Deloitte.

30/07/2022

# Federated Learning and Decentralized Data

This paper analyses a learning technique that allows users to collectively gather the benefits of shared models trained from data, without the need to centrally store it. This approach is called Federated Learning (FL). It is also exposed how this technique tries to guarantee and preserve the respect of privacy according to the GDPR.

> *"Somewhere, something incredible is waiting to be known"*
>
> ***Carl Sagan***

## Federated Learning: A Distributed Collaborative Learning Approach

As data breach becomes a major concern, more and more governments establish regulations to protect users' data, such as GDPR in European Union and CCPA in the US. Under the above circumstances, Federated Learning, a decentralized Machine Learning framework that can train a model without direct access to users' private data, has drawn increasingly attention nowadays. Many Machine Learning algorithms require large amounts of data to be trained, and, often data are dispersed over different organizations under the protection of privacy restrictions. Due to these factors, Federated Learning has become a newsworthy research topic in Artificial Intelligence and Machine Learning field.

An exemple situation, data of different financial institutions are isolated and become *data islands*. Since each *data island* has limitations in size and discriminant power, a single financial institution may not be able to train a high-quality model that has a good predictive metrics for a specific task. Ideally, financial institutions can benefit more if they can collaboratively train a Machine Learning model on the union of their data. However, the data can't simply be shared among the financial institutions due to various policies and regulations, policies such as General Data Protection Regulation (GDPR). The problem can be overcome by introducing Federated Learning, that is, a distributed Machine Learning technique where models move across Institutions to learn from datasets separately and adjust their parameters as they learn from more and more data. However, the exchanging of model parameters, or hyperparameters in Machine Learning terminology, may cause data leakage. Therefore, a series of techniques (subsequently exposed) aimed at preserving privacy are essential for Federated Learning to protect privacy from cyber attack.

## A Technical Overview: Federated Learning Workflow Cycle

Federated Learning, a learning paradigm in which multiple parties train collaboratively without the need to exchange or centralize datasets, can have two types of architecture process workflows: Aggregation Server and Peer to Peer.

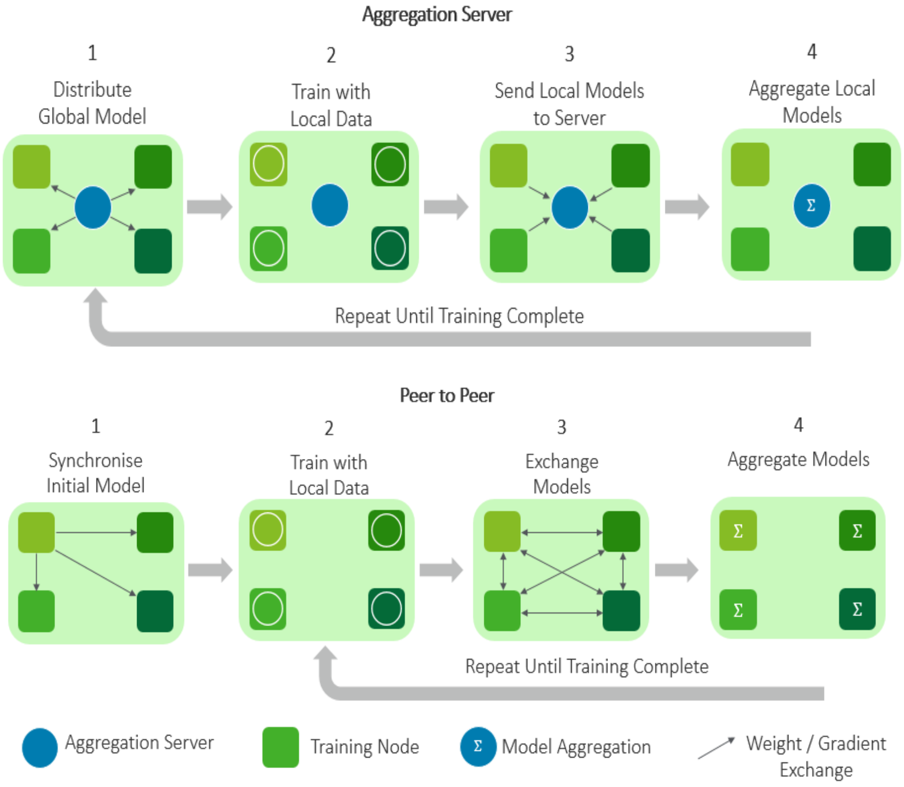
The Aggregation Server model consists of a single and central aggregator, surrounded by nodes. The federation of training nodes receive the global model from the aggregation server, then each node, independently, start training model with their own data. After, they resubmit their partially trained models to a central server for aggregation, and then, continue training after receiving the consensus from the central aggregator.

This type of architecture is useful when an *entity* needs to train its Machine Learning model on various non-communicating nodes (e.g. clients, other institutions) which are acting as trainers.

The other type of architecture is a Peer-to-Peer model. In this version, each node benefits from the trained models without insisting on an exchange of data. The lack of central aggregation servers makes the model slightly more complex for management since the first step should be started from one node. After the synchronization has taken place, each node begins to train the model with its own data, and then, exchanging their hyperparameters. Finally, each node performs their own aggregation of the parameters.

This type of architecture is useful when all nodes are both trainers and users of the final model. One of the applications where it is possible to apply this model, in the financial field, is for the collection of data aimed at training a Machine Learning model for AML purpose.

The problem with the current AML approach lies in the limitations of any one bank's data. Collectively, the financial data processed by banks across some region is enough to train an efficient AML machine. But data privacy rules, security issues and technological limitations do not allow for information sharing between enterprises. This problem could be overcome with a Federated Learning approach.



## Privacy and GDPR

The privacy preservation advantage of Federated Learning, compared to the traditional centralized Machine Learning approaches, is undeniable: It enables to train a Machine Learning model whilst retaining personal training data on nodes. Nevertheless, such model parameters still enclose some sensitive features that can be exploited to reconstruct or to infer related.

### Privacy Preserving Techniques in Machine Learning

In general, privacy preservation techniques for a distributed learning system target two main objectives: privacy of the training set and privacy of the local model parameters which are exchanged with other nodes and/or a centralised server. Regarding this topic, notable privacy-preserving techniques in Machine Learning are Data Anonymization, Differential Privacy, Secure Multi-party Computation (SMC), and Homomorphic Encryption.

- Data Anonymization
  Data anonymization or de-identification is a technique to mask or remove sensitive attributes, such as Personally Identifiable Information (PII), so that a data subject can't be identified within the modified dataset. Therefore, data anonymization should balance well between privacy guarantee and utility because masking or removing information may reduce the discriminating power of the dataset. Furthermore, when combined with auxiliary information from other anonymous datasets, a data subject might be reidentified, subjected to a privacy attack called linkage attack. To prevent from linkage attack, numerous techniques have

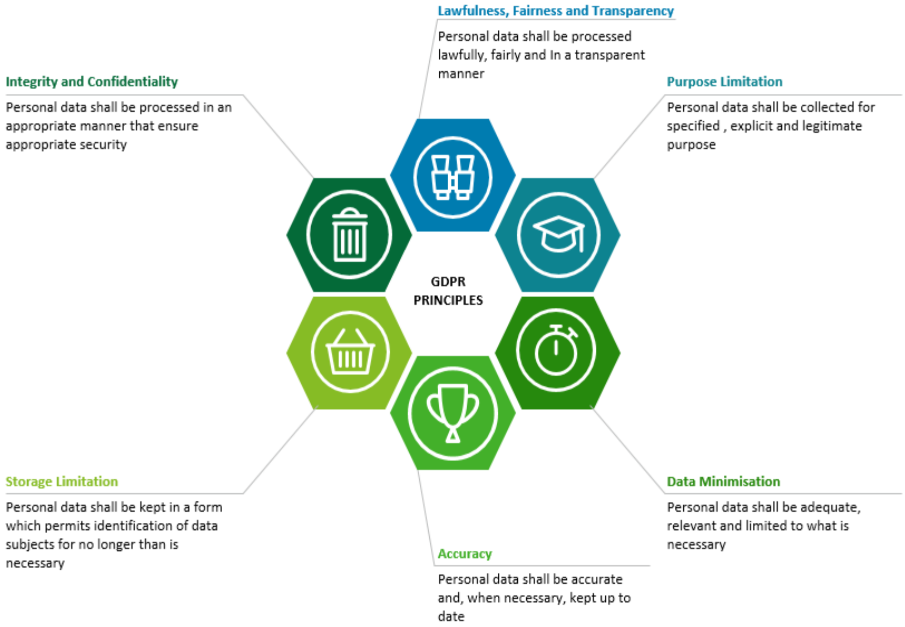
been proposed such as k-anonymity (suppression and generalization of some attributes), l-diversity (a further reduction in granularity than k-anonymity and tcloseness- a technique built on both k-anonymity and l-diversity. Unfortunately, such privacy preserving techniques can’t defend against linkage attacks whose adversaries possess some knowledge about the sensitive attributes. This weakness in the k-anonymity-based methods calls for different approaches that offer rigorous privacy guarantee such as differential privacy, Multi-party Computation and Homomorphic Encryption.

- **Differential Privacy** is an advanced solution of the perturbation privacy preserving technique where random noise is added to true outputs.
  As a result, it’s 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. 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. Differential privacy technique has been widely employed in various Machine Learning algorithms such as Logistic Regression, Support Vector Machine (SVM) and Deep Learning.

- **Multi-party Computation** (MPC) or privacy-preserving computation is a method that can be collectively computed over a dataset owned by multiple parties using their own inputs. so that any party learns nothing about others’ data except the outputs. MPC is beneficial to data privacy preservation in distributed learning wherein compute nodes collaboratively perform model training on their local dataset without revealing such dataset to others (Peer to Peer architecture).

- **Homomorphic Encryption** is another approach to preserve data privacy specially for Aggregation Server architecture. This technique enables the ability to perform computation on an encrypted form of data without the need for the secret key to decrypt the ciphertext. Results of the computation are in encrypted form and can only be ecrypted 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.

## The GDPR Perspective



Lawfulness, Fairness and Transparency

Personal data shall be processed lawfully, fairly and In a transparent manner

Integrity and Confidentiality

Personal data shall be processed in an appropriate manner that ensure appropriate security

Purpose Limitation

Personal data shall be collected for specified , explicit and legitimate purpose

Storage Limitation

Personal data shall be kept in a form which permits identification of data subjects for no longer than is necessary

Data Minimisation

Personal data shall be adequate, relevant and limited to what is necessary

Accuracy

Personal data shall be accurate and, when necessary, kept up to date

Federated Learning system still retains within the GDPR and is liable for complying with obligatory requirements. In order to achieve these requisites, the 6 GDPR core principles that should be respected are:

- **Lawfulness, Fairness and Transparency**, personal data shall be processed lawfully, fairly and in a transparent manner;
- **Purpose Limitation,** personal data shall be collected for specified, explicit and legitimate purpose;
- **Data Minimisation,** personal data shall be adequate, relevant, and limited to what is necessary;
- **Accuracy**, personal data shall be accurate and, when necessary, kept up to date;
- **Storage Limitation,** personal data shall be kept in a form which permits identification of data subjects for no longer than is necessary;
- **Integrity and Confidentiality,** personal data shall be processed in an appropriate manner that ensure appropriate security.

In addition to the core principles, a number of rights must also be respected such as Rights of Data Subject, Right to be informed, Rights in relation to automated decision making and profiling to maintain and validate the integrity of the GDPR policies.

Federated learning proves to be a technique with very high potential as it is able to mitigate problems present in the field of Machine Learning such as data scarcity and the maintenance of privacy from data from different sources.
One of the challenges of the present and future will be to develop a Federated Learning system with good predictive metrics while observing policies and regulations to protect privacy.

## Contact us:

**Filippo Finocchiaro**
**Equity Partner**
Email: ffinocchiaro@deloitte.it

**Simone Francesia**
**Director**
Email: sfrancesia@deloitte.it

**Carlo Scialpi**
**Director**
Email: cscialpi@deloitte.it

Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited ("DTTL"), its global network of member firms, and their related entities (collectively, the "Deloitte organization"). DTTL (also referred to as "Deloitte Global") and each of its member firms and related entities are legally separate and independent entities, which cannot obligate or bind each other in respect of third parties. DTTL and each DTTL member firm and related entity is liable only for its own acts and omissions, and not those of each other. DTTL does not provide services to clients. Please see www.deloitte.com/about to learn more.