sustainability

MDPI

Article

Examining Compliance with Personal Data Protection Regulations in Interorganizational Data Analysis

Szu-Chuang Li [1], Yi-Wen Chen [1,*] and Yennun Huang [2]

[1] Department of Information and Communication, Tamkang University, New Taipei City 237, Taiwan; 156671@mail.tku.edu.tw
[2] CITI, Academia Sinica, Taipei 115, Taiwan; yennumhuang@citi.sinica.edu.tw
* Correspondence: 137186@mail.tku.edu.tw

Abstract: The development of big data analysis technologies has changed how organizations work. Tech giants, such as Google and Facebook, are well positioned because they possess not only big data sets but also the in-house capability to analyze them. For small and medium-sized enterprises (SMEs), which have limited resources, capacity, and a relatively small collection of data, the ability to conduct data analysis collaboratively is key. Personal data protection regulations have become stricter due to incidents of private data being leaked, making it more difficult for SMEs to perform interorganizational data analysis. This problem can be resolved by anonymizing the data such that reidentifying an individual is no longer a concern or by deploying technical procedures that enable interorganizational data analysis without the exchange of actual data, such as data deidentification, data synthesis, and federated learning. Herein, we compared the technical options and their compliance with personal data protection regulations from several countries and regions. Using the EU's GDPR (General Data Protection Regulation) as the main point of reference, technical studies, legislative studies, related regulations, and government-sponsored reports from various countries and regions were also reviewed. Alignment of the technical description with the government regulations and guidelines revealed that the solutions are compliant with the personal data protection regulations. Current regulations require "reasonable" privacy preservation efforts from data controllers; potential attackers are not assumed to be experts with knowledge of the target data set. This means that relevant requirements can be fulfilled without considerably sacrificing data utility. However, the potential existence of an extremely knowledgeable adversary when the stakes of data leakage are high still needs to be considered carefully.

Keywords: personal data protection; privacy; federated learning; data deidentification

Citation: Li, S.-C.; Chen, Y.-W.; Huang, Y. Examining Compliance with Personal Data Protection Regulations in Interorganizational Data Analysis. *Sustainability* **2021**, *13*, 11459. https://doi.org/10.3390/su132011459

Academic Editors: Lucia Porcu and Nuria Rodríguez-Priego

Received: 9 August 2021
Accepted: 14 October 2021
Published: 16 October 2021

1. Introduction

With the rapid development of machine learning and deep learning, public and private organizations have increasingly leveraged big data technologies to, for example, uncover novel solutions or provide evidence for business decisions that are otherwise based on intuition, according to Harvard business review. Scholars are also increasingly using big data for exploratory studies [1,2], and governments have been making government data more transparent to citizens in the hope that such data can be used to add value [3].

For small- and medium-sized enterprises (SMEs), big data presents significant opportunities and challenges. SMEs know their customers well but have limited capacity for data collection and analysis [4,5]. Moreover, the benefit of using big data technology becomes more evident as the volume and variety of data increase [4]. For small businesses or organizations, independently achieving the four Vs of big data, namely velocity, volume, variety, and veracity [5], is challenging. The substantial technological capacity, number of users, and variety of services of tech giants such as Google and Facebook better equip these enterprises to attain these goals.

Despite these obstacles, SMEs can find methods for keeping up with their larger counterparts in the big data race. To overcome resource and capacity limitations, delegating activities in the big data value chain [6] such as analysis and visualization to organizations with relevant expertise constitutes a viable strategy. To increase data volume and variety, data sharing or collaborative data analysis among data controllers can benefit the group as a whole [7]. In short, small organizations can enhance their competitiveness through interorganizational data analysis.

However, such collaboration requires data to be transferred from each involved party, and privacy regulations must be respected when personal data are involved. Incidents such as Cambridge Analytica' s illegal use of personal data, Amazon' s improper use of customer data from their customer database [8], and frequent data leakage incidents from various Internet platforms have resulted in the implementation of stricter personal data protection regulations worldwide. The General Data Protection Regulation (GDPR) of the European Union (EU) covers multiple facets of personal data protection, including the right to erasure (Article 17) and the right not to be subject to a decision based solely on automated processing (Article 22). The most essential facet is the prohibition of the sharing of user data across organizational and geographical borders without user consent, and similar rules are commonly found in the privacy regulations of various countries or regions. Regulations such as the GDPR make companies of all sizes more cautious about privacy. However, they also make it harder for SMEs to create synergy from their data collection efforts.

Technical options for facilitating interorganizational data analysis without violating privacy regulations, either by circumventing or conforming to them, are available. Data anonymization is one such feasible approach. Anonymized data are usually exempt from personal data protection regulations and can be treated as nonpersonal data. If data are processed to the point where no specific individuals can be identified, they are no longer personal data and are thus not covered by privacy regulations. The application of concepts such as *k*-anonymity [9,10] and noninteractive privacy [11] is often regarded as a solution to data sharing restrictions. These two methods can be used to produce data sets that retain the informational value of the original data sets while ensuring anonymity. Data sets that fit those criteria can then be processed without restriction and exchanged between organizations, given that they are no longer considered personal data.

Federated learning technology presents another opportunity to interorganizational data analysis. Developed by Google to improve the performance of keyboard input prediction, this mechanism allows machine learning to proceed locally on mobile devices and for model parameters to be sent a central server such that a shared model can be constructed. Thereafter, the model is sent back to the mobile devices to produce predictions. In the process, no raw data in any form is transmitted to the central server, thereby eliminating the concern regarding the transfer of personal data across organizations or geographical borders and thus conforming to privacy regulations. Federated learning enables multiple devices to train a shared model without the original data ever having to leave the device, thereby opening up the opportunity for secure interorganizational model training—albeit without the possibility of conducting conventional analyses such as descriptive statistics analysis.

For SMEs to remain competitive vis-à-vis tech giants, they must have the capability to utilize big data technology. That said, interorganizational data analysis must comply with personal data regulations to satisfy the societal expectation for personal data protection, or the practice cannot be sustained in the long term. Data anonymization and federated learning are promising technical procedures, but questions regarding whether such processes are compliant with relevant regulations remain. The literature mostly focuses on either the technological or regulatory aspects of personal data protection in a data exchange scenario rather than their congruity. In this paper, we answer these questions by integrating insights from technological and privacy regulation studies as well as from the content of privacy regulations and government guidelines.

The remainder of this paper is organized as follows. In Section 2, the research questions and research method are presented. Next, recent trends in personal data protection and the pivotal role of interorganizational data analysis are revisited to demonstrate the importance of performing interorganizational data analysis without violating privacy regulations. Integral roles in personal data protection that are defined in the GDPR are also explained. Subsequently, we describe the requirements that must be fulfilled before interorganizational data analysis can be conducted and the approaches to circumvent them or conform with them. Potential solutions for interorganizational data analysis are then introduced, and their compliance with the privacy regulations are examined. Section 6 discusses the paper. Section 7 concludes the paper and outlines directions for future research.

2. Research Questions and Research Method

In this study, we determined whether current technical solutions for interorganizational data analysis comply with relevant privacy regulations. The research questions were as follows:

- How do current privacy regulations complicate the performance of interorganizational data analysis?
- What are the current technical options for enabling interorganizational data analysis? Do they comply with relevant privacy regulations?

To answer these questions, reference materials from various sources were reviewed. They include:

- Privacy regulations and government initiatives from various parts of the world, including the EU, the United Kingdom, Japan, and Taiwan. The GDPR was the main point of reference;
- Studies regarding technical procedures that facilitate interorganizational data analysis in compliance with privacy regulations;
- Studies (including real-world case studies) conducted in Japan and Taiwan that address the need for interorganizational data analysis in adherence to privacy regulations.

By surveying the technologies, privacy regulations, and privacy protection practices of several countries and regions, we identified the determinants of a specific strategy can enable interorganizational data analysis without violating privacy regulations. We first introduced the current technologies that facilitate interorganizational data analysis from the technical perspective, and picked three for further analysis due to their broader application in practice. We then determined whether these technologies comply with relevant privacy regulations. Subsequently, we compared these technologies so that enterprises can choose one that best suits their use case to enable interorganizational data analysis.

3. Personal Data Protection, Interorganizational Data Analysis, and Business Sustainability

Long before big data was known as such, enterprises were leveraging customer data to gain a competitive advantage. With the widespread adoption of big data technology, personal data protection has become an even more pressing concern.

Data breach incidents [12] and the misuse of user data [13] are no longer unexpected to the average consumer and have created tensions between users and vendors. Countries and regions such as the United States, the United Kingdom, the EU, Japan, and Taiwan all have regulations regarding personal data protection to urge enterprises to respect their customers' privacy and provide better protection. One example is that, through GDPR, the EU can impose large fines for data privacy violations amounting either up to 4% of an enterprise's annual global revenue or up to EUR 20 million. In the GDPR, several articles ensure that a user has a substantial amount of control over their data in specific circumstances. Two such articles are listed as follows.

- Article 17 states that each user has the " right to be forgotten," which means a consumer can ask service providers to erase their data from their database if they so wish;
- Article 22 states that each user has the right not to be subject to a decision based solely on automated processing, including profiling, that produces a legal effect concerning them.

Vendors must pay more attention to user privacy, not only to avoid the hefty fines that may influence their profitability but also to maintain sustainable relationships with their users. Chen [14] reported that after the Facebook–Cambridge Analytica data scandal, Facebook users began tightening up their privacy settings on the platform, which complicated Facebook's ability to leverage user data. Implementing more stringent privacy protection policies and practices allows vendors to maintain sustainable relationships with users, which in turn helps vendors obtain a sustainable competitive advantage from big data. For this reason, an increased number of vendors have begun treating user privacy seriously [15]. Failure to conform to personal privacy protection regulations may not only pose a financial risk to vendors but also cause them to lose their credibility with customers.

The ability to use big data has become an integral part of the decision-making process for businesses, and all enterprises are seeking opportunities to access more data. Tech giants such as Google and Facebook possess abundant resources, including enormous amounts of data, which give them a considerable headstart over other companies. To stay competitive, other companies must work together to overcome their disadvantages through data sharing or by distributing various activities on the big data value chain to various parties. If a business cannot find a means to leverage big data technology long term, its sustainability will suffer.

Personal data protection regulations usually stipulate that users must consent to a business sharing their data with other partners. According to Article 5(1b) of the GDPR, personal data "shall be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes." Under these restrictions, vendors can perform statistical analysis on their data. However, collaborative analysis entails data exchange in some form. The only way to enable this without obtaining explicit consent from individual users was to conduct adequate data anonymization.

Recently, federated learning has opened up a path to interorganizational data analysis without the data leaving the organization that owns them. This constitutes a novel approach for interorganizational data analysis. Data anonymization and federated learning could be enablers of business sustainability in the era of big data and data privacy. We introduce and compare the technologies and examine their compliance with current personal data protection regulations starting from the Section 4.

As defined in Sections 3 and 4 of the GDPR, the following terms are used to refer to various roles in interorganizational data analysis in the remainder of this paper.

- Data subjects: users or consumers from whom personal data are collected;
- Data controller: the entity that collects and manages data;
- Data processor: the entity that processes data on behalf of the data controller.

4. Approaches Facilitating Interorganizational Data Analysis under the Requirements Outlined in Personal Data Protection Regulations

4.1. GDPR Restrictions on Personal Data Processing

As mentioned, SMEs must engage in interorganizational data analysis to remain sustainable in the era of big data. However, an approach to sharing data, whether raw data or aggregated statistics, must be present if the SME is to collaborate with a third party in data analysis. Personal data protection regulations often prohibit the transfer of personal data without user consent. For example, as defined in GDPR Article 6(I), personal data can be processed lawfully if the following conditions are met :

- The data subject has given consent to the processing of their or her personal data for one or more specific purposes;

- Processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract;
- Processing is necessary for compliance with a legal obligation to which the controller is subject;
- Processing is necessary in order to protect the vital interests of the data subject or of another natural person;
- Processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller;
- Processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child.

These conditions indicate that, except in some extreme scenarios, user consent is almost always required if one vendor wants to transfer personal data to another vendor to facilitate collaborative data analysis. However, it is impractical to expect consent for this to be granted for every collaborative project. Currently, two routes can be taken to circumvent this limitation and achieve interorganizational data analysis when personal data are involved. First, a data controller can process the data such that no one can identify a specific person therefrom. Given that the data are no longer personal data, they can be used freely, with no restrictions under privacy protection regulations such as the GDPR. The second option is to employ technologies that enable various data controllers to perform machine learning or statistical analysis collaboratively without sharing actual data. The algorithms only require the transfer of calculation results such as machine learning parameters or statistical estimates and thus comply with privacy protection regulations. These technologies are described as follows.

4.2. Data Anonymization: Making Personal Data Nonpersonal

One of the objectives of personal data protection regulations such as the GDPR is to prevent the leakage of sensitive personal information, and the purpose of interorganizational data analysis is to generate insights from data collected from various sources. When personal data are involved, collaborative analysis through data exchange is prohibited. However, if the data can be processed to reduce or even eliminate the possibility that a specific individual is reidentified, they are regarded as nonpersonal and can thus be processed as necessary by the data controller. Such processes include the transfer of the data to other parties.

The most straightforward means to identify an individual is through "direct" identifiers; thus, attributes such as names and social security numbers or online data such as email addresses and identity documents on social media that can be employed to pinpoint a specific individual must first be removed. Direct identifiers are protected by personal data protection regulations in countries and regions such as the USA, the EU, the UK, Japan, and Taiwan. Masking or removing direct identifiers from a data set that is to be released to the public or shared with another organization has become the norm.

However, these measures are but a first step. If a data processor has access to several data sources, it has a greater likelihood of identifying specific individuals by linking the other attributes. For instance, from government-released open data, a data processor can obtain a data set with no direct identifiers but containing some personal attributes such as age, gender, and residential area, as well as sensitive health information. Suppose the same processor can also obtain another data set with the same personal attributes but with direct identifiers. In that case, it can link a health status with a specific individual with great confidence. In this scenario, attributes such as age, gender, and residential area are called indirect identifiers or quasi-identifiers [10].

The example also indicates that indirect identifiers are difficult to define. For instance, clinics commonly have patients' personal data, which include medical information such

as their blood type and body weight. Therefore, to prevent potential data linkage–related attacks, all attributes must be treated as indirect identifiers. This is because we cannot exclude the possibility that some organizations hold data sets containing information to allow a data processor to identify a specific individual by linking the data set it possesses with the data set obtained from a public or shared data source.

To overcome these limitations, data deidentification techniques such as *k*-anonymity can reduce or eliminate the possibility of identification by data linkage. Quasi-identifiers are processed or removed to the point where the probability of guessing the correct piece of sensitive information is greatly reduced. However, data utility is compromised when many indirect identifiers are present (because information pertaining to these identifiers cannot be overly specific). Data anonymity and data utility thus necessarily trade off each other [16].

In addition to *k*-anonymity, data synthesis has also been used as a means to generate anonymized data in recent surveys [17,18]. Such anonymization approaches involve capturing the statistical properties of the original data and reproducing a new data set from them. The resulting data set is called the synthesized data set. If the statistical properties can be captured perfectly, the resulting data set could theoretically be the same as the original data set. Noise addition in the process of data set generation is thus often the critical element in the construction of such a privacy-preserving data set. The privacy metric often influences the amount of random noise added and is negatively related to data utility. Given that the resulting data set is often generated from a stochastic mechanism, that the resulting and original data sets contain no identical records is not guaranteed. The most popular synthetic data set generation methods are those based on noninteractive differential privacy. In contrast to those in *k*-anonymity, the attributes need not be classified as sensitive or not sensitive.

In the Section 4.3, two methods of data anonymization are introduced and their compliance with current personal data protection regulations are examined.

4.3. Interorganizational Analysis without the Exchange of Raw Data

Instead of deidentifying the original data, parties can conduct certain limited data analysis tasks together without exchanging data sets in any form. Data exchange remains necessary for interorganizational data analysis. However, if statistical results are the subject of data exchange, they have a much higher likelihood of conforming with privacy regulations because the chances of identifying a specific individual through aggregated statistics is considerably lower. Federated learning, which fits this criterion, is one of the fastest-growing research streams given that machine learning and deep learning are subjects of extensive discussion.

Before the emergence of federated learning, local differential privacy was the go-to model when aggregated statistics were required but the collection of raw user data was not allowed. Randomized Aggregatable Privacy Preserving Ordinal Responses was developed by Google [19] for the transfer of anonymous data on browser usage behavior. Data from end devices are first processed by several hash functions and Bloom filters, and, subsequently, permanent and instantaneous noise complying with differential privacy are added before the data are sent to a central server. The central server can then employ statistical techniques in estimating the true frequencies of the strings sent from all the end devices. The method, termed randomized response, can be traced back to a 1965 social scientific study [20]. In that study, respondents were asked whether they were drug users and were instructed to secretly flip a coin before answering. Respondents were prompted to provide a fixed answer (e.g., "yes") if the coin flip yielded heads; therefore, the researcher could not be sure about whether a specific respondent's answer was true. However, after a large number of responses were collected, the researcher could infer the true percentage of drug users.

Researchers have performed various statistical tasks using local differential privacy, including the estimation of frequency, mean value, and range queries. Scholars have also

developed supervised learning and unsupervised learning methods [21]. As is the case in federated learning, no data in the local differential privacy model are in the same form as the original data transferred to a central server; thus, the data are expected to conform to privacy regulations. Because the statistical and machine learning tasks this model can manage are limited and highly specific, they are not discussed herein. Notably, researchers are attempting to leverage local differential privacy in addressing advanced security issues in federated learning [22], and these approaches can work in concert in the future.

Homomorphic encryption can be applied to secure aggregation, enabling calculations between encrypted data. Although a tremendous amount of computing power is required, homomorphic encryption is becoming an increasingly practical option. Machine learning tasks such as can be performed on an encrypted data set [23] through homomorphic encryption. Because the transfer of an encrypted data set is required, the bandwidth overhead is higher than those in federated learning and the local differential privacy model.

Federated learning, local differential privacy, and homomorphic encryption are similar in that they cannot be combined with other statistical or machine learning tools, unlike anonymized data sets. Special algorithms must be developed if a data analyst wishes to use those approaches. By contrast, anonymized data sets can be analyzed by any off-the-shelf software or the ready-built modules in popular programming languages. Due to its recent popularity, we introduce federated learning herein and examine its compliance with current personal data protection regulations.

5. Potential Technologies for Interorganizational Data Analysis and Their Compliance with Personal Data Protection Regulations

Consulting recent surveys on data anonymization [17,18] (Rao, 2018; Majeed, 2021), we revisited the technical aspects of *k*-anonymity and noninteractive differential privacy to analyze their adherence to current personal data regulations. We aligned our discussion with articles on data privacy regulations (e.g., the GDPR) and with various concepts (e.g., the motivated intruder test). Federated learning serves as a potential alternative for and was thus also introduced for later comparison with data anonymization techniques for privacy-preserving interorganizational data analysis.

5.1. k-Anonymity for Data Anonymization

5.1.1. Technical Overview

The concept of *k*-anonymity was first proposed by Samarati and Sweeney [9]. Inspired by the public voter list in certain US states, they investigated the possibility of identifying a specific individual using indirect identifiers. In some cases, combining data sources and identifying an individual's sensitive attributes therefrom were extremely easy. When some medical information is released to the public with the direct identifiers removed but the indirect identifiers retained, an adversary can compare it with a public voter list. If the city or town the person lives in is small, at least one other individual could conceivably have the same value combination of indirect identifiers such as age, gender, and ethnicity. When this is the case, the individual's sensitive data have been leaked.

A solution to this problem is *k*-anonymity, a property that ensures that the value combination of indirect identifiers cannot be used to single out any individual. Two procedures are employed to guarantee that a data set with the same combinations of indirect identifiers contains at least *k* records such that no individual can be singled out. They are described as follows.

- Generalization: By combining values into a single value in an attribute, people with similar attribute combinations can be "hidden" in a cluster. For example, if the data set features only one musician, the data controller can combine the musician and writer categories into a new category called "artist" to ensure that the musician will not be singled out. For numerically coded attributes, a range of values can be regarded as the

same value. To increase the value of k, the data controller can give assign all numerical values in this range with a value of 21 to 25.

- Suppression: Another approach to avoid a small k value is to remove the attribute or the data record altogether. This is usually accomplished by removing data from the data set or by marking certain parts of the data with an asterisk. Although data utility is more greatly affected by suppression than by generalization, suppression can be used as a last resort when reaching the necessary k value is difficult.

The k-anonymity property is not without its problems. Consider the case of medical information release. If the k people in the data set all tested positive for a certain disease, their sensitive information would still be leaked; people would know that they were sick. This problem can be solved using variations of k-anonymity such as l-diversity [24] and t-closeness [25]. Specifically, l-diversity seeks to prevent this problem from occurring by ensuring that, given a group of people with the same combination of quasi-identifiers, variations for sensitive data attributes are present. The concept of t-closeness takes this one step further in ensuring that the distance between the distribution of a sensitive attribute in a group of people with the same combination of quasi-identifiers and the whole table does not exceed a threshold of t. This makes it even more challenging for an adversary to make a well-educated guess on the sensitive attribute of a specific individual.

5.1.2. Analysis of Compliance with Privacy Regulations

The key to privacy regulation compliance is that under the property of k-anonymity, the original data must be converted to a form wherein the risk of reidentification is reduced. As discussed in Section 4.2, the choice of quasi-identifiers is vital in terms of preventing data linkage attacks, but very little data utility would be preserved in the interest of ensuring maximal data protection. Recognizing that perfect anonymization may not be practical, the United Kingdom's Information Commissioner's Office (ICO) clarified that the prevention of data linkage should extend to data sources that are publicly accessible (e.g., government data) from, for example, the Internet. Private organizations should refrain from using their data sets to create, share, or release a data set in violation of data privacy regulations. In summary, before an organization releases a data set to the public or its partners, it should first remove or mask the direct identifiers and process the indirect identifiers to the point where no one can identify a specific individual therefrom by linking it with a publicly accessible data set.

The UK ICO implements the principles as the motivated intruder test [26], which is detailed as follows. On the one hand, the intruder has no prior knowledge of the data, performs no illegal activity, and is not someone whose profession is to compromise data privacy. On the other hand, the intruder has some motivation to intrude on a person's privacy and can use other public data sources, such as open government data, libraries, and the Internet, to reidentify an individual. As long as an anonymized data set ensures that no person can be reidentified in this clearly defined scenario, the data set is free from the restrictions of personal data protection regulations. This more relaxed scenario is thus used to ensure data utility while protecting privacy. Notably, organizations possessing sensitive personal data should take precautions to prevent them from being linked to publicly accessible sources. Moreover, they should be held responsible for any privacy leak caused by the mismanagement of sensitive personal data. In addition to the United Kingdom, and the EU, Australia also employs the motivated intruder test in gauging the adequacy of an anonymization process [27]. In Japan, a principle similar to this test is used [28].

Among the technical solutions for interorganizational data analysis, k-anonymity is the concept most favored by the nontechnical community. The relevant concepts have been incorporated into the regulations of some regions and countries. For example, indirect or quasi-identifiers and k values feature in the draft of the Verification Guide for the Process of Personal Data Deidentification, which was published by the National Standard Technology Council [29] of Taiwan. Furthermore, the concept of indirect identifiers is widely used

in the UK ICO's code of practice for data anonymization. The reason concepts related to *k*-anonymity are more widely accepted than their alternatives is that *k*-anonymity is considerably easier to explain and understand for people without a technical background.

By carefully eliminating direct identifiers and selecting quasi-identifiers for generalization and suppression, a data set processed under the property of *k*-anonymity should pass the motivated intruder test and satisfy the requirements of personal data protection regulations. This is because the prevention of data linkage to other publicly available data sources protects data from reidentification. A thorough scan for potential data linkages must be conducted before certain data set attributes are selected as indirect identifiers. Coupled with the fact that its concepts are widely adopted by government published guidelines in various countries and regions, *k*-anonymity is compliant with privacy regulations .

5.2. Noninteractive Differential Privacy for Data Anonymization

5.2.1. Technology Overview

One commonly held belief in the past was that the release of sample-level statistics, such as the mean and standard deviation of a data set, would not hurt privacy. However, privacy studies have refuted this belief. Suppose a billionaire is planning to move from one small town to another. This event will significantly influence the average resident income of the town. Therefore, anyone with access to the statistics of one of the towns can make a confident guess about whether the billionaire has moved. The authors of [11] took this assumption a step further by introducing the concept of differential privacy. Suppose one party possesses a dataset D^1, and the other possesses a data D^2, which is one record more or less than D^1. In that case, it is trivial to calculate and determine which record is left out in one of the datasets by comparing a statistics, such as mean. To prevent this from happening, the data controller needs to add calculated noise to the statistics before release, and this definition of privacy is called differential privacy. Differential privacy assumes that the adversary is very knowledgeable with the dataset, to the point that he possesses an almost identical dataset. To make the ouput statistics protected by differential privacy, the data controller can use a stochastic function $\mathcal{M}$ to add noise to it. We define the function as providing ε-DP if for any two datasets D^1 and D^2, differing by at most one record, and any dataset of possible outputs $S \subseteq Range(\mathcal{M})$, where the probability $\mathbb{P}$ depends on $\mathcal{M}$'s randomness, as defined in Equation (1):

$$\frac{\mathbb{P}[\mathcal{M}(D^1) \in S]}{\mathbb{P}[\mathcal{M}(D^2) \in S]} \leq e^{\varepsilon} \tag{1}$$

The parameter ε can be used to fine-tune the tradeoff between data privacy and data utility. Larger ε means less noise needs to be added by $\mathcal{M}$, which also means better data utility and worse privacy. Differential privacy is mathematically provable, and some US vendors [30] and government agencies [31] are beginning to adopt it.

The aforementioned scenario pertains to database queries and is called interactive differential privacy. To release or share data, the data controller can perform a series of data queries on the data set and use the results to rebuild a data set; this is called noninteractive differential privacy. Given that the data set is established from noise-added statistics, it is protected using differential privacy. Studies have constructed such a data set with a variety of statistical methods such as principal component analysis [32] and deep learning models [33]. Herein, we employ the DPTable algorithm [34] as an example to illustrate the process of noninteractive differential privacy.

The basic concept underlying DPTable is the making of a contingency table that contains the counts of all unique value combinations in the data set, adding of noise to those counts, and rebuilding of a data set from the perturbed table. The upper left hand side of Figure 1 illustrates four records with two attributes X and Y in the original data, and the right hand side of the figure is the contingency table, which contains the counts of various combinations of the (X,Y) value pair. Each count value is equivalent to a database query on the data set, and the addition of calculated noise allows those perturbed query

results (counts) to meet the requirement of differential privacy. Some count values in the Table are changed after noise is added, and the perturbedTable is then used to generate the new data set (bottom left of the figure). The procedure ensures that the synthetic data sets are protected by differential privacy. When the synthetic data sets are employed in machine learning tasks, they perform similarly to the original data sets, indicating favorable data utility. The degree of privacy preservation depends on the ε value the data controller selects for generating the synthetic data sets. When ε tends toward infinity, the noise added tends toward 0 and the content of the rebuilt data set is very similar to that of the original data set. In summary, it is essential to select suitable ε values for different cases [35,36] (see Figure 1).

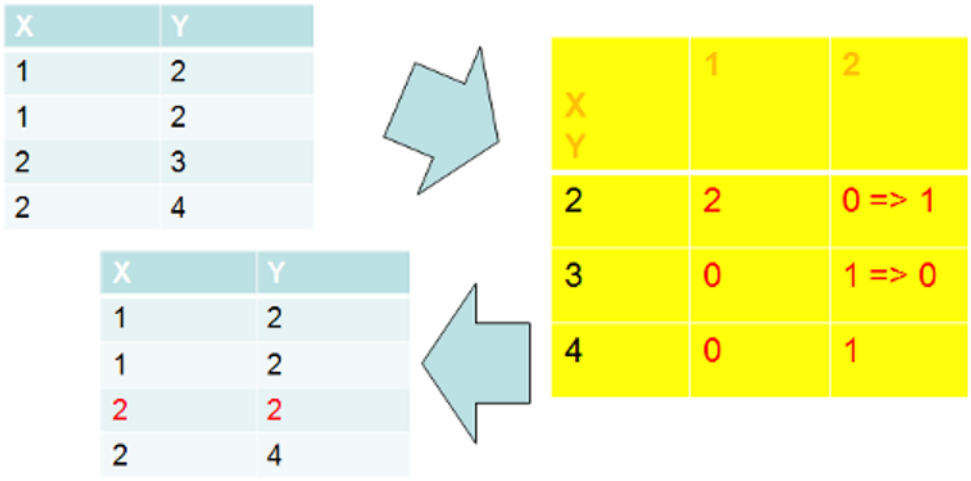

Figure 1. The basic concept of DPTable.

Notably, when the dimensionality of a data set is high, the tradeoff between data privacy and utility becomes substantially greater because a large amount of noise must be added to the queries according to the concept of differential privacy. DPTable uses some advanced mathematical techniques to synthesize high-dimensional tables. Nonetheless, the simplified illustration of DPTable provides an understanding of how noninteractive differential privacy functions.

5.2.2. Analysis of Compliance with Privacy Regulations

Data sets generated through approaches based on noninteractive differential privacy differ considerably from those processed through *k*-anonymity-based approaches. They are called synthetic data sets because they are generated from a series of queries on the original data set rather than by aggregating or removing data from it. DPTable generates each record from the stochastically perturbed contingencyTable and can ensure, to a reasonable extent, that no record can be associated with a particular record in the original data set. When a synthetic data set is generated, all records are sampled from a probability distribution. Thus, none of the synthetic data records are derived from a specific record or associated with a specific individual. In this respect, data sets generated by noninteractive differential privacy mechanisms are expected to be compliant with current personal privacy regulations.

However, as mentioned, the selected ε value greatly affects the similarity between the original and synthetic data sets. When an extremely large ε is chosen, the likelihood of minimal noise addition increases and the content of the synthetic data set can be very similar to that of the original data set. Comparing the two data sets, the synthetic data set may contain records identical to those in the original data set. In this context, if ε is large, the situation is exacerbated. Moreover, a data subject could claim that their private information has been linked. In this case, an adversary can make an "educated guess" but

not a certain one, and there are some descriptions of such situations in the UK ICO's Guide to Data Protection.

The " Special Category Data" of the UK ICO's Guide to Data Protection states that an "educated guess," even when on target, does not constitute a data leak [37]. However, if inferences can be made from certain linkages between data sets, the data records should be regarded as compromised. Data sets generated through a stochastic process under the concept of noninteractive differential privacy provide no information that can be linked with certainty to other data. Provided the synthetic data set can pass the motivated intruder test, it should be good for sharing without violating privacy regulations.

5.3. Federated Learning: Collaborative Machine Learning without Data Sharing

5.3.1. Technical Overview

Federated learning, first proposed by Google to improve keyboard input efficiency for mobile phone users [38], was designed to simultaneously enable distributed machine learning and preserve data privacy. Each mobile phone has a local model and performs model training locally. Parameters from mobile phones are sent to a central server and processed by a custom-designed algorithm to update the parameters of the central model. The results will then be transmitted back to the mobile phones to improve local models such that all participants in federated learning can benefit.

Because federated learning enables collaborative machine learning without data aggregation, researchers and practitioners have regarded it as a means for enabling interorganizational data analysis. Such a scenario was termed cross-silo machine learning (as opposed to cross-device machine learning) [39]. Cross-device applications emphasize effective model training on low-powered devices and low-bandwidth network connections without compromising the user experience. By contrast, cross-silo applications have no such limitations. Most studies on cross-silo federated learning have focused on the performance of distributed machine learning.

Federated learning can be categorized as follows, according to [40] (to the left shuld be a citation):

- Horizontal federated learning: In this representative setting of federated learning, the nodes all possess the same data attributes and labels;
- Vertical federated learning: Because the nodes in federated learning are assigned different attributes possessed by the same object of analysis (e.g., different attributes of the same individual), data attributes must be aligned using a certain type of encryption mechanism;
- Transfer federated learning: The results of a machine learning task are applied to another machine learning task to enhance performance. For instance, in text mining, models representing the relationship between terms in documents are typically constructed, and those representations are in turn used in other studies.

Federated learning allows organizations to train machine learning models collaboratively without data sharing, making it a promising solution to privacy problems concerning interorganizational data analysis. Furthermore, the distributed nature of the mechanism saves bandwidth and distributes the computational load between nodes, making it especially suitable for Internet of Things devices with limited system resources and unstable network connections. In Taiwan, where most companies are SMEs, federated learning seems to be a favorable option for businesses to work together in facing the challenges of big data. Taiwan's government has set up an alliance to coordinate efforts from various sectors to promote federated learning [41].

5.3.2. Analysis of Compliance with Privacy Regulations

Personal information is the subject of protection in data privacy regulations. As long as the data contain no information from which a specific individual can be identified, the data should not be categorized as personal data and is thus safe for release or exchange. In federated learning, model parameters, which contain neither direct nor indirect

identifiers (as is the case with *k*-anonymity), are sent to a central server. In summary, an organization using the data they own to participate in federated learning is not in violation of personal data protection regulations. This is why federated learning has been regarded as the solution to interorganizational data analysis. Moreover, a recent study endorsed the compliance of federated learning to the GDPR [42].

Of course, federated learning is not free of privacy concerns. When investigating the privacy concerns of federated learning, we often assume that the central server is a curious one, and privacy concerns may arise if we examine the problem from the perspective of differential privacy [39]. Suppose the central server possesses the same data set that one of the data nodes possesses with the exception of one record and that it is fully aware of the algorithm that is employed by the data nodes. The server can compare the parameters to identify this data record. This scenario demonstrates that federated learning is certainly not perfect when it comes to privacy protection. Various studies have addressed this problem [20,43,44]. Although the assumption that the central server possesses all the data of a data node is fairly extreme, it necessarily makes the transferred parameters fit the definition of an indirect identifier to a certain extent. However, as mentioned, the UK ICO indicates that when determining whether a data attribute is an indirect identifier, the assumption is that the professionals holding the personal data protect them from improper use. To assume otherwise would negatively affect the utility of shared or released data to an excessive degree. In summary, federated learning mechanisms must enable collaborative machine learning while conforming to privacy regulations, and future technical developments are expected to mitigate the data privacy concerns.

6. Discussion

6.1. Comparison of Interorganizational Data Analysis Mechanisms

Table 1 lists the characteristics of data deidentification methods and federated learning for comparison as approaches to interorganizational data analysis. Local differential privacy [19] and homomorphic encryption–based machine learning [23] are also listed for comparison. The advantage of federated learning is that because only machine learning parameters are transmitted, compliance with privacy regulations is higher and bandwidth consumption is lower. Low bandwidth consumption is imperative for cross-device applications but not for cross-silo applications. Federated learning systems enable distributed machine learning with predefined algorithms. If the algorithm is changed, the model training process must be restarted (see Table 1).

Table 1. A comparison between *k*-anonymity, non-interactive differential privacy and federated learning for interorganizational data analysis.

	***k*-Anonymity**	**Non-Interactive Differential Privacy**	**Federated Learning (Horizontal)**
Data to transfer	Data that has been processed to remove direct and indirect identifiers.	Data that generated from a series of randomized queries that is compliant to differential privacy.	Parameters of local machine learning model.
Tools for analysis	The analyst can use any tool that he or she is familiar with to process the deidentified data.	The analyst can use any tool that he or she is familiar with to process the synthetic data.	The collaborators needs to choose a machine learning method first.
Output	A dataset that has the same attributes as the original data.	A dataset that has the same attributes as the original data.	A global model that can be shared across nodes.
Type of analytic tasks	Statistical and machine learning tasks.	Statistical and machine learning tasks.	Machine learning tasks, but limited to those that have been implemented in the context of federated learning.
Privacy risk	Deidentified data can still be compromised via data linkage using alternative indirect identifiers.	Synthetic data can still contain records with identical attribute combination as in the original dataset especially when ε is large.	A "curious" central server with rich background knowledge as defined in differential privacy can make a good guess regarding if a person is in the dataset of a node or not.

By contrast, data deidentification techniques output data sets in the same format as that of the original data set. Therefore, data processors can use whatever tool they wish for analysis. The disadvantage of data deidentification is that the amount of data to be transferred is considerably larger. Thus, data leaks can occur from a greater number of points than they can in federated learning. For example, under the property of *k*-anonymity, attributes that the data controller regards as indirect identifiers may be processed, but a certain adversary can possess a data set with attributes that the data controller does not expect and perform a successful linkage attack. In a scenario involving noninteractive differential privacy, if the ε value chosen is large—that is, if little noise is added—numerous records may have the same combination of data attributes as those in the original data sets. In summary, therefore, data controllers must be extremely cautious when conducting data deidentification.

In cross-silo applications, when collaborators have reached a consensus on which machine learning task they wish to perform, federated learning is the better option. If the data processors are still in the process of selecting algorithms by implementing them on shared data, data deidentification is the preferable approach (see Table 2).

Table 2. Additional comparison about local differential privacy and homomorphic encryptionn-based machine learning with methods in Table 1.

	Local Differntial Privacy	**Homomorphic Encryption-Based Machinne Learning**
Data to transfer	Hashed data with provable privacy guarantee via perturbation	Encrypted data.
Tools for analysis	A central server that is able to aggregate the data and estimate statitics from them	Algorithms need to be developed for specific machine learning tasks.
Output	Aggregated statistical estimates such as counts	A model generated from encrypted data.
Type of analytic tasks	Mostly simple statistical results.	Specific machine learning tasks that have been implemented in the context of homomorphic encryption.
Privacy risk	Low privacy risk as only perturbated hashed data were transferred in the process.	Low privacy risk as the training data, prediction model, and data used for prediction are all encrypted.

6.2. Regulation Compliance versus Actual Privacy risk

Scholars have formulated methods for protecting and reidentifying personal data. The determination of whether personal data in a data set are well protected depends on how we define privacy. For example, data privacy as defined in *k*-anonymity requires data controllers to select certain attributes as indirect identifiers and process them to eliminate the possibility that a specific individual can be identified through data linkage. However, the possibility that an adversary can launch such a linkage attack despite possessing data with attributes not marked as indirect identifiers remains. Although such a situation falls outside the purview of institutions such as the UK ICO, it nevertheless constitutes a privacy risk.

Synthetic data sets generated through approaches based on noninteractive differential privacy stochastically produce artificial data sets from the probability distribution of the original data set. Although the synthetic data set technically contains no real personal data, some generated data records could coincidentally have all the attributes constituent of a real person. Furthermore, if a large ε value is selected when a synthetic data set is generated, the likelihood of this occurring skyrockets. Federated learning, when evaluated using the standard of differential privacy, is also vulnerable to similar attacks. In short, no mechanism both provides perfect privacy protection and retains high data utility.

Adequately anonymized data, synthetic data, and federated learning can pass the test of privacy risk assessment practices such as the motivated intruders test, but data controllers must be aware that bypassing or complying with privacy regulations does not mean that data sharing or release is perfectly safe from privacy leaks from a technical standpoint. Furthermore, data controllers must understand what scenarios can occur

and be prepared for them. In some circumstances, the adoption of privacy protection techniques far exceeding the requirements of data privacy regulations may be necessary. For instance, if a data set contains critical personal data that if leaked can cause serious legal or financial risks, substantially higher standards of privacy protection should be imposed in technology and parameter selection.

7. Conclusions and Directions for Future Research

Advancements in big data technology have made it not only a notable source of competitive advantage for enterprises but also a contributor to various social and scientific breakthroughs. The requirement of a large amount of data and rich data attributes creates a tension between data processing and privacy when personal data are involved. Moreover, in view of tightened personal data privacy regulations, big data analysis might become exclusive to tech giants if SMEs have no means of leveraging the data they own collaboratively. The methods reviewed herein provide options to enable such organizations to aggregate their data collection efforts through synergistic efforts, thereby allowing them to stay competitive in the race of big data.

SMEs often do not possess data of sufficient volume or variety for leveraging the power of big data analysis. In this study, we provided solutions for such organizations to analyze data collaboratively without violating data privacy regulations. Our close examination of articles in the GDPR, as well as studies and government documents from countries and regions such as the United Kingdom, Japan, and Taiwan, revealed that the motivated intruder test is the current standard. When a data controller wishes to share their data, be it raw data or statistical results, with its partners, the data must first be adequately processed. An adversary should not be able to identify a specific individual just by viewing the data entries or by linking the data to a public data source (e.g., those on the Internet). In case of privacy infringement caused by the integration of the shared data with another private source of sensitive data, the party that failed to guard or process the data adequately must be held accountable, not the technologies that facilitate interorganizational data analysis. In sum, the technologies introduced herein that can pass the relevant tests are all effective tools that can meet the requirements of data privacy regulations.

This study has several limitations. First, the literature review did not yield a high number of real-world examples of court decisions regarding privacy issues surrounding interorganizational data analysis. For example, Japan's Act on the Protection of Personal Information contains references to articles related to data deidentification starting from 2017. The Act mandates that the organizations performing data deidentification must set up a contact window for the resolution of potential disputes. As of 2018, no complaint had been filed despite the fact that 380 organizations had conducted data deidentification by that point [28]. In lawsuits regarding personal data leakage filed in Taiwan, the judges involved dismissed the the notion that leaks of indirect-identifier data constitute a privacy breach; this indicates that legal practitioners lack a clear understanding of data anonymization [45]. We aim to investigate any privacy disputes in the context of interorganizational data analysis should they arise. Second, local differential privacy and other approaches that can be employed in interorganizational data analysis were not examined herein because of their still-limited application. Future studies can undertake a more comprehensive technical review featuring a greater number of technical solutions or an analysis of their compliance with privacy regulations.

Author Contributions: Conceptualization, S.-C.L. and Y.-W.C.; methodology, S.-C.L.; investigation, S.-C.L. and Y.-W.C.; resources, Y.-W.C. and Y.H.;writing—original draft preparation, S.-C.L.; writing—review and editing, Y.-W.C. and Y.H.; supervision, Y.H.; project administration, S.-C.L.; funding acquisition, Y.-W.C. and Y.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Ministry of Science and Technology grant number MOST107-2221-E-001-016-MY2 and MOST109-2410-H-032-039-.

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

References

1. McAfee, A.P.; Brynjolfsson, E. Big data: The management revolution. *Harv. Bus. Rev.* **2012**, *10*, 142–149.
2. Kitchin, R. Big Data, new epistemologies and paradigm shifts. *Big Data Soc.* **2014**, *1*, 1–12. [CrossRef]
3. Open Government Data. Available online: https://www.oecd.org/gov/digital-government/open-government-data.htm (accessed on 14 May 2021).
4. Del Vecchio, P.; Di Minin, A.; Messeni Petruzzelli, A.; Panniello, U.; Pirri, S. Big data for open innovation in SMEs and large corporations: Trends, opportunities, and challenges. *Creat. Innov. Manag.* **2018**, *27*, 6–22. [CrossRef]
5. Wang, S.; Wang, H. Big data for small and medium-sized enterprises (SME): A knowledge management model. *J. Knowl. Manag.* **2020**, *24*, 881–897. [CrossRef]
6. Miller, H.; Mork, P. From Data to Decisions: A Value Chain for Big Data. *IT Prof.* **2013**, *15*, 57–59. [CrossRef]
7. Durrant, A.; Markovic, M.; Matthews, D.; May, D.; Enright, J.; Leontidis, G. The Role of Cross-Silo Federated Learning in Facilitating Data Sharing in the Agri-Food Sector. *arXiv* **2021**, arXiv:2104.07468.
8. Amazon Gets Record $888 Million EU Fine over Data Violations. Available online: https://www.bloomberg.com/news/articles/2021-07-30/amazon-given-record-888-million-eu-fine-for-data-privacy-breach (accessed on 4 August 2021).
9. Samarati, P.; Sweeney, L. Protecting Privacy when Disclosing Information: *k*-Anonymity and its Enforcement through Generalization and Suppression. In *Technical Report SRI-CSL-98-04*; Computer Science Laboratory, SRI International: Menlo Park, CA, USA, 1998.
10. Sweeney, L. *k*-Anonymity: A model for protecting privacy. *Int. J. Uncertainty Fuzziness Knowl. Based Syst.* **2002**, *10*, 557–570. [CrossRef]
11. Dwork, C. Differential Privacy. In Proceedings of the 33rd International Conference on Automata, Languages and Programming, Venice, Italy, 10–14 July 2006; Springer: Berlin/Heidelberg, Germany, 2006.
12. The 56 Biggest Data Breaches. Available online: https://www.upguard.com/blog/biggest-data-breaches (accessed on 14 May 2021).
13. Cambridge Analytica and Facebook: The Scandal and the Fallout So Far. Available online: https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html (accessed on 14 May 2021).
14. Chen, Y.N.K.; Wen, C.H.R. A study of the privacy behavior of Taiwan Facebook users after the Cambridge Analytica scandal. *Commun. Soc.* **2020**, *54*, 27–57.
15. Data Privacy and Protection in the ESG Era. Available online: https://www.alpha-sense.com/blog/data-privacy-esg/ (accessed on 15 May 2021).
16. Li, T.; Li, N. On the tradeoff between privacy and utility in data publishing. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, 28 June–1 July 2009.
17. Rao, P.R.; Krishna, S.; Kumar, A.S. Privacy preservation techniques in big data analytics: A survey. *J. Big Data* **2018**, *5*, 1–12.
18. Majeed, A.; Lee, S. Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey. *IEEE Access* **2021**, *9*, 8512–8545. [CrossRef]
19. Erlingsson, Ú.; Korolova, A.; Pihur, V. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. In Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, Scottsdale, AZ, USA, 3–7 November 2014.
20. Wei, K.; Li, J.; Ding, M.; Ma, C.; Yang, H.H.; Farhad, F.; Jin, S.; Quek, T.; Poor, H. Federated Learning with Differential Privacy: Algorithms and Performance Analysis. *IEEE Trans. Inf. Forensics Secur.* **2020**, *15*, 3454–3469. [CrossRef]
21. Yang, M.; Lyu, L.; Zhao, J.; Zhu, T.; Lam, K. Local Differential Privacy and Its Applications: A Comprehensive Survey. *arXiv* **2020**, arXiv:2008.03686.
22. Arachchige, P.C.; Bertók, P.; Khalil, I.; Liu, D.; Çamtepe, S.; Atiquzzaman, M. Local Differential Privacy for Deep Learning. *IEEE Internet Things J.* **2020**, *7*, 5827–5842. [CrossRef]
23. Han, K.; Hong, S.; Cheon, J.; Park, D. Logistic Regression on Homomorphic Encrypted Data at Scale. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019.
24. Machanavajjhala, A.; Gehrke, J.; Kifer, D.; Venkitasubramaniam, M. L-diversity: Privacy beyond k-anonymity. In Proceedings of the 22nd International Conference on Data Engineering, Dallas, TX, USA, 20–24 April 2006.
25. 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, 11–15 April 2007; pp. 106–115.
26. Anonymisation: Managing Data Protection Risk Code of Practice. Available online: https://ico.org.uk/media/1061/anonymisation-code.pdf (accessed on 16 May 2021).
27. Hsiang, C. *A Study on Open Data and Its Influence on Governance of Government and Personal Privacy*; The Research Center for Digital Governance, National Development Council: Taipei, Taiwan, 2015.
28. Fan-Chiang, C.; Chou, Y. *Final Report: Data De-Identification Regulations in Japanese Personal Data Protection Act*; National Development Council: Taipei, Taiwan, 2019.
29. Verification Guide for the Process of Personal Data De-Identification (Draft). Available online: https://www.bsmi.gov.tw/wSite/public/Data/f1456791848684.pdf (accessed on 16 May 2021).
30. Differential Privacy. Available online: https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf (accessed on 16 May 2021).

31. Differential Privacy and the US Census. Available online: https://www.youtube.com/watch?v=NNTBQ_K4h7c (accessed on 17 May 2021).
32. Kitchin, R. Differential-Private Data Publishing Through Component Analysis. *Trans. Data Priv.* **2013**, *6*, 19–34.
33. Abay, N.C.; Zhou, Y.; Kantarcioglu, M.; Thuraisingham, B.; Sweeney, L. Privacy Preserving Synthetic Data Release Using Deep Learning. In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Dublin, Ireland, 10–14 September 2018.
34. Chen, R.; Xiao, Q.; Zhang, Y.; Xu, J. Differentially Private High-Dimensional Data Publication via Sampling-Based Inference. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sidney, Australia, 10–13 August 2015.
35. Lee, J.; Cliffton, C. How much is enough? choosing ε for differential privacy. In Proceedings of the 14th International Conference on Information Security, Xi'an, China, 26–29 October 2011.
36. Chen, R.; Xiao, Q.; Zhang, Y.; Xu, J. Differential Privacy: An Economic Method for Choosing Epsilon. In Proceedings of the IEEE 27th Computer Security Foundations Symposium, Vienna, Austria, 19–22 July 2014.
37. Guide to Data Protection. Available online: https://ico.org.uk/for-organisations/guide-to-data-protection/ (accessed on 24 May 2021).
38. Yang, T.; Andrew, G.; Eichner, H.; Sun, H.; Li, W.; Kong, N.; Ramage, D.; Beaufays, F. Applied federated learning: improving google keyboard query suggestions. *arXiv* **2018**, arXiv:1812.02903.
39. Kairouz, P.; McMahan, H.B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A.; Bonawitz, K.; Charles, Z.B.; Cormode, G.; Cummings, R.; et al. Advances and Open Problems in Federated Learning. *arXiv* **2021**, arXiv:1912.04977.
40. Yang, Q.; Liu, Y.; Chen, T.; Tong, Y. Federated Machine Learning: Concept and Applications. *ACM Trans. Intell. Syst. Technol. (TIST)* **2019**, *10*, 1–19. [CrossRef]
41. Taiwan AI Federated Learning Alliance. Available online: https://www.cio.com.tw/taiwan-joint-learning-and-operation-alliance-launched/ (accessed on 17 May 2021).
42. Truong, N.; Sun, K.; Wang, S.; Guitton, F.; Guo, Y. Privacy Preservation in Federated Learning: An insightful survey from the GDPR Perspective. *Cryptogr. Secur.* **2021**, *110*, 102402.
43. Seif, M.; Tandon, R.; Li, M. Wireless Federated Learning with Local Differential Privacy. In Proceedings of the IEEE International Symposium on Information Theory, Los Angeles, CA, USA, 21–26 June 2020.
44. Hu, R.; Guo, Y.; Li, H.; Pei, Q.; Gong, Y. Personalized Federated Learning with Differential Privacy. *IEEE Internet Things J.* **2020**, *7*, 9530–9539. [CrossRef]
45. Sung, H. Legal Risks and Management Implications of Big Data Transactions—Focusing on the Reidentification of Personal Data. *Manag. Rev.* **2018**, *37*, 37–51.