DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis
for the Attainment of the Degree
Master of Science
at the TUM School of Management
of the Technische Universität München

Identification and Evaluation of Concepts for Privacy-Enhancing Big Data Analytics Using De-Identification Methods on Wrist-Worn Wearable Data

Author:	Kevin Baumer
	Matriculation Number: 03707604
Course of Study:	Management & Technology
Supervisor:	Prof. Dr. Florian Matthes
	Chair for Software Engineering
	for Business Information Systems
Advisor:	Gloria Bondel, M.Sc.
Submission Date:	June 12, 2020

Declaration of Authorship

I hereby declare that the thesis submitted is my own unaided work. All direct or indirect sources used are acknowledged as references.

I am aware that the thesis in digital form can be examined for the use of unauthorized aid and in order to determine whether the thesis as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of my work with existing sources I agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future theses submitted. Further rights of reproduction and usage, however, are not granted here.

This paper was not previously presented to another examination board and has not been published.

Munich, June 12, 2020 Kevin Baumer

Abstract

Innovative technologies have led to enormous growth in the volume and variety of information being available. The information is stored as large data sets that are often processed in cloud environments. Organizations leverage the data to improve their products and services and to gain additional insights, leading to the realization of new economic potentials. However, when personal data is involved, privacy issues arise. Privacy laws, like the General Data Protection Regulation, aim to regulate the processing of personal data. Thus, it leads to a trade-off between the fulfillment of privacy requirements while at the same time enabling the capabilities of data analytics.

Wrist-worn wearables collect large amounts of highly sensitive health and movement data of their users. While this data can be used to identify health issues and to enhance the services offered, the data also poses significant privacy risks. Hence, there is a need to address these risks appropriately.

This work identifies and evaluates concepts to address the trade-off between an individual's privacy and the utility of data by applying de-identification methods. We provide a comprehensive overview and classification of applicable de-identification methods and classify them accordingly. Twelve expert interviews were conducted to derive ten requirements for a concept for the use-case of wrist-worn wearable data. Furthermore, we propose a local probabilistic k-anonymity concept that involves the local application of de-identification methods and the calculation of privacy estimates. After elaborating on four different scenarios, we suggest the formation of privacy clusters that combine users with the same privacy desire as the most promising approach.

Contents

List of Figures

List of Tables

Acronyms

AVC attribute value combinations. 61, 67, 73, 78–80, 82, 85, 86, 89

DFD data-flow diagram. 51–54

EI explicit identifier. 6, 59, 60, 70

FHE fully homomorphic encryption. 45, 46

GDPR General Data Protection Regulation. 1, 11, 63–67

HIPAA Health Insurance Portability and Accountability Act. 11, 12

MPC secure multi-party computation. 19, 20

NSA non-sensitive attribute. 7, 59, 60

PHE partially homomorphic encryption. 45

QI quasi identifier. 6, 31, 33, 34, 59, 60, 70

SA sensitive attribute. 7, 59, 60

SHE somewhat homomorphic encryption. 45

TEE Trusted Execution Environments. 20, 21

1. Introduction

1.1. Motivation

> "Historically, privacy was almost implicit, because it was hard to find and gather information. But in the digital world [...] we need to have more explicit rules [...]" (Gates, 2013).

Advancements in innovative technologies have led to an enormous growth in generated data. The vast amounts of data serve as the basis for emerging technologies like artificial intelligence, the Internet of Things, 5G, and Cloud Computing. Organizations and corporations make use of this data by applying big data analytics to enhance data-driven decision making and to improve their products and services (Bondel et al., 2020). Therefore, the data is being entrusted to service providers in the cloud, which enables the creation of new business opportunities. However, the collection and processing of personal data is also linked to the issue of privacy (Bondel et al., 2020). Regulations regarding data privacy are being tightened almost worldwide. The European General Data Protection Regulation (GDPR) is pioneering in this development. Also, the topic has received increased attention among individuals through several privacy breaches, e.g., at Dropbox or iCloud, that occurred (Gibbs, 2016; Lewis, 2014). As a result, people's privacy is lacking and there is a need to improve such services.

Wrist-worn wearables like smartwatches do not only provide opportunities for the tracking of sports activities. Additionally, individuals can also perform several health-related measurements. Sensors collect information about a user's heart rate, blood oxygen saturation, and new devices are even capable of conducting clinically tested ECGs (Bondel et al., 2020). Health-related data is among the most sensitive data types a human being can reveal. Hence, lots of sensitive data points are stored and processed on the provider's side, which represents a central point of attack. The data is used to detect serious health issues like cardiac arrhythmia, atrial fibrillation, and sleep apnea of an individual. Additionally, the service providers use analytics on the aggregated data sets to derive insights in order to improve their products and services, i.e., to provide users with optimized training and health recommendations (Bondel et al., 2020). However, the service providers cannot generally be assumed to be trustworthy and the transfer of data to third-party providers poses a further risk. As a consequence, there is a substantial trade-off between an individual's data privacy and the utility of the data.

To address this trade-off, a concept based on de-identification methods is envisioned. The

goal of the application of these measures on wrist-worn wearable data is to ensure analytical capabilities while at the same time achieving high levels of privacy. Hence, we aim to fulfill the privacy requirements of an individual while the organizations will benefit from privacy law compliance and the ability to continuously enhance their products and services.

1.2. Research Questions

To achieve the objectives and address these problems, we aim to answer the research questions stated in this section. On a high level, the goal of this thesis is to investigate the state of the art of de-identification methods. Additionally, requirements for the use case of wrist-worn wearables are examined. These insights are used to develop a concept which enables data privacy using de-identification methods on wrist-worn wearable data.

RQ 1: What is the state of the art of approaches using de-identification methods for privacy-enhancing Big Data Analytics and how can they be distinguished from other approaches?

This research question aims to identify the current state of research in de-identification methods. Through an extensive literature review, we will contribute with a comprehensive overview and classification of existing de-identification methods. The benefits, shortcomings, and possible application scenarios will be investigated for each of these methods. The de-identification methods will be compared and brought into context with alternative approaches for the enhancement of privacy in data analytics. Therefore, this research question aims to set a comprehensive theoretical basis for this work.

RQ 2: What are requirements for privacy-enhancing analytics of wrist-worn wearable data in the cloud?

This research question is designed to identify requirements for a concept using de-identification methods targeted explicitly to the use case of wrist-worn wearable data. Therefore, we combine insights derived throughout the literature review alongside with the findings of 12 semi-structured interviews. The interviews will provide practical knowledge and insightful discussions with experts within the domains of data privacy and data analytics. Specific requirements will be formulated to serve as the groundwork for the targeted concept.

RQ 3: What are concepts enabling data privacy for wrist-worn wearable data in the cloud based on de-identification methods?

The third research question aims to result in a concept that facilitates de-identification methods on wrist-worn wearable data to enhance the data privacy of individuals. Therefore, we develop and evaluate different concept scenarios within multiple iterations. The results of the first two research questions are guiding this development. Additionally, a simulation

approach is proposed to support the evaluation of a local application of de-identification methods.

1.3. Structure

This section gives an overview of each chapter's content to facilitate a more structured understanding for the reader, starting with the second chapter.

Chapter 2: Theoretical Foundations provides an overview of the most relevant terms and theoretical concepts for this work. A basic understanding of the domains Big Data Analytics and data privacy is facilitated.

Chapter 3: Related Work introduces related literature that covers similar topics and goals as this work.

Chapter 4: Research Approach states the research approach that is used as part of this thesis. The procedure of the literature research and expert interviews are described.

Chapter 5: Approaches for Privacy-enhancing Data Analytics in Cloud Environments gives an overview and introduction of existing concepts to enhance privacy within cloud environments.

Chapter 6: De-Identification Methods provides a comprehensive overview and categorization of de-identification methods and related privacy models. Each method is described, analyzed, and classified.

Chapter 7: Use Case of Wrist-Worn Wearable Data describes the specific use case that is considered in this work. Additionally, a generic data model is developed.

Chapter 8: Requirements for Privacy-Enhancing Analytics of Wrist-Worn Wearable Data investigates and defines ten concept requirements based on the literature review and expert interviews.

Chapter 9: Concepts Using De-Identification Methods on Wrist-Worn Wearable Data explains the development and evaluation of different concepts using de-identification methods on wrist-worn wearable data. It presents the overall results of this work.

Chapter 10: Conclusion & Outlook summarizes the results of this work based on the research questions. Finally, related limitations and possible future work are explained.

2. Theoretical Foundations

This chapter provides the theoretical foundations for this work and a shared understanding of specific terminologies. The most important definitions in the areas of Big Data analytics and data privacy that are relevant for later references are covered.

2.1. Big Data Analytics

2.1.1. Terminology

The term Big Data is used for amounts of data that cannot be processed by traditional database management systems (Chakraborty & Patra, 2014). This circumstance is due to their large volume, complexity, and partially non-structured form (R. Lu et al., 2014). Especially in recent years, Big Data has become popular with technologies like Hadoop, NoSQL, and MapReduce appearing (Soria-Comas & Domingo-Ferrer, 2016).

The emergence of Big Data mainly started due to its extreme size. Nowadays its characteristics are often referred to as three to five Vs, which are described below and illustrated in Figure 2.1 (Dautov & Distefano, 2018). Not all properties need to be satisfied, but some of them are required for the classification as Big Data (Soria-Comas & Domingo-Ferrer, 2016).

- *Volume* refers to the large amount of data generated by organizations and individuals (R. Lu et al., 2014; Soria-Comas & Domingo-Ferrer, 2016).
- *Variety* describes the proliferation of heterogeneous data forms and formats (Dautov & Distefano, 2018; R. Lu et al., 2014).
- *Velocity* refers to the high speed of data generation and processing (Soria-Comas & Domingo-Ferrer, 2016).
- *Veracity* is about the trustworthiness and accuracy of Big Data (R. Lu et al., 2014).
- *Value* describes the ability to transform the data into valuable output (Terzi et al., 2015).

Regardless of the consideration of these five characteristics, one of the biggest challenges of Big Data lies in its privacy and security (Alloghani et al., 2019; R. Lu et al., 2014). Analytics of Big Data is often infeasible to be processed locally due to the high costs and computing powers needed (El-Yahyaoui & Ech-Chrif El Kettani, 2018). For this reason, distributed

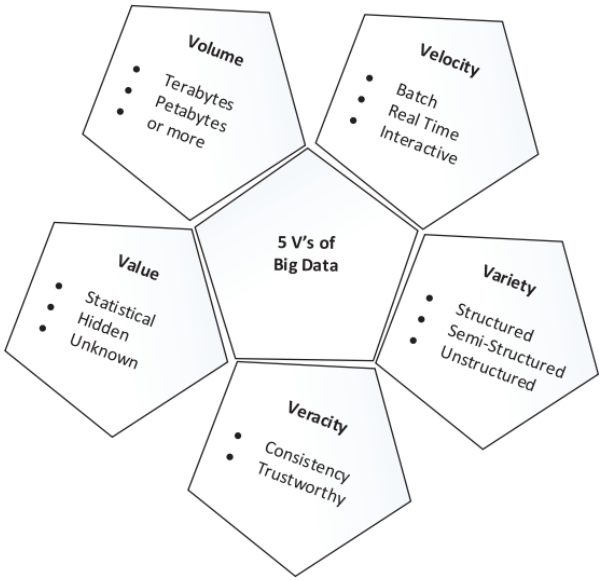


Figure 2.1.: 5 Vs of Big Data
Source: Terzi et al. (2015)

systems and technologies like cloud computing are used extensively (Chakraborty & Patra, 2014; Soria-Comas & Domingo-Ferrer, 2016). This can reduce cost and improve scalability but also entails critical privacy and security concerns (Domingo-Ferrer et al., 2019; Sidorov & Ng, 2016). These risks are highly dependent on the respective providers and especially when personal and sensitive data is stored, one needs to be careful and ensure appropriate protection measures (R. Lu et al., 2014; Will & Ko, 2015).

2.1.2. Specific Definitions related to Data

Throughout this work, a lot of different terms that are related to data are used. This subsection serves to provide appropriate definitions. We distinguish between general data related and specifically personal data related definitions. If possible, the definitions of "ISO/IEC 20889 Privacy enhancing data de-identification terminology and classification of techniques" (ISO, 2018a) are used.

Data related definitions

- *Data set / data table*: "[...] a collection of records, where each record is comprised of a set of attributes." (ISO, 2018a, p. 6) (see Table 2.1)
- *Attribute*: "Each column is called an *attribute* and denotes a field or semantic category of information that is a set of possible values [...]." (Sweeney, 2002) (see Table 2.1)
- *Attribute value*: Each cell of a table is referred to as an attribute value. (see Table 2.1)

- *Record*: "Each record is related to one data subject and is composed of a set of values [...] for each attribute [...]." (Article 29 DP Working Party, 2014, p. 12) (see Table 2.1)
- *Numerical data*: a value expressed by a number.
- *Non-numerical data*: a value expressed by characters rather than numbers.
- *Categorical data*: non-numerical data can take a limited number of values (categories).
- *Plaintext*: "[...] any information that a sender desires to transfer to a receiver." (Ogburn et al., 2013)
- *Ciphertext*: "[...] data that has been encrypted and is unreadable until it has been decrypted with a key." (Ogburn et al., 2013)
- *Data controller*: "[...] the natural or legal person, public authority, agency or other body which [...] determines the purposes and means of the processing of personal data [...]." (Art. 4 para. 7; European Parliament and Council of the European Union, 2016, p. 33)
- *Adversary / attacker*: "[...] a third party (i.e., neither the data controller nor the data processor) accessing the original records whether accidentally or intentionally." (Article 29 DP Working Party, 2014, p. 12)

Attribute A	Attribute B	Attribute C
Value A1	Value B1	Value C1
Value A2	Value B2	Value C2
Value A3	Value B3	Value C3

Record

Table 2.1.: Data set example

Personal data related definitions

- *Data subject / data principle*: identifiable natural person to which the data relates (European Parliament and Council of the European Union, 2016; ISO, 2018a). (see Table 2.2)
- *explicit identifier (EI) / direct identifier*: "attribute that alone enables unique identification of a data principal [...]." (ISO, 2018a, p. 2). Examples are the passport number or social security number. (see Table 2.2)
- *quasi identifier (QI)*: "attribute in a dataset that, when considered in conjunction with other attributes in the dataset, singles out a data principal." (ISO, 2018a, p. 4) Examples are attributes like gender, age and the zip code. (see Table 2.2)

- *sensitive attribute (SA)*: "attribute in a dataset that, depending on the application context, merits specific, high-level protection against potential re-identification attacks enabling disclosure of its values, its existence, or association with any of the data principals." (ISO, 2018a, p. 4). Examples for such attributes are health related or financial data (Gerl, 2020, p. 28). (see Table 2.2)
- *non-sensitive attribute (NSA)*: Attributes that do not belong to the three categories defined above and are therefore neither identifying nor sensitive (Domingo-Ferrer et al., 2019).

Explicit Identifier A	Quasi Identifier B	Quasi Identifier C	Sensitive Attribute D
Value A1	Value B1	Value C1	Value D1
Value A2	Value B2	Value C2	Value D2
Value A3	Value B3	Value C3	Value D3

Data subject

Table 2.2.: Personal data set example

2.2. Data Privacy

2.2.1. Definition and Differentiation to Security

Before we can discuss data privacy, it is crucial to understand its definitions, implications, and related terms. We will first have a look at the term security with a particular focus on information security, which is often seen as the connection between privacy and security. Then, we present a for this context suitable definition of privacy.

Security can be defined as "[...] a set of measures to ensure that a system will be able to accomplish its goal as intended, while mitigating unintended negative consequences" (Song et al., 2018, p. 1). Concerning the IT area, it aims to counter vulnerabilities to software and hardware like natural disasters, malicious attacks, accidental disruptions, and the unintended use of computational resources (Hurlburt et al., 2009). Information security, which is a more specific term, can be considered as the link between privacy and security. It is about protecting different kinds of information and data from destructive forces and unwanted actions (Mukherjee et al., 2015). Information security is characterized by three principles that need to be achieved: confidentiality, integrity, and availability (Domingo-Ferrer et al., 2019). Together these attributes are also known as the *CIA triad* and are regarded as the heart of information security (Song et al., 2018, p. 2). They are the three fundamental elements of information security and have been widely used and adopted both in practice and in academic literature (Samonas & Coss, 2014). Figure 2.2 shows a visualization of the CIA triad. The terms can be defined as follows:

- *Confidentiality* means that "[...] information is not made available or disclosed to

unauthorized individuals, entities, or processes" (ISO, 2018b, p. 2). Hence, it implies full trust and reliance and ensures restrictive access for authorized parties (Samonas & Coss, 2014; Song et al., 2018).

- *Integrity* is a "property of accuracy and completeness" (ISO, 2018b, p. 5). It means that information can only be modified by authorized parties or in authorized ways (Song et al., 2018, p. 2).
- *Availability* means to be "[...] accessible and usable on demand by an authorized entity" (ISO, 2018b, p. 2).

These elements do not only support and shape the theoretical understanding of information security, they are also often used as a basis for defining privacy rules and for protecting electronic health information (Samonas & Coss, 2014).

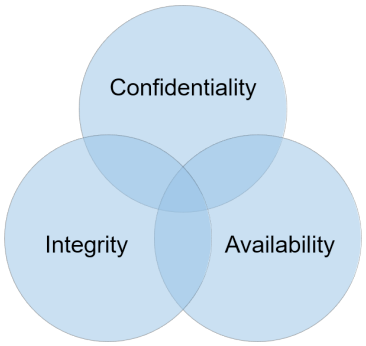


Figure 2.2.: CIA triad
Source: based on Samonas and Coss (2014)

The relation between privacy and security cannot be defined precisely. Hurlburt et al. even state that there exist different viewpoints on how the two can relate to each other: They can be interpreted either in an overlapping manner or in such a way that one concept is trumping the other one (Hurlburt et al., 2009). An often mentioned understanding is that privacy is seen as an aspect of security because some security methods have a direct effect on privacy (Song et al., 2018, p. 2). Both - privacy and security - have in common that they deal with the appropriate use and protection of information. However, their scope and reason of protection varies widely (Hurlburt et al., 2009; Song et al., 2018).

For the term of privacy, there are a lot of different definitions and interpretations. Two commonly cited and historical definitions are the following:

- "Privacy is the right to be let alone" (Warren & Brandeis, 1890).
- "Privacy is the claim of individuals, groups or institutions to determine for themselves when, how, and to what extent information about them is communicated to others" (Westin, 1967, p. 7).

Generally, privacy is often perceived as "[...] freedom from observation, disturbance, or unwanted public attention [...]" (Song et al., 2018, p. 2). It empowers an individual or a group of them to limit the level of their self-expression (Song et al., 2018, p. 2). The state of privacy implies that one might have knowledge about a person's identity, but without being aware of any associated personal facts (Jeffrey M. Skopek, 2013, p. 1755). Hence, it can be understood as a social value (Solove, 2015, p. 79).

However, the usage and interpretation of the term privacy are highly dependent on the context in which it is used. The scope of the definitions stated above is too broad, and therefore they are not suitable in our context. For a deeper understanding of privacy, the approaches of Solove and Wu are more appropriate and described in the following paragraph.

Both authors use a threat-based or attack-based approach to conceptualize privacy. Solove states that privacy cannot be seen as one single thing. Instead, it is more a combination of a plurality of many distinct things that are related in different ways (Solove, 2015, p. 74). Privacy is about identifying and characterizing relevant privacy threats, hence to protect against a variety of harms and problems (Solove, 2015; Wu, 2012). It then also serves as a basis for identifying mitigation strategies (Wu, 2012). The resulting (social) value of privacy is highly dependent on the nature of the problem that is addressed (Solove, 2015, p. 80). The following two statements illustrate how we will approach privacy in this context:

- Privacy "[...] is defined not by what it is, but by what it is not - it is the absence of a privacy breach that defines a state of privacy." (Wu, 2012, p. 1147)
- "Privacy is a set of protections against a related set of problems. These problems are not all related in the same way, but they resemble each other." (Solove, 2015, p. 80)

Thus, it is important to specify what possible privacy breaches are. The problems to be protected are referred to as privacy properties. An adversary is imagined to attack a system in order to accomplish a specific goal. In case of success for the attacker, the system does not sufficiently protect privacy (Wu, 2012). Therefore, privacy is about regulating information flows, ensuring responsible usage of information, and exercising control over the information to ensure that respective privacy properties are satisfied (Solove, 2015, p. 73). In the following Table 2.3, the privacy properties by Deng et al., Bieker et al., and Pfitzmann et al. are compared. A listing in the same line implies equality or synonymity of the terms.

A total of ten distinct properties were identified. Those properties or protection goals "[...] represent the perspective of the data subject whose rights are at stake" (Bieker et al., 2016). Deng et al. even distinguish between hard and software properties, where the first five count as hard privacy properties and the last two (*user content awareness* and *policy and consent compliance*) are considered to be soft ones. In Table 2.4, the definitions for the different attributes are stated. Bieker et al. also incorporate the classical information security attributes, namely *confidentiality*, *integrity*, and *availability* (Bieker et al., 2016). Since they were already explained in this section, they are not considered in the table.

Deng et al., 2011	Bieker et al., 2016	Pfitzmann et al., 2010
Unlinkability	Unlinkability	Unlinkability
Anonymity & Pseudonymity		Anonymity & Pseudonymity
Plausible Deniability		
Undetectability & Unobservability		Undetectability & Unobservability
Confidentiality	Confidentiality	
User content awareness	Transparency	
Policy and consent compliance		
	Intervenability	
	Integrity	
	Availability	

Table 2.3.: Privacy properties

Term	Definition
Unlinkability	"[...] of two or more items of interest (IOIs, e.g., subjects, messages, actions, ...) from an attacker's perspective means that within the system [...], the attacker cannot sufficiently distinguish whether these IOIs are related or not" (Pfitzmann et al., 2010)
Anonymity	"[...] of a subject from an attacker's perspective means that the attacker cannot sufficiently identify the subject within a set of subjects, the anonymity set" (Pfitzmann et al., 2010)
Pseudonymity	"[...] is the use of pseudonyms as identifiers" (Pfitzmann et al., 2010)
Plausible Deniability	"[...] from an attackers perspective means that an attacker cannot prove a user knows, has done or has said something" (Deng et al., 2011)
Undetectability	"[...] of an item of interest (IOI) from an attacker's perspective means that the attacker cannot sufficiently distinguish whether it exists or not" (Pfitzmann et al., 2010)
User content awareness	"[...] is proposed to make sure that users are aware of their personal data and that only the minimum necessary information should be sought and used to allow for the performance of the function to which it relates" (Deng et al., 2011)
Policy and consent compliance	"[...] requires the whole system [...] as data controller to inform the data subject about the system's privacy policy, or allow the data subject to specify consents in compliance with legislation, before users accessing the system" (Deng et al., 2011)
Transparency	"[...] means that the data subjects have knowledge of all relevant circumstances and factors regarding the processing of their personal data" (Bieker et al., 2016)
Intervenability	"[...] entails the control of the data subjects, as well as the controller or supervisory authority over the personal data" (Bieker et al., 2016)

Table 2.4.: Privacy property definitions

2.2.2. Regulations

The concepts of two regulations related to data privacy are illustrated in this subsection. We look at the General Data Protection Regulation (GDPR), as it is the most important one within the European Union, and the Health Insurance Portability and Accountability Act (HIPAA), as it is a very practical and often cited regulation from the US.

There are other relevant regional regulations regarding data privacy such as the German Federal Data Protection Act and the Clarifying Lawful Overseas Use of Data Act in the US. Due to time constraints and their limited direct implications on this work, we will not further elaborate these.

General Data Protection Regulation (GDPR)

The GDPR came into force in 2018. It has the goal of unifying personal data protection within the whole European Union (Tamburri, 2020). The most relevant articles and principles are summarized in the following paragraphs.

Article 5 specifies that personal data shall follow the seven key principles (a) lawfulness, fairness and transparency, (b) purpose limitation, (c) data minimization, (d) accuracy, (e) storage limitation, and (f) integrity and confidentiality (Art. 5 para. 1, European Parliament and Council of the European Union, 2016, p. 36).

Article 9 prohibits the processing of special categories of personal data which reveal information about racial/ethnic origin, political opinions, religious/philosophical beliefs, memberships, genetic/biometric/health data or sexual orientation. This prohibition is lifted if (a) the data subject gives explicit consent for the processing, (g) the processing is in substantial public interest or (h) the processing serves medical or health care purposes (Art. 9 para. 1/2, European Parliament and Council of the European Union, 2016, p. 38).

Article 20 gives an individual the right to receive his or her personal data that was provided to a controller (Art. 20 para. 1, European Parliament and Council of the European Union, 2016, p. 45).

Article 25 incorporates the two principles of privacy by design and privacy by default. Privacy by design means that a data controller should "[...] implement appropriate technical and organizational measures [...], which are designed to implement data-protection principles [...], in an effective manner [...] in order to [...] protect the rights of data subjects." (Art. 25 para. 1, European Parliament and Council of the European Union, 2016, p. 48) Privacy by default specifies that such measures process, by default, only necessary data with regards to the amount of data, the extent of processing, the storage period, and the accessibility (Art 25 para. 2, European Parliament and Council of the European Union, 2016, p. 48).

Article 32 states that appropriate technical and organizational measures shall be implemented to ensure a level of security and privacy appropriate to the risk (Art 32 para. 1, European Parliament and Council of the European Union, 2016, p. 51).

Article 35 describes that a data protection impact assessment needs to be carried out in case the processing can result in a high risk to an individual's rights and freedoms (Art 35 para. 1, European Parliament and Council of the European Union, 2016, p. 51).

Health Insurance Portability and Accountability Act (HIPAA)

Health Insurance Portability and Accountability Act (HIPAA) is a health-related regulation in the US which was introduced in 1996 and led to initial challenges for everyone processing health data (Alshugran et al., 2015; Pentecost, 2004; Schoppmann & Sanders, 2004). It is especially popular due to the HIPAA Safe Harbor provision which defines 18 attributes that need to be removed or altered in health data (Prasser et al., 2018; Wu, 2012). However, it is also often criticized as it does not take into account other available data and advanced re-identification methods (Nelson, 2015). The respective attributes are the following: (1) names, (2) geographic divisions smaller than state, except 3 digit zip codes, (3) dates of birth, death, admission, and ages greater than 89 years, (4) Driver's license or car license numbers, (5) social security numbers, (6) numbers of medical records, (7) health plan numbers, (8) account numbers, (9) phone numbers, (10) fax numbers, (11) e-mail addresses, (12) license numbers, (13) vehicle identification numbers, (14) medical device or serial numbers, (15) internet URLs, (16) IP addresses, (17) biometric identifiers and (18) any other unique identifier, characteristic, or code (Lavin, 2006).

2.2.3. Data Privacy in Practice

The implementation of data privacy in practice is often challenging as one is likely to underestimate the risk. In 2000, it was shown that around 87% of all US citizens could be uniquely identified only by knowing the ZIP code, the date of birth and the gender of a person (Sweeney, 2000). Sweeney also used these attributes to link a medical data set that was published for research purposes and a purchasable voting list of a city in Massachusetts. The combination is illustrated in Figure 2.3. As a result, the disclosure of sensitive health data for particular individuals was facilitated (Sweeney, 2000). It shows that only a few detailed characteristics are needed to identify a person.

The number of data breaches in general, but also regarding health data, is growing. This is mainly due to the sharing of data for research purposes (Prasser et al., 2018). Two popular privacy breaches with large corporations involved are the ones from AOL and Netflix. In 2006, AOL released a dataset containing 20 million search queries for 650,000 users collected over three months. Identifying attributes were replaced with a unique identification number. It led to the identification and location of users as researchers were able to correlate different search terms to individuals (Article 29 DP Working Party, 2014; Ohm, 2009). In the same year, Netflix published data about more than 100 million movie ratings by 500,000 users as part of a recommendation system contest. Only the movie names, the rating values, the dates of rating, and the respective user ids were released with some noise added on top. However, researchers were still able to identify 99% of the users just be knowing eight ratings and the

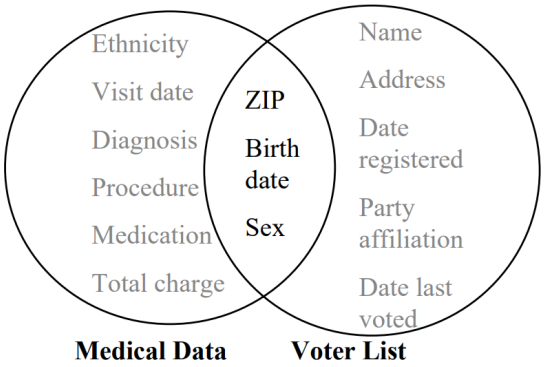


Figure 2.3.: Linking of medical data and voter list
Source: Sweeney (2000)

corresponding dates (Article 29 DP Working Party, 2014; Ohm, 2009; Wu, 2012). These two examples show that personal information is increasingly exposed to purposes that are useful in a particular way (Solove, 2015; Wu, 2012). This leads us to the content of the following subsection.

2.2.4. Privacy vs. Utility

Privacy and utility are two important elements that need to be considered regarding a data set. The utility represents the usefulness of a data set for a specific purpose (Gerl, 2020, p. 45). In an ideal case, a data set would allow a maximum of analytical functions while at the same time achieving a high level of privacy. However, the two goals are in tension with each other, and there needs to be a balance between them (Venkataramanan & Shriram, 2016, p. 16). Some argue that they are fundamentally incompatible with each other, while others say that both can simultaneously be achieved when undertaking the correct actions (Wu, 2012, p. 1117). This trade-off is, along with the performance, also often seen as a key problem for cloud service providers (Will & Ko, 2015). Figure 2.4 visualizes an exemplary dependency between privacy and utility.

It is illustrated that an increase in privacy generally leads to a decrease in utility. Different types of functions have different effects on the two variables, which leads to varying gradients in the displayed curve. In a cryptographic function, the privacy and utility are either zero or one, whereas other anonymization methods lie in between (Wu, 2012, p. 1125). Therefore, the anonymization process of data can be seen as a constrained optimization problem which, if designed properly, leads to a reasonable balance between these two goals (Venkataramanan & Shriram, 2016, p. 17).

The value of utility itself can be measured in various ways. A decision for one should always be dependent on the context (Wu, 2012, p. 1117). Existing utility measures are for example generalization height, discernibility metric, average locations appearance ratio, average relative error, pairs lost, accuracy, completeness, entropy, ambiguity, and normalized

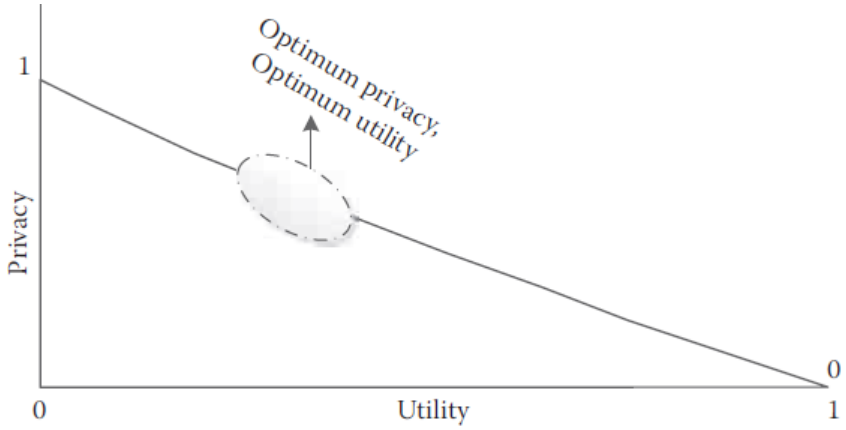


Figure 2.4.: Trade-off between privacy and utility
Source: Venkataramanan and Shriram (2016, p. 16)

average equivalence class size (Gerl, 2020; Machanavajjhala et al., 2007; Prasser et al., 2018; Terrovitis et al., 2017; Tomashchuk et al., 2019).

3. Related Work

In this chapter, we introduce previous work with topics and goals that are similar or related to ours.

De-identification overviews: In 2014, the Article 29 DP Working Party published the "Opinion 05/2014 on Anonymisation Techniques" where existing anonymization techniques were analyzed regarding three risk criteria. The publication compares the strengths and weaknesses of single techniques to recommendations for an adequate anonymization process in a given context (Article 29 DP Working Party, 2014). The "ISO/IEC 20889" is the official ISO standard for privacy-enhancing data de-identification techniques. It provides a terminology and classification of different methods but does not offer much practical guidance on how and when to apply them.

De-identification approaches: Next to these overviews, there also exist concepts and approaches similar to the one we aim for. Kim et al. developed methods to collect sensitive health lifelogs from a smartwatch in a privacy-protecting manner. A histogram-based data collection approach and a concept using local differential privacy are proposed and implemented (Kim et al., 2019). Terrovitis et al. present four approaches to preserve privacy in the publication of location sequences captured by credit cards. Suppression and splitting techniques are used to prevent privacy breaches while enabling aggregated analysis (Terrovitis et al., 2017). Related to that, Y. Li et al. developed an algorithm based on differential privacy for transit smart card data to eliminate privacy concerns on published data (Y. Li et al., 2020). Prasser et al. propose a de-identification solution for high-quality health data in data sharing environments. The authors use suppression and generalization methods to protect an individual's privacy by focusing on the restriction of unique data characteristics (Prasser et al., 2018). A privacy-preserving outsourced calculation toolkit named Pockit is using homomorphic encryption schemes to let data owners outsource their data to cloud storage (Liu et al., 2019). Chatfield et al. presented a tool specifically designed for the US regulation HIPAA. The ARX Data Anonymization Tool is a popular open-source software for the application of de-identification methods. In their paper, Prasser et al. describe the current development state of this tool, practical experiences they have encountered and the remaining issues and challenges they are facing (Prasser et al., 2020).

4. Research Approach

This chapter describes the research approach used in this work to solve the proposed research questions. The theoretical basis was set throughout an extensive literature review. Further, we conducted semi-structured interviews with experts in the domains data privacy and data analytics to elicit requirements in chapter 8. Using these findings, we developed a privacy-enhancing concept for wrist-worn wearable data. The development, refinement, and validation of the concept, as it is described in chapter 9, was developed in multiple iterations supported by discussions with three researchers in this area.

4.1. Literature Research

The research part of this work was conducted as an extensive literature review. The main areas being covered though the review were existing methods and approaches for privacy-enhancing data analytics, privacy models, de-identification methods, the processing and protection of health-related data, and models to evaluate privacy risks. Therefore, we searched the following databases:

- ScienceDirect
- IEEE Xplore Digital Library
- Web of Science
- ACM Digital Library
- SpringerLink

Additionally, we also considered cited literature from the papers that were identified. The literature research was primarily used for chapter 5 and chapter 6, partially also for chapter 7 and chapter 8.

4.2. Expert Interviews

The goal of this section is to explain the qualitative data collection approach used in the expert interviews and to provide an overview of the consulted interview partners.

The expert interviews are based on the approach of Gläser and Laudel (2009). The authors propose to guide through the interview with a pre-defined catalog of questions based on the research questions. The approach consists of recording and transcribing the interviews as well as the subsequent analysis. As part of this work, we used a semi-structured interview guide based on our research questions.

The interview guide (Appendix B) starts with a general introduction to the thesis topic and the request for consent to record the interview. Then the role of the interviewee within his/her organization and his/her relevant experience are examined. The central part consists of five open questions. The goal is to derive requirements, practical insights, and implications for the described concept for the application of de-identification methods on wrist-worn wearable data. Hence, it targets the second and third research question. We aim to understand how data privacy is handled within different organizations, how risks can be identified and evaluated, and which measures are taken to mitigate them. Application potentials of de-identification methods and the interviewee's view on our use case are also investigated. The interviews close with further remarks on the topic and a discussion.

A total of 12 interviews were conducted over a period of about one month. The interviewees are from nine different companies with a broad and diverse range in terms of their number of employees. They serve different roles in their organizations, but all with close touching points to the areas of data privacy or information security. Among them are employees from specific data privacy departments, data analysts with a focus on privacy as well as consultants in these domains. In three cases, the interview partners were able to point to other people in their organization with relevant experience in the investigated areas. All 12 interview partners agreed to be recorded, and therefore all interviews could be transcribed. On average, an interview lasted about 40 minutes. Table 4.1 gives an overview of the interviewees. We use the ID in the first column to refer to specific interviewees from now one.

ID	Role	Relevant experience (in years)	No. of employees	Duration (hh:mm:ss)
I1	Consultant Data Privacy & Information Security	>5	1-10	01:04:02
I2	Consultant Data Security & Data Privacy	13	1-10	00:56:49
I3	Head of Data Privacy	20	10,001-50,000	00:28:23
I4	Managing Director & Lawyer	11	1-10	00:42:57
I5	Researcher Digital Health	2	11-50	00:28:13
I6	Data-driven Development & Data Privacy Expert	8	>100,000	0:33:43
I7	Information Security Officer & Data Protection Officer	20	10,001-50,000	00:35:51
I8	Head of Data Privacy	23	>100,000	00:39:57
I9	Head of Data Privacy	19	>100,000	00:28:58
I10	Consultant Data Privacy	7	51-250	00:34:30
I11	Chief Information Security Officer	8	251-1,000	00:32:55
I12	Key Expert Data Privacy	6	>100,000	00:52:21

Table 4.1.: Interview participants

5. Approaches for Privacy-Enhancing Data Analytics in Cloud Environments

This chapter provides an overview of existing approaches that support the enhancement of privacy for Big Data analytics in cloud environments. In the literature, such approaches are often referred to as privacy-enhancing technologies (PETs) (Heurix et al., 2015). We will introduce such technologies, demonstrate how they can be distinguished from de-identification methods, and illustrate why we focus on the latter.

5.1. General Overview

An overview of relevant approaches is provided in Figure 5.1, divided into two categories. Data-centric techniques relate directly to the data they protect and therefore enforce changes in storage or computation of that data (Grandison et al., 2017; Mansfield-Devine, 2014). Indirect methods, on the other hand, imply changes on an infrastructure level. The four different techniques will be described along with its advantages and shortcomings in the following subsections.

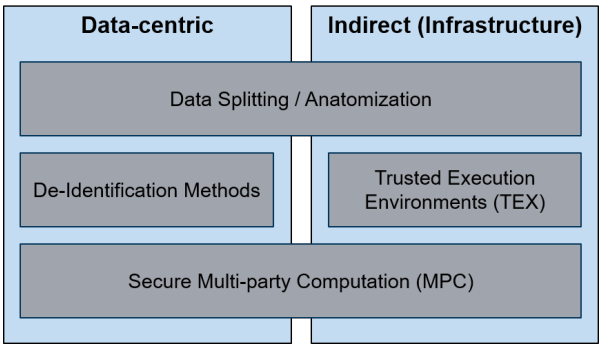


Figure 5.1.: Approaches for privacy-enhancing data analytics

5.2. Secure Multi-Party Computation

Secure multi-party computation (MPC) is a subfield of cryptography that deals with computations of combined data while preventing the different parties to reveal their private input.

Such calculations are possible through the distribution of encrypted messages (shares) among the parties, which make it possible to derive correct results without sharing sensitive data between the parties (McWaters et al., 2019; UN Global Working Group, 2019). Therefore, a trusted intermediary is not needed anymore (The Royal Society, 2019). The technology was classified into data-centric as well as indirect methods because infrastructural changes are needed and specific data needs to be passed. The complete logic behind MPC is rather complex, but Figure 5.2 illustrates a conceptualization of this technique where three parties share messages with each other to calculate the overall average of their salary.

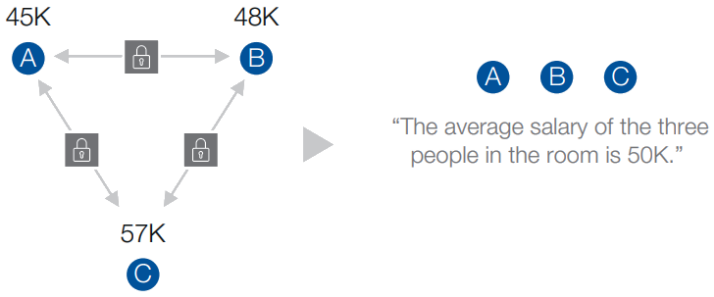


Figure 5.2.: Secure multi-party computation
Source: McWaters et al. (2019)

MPC allows the joint computation on sensitive data without the need to trust in third parties as only the results are revealed. However, these computations can be relatively expensive, but the times are currently significantly increasing with further developments. Right now, a slowdown factor of about 10,000 can be estimated, dependent on the operation. The communication costs for the exchange of messages must also be taken into account because the parties need to communicate with each other (McWaters et al., 2019; UN Global Working Group, 2019).

5.3. Trusted Execution Environments

Trusted Execution Environmentss (TEEs) are a fully hardware-based approach and are therefore part of the indirect technologies. A TEE is an isolated part of a processor that is also known as an enclave (The Royal Society, 2019). This part is designed such that the rest of the system cannot access it to provide confidentiality. The memory or execution state is not visible to other processes on the processor such that the data in the TEE does not need to be encrypted, and a secure computation capability is provided. As a result, a secret code can be executed isolated from the rest of the system and in a privacy-preserving manner (UN Global Working Group, 2019).

TEEs are used to outsource computations to a server or cloud environment without the need for cryptographic solutions and therefore without loss of information and utility (The Royal

Society, 2019). However, one needs to fully trust the hardware when it comes to potential errors and security vulnerabilities. The immunity of such side-channel attacks can be difficult to prove (Papadimitriou et al., 2016).

Commercial solutions of TEEs are widely available today (Papadimitriou et al., 2016). Common practical implementations are ARM Trustzone[1], Intel Software Guard Extensions[2], Intel Trusted Execution Technologie[3], IBM Secure Execution[4], and AMD Secure Processor[5].

5.4. Data Splitting

Data splitting is a technique that involves fragmenting sensitive data into chunks that are stored in separate locations. The splitting is done in a way such that single parts do not disclose identities or reveal sensitive information (Domingo-Ferrer et al., 2019; Sánchez & Batet, 2017). This approach facilitates storage distribution through multi-cloud environments and therefore minimizes the consequences a potential data leakage can have (Alqahtani & Sant, 2016; Sánchez & Batet, 2017). The metadata containing the splitting criterion and the storage locations needs to be stored in a trusted database. Figure 5.3 shows an exemplary workflow for data splitting including this metadata in a multi-cloud scenario (Domingo-Ferrer et al., 2019).

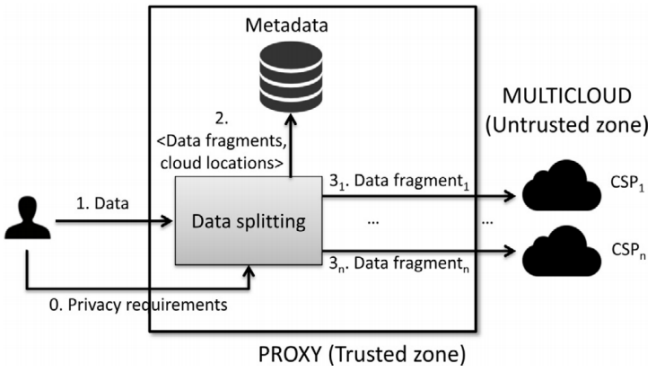


Figure 5.3.: Data splitting workflow
Source: Domingo-Ferrer et al. (2019)

By distributing to multiple cloud environments, issues regarding the file size are avoided and load balancing is facilitated (Alqahtani & Sant, 2016). No information loss occurs and lots of computational functionalities are supported by transferring the queries to the different storage locations (Sánchez & Batet, 2017). However, the metadata storage represents a

[1]https://www.arm.com/why-arm/technologies/trustzone-for-cortex-a
[2]https://www.intel.de/content/www/de/de/architecture-and-technology/software-guard-extensions.html
[3]https://www.intel.de/content/www/de/de/support/articles/000025873/technologies.html
[4]https://developer.ibm.com/blogs/technical-overview-of-secure-execution-for-linux-on-ibm-z
[5]https://www.amd.com/en/technologies/security

single point of failure and attack, as it is a requirement for accessing and reconstructing the data (Domingo-Ferrer et al., 2019).

5.5. De-Identification

De-identification methods are methods to transform a data set with the goal to preserve an individual's privacy while at the same time maintaining as much analytical functionalities as possible (Tomashchuk et al., 2019). They will be analyzed in detail in the following chapter.

We focus on de-identification methods as they are the only full data-centric approach that was obtained in this overview. Therefore, it does not require a direct change of the infrastructure, and the cloud providers do not need to be trusted. They can realize multiple levels of security and privacy, which we will elaborate as part of this work.

Several tools for de-identification methods exist. Their application is, however, always very use case dependent. Such open-source tools are ARX Data Anonymization[6], UTD Anonymization ToolBox[7], μ-ARGUS[8], τ-ARGUS[9] and sdcMicro[10]. Commercially offered products are Privacy Analytics Eclipse[11] , Google Cloud Healthcare API[12], Anonos BigPrivacy[13], IBM InfoSphere Optim Data Privacy[14], IBM Guardium Data Protection[15], Data Masking by DataSunrise[16], Oracle Data Masking and Subsetting Pack[17], and Informatica Data Masking[18].

[6]https://arx.deidentifier.org
[7]http://www.cs.utdallas.edu/dspl/cgi-bin/toolbox
[8]http://research.cbs.nl/casc/mu.htm
[9]http://research.cbs.nl/casc/tau.htm
[10]https://github.com/sdcTools/sdcMicro
[11]https://privacy-analytics.com/software/privacy-analytics-eclipse
[12]https://cloud.google.com/healthcare
[13]https://www.anonos.com/bigprivacy
[14]https://www.ibm.com/products/infosphere-optim-data-privacy
[15]https://www.ibm.com/products/ibm-guardium-data-protection
[16]https://www.datasunrise.com/data-masking
[17]https://www.oracle.com/de/database/technologies/security/data-masking-subsetting.html
[18]https://www.informatica.com/products/data-security/data-masking.html

6. De-Identification Methods

The following chapter deals with different types of de-identification methods. First, the term itself is defined and delimited from other terms. Then, a general review of methods in the existing literature is pursued. As a result, this chapter provides a comprehensive overview and categorization of applicable de-identification methods and their corresponding privacy models.

6.1. Definition & Terminology

In the existing literature, there is no unified and consistent definition of the term *de-identification*. Also, the term is often used together and sometimes interchangeably with *data anonymization* and *pseudonymization*. These all describe similar concepts about a process or action to bring data in a more anonymous or pseudonymous state (Tomashchuk et al., 2019). However, there is no agreed-upon consensus about their relation.

De-Identification A commonly used definition for the term *de-identification* is that it refers to a process of removing the associations between

- "[...] a set of identifying attributes [...] and the data principal" (ISO, 2018a, p. 2)
- or "[...] data and identifying elements of individual data subjects." (Tomashchuk et al., 2019, p. 63)

The data principal (or data subject) describes the person to whom the data refers. Identifying attributes in this context refer to direct identifiers and quasi identifiers as they were defined in subsection 2.1.2. The general objective of de-identification is to preserve the privacy of those individual data subjects, hence to minimize the risk of unintended identity and information disclosure (Nelson, 2015; Tomashchuk et al., 2019). It is important to understand that de-identification does not only describe one single method or technique. It is more a broad "[...] collection of approaches, algorithms, and tools that can be applied to different kinds of data with differing levels of effectiveness" (Garfinkel, 2015, p. 1). These methods serve to modify data such that the vulnerability of re-identification attacks is reduced. They also help to achieve different privacy requirements, which are often represented in privacy models (Tomashchuk et al., 2019). Hence, those models (see section 6.3) are closely related to the application of de-identification methods. Generally, one can say a more aggressive

application of the methods leads to improved privacy protection, but also less remaining utility in the resulting dataset (Garfinkel, 2015, p. 1).

Data anonymization The term *data anonymization* sometimes is used as an interchangeable synonym for *de-identification*. Hence, the same definition as illustrated in the last paragraph is applied. In contrast to that, other researchers define *data anonymization* as a real subset and specific kind of *de-identification* (Garfinkel, 2015; Tomashchuk et al., 2019). In that case, the additional condition of irreversibility is included, which indicates that the process cannot be reversed and the data cannot be re-identified. Two exemplary definitions for the latter case are as follows:

- "[...] data anonymization irreversibly masks data in a privacy-preserving way" (Domingo-Ferrer et al., 2019, p. 43)
- anonymization is a "[...] process by which personal data [...] is irreversibly altered in such a way that a data subject can no longer be identified directly or indirectly, either by the data controller alone or in collaboration with any other party" (ISO, 2017, p. 2)

The usage of the terms *irreversibility* and *reversibility* however also depends on how it is defined and interpreted. This is why the second definition above explicitly describes irreversibility as a measure to disallow the controlled re-identification, including the combination with other parties or data sources (ISO, 2018a, p. 31). The application of a definition like this is somewhat subjective because one can not always say for sure that re-identification is not possible anymore since one might not be aware of all combinable datasets. In contrast to that, *irreversibility* can also relate to the property of mathematical functions. In this case, the definition focuses only on the functionality of the method itself and not on the output and its identifiability.

Pseudonymization The usage of the term *pseudonymization* and especially its relation to the previously described terms are also not consistent in the existing literature. In some contexts, *pseudonymization* is treated equivalently to *de-identification* itself (Garfinkel, 2015, p. 2). Others describe it as a specific subset of anonymization "[...] that both removes the association with a data subject and adds an association between a particular set of characteristics relating to the data subject and one or more pseudonyms" (ISO, 2017, p. 6). Yet others refer to it as not being a subtype of anonymization and merely reducing the possibility to link data with the original identity of data subjects (Article 29 DP Working Party, 2014, p. 3). A for this work suitable and comprehensive definition is the following: "Pseudonymization means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information [...]" (European Parliament and Council of the European Union, 2016, p. 33).

Masking Another terminology that needs to be considered in this context is the term *masking* or *data masking*. Generally, masking means to modify data based on predetermined rules (like transformation algorithms). The objective of this technique is to retain the structure and, therefore, the functional usability of the data, as far as it is possible (Nelson, 2015). Masking is an irreversible process because the data can be transformed only in one way (Meunier et al., 2019). The most widely used example is to mask a credit card number by replacing several digits with Xs (e.g. 1234 5678 1234 5678 -> XXXX XXXX XXXX 5678). Nelson states that de-identification methods can generally be described as masking methods (Nelson, 2015, p. 15).

Implications for this work As shown at the beginning of this section, the terms *de-identification*, *data anonymization*, and *pseudonymization* cannot be distinguished from each other in a unified way in the existing research. Because of these inconsistencies, we use a combination of the approaches of Garfinkel and Tomashchuk et al. with the following criteria:

- Firstly, we use the term *de-identification* as a "[...] concept of a higher level, which covers both anonymization and pseudonymization [...]" (Tomashchuk et al., 2019). The term therefore describes and comprises all concepts which were mentioned in this section. Hence, each method that falls under the scope of either one of the concepts is considered as a de-identification method.
- Secondly, we avoid the terms *data anonymization* and *pseudonymization* except for in this section. Instead, only the term *de-identification* is used with its generic meaning explained above.
- Thirdly, within the definition of *de-identification methods*, we do not use the term of *irreversibility*, such that the methods are not restricted and can be taken into account in a broad context.

Summarizing these criteria, the following overall definition is used within the scope of this work:

> A de-identification methods is a "method for transforming a dataset [...] with the objective of reducing the extent to which information is able to be associated with individual data principals [...]" (ISO, 2018a, p. 2).

6.2. General Overview

To obtain a general overview of existing de-identification methods, an extensive literature review was performed. The goal of this review was first to identify different methods in the existing literature and then to categorize and cluster them. As a result, we will provide a comprehensive overview of available methods and their application scenarios. As an initial

step of the literature review, the following databases were searched for overviews of methods within the area of *de-identification*, *data anonymization*, and *pseudonymization*:

- ScienceDirect
- IEEE Xplore Digital Library
- Web of Science
- ACM Digital Library
- SpringerLink

As a result, seven different sources were identified. They are listed in Table 6.1 below, indicating the term they are referring to. The ID value in the first column serves as an identifier of the source for the following table. These sources are the starting point for the subsequent categorization of de-identification methods.

ID	Source	Term referred to
1	Domingo-Ferrer et al., 2019	Data anonymization
2	Mansfield-Devine, 2014	Data masking
3	Nelson, 2015	De-identification
4	Tomashchuk et al., 2019	De-identification
5	Article 29 DP Working Party, 2014	Anonymization techniques
6	Bourka and Drogkaris, 2018	Pseudonymization techniques
7	ISO, 2018a	De-identification

Table 6.1.: Identified sources for overviews of de-identification methods

In the next step, the mentioned sources were analyzed with regard to the different methods they are describing. Table 6.2 denotes an overview of the sources and their corresponding methods. The first column indicates the ID of the source, which can be mapped to Table 6.1. The header row shows the names of the methods as they are used in the literature. Techniques that are already described as synonyms in one paper are summarized in one field (e.g. generalization and global recoding). We use the term *creating pseudonyms* instead of *pseudonymization* to prevent misleading regarding the definitions stated earlier in this chapter. Also, terms that are used to define categories or to cluster multiple methods are not considered since they do not represent an applicable method for themselves. An *X* indicates whether the method is covered in the corresponding source using the specified term. However, this table does not take method overlaps and equivalences into account. Hence, it solely showcases which source paper references which methods.

In order to enable a clean classification, a set of distinct techniques needs to be created. To further summarize the listed terms, we determine appropriate exclusions, synonyms and variants.

The method anatomization, which refers to data splitting, does not provide any value for

Source	(Sub-) Sampling	Local suppression	Record suppression	Cell suppression	Generalization / Global recoding	Local generalization	Noise addition / Variance masking	Data swapping / Shuffling	Microaggregation	Substitution / Randomization	Deterministic encryption	Homomorphic encryption	Nulling Out	Character masking	Creating pseudonyms	Character scrambling	Truncation	Redaction	Encoding	Blurring	Masking	Perturbation	Top and bottom coding	Rounding	Permutation	Anatomization	Differential privacy
(1)	X	X			X		X	X	X																		
(2)							X	X		X	X		X	X													
(3)	X		X	X	X		X	X		X				X	X	X	X	X	X	X	X	X					
(4)		X					X		X		X				X								X	X			
(5)					X		X								X										X		X
(6)											X	X		X	X	X				X							
(7)	X	X	X		X	X	X	X	X		X	X			X						X		X	X	X	X	X

Table 6.2.: Overview of de-identification methods in existing research

reducing the association of information with individual data subjects for itself. This is because different access rights need to be assigned to the split tables. Hence, we also exclude anatomization in the context of this consideration. Differential privacy is also listed in two of the sources. However, we refer to it as a privacy model (see section 6.3) and not as a de-identification method itself.

The methods in Table 6.2 were investigated with regard to their overlap with the objective to derive synonyms and variants. As a result, 15 unique de-identification methods were identified. The following Table 6.3 shows these methods and specifies the term we will be using from now on (first column), the synonyms for each method, and possible variants. The table contains all terms out of Table 6.2 besides blurring, masking, and perturbation. These cannot be unambiguously allocated to one of these methods and instead serve as synonyms for multiple methods. As a result, they were omitted in order to avoid confusion.

To achieve a meaningful categorization, the methods were examined concerning different characteristics. First of all, we use the classification approach of Domingo-Ferrer et al. to distinguish between perturbative and non-perturbative methods. Non-perturbative methods do not alter the truthfulness of the original data but instead reduce their accuracy. Whereas, the resulting data of perturbative methods is not truthful in general but statistical properties of the original data may be preserved. (Domingo-Ferrer et al., 2019) A method can always be assigned to one of them. Hence, one is either perturbative or non-perturbative. Additionally, the applicability of the techniques on numerical and on categorical (non-numerical) data is

Method	Synonyms	Variants
(Sub-) Sampling	-	-
Suppression	Nulling out	Local (cell) / global suppression, record / attribute suppression (redaction)
Generalization	Recoding	Local / global generalization
Rounding	-	-
Top and bottom coding	-	-
Noise addition	Variance masking	-
Data swapping	Shuffling, Permutation	-
Microaggregation	Averaging	-
Randomization	Substitution, Encoding	-
Character masking	-	-
Creating pseudonyms	Pseudonymization	-
Character scrambling	-	-
Truncation	-	-
Deterministic encryption	-	Order- / format-preserving encryption
Homomorphic encryption	-	-

Table 6.3.: Synonyms and variants of de-identification methods

assessed. Table 6.4 shows the respective results. An *X* means the method is always applicable to that data type. An *(X)* indicates that the method is not always or only with difficulty applicable in that case. Lastly, a - shows that the respective method is not applicable at all.

	(Sub-) Sampling	Suppression	Generalization	Rounding	Top and bottom coding	Noise addition	Data swapping	Microaggregation	Randomization	Character masking	Creating pseudonyms	Character scrambling	Truncation	Deterministic encryption	Homomorphic encryption
Perturbative (P) Non-perturbative (N)	N	N	N	N	N	P	P	P	P	N	P	P	N	P	P
Numerical data	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Categorical data	X	X	X	-	(X)	(X)	X	(X)	X	X	X	X	X	X	X

Table 6.4.: Classification of de-identification methods

Based on the information in Table 6.4, the 15 identified de-identification methods were classified. The final result is displayed in Figure 6.1. This classification approach is the result of multiple iterations and discussions with three researchers. The perturbation characteristic serves as input for the two main categories in which the methods are placed. Furthermore, four subcategories, which are independent of the perturbation category, were derived. They are defined as follows:

- Data type independent methods can be applied to any type of data.
- Numerical methods are always applicable to numerical data, but not or only partially to categorical data.
- Deletion classifies methods that incorporate the removal of values.
- Generalizing methods describe methods that reduce the accuracy and granularity of values.

An interesting finding is that there are no methods that are only applicable to categorical data. Thus, all methods can be applied to numerical values.

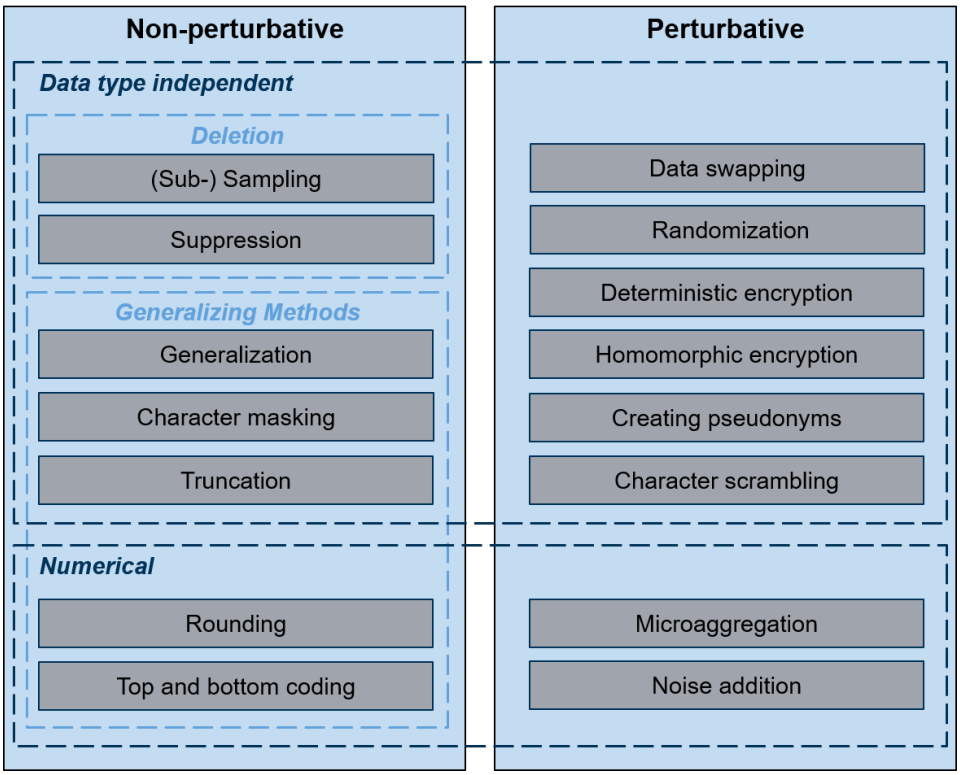


Figure 6.1.: Classification of de-identification methods
Source: Bondel et al. (2020)

6.3. Privacy Models

Privacy models aim to mathematically conceptualize privacy for data sets. The models specify conditions and requirements the data set must satisfy to control disclosure risks. They do not determine specific transformations (Soria-Comas & Domingo-Ferrer, 2016). Thereby, they guarantee the privacy of individuals to a certain degree and preserve utility at the same time (Gerl, 2020, p. 38). By using such models, one "[...] may obtain a clearer idea on the type and level of protection achieved for the data outsourced to the cloud, no matter their size, type or structure." (Domingo-Ferrer et al., 2019) The literature proposes several different privacy models. The most cited and used ones, being k-anonymity, l-diversity, t-closeness, and differential privacy, will be further explained in this work. For each model, we describe the general definition, provide an example showing the achievable value, and state the shortcomings of the approach.

6.3.1. k-anonymity

K-anonymity was published as one of the first formal privacy protection models by Sweeney in 2002. The author itself defines the model as follows:

> "A release provides k-anonymity protection if the information for each person contained in the release cannot be distinguished from at least k-1 individuals whose information also appears in the release." (Sweeney, 2002)

This definition focuses solely on people, but the concept can generally be applied to every kind of record within data tables. It implies the requirement that every combination of values of the quasi-identifiers appears at least k times in the whole table (Machanavajjhala et al., 2007). When fulfilling this criterion, a data set is denoted as k-anonymous (Gerl, 2020, p. 39). The set of records with the same values of quasi identifiers (whose minimum size is k) is referred to as an equivalence class (ISO, 2018a; Soria-Comas & Domingo-Ferrer, 2016). The model intends that each record can hide within a group of records, its equivalence class. A potential adversary can never narrow down the set of records to less than k elements (Domingo-Ferrer et al., 2019; Wu, 2012). Thus, the higher the k, the higher is the achieved privacy level. To achieve k-anonymity on a data set, the application of different transformation methods is required. The de-identification methods are described more precisely in the following sections of this chapter.

Table 6.5 illustrates an example of k-anonymity applied on a data table. The left side shows the original raw data table with the ZIP code, the age, and the nationality as the non-sensitive quasi identifiers and the health condition as the corresponding sensitive attribute. In this table, each record is unique. The table on the right-hand side shows the transformed data set fulfilling the k-anonymity constraint with k=4. Hence, the table is 4-anonymous. Each record now cannot be distinguished from three other records by only looking at the quasi identifiers.

The main advantage of k-anonymity is protection against identity disclosure. Due to the

Non-sensitive QI			Sensitive
Zip Code	Age	Nationality	Condition
13053	28	Russian	Heart Disease
13068	29	American	Heart Disease
13068	21	Japanese	Viral Infection
13053	23	American	Viral Infection
14853	50	Indian	Cancer
14853	55	Russian	Heart Disease
14850	47	American	Viral Infection
14850	49	American	Viral Infection
13053	31	American	Cancer
13053	37	Indian	Cancer
13068	36	Japanese	Cancer
13068	35	American	Cancer

(a) Original table

Non-sensitive QI			Sensitive
Zip Code	Age	Nationality	Condition
130**	< 30	*	Heart Disease
130**	< 30	*	Heart Disease
130**	< 30	*	Viral Infection
130**	< 30	*	Viral Infection
1485*	≥ 40	*	Cancer
1485*	≥ 40	*	Heart Disease
1485*	≥ 40	*	Viral Infection
1485*	≥ 40	*	Viral Infection
130**	3*	*	Cancer
130**	3*	*	Cancer
130**	3*	*	Cancer
130**	3*	*	Cancer

(b) k-anonymous table with k=4

Table 6.5.: Example for k-anonymity
Source: based on Machanavajjhala et al. (2007)

indistinguishability from other k-1 records, an individual cannot be linked to less than k records based on the quasi identifiers (N. Li et al., 2007). This makes the connection with other data sets more complicated. Precisely, the probability to correctly identify a record is at most 1/k (Domingo-Ferrer et al., 2019).

However, there are also several shortcomings with this model, which have been addressed by recent publications. On the one hand side, it is clear that the privacy risk is not reduced to completely zero. On the other side, this approach may not provide enough data quality for common usage scenarios (Prasser et al., 2018). The biggest threat is the so-called homogeneity attack (Machanavajjhala et al., 2007). This one allows deriving sensitive information about a person or record without knowing precisely which exact record corresponds to it. This is the case when all records of an equivalence class share the same characteristic of a sensitive attribute (Wu, 2012, p. 1142). Referring to our example in Table 6.5: By knowing that a person is part of the data set and falls within the 3rd equivalence class (Zip code: 130**; Age: 3*; Nationality: *), one can easily conclude that the person has cancer because it is the only attribute value within this class. Another shortcoming is that the privacy level heavily depends on the adversary's background information, which might lead to the discovery of additional sensitive information (Wu, 2012, p. 1143). This means for our example (4-anonymous table in Table 6.5): If an attacker knows that a person is 21 years old, lives in a 13068 ZIP code and is Japanese, he cannot be sure whether he has a heart disease or caught a virus. Nevertheless, when the attacker is additionally considering that Japanese people have a relatively low incidence of heart disease, he can conclude that the person most likely has a viral infection (Machanavajjhala et al., 2007). Sweeney formulated further attacks against

k-anonymity like the unsorted matching attack, the complementary release attack, and the temporal attack (Sweeney, 2002). However, the two explained attack possibilities are generally perceived as the most crucial ones.

6.3.2. l-diversity

The l-diversity principle was introduced by Machanavajjhala et al. in 2007 as an extension of k-anonymity to address its shortcomings. The requirement for l-diversity is defined as follows:

> "An equivalence class is said to have l-diversity if there are at least l 'well-represented' values for the sensitive attribute. A table is said to have l-diversity if every equivalence class of the table has l-diversity." (N. Li et al., 2007)

However, the original paper does not state what exactly 'well-represented' means (Machanavajjhala et al., 2007). The most straightforward understanding is the so-called distinct l-diversity, which we will use to describe the principle here. In this case, it has the same meaning as distinct and therefore implies that there are at least l distinct values of the sensitive attribute in each equivalence class (N. Li et al., 2007). L-diversity is an enhancement of k-anonymity since it is designed to protect against attribute disclosure through homogeneity and background knowledge attacks (ISO, 2018a, p. 21). These improvements are achieved by ensuring diversity of sensitive attributes within the equivalence classes (Domingo-Ferrer et al., 2019; Machanavajjhala et al., 2007).

By reusing the previous example, Table 6.6 shows how the principle can be applied and which benefit comes with it. The left-hand side (table a) shows the 4-anonymous table which resulted in the k-anonymity explanation above. Since the records of the third equivalence class (the last four records) always have the same sensitive value, the table only fulfills l-diversity with l=1 and is therefore 1-diverse. Table b represents the same records and values but in a different order and with partially different representations of ZIP code and age. The table now is 3-diverse because all equivalence classes have three different values for the sensitive attribute. This measure addresses the homogeneity attack of k-anonymity in a very effective way. Additionally, the background knowledge attack also becomes more difficult (Machanavajjhala et al., 2007). The attacker from before (targeting the 21 year old Japanese from Zip code 13068) cannot conclude the condition of its victim anymore. Even with background knowledge, he is left with a significant amount of uncertainty now (Article 29 DP Working Party, 2014). A larger l again leads to higher privacy protection.

The two main disadvantages of l-diversity are represented by the similarity attack and the skewness attack. The problem of similarity appears when attributes within an equivalence class are unevenly distributed because the semantical closeness of these values is not taken into account (ISO, 2018a; N. Li et al., 2007). For our example, this could appear when there are three pretty similar diseases. The skewness attack can appear when the overall distribution is skewed (Gerl, 2020, p. 42). Assuming two different equivalence classes: the first one has

Non-sensitive QI			Sensitive
Zip Code	Age	Nationality	Condition
130**	< 30	*	Heart Disease
130**	< 30	*	Heart Disease
130**	< 30	*	Viral Infection
130**	< 30	*	Viral Infection
1485*	≥ 40	*	Cancer
1485*	≥ 40	*	Heart Disease
1485*	≥ 40	*	Viral Infection
1485*	≥ 40	*	Viral Infection
130**	3*	*	Cancer
130**	3*	*	Cancer
130**	3*	*	Cancer
130**	3*	*	Cancer

(a) l-diverse table with l=1

Non-sensitive QI			Sensitive
Zip Code	Age	Nationality	Condition
1305*	≤ 40	*	Heart Disease
1305*	≤ 40	*	Viral Infection
1305*	≤ 40	*	Cancer
1305*	≤ 40	*	Cancer
1485*	> 40	*	Cancer
1485*	> 40	*	Heart Disease
1485*	> 40	*	Viral Infection
1485*	> 40	*	Viral Infection
1306*	≤ 40	*	Heart Disease
1306*	≤ 40	*	Viral Infection
1306*	≤ 40	*	Cancer
1306*	≤ 40	*	Cancer

(b) l-diverse table with l=3

Table 6.6.: Example for l-diversity
Source: based on Machanavajjhala et al. (2007)

an equal number of positive and negative records, the second one has 99% positive and 1% negative records. Even though both equivalence classes are 2-diverse, the records which are part of the first class have a much higher probability of having a positive attribute (N. Li et al., 2007).

6.3.3. t-closeness

The principle of t-closeness extends the privacy model of l-diversity. It was proposed by N. Li et al. in 2007. The general idea is that the distribution of an attribute in an equivalence class should mirror the initial distribution of the attribute in the whole table (Article 29 DP Working Party, 2014; N. Li et al., 2007). It is formerly defined as follows:

> "An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness." (N. Li et al., 2007)

Hence, the concept ensures that the distance between the distributions stays below a specific threshold t (ISO, 2018a). Contrary to k-anonymity and l-diversity, higher privacy protection is, in this case, achieved by a lower value of the parameter t. However, a big issue is to measure the distance between the distributions in a senseful way. The variational distance and the Kullback-Leibler distance are two approaches for numerical attributes, but they do not reflect semantic distances between values (N. Li et al., 2007). Therefore, the Earth Mover's Distance (EMD) was proposed, which is suitable for numerical and categorical attributes. The

calculation is based on the amount of work needed to transform one of the distributions into another one (N. Li et al., 2007).

Non-sensitive QI		Sensitive	
Zip Code	Age	Salary	Disease
476**	2*	3K	gastric ulcer
476**	2*	4K	gastritis
476**	2*	5K	stomach cancer
4790*	≥ 40	6K	gastritis
4790*	≥ 40	11K	flu
4790*	≥ 40	8K	bronchitis
476**	3*	7K	bronchitis
476**	3*	9K	pneumonia
476**	3*	10K	stomach cancer

(a) l-diverse table with l=3

Non-sensitive QI		Sensitive	
Zip Code	Age	Salary	Disease
4767*	≤ 40	3K	gastric ulcer
4767*	≤ 40	5K	stomach cancer
4767*	≤ 40	9K	pneumonia
4790*	≥ 40	6K	gastritis
4790*	≥ 40	11K	flu
4790*	≥ 40	8K	bronchitis
4760*	≤ 40	4K	gastritis
4760*	≤ 40	7K	bronchitis
4760*	≤ 40	10K	stomach cancer

(b) table with 0.167-closeness (salary) and 0.278-closeness (disease)

Table 6.7.: Example for t-closeness
Source: based on N. Li et al. (2007)

In Table 6.7, the improvement of t-closeness compared to l-diversity is illustrated. The table on the left serves 3-diversity: By knowing that a record belongs to the first equivalence class, an attacker can derive that the person has a relatively low salary (between 3k and 5k) and has some stomach-related problems (gastric ulcer, gastritis or stomach cancer) (N. Li et al., 2007). By considering t-closeness, this issue can be solved, which is shown in the right-hand table. The classification into the three equivalence classes results in 0.167-closeness with regards to the salary and 0.278-closeness with regards to the disease. Thus an attacker can no longer infer that a person has a high salary or issues with its stomach (N. Li et al., 2007). Hence, t-closeness is a useful mitigation measure against inference attacks.

However, t-closeness does not mitigate risks of identity disclosure and unlinkability compared to k-anonymity and l-diversity. In addition, the model can result in a significant loss of data utility because correlations within the data (e.g. between quasi-identifiers and sensitive attributes) are eliminated (ISO, 2018a).

6.3.4. Differential Privacy

Differential privacy is another concept targeting privacy, which was introduced by Dwork in 2006. It recently received strong attention and is widely accepted within the privacy community, which led to an increasing number of publications (Kim et al., 2019). The model is not directly built upon the previously mentioned models. However, it is related to t-closeness (ISO, 2018a). It copes with privacy from a different perspective and can be defined as follows:

> "Differential privacy mathematically guarantees that anyone seeing the result of a differentially private analysis will essentially make the same inference about any individual's private information, whether or not that individual's private information is included in the input to the analysis." (Wood et al., 2018, p. 212)

Differential privacy is achieved by setting a boundary on the probability to infer the presence or absence of a data subject from the dataset. This approach even incorporates that the attacker might have access to other linkable datasets (ISO, 2018a, p. 21). Wu explains that theoretically, it is always possible that a data set reveals additional information about an individual. When an adversary knows that a person is exactly two centimeters shorter than an average Lithuanian woman, a data set with that information would reveal information about the person (Wu, 2012, p. 1137). The information about the average height would be roughly the same when one individual did not appear in the data set. This explains the concept of differential privacy fairly well since its intention is only to reveal information that does not significantly depend on individuals (Wu, 2012, p. 1138). Dwork has set the following mathematical definition for differential privacy:

> "A randomized function $\mathcal{K}$ gives ϵ-differential privacy if for all data sets D_1 and D_2 differing on at most one element, and all S $\subseteq$ Range($\mathcal{K}$)." (Dwork, 2006)

$$\Pr\left[\mathcal{K}\left(D_1\right) \in S\right] \leq e^{\epsilon} \times \Pr\left[\mathcal{K}\left(D_2\right) \in S\right]$$

Hence, the parameter ϵ controls the amount of information which is leaked. A small ϵ implies that the affect of an individual's information being part in the data set is significantly low (X. Lu & Au, 2017). Figure 6.2 serves as an illustration of the mentioned definitions.

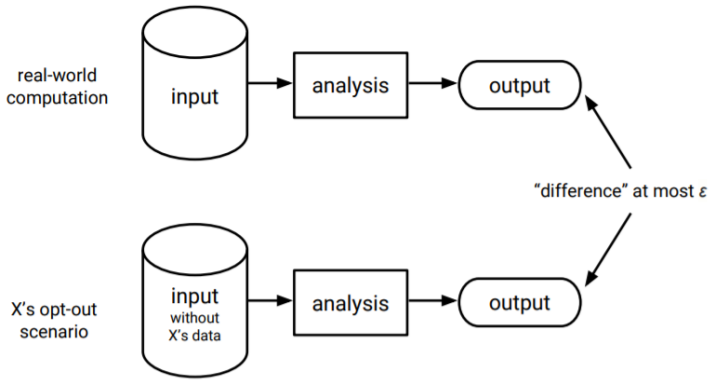


Figure 6.2.: Differential privacy concept
Source: Wood et al. (2018, p. 235)

The method usually used to achieve differential privacy is *noise addition*, for which the model sets different requirements and constraints for the needed noise. The more noise is added, the more privacy can be guaranteed by the concept. However, from a utility point-of-view, this is not beneficial since an increase of noise also results in less data utility (X. Lu & Au, 2017).

The big challenge is to come up with the proper amount of noise to preserve the usefulness of the results on the one hand, and protect an individual's privacy on the other hand (Article 29 DP Working Party, 2014). Generally, any algorithm that meets the formal definition is called differentially private (ISO, 2018a). Regarding the application, there are two different approaches (Soria-Comas & Domingo-Ferrer, 2016):

- Creating a synthetic data set for a specific purpose while keeping the original data.
- Masking the values of the original records by adding noise.

The first approach was the initial idea of differential privacy and is based on the assumption that the creation of the synthetic data set is dependent on the specific query to be pursued. Thus, for different queries, different data sets are created instead of releasing one single data set. However, this also leads to the biggest shortcoming of this approach: By using and combining multiple query requests, it might be possible to derive information that should not be disclosed. Hence, it is crucial to retain a query history to detect and limit such attacks (Article 29 DP Working Party, 2014).

6.4. Non-Perturbative Methods

Earlier in this chapter, non-perturbative de-identification methods were characterized as methods that do not alter the truthfulness of the original data, but instead reduce their accuracy (Domingo-Ferrer et al., 2019). Within this section, the identified methods of this category are presented, covering an explanation, examples, possible applications, and shortcomings.

6.4.1. Sampling

Sampling means to release only a small subset (sample) of all available records (Domingo-Ferrer et al., 2019; Nelson, 2015). For example, one could decide to share just 20% of a data set's records with a third party. This has the effect that an attacker does not know if a unique record of the released data set is also unique in the original data. Therefore, it is not possible to imply that a record corresponds to a specific data subject because it is unknown if the subject is part of the sample at all. Hence, it mitigates several possible attack scenarios. A smaller amount of records within the sample data leads to more protection and a higher privacy level (Domingo-Ferrer et al., 2019). At the same time, it has to be considered that the choice of the sample is crucial since it is used to represent the data set as a whole for analysis and pattern recognition (ISO, 2018a). In Table 6.8, it is shown how a sampling method can be applied.

The most straightforward algorithm to pursue sampling is probability sampling, which incorporates random numbers to select the records. It adds uncertainty about the data on the one hand, but it also might destroy statistical properties and, therefore, the utility of the

Nationality	Age	Gender	Disease
Argentina	27	Male	None
American	49	Female	Cancer
Japanese	43	Female	Heart Disease
German	24	Male	Cancer
Dutch	22	Female	None
American	29	Male	Heart Disease
Japanese	43	Male	Cancer
Argentina	39	Female	Heart Disease
American	38	Male	None
German	32	Male	None
⋮	⋮	⋮	⋮

Released data

Table 6.8.: Example for sampling

data (ISO, 2018a). It makes sense to ensure that the proportion of several attribute values stays the same as in the original (Nelson, 2015). This allows various statistical properties to be retained. Another approach can be to make use of the k-anonymity model and to incorporate only the records of equivalence classes with a specific minimum value of k. This would improve the protection through k-anonymity but also requires a little more computational effort compared to random sampling. But this approach is then identical to record suppression (see the following subsection), whereas we only consider the random probability sampling as part of this method.

The most significant benefit of this method is that correct values remain. However, it can result in a substantial utility loss due to the removal of records (Domingo-Ferrer et al., 2019). If applied appropriately, it can be an efficient de-identification method. Combining it with generalization and randomization methods can increase the effectiveness even more (ISO, 2018a).

6.4.2. Suppression

Suppression (sometimes also referred to as nulling out) is about removing certain attribute values of the data set either for some records (local suppression) or for all records (global suppression) (Domingo-Ferrer et al., 2019; Samarati & Sweeney, 1998). The values are then either deleted or replaced with null values. This method is most suitable for categorical data, but it generally works on all data types (ISO, 2018a, p. 14). There are different variants:

- Local suppression (also cell suppression) involves removing specific attribute values from selected records that could lead to the identification of a data subject (ISO, 2018a).
- Global suppression refers to suppressing specific attribute values from all records

(globally) (Terrovitis et al., 2017).

- Record suppression is about deleting an entire record (for example, because he contains multiple rare attributes) (ISO, 2018a).
- Attribute suppression (also: redaction) entails the removal of an attribute with all values.

The basic idea of this method is to identify and delete rare attribute values or rare combinations of attribute values to prevent a possible re-identification of the data subjects. Suppression may as well be applied in combination with k-anonymity: One can drop all records of equivalence classes with a number of entries below a certain threshold. Also, the deletion of a single value can lead to a higher number of records sharing the value combinations and, consequently, to a higher k-anonymity protection (Domingo-Ferrer et al., 2019; Nelson, 2015).

Table 6.9 shows an exemplary application of local suppression in the left table (a). Since there is only one record with an age between 40 and 49, that value is suppressed. Furthermore, the 5th record is not only in a unique age range, but also suffers a relatively rare disease. Hence, it makes sense to suppress the whole record and delete it from the table.

Age	Gender	Disease
30-39	Male	None
30-39	Female	Heart Disease
30-39	Female	Heart Disease
30-39	Male	Heart Disease
60-69	Female	Cancer
20-29	Male	Heart Disease
20-29	Male	None
20-29	Female	Heart Disease
40-49	Female	None
20-29	Male	Heart Disease

Suppressed values

(a) local suppression

Zip Code	Age	Nationality	Condition
13053	28	Russian	Heart Disease
13068	29	American	Heart Disease
13068	21	Japanese	Viral Infection
14853	50	Indian	Cancer
14853	55	Russian	Heart Disease
13055	21	Japanese	Heart Disease
12003	42	Russian	Heart Disease
13068	35	American	Cancer
⋮	⋮	⋮	⋮

Suppressed attribute

(b) attribute suppression

Table 6.9.: Example for suppression
Source: based on Machanavajjhala et al. (2007)

The right table (b) shows the application of attribute suppression. The complete deletion of the zip code will protect the individual's privacy tremendously, but it can only be applied if the zip code is no longer necessary for further use. Attribute suppression is therefore best suited for explicit identifiers like names or unique numbers, but can also be applied to quasi identifiers in case they are not needed anymore.

Samarati et al. state that the application of suppression is especially useful in combination with generalization. Rare attribute values can require a significant amount of generalization, and suppression can be used to moderate this process by deleting those values. The author

recommends the joint application of both. The data owner has to find the right balance between generalization, at the cost of less precision, and suppression, at the cost of completeness (Samarati & Sweeney, 1998). Our previous example already showed this combination since a data set with a generalized attribute (age) was used.

The benefit of suppression is that only true and exact values are kept. However, the application leads to missing values resulting in a substantial utility loss. Especially for small sub-groups of the datasets, the utility loss has its most significant impact (Domingo-Ferrer et al., 2019; Nelson, 2015). Nevertheless, if applied in the right way and in combination with generalization, suppression represents a promising de-identification method. Primarily, attribute suppression on the right attributes can lead to a substantial effect in privacy protection.

6.4.3. Generalization

Generalization is one of the most popular de-identification methods. It means to substitute attribute values with more general categories to reduce detail (Domingo-Ferrer et al., 2019). The method is used to make combinations of quasi identifiers less rare, and the application is especially useful in combination with privacy models like k-anonymity and its extensions (Article 29 DP Working Party, 2014; Nelson, 2015).

Generalization methods work on all kinds of attributes. Numerical values can be transformed into categorical intervals to reduce granularity (e.g., 37 $\rightarrow$ 30-39) (Domingo-Ferrer & Mateo-Sanz, 2002). For non-numerical and categorical attributes, a suitable generalization hierarchy is needed. All attributes can potentially be arranged in such a hierarchy in order to enable generalization. The creation of this generalization hierarchy is a crucial part of this method because it supports the process based on predefined transformation rules. An example would be to replace a specific street name with the name of the city, which is a more general category (e.g. '221B Baker Street London $\rightarrow$ 'London') (ISO, 2018a). Other possibilities are to generalize names to genders or words to their first letters.

Job	Age	Gender	Disease
Engineer	35	M	HIV
Engineer	33	M	HIV
Lawyer	29	M	Flu
Lawyer	33	M	Flu
Writer	42	F	Cancer
Singer	45	F	Cancer
Writer	42	F	Cancer

(a) Original table

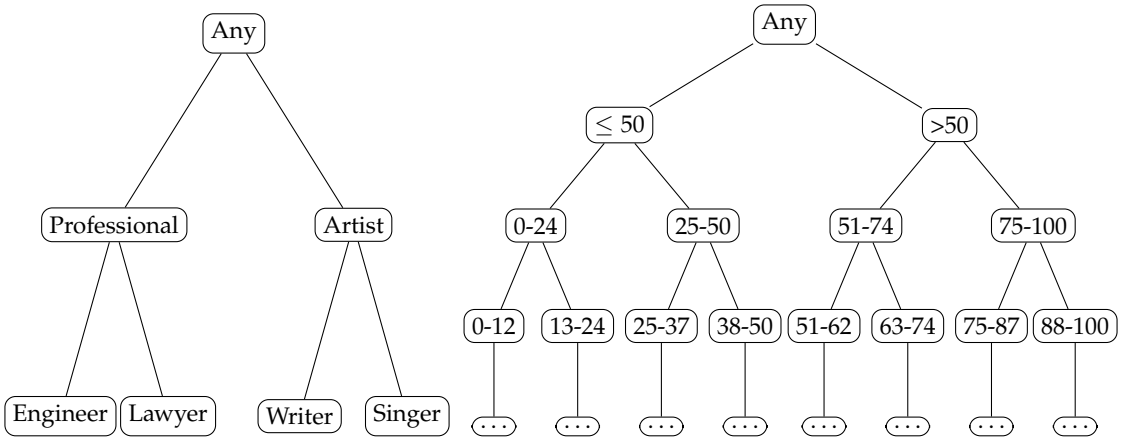
Job	Age	Gender	Disease
Professional	25-37	M	HIV
Professional	25-37	M	HIV
Professional	25-37	M	Flu
Professional	25-37	M	Flu
Artist	38-50	F	Cancer
Artist	38-50	F	Cancer
Artist	38-50	F	Cancer

(b) Generalized 3-anonymous table

Table 6.10.: Example for generalization
Source: based on Gerl (2020, p. 41)

Table 6.10 shows an example of how generalization is applied on a numerical and a non-

numerical attribute to achieve k-anonymity with k=3. The attributes job and age were recoded such that each equivalence class has at least three corresponding records. The choice of the generalization hierarchy and the numerical intervals depends on the single values of the considered attributes. In Figure 6.3, the for this example applied generalization hierarchies for the two attributes are visualized. The levels at the very bottom (level 0) represent the most granular attribute values in the original data set. With every further level up granularity and therefore also the utility of the attribute is reduced, whereas the achieved privacy level improves. By bringing any value of an attribute to the highest level and using the value *Any*, the utility is completely lost, and the effect is analogous to a deletion of the column (Samarati & Sweeney, 1998). The number of applicable generalization hierarchy levels is called hierarchy height. Hence, the attribute job has a hierarchy height of 3, the attribute age one of 5. Due to its aggregating nature, generalization directly influences k-anonymity, l-diversity, and t-closeness dependent on the selected generalization hierarchy. Therefore, the intention should always be to increase the size of the clusters (equivalence classes) but at the same time to keep as much utility as possible.



(a) Hierarchy of attribute 'job' (b) Hierarchy of attribute 'age'

Figure 6.3.: Generalization hierarchy
Source: based on Gerl (2020, p. 36+40)

The method also offers some specific variants which are sometimes even presented as independent methods:

- Global generalization refers to the method shown in the example above. It includes generalizing all single values of an attribute.
- Local generalization means to generalize only specific attribute values from selected records. The intention is to remove rare values while keeping the remaining ones unmodified (ISO, 2018a).

The method has the benefit that only true values are released. Limitations are the significant loss in granularity and, consequently, utility where one needs to find the right

balance (Domingo-Ferrer et al., 2019). Hence, the achieved value is highly dependent on the selected generalization hierarchy which offers much customizability.

6.4.4. Character Masking

Character masking is a method which operates completely character-based and involves replacing predetermined characters with Xs or other characters (Mansfield-Devine, 2014; Nelson, 2015). The following table shows three examples of how the method can be applied to different types of attributes. Character masking theoretically can always be applied, but the data structure and the formatting of the attribute values have to be considered. The method has substantial restrictions regarding its applications because the utility of the transformed values is often lost.

Attribute	Original value	Transformed value
Credit card number	4678 3412 5100 5239	XXXX XXXX XXXX 5239
Last name	Smith	Smit*
Phone number	+49 160 12345667	+49 176 XXXXXXXX

Table 6.11.: Example for character masking
Source: based on Bourka et al. (2018), Nelson (2015)

6.4.5. Truncation

Truncation is a very specific method where the nth character of an attribute value is removed or cut. An example is to replace 'SMITH' with 'SMI' (Nelson, 2015). The method has commonalities with character masking since both delete certain characters. However, truncation does not indicate if and which characters are missing. Reversing this process is harder than for character scrambling or character masking because it is not known if and at which point letters were removed. The utility can also be quite low, but it can be a suitable method when a part of a string reveals too much information or is simply not needed.

6.4.6. Rounding

Rounding is a method that is part of the generalizing methods but limited to numerical attributes. It involves the rounding of numerical values based on a predetermined rounding base (Bourka & Drogkaris, 2018). It reduces granularity and can help to achieve a specific level of k-anonymity while remaining truthfulness of the data. Examples that incorporate different rounding bases are $12.734 \rightarrow 13$ and $56{,}899.5 \rightarrow 60{,}000$.

6.4.7. Top and Bottom Coding

Top and bottom coding is also part of the category of generalizing methods. With this method, only a specific subset of all attribute values are transformed. It is about setting a threshold for the largest and/or smallest value of an attribute. All values that do not fall within the given interval are replaced with the respective top and bottom values. An example is to hide large salaries by only indicating that a value is above a threshold (>100,000) (ISO, 2018a). This method is useful if extremely low or high values of a quasi identifier are rather rare. In this case, the method can help to achieve k-anonymity constraints. For the application on sensitive values, top and bottom coding can help to protect the disclosure of sensitive outliers.

6.5. Perturbative Methods

Perturbative de-identification methods were defined as methods that change the truthfulness of single values within the data set. However, some statistical properties of the original data may be preserved (Domingo-Ferrer et al., 2019). Such methods are preferable if the goal is to perform mainly aggregation computations. This section will explain the identified perturbative methods together with examples, possible applications, and shortcomings.

6.5.1. Data Swapping

Data swapping (also referred to as shuffling or permutation) means to randomly swap data between records regarding a specific attribute (Mansfield-Devine, 2014). It has the effect that values are artificially linked to different data subjects while univariate distributions of the values are exactly preserved (Article 29 DP Working Party, 2014; Domingo-Ferrer et al., 2019).

The method is datatype independent and can be applied to numerical and non-numerical values. As part of the application, it should be considered that the swapping algorithm cannot be reconstructed (ISO, 2018a). For numerical values, specific conditions can be set in order to swap values only within a specific range. This allows a more or less stable variance-covariance matrix, which is also applicable for ranked categorical attributes (Domingo-Ferrer et al., 2019).

In Table 6.12, the method is applied to the *income* attribute. However, it shows that another correlating attribute (in this case, the job), which is not swapped could still make it possible to draw inferences on the income. Hence, this has to be considered carefully. Generally, an application on quasi identifiers as well as sensitive attributes is possible to prevent attribute disclosure (Domingo-Ferrer et al., 2019). A widespread use case, for example, is software testing because the method provides real values (Nelson, 2015).

Similar to other methods, data swapping alone is not enough in most cases, so it should be combined with other methods. Another shortcoming is that an attacker can draw wrong

Year	Gender	Job	Income
1957	M	Engineer	45k
1957	M	CEO	70k
1957	M	Unemployed	5k
1964	M	Engineer	43k
1964	M	Manager	100k

(a) Original table

Year	Gender	Job	Income
1957	M	Engineer	70k
1957	M	CEO	5k
1957	M	Unemployed	43k
1964	M	Engineer	100k
1964	M	Manager	45k

(b) Transformed table

Table 6.12.: Example for data swapping
Source: based on Article 29 DP Working Party (2014)

inferences on a data subject. However, this inference might only be probabilistic since it is mostly not known which attributes have been swapped (Article 29 DP Working Party, 2014).

6.5.2. Randomization

Randomization (also called substitution or encoding) is about replacing data with simulated random values. Those values can but do not have to match the format of the original data (Mansfield-Devine, 2014). A simple example would be to replace *SMITH* with *X&T%#*. A suitable application scenario is, for example, to replace the unique or rare values of an attribute by randomization while retaining the other ones (Nelson, 2015). Hence, randomization is a useful privacy measure because it reduces the risk of linkability. However, the utility of the values is also completely lost, so the data is not of any value anymore. The use of the method, therefore, has to be critically evaluated.

6.5.3. Deterministic Encryption

Deterministic encryption describes a non-randomized encryption technique, which means that the same value always is transformed into the same ciphertext when using the same encryption key (ISO, 2018a). This property also qualifies the method as a suitable de-identification method. Its application can be targeted on explicit identifiers, quasi identifiers, and sensitive attributes. However, its analytical usage is often very limited or non-existent. Two variants that enable the preservation of specific utilities (namely order-preserving encryption and format-preserving encryption) are described further below.

The application of this method generally limits the analytical functionalities on equality checking or search functions. By knowing the key, one can encrypt the search term and compare the resulting ciphertext with other values (ISO, 2018a). Due to its non-randomized nature, it is also still possible to join different tables together based on the same attributes. Encryption techniques are often computationally complex, but the algorithms are getting less and less expensive which supports the adoption of this method (Branco et al., 2016; Ciriani

et al., 2010; Prasser et al., 2018). Deterministic encryption offers a suitable measure against all kinds of attacks because an adversary needs to have access to the appropriate encryption key. (ISO, 2018a). However, this is also linked to the biggest shortcoming: the management of the keys at the human side can be a tough and delicate process. A loss of keys can cause significant consequences if it falls into the wrong hands (Branco et al., 2016; Ciriani et al., 2010). Hence, it has to be carefully evaluated if the application of deterministic encryption makes sense.

Order-preserving Encryption

Order-preserving encryption is a specific type of deterministic encryption that has the property of retaining the order of values before and after encryption when using the same encryption key (ISO, 2018a). Thus, it enables a higher utility of the data. It extends the capabilities of deterministic encryption with range searches and the analysis of frequencies (ISO, 2018a). However, possible applications are limited to scenarios where the ordering of values is of high importance. An exemplary scheme for order-preserving encryption is called OPES (Agrawal et al., 2004).

Format-preserving Encryption

Format-preserving encryption is another sub-type of encryption that retains the data format and lengths of the original data after the transformation. For example, when encrypting a 9-digit social security number, another sequence of 9 digits is obtained (ISO, 2018a). This method can be used for systems where specific data formats are required. In general, the analytical usefulness and utility of the data is very limited.

6.5.4. Homomorphic Encryption

Homomorphic encryption is a specific type of encryption that allows computations on the encrypted data. The results can then be accessed after decrypting the calculated values. (El-Yahyaoui & Ech-Chrif El Kettani, 2018). This leads to several interesting possibilities, especially in the field of cloud computing. The data will not need to be unencrypted in non-trusted environments, but computations can still be performed. Hence, cloud environments can be leveraged, while data privacy is still preserved (El-Yahyaoui & Ech-Chrif El Kettani, 2018; Will & Ko, 2015). Because of its properties, homomorphic encryption is often referred to as the holy grail of encryption (Micciancio, 2010; Sidorov & Ng, 2016). However, the technique is, especially due to its inefficiencies, not yet really suitable for most practical use cases (Alloghani et al., 2019).

Figure 6.4 shows the general functionality in a diagram. It shows that one can get from a message m to a calculated value f(m), either by executing the function, or by encrypting, evaluating (executing the function on the encrypted message) and decrypting again (Alloghani

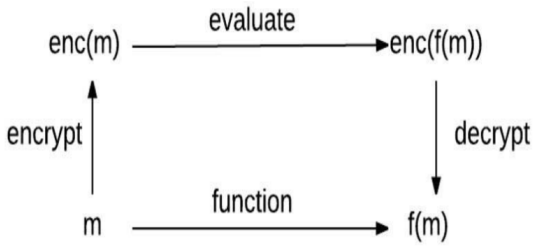


Figure 6.4.: Homomorphic encryption diagram
Source: El-Yahyaoui et al. (2018)

et al., 2019; Micciancio, 2010). There are two different types: fully homomorphic encryption (FHE) and partially homomorphic encryption (PHE).

FHE is theoretically the best approach and is capable of solving multiple problems in the areas of privacy and security. It allows one to perform any type and number of computations over the encrypted data (El-Yahyaoui & Ech-Chrif El Kettani, 2018). In research, a distinction is made between additive homomorphic encryption, which allows additive operations, and multiplicative homomorphic encryption, which allows multiplicative operations on data. Only if both properties are satisfied simultaneously, an algorithm is called fully homomorphic (Tebaa & Hajji, 2014). Also, the number of operations is not restricted in any way (Gaidhani, 2017). The potential of this technique is immense since it would allow users to perform operations on data in cloud data centers while fully preserving privacy (Chatterjee & Sengupta, 2018; El-Yahyaoui & Ech-Chrif El Kettani, 2018). However, as already mentioned, due to the low computational efficiency, it has not been widely applied yet (Wang et al., 2017). Somewhat homomorphic encryption (SHE) is another subclass of FHE, which has a limit on the number of operations that can be performed (Gaidhani, 2017).

PHE allows only one possible operation like multiplication or addition, but not both (Ogburn et al., 2013). Multiple schemes for PHE exist that are usable in practice. Due to its focus on only one operation, it is already more widely adopted than FHE.

There are several different schemes for homomorphic encryption, with the most important ones being the one from Rivest in 1978 and the one from Gentry in 2009 (Wang et al., 2017; Will & Ko, 2015). The most mentioned ones are the following:

- PHE (additive): Paillier, Goldwasser-Micali (Tebaa & Hajji, 2014)
- PHE (multiplicative): RSA, El Gamal (Tebaa & Hajji, 2014)
- SHE: Boneh-Goh-Nissim, SYY (Gaidhani, 2017; Tebaa & Hajji, 2014)
- FHE: Gentry, AHEE, EHES (Gaidhani, 2017; Khalid El Makkaoui et al., 2016; Tebaa & Hajji, 2014)

However, these schemes are very inflexible because they were tailored to specific use cases. There is no algorithm that fits for all scenarios. Therefore, it always has to be adapted on the

specific needs. Besides that, time and space complexity is the most crucial shortcoming of these techniques (Gaidhani, 2017; Prasser et al., 2018). Will et al. describe the methods as "[...] a balancing act between utility, protection, and performance." (Will & Ko, 2015) These reasons lead to the fact that real applications (especially of FHE) are years away from a scalable use in the cloud. This will require a significant advancement in this field through research (Will & Ko, 2015).

6.5.5. Creating Pseudonyms

Creating pseudonyms, often also referred to as pseudonymization, means to replace the values of an attribute with other ones (the pseudonyms) (Article 29 DP Working Party, 2014; Pfitzmann & Hansen, 2010). This process is often done on unique attributes or unique identifiers. Generally, this can be performed in two ways: reversible, which means the pseudonym is dependent on the input value, and irreversible, which means it is not derived from the original value (Article 29 DP Working Party, 2014; Nelson, 2015). There are several different techniques for the creation of such pseudonyms:

- Tokenization: describes a process where the pseudonym is a randomly-generated value (token), which has no mathematical relationship with the original value. For some specific contexts, which require synchronized tokens across multiple systems, this technique might not be suitable (Bourka & Drogkaris, 2018, p. 28).
- Hash function: refers to a function that transforms one value to another one (hash) without the possibility to reverse the process. However, by hashing all possible input values this can potentially be reversed (Article 29 DP Working Party, 2014).
- Keyed hash function: describes hash functions whose output also depends on the addition of a secret key which increases the protection level. The key might also be deleted, which would make the function equivalent to tokenization (Bourka & Drogkaris, 2018, p. 23).
- Symmetric encryption (secret key): refers to the creation of a pseudonym by encrypting the value with a secret key. The key holder can then always obtain the original value with a simple decryption process (Article 29 DP Working Party, 2014). When the data controller holds the secret key, he always has the possibility to derive the original data.
- Public key (asymetric) encryption: involves two keys in the process, the public and the private key. Anyone can encrypt data with the public key. However, only the secret key owner (usually the data subject) can decrypt the data (Bourka & Drogkaris, 2018, p. 26).

Next to these techniques, other de-identification methods, like character masking, generalization, character scrambling, or truncation, can theoretically also be used for creating pseudonyms. Due to the lack of application scenarios we will not consider these in this context.

One of the shortcomings of creating pseudonyms is that the method does not reduce the risk of singling out a data subject when quasi-identifiers remain (Article 29 DP Working Party, 2014). It only mitigates the risk regarding the specific attribute on which it is applied with regards to linkability. Therefore, it usually has to be combined with other de-identification methods.

6.5.6. Character Scrambling

Character scrambling means to mix or rearrange the order of characters of an attribute value (Bourka & Drogkaris, 2018, p. 28). Examples are to transform a credit card number from '4678 3412 5100 5239' to '0831 6955 0734 4122' or to change the name *SMITH* to *TMHIS*. This process can be easily reversed and is, therefore, not an effective way to protect an attribute value (Bourka & Drogkaris, 2018; Nelson, 2015). Also, the utility of the transformed data is almost zero (with the exception of counting characters), whereas the application of this method has to be critically assessed.

6.5.7. Microaggregation

Microaggregation (also: averaging) is a method that replaces individual values with aggregated values (the average) in a certain way (ISO, 2018a, p. 19). Different groups of the original values are formed, for which the aggregated values are then calculated. The intention is that the elements of one group or cluster are as similar as possible (Domingo-Ferrer & Mateo-Sanz, 2002).

Microaggregation is typically applied to quasi identifiers, and it is mainly used on numerical values. For numerical data, different conditions on group size and statistical characteristics can be set. For non-numerical attributes, the application is not that straightforward because it is harder to define appropriate aggregation operators. However, it is generally also possible if it is conceptualized in the right way (Martínez et al., 2012). The records with the closest values of the attribute should be in the same group, and for each group, there should be at least k records to fulfill k-anonymity. Hence, there is a condition on the group size but not on the number of groups (Domingo-Ferrer & Mateo-Sanz, 2002; ISO, 2018a). Table 6.13 shows the exemplary application of microaggregation on the attribute *age* with k=3. Three different groups with similar age ranges were created, all of them with at least three records. The values are replaced with the respective averages of each group.

Microaggregation is, in combination with k-anonymity, an advantageous method to preserve an individual's privacy. It is highly dependent on the clustering. More similar records in a group lead to higher remaining utility, and groups should be chosen such that no individual dominates (Domingo-Ferrer & Mateo-Sanz, 2002). Considering numerical values, this method comes pretty close to generalization. Despite the perturbation of values, attribute means are preserved, and the covariance is only moderately damaged. However, the

Age	Gender	Disease
27	Male	None
49	Female	Cancer
43	Female	Heart Disease
24	Male	Cancer
22	Female	None
29	Male	Heart Disease
43	Male	Cancer
39	Female	Heart Disease
38	Male	None
32	Male	None

(a) Original table

Age	Gender	Disease
25.5	Male	None
45	Female	Cancer
45	Female	Heart Disease
25.5	Male	Cancer
25.5	Female	None
25.5	Male	Heart Disease
45	Male	Cancer
36.3	Female	Heart Disease
36.3	Male	None
36.3	Male	None

(b) Transformed table with microaggregation on age with k=3

Table 6.13.: Example for microaggregation

application is computationally more expensive because the clustering leads to a quadratic complexity (Domingo-Ferrer et al., 2019).

6.5.8. Noise Addition

Noise addition describes a method that modifies data by adding random noise on selected continuous attributes. The noise is added in a way such that properties like mean, variance, standard deviation, and covariance are retained as much as possible (ISO, 2018a). It is mostly applied to numerical and date values. For non-numerical attributes, it is fairly hard to achieve (Mansfield-Devine, 2014).

There are a lot of different noise addition algorithms. However, the added noise should always be dependent of the value of an attribute. Examples are to change a student's grade from 3.33 to 3.53 or to adapt an individual's height accuracy to +/-10cm (Article 29 DP Working Party, 2014; ISO, 2018a). The privacy model which is usually used in combination to determine the level of noise is differential privacy (see subsection 6.3.4).

Noise addition can be applied to quasi identifiers and sensitive attributes. Hence, it can avoid identity as well as attribute disclosure. It has linear computational complexity and is an efficient de-identification method for preserving statistical features (Domingo-Ferrer et al., 2019). However, the achieved value highly depends on the added level of noise. The transformed values are not truthful anymore and they deviate with more noise added (Article 29 DP Working Party, 2014, p. 12). Furthermore, it might be possible to recreate the original values since there exist methods in the area of signal processing to remove the noise (Nelson, 2015).

6.6. Applying De-Identification Methods & Privacy Models

In this chapter, 14 different de-identification methods and the four most common privacy models were identified and described. As these should be applied together, Table 6.14 summarizes the methods and their respective models.

	(Sub-) Sampling	Suppression	Generalization	Character masking	Creating pseudonyms	Data swapping	Randomization	Deterministic encryption	Homomorphic encryption	Rounding	Top and bottom coding	Microaggregation	Noise addition	Character scrambling	Truncation
k-anonymity	-	X	X	X	-	-	-	-	-	X	X	X	-	-	X
l-diversity	-	X	X	X	-	-	-	-	-	X	X	X	-	-	X
t-closeness	-	X	X	X	-	-	-	-	-	X	X	X	-	-	X
Differential privacy	-	-	-	-	-	-	-	-	-	-	-	-	X	-	-

Table 6.14.: De-identification methods & privacy models

The de-identification methods were described with their strengths, weaknesses, and possible application scenarios to serve as a guide for choosing the right ones. However, this decision is a non-trivial task and is highly dependent on the given context and use-case (Article 29 DP Working Party, 2014; Tomashchuk et al., 2019). Especially the desired levels of privacy and utility, which act as contradicting elements, have an essential effect on this decision process.

7. Use Case of Wrist-Worn Wearable Data

In this section, we describe the use case for which we aim to develop and evaluate a concept for the application of de-identification methods. The first section serves as a general description of the idea behind wrist-worn wearable data. Then we identify relevant privacy threats and investigate which ones can be solved with de-identification methods. Lastly, we present a generic data model for wrist-worn wearable data and analyze its identifiers. The data model will serve as the basis for the developed concept.

7.1. General Description

Wrist-worn wearables are devices like smartwatches that individual people wear on their wrist to collect information through various sensors. They steadily collect sensitive data about the user's heart rate, blood oxygen saturation, sleep, and steps. Recent devices can even conduct clinically tested ECGs (Bondel et al., 2020). Besides that, they can be used to track trajectories during sports activities.

In combination with specific information about the user, this data is often transferred and stored in platforms offered by service providers. Those providers can be the manufacturers of such devices themselves, like Garmin, Runtastic, and Fitbit, but also ones that specialize specifically on such platforms like Strava (Statista, 2019). At the platform, the data of all users is combined and centrally stored, thus endangering the privacy of individuals. From a user perspective, the platform can be used to analyze the collected information, compare it with other users, and detect health issues like cardiac dysrhythmia, atrial fibrillation, and sleep apnea. On the other side, the service providers make use of the data to achieve improvements for its products and the platform itself.

The most significant risk of such a platform is that it can lead to a disclosure of personal information, including sensitive health data, as well as location data that can allow one to identify an individual's home and working place. The platform providers can generally not be considered trustworthy and despite very general privacy policies, there is limited information about the storage and processing of data. The providers might even sell the data to third parties in the healthcare or insurance industry.

We will develop a concept to apply de-identification methods on the data that wrist-worn wearables generate. This concept shall ensure high levels of privacy while at the same time enabling the benefits of data analytics.

7.2. Identification of Privacy Threats

The purpose of this section is to identify possible privacy threats for the use case of wrist-worn wearables. For these privacy threats, we then analyze whether they can be solved or mitigated by the application of de-identification methods.

In section 2.2, a threat-based approach and definition to conceptualize privacy was presented. This section builds on top of this approach to identify relevant threats for our specific use case. Therefore, we make use of the threat modeling methodology LINDDUN, which was proposed by Deng et al. The framework recently gained attention in the privacy community as a useful tool for threat modeling (Wuyts et al., 2018). There are several advantages over other methodologies. First, it is threat-based and therefore fits well with our privacy definition. Secondly, it is a systematic and step-wise methodology, and lastly, it also supports with an extensive knowledge base on common threats in the form of a threat tree catalog (Wuyts et al., 2018).

General procedure

The LINDDUN methodology is based on STRIDE - an approach for security threat modeling by Microsoft - and consists of six consecutive steps (Robles-González et al., 2020). They are visualized in Figure 7.1. Within the scope of this thesis, we will focus solely on the first three steps, which are situated in the problem space. Hence, their focus lies on the identification of threats (Wuyts et al., 2018). The steps within the solution space are not considered here. Instead, we are using the resulting threat scenarios and trees to investigate whether they are potentially solvable with de-identification methods or not.

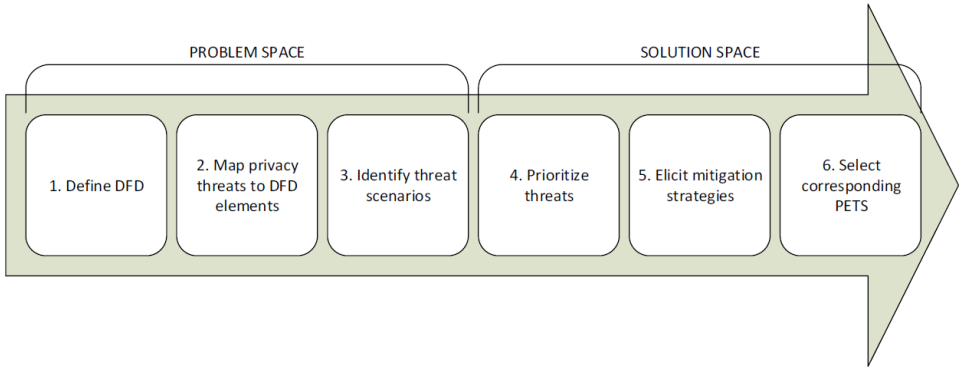


Figure 7.1.: Process of LINDDUN methodology
Source: Robles-González et al. (2020)

In the first step, an information flow oriented model of the use case needs to be defined. The framework proposes the data-flow diagram (DFD), which is a standardized notation to visualize an information system (Deng et al., 2011). This model is then leveraged to map privacy threats to the different elements of the data-flow diagram. Secondly, the methodology

provides a predefined list of privacy threat types and also a mapping table between the threats and the elements (Deng et al., 2011). For the third step, LINDDUN then proposes an extensive catalog of threat tree patterns related to privacy, which can be used to analyze the threats in more detail (Deng et al., 2011).

The term LINDDUN is an acronym, and each letter symbolizes one specific privacy threat. Those threats are obtained by negating privacy properties (Deng et al., 2011). Table 7.1 lists the seven considered privacy threat types and the corresponding privacy properties. Their definitions were already stated in Table 2.4 in section 2.2.

Privacy properties	Privacy threats
Unlinkability	Linkability
Anonymity & Pseudonymity	Identifiability
Plausible deniability	Non-repudiation
Undetectability & unobservability	Detectability
Confidentiality	Disclosure of information
Content awareness	Content unawareness
Policy and consent compliance	Policy and consent non-compliance

Table 7.1.: Privacy properties & privacy threats
Source: Deng et al. (2011)

Modeling the Use Case with a Data-flow diagram

The graphical representation of our use case of wrist-worn wearables using a DFD is represented in Figure 7.2. Generally, a DFD consists of four different types of elements which are shown in the legend of the figure: "data flows (i.e. communication data), data stores (i.e. logical data or concrete databases, files, and so on), processes (i.e. units of functionality or programs), and external entities (i.e. endpoints of the system like users, external services, and so on)" (Deng et al., 2011).

The specific elements used in the shown diagram can be described as follows:

- **User (entity):** The user describes an individual person wearing and using a wrist-worn wearable (e.g. smartwatch). The user refers to the data subject whose privacy we aim to protect.
- **Wearable (entity):** The wearable itself is the physical device the user is wearing on its wrist to collect different types of user-related data.
- **Platform & service (process):** This element is fully operated by a platform provider and contains the analysis and aggregation of the data. It can be represented by a browser-based web portal and/or a mobile application.
- **Provider database (data store):** The database stores all the information collected from

the user and its wearable. It is the bases of the operations within the "platform & service" element.

- **User-wearable data stream (data flow):** This data flow symbolizes the physical connection between user and wearable (e.g. heart rate tracking) as well as the user inputs while using the device.
- **User-platform data stream (data flow):** This element represents the direct information exchange between the user and the wearable platform, including manually entering information and requesting analyses.
- **Wearable-platform data stream (data flow):** The wearable communicates the collected sensor data to the platform to enable the analysis and aggregation of the data.
- **Platform-database data stream (data flow):** This data flow represents the communication between the platform and its central database.

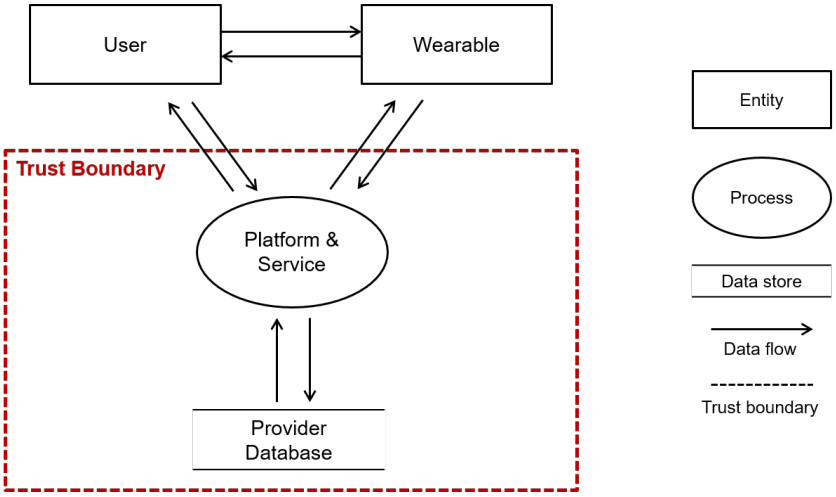


Figure 7.2.: Data-flow diagram of wearable use case

Additionally, a trust boundary is introduced to indicate trustworthy and untrustworthy elements (Deng et al., 2011). The 'platform & service' process and the 'provider database' both lie within the trust boundary since they are entirely operated and controlled by the platform provider. They might be located in a public cloud, so we cannot trust these elements at all. The user and the wearable itself are assumed to be fully trustworthy until their data gets shared with the platform.

Mapping privacy threats to the DFD

The LINDDUN methodology provides an identification of the privacy threat categories for each DFD element by following a mapping table (Deng et al., 2011). Table 7.2 shows the

predefined mappings applied to the DFD elements of the wrist-worn wearable use case. Each *X* indicates a potential privacy threat in the system.

	Threat target	L	I	N	D	D	U	N
Data Store	Provider Database	X	X	X	X	X		X
Data flow	User-Wearable data stream	X	X	X	X	X		X
	User-Platform data stream	X	X	X	X	X		X
	Wearable-Platform data stream	X	X	X	X	X		X
	Platform-Database data stream	X	X	X	X	X		X
Process	Platform & Service	X	X	X	X	X		X
Entity	User	X	X				X	
	Wearable	X	X				X	

Table 7.2.: Threat mapping for the DFD elements
Source: based on Deng et al. (2011)

Analyzing privacy threats via threat trees

For each privacy threat in the depicted table, LINDDUN presents an extensive catalog of threat tree patterns, which serves for a more detailed investigation of these threats within a realistic system (Deng et al., 2011). For the scope of this thesis, we make use of the publication "LIND(D)UN privacy threat tree catalog: CW Reports" by Wuyts et al., which provides the most recent version of the threat three catalogs. For each threat tree, we analyze whether the application of our de-identification approach with the assumptions stated above will either solve or mitigate this threat. Therefore, all concrete threats within the trees were investigated.

All for this analysis relevant threat trees are displayed in section A.2 of the appendix. In the following, the analysis of one exemplary threat tree (linkability of a data store) is explained in detail. The respective tree is shown in Figure 7.3.

From the tree, one can derive that the linkability threat occurs in a data store when there exists weak access control to the database together with insufficient minimization of the data. The second leaf node (insufficient minimization) can be mitigated by applying de-identification methods because that is exactly what these methods aim for. De-identification can not only reduce the linkability of the data to other external or internal databases, but it is also capable of reducing the overall amount of data available (e.g., by applying sampling, suppression, or redaction). Both are depicted as root causes within this threat tree. The green box in the figure visualizes the mitigation potential by de-identification that was mentioned above. In contrast to that, weak access control and the resulting disclosure of information is not directly addressed by de-identification. However, due to their conjunctive relation, a mitigation of the minimization threat indirectly reduces the potential impact of information disclosure. The orange box in the figure indicates this finding.

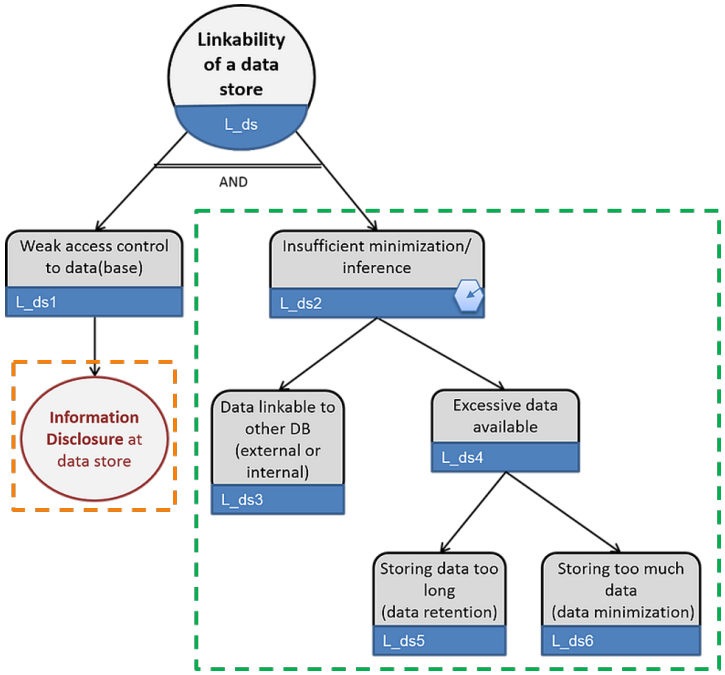


Figure 7.3.: Linkability of data store (provider database)
Source: Wuyts et al. (2014)

The same procedure was applied to all threat trees. Hence, the same annotation with green and orange boxes is also used in the further figures in section A.2 (analogous to the legend of the following table). Table 7.3 summarizes the findings of this analysis and indicates the threats which can be directly or indirectly influenced by de-identification methods. Threats regarding the data flow "user data stream" (between user and wearable) cannot be mitigated since the de-identification will take place after the data is collected (more details can be found in the following chapters). The same also accounts for the processes "user" and "entity" because these might be harmed in other ways, e.g., through re-used logins or weak passwords/usernames.

In total, ten directly solvable and eight indirectly influenceable threats were identified (considering that *policy and consent non-compliance* applies to the system as a whole). The table shows that de-identification methods can directly mitigate the threats linkability and identifiability, whereas disclosure and non-compliance are affected more indirectly.

	Threat target	L	I	N	D	D	U	N
Data Store	Provider Database	X	X	X	X	X		X
Data flow	User-Wearable data stream	X	X	X	X	X		X
	User-Platform data stream	X	X	X	X	X		X
	Wearable-Platform data stream	X	X	X	X	X		X
	Platform-Database data stream	X	X	X	X	X		X
Process	Platform & Service	X	X	X	X	X		X
Entity	User	X	X				X	
	Wearable	X	X				X	

X Mitigation through de-identification possible

X Mitigation through de-identification not possible, but the impact of the privacy threat is indirectly reduced through mitigation of other threats

X Mitigation through de-identification not possible

Table 7.3.: Privacy threat mitigation with de-identification methods

7.3. Wrist-Worn Wearable Data Model

For the development of an approach for de-identification, a data set or data model is needed to conceptualize the application of this methods. As data sets of wrist-worn wearable data are not publicly available, a generic wrist-worn wearable data model was created to support the concept development. The basis for this data model was formed by requesting data exports of individual user accounts from the platforms of Garmin, Apple Watch, and Fitbit. These exports contain information relating to the data stored of one single individual. This gave us detailed insights into how the wearable information is stored and how it is formatted. Out of this information, a generic data model for the wrist-worn wearable use case was derived. The data model is shown in Figure 7.4. Each box refers to one table with a specific structure. There are seven tables in total: one master data table, and six raw sensor data tables. The connections between the tables visualize how they can be joined or linked together. The master data tables store general information about the user itself, which is often entered directly by himself. This data does not change frequently. In contrast to that, the raw sensor data tables contain the sensor data about the user, which is frequently collected by the wrist-worn wearable. An overview of these seven data tables and their columns, corresponding data types, and example values can be found in section A.1 of the appendix.

The user data table contains the following user-specific information: a unique ID, email address, first and last name, the country where the user lives, gender, date of birth, body height, handedness (indicating whether the user is right- or left-handed), a profile image, the gear he is currently using (e.g., running shoes), the creation date of the account, body weight, and the maximal oxygen consumption (VO_2max). These attributes are mainly used to compare

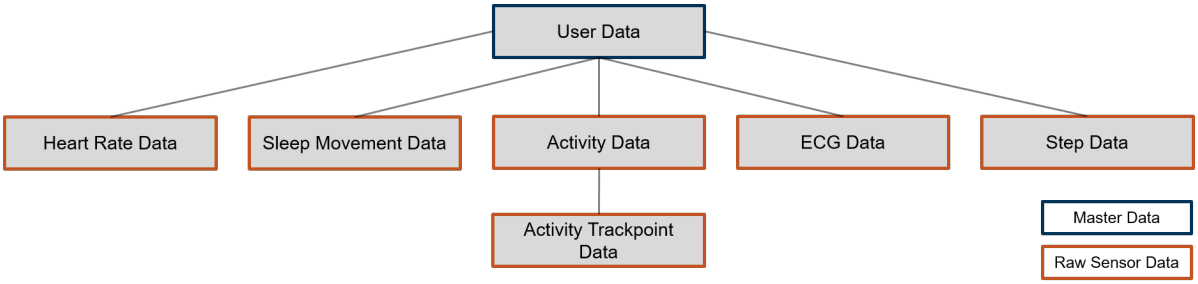


Figure 7.4.: Wrist-worn wearable data model

different users with each other based on different criteria. The reveal of information like body weight and maximal oxygen consumption, which indicates one's endurance fitness, may be undesirable for a user. The information in the heart rate data table provides information about the user's heart rate at a given time, which may imply details about their activities. The user's sleep movement is tracked by indicating a relative activity or movement level for one minute. This can be leveraged to detect sleep apnea but also to determine one's sleeping hours. Information about tracked activities is stored in the activity data and activity trackpoint data table. They contain information about the type of the activity, as well as time, GPS coordinates, elevation, heartrate, and cadence of each recorded trackpoint. The user benefits with detailed insights into his activities, but it could also be used to determine his home or working place. In case an ECG is performed, a time series of around 15,000 data points indicating the voltage and derivation related to the hearth rhythm are collected. This data is used to detect cardiac dysrhythmia and atrial fibrillation, which is very sensitive information for an individual. Lastly, the step data table provides information about the steps and floors a user covered within one minute.

Table 7.4 indicates the number of records collected per user for each column in the data tables. When the data exports of the considered providers were using different frequencies (e.g., Garmin stores the heart rate for each 2 minute interval, whereas Fitbit saves this information every 5 seconds), the approach with the higher frequency was chosen.

Data table	Type	# of columns	# of attributes per User
User data	Master data	14	1
Heart rate	Raw sensor data	3	1 per 5 sec
Sleep movement	Raw sensor data	4	1 per min during sleep
Activity	Raw sensor data	4	1 per activity
Activity trackpoint	Raw sensor data	7	1 per sec during activity
ECG	Raw sensor data	4	15,258 per ECG
Step	Raw sensor data	4	1 per min

Table 7.4.: Overview of Data Tables

To illustrate the amount of data which is collected for a user during one day, the information from Table 7.4 was added up. Since the amount of data in the master data tables is relatively

small and stays stable over time, these tables were not considered in this calculation. The following formula expresses the number of attributes a_{day} per day and user. It is dependent on the number of sleeping hours d_s, the number of activities per day a, the duration of an activity in minutes d_a, and the number of ECGs per day e.

$$a_{day} = 57,600 + 240 * d_s + 4 * a + 420 * a * d_a + 61,032 * e$$

Assuming that a user sleeps 8 hours ($d_s = 8$), tracks two activities a week ($a = 2/7$) with a duration of 60 minutes ($d_a = 60$) and performs one ECG per week ($e = 1/7$), this leads to 75,440 collected attributes a day. Calculating with an average attribute storage size of 20 Bytes, this leads to 1.43 Megabytes per user and day. This number may sound rather small, but with an exemplary number of users of 50 million, it already sums up to more than 68 Terabytes per day. It also has to be taken into account that only the raw sensor values without any calculation or analysis are considered.

7.3.1. Determination of Identifiers and Sensitive Attributes

The goal of this subsection is to identify *Explicit Identifiers*, *Quasi Identifiers*, *Sensitive attributes*, and *non-sensitive attributes* within the defined data model. Jung et al. propose a suitable determination scheme for quasi identifiers. Its original purpose is based on clinical data. However, the classification approach can be applied in a more general way. The first step of this scheme includes classifying the data columns based on the process shown in Figure 7.5.

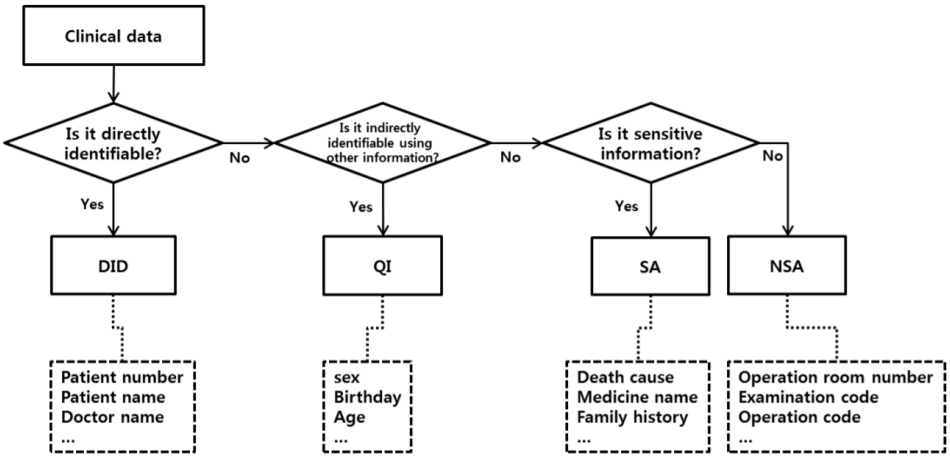


Figure 7.5.: Data classification process
Source: Jung et al. (2020)

Jung et al. use the term *Direct Identifiers* and its abbreviation *DID*. However, we use *EI* and *Explicit Identifiers* instead, like it was defined in section 2.1. Strictly following the process indicates that a data column always gets uniquely classified into one of the four categories and hence that those are non-overlapping. However, we propose the assumption that the

two identifying categories (EI and QI) and the two sensitivity categories (SA and NSA) need to be considered separately from each other. This is because some attributes can be quasi identifiers while being sensitive at the same time. An example is the attribute *weight*, which we classify as both. Therefore, after classifying information as an *Explicit Identifier* or a *Quasi Identifier*, we nevertheless check if the attribute can also be referred to as a *sensitive attribute*. If this is not the case, the category *non-sensitive attribute* is not considered since it does not bring additional value. Hence, an attribute can only be classified as a NSA exclusively.

In the proposed generic data model, some attributes serve only as QI or SA when combined with other specific attributes, whereas they alone are classified as NSA. Therefore, we replace these attributes with matrices containing the related elements as part of this process. Table 7.5 shows the resulting classifications for the data columns of the wrist-worn wearable data model.

Data table	Column	Classification
User data	userID	EI
	eMail	EI
	firstName	QI
	lastName	QI
	country	QI
	gender	QI
	birthDate	QI
	height	QI
	handedness	QI
	profileImage	QI
	currentGear	QI
	createdDate	QI
	weight	QI & SA
	vo2Max	SA
Heart rate	userID	EI
	$\begin{pmatrix} \text{timestamp} \\ \text{heartrate} \end{pmatrix}$	SA
Sleep movement	userID	SA
	$\begin{pmatrix} \text{startTime} \\ \text{endTime} \\ \text{activityLevel} \end{pmatrix}$	SA
Activity	userID	EI
	activityID	NSA
	time	QI & SA
	type	QI & SA

continued on next page

continued from previous page

Data table	Column	Classification
Activity trackpoint	activityID	NSA
	$\begin{pmatrix} \text{time} \\ \text{latitude} \\ \text{longitude} \\ \text{elevation} \end{pmatrix}$	QI & SA
	heartrate	SA
	cadence	SA
ECG	userID	EI
	startTime	SA
	voltage	SA
	derivation	SA
Step	userID	EI
	$\begin{pmatrix} \text{time} \\ \text{steps} \\ \text{floors} \end{pmatrix}$	SA

Table 7.5.: Classification of wrist-worn wearable data attributes

7.3.2. Investigation of Quasi Identifiers

In the previous subsection, 14 quasi identifiers in four different data tables were identified. This section serves to identify and evaluate the identifiability of these quasi identifiers.

Jung et al. propose two meaningful measures to calculate re-identifiability scores for an attribute. The *uniqueness value* of an attribute is expressed by the ratio of the number of unique values to the total number of different values (Dankar et al., 2012; Jung et al., 2020). Hence, if the uniqueness value is 0, a specific individual cannot be identified just with this information. If the value is not 0, the attribute has at least one distinct value (Jung et al., 2020). The *inference value* compares the number of equivalence classes (records with the same values on the quasi identifiers) when a specific attribute is excluded to the number of classes of the entire table. A significant difference indicates a high level of influence of that attribute which leads to an increasing re-identification risk (Jung et al., 2020). El Emam suggests two risk metrics that apply to a complete record of a data set instead of a single attribute. If a potential adversary knows whether a data subject is in that data set, the *prosecutor risk* can be used. It describes the probability of re-identification and is calculated by dividing through the size of the matching equivalence class in the data set (El Emam, 2013, p. 186). If it is not known whether the individual is in the data set, the *journalist risk* applies which is calculated by dividing through the number of records in an identification database (a superset of the database) (El Emam, 2013, p. 186-188). These two concepts, which are closely related to k-anonymity, are also used in implementations and tools within organizations for risk assessment purposes (I12). Table 7.6 summarizes the different risk metrics.

Metric	Calculation	Level
Uniqueness value	$\frac{\text{number of unique values of attribute}}{\text{total number of different values of attribute}}$	attribute
Influence value	$1 - \frac{\text{number of equivalence classes without attribute}}{\text{number of equivalence classes}}$	attribute
Prosecutor Risk	$\frac{1}{\text{size of matching equivalence class}}$	record
Journalist Risk	$\frac{1}{\text{size of equivalence class in identification database}}$	record

Table 7.6.: Metrics for risk analysis
Source: based on Jung et al. (2020), El Emam (2013)

However, these measures can only be used to evaluate the risks of quasi identifiers when there is a complete data set available. Since our basis is a data model with a defined structure but without explicit values, we cannot calculate measures like the uniqueness and inference value. To cope with that issue, we identified the amount of distinct values an attribute can take (also called cardinality) as a suitable alternative measure. It takes into account the definition and properties of an attribute and is related to the mentioned metrics for risk analysis. The numbers of distinct values for the attributes of the examined use-case are presented in Table 7.7.

Generally, a higher number of distinct values also implies a higher risk regarding a possible re-identification. This is based on the fact that the number represents the level of detail that can be associated with a data subject. If we assume a uniform distribution of the values within an attribute, a higher cardinality leads to a higher uniqueness value and also a higher prosecutor and journalist risk. Of course, this distribution is not entirely accurate in practice. Taking the date of birth into account, it is unlikely that there are as many smartwatch users with the age of 60 as with the age of 25. Considering the activity time, there will be more activities pursued during day time than in the night. However, we will still follow up on this approach to model the risk values as it is suitable to represent the level of detail in an unknown data set.

The so-called number of minimal sample uniques is another approach for risk estimation. It is based on the number of unique patterns of attribute values which can lead to the identification of individuals within a data set (Manning et al., 2008; Prasser et al., 2020). As it is only usable in case of availability of a full data set, we propose the following metric which is closely related to it:

> The attribute value combinations (AVC) represent the total number of possible value combinations for all attributes in a data set.

Hence, it is calculated by multiplying the numbers of distinct values (or value combinations)

Data table	Column	Distinct values	Assumption
User data	firstName	$\sim \infty$	
	lastName	$\sim \infty$	
	country	195	Source: xyz
	gender	3	
	birthDate	15,695	Users between 18 and 60 years old
	height	51	Heights between 150cm and 200cm
	handedness	2	
	profileImage	$\sim \infty$	
	currentGear	2,000	20 brands with 100 models each
	createdDate	3,650	Days within the last 10 years
	weight	81	Weights between 40kg and 120kg
Activity	time	5,256,000	Minutes within the last 10 years
	type	10	
Activity trackpoint	$\begin{pmatrix} \text{time} \\ \text{latitude} \\ \text{longitude} \\ \text{elevation} \end{pmatrix}$	$\sim \infty$	

Table 7.7.: Distinct values of Quasi Identifiers

of each attribute of the data set. It indicates the maximum possible unique records, which would be the case if each possible attribute value combination existed exactly once.

8. Requirements for Privacy-Enhancing Analytics of Wrist-Worn Wearable Data

This chapter deals with the identification of requirements for a concept for privacy-enhancing analytics of wrist-worn wearable data. We present general insights gained in the expert interviews. Afterwards, the resulting requirements are presented to address the second research question.

8.1. Findings of the Expert Interviews

The following section will give an overview of the obtained results in the conducted expert interviews. The insights are structured into different areas. Further results are stated also in the following chapter.

Impact of Data Privacy

Data privacy has an increasing impact on organizations and individuals already, primarily through the release of GDPR. However, it certainly does not yet have the status that it needs everywhere (I1). Especially for consumer goods, like wearables or smart speakers, users often see the great things that are possible, but of course, there are some major drawbacks in terms of privacy (I2). "It is not about the sale of such a smartwatch, it is about the data that is collected and with which business is to be done." (I1) Furthermore, such wearable service providers often operate in a grey area as they choose a location where they try to evade the enforceability of GDPR (I4).

Tools can help to identify and address privacy issues within IT systems. In practice, these are often simple so that different questionnaires are provided. Based on the results, a recommendation regarding the risks and potential measures is then given (I7). There are several tools and software on the market to support the compliance of data privacy. However, a proper risk analysis should always contain contextual information, whereas it is hard to implement a generic solution (I1). It would be beneficial, and it is an important topic to find such a solution for large amounts of data to prevent the possibility of drawing conclusions about individuals. Especially for topics like the Internet of Things, Artificial Intelligence and Cloud Computing, this often leads to problems (I3).

Potential of De-Identification

The potential of a concept using de-identification methods to preserve an individual's privacy is perceived to be very large (I1, I4, I7). Simple pseudonymization is not sufficient in most cases, so there is a need for this approach, and such solutions are on the rise (I4). Some people do not realize the extent of personal data they provide to cloud providers, which then can potentially be accessed by the cloud provider itself but also by unauthorized third parties, i.e., adversaries (I2). The ultimate goal of such a solution would be to transform the data in a way that it can no longer be considered as personal data with regards to the GDPR. As a result, there would be no data protection restrictions, and the individuals would no longer have a risk of abuse anymore (I4). Hence, the de-identification methods can be used to achieve a regulation compatible purpose (I9). In most cases, the full raw data is not needed anyway to obtain trends or statistics. To relate to the example of a smartwatch: It may only be interesting to know that someone has run three miles, but not where he has run those (I7). The key of such de-identification methods is to identify how to set the re-identification risk to zero (I3). The combination of methods on attributes given the constraints for privacy and information loss might by feasible to solve in an optimization problem (I12).

Suitable use cases can be found in healthcare, telecommunications, and banking for example (I4). When it comes to autonomous driving, such concepts will be needed very soon due to the mass of sensors and personal data which is collected (I7). However, real applications are rather rare up to now (I1, I2, I3, I7). Initial concepts already exist in the health and insurance sector (I1, I2). To drive the usage of these concepts a lot of sensitization will be needed (I1).

Amount of Data

The available amount of data is a crucial factor for risk evaluation. The more (anonymous, pseudonymous, or de-identified) data one collects, the easier it gets to link it to a person again (I7, I1). Combining the data with published data sets is one way to contextualize it. This allows one to narrow down eligible persons by detailing their profile (I1). If many more records of personal data are processed compared to records with special categories of data, the former ones can cause much more damage (I10). This is why the principle of data minimization should be considered to mitigate such risks (I10). Also, if the wearable data can be assigned to a smartphone, it might be easily linkable to information from other applications which were not designed in a privacy-friendly way (I7, I2).

Classification of data

Different approaches for the classification of data based on their criticality and risk of abuse exist (I1). The distinction of GDPR into personal data and special categories of personal data like health and politic related information often serves as a basis (I3, I6, I8). Our use case falls into the latter one as it includes health-related as well as GPS motion data. It is

essential to distinguish between unique information, like names, and additional information that can help to identify individuals. From a GDPR point of view, sensitive and non-sensitive attributes of personal data are treated equally. Former ones merely have a higher threshold to be processed (I3). Risk-based classifications into stage models (from high to low risk or from major to marginal impact) are often used to conceptualize measures accordingly (I6, I8). An evaluation of the exact content of an attribute is crucial for this purpose (I12). However, the classification has to be carefully considered together with the amount of data that is available (I1).

Risk evaluation

"The guiding principle is: no conclusions on individuals unless I have the appropriate consent" (I3). However, it is often not clear which methods exactly are sufficient for a de-identification (I5). Often, it is at an individual's discretion which data is critical. For some users, it may be valid to share GPS motion data, for others not. Ideally, there would be mature users that decide for themselves what they want to share and what not (I1). The central question should be: What implications does it have for individuals in case of a disclosure (I8, I12)?

Impact of Regulations

Regulations regarding data privacy are gaining more influence and are being tightened worldwide. The GDPR is pioneering in this development and serves as a basis for laws in China, Thailand, India, and California, which emerged in recent years (I6, I12). Especially the threat of punishment and the increase in fines have caused GDPR receiving great attention for organizations of all kinds (I1, I3). Hence, it is also highly relevant for providers of wrist-worn wearable services.

Generally, it is essential to distinguish between the two data categories personal data and special categories of personal data (I2, I4, I9). As the wearable data falls under the latter, processing requires explicit and demonstrable consent of the individual (I2, I4). Additionally, the data can only be processed if it is necessary for a specific purpose, which is in our case the platform service, including the analytical possibilities (I4, I9). Purpose limitation "[...] is part of the holy grail of data protection principles." (I9) Other important principles are need-to-know and data minimization that both advice to limit the amount of data as far as possible (I1, I3). An important aspect with regards to de-identification methods is the integrity of data, which implies that the data needs to be correct and accurate (I2, I12). For perturbative methods, this can lead to a challenge as these methods might alter the truthfulness of single values. The decisive question here is: does the application of a perturbative method like noise addition change the accuracy and correctness of data in such a way that it violates the GDPR? There is no precise answer to that as the regulation itself does not define when data can be considered as incorrect. One can argue that through the perturbation, it is not personal data

anymore. However, one needs to be careful regarding the application of perturbative methods and investigate a potential violation for each specific use case.

Two core principles of GDPR that are also crucial for the wearable use case are privacy by design and privacy by default (I1, I3, I4, I8, I10). Privacy by design indicates that privacy issues should be addressed in the product development of a software or tool. Hence, our concept should also mitigate possible risks by design, if possible. Privacy by default suggests that products should be configured by default in a way that the privacy of individuals is preserved (I3, I8). Ideally, these two principles are integrated in such a way that the data "[...] can no longer be considered as personal data under the GDPR." (I4)

For modern data privacy protection, it will be necessary to also bring the algorithms itself under control. This implies that these should be transparent to enable more straightforward impact and risk assessments. There is a lack of regulations with regards to this partial aspect of algorithms, but this will most likely come in the future (I1). Hence, we will include transparency as a requirement for the development of the concept.

8.2. Derivation of Requirements

In this section, the requirements of a concept for privacy-enhancing analytics of wrist-worn wearable data are specified and illustrated. In total, ten requirements were identified. They originate from

- the literature review on de-identification methods and privacy models (chapter 6),
- the conducted interviews and discussions with experts within the domain data privacy,
- and the findings related to the generic wrist-worn data model described in the previous chapter.

Requirement 1: Local transformation of data
The de-identification methods shall be applied locally on the wrist-worn wearable before transferring the data to the service provider.

Due to the specific nature of our use case, which incorporates multiple wearables as data sources, we aim to apply the de-identification techniques on the wearable devices itself. This implies that the source data is not shared with the service provider. Hence, the risk that this data gets disclosed is reduced (I7, I12). Only the transformed and de-identified data shall be transferred to the service provider. In the following chapter, we will further specify the implications that are caused by this requirement.

Requirement 2: k-anonymity enforcement
The concept shall incorporate the privacy model k-anonymity to conceptualize the re-identification risk.

The k-anonymity model is the basis for its extensions l-diversity, and t-closeness and it has not been thoroughly researched yet how it can be applied on a local basis. For this reason, we aim for a concept that incorporates local enforcement of k-anonymity. De-identification methods that work along with this privacy model, are preferably selected. Metrics like the *number of distinct values* and the *attribute value combinations (AVC)* will support the realization of this requirement.

Requirement 3: Privacy levels
The concept shall provide different levels of privacy dependent on the choice of the user which reflects his preference for the trade-off between privacy and utility.

The idea of different privacy levels is that every user might have a different perspective about how strong he wants to be protected and which kind of analytical functionalities he wants to use. Each user should decide for a specific level based on which de-identification methods are applied (I1, I8). However, a minimum level should be defined and set by default (I1).

Requirement 4: Generic wrist-worn wearable data model
The concept shall be explicitly designed for the generic wrist-worn wearable data model.

The wrist-worn wearable data model, which was described in chapter 7, will be used to conceptualize and evaluate the de-identification approach. The basis of this concept is required to be adaptable to other use cases as well.

Requirement 5: Compliance with regulations
The concept shall reflect the principles of privacy by design and privacy by default of the General Data Protection Regulation. The recommendations of the HIPAA Safe Harbor method shall also be taken into account.

The impact of regulations on data privacy was already illustrated earlier. The principles of privacy by design and privacy by default of the GDPR and the HIPAA Safe Harbor method were identified as the most critical aspects for this use case. The former ones relate closely to the provision of different levels and the definition of a minimum level in requirement 3 (I1, I3, I10). Due to its practical nature and its focus on health-related data, the HIPAA Safe Harbor method will also be considered (see subsection 2.2.2).

Requirement 6: Low performance overhead
The concept shall focus on de-identification methods with rather low computation times (linear complexity) to limit the performance overhead and enable scalability for Big Data.

The computation times are a significant factor for the choice and comparison of applicable methods (I5). De-identification methods with low performance overhead will be preferred to allow scalability and usability for Big Data sources. Furthermore, the processing power of wearables might not be able to deal with more complex computations like homomorphic

encryption.

Requirement 7: Constraints from a privacy perspective
The selection of appropriate de-identification methods shall be carried out based on constraints and requirements out of a privacy perspective instead of an analytical perspective.

In chapter 2, the trade-off between privacy and utility was illustrated. On the one hand, one wants to promote data privacy and limit the relation of data to persons. On the other hand, utility should be preserved as much as possible (I9). The concept shall be developed from a privacy point of view to ensure appropriate protection of the individual's privacy. Hence, the requirements and constraints are set to provide sufficient privacy. The analytical operations need to adapt to these constraints. This will create higher confidence among the users, which might again lead to more data that can be collected (I11).

Requirement 8: Protection against complete disclosure to any adversary
The concept shall protect against disclosure of the entire data set, i.e., internal and external adversaries that have access to the full data.

The concept is supposed to protect a scenario that involves disclosure or leakage of the complete data set, including all attributes and records. The type of adversaries is also not limited and includes internal as well as external adversaries of the system provider.

Requirement 9: Transparency
The combination of de-identification methods applied to different privacy levels shall be disclosed to enable full transparency for a user.

The Article 29 DP Working Party advices to disclose implemented anonymization techniques if a data set is released (Article 29 DP Working Party, 2014, p. 25). We follow this proposal and aim to provide a transparent overview of applied de-identification methods to the users. This will help the users to get an idea of the impact of these to determine which privacy level to choose.

Requirement 10: Transformation of identifiers
The concept shall focus on the transformation of explicit and quasi identifiers to reduce the re-identification risk.

Generally, de-identification methods can be applied to identifiers as well as sensitive attributes. The concept we aim for transforms the identifiers in a way that the re-identification risk is mitigated. The sensitive attributes will not be modified with this approach as the risk of allocation to an individual is decreased.

9. Concepts Using De-Identification Methods on Wrist-Worn Wearable Data

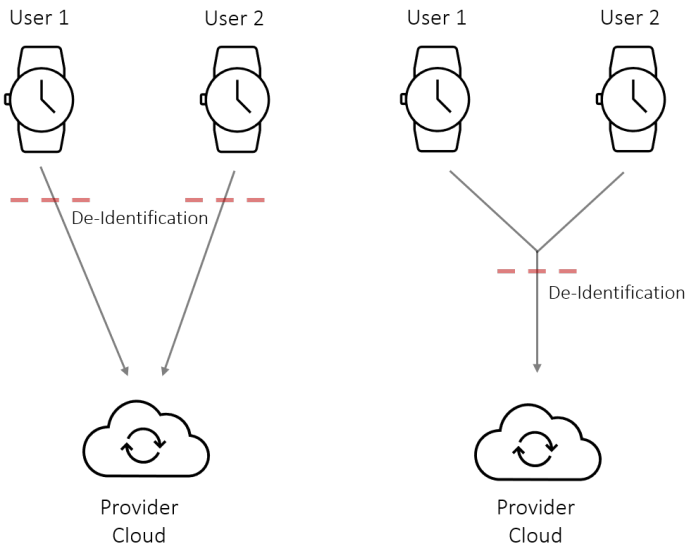
This chapter describes the development and evaluation of a concept for the application of de-identification methods on wrist-worn wearable data. The comprehensive literature review on de-identification methods in chapter 6, the description of the use case along with its generic data model in chapter 7, and the identified requirements in chapter 8 serve as the input for this concept. At first, we describe a suitable technical architecture and evaluate the applicability of different de-identification methods. We then develop a new local probabilistic k-anonymity concept that allows applying the privacy model at a local level. We propose a Monte Carlo simulation to verify certain privacy levels. Lastly, a critical evaluation of this concept is pursued by the analysis of different scenarios.

9.1. Technical Architecture

In Figure 9.1, two possible variants of a simplified technical architecture for a de-identification approach on wrist-worn wearable data are presented. Both represent the data flow of two smartwatches to a provider in the cloud. The red dotted lines indicate the point of de-identification.

For variant A, the de-identification takes place on the wearable itself before the data is transferred to the cloud provider. This has the advantage that no additional location is needed where the data is temporarily stored and processed. Hence, it would work with existing infrastructures. The computations would need to be done on the wearables or its corresponding smartwatches which should not be an issue anymore since only the data of one single wearable would need to be handled (I12). The disadvantage, however, is that one can use privacy models only in a limited way since the complete data set of other wearables is not available at the point of de-identification. For differential privacy, concepts coping with that issue (called local differential privacy) exist (Cormode et al., 2018; Fan et al., 2020). For k-anonymity and its extensions l-diversity and t-closeness, a local application is more limited due to the uncertainty of the complete data set.

With variant B, the de-identification takes place at a trustworthy point between the wearable and the cloud provider. The data from different wearables is merged and de-identification methods are applied on this complete data set. The data is transferred to the provider afterwards. This allows the best possible application of de-identification methods with all



(a) Variant A (b) Variant B

Figure 9.1.: Technical architecture variants

privacy models because the operations can be performed on the full data set. However, this additional point, which serves as an intermediary, has to be utterly trustworthy as it operates on the whole data and therefore has access to the individual's raw data. Furthermore, it represents an additional point of attack for adversaries. The requirements with regards to computation power are also very high since calculations on large amounts of data have to be performed.

Based on scientific publications and current implementations of de-identification methods, variant B is often assumed. But, due to the ease of compatibility with existing infrastructure in the wearable use case, the avoidance of an additional point of attack and the existing research gap for local k-anonymity, we will focus on variant A. This architecture is promising as the full source data is not transferred and wearables should be capable of performing such transformative operations (I7, I12). It also harmonizes with requirement 1 where a local transformation approach was identified.

9.2. Evaluation of Applicable De-Identification Methods

In this section, the local applicability of de-identification methods on the generic wrist-worn wearable data model is evaluated. As stated in requirement 10, we focus on the transformation of identifiers rather than sensitive attributes. Table 9.1 shows the de-identification methods that are applicable to the EIs and QIs of the wrist-worn wearable data model. We focus on the local application of the methods as stated in the previous section. In case it generally

makes sense to apply a method, the entry is marked with an *X*.

The *userID* is not listed as we decide to leave it untouched. We assume it to be only used in this context, so it is neither identifiable, nor sensitive. As the *eMail* is used as a login parameter, a transformation has to remain its uniqueness. *Latitude* and *longitude* are considered together since they represent one location point in combination, and they have the same properties regarding possible methods. The same accounts for *firstName* and *lastName*. For the applicability of generalization methods, we investigated if a suitable generalization hierarchy exists for the respective data column. In case there are only a few possible values (e.g., *gender*, *handedness*) and the next generalization step would lead to one category containing all values, the method is considered as not applicable. This accounts for all attributes with a generalization height of 2. All date columns can be easily transformed into numerical values. Therefore, several numerical methods are applicable to them. The same procedure applies for time columns. Attributes with values of only one character (e.g., *gender*) and columns with only a few possible attributes (e.g., *handedness*, *activity type*) are considered not to be applicable for character masking. Also, all data columns with a non-consistent length (e.g., *profileImage*, *currentGear*) are not considered applicable for character masking and the creation of pseudonyms. When an attribute has only one value per user, data swapping, sampling and microaggregation are not useful due to the local de-identification approach. In this case, swapping is not possible, sampling would be equivalent to a deletion, and microaggregation would not lead to any change of the value. Therefore, these methods cannot be applied to the attributes of the user data table. The creation of synonyms only makes sense for free-text fields like *firstName* and *lastName*.

Randomization, character scrambling and truncation can be applied on all data columns. However, we do not consider randomization because the remaining utility is limited and we suggest to use deletions instead as they do not introduce any false values. Character scrambling is also not useful for our use case for similar reasons. This method is only beneficial in very specific cases such as counting characters or creating test data. For truncation, we did not encounter suitable application possibilities targeted to wrist-worn wearable data. It does not provide adequate utility and is replaceable with other methods. As a result, we do not incorporate these three methods for our concept due to the mentioned shortcomings.

Noise addition is listed as applicable in many cases but will, however, not be further considered due to its unsuitability with the k-anonymity privacy model. Theoretically, homomorphic encryption is applicable to the attributes. Since the computations on that data are not yet practically applicable due to their complexity and immaturity, we do not incorporate them within this concept. The same accounts for deterministic encryption as well because we aim to avoid the necessity of storing a key.

Table	Column	(Sub-) Sampling	Suppression	Data swapping	Randomization	Deterministic encryption	Homomorphic encryption	Generalization	Rounding	Top and bottom coding	Microaggregation	Noise addition	Character masking	Creating pseudonyms	Character scrambling	Truncation
User data	eMail	-	X	-	X	X	X	-	-	-	-	-	-	X	X	X
	firstName lastName	-	X	-	X	X	X	-	-	-	-	-	X	X	X	X
	country	-	X	-	X	X	X	X	-	-	-	-	X	-	X	X
	gender	-	X	-	X	X	X	-	-	-	-	-	-	-	X	X
	birthDate	-	X	-	X	X	X	X	X	X	-	X	-	-	X	X
	height	-	X	-	X	X	X	X	X	X	-	X	-	-	X	X
	handedness	-	X	-	X	X	X	-	-	-	-	-	-	-	X	X
	profileImage	-	X	-	X	X	X	-	-	-	-	-	-	-	X	X
	currentGear	-	X	-	X	X	X	X	-	-	-	-	-	-	X	X
	createdDate	-	X	-	X	X	X	X	X	X	-	X	-	-	X	X
	weight	-	X	-	X	X	X	X	X	X	-	X	-	-	X	X
Activity	time	X	X	X	X	X	X	X	X	X	X	X	-	-	X	X
	type	X	X	X	X	X	X	-	-	-	-	-	-	-	X	X
Activity trackpoint	time	X	X	X	X	X	X	X	X	X	X	X	-	-	X	X
	latitude longitude	X	X	X	X	X	X	X	X	X	X	X	-	-	X	X
	elevation	X	X	X	X	X	X	X	X	X	X	X	-	-	X	X

Table 9.1.: Applicable de-identification methods on identifiers

9.3. Local Probabilistic k-anonymity

Due to the local de-identification approach, the incorporation of the privacy model k-anonymity (requirement 2) is not as straightforward as when it is applied globally. This is caused by uncertainty about the properties of the entire data set. In this section, we propose a local probabilistic k-anonymity concept to address this uncertainty. The concept will enable one to draw inferences about the expected privacy level within an unknown data set. A Monte Carlo simulation is used to verify the privacy levels.

The level of k-anonymity is dependent on three major factors: the number of records, the number of quasi identifiers, and the distribution of values within the attributes. For this model, we assume that the number of records, quasi identifiers, and distinct values per quasi identifier are known. Hence, the attribute value combinations (AVC) can be calculated by multiplying the number of distinct values of each attribute. For a number of quasi identifiers a, this accounts to

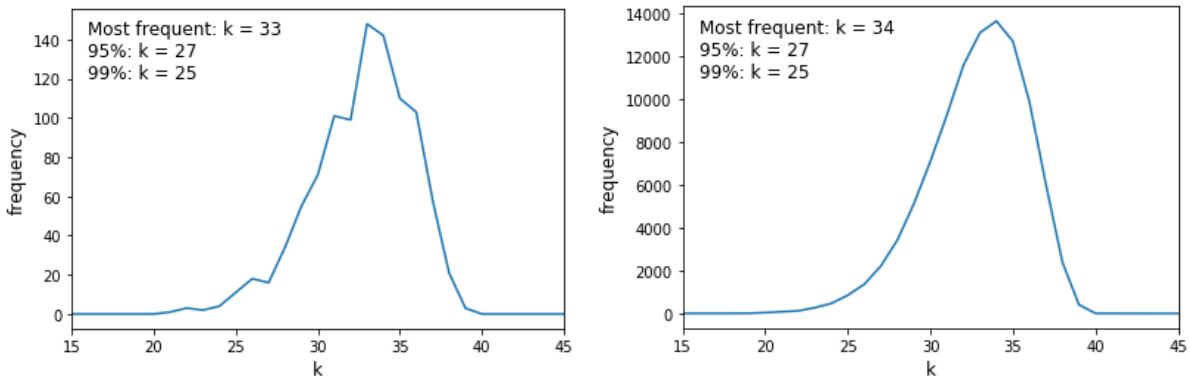
$$AVC = \prod_{i=1}^{a} DistinctValues(Attribute_i)$$

The distribution of values within an attribute is unknown in the local scenario. Within this simulation, we model it as an equal distribution. Each value is assumed to occur with the same probability.

The Monte Carlo simulation is carried out by randomly assigning r records to g groups, whereas the parameter g represents the AVC. The size of the smallest occupied group is then equivalent to the achieved k of the k-anonymity model. This process is repeated n times such that a frequency distribution of k can be obtained. We calculate three different measures, giving implications about the resulting privacy level: The most frequent value of k and the k with which at least 95% and 99% of all records serve k-anonymity. We define the anonymity threshold as a percentage that indicates the risk that the respective k-value is not reached. For the scope of this work, we assume an anonymity threshold of 1% as reasonable. This implies that we will further use the k for which 99% of the simulations result in an equal or larger k (99% rule). An implementation of this simulation in Python, which was used to calculate the results in this work, can be found in section C.1 within the appendix.

An exemplary simulation was carried out with five groups (g = 5) and 200 different records (r = 200), the results of which are shown in Figure 9.2. Two different amounts of iterations (n = 1000 and n = 100,000) are compared. It shows that more iterations lead to a smoothing of the curve and an approximation to a Gauss distribution. This will generally also lead to more accurate and usable results. Thus, we aim to use a high n to obtain usable results. However, more iterations also cause an increase in computation time. For the relative small parameters of g and r, this was still feasible as the computation times of 0.5s and 36s show. However, for higher ones, this can cause relatively long-lasting simulations. The right-hand graphic shows that in most cases, a k of 34 is achieved. In more than 99% of all iterations, k-anonymity with

a k of at least 24 was achieved. We will use the respective 99% thresholds as the relevant indicator of the privacy level as part of the evaluation of the concept for wrist-worn wearable data.



(a) Iterations: n = 1000
Computation time: 0.5s

(b) Iterations: n = 100,000
Computation time: 36s

Figure 9.2.: simulation with r = 200 and g = 5

This local probabilistic k-anonymity concept can be considered as a beneficial approach to investigate the potential privacy in an unknown data set for some given parameters. However, it is also related to some critical shortcomings. The model, like it is designed, applies a random distribution for the values of an attribute. This is not always representative for a data set as there can exist values that occur very frequently and rather rare ones. These rare values affect the k-anonymity model as it will most likely lead to a lower k. Additionally, the simulation does not take into account correlations between attributes, so the simulation is more risk-averse in this case. The implementation has the potential to be extended for these distribution and correlation purposes. Another drawback is the long computation time when high values are assigned to the input parameters g and r. Nevertheless, in the optimal case, the simulation only needs to be executed only once to obtain an appropriate risk assessment.

9.4. De-Identification of Identifiers

To analyze the de-identification of identifiers, we first differentiate between single value and multiple value attributes. The former ones indicate that exactly one value is assigned to each record or user, whereas there are more values assigned to the latter ones. The user data table contains only single value attributes, while the other tables contain multiple value ones. In this section, we will investigate the methods for quasi identifiers, propose suitable combinations for the application and evaluate them with the proposed k-anonymity Monte Carlo simulation.

9.4.1. Methods on Single Value Attributes

Single value attributes have the advantage that their risk and utility can be easily investigated by the number of distinct attribute values. Table 9.2 shows these values in increasing order along with an indication whether the attribute shall be removed or generalized according to HIPAA Safe Harbor (see chapter 2). For the date values, HIPAA recommends removing everything besides the year. For others, which are also marked with an *X*, it means that they should be completely removed. The table illustrates that the number of distinct values correlates with a removal according to HIPAA. We use this table as a risk measure to determine the application of de-identification methods.

	handedness	gender	height	weight	country	currentGear	createdDate	birthDate	eMail	firstName	lastName	profileImage
Distinct values	2	3	51	81	195	2000	3650	15695	$\sim \infty$	$\sim \infty$	$\sim \infty$	$\sim \infty$
HIPAA Safe Harbor	-	-	-	-	-	-	X	X	X	X	X	X

Table 9.2.: Risk of single value attributes

Attribute suppression

Attribute suppression (the deletion of the attribute) is applicable to all attributes of the user data table. However, the *eMail* attribute is not feasible for that as it is used as a login parameter of the user within the system. There are three attributes with a very large number of distinct values that do not allow generalization at the same time: *firstName*, *lastName* and *profileImage*. We propose their redaction as they lead to a high re-identification risk and offer just a small utility.

Suppression(firstName); Suppression(lastName); Suppression(profileImage)

Generalization

We observed a possible applicability of generalization on the attributes *country*, *height*, *weight*, *currentGear*, *createdDate*, and *birthDate*. In Table 9.3, we suggest respective generalization hierarchies, provide example values, and indicate the resulting number of unique values (last line). The term *any* means that on this level there is no differentiation between values anymore, leading to only one distinct value.

Level	Legend	country	height	weight	currentGear	createdDate	birthDate
0	domain	country	cm	kg	model	day	day
	(example)	(DE)	(181)	(81)	(NB Rubix)	(2018-09-12)	(1994-05-04)
	distinct values	195	51	81	2,000	3,650	15,695
1	domain	continent	5cm	5kg	brand	month	month
	(example)	(Europe)	(180-184)	(80-84)	(New Balance)	(2018-09)	(1994-05)
	distinct values	6	10	16	20	120	516
2	domain		10cm	10kg		year	year
	(example)	Any	(180-189)	(80-89)	Any	(2018)	(1994)
	distinct values	1	5	8	1	10	43
3	domain		20cm	20kg		5 years	5 years
	(example)		(180-199)	(80-99)		(2015-2019)	(1990-1994)
	distinct values		3	4		2	9
4	domain						10 years
	(example)		Any	Any		Any	(1990-1999)
	distinct values		1	1		1	5
5	domain						
	(example)						Any
	distinct values						1

Table 9.3.: Generalization hierarchies with example values and number of unique values

Figure 9.3 shows the relative influence of the generalization level on the number of distinct values (0=minimum generalization level, 1=maximum generalization level). The y-axis is based on a logorithmic scale where each interval section shows a decrease by a factor of 10. One can conclude that the most significant improvements of privacy can be done by generalization of *birthDate* and the lowest with *height*. This also matches with their initial number of distinct values.

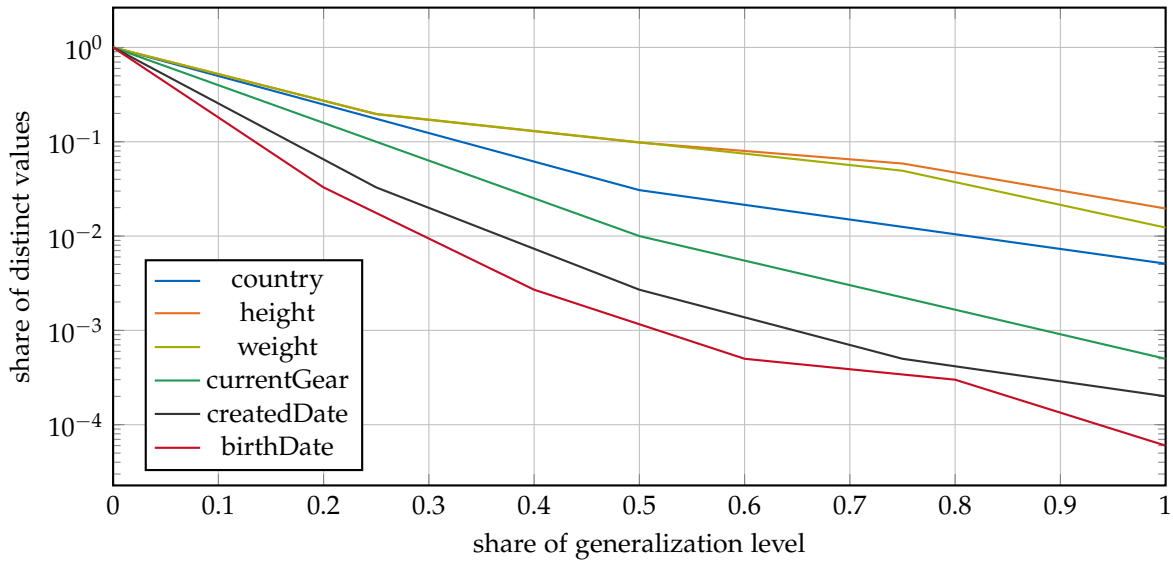


Figure 9.3.: Distinct values for generalization

As a result, we propose the application of all six attributes dependent on the desired level of privacy. The parameter *X* indicates the respective level of generalization, which will be used later on. For *birthDate* and *createdDate* the lowest generalization level shall be 2 due to the HIPAA safe harbor rules.

Generalization(birthDate, X); Generalization(createdDate, X); Generalization(currentGear, X); Generalization(country, X); Generalization(weight, X); Generalization(height, X)

Noise addition

We do not consider noise addition for the single value attributes as it is not suitable because it does not improve privacy according to the k-anonymity model. Instead, it is more feasible for privacy models like differential privacy.

Character masking

This method is applicable for the attributes *eMail*, *firstName*, *lastName*, and *country*. We do not take this option into account as it destroys the utility, which is why a deletion in this case is more suitable.

Creating pseudonyms

For the attributes *eMail*, *firstName*, *lastName*, the creation of pseudonyms is very similar. Since the email address often contains an individual's name and is used as a login credential, we propose to replace it with a pseudonym and delete the name attributes.

Creating pseudonyms (eMail)

9.4.2. Evaluation of Single Value Attributes

The evaluation of suitable combinations of de-identification methods on single-value attributes is done in multiple iterations. For each iteration, the k-anonymity Monte Carlo simulation is performed. The simulation is carried out based on three user amounts which we assume to be reasonable and representative for a wrist-worn wearable platform: 20 million users, 2 million users and 100,000 users.

We start the evaluation with the highest user amount (50 million) until adequate privacy protection is reached. The threshold for adequate privacy protection is illustrated later on. Then the concept is further extended to 2 million and 100,000 users, respectively. This is based on the fact that more users lead to higher privacy protection based on k-anonymity.

First iteration

Based on the findings in the previous section, we propose the methods shown in Table 9.4 as the first iteration of the privacy levels and calculate the respective AVCs.

The highest AVC value of about 67 million is still higher compared to the largest amount of assumed users (50 million). This ratio will not lead to sufficient privacy protection considering the assumptions of the k-anonymity Monte Carlo simulation. An AVC larger than the number of records will, in almost all cases, lead to k-anonymity with k = 1. Therefore, these combinations are not considered sufficient since they will lead to unique records in the data set.

#	Suppression	Creating pseudonyms	Generalization	AVC
3	firstName, lastName, profileImage	eMail	birthDate(3), createdDate(3), height(1), weight(1), currentGear(1)	$67.4 * 10^6$
2	firstName, lastName, profileImage	eMail	birthDate(2), createdDate(2), height(1), weight(1), currentGear(1)	$1.6 * 10^9$
1	firstName, lastName, profileImage	eMail	birthDate(2), createdDate(2)	$4.15 * 10^{12}$

Table 9.4.: Concept iteration 1

Second iteration

We use level 3 of the first iteration as a basis for the second iteration and add further methods for the reduction of the AVC, as shown in Table 9.5. As the *handedness* attribute (2 distinct values) does not offer large analytical value, it will be deleted from now on.

The k-anonymity Monte Carlo simulation is applied to the calculated AVC values. It was performed on the three described user amounts, whereas

- (1) relates to 50 million users,
- (2) relates to 2 million users and
- (3) relates to 100,000 users.

For these scenarios, the resulting 99% thresholds, meaning that in 99% of all cases a specific k-anonymity level is achieved, are used as evaluation criteria. We refer to this value as *k (99% rule)* in Table 9.5 and the following ones. The full results of these simulations, including the corresponding frequency distributions, can be found in section C.2 in the appendix. Due to

time constraints and the computational complexity, not all combinations were simulated. In case a k of 1 was retrieved, all combinations with the same parameters but either a higher AVC or a lower number of records were also set to 1.

The choice for an appropriate k and therefore the interpretation of these k values is rather difficult. It is not possible to give a minimum recommendation for that parameter. Instead, it should always be chosen according to the specific use case (Article 29 DP Working Party, 2014; Kiyomoto & Miyake, 2014). This is because it is hard to assess what exactly is sufficient for de-identification, and also privacy laws do not provide a specific suggestion (I5). In the field of health data, a k between 5 and 15 is often used, but it is still an arbitrary value (Desfontaines, 2017). Based on this, we propose a k of 10 as the first value for a minimum privacy protection. Values above that threshold are framed with a green dashed line in the table. Within this iteration, two de-identification approaches that serve appropriate privacy protection for the user amount of 50 million were identified.

#	Suppression	Creating pseudonyms	Generalization	AVC	k (99% rule) (1)	(2)	(3)
4	firstName, lastName, profileImage, handedness, createdDate	eMail	birthDate(3), height(1), weight(1), currentGear(1), country(1)	518,400	48	1	1
3	firstName, lastName, profileImage, handedness, createdDate, currentGear	eMail	birthDate(3), height(1), weight(1)	842,400	22	1	1
2	firstName, lastName, profileImage, handedness, createdDate	eMail	birthDate(3), height(1), weight(1), currentGear(1)	16,848,000	1	1	1
1	firstName, lastName, profileImage, handedness	eMail	birthDate(3), createdDate(3), height(1), weight(1), currentGear(1)	33,696,000	1	1	1

Table 9.5.: Concept iteration 2

Third iteration

For the third iteration, we add additional methods of generalization and suppression on top of the proposed combinations in iteration 2. This leads to a further decrease in AVC. Table 9.6 shows the individual results that were obtained in the simulations. The combinations now lead to sufficient k values throughout all three user groups. It is recognizable that a lower AVC and a higher amount of users always lead to better privacy protection in terms of k-anonymity. We stop at this iteration step as even for the smallest user group, an appropriate k value is reached.

#	Suppression	Creating pseudonyms	Generalization	AVC	k (99% rule) (1)	(2)	(3)
5	firstName, lastName, profileImage, handedness, createdDate, currentGear, country, weight	eMail	birthDate(3), height(2)	135	368,110	14,367	640
4	firstName, lastName, profileImage, handedness, createdDate, currentGear, country	eMail	birthDate(3), height(2), weight(2)	1,080	45,418	1,672	55
3	firstName, lastName, profileImage, handedness, createdDate, currentGear, country	eMail	birthDate(3), height(1), weight(1)	4,320	11,085	366	5
2	firstName, lastName, profileImage, handedness, createdDate, currentGear	eMail	birthDate(3), height(2), weight(2), country(1)	6,480	7,307	232	1
1	firstName, lastName, profileImage, handedness, createdDate, currentGear	eMail	birthDate(3), height(1), weight(1), country(1)	25,920	1,722	39	1

Table 9.6.: Concept iteration 3

9.4.3. Evaluation of Multiple Value Attributes

Quasi identifier attributes with more than one value per user are harder to evaluate in terms of re-identification risk. This is mainly due to the number of values per user which is another unknown component that needs to be considered. Table 9.7 shows these attributes, along with their number of distinct values.

Table	Activity		Activity trackpoint
Attribute	type	time	$\begin{pmatrix} \text{time} \\ \text{latitude} \\ \text{longitude} \\ \text{elevation} \end{pmatrix}$
Distinct values	5	5,256,000	∞
HIPAA Safe Harbor	-	-	-

Table 9.7.: Risk of multiple value attributes

All attributes in this category are related to the tracking of activities. The 4-tuple in the activity trackpoint table is the most critical one, as it indicates the exact location of an individual (expressed by *latitude*, *longitude*, and *elevation*) at a specific point of time. Since many activities are assumed to start or end at home or at work, this can easily be used to identify an individual or to narrow down possible records. Transformative methods on single elements of the tuple are not beneficial as the relation of remaining utility and promised privacy is not justifiable. One option is the suppression of GPS coordinates that fall within a certain radius, but since a part of an activity's location data is then deleted, the utility is questionable. Generalizing, aggregating, swapping, or adding noise to the location points also does not provide any value in terms of utility and leads to a remaining identification risk. As a result, we propose not to include the activity trackpoint tuples as part of the user-related data, which is transferred to the provider's system. An alternative can be to transfer the activity trackpoints separately, without linking it to specific users. This would still allow the support of transport infrastructure planning, for example. The information that is calculated on the trackpoints, like the distance and pace of an activity, can then be added to the general activity table as long as they are not classified as quasi identifiers.

The activity table itself, containing activity time and type, also implies a substantial re-identification risk, mainly because each user can relate to multiple entries of the table. In case an adversary knows that an individual tracked activities on two specific days, he can narrow down eligible records. In this case, let us assume a period of one month as valid to be accountable as quasi identifiers. If individuals pursue between 0 and 15 activities in a month with 30 days, each one on a separate day, the number of possible combinations is

$$\sum_{i=1}^{15} \frac{30!}{(30-i)! \cdot i!} = 614,429,671$$

This number represents a substantial risk. However, the applicability of this risk is highly dependent on the type of adversary and his knowledge. An insider, who has a close connection to the data subject, can know significantly more about an individual (like the days on which activities were performed) than the general public does (Wu, 2012, p. 1154). Hence, it relies on a subjective decision which specific knowledge an adversary can obtain.

Since the number of possible combinations is very high compared to the number of distinct values of the previously analyzed single value attributes, we propose to suppress the activity related attributes completely. Even after reducing the maximum possible activities per month to ten, the number of possible combinations is still around 53 million (calculated by adapting the above formula). As this is higher than the largest user group (50m users), it will very likely lead to 1-anonymity and to the potential identification of individuals.

As a result, the concept iterations performed in the previous subsection are used as the overall concept for the de-identification of quasi identifiers. The multiple value attributes are suppressed additionally, which does not change the previous risk estimation.

9.5. Evaluation of Scenarios

In the last sections, we introduced the concept of local probabilistic k-anonymity and showed how it can be applied to the introduced wrist-worn wearable data model. In this section, we will evaluate the realization of different privacy levels based on the choice of the users. This condition was identified as one of the requirements in chapter 8. The evaluation will be carried out using four different scenarios.

Figure 9.4 shows a diagram summarizing all results that were obtained through the simulations. More detailed results can be found in section C.2. The x-axis shows the attribute value combinations which is the first input parameter of the simulation. The second one is illustrated through the colored lines which represent different user amounts. The k value (99% rule) is displayed on the y-axis, and each dot represents one simulation result. Both axes are represented on a logarithmic scale. The obtained lines are all monotonously falling. The increase of AVC results in a decrease of k until the minimum value of 1 is reached. A higher user amount leads to a shift of the points in positive a direction in terms of both x- and y-axis, which indicates a better privacy protection.

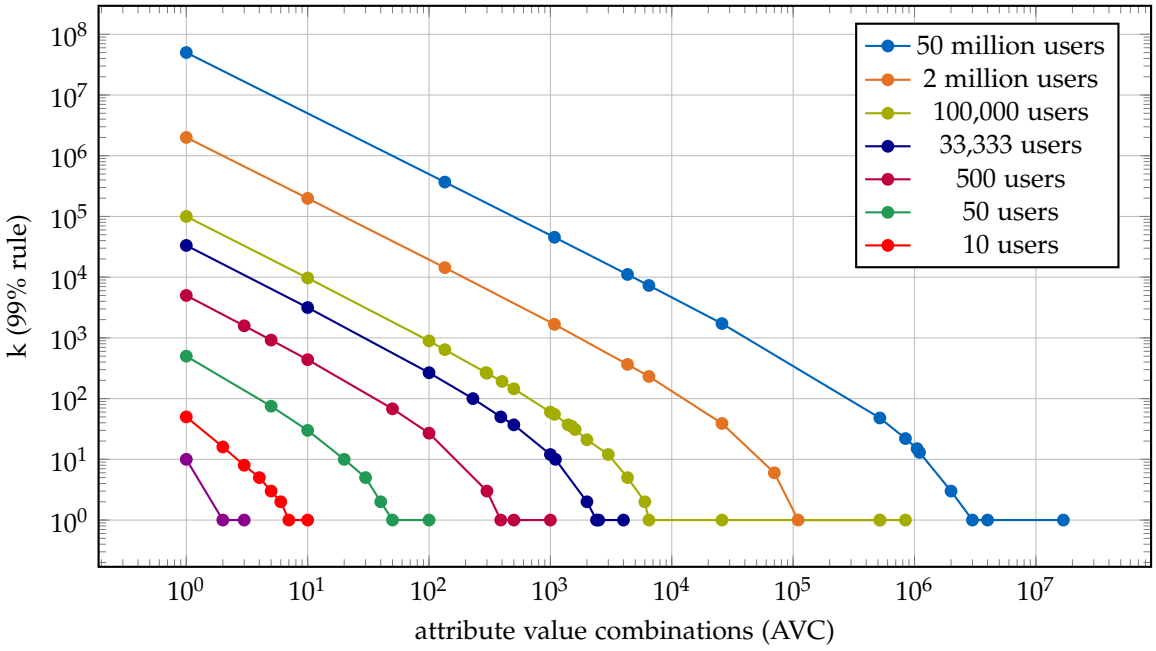


Figure 9.4.: local probabilistic k-anonymity simulation results

9.5.1. Scenario A: Common Privacy Level

By a common privacy level, we mean that all users within a system are tied to the same level of privacy, which is illustrated in Figure 9.5. This implies that the same de-identification methods are applied to all of them. In this case, the k-anonymity simulations stated in

subsection 9.4.2 can be taken into account as risk indicators. This scenario is beneficial because all users can be considered together for risk estimation. As Figure 9.4 shows, a high user number leads to better privacy. However, an independent privacy level based on the choice of every single user was identified as one of the requirements. As this scenario only provides one single level, it is not applicable.

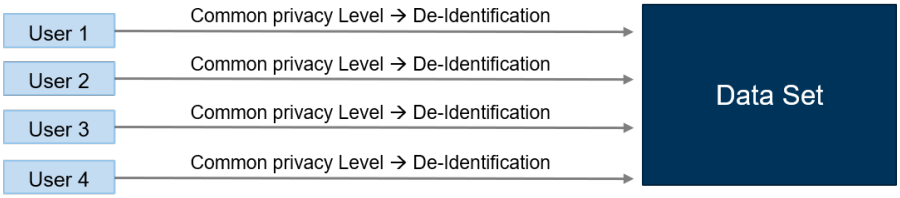


Figure 9.5.: Common privacy level

9.5.2. Scenario B: Independent & Individual Privacy Levels

This scenario assumes that each user can choose his desired privacy level completely independent, leading to various individual combinations of de-identification methods. Figure 9.6 shows the corresponding data flow.

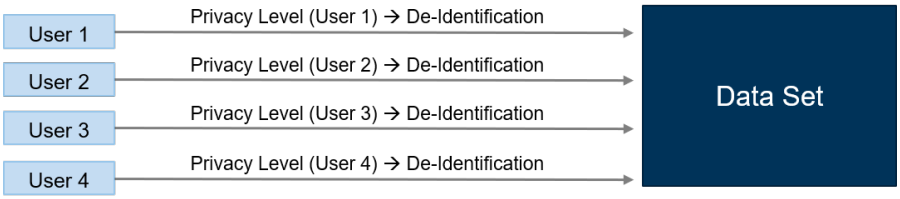


Figure 9.6.: Independent & individual privacy levels

In the following example, we will show the effect of this scenario on the achieved privacy of individuals. Let us first assume we have a group of 4 users and apply de-identification methods independently on them. Thus, each user selects his individual level of privacy, and based upon this, several methods are applied. However, the desired privacy levels of other users are unknown. The single records are transformed and then combined in one data set. Table 9.8 (a) shows an exemplary source data table with four values of the attribute age. Each age is unique, resulting in k=1.

In variant (b), all values are generalized to the same level of a 10-year interval. This represents the case that all users choose the same privacy level independently from each other. For this example, it results in one equivalence class and 4-anonymity.

Now, let us assume the first user aims for a rather low privacy level, leading to no generalization at all. The other seven users still use the same level from the previous example and generalize their age in a 10-year interval. The result, after combining the records, is shown

in table (c). The first user now cannot be distinguished from the second, third and fourth records because they all could be 22 years old. Hence, his privacy protection does not change compared to example (b). However, the next three users are less protected as the value 22 is not part of their equivalence class anymore. This can be illustrated by a potential adversary who is looking for a record with the age of 27. He can narrow down the table to three records. Hence, 3-anonymity is achieved.

The values in table (d) show the contrary situation. Only the first user is aiming for the mentioned privacy level while the others aim for low protection, whereby their exact age remains. As a result, the first user is protected with k=1 since because he is the only one that can be of age 22. In contrast, the other users are all in equivalence classes of size two, i.e., two records are possible to have the age of 28 (first and last user).

Age	Age	Age	Age
22	20-29	22	20-29
23	20-29	20-29	23
27	20-29	20-29	27
28	20-29	20-29	28
(a) Source Table k=1	(b) k=4	(c) k=3	(d) k=1

Table 9.8.: Local & independent de-identification

In conclusion, an increase of one user's individual privacy level would not improve his privacy, but the remaining users would benefit from such action. Analogously, a decreasing privacy level would enhance one user's utility but lead to a decrease in the other user's privacy protection. As a consequence, a system like that would make everyone choose the lowest k possible and therefore harm the privacy protection of each individual. Hence, it is not suitable for our approach.

9.5.3. Scenario C: Privacy Clusters with Equal Distribution

The third scenario describes the combination of users seeking the same privacy level into so-called privacy clusters. Within the clusters, the same de-identification methods are applied. For each cluster, the k-anonymity estimation is performed with regards to the number of users within this cluster instead of the amount of all users. Through a limited amount of available levels, it is ensured that the user numbers of one group are not too small. A combination of the different clusters can then only lead to an improvement of privacy, so there is no mutual weakening. Figure 9.7 shows a visualization of the scenario.

Let us now presume we have a system with 100,000 users, and three different levels of privacy are offered, resulting in the formation of three clusters. Let us say the users split up evenly to the different levels. Thus, every cluster has 33,333 users. Based on the previously proposed

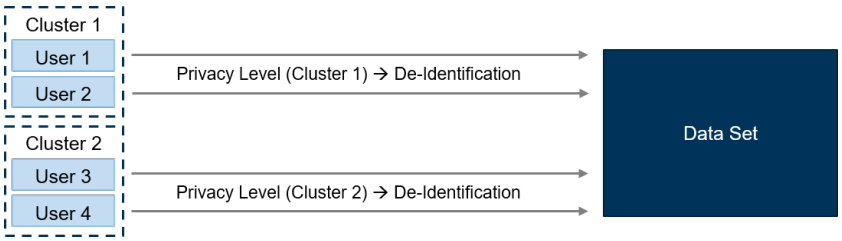


Figure 9.7.: Privacy clusters

minimum k value of 10, we arbitrarily define the different levels to a k of 10, 50, and 100. We refer to them as the privacy levels A, B and C. Throughout the k-anonymity simulation we obtain the following results for this example:

- Cluster 1 - 33,333 users, 1,100 AVC: k (99% rule) = 10 (privacy level A)
- Cluster 2 - 33,333 users, 390 AVC: k (99% rule) = 50 (privacy level B)
- Cluster 3 - 33,333 users, 230 AVC: k (99% rule) = 100 (privacy level C)

Now let us compare the achieved results with an approach that combines all 100,000 users together (scenario A) and leads to the same privacy levels:

- 100,000 users, 3,200 AVC: k (99% rule) = 10 (privacy level A)
- 100,000 users, 1,150 AVC: k (99% rule) = 50 (privacy level B)
- 100,000 users, 680 AVC: k (99% rule) = 100 (privacy level C)

The results show that the user amount can have a significant impact on the remaining utility, which is represented by the AVC value. Instead of realizing the three privacy clusters and serving 10-, 50- and 100-anonymity, we could combine all users in one group and obtain 50-anonymity (privacy level B). This yields to a higher utility for all users (1,150 $>$ 1,100, 390, 230). While the clusters 1 and 2 are protected with the same or a higher level, only the third cluster is harmed as it achieves a lower k value than with a separation of the clusters. To counteract this, we only merge the first two clusters and leave cluster 3 separately. This leads to the following results:

- Cluster 1+2 - 66,666 users, 790 AVC: k (99% rule) = 50 (privacy level B)
- Cluster 3 - 33,333 users, 230 AVC: k (99% rule) = 100 (privacy level C)

Cluster 1 benefits from better privacy with a lower utility, while the users from cluster 2 achieve a significantly better utility. A combination of the clusters is highly dependent on the preferences of the users and cannot be generalized for this reason. Cluster 1 will need to decide if the trade-off is advantageous for them, whereas cluster 2 can only benefit from a merge with them.

This scenario shows that it can make sense to combine multiple clusters. However, it is also

dependent on the user's perception and preferences. The user numbers in one cluster are an essential factor to counter the trade-off between privacy and utility.

9.5.4. Scenario D: Privacy Clusters with Unequal Distribution

In scenario D, we assume that the users are unequally distributed to the clusters. This represents a more realistic approach since an equal distribution is rather unlikely. For three different clusters, we assume that the majority of users choose the medium privacy level, whereas only 10% choose level A and C each. For 100,000 total users and the targeted k values from the previous scenario, we obtain the following results through the simulation:

- Cluster 1 - 10,000 users, 350 AVC: k (99% rule) = 10 (privacy level A)
- Cluster 2 - 80,000 users, 920 AVC: k (99% rule) = 50 (privacy level B)
- Cluster 3 - 10,000 users, 72 AVC: k (99% rule) = 100 (privacy level C)

As indicated by the AVC, cluster 2 has by far the highest remaining utility. This is a result of the large user amount which is allocated to this group. Compared to that, the first cluster has a lower utility as well as a lower privacy level. Hence, it is only beneficial to combine the first two cluster like the following result shows:

- Cluster 1+2 - 90,000 users, 1,050 AVC: k (99% rule) = 50 (privacy level B)

The third cluster, however, will need to be kept separated from the other ones as the high desired level of privacy cannot be achieved without other shortcomings. Therefore, the users either have to cope with the relatively low utility or decrease their privacy level. The latter will lead to a combination with the other clusters resulting in substantially higher utility value.

9.5.5. Implications for Wrist-Worn Wearable Data

Based on the arguments stated in this chapter, we propose the approach of privacy clusters for the wrist-worn wearable use case. This scenario commonly leads to the best results for individual users because it leads to the achievement of privacy guarantees through consistent de-identification methods. The service provider needs to take into account that the number of distinct clusters should not be too high and therefore limited as this negatively affects the general utility of the data. Therefore, we propose three different levels and clusters as reasonable to cope with the user's different requirements for privacy. As shown in the scenarios C and D, it can be beneficial to combine different clusters together into one. This can yield better results for both groups. However, this dependents heavily on how the users distribute among the single clusters. A possible cluster combination can be determined and solved as part of an optimization problem by the service provider. It is also necessary to consider possible variants and fluctuations between the different privacy clusters. Therefore,

we propose conducting the simulation with a lower user amount that is deemed to be certain. This might be around 80% of the actual users.

In combination with the local probabilistic k-anonymity approach, the concept of privacy clusters will lead to promising results to make predictions for the individual user's privacy levels. However, the simulation approach, as it was implemented in this work, brings also some shortcomings which need to be considered. First of all, a random or equal distribution for the values of the attributes is assumed. This is not always realistic, but the simulation can be extended by a different distribution approach in future work. For this purpose, standard attribute value distributions within data sets would need to be analyzed. These insights could then serve as the basis to assign attribute values as part of the simulation. While this fact is an underestimated risk in the simulation, the missing consideration of correlations between attributes leads to an overestimated risk as some attribute value combinations might not be possible. Also, dependencies between the records (users) are not considered, as in our case a homogeneous distribution is assumed. The anonymity threshold was set to 1% in this work. This can be varied for different use cases in order to set a higher or lower acceptable risk limit. Overall, we argue that the proposed simulation is an beneficial approach to estimate k-anonymity and, thus, the achieved level of privacy. It is a practical way to apply k-anonymity locally without knowing the overall distribution. Furthermore, the needed input parameters are reasonable to know even in case of a local application.

10. Conclusion & Outlook

This final chapter summarizes the research results of this work. Additionally, limitations as well as topics for future research that are related to this thesis are presented.

10.1. Summary & Discussion

In this section, we will present the three research questions alongside with the answers and contributions this work provides.

RQ 1: What is the state of the art of approaches using de-identification methods for privacy-enhancing Big Data Analytics and how can they be distinguished from other approaches?

Through an extensive literature review, we provide a comprehensive analysis of the current state of the art of approaches using de-identification methods. We illustrate that a clear and precise delimitation of the terms *de-identification*, *anonymization*, and *pseudonymization* does not exist in current research. Therefore, a more generic terminology that covers all three terms was used and proposed for the scope of this work. We contribute to existing research with a complete and comprehensive classification of de-identification methods, which is summarized in Figure 6.1. Existing overviews in research either lack in such a classification or its completeness of methods. A total of 15 distinct methods were identified and then classified into two categories and four subcategories. We classified them into non-perturbative and perturbative de-identification methods and distinguished between data type independent and numerical methods. The overview provides three essential benefits to researchers: First and foremost, it helps to get a clear picture and impression of available de-identification methods. Secondly, it illustrates how the methods are related to each other and how they can be differentiated. Lastly, the classification supports the decision process for the choice of such a method. Additionally, we provide definitions, explanations, and exemplary application scenarios for all identified de-identification methods.

RQ 2: What are requirements for privacy-enhancing analytics of wrist-worn wearable data in the cloud?

Wrist-worn wearables collect large amounts of highly sensitive data and therefore pose a significant risk to the privacy of their users. In chapter 7, we provide an analysis of how

this specific use case can be impacted by the application of de-identification methods. By an analytical threat model, it was shown that linkability and identifiability threats can be mitigated. Additionally, a generic wrist-worn wearable data model was developed based on data exports of multiple service providers. It contains a data structure for the use case that can be leveraged for investigations. Ten requirements for a concept using de-identification methods were identified and described. They were the result of the conducted literature research in combination with 12 expert interviews. They address the second research question and serve as guidance in the phase of concept development.

RQ 3: What are concepts enabling data privacy for wrist-worn wearable data in the cloud based on de-identification Methods?

Throughout chapter 9, a specific concept for the application of de-identification methods on wrist-worn wearable data was developed. We propose a local probabilistic k-anonymity concept that uses a Monte Carlo simulation to calculate privacy estimates. Thus, it closes the existing research gap for a local application of k-anonymity. It is based on reasonable parameters that shall be available for a local user. We illustrated how this could be applied to the wrist-worn wearable data model and show the impact that different method combinations as well user numbers have on the obtained privacy level. A decrease in attribute value combinations (AVC) and an increase in the user amount were identified as crucial measures to improve privacy. Furthermore, different scenarios for the choice of privacy levels were constructed and investigated. The evaluation and comparison led to the conclusion that a concept based on privacy clusters is the most promising one. In this concept, users with the same privacy desire are combined, so that the same de-identification methods are applied on their data. We show that this approach prevents the mutual weakening of each other's privacy and leads to the best outcome compared to the other scenarios. We argue that three distinct privacy levels are reasonable to cope with different user's needs and to limit the number of overall clusters. Based on the distribution of all users to these levels, an appropriate number of AVC and therefore proper de-identification methods can be identified.

10.2. Limitations

The results of this work were influenced by a few limitations which will be explained in this section.

Firstly, the lacking availability of a complete data set of wrist-worn wearable data limited the findings and their corresponding evaluation. We circumvented this limitation by the creation of a generic data model for the specific use case of wrist-worn wearable data. However, this data model only shows which attributes are stored and how they are related. It does not take into account how single attribute values look like and how they relate to each other in a real data model. Therefore, distributions within values and correlations between attributes could not be obtained. Thus, the current local probabilistic k-anonymity concept is limited

to some factors which can be improved by incorporating an existing data set. This data set could also help to validate and further improve the concept through continuous testing and application of de-identification methods, which was not possible in this work due to time constraints.

Secondly, the possibility of validation in a real application scenario also limited the results. This would be especially beneficial for the proposed scenario with privacy clusters. In a real application scenario, one could derive further insights into how users will allocate and distribute to different clusters and how frequent switches between privacy levels will happen. Additionally, the scenario of an increasing number of users over some time was also not considered.

Thirdly, the difficulty of finding suitable domain experts for the pursued interviews was another constraint. As the research area is located at the intersection of privacy and data analytics, we addressed experts of those domains. However, we realized that de-identification is generally known but not often applied in organizations yet. Therefore, it was hard to gather practical experiences and knowledge directly tied to the application of such methods. We were able to obtain broad and usable insights into how data privacy is handled and perceived in organizations. However, only two interview partners had a deeper knowledge of the practical application of de-identification methods and privacy models. This limitation is mainly due to the fact that the use of de-identification is still rather rare in practice.

10.3. Future Work

This work provides some topics that were not covered and therefore leave room for further investigations.

The proposed local probabilistic k-anonymity concept can be extended with a more realistic distribution of attribute values. We proposed an equal distribution as a starting point, but an extension will be beneficial to obtain more valid results. Therefore, we propose to analyze existing data sets for a specific use case with regards to their distribution of attribute values. This distribution could then be used as an allocation basis for the simulation. An integration of l-diversity and t-closeness can also be taken into consideration as these solve reasonable shortcomings of k-anonymity.

Furthermore, the evaluation and testing of the proposed concept is another area of future work. First of all, the impact of different anonymity thresholds and the choice of a reasonable one can be investigated. Secondly, the validity of the obtained simulation results can be performed. Additionally, a benchmark of local probabilistic k-anonymity against local differential privacy will be from general research interest.

A. Wrist-Worn Wearable Use Case

A.1. Wrist-Worn Wearable Data Tables

userID	eMail	firstName	lastName	country	gender	birthDate
(int)	(string)	(string)	(string)	(string)	(string)	(string)
80248072	john.doe@mail.com	John	Doe	DE	m	1994-05-04

Table A.1.: User data part 1

height	handedness	profileImage	currentGear	createdDate	weight	vo2Max
(int)	(string)	(string)	(string)	(string)	(int)	(int)
184	right	image.png	New Balance Rubix	2018-09-12	81	53

Table A.2.: User data part 2

userID	timestamp	heartrate
(int)	(int)	(int)
80248072	1540591440000	67
80248072	1540591560000	68

Table A.3.: Heart rate data

userID	startTime	endTime	activityLevel
(int)	(string)	(string)	(float)
80248072	2019-01-06T22:00:00Z	2019-01-06T22:01:00Z	5.799102389
80248072	2019-01-06T22:01:00Z	2019-01-06T22:02:00Z	4.425838591

Table A.4.: Sleep movement data

userID	activityID	time	type
(int)	(int)	(string)	(string)
80248072	47964342	2020-01-01T08:03Z	running

Table A.5.: Activity data

activityID	time	latitude	longitude	elevation	heartrate	cadence
(int)	(string)	(float)	(float)	(float)	(int)	(int)
47964342	2020-01-01T08:48:46Z	48.172334	11.5538459	529.40	163	79
47964342	2020-01-01T08:48:47Z	48.172357	11.5538698	529.20	164	80

Table A.6.: Activity trackpoint data

userID	startTime	voltage	derivation
(int)	(string)	(int)	(int)
80248072	2020-01-15T08:48:00Z	-40	438
80248072	2020-01-15T08:48:00Z	-44	887

Table A.7.: ECG data

userID	time	steps	floors
(int)	(string)	(int)	(int)
80248072	2020-01-18T14:12:00Z	14	0
80248072	2020-01-18T14:13:00Z	23	1

Table A.8.: Step data

A.2. LINDDUN Threat Trees

A.2.1. Linkability Threat Trees

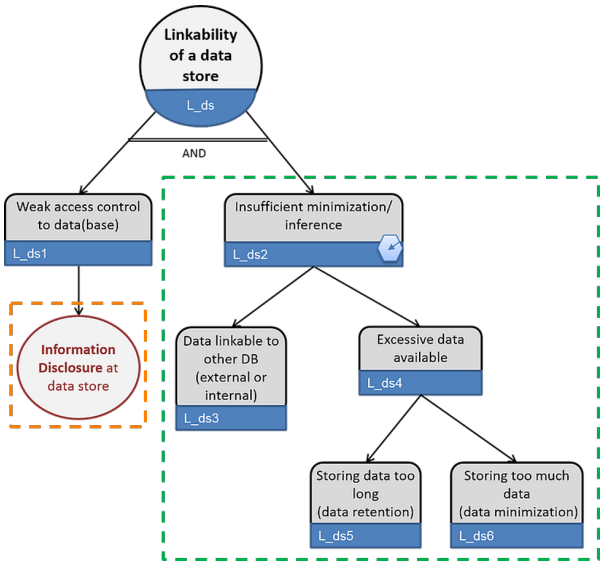


Figure A.1.: Linkability of data store (provider database)
Source: based on Wuyts et al. (2014)

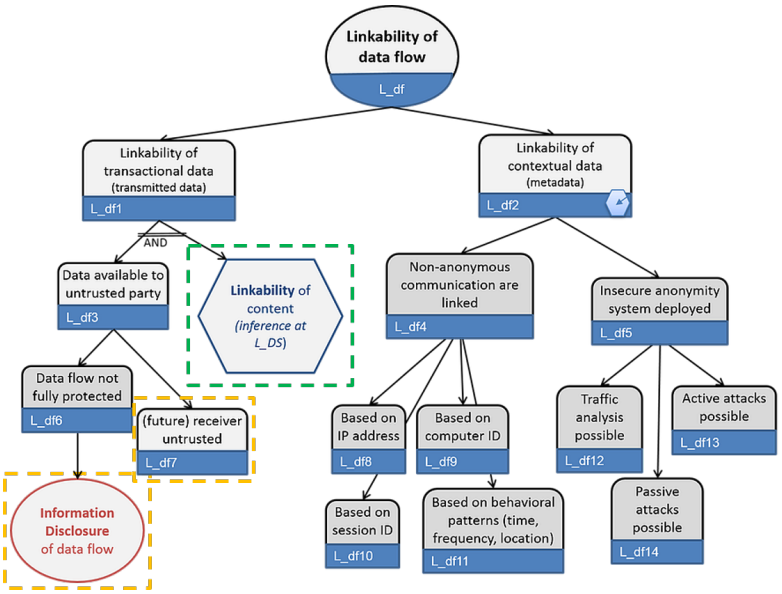


Figure A.2.: Linkability of data flow (platform & wearable data stream)
Source: based on Wuyts et al. (2014)

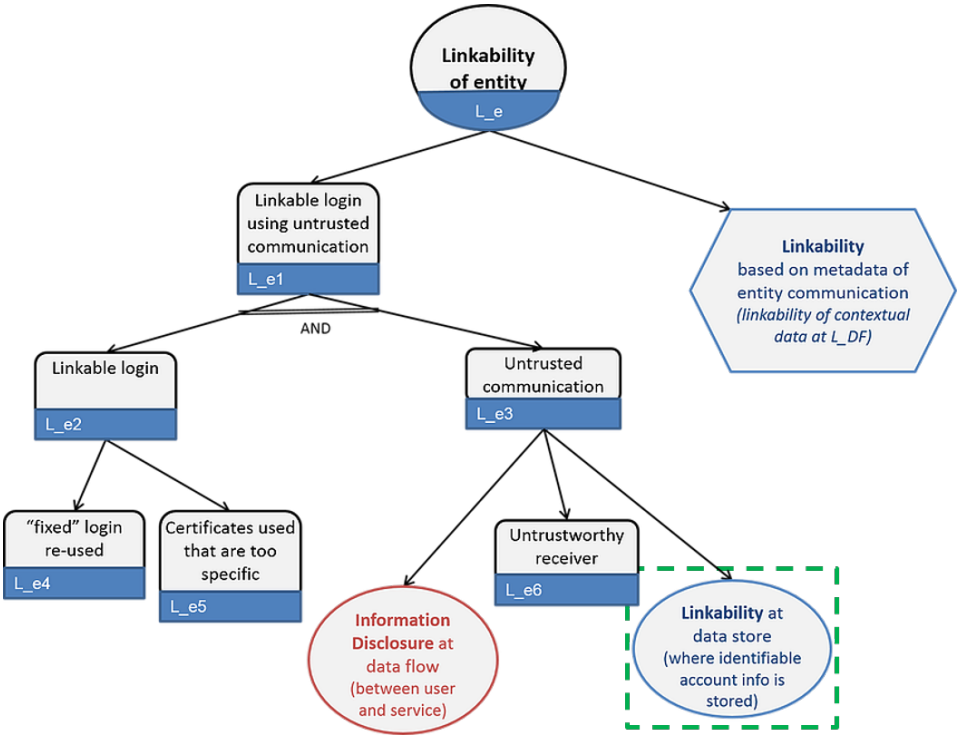


Figure A.3.: Linkability of entity (user & wearable)
Source: based on Wuyts et al. (2014)

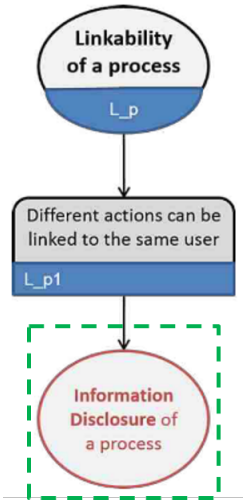


Figure A.4.: Linkability of process (platform & service)
Source: based on Wuyts et al. (2014)

A.2.2. Identifiability Threat Trees

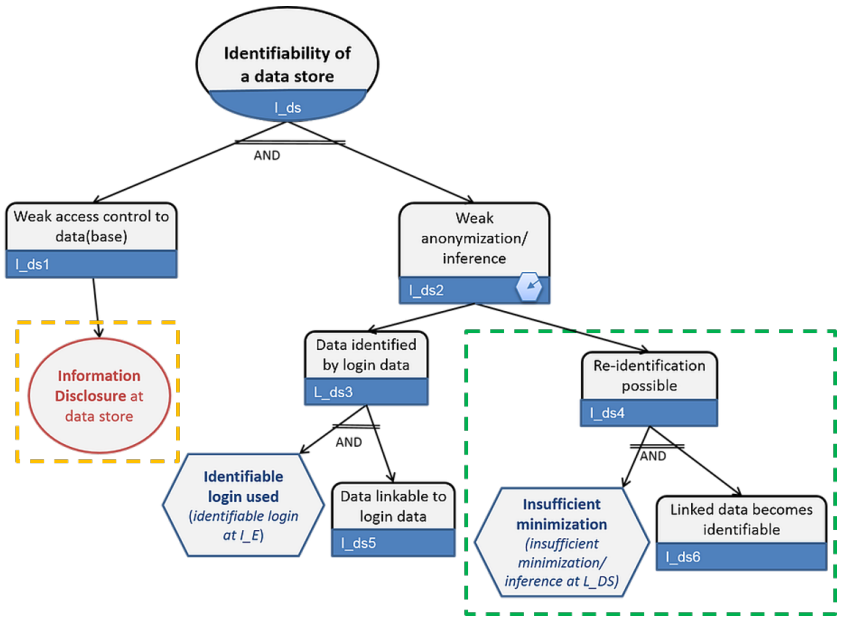


Figure A.5.: Identifiability of data store (provider database)
Source: based on Wuyts et al. (2014)

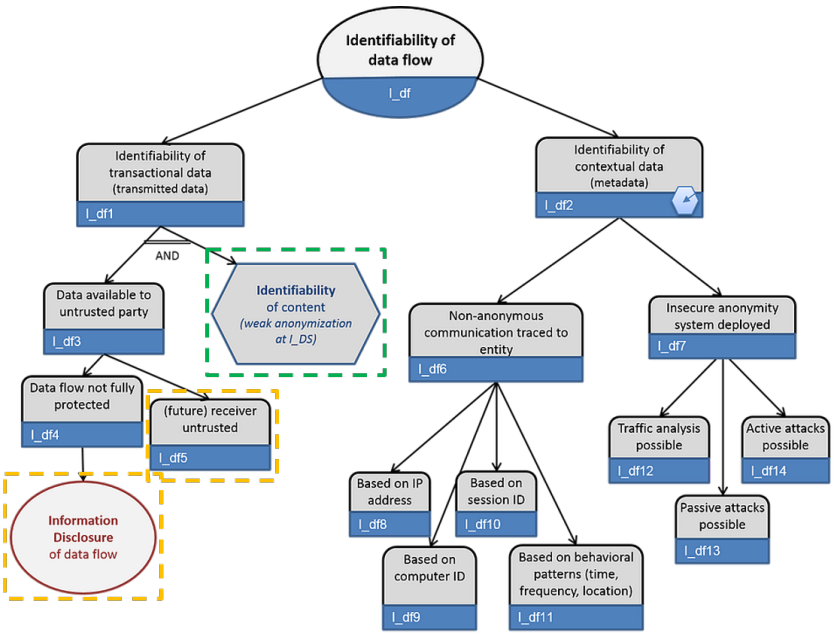


Figure A.6.: Identifiability of data flow (platform & wearable data stream)
Source: based on Wuyts et al. (2014)

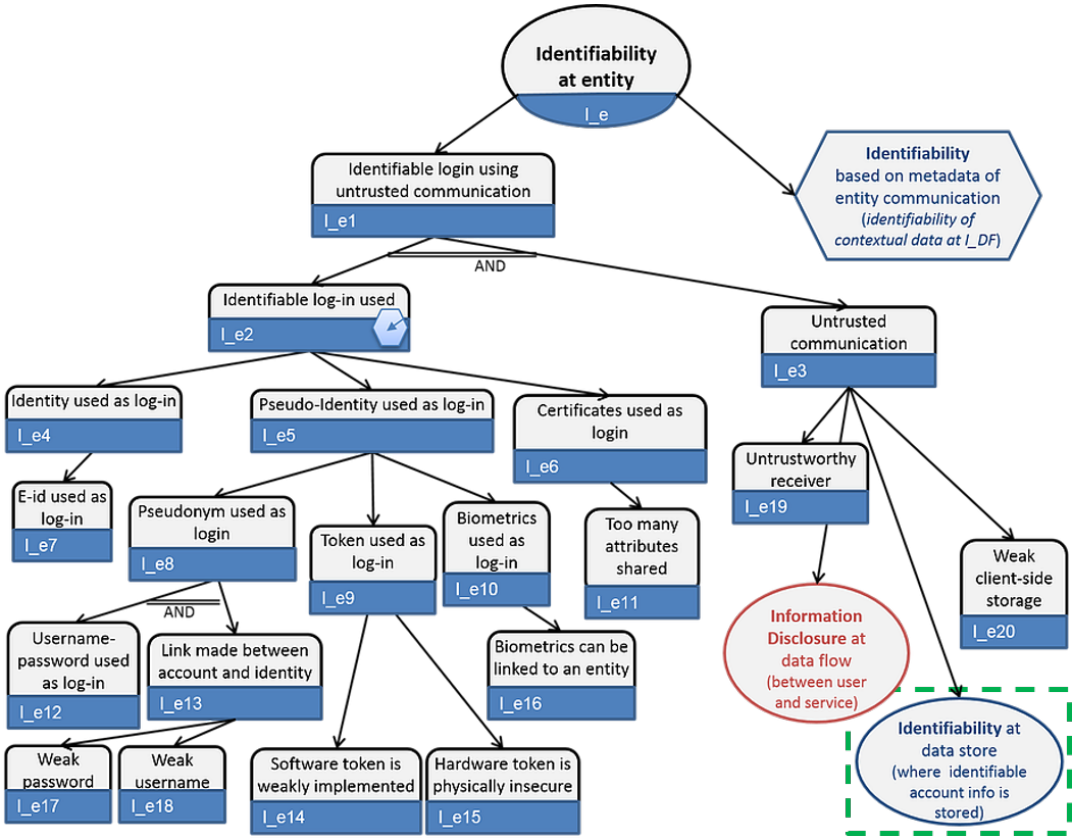


Figure A.7.: Identifiability of entity (user & wearable)
Source: based on Wuyts et al. (2014)

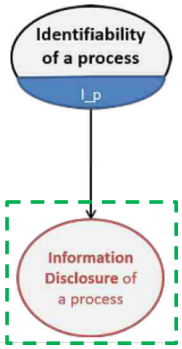


Figure A.8.: Identifiability of process (platform & service)
Source: based on Wuyts et al. (2014)

A.2.3. Unawareness & Non-Compliance Threat trees

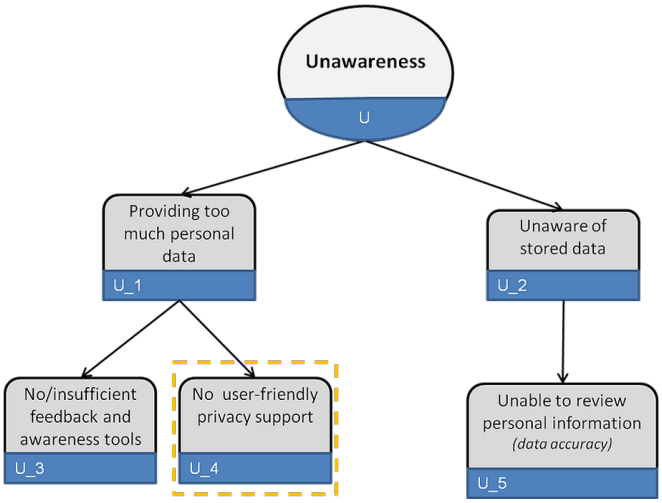


Figure A.9.: Unawareness of entity (user & wearable)
Source: based on Wuyts et al. (2014)

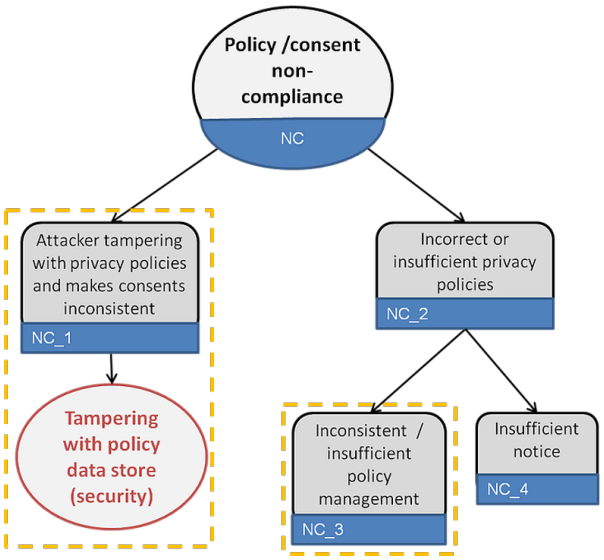


Figure A.10.: Policy and consent non-compliance (whole system)
Source: based on Wuyts et al. (2014)

B. Interview Guide

Einführung

- Einwilligung zur Aufnahme des Interviews
- Einführung ins Thema der Arbeit

Infos zum Befragten

- Rolle im Unternehmen/Organisation
- Relevante Erfahrung

Welchen Einfluss hat das Thema Data Privacy in ihrem Unternehmen?

a) Speicherung / Verarbeitung personenbezogener Daten
b) Einfluss auf Datenanalysen
c) Trends (Bezug auf Analytics / Cloud)

Wie erkennen und bewerten sie Datenschutz-Risiken in ihrem Unternehmen? (Tools / Prozesse / Methoden)

a) Spezialfall der Speicherung personenbezogener Daten in der Cloud
b) Klassifizierung und Bewertung identifizierbarer Informationen
c) Klassifizierung und Bewertung sensibler Informationen
d) Bewertung möglicher Risiken in Bezug auf sensible Daten

Welche Maßnahmen setzen sie ein um Data Privacy zu gewährleisten?

a) Anonymisierung / Pseudonymisierung von Daten
b) Bewertung der Maßnahmen
c) Wann sind Maßnahmen ausreichend?

Sind ihnen De-Identification Methoden bekannt? Wie schätzen sie diese ein?

a) Potential & Anwendungsfälle
b) Risiken & Nachteile

Untersuchung des Anwendungsfalls von Gesundheits-/Aktivitätsdaten, die von einer Smartwatch erfasst werden: Welche Anforderungen & Besonderheiten bringt dies aus ihrer Sicht mit sich?

a) Erfassung persönlicher Gesundheits-/Aktivitätsdaten
b) Erfassung über Smartwatch und Speicherung / Verarbeitung in der Cloud
c) Einfluss gesetzlicher Regularien

Abschluss

- Weitere Anmerkungen / Hinweise zu dem Thema
- Kontakte als weitere Interview-Partner

C. Local Probabilistic k-anonymity Simulation

C.1. Python Script

```
from random import randint
import matplotlib.pyplot as plt
import collections
from tqdm import tqdm_notebook as tqdm

#Input
g = 3000000         #number of groups or distinct values
r = 50000000        #number of records
n = 400             #number of iterations

#list in which the calculated k values will be stored in
k_values = []

#looping through n iterations
for iter in tqdm(range(0,n)):

    #defining an empty list with g+1 items which are initialized with 0. This list will the
    ↪  number of occurences of each group
    counting = [0]*(g+1)

    #Assigning values of range g to r records
    for i in range(0,r):
        #instead of saving the random number, we directly increase the counting value of the
        ↪  corresponding list entrie
        counting[randint(1,g)] += 1

    #deleting the 0 values from the counting list. The k-anonymity value cannot be 0
    count_values_0 = [i for i in counting if i!=0]
    #determine value of k (k-anonymity) and append the k_values list with it
    k_values.append(min(count_values_0))

#list to count the occurence of different k values
count_k_values = []
#looping through all possible k values (the maximum is r)
```

```
33 for i in range(0,r+1):
34     #save number of respective k_value in a list which stores all counts
35     count_k_values.append(k_values.count(i))
36
37
38 count_k_values_inc = []
39 #calculate floating sum of occurences to determine the 95% and 99% threshold values
40 count_k_values_inc.append(n-count_k_values[0])
41 for i in range(1,r):
42     count_k_values_inc.append(count_k_values_inc[i-1]-count_k_values[i])
43
44 print("Most frequent k value:", count_k_values.index(max(count_k_values)))
45 print("95% serve k anonymity with k =", count_k_values_inc.index(list(filter(lambda i: i <
   ↪  0.95*n, count_k_values_inc))[0]))
46 print("99% serve k anonymity with k =", count_k_values_inc.index(list(filter(lambda i: i <
   ↪  0.99*n, count_k_values_inc))[0]))
47
48
49 ##########Plotting and saving the result##########
50
51 text1 = "Most frequent: k = " + str(count_k_values.index(max(count_k_values)))
52 text2 = "\n95%: k = " +str(count_k_values_inc.index(list(filter(lambda i: i < 0.95*n,
   ↪  count_k_values_inc))[0]))
53 text3 = "\n99%: k = " +str(count_k_values_inc.index(list(filter(lambda i: i < 0.99*n,
   ↪  count_k_values_inc))[0]))
54 plt.plot(count_k_values)
55 plt.xlim(0,5)
56 plt.xlabel('k', fontsize=12)
57 plt.ylabel('frequency', fontsize=12)
58 plt.text(0.5,95,text1+text2+text3,fontsize=12)
59 plt.savefig('simu.png',bbox_inches='tight')
60 plt.show()
```

C.2. Simulation Results

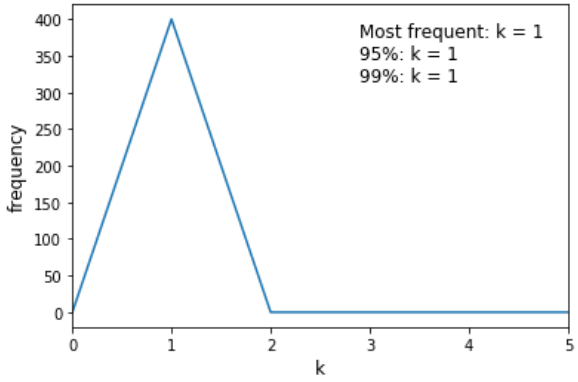


Figure C.1.: simulation with r = 50,000,000, g = 16,848,000 and n = 400
Computation time: 7:18h

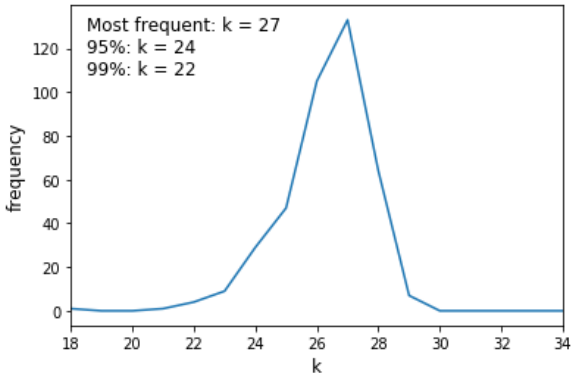


Figure C.2.: simulation with r = 50,000,000, g = 842,400 and n = 400
Computation time: 6:15h

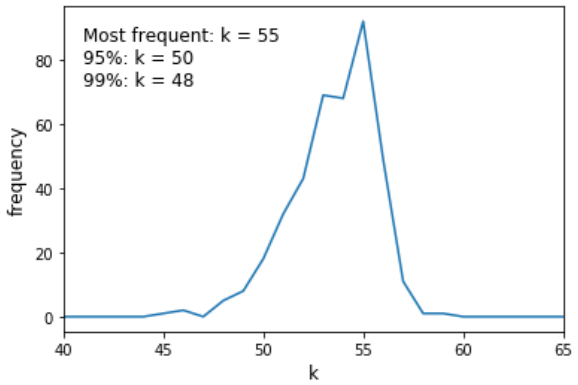


Figure C.3.: simulation with r = 50,000,000, g = 518,400 and n = 400
Computation time: 8:22h

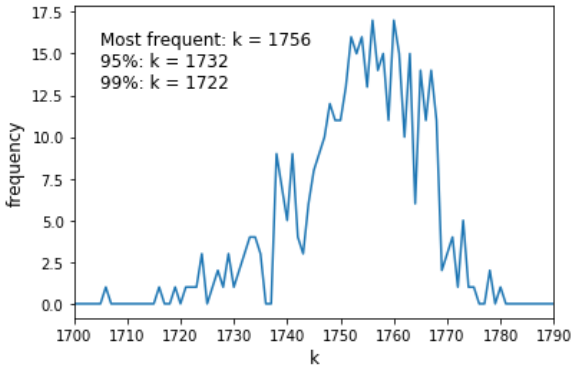


Figure C.4.: simulation with r = 50,000,000, g = 25,920 and n = 400
Computation time: 6:05h

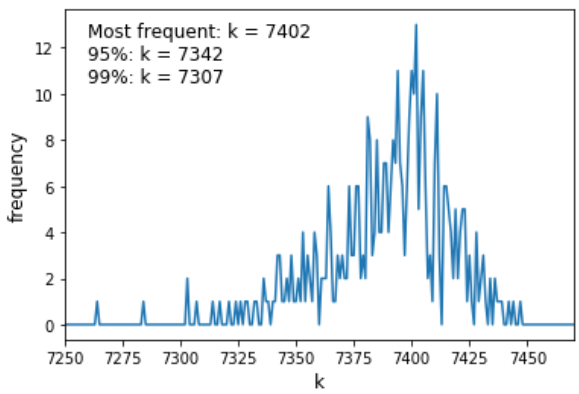


Figure C.5.: simulation with r = 50,000,000, g = 6,480 and n = 400
Computation time: 6:03h

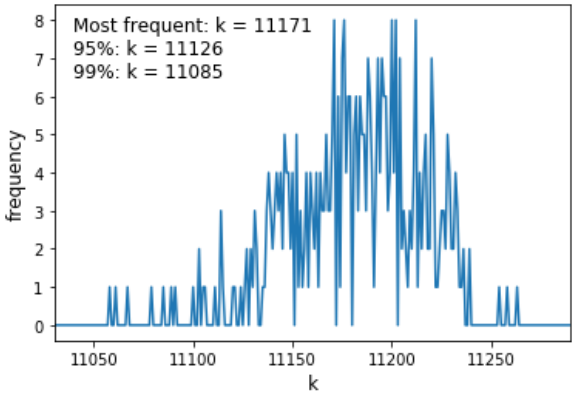


Figure C.6.: simulation with r = 50,000,000, g = 4,320 and n = 400
Computation time: 6:33h

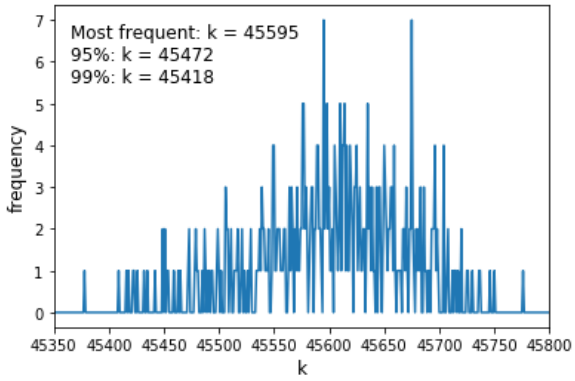


Figure C.7.: simulation with r = 50,000,000, g = 1,080 and n = 400
Computation time: 10:33h

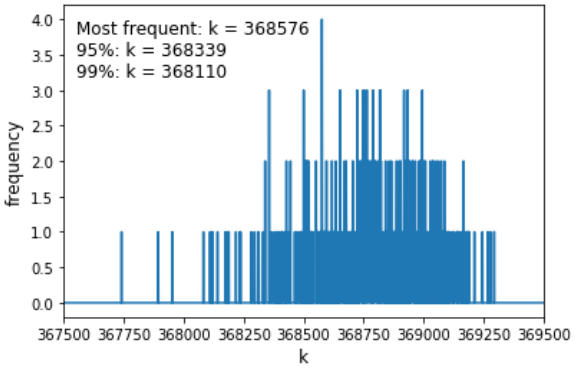


Figure C.8.: simulation with r = 50,000,000, g = 135 and n = 400
Computation time: 6:18h

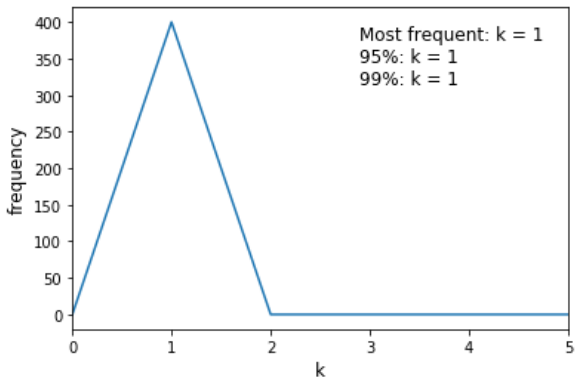


Figure C.9.: simulation with r = 2,000,000, g = 518,400 and n = 400
Computation time: 19min

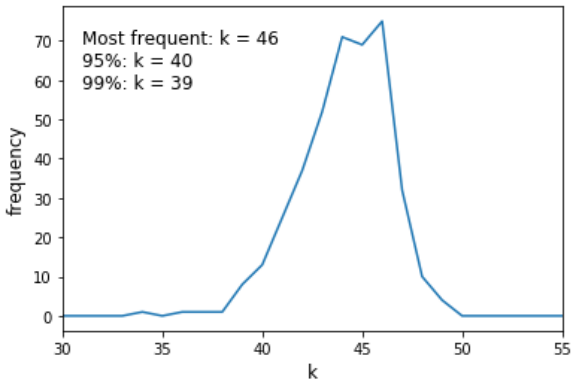


Figure C.10.: simulation with r = 2,000,000, g = 25,920 and n = 400
Computation time: 24min

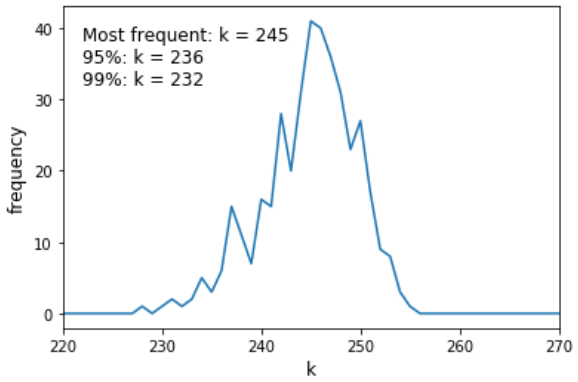


Figure C.11.: simulation with r = 2,000,000, g = 6,480 and n = 400
Computation time: 42min

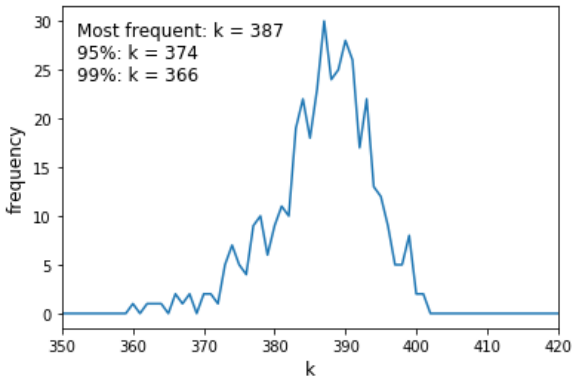


Figure C.12.: simulation with r = 2,000,000, g = 4,320 and n = 400
Computation time: 22min

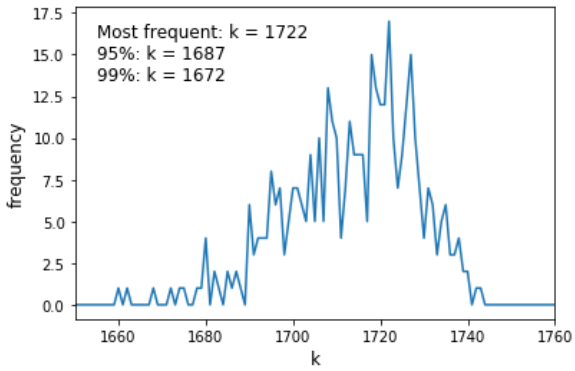


Figure C.13.: simulation with r = 2,000,000, g = 1,080 and n = 400
Computation time: 30min

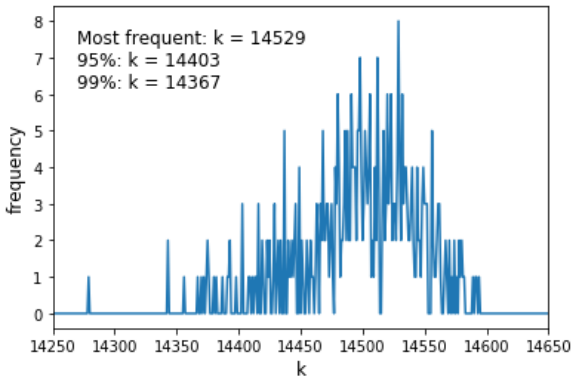


Figure C.14.: simulation with r = 2,000,000, g = 135 and n = 400
Computation time: 22min

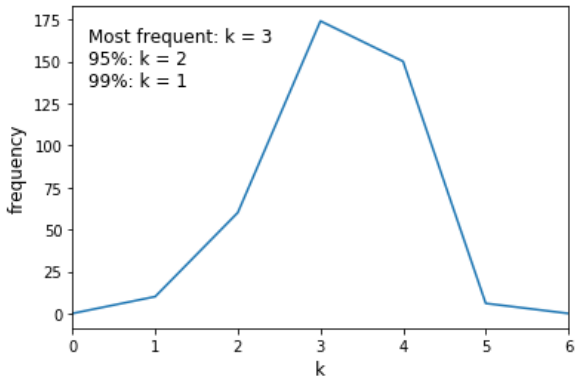


Figure C.15.: simulation with r = 100,000, g = 6,480 and n = 400
Computation time: 1min

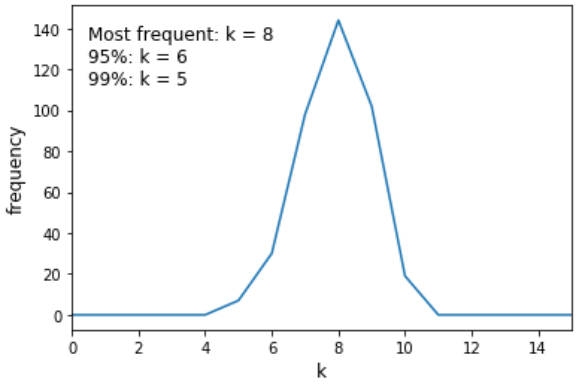


Figure C.16.: simulation with r = 100,000, g = 4,320 and n = 400
Computation time: 1min

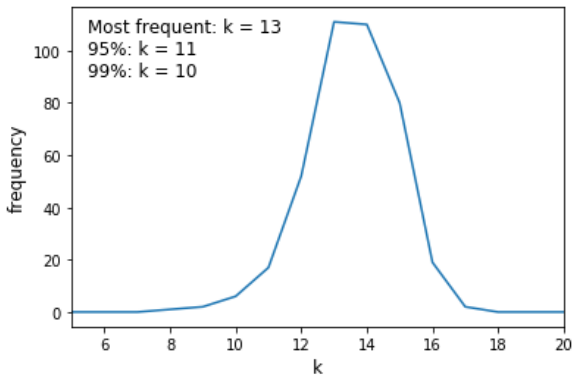


Figure C.17.: simulation with r = 100,000, g = 3,250 and n = 400
Computation time: 1min

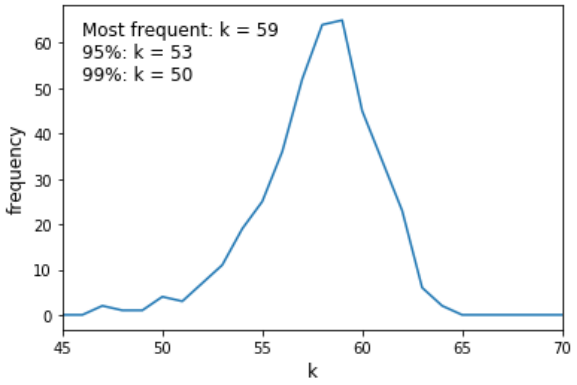


Figure C.18.: simulation with r = 100,000, g = 1,150 and n = 400
Computation time: 1min

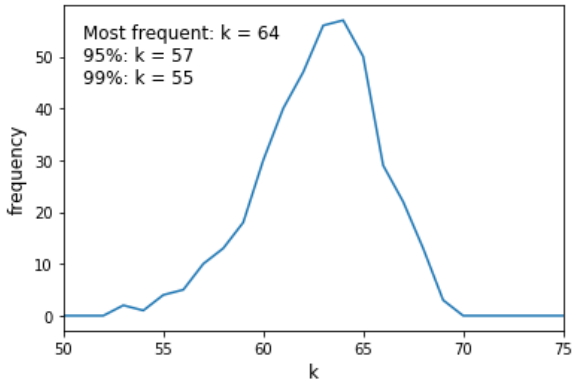


Figure C.19.: simulation with r = 100,000, g = 1,080 and n = 400
Computation time: 1min

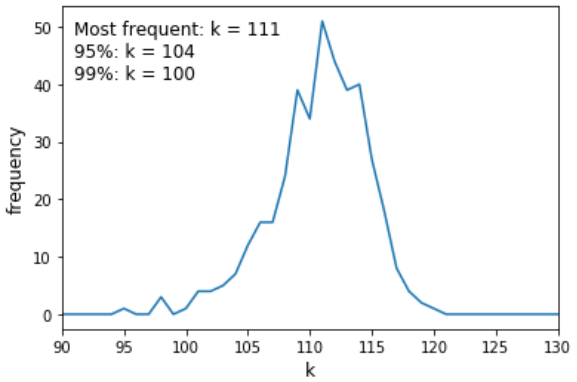


Figure C.20.: simulation with r = 100,000, g = 680 and n = 400
Computation time: 1min

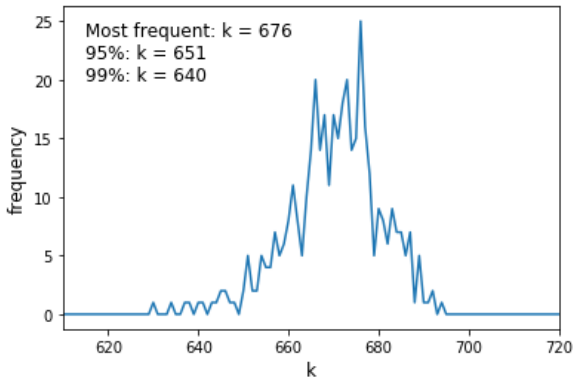


Figure C.21.: simulation with r = 100,000, g = 135 and n = 400
Computation time: 4min

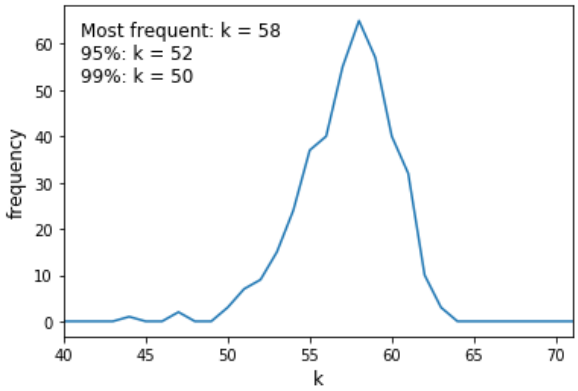


Figure C.22.: simulation with r = 90,000, g = 1,050 and n = 400
Computation time: 30s

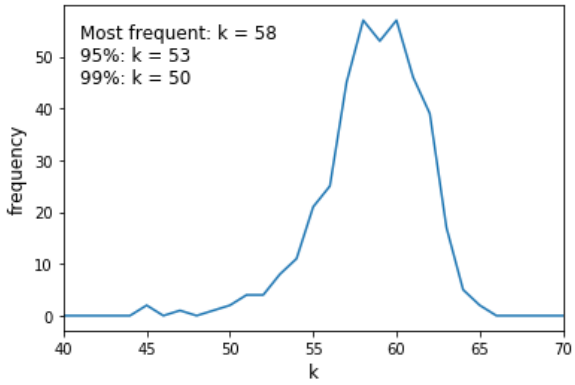


Figure C.23.: simulation with r = 80,000, g = 920 and n = 400
Computation time: 1min

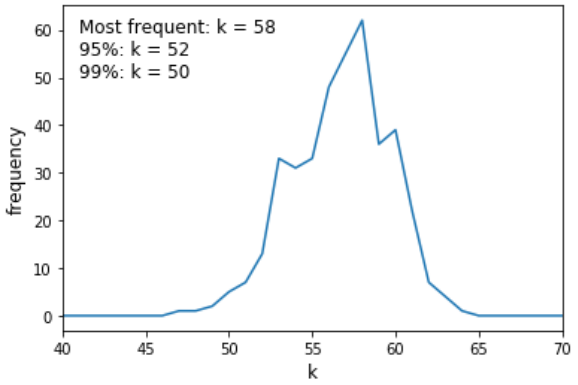


Figure C.24.: simulation with r = 66,666, g = 790 and n = 400
Computation time: 1min

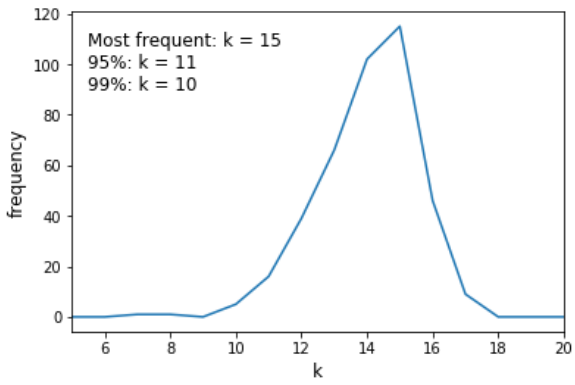


Figure C.25.: simulation with r = 33,333, g = 1,100 and n = 400
Computation time: 1min

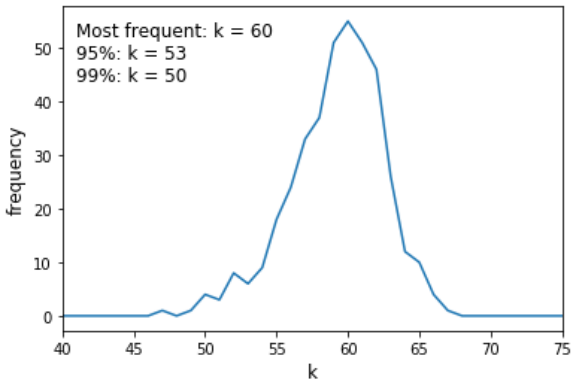


Figure C.26.: simulation with r = 33,333, g = 390 and n = 400
Computation time: 1min

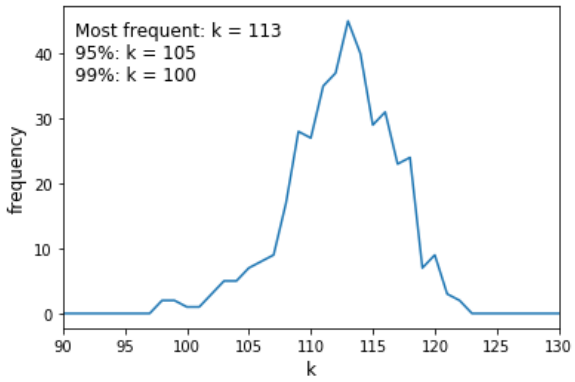


Figure C.27.: simulation with r = 33,333, g = 230 and n = 400
Computation time: 1min

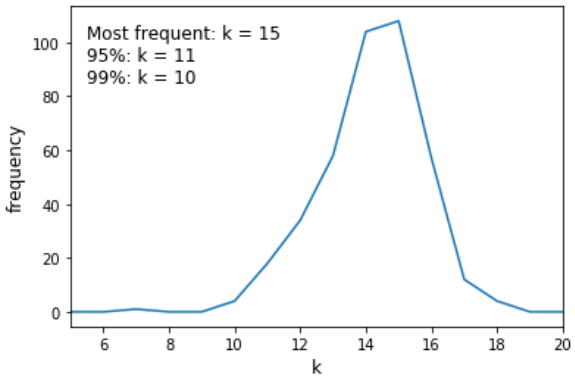


Figure C.28.: simulation with r = 10,000, g = 350 and n = 400
Computation time: 30s

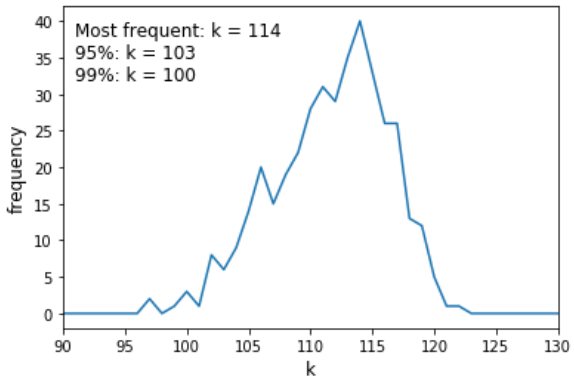


Figure C.29.: simulation with r = 10,000,
g = 72 and n = 400
Computation time: 20s

Bibliography

Agrawal, R., Kiernan, J., Srikant, R., & Xu, Y. (2004). Order preserving encryption for numeric data (P. Valduriez, G. Weikum, A. C. König, & S. Dessloch, Eds.). In P. Valduriez, G. Weikum, A. C. König, & S. Dessloch (Eds.), *Proceedings of the 2004 acm sigmod international conference on management of data*, New York, New York, USA, ACM Press. https://doi.org/10.1145/1007568.1007632

Alloghani, M., M. Alani, M., Al-Jumeily, D., Baker, T., Mustafina, J., Hussain, A., & J. Aljaaf, A. (2019). A systematic review on the status and progress of homomorphic encryption technologies. *Journal of Information Security and Applications*, *48*, 1–9. https://doi.org/10.1016/j.jisa.2019.102362

Alqahtani, H. S., & Sant, P. (2016). A multi-cloud approach for secure data storage on smart device, In *2016 sixth international conference on digital information and communication technology and its applications (dictap)*, IEEE. https://doi.org/10.1109/DICTAP.2016.7544002

Alshugran, T., Dichter, J., & Faezipour, M. (2015). Formally expressing hipaa privacy policies for web services, In *2015 ieee international conference on electro/information technology (eit)*, IEEE. https://doi.org/10.1109/EIT.2015.7293356

Article 29 DP Working Party. (2014). Opinion 05/2014 on anonymisation techniques (Article 29 DP Working Party, Ed.).

Bieker, F., Friedewald, M., Hansen, M., Obersteller, H., & Rost, M. (2016). A process for data protection impact assessment under the european general data protection regulation (S. Schiffner, J. Serna, D. Ikonomou, & K. Rannenberg, Eds.). In S. Schiffner, J. Serna, D. Ikonomou, & K. Rannenberg (Eds.), *Privacy technologies and policy*, Cham, Springer International Publishing.

Bondel, G., Munilla Garrido, G., Baumer, K., & Matthes, F. (2020). Towards a privacy-enhancing tool based on de- identification methods (to be published in june 2020), In *Twenty-third pacific asia conference on information systems*.

Bourka, A., & Drogkaris, P. (2018). *Recommendations on shaping technology according to gdpr provisions: An overview on data pseudonymisation: European union agency for network and information security*. Heraklion, ENISA.

Branco, E. C., Monteiro, J. M., Reis, R., & Machado, J. C. (2016). A flexible mechanism for data confidentiality in cloud database scenarios, In *Proceedings of the 18th international conference on enterprise information systems*, SCITEPRESS - Science and and Technology Publications. https://doi.org/10.5220/0005872503590368

Chakraborty, N., & Patra, G. K. (2014). Functional encryption for secured big data analytics. *International Journal of Computer Applications*, *107*(16), 19–22. https://doi.org/10.5120/18836-0359

Chatfield, A. A., Parker, J. L., & Egeler, P. W. (2018). Zippy safe harbor de-identification macros: Grand valley state university and spectrum health office of research administration, In *Sas conference proceedings*.

Chatterjee, A., & Sengupta, I. (2018). Translating algorithms to handle fully homomorphic encrypted data on the cloud. *IEEE Transactions on Cloud Computing*, *6*(1), 287–300. https://doi.org/10.1109/TCC.2015.2481416

Ciriani, V., Di Vimercati, S. D. C., Foresti, S., Jajodia, S., Paraboschi, S., & Samarati, P. (2010). Combining fragmentation and encryption to protect privacy in data storage. *ACM Transactions on Information and System Security*, *13*(3), 1–33. https://doi.org/10.1145/1805974.1805978

Cormode, G., Jha, S., Kulkarni, T., Li, N., Srivastava, D., & Wang, T. (2018). Privacy at scale (G. Das, C. Jermaine, & P. Bernstein, Eds.). In G. Das, C. Jermaine, & P. Bernstein (Eds.), *Proceedings of the 2018 international conference on management of data - sigmod '18*, New York, New York, USA, ACM Press. https://doi.org/10.1145/3183713.3197390

Dankar, F. K., El Emam, K., Neisa, A., & Roffey, T. (2012). Estimating the re-identification risk of clinical data sets. *BMC medical informatics and decision making*, *12*, 66. https://doi.org/10.1186/1472-6947-12-66

Dautov, R., & Distefano, S. (2018). Vs-driven big data process development (S. Balsamo, A. Marin, & E. Vicario, Eds.). In S. Balsamo, A. Marin, & E. Vicario (Eds.), *New frontiers in quantitative methods in informatics*, Cham, Springer International Publishing.

Deng, M., Wuyts, K., Scandariato, R., Preneel, B., & Joosen, W. (2011). A privacy threat analysis framework: Supporting the elicitation and fulfillment of privacy requirements. *Requirements Engineering*, *16*(1), 3–32. https://doi.org/10.1007/s00766-010-0115-7

Desfontaines, D. (2017). K-anonymity, the parent of all privacy definitions. Retrieved May 10, 2020, from https://desfontain.es/privacy/k-anonymity.html

Domingo-Ferrer, J., & Mateo-Sanz, J. M. (2002). Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering*, *14*(1), 189–201. https://doi.org/10.1109/69.979982

Domingo-Ferrer, J., Farràs, O., Ribes-González, J., & Sánchez, D. (2019). Privacy-preserving cloud computing on sensitive data: A survey of methods, products and challenges. *Computer Communications*, *140-141*, 38–60. https://doi.org/10.1016/j.comcom.2019.04.011

Dwork, C. (2006). Differential privacy (D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, M. Bugliesi, B. Preneel, V. Sassone, & I. Wegener, Eds.). In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, M. Bugliesi, B. Preneel, V. Sassone, &

I. Wegener (Eds.), *Automata, languages and programming*. Berlin, Heidelberg, Springer Berlin Heidelberg. https://doi.org/10.1007/11787006_1

El Emam, K. (2013). *Guide to the de-identification of personal health information*. Hoboken, CRC Press.

El-Yahyaoui, A., & Ech-Chrif El Kettani, M. D. (2018). Data privacy in cloud computing, In *2018 4th international conference on computer and technology applications (iccta)*, IEEE. https://doi.org/10.1109/CATA.2018.8398650

European Parliament and Council of the European Union. (2016). Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation): General data protection regulation. Retrieved June 9, 2020, from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679

Fan, W., He, J., Guo, M., Li, P., Han, Z., & Wang, R. (2020). Privacy preserving classification on local differential privacy in data centers. *Journal of Parallel and Distributed Computing*, *135*, 70–82. https://doi.org/10.1016/j.jpdc.2019.09.009

Gaidhani, D. (2017). A survey report on techniques for data confidentiality in cloud computing. *International Journal of Advanced Research in Computer Science*, *8*(8), 389–394. https://doi.org/10.26483/ijarcs.v8i8.4746

Garfinkel, S. L. (2015). De-identification of personal information: Nistir 8053. https://doi.org/10.6028/NIST.IR.8053

Gates, B. (2013). Bill gates and president bill clinton on the nsa, safe sex, and american exceptionalism. interview by steven levy. Retrieved May 24, 2020, from https://www.wired.com/2013/11/bill-gates-bill-clinton-wired/

Gerl, A. (2020). *Modelling of a privacy language and efficient policy-based de-identification* (Doctoral dissertation). Universität Passau. Retrieved June 9, 2020, from https://opus4.kobv.de/opus4-uni-passau/frontdoor/index/index/docId/767

Gibbs, S. (2016). Dropbox hack leads to leaking of 68m user passwords on the internet. Retrieved May 24, 2020, from https://www.theguardian.com/technology/2016/aug/31/dropbox-hack-passwords-68m-data-breach

Gläser, J., & Laudel, G. (2009). *Experteninterviews und qualitative inhaltsanalyse als instrumente rekonstruierender untersuchungen* (3., überarb. Aufl.). Wiesbaden, VS Verlag für Sozialwissenschaften.

Grandison, T., Bilger, M., O'Connor, L., Graf, M., Swimmer, M., Schunter, M., Wespi, A., & Zunic, N. (2017). Elevating the discussion on security management: The data centric paradigm, In *2007 2nd ieee/ifip international workshop on business-driven it management*, IEEE. https://doi.org/10.1109/BDIM.2007.375015

Heurix, J., Zimmermann, P., Neubauer, T., & Fenz, S. (2015). A taxonomy for privacy enhancing technologies. *Computers & Security*, *53*, 1–17. https://doi.org/10.1016/j.cose.2015.05.002

Hurlburt, G. F., Miller, K. W., Voas, J. M., & Day, J. M. (2009). Privacy and/or security: Take your pick. *IT Professional*, *11*(4), 52–55. https://doi.org/10.1109/MITP.2009.81

ISO. (2017). Iso 25237:2017 health informatics - pseudonymization. *Berlin*, Beuth Verlag GmbH. https://doi.org/10.31030/2555889

ISO. (2018a). Iso/iec 20889 privacy enhancing data de-identification terminology and classification of techniques.

ISO. (2018b). Iso/iec 27000 information technology - security techniques - information security management systems — overview and vocabulary.

Jeffrey M. Skopek. (2013). Anonymity, the production of goods, and institutional design. *Fordham Law Review*, (82), 1751–1809.

Jung, J., Park, P., Lee, J., Lee, H., Lee, G. K., & Cha, H. S. (2020). A determination scheme for quasi-identifiers using uniqueness and influence for de-identification of clinical data. *Journal of Medical Imaging and Health Informatics*, *10*(2), 295–303. https://doi.org/10.1166/jmihi.2020.2966

Khalid El Makkaoui, Abdellah Ezzati, & Abderrahim Beni Hssane. (2016). Homomorphic encryption as a solution of trust issues in cloud applications: Selected papers, tetuan, morocco, may 25-26, 2015 (Mohammed Al Achhab, Mohamed Lazaar, & Youness Tabii, Eds.). In Mohammed Al Achhab, Mohamed Lazaar, & Youness Tabii (Eds.), *Proceedings of the international conference on big data, cloud and applications: Selected papers, tetuan, morocco, may 25-26, 2015*, CEUR-WS.org. http://ceur-ws.org/Vol-1580/id6.pdf

Kim, J. W., Lim, J. H., Moon, S. M., & Jang, B. (2019). Collecting health lifelog data from smartwatch users in a privacy-preserving manner. *IEEE Transactions on Consumer Electronics*, *65*(3), 369–378. https://doi.org/10.1109/TCE.2019.2924466

Kiyomoto, S., & Miyake, Y. (2014). How to find an appropriate k for k-anonymization, In *2014 eighth international conference on innovative mobile and internet services in ubiquitous computing*, IEEE. https://doi.org/10.1109/IMIS.2014.34

Lavin, R. P. (2006). Hipaa and disaster research: Preparing to conduct research. *Disaster management & response : DMR : an official publication of the Emergency Nurses Association*, *4*(2), 32–37. https://doi.org/10.1016/j.dmr.2006.01.003

Lewis, D. (2014). Icloud data breach: Hacking and celebrity photos. Retrieved May 24, 2020, from https://www.forbes.com/sites/davelewis/2014/09/02/icloud-data-breach-hacking-and-nude-celebrity-photos

Li, N., Li, T., & Venkatasubramanian, S. (2007). T-closeness: Privacy beyond k-anonymity and l-diversity, In *2007 ieee 23rd international conference on data engineering*, IEEE. https://doi.org/10.1109/ICDE.2007.367856

Li, Y., Yang, D., & Hu, X. (2020). A differential privacy-based privacy-preserving data publishing algorithm for transit smart card data. *Transportation Research Part C: Emerging Technologies*, *115*, 102634. https://doi.org/10.1016/j.trc.2020.102634

Liu, X., Deng, R., Choo, K.-K. R., Yang, Y., & Pang, H. (2019). Privacy-preserving outsourced calculation toolkit in the cloud. *IEEE Transactions on Dependable and Secure Computing*, 1. https://doi.org/10.1109/TDSC.2018.2816656

Lu, R., Zhu, H., Liu, X., Liu, J. K., & Shao, J. (2014). Toward efficient and privacy-preserving computing in big data era. *IEEE Network*, *28*(4), 46–50. https://doi.org/10.1109/MNET.2014.6863131

Lu, X., & Au, M. H. (2017). An introduction to various privacy models (G. Kessler, M.-H. Au, & R. K.-K. Choo, Eds.). In G. Kessler, M.-H. Au, & R. K.-K. Choo (Eds.), *Mobile security and privacy*. Cambridge, MA, Syngress.

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). L-diversity. *ACM Transactions on Knowledge Discovery from Data*, *1*(1), 3–es. https://doi.org/10.1145/1217299.1217302

Manning, A. M., Haglin, D. J., & Keane, J. A. (2008). A recursive search algorithm for statistical disclosure assessment. *Data Mining and Knowledge Discovery*, *16*(2), 165–196. https://doi.org/10.1007/s10618-007-0078-6

Mansfield-Devine, S. (2014). Masking sensitive data. *Network Security*, *2014*(10), 17–20. https://doi.org/10.1016/S1353-4858(14)70104-7

Martínez, S., Sánchez, D., Valls, A., & Batet, M. (2012). Privacy protection of textual attributes through a semantic-based masking method. *Information Fusion*, *13*(4), 304–314. https://doi.org/10.1016/j.inffus.2011.03.004

McWaters, J., Blake, M., Galaski, R., Soni, H., & Majumdar, I. (2019). The next generation of data-sharing in financial services: Using privacy enhancing techniques to unlock new value. Retrieved June 9, 2020, from https://www.weforum.org/whitepapers/the-next-generation-of-data-sharing-in-financial-services-using-privacy-enhancing-techniques-to-unlock-new-value

Meunier, M.-A., Pirzada, Z., & Gardner, D. (2019). Market guide for data masking. Retrieved June 9, 2020, from https://www.gartner.com/en/documents/3975500/market-guide-for-data-masking

Micciancio, D. (2010). A first glimpse of cryptography's holy grail. *Communications of the ACM*, *53*(3), 96. https://doi.org/10.1145/1666420.1666445

Mukherjee, J., Datta, B., Banerjee, R., & Das, S. (2015). Dwt difference modulation based novel steganographic algorithm (S. Jajodia & C. Mazumdar, Eds.). In S. Jajodia & C. Mazumdar (Eds.), *Information systems security*. Cham, Heidelberg, New York, Dordrecht, London, Springer.

Nelson, G. S. (Ed.). (2015). *Practical implications of sharing data: A primer on data privacy, anonymization, and de-identification*, Vol. 1884-2015. Retrieved June 9, 2020, from https://support.sas.com/resources/papers/proceedings15/1884-2015.pdf

Ogburn, M., Turner, C., & Dahal, P. (2013). Homomorphic encryption. *Procedia Computer Science*, *20*, 502–509. https://doi.org/10.1016/j.procs.2013.09.310

Ohm, P. (2009). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, *57*.

Papadimitriou, A., Bhagwan, R., Chandran, N., Ramjee, R., Haeberlen, A., Singh, H., Modi, A., & Badrinarayanan, S. (2016). Big data analytics over encrypted datasets with seabed, In *Proceedings of osdi '16: 12th usenix symposium on operating systems design and implementation*. Berkeley, CA, USENIX Association.

Pentecost, M. J. (2004). Hipaa and the law of unintended consequences. *Journal of the American College of Radiology : JACR*, *1*(3), 164–165. https://doi.org/10.1016/j.jacr.2003.12.023

Pfitzmann, A., & Hansen, M. (2010). A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management. Retrieved June 9, 2020, from https://dud.inf.tu-dresden.de/literatur/Anon_Terminology_v0.34.pdf

Prasser, F., Eicher, J., Spengler, H., Bild, R., & Kuhn, K. A. (2020). Flexible data anonymization using arx—current status and challenges ahead. *Software: Practice and Experience*. https://doi.org/10.1002/spe.2812

Prasser, F., Kohlmayer, F., Spengler, H., & Kuhn, K. A. (2018). A scalable and pragmatic method for the safe sharing of high-quality health data. *IEEE journal of biomedical and health informatics*, *22*(2), 611–622. https://doi.org/10.1109/JBHI.2017.2676880

Robles-González, A., Parra-Arnau, J., & Forné, J. (2020). A linddun-based framework for privacy threat analysis on identification and authentication processes. *Computers & Security*, 33. https://doi.org/10.1016/j.cose.2020.101755

Samarati, P., & Sweeney, L. (1998). Protecting privacy when disclosing information: K-anonymity and its enforcement through generalization and suppression. Computer Science Laboratory, SRI International. Retrieved June 1, 2020, from http://www.csl.sri.com/papers/sritr-98-04/

Samonas, S., & Coss, D. (2014). The cia strikes back: Redefining confidentiality, integrity and availability in security. *Journal of Information System Security*, *Volume 10*(3), 21–45.

Sánchez, D., & Batet, M. (2017). Privacy-preserving data outsourcing in the cloud via semantic data splitting. *Computer Communications*, *110*, 187–201. https://doi.org/10.1016/j.comcom.2017.06.012

Schoppmann, M. J., & Sanders, D. L. (2004). Hipaa compliance: The law, reality, and recommendations. *Journal of the American College of Radiology : JACR*, *1*(10), 728–733. https://doi.org/10.1016/j.jacr.2004.03.017

Sidorov, V., & Ng, W. K. (2016). Towards performance evaluation of oblivious data processing emulated with partially homomorphic encryption schemes, In *2016 ieee 2nd international conference on big data security on cloud (bigdatasecurity)*, IEEE. https://doi.org/10.1109/BigDataSecurity-HPSC-IDS.2016.36

Solove, D. J. (2015). The meaning and value of privacy (B. Roessler & D. Mokrosinska, Eds.). In B. Roessler & D. Mokrosinska (Eds.), *Social dimensions of privacy*. Cambridge, Cambridge University Press. https://doi.org/10.1017/CBO9781107280557.005

Song, H., Fink, G. A., & Jeschke, S. (Eds.). (2018). *Security and privacy in cyber-physical systems: Foundations, principles, and applications* (First edition). Hoboken, NJ, Wiley IEEE Press.

Soria-Comas, J., & Domingo-Ferrer, J. (2016). Big data privacy: Challenges to privacy principles and models. *Data Science and Engineering*, *1*(1), 21–28. https://doi.org/10.1007/s41019-015-0001-x

Statista (Ed.). (2019). Statista global consumer survey. Retrieved May 16, 2020, from https://de.statista.com/prognosen/999765/umfrage-in-deutschland-zu-beliebten-smartwatch-marken

Sweeney, L. (2002). K-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(05), 557–570. https://doi.org/10.1142/S0218488502001648

Sweeney, L. (2000). Simple demographics often identify people uniquely: Privacy working paper 3. https://doi.org/10.1184/R1/6625769.v1

Tamburri, D. A. (2020). Design principles for the general data protection regulation (gdpr): A formal concept analysis and its evaluation. *Information Systems*, *91*, 101469. https://doi.org/10.1016/j.is.2019.101469

Tebaa, M., & Hajji, S. (2014). Secure cloud computing through homomorphic encryption. Retrieved May 20, 2020, from https://arxiv.org/abs/1409.0829

Terrovitis, M., Poulis, G., Mamoulis, N., & Skiadopoulos, S. (2017). Local suppression and splitting techniques for privacy preserving publication of trajectories. *IEEE Transactions on Knowledge and Data Engineering*, *29*(7), 1466–1479. https://doi.org/10.1109/TKDE.2017.2675420

Terzi, D. S., Terzi, R., & Sagiroglu, S. (2015). A survey on security and privacy issues in big data, In *2015 10th international conference for internet technology and secured transactions (icitst)*, IEEE. https://doi.org/10.1109/ICITST.2015.7412089

The Royal Society (Ed.). (2019). Protecting privacy in practice: The current use, development and limits of privacy enhancing technologies in data analysis. *London*, The Royal Society. Retrieved May 17, 2020, from https://royalsociety.org/topics-policy/projects/privacy-enhancing-technologies/

Tomashchuk, O., van Landuyt, D., Pletea, D., Wuyts, K., & Joosen, W. (2019). A data utility-driven benchmark for de-identification methods (S. Gritzalis, E. R. Weippl, S. K. Katsikas, G. Anderst-Kotsis, A. M. Tjoa, & I. Khalil, Eds.). In S. Gritzalis, E. R. Weippl, S. K. Katsikas, G. Anderst-Kotsis, A. M. Tjoa, & I. Khalil (Eds.), *Trust, privacy and security in digital business*. Cham, Springer International Publishing. https://doi.org/10.1007/978-3-030-27813-7{\textunderscore}5

UN Global Working Group (Ed.). (2019). Un handbook on privacy-preserving computation techniques. Retrieved May 17, 2020, from https://marketplace.officialstatistics.org/privacy-preserving-techniques-handbook

Venkataramanan, N., & Shriram, A. (2016). *Data privacy: Principles and practice*. Boca Raton, CRC Press.

Wang, D., Guo, B., Shen, Y., Cheng, S.-J., & Lin, Y.-H. (2017). A faster fully homomorphic encryption scheme in big data, In *2017 ieee 2nd international conference on big data analysis (icbda)(*, IEEE. https://doi.org/10.1109/ICBDA.2017.8078836

Warren, S. D., & Brandeis, L. D. (1890). The right to privacy. *Harvard Law Review*, *4*(5), 193. https://doi.org/10.2307/1321160

Westin, A. F. (1967). *Privacy and freedom*. New York, IG Publishing.

Will, M. A., & Ko, R. K. (2015). A guide to homomorphic encryption, In *The cloud security ecosystem*. Elsevier. https://doi.org/10.1016/B978-0-12-801595-7.00005-7

Wood, A., Altman, M., Bembenek, A., Bun, M., Gaboardi, M., Honaker, J., Nissim, K., O'Brien, D. R., Steinke, T., & Vadhan, S. (2018). Differential privacy: A primer for a non-technical

audience. *Vanderbilt Journal of Entertainment & Technology Law*, *21*(1), 209–275. Retrieved April 20, 2020, from http://www.jetlaw.org/journal-archives/volume-21/volume-21-issue-1/differential-privacy-a-primer-for-a-non-technical-audience/

Wu, F. T. (2012). Defining privacy and utility in data sets. *84 University of Colorado Law Review 1117 (2013); 2012 TRPC*, 1117–1177. https://doi.org/10.2139/ssrn.2031808

Wuyts, K., van Landuyt, D., Hovsepyan, A., & Joosen, W. (2018). Effective and efficient privacy threat modeling through domain refinements (H. M. Haddad, R. L. Wainwright, & R. Chbeir, Eds.). In H. M. Haddad, R. L. Wainwright, & R. Chbeir (Eds.), *Proceedings of the 33rd annual acm symposium on applied computing - sac '18*, New York, New York, USA, ACM Press. https://doi.org/10.1145/3167132.3167414

Wuyts, K., Scandariato, R., & Joosen, W. (2014). Lind(d)un privacy threat tree catalog: Cw reports. Retrieved June 5, 2020, from https://lirias.kuleuven.be/1656217