Research Project

Impact of privacy-preserving-software-created trust on private-data-enabled business models of Big Tech

Decloedt, Thomas
thomas.decloedt@ugent.be

De Bock, Jens
jensdboc.debock@ugent.be

Voet, Ysabeau
ysabeau.voet@ugent.be

February 20, 2023

Abstract

The distribution of personal data is undeniably an important aspect in communication nowadays due to the rise of social media and the enhancements in technology. Many Privacy-Preserving Methods (PPMs) have already been established to offer protection of personal data. Consequently, businesses must adapt to this explosion of data circulation and all the privacy concerns that come with it.

A rigid structure for the classification of both Business Models (BMs) and the concept of privacy is not yet in place. This is due to the rapid evolution of communication and market technology.

An overview of PPMs is given to evaluate the technical difficulties when utilizing personal data. Furthermore, we categorize Data-Driven Business Models (DDBMs) and evaluate which impact PPMs exert on these models.

Generally speaking, the impact of applying PPMs to the models is exceptionally noticeable in gaining users' trust, which is an essential factor for a successful BM.

A trade-off between the usability of personal data and the protection remains relevant in the foreseeable future.

Keywords: Privacy, Data-Driven Business Model, Big Tech, Privacy-Preserving Methods, Trust

1 Introduction

1.1 Privacy as personal data

The purpose of this paper is to discuss different PPMs and their possible influence on current business models. *Privacy* is the central concept in this process, but it is challenging to define. Cambridge Dictionary [1] defines privacy as *someone's right to keep their personal matters and relationships secret.* However, the definition of privacy heavily depends on the environment, as it is inherently related to a certain culture. The way privacy is experienced has roots in philosophy, legality, sociology, politics, and economy [2].

In the past, the concept of privacy was related to one's financial status and wealth. Families living in poverty do not possess the same amount of personal belongings compared to the elite, which meant sharing common items and less privacy. But privacy has undergone an enormous evolution, ranging from the first privacy-shame dynamic to currently leaning more towards the importance of personal data [3]. Personal data is used as a way to describe privacy nowadays and is in essence information. This information leads to two crucial and sensitive features: money and power [3].

With this in mind, K. D. Martin and P. E. Murphy [4] propose that it could be considered as a form of currency in exchange for convenience. Since this is precisely what is required for further discussion, we consider this definition of privacy. The usage of personal data is regulated and standardized to some extent due to the introduction of the General Data Protection Regulation (GDPR) in 2018. Big Tech companies (Section 2.1) such as Microsoft and Apple, stated in response to this GDPR that privacy, defined as the protection of personal data, is a fundamental human right [5].

1.2 The importance of trust

The sharing of personal data brings numerous challenges. On the one hand, it is important to be able to protect oneself without having to be accountable for everything. On the other hand, the convenience that comes with sharing your data cannot be underestimated [3]. Moreover, by encroaching on privacy, some benefits can (more easily) be achieved, e.g., video surveillance to provide another type of safety which would not be possible if privacy regulations prevented this [3]. Creating a trustworthy environment proves to be essential for an optimal way of processing, analysing and publishing personal data. This is especially true for a few specific technological companies that are known as Big Tech. These are the largest and most influential technology compa-

nies (Alphabet, Amazon, Apple, Meta and Microsoft in particular), which are further discussed in Section 2.1. If citizens were to decide not to utilize their services any more due to trust issues, this would result in economic consequences, because their main revenue is based on the personal data of users [6].

A number of measures have to be met to ensure the user's trust [7]:

- Confidentiality
- Authenticity
- Integrity
- Accountability
- Antitrust

Confidentiality prevents the disclosure of personal data towards unauthorized parties. Users only want to share their personal data with agreed upon parties. Companies claim that certain values, privacy in this case, are highly important for them. Authenticity and integrity, which are closely related, define to which extent these values are fulfilled. In addition, accountability assures that companies act responsible. As a result, users are more inclined to put their trust into this company. These measures are exactly why PPMs are introduced, which are further discussed in Section 3.

The antitrust law is created as a regulation against the power of singular companies and effectively reducing the power of monopolies [8]. This limits the Big Tech companies in their quest for market domination and power by forcing them to split up [9]. Although there is plenty of discussing when it comes to antitrust [10, 11], this reduces the risks of privacy violations.

Lastly, despite the general belief that people attach great importance to their privacy, some sources seem to disagree. People might feel confident that they care, but in reality, they are not convinced that this is an issue [6].

1.3 Overview

Firstly, the importance of personal data for the Big Tech companies is discussed in Section 2.1. Hereafter, an overview of methods used to preserve the privacy of users is described (Section 3). Section 4 touches upon the different business models currently on the market, with a focus on the Data-Driven (DD) methodologies. Finally, the last section, Section 4.4, evaluates the relationship between these methods and business models.

2 The role of data for big tech

The current digital age offers many possibilities, which companies belonging to Big Tech, take great advantage of. Companies such as Alphabet (Google), Amazon and Apple define their place on the market by means of technological innovations. Because of their influence on present markets, they cannot be removed from today's view of society.

2.1 Big Tech's advantage

One of the big advantages Big Tech companies have, lies in the knowledge gained from personal data, which can improve customer relationships. Data analysts often use machine learning and artificial intelligence to perform the necessary investigations [9]. On the contrary, conventional markets can only estimate the customers' interests based on the current market price. This includes smaller businesses that do not reside on the internet and do not strongly depend on personal data. This market price provides only so much information, and it is often not trivial to derive the correct information. On the opposite side, the Big Tech companies are ahead with this valuable data at their disposal. They can be more accurate and direct using the aforementioned techniques, which guide the customers in their choices.

With Big Tech companies having the advantage of already accumulated data, it is challenging for small companies to compete in the long run. Newer products similar in functionality compared to the ones

of Big Tech companies can easily be suppressed by these companies. There have been proposals to level the playing field, e.g., initiating a progressive data-sharing mandate resulting in smaller companies having access to the same tools [12].

We can conclude that Big Tech is as big as it is now as a consequence of the large boost in data collection and processing. The influence of these companies is noticeable in a variety of domains, from an economic standpoint, but also from a democratic one [10, 11].

3 Privacy-preserving methods

This section clarifies methods used for privacy-preserving software. Each subsection goes over a different step in the data life cycle, except Subsection 3.5, which goes more into depth on operations that are used for the described techniques. The life cycle is covered from capturing the data (3.1) until its actual use (3.4).

The following typical abstraction is used to explain the data life cycle:

- A Data Owner (DO) is an individual whose data and ultimately privacy is preferably protected [15, 13, 14]. Sometimes called a record owner [16].
- A Data Collector (DC) is an individual, a company, a government, an organization, etc. who amasses the DOs data [16].
- A Data Publisher (DP) is comparable to DCs, but instead of collecting data, the DP publishes, for example, by sharing a data set on the web or making it accessible through an API [15, 16].
- A Data Recipient (DR) is a third-party that executes mining on data sets published by DPs [16, 17, 18].

Figure 1 illustrates this abstraction: Alice, Bob, Cathy, and Doug are DOs. In this case, the DC and DP are the same.

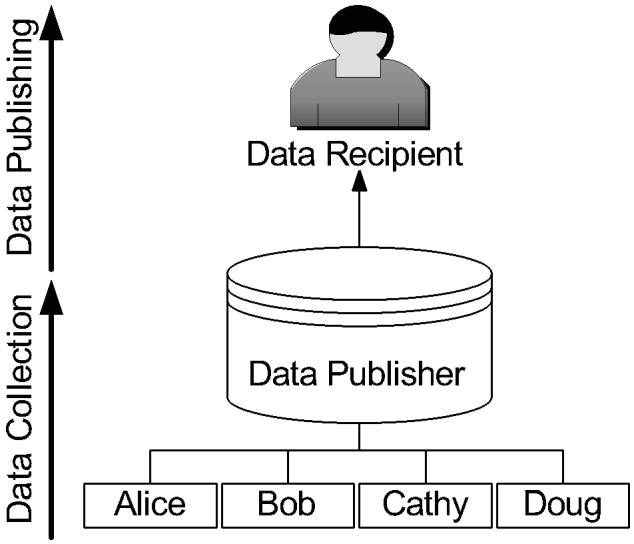

Figure 1: Data collection and data publishing [16]

Furthermore, the following common terminology is used concerning the attributes of records in data sets:

- The identifier attributes or explicit identifiers of a record contain information that can directly identify a DO, e.g. full name and home address [13, 14, 16, 17, 20, 19].
- A Quasi-ID (QID) is a set of attributes that could lead to re-identification by combining either with or without other data from outside the data set. These attributes do not explicitly identify the DO such as age and date of birth [13, 14, 16, 20, 19].
- Records have some attributes that contain personal information that does not explicitly identify the DOs nor are they QIDs. These are called Sensitive Attributes (SA), for example, sickness and religion [13, 14, 16, 20, 19].
- Any other attribute containing harmless information is considered insensitive or non-sensitive [13, 14, 16, 20, 19].

3.1 Data collection and generation

Data collection and generation is the first step of the data life cycle [19, 21]. In this phase, the data gets collected from the DO. This process can be active or passive. When the DO knowingly gives data to the DC, this is active data generation. In passive data generation, the DC gets the data from the DO without asking for it, generally, without the DO's knowledge. Several methods and techniques can provide privacy protection during the data collection phase. The DO could refuse to give data to the DC or use several tools to confirm that passively collecting the data is impossible. This method is called access restriction [13].

Sometimes data cannot be protected by access restrictions because the data is needed to perform specific actions. For example, online shopping requires users to enter their email addresses and phone numbers. Another method for protecting private data exists for these cases: falsifying the data. The data is distorted such that the original information cannot be retrieved. To this end, *Sockpuppet* can be used and tools such as *MaskMe* [22] are available. This protects the online identity of the DO by, among other things, providing a fake profile for the DO. This allows certain data to be used, without linkage to real users [13].

To ensure privacy from the collector's point of view, noise or randomization [19] can be used. Section 3.5.2 goes further into depth on noise. Making sure that raw data can never be stored by the collector is an important aspect, and the usage of noise prevents this. Otherwise, the original data could be retrieved afterwards.

Additive noise is not significantly effective for maintaining privacy, as noise-reducing techniques are used to recover an accurate approximation of the original data. Noise multiplication is a more effective method since it is harder to restore the original data. A disadvantage of both methods is their sensitivity to outliers. More noise is needed to mask outliers, which hurts the data quality.

3.2 Data storage

The next step in the data life cycle is storage. During this phase, stored data needs to be protected.

Big Tech often opts for cloud computing because the amount of data they need to store is substantial. This provides privacy benefits, such as distributed storage, but it can also come with privacy issues [21]. Two examples:

- Even though DOs could lose control over their data through outsourcing to the cloud, this can be remedied by providing a secure computing environment and secure data storage.

- Multi-tenancy, a practice where data from different customers is located on one shared physical storage device, can make it easier for malicious users to access other users' data.

There are several ways to protect privacy concerning stored data. The discussed techniques make sure that only authorized users can access their data.

3.2.1 Identity based encryption

The first encryption method that helps to preserve data privacy during storage is Identity Based Encryption (IBE) [23]. This method is similar to Public Key Encryption (PKE), but the public key depends on the identity of the user, e.g. it is based on the user's email address. In PKE, the user can encrypt text with a public key, resulting in what is known as a ciphertext. This ciphertext can only be decrypted with the user's private key. In contrast, IBE requests the user's identity using a username and password when decrypting. A private key generator gives the correct private key to the user to decrypt the ciphertext. One of the problems that occur is that this method can be time-consuming when managing a wide variety of identities or huge chunks of data. Another disadvantage is that data can only be accessed by the DO, while it can sometimes be useful to grant access to other users without them having the private key from the DO [21, 23].

3.2.2 Attribute based encryption

The second method that is discussed is Attribute Based Encryption (ABE) [13, 21]. This belongs to PKE where the private key depends on attributes of the DO. ABE is extremely similar to IBE, but different private keys, can be used to decrypt data that has been encrypted with the same public key. The public key is created based on a collection of attributes instead of an identity. Users who have these attributes are able to obtain a private key from the private key generator that can decrypt the data. The main advantage of ABE over IBE is that the same data can be viewed by multiple authorized people without them having to share the same private key. ABE, just like IBE, has a considerable amount of computational overhead and is consequently quite slow [13, 21].

3.2.3 Homomorphic encryption

A third method of protecting privacy in data storage is Homomorphic Encryption (HE) [13, 21]. HE allows modifications of and computations on encrypted data without decrypting it first, contrary to IBE and ABE. The output of a sequence of operations on a homomorphically encrypted data set is thus the same as if this sequence of operations were done on the data set, while it was not encrypted. This is beneficial as unnecessarily decrypting the data can pose privacy issues. Again, like IBE and ABE, HE requires plenty of processing power [13, 21].

3.2.4 Storage path encryption

Storage path encryption [13, 24] is a technique that protects private data stored on the cloud. The data is partitioned and stored on different physical storage media, and the path to all these parts is required to obtain the data. Data paths, rather than the data itself, are encrypted by making use of a trapdoor function to protect this data. This function is to calculate in one direction, but additional information is needed to calculate the inverse. The DO keeps track of this information, allowing them to decrypt the path to the data and retrieve it.

A big advantage of this technique is that it can retrieve the data rapidly because the data is divided into smaller blocks that have to be collected from the cloud. Encrypting the path is more efficient than encrypting the data itself because of the reduced size of the path. Consequently, the computational overhead is much smaller. By providing the additional storage path information, it is possible to share the data with other users, who may access the data by decrypting the proper storage path afterwards [13, 24].

3.3 Data publishing

Data collection is followed by data publishing. Fung et al. [16] describe Privacy-Preserving Data Publishing (PPDP) as an "approach to information sharing, while preserving individual privacy and protecting sensitive information".
In DP, DOs share their data with either an untrusted or trusted DP. For the untrusted case, methods and techniques are provided in Section 3.1. Data is collected anonymously, i.e., the identity of the DOs is undisclosed. In the case of a trusted DP, care must still be taken because the trust of the DOs in the DP is usually not transient to any DR. This means that, while an individual might entrust their data to a DP, this trust is not implicit to DRs. Thus, irrespective of the model, because publishing data may result in privacy leakage, techniques and methods have been proposed to prevent this. For the remainder of this section, trusted DPs are assumed.

There is no single way to categorize the techniques that follow. One way [16] is to present the techniques based on the attack models they counter. The techniques are then referred to as privacy models. Another way [20] is by the type of protection offered by these Privacy-Preserving (PP) techniques, which are then alternatively called privacy risk measures. Although these two differ in perspective, they are quite related. This section lists the most important methods and then explain what types of attacks they protect against.

Before going into depth, some more clarification is presented. This survey [20] lists three common threats:

1. singling out
2. linkability
3. inference

as well as four types of disclosure risks:

1. identity disclosure
2. attribute disclosure
3. inferential disclosure
4. membership disclosure

Section 3.5 about data modification explains how the anonymization approaches are realized by digging into the operations used. Table 1 summarizes many PPDP techniques.

Finally, an attacker or adversary is an individual or an organization that seeks to violate the privacy of one or multiple DOs, called the victims [18].

3.3.1 k-Anonymity

k-Anonymity is a type of privacy protection that can be used during the data publishing phase of the data life cycle. A data set satisfies k-anonymity if and only if for each set of QIDs there are at least k unique records [17]. Groups are at least of size k, thus the probability of linking a victim to a record is no more than $1/k$, assuming the QID of the victim is known to the attacker and each record in the data set represents a unique individual [16]. If this would not be the case, multiple records would correspond to one individual and the probability of linking would be higher, since k records represent less than k individuals.

Record linkage attacks A linkage attack consists of an attacker being able to link a record owner to another published record, attribute or table. The assumption is that the attacker has some QID of the victim.

A record linkage attack can arise when the published data can be narrowed down to a small subset, called a group: records with matching QIDs. If the attacker has extra information, the victim's record could be identified from the group [16].

Similar techniques Improvements on k-anonymity exist, for example (X, Y)-anonymity [16] which abstracts the concept of k-anonymity. Each subset X, not just the QIDs, is linked to at least k distinct records in the data set, with Y being the sensitive attribute.

MultiRelational k-anonymity [16] extends the concept to databases consisting of multiple tables.

3.3.2 l-Diversity

l-Diversity is an improvement on k-anonymity. Data is l-diverse if there are at least l "well-represented" values that exist for sensitive attributes. "Well-represented" can have different meanings that illustrate different types of l-diversity. For example, one of the interpretations is that there should be at least l different values in the data [19, 26].

Attribute linkage attacks l-Diversity protects against both record and attribute linkage attacks. In these attacks, the attacker may infer information from a group, as defined in Section 3.3.1.

Thus, a group has a set of sensitive values associated with it: all the sensitive attribute values from the records contained in the group. An attribute linkage attack occurs when an attacker can infer some information about the victim from this set of sensitive values. Moreover, this type of attack may still be possible if not protected against, even if record linkage attacks are prevented [16].

Similar techniques An improvement on l-diversity is (c, l)-diversity. This technique has the same conditions on the data as l-diversity, but also demands that the most common values do not appear too much and the least common values are still sufficiently represented [16].

3.3.3 ϵ-Differential privacy

The next method to ensure privacy in the publishing phase is ϵ-differential privacy. Differential privacy is a technique that adds randomness to the data. The

Privacy model	Sanitization methods	Record linkage	Attribute linkage	Table linkage	Probabilistic attack	References
k-anonymity	Generalization, Suppression	x				[16],[17],[19]
MultiRelational k-anonymity	Generalization, Suppression	x				[16][25]
l-diversity	Generalization, Suppression	x	x			[16],[19],[26]
(X,Y)-anonymity	Generalization, Suppression	x				[16],[25]
(c,l)-diversity	Generalization, Suppression	x	x			[16],[19]
t-closeness	Generalization, Suppression		x			[16],[19],[27]
ϵ-differential privacy	Randomization			x	x	[16],[19],[27]

Table 1: PPDP techniques

data is still accurate enough to make overall analyses, but some information may have been adjusted. Hereby ϵ is the privacy parameter that regulates the trade-off between privacy and accuracy. For every entry in the data, the method decides if the value is changed. As a consequence, there is no certainty that the values are correct and privacy is preserved [19, 27].

Table linkage attacks ϵ-Differential privacy protects against table linkage attacks. These types of linkage attacks do not assume that a victim's record is present in the published data set. Contrary to record and attribute linkage, this type of attack does not have the intent to infer some information. Instead, an attacker tries to infer the presence or absence of a victim in a data set [16].

3.3.4 t-Closeness

The last method to provide privacy in data publishing that is discussed is t-closeness. In t-closeness, the difference between the distribution of a SA in a group and the distribution of the SA in the database must be smaller than a certain value t [26, 27].

Probabilistic attack t-Closeness protects against probabilistic attacks [16]. This associated family of privacy models differs from linkage, and some mathematics are needed for clarification. The attacker aims to improve their probabilistic belief about a sensitive attribute of the victim by consulting the data set. The attacker's belief is given by equation 1. The models combat this by adhering to the uninformative principle. It aims to reduce the difference between the prior and posterior, see equation 2.

Bayes' theorem states:

$$P(C|\mathbf{x}) = \frac{P(C)p(\mathbf{x}|C)}{p(\mathbf{x})} \tag{1}$$

or using terminology:

$$posterior = \frac{prior \cdot likelihood}{evidence} \tag{2}$$

The prior is the probability distribution of a random variable before evidence is taken into account. An uninformative prior has minimal influence on inference, thus making the attacker's inference harder. This explains the principle: adherence to it makes it harder for the attacker to change (i.e., improve) their belief [16].

3.4 Data mining

Data Mining (DM) is the process of extracting information from a data set. Since DM is a buzzword, hereafter follows a delineation of the extent to which the topic is handled. From the multitude of ways DM may be interpreted, the focus of this paper is on:

- statistics that can be done on a data set

- the training of Machine Learning (ML) models on a data set

The output of a ML model, i.e., a classifier or regressor, must also be taken into consideration when privacy preservation is important. The output of a model could be offered to the public, another organization and other DRs. Attackers could try to infer sensitive information from a model. In the following sections, techniques are presented to protect against nefarious users. This protection can be achieved at the data or application level [19].

3.4.1 Data level

Association rule hiding Association Rule Hiding (ARH) [19, 28] is a technique that protects privacy in association rule data mining. The latter is a method to identify patterns and associations in data, so rules can be formed to predict implications within the data set. The rules may expose private information. The main idea of ARH is that rules that expose private information are not used, only association rules that do not disclose private information are mined. There are multiple approaches to apply ARH, such as heuristic-based ARH, which can be divided into a perturbation-based and a blocking-based method, reconstruction-based ARH and metaheuristic-based ARH. These approaches are not further discussed here, but L. Zhang, W. Wang, and Y. Zhang [28] discusses them in more detail.

Downgrading classifier effectiveness Classification [19, 29] is a method that organizes data points into different groups based on their features. It links data within a group, so it can assign the right group for the specific data. Decision trees, clustering algorithms, Bayesian networks, and association rule mining are all processes that can be used for classifying data. If users can discover in which group some data is classified and what the needed features are to be classified in this group, private data features may be found by nefarious users. The effectiveness of the classifier can be downgraded, making it less likely that data features can be retrieved by users.

3.4.2 Application level

Query auditing When queries can be directly executed on the original dataset, there are some constraints so that users can only query aggregate data and not personal data. There may still be some situations in which certain queries or a certain combination of queries make it possible to discover private personal data. Query auditing [19] is a method to avoid such situations. Some queries are denied in query auditing, to protect the private data that would be returned by those queries.

Query inference control Query inference control [19] is also an approach to preserve privacy when queries are executed on a dataset. This technique tries to solve the same problem as query auditing, but instead of prohibiting (combinations of) queries, the output of the query is distorted, e.g. by adding noise, so the original private data cannot be retrieved by the queries.

3.5 Data modification

Many techniques or operations exist to modify data sets, in the context of this paper their point is to preserve the privacy of DOs. These operations can enable de-identification, satisfying certain PP measures discussed in Section 3.3, while keeping a data set valuable to a DR. It is this that enables data modification to foster trust between DOs and DRs, e.g., Big Tech companies [18].

Through data modification:

- The collection of data from DOs can be done in a PP way, as discussed in Section 3.1.
- The publishing of sensitive data can be done in a PP way. For example, a hospital acting as a DP could share the data with a DR for data mining purposes. While guaranteeing a certain level of privacy to the DOs through the use of PP measures discussed in Section 3.3.

A taxonomy is given with a few examples for each category, it is largely based on a survey [20]. An enumeration of existing operations, even if non-exhaustive,

is out of the scope of this paper. Many techniques are listed in table 2 and a few are explained.

When these techniques are used in the context of Section 3.3, they are called sanitization methods [19] or anonymization operations [16].

3.5.1 Non-perturbative

Despite the name of this section, this category of techniques does not modify the data. Instead, the data set's level of detail is reduced or information is partially suppressed, thus lowering the amount of information [20].

Global recoding Global recoding, also called (full-domain) generalization [16, 20], changes values to be less accurate. Numerical values can be mapped to intervals, and the same concept can be applied to categorical values as well. Though for categorical values, this is less straightforward: a topology of all possible values for the attribute has to be constructed. More specifically, this topology needs to be a tree to have a notion of what it means for a value to be more or less general. A parent node is then more general than its child nodes, take hobbies for example:

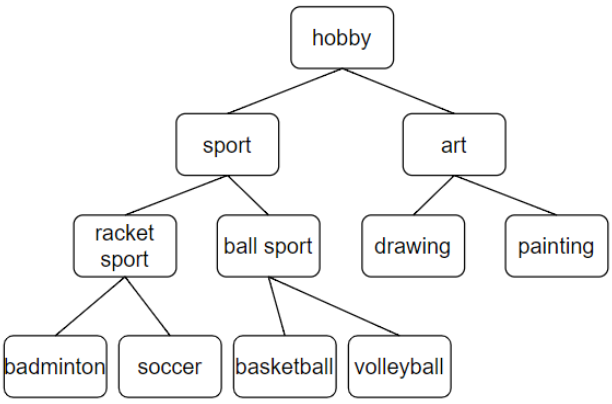

Figure 2: Classification of hobbies

In essence, this technique combines values for a specific attribute into coarser categories. This survey [16] enumerates more generalization schemes: full-domain, subtree, sibling, cell, and multidimensional.

Suppression Suppression or blocking consists of turning attribute values into unknowns (NaNs) or special characters ('*', '?', etc.) [20, 28]. There are multiple suppression schemes as well: tuple or record suppression, value suppression, cell or local suppression, and attribute suppression. Fung et al. [16] and Lu et al. [20] discuss these in more detail.

3.5.2 Perturbative

Perturbation-based methods are widely used heuristic techniques [19]. For some values of certain attributes, zeros and ones are flipped with the intent of lowering the support of sensitive rules (see Section 3.4.1) in a way that the utility of the data set remains equal to a specific maximum value [19]. Data is distorted; in the context of DP, this is of course done before release [20].

Noise or randomization In randomization, data is distorted by adding random noise to each element of the data set before sending it to the DC [19]. Two types of noise exist: additive and multiplicative noise [20, 19].

For additive noise, a noise distribution Y, which is most often a Normal or Gaussian distribution, is added to the data X. The following formula applies:

$$Z = X + Y \tag{3}$$

Furthermore, four procedures exist for additive noise [20]:

- uncorrelated noise addition,
- correlated noise addition,
- noise addition and linear transformation (e.g., Laplace noise addition)
- noise addition and non-linear transformation

For multiplicative noise, a distribution gets multiplied with the data instead of added.

Category	Technique	References
Non-perturbative	Global recoding or generalization	[16],[20]
	Local recoding	[20]
	Top-and-bottom coding	[20]
	Suppression or blocking	[16],[20]
	(Sub)sampling	[20]
Perturbative	Swapping	[16],[20]
	Re-sampling	[20]
	Noise	[16],[20]
	Aggregation or merging	[16],[20]
	Rounding	[20]
	PRAM	[20]
	Shuffling	[20]
	Synthetic data generation	[16]
De-associative	Bucketization or permutation	[20],[28]
	Anatomization	[16],[20]
	Angelization	[20]
	Slicing	[20]

Table 2: PP Data Modification

Synthetic data generation New data based on statistical properties of the original data set is generated [16].

3.5.3 De-associative

The primary intent of these techniques is to disassociate the relationship between QIDs and SAs [16, 20]. The values are not changed, instead, relations are modified. This dissociation ameliorates the privacy of the DOs. The basis for de-associative methods is groups. These are records with identical QIDs. Each group then has particular distributions for the SAs.

Bucketization Bucketization or permutation shuffles the sensitive values within each group, thus keeping the distribution, yet breaking the association between QID and SA [16].

Anatomization Anatomization is a variant of bucketization that does not permute sensitive values for each group, but splits the data into two separate tables: One table contains the QIDs and another the sensitive attributes. Both tables get a new attribute to link them: a group ID.

The sensitive attributes' table contains the group ID, a SA, and the number of occurrences of that SA within the group. If an attacker acquires the QID of a victim, the probability of inferring a sensitive attribute is given by the probability distribution of that specific attribute.

Anatomization is not possible when a DP wants to continuously publish more data to the data set. DM discussed in Section 3.4 is not possible with standard techniques and more research is needed [16].

4 Data driven business models

4.1 Introduction

Companies provide value to their customers by employing a BM, which is a relevant but not standardized topic. Due to the competitive nature of current markets, BMs are used to maintain an advantage over competitors [30]. They are utilized to evaluate business strategies, prepare for future changes, and

develop a strategy to stand out in the market. A. Sorescu [31] proposes three main attributes:

- Value creation
- Value delivery
- Value appropriation

Value creation covers the development process of the services or products the company offers. This can be done through providing a unique product or enhancing the efficiency of existing ones. In the case of value delivery, the focus lies on the products and the environment itself. How can the customer interact with the offered product or service? Finally, value appropriation takes the financial situation into account. The purpose of this attribute is to evaluate the offered values and adapt to customer's needs.

Another way to look at BMs is described by the following attributes [32]:

- Participants
- Relationship between partners and customers
- Flow of products

Clearly, this view is more socially driven in comparison to the first proposal. The focus does not only lie on the offered value, but also takes the fluency of value exchanges in to account. A BM adapts according to the needs of the company, the competitors, and the customers. Although there are a wide range of BMs, our focus lies on the DD approach because of the relevancy with the aforementioned methods. The value creation in these models is firmly correlated to the personal data of their target group [32].

4.2 Types of data-driven business models

Up to this point, there is no agreement on the representation and classification of DDBMs. Although different models exist and function wonderful in their specific environment, all of them need to define four main attributes, as depicted in Figure 3 [33].

4.2.1 Value offered

Categorizing the models according to the value they offer proves to be a fitting first breakdown.

- Product
  - Raw Data
  - Applications
- Service
  - Information
  - Insights

Service-based models can be distinguished from product-based models. The former is responsible for the creation of an environment or platform that offers benefits to their target group, while product-based models provide usable products [34, 35]. Businesses with a focus on big data opt to utilize a service-based model, although offering raw data or applications is a common practise. The reason for the high demand of raw data is explained by the fact unprocessed personal data contains more detailed information which can be used more efficiently. This however violates the expected privacy principals. The value lies in the fact that an enormous amount of knowledge can be gained from analysing and distributing personal data. This information offers feedback as value and attempts to provide useful insights.

4.2.2 Resources used

Since data is the main component in constructing a DDBM, the chosen strategy to obtain the required data is the foundation of the model. Hence, the reason for offering raw data in the first place.

- Collection
  - Web Crawlers [36]
- Generation
  - Crowdsourcing [37]

As mentioned in Section 3.1, data can be acquired through collection or generation of data. Even though utilizing exclusively available data is certainly

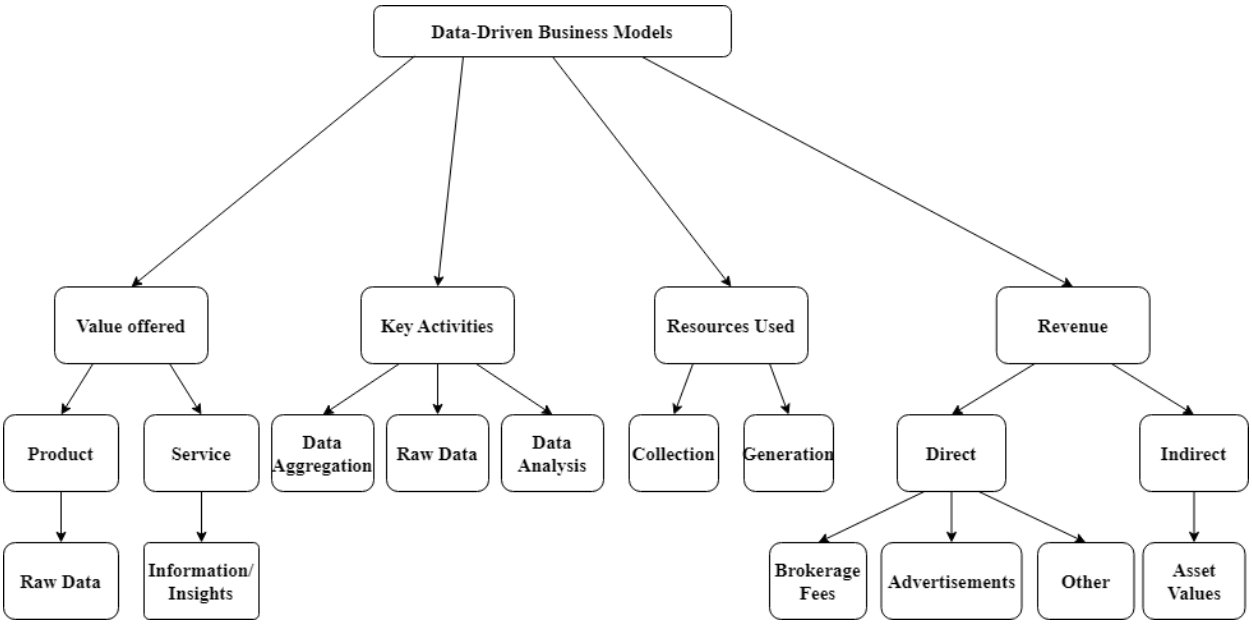

Figure 3: A structured view on different aspects of DDBMs [33]

a valid strategy, some companies obtain their data directly from customers or partners. Combining both ways results in models that heavily rely on data acquired from social media. This is of great importance for the subject of this paper, since Big Tech companies are focused on acquiring personal data through similar means. Lastly, current companies resort to crowdsourcing or web crawlers [33, 36, 37].

4.2.3 Key activities

The fundamental distinctions of DDBMs are prominent in their key activities.

- Data Aggregation
- Data Analysis
- Raw Data
- Processing

The first approach is data aggregation, consisting of summarizing data from various sources, processing it subsequently, and finally providing a comprehensible presentation [38]. Another crucial approach for different BMs is data analysis. By performing a data analysis, a significant part of the work load reduces and simplifies the further process. Because there is a need to fuel the data aggregation and analysis models with data, other models focus on data generation, providing more or specific data for their partners or customers. In some cases, the customer prefers already processed data.

4.2.4 Revenue

Seeing that the values contribute to a sizeable part of the model, it also has to feature a defined way to generate revenue in exchange for the offered services.

- Direct
  - Brokerage Fees
  - Advertisements
  - Other

- Indirect
  - Asset Values

On the one hand, BMs could focus on direct revenue, which refers to an instantaneous gain of income. Brokerage fees are revenue creating for the intermediary company. Their function is to pass information and/or data to partners in exchanges for a fee [33]. Assuming that the company resides on the internet, which is the case for most Big Tech companies, they could make use of advertisements on their site to make a passive income. This is often done in combination with a service based strategy, e.g., Amazon offering a central platform for selling products. Other possibilities such as subscription based services and licensing- or usage fees are common practise [33].

DDBMs are also known to rely on indirect revenue, which describes asset values. Even though the purpose of revenue is often thought of as money, knowledge, and power can be equally valuable (if not more). Asset values take the actual value of a resource into account, in this case personal data [33]. The following Section 4.3 goes into more detail.

4.3 Advantages for Big Tech

As touched upon in Section 2.1, Big Tech has some advantages over other companies due to the previously collected data. One of these advantages is the fact that Big Tech mainly consists of already established businesses that were able to accumulate personal data in the past years and by doing so, they are able to control markets. Smaller businesses wanting to compete with Big Tech often shift their focus of key activities. Instead of providing data, their focus lies on the analytics of already existing data [33].

4.4 Relation with PPMs

Before concluding this survey, the link between the main topics that were touched upon are clarified. Namely, how the discussed PPMs can be utilized in the context of DDBMs to create trust in Big Tech. Besides PPMs, privacy disclosures for example have already been shown to increase trust and decrease concern of consumers [39], i.e., DOs in the case of DDBMs.

According to Barbosa et al. [18] companies and organizations, not just Big Tech, can now have a competitive advantage by providing privacy-friendly services and products. They even propose a novel framework to develop privacy-friendly software, that includes the measures, methods, and models discussed in Section 3.

Moreover, Privacy by Evidence (PbE) is introduced [18], an evidence-based methodology to evaluate the level of privacy of software. Examples of evidence then include the implementation of: noise addition (3.5.2) and homomorphic encryption (3.2.3). It is this proof that fosters trust and which would lead to a competitive advantage.

PbE builds on the older Privacy by Design (PbD) paradigm, which has already been recommended to companies in the EU and US, though it lacks in clarity regarding the actual methodologies that are supposed to realize the principles [40]. PbE is more concrete.

Unfortunately, no papers were found specifically targeted towards the impact of the discussed PPMs on consumer trust in DDBMs. Though, by piecing the above findings together, it can be concluded that Big Tech companies, among others, can enhance trust in their DDBMs through PPMs. Finally, the competitive advantage is an incentive for Big Tech to actually use PP techniques.

5 Conclusion

Since the rise of Big Tech and its business models, respecting customers' privacy becomes an important attribute.

After establishing the need to guarantee the preservation of customers' privacy, this paper discussed a plethora of PPMs and their implications. These consist of five separate steps. Starting with data collection and generation, the basis of everything discussed

so far. Next, we have looked at ways to store data, followed by available protections during the publishing phase. Data modification features a wide range of different methods. Finally, extracting information and knowledge from data is the most valuable for company strategies.

Furthermore, these methods are related to the most common DDBMs. Data mining has proven to be a valuable asset in gaining knowledge as a competitive advantage.

Unfortunately, this comes at the cost of risking the trust of users.

Further research is needed to relate PPMs and their impact on consumer trust in DDBMs. A valid approach could be measuring the impact of PbE on trust.

This survey paper was written by three computer science engineering students of Ghent University for one of their master courses. They do not conduct any research themselves. The intent of this paper is for it to be their first exploration in academic research. None of the authors claim to have expert knowledge on the discussed topics, and this survey was entirely based on research of others.

Acknowledgement

We would like to express our appreciation to prof. dr. ir. Sofie Verbrugge for her proposal of this topic, and her guidance throughout the writing process. We found it to be an interesting and formative experience. We extend our gratitude to doctoral researcher Maarten de Mildt as well, for his iterative feedback and swift communication. Finally, we want to thank prof. dr. ir. Joris Walraevens for giving the course for which this paper was written. We are convinced the skills we acquired will prove to be valuable for further ventures.

References

[1] https://dictionary.cambridge.org/dictionary/english/privacy.

[2] Kobbi Nissim and Alexandra Wood. "Is privacy privacy?" In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376 (2128 2018). ISSN: 1364503X. DOI: 10.1098/rsta.2017.0358.

[3] Karl De Leeuw, Jan Bergstra, and Jan Holvast. *The History of Information Security: A Comprehensive Handbook HISTORY OF PRIVACY*. 2007.

[4] Kelly D. Martin and Patrick E. Murphy. "The role of data privacy in marketing". In: *Journal of the Academy of Marketing Science* 45 (2 Mar. 2017), pp. 135–155. ISSN: 00920703. DOI: 10.1007/s11747-016-0495-4.

[5] Paul M Schwartz. *Global Data Privacy: The EU Way*. 2019, pp. 771–818. URL: https://www..

[6] Tamara Dinev. *Why would we care about privacy?* 2014. DOI: 10.1057/ejis.2014.1.

[7] Barbara Guttman and E Roback. *An Introduction to Computer Security: the NIST Handbook*. en. 1995-10-02 1995. DOI: https://doi.org/10.6028/NIST.SP.800-12r1.

[8] David M Hart. *Antitrust and technological innovation in the US: ideas, institutions, decisions, and impacts, 1890-2000*. 2001, pp. 923–936.

[9] Olivia T Creser. *In Antitrust We Trust?: Big Tech Is Not the Problem-It's Weak Data Privacy Protections*. 2021, p. 289. URL: https://perma.cc/8PUW-YZTC%5D..

[10] Francis Fukuyama. "Making the internet safe for democracy". In: *Journal of Democracy* 32 (2 Apr. 2021), pp. 37–44. ISSN: 10863214. DOI: 10.1353/jod.2021.0017.

[11] Dipayan Ghosh and Ramesh Srinivasan. "The future of platform power reining in big tech". In: *Journal of Democracy* 32 (3 July 2021), pp. 163–167. ISSN: 10863214. DOI: 10.1353/jod.2021.0042.

[12] Viktor Mayer-Schinberger and Thomas Ramge. *A Big Choice for BigTech Share Data or Suf-*

fer the Consequences. 2018. URL: https://heinonline.org/HOL/Page?collection=journals&handle=hein.journals/fora97&id=891&men_tab=srchresults.

[13] Priyank Jain, Manasi Gyanchandani, and Nilay Khare. "Big data privacy: a technological perspective and review". In: *Journal of Big Data* 3 (1 Dec. 2016). ISSN: 21961115. DOI: 10.1186/s40537-016-0059-y.

[14] Naga Prasanthi Kundeti and M. V.P. Chandra Sekhara Rao. "Accuracy and utility balanced privacy preserving classification mining by improving K-anonymization". In: *International Journal of Simulation: Systems, Science and Technology* 19 (6 Dec. 2018), pp. 51.1–51.7. ISSN: 1473804X. DOI: 10.5013/IJSSST.a.19.06.51.

[15] Yousra Abdul Alsahib S. Aldeen, Mazleena Salleh, and Mohammad Abdur Razzaque. "A comprehensive review on privacy preserving data mining". In: *SpringerPlus* 4 (1 Dec. 2015), pp. 1–36. ISSN: 21931801. DOI: 10.1186/s40064-015-1481-x.

[16] Benjamin C.M. Fung et al. "Privacy-preserving data publishing: A survey of recent developments". In: *ACM Computing Surveys* 42 (4 June 2010). ISSN: 03600300. DOI: 10.1145/1749603.1749605.

[17] Latanya Sweeney. *k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY 1*. 2002, pp. 557–570. URL: www.worldscientific.com.

[18] Pedro Barbosa, Andrey Brito, and Hyggo Almeida. "Privacy by Evidence: A Methodology to develop privacy-friendly software applications". In: *Information Sciences* 527 (July 2020), pp. 294–310. ISSN: 00200255. DOI: 10.1016/j.ins.2019.09.040.

[19] Ricardo Mendes and Joao P. Vilela. "Privacy-Preserving Data Mining: Methods, Metrics, and Applications". In: *IEEE Access* 5 (June 2017), pp. 10562–10582. ISSN: 21693536. DOI: 10.1109/ACCESS.2017.2706947.

[20] Zhicong Lu et al. "Survey on Privacy-Preserving Techniques for Data Publishing". In: *Proceedings of the ACM on Human-Computer Interaction* 5 (CSCW1 Apr. 2021). ISSN: 25730142. DOI: 10.1145/1122445.1122456.

[21] Abid Mehmood et al. "Protection of big data privacy". In: *IEEE Access* 4 (2016), pp. 1821–1834. ISSN: 21693536. DOI: 10.1109/ACCESS.2016.2558446.

[22] Sarah Downey. *Introducing maskme: Why ever give away your real personal info online?* July 2013. URL: https://www.abine.com/blog/2013/introducing-maskme/.

[23] Daniel Kats. *A gentle introduction to attribute-based encryption*. Mar. 2019. URL: https://medium.com/@dbkats/a-gentle-introduction-to-attribute-based-encryption-edca31744ac6.

[24] Hongbing Cheng et al. "Secure big data storage and sharing scheme for cloud tenants". In: *China Communications* 12 (6 June 2015), pp. 106–115. ISSN: 16735447. DOI: 10.1109/CC.2015.7122469.

[25] Yang Xu et al. "A survey of privacy preserving data publishing using generalization and suppression". In: *Applied Mathematics and Information Sciences* 8 (3 May 2014), pp. 1103–1116. ISSN: 19350090. DOI: 10.12785/amis/080321.

[26] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. *t-Closeness: Privacy Beyond k-Anonymity and-Diversity*.

[27] Justin Hsu et al. *Differential privacy: An economic method for choosing epsilon*. Nov. 2007. DOI: 10.1109/CSF.2014.35.

[28] Lili Zhang, Wenjie Wang, and Yuqing Zhang. "Privacy Preserving Association Rule Mining: Taxonomy, Techniques, and Metrics". In: *IEEE Access* 7 (2019), pp. 45032–45047. ISSN: 21693536. DOI: 10.1109/ACCESS.2019.2908452.

[29] Maria Eleni Skarkala et al. "PPDM-TAN: A privacy-preserving multi-party classifier". In: *Computation* 9 (1 Jan. 2021), pp. 1–25. ISSN: 20793197. DOI: 10.3390/computation9010006.

[30] Fabian Hunke et al. "Towards a process model for data-driven business model innovation". In:

vol. 1. Institute of Electrical and Electronics Engineers Inc., Aug. 2017, pp. 150–157. ISBN: 9781538630341. DOI: 10.1109/CBI.2017.43.

[31] Alina Sorescu. "Data-Driven Business Model Innovation". In: *Journal of Product Innovation Management* 34 (5 Sept. 2017), pp. 691–696. ISSN: 15405885. DOI: 10.1111/jpim.12398.

[32] Christoph Zott, Raphael Amit, and Lorenzo Massa. *The business model: Recent developments and future research*. July 2011. DOI: 10.1177/0149206311406265.

[33] Philipp Max Hartmann et al. *Capturing Value from Big Data-A Taxonomy of Data-Driven Business Models Used by Start-Up Firms Journal: International Journal of Operations and Production Management*. Jan. 2016. DOI: 10.1108/IJOPM-02-2014-0098.

[34] Andreas Zolnowski, Jürgen Anke, and Jan Gudat. *Towards a Cost-Benefit-Analysis of Data-Driven Business Models*.

[35] Ivanka Visnjic, Frank Wiengarten, and Andy Neely. "Only the Brave: Product Innovation, Service Business Model Innovation, and Their Impact on Performance". In: *Journal of Product Innovation Management* 33 (1 Jan. 2016), pp. 36–52. ISSN: 15405885. DOI: 10.1111/jpim.12254.

[36] https://en.wikipedia.org/wiki/Web_crawler.

[37] Daren C. Brabham. "Crowdsourcing as a model for problem solving: An introduction and cases". In: *Convergence* 14 (1 2008), pp. 75–90. ISSN: 13548565. DOI: 10.1177/1354856507084420.

[38] https://www.techtarget.com/searchdatamanagement/definition/data-aggregation.

[39] Yue Pan and George M. Zinkhan. "Exploring the impact of online privacy disclosures on consumer trust". In: *Journal of Retailing* 82 (4 2006), pp. 331–338. ISSN: 00224359. DOI: 10.1016/j.jretai.2006.08.006.

[40] Anna Monreale et al. *Privacy-by-design in big data analytics and social mining*. 2014, p. 10. URL: http://www.epjdatascience.com/content/2014/1/10.

[41] Paolo Campanella et al. *The impact of electronic health records on healthcare quality: A systematic review and meta-analysis*. Feb. 2016, pp. 60–64. DOI: 10.1093/eurpub/ckv122.

[42] Rui Guo et al. "An efficient and provably-secure certificateless public key encryption scheme for telecare medicine information systems". In: *Journal of Medical Systems* 37 (5 Oct. 2013). ISSN: 01485598. DOI: 10.1007/s10916-013-9965-0.

[43] R. S. Evans. "Electronic Health Records: Then, Now, and in the Future". In: *Yearbook of medical informatics* (May 2016), S48–S61. ISSN: 23640502. DOI: 10.15265/IYS-2016-s006.

[44] Sasikanth Avancha, Amit Baxi, and David Kotz. *Privacy in mobile technology for personal healthcare*. Nov. 2012. DOI: 10.1145/2379776.2379779.

[45] I. Glenn Cohen and Michelle M. Mello. *Big Data, Big Tech, and Protecting Patient Privacy*. Sept. 2019. DOI: 10.1001/jama.2019.11365.

[46] Ashutosh Dhar Dwivedi et al. "A decentralized privacy-preserving healthcare blockchain for IoT". In: *Sensors (Switzerland)* 19 (2 Jan. 2019). ISSN: 14248220. DOI: 10.3390/s19020326. URL: https://www.mdpi.com/1424-8220/19/2/326.

[47] Shekha Chenthara et al. "Security and Privacy-Preserving Challenges of e-Health Solutions in Cloud Computing". In: *IEEE Access* 7 (2019), pp. 74361–74382. ISSN: 21693536. DOI: 10.1109/ACCESS.2019.2919982. URL: https://ieeexplore.ieee.org/abstract/document/8726303.

[48] Maureen Mckelvey. "The Economic Dynamics Of Software: Three Competing Business Models Exemplified Through Microsoft, Netscape And Linux". In: *Economics of Innovation and New Technology* 10.2-3 (2001), pp. 199–236. DOI: 10.1080/10438590100000009. eprint: https://doi.org/10.1080/10438590100000009. URL: https://doi.org/10.1080/10438590100000009.

[49] Alexander Bleier, Avi Goldfarb, and Catherine Tucker. "Consumer privacy and the future of

data-based innovation and marketing". In: *International Journal of Research in Marketing* 37 (3 Sept. 2020), pp. 466–480. ISSN: 01678116. DOI: 10.1016/j.ijresmar.2020.03.006.

[50] Mohamed Zaki, Andy Neely, and Florian Urmetzer. "Data-Driven Business Models: A Blueprint for Innovation Customer Loyalty View project CX Analytics: A Data-Driven Measurement System for Customer Experience View project". In: (2015). DOI: 10.13140/RG.2.1.2233.2320. URL: https://www.researchgate.net/publication/276272305.

[51] Vassilios S Verykios et al. *State-of-the-art in Privacy Preserving Data Mining **. 2004.

[52] https://en.wikipedia.org/wiki/Big_Tech.

[53] Shubha U Nabar et al. *A SURVEY OF QUERY AUDITING TECHNIQUES FOR DATA PRIVACY*.

[54] Zhan Liu et al. "Privacy-friendly business models for location-based mobile services". In: *Journal of Theoretical and Applied Electronic Commerce Research* 6 (2 2011), pp. 90–107. ISSN: 07181876. DOI: 10.4067/S0718-18762011000200009.

[55] Jeffrey T. Prince and Scott Wallsten. "How much is privacy worth around the world and across platforms?" In: *Journal of Economics and Management Strategy* 31 (4 Nov. 2022), pp. 841–861. ISSN: 15309134. DOI: 10.1111/jems.12481.

[56] Kean Birch, D. T. Cochrane, and Callum Ward. "Data as asset? The measurement, governance, and valuation of digital personal data by Big Tech". In: *Big Data and Society* 8 (1 2021). ISSN: 20539517. DOI: 10.1177/20539517211017308.