Machine Learning Meets Data Modification

The Potential of Pre-processing for Privacy Enchancement

Giuseppe Garofalo[1(✉)], Manel Slokom[2], Davy Preuveneers[1], Wouter Joosen[1], and Martha Larson[2,3]

[1] imec-Distrinet, KU Leuven, Celestijnenlaan 200A, 3001 Heverlee, Belgium
{giuseppe.garofalo,davy.preuveneers,wouter.joosen}@cs.kuleuven.be
[2] Delft University of Technology, Delft, The Netherlands
m.slokom@tudelft.nl
[3] Radboud University, Nijmegen, The Netherlands
M.Larson@cs.ru.nl

Abstract. We explore how data modification can enhance privacy by examining the connection between data modification and machine learning. Specifically, machine learning "meets" data modification in two ways. First, data modification can protect the data that is used to train machine learning models focusing it on the intended use and inhibiting unwanted inference. Second, machine learning can provide new ways of creating modified data. In this chapter, we discuss data modification approaches, applied during data pre-processing, that are suited for online data sharing scenarios. Specifically, we define two scenarios "User data sharing" and "Data set sharing" and describe the threat models associated with each scenario and related privacy threats. We then survey the landscape of privacy-enhancing data modification techniques that can be used to counter these threats. The picture that emerges is that data modification approaches hold promise to enhance privacy, and can be used alongside of conventional cryptographic approaches. We close with an outlook on future directions focusing on new types of data, the relationship among privacy, and the importance of taking an interdisciplinary approach to data modification for privacy enhancement.

1 Introduction

The importance of data in for gaining insight and supporting decision making has long been appreciated. However, recently recognition has grown of other aspects of data, both positive and negative. On the positive side, data are useful for training machine learning (ML) models that guide the development of new products and enable new services. ML has lead to a growing demand for data by businesses and other organizations looking to create value, to reduce costs or to boost profits. On the negative side, data can be dangerous. Large, centralized collections are susceptible to breaches and give rise to privacy and security risks. Moreover, ML algorithms introduce novel attack surfaces, opening the door to function creep by service providers and putting privacy at risk.

L. Batina et al. (Eds.): Security and Artificial Intelligence, LNCS 13049, pp. 130–155, 2022.
https://doi.org/10.1007/978-3-030-98795-4_7

It has become apparent that we need to understand how to derive benefit from data without running serious risks. Conventional approaches use encryption, or multiple layers of system security, to protect data. Such approaches are effective, but also have specific drawbacks. They are technically complex to implement and must be continuously monitored for breaches. Approaches to protecting privacy that do not suffer these drawbacks would clearly be advantageous.

In this chapter, we take a look at a set of less conventional approaches that involve an alternative process: data modification. Data modification is the practice of changing raw data into a transformed form for the purpose of protection. One commonality between conventional approaches to data protection, is that they assume that data must be maintained in its original form in order to be useful. Although, this might be the case for some applications, with the rise of machine learning there are an increasing number of cases for which the original data is not necessarily. Approaches like machine learning that work probabilistically can tolerate variation in the data, especially in cases where that variation does not impact aspects of the data most important for the task at hand.

When data modification is integrated into a data pipeline, it is usually integrated as a pre-processing step. In contrast, conventional data protection approaches can be applied multiple places along the pipeline. We use the term "pre-processing" to refer to a transformation applied to raw data, possibly during the phases of cleaning or feature extraction. Data modification at the beginning of the pipeline can be combined with other forms of encryption or security anywhere along the pipeline to add extra protection. In this chapter, however, we focuses specifically on data modification.

The result of data modification is a data set that can be shared and further used without needing to reverse the modification, as opposed to encryption, where generally only decrypted data can be used in a meaningful way. After modification, data no longer offer a viable opportunity to threaten privacy or attack security. The overhead of managing encryption keys or of monitoring system level security can be spared. Data modification can protect against data misuse by an internal party and can limit the damage done by a breach.

This chapter discusses how ML and data modification are related, and how the modification of data is growing in importance as a method for privacy enhancement. The chapter follows two major themes, corresponding to two ways in which machine learning can be said to "meet" data modification. First, data modification protects the data fed into ML algorithms. Second, ML can be used to data in order to create protection. Our chapter provides a literature survey that covers work on data modification techniques that fit into these two themes. We argue that the relationship between ML and data modification is not static, but can be anticipated to evolve in the future. Specifically, interest in data modification is driven by the growth of ML, due to both the risks associated with ML as well as the specific opportunities that it presents. Next, we turn to further discuss this effect, in order to provide important background and motivation for the use of data modification for privacy enhancement.

1.1 Risks and Opportunities of Machine Learning

Growth and Uncertainty: Machine learning applications are trained using large amounts of data. The data are often collected from people, and contain detailed information reflecting those people's identities, attributes, activities, and habits. As ML becomes central to the way that businesses produce value, more and more data are collected. More data means not only more private information, but also greater challenges in data management and storage. Data must be transferred, stored remotely, and processed using cloud services, increasing the opportunities for privacy violations. It starts becoming uncertain how data can be found in the future from the moment the data are collected. Furthermore, Machine-learning-as-a-service (MLaaS) has recently emerged, making results generated by complex models available to a wide public. MLaaS interfaces are easy and cheap. However, widely exposing models increases the risk of inference of properties of the data used to train those models.

Conventional ways of protecting data often assume top-down planning of data management rather than organic expansion, or careful control over the use of the products of data. Data modification becomes increasingly interesting as a means of privacy enhancement in conditions that cannot be fully anticipated or controlled.

Shifting Incentive Structures: Data have long been valuable, but the rise of machine learning has seen a further increase in that value. This value changes the incentive structures surrounding data that have been collected by companies. Specifically, the temptation arises to use data in ways that were not intended when the data were collected. It is not always the case that the change that triggers data to be used for an unexpected new purpose is a sudden change. It may be that the purpose for which the data are used slowly evolves away from the original purpose, a process commonly referred to as *function creep*. These issues are described by an *honest-but-curious* party. This party has the right to use the data, but is driven by an incentive structure to use it in ways inconsistent with the original purpose. The concept of curiosity should be understood with a broad interpretation that covers both the situation of greed (because data can be used to create value) and the situation of neglect (because it is easier and less expensive than to take care of data properly).

The incentive is strong to cut corners when managing or processing data, as illustrated by recent high-profile scandals. The Cambridge Analytica scandal is an example of a failure of data control [52]. It serves to illustrate that complex data environments can give rise to new ways in which data can end up where they should not be, serving a purpose that they should not serve. Another issue is that the task remains the same, but the way that the data are processed suddenly changes. A recent inquiry showed how Amazon employees were instructed to perform manual inspection and transcription of voice signals [78].

The Purpose of Data: The rise of machine learning has seen a focus on the purpose of data. Companies that collect data have a business model and train

machine learning models that help them to create value within that model. Moving forward, we expect that the collection of data will be more tightly linked to purpose. In Europe, the General Data Protection Regulations (GDPR) tackles the dangers of data by enforcing the *data minimization* principle: only data that are useful for the task to be carried out can be obtained, upon consent by the user, and for a limited time.

The rise of data sets with a purpose opens the door for a new kind of data modification: *purpose-aware modification*. Purpose-aware modification changes the data so that it is still useful for some purpose (training a particular type of model) but contains minimal privacy-sensitive information. Such data modification has the potential to be particularly effective by introducing changes along any dimensions that is (nearly) orthogonal to the relevant features. Furthermore, data minimization does not always imply removing data. Data minimization should also focus on reducing the information that the data contains. Understanding purpose-aware data modification will help us understand how to more effectively minimize data.

1.2 Scope and Outline

The data modification techniques that we cover in this chapter fulfill two prerequisites. First, they do not use cryptography. In other words, Fully Homomorphic Encryption (FHE) and Secure Multi-Party Computation (MPC) are out of scope. We refer to [86] for a thorough discussion of the crypto-oriented landscape of privacy-preserving ML.

Avoiding cryptography cuts computational requirements drastically, and avoids issues such as key management.

Second, we assume a centralized scenario. In other words, it is not possible to avoid that the modified data are at some point held by a single entity. Thus, we exclude scenarios of distributed training, also known as distributed machine learning (DML). An example is federated learning which consists of random nodes being assigned a small training task to be carried out locally. They will optimize the global model and send the gradient update to the central node for aggregation. Secure parameter aggregation, involving differential privacy (DP) and MPC, protects against the privacy leakage resulting from the loss of the model. [86] gives an overview of the privacy-preserving DML techniques in this area.

The next section in this chapter provides a characterization of two important scenarios in which data are used, “user data sharing” and “data set sharing”. These scenarios are chosen because they illustrate the types of privacy risks that can be addressed by data modification. After describing the scenarios, we then provide threat models that capture the nature of the privacy risks. Then, we give an overview of the state of the art of data modification techniques that have been proposed to enhance privacy. The techniques have two distinct relations to machine learning: first, machine learning can be used to create data modification, and, second, the modified data can be used by machine learning algorithms.

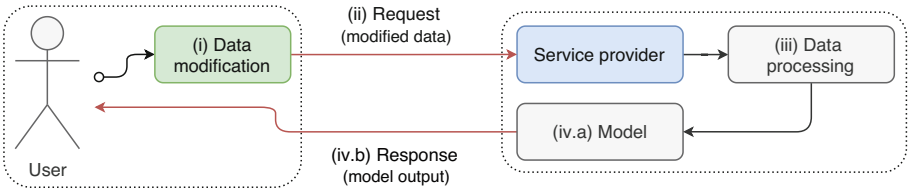

Fig. 1. Scenario 1: user data sharing.

These relations are highlighted in our overview. The chapter finishes with a discussion of challenges and open directions.

2 Scenarios and Requirements

Our data sharing scenarios are inspired by the growing interest in cloud computing technologies [15]. As data becomes cheap, both for centralized entities to collect and edge devices to share, novel business models are popping up that take advantage of this abundance. Considering this relationship between the user (at the edge) and the service provider (at the centre), we obtain abstractions of two scenarios for sharing data. Making reference to these two sharing scenarios allows us to focus on the specific challenges posed by the collection of data for ML pipelines in the context of existing regulatory frameworks [44], and the potential for pre-processing data prior to their release.

In the first scenario, a user shares sensitive data with a service provider (SP) to train a ML model, receive a prediction, or carry out an analysis. In the second scenario, a data-collector has received sensitive data and wants to enable third-parties to perform data analysis. We cannot assume that the channel by which data is shared is reliable. These scenarios allow for a more in-depth discussion on the threats and defenses covered throughout the chapter.

2.1 Scenario 1: User Data Sharing

In this scenario, the user sharing the data is the person who produced the data. The user shares the data in order to receive a certain output, but at the same time does not want the data to be used for a purpose that they do not approve of. The sharing of data serves to feed a ML model or to perform statistical data analysis. This scenario is important due to the rise of cloud-based computer vision APIs, which make it possible for any business/user to build a state-of-the-art model merely by sharing data [2]. This scenario allows a user to benefit from the model while controlling privacy risks. For example, a user can share an image on a social media platform, and agree that platform analyzes that image for the purpose of producing recommendations, but not agree that the platform uses it for other purposes, e.g., training a facial recognition system.

The ML pipeline that we identify as a final goal of this scenario is conventionally divided into modules, which are shown in Fig. 1. Data are first collected and pre-processed by the users themselves, then *Data modification* is performed. Performing modification right after collection is important for several reasons, including cleaning data and minimising storing and network requirements. Technically, modification could also be performed after sharing. However, we focus on applications that apply modification before sharing for privacy-preserving purposes. Next, a *Data processing* module carries out pre-processing and, if needed, extracts relevant features from the data. Feature extraction might be directly integrated into the machine learner, or be carried out as a separate step. For example, the machine learner might be a classifier that uses the data to learn how to label images of animals as *cat* or *dog*. The phase in which data is presented to the classifier for the purpose of learning is called "training". The phase in which a new, yet-unseen data sample is presented to the classifier to obtain a label and/or score is called "inference".

Combing the ML pipeline just described with the data sharing steps, we arrive at a scenario that describes how a *User* and a *Service provider (SP)* interact in a data-value exchange protocol. We divide the procedure in four main steps illustrated in Fig. 1:

i Data are generated/captured and processed to obtain a sample.
ii The sample is shared with the SP, typicality through an unreliable channel (the internet).
iii The sample is further (optionally) processed and fed to the ML model.
iv The SP returns an answer to the user.

In general, the two communicating parties have competing needs: one the one hand, the user wants to be protected; on the other hand, the SP aims to maximize utility. In line with the data quality principles of the GDPR (Art. 5), we define the following requirements:

1. Data confidentiality: by protecting data, we minimize the risks of sharing sensitive information with, generally untrustworthy, third-parties.
2. Data minimization: only data that is needed for the primary learning task is sent via the communication channel.
3. Purpose limitation: data collection and data processing are limited to used in a clearly defined ML task.
4. Usefulness: the ML process preserves the primary utility of the service and the value for the users.

2.2 Scenario 2: Data Set sharing

In this scenario, a pool of people, i.e., the *Data subjects*, have already shared sensitive data with a central node, i.e., the *Data collector*. This might be a hospital who collects the digital clinical diary of their patients [28]. The central node is trusted to be the only entity that is allowed to manage sensitive information,

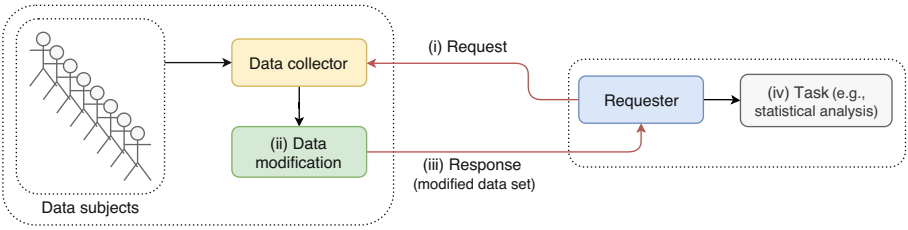

Fig. 2. Scenario 2: Data set sharing.

hence no other party is granted direct access permissions. An external entity, i.e., the *Requester*, performs an access request to obtain a modified version of the data set. The Requester might be a benign organization such as a research institute that aims to performs in-depth analysis or train a ML model.

The interaction between a requester and a data collector can be divided into four steps (Fig. 2):

i The *Requester* forwards a request to the data collector.
ii The data set is processed by the *Data collector*.
iii The modified version of the data set is shared to the *Requester*.
iv The *Requester* performs some defined tasks, e.g., statistical analysis, inference on the received data set.

In this scenario, the data collector manages highly sensitive data that must not be leaked to untrusted parties. Nonetheless, benign requesters can greatly benefit from the sharing of this asset. The main requirements that arise from this scenario include:

1. Data confidentiality: the original data can only be accessed by the *Data collector*.
2. Data privacy: the released data set has to remain anonymous for the *Requestor* and/or suppress sensitive attributes.
3. Usefulness: the usefulness of the released data set is preserved for the inferential task carried out by the *Requester*.

We can observe that two scenarios are similar in that they represent a relationship between a sharer and a receiver. However, they differ with respect to the information that is sent through the unreliable communication channel. In Scenario 1, the data are relevant to an individual, and in Scenario 2, they are relevant for an entire group. Additionally, in Scenario 1, the data is shared with a particular model with a particular function as the target. In Scenario 2, the use of the data is not necessarily limited to a single purpose. Keeping these differences in mind will make it easier to understand the structure of the landscape of threats and countermeasures.

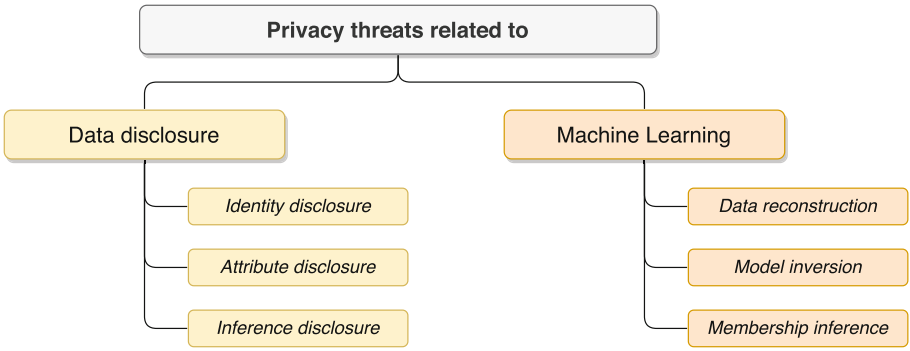

Fig. 3. Privacy threats specific to the contexts of data sharing and machine learning.

3 Threat Model

We define our threat actors based on a standard scheme used in the literature [70]. An attacker has a clear *goal* to carry out, the extent to which the goal is pursued, however, depends on the attacker's *resources*. The resources might include prior knowledge about the target as well as technological and economic resources. The resources deployed are dependent on the *risks* an attacker is willing to take and *countermeasures* used by the target. Based on this scheme, in this section, we model adversaries by presenting threat models that are related to our scenarios. The presentation of the threat models is followed by a description of the main privacy threats linked to our malicious users (Fig. 3).

3.1 Scenario 1 Threat Model

In Scenario 1, "User data sharing", we have two interacting entities: the *Service provider* (SP) and the *User*. Both stakeholders can act maliciously in order to maximize their value at the expense of other actors. Hence, we describe two malicious actors that are directly derived from the previous ones: the *honest-but-curious service provider*, and the *malicious user*.

The honest-but-curious service provider is the coordinator of the communication, i.e., the SP of Fig. 1. Its primary intent is to run the service smoothly while behaving honestly by following the protocol as expected. Accordingly, it receives requests (Fig. 1(ii)) and produces rightful answers (Fig. 1(iv.b)). However, it is *curious* in the sense that it aims at gathering as much information as possible about its users. Specifically, it can modify step (iii) of Fig. 1 to gain out-of-context knowledge that the *User* does not intend the SP to gain. As the SP can be a big corporation, we assume they have the power and resources to obtain additional knowledge about their customers, train large networks, and coordinate an attack between many subsidiaries if needed. In this context, data re-purposing poses a severe threat to the user. The SP performs a task beyond the original one agreed

with the users and, potentially, sells these data to third-parties without explicit consent.

The malicious user follows the protocol but might try to get additional information about other users. Malicious users have black-box access to the ML model trained by the SP: they can query the system (Fig. 2(ii)) and observe the output (Fig. 2(iv.b)), but no access is given to the internal parameters or data used to train the model. An possible extension of the malicious user is a group of users acting in a coordinated fashion (e.g., other companies) that makes use of their greater computation capabilities to carry out larger campaigns. These capabilities can be used, for example, to learn sensitive information about data subjects, which leads to discriminative and abusive behaviours, or to steal the intellectual property of the SP (e.g., the ML model internals).

3.2 Scenario 2 Threat Model

Like Scenario 1, Scenario 2, "Data set sharing", is characterized by a two-party interactive exchange. The two main attackers, as before, are adversarial versions of the attackers participating in the exchange: the *malicious requester* and the *honest-but-curious data collector*.

The malicious requester aims to de-anonymize, de-obfuscate or carry out unwanted inference on the received data set. As a requester, it requests access to a data set (Fig. 2(i)) acting as a trustworthy party and obtains a modified version of the target data set (Fig. 2(iii)). At this point, it carries out the inference step (Fig. 2(iv)) targeting users' private attributes, having full access to the original data set. In the worst case, the malicious requester can be a powerful entity capable of obtaining further data about the victims – like the powerful malicious user of Scenario 1 – which enhances its knowledge about the target data distribution. Anonymization techniques could protect users' privacy in this setting. Unfortunately, data are easy to de-anonymize by harnessing correlations between variables and linking different data sources [60].

The honest-but-curious data collector is a *Data collector* (Fig. 2) that works on behalf of its data subjects. It may attempt to infer sensitive information about its customers based on the collected data combined with additional background knowledge. This actor partially overlaps with the honest-but-curious SP of Scenario 1 in that it behaves honestly but leverages its position to maximize its profits, e.g., by collecting and/or retaining data more than necessary. In order to avoid redundancy, we consider the data collector as a trustworthy entity that is required to protect the data before being released to third-parties (i.e., the *requester*).

3.3 Privacy Threats in the Context of ML

Next, we discuss the privacy threats that are related directly to the machine learning model (Fig. 3, right side). These can be understood as mainly related to Scenario 1, "User data sharing", in which we are concerned about the path traveled by the data of an individual user. These threats are related to data modification because data modification can be used to counter them. The gravity of the threat depends on the sensitivity of the data and the threat model of the attacker. We mention the following set of threats [8,19]:

Data Reconstruction [7,8,72]. Several methods aim at reconstructing private data from a processed version of the data. For example, trying to reconstruct private information after the original data has been transformed into feature vectors. In data reconstruction, the attacker can exploit various levels of knowledge. A malicious user is constrained by the output returned by the SP and the information they know about other users. In contrast, an SP attempting to reconstruct data is constrained by the modification performed locally by the user, but has access to the internals of their own models. Assuming that data samples undergo a modification, the knowledge about the procedure represents an additional tool. The trade-off between the amount of distortion is the data and the usefulness that is preserved determines the amount of protection for both the user and the SP.

Model Inversion [27]. A ML model can undercover statistical correlations between publicly known variables and sensitive attributes. Model inversion allows malicious users to query the system in order to infer sensitive attributes about a target user if they know something about the target user from another source. For example, given a picture of the target, which is usually available on social media, the attacker (a malicious user) requests the SP to output the probability that the target has of developing skin cancer. At the same time, the SP can abuse its power to sell sensitive data or the access to the model to third-parties interested in these valuable information.

Membership Inference [10,26,73]. The attacker seeks to infer whether an individual was a member of the training set used to build the ML model. In Scenario 1, the membership inference threat is the threat that malicious user aims to learn whether a target user's data was used in building SP's model. The malicious user queries the system and uses the received output to carry out an inferential analysis.

3.4 Privacy Threats in the Context of Data Sharing

Now, we move to discuss the privacy threats that are related directly to data disclosure (Fig. 3, left side). Again, these threats are related to data modification because data modification can be use to counter them. In the context of data set publishing, which falls under Scenario 2, a major threat is represented

by the *disclosure* of sensitive information. However, data disclosure is also a threat for the user in Scenario 1. Generally, we distinguish three major types of disclosure [77,80]:

Identity Disclosure. Identity disclosure occurs when a malicious requester successfully links their target with data obtained from the Data collector. The linkage can be made using a small set of variables. If successful, the adversary has access to sensitive attributes in the shared data. An example of identity disclosure was the case of the Netflix Prize competition, where an anonymized data set was released where each record was a tuple containing an anonymous user ID, a movie, the rating given by the user to the movie and the date of the grade [12]. Using the Internet Movie Database as the source of background knowledge, this data set was successfully de-anonymized and Netflix records of known users were identified [60].

Attribute Disclosure. Attribute disclosure if some key variables about the user are already known, and a malicious requester is able to infer additional characteristics (attributes) of the targets from a data set leveraging these variables. In [77], authors used an example where every person with "race = black", "aged 50–60", "living in region ZIP = 1234" in the data set has the same sensitive variable "religion = roman/catholic". Therefore, if the adversary knows that an individual has the characteristics "race = black", "aged 50–60" and "ZIP = 1234", the sensitive variable "religion" is easily inferred.

Inference Disclosure. Sensitive information disclosure occurs when an malicious requester is able to determine characteristics of the target more accurately by making use of the released data [77] and the process of inference. With inference disclosure, individuals are threatened not merely due to the information in their records, but by statistical properties of the entire database [32]. An example would be when you are one of the two richest people in a country. The aggregated information on a survey regarding the income of everyone in the country has been released. You can now easily estimate the wealth of the other rich person by using the published information.

4 Overview of Data Modification Techniques

In this section, we provide an overview of the relevant techniques relevant to the scenarios in Sect. 2 and of use for countering the privacy threats presented in Sect. 3. We provide a categorization in Fig. 4, which summarizes the approaches covered throughout the section and indicates their relation to the scenarios. Many techniques are applicable to both reference scenarios some are more closely connected to Scenario 2, "Data set sharing".

Techniques to pre-process data before sharing can be divided into non-perturbative, perturbative, and synthetic data generation. Non-perturbative techniques achieve privacy protection by applying masks or generalizing given

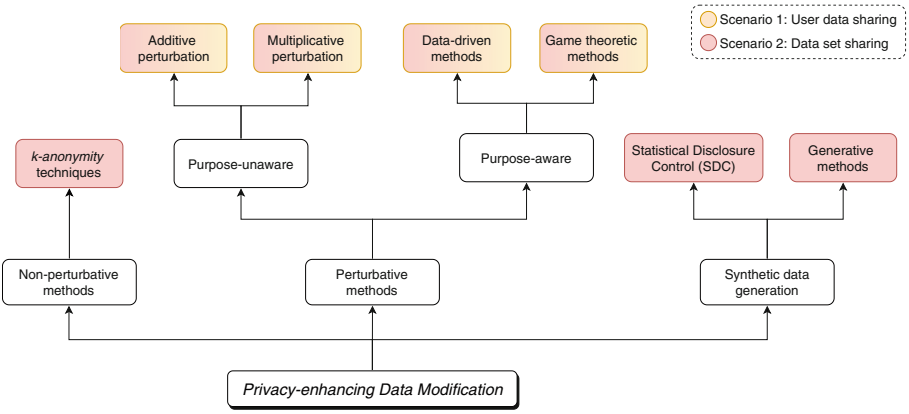

Fig. 4. Taxonomy of the landscape of privacy-enhancing data modification techniques. The boxes designating the methods are shaded to link them to our two scenarios.

attributes. Perturbative methods apply a transformation as a means to hide/obfuscate the sensitive attributes. A further division separates perturbative methods into techniques that apply indiscriminate noise, i.e., purpose-unaware, and techniques that incorporate some knowledge about what the data is to be used for, i.e., purpose-aware. Lastly, it is possible to generate synthetic data that preserve statistical properties of the target distribution while protecting the privacy of users whose data was in the original data set.

There are further differences related to the techniques that we present here that depend on the kind of data that is given as input. In the rest of the chapter, we differentiate between techniques appropriate for structured and for unstructured data. A *structured data set* corresponds to a set of records (rows) composed of well-defined attributes (columns). This data typically resides in a relational database. By contrast, an *unstructured data set* is not organized in a way that directly encodes meaningful relationships between data points. Unstructured data includes many forms of text, images, and audio. A *semi-structured data set* falls between the two. It defines hierarchies/groups of samples without being fundamentally tabular in nature, essentially adding semantic categorization on top of an unstructured data set.

4.1 Non-perturbative Techniques

Non-perturbative techniques adjust the original data so that it is less precise, but do not otherwise change the data [21]. Traditional indistinguishability approaches modify the data so as to prevent identification of individuals within a data set. By generalizing or suppressing specific attributes within a data set, we can achieve properties such as k-anonymization, t-closeness and l-diversity.

However, these approaches suffer from several shortcomings: First, they are not always suitable for releasing large data sets since they may allow the inference

of sensitive attributes on the basis of attributes that are publicly available [12]. Second, despite being applied to unstructured or semi-structured data – such as face images [34] – these techniques are mostly confined to structured data set release. Third, the privacy protection is bounded to the number of attributes present in the data set and their uniqueness.

In the context of images and videos, a non-perturbative technique that has been improving over the years is K-same [35,53,66]. K-same (in addition to pixelation or blurring) aims to obfuscate some parts in the images. It protects the privacy of individuals by de-identifying faces such that some facial appearances remain but the face cannot be recognized [61]. The basis of privacy protection of K-same is a non-perturbative k-anonymity algorithm proposed by Sweeney et al. in [76].

K-Same-Net [53] and K-Same-Siamese-GAN [66] are different privacy protection amelioration of K-same against face de-identification. K-Same-Net is a combination of recent generative neural networks (GNN) with k-anonymity mechanisms. It generates synthetic surrogate face images by combining the characteristics of the identities used to form the model. K-Same-Siamese-GAN combines the power of K-same anonymity mechanism with generative adversarial network and hyperparameter tuning. We can consider K-Same-Siamese-GAN to combine non-perturbative methods and synthetic data generation.

4.2 Pertubative Techniques

Perturbative techniques introduce distortions into the data [21] and can be either purpose-unaware or purpose-aware. Purpose-unaware techniques aim to modify the data in a way that contributes to protecting privacy, while at the same time maintaining the usefulness of the data for general purposes. Purpose-aware techniques make use of advance knowledge of the function that the modified data is intended to serve, and modify the data the data in a way that maintains the usefulness of the data for that function.

Purpose-Unaware Techniques. Among traditional pertubative techniques, data swapping and rank swapping exchange confidential attributes between different records [21], data shuffling shuffles the values of the confidential variables among observations [59]. The perturbations are constrained such that the usefulness of the data is maintained. However, while these techniques are sufficient for simple, structured, data sets, they are not suited to address the limitations of large collections of unstructured data. In the following, we delve into recent work on additive and multiplicative perturbation as purpose-unaware techniques.

In the domain of *additive perturbation*, differential privacy (DP) is the de-facto standard for anonymization and attribute hiding [20]. DP, in its most basic form [25], defines formal privacy guarantees that a set of algorithms usually implement via noise addition. These are guarantees on the amount of sensitive information leaked by publishing two 'close' data sets. The privacy loss is measured by ϵ. DP has several advantages w.r.t. previous techniques. First, it

makes it possible to quantify the privacy loss via ϵ and tune the utility-privacy trade-off accordingly. Second, it models a worst-case adversary whose aim is to learn a target variable of the target user. Third, because it is not property of the data set, like *k-anonymity*, but rather a property of the process [20], DP can be combined with several techniques and can be applied at different stages. In its *local* configuration, for example, DP permits the local processing of data before sharing with an untrusted party. Random noise addition, however, is a double-edged sword. It protects against reconstruction attacks but does not offer the possibility to balance privacy and usefulness in a satisfactory manner.

Within the ML landscape, DP can be applied at different stages of the pipeline: beyond the protection of input and output, several techniques target the training of a model. The goal is to train a model on sensitive data while guaranteeing DP. Abadi et al. [3] introduced a variation on the stochastic gradient descent (SGD) algorithm, commonly used to train deep neural networks. In particular, they propose to modify the gradient computation by clipping and adding noise. This method is also referred to as *Moments Accountant* since its formal guarantees originate from privacy loss being accounted for at each step of the training procedure. This randomization can be moved to users' devices, as proposed by Arachchige et al. [9]. We can achieve DP training of a deep learning model without trusting a central node, i.e., no sensitive data leave the device.

A second form of noise addition is based on random projections, i.e., *multiplicative perturbation*. Multi-dimensional projections can be used on structured input to preserve the distance between the samples in a lower-dimensional space [6]. This technique makes it possible to run analytics as well as train a regression or classification model on the modified data. Differently from DP-like techniques, projections are prone to reconstruction attacks and sensitive leakage. Recently, Jiang et al. [42] proposed a method to apply individual Gaussian random projections locally that also protect against common attacks. In contrast to previous techniques, they harness the capabilities of deep learning to learn complex patterns and find a projection that better suits semi-structured or unstructured data.

Purpose-Aware Techniques. Next we turn to discuss perturbative techniques that protect data, while maintaining usefulness for a particular function. This function (i.e., "purpose") is known before data modification is applied, and the process used to modify the data is specifically designed so that the modified data can still fulfill this function.

We start with an example that is well suited to illustrate the basic principle of purpose-aware techniques. The example is drawn from the area of recommender systems and is a close fit with Scenario 2 "Data set sharing". In this example, a company (acting as a *Data collector*) shares data with an external researcher (acting as a *Requester*), who is carrying out recommender system research (the *Task*). In this case, the data consists of a so-called user-item matrix in which each row corresponds to a user and contains information on the interactions that the user has had with a set of items, corresponding to the columns. Slokom

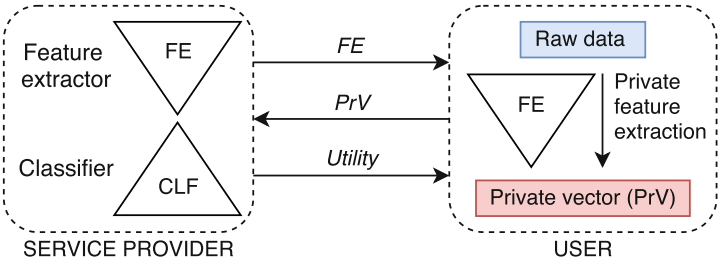

Fig. 5. The split between feature extraction and training in a user data sharing scenario [65]. The feature extractor can be trained with the utility of the classifier, i.e., purpose-aware, or being decoupled from the classifier, i.e., purpose-unaware, depending on the task at hand.

et al. [74] introduced a data masking approach called *Shuffling Non-Nearest-Neighbors* (Shuffle-NNN) that modifies the data so that it no longer contains precise information about which user has interacted with which item. At the same time, Shuffle-NNN aims to maintain the usefulness of the data for the purpose of training and testing recommender systems, which is necessary to carry out research. Shuffle-NNN generates a masked data set by changing a large portion of values of the preferences in a user's profile. Shuffle-NNN can be considered a Data-driven method used to create purpose-aware data modification, because it uses patterns in the data in order to decide which changes to make. Specifically, Shuffle-NNN aims to preserve item-item similarity information, based on the assumption that this information is the most important pattern that needs to be present in the data in order to train and test a recommender system algorithm and move forward recommender system research. Shuffle-NNN applies a data shuffling technique hides (i.e., changes) preferences of users for individual items. Shuffle-NNN occurs in two steps: neighborhood selection and value swapping. Neighborhood selection determines the neighborhoods of every item based on the K-nearest neighbor algorithm and then joins these neighborhoods in order to find a set of critical items. All items not in this set are considered "non-nearest neighbors" and are shuffled. The protection level is judged by the number of ratings that were hid.

Other purpose-aware approaches differ from this basic example along a number of different dimensions. First, they can modify feature representations extracted from the data, rather than the original data itself. Second, they can use machine learning to determine how to modify the data and/or how to preserve patterns in the data needed to maintain usefulness. Third, they can seek to provide privacy guarantees, whereby it is important to keep in mind that whether or not a guarantee holds depends on the threat model. We will now go on to cover series of more sophisticated purpose-aware techniques that exhibit these various dimensions.

Among approaches that use machine learning, techniques based on multi-objective optimization are important. Here, we discuss two examples of work that has investigated the possibility to perform data-driven data modification at the edge. Liu et al. [49] combine deep networks and noise to obtain the best trade-off between privacy and utility. They find a candidate subset of features for noise addition by harnessing a deep auto-encoder: an architecture comprising a compressing part (i.e., the encoder) and a reconstructing part (i.e., the decoder). On top of it, DP-noise is added to obtain measurable privacy guarantees.

Osia et al. [65] propose a hybrid framework in which a feature extractor is trained by a provider with privacy guarantees and shared with the user (cf. Fig. 5). The user can then derive a private representation that is shared with the provider, hence moving the privacy-preserving modification to the edge. The extracted features, or *private vector*, is designed to only contain relevant information for the primary learning task, thus adaptable on the learning task and the given privacy objective. The framework [65] makes use of a fine-tuning strategy. A cloud provider fine-tunes a pre-trained model with two objectives: the primary classification loss function and a contrastive loss. While the first term accounts for the utility of the process, the second one is directly applied to an intermediate layer, such that two samples with the same label end up being close in the feature space and two samples with different labels are separated as much as possible. After training, the classification block is discarded and the feature extractor is shared with the users. In this specific case, data are not shared by the user to perform training, rather for carrying out inference on a trained model.

Also among approaches using machine learning, an increasing among of work makes use of generative adversarial networks [33] (GANs). A GAN models a minmax game between a generator G and a discriminator D: while G is being trained to approximate a target data distribution, D tries to distinguish between a real sample and a generated one. Because a GAN realizes a minmax game, we refer to GAN-based data modification approaches as "Game theoretic methods". Normally, G and D are implemented using two deep neural networks that are trained *adversarially*. The output is a generator that (1) provides realistic data samples and (2) is able to deceive the discriminator. Adversarial learning overcomes the difficulty of modelling an underlying distribution. For this reason, it can be particularly useful when we cannot formally define our privacy objective, because it provides a data-driven way to characterize the private and the target variable distributions.

An ambitious line of work has considered the possibility of bridging the gap between generative networks, adversarial learning, and privacy guarantees. Huang et al. first proposed GAP [38]: a framework to achieve an optimal privacy mechanism inspired by GANs. Here, the generator becomes a *privatizer* that protects against attribute leakage, and the discriminator becomes the *adversary* competing with the latter by trying to infer the protected attribute. The learning strategy is defined as a constrained minmax optimization process that infers the distribution from the data set. This greatly improves the practicality of the approach compared to information-theoretic strategies based on Mutual Infor-

mation (MI) – a measure of the dependence between two random variables – that is often deemed as an intractable problem.

The seminal work on GAN-based methods had been adapted to different domains in recent years. Different data distributions and different requirements on the private attributes that must be protected require different approaches, both from an architectural and an optimization perspective. Biometric data represent a tough challenge. Applications using biometric data require that the modified data is useful for a specify purpose, e.g., identifying a user. At the same time, sensitive attributes must be protected. The challenge arises because cues of identity and cues related to sensitive attributes are tightly tangled in the data. The information that must be maintained in the data, and the information that must be protected differs from use-case to use case, but the challenge arising from entanglement remains.

In the area of biometric data, some work has tackled the data anonymization problem [30,40,47,51,58,67] and other focused more on the selective compression of data to retain pre-determined attributes [16,31,56,63]. However, techniques working in image domain present substantial differences w.r.t. techniques applied to motion data from inertial sensors. Ren at al. [67] introduce a model trained with an adversarial regularizer and an action recognition network. This data-driven strategy aims at finding the right perturbation that preserves action recognition performance in videos. Li et al. [47] take a different approach by using a conditional generative networks (CGANs). The face is first identified and blurred. A CGAN then generates a new face image by fixing key features – such as the head pose. A similar approach based on GANs and swapping is presented in [40] where the head pose and the background are the only preserved attributes. Beyond identity obfuscation (or anonymization), Chhabra et al. [16] tackle the soft biometric privacy problem. Their proposed algorithm searches for a sub-optimal perturbation that preserves one attributes but hides multiple sensitive attributes. Similarly, PrivacyNet [57] uses a GAN-like training procedure to achieve controllable privacy w.r.t. several sensitive attributes – such as gender and age. In the context of motion sensor data, a few approaches have been proposed that tackle the anonymization of the input trace [51] or the selective hiding of private attributes [31].

Another application domain is online image sharing. Here, the goal is to maintain the usefulness of the images from the point of view of people looking at the images. Images should retain their quality after data modification. At the same time, modified images should offer privacy protection. Oh et al. [62] investigate person recognition, and propose a framework formulated as a game between a social media user and a recognizer. The user attempts to perturbs the image to protect the identity of the person it depicts and the recognizer attempts to break the protection using a countermeasure. Larson et al., [45] formulate a benchmarking task to encourage work on techniques that protect sensitive information in images going beyond faces in people, starting with protecting sensitive scene information in images. Whereas in [62], the assumption is made that the adversarial techniques will preserve the quality of the images, in [45],

preserving image usefulness for sharing is specified explicitly as a goal of the data modification.

Zhao et al. [85] is an example of an approach that maintains quality and can be used to protect against unwanted inference of a classifier. Alternation between enlarging perturbations informed by the classification loss and minimizing perturbations informed by perceptual color distance is shown to result in efficient and effective adversarial examples. Shan et al. [71] propose Fawkes, which applies a cloak to images to protect users against unwanted face recognition. Fawkes attempts to move the latent representation of a user towards a second user. More work is needed on broadening the threat model under which such approaches offer protection, especially to include countermeasures deployed by the attacker. More information on adversarial examples can be found in [87].

Our main emphasis is on techniques that enhance privacy by striving to limit the information that can be derived from modified data. For completeness we mention another goal, namely, protecting data from being used in an unwanted fashion. Huang et al. [39] proposed a method for making user data unusable for training machine learning models. Whereas standard strategies seek to maximize error inducing noise, [39] pursues the strategy of finding small noise that minimizes the model's error via a min-min optimization process.

4.3 Synthetic Data Generation

Synthetic data preserves specific statistical properties or relationships between attributes in the original space, without exposing users. Synthetic data generation methods work by first constructing a model of the target data distribution and then generating synthetic surrogates. In this section, we discuss synthetic data generation going from statistical disclosure control [21,79] to deep learning [4,81,83].

Synthetic data is first proposed for the Statistical Disclosure Control (SDC), or inference control methods. SDC seeks to protect the users' data from being disclosed/linked to a specific user [22,41]. The main purpose of SDC is to release protected data to minimize *disclosure risk*, i.e., the risk that a malicious user uses data to determine sensitive variables of a victim user. To retain data utility, the statistical analysis on protected data and original data must yield similar results [41]. Synthetic data generation is one of the methods which can be used for SDC. Several approaches have been proposed in the literature for generating synthetic data for SDC, such as data distortion by probability distribution [48], synthetic data by multiple imputation [68] and synthetic data by Latin Hypercube Sampling [18]. Recent techniques for generating synthetic data fall into three basic categories [22,24]: fully synthetic, partially synthetic and hybrid techniques.

Fully synthetic data sets keep the original data private since they are obtained as a replacement set created entirely anew. [22]. We note the disclosure risk for fully synthetic data sets is low, as all values are synthetic. Differently, partially synthetic data sets contain a mix of original and synthetic values [22]. Techniques to achieve partial synthesis replace only observed values for variables that bear

a high risk of disclosure (i.e., key variables) [23]. The disclosure risk for partially synthetic data sets is higher than for fully synthetic data sets, since some true values remain in the data set. The disclosure risk significantly increases if the adversary knows which records are present in the data. However, partially synthetic data sets typically have a higher data utility compared to the fully synthetic data sets. Third, hybrid masking techniques generate masked data as a combination of original and synthetic data sets [18]. The value in the original data set is linearly matched with the value in the synthetic data set and are then added together or multiplied to create the published value [84]. This combination allows for better control over individual characteristics [18].

The difference between partially synthetic data sets and hybrid masking is the following: with partially synthetic data sets, an individual variable is either replaced by a synthetic record or the record is kept original, while in hybrid masking the values in each record are added or multiplied with the corresponding value in the synthetic data set.

The domain of synthetic data generation has been evolving over the years. A more recent line of research focuses on deep learning-based *synthetic data generation*. The generated data retains the same statistical properties as the original data while being private for the users. In [4], Abay et al. propose a generative deep learning technique that produces synthetic data from an original data while preserving the utility. An auto-encoder is used to partition the original data into groups. For each group, they build a private generative auto-encoder called *DP-SYN*. The auto-encoder first learns the latent representation for each group, and then uses the expectation maximization algorithm to simulate them. In [46], a variational auto-encoder (VAE) is used as a generative model. The first step is to feed the encoder with the original data and the model outputs a reconstructed data. The second step is to feed the decoder with Gaussian random data. Then, it generates new data from the Gaussian distribution. They showed that VAE succeeds to generate an artificial data that closely mimics the original data while maintaining good accuracy. In addition to auto-encoders and variational auto-encoders, generative adversarial network [33] (GAN) has been widely used for generating synthetic data [5,17,50,81,83]. As discussed above, GANs are composed of two networks: a generator and a discriminator. The generator attempts to produce a realistic looking data based on the learned data distribution and the discriminator seeks to differentiate between the real data from the original data and the synthetic data from the generator. Bindschaedler et al. [13] propose a new approach for releasing privacy preserving synthetic through plausible deniability data while maintaining statistical properties of the data. It is based on the fact that there are at least k ($k > 0$) input records that could have generated the observed output with similar probability. Plausible deniability has two main steps [13]: First, the *generative step* consists of constructing a utility preserving data model. Second, the *privacy test step* aims to protect the privacy of users whose data records are in the input data set. It ensures that every released output can be plausibly deniable.

5 Summary and Future Directions

In this chapter, we have provided an overview of data modification for privacy enhancement based on two main scenarios: user data sharing (Sect. 2.1) and data set sharing (Sect. 2.2). We discussed the related threat models, which describe risks and sources of privacy leakages. We then provided an overview of different approaches for privacy-enhancing data modification (Sect. 4). In the following, we point out important discussions and sketch directions for future work.

5.1 New Types of Data

Data are at the center of research on approaches to privacy enhancement. While structured data lend themselves to the use of traditional techniques based on k-anonymity or DP, semi-structured and unstructured data need a different set of approaches. Moving to new types of data requires careful attention to both the potential and the challenges that are related to the use of machine learning for data modification. Specifically, approaches that extract privacy representations are promising (cf. Sect. 4), but present a future challenge since machine learning research does not study feature extraction to the same depth across different types of data. Extracting features using a neural network is common for face images. However, for sensor data for activity recognition or fingerprint authentication it is generally necessary to rely on manual feature engineering for feature extraction [69]. In some contexts, static features provide a level of interpretability that is currently lacking when using dynamic features. This is important when outsourcing the feature extraction process (Fig. 5), since dynamic features can introduce a new attack surface. For example, it is non-trivial to define the relationship between 128 features extracted from a deep learning model for face recognition as it is to exactly define 64 statistical attributes derived from a motion trace. As a consequence, hand-crafted feature engineering aid the sanitation of data – by imposing constraints on the validity of the features – and can leverage human intelligence in a human-in-the-loop learning process [37]. Nonetheless, deep learning has been demonstrated to be a great ally in solving problems in which traditional ML falls short due to unstructured and complicated data [55], and model explanations can provide an adversary with additional information that hinder the privacy-preserving mechanisms [72].

5.2 Privacy and Fairness

Harmful social bias in machine learning can originate from data sets, algorithms, and processes. Recently, increasing amounts of research have been devoted to the analyses of discrimination and the embedding of fairness into the automatic decision making process [54]. In many cases, the attributes that are deemed sensitive for the user are the ones which drive the unwanted discrimination. Suppressing them, however, is not enough to obtain a fair representation. As with *sensitive attribute disclosure*, the correlation among variables retains the source of bias within our target data distribution, even after attribute suppression.

There is a substantial overlap between work investigating algorithmic fairness and private data modification. Beyond the overlap of sensitive attributes, techniques are applied at similar steps of the pipeline: prior to feeding the algorithm (pre-processing) [14,43], during processing [14] and post-processing [54]. Often, the obfuscation targets the membership in a *protected* group while preserving the utility, which can be modelled as the minmax optimization process seen with purpose-aware data modification. Recent efforts towards the realization of a framework for controllable and measurable fairness include open source libraries that implement the proposed techniques [11].

5.3 Interdisciplinarity

Currently, the domain of privacy-preserving techniques is fragmented across different research communities. Machine learning researchers might approach the problem from a learning perspective, focusing on the model and its optimization. By contrast, the privacy community relies on well-established formal definitions and thoroughly studied solutions. Privacy is a broad field encompassing objective metrics and legal requirements that are often either detached or incompatible [82]. This separation widens further if we consider that the formal techniques are applied in a multitude of systems and that specific domains, e.g., image classification vs. recommender systems, require domain-targeted solutions.

In order to advance research on data modification for privacy enhancement, it is important to bring different disciplines together. Examples of successful collaboration in related areas includes bridging the gap between science and society [64], ethics and big data [29,36], privacy and data quality [1,75]. Here we mention two reasons why we find interdiscplinary approaches to be particularly important. First, machine learning technology is developing rapidly. As a result, the ways in which machine learning meets data modification are constantly changing. Machine learning experts and privacy experts must work together to identify how data modification can address new threats of machine learning and also how machine learning can enable new methods for data modification. Second, data modification is often well-suited for general privacy enhancement, but not for well-defined guarantees of privacy protection in real-world use scenarios. Tackling this challenge will require development of threat models that help to define where data modification could be most helpful, and how it could be combined with other approaches. A benefit of data modification is that it can be used in a decentralized way, in other words, applied at the edge, i.e., on a user's personal device before the user shares the data. Such scenarios must also be incorporated into threat models. Research dedicated to developing the threat models must involve experts in machine learning, distributed systems, human factors, privacy, and the law.

Acknowledgements. This research is partially funded by the Research Fund KU Leuven, and by the Flemish Research Programme Cybersecurity.

References

1. This thing called fairness: disciplinary confusion realizing a value in technology. Proc. ACM Hum.-Comput. Interact. **3**(CSCW), 1–36 (2019)
2. Amazon Rekognition: Automate your image and video analysis with machine learning. (2020). https://aws.amazon.com/rekognition/. Accessed 07 Feb 2021
3. Abadi, M., et al.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 308–318 (2016)
4. Abay, N.C., Zhou, Y., Kantarcioglu, M., Thuraisingham, B., Sweeney, L.: Privacy preserving synthetic data release using deep learning. In: Berlingerio, M., Bonchi, F., Gärtner, T., Hurley, N., Ifrim, G. (eds.) ECML PKDD 2018. LNCS (LNAI), vol. 11051, pp. 510–526. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-10925-7_31
5. Acs, G., Melis, L., Castelluccia, C., De Cristofaro, E.: Differentially private mixture of generative neural networks. IEEE Trans. Knowl. Data Eng. **31**(6), 1109–1121 (2018)
6. Aggarwal, C.C., Philip, S.Y.: A survey of randomization methods for privacy-preserving data mining. In: Aggarwal, C.C., Yu, P.S. (eds.) Privacy-Preserving Data Mining. Advances in Database Systems, vol. 34, pp. 137–156. Springer, Boston (2008). https://doi.org/10.1007/978-0-387-70992-5_6
7. Al-Rubaie, M., Chang, J.M.: Reconstruction attacks against mobile-based continuous authentication systems in the cloud. IEEE Trans. Inf. Forensics Secur. **11**(12), 2648–2663 (2016)
8. Al-Rubaie, M., Chang, J.M.: Privacy-preserving machine learning: threats and solutions. IEEE Secur. Priv. **17**(2), 49–58 (2019)
9. Arachchige, P.C.M., Bertok, P., Khalil, I., Liu, D., Camtepe, S., Atiquzzaman, M.: Local differential privacy for deep learning. IEEE Internet Things J. **7**(7), 5827–5842 (2019)
10. Backes, M., Berrang, P., Humbert, M., Manoharan, P.: Membership privacy in microRNA-based studies. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, pp. 319–330 (2016)
11. Bellamy, R.K., et al.: AI fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:1810.01943 (2018)
12. Bennett, J., Lanning, S., et al.: The Netflix prize. In: Proceedings of the Annual Knowledge Discovery and Data Mining Cup and Workshop, p. 35 (2007)
13. Bindschaedler, V., Shokri, R., Gunter, C.A.: Plausible deniability for privacy-preserving data synthesis. Proc. Very Large Data Base Endow. **10**(5), 481–492 (2017)
14. Calmon, F.P., Wei, D., Vinzamuri, B., Ramamurthy, K.N., Varshney, K.R.: Optimized pre-processing for discrimination prevention. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 3995–4004. Curran Associates Inc. (2017)
15. Chen, D., Zhao, H.: Data security and privacy protection issues in cloud computing. In: The IEEE International Conference on Computer Science and Electronics Engineering, vol. 1, pp. 647–651 (2012)
16. Chhabra, S., Singh, R., Vatsa, M., Gupta, G.: Anonymizing k-facial attributes via adversarial perturbations. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence, pp. 656–662 (2018)

17. Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W.F., Sun, J.: Generating multi-label discrete patient records using generative adversarial networks. Proc. Mach. Learn. Res. **68**, 286–305 (2017)
18. Dandekar, R.A., Cohen, M., Kirkendall, N.: Sensitive micro data protection using latin hypercube sampling technique. In: Domingo-Ferrer, J. (ed.) Inference Control in Statistical Databases. LNCS, vol. 2316, pp. 117–125. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-47804-3_9
19. De Cristofaro, E.: An overview of privacy in machine learning. arXiv preprint arXiv:2005.08679 (2020)
20. Desfontaines, D., Pejó, B.: SoK: differential privacies. Proc. Priv. Enhanc. Technol. **2020**(2), 288–313 (2020)
21. Domingo-Ferrer, J.: A survey of inference control methods for privacy-preserving data mining. In: Aggarwal, C.C., Yu, P.S. (eds.) Privacy-Preserving Data Mining. Advances in Database Systems, vol. 34, pp. 53–80. Springer, Boston (2008). https://doi.org/10.1007/978-0-387-70992-5_3
22. Drechsler, J.: Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation, vol. 201. Springer, Heidelberg (2011). https://doi.org/10.1007/978-1-4614-0326-5
23. Drechsler, J., Bender, S., Rässler, S.: Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB establishment panel. Trans. Data Priv. **1**(3), 105–130 (2008)
24. Drechsler, J., Reiter, J.P.: An empirical evaluation of easily implemented, non-parametric methods for generating synthetic datasets. Comput. Stat. Data Anal. **55**(12), 3232–3243 (2011)
25. Dwork, C., Roth, A.: The algorithmic foundations of differential privacy. Found. Trends® Theor. Comput. Sci. **9**(3–4), 211–407 (2014)
26. Dwork, C., Smith, A., Steinke, T., Ullman, J., Vadhan, S.: Robust traceability from trace amounts. In: 2015 IEEE 56th Annual Symposium on Foundations of Computer Science, pp. 650–669 (2015)
27. Fredrikson, M., Jha, S., Ristenpart, T.: Model inversion attacks that exploit confidence information and basic countermeasures. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, pp. 1322–1333 (2015)
28. Fung, B.C.M., Wang, K., Chen, R., Yu, P.S.: Privacy-preserving data publishing: a survey of recent developments. ACM Comput. Surv. **42**(4), 1–53 (2010)
29. Gambs, S.: Privacy and ethical challenges in big data. In: Zincir-Heywood, N., Bonfante, G., Debbabi, M., Garcia-Alfaro, J. (eds.) FPS 2018. LNCS, vol. 11358, pp. 17–26. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-18419-3_2
30. Garofalo, G., Van hamme, T., Preuveneers, D., Joosen, W.: A siamese adversarial anonymizer for data minimization in biometric applications. In: IEEE European Symposium on Security and Privacy Workshops, pp. 334–343 (2020)
31. Garofalo, G., Preuveneers, D., Joosen, W.: Data privatizer for biometric applications and online identity management. In: Friedewald, M., Önen, M., Lievens, E., Krenn, S., Fricker, S. (eds.) Privacy and Identity 2019. IAICT, vol. 576, pp. 209–225. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-42504-3_14
32. Gomatam, S., Karr, A.F., Reiter, J.P., Sanil, A.P.: Data dissemination and disclosure limitation in a world without microdata: a risk-utility framework for remote access analysis servers. Stat. Sci. **20**, 163–177 (2005)
33. Goodfellow, I., et al.: Generative adversarial Nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, vol. 27, pp. 2672–2680. Curran Associates, Inc. (2014)

34. Gross, R., Airoldi, E., Malin, B., Sweeney, L.: Integrating utility into face de-identification. In: Danezis, G., Martin, D. (eds.) PET 2005. LNCS, vol. 3856, pp. 227–242. Springer, Heidelberg (2006). https://doi.org/10.1007/11767831_15
35. Gross, R., Sweeney, L., De la Torre, F., Baker, S.: Model-based face de-identification. In: International Computer Vision and Pattern Recognition Workshop, p. 161 (2006)
36. Hagendorff, T.: The ethics of AI ethics: an evaluation of guidelines. Minds Mach. **30**, 1–22 (2020)
37. Holzinger, A.: Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Inform. **3**(2), 119–131 (2016). https://doi.org/10.1007/s40708-016-0042-6
38. Huang, C., Kairouz, P., Chen, X., Sankar, L., Rajagopal, R.: Context-aware generative adversarial privacy. Entropy **19**(12), 656 (2017)
39. Huang, H., Ma, X., Erfani, S.M., Bailey, J., Wang, Y.: Unlearnable examples: making personal data unexploitable. In: International Conference on Learning Representations (2021)
40. Hukkelås, H., Mester, R., Lindseth, F.: DeepPrivacy: a generative adversarial network for face anonymization. In: Bebis, G., et al. (eds.) ISVC 2019. LNCS, vol. 11844, pp. 565–578. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-33720-9_44
41. Hundepool, A., et al.: Statistical Disclosure Control. Wiley, Hoboken (2012)
42. Jiang, L., Tan, R., Lou, X., Lin, G.: On lightweight privacy-preserving collaborative learning for internet-of-things objects. In: Proceedings of the International Conference on Internet of Things Design and Implementation, pp. 70–81 (2019)
43. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowl. Inf. Syst. **33**(1), 1–33 (2012)
44. Kop, M.: Machine learning & EU data sharing practices. Stanford-Vienna Transatlantic Technology Law Forum, Transatlantic Antitrust (2020)
45. Larson, M., Liu, Z., Brugman, S., Zhao, Z.: Pixel privacy. Increasing image appeal while blocking automatic inference of sensitive scene information. In: Working Notes Proceedings of the MediaEval Workshop (2018)
46. Li, S.C., Tai, B.C., Huang, Y.: Evaluating variational autoencoder as a private data release mechanism for tabular data. In: 24th IEEE Pacific Rim International Symposium on Dependable Computing, pp. 198–1988 (2019)
47. Li, T., Lin, L.: AnonymousNet: natural face de-identification with measurable privacy. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (2019)
48. Liew, C.K., Choi, U.J., Liew, C.J.: A data distortion by probability distribution. ACM Trans. Database Syst. **10**(3), 395–411 (1985)
49. Liu, C., Chakraborty, S., Mittal, P.: DEEProtect: enabling inference-based access control on mobile sensing applications. arXiv preprint arXiv:1702.06159 (2017)
50. Lu, P.H., Wang, P.C., Yu, C.M.: Empirical evaluation on synthetic data generation with generative adversarial network. In: Proceedings of the 9th International Conference on Web Intelligence, Mining and Semantics, pp. 1–6 (2019)
51. Malekzadeh, M., Clegg, R.G., Cavallaro, A., Haddadi, H.: Mobile sensor data anonymization. In: ACM Proceedings of the International Conference on Internet of Things Design and Implementation, pp. 49–58 (2019)
52. McNamee, R., Parakilas, S.: The Facebook breach makes it clear: data must be regulated. The Guardian (2018). https://www.theguardian.com/commentisfree/2018/mar/19/facebook-data-cambridge-analytica-privacy-breach. Accessed 07 Feb 2021

53. Meden, B., Emeršič, Ž, Štruc, V., Peer, P.: K-same-net: K-anonymity with generative deep neural networks for face deidentification. Entropy **20**(1), 60 (2018)
54. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635 (2019)
55. Miotto, R., Wang, F., Wang, S., Jiang, X., Dudley, J.T.: Deep learning for healthcare: review, opportunities and challenges. Brief. Bioinform. **19**(6), 1236–1246 (2018)
56. Mirjalili, V., Raschka, S., Ross, A.: Gender privacy: an ensemble of semi adversarial networks for confounding arbitrary gender classifiers. In: the 9th IEEE International Conference on Biometrics Theory, Applications and Systems, pp. 1–10 (2018)
57. Mirjalili, V., Raschka, S., Ross, A.: PrivacyNet: semi-adversarial networks for multi-attribute face privacy. IEEE Trans. Image Process. **29**, 9400–9412 (2020)
58. Mirjalili, V., Raschka, S., Namboodiri, A., Ross, A.: Semi-adversarial networks: convolutional autoencoders for imparting privacy to face images. In: International Conference on Biometrics, pp. 82–89. IEEE (2018)
59. Muralidhar, K., Sarathy, R.: Data shuffling: a new masking approach for numerical data. Manage. Sci. **52**(5), 658–670 (2006)
60. Narayanan, A., Shmatikov, V.: Robust de-anonymization of large sparse datasets. In: IEEE Symposium on Security and Privacy, pp. 111–125 (2008)
61. Newton, E.M., Sweeney, L., Malin, B.: Preserving privacy by de-identifying face images. IEEE Trans. Knowl. Data Eng. **17**(2), 232–243 (2005)
62. Oh, S.J., Fritz, M., Schiele, B.: Adversarial image perturbation for privacy protection – a game theory perspective. In: International Conference on Computer Vision (ICCV) (2017)
63. Oleszkiewicz, W., Kairouz, P., Piczak, K., Rajagopal, R., Trzciński, T.: Siamese generative adversarial privatizer for biometric data. In: Jawahar, C.V., Li, H., Mori, G., Schindler, K. (eds.) ACCV 2018. LNCS, vol. 11365, pp. 482–497. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20873-8_31
64. Olhede, S.C., Wolfe, P.J.: The growing ubiquity of algorithms in society: implications, impacts and innovations. Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci. **376**(2128), 20170364 (2018)
65. Osia, S.A., et al.: A hybrid deep learning architecture for privacy-preserving mobile analytics. IEEE Internet Things J. **7**, 4505–4518 (2020)
66. Pan, Y.L., Haung, M.J., Ding, K.T., Wu, J.L., Jang, J.S.: k-Same-Siamese-GAN: k-same algorithm with generative adversarial network for facial image deidentification with hyperparameter tuning and mixed precision training. In: IEEE proceedings of the 16th International Conference on Advanced Video and Signal Based Surveillance, pp. 1–8 (2019)
67. Ren, Z., Jae Lee, Y., Ryoo, M.S.: Learning to anonymize faces for privacy preserving action detection. In: Proceedings of the European Conference on Computer Vision, pp. 620–636 (2018)
68. Rubin, D.B.: Discussion statistical disclosure limitation. J. Off. Stat. **9**(2), 461 (1993)
69. Rui, Z., Yan, Z.: A survey on biometric authentication: toward secure and privacy-preserving identification. IEEE Access **7**, 5994–6009 (2018)
70. Salter, C., Saydjari, O.S., Schneier, B., Wallner, J.: Toward a secure system engineering methodolgy. In: Proceedings of the 1998 Workshop on New Security Paradigms, pp. 2–10. ACM (1998)

71. Shan, S., Wenger, E., Zhang, J., Li, H., Zheng, H., Zhao, B.Y.: Fawkes: protecting personal privacy against unauthorized deep learning models. In: Proceeding of USENIX Security (2020)
72. Shokri, R., Strobel, M., Zick, Y.: Privacy risks of explaining machine learning models. arXiv preprint arXiv:1907.00164 (2019)
73. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: Symposium on Security and Privacy, pp. 3–18. IEEE (2017)
74. Slokom, M., Larson, M., Hanjalic, A.: Data masking for recommender systems: prediction performance and rating hiding. In: Late Breaking Results, in Conjunction with the 13th ACM Conference on Recommender Systems (2019)
75. Srivastava, D., Scannapieco, M., Redman, T.C.: Ensuring high-quality private data for responsible data science: vision and challenges. J. Data Inf. Qual. **11**(1), 1–9 (2019)
76. Sweeney, L.: Achieving k-anonymity privacy protection using generalization and suppression. Internat. J. Uncertain. Fuzziness Knowl.-Based Syst. **10**(05), 571–588 (2002)
77. Templ, M.: Statistical Disclosure Control for Microdata: Methods and Applications in R. Springer, Heidelberg (2017). https://doi.org/10.1007/978-3-319-50272-4
78. Tim, V., Denny, B., Lente, V.H., Ruben, V.D.H.: Google employees are eavesdropping, even in your living room VRT NWS has discovered (2019). https://www.vrt.be/vrtnws/en/2019/07/10/google-employees-are-eavesdropping-even-in-flemish-living-rooms/. Accessed 07 Feb 2021
79. Torra, V.: Privacy in data mining. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, pp. 687–716. Springer, Boston (2009). https://doi.org/10.1007/978-0-387-09823-4_35
80. Torra, V.: Masking methods. In: Torra, V. (ed.) Data Privacy: Foundations, New Developments and the Big Data Challenge. Studies in Big Data, vol. 28, pp. 191–238. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-57358-8_6
81. Tripathy, A., Wang, Y., Ishwar, P.: Privacy-preserving adversarial networks. In: 57th IEEE Annual Allerton Conference on Communication, Control, and Computing, pp. 495–505 (2019)
82. Wu, F.T.: Defining privacy and utility in data sets. Univ. Colorado Law Rev. **84**, 1117 (2013)
83. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional GAN. In: Wallach, H., Larochelle, H., Beygelzimer, A., dÁlché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 32, pp. 7335–7345 (2019)
84. Yu, T., Jajodia, S.: Secure Data Management in Decentralized Systems, vol. 33. Springer, Boston (2007). https://doi.org/10.1007/978-0-387-27696-0
85. Zhao, Z., Liu, Z., Larson, M.: Towards large yet imperceptible adversarial image perturbations with perceptual color distance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020
86. Rechberger, C., Walch, R.: Privacy-preserving machine learning using cryptography. In: Batina, L., Bäck, T., Buhan, I., Picek, S. (eds.) Security and Artificial Intelligence. LNCS, vol. 13049, pp. 109–129. Springer, Cham (2022)
87. Hernández-Castro, C.J., Liu, Z., Serban, A., Tsingenopoulos, I., Joosen, W.: Adversarial machine learning. In: Batina, L., Bäck, T., Buhan, I., Picek, S. (eds.) Security and Artificial Intelligence. LNCS, vol. 13049, pp. 287–312. Springer, Cham (2022)