TechDispatch

FEDERATED LEARNING

EDPS
European Data Protection Supervisor

aepd agencia española protección datos

Luxembourg: Publications Office of the European Union, 2025

PDF ISBN 978-92-9242-892-1 ISSN 2599-932X doi:10.2804/5357101 QT-01-24-000-EN-N

CONTENTS

1. Executive Summary

Federated Learning (FL) presents a promising approach to machine learning (ML) by allowing multiple sources of data (devices or entities) to collaboratively train a shared model while keeping data decentralised. This approach mitigates privacy risks as raw data remains locally on the sources, which is particularly beneficial in scenarios where data sensitivity or regulatory requirements make data centralisation[1] impractical. Applications of FL are diverse, spanning personalised recommendations, healthcare data analysis, data spaces, and autonomous transport systems, where ensuring privacy and data protection is paramount.

From a personal data protection perspective, FL offers significant benefits by minimising personal data sharing. This decentralised approach aligns with the core principles of data protection, such as data minimisation and purpose limitation, by ensuring that personal data remains under the control of the controller and is not exposed to external parties. Furthermore, FL improves accountability and auditability, as data controllers have clearer oversight of how personal data is processed. Additionally, by keeping raw data on local devices/servers and only sharing models or model updates (gradients or weights), FL can enhance the confidentiality of personal data limiting the need for its centralisation and reducing the impact of large-scale data breaches.

Despite its advantages, FL presents some challenges that are still not fully solved. One of the primary concerns is the potential for data leakage through model updates, as even without direct access to raw data, an attacker could infer sensitive information by analysing the gradients or weights shared between devices (and the central server where there is one). This vulnerability opens the door to membership inference attacks, where adversaries can determine whether specific data points were part of the training set. Additionally, security has to be implemented across the whole ecosystem or attackers would have the opportunity to attack the weakest link and then compromise the whole system. Furthermore, FL must put in place specific distributed training data quality assurance measures, and be free of bias, when data is processed for an intended purpose. Compared to non-FL architectures, FL has different threat vectors[2] that can affect the integrity of the data and appropriate measures should be devised and implemented.

It should not be assumed that data exchanged among the client devices and the resulting ML models can be treated as anonymous data; a careful technical and legal analysis has to be done to analyse the nature of the data, the risks associated with the model updates and the measures that should be applied to mitigate such risks.

1 i.e. centralising data in a single location
2 Various methods or pathways that attackers use to gain unauthorised access to data

To fully leverage the benefits of FL while addressing its challenges, a holistic approach to how personal data is actually processed is essential. This includes implementing system architectures that prioritise data protection by design and by default, ensuring that data access among federated parties is carried out balancing the level of risk of the processing, accuracy and usefulness of the resulting model.

By focusing on privacy and the protection of personal data, FL can be effectively utilised to develop AI systems that are both powerful and respectful of user's rights and freedoms.

2. AI and Privacy Enhancing Technologies

In recent years, thanks to the availability of massive computations power and access to massive amounts of data, the success of artificial intelligence (AI) systems[3] has increased with the use of machine learning (ML) development techniques. As any other product, AI systems should follow a design, development, validation and testing process that guarantees the performance requirements for a specific purpose and context, and be in accordance with, among others, personal data protection legislation.

The ML system development process includes the following phases (in addition to the regular development of non-ML[4] systems).

1. **The training of the system:** training an AI requires large quantities of data that is relevant for the purpose/objective of the AI. For example, for a large language model (AI designed to understand and generate human language), the training data should comprise a large volume of machine-readable text[5] (currently the training of some of these systems involves hundreds of billions or trillions of words[6]) so that the AI can provide outputs that look like human readable text.
2. **Automatic evolution of the system:** Some[7] AI systems include algorithms that allow the evolution of the system during the implementation[8] phase; i.e. the capability to use the operation data for continuous training, which implies the need for a continuous validation and testing.

3 The AI Act (Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance) defines an AI system as a "machine-based system designed to operate with varying levels of autonomy, that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments".

4 Details are provided in ISO 22989:2022

5 Information or data that is in a format that can be easily processed by a computer without human intervention

6 How much data from the public Internet is used for training LLMs? , Michael Humour, 2023

7 A Comprehensive Survey of Continual Learning: Theory, Method and Application, Liyuan Wang et al., 2024

8 ISO/IEC/IEEE 15288:2023— Systems And Software Engineering – System Life Cycle Processes

Advancements in AI[9,10] are rapidly changing the state of the art. The amount and diversity of data[11] used for the learning process of AI systems are growing at a fast pace to keep up with the different and more sophisticated uses of this new technology.

Many AI applications use personal data[12]. Thus, to face the ever-increasing legal and privacy-related risks posed by AI when it uses personal data, the application of appropriate data protection by design and by default[13] measures is essential for protecting the fundamental rights of individuals. In this context, Privacy Enhancing Technologies, commonly known as PETs, could play an increasing role.

PETs[14] are a variety of techniques to improve privacy and control over personal data and can be put in place, in the training phase of AI's development, amongst others. In this context, Federated Learning could be considered as a form of PET and, if applied correctly, could be used in combination with some other PETs to provide further protection whenever processing personal data.

3. Federated learning

3.1 What is Federated learning?

Federated learning (FL) is a type of machine learning where multiple sources of data (devices or servers) collaborate to train a shared model while keeping data decentralised. Instead of sending raw data to a central server (when there is one), each source processes its own data locally and only shares model updates (e.g. gradients or weights[15]).

9 Defining Artificial Intelligence 2.0, Samoili, S. et al., 2021

10 The European's Commission's High-Level Expert Group on Artificial Intelligence - A Definition of AI: Main Capabilities and Scientific Disciplines, 2018

11 For example, image classification can require millions of images; large language models are typically trained on billions or trillions of tokens...

12 The impact of the General Data Protection Regulation (GDPR) on artificial intelligence, Professor Sartor and Dr Francesca Lagioia, 2020

13 Article 27 of the Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data, and repealing Regulation (EC) No 45/2001 and Decision No 1247/2002/EC and Article 25 of the Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

14 OECD (2023), Emerging privacy-enhancing technologies: Current regulatory and policy approaches, OECD Digital Economy Papers, No. 351, OECD Publishing, Paris, 2023

15 Gradients represent the direction and rate of change of a function, specifically related to how a small change in input affects the output. Weights are numerical parameters that determine the strength of the connections between neurons in a neural network.

Data and the learning process are two key pieces of building AI systems.

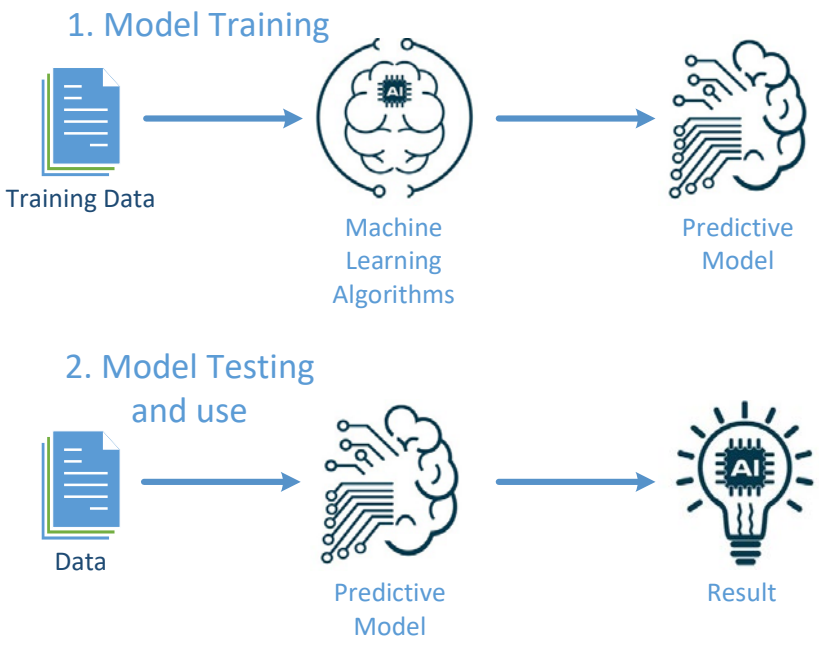


Figure 1 Machine Learning[16]

Originally, in ML, the data and the learning process were centralised[17] i.e. the data was located in a specific data centre where the learning process occurs.

Over time, the learning process became a too-costly process in time, computing power and storage to be centralised on a single machine. Therefore, developers ended up uploading and storing their datasets in cloud services. Such services are optimised to distribute data and processing across multiple machines.

Federated Learning with central server coordination

In a FL setting with central server:

1. A central server or service provider sends a pre-trained or initial ML model to each federated client device.
2. Each of these client devices locally trains its ML model with its own data resulting in multiple locally trained ML models.

16 Icons courtesy of https://vecteezy.com
17 Federated learning, Qiang Yang and al., Morgan and Claypool Publishers Preface (xiii) and Introduction (p1)

3. Each of the locally trained models sends back to the central server either its parameters or the updates of those parameters[18].
4. The central server compiles all the data received from local models to produce one combined model. In some FL cases[19], a voting system is implemented to select the local models that will be integrated (the ones that produce the best results).
5. The central server then sends this combined model to all the multiple client devices again.
6. This process is repeated a fixed number of times or until the central model performance reaches a performance threshold.

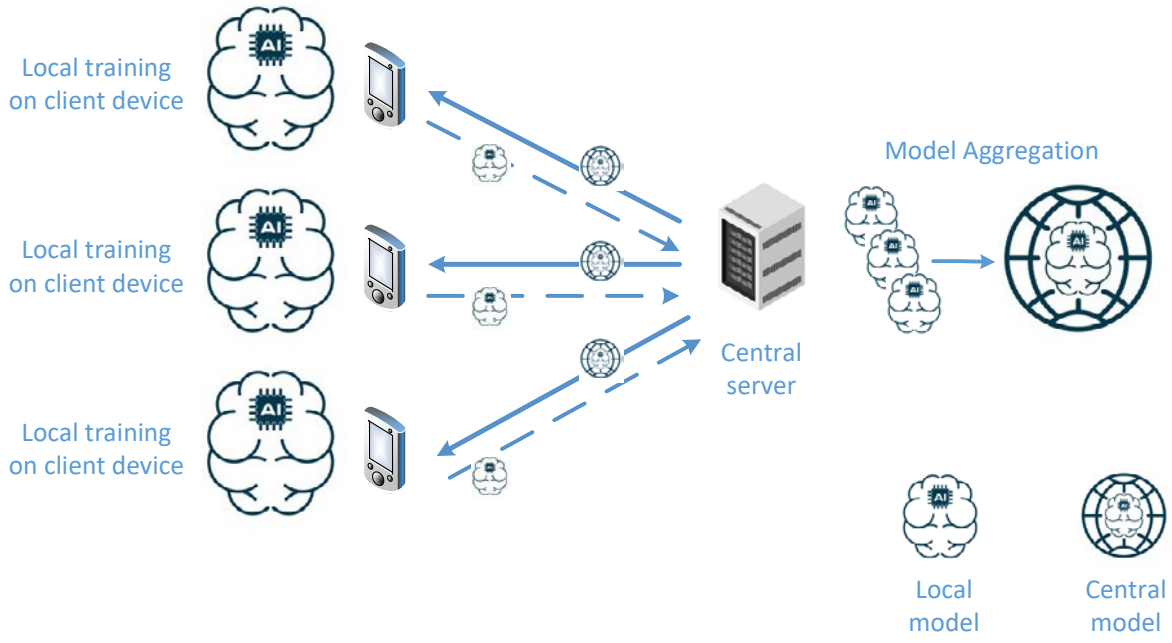


Figure 2: Federated Learning[20]

It is an application of the compute-to-data strategy, which means taking the process to the data, instead of taking (or moving) the data to the process, using techniques similar to those used in large cloud service providers for increasing efficiency.

18 A systematic review of federated learning from clients' perspective: challenges and solutions, , Yashothara Shanmugarasa et al., 2023
19 Robust federated learning with voting and scaling, Xiang-Yu Liang et al., 2024
20 Icons courtesy of https://vecteezy.com

Federated Learning without central server coordination

FL can also be completely decentralised (Decentralised Federated Learning (DFL)). In this case, there is no central server. Each client device builds its local model and exchanges its parameters directly with other client devices via a peer-to-peer network architecture that does not include a central server.

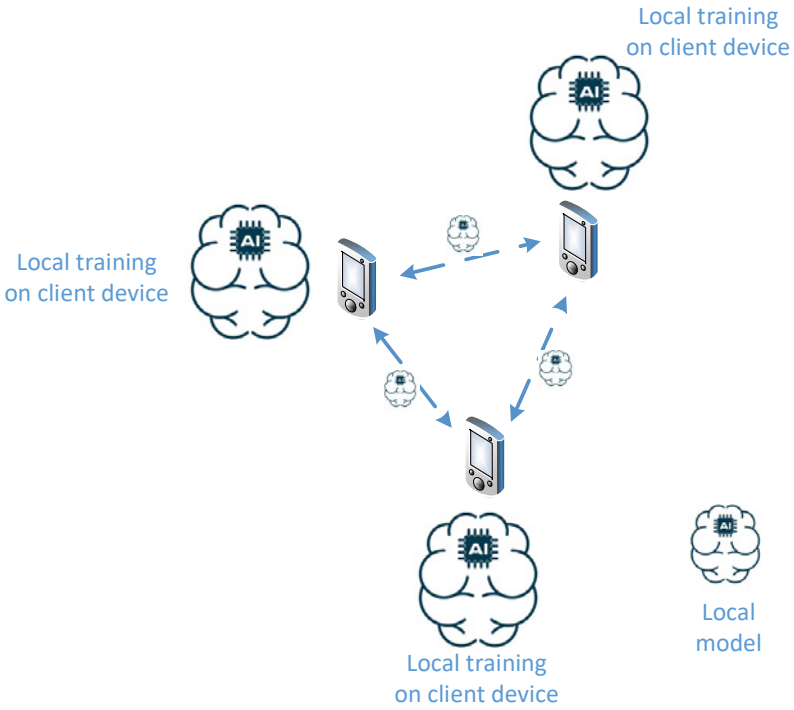


Figure 3: Decentralised Federated Learning[21]

3.2 How can Federated Learning Models be classified?

3.2.1 Horizontal vs Vertical

FL models can be classified as per the following[22].

- **Horizontal Learning:** In horizontal learning, the data held by the different client devices share the same features, i.e. every client device uses the same data structure (see Figure 4: Horizontal Learning Data).

21 Icons courtesy of https://vecteezy.com
22 Understanding the types of Federated Learning, Openminded blog

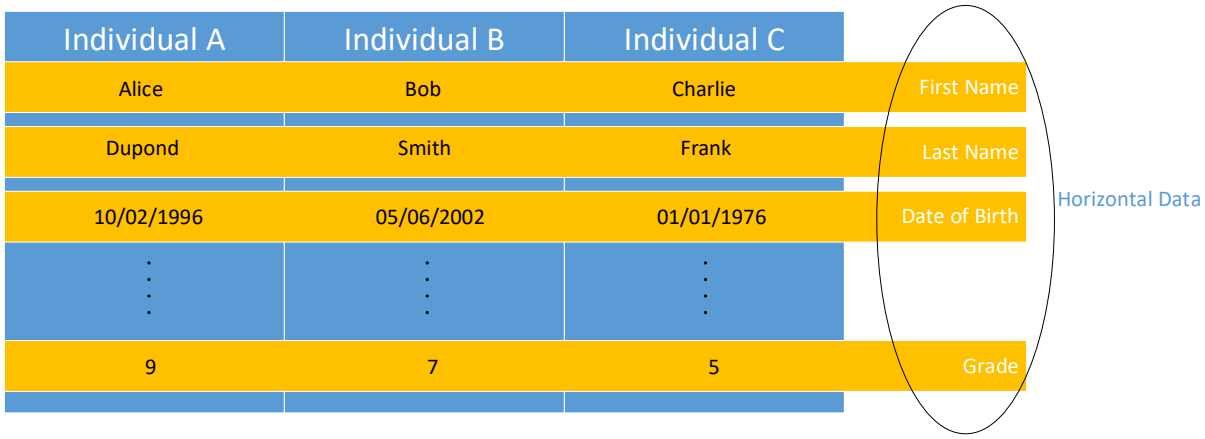


Figure 4: Horizontal Learning Data

- **Vertical Learning:** In vertical learning, data across devices can hold data on the same entity (e.g. individuals) albeit with different types of information as shown in Figure 5: Vertical Learning Data.

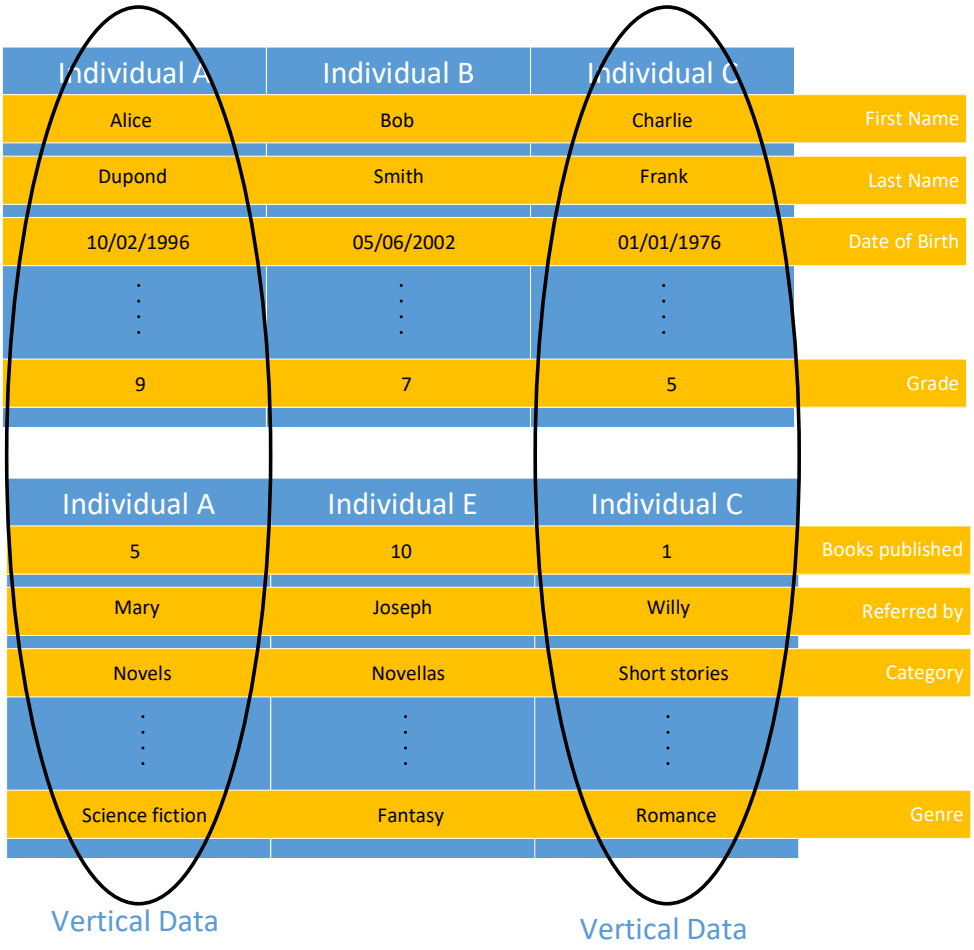


Figure 5: Vertical Learning Data

3.2.2 Cross devices vs Cross silos

Furthermore, it is possible to classify FL systems depending on their type of clients.

- Cross devices: Clients are individuals with personal devices (e.g. smartphones, wearables...) in large numbers. Data held in each device is limited to those generated by its own user. In some cases, new data is generated dynamically while older data is removed. For some use cases, like the adjustment of a model to the specificity of an individual, cross device FL fits well.

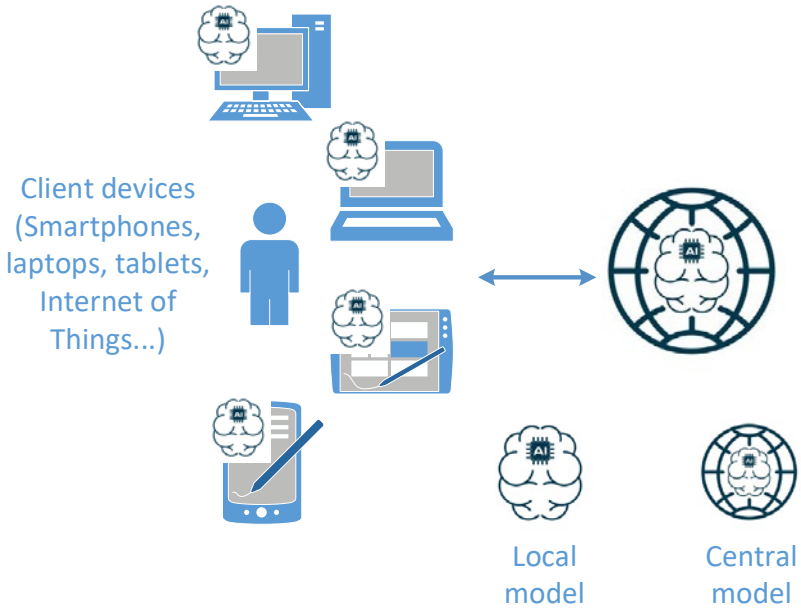


Figure 6: Cross Devices FL (clients are individuals within an organisation or not)[23]

- Cross silos: The clients are organisations (e.g. banks, hospitals...) in small number. The data held by each is big in quantity and one set of data from one organisation can (but not necessarily) have the same overall characteristics as all data across all organisations.

23 Icons courtesy of https://vecteezy.com

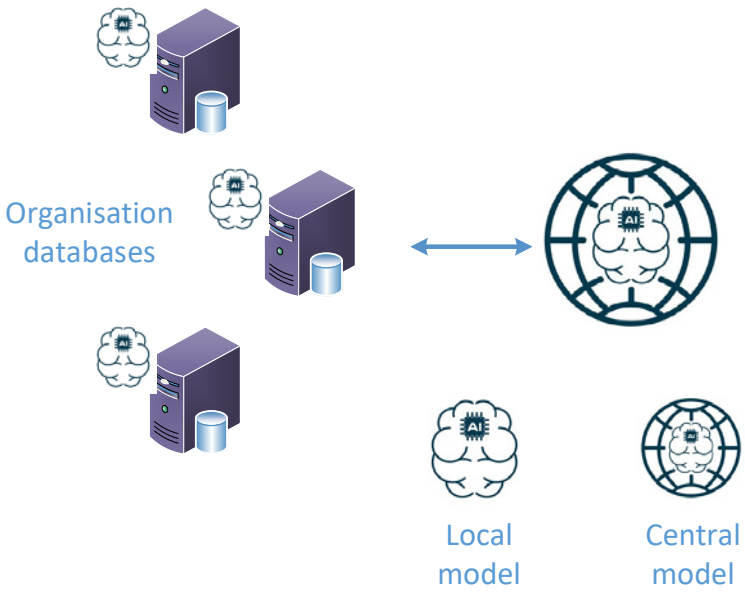


Figure 7: Cross Silos: clients are organisations and not individuals[24]

3.3 Examples of Federated Learning use cases

As mentioned previously, FL is an interesting alternative to centralised learning when development of AI systems requires data from different sources but it is not possible or desirable to share this data. The applications can be diverse; some of the most common use cases are presented here in order to exemplify the potential of the technology.

3.3.1 Healthcare AI Models

Organisations in the field of healthcare (such as hospitals, medical research institutions...) can build an AI system using FL to avoid sharing sensitive data (e.g. patient medical data) with third parties. In this context, getting big enough training datasets may be difficult[25,26,27,28] due to privacy concerns (medical data is a special category of data, subject to a particular legal regime) or low number of samples in each organisation (e.g. certain hospitals might not have a lot of patients with a specific disease that needs to be studied and for which AI could help). Where AI systems cannot get sufficient training data, FL can help improve the performance

24 *ibid*
25 Bridging federated learning theory and practice with real-world healthcare data, Jean Ogier du Terrail et al., 2022
26 A Systematic Review of Federated Learning in the Healthcare Area: From the Perspective of Data Properties and Applicationst, Prayitno et al., 2021
27 Federated learning project connects pharma with university to train AI model
28 Federated learning for medical imaging radiology, Muhammad Habib ur Rehman et al., 2023

of AI i.e. to properly train an AI system on data relating to a specific disease coming from different health services. For example, this was achieved to help fight against cancer[29] (brain tumour) where the heterogeneity of the disease, the tissue preparation and staining processes made it impossible to find a single central location containing sufficient data for training AI systems. Pooling resources together using FL solved this issue without the necessity to move all data to the central location.

3.3.2 Speech models

FL has also been used to train a speech model[30,31] using as client devices smartphones and wearables (e.g. smart watches) to avoid sharing their voice data with the service provider (e.g. voice recognition or predictive text).

3.3.3 Autonomous transport systems

Yet another example is in the field of autonomous transport systems (e.g. cars) where each vehicle can send the local models' parameters to a central model without revealing the data subject's personal data (such as location data) in order to improve an AI model used for purposes such as object detection and route planning[32]. This was done to forecast traffic conditions, identify pedestrian behaviour, and assist drivers in making decisions.

3.4 Technical challenges for the implementation of FL systems

Computing resources. For some situation in the cross devices scenario, the training process could be taxing in terms of computing resources for the client devices if their computational power is limited. By design, some client devices might have less processing power than servers, which can integrate multiple processing units (CPUs, GPUs, and TPUs), elements performing calculations in a computer[33].

Efficiency. Depending on the use case of FL, communication performance could be an advantage or disadvantage. When FL is used in a scenario with several massive databases in

29 A Systematic Review of Federated Learning in the Healthcare Area: From the Perspective of Data Properties and Applications, Prayitno et al., 2021
30 https://support.google.com/assistant/answer/11140942?hl=en#zippy=%2Cfederated-learning
31 Federated Learning for mobile keyboard prediction, A. Hard et al., 2019
32 Real-time End-to-End Federated Learning: An Automotive Case Study, Hongyi Zhang et al., 2021
33 A Survey of Three Types of Processing Units: CPU, GPU and TPU, Goran S. Nikolić et al., 2022

different clients (usually corporations or public bodies), and the local model is smaller than such databases, the communication can be more efficient because there is no need for massive data communications. Otherwise, in cases of collection of real time data for natural persons, communication can be less efficient depending on the size of the model. In both scenarios, the training process could be more efficient because there is no need for centralised high processing power; the processing is distributed across multiple machines (in case of physical users, using the edge-computing paradigm).

Complexity. Client devices can vary in terms of size, computing power, communication means, architecture..., which makes establishing a FL training a complex task[34]. Furthermore, scaling FL to a large number of clients can introduce additional complexities, including managing client participation (e.g. client dropouts or failures during the training process), handling stragglers (slow or unreliable clients) and balancing load.

Convergence. Achieving fast and stable convergence in FL settings should be managed due to the possible asynchronous nature of updates and the non-Independent and Identically Distributed[35,36] (IID) distribution of data.

4. Where in a FL architecture can personal data processing occur?

Given the architecture of FL systems, there are three stages where personal data might be present.

- Personal data might be processed **within each device** (the training data might contain personal data).
- Personal data might be **exchanged** between devices (the data shared between devices are the weights and/or gradients of the ML models).
- Personal data might be processed **within the ML models** (both in the central model when it exists and in local models).

34 Heterogeneous Federated Learning: State-of-the-art and Research Challenges, Mang Ye et al., 2023
35 I.e. there are no overall trends.
36 Independent and Identically Distributed (IID) Data Assessment in Federated Learning, Arafeh, Mohamad & Hammoud et al. 2022

4.1 Is personal data processed locally in each device?

In FL, the training data is collected and processed on each participating device to train a local model (when there is a central server, a baseline model is provided first by this central server). If this training data includes personal data, that personal data could potentially also be partially memorised in the resulting local model.

The decentralised nature of FL does not exempt AI system providers using FL from ensuring compliance with the applicable personal data legislation. Consequently, providers using FL will need to consider appropriate safeguards to protect personal data on each device.

Any local processing of personal data in a FL setup needs to be performed, among others, in such a way as to guarantee a valid legal basis, provide transparency, ensure data minimisation, and implement robust security measures to protect the personal data from unauthorised access or modifications.

4.2 Potential personal data exchange among the devices in the federation

In FL, communication between devices inherently involves the transmission of knowledge derived from the training data present in each device. The data exchanged consists of gradients and/or weights that encapsulate the knowledge and patterns learned from the training data. These gradients and/or weights are necessary for aggregating local updates into a comprehensive global model. However, the question whether this data (gradients and weights) transmitted among the devices enables the processing of personal data needs to be assessed on a case by case basis by the controller.

Given that only weights and/or gradients are shared, FL has lower risks from a personal data protection point of view than exchanging the full training datasets. Reconstructing training data from data exchanged in a FL setting is complicated and will only work for a fraction of the training data (the difficulty level and rate of success depend on the FL setup).

To verify if weights and gradients would enable the processing of personal data, it is necessary to answer the question: "Can information related to an individual whose personal data was used during the training be extracted from the information exchanged among the devices?" The answer is not always evident. Reconstructing the original training data from gradients and weights will not be straightforward and in most cases may not be possible.

This risk needs to be determined at the inception of the system. This is important since, if weights and gradients allow reconstructing personal data used during the training phase, appropriate safeguards must be in place on a case-by-case basis[37].

4.3 Can information related to an individual be extracted from the resulting model?

As with the exchanged information, the response is not straightforward and requires a case-by-case analysis. ML models can retain features and correlations from training data samples[38]; they can be attacked to reconstruct personal data used in the training phase (extraction attacks) or to infer if specific data samples were present in the training dataset[39,40] (membership inference attack[41]). For example, if we ask a Large Language Model (LLM) about a public figure it could be possible to retrieve personal information and sometimes even produce an image of that person.

Given that these reconstructions and extraction of personal data from ML models are possible, it can be concluded there is a risk that part of the personal data used in the training could be extracted from the resulting ML models. Thus, an assessment should be performed on a case-by-case basis to determine whether the ML models need to be considered as personal data.

For more information on the circumstances under which AI models could be considered anonymous and the related demonstration, see section 3.2 of the EDPB Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models[42].

37 A review of federated learning: taxonomy, privacy and future directions, Ratnayake, H. et al., 2023

38 Wei, J., Zhang, Y., Zhang, L. Y., Ding, M., Chen, C., Ong, K. L., ... & Xiang, Y. (2024). Memorization in deep learning: A survey. arXiv preprint arXiv:2406.03880

39 Fang, H., Qiu, Y., Yu, H., Yu, W., Kong, J., Chong, B., ... & Xia, S. T. (2024). Privacy leakage on DNNs: A survey of model inversion attacks and defences. arXiv preprint arXiv:2402.04013.

40 Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., ... & Lee, K. (2023). Scalable extraction of training data from (production) language models. arXiv preprint arXiv:2311.17035.

41 Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P. S., & Zhang, X. (2022). Membership inference attacks on machine learning: A survey. ACM Computing Surveys (CSUR), 54(11s), 1-37.

42 EDPB Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models https://www.edpb.europa.eu/our-work-tools/our-documents/opinion-board-art-64/opinion-282024-certain-data-protection-aspects_en

5. What are the data protection benefits and challenges of FL?

Where personal data is used for training, in non-federated models, each device will collect and transmit directly this data, while in FL systems each device will train a local model and will only transfer the result of the training (weights and parameters). This avoids both the exchange and the direct processing of personal data by the central system or other federated devices and mitigates the relevant data protection risks.

FL can potentially bring certain advantages from a personal data protection perspective in comparison with ML centralised processes. However, it should not be taken for granted that FL solves all the problems as some risks will persist.

5.1 Benefits of FL over centralised ML systems from a personal data protection point of view

5.1.1 Transfer/Data Minimisation

Due to its nature, FL may contribute to implementing the principle of data minimisation (a core principle of data protection) because instead of sending the whole dataset[43] to other parties only the model parameters or their updates are transmitted.

5.1.2 Enhanced Accountability

FL may supports controllers in implementing the accountability principle: they would potentially be able to better control the access to personal data, thus also avoiding any possible unlawful re-purposing of the processing.

43 However, personal data still needs to be processed on the client devices. FL does not influence the amount of training data to be used locally.

5.1.3 Safer Sensitive Data Processing (including special categories of data)

FL allows the processing of different categories of personal data (including sensitive data) without the need to share data with the other parties. Due to the local processing and no data sharing, FL helps to reduce risks for individuals' rights and freedoms, mainly in cases of massive processing of special categories of personal data (Article 9 GDPR[44]), to get a more positive assessment of the proportionality principle in the context of the DPIA to be carried out pursuant to Article 35.7.b GDPR and to ensure accountability. However, to ensure fair processing, and avoid bias from existing patterns in training data, safeguards are needed to detect and mitigate bias present in the source data[45].

5.1.4 Consent management

FL allows having a better control of a data subject's personal data increasing transparency (what data, what for, when). Ideally, the data subject will be able to better verify what use is made with their personal data as FL enhances control and sovereignty over their own environment. Thus, as the training data remains on the devices, FL simplifies consent management.

5.1.5 Data Security

In a cross-silo scenario, FL could help reduce the reluctance of organisations with large volumes of data to reveal private information. This could help implement collaborative data sharing scenarios e.g. data spaces, reduce risks and thus better leverage potential benefits. Additionally, given that there is no central storage of personal data, it is very unlikely that a personal data breach would affect all personal data used in the training of the model.

However, given that there are risks that personal data is reconstructed using the gradients or weights, or the local/central models (see Chapter 4), an analysis should be performed and the individual be appropriately informed of these risks.

44 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

45 EDPS's Generative AI and the EUDPR. First EDPS Orientations for ensuring data protection compliance when using Generative AI systems.

5.2 Challenges of FL over centralised ML systems from a personal data protection point of view

5.2.1 Training data quality management

In software engineering, data quality could be defined as the degree to which data satisfies the requirements of its intended purpose[46]. This means that a ML training dataset has enough quality if it is possible to develop a ML system fulfilling its performance requirements. It is important not to confuse the data quality of the training data set (in particular, the characteristics of accuracy and precision) with the GDPR accuracy principle. For example, the use of anonymisation techniques on a training dataset might render some of its data inaccurate (from the data protection point of view) but the dataset might still have enough quality to be an input in the ML training process.

In a non-FL setting (setting where data is centralised), before starting the training process, a distributed assessment of each source of data should be performed in order to assess its quality level. This could be done by checking some quality characteristics like completeness, credibility of the sources, currentness[47], compliance and others[48], and checking whether it is even possible to carry out a successful process to increase the data quality level. Data that does not reach a certain level of quality are not necessary for the training process. Once the training dataset is consolidated, other quality assessments (like consistency between records) could be carried out.

In an FL setting, checking data quality is more difficult as the data sources are not centralised and not transmitted. Thus, each source of data cannot be compared against the other data sources (no cross-source data quality checks), there are no possibilities to check the data quality of all the training data as a whole and it might be difficult to check the credibility of each data source. In FL, specific distributed data quality management procedures should necessarily be implemented.

46 ISO/IEC 25012 data quality model
47 Currentness, as defined in ISO 25012:2008, means the degree to which data has attributes that are of the right age in a specific context of use
48 ISO/IEC 25012 data quality model

Some methods for assessing and improving FL data quality include the following[49].

- Data distribution: The central server (when there is one) can request statistical information of the local data sets (that were used to train local models) to determine whether to give more weight (i.e. relevance) to the local models' parameters with higher value. For example, the Private Set Intersection (PSI) method[50] allows calculating the common elements of different data sets without exchanging these data sets[51]. To a certain extent, the higher the statistical similarity, the greater the assurance the central server has that the data used to train the models is of sufficient quality.
- Model utility: The quality of a local model can be assessed based on its utility i.e. how much the local model impacts the quality of the predictions of the next iteration central model (the central model resulting from the use of local models' parameters). To achieve this, the impact on the quality of the current central model is assessed. A local model's parameters are integrated and the central model is reassessed. If the quality of this new central model has improved, then it can be asserted that the local model was trained on quality data.
- Statistical metrics: Local models can be evaluated based on statistical metrics. Usually, this is done calculating the distance between model parameters before and after rounds of training; some argue that if the model parameters distance is higher when using Independent and Identically Distributed (IID) data[52], then the model can be considered of better quality; others[53] show that for non-IID data the opposite may be true.

In any case, when the ongoing training data is collected in data streams (i.e. each new data is immediately added to the training dataset) and processed in real-time, other specific solutions should be adopted to guarantee the data quality in both centralised and FL settings.

49 A Survey of Federated Evaluation in Federated Learning, Behnaz Soltani et al., 2023
50 Private set intersection: A systematic literature review, Daniel Morales et al., 2023
51 Practical Private Set Intersection Protocols with Linear Computational and Bandwidth Complexity, E. De Cristofaro, 2009
52 The logic is that if the distance between the model parameters before and after a training round is larger, the model has adjusted more significantly, which might indicate that it is learning effectively from the data. Conversely, small parameter updates could suggest that the model is converging or that the data might not be providing new, informative updates. However, this is a nuanced metric and can depend on the specific context. In some cases, excessively large parameter changes might indicate instability or noisy gradients, which could harm the model's convergence.
53 Federated Learning with Non-IID Data, Yue Zhao et al., 2022

5.2.2 ML output accuracy and bias

Both in FL and non-FL ML training, developers should ensure that the final ML model is free of bias and remain free of bias (this is a continuous process). In case of FL, the difficulty comes from the implementation of a distributed training data quality management process. Some of the available mitigation techniques are:

- ensuring that extract and transform data operations work properly in each site (like sensors or format translators);
- ensuring that sampling and normalisation processes for each local ML model are consistent;
- monitoring the statistical distribution of local training data and locally rebalancing their statistical representativeness; this need to be done until some uniformity is reached in a set of participants before the training or update of the global model can start.

5.2.3 Integrity

In a FL environment, ensuring that data is not unduly modified is necessary in order for the resulting central (or distributed) model to be accurate. Compared to non-FL architectures, FL has different threat vectors[54] that can affect the integrity of the data. This is because, in FL architectures, there are multiple devices participating in the overall system and thus, there are multiple models and data transfers to defend (the local models and, where applicable, the central model).

One way to attack the local or central models is to perform Data Poisoning[55,56]. Data poisoning is an attack where false data is injected into the training process of any device to bias an AI system as a whole and reduce its performance. Usually, this can be mitigated by detecting outliers (by analysing the local model updates received from devices for statistical anomalies)[57,58].

Unduly modifying local model updates (Model Poisoning), on the client devices or in transit, would also have detrimental effect on the global model in terms of integrity.

54 Various methods or pathways that attackers use to gain unauthorised access to data.
55 Robustness and Explainability of Artificial Intelligence, Hamond R. et al., 2020
56 Detection and Prevention Against Poisoning Attacks in Federated Learning, V. Valadi et al., 2022
57 Precision Guided Approach to Mitigate Data Poisoning Attacks in Federated Learning, K. N. Kumar et al., 2024
58 Detection and Prevention Against Poisoning Attacks in Federated Learning, V. Valadi et al., 2022

In response to poisoning attacks, researchers propose passive and active defences. Passive defences start by analysing the aggregation of the models on the server side (designing relevant aggregation model strategies), thereby improving the global model performance. Active defences eliminate the impact of the poisoning model on the global model by detecting the performance of the local model and eliminating the poisoned model. Currently, active defences seem to be the most promising trend[59].

Protecting the client devices remains difficult in cross device FL as, in general, participants owning those devices lack the expertise, resources and maturity that organisations can put into securing their FL settings.

5.2.4 Confidentiality

FL offers improved confidentiality for the controller of each of the devices since it does not require sharing the raw training data with the rest of the devices in the ecosystem. At the same time, a FL setting offers the opportunity to attack the local models as they are trained by the devices and attacking the weakest link can put at risk the whole structure. FL local models need to be stored in the original location (devices) and later transmitted to the central location[60]. They then could be hacked on the devices, or they could be analysed in transit or at destination. After receiving the initial pre-trained model, the local models start by being trained on local data sets and are more susceptible to disclose what data (or a subset of this data, including potentially personal data) was used to train them. This can happen because the local models can preserve characteristics and correlations from training data samples that attackers could use to reconstruct or extract records. Thus, it is possible to build an attacker model[61] that tries to guess the training data of the local models or observe the model changes over time to help determine the personal data training data sets.

In order to protect against such attacks, safeguards should be implemented on

- the data at rest and the models on the client devices;
- the communication between client devices and the central server (or between client devices in a DFL approach);
- the central server itself (when there is one), containing the intermediary and final models.

59 Challenges and Countermeasures of Federated Learning Data Poisoning Attack Situation Prediction, Wu J et al., 2024
60 in a non-DFL architecture
61 Leak and Learn: An Attacker's Cookbook to Train Using Leaked Data from Federated Learning, Joshua C. Zhao et al., 2024

These safeguards can include the following.

- The use of **encryption** on the data at rest on client devices in order to mitigate attacks that could directly compromise those devices.
- The use of **Secure Multi-party Computation (SMPC) or Immediate and Secure Aggregation**[62,63,64] to limit data exposure. These are cryptographic methods that allow multiple parties to compute on distributed data without revealing individual data points. Calculations are done on encrypted parameters without ever revealing the parameters.
- The use of **Trusted Execution Environments (TEE)**[65], which allows the data to be processed within a "secure piece of hardware" and uses cryptographic protections to enable a protected computing environment. These are used for example in secure payment transactions.
- The use of **Differential Privacy**. It consists in adding noise to the data to reduce risk that any individual person can be identified during the training phase of the local models.

Due to the complexity of FL settings and given that no PET is a silver bullet, controllers should consider implementing available "classic" security measures to protect data as they would in any other processing operation, in order to minimise the risks.

62 Fair and Secure Multi-Party Computation with Cheater Detection, Minhye Seo, 2021
63 https://securecomputation.org/
64 SMPAI: Secure Multi-Party Computation for Federated Learning, Vaikkunth Mugunthan et al., 2019
65 Trusted Execution Environments: Applications and Organizational Challenges, Tim Geppert et al., 2022

6. Conclusion

FL offers a promising approach to machine learning by enabling multiple devices to collaboratively train a shared model while keeping data (including personal data where applicable) decentralised. This method is particularly advantageous for scenarios involving the processing of sensitive personal data or regulatory requirements, as it mitigates privacy risks by ensuring that raw personal data remains on local devices. By keeping personal data decentralised, FL aligns with core data protection principles such as data minimisation, accountability and security, and reducing the risk of large-scale personal data breaches.

In non-FL environment, a case-by-case assessment is needed to determine the risk of re-identification attacks (as membership attacks) in the final models, but in FL environments, such assessment should be done in the local models interchanged too.

FL presents challenges that need to be addressed to ensure effective protection of data. One major concern is the potential for data leakage through model updates, where attackers might infer information from gradients or weights shared between devices and central servers. This risk, along with potential membership inference attacks and the difficulty in detecting and mitigating bias or ensuring data integrity, highlights the need for robust security measures throughout the FL ecosystem and for the combination of FL with some other PETs.

7. Recommended Reading

- L. Tian, A. Kumar Sahu, A. S. Talwalkar and V. Smith, **Federated Learning: Challenges, Methods, and Future Directions**, IEEE Signal Processing Magazine 37, 2020.

- Q. Li, W. Zeyi, H. Bingsheng, **A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection**, ArXiv abs/1907.09693, 2021.

- P. Kairouz et al, **Advances and Open Problems in Federated Learning, Foundations and Trends in Machine Learning Vol 4 Issue 1**, 2021.

- Nicole Mitchell and Adam Pearce, **How Federated Learning Protects Privacy**, November 2022 – available at https://pair.withgoogle.com/explorables/federated-learning/

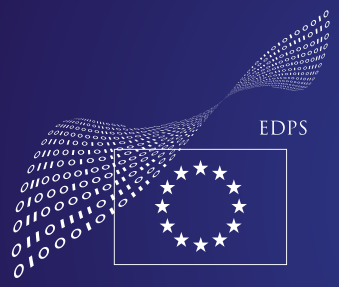
This publication is a brief report produced by the Technology and Privacy Unit of the European Data Protection Supervisor (EDPS) and the Technological Innovation Division (División de Innovación Tecnológica) of the Spanish Data Protection Agency (Agencia Española de Protección de Datos, AEPD). It aims to provide a factual description of an emerging technology and discuss its possible impacts on privacy and the protection of personal data. The contents of this publication do not imply a policy position of the EDPS.

Issue Authors: Andy Goldstein, Miguel Peñalba, Luis de Salvador Carrasco Editors: Luis Velasco, Luis de Salvador Carrasco, Massimo Attoresi and Xabier Lareo.

Contact: techmonitoring@edps.europa.eu

To subscribe or unsubscribe to TechDispatch publications, please send a mail to techmonitoring@edps.europa.eu.

The data protection notice is online on the EDPS website.



edps.europa.eu

www.aepd.es