WERKSCHULHEIM FELBERTAL
ZUKUNFTS>CAMPUS

Data as an Asset

On the Economical Value of Data Collection in the Digital Age

Final thesis

by

Simon Michael Wimmer

13th grade

Supervising teacher:
Robert Mutter

Ebenau, February 2025

Werkschulheimstraße 11, 5323 Ebenau

This English version is based on the original German paper. The translation was supported by AI tools and carefully reviewed and edited by the author.

Abstract

With the advent of digital commerce and advertising, user data has become a valuable corporate resource that is now traded through dedicated brokers and whose uses extend far beyond the realm of marketing.

The aim of this paper is to explain the principles and functioning of this market to the reader in a scientific manner, i.e., objectively and soberly, without making premature accusations.

To this end, a number of key research questions were defined: What are the fundamental differences between the various types of data and how are they collected? What legal regulations apply to their collection? What economic standards does the data trading market achieve? And what are the advantages and disadvantages of this type of information trading?

To answer these questions, extensive specialist literature was consulted and critically evaluated. The specialist literature on this topic was often published by pure business economists on the one side or pure data protection advocates on the other. The aim of this paper is to build bridges and present the facts as neutrally as possible. To this end, the data brokers' prospectuses were also consulted and examined. This is rounded off by the author's own connections and theories.

The work highlights the growing importance of data trading in the modern economy, but does not forget to point out the potential risks.

Foreword

Modern data processing and the inevitable collection of data that goes with it have become omnipresent and indispensable in our modern world. Long before I began my technical training, I was interested in the fundamental concepts of data processing and its profound impact on our world.

At the same time, I developed a strong interest in economics, which led me to look at data collection not only from a technical perspective, but also to explore its macroeconomic dimension.

I would like to take this opportunity to express my gratitude for the technically and universally scientific education I received at the Werkschulheim Felbertal, for the highly exciting technical discussions I was able to have with my training instructor, and to all those who made this path possible for me. Anyone who feels addressed is most likely meant.

Braunau am Inn, February 16, 2025 Simon Michael Wimmer

Contents

Introduction

Customer and user data are no longer an inconspicuous by-product of economic activity, but rather essential guides in all processes. From conception and development to the marketing of goods and services, they play a key role in identifying customer problems, optimizing existing applications, and presenting the product to the right target group.

Datafication is the term used to describe the increasing description of all measurable phenomena through digital data. Since the invention of computer-assisted data processing, the quantification of our world has been advancing inexorably. This is not surprising when you consider that measuring our world has played a role since the emergence of Homo sapiens.

The ultimate goal of datafication is, of course, not the mere aggregation of petabytes of information, but the pursuit of insights into humans and their environment that the analysis of such large amounts of data can provide us with. Insights into people's problems, hopes, and dreams are important for the economy, as the founding intention of every company is to satisfy a human need.

Business leaders around the globe recognized that it is difficult for a single company to collect the amount of data needed to gain such insights. In keeping with the spirit of *the sharing economy*, the solution was found in compiling data from a wide variety of sources. In exchange for monetary compensation, companies make the data they collect available to third parties. These third parties accumulate data from a wide variety of sources and can then resell data sets of impressive granularity to individual companies. This is how the business concept of data brokers was born.

Data broking is now a billion-dollar industry that has gained a foothold in all geographical markets. They are sought after by mega-corporations and medium-sized companies alike when they suffer from an information deficit.

But despite its obviously significant role in the global economy, this market receives little public attention. The aim of this paper is therefore to shed light on the dark server rooms of these data giants.

To this end, we first address fundamental questions such as: What is data and what types of data are there? Once this has been clarified, the various methods of collecting data will be discussed. Once the data has been collected, we can move on to the actual

focus of this work: the role of data as an asset, the structure and functioning of the market surrounding it, the players involved, and their economic significance. Of course, the legal requirements in the various phases must not be ignored. The most important framework conditions in the two largest political markets (the US and the EU) are presented. Last but not least, an economic analysis must, of course, include a comparison of the biggest advantages and disadvantages for both companies and consumers.

The knowledge required to answer the questions raised comes from numerous works of specialist literature and legal texts. It was important to the author to overcome the sometimes narrow perspectives of purists from business administration or data protection, which are sometimes reminiscent of a déformation professionnelle, in order to do justice to the claim of a scientifically neutral work.

At the same time, it should be noted on behalf of the author that a comprehensive analysis of all aspects of this market would go far beyond the scope of a single work. More than one work could be written on supposed sub-topics alone, such as the legal situation within the EU or industry analysis. The intention of this work is to provide a meaningful overview of the market as a holistic construct.

0. Fundamentals of the data world

0.1 Definition of the term "data"

In order to answer the introductory, essential question of how data should be categorized, it is first necessary to clarify the meaning behind the frequently used yet meaningless umbrella term "data."

Etymologically speaking, data is the plural form of *datum*. The loanword comes from the Latin past participle (PPP) of "dare," which can be translated as "to give"[1] . Documents that were handed over were marked with this introductory phrase, which included the place and time.[2]

The current definition of the plural noun "data" is as follows: "collected information; details about place, time, and other facts"[3] This is an interesting form of polysemy, in which, at least in colloquial language, a strict distinction is made between the meaning of the singular and that of the plural.[4] To make it easier to distinguish between the original definition in the temporal sense and the new definition in the technical sense, synonyms are used for the Numeri of the different meanings. For example, the singular of "data" as general information is often expressed as "data point."

So much for the meaning of the word in colloquial usage. In information technology, however, a more precise breakdown is necessary, as a clearer distinction must be made between data, information, and knowledge. In everyday use, these terms are used almost synonymously, as can be seen from the dictionary definition. This is also logical when you consider that we mostly speak in whole sentences in which data points are used to form information and knowledge is created through the context of a conversation. The boundaries become blurred, which also makes it difficult to differentiate between them.

When storing information digitally, however, knowledge is inevitably broken down into its atomic components. For example, when you fill out a form, you are passing on your data. However, information can only be said to exist when the data point entered is linked to another data point that describes it. The second type of data is called metadata and is discussed in more detail in the following chapter. The entry "Wimmer"

---

[1] Lošek 2016 , p. 198
[2] Kluge und Seebold 1999 , p. 163
[3] Pabst, Fussy und Steiner (Red.) 2016 , p. 155
[4] Pabst, Fussy und Steiner (Red.) 2016 , p. 156

in a data field is meaningless and does not yet constitute information. It could just as easily be a street or product name instead of a surname. Only when the attribute description, in this case "surname," is added can it be referred to as information. The networking of information can ultimately be understood as knowledge.

German business information scientist Freimut Bodendorf has developed a four-stage model for defining data in the information technology sense, which arranges the terms characters, data, information, and knowledge hierarchically and describes how contextualization, semantics, and networking ultimately create knowledge from simple characters.

According to him, data is *"formed from characters in a character set according to defined syntax rules [...]. Data becomes information when it is assigned a meaning (semantics)."*[5]

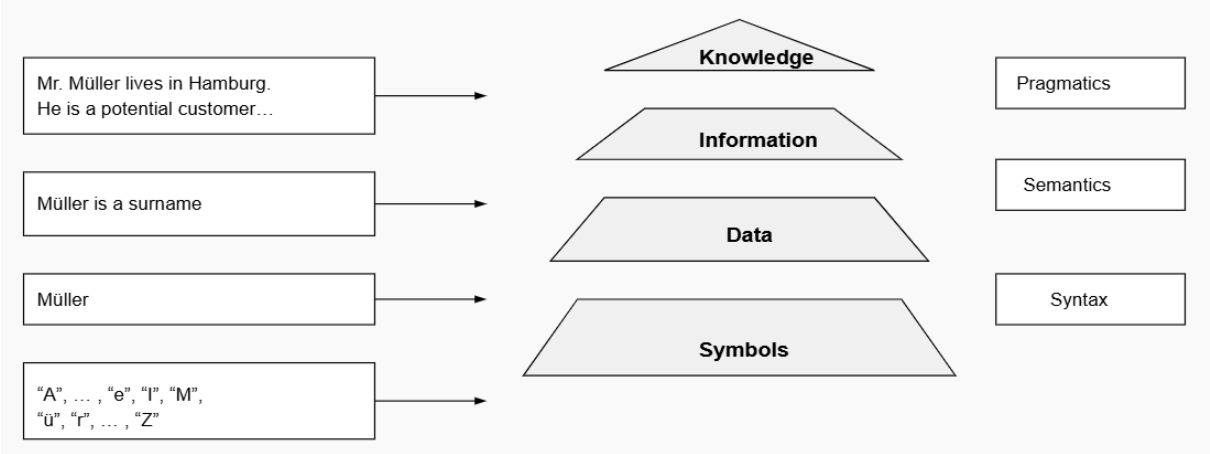

Fig.1: Concept hierarchy. Bodendorf 2006, p. 1

[5] Bodendorf 2006 , p. 1

0.2 Categorization of data

The types of information and facts that fall under the umbrella term data are manifold. It is therefore essential to differentiate between types of data with different characteristics and to categorize them accordingly. Especially for later discussions in which the business advantages are weighed against the potential individual disadvantages, it is essential to define clear names for incomparable types of data so that the opposing parties are always talking about the same concepts during the discourse.

However, the main problem in categorizing data lies in the fact that with the advent of electronic data processing and its introduction into almost all areas of work, a wide variety of categorization systems have become established in the various fields of work, some of which have parallels but are almost never identical. None of these systems can be objectively judged as wrong or right; rather, their essence lies in reflecting the way data is handled in the specific work domain and the purposes it serves there.

0.2.1 Information technology

In the field of information technology, i.e., in IT departments or among database administrators, for example, the type of storage and collection plays a primary role. When it comes to storage, a concrete distinction is made between three structural forms: structured, semi-structured, or unstructured.

As the name suggests, structured data has a clearly defined structure. The form in which the data is stored is defined in advance. Examples of this would be relational databases or CSV documents, in which the data is entered into tables, with each field representing a specific data point and each row representing a data record. Each data point is contextualized by another descriptive data point, which refines it into information. In the example above, this would be the title of the table column. These descriptive data points are also referred to as "metadata." The from ancient Greek borrowed prefix *"meta*,*"* symbolizes that the element is hierarchically one level above.[6] This is data that characterizes, organizes, or identifies other data.[7] It forms another subcategory of data whose existence is necessary for all structured and semi-structured data. Structured data can be easily processed by machines, which is why it is considered essential in data analysis.

[6] Duden (Hrsg.) o.J.
[7] Kranz 2024

Unstructured data is an accumulation of data without metadata. A priori, no information can be extracted from it. Instead, it must be contextualized extrinsically. Examples of unstructured data are graphics or this text. An extrinsic force can be a person who uses their prior linguistic knowledge in the areas of syntax and semantics to understand this text. Machines can contextualize unstructured data using special algorithms.[8]

Semi-structured data is a hybrid of the two structures mentioned above. A file which contains both data points that have been contextualized using metadata and unstructured data. An example of this would be an email, which has clearly defined fields for recipients, senders, and timestamps on the one hand, and unstructured text content on the other. Other common file types for semi-structured data are XML and JSON.[9]

While humans in their pure form of existence produce only unstructured data, when interacting with information technology systems, this data is almost always transformed into semi-structured data. Most application programs independently create metadata for the unstructured data entered. In most cases, this is at least statistical data such as the time of data entry, the user who entered it, the volume of the data entered, and the like.

Structured data is usually generated independently by machines (e.g., reading from sensors) or when machines force the user to input data in a structured way, as is the case with forms.

Furthermore, when storing data, it is possible to distinguish not only the structure in which the data was stored, but also the data types.

Data types in the programming sense are structures within computer software that are assigned a defined storage space in the computing system and, as a result, a possible value range.[10] Furthermore, operations (e.g., addition, multiplication, negation, etc.) that can be performed with them are assigned to the data types.[11]

Data types can then be broken down into standard types and derived types.[12] Standard types are specified by the programming language and have become established

[8] Wuttke 2024
[9] Seiter 2023 , p. 24
[10] Microsoft (Hrsg.) 2023
[11] Lackes und Siepermann 2018
[12] Langer 1993 , p. 21

across individual languages in the entire field of computer science.[13] It should be noted, however, that despite the familiar trivial names and comparable value ranges and operation options, the exact ranges and options can vary between different programming languages. Derived types are defined independently by the programmer. The memory space and value range can be freely selected within the limits of the language, and it is even possible to develop your own operations.

[13] Lange und Stegemann 1985 , p. 31

Below is an overview of the most common data types and their characteristic value ranges:[14]

Data type	Description	Characteristic storage space requirement [bits]	Characteristic value range
boolean	Boolean value	1	True or false
byte	Whole number	8	-128 to 127
date	Date	8	-657,434 (January 1, 100) to 2,958,465 (December 31, 9999)
char	Characters	16	Unicode character
int	Whole number	32	$-2^{31}\ bis + 2^{31} - 1$
float	Floating point number	32	$\pm 3{,}40282347 \times 10^{38}$

Table1: Common data types. Own table.

0.2.2 Analytical

The analytical domain involves information technology specialists who use data specifically to gain insights from it. The proximity of the analytical domain to the information technology domain is also reflected in the similar categorization of data.

Computer-based algorithms are most commonly used in this domain for analytical knowledge discovery, and the programming language expects specific data types for their input and output variables. When entering data into an algorithm, it must therefore always be ensured that the data to be entered matches the data types expected by the programming language.

In addition, another meta-categorization has become established in the field of computer-aided analytics: the distinction between "non-dependency-oriented data" and "dependency-oriented data." According to Aggarwal, this categorization deals with

[14] Fässler, Scheuner und Sichau 2024 , p. 6

whether there is a relationship between the individual data points. Furthermore, a distinction is made between implicit and explicit relationships. An implicit relationship exists when the connection between the data points is not directly specified but can be inferred from the nature of the data. An example of this is sensor data that is recorded in rapid succession, for example once per second. Neighboring values within such a time series, such as the temperature values of a machine, are implicitly related because it is assumed that they do not show any abrupt changes. A significant deviation therefore indicates a relevant case. An explicit relationship, on the other hand, exists when the connection between data points is directly defined, as is the case with network data. A typical example of this is contacts between members on social media platforms, where connections are explicitly represented by friendships or follower relationships.[15]

Such relationships between data points are highly relevant for analysts, as they must be taken into account when interpreting the analysis results, and the choice of the most suitable algorithm depends heavily on the type of data. For a meaningful outlier analysis, there must be an implicit relationship between the data points.

0.2.3 Business management

In the context of data-driven corporate management in particular, there has been a need to develop a data categorization method that provides information about the business purpose of the data for a company's operational activities. In popular management literature, especially in German-speaking countries, a categorization method with seven data types has become established. Depending on the source, the names and definitions of these categories vary, and in some cases there is overlap between different categories. In view of this, this method should be regarded less as a scientific approach and more as an attempt by management consultants to familiarize executives with the world of data management. This impression is reinforced by the fact that much of the available documentation on this method is based on publications by management consultancies.

In the following, the author has attempted to find a common denominator for the different interpretations of this model and thus provide an insight into this methodology based on a high level of abstraction. Of course, differences in individual specifics may

[15] Seiter 2023 , p. 57

arise when consulting individual papers. This structure is most similar to the description by Daniel Liebhardt.[16]

1. **Metadata:** as already explained, metadata describes other data. A classic example of this is the title of a table column.
2. **Reference data:** categorizes general business entities. As with metadata, this is data that classifies other data. Unlike metadata, however, the values here rarely change and are mostly obtained from external sources. Examples include airport codes or a list of the states in a country. This is often referred to as "stable information."
3. **Company-wide structural data:** represents the organization and structure of a company, including its products, services, and responsibilities.
4. **Transaction structure data:** describes the structure and basic framework of business transactions. This typically includes the business entities involved, such as customers and products.
5. **Inventory data:** describes a company's assets and their quantity, such as stock levels, account balances, or the number of properties.
6. **Transaction data:** describes the commercial and legal business activities of a company. This is always based on an exchange of intellectual or physical goods. Examples of this would be invoices (incoming and outgoing), securities purchases, returns, or credit notes.
7. **Audit data:** records the individual steps of a transaction and is used for later verification. It ensures transparency with regard to data generation and provides a means of verifying correctness.[17]

0.2.4 Regulatory

A fourth domain, which will take on special significance later in this work, is the regulatory domain. This refers to the legislative power which, in the course of the increasing commercialization of user data from private individuals, has recognized that restrictions on data collection must be imposed in the interests of personal rights.

---

[16] Liebhart 2010 , p. 21
[17] dataspot. GmbH kein Datum

This categorization method is probably the most popular and is also the one most frequently used not only by lawyers or politicians, but a majority of the general public. They are particularly interested in which personal information can be expressed by means of data. This represents a shift away from the pure form or type of collection towards the insights that can be gained from it.

Most important in this domain is the definition of personal data. This includes all data records that contain information about natural persons, which can be clearly assigned to individual persons through identification. Identifiability is defined very broadly in the General Data Protection Regulation: *"A natural person is considered identifiable if they can be identified, directly or indirectly, in particular by association with an identifier such as a name, an identification number, location data, an online identifier or one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person*."[18]

Furthermore, the GDPR specifies certain subcategories of personal data that are considered particularly sensitive. These include, among others:

- sensitive personal data, the processing of which *"[...] may pose significant risks to fundamental rights and freedoms. Such personal data should include personal data revealing racial or ethnic origin [...]"*[19] . This data may only be processed in extremely exceptional cases.
- *"'genetic data' [...] relating to inherited or acquired genetic characteristics [...] which provide unique information about the physiology or health of that natural person and which result in particular from an analysis of a biological [...];"*
- *"'biometric data' [...] relating to the physical, physiological, or behavioral characteristics of a natural person that enable the unique identification of that natural person [...]"*
- *"'Health data' [...] relating to the physical or mental health of a natural person, including the provision of health services, and revealing information about their health status."*[20]

The regulatory perspective has been strongly influenced by the work of various data protection experts and reflects their way of thinking: the danger is not created by the

[18] Article 4(1) of Regulation (EU) 679/2016
[19] Recital 51 of Regulation (EU) 679/2016
[20] Article 4, paragraphs 13–15 of Regulation (EU) 679/2016

data per se, but by the conclusions that can be drawn from the networking and analysis of various data attributes. Countless data protection advocates have devised a multitude of categorization methods, the detailed examination of which would go beyond the scope of this work. In the following, a particularly simple but equally widespread method, that of NSA Chief Counsel April Falcon Doss, is presented as an example.

1. **Property data** – "what we have": this is data about our digital and analog possessions. The purchase history of a user who buys goods from an online retailer belongs to this category, as does a list of the applications they have installed on their mobile phone or the model of the digital device they use to access the internet.
2. **Behavioral data** – "what and how we act": includes patterns of our everyday actions, which can also be used to predict our future behavior. This is mostly data derived from property data. Retailers can use loyalty programs, for example, to record the purchasing behavior of their customers. Navigation applications can reconstruct our daily routine from movement data.
3. **Identification data** – "who we are": makes us uniquely identifiable as natural persons. On the one hand, this refers to alphanumeric identifiers that are assigned only once per person (within their scope of validity, e.g., within a state, an organization, etc.). These include social security numbers, account numbers, license plate numbers, and telephone numbers. But it also includes general physical characteristics that can be assigned to a specific person. This includes, in particular, biometric data (fingerprints, iris recognition, etc.) as well as genetic sequences.
4. **Psychographic data** – "what we believe, know, and think": The distinction from behavioral data is made on a psychological level. While behavioral data is usually understood to mean subconscious, regular processes that can be freely observed in the analog world by fellow human beings, psychographic data refers to cognitive decisions or personal characteristics that can be consciously communicated or concealed. In psychology, this refers to the personality aspect of the *private individual*, i.e., aspects whose disclosure is within the person's own sphere of decision-making.[21] This group includes political attitudes, sexual preferences, and values. [22]

[21] Lahmer 2023 , p. 146
[22] Doss 2020 , pp. 11–31

These groups are by no means mutually exclusive. The boundaries are transitional, and information that can be assigned to one group is often obtained by processing data from another category. In most cases, the raw data generated by data collection is property data that, when analyzed, indicates behavioral patterns or psychographic data. Furthermore, with sufficient data and appropriate algorithms, it is possible to identify a natural person using non-identifying data. The intimacy of the data and the potential risks of misuse increase with the numbering.

0.3 Personal data and its correlation with private data

Although there is no clear definition of the term "private data," it is used below to refer to information whose potential risks of misuse are so extensive that the average user has concerns about its publication and therefore prefers to keep it private.

It is worth noting that our understanding of private data has changed fundamentally with increasing digitalization, and much data that is considered public in the analog world is classified as private in digital environments.

While general biometric data such as facial features or voice profiles are freely visible and audible to anyone passing by in the analog world and thus belong to publicly accessible data, they are listed in the GDPR as particularly sensitive.

While it is generally considered perfectly legitimate for a local baker to make personalized recommendations and offers to customers based on their previous purchasing behavior, there are media debates about the legitimacy of personalized ads on the websites of online retailers.

The answer to the question of why, in the two examples presented, identical data is assigned different risks in different situations (computer-assisted or human processing) is easy to find: While it is unlikely that a passerby would be able to imitate a voice they hear in such a way that they could undoubtedly impersonate that person to a third party, even laypeople are now able to generate so-called deepfakes that do exactly that. Furthermore, cloud computing gives companies the ability to aggregate vast amounts of user data and analyze it collectively. In a matter of seconds, this can be used to generate forecasts of customer behavior on a scale that the local baker could not even dream of.

It can therefore be concluded that the potential risks for users do not stem from the data per se, but from novel analysis and prediction models, as well as the technical platforms that enable their execution.[23]

What is also astonishing, however, is that a contrary development has taken place in parallel with some data. In the world of data protection, the term "data protection paradox" has become established for this phenomenon. It describes the phenomenon

[23] Doss 2020 , pp. 35 - 36

that, despite their (sometimes blatant) concerns about privacy in the digital space, users are simultaneously willing to share personal data online that they would be extremely reluctant to disclose in analog life.[24]

This particularly affects information about medical conditions, sexual preferences, or political opinions. Direct comparative studies – which meet a fair amount of scientific standards – between information openness in digital versus analog environments are difficult to conduct. However, researchers at *the Brookings Institution* have uncovered interesting correlations in this thematic by analyzing Google search statistics or sales data from online retailers, which at least create room for speculative theories.[25]

This finding, which at first glance seems counterintuitive, can be attributed to the fact that many people feel more anonymous online. Privacy in digital life is much more abstract than its counterpart in real life. While in analog life we often find ourselves wondering whether we would feel comfortable if the person we are looking in the eye were to receive certain information about us, we often do not care whether a web server stores the same information about us. Even if a natural person analyzes this information, we will most likely never come into direct contact with them.[26]

Finally, on the subject of private data, it should be noted that regulatory authorities have also become aware of the inconsistency of regulating data that is considered public in analog life in digital life. This is because the GDPR now applies to the local baker who keeps manual lists of his customers' purchasing behavior, just as it does to the internationally active online mail order company. Only human memory and the thoughts that arise from it remain free, of course, which means that the right to be forgotten reaches its limits here.

[24] Gerber, Gerber und Volkamer 2018 , p. 226
[25] Wittes und Liu 2015 , pp. 11–20
[26] Doss 2020 , p. 83

0.4 Data quality

The quality of data is of paramount importance for its business usability. Alongside data volume, this is the second major factor that determines the value of a data set, which will be discussed in more detail later.

Traditionally, data quality is divided into three dimensions: completeness, correctness, and consistency.[27]

- In the case of structured data, **completeness** means that as many possible attributes of a data record are defined. Using the example of a data table, this means that as many of the fields created as possible are filled in. Furthermore, completeness implies the creation of sufficient attributes to do justice to the complexity of the analysis task. Even with unstructured data, it is assumed that the data collection contains enough data points to adequately solve the analysis task. A well-known practical example is mandatory fields in forms, which serve to increase completeness. However, there are also exceptions where missing attributes can be an indication of a relevant case in terms of the analysis.
- **Correctness** is probably the easiest quality dimension to explain, but the most difficult to implement. Factually correct data is essential in order to derive correct analysis results. Validation rules, such as the requirement that an email address must contain the @-symbol and a top-level domain, ensure that obviously incorrect information cannot be entered in the first place. If data is primarily generated through manual entries, quality training for the persons entering the data is recommended.[28]
- **Consistency** ensures linkability with other data sources and the use of analysis algorithms. This refers to formal correctness, i.e., storage in the correct (predefined) data types.[29]

In the specialist literature, the definitions of correctness and consistency often overlap, and in some cases these two aspects are combined, which is also a perfectly legitimate approach. However, the author advocates a distinction between intrinsic (content-related) and extrinsic (form-related) correctness. Technical literature also

[27] Seiter 2023 , p. 25
[28] Schlageter und Stucky 1983 , p. 287
[29] Kähler 1990 , p. 144

contains various synonyms for the term consistency, including uniformity and accessibility, to name just a couple.

Another quality criterion that is often underestimated and consequently ignored is timeliness. The importance of this criterion depends heavily on the respective application, but increases with the aging process of our digital age. In order to incorporate this aspect, the principle of mathematical-biological decay is sometimes implemented. In this process, numerical values are multiplied by a natural exponential function.

$$Q(w, A) := e^{-Verfall(A) \times Alter(w,A)}$$

w … Attribute value in a data matrix Q

A … Attribute of the data matrix Q

Decay(A) … Decay rate of attribute A

Age(w,A) … Age of the attribute value w

For metrics with unlimited lifetime, decay assumes the value 0, whereby the exponential function becomes a constant with a value of 1.[30]

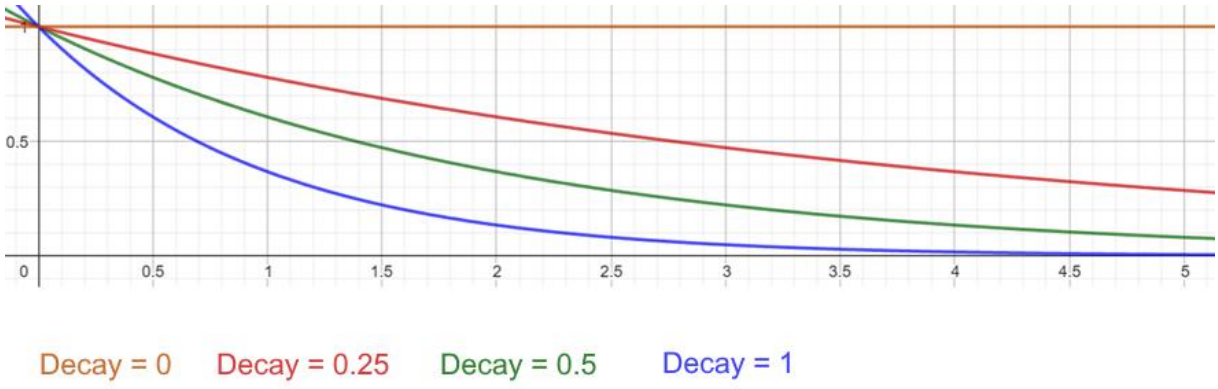

Fig.2: Decay functions. Own graphic.

[30] Seiter 2023 , pp. 62–63

Three basic strategies can be named to improve data quality in terms of completeness:

- **Follow-up collection:** To increase completeness, research efforts are initiated to obtain the correct data, at least for particularly relevant attributes. This strategy often involves a high expenditure of resources, as renewed field research is necessary.
- **Data exclusion:** Instances with low informative value (due to many missing attributes) are removed from the data set. This is the simplest strategy, but it reduces the amount of data and thus also the informative value of analyses based on it.
- **Imputation:** This involves attempting to derive a probable value for the missing value from other attributes in the data set. The usefulness of this approach depends largely on the context of the data set. For homogeneous data collections, this can be a valid strategy that is also cost-effective, but if data sets are artificially homogenized, this leads to incorrect analysis results.[31]

[31] Aggarwal 2015 , p. 35

1. Data collection

1.1 Methods for collecting user data

The first step in any data analysis is the procurement of raw data. Now that we have clarified the basic terminology in the world of data, we will turn our attention to this fundamental step.

Fundamentally, a distinction can be made between primary and secondary collection in computer-assisted data acquisition. In primary collection, the operator of an online service acts as the data collector. In secondary collection, an outside entity collects the data, extracting it from the service with or without the knowledge of the service operator. Secondary collection always requires that primary collection by the operator has taken place beforehand.

In primary data collection, a distinction is also made between supervised and unsupervised collection. In supervised collection, the user is aware that all of their actions serve to generate data. In unsupervised collection, data generation is a by-product of actions with other basic intentions. The terms supervised and unsupervised stem from the fact that unsupervised collection takes place almost exclusively in an automated manner, while there are some supervised collection methods in which the data collector accompanies and monitors the data provider throughout the entire collection process. It should also be noted that supervised surveys are often combined with unsupervised survey methods, but in these cases, they are referred to exclusively as supervised surveys, as the person being analyzed is aware of the analysis.[32]

[32] Höchstötter 2009 , p. 176

1.1.1 Primarily supervised (active)

Questionnaires: As with their analog counterparts, digital questionnaires ask people to answer a series of quantitative or qualitative questions. The advantage of this method lies in the nature of the direct questions, which allow extremely specific data to be obtained.

Computer lab experiments: The aim here is to find out how users behave when using digital devices. The test subjects are given a variety of tasks to complete in a computer lab using the technologies available to them. They are under constant observation by the experimenters.[33] Since this method does not usually generate any personal data, it is not the focus of this work, but it should nevertheless be mentioned as a general method for understanding user behavior.

1.1.2 Primarily unsupervised (passive)

Log files: Web servers and other computer programs usually log the actions they perform, as well as the information provided by clients, in so-called log files.[34] Uniform standards have been developed for the generation of such log files. Among the best known and most frequently used are the *"NCSA Common Logfile Format"* and the *"W3C Extended Log Format*." In this way, it is possible to collect information such as when a user accessed a particular subpage, how they got there (referral), or which browser they used to access it.[35] The data is usually stored chronologically in text files. The data in such log files can be used for statistical evaluations of the use of an online service.

User profile data: While log files are strictly limited to server actions by default and are stored in standardized formats, user profile data is an umbrella term for all data that is logged when using an online service and directly assigned to a specific user. The data is usually stored in databases where the collected data is linked to user profiles. Typically, this includes information about items viewed or purchased, transactions completed, or reviews submitted.

---

[33] Höchstötter 2009 , p. 177
[34] Hegewald 2017 , p. 22
[35] Höchstötter 2009 , p. 180

Interaction analysis: Describes all methods that automatically test how a user interacts with an online service. One of the most common tactics is to use tracking pixels to measure the paths taken with the mouse on a website. Tracking pixels are small, transparent graphics that trigger a call as soon as the mouse pointer is moved over them. On the one hand, interaction analyses are used in a similar way to computer lab experiments to find out how intuitive digital products are for end users. On the other hand, it is conceivable to draw conclusions about the user's demographic data (such as age) from key figures such as click speed and the like.

1.1.3 Secondary

API: The acronym stands for *"application programming interface"*. In a nutshell, APIs are standardized protocols that enable structured data exchange between different software applications. Without APIs, the diversity and interoperability of modern IT infrastructures would be virtually inconceivable. For the most part, they serve to enable efficient communication between programs and enable functions such as access to cloud services or the integration of external systems. In the context of secondary data collection, however, they also allow access to existing data sets from third-party providers by enabling automated queries to external platforms or databases. This facilitates the aggregation and analysis of information from various sources.[36] This form of secondary data collection is illustrated by the example of *Meta* and *Cambridge Analytica*, which has now also gained notoriety in popular media. Meta's *Facebook Open API* enabled *Cambridge Analytica* to read information from users' Facebook profiles until May 2015.[37]

Web scraping: This includes all methods and technologies for the (automatic) extraction of content from an online service, with the exception of APIs. While APIs are made available to third parties by online service operators for the purpose of collecting data, technologies that fall into this category are often used without the operator's knowledge. Web scraping is a cost-effective and simple method for collecting large

[36] Goodwin 2024
[37] Isaak und Hanna 2018 , pp. 56–59

data sets (big data) whose information has already been made public by others. However, the use of this technology often raises legal and ethical concerns, primarily in the context of copyright law.[38]

Open data: Research institutions, non-profit organizations, and government bodies are increasingly publishing raw data for free distribution. These data sets, known as "open data," are mostly suitable for statistical analysis purposes. In contrast to APIs or web scraping, the data available in this way is much more selective, as distributors usually publish the data with a clear intention.[39]

Third-party cookies: Cookies are pieces of information that are stored locally in the client's browser by an online service. They were originally designed to allow information and settings that a user has entered in a web service to be stored for the duration of a single session without the user having to log in. For example, a website can remember which language the user has selected, whether display options have been changed, or which items have been placed in a shopping cart. The functions behind this basic idea are now referred to as necessary or functional, and their cookies as *"first-party cookies."* It is also important to note that cookies can only be read and edited by the domain from which they were set. First-party cookies are set by the operator of the website visited.

With the commercialization of the internet and the increasing spread of digital advertising, a new application field for cookies has emerged. Website operators provide advertising space to third parties who integrate their own content in exchange for financial compensation. These third-party providers are able to set so-called third-party cookies via their embedded content. Unlike first-party cookies, which originate from the domain visited, third-party cookies are managed by external entities. They enable user activities to be tracked across different websites by storing, among other things, the domains visited or interactions. However, this requires the third-party provider to have advertising placements on several independent websites. Only then can it set its cookies on different pages and recognize users. This may sound like a major obstacle, but in fact less than 1% of advertising networks have access to about 75% of websites with advertising placements.[40] Combining this data creates a detailed usage profile that is used in particular for targeted advertising. On the one hand, advertising networks can

[38] Krotov, Johnson und Silva 2020 , pp. 555–557
[39] Murray-Rus 2008 , pp. 1–2
[40] Cahn, et al. 2016 , p. 891

analyze the collected surfing behavior to provide individualized ads in real time in order to optimize the effectiveness of their advertising campaigns. On the other hand, the user data obtained in this way can also be resold.[41]

1.1.4 Overview

Below is a schematic illustration of the methods mentioned above as a suggestion for easy assignment:

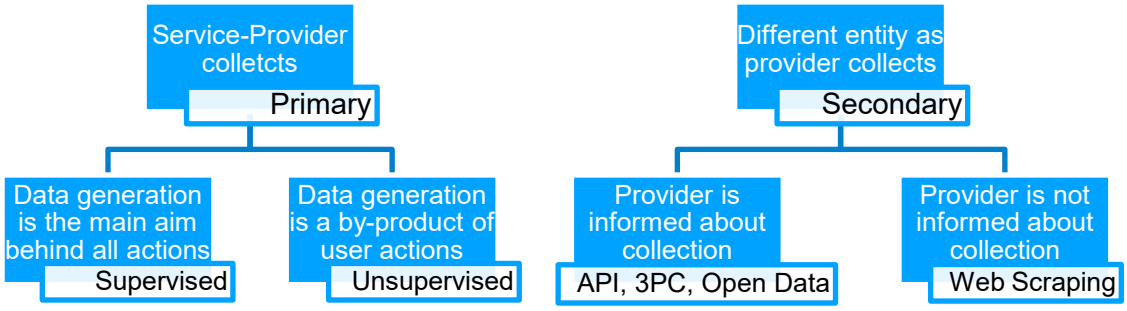

Fig.3: Schematic classification of data collection methods. Own graphic.

[41] Moryl 2024

1.2 Comparison of the strengths and weaknesses of the methods

Supervised primary collection is mostly suitable for specific questions, as the required data can be explicitly requested through direct interaction with the data providers. On the other hand, the costs are high, as these collections are rarely standardized and the necessary forms or experiments must first be designed by the researchers.

Another difficulty is the generation of large amounts of data, as these are not produced passively as a by-product of other activities, as is the case with other unsupervised methods, but are actively generated by the users. Data providers must therefore be incentivized to participate, which represents a further cost factor.

In addition, the psychological Hawthorne effect (observer effect) must be taken into account when evaluating the results. This effect states that people who know they are being observed or that their input is being analyzed change their behavior or present their opinions differently in order to better match their self-image and the expectations of a social group.[42] For example, when it comes to social or financial issues, there is often a need to present oneself in a better light, which is why the truth is distorted.

Furthermore, this collection method allows for the deliberate provision of false information. For example, participants who do not want to disclose their address or telephone number can simply enter false information.

[42] Anteby und Khurana o.J.

On the other hand, unsupervised methods are extremely cost-effective and reliable due to their almost exclusively automated collection. While any lay user can enter a false address in a form field, it is much more difficult for them to change their IP address, which also provides information about their location.

Due to the inability to opt out of data collection in log files or transaction data, large amounts of data can be collected on virtually every user.

The biggest disadvantage arises from the limited context. While survey respondents can explain their actions rationally, researchers are faced with the problem of having to interpret the actions correctly.

Secondary collection methods allow large amounts of data to be obtained in a very short time with less resource expenditure. However, the reliability of the data depends heavily on the trustworthiness of the source, and there are additional legal concerns, such as copyright issues.

1.3 Legal restrictions on the collection of user data

Following numerous scandals involving data leaks and widespread public debate on the ethical acceptability of collecting sensitive private data, more and more regulatory bodies around the world are recognizing the need to define strict requirements to protect our loose understanding of privacy.

The approaches differ greatly from one jurisdiction to another. It is not the intention of this work to provide a detailed guide to dealing with the various regulatory frameworks. Rather, it aims to give readers an insight into the different approaches so that they can gain a basic understanding of the wide range of strategies available.

1.3.1 USA

As the country of origin of many technical achievements, this judiciary deserves to be examined first, especially since many of the products we use every day are developed with this data protection perspective in mind.

As an introduction, it is worth noting that US law in its original form did not recognize the concept of privacy. The term did not appear in any law published in the Declaration of Independence. Rather, privacy was indirectly guaranteed by property rights. No one was allowed to observe what someone else was doing within their own four walls, as trespassing on someone else's property was prohibited. No uninvolved third party was allowed to read the contents of a letter from strangers, as this would have constituted theft of property in the form of a letter. The notion that privacy can be infringed upon even without crossing physical boundaries was only incorporated into law books later on.[43]

However, these piecemeal additions have always been very narrowly formulated. For example, the *Health Insurance Portability and Accountability Act (HIPAA)* only requires specific categories of companies or organizations (such as hospitals or insurance companies) to protect the health information disclosed to them. Companies that do not fall into one of the defined categories but still store health information, such as smartwatch providers, are not required to comply with the strict regulations.[44]

[43] Doss 2020 , pp. 42–49
[44] Doss 2020 , pp. 24–25

Another specific feature of US law is its federalist approach. Internet privacy rights differ from state to state. In California, the American hub of the IT industry, the *California Consumer Privacy Act (CCPA)* came into force in 2020. This is a wide-ranging law that is apparently also based on the principles of the European GDPR.[45]

1.3.2 EU

To understand the European perspective on privacy, one must remember Europe's dark past of discrimination and exclusion. In particular, the data collection practices of the Nazis, who persecuted people based on their religious affiliation, political views, sexual orientation, or similar personal characteristics, serve as a grim reminder of the abuse of personal information. In light of this, it is not surprising that such sensitive data may only be processed in extremely exceptional cases. In addition, every person residing in the EU has the right to obtain information about the data stored about them and to have it changed or deleted if necessary.

Also worth mentioning is Europe's holistic ("*one size fits all"*) approach. Data must be protected equally by every entity according to its sensitivity. Critics, however, accuse this approach of being bureaucratic overregulation that jeopardizes Europe's position in technical developments. The hotly debated *right to be forgotten* also raises ethical questions regarding freedom of expression and freedom of the press. However, these discussions are beyond the scope of this work.[46]

[45] Doss 2020 , pp. 62–63
[46] Doss 2020 , pp. 235–242

2. Data trading

2.1 Data becomes an asset

For most companies, the collection of user data begins with the intention of improving or better marketing the services and products they offer through data-driven analysis. The return on investment (ROI), i.e., the amortization of the capital invested in data collection,[47] is therefore generated by an increase in sales as a result of more competitive products or the development of new customer segments.

However, with the breakthrough of online marketing, i.e., advertising on digital platforms that focuses primarily on personalizing the advertising messages displayed, a market for data per se has emerged. Whereas the marketing industry used to develop advertising at the macro level by conducting representative market research in order to appeal to as large a share of the audience as possible, today it relies on modifications at the micro level to show individual recipients exactly the type of advertising to which they are most likely to respond.

In this context, it is not enough to have representative sample data that reflects the greatest common denominator of the target group; personally generated data from each potential consumer must also be available. The granularity of the data collected, i.e., its depth of detail and accuracy in relation to individual user preferences and behavior patterns, is essential. The finer the data is segmented, the more targeted advertising messages can be personalized and adapted in real time.[48]

For a single company, it would be a mammoth task to carry out such granular data collection alongside its actual business activities. This gap in the market has given rise to companies that have made it their dedicated task to collect data from a wide variety of sources and resell it to customers. These companies are known as "data brokers."

In the financial world, the term "broker" is the Anglo-American term for trading companies that act as business intermediaries between buyers and sellers. These are mostly assets such as securities or commodities.[49]

[47] SRH Fernhochschule GmbH o. J.
[48] Doss 2020 , pp. 82–86
[49] Heldt 2018

The keyword "asset" brings us to the rather rhetorical key question of this paper: In general, an asset is any item of property that can be sold on various markets. More specifically, it is also a technical term in accounting that describes an item on the assets side with certain characteristics. These characteristics include that this resource *"was created by a previous business activity, is controlled by the company, and is expected to provide future benefits to the company."* In addition, unlike the partially related term "property," it is not assumed that it can be sold individually.[50]

These characteristics clearly apply to the resource "data" as defined and categorized in the zero chapter of this work. They are generated by the business activities listed in 1.1, are intangible assets controlled by the owning company, and the insights gained through analysis result in a competitive advantage that can have a monetary impact. Furthermore, although not on the scale of a single data point, they can be sold via data brokers. Clarifying the details of this possibility is a focus of this paper and will be discussed below.

[50] Hinz und Weiß 2020

2.2 Origin of the traded data

Data is always collected using the methods described in Chapter 1.1. The data offered for sale on data broker portals has therefore been collected in the same way. The special feature of data brokers is that they mainly use secondary forms of collection. They draw on a variety of sources, compile the information obtained there, and then offer it for resale or use it to fulfill specific orders. In summary, they act as a central information agency for companies with an information deficit.

Specifically, three common categories of sources can be distinguished:

2.2.1 Acquisition from external companies

As briefly mentioned in the introduction to this chapter, a considerable number of companies, primarily those that operate purely online, now collect customer data to improve their own product or service range. The sale of general data, the disposal of which does not entail any competitive disadvantage for the company itself, represents a lucrative sideline. In professional circles, this is often referred to as *data commercialization*.

2.2.2 Public sources

Government institutions in particular publish a great deal of personal information in the spirit of democratic transparency. This includes company registers and trade registers, the land register, and the insolvency register. However, private entities such as *Herold Business Data GmbH* also publish personal data in their publicly accessible telephone directory.[51]

2.2.3 Own surveys

The only primary collection method mentioned here is probably also the least common, reserved only for larger data brokers with the necessary resources. However, this source is becoming increasingly important for those data brokers, and so, for example, the data broker Epsilon conducted the largest customer-specific survey in North America with 20 million participating households.[52]

These are methods by which data can be collected from users in principle. However, they are not the only way in which data about users can be generated.

---

[51] Neally 2019 , pp. 30–46
[52] Epsilon Data Management LLC 2025

2.2.4 Inferential data generation

An inference is a conclusion formed from a logical system of rules. Inferential data generation attempts to derive further attributes of a subject under consideration from known, collected data through statistical analysis. For example, a consumer who frequently purchases men's grooming products could be inferred to be a male individual. In the case of a buyer of a certain video game title, the age could be estimated to be under 25.[53]

Since this is not an actual collection of user data, but rather the result of an analysis of such data, this possibility was not mentioned in Chapter 1.1. Nevertheless, it is of immense importance in the context of considering the offerings of data brokers, as the core business of many leading brokers is now the provision of such insights about people.

This type of data is certainly the most interesting of those available to many business administrators. Nevertheless, it should be used with caution. Inferential data often attempts to reflect information about age, gender, health, sexual orientation, political views, hobbies, and the like. In fact, however, there is no algorithm that can accurately make such assignments, and the use of certain assignments also raises ethical concerns.[54] This is especially true when the algorithms are based on common prejudices. Reviews of the demographic user data generated by predictive analyses from a market-leading data broker have shown that these are only 50% accurate.[55] However, it can be assumed that accuracy will increase with increasing datafication and longer histories.

[53] Doss 2020 , pp. 68–69
[54] Venkatadri, et al. 2019 , pp. 1920–1930
[55] Ruschemeier 2023 , p. 30

2.3 Use of traded data

The purchase of data, or rather the information resulting from it, is always preceded by an information deficit on the part of the buyer, which can be mitigated or resolved by the seller. The types of information deficits are diverse, but traditionally, providers limit themselves to two main categories: finance and marketing.

2.3.1 Finance

Providers operating in this area typically collect information on the identity of individuals (full name, address, date of birth, gender, social security number, etc.), purchasing behavior (which products are purchased in what quantities on which platforms), and assets.

On the one hand, this information is used to offer identification tools for verifying the identity of customers and checking the data provided in order to protect against fraud. On the other hand, it is used to derive forecasts for assessing creditworthiness.[56]

2.3.2 Marketing

If, for example, additional personal data such as that from social media profiles is added to the aforementioned data on purchasing behavior, conclusions can be drawn about interests and hobbies. Based on this, advertising profiles can be generated, which can be used to deliver personalized advertising to individuals. These advertising profiles are usually divided into meaningful segments, and the customer can choose a segment that best matches their target group.

However, a further distinction must be made here as to whether the collected user data is actually sold in raw form and the buyer places the advertising themselves, or whether the buyer commissions the data broker to place advertising in this customer segment. Meanwhile, the largest data brokers often no longer act as pure data traders, but have transformed themselves into advertising agencies that serve the entire value chain of an advertising campaign.

In the case of an actual data exchange, as defined in the original meaning of a data broker, the data broker provides the buyer with an extensive data list, which usually contains unique identifiers (full name, address, etc.) as well as advertising-relevant data such as personal interests, level of education, or assets.[57]

---

[56] Beckmann 2024
[57] Doss 2020 , 85–88

The American data broker *Epsilon Data Management LLC* gives the following example of a user profile:

Demographics & Lifestyle	34 years old, married, female, two children
Health & Wellness	Active in yoga, Pilates, and running. Frequently tries out different organic diets. Suffers from allergies and high blood pressure and does not have supplemental dental insurance.
Finances & Assets	Homeowner, household income between $125k–$150k, net household assets between $250k–$500k, active investor, two vehicles owned by household: Audi (2015) and Chrysler (2016)
Retail-level purchase data	Spends over $5,000 at United Airlines, frequent shopper at Brooks Brothers and Marshalls, frequently purchases vitamins from Tide, Crest, and Nature's Way
Purchase data from Abacus Cooperative [Note: this is a transaction database initiated by Epsilon and operated by over 3,000 participating companies][58]	Frequent buyer of mid-priced women's clothing, spent over $1,000 in 15 transactions last year
Preferences & intentions	Looking for a new luxury car, likely wants to buy organic food, switch phone providers, take out multiple insurance policies, go on a family vacation, and donate to children's projects.
Attitudes & preferences	Tech early adopter, frequently shops at Amazon and other online retailers

Table2: Example of a data broker user profile. Own table with information taken from Epsilon Data Management LLC 2025

[58] Epsilon Data Management LLC 2025

According to its own information, Epsilon Data Management LLC has extensive data lists on 250 million US citizens, covering almost the entire consumer population of the US. Each of these data records is said to be linked to a verified real name.[59]

[59] Epsilon Data Management LLC 2025

2.4 Leading data brokers and their market position

According to a report by *Maximize Market Research*, Acxiom (US), Experian (Ireland), Equifax (US), CoreLogic (US), and TransUnion (US) are the data brokers with the largest market share. It is estimated that there are up to 5,000 active data brokers globally. Overall, the market volume of the industry was estimated at USD 270.4 billion in 2024, with a compound annual growth rate (CAGR) of 7.25% forecast through 2032, bringing the industry's market volume to USD 473.35 billion that year.

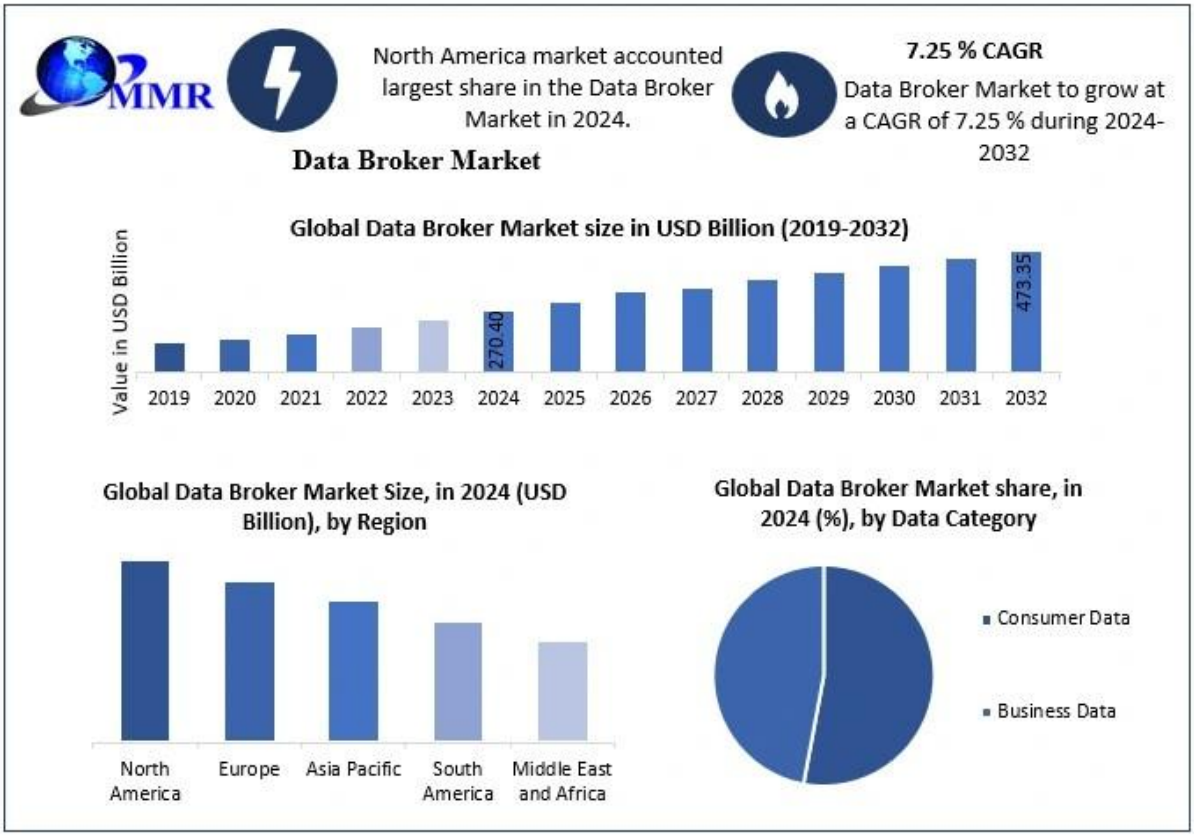

Fig.4: Economic indicators for the data broker industry. Maximize Market Research 2025, retrieved from https://www.maximizemarketresearch.com/wp-content/uploads/2020/04/Data-Broker-Market-1.webp

By way of comparison, the market volume of the mobile phone sector, the primary device for generating user data, was estimated at USD 649.13 billion for 2024. According to forecasts, it is expected to grow at an average annual rate of 6.8% to around USD 1.1 trillion by 2032.[60] This means that the volume of the mobile phone market is

[60] SkyQuest Technology Group 2025

more than double that of data trading, but has an annual growth rate that is 0.45 percentage points lower until 2032. The telecommunications industry has a volume of USD 1.99 trillion and a projected CAGR of 6.5% until 2030.[61]

With a market share of 34.68%, North America is the most strongly represented geographical region, closely followed by Europe and Asia. However, it is primarily in these regions that the industry is facing legislative headwinds. In Europe, this is mainly due to the well-known *General Data Protection Regulation* (GDPR). In the United States, regulation of the industry is progressing more slowly due to the federalist approach, but individual states such as California have already made advances with the *California Consumer Privacy Act (CCPA)*, for example.

This contrasts with the growing interest among businesses in monetizing the annually increasing amounts of data beyond product improvements and in not losing knowledge about individual customers during the transition from the analog to the digital marketplace.[62]

Another report by *Research and Markets* even speaks of an estimated market volume of $389.765 billion and sees an annual growth opportunity of 7.58% until 2029.[63]

[61] Grand View Research Inc. 2024
[62] Maximize Market Research PVT. LTD. 2025
[63] Research and Markets 2024

2.5 The value of data

The value of traded data is extremely difficult to measure, as the big players in the industry in particular are increasingly focusing on subscription-based services or holistic marketing solutions. For example, the subscription price for identity verification services depends on the size of the customer, the average number of requests per billing period, and the complexity of the queries. In particularly difficult cases, the leading data brokers also offer to make paid inquiries to government institutions or conduct their own research if the person requested is not found in the existing data sets. This approach leads to an extremely variable price, which in most cases is negotiated individually with the customer. At the time of research for this work, none of the market-leading data brokers quote flat rates for their services.

Data brokers who resell complete data sets in raw form generally do not disclose their prices. This is likely due, among other things, to the fact that they have been the focus of increased investigative research in recent years. Forbes Magazine reported that *MEDbase*, a US data broker specializing in health data, offers *200* lists of sensitive health information for $79 per 1,000 data records, or less than 8 cents per person.[64]

Another important indicator is the value of such information. Here, too, the principle applies that exact values are difficult to specify due to the high level of abstraction. However, in 2012, the data broker Experian estimated the value of an email address at £84.50. This figure is based on the revenue generated by email advertising over the lifetime of the address.[65]

[64] Hill 2013
[65] Jenkings 2012

2.6 Legal restrictions on the resale of data

2.6.1 USA

Trading takes place in legally secure waters, especially in the country where most market-leading brokers are based. Even though case law is becoming increasingly restrictive, as explained in Chapter 1.3, and some providers have had to justify their business activities in *Senate hearings*[66] , US law continues to apply the principle of not minimizing data collection in general, but rather of better informing users about it or giving them the right to opt out of data collection on an individual basis.[67]

For example, the *California Consumer Privacy Act (CCPA)*, a pioneering law in American data protection, grants users the right to obtain information about the data collected, to refuse its resale to third parties, or to have the collected data deleted.[68] However, there is no general requirement for *purpose limitation* or *data minimization*[69] , two concepts that have been clearly defined in the European General Data Protection Regulation. The applicability of the law to data brokers is explicitly mentioned by the California Attorney General's Office, but it merely states that users can assert the same rights with them and that a list of all registered brokers can be found on the Attorney General's website.[70]

The mantra here is often that a balance must be found between data protection and economic efficiency. Having both is unrealistic, having only one is not desirable. This attitude is reinforced by the size of the technology industry in the US and its relevance to the economy.

2.6.2 EU

The picture is different within the European Union, where data protectionists have identified illegalities in the business practices of data brokers since the publication of the GDPR at the latest.

Specifically, sub-points b and c of the principles for the processing of personal data are often cited. These state that personal data may only *be collected "for specified, explicit and legitimate purposes and [...] not further processed in a manner that is incompatible with those purposes [...]"* and must be *"adequate, relevant and limited to*

---

[66] U.S. Senate Committee on Commerce, Science, and Transportation 2015
[67] Doss 2020 , pp. 277–283
[68] §§ 1798.105, 1798.110, 1798.120 California Consumer Privacy Act (CCPA), Regulation (California) 2018
[69] Article 5, paragraph 1 of Regulation (EU) 679/2016
[70] State of California Department of Justice Office of the Attorney General 2024

what is necessary in relation to the purposes for which they are processed [...]"[71]. Critics accuse brokers of mass data aggregation, which is the core competence of every data broker, being incompatible with this principle. The other side argues that the data collected is necessary and legitimate for advertising measures or financial assessment, for example, and that its use is clearly defined in the partners' general terms and conditions.

Another problem is that data brokers are not mentioned at all in the GDPR, which leaves room for interpretation on both sides. There are no specific requirements for trading, as the scope of application according to Article 2 simply refers to any processing. However, this also includes trading, as processing includes *"the collection, recording, organization, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction"*[72] .

The first and to date largest criminal case in the EU against a data broker itself was conducted in Poland in 2019. Among other things, the broker collected address and contact details of Polish residents from public sources, but did not inform the persons for whom no email address was available about this collection.[73] According to the broker, this decision was made because sending a postal notification would have been too costly. The broker was fined the equivalent of approximately €220,000 for this action. However, the case also gained notoriety because the data was collected using web scraping and a large part of the data came from public registers maintained by government organizations.[74]

[71] Article 5, paragraph 1 of Regulation (EU) 679/2016
[72] Article 4, Paragraph 2 of Regulation (EU) 679/2016
[73] Urząd Ochrony Danych Osobowych (UODO) 2023
[74] Doss 2020 , p. 271

3. Added value and risks

3.1 Advantages for companies

In terms of marketing, the biggest advantage for businesses is undoubtedly that potential customers with a high probability of purchasing can be addressed directly, thereby significantly reducing the capital invested in unprofitable advertising.

However, as discussed in section 2.3, the services offered by data brokers, and thus the advantages for businesses, go far beyond marketing.

With the help of detailed credit assessments for virtually every customer, business risk can be drastically reduced. Services such as identity verification enable the secure execution of capital-intensive business transactions without direct customer contact. And by constantly checking the user information stored in the data broker's comparison database, incorrect entries and inconsistencies can be detected and corrected.

Last but not least, access to aggregated market data opens up strategic competitive advantages. Companies can analyze purchasing behavior, demographic trends, or regional characteristics in order to tailor their products and services to the needs of the market. This data-driven decision-making enables them to efficiently tap into new markets, optimize pricing models, or identify potential growth opportunities at an early stage.

3.2 Advantages for consumers

In the best case scenario, customized advertising means that users only see ads that are relevant to their lifestyle. If we assume, in our capitalist economic world, that (healthy) consumption is essential for the existence and success of the system, then this can also be seen as an advantage for the end customer.

However, the other data broker services mentioned also bring added value not only for the entrepreneur. As explained in 3.1, it is only through identity verification and fraud detection that transactions can take place digitally in a secure manner. Digitized individuals, who generally enjoy the freedom and flexibility afforded by the ability to conduct business transactions via their mobile phones, appreciate the fact that more and more business transactions can be completed digitally.

3.3 Disadvantages for consumers

The information predicted and stored by some data brokers can be profound and contain content that people do not want to make completely public for legitimate reasons. This includes, in particular, information on political views, religious affiliation, or sexual orientation.[75]

Furthermore, incorrect information or attributions can lead to serious decisions on the part of companies towards their customers. Especially in the case of critical services such as credit scoring, there are too few legal regulations to ensure high-quality analysis. The lack of standardization means that the weighting of individual (sometimes incomprehensible) variables introduces subjectivity into models that are touted as objective. In addition, there are no requirements regarding the timeliness of the data used and there is a lack of transparency towards customers.[76]

The possibility that private individuals could purchase such data sets in order to obtain information about people in their immediate environment is also a cause for concern. Numerous data protection experts warn of a new era of digital stalking.[77] In particular, the fact that the data broker *MEDbase 200*, already mentioned in this paper, offered de-anonymized lists of HIV-infected individuals or rape victims for sale is far more than simply questionable in this context.[78] However, according to the latest research findings, the anonymization of sold data sets is not a satisfactory solution either. An American experiment confirmed that it is possible to re-identify anonymized individuals with a success rate of 99.98% using only 15 demographic attributes.[79]

[75] Beckmann 2024
[76] Rothmann, Sterbik-Lamina und Peissl 2014 , p. 57
[77] Rostow 2017 , p. 667
[78] Hill 2013
[79] Rocher, Hendrickx und Montjoye 2019 , p. 1

Summary

Data is an asset, and the industry that trades in it is estimated to be worth hundreds of billions. Positive development with growth opportunities of around 7% per year seems assured as long as the seemingly unstoppable trend away from the analog to the digital marketplace continues.

With this transformation, the services offered by data brokers are becoming increasingly relevant for businesses. Not only for the purposes of targeted advertising, but also for purposes that are taken for granted in analog life, such as identity verification or business acumen assessments.

At the same time, legal and ethical difficulties must not be ignored. In the US, openness to technology seems to continue to prevail in legislation. The situation is different within the European Union, where the business practices of many data brokers already fall into the gray areas of applicable law (GDPR) and data protection initiatives are gaining momentum.

If data collection gets out of control and neither the actors nor their purposes are transparent anymore, the transformation into transparent humans, with all its civil and political consequences, can no longer be stopped.

It is therefore essential to promote an open dialogue about the functioning, size, and goals of this socially little-known market. Only those who are informed can make informed decisions. This applies equally to companies that rely on the services of data brokers and to the people whose data is being traded.

List of figures

List of tables

Bibliography

Aggarwal, Charu. Data Mining. The Textbook. Heidelberg: Springer, 2015.

Anteby, Michel, and Rakesh Khurana. Harvard Business School. o.J. https://www.library.hbs.edu/hc/hawthorne/09.html#nine accessed Februar 04, 2025.

Beckmann, Werner. nordvpn s.a. Februar 09, 2024. https://nordvpn.com/de/blog/was-ist-datenbroker/ accessed Februar 13, 2025.

Bodendorf, Freimut. Daten- und Wissensmanagement. (2. Aufl.). Berlin: Springer, 2006.

Cahn, Aaron, Scott Alfeld, Paul Barford, and Shan Muthukrishnan. "An Empirical Study of Web Cookies." WWW '16: Proceedings of the 25th International Conference on World Wide Web. Montréal: ACM, 2016. 891 - 901.

dataspot. GmbH. dataspot.at. n.d. https://www.dataspot.at/glossary/datenarten/ accessed Januar 13, 2025.

Doss, April Falcon. Cyber Privacy. Who Has Your Data and Why You Should Care. Dallas, 2020.

Duden (Hrsg.). duden.de. o.J. https://www.duden.de/rechtschreibung/meta_ accessed Dezember 16, 2024.

Epsilon Data Management LLC. Epsilon. 2025. https://www.epsilon.com/us/products-and-services/data, https://www.epsilon.com/abacus accessed Februar 13, 2025.

Fässler, Lukas, Barbara Scheuner, and David Sichau. "ETH Zürich." November 29, 2024. https://doc.et.ethz.ch/latest/p_java1_TH.pdf accessed Dezember 17, 2024.

Gerber, Nina, Paul Gerber, and Melanie Volkamer. "Explaining the privacy paradox: A systematic review of literature investigating privacy attitude and behavior." Computers & Security, August 2018: 226-261.

Goodwin, Michael. IBM. April 09, 2024. https://www.ibm.com/think/topics/api accessed Februar 03, 2025.

Grand View Research Inc. Dezember 2024. https://www.grandviewresearch.com/industry-analysis/global-telecom-services-market accessed Februar 13, 2025.

Hegewald, Tony. „Im Internet weiß niemand, dass du ein Hund bist“ – Personalisierung von Onlinewerbung in Deutschland. Mittweida: Hochschule Mittweida, 2017.

Heldt, Cordula. Gabler Wirtschaftslexikon. Februar 19, 2018. https://wirtschaftslexikon.gabler.de/definition/broker-27861/version-251503 accessed Februar 13, 2025.

Hill, Kashmir. Dezember 19, 2013. https://www.forbes.com/sites/kashmirhill/2013/12/19/data-broker-was-selling-lists-of-rape-alcoholism-and-erectile-dysfunction-sufferers/ accessed Februar 13, 2025.

Hinz, Michael, and Gregor Weiß. Gabler Banklexikon. April 08, 2020. https://www.gabler-banklexikon.de/definition/asset-55743/version-377253 accessed Februar 13, 2025.

Höchstötter, Nadine. "Methoden der Erhebung von Nutzerdaten und ihre Anwendung in der Suchmaschinenforschung." In Handbuch Internet-Suchmaschinen, by Dirk Lewandowski, 175 - 203. Heidelberg: AKA Verlag, 2009.

Isaak, Jim, and Mina Hanna. "User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection." Computer, August 14, 2018: 56 - 59.

Jenkings, Richard. The Drum. April 04, 2012. https://www.thedrum.com/opinion/2012/04/04/how-much-your-email-address-worth accessed Februar 13, 2025.

Kähler, Wolf-Michael. SQL. Bearbeitung relationaler Datenbanken. Wiesbaden: Vieweg+Teubner, 1990.

Kluge, Friedrich, and Elmar Seebold. Etymologisches Wörterbuch der deutschen Sprache (23. Auflage). Berlin: de Gruyter, 1999.

Kranz, Garry. ComputerWeekly.de. September 2024. https://www.computerweekly.com/de/definition/Metadaten accessed Dezember 16, 2024.

Krotov, Vlad, Leigh Johnson, and Leiser Silva. "Tutorial: Legality and Ethics of Web Scraping." Communications of the Association for Information Systems, Dezember 15, 2020: 555 - 581.

Lackes, Richard, and Markus Siepermann. Gabler Wirtschaftslexikon. Februar 19, 2018. https://wirtschaftslexikon.gabler.de/definition/datentyp-29319/version-252929 accessed Dezember 17, 2024.

Lahmer, Karl. Kernbereiche Psychologie & Philosophie. Wien: westermann, 2023.

Lange, Otto, and Gerhard Stegemann. Datenstrukturen und Speicherarten. Braunschweig: Vieweg+Teubner, 1985.

Langer, Erasmus. Programmieren in Fortran. Wien: Springer, 1993.

Liebhart, Daniel. "Die glorreichen sieben Datenarten." netzwoche, 2010: 21.

Lošek, Firtz (Hrsg.). Stowasser. Lateinisch-deutsches Schulwörterbuch. (1. Aufl. 4. Druck). München: Oldenbourg Schulbuchverlag, 2016.

Maximize Market Research PVT. LTD. Januar 2025. https://www.maximizemarketresearch.com/market-report/global-data-broker-market/55670/ accessed Februar 13, 2025.

Microsoft (Hrsg.). Microsoft Corporation. August 03, 2023. https://learn.microsoft.com/de-de/office/vba/language/reference/user-interface-help/data-type-summary accessed Dezember 17, 2024.

Moryl, Beata. Piwik PRO. Oktober 10, 2024. https://piwikpro.de/blog/was-ist-der-unterschied-zwischen-first-party-cookies-und-third-party-cookies/ accessed Februar 04, 2025.

Murray-Rus, Peter. "nature.com." Januar 18, 2008. https://www.nature.com/articles/npre.2008.1526.1 accessed Februar 03, 2025.

Neally, Daniel. "Data Brokers and Privacy: An Analysis of the Industry and How It's Regulated." Adelphia Law Journal, 2019: 30 - 46.

Pabst, Christiane M., Herbert Fussy, and Ulrike Steiner (Red.). Österreichisches Wörterbuch (43. Aufl.). Wien: Österreichischer Bundesverlag Schulbuch GmbH, 2016.

Research and Markets. Dezember 2024. https://www.researchandmarkets.com/reports/5987173/data-broker-market-forecasts accessed Februar 13, 2025.

Rocher, Luc, Julien Hendrickx, and Yves-Alexandre Montjoye. "Estimating the success of re-identifications in incomplete datasets using generative models." nature Communications, 2019.

Rostow, Theodore. "What Happens When an Acquaintance Buys Your Data?: A New Privacy Harm in the Age of Data Brokers." Yale Journal on Regulation, 2017: 667 - 707.

Rothmann, Robert, Jaro Sterbik-Lamina, and Walter Peissl. Credit Scoring in Österreich. Studie, Wien: AK Wien, 2014.

Ruschemeier, Hannah. "Data Brokers and European Digital Legislation." European Data Protection Law Review (EDPL), 2023: 27 - 38.

Schlageter, Gunter, and Wolffried Stucky. Datenbanksysteme: Konzepte und Modelle. Wiesbaden: Vieweg+Teubner, 1983.

Seiter, Mischa. Business Analytics. Wie Sie Daten für die Steuerung von Unternehmen nutzen. (3., überarbeitete Auflage). München: Franz Vahlen, 2023.

SkyQuest Technology Group. SKYQUEST. Januar 2025. https://www.skyquestt.com/report/smartphones-market accessed Februar 13, 2025.

SRH Fernhochschule GmbH. SRH Fernhochschule - The Mobile University. o. J. https://www.mobile-university.at/studium/return-on-investment/ accessed Februar 13, 2025.

State of California Department of Justice Office of the Attorney General. März 13, 2024. https://oag.ca.gov/privacy/ccpa accessed Februar 15, 2025.

U.S. Senate Committee on Commerce, Science, and Transportation. What Information Do Data Brokers Have on Consumers, and How Do They Use It? Senatsanhörung, Washington D.C.: U.S. Government Publishing Office, 2015.

Urząd Ochrony Danych Osobowych (UODO). September 20, 2023. https://uodo.gov.pl/en/553/1572 accessed Februar 15, 2025.

Venkatadri, Giridhari, Piotr Sapiezynski, Elissa Redmiles, Alan Mislove, Oana Goga, and Michelle, Gummadi, Krishna Mazurek. "Auditing Offline Data Brokers via

Facebook's Advertising Platform." WWW '19. San Francisco: Association for Computing Machinery, 2019. 1920 - 1930.

Wittes, Benjamin, and Jodie Liu. The privacy paradox: The privacy benefits of privacy threats. Washington: The Brookings Insitution, 2015.

Wuttke, Lorenz. Datasolut. Juni 27, 2024. https://datasolut.com/wiki/daten-definition/ accessed Dezember 16, 2024.

Affidavit

I, Simon Michael Wimmer, declare that I have completed this final thesis independently and have only used the sources, materials, and resources listed in the bibliography.

(Braunau am Inn, February 16, 2025)