information

Review

Anonymization Procedures for Tabular Data: An Explanatory Technical and Legal Synthesis

Robert Aufschläger [1,*], Jakob Folz [1], Elena März [2], Johann Guggumos [2], Michael Heigl [1], Benedikt Buchner [2] and Martin Schramm [1]

[1] Technology Campus Vilshofen, Deggendorf Institute of Technology, 94474 Vilshofen an der Donau, Germany; jakob.folz@th-deg.de (J.F.); michael.heigl@th-deg.de (M.H.); martin.schramm@th-deg.de (M.S.)
[2] Faculty of Law, University of Augsburg, 86159 Augsburg, Germany; elena.maerz@jura.uni-augsburg.de (E.M.); johann.guggumos@jura.uni-augsburg.de (J.G.); benedikt.buchner@jura.uni-augsburg.de (B.B.)
* Correspondence: robert.aufschlaeger@th-deg.de

Abstract: In the European Union, Data Controllers and Data Processors, who work with personal data, have to comply with the General Data Protection Regulation and other applicable laws. This affects the storing and processing of personal data. But some data processing in data mining or statistical analyses does not require any personal reference to the data. Thus, personal context can be removed. For these use cases, to comply with applicable laws, any existing personal information has to be removed by applying the so-called anonymization. However, anonymization should maintain data utility. Therefore, the concept of anonymization is a double-edged sword with an intrinsic trade-off: privacy enforcement vs. utility preservation. The former might not be entirely guaranteed when anonymized data are published as Open Data. In theory and practice, there exist diverse approaches to conduct and score anonymization. This explanatory synthesis discusses the technical perspectives on the anonymization of tabular data with a special emphasis on the European Union's legal base. The studied methods for conducting anonymization, and scoring the anonymization procedure and the resulting anonymity are explained in unifying terminology. The examined methods and scores cover both categorical and numerical data. The examined scores involve data utility, information preservation, and privacy models. In practice-relevant examples, methods and scores are experimentally tested on records from the UCI Machine Learning Repository's "Census Income (Adult)" dataset.

Keywords: emerging technologies and applications; multimedia content management; privacy and trust

Citation: Aufschläger, R.; Folz, J.; März, E.; Guggumos, J.; Heigl, M.; Buchner, B.; Schramm, M. Anonymization Procedures for Tabular Data: An Explanatory Technical and Legal Synthesis. *Information* **2023**, *14*, 487. https://doi.org/10.3390/info14090487

Academic Editors: Jose de Vasconcelos, Hugo Barbosa and Carla Cordeiro

Received: 2 August 2023
Revised: 22 August 2023
Accepted: 29 August 2023
Published: 1 September 2023

1. Introduction

Working with personalized data is a highly risky task. Not only in sensitive sectors like health and finance, personal data has to be protected. Personal data can occur in vast varieties. Nevertheless, in practice, personal data are often stored in structured tabular datasets, and this work focuses on tabular datasets as objects of study.

Violating the regulations in force, such as the General Data Protection Regulation (GDPR) by the European Union (EU), can lead to severe penalties. More importantly, from an ethical perspective, data leakage can cause irreversible and irreparable damage.

However, removing personal information, i.e., called anonymizing, is a challenging task that comes with a trade-off. On the one hand, after anonymizing, no personal references should be possible. This can only be achieved by manipulating or even deleting data. On the other hand, the data utility should be maintained. Hereby, we refer to "data utility" as any measure to rate how useful data are for given tasks.

Furthermore, anonymization is highly task-dependent, and due to the lack of specialized Open Data, Data Controllers and Data Processors cannot rely on given experiences.

In the following, this article looks at the anonymization of tabular data from the legal perspective of the GDPR. We describe practice-relevant anonymization terms, methods, and scores for tabular data in a technical manner while enforcing common terminology and explaining the legal setting for anonymizing tabular data.

This explanatory synthesis aims to distill and organize the wealth of information from a multitude of versatile sources in the context of anonymizing tabular data. We aim to bring the information into a clear and structured format to grasp the key concepts, trends, and current ambiguities. Our approach seeks to ensure both comparability and broad applicability, focusing on achieving general validity in practical use cases.

The main contributions of this review paper can be summarized as follows:

1. Terminology and taxonomy establishment of anonymization methods for tabular data:
   This review introduces a unifying terminology for anonymization methods specific to tabular data. Furthermore, the paper presents a novel taxonomy that categorizes these methods, providing a structured framework that enhances clarity and organization within tabular data anonymization.
2. Comprehensive summary of information loss, utility loss, and privacy metrics in the context of anonymizing tabular data:
   By conducting an extensive exploration, this paper offers a comprehensive overview of methods used to quantitatively assess the impact of anonymization on information and utility in tabular data. By providing an overview of the so-called privacy models, along with precise definitions aligned with the established terminology, the paper reviews and explains the trade-offs between privacy protection and data utility, with special attention to the Curse of Dimensionality. This contribution facilitates a deeper understanding of the complex interplay between anonymization and the quality of tabular data.
3. Integration of anonymization of tabular data with legal considerations and risk assessments:
   Last but not least, this review bridges the gap between technical practices and legal considerations by analyzing how state-of-the-art anonymization methods align with case law and legislation. By elucidating the connection between anonymization techniques and the legal context, the paper provides valuable insights into the regulatory landscape surrounding tabular data anonymization. This integration of technical insights with legal implications is essential for researchers, practitioners, and policymakers alike, contributing to a more holistic approach to data anonymization. The paper conducts a risk assessment for privacy metrics and discusses present issues regarding implementing anonymization procedures for tabular data. Further, it examines possible gaps in the interplay of legislation and research from both technical and legal perspectives. Based on the limited sources of literature and case law, conclusions on the evaluation of the procedures were summarized and were partially drawn using deduction.

In summary, these three main contributions collectively provide interdisciplinary insights for assessing data quality impact and promote a well-informed integration of technical and legal aspects in the domain of tabular data anonymization.

2. Background

This article does not consider the anonymization of graph data or unstructured data, where high dimensionality adds additional constraints [1]. We solely focus on tabular data that can be extracted from relational databases. Due to their reliability and widespread tools, relational databases are used in a wide range of applications across various industries. Thus, anonymizing tabular data in relational databases is a practice-relevant task. In this matter, protecting privacy is the main goal. Further, it facilitates the development of new applications with the possible publishing of Open Data.

We only consider data that have string or atomic data types, e.g., Boolean, integer, character, and float, as attribute data types. From a conceptual point of view, we only distinguish between categorical and numerical attributes, which can be reduced to the data types of string and float in implementations. Characters and integers might be typecast, respectively. We define records as single entries in the database. Individuals might be related to more than one record. This happens when records are created by user events, such as purchase records. Though we relate to relational databases and their taxonomy, to emphasize the anonymization task, instead of using the term "primary key", we use the term Direct Identifier. Instead of talking about a "super key", we say Quasi-Identifier (QI). A QI refers to a set of attributes where the attributes are not identifiers by themselves, but together as a whole might enable the unique identification of records in a database. The QI denotes the characteristics on which linking can be enforced [2]. The QI contains the attributes that are likely to appear in other known datasets, and in the context of privacy models, there is the assumption that a data holder can identify attributes in their private data that may also appear in external information and thus can accurately identify the QI [3]. Further, by considering the same attribute values of a QI, the dataset of records is split into disjunct subsets that form equivalence classes. In the following, we call these equivalence classes groups. If a group consists of $k \in \mathbb{N}$ entries, we call the group a k-group. Besides Direct Identifiers and (potentially more than one) QIs, there are the so-called Sensitive Attributes (SAs), which, importantly, should not be assignable to individuals after applying anonymization. In Section 4, we give the mathematical setting for the data to study. In contrast to pseudonymization, where re-identification is possible but is not within the scope of this article, anonymization does not allow Direct Identifiers at all. For this reason, in anonymization, removing Direct Identifiers is always the first step to take (e.g., in [4]). For the sake of simplicity, we assume that this step is already performed and define the data model on top of it. For the sake of consistency and comparability, throughout the article, we use the Adult dataset from the UCI Machine Learning Repository [5] ("Census income" dataset) for visualizing examples.

3. Related Work

Related work can be categorized into several categories depending on the data format, the perspective (technical or legal), and the use case. The first listed works take a technical perspective and deal with different data types and use cases in anonymization.

The survey [6] by Abdul Majeed et al. gives a comprehensive overview of anonymization techniques used in privacy-preserving data publishing (PPDP) and divides them into the anonymization of graphs and tabular data. Although anonymization techniques for tabular data are presented, the focus of the survey is on graph data in the context of social media. The survey concludes that privacy guidelines must be considered not only at the anonymization level, but in all stages, such as collection, preprocessing, anonymization, sharing, and analysis.

In the literature, most often, the approaches to anonymization are context-sensitive.

Another example is [7], where the authors discuss anonymizing Public Participation Geographic Information System (PPGIS) data by first identifying privacy concerns, referring to the European GDPR as the legal guideline. The authors claim to have reached a satisfactory level of anonymization after applying generalization to non-spatial attributes and perturbations to primary personal spatial data.

Also in [8], by Olatunji et al., anonymization methods for relational and graph data are the focus but with an emphasis on the medical field. Further, in addition to the various anonymization methods, an overview of various attack methods and tools used in the field of anonymization is given. The evaluation is focused on two main objectives, which are performed on the Medical Information Mart for Intensive Care (MIMIC-III) dataset anonymized with the ARX data anonymization tool [9]. In the anonymization procedure, the differences in the accuracy of the predictions between anonymized data and

de-anonymized data are shown. In this use case, generalization has less impact on accuracy than suppression, and it is not necessary to anonymize all attributes but only specific ones.

Again—considering anonymization procedures—in [10], Jakob et al. present a data anonymization pipeline for publishing an anonymized dataset based on COVID-19 records. The goal is to provide anonymized data to the public promptly after publication, while protecting the dataset consisting of 16 attributes against various attacks. The pipeline itself is tailored to one dataset. All Direct Identifiers were removed, and the remaining variables were evaluated using [11] to determine whether they had to be classified as QIs or not.

In [12], the authors examine privacy threats in data analytics and briefly list privacy preservation techniques. Additionally, they propose a new privacy preservation technique using a data lake for unstructured data.

In the literature review in [13], the authors list 13 tools and their anonymization techniques. They identify Open Source anonymization tools for tabular data and give a short summary for each tool. Also, they give an overview of which privacy model is supported by which tool. However, they focus on a literature review and do not give in-depth evaluations of the tools. Last but not least, they derive recommendations for tools to use for anonymizing phenotype datasets with different properties and in different contexts in the area of biology. Besides anonymization methods, some of the literature focuses on the scoring of anonymity and privacy.

In the survey [14], the authors list system user privacy metrics. They list over 80 privacy metrics and categorize into different privacy aspects. Further, they highlight the individuality of single scenarios and present a method for how to choose privacy metrics based on questions that help to choose privacy metrics for a given scenario. Whereas the authors unify and simplify the metric notation when possible, they do not focus on the use case of tabular data and do not describe anonymization methods for tabular data (in a unifying manner). Further, they do not consider the legal perspective.

The following works take a legal perspective but do not fill the gap between legal and technical requirements. The legal understanding is not congruent with technology development, and there are different definitions of identifiable and non-identifiable data in different countries.

In [15], the authors discuss different levels of anonymization of tabular health data in the jurisdictions of the US, EU, and Switzerland. They call for legislation that respects technological advances and provides clearer legal certainty. They propose a move towards fine-grained legal definition and classification of re-identification steps. In the technical analysis, the paper considers only two anonymization methods, removal of Direct Identifiers and perturbation, and gives a schematic overview of classification for levels of data anonymization. The data are classified into identifying data, pseudonymized data, pseudo-anonymized data, aggregated data, (irreversibly) anonymized data, and anonymous data.

In [1], the authors consider the even more opaque regulations regarding anonymizing unstructured data, such as text documents or images. They examine the identifiability test in Recital 26 to understand which conditions must be met for the anonymization of unstructured data. Further, they examine both approaches that will be discussed in Sections 6.3 and 6.4.

From a conceptual perspective, in [16], the authors call for a paradigm shift from anonymization towards transparency, accountability, and intervenability, because full anonymization, in many cases, is non-feasible to implement, and solely relying on anonymization often leads to undesired results.

In summary, it can be seen that there is an increasing demand for practical anonymization solutions due to the rising number of privacy data breaches and the increasing number of data. With the establishment of new processing paradigms, the relevance of user data anonymization will continue to increase. However, current approaches need significant improvement, and there is a need to develop new practical approaches that enable the balancing act between privacy and utility.

4. Technical Perspective

The following model omits the existence of Direct Identifiers and just deals with one QI and several SAs. Furthermore, to make the setting comprehensible, we use the terms table, database, and dataset interchangeably. Let $D = \{R_1, R_2, \ldots, R_n\}$ be a database modeled as a multiset with $n \in \mathbb{N}$ not necessarily distinct records, where $R_i \in A_1 \times A_2 \times \cdots \times A_r \times A_{r+1} \times \ldots \times A_{r+t}$, $i = 1, \ldots, n$, are database entries composed of attribute values; $r \in \mathbb{N}$ is the number of attributes that are part of the QI; $t \in \mathbb{N}_0$ is the number of non QI attributes; A_j, $j = 1, \ldots, r+t$, is the set of possible attribute values of the attribute indexed by j; and the first r attributes represent the QI. In the following, let $|\cdot|$ denote the cardinality of a set, and more specifically, let $|D|$ denote the number of distinct records in database D. As several records can potentially be assigned to one individual, n records correspond to $m \leq n$ individuals with QI attributes $\{U_1, U_2, \ldots, U_m\}$, where $U_i \in A_1 \times A_2 \times \ldots \times A_r$, $i = 1, \ldots, m$. We assume that given data are preprocessed and individuals can only be assigned to one individual, i.e., $|D| = m = n$. Further, let $SA \subseteq \{A_1, \ldots, A_{r+t}\}$ denote the SAs as a subset of all attributes. For the sake of simplicity, in the article, without loss of generality, we restrict the numerical attributes to $A_i \subset \mathbb{R}$ and the categorical attributes to $A_i \subset \mathbb{N}$, $i = 1, \ldots, r+t$. Let R_i, $i \in \{1, \ldots, n\}$ denote the i-th entry and $R_i(j)$, $j \in \{1, \ldots, r+t\}$ denote the value of the j-th attribute of the i-th entry in the database. Figure 1 visualizes the data structure to be studied.

	attr. 1	...	attr. r	attr. r + 1	...	attr. r + t
R_i	$R_i(1)$	...	$R_i(r)$	$R_i(r+1)$	...	$R_i(r+t)$

Figure 1. The considered data model. The first r attributes form a QI. All attributes indexed from 1 to $r+t$ are potentially SAs. The considered data model does not contain Direct Identifiers.

Before scoring certain levels of anonymity for a dataset with personal data, we give an overview of common anonymization methods. We aim to cover relevant methods for tabular data in as detailed a manner as necessary. We are aware that not all methods are described in detail and that research is being carried out on newer approaches. However, in this article, we focus on the most important methods that are state-of-the-art and/or common practice. Some anonymization methods use the information given by the QI. In that case, it is important to note that there might be more than one QI (super key) in a database, and often, several choices of QI have to be considered to score anonymization. For the sake of simplicity and because the following definitions do not limit the use of multiple QIs, where needed, we use a fixed set of attributes as a single QI. In the following, we categorize anonymization methods in seven categories (Sections 4.1–4.7), where not all are necessarily based on QIs. The considered methods are given in the taxonomy in Figure 2. This taxonomy represents a hierarchical structure that classifies anonymization methods into different levels of categories and subcategories, reflecting their relationships.

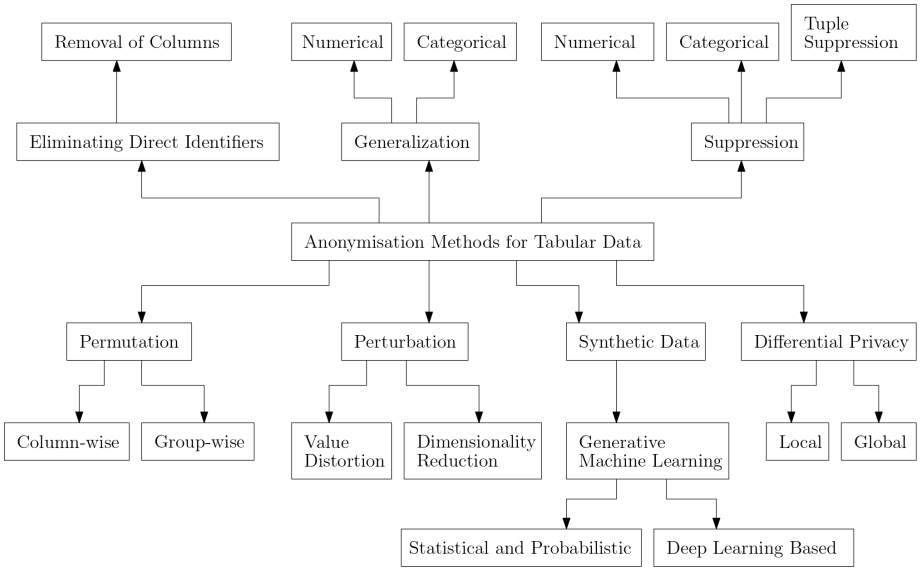


Figure 2. Taxonomy of anonymization methods for tabular data.

4.1. Eliminating Direct Identifiers

Direct Identifiers are attributes that allow for the immediate re-identification of data entries. Therefore, due to the GDPR definition of anonymization, removing the Direct Identifier is compulsory and usually the first step in any anonymization of personal tabular data. Direct Identifiers, often referred to as IDs, do not usually contain valuable information and can simply be removed. A more detailed description can be found in the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule [17] by the United States Department of Health and Human Services, which specifies a Safe Harbor method that requires certain Direct Identifiers of individuals to be removed. The 18 Direct Identifiers that are required to be removed according to the Safe Harbor method can be found in Table 1. To the best of our knowledge, there is no EU counterpart to the HIPAA.

Table 1. Direct Identifiers in the HIPAA Safe Harbor method.

No.	Direct Identifier	No.	Direct Identifier
1	Names	10	Social security numbers
2	All geographic subdivisions smaller than a state	11	IP addresses
3	All elements of dates (except year) directly related to an individual	12	Medical record numbers
4	Telephone numbers	13	Biometric identifiers, including finger and voice prints
5	Vehicle identifiers and serial numbers	14	Health plan beneficiary numbers
6	Fax numbers	15	Full-face photographs and any comparable images
7	Device identifiers and serial numbers	16	Account numbers
8	Email addresses	17	Any other unique identifier
9	URLs	18	Certificate/license numbers

4.2. Generalization

In generalization, the level of detail is coarsened. As a result, given the attributes of individuals, re-identification in the dataset should be impossible. Further, generalization limits the possibility of finding correlations between different attribute columns and datasets. This also makes it difficult to combine and assign records to an individual. There are several types of generalizations, such as subtree generalization, full-domain generalization, unrestricted subtree generalization, cell generalization, and multi-dimensional generalization [6]. generalization for categorical attributes can be defined as follows (c.f. [18]): Let $\overline{A}_j \subseteq \mathcal{P}(A_j)$ be a set of subsets of A_j.

A mapping

$$g : A_1 \times \ldots \times A_r \to \overline{A}_1 \times \ldots \times \overline{A}_r \tag{1}$$

is called a record generalization if and only if for any record's QI $(b_1, \ldots, b_r) \in A_1 \times \ldots \times A_r$ and $(B_1, \ldots, B_r) := g(b_1, \ldots, b_r) \in \overline{A}_1 \times \ldots \times \overline{A}_r$, it holds that $b_j \in B_j$, $j = 1, \ldots, r$.

Let

$$g_i : A_1 \times \ldots \times A_r \to \overline{A}_1 \times \ldots \times \overline{A}_r, \; i = 1, \ldots, n \tag{2}$$

be record generalizations. With $\overline{R}_i := g_i(R_i)$, $i = 1, \ldots n$, we call $g(D) := \{\overline{R_1}, \ldots, \overline{R_n}\}$ a generalization of database D.

The trivial generalization of an attribute is defined as

$$g : A \to \overline{A}, \; b \mapsto \{b\}. \tag{3}$$

Often, generalization is achieved by generalizing attribute values by replacing parts of the value with a special character, for example, "$*$".

Generalization is sometimes also named recoding and can be categorized according to the strategies used [19]. There is a classification in global or local recoding. Global recoding refers to the process of mapping a chosen value to the same generalized value or value set across all records in the dataset. In contrast, local recoding allows the same value to be mapped to different generalized values in each anonymized group. For the sake of simplicity, we use the word generalization instead of recoding. Generalization offers flexibility in data anonymization, but it also requires more careful consideration to ensure that the privacy of individuals is still protected. Further, there is the classification into single- and multi-dimensional generalizations. Here, single-dimensional generalization involves mapping each attribute individually.

$$g : A_1 \times \ldots \times A_r \to \overline{A}_1 \times \ldots \times \overline{A}_r, \tag{4}$$

In contrast, multi-dimensional generalization involves mapping the Cartesian Product of multiple attributes.

$$g : A_1 \times \ldots \times A_r \to \overline{B}_1 \times \ldots \times \overline{B}_s, \; s < r, \tag{5}$$

where B_i, $i = 1, \ldots, s$, is a set in $\{A_1, \ldots, A_r\}$ or is a Cartesian Product of sets $A_{k_1} \times \ldots \times A_{k_l}$, $1 < l \leq r$. When dealing with numerical attributes, generalization can be implemented using discretization, where attribute values are discretized into same-length intervals. The approach is also referred to as value-class membership [20]. Let $L \in \mathbb{N}$ be the interval size. Then, discretization can be defined as

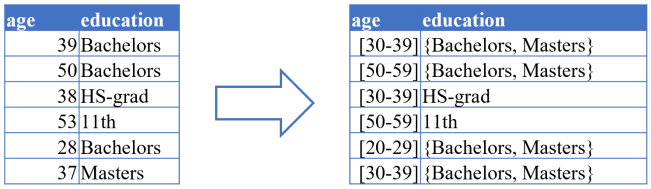
$$g : \mathbb{R} \to \{[a, b) \mid a, b \in \mathbb{R}, \; a < b\}, \; \lambda \mapsto I, \tag{6}$$

where g maps the real number λ to half-open real interval

$$I = [lower, \; upper) := \left[\left\lfloor \frac{\lambda}{L} \right\rfloor L, \; \left\lfloor \frac{\lambda}{L} \right\rfloor L + L \right), \tag{7}$$

where I has length L and $\lfloor \cdot \rfloor$ represents the floor function, which rounds down to the nearest integer. If one wants to discretize to tenths or even smaller decimal places, one can multiply L and the attribute values in the corresponding column with $10, 100, \ldots$ before applying discretization and with the multiplicative inverse of $10, 100, \ldots$ after applying discretization. In practice, due to the often vast possibilities of generalizing tabular data, a generalization strategy has to be found. Note that data consisting of categorical and numerical attributes can incorporate different generalizations for different attributes and different database entries (Equations (2)–(6)).

An example for applying generalization and discretization to the Adult dataset is given in Figure 3.



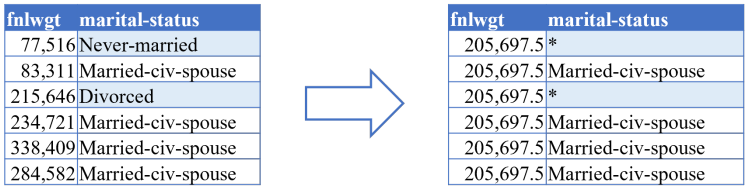
age	education
39	Bachelors
50	Bachelors
38	HS-grad
53	11th
28	Bachelors
37	Masters

age	education
[30-39]	{Bachelors, Masters}
[50-59]	{Bachelors, Masters}
[30-39]	HS-grad
[50-59]	11th
[20-29]	{Bachelors, Masters}
[30-39]	{Bachelors, Masters}

Figure 3. Example. Visualizing both generalization and discretization by projecting the first six records of Adult on the columns `age` and `education`. In the categorical attribute column `education`, the attribute values "Bachelors" and "Masters" are summarized to a set with both values. In the numerical attribute column `age`, the values for `age` are discretized in intervals of size 10.

4.3. Suppression

Suppression (or Data Masking) can be defined as a special type of generalization [18]. To be specific, suppression using generalization resp. total generalization can be achieved by applying $g(b_1, \ldots, b_r) = (\overline{b_1}, \ldots, \overline{b_r})$, $\overline{b_j} \in \{b_j, *\}$ for every database record $(b_1, \ldots, b_r) \in A_1 \times \ldots \times A_r$, where $* := A_j$ or $* := \emptyset$, when suppressing categorical attribute values. To suppress numerical attribute values, we can define $* := f(A_j)$ with $f : \mathbb{R} \to \mathbb{R}$, where f is a statistical function such as mean, sum, variance, standard deviation, median, mode, min, and max. An example of suppression is given in Figure 4.



fnlwgt	marital-status
77,516	Never-married
83,311	Married-civ-spouse
215,646	Divorced
234,721	Married-civ-spouse
338,409	Married-civ-spouse
284,582	Married-civ-spouse

fnlwgt	marital-status
205,697.5	*
205,697.5	Married-civ-spouse
205,697.5	*
205,697.5	Married-civ-spouse
205,697.5	Married-civ-spouse
205,697.5	Married-civ-spouse

Figure 4. Example. Visualizing suppression of the numerical attribute column `fnlwgt` (final weight: number of units in the target population that the responding record represents) by replacing every column value with the mean value of all column values. Visualizing suppression of the categorical attribute column `marital-status` by replacing the values with $*$, which denotes all possible values or the empty set.

Another concept of suppression is tuple suppression, which can be used to deal with outliers. Thereby, given a positive $k \in \mathbb{N}$ for the desired k-anonymity, the database entries in groups with less than k-entries are deleted [21].

4.4. Permutation

With permutation, the order of an individual QI's attribute values within a column is swapped. Mathematically, a permutation is defined as a bijective function that maps a finite set to itself. Let

$$\begin{aligned} \sigma : &\{1, 2, \ldots, n\}^n \to \{1, 2, \ldots, n\}^n, \\ &(i_1, i_2, \ldots, i_n) \mapsto (\sigma(i_1), \sigma(i_2), \ldots, \sigma(i_n)) \end{aligned} \tag{8}$$

be a permutation of record indices.

Considering only column j of the records of a database, we define a column permutation as

$$\pi : A_j^n \to A_j^n, \ (R_i(j))_{i=1,\ldots,n} \mapsto \left(R_{\sigma(i)}(j)\right)_{i=1,\ldots,n}. \tag{9}$$

This reassigns information among columns, potentially breaking important relationships among attributes. This can result in a subsequent deterioration of analyses where the relationships are relevant. An example of column permutation is given in Figure 5.

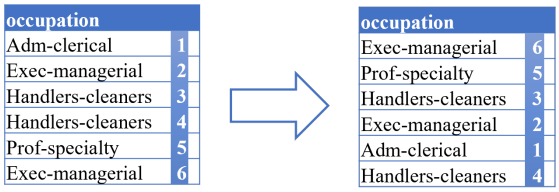


Figure 5. Example. Visualizing permutation of the column `occupation` in the cutout of the first six rows in the Adult dataset. The attached indices point out the change in order by applying permutation. No attribute values are deleted, but the ordering inside the column is very likely destroyed.

4.5. Perturbation

In perturbation, additive or multiplicative noise is applied to the original data. However, without a careful choice of noise, there is the possibility that utility is hampered. On the contrary, especially in the case of outliers, applying noise might not be enough to ensure privacy after anonymization achieved using perturbation. Perturbation is mainly applied to SAs. In [20], the perturbation approaches provide modified values for SAs. The authors consider two methods for modifying SAs without using information about QIs. Besides value-class membership or discretization, which is here explained in generalization (Section 4.2), the authors use value distortion as a method for privacy preservation in data mining. Hereby, for every attribute value $R_i(j)$, $i = 1, \ldots, n$, in an attribute column j, the value $R_i(j)$ is replaced with the value $R_i(j) + \rho$, where $\rho \in \mathbb{R}$ is additive noise drawn from a random variable with continuous uniform distribution $r \sim U(-a, a)$, $a > 0$, or with normal distribution $r \sim \mathcal{N}(\mu, \sigma)$ with mean $\mu = 0$ and standard deviation $\sigma > 0$.

Probability distribution-based methods might also be referred to as perturbation. However, because these methods replace the original data as a whole, we list these approaches in Synthetic Data (Section 4.7). The same applies to dimensionality reduction-based anonymization methods, which we also list in Synthetic Data.

Section 4.6 studies a more sophisticated field of perturbations, namely, Differential Privacy (DP), which is the state of the art in privacy-preserving ML.

4.6. Differential Privacy

Differential Privacy, introduced by Cynthia Dwork in [22], is a mathematical technique that allows for the meaningful analysis of data while preserving the privacy of individuals in a dataset. The idea is to add random noise to data in such a way that—as it is the goal in anonymization—no inferences can be made about personal and sensitive data. DP is implemented in different variants depending on the use case, where anonymization is only a sub-task in a vast variety of use cases. Generally, there is a division into local [23] and global DP [24]. The local DP model does not require any assumptions about the server, whereas the global DP model is a central privacy model that assumes the existence of a trusted server. As a result, the processing frameworks for global and local DP differ significantly. However, the definition of local DP can be embedded into the definition of global DP as a special case where the number of database records equals one.

Common techniques to implement local DP are the Laplace and Exponential mechanisms [24], and Randomized Response [25].

In the context of global DP, there are novel output-specific variants of DP for ML training processes, where ML models are applied to sensitive data and model weights are manipulated in order to preclude successful membership or attribute inference attacks. For example, in Differentially Private Stochastic Gradient Descent (DP-SGD) [26], instead of adding noise to the data themselves, gradients (i.e., the multi-variable derivative of the loss function with respect to the weight parameters) are manipulated to obtain privacy-preserving Neural Network models. Adapting the training process is also referred to as private training. Whereas private training only adjusts the training process and leads to private predictions, private prediction itself is a DP technique to prevent privacy violations by limiting the amount of information about the training data that can be obtained from

a series of model predictions. Whereas private training operates on model parameters, private prediction perturbs model outputs [27]. The privacy models' k-anonymity, l-diversity, and t-closeness rely on deterministic mechanisms and can be calculated given the database and a QI. On the contrary, global DP does not depend only on the QI but also on the whole database and a randomized mechanism M in connection with a data-driven algorithm, such as database queries, statistical analysis, or ML algorithms. The most basic definition of the so-called (ϵ, δ)-DP includes the definition of a randomized algorithm, probability simplex, and the distance between two databases based on the ℓ_1-norm of the difference of histograms. This definition of DP requires that for every pair of "neighbouring" databases X, Y (given as histograms), it is extremely unlikely that, ex post facto, the observed value $M(X)$ resp. $M(Y)$ is much more or much less likely to be generated when the input database is X than when the input database is Y [24]. Two databases are called neighbors if the resulting histograms $x, y \in \{0,1\}^{|\mathcal{X}|}$ only differ in at most one record, where in our setting, $x_i, y_i \in \{0,1\}$, $i = 1, \ldots, |\mathcal{X}|$ is the number of non-duplicate records with the same type in X resp. Y, where $\mathcal{X} \supseteq D$ is the record "universe". More in detail, the (ϵ, δ)-DP for a randomized algorithm M with domain $\{0,1\}^{|\mathcal{X}|}$ is defined by the inequality below, where $\epsilon > 0$, $\delta \geq 0$ are privacy constraints.

For all $S \subseteq range(M)$ (subset of the possible outputs of M) and $x, y \in \{0,1\}^{|\mathcal{X}|}$, such that $\|x - y\|_1 \leq 1$, we have

$$Pr[M(X) \in S] \leq e^{\epsilon} Pr[M(Y) \in S] + \delta. \tag{10}$$

The smaller the value of the so-called privacy budget ϵ, the stronger the privacy guarantee. Additionally, parameter δ is a small constant term that is usually set to a very small value to ensure that the formula holds with high probability. In summary, DP guarantees that the output of a randomized algorithm does not reveal much about any individual in the dataset, even if an adversary has access to all other records in the database. There are promising approaches, such as in [28], where the authors propose a general and scalable approach for differentially private synthetic data generation that also works for tabular data.

4.7. Synthetic Data

Whereas the above approaches directly manipulate dataset entries, with synthetic approaches, new data are generated based on extracted and representative information from the original data. For the sake of simplicity, the following synthetic approaches to generate data are only described for numerical data. However, by using a reasonable coding method (such as one-hot encoding), categorical data might be converted into numerical data, and vice versa.

In [29], to improve anonymization using generalization for k-anonymity, the so-called condensation method was introduced. The approach is related to probability distribution-based perturbation methods. Thereby, the resulting numerical attribute values closely match the statistical characteristics of the original attribute values, including inter-attribute correlations (second order) and mean values (first order). Condensation does not require hierarchical domain generalizations and fits both static data (static condensation) and dynamic data streams (dynamic condensation). In summary, this approach condenses records into groups of predefined size, where each group maintains a certain level of statistical information (mean, covariance). The authors test the accuracy of a simple K-Nearest Neighbor classifier on different labeled datasets and show that condensation allows for high levels of privacy without noticeably compromising classification accuracy. Further, the authors find that by using static condensation for anonymization, in many cases, even better classification accuracy can be achieved. This is because the implied removal of anomalies cancels out the negative impact of adding noise. In summary, condensation produces synthetic data by creating a new perturbed dataset with similar dataset characteristics. The mentioned paper states the corresponding algorithm to calculate statically condensed group statistics: first-order and second-order sum per attribute and

total number of records. Afterwards, given the calculated group statistics, by building the covariance matrix of attributes for every group, the eigenvectors and eigenvalues of the covariance matrix can be calculated using eigendecomposition. To construct new data, the authors assume that the data within each group is independently and uniformly distributed along each eigenvector with a variance equal to the corresponding eigenvalue.

Another approach to improving privacy preservation when creating synthetic data is to bind Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA) with DP [30]. In the paper, considering the PCA-based approach, a perturbed covariance matrix (real and symmetric) is decomposed into eigenvalues and eigenvectors, and Laplace noise is applied on the resulting eigenvectors to generate noisy data. The introduced differential PCA-based privacy-preserving data publishing mechanism satisfies ϵ-Differential Privacy and yields better utility in comparison to the Laplace and Exponential mechanisms, even when having the same privacy budget.

In [31], the authors propose a sparsified Singular Value Decomposition (SVD) for data distortion to protect privacy. Given the dataset—often a sparse—matrix $D \in \mathbb{R}^{n \times m}$, the SVD of D is $D =: U\Sigma V^T$, where U is an $n \times n$ orthonormal matrix; $\Sigma := diag[\sigma_1, \sigma_2, \ldots, \sigma_s]$, $\sigma_i \geq 0$, $1 \leq i \leq s$ with $s := \min\{m, n\}$ is an $n \times m$ diagonal matrix whose non-negative diagonal entries are in descending order; and V^T is an $m \times m$ orthonormal matrix. Due to the property of descending variation in $\sigma_1, \ldots, \sigma_s$, data can be compressed to lower dimensionality while preserving utility. This is achieved by using only the first $1 \leq d \leq s$ columns of U, the $d \times d$ upper left submatrix of Σ, and the first d rows from V^T: U_d, Σ_d, V_d^T. The matrix $D^d := U_d \Sigma_d V_d^T$ to represent D can be interpreted as a reduced dataset of D that can be used for mining on the original dataset, D. In contrast to SVD, in sparsified SVD, entries in U_d and V_d^T that are below a threshold are set to zero to obtain a sparsified data matrix $\overline{D}^d$. By thresholding values in U_d and V_d^T to zero and by dropping less important features in D, data are distorted, which makes it harder to estimate values and records in D. However, the most important features are kept. Therefore, the approach aims to maintain the utility of the original dataset, D.

Overall, from a technical perspective, when considering eigenvector-based approaches to generate synthetic data, a numerically stable algorithm including suitable matrix preprocessing for the eigenvalue problem at hand has to be selected. Last but not least, eigenvector-based approaches can also help mitigate the Curse of Dimensionality in data anonymization [32]. The Curse of Dimensionality and its relation to anonymization methods are explained in more detail in Section 5.5.

More recent generative ML models that are often based on deep learning can effectively create synthetic and anonymous data. Generative models aim to approximate a real-world joint probability distribution, such that the original dataset only represents samples pulled from the learned distribution. One common use case of generative models is to fix class imbalances or to apply domain transfer. However, generative approaches can also be used to generate anonymous data. Importantly, considering privacy preservation, the generated data should not allow for (membership/attribute) inferences about specific training data. When it comes to tabular data, in [33], the authors create synthetic tabular data by adapting a Generative Adversarial Network (GAN) that incorporates a Long Short-Term Memory (LSTM) Neural Network in the generator and a Fully Connected Neural Network in the discriminator. Other examples for synthetic tabular data based on GANs can be found in the papers [34,35]. However, just considering a generative ML model by itself does not imply the privacy preservation of training data. Therefore, generative ML might be combined with DP as a potential way out [36]. This again also applies to tabular data; c.f. [37].

5. Utility vs. Privacy

In anonymization, there is always the trade-off of removing information vs. keeping utility. In the literature, two main concepts are used to model the change in utility when applying anonymizing: information loss (Section 5.1) and utility loss (Section 5.2).

To give an overview, we categorize and list the studied anonymization scores in Section 5 in Table 2.

Table 2. Overview of information losses, utility losses/measurements, and privacy models when applying anonymization methods to tabular data

Measurement	Method
Information loss	Conditional entropy [18] Monotone entropy [18] Non-uniform entropy [18] Information loss on a per-attribute basis [38] Relative condensation loss [39] Euclidean distance [40]
Utility loss	Average group size [41] Normalized average equivalence class size metric [42] Discernibility metric [21,42,43] Proportion of suppressed records ML utility Earth Mover Distance [44] z-Test statistics [7]
Privacy models	k-Anonymity [3] Mondrian multi-dimensional k-anonymity [42] l-Diversity [45] t-Closeness [46] Privacy probability of non-re-identification [47]

In the following subsections, we explain the measurements and methods in greater detail. Further, we give insights into the occurring phenomena of the so-called Curse of Dimensionality in the context of anonymizing tabular data.

5.1. Information Loss

Conditional entropy assesses the amount of information that is lost with anonymization in terms of generalization and suppression of categorical attributes. In [18], the authors study the problem of achieving k-anonymity using generalization and suppression with minimal loss of information. As a solution to the problem, they prove that the stated problem is NP-hard and present an algorithm with an approximation guarantee of $O(\ln\ k)$-anonymity. The calculation of information loss based on entropy builds on probability distributions for each of the attributes. Let X_j denote the categorical value of attribute A_j, $j = 1, \ldots r$, in a randomly selected record from a dataset D consisting of only categorical data. Then, for $a \in A_j$, $j \in \{1, 2, \ldots, r\}$,

$$Pr[X_j = a] := \frac{|\{1 \leq i \leq n : R_i(j) = a\}|}{n}. \tag{11}$$

Let $B_j \subseteq A_j$. Then, the conditional entropy of X_j given B_j is defined as follows:

$$H(X_j|B_j) := \\ -\sum_{b_j \in B_j} Pr[X_j = b_j | X_j \in B_j] \log_2\big(Pr[X_j = b_j | X_j \in B_j]\big). \tag{12}$$

Loosely speaking, conditional entropy measures the average amount of uncertainty in X_j given the knowledge that X_j takes values from B_j.

Given $g(D) = \{\overline{R_1}, \overline{R_2}, \ldots, \overline{R_n}\}$, a generalization of D, the entropy measure of the loss of information caused by generalizing D into $g(D)$ is defined as

$$\Pi_e(D, g(D)) := \sum_{i=1}^{n} \sum_{j=1}^{r} H(X_j | \overline{R}_i(j)). \tag{13}$$

If $\overline{R}_i$, $i \in \{1, \ldots, n\}$, is no generalization at all, i.e., $|\overline{R}_i(j)| = 1$, we have $H(X_j|\overline{R}_i(j)) = 0$, and there is no uncertainty. On the other hand, if $\overline{R}_i(j) = A_j$, there is maximal uncertainty. An example of entropy information loss is given in Figure 6.

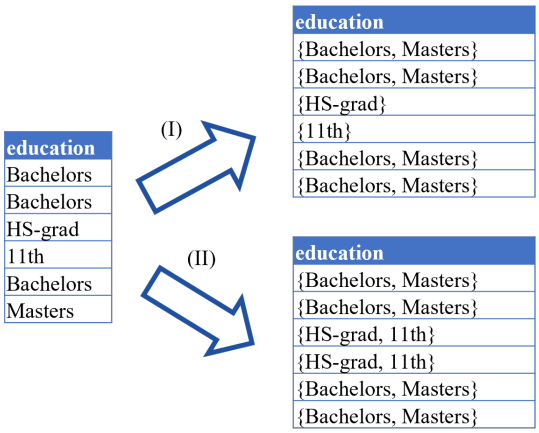


Figure 6. Example. Entropy information loss when generalizing the column `education` of the cutout of the first six rows in the Adult dataset. In generalization (I), we obtain $\Pi_e(D, g(D)) \approx 3.25$, which means lower information loss than in generalization (II), where $\Pi_e(D, g(D)) \approx 5.25$.

In [18], the authors also use other variants of entropy measures, namely, the so-called monotone entropy measure and non-uniform entropy measure, with different characteristics. However, the authors claim that the entropy measure is a more appropriate measure when it comes to privacy.

Given a dataset $D = \{R_i \mid i = 1, \ldots, n\}$ consisting of only numerical attributes and a discretization $g(D) = \{\overline{R}_i(j) \mid i = 1, \ldots, n,\ j = 1, \ldots, r\}$ of D, the information loss on a per-attribute basis can be calculated with the following formula [38]:

$$\Pi(D, g(D)) := \frac{1}{n \cdot r} \sum_{i=1}^{n} \sum_{j=1}^{r} \frac{upper_{ij} - lower_{ij}}{\max_j - \min_j}, \tag{14}$$

where $upper_{ij}$ and $lower_{ij}$ are the upper and lower bounds of generalized attribute value interval $\overline{R}_i(j)$, and $\min_j := \min_{i=1,\ldots,n}\{R_{ij}\}$ and $\max_j := \max_{i=1,\ldots,n}\{R_{ij}\}$, i.e., the minimum and maximum attribute values before generalization.

Based on condensation (Section 4.7) for k-anonymity, in [39], the so-called relative condensation loss is defined to score information loss in anonymization. Given anonymized tabular data $\overline{D}$, the relative condensation loss is group-wise-defined and represents a minimum level of information loss. For $g \in groups$, where $groups$ are the groups of anonymized data $\overline{D}$,

$$\mathcal{L}(g) := \frac{\max_{\overline{R}_i, \overline{R}_k \in g,\ i \neq k} \|\overline{R}_i - \overline{R}_k\|_2}{\max_{\overline{R}_i, \overline{R}_k \in \overline{D},\ i \neq k} \|\overline{R}_i - \overline{R}_k\|_2} \in (0, 1], \tag{15}$$

where $||\cdot||_2$ denotes the 2-norm and anonymized entries $\overline{R}_i$, $i = 1, \ldots, n \in \mathbb{R}^d$, are quantified as real vectors of dimension $d \in \mathbb{N}$, $d \geq r$. Different values of $\mathcal{L}(g)$ for the different $g \in groups$ can be aggregated (avg, max, ...) to a total information loss $\mathcal{L}(\overline{D})$.

Last but not least, in [40], the authors use the average Euclidean distance to measure information loss:

$$IL(D, g(D)) := \frac{1}{n} \sum_{i=1}^{n} dist\left(R_i, \overline{R}_i\right), \tag{16}$$

where *dist* defines the Euclidean distance between data records. Note that in the case of non-real-valued attributes in the dataset, the records have to be vectorized before applying *dist*. An example of numerical information loss is given in Figure 7.

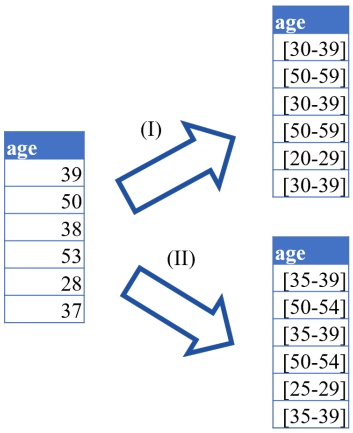


Figure 7. Example. Numerical information loss when generalizing the column age of the cutout of the first six rows in the Adult dataset. In generalization (I), we obtain $\Pi(D, g(D)) \approx 0.36$ and $IL(D, g(D)) \approx 3.33$, which means higher information loss than in generalization (II), where $\Pi(D, g(D)) = 0.16$ and $IL(D, g(D)) \approx 1.17$. In this example, to apply ID, intervals are vectorized by calculating the mean of the minimum and maximum values.

If there is a mixture of categorical and numerical attributes in D, the summands of the combined sum have to be weighted accordingly. Relative condensation loss can be used for both categorical and numerical data by defining feature embeddings for categorical data.

5.2. Utility Loss

As mentioned above, the entropy measure can only be used for processing categorical attributes. However, they lack the capability to deal with numerical data. By designing a utility loss that can deal with both categorical and numerical attribute values, we can overcome this downside. In [44], the authors quantify utility by calculating the distance between the relative frequency distributions of each data attribute in the original data and the sanitized data. The distance is based on the Earth Mover Distance (EMD). Further, z-test statistics can be utilized to examine whether significant differences exist between variables in the original and the anonymized data [7]. Another method to score the utility of anonymization that can be used for evaluations is the average size of groups [41],

$$group_{AVG}(D) := \frac{|D|}{|groups|}, \tag{17}$$

or the normalized average equivalence class size metric [42], defined by the formula

$$C_{AVG}(D) := \frac{|D|}{|groups| \cdot k}, \tag{18}$$

or the so-called, commonly used discernibility metric, which scores the number of database entries that are indistinguishable from each other [21,42,43] and penalizes large group sizes,

$$C_{DM}(D) := \sum_{group \in groups} |group|^2. \tag{19}$$

The listed group-size-based metrics $group_{AVG}$, C_{AVG}, and C_{DM} should be minimized to maintain utility while aiming for k-anonymity with k greater than or equal to a predefined positive integer.

Taking into account record suppression (Section 4.3), the proportion of suppressed records in the total number of records before anonymization can also be used to measure the loss of utility. However, applying record suppression to obtain k-anonymity extends group sizes and thus group-size-based metrics.

In contrast to the above approaches, when the context is known in advance, there is the possibility to measure the data utility by scoring the output of ML algorithms that use anonymized data for training. For example, in [38], anonymized labeled data are scored by calculating the F-measure after applying K-Nearest Neighbor to classify molecules that are given as numerical attributes. Considering the Adult dataset, in [48], the authors apply different ML algorithms (K-Nearest Neighbor, Random Forest, Adaptive Boosting, Gradient Tree Boosting) to anonymized data. However, they just apply record suppression for anonymization. In the following, we call this type of score ML utility.

5.3. Privacy Models

There are common models to determine if records in a dataset can be re-identified. Yet, the models have weaknesses that can potentially be exploited by attackers. In the following, we solely focus on the definitions and give examples. In Section 6.7, we list the models' weaknesses and embed the definitions in a legal context.

5.3.1. k-Anonymity

The so-called k-anonymity, first introduced in [3], $k \in \mathbb{N}^+$, $k \leq n$, is a dataset property for anonymization that considers a QI. If the attributes of the QI for each record in the dataset are identical to at least $k-1$ other records in the dataset, the dataset is called k-anonymous. When having k-anonymity, groups consist of at least k-records. Technically, k-anonymity is defined by

$$k := \min_{group \in groups} |group|.$$

To give an example, Figure 8 shows a database R, where the four attributes `education`, `education-num`, `capital-loss`, `native-country` build a QI and the attribute `age` is an SA. In Figure 8, generalization and discretization are applied, affecting the attributes `education`, `education-num`, `native-country` in such a way that at least two records in the table always have the same QI, leading to k-anonymity with $k = 2$. To be precise, the data are split into two groups: $\{R_1, R_2, R_5, R_6\}$ and $\{R_3, R_4\}$.

The privacy metric of k-anonymity might be combined with different metrics. For example, the authors in [42] introduce the so-called Mondrian multi-dimensional k-anonymity as a multi-dimensional generalization model for k-anonymity. The paper proposes a greedy metric approximation algorithm that offers flexibility and incorporates general-purpose metrics such as the discernibility metric or the normalized average equivalence class size metric (Section 5.2).

age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	income
39	State-gov	77,516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2,174	0	40	United-States	<=50K
50	Self-emp-not-inc	83,311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
38	Private	215,646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
53	Private	234,721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
28	Private	338,409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
37	Private	284,582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K

age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	income
39	State-gov	77,516	{'Bachelors', 'Masters'}	(10.0, 14.0)	Never-married	Adm-clerical	Not-in-family	White	Male	2,174	0	40	{United-States, Cuba}	<=50K
50	Self-emp-not-inc	83,311	{'Bachelors', 'Masters'}	(10.0, 14.0)	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	{United-States, Cuba}	<=50K
38	Private	215,646	{'HS-grad', '11th'}	(5.0, 9.0)	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	{United-States, Cuba}	<=50K
53	Private	234,721	{'HS-grad', '11th'}	(5.0, 9.0)	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	{United-States, Cuba}	<=50K
28	Private	338,409	{'Bachelors', 'Masters'}	(10.0, 14.0)	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	{United-States, Cuba}	<=50K
37	Private	284,582	{'Bachelors', 'Masters'}	(10.0, 14.0)	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	{United-States, Cuba}	<=50K

Figure 8. Example. The first six rows of the Adult dataset, where the blue-background attributes `education, education-num, capital-loss, native-country` define a QI (just artificially chosen as the QI for demonstration purposes!). Column sorting can be applied to fit the data scheme (Figure 1). The transformed six-row database fulfills k-anonymity with $k = 2$, whereas before discretization in the column `education-num` and generalizations in the columns `education` and `native-country`, the groups had a minimum group size of one. The background colors (orange and yellow) visualize group correspondence, where the attributes in the chosen QI are identical for every record in the group.

5.3.2. l-Diversity

l-Diversity, introduced in [45], a second common model for anonymization, considers SAs and gives additional privacy protection to k-anonymity. Again, it considers groups of records with the same QI. When having distinct l-diversity, $l \in \mathbb{N}^+$, $l \leq n$, each group has at least l different attribute values for every SA. Therefore, it is not possible to assign a single attribute value to all records of a group, and group membership does not imply assigning a unique SA to a person. Utilizing l-diversity for scoring anonymity can be challenging, as it depends on the variety of values an SA can have. Technically, l-diversity is defined as

$$l := \min_{group \in groups} |\{R(j) \mid R \in group\}|, \tag{20}$$

where $j \in \{1, \ldots, r+t\}$ denotes the column index of the SA. Given the example at the bottom of Figure 8 and the SA `age` in every group, all values of `age` are diverse, and each group consists of two records. Therefore, we have l-diversity with $l = 2$. For the SA `workclass`, there would be l-diversity with $l = 1$.

5.3.3. t-Closeness

t-Closeness [46] again takes into account SA values. Whereas l-diversity considers the variety of SA values in single groups, t-closeness checks the granularity of SA values in a single group in comparison to the overall value distribution in the dataset. A group is said to have t-closeness if the EMD between the relative frequency distribution of an SA in this group and the relative frequency distribution of the attribute in the whole dataset is no more than a threshold $t > 0$. A dataset is said to have t-closeness if all equivalence classes have t-closeness. Originally, the authors considered the EMD for this purpose (for comparison, see Section 5.2). The distance is calculated differently for integer, numerical, and categorical attributes. Given a dataset D with an SA at index $s \in \{1, \ldots, r+t\}$, the t-closeness of the dataset is defined as

$$t(D) := \max_{group \in groups} EMD(P, Q_{group}), \tag{21}$$

where the following apply:

- D is the dataset;
- P is the relative frequency distribution of all attribute values in the column of the SA in dataset D;

- Q_{group} is the relative frequency distribution of all attribute values in the column of the SA within $group$ that is an equivalence class of dataset D and is obtained by a given QI;
- $EMD(P, Q)$ is the EMD between two relative frequency distributions and depends on the attributes' value type.

Given two ordered relative frequency distributions P and Q of integer values, the ordered EMD is defined as follows:

$$EMD(P, Q) := \frac{1}{o-1} \sum_{i=0}^{o-1} \left| \sum_{j=0}^{i} (P - Q)_j \right|, \tag{22}$$

where the following apply:

- o is the number of distinct integer attribute values in the SA column;
- P and Q are two relative frequency distributions as histograms (integers are ordered in ascending order).

Given two ordered relative frequency distributions P and Q of categorical values, the equal EMD is defined as follows:

$$EMD(P, Q) := \frac{1}{2} \sum_{i=0}^{o-1} |(P - Q)_i|, \tag{23}$$

where the following apply:

- o is the number of distinct categorical attribute values in the SA column;
- P and Q are two relative frequency distributions as histograms (integers are ordered in ascending order).

Given the example at the bottom of Figure 8 and the sensitive integer attribute `age`, there would be t-closeness with $t = 0.2$, due to

$$EMD(P_1, Q) = 0.1$$

and

$$EMD(P_2, Q) = 0.2,$$

where P_1 is the orange group with four records and P_2 is the yellow group with two records.

5.4. Re-Identification Risk Quantification

Besides information loss, utility scoring, and privacy models, there is a fourth important method to score anonymization, namely, quantifying the probability of re-identification risk. Privacy models can only be calculated given the anonymized tabular dataset, and information loss and utility scores evaluate the application of anonymization regarding utility preservation. Re-identification risk can be calculated given an anonymized dataset plus an individual's attribute value(s) as background knowledge. The re-identification risk method particularly takes into account the very realistic danger of the so-called inference attacks. For example, in [47], the authors define a score that incorporates the uniqueness, uniformity, and correlation of attribute values. They quantify the re-identification risk by calculating a joint probability of the non-uniqueness and non-uniformity of records. From a technical perspective, the re-identification risk is modeled as a Markov process. We adapt the definition of the probability (PR) of re-identifying a record R to our setting assuming a unit record dataset D, i.e., not having event data. Further, we restrict the definition to attributes that are part of the QI, i.e., to the first r attributes in the dataset. We define the probability (PR) of re-identifying a record R given its attribute values at indices $J \subseteq \{1, \ldots, r\}$ as follows:

$$PR(R(J)) := 1.0 - PP(R(J)) \cdot n), \tag{24}$$

where n is the total number of records in the dataset and $PP(R(J))$ is the privacy probability of non-re-identifying record R in dataset D with a subset of attribute values of record R at attribute indices J, i.e., $R(J)$. PP is calculated by utilizing the Markov Model risk score. Without loss of generality, we re-index the ordered set of attribute values $\{R(1), \ldots, R(r)\}$, define the ordered set $\{R(2), \ldots, R(m)\} := \{R(1), \ldots, R(r)\} \setminus R(J)$, and let $R(1) := R(J)$. Then, the privacy probability of non-re-identifying record R in dataset D with a subset of attribute values of record R at attribute indices J is defined as

$$PP(R(J)) := P(R(J)) \cdot (1 - P(R|R(J))) \cdot \prod_{1 \leq j \leq m-1} P(R(j+1)|R(j))(1 - P(R|R(j+1))), \tag{25}$$

where the following apply:

- $P(R(J)) := Pr[X_j = R(j), j \in J]$;
- $P(R|R(J)) := Pr[X_i = R(i), i \notin J | X_j = R(j), j \in J]$;
- $P(R(j+1)|R(j)) := Pr[X_{j+1} = R(j+1) | X_j = R(j)]$;
- $P(R|R(j+1)) := Pr[X_i = R(i), i \notin J | X_{j+1} = R(j+1)]$.

Calculating the average PR for all records in the dataset yields

$$PR(D, J) := \frac{1}{n} \sum_{i=1}^{n} PR(R_i(J)). \tag{26}$$

Considering the dataset given in Figure 9 as an example, given the attribute value "Bachelors" for education in dataset record R_1, the privacy probability of re-identifying the record is $PR(R_1(\{1\})) = 0.9$. The calculation of the start probability, i.e., attribute uniqueness, $P(R_1(\{1\})) \approx 0.386$, is equivalent to the re-identification-risk score, RIR, which is efficiently calculated with CSIRO's R4 tool [49]. Given the attribute value "HS-grad" for education in dataset record R_3, the privacy probability of re-identifying this record is the highest, as $PR(R_3(\{1\})) = 1.0$, and the RIR score is $P(R_3(\{1\}) = 1.0$. Whereas the RIR score does not depend on the order of attributes, PR depends on the attribute indices and also takes into account inter-attribute relations. Besides the average privacy probability of re-identifying records, the paper [47] describes the minimum, maximum, median, and marketer re-identification risk based on the calculated PR values of all dataset records to score the re-identification risk of a dataset.

	education	sex	hours-per-week
R_1	Bachelors	M	40
R_2	Bachelors	M	13
R_3	HS-grad	M	40
R_4	11th	M	40
R_5	Bachelors	F	40
R_6	Masters	F	40

Figure 9. Example. Projecting the first six rows of the Adult set on the attributes `education`, `sex`, `hours-per-week`. The PR score assumes that attribute values are known and subsequently calculates the risk of re-identifying a single record (in the case of unit record data). Having knowledge about different values of the attribute `education` (yellow resp. orange) leads to different privacy probabilities of re-identifying a record (record R_1 resp. R_3).

5.5. Curse of Dimensionality

The phenomena of the Curse of Dimensionality, first mentioned in [50] in the context of linear equations, refer to the increase in computational complexity and requirements for data analysis as the number of variables (dimensions/attributes) grows. This increase makes it more and more difficult to find optimal solutions for high-dimensional problems. Considering anonymization, most privacy models on multivariate tabular data lead to poor utility if enforced on datasets with many attributes [32]. Aggarwal has already shown

in [39] that large-sized QIs lead to difficult anonymization, having previously presented condensation [29] (described in Section 4.7) as a synthetic approach to anonymization to achieve k-anonymity. Besides showing the openness inference attacks in terms of probability when having high-dimensional data, in an experimental analysis, it is visualized that anonymizing high-dimensional data, even for only 2-anonymity, leads to unacceptable information loss. However, high-dimensional data potentially have inter-attribute correlations that—despite the theoretic Curse of Dimensionality—can be used to better anonymize them in terms of utility preservation. Therefore, to overcome the Curse of Dimensionality in anonymization, in the so-called Vertical Fragmentation, the data are first partitioned into disjoint sets of correlating attributes and subsequently anonymized and assembled after the anonymization step [38]. This approach is method-agnostic, as it can be used with all anonymization methods described in Section 4. Given the attributes $A_1, \ldots, A_{r+t}$, a vertical fragmentation $\mathcal{F}$ of the attributes is a partitioning of the attributes in fragments $\mathcal{F} = \{F_1, \ldots, F_f\}$ s.t. $\forall i \in \{1, \ldots, f\} : F_i \subseteq \{A_1, \ldots, A_{r+t}\}$, $F_i \cap F_j = \emptyset$, $i \neq j$ and $\bigcup_{i=1,\ldots,f} F_i = \{A_1, \ldots, A_f\}$, where $i, j \in \{1, \ldots, r+t\}$. Considering the single fragments, groups can be formed, and k-anonymity, calculated. However, there are a vast number of possibilities for vertical fragmentation depending on the number of attributes. Therefore, systematic vertical fragmentation that takes into account inter-attribute correlations and post-utility after anonymization has to be chosen. The approach in [38] focuses on classification problems and attempts to maximize the amount of non-redundant information contained in single fragments while also striving for high utility of fragments to conduct the classification task. The authors propose the so-called Fragmentation Minimum Redundancy Maximum Relevance (FMRMR) metric to head into beneficial fragmentation. In the following, let F^j, $j = 1, \ldots, |F|$, denote indexed attributes of fragment F and A^C be the class attribute in the database. The "supervised" FMRMR metric is calculated with the formula

$$FMRMR(\mathcal{F}) := \sum_{F \in \mathcal{F}} (V_F - W_F), \tag{27}$$

where

$$V_F := \frac{1}{|F|} \sum_{j=1}^{|F|} I\left(A^c, F^j\right) \tag{28}$$

is the total mutual information between the attributes and class attribute A^C in fragment F of fragmentation $\mathcal{F}$ and

$$W_F := \frac{1}{|F|^2} \sum_{k=1}^{|F|} \sum_{j=1}^{|F|} I\left(F^k, F^j\right), \tag{29}$$

is the total pairwise mutual information between the attributes in fragment F of fragmentation $\mathcal{F}$. The formula [51]

$$I(A_k, A_j) := \sum_{a_k \in R(k)} \sum_{a_j \in R(j)} Pr\left[X_k = a_k,\ X_j = a_j\right] \log_2\left(\frac{Pr\left[X_k = a_k,\ X_j = a_j\right]}{Pr\left[X_k = a_k\right] Pr\left[X_j = a_j\right]}\right) \tag{30}$$

defines the mutual information between attributes A_k and A_j, where X_k, X_j are discrete random variables as defined in Section 5.1 and the joint probability distribution is defined as

$$Pr[X_k = a, X_j = b] := \frac{|\{1 \leq i \leq n : R_i(k) = a,\ R_i(j) = b\}|}{n}, \tag{31}$$

where $a \in R(k)$, $b \in R(j)$ are values of the corresponding column. Note that if X_k and X_j are independent random variables, we have $I(A_k, A_j) = 0$, and the columns are non-redundant.

With Equation (27), the fragment utility for the classification task at hand is maximized (Equation (28)) while minimizing the mutual information and redundancy of attributes inside the fragment (Equation (29)). Above, we described the procedure in the context of a supervised application. However, vertical fragmentation can also be used in the context of an unsupervised application by adding one or more common attributes to the single fragments to enforce correspondence between fragments. Therefore, when having an unsupervised task at hand, an "unsupervised" FMRMR metric might be defined by adapting Equation (27):

$$uFMRMR(\mathcal{F}_{ext}) := -\sum_{Fe \in \mathcal{F}_{ext}} W_{Fe}, \tag{32}$$

where $\mathcal{F}_{ext} := \{Fe_1, \ldots, Fe_f\}$ is obtained from fragmentation $\mathcal{F} = \{F_1, \ldots, F_f\}$ by adding one or more common attribute(s) $A \subset \{A_1, \ldots, A_{r+t}\}$ to each fragment: $\forall i = 1, \ldots, f : Fe_i := F_i \cup A$.

To sum up, the vertical fragmentation approach aims to alleviate the negative effects of the Curse of Dimensionality. By choosing suitable discrete or continuous probability distributions depending on the given data, after possibly necessary preprocessing like discretizing values, the approach can be used in principle for both categorical and numerical data. Figure 10 visualizes the mutual information of all attribute pairs of the Adult dataset in a symmetric matrix.

The Curse of Dimensionality also occurs in DP. For example, in [52], the authors state that Randomized Response suffers from the Curse of Dimensionality. There is a trade-off between applying Randomized Response to single attributes and applying Randomized Response to a set of attributes simultaneously. Depending on the number of records, the latter might lead to poor utility of the estimated distribution of the original data, and applying Randomized Response to single attributes implies a poor estimated joint distribution of the original data. The authors propose an algorithm to cluster attributes with high mutual dependencies and apply Randomized Response to single clusters jointly. Their measure of dependency between two attributes A_k, A_j is based on the absolute value of the Pearson Correlation and Cramér's V Statistic $V(A_k, A_j)$. In Randomized Response, $|Corr(A_k, A_j)|$ can be calculated given discretized numerical attributes A_k, A_j, and Cramér's V Statistic $V(A_k, A_j)$ can be calculated given categorical attributes A_k, A_j that have no ordering. In their experimental results, they empirically evaluate the phenomenon on the multivariate Adult dataset.

The Pearson Correlation of attributes A_j, A_k is defined as

$$|Corr(A_k, A_j)| := \left| \frac{\sum_{i=1}^{n} \left(A_j^{(i)} - \bar{A}_j\right)\left(A_k^{(i)} - \bar{A}_k\right)}{\sqrt{\sum_{i=1}^{n} \left(A_j^{(i)} - \bar{A}_j\right)^2 \sum_{i=1}^{n} \left(A_k^{(i)} - \bar{A}_k\right)^2}} \right|, \tag{33}$$

where $\bar{A}_j$ resp. $\bar{A}_k$ denote the mean value of attributes A_j resp. A_k.

Let r^j be the number of categories of attribute A_j and r^k be the number of categories of attribute A_k. In the scope of the following formula, let $\{1, \ldots, r^j\}$ be the set of categories of attribute A_j and $\{1, \ldots, r^k\}$ be the set of categories of attribute A_k.

Then, Cramér's V Statistic of attributes A_j, A_k is defined as

$$V_{jk} = \sqrt{\frac{\chi^2_{jk}/n}{\min(r^j - 1, r^k - 1)}}, \tag{34}$$

where

$$\chi^2 := \sum_{p=1}^{r^j} \sum_{q=1}^{r^k} \frac{(n \cdot Pr[X_j = p, X_k = q] - n \cdot Pr[X_j = p] \cdot Pr[X_k = q])^2}{n \cdot Pr[X_j = p] \cdot Pr[X_k = q]} \tag{35}$$

is the chi-squared independence statistic.

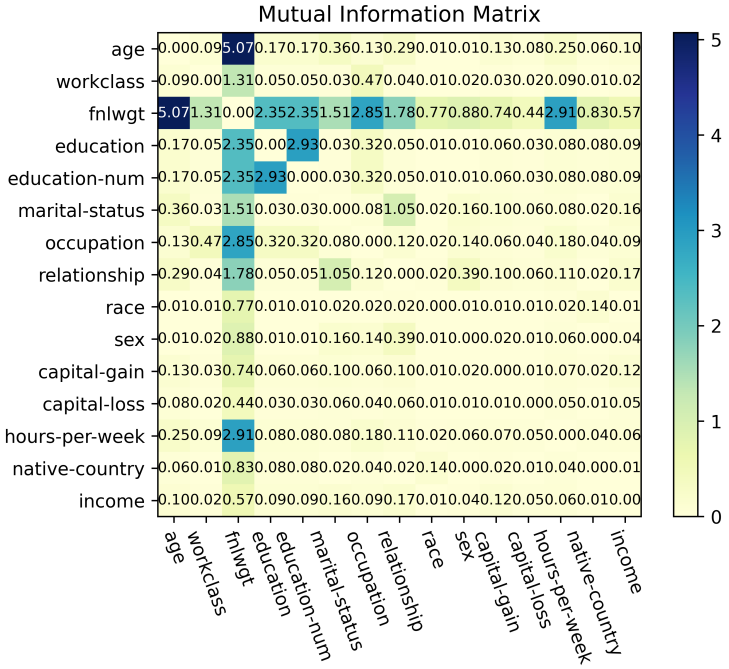


Figure 10. Example. Considering the Adult dataset as an example, this dataset can be used for the supervised training of a machine learning algorithm to classify persons having income $\leq$USD 50 K. The categorical attributes `education` and `education-num` contain highly mutual information ($I(A_{education}, A_{education-num}) \approx 2.93$) and might be part of different fragments, whereas the categorical attributes `race` and `sex` do not contain highly mutual information ($I(A_{race}, A_{sex}) \approx 0.01$) and can be part of the same fragment in vertical fragmentation. The calculated mutual information values are based on the training dataset (without the test data) of the Adult dataset. The matrix is symmetric because the function in (30) is symmetric. The values are rounded to two decimal places.

Figure 11 shows an example where the absolute value of the Pearson Correlation and Cramér's V Statistic are calculated for numerical resp. categorical attributes in the Adult dataset.

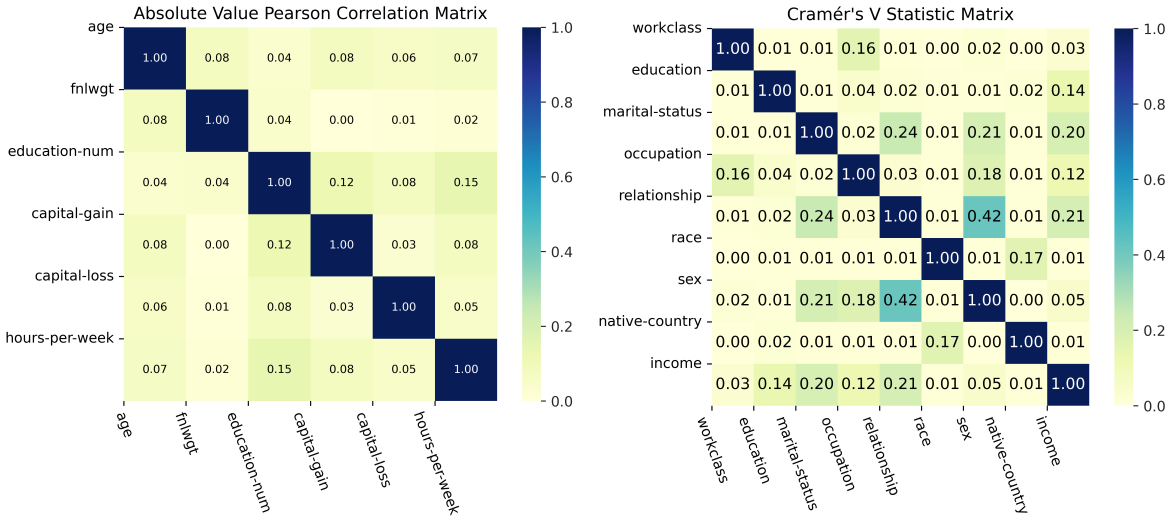


Figure 11. Example. Absolute values of Pearson Correlation coefficients and Cramér's V Statistic coefficients in the Adult dataset. Both matrices are symmetric. The values are rounded to two decimal places.

6. Legal Perspective

Sections 4 and 5 have presented technical procedures, and the consequences of the anonymization of tabular datasets have been worked out. To comply with the legal requirement for anonymization in the EU, especially concerning the GDPR, the legal basis and prerequisites must be elaborated. Based on this, conclusions about the legally secure and robust anonymization of tabular data can be drawn. In general, the legal literature on anonymization is not restricted to structured data.

However, the literature discussed in this review can be straightforwardly related to tabular data but not to unstructured data.

Firstly, we look at the legal aspects of data anonymization in general. The legal framework and requirements for handling anonymized data are analyzed. Subsequently, the problem of anonymizing tabular data is addressed, and existing legislation, analyzed. Particular attention is paid to the GDPR, which must be interpreted as the legal basis for this problem. Furthermore, different approaches to anonymizing data are considered. Especially, the absolute and relative theories of anonymization are discussed, and the different legal interpretations are highlighted. Lastly, an evaluation of the privacy models is carried out with an individual evaluation of the *k*-anonymity, *l*-diversity, and *t*-closeness privacy models, which serve as common approaches to anonymizing tabular data. Relevant factors such as the effectiveness and security of anonymization techniques are considered.

6.1. Synopsis of the Problem

When publishing data, the GDPR sets the framework and requirements for lawful publication. The aim of this law is to protect the individual's right to informational self-determination, i.e., the individual's own influence on the dissemination and collection of personal data is to be preserved [53].

The European GDPR refers in its scope exclusively to personal data. This means that all data that cannot be traced back to an identifiable person fall outside the scope of protection and are generally available as Open Data. Despite the considerable importance of the distinction between personal reference and anonymity, the GDPR does not regulate this but merely presupposes the concept of anonymity as a counterpart to personal data.

According to Art. 4 (1) GDPR "personal data means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier, or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural, or social identity of that natural person". In this context, the Article 29 Data Protection Working Party stated, in Opinion 4/2007 WP 136, that identification is normally achieved using particular pieces of information, which are called "identifiers" [54]. They are distinguished in "directly" and "indirectly" identifiers.

Thereby—in the context of tabular data—in our terminology, "directly" identifier refers to a Direct Identifier and "indirectly" identifier refers to an attribute that is part of a QI. A person may be directly identified by name, whereas they may be identified indirectly by a telephone number, car registration, or by a combination of significant criteria, which allows them to be recognized by narrowing down the group to which they belong (age, occupation, place of residence) [54].

Particularly with regard to Indirect Identifiers, the issue arises when a reference to a person still exists. Some characteristics are so unique that someone can be identified with no effort ("present Prime Minister of Spain"), but a combination of several different details may also be specific enough to narrow it down to one person, especially if someone has access to additional information [54]. According to this, sufficient anonymization only exists if this personal reference is removed and is not traceable [55].

Hereby, it should be pointed out that in [54], "[. . .] it is not necessary for the information to be considered as personal data that it is contained in a structured database or file".

However, the given examples mostly refer to structured data, as they are given in tabular datasets.

Further, as Recital 26 to the GDPR states, "the principles of data protection should therefore not apply to anonymous information, namely, information, which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable".

Anonymization occurs when personal data are changed in such a way that the person behind them can no longer be identified by personal and factual circumstances [56]. This also applies to the remaining or otherwise related datasets in their entirety [57]. The complexity of anonymity, therefore, lies in the definition, which is difficult to delimit and determine, of which datasets have which attributes that are sufficiently related to a person. This can only be performed with an intensive examination of the type and scope of the existing data and the data to be anonymized [57].

To obtain meaningful Open Data, a careful and difficult balance between sufficient information and effective anonymization to protect data subjects is necessary. Basically, anonymization must be distinguished from pseudonymization, which is essentially characterized by the fact that the data and persons can be identified again by using a code or key [55]. So far, pseudonymization has been considered insufficient and treated as personal data [56]. However, the European General Court (EGC) recently ruled that under certain circumstances, pseudonymous data may not fall under the scope of the GDPR if the data recipient lacks means for re-identification. The critical factor is whether the recipient has access to the decryption key or can obtain it. If not, the data are not considered personal data and thus do not fall under the GDPR [58].

6.2. Recital 26

Recital 26 to the GDPR further demands "to ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments".

In determining the relevant knowledge and means, Recital 26, therefore, requires a risk analysis to evaluate the likelihood of the risk of re-identification. In this analysis, an objective standard must be used, and in principle, a purely abstract standard of measurement must be applied, not the subjective interests and motivation for the use of such data. Under certain circumstances, however, these must also be included in the assessment criteria [53].

The risk of re-identification must, therefore, be assessed on a case-by-case basis. However, the interpretation of these requirements and the extent to which the available knowledge and means of third parties are to be taken into account are controversial. In this context, the spectrum of opinions is divided with regard to the requirements for the feasibility of establishing a connection to a person. It is questionable whether it depends on the respective Data Controller (relative personal reference) or whether anybody can establish the personal reference (absolute personal reference) [53].

6.3. Absolute Personal Reference/Zero-Risk Approach

The absolute approach shows two main considerations. On the one hand, it is about the group of people who must be considered potential de-anonymizers. The other is the re-identification risk that still exists due to the means available to this group of people. According to the absolute personal reference approach, a person becomes identifiable if anybody at all can re-establish the personal reference. All means available to this third party must be deliberated over. Hence, this approach can only be met if all anonymization is fully and completely irreversible and the capability of de-anonymization is eliminated [59]. In this regard, it is sometimes demanded that the original and thus still personal data records are deleted after anonymization has been implemented [60]. This refers to Tuple

Suppression, which is explained in Section 4.3. According to this approach, they are still personal data when a Data Controller does not delete the original data and hands over the anonymized dataset [61]. Accordingly, all possibilities for reversing the anonymization process must be taken into contemplation. This also includes illegal means of obtaining special knowledge as well as potential infringements of professional confidentiality [62].

To a greater extent, nevertheless, such a scale should not be required and is simply not feasible, according to the state of the art [63]. This also reflects both the telos of the law and the wording of Recital 26. Recital 26 states that "all the means reasonably likely to be used" should be deliberated. Hence, the GDPR does not consider all and every possibility of de-anonymization. It more likely supports a risk-based approach to it, which must be evaluated on the basis of the circumstances of the individual case.

Furthermore, following this absolute approach would mean that most data must still be considered personal, making true anonymization practically impossible. The main issue lies in the fact that there can never be complete certainty that no one else possesses additional knowledge or data that could potentially lead to re-identification [64].

6.4. Relative Personal Reference/Risk-Based Approach

The relative approach also exhibits two considerations that run parallel to those of the absolute approach. First, the circle of persons who need to be focused on is tighter. Secondly, the relative approach acknowledges a certain risk of de-identification [62,64]. Moreover, when dealing with Open Data, the choice between the relative and absolute approaches becomes largely inconsequential. The very nature of Open Data dictates that they should be accessible to a broad and diverse audience, opening the data to virtually anybody interested in utilizing them. As a result, the practical reality of Open Data means that considerations must extend to any potential data recipient, since they all have access to the shared data. Therefore, it is necessary to consider anybody as a potential de-anonymizer. The absolute and relative approaches thus lead to the same result. However, a key distinction between the relative approach and the absolute one emerges concerning the treatment of re-identification risk. While the absolute approach aspires to eliminate any possibility of re-identification, the relative approach recognizes that a certain level of re-identification risk may persist. The decisive factor is then the assessment of the risk and the inclusion of risk factors.

6.5. Tightened Relative Personal Reference of the EU's Court of Justice

The EU's Court of Justice (ECJ) developed a conciliatory, relative approach to establishing the reference to persons in the context of a preliminary ruling in 2016. In this respect, the ECJ dealt with the question of the extent to which the knowledge and means of third parties should be included in accordance with Recital 26, so we are referring to anonymized data. The decisive issue was whether dynamic IP addresses constitute personal data. The crucial question was which conditions must be met for a Data Controller to "reasonably" have access to the data held by a third party [64]. As General Advocate Sanchez pointed out, Recital 26 does not refer to any means that may be used by anybody but constrains these means to "likely reasonably to use".

Therefore, a risk-based approach is more in line with the wording. Third parties are persons to whom any person may reasonably turn to obtain additional data or knowledge for the purpose of identification. After all, the General Advocate set forth that "otherwise [...] it would be virtually impossible to discriminate between the various means, since it would always be possible to imagine the hypothetical contingency of a third party who, no matter how inaccessible to the [data controller], could—now or in the future—have additional relevant data to assist in the identification of a [person]" [64].

This restriction of the absolute theory and tightening of the relative theory have been endorsed by the ECJ. In this respect, the absolute theory is limited to the extent that additional knowledge, which can only be gained using illegal methods or is practically

impossible on account of the fact that it requires a disproportionate effort in terms of time, cost, and manpower [64]. Thus, the risk of identification appears to be negligible [64].

The relative approach, on the other hand, is tightened to the effect that they are still to be considered personal data if there are legal means that can be used to obtain additional knowledge from a third party that can enable the identification of a person [64]. However, the extent to which such legal means are available and whether it is reasonable to expect them to be used remains an open question. This concretization work is, therefore, incumbent on the national courts [64].

6.6. Evaluation Standards for the Risk Assessment of the Techniques

The Art. 29 Data Protection Working Party sets out various criteria for assessing the risk of individuals being identifiable or determinable when personal data are anonymized. The individual risk groups are merely a framework for evaluating the risk of identification. These principles should always be applied to the individual case and require a thorough evaluation. According to the idea of the data protection authority, Data Controllers should submit a final risk evaluation to the relevant authority. This is recommended as a general concept that a Data Controller drafts for his existing and expected datasets.

The first aspect of risks is singling out individuals from datasets [61]. The initial point is anonymized data records that have been generalized, for example. The aim of a legally secure anonymization process is to form these groups on such a scale that an individual assignment of attributes to a single person is no longer possible [65]. This is to be achieved by ensuring that the combined group has several identical attributes. The danger of singling out, therefore, exists within small group formations as well as with extreme attributes, since these are easier to assign. If persons in group formations still have unique characteristics of attributes, this favors classification. In order to prevent singling out, an appropriately large number of similar attributes must be chosen based on the evaluation of the individual case and the dataset. In this evaluation process, special attention should be paid to preserving the information content [61]. Consequently, if the *k*-groups are becoming too large, the information value can be reduced or falsified. Therefore, the information content of the dataset should always be taken into account, as this can result in data being rendered unrecognizable or falsified. In this way, the Data Controller can maintain the information content of other attributes and still guarantee anonymity.

The second risk factor relates to the linkability of data [61]. In relation to an anonymous dataset, this must be considered in combination with two individually anonymous datasets. If a Data Controller publishes several anonymized datasets, these must also preserve anonymity in their entirety. If individual persons can be determined from the combination of these two datasets, because individual attributes can now be linked together, the data are still to be considered personal [66]. In this respect, this approach has substantial uncertainty. It is questionable, and not yet clarified, which data are to be considered for this purpose. Certainly, the entirety of the publication is to be taken into account, but it is debatable whether data already published by third parties are also to be included [67]. Or, what probably leads to the widest extension, whether third parties have data at their disposal with which a linkage leads to the identifiability of individuals. Again, the jurisprudence of the ECJ can be used, that is, only additional knowledge that can be obtained by legal means is taken into consideration.

The last criterion set by the Art. 29 Data Protection Working Party is the so-called inference [61]. This is the most difficult requirement to circumvent. Basically, it means that conclusions can be drawn from datasets for the entirety of persons. In view of the challenges of anonymization, it rather demands that no conclusions that could be used to infer an individual person can be drawn from the published dataset. Here, too, there is a lack of concreteness in differentiation from singling out. However, reference attributes are probably more limited to the individual dataset from which assumptions could be drawn.

In the further outlook, each anonymization concept and method is, therefore, examined with regard to these three risk factors [61]. Other aspects may also be included as risks

in the evaluation, so the standard for these three aspects from the perspective of the "motivated intruder" must always be set. This "motivated intruder test" is intended to test the anonymization carried out for its stability and, as above, is based on the individual case. The motivation of the intruder is inevitably measured according to the value and information content of the dataset.

6.7. Legal Evaluation

This subsection conducts a legal evaluation by embedding technical terms such as privacy models in a legal context.

6.7.1. Identifiers, Quasi-Identifiers, and Sensitive Attributes

In the process of anonymization using the individual models, the QIs are to be determined and evaluated. For example, these might include dates of specific events (death, birth, discharge from a hospital, etc.), postal codes, sex, ethnicity, etc. [68]. One can orient oneself towards an assessment system that evaluates and assesses the attributes. This should essentially identify all SAs, also in the sense of the GDPR. For this purpose, all variables are listed and evaluated within the framework of three case groups. The assessment ranges from low (1) to medium (2) to high (3). The first category for the individual variables is "replication", in which the information is assessed according to how consistently it appears in connection with a person. A low score is given to measured blood pressure, while a high score is given to a person's date of birth. The second group is concerned with the "availability" of the information. The decisive factor, here, is how available this information or variable is for third parties to re-identify. As already shown above, the ECJ's standard also affects this assessment as to how far-reaching additional knowledge is to be taken into account. Therefore, the laboratory values of a person are difficult to obtain, whereas, as in the example of the "Breyer" case, the person behind an IP address can certainly be obtained by legal means if there is a legitimate interest. This should also be considered for public registers, such as the land registry. The last category concerns "distinguishability", according to which it is possible to assess how people can be distinguished from each other by means of individual values. For example, a ZIP code with a complete reproduction is to be classified as higher than one with a shortened reproduction [11].

6.7.2. *k*-Anonymity

The privacy model *k*-anonymity, which is defined in Section 5.3.1, ensures that given a QI, each record is indistinguishable from at least $k-1$ other records, making it more difficult for attackers to identify individuals by their attributes [3]. The degree of privacy protection depends on the quality and quantity of attributes in the dataset and the choice of k. The larger k, the larger its group, and the more securely an individual is protected from re-identification.

Singling out within a *k*-group is made more difficult by the fact that all individuals have the same QI and are indistinguishable based on them, such that individuals can hide behind the *k*-group.

However, Data Processors must also consider the risk of attribute disclosure, where an attacker can infer sensitive information about an individual even if they cannot directly re-identify them. This may still be possible with linkability and inference. Linkability of records may still be possible, because the probability of $1/k$ with small k is sufficient to make correlations about affected individuals among records in a *k*-group.

Another deficit of the *k*-anonymity model is that attacks are not closed with inference techniques [65]. If all *k*-individuals belong to the same group and it is known to which group an individual belongs, it is very easy to determine the value of a property. Attackers are able to extract information from the dataset and make inferences about the affected individuals, whether it is included in the dataset or not.

Therefore, whether this model alone ensures compliance with the anonymization requirement of the GDPR is largely negated. To achieve robust anonymization, additional models such as *l*-diversity or *t*-closeness can be used.

Nevertheless, the model is used in anonymization applications because it provides the basic structure for anonymization when values are not to be corrupted, as it is the case with perturbation. The LEOSS cohort study [10] uses an anonymization pipeline built on *k* equal to 11 by applying the ARX tool [9]. Thus, they follow the recommendation of the Art. 29 Data Protection Working Party (WP216) [61], which evaluates a *k*-value less than or equal to 10 as insufficient. The *k*-value depends, among other things, on the number of aggregated attributes [57] used in a QI. In the NAPKON study, the qualitative analysis of the attributes included in the dataset was controlled for the risk of linkage or selection by reducing the uniqueness of the combinations of the variables age, sex, quarter, and year of diagnosis and cohort [69].

6.7.3. *l*-Diversity

The privacy model *l*-diversity, which is defined in Section 5.3.2, was introduced as an extension of *k*-anonymity to compensate one of its major shortcomings: the failure to account for the distribution of SAs within each group of *k*-indistinguishable individuals [45]. This deficiency can lead to the disclosure of SAs resulting from the merging to *k*-groups. The advancement aims to ensure that deterministic attacks using inference techniques are no longer possible by guaranteeing that the individual attributes in each equivalence class have at least *l* different values, so that attackers are always guaranteed significant uncertainty about a particular affected individual [61].

Thus, the evaluation in [68] shows two different shortcomings of *l*-diversity, when the *l* values for each SA are not well represented. A similarity attack can be performed when the SAs fulfill the criterion of *l*-diversity but are semantically similar. Despite meeting the requirement of *l*-diversity, it is possible to learn that someone has cancer when every attribute value is a specific form of cancer. An attack on skewness can be made when the overall distribution is skewed. Then, *l*-diversity cannot prevent attribute disclosure. This is the case when the distribution of attribute values in a dataset consists predominantly of one of two possible values and a *k*-group has the other value except for one entry. This allows assumptions to be derived about this group that an attacker can use.

Despite possible protection from inference techniques, linkability may still be possible even with diversification because this risk still remains on *k*-anonymity settings. Only the risk of singling out can be prevented when implementing *l*-Diversity as an extension of *k*-anonymity. *l*-diversity processes just the SAs that were initially unaffected. Unlike *k*-anonymity, there is no recommendation from WP216 for a threshold of *l*.

This privacy model is suitable for protecting data from attacks using inference techniques when the values are well distributed and represented. However, it should be noted that this technique cannot prevent information leakage if the attribute values within a group are inconsistently distributed, have low bandwidth, or are semantically similar. Eventually, the concept of *l*-diversity provides room for attacks using inference techniques [61].

6.7.4. *t*-Closeness

The privacy model *t*-closeness, which is defined in Section 5.3.3, deals with a new measure of security and complements *l*-diversity [46]. It takes into account the unavoidable gain in knowledge of an attacker when considering all SA values in the entire dataset. *t*-Closeness represents a measure of minimal knowledge gain that results from considering a generalized *k*-group compared with the entire dataset. This also means that any group of individuals, indistinguishable on the basis of the QI, behind which a person is anonymized, can hardly be distinguished from any other group with respect to their SA values by the *t*-closeness-defined measure. Thus, a person's data are better protected in their anonymizing group than was the case with *l*-diversity, since this group hardly reveals more information than the entire distribution.

In the specific case where the attribute values within a group are non-uniformly distributed, have a narrow range of values, or are semantically similar, an approach known as *t*-closeness is applied. This represents a further improvement in anonymization using generalization and consists of a procedure in which the data are partitioned into groups in such a way that the original distribution of the attribute values in the original dataset is reproduced as far as possible [61]. However, WP216 has not given any recommendation for the *t*-value, so it depends on case-by-case consideration. One approach would be to incrementally increase the *t*-value if re-identification by an attacker with the current value is still possible.

With *t*-closeness, a dataset processed with *k*-anonymity is improved regarding the risk of inference and was implemented in the LEOSS cohort study [10], with *t* equal to 0.5.

Nevertheless, data anonymized using *k*-anonymity and *t*-closeness are still vulnerable to inference techniques and have to reviewed case by case. Whereas in *k*-anonymity and *l*-diversity, large values mean better privacy, in *t*-closeness, small values mean better privacy.

6.7.5. Differential Privacy

DP, which is defined in Section 4.6, applied as a randomized process, manipulates data in such a way that the direct link between data and the data subject can be removed [61]. There are several mechanisms that satisfy the defined anonymity criterion and are applicable to different types of data. The method ensures the protection of individual data by modifying the results by adding random noise. This can limit a potential attacker's ability to draw conclusions about the attribute value of a single data point, even if they know all the attribute values of the other data points. By adding random noise, the influence of a single data point on the statistical result is hidden [70]. With regard to the risk criteria, it can be seen that singling out can be prevented under certain circumstances. Linking and inference can still be possible with multiple applications and are thus dependent on the so-called privacy budget, which refers to parameter ϵ in Section 4.6.

6.7.6. Synthetic Data

As explained in Section 4.7, synthetic approaches can be used as a workaround to anonymize tabular data. Artificially generated synthetic data retain the statistical characteristics of the original data. This process can involve utilizing a machine learning model that comprehends the structure and statistical distribution of the original data to create synthetic data. Preserving the statistical properties of the original data is vital, as it enables data analysts to derive significant insights from the synthetic data, treating them as if they were drawn directly from the original dataset. To introduce a diverse range of data, the generation process may incorporate a certain level of unrelated randomness into synthetic data [71].

Synthetic data can help to ensure that an individual's records are not singled out or linked. However, if an adversary knows of the presence of a person in the original dataset, even if that person cannot be individualized, sensitive inferences such as attribute disclosure may still be possible, as shown in [72]. Moreover, machine learning models can be exposed to privacy attacks by the so-called Membership Inference Attacks or Model Inversion Attacks [73].

6.7.7. Risk Assessment Overview

Based on the findings in Sections 6.7.2–6.7.6, Table 3 gives an overview of risk assessments of the discussed privacy models and privacy-enhancing technologies for anonymizing tabular data. We only rate with respect to the attack scenarios that are described by the Art. 29 Data Protection Working Party: singling out, linkability, and inference.

Table 3. Risk assessment for anonymization methods of tabular data. (1): Risk depends on chosen k. (2): It does not take into account similarity attacks. (3): Based on k-anonymity. (4): Risk depends on value distribution of Sensitive Attributes. (5): Risk depends on privacy budget. (6): Might be combined with DP. +: The method can be considered a strategy to defend against the attack scenario. −: The method cannot solely be considered a defense strategy against the attack scenario.

	Singling Out	**Linkability**	**Inference**
k-Anonymity	+	− (1)	− (2)
l-Diversity	+ (3)	− (1,3)	+ (2,4)
t-Closeness	+ (3)	− (1,3)	+ (2,4)
DP	+	+ (5)	+ (5)
Synthetic data	+	+	− (6)

7. Discussion

In our exploration of anonymization methods and scores for tabular data, some unclarities and issues are present.

Foremost is the uncertainty surrounding the choice of QIs and thresholds for privacy models. A fundamental challenge is the inability to make a priori assumptions about the knowledge an adversary possesses regarding records in tabular data. Often, there is a vast array of potential QIs that could be exploited, which goes hand in hand with the lack of context understanding.

This issue is further complicated by the fact that the privacy models adopted only cover specific scenarios, leaving room for specific attack scenarios to succeed.

Further, to maximize privacy protection, we may compromise the data utility. A potential solution might be found in combining different anonymization methods, each addressing specific weaknesses. For instance, use-case-specific DP can be applied to provide an additional layer of security. However, implementation details and the actual compatibility of methods are yet to be thoroughly studied. As an example, the interaction between t-closeness and group formation has shown that the elimination of group records to achieve certain t-closeness, k-anonymity, and l-diversity can unintentionally lead to higher t. This can potentially compromise the achieved anonymization.

Moreover, the structure and composition of the dataset themselves poses a challenge. Often, SAs are the target variables, thereby making their concealment problematic. Privacy models, such as l-diversity, depend on the number of attribute values for the SA, meaning that the effectiveness of the method varies based on the characteristics of the dataset. When it comes to anonymizing high-dimensional tabular data, as described in Section 5.5, one also has to deal with the Curse of Dimensionality.

Anonymizing the Adult dataset into k-anonymity with $k > 10$ still yields comparable utility for different ML models, but this is data- and task-dependent and DP might additionally be applied in model inference [48].

As Wagner et al. [14] have recommended, a selection of multiple metrics to cover multiple aspects of privacy should be pursued. This approach allows for more robust privacy protection, minimizing the chances of oversights and weaknesses.

The implementation of these privacy protection measures presents its own set of challenges. To begin with, different types of data, such as categorical and numerical, necessitate different approaches. Some attributes might even possess dual characteristics, complicating the anonymization procedure. Different possible definitions and ways of implementing these methods add to the complexity. Privacy models must also be adapted to data types, with a clear understanding of the differences between integers and floating-point numbers, or categorical versus numerical data types. Additionally, applying these methods often involves a trial-and-error process. Multi-stage anonymization is a potential strategy that might yield better results, though the complexity and difficulty of execution cannot be underestimated. For example, achieving certain k-anonymity using generalization and suppression with minimal loss of information [18] is an NP-hard problem. This implies

that execution time could be exponential in the worst-case scenarios—a factor that needs to be tested and considered in the implementation phase.

Last but not least, the context of data—whether they are fixed or streaming—poses another challenge. Privacy protection measures for streaming or online data may require a different approach, considering the time and space complexity involved.

Future research should focus on addressing these issues, providing a more comprehensive and effective solution to data anonymization of tabular data.

8. Conclusions

In conclusion, this article has examined the technical and legal considerations of data anonymization and explored different approaches to solving this problem.

From the legal perspective, based on our analysis and legal evaluation, the following conclusions can be drawn. The risk-based approach, in alignment with the ECJ case law in the "Breyer" case, highlights the importance of considering legally obtainable additional knowledge when assessing the acceptable re-identification risk. This approach enhances the understanding of data anonymity by taking into account relevant information that can potentially lead to re-identification. Due to the missing legal requirements for robust anonymization, a recommendation for k-anonymity with k greater than 10 was made by the Article 29 Data Protection Working Party in WP216 [61]. Prior to implementing k-anonymity, it is crucial to identify the QIs using the evaluation table and the provided evaluation system. Furthermore, the opinion suggests the use of t-closeness. Similarly, there are no legal requirements at this point to ensure legally compliant anonymization. Only in [10], a t-value set at 0.5 was considered to be a high level of privacy protection. However, since the risk-based approach is based on individual-case assessment, it must be considered that these values should not be considered universally applicable. The ongoing uncertainty makes anonymization still a challenging endeavor. In addition, it is important to note that for anonymized data, future consideration of the EU Data Governance Act, particularly in relation to data rooms and the security of such data, becomes crucial. The Data Governance Act aims to establish a framework for secure and responsible data sharing that ensures data protection and governance in data rooms.

Future research and advancements in the field should continue to explore the legal and technical aspects of data anonymization, taking into account evolving legislation, court rulings, and emerging best practices. By staying abreast of these developments and adhering to appropriate standards, a data-driven environment that respects privacy, safeguards personal information, and promotes responsible data sharing practices can be fostered.

Anonymization procedures can support the creation of Open Data. Similar to Open Source, Open Data represent an economically and socially relevant concept. For example, it is part of the digital strategy resp. the Open Data strategy of the current resp. the previous federal government in Germany. However, a challenge may be that under the current European regulations, in the near future, all data might be classified as personal data as a result of moving forward into a data-driven world. In [74], this is named the Law of Everything. The reason for this is the widely defined rules on data protection and the definition of the terms "information" and "personal data" by the GDPR. This is accelerated by the rapid advances in technology, which enable ever greater interpretability of data as well as the increased collection of information in real time. The Law of Everything is an approach with a worthy goal but not one that can be implemented sustainably with current procedures.

Author Contributions: Conceptualization, R.A.; Methodology, R.A., J.F., J.G. and E.M.; Software, R.A.; Validation, R.A., J.F. and M.H.; Formal analysis, R.A.; Investigation, R.A. and J.F.; Resources, R.A., J.F., J.G. and E.M.; Data curation, R.A.; Writing—original draft preparation, R.A., J.G. and E.M.; Writing—review and editing, R.A., J.F., M.H., B.B. and M.S.; Visualization, R.A. and J.F.; Supervision, M.H., B.B. and M.S.; Project administration, M.H., B.B. and M.S.; Funding acquisition, M.H. and M.S. All authors have read and agreed to the published version of the manuscript.

Funding: The research project EAsyAnon ("Verbundprojekt: Empfehlungs- und Auditsystem zur Anonymisierung", funding indicator: 16KISA128K) is funded by the European Union under the umbrella of the funding guideline "Forschungsnetzwerk Anonymisierung für eine sichere Datennutzung" from the German Federal Ministry of Education and Research (BMBF).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset Adult used in this study for experiments is openly available to download from the UCI Machine Learning Repository. Data can be found at https://archive.ics.uci.edu/dataset/2/adult (accessed on 15 May 2023).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DP	Differential Privacy
DP-SGD	Differentially Private Stochastic Gradient Descent
ECJ	European Court of Justice
EGC	European General Court
EU	European Union
FMRMR	Fragmentation Minimum Redundancy Maximum Relevance
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
HIPAA	Health Insurance Portability and Accountability Act
LDA	Linear Discriminant Analysis
LSTM	Long Short-Term Memory
MIMIC-III	Medical Information Mart for Intensive Care
PCA	Principal Component Analysis
PPDP	Privacy-preserving data publishing
PPGIS	Public Participation Geographic Information System
QI	Quasi-Identifier
SA	Sensitive Attribute
SVD	Singular Value Decomposition

References

1. Weitzenboeck, E.M.; Lison, P.; Cyndecka, M.; Langford, M. The GDPR and unstructured data: Is anonymization possible? *Int. Data Priv. Law* **2022**, *12*, 184–206. [CrossRef]
2. Samarati, P.; Sweeney, L. Protecting privacy when disclosing information: K-anonymity and its enforcement through generalization and suppression. In Proceedings of the IEEE Symposium on Security and Privacy, Oakland, CA, USA, 3–6 May 1998; pp. 1–19.
3. Sweeney, L. K-Anonymity: A Model for Protecting Privacy. *Int. J. Uncertain. Fuzziness-Knowl.-Based Syst.* **2002**, *10*, 557–570. [CrossRef]
4. Ford, E.; Tyler, R.; Johnston, N.; Spencer-Hughes, V.; Evans, G.; Elsom, J.; Madzvamuse, A.; Clay, J.; Gilchrist, K.; Rees-Roberts, M. Challenges Encountered and Lessons Learned when Using a Novel Anonymised Linked Dataset of Health and Social Care Records for Public Health Intelligence: The Sussex Integrated Dataset. *Information* **2023**, *14*, 106. [CrossRef]
5. Becker, B.; Kohavi, R. Adult. UCI Machine Learning Repository. 1996. Available online: https://archive-beta.ics.uci.edu/dataset/2/adult (accessed on 15 May 2023).
6. Majeed, A.; Lee, S. Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey. *IEEE Access* **2021**, *9*, 8512–8545. [CrossRef]
7. Hasanzadeh, K.; Kajosaari, A.; Häggman, D.; Kyttä, M. A context sensitive approach to anonymizing public participation GIS data: From development to the assessment of anonymization effects on data quality. *Comput. Environ. Urban Syst.* **2020**, *83*, 101513. .: 10.1016/j.compenvurbsys.2020.101513. [CrossRef]
8. Olatunji, I.E.; Rauch, J.; Katzensteiner, M.; Khosla, M. A review of anonymization for healthcare data. In *Big Data*; Mary Ann Liebert, Inc.: New Rochelle, NY, USA, 2022.
9. Prasser, F.; Kohlmayer, F. Putting statistical disclosure control into practice: The ARX data anonymization tool. In *Medical Data Privacy Handbook*; Springer: Cham, Switzerland, 2015; pp. 111–148.

10. Jakob, C.E.M.; Kohlmayer, F.; Meurers, T.; Vehreschild, J.J.; Prasser, F. Design and evaluation of a data anonymization pipeline to promote Open Science on COVID-19. *Sci. Data* **2020**, *7*, 435. [CrossRef]
11. Malin, B.; Loukides, G.; Benitez, K.; Clayton, E.W. Identifiability in biobanks: Models, measures, and mitigation strategies. *Hum. Genet.* **2011**, *130*, 383–392. [CrossRef]
12. Ram Mohan Rao, P.; Murali Krishna, S.; Siva Kumar, A. Privacy preservation techniques in big data analytics: A survey. *J. Big Data* **2018**, *5*, 33. [CrossRef]
13. Haber, A.C.; Sax, U.; Prasser, F.; the NFDI4Health Consortium. Open tools for quantitative anonymization of tabular phenotype data: literature review. *Briefings Bioinform.* **2022**, *23*, bbac440. [CrossRef]
14. Wagner, I.; Eckhoff, D. Technical Privacy Metrics. *ACM Comput. Surv.* **2018**, *51*, 1–38. [CrossRef]
15. Vokinger, K.; Stekhoven, D.; Krauthammer, M. Lost in Anonymization—A Data Anonymization Reference Classification Merging Legal and Technical Considerations. *J. Law Med. Ethics* **2020**, *48*, 228–231. [CrossRef] [PubMed]
16. Zibuschka, J.; Kurowski, S.; Roßnagel, H.; Schunck, C.H.; Zimmermann, C. Anonymization Is Dead—Long Live Privacy. In Proceedings of the Open Identity Summit 2019, Garmisch-Partenkirchen, Germany, 28–29 March 2019; Roßnagel, H., Wagner, S., Hühnlein, D., Eds.; Gesellschaft für Informatik: Bonn, Germany, 2019; pp. 71–82.
17. Rights (OCR), Office for Civil. Methods for De-Identification of PHI. HHS.gov. 2012. Available online: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html (accessed on 21 July 2023).
18. Gionis, A.; Tassa, T. k-Anonymization with Minimal Loss of Information. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 206–219. [CrossRef]
19. Terrovitis, M.; Mamoulis, N.; Kalnis, P. Local and global recoding methods for anonymizing set-valued data. *VLDB J.* **2011**, *20*, 83–106. [CrossRef]
20. Agrawal, R.; Srikant, R. Privacy-Preserving Data Mining. In Proceedings of the SIGMOD '00: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, TX, USA, 16–18 May 2000; Association for Computing Machinery: New York, NY, USA, 2000; pp. 439–450. [CrossRef]
21. Bayardo, R.; Agrawal, R. Data privacy through optimal k-anonymization. In Proceedings of the 21st International Conference on Data Engineering (ICDE'05), Tokyo, Japan, 5–8 April 2005; pp. 217–228. [CrossRef]
22. Dwork, C. Differential Privacy. In *Automata, Languages and Programming, Proceedings of the 33rd International Colloquium on Automata, Languages and Programming, Part II (ICALP 2006), Venice, Italy, 10–14 July 2006*; Springer: Berlin/Heidelberg, Germany, 2006, Volume 4052, pp. 1–12.
23. Wang, T.; Zhang, X.; Feng, J.; Yang, X. A Comprehensive Survey on Local Differential Privacy toward Data Statistics and Analysis. *Sensors* **2020**, *20*, 7030. [CrossRef]
24. Dwork, C.; Roth, A. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* **2014**, *9*, 211–407. [CrossRef]
25. Wang, Y.; Wu, X.; Hu, D. Using Randomized Response for Differential Privacy Preserving Data Collection. In Proceedings of the EDBT/ICDT Workshops, Bordeaux, France, 15 March 2016.
26. Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H.B.; Mironov, I.; Talwar, K.; Zhang, L. Deep Learning with Differential Privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, 24–28 October 2016; pp. 308–318. [CrossRef]
27. van der Maaten, L.; Hannun, A.Y. The Trade-Offs of Private Prediction. *arXiv* **2020**, arXiv:2007.05089.
28. McKenna, R.; Miklau, G.; Sheldon, D. Winning the NIST Contest: A scalable and general approach to differentially private synthetic data. *arXiv* **2021**, arXiv:2108.04978.
29. Aggarwal, C.C.; Yu, P.S. A condensation approach to privacy preserving data mining. In *Advances in Database Technology-EDBT 2004, Proceedings of the International Conference on Extending Database Technology, Crete, Greece, 14–18 March 2004*; Springer: Berlin/Heidelberg, Germany, 2004, pp. 183–199.
30. Jiang, X.; Ji, Z.; Wang, S.; Mohammed, N.; Cheng, S.; Ohno-Machado, L. Differential-Private Data Publishing Through Component Analysis. *Trans. Data Priv.* **2013**, *6*, 19–34.
31. Xu, S.; Zhang, J.; Han, D.; Wang, J. Singular value decomposition based data distortion strategy for privacy protection. *Knowl. Inf. Syst.* **2006**, *10*, 383–397. [CrossRef]
32. Soria-Comas, J.; Domingo-Ferrer, J. Mitigating the Curse of Dimensionality in Data Anonymization. In Proceedings of the Modeling Decisions for Artificial Intelligence: 16th International Conference, MDAI 2019, Milan, Italy, 4–6 September 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 346–355.
33. Xu, L.; Veeramachaneni, K. Synthesizing Tabular Data using Generative Adversarial Networks. *arXiv* **2018**, arXiv:1811.11264.
34. Park, N.; Mohammadi, M.; Gorde, K.; Jajodia, S.; Park, H.; Kim, Y. Data Synthesis based on Generative Adversarial Networks. *arXiv* **2018**, arXiv:1806.03384.
35. Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; Veeramachaneni, K. Modeling Tabular data using Conditional GAN. *arXiv* **2019**, arXiv:1907.00503.
36. Xie, L.; Lin, K.; Wang, S.; Wang, F.; Zhou, J. Differentially Private Generative Adversarial Network. *arXiv* **2018**, arXiv:1802.06739.
37. Kunar, A.; Birke, R.; Zhao, Z.; Chen, L. DTGAN: Differential Private Training for Tabular GANs. *arXiv* **2021**, arXiv:2107.02521.
38. Zakerzadeh, H.; Aggrawal, C.C.; Barker, K. Towards Breaking the Curse of Dimensionality for High-Dimensional Privacy. In Proceedings of the 2014 SIAM International Conference on Data Mining, Philadelphia, PA, USA, 24–26 April 2014.

39. Aggarwal, C.C. On K-Anonymity and the Curse of Dimensionality. In Proceedings of the VLDB '05: 31st International Conference on Very Large Data Bases, Trondheim, Norway, 30 August–2 September 2005; pp. 901–909.
40. Salas, J.; Torra, V. A General Algorithm for k-anonymity on Dynamic Databases. In Proceedings of the DPM/CBT@ESORICS, Barcelona, Spain, 6–7 September 2018.
41. Xu, J.; Wang, W.; Pei, J.; Wang, X.; Shi, B.; Fu, A. Utility-based anonymization for privacy preservation with less information loss. *SIGKDD Explor.* **2006**, *8*, 21–30. [CrossRef]
42. LeFevre, K.; DeWitt, D.; Ramakrishnan, R. Mondrian Multidimensional K-Anonymity. In Proceedings of the 22nd International Conference on Data Engineering (ICDE'06), Atlanta, GA, USA, 3–8 April 2006; p. 25. [CrossRef]
43. Elabd, E.; Abd elkader, H.; Mubarak, A.A. L—Diversity-Based Semantic Anonymaztion for Data Publishing. *Int. J. Inf. Technol. Comput. Sci.* **2015**, *7*, 1–7. [CrossRef]
44. Wang, X.; Chou, J.K.; Chen, W.; Guan, H.; Chen, W.; Lao, T.; Ma, K.L. A Utility-Aware Visual Approach for Anonymizing Multi-Attribute Tabular Data. *IEEE Trans. Vis. Comput. Graph.* **2018**, *24*, 351–360. [CrossRef]
45. Machanavajjhala, A.; Gehrke, J.; Kifer, D.; Venkitasubramaniam, M. L-diversity: Privacy beyond k-anonymity. In Proceedings of the 22nd International Conference on Data Engineering (ICDE'06), Atlanta, GA, USA, 3–8 April 2006; p. 24. [CrossRef]
46. Li, N.; Li, T.; Venkatasubramanian, S. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering, Istanbul, Turkey, 15 April 2006–20 April 2007; pp. 106–115. [CrossRef]
47. Vatsalan, D.; Rakotoarivelo, T.; Bhaskar, R.; Tyler, P.; Ladjal, D. Privacy risk quantification in education data using Markov model. *Br. J. Educ. Technol.* **2022**, *53*, 804–821. [CrossRef]
48. Díaz, J.S.P.; García, Á.L. Comparison of machine learning models applied on anonymized data with different techniques. *arXiv* **2023**, arXiv:2305.07415.
49. CSIRO. Metrics and Frameworks for Privacy Risk Assessments, CSIRO: Canberra, Australia, Adopted on 12 July 2021. 2021. Available online: https://www.csiro.au/en/research/technology-space/cyber/Metrics-and-frameworks-for-privacy-risk-assessments (accessed on 4 June 2023).
50. Bellman, R. *Dynamic Programming*, 1st ed.; Princeton University Press: Princeton, NJ, USA, 1957.
51. Ding, C.; Peng, H. Minimum redundancy feature selection from microarray gene expression data. In Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003, Stanford, CA, USA, 11–14 August 2003; pp. 523–528. [CrossRef]
52. Domingo-Ferrer, J.; Soria-Comas, J. Multi-Dimensional Randomized Response. *arXiv* **2020**, arXiv:2010.10881.
53. Kühling, J.; Buchner, B. (Eds.) *Datenschutz-Grundverordnung BDSG: Kommentar*, 3rd ed.; C.H.Beck: Bayern, Germany, 2020.
54. Article 29 Data Protection Working Party. Opinion 4/2007 on the Concept of Personal Data, WP136, Adopted on 20 June 2007. 2007. Available online: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf (accessed on 5 May 2023).
55. Auer-Reinsdorff, A.; Conrad, I. (Eds.) Früher unter dem Titel: Beck'sches Mandats-Handbuch IT-Recht. In *Handbuch IT-und Datenschutzrecht*, 2nd ed.; C.H.Beck: Bayern, Germany, 2016.
56. Paal, B.P.; Pauly, D.A.; Ernst, S. *Datenschutz-Grundverordnung, Bundesdatenschutzgesetz*; C.H.Beck: Bayern, Germany, 2021.
57. Specht, L.; Mantz, R. Handbuch europäisches und deutsches Datenschutzrecht. In *Bereichsspezifischer Datenschutz in Privatwirtschaft und öffentlichem Sektor*; C.H.Beck: München, Germany, 2019.
58. *Case T-557/20*; Single Resolution Board v European Data Protection Supervisor. ECLI:EU:T:2023:219. Official Journal of the European Union: Brussel, Belgium, 2023. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:62020TA0557 (accessed on 1 July 2023).
59. Groos, D.; van Veen, E.B. Anonymised data and the rule of law. *Eur. Data Prot. L. Rev.* **2020**, *6*, 498. [CrossRef]
60. Finck, M.; Pallas, F. They who must not be identified—distinguishing personal from non-personal data under the GDPR. *Int. Data Priv. Law* **2020**, *10*, 11–36. [CrossRef]
61. Article 29 Data Protection Working Party. *Opinion 5/2014 on Anonymisation Techniques*; WP216, Adopted on 10 April 2014; Directorate-General for Justice and Consumers: Brussel, Belgium, 2014. Available online: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf (accessed on 1 July 2023).
62. Bergt, M. Die Bestimmbarkeit als Grundproblem des Datenschutzrechts—Überblick über den Theorienstreit und Lösungsvorschlag. *Z. Datenschutz* **2015**, *365*, 345–396.
63. Burkert, C.; Federrath, H.; Marx, M.; Schwarz, M. Positionspapier zur Anonymisierung unter der DSGVO unter Besonderer Berücksichtigung der TK-Branche. Konsultationsverfahren des BfDI. 10 February 2020. Available online: https://www.bfdi.bund.de/SharedDocs/Downloads/DE/Konsultationsverfahren/1_Anonymisierung/Positionspapier-Anonymisierung.html (accessed on 11 May 2023).
64. *Case C-582/14*; Patrick Breyer v Bundesrepublik Deutschland. ECLI:EU:C:2016:779. Court of Justice of the European Union: Brussel, Belgium, 2016. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:62014CJ0582 (accessed on 1 July 2023).
65. Schwartmann, R.; Jaspers, A.; Lepperhoff, N.; Weiß, S.; Meier, M. Practice Guide to Anonymising Personal Data; Foundation for Data Protection, Leipzig 2022. Available online: https://stiftungdatenschutz.org/fileadmin/Redaktion/Dokumente/Anonymisierung_personenbezogener_Daten/SDS_Practice_Guide_to_Anonymising-Web-EN.pdf (accessed on 10 June 2023).
66. Bischoff, C. Pseudonymisierung und Anonymisierung von personenbezogenen Forschungsdaten im Rahmen klinischer Prüfungen von Arzneimitteln (Teil I)-Gesetzliche Anforderungen. *Pharma Recht* **2020**, *6*, 309–388.

67. Simitis, S.; Hornung, G.; Spiecker gen. Döhmann, I. *Datenschutzrecht: DSGVO mit BDSG*; Nomos: Baden-Baden, Germany, 2019; Volume 1.
68. Csányi, G.M.; Nagy, D.; Vági, R.; Vadász, J.P.; Orosz, T. Challenges and Open Problems of Legal Document Anonymization. *Symmetry* **2021**, *13*, 1490. [CrossRef]
69. Koll, C.E.; Hopff, S.M.; Meurers, T.; Lee, C.H.; Kohls, M.; Stellbrink, C.; Thibeault, C.; Reinke, L.; Steinbrecher, S.; Schreiber, S.; et al. Statistical biases due to anonymization evaluated in an open clinical dataset from COVID-19 patients. *Sci. Data* **2022**, *9*, 776. [CrossRef]
70. Dewes, A. Verfahren zur Anonymisierung und Pseudonymisierung von Daten. In *Datenwirtschaft und Datentechnologie: Wie aus Daten Wert Entsteht*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 183–201. [CrossRef]
71. Giomi, M.; Boenisch, F.; Wehmeyer, C.; Tasnádi, B. A Unified Framework for Quantifying Privacy Risk in Synthetic Data. *arXiv* **2022**, arXiv:2211.10459.
72. López, C.A.F. On the legal nature of synthetic data. In Proceedings of the NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research, New Orleans, LA, USA, 2 December 2022.
73. Veale, M.; Binns, R.; Edwards, L. Algorithms that Remember: Model Inversion Attacks and Data Protection Law. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* **2018**, *376*, 20180083. [CrossRef]
74. Purtova, N. The law of everything. Broad concept of personal data and future of EU data protection law. *Law Innov. Technol.* **2018**, *10*, 40–81. [CrossRef]