in the supply chain, e.g., with suppliers, vendors, and partners (E07, E13), or distributors of their own product (E12). Yet, also other companies between which no strong link exists can share data, e.g., Expert 05 reported that they currently have two use cases in operation that rely on data sharing with non-competing companies. Moreover, experts from various industries reported on data sharing projects between competing companies with the goal to drive research in their respective areas and improve the insights they can take from their models (E02, E09, E11). All of these sharing approaches have in common that they initially seek non-monetary benefits, let it be a research advantage or a more efficient supply chain. Some use cases need a lot of data and are common to an industry. The solution to those concepts is a shared platform. All interested companies bring their own data in and benefit from accessing all other's data. Thereby, a very much larger database is at disposal to solve use cases. The companies can directly profit from access to a larger database. E04 reported that such a data marketplace has been developed and is already available for a longer time, but still in a maturing state. Further examples from practice include CDQ and Catena-x.

CDQ[21] is a data sharing initiative with the goal of high-quality master data on customers and vendors. This master data is crucial but changes quickly. At the same time, the master data demand of many companies overlaps quite severely. CDQ shares that data, thus mitigating redundant processes across organizations. Thereby, all companies participating can achieve significant savings. The platform already hosts 180+ million datasets. Participating companies from our company set include Bayer, BASF, Nestle, Novartis, and Siemens.

[21] cdq.com

Catena-X[22] is an initiative from the automobile sector focusing on the establishment of a collaborative data ecosystem. It aims at enabling data exchange in the value chain based on sovereign, transparent, trustworthy, decentralized, yet standardized solutions. Exemplary use cases are the traceability of parts through the supply chain in order to track sustainability and comply with legal obligations and Corporate Social Responsibility (CSR) requirements regarding the supply chain or quality control across multiple stages of a supply chain. Among the companies working on this development are Siemens, Telekom, BASF, and Daimler.

Competitive data exchange makes data ready for sale. Companies can directly gain a monetary profit from selling data to interested parties. While collaborative exchange requires a quid-pro-quo situation in which only those companies participate that can get more out of the shared data than they have to put in, data sale enables data exchange even in more asymmetric situations. The data sold can be of higher or lower processed forms. The sale of information in highly processed forms has a long history. A dictionary is nothing else than a very big data set full of structured information. Today information is more and more sold in machine-readable forms. The frontier of research and business is the enabling of less processed and less aggregated data products. Instead of selling polished reports, the question nowadays is, how to sell raw sensor data. Several of our experts reported that their companies are also interested in opportunities to sell raw data or insights from specific analyses (E01, E03, E08, E10). Though direct monetization of data is a topic often discussed in data strategies, it is rarely put into practice (E01). However, some experts report that their companies already operate services where limited data and insights from their companies can be purchased (E08, E10). It is an issue on its own to determine the granularity and scope of such an offer.

One option in competitive data sharing is the direct monetization of data to customers, often in wrapped form around a physical core product and marketed as digital solutions. Exemplary, Caterpillar has developed a set of data-driven services and built an ecosystem around their construction equipment to improve the operations of their customers in terms of efficiency, safety, sustainability, etc., and making informed decisions. Exemplary services include tracking proximity and recommending maintenance. The services themselves are relying on data measured and collected at the customer but require knowledge from data beyond a single customer. Caterpillar has published a statement on its data governance, indicating that data is used not only to develop but also to provide services to others. Withal Caterpillar benefits from having access to a lot of data from their customers running their services enabling them to provide those services.[23]

[22] `catena-x.net/de/`

[23] `peoriamagazines.com/ibi/2017/jun/dirt-digital-technologies` and `caterpillar.com/en/legal-notices/data-governance-statement.html`

Unfortunately, we have not found any public sources indicating the direct sale of data sets or access in a rather raw format. Notwithstanding, we hypothesize that B2B negotiations and contracts on data bought and sold are kept confidential. Press releases dealing with a deeper collaboration between two or more companies usually treat the collaboration and the goals, but not the terms and forms of data exchange. Therefore, we believe, it would be a false conclusion from the non-existence of evidence in public sources that no such monetary data sharing exists. For example, Amgen and GE exchange data along the supply chain to ensure reliability and increase manufacturing efficiency in the production of pharmaceuticals. From the exchange, the companies strive for a better understanding of the interrelationship between raw material variability and process performance. Both companies expect mutual benefits from this cooperation. However, from the article, it remains unclear if this data is shared collaboratively or competitively.[24] Yet some forms of data offers for sale that are directed at many potential interested parties have been found.

Caruso[25] is an initiative of the German automotive sector, especially the independent aftermarket, and aims at the industry-wide availability of vehicle data across brands. It covers 50% of the connected car park in Europe, which are 22 million cars in total. It grants access to a wide range of car data, status-related as well as service-related. Possible use cases include maintenance, insurance, fleet management, etc. The data on the Caruso marketplace is sold. Specifically, consumers pay a subscription fee and a data fee that is – dependent on the choice of the seller – variable in time and cars or requests. The platform primarily solves the problem of in-vehicle data access for all independent service providers, i.e., anyone who wants to offer car-related services but is not the car manufacturer itself. One of the car manufacturers selling data of their connected cars through Caruso is Daimler AG.[26]

Telekom offers the Data Intelligence Hub (DIH)[27] on which open and commercial data can be shared and sold. The dataplace prioritizes the security of use and meets data sovereignty requirements. Data providers are able to see and determine access, usage period, and purposes. The marketplace can be used for multiple use cases, e.g., quality control along the supply chain[28] or planning purposes with third-party data[29]. Telekom itself as well as many other companies have open data offers as well.

Last, open data is the approach of giving the data away for free. This is the least restrictive approach and does not bear any direct benefit for the company, but it is possible to make indirect profit from open data, e.g., via third-party innovation or

[24] biopharma-reporter.com/Article/2019/03/13/GE-and-Amgen-partner-in-digital-data-exchange

[25] www.caruso-dataplace.com

[26] caruso-dataplace.com/daimler-ag-data-now-available-on-our-marketplace/

[27] dih.telekom.net

[28] dih.telekom.net/industrie-4-0-loest-qualitaetsprobleme/

[29] dih.telekom.net/mobility/

increased reputation. However open data is something done to a certain degree and the companies, who engage in this, pursue a certain agenda. Summarizing the Open Data Approaches aim at making a societal impact and stimulating the data economy in a specific area. Of our experts, only one reported that their company actively pushes open data to fuel the data market (E07).

Euler Hermes, a subsidiary of Allianz SE that offers trade credit insurance among other services, has set up an open data portal containing anonymized B2B trade data from three years. Their stated goal is to build a community around B2B trade data, to empower others to solve business challenges and embrace their own corporate social responsibility. Ultimately, they hope to inspire others to open up their data as well.[30]

Johnson&Johnson is sharing clinical trial data for research purposes in alliance with the Yale University Open Data Access (YODA) project. The advantage of participating in YODA over processing requests separately is that it ensures a uniform and objective evaluation of researchers' requests and attracts more researchers from the outset. Among the goals are in particular, to ensure that data, once collected, can continue to contribute to public health and medical science, as well as to pave the way for better healthcare data sharing practices.[31] However, Johnson&Johnson is just an example of something which has established itself as state of art in medical research. Novo Nordisk has its own platform for the dissemination of clinical trials.[32] Bayer participates in the cross-provider *clinicalstudydatarequest.com*.[33]

Regardless of the specific form in which a company chooses to share data, companies have to pay attention to certain aspects in order to reap the benefits. Data sharing increases data awareness and availability. Sharing data makes good data quality a necessity and provides an opportunity to push for action to maintain and achieve better quality (E02). Turning a profit with data has a positive consequence for the attitude towards data and supports the cultural change in a company (E05). Data sharing and holistic business problem solving go hand in hand in a collaborative approach across organizational boundaries (E05), ultimately allowing end-to-end connectivity and solutions (E06). This creates new opportunities to tackle business challenges. Participation in data sharing initiatives may also be a strategic imperative to remain relevant to customers who require their suppliers to provide certain data (E06). A common psychological barrier to sharing data is the perception of the own data as incomplete and insufficient even for the own use cases (E01). However, exchanging data with others enlarges the foundation for previously unavailable data. Thus, it brings the opportunity to improve statistical models and generate new insights (E07, E11). There is also a fear of being disrupted by third parties who are better at handling data in the first

[30] allianz-trade.com/en_global/news-insights/news/euler-hermes-launches-its-open-data-portal.html

[31] jnj.com/innovation/yale-open-data-access-project

[32] novonordisk-trials.com/en/how-access-clinical-trial-datasets/

[33] clinicaltrials.bayer.com/transparency-policy

place (E03, E05, E09, E10). This fear especially plays a role in exchanging data with competitors (E05, E09, E10). Additionally, there is legal uncertainty regarding privacy laws (E06, E10) and regarding competition law (E07). Moreover, reputation risk is involved if the personal data of customers is shared even if it is completely legal within privacy regulation (E04). Therefore it is of uttermost importance that you know what you share and you are certain about the purposes of sharing (E10). Expert 06 lines out that a company needs to define its strategy towards data sharing and ensure strategic alignment of single initiatives. This can prevent a department from sharing data that is beneficial to their domain, but detrimental to the company in a larger strategic perspective (E06).

3.9 Limitations and Conclusion

Our research has highlighted core components of data strategies for practitioners. These are major action fields and our research results include guidance on trends and best practices. The core components and practices are of even more interest to researchers as they constitute future research opportunities. Many of the described measures have been understudied and only insufficiently understood in their effect on organizations. While this study was purely descriptive based on codified practice experience, questions of the quantifiable effects of suggested measures under different circumstances emerge.

At the outset of our company study, we had hoped to find a well-documented data strategy as part of the companies' overarching disclosed strategy or information published by Investor Relations. This was not the case. Instead, our findings refer to press releases, interviews, and news articles. While it was clear from the beginning that it would not have been possible to restrict us to the use of peer-reviewed materials, this adds another layer of uncertainty. Reporters writing news articles can err in their interpretation. Single statements and businesses might be individual cases, which are not representative of the whole strategy. Last but not least, the published information is necessarily filtered by companies. The actual reality can differ considerably from the proclaimed one. These issues persist in this work, while at the same time the expert study added plausibility to our results. However, interviewing business professionals who are experts in the field will likewise not guarantee that their opinions and thoughts are correct, objective, or exhaustive to a topic.

To summarize, the learning for future implementation of data strategies is, to first understand data as an asset class in its own right. It needs to be managed and monitored. This requires creating an oversight, e.g., in a data catalog of appropriate scope. Further, it is critical that a person and/or division in the organization takes responsibility for data governance. For them to succeed, there must be a congruence of responsibilities, tasks, and skills. It is also beneficial to distinguish between traditional IT tasks and data management. Data architectures have the task of storing and making available data efficiently and flexibly accessible and usable at the same time. On the one hand, this must be done for the entire company, but on the other hand, it must also be done for a lot of data and without overloading central teams. The crucial task is to ensure uniformity but allow diversity. This creates a constant tension between centralization and decentralization. New concepts such as the DLH and the Data Mesh attempt to resolve these conflicts. Next, the skills of the workforce continue to be of critical importance. These are technical skills in handling data on the one hand and domain knowledge about the application on the other. In a tight labor market, experts in both areas are hard to find, so it is important to be clear about job descriptions and offer appropriate training programs. Partnerships and acquisitions can also be compelling in order to bring missing expertise into the company.

Finally, the value of data is highly dependent on access to further data, so data sharing is becoming increasingly important. However, there are also many risks associated with it, which can be minimized if you know well what you are sharing, with whom and why. Keeping the data closed is a viable approach for some data, but not for all. If data exchange is strategically desirable, there exist three different strategic approaches, which were described in this work including examples. Knowing the data assets and gaps in the business, understanding their value, and identifying the right partners will help in choosing the right strategy for the business.

Chapter 4

Challenges of Data Pricing

> Where there is no free market, there is no pricing mechanism; without a pricing mechanism, there is no economic calculation.
>
> Ludwig von Mises

Data has tremendous value that can be levied via adequate management as pointed out in the last Chapter. However, even today, too much of this value is lost because data is not made accessible and reusable outside of its original context. An important key to taking full advantage of more and better data is breaking down silos and sharing data. The competitive exchange of data is one relevant approach to do so. In this view, data should become an asset that companies can sell and buy, accessible to anyone with an idea of how to use it. The main advantage over a collaborative approach that exchanges data on a quid pro quo base is that sharing is possible even if no adequate data for mutual exchange exist.

In addition, it provides the possibility to finely design payments. This is at the same time a great barrier because one must first address the question of how to price data before a competitive data market can thrive. In this Chapter, we have structurally examined the existing body of research on this issue and highlighted the challenges of data pricing from a scientific perspective. In this way, we gain a comprehensive overview of the state of the art and prepare the ground for future research.

The contents of this chapter are adopted or taken from the paper: Wolfgang Badewitz, Christoph Hengesbach, and Christof Weinhardt (2022). "Challenges of pricing data assets: a literature review". In: *2022 IEEE 24th Conference on Business Informatics (CBI)*. vol. 01, pp. 80–89. DOI: 10.1109/CBI54897.2022.00016.
See Appendix A for further details.

4.1 Introduction

The value of data is difficult to determine. Its value is mainly and for most use cases driven by what information and insights to get on very specific issues. This in turn plays into many characteristics of data that distinguish it from conventional digital goods. Multiple quality dimensions, such as interpretability, accuracy, and timeliness, are of major importance, and value scales with them over a greater range. Moreover, data assets are highly characterized by complementing and substituting each other. Whatever can be learned from one data set, might most probably also be learned from some other data set. The data sets could replace each other or their combination would lead to an even better understanding. Therefore, it is worth investigating other data sources and looking for options to combine the data one has with outside data. Of course, this also applies the other way around. One's own data assets might just be what someone else is missing and cannot economically collect themselves. At the same time, duplication of data is almost free, which opens up the possibility of making profits by buying and selling data.

The arguments for trading data are evident, but it is not yet clear how to facilitate that trade. One pressing problem is to determine a fair price in an effective way. To date, there is great uncertainty about how to recognize the value of a data asset and how to price it. Yet it is apparent that data pricing will become more and more relevant in the future. Both data consumers and data providers wish for a comprehensible and fair way to assess prices for data assets. Also, market owners are in need of clear processes to facilitate smooth and easy transactions and guarantee the success of their platforms. A sound pricing mechanism would solve those issues.

Many scientists have already been engaged with this topic considering many different aspects of data pricing. At this point, several research findings have accumulated that require systematization. Therefore, we conducted a structured literature review of scientific works which feature mathematically expressed models of how to calculate prices for data. We exclude purely conceptual work and frameworks on this topic. Instead, we concentrate on papers that describe formalized methods to set, update or infer prices for data that could be implemented in a real market, at least in principle. Our comprehensive review provides an analysis and structure to data pricing literature which offers a) an overview of the existing approaches and models, b) a summary of challenges, which should or could be reflected in data pricing mechanisms, and c) impulses for future research. This overview of the existing knowledge will constitute a great assistance to future researchers in developing new and better mechanisms. Concisely we formulated the following Research Questions onto the research corpus:

RQ 4.1 *What objectives are pursued by current research?*

RQ 4.2 *What are the underlying data and market structures?*

RQ 4.3 *What are the common approaches to data pricing in commercial environments?*

RQ 4.4 *What challenges do data pricing approaches face?*

4.2 Method

In order to answer our research questions, we conducted a structured literature review according to established scientific conventions (Webster and Watson, 2002; Brocke et al., 2009). The initial literature base was retrieved from the interdisciplinary research databases *ScienceDirect / Scopus* and *Web of Science (WoS)*. We queried those two databases with different search terms in title, abstract, or keywords. A comprehensive overview of search terms and results is given in Table 4.1. By 2021, 16th of November and after removing duplicates this yields a total of 357 unique papers. We analyzed title, abstract, and keywords and excluded publications, which did not focus on data pricing in commercial settings or lack relevance in general. Criteria for exclusion were (i) papers not focusing on pricing for data, e.g. frameworks for data marketplaces, business models to turn data assets into profit, pricing of related goods, such as IoT-Services (ii) redundancy, multiple versions, early-stage drafts (iii) non-scientific publications (iv) another language than English. This resulted in a set of 55 papers, of which 46 papers were relevant after reading the whole article. Subsequently, we performed two iterations of a forward and backward search and identified 24 additional relevant papers. In the end, a total of 70 publications were part of the review. According to our research questions, we analyzed our literature in a concept-centric approach. The derived concepts for analysis were the objectives of data pricing, the underlying structure of the market on the economic side and of the data as the transactional object in question, the applied approach to pricing, and the challenges addressed by the research. An overview of results is given in Table 4.2.

Table 4.1: Search Results for Data Pricing Literature 16.11.2021

	# Paper	
Search String	Scopus	WoS
“data marketplace” AND pric*	35	10
“data monetization” OR “data monetization”	62	20
“data pric*”	262	98
	332	95
	⇒ 357 unique papers	

Table 4.2: Concept Matrix on Data Pricing Literature

Source	Objective	Market Structure	Data Structure	Approach	Challenges*
Agarwal, Dahleh, and Sarkar (2019)	Profit	n:1:m	Not specified	Query-based; Auction	Q, AF, RSh, C, IR/IC/BB
An et al. (2017)	Welfare	1:n	Not specified	Auction	FP, C, IR/IC/BB
Bataineh et al. (2020)	Welfare	n:1:m	Not specified	Programming Model	Q, PP
Bergemann and Bonatti (2015)	Profit	n:m	Cookies	Programming Model	Q
H. Cai, Zhu, et al. (2019)	Welfare	n:1:m	Not specified	Double Auction	Q, CoI, C, IR/IC/BB
H. Cai, F. Ye, et al. (2022)	Profit	n:1:m	Relational	Auctions; Query-based	Q, PP, FG, C, IR/IC/BB
X. Cao, Y. Chen, and K. J. R. Liu (2017)	Welfare	n:k:m	Not specified	Auction	PP, CoI, IR/IC/BB
Chawla et al. (2019)	Profit	1:n	Relational	Programming Model	AF, FG, C
Y.-J. Chen and Seshadri (2007)	Profit	1:n	Not specified	Programming Model	Q, PD, IR/IC/BB
L. Chen, Koutris, and Kumar (2019)	Profit	n:1:m	Relational	NYOP	Q, AF, FG, PD, C

Table 4.2: Concept Matrix on Data Pricing Literature

Source	Objective	Market Structure	Data Structure	Approach	Challenges*
Chuang et al. (2020)	Welfare	n:m	Not specified	Programming Model	FP, C
Dandekar, Fawaz, and Ioannidis (2014)	Profit	n:1	Relational	Auction	Q, PP, IR/IC/BB
Deep and Koutris (2017a)	Consistency	n:m	Relational	Query-based	AF, FG, C
Deep and Koutris (2017b)	Consistency	1:n	Relational	Query-based	AF, C
Fleischer and Lyu (2012)	Profit	n:1	Not specified	Auction	Q, PP, IR/IC/BB
G. Gao et al. (2020)	Profit	n:1:m	Not specified	Auction	PP, IR/IC/BB
Ghosh and A. Roth (2011)	Profit	n:1:m	Relational	Auction; NYOP	Q, PP, IR/IC/BB
Gkatzelis, Aperjis, and Huberman (2015)	Profit	n:1:m	Not specified	Auction	PP, IR/IC/BB
Goncalves, Pinson, and Bessa (2021)	Welfare	n:m	Relational	Auction	Q, RSh, IR/IC/BB
P. Gupta et al. (2020)	Profit	n:m	Not specified	Auction	Q, RSh, T
Huang, Milani, and Chiang (2020)	Consistency	1:1	Relational	Query-based	Q, AF, PP, C
Jia et al. (2019)	Consistency	n:1	Not specified	Query-based	RSh, C
Jiao et al. (2018)	Profit	n:1:1:m	Not specified	Auction	Q, C, IR/IC/BB
Jung et al. (2019)	Welfare	n:1:m	Not specified	Query-based	PP, IR/IC/BB
Koutris et al. (2015)	Consistency	n:m	Relational	Query-based	AF, C

Table 4.2: Concept Matrix on Data Pricing Literature

Source	Objective	Market Structure	Data Structure	Approach	Challenges*
Koutsopoulos, Gionis, and Halkidi (2015)	Profit	n:1	Relational	Auction	PP, IR/IC/BB
C. Li and Miklau (2012)	Consistency	1:n	Relational	Query-based	AF, NI, FG
C. Li, D. Y. Li, et al. (2014)	Consistency	n:1:m	Relational	Query-based	Q, AF, PP, C, IR/IC/BB
Xijun Li et al. (2017)	Consistency	1:n	Relational	Query-based	Q, AF
Q. Li et al. (2021)	Profit	1:n	Not specified	Programming Model	Q, FG, T, IR/IC/BB
Lin and Kifer (2014)	Consistency	1:n	Not specified	Query-based	AF, NI, FG, C
K. Liu et al. (2019)	Welfare	n:1:m	Not specified	Programming Model	T, IR/IC/BB
Jinfei Liu et al. (2021)	Profit	n:1:m	Not specified	Programming Model	Q, AF, PP, FG, C
Mao, Z. Zheng, and F. Wu (2019)	Profit	1:n	IOT Sensor Data	Programming Model	Q, PD, C, IR/IC/BB
Sameer Mehta et al. (2019)	Profit	1:n	Relational	Auction	Q, C, IR/IC/BB
Niu, Z. Zheng, S. Tang, et al. (2019)	Profit	n:1:m	Relational	Query-based	Q, AF, PP, FG
Niu, Z. Zheng, F. Wu, et al. (2020)	Profit	n:1:m	Not specified	Query-based	Q, PP, C
Niyato, D. T. Hoang, et al. (2016)	Profit	n:m	Not specified	Auction	FG, PD, FP, IR/IC/BB

Table 4.2: Concept Matrix on Data Pricing Literature

Source	Objective	Market Structure	Data Structure	Approach	Challenges*
Niyato, Alsheikh, et al. (2016)	Profit	n:1:m	Not specified	Programming Model	Q
Ray, Menon, and Mookerjee (2020)	Profit	1:1	Not specified	Programming Model	Q, T, IR/IC/BB
Sakr (2018)	Consistency	n:m	Spatial Data	Query-based	AF, NI, FG, T
Yuncheng Shen et al. (2016)	Consistency	n:m	Not specified	Query-based	Q, PP
Yuncheng Shen, Guo, Yan Shen, F. Wu, et al. (2019)	Consistency	1:n	Relational	Query-based	AF, PP, C
B. Shen, Yulong Shen, and Ji (2019)	Profit	1:1:1	Not specified	Programming Model	Q, C
Yuncheng Shen, Guo, Yan Shen, Duan, et al. (2019)	Consistency	n:m	Relational	Query-based	Q, AF
Stahl and Vossen (2017)	Welfare	n:m	Relational	NYOP	Q, PD, T
R. Tang, H. Wu, Bao, et al. (2013)	Consistency	1:n	Relational	Query-based	AF, FG, C
R. Tang, Shao, et al. (2013)	Consistency	n:1:m	Relational	NYOP	Q
R. Tang, H. Wu, X. He, et al. (2015)	Consistency	1:n	Relational	Query-based	C
R. Tang, Amarilli, et al. (2016)	Consistency	1:n	XML	NYOP	Q, AF, C

Table 4.2: Concept Matrix on Data Pricing Literature

Source	Objective	Market Structure	Data Structure	Approach	Challenges*
Tian et al. (2022)	Profit	n:1:m	Not specified	Programming Model	Q, AF, PD, RSh
X. Wang, Wei, Y. Liu, et al. (2018)	Welfare	n:1:m	Relational	Query-based	Q, AF, C
X. Wang, Wei, S. Gao, et al. (2019)	Welfare	1:n	Relational	Auction	C, IR/IC/BB
Z. Xiao, D. He, and Du (2021)	Welfare	n:1:1:m	Not specified	Programming Model	IR/IC/BB
L. Xiong and H. Zheng (2019)	Profit	2:1	Not specified	Programming Model	Q
W. Xiong and L. Xiong (2021)	Profit	1:n	Not specified	Auction	FP, T, IR/IC/BB
C. Xu et al. (2020)	Profit	n:1:m	Not specified	Programming Model	Q, PP
J. Yang and Xing (2019)	Profit	n:1:m	Relational	Programming Model	PP, PD, T, IR/IC/BB
J. Yang, C. Zhao, and Xing (2019)	Profit	n:1:m	Relational	Programming Model	Q, PD
Yassine, Nazari Shirehjini, and Shirmohammadi (2015)	Welfare	n:1:m	Classification	NYOP	Q, PP, IR/IC/BB

Table 4.2: Concept Matrix on Data Pricing Literature

Source	Objective	Market Structure	Data Structure	Approach	Challenges*
Y. Ye et al. (2021)	Consistency	1:n	Not specified	Query-based	FG
You et al. (2021)	Consistency	n:1	Classification	NYOP	Q, RSh
H. Yu and M. Zhang (2017)	Profit	1:n	Not specified	Programming Model	Q, PD
J. Zhang et al. (2019)	Consistency	n:1:m	Not specified	Programming Model	IR/IC/BB
Y. Zhang et al. (2020)	Profit	1:n	Not specified	Programming Model	PD, IR/IC/BB
M. Zhang, Beltran, and Jiamou Liu (2020)	Consistency	n:1	Graph Data	Auction	Q, PP, IR/IC/BB
Z. Zhang, Song, and Yuan Shen (2021)	Welfare	n:m	Not specified	Query-based	Q, PP
Y. Zhao et al. (2020)	Profit	1:n	Not specified	Auction	IR/IC/BB
Z. Zheng et al. (2020)	Profit	n:1:m	Spatial Data	Auction	Q, AF, FG, PD, RSh
X. Zheng (2020)	Consistency	n:m	Not specified	Query-based	Q, PP

* Legend: Quality Issues (Q); Arbitrage Freeness (AF); Provide Privacy (PP); Fine Granularity (FG); Negative Information (NI); Fraud Prevention (FP); Revenue Sharing (RSh); Transparency (T); Complexity (C); Price Differentiation (PD); Individual Rationality, Incentive Compatibility, Budget Balance (IR/IC/BB)

4.3 Objectives

We differentiate between two objectives, which are sought to be reached in the current literature, profit maximization, and social welfare. A third line of work does not focus on a specific outcome but tries to reach consistency with posed challenges.

First, profit maximization deals with the optimization of the profit as revenue minus cost or just the revenue (while assuming cost as fixed and zero) of some participants in the market. The actor in question varies across papers. There is a connection to the data structure. In a one-to-many market, optimization is usually performed from the perspective of the data seller, while in a many-to-one market it is performed by the data buyer, and in a many-to-one-to-many market by the data broker. There exist more special market structures on which no general statement can be made. Closely coupled with profit maximization is the challenge of price differentiation.

Second, social welfare research is not concerned about one market participant in particular but tries to improve the situation for all participants at once. Not all the papers explicitly model the welfare of those agents but clearly state the intent to generate value for all participants in the data market.

Third, consistency is about the creation of a pricing scheme given one or more constraints that are deemed necessary for pricing data. The difference between the first two objectives is that the solution space is bigger. There are more pricing formulas that solve certain challenges such as arbitrage freeness than pricing formulas that solve the challenges and guarantee the profit to be optimal or at least inside a competitive bound.

4.4 Underlying Market Structure

The market settings in the papers were differentiated between roles and the number of actors in each role. We differentiate between one-to-many (1:n), many-to-many (n:m), and many-to-one (n:1) settings in the absence of a data broker and mostly many-to-one-to-many (n:1:m) in the presence of a data broker. Other structures exist and are denoted accordingly. The most prevalent setting is many-to-one-to-many, followed by a one-to-many structure. Almost all research regarding social welfare is done in a n:m or n:1:m setting. Nonetheless, by far most papers incorporating a broker aim at profit maximization. The broker role can be introduced to divide a many-to-many market into a many-to-one and a one-to-many market. This provides the opportunity to have different prices for sale and purchase. For example, compensate data owners for privacy loss via an auction approach and propose query-based prices to consumers (Ghosh and A. Roth, 2011). Another role of the broker is the one of a processor of data adding value by different operations, such as feature engineering, dimension reduction, or predictive analytics (Agarwal, Dahleh, and Sarkar, 2019; Z. Xiao, D. He, and

Du, 2021), or to provide privacy protection by adding noise (C. Li, D. Y. Li, et al., 2014; Niu, Z. Zheng, F. Wu, et al., 2020).

The 1:n setting can be interpreted as an investigation in direct data monetization, where the agent in focus wants to sell his data to customers. Profit maximization is a common objective in this setting as well but closely followed by consistency. For consistency research, a 1:n setting offers the advantage of a relatively simple market environment. Only two papers research social welfare in a one-to-many setting.

Special structures include for example the three-tier big data market model in (Z. Xiao, D. He, and Du, 2021), where multiple service users are served by a service provider that receives data from a data vendor that acts as an aggregator for multiple original data sources, and the 1:1 model in (Ray, Menon, and Mookerjee, 2020), which focuses especially on data demonstration and reduces all unnecessary complexity.

4.5 Dependence on Data Structure

Data is the transaction object in focus. However, data is various in its structure. Panel data varies from geospatial data, from pictures, from natural language, and so on. This makes it hard to come up with a general mechanism for data. Our initial assumption was that pricing should depend on the type of data. Unstructured data such as pictures, videos, or free text files pose other difficulties than structured data. Also, different structures such as linked data or relational databases pose different opportunities to assess the value of the data. Curiously, literature focuses on relational data or does not specify the kind of data traded at all. Only very few papers considered other types of data and exploit their properties such as XML-data, e.g., (R. Tang, Amarilli, et al., 2016). 'Not specified' hereby does not necessarily mean that the proposed pricing mechanism is universally deployable, e.g., (Yuncheng Shen et al., 2016) requires that Shannon Entropy is calculable.

4.6 Pricing Approaches

In this work, we differentiate between four pricing approaches: query-based, programming model, name your own price, and auctions. This differentiation is motivated by the basic principle according to which the price is calculated. Another option would be to divide pricing into raw data pricing and model-based pricing (L. Chen, Koutris, and Kumar, 2019). This division is based on the data, resp. information product in question and not the basic idea of how to determine the price. Basically, it is a division of how flexible, fine-grained, and accurate the demand can be posted on the data market. Model-based pricing is a more complicated query on a database, i.e. a query which returns a model but is not in itself different from other queries, which return variables.

4.6.1 Query-based Pricing

In query-based pricing, the data consumers specify, what they want in the form of a query Q and get a price. Especially, the price is not dependent on any form of valuation of the customer, transmitted or assumed, but calculated on properties of the query, the database, or both. Subsequently, one can further differentiate between instance-independent and answer-dependent, and data-dependent pricing (Deep and Koutris, 2017a). In instance-independent pricing, prices are dependent on the query, but not on the database. In answer-dependent pricing, the prices depend on Q and the answer $Q(D)$, in data-dependent prices depend on both Q and D.

Query-based prices are independent of the data buyer and his valuation of the data. As everyone with the same query receives the same price and data product, there is no opportunity for price discrimination. On the other hand, query-based pricing has no possibility for strategic behavior because prices do not reflect revealed valuations. A challenge is to determine the value of data since the prices have to be set in some kind of way. The prices are more driven by the valuations the data seller thinks his data has. As a matter of fact, many query-based pricing models work with some predetermined prices that are used to determine prices for other queries, e.g., (Yuncheng Shen et al., 2016; Koutris et al., 2015; R. Tang, H. Wu, X. He, et al., 2015). Another common option is to base query prices on certain aspects of the data, especially quality, e.g., (C. Li, D. Y. Li, et al., 2014; Yuncheng Shen, Guo, Yan Shen, Duan, et al., 2019), or privacy, e.g., (X. Zheng, 2020; Z. Zhang, Song, and Yuan Shen, 2021). Almost all researchers in this stream focus on consistency. There are some exceptions. (X. Wang, Wei, Y. Liu, et al., 2018) minimize the cost of processing queries by taking resource consumption into account. (Niu, Z. Zheng, F. Wu, et al., 2020) maximize profits by following a dynamic, query-based approach, and learn the market price via iterated try-and-error.

4.6.2 Programming Model

In a programming model, assumptions are made about the market participants and an objective is maximized mathematically. This can be social welfare, e.g., (Z. Xiao, D. He, and Du, 2021), or profit, e.g., (Chawla et al., 2019; Mao, Z. Zheng, and F. Wu, 2019; Y. Zhang et al., 2020). This also includes also some game-theoretic approaches, such as Stackelberg Games in which the market participants maximize their respective profit after each other, e.g., (B. Shen, Yulong Shen, and Ji, 2019; Niyato, Alsheikh, et al., 2016). Some authors deployed genetic algorithms to solve the formulated optimization problems (H. Yu and M. Zhang, 2017; J. Yang and Xing, 2019).

4.6.3 Name Your Own Price

In "Name Your Own Price" concepts, customers decide for themselves what they want to pay, and the transaction occurs if they are above a certain threshold set by the seller. They have in common that the prices customers receive depend on their own transmitted willingness-to-pay (WTP), but not on those of others. We include also approaches in which the data product reflects on the WTP and the quality of the product is somehow degraded due to the WTP. For example, Stahl and Vossen (2017) is regarding social welfare and let consumers set a price and maximize the quality over multiple dimensions, thereby enabling also consumers with low financial possibilities to participate in the market. L. Chen, Koutris, and Kumar (2019) introduces a mechanism in which buyers have three possibilities: name their price and obtain the best quality of data; name their quality and obtain the cheapest price; name a combination of error and price. Still, they focus on the profit of the data sellers (L. Chen, Koutris, and Kumar, 2019). One challenge in this approach is dealing with the strategic behavior of consumers as the possibility exists to claim a lower WTP than one actually has.

4.6.4 Auctions

In auctions, data consumers and/or data providers are calling their respective valuations on the data. The difference to NYOP approaches is that data consumers and/or providers influence not only their own but also the prices of everyone else. Auctions are well-known to maximize profits for many different goods, where strong information asymmetries about valuations pose an obstacle to finding a fixed price. In data markets, participants are faced with severe information asymmetry, since the seller cannot access the utility of its data for the data customer. This value is dependent on his use case, data maturity, the prevalence of data assets etc. Auctions are an effective way to realize a good allocation in such situation (W. Xiong and L. Xiong, 2021). With limited supply, a second-price mechanism, i.e. the one calling the highest valuation paying the price of the second-highest bid, is an efficient way to retrieve the most profit out of a sale (Myerson, 1981). Data on the other hand has an unlimited supply. Everyone can be served with data, thus it would be optimal to serve everyone for the price he is willing to pay. This would create a strong incentive to bid zero. In theory that can be solved by assuming a distribution over valuation (Myerson, 1981). In practice, the optimal price has to be calculated by the bids. This is a fundamental issue in auctions with unlimited supply. However one can design competitive auctions which guarantee profits to be inside a competitive ratio to the optimal, but not truthful mechanisms (Goldberg et al., 2006). Such competitive auctions in data pricing were designed by, e.g., (Y. Zhao et al., 2020).

4.7 Challenges

Our fourth research question aims at the challenges of mechanism design in data pricing. In this section we comprehensively describe the most important challenges specific to data pricing, leaving out general challenges such as price differentiation and the classical triad of Individual Rationality, Incentive Compatibility, and Budget Balance. Some of the challenges come with more or less mathematically concise definitions, which might pose strict requirements on pricing mechanisms, e.g., arbitrage freeness or privacy guarantees, others are of transcendental nature as is especially the case for transparency.

4.7.1 Infer Value from Quality

Data is not only easy to duplicate but also easy to falsify. This makes it easy to influence the quality of data products. Noise injection, also known as perturbation, can be used to control exactly how much the quality is reduced. Deliberately altering the quality can bring many benefits, such as applying price differentiation, e.g., (Mao, Z. Zheng, and F. Wu, 2019; L. Chen, Koutris, and Kumar, 2019; Z. Zheng et al., 2020), dealing with data privacy, e.g., (C. Li, D. Y. Li, et al., 2014; Jinfei Liu et al., 2021), enforcing Incentive Compatibility, e.g., (Agarwal, Dahleh, and Sarkar, 2019; Sameer Mehta et al., 2019). Quality and value are often used interchangeably, but while value has a clear meaning and interpretation as a measure of the amount of money one is willing to spend for a product quality is less concisely understood. One main problem in determining a price for data is indeed the uncertainty of its quality for both seller and buyer. This problem is very hard for data because quality is a very fuzzy term for data stemming from multiple dimensions. For example, (Naumann, 2002) points out, that data has many dimensions which can be referred to as quality. This includes content-related (accuracy, completeness, ...); technical (availability, timeliness, ...); intellectual (believability, reputation, ...); and instantiation-related criteria (amount of Data, verifiability, ...). Some papers focus on one specific relevant quality parameter,e.g. timeliness (Y. Zhao et al., 2020), e.g. accuracy (L. Chen, Koutris, and Kumar, 2019), e.g. completeness (R. Tang, Amarilli, et al., 2016), e.g. Volume (Niyato, Alsheikh, et al., 2016). The two most important ways to measure quality are information entropy, e.g., (Deep and Koutris, 2017b; Z. Zhang, Song, and Yuan Shen, 2021; Yuncheng Shen, Guo, Yan Shen, Duan, et al., 2019; Xijun Li et al., 2017), and variance, e.g., (C. Li, D. Y. Li, et al., 2014; Agarwal, Dahleh, and Sarkar, 2019; X. Wang, Wei, Y. Liu, et al., 2018; L. Chen, Koutris, and Kumar, 2019). A reasonable amount of papers take into account that data has multiple sources of quality which all are important. Moreover, the quality of data is not objectively determined. Independent economic entities can vary in their assessment of the quality of the same data. This is used by

(Stahl and Vossen, 2017), who combined seven utility dimensions in a knapsack pricing, which maximizes the quality of a query answer given a budget. (J. Yang, C. Zhao, and Xing, 2019) linearly weight quality attributes, such as accuracy, completeness, and redundancy into one quality score. (H. Yu and M. Zhang, 2017) optimize over multiple quality dimensions with individually different marginal WTP, saturation, and reservation qualities. Thereby they also account for the fact, that the quality dimension does not necessarily scale linearly. (Yuncheng Shen et al., 2016) additionally includes credibility by a data reference index, which is inspired by the H-index and price data according to this and other relevant quality criteria, thereby involving an intellectual quality criterion, which is typically difficult to influence (Stahl and Vossen, 2017).

Moreover, the quality of a data asset has to be translated into a utility for the data consumer. A simple approach is to linearly translate quality into value $U_i(q_j) = q * \theta$ (Niu, Z. Zheng, F. Wu, et al., 2020; Y.-J. Chen and Seshadri, 2007). θ may be a universal scaling parameter, but may also be individual and reflect on heterogeneous WTP between customers (Y.-J. Chen and Seshadri, 2007). With multiple regarded quality criteria, this is also possible as weighted sum $U(q) = \sum_{(i)} q_i \cdot \theta_i$. However, a more elaborate utility function on the quality parameter can be used to reflect non-linear valuation. In this case $U(q)$ should be non-negative, concave, and increasing in q (J. Yang, C. Zhao, and Xing, 2019). Moreover, quality might be dependent on how well the data matches desired properties. H. Cai, Zhu, et al. (2019) first compares the quality attributes with the desired attributes before linear weighting. Sameer Mehta et al. (2019) incorporates the distance to ideal records, which is useful if certain entries are sought in a database.

4.7.2 Guarantee Arbitrage Freeness

One of the most important properties of data and information is, that you can combine it to generate even more information and that you can use data from several sources to obtain knowledge about something else. This is a big problem for the pricing of data and information products as it opens the way for arbitrage. Arbitrage is generally understood as the exploitation of price differences at the same point in time, e.g. buying a stock at a lower price in London and selling it at a higher price in New York. Arbitrage in data products is more complicated because it does not reside in the simultaneous purchase and sale but in the tuning of orders and subsequent processing of the purchased data, e.g. by concatenating searches, exploiting strong correlations between variables, or repeating low-quality queries to get a high-quality answer on average. For commodities characterized by physical scarcity, arbitrage regulates itself through increased demand at the less expensive location and increased supply at the more expensive location. No such self-regulation exists for digital goods. Thus a mechanism to prevent arbitrage must be explicitly included in the pricing for data.

Fundamentally, arbitrage freeness states that the price of a single query should always be less or equal to the total price of multiple queries from which the answer to the single query can be deduced, i.e. If some operation f exists such that $Q(D) = f(Q'(D), Q''(D), ...)$, then $p(Q) \leq p(Q') + p(Q'') + ...$, where D denotes a data set; Q, Q', and Q'' denote different Queries, and p denotes the price. C. Li and Miklau (2012) represent the view, that it is only arbitrage if the answer to Q is implied by $Q', Q'', ...$ on every database instance. If arbitrage is made by chance (Lin and Kifer, 2014) talk off serendipitous arbitrage.

One can distinguish between several forms of arbitrage, i.e. information and combination arbitrage (Chawla et al., 2019), or separate account, post-processing, serendipitous, and almost certain arbitrage (Lin and Kifer, 2014). Most importantly one should differentiate between three forms of arbitrage, which are the following:

Combination Arbitrage This covers fairly simple situations in which query answers are concatenated, e.g., to query temperature and pressure separately instead of together.

Prediction Arbitrage This covers situations in which a query answer can be deduced from a different query, e.g. temperature in Celsius from the temperature in Fahrenheit, but also acceleration from a time-stamped series of speed, or the income from education data. The first of course can be calculated exactly, while the latter is predicted with quite high uncertainty, which leads to the last form of arbitrage.

Quality Arbitrage This covers situations in which a higher quality to a query answer by making multiple low-quality queries, e.g. reducing the variance of a query answer to v/n by requesting n-times v variance.

In general, one can state that less informative queries should be less expensive, and uninformative queries, such as the zero query and the infinite noise query should be free (C. Li, D. Y. Li, et al., 2014). Also, the concept of subadditiveness is very important. The concise mathematical properties of pricing functions depend on the underlying model. C. Li, D. Y. Li, et al. (2014) describes properties for variance perturbation. L. Chen, Koutris, and Kumar (2019) describes properties for the Gaussian mechanism specifically. However, depending on the basic model of how quality is translated into value, other properties might be necessary. Moreover, the computation of arbitrage-free and optimal price functions has shown to be computationally hard, so an efficiently computable approximation proves to be necessary for practice (L. Chen, Koutris, and Kumar, 2019).

4.7.3 Avoid Negative Information

Prices contain information. With traditional products, this does not cause problems. Deducing that a pair of sneakers is rare because it is traded at high prices or that a movie is unpopular because it is cheap to buy online, does not affect the trade with those goods. However, in dealing with information goods, this becomes a problem. If the price of information reveals that information, trade is impossible. For example, if the price of credit ratings depends on the number of registered overdue reminders and unpaid invoices, a high price can be used to infer a poor rating, and it becomes unnecessary to buy the credit rating itself. If different data labels have different worth, the price must not reflect that difference or the labels are given away for free (C. Li and Miklau, 2012; Sakr, 2018; Lin and Kifer, 2014). Lin and Kifer (2014) countered that issue in query-based pricing with the concept of a delayed pricing scheme. They propose to show only an instance-independent price in advance and later charge a data-dependent price. This way no information can be inferred beforehand.

4.7.4 Provide Privacy Protection

A rather big branch of data pricing literature is concerned with data privacy aiming to pay for privacy so that there are incentives to open personal data up by consent, complying with legal requirements, and minimizing risks of privacy losses for owners of personal data. The state-of-the-art approach to measure privacy loss quantitatively is differential privacy, which was introduced by Dwork, McSherry, et al. (2006) in 2006.

Definition 4.1 (Differential Privacy as in (Dwork, 2008)). *A randomized algorithm M is ϵ-differentially private if, for any two databases D_1 and D_2 that differ in at most one entry, and for all $S \subseteq Range(M)$:*

$$\mathbf{P}[M(D_1) \in S] \leq \exp(\epsilon) \cdot \mathbf{P}[M(D_2) \in S]$$

In essence, differential privacy limits how much a data owner's personal data affects the outcome of a query. Thereby, ϵ is a measure of privacy loss. If ϵ equals zero, the probability of receiving a certain answer to a query is the same whether one data owner's personal data is included or not. The higher ϵ, the more the probabilities of results diverge between the two data sets, eventually allowing to draw conclusions about that data owner.

Example 1: Alice is applying for a loan and the bank makes a query to some database in order to find out whether she is worthy of credit. The answer to that query is $M(D)$ and Alice will get the loan if $M(D)$ is in S, e.g. $M(D)$ estimates her credit rating between 0 and 1, and she will only get the loan if $M(D)$ is higher than 0.9, i.e. $S = (0.9, 1]$. $M(D)$ is randomized, which means that the answer to the query is not determined by D. There is some possibility that a repeated query on the same data will produce a different answer. The Probability that she gets the loan is $P[M(D) \in S]$. Now let there be two versions of this database. D_1 contains Alice's data, while D_2 does not. Differential Privacy now requires that $\mathbf{P}[M(D_1) \in S] \leq \exp(\epsilon) \cdot \mathbf{P}[M(D_2) \in S]$ as well as $\mathbf{P}[M(D_2) \in S] \leq \exp(\epsilon) \cdot \mathbf{P}[M(D_1) \in S]$ since it holds true for every two databases D and D' which differ in one entry and does not specify which one contains the other. With $\epsilon = 0$, the probability that Alice gets the loan is identical whether she is in the database or not. With $\epsilon > 0$ the difference in probabilities is bounded. The probability that she gets the loan if she is not in the data set ($\mathbf{P}[M(D_2) \in S]$) is neither much higher nor much lower than if she would be in the data set.

Example 2: Imagine a list of grades. Alice wants to know the grade of Bob. So she poses a question like "What is the mean grade of persons named Bob?" knowing that there is only one Bob in the data set. She gets a randomized grade. With epsilon = 0, the probability that that grade is 1.3, i.e., $S = 1.3$, is the same on the data set containing only bob, containing no one, and containing Bob and somebody else. So she has not learned anything about bob. With a higher epsilon, she is able to learn something about bob, and the more the higher epsilon.

Most pricing mechanisms concerned with privacy issues build on that notion and its varieties, such as (ϵ, δ), which is a relaxation for practical purposes (Dwork and A. Roth, 2013). Multiple authors build pricing models based on ϵ- or (ϵ, δ)-differential privacy respectively. A common approach is to perturb data to steer privacy loss and compensate data owners accordingly (H. Cai, F. Ye, et al., 2022; Jinfei Liu et al., 2021; C. Li, D. Y. Li, et al., 2014). For some kinds of data, it is necessary to adapt special forms of differential privacy, e.g. node and edge differential privacy on social network data (M. Zhang, Beltran, and Jiamou Liu, 2020). A big issue is to overcome the fact that privacy loss can be bigger when data is combined with outside data (Z. Zhang, Song, and Yuan Shen, 2021). This is true for background data but also for historic purchases. Also, it is a challenge to cope with correlated queries since their privacy loss does not sum up, but this also provides the opportunity to achieve greater accuracy while maintaining the privacy guarantees via a matrix approach (H. Cai, F. Ye, et al., 2022).

A pricing scheme should not only incorporate privacy loss but also reflect individual privacy attitudes (Ghosh and A. Roth, 2011; Gkatzelis, Aperjis, and Huberman, 2015; Koutsopoulos, Gionis, and Halkidi, 2015; J. Yang and Xing, 2019). (H. Cai, F. Ye, et al., 2022) point out that a fixed compensation to data owners may lead to biased results as more privacy-aware persons drop out of the data market. Therefore the monetary compensation for privacy loss has to be individually different to provide incentives to participate in the market to everyone but it must not set any incentives to claim a higher privacy awareness than what is true. (Ghosh and A. Roth, 2011) developed a truthful auction to solve this problem in which data owners bid their valuation of a marginal increase in differential privacy. Data owners may give an individual privacy cost c_i that linearly increases the payment: $p_i = c_i * \epsilon$ (Koutsopoulos, Gionis, and Halkidi, 2015; C. Li, D. Y. Li, et al., 2014). Other approaches are personalized functions $s(\epsilon)$ which capture nonlinear privacy sensitivities (Jinfei Liu et al., 2021). By adapting the basic concept, personalized differential privacy (Jorgensen, T. Yu, and Cormode, 2015) as applied in (Nget, Y. Cao, and Yoshikawa, 2017) allows every user to set a maximum tolerable privacy loss.

Moreover, privacy loss might not be perceived linearly but is dependent on different risk attitudes (C. Li, D. Y. Li, et al., 2014; J. Yang and Xing, 2019). Risk-neutral data owners would expect a small compensation for small privacy losses and a significantly increasing compensation for bigger privacy losses, whereas risk-averse data owners would like a big compensation already for any privacy losses without a relevant increasing dynamic. This eventually leads to a situation where a big privacy loss is more expensive for risk-neutral data owners than for risk-averse ones.

An alternative to differential privacy is privacy cost. In Koutsopoulos, Gionis, and Halkidi (2015) owners of private data can state a willingness to sell and their data is auctioned to possible data consumers. A more complex setting is investigated by J. Yang and Xing (2019), who differentiate between risk takers and risk averse owners of private data and reflect paid compensations on the privacy attitude.

4.7.5 Enable Fine Granularity

Not all data is equally valuable. Often, only a subset of the data is relevant to the consumer. For example, one customer is only interested in the data from a specific region, another is only interested in a specific variable, and a third does not want to see the raw data at all, but only wants to perform a calculation on it. All these customers are forced to buy the data as a whole, including all the information they are not interested in. This leads to losses in the market, as some customers would have bought the desired excerpt, but the WTP for the purchase of the entire data is not high enough and they, therefore, refrain from buying (Chawla et al., 2019; Z. Zheng et al., 2020). It is therefore desirable to offer fine-grained data purchases and offer flexibility in the way queries can be formulated. Enabling fine-grained queries is also the imperative of model-based pricing (L. Chen, Koutris, and Kumar, 2019; Jinfei Liu et al., 2021). The granularity also extends to differences in the requested quality, i.e. the accuracy of query answers.

Fine granularity also brings the requirement to price data incrementally, which means that past purchases should be reflected in pricing. A customer should have no regret if he buys the data piece by piece instead of all at once (C. Li and Miklau, 2012; Sakr, 2018). However, while the benefits for data consumers are clear, data providers may be worse off offering more flexibility (Q. Li et al., 2021). Moreover, Fine granularity aggravates the challenges of arbitrage freeness and privacy preservation, since the more options a consumer has to choose from, the more opportunities for combining queries and extracting extra information from purchased small data chunks exist. As a result, privacy protection becomes harder, because many fine-grained queries can have a bigger privacy loss than the aggregated privacy loss of individual queries (H. Cai, F. Ye, et al., 2022) and will also require a more meticulous approach to arbitrage freeness (Jinfei Liu et al., 2021).

4.7.6 Deal with Complexity

Pricing data can quickly become a very complex problem. Not atypical are situations in which information from n sellers must be processed into quality perturbation and prices that simultaneously ensure privacy and arbitrage freedom and maximize a profit. By all means, it is important that the market is able to scale the number of customers and transactions. Otherwise, the mechanism is unable to deal with the enormous growth and volume of a future data market. Therefore it is very important to evaluate the computational complexity of pricing (Mao, Z. Zheng, and F. Wu, 2019; An et al., 2017; H. Cai, Zhu, et al., 2019) and of queries themselves (X. Wang, Wei, Y. Liu, et al., 2018; X. Wang, Wei, S. Gao, et al., 2019; B. Shen, Yulong Shen, and Ji, 2019). Further one has to consider requirements from the application field. In many IoT applications, for instance, real-time data streams are required which makes it impractical to calculate the prices during run time Mao, Z. Zheng, and F. Wu (2019). However, even if prices are calculated beforehand, the calculation has to be in polynomial time. Some suggested mechanisms, which are NP-hard, can be solved alternatively via a heuristic, e.g., done by R. Tang, Amarilli, et al. (2016).

Besides time complexity (Mao, Z. Zheng, and F. Wu, 2019; An et al., 2017; H. Cai, F. Ye, et al., 2022), space complexity also plays an important role (Chawla et al., 2019; Niu, Z. Zheng, F. Wu, et al., 2020; Deep and Koutris, 2017a), especially when maintaining a set of knowledge about consumers (Niu, Z. Zheng, F. Wu, et al., 2020). In developing mechanisms for real-world applications, a tradeoff must be made between space and time complexity as well as how far challenges are addressed, e.g., how fine-grained queries can be (Deep and Koutris, 2017a) or only approximately guaranteeing arbitrage freeness (L. Chen, Koutris, and Kumar, 2019). Even if an exact price mechanism is available, it may be necessary to resort to heuristics due to its complexity (R. Tang, H. Wu, Bao, et al., 2013; R. Tang, Amarilli, et al., 2016). Another important aspect is to consider the resource consumption of answering to a query. The processing of a complex query can incur non-negligible costs for the database owner, and a large number of queries can also lead to an excess load on computing resources. This might be a reason to only answer approximately(X. Wang, Wei, Y. Liu, et al., 2018) or include resource consumption in pricing (X. Wang, Wei, Y. Liu, et al., 2018; X. Wang, Wei, S. Gao, et al., 2019).

4.7.7 Create Transparency

Transparency is a very important feature of markets, but a hard-to-define concept. The importance of transparency to all kinds of markets stems from its direct connection to trust. Transparency means that market participants are able to understand what they get, and what they pay, this is they can anticipate the outcomes of a transaction beforehand. This lead to more realistic expectations and subsequently to more satisfaction with the market as such. It also supports the individual rationality of market participants as no one is fooled into a transaction that turns out to be mischievous.

In data markets transparency is twofold. First, the price must be predictable for both consumers and suppliers, and it should be intuitive and understandable how the price is derived (Balazinska, Howe, and Suciu, 2011; Sakr, 2018). This is especially important for dealing with private data. If data owners have to call their privacy cost or select a privacy type, they have to understand the consequences that arise from that decision (J. Yang and Xing, 2019). Transparency can also mean non-repudiation, which can be achieved via smart contracts (W. Xiong and L. Xiong, 2021; K. Liu et al., 2019; P. Gupta et al., 2020). Second, the data product itself must be clear in the sense that it is unambiguous about what data of what quality is traded. If Data Quality is misjudged by consumers the market can be significantly inefficient. Since data is an experience good, it is hard to evaluate quality beforehand (Q. Li et al., 2021). Revealing quality without revealing the data itself is complicated. A solution to this problem is data demonstration. A market could provide samples, i.e. excerpts of the real data, or synthetic data, i.e. data that has the structure of the real data but is made up and therefore uninformative. Pricing mechanisms that deploy Differential Privacy or any form of perturbation have to explain to customers how the data is altered and the consequences for usage and utilization. A data demonstration thereby can fulfill two jobs: First, it can reduce the uncertainty about the value, and second, it can correct a bias, i.e. let a data buyer which is under- or overestimating data value get a more realistic view (Ray, Menon, and Mookerjee, 2020; Q. Li et al., 2021). (Q. Li et al., 2021) found that a provider has no interest in correcting optimistic buyers, while (Ray, Menon, and Mookerjee, 2020) have shown that a data demonstration is still advisable if the buyer has a good enough outside option. In all cases, care must be taken not to reveal too much or even the all of the actual complete database.

4.7.8 Revenue Division

Some papers have researched how to divide the revenue of a query that aggregates data from multiple providers. Shapley values are the foremost suggestion to do so (Agarwal, Dahleh, and Sarkar, 2019; Z. Zheng et al., 2020; Jinfei Liu et al., 2021; Jia et al., 2019; Tian et al., 2022). This approach is promising because it rests on a solid foundation of cooperative game theory and, in particular, axiomatically guarantees a fair division. This is also a compelling approach to compensate the sellers in order to provide incentives to data provision in high-quality (Badewitz, Kloker, and Weinhardt, 2020) as one motivation to split revenue according to Shapley Values is the difference in quality among data providers (Jia et al., 2019).

However, there are important issues with that approach, which also resulted in some authors defining adapted versions. First, Shapley Values are not robust to replication and partition of data (Agarwal, Dahleh, and Sarkar, 2019; You et al., 2021). Second, Shapley prices are not computationally efficient to calculate, thus it is necessary to use appropriate heuristics (Goncalves, Pinson, and Bessa, 2021; Ghorbani and Zou, 2019; Jia et al., 2019). Shapley Values are not the only method to divide revenues. Yan and Procaccia (2021) suggest using the concept of the least core instead of Shapley Values, which are yet dominant in research.

4.7.9 Fraud Prevention

Another problem for markets in general and data markets in specific is the prevention of collusion. Multiple data providers or consumers can collude with the intention to gain a higher profit. Especially in auction settings, bidders have to be restrained from coordinating their bidding behavior in order to maximize their profits. Consequently, some authors have worked on collusion-proof auctions to prevent bidders from colluding with each other (W. Xiong and L. Xiong, 2021). Another issue is consumers creating second identities and bidding multiple times, i.e. false name attacks, which are dealt with in the work of An et al. (2017). P. Gupta et al. (2020) deploy a reputation score, that is dependent on a rating and various measures to assess the credibility of the rating, to prevent actors from fraud and collusion.

Further opportunities for fraud reside in the fact, that data is easily copied. So a data provider can easily act as multiple data providers and sell duplicated and thus redundant data to one buyer (Agarwal, Dahleh, and Sarkar, 2019), as well as a customer, can easily engage in the resale of data (H. Cai, Zhu, et al., 2019).

4.7.10 Conflict of Interest

The value of data is higher if one gains exclusive rights. The more direct competitors also have access to the same or similar data, the weaker the competitive advantage one can gain from the data. However, sharing with non-competitors is merely a little problem. This will become a major problem, and one that is largely unresearched, as more and more sensitive data is traded. H. Cai, Zhu, et al. (2019) suggested an advanced solution by building a conflict graph that indicates reported conflicts between customers. This way they can determine the conflict-free groups and include conflict-behavior and conflict-induced behavior in their pricing mechanism.

4.8 Related Work

There are already some works that have dealt with creating an overview and systematization of data pricing that are complementary reads to this Chapter. Bohli, Sorge, and Westhoff (2009) were one of the first to describe a vision and the path toward a market for data and information goods. Recently, Pei (2022) discussed fundamental principles in both digital and data products and dealt thoroughly with the streams of revenue for digital products. Luong et al. (2016) provides an extensive overview of pricing models in sensing networks for a bunch of different issues such as topology formation, resource allocation, and coverage. M. Zhang and Beltrán (2020) surveyed data pricing mechanisms and take market structures, information symmetry, granularity, and privacy into account. Liang et al. (2018) addressed issues in big data trading and give a detailed analysis of eleven data pricing models. Luong et al. (2016) provides an extensive overview of economic and pricing models as well as market structures in sensing networks for a bunch of different issues such as topology formation, resource allocation, and coverage.

4.9 Conclusion

We have provided a comprehensive overview of the existing literature on data pricing, reviewing 70 current and relevant articles on that topic and extracting the most prominent challenges researchers have already taken care of. We have worked out that a well-suited valuation model based on quality, that guarantees arbitrage freeness in a fine-grained and privacy-protective setting is the gold standard. To fulfill these critical success factors of data pricing, future research can draw on many sound results. From here on, researchers can take four paths. First, most existing research is done for arbitrary environments. There are opportunities and challenges in adapting those general approaches to the needs of specific industries and applications. Second, while the big challenges are being explored more and more, the small challenges are gaining importance. Solving conflicts of interest, developing robust and efficient ways to divide revenues, and incorporating measures to detect and prevent fraud are future fields. Third, although a broad range of research exists for arbitrage freeness as one of the biggest challenges, this remains a diverse challenge and no unique and easily applicable formula is yet found. Fourth, economic interrelations among actors and questions of market structure and power are almost unresearched. The effect of competing data providers or customers on data markets remains an open field.

Chapter 5

Quality-aware Revenue Sharing

> It is not from the benevolence of the butcher, the brewer, or the baker, that we expect our dinner, but from their regard to their own interest.
>
> Adam Smith

Data sharing between companies enables formerly unexploited potential for data-driven business applications. In many cases, it is clear how the owner of the business application will profit from the shared data, but it is still in question how data providers can benefit from data exchange. In the past chapters, competitive pricing was introduced and investigated as a solution to this problem. In this Chapter, we examine closer how pricing can be designed when data quality is not fixed. Therefore, we suggest and analyze two pricing mechanisms that transfer the revenue from data consumers to data providers and give incentives for welfare-optimal data collection.

In order to do so, we developed the Data Provision Game based on a synthesized view of data value chains as a mathematical framework to analyze the economic interactions and incentives in a collaborative data network. We thereby emphasize revenue sharing as an important managing activity as everyone who contributes to a business application should profit from it in order to set the right incentives. Further, we differentiate between entity-borne and network-borne activities. Finally, we formulated Shapley Pricing and Leave-One-Out Pricing as two options for revenue sharing mechanisms, which set incentives for welfare-optimal participation and data quality.

The contents of this chapter are adopted or taken from the paper: Wolfgang Badewitz, Simon Kloker, and Christof Weinhardt (2020). "The Data Provision Game: Researching Revenue Sharing in Collaborative Data Networks". In: *2020 IEEE 22nd Conference on Business Informatics (CBI)*. vol. 01, pp. 191–200. DOI: `10.1109/CBI49978.2020.00028`.
See Appendix A for further details.

5.1 Introduction

Over the last decade, data has become a decisive asset for businesses (Michael E Porter and Heppelmann, 2015). The EU Data Market Study estimates the overall impact of data on the Economy to amount to 477 Bn Euro, which makes up for 3.2% of European GDP (*Data Landscape* 2020). Products become “smart” and are embedded in “smart” services (Pflaum and Gölzer, 2018). More than that, the manufacturing process itself is becoming “smart”. Processes that formerly only consisted of a physical supply chain are now accompanied by a digital data value chain (Hofmann and Rüsch, 2017). Industry 4.0 is designed to be interconnected and information transparent (Hermann, Pentek, and Otto, 2016). Previously separate systems collaborate. They use data from others and share data with others (J. Lee, Kao, and S. Yang, 2014). The more data sources from different companies are included in a data driven business application, the more important the economical aspects of collaborative data value chains become.

How to create Business Value out of Big Data is well studied on a meta-level. While this branch of research structures activities (Curry, 2016; Crié and Micheaux, 2006; Latif et al., 2009) or provides frameworks to create a business case (Poeppelbuss and Durst, 2019; Latif et al., 2009), the economic perspective on data sharing is widely unattended. Companies are reluctant to provide data for a data-driven business application for another company if they are not compensated for their expenses. Pujol Priego, Osimo, and Wareham (2019) examined the approach of 102 companies towards the big data ecosystem and found a vast majority of 89 companies to stay closed. Broek and Veenstra (2015) analyzed the mode of governance in established data collaborations and found forms based on trust or direct mutual benefit (data for data), but no market-based mechanisms. Under these circumstances, the economic potential of applications, which require or could benefit from external data often remains untapped. From the data owners’ view, data monetization is an opportunity to generate an additional stream of revenue. If data is shared inside the supply chain, this will double pay off as the overall competitiveness of the supply chain increases. At its heart, this is an incentive issue as data owners must be compensated for their costs in data collection and the risk they take by data sharing. Otherwise, it would not be individually rational for them to provide data to parties further down the data value chain. If unsolved, that results in an inferior system state.

It is known that revenue sharing is a well-working tool to achieve system efficiency in supply chains (Giannoccaro and Pontrandolfo, 2004). Researchers in the area of Industry 4.0 have addressed the need for revenue sharing and pricing of data (Bucherer and Uckelmann, 2011; Uckelmann and Scholz-Reiter, 2011). Also, researchers have engaged in the design of mechanisms to price data in game-theoretic settings in order to achieve high-quality (Y. Cai, Daskalakis, and Papadimitriou, 2015), to consider privacy (C. Li, D. Y. Li, et al., 2014), or to foster truthfulness (Farokhi, Shames, and Cantoni, 2018). However, research in this area is still in its infancy and currently, there exists no general framework to investigate incentives for all collaborating parties out of a revenue sharing view to optimize overall system performance.

In Section 5.3, we synthesize a framework of collaborative data networks from existing data value chain models from the literature. In our framework, we define the most important business roles in such networks. We focus on a simple structure that emphasizes the role of data providers. In Section 5.4, we formulate a cooperative game, which allows us to economically analyze the incentives in situations where the data provider and the data consumer are not part of the same economic unit. To ensure collaboration inside the network and solve issues regarding incentives, we propose a competitive approach to price data between the collaborating providers and consumers of data. In Section 5.5, we deduce the properties a welfare-optimal revenue sharing mechanism must fulfill and propose two possible mechanisms: Shapley Pricing and Leave-One-Out Pricing. The former is deduced from economic theory, while the latter is a less robust simplification to reduce computational complexity.

The core contribution of this Chapter is twofold. First, the model of a data value chain and the Data Provision Game based on it are frameworks to analyze the economic interactions in collaborative data networks, which can be extended to cover more complex situations than those analyzed in this Chapter. Second, Shapley Pricing and Leave-one-out-Pricing pave the way towards a benefit-based and incentive-compatible pricing of data, which will align the interests of different parties in a data-driven business application by applying a competitive exchange paradigm.

5.2 Related Work

Value chains are an important tool for the economic analysis of companies. A value chain represents value-adding activities and their interlinks in a structured way and thus makes the basic structure and the path from raw materials to finite products visible. This promotes a view on the effectiveness of the whole system (Kaplinsky and Morris, 2000). Historically, product-centered business models dominated and most value chains have been physical. Virtual value chains gained first attention at the end of the 20th century, dealing with the value creation from information (Rayport and Sviokla, 1995). With the increased pace of digital transformation, this topic gained further research attention. To identify the current perception of revenue sharing in data value chains, we carried out a literature review with the keyword *data value chain* and *data supply chain* on the Web of Science (WoS) Core Collection and conducted backward/forward search from the relevant results. A paper was considered relevant if it proposed a model for the data value chain. This way we identified eight contributions for detailed contemplation.

Table 5.1: Overview of Selected Data Value Chain Models

	Activities							
Reference	Collection	Interpretation	Exploitation	Curation	Distribution	Linear Process	Collaboration	Citations (03/2020)
Attard, Orlandi, and Auer (2017)	0	0	0	0	0	No	Yes	6
Crié and Micheaux (2006)	1	3	4	2	-	Yes	No	73
Curry (2016)	1	2	5	3	4	Yes	No	79
H. Hu et al. (2014)	1,2	4	-	-	3	Yes	No	884
Kasim, Hung, and Xiaorong Li (2012)	1	4,5	5,6	2	3	Yes	No	6
Latif et al. (2009)	1	2,3	4	-	-	Yes	Yes	86
Lim et al. (2018)	1	2	4	-	3	Yes	Yes	72
Miller and Mork (2013)	1	3	3	2, 4	2	Yes	No	200
Otto et al. (2019)	1	2	3	0	0	Yes	Yes	2
Faroukhi et al. (2020)	1,2	3,5	6,7	0	4	Yes	Yes	82

Table 5.1 gives an overview of the activities covered in the respective data value chain model, as well as if the activities are ordered in a linear process. If so, a number indicates the position of the activity in the respective model. Otherwise, a zero indicates that no order is induced by the authors. Further, we give in the collaboration column, whether the model features the participation of multiple parties. Finally, we report the number of citations as an indicator of scientific dissemination.

Activities that conceptually appeared to be consistent throughout all models serve as categories and will consequently be used in the synthesized model (see Section 5.3). Collection, interpretation, and exploitation – although called slightly differently in the literature and sometimes were further compartmentalized – are consensual parts of any data value chain. Only the paper by (H. Hu et al., 2014) lacks the exploitation activity. The reason behind this is that they focus on the technologies behind Big Data Analytics and therefore do not trace the data further than interpretation. On the other hand, not all authors do acknowledge the more technical activities such as curation and distribution. Regarding roles inside the data value chain, the IDSA Reference Architecture Model establishes a formal standard for data sharing between organizations in company networks (Otto et al., 2019). The standard includes a detailed description of multiple roles in the data ecosystem in its business layer. We will take over their naming in our model. Surprisingly, no model does consider the issue of revenue sharing yet, although it becomes more important in light of current developments, especially the increasing collaboration within production and service networks. For this reason, we synthesized the existing models into a new model to illustrate its impact.

The most differentiating property of our proposed data value chain to the related work is topology. Topology adds structural information to the value chain. For instance, Porter's Value Chain, which is commonly used in business administration, differentiates between primary and secondary activities (Michael Eugene Porter, 1985). The selected data value chain models exhibit the activities ordered in linear processes. However, there is hardly any consensus regarding the sorting of activities. It is especially complicated to sort the more technical activities into the value chain, e.g., data curation is placed before analysis in the model of (Kasim, Hung, and Xiaorong Li, 2012) and after analysis in the model of (Curry, 2016). Attard, Orlandi, and Auer (2017) argue that the activities cannot be ordered in a linear process at all but build fluid networks. Nonetheless, they agree that data has to be generated before it can be used, inducing a natural order on at least two fundamental activities. Value chains in the domain of knowledge management already implement the topology of Porter's Value Chain (C. L. Wang and Ahmed, 2005; Holsapple and Singh, 2001). The advantage lies in the differentiation between goal-oriented activities ordered into a process and managing activities, which support this process and coordinate between the primary activities. We will use this kind of topology to sort curation and distribution into the value chain.

There is some literature concerning the pricing of data and information goods. Niyato, X. Lu, et al. (2016) analyzed an IoT service market on which sensors only contribute if they are paid more than their reservation wage. Quality of the service increases with the number of participating sensors. Dobakhshari, N. Li, and V. Gupta (2016) analyzed a setting in which a data consumer is interested in a target variable, which can be measured by many strategic sensors. The sensors can bring more effort to achieve a better measurement. The strategic sensors report the target value and variance but will lie if this yields a higher reward. Westenbroek et al. (2018) analyzed a model where multiple data consumers want to build the best model to predict a target variable. The location, i.e., predictors, of each data supplier is fixed but the variance of its reported target variable is dependent on its effort. Y. Cai, Daskalakis, and Papadimitriou (2015) analyzed a similar model with only one data consumer and variable locations. All these works have in common that data sources provide target variables not input data for an analytics service.

In this Chapter, we propose the use of Shapley Values to price data in an incentive-compatible mechanism. Shapley Values have been introduced before to the realm of Machine Learning (ML). Lundberg and S.-I. Lee (2017) introduced the SHapley Additive exPlanations (SHAP) as an feature importance measure. Besides its attention in explanatory ML, they soon raised the interest to assess the economic value of data, e.g., in Jia et al. (2019). Further researchers used Shapley values and derivations to divide the revenue made among data providers (Agarwal, Dahleh, and Sarkar, 2019; Tian et al., 2022) or to calculate fair compensations for privacy loss in a personal data market (Jinfei Liu et al., 2021). The prime complications, however, are complexity (Jia et al., 2019), and a lack of robustness against data replication (You et al., 2021). These issues have been addressed by adapting Shapley values, e.g., in Agarwal, Dahleh, and Sarkar (2019), or the use of heuristics, e.g., in Jia et al. (2019). While Shapley Values are the most important concept for revenue division, leave-one-out (Jinfei Liu et al., 2021) and the least kernel (Yan and Procaccia, 2021) have also been discussed by other researchers. The work in this Chapter is specifically concerned with providing incentives for good data quality during collection, which is only touched upon in the literature but, to our knowledge, has never been a focus.

5.3 Collaborative Data Network

We focus on collaborative data-driven business applications, where many decisions have to be made in high frequency. Based on the related work, we synthesized a new model of a data value chain (see Figure 5.1). We designed this model to explicitly address concerns of collaborative data-driven business applications. Therefore, we differentiate between entity-borne activities and network-borne activities. Primary activ-

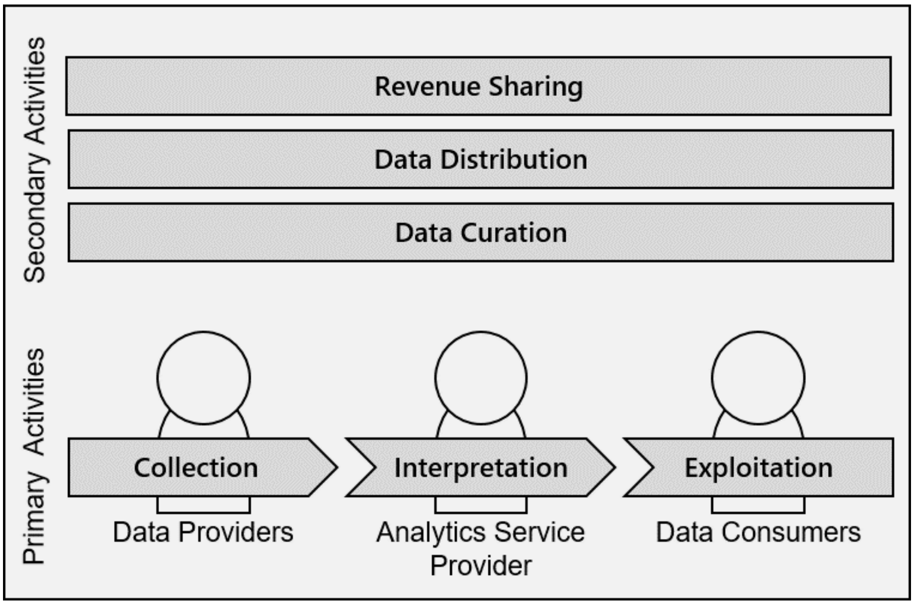


Figure 5.1: The Synthesized Data Value Chain Model

ities are those performed by one entity and immediately serve the purpose of adding or generating value. They are ordered along a primary process, which is traversed by data artifacts. The entities are also the stakeholders of the application and serve as a foundation for the economic agents included in the game-theoretic model. Secondary activities are those performed by the network as a whole and support the primary process. Thereby, secondary activities enable value creation in the first place. Intentionally, topology was chosen to reflect the structure of the popular value chain by Michael Eugene Porter (1985), although we do not claim to reassemble it. Beyond activities found in the body of literature, we incorporate the activity of revenue sharing, which has formerly not been sufficiently considered. This activity is crucial in collaborative scenarios as it ensures that the interests of the different stakeholders are adequately taken into account.

5.3.1 The Primary Process of Value-Enhancing Activities

To think of data as the new oil is a common analogy. Like crude oil is lifted out of the ground and then refined before it literally fuels our economy, data is raw in the beginning and has to be processed and interpreted before it is applicable in decision-making or autonomous application running. This transformation is widely reflected in the literature. Latif et al. (2009) lined out the transformation from raw data to linked data to human-readable data. Lim et al. (2018), as well as Crié and Micheaux (2006), are referring to the data-information-knowledge hierarchy for the fundamental transformation process. In this view, data are mere symbols, whereas information already can be used to answer questions. Knowledge is actionable information; the ability of an entity to use information in order to achieve some end (Rowley, 2007). We also apply that definition. The proposed model distinguishes between three stages in the primary process. The categorization of activities in these stages is based on their in-/output structure. Originally, Collection brings data into existence. Interpretation turns data and information into more information and ultimately knowledge. Finally, Exploitation turns the gained knowledge into business value rather than into other information. Data, information, and knowledge are the artifacts, which can be exchanged between the activities.

However, there have to be activities in which data come into existence and which turn information into business value rather than into other information. The primary process consists of three value-enhancing stages: (i) Collection, (ii) Interpretation, and (iii) Exploitation. Collection activities provide data. The entities performing collection are data providers. Data collection does not require data input. A typical task in a smart manufacturing context would be measuring or sensing (J. Lee, Kao, and S. Yang, 2014). In social science, surveying is the typical mean of data collection (Vaus, 2013). Other data, e.g., sales data, has to be recorded, which is also a collection activity. Interpretation activities aim at refining their inputs into more valuable information. The entities performing interpretation are the analytics service providers. They take data or information as input and provide new information. The activities in the interpretation step are based on the methods from analytics and ML. Exploitation activities finally generate revenue. The entities performing exploitation are data consumers. They take information as input and perform actions based on that information. Exploitation turns a profit by making the enterprise more efficient and effective, e.g., through adding volume and growth, optimizing risks, or reducing costs (Gillon et al., 2014). Without the exploitation activities, the information would not provide any value and only causes cost in measuring and storing. Exploitation can be done autonomously or need human interaction.

Exploitation activities themselves do not generate any new data or information. Nonetheless, another collection activity can be installed to gather data about the result indicators of the exploitation activity. Indeed the process might as well be circular. In reinforcement learning approaches data is collected, interpreted, and used. Then data about the results of the exploitation is collected, interpreted, and used for the next decision. This shows, that exploitation activities and collection activities can be parallelized. Nonetheless, they are two different kinds of activities, have different aims, methods, and requiring different types of expertise. Collection activities deal with how to get data in the first place, while exploitation activities are making business value out of existing data.

5.3.2 The Secondary Activities in Auxiliary Dimensions

All three types of primary activities are sovereign by themselves and can be performed by autonomous entities. This raises the question of integration into the overall process. Secondary activities are auxiliary in nature and integrate the primary activities into the data value ecosystem by coordinating between them. Based on the previously identified models from the literature and our considerations on economic incentives, we identified three secondary activities distribution, curation, and revenue sharing.

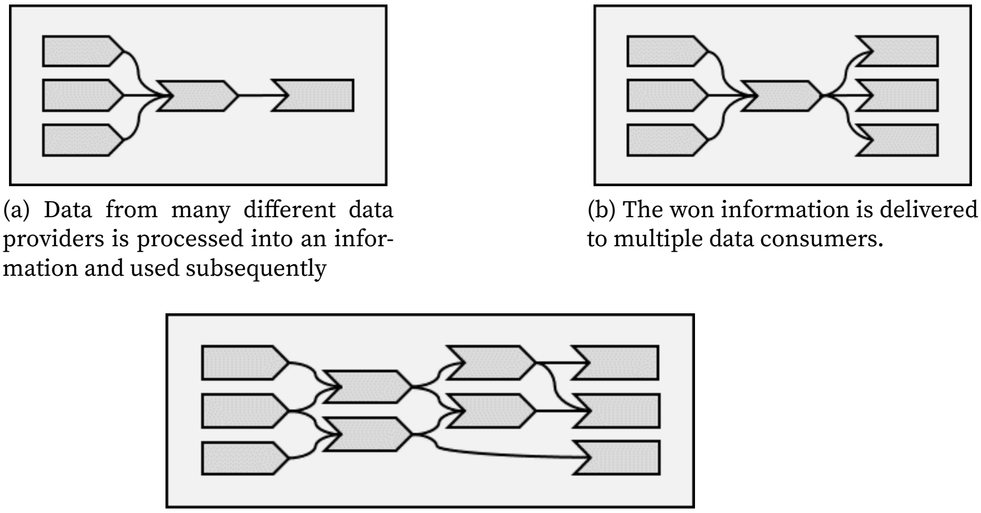
Data distribution activities deal with the aspects of enabling and restricting data access. To be accessible, data has to be stored and transmitted (Lim et al., 2018). Therefore, distribution deals with providing the technical infrastructure and standardized formats. Once data access is technically feasible, data distribution has to ensure data security and privacy (D. Chen and H. Zhao, 2012). Data security restricts access to legitimate users. Data privacy restricts knowledge gain over specific entities by legitimate users. Thus, distribution has also an organizational dimension.

The huge number of various collection and interpretation activities adding huge amounts of highly variable data requires the definition of principles and scalable approaches to solving data quality issues (Freitas and Curry, 2016). Data curation activities are concerned with maintaining and increasing data quality in terms of accuracy, completeness, and consistency. It also has to provide the relevant metadata in order to enable future reuse and preservation in a distributed network (Kasim, Hung, and Xiaorong Li, 2012). Consequently, curation is also closely linked to the achievement of the Fair Principles (Wilkinson et al., 2016).

All activities have a cost and an investment. However, revenues are only achieved in the exploitation activity. Therefore, the expenses of all other activities have to be compensated by the revenue made in the exploitation activity in such a way that every participating entity gains a non-negative profit. This can be done indirectly by offering some kind of service in exchange for data. Another way to share revenue is through direct monetary compensation (Woerner and Wixom, 2015). The data consumers pay the other participating entities for their contribution of data. The advantage of this direct revenue sharing is clear. Services have to be developed, deployed, and demanded. Its configuration is difficult and it does not allow sending as clear incentives as price signals are. Payments on the other hand are easy to handle and applicable to many situations without a lot of thought.

5.3.3 Structures

The minimal primary process consists of one collection and one exploitation activity. The interpretation activity can be skipped, if the collected raw data contains enough information to be actionable. An example from the industrial context are simple security zones. These are set up around some machines in order to prevent injuries. A



(a) Data from many different data providers is processed into an information and used subsequently

(b) The won information is delivered to multiple data consumers.

(c) A complex data value network consists of many interlinked interpretation activities. Information can be used by further Analytic Service Providers and users.

Figure 5.2: Different network structures are possible configurations of the data value chain and vary in complexity for analysis.

light barrier can detect if someone or something enters the security zone, and without further processing, this data can be used to turn the respective machine off. On the other hand, the structure can become arbitrarily complex.

We focus on the collaborative case with multiple providers (see Figure 5.2a). Here many independent data providers deliver data to an analytics service provider, which calculates the information for one data consumer. The same information could also be supplied to multiple data consumers (see Figure 5.2b). We will cover this by aggregating the data consumers into one virtual data consumer. The challenge is to design the revenue sharing in a way such that the providers are incentivized to participate and deliver data in adequate quality.

In the long run, we want to turn towards complex collaborative data value networks as sketched in Figure 5.2c. Many data consumers base their business applications on different information products from various interpretation steps, which may have other interpretation steps as input. Our vision is to develop a pricing framework, which allows tracing the generated business value in the utilization activities back to the data providers. This way, they can be paid meaningful prices, which incentivizes participation and the provision of good data quality, which is commensurate with its contribution to the generated revenue.

5.4 The Data Provision Game

The Data Provision Game transfers the collaborative data network in the simple form with one focal analytics service provider into a mathematical framework. It is a cooperative game with transferable utility, which enables the assessment of revenue sharing mechanisms and their impact on data governance, especially data quality. It allows for raising and answering questions on how to price data and propagate data-generated revenue through the collaborative data network, such that the data providers receive meaningful and effective price signals on their data importance and quality. It can be extended to more complex scenarios than the one covered here.

5.4.1 General Mathematical Model

A data consumer is interested in an information Y. Let $m(X)$ denote an estimator for Y. Thus, Y is referred to as the target variable. X is a set of several independent variables, which will be called predictors.

$$Y = m(X) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{5.1}$$

The j-th predictor is denoted as $X^{(j)}$. For each predictor, the expected value μ_j and the variance σ_j^2 are known. Unfortunately, the data consumer cannot observe the predictors X himself but has to rely on strategic data providers. Each predictor j can be measured by exactly one strategic data provider, which will also be referred to as j. For simplicity, we assume that all measurements are unbiased, but still afflicted with uncertainty. We denote the quantities, which rely on measurements with a tilde. The measurement error on predictor $X^{(j)}$ is denoted as ϵ_j. The measurement uncertainty is given by the variance of the measurement error and denoted as s_j^2.

$$\tilde{X}^{(j)} = X^{(j)} + \epsilon_j,\ X^{(j)} \text{ i.i.d. with } \mathbf{E}\left[X^{(j)}\right] = \mu_j \tag{5.2}$$

$$\mathbf{V}\left[X^{(j)}\right] = \sigma_j \tag{5.3}$$

$$\epsilon_j \sim \mathcal{N}(0, s_j^2) \tag{5.4}$$

A data provider may determine the degree of uncertainty of his measurements by the effort he puts into measuring. A higher effort will result in a better accuracy of the measurement but causes higher cost. The effort is not explicitly modeled. Instead, the measurement error variance s_j^2 is considered as a direct decision variable of the data provider, and a continuously differentiable cost function is given. Additionally, we assume costs to increase with better precision (*condition on the 1st derivative*) and to increase faster at higher levels of precision (*condition on the 2nd derivative*).

$$c(s_j^2) : \mathbf{R}^+ \longrightarrow \mathbf{R}_0^+ \tag{5.5}$$

$$c'(s_j^2) \leq 0 \tag{5.6}$$

$$c''(s_j^2) \geq 0 \tag{5.7}$$

Moreover, every data provider has the choice not to participate, i.e., perform no measurements. If a data provider j decides not to participate, the value of $\tilde{X}^{(j)}$ is set to the expected value of the corresponding predictor μ_j. If not all providers are active, quantities will be indexed with a set of active providers, i.e., $\tilde{X}_A$ and $\tilde{Y}_A$. If data provider j is the only data provider, who is not active, we will write '$-j$' as a shorthand for $\{1, \dots, J\} \setminus \{j\}$.

$$\tilde{X}_A = \left(\tilde{x}_j \mathbf{1}_{\{j \in A\}} + \mu_j \mathbf{1}_{\{j \notin A\}}\right) \tag{5.8}$$

$$\tilde{Y}_A = m(\tilde{X}_A) \tag{5.9}$$

Using the measured predictors an estimate $\tilde{Y} = m(\tilde{X})$ is calculated. This estimate is available to the business application of the data consumer, which will result in a benefit for him. The benefit and thus the revenue of the estimate depends on its quality. Note, that the estimate will be more accurate if the data providers had chosen a higher precision for their measurements. The benefit is at its maximum if the estimate equals the true value, i.e., $\tilde{Y} = Y$, resulting in a revenue of r_{max}. An incorrect estimate causes costs in relation to this optimum. This constitutes a close relationship between the economic revenue of an estimate and the loss used in statistical and ML models. Exploiting this analogy the loss function is used to model revenue.

$$r(\tilde{Y}, Y) : \qquad \mathbf{R}^2 \longrightarrow \mathbf{R} \tag{5.10}$$

$$r(\tilde{Y}, Y) = r_{max} - Loss(\tilde{Y}, Y) \tag{5.11}$$

We shortly discuss two exemplary use cases. Figure a company, which produces a chemical. The target variable could be the degree of contamination in this chemical. If the contamination is too high, the chemical is junk. If this could be estimated before production, the decision maker would be able to take countermeasures and save excess costs. External information, which might be of use in this data analytics task are regarding the production parameters or quality control data from chemical suppliers. Here the data value chain is along the physical value chain. In another use case of this company, the target variable would be the sales volume. Knowing the demand for the chemical allows for better production planning. External information important to this target variable are sales data from its customers. Here the data value chain is directed against the physical supply chain.

5.4.2 Timing of the Data Provision Game

In order to better understand how the Data Provision Game runs, the timing of the game is structurally displayed in Figure 5.3. The timing is important as it defines, which decisions have to be made when and which random variables are already known at this moment. Also, the timing helps to better understand how participants interact with each other.

The game is set up at $t = 0$. At this point in time, the pricing mechanism is specified. The pricing mechanism specifies the payments between the different participants at the end of each round and thus influences the distribution of costs and benefits. Prices are paid from the data consumer to the active data providers. At the start of each round, the data providers can decide whether they participate and if so, with what measurement error variance. During the round, the measurements are performed, and the target variable is estimated and used by the data consumer. This incurs costs for the data providers and profits for the data consumer.

The application in focus should support decisions "that repeat, especially at massive scale", which is one of the two prominent types of data-driven applications (Provost and Fawcett, 2013). Therefore, we model it as an ongoing process. In every round t, all random variables take on new realizations. The data providers have to perform new measurements and the analytics service provider calculates a new estimate. Although we use t as an identifier for a round, we do not focus on time series.

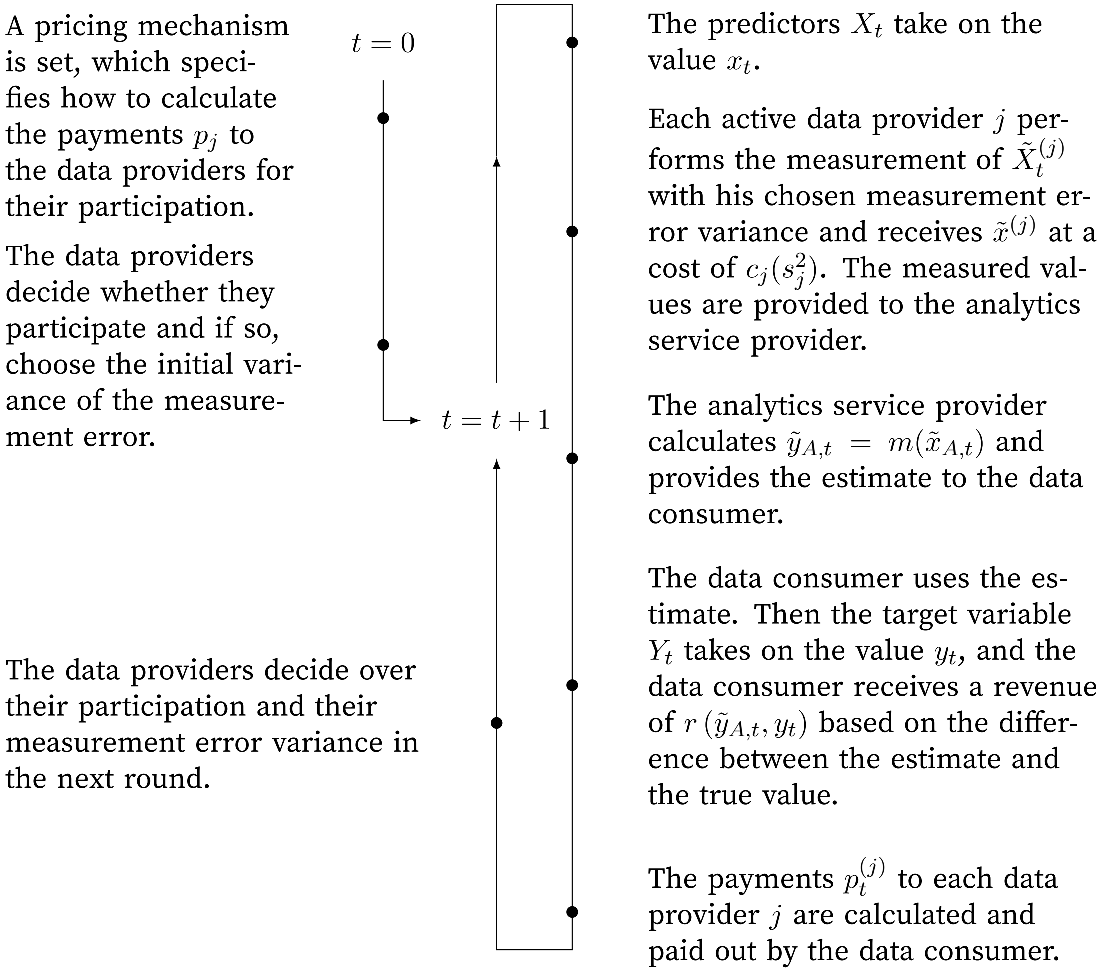
5.4.3 Objectives of the Participants

Data consumers as well as data providers are rational and risk-neutral. They seek to maximize their respective expected payoff from the single rounds of the data application. This is a realistic assumption as payoffs in the single rounds will be small, and the game consists of many independent rounds. Due to the law of large numbers, the mean payoff of rounds during the whole game will tend toward the expected payoff.

The Data Consumers The data consumers owns business applications driven by the information in question. The more accurate information they receive from the analytics service provider, the more profitable the application will be. Their goal is to optimize the benefits from the application minus the due payments to the data providers. In the case of multiple data consumers, they can be aggregated into one data consumer by summing up all individual revenues.

$$\max \mathbf{E}\left[r\left(\tilde{Y}, Y\right) - \sum_{j=1}^{J} \mathbf{1}_{\{j \in A\}}\, p_j\right] \tag{5.12}$$

The Data Providers Each data provider has the goal to maximize their expected profit from participation. The data providers maximize over the measurement error variance s_j^2 and the decision whether to participate at all. If their expected profit turns negative, they do not participate. No participation is equivalent to $s_j^2 = \sigma_j^2$, which is associated with no costs. Since μ_j is generally known and has an error variance of σ_j^2,



A pricing mechanism is set, which specifies how to calculate the payments p_j to the data providers for their participation.

The data providers decide whether they participate and if so, choose the initial variance of the measurement error.

The data providers decide over their participation and their measurement error variance in the next round.

The predictors X_t take on the value x_t.

Each active data provider j performs the measurement of $\tilde{X}_t^{(j)}$ with his chosen measurement error variance and receives $\tilde{x}^{(j)}$ at a cost of $c_j(s_j^2)$. The measured values are provided to the analytics service provider.

The analytics service provider calculates $\tilde{y}_{A,t} = m(\tilde{x}_{A,t})$ and provides the estimate to the data consumer.

The data consumer uses the estimate. Then the target variable Y_t takes on the value y_t, and the data consumer receives a revenue of $r\left(\tilde{y}_{A,t}, y_t\right)$ based on the difference between the estimate and the true value.

The payments $p_t^{(j)}$ to each data provider j are calculated and paid out by the data consumer.

Figure 5.3: Timing of the Data Provision Game

this strategy indeed does not require any action by the data provider at all. Note that, the costs induced by a chosen measurement error variance are deterministic.

$$\max_{s_j^2} \left(\mathbf{E}\left[p_j\right] - c_j\left(s_j^2\right)\right) \quad s.t.\ s_j^2 > 0 \tag{5.13}$$

The Analytics Service Provider The analytics service provider performs the interpretation step from data into information. They are focal in our setting and also serve as the clearing house; calculating the prices and conducting the clearing. However, economically they are neglected in the current state of the game. They neither face costs nor do they benefit.

5.5 Revenue Sharing Design

Based on the Data Provision Game, we can mathematically define objectives for the data pricing mechanisms and deduce the properties that welfare-optimal revenue sharing mechanisms must fulfill. We then show for Shapley Pricing – an application of Shapley Values (Shapley, 1953), which are well-founded in economic theory – that they are welfare-optimal. Due to their computational complexity, we suggest Leave-one-Out-Pricing as a less complex alternative, which is also welfare-optimal. One can distinguish between various overall goals of the pricing mechanism. For instance, one could aim for maximizing the profit of the data consumer. This would be reasonable in a setting, where the data consumer is focal, e.g., because of market power. However, in this work, we take on the perspective of the network. The focal participant in our set-up is the analytics service provider, which is economically neglected. Therefore, we take on a systems perspective and are interested in maximizing welfare.

Definition 5.1 (Welfare).

$$W = r\left(\tilde{Y}, Y\right) - \sum_{j=1}^{J} c_j\left(s_j^2\right)$$

Subordinated to this overall goal, the pricing mechanisms should exhibit some properties, which are Individual Rationality, Self-Selection, and Truthfulness (Faltings and Radanovic, 2017). Truthfulness requires all data providers to share their predictor to the best of their knowledge. In this game, this is given by design. The other two properties are defined in the following:

Definition 5.2 (Individual Rationality). *All data providers, who can provide useful data, and the decision maker have a utility greater or equal zero.*

$$\begin{aligned}&\forall j \in \{1, \ldots, J\}\, \forall A \subseteq \{1, \ldots, J\} \setminus \{j\} \ : \\ &\mathbf{E}\left[W_{A \cup \{j\}}\right] > \mathbf{E}\left[W_A\right] \Rightarrow \exists s_j^2 \ : \ \mathbf{E}\left[p_j - c_j(s_j^2)\right] \geq 0\end{aligned} \tag{5.14}$$

$$\begin{aligned}&\forall A \subseteq \{1, \ldots, J\} \ : \\ &\mathbf{E}\left[v\left(\tilde{Y}_A, Y\right) - \sum_{j \in A} p_j\right] \geq 0\end{aligned} \tag{5.15}$$

Definition 5.3 (Self-Selection). *All data providers, who cannot provide useful data, have a utility less or equal zero.*

$$\begin{aligned}&\forall j \in \{1, \ldots, J\}\, \forall A \subseteq \{1, \ldots, J\} \setminus \{j\} \ : \\ &\mathbf{E}\left[W_{A \cup \{j\}}\right] \leq \mathbf{E}\left[W_A\right] \Rightarrow \forall s_j^2 \ : \ \mathbf{E}\left[p_j - c_j(s_j^2)\right] \leq 0\end{aligned} \tag{5.16}$$

It is important to notice that a trivial payment scheme does not provide any incentives to measure at all. If a constant payment p_j is paid, the payoff of data providers is independent of their effort. Thus, they would only participate with the minimum possible effort. In practice, this means data is shared in the quality in which it is measured anyway because of regulation or internal usage of the data provider. The worst case would be data fraud. A data provider, who pretends to measure, but actually just shares random data in order to get the constant payment.

In this section, we will first investigate the welfare-optimal measurement error variance and secondly show two pricing mechanisms that are sensible for the quality choices of the data providers and set the incentives to choose welfare-optimal. Both pricing mechanisms are based on the idea of paying a data provider on the basis of its added value for forecasting accuracy. Shapley Pricing is an application of Shapley Values, see (Shapley, 1953), which is a common approach from economic theory. As this is computationally complex, we will define Leave-one-out Pricing, which is less robust and not theoretically founded. Its advantage is to be applicable also in the presence of a huge amount of data providers. In the following, we will also assume that

- the dependence between the predictors and the target variable is a known linear model $m(X) = X\beta$.
- the loss function is given by the squared error of the prediction.
- all predictors are uncorrelated.

5.5.1 Welfare-optimal Measurement Error Variance

First, we calculate the formula for expected welfare in the Data Provision Game under the given assumptions. As welfare is total revenue minus total costs, and revenues are dependent on the loss, we will provide three lemmas, which build up on each other, starting with the expected squared error loss.

Lemma 5.1 (Expected Squared Error).

$$\mathbf{E}\left[\left(\tilde{Y} - Y\right)^2\right] = \sigma^2 + \sum_{j=1}^{J} \beta_j^2\, s_j^2 \tag{5.17}$$

Lemma 5.2 (Expected Revenue).

$$\mathbf{E}\left[r\left(\tilde{Y}, Y\right)\right] = r_{max} - \sigma^2 - \sum_{j=1}^{J} \beta_j^2\, s_j^2 \tag{5.18}$$

Lemma 5.3 (Expected Welfare).

$$\mathbf{E}\left[W\right] = r_{max} - \sigma^2 - \sum_{j=1}^{J} \beta_j^2\, s_j^2 + c_j\left(s_j^2\right) \tag{5.19}$$

The expected revenue is solely dependent on the parameters of the game set-up, that is (r_{max}, σ^2, β, c) and the measurement error variance choice of the data providers. Next, we compute the condition the measurement error variance must fulfill in order to be welfare-optimal.

Theorem 5.1 (Welfare-optimal measurement error variance)**.** *In order to maximize the system's welfare, the marginal costs of increased precision in variable j have to equal the corresponding squared coefficient*

$$\forall j \in \{1, ..., J\} : \qquad c_j'\left(s_j^2\right) = -\beta_j^2 \tag{5.20}$$

Proof:

$$\max_{s_j^2} \mathbf{E}\left[W\right]$$

$$\iff \max_{s_j^2} r_{max} - \sigma^2 - \sum_{j=1}^{J} \beta_j^2\, s_j^2 + c_j\left(s_j^2\right)$$

Necessary condition

$$\forall j \in \{1, ..., J\} : \qquad \frac{\mathrm{d}\mathbf{E}\left[W\right]}{\mathrm{d}s_j^2} = -\beta_j^2 - c_j'\left(s_j^2\right) = 0$$

$$\Leftrightarrow \qquad c_j'\left(s_j^2\right) = -\beta_j^2$$

Sufficient condition

$$\forall j \in \{1, ..., J\} : \qquad \frac{\mathrm{d}^2\mathbf{E}\left[W\right]}{\mathrm{d}^2 s_j^4} = \qquad - c_j''\left(s_j^2\right) < 0$$

■

From the functional dependence between marginal costs and the squared coefficients, we can deduce the following four properties of welfare-optimal measurement error variances:

- If the coefficient is zero, the marginal costs are zero. This means the respective data provider does not participate. Only variables with an impact on the target variable are measured, which is known as Self-Selection.
- s_j^2 is symmetrically around a zero coefficient. The same absolute coefficients will lead to the same welfare-optimal measurement error variances.
- s_j^2 is strictly monotonic decreasing in absolute coefficients. Higher coefficients will lead to more precise measurements.
- If the coefficient is not zero, the marginal costs are greater than zero. This means the respective data provider does participate. All variables with an impact on the target variable are measured, which is known as Individual Rationality.

5.5.2 Welfare-optimal Pricing Mechanism

Based on the description of the welfare-optimal state from theorem 1, we can now deduce the condition that any revenue sharing mechanism must meet in order to incentivize the data providers to choose welfare-optimal measurement error variances. The mechanisms which do so will also inherit the properties of being self-selective and individually rational.

Theorem 5.2 (Welfare-optimal Pricing). *A pricing scheme incentivizes the welfare-optimal measurement error variances if and only if*

$$\forall j \in \{1, ..., J\}: \qquad \frac{\mathrm{d}\mathbf{E}\left[p_j\right]}{\mathrm{d}s_j^2} = -\beta_j^2 \tag{5.21}$$

Proof: The objective of each data provider j is to maximize his respective profit:

$$\begin{aligned} &\max_{s_j^2} \mathbf{E}\left[p_j - c_j\left(s_j^2\right)\right] \\ \Longleftrightarrow &\max_{s_j^2} \mathbf{E}\left[p_j\right] - c_j\left(s_j^2\right) \end{aligned}$$

The necessary condition for each data provider j:

$$\begin{aligned} \frac{\mathrm{d}\mathbf{E}\left[p_j\right] - c_j\left(s_j^2\right)}{\mathrm{d}s_j^2} &= \frac{\mathrm{d}\mathbf{E}\left[p_j\right]}{\mathrm{d}s_j^2} - c_j'\left(s_j^2\right) = 0 \\ \Longleftrightarrow & \qquad \frac{\mathrm{d}\mathbf{E}\left[p_j\right]}{\mathrm{d}s_j^2} = c_j'\left(s_j^2\right) \end{aligned}$$

From Theorem 5.1, we know that following has to hold for welfare optimal measurement error variance choice

$$\forall j \in \{1, ..., J\}: \qquad c_j'\left(s_j^2\right) = -\beta_j^2$$

■

In order to make the price for data provider j dependent on the corresponding coefficient β_j and its measurement error variance s_j^2, we will compare the welfare with and without this data provider. Therefore, we next give the expected squared error on an active subset.

Lemma 5.4 (Expected Squared Error on active subsets).

$$\mathbf{E}\left[\left(\tilde{Y}_A - Y\right)^2\right] = \sigma^2 + \sum_{k \in A} \beta_k^2 s_k^2 + \sum_{k \notin A} \beta_k^2 \sigma_k^2 \tag{5.22}$$

Shapley Pricing is an application of Shapley Values, which were developed by (Shapley, 1953) and are well-founded in economic theory to solve coalition games. The idea is to calculate the average added value from the contribution of a provider to all possible combinations of other providers. In this application, this means to compare the gain in precision of the estimate when adding the data from provider j for every possible set of other providers. We then show that Shapley-Pricing is indeed incentivizing welfare-optimal measurement error variances.

Definition 5.4 (Shapley Pricing).

$$\begin{aligned} p_j^{Shapley} &= \sum_{A \subseteq \{1,\ldots,J\}\setminus\{j\}} w_{A,j} \Delta_{A,j,r} \\ \text{, where } w_{A,j} &= \frac{|A|!(J-|A|-1)!}{J!} \\ \Delta_{A,j,r} &= r\left(\tilde{Y}_{A\cup\{j\}}, Y\right) - r\left(\tilde{Y}_A, Y\right) \end{aligned}$$

Corollary 5.2.1. *Shapley Pricing is a welfare-optimal pricing.*

Proof:

$$\mathbf{E}\left[p_j^{Shapley}\right] = \sum_{A \subseteq \{1,\ldots,J\}\setminus\{j\}} w_{A,j} \mathbf{E}\left[\Delta_{A,j,r}\right]$$

It holds:

$$\begin{aligned} \sum_{A \subseteq \{1,\ldots,J\}\setminus\{j\}} w_{A,j} &= 1 \\ \forall j\ : \mathbf{E}\left[\Delta_{A,j,r}\right] &= \beta_j^2\,\sigma_j^2 - \beta_j^2\,s_j^2 \end{aligned}$$

and thus

$$\begin{aligned} \mathbf{E}\left[p_j^{Shapley}\right] &= \beta_j^2\,\sigma_j^2 - \beta_j^2\,s_j^2 \\ \frac{\mathrm{d}\mathbf{E}\left[p_j^{Shapley}\right]}{\mathrm{d}s_j^2} &= -\beta_j^2 \end{aligned}$$

■

Due to the fact, that Shapley Pricing requires calculating estimates for all 2^J possible subsets of active data providers, it has the computational complexity of $\mathcal{O}(2^J)$. This is not applicable if J is a big number. As can be seen in the proof, not only the Shapley prices constitute a welfare-optimal pricing. Also, every single term $\mathbf{E}\left[\Delta_{A,j,r}\right]$ constitutes a welfare-optimal pricing. It is therefore sufficient to only calculate the added value of provider j to the set of all providers without him. This enables the calculation of welfare-optimal prices within $\mathcal{O}(J)$. We will call this method Leave-One-Out-Pricing.

Definition 5.5 (Leave-One-Out Pricing).

$$p_j^{LOO} = \Delta_{\{1,\ldots j-1,j+1,\ldots,J\},j,r}$$

Corollary 5.2.2. *Leave-One-Out Pricing is a welfare-optimal pricing.*

Being computationally less complex though has a downside. Leave-One-Out-pricing is not robust to the correlation between the predictors. For example, if one predictor can be measured by two providers. Leave-one-out pricing would pay them solely on the added value to an estimation in which the other provider of the same predictor is active and therefore paying significantly less than Shapley Prices would do. Leave-One-Out-Pricing would not be welfare-optimal pricing if we do not assume the predictors to be independent.

Numerical Example

We illustrate the proposed pricing mechanism in a short numerical example. Suppose there are two predictors, $X^{(1)}$ and $X^{(2)}$, measured by data providers. The true data-generating process is given by the following formula:

$$Y = 5X^{(1)} - 3X^{(2)} + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, 1) \tag{5.23}$$

Further assume $X^{(1)} \sim \mathcal{N}(0, .5)$ and $X^{(2)} \sim \mathcal{N}(0, .3)$. Both means are set to zero because they would not change the variance of the target variable or the accuracy of its estimate. A nonzero mean could be simply subtracted via an intercept. If both providers deliver their predictors with the same measurement error variance, data provider 1 should be paid more than data provider 2 for two reasons:

1. $|\beta_1| = 5 > 3 = |\beta_2|$, i.e., $X^{(1)}$ has a bigger influence on the target variable than $X^{(2)}$.
2. $\sigma_1^2 = .5 > .3 = \sigma_2^2$, i.e., an unknown $X^{(1)}$ induces more variance on Y than an unknown $X^{(2)}$.

This is reflected in Shapley-Pricing (Def. 5.4) and Leave-One-Out-Pricing (Def. 5.5). With r_{max} =15 ct, Squared Error as Loss in ct, and both providers measuring with a measurement error variance of 0.1; data provider 1 receives an expected payment of 10 ct, while data provider 2 receives an expected payment of 1.8 ct. Note that the expected prices of both mechanisms are identical.

Provider 1 now decreases its measurement error variance to 0.05. This increases its expected payment to 11.25 ct. If the improvement causes less cost than 1.25 ct per measurement, it is fully covered by the mechanism. The expected payment of provider 2 does not change and still is 1.8 ct. This means provider 2 does neither benefit nor come off worse. This prevents data providers from free-riding as well from the danger to be out-competed.

Provider 2 has costs of 2 ct per measurement, thus making a minus with every measurement. It would be inclined to put less effort into data collection. On the other hand, in a collaborative data value network, as shown in Figure 5.2c, $X^{(2)}$ might also be needed in another data analytics service. The income from both justifies the effort of the precise measurement. As both services pay more for higher quality, provider 2 also provides the data to both services to the best of its knowledge. Further, the incentives of different data analytics services also add up. A market using this pricing mechanism has network effects. Services, which share predictors, will profit from each other because the respective providers get incentives to measure and deliver data of a higher quality to both.

5.6 Conclusion & Outlook

We have synthesized the important activities and roles in a data-driven business application and emphasized the need for research on the economic structure of collaborative data networks. Based on this qualitative work, we built a general game-theoretic model for a collaborative data network with one focal analytics service provider and many independent data providers, and one or many aggregated data consumers. This enabled us to quantitatively assess approaches toward data pricing. Our proposed revenue sharing mechanisms are benefit-based. Our suggested Shapley Prices and Leave-One-Out Prices are both solving incentive issues in data sharing and can foster decisions in favor of welfare-optimal data quality by the data providers. While both mechanisms are fulfilling their purpose under our assumptions we already pointed out, that there will be a trade-off between computational complexity and robustness to loosening the assumptions.

Our work is yet restricted to a known linear model and uncorrelated predictors. In future research, we will further assess mechanisms to price data under different models and with fewer assumptions on the predictors. The Data Provision Game is intentionally kept very general around a regression kernel. It will be possible to incorporate statistical learning of the linear regression as a next step. Also, it is possible to incorporate completely different regression models, e.g., regression trees, nonparametric regression, or neural networks. Further analysis of pricing mechanisms in different set-ups will lead to veritable insights into how to design monetary incentive schemes in real-world applications. An evident limitation of the model is, that it is only suited for supervised learning approaches and thus does not cover data-driven business applications, which provide benefits from exploratory analytics. Also, important confounding factors, like security and privacy concerns are yet not built-in and have to be considered in further work.

Chapter 6

Data Utilization in Networks

> Ideas are useless unless used. The proof of their value is in their implementation. Until then, they are in limbo.
>
> Theodore Levitt

In this Chapter, we turn to an application of collaborative data in a value network. We present a Decision Support System (DSS) based on shared ERP data to solve the Capability Matchmaking (CM) problem in supply chain planning. CM is the task to find machines that are capable of producing specific parts. The complexity in this area arises from the high diversity of products and machines and the immense number of possible configurations in a Global Production Network (GPN). Today, this is done by experienced engineers who have the necessary knowledge to assess the feasibility and efficiency of solutions. However, they have limited knowledge of all available machines in the network and are strongly influenced by their personal familiarity with specific products, machines, and locations.

The presented DSS is based on an implicit Recommender System (RS) that predicts possible machine types for parts based on their historical production patterns. System quality will rise given more data and benefit those who receive good recommendations. Since the data is owned by those who consume the information, this is a viable case for collaborative data exchange. The approach constitutes an effective and lightweight option for CM in brownfield settings and a novel application field for the RS-technology.

The contents of this chapter are adopted or taken from the paper: Wolfgang Badewitz, Florian Stamer, et al. (2021). "Recommender Systems for Capability Matchmaking". In: *2021 IEEE 23rd Conference on Business Informatics (CBI)*. vol. 02, pp. 87–96. DOI: `10.1109/CBI52690.2021.10059`.
See Appendix A for further details.

6.1 Introduction

Today, manufacturing takes place in networks of a global scale. A GPN consists of globally distributed production machines connected by their material, information, and financial flows (Lanza et al., 2019). GPNs are characterized by the multitude and variance of machines affected by different factors, e.g., labor cost, automation, culture (Abele et al., 2008), and the resulting complexity. At the same time, supply chains are under pressure to become more flexible and agile due to the trend of modern markets towards smaller batch sizes, shorter lead times, faster reconfiguration, and lower costs (Lasi et al., 2014). Thus, one vital task in the management of these networks' operations is the allocation of production orders to resources. An important step in making allocation decisions is to assess the machines' ability to produce specific products. This step is rather complex considering the extensive quantity and variety of machines in a GPN. It requires engineers with appropriate skills and experience in this context and is a time-consuming and exhaustive task. Moreover, the engineers' search for a feasible allocation is driven by their individual experience and thus results are strongly influenced by their personal familiarity with certain products, machines, and sites. Therefore, it would be desirable to automate or digitally support this task. If successful, the search space would be larger, which would improve allocation decisions, and also less time would be spent on unfeasible allocations. The GPN would benefit from a faster and easier reconfiguration of the supply chain and eventually from a better allocation. Thus, CM fosters more efficient production (Siltala, Järvenpää, and M. Lanz, 2018).

The problem of CM can be solved automatically by leveraging capability ontologies (X. L. Hoang, Hildebrandt, and Fay, 2018; Järvenpää et al., 2017). Capability ontologies provide a framework for saving structured data about the skills of machines, which can be compared against the requirements of products. This solution is accurate and reliable (Järvenpää et al., 2017; Bildstein, Feng, and Bauernhansl, 2018). However, setting up such frameworks has proven to be both complex and expensive (Köcher et al., 2020; Perzylo et al., 2019). In this Chapter, we propose to leverage transactional data on historical production processes to develop a simple and low-effort approach to solve the CM problem. We develop a RS trained on data stemming from ERP. This data is already available, clean, and continuously updated for the entire GPN. Thereby, we avoid creating and maintaining structured, high-quality capability ontologies for all machines and products in the GPN. Primarily, we do not need to populate ontologies with all of the existing GPN infrastructure, which predestines our approach for applications in the brownfield. Finally, the RS is embedded in a DSS that assists engineers in making allocation decisions, but not in an autonomous system making the decisions itself. Against this backdrop, we raise the research questions:

RQ 6.1 *How well suited are RS for supporting CM tasks?*

RQ 6.2 *Is collaborative exchange of ERP data to this end a valid strategy?*

To approach these questions, we implemented an RS using real-world data from a German manufacturer that operates several production sites worldwide. Evaluation results show that our approach is capable of identifying relevant machines for specific production tasks which have never been used for it before. Using this implementation, we built a prototype of a DSS and outlined the next steps to develop a system that can be used in production. Finally, we discuss the theoretical and practical implications, as well as limitations.

6.2 Related Work

Before we focus on the specific problem at hand, Section 6.2.1 elaborates on existing approaches to CM and how related tasks would be solved using state-of-the-art approaches. Existing approaches to machine/ product matchmaking are profoundly automated and exact (X. L. Hoang, Hildebrandt, and Fay, 2018; Järvenpää et al., 2017). However, their setup process entails a high level of complexity (Köcher et al., 2020). Consequently, they are a good fit for newly installed production lines, but unfortunately, they are only to a limited extent suitable for applications on already existing and used machine parks, the so-called brownfield. Nevertheless, the simple fact that these machines are already in operation offers a great opportunity: An extensive database of historic production processes. Instead of approaches solving CM problems using ontologies, our RS approach utilizes this existing data. RS find patterns in data sets and predict good combinations utilizing Machine Learning (ML) and information retrieval methods. Section 6.2.2 presents the technical foundations of RS and discusses their current dissemination and application with a focus on Manufacturing.

6.2.1 Capability Ontologies and Matchmaking

The full-fledged engineering approach matches product requirements with machine capabilities. To do so, capabilities, as well as interfaces and other properties of machines, are formally described in a resource description concept (Siltala, Järvenpää, and M. Lanz, 2018). Various capability models have been introduced by scholars with the aim of providing a sound basis for matching product requirements and machine capabilities, e.g., (Köcher et al., 2020; Perzylo et al., 2019; Himmelhuber et al., 2020). The majority of these follow an ontological approach.

Järvenpää et al. (2017) presented an approach to CM based on an ontological representation of products and resources. Therefore, product requirements and resource capabilities are compared with each other on skill level. It supports the search for feasible resources and the detection of discrepancies in existing configurations. In this way, the engineers in charge can perform their tasks more efficiently.

Next, X. L. Hoang, Hildebrandt, and Fay (2018) describes an approach that strictly separates between product, process, and resource. In their approach, it is obsolete to derive skills from product descriptions. Instead, they focus on comparing property spaces. Thereby, they are able to reduce complexity, communication, and planning efforts. However, complexity may arise from not sufficiently standardized products.

The approaches have in common that they require a high level of expertise and rely on structured, clean, and accessible metadata about machines and products, which has to be maintained and updated. Modeling capabilities in-depth is a complex and time-consuming task (Perzylo et al., 2019) and rules have to be simplified, potentially causing crude results (Järvenpää et al., 2017). Product designers may also lack the ability to translate product specifications into specific production requirements (X. L. Hoang, Hildebrandt, and Fay, 2018). The engineering approach also demands high invest and high effort in its set-up, which will only be amortized over the long run (Köcher et al., 2020). The advantages on the other hand are high accuracy and reliability, and thus ontological approaches are suited for building autonomous systems (Bildstein, Feng, and Bauernhansl, 2018; Järvenpää et al., 2017).

6.2.2 Recommender Systems

In decision-making where there is a wide variety of choices, RS provide a suitable means to predict eligible item/ user combinations (Bobadilla et al., 2013; Ricci, Rokach, and Shapira, 2011). To do so, RS analyze historic transactions between users with diverse goals and interests with items, the 'objects to recommend' (Aguilar, Valdiviezo-Díaz, and Riofrio, 2017). They are proven to reduce information overload, for instance by drawing attention to well fitting niche items (Oestreicher-Singer and Sundararajan, 2012), and, ultimately, improve decision quality (Xiao and Benbasat, 2007; J. D. Xu, I. Benbasat, and Cenfetelli, 2014).

Technological Foundations

RS can be classified based on the technique the data has been processed with. Among those, the most common are Content-Based Filtering (CBF), Knowledge-Based Filtering (KBF), and Collaborative Filtering (CF)(J. Lu et al., 2015). CBF approaches use descriptive properties of users and items to find commonalities (Aggarwal, 2016). Next, KBF approaches require existing knowledge about the scouted solution, e.g., how many machining axes are required and filter the possible choices accordingly (Aggarwal, 2016). Last, CF approaches rely solely on the rating matrix. The rating matrix has an entry for each user/ item combination, which may be 'none' if the specific combination is not rated. A common approach for model-based CF is dimensionality reduction, especially matrix factorization. In this, the rating matrix is approximated via the multiplication of two lower-dimensional factor matrices that represent latent factors (Lü et al., 2012). The entries in the rating matrix may represent a definitive preference score or an implicit confidence score. Explicit information contains ratings of past choices, e.g., user A gave item x five stars, while implicit information contains only observations of the past behavior, e.g., user A bought item x 32-times (Bobadilla et al., 2013). Due to these differences, explicit feedback states preferences directly, while implicit data can only be used to derive confidence (Y. Hu, Volinsky, and Koren, 2008). A preference score is an ordinal feature stating how good a combination is, which can be used as ground truth for the quality of a combination. A confidence score, on the other hand, does not contain any information about how good or bad a recommendation is, but rather how probable a recommendation is good. High confidence implies that it is more probable that a combination is good, while low confidence implies that it is less probable. This adds another layer of uncertainty to the RS and requires that the rating matrix is processed differently. Subsequently, the interpretation of the recommendations is also affected and must be done accordingly.

Applications in Smart Manufacturing

RS find widespread application in various domains, such as e-commerce, e-resource, and e-government (J. Lu et al., 2015), but so far, they have stayed within customer-centric and marketing-oriented areas. Only few research is conducted towards aligning such RS with objectives from manufacturing or supply chain operations (C. Zhang et al., 2019; Dadouchi and Agard, 2020). Our approach takes things one step further. We deploy an RS to find well fitting machines for parts. Contrary to the traditional areas in which RS are deployed, the user makes the decision, but his personal preferences are negligible. Instead, the objective requirements of the parts are decisive. To the best of our knowledge, RS have not been applied to CM so far, neither they find wide dissemination in Smart Manufacturing or other domains to support knowledge workers in challenging tasks of a similar kind. However, the potential of RS in similar problem areas is recognized (Khakifirooz, Fathi, and K. Wu, 2019; Alinani et al., 2019), and some research is aimed precisely in this direction (X. Chen and Jin, 2020; Lehmann, Shamiyeh, and Ziemer, 2019; Z. Liu et al., 2021; Simeone, Zeng, and Caggiano, 2021).

RS have been used to retrieve text-based knowledge from a vast corpus. Lehmann, Shamiyeh, and Ziemer (2019) propose ontological RS to increase the productivity of knowledge workers in compiling technical reports. In providing references to relevant files, double work should be avoided and existing knowledge reused. Technologically, their case is founded on semantic modeling and text processing. A. Trappey,

C. V. Trappey, and Hsieh (2021) developed an RS to identify relevant patents in the smart machinery domain using natural language processing. Their approach leads to an enhancement of technology mining and trend analysis. Oboe (C. Yang et al., 2019) and AdaPipe (X. Chen and Jin, 2020) used RS for automated ML, the latter explicitly in the context of industrial cyber-physical systems. These tools reduce the required sophistication of specific manufacturing processes and data-driven analysis in order to design good computational services. This might considerably affect the efficacy and efficiency of knowledge workers in this field. Closest to our use case is the application of RS in Cloud Manufacturing (CMft). Analogously to other cloud technologies, CMft enables 'ubiquitous, convenient, on-demand network access to a shared pool of configurable manufacturing resources' (X. Xu, 2012). One of the main issues of CMft is optimal resource allocation, which requires, among other things, intelligent matching between resources and tasks (L. Zhang et al., 2014). To help customers filter out the best manufacturing service, RS have recently been suggested (Alinani et al., 2019) and prototyped (Z. Liu et al., 2021; Simeone, Zeng, and Caggiano, 2021).

Similar to our work, the relevance of a recommendation in all these applications does not stem from mere user preference but from the task, a user has to complete. Also, the aim is to support knowledge workers in their daily business and increase the efficiency and efficacy in which they perform their respective tasks. Notwithstanding, the developed DSS in this Chapter differs from the solutions available in the literature in the specific application, i.e., CM, the utilized data, i.e., ERP data), and the used methods, i.e., implicit CF.

6.3 Development of the Decision Support System

Our goal is to create an RS-based DSS that enables more flexibility in production planning. This should be achieved by making the process of allocating products to resources more efficient, thereby accelerating decision-making and reducing organizational cost. The main issue to solve is determining the technical feasibility of supply chain scenarios. Therefore, we must be able to identify machines that are capable of producing certain products automatically. The method for doing so has to be effective in terms of decision quality and build upon existing and accessible data. We apply RS to the CM task in order to fulfill these design goals. The production orders in consideration contain a list of parts. Each part is a plain mechanic component, which can be produced by one production entity alone. The search for a machine type capable of producing a specific part from this list poses the same problems and needs as RS are trying to solve: From a huge quantity of machines, the engineers in charge are looking for machines that are (i) relevant, i.e., a machine, which can produce this part; (ii) novel, i.e., a machine, which never produced this part before and that was not even on the radar of the engineers; and (iii) serendipitous, i.e., unconventional, but possible solutions to produce a part, which most probably would not have come up else and can possibly prove very useful (Aggarwal, 2016). The difference to traditional RS applications is, that there is no need to predict the preference of the user, i.e., the engineer. We implemented an RS to give meaningful recommendations which support the engineers in solving the CM. We trained and evaluated it with an excerpt of the ERP data obtained in the GPN of a German manufacturing company. For the development of the DSS, and the structure of the remainder, we followed the CRISP-DM Cycle (Shearer, 2000).

6.3.1 Business Understanding

Our use case partner is a global company headquartered in Germany, mainly active in process and automation engineering. It is a large manufacturer of a wide variety of products produced in a GPN with several production sites worldwide. Besides the sites in Germany, production facilities are located in Eastern Europe, Asia, and America. Products are assembled from simple machined parts. These parts are produced in one production step on one machine. However, multiple machines from different machine types are, in principle, feasible choices for production. Among other responsibilities, the strategic Supply Chain Management (SCM) department prepares and implements the decisions of what parts to produce on which machines. A recurring challenge here is not knowing which machines can produce which parts when considering all available machines in the GPN. This challenge can be illustrated with two common tasks; the *introduction of new products* and the *reaction to demand spikes*.

Introduction of New Products: When a new product is to be introduced into the GPN, a bill of materials is created consisting of the individual parts. In the first step, supply chain engineers search for machines that can produce these parts and create supply chain scenarios. The scenarios map the bill of materials to specific production sites and machines. Based on the capable machines found, different scenarios are created and compared in terms of reliability, cost, time, and other criteria. A central authority selects the scenario to be executed. At this stage, a manual check is always required. Finally, the production sites and machines receive the production order and manufacture the parts. Since the scenarios can only consist of machines that the engineers identified in the first step, it is crucial that the search scope for machines is large enough, yet the search itself must remain efficient. In addition, the risks for investments can be reduced if product startups can initially be launched on existing machines and product- or line-specific machines are only purchased later as quantities increase.

Reaction to demand spikes: In existing production lines, the company is challenged with the occasional occurrence of demand spikes. Most products are produced in low and often constant numbers, but every once in a while, a customer buys products in quantities that are an order of magnitude higher than the norm. The generic supply chain scenario for this product then runs in a capacity bottleneck. In such cases, the strategic supply chain department looks for alternative production possibilities. Here, it is important to quickly develop a feasible solution, which may consist of a different machine from a different machine type to produce a part.

The DSS supports engineers in both tasks. So far engineers manually check machines whether they are capable of producing a certain part with little data-driven decision support. This process is time-consuming and exhausting as machine capabilities are not stored in a central database and intensive communication with decentral experts is required to design a feasible production concept. Thus, engineers are strongly influenced by their personal familiarity with parts, production sites, and machines. A DSS will open up the search space and help to find a suitable solution more efficiently as it gives the engineers a list of promising machines at hand. A list of machines from the production history would already have the potential to support the engineer if he were not aware of some of these machines. Even better and more applicable would be a system that judges every machine in the GPN and provides the engineer with the most promising candidates. This way novel machines that have never been used for producing a specific part are included in the recommendations as well. Note that the DSS should not become an autonomous system for matchmaking and will always require a manual check and confirmation. Nonetheless, the work of engineers is usefully supported, and their workload is reduced.

Moving beyond the scope of a single company, it is possible to merge data from multiple companies to increase the performance of the DSS or to enable it in the first place if a single company has too little data to learn meaningful recommendations, e.g., a small contract manufacturer in a CMft environment. However, this data contains sensible corporate information, for instance about the order situation. Therefore, any unnecessary information has to be removed and data is best to be processed by a trustworthy third party that only hands out recommendations. If the risk of leakage is less than the value of the application, companies should participate. However, companies could try to free ride, i.e., they could use the application but share manipulated or extracted data. This can be prevented by organizational measures of the analytics service provider, e.g., fraud detection and minimum quotas on sending data volumes, while accepting that data from particularly critical processes will be withheld. In addition, companies harm themselves through malicious behavior because they reduce system quality, especially with regard to their own machines and parts.

6.3.2 Data Understanding and Preparation

The data at hand is comparatively straightforward. We have retrieved a representative excerpt from the ERP system containing about 200,000 processes from four production facilities across two years. Table 6.1 provides an example of the data structure with toy entries. Each entry contains the ID of the production process, a timestamp, the PartID, and the MachineID. A MachineID represents a specific machine of a machine type, which can also be assessed. The company owns several machines of the same machine types. Every machine of a certain machine type has the same capabilities. Other accessible information about a machine's type is, for example, whether it is a milling or a turning machine. However, this additional accessible information is much too coarse to help solve the CM task, e.g., some milling machines have to turn axes. That information is consequently are omitted. Further data about machines, e.g., on their respective site, may have supply chain implications and are of importance to further planning but are not relevant to solve the CM task in the first place. Hence we do not use this data to train the RS, but display them in the DSS.

The data contains slightly more than 200 different machine types and almost 20,000 different parts. Only historic production processes are recorded. Out of over four million theoretical combinations of parts and machine types, approximately 25,000 different combinations have been documented, making the data very sparse. Moreover, there is no explicit feedback of machine-part-combinations, neither are evaluations by human experts available nor is a quality or cost evaluation derivable from the ERP logs. Thus, the data are purely implicit (Oard and Kim, 1998; Y. Hu, Volinsky, and Koren, 2008). This means we can neither judge whether the production of a part on a machine type has been efficient in the past or has caused high costs and infe-

Table 6.1: ERP-Extract (Fictitious Example)

ID	Date	PartID	MachineID	MachineType	Technology
1	2019/03/12	189853	2354	TBAS RZ3	Milling
2	2019/03/12	42343	9375	SDA X2-300	Turning
3	2019/03/12	231093	5235	TBAS RZ3	Milling

Table 6.2: Prepared Data (Fictitious Example)

PartID	MachineType	Interactions
189853	TBAS RZ3	12
231093	TBAS RZ3	3
42343	SDA X2-300	17

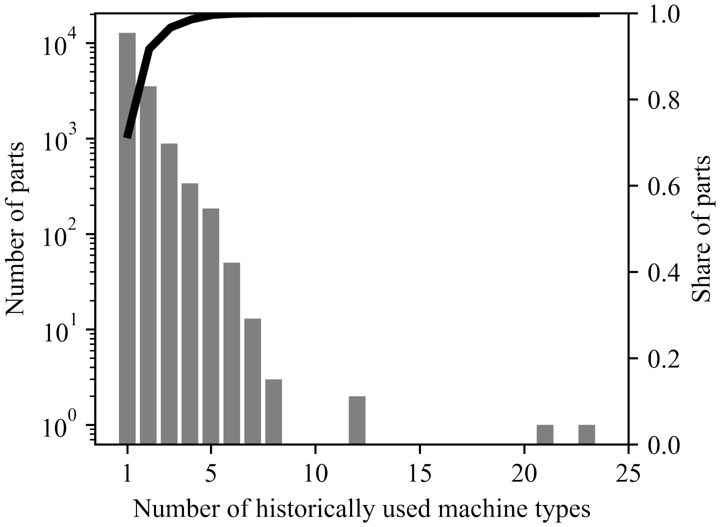


Figure 6.1: Number and share of parts per number of used machine types. Depiction taken from Badewitz, Stamer, et al. (2021)

rior quality, nor whether a combination of a part and a machine type that we cannot observe is technically impossible or uneconomical. We use the number of observed combinations of the same part and machine type as confidence that the combination is satisfactory. Table 6.2 shows an example of the prepared data. A specific part and a machine-type combination, that is observed frequently, can be presumed to constitute an effective and efficient choice, simply because it is chosen often. On the other hand, rarely observed combinations are most likely less effective and efficient, but it is quite possible that some of them would be equivalent or even better suited from a capability perspective. Production decisions are complex with many restrictions. The best machine to produce a part might not have enough free capacity, such that a sub-optimal choice appears more frequently in the data. At the same time, a rare combination can be a hint of a makeshift solution because the optimal choice is temporarily not available. If a combination only exists once in the data, such as most combinations, this may be due to a failed trial, a mistake in data entry, or simply because this specific part was only produced once in the recorded time frame.

Figure 6.1 shows how many parts were produced on how many different machine types. The solid black line gives the percentage of parts that are produced on less or equal different machine types. 72% of parts were produced on only one machine type. Over 95% of parts were produced on three or fewer different machine types. Over 99% of parts were produced on five or fewer machine types. The allocation of parts to machines currently is relatively rigid, producing most parts with little flexibility always on the same or only a very small set of machine types. It also poses the requirement

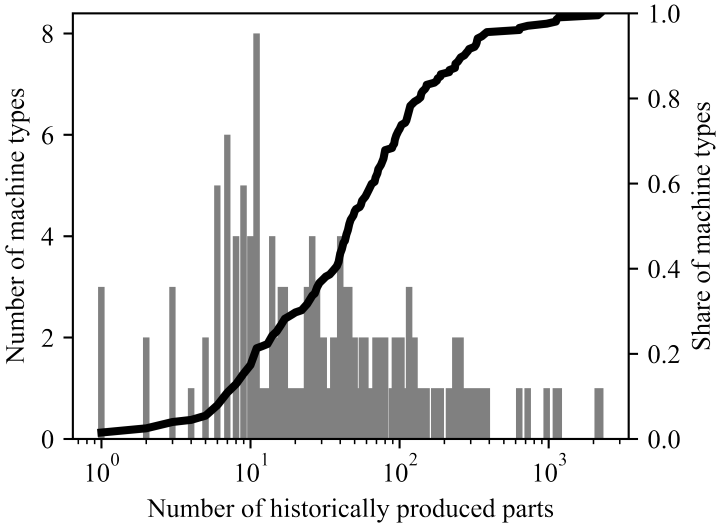


Figure 6.2: Number and share of machine types per number of produced parts. Depiction taken from Badewitz, Stamer, et al. (2021).

for the RS to give meaningful recommendations for parts with only one historically used machine and to cope with this form of sparsity. The difficulty is comparable to recommending a movie to a user you only know they have seen 'Avatar' but do not even know if they liked it. Figure 6.2 shows the frequencies from the machine type perspective, i.e., how many machine types have produced how many different parts. Most machines produce between 10 and 140 parts. Only 20% of machines produce less than 10 parts. Thereof it is seen that machine types are not rigidly used for one specific part only, but used for the production of many parts. Also, this presents the opportunity for the RS to learn in the machine-type dimension since the data is not as sparse as in the part's dimension.

6.3.3 Modeling and Evaluation

We have built two RS following a model-based CF approach. A CBF approach was infeasible as the descriptive properties of both parts and machines are not stored in a structured way and, therefore, cannot be extracted easily and accurately. A neighborhood-based approach was infeasible because neighborhoods of numerous parts are nearly empty due to high data sparsity. We consider a dimensional reduction model particularly well suited for parts with only one machine in the production history to estimate those with latent factors. As a further requirement, the model must be able to handle implicit ERP data. Therefore, we use the established Alternating Least Squares (ALS) algorithm (Y. Hu, Volinsky, and Koren, 2008), and the Logistic Matrix Factorization

(LMF) algorithm (Johnson, 2014).[1]

Let $r_{i,j}$ be the number with which a part i was historically produced on a machine type j. ALS (Y. Hu, Volinsky, and Koren, 2008) models each combination with a preference $p_{i,j}$ of 1, if it is historically observed, i.e., $r_{i,j} \geq 1$, and 0 otherwise, i.e., $r_{i,j} = 0$, and the confidence in observing this preference with $c_{i,j} = 1 + \alpha r_{i,j}$. It tries to find good vectors $x_i \in \mathbf{R}^f$ for each part and $y_j \in \mathbf{R}^f$ for each machine type, where f denotes the number of factors, such that the preference is estimated by the inner product $x_i^\top y_j$ weighted by the confidence and regularized:

$$min_{x,y} \sum_{i,j} c_{i,j}(p_{i,j} - x_i^\top y_j)^2 + \lambda(\sum_i ||x_i||^2 + \sum_j ||y_j||^2) \tag{6.1}$$

LMF (Johnson, 2014) assumes the preference to be distributed according to a logistic function (see Equation (6.2)) parameterized with $x_i^T y_j + \beta_i + \beta_j$, where β is a bias for parts and machine types respectively. The algorithm maximizes the log-likelihood regularized by a Gaussian prior on part and machine-type vectors.

$$\mathbf{P}(p_{i,j} = 1) = \frac{exp(x_i^\top y_j + \beta_i + \beta_j)}{1 + exp(x_i^\top y_j + \beta_i + \beta_j)} \tag{6.2}$$

When evaluating explicit data, distinct metrics based on the distance between predicted and given preference can measure predictive success. This is impossible for implicit data because no ground truth for preference is available (Y. Hu, Volinsky, and Koren, 2008). Instead, we fall back on established Information Retrieval evaluation measures to estimate RS machine-type recommendation quality. We define a machine type as relevant for a specific part if it has historically been used to produce a part. The employed evaluation criteria include:

Mean Reciprocal Rank (MRR): The reciprocal value of the harmonic mean over the ranks of the first relevant recommendation for a query, e.g., if the third recommendation is a relevant machine type, the MRR is $1/3$.

Precision@k: The fraction of relevant machine types among the top-k recommended machine types.

Recall@k: The fraction of top-k-recommended, relevant machine types among all relevant machine types.

F-Score@k: The harmonic mean of Precision@k and Recall@k.

For all evaluations, we regarded the first five recommendations. In order to parameterize the model with the optimal number of factors and avoid overfitting, we

[1] For both, we use the python implementation of the ‘implicit’ package by Frederickson (2017).

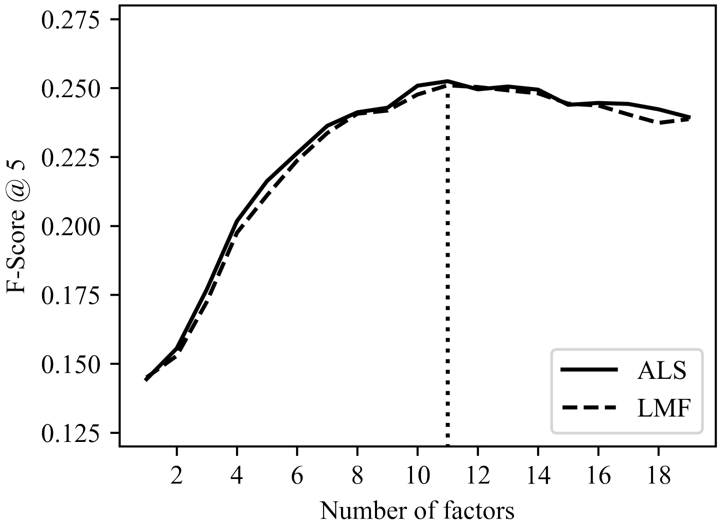


Figure 6.3: Mean F-Score@5 over 20 iterations of 70-30 train-test splits dependent on the number of factors used for ALS and LMF respectively. Note that the maximum achievable F-Score@5 in this setting is 0.38. Depiction is taken from Badewitz, Stamer, et al. (2021).

ran 20 iterations of each factor between 1 and 20. In each iteration, we applied a random 70-30 train-test split. We evaluated the top-five recommendation for all parts with at least one machine type in the test set. The recommendations of a part were filtered for the machine types in the training set, such that evaluation only refers to 'novel' machine types. In this setting, the withheld machine-types of a part were the relevant machines. This way the metrics purely indicate how well unknown machine types can be recommended and do not consider how well known machine types are reproduced. The F-Score@5 for varying numbers of factors are shown in Figure 6.3. Both models perform about equally well and reach their maximum F-score at 11 factors with 0.25. Note that a part having only one machine type in the test set can achieve a maximum Precision@5 of $1/5$. In our data, this case applies on average to about 60% of the parts. Therefore, the maximum achievable Precision@5 in this setting is 0.2405. This also influences the maximum achievable F-Score, which is 0.3877. As the number of factors increases, the F-Scores increase until eleven factors are reached and then decrease because of overfitting to the known machine types from the training set.

The exact results for all evaluation criteria at 11 factors are given in Table 6.3. The last line of the table sketches the achievable optimum of the criteria for the run performed. Both models perform about equally well regarding every criterion, with ALS only slightly better in each. First, we can state that almost always the first or the second recommendation is already a relevant machine type. Second, round about 65% of all relevant machine types are recommended in the top five. Third, the Precision@5 is about 2/3 of its maximum value. Also, note that all machine types, which are recommended but marked as non-relevant, might as well be relevant. Therefore, we assume our results as a lower bound to the real precision, since we have conservatively marked all combinations without prior data as irrelevant. In the remainder of this Chapter and for the purpose of the DSS itself, we deploy the ALS algorithm.

It is of vital importance for the RS to perform well on parts that are historically only produced on very few machine types, especially on those produced on only one machine type. For those parts, having access to good recommendations of novel machine types has the most considerable effect. To evaluate how good the RS recommend machine types for parts with *n* known machine types, we used the following approach: For every part that used more than *n* machine types historically, we learned a model with randomly chosen *n* machine types of this part for training and tried to predict the missing machine types.

Table 6.4 shows the evaluation metrics in dependence on how many entries a part in the training set has. Parts, which have fewer machines in the training set perform worse than those with more training entries. However, the RS already performs well on parts with only one entry in the training set, retrieving almost 60% of relevant machine types in the top-five recommendations. In interpreting the precision, which stays the same for parts with two and three training entries and then even decreases, note again that with an increasing number of entries in the training set, fewer parts are in the test set, making it harder to achieve a good precision at a constant number of recommendations. These results are promising insofar as they outline the ability of RS to create relevant and novel machine types for the majority of parts, which were only produced on one machine type hitherto.

6.3.4 Deployment

The DSS is implemented and deployed as a Python-based web application accessible within the manufacturer's intranet. The application is designed as a DSS to help engineers configure the supply chain scenario. Figure 6.4 depicts an illustration of the current development status. Note that all identifiers of parts and machines are pseudonymized. On the left side, the view shows a list of recommended, never-used

Table 6.3: Evaluation Results of Different Approaches

Model	MRR	Precision@5	Recall@5	F-Score@5
ALS	0.4653	0.1560	0.6618	0.2525
LMF	0.4647	0.1550	0.6588	0.2510
Maximum achievable	1.0000	0.2405	0.9994	0.3877

machine types yielded by the RS. At the top right side, the machine types that have been used in the past to produce are listed. Specific machines and their associated costs in the supply chain are further valuable information for a production decision and are outlined here. The costs can be calculated, but are randomly generated for publication. Additionally, the view provides information regarding specific existing machines of this type and their related costs. The provision of explanations to the recommendations is highly relevant since explainability has proven to be a vital acceptance fostering element in adopting novel ML solutions (Goebel et al., 2018). For each recommended machine type, an explanation can be displayed by clicking the question mark next to it. The explanations consist of the top contributing already used machine types that have led to the recommendation of a particular machine type, which can be identified by measuring the contribution of each past combination as presented by Y. Hu, Volinsky, and Koren (2008). Additionally, the bottom right area shows similar parts based on production history to increase explainability. To ensure continuous re-training and optimization of the RS, the view also incorporates a feedback mechanism. For this, engineers may evaluate provided recommendations by means of ordinary thumbs surveys. Thereby, we incorporate the option to provide feedback on both historical machine-part combinations and recommendations at run time. The goal is to collect as much feedback as possible while ensuring that the feedback routine does not negatively affect the perceived value and usability of the tool.

The prototype has been tested by three engineers – the explicit target group of the DSS. During the evaluation, the engineers always considered the first five recommendations for a part number. Their general conclusion was that the prototype could provide decision support for their daily business in the future. In its present form, however, the prototype still does not provide enough information on the alternatives. In particular, the engineers demanded more detailed information regarding machine types, e.g., how many machining axes it has, and specific machines, e.g., where are they located and what is their current utilization. While it would be easy to supplement the recommendations with some of the desired information, such as locations,

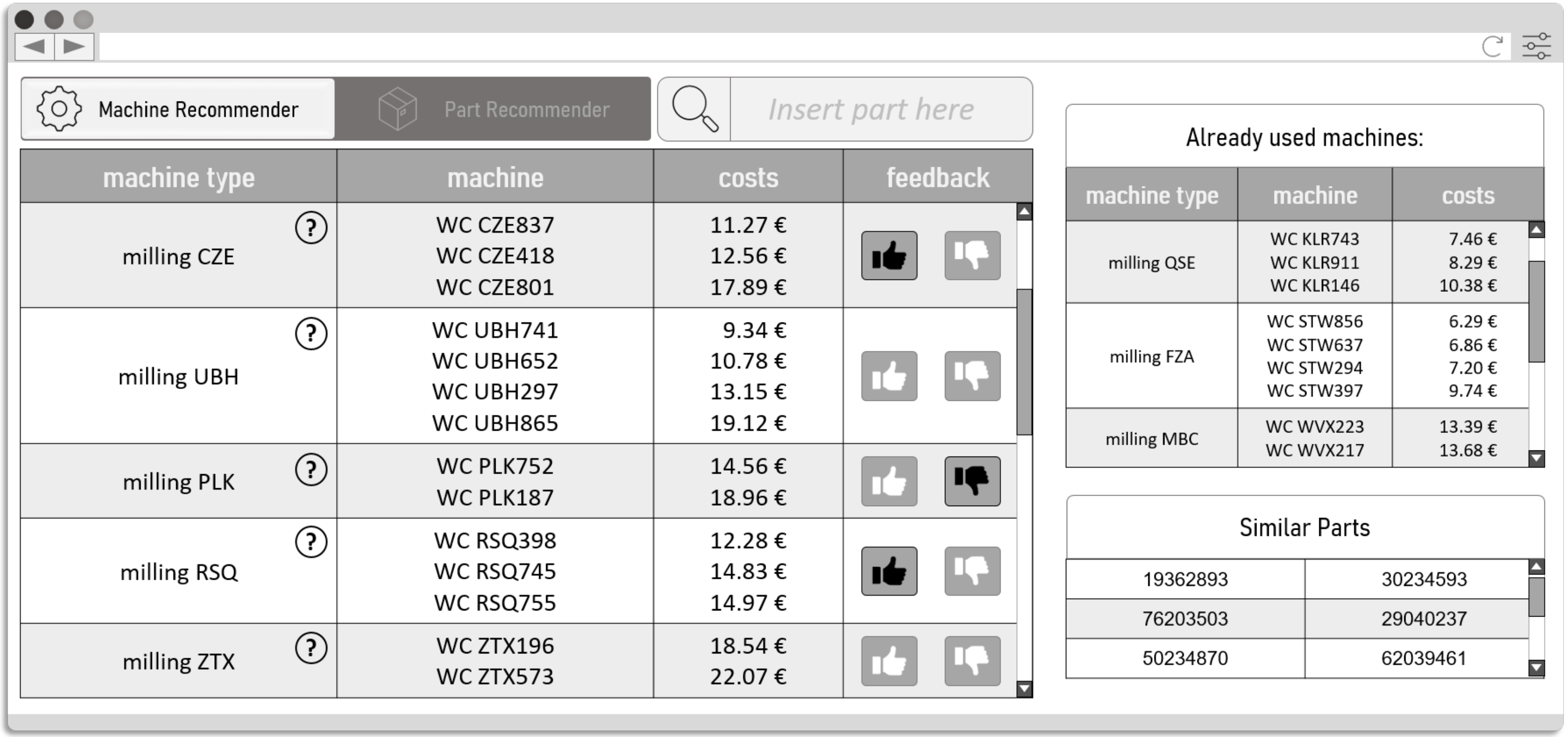
Table 6.4: Evaluation results of parts
with a certain number of known machine types

	Number of known machine types			
	1	2	3	4
MRR	0.4543	0.5468	0.5703	0.6231
Precision@5	0.1648	0.1971	0.2047	0.1906
Recall@5	0.5839	0.6409	0.6732	0.7336
F-Score@5	0.2571	0.3015	0.3139	0.3026

other data, such as machining axes, can only be brought into a standardized format with great effort. A look at the recommendations showed that they mostly seem to be good in principle but need clarification in detail (approx. 50% of the assessed recommendations). On the other hand, some recommendations would be possible per se but are foreseeable technically sub-optimal or uneconomical (approx. 35%). The latter can be attributed to the fact that the recommendations are based on historical data that represent grown structures. A relevant number of machine types that historically produce the parts have also been placed in this category. This is most probably caused by the aggregation of machines into their respective machine-types. Machines that are already set up for a particular mass production task can perform that task economically, while another machine of the same type may not be suitable for it in the short term. A minority of the recommendations have been classified as fundamentally unusable, either because they were technically impossible or very uneconomical (approx. 15%).

6.4 Discussion

This Chapter shows that RS can provide valuable decision support for CM in production planning – a scenario outside of traditional RS application, which commonly predicts customer preferences, e.g., in e-commerce. We are confident that RS are suitable for various matchmaking tasks in which data about historic matches exist, but preferences and personal familiarity of decision-makers do not have to be considered. However, our results are subject to several limitations. First, the RS is restricted due to the underlying data. We have not used any data about machines or parts and had no explicit feedback on machine-part combinations. Both kinds of data would most probably improve recommendations. Moreover, the ERP data represents historically grown structures, which may not represent the most suitable choice. Therefore, the recommendations might perpetuate inefficiencies from the past. Second, the RS has to deal with the use-case specific situation of a highly differentiated manufacturer that offers a wide variety of parts, of which most are produced in rather low numbers. At the same time, the manufacturer has an extensive machine park with many different types of machines, which leads to a high sparsity of data. We expect a similar system to achieve even better results for manufacturers with a less differentiated product range. Third, expert feedback on our recommendations is costly. Though we have gathered feedback from the engineers and prospective users on the historic and recommended machines, this feedback is restricted to a small sample of parts, and it was not possible to control for aspects such as inter-rater and intra-rater reliability.



machine type	machine	costs	feedback
milling CZE ?	WC CZE837 WC CZE418 WC CZE801	11.27 € 12.56 € 17.89 €	👍 👎
milling UBH ?	WC UBH741 WC UBH652 WC UBH297 WC UBH865	9.34 € 10.78 € 13.15 € 19.12 €	👍 👎
milling PLK ?	WC PLK752 WC PLK187	14.56 € 18.96 €	👍 👎
milling RSQ ?	WC RSQ398 WC RSQ745 WC RSQ755	12.28 € 14.83 € 14.97 €	👍 👎
milling ZTX ?	WC ZTX196 WC ZTX573	18.54 € 22.07 €	👍 👎

Already used machines:

machine type	machine	costs
milling QSE	WC KLR743 WC KLR911 WC KLR146	7.46 € 8.29 € 10.38 €
milling FZA	WC STW856 WC STW637 WC STW294 WC STW397	6.29 € 6.86 € 7.20 € 9.74 €
milling MBC	WC WVX223 WC WVX217	13.39 € 13.68 €

Similar Parts

19362893	30234593
76203503	29040237
50234870	62039461

Figure 6.4: A comprehensive Mock-Up of the prototype DSS. Recommended and historic machine types are shown separately. There are options included to show explanations and give feedback to each recommendation. Specific machines and respective costs are examples for additional information. Depiction taken from Badewitz, Stamer, et al. (2021).

Practically, the RS is founded on ERP data, which is available and accessible in almost every company. This enables easy and low-effort decision support for CM in the brownfield and represents an entry point for designing more complex services without the need to populate ontologies with the existing machine park. Yet, the prototype only solves the very limited task of finding capable machines, but much more is needed to fulfill the primary goal of a more flexible and responsive supply network. Two critical success factors have to be improved as the DSS is maturing: *system quality*, i.e., the technically oriented performance, and *adequate information supply*, i.e., "ensure [that] the right analytical information [is provided] to the right people in the right place at the right time" (Dinter, 2013, p. 1211). The latter is important as production decisions are not solely dependent on capability issues but also on further criteria, e.g., costs of production and costs of logistic network formation. To decide adequately, an engineer will require more information about machine types and machines, their location, and follow-up issues of his decisions. These have to be incorporated and accessible from within the DSS. It is necessary to take this into account when developing the presented approach into an operative used software. Regarding the system quality, our proposed solution will face two major problems:

First, we have to handle the cold start problem, i.e., making recommendations for parts that have never been produced before. In a typical GPN, several hundred new parts are produced each year, and new machines are installed regularly. Thus, it will be of vital importance to handle the cold-start problem for parts, which have no production history at all. A promising approach would be to recommend machine types for a new part by looking at the recommendations from similar parts. As we described earlier, we do not have the required metadata and knowledge to do so automatically. However, the engineers most probably know similar parts. The web application could contain a specific GUI, which asks the engineer to enter similar parts resulting in a list of machine types, which historically produced all similar parts or are believed to be capable of doing so by the recommender.

Second, we have to expect to run out of good recommendations as more and more capable machine types will have been used at least once at some time in the future. When every capable machine type was already used, it is impossible to give any good, 'new' recommendations. Thus, in the future, it will become necessary to abstain from separating recommendations of capable machines and listings of used machines and turn towards a recommendation of what is the best choice from all machines, including historically used ones. For this purpose, it would be worth leveraging explicit feedback on machine product allocations, which is already gathered in the prototype DSS. Further research also has to answer the questions, of how to best combine explicit *and* implicit RS in this task. Hybrid RS merge different approaches and are known to improve the performance (Burke, 2002; Haubner and Setzer, 2020) while relying on the same type of feedback. Consequently, evaluating multi-criteria RS, which combine feedback from several sources (Nilashi, Ibrahim, and Ithnin, 2014) appears to be a natural next step of research.

6.5 Conclusion

In this Chapter, we explored the question of how to solve the CM problem with a simple and low-effort approach given a large dataset of historical production decisions is available. Specifically, we asked whether and how well RS can be utilized to support engineers in this task. We have shown the needs to be considered in the design of the RS and how such a system can be built. The two main challenges in the application are (1) to deal with implicit instead of explicit data in both training and evaluation of the models, and (2) to handle the pronounced sparsity in the part's dimension, as a majority of parts were produced on merely one machine type in the past. With optimized parameters, the RS we implemented achieved an F-score of about two-thirds of the achievable optimum and an MRR of just below 0.5. Even for parts that have historically been manufactured on only one machine type, the RS provides reliable recommendations. Feedback from engineers who have been provided with the prototype for testing suggests that the recommendations are helpful but need manual review. Furthermore, it has been clarified that a good production decision requires more information than just whether a machine is capable of producing a part. Summarizing, RS are a suitable means to solve the CM task in brownfield settings and can be deployed to support engineers in their production decisions. Our results further indicate, that established manufacturing companies should take advantage of the opportunity presented by their large transactional data sets.

Chapter 7

Finale

> All you need are these: certainty of judgment in the present moment; action for the common good in the present moment; and an attitude of gratitude in the present moment for anything that comes your way.
>
> Marcus Aurelius

This dissertation examines the economic aspects of data exchange between companies and lies a focus on established companies in the manufacturing sector. It contributes from the strategic to the operational level and combines theoretical insights with practical application-oriented improvements. A strategic variety of data exchange modes were formulated to fit and support different situations in value networks, and best practices for the digital transformation of manufacturing companies and networks were elaborated. The challenges of data pricing mechanisms have been explored. In this context, the Data Provision Game serves as a framework for future research on the specific challenge of data quality incentivization and revenue sharing. Two pricing mechanisms to achieve a welfare-optimized solution were proposed. Lastly, a specific problem in value networks was solved with an approach founded on collaborative shared data. The methods applied to achieve these results showed a great variety from structured reviews, qualitative interviews, game-theoretic modeling, and mathematical analysis, to Machine Learning. In addition to insights, all of these approaches also bring limitations that open up spaces for further research in two fields: First, the scientific monitoring of practical implementations according to the managerial implications. Second, the further development of methods and solutions to overcome their recognized shortcomings.

Far from answering all open questions regarding data exchange in value networks, this dissertation serves as a piece of the puzzle and provides an impetus for the development of solutions. Thereby, it supports the emergence of shared data utilization between production companies and contributes to the academic examination of the data economy.

7.1 Research Conclusion

Concluding on the overarching research objective from the introduction, this dissertation provides valuable insight into novel approaches to data exchange in value networks. Thereby, it fosters the comprehension of the strategic dimension of becoming data-driven as a manufacturing company and supports the strategic decision making on the mode of data exchange. It highlights the challenges in mechanism design when it comes to the competitive exchange of data and shows its potential to solve incentivization issues regarding data quality along a vertical data value chain. Moreover, it also proposes a novel data-based solution to the capability matchmaking problem, a common issue in supply chain management. From a market perspective, this solution is a horizontal data value chain allowing for collaborative data exchange. In the following, I summarize the results from the research chapters that followed the introduction and foundation to the topic in Chapters 1 and 2.

Result 1 *Structured overview of the data strategies of incumbent companies and adesign of a framework of strategic alternatives for data sharing.*

In Chapter 3, data strategies were decomposed into the four core components of (i) *creating responsibilities for data*, (ii) *designing the technical architecture*, (iii) *gaining data and AI capabilities*, and (iv) *approaching data exchange*. Based on an investigation from three angles, the strategic approaches to these topics in the management of data were approached. The angles were literature, the practice of incumbent companies, and insights from domain experts. This ensures that an exhaustive view of recent developments and longer trends had been gathered. First, it carved out the significance of a data asset mindset, and according to professionalization in the management of data via a catalog and a dedicated executive, the Chief Data Officer (CDO). Second, it detected the trend from central to decentral data management, which is enabled by modern and novel data architectures such as the data lakehouse or the data mesh. Third, it pointed out the importance of the data workforce and three ways of increasing the responsive capabilities by upskilling programs, partnerships, or acquisitions. Fourth, it introduced a new framework to categorize how data is shared and motivated with findings from practice and statements of domain experts. The four strategic alternatives to sharing data are closed, collaborative, competitive, and open. While a closed strategy permits the exchange of data, an open strategy makes data freely available. Thus, those two mark opposite extremes. In between are collaborative and competitive data exchange. Both share access to data for something in return. Collaborative sharing emphasizes mutual sharing and the reuse of data. In collaborative sharing, the incentives to share have to be sought in the use case. Contrary, competitive sharing exchanges data against money or monetary benefits. In competitive sharing data can be bought. Thus, it is not necessary to find data assets of more or less the same value. In collaborative as well as competitive sharing, mechanisms must be designed to guarantee a smooth and beneficial exchange process and to avoid misbehavior such as the sharing of unusable data because of either lacking effort in collection and preprocessing or a deliberate corruption.

Result 2 *Overview of the state of the art in data pricing mechanisms and comprehension of challenges within their design.*

In Chapter 4, the state of the art in pricing data on the basis of a structured literature review containing 70 papers and reached two important goals was investigated. First, it described and classified pricing mechanisms based on their objective, market structure, data structure, and general approach. Secondly, the almost more important point, it pointed out what challenges exist and how the various researchers have tackled them. Moreover, it could recourse to the former classification to find recurring patterns. Our research thus provides a comprehensive overview of existing solutions to price data, but shows as well their shortcomings, i.e., which challenges are not addressed by the authors. Additionally, the analysis gives insights into the focal points of past research as well as blind spots for future investigation. Noticeable challenges were how to incorporate *quality aspects*, how to guarantee *arbitrage freeness*, how to avoid *negative information*, how to provide *privacy protection*, how to enable *fine granularity*, how to deal with *computational complexity*, and how to create *transparency*. We found that the *division of revenue* between many providers of data to an aggregate query, the *conflict of interest* between multiple buyers that are interested in exclusive data access, and the *prevention of fraud* are rising concerns in the design of pricing mechanisms. Further, it identified the effect of competing data providers and consumers as a future field of research.

Result 3 *Specification of a data value chain and design of a Data Provision Game including an incentive-compatible revenue-sharing mechanism.*

Chapter 5 examined incentivization issues in the exchange of data and suggests a viable approach to solve them via a competitive pricing mechanism. Modern value networks do not consist primarily of the physical flow of goods but to an equal extent of a digital counterpart. The data flow has to be organized along and between many stakeholders that are, at least to some extent, economically independent of each other and have to keep their own interests in mind. This leads inevitably to conflicts of interest and is one of the main reasons why applications that rely on shared data often fail to realize their full potential. This description, however, is not sufficient to understand and resolve the specific conflicts. Therefore, it furthermore deduced a version of the *data value chain* from literature, that emphasizes organizational responsibilities. This allows pinning down the relevant conflict among the involved parties. With a specific industrial case in mind, it has found that data accuracy is an important issue if the data consumer requires data of a higher quality than the data provider does. The latter has no reason to invest in more precise data collection unless the consumer compensates him for his increased expenses. The consumer however can only indirectly control whether the data fulfills his requirements. We suggest solving this incentivization problem with competitive data exchange, i.e., by designing a mechanism to price the transferred data based on its usefulness. The usefulness in turn depends on the accuracy of the data. This is to make the data provider invest in data quality to maximize its expected profit. We defined the *Data Provision Game*, a game-theoretic model which allows analyzing the situation and the decisions of the main stakeholders. We suggested applying Shapley values to determine the price and could prove that they – under further, quite realistic, model assumptions regarding cost, revenue, and data model – provide incentives for system-optimal data quality. Further, it showed that leave-one-out pricing also provides these incentives, but is computationally far less complex. However, it brings other downsides and requires a careful and thorough understanding of dependencies in order to avoid fraud. The practical contribution is on hand. With the necessary adaptions made the orchestrator of a value network can use the found theoretical result in order to deduct transfer prices of data between data providing and data using parties inside his network. The main academic contribution is the data provisioning game, which can serve as a general framework for studying the outcomes of data pricing mechanisms in the presence of incentive problems.

Result 4 *Proof of a novel method to support capability matchmaking in the brownfield via collaborative data exchange.*

Chapter 6 deals with the solution of a specific collaborative use case from supply chain management in value networks. The problem at hand is Capability Matchmaking (CM) in a Global Production Network (GPN). CM attempts to answer whether a particular production order for a part can be produced on a particular production machine, and in more advanced settings, which machines are better suited to perform a production order. CM is a difficult task and today is often performed by a skilled workforce. However, human labor is expensive, and the task can be greatly improved by computer-aided decision support systems. The current approach to developing such systems involves extensive modeling of the technical characteristics of machines and parts. This approach has several drawbacks, most notably the brown-field problem, i.e., the large effort required to model all existing machines and parts in a GPN, which is a substantial number considering that these systems are operated worldwide for decades. Our approach is consequently a collaborative data exchange. Based on historical production data of which production orders were produced for which parts, a Recommender System (RS) that suggests machines to the engineers who are still left as the last authority has been trained. A technical evaluation as well as a small study with the engineers showed the suitability and makes this approach an lightweight viable solution for the future. The big advantage is the low effort for decision support in the brownfield. The foundation data is already available in good quality since the tracking of production orders is part of ERP systems. In this case, collaborative data sharing is a perspective solution because there is no incentive problem with data quality - the data is already there - and everyone benefits from the solution. A free-rider problem can be ruled out because transferring more data improves system quality but does not increase costs. However, this assumes that there is no leakage of key performance indicators, such as the overall efficiency of the assets of a data sharing party.

7.2 Research Limitations

The dissertation has, as every research work, some limitations. These limitations are mainly attributable to three reasons. The applied methods have limitations in themselves, which were already known and carefully considered in the research design. During the research, earlier uncertain conditions about the framework of the research object became known, which pose further limitations on the general applicability of the research results. In completing the research, time and effort constraints had to be taken into account. Future work may set other priorities to address specific problems in more depth.

Regarding the data strategy of incumbents, the main idea was to investigate well documented publicly available company material and cross-check it with semi-structured, qualitative interviews with domain experts. This provides a good overview of the state of the art, which is why this method was chosen, but limits the results by being descriptive rather than normative. Moreover, it relies on public self-disclosure, which is likely to sketch a brighter picture of reality than a peer-reviewed assessment of the situation would bear. Unfortunately, it was found that most companies do not reveal the data strategy in their investor relations documents. Therefore, the analysis was based on more news articles and press releases than initially intended. These only allow for indirect conclusions to be drawn. However, they are less likely to suffer from the limitations of self-disclosure. The opinions of experts are likewise not necessarily objective and exhaustive. We have interviewed thirteen domain experts from eleven companies. That number was satisfying, but a higher rate of participation among requested experts would have been desirable.Due to the qualitative nature of interviews, it was possible to explore the topics in question, but it was not possible to draw quantitative figures that describe the situation statistically. Unfortunately, the domain experts also have a regional bias towards Europe, whereas the company sample is balanced between European and American companies.

The insights on state-of-the-art data pricing mechanisms were revealed in a structured literature review. That is a renowned method to comprise previous research results and has the advantage of being transparent and comprehensible. However, the won insights into the topic are merely from the realm of science. Practical evidence from existing real-world data marketplaces is only reflected if there existed a peer-reviewed paper in this marketplace, which was almost never the case. We have taken certain decisions in advance of the search process. This obviously concerns the search string, which was wide. More importantly, it restricted the research to mathematically described mechanisms, which excludes all forms of simple mechanisms, that can be described verbally, i.e., lump fees or subscriptions. In addition, it was found that the corpus of articles suitable for answering the research question is not published only or even mostly in high-impact journals. Therefore, articles of lower ranked outlets leading to high heterogeneity in the quality of reviewed articles had been included. At the same time, some of the most distinguishing papers on that issue are from small but unrenowned outlets. The conceptualization of the research and the classification of mechanisms to concepts were done by two researchers but not independently cross-checked. A special focus of the review was the challenges of pricing mechanisms. The possible challenges of a pricing mechanism were not given by an exhaustive list and seldom explicitly stated in a paper. Therefore, tackled but unmentioned challenges might have been overlooked.

The incentivization problem of data quality was addressed through a game-theoretic analysis. This in turn is based on a qualitative analysis of use in terms of a data value chain, which was deduced from the literature. The data value chain is therefore only reflecting the academic view on the topic possibly neglecting practice-relevant aspects. Further, it results in a high degree of abstraction. The Data Provision Game was developed with a certain industry case in mind, which might decrease its generalizability although it was intentionally designed to cover a broad range of similar situations. Moreover, the Data Provision Game is one-dimensional, meaning it only regards data quality issues and ignores certain other present challenges such as privacy and information security concerns. The suggested pricing mechanisms were analytically checked for conformance with desired properties, i.e., their welfare optimality. By choosing an analysis with rational agents, the mathematical properties of the mechanisms can be shown. However, their applicability in practice and with real decision-makers should be researched in the next step. Additionally, certain model assumptions, particularly linearity, limited the analysis and pose an opportunity to extend the work.

Solving the capability matchmaking problem was approached using RS, which is a novel way to address this problem, but common in general. A relevant barrier to the modeling and technical evaluation of the proposed solution concept was the missing of explicit feedback in part-machine combinations. Moreover, the technical evaluation is in itself directed backward and measures how well historical configurations can be suggested. These, however, might be an insufficient benchmark because they resemble a grown structure that is possibly far from optimal. The system was developed and evaluated on data from an industry partner. There might be individual effects due to the specific conditions in his value network. A data set from another company might have resulted in different figures. Three potential users tested the solution in the aftermath. However, this user study was too small to draw statistically significant conclusions. In addition, the solution was developed to work on a brownfield scenario for which no state-of-the-art approach was available as a benchmark. Therefore, the conclusions had to be drawn against a system with assumed 100% accuracy.

7.3 Research Outlook

Considering research objectives, results, and the limitations of the applied methods, opportunities for future research can be deduced. Placed in a middle ground between application-oriented and theoretical research, the opportunities point in both directions calling for a more thorough analytical comprehension in terms of mechanisms designed as well as for a more rigid practical evaluation of existing mechanisms.

In the investigation on data strategy, practitioners can be provided with a comprehensive overview of best practices in five core components of a data strategy. These results reproduce the state of the art in incumbent companies and the views of domain experts. Although the results are linked to the academic context, the results must be qualified against this background when used as recommendations for action. In order to provide the derived implications with a more solid footing and increased applicability two things have to happen. For one, the high-level implication has to be operationalized in specific action items and low-level capabilities. This is, for instance, the development of a decentralized technical architecture via a design cycle approach. On the other, the efficiency and effectiveness of implications have to be evaluated scientifically via concise studies, which compare organizations, which take different routes. This is, for example, the comparison of organizations with data dictionaries of varying depth and coverage in terms of their degree of data reuse.

Our literature review on data pricing mechanisms revealed many captivating insights into the current direction of research. First, it is apparent that most research focuses on a small subset of existing challenges. While one has to realize that every challenge is demanding by itself, it is necessary to consider more challenges simultaneously in the further design of mechanisms. Second, it was found that certain challenges, such as arbitrage freeness and differential privacy, have been well studied, while others, such as revenue division and conflict of interest were much less established. For the success of mechanisms, however, the latter is as important as the first. Last, it is notable that mechanisms are usually researched in a purely scientific context that abstracts from specific real-world applications or data markets. Therefore, a logical next step is to review data pricing mechanisms that already are applied today in the existing marketplaces. A compelling question to ground the research is whether these already account for the posed challenges, and if not, why. Moreover, there is nearly inexhaustible potential to design data exchange mechanisms for specific applications, which consider the peculiarities in data and market structures of the respective cases.

We researched a competitive pricing mechanism to target incentivization issues in a many-to-one setting but restricted ourselves to a special case. The peculiarities of this case are (i) a linear relationship between provided data and exploited information, (ii) a many-to-one setting, and (iii) a non-consideration of analytics service providers. Further work must think beyond these characteristics. It is currently an open question whether the results can be generalized to arbitrary relationships between predictors and target variables. Therefore, more generalized relationships have to be examined. The research has to be extended to cover more complex network situations, which includes especially multi-level data value chains, and the inclusion of analytics service providers. The practicability of proposed mechanisms needs to be investigated beyond the model scope. As is, the model reduces the reluctance to share to an accuracy-based cost factor that cost might incorporate the risk-taking of data providers but is not designed to. Further research has to verify whether non-cost-based obstacles to data sharing can be appropriately modeled in such a cost factor. Also, the mechanisms are only analyzed on an analytical level. The mechanisms have to be assessed when applied to real decision-makers. A simulation study and a subsequent field study on whether the price signals are understood by decision-makers are necessary to find out whether mechanisms would work in practice. In addition, further research should be undertaken in order to reduce the computational complexity of the proposed solution and to better fit Shapley values to the diversity of challenges in data pricing.

Last an RS-based solution to the supply chain problem of capability matchmaking based on collaborative data exchange was developed. While the research showed the general applicability and validity of the proposed solution, it is still a long road to its use in corporate practice. Further research and development are needed on technical as well as on organizational aspects. Technically, the most challenging upcoming task is to prioritize recommendations. As more and more possible combinations become known, it becomes more important to weigh between them than to find good new combinations. Therefore, the implicit recommender has to be combined with an explicit recommender. However, the hybridization of both methods is complicated due to the differences in data representations and interpretations of the given recommendations. Another challenge is to cover the cold start problem. New parts and machines are constantly added to the production network and must be handled by the recommendation system, which has never seen those before. Several ways are imaginable to onboard those new items into the RS, e.g., a part could be entered including similar parts or suitable machines. Organizationally, a more intense investigation of barriers to sharing the necessary ERP data across company borders has to be undertaken. It would be advisable to conduct expert interviews with data owners before implementing a cross-organizational prototype.

7.4 Final Remarks

Ultimately, this dissertation made its contribution to the evolving field of data economics. The proposed solutions are far from the be-all and end-all, but they are also more than just directional. The scientific publishing of parts of the work and of this dissertation as a whole will help future researchers to comprehend and analyze the peculiarities and difficulties of dealing with data in value networks. More than that, the presented results have already contributed to the design of data-driven collaborations among the industry partners I had the honor to work with. Further, this dissertation ideally has valuable implications for practitioners, who strive for the same. It is my reasonable hope that my work will serve as an inspiration to others and provide a small impetus for the development of data markets in practice and research.

Appendix A

Disclosure of Own Contributions

The research presented in this thesis emerged through discussion, exchange, and sometimes collaboration with other researchers; supervisors, colleagues, partners, and students. Their engagement and support of my research has eventually resulted in co-authorship for some of my papers. Their engagement reached from enabling to enhancing activities. Inter alia, they sparred ideas with me, they formulated problems, they provided data, they conducted and reviewed parts of the research. While this work would not have been possible without their contribution, I would like to point out my very own contribution to each of the papers in the following and indicate the changes I have made for the dissertation.

"The Data Provision Game: Researching Revenue Sharing in Collaborative Data Networks" is a joint paper with Dr. Simon Kloker, and Prof. Dr. Christof Weinhardt, published as a Full Paper in the Proceedings of the 22$^{\text{nd}}$ Conference on Business Informatics in 2020. My contributions consisted of:

- The conceptualization.
- The research design.
- The literature review.
- The creation of the Data Value Chain.
- The creation of the Data Provision Game.
- The proposal of the pricing mechanisms.
- The formal analysis.
- The visualization.
- The writing of all sections.

The changes included:

- Reworked first page (Abstract)

- Extension on related work
- Reworked Figures
- Revision of wording in several instances
- Minor changes

"Recommender Systems for Capability Matchmaking" is a joint Paper with Florian Stamer, Johannes Linzbach, Dr. Sebastian Lichtenberger, Dr. David Dann, and Prof. Dr. Christof Weinhardt, published as a Full Paper in the Proceedings of the 23rd Conference on Business Informatics in 2021. My contributions consisted of:

- The conceptualization.
- The research design.
- The review of the state of the art.
- The review of related work.
- The co-curation of research data and code.
- The development of the decision support system.
- The co-implementation of the decision support system.
- The formal evaluation of the decision support system.
- The discussion of limitations and implications.
- The visualization.
- The writing of all sections.

The changes included:

- Reworked first page (Abstract)
- Added paragraphs on the collaborative exchange of data
- Revision of wording in several instances
- Minor changes

"Challenges of pricing data assets: a literature review" is a joint paper with Christoph Hengesbach, and Prof. Dr. Christof Weinhardt, published as a Research Paper in the Proceedings of the 24th Conference on Business Informatics in 2022. The paper was inspired by Christoph Hengesbach's bachelor's thesis, which was written under my supervision, and built upon it. My contributions consisted of:

- The conceptualization.
- The research design.

- The review of related work.
- The co-creation of the literature corpus.
- The creation of concepts.
- The derivation of challenges.
- The discussion of limitations and implications.
- The visualization.
- The writing of all sections.

The changes included:

- Reworked first page (Abstract)
- Extension on challenges
- Revision of wording in several instances
- Minor changes

"Data Strategies of Industrial Incumbents: Worth a look into practice" is a joint paper with Carl-Philipp Wachter and Prof. Dr. Christof Weinhardt, currently a working paper. The paper was inspired by Carl-Philipp Wachter's master's thesis, which was written under my supervision, and built upon it. My contributions consisted of:

- The conceptualization.
- The research design.
- The review of related work.
- The co-conduction of the company study.
- The co-analysis of the interviews.
- The derivation of managerial implications.
- The development of the data exchange framework.
- The discussion of limitations and implications.
- The visualization.
- The writing of all sections.

The changes included:

- Reworked first page (Abstract)
- Extension on literature background
- Revision of wording in several instances
- Minor changes

Funding Note

The research of the two papers Badewitz, Kloker, and Weinhardt (2020) *"The Data Provision Game: Researching Revenue Sharing in Collaborative Data Networks"* and Badewitz, Stamer, et al. (2021) *"Recommender Systems for Capability Matchmaking"* were performed in scope of the ReKoNet project funded by the German Federal Ministry of Education and Research (BMBF) within the "Innovations for Tomorrow's Production, Services, and Work" Program and implemented by the Project Management Agency Karlsruhe (PTKA). The authors are responsible for the content of this publication.

Bibliography

Abbas, Antragama Ewa, Wirawan Agahari, Montijn van de Ven, Anneke Zuiderwijk, and Mark de Reuver (2021). “Business Data Sharing through Data Marketplaces: A Systematic Literature Review”. In: *Journal of Theoretical and Applied Electronic Commerce Research* 16.7, pp. 3321–3339. DOI: 10.3390/jtaer16070180.

Abele, Eberhard, Tobias Meyer, Ulrich Näher, Gernot Strube, and Richard Sykes (2008). *Global Production*. Ed. by Eberhard Abele, Tobias Meyer, Ulrich Näher, Gernot Strube, and Richard Sykes. Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-540-71653-2.

Agarwal, Anish, Munther Dahleh, and Tuhin Sarkar (2019). “A Marketplace for Data: An Algorithmic Solution”. In: *Proceedings of the 2019 ACM Conference on Economics and Computation*. Phoenix AZ USA: ACM, pp. 701–726. DOI: 10.1145/3328526.3329589.

Aggarwal, Charu C (2016). *Recommender Systems*. Springer International Publishing, Cham. DOI: 10.1007/978-3-319-29659-3.

Aguilar, Jose, Priscila Valdiviezo-Díaz, and Guido Riofrio (2017). “A general framework for intelligent recommender systems”. In: *Applied Computing and Informatics* 13.2, pp. 147–160. DOI: 10.1016/j.aci.2016.08.002.

Akerlof, George A (1978). “The market for “lemons”: Quality uncertainty and the market mechanism”. In: *Uncertainty in economics*. Elsevier, pp. 235–251. DOI: 10.1016/b978-0-12-214850-7.50022-x.

Akter, Shahriar and Samuel Fosso Wamba (2016). “Big data analytics in E-commerce: a systematic review and agenda for future research”. In: *Electronic Markets* 26.2, pp. 173–194. DOI: 10.1007/s12525-016-0219-0.

Alinani, Karim, Deshun Liu, Dong Zhou, and Guojun Wang (2019). “Recommender System for Decentralized Cloud Manufacturing”. In: *Communications in Computer and Information Science*. Vol. 1123 CCIS. Springer, pp. 170–179. DOI: 10.1007/978-981-15-1304-6_14.

An, Dou et al. (2017). “Towards Truthful Auction for Big Data Trading”. In: *2017 IEEE 36th International Performance Computing and Communications Conference (IPCCC)*. San Diego, CA: IEEE, pp. 1–7. DOI: 10.1109/PCCC.2017.8280501.

Arcondara, Jonathan, Khaled Himmi, Peiqing Guan, and Wei Zhou (2017). “Value oriented big data strategy: analysis & case study”. In: *Proceedings of the 50th Hawaii international conference on system sciences*. URL: `http://hdl.handle.net/10125/41277`.

Armbrust, Michael, Ali Ghodsi, Reynold Xin, and Matei Zaharia (2021). “Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics”. In: *Proceedings of the 11th Annual Conference on Innovative Data Systems Research (CIDR ’21)*.

Asswad, Jad and Jorge Marx Gómez (2021). “Data Ownership: A Survey”. In: *Information* 12.11. DOI: `10.3390/info12110465`.

Attard, Judie, Fabrizio Orlandi, and Sören Auer (2017). “Exploiting the Value of Data through Data Value Networks”. In: *Proceedings of the 10th International Conference on Theory and Practice of Electronic Governance*. Ed. by Rehema Baguma, Rahul Dé, and Tomasz Janowski. ICPS. New York New York: The Association for Computing Machinery, pp. 475–484. DOI: `10.1145/3047273.3047299`.

Bach, Norbert, Carsten Brehm, Wolfgang Buchholz, and Thorsten Petry (2012). *Wertschöpfungsorientierte Organisation*. Gabler Verlag. DOI: `10.1007/978-3-8349-3691-2`.

Badewitz, Wolfgang, Christoph Hengesbach, and Christof Weinhardt (2022). “Challenges of pricing data assets: a literature review”. In: *2022 IEEE 24th Conference on Business Informatics (CBI)*. Vol. 01, pp. 80–89. DOI: `10.1109/CBI54897.2022.00016`.

Badewitz, Wolfgang, Simon Kloker, and Christof Weinhardt (2020). “The Data Provision Game: Researching Revenue Sharing in Collaborative Data Networks”. In: *2020 IEEE 22nd Conference on Business Informatics (CBI)*. Vol. 01, pp. 191–200. DOI: `10.1109/CBI49978.2020.00028`.

Badewitz, Wolfgang, Florian Stamer, et al. (2021). “Recommender Systems for Capability Matchmaking”. In: *2021 IEEE 23rd Conference on Business Informatics (CBI)*. Vol. 02, pp. 87–96. DOI: `10.1109/CBI52690.2021.10059`.

Badewitz, Wolfgang, Carl-Philipp Wachter, and Christof Weinhardt (n.d.). “Data Strategies of Industrial Incumbents: Worth a look into practice”. In: *Research Paper in work*.

Baecker, Julius, Martin Engert, Matthias Pfaff, and Helmut Krcmar (2020). “Business Strategies for Data Monetization: Deriving Insights from Practice.” In: *Proceedings of the 15th International Conference on Wirtschaftsinformatik*, pp. 972–987.

Balazinska, Magdalena, Bill Howe, and Dan Suciu (2011). “Data Markets in the Cloud: An Opportunity for the Database Community”. In: *Proceedings of the VLDB Endowment* 4.12, pp. 1482–1485. DOI: `10.14778/3402755.3402801`.

Bataineh, Ahmed Saleh, Rabeb Mizouni, Jamal Bentahar, and May El Barachi (2020). “Toward Monetizing Personal Data: A Two-Sided Market Analysis”. In: *Future Generation Computer Systems* 111, pp. 435–459. DOI: 10.1016/j.future.2019.11.009.

Bergemann, Dirk and Alessandro Bonatti (2015). “Selling Cookies”. In: *American Economic Journal: Microeconomics* 7.3, pp. 259–294. DOI: 10.1257/mic.20140155.

— (2019). “Markets for information: An introduction”. In: *Annual Review of Economics* 11, pp. 85–107. DOI: 10.1146/annurev-economics-080315-015439.

Bildstein, Andreas, Junkang Feng, and Thomas Bauernhansl (2018). “Information Flow-based Capability Matching Service for Smart Manufacturing”. In: *Procedia CIRP*. Vol. 72. Elsevier BV, pp. 1015–1021. DOI: 10.1016/j.procir.2018.03.147.

Bobadilla, J., F. Ortega, A. Hernando, and A. Gutiérrez (2013). “Recommender systems survey”. In: *Knowledge-Based Systems* 46, pp. 109–132. DOI: 10.1016/j.knosys.2013.03.012.

Bohli, Jens-Matthias, Christoph Sorge, and Dirk Westhoff (2009). “Initial Observations on Economics, Pricing, and Penetration of the Internet of Things Market”. In: *SIGCOMM Comput. Commun. Rev.* 39.2, pp. 50–55. DOI: 10.1145/1517480.1517491.

Brackenbury, Will et al. (2018). “Draining the data swamp: A similarity-based approach”. In: *Proceedings of the workshop on human-in-the-loop data analytics*. ACM, pp. 1–7. DOI: 10.1145/3209900.3209911.

Braud, Arnaud, Gaël Fromentoux, Benoit Radier, and Olivier Le Grand (2021). “The Road to European Digital Sovereignty with Gaia-X and IDSA”. In: *IEEE Network* 35.2, pp. 4–5. DOI: 10.1109/MNET.2021.9387709.

Brocke, Jan vom et al. (2009). “Reconstructing the Giant: On the Importance of Rigour in Documenting the Literature Search Process”. In: *ECIS 2009 Proceedings*. URL: https://aisel.aisnet.org/ecis2009/161.

Broek, Tijs van den and Anne Fleur van Veenstra (2015). “Modes of Governance in Inter-Organizational Data Collaborations”. In: *ECIS 2015 Completed Research Papers*. DOI: 10.18151/7217509.

Bucherer, Eva and Dieter Uckelmann (2011). “Business Models for the Internet of Things”. In: *Architecting the Internet of Things*. Ed. by Dieter Uckelmann, Mark Harrison, and Florian Michahelles. Springer, Berlin, Heidelberg, pp. 253–277. DOI: 10.1007/978-3-642-19157-2_10.

Burke, Robin (2002). “Hybrid recommender systems: Survey and experiments”. In: *User Modelling and User-Adapted Interaction* 12.4, pp. 331–370. DOI: 10.1023/A:1021240730564.

Cai, Hui, Fan Ye, et al. (2022). “Online Pricing and Trading of Private Data in Correlated Queries”. In: *IEEE Transactions on Parallel and Distributed Systems* 33.3, pp. 569–585. DOI: 10.1109/TPDS.2021.3095238.

Cai, Hui, Yanmin Zhu, Jie Li, and Jiadi Yu (2019). “Double Auction for a Data Trading Market with Preferences and Conflicts of Interest”. In: *The Computer Journal* 62.10, pp. 1490–1504. DOI: 10.1093/comjnl/bxz025.

Cai, Yang, Constantinos Daskalakis, and Christos Papadimitriou (2015). “Optimum Statistical Estimation with Strategic Data Sources”. In: *Proceedings of the 28th Conference on Learning Theory*. Paris: PMLR, pp. 280–296. URL: https://proceedings.mlr.press/v40/Cai15.html.

Camerer, Colin (1999). “Behavioral economics: Reunifying psychology and economics”. In: *Proceedings of the National Academy of Sciences* 96.19, pp. 10575–10577.

Cao, Xuanyu, Yan Chen, and K. J. Ray Liu (2017). “Data Trading With Multiple Owners, Collectors, and Users: An Iterative Auction Mechanism”. In: *IEEE Transactions on Signal and Information Processing over Networks* 3.2, pp. 268–281. DOI: 10.1109/TSIPN.2017.2668144.

Chawla, Shuchi, Shaleen Deep, Paraschos Koutrisw, and Yifeng Teng (2019). “Revenue Maximization for Query Pricing”. In: *Proceedings of the VLDB Endowment*. Vol. 13. 1, pp. 1–14. DOI: 10.14778/3357377.3357378.

Chen, Deyan and Hong Zhao (2012). “Data Security and Privacy Protection Issues in Cloud Computing”. In: *International Conference on Computer Science and Electronics Engineering (ICCSEE), 2012*. Piscataway, NJ: IEEE, pp. 647–651. DOI: 10.1109/ICCSEE.2012.193.

Chen, Lingjiao, Paraschos Koutris, and Arun Kumar (2019). “Towards Model-based Pricing for Machine Learning in a Data Marketplace”. In: *Proceedings of the 2019 International Conference on Management of Data*. Amsterdam Netherlands: ACM, pp. 1535–1552. DOI: 10.1145/3299869.3300078.

Chen, Xiaoyu and Ran Jin (2020). “AdaPipe: A Recommender System for Adaptive Computation Pipelines in Cyber-Manufacturing Computation Services”. In: *IEEE Transactions on Industrial Informatics*, pp. 1–1. DOI: 10.1109/tii.2020.3035524.

Chen, Ying-Ju and Sridhar Seshadri (2007). “Product Development and Pricing Strategy for Information Goods Under Heterogeneous Outside Opportunities”. In: *Information Systems Research* 18.2, pp. 150–172. DOI: 10.1287/isre.1070.0119.

Chuang, I-Hsun, Shih-Hao Huang, Wei-Chu Chao, Jen-Sheng Tsai, and Yau-Hwang Kuo (2020). “TIDES: A Trust-Aware IoT Data Economic System With Blockchain-Enabled Multi-Access Edge Computing”. In: *IEEE Access* 8, pp. 85839–85855. DOI: 10.1109/ACCESS.2020.2991267.

Clauss, Thomas et al. (2021). “Organizational ambidexterity and competitive advantage: The role of strategic agility in the exploration-exploitation paradox”. In: *Journal of Innovation & Knowledge* 6.4, pp. 203–213. DOI: 10.1016/j.jik.2020.07.003.

Collins, Virginia and Joel Lanz (2019). “Managing Data as an Asset”. In: *The CPA Journal* 89.6, pp. 22–27. URL: https://www.proquest.com/scholarly-journals/managing-data-as-asset/docview/2239576675/se-2.

Crié, Dominique and Andrea Micheaux (2006). “From customer data to value: What is lacking in the information chain?” In: *Journal of Database Marketing & Customer Strategy Management* 13.4, pp. 282–299. DOI: 10.1057/palgrave.dbm.3240306.

Cukier, Kenneth and Viktor Mayer-Schoenberger (2013). “The rise of big data: How it’s changing the way we think about the world”. In: *Foreign Affairs* 92, p. 28.

Curry, Edward (2016). “The Big Data Value Chain: Definitions, Concepts, and Theoretical Approaches”. In: *New horizons for a data-driven economy*. Ed. by José María Cavanillas, Edward Curry, and Wolfgang Wahlster. Springer Open, Cham, pp. 29–37. DOI: 10.1007/978-3-319-21569-3_3.

Dadouchi, Camélia and Bruno Agard (2020). *Recommender systems as an agility enabler in supply chain management*. DOI: 10.1007/s10845-020-01619-5.

DalleMule, Leandro and Thomas H Davenport (2017). “What’s your data strategy”. In: *Harvard Business Review* 95.3, pp. 112–121.

Dandekar, Pranav, Nadia Fawaz, and Stratis Ioannidis (2014). “Privacy Auctions for Recommender Systems”. In: *ACM Trans. Econ. Comput.* 2.3, pp. 1–22. DOI: 10.1145/2629665.

Data Landscape (2020). http://datalandscape.eu//. Accessed: 2020-03-20.

Datar, Srikant M., Sarah Mehta, and Paul Hamilton (2020). *Applying Data Science and Analytics at P&G*. URL: https://www.hbs.edu/faculty/Pages/item.aspx?num=58380 (visited on 12/13/2022).

Deep, Shaleen and Paraschos Koutris (2017a). “QIRANA: A Framework for Scalable Query Pricing”. In: *Proceedings of the 2017 ACM International Conference on Management of Data*. Chicago Illinois USA: ACM, pp. 699–713. DOI: 10.1145/3035918.3064017.

— (2017b). “The Design of Arbitrage-Free Data Pricing Schemes”. In: *20th International Conference on Database Theory (ICDT 2017)*. Vol. 68. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum für Informatik. DOI: 10.4230/LIPICS.ICDT.2017.12.

Dinter, Barbara (2013). “Success factors for information logistics strategy — An empirical investigation”. In: *Decision Support Systems* 54.3, pp. 1207–1218. DOI: 10.1016/j.dss.2012.09.001.

Dobakhshari, Donya G., Na Li, and Vijay Gupta (2016). "An incentive-based approach to distributed estimation with strategic sensors". In: *2016 IEEE 55th Conference on Decision and Control (CDC)*. IEEE, pp. 6141–6146. DOI: 10.1109/CDC.2016.7799213.

Dwork, Cynthia (2008). "Differential Privacy: A Survey of Results". In: *Theory and Applications of Models of Computation. TAMC 2008. Lecture Notes in Computer Science*. Ed. by Manindra Agrawal, Dingzhu Du, Zhenhua Duan, and Angsheng Li. Vol. 4978. Springer, Berlin, Heidelberg, pp. 1–19. DOI: 10.1007/978-3-540-79228-4_1.

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith (2006). "Calibrating Noise to Sensitivity in Private Data Analysis". In: *Theory of Cryptography*. Ed. by Shai Halevi and Tal Rabin. Red. by David Hutchison et al. Vol. 3876. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 265–284. DOI: 10.1007/11681878_14.

Dwork, Cynthia and Aaron Roth (2013). "The Algorithmic Foundations of Differential Privacy". In: *Foundations and Trends in Theoretical Computer Science* 9.3-4, pp. 211–407. DOI: 10.1561/0400000042.

Earley, Seth (2017). "The Evolving Role of the CDO". In: *IT Professional* 19.1, pp. 64–69. DOI: 10.1109/MITP.2017.4.

European Commission, Directorate-General for Communications Networks, Content and Technology (2018). *COMMISSION STAFF WORKING DOCUMENT Guidance on sharing private sector data in the European data economy Accompanying the document Communication from the Commission to the European Parliament, the Council, the European economic and social Committee and the Committee of the Regions "Towards a common European data space"*. URL: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52018SC0125 (visited on 12/13/2022).

— (2020). *A European strategy for data*. URL: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52020DC0066 (visited on 12/13/2022).

Faltings, Boi and Goran Radanovic (2017). "Game Theory for Data Science: Eliciting Truthful Information". In: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 11.2, pp. 1–151. DOI: 10.2200/S00788ED1V01Y201707AIM035.

Farokhi, Farhad, Iman Shames, and Michael Cantoni (2018). "Optimal contract design for effort-averse sensors". In: *International Journal of Control*, pp. 1–8. DOI: 10.1080/00207179.2018.1486041.

Faroukhi, Abou Zakaria, Imane El Alaoui, Youssef Gahi, and Aouatif Amine (2020). "Big data monetization throughout Big Data Value Chain: a comprehensive review". In: *Journal of Big Data* 7.1. DOI: 10.1186/s40537-019-0281-5.

Fernandez, Raul Castro, Pranav Subramaniam, and Michael J. Franklin (2020). "Data Market Platforms: Trading Data Assets to Solve Data Problems". In: *Proceedings of the VLDB Endowment*. Vol. 13. 12. VLDB Endowment, pp. 1933–1947. DOI: 10.14778/3407790.3407800.

Fleckenstein, Mike and Lorraine Fellows (2018a). "Implementing a data strategy". In: *Modern data strategy*. Springer, pp. 35–54. URL: 10.1007/978-3-319-68993-7_6.

— (2018b). *Modern Data Strategy*. Springer International Publishing. DOI: 10.1007/978-3-319-68993-7.

Fleischer, Lisa K. and Yu-Han Lyu (2012). "Approximately Optimal Auctions for Selling Privacy When Costs Are Correlated with Data". In: *Proceedings of the 13th ACM Conference on Electronic Commerce - EC '12*. Valencia, Spain: ACM Press, p. 568. DOI: 10.1145/2229012.2229054.

Frederickson, Ben (2017). *Implicit*. URL: https://implicit.readthedocs.io/en/latest/ (visited on 04/07/2021).

Freitas, André and Edward Curry (2016). "Big Data Curation". In: *New horizons for a data-driven economy*. Ed. by José María Cavanillas, Edward Curry, and Wolfgang Wahlster. Springer Open, Cham, pp. 87–118. DOI: 10.1007/978-3-319-21569-3_6.

Gao, Guoju et al. (2020). "DPDT: A Differentially Private Crowd-Sensed Data Trading Mechanism". In: *IEEE Internet of Things Journal* 7.1, pp. 751–762. DOI: 10.1109/JIOT.2019.2944107.

Gartner Inc. (2017). *Turn Your Big Data into a Valued Corporate Asset*. URL: https://www.forbes.com/sites/gartnergroup/2017/11/13/turn-your-big-data-into-a-valued-corporate-asset/ (visited on 12/13/2022).

Ghorbani, Amirata and James Zou (2019). "Data shapley: Equitable valuation of data for machine learning". In: *International Conference on Machine Learning*. PMLR, pp. 2242–2251.

Ghosh, Arpita and Aaron Roth (2011). "Selling Privacy at Auction". In: *Proceedings of the 12th ACM Conference on Electronic Commerce - EC '11*. San Jose, California, USA: ACM Press, p. 199. DOI: 10.1145/1993574.1993605.

Giannoccaro, Ilaria and Pierpaolo Pontrandolfo (2004). "Supply chain coordination by revenue sharing contracts". In: *International Journal of Production Economics* 89.2, pp. 131–139. DOI: 10.1016/S0925-5273(03)00047-1.

Gillon, Kirstin, Sinan Aral, Ching-Yung Lin, Sunil Mithas, and Mark Zozulia (2014). "Business Analytics: Radical Shift or Incremental Change?" In: *Communications of the Association for Information Systems* 34.1. DOI: 10.17705/1CAIS.03413.

Gimpel, Henner, Nicholas R Jennings, Gregory E Kersten, Axel Ockenfels, and Christof Weinhardt (2008). "Market engineering: A research agenda". In:

Negotiation, auctions, and market engineering, pp. 1–15. DOI: 10.1007/978-3-540-77554-6_1.

Gkatzelis, Vasilis, Christina Aperjis, and Bernardo A. Huberman (2015). “Pricing Private Data”. In: *Electron Markets* 25.2, pp. 109–123. DOI: 10.1007/s12525-015-0188-8.

Glennon, Mike et al. (2022). *European data market study 2021 - 2023 : d2.1 first report on facts and figures*. European Commission.

Goebel, Randy et al. (2018). “Explainable AI: The new 42?” In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 11015 LNCS. Springer Verlag, pp. 295–303. DOI: 10.1007/978-3-319-99740-7_21.

Goldberg, Andrew V., Jason D. Hartline, Anna R. Karlin, Michael Saks, and Andrew Wright (2006). “Competitive Auctions”. In: *Games and Economic Behavior* 55.2, pp. 242–269. DOI: 10.1016/j.geb.2006.02.003.

Goncalves, Carla, Pierre Pinson, and Ricardo J. Bessa (2021). “Towards Data Markets in Renewable Energy Forecasting”. In: *IEEE Transactions on Sustainable Energy* 12.1, pp. 533–542. DOI: 10.1109/TSTE.2020.3009615.

Gupta, Pooja, Volkan Dedeoglu, Kamran Najeebullah, Salil S. Kanhere, and Raja Jurdak (2020). “Energy-Aware Demand Selection and Allocation for Real-time IoT Data Trading”. In: *2020 IEEE International Conference on Smart Computing (SMARTCOMP)*. Bologna, Italy: IEEE, pp. 138–147. DOI: 10.1109/SMARTCOMP50058.2020.00038.

Gür, Inan, Markus Spiekermann, Michael Arbter, and Boris Otto (2021). “Data Strategy Development: A Taxonomy for Data Strategy Tools and Methodologies in the Economy”. In: *Innovation Through Information Systems. WI 2021. Lecture Notes in Information Systems and Organisation*. Vol. 46. Springer International Publishing, Cham, pp. 448–461. DOI: 10.1007/978-3-030-86790-4_30.

Hagiu, Andrei and Julian Wright (2020). “When data creates competitive advantage”. In: *Harvard Business Review* 98.1, pp. 94–101.

Hai, Rihan, Sandra Geisler, and Christoph Quix (2016). “Constance: An intelligent data lake system”. In: *Proceedings of the 2016 international conference on management of data*, pp. 2097–2100. DOI: 10.1145/2882903.2899389.

Hartline, Jason D. and Anna R. Karlin (2007). “Profit Maximization in Mechanism Design”. In: *Algorithmic game theory*. Ed. by Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V Vazirani. Cambridge, England: Cambridge University Press. ISBN: 978-0-521-87282-9.

Hartmann, Philipp Max, Mohamed Zaki, Niels Feldmann, and Andy Neely (2016). “Capturing value from big data–a taxonomy of data-driven business models used by start-up firms”. In: *International Journal of Operations & Production Management* 36.10. DOI: 10.1108/ijopm-02-2014-0098.

Haubner, Nicolas and Thomas Setzer (2020). “Applying Optimal Weight Combination in Hybrid Recommender Systems”. In: *Proceedings of the 53rd Hawaii International Conference on System Sciences*. Hawaii International Conference on System Sciences. DOI: 10.24251/hicss.2020.191.

Hax, Arnoldo C. (1990). “Redefining the concept of strategy and the strategy formation process”. In: *Planning Review* 18.3, pp. 34–39. DOI: 10.1108/eb054290.

Hermann, Mario, Tobias Pentek, and Boris Otto (2016). “Design Principles for Industrie 4.0 Scenarios”. In: *2016 49th Hawaii International Conference on System Sciences (HICSS)*. IEEE, pp. 3928–3937. DOI: 10.1109/HICSS.2016.488.

Hess, Charlotte and Elinor Ostrom (2003). “Ideas, artifacts, and facilities: information as a common-pool resource”. In: *Law and contemporary problems* 66.1/2, pp. 111–145. URL: https://www.jstor.org/stable/20059174.

Himmelhuber, Anna, Stephan Grimm, Thomas Runkler, and Sonja Zillner (2020). “Ontology-Based Skill Description Learning for Flexible Production Systems”. In: *2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*. IEEE, pp. 975–981. DOI: 10.1109/ETFA46521.2020.9211906.

Hoang, Xuan Luu, Constantin Hildebrandt, and Alexander Fay (2018). “Product-oriented description of manufacturing resource skills”. In: *IFAC-PapersOnLine* 51.11, pp. 90–95. DOI: 10.1016/j.ifacol.2018.08.240.

Hofmann, Erik and Marco Rüsch (2017). “Industry 4.0 and the current status as well as future prospects on logistics”. In: *Computers in Industry* 89, pp. 23–34. DOI: 10.1016/j.compind.2017.04.002.

Holsapple, C W and M Singh (2001). “The knowledge chain model: activities for competitiveness”. In: *Expert Systems with Applications* 20.1, pp. 77–98. DOI: 10.1016/S0957-4174(00)00050-6.

Hu, Han, Yonggang Wen, Tat-Seng Chua, and Xuelong Li (2014). “Toward Scalable Systems for Big Data Analytics: A Technology Tutorial”. In: *IEEE Access* 2, pp. 652–687. DOI: 10.1109/ACCESS.2014.2332453.

Hu, Yifan, Chris Volinsky, and Yehuda Koren (2008). “Collaborative filtering for implicit feedback datasets”. In: *Proceedings - IEEE International Conference on Data Mining, ICDM*, pp. 263–272. DOI: 10.1109/ICDM.2008.22.

Huang, Yu, Mostafa Milani, and Fei Chiang (2020). “Privacy-Aware Data Cleaning-as-a-Service”. In: *Information Systems* 94, p. 101608. DOI: 10.1016/j.is.2020.101608.

Hummel, Patrik, Matthias Braun, and Peter Dabrock (2021). “Own data? Ethical reflections on data ownership”. In: *Philosophy & Technology* 34.3, pp. 545–572. DOI: 10.1007/s13347-020-00404-9.

Hummel, Patrik, Matthias Braun, Max Tretter, and Peter Dabrock (2021). “Data sovereignty: A review”. In: *Big Data & Society* 8.1. DOI: 10.1177/2053951720982012.

Hurwicz, Leonid and Stanley Reiter (2006). *Designing Economic Mechanisms*. Cambridge University Press. DOI: 10.1017/cbo9780511754258.

International Data Spaces Association (2022). *Data spaces overview: Let’s build the future data economy!* URL: https://internationaldataspaces.org//wp-content/uploads/220513_Use-Case-Brochure.pdf (visited on 12/13/2022).

Järvenpää, Eeva, Niko Siltala, Otto Hylli, and Minna Lanz (2017). “Capability Matchmaking Procedure to Support Rapid Configuration and Re-configuration of Production Systems”. In: *Procedia Manufacturing* 11, pp. 1053–1060. DOI: 10.1016/j.promfg.2017.07.216.

Jia, Ruoxi et al. (2019). “Towards Efficient Data Valuation Based on the Shapley Value”. In: *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*. Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, pp. 1167–1176. URL: https://proceedings.mlr.press/v89/jia19a.html.

Jiao, Yutao, Ping Wang, Shaohan Feng, and Dusit Niyato (2018). “Profit Maximization Mechanism and Data Management for Data Analytics Services”. In: *IEEE Internet Things J.* 5.3, pp. 2001–2014. DOI: 10.1109/JIOT.2018.2819706.

Johnson, Christopher C (2014). “Logistic Matrix Factorization for Implicit Feedback Data”. In: *Advances in Neural Information Processing Systems* 27.78.

Jorgensen, Zach, Ting Yu, and Graham Cormode (2015). “Conservative or Liberal? Personalized Differential Privacy”. In: *2015 IEEE 31st International Conference on Data Engineering*. Seoul, South Korea: IEEE, pp. 1023–1034. DOI: 10.1109/ICDE.2015.7113353.

Jung, Kangsoo, Junkyu Lee, Kunyoung Park, and Seog Park (2019). “PRIVATA: Differentially Private Data Market Framework Using Negotiation-based Pricing Mechanism”. In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. Beijing China: ACM, pp. 2897–2900. DOI: 10.1145/3357384.3357855.

Jurksiene, Lolita and Asta Pundziene (2016). “The relationship between dynamic capabilities and firm competitive advantage”. In: *European Business Review* 28.4, pp. 431–448. DOI: 10.1108/ebr-09-2015-0088.

Kaiser, Robert (2014). *Qualitative Experteninterviews*. Springer Fachmedien Wiesbaden. DOI: 10.1007/978-3-658-02479-6.

Kamilaris, Andreas, Andreas Kartakoullis, and Francesc X. Prenafeta-Boldú (2017). “A review on the practice of big data analysis in agriculture”. In: *Computers and Electronics in Agriculture* 143, pp. 23–37. DOI: 10.1016/j.compag.2017.09.037.

Kaplinsky, Raphael and Mike Morris (2000). *A handbook for value chain research.* Vol. 113. Brighton: University of Sussex, Institute of Development Studies.

Kasim, Henry, Terence Hung, and Xiaorong Li (2012). "Data Value Chain as a Service Framework: For Enabling Data Handling, Data Security and Data Analysis in the Cloud". In: *2012 IEEE 18th International Conference on Parallel and Distributed Systems*. IEEE. DOI: 10.1109/icpads.2012.131.

Khakifirooz, Marzieh, Mahdi Fathi, and Kan Wu (2019). "Development of Smart Semiconductor Manufacturing: Operations Research and Data Science Perspectives". In: *IEEE Access* 7, pp. 108419–108430. DOI: 10.1109/ACCESS.2019.2933167.

Kitchin, Rob and Gavin McArdle (2016). "What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets". In: *Big Data & Society* 3.1. DOI: 10.1177/2053951716631130.

Köcher, Aljosha, Constantin Hildebrandt, Luis Miguel Vieira da Silva, and Alexander Fay (2020). "A Formal Capability and Skill Model for Use in Plug and Produce Scenarios". In: *2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*. IEEE, pp. 1663–1670. DOI: 10.1109/ETFA46521.2020.9211874.

Kolmar, Martin (2021). "Externalities and the Limits of Markets". In: *Principles of Microeconomics*. Springer International Publishing, pp. 103–153. DOI: 10.1007/978-3-030-78167-5_6.

Koutris, Paraschos, Prasang Upadhyaya, Magdalena Balazinska, Bill Howe, and Dan Suciu (2015). "Query-Based Data Pricing". In: *Journal of the ACM* 62.5, pp. 1–44. DOI: 10.1145/2770870.

Koutroumpis, Pantelis, Aija Leiponen, and Llewellyn D W Thomas (2020). "Markets for data". In: *Industrial and Corporate Change* 29.3, pp. 645–660. DOI: 10.1093/icc/dtaa002.

Koutsopoulos, Iordanis, Aristides Gionis, and Maria Halkidi (2015). "Auctioning Data for Learning". In: *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. Atlantic City, NJ, USA: IEEE, pp. 706–713. DOI: 10.1109/ICDMW.2015.175.

Lanza, Gisela et al. (2019). "Global production networks: Design and operation". In: *CIRP Annals* 68.2, pp. 823–841. DOI: 10.1016/j.cirp.2019.05.008.

Lasi, Heiner, Peter Fettke, Hans-Georg Kemper, Thomas Feld, and Michael Hoffmann (2014). "Industry 4.0". In: *Business & Information Systems Engineering* 6.4, pp. 239–242. DOI: 10.1007/s12599-014-0334-4.

Latif, Atif, Anwar Us Saeed, Patrick Hoefler, Alexander Stocker, and Claudia Wagner (2009). "The Linked Data Value Chain: A Lightweight Model for Business Engineers". In: *I-SEMANTICS*, pp. 568–575.

Lee, Jay, Hung-An Kao, and Shanhu Yang (2014). “Service Innovation and Smart Analytics for Industry 4.0 and Big Data Environment”. In: *Procedia CIRP* 16, pp. 3–8. DOI: 10.1016/j.procir.2014.02.001.

Lehmann, Jos, Michael Shamiyeh, and Sven Ziemer (2019). “Challenges of Modeling and Evaluating the Semantics of Technical Content Deployed in Recommendation Systems for Industry 4.0”. In: *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, pp. 359–366. DOI: 10.5220/0008348503590366.

Leimeister, Jan Marco (2015). “Führungsaufgaben des IT-Managements”. In: *Einführung in die Wirtschaftsinformatik*. Springer, Berlin, Heidelberg, pp. 179–229. ISBN: 978-3-540-77847-9. DOI: 10.1007/978-3-540-77847-9_4.

Li, Chao, Daniel Yang Li, Gerome Miklau, and Dan Suciu (2014). “A Theory of Pricing Private Data”. In: vol. 39. 4. ACM, pp. 1–28. DOI: 10.1145/2691190.2691191.

Li, Chao and Gerome Miklau (2012). “Pricing Aggregate Queries in a Data Marketplace”. In: *WebDB*.

Li, Qinya et al. (2021). “Capitalize Your Data: Optimal Selling Mechanisms for IoT Data Exchange”. In: *IEEE Transactions on Mobile Computing*, pp. 1–1. DOI: 10.1109/TMC.2021.3113387.

Li, Xijun, Jianguo Yao, Xue Liu, and Haibing Guan (2017). “A First Look at Information Entropy-Based Data Pricing”. In: *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*. Atlanta, GA, USA: IEEE, pp. 2053–2060. DOI: 10.1109/ICDCS.2017.45.

Liang, Fan et al. (2018). “A Survey on Big Data Market: Pricing, Trading and Protection”. In: *IEEE Access* 6, pp. 15132–15154. DOI: 10.1109/ACCESS.2018.2806881.

Lim, Chiehyeon et al. (2018). “From data to value: A nine-factor framework for data-based value creation in information-intensive services”. In: *International Journal of Information Management* 39, pp. 121–135. DOI: 10.1016/j.ijinfomgt.2017.12.007.

Lin, Bing-Rong and Daniel Kifer (2014). “On Arbitrage-Free Pricing for General Data Queries”. In: *Proceedings of the VLDB Endowment* 7.9, pp. 757–768. DOI: 10.14778/2732939.2732948.

Liu, Jinfei et al. (2021). “Dealer: An End-to-End Model Marketplace with Differential Privacy”. In: *Proceedings of the VLDB Endowment* 14.6, pp. 957–969. DOI: 10.14778/3447689.3447700.

Liu, Kang, Xiaoyu Qiu, Wuhui Chen, Xu Chen, and Zibin Zheng (2019). “Optimal Pricing Mechanism for Data Market in Blockchain-Enhanced Internet of Things”. In: *IEEE Internet of Things Journal* 6.6, pp. 9748–9761. DOI: 10.1109/JIOT.2019.2931370.

Liu, Zhengchao, Lei Wang, Xixing Li, and Shibao Pang (2021). “A multi-attribute personalized recommendation method for manufacturing service composition with combining collaborative filtering and genetic algorithm”. In: *Journal of Manufacturing Systems* 58, pp. 348–364. DOI: 10.1016/j.jmsy.2020.12.019.

Lu, Jie, Dianshuang Wu, Mingsong Mao, Wei Wang, and Guangquan Zhang (2015). “Recommender system application developments: A survey”. In: *Decision Support Systems* 74, pp. 12–32. DOI: 10.1016/j.dss.2015.03.008.

Lü, Linyuan et al. (2012). “Recommender systems”. In: *Physics Reports* 519.1, pp. 1–49. DOI: 10.1016/j.physrep.2012.02.006.

Luban, Daniel (2019). “What Is Spontaneous Order?” In: *American Political Science Review* 114.1, pp. 68–80. DOI: 10.1017/s0003055419000625.

Lundberg, Scott M and Su-In Lee (2017). “A Unified Approach to Interpreting Model Predictions”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc.

Luong, Nguyen Cong et al. (2016). “Data Collection and Wireless Communication in Internet of Things (IoT) Using Economic Analysis and Pricing Models: A Survey”. In: *IEEE Commun. Surv. Tutorials* 18.4, pp. 2546–2590. DOI: 10.1109/COMST.2016.2582841.

Machado, Inês Araújo, Carlos Costa, and Maribel Yasmina Santos (2022). “Data Mesh: Concepts and Principles of a Paradigm Shift in Data Architectures”. In: *Procedia Computer Science* 196, pp. 263–271. DOI: 10.1016/j.procs.2021.12.013.

Mao, Weichao, Zhenzhe Zheng, and Fan Wu (2019). “Pricing for Revenue Maximization in IoT Data Markets: An Information Design Perspective”. In: *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*. Paris, France: IEEE, pp. 1837–1845. DOI: 10.1109/INFOCOM.2019.8737571.

Marr, Bernard (2021). *Data Strategy : How to Profit From a World of Big Data, Analytics and Artificial Intelligence*. Kogan Page. ISBN: 9781398602588.

Medeiros, Mauricius Munhoz de, Antonio Carlos Gastaud Maçada, and José Carlos da Silva Freitas Junior (2020). “The effect of data strategy on competitive advantage”. In: *The Bottom Line* 33.2, pp. 201–216. DOI: 10.1108/bl-12-2019-0131.

Mehta, Sameer, Milind Dawande, Ganesh Janakiraman, and Vijay Mookerjee (2019). “How to Sell a Dataset? Pricing Policies for Data Monetization”. In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.3333296.

Miller, H Gilbert and Peter Mork (2013). “From Data to Decisions: A Value Chain for Big Data”. In: *IT Professional* 15.1, pp. 57–59. DOI: 10.1109/MITP.2013.11.

Mittal, Sameer, Muztoba Ahmad Khan, David Romero, and Thorsten Wuest (2017). “Smart manufacturing: Characteristics, technologies and enabling factors”. In:

Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture 233.5, pp. 1342–1361. DOI: 10.1177/0954405417736547.

Moody, Daniel L and Peter Walsh (1999). "Measuring the Value of Information-an Asset Valuation Approach." In: *European Conference on Information Systems (ECIS '99)*, pp. 496–512.

Myers, Michael D (2019). *Qualitative research in business and management*. Sage. ISBN: 978-1-4739-1232-8.

Myerson, Roger B. (1981). "Optimal Auction Design". In: *Mathematics of OR* 6.1, pp. 58–73. DOI: 10.1287/moor.6.1.58.

Nagel, Lars and Douwe Lycklama (2021). *Design Principles for Data Spaces - Position Paper*. Version 1.0. DOI: 10.5281/zenodo.5105744.

Najjar, Mohammad S and William J Kettinger (2013). "Data Monetization: Lessons from a Retailer's Journey." In: *MIS Quarterly Executive* 12.4. URL: https://aisel.aisnet.org/misqe/vol12/iss4/4.

Narahari, Yadati (2014). *Game theory and mechanism design*. IISc lecture notes series ; v. 4. Singapore: World Scientific Pub. Co. ISBN: 978-981-4525-04-6.

Naumann, Felix (2002). "Completeness-Driven Query Optimization". In: *Quality-Driven Query Answering for Integrated Information Systems. Lecture Notes in Computer Science*. Ed. by Felix Naumann. Vol. 2261. Springer, Berlin, Heidelberg, pp. 123–149. DOI: 10.1007/3-540-45921-9_8.

Nget, Rachana, Yang Cao, and Masatoshi Yoshikawa (2017). "How to Balance Privacy and Money through Pricing Mechanism in Personal Data Market". In: *Proceedings of the SIGIR 2017 Workshop on eCommerce*. Tokyo, Japan: ACM. URL: http://ceur-ws.org/Vol-2311/paper_15.pdf.

Nie, Yu, John Talburt, Xinming Li, and Zhongdong Xiao (2018). "Chief data officer (CDO) role and responsibility analysis". In: *Journal of Computing Sciences in Colleges* 33.5, pp. 4–12. ISSN: 1937-4771.

Nilashi, Mehrbakhsh, Othman Bin Ibrahim, and Norafida Ithnin (2014). "Multi-criteria collaborative filtering with high accuracy using higher order singular value decomposition and Neuro-Fuzzy system". In: *Knowledge-Based Systems* 60, pp. 82–101. DOI: 10.1016/j.knosys.2014.01.006.

Niu, Chaoyue, Zhenzhe Zheng, Shaojie Tang, Xiaofeng Gao, and Fan Wu (2019). "Making Big Money from Small Sensors: Trading Time-Series Data under Pufferfish Privacy". In: *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*. Paris, France: IEEE, pp. 568–576. DOI: 10.1109/INFOCOM.2019.8737579.

Niu, Chaoyue, Zhenzhe Zheng, Fan Wu, Shaojie Tang, and Guihai Chen (2020). "Online Pricing with Reserve Price Constraint for Personal Data Markets". In:

2020 IEEE 36th International Conference on Data Engineering (ICDE). Dallas, TX, USA: IEEE, pp. 1978–1981. DOI: 10.1109/ICDE48307.2020.00218.

Niyato, Dusit, Mohammad Abu Alsheikh, Ping Wang, Dong In Kim, and Zhu Han (2016). "Market Model and Optimal Pricing Scheme of Big Data and Internet of Things (IoT)". In: *2016 IEEE International Conference on Communications (ICC)*. Kuala Lumpur, Malaysia: IEEE, pp. 1–6. DOI: 10.1109/ICC.2016.7510922.

Niyato, Dusit, Dinh Thai Hoang, et al. (2016). "Smart Data Pricing Models for the Internet of Things: A Bundling Strategy Approach". In: *IEEE Network* 30.2, pp. 18–25. DOI: 10.1109/MNET.2016.7437020.

Niyato, Dusit, Xiao Lu, Ping Wang, Dong In Kim, and Zhu Han (2016). "Economics of Internet of Things: An information market approach". In: *IEEE Wireless Communications* 23.4, pp. 136–145. DOI: 10.1109/MWC.2016.7553037.

O'Cass, Aron, Nima Heirati, and Liem Viet Ngo (2014). "Achieving new product success via the synchronization of exploration and exploitation across multiple levels and functional areas". In: *Industrial Marketing Management* 43.5, pp. 862–872. DOI: 10.1016/j.indmarman.2014.04.015.

O'Donovan, Peter, Kevin Leahy, Ken Bruton, and Dominic T. J. O'Sullivan (2015). "Big data in manufacturing: a systematic mapping study". In: *Journal of Big Data* 2.1. DOI: 10.1186/s40537-015-0028-x.

Oard, Douglas W and Jinmook Kim (1998). "Implicit Feedback for Recommender Systems". In: *Proceedings of the AAAI workshop on recommender systems*, pp. 81–83.

Oestreicher-Singer and Sundararajan (2012). "Recommendation Networks and the Long Tail of Electronic Commerce". In: *MIS Quarterly* 36.1, pp. 65–83. DOI: 10.2307/41410406.

Oreščanin, Dražen and Tomislav Hlupić (2021). "Data Lakehouse - a Novel Step in Analytics Architecture". In: *2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO)*, pp. 1242–1246. DOI: 10.23919/MIPRO52101.2021.9597091.

Ostrom, Vincent and Elinor Ostrom (2019). "Public goods and public choices". In: *Alternatives for delivering public services*. Routledge, pp. 7–49. ISBN: 9780429047978.

Otto, Boris, Sebastian Steinbuß, Andreas Teuscher, and Steffen Lohmann (2019). *Reference Architecture Model Version 3.0*. Ed. by Sebastian Steinbuß. International Data Spaces Association. URL: https://internationaldataspaces.org/wp-content/uploads/IDS-Reference-Architecture-Model-3.0-2019.pdf (visited on 12/13/2022).

Park, YoungKi, Paul A. Pavlou, and Nilesh Saraf (2020). "Configurations for Achieving Organizational Ambidexterity with Digitization". In: *Information Systems Research* 31.4, pp. 1376–1397. DOI: 10.1287/isre.2020.0950.

Pei, Jian (2022). “A Survey on Data Pricing: From Economics to Data Science”. In: *IEEE Transactions on Knowledge and Data Engineering* 34.10, pp. 4586–4608. DOI: 10.1109/TKDE.2020.3045927.

Perrons, Robert K. and Jesse W. Jensen (2015). “Data as an asset: What the oil and gas sector can learn from other industries about "Big Data"”. In: *Energy Policy* 81, pp. 117–121. DOI: 10.1016/j.enpol.2015.02.020.

Perzylo, Alexander et al. (2019). “Capability-based semantic interoperability of manufacturing resources: A BaSys 4.0 perspective”. In: *IFAC-PapersOnLine* 52.13, pp. 1590–1596. DOI: 10.1016/j.ifacol.2019.11.427.

Pflaum, Alexander A and Philipp Gölzer (2018). “The IoT and Digital Transformation: Toward the Data-Driven Enterprise”. In: *IEEE Pervasive Computing* 17.1, pp. 87–91. DOI: 10.1109/MPRV.2018.011591066.

Poeppelbuss, Jens and Carolin Durst (2019). “Smart service canvas - A tool for analyzing and designing smart product-service systems”. In: *Procedia CIRP*. Vol. 83. Elsevier BV, pp. 324–329. DOI: 10.1016/j.procir.2019.04.077.

Porter, Michael E and James E Heppelmann (2015). “How smart, connected products are transforming companies”. In: *Harvard business review* 93.10, pp. 96–114.

Porter, Michael Eugene (1985). *Competitive advantage: Creating and sustaining superior performance*. New York: Free Press. ISBN: 0029250900.

Provost, Foster and Tom Fawcett (2013). “Data Science and its Relationship to Big Data and Data-Driven Decision Making”. In: *Big Data* 1.1, pp. 51–59. DOI: 10.1089/big.2013.1508.

Pujol Priego, Laia, David Osimo, and Jonathan Douglas Wareham (2019). “Data Sharing Practice in Big Data Ecosystems”. In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.3355696.

Ray, Jyotishka, Syam Menon, and Vijay Mookerjee (2020). “Bargaining over Data: When Does Making the Buyer More Informed Help?” In: *Information Systems Research* 31.1, pp. 1–15. DOI: 10.1287/isre.2019.0872.

Rayna, Thierry (2008). “Understanding the challenges of the digital economy: The nature of digital goods”. In: *Communications & Strategies* 71, pp. 13–16. URL: https://ssrn.com/abstract=1353583.

Rayport, Jeffrey F and John J Sviokla (1995). “Exploiting the virtual value chain”. In: *Harvard business review* 73.6, p. 75.

Ricci, Francesco, Lior Rokach, and Bracha Shapira (2011). “Introduction to Recommender Systems Handbook”. In: *Recommender Systems Handbook*. Springer US, pp. 1–35. DOI: 10.1007/978-0-387-85820-3_1.

Richter, Heiko and Peter R. Slowinski (2018). “The Data Sharing Economy: On the Emergence of New Intermediaries”. In: *IIC - International Review of Intellectual Property and Competition Law* 50.1, pp. 4–29. DOI: 10.1007/s40319-018-00777-7.

Roche, Dominique G. et al. (2014). “Troubleshooting Public Data Archiving: Suggestions to Increase Participation”. In: *PLoS Biology* 12.1. Ed. by Jonathan A. Eisen. DOI: 10.1371/journal.pbio.1001779.

Roth, Alvin E. (2002). “The Economist as Engineer: Game Theory, Experimentation, and Computation as Tools for Design Economics”. In: *Econometrica* 70.4, pp. 1341–1378. DOI: https://doi.org/10.1111/1468-0262.00335.

Rowley, Jennifer (2007). “The wisdom hierarchy: representations of the DIKW hierarchy”. In: *Journal of Information Science* 33.2, pp. 163–180. DOI: 10.1177/0165551506070706.

Sagiroglu, Seref and Duygu Sinanc (2013). “Big data: A review”. In: *2013 International Conference on Collaboration Technologies and Systems (CTS)*. IEEE. DOI: 10.1109/cts.2013.6567202.

Sakr, Mahmoud (2018). “A Data Model and Algorithms for a Spatial Data Marketplace”. In: *International Journal of Geographical Information Science* 32.11, pp. 2140–2168. DOI: 10.1080/13658816.2018.1484124.

Scaria, E, A Berghmans, M Pont, C Arnaut, and S Leconte (2018). *Study on data sharing between companies in Europe : final report*. Publications Office. DOI: doi/10.2759/354943.

Schröer, Christoph, Felix Kruse, and Jorge Marx Gómez (2021). “A Systematic Literature Review on Applying CRISP-DM Process Model”. In: *Procedia Computer Science* 181, pp. 526–534. DOI: 10.1016/j.procs.2021.01.199.

Schulz, Michael et al. (2020). “Introducing DASC-PM: A Data Science Process Model”. In: *Australasian Conference on Information Systems (ACIS) Proceedings*. URL: https://aisel.aisnet.org/acis2020/45.

Shapley, Lloyd S (1953). “A value for n-person games”. In: *Contributions to the Theory of Games* 2.28, pp. 307–317.

Shearer, Colin (2000). “The CRISP-DM model: the new blueprint for data mining”. In: *Journal of data warehousing* 5.4, pp. 13–22.

Shen, Bo, Yulong Shen, and Wen Ji (2019). “Profit Optimization in Service-Oriented Data Market: A Stackelberg Game Approach”. In: *Future Generation Computer Systems* 95, pp. 17–25. DOI: 10.1016/j.future.2018.12.072.

Shen, Yuncheng, Bing Guo, Yan Shen, Xuliang Duan, et al. (2019). “Pricing Personal Data Based on Information Entropy”. In: *ICSIM 2019*.

Shen, Yuncheng, Bing Guo, Yan Shen, Fan Wu, et al. (2019). “Pricing Personal Data Based on Data Provenance”. In: *Applied Sciences* 9.16, p. 3388. DOI: 10.3390/app9163388.

Shiyal, Bhadresh (2021). “Modern Data Warehouses and Data Lakehouses”. In: *Beginning Azure Synapse Analytics: Transition from Data Warehouse to Data*

Lakehouse. Berkeley, CA: Apress, pp. 21–48. ISBN: 978-1-4842-7061-5. DOI: 10.1007/978-1-4842-7061-5_2.

Siltala, Niko, Eeva Järvenpää, and Minna Lanz (2018). "Value proposition of a resource description concept in a production automation domain". In: *Procedia CIRP* 72, pp. 1106–1111. DOI: 10.1016/j.procir.2018.03.154.

Simeone, Alessandro, Yunfeng Zeng, and Alessandra Caggiano (2021). "Intelligent decision-making support system for manufacturing solution recommendation in a cloud framework". In: *International Journal of Advanced Manufacturing Technology* 112.3-4, pp. 1035–1050. DOI: 10.1007/s00170-020-06389-1.

Sirén, Charlotta A., Marko Kohtamäki, and Andreas Kuckertz (2012). "Exploration and exploitation strategies, profit performance, and the mediating role of strategic learning: Escaping the exploitation trap". In: *Strategic Entrepreneurship Journal* 6.1, pp. 18–41. DOI: 10.1002/sej.1126.

Smichowski, Bruno Carballa (2016). *Data as a common in the sharing economy: a general policy proposal.* URL: https://hal.archives-ouvertes.fr/hal-01386644 (visited on 12/13/2022).

Spiekermann, Markus (2019). "Data Marketplaces: Trends and Monetisation of Data Goods". In: *Intereconomics* 54.4, pp. 208–216. DOI: 10.1007/s10272-019-0826-z.

Stahl, Florian and Gottfried Vossen (2017). "Name Your Own Price on Data Marketplaces". In: *Informatica* 28.1, pp. 155–180. DOI: 10.15388/Informatica.2017.124.

Stefansson, Gunnar (2002). "Business-to-business data sharing: A source for integration of supply chains". In: *International Journal of Production Economics* 75.1-2, pp. 135–146. DOI: 10.1016/s0925-5273(01)00187-6.

Stigler, George J. (1961). "The Economics of Information". In: *Journal of Political Economy* 69.3, pp. 213–225. DOI: 10.1086/258464.

Strohm, Oliver and Eberhard Ulich (1997). *Unternehmen arbeitspsychologisch bewerten: ein Mehr-Ebenen-Ansatz unter besonderer Berücksichtigung von Mensch, Technik und Organisation*. vdf Hochschulverlag an der ETH Zürich. DOI: 10.3218/3951-1.

Sugden, Robert (1989). "Spontaneous Order". In: *Journal of Economic Perspectives* 3.4, pp. 85–97. DOI: 10.1257/jep.3.4.85.

Tang, Ruiming, Antoine Amarilli, Pierre Senellart, and Stéphane Bressan (2016). "A Framework for Sampling-Based XML Data Pricing". In: *Transactions on Large-Scale Data- and Knowledge-Centered Systems XXIV. Lecture Notes in Computer Science.* Vol. 9510. Springer, Berlin, Heidelberg, pp. 116–138. DOI: 10.1007/978-3-662-49214-7_4.

Tang, Ruiming, Dongxu Shao, Stéphane Bressan, and Patrick Valduriez (2013). "What You Pay for Is What You Get". In: *Database and Expert Systems Applications. DEXA 2013. Lecture Notes in Computer Science*. Ed. by Hendrik Decker, Lenka Lhotská, Sebastian Link, Josef Basl, and A. Min Tjoa. Vol. 8056. Springer, Berlin, Heidelberg, pp. 395–409. DOI: 10.1007/978-3-642-40173-2_32.

Tang, Ruiming, Huayu Wu, Zhifeng Bao, Stéphane Bressan, and Patrick Valduriez (2013). “The Price Is Right”. In: *Database and Expert Systems Applications. DEXA 2013. Lecture Notes in Computer Science*. Ed. by Hendrik Decker, Lenka Lhotská, Sebastian Link, Josef Basl, and A. Min Tjoa. Vol. 8056. Springer, Berlin, Heidelberg, pp. 380–394. DOI: 10.1007/978-3-642-40173-2_31.

Tang, Ruiming, Huayu Wu, Xiuqiang He, and Stephane Bressan (2015). “Valuating Queries for Data Trading in Modern Cities”. In: *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. Atlantic City, NJ: IEEE, pp. 414–421. DOI: 10.1109/ICDMW.2015.11.

Taylor, Linnet, Hellen Mukiri-Smith, Tjaša Petročnik, Laura Savolainen, and Aaron Martin (2020). “(Re)making data markets: an exploration of the regulatory challenges”. In: DOI: 10.31235/osf.io/pv98s.

Tian, Yingjie, Yurong Ding, Saiji Fu, and Dalian Liu (2022). “Data Boundary and Data Pricing Based on the Shapley Value”. In: *IEEE Access* 10, pp. 14288–14300. DOI: 10.1109/ACCESS.2022.3147799.

Trappey, Amy, Charles V. Trappey, and Alex Hsieh (2021). “An intelligent patent recommender adopting machine learning approach for natural language processing: A case study for smart machinery technology mining”. In: *Technological Forecasting and Social Change* 164, p. 120511. DOI: 10.1016/j.techfore.2020.120511.

Treder, Martin (2020). *The Chief Data Officer Management Handbook*. Apress. DOI: 10.1007/978-1-4842-6115-6.

Uckelmann, Dieter and Bernd Scholz-Reiter (2011). “Integrated Billing Solutions in the Internet of Things”. In: *Architecting the Internet of Things*. Ed. by Dieter Uckelmann, Mark Harrison, and Florian Michahelles. Springer, Berlin, Heidelberg, pp. 229–251. DOI: 10.1007/978-3-642-19157-2_9.

Vassilakopoulou, Polyxeni, Espen Skorve, and Margunn Aanestad (2018). “Enabling openness of valuable information resources: Curbing data subtractability and exclusion”. In: *Information Systems Journal* 29.4, pp. 768–786. DOI: 10.1111/isj.12191.

Vaus, David de (2013). *Surveys In Social Research*. Routledge. DOI: 10.4324/9780203519196.

Vernon, Richard (1979). “Unintended Consequences”. In: *Political Theory* 7.1, pp. 57–73. DOI: 10.1177/009059177900700104.

Visconti, Roberto Moro, Alberto Larocca, and Michele Marconi (2017). “Big Data-Driven Value Chains and Digital Platforms: From Value Co-Creation to Monetization”. In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.2903799.

Vlijmen, Herman van et al. (2020). “The Need of Industry to Go FAIR”. In: *Data Intelligence* 2.1-2, pp. 276–284. DOI: 10.1162/dint_a_00050.

Walker, Russell (2015). *From big data to big profits: Success with data and analytics*. New York: Oxford University Press. ISBN: 978-0-199-37832-6.

Wang, Catherine L and Pervaiz K Ahmed (2005). “The knowledge value chain: a pragmatic knowledge implementation network”. In: *Handbook of Business Strategy* 6.1, pp. 321–326. DOI: 10.1108/08944310510558115.

Wang, Gang, Angappa Gunasekaran, Eric W.T. Ngai, and Thanos Papadopoulos (2016). “Big data analytics in logistics and supply chain management: Certain investigations for research and applications”. In: *International Journal of Production Economics* 176, pp. 98–110. DOI: 10.1016/j.ijpe.2016.03.014.

Wang, Xingwang, Xiaohui Wei, Shang Gao, Yuanyuan Liu, and Zongpeng Li (2019). “A Novel Auction-Based Query Pricing Schema”. In: *International Journal of Parallel Programming* 47.4, pp. 759–780. DOI: 10.1007/s10766-017-0534-x.

Wang, Xingwang, Xiaohui Wei, Yuanyuan Liu, and Shang Gao (2018). “On Pricing Approximate Queries”. In: *Information Sciences* 453, pp. 198–215. DOI: 10.1016/j.ins.2018.04.036.

Wauters, P et al. (2018). *Study on emerging issues of data ownership, interoperability, (re-)usability and access to data, and liability : final report*. Publications Office. DOI: doi/10.2759/781960.

Webster, Jane and Richard T. Watson (2002). “Analyzing the Past to Prepare for the Future: Writing a Literature Review”. In: *MIS Quarterly* 26.2, pp. xiii–xxiii. ISSN: 02767783. URL: https://www.jstor.org/stable/4132319.

Weinhardt, Christof, Carsten Holtmann, and Dirk Neumann (2003). “Market-Engineering”. In: *Wirtscahftsinformatik* 45.6, pp. 635–640. DOI: 10.1007/bf03250926.

Westenbroek, Tyler, Roy Dong, Lillian J. Ratliff, and S. Shankar Sastry (2018). “Statistical estimation with strategic data sources in competitive settings”. In: *2017 IEEE 56th Annual Conference on Decision and Control, CDC 2017*. Vol. 2018-Janua. Institute of Electrical and Electronics Engineers Inc., pp. 4994–4999. DOI: 10.1109/CDC.2017.8264398.

Wilberg, Julian, Isabell Triep, Christoph Hollauer, and Mayada Omer (2017). “Big Data in Product Development: Need for a Data Strategy”. In: *2017 Portland International Conference on Management of Engineering and Technology (PICMET)*. IEEE. DOI: 10.23919/picmet.2017.8125460.

Wilkinson, Mark D. et al. (2016). “The FAIR Guiding Principles for scientific data management and stewardship”. In: *Scientific Data* 3.1. DOI: 10.1038/sdata.2016.18.

Wixom, Barbara H. (2014). “Cashing in on your data”. In: *Center for Information Systems Research, Sloan School of Management, Cambridge, MA: Massachusetts Institute of Technology. Research Briefing* 14.8. URL: https://cisr.mit.edu/publication/2014_0801_DataMonetization_Wixom (visited on 12/13/2022).

Wixom, Barbara H., Gabriele Piccoli, and Joaquin Rodriguez (2021). “Fast-Track Data Monetization With Strategic Data Assets”. In: *MIT Sloan Management Review* 62.4, pp. 1–4.

Wixom, Barbara H. and Jeanne W Ross (2017). “How to monetize your data”. In: *MIT Sloan Management Review* 58.3.

Wixom, Barbara H., Ina M. Sebastian, and Robert W. Gregory (2020). *Data Sharing 2.0: New Data Sharing, New Value Creation.* URL: https://cisr.mit.edu/publication/2020_1001_DataSharing_WixomSebastianGregory (visited on 12/13/2022).

Woerner, Stephanie L. and Barbara H. Wixom (2015). “Big Data: Extending the Business Strategy Toolbox”. In: *Journal of Information Technology* 30.1, pp. 60–62. DOI: 10.1057/jit.2014.31.

Xiao and Benbasat (2007). “E-Commerce Product Recommendation Agents: Use, Characteristics, and Impact”. In: *MIS Quarterly* 31.1, p. 137. DOI: 10.2307/25148784.

Xiao, Zheng, Dan He, and Jiayi Du (2021). “A Stackelberg Game Pricing Through Balancing Trilateral Profits in Big Data Market”. In: *IEEE Internet of Things Journal* 8.16, pp. 12658–12668. DOI: 10.1109/JIOT.2020.3001010.

Xiong, Li and Huixian Zheng (2019). “Data Products Pricing Mechanism: A Harmonious and Mutual-beneficial Perspective”. In: *IOP Conference Series: Materials Science and Engineering* 677.3, p. 032008. DOI: 10.1088/1757-899X/677/3/032008.

Xiong, Wei and Li Xiong (2021). “Anti-Collusion Data Auction Mechanism Based on Smart Contract”. In: *Information Sciences* 555, pp. 386–409. DOI: 10.1016/j.ins.2020.10.053.

Xu, Chengzhen, Kun Zhu, Changyan Yi, and Ran Wang (2020). “Data Pricing for Blockchain-based Car Sharing: A Stackelberg Game Approach”. In: *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*. Taipei, Taiwan: IEEE, pp. 1–5. DOI: 10.1109/GLOBECOM42002.2020.9322221.

Xu, Jingjun David, Izak Benbasat, and Ronald T Cenfetelli (2014). “The Nature and Consequences of Trade-Off Transparency in the Context of Recommendation Agents”. In: *MIS Quarterly* 38.2, pp. 379–406. DOI: 10.25300/MISQ/2014/38.2.03.

Xu, Xun (2012). “From cloud computing to cloud manufacturing”. In: *Robotics and Computer-Integrated Manufacturing* 28.1, pp. 75–86. DOI: 10.1016/j.rcim.2011.07.002.

Yan, Tom and Ariel D. Procaccia (2021). “If You Like Shapley Then You’ll Love the Core”. In: vol. 35. 6. Association for the Advancement of Artificial Intelligence (AAAI), pp. 5751–5759. DOI: 10.1609/aaai.v35i6.16721.

Yang, Chengrun, Yuji Akimoto, Dae Won Kim, and Madeleine Udell (2019). “OBoe: Collaborative filtering for automl model selection”. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, pp. 1173–1183. DOI: 10.1145/3292500.3330909.

Yang, Jian and Chunxiao Xing (2019). “Personal Data Market Optimization Pricing Model Based on Privacy Level”. In: *Information* 10.4, p. 123. DOI: 10.3390/info10040123.

Yang, Jian, Chongchong Zhao, and Chunxiao Xing (2019). “Big Data Market Optimization Pricing Model Based on Data Quality”. In: *Complexity* 2019, pp. 1–10. DOI: 10.1155/2019/5964068.

Yassine, Abdulsalam, Ali Asghar Nazari Shirehjini, and Shervin Shirmohammadi (2015). “Smart Meters Big Data: Game Theoretic Model for Fair Data Sharing in Deregulated Smart Grids”. In: *IEEE Access* 3, pp. 2743–2754. DOI: 10.1109/ACCESS.2015.2504503.

Ye, Yazhen, Yao Zhang, Guohua Liu, and Yangyong Zhu (2021). “A Measure Based Pricing Framework for Data Products”. In: *Web Intelligence* 18.4, pp. 249–260. DOI: 10.3233/web-210446.

You, Zhaoyang, Xinya Wu, Kexuan Chen, Xinyi Liu, and Chao Wu (2021). “Evaluate the Contribution of Multiple Participants in Federated Learning”. In: *Database and Expert Systems Applications. DEXA 2021. Lecture Notes in Computer Science*. Springer International Publishing, Cham, pp. 189–194. DOI: 10.1007/978-3-030-86475-0_19.

Yu, Haifei and Mengxiao Zhang (2017). “Data Pricing Strategy Based on Data Quality”. In: *Computers & Industrial Engineering* 112, pp. 1–10. DOI: 10.1016/j.cie.2017.08.008.

Yuncheng Shen et al. (2016). “A Pricing Model for Big Personal Data”. In: *Tsinghua Science and Technology* 21.5, pp. 482–490. DOI: 10.1109/TST.2016.7590317.

Zhamak, Dehghani (2022). *Data Mesh*. O’Reilly Media. ISBN: 9781492092391.

Zhang, Cheng, Daindi Chen, Fei Tao, and Ang Liu (2019). “Data driven smart customization”. In: *Procedia CIRP*. Vol. 81. Elsevier BV, pp. 564–569. DOI: 10.1016/j.procir.2019.03.156.

Zhang, Jiajia et al. (2019). “Negotiation Game Model for Big Data Transactions”. In: *2019 IEEE Fourth International Conference on Data Science in Cyberspace (DSC)*. Hangzhou, China: IEEE, pp. 162–167. DOI: 10.1109/DSC.2019.00032.

Zhang, Lin et al. (2014). “Cloud manufacturing: a new manufacturing paradigm”. In: *Enterprise Information Systems* 8.2, pp. 167–187. DOI: 10.1080/17517575.2012.683812.

Zhang, Mengxiao, Fernando Beltran, and Jiamou Liu (2020). “Incentive Mechanism for Social Network Data Pricing under Privacy Preservation”. In: *Proceedings of the 2nd ACM International Symposium on Blockchain and Secure Critical Infrastructure*. Taipei Taiwan: ACM, pp. 85–95. DOI: 10.1145/3384943.3409425.

Zhang, Mengxiao and Fernando Beltrán (2020). “A Survey of Data Pricing Methods”. In: *SSRN Journal*. DOI: 10.2139/ssrn.3609120.

Zhang, Yanru, Dusit Niyato, Ping Wang, and Zhu Han (2020). “Data Services Sales Design With Mixed Bundling Strategy: A Multidimensional Adverse Selection Approach”. In: *IEEE Internet of Things Journal* 7.9, pp. 8826–8836. DOI: 10.1109/JIOT.2020.2999824.

Zhang, Zheng, Wei Song, and Yuan Shen (2021). “A Reasonable Data Pricing Mechanism for Personal Data Transactions with Privacy Concern”. In: *Web and Big Data. APWeb-WAIM 2021. Lecture Notes in Computer Science*. Ed. by Leong Hou U, Marc Spaniol, Yasushi Sakurai, and Junying Chen. Vol. 12859. Springer International Publishing, Cham, pp. 64–71. DOI: 10.1007/978-3-030-85899-5_5.

Zhao, Yi et al. (2020). “Auction-Based High Timeliness Data Pricing under Mobile and Wireless Networks”. In: *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*. Dublin, Ireland: IEEE, pp. 1–6. DOI: 10.1109/ICC40277.2020.9149197.

Zheng, Xiao (2020). “Data Trading with Differential Privacy in Data Market”. In: *Proceedings of 2020 the 6th International Conference on Computing and Data Engineering*. Sanya China: ACM, pp. 112–115. DOI: 10.1145/3379247.3379271.

Zheng, Zhenzhe, Yanqing Peng, Fan Wu, Shaojie Tang, and Guihai Chen (2020). “ARETE: On Designing Joint Online Pricing and Reward Sharing Mechanisms for Mobile Data Markets”. In: *IEEE Transactions on Mobile Computing* 19.4, pp. 769–787. DOI: 10.1109/TMC.2019.2900243.