WIK-Consult • Report

Study for Facebook

# European Data Economy: Between Competition and Regulation

## Final Report

Authors:
Dr René Arnold
Christian Hildebrandt
Serpil Taş

WIK-Consult GmbH
Rhöndorfer Str. 68
53604 Bad Honnef
Germany

Bad Honnef, January 2020

# Imprint

WIK-Consult GmbH
Rhöndorfer Str. 68
53604 Bad Honnef
Germany
Phone: +49 2224 9225-0
Fax: +49 2224 9225-63
Email: info@wik-consult.com
www.wik-consult.com

**Person authorized to sign on behalf of the organization**

| | |
|---|---|
| General Manager | Dr. Cara Schwarz-Schilling |
| Director<br>Head of Department<br>Postal Services and Logistics | Alex Kalevi Dieke |
| Director<br>Head of Department<br>Networks and Costs | Dr. Thomas Plückebaum |
| Director<br>Head of Department<br>Regulation and Competition | Dr. Bernd Sörries |
| Head of Administration | Karl-Hubert Strüver |
| Chairperson of the Supervisory Board | Dr. Daniela Brönstrup |
| Registered at | Amtsgericht Siegburg, HRB 7043 |
| Tax No. | 222/5751/0926 |
| VAT-ID | DE 123 383 795 |

## Contents

## Your quick link to key findings

## Executive summary

Data and its economic impact permeates all sectors of the economy. The data economy is not a new sector, but more like a challenge for all firms to compete and innovate as part of a new wave of economic value creation.

Our report findings are contrary to the typically held view—we find that data access is not the main challenge to a thriving data economy in Europe. A much more pressing challenge is the lack of a common "data language" that can facilitate data exchanges. A common data language would help to transfer data from one context to another to create datasets that are usable by other firms. So-called "reference architectures" provide exactly this common language.

Policymakers across Europe should promote the development and adoption of such reference architectures as a way to increase the quantity and quality of data exchange through data sharing and data pooling between firms. Obligations to share data are found to be of limited benefit and can even prove detrimental, especially without the underlying reference architectures in place.

**Promoting unified data architectures to facilitate the European data economy**

Acclaimed digital leaders such as Estonia and South Korea have built their success on unified (underlying) reference architectures that facilitate data exchange and data reuse. These two countries have benefited from the socioeconomic potential of data more than others.

With data playing an increasingly important role across all sectors of the economy, the results of this report point European policymakers to promote the development and adoption of unified reference architectures. These architectures constitute a technology-neutral and cross-sectoral approach that will enable companies small and large to compete and to innovate—unlocking the economic potential of data capture in an increasingly digitized world.

Data access appears to be less of a hindrance to a thriving data economy due to the net increase in capabilities in data capture, elevation, and analysis. What does prove difficult for firms is discovering existing datasets and establishing their suitability for achieving their economic objectives. Reference architectures can facilitate this process as they provide a framework to locate potential providers of relevant datasets and carry sufficient additional information (metadata) about datasets to enable firms to understand whether a particular dataset, or parts of it, fits their purpose.

Whenever there is an existing dataset that can be used, accessing this dataset as a third party is likely to be more efficient than first-party data capture. In such situations, profit-maximizing firms should have a preference for data exchange over data capture. This is underscored by the fact that companies already frequently exchange data in both horizontal data-sharing agreements and vertical data-pooling schemes such as the Industrial Data Space. Based on this premise, reference architectures, as they become available and are adopted across sectors, should increase the frequency and quality of data exchanges for companies big and small across many sectors.

**Economic actors, not policymakers, should decide on the value of a data exchange**

Whether third-party data access is suitable to solve a specific business task in the first place ought to be a decision at the discretion of the economic actors involved. As our report underscores, data captured in one context with a specific purpose may not be fit for another context or another purpose. Consequently, a firm has to evaluate case-by-case whether first-party data capture, third-party data access, or a mixed approach is the best solution. This evaluation will naturally depend on whether there is any other firm capturing data suitable for the task that is willing to negotiate conditions for third-party access to this data. Unified data architectures may also lower the barriers for a firm capturing suitable data to engage in negotiations, since its adoption will lower the costs of making the data ready for a successful exchange. Such architectures may further integrate licensing provisions ensuring that data, once exchanged, is not used beyond the agreed purpose. It can also bring in functions that improve the discoverability of potential data providers.

**A data-sharing regulation is unlikely to resolve concerns around the data economy**

Data exchanges across all sectors represent a critical stepping stone for a thriving European data economy. However, concerns of data concentration with only a few large firms have recently entered the policy debate. These concerns revolve around (1) consumers being locked-in by service providers and (2) in particular small- and medium-sized enterprises (SMEs) not being able to compete effectively due to a lack of access to data.

To mitigate **consumer lock-in**, data portability has been established as part of the General Data Protection Regulation (GDPR). However, while data extraction is manageable, importing these extracted (personal) data to another service provider remains a significant challenge for the data economy as a whole. Moreover, regulatory specifications on data interoperability can facilitate successful data porting. Private initiatives such as the Data Transfer Project have emerged to address this challenge.

We find that **a lack of data access** is neither the only nor the most-important success factor for businesses. Fundamentally, a lack of data access is difficult to conceive. Virtually any data point (datum) can be captured or otherwise determined by various methods and so a firm that is unable to capture data with the same procedure as its competitor may still be able to capture the required data some other way.

Against this backdrop, the assumption of (personal) data being an essential facility in the same sense as (physical) infrastructure such as a railroads or telecommunications networks is misleading. In order to qualify as an “essential” facility or infrastructure, (personal) data would have to fulfill two necessary conditions: 1) market entry to the complementary market is not effectively possible without access to this facility; and 2)  a supplier on the complementary market cannot duplicate this facility with reasonable effort and there is no substitute. It is obvious that (personal) data does not fulfill either condition. Regarding the first condition, it requires that one firm is more cost-efficient than alternatives. However, every firm in the data economy can gather any (personal) data at often negligible cost. Consequently,  there is no monopoly when it comes to (personal) data and so the first condition of being an essential facility is not satisfied. With respect to the second condition, (personal) data can always be duplicated or replicated with reasonable effort, even if it requires the consent of the end-user. And so, (personal) data cannot be an essential facility.

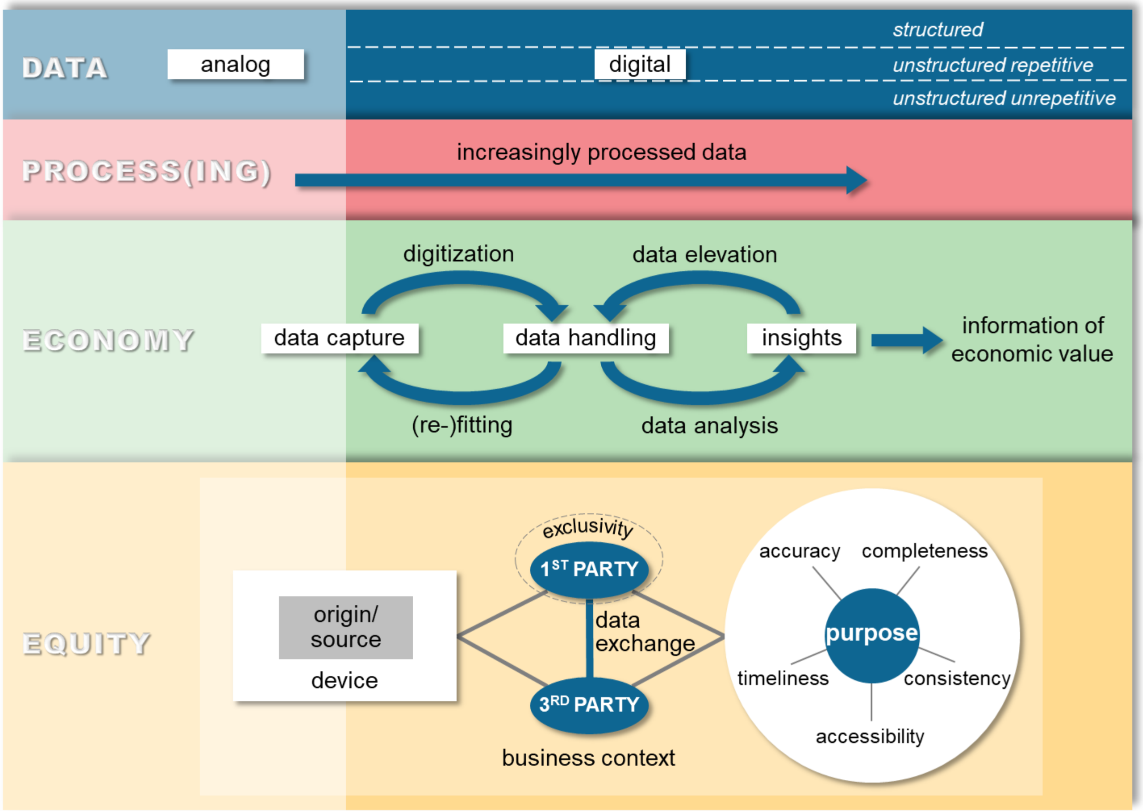
A general regulatory obligation to share or to pool data is therefore difficult to justify. In practice, generating economic value from (big) data is a rather complex and varied process. So, even if a regulatory obligation was justifiable, such an obligation would have to be flexible enough to adapt to these different circumstances. However, this would likely include a high administrative burden passed on to firms and may therefore ultimately do more harm than good to the European data economy.

The burden of transaction costs without a unified data architecture in place would likely eliminate any incentive to request data exchanges. If personal data were to be exchanged, the concept of consent under GDPR (General Data Protection Regulation) would have to be amended as such exchanges would almost inevitably entail a change in the purpose for which the data were originally captured. It is difficult to conceive how competent authorities would be able to enforce such a regulatory obligation without full transparency about all data, its properties, contexts, and purposes held by all parties involved. Finally, such a market intervention would inevitably pick winners and losers, without any convincing proof of economic harm, which renders any justification ineffective.

**False assumptions promoted a misleading popular narrative around data and its socioeconomic impact**

False assumptions about the nature of data and the data economy appear in the current popular narratives of the socioeconomic impact of data and data-sharing regulation. Our report aims to educate the public debate by shedding light on the key characteristics of data and the role it plays for businesses, innovation, and competition. To achieve this, we developed a layered framework to make sense of data (Figure 2-1, reproduced here).

Figure 0-1: A layered framework to make (economic) sense of data



Source: WIK-Consult.

Our framework highlights critical properties of data as well as various shortcomings and fallacies of popular data economy narratives. Consisting of four layers, the first layer represents the conversion of **analog data** (signals from the real world) into **digital data**. On this **data layer**, digital data can be further split into **structured**, **unstructured repetitive**, and **unstructured unrepetitive** digital data. While other ways of representing data may be useful in some contexts, they consistently fail to inform about the nature of data and its implications for data utilization and this is particularly important since the vast majority of digital data available today is unstructured (e.g., video, pictures, text, natural language, social media streams).

The **process(ing) layer** highlights the fact that there is no raw digital data. Starting with digitization, conscious decisions have already been taken about which analog signals shall be converted to digital bits as well as how exactly this conversion will happen (e.g., which sensors or interfaces are employed). The degree of processing usually increases as digital data is further utilized. Contrast this with the popular belief that digital data are objective facts that therefore have the same value to any economic actor.

The **economy layer** illustrates that turning data into **information of economic value** is not necessarily a straightforward process. Only when in combination with a specific context can data become information of economic value. Logically, the digitization happens during **data capture,** where devices with sensors and/or interfaces are used and data is transferred to **data handling** consisting of pre-processing, storage, and in many cases immediate feedback to the data capture function in the form of **re-fitting**. To create **insights** from data, further **data analysis** is necessary. Insights can be used to elevate data to a new level, resubmitting it for further analysis. This process enables the emergence of **information of economic value**. In contrast to the popular narrative, this layer highlights that data—no matter the quantity that is available—has little to no economic value. In fact, businesses have to put substantial effort in to turn the data into **information of economic value**.

On the **equity layer** of the diagram, the role of the **device** is highlighted as it surrounds the origin and source of the analog signal and constitutes the first filter through which information ultimately finds its way into digital datasets. This data can be accessed by **first-party organizations**: the ones having control over the device and/or the software running on the device, providing them with access to the source and (some) control over what data is captured and in what manner. Such first-party organizations can decide to make this data available to **third-party organizations,** or keep it to themselves (**exclusivity**). Third-party access to data happens through **data exchange**. The key take-away educating the popular debate here is that there is a trade-off for firms when choosing between first- and third-party access in which these firms have to balance their control over the data capture against the potential cost savings of not having to develop processes to capture and handle data.

In line with this, the equity layer illustrates that data capture is never independent from the business context and the purpose of data use. Notably, the intended purpose does not preclude the data from being used in other contexts. The purpose is critical to the data quality, having influence on the **accuracy**, **completeness**, **consistency**, **accessibility**, and **timeliness** of the data.

### The data economy—An important part of the digital single market

The data economy is not an emergent sector, but is in fact a new wave of innovation and competition challenging established business practices in all sectors. While data is fueling this new wave of economic activity, **more data does not necessarily translate into more value.** Very much unlike oil or a currency, data is non-rivalrous and intangible. In the same vein, value creation from data does not resemble traditional value chains—instead, economic exchanges perpetuate circularly, fostering multi-sided markets.

# 1 Introduction

Recent innovations have enabled us to capture, analyze and store (digitized) data at an unprecedented scale. Although varying sources arrive at different specific figures[1], it is undeniable that there is huge economic potential associated with what is typically referred to as the "data economy".

Digitalization and digital data are affecting all sectors of the economy, it is hard to imagine that in the long-term any sector will not somehow have to increase its efforts to take advantage of digital data. While the data can be just as varied as the sectors themselves, the popular debate seems preoccupied with the use of personal data mainly with a view to monetize services offered on the internet through targeted advertising. Our report seeks to broaden the scope of the discussion and to educate it.

In **Chapter 2**, we seek to **shed light on some of the underlying properties of data.** Within this, we test some of the potentially inflated expectations and assumptions regarding the power of data to facilitate economic activity. Our report explores the role of data quality alongside data quantity, which is usually the focus of interest.

**Chapter 3 explores how the data economy works and explains the role of data in digital business models**, most notably digital platforms. Data plays an integral role in innovation; however, so far it seems unclear which factors affect data's role in innovation and whether a lack of data access does actually impede innovation.

**Chapter 4 focuses on the role of data for competition.** It reflects on data as an input and as a key competitive resource, in particular with respect to data quality. This chapter considers the powerful economic effects underlying the structure of the data economy and the potential for data-driven market power and barriers to entry, before suggesting measures to promote effective competition in the data economy.

**Chapter 5 highlights several approaches to data exchange,** reflecting on each of them and elaborating on their respective advantages and disadvantages.

**Chapter 6** draws on the insights gained in the report to provide a summary of the **challenges of the data economy**, bringing together high-level **implications for policymakers** and regulators and suggesting a way forward to facilitate **more-frequent and higher-quality data exchanges** through the use of **reference architectures.**

**Chapter 7 concludes** the report.

[1] For example, several years ago, Manyika J, Chui M, Groves P, Steve F, Kuiken V, Doshi EA. 2013. Open data: Unlocking innovation and performance with liquid information, McKinsey Global Institute, rated the annual value enabled by open data in seven different domains at about $3 trillion globally. IDC, Open Evidence. 2017. European Data Market SMART 2013/0063 - Final Report. A study prepared for the European Commission, IDC, Open Evidence, estimated the value of the data economy at about €300 billion in 2016 across Europe. For the future, according to Hogan O, Holdgate L, Jayasuriya R. 2016. The Value of Big Data and the Internet of Things to the UK Economy, Cebr, London, the contribution of big data as well as the IoT will reach £62 billion by 2020 in the UK alone.

## Key Findings of Chapter 2

- *To understand the economic value of data, it is crucial to realize that data cannot be considered to be objective facts, but rather consciously selected representations of reality from the context and the purpose for which they were captured.*
- *Processing can elevate data's economic value in one context, but render it null in others.*
- *Data quantity and data quality are closely intertwined. One must not be considered without the other to discern the economic value of data.*
- *Data is intangible and non-rivalrous. Most importantly, data varies not only in form and format, but also with regards to context and purpose.*
- *A specific datum can be captured and accessed in many ways. True exclusivity of data access is difficult to establish in an increasingly digitized world.*
- *There is a diminishing return when scaling up data quantity. However, a minimum quantity of data is required in most contexts. The size of this minimum quantity depends heavily on the context and purpose of data utilization.*
- *Data is not a magic potion. Theory, knowledge and skill are (at least) as important inputs to the economic value of data.*
- *Data typologies can only cover part of the process from analog signals over digital data to information of economic value. An holistic framework is required for a comprehensive understanding of the factors influencing data and data value.*
- *To typologize data effectively, it is essential to consider the specific context and the intended purpose of the data. In general, it is helpful to differentiate between 1) structured data, 2) unstructured repetitive data, and 3) unstructured unrepetitive data.*

# 2 Making sense of data

With the internet, and an increasing number of connected devices, our ability to transfer analog signals into digital bits (i.e., our ability to digitize) has increased. Typically, when we speak of data we think of digital bits that store information, rendering them accessible for analysis, which in turn enables the information to be used in various contexts. This process is at the heart of the digital transformation that businesses and our society at large are going through. Substantial improvements in data storage and analytics have enabled innovative applications of digital data. To begin our discussion of the data economy, we provide a brief overview of a framework that we recommend, based on our research, as an integrated approach to pinpoint key challenges, which will be addressed in detail in later sections. Our framework is described in the following section.

## 2.1 A layered framework

> ***Insights****: Only an holistic understanding of the whole process: moving from analog data to digital data and then on to the creation of information of economic value can truly educate the data economy debate. A single typology for data is almost always misleading and by default incomplete.*

Our ability to capture and utilize data has dramatically increased over recent decades, mainly through digital technologies comprising systems, devices, and resources that generate, store, process, exchange, or use digital data. The process of the conversion of an analog signal conveying information (e.g., sound, image, printed text) to binary bits is called **digitization**. The application or increase in the use of digital technologies by an organization, industry, or country, transforming existing tasks, or enabling new ones is called **digitalization**. This concept refers to how digitization affects the economy or society (OECD & Eurostat 2018). For the remainder of the report, we follow this terminology. In line with this, digitization is a prerequisite to the data economy.

Our layered framework aims to make (economic) sense of data and consists of four layers. The first layer represents the conversion of **analog data** (signals from the real world) into **digital data**. On this **data layer**, digital data can be further split into **structured**, **unstructured repetitive**, and **unstructured unrepetitive** digital data. While other typologies may be useful, depending on the context, they consistently fail to inform about the nature of the data and its implications for data utilization. This is particularly important, since the vast majority of digital data available today is unstructured (e.g., video, pictures, text, natural language, social media streams).

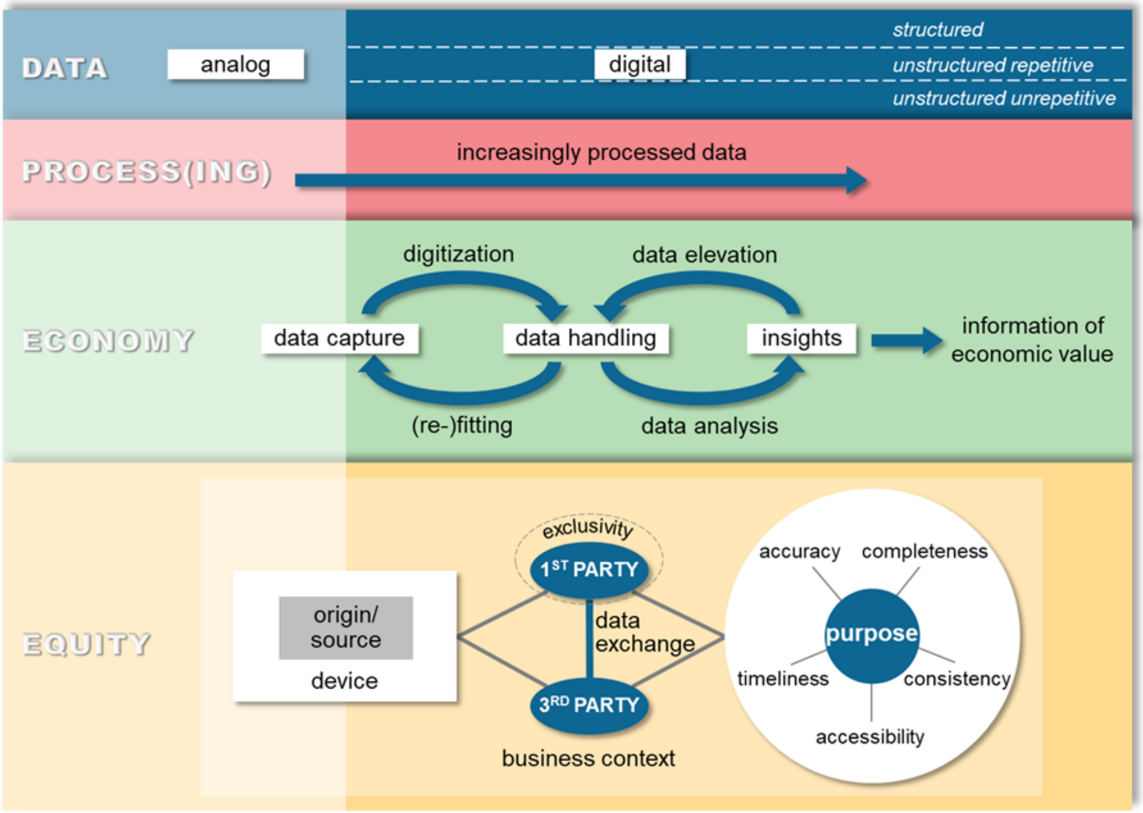
The **process(ing) layer** highlights the fact that there is no raw digital data. Starting with digitization, conscious decisions have already been taken about which analog signals shall be converted to digital bits as well as how exactly this conversion will happen, e.g., which sensors or interfaces are employed. The degree of processing usually increases as digital data is further utilized.

The **economy layer** illustrates how the turning of data into information of economic value is not necessarily a straightforward process. It is only in combination with a specific context that data can become information of economic value. Logically, the digitization happens during **data capture,** for which devices with sensors and/or interfaces are used and data is transferred to **data handling**, which consists of pre-processing, storage, and, in many cases, immediate feedback to the data capture function in the form of **re-fitting**. To create **insights** from data, further **data analysis** is necessary. Insights in themselves can be used to **elevate data** to a new level, resubmitting them for further analysis. This iterative process enables **information of economic value** to emerge. In contrast to the popular narrative, this layer highlights that data—no matter the quantity that is available—inherently has little or no value. In fact, businesses have to put substantial effort into turning them into information of economic value.

On the **equity layer**, the role of the **device** is further highlighted as it surrounds the **origin and source** of the analog signal. Thus, it constitutes the first filter through which information ultimately finds its way into digital datasets. This data can be accessed by **first-party organizations**: the ones having control over the device and/or the software running on the device, providing them with access to the source and (some) control over what data is captured and in what manner. Such first-party organizations can decide to make this data available to **third-party organizations,** or keep it to themselves (**exclusivity**). Third-party access to data happens through **data exchange**.

In line with this, the equity layer illustrates that data capture is never independent from the **business context** and the **purpose** of data use. Notably, the intended purpose does not preclude the data from being used in other contexts. However, by tailoring the data processing for one specific purpose, the data’s usefulness for other purposes might be compromised. The purpose is critical to the data quality, having influence on the **accuracy**, **completeness**, **consistency**, **accessibility**, and **timeliness** of the data. Notably, the context can vary between “data-rich” and “data-poor” environments, depending on the degree of digitalization in that particular context.

Figure 2-1: A layered framework to make (economic) sense of data



Source: WIK-Consult.

In a nutshell, our framework highlights various shortcomings and fallacies of popular narratives around the data economy.

- “Data suppliers” and “data users” cannot be distinguished in a straightforward and clear-cut manner as suggested by the DataLandscape[2] project. Instead, due to data capture having become an integral part of virtually all organizations, any organization will, for some data capture or data elevation instances, have first-party access, while only third-party access will be possible for data not captured or elevated directly by the organization.
- Data heterogeneity (variation) is not only due to it coming from different sources, but is a combination of differences in context, purpose, source, device, and data handling. This scope of heterogeneity is often neglected in other reports (e.g., Arnaut et al (2018), Crémer et al (2019), Furman et al (2019), Morton et al (2019) Feld (2019)).

[2] http://datalandscape.eu/
The website, as well as the reports and data presented there, are part of the EU Data Market study contracted by the European Commission under *SMART* 2013/0063.

- In contrast to the message of several reports, the use of big data is not a self-propelling tool enabling economic success in the data economy. In fact, finding the right balance between data quality and data quantity is a rather complex task. Depending on the context and purpose of data use, firms have to take into account several dimensions of data quality such as accuracy, completeness, and timeliness. While there is a minimum required quantity of data, there are also diminishing benefits for increasing dataset sizes. Therefore, businesses using big data have to deal with several trade-offs, pointing toward the importance of relevant expertise (e.g., data scientists) and experience (e.g., learning-by-doing feedback loops).
- The discussion on the relevance of data seems to assume that data itself has a value. However, data as such is worthless. In fact, in order to enable an educated discussion on the data economy, it is very important to understand and recognize that only data in combination with context becomes information, which in turn enables value creation.

## 2.2 Defining data

> ***Insights**: Only data in combination with context becomes information, which in turn enables value creation. Reuse of data depends on knowledge about the context and purpose for which the data was originally captured.*

Despite its obviously increasing importance, the term “data” and its definition have received very little attention from researchers. Surprisingly, academic disciplines such as law and economics have made virtually no effort to truly understand the nature of data while they have invested substantial effort in exploring and discussing the impact of data on welfare, economic growth, and competition. As Furner (2016) stated: “*A source of misunderstanding in contemporary discussions of data science and big data is a tendency to conflate three related but distinct interpretations: data as evidence, data as typically numeric attribute-values and data as bits.*” (p. 298) Consequently, our contribution sets off by spelling out the specific definition of the term that we adhere to.

For the present contribution, we will use the word datum for the individual data point. Data will be the term that we used to describe a multitude of individual data points. As regards the definition of data as such, we concur with the definition put forward by Kaase (2001) and suggested by Hjørland (2019) as the most fruitful one based on his review of current definitions of the term:

“*Data are information on properties of units of analysis*.”

This definition appears to be superior to other definitions as it does not confuse data with documents.[3] It includes the unit-phenomena captured in the data, which is a necessary condition to make data interpretable.[4] This is furthermore important as the definition acknowledges that the choice, understanding and description of a "unit" depends on the context of data capture, elevation, or interpretation. Most importantly, this definition acknowledges that data does not speak for itself in the sense of being objective facts (Hjørland 2019). Indeed, only data in combination with context becomes information, which in turn enables value creation. However, it is useful to explore how the understanding of data has developed over time in order to shed light on some of the key shortcomings of the current debate around what is typically referred to as the data economy.

The term "data" can be traced back to the Latin.[5] *Datum* (i.e., the singular) in the literal sense of the word means "that which is given" and *data* (i.e., plural) "things given" or simply "gifts". This origin shaped how early mathematicians like Euclid used the term data to refer to given facts within an equation such as the length of the sides and angles within triangles of which some may be given while others can be calculated based on mathematic principles. This understanding of data persisted until the mid of the 19th century, when, for instance, Worcester's dictionary of the English language defines data as *"Truths or premises given or admitted from which to deduce conclusions, the facts from which an inference is drawn."* This understanding shifted however in the second half of the 19th century when, alongside the emerging disciplines of statistical and social sciences, the number of tables increased, which provided systematically organized recording and reporting of frequencies and quantities resulting from observations and measurements. Suddenly, these tables—once collected and fixated in written form—were treated as "given", thus becoming the raw input for novel forms of quantitative analysis. Eventually, these inputs were referred to as data (Furner 2016).[6]

[3] For an elaboration on the difference between data and documents see Furner J. 2016. "Data": The data In *Information Cultures in the Digital Age - A Festschrift in Honor of Rafael Capurro*, ed. M Kelly, J Bielby, pp. 287-306. Wiesbaden: Springer.

[4] For further elaboration on this see Jensen HE. 1950. Editorial Note In *Through Values to Social Interpretation: Essays on Social Contexts, Actions, Types and Prospects*, ed. H Becker, pp. vii-xi. Durham, NC: Duke University Press cf. Hjørland B. 2019. Data (with Big Data and Database Semantics). *KO Knowledge Organization* 45: 685-708.

[5] Specifically, the term originates from the verb's present active indicative dō ("I give") whose perfect participle is datus ("given"). The two forms used most commonly today datum and data are the respective participle's nominative neuter singular form and its nominative neuter plural as well as the nominative feminine singular.

[6] Furner J. 2016. "Data": The data In *Information Cultures in the Digital Age - A Festschrift in Honor of Rafael Capurro*, ed. M Kelly, J Bielby, pp. 287-306. Wiesbaden: Springer provides a much more detailed and highly insightful history of the terms data and datum throughout the centuries. We have only summarized his book chapter here very briefly.

Treating such frequencies and quantities stemming from observations and measurements as “given” or even facts is a key fallacy rooted in a naïve realist ontological assumption that phenomena exist independently of any observers and data can be read off fully objectively from a single objective reality. As Drucker (2011) points out, such a conception of data completely neglects the “situated, partial, and constitutive” character of knowledge creation. Consequently, she suggests to replace *data* with *capta*, i.e., to replace the “given” with the “taken” in the literal sense of the word. Checkland (1999) and Capurro (1978) as quoted in Zins (2007) concur with this notion.

In the same vein, Hjørland (2019) highlights that *“[…], documented data are considered as being facts for the tasks they are produced to serve, they represent sufficient facts in a given social context. If they were not, it would be impossible to act on the basis of data; for example, it would be impossible to construct family trees, and there would be no reason to issue such documents (e.g., birth certificates). […] The datum “X is child of Y” may, for example, be obtained from three kinds of documents:*

*(1) an interview with Y;*

*(2) the birth certificate of X;*

*(3) and DNA analysis based on DNA from X and Y.*

*Normally all three documents will be considered reporting this datum as a fact. The DNA report is the most reliable source today, but all three contain the datum”* (p. 688).

All these authors point to an obvious shortcoming of the current debate around the data economy, which seems to be driven by the implicit presumption that data are discrete, objective facts or observations when in fact they are not and cannot be either of these, or, as Gitelman (2013) put it: *“‘Raw data’ is an oxymoron.”*[7]

[7] Somewhat counter-intuitively, big data—typically thought of as a highly immediate capture of real-world actions of human and non-human actions due to an implicitly assumed objectivity of the sensors themselves—appears to be particularly prone to the epistemological challenges outlined here. First and foremost, in particular streamed big data usually undergoes automatic quality checking whose main purpose is to identify and correct “outliers” originating from the data sources used in the specific circumstance. While there are different ways to achieve this correction, all of these ways have in common that they develop some kind of prediction based on historical data from the same source against which each new datum is benchmarked and if the variation exceeds a predefined range, the datum is usually replaced by a datum more in line with the historical data from that source, e.g., the corresponding minimum, maximum, or median. Alternatively, the datum is disregarded altogether. Thus, instead of treating the input as objective, it is always automatically checked and may be corrected. Some of these systems rely on deep learning algorithms which may even render the procedure of how and why a specific datum was replaced opaque to human actors. Secondly, the heterogeneous nature of big data makes it susceptible to epistemological challenges referring to inconsistent or even plainly contradictory documents referring to one and same datum. How the choice is made, which of these documents holds the true datum (and therefore becomes part of the dataset) is also usually opaque. Mirzaie M, Behkamal B, Paydar S. 2019. Big Data Quality: A systematic literature review and future research directions. *arXiv preprint arXiv:1904.05353* .

Understanding the situated and purposeful definition and collection of data is of essential importance to the data economy. First, this opposes a view held by some people that data is merely a byproduct of digitalization.[8] Second, understanding the fundamental nature of data can critically inform the emerging debate around access to and reuse of data, e.g., in the form of data exchange. As Borgman (2010) states: *"All too rarely do those promoting the sharing and curation of data define 'data' explicitly or acknowledge the diversity of forms that data may take."* (p. 2).

A large part of the expectations linked to the data economy and the socioeconomic potential of exchanging or otherwise making data accessible for (re-)use appears to build on the concept of big data and the so-called "fourth paradigm"[9] enabled by big data.[10] Essentially, the fourth paradigm proclaims a new age of discovery based on big data in which theory is no longer the starting point for the search for knowledge, but the data as such takes that role (Dhar 2013). The aim is to identify "*interesting and robust patterns that satisfy data*" and are expected to occur in the future (Dhar 2013). The insights gained based on such patterns are expected to have enormous potential for firms in creating new businesses, developing new services and products, and improving business operations (Lee 2017). Cao (2017) is somewhat more cautious in defining data products: "*A data product is a deliverable from data, or is enabled or driven by data, and can be a discovery, prediction, service, recommendation, decision-making insight, thinking, model, mode, paradigm, tool, or system. The ultimate data products of value are knowledge, intelligence, wisdom, and decision*" (p. 51).

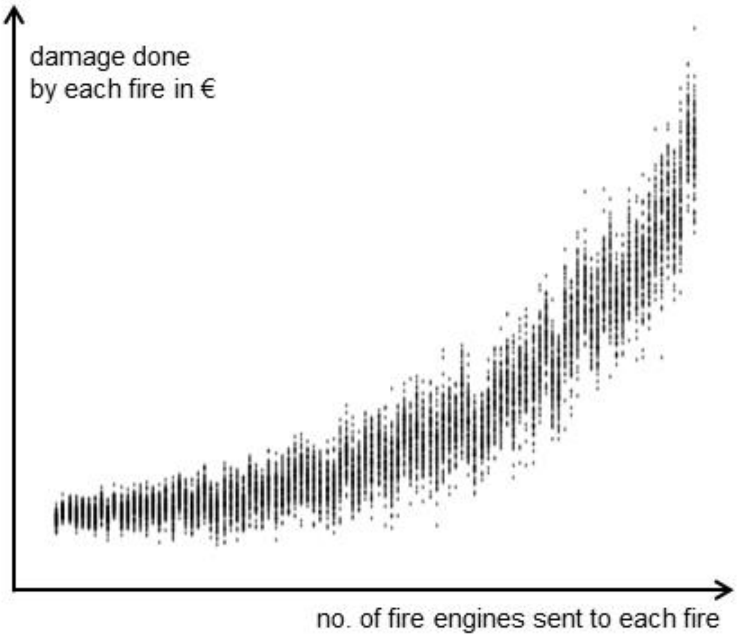
[8] Arguably, whether data can be considered a byproduct or not may depend on the perspective. From the perspective of a data subject—which can be an individual or an organization—data can be a byproduct in the sense that they have little or no influence over data that may be captured as part of actions they take. From the perspective of the entity that captures and may eventually store and utilize the data, an active decision to do so is required. As such data cannot be considered a byproduct in data value creation. Here, the purposeful and contextualized capture of data is always the starting point. This does however not preclude data, once captured, from being utilized for purposes other than the one(s) originally intended.

[9] The concept of the "fourth paradigm" was originally coined by Microsoft Research. See Hey T, Tansley S, Tolle K. 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmont, WA: Microsoft Research.

[10] For our discussion of data quality versus quantity see Section 2.5.

**Example**

Figure 2-2: Example for a potentially deceiving correlation



The above data about many fires indicates an obvious positive correlation between the damage done by a fire and the number for fire engines that were sent to this fire. Based solely on the data, one would have to arrive at the conclusion that sending fewer fire engines to a fire must reduce the damage done. Naturally, the opposite is true. So, without theoretical knowledge about the causal linkages in the observed phenomena it can be very easy to arrive at the wrong conclusion. Consequently, one ought to be careful to take patterns in data as self-evident.

From both an epistemological and practical data science perspective, the premise that data can replace theory is considered to be false. Boyd and Crawford (2012) integrate their critique of this notion even in their definition of big data, when they treat it as a "*cultural, technological, and scholarly phenomenon that rests on the interplay of:*

*(1) Technology: maximizing computation power and algorithmic accuracy to gather, analyze, link, and compare large datasets.*

*(2) Analysis: drawing on large datasets to identify patterns in order to make economic, social, technical, and legal claims.*

*(3) Mythology: the widespread belief that large datasets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy*" (p. 663).

Many other researchers underscore this critique. For example, Frické (2009) in his critique of the Data-Information-Knowledge-Wisdom (DIKW) hierarchy points out that the idea of data being (automatically) the root of information and that information answers questions encourages the mindless and meaningless collection of data. This collection is driven by the hope that this data will ascend to information, which may never happen.[11] Hjørland (2019) observes that the "fourth paradigm *... causes a block for better-understanding theoretical problems related to data and knowledge organization. Therefore, although big data is the background for 'e-science,' e-science does not define big data and cannot do without theory"* (p. 701). Indeed, much of the excitement around big data seems to be rooted in a problematic (naïve) empiricism.[12] Criticizing this empiricism, however, does not invalidate data or the data economy, it only highlights the need to understand the nature of data and to explore the undeniable socioeconomic potential based on other background assumptions.

Beyond a fundamental understanding of data as such, in the context of the data economy, it is particularly important to also elaborate on the economic characteristics of data. We elaborate on those in the following section.

## 2.3 Characterizing data from an economic perspective

> ***Insights**: Data is non-rivalrous in its use, it can be kept private (exclusivity of first-party organization) or made public (e.g., published via the internet), and its value depends on the timeliness, context, and intended purpose.*

Every discipline appears to inflict its very own preconceptions, knowledge and traditions on the term data and its use. In Section 2.4, we show that this also emerges from the respective categorizations of data drawn from the various disciplines. We approach from an economic perspective while not disregarding important insights from other disciplines and streams of research.

[11] Consequently, the concepts of big data and data mining are frequently the object of controversial discussions in information and data science. See e.g., Austin PC, Goldwasser MA. 2008. Pisces did not have Increased Heart Failure: Data-driven Comparisons of Binary Proportions between Levels of a Categorical Variable can Result in Incorrect Statistical Significance Levels. *Journal of Clinical Epidemiology* 61: 295-300, Austin PC, Mamdani MM, Juurlink DN, Hux JE. 2006. Testing Multiple Statistical Hypotheses Resulted in Spurious Associations: a Study of Astrological Signs and Health. Ibid.59: 871-72, Frické M. 2015. Big Data and its Epistemology. *Journal of the Association for Information Science and Technology* 66: 651-61.

[12] This also sheds doubts on the ultimate success of what is known as computational theory discovery see e.g. Berente N, Seidel S, Safadi H. 2018. Research Commentary—Data-Driven Computationally Intensive Theory Development. *Information Systems Research* 30: 50-64.

From an economic perspective, data is an intangible good and can therefore be used many times and for different purposes at different times (Floridi 2010, Hildebrandt & Arnold 2016, Schepp & Wambach 2016). Data is non-rivalrous in its use, as it can be used as an input factor many times, simultaneously or sequentially. If data is captured or elevated by one firm, this does not prohibit other firms from collecting the same data. As pointed out in the above example made by Hjørland (2019), one and the same datum can be collected in various ways. Thus, all firms interested in it can either follow the same or different ways to capture it.

Another key property of data—once it has been captured—is that other actors can be excluded from its use. On the one hand, if there is a constellation of non-rivalry and exclusivity, data in economic terms can be a "club commodity" (Buchanan 1965). On the other hand, if there is non-exclusivity, for instance, through the publication of data on the internet (e.g., open access), then data can also become a "public good" (Cornes & Sandler 1986). Consequently, popular descriptions of data as oil or a currency are misleading, since the value of data depends critically on its context, accuracy, and timeliness. For instance, unlike oil, data is not consumed when it is used. As the value of data depends on the context and intended purpose, its usage also makes it extremely difficult to put a specific price on data. What is clear, however, is that an individual datum typically has little to no monetary value on its own.

Regarding the timeliness of data, the information that data conveys has limited value most of the time. For instance, weather data is very relevant for today or tomorrow, but as soon as the corresponding day is over, it becomes historical data that is much less valuable. Furthermore, current income data is also significantly more valuable than historical income data when, for instance, targeting advertisements at consumers. However, historical data can also have a significant value, depending on the business model, but in most cases the so-called half-life period is considered to be very short (Feijóo et al 2016).

This also implies that very large amounts of data can be worthless after a very short period of time. Moreover, there are huge differences in the capability and success of businesses in gaining insights (relevant information, patterns, structures and trends) from the data (Banko & Brill 2001, Junqué de Fortuny et al 2013, Tucker 2010).[13] While these properties should set the frame for the popular debate around the data economy, other—more, or less, useful—data typologies tend to override them. In the following section, we shed light on these data typologies and their contribution to the debate.

[13] See also Cao's data products in Section 2.2.

## 2.4 Data typologies

> ***Insights****: To typologize data effectively, it is essential to consider the specific context and the intended purpose of the data. In general, it is helpful to differentiate between 1) structured data, 2) unstructured repetitive data, and 3) unstructured unrepetitive data.*

Data typologies are a crucial part of data exchange, essential to the data economy, particularly in identifying potential constraints for data capture and utilization. The most obvious example of such constraints is the legal framework governing the capture and utilization of personal data. Other typologies such as volunteered, inferred, and observed data, recently brought forward by Crémer et al (2019),[14] may be instrumental in the development of a framework to foster a thriving European data economy and so merit further investigation. We focus on a selection of data typologies that appear to be particularly relevant for the current debate around the socioeconomic impact of data and data's influence on competition in the digital era.

The most often used typology of data originates from the context of privacy, in particular the recently enforced EU General Data Protection Regulation (GDPR).[15] Here, the distinction is made between personal data and non-personal data.[16] This dichotomy is instrumental in many other contexts. Notably, it is also relevant for the Free Flow of Data Regulation,[17] which explicitly addresses non-personal data.[18] In the context of ever-increasing quantities of data and improvements in our ability to combine and analyze various datasets, one may doubt that this dichotomy still exists. In particular, pseudonymization and anonymization of datasets containing obviously personal data appear to be closer to a delusion than ever before. If anything, there is an arms race between technologies promising de-personalization of data and others reversing this

[14] Notably, Davis M, Martinez R, Kalaboukis C. 2010. Rethinking Personal Information – Workshop Pre-read, Invention Arts and World Economic Forum, Cologny first introduced this categorization with respect to the ways in which organizations can capture personal data. World Economic Forum. 2011. Personal Data: The Emergence of a New Asset Class, WEF, Cologny. Thus, the categorization was never meant to describe different "forms" of data, nor was it meant to be used beyond the realm of personal data.

[15] Regulation (EU) 2016/679.

[16] Personal data is defined as "*any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier, or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person*."

[17] Regulation (EU) 2018/1807.

[18] The European Commission defines non-personal data in the context of the Free Flow of Data Regulation in contrast to the personal data definition in the GDPR: *"Where the data are not 'personal data as defined in the General Data Protection Regulation, they are non-personal. The non-personal data can be categorized by origin as: Firstly; data which originally did not relate to an identified or identifiable natural person, such as data on weather conditions generated by sensors installed on wind turbines or data on maintenance needs for industrial machines. Secondly; data which were initially personal data, but were later made anonymous. The 'anonymization' of personal data is different to pseudonymization (see above), as properly anonymized data cannot be attributed to a specific person, not even by use of additional data16and are therefore non-personal data."* See European Commission (2019): Guidance on the Regulation on a framework for the free flow of non-personal data in the European Union. COM(2019)250 final. p. 5–6.

process. Toward the end of the 2000s, de-anonymization was often possible (Ohm 2010). A recent study by Rocher et al (2019) indicates that using 15 demographic attributes, 99.98% of Americans could be correctly reidentified in any dataset. Their results shed doubts on whether even heavily sampled anonymized datasets can satisfy modern standards for anonymization set forth, e.g., by the General Data Protection Regulation (GDPR). Thus, they also challenge the technical and legal adequacy of the de-identification release-and-forget model.

The typology of volunteered, inferred, and observed data, as it is put forward by Crémer et al (2019), also has some gray areas as the authors admit (p. 25). This typology does not qualify for developing a regulatory framework for the data economy for a few reasons. As illustrated by Hjørland (2019),[19] the same datum can be volunteered in one context of data capture, inferred from other data, or observed in yet another context. Also, this typology refers (as it was originally intended by Davis et al (2010)[20]) solely to the process of *how* data is captured. In light of the works by Drucker (2011) and Gitelman (2013), it may even be argued that all digital data should effectively be considered inferred data. Beyond such fundamental considerations of the validity of such a typology, it may be useful to consider the process of data capture with regard to the quality of data,[21] which has obvious implications for competition.[22]

There are numerous approaches to data typology, including but not necessarily limited to the following (e.g. Kitchin 2014):

- quantitative (numeric, discrete or continuous) and qualitative (nonnumeric) data.[23]
- nominal, ordinal, and interval data.
- primary, secondary, and tertiary data.
- captured, exhaust, transient, and derived data.
- indexical, attribute, and meta data.
- unprocessed and processed data.
- relational and multidimensional data.
- streamed, batched, and stored data.
- top secret, highly sensitive, sensitive, private and public (unclassified) data.

[19] See Section 2.2.

[20] Cf. World Economic Forum. 2011. Personal Data: The Emergence of a New Asset Class, WEF, Cologny; see also footnote 7 in the above.

[21] We discuss this further in the following section.

[22] We discuss these implications further in Chapter 4.

[23] In light of the discussion in Section 2.1, one may also consider documents here the superior descriptive term.

A useful typology of different kinds of data always depends on the specific context and the intended purpose. Nonetheless, we suggest, in line with the approach of information science, to start thinking of data first and foremost in terms of 1) structured data, 2) unstructured repetitive data, and 3) unstructured unrepetitive data (Salinas & Lemus 2017). Table 2-1 summarizes some key features of this fundamental data typology.

Table 2-1: Structured and unstructured data

| Data type | Characteristics | Scope | Processing | Store | User |
|---|---|---|---|---|---|
| Structured data (SD) | This data is capable of being represented by predefined structures (vectors, graphs, tables, among others).<br>The structure can be generalized. | This data belongs to the domain of traditional database systems and data warehouses. | This data can be stored by data structures such as tables or arrays and managed through widely distributed languages such as SQL. | This data is usually stored and managed through relational databases. | Business analysts |
| Unstructured repetitive data (US- RD) | This data does not have predefined structure, is recurrent in time, is generally massive. Not all of this type of data has a value for the analyses, so you can use samples or portions of these. | This data comes from electronic sensors whose objective is the analog analysis of the signal, such as: vital signs, seismic movements, positioning, biological and chemical processes, among others. | Generally, there are defined algorithms for the treatment of this type of data, like Fourier analysis for the signals. This type of data is susceptible to repetition and reuse. | This data is stored raw and free of context; this is done using NoSQL databases (document-oriented, key-value, among others) and flat files. | Data mining experts applied to different domains |
| Unstructured unrepetitive data (US- URD) | This data does not have a single structure. | It includes textual information, image analysis, dialog analysis, video content analysis, and string analysis. | The algorithms for processing this type of data are not reusable and the mere fact of predicting its structure is already a complex task.<br>Different processing is required depending on the type of data, such as natural language processing and computational linguistics techniques for text-type data. | They are stored raw and context-free in NoSQL databases and flat files. | Data Science Experts |

Source: Salinas and Lemus (2017).

While these three data types already imply some information about the quality of the data considered (see Table 2-1), there are many other aspects that have to be considered with regards to data quality. The following section provides a detailed account of such aspects and considers potential trade-offs between data quality and data quantity.

## 2.5 Data quantity versus quality?

> ***Insights***: *Finding the right balance between data quality and data quantity is a rather complex task. Depending on the context and purpose of data use, firms have to take into account several dimensions of data quality such as accuracy, completeness, and timeliness, with a minimum required quantity of data alongside diminishing benefits for increasing dataset sizes. Therefore, businesses using big data have to deal with several trade-offs.*

**Data quantity** appears to be at the heart of the popular discourse about data and its economic utilization. Thus, at least implicitly, an increase in the quantity of data that an organization can access—be it by capturing data, elevating data themselves or through a third party—is perceived to be a positive. More often than not, data quality is ignored, or simply taken for granted. This section reflects upon these assumptions and highlights potential interactions between data quality and data quantity.

In practice, **data quality** appears to be a much more pressing issue than data quantity. According to Schroeder (2016), most data scientists spend between 75% and 90% of their time cleaning, manipulating, transforming, and preparing data for analysis, and yet poor data quality still has a detrimental effect on the economy. IBM claims that the economic damage of poor data accounts to US$3.1 trillion annually in the US (Redman 2016). Poor data quality has a huge impact on businesses. For example, decisions based on poor data can lead directly to customer dissatisfaction, increased costs, and reduced employee job satisfaction, which can ultimately impact overall company performance and revenue (Haug et al 2011, Redman 1998).

In stark contrast to widely held beliefs, innovative algorithms belonging to the realm of artificial intelligence cannot overcome data deficiency. As Qi et al (2018) show, different anomalies in data require different adaptions in procedure, even for basic standard machine learning tasks like classification or clustering, which in turn are only possible with knowledge of the underlying process responsible for the anomaly. While many problems are related to missing or “dirty” data, solutions can vary substantially in complexity, which can render them infeasible with the currently available computing power. Their experiments show that in addition to the data itself, metadata, domain knowledge, and methodological expertise are equally important for firms utilizing data analytics involving machine learning.

It is clear that data quality is multidimensional (Pipino et al 2002, Sidi et al 2012, Taleb et al 2015, Taleb et al 2018, Wang & Strong 1996).[24] Researchers tend to concur with the seminal paper by Wang and Strong (1996) on what these dimensions are; they are summarized in Table 2-2, amended to select the most relevant dimensions by drawing on the results by Cichy and Rass (2019), who investigated the number of data quality frameworks supporting individual dimensions of data quality (see Figure 2-3).

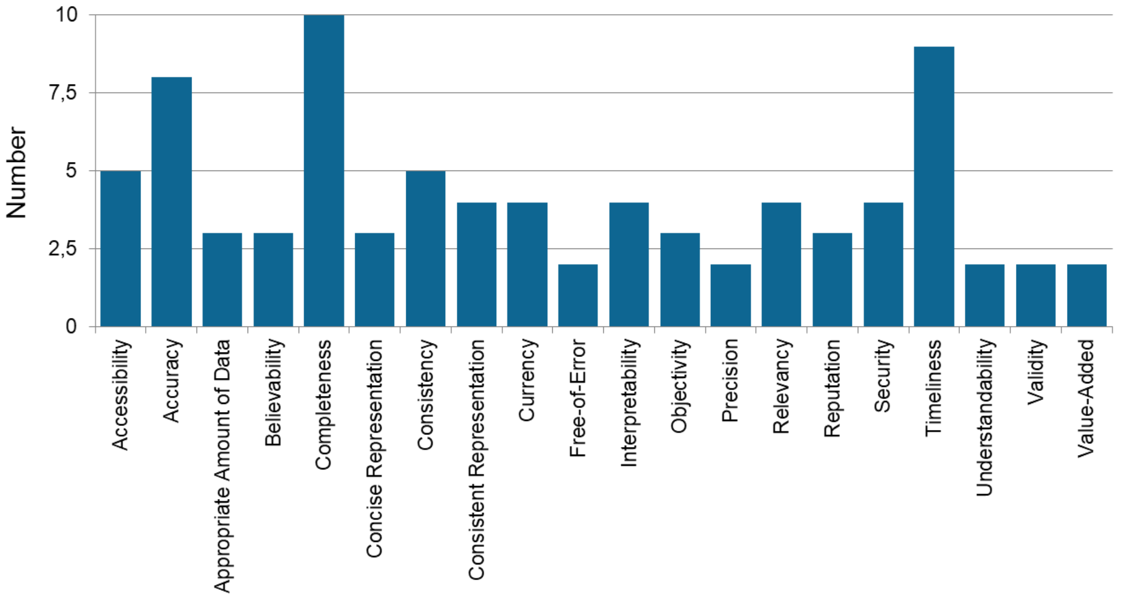
Table 2-2: Dimensions of data quality

| | |
|---|---|
| **Access security** | Access to data can be restricted |
| **Accessibility** | Data is available or easily retrievable |
| **Accuracy** | Data is correct and reliable |
| **Appropriate amount of data** | Amount of available data is neither too low nor too high |
| **Believability** | Data is considered true and credible |
| **Completeness** | Scope and level of detail of the data is adapted to the task |
| **Concise representation** | Data is stored in a compact yet complete form |
| **Ease of understanding** | Data is clear and easy to understand |
| **Interpretability** | Data is clearly defined and presented in the same language and unit |
| **Objectivity** | Data is unbiased and neutral |
| **Relevance** | Data can be used for a specific task |
| **Representational consistency** | Data is in the same format and compatible with previous data |
| **Reputation** | Sources of data have a high trustworthiness |
| **Timeliness** | Age of the data is adapted to the purpose |
| **Value-added** | Data provides added value |

Source: Adapted from Wang and Strong (1996).

[24] In total, more than 170 relevant dimensions for data quality have been identified in the academic literature. Their specific relevance depends on the context and the purpose of data utilization. Sidi F, Panahy PHS, Affendey LS, Jabar MA, Ibrahim H, Mustapha A. *2012 International Conference on Information Retrieval & Knowledge Management2012*: 300-04. IEEE. underscore this dependency in their overview of data quality dimensions that directly hinge on the specific task that shall be solved with the help of the data. Such dimensions could be broadly subsumed under the label "suitability" and comprise, for instance, usefulness, coverage, and data specification.

Figure 2-3: Number of data quality frameworks and corresponding data quality dimensions employed



Source: Cichy and Rass (2019).

Taleb et al (2015) and Taleb et al (2018) drew on these insights to devise an holistic data quality management model for the Big Data value chain. It covers eight stages from data inception to visualization. Data quality can be compromised at each of these stages. It is also obvious that the stages depend on each other and the data quality can only be as good as the weakest link in this chain. This implies that if there is little knowledge about the specific data capture process, an organization will likely face challenges in putting the data to use. The same is true if data transmission is unreliable, e.g., high packet loss rate. Lenart et al (2018) point out that even a single sensor can perform differently depending on the (sub-) task it is performing. Within a network of numerous sensors, establishing the credibility of a specific datum becomes even more difficult. Whilst consistent results across various sensors re-enforce the credibility of the data, conflicting information weakens it. The challenge is then to decide which of the sensors is not performing correctly and why.[25]

Within a given context and purpose of data analysis, **data quantity** can enable organizations to disregard or manipulate data that does not meet the quality criteria set for the task. More data usually goes hand-in-hand with providing more information and so the learning curve of a data-based business model shows a more exponential increase compared to that of traditional business models (Junqué de Fortuny et al 2013). This, of course, depends on the context of data capture and utilization.

[25] In research contexts, these questions refer to concepts of internal and external reliability as well as partly to internal and external validity of data.

However, the evolution of the concept of big data from a purely quantity-driven concept to a much broader concept accepted today suggests that quantity as such is strongly intertwined with data's qualitative characteristics. While for some time the predominant concept of big data was directly (and solely) linked to the quantity of data that is captured and analyzed or otherwise used, both academic and industry sources concur today that characteristics of big data may begin with quantity, but certainly entail many other factors.

To broaden the perspective on big data, de Mauro et al (2016) provide a definition that summarizes the complexity and richness of big data: "*Big data is the information asset characterized by such a high volume, velocity, and variety to require specific technology and analytical methods for its transformation into value*."[26] They define big data not only by quantity of data but also acknowledge the different types of existing data and the speed at which data is captured and analyzed. Furthermore, the definition recognizes the importance of technologies and methods necessary to make this data usable and valuable.[27] However, big data creates unique challenges regarding the quality of data. Specifically, with the large quantity of data generated, the high speed of the incoming data and the large variety of the data, imperfect data quality is assumed.

In fact, one of the major problems associated with big data is the problem of information noise. In this context, Liu et al (2016) conclude: "*although big data contain information in a […] detailed manner, they also record random variations, fluctuations, and even noise during the measurement. When applying […] machine learning to analyse big data, researchers can oftentimes run into the phenomenon of over-fitting, where the machine learning algorithm learns from the noise embedded in the fine-grained big data and predicts based on the noised information*" (p. 138). One of the major consequences is that false patterns or correlations may be recognized. Even without distortions in the dataset, due to the sheer amount of data, correlations can be found between individual variables in the dataset that should theoretically be uncorrelated (Boyd & Crawford 2012, Fan et al 2014). Liu et al (2016) point to another related problem, namely the origin of data. Data acquired from commercial data providers, for example, often lacks usefulness. They tend not to use scientific methods for the selection and collection of data. In most cases, the populations depicted are a small group of people who use a particular service or device (a biased sample). Studies that ignore this are tempted to deduce assumptions that do not apply to the entire population. This is particularly true for data originating from social media (Blank & Lutz 2017). Another central challenge associated with such data is that they are void of quality references. This suggests that "*data must be profiled and provided with certain quality information at the inception*

[26] There are studies that record further characteristics such value and veracity (see e.g. White).

[27] For a definition featuring the typical Vs of big data see e.g. Gandomi A, Haider M. 2015. Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management* 35: 137-44. Notably, Boyd D, Crawford K. 2012. Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. *Information, Communication & Society* 15: 662-79 provide a more critical account of big data than popular industry narratives. We cited it in Section 2.2 and refrain from reproducing it here.

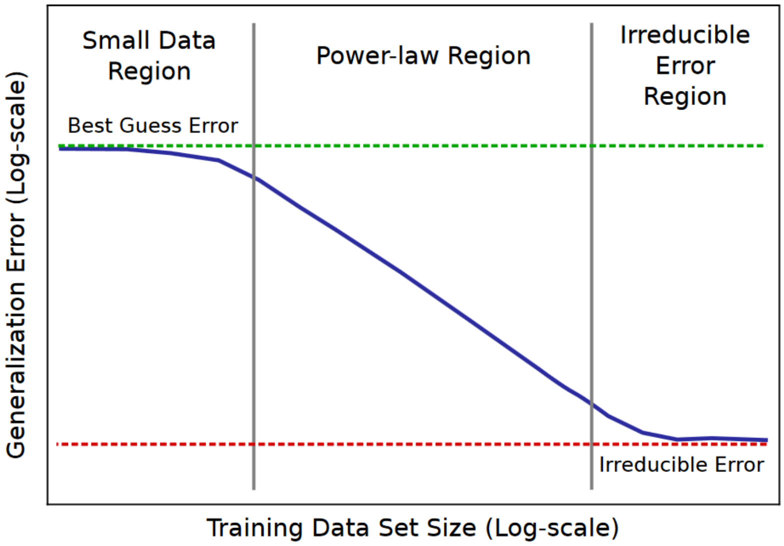
*phase. This also means that data attribute quality must be assessed, improved and controlled all along its lifecycle as it directly impacts the results of the analysis phase*" (Chen et al 2014).

Irrespective of the quantity and quality, it should be noted that "*Big data models are based on correlation rather than causation, so they cannot be extrapolated outside the range of the data. Moreover, it is difficult to separate data error, uncertainty, and measurement noise from actual phenomena. Plant data can be and has been used to tune first-principles models. A properly tuned first-principles (causation) model is always better than a correlation model developed from big data analytics*" (Saudagar et al 2019).

In relation to the amount of data required, Pipino et al (2002) point out that it should neither be too little nor too much. However, they refrain from providing a specific quantity of data, as the sufficient amount of data units will inevitably vary from one context and purpose to another. It is not necessarily obvious from the beginning exactly how much data is required to solve a specific task or problem. Also, with digitalization enabling access to potentially extremely large datasets, quantity may matter less today than it did at the beginning of the 2000s. Today, the public debate focuses on the minimum required quantity of data for specific tasks and whether there are diminishing returns to scale for data. While the popular narrative dictates that more data is always better, research points to diminishing benefits for increasing dataset sizes (Li et al 2016).

Recently, Hestness et al (2017) compared the impact of additional data on four common deep learning applications: (1) machine translation, (2) language modeling, (3) image processing and (4) speech recognition. They find a similar pattern for the effect of additional data on model accuracy improvements depicted in Figure 2-4. Their results suggest that there is indeed a minimum threshold of training data quantity for deep learning models to enter the power-law region of the learning curve. As training datasets become very large, the learning curve flattens. The authors describe this area as the irreducible error region. This lower bound originates from statistical (Bayes) error and also from issues relating to data quality such as mislabeled samples in the training or validation data. The exact exponent of the power-law region depends strongly on the context and purpose of the deep learning task as well as the data quality. Finally, they point out that larger training datasets can offset a large part of the accuracy potentially lost by lower-precision computation. However, data quantity cannot recover fundamental issues with data quality as suggested by the irreducible error region. Beyond these considerations, Schwartz et al (2019) raise concerns that ever-increasing training data unduly increases the carbon footprint of artificial intelligence (AI) and essentially renders advanced applications of AI as an activity of a select group of elite universities and large corporations.

Figure 2-4: Sketch of power-law learning curves



Source: Hestness et al (2017).

Beyond these considerations, (Schwartz et al 2019) raise concerns that ever-increasing training data unduly increases the carbon footprint of artificial intelligence (AI) and essentially renders advanced applications of AI an activity of a select group of elite universities and large corporations. With this in mind, researchers engaged in training algorithms should optimize the size of the datasets employed.

## Key Findings of Chapter 3

- *Data permeates virtually all economic activity.*
- *The data economy is not a sector that can be neatly delineated from other (traditional) sectors.*
- *Value is created in circles and networks rather than sequential interactions along a value chain.*
- *The process of value creation, starting with data, can take place entirely within a single undertaking or can involve many actors.*
- *There is a trade-off between decreasing costs of capturing data and increasing costs of effective control over the data.*
- *Businesses have to go through substantial organizational changes in order to reap the full benefits of (big) data.*
- *Big data use can enable pattern recognition, predictive analytics, and new insights.*
- *Data in itself does not make a digital platform successful. There are other factors playing an equally important role.*
- *The role of data in innovation is threefold: (1) data can be a driver of product and process innovations; (2) data is an integral part of knowledge-capturing product innovations; and (3) continuous real-world data can contribute new means to measure innovation activity and success in official statistics.*
- *The major challenge for utilizing data in innovation activity is typically not access to data as a first or third party, but overcoming the technological and process-inherent challenges in utilizing the data.*
- *The role of data within innovation differs between “data-poor” and “data-rich” contexts.*

# 3 Value creation from data

The value creation from data is naturally at the heart of the data economy. In this chapter, we first shed light on the basic functioning of the data economy which, due to the fluidity of data, differs from traditional concepts of economic value creation. We continue with an exploration of actual business use of data and the various challenges that have to be overcome in order to reap the full benefits of data utilization. Finally, we explore data's role for innovation.

## 3.1 The data economy

> ***Insights**: Due to the fluidity of data, the data economy consists of numerous complex interactions between market actors. As such it does not fit traditional value chain concepts, but instead follows a circular logic with various circles of value creation being intertwined into value networks.*

As explained in the preceding chapter, information of economic value can be extracted from captured and/or elevated data or a combination thereof through generating insights by means of data analysis. Enabling this core of value creation from data, a number of surrounding service and infrastructure providers exist, which support the processes of data capture, transmission, storage, analysis, and exploitation of insights.[28]

This process happens at different levels of sophistication in all businesses in all sectors. Thus, data utilization transcends traditional sector boundaries as, with increasing digitalization, the capabilities formerly (largely) contained in the information and communication technology (ICT) sector permeate virtually all sectors.[29] As a consequence, narrow definitions of the data economy such as the one suggested by the DataLandscape[30] project and adopted by BEREC (2019)[31] likely fall short of reality.

[28] For further descriptions of these economic activities and their interactions see e.g. BVDW. 2018. Datenwertschöpfung und Qualität von Daten, Bundesverband Digitale Wirtschaft (BVDW) e.V., Düsseldorf,GSMA. 2018. The Data Value Chain, GSMA, Curry E. 2016. The Big Data Value Chain: Definition, Concepts, and Theoretical Approaches In *New Horizons for a Data-Driven Economy - A Roadmap for Usage and Exploitation of Big Data in Europe*, ed. JM Cavanillas, E Curry, W Wahlster, pp. 29-38: SpringerOpen, Attard J, Orlandi F, Auer S. *International Conference om Theory and Practice of Electronic Governance, New Delhi, India, 2017*: 475-784..

[29] As Calvino F, Criscuolo C, Marcolin L, Squicciarini M. 2018. A taxonomy of digital intensive sectors, Organisation for Economic Co-operation and Development, Paris show, sectors of the economy vary as regards their investment in and likely capacity of utilizing digital technologies and data. Similar findings are reflected in Arnold R, Schiffer M, Pols A. 2013. Wirtschaft Digitalisiert - Welche Rolle spielt das Internet für die deutsche Industrie und Dienstleister?, IW Consult and BITKOM, Cologne, Berlin.

[30] http://datalandscape.eu/ The website, as well as the reports and data presented there, are part of the EU Data Market study contracted by the European Commission under *SMART* 2013/0063.

[31] The term "Data Economy" encompasses the (increase in the) availability of data, the related business opportunities, as well as the (potential) social value of the insights that can be generated. According to

Just as traditional sector boundaries fail to capture the data economy, the traditional concept of the value chain is equally not applicable to the data economy. As described in Section 2.3, data has many economic properties that are not concurrent with traditional concepts of exclusivity, consumption, or product value. Most notably, data is non-rivalrous in use and intangible and so, the data economy should not be thought of as a typical value chain at whose end a product is purchased and consumed or used.[32] In line with Arnold and Waldburger (2015), the data economy should be thought of a circle of value creation (see Figure 3-1).

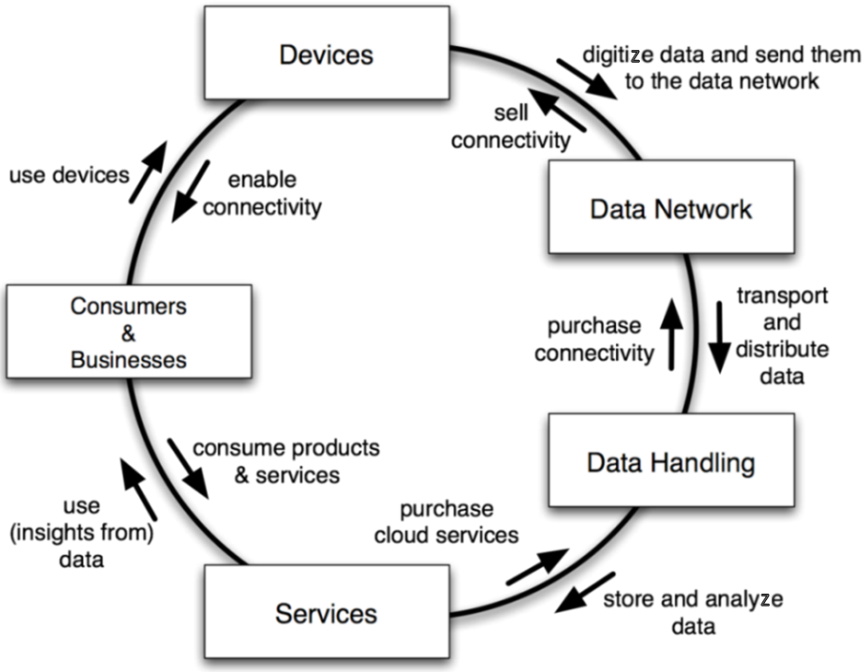
In their concept called the "data value circle," users (i.e., consumers and/or businesses) are both the end point and the starting point of value creation. On the one hand, they are the sources of data that enable value creation over the course of the circle. On the other hand, they are ultimately the recipients of services and products based on this data. In between, there are various stages of value creation. First, there are digital devices that capture data.[33] Then, the data is transmitted via communications networks. The data handling stage comprises all services that deal with the storage, analysis, and other processing of data. Within the service level, the information generated by data analytics is used to improve or more effectively design existing processes, products, and services, or to develop new ones.

the EC report Building a European Data Economy, the *"data economy measures the overall impacts of the data market—i.e. the marketplace where digital data is exchanged as products or services derived from raw data—on the economy as a whole. It involves the generation, collection, storage, processing, distribution, analysis, elaboration, delivery, and exploitation of data enabled by digital technologies. A key development in the data economy in recent years has been the increase in the variety and volume of data being generated through online activities"* (p. 7). See also: European Commission. 2017. Communication from the Commission Commission to the European Economic and Social Committee and the Committee of the Regions - "Building a European Data Economy" (SWD(2017) 2 final), European Commission, Brussels.

[32] The traditional value chain concept is usually traced back to Porter ME. 1985. *Competitive Advantage: Creating and sustaining superior performance*. New York: The Free Press.

[33] As already pointed out in Section 2.1, data capture can also happen through software. While devices remain the ultimate interface of data input, the software in the background runs the processes necessary to digitize, pre-process, and store data. As such, access to the software may suffice in order to gain access to the data captured by a device.

Figure 3-1: Data value circle



Source: Arnold and Waldburger (2015).

The data value circle illustrates the central components and interactions of various organizations across the five stages. However, all stages in the data value circle can be part of one and the same undertaking. Obviously, large online platforms have entered various stages of the data value circle as they have stretched their presence from devices (smartphones, tablets, set-top boxes, etc.) to data networks, including undersea cables, and over data handling with sophisticated cloud infrastructures and services for consumers and businesses.

Large companies from traditional industrial sectors have behaved in similar ways. They have also integrated various stages of the data value circle into their business models. For instance, original equipment manufacturers (OEMs) from the automotive sector have integrated connectivity into their cars, built their own data centers and data analytics tools and generate an increasing share of their revenues from services. We observe similar trends with agriculture, machinery, and chemicals.[34]

[34] The increasing share of service-related value creation in traditional industrial sectors has been observed since the second part of the 2000s. Data, digitization and digitalization have enabled this process. Early accounts of this trend can be found in Kempermann H, Lichtblau K. 2012. Definition und Messung von hybrider Wertschöpfung. *IW Trends* 39: 1-20 and Lichtblau K, Arnold R. 2012. Smart Industry – Intelligente Industrie: Eine neue Betrachtungsweise der Industrie. Ergebnisse einer Studie der Institut der deutschen Wirtschaft Köln Consult GmbH für das Land Hessen, Initiative Industrieplatz Hessen, Neu-Isenburg.

Without any presumption about the nature of the data that is captured by devices, the data value circle sheds light on the concept of data equity used by Carrière-Swallow and Haksar (2019).[35] Any organization that decides to capture or elevate data for their purposes will likely have an interest in controlling access to this data. Their control may decrease as other organizations can equally use the same (or sufficiently similar) devices to capture data for their purposes. As a contemporary device will likely be a combination of parts supplied by various firms and equally a combination of software by various providers, it is also likely that each of these organizations may be able and interested in capturing and utilizing some of the data that the device captures.

Capturing and integrating the data from devices across multiple supply chains is at the heart of the vision of Industry 4.0. While this vision promises substantial economic gains, decreasing control by organizations over their data and a lack of trust among competitors may eventually thwart its success. Ultimately, the successful achievement of Industry 4.0 hinges on whether the cost of data capture and the economic value of the information generated can offset the costs incurred by the increasing complexity of control over the data for the organizations involved.[36]

Value networks present another integral part of the Industry 4.0 vision. They are enabled by the fluidity of data and can be thought of as the combination of various (parts of the) data value circle(s). Attard et al (2016, 2017) illustrate this in their concept of data value creation: The Data Value Network (DVN).[37] A DVN is defined as *"a set of independent activities having the aim of creating value upon data in order to exploit it as a product where different actors […] can participate by executing one or more activities […] and each activity can consist of a number of actions or value creation techniques […]. In turn, each action can consist of one or more data value chains, since they might need a series of processes to be executed in order." (Attard et al 2017)*.

In essence, data as such has little or no value. Instead, it can be turned into information carrying economic value which has to be utilized within business models. The following section takes a closer look at such data-driven business models.

[35] The view of Carrière-Swallow Y, Haksar V. 2019. The Economics and Implications of Data - An Integrated Perspective, Washington, DC remains limited to personal data revolving around the data subject and potential further utilization of personal data by the data collector and the data processor.

[36] It is conceivable that eventually, costs of protecting data access effectively may outrun the potential expected return from capturing the data in the first place. So, there is also a possibility that the growth observed in the data economy is not necessarily a one-way street.

[37] Features of DVN are: non-tangible data product, non-sequential, multiple actors, nested value chains, recurring value network, independent activities.

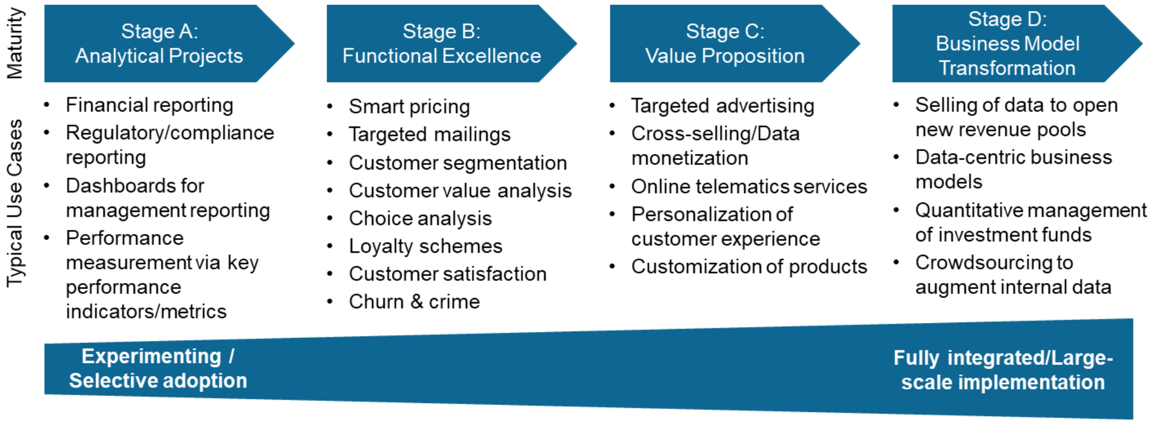
## 3.2 Business models in the data economy

> ***Insights**: Utilizing data to the benefit of the business requires significant effort on the business' side that goes well beyond data access. In fact, to reap the full benefit from data, businesses have to undergo substantial transformation up to a full transformation of their business model.*

Digitalization has provided businesses with new opportunities to improve their conduct and access to data is central to this, with leaders in digitalization tending to be leaders in data utilization. The OECD (2015) survey on ICT usage provides some business rationales and potential impacts of data analytics in firms. An organization can utilize data analytics for (1) the identification of potential customers, (2) to increase customers' spending by targeting offers and discounts, (3) to tailor products (goods and services) to customers' needs and (4) to gain effectiveness in internal production and/or organization. The (positive) impacts of data analytics for businesses are thus likely a mix of (1) potential cost savings, (2) sales growth and (3) enhancements in business organization.

In a similar vein, Grover et al (2018) identify four main goals for the use of big data in companies. First, big data can create value by improving organizational decision-making. This can be accomplished by providing broad and consistent access to data across an organization, complemented with empowerment structures to act on the data, or through decision models that augment human decision-making or are built into business processes. Second, big data can create value by improving the effectiveness, efficiency, and productivity of business processes, which leads to better execution and less time spent on process breakdowns. Third, big data can create value for product and/or service innovation. Fourth, big data can also deliver a better customer experience and more competitive services, resulting in higher customer satisfaction and retention. The maturity model for big data developed by Tiefenbacher and Olbrich (2015) goes even further as it commends (new or augmented) value propositions and even full business model transformation as stages within the use of big data in businesses (see Figure 3-2).

Figure 3-2: Maturity model for big data

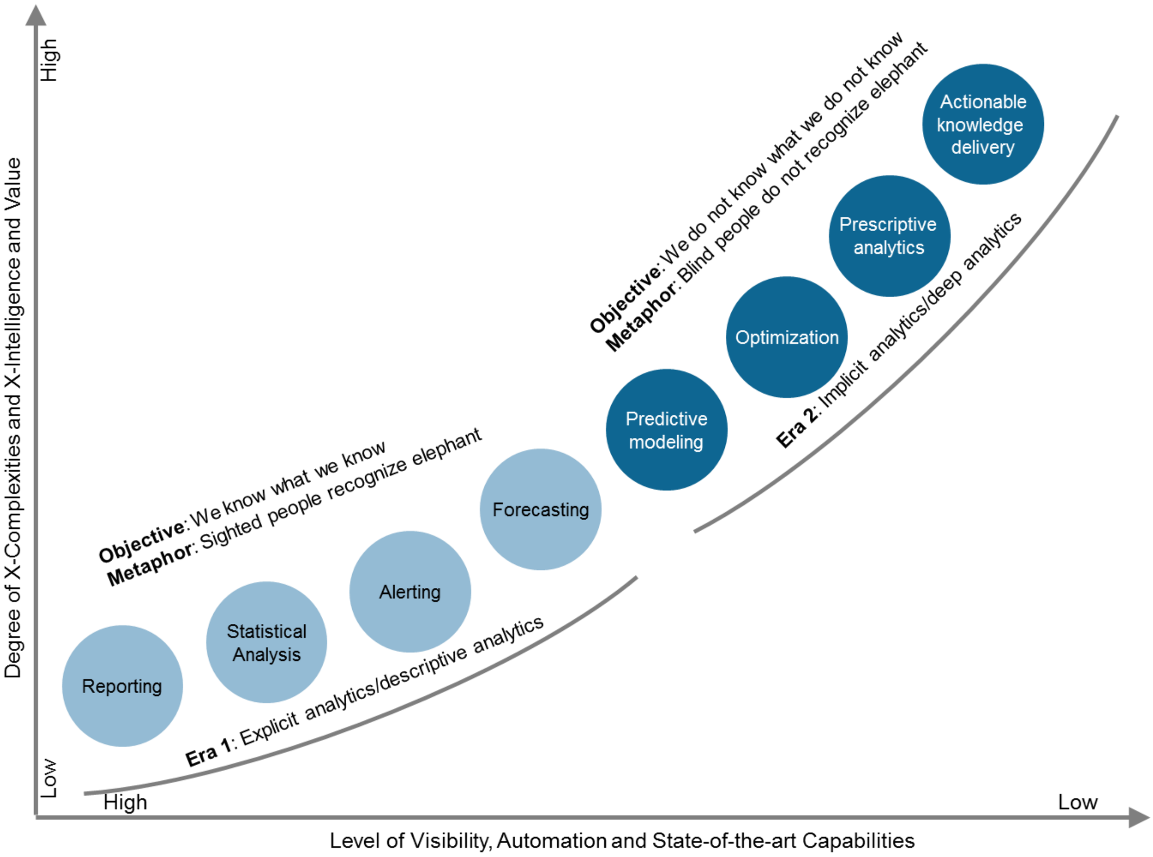

Maturity
Typical Use Cases
Stage A: Analytical Projects
Stage B: Functional Excellence
Stage C: Value Proposition
Stage D: Business Model Transformation
• Financial reporting
• Regulatory/compliance reporting
• Dashboards for management reporting
• Performance measurement via key performance indicators/metrics
• Smart pricing
• Targeted mailings
• Customer segmentation
• Customer value analysis
• Choice analysis
• Loyalty schemes
• Customer satisfaction
• Churn & crime
• Targeted advertising
• Cross-selling/Data monetization
• Online telematics services
• Personalization of customer experience
• Customization of products
• Selling of data to open new revenue pools
• Data-centric business models
• Quantitative management of investment funds
• Crowdsourcing to augment internal data
Experimenting / Selective adoption
Fully integrated/Large-scale implementation

Source: Tiefenbacher and Olbrich (2015).

When it comes to (big) data value realization, there are two socio-technical features that shape how organizations realize value from data, that is portability and interconnectivity (Günther et al., 2017). Portability refers to the ability to access and to transfer digital data from one context of application to be used in another context. Thus, data can be transferred and accessed across platforms and organizational boundaries. Interconnectivity refers to the ability to combine data from various sources, thereby enabling decision makers and analysts to arrive at more insights by exploring links, patterns, and relationships. As a consequence, new value propositions can be discovered by the use of big data.

Cao (2017) focuses closely on the specific value that data analytics can add in businesses. Explicitly, he distinguishes two eras. There is the era of explicit and largely descriptive analytics, which (still) relates to most organizations' data analysis capabilities. The era of implicit and deep analytics on the other hand extends the capabilities of data analytics beyond the descriptions of things already known and into the why and how of real-word phenomena. According to Cao (2017), insights gained from implicit and deep analytics can be used to determine the next-best or worst situation compared to the current state, enabling optimal intervention strategies within or across organizations to be devised. In light of the discussion we presented in Sections 2.2 and 2.5, these expectations seem somewhat overblown, as they at least implicitly support a theory-free generation of insights from data, which is not possible. Nonetheless, the evolutionary path of data analytics in general, shown in Figure 3-3, highlights an important point in relation to the visibility of data analytics, where visibility decreases with increasing complexity, a result that is somewhat counter-intuitive. In other words, the more sophisticated and intensive the use of data analytics, the more opaque the actual insight generation becomes. Algorithms become increasingly complex and the degree of automation of data analysis increases with deployment of tools like deep learning. Notably, a linear evolutionary path of data analytics within any organization is unlikely as Cao (2017) points out that there are many hurdles on the way resulting in backward and forward iterations, including a sizable number of trial and error cycles within data science teams.

Figure 3-3: Explicit to implicit analytics spectrum and evolution


High
Low
Degree of X-Complexities and X-Intelligence and Value
High
Low
Level of Visibility, Automation and State-of-the-art Capabilities
Objective: We know what we know
Metaphor: Sighted people recognize elephant
Objective: We do not know what we do not know
Metaphor: Blind people do not recognize elephant
Reporting
Statistical Analysis
Alerting
Forecasting
Predictive modeling
Optimization
Prescriptive analytics
Actionable knowledge delivery
Era 1: Explicit analytics/descriptive analytics
Era 2: Implicit analytics/deep analytics

Source Cao (2017).

Naturally, all of these (positive) outcomes only come to fruition if the organization is able to purposefully unlock the information value from the data it captures through data analysis and elevation. Complex organizational changes are required to make full use of (big) data within a business, according to McKinsey (2016). Specifically, the authors describe five stages: (1) development of use cases; (2) building a data ecosystem; (3) acquisition of analytic capabilities needed to derive insights from data; (4) changing business processes to incorporate data insights into actual workflow; (5) building capabilities of executives and mid-level managers to understand how to use data-driven insights within the current business model of the firm or with the aim of developing a new business model. In order to reap the full benefits of (big) data, Günther et al (2017) point out that reaching functional excellence within the organization can only be the first step to business model transformation.

Transforming (parts of) the business into a digital platform can be one possible aim of business model transformations outlined in the above. As digital platforms also form a focal point of the current public debate around data access and utilization, they merit some further elaboration here and we focus on the role that data plays within digital platforms.

This role critically depends on the specific concept of digital platforms. Three main concepts of digital platforms typically form part of the public debate: (1) digital platforms as an economic concept, (2) digital platforms as a technological platform and (3) digital platforms as aggregators and curators of (media) content. Due to its stages being organized in a circle, the data economy lends itself naturally to the first concept of multi-sided digital platforms.[38]

According to the economic literature on platforms as an economic concept (Armstrong 2006, Caillaud & Jullien 2003, Evans & Noel 2005, Evans & Schmalensee 2007, Hagiu 2007, Rochet & Tirole 2003, Rysman 2009), they share the following characteristics:

(1) Digital platforms act as intermediaries, enabling the interaction of different user groups. Digital platforms' services (mediation, transaction, exchange, comparison) are oriented toward the behavior and usage patterns of the respective user group. A direct involvement of the platform in the interaction is not mandatory. Capturing and utilizing data are however clearly necessary to fulfill this key function of a digital platform.

(2) These interactions are characterized by so-called direct and indirect network effects. As outlined in Section 2.3, these effects likely augment the ability of the firm to gather an increasing quantity and potentially better quality of data which, in turn, can be exploited to further improve the services offered as part of the digital platform, especially with a view to tailor them to specific target groups as mentioned in (1).

(3) The pricing of digital platforms is usually a function of the price elasticities of the demand of different user groups of the platform. The elasticities also reflect the mutual dependencies of the platform sides. The side(s) of platforms with low price elasticity tend to pay a relatively high price and thus predominantly pay for the costs of the respective digital platform, while the side(s) of platforms with a high price sensitivity tend to pay a low or zero price for platform use (i.e., cross-subsidy). Remuneration can be in terms of money, data and attention to advertisements.

[38] This is further elaborated in Arnold R, Waldburger M. 2015. The Economic Influence of Data and their Impact on Business Models In *Trends in Telecommunication Reform 2015 - Getting Ready for the Digital Economy*, ed. ITU, pp. 153-83. Geneva: International Telecommunication Union.

Digital platforms as technological platforms are essentially defined with reference to their functionalities. The platform represents a uniform technical basis on which components can be connected and with which (software) programs can be operated. In general, a distinction can be made between a hardware platform and a software platform: a hardware platform (also called machine level) consists of a computer architecture and the underlying command and processor structures. In contrast, a software platform (also called application level) forms the basis on which operating systems and application programs can operate. In this sense, a digital platform plays a critical role in enabling data capture in standardized ways. Such technological digital platforms and the data they capture can be an integral building block of a business model, especially if they provide first-party access to data that can create information of economic value.

The concept of digital platforms as content aggregators is based on the platform concept of the media industry. Their key functionality is providing digital transmission capacities or digital data streams (including from third parties) with the aim of making these offers accessible as a complete package to users of the service. They also decide on the selection for the aggregation. The quality of the platform is therefore essentially determined by how a company decides on the compilation of the content. Data is therefore an integral part of the business model in such platforms.[39]

Throughout the three types of digital platforms, it emerges that while data plays a role in each case, this role differs substantially depending on the context. Furthermore, it is obvious that besides data access, other factors of business model development decide over the success or failure of the business.[40] Innovation is a key ingredient for long-term success and is the focus of the following section.

[39] We elaborate this in Section 3.3.

[40] We elaborate this further in Arnold R, Bott J, Hildebrandt C, Schäfer S, Tenbrock S. 2016. Internet-basierte Plattformen und ihre Bedeutung in Deutschland, Wissenschaftliches Institut für Infrastruktur und Kommunikationsdienste (WIK), Bad Honnef and Arnold R, Hildebrandt C. 2017. The Socio-Economic Impact of Online Platforms, Wissenschaftliches Institut für Infrastruktur und Kommunikationsdienste (WIK), Bad Honnef.

## 3.3 Innovation in the data economy

***Insights:*** *The role of data in innovation is threefold: (1) data can be a driver of product and process innovations; (2) data is an integral part of knowledge-capturing product innovations; and (3) continuous real-world data can contribute new means for measuring innovation activity and success in official statistics. The major challenge for utilizing data in innovation activity is typically not access to data as a first or third party, but overcoming the technological and process-inherent challenges to utilize them.*

Innovation researchers concur that innovation activity and, ultimately, successful innovations depend on complex interactions between private and public agents; this is considered to form an "innovation system".[41] This inherent complexity renders it almost impossible to pinpoint specific drivers for successful innovations. Causal links between a specific input and output are even more elusive.

Nonetheless, it is obvious that data has been and will remain integral to virtually all innovation activity. Consequently, OECD and Eurostat, in their latest edition of the Oslo Manual[42] (OECD & Eurostat 2018), point out that there is no added value for innovation statistics in monitoring whether data was involved in a specific innovation or not. Instead, the OECD and Eurostat recommend measuring the general digital capability of each enterprise and correlating this with its innovation activity and success. This recommendation points to an important insight for the present debate around data access and sharing. Data is an integral part of any innovation activity, but can only add value if the firm is sufficiently capable of using it. Consequently, data access is a necessary, but not sufficient condition for fostering innovation activity and, ultimately, competition for the best products and services.

Besides other factors of digital capability, the Oslo Manual recommends the use of indicators of an enterprise's capability regarding *"access to and ability to use data analytics to design, develop, commercialize and improve products, including data about the users of the firm's products and their interactions with such products."* (p. 123) At least the "access to" data critically depends on the specific context in which the enterprise is active, as the **data-richness** varies strongly between contexts (see Section 2.5). Notably, *"the type of data needed differs across sectors and often across specific sectoral applications. The availability and access challenges as well as data*

[41] *"Innovation systems is not an economic theory in the same sense as neo-classical or evolutionary economics rather the concept integrates theoretical perspectives and empirical insights based on several decades of research. Within this approach, innovation is both a cumulative, path-and context-dependent process, and an interactive process."* see Lundvall B-Å, Borrás S. 2005. Science, technology and innovation policy In *The Oxford handbook of innovation*, ed. J Fagerberg, DC Mowery, RR Nelson, pp. 599-631. Oxford: Oxford University Press cf. Bertenrath R, Arnold R, Koppel O, Lang T. 2011. Innovation Policy and the Business Cycle: Innovation Policy's Role in Addressing Economic Downturn - INNO-Grips Policy Brief No. 1, European Commission, Cologne/Brussels.

[42] The Oslo Manual is the go-to source of information on how understand and measure innovation in official statistics.

*quality and the ease of integrating multiple databases also differ"* (Paunov & Planes-Satorra 2019). Fundamentally, we can distinguish data-rich from data-poor contexts.

In **data-rich environments** such as online services, an enterprise with innovation activity can typically access a wealth of data (continuously) captured as part of the largely digital interactions with the users of its own services. Through cookies, device or browser fingerprinting, enterprises can even gain insights about consumer behavior with online services other than their own.[43] With regards to innovation activity in the context of online services, a particularly relevant impact of digitalization and data is A/B testing of (new or improved) features in (pseudo) experiments. Given that A/B testing, and other innovation activities relying on data, can be conducted at negligible cost in data-rich contexts, alternative innovation activities are both more costly and less effective. A firm competing in a data-rich context will have to make data an integral part of its innovation process. This is neither costly nor difficult. However, it will likely be critical that the enterprise can capture specifically the data that it requires. It is questionable whether such a firm would be able to gain a substantial advantage from accessing the data captured by another (competing) firm.[44] Consequently, enterprises in data-rich environments need to collect their own data for their innovation activity, because they have to have full control over the source and origin of the data and the devices (or rather the measurement tools) that capture the data.[45] Moreover, the (added) value of third-party data access appears limited.

In **data-poor environments** such as logistics and transport or waste and recycling, the task of accessing data is exponentially more complex and costly than in data-rich contexts, since many interactions among actors still happen offline and few processes are digitized. Consequently, in data-poor environments, creating data access can be considered an innovation activity of a firm in itself. Also, while the costs of accessing data in the first place will be higher in such contexts, the returns of an innovation that introduces a novel level of data access to the market may be substantially higher, too. In fact, such innovations may be disruptive to a traditionally data-poor sector. Consequently, an enterprise with innovation activity involving new means of data access likely has a strong incentive to protect its data as they probably constitute the competitive advantage of the innovation itself.

[43] There is abundant literature covering the various opportunities to track consumers online. For an overview see Boerman SC, Kruikemeier S, Zuiderveen Borgesius FJ. 2017. Online Behavioral Advertising: A Literature Review and Research Agenda. *Journal of Advertising* 46: 363-76. For a perspective on how these approaches are integrated into business models, see Hildebrandt C, Arnold R. 2016. Big Data und OTT-Geschäftsmodelle sowie daraus resultierende Wettbewerbsprobleme und Herausforderungen bei Datenschutz und Verbraucherschutz - WIK-Diskussionsbeitrag Nr. 414, Wissenschaftliches Institut für Infrastruktur und Kommunikationsdienste (WIK), Bad Honnef or Bott J, Hildebrandt C, Arnold R. 2018. Die Nutzung von Daten durch OTT-Dienste zur Abschöpfung von Aufmerksamkeit und Zahlungsbereitschaft: Implikationen für Wettbewerb, Regulierung sowie Daten- und Verbraucherschutz - WIK-Diskussionsbeitrag Nr. 431, Wissenschaftliches Institut für Infrastruktur und Kommunikationsdienste (WIK), Bad Honnef.

[44] A/B testing is a good case in point. Obviously, an enterprise can gain valuable insights for an improvement of its own offering based on data captured about its users. The value of insights gained from data about users of other services will likely be lesser or even zero for the same purpose.

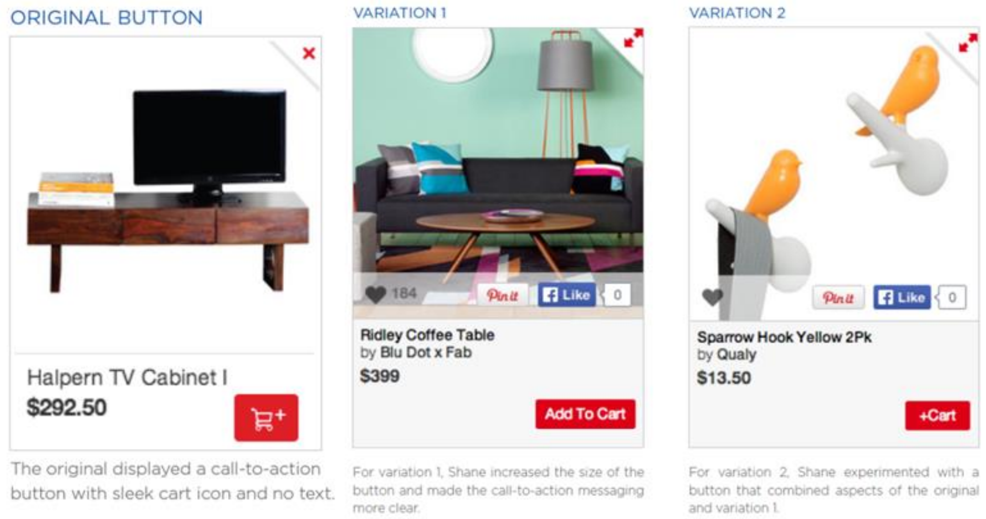
[45] This tends to be a software rather than a hardware device.

## Innovation in data-rich environments

In data-rich contexts, innovations driven by data tend to be incremental. This is mainly due to the fact that companies have easy access to large amounts of data, which in turn reduces the cost of constantly and innovatively adapting to even the smallest market changes. Furthermore, especially in the knowledge-base, de-materialized digital economy, even minor innovations have a crucial effect on entrepreneurial success and competitiveness. One of the central methods that companies use in data-rich contexts to innovate is A/B testing.

A/B testing is often used by website owners to constantly improve their website design to increase user engagement. For example, Fab (https://fab.com/) used A/B testing in order to test different types of “Add to Cart” buttons. Fab is an e-commerce platform for both selling and buying apparel, home goods, accessories, and more. The testing included three types of buttons, one icon-focused button and two that were mostly text-based (Figure 3-4).

Figure 3-4: Example of A/B testing



The original displayed a call-to-action button with sleek cart icon and no text.

For variation 1, Shane increased the size of the button and made the call-to-action messaging more clear.

For variation 2, Shane experimented with a button that combined aspects of the original and variation 1.

Changing the button had a non-negligible influence on click-rates. Variation 1 increased performance by 49%.[46] AutoScout24, an online marketplace for new and used cars, also applied A/B testing to increase their form's conversion rate. Again, several designs were tested against the original. The result was that the new variants were superior to the original in terms of conversion rate.[47]

[46] A comprehensive description of the case can be found at https://s3.amazonaws.com/optimizely-marketing/customer-stories/fab-casestudy.pdf [Last accessed 22/10/2019].

[47] A comprehensive description of the case can be found at https://www.optimizely.com/de/customers/autoscout24/ [Last accessed 22/10/2019].

## Innovation in data-poor environments

One sector that can still be classified as data-poor is the logistics sector. However, particularly innovative firms in this sector have realized the value that digital data can have. These firms focus their innovative activity on creating digital data representing formerly analog processes with predominantly manual input. One example of such a firm is **Carrypicker** (https://www.carrypicker.com/). Carrypicker is a start-up aiming to increase efficiency of freight forwarding companies. A major problem in this area is that the entire planning process, despite its high complexity, is still carried out manually, even in the largest freight forwarders. In order to achieve revenue maximization while simultaneously reducing idle capacities, the latest methods from the fields of artificial intelligence, machine learning, and predictive analytics are used. The objective is to develop a dynamic online pricing platform and route optimization and assignment.[48]

**Cirplus** (https://www.cirplus.io/) is another start-up seeking to innovate data access in a data-poor sector. Their focus is on the recycling industry and they are currently developing a B2B marketplace for recycled plastic. Not only is the recycling rate for plastic packaging extremely low, but also most disposable products are made of different plastic materials, making it difficult to recycle and produce high-quality recyclates. At the same time, the digitalization of this industry is still in its infancy, which makes the process more difficult for both buyers and sellers. The independent software service provider has therefore made it its mission to network the entire plastics and waste disposal industry—not only providing a platform, but also introducing a new standard for labeling and the exchange of information on recyclates.[49]

Source: WIK-Consult.

Having established the fundamental differences in the roles that data plays depending on the data-richness of the innovation context, it is useful to zoom in on the specific impact of data on innovation. First and foremost, increasing digitalization enables innovation. In line with the data value circle introduced at the beginning of this chapter, Paunov and Planes-Satorra (2019) point out, *“smart and connected devices are a rich source of innovations across all sectors. They gather and transmit data on processes, use, and environmental conditions, allowing for process optimization, predictive analytics/diagnostics and in their most advanced stages the autonomous operation of products as would be the case for self-driving cars”* (p. 11). Furthermore, this data can be an enabler of additional innovations as it offers new ways of differentiating products and services or altering the value proposition of suppliers altogether, i.e., turn them from

[48] A comprehensive description of the case can be found at https://www.bmvi.de/SharedDocs/DE/Artikel/DG/mfund-projekte/carrypicker.html [Last accessed 22/10/2019].

[49] For more information see https://www.capital.de/wirtschaft-politik/dieses-start-up-koennte-die-recyclingindustrie-revolutionieren [Last access 22/10/2019], https://www.cirplus.io/pilot-program [Last accessed 22/10/2019].

producers into service providers.[50] Finally, the fluidity of data enables them to “seep out” of formerly contained sector silos. This enables new entrants to capture a significant part of the value created in a sector, if only they are able to access and utilize data in a more innovative way than established services, products, or processes in the sector.

**Case in point – Uber**

Uber is an obvious example of a company that exploits existing infrastructures and technologies to provide a service that has been around for decades in an innovative way. Uber utilized the existing infrastructure of devices such as smartphones and data transmission via telecommunications networks to establish an interface for data exchange between drivers and people seeking transport. This interface enables the digital completion of service delivery.

Beyond the continuous data stream from devices, Uber has also drawn on existing and largely static data, such as maps. This combination led to reduced waiting and driving times. As the business started growing, Uber replaced or at least augmented, public data by their own data that they were able to capture due to substantial investment in new technologies.

Beyond these three basic impacts of digital data on innovation, their effects can be observed in different types of innovation as defined in the Oslo Manual (OECD & Eurostat 2018): Fundamentally, innovations of business enterprise can be split into **product innovations** and **process innovations**. Both types are further split into sub-types, which will be discussed in the following paragraphs, focusing on the impact that digital data has within each of them.

The Oslo Manual (OECD & Eurostat 2018) distinguishes two types of product innovation: innovations in (1) goods and (2) services.[51] Notably, the two types can be difficult to delineate, e.g., considering rental of durable goods, bundling of goods and services or the inclusion of insurance in purchases of goods. We already discussed how

[50] This is also known as hybrid value creation. See Kempermann H, Lichtblau K. 2012. Definition und Messung von hybrider Wertschöpfung. *IW Trends* 39: 1-20 and Lichtblau K, Arnold R. 2012. Smart Industry – Intelligente Industrie: Eine neue Betrachtungsweise der Industrie. Ergebnisse einer Studie der Institut der deutschen Wirtschaft Köln Consult GmbH für das Land Hessen, Initiative Industrieplatz Hessen, Neu-Isenburg.

[51] The following definitions apply: Goods: *“Goods include tangible objects and some knowledge-capturing products (see below) over which ownership rights can be established and whose ownership can be transferred through market transaction”;* Services: *“Services are intangible activities that are produced and consumed simultaneously and that change the conditions (e.g. physical, psychological, etc.) of users. The engagement of users through their time, availability, attention, transmission of information, or effort is often a necessary condition that leads to the co-production of services by users and the firm. The attributes or experience of a service can therefore depend on the input of users.”* OECD, Eurostat. 2018. *Oslo Manual 2018: Guidelines for Collecting, Reporting and Using Data on Innovation, 4th Edition, The Measurement of Scientific, Technological and Innovation Activities*. Paris and Luxembourg: OECD Publishing and Eurostat.

data is integral to virtually any product innovation, and arguably the most-important impact of data on product innovation activity relates to how such activities are conducted today. In particular, virtualization has reduced the time and budget required for prototyping, testing, and deploying new products and services. Data requirements for these processes are relatively complex and typically draw on numerous sources of data for a highly specific purpose. In some instances, synthetic data is required.[52]

Data's impact is particularly pronounced when data is integrated into digital technologies. This can turn them into a knowledge-capturing product,[53] not only an enabler of innovations, but also a major building block of the innovation itself. Games, music, and video streaming services can be considered as innovative knowledge-capturing products. While these services can be deemed to be innovations in themselves, it is clear that they could not be provided without innovation activity involving data at various levels. For instance, such services often rely on innovative data formats as well as compression technologies and similar techniques that enable an enjoyable user experience. They often feature personal recommendation systems that further augment the user experience based on choices made or general preferences captured by the service. Lastly, structuring the usually large amounts of data about titles, genres, interprets, publishers, etc., requires innovative data handling. While these three examples can only be illustrative of the various roles that data play as part of the innovation activity within and around the provision of digital services, it emerges that for a successful knowledge-capturing product innovation, data usually has to be combined with other technological innovations in order to be successful.

Process innovations as defined in the Oslo Manual (OECD & Eurostat 2018) cover the full breadth of business processes.[54] Data captured by pervasive digital technologies and practices in all parts of modern businesses inform and ultimately enable such process innovations. While the most-important data will likely originate from the digital

[52] See Section 2.5 for more information.

[53] Knowledge-capturing products are defined in the System of National Accounts*: "Knowledge-capturing products concern the provision, storage, communication, and dissemination of information, advice, and entertainment in such a way that the consuming unit can access the knowledge repeatedly. The industries that produce the products are those concerned with the provision, storage, communication, and dissemination of information, advice, and entertainment in the broadest sense of those terms including the production of general or specialized information, news, consultancy reports, computer programs, movies, music, etc. The outputs of these industries, over which ownership rights may be established, are often stored on physical objects (whether on paper or on electronic media) that can be traded like ordinary goods. They have many of the characteristics of goods in that ownership rights over these products can be established and they can be used repeatedly. Whether characterized as goods or services, these products possess the essential common characteristic that they can be produced by one unit and supplied to another, thus making possible division of labor and the emergence of markets."* European Commission, IMF, OECD, UN, World Bank. 2009. System of National Accounts 2008, European Commission, International Monetary Fund, Organisation for Economic Co-operation and Development, United Nations and the World Bank, New York, NY

[54] According to the Oslo Manual OECD, Eurostat. 2018. *Oslo Manual 2018: Guidelines for Collecting, Reporting and Using Data on Innovation, 4th Edition, The Measurement of Scientific, Technological and Innovation Activities*. Paris and Luxembourg: OECD Publishing and Eurostat. process innovations comprise: (1) Production of goods or services; (2) distribution and logistics; (3) marketing and sales; (4) information and communication systems; (5) administration and management; (6) product and business process development.

technologies and practices employed within the specific firm, it can be necessary to augment this data with data accessed from other private or public entities.[55]

### Case in point: Weather data helps to optimize predictions

The weather has a fundamental impact on various processes and decisions in companies. Weather data is multifaceted: it contains information on time and location that is not only available for the past but also for the future, and which helps to anticipate the future course of actions. Therefore, access to public weather data can help to improve processes and make adequate informed decisions during different process stages. This is particularly true in the agricultural sector, where harvests may actually depend on the weather. In general, agricultural companies can use historical data and forecasts in the pre-planting phase to select seed types. In the growing phase, data can be used for decisions on irrigation and fertilization.[56]

A number of companies provide innovative solutions that use this data to make business processes more efficient. For instance, Agrivi (https://www.agrivi.com/en/) provides a farm management software for planning, monitoring, and analyzing each process. The software also uses weather data to inform farmers about optimal times for spraying and pest control measures.[57] Tracker.com also provides farm management software that helps farmers to coordinate processes. Within the project “Big Data Agricultural Platform," tracker.com intends to expand its software by integrating further data sources such as weather and satellite data. Particularly with regard to pest control, the software intends to link the documented data with weather data to enable a forecast of migration and the development of pest and weed infestation.[58]

The vision of Industry 4.0 hinges on expectations about substantial process innovations based on the comprehensive digitalization of industrial value chains, effectively turning them into value networks. This entails nothing less than a paradigm shift. It is expected

[55] To facilitate access and exchange of data from both private and public entities, the European Commission has put in place a Regulation on a framework for the free flow of non-personal data in the European Union (Regulation (EU) 2018/1807) and updated the Directive on open data and the reuse of public sector information (Directive (EU) 2019/1024). For further insights on these legislative measures and their impact please refer to Barbero M, Cocoru D, Graux H, Hillebrand A, Linz F, et al. 2018b. Study on emerging issues of data ownership, interoperability, (re-)usability and access to data, and liability, Deloitte, Brussels and Barbero M, Bartz K, Linz F, Mauritz S, Wauters P, et al. 2018a. Study to support the review of Directive 2003/98/EC on the re-use of public sector information, Deloitte, Brussels.

[56] The explanations are mainly based on https://www.ibm.com/blogs/insights-on-business/gbs-strategy/weather-means-business/ [Last accessed 22/10/2019].

[57] See https://www.agrivi.com/en/ [Last accessed 22/10/2019] and https://zenodo.org/record/1406945/files/Report%20on%20successful%20innovation%20processes%20and%20best%20practices%20_20180831.pdf?download=1 [Last accessed 22/10/2019].

[58] See https://www.trecker.com/index.php/eu-projekt-big-data-agrarplattform/ [Last accessed 22/10/2019].

that the so-called digital twin[59] will be instrumental in this. The digital twin shall ultimately enable "self-thinking supply chains" (Srai et al 2019) that autonomously allocate production capacity within and across firms in value networks. Commonly, a straightforward cause and effect relationship is stipulated where the link between increasing digitization—in particular digital twins—and a supply chain paradigm shift is self-evident. Recently, critical assessments of this naïve cause and effect assumption have emerged. It appears that data captured from sensors as part of production processes are, in practice, rarely revisited. Thus, *"availability of technical data is not the problem, rather, the problem is finding the time, tools, and expertise to analyse it"* (Saudagar et al 2019). Srai et al (2019) point out that the main challenge for a fully connected (potentially autonomous) value network is not the data, i.e., the digital twin itself, but rather, it is in the capabilities of the software behind business processes. In fact, *"challenges encountered […] are not purely technological, rather, they arise in trying to endow data with meaning, and putting the insights obtained in action […]"* (Srai et al 2019). A truly innovative business process befitting a firm with a sizable comparative advantage likely requires significantly more than access to data.

From the literature and cases reviewed in the above, it transpires that control over the data capture process and thus in-depth knowledge about the content and characteristics of data contributes highly important information to both product and process innovations. Third-party data access arguably reduces the firm's ability to fully know the characteristics of the data. In turn, this lack of knowledge may impede their ability to develop innovations that yield a competitive advantage. As such, it appears sensible that only four out of the seven national innovation policies analyzed by Planes-Satorra and Paunov (2019) mandate data access above and beyond the legislative frameworks established at the European level.[60] Furthermore, data access mandates in these innovation policies remain either vague in their approaches, aiming to improve regulatory frameworks (Germany), explore new ways of data sharing (UK), and fostering open data (China), or limited to sector-specific platforms to compile and share data (France).

Within the innovation policy context, data captured continuously in the real world can eventually be used to augment traditional official statistics to provide better insights into innovation activity and outcomes across sectors. Such improvements in the understanding of innovation systems could critically inform future policy interventions, increasing their specificity.[61]

[59] For an overview see Tao F, Zhang H, Liu A, Nee AY. 2018. Digital twin in industry: state-of-the-art. *IEEE Transactions on Industrial Informatics* 15: 2405-15.

[60] See Footnote 55.

[61] For contemporary discussions of the impact of (big) data on official statistics see e.g. Salgado D, Esteban E, Saldana S, Oancea B, Sakarovitch B, et al. 2018. *Estimation of population counts combining official data and aggregated mobile phone data.* Presented at European Conference on Quality in Official Statistics, Kraków and Wiengarten L, Zwick M. 2018. Neue digitale Daten in der amtlichen Statistik. *WISTA* 2017: 43-60.

## Key Findings of Chapter 4

- *Big data is not necessarily the best approach. Small datasets may be more efficient than large datasets, depending on the specific application.*
- *When data is a valuable input, there is a tendency that the higher the quality and quantity of data available, the more opportunities are available to discover relationships and patterns and to gain new insights, which in turn enables more efficient processes, product improvements, and service innovation.*
- *Economies of scale and scope on the supply side, re-enforced by network effects on the demand side, with additional increasing returns due to the reuse of the data, can result in data-driven market power and market concentration. However, this does not necessarily preclude competitive pressure for incumbents from potential market entrants.*
- *The collection of and the exclusive control over specific data by some firms may give rise to competition concerns in terms of access to data (barriers to entry), but data as such is worthless if firms are unable to extract knowledge that they can use to improve and/or monetize their products and services.*
- *Data quality is a key competitive resource in the data economy.*
- *From an Industry 4.0 perspective, coopetition as a mix of competition and cooperation between firms (e.g., data exchange) is the dominant paradigm.*
- *(Personal) Data don't satisfy the relevant criteria for an essential facility.*
- *Transparency regarding data management and terms and conditions enables business users and end users to weigh their preference for confidentiality and privacy against the advantages of data disclosure.*
- *Technical transparency can help to reduce transaction costs.*
- *Access to open data enables data (re-)users to promote a collaborative service-offering from several third-party service providers.*
- *Transparency obligations and unrestricted access to open data are suitable tools to effectively promote competition in the data economy.*

# 4 Competition in the data economy

The preceding sections have explored the business impact of data. It is obvious that data and the insights gained from data can have a substantial impact on business processes and profitability. So, a firm that is able to access, capture, and utilize relevant data to its advantage is arguably more competitive than a firm which cannot access, capture, and utilize relevant data.

As the data economy as a phenomenon cross-cutting traditional sectors gains traction, concerns about data-driven market power and stifled (future) competition due to inadequate access to data have emerged. In essence, policymakers suspect that data access may turn into a critical barrier to entry. Obviously, such concerns merit further investigation and this section looks at issues related to competition in the data economy.

## 4.1 Role of data for competition

> ***Insights**: It is not necessary for all kinds of applications to use big data as it can be more efficient to have specific (small) data. However, when data represents a valuable input, there is a tendency for better availability of high-quality data, suitable within the specific context and for the intended purpose, increases the odds of achieving better results—implementing more efficient processes, product improvements, and service innovation—than they could without such data. In order to reap the benefits of data access, firms need to put in significant additional effort and capabilities.*

In order to be able to understand the role of data for competition in the data economy, it is helpful to compare the previous era of “small data” and the new era of “big data” to emphasize some qualitative differences. First and foremost, for current applications, continuous data flows play a more central role than data stocks (Davenport et al 2012). While the quantity of data previously covered a range from limited to large, it can now be characterized as very large. Recently, there was scarcity regarding digital data and a strong need for access to varied data sources; one challenge was in obtaining the right sample. Today, all kind of devices, individuals, firms, and institutions create all kinds of digital data. Consequently, identifying the most suitable data—sifting through the huge quantity of continuously captured data—is often the key challenge for firms.

Regarding the information systems in use, there is a huge progression from low to high scalability and flexibility in the big data context. As storage capacities increase, allowing for even bigger datasets to be captured, the concept of big data is continuously shifting. So, whatever may be deemed big data today, may not meet the concept in the future. In addition, the type of data also defines what is meant by “big,” for instance, video needs more processing and storage capacity than text. Clickstream data from the web, video streaming data, and data flows from social media require individual handling techniques for data feeds.

Big data offers new opportunities based on the recognition of patterns with machine learning approaches to deliver new and valuable insights to different economic agents. Combining different datasets in order to infer or determine new information that has economic value in a particular context is central to the data economy. However, it is not necessary for all kinds of applications to use big data, as it can be more efficient to have specific (small) data. No matter the size of the dataset, its suitability for a specific purpose when accessed by a third party will always depend on the original context and purpose. Firms with first-party access to data naturally tend to have more control over the context and purpose of the original data capture. The layered framework developed in Section 2.1 explains this in detail.

The insights that can be extracted from data assign an economic value to specific datasets. While this seems to be obvious, in practice this turns out to be a very challenging task (Feijóo et al 2016). As data of different origin and type can be used in different contexts for a different purpose, the specific value of a particular dataset also depends on the specific task and objective (Bründl et al 2015). In general, the economic value of data can be determined from the supplier perspective on the one side, and from the user perspective on the other. For instance, an indicator of the value of personal data in the context of an advertising-financed business model is the advertising revenue per user (ARPU). Approaches to determine the value of other datasets are highly context-dependent, including the business model, type of data, the product, industry, etc. The price of data can also be determined from an (external) data intermediary's point of view (Anthes 2015, Feijóo et al 2016, FTC 2014). Here, the price per dataset is a function of the survey costs, the revenue potential at present and in the future, the competitive use of the information, and the business perspectives, as well as the overall development of the corresponding industry.

A widespread (implicit) assumption is that data is rather homogeneous (consistent), assuming it to be an important production factor similar to labor, capital, and human capital  (e.g., Farboodi & Veldkamp 2019, Jones & Tonetti 2019). However, data is heterogeneous (varied), making it difficult to define the legal status and economic value of different types of data. This is further exemplified by the discussion on data ownership issues (cf Dosis & Sand-Zantman 2019, Duch-Brown et al 2017).

Data access itself is not the main requirement to gain a (data-driven) competitive advantage, but rather it is access to data of appropriate quality that makes the difference. Data quality can be defined as "*data that are fit for use by data consumers*" (Wang & Strong 1996). The extent to which the quality of data needs to be evaluated depends first and foremost on the context of data use. This evaluation can be based on more than 170 dimensions described among others by Wang and Strong (1996).

According to Cichy and Rass (2019), the most-important objective dimensions of data quality are:

- **Completeness**: The extent to which data is of sufficient breadth, depth, and scope for the task at hand.
- **Accuracy**: The extent to which data is correct, reliable, valid, and certified.
- **Timeliness**: The extent to which the age of the data is appropriate for the task at hand.
- **Consistency**: The extent to which data is presented in the same format and is compatible with previous data.
- **Accessibility**: The extent to which information is available, or easily and quickly retrievable.

Data quality can be considered to be a key competitive resource in the data economy, where all kinds of businesses systematically collect, store, process, and use different types of data(sets). When different types of businesses compete on the basis of data-driven products and services, there is a tendency that the higher the quality and quantity of data available, the more opportunities are available to discover relationships and patterns and to gain new insights, which in turn enable product improvements and service innovation (Junqué de Fortuny et al 2013, Martens 2016).

Evaluating big data from the “resource-based view of the firm” (Barney 1991), which states that, for big data to provide a comparative advantage, it has to be inimitable, rare, valuable, and non-substitutable, Lambrecht and Tucker (2015) find that 1) big data is not inimitable or rare, 2) substitutes exist, 3) by itself, big data is unlikely to be valuable, and 4) there are many alternative sources for data available to firms. This also corresponds to Tucker and Wellford (2014), who argue that big data is one of many information inputs into the services that online businesses provide, with most firms also self-generating relevant information. They conclude that big data is neither a product in the antitrust sense nor the type of input that businesses need to obtain from others in order to compete effectively. Nevertheless, comprehensive (user and usage) data can be a valuable input in the data economy as it enables 1) customized offers, 2) personalized recommendations, and 3) targeted advertisements (Fast et al 2019). Moreover, there are data marketplaces enabling the trade of datasets. Next, we provide a summary of the outcomes of some main sources of digital data: customized services, recommender systems, targeted advertising, and data marketplaces:

> **Customized services** can lead to greater satisfaction and customer loyalty, increase switching costs, cross-selling opportunities, and willingness-to-pay (Ansari & Mela 2003, Benlian 2015, Peppers et al 1999, Pine et al 1995, Tam & Ho 2006). However, data-driven services may also raise concerns with respect to privacy and trust, depending on the reputation and operating

context of the firm and the type of customization (Awad & Krishnan 2006, Chellappa & Sin 2005, Thirumalai & Sinha 2013).

**Recommender systems** can lead to higher sales compared to businesses without recommender systems, provided that sufficient, accurate, and current data is available. Using a customer’s history (e.g., from searches, purchases, service use) to provide concise recommendations requires a high level of data quality and data quantity (O'Mahony et al 2006, Pipino et al 2002). Moreover, this depends on customers’ intention to disclose data, their willingness-to-pay, and a firm’s turnover and cross-selling opportunities (Adomavicius et al 2017, Hinz & Eckert 2010, Karwatzki et al 2017, Schafer et al 2001). In general, the quality and reputation (credibility) of recommendations depends on the timing, recommendation neutrality, and transparency to customers as well as their level of trust for the data-driven service (Benbasat & Wang 2005, Ho et al 2011, Karwatzki et al 2017, Sinha & Swearingen 2002, Wang et al 2018).

**Targeted advertising** improves the ad effectiveness in many contexts. It is dependent on click-through rates, view-through rates, purchase intention, and purchase decision of users (Bleier & Eisenbeiss 2015a, Bleier & Eisenbeiss 2015b, Goldfarb & Tucker 2011a, Goldfarb & Tucker 2011b, Kim et al 2019, Lambrecht & Tucker 2013, Tucker 2014). Important aspects that advertisers have to consider are timing and placement of ads, ad justification, trust, perceived control, transparency and privacy regulations (Aguirre et al 2015, Bleier & Eisenbeiss 2015a, Bleier & Eisenbeiss 2015b, Kim et al 2019, Samat et al 2017, Schumann et al 2014).

**Data marketplaces** are also emerging where industry data is not exclusively assigned to any of the parties involved, and can therefore, in principle, be marketed by anyone (e.g., Krämer & Wohlfarth 2018).

In essence, all these examples typically referred to in the current literature as potential positive outcomes of data access and utilization in business contexts show that data alone represents only one of many building blocks that lead to sustained business success.

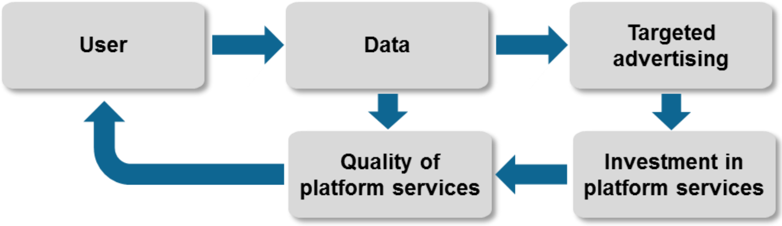
## 4.2 Data-driven market power and barriers to entry

> ***Insights***: *Economies of scale and scope on the supply side, re-enforced by network effects on the demand side with additional increasing returns due to the reuse of the data can result in data-driven market power and market concentration. However, this does not necessarily exclude competitive pressure for incumbents from potential market entrants. The collection of and the exclusive control over specific data by some firms may give rise to competition concerns in terms of access to data (barriers to entry), but data as such is worthless if firms are unable to extract knowledge that they can use to improve and/or monetize their products and services. Thus, data quality is a key competitive resource. From an Industry 4.0 perspective, coopetition is the dominant paradigm in the data economy.*

Digital technologies are changing economic decision-making and business models by shifting the costs of search, duplication, transportation, tracking, and verification (Goldfarb & Tucker 2019). As a consequence, drastically reduced costs of collecting and storing data, as well as advances in analytical techniques, allow for increased efficiency of firms from varying backgrounds (OECD 2018). This is also likely to be associated with a change in the competitive landscape. In contrast to traditional economic sectors characterized by “competition in the market,” the data economy covers a broad range of firms with data-driven products and services acting across sectors and, therefore, more often than not can be characterized by “competition for the market” (OECD 2016).

The competitive effects of data collection are often referred to as “positive feedback loops”. As the collection of data can lead to significant improvements of services, these services may attract more customers, which in turn enables the firms to collect even more data which, once again, can be used to improve their services. These positive feedback loops enabled by data may make it difficult for potential competitors to match the quality of the incumbent, thus reinforcing its arguably strong market position, leading to market concentration or even market dominance.

Figure 4-1: User and monetization feedback loops



Source: OECD (2016).

In essence, these positive feedback loops refer to economies of scale enabled by data. Economies of scale in data analytics means that the costs associated with the collection and processing of a datum (i.e., the cost per unit of output) decreases with larger amounts of data (Junqué de Fortuny et al 2013, Lewis & Rao 2015, Li et al 2016, Moore 1959). This is based on the fact that the fixed costs associated with the infrastructure are high, while the marginal costs of collecting and processing data verge toward (near) zero. The magnitude of this beneficial cost structure may also depend on whether there is first-party or third-party data collection.

Economies of scale are quite common in many industries and, from a competition perspective, it is crucial whether ever-increasing returns to scale exist or whether these diminish when a firm has achieved a sufficiently large amount of data (Lerner 2014). The latter case of diminishing returns to scale implies that the marginal value of more data declines at some point and thus the positive feedback loops are limited. The extent of economies of scale may also differ from one data-driven service to another.

There may also be economies of scope if a firm offers a variety of services that collect data, e.g., bundling of different data-driven services (Panzar & Willig 1981). Linking these data together may provide firms with more insights and enable them to improve their services and reinforce their market position more effectively than firms with fewer services. Therefore, the more data a firm can combine, the better its odds of gaining insights that could be used to strengthen its market position.

Economies of scale and scope on the supply side can be reinforced by direct and indirect network effects on the demand side. Direct network effects exist if the utility of a service increases with the number of customers, i.e., if there is a linear increase in costs associated with an exponential increase in the value of a service (Katz & Shapiro 1985). Indirect network effects can exist if there are (positive) spill-overs between the different sides of a (platform) service (Parker & van Alstyne 2005). With data-driven indirect network effects, a firm can use the data accumulated on one side of the platform not only to improve its service for this user group, but also to improve its services offered to other (user groups) sides (Prüfer & Schottmüller 2017).

The reuse of data generates additional returns to scale and scope and reinforces the market position of the service. Economies of scale and scope, as well as direct and indirect network effects, can result in market concentration; however, this does not necessarily exclude competitive pressure, (potential competition from new market entrants). This is particularly true for data-driven services that operate as multi-sided platforms.

An important factor that might limit the competitive advantage of data-rich firms is the timeliness of data. As the degree of information firms can extract from data typically depends on its accuracy, the ability to collect and process highly relevant data might be of higher importance than the mere size of a dataset, which might also include outdated data. In particular, this limitation is of significant importance for services that heavily rely on the timeliness of data as, for instance, in the case of targeted advertising. As a consequence, potential competitors do not necessarily have to build a dataset equivalent to the size of the incumbent's, they rather need to find ways to accumulate highly relevant data to build a "competitive dataset," which is not necessarily the "same dataset" (Schepp & Wambach 2016).

Superior data, i.e., data of the highest quality available, may lead to a non-transient data-driven competitive advantage (Krämer & Wohlfarth 2018). However, customers may benefit from a single service provider, e.g., if network effects are important. Additionally, the possibility of supply-side substitution and the need to innovate at a steady pace to retain a competitive advantage may reduce the potential to exploit a dominant market position.

While it cannot be ruled out that the collection of and exclusive control over specific data by some firms may give rise to competition concerns, data as such is worthless if firms are unable to extract knowledge that they can use to improve and/or monetize their products and services. Thus, important factors of success are, among other things, the professionals, such as data scientists, and the technology used to analyze the accumulated amounts of data. Furthermore, data-driven firms also have to anticipate diminishing benefits for increasing dataset sizes (Li et al 2016) as well, as they have to deal with a minimum efficient scale of data use (Lewis & Rao 2015).[62]

There are conditions in which it is possible for a data-driven firm to sustain and enhance market power based on its collection and use of data (Fast et al 2019). While network effects on their own promote competition for the market, their combination with other factors are likely to foster monopolies, essentially acting as a barrier to entry that protects the incumbent firm (Schweitzer 2019). Reputation effects from a brand or an established firm are likely to favor large firms at the expense of smaller firms when it comes to the collection of consumer data in the context of personalized content and services (Chellappa & Sin 2005).

Dominant services may foreclose competition by raising barriers to entry in the large-scale collection of user data (Haucap 2019, Haucap & Stühmeier 2016). This may give rise to access problems for competitors and new entrants that need access to data gathered by dominant services in order to provide competing or complementary services (Graef et al 2015). A study by Rubinfeld and Gal (2017) analyses the different types of access barriers that limit entry into the different links of the data value chain and they find that the unique characteristics of big data have an important role to play. In contrast, Mahnke (2015) argues that it does not matter how much data a firm has, but instead that it is decisive what they do with the data when they design and improve their products and services. As a consequence, competition analysis has to take into account that while there may be strong competition in some data markets, this does not mean that there are no barriers to entry in others. Schweitzer (2019) argues that access to

[62] We elaborate on this further in Section 2.5.

data can only be covered via competition law if there is (1) market power in primary markets, or (2) market power because of exclusive control over relevant data in secondary markets.

Industry 4.0 is a paradigm change whereby traditional value chains are no longer the dominant organizational structure (e.g., firms operate at (a) certain layer(s) of value creation). Instead, there are data value networks, where firms increasingly participate in different value creation processes simultaneously.[63] However, industry data currently faces a lack of legal certainty regarding data ownership and (re-)use of data. From a competition perspective, the dominant paradigm will be coopetition, which characterizes the relationship between firms as a mixture of competition and cooperation at the same time.

Overall, at this early stage of development of the European data economy, there is no structural problem leading to market failure, which could render sector-specific regulation necessary. Indeed, numerous firms engage in different types of data exchanges in the data economy based on bilateral and multilateral contractual arrangements, while also being in competition with others. As an ongoing research project initialized by the European Commission shows,[64] "*a significant share of firms in Europe and elsewhere have been exchanging data incentivized by the market instead of data sharing regimes*." However, policymakers and regulators aiming to strengthen the data economy may focus on promoting such data exchanges with suitable standardization of data formats, data portability, interoperability, and provision of legal certainty regarding reference architectures. We outline this further in Chapters 5 and 6 of this report. Table 4-1 summarizes the main competition concerns and the main assessment criteria, which have to be scrutinized on a case-by-case basis.

Table 4-1: Competition concerns and assessment criteria

| Main competition concerns | Main assessment criteria |
|---|---|
| • Is there exclusive control of certain data creating a significant barrier to entry?<br>• Is there a leverage of market power into adjacent markets?<br>• Is there a lack of competition over non-price features such as privacy?<br>• Are there information asymmetries between users (i.e., consumers, businesses) and the service provider resulting from a specific position with access to a (very) comprehensive amount and variety of (timely) data? | • Economic properties of relevant data<br>• Context and purpose of data use<br>• Access to data sources<br>• Data quality<br>• De facto exclusivity of data<br>• Opportunities for replication of data<br>• Data-based economies of scale & scope<br>• Context-dependency of data for the corresponding product/service<br>• Value of the data, i.e., information value |

Source: WIK-Consult.

[63] For discussion of this development see Chapter 3.
[64] See http://datalandscape.eu/companies.

In the data economy, competition revolves around direct access to the (business and/or end-) user via interfaces such as application programming interfaces (APIs). This enables direct interactions between firms and with customers for service customization in (near) real time. Accordingly, such interfaces are developed or made available when necessary, resulting in market-driven data exchanges.

However, a frequent question in competition policy circles is whether a further lowering of the barrier to data sharing is likely to strengthen competition in the data economy. In answering this question, it is helpful to differentiate two cases: (1) facilitating data exchanges by the end-user, and (2) facilitating data exchanges by businesses.

While an artificial lowering of the entry barrier to data sharing of personal data raises further privacy concerns, it may also undermine trust in data exchange practices and can be considered detrimental to competition in the data economy. The second case, facilitating data exchanges by businesses, is relevant from a competition perspective, since an uneven playing field between large and small firms may hamper the competitive process and so lowering the barrier to entry to data sharing may be helpful when it comes to public data and third-party data, but not confidential data (e.g., data encompassing business and trade secrets).

Another frequently discussed issue in competition policy circles is whether (personal) data can be an “essential facility”. An essential facilities doctrine (EFD) specifies when the owner(s) of an “essential” or “bottleneck” facility is mandated to provide access to that facility at a “reasonable” price (OECD 1996). The concept of “essential facilities” requires two markets, often expressed as an upstream market and a downstream market (e.g., two complementary products/services). Typically, one firm is active in both markets and other firms are active or wish to become active in the downstream market. A downstream competitor wishes to buy an input from the integrated firm, but is refused. An EFD defines those conditions under which the integrated firm will be mandated to supply access to its facilities (Sidak & Lipsky 1999).

However, in order to qualify as a potential candidate for an “essential” facility or infrastructure, (personal) data would have to fulfill two necessary conditions (Areeda & Hoverkamp 1988):

1) market entry to the complementary market is not effectively possible without access to this facility, and
2) a supplier on the complementary market cannot duplicate this facility with reasonable effort; substitutes do not exist.

It is obvious that (personal) data does not fulfill either condition. Regarding the first condition, it requires that one firm is more cost-effective than alternatives. However, every firm in the data economy can gather any (personal) data at often negligible (near-to-zero) marginal cost such that there is no monopoly when it comes to (personal) data;

so the first condition is not satisfied. With respect to the second condition, (personal) data can always be duplicated or replicated with reasonable effort, even if it requires the consent of the end-user at hand. And so the second condition is not satisfied. Overall, there are no cases where (personal) data could qualify as an essential facility.

In order to educate the current debate on data sharing, it is important to differentiate (1) data as a (platform) service, where there are data-sharing incentives for established firms, and (2) data as the originally captured data,[65] where there are no sharing incentives for established firms.[66]

[65] Often referred to as “raw data”.

[66] We elaborate this further in Chapter 5 of this report.

## 4.3 Promoting competition with transparency and access to open data

> ***Insights***: *Transparency regarding data management and terms and conditions enables business users, as well as end users, to weigh their preference for confidentiality and privacy against the advantages of data disclosure. Technical transparency can help to reduce transaction costs. Access to open data enables data (re-)users to promote a collaborative service-offering from several third-party service providers. Thus, transparency obligations and unrestricted access to public open data are suitable tools to effectively promote competition in the data economy.*

Pursuing the next steps toward a unified European data economy may include an informed design of data policy frameworks. Clearly, regulations matter, as they define the available access points to data and determine exclusivity and plurality with respect to the control of data. If there is an information asymmetry between data-driven businesses and business users, as well as end users, regarding the collection and use of data, then it can be reduced with more transparency.

Transparency obligations enable business users as well as end users to weigh their preference for confidentiality and privacy against the advantages of data disclosure. Moreover, technical transparency can help to reduce transaction costs in the data economy (Tsai et al 2011). Therefore, effective transparency obligations can strengthen trust in data-driven services and may enable them to compete on confidentiality and privacy dimensions as a way of product and service differentiation (Casadesus-Masanell & Hervas-Drane 2015).[67]

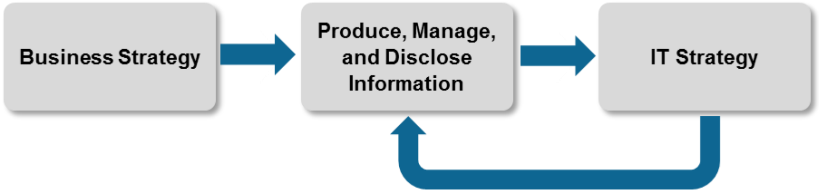
In line with this approach is the proposed "New Deal for Consumers" Directive[68] in Europe. Digital platforms such as online marketplaces have to provide transparency regarding the most-important parameters within their rankings. As a consequence, online services must also provide information on whether third parties pay a fee for a better ranking or for inclusion in the result lists. In addition, according to the proposed

[67] However, there are three types of paradoxes, which are currently subject to further research: 1) privacy paradox Awad NF, Krishnan MS. 2006. The Personalization Privacy Paradox: An Empirical Evaluation of Information Transparency and the Willingness to Be Profiled Online for Personalization. *MIS Quarterly* 30: 13-28, Norberg PA, Horne DR, Horne DA. 2007. The Privacy Paradox: Personal Information Disclosure Intentions versus Behaviors. *Journal of Consumer Affairs* 41: 100-26, 2) control paradox Brandimarte L, Acquisti A, Loewenstein G. 2012. Misplaced Confidences: Privacy and the Control Paradox. *Social Psychological and Personality Science* 4: 340-47 and 3) transparency paradox Bernstein ES. 2012. The transparency paradox: A role for privacy in organizational learning and operational control. *Administrative Science Quarterly* 57: 181-216, Nissenbaum H. 2011. A Contextual Approach to Privacy Online. *Daedalus* 140: 32-48. All three paradoxes revolve around cognitive biases in individual decision-making. They essentially undermine the positive effects of transparency obligations and other regulatory approaches. As long as these issues are not resolved (e.g., implications of these phenomena), it remains an ineffective task to make sound recommendations to regulators and policymakers.

[68] Proposal for a Directive as regards better enforcement and modernization of EU consumer protection rules, COM(2018)0185 final-2018/090 (COD).

"Platform-to-Business" Regulation,[69] digital platforms such as search engines have to make accessible certain relevant criteria for their ranking to the benefit of other business users and consumers.

Figure 4-2: Transparency strategy



Source: Granados and Gupta (2013).

Anticipating these new regulations, firms in the data economy should develop a transparency strategy (see Figure 4-2 above) in order to be able to make decisions about information disclosure within and outside the firm by selectively disclosing relevant information to other business users as well as to consumers (Granados & Gupta 2013). Overall, ensuring an appropriate level of transparency helps to foster fair competition while also benefiting business users and consumers.

Table 4-2: Principles for open data

| | |
|---|---|
| **Data must be complete** | All data is made available, subject to statutes of privacy, security, or privilege limitations. |
| **Data must be primary** | Data is published as collected at the source, with the finest possible level of granularity, not in aggregate or modified form. |
| **Data must be timely** | Data is made available as quickly as necessary to preserve the value of the data. |
| **Data must be accessible** | Data is available to the widest range of users for the widest range of purposes. |
| **Data must be machine-processable** | Data is reasonably structured to allow automated processing of it. |
| **Access must be non-discriminatory** | Data is available to anyone, with no requirement of registration. |
| **Data formats must be non-proprietary** | Data is available in a format over which no entity has exclusive control. |
| **Data must be license-free** | Data is not subject to any copyright, patent, trademark, or trade secret regulation. Reasonable privacy, security, and privilege restrictions may be allowed as governed by other statutes. |
| **Compliance must be reviewable** | A contact person must be designated to respond to people trying to use the data or complaints about violations of the principles and another body must have the jurisdiction to determine if the principles have been applied appropriately. |

Source: http://www.opengovdata.org.

[69] Proposal for a Regulation on promoting fairness and transparency for business users of online intermediation services, COM(2018)238/974102.

Another approach to promote competition is to enable open data, which implies that a public data supplier makes available its data to an open range of data (re-)users in order to promote the evolution of an ecosystem with several third-party service providers (Argenton & Prüfer 2012). Public organizations may enable and ensure access to open data in order to effectively promote competition in the data economy.

## Key Findings of Chapter 5

- *Financial data providers, data brokers, and online aggregators are the main economic agents enabling trade of data on data markets. Their products and services can increase customers' choice, and lower search and transaction costs. However, they also raise privacy concerns.*
- *Horizontal data sharing is mostly happening on the basis of bilateral contractual agreements and common initiatives between firms. Value creation regarding horizontal data sharing is determined in the transition from data to information (context-dependent).*
- *Vertical data pooling leverages state-of-the-art technologies and standards to create value in the transition from data gathering and data processing, through gaining insights to information provision and information use. Data pooling therefore enables a high level of innovation, in particular in the Industry 4.0 context.*
- *The overall welfare effects are ambiguous depending on the specific design of the data exchange architecture.*
- *Data portability and interoperability (compatibility) are central to innovation.*
- *While data portability refers mostly to open specifications, interoperability is linked to formalized standards.*
- *Data portability is currently lacking uniform technical standards as regards data formats and processing protocols for data extraction and implementation, rendering the concept ineffective.*
- *Interoperability may help to weaken the power of network effects (i.e., the market power of incumbent firms) but at the expense of economic efficiency.*
- *Depending on their specific design, both tools may result in lower levels of confidentiality and data protection.*

# 5 Data-sharing approaches

A key element of the popular narrative around the data economy is an apparent lack of data exchanges. This chapter informs considers the various types of data exchanges that are already happening.

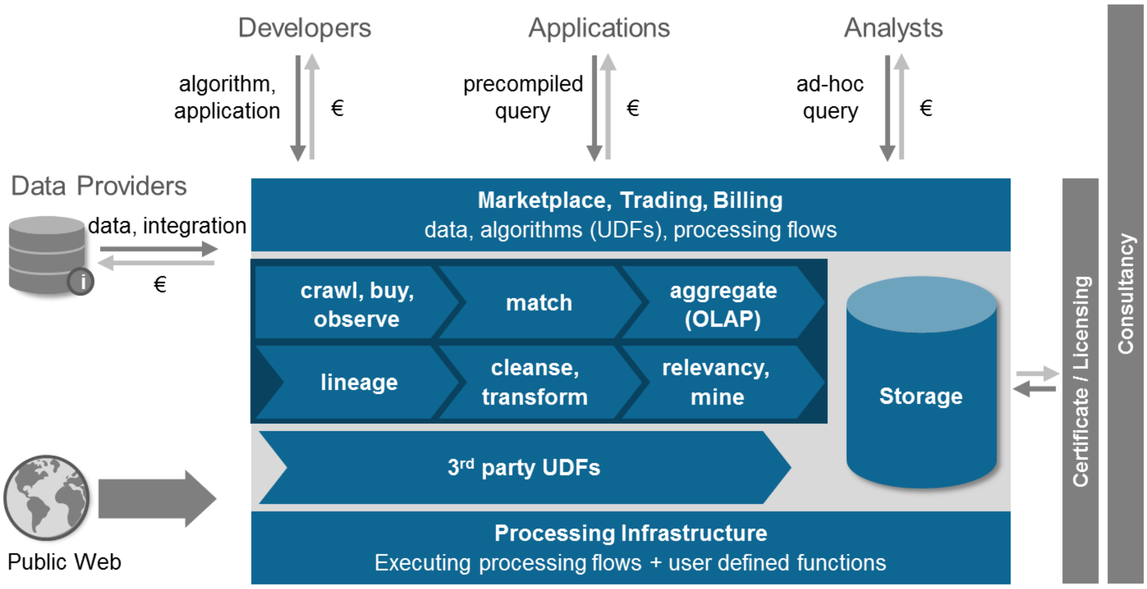
## 5.1 Data markets and trade of data

> ***Insights**: Financial data providers and credit rating agencies calculating scores on users, data brokers compiling huge databases on individual consumers, and online aggregators mining publicly accessible data to create consumer profiles represent the main economic agents enabling data markets and trade of data. They provide different products and services, increasing customers' choice, and lowering search and transaction costs. However, they also raise privacy concerns.*

There are basically three types of data vendor engaged in commercial data markets (Bergemann & Bonatti 2019). Depending on the source of the data it can be distinguished between: 1) financial data providers (e.g., Bloomberg, Thomson Reuters) and credit rating agencies (Moody's, Standard & Poors, Fitch), calculating scores on users; 2) data brokers (e.g., Acxiom, Palantir, LexisNexis) compiling huge databases on individual consumers; and 3) online aggregators (e.g., Intelius, Spokeo) mining publicly accessible data to create consumer profiles. According to a study by Pew Research Center, consumers can typically engage in data trade in three ways: 1) data against services; 2) data against remuneration; 3) data as a donation.[70]

[70] The Pew Research Center survey of American consumers on data sharing: https://www.pewinternet.org/2016/01/14/privacy-and-information-sharing/

Figure 5-1: Data marketplace



Source: Muschalle et al (2012).

Figure 5-1 illustrates how a data marketplace works. As this example of integrating public web data with other data sources shows, it includes components for data extraction, data transformation, and data loading, as well as metadata repositories describing data and algorithms (Muschalle et al 2012). The data marketplace has interfaces for data integration, methods for optimizing, and components for third-party trading and billing. The data marketplace operator receives a monetary fee from its clients (e.g., developers, analysts, applications), while relying on a scalable infrastructure for processing and indexing data.

A commercial data markets investigation by the US Federal Trade Commission (FTC 2014) has analyzed nine large data brokers, their business models, and practices. The investigation found that data brokers compile information on users from sources such as public posts, online purchases, browsing history, and warranty cards. While collecting data primarily from open government sources, they also use social media, blogs, and other commercial data sources. These data brokers are found to provide basically three types of products and services:

- online and offline marketing analytics;
- risk mitigation, e.g., scoring, verification of identities, fraud detection;
- people search, e.g., competitive intelligence services, finding old friends.

Data brokers provide users with more choice, and they lower search and transaction costs; however, their business model does raise privacy concerns. While multiple organizations hold data that can be considered inconclusive for individuals, they can be useful for large (global) organizations to solve complex problems from a societal perspective. Due to strategic, legal, and policy concerns, there is a hesitation to exchange data. In a secure data market, exchanged data is secure and is available in granular or aggregate form based on the specific requirements of vendors and customers. While data trade is considered to be essential for the data economy, there are multiple layers of data brokers (increasingly processing and refining the data) between the originally captured data and the data sold to customers, which makes it difficult for the user to trace the data used by the clients of a data broker (Duch-Brown et al 2017).

Most data brokers, such as Palantir, provide different types of products and services, e.g., captured data, analytic results, data analytics software and courses, and consultancy services. From an economics perspective, data brokers typically face high fixed costs to create their data-driven products and services and low marginal costs per data unit. Thus, there are economies of scale and scope in data markets. While this cost structure does not result in cost-based pricing, it does lead to pricing according to the value to customers, e.g., versioning of information goods (Shapiro & Varian 1999). Their pricing strategies on different data markets are largely unknown but may include: free data obtained from public authorities such as statistical data; usage-based pricing; package pricing; subscriptions; two-part tariffs consisting of a fixed fee and a variable fee per unit sold; and freemium, a combination of a free version with basic services and a paid version with premium services.

## 5.2 Horizontal data sharing and vertical data pooling

> ***Insights**: Horizontal data sharing commonly happens on the basis of bilateral contractual agreements and common initiatives among firms in which value creation occurs on the progression from data to information (context-dependent). In contrast, vertical data exchange, in terms of data pooling, leverages state-of-the-art technologies and standards to create value on the progression from data gathering, data processing, and gaining insights, to information provision and use. Data pooling therefore enables a high level of innovation—in the Industry 4.0 context in particular.*

The most popular means of **horizontal** data transfer in terms of **data sharing** is through bilateral contractual agreements and common initiatives among firms, as the study by Barbero et al (2018b) shows. A market research study by Pauer et al (2018) of different companies from all industries on data exchanges finds that 75% of executives consider the opportunities for improving customer relations, customer contacts, and services to be very good. Moreover, more than two-thirds consider optimizing both the firm’s processes and the supply chain to be very big opportunities. A survey commissioned by the European Commission on business-to-business (B2B) data sharing finds that: (1) firms share and reuse data already; (2) their share will grow in the near future; (3) it enhances business opportunities and improves internal efficiency; (4) investments in real time data access or localization data may have a positive impact on a firms business; (5) most data suppliers and users appear to share and reuse data within their own business sector; (6) data holders share only a small proportion of the data they hold; (7) technical and legal obstacles are hindering B2B data sharing, while denial of access is a common barrier among firms re-using data; and (8) trust and simplicity would help firms seeking to share data (Arnaut et al 2018).

Regarding horizontal data sharing and its welfare effects, Jentzsch et al (2013) show that the incentives for data sharing among firms depend on the type of customer data and on customer variability. The incentives to share data are stronger if customers are all relatively similar to one another. Customer data sharing is most likely to be detrimental to consumer surplus, while the effect on social welfare can be positive. Making rival firms share their customer-specific data may require sufficient firm asymmetry (Liu & Serfes 2006). For instance, a low-quality firm may sell its customer database to a high-quality firm. However, the high-quality firm may never sell its data to the low-quality rival.

The continuous or occasional transfer of data between firms in a horizontal relationship with specified purposes principally enables increasing efficiency of business processes, product improvements, service innovation, customization of services, and recommendations. Coopetition based on contractual agreements on horizontal data sharing is likely to create an increase in social welfare. Market-driven limitations to horizontal data-sharing agreements depend on whether the value of the shared data comes from a datum or rather from its agglomeration and subsequent analysis. Furthermore, competition law may restrict (data) cooperation between (rival) firms, e.g., regarding sales, cartels, data-driven barriers to entry.[71] Thus, value creation regarding horizontal data sharing occurs during the progression from data to (context-dependent) information.

In contrast, **vertical** data exchange in terms of **data pooling** creates value in the process from (1) data gathering, (2) data processing, (3) gaining insights, (4) information provision, and (5) information use. Establishing data pooling in order to exchange and use data by firms in a vertical relationship plays an increasingly important role in the data economy. Three types of economic agents may have a legitimate claim on the returns it generates: 1) the data subject, who the data is about; 2) the data collector, who pays the cost of protecting the data; and 3) the data processor, who analyses the data to extract insights (Carrière-Swallow & Haksar 2019). Each of these agents has their own interests, and their decisions with respect to the data-pooling requirements are going to affect the others.

A promising example of an alliance of private firms to implement a data pool is the Industrial Data Space (Otto et al 2016, Otto et al 2018). This project aims to standardize internet of things applications and basically consists of five layers, each specifying the elements necessary for effective data pooling:

Table 5-1: Layers of the Industrial Data Space (IDS)

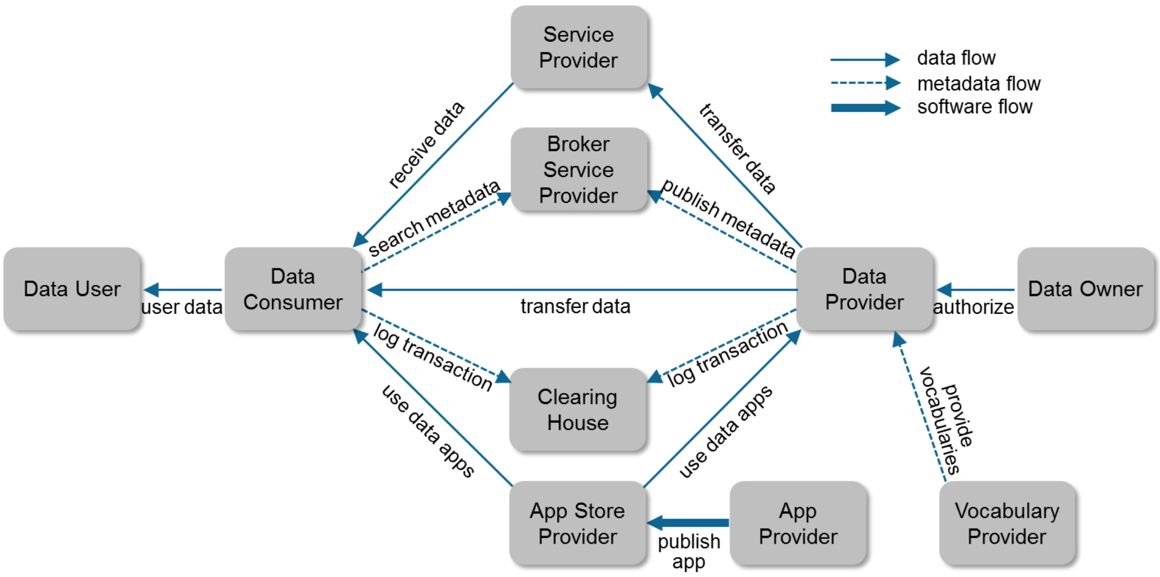
| Layer | Description |
|---|---|
| **System** | Maps the roles of the business layer onto a data and service architecture; a technical core based on three major components: connector, broker, and app store. |
| **Information** | Specifies the compatibility and interoperability of the data pool. |
| **Process** | Specifies the interactions between different components; providing data, exchanging data, publishing, and using data apps. |
| **Functional** | Defines the functional requirements and features to be implemented. |
| **Business** | Specifies the different roles which participants may assume. |

Source: Otto et al (2018).

In general, from a business perspective this project foresees several roles: data owner, data provider, data consumer, data user, broker service provider, clearing house, identity provider, app store, app provider, vocabulary provider, software provider, service provider, and a certification body and evaluation facility. Depending on the status, each role receives and/or sends (content) data flows, meta data flows, and software flows. Figure 5-2 illustrates these roles and their interactions at the business layer.

[71] There are also legal exceptions such as the EU Block Exemption Regulations for horizontal and vertical agreements.

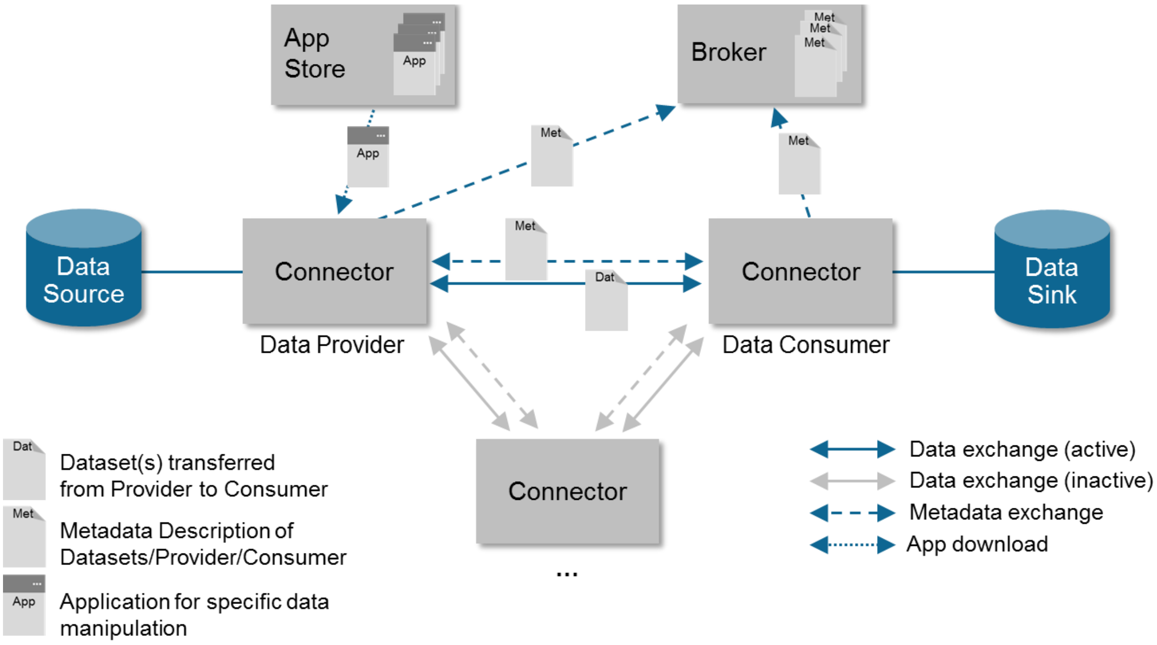
Figure 5-2: Roles and interactions (business layer)



Source: Otto et al (2018).

At the functional layer, the individual requirements are grouped into six functional entities: trust, security and data sovereignty, ecosystem of data, standardized interoperability, value-adding apps (data processing software), and data markets. Finally, the interaction of technical components, including data usage control make up the system layer as shown in Figure 5-3.

Figure 5-3: Interaction of technical components including data usage control (system layer)

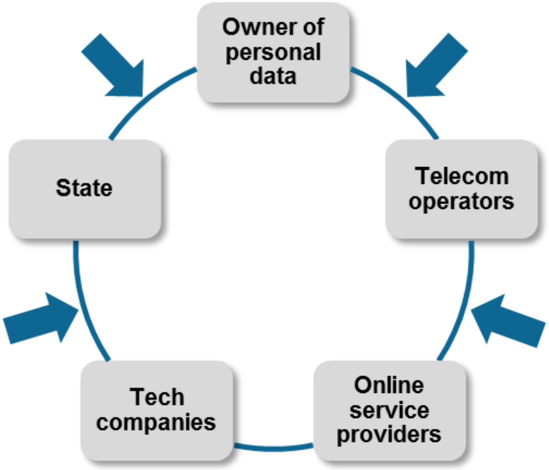


Source: Otto et al (2018).

Overall, “*the Industrial Data Space is a virtual data space leveraging existing standards and technologies, as well as accepted governance models for the data economy, to facilitate the secure and standardized exchange and easy linkage of data in a trusted business ecosystem. It thereby provides a basis for smart service scenarios and innovative cross-company business processes, while at the same time making sure data sovereignty is guaranteed for the participating data owners*” (Otto et al 2018).

An example of a public institution representing the data-pool provider is the German Data Trust project (Lind & Suckfüll 2013). The architecture strives for a proper treatment and use of personal data by a German data-pool provider based on the idea of GEMA, a well-known data-pool provider in the music industry. The leading roles in this model are illustrated in Figure 5-4.

Figure 5-4: German Data Trust



Source: Lind and Suckfüll (2013).

The German Data Trust allows rules for personal digital data use and monetization to be defined. The data trust entity considers all interests and preferences of the participants involved, striving for a balance of ownership and usage rights to which all economic agents agree. The data trust (as custodian) also manages the usage fees received from the data processors and ensures these are passed on to the data subjects. A benefit of this approach is that it provides control over the personal data, stored in a coded format within the data trust entity, which has by definition no economic interests. According to Heumann and Jentzsch (2019), such data pooling can be an enabler for innovation. However, because of regulatory uncertainty and a lack of initiatives by small to mid-sized firms, it is so far primarily larger firms that have engaged in this approach.

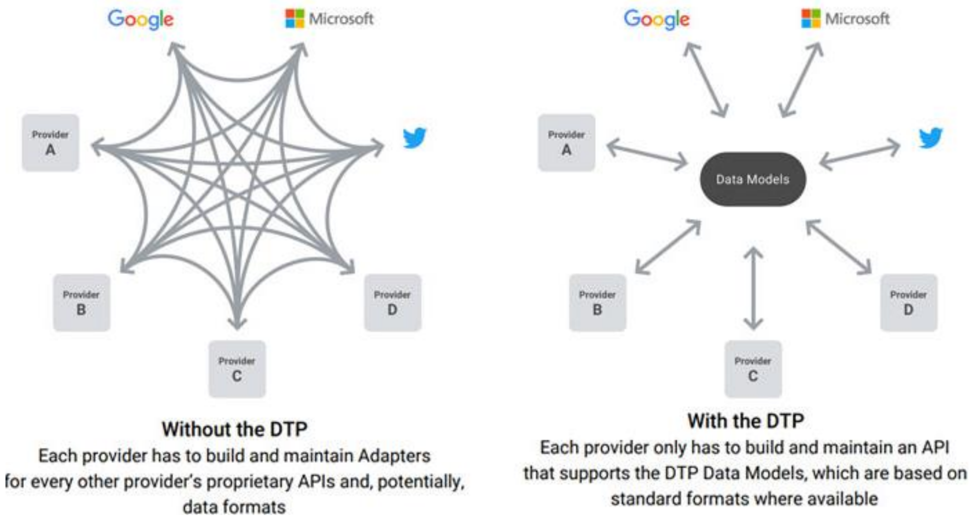
## 5.3 Static data portability and dynamic interoperability

> ***Insights***: *Data portability and interoperability are central to innovation. While data portability refers mostly to open specifications, interoperability is linked to formalized standards. Data portability is currently lacking technical standards regarding data formats and processing protocols for data extraction and implementation, rendering the concept ineffective. Interoperability standards may help to weaken the power of network effects and therefore market power of incumbent firms, but at the expense of economic efficiency. They may also result in lower levels of confidentiality and privacy.*

**Data portability,** according to Article 20 of the EU General Data Protection Regulation (GDPR), is defined as "*the right to receive the personal data concerning him or her, which he or she has provided to a controller, in a structured, commonly used and machine-readable format and have the right to transmit those data to another controller without hindrance from the controller to which the personal data have been provided*." As a consequence, the process of data portability consists of two stages: 1) extraction of the personal dataset from service A; and 2) implementation of the personal dataset to service B. However, while the first part (data extraction) works fine with the current data protection regulation in Europe, the main problem arises when it comes to the implementation of the dataset on another (platform) service. Today, there is no technical regulation requiring each service provider to use the same, uniformly applied, data formats and processing protocols (standards). While experts agree that data portability helps consumers to control their data, data portability is currently lacking the technical capabilities to be effective in practice (Egan 2019). Therefore, a law that requires the opportunity for data transfers can only be considered a first step toward true data portability.

In an effort to overcome these issues, Apple, Deezer, Facebook, Google, Mastodon, Microsoft, Solid, and Twitter initiated the "Data Transfer Project (DTP)".[72] The alliance is working on standards and technologies that ensure a proper realization of data portability worldwide. Thus, all participants of the DTP commit themselves to use the same data models when it comes to customers wanting to transfer their personal data from platform A to B in near-to-real time.

Figure 5-5: The Data Transfer Project


Google
Microsoft
Provider A
Provider B
Provider C
Provider D
Without the DTP
Each provider has to build and maintain Adapters for every other provider's proprietary APIs and, potentially, data formats
Google
Microsoft
Provider A
Data Models
Provider B
Provider C
Provider D
With the DTP
Each provider only has to build and maintain an API that supports the DTP Data Models, which are based on standard formats where available

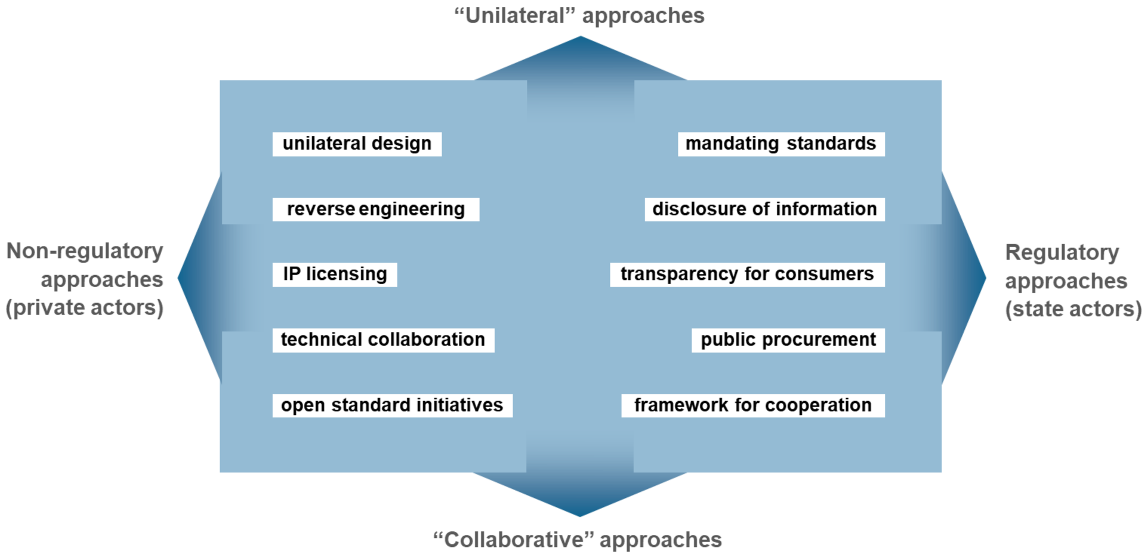
Source: https://datatransferproject.dev/

[72] See https://datatransferproject.dev/.

Such data portability makes switching of service providers by consumers more convenient and may facilitate the entry of new firms. However, the GDPR applies only to data "provided by" the consumer (data subject) such as purchasing data. Data "derived by" a firm (data controller) with the help of data analytics, such as recommendations derived from purchasing data does not fall under the GDPR. Lam and Liu (2018) show that without data analytics, data portability can indeed facilitate switching, but with data analytics, data portability may hinder switching because consumers anticipate easier data portability and thus are more willing to provide data to the incumbent, which strengthens the incumbency advantage. Swire and Lagos (2013) argue that the right to data portability imposes substantial costs on suppliers of software and apps. Moreover, the risks of data portability may be too high, as one moment of identity fraud can turn into a lifetime breach of personal data. Wohlfarth (2019) shows that although data portability is designed to protect users, they may be hurt because market entrants have an incentive to increase the amount of collected data compared to a regime without data portability. However, profits for new services and total surplus increase if the costs for implementation are not too large. This is likely to improve innovation and service variety.

While data portability refers mostly to open specifications, **interoperability** is linked to formalized standards. Interoperability can be defined as the "*ability to transfer and render useful data and other information across systems, applications, or components*" (Gasser & Palfrey 2007). However, there are several approaches to interoperability, as Figure 5-6 shows.

Figure 5-6: Approaches to interoperability


"Unilateral" approaches
unilateral design
reverse engineering
IP licensing
technical collaboration
open standard initiatives
mandating standards
disclosure of information
transparency for consumers
public procurement
framework for cooperation
Non-regulatory approaches (private actors)
Regulatory approaches (state actors)
"Collaborative" approaches

Source: Gasser and Palfrey (2007).

Interoperability types can be differentiated by platform, device, service, data format, (processing) protocol, and functionality. Depending on the type of interoperability, there can be many implications for competition, innovation, and privacy. For instance, according to Casadesus-Masanell and Ruiz-Aliseda (2009), incompatibility changes the balance, and a dominant platform may out-perform competitors to a greater extent than it could under compatibility (i.e., interoperability). Interoperability may therefore help to weaken the economic power of network effects and the market power of incumbent firms (Adner et al 2016). However, at the same time, this also decreases economic efficiency and, depending on the design of the interoperability standards, they may also result in lower levels of confidentiality and data protection.

It is therefore very important to define exactly what type of interoperability is considered for each particular objective. Interoperability can be relevant on different layers (Horak 2008): syntactic interoperability refers to the possibility that systems can physically connect to each other and exchange data, whereas semantic interoperability refers to the ability of systems to understand the meaning of the information exchanged. Horizontal interoperability (i.e., competing services such as mobile telecommunications services) and vertical interoperability (i.e., complementary services such as a web browser and an operating system) are open by design if data/information can be shared and/or accessed by complementary or rival services, respectively. Thus, in practice there is a continuum between full and no interoperability, based on the number of functionalities specified.

In the data economy, interoperability of a variety of platforms and networks may be key, in particular for data value networks in the context of Industry 4.0. However, such market interventions are associated with benefits and costs and interoperability and openness can also be symmetric or asymmetric for different market actors. As a consequence, they have to be designed such that trade-offs revolving around competition, innovation, and privacy are taken into account.

## Key Findings of Chapter 6

- *The key challenge for the data economy does not appear to be data access as such, but instead the ability to discover, understand, and integrate third-party datasets.*
- *Policymakers should promote the development and adoption of common data architectures as a means to facilitate data exchanges.*
- *Reference architectures—if adopted widely—will likely increase the quantity and quality of data exchanges.*
- *Firms should prefer third-party data as it is generally cheaper than first-party data capture.*
- *Policymakers should refrain from deciding which data exchanges create economic value (avoid picking winners and losers). Data exchanges should be market driven.*

# 6 Implications for economic policy and regulation

Thus far, we called for increased attention to detail in the debate around the data economy. Unlike other assets, data is extremely heterogeneous as it is captured within a specific context for a particular purpose using a selected sensor or application programming interface (API). As it is then analyzed to extract information of economic value, data is further processed and may be reintroduced to further cycles of analysis as elevated data. The circular character of value creation in the data economy and the multiple interactions among (often multi-sided) businesses partaking in the data economy add further opacity.

It is clear, however, that data plays an integral role for innovation and competition. For both, whether (and which) data can ultimately generate a competitive advantage depends on the specific use case. With increasing digitalization and various existing opportunities for organizations to exchange and reuse data, a fundamental hindrance to data access is difficult to conceive. However, it transpires also from our findings that a lack of knowledge about the details of data capture, context, and purpose, the (pre-) processing conducted and other characteristics curb data reuse for businesses, researchers, and public servants alike. Increasing the transparency of these characteristics would likely increase data's fluidity within and across sectors. In turn, stakeholders could benefit from a higher quantity of data that becomes reusable, as well as a better efficiency of data reuse.

While the idea of lowering the barriers to data exchange does not necessarily strengthen competition in the data economy (as it depends on the data itself and the market conditions), regulatory obligations for porting platforms regarding personal data may help to overcome the problems with data implementation that are currently rendering data portability ineffective in practice. Thus, standardization of data formats can be considered necessary in order to strengthen the data economy. However, any type of data exchange is a non-zero-sum game. In situations where there are gains from data sharing, profit-maximizing firms will engage in such activities. In contrast, a data-sharing regulation would fall short of the complexity involved and would require case-by-case decision-making. Thus, market-driven bilateral and multilateral agreements between firms are superior to any general (sectoral) regulation.

A general data-sharing regulation would also not solve the privacy problem. The fundamental problem of European data protection is the instrument of consent(ing) itself, which has to be resolved.

In essence, the complex nature of the data economy ought to caution policymakers against an overly interventionist approach. There is a substantial risk of policymakers (unintentionally) picking winners and losers. Establishing a framework that can overcome the central challenge of transparency in the data economy in a manner that adheres to the fundamental policy principle of technology neutrality appears appropriate in light of the insights presented in this report. We feel that such a framework can also mitigate other challenges emerging in the data economy. Consequently, this chapter commences with a short summary of apparent challenges in the data economy and continues to suggest a technology-neutral policy approach to facilitate a European data economy.

## 6.1 Policy-relevant challenges for the data economy

> ***Insights**: Data accessibility in itself is not the main hindrance for a thriving data economy. Transparency, discoverability, and re-usability of data are more critical challenges to solve. Interventions need to be technology-neutral and enable all stakeholders to exchange data.*

This report sheds light on the key challenges for a thriving data economy. As highlighted in Chapter 2, even a common understanding of what constitutes data is lacking from the public debate. This not only impedes a nuanced understanding of the economic potential of the data economy, but also creates challenges for exchanging data and its subsequent reuse. Such challenges have already become obvious in the debate around the reuse of research data. As Borgman (2010) indicates, the substantial variation of the notion of “data” among collaborators, and even more so across disciplines, renders the reuse of research data difficult. In the same vein, sector silos, and even boundaries between different IT systems may block the potential of data exchanges.

To transfer data from one context to another context and to then make it usable in the new context might well require substantial investments in processing, structuring, or tagging of the data in a compatible format. Within this, there is always a risk of not being able to use the data, ultimately due to a lack of knowledge about the specific characteristics of the data stemming from the processing they underwent during capture, or due to inadequacies in data completeness or its frame of reference. Indeed, integration of datasets from various sources is a key challenge for a thriving data economy and in particular for innovation (Paunov & Planes-Satorra 2019).

These investments, combined with the non-zero cost of gaining third-party access to relevant data, may incentivize firms to rely on their own capabilities to capture the data that they need. Having full control over their means of data capture mitigates a substantial part of the costs involved in data conversion, as well as the risks involved in not having full knowledge over the process of data capture. In data-rich environments, firms will likely prefer their own data over data from a third party whenever that is possible. In data-poor environments, where data capture is much more expensive than in data-rich ones, the overall efficiency gains from data exchange and reuse are significantly greater. On the one hand, since some of the most-important industrial sectors in Europe feature such data-poor environments, unlocking the data equity by setting framework conditions that can facilitate data exchanges promises substantial economic impact, e.g., by freeing up resources to collect additional data instead of reproducing data collected by others or even worsening their mistakes in utilizing data.[73] On the other hand, in particular in data-poor environments, creating access to data can be considered an innovation in itself, resulting in a competitive advantage.

Against this backdrop, policymakers should facilitate data exchange by ensuring that reference architectures and standards are developed and agreed upon across sectors. Within such an effort, the trade-off between transparency and data access should be considered, alongside the cost and complexity of still being able to preserve important trade secrets for the organizations involved. Notably, the complexity of safeguarding an organization's secrets will increase as digitalization progresses.

If data is exchanged on an even broader scale than is the case today, then organizations may also be incited to invest more in high-quality data capture and more comprehensive metadata, since this will increase the value of their data for subsequent exchanges. A framework to facilitate data exchange may therefore also improve the availability of high-quality data. However, with more and more varied data becoming widely available, discovery of data will play an increasingly important role for a thriving data economy. Even today, matching problems between first parties which stream or store data required by another (third) party frequently occur. Any policy measure to foster the European data economy should include a framework that enables efficient data discovery.

Finally, organizations will need access to talent with the skills necessary for data analytics. The apparent shortage of these skills may curb the evolution of the European data economy. Boyd and Crawford (2012) point out that there is a particularly pronounced lack of practice with big data at universities, especially when only elite universities are able to afford access to big data.

Based on the insights gathered in this report, we feel that promoting reference architectures can address many of the challenges outlined. The following section discusses the proposition in detail, highlighting relevant examples of such reference architectures.

[73] This is particularly relevant when innovation is based on data that is costly to attain, e.g., in pharmaceutical trials. See e.g. Federico G, Morton FS, Shapiro C. 2019. Antitrust and Innovation: Welcoming and Protecting Disruption  In *Innovation Policy and the Economy, Volume 20*: University of Chicago Press.

## 6.2 Reference architectures as a way forward

> ***Insights**: Reference architectures provide a technology-neutral framework that can mitigate the key hindrances to frequent and high-quality data exchanges. Policymakers should promote the development of such architectures across sectors to facilitate the European data economy.*

Based on the insights gathered in the present study and, in particular, the key challenges for a data economy outlined in the preceding section, we suggest the promotion of reference architectures as a means to set the right framework conditions to enable an increase in the frequency and quality of data exchanges within and across sectors.

Such reference architectures essentially provide a consistent representation of the complex relationships within value networks emerging in the data economy as outlined in Section 3.1. In other words, a reference architecture comprises several layers and is a common language for data, processes, and interfaces that transcends departmental, company, sector, and industry boundaries. Such a common language helps to save transaction costs and enables higher efficiency than fragmented IT infrastructures commonly found today. Notably, the Industrial Data Space described in Section 5.2 already implements this logic in a vertical data-pooling setting.

The need for a common language for data exchanges is obvious: Srai et al (2019) argue that "*supply chain ontologies, while easily overlooked, provide the backbone of an implementable [digital twin] supply chain*" (p. 3). Cuquet and Fensel (2018) argue that data semantics, as a research and innovation topic, should constitute a central building block for reaping a large and positive impact of big data in the future in Europe. Data semantics play a critical role for improved efficiency, innovation, changing business models, employment, public funding as well as awareness building. By and large, academic and industry experts agree that compatibility and interoperability standards are required in order to create economic value from data and to strengthen a European data economy more widely. It is important that actors have the capability and ability to link and aggregate relevant datasets. This requires the development and implementation of uniform standards that enable the interoperability of data (HM Treasury 2018). Standards as such may, however, not be sufficient.

An holistic framework in the form of a reference architecture may be more appropriate.[74] Reference architectures can enable data transfer from one context to another. Since they incorporate relevant metadata in standardized formats, participating

[74] Notably, such a framework cannot anticipate every possible future reuse of data. A perfect framework is just as elusive as a perfect language. There will always be ambiguity: Laporte S. 2018. Ideal language. *KO KNOWLEDGE ORGANIZATION* 45: 586-608. Nonetheless, substantial improvements for data exchange and value creation from data can be expected as the following examples in the main text show.

organizations can use this information to understand how data was captured and may be able to trace it back to its source if necessary. Once established, reference architectures can significantly reduce the cost of producing such metadata due to their then-standardized nature. If they include information on the licensing of the exchanged data, they could also address concerns emerging with respect to the safeguarding of relevant trade secrets.[75] For instance, Carrière-Swallow and Haksar (2019) suggest licensing schemes that would be able to mitigate these concerns as they could clearly state the purposes for which the data may be used.[76] They can also improve the discoverability of potential collaborators and their capabilities as well as discovering opportunities for data offers, since one main function of a reference architecture is to provide logical positioning and classifications of new technologies, standards, and actors within value networks.

Reference architectures are therefore a critical step toward realizing the Industry 4.0 vision. They are able to coordinate the IT systems, technical resources, and capabilities and capacities of different companies and industries within a value network. A reference architecture has the task of providing a framework for structuring, developing, and implementing a value-added network and can also assist the integration as well as operation of the information systems. An overarching reference architecture—blurring the boundaries between industrial value chains and, eventually, industrial sectors—can be a way forward.

To understand the impact of reference architecture models, policymakers can look to the automobile sector, which was among the first sectors to implement such an architecture.[77] The reference architecture is known as “AUTOSAR”. It started out in 2003 as a partnership of significant actors in the automotive sector interested in developing open industry standards for an electrical/electronic (E/E) architecture. During the initial phase, BMW Group, Bosch, Continental, DaimlerChrysler, Ford Motor Company, PSA, Siemens VDO, Toyota Motor Company and Volkswagen were members of the partnership (Heinecke et al 2004). They had recognized that automotive E/E systems would continue to add new functions and that this, in turn, would increase complexity and expenses exponentially for each of them if they stuck to the plethora of proprietary standards that prevailed at the time. It took until 2005 for first release of the (now classic) AUTOSAR platform.[78] At the start of Phase 2, 166 partners had joined AUTOSAR and by December 2016 there were 191 partners. The platform was developed further into a reference architecture and is currently revised and

[75] Smart contracts, e.g., based on blockchain technology could be instrumental in such a framework as outlined by Roman D, Stefano G. 2016. *Towards a Reference Architecture for Trusted Data Marketplaces: The Credit Scoring Perspective.* Presented at 2nd International Conference on Open and Big Data (OBD), Vienna.

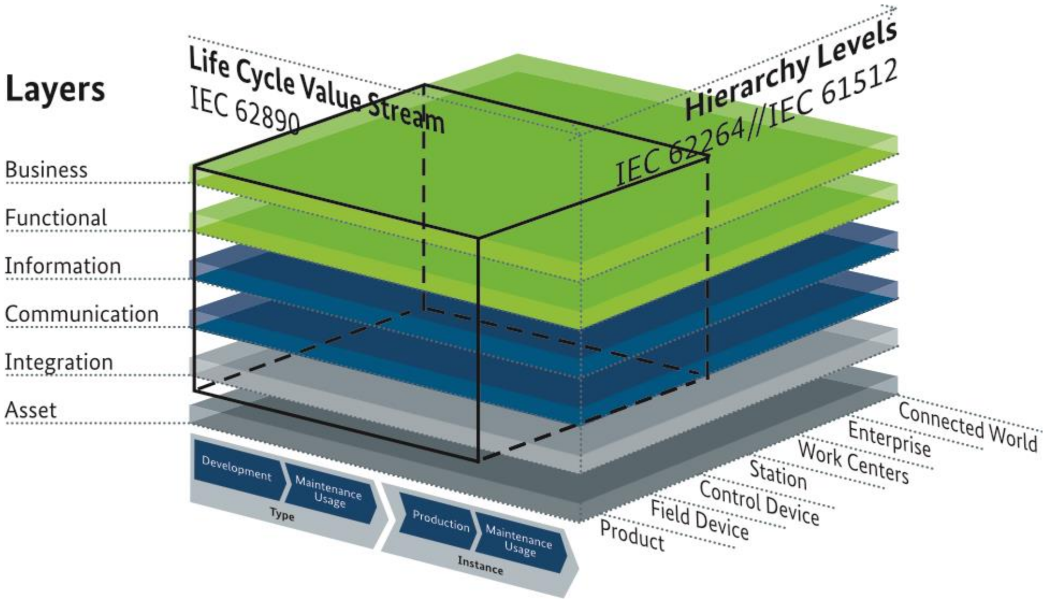
[76] Notably, enforcing such schemes might be difficult in practice due to non-rivalry and virtual non-traceability of data’s misuse. A thriving data economy consequently has to build on trust among the agents involved.

[77] Industry reference architectures have also been established in agriculture “ISOBUS,” for smart grids and in the health sector.

[78] Release 1.0.0 was published on June 26, 2005.

expanded into an adaptive platform.[79] A literature review by Dersten et al (2011) indicates both positive and negative impacts across various areas. For instance, the implementation of AUTOSAR reduced complexity, enabled improved quality management, and increased software reliability. It also improved the position of suppliers, who could develop solutions for various OEMs more easily. However, AUTOSAR also led to performance risks and a trade-off between memory and response times in E/E systems. Within companies, new processes had to be developed requiring substantial investments and internal frictions. The continued success of an integral architecture in the automotive sector is, however, underscored by recent analysis by Continental, who claim that advanced architectures sustain innovation in the automotive sector. For instance, around a third of the wires within a modern vehicle can be saved by applying a vehicle system architecture, resulting in significant weight reduction and thus reduction of fuel consumption and emissions (Locks & Winkler 2017).

Figure 6-1: RAMI4.0 as an illustration of a reference architecture for Industry 4.0



Source: DIN SPEC 913 cf. Arnold and Liebe (2018).

While traditional reference architectures remained somewhat limited to individual sectors, an overarching reference architecture model for Industry 4.0 (RAMI 4.0)[80] was developed in Germany within the framework of the Industry 4.0 platform.[81] Figure 6-1

[79] https://www.autosar.org/about/history/.

[80] For more detail on RAMI 4.0 and its impact, see Arnold R, Liebe A. 2018. Digitale Wertschöpfungsnetzwerke und RAMI 4.0 im hessischen Mittelstand, Hessisches Ministerium für Wirtschaft, Energie, Verkehr und Landesentwicklung, Wiesbaden.

[81] The following stakeholders were involved: BITKOM, VDMA, ZVEI and VDI associations cooperated with various research institutions and business partners under the coordination of the Industry 4.0 platform working group "Reference Architectures, Standards and Standardization."

depicts the layered model. Eventually, RAMI 4.0 will provide even small- and medium-sized enterprises (SMEs) with a tool that opens up and promotes the introduction and implementation of manufacturer-independent IT-solutions and data exchange. Agreement on standards is a critical aspect of cross-company data exchange, which in turn enables communication between machines or between components and machines. Only if there is a common semantics of the data, can it work as a network in a horizontal manner.[82]

As already mentioned in Section 3.3 on innovation in the data economy, the concept of the digital twin is central to the vision of Industry 4.0. It provides a unified digital representation of the asset characteristics within a complex value network. This is in clear contrast to today's application-centered solutions, which are mostly based on highly heterogeneous proprietary systems. When data is exchanged across such systems, media breaks are common, as are inconsistent and redundant data. Low-quality data and a lack of efficiency of data exchanges are the result. The concept of digital twin is taken up within RAMI 4.0 as the administration shell. The administration shell makes available the full digital representation of physical entities both in horizontal networking, i.e. across companies, and in vertical networking, i.e. within an entity.

As an extension, reference architectures provide the basic structure and the framework for industrial data platforms (IDPs). Examples of IDPs include the Cooperative ITS for Mobility in European Cities (CIMEC) and openEASE. IDPs provide a way to pool sector-relevant data and are of particular value for the development of new services, artificial intelligence (AI) applications, and innovative business processes. The key success factor of these projects will be striking a balance between sharing data effectively and holistically while preserving the digital sovereignty of the participants, which is one of the key barriers for many stakeholders to participate in data exchanges.

However, the potential impact of such reference architectures goes well beyond industrial applications, as examples from Estonia and South Korea highlight. The success of the Estonian eGovernment implementation rests in part on the reference architecture they employed. Already, in 2005, the Estonian Ministry of Economic Affairs issued a strategic document on enforcing semantic interoperability of state information technology. The strategy focused on data objects and input/output parameters of data services, all of which are semantically described within the system. The government supports this effort continuously with corresponding policies, guidelines, tools, and educational and promotional activities. The reference architecture was ultimately set up in 2009. Initially, there was relatively limited success. A key hurdle was the complexity of existing ontologies with which public services were unable to cope; and so, a specific ontology was developed. By 2011, 240 million online data requests were being made

[82] An essential step in this direction is the integration of the Open Platform Communications Unified Architecture (OPC UA) in RAMI 4.0. The OPC UA standard fits seamlessly into RAMI 4.0. With the successive integration of standards, RAMI 4.0 gains both in importance and impact.

annually by Estonian citizens and public services, enabled by the reference architecture (Haav & Küngas 2013). Among them, 94% of tax returns were made online (cf Kütt & Priisalu 2014). Today, 99% of services are online.[83]

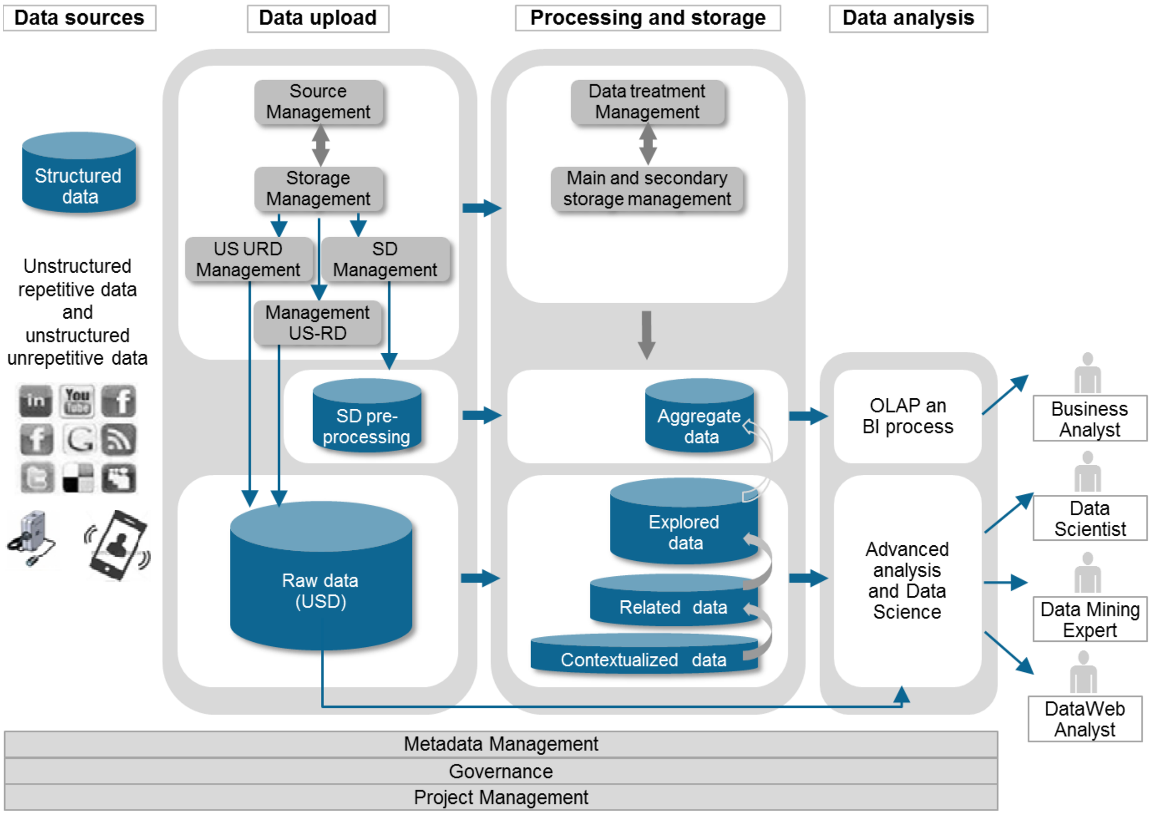
The Korean National Standard Reference Data (SRD) Program is somewhat less comprehensive than the architecture developed in Estonia and can only be considered as part of a reference architecture. Nonetheless, it points to a potential best practice of the government, facilitating data exchange through promoting standardization and establishing a framework for trusted data exchanges. The SRD was legislated in 1999 and implemented in 2006. A key component of the program was the setting-up of clear data (quality) evaluation guidelines. Over the past decade, 43,000 databases have been developed and have become part of the SRD across scientific, social and industrial fields (Lee 2019).

In essence, our insights gathered in this report, as well as the examples of reference architectures and similar data exchange architectures serving a similar purpose, highlight that data shared without its metadata may in fact be much less valuable than it could be, if it were embedded in a clearly defined ontology. Access to large pools of users can mitigate some of these issues, as the users can be asked to “tag” data with reference to specific content or other interpretations. The rationale is that they will eventually generate the correct results, because a majority of users will provide the right interpretation of the piece of data. However, this approach may work well in some contexts, but not in others, as the users might bring their cultural bias to the task, which may or may not be valuable for the purpose that the data is to be used for (Caracciolo et al 2018, Stuckenschmidt 2012). A further advantage of integrating specific ontologies into reference architectures is that the metadata can remain machine-readable and linked to the data on any information system. This is not true for most implementations of existing major big data solutions, which typically embed the metadata in the content data (Marshall 2012).

When promoting such an approach, policymakers should pay attention to the lessons already learned in data warehousing as, for instance, brought forward by Inmon (2006), Inmon et al (2010) or Kimball (2011), with respect to the emerging requirements for big data. Incorporating these insights, Salinas and Lemus (2017) put forward a multilayer staggered architecture model for big data (see Figure 6-2). The model allows for handling of all kinds of data: acquisition, cleaning, integration, identification, analysis, and management of data quality. It also includes transversal components for data storage, metadata, lifecycle, and security handling.

[83] https://e-estonia.com/

Figure 6-2: A multilayer staggered architecture model for big data



Source: Salinas and Lemus (2017).

# 7 Conclusion

This study set out to educate the debate around the data economy in Europe, with a specific focus on exploring potential concerns regarding control over data and the introduction of a potential mandated data access. We find that the popular debate lacks nuance and imposes sometimes wrong assumptions on the data economy. In particular, the wide variations in data and the role that data quality plays are neglected in favor of a focus on data quantity.

To educate the debate in a meaningful way, we developed a layered framework providing a primer on how data is captured and handled, and how insights are generated such that information of economic value can be extracted. Our framework highlights the complexity and the various interactions that are part of the process. It further highlights that data collection is inherently constitutive, thus the claim that data is objective fact cannot be upheld. This has implications on the value of data as such, as data commonly only generates value in a specific context and for a certain purpose for which it was captured. While this does not preclude data from being used in another context, it illustrates that there will likely be some costs involved in making the data suitable for another context and purpose.

Due to the rapidly increasing digitalization and established means of data exchange between first parties (collectors of data) and third parties (firms which are granted access to data, but without immediate control over the process of data capture), data access as such does not appear to be a hindrance to the data economy in general. Fostering data exchanges would nonetheless have a positive economic impact, as businesses could operate more efficiently and more varied data could be captured as duplicate efforts could be avoided.

However, an interventionist policy approach does not seem to be appropriate. Instead, we recommend actively facilitating the development of reference architectures to establish compatibility and interoperability of IT infrastructures, along with data formats and processing protocols. Positive examples from industry (AUTOSAR and RAMI 4.0) as well as from governments (data management architectures in Estonia and South Korea) underscore the potential success of such frameworks. Ultimately, the development of such reference architectures should naturally remain with market actors and standard-setting organizations such as the IEEE.

Naturally, such a facilitation of industry engagement would have to be complemented by efforts to increase the talent pool and to foster a generally supportive business environment, including sufficient venture capital, R&D support and appropriate regulation of privacy and security. To this end, a concerted effort of institutions and authorities on the Member State and European level is necessary.

## References

Adner R, Chen J, Zhu F. 2016. Frenemies in Platform Markets: The Case of Apple's iPad vs. Amazon's Kindle

Adomavicius G, Bockstedt JC, Curley SP, Zhang J. 2017. Effects of online recommendations on consumers' willingness to pay. *Information Systems Research* 29: 84-102

Aguirre E, Mahr D, Grewal D, de Ruyter K, Wetzels M. 2015. Unraveling the Personalization Paradox: The Effect of Information Collection and Trust-Building Strategies on Online Advertisement Effectiveness. *Journal of Retailing* 91: 34-49

Ansari A, Mela CF. 2003. E-customization. *Journal of Marketing Research,* 40: 131-45

Anthes G. 2015. Data Brokers are Watching You. *Communications of the ACM* 58: 28-30

Areeda P, Hoverkamp H. 1988. An Analysis of Antitrust Principles and Their Application. *Antitrust Law* 736: 675-701

Argenton C, Prüfer J. 2012. Search engine competition with network externalities. *Journal of Competition Law & Economics* 8: 73-105

Armstrong M. 2006. Competition in Two-Sided Markets. *The RAND Journal of Economics* 37: 668-91

Arnaut C, Pont M, Scaria E, Berghmans A, Leconte S. 2018. Study on data sharing between companies in Europe. A study prepared for the European Commission, European Commission; everis, Luxembourg

Arnold R, Bott J, Hildebrandt C, Schäfer S, Tenbrock S. 2016. Internet-basierte Plattformen und ihre Bedeutung in Deutschland, Wissenschaftliches Institut für Infrastruktur und Kommunikationsdienste (WIK), Bad Honnef

Arnold R, Hildebrandt C. 2017. The Socio-Economic Impact of Online Platforms, Wissenschaftliches Institut für Infrastruktur und Kommunikationsdienste (WIK), Bad Honnef

Arnold R, Liebe A. 2018. Digitale Wertschöpfungsnetzwerke und RAMI 4.0 im hessischen Mittelstand, Hessisches Ministerium für Wirtschaft, Energie, Verkehr und Landesentwicklung, Wiesbaden

Arnold R, Schiffer M, Pols A. 2013. Wirtschaft Digitalisiert - Welche Rolle spielt das Internet für die deutsche Industrie und Dienstleister?, IW Consult and BITKOM, Cologne, Berlin

Arnold R, Waldburger M. 2015. The Economic Influence of Data and their Impact on Business Models In *Trends in Telecommunication Reform 2015 - Getting Ready for the Digital Economy*, ed. ITU, pp. 153-83. Geneva: International Telecommunication Union

Attard J, Orlandi F, Auer S. *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI), 13-16 Oct. 20162016*: 453-56. IEEE.

Attard J, Orlandi F, Auer S. *International Conference om Theory and Practice of Electronic Governance, New Delhi, India, 2017*: 475-784.

Austin PC, Goldwasser MA. 2008. Pisces did not have Increased Heart Failure: Data-driven Comparisons of Binary Proportions between Levels of a Categorical Variable can Result in Incorrect Statistical Significance Levels. *Journal of Clinical Epidemiology* 61: 295-300

Austin PC, Mamdani MM, Juurlink DN, Hux JE. 2006. Testing Multiple Statistical Hypotheses Resulted in Spurious Associations: a Study of Astrological Signs and Health. *Journal of Clinical Epidemiology* 59: 871-72

Awad NF, Krishnan MS. 2006. The Personalization Privacy Paradox: An Empirical Evaluation of Information Transparency and the Willingness to Be Profiled Online for Personalization. *MIS Quarterly* 30: 13-28

Banko M, Brill E. *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics2001*: 26-33. Association for Computational Linguistics.

Barbero M, Bartz K, Linz F, Mauritz S, Wauters P, et al. 2018a. Study to support the review of Directive 2003/98/EC on the re-use of public sector information, Deloitte, Brussels

Barbero M, Cocoru D, Graux H, Hillebrand A, Linz F, et al. 2018b. Study on emerging issues of data ownership, interoperability, (re-)usability and access to data, and liability, Deloitte, Brussels

Barney J. 1991. Firm resources and sustained competitive advantage. *Journal of management* 17: 99-120

Benbasat I, Wang W. 2005. Trust in and adoption of online recommendation agents. *Journal of the association for information systems* 6: 4

Benlian A. 2015. Web personalization cues and their differential effects on user assessments of website value. *Journal of Management Information Systems* 32: 225-60

BEREC. 2019. BEREC Report on the Data Economy, Body of European Regulators for Electronic Communications

Berente N, Seidel S, Safadi H. 2018. Research Commentary—Data-Driven Computationally Intensive Theory Development. *Information Systems Research* 30: 50-64

Bergemann D, Bonatti A. 2019. Markets for Information: An Introduction. *Annual Review of Economics* 11: 85-107

Bernstein ES. 2012. The transparency paradox: A role for privacy in organizational learning and operational control. *Administrative Science Quarterly* 57: 181-216

Bertenrath R, Arnold R, Koppel O, Lang T. 2011. Innovation Policy and the Business Cycle: Innovation Policy's Role in Addressing Economic Downturn - INNO-Grips Policy Brief No. 1, European Commission, Cologne/Brussels

Blank G, Lutz C. 2017. Representativeness of social media in Great Britain: investigating Facebook, LinkedIn, Twitter, Pinterest, Google+, and Instagram. *American Behavioral Scientist* 61: 741-56

Bleier A, Eisenbeiss M. 2015a. The importance of trust for personalized online advertising. *Journal of Retailing* 91: 390-409

Bleier A, Eisenbeiss M. 2015b. Personalized online advertising effectiveness: The interplay of what, when, and where. *Marketing Science* 34: 669-88

Boerman SC, Kruikemeier S, Zuiderveen Borgesius FJ. 2017. Online Behavioral Advertising: A Literature Review and Research Agenda. *Journal of Advertising* 46: 363-76

Borgman CL. 2010. Research Data: Who will share what, with whom, when, and why? - (RatSWD Working Paper Series Nr. 161, Berlin

Bott J, Hildebrandt C, Arnold R. 2018. Die Nutzung von Daten durch OTT-Dienste zur Abschöpfung von Aufmerksamkeit und Zahlungsbereitschaft: Implikationen für Wettbewerb, Regulierung sowie Daten- und Verbraucherschutz - WIK-Diskussionsbeitrag Nr. 431, Wissenschaftliches Institut für Infrastruktur und Kommunikationsdienste (WIK), Bad Honnef

Boyd D, Crawford K. 2012. Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. *Information, Communication & Society* 15: 662-79

Brandimarte L, Acquisti A, Loewenstein G. 2012. Misplaced Confidences: Privacy and the Control Paradox. *Social Psychological and Personality Science* 4: 340-47

Bründl S, Matt C, Hess T. 2015. Wertschöpfung in Datenmärkten-eine explorative Untersuchung am Beispiel des deutschen Marktes für persönliche Daten*2199-8914*, Forum Privatheit und selbstbestimmtes Leben in der digitalen Welt, Karlsruhe

Buchanan JM. 1965. An Economic Theory of Clubs. *Economica* 32: 1-14

BVDW. 2018. Datenwertschöpfung und Qualität von Daten, Bundesverband Digitale Wirtschaft (BVDW) e.V., Düsseldorf

Caillaud B, Jullien B. 2003. Chicken & Egg: Competition Among Intermediation Service Providers. *The RAND Journal of Economics* 34: 309-28

Calvino F, Criscuolo C, Marcolin L, Squicciarini M. 2018. A taxonomy of digital intensive sectors, Organisation for Economic Co-operation and Development, Paris

Cao L. 2017. Data science: a comprehensive overview. *ACM Computing Surveys (CSUR)* 50: 43

Capurro R. 1978. *Information. Ein Beitrag zur etymologischen und ideengeschichtlichen Begründung des Informationsbegriffs*. München: Saur Verlag.

Caracciolo C, Aubin S, Whitehead B, Zervas P. *Research Conference on Metadata and Semantics Research2018*: 340-45. Springer.

Carrière-Swallow Y, Haksar V. 2019. The Economics and Implications of Data - An Integrated Perspective, Washington, DC

Casadesus-Masanell R, Hervas-Drane A. 2015. Competing with privacy. *Management Science* 61: 229-46

Casadesus-Masanell R, Ruiz-Aliseda F. 2009. Platform competition, compatibility, and social efficiency. In *Working Paper*: Harward Business School

Checkland P. 1999. Systems thinking. *Rethinking management information systems: An interdisciplinary perspective*: 45-56

Chellappa RK, Sin RG. 2005. Personalization versus privacy: An empirical examination of the online consumer's dilemma. *Information Technology and Management* 6: 181-202

Chen M, Mao S, Liu Y. 2014. Big data: A survey. *Mobile networks and applications* 19: 171-209

Cichy C, Rass S. 2019. An Overview of Data Quality Frameworks. *IEEE Access* 7: 24634-48

Cornes R, Sandler T. 1986. *The Theory of Externalities, Public Goods, and Club Goods*. Cambridge, MI: Cambridge University Press.

Crémer J, de Montjoye Y-A, Schweitzer H. 2019. Competition policy for the digital era - Final report, European Commission - DG COMP, Brussels

Cuquet M, Fensel A. 2018. The societal impact of big data: A research roadmap for Europe. *Technology in Society* 54: 74-86

Curry E. 2016. The Big Data Value Chain: Definition, Concepts, and Theoretical Approaches In *New Horizons for a Data-Driven Economy - A Roadmap for Usage and Exploitation of Big Data in Europe*, ed. JM Cavanillas, E Curry, W Wahlster, pp. 29-38: SpringerOpen

Davenport TH, Barth P, Bean R. 2012. How 'big data' is different. *MIT Sloan Management Review* 54: 22-24

Davis M, Martinez R, Kalaboukis C. 2010. Rethinking Personal Information – Workshop Pre-read, Invention Arts and World Economic Forum, Cologny

de Mauro A, Greco M, Grimaldi M. 2016. A Formal Definition of Big Data based on its Essential Features. *Library Review* 65: 122-35

Dersten S, Axelsson J, Fröberg J. 2011. *Effect Analysis of the Introduction of AUTOSAR: A Systematic Literature Review*. 239-46 pp.

Dhar V. 2013. Data science and prediction. *Communications of the ACM* 56: 64-73

Dosis A, Sand-Zantman W. 2019. The ownership of data. *Toulouse School of Economics. Working Paper*

Drucker J. 2011. Humanities approaches to graphical display. *Digital Humanities Quarterly* 5: 1-21

Duch-Brown N, Martens B, Mueller-Langer F. 2017. The economics of ownership, access and trade in digital data. JRC Digital Economy Working Paper 2017-01, Joint Research Centre, Seville

Egan E. 2019. Charting a Way Forward: Data Portability and Privacy, Facebook

European Commission. 2017. Communication from the Commission Commission to the European Economic and Social Committee and the Committee of the Regions - "Building a European Data Economy" (SWD(2017) 2 final), European Commission, Brussels

European Commission, IMF, OECD, UN, World Bank. 2009. System of National Accounts 2008, European Commission, International Monetary Fund, Organisation for Economic Co-operation and Development, United Nations and the World Bank, New York, NY

Evans DS, Noel M. 2005. Defining Antitrust Markets When Firms Operate Two-Sided Platforms. *Columbia Business Law Review* 3: 101-34

Evans DS, Schmalensee R. 2007. Antitrust Analysis of Multi-sided Platforms: The Industrial Organization of Markets with Two-sided Platforms. *Competition Policy International* 3: 151-79

Fan J, Han F, Liu H. 2014. Challenges of Big Data analysis. *National Science Review* 1: 293-314

Farboodi M, Veldkamp L. 2019. A Growth Model of the Data Economy, Working Paper, Columbia Business School, New York, June 20, New York. NY

Fast V, Schnurr D, Wohlfarth M. 2019. *Data-driven Market Power: An Overview of Economic Benefits and Competitive Advantages from Big Data Use.* Presented at 47th Research Conference on Communications, Information, and Internet Policy (TPRC), Washington, DC

Federico G, Morton FS, Shapiro C. 2019. Antitrust and Innovation: Welcoming and Protecting Disruption  In *Innovation Policy and the Economy, Volume 20*: University of Chicago Press

Feijóo C, Gómez-Barroso J-L, Aggarwal S. 2016. Economics of Big Data  In *Handbook on the Economics of the Internet*, ed. JM Bauer, M Latzer, pp. 510-25. Cheltenham: Edward Elgar Publishing

Feld H. 2019. The Case for the Digital Platform Act: Market Structure and Regulation of Digital Platforms, Roosevelt Institute

Floridi L. 2010. Information: a very short guide. Oxford University Press, Oxford

Frické M. 2009. The knowledge pyramid: a critique of the DIKW hierarchy. *Journal of information science* 35: 131-42

Frické M. 2015. Big Data and its Epistemology. *Journal of the Association for Information Science and Technology* 66: 651-61

FTC. 2014. Data Brokers: A Call for Transparency and Accountability, Washington, DC

Furman J, Coyle D, Fletcher A, McAuley D, Marsden P. 2019. Unlocking digital competition - Report of Digital Competition Expert Panel, HM Treasury, London

Furner J. 2016. “Data”: The data  In *Information Cultures in the Digital Age - A Festschrift in Honor of Rafael Capurro*, ed. M Kelly, J Bielby, pp. 287-306. Wiesbaden: Springer

Gandomi A, Haider M. 2015. Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management* 35: 137-44

Gasser U, Palfrey J. 2007. Breaking Down Digital Barrieres - When and How ICT Interoperability Drives Innovation

Gitelman L. 2013. *“Raw data” is an oxymoron*. Cambridge, MA: MIT press.

Goldfarb A, Tucker C. 2019. Digital Economics. *Journal of Economic Literature* 51: 3-43

Goldfarb A, Tucker CE. 2011a. Online Display Advertising: Targeting and Obtrusiveness. *Marketing Science* 30: 389-404

Goldfarb A, Tucker CE. 2011b. Privacy Regulation and Online Advertising. *Management Science* 57: 57-71

Graef I, Wahyuningtyas SY, Valcke P. 2015. Assessing Data Access Issues in Online Platforms. *Telecommunications Policy* 39: 375-87

Granados N, Gupta A. 2013. Transparency strategy: competing with information in a digital world. *MIS quarterly* 37: 637-41

Grover V, Chiang RHL, Liang T-P, Zhang D. 2018. Creating Strategic Business Value from Big Data Analytics: A Research Framework. *Journal of Management Information Systems* 35: 388-423

GSMA. 2018. The Data Value Chain, GSMA

Günther WA, Mehrizi MHR, Huysman M, Feldberg F. 2017. Debating big data: A literature review on realizing value from big data. *The Journal of Strategic Information Systems* 26: 191-209

Haav H-M, Küngas P. 2013. Semantic data interoperability: the key problem of big data. *Big Data Computing*: 245

Hagiu A. 2007. Merchant or two-sided platform? *Review of Network Economics* 6: 115-33

Haucap J. 2019. Competition and Competition Policy in a Data-Driven Economy. *Intereconomics*: 201-08

Haucap J, Stühmeier T. 2016. Competition and Antitrust in Internet Markets  In *Handbook on the Economics of the Internet*, ed. JM Bauer, M Latzer, pp. 183-210. Cheltenham, UK: Edward Elgar Publishing

Haug A, Zachariassen F, Liempd D. 2011. The costs of poor data quality. *Journal of Industrial Engineering and Management* 4: 168-93

Heinecke H, Schnelle K-P, Fennel H, Bortolazzi J, Lundh L, et al. 2004. Automotive open system architecture-an industry-wide initiative to manage the complexity of emerging automotive e/e-architectures, SAE Technical Paper

Hestness J, Narang S, Ardalani N, Diamos G, Jun H, et al. 2017. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*

Heumann S, Jentzsch N. 2019. Wettbewerb um Daten. Über Datenpools zu Innovationen. Berlin: Stiftung Neue Verantwortung e. V.

Hey T, Tansley S, Tolle K. 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmont, WA: Microsoft Research.

Hildebrandt C, Arnold R. 2016. Big Data und OTT-Geschäftsmodelle sowie daraus resultierende Wettbewerbsprobleme und Herausforderungen bei Datenschutz und Verbraucherschutz

- WIK-Diskussionsbeitrag Nr. 414, Wissenschaftliches Institut für Infrastruktur und Kommunikationsdienste (WIK), Bad Honnef

Hinz O, Eckert J. 2010. The impact of search and recommendation systems on sales in electronic commerce. *Business & Information Systems Engineering* 2: 67-77

Hjørland B. 2019. Data (with Big Data and Database Semantics). *KO Knowledge Organization* 45: 685-708

Ho SY, Bodoff D, Tam KY. 2011. Timing of adaptive web personalization and its effects on online consumer behavior. *Information Systems Research* 22: 660-79

Hogan O, Holdgate L, Jayasuriya R. 2016. The Value of Big Data and the Internet of Things to the UK Economy, Cebr, London

Horak R. 2008. *Telecommunications and data communications handbook*. Hoboken: John Wiley & Sons.

IDC, Open Evidence. 2017. European Data Market SMART 2013/0063 - Final Report. A study prepared for the European Commission, IDC, Open Evidence

Inmon WH. 2006. DW 2.0; Architecture for the Next Generation of Data Warehousing. *Information Management* 16: 8

Inmon WH, Strauss D, Neushloss G. 2010. *DW 2.0: The architecture for the next generation of data warehousing*. Amsterdam: Elsevier.

Jensen HE. 1950. Editorial Note In *Through Values to Social Interpretation: Essays on Social Contexts, Actions, Types and Prospects*, ed. H Becker, pp. vii-xi. Durham, NC: Duke University Press

Jentzsch N, Sapi G, Suleymanova I. 2013. Targeted Pricing and Customer Data Sharing Among Rivals. *International Journal of Industrial Organization* 31: 131-44

Jones CI, Tonetti C. 2019. Nonrivalry and the Economics of Data*0898-2937*, National Bureau of Economic Research

Junqué de Fortuny E, Martens D, Provost F. 2013. Predictive Modeling with Big Data: Is Bigger Really Better? *Big Data* 1: 215-26

Kaase M. 2001. Databases, Core: Political Science and Political Behavior In *International Encyclopedia of the Social and Behavioral Sciences*, ed. NJ Smelser, PB Baltes, pp. 3251-55. Amsterdam: Elsevier

Karwatzki S, Dytynko O, Trenz M, Veit D. 2017. Beyond the Personalization–Privacy Paradox: Privacy Valuation, Transparency Features, and Service Personalization. *Journal of Management Information Systems* 34: 369-400

Katz ML, Shapiro C. 1985. Network Externalities, Competition, and Compatibility. *The American Economic Review* 75: 424-40

Kempermann H, Lichtblau K. 2012. Definition und Messung von hybrider Wertschöpfung. *IW Trends* 39: 1-20

Kim T, Barasz K, John LK. 2019. Why Am I Seeing This Ad? The Effect of Ad Transparency on Ad Effectiveness. *Journal of Consumer Research* 45: 906-32

Kimball R. 2011. The evolving role of the enterprise data warehouse in the era of big data analytics - White Paper

Kitchin R. 2014. *The data revolution: Big data, open data, data infrastructures and their consequences*. London: Sage.

Krämer J, Wohlfarth M. 2018. Market power, regulatory convergence, and the role of data in digital markets. *Telecommunications Policy* 42: 154-71

Kütt A, Priisalu J. *Proceedings of the International Conference on e-Learning, e-Business, Enterprise Information Systems, and e-Government (EEE)2014*: 1. The Steering Committee of The World Congress in Computer Science, Computer ….

Lam WMW, Liu X. 2018. Does Data Portability Facilitate Entry?

Lambrecht A, Tucker CE. 2013. When Does Retargeting Work? Information Specificity in Online Advertising. *Journal of Marketing Research* 50: 561-76

Lambrecht A, Tucker CE. 2015. Can Big Data Protect a Firm from Competition? : SSRN

Laporte S. 2018. Ideal language. *KO KNOWLEDGE ORGANIZATION* 45: 586-608

Lee D. 2019. Big Data Quality Assurance Through Data Traceability: A Case Study of the National Standard Reference Data Program of Korea. *IEEE Access* 7: 36294-99

Lee I. 2017. Big data: Dimensions, evolution, impacts, and challenges. *Business Horizons* 60: 293-303

Lenart M, Bielecki A, Lesot M-J, Petrisor T, d'Allonnes AR. *International Conference on Artificial Intelligence and Soft Computing2018*: 579-91. Springer.

Lerner AV. 2014. The Role of 'Big Data' in Online Platform Competition. SSRN 2482780

Lewis RA, Rao JM. 2015. The Unfavorable Economics of Measuring the Returns to Advertising. *The Quarterly Journal of Economics* 130: 1941-73

Li X, Ling CX, Wang H. 2016. The convergence behavior of naive Bayes on large sparse datasets. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 11: 10:1-10:24

Lichtblau K, Arnold R. 2012. Smart Industry – Intelligente Industrie: Eine neue Betrachtungsweise der Industrie. Ergebnisse einer Studie der Institut der deutschen Wirtschaft Köln Consult GmbH für das Land Hessen, Initiative Industrieplatz Hessen, Neu-Isenburg

Lind H-G, Suckfüll H. 2013. Die Initiative zu einer Deutchen Daten Treuhand (DEDATE) als ultima ratio der persönlichen digitalen Datenwirtschaft (PDD) - Ansatätze und Strukturen für eine gezielte Verwertung persönlicher Daten unter Berücksichtigung aller Interessengruppen - Dateneigentümer, Wirtschaft und Staat, Fraunhofer, HSBD GmbH, Leipzig

Liu J, Li J, Li W, Wu J. 2016. Rethinking big data: A review on the data quality and usage issues. *ISPRS journal of photogrammetry and remote sensing* 115: 134-42

Liu Q, Serfes K. 2006. Customer information sharing among rival firms. *European Economic Review* 50: 1571-600

Locks O, Winkler G. 2017. *Future Vehicle System Architecture.* Presented at ASAM General Assembly,

Lundvall B-Å, Borrás S. 2005. Science, technology and innovation policy In *The Oxford handbook of innovation*, ed. J Fagerberg, DC Mowery, RR Nelson, pp. 599-631. Oxford: Oxford University Press

Mahnke RP. 2015. Big Data as a Barrier to Entry. *Antitrust Chronicle* 12: 1-6

Manyika J, Chui M, Groves P, Steve F, Kuiken V, Doshi EA. 2013. Open data: Unlocking innovation and performance with liquid information, McKinsey Global Institute

Marshall P. 2012. What you need to know about big data. *Government Computer News*, 07. February 2012:

Martens B. 2016. An Economic Policy Perspective on Online Platforms, European Commission

McKinsey. 2016. The age of analytics: Competing in a data-driven world

Mirzaie M, Behkamal B, Paydar S. 2019. Big Data Quality: A systematic literature review and future research directions. *arXiv preprint arXiv:1904.05353*

Moore FT. 1959. Economies of scale: Some statistical evidence. *The Quarterly Journal of Economics* 73: 232-45

Morton FS, Bouvier P, Ezrachi A, Jullien B, Katz R, et al. 2019. Stigler Committee on Digital Platforms - Final report, Stigler Center

Muschalle A, Stahl F, Löser A, Vossen G. *international workshop on business intelligence for the real-time enterprise2012*: 129-44. Springer.

Nissenbaum H. 2011. A Contextual Approach to Privacy Online. *Daedalus* 140: 32-48

Norberg PA, Horne DR, Horne DA. 2007. The Privacy Paradox: Personal Information Disclosure Intentions versus Behaviors. *Journal of Consumer Affairs* 41: 100-26

O'Mahony MP, Hurley NJ, Silvestre G. *Proceedings of the 11th international conference on Intelligent user interfaces2006*: 109-15. ACM.

OECD. 1996. The Essential Facilities Concept, Paris

OECD. 2015. The OECD Model Survey on ICT Usage by Businesses - 2nd revision, Organization for Economic Co-operation and Development, Paris

OECD. 2016. Big Data: Bringing Competition Policy to the Digital Era, OECD

OECD. 2018. Rethinking Antitrust Tools for Multi-Sided Platforms, OECD

OECD, Eurostat. 2018. *Oslo Manual 2018: Guidelines for Collecting, Reporting and Using Data on Innovation, 4th Edition, The Measurement of Scientific, Technological and Innovation Activities*. Paris and Luxembourg: OECD Publishing and Eurostat.

Ohm P. 2010. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *UCLA Law Review* 57: 1701-77

Otto B, Auer S, Cirullies J, Jürjens J, Menz N, et al. 2016. Industrial data space: digital sovereignty over data, Fraunhofer, München

Otto B, Lohmann S, Steinbuß S, Teuscher A. 2018. IDS reference architecture model, industrial data space, version 2.0. *International Data Spaces Association & Fraunhofer*

Panzar JC, Willig RD. 1981. Economies of scope. *The American Economic Review* 71: 268-72

Parker GG, van Alstyne MW. 2005. Two-sided network effects: A theory of information product design. *Management science* 51: 1494-504

Pauer A, Nagel L, Fedkenhauser T, Fritzsche-Sterr Y, Resetko A. 2018. Data exchange as a first step towards data economy, London

Paunov C, Planes-Satorra S. 2019. How are digital technologies changing innovation? Evidence from agriculture, the automotive industry and retail, Organisation for Economic Co-operation and Development, Paris

Peppers D, Rogers M, Dorf B. 1999. Is your company ready for one-to-one marketing. *Harvard Business Review* 77: 151-60

Pine BJ, Peppers D, Rogers M. 1995. *Do you want to keep your customers forever?* Boston: Harvard Business Press.

Pipino LL, Lee YW, Wang RY. 2002. Data quality assessment. *Communications of the ACM* 45: 211-18

Planes-Satorra S, Paunov C. 2019. The Digital Innovation Policy Landscape in 2019. OECD Science, Technology and Innovation Policy Papers No. 71, Organisation for Economic Co-operation and Development, Paris

Porter ME. 1985. *Competitive Advantage: Creating and sustaining superior performance*. New York: The Free Press.

Prüfer J, Schottmüller C. 2017. Competing with big data. *Tilburg Law School Research Paper*

Qi Z, Wang H, Li J, Gao H. 2018. Impacts of dirty data: and experimental evaluation. *arXiv preprint arXiv:1803.06071*

Redman TC. 1998. The impact of poor data quality on the typical enterprise. *Communications of the ACM* 41: 79-82

Redman TC. 2016. Bad data costs the US $3 trillion per year. *Harvard Business Review*, 22. September 2016:

Rocher L, Hendrickx JM, de Montjoye Y-A. 2019. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications* 10: 3069

Rochet JC, Tirole J. 2003. Platform Competition in Two-Sided Markets. *Journal of the European Economic Association* 1: 990-1029

Roman D, Stefano G. 2016. *Towards a Reference Architecture for Trusted Data Marketplaces: The Credit Scoring Perspective.* Presented at 2nd International Conference on Open and Big Data (OBD), Vienna

Rubinfeld DL, Gal MS. 2017. Access Barriers to Big Data. *Arizona Law Review* 59

Rysman M. 2009. The Economics of Two-Sided Markets. *The Journal of Economic Perspectives* 23: 125-43

Salgado D, Esteban E, Saldana S, Oancea B, Sakarovitch B, et al. 2018. *Estimation of population counts combining official data and aggregated mobile phone data.* Presented at European Conference on Quality in Official Statistics, Kraków

Salinas SO, Lemus AC. 2017. Data warehouse and big data integration. *Int. Journal of Comp. Sci. and Inf. Tech* 9: 1-17

Samat S, Acquisti A, Babcock L. *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017), Santa Clara, 2017*: 299-319.

Saudagar M, Ye M, Al-Otaibi S, Al-Jarba K. 2019. Smart Manufacturing: Hope or Hype? *CEP* 115: 43-48

Schafer JB, Konstan JA, Riedl J. 2001. E-commerce recommendation applications. *Data Mining and Knowledge Discovery* 5: 115-53

Schepp N-P, Wambach A. 2016. On Big Data and its Relevance for Market Power Assessment. *Journal of European Competition Law & Practice* 7: 120-24

Schroeder R. 2016. Big data business models: Challenges and opportunities. *Cogent Social Sciences* 2

Schumann JH, von Wangenheim F, Groene N. 2014. Targeted online advertising: Using reciprocity appeals to increase acceptance among users of free web services. *Journal of Marketing* 78: 59-75

Schwartz R, Dodge J, Smith NA, Etzioni O. 2019. Green AI. *arXiv preprint arXiv:1907.10597*

Schweitzer H. 2019. Datenzugang in der Datenökonomie: Eckpfeiler einer neuen Informationsordnung. *GRUR* 121: 569-80

Shapiro C, Varian HR. 1999. *Information Rules. A Strategic Guide to the Network Economy*. Boston, MA: Harvard Business School Press.

Sidak JG, Lipsky AB. 1999. Essential Facilities. *Stanford Law Review* 51: 1187-249

Sidi F, Panahy PHS, Affendey LS, Jabar MA, Ibrahim H, Mustapha A. *2012 International Conference on Information Retrieval & Knowledge Management2012*: 300-04. IEEE.

Sinha R, Swearingen K. *CHI'02 Extended Abstracts on Human Factors in Computing Systems, Minneapolis, 2002*: 830-31.

Srai JS, Settanni E, Tsolakis N, Aulakh PK. 2019. *Supply Chain Digital Twins: Opportunities and Challenges Beyond the Hype.* Presented at 23rd Cambridge International Manufacturing Symposium, Cambridge

Stuckenschmidt H. 2012. Data semantics on the web. Springer

Swire P, Lagos Y. 2013. Why the Right to Data Portability Likely Reduces Consumer Welfare: Antitrust and Privacy Critique. *Maryland Law Review* 72: 353-80

Taleb I, Dssouli R, Serhani MA. *2015 IEEE international congress on big data2015*: 191-98. IEEE.

Taleb I, Serhani MA, Dssouli R. *2018 International Conference on Innovations in Information Technology (IIT)2018*: 69-74. IEEE.

Tam KY, Ho SY. 2006. Understanding the impact of web personalization on user information processing and decision outcomes. *MIS Quarterly* 30: 865-90

Tao F, Zhang H, Liu A, Nee AY. 2018. Digital twin in industry: state-of-the-art. *IEEE Transactions on Industrial Informatics* 15: 2405-15

Thirumalai S, Sinha KK. 2013. To personalize or not to personalize online purchase interactions: implications of self-selection by retailers. *Information Systems Research* 24: 683-708

Tiefenbacher K, Olbrich S. *ECIS, Münster, Germany, 2015*.

Tsai JY, Egelman S, Cranor L, Acquisti A. 2011. The effect of online privacy information on purchasing behavior: An experimental study. *Information Systems Research* 22: 254-68

Tucker CE. 2010. The Economic Value of Online Customer Data. *Economics of Personal Data and Privacy* 30

Tucker CE. 2014. Social networks, personalized advertising, and privacy controls. *Journal of Marketing Research* 51: 546-62

Tucker DS, Wellford HB. 2014. Big Mistakes Regarding Big Data. In *Antitrust Source, American Bar Association*

Wang RY, Strong DM. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems* 12: 5-33

Wang W, Xu J, Wang M. 2018. Effects of recommendation neutrality and sponsorship disclosure on trust vs. distrust in online recommendation agents: Moderating role of explanations for organic recommendations. *Management Science* 64: 5198-219

Wiengarten L, Zwick M. 2018. Neue digitale Daten in der amtlichen Statistik. *WISTA* 2017: 43-60

Wohlfarth M. 2019. Data Portability on the Internet. *Business & Information Systems Engineering* 61: 551-74

World Economic Forum. 2011. Personal Data: The Emergence of a New Asset Class, WEF, Cologny

Zins C. 2007. Conceptual approaches for defining data, information, and knowledge. *Journal of the American society for information science and technology* 58: 479-93