KONRAD
ADENAUER
STIFTUNG

The Emerging Global Data Economy

Implications and Options for the Transatlantic Relationship

Olaf J. Groth, PhD, CEO, Cambrian LLC

Tobias Straube, VP Analysis, Cambrian LLC

with contributions by

Prof Andreas Moring, Senior Analyst, Cambrian LLC

Carl Lange, Junior Analyst, Cambrian LLC

CAMBRIAN
FUTURES

www.cambrian. ai

APRIL 2022

www.kas.de

Table of Contents

Executive Summary

Digital technologies are transforming all parts of economic and social life and data is at the core of this transformation. **Data is becoming a production factor, similar to labor or energy, and increasingly drives value creation in the public and private sectors**. The EU Commission projects that the data economy in its member states will be worth EUR 829 billion in 2025, which reflects 2.4 per cent of the EU's GDP. The U.S. Department of Commerce calculated that the digital economy of the U.S. accounted for nine percent of its GDP in 2018. The value created by digitally transformed enterprises is estimated to surpass the value creation from non-digitally transformed enterprises by 2023. Moreover, data-powered innovations open **new horizons for individual and communal human development** in areas such as health care, education, transport and climate change. Considering the efficiencies that digital services bring to the traditional economy and the unpredictable innovations to come, these estimates might even be understatements. Thus, it is no surprise that **thinking about data has evolved from a research and business domain to a (geo-)political and policy domain**.

Hence, it is imperative that leaders and decision-makers in government, private sector and civil society on both sides of the Atlantic understand the current shifts in the datasphere in order to agree on objectives and strategies to enable its transformation for economic and social benefit. **Contributing to this understanding by providing insights** on the **nature, trends and drivers** of the global datasphere, as well as the **technical and legal questions** around it **is the purpose of this study**.

The study builds on a **value chain framework of the global datasphere** that structures the different **types of generated data** (personal/non-personal, real/synthetic) and the actors and processes involved in the subsequent value creation steps from data storing, pooling, curation and data brokerage to the design of digital applications.

The Internet of Everything (IoE), which refers to physical objects such as cars or mobile phone cameras, produces data and represents in its entirety the data sphere in terms of hardware and software. In 2020, 64.2 zettabytes (ZB) of data was created or replicated, up from 2 ZB in 2010 and 18 ZB in 2016.[1] **Chapter 1** shows how the **global datasphere continues to grow in quantity** at a compound annual growth rate (CAGR) of 23% through 2025[2] as an increasing amount of both personal and non-personal data is created and exchanged through the internet and stored and processed through cloud services. This is driven by **increased internet connectivity**, especially in the Global South and **expanding broadband capacities** coupled with a **steep increase of**

personal and IoT device availability. The Covid-19 pandemic has certainly accelerated the use of personal end devices. In 2021, 23 billion photos were uploaded on Instagram alone. Every second, close to 100 000 searches are conducted on Google, and about four million smartphones are sold every day. But not just people share text, image, and video data on internet-based platforms. Also machines are increasingly creating and exchanging **non-personal data** with each other in the **Internet of Everything**. In fact, **machine-to-machine (M2M) connections will make more than 50 percent of all global internet connections by 2023**. The number of IoT devices that are connected to the internet has reached 8.7 billion in 2021 and will continue to grow exponentially in the 2020s. To date, Global North countries produce the largest volumes of non-personal data estimated by IoT devices. However, countries in the Global South, especially in **Africa, demonstrate steep growth** curves. This provides an **opportunity for European (e.g. Orange, Vodacom) and North American key players** to provide such infrastructure and broadband services.

The existence of data alone does not drive value. Hence, in the data value chain, making this data accessible represents the next step beyond data generation. The accessibility of data depends on how data is stored and processed, either in a data center or on the end device. General and industry-specific platforms then aggregate and integrate data, enabling an ecosystem of innovators and service providers to develop products and services that ultimately create value by driving digital transformation in business and society. In the logic of the data value chain, **chapter 2 analyzes the trends in data pooling, curation, trade and access**. Today, much data is trapped either in organizational, industry or platform silos. **Data sharing, especially among organizations, is still in its infancy, prompting calls for regulatory and technological innovation to promote data sharing and, increasingly, data trading** – similar to the market mechanisms that already exist for other economic inputs such as energy or commodities. Increasingly, platform companies such as Amazon Web Services or Alphabet are creating these environments as they scale and commercialize their data collection and processing capabilities while using their capital to enter new sectors such as education or code-sharing. **Platforms become super-platforms**. Consequently, the tech giants are becoming the central gatekeepers for access to data, compute power and knowledge, thus increasing entry-barriers for smaller companies and the risk of abusive data use. Conversely, new technological developments and approaches in the **Web 3.0 promise a decentralized data-sharing infrastructure**: through distributed ledgers at large, and blockchain in particular. Finally, due to recent governance and technological developments, data is likely to become a corporate asset presented on corporate balance sheets. The discussion around "**data productization**" will facilitate its trade like any other economic good.

Looking at the top of the data value chain, the availability of large quantities of data, commonly referred to as Big Data, has been considered a prerequisite for creating value by building intelligent digital innovations and services. It is widely assumed that the more data is available, the more patterns, trends and associations can be derived from it. However, as the study shows in **chapter 3**, this is about to change. Technical innovations (e.g., automated data cleaning) in the field of artificial intelligence (AI) enable a shift towards data efficient algorithms. This is not only necessary due to **data scarcity that remains prevalent in critical areas**, such as medicine and drug discovery, but also due to a need for more **energy efficiency in algorithm training to reduce related greenhouse gas emissions**. Coupled with the **spread of Automated Machine Learning (AutoML) approaches and low-code and no-code applications**, which makes the development of AI applications possible even for non-experts, a new wave of data and code efficient AI solutions can be expected. The consequent **shift from big data to small data** and the **democratization of AI** will positively change the cost-benefit ratio of AI projects in organizations.

The final **chapter 4** draws on the analysis of the trends and provides concrete policy recommendations on how to enable **more data flow among the transatlantic partners** and beyond while ensuring data privacy and empowering individuals and smaller companies to take back control over their data. Recommendations further advocate for smart regulation to data trading and monetization, as well as for emerging phenomena of the Web 3.0. In order to ensure equitable and sustainable growth in the new phase of the data economy, technology should be added to ESG frameworks and organizations need to rethink how to build the right talent pool. Finally, for acquiring a leadership position in the evolving global data sphere, actors need to think about “cognification” as the next frontier.

Chapter 1:

Growth of the Datasphere

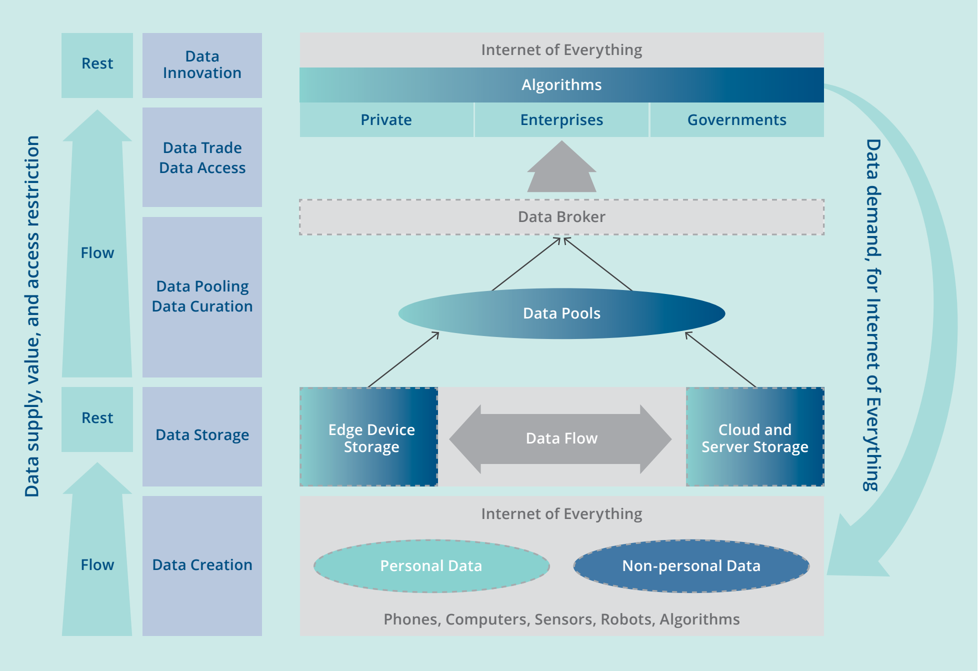
1.1 Introduction

The global datasphere refers to all digital data that is created and replicated worldwide, as well as the environment in which it is stored and processed. The size and shape of the datasphere is the starting point for the digital economy, as it determines the opportunities in the subsequent value creation steps, which include data storing, data pooling and curation, data brokerage and digital applications that transform businesses and society. The value creation by digitally transformed enterprises made up 37 per cent of global GDP in 2020 already and is estimated to surpass the value creation from non-digitally transformed enterprises in 2023.[3]

To gain insight into the relationship between the digital economy and data, it is worth taking a look at the data value chain (see chart 1). Data is the source of value creation in the digital economy. According to the data value chain, the Internet of Everything (IoE), which refers to physical objects such as cars or mobile phone cameras, produces data and represents in its entirety the data sphere in terms of hardware and software. In 2020, 64.2 zettabytes (ZB) of data were created or replicated, up from 2 ZB in 2010 and 18 ZB in 2016.[4] This data sphere is to grow at a compound annual growth rate (CAGR) of 23% through 2025.[5] The existence of data alone, however, does not drive value. Hence, in the data value chain, making this data accessible represents the next step beyond data generation in the IoE. The accessibility of data depends on how data is stored and processed, either in a data center or on the end device. However, many data storage systems are silos, limited in scope and do not allow correlating data. This is where platform providers come in. From software development, as in the case of Salesforce, to agriculture, as in the case of U.S. agricultural machinery manufacturer John Deere, platforms aggregate and integrate data, enabling an ecosystem of innovators and service providers to develop products and services that ultimately create value by driving digital transformation in business and society.

However, digital innovations that are creating social or economic value are only the end result of the data economy value chain. It begins with the creation of data by either end devices **(real data)** or algorithms **(synthetic data)**, together constituting the Internet of Everything. Contrary to popular belief, the size of the datasphere is merely one factor that determines its value and use downstream the data value chain. **Small data**, rather than big data, refers to the ability of new algorithms to derive insights

Chart 1:
The Data Value Chain



and recognize patterns from small data pools (see chapter 3.2). Another distinction is made between data that contains information about specific, identifiable individuals (**personal data**) and **non-personal data** that does not contain personal information, including anonymized personal data. This distinction is important because use cases and data flows are restricted by different legal frameworks, depending on whether data is considered personal or not. The EU General Data Protection Regulation (GDPR), for example, explicitly refers only to personal data and not to non-personal data.

After its creation, the data is **"in flow"** both between countries and regions while it is transferred to be stored centrally in data centers and clouds or decentrally on end devices. When data reaches its storage destination it is **"at rest"** until it is transferred again to the next segment in the value chain. Interestingly, only about 10 percent of the created data is actually stored somewhere, according to IDC.[6] Depending on the use case, the data is then pooled in **data lakes** or curated in **data warehouses**, either within organizations or by **platform companies** (see more on these concepts in Chapter 3.2). This bundling and processing of data is the prerequisite for innovators to use data directly or indirectly via intermediaries. Such intermediaries can be **data brokers**, which are either devices or organizations facilitating the access, flow, and trade of data towards public and private users which in turn generate digital innovations from it. Data brokers are all those organizations that establish access to stored data as a service, for example via an Application Programming Interface (API) or a platform.

Economic or political actors who want access to data face two types of challenges: technical access and legal access. Technical access describes the actual technology with which the data broker grants the data user access to the stored data. Legal access determines whether the data user actually has the right to access stored data. Given that access challenges are solved, innovators can then build applications that in turn create new data. This creates a circle of data flows that starts and ends with the Internet of Everything generating and using the data.

In the following chapters, we build on this framework to analyze the growth of the datasphere, the trends in the Internet of Everything, i.e., data sources, and global data flows and how these trends contribute to increases in value generation from data.

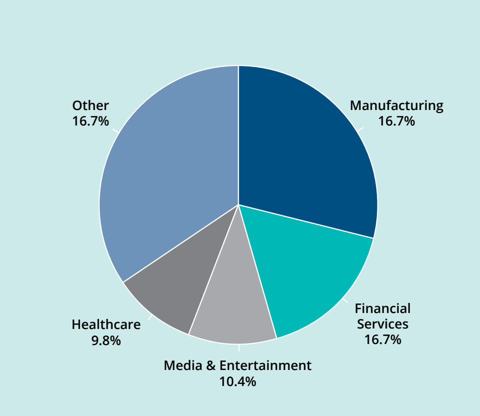
1.2 Growth, flow and drivers of the personal datasphere

In 2021, 23 billion photos were uploaded on Instagram alone. Every second, close to 100,000 searches are conducted on Google, and about 4 million smartphones are sold every day.[7] Between 2016 and 2020, mobile internet traffic increased by a factor of 10 globally which is strongly associated with a growing personal datasphere. From 2018 to 2020, every internet user used 300 percent more bandwidth. This increase is twice as large as the one from 2016 to 2018.[8] Meanwhile, an additional 7 percent of the world's population became internet users.[9] This contributes to rapid growth of personal data worldwide. Global data flows provide benefits to all industries and regions. Nevertheless, the number of data localization measures and personal data flow restrictions enacted by countries all over the world grows increasingly fragmented despite their negative impact on human and economic growth driven by digital technologies.

While the COVID-19 pandemic has certainly accelerated these ongoing trends as aspects of daily life moved to the digital space, growth patterns of the global datasphere across global regions vary greatly. This can be assessed by using the datasphere framework above. Trends in the availability of end devices such as computers, smart phones, and Internet of Things devices (e.g., smart home) mean more personal data is created. Trends in internet traffic capacity, broadband speed, and decreasing internet volume prices on the other hand are indications for both, data creation and data storage. An analysis along these indicators shows that the data sector is most mature in the U.S., Europe and Asia, and yet continues to grow.

Generally, growth in internet connectivity and internet volume prices show only small annual increases in the Global North e.g., in the U.S. and in the EU.[10] However, this high saturation does not represent a growth limit for the datasphere but will allow for a deepening and broadening of it. The high level of internet penetration contributes to the fact that more and more end devices are connected to the internet in Europe and the U.S. with 485 and 302 million users respectively in 2021.[11] Moreover, each individual mobile phone created an average of 70 percent more data from 2017 to 2019. Some industries stand out in growing the sphere of personal data. From the personal-data-heavy industry sectors, the healthcare sector and the media and entertainment sector each generated about 1.2 ZB in 2018. In contrast, the non-personal-data-heavy sectors manufacturing and financial services placed first and second overall in terms of generated data, with 3 ZB and 2 ZB each[12] (see chart 2). The healthcare sector is projected to grow the fastest with a CAGR of 36 per cent from 2018-2025 (the global

Chart 2:
Data Creation by Industry Sectors in 2018



datasphere has a projected CAGR of 27 per cent in that time frame).[13] This steep growth is partially due to the digitization of diagnostics and medical imagery analysis in the cloud. However, the trend shows that all industries are on the move to benefit from and create large amounts of data. This is caused by cross-industry, data-heavy services such as internationally spread cloud computing centers.[14]

But important consumer growth markets can also be found outside the datasphere in Europe, the US and China. Interestingly, the countries with the highest data creation per mobile phone are predominantly small, Global South countries. A possible explanation for that large per-device data creation is the low, yet increasing, mobile phone and computer availability in some of these countries. In Kenya for example, every internet user occupied about 4000 Mbit/s of internet bandwidth in 2017. This is the tenth highest bandwidth usage per user in the world — about 600 percent more than the bandwidth usage of the average German internet user. At the same time, only seven percent of all Kenyian households had access to a computer, in contrast to 88 percent in Germany, and only 33 percent of the Kenyan population had a mobile broadband subscription. Furthermore, Kenya exhibited low internet connectivity, with only 17 percent using the internet in 2017.[15] The combination of low availability of computers and connectivity on the one hand, but high availability of mobile data and mobile phones on the other, suggests that the dataspehere's potential lies mainly in the consumption of digital services rather than in their production. This is also evident when other growth indicators are taken into account.

Generally, Global South countries have the highest growth potential in terms of internet users. Eight of the ten countries with the lowest internet connectivity rates in 2019 worldwide were Sub-Saharan countries.[16] In addition, populations in these countries grow rapidly. Brazil and Nigeria are already the fifth- and sixth-largest countries in the world. Nigeria is predicted to be the third largest country by 2100 with an increase in population by 530 million people.[17] When larger shares of rapidly growing populations become connected to the internet their data creation volume will increase accordingly. This observation caused Kai Fu Lee, a prominent Taiwanese-born AI scientist, executive, investor and author in China, to conclude that: "Whatever company wants to lead in AI and wants to become the next Facebook or Google needs to have a strategy to tap into the markets of developing countries – this is where the consumers of tomorrow live."[18]

Wherever data is created, it may be used in another location in the global data economy. This requires the transfer of data via the internet, putting the data at flow. Large flows of data indicate a large volume of data being created and used for economic value generation. Out of all regions, Europe occupied the most internet bandwidth with 503 terabytes per second (Tb/s) in 2021[19] (on average, 503 TB of data were transferred via internet broadband in Europe every second in 2021). Flows show that only about 25 percent of this data was relocated outside the region. This explains Europe's large internet bandwidth use, yet small international data flow. In contrast, North America leads with respect to interregional data flow as more than 80 percent of its broadband bandwidth connects to the regions Latin America, Europe, and Asia. Three of the five largest data flows come in and out of North America (see chart 3). Apart from Europe, Asia was the only other region with more than 50 percent of the data traffic being intraregional. The Global South does not play a large role in interregional data flows yet but will do so in the future: between 2017 and 2022, Africa's bandwidth usage grew the most in relative terms with, 45 percent on average per year. This is followed by Asia's bandwidth which increased by 37 percent each year. North America's bandwidth experienced the least relative growth out of all regions at a CAGR of 23 percent.[20]

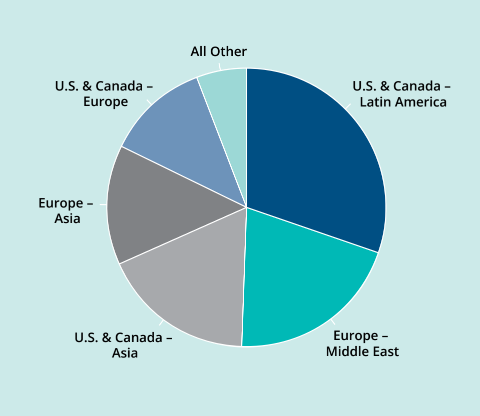


Chart 3:
Largest inter-regional data flows in internet bandwidth in 2021

These global cross-border data flows of personal data are still hindered by disparate data regulations. Data regulations globally increased to 144 regulations in 2021.[21] Although the European GDPR for personal data created a legal framework for all EU member states, fragmentation remains as it is enforced at national levels. Between the EU and the US, the flow of personal data is hindered by the repeal of the Privacy Shield (a framework for GDPR-compliant data transfers between the US and the EU) in 2020. As a result, companies will have to use standard contractual clauses (SCCs) when transferring data. SCCs force businesses to evaluate their data flows on a tedious and costly case-to-case basis in a complex legal environment. The New Economics Foundation estimates that an affected small business could experience costs of up to 13,500 USD.[22] While the EU and US lack a regulatory framework for data sharing between them, other markets tend to overregulate data flows. China, India, Russia, and Turkey form the four most restrictive countries, enacting 57 data regulations in total. Most of their regulations have a protectionist motivation and concern the localization of non-personal financial and accounting data and personal data about citizens.[23] China's Personal Information Protection Law ("PIPL") is the latest piece of a full data protection framework at the time of release of this report, covering not only personal data, but also restricting storage and export of certain kinds of non-personal data, e.g., on essential infrastructure and mapping services.[24] This will impact foreign and domestic businesses in China as, for example, self-driving car manufacturers rely on high-resolution mapping data for their products. Conversely, a large body of econometric literature shows that data flow restrictions decrease a country's total productivity and trade output.[25] Hence, reducing restrictions on data flow overall and creating data flow protection standards with like-minded countries should be the goal for the transatlantic partners.

A second factor for data flow is data creation versus data demand. Europe's largest outgoing flow of data is to the U.S, which had the second highest traffic volume in 2018. Little flow between the EU and Africa is due to the little data storage and flow infrastructure via broadband in Africa, which registered 27 TB/s of bandwidth in 2021.[26] More than 75 percent of Africa's used broadband bandwidth flows to Europe. The 2nd largest share of broadband bandwidth is intra-regional within Africa. However, as growing populations in Asia and Africa will create more data within the next few years, global data flows may change. For the EU to be able to profit from the newly created data, local data regulations have to allow for interregional data flow.

Today's patterns of data flows may change through increases in data creation and data regulation policies. Internet connectivity, broadband speed and end device availability are all drivers of the personal datasphere for which Global South countries score poorly. However, taking the high per-user data creation rate, few data regulations and population growth trends of these countries into account, Global South countries will be important contributors to the personal datasphere in the future. The magnitude of their potential critically depends on offering end devices and network services for prices that are affordable and accessible. While this might at first conflict with production or operation costs, it should be seen as an investment in enabling more end devices and network usage which fuels the growth of the personal data pillar.

1.3 Growth trends and drivers of the non-personal datasphere

In 2020, The European finance hubs Frankfurt and London created 185 GB of internet traffic every second, ranking them first and second worldwide.[27] Machine-to-machine (M2M) connections will make up more than 50 percent of all global internet connections by 2023[28] (see chart 4). The number of IoT devices that are connected to the internet reached 8.7 billion in 2021 and is projected to scale to 55.7 billion by 2025. This fleet of IoT devices alone could produce 73.1 ZB, more data than the global datasphere in 2020, according to IDC analysts.[29] A steep increase of the number of IoT devices connected to the internet combined with increased broadband availability, both in terms of infrastructure and affordability, are the main drivers and enablers of non-personal data creation. Cloud availability and rules on data sharing in turn determine the chance of value creation from the generated data. Strong regulation on data sharing may inhibit interregional data exchange, hence limiting the chance of driving digital innovation.

Common sources of non-personal data include sensors, manufacturing robots, or stock market servers. Non-personal data is useful for a wide range of intelligence, from business processes and predictive maintenance in Industry 4.0 to energy efficiency ratings of real estate. Within the Internet of Everything, smart devices and sensors increasingly gather data, e.g., on business processes like automated manufacturing, financial transactions, as well as cloud data flow. IDC forecasts that in 2021 Enterprise IoT, used for improving business processes, would be the largest application area of IoT in terms of spending, before Industrial IoT and IoT consumer applications like smart homes.[30] (see chart 5).

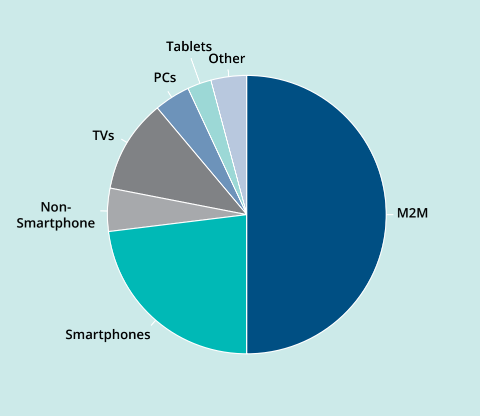


Chart 4:
Share of Device Connections predicted for 2023 by CISCO

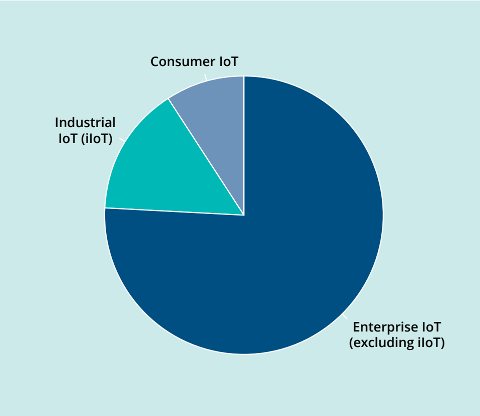


Chart 5:
IoT spendings in 2021 by sub-market

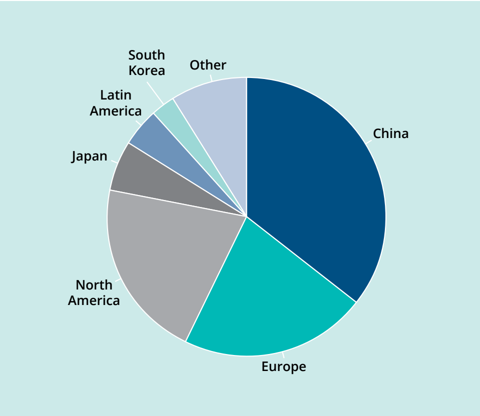


Chart 6:
Number of connected IoT devices in 2019 by region

This indicates that the growing IoT market is the major contributor to the growing non-personal datasphere. Breaking this datasphere into regions, APACxJ[31] region contributed the largest share to global IoT spendings in 2018 with 35 percent, followed by North America and EMEA[32]. Within these broader geographic categories, China, Europe and the U.S. lead the IoT device market with registered devices, respectively (see chart 6). While China, the U.S., and Europe hold strong positions in the IoT market and therefore IoT data creation, other regions show the largest growth rates. Between 2019 and 2030, IoT device connections will experience the largest relative growth in India and South Asia (19 percent CAGR) and Sub-Saharan Africa (18 percent CAGR), as predictions show.[33] In contrast, today's leading regions register CAGRs between 9 and 11 percent. This large growth can be explained by those regions having few connected IoT devices in absolute number and thus a large potential for growth. Despite smaller CAGRs, China, the U.S., and Europe are forecasted to remain the region's leading in IoT device connections. This data can be interpreted as Global North countries producing the largest volumes of non-personal data estimated by IoT devices. However, countries in the Global South, especially in Africa, demonstrate steep growth curves and also spend a larger share of their GDP on IoT.

Creation of data from more IoT devices is only the first step to value generation. To make non-personal data valuable, businesses have to be able to upload and store the data. Large upload volumes require broadband speed and cloud service availability. Fixed-broadband connection is the most common internet connection type among enterprises and therefore functions as an indicator of non-personal data creation.[34] According to the Net Vitality Index, a ranking of countries based on multiple broadband-related statistics and measurements, China, the US, Germany, the UK, and Canada lead the global broadband ecosystem.[35] In 2019, China led the world in annual fixed-broadband usage in absolute numbers with 600 exabytes or 0.6 zettabytes. In comparison, Germany used only 63 exabytes of fixed broadband in 2019.[36] In 2019, Europe among all regions used the most fixed-broadband internet bandwidth, a good indicator of non-personal data volume.[37] However, Asia had the most interregional data flow. The majority of data created in Europe remained in European networks and storage facilities, while data exchange between North America and Asia was flourishing. Fixed-broadband traffic volume shows a very clear dichotomy between Global North and Global South countries: in 2019, no African country was among the top 70 countries in terms of fixed-broadband speed or top 30 countries in terms of fixed-broadband volume. This provides an opportunity for European and North American key players to provide such infrastructure and broadband services. The resulting decrease in internet traffic costs may elevate African businesses to a higher level in the data economy value chain.

Beside fixed-broadband speed for uploading data, the availability of cloud services for storing created data is an important factor for growth in the non-personal datasphere. Regional cloud storage availability can be approximated by the number of cloud on-ramps of the cloud provider. Cloud on-ramps are the technological gateway and access to cloud storage for end users and businesses. To utilize large quantities of data, companies rely on cloud providers such as Amazon Web Services (AWS), Alibaba Cloud, and Microsoft Azure. Different cloud providers have different regional focuses in terms of infrastructure and on-ramps. For instance, U.S.-founded companies such as AWS, Google Cloud, and Microsoft Azure record a fairly even distribution of on-ramps between the EU and U.S., as well as China and Japan. Asian e-commerce and tech giant Alibaba however focuses its cloud service mostly on Asia.[38] 69 percent of the global cloud market is occupied by U.S. companies such as Amazon, Microsoft, and Google. Asian providers such as Alibaba and Tencent cover the 4th and 7th largest market shares.[39] 88 percent of all global cloud data centers are located in Asia, Europe, and the U.S.[40] Regions with few cloud availability have little access to the benefits of cloud computing and storage for businesses. So far, European cloud service providers do not play a significant role in the storage of the data value chain. How this shortcoming of European cloud providers could be compensated by a shift towards edge computing will further be discussed in chapter 2.3. Notably, the distribution of cloud on-ramps across cloud providers globally correlates well with regional fixed-broadband use. In Africa and South America there are only few on-ramps of the big cloud players while there are no competitive local cloud services. This provides a disadvantage for business firms. Even if they create large volumes of non-personal data, these data cannot be stored and utilized efficiently and at scale, leading to less growth of the non-personal datasphere.

The forthcoming European Data Act may be part of a strategy to facilitate further data sharing and increase value creation from data in the EU. The Data Act aims to promote B2B and B2G data sharing by several means, such as ensuring contractual compliance for all businesses and incentives for businesses to share their data. Furthermore, the Data Act plans to build data spaces in which data can be aggregated and pooled safely to increase its value[41] (see Data Pooling in the datasphere framework). For the EU to benefit the most, a regulatory framework should not only be agreed upon by today's main datasphere players like the EU itself and the U.S. It should also involve the datasphere drivers of tomorrow, e.g., Nigeria and Brazil. Thus, EU's project GAIA-X may be a step in the right direction, provided that the project manages to achieve its desired goal of building a trustworthy and competitive cloud infrastructure by finding consensus between the diverse members. Though the project took off fast in 2020 and now has more than 850 members,[42] potential customers complain about delayed progress and bureaucracy and complexity hurdles, which is exactly what the project was intended to avoid. In the medium to long term, it can have the downside of indirectly denying access to rapidly growing data markets of the future by enforcing de-facto localization of data within member regions. In parallel, the African Continental Free Trade Area (AfCFTA) project of the African Union strives for a common framework on the exchange of both physical and digital goods and services across the continent.[43] This provides an opportunity for the EU to define inter-regional frameworks on data sharing, as well. Effort should be invested in propagating GAIA-X cloud standard as the European way of cloud computing and making this standard attractive for growing datasphere regions of Global South countries. If these countries decide to constrain internationally outflowing data severely like China or South Africa, their growing non-personal datasphere could become inaccessible for the EU.

Chapter 2:

Accessibility and the Evolution of Data Sharing

2.1 Introduction

The global datasphere drives the digital economy. Data-driven organizations and platforms serve as a catalyst by aggregating or processing data so that innovators can develop economic use cases and applications from it. In the digital economy data is an economic input, a factor of production. However, for many organizations data is not accessible in the same way as other production goods are, such as energy, raw materials, or labor. Thus, how access to data is organized plays a central role in the competitiveness of organizations and economies. In terms of the data value chain, data access essentially depends on where and how data is stored, pooled, or curated, in short, how data is shared.

The Support Centre for Data Sharing defines data sharing as a "collection of practices, technologies, cultural elements and legal frameworks that are relevant to transactions in any kind of information digitally, between different kinds of organizations".[44] Such practices have evolved over time. In fact, entire scientific disciplines, such as macroeconomics, social science or meteorology have only been able to emerge and develop over the past 100 years based on data being shared between governments and researchers and within the national and international research communities. The introduction of the internet made certain data accessible not only to research communities but to the economy at large. Today, organizations exchange data within or with others on a bilateral basis or via platforms. Usually, it is less the raw data that is exchanged but the insights derived from it. In the current data sphere, there are still significant limitations by way of centralization and filtering of data that prevent a ubiquitous flow of data for greater benefit. This is met with significant distrust vis-a-vis large platform companies, in terms of their data handling. Issues around privacy, agency, transparency, and objectivity have not yet been sufficiently addressed by actors in the data ecosystem. At the same time, new decentralized structures and institutions for data sharing emerge with the proliferation and advancement of digital technologies.

Thus, this chapter looks at how data sharing is organized within and between organizations, analyzes the changing role of platforms in collecting and commercializing data and examines the trend of distributed ledger technologies for decentralized data sharing. These different dimensions are both enablers and disrupters of each other. Some companies are progressively transforming into data-driven organizations that form data partnerships with other companies, or successfully adopt platform approaches. Others are still struggling to even implement the basics: sharing data across the enterprise. At the same time, digital startup trailblazers are already looking at the next evolution: blockchain-based data sharing. One implication of the evolving data sharing landscape is the increased recognition of data as an economic good, which can be traded.

2.2 Breaking up data silos within and between organizations

Digital data has become an important business asset and driver of profitability. In order to create value from data, organizations need to break up data silos within, but also need an environment in which data sharing with other organizations is safe and affordable.

Research by McKinsey suggests that companies that intensively analyze customer data are 23 times more likely to outperform their competitors in terms of new customer acquisition. Similarly, achieving above-average profitability is almost 19 times higher for customer-analytics champions than for laggards.[45] Consequently, an increasing number of companies invest in their ability to collect, analyze, and interpret data to become data-driven organizations to optimize their business model. Yet, there are many companies that recognize the value of data, but do not use it strategically or operationally. In a global survey of 900 companies, 84% of respondents said it is very important or critical to put data at the center of their key business decisions and strategies but only 56% said their companies consistently use data to drive business decisions.[46] Organizations must have the processes, structures, and people in place to identify data needs, collect or obtain the necessary data, and prepare the data for processing, for example, by structuring or cleansing it. Only with these foundations in place, companies can develop products or services and make decisions based on data.

One of the core constraints to leverage data sharing within an organization are data silos. Data silos are collections of data held by one department or group within a company, which are not easily or fully accessible by other groups in the same organization. Breaking up such silos and building a data-centric organization is both a strategic and a cultural issue. Data-centric organizations are characterized by treating data as the core intellectual property (IP) of the enterprise, while most companies use it only as a value-added tool to validate the effectiveness of sales strategies, for example. One way to drive the process towards becoming a data-driven organization is to introduce so-called data stewards. Data stewards are specific supervisory or data governance positions in organizations and companies that are generally responsible for ensuring the quality and adequacy of the organization’s data assets. Where they exist, data stewards are often located in IT departments. However, organizations will have to understand data stewardship as an interdepartmental task, if they wish to fully tear down data silos.

On top of that, organizations need a technical data architecture for managing data. Two approaches to modern data architecture have become established for this purpose: data lakes and data warehouses, both of which cover different needs and require different capabilities yet can be combined. Data lakes are data storages where any kind of data, such as text data, voice, picture or sensor data, can enter without meeting any quality or formatting requirements. As a result, data lakes are essentially unstructured data pools. Thus, making sense of data lakes requires more specialized experts, such as data scientists, data developers and business analysts. Data warehouses, on the other hand, seek to structure data into queryable components which are consistently governed and easy to consume or use for a scalable audience. This means that data is curated and structured, serving as a central version of the truth. This makes data consumable for data dashboards and purpose-built and scalable tools that can be used also by less data science-savvy business analysts. Both approaches are not necessarily competing. Data-centric companies often have both data lakes and data warehouses. Data lakes serve the collection of data from across the organization and as sources of “raw material” for data warehouses, which then fulfill more specific needs. The global growth forecasts show above all the growing importance of data lakes compared to data warehouses. The global data lake market is expected to grow at a compound annual growth rate (CAGR) of 20.6 percent from 2020 to 2027,[47] outpacing the growth rate of the data warehouse market, which is expected to be 10.7 percent from 2020 to 2028.[48] The need for data lakes is driven by the rapidly growing data sphere of unstructured data that results e.g. from increased IoT device availability and smart city initiatives, as well as the increased ability and application of AI to make sense of it.

While these trends show that data is increasingly being made available in the corporate world, leveraging economic and social assets through data also requires breaking down barriers that prevent data sharing between organizations. Fortunately, the technology for this is available. Application Programming Interfaces, in short APIs, make data sharing between software and applications possible. APIs, therefore, allow organizations to thrive on interconnectivity between applications, devices, and organizations and in doing so access to data and capabilities beyond organizational silos. Accordingly, APIs are understood to be the “digital reflection of an organization”.[49] The Postman API platform for example, used by more than 17 million developers in 800,000 plus organizations worldwide, provides data points and folders where API developers aggregate their API requests (Postman Collections). The number of Postman Collections skyrocketed from less than half a million to nearly 35 million between 2016 and 2020 alone,[50] pointing to the rapidly growing use of APIs. The State of the API Economy Report 2021 by Google Cloud adds to the picture by differentiating the types of APIs. According to that report, 50 percent of APIs are designed solely for uses within organizations, 44 percent were designed for both internal and external use, and 6 percent were exclusively designed for partners or developers outside the company.[51] The last data point underscores that sharing data between organizations is not just a technology issue, but one that depends on other factors, such as regulation, confidentiality, security, public perception, business model and data trading market design, to name just a few.

However, even if the technology is in place, cultural, legal, and business strategy barriers often block effective data sharing between organizations. For the past five years or so, NYU GovLab has been trying to change that by advocating for so-called “data collaboratives.” The concept of data collaboratives refers to partnerships between private and public entities to exchange their data to create public value. These can take the form of data pools, public interfaces, or research partnerships. Data pools are unified collections of datasets accessible by multiple parties, sometimes managed and maintained by an independent and trusted intermediary. Public interfaces refer to companies that grant open access to specific data sets, thereby allowing the independent use of the data by external parties. Research partnerships on the other hand are public-private partnerships, in which companies share certain proprietary data assets to derive insights for public value.[52] In 2021, the Data Collaborative Explorer,[53]a data collaborative repository managed and maintained by the NYU GovLab, counts 230 examples, ranging from data collaboratives in areas such as healthcare, environment, transportation and social inclusion. However, many data collaboratives remain one-off initiatives that do not scale or have expiration dates. Moreover, power imbalances between data owners and data users often impose transaction costs on the part of the data users that make data access very expensive or prohibit it altogether.

2.3 Platforms: Aggregators and gatekeepers of the datasphere

The digital economy is built to a large extent around platform companies that have established themselves as aggregators and gatekeepers to the datasphere. Many platform companies have experienced a countertrend to the global economy in the Corona pandemic. While the general economy suffers from the burden of pandemic-related constraints, a small group of digital platform companies outperforms on the stock markets. The oligopolistic power of these digital platforms shows no sign of abating, with governments taking reluctant steps to rein them in.[54] Meanwhile, platform companies spread from B2C business models to B2B, in the digital as well as in the digital-physical hybrid economy. In doing so, they are evolving their data collection and processing capabilities by commercializing their data center assets, by evolving from industry-specific platforms to super-platforms, or both. This section describes the emergence of the platforms and analyzes the development trends of the platforms towards data center providers and super platforms.

In the beginning of the internet (Web 1.0), users were mainly consumers of what were at the time rather static services and websites. Consumers could use e-mail communication, photo, and file storage, and read online newspapers. But the underlying infrastructure did not allow for much interaction or collaboration. This has been changing with the rise of the platform businesses from around 2006. Thanks to ever-improving AI and an abundance of computing power, platform companies were able to develop services that enabled sharing, collaboration, and interaction between different users. From the growing amount of user data that could be generated from these interactions, the platforms were able to gain insights and market them to advertisers. Consequently, the users of the Web 1.0 became data producers in what became known as Web 2.0. Today, platforms generate value by connecting data producers with data consumers. Social media platforms, for example, connect end users with the advertising companies. Until recently, however, unlike in China, platform companies in the U.S. and Europe have largely focused on vertical industries, (e.g., Uber), graphic design and content development services (e.g., Upwork), home services (e.g. Urban Clap) real estate (e.g. Airbnb), e-commerce (e.g. eBay), payments (e.g. Stripe) or software-as-a-service development (e.g. Salesforce), to name a few.

ZOOM IN: Understanding the superiority of platform businesses.

Of the 10 most valuable publicly listed companies in 2021, five are considered platform businesses (Apple, Microsoft, Alphabet, Amazon, Meta), compared to two in 2011 (Apple and Microsoft).[55] Platform businesses also outperform other public benchmarks such as Dow Jones Industrial and the Nasdaq Composite since 2016.[56] The success of platform businesses is built on two pillars. The underlying economic model of platforms and so-called network effects.

As an actor between the manufacturer or provider and the consumer, platforms create value by owning the means of connection, rather than the means of production as common in traditional business models. In linear business models, input factors such as raw materials or related services are considered expenses, and revenue is generated only from downstream value creation. In a platform business model, on the other hand, income can be generated from both or multiple sides of the platform. Apple's App Store, for example, has the ability to charge a fee for App developers and App users.

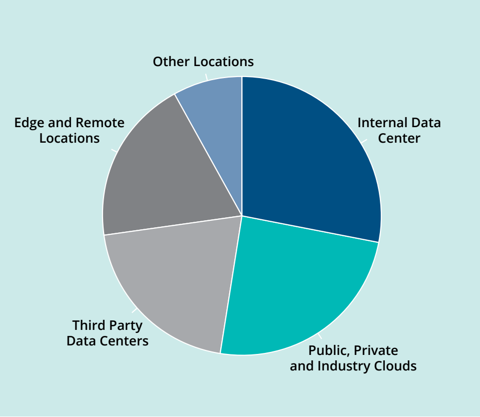
To attract a sufficient number of users and to be able to charge a premium, these companies make use of network effects. Network effects refer to business models where the value they provide to their customers increases as they scale and acquire more users. This is especially true for platform companies like ride-hailing, which are only as attractive to riders as the platform succeeds in attracting customers, and vice versa. Establishing and leveraging network effects requires strategies to enlist (getting the attention of), acquire (sign up), engage (transact with) and monetize (receive revenue from) users on and between different sides of the platform. The basic prerequisite for this is a convincing value proposition of the platform for every user. In the digital economy, this value proposition is the ability to analyze, predict and match user behavior and needs - based on the data provided by the different platform users, processed through AI and compute power. This also explains why some B2C social media platforms thrive off the creative or political tension between different user groups, as negative or emotion-invoking headlines stimulate engagement.

With these key vertical industries and markets exploited, platform companies have started to look for new growth markets in the data economy. One such market is the one for data storage and processing in the form of cloud-computing services. Cloud computing is the on-demand availability of computer system resources for data storage and processing over the Internet, without direct active management by the user. This allows organizations or individuals to outsource their data management to specialized cloud providers. It means the data is centrally stored or processed in the cloud provider's data center rather than in an on-site center hosted by the organization itself. The cloud providers are also responsible for the management and maintenance of the data center, allowing their customers to focus on their core business processes. This is a departure from the decentralized data storage between 1980 and 2010, a period in which data was mainly stored on client servers. While a 2019 study shows that about 65% of workloads will continue to be hosted in private data centers and managed by internal infrastructure teams over the next few years,[57] the growth trajectory of public clouds continues to be strong. The global cloud computing market is expected to grow by 19.1 percent between 2021 and 2028, in part because smaller companies are moving their work to the cloud as the Asia-Pacific region expands its digital initiatives.[58] Amazon was the first company to transform the data centers it needed for its own operations from an operating expense into a revenue stream by first perfecting its data center capabilities and then marketing them to third parties. Other platform companies followed suit and have become the largest providers of public clouds today. As of 2021, Amazon's AWS controls a share of 32 percent of the global cloud market, Microsoft's Azure 20 percent, Google's Cloud 9 percent and Alibaba's Cloud 6 percent.[59] European cloud providers on the other hand have not been able to secure a significant share in this growth market so far.

While cloud computing services will continue to shape how data is shared and processed, the next wave of innovation is already underway. The emergence of edge computing heralds another phase of decentralized data storage and processing. Instead of computing and storing data in a central cloud, edge computing refers to data processing and storage close to the data source, e.g., on the end device such as a mobile phone. Edge computing is especially becoming critical for latency sensitive applications, such as autonomous vehicles. This means for companies and organizations that they have to increasingly manage their data across multiple locations and architectures. IDC and Seagate estimate that by 2022, the largest share of the global datasphere will be managed in internal data centers (570 TB), followed by public, private or industry clouds (498TB), third party data centers (407TB), edge and remote locations (390TB) and other locations (160TB).[60] See chart 7. However, despite the growth and diversification of data storage and processing capabilities, the growth of the datasphere worldwide is outpacing the growth of storage capacities. Of the 175ZB of data generated in 2025, only 17ZB will be stored[61] — a paradox, when considering that data is the economic input factor in the data and digital economy. Bringing down the costs of data storage and complexity of managing data across many different architectures are the key for resolving this paradox. The established platform companies will work on solutions to fill these gaps in the market.

The data processing and storage or cloud market is not the only growth market for platform companies. Integrated data-fueled super platforms are the other. If the

Chart 7:
Where is the global data sphere stored



established platform companies Apple, Amazon, Google and Facebook were to continue their growth journeys, they would jointly add USD 1 trillion to their revenue between 2021 and 2026.[62] Thus, a second growth strategy is the horizontal integration of other industries and data entities to support the development of applications that address new sets of use cases. In other words, existing platforms are adding features that are transforming entire industries outside their initial core businesses. This trend is already emerging in China and, to a lesser degree, in the US. Alibaba and its financial services provider Ant Group are using cross-industry customer data from e-commerce and logistics to offer new financial services.[63] Baidu, on the other hand, is venturing into autonomous driving. The company's open platform Apollo will draw on Baidu's compute power and AI-capabilities to integrate autonomous driving, entertainment offerings, advertisement and, in the future, ride-hailing service. Apple taps into new sources of data, such as wearables, and increasingly penetrates new industries, such as healthcare and entertainment. But it is not only digital companies that are leveraging the datasphere through super platforms. John Deere, for example, founded in 1837 and the world's largest farm machinery manufacturer in 2020, is complementing its business model around the production of machinery with a super-platform for different segments in agriculture. Since 2012, the company has been developing software products that networked John Deere machines with other machines, owners, operators, dealers, and agricultural consultants.[64] This software analyzes data from satellites, weather stations, and machines with integrated sensors, allowing farmers to make sense of an increasingly large and complex datasphere. But the company has not only incorporated new technologies into its products to increase yield per unit of farm, it has also repositioned itself in the ecosystem of customers and suppliers. In 2013, the company opened its cloud-based platform "MyJohnDeere" to third-party parties, including suppliers, agriculture-tech companies, and others, to facilitate the sharing of data and services. The SeedStar Mobile app, for example, captures row-by-row planting data that can help optimize planter settings, diagnose potential problems and field inspections.[65] The potential of these apps increases when farmers use John Deere's platform to share their data not only with other farmers but with the entire agricultural ecosystem — from the various suppliers and manufacturers to adjacent experts such as scientists around the world.

As digital transformation initiatives roll out across industries and data and AI capabilities expand, more companies, in particular in more traditional industries, will adopt platform business models within their industries and integrate across industries. This trend from product-centric to platform-centric companies, as a strategy of building businesses in the digital age is not new, but it will continue to impact the way data is being shared: Data provided by an ecosystem of users is aggregated, stored and processed by centralized platforms using cloud systems so that other users of the platform can create value from that data. This makes platform businesses progressively the gatekeepers to key data pools. But there are also opportunities for new players and innovations, e.g., due to diminishing trust in platform companies and the fragmentation of the data storage landscape, as we explore in the next chapter 2.4.

2.4 Web 3.0: The rise of data ownership and decentralized data sharing

For many, the rise of platform companies held the promise of democratizing the digital economy, by taking away access barriers to digital goods and services. However, the same companies that made data, compute power and knowledge accessible also consolidated with their dominance the very centralization of these goods. As outlined in the previous section, compute power is already in the hands of just a few players. The same applies for code collaboration platforms. Github was acquired by Microsoft in 2018, Kaggle by Google in 2017 and Wise.io by GE in 2016. While knowledge platforms such as Udacity and Coursera are not yet in the control of major platforms, some observers expect that education is one of their next target markets for horizontal integration.[66] The process of centralization of data and hence the control over it is eroding trust in the digital economy, limits competition and led to public outcry and legal action. This is because platforms that collect and hold data tend to overuse – and even abuse – the data they have.[67] Moreover, the gatekeeper role of platforms translates into an economically suboptimal allocation of data, making it difficult for smaller actors to participate in digital value creation. Coupled with the lack of approaches to scaling data sharing between organizations, as described in the "Breaking up data silos" chapter 2.2, private and public value growth remains constrained. Simultaneously, a new technological development allows for a decentralized data-sharing infrastructure: distributed ledgers at large, and blockchain in particular.

Blockchain is a distributed ledger that is open to everyone and thus contrasts with organized systems that are controlled by a central authority. The key value proposition is twofold: Once data is recorded in a blockchain, it is almost immutable. And the distributed nature makes it nearly impossible, even for a "legitimate" authority, to exert unilateral control over the transactions conducted on the chain. Bitcoin, as the first major blockchain application, still does not have much use except for serving as a currency, speculative object, or store of value. Other blockchain innovations offer more potential applications. The Ethereum blockchain, for example, is a kind of do-it-yourself platform or operating system for writing decentralized programs built on smart contracts. This and other blockchain platforms have driven innovation and have the potential to impact a wide array of industries providing the backdrop for what is commonly referred to as Web 3.0. Opponents are quick to point out that also the blockchain economy is only decentralized in theory but not in practice, as it is shaped by actors exerting significant power, such as the mining companies or crypto exchanges, centralized clearinghouses that facilitate the trade of listed tokens.[68]

Although some of the critiques have merit, the focus on comparisons between crypto's boom-and-bust cycles and previous bubbles too often overlook a crucial point. In the shadow of bitcoin and ethereum, a vibrant blockchain ecosystem is emerging, producing a growing number of infrastructural innovations, use cases and applications, increasingly connecting with the traditional economy. These include also use cases for the decentralization of data storage and processing, that have the potential to disrupt the businesses of centralized cloud infrastructures, and the development of new business models. Some of the many use cases and applications will undoubtedly fail, but much of which will last — amongst them Decentralized Finance (DeFi) and associated innovations such as Non-Fungible Token (NFT) or the rise of Decentralized Autonomous Organizations (DAO). DeFi aims to eliminate all points of central control or authority for virtually all types of financial services. Combined with decentralized money, such as certain cryptocurrencies, developers and businesses can set up exchanges or offer loans, insurance, and other similar services without any central authority overseeing or controlling it.[69] NFT, on the other hand, could provide the grounds for new kinds of social media platforms, music services or marketing, as it secures ownership over unique digital content, e.g. a piece of art of a person's digital identity. Instead of ceding the rights over this data and its commercialisation to platforms, as in the case of Facebook, Twitter or Spotify, NFTs provide the basis for market structures in which it can be traded peer to peer. Governing this emerging decentralized data economy are a wide variety of DAO. DAOs are web-based organizations, some of them even without a legal registration in any jurisdiction. Hence, they are often also referred to as data- or internet-native organizations. Their rules are encoded in the Blockchain and controlled solely by the members that are invested in them via so-called governance tokens. Already today, DAOs deploy venture funding or donations for business, crypto or social purpose projects on behalf of their constituencies and manage art, music or projects. Within the crypto community, it is widely believed that DAO will become major political and economic actors for governing digital and non-digital issues, as well as allocating resources (also in the traditional economy). However, considering that almost all decentralized exchanges have been affected by hacks in recent months, DeFi is still three to five years behind the resilience developments of the larger crypto ecosystem. Corresponding time will therefore be needed before DeFi will reach the broader economy, if policy developments do not stop it. However, the underlying technological innovations in the distributed leger infrastructure are bearing fruit.

We are already seeing a number of early distributed ledger and blockchain-based approaches for data sharing, such IOTA or the Ocean Protocol. IOTA for example is a cryptographic approach similar to the blockchain to facilitate data exchange between machines. It does not require miners, as we know them from other blockchains. Its features also compensate for other shortcomings of blockchains: the speed of transaction time and scalability essentially enables microtransactions, which is a critical use case when considering the low value of some machine-to-machine data transactions, which necessitates scale economies. IOTA currently focuses on facilitating data sharing in application areas such as digital identity management, mobility, smart cities and global commerce. And it is gaining in popularity: IOTA reached a market cap of USD 3.4 billion in October 2021, up from USD 1 billion shortly after its inception in June 2017. Another example is Ocean Protocol (**www.oceanprotocol.com**), a blockchain-based ecosystem designed to allow organizations and individuals to exchange data and monetize it. The data exchanges take place in the Ocean Protocol marketplaces. Every data service is represented through a unique token which holds a data set or data service. Interestingly, it allows for working on data that remains under the control of the data provider, thereby securing ownership and privacy. Like IOTA, the Ocean Protocol has seen a spike in market capitalization that has now reached USD 455 million in October 2021, up from USD 13.7 million in January 2020.

It is just a matter of time until these innovations will converge with decentralized marketplace approaches, allowing for truly decentralized data exchanges. The smart combination of the traditional economy and the technologies of the decentralized distributed ledger economy presents opportunities especially for fragmented markets like the EU. As there are no significant platforms but a multitude of smaller actors and a cultural aversion to sharing personal data, a platform-driven healthcare system for example that is emerging in the US, is unlikely to develop in the EU. Privacy-assured and decentralized data marketplaces could be the solution in this situation — avoiding a paralysis in the health system based on the fragmentation and high data protection standards, and rather strengthening the system through the secure sharing of data with doctors or the commercialized trading of anonymized data access with researchers or healthcare innovators.

2.5 Data trading: A necessity in the evolution of data sharing?

While legal uncertainty or even barriers to the flow of data between countries and organizations persist, organizational and technical platforms favor the tradability of data. This is accompanied by the recognition of data as an asset class or factor of production. For companies, data is likely to become a corporate asset presented on corporate balance sheets in the future. While the concepts for data trading and monetization are not new, they are receiving increasing attention thanks to recent governance and technological developments. Analytics firm Gartner projects that it will take 3 to 6 years to bring “data productization” to market beginning in 2021.[70] Data productization essentially means treating data like a product, forcing the definition of “product” requirements, a step toward trading data as an economic good. This section describes the characteristics of data as an economic good and assesses the trends that indicate or hinder the trading of data.

Data is not the “new oil” as suggested in popular media, because it is a non-rivalrous asset. This means that an algorithm or a company can use the same data set multiple times without it losing its value or quality. This is different from, for example, a barrel of oil, which can only be burned once. Data is also a non-excludable asset. Although there might be proprietary data, in theory data can be easily shared and made available to anyone with a computer. In fact, data is a changing representation of values depending on the situation, the combination with other data types, and the use case. Data can also be fungible *and* non-fungible. This means, some data is replaceable or interchangeable, e.g., like the digital currency bitcoin, and other data is non-fungible, like a digital identity. This makes the trading of data difficult, because securing ownership and defining a price for it is a highly complex undertaking requiring sophisticated talent and algorithms. The world, however, has both and the authors of this report believe that this is a promising frontier for both economic and social prosperity.

Traditional ownership rights depend on the exclusivity of the goods to be protected. The European Union has set the tone for defining and clarifying the ownership of personal data with the General Data Protection Regulation (GDPR), however with apparent shortcomings. While protecting the ownership of data on paper, effective privacy management is essentially impossible due to tedious, complicated, and piecemeal “user privacy settings” in operating systems, browsers, apps, and other websites and services. Currently, users who want to understand and effectively manage their privacy settings must work through up to 900 pages spending up to 34 hours to read all the terms and conditions of popular apps on an average user’s phone, which frequently change and get updated.[71] Understanding the legal text is difficult, if not impossible, for most internet users, with the complexity of many privacy policies exceeding even collegiate standards of reading comprehension.[72] This explains the privacy paradox, a dichotomy between a person’s intention to protect their online privacy and their actual online behavior. So, the forthcoming revision of the GDPR might embrace a more user centric approach to privacy management. Despite their shortcomings, the GDPR provides basic data protections, but ways must now be found to implement them in a user-centered way. However, the same cannot be said for anonymized or non-personal data.

In its efforts to create an EU data market, the legal community considered the introduction of intellectual property rights as a means to establish “the right of data producers to non-personalized or anonymized data.” This was intended to achieve the sharing of such data as well as to create incentives to protect investments and assets. This idea, however, hasn’t found its entry into the latest EU Data Strategy, which also makes it an unlikely subject of the forthcoming EU Data Act, due to twofold reasons: First, because data producers can claim de facto exclusivity due to the way the data is stored, and second, because existing database law (the content of a database) and copyright law (the structure of a database)[73] in combination with case law provides sufficient legal instruments to protect data assets.[74] This shifts the onus from a legal discussion to a technological discussion around the question of how to package data for trading and portability between different applications and platforms. Imagine being able to port your content and followers from Twitter to another social media platform where you get better features. Or picture how you, as a musician, can share your music directly with your fans. This is what property rights of data in the data economy could mean. Blockchain tokens can make these ideas a reality. Blockchain tokens make it possible to package data and digital content and ensure ownership and transferability through the blockchain infrastructure. This moves the economic power from centralized platforms to the edges, the users. However, in order to still get AI-supported services, which is currently only possible because we transfer our data to central platforms where it's used to train AI algorithms, we need further technological innovations. Innovations that bring AI to the data instead of the data to the AI, as in the current model. Federated learning could be the solution.

Federated learning is an AI technique that allows data to be processed in a space controlled by the user. First introduced in 2017 by Google,[75] federated learning heralds a fundamental paradigm shift in how data is processed. Instead of bringing the data to the algorithms, as is the reason for centralizing data pools in the current platform economy, the algorithms are brought to the data. While most AI systems require a central data set for training, federated learning models train an algorithm across multiple edge devices, each of which retains its data and does not share it. In 2019, Apple, for example, changed the underlying architecture of its voice assistant Siri to federated learning, which deployed the actual machine learning inference right on the phone, rather than in the cloud, thus proving the concept of data security-by-design. Federated learning therefore promises to increase data privacy and reduce data leakage. Put simply, data that never makes it to the cloud but remains with its owner cannot be appropriated or stolen in hacks that typically target huge aggregations of centrally stored data.[76] However, while the data remains on the end devices, the Federated Learning model still operates from a central server, which leads to increased questioning of the security of federated learning, because malicious clients or central servers can still attack the global model. For this reason, proponents are calling for a decentralized federated learning framework based on blockchain, in which smart contracts built on the decentralized network of blockchains replaces the central server used by the Federated Learning Model.[77] The first use cases that exemplify this convergence of federated learning AI and blockchain already exist. Raven Protocol (**www.ravenprotocol.com**), a blockchain-based, decentralized and distributed deep-learning training protocol, adds this functionality to the market created by the Ocean Protocol discussed in the previous section.[78] Other use case areas for this convergence of Blockchain, AI and data science, in addition IoT and cloud computing, are likely to come out of the area of healthcare and autonomous vehicles in the near future.[79]

However, as with other areas in the economy, few things move without incentives, which require valuation and pricing. Pricing of data is another difficult issue. At present, data brokers trade general information about a person – such as age, gender and GPS location – for a mere USD 0.0005 per person (or 50 cents per 1,000 people).[80] Yet, when asked about the price they put on their "loss of privacy," internet users ascribe as much as USD 36 to their personal identifiable data.[81] The reason for this difference is that the current digital economy is based on the model of trading data for a presumably free service. But if data now becomes an economic input factor, a transparent marketplace is needed in which data producers can negotiate value.[82] But how do you price something that can be multiplied infinitely without losing its value? Traditionally, pricing depends on the availability or in other words on the scarcity of a commodity and its features. The prerequisite for setting prices for data sets is therefore the possibility of making them scarce or restricting their use by claiming property rights. However, the latter is not currently possible and does not provide any leverage to renegotiate the data value in today's digital economy controlled by platforms, as discussed above. Fortunately, innovators are clustering around this issue: Jaron Lanier, a computer philosopher and "OCTOPUS" (Office of the CTO Prolific Ubiquitous Scientist) at Microsoft, was one of the first to advocate for a union-like organization called Mediators of Individual Data, or MID, which negotiates the value of data on behalf of its members in order to commoditize data and reorder the balance of power among those who produce and use data.[83]

The Ocean Protocol (see chapter 2.4 above), on the other hand, allows for the data service to be purchased, rather than the data itself. Thanks to Federated Learning and Raven Protocol, it is technically possible to allow the algorithm to work with the data for a specific purpose only and then deny access to the data again. But even if the problem of data scarcity is solved, there are other problems in the pricing of data. This has to do with their different features. For example, a person's health record may have value in itself, but not a person's mobility data footprint, which in turn only has value in aggregate with the mobility footprint of many others. And there are multiple other features, which determine the price of datasets. But just as AirBnb succeeds in recommending prices for flats and houses with almost infinitely different characteristics, a clearing house of a data marketplace with a similar pricing engine — with an auction model during the market entry phase — should also be able to determine the price of different data sets in varying configurations.[84]However, this will only work if the digital economy gets the impetus to change the paradigm of how data is currently valued and traded. Only recently, institutions such as the IEEE, the Japanese Government and the World Economic Forum have promoted ways to trade and monetize data. While the results are not yet accessible, IEEE, the world's largest professional engineering association and known for developing standards for the global digital economy, started a data trading initiative in 2020.[85] However, some of the most recent and trailblazing initiatives are coming from Japan. For example, the Japanese government has established guidelines for "personal data trust banks," with the objective to establish organizations that store data from customers currently held by companies and public entities. If an individual consents to the data being shared, the bank would provide the information to businesses in exchange for a fee.[86] The Federation of Japanese IT Associations manages the certification system for these data banks, called information banks. The first certified databases were announced in July 2019. At the G7 Summit 2019, the Government of Japan put the vague concept of "data free flow with trust" (DFFT) on the global agenda.[87] More recently, the World Economic Forum's Center for the Industrial Revolution in Japan initiated the Data Common Purpose Initiative (DCPI)[88] which calls for data marketplaces and exchanges as a means for incentivizing data sharing. In 2021, the Center published a governance framework for such marketplaces.[89] Given the recent US-Japan Digital Trade Agreement and the likely expansion of the Digital Economy Partnership Agreement (DEPA), which currently counts Singapore, New Zealand, Chile, and South Korea as members and may add Japan soon, Japan's push for data monetization is expected to take a multilateral shape.[90]

Chapter 3:

Changing Imperatives in the Data Economy

3.1 Introduction

“The world is one big data problem”, is a famous quote by Andrew McAfee, co-director of the MIT Initiative on the Digital Economy. The quote reflects the common belief that more data means better insights, foresight, or cognition and by extension also, cognitive innovation — the last step in the data value creation chain. Until now, all commercially successful innovations in the digital economy have been geared towards collecting as much data as possible, whether via search engines, e-commerce or social networks. The secret of success used to be: performance grows as data expands. The most complex and powerful algorithms usually are more demanding in terms of data. Those powerful algorithms create powerful AI. AI leads to better products, increased productivity, and superior customer experiences, in turn leading to more customers who share more data, producing even smarter AI. But the era of amassing data to what is referred to as “big data” will not last forever.

First, because data scarcity remains prevalent in critical areas, e.g., medicine and drug discovery, despite the flood of data elsewhere, e.g., mobility or finance. Second, advances in AI that make it possible to work with ever smaller datasets, responding to industry needs for cost-effective learning and academic ambitions in the field of strong or general AI. Third, big data and algorithm training are highly energy consuming and climate change mitigation policies will bring more attention and restrictions to this fact. This paradigm shift towards data efficient algorithms accelerated during the COVID-19 pandemic and promises many benefits for companies and the digital economy at large. Coupled with the spread of low-code and no-code applications, which makes the development of AI applications possible even for non-experts, a new wave of data and code efficient AI solutions can be expected. The shift from big to small data and the democratization of AI has consequences: When data volumes are no longer the decisive value driver, entire business models change, and completely new ones emerge. This chapter will analyze these trends and assess what it means for businesses and the data economy at large.

3.2 The dwindling importance of big data

Due to innovations in machine learning approaches, data generating and processing methods as well as in data use, we currently witness a paradigm shift from big data to small data providing huge opportunities for organizations to create value in the top layer of the datasphere. In general, the term “small data” refers to approaches that require less data but still provide useful insights. In 2021, analytics firm Gartner predicted that by 2025, 70 percent of all enterprises will shift their focus from big data to small data, making artificial intelligence less data hungry.[91] This paradigm shift is made possible by three converging trends: First, thanks to automated data cleansing, companies and organizations can increasingly work with cleaner and therefore smaller data sets. Second, there is a rise of new machine learning approaches such as Few Shot Learning and Self-Supervised Learning that work with small data sets. Third, there is an increase in synthetically generated data, which is used for areas where there is not enough real-world data. This chapter will analyze the trends towards small data.

“Garbage in, garbage out” is a common quote within the AI community which means that the quality of the output of an AI algorithm is determined by the quality of the input, hence the data. Data quality is a key imperative in the data economy. According to IBM estimates, USD 3.1 trillion was the yearly cost of poor-quality data, in the U.S. alone, in 2016.[92] Generally, data quality is defined and described by dimensions such as accuracy, consistency, timeliness, uniqueness and validity. As a rule of thumb, the higher the quality, the smaller a data set can be from which the AI can draw conclusions. But good data quality is not a given. Therefore, data and data sets usually have to be prepared in an elaborate way so that they can be evaluated and processed. These activities are often underestimated. On average, companies and data analysts spend 60 to 70 percent of their working time sorting, cleaning and preparing data.[93] This problem is pervasive. A study from 2019 shows that 96 percent of enterprises encounter data quality and labeling challenges in machine learning (ML) projects.[94] For this reason, automatic data cleaning is a subject of research and an attractive field for innovators. Small and big companies alike are now offering solutions for automated data cleaning.

Zoom in: Automating the tedious task of data cleaning

Making automated data cleaning possible is an approach called Augmented Data Science. Augmented Data Science automates the identification of identifying errors in a data set and suggestions for fixing data quality issues, based on the metadata of datasets.[95] Metadata refers to logs, query history, usage statistics etc. of data. This means that active metadata platforms are emerging that are continually collecting metadata at every stage of the modern data stack. Augmented Data Science is currently still mainly in the experimental and development stage, but its results are already very promising. It is likely to become mainstream in enterprises in the next two to five years.

With automated data cleaning, companies can radically shorten time-consuming but non-productive work. This saves costs and opens new potential and capacities. When highly qualified data scientists no longer have to spend two-thirds of their working time on data cleaning, the productivity of an entire company increases significantly[96] In addition, automated data cleaning also makes data stocks manageable that were previously not used due to the amount of work involved, for example vast amounts of historical data that lie in databases and archives of companies and government institutions.

Moreover, the trend from big data to small data is also driven by new machine learning approaches, such as Few-shot Learning (FSL). Few-shot Learning is about making predictions based on a limited number of samples.[97] For example, in the field of image recognition, classic ML systems need several million images as examples to learn. Few Shot Learning, on the other hand, can get by with just a few hundred or a few thousand examples and achieve high performance and reliability. This is necessary in cases for which it is simply impossible to collect sufficient data. Research into rare diseases best highlights the need for small data approaches, especially in the light of only a few functioning data sharing and pooling mechanisms.[98] But also in areas such as language processing, autonomous driving, industrial robots or drug discovery, FSL approaches offer new possibilities, while reducing costs associated with collecting and storing data for companies and organizations.[99]

While small data concepts offer a solution for using AI even with small or scattered datasets, they still require real world data. In particular, small organizations and companies might still face challenges in compiling datasets large enough to train AI systems. Legacy infrastructures, siloed data systems, strict regulation on personal data or security concerns may cause data unavailability. One approach for dealing with these challenges is so-called “synthetic data.” This approach has only arisen in the midst of the 2010 decade but developed quickly because of its technological simplicity and will certainly become mainstream in the next two to three years. Synthetic data is created with the help of AI rather than collected from or measured in the real world. It is therefore artificial or „synthetic,“ but it reflects real-world data mathematically or statistically.[100] Take the example of training AI for autonomous driving. Instead of recording millions of hours of video data from real roads, an AI can “extrapolate” an unlimited amount of synthetic road data from a small video data set, which in turn can be used to train the driving AI. The cost advantages of such an approach are enormous. Other examples for synthetic data can be different variations of movements of humans in a room or an entire city; variations of how people touch an analogue product or a device and use its services. The possibilities of variations and their connections are practically unlimited. Another advantage here is, that there is no need for “surveillance” of people and their behavior to collect enough data for ML/AI Training. For machine learning, however, this synthetic data is always “new” data and examples to learn from.[101] Sometimes, synthetic data may even be better than real world data.[102] Hence, synthetic data delivers multiple advantages, e.g. when privacy requirements limit data availability or limit the use of data, when data is needed for testing new products, but that data does not exist or is not accessible or when training data for machine learning is needed, but too expensive to generate in real life.[103] This presents companies with huge cost-saving advantages, e.g. when testing novel methods and processes in industrial production. Instead of lengthy series of trials and tests in the real world, the methods are simulated and tested virtually in all possible environments and circumstances. This significantly shortens innovation cycles, reduces cost and time to market, especially in industries with restrictive regulations, such as the financial sector or medicine.

Despite the advantages, there are definitely challenges with synthetic data: The synthetic data is only as good as the quality of the original data, which must still be annotated by humans. Flaws in the original data are then copied into synthetic data. Another challenge comes with high-quality synthetic data in the form of so-called "deep fakes". Contrary to what the term suggests, deep fakes look extremely real and convincing for the audience nowadays. Synthetically created audio files have tricked a CFO in making wire transfers to wrong recipients, synthetically created video clips have spread fake messages of real political figures.[104] A number of companies are developing deep fake detectors, which use AI to spot AI edits of videos of famous persons, by tracking small facial movements unique to each individual. These markers are known as "soft biometrics" and are too subtle for the AI to currently mimic.[105] The result is an arms race between the AI that gets better at generating artificial data and the AI that keeps pace with the task of detecting the fakes. This arms race in AI capabilities of creating and detecting deep fakes is reaching the next level with an innovation called "Third Generation Generative Pre-trained Transformer" (GPT-3). It is a machine learning model trained to use internet data to generate any type of text.[106] It requires a small amount of input text to generate large volumes of relevant and sophisticated machine-generated text. In November 2021, Microsoft announced to connect its Azure platform with GPT-3 to improve the speech processing of its enterprise platform. OpenAI, the developers of GPT-3, announced to publish the code of the model for the public, because they are sure to have developed and modified the model so well that it cannot be misused. Open AI has developed GPT-3 to, for example, convert natural language into programming language, to automatically create text, or to summarize large amounts of text.

While data preparation is being automated and data management is gaining strategic importance, the interpretation and integration of the results into companies' own operations, processes and business models are becoming a table-stake capability. It is not only the machines and technologies anymore that make the difference. It is rather a matter of the right division of tasks between human and machine as opposed to a traditional industrial logic in which people were interchangeable as operators of the machine, as long as the machines were running. In the dawning age of AI, people and machines will become more of a symbiotic team. Machines help to organize and evaluate the data, and in some cases automated AI is even able to make recommendations or suggestions for interpretation, as will be described later. This allows human workers to focus more on value creating, strategic or design tasks.

3.3 The impact of automated AI engines

Although data is becoming a key production factor for the digital economy downstream, most of the money is still made further midstream in the data value chain with data storage, data brokerage, etc, rather than upstream at the source (raw data). This might change with the emergence of data marketplaces and trading mechanisms expected to evolve in the 2020s. While downstream applications based on data reduce cost or create new revenue streams over time, they first generate higher costs. Hence, the cost-benefit analysis of AI projects is not always clear at first. In fact, 76 % of companies in the US barely break even on their AI investments, a 2020 survey showed.[107] This is because AI modeling requires expensive data science and AI design talent,[108] data lakes and pools need to get structured and cleaned, AI models need to get custom-designed and trained for each use case before they can go into operation. Moreover, data scientists and AI specialists are rare in the labor-market, but are increasingly in demand, making their services expensive. AI models need to be trained before they are ready for use. This can take time and requires appropriate know-how. Both cost money. Furthermore, large amounts of data are needed for training (to date). These must either be purchased or collected. In addition, there is the effort already described to make the data usable (cleaning). But most importantly, AI solutions so far tend to have been tailored to the respective company. They are therefore unique and scale economies are hard to achieve.

However, due to recent innovations, the cost-benefit ratio of AI projects will change. During the last decade, there has been a significant rise in investments flowing into Automated Machine Learning (AutoML) as well as low-code and no-code applications. These innovations promise to bring down the costs of AI applications by making large AI teams obsolete. AutoML accelerates and simplifies machine learning, because individual steps of the learning process are automated, and no human experts are required.

Innovative AutoML solutions allow automating the process of architecture selection and neural net design, upending the potential skill requirements for low intensity AI application deployment. This may reduce the cost for AI deployment across organizations, as many best-in-class open-source algorithms already exist. It also reduces the skill requirement or talent barrier for firms looking to add predictive capabilities to their organization. AutoML can provide all steps in a development and implementation process of AI models from preparation of data, selection of models and algorithms to the deployment of the model integrated in an application.[109] AutoML thus reduces the time and labor required to use machine learning, as humans only need to understand and work out the bridge from the business to the technical problem. This is followed by testing, deploying and monitoring the ML model in the productive environment. The challenge for companies in the future will therefore mainly be to define the tasks that AutoML will take on.

Just as AutoML radically simplifies the training of AI models, no-code and low-code solutions will simplify the building of concrete applications and products built on top of AutoML applications. Low- or no-code means that applications can be created in a modular system without requiring programming knowledge.[110] There is an increasing number of no-code platforms through which programmers are provided with a graphical user interface with drag-and-drop functions and easy-to-understand building blocks for important steps. The advantage: No-code development environments make it possible, especially for employees who have no connection to technology but have in-depth technical know-how, to create their own applications or program sequences without having to struggle with programming languages. For example, Zapier and automate.io allow the automation of processes via no-code solutions. Obviously.ai and Mixpanel offer analytics without having to write their own code. PrimerAI is offering Natural Language Processing Services on a low-code platform. These applications are ready for immediate use and can be combined like building blocks to form a finished application without having to constantly redevelop individual components. This results in time, resource, security, and cost advantages. In the last two to three years, not only have many no-code/low-code startups emerged in the US and Europe, but big platform companies such as Google and Microsoft are also entering the market. Gartner estimates that low-code tools will make up 65 percent of application development activity by 2024 and that most SME will have adopted no-code tools by then.[111]

In the next three to five years, these approaches will become established because they will give small and medium-sized enterprises in particular the opportunity to build and deploy concrete AI-based applications themselves for the first time at low cost and with manageable effort. However, while no-code and low-code applications are likely to spread across all industries and use cases and democratize the usability of AI, they do not spell the end of AI teams in companies. Quite the contrary. This is because purchasable AI solutions, such as no-code or low-code applications, no longer represent an advantage over competitors who have access to the same solutions. It is much more a question of companies having to weigh carefully for which use cases they buy turn-key solutions and for which they develop their own solutions in order to create a competitive advantage. Boston Consulting Group (BCG), a global consulting firm, provides a helpful framework. Here, “commodities” are AI solutions with a low degree of value potential and differentiated data access. Solutions and options with a high degree of value potential but a low degree of differentiated data access are “danger zones”. Data access and high-quality reliable data is crucial for success by differentiation. Options with a high degree of differentiated data access and low degree of value potential are defined as “hidden opportunities”. Companies can generate quick wins and insights to build up knowledge and experience with AI and data analytics. So-called “gold mines” in turn are characterized by a high degree of value potential and data access. These AI solutions should be developed by the company itself, while in all other cases no-code or AutoML solutions should be used. Only if the added value is high and there is access to high-quality data will it still be worthwhile to invest in in-house AI developments in the future.

Hence, it will be crucial to adjust the composition of AI teams. The typical pyramid structure of AI teams, with ML specialists at the top, software developers in large numbers at the bottom, and data-scientists and -engineers in between,[112] will likely change. Product strategists who know the landscape of turnkey, low-code and no-code solutions and who can assess how differentiated access to specific data can be ensured to establish a competitive advantage in proprietary AI solutions will become more important. In areas where turnkey and low-code or no-code solutions can be used, software-savvy developers will probably suffice rather than cutting-edge data scientists and ML engineers.

Chapter 4:

Recommendations

The previous chapters laid out the broad lines that are shaping or will shape the data sphere and, consequently, the creation of value from it. The implications for business executives and policy makers on both sides of the Atlantic and beyond can be summarized in a Strategy and Policy Exploration Agenda as follows:

Recommendation 1 – Foster cross-border data flows between the EU and the US: by renegotiating the Privacy Shield to promote the core of a global data marketplace: The unhindered cross-border flow of personal and non-personal data is an important growth driver for the digital economy. To this end, a common data space should be created between the EU and the US and later expanded to other regions of the world. A key prerequisite for this is the renegotiation of a new US-EU data protection shield (Privacy Shield Framework) with the aim of reducing the legal costs of transatlantic data transfers while ensuring data protection. This provides the basis for the following two recommendations.

Recommendation 2 – Enforce data protection by creating a supranational data protection agency: Within the EU, efforts to harmonize data regulation with the US should be complemented by the establishment of a pan-European oversight mechanism. Although the EU has a common legal framework (GDPR), it is enforced at national levels, adding complexity for companies that operate in different EU countries. The institutional foundations for a pan-European regulator are already in place, with the G29 network (network of data protection authorities) and the European Data Protection Supervisor (EDPS) already on the way. Introducing such policies and regulators will foster cross-border data flows and further spur growth and competition in the digital economy on both sides of the Atlantic. Moreover, European and US policy makers and regulators should harmonize or at least render interoperable the administrative procedures and mechanisms that digital innovation entrepreneurs have to navigate to start and scale projects and ventures. This should include regulatory experimentation sandboxes for data-driven business models that enable projects rather than prohibiting them and are safeguarded by legal and ethics experts. This requires a positivist mindset, rather than a protect-and-prohibit one.

Recommendation 3 – Create a Transatlantic Free Data Trade community (FDaT) governed by a Multilateral Data Agency (MDA): The combined data market of the US, Canada and the EU encompasses around 733M internet users. This presents a vast set of opportunities for insight generation and solution design in critical areas, from decarbonization to healthcare, education, transportation and migration management. While the three partners follow different approaches to business creation, economic growth and the appropriate governance and regulatory mechanisms, they are the most closely aligned in the global economy in terms of their basic values. In parallel, the transatlantic partners should not just facilitate integration among them but integrate with like-minded partners in Asia. Both from a value and an economic perspective, this is desirable, as Asia is emerging as the most dynamic region for new data generation and data business models. As a first step, an opportunity awaits for the transatlantic partners to make data transfer protocols interoperable and harmonize data market regulations, including privacy and other ethics safeguards. This would not only include a renegotiation of the Privacy Shield, but also the design of a multilateral data agency (MDA). Such an agency would review and certify data trading parties and arrangements, oversee and mediate issues of data property rights and privacy. The MDA had the right to audit such arrangements to verify compliance and to make recommendations to all parties on how to increase the efficacy of the trading flows. In parallel, we recommend that trans-Atlantic policy makers begin to negotiate a Free Data Trade community (FDaT) and its expansion to additional countries that have advanced digital programs and democratic political systems, such as Israel, Australia, Japan and India, as well as Global South democracies, such as Senegal, Ghana or Nigeria. That expansion should also include a data federation with the evolving Digital Economy Partnership Agreement (DEPA), which currently counts Singapore, New Zealand, Chile, and South Korea. In a third staggered but overlapping negotiation, the FDaT-DEPA data federation could include China as well, both a key data generator and digital innovator. Negotiations to that end could start immediately after the formation of the core FDaT group. It is important to engage in these processes quickly as the careful navigation on the political use of data, the rights of individuals, and the compatibility of regulations and laws will be difficult and hence time-consuming. Full data market integration with China may be an elusive goal and may not reach beyond data market interoperability rules. But the goal must be to avoid a fragmentation of the data sphere into parallel blocks and data universes.

Recommendation 4 – Establish human agency over personal data and flip data market power by smart regulation of platforms: Rebalancing economic power in today’s digital economy is essential to restore trust and improve data sharing. However, privacy exists only on paper, not in implementation. Privacy declarations and user agreements of digital services are too technical and complex and change too frequently for users to be able to manage and protect their privacy effectively. Instead of expanding the regulatory framework, the focus should be on procedures and processes to incentivise data actors to ensure frictionless data agency. Efforts could take the form of incentives for more user-friendly legal language and support for technical solutions that centralize privacy management in user-centric privacy charters. Instead of leaving it to the digital app and service providers to dictate their terms for privacy settings, individual users should be empowered to define a central individual charter per user to which all digital services must then adhere. Legal frameworks allowing for such an approach could give users back the agency over their data. It is in this context that decentralized economic models created by blockchain innovations offer opportunities for platform regulation that is smarter than traditional antitrust approaches. Instead of splitting platforms by market power, smart regulation should enable the transferability of digital identities and digital content between platforms, e.g., by using crypto tokens. This in turn would reposition platforms as infrastructure providers instead of data barons while giving back users control over their data. This will increase competition amongst platforms, promote innovation and support an optimal allocation of data and its value.

Recommendation 5 – Productize data and set up a data market reference design to govern monetization, valuation, and trading mechanisms: Data trade is key to fostering an integrated data market within the EU and between the US and the EU, as it would make sourcing data to train AI nothing more than a cost for innovators. Thus, the paradigm shift towards data as a tradable economic input or production factor promises benefits especially for start-ups and SMEs that currently have insufficient data to develop or train their own digital innovations. Achieving this paradigm shift requires a governance framework for emerging data marketplaces and data valuation that must be shaped in partnership by all actors along the data value chain (data creators, repositories, curators, brokers and users, as well as users of data insights). A governance framework should be accompanied by or include a reference design for a data trading market. This combination should entail institutional mechanisms and goals for data valuation and combinatory data set pricing, data footprint protection, privacy sanitizing, portable digital identity management, a review of digital content ownership and property rights and data usage. The work of the World Economic Forum or the Japanese government can serve as a basis for this. Such governance frameworks for data marketplaces can initially be industry-specific in design. In the EU, the healthcare sector is a good example, as it is particularly fragmented and unlike in the US, there are no platforms emerging as data aggregators.

Recommendation 6 – Explore the Web 3.0 and experiment with policy frameworks for Web 3.0 projects and stakeholders: The previous two recommendations should culminate in a larger policy framework for Web 3.0, which promises an evolution from the current platform-driven data sphere and digital economy to a creator-driven data economy. While greater decentralization is essential to regain trust in and foster the innovative power of the digital economy, the right balance between centralized and decentralized elements remains to be found. Web 3.0 is more than cryptocurrencies or decentralized finance and an exploration of this emerging space must take into account the versatility of Web 3.0 projects and actors. Besides crypto protocols, apps or app developers, Decentralized Autonomous Organizations (DAO) should be considered in particular. A policy framework for DAOs should allow them to enter into traditional contracts, enjoy the same benefits of corporations, but also to pay taxes. New incentives and policy approaches should also be considered for Web 3.0 platforms to incorporate the transparency and authentication advantages and issues they bring. For example, financial reports can be produced hourly rather than quarterly, and their reporting structure can be aligned with user needs instead of legacy legal requirements. But also, delicate issues need to be explored, such as how the strength of U.S. Dollar and Euro currencies remain secure in the Web 3.0 world — for example through decentralized Euro or U.S. Dollar stablecoin projects, in parallel to Central Bank Digital Currencies (CBDC).

Recommendation 7 – Lay the foundation for more equitable growth by adding a "T" for Technology to ESG frameworks. Small data approaches and less data hungry algorithms in AI reduce the energy footprint of the respective applications, increase resilience, and thus support an organization's practices in the context of the environment dimension of ESG frameworks. Compliance with protective data governance to ensure its responsible use falls under the governance dimension of these frameworks. However, there are multiple dimensions related to technology and data that do not fall into any of the environmental ("E"), social ("S") or governance ("G") buckets, such as cyber security, data mining, or data and AI geopolitics.[113] Adding a "T" (for technology and data) to the evolving spectrum of ESG initiatives designed by corporations, investors and policy makers is therefore an imperative for the 2020's. Only through a holistic approach on emerging data and technology dimensions, individuals can keep control over digital spaces, be secured and shape the form that new trends such as the metaverse will take. Only by reflecting the "T" for corporate behavior, the sustainability and equity of the next phase of the digital economy can be ensured.

Recommendation 8 – Build cognitive design and insights teams by creating talent pools in response to advances in AI and data science: Rather than top-notch AI experts, machine learning (ML) engineers and data scientists, it is hands-on cognitive design and insight teams that organizations need most in this next phase of the digital economy. New skill sets are needed with the advance of no-code, low-code and AutoML, which are shifting from programming and modeling to technical and design requirements to create data-driven and intelligent applications. Such cognitive and insight teams should include a wide range of data wranglers, AI engineers, and digitally savvy product developers. However, the creation of such a talent pool requires industry specific in-company AI training, vocational training programs, technology-oriented business schools and the inclusion of AI as a cross-cutting theme in university programs. These training-oriented offers should not replace but complement university education or postgraduate research in computer science and cognitive technologies.

Recommendation 9 – Thinking ahead about cognification not digital transformation. The emerging commoditization of AI and data science will converge with other breakthrough cognitive technologies that enable a much greater degree of automated sense making. We will understand more, unleash new opportunities and explore creativity in novel ways. This requires overhauling existing business or political science curricula in higher education to consider changes in technology strategy. The emergence of no-code, low-code and AutoML will shift the strategic considerations of companies and public institutions. Instead of creating a strategic differentiator through AI and data use, both will become table-stake capabilities. When to buy existing solutions and when to build new ones will increasingly become the decisive point. At the same time, the frontier and thus the first mover advantage is shifting from digital and data technologies to cognitive technologies, such as AI, quantum computing, VR/AR or brain computer interfaces (BCI). Companies and organizations should already explore the new possibilities and capabilities, which these technologies promise.

Endnotes

[1] Holst, A. (2021). Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025. https://www.statista.com/statistics/871513/worldwide-data-created/. 08.01.2022

[2] Business Wire (. (2021): Data Creation and Replication Will Grow at a Faster Rate Than Installed Storage Capacity, According to the IDC Global DataSphere and StorageSphere Forecasts. https://www.businesswire.com/news/home/20210324005175/en/Data-Creation-and-Replication-Will-Grow-at-a-Faster-Rate-Than-Installed-Storage-Capacity-According-to-the-IDC-Global-DataSphere-and-StorageSphere-Forecasts 12.02.2022

[3] https://www.statista.com/statistics/871513/worldwide-data-created/

[4] https://www.businesswire.com/news/home/20210324005175/en/Data-Creation-and-Replication-Will-Grow-at-a-Faster-Rate-Than-Installed-Storage-Capacity-According-to-the-IDC-Global-DataSphere-and-StorageSphere-Forecasts

[5] Liu, S. (2021). Nominal GDP driven by digitally transformed and other enterprises worldwide from 2018 to 2023. https://www.statista.com/statistics/1134766/nominal-gdp-driven-by-digitally-transformed-enterprises/. 08.01.2022

[6] Seagate. (2020). Rethink data. https://www.seagate.com/de/de/our-story/rethink-data/. 08.01.2022

[7] Internet Live Stats. (2021). Internet Live Stats. https://www.internetlivestats.com/. 09.01.2022

[8] Additional internet users contribute to both, personal and non-personal data creation.

[9] International Telecommunication Union. (2021). World Telecommunication/ICT Indicators Database 2021. https://www.itu.int/en/ITU-D/Statistics/Pages/publications/wtid.aspx. 09.01.2022

[10] All northern Atlantic countries are counted towards the Global North, as well as Japan and Australia. From a datasphere point of view, China also falls under Global North, because it scores similarly to European countries across many datasphere metrics.

11 Newzoo. (2021). Global Mobile Market Report. https://newzoo.com/key-numbers#mobile. 08.01.2022 Newzoo. (2021).

12 Reinsel, D., Gantz, J., & Rydning, J. (2018). The Digitization of the World From Edge to Core (pp 22-23). https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf. 11.01.2022

13 Global Mobile Market Report. https://newzoo.com/key-numbers#mobile. 08.01.2022

14 Castro, D., & McQuinn, A. (2015). Cross-Border Data Flows Enable Growth in All Industries (pp 2). https://www2.itif.org/2015-cross-border-data-flows.pdf. 14.01.2022

15 International Telecommunication Union. (2021). World Telecommunication/ICT Indicators Database 2021. https://www.itu.int/en/ITU-D/Statistics/Pages/publications/wtid.aspx. 09.01.2022

16 International Telecommunication Union. (2021). World Telecommunication/ICT Indicators Database 2021. https://www.itu.int/en/ITU-D/Statistics/Pages/publications/wtid.aspx. 09.01.2022

17 PEW Research Center. (2019). For World Population Day, a look at the countries with the biggest projected gains – and losses – by 2100. https://www.pewresearch.org/fact-tank/2019/07/10/for-world-population-day-a-look-at-the-countries-with-the-biggest-projected-gains-and-losses-by-2100/. 09.01.2022

18 Lee, K. F. (2018). AI superpowers : China, Silicon Valley, and the new world order.

19 TeleGeography. (2022). Global Internet Map 2022. https://global-internet-map-2022.telegeography.com/. 10.01.2022

20 These intra- and interregional data flows refer to all internet bandwidth usage, personal and non-personal combined.

21 Cory, N., & Dascoli, L. (2021). How Barriers to Cross-Border Data Flows Are Spreading Globally, What They Cost, and How to Address Them. https://itif.org/publications/2021/07/19/how-barriers-cross-border-data-flows-are-spreading-globally-what-they-cost. 13.01.2022ceme

22 McCann, D., Patel, O., & Ruiz, J. (2020). The Cost of Data Inadequacy. https://neweconomics.org/2020/11/the-cost-of-data-inadequacy. 14.01.2022

23 Cory, N. & Dascoli, L. (2021). How Barriers to Cross-Border Data Flows Are Spreading Globally, What They Cost, and How to Address Them. https://itif.org/publications/2021/07/19/how-barriers-cross-border-data-flows-are-spreading-globally-what-they-cost. 09.01.2022

24 Luo, D. & Wang, Y. (2021). China - Data Protection Overview. https://www.dataguidance.com/notes/china-data-protection-overview. 09.01.2022

25 Cory, N. (2017). Global Digital Trade I: Market Opportunities and Key Foreign Trade Restrictions (pp 7-9). https://www2.itif.org/2017-usitc-global-digital-trade.pdf. 09.01.2022

26 TeleGeography. (2022). Global Internet Map 2022. https://global-internet-map-2022.telegeography.com/. 10.01.2022

27 TeleGeography. (2021). Global Internet Map 2021. https://global-internet-map-2021.telegeography.com/. 09.01.2022

28 Cisco. (2021). Cisco Annual Internet Report (2018-2023). https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html. 13.01.2022

29 IDC. (2020). IoT Growth Demands Rethink of Long-Term Storage Strategies, says IDC. https://www.idc.com/getdoc.jsp?containerId=prAP46737220. 14.01.2022

30 Vailshery, L. S. (2021). Forecast for Internet of Things (IoT) spending share worldwide 2021, by category. https://www.statista.com/statistics/270843/distribution-of-internet-of-things-revenue/. 14.01.2022

31 Asia, Pacific, and China excluding Japan

32 Europe, Middle East, and Africa

33 Holst, A. (2021). Number of IoT connected devices 2019-2030, by region. https://www.statista.com/statistics/1194677/iot-connected-devices-regionally/. 14.01.2022

34 OECD. (2021). Business use of broadband (indicator). https://data.oecd.org/broadband/business-use-of-broadband.htm#indicator-chart. 09.01.2022

35 Brotman, S. N. (2018). The Net Vitality Index In Detail (pp 5). https://www.trpiresearch.org/uploads/1/9/2/9/19298259/net_vitality_2.0_index_in_detail.pdf. 09.01.2022

36 International Telecommunication Union. (2021). World Telecommunication/ICT Indicators Database 2021. https://www.itu.int/en/ITU-D/Statistics/Pages/publications/wtid.aspx. 09.01.2022

37 Fixed-broadband is a type of internet connection to a fixed location which is usually used by businesses because of its very high speed. Because it is mainly used by businesses, it serves as a good indicator for how much data businesses create and upload, i.e. non-personal data.

38 TeleGeography. (2021). Global Internet Map 2021. https://global-internet-map-2021.telegeography.com/. 09.01.2022

39 Richter, F. (2021). Amazon Leads $150-Billion Cloud Market. https://www.statista.com/chart/18819/worldwide-market-share-of-leading-cloud-infrastructure-service-providers/. 13.01.2022

40 TeleGeography. (2022): Global Internet Map 2022. **https://global-internet-map-2022.telegeography.com/**. 13.01.2022

41 Hunton Andrews Kurth. (2021). EU Parliament and Council of the EU Reach Agreement on Data Governance Act. **https://www.natlawreview.com/article/eu-parliament-and-council-eu-reach-agreement-data-governance-act**. 11.01.2022

42 Bonfiglio, F. (2021). Vision & Strategy (pp 3-5). **https://gaia-x.eu/sites/default/files/2021-12/Vision%20%26%20Strategy.pdf**. 09.01. 2022

43 African Continental Free Trade Area. (2022). About the African Continental Free Trade Area (AfCFTA). **https://afcfta.au.int/en/about**. 14.01.2022

44 Support Centre for Data Sharing (n.D.): What is data sharing? **https://eudatasharing.eu/what-data-sharing** 11.02.2022

45 Bokman, A.; Fiedler, L.; Perrey, J.; and Pickersgill, A. (2014): Five facts: How customer analytics boosts corporate performance. McKinsey. **https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/five-facts-how-customer-analytics-boosts-corporate-performance** 11.02.2022

46 Collibra (2020): Survey Shows Data-centric Businesses are 58% More Likely to Exceed Revenue Goals. **https://www.prnewswire.com/news-releases/survey-shows-data-centric-businesses-are-58-more-likely-to-exceed-revenue-goals-301060363.html** 11.02.2022

47 Grand View Reserach (2020): Data Lake Market Size, Share & Trends Analysis Report By Type (Solution, Services), By Deployment (Cloud, On-premise), By Vertical (IT, BFSI, Retail, Healthcare), By Region, And Segment Forecasts, 2020 - 2027 **https://www.grandviewresearch.com/industry-analysis/data-lake-market** 11.02.2022

48 Borasi, P.; Khan, S.; Kumar, V. (2021): Data Warehousing Market by Type of Offering (ETL Solutions, Statistical Analysis, Data Mining, and Others), Type of Data (Unstructured and Semi-Structured & Structured), Deployment Model (On-Premise, Cloud, and Hybrid), Enterprise Size (Large Enterprises and Small & Medium Enterprises), and Industry Vertical (BFSI, IT & telecom, Government, Manufacturing, Retail, Healthcare, Media & Entertainment, and Others): Global Opportunity Analysis and Industry Forecast, 2021–2028. Allied Market Research. **https://www.alliedmarketresearch.com/data-warehousing-market** 11.02.2022

49 Narain, R.; Merrill, A.; Lesser, E. (2016): Innovation in the API economy. IBM Institute for Business Value. **https://www.ibm.com/downloads/cas/OXV3LYLO** 11.02.2022

50 Bettendorf, M. (2020): API Growth Continues to Skyrocket in 2020 and into 2021. Postman. **https://blog.postman.com/api-growth-rate/** 11.02.2

51 Google Cloud (2020). State of API Economy 2021 Report. **https://pages.apigee.com/api-economy-report-confirmation-ty.html** 11.02.2022

52 Verhulst, S. G.; Young, A.; Winowatan, M.; Zahuranec, A. (2019): Data Collaboratives. Leveraging Private Data for Public Good. Page 11. GovLab. **https://datacollaboratives.org/static/files/existing-practices-report.pdf**

53 GovLab (n.D.): Data Collaborative Explorer. **https://datacollaboratives.org/explorer.html** 11.02.2022

54 Groth, O.; Esposito, M.; Tse, T. (2020): Corona-blues: The next hollowing-out of the economy. California Mangement Review. **https://cmr.berkeley.edu/2020/10/corona-blues/** 11.02.2022

55 The Original Platform Fund (n.D.): Platform-Index. **https://www.theoriginalplatformfund.de/plattform-index** 11.02.2022

56 Atali, A.; Gnanasambandam, C.; Srivathsan, B. (2019): Transforming infrastructure operations for a hybrid-cloud world. McKinsey. **https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/transforming-infrastructure-operations-for-a-hybrid-cloud-world** 12.02.2022

57 Schmidt, H. (2021): Die zehn wertvollsten Unternehmen der Welt. Twitter. **https://twitter.com/HolgerSchmidt/status/1434798781103542276/photo/1** 11.02.2022

58 59. Business Wire (2021): Global Cloud Computing Market (2021 to 2028) - Size, Share & Trends Analysis Report - ResearchAndMarkets.com **https://www.businesswire.com/news/home/20210810005902/en/Global-Cloud-Computing-Market-2021-to-2028---Size-Share-Trends-Analysis-Report---ResearchAndMarkets.com** 12.02.2022

59 **https://www.statista.com/chart/18819/worldwide-market-share-of-leading-cloud-infrastructure-service-providers/**

60 Seagate (2020): Rethink Data - Bessere Nutzung von mehr Unternehmensdaten – vom Netzwerkrand bis hin zur Cloud

61 Seagate (2020): Rethink Data - Bessere Nutzung von mehr Unternehmensdaten – vom Netzwerkrand bis hin zur Cloud

62 Galloway, S. (2020): Post Corona: From Crisis to Opportunity. Portfolio. Book. Page 65

63 Lanteri, A.; Esposito, M., Tse, T. (2021): From fintechs to banking as a service: global trends banks cannot ignore. LSE Blog. **https://blogs.lse.ac.uk/businessreview/2021/01/19/from-fintechs-to-banking-as-a-service-global-trends-banks-cannot-ignore/** 12.02.2022

64 Perlman, C. (2017): From product to platform: John Deere revolutionizes farming. Harvard Business School, Digital Initiative.

65 Deere (2017): Mobile App helps farmers grow. Press Release. **https://www.deere.com.au/en/our-company/news-and-announcements/press-releases/2017/february/Connect-Mobile-App-helps-growers.html** 20.09.2021

66 Galloway, S. (2020): Post Corona: From Crisis to Opportunity. Portfolio. Book.

67 Jones, C., Tonetti, C. (2019): Nonrivalry and the Economics of Data. Stanford Business Working Paper

68 Roubini, N. (2021): Nouriel Roubini: bitcoin is not a hedge against tail risk. Financial Times. https://www.ft.com/content/9be5ad05-b17a-4449-807b-5dbcb5ef8170 12.02.2022

69 Groth, O.; Straube, T.; Zehr, D. (2021): A Crypto Cataclysm? The Case for The Long View. California Management Review. https://cmr.berkeley.edu/2021/04/crypto-cataclysm/ 12.02.2022

70 Nguyen, T. (2021): 4 Impactful Technologies From the Gartner Emerging Technologies and Trends Impact Radar for 2021. Gartner. https://www.gartner.com/smarterwithgartner/4-impactful-technologies-from-the-gartner-emerging-technologies-and-trends-impact-radar-for-2021 12.02.2022

71 Lützow-Holm Myrstad, F. (2018): How Tech Companies deceive you into giving up your Data and

72 To be successful in college, students need to be able to understand texts with a score of 1300 on the Lexile test developed by the education company Metametrics. The test measures a text's complexity based on factors such as sentence length and the difficulty of vocabulary. Many privacy policies exceed that 1300 threshold.

Source: Litman-Navarro, k. (2019): We Read 150 Privacy Policies. They Were an Incomprehensible Disaster. https://www.nytimes.com/interactive/2019/06/12/opinion/facebook-google-privacy-policies.html 12.02.2022

73 Europe (n.D.): Database protection. https://europa.eu/youreurope/business/running-business/intellectual-property/database-protection/index_en.htm 12.02.2022

74 Stepanov, I. (2020): Introducing a property right over data in the EU: the data producer's right – an evaluation, International Review of Law, Computers & Technology, 34:1, 65-86, DOI: 10.1080/13600869.2019.1631621 12.02.2022

75 McMahan, B.; Ramage, D. (2017): Federated Learning: Collaborative Machine Learning without Centralized Training Data. https://ai.googleblog.com/2017/04/federated-learning-collaborative.html 12.02.2022

76 Groth, O.; Straube, T. (2021): Analysis of current global AI developments with a focus on Europe. Konrad Adenauer Foundation. https://www.kas.de/en/single-title/-/content/analysis-of-current-global-ai-developments-with-a-focus-on-europe 12.02.2022

77 Li, Y.; Chen, C.; Liu, N.; Huang, H.; Zheng, Z.; Yan, Q. (2021): A Blockchain-Based Decentralized Federated Learning Framework with Committee Consensus. IEEE Network, vol. 35, no. 1, pp. 234-241, January/February 2021. https://ieeexplore.ieee.org/document/9293091 12.02.2022

78 Ocean Protocol Team (2021): Ocean Protocol works with Raven Protocol to add Federated Learning to Ocean Market via Compute-to-Data. https://blog.oceanprotocol.com/ocean-protocol-and-raven-protocol-to-add-federated-learning-to-ocean-market-via-compute-to-data-f0314575d3c8 12.02.2022

79 Lee, H.; Kim, J. (2021): Trends in Blockchain and Federated Learning for Data Sharing in Distributed Platforms. https://arxiv.org/pdf/2107.08624.pdf 12.02.2022

80 Steel, E., Locke, C., Cadman, E., Freese, B. (2018): How much is your personal data worth? https://ig.ft.com/how-much-is-your-personal-data-worth/ 12.02.2022

81 Li, C., Li, D. Y., Miklau, G., Suciu, D. (2012). A Theory of Pricing Private Data. https://dl.acm.org/doi/10.1145/2448496.2448502 12.02.2022

82 Groth, O.; Straube, T.; Zehr, D. (2020): Personal Data is valuable,. https://www.wired.com/story/opinion-give-data-pricing-power-to-the-people 12.02.2022

83 Lanier, J.; Weyl, E. G. (2018): A Blueprint for a Better Digital Society. Harvard Business Review. http://eliassi.org/lanier_and_weyl_hbr2018.pdf 12.02.2022

84 Groth, O.; Straube, T., Zehr, D. (2019): A Privacy-Assured Market Design for the Emerging Trillion Dollar Asset Class of Data. https://www.cambrian.ai/data-marketplace 12.02.2022

85 IEEE (n.D.): Data trading initiative. https://standards.ieee.org/industry-connections/datatradingsystem.html 12.02.2022

86 Nikkei Asia (2017): Japan takes step toward enormous bank of personal data. https://asia.nikkei.com/Economy/Japan-takes-step-toward-enormous-bank-of-personal-data 12.02.2022

87 G20 Japan (2019): G20 Osaka Leaders' Declaration. www.mofa.go.jp/policy/economy/g20_summit/osaka19/en/documents/final_g20_osaka_leaders_declaration.html 12.02.2022

88 World Economic Forum (2021): Data for Common Purpose Initiative (DCPI). https://www.weforum.org/projects/data-for-common-purpose-initiative-dcpi 12.02.2022

89 World Economic Forum (2021): Developing a Responsible and Well-designed Governance Structure for Data Marketplaces. Briefing Paper. www.weforum.org/docs/WEF_DCPI_Governance_Structure_Towards_Data_Exchanges_2021.pdf 12.02.2022

90 Cutler, W.; Meltzer, J. P. (2021): Digital trade deal ripe for the Indo-Pacific. Brookings. www.brookings.edu/opinions/digital-trade-deal-ripe-for-the-indo-pacific/ 12.02.2022

91 Gartner (2021): Gartner Says 70% of Organizations Will Shift Their Focus From Big to Small and Wide Data By 2025. Press Release. www.gartner.com/en/newsroom/press-releases/2021-05-19-gartner-says-70-percent-of-organizations-will-shift-their-focus-from-big-to-small-and-wide-data-by-2025 14.02.2022

92 Redman, T. (2016): Bad Data Costs the U.S. $3 Trillion Per Year. Harvard Business Review. https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year 14.02.2022

93 Neutatz, Felix et al.: Towards automated data cleaning Workflows, Berlin 2019, LWDA Conference 2019. www.researchgate.net/publication/335136628_Towards_Automated_Data_Cleaning_Workflows 14.02.2022

94 Dimensional Research (2019): Artificial Intelligence and Machine Learning Projects Are Obstructed by Data Issues. https://cdn2.hubspot.net/hubfs/3971219/Survey%20Assets%201905/Dimensional%20Research%20Machine%20Learning%20PPT%20Report%20FINAL.pdf 14.02.2022

95 Cerliani, Marco: Automate data Cleaning with unsupervised learning, in: towardsdatascience, 09/2019, https://towardsdatascience.com/automate-data-cleaning-with-unsupervised-learning-2046ef59ac17 14.02.2022

96 Brownlee, J. (2020): How to Perform Data Cleaning for Machine Learning with Python. https://machinelearningmastery.com/basic-data-cleaning-for-machine-learning/ 14.02.2022

97 Ye, Han-Jia et al: Few shot learning with a strong teacher, Journal of Latex class files, 2021. www.researchgate.net/publication/353056915_Few-Shot_Learning_with_a_Strong_Teacher 14.02.2022

98 Philippakis, Anthony et al., "The Matchmaker Exchange: A Platform for Rare Disease Gene Discovery," Human Mutation, Volume36, Issue 10, October 2015, 915–921, https://onlinelibrary.wiley.com/doi/full/10.1002/humu.22858 14.02.2022

99 Altae-Tran, H.; Ramsundar, B.; Pappu, A. S., Pande, V. (2017): Low Data Drug Discovery with One-Shot Learning. ACS Central Science 2017 3 (4), 283-293. https://pubs.acs.org/doi/10.1021/acscentsci.6b00367

100 The Real promise of synthetic data, MIT News October 2020, https://news.mit.edu/2020/real-promise-synthetic-data-1016; Raghunatan, Trivellore: Synthetic data, Annual review of statistics and its applications, 2021, www.researchgate.net/publication/345364616_Synthetic_Data 14.02.2022

101 Mayer, Nikolaus et al: What makes good synthetic data for learning disparity and optical flow estimation?, International journal of computer vision 2018, https://lmb.informatik.uni-freiburg.de/projects/synthetic-data/ 14.02.2022

102 Fintz, Matan et al: Synthetic data for model selection, Workshop on Synthetic Data Generation at ICLR 2021, www.researchgate.net/publication/351298240_Synthetic_Data_for_Model_Selection 14.02.2022

103 Tiwald, Paul et al: Representative and fair synthetic data, www.researchgate.net/publication/350720154_Representative_Fair_Synthetic_Data 14.02.2022

104 Damiani, J. (2019): A Voice Deepfake Was Used To Scam A CEO Out Of $243,000. Forbes. www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/?sh=7e0e76d02241 14.02.202

105 Vincent, J. (2019): Deepfake detection algorithms will never be enough. The Verge. www.theverge.com/2019/6/27/18715235/deepfake-detection-ai-algorithms-accuracy-will-they-ever-work 14.02.2022

106 OpenAI, Pilipiszyn, A. (2021): GPT-3 Powers the Next Generation of Apps. OpenAI. https://openai.com/blog/gpt-3-apps/ 14.02.2022

107 Likens, S.; Shehab, M.; Rao, A. (n.D.): AI Predictions 2021. PwC. www.pwc.com/us/en/tech-effect/ai-analytics/ai-predictions.html 14.02.2022

108 Tse, Terence (2019): Four reasons why your business isn't using AI. California Review Management https://cmr.berkeley.edu/2019/07/why-your-business-isnt-using-ai/ 14.02.2022

109 Li, Yaliang et al: AutoML: From methodology to application, Conference CIKM 2021, www.researchgate.net/publication/356249067_AutoML_From_Methodology_to_Application; van der Blom, Koen et al: AutoML Adoption in ML Software, 8th ICML Workshop on automated machine learning 2021, www.researchgate.net/publication/355153049_AutoML_Adoption_in_ML_Software 14.02.2022

110 Ploder, Christian et al: The future use of low code/no code platforms by knowledge workers, International conference on knowledge management in organizations, 2019, www.researchgate.net/publication/333716085_The_Future_Use_of_LowCodeNoCode_Platforms_by_Knowledge_Workers_-_An_Acceptance_Study 14.02.2022

111 Gillin, P. (2020): Low-code and no-code tools may finally usher in the era of 'citizen developers'. Silicon Angle. https://siliconangle.com/2020/10/06/low-code-no-code-tools-may-finally-usher-era-citizen-developers/ 14.02.2022

112 Lorenz, P.; Saslow, K. (2019): Demystifying AI & AI Companies. Stiftung Neue Verantwortung. Page 19. www.stiftung-nv.de/sites/default/files/demystifying_ai_and_ai_companies.pdf 14.02.2022

113 Bonime-Blanc, A. (2020): It's time we added a letter to ESG. Here's why. World Economic Forum. www.weforum.org/agenda/2020/10/its-time-we-added-a-letter-to-esg-heres-why/ 14.02.2022

Authors

CAMBRIAN
FUTURES

Olaf J. Groth, PhD
CEO, Cambrian LLC

Global Strategist, Entrepreneurial Thinktank Founder, Trusted Advisor and Serial Author for Cognitive and DeepTech Disruption, Discontinuities and Strategies in the Global Economy.

Olaf has 25 years of experience as an executive and adviser building strategies, capabilities, programs and ventures across 35+ countries (including UAE) with multinationals (e.g. AirTouch, Boeing, Chevron, GE, Qualcomm, Q-Cells, Vodafone, Volkswagen, etc.), consultancies, startups, VCs, foundations, governments and academia. He is the founding CEO of advisory thinktank Cambrian Futures and of concept development firm Cambrian Designs. Olaf serves as Professional Faculty for strategy, technology, innovation and futures at UC Berkeley's Haas School of Business. Professor of Practice for global strategy, innovation, economics & futures at Hult International Business School teaching across campuses in the US, Europe, Middle East and China. He is a member of the Global Expert Network for the 4th Industrial Revolution and Positive AI Economy Futures at the World Economic Forum, Visiting Scholar at UC Berkeley's Roundtable on the International Economy (BRIE) and its program Working with Intelligent Tools & Systems (WITS), a member of the Innovation Policy Committee for Biden/Harris and the CleanTech For Obama (CT4O) steering committee.

Olaf is lead co-author of the 2018 AI book *Solomon's Code: Humanity in a World of Thinking Machines* and its 2021 paperback version The AI Generation: Shaping Our Global Future With Thinking Machines). He is also co-author for the forthcoming 2023 MIT Press book *The Great Remobilization: Designing A Smarter Future*. He is a frequent commentator on ABC, CBS and NPR (USA), Deutsche Welle, ARD and ZDF (Germany) and contributor for outlets like *WIRED*, *Financial Times*, *The Hill*, *Harvard Business Review* (USA, Germany, France, Italia, Spain, Arabia), *California Management Review*, *Quartz*, *FOCUS*, *Die Zeit*, *World Economic Forum*, *Huffington Post*, *Peter Drucker Forum*, *LSE*, Today's CFO, Thunderbird International Business Review, World Financial Review, *European Business Review*, *Roubini EconoMonitor*, and *Duke CE Dialogue*.

Olaf holds PhD & MALD degrees in global affairs with business, economics and technology focus from the Fletcher School at Tufts University, MAIPS & BA degrees with economics focus from the Middlebury Institute at Monterey, studied negotiation at Harvard, economics at Georgetown, finance at Berkeley, and strategic leadership at the Center for Creative Leadership.

A naturalized immigrant from Germany, Olaf lives and works globally (including the Middle East) from the San Francisco Bay Area with his wife and two daughters.

Tobias Straube
VP Analysis, Cambrian LLC

Advisor and entrepreneur in the fields of DeepTech and international economic development with 10+ years working experience in Europe, Africa, Latin America, Asia and USA.

CAMBRIAN
FUTURES

Tobias Straube is VP Analysis at Cambrian, board member at Digital Waves, assistant instructor at UC Berkeley Executive Education and co-founder of Scio Network. At Cambrian, a Berkeley-based lab for product and strategy design for the human-machine economy, Tobias has led and co-designed international strategy and research projects, such as the development of an AI strategy for a German medium-sized company, a case study on Baidu Apollo, an analysis of thirteen national AI strategies for the Konrad Adenauer Foundation, an digital economy assessment of UAE's AI Ministry, and the development of an AI governance checklist for legal counsels in collaboration with LexMundi and Microsoft Europe, to name a few. He has also co-authored eight patents on data protection management and data marketplaces. At Digital Waves, an investment service firm based in Switzerland, Tobias is leading the venture building and strategy development of investment products in alternative and digital assets, such as the equity disruptive tech startups or DeFi token. At UC Berkeley Executive Education and Hult International Business School, Tobias supports Olaf J Groth, PhD, in delivering executive education course on Future of Technology, Disruption and the Global Economy.

Tobias brings deep experience in governance, higher education, and entrepreneurship projects across the Global South. As an adviser to the German Agency for International Cooperation (GIZ), Mr. Straube co-designed and co-led projects with combined budgets of $40 million. Those efforts included key advisory and leadership roles in the creation of the Pan African University, the establishment of a global "AI for All" Lab and the tech entrepreneurship initiative "Make IT Africa".

Tobias frequently gives webinars and speaks at Cambrian and client events, often jointly with Olaf Groth, covering subjects such as deep technologies, innovation management, university-industry collaboration, and governance. He also publishes, amongst others for *Wired*, *California Management Review*, *European Business Review*, *Global Solution Magazine*, *ZeitOnline* and *Hult Blog*.

Tobias holds an Executive MBA from the Hult International Business School and a BA degree in international political management from the Bremen University of Applied Sciences (both with distinction). He is an alumnus of the German Merit Foundation.

Tobias lives in Hamburg, Germany, but works globally.

KONRAD
ADENAUER
STIFTUNG

The Konrad-Adenauer-Stiftung (KAS) is a German political foundation and think tank. Worldwide, KAS is in charge of over 200 projects in more than 120 countries, where it promotes freedom and liberty, peace, and justice. KAS also focuses on consolidating democracy, the unification of Europe and the strengthening of transatlantic relations, as well as on development cooperation. The KAS office in Washington DC was established in the late 1970s and to this day it stands for and sees itself as a promoter of these values.

Published by:

Konrad-Adenauer-Stiftung USA
1233 20th Street, NW
Suite #610
Washington, DC 20036
U.S.A.
Tel.: +1(202) 464-5840
www.kas.de/usa

Disclaimer:

The views, conclusions and recommendations expressed in this report are solely those of its author(s) and do not reflect the views of the Konrad-Adenauer-Stiftung, or its employees. This publication of the Konrad-Adenauer-Stiftung e. V. is solely intended for information purposes. It may not be used by political parties or by election campaigners or supporters for the purpose of election advertising. This applies to federal, state and local elections as well as elections to the European Parliament.

Cover illustration: ©Carmel Steindam Graphic Design

KONRAD
ADENAUER
STIFTUNG

Konrad-Adenauer-Stiftung USA
1233 20th Street, NW
Suite #610
Washington, DC 20036
U.S.A.
www.kas.de/usa