Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/cose

Computers & Security

TC 11 Briefing Papers

Privacy preservation in federated learning: An insightful survey from the GDPR perspective

Nguyen Truong[a,*], *Kai Sun*[a], *Siyao Wang*[a], *Florian Guitton*[a], *YiKe Guo*[a,b,**]

[a] *Data Science Institute, South Kensington Campus, Imperial College London, London SW7 2AZ, United Kingdom*
[b] *Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong*

ARTICLE INFO

Article history:
Received 7 April 2021
Revised 17 June 2021
Accepted 12 July 2021
Available online 17 July 2021

Keywords:
Federated learning
Data protection regulation
GDPR
Personal data
Privacy
Privacy preservation

ABSTRACT

In recent years, along with the blooming of Machine Learning (ML)-based applications and services, ensuring data privacy and security have become a critical obligation. ML-based service providers not only confront with difficulties in collecting and managing data across heterogeneous sources but also challenges of complying with rigorous data protection regulations such as EU/UK General Data Protection Regulation (GDPR). Furthermore, conventional centralised ML approaches have always come with long-standing privacy risks to personal data leakage, misuse, and abuse. Federated learning (FL) has emerged as a prospective solution that facilitates distributed collaborative learning without disclosing original training data. Unfortunately, retaining data and computation on-device as in FL are not sufficient for privacy-guarantee because model parameters exchanged among participants conceal sensitive information that can be exploited in privacy attacks. Consequently, FL-based systems are not naturally compliant with the GDPR. This article is dedicated to surveying of state-of-the-art privacy-preservation techniques in FL in relations with GDPR requirements. Furthermore, insights into the existing challenges are examined along with the prospective approaches following the GDPR regulatory guidelines that FL-based systems shall implement to fully comply with the GDPR.

1. Introduction

We are now living in a data-driven world where most applications and services such as healthcare and medical services, autonomous cars, and finance applications are based on artificial intelligence (AI) technology with complex data-hungry machine learning (ML) algorithms. AI has been showing advances in every aspect of lives and expected to *"change the world more than anything in the history of mankind. More than elec-*

* Corresponding author.
** Principal corresponding author
E-mail addresses: n.truong@imperial.ac.uk (N. Truong), k.sun@imperial.ac.uk (K. Sun), s.wang18@imperial.ac.uk (S. Wang), f.guitton@imperial.ac.uk (F. Guitton), y.guo@imperial.ac.uk (Y. Guo).
https://doi.org/10.1016/j.cose.2021.102402

tricity."[1]. However, the AI technology is yet to reach its full potential, also the realisation of such AI/ML-based applications has been still facing long-standing challenges wherein centralised storage and computation is one of the critical reasons.

In most of the real-world scenarios, data, particularly personal data, is generated and stored in data silos, either end-users' devices or service providers' data centres. Most conventional ML algorithms are operated in a centralised fashion, requiring training data to be fused in a data server. Essentially, collecting, aggregating and integrating heterogeneous data dispersed over various data sources as well as securely managing and processing the data are non-trivial tasks. The challenges are not only due to transporting high-volume, high-velocity, high-veracity, and heterogeneous data across organisations but also the industry competition, the complicated administrative procedures, and essentially, the data protection regulations and restrictions such as the EU General Data Protection Regulation (GDPR)[2] Horvitz and Mulligan (2015). In traditional ML algorithms, large-scale data collection and processing at a powerful cloud-based server entails the single-point-of-failure and the risks of severe data breaches. Foremost, centralised data processing and management impose limited transparency and provenance on the system, which could lead to the lack of trust from end-users as well as the difficulty in complying with the GDPR Truong et al. (2019).

To overcome such challenges, Federated Learning (FL), proposed by Google researchers in 2016, has appeared as a promising solution and attracted attention from both industry and academia Konečnỳ et al. (2016a,b); McMahan et al. (2017a, 2016). Generally, FL is a technique to implement an ML algorithm in decentralised collaborative learning settings wherein the algorithm is executed on multiple local datasets stored at isolated data sources (i.e., local nodes) such as smartphones, tablet, PCs, and wearable devices without the need for collecting and processing the training data at a centralised data server. FL allows local nodes to collaboratively train a shared ML model while retaining both training dataset and computation at internal sites Konečnỳ et al. (2016a). Only results of the training (i.e., parameters) are exchanged at a certain frequency, which requires a central server to coordinate the training process (centralised FL) or utilises a peer-to-peer underlying network infrastructure (i.e., decentralised FL) to aggregate the training results and calculate the global model.

The natural advantage of FL compared to the traditional cloud-centric ML approaches is the ability to reassure data privacy and (presumably) comply with the GDPR because personal data is stored and processed locally, and only model parameters are exchanged. In addition, the processes of parameters updates and aggregation between local nodes and a central coordination server are strengthened by privacy-preserving and cryptography techniques, which enhance data security and privacy Bonawitz et al. (2016, 2017); Geyer et al. (2017); Phong et al. (2018); Wei et al. (2020). The FL capability could potentially inaugurate new opportunities for service providers to implement some sorts of ML algorithms for their applications and services without acquiring clients' personal data, hence naturally complying with data protection regulations like the GDPR. Unfortunately, despite the distributed collaborative learning model of FL empowered by privacy-preserving measures, personal information can be stealthily extracted from local training parameters Aono et al. (2017); Hitaj et al. (2017); Melis et al. (2019); Phong et al. (2018); Zhu et al. (2019). As a consequence, FL-based service providers still stay within the regulatory personal data protection framework and are still liable for implementing GDPR-compliant mechanisms when dealing with EU/UK citizens.

In this article, we conduct a survey on existing FL studies with an emphasis on privacy-preserving techniques from the GDPR-compliance perspective. Firstly, we briefly review the challenges on data privacy preservation in conventional centralised ML approaches (*Section 2*) and dummyTXdummy- introduce FL as a potential approach to address the challenges (*Section 3*). Secondly, the state-of-the-art privacy-preserving techniques for centralised FL are described with the analysis of how these solutions can mitigate data security and privacy risks (*Section 4*). Thirdly, we provide an insightful deliberation with potential solution approaches of how an FL system can be implemented in order to comply with the EU/UK GDPR (*Section 5*). Unsolved challenges hindering an FL system from complying with the GDPR are also specified along with the future research directions.

2. Privacy preservation and GDPR-Compliance in ML-based systems

2.1. *Fundamental background*

ML is a disruptive technology for designing and building intelligent systems that can automatically learn and improve from experience to accomplish a task without being explicitly programmed. For this purpose, an ML-based system builds up a mathematical model (i.e., model training process) based on a sample set (i.e., training data) whose parameters are to be optimised during this training process. As a result, the system can perform better predictions or decisions on a new, unseen task. Typically, an ML task can be formulated as a mathematical optimisation problem whose goal is to find the extremum of an objective function. Thus, an optimisation method is of paramount importance in any ML-based systems.

2.1.1. *Gradient descent algorithm*

One of the most widely used optimisation methods for ML, which is also the core of FL, is gradient descent. It is a first-order iterative optimisation algorithm for finding a local minimum of an objective function $f(\theta)$ parameterised by a set of parameters $\theta \in \mathbb{R}^d$ Ruder (2016). Consider a samples set $\mathcal{D} = (x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$, and the objective function $f(\theta)$; a model training process uses the gradient descent method to update each parameter in the opposite direction of the gradient of the objective function $\nabla f(\theta)$ regarding to the parame-

[1] Dr. Kai-Fu Lee, former vice president at Google, https://www.cnbc.com/2019/01/14/the-oracle-of-ai-these-kinds-of-jobs-will-not-be-replaced-by-robots-.html

[2] https://gdpr-info.eu/.

ters by the following equation:

$$w_j \leftarrow w_j - \eta \nabla \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}(f(x_i) - y_i) \quad (1)$$

where w_j refers to the j^{th} parameter of θ, and η refers to the learning rate hyper-parameter, i.e., the size of steps to reach the optimal. $\mathcal{L}$ represents a loss function such as mean-square error (MSE) and cross-entropy loss. The parameters update process using Eq. 1 is iteratively carried out until either an acceptable local minimum is found or the difference of the loss between two consecutive steps is negligible.

2.1.2. *Gradient descent variants*

Generally, there are three gradient descent methods that are categorised based on the amount of training data used in the gradient calculation of the objective function $f(\theta)$ Ruder (2016). The first category is *batch gradient descent*, in which the gradients are computed over the entire training dataset $\mathcal{D}$ for one update. The second category is *stochastic gradient descent (SGD)*, that, in contrast to batch gradient descent, randomly selects a sample (or a subset) from $\mathcal{D}$ and performs the parameters update based on the gradient of this sample only (one sample per step, the whole process sweeps through the entire dataset). The third one is *mini-batch gradient descent* in which the dataset is subdivided into mini-batches of n training samples (n is the batch-size); the parameters update is then performed on every mini-batch (single mini-batch per step).

There is a trade-off between the accuracy of parameters update and the efficiency of the computation in each step of gradient descent. Generally, mini-batch gradient descent mitigates the problem of inefficiency in batch gradient descent and gradient oscillation in SGD. However, it introduces the extra hyper-parameter batch-size n, which requires expertise and extensive trial and error and sometimes needs to be manually adjusted Keskar et al. (2016). The gradient descent normally comes along with optimisers, which are techniques for controlling the learning rate η logistically and accurately. Such optimisers tie together with the model parameters θ and the loss function $\mathcal{L}$ in order to adjust the learning rate η in response to the output of the loss function. The most common gradient-based optimisers include Momentum Qian (1999), Adam Kingma and Ba (2014), RMSprop Tieleman and Hinton (2012), and Adagrad Duchi et al. (2011).

2.1.3. *Gradient descent in distributed learning*

Although gradient descent-based optimisation methods were successfully engaged in various ML algorithms, they have recently re-gained much attention since the emergence of large-scale distributed learning, including FL Bottou (2010); Dean et al. (2012). In these scenarios, a complex model, e.g., a deep neural network (DNN) with millions of parameters, is trained on a very large dataset across multiple nodes. These nodes are called *compute nodes* and grouped into *clusters*. For efficiency, the calculations in the training process should be parallelised using concurrency methods such as *model parallelism* and *data parallelism* Chen and Lin (2014). Model parallelism distributes an ML model into different computing blocks; available computing nodes are then be assigned to compute some specific blocks only. Model parallelism requires mini-batch data to be replicated at computing nodes in a cluster, as well as regular communication and synchronisation among such nodes Dean et al. (2012). Data parallelism, instead, keeps the completeness of the model on each computing node but partitions the training dataset into smaller equal size shards (also known as *sharding*), which are then distributed to computing nodes in each cluster Ben-Nun and Hoefler (2019). The computing nodes then train the model on their subset as a mini-batch, which is especially effective for SGD variants because most operations over mini-batches are independent in these algorithms. Data parallelism can be found in numerous modern ML frameworks including *TensorFlow*[3] and *Pytorch*[4]. The two parallelism techniques can also be combined (so-called Hybrid parallelism) to intensify the advantages while mitigating the drawbacks of each one; as a result, a hybrid system can achieve better efficiency and scalability Chilimbi et al. (2014).

2.2. *Privacy preserving techniques in ML*

Generally, privacy preservation techniques for a distributed learning system target two main objectives: *(i)* privacy of the training dataset and *(ii)* privacy of the local model parameters (from an optimisation algorithm such as a gradient descent variant) which are exchanged with other nodes and/or a centralised server Shokri and Shmatikov (2015). In this respect, prominent privacy-preserving techniques in ML include data anonymisation Narayanan and Shmatikov (2008), differential privacy Dwork et al. (2006), secure multi-party computation (SMC) Yao (1986), and homomorphic encryption Gentry (2010).

2.2.1. *Data anonymisation*

Data anonymisation or de-identification is a technique to hide (e.g., hashing) or remove sensitive attributes, such as personally identifiable information (PII), so that a data subject cannot be identified within the modified dataset (i.e., the anonymous dataset) Narayanan and Shmatikov (2008). As a consequence, data anonymisation has to balance well between privacy-guarantee and utility because hiding or removing information may reduce the utility of the dataset. Furthermore, when combined with auxiliary information from other anonymous datasets, a data subject might be re-identified, subjected to a privacy attack called *linkage attack* Fung et al. (2010). To prevent from linkage attack, numerous techniques have been proposed such as *k-anonymity* Sweeney (2002), *l-diversity* Machanavajjhala et al. (2007), a *k-anonymity*-based method, and *t-closeness* - a technique built on both *k-anonymity* and *l-diversity* that preserves the distribution of sensitive attributes in a dataset so that it reduces the risk of re-identifying a data subject in a same quasi-identifier group Li et al. (2007).

Unfortunately, such privacy-preserving techniques cannot defend against linkage attacks whose adversaries possess some knowledge about the sensitive attributes. This deficiency in the *k-anonymity*-based methods calls for different approaches that offer rigorous privacy-guarantee such as *differential privacy*.

[3] https://www.tensorflow.org/.
[4] https://pytorch.org/.

2.2.2. *Differential privacy*

Proposed by Dwork *et al.* in 2006, differential privacy Dwork et al. (2006) is an advanced solution of the perturbation privacy/preserving technique in which random noise is added to true outputs using rigorous mathematical measures Fung et al. (2010). As a result, it is statistically indistinguishable between an original aggregate dataset and a differentially additive-noise one. Thus, a single individual cannot be identified as any (statistical) query results to the original dataset is practically the same regardless of the existence of the individual Dwork (2008); Dwork et al. (2006, 2014). However, there is a trade-off between privacy-guarantee and utility as adding too much noise and improper randomness will significantly depreciate reliability and usability of the dataset Dwork (2008); Dwork et al. (2014); Fung et al. (2010).

Differential privacy technique has been widely employed in various ML algorithms such as linear and logistic regression Chaudhuri and Monteleoni (2009), Support Vector Machine (SVM) Rubinstein et al. (2012) and deep learning Abadi et al. (2016); Chaudhuri et al. (2011), as well as in ML-based applications such as data mining Friedman and Schuster (2010) and signal processing with continuous data Sarwate and Chaudhuri (2013).

2.2.3. *Secure multi-party computation*

SMC, also known as multi-party computation (MPC) or privacy-preserving computation, was firstly introduced by Yao in 1986 Yao (1986) and further developed by numerous researchers. Its catalyst is that a function can be collectively computed over a dataset owned by multiple parties using their own inputs (i.e., a subset of the dataset) so that any party learns nothing about others data except the outputs Canetti et al. (1996); Cramer et al. (2000); Goldreich (1998). Specifically, n parties $P_1, P_2, ., P_n$ own n pieces of private data $X_1, X_2, \ldots, X_n$, respectively to collectively compute a public function $f(X_1, X_2, ., X_n) = (Y_1, Y_2, ., Y_n)$. The only information each party can obtain from the computation is the result $(Y_1, Y_2, ., Y_n)$ and its own inputs X_i. Classical secret sharing such as Shamir secret sharing Brickell (1989); Shamir (1979) and verifiable secret sharing (VSS) schemes Chor et al. (1985) are the groundwork for most of the SMC protocols.

Although the idea of SMC has been investigated and shown feasible since late 1980s Goldreich et al. (2019); Yao (1986), its practicality still remains a long-standing challenge. SMC groundwork protocols Canetti et al. (1996); Cramer et al. (2000); Goldreich (1998); Yao (1986), which are based on zero-knowledge proofs, were shown to be inefficient and impractical under the presence of malicious adversaries Jarecki and Shmatikov (2007). These protocols are built upon the *Yao garbled circuits* idea in Yao (1986) which are only efficient in semi-honest settings. Significant research effort to achieve security against malicious adversaries while being efficient has been carried out; and one of the notable technique is based on *cut-and-choose* paradigm Lindell and Pinkas (2007). In the cut-and-choose approach, as a large number of circuits are processed in order to prevent from adversaries, significant overheads, both in computation and in communication, are introduced. To overcome this challenge, some efficient SMC protocols based on the cut-and-choose paradigm have been proposed and shown to be practical while achieving the same level of security in the malicious adversaries settings Huang et al. (2013); Lindell (2016).

With such efficient SMC protocols, it is feasible to achieve data privacy in distributed learning wherein compute nodes collaboratively perform model training on their local dataset without revealing such dataset to others. Indeed, SMC has been employed in numerous ML algorithms such as secure two-party computation (S2C) in linear regression Du et al. (2004), Iterative Dichotomiser-3 (ID3) decision tree learning algorithm Lindell and Pinkas (2000), and *k-means* clustering algorithm for distributed data mining Jagannathan and Wright (2005). However,most of SMC protocols impose non-trivial overheads which require further efficiency improvements with practical deployment.

2.2.4. *Homomorphic encryption*

Another approach to preserve data privacy and security in ML is to utilise homomorphic encryption techniques, particularly in centralised systems, e.g., cloud servers, wherein data is collected and trained at a server without disclosing the original information. Homomorphic encryption enables the ability to perform computation on an encrypted form of data without the need for the secret key to decrypt the cipher-text Gentry (2010). Results of the computation are in encrypted form and can only be decrypted by the requester of the computation. In addition, homomorphic encryption ensures that the decrypted output is the same as the one computed on the original unencrypted dataset.

Depending on encryption schemes and classes of computational operations that can be performed on an encrypted form, homomorphic encryption techniques are divided into different categories such as partial, somewhat (SWHE), and fully homomorphic encryption (FHE)Acar et al. (2018). Some classic encryption techniques, including *Rivest æShamir æAdleman (RSA)*, is SWHE wherein simple addition and multiplication operations can be executed Acar et al. (2018). FHE, firstly proposed by Graig *et al.* in Gentry and Boneh (2009); Gentry and Halevi (2011), enables any arbitrary operations (thus, enables any desirable functionality) over cipher-text, yielding results in encrypted forms. In FHE, computation on the original data or the cipher-text can be mathematically transferred using a decryption function without any conflicts.

Even though homomorphic encryption offers rigorous privacy-guarantee to individuals as the original data in plain-text has never been disclosed, there is a practical limitation in performing computation over cipher-text due to the tremendous computational overhead. As a consequence, employing homomorphic encryption in large-scale data training remains impractical Gilad-Bachrach et al. (2016).

2.3. *The GDPR*

The new GDPR legislation has come into force from May 2018 in all European Union (EU) countries which is a major update to the EU Data Protection Directive (95/46/EC) (DPD-95) introduced in the year 1995. The GDPR aims to protect personal data (more comprehensive range depicted in *"Which?"* - Fig. 1) with the impetus that *"personal data can only be gathered legally, under strict conditions, for a legitimate purpose"*. The

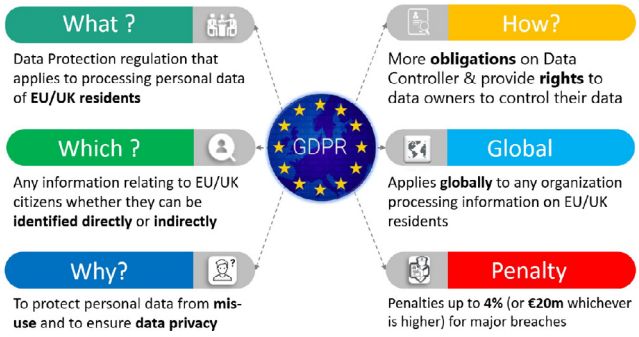

Fig. 1 – The GDPR legislation in a nutshell.

full regulation is described in detail across 99 articles covering principles, and both technical and admin requirements around how organisations need to process personal data. The GDPR creates a legal data protection framework throughout the EU/UK member states which has impacted commercial and public organisations worldwide processing EU/UK residents' data (*"Global"* in Fig. 1).

The GDPR clearly differentiates three participant roles, namely: Data Subject, Data Controller and Data Processor, along with associated requirements and obligations under the EU/UK data protection law. While serving as a better privacy and security framework, the GDPR also aims at protecting data ownership by obligating Data Controllers to provide fundamental rights for Data Subjects to control over their data (*"How?"* in Fig. 1). For these purposes, the GDPR introduces and sets high-standard for the consent lawful basis in which Data Controller shall obtain consent from Data Subject in order to process data. Data Controller takes full responsibility to regulate the purposes for which and the methods in which, personal data is processed under the Terms and Conditions defined in the consent.

2.4. *Challenges on complying with the GDPR*

To meet stringent requirements of the GDPR, conventional ML-based applications and services are required to implement measures that effectively protect and manage personal data adhering to the six data protection principles in the GDPR, as well as to provide mechanisms for data subjects to fully control their data. Although ML-based systems are strengthened by several privacy-preserving methods, implementing these obligations in a centralised ML-based system is non-trivial, sometimes technologically impractical Greengard (2018); Wachter et al. (2017).

Large-scale data collection, aggregation and processing at a central server in such ML-based systems not only entail the risks of severe data breaches due to single-point-of-failure but also intensify the lack of transparency, data misuse and data abuse because the service providers are in full control of the whole data lifecycle Truong et al. (2019). In addition, as ML algorithms operate in a black-box manner, it is also challenging to provide insightful interpretation of how the algorithms execute and how certain decisions are made Mehrabi et al. (2019); Murdoch et al. (2019). Consequently, most of the ML-based systems find it difficult to satisfy the requirements of transparency, fairness, and automated decision-making in the GDPR.

Furthermore, the requirements of purpose limitation and data minimisation are not always feasibly carried out in ML-based systems. The majority of ML algorithms heavily rely on data quality and quantity, thus researchers tend to collect as much related data as possible. Therefore, determining 1) the purposes of data collection as well as 2) what data is adequate, limited, and relevant only to the claimed purposes before executing such ML algorithms are problematic challenges. These requirements overly restrict the natural operations of ML-based services and applications to a smaller range than ever before.

Finally, ML algorithms are essentially designed for optimising performance, whereas privacy preservation measures remain to be a simple disclaimer. With rigorous requirements of the GDPR, such ML algorithms shall be redesigned internally at the algorithm level in order to accommodate sufficient privacy-preserving techniques. This system redesign requires enormous, or even infeasible, efforts in terms of both technological resolution and human and financial resources. In addition, the trade-off between efficiency and privacy-guarantee is apparently a serious issue for many service providers as sacrificing system performance might lead to the inability to handle their existing services.

3. Federated learning: A Distributed collaborative learning approach

In many scenarios, the traditional cloud-centric ML approaches are no longer suitable due to the challenges of complying with strict data protection regulations on vast aggregation and processing of personal data. By nature, most personal data is generated at the edge by end-users' devices (e.g., smartphones, tablets, and wearable devices) which are equipped with increasingly powerful computing capability and Internet connectivity. Given the pervasiveness of such personal devices along with the growing privacy concerns, the trend of decentralised AI has naturally risen which converges the mobile edge computing (MEC) Hu et al. (2015) with AI/ML techniques to migrate the intelligence from the cloud to the edge Wang et al. (2020).

In this regard, FL is an alternative for the cloud-centric ML technique that facilitates an ML model to be trained collaboratively while retaining original personal data on their devices, thus potentially mitigates data privacy-related vulnerabilities. It is a cross-disciplinary technique covering multiple computer science aspects including ML, distributed computing, data privacy and security that enables end-users' devices (i.e., local nodes) to locally train a shared ML model on local data. Only parameters in the training process are exchanged for model aggregation and updates. The difference between FL and standard distributed learning is that in distributed learning, local training datasets in compute nodes are assumed to be *independent and identically distributed data* (IID) whose sizes are roughly the same. FL is, thus, an advancement of distributed learning as it is designed to work with unbalanced and *non-independent identically-distributed data* (non-IID) whose sizes may span several orders of magnitude. Such het-

erogeneous datasets reside at a massive number of scattering mobile devices under unstable connectivity and limited communication bandwidth Kairouz et al. (2019); McMahan et al. (2017a, 2016).

3.1. Model training in federated learning

FL is well-suited for sorts of ML models that are formulated as minimisation of some objective functions (loss functions) on a training dataset for parameter estimation, particularly for gradient-based optimisation algorithms Konečnỳ et al. (2016a). The minimisation objective can be formulated as follows:

$$\min_{w\in\mathbb{R}^d} f(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w) \tag{2}$$

where the training dataset is in form of a set of input-output pairs $(x_i, y_i), x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}, \forall i \in \{1, 2, .., n\}$. In Eq. 2, n is the number of samples in the dataset, $w \in \mathbb{R}^d$ is the *parameter vector*, and $f_i(w)$ is a loss function. This formulation covers both linear and logistic regressions, support vector machines, as well as complicated non-convex problems in Artificial Neural Networks (ANN) including Deep Learning Konečnỳ et al. (2016a). This problem requires an optimisation process that can be efficiently computed by using a gradient descent algorithm with back-propagation technique Rezende et al. (2014); Rumelhart et al. (1985) for minimising the overall loss with respect to each model parameters.

In traditional ML approaches, this sort of algorithms performs a vast number of fast iterations over a large dataset homogeneously partitioned in data servers. Such algorithms require super low-latency and high-throughput connections to the training data McMahan et al. (2017a). Therefore, solving this optimisation problem in the context of FL is different from the traditional ML approaches as such conditions do not hold in FL settings. Training data in FL is unbalanced and non-IID, which is scattered across millions of personal mobile devices with significant higher-latency, lower-throughput connections compared to the traditional techniques working on a cloud-centric data server. In addition, the data and computing resources in personal devices are only intermittently available for training. Therefore, to actualise FL, optimisation algorithms must be well adapted and efficiently performed for federated settings (i.e., federated optimisation Konečnỳ et al. (2016a)).

3.2. Federated optimisation

One of the fundamentals of FL is efficient optimisation algorithms for federated settings wherein training data is non-IID, massively and unevenly distributed across local nodes, first introduced by Konečnỳ *et al.* in 2016 Konečnỳ et al. (2016a). The distributed settings for the federated optimisation is formulated as follows. Let K be the number of local nodes, $\mathbb{P}_k$ be the set of data samples stored on node $k \in \{1, 2, .., K\}$, and $n_k = |\mathbb{P}_k|$ be the number of data samples stored on node k. As personal data in each local node is different, we can assume that $\mathbb{P}_k \cap \mathbb{P}_l = \varnothing$ if $k \neq l$ and $\sum_{k=1}^{K} n_k = n$. The distributed problem formulation for the minimisation objective is defined as:

$$\min_{w\in\mathbb{R}^d} f(w) = \frac{n_k}{n}\sum_{k=1}^{K} F_k(w) \tag{3}$$

where the local empirical loss function $F_k(w)$ is defined as:

$$F_k(w) = \frac{1}{n_k}\sum_{i\in\mathbb{P}_k} f_i(w) \tag{4}$$

Here, the $f(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$ defined in Equation (1) as a convex combination of the local empirical losses $F_k(w)$ available locally to node k.

In this federated setting, minimising the number of iterations in the optimisation algorithms is paramount of importance as there is limited communication capability of the local nodes. In the same paper, Konečnỳ *et al.* proposed a novel distributed gradient descent by combining the Stochastic Variance Reduced Gradient (SVRG) algorithm Johnson and Zhang (2013); Konečnỳ and Richtárik (2017) with the Distributed Approximate Newton algorithm (DANE) Shamir et al. (2014) for distributed optimisation called Federated SVRG (FSVRG) Konečnỳ et al. (2016a). The FSVRG computes gradients based on $\mathbb{P}_k$ data on each local node k, obtains a weighted average of the parameters from all the K local nodes, and updates new parameters for each node after the round. This algorithm is then experimented based on public Google+ posts, clustered by about 10,000 users as local nodes, for predicting whether a post will receive any comments. The results show that the FSVRG outperforms the native gradient descent algorithm as it converges to the optimum within only 30 iterations.

It is worth noting that standard distributed ML algorithms are generally designed to train independent identically/distributed (IID) data, and this assumption does not hold in federated settings due to the significant differences in the number of data samples and data distributions among personal mobile devices. Training over non-IID data has been shown to be much less accurate as well as slower convergence than IID data in federated settings Zhao et al. (2018). Konečnỳ with his colleagues at Google went further on improving the efficiency of the FSVRG algorithms in distributed settings by minimising the information in a parameter update to be sent to an orchestration server Konečnỳ et al. (2016b). Two types of updates are considered called *structured updates* and *sketched updates* in which the number of variables used in an ML model is minimised as many as possible, along with the compression of the information in the full model updates. Another ambitious federated optimisation approach is that local nodes are independently trained in different ML models as a task in a multi-learning objective simultaneously Smith et al. (2017). Generally, local nodes generate data under different distributions which naturally fit separate learning models; however, these models are structurally similar resulting in the ability to model the similarity using a multi-tasking learning (MTL) framework. Therefore, this approach improves performance when dealing with non-IID data as well as guarantees the learning convergence Smith et al. (2017).

Standing on these federated optimisation research works, McMahan *et al.* proposed a variation of the SGD called *FederatedSGD* along with the *Federated Averaging* algorithm that can

train a deep network at 100 times fewer communications compared to the naive FSVRG McMahan et al. (2017a, 2016). The catalyst of such algorithms is to leverage the increasingly powerful processors in modern personal mobile devices to perform high-quality updates than simply calculating gradient steps. Specifically, each client not only calculates the gradients but also computes the local model multiple times; the coordination server only performs aggregation of the local models from the clients. This results in fewer training rounds iterations (thus fewer communications) while producing a decent global model. These proposed algorithms well suited for scenarios that are highly limited communication bandwidth with high jitter and latency. In these scenarios, the naive FSVRG algorithms proposed in Konečnỳ et al. (2016a,b) are not efficient enough. Indeed, the algorithms are utilised for a real-world application for text prediction in Google keyboard in Android smartphones (i.e., G-board)[5] Yang et al. (2018). In this system setting, the *FederatedSGD* is executed locally on the smartphone to compute gradient descent using local data. The gradient is then sent to an aggregation server. This server performs the *FederatedAveraging* algorithm which randomly selects a fraction of smartphones for each training round, and takes the average of all gradients sent from the selected participants to update the global model. This updated global model is distributed to all participants; the local nodes will then update their local models accordingly.

3.3. *Centralised vs. decentralised architecture*

The architecture of a distributed learning and FL-based system can be centralised (e.g., *master-slave*) or decentralised (e.g., *ring*) Lian et al. (2017). In a centralised architecture, slaves (i.e., workers) only compute gradients; a master (i.e., a parameter server) obtains the parameters from all workers and disseminates the latest global parameters back to the workers to be updated in the next training round. This centralised distributed learning requires high-communication cost between workers and a server Dean et al. (2012). In a ring architecture, there is no centralised server to coordinate the parameter update; instead, each node both locally computes gradients and performs parameter aggregation by communicating with other nodes using a *Gossip* algorithm Daily et al. (2018); Koloskova et al. (2019); Ram et al. (2009). The ring architecture requires an efficient asynchronous updates strategy among compute nodes; otherwise, *model consistency* cannot be achieved Ben-Nun and Hoefler (2019); Lian et al. (2018).

Nevertheless, both centralised and decentralised architectures are required to acquire model consistency, particularly when data parallelism is employed. There are numerous strategies to update parameters in order to maintain the consistency of a global model, respected to a synchronisation model among compute nodes. In this regard, Asynchronous Parallel (ASP) Dean et al. (2012); Recht et al. (2011), Bulk Synchronous Parallel (BSP) Gerbessiotis and Valiant (1994), and Stale Synchronous Parallel (SSP) Ho et al. (2013) are the most common approaches to update parameters in a distributed learning system. The BSP and the ASP update parameters once receiving all gradients from a bulk of compute nodes (barrier synchronisation) and from just any node (no synchronisation), respectively. Generally, the BSP is relatively slow due to the stall time of waiting whereas ASP is faster as it does not perform any synchronisation; as a trade-off, the convergence in BSP is guaranteed but uncertain in the ASP Zhou et al. (2018). The SSP is as an intermediate solution balancing between the BSP and the ASP that performs relaxed synchronisation. In the SSP, compute nodes continue to the next training iteration only if it is not faster than the slowest node by β steps, (i.e., the progress gap between the fastest node and the slowest node is not too large), which guarantees the convergence although the number of iterations might be large. However, as a trade-off, the SSP introduce the β hyper-parameter which is non-trivial to be fine-tuned Ho et al. (2013).

3.4. *Federated learning workflow cycle*

Inspired by the research Bonawitz et al. (2016, 2017); Konečnỳ et al. (2016a,b); Konečnỳ and Richtárik (2017); McMahan et al. (2017a, 2016) and the real-world application (i.e., G-board) by the Google team, most of the existing FL-related research works have focused on the centralised FL framework (i.e., centralised FL) wherein an orchestration server plays as a controller requesting and aggregating training results to/from local nodes. However, it does not necessarily require a centralised server for reconstructing a global model; instead, local nodes can directly exchange their training results in a peer-to-peer manner (i.e., decentralised FL) He et al. (2018). This decentralised training approach requires a local updating scheme in which a synchronisation scheme among local nodes must be implemented Ferdinand et al. (2020); Reisizadeh et al. (2019) - which is not always feasible in federated settings. Research on decentralised FL is still in its early stage which is either restricted to simple learning models (e.g., linear models) or with the assumption of full or part synchronisation among participants He et al. (2018); Li et al. (2020).

In this paper, we examine the centralised FL in which there exists a centralised server (i.e., service provider) requests to coordinate the whole training process. Specifically, this coordination server (i) determines a global model to be trained, (ii) selects participants (i.e., local nodes) for each training round, (iii) aggregates local training results sent by the participants, (iv) updates the global model based on the aggregated results, (v) disseminates the updated model to the participants, and (vi) terminates the training when the global model satisfies some requirements (e.g., accurate enough). Local nodes passively train the model over their local data as requested, and send the training results back to the server whenever possible. The workflow cycle in a centralised FL framework consists of four steps (illustrated in Fig. 2) as follows:

1. *Participant Selection and Global Model Dissemination*: The server selects a set of participants that satisfy requirements to be involved in the training process. It then broadcasts a global ML model (or the global model updates) to the participants for the next training round.
2. *Local Computation*: Once receiving the global ML model from the server, the participants update their current local ML model and then trains the updated model using the local

[5] https://ai.googleblog.com/2017/04/federated-learning-collaborative.html.

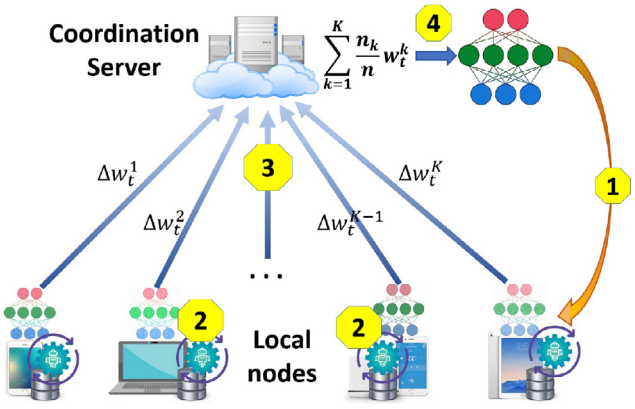

Fig. 2 – Workflow cycle in a centralised FL framework comprising of four steps.

dataset resided in the device. This step is operated at local nodes, and it requires end-users' devices to install an FL client program to perform training algorithms such as *FederatedSGD* and *Federated Averaging*, as well as to receive the global model updates and send the local ML model parameters from/to the server.

3. *Local Models Aggregation*: The server aggregates a sufficient number of the locally trained ML models from participants in order to update the global ML model (the next step). This aggregation mechanism is required to integrate some privacy-preserving techniques such as secure aggregation, differential privacy, and advanced encryption methods to prevent the server from inspecting individual ML model parameters.
4. *Global Model Update*: The server performs an update on the current global ML model based on the aggregated model parameters obtained in step 3. This updated global model will be disseminated to participants in the next training round.

This 4-step cycle is repeated until the global model has reached sufficient accuracy.

It is worth emphasising that the separation of the four steps in the cycle is not a strict requirement in every training round. For instance, an asynchronous SGD algorithm can be used in which results of the local training can be immediately applied to update the local model before obtaining updates from other participants Chen et al. (2016). This asynchronous approach is typically utilised in distributed training for deep learning models on a large-scale dataset as it maximises the rate of updates Chilimbi et al. (2014); Dean et al. (2012). However, in FL settings, the synchronous approach, which requires the coordination from a centralised server, has substantial advantages over the asynchronous ones in terms of both communication efficiency and security because it allows advanced technologies to be integrated such as aggregation compression, secure aggregation with SMC, and differential privacy Hardy et al. (2017); Konečnỳ et al. (2016b); McMahan et al. (2017a); Wang et al. (2019a).

3.5. *Applications*

Being able to train a global, united ML model on data from multiple participants without compromising the privacy and security of those training data; FL enables a variety of applications in smartphone services, healthcare industry, and financial services wherein the aggregation of data into a centralised data server is infeasible due to factors such as the restriction on data collection and transfer, intellectual property rights, as well as the rigour of complying with data protection regulations, e.g., the GDPR.

Personalised smart services is a prospective application of FL in which a variety of services can be customised in line with individual characteristics and preferences. A typical example of this application is the text prediction service for Google Android Keyboard Yang et al. (2018). Apple also utilises FL to improve Siri's voice recognition service in iPhone[6]. We believe FL can be employed to improve a variety of existing smart services including smart retail (e.g., product recommendation and sales services) and smart healthcare (e.g., daily activity, nutrition, sleep monitoring and recommendation). In these scenarios, miscellaneous types of data resided in end-users' smartphones and/or wearable devices can be utilised to boost the ML models in the existing services that serve the individual client the best.

Healthcare research and industry is also a potential domain that could greatly benefit from FL. Medical data such as patient information, disease symptoms, gene sequences, and different types of medical reports are dispersed in isolated clinical, medical centres and research institutes; and sharing such healthcare information is critical challenging with rigorous data protection regulations including the GDPR (in UK/EU) or HIPPA (in USA). Generally, it is impractical to fuse such data into one single data centre for ML training purposes. Indeed, the insufficiency of data samples have led to the poor performance of ML-based services and is the bottleneck that prevents the smart healthcare industry to reach its full potential. In this respect, FL enables a new technique to train ML model on a vast and varied medical dataset without the need for aggregating such e-health record; as well as further improves the performance of the conventional-trained ML models. This would open a new opportunity in the development of smart healthcare and might take it to a whole new level.

4. Privacy-Preservation in centralised federated learning framework

As an ML model can be cooperatively trained while retaining training data and computation on-device, FL naturally offers privacy-guarantee advantages compared to the traditional ML approaches. Unfortunately, although personal data is not directly sent to a coordination server in its original form, the local ML model parameters still contain sensitive information because some features of the training data samples are inherently encoded into such models Aono et al. (2017); Ateniese et al. (2015); McMahan et al. (2016);

[6] https://www.technologyreview.com/2019/12/11/131629/apple-ai-personalizes-siri-federated-learning/.

Melis et al. (2019); Phong et al. (2018). For example, authors in Ateniese et al. (2015) have shown that during the training process, correlations implied in the training data are concealed inside the trained models, and personal information can be subsequently extracted. Melis *et al.* have also pointed out that modern deep-learning models conceal internal representations of all kinds of features, and some of them are not related to the task being learned. Such *unintended features* can be exploited to infer some information about the training data samples. FL systems, consequently, is vulnerable to *inference attacks* (i.e., membership and reconstruction attacks Dwork et al. (2017)).

Furthermore, local nodes not only passively contribute local training results but also get updated about intermediate stages of a global training model from a coordination server. This practice enables the opportunity for malicious participants to manipulate the training process by providing arbitrary updates in order to poison the global model Bhagoji et al. (2019); Fung et al. (2018), which calls for an investigation on security models along with insightful analysis of privacy guarantees for a centralised FL framework. Accordingly, the FL framework then needs to be strengthened by employing further privacy and security mechanisms to protect personal data effectively and to comply with intricate data protection legislation like the GDPR. A summary of related articles in terms of attack models with associated privacy preservation methods in centralised FL is depicted in Table 1. Detailed descriptions along with analysis are carried out in the following sub-sections.

4.1. *Attack models on FL*

4.1.1. *Inference attacks on FL*

As aforementioned, a trained ML model contains unintended features that can be utilised to extract personal information. Thus, local ML model parameters from a federated optimisation algorithm can be exploited by an adversary to infer personal information, particularly when combining with related information such as model data structure and meta-data. This information can be either original training data samples (i.e., *reconstruction attack*) Aono et al. (2017); Bagdasaryan et al. (2020); Fredrikson et al. (2015); Geiping et al. (2020); Hitaj et al. (2017); McMahan et al. (2016); Nasr et al. (2018); Phong et al. (2018); Shokri and Shmatikov (2015); Shokri et al. (2017); Zhu et al. (2019) or *membership tracing* (i.e., to check if a given data point belongs to a training dataset) Bonawitz et al. (2017); Melis et al. (2019); Shokri et al. (2017).

Attackers might carry out model inversion (MI) attack to extract sensitive information contained in training data samples, for instance, by reconstructing representatives of classes which characterising features in classification ML models Fredrikson et al. (2015). MI attacks do not require the attacker to actively participate in the training process (i.e., black-box or passive attacks). For example, it is possible to recover images from a facial recognition model for a particular person (i.e., all class members depict this person) using MI by deriving a correct weighted probability estimation for the target feature vectors Geiping et al. (2020); Shokri et al. (2017). In this scenario, the experiment results show that this MI attack can reconstruct images that are visually similar to the victim's photos Fredrikson et al. (2015).

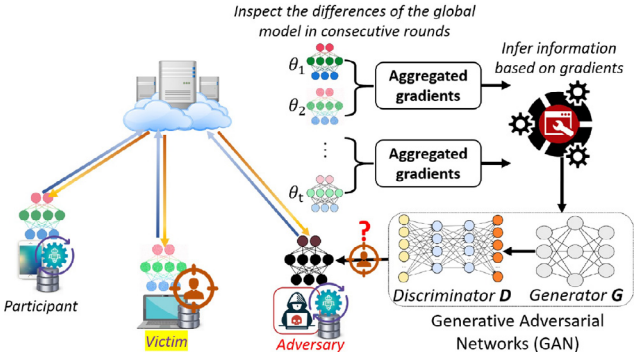

Fig. 3 – High-level concept of inference attacks against FL based on GANs.

In an FL framework, attackers are not only able to observe the trained model parameters but also participate in the training process to inspect the changes in the updated global models in some consecutive training rounds (i.e., white-box or active attacks), which will intensify the attack (Fig. 3). It is shown that MI attacks based on class representation are more challenging than reconstructing from gradients for classification models Geiping et al. (2020). In this regard, numerous reconstruction attacks were proposed based on Generative Adversarial Networks (GANs) Goodfellow et al. (2014); Salimans et al. (2016) to synthesise fake samples which have the same statistics (e.g., distribution) to those in the training set without having access to the original data. For instance, Hitaj *et al.* based on GANs have developed an attack at the user-level which allows an insider to infer information from a victim just by analysing the shared model parameters in some consecutive training rounds Hitaj et al. (2017). This attack can be accomplished at the client-side without interfering the whole FL procedure, even when the local model parameters are obfuscated using the DP technique. A malicious coordination server can also recover partial personal data by inspecting the proportionality between locally trained model parameters sent to the server and the original data samples Aono et al. (2017); Wang et al. (2019b).

Reconstruction attacks using MI and GANs are only feasible if and only if all class members in an ML model are analogous which entails a similarity between the MI/GAN-reconstructed outputs and the training data (e.g., facial recognition of a specific person or MNIST dataset for handwritten digits[7] used in Aono et al. (2017)). Fortunately, this precondition is less practical in most FL scenarios.

However, it is not necessary to fully reconstruct the trained data; instead, inferring attributes or membership of the original trained data from local model parameters can also induce serious privacy leakage Ganju et al. (2018); Melis et al. (2018, 2019); Nasr et al. (2018, 2019) (e.g., an attacker can figure out whether a specific data sample (of a patient) is used to train a model of a disease). This is the baseline for the membership attack. Authors in Melis et al. (2018, 2019);

[7] http://yann.lecun.com/exdb/mnist/.

Table 1 – Summary of Attack Models vs. Privacy Preservation Methods in centralised FL.

Attack Models		Privacy-preserving Techniques employed at Server-side	Privacy-preserving Techniques employed at Client-side
Inference Attacks	Reconstruction Attacks Aono et al. (2017); Bagdasaryan et al. (2020); Fredrikson et al. (2015); Hitaj et al. (2017); McMahan et al. (2016); Phong et al. (2018); Shokri and Shmatikov (2015); Shokri et al. (2017) Geiping et al. (2020); Goodfellow et al. (2014); Nasr et al. (2018); Salimans et al. (2016); Wang et al. (2019b); Zhu et al. (2019)	SMC & Secure Aggregation Bonawitz et al. (2019, 2016, 2017); McMahan et al. (2017a, 2016, 2017b) Homomorphic Encryption Phong et al. (2018); Salem et al. (2019)	SMC & Secure Aggregation Bonawitz et al. (2019, 2016, 2017); McMahan et al. (2017a, 2016, 2017b); Pathak et al. (2010) Homomorphic Encryption Phong et al. (2018); Salem et al. (2019) Batch-level DP Abadi et al. (2016); Hitaj et al. (2017); Pathak et al. (2010); Shokri and Shmatikov (2015) User-level DP Bhowmick et al. (2018); Geyer et al. (2017); Hitaj et al. (2017); McMahan et al. (2017b); Pathak et al. (2010); Sun et al. (2020)
Membership Tracing Aono et al. (2017); Bonawitz et al. (2017); Goodfellow et al. (2014); Melis et al. (2018, 2019); Nasr et al. (2019); Salimans et al. (2016); Shokri et al. (2017); Wang et al. (2019b)			
Poisoning	Data Poisoning Biggio et al. (2012); Chen et al. (2017); Jagielski et al. (2018); Koh and Liang (2017); Mei and Zhu (2015); Xiao et al. (2015) Model Poisoning Bagdasaryan et al. (2020); Bhagoji et al. (2019); Blanchard et al. (2017); Chen et al. (2018); Fung et al. (2018); Mhamdi et al. (2018)	Model Anomaly Detection* Fung et al. (2018); Jagielski et al. (2018) ***This solution is not feasible****if Secure Aggregation is employed*	***None***

Nasr et al. (2019) have investigated membership attacks in FL and demonstrated the capability of these attacks in both passive and active approaches. For instance, the gender of a victim can be inferred with very high accuracy of 90% when conducting this attack in a binary gender classifier on the FaceScrub dataset[8]. Other features, which are uncorrelated with the main task, can also be inferred such as race and facial appearance (e.g., whether a face photo is wearing glasses) Melis et al. (2019). Nasr *et al.* proposed an active attack approach called *gradient ascent* by exploiting the privacy vulnerabilities of SGD optimisation algorithms. This attack based on the correlation between the local gradients of the loss and the direction and the amount of changes of model parameters when minimising the loss to fit a model to train data samples in the SGD algorithms. This active membership attack was conducted on the CIFAR100 dataset[9] showing a high accuracy of 74% compared to only 50% in passive attack Nasr et al. (2018, 2019).

4.1.2. *Poisoning attacks on FL*

One of the privacy-preserving objectives of centralised FL is that a coordination server is unable to inspect the data or administer the training process at a local node. This, however, prohibits the transparency of the training process; thus, imposes a new vulnerability of a new type of attack called *model poisoning* Bagdasaryan et al. (2020); Bhagoji et al. (2019); Blanchard et al. (2017); Chen et al. (2018); Fung et al. (2018); Mhamdi et al. (2018). Generally, model poisoning attacks aim at manipulating the training process by feeding poisoned local model updates to a coordination server. This type of

[8] http://vintage.winklerbros.net/facescrub.html.

[9] https://www.cs.toronto.edu/~kriz/cifar.html.

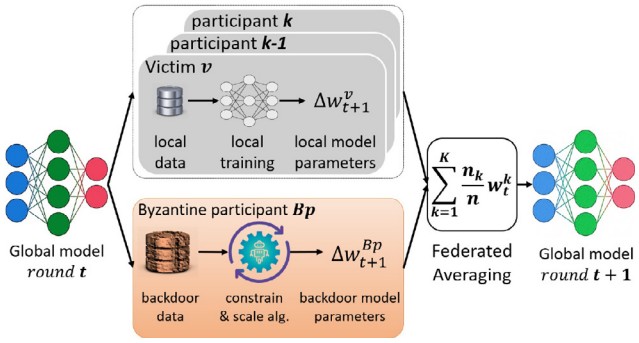

Fig. 4 – High-level concept of model poisoning using backdoor attack against FL.

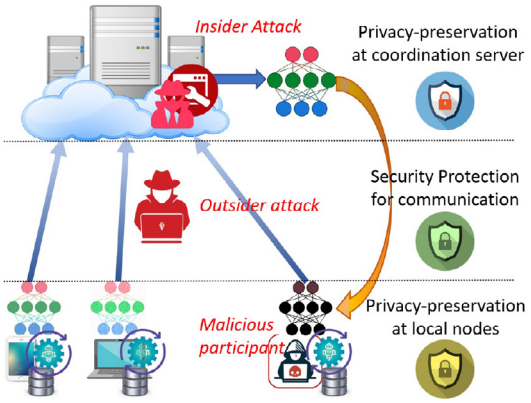

Fig. 5 – Overview of the Privacy and Security employed in a centralised FL framework.

attack is different from *data poisoning* Biggio et al. (2012); Chen et al. (2017); Jagielski et al. (2018); Koh and Liang (2017); Mei and Zhu (2015); Xiao et al. (2015), which is less effective in FL settings Bagdasaryan et al. (2020); Bhagoji et al. (2019) because the original training data is never shared with a server. Thus, this section is mainly dedicated to analysing the model poisoning attacks in FL.

Generally, model poisoning is conducted at the client-side wherein an adversary controls a fraction of participants for a common adversarial goal, either (*i*) corrupting the global model so that it converges to a *sub-optimal* which is an incompetent, ineffective one (i.e., random attack) Blanchard et al. (2017); Chen et al. (2018); Mhamdi et al. (2018), or (*ii*) replace it to a targeted model (i.e., replacement attack) Bagdasaryan et al. (2020); Bhagoji et al. (2019).

Poisoned model parameters sent to a coordination server can be generated by injecting a hidden backdoor model intentionally, as illustrated in Fig. 4. Compromised participants analyse the targeted global model; the poisoned model is then trained on backdoor data samples using dedicated techniques such as *constrain-and-scale* accordingly, and feed the parameters to a coordination server as other honest participants. The objective of this attack is that the global model is replaced by a joint model consisting of both the original task and the injected backdoor sub-task while retaining high accuracy on the two. The backdoor training at the adversary can be empowered by modifying minimisation strategies such as *constrain-and-scale*, which optimises both gradients of the loss and the backdoor objective Bagdasaryan et al. (2020). A parameter estimation mechanism is then used for generating parameters submitted to the coordination server for honest participants' updates. As secure aggregation is used for preventing the server from inspecting individual models, this poisoning model is unable to detect Bagdasaryan et al. (2020); Bhagoji et al. (2019).

4.2. *Threat model in a centralised FL framework*

As the target of both inference and model poisoning attacks, a centralised FL framework needs to be well designed to withstand potential adversaries. As illustrated in Fig. 5, the security and privacy threats are classified into three categories: (1) Threats at the coordinator server by insider attackers, (2) Threats at communication medium by outsider attacker, and (3) Threats due to malicious participants.

4.2.1. *Malicious coordination server*

The coordination server is assumed to be malicious as there exist insider attackers who can carry out inference attacks to infer information of a target client illegitimately. These attacks are feasible at the server-side by analysing periodic parameters updates obtained from related local nodes including the victim (i.e., passive attack), or even purposely requesting the victim to train modifying models with adversarial influence (i.e., active attack) Wang et al. (2019b).

4.2.2. *Secure communication medium*

It is assumed that the communication medium for information exchange between local nodes and a coordination server is secure regardless the information is in plaintext McMahan et al. (2017a) or encrypted Mohassel and Zhang (2017). Secure communications protocols such as SSL/TLS and HTTPS are readily integrated into the FL framework to prevent man-in-the-middle attacks, eavesdropping and tampering. Thus, in a centralised FL framework, privacy and integrity of the exchanged information are assured while in transit.

4.2.3. *Byzantine participants*

In most FL scenarios, local nodes are assumed to be malicious, meaning that there is a possibility that there exists an adversary controlling a fraction of local nodes to perform *model poisoning*. Moreover, such malicious participants might operate in a Byzantine fashion, meaning that they send arbitrary training model updates to shape the global model in a targeted manner (i.e., either demolish the global model or be replaced by a vicious one).

Furthermore, an inference attack can also be carried out by a malicious participant as the adversary can commit its local update and observe the changes in the updated global model Melis et al. (2019). Instead, the active inference attack is only accomplished by a malicious server.

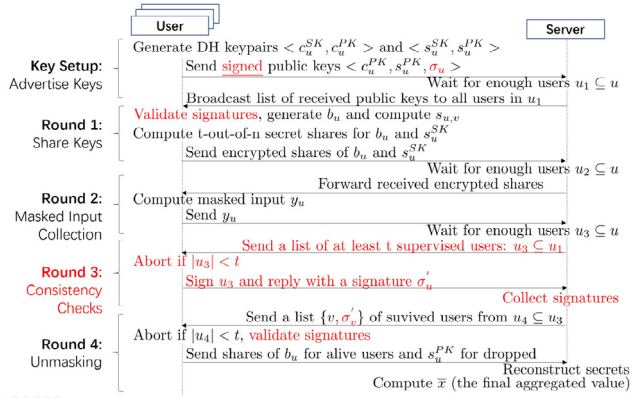

Fig. 6 – Sequence diagram of the Secure Aggregation. Red-color processes are required to guarantee the security of the protocol against malicious server and participants Bonawitz et al. (2017).

4.3. *Privacy-preservation solutions for coordination server*

Most of the existing privacy-preserving techniques for FL systems are built upon advanced cryptographic protocols, including SMC and differential privacy. On the server-side, such techniques are employed in order to (*i*) prevent insiders at the server from conducting inference attacks, and (*ii*) prevent Byzantine participants from conducting model poisoning.

4.3.1. *Inference attacks prevention*

Several solutions have been proposed to tackle inference attacks at the server-side following the same purpose of preventing the coordinate server from inspecting parameters sent from a particular user during the global model training process. Specifically, in the aggregation process, parameters sent from the clients (*gradients* in Federated SGD or *local model weights* in Federated Averaging) can be protected based on SMC called Secure Aggregation protocol, first proposed by Bonawitz in Bonawitz et al. (2016, 2017). The baseline of the protocol is SMC in which cryptography techniques are leveraged that enable participants to jointly compute the average of the model parameters without revealing their inputs. As illustrated in Fig. 6, the protocol comprises of four interactive rounds between participants and a coordinate server including public-keys advertisement and sharing (round 1), masked inputs computation at client-side once getting an independent response from the server (round 2), consistency check that the model has at least t participants involved in the training process (round 3), and unmasking once at least t participants reveal sufficient cryptographic secrets so that the coordination server is able to unmask the global model update (round 4). Round 3 of the protocol is required if the server is malicious but not necessary for an honest-but-curious one. As a trade-off, this protocol results in increasing communication overhead and computation complexity at both clients and a coordination server. It is worth noting that the Secure Aggregation protocol has already been integrated into the TensorFlow Federated framework[10], developed by Google Bonawitz et al. (2019), to facilitate research and real-world experimentation with FL.

Secure Aggregation protocol is based on the fact that it only requires to calculate the averages of the local model weights from a random subset of participants to perform SGD and compute global model updates. The coordination server, thus, does not need to acquire local updates from individual participants. This would prevent the server from observing individual users and carrying out inference attacks. Along with Federated Averaging, Secure Aggregation protocol facilitates secure SGD execution with robustness to failures and less communication overhead in a server with limited trust. However, this SMC-based technique only works effectively in scenarios of honest participants. There is no guarantee for the availability and correctness of the protocol in the case of Byzantine participants, particularly when such Byzantine participants collude with the malicious server to disclose inputs of a targeted client. In case of the client-server collusion, the protocol can only tolerate up to $[\frac{n}{3}] - 1$ Byzantine participants whereas the number of total participants involved in the training process should be at least $[\frac{2n}{3}] + 1$, ensuring the robustness up to $[\frac{n}{3}] - 1$ dropping out participants Bonawitz et al. (2017).

4.3.2. *Model poisoning prevention at server-side*

Model poisoning attacks are always inherent in collaborative learning including FL. As shown by Bagdasaryan *et al.* in Bagdasaryan et al. (2020), just by controlling less than 1% Byzantine participants, an adversary can successfully insert a backdoor functionality into a global model without reducing much accuracy, preventing the coordination server from detecting the attack. Solutions to mitigate model poisoning attack at the server-side have to detect and filter out poisoned model updates from malicious clients (i.e., model anomaly detection) Fung et al. (2018); Jagielski et al. (2018). For this purpose, the server needs to access either participants training data or parameter model updates, which breaks the privacy-preservation catalyst of FL. Besides, Secure Aggregation protocol is assumed to be implemented at both client- and server-side, which prevents the server from inspecting individual model updates; consequently, ruling out any solutions for model poisoning attacks Fung et al. (2018). Indeed, no resolutions have been proposed that effectively tackle model poisoning attacks at the server-side, which imposes as a critical research topic for FL.

4.4. *Privacy-preservation solutions for local nodes*

Local nodes, along with a coordination server, should implement Secure Aggregation protocol to mitigate the risk of privacy leakage in case there exists an inside attacker carrying out inference attacks at the server Bonawitz et al. (2016, 2017). This SMC-based aggregation protocol can also be strengthened with Homomorphic Encryption to encrypt local model parameters from all participants for secure multi-party deep learning in FL settings Zhang et al. (2017). The coordination server, hence, receives an encrypted global model which can

[10] https://www.tensorflow.org/federated.

only be decrypted if and only if a sufficient number of local models have been aggregated. As a result, the privacy of individual contributions to the global model is guaranteed.

Furthermore, the local nodes can leverage the perturbation method to prevent a coordination server and other adversaries from disclosing model parameters updates and original training dataset. The idea of employing the perturbation technique to FL is that a local node adds random noise to its local model parameters in order to obscure certain sensitive attributes of the model before sharing. As a result, adversaries, in case it can successfully derive such model parameters, is unable to accurately reconstruct the original training data or infer some related information. In other words, the perturbation method could prevent adversaries from carrying out inference attacks on a local model trained by a particular client. This privacy-preservation method typically adopts differential privacy technique that adds random noises to either training dataset or model parameters, offering statistical privacy guarantees for individual data Bassily et al. (2014); Dwork (2008); Dwork et al. (2014). Indeed, before the proposal of FL, differential privacy with SMC has been suggested as a privacy-preserving technique for the aggregation of independently trained neural networks in Pathak et al. (2010). Since then, this technique has been improved to return statistically indistinguishable results among participants while ensuring that such noise-added model parameters do not affect much on the accuracy of the global model in FL settings Abadi et al. (2016); Aono et al. (2017); Geyer et al. (2017); McMahan et al. (2017b); Shokri and Shmatikov (2015); Song et al. (2013). As a consequence, adversaries cannot distinguish individual records in the FL training process and do not know whether or not a targeted client participating in the training; thus, preserving data privacy and protecting against inference attacks. Generally, there are two types of employing differential privacy techniques for local nodes in FL settings: *batch-level* and *user-level* where random noise is added by measuring parameters' sensitivity from data points in a mini-batch and users themselves, respectively.

4.4.1. *Batch-level differential privacy approach*

Shokri and Shmatikov in Shokri and Shmatikov (2015) have proposed a communication efficient privacy-preserving SGD algorithm for deep learning in distributed settings in which local gradient parameters are asynchronously shared among participants with an option of adding noise to such updates for the differentially private protection of the individual model parameters. In this algorithm, participants can choose a fraction of parameters (randomly selected or following a strategy) to be updated at each round so that their local optimal can converge faster while being more accurate. In order to integrate differential privacy technique into the algorithm, the ε total privacy budget parameter and the sensitivity of gradient Δf_i for each parameter f_i are taken into account to control the trade-off between the differential privacy protection and the model accuracy. *Laplacian mechanism* is used to generate noise during both parameter selection and exchange processes based on the estimation of the Δf_i sensitivity and the allocated ε privacy budget. The proposed algorithm has experimented on *MNIST* and *SVHN* datasets showing the trade-off between strong differential privacy guarantees and the high accuracy of the training model. However, with a large number of participants sharing a large fraction of gradients, the accuracy of the proposed algorithm with differential privacy is better than the standalone baseline. It is worth noting that in this algorithm, local gradients can be exchanged directly or via a central server, which can feasibly be implemented in the FL settings.

The authors in Abadi et al. (2016) have proposed an SGD algorithm integrated with differential privacy performing over some batches (a group) of data samples. This algorithm estimates the gradient of the group by taking the average of the gradient loss of these batches and adds noise (generated by *Gaussian mechanism*) to the group to protect privacy. This algorithm is implemented to train on the *MNIST* and *CIFAR-10* datasets showing sensible results as it achieves only 1.3% and 7% less accurate compared to the non-differentially private conventional baseline algorithms on the same datasets, respectively. Similar to the mechanism proposed by Shokri and Shmatikov in Shokri and Shmatikov (2015), the authors have proposed a mechanism to monitor the total privacy budget (i.e., privacy accounting) as accumulated privacy loss by observing privacy loss random variables. Based on the experiment, the authors also indicate that privacy loss is minimal for a large group size (with a large number of datasets).

4.4.2. *User-level differential privacy approach*

Geyer *et al.* in Geyer et al. (2017) have developed another method to implement differential privacy for federated optimisation in FL settings that conceals the participation of a user in a training task; as a result, the whole local training dataset of the user is protected against differential attacks. This approach is different from the *batch-level* one, which aims at protecting a single data point in a training task. The proposed method utilises a similar concept of privacy accounting from Abadi et al. (2016) that allows a coordination server to monitor the accumulated privacy budget by observing the moment accountant and privacy loss proposed in Abadi et al. (2016). The training process is halted once the accumulated privacy budget reaches a pre-defined threshold, implying that the privacy guarantee is no further tolerated. The Gaussian mechanism is also used to generate random noise which is then added to distort the sum of gradients updates to protect the whole training data. The proposed method has experimented on *MNIST* dataset, and the results show that with a sufficiently large number of participants (e.g., about 10,000 clients), the accuracy of the FL trained model almost achieves as high as the non-differential-privacy baseline while a certain level of privacy guarantee over the local training data still holds.

Similarly, McMahan *et al.* in McMahan et al. (2017b) have leveraged the privacy accounting and moment privacy proposed in Abadi et al. (2016) to integrate *user-level* differential privacy into a federated averaging mechanism previously proposed in McMahan et al. (2016) in order to protect local model parameters sharing with a coordination server. The proposed mechanism is a noise-added version of the federated averaging algorithm in FL which was deployed to train deep recurrent models like Long Short-Term Memory (LSTM) recurrent neural networks (RNNs). They have implemented the mecha-

nism to train the LSTM RNNs tuned for language modelling in a mobile keyboard. The experimental results indicate that the integration of differential privacy only causes a minor effect on predictive accuracy; however, it could induce a qualitative effect on word predictions and tends to bias the model away from uncommon words. This potential bias in the mechanism calls for further research on adaptive tuning mechanisms for the clipping and noise in order to balance between utility and privacy in FL. Bhowmick *et al.* in Bhowmick et al. (2018) and Sun *et al.* in Sun et al. (2020) have also proposed similar *user-level* differential privacy in FL settings with some improvements such as employing a better estimation on total privacy budget (in Bhowmick et al. (2018)), and adding a *splitting & shuffling* mechanism for local model parameters before sending to a coordination server (in Bhowmick et al. (2018)).

As aforementioned, Hitaj *et al.* have successfully carried out inference attacks at the client-side based on GAN Hitaj et al. (2017). In this paper, they have also shown that an FL training task with differential privacy employed at *batch-level* is still susceptible to the attacks; however, the *user-level* differential privacy approach could protect against such attacks.

5. GDPR-Compliance In centralised federated learning systems

FL emerges a new approach to tackle data privacy challenges in ML-based applications by decoupling data storage and processing (i.e., local model training) at end-users' devices (i.e., local nodes) and the aggregation of a global ML model at a service provider's server (i.e., a coordination server). The privacy-preservation advantage of FL compared to the traditional centralised ML approaches is undeniable: It enables to train an ML model whilst retaining personal training data on end-users' devices. Only locally trained model parameters, which contain the essential amount of information required to update the global model, are shared with a coordination server. Nevertheless, such model parameters still enclose some sensitive features that can be exploited to reconstruct or to infer related personal information as depicted in *Section* 4. Subsequently, an FL system still retains within the GDPR and is liable for complying with obligatory requirements. This section closely examines whether a GDPR requirement should be complied with or inapplicable and should be waived in FL settings. Unsolved challenges on fully complying with the GDPR are also determined and discussed.

5.1. *Roles and obligations*

The GDPR differentiates three participant roles, namely Data Subject, Data Controller and Data Processor, and designates associated obligations for these roles under the EU data protection law. Data Controllers are subject to comply with the GDPR by determining the purposes for which, and the method in which, personal data is processed by Data Processors - who will be responsible for processing the data on behalf of Data Controllers. Furthermore, Data Controllers should take appropriate measures to provide Data Subjects with information related not only to how data is shared but also to how data is processed in the manner ensuring security and privacy of personal data. The GDPR also clearly specifies the rights of Data Subjects, giving data owners the rights to inspect information about how the personal data is being processed (e.g., Right to be informed and Right of access) as well as to fully control the data (e.g., Right of rectification and erasure, and Right to restriction of processing).

Table 2 – GDPR Roles in traditional centralised ML-based and centralised FL-based applications and services.

GDPR Roles	Traditional ML-based services	Centralised FL-based services
Personal Data	Original training data	Local model parameters
Data Subject	End-users	End-users
Data Controller	Service Provider	Service Provider
Data Processor	Service Provider, Third-parties	Service Provider

As depicted in Table 2, in FL settings, personal data is regarded as local model parameters, not the original data samples as in traditional cloud-based ML systems. A service provider, who implements an FL system, is Data Controller and Data Processor combined as the service provider (*i*) dictates end-users (i.e., Data Subject) to train an ML model using their local training data and to share such locally trained model, (*ii*) processes the local model parameters sent from end-users (i.e., aggregates and updates the global model), and (*ii*) disseminates the global models to all end-users and requests the end-users to update their local models. Furthermore, in centralised FL settings, a service provider can only share a global ML model, which can be considered as anonymous information, with third-parties as it does not possess any other personal data (e.g., original training data as in traditional ML systems). Therefore, Data Processors in FL settings are also the service providers, but not other players (i.e., third-parties). The processing mechanisms in FL are also uncomplicated compared to the traditional ones as they are only related to the aggregation of the local ML models as well as the update of the global ML model.

5.2. *GDPR Principles*

The GDPR defines 6-core principles as rational guidelines for service providers to manage personal data as illustrated in Fig. 7 (The GDPR Articles 5–11). These principles are broadly similar to the principles in the Data Protection Act 1998 with the accountability that obligates Data Controllers to take responsibility for complying with the principles and implementing appropriate measures to demonstrate compliance.

5.2.1. *Lawfulness, fairness and transparency*

According to the first principle, a service provider providing an FL application, as a Data Controller, must specify its legal basis in order to request end-users to participate in the FL training. There are a total of six legal bases required by the GDPR namely (1) Consent, (2) Contract, (3) Legal Obligation, (4) Vital Interest, (5) Public Task, and (6) Legitimate Interest (defined in

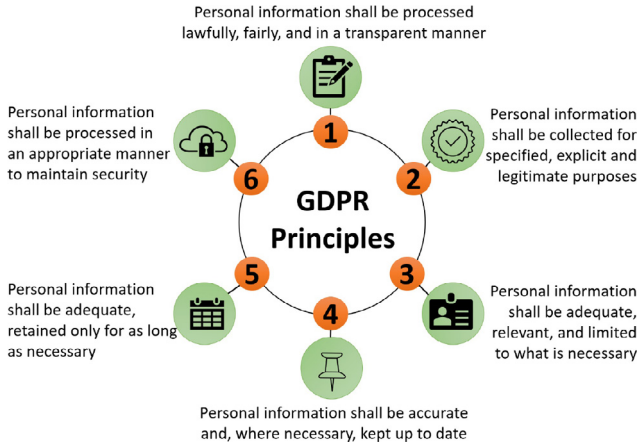

Fig. 7 – 6-core principles in GDPR.

of the GDPR in detail). These lawful bases might need to come along with other separate conditions for lawfully processing some special category data including healthcare data, biometric data, racial and ethnic origin. Depending on the specific purposes and context of the processing, the most appropriate one should be determined and documented before starting to process personal data.

To ensure privacy, an FL system is designed in a way that does not let the service provider (i.e., the coordination server) directly access and obtain either original training data or locally trained ML models at end-users' devices. Instead, end-users, as participants in the FL system, will only send the results back to the coordination server when they are ready. An FL client-side application should offer several options for clients to participate in the training process proactively that allows a client to fully control the local training as well as of the sending/receiving ML model updates to/from a coordination server. Furthermore, FL systems only process data (i.e., local ML model parameters) for an explicit purpose (i.e., aggregates results and updates a global model), which is in ways that clients would reasonably expect whilst having minimal privacy impact. For these reasons, either *Consent* or *Legitimate Interest* legal basis can be appropriate for an FL application[11].

Regarding the *Fairness* and *Transparency* requirements, as AI/ML algorithms like deep learning are normally operated in a black-box fashion, it is limited of transparency of how certain decisions are made, as well as limited understanding of the bias in data samples and training process Ananny and Crawford (2018); Doshi-Velez and Kim (2017); Mehrabi et al. (2019); Murdoch et al. (2019). An FL system is not an exception. Generally, if the training data is poorly collected or intentionally prejudicial and fed to an ML, including FL, system, the results turn out to be biased. If the trained model is then utilised for an automated decision-making system, then it probably leads to discrimination and injustice. Furthermore, the nature of preventing service providers from accessing the original training dataset as well as the inability to inspect individuals' locally trained ML model due to the Secure Aggregation mechanism amplifies the lack of transparency and fairness in FL systems. As a result, an FL system finds it problematic to transparently execute the training operations as well as to ensure any automated decisions from the system are impartially performed. This, consequently, induces impracticality for any FL systems and fails to fully comply with the GDPR requirements of fairness and transparency.

These unsolved challenges appoint much more research on transparency, interpret-ability and bias for AI/ML algorithms as well as demand the GDPR supervisory boards to relax the requirements on AI/ML including FL systems. Another promising solution to comply with this GDPR principle is to design a new type of ML models with associated algorithms that are inherently interpretable, which encourages responsible ML governance Harder et al. (2020); Li et al. (2017); Molnar (2020); Rudin (2019).

5.2.2. *Purpose limitation*

This purpose limitation principle can be interpreted that an FL service provider needs to clearly inform clients about the purpose of a global ML model training as well as how clients' local personal data and devices' computation are used to locally train a requested ML model provided by the service provider. The principle also states that the service provider can further process the data for other compatible purposes. In this respect, FL systems fully satisfy with the principles if sufficient privacy-preserving mechanisms such as Secure Aggregation and differential privacy are readily implemented into the systems. This is because locally trained ML models from clients are aggregated only for the global model updates and cannot be individually extracted and exploited (by the coordination server) for other purposes.

However, as described in *Section* 4.1, a malicious service provider or Byzantine participants can inject a hidden backdoor model for an unauthorised training purpose. Currently, there is no solution for any model anomaly detection mechanism at the server-side for this type of attack due to the use of secure aggregation in centralised FL; this, as a consequence, remains an unsolved challenge for an FL system to fully comply with the GDPR.

5.2.3. *Data minimisation*

The data minimisation principle in the GDPR necessitates a Data Controller (e.g., a service provider) to collect and process personal data that is adequate, limited, and relevant only to claimed purposes. In traditional centralised ML algorithms, this data minimisation requirement is a challenge as it is not always possible to envision what data and the minimal amount of data are necessary for training an ML model. In this regard, FL appears as a game-changer as an FL system does not need to collect and process original training data; instead, a service provider only needs to gather local ML models from participants for assembling the global model. Generally, with privacy-preserving techniques introduced in *Section* 4, an FL system can assure that the coordination server obtains aggregated local model parameters from participants for global model updates only (i.e., the claimed purposes) while acquiring nothing about an individual's contribution. The aggregation mechanism also assures that the global model itself con-

[11] https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/.

tains no individual sensitive features that can be exploited by adversaries to extract or infer any personal information.

Similar to the purpose limitation principle, back-door attacks are feasibly carried out to inject an unauthorised purpose. In this scenario, local ML model parameters obtained from participants is no longer minimal for the original purpose but also another unauthorised sub-task. This injected sub-task might be exploited to expose the personal information of the participant, imposing an unsolved challenge for FL systems.

5.2.4. *Accuracy*

The purpose of this principle is to ensure that a Data Controller should keep personal data correctly, updated, and not misleading any matter of fact. In centralised FL settings, a coordination server does not store any individual locally trained ML model parameters except the aggregated results from a batch of participants, and the anonymised global ML model. This information is stored and processed (i.e., for updating the global model) in its original form without any changes, and updated for every training round. For these reasons, FL systems automatically satisfy the GDPR accuracy principle.

5.2.5. *Storage limitation*

This principle ensures that a Data Controller does not keep personal data for longer if the data is no longer needed for the claimed purposes. In this case, data should be erased or anonymised. There is an exception for data retention only if the Data Controller keeps the data for public interest archiving, scientific or historical research, or statistical purposes.

Regarding the centralised FL settings, an FL system implementing Secure Aggregation does not store any individual ML model updates from participants except the global model - which can be assured to contain no individual sensitive features to be exploited for inference attacks. Even in the case of a malicious server holding aggregated contributions from many FL training rounds for further analytic (e.g., inference attacks), with secure aggregation and differential privacy integration, such aggregated information is protected and pseudo-anonymised. In other words, an FL system with appropriate privacy-preserving mechanisms can be fully compliant with the storage limitation principle.

5.2.6. *Integrity and confidentiality (security)*

This principle obligates Data Controllers to implement appropriate measures in place to effectively protect personal data. Thus, in order to comply with this principle, a centralised FL system requires to implement security and privacy mechanisms not only at a coordination server but also at end-users' devices as the FL system itself does not guarantee security and privacy.

Along with the privacy-preserving techniques such as Secure Aggregation, differential privacy, and Homomorphic Encryption designated for protecting local ML parameters, the FL client application installed at end-users' devices must be secure to prevent unauthorised access, cyber-attack, or data breach directly from the devices or from the communications between the users' devices and a coordination server. This precondition is the same as any other systems in which a variety of security and privacy techniques are readily integrated into FL applications, as well as secure communications protocols such as IPSec, SSL/TLS and HTTPS to protect data in transit between clients and the server.

5.3. *Rights of data subject*

The GDPR requires Data Controllers to provide the following rights for Data Subjects if capable (The GDPR Articles 12–23): (1) Right to be informed, (2) Right of access, (3) Right to rectification, (4) Right to erasure (Right to be forgotten), (5) Right to restrict processing, (6) Right to data portability, (7) Right to object, and (8) Rights in relation to automated decision making and profiling.

5.3.1. *Right to be informed*

The challenge to provide this right to Data Subjects is that the GDPR demands the Data Controller to concisely, intelligibly, and specifically specify what and how the local ML model is used in the FL training, along with expected outputs of the mechanism[12]. Same as many complex ML mechanisms, FL is as a black-box model; thus, it cannot be precisely interpreted of how it works and predicting the outcomes. The GDPR supervisory board recognises the challenges and relaxes the requirement for AI/ML mechanisms by accepting a general explanation as an indication of how and what personal data is going to be processed. As a result, for an FL system, the right to be informed is achieved as *privacy information* including purposes for processing local ML model (i.e., to build a global ML model), retention periods (i.e., no longer in use after each training round), and who it will be shared with (only the coordination server) can be determined as in Terms and Conditions when a client accepts to participate in an FL system.

5.3.2. *Rights in relation to automated decision making and profiling*

A Data Subject is assumed to have the right "not to be subject to a decision based solely on automated processing, including profiling" - (1), the GDPR. Therefore, an FL client, as a Data Subject, has the right to receive meaningful information and explanation about whether the result of the processing (i.e., a global ML model) used in an automated decision-making system will produce legal effects concerning the client or similarly significantly affects the client. Unfortunately, due to the black-box operation model and the limitation of the transparency in ML, including FL, training process, results (e.g., a global ML model in FL) are generally generated without any proper explanation Wachter et al. (2017). Thus, it is infeasible to predict whether outcomes of an ML model might affect the *legal status* or *legal rights* of the Data Subject, or negatively impact its circumstances, behaviour or choices. Consequently, any FL system fails to comply with the GDPR requirements of the data subject's right to control automated decision making. Fortunately, this requirement can be neglected if a Data Controller explicitly mentions the lack of automated decision making and profiling right when asking for Data Subject's consent to process personal data.

[12] https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/right-to-be-informed.

5.3.3. *Other rights*

The nature of decoupling between data storage and processing at client-side and global ML model aggregation at server-side in centralised FL leads to the unnecessity of providing the (2) Right of access, (3) Right to rectification, (4) Right to erasure, (5) Right to restrict processing, (6) Right to data portability, and (7) Right to object. For instance, regarding the "Right to erasure", if a user requests to delete its data (i.e., local ML model parameters sent to an FL server), literally, the only way to fulfil the user's request is to thoroughly re-train the global model without using user's data from the round that the user first participates Ginart et al. (2019). This is unnecessary and impractical in FL settings as only local ML model parameters (possibly privacy guarantee-strengthened with differential privacy) in aggregated encrypted forms (by using Secure Aggregation and other advanced cryptography techniques) are shared with a coordination server. Consequently, it is worthless for a Data Subject to have control over its local ML model as *(i)* the model parameters are protected by privacy-preserving techniques from inference attacks; *(ii)* the server is unable to separate the user's data from the others, the server also does not store the model once it is aggregated to update the global model; and *(iii)* the global model is wholly anonymised and cannot be exploited to extract or infer any individual information.

5.4. *GDPR-Compliance investigation and demonstration*

The GDPR establishes supervisory authorities in each member state which are independent public authorities called Data Protection Authorities (DPAs). DPAs are responsible for supervising and inspecting whether a Data Controller is compliant with the data protection regulations whilst the Data Controller is responsible for demonstrating the compliance. The questions are judiciously raised: How can an FL system be investigated and validated by DPAs, and how can it demonstrate compliance?

5.4.1. *DPA'S supervisory competence*

As illustrated in Fig. 8, the investigation of non/compliance and decision of punishment are carried out by DPAs once there is a suspicion or a claim filed by a customer. The compliance inspection will conduct some analysis to see whether a suspicious organisation follows the legal requirement of Privacy&Security-by-design approach and satisfies some standard assessments such as Data Protection Impact Assessment (DPIA) and Privacy Impact Assessment (PIA), which are essential parts of the GDPR accountability obligations.

The GDPR establishes heavy punishment for non/compliance as failing to comply with the GDPR can be penalised by both financial fine (up to € 20M, or 4% of global annual turnover, whichever is higher) and reprimand, ban or suspension of the violator's business (Fig. 8). A number of criteria specifically defined by the GDPR (Articles 77–84) are taken into account when determining the punishment such as the level of cooperation during the investigation, type of personal data, any previous infringement, and the nature, gravity, and duration of the current infringement. For instance, Facebook and Google were hit with a collective

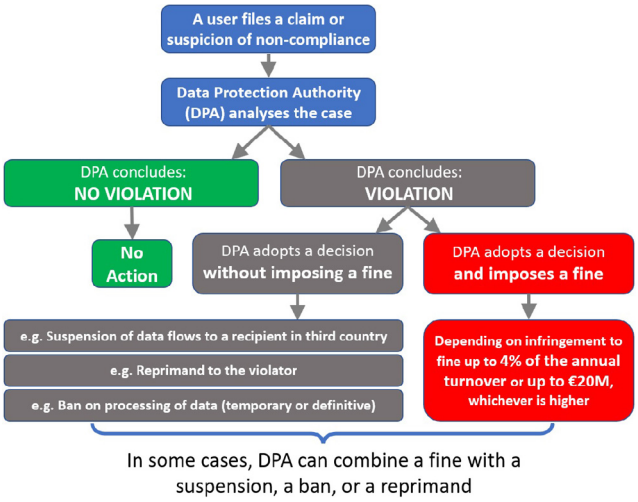

Fig. 8 – Workflow of the GDPR-compliance inspection and punishment procedure.

$8.8 billion lawsuits (Facebook, 3.9 billion euro; Google, 3.7 billion euro) by Austrian privacy campaigner, Max Schrems, alleging violations of GDPR as it pertains to the opt-in/opt-out clauses. Specifically, the complaint alleges that the way these companies obtain user consent for privacy policies is an "all-or-nothing" choice, asking users to check a small box allowing them to access services. It is a clear violation of the GDPR's provisions per privacy experts and the EU/UK. A list of fines and notices (with non-compliance reasons) issued under the GDPR can be found on Wikipedia[13]

Normally, DPAs might require a variety of information with a detailed explanation from the Data Controller to perform the analysis including documents of organisational and technical measures related to the implementation of the GDPR requirements as well as independent DPIA and PIA reports frequently conducted by the Data Controller. DPAs may also require to be given access to data server infrastructure and management system including personal data that is being processed. In this respect, besides the legal basis such as consents from end-users, an FL service provider can only provide documentation of how FL-related mechanisms are implemented along with privacy-preserving technical measures such as secure aggregation, differential privacy, and homomorphic encryption. Other inquiries from DPAs such as direct access to the FL model training operations and inspection of individual local model parameters from a particular end-user are technically infeasible for any FL systems.

5.4.2. *Compliance demonstration*

In order to build and demonstrate the GDPR compliance, AI/ML-based service providers should realise DPIA and PIA from the beginning of the project and document the processes accordingly which are designed to describe and clarify the whole data management processes along with the necessity and proportionality of these processes. Such assessments are important tools for accountability and essential to efficiently

[13] https://en.wikipedia.org/wiki/GDPR_fines_and_notices.

manage the data security and privacy risks, to demonstrate the compliance, as well as to determine the measure have been taken to address the risks. However, carrying out a DPIA or PIA is not mandatory for every data processing operation. It is only required when the operation is *"likely to result in a high risk to the rights and freedoms of natural persons"* (NS/>(1)). The guideline for the criteria on the DPIA/PIA obligatory is described under (3), 35(4) which are adopted by DPAs to carry out such assessments.

In this respect, any FL service providers should perform the following steps for the DPIA/PIA to ensure the GDPR-compliance as well as to demonstrate the compliance once required by DPAs:

1. A systematic description of data processing operations, associated purposes, along with clarification and justification of the operations. For instance, the operation of asking the Data Subject's consent for local ML training and sending the ML model parameters to a coordination server should be documented in detail.
2. An assessment of the necessity and proportionality of each operation, given its associated purposes. For instance, a Secure Aggregation mechanism is necessary to implement whereas a differential privacy mechanism is proportionally required.
3. An assessment of the data security and privacy risks that might be induced by each operation, along with the technical measures implemented to mitigate and manage the risks. For instance, in an FL system, the operation of sending local ML model parameters to a coordination server for global ML model update might be the target of inference attacks, thus, inducing privacy leakage. The measures called Secure Aggregation and Homomorphic Encryption mechanisms are implemented along with the technical report. Even though such privacy-preserving methods are implemented to strengthen FL systems, there exist some risks that can be exploited for illegitimate purposes such as model poisoning with back-door sub-tasks. These possible attacks, which lead to non-compliance with the GDPR, should be addressed.

Foremost, same as any AI/ML-based system, an FL system is based on black-box complex ML models (e.g., deep learning and neural networks) with limited transparency, making it troublesome for both service providers and DPAs to comprehend and to inspect hidden operations taking place inside the system. Therefore, conducting DPIA/PIA on an FL system seems to be superficial, which requires much effort to discover breaches of the regulations, so as to avoid risky operations and to impose better privacy-preserving measures.

6. Recap and outlook

AI/ML-based applications and services are high on the agenda in most sectors. However, the unregulated use or misuse of personal data is dramatically spreading, resulting in severe concerns of data privacy. A series of severe personal data breaches such as Facebook's Cambridge Analytica scandal, along with urgent mobile applications during the SARS-CoV2 pandemic for large-scale contact tracing and movement tracking Ienca and Vayena (2020) trigger worldwide attention respecting to a variety of privacy-related aspects including algorithm bias, ethics, implications of politic settings, and legal responsibility. This leads to a critical demand for effective privacy-preserving techniques, particularly for *"data-hungry"* AI/ML-based systems, wherein FL is a prospective solution. The decoupling between local storage and processing at end-users devices and the aggregation of processing results at the server-side in FL undoubtedly mitigate the risk of massive data breaches in a traditional centralised system while giving full control of personal data back to the users.

Although FL is in its infancy, we strongly believe that the collaborative computation with decentralised data storage as in FL systems has tremendous advantages to facilitate a variety of AI/ML-based applications without directly accessing end-users' data. Thus, FL systems are presumed to naturally comply with strict data protection legislation such as the GDPR. However, such FL systems still stay within the GDPR regulatory data protection framework as the local processing results sent to a server from end-users (e.g., locally trained ML model parameters) conceal some sensitive features that can be exploited to infer personal information of the end-users. Accordingly, FL systems are the target of some types of attack such as inference attacks and model poisoning, which could lead to infringements of the GDPR. Therefore, the systems must be strengthened by applicable privacy mechanisms such as SMC, differential privacy, and encrypted transfer learning methods Salem et al. (2019). We present a systematic summary of the threat models, possible attacks, and the privacy-preserving techniques in FL systems, along with the analysis of how these techniques can mitigate the risk of privacy leakages. Furthermore, insightful analysis of how an FL system complies with the GDPR is also provided. Obligations and appropriate measures for a service provider to implement a GDPR-compliant FL system are examined in details following the rational guidelines of the GDPR six principles.

As FL is in the early stage, a fruitful area of multi/disciplinary research is commenced in order to flourish the technology and to comply with the GDPR fully. Firstly, efficient cryptographic and privacy primitives for decentralised collaborative learning must be further developed, particularly for counteracting model poisoning and inference attacks. Furthermore, as these privacy-preserving techniques such as SMC impose non-trivial performance overheads, further effort on how to efficiently utilise such techniques on FL applications are required. Secondly, research on transparency, interpret-ability and algorithm fairness in FL systems should be profoundly carried out. Even though a substantial amount of research has been conducted in centralised AI/ML settings, there is still an open question of whether these approaches could be employed and how to sensibly adapt them to the decentralised settings where training data is highly skewed *non-IID* and unevenly distributed across sources. The sampling constraints should be investigated to see how much extend they affect and how to mitigate the bias of the global training model. For instance, the agnostic FL framework introduced in Mohri et al. (2019) naturally yields *good-intent fairness* as it modelled the target distribution as an unknown mixture of the distributions instead of the uniform distri-

bution in typical FL training algorithms. This agnostic FL framework, as a result, can control for bias in the training objective. Thirdly, it requires more research on interpretable and unbiased ML models and algorithms that can be employed over encrypted settings to well consolidate with advanced encryption schemes in FL systems. Besides, the trade-offs between privacy utility, accuracy, interpretability, and fairness in an FL framework need to be thoroughly explored.

If these requisites are successfully settled, it will assure to inaugurate responsible, auditable and trustworthy FL systems; as a result, complying with stringent requirements of the GDPR whilst bolstering the universal recognition of the secure decentralised collaborative learning solutions by both end-users and policymakers, including the GDPR supervisory authority.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Nguyen Truong: Investigation, Conceptualization, Methodology, Writing – original draft. **Kai Sun:** Investigation, Conceptualization, Supervision, Writing – review & editing. **Siyao Wang:** Writing – review & editing. **Florian Guitton:** Conceptualization, Methodology. **YiKe Guo:** Investigation, Supervision.

Acknowledgement

This research was supported by the HNA Research Centre for Future Data Ecosystems at Imperial College London and the Innovative Medicines Initiative 2 IDEA-FAST project under grant agreement No 853981.

REFERENCES

Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, Zhang L. Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security; 2016. p. 308–18.

Acar A, Aksu H, Uluagac AS, Conti M. A survey on homomorphic encryption schemes: theory and implementation. ACM Computing Surveys (CSUR) 2018;51(4):1–35.

Ananny M, Crawford K. Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. new media & society 2018;20(3):973–89.

Aono Y, Hayashi T, Wang L, Moriai S, et al. Privacy-preserving deep learning: Revisited and enhanced. In: International Conference on Applications and Techniques in Information Security. Springer; 2017. p. 100–10.

Ateniese G, Mancini LV, Spognardi A, Villani A, Vitali D, Felici G. Hacking smart machines with smarter ones: how to extract meaningful data from machine learning classifiers. Int. J. Secur. Netw. 2015;10(3):137–50.

Bagdasaryan E, Veit A, Hua Y, Estrin D, Shmatikov V. How to backdoor federated learning. In: International Conference on Artificial Intelligence and Statistics; 2020. p. 2938–48.

Bassily R, Smith A, Thakurta A. Private empirical risk minimization: Efficient algorithms and tight error bounds. In: 2014 IEEE 55th Annual Symposium on Foundations of Computer Science. IEEE; 2014. p. 464–73.

Ben-Nun T, Hoefler T. Demystifying parallel and distributed deep learning: an in-depth concurrency analysis. ACM Computing Surveys (CSUR) 2019;52(4):1–43.

Bhagoji AN, Chakraborty S, Mittal P, Calo S. Analyzing federated learning through an adversarial lens. In: International Conference on Machine Learning; 2019. p. 634–43.

Bhowmick A, Duchi J, Freudiger J, Kapoor G, Rogers R. Protection against reconstruction and its applications in private federated learning. arXiv preprint arXiv:1812.00984 2018.

Biggio B, Nelson B, Laskov P. Poisoning attacks against support vector machines. In: 29th International Conference on Machine Learning. ArXiv e-prints; 2012. p. 1807–14.

Blanchard P, Guerraoui R, Stainer J, et al. Machine learning with adversaries: Byzantine tolerant gradient descent. In: Advances in Neural Information Processing Systems; 2017. p. 119–29.

Bonawitz K, Eichner H, Grieskamp W, Huba D, Ingerman A, Ivanov V, Kiddon C, Konečnỳ J, Mazzocchi S, McMahan HB, et al. Towards federated learning at scale: system design. arXiv preprint arXiv:1902.01046 2019.

Bonawitz K, Ivanov V, Kreuter B, Marcedone A, McMahan HB, Patel S, Ramage D, Segal A, Seth K. Practical secure aggregation for federated learning on user-held data. arXiv preprint arXiv:1611.04482 2016.

Bonawitz K, Ivanov V, Kreuter B, Marcedone A, McMahan HB, et al. Practical secure aggregation for privacy-preserving machine learning. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security; 2017. p. 1175–91.

Bottou L. Large-scale Machine Learning with Stochastic Gradient Descent. In: Proceedings of COMPSTAT'2010. Springer; 2010. p. 177–86.

Brickell EF. Some ideal secret sharing schemes. In: Workshop on the Theory and Application ofCryptographic Techniques. Springer; 1989. p. 468–75.

Canetti R, Feige U, Goldreich O, Naor M. Adaptively secure multi-party computation. In: Proceedings of the twenty-eighth annual ACM symposium on Theory of computing; 1996. p. 639–48.

Chaudhuri K, Monteleoni C. Privacy-preserving logistic regression. In: Advances in neural information processing systems; 2009. p. 289–96.

Chaudhuri K, Monteleoni C, Sarwate AD. Differentially private empirical risk minimization.. Journal of Machine Learning Research 2011;12(3).

Chen J, Pan X, Monga R, Bengio S, Jozefowicz R. Revisiting distributed synchronous sgd. arXiv preprint arXiv:1604.00981 2016.

Chen L, Wang H, Charles Z, Papailiopoulos D. Draco: byzantine-resilient distributed training via redundant gradients. arXiv preprint arXiv:1803.09877 2018.

Chen X, Liu C, Li B, Lu K, Song D. Targeted backdoor attacks on deep learning systems using data poisoning. arXiv preprint arXiv:1712.05526 2017.

Chen X-W, Lin X. Big data deep learning: challenges and perspectives. IEEE Access 2014;2:514–25.

Chilimbi T, Suzue Y, Apacible J, Kalyanaraman K. Project adam: Building an efficient and scalable deep learning training system. In: 11th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 14); 2014. p. 571–82.

Chor B, Goldwasser S, Micali S, Awerbuch B. Verifiable secret sharing and achieving simultaneity in the presence of faults. In: 26th Annual Symposium on Foundations of Computer Science (sfcs 1985). IEEE; 1985. p. 383–95.

Cramer R, Damgård I, Maurer U. General secure multi-party computation from any linear secret-sharing scheme. In: International Conference on the Theory and Applications of Cryptographic Techniques. Springer; 2000. p. 316–34.
Daily J, Vishnu A, Siegel C, Warfel T, Amatya V. Gossipgrad: scalable deep learning using gossip communication based asynchronous gradient descent. arXiv preprint arXiv:1803.05880 2018.
Dean J, Corrado G, Monga R, Chen K, Devin M, Mao M, Ranzato M, Senior A, Tucker P, Yang K, et al. Large scale distributed deep networks. In: Advances in neural information processing systems; 2012. p. 1223–31.
Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 2017.
Du W, Han YS, Chen S. Privacy-preserving multivariate statistical analysis: Linear regression and classification. In: Proceedings of the 2004 SIAM international conference on data mining. SIAM; 2004. p. 222–33.
Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization.. Journal of machine learning research 2011;12(7).
Dwork C. Differential privacy: A survey of results. In: International conference on theory and applications of models of computation. Springer; 2008. p. 1–19.
Dwork C, Kenthapadi K, McSherry F, Mironov I, Naor M. Our data, ourselves: Privacy via distributed noise generation. In: Annual International Conference on the Theory and Applications of Cryptographic Techniques. Springer; 2006. p. 486–503.
Dwork C, Roth A, et al. The algorithmic foundations of differential privacy.. Foundations and Trends in Theoretical Computer Science 2014;9(3–4):211–407.
Dwork C, Smith A, Steinke T, Ullman J. Exposed! a survey of attacks on private data. Annu Rev Stat Appl 2017;4(1):61–84.
Ferdinand N, Al-Lawati H, Draper SC, Nokleby M. Anytime minibatch: exploiting stragglers in online distributed optimization. arXiv preprint arXiv:2006.05752 2020.
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; 2015. p. 1322–33.
Friedman A, Schuster A. Data mining with differential privacy. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining; 2010. p. 493–502.
Fung BC, Wang K, Chen R, Yu PS. Privacy-preserving data publishing: a survey of recent developments. ACM Computing Surveys (Csur) 2010;42(4):1–53.
Fung C, Yoon CJ, Beschastnikh I. Mitigating sybils in federated learning poisoning. arXiv preprint arXiv:1808.04866 2018.
Ganju K, Wang Q, Yang W, Gunter CA, Borisov N. Property inference attacks on fully connected neural networks using permutation invariant representations. In: Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security; 2018. p. 619–33.
Geiping J, Bauermeister H, Dröge H, Moeller M. Inverting gradients–how easy is it to break privacy in federated learning? arXiv preprint arXiv:2003.14053 2020.
Gentry C. Computing arbitrary functions of encrypted data. Commun ACM 2010;53(3):97–105.
Gentry C, Boneh D, 20. Stanford university Stanford; 2009.
Gentry C, Halevi S. Implementing gentrys fully-homomorphic encryption scheme. In: Annual international conference on the theory and applications of cryptographic techniques. Springer; 2011. p. 129–48.
Gerbessiotis AV, Valiant LG. Direct bulk-synchronous parallel algorithms. J Parallel Distrib Comput 1994;22(2):251–67.
Geyer RC, Klein T, Nabi M. Differentially private federated learning: a client level perspective. arXiv preprint arXiv:1712.07557 2017.
Gilad-Bachrach R, Dowlin N, Laine K, Lauter K, Naehrig M, Wernsing J. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In: International Conference on Machine Learning; 2016. p. 201–10.
Ginart A, Guan M, Valiant G, Zou JY. Making ai forget you: Data deletion in machine learning. In: Advances in Neural Information Processing Systems; 2019. p. 3518–31.
Goldreich O. Secure multi-party computation. Manuscript. Preliminary version 1998;78.
Goldreich O, Micali S, Wigderson A. How to Play Any Mental Game, or a Completeness Theorem for Protocols with Honest Majority. In: Providing Sound Foundations for Cryptography: On the Work of Shafi Goldwasser and Silvio Micali; 2019. p. 307–28.
Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Advances in neural information processing systems; 2014. p. 2672–80.
Greengard S. Weighing the impact of gdpr. Commun ACM 2018;61(11):16–18.
Harder F, Bauer M, Park M. Interpretable and differentially private predictions.. In: AAAI; 2020. p. 4083–90.
Hardy C, Le Merrer E, Sericola B. Distributed deep learning on edge-devices: feasibility via adaptive compression. In: 2017 IEEE 16th International Symposium on Network Computing and Applications (NCA). IEEE; 2017. p. 1–8.
He L, Bian A, Jaggi M. Cola: Decentralized linear learning. In: Advances in Neural Information Processing Systems; 2018. p. 4536–46.
Hitaj B, Ateniese G, Perez-Cruz F. Deep models under the gan: information leakage from collaborative deep learning. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security; 2017. p. 603–18.
Ho Q, Cipar J, Cui H, Lee S, Kim JK, Gibbons PB, Gibson GA, Ganger G, Xing EP. More effective distributed ml via a stale synchronous parallel parameter server. In: Advances in neural information processing systems; 2013. p. 1223–31.
Horvitz E, Mulligan D. Data, privacy, and the greater good. Science 2015;349(6245):253–5.
Hu YC, Patel M, Sabella D, Sprecher N, Young V. Mobile edge computinga key technology towards 5g. ETSI white paper 2015;11(11):1–16.
Huang Y, Katz J, Evans D. Efficient secure two-party computation using symmetric cut-and-choose. In: Annual Cryptology Conference. Springer; 2013. p. 18–35.
Ienca M, Vayena E. On the responsible use of digital data to tackle the covid-19 pandemic. Nat. Med. 2020;26(4):463–4.
Jagannathan G, Wright RN. Privacy-preserving distributed k-means clustering over arbitrarily partitioned data. In: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining; 2005. p. 593–9.
Jagielski M, Oprea A, Biggio B, Liu C, Nita-Rotaru C, Li B. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In: 2018 IEEE Symposium on Security and Privacy (SP). IEEE; 2018. p. 19–35.
Jarecki S, Shmatikov V. Efficient two-party secure computation on committed inputs. In: Annual International Conference on the Theory and Applications of Cryptographic Techniques. Springer; 2007. p. 97–114.
Johnson R, Zhang T. Accelerating stochastic gradient descent using predictive variance reduction. In: Advances in neural information processing systems; 2013. p. 315–23.
Kairouz P, McMahan HB, Avent B, Bellet A, Bennis M, Bhagoji AN, Bonawitz K, Charles Z, Cormode G, Cummings R, et al. Advances and open problems in federated learning. arXiv preprint arXiv:1912.04977 2019.

Keskar NS, Mudigere D, Nocedal J, Smelyanskiy M, Tang PTP. On large-batch training for deep learning: generalization gap and sharp minima. arXiv preprint arXiv:1609.04836 2016.
Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 2014.
Koh PW, Liang P. Understanding black-box predictions via influence functions. In: International Conference on Machine Learning; 2017. p. 1885–94.
Koloskova A, Stich SU, Jaggi M. Decentralized stochastic optimization and gossip algorithms with compressed communication. arXiv preprint arXiv:1902.00340 2019.
Konečnỳ J, McMahan HB, Ramage D, Richtárik P. Federated optimization: distributed machine learning for on-device intelligence. arXiv preprint arXiv:1610.02527 2016.
Konečnỳ J, McMahan HB, Yu FX, Richtárik P, Suresh AT, Bacon D. Federated learning: strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492 2016.
Konečnỳ J, Richtárik P. Semi-stochastic gradient descent methods. Frontiers in Applied Mathematics and Statistics 2017;3:9.
Li N, Li T, Venkatasubramanian S. t-closeness: Privacy beyond k-anonymity and l-diversity. In: 2007 IEEE 23rd International Conference on Data Engineering. IEEE; 2007. p. 106–15.
Li O, Liu H, Chen C, Rudin C. Deep learning for case-based reasoning through prototypes: a neural network that explains its predictions. arXiv preprint arXiv:1710.04806 2017.
Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: challenges, methods, and future directions. IEEE Signal Process Mag 2020;37(3):50–60.
Lian X, Zhang C, Zhang H, Hsieh C-J, Zhang W, Liu J. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In: Advances in Neural Information Processing Systems; 2017. p. 5330–40.
Lian X, Zhang W, Zhang C, Liu J. Asynchronous decentralized parallel stochastic gradient descent. In: International Conference on Machine Learning. PMLR; 2018. p. 3043–52.
Lindell Y. Fast cut-and-choose-based protocols for malicious and covert adversaries. Journal of Cryptology 2016;29(2):456–90.
Lindell Y, Pinkas B. Privacy preserving data mining. In: Annual International Cryptology Conference. Springer; 2000. p. 36–54.
Lindell Y, Pinkas B. An efficient protocol for secure two-party computation in the presence of malicious adversaries. In: Annual international conference on the theory and applications of cryptographic techniques. Springer; 2007. p. 52–78.
Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. L-diversity: privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD) 2007;1(1) 3-es.
McMahan B, Moore E, Ramage D, Hampson S, y Arcas BA. Communication-efficient learning of deep networks from decentralized data. In: Artificial Intelligence and Statistics; 2017. p. 1273–82.
McMahan HB, Moore E, Ramage D, y Arcas BA. Federated learning of deep networks using model averaging. arXiv preprint arXiv:1602.05629 2016.
McMahan HB, Ramage D, Talwar K, Zhang L. Learning differentially private recurrent language models. arXiv preprint arXiv:1710.06963 2017.
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.
Mei S, Zhu X. Using machine teaching to identify optimal training-set attacks on machine learners.. In: AAAI; 2015. p. 2871–7.
Melis L, Song C, De Cristofaro E, Shmatikov V. Inference attacks against collaborative learning. arXiv preprint arXiv:1805.04049 2018;13.
Melis L, Song C, De Cristofaro E, Shmatikov V. Exploiting unintended feature leakage in collaborative learning. In: 2019 IEEE Symposium on Security and Privacy (SP). IEEE; 2019. p. 691–706.
Mhamdi EME, Guerraoui R, Rouault S. The hidden vulnerability of distributed learning in byzantium. arXiv preprint arXiv:1802.07927 2018.
Mohassel P, Zhang Y. Secureml: A system for scalable privacy-preserving machine learning. In: 2017 IEEE Symposium on Security and Privacy (SP). IEEE; 2017. p. 19–38.
Mohri M, Sivek G, Suresh AT. Agnostic federated learning. In: 36th International Conference on Machine Learning, ICML 2019. International Machine Learning Society (IMLS); 2019. p. 8114–24.
Molnar C. Interpretable machine learning. Lulu.com; 2020.
Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Interpretable machine learning: definitions, methods, and applications. arXiv preprint arXiv:1901.04592 2019.
Narayanan A, Shmatikov V. Robust de-anonymization of large sparse datasets. In: 2008 IEEE Symposium on Security and Privacy (sp 2008). IEEE; 2008. p. 111–25.
Nasr M, Shokri R, Houmansadr A. Comprehensive privacy analysis of deep learning: stand-alone and federated learning under passive and active white-box inference attacks. arXiv preprint arXiv:1812.00910 2018.
Nasr M, Shokri R, Houmansadr A. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In: 2019 IEEE Symposium on Security and Privacy (SP). IEEE; 2019. p. 739–53.
Pathak M, Rane S, Raj B. Multiparty differential privacy via aggregation of locally trained classifiers. In: Advances in Neural Information Processing Systems; 2010. p. 1876–84.
Phong LT, Aono Y, Hayashi T, Wang L, Moriai S. Privacy-preserving deep learning via additively homomorphic encryption. IEEE Trans. Inf. Forensics Secur. 2018;13(5):1333–45.
Qian N. On the momentum term in gradient descent learning algorithms. Neural networks 1999;12(1):145–51.
Ram SS, Nedić A, Veeravalli VV. Asynchronous gossip algorithms for stochastic optimization. In: Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference. IEEE; 2009. p. 3581–6.
Recht B, Re C, Wright S, Niu F. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In: Advances in neural information processing systems; 2011. p. 693–701.
Reisizadeh A, Taheri H, Mokhtari A, Hassani H, Pedarsani R. Robust and communication-efficient collaborative learning. In: Advances in Neural Information Processing Systems; 2019. p. 8388–99.
Rezende DJ, Mohamed S, Wierstra D. Stochastic backpropagation and approximate inference in deep generative models. arXiv preprint arXiv:1401.4082 2014.
Rubinstein BI, Bartlett PL, Huang L, Taft N. Learning in a large function space: privacy-preserving mechanisms for svm learning. Journal of Privacy and Confidentiality 2012;4(1):65–100.
Ruder S. An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747 2016.
Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 2019;1(5):206–15.
Rumelhart DE, Hinton GE, Williams RJ. In: Technical Report. Learning internal representations by error propagation. California Univ San Diego La Jolla Inst for Cognitive Science; 1985.
Salem M, Taheri S, Yuan J-S. Utilizing transfer learning and homomorphic encryption in a privacy preserving and secure biometric recognition system. Computers 2019;8(1):3.

Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. Improved techniques for training gans. In: Advances in neural information processing systems; 2016. p. 2234–42.
Sarwate AD, Chaudhuri K. Signal processing and machine learning with differential privacy: algorithms and challenges for continuous data. IEEE Signal Process Mag 2013;30(5):86–94.
Shamir A. How to share a secret. Commun ACM 1979;22(11):612–13.
Shamir O, Srebro N, Zhang T. Communication-efficient distributed optimization using an approximate newton-type method. In: International conference on machine learning; 2014. p. 1000–8.
Shokri R, Shmatikov V. Privacy-preserving deep learning. In: Proceedings of the 22nd ACM SIGSAC conference on computer and communications security; 2015. p. 1310–21.
Shokri R, Stronati M, Song C, Shmatikov V. Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy (SP). IEEE; 2017. p. 3–18.
Smith V, Chiang C-K, Sanjabi M, Talwalkar AS. Federated Multi-task Learning. In: Advances in Neural Information Processing Systems 30. Curran Associates, Inc.; 2017. p. 4424–34.
Song S, Chaudhuri K, Sarwate AD. Stochastic gradient descent with differentially private updates. In: 2013 IEEE Global Conference on Signal and Information Processing. IEEE; 2013. p. 245–8.
Sun L, Qian J, Chen X, Yu PS. Ldp-fl: practical private aggregation in federated learning with local differential privacy. arXiv preprint arXiv:2007.15789 2020.
Sweeney L. K-anonymity: a model for protecting privacy. Int. J. Uncertainty Fuzziness Knowledge Based Syst. 2002;10(05):557–70.
Tieleman T, Hinton G. Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning 2012;4(2):26–31.
Truong NB, Sun K, Lee GM, Guo Y. Gdpr-compliant personal data management: ablockchain-based solution. IEEE Trans. Inf. Forensics Secur. 2019;15:1746–61.
Wachter S, Mittelstadt B, Floridi L. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. International Data Privacy Law 2017;7(2):76–99.
Wang S, Tuor T, Salonidis T, Leung KK, Makaya C, He T, Chan K. Adaptive federated learning in resource constrained edge computing systems. IEEE J. Sel. Areas Commun. 2019;37(6):1205–21.
Wang X, Han Y, Leung VC, Niyato D, Yan X, Chen X. Convergence of edge computing and deep learning: acomprehensive survey. IEEE Communications Surveys & Tutorials 2020;22(2):869–904.
Wang Z, Song M, Zhang Z, Song Y, Wang Q, Qi H. Beyond inferring class representatives: User-level privacy leakage from federated learning. In: IEEE INFOCOM 2019-IEEE Conference on Computer Communications. IEEE; 2019. p. 2512–20.
Wei K, Li J, Ding M, Ma C, Yang HH, Farokhi F, Jin S, Quek TQ, Poor HV. Federated learning with differential privacy: algorithms and performance analysis. IEEE Trans. Inf. Forensics Secur. 2020.
Xiao H, Biggio B, Brown G, Fumera G, Eckert C, Roli F. Is feature selection secure against training data poisoning?. In: International Conference on Machine Learning; 2015. p. 1689–98.
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 preprint arXiv:1812.02903 2018.
Yao AC. How to generate and exchange secrets. In: 27th Annual Symposium on Foundations of Computer Science (sfcs 1986). IEEE; 1986. p. 162–7.
Zhang X, Ji S, Wang H, Wang T. Private, yet practical, multiparty deep learning. In: 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS). IEEE; 2017. p. 1442–52.
Zhao Y, Li M, Lai L, Suda N, Civin D, Chandra V. Federated learning with non-iid data. arXiv preprint arXiv:1806.00582 2018.
Zhou Z, Mertikopoulos P, Bambos N, Glynn P, Ye Y, Li L-J, Fei-Fei L. Distributed asynchronous optimization with unbounded delays: How slow can you go?. In: International Conference on Machine Learning. PMLR; 2018. p. 5970–9.
Zhu L, Liu Z, Han S. Deep leakage from gradients. In: Advances in Neural Information Processing Systems; 2019. p. 14774–84.

Dr. Nguyen B.Truong is currently a Research Associate at Data Science Institute, Imperial College London, United Kingdom. He received his Ph.D, MSc, and BSc degrees from Liverpool John Moores University, United Kingdom, Pohang University of Science and Technology, Korea, and Hanoi University of Science and Technology, Vietnam in 2018, 2013, and 2008, respectively. He was a Software Engineer at DASAN Networks, a leading company on Networking Products and Services in South Korea in 2012–2015. His research interest is including, but not limited to, Data Privacy, Security, and Trust, Personal Data Management, Distributed Systems, and Blockchain.

Dr. Kai Sun is the Operation Manager of the Data Science Institute at Imperial College London. She received the MSc degree and the Ph.D degree in Computing from Imperial College London, in 2010 and 2014, respectively. From 2014 to 2017, she was a Research Associate at the Data Science Institute at Imperial College London, working on EU IMI projects including U-BIOPRED and eTRIKS, responsible for translational data management and analysis. She was the manager of the HNA Centre of Future Data Ecosystem in 2017–2018. Her research interests include translational research management, network analysis and decentralised systems.

Mr. Siyao Wang is a PhD student of the Data Science Institute at Imperial College London. He received the BSc degree in Computer Science and Technology from the University of Chinese Academy of Sciences in 2018. He received the MRes degree in Medical Robotics and Image-Guided Intervention from Imperial College London in 2019. His research interests include machine learning, deep learning, computer vision and artificial intelligence applications in healthcare.

Mr. Florian Guitton received a BSc in Software Engineering from Epitech (France) in 2011 and a MSc in Advanced Computing from the University of Kent (United Kingdom) in 2012. In 2012 he joined the Discovery Sciences Group at Imperial College London where he became Research Assistant working on iHealth, eTRIKS and IDEA-FAST EU programs. He is currently a PhD candidate at Data Science Institute, Imperial College

London working on distributed data collection and analysis pipeline in mixed-security environments with the angle of optimising user facing experiences.

Dr. Yike Guo (FREng, MAE) is the director of the Data Science Institute at Imperial College London and the Vice-President (Research and Development) of Hong Kong Baptist University. He received the BSc degree in Computing Science from Tsinghua University, China, in 1985 and received the Ph.D in Computational Logic from Imperial College London in 1993. He is a Professor of Computing Science in the Department of Computing at Imperial College London since 2002. He is a fellow of the Royal Academy of Engineering and a member of the Academia Europaea. His research interests are in the areas of data mining for large-scale scientific applications including distributed data mining methods, machine learning and informatics systems.