https://doi.org/10.51483/IJDSBDA.1.2.2021.1-13 ISSN: 2710-2599

SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Data Science and Big Data Analytics

Publisher's Home Page: https: / / www.svedbergopen.com /

Research Paper **Open Access**

Federated learning for privacy-preserving data access

Malgorzata Smietanka[1*], Hirsh Pithadia[2] and Philip Treleaven[3]

[1]University College London, Gower St, Bloomsbury, London WC1E 6BT, United Kingdom. E-mail: malgorzata.wasiewicz.17@ucl.ac.uk
[2]University College London, Gower St, Bloomsbury, London WC1E 6BT, United Kingdom. E-mail: H.Pithadia@cs.ucl.ac.uk
[3]University College London, Gower St, Bloomsbury, London WC1E 6BT, United Kingdom. E-mail: p.treleaven@ucl.ac.uk

Article Info

Volume 1, Issue 2, May 2021
Received : 20 December 2020
Accepted : 13 April 2021
Published : 05 May 2021
doi: 10.51483/IJDSBDA.1.2.2021.1-13

Abstract

Federated learning is a pioneering privacy-preserving data technology and also a new machine learning model trained on distributed data sets. Companies collect huge amounts of historic and real-time data to drive their business and collaborate with other organizations. However, data privacy is becoming increasingly important because of regulations (e.g., EU GDPR) and the need to protect their sensitive and personal data. Companies need to manage data access: firstly within their organizations (so they can control staff access), and secondly protecting raw data when collaborating with third parties. What is more, companies are increasingly looking to 'monetize' the data they've collected. However, under new legislations, utilizing data by different organization is becoming increasingly difficult (Yu, 2016). Federated learning pioneered by Google is the emerging privacy- preserving data technology and also a new class of distributed machine learning models. This paper discusses federated learning as a solution for privacy-preserving data access and distributed machine learning applied to distributed data sets. It also presents a privacy-preserving federated learning infrastructure.

Keywords: Federated Learning, Privacy, Machine Learning, Data Science, InsurTech

1. Introduction

1.1. Privacy-preserving data challenges

As discussed, companies are increasingly collecting huge quantities of historic and real time data (e.g., business, financial, economic, social media, alternative[1]) and also increasingly collaborating on joint data analytics with third parties. Privacy of this very valuable data poses many challenges:

- **Internal data management:** Companies need to control data-access and be able to store data in a way that will enable them to do analytics on those data.
- **Collaboration with third parties:** Companies recognize the value of collaboration. They need a way to do that without compromising sensitive 'raw' data access.

[1] Alternative data – Data gathered from non-traditional information sources (e.g., mobile/wearable devices, IoT devices, transportation, satellites, etc.).

* *Corresponding author: Malgorzata Smietanka, University College London, Gower St, Bloomsbury, London WC1E 6BT, United Kingdom. E-mail: malgorzata.wasiewicz.17@ucl.ac.uk*

- **Monetizing the data:** Companies are increasingly looking to monetize (anonymized) data that they've collected through business operations.
- **Data analytics:** Need to train AI (machine learning) models on multiple (distributed) data sources.
- **Legislation:** Demand to treat user data securely and comply with national user-privacy laws.

1.2. Data the new 'gold'

Often used quotes are: 'data is the new gold', or 'data is the new oil'. The first acknowledges the monetary value of data; the second that data increasingly 'oils the wheels' of commerce. Google, Amazon, Alibaba and Tencent (WeChat) 'harvest' their users' data. Bloomberg, Thomson Reuters, Refinitiv, etc. are data businesses. In addition, service sector companies are increasingly seeing the commercial potential of selling anonymized data; examples include credit card transactions, loyalty card data, and public transport travel data (Derwisch, 2019).

A notable development is China's social credit system. It is an ambitious initiative to build a database that monitors individual, corporate and government behavior across the country in real time. According to the Chinese government, the system will use big data to build a high-trust society where individuals and organizations follow the law (Koty, 2019).

Data being digital—unlike gold or oil—makes it hugely vulnerable to misuse and theft. Hence, the importance of privacy-preserving data technologies.

1.3. Internal data management

Controlling access to data is a growing priority for organizations. Across industry and government, if an employee is accessing user data it is tracked and monitored. What is more, companies reserve the right to not only terminates an employee for breach of trust (when it comes to data access) but also to sue that employee. Subsequently, global institutions (e.g., Facebook) are motivated to protect client trust and guard against reputational damage from data breaches (Facebook_data_scandal, 2020).

1.4. Collaboration with third parties

Today, companies recognize the value of collaboration in data analysis with partner companies in different sectors. In healthcare, for example, building knowledge on complex medical problems and creating drugs requires companies to gather data from diverse organizations like medical institutions, insurance companies, gyms, and even supermarkets. In addition, collaboration between healthcare institutions and the insurance sector may lead to more accessible insurance products for those who suffer from rare or incurable diseases (Turea, 2019; MELLODDY, 2020).

1.5. Monetizing the data

Companies are increasingly looking to monetize data that they've collected through business operations. This includes financial services, transport companies, energy providers, telecommunications companies, and retailers. An example is credit card companies selling real-time anonymous credit transactions to investment companies (Cohan, 2018). We can distinguish internal and external data monetization depending on how data is used and distributed to produce economic benefit.

1.6. Data analytics

Large amounts of data are often required to train and deploy useful machine learning models in industry. Many enterprises (especially smaller ones) do not have the luxury of gathering enough data themselves for machine learning models. Nowadays however, the total amount of information available online is incalculable, but there are other challenges associated with data (e.g., security, computational resources). To address those challenges distributed machine learning techniques are adopted.

1.7. Legislation

Lastly, there is increased demand to treat user data securely and comply with national user-privacy laws. In addition, so called Data Sovereignty—ensuring data is subject to the laws within the nation where it is collected—is become a major issue for governments (GDPR, 2018).

Examples are GDPR (General Data Protection Regulation) introduced by the European Commission (GDPR, 2018), California Consumer Privacy Act (CCPA), Singapore Personal Data Protection Act (PDPA) or FTC (Federal Trade Commission) in the USA. Sensitive data like financial transactions or medical records need to be stored and maintained by the data owners. They usually exist in isolated silos and free data circulation is prohibited. Failure to adequately address the problem of data fragmentation and isolation while complying with the privacy-protection laws will likely lead to a new AI winter (Yang *et al.*, 2019).

2. Data-driven commerce

Many business are on a 'data' journey: (a) we don't collect data; (b) we collect data and do some analytics; (c) we collect data comprehensively and do detailed analysis; (d) we have totally transformed our business to make it data-driven and digital (e.g. Amazon). Adding to this 'journey'; (e) we now share data with our business partners and collaborate on analysis; and (f) we are now a data provider company (e.g., Credit card companies) selling anonymous company data (Cohan, 2018).

Successful companies are increasingly data-driven; collecting huge amounts of historic and real-time data (e.g., business, economic, social media, alternate) to drive business decisions; but are also opening up their data.

Due to the so-called Black Swan[2] events such as Covid-19, data-driven commerce and automated trading is critical for companies to keep business sustainable. Companies now realize the need to embrace automation with digital infrastructures:

- **Data management solutions** – Effective data management requiring a so-called Big data strategy and infrastructure is essential to facilitate increasingly sophisticated AI-based analytics. Most data exist in isolated silos and are not well managed, while AI projects involve multiple types of data.
- **Automated trading platform** – Digital marketplaces and platforms provide direct engagement with clients, support automation and drive down costs.

To provide context for federated learning we next look at the technologies driving the data revolution.

2.1. Technologies driving the data revolution

Data science technologies divide into:

- **Data technologies –** Includes solutions for data management and collection, as well as services that are based on data generated by both human and machines (e.g., Big data, Internet of Things, Chatbots).
- **Algorithm technologies –** New forms of 'statistics', such as machine learning, computational statistics, and complex systems (e.g., deep neural networks, federated machine learning, Monte Carlo simulation).
- **Analytics technologies –** Covering the application of the data technologies (e.g., natural language, sentiment analysis and behavioral analysis).
- **Infrastructure technologies –** Providing the infrastructure for information management and automation (e.g., Federated Learning, Blockchain-based digital marketplace, computable contracts).

For the detailed description of other data science technologies we refer to (Śmietanka, *et al*., 2020).

We now describe in detail federated learning.

3. Federated learning

Federated learning is typically described as new class of distributed machine learning models, often referred as federated machine learning. However, it is also increasingly acknowledged as a privacy-preserving data technology or infrastructure (Yang *et al*., 2019):

- **Federated machine learning** – Decentralized training approach which enables to collaboratively learn a machine learning model while keeping data sources in their original location. For example, mobile phone users can benefit from obtaining well-trained model without sending their personal data to the cloud.
- **Privacy-preserving data infrastructure** – Framework for building collaborative models, allowing secure communication with collaborating parties (such that data don't leave the owner)**.**

3.1. Federated machine learning

The traditional machine learning model is to gather raw data together (e.g., Cloud) for training. This is characterized as 'taking the data to the algorithm'. In contrast, federated learning is 'taking the algorithm to the data'.

Federated learning by definition aims to build a joint ML model based on data located at multiple sites. The concept was proposed by Google in 2016 in the context of privacy-preserving multiple mobile devices model training (Konečný *et al*., 2017). Google's work concentrate on mobile devices model optimization, where data is distributed over an extremely large number of nodes.

---

[2] Black Swan – A massive unpredictable event that has potentially severe consequences.

There are two processes in federated machine learning: (i) model training; and (ii) model inference. In the process of model training, anonymized information can be exchanged between parties but not the sensitive raw data. The exchange does not reveal any protected private portions of the data at each site. The trained model can reside at one party or be shared among multiple parties. At the model inference stage, the model is applied to a new data instance. Using an insurance example, federated-fraud detection system may receive a new claim from a policyholder insured in a different company. Parties collaborate in classifying the claim as legitimate or fraudulent and on predicting the total future claim amount from this claim.

One way to classify federated learning is according to how data is partitioned among parties contributed to the global ML model (Figure 1) (Yang *et al*., 2019):

- **Horizontal federated learning (HFL) –** Assumes that datasets from different participants share the same feature space but may not share the same sample ID space.
- **Vertical federated learning (VFL) –** Participants share the same ID space but have different feature spaces, while label information is owned by one participant. One possible use case is two e-commerce companies and a bank which collaborate on training a model to recommend personalized loans for users based on their online shopping behaviors through VFL (Yang *et al*., 2019). This use case can easily be adopted to the insurance domain, where NHS collaborates with medical insurance providers to provide personalized medical insurance products.
- **Federated transfer learning (FTL) –** Applies to the scenarios when two data sets differ both in samples and in feature space.

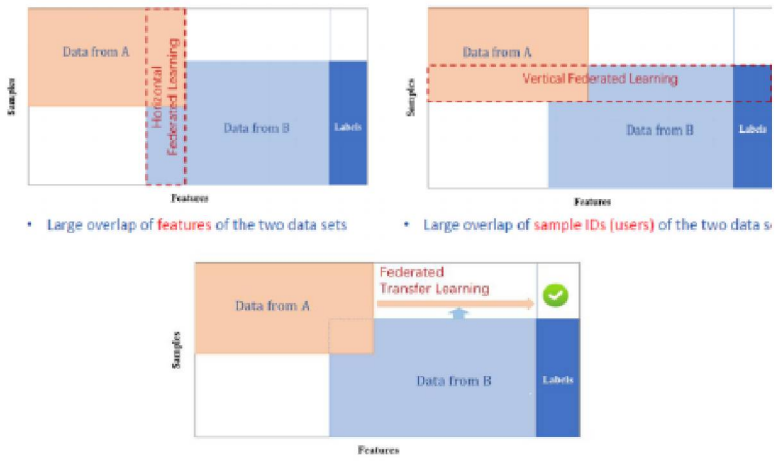

Figure 1: Federated learning HFL vs. VFL vs. FTL

Source: Yang et al. (2019)

3.2. Privacy-preserving data infrastructure

Federated learning is also an algorithmic framework for building ML models characterized by the following features:

- There are at least two parties interested in jointly building an ML model. Each party holds some data that will be contributed to training the model.
- In the model-training process, the data does not leave the party.
- The model can be transferred in a secured way (other party cannot re-engineer the data) under an encryption scheme.
- The resulting model is a good approximation of the ideal model with all data centralized.

Further, we can distinguish three classes of federated learning infrastructure:

- **Closed homogeneous FL systems** – Federated learning across a 'private' network of homogeneous devices (e.g., Google's mobile Android smart phones).

- **Closed heterogeneous FL systems** – Federated learning across a 'private' network of trusted collaborating institutions. An example might be (a) a group of Insurance Companies working on fraud detection model; and (b) banks want to offer insurance product for their credit card holders.
- **Open heterogeneous FL systems** – Federated learning services across a 'public' network of untrusting institutions, where each institution is selling privacy-preserving (analyzed) data based on its sensitive 'raw' data. Examples include: (a) supermarkets selling analyzed loyalty card data; (b) a transport authority selling anonymized travel data of its passengers; (c) telecoms companies selling anonymized data on its subscribers; and (d) hospitals giving analyzed medical data for insurance companies to make rare diseases products more accessible.

4. Federated learning techniques

Federated learning is a Privacy-Preserving Machine Learning (PPML) technique. This is not to be confused with secure machine learning (secure ML). The fundamental difference between the two is as follows (Barreno *et al*., 2006; Yang *et al*., 2019):

- **PPML** assumes that the adversary breaches the confidentiality and the privacy of the system.
- **Secure ML** assumes that the adversary breaches the integrity and availability of the system.

The fundamental differences between these notions are as follows (Yang *et al*., 2019):

- **Integrity** – An adversarial attack on the integrity means the veracity of the system is compromised, e.g., the system may false negative outputs by the system as normal.
- **Availability** – A systemic adversarial attack that renders the system unusable and can lead to classification errors.
- **Confidentiality** – An adversarial results in the leakage of sensitive information, e.g., training data.
- **Privacy** – An adversarial attack results in the leakage of identifiable and attributable information, e.g., identifiable features such as person name or company name in a data set.

A number of cryptographic techniques have been used to mitigate the risk of privacy and confidentiality attacks and facilitate PPML, this includes (a) Secure Multi-Party Computation (SMPC); (b) Homomorphic Encryption; and (c) Differential Privacy.

4.1. Secure Multi-Party Computation (SMPC)

SMPC also known as secure function evaluation (Yao, 1986), involves jointly computing a function from the private input by each party without revealing the value of these private inputs to other parties. In ML terms this "function" could be a model's loss function during training or the model itself (during inference). Cryptographic schemes such as oblivious transfer (Keller *et al*., 2016) and threshold homomorphic encryption (Cramer *et al*., 2000) schemes can be used to facilitate SMPC. However, the most common scheme currently used in PPML is Secret Sharing schemes (Shamir, 1979; Damgård *et al*., 2012).

Secret sharing schemes involve the hiding of a secret value by splitting it into parts and randomly distributing it to the parties involved in the multi-party compute such that each part has only one share of this secret. A threshold number of these individual shares are needed to fully reconstruct the entire secret. Arithmetic secret sharing is the most commonly used in existing SMPC PPML systems.

Most SMPC based PPML consists of two parts: an online and offline phase. The offline phase involves the bulk of the cryptographic operations such as the generation of triples. The online phase involves the training of the ML Model (e.g., using the triples generated in the offline phase). SMPC based PPML is commonly used for a number of reasons (OpenMined, 2020):

- It is less computationally expensive than fully homomorphic encryption.
- It isn't vulnerable to computationally powerful adversaries.
- It can be used to perform inference directly on encrypted data, i.e., without allowing the model owner to see the private data of the owner.

It however has a number of limitations:

- There is a networking and communications overhead.
- It assumes that individual parties are not colluding or at most the n parties are colluding such that $n <$ threshold distributed secret shares, assuming each party has at most one share.

4.2. Homomorphic Encryption

Homomorphic Encryption (HE), first proposed by (Rivest *et al*. (1978) involves the direct computation over a ciphertext, without decrypting the ciphertext. There are a number of HE schemes, often bucketed into three: (a) partially HE schemes, (b) somewhat HE schemes and (c) fully HE schemes that have been used for PPML (their computational complexity grows as the HE functionality grows). Fully HE schemes include the use of Paillier's scheme in the training of logistic regression models through secure gradient descent (Hardy *et al*., 2018) or the use of secure inference of encrypted queries over trained neural networks (Gilad-Bachrach *et al*., 2016). The inference by clients is classified securely by the neural network without inferring information from the query itself.

The biggest advantage of HE is that it can be used to perform inference directly on encrypted data without revealing any information.

It however has the following limitations:

- HE schemes are extremely slow ad require large computational and memory overheads compared to other techniques.
- HE schemes are restrictive as a limited set of computational operations can be performed, these might not restrict the degree to which they can be used for ML.

4.3. Differential Privacy

The premise of Differential Privacy (DP) is to confuse an adversary that may be trying to access individual information from a database such that they cannot distinguish any individual-level sensitivity from it. In the PPML context DP can be used at the local level (Local Differential Privacy, LDP). Each party perturbs their dataset and releases this obfuscated data for model training. DP schemes are usually considered to have a relatively smaller computational overhead compared to SMPC/ HE-based schemes and a number of different techniques have been explored (Papernot *et al*., 2020; McMahan, *et al*., 2017). However, some of the limitations include noise and sensitivity when it comes to very skewed datasets. Moreover trade-offs need to be made especially when considering the privacy loss value, ε[3].

Further, we will describe federated learning architecture.

5. Federated learning architectures

Given that federated learning is a distributed ML process, where private datasets residing within edge devices are used to locally train models (before aggregation), they tend to be orchestrated within distributed computation clusters that often have a peer to peer topology. Within these there are a number of administration paradigms each characterized by the entity that initiates the training process. These entities are the collaborators that want to engage in a FL process. They include (but is not restricted to) the dataset owners and data scientists. Although both want to create more informed models, through the use of private data, they both may have separate requirements from a governance and control perspective.

Consider the perspective of the data scientist, they may want to create more informed models but-because of business risks (not wanting to provide access to datasets that may provide their competitors an edge), or even residual privacy and confidentiality risks, they may not want to share their private datasets with just anyone—they would like to administer the right to access and train models. Assuming that pre-trained models are available, the dataset owner in this case will initiate the training process on their private data. This is a model-centric process where a pre-trained model is used to initiate the process.

Now consider the role of the data scientist, they may want to be able to easily identify and use datasets to solve their problem. Assuming that the data scientist is aware of the existence of a given private dataset readily available for training, they could initiate the training process and train a model themselves. Notice the distinction between the two with respect to administration, although subtle, this has implications when it comes to data governance and policies (such as GDPR).

In the subsection below we discuss these paradigms in detail with respect to one such open source architecture PyGrid (PyGrid, 2020) and further give examples of some Federated Learning workflows and how they interact with the infrastructure.

---

[3] The ε value is a metric of privacy loss at a differentially change in data. A smaller value provides better privacy protection at the cost of noisiness.

5.1. PyGrid

Most data architectures use a peer-to-peer based topology, where a peer consists of one of the organizations/individual that may either want to share their data sets or train models. PyGrid (PyGrid, 2020) is one such open source architecture. It consists of three components:

- **Network** – An application used to manage/monitor and facilitate FL through the network.
- **Node** – An application that is used to store private data/models and control workers.
- **Workers** – These are ephemeral compute instances for carrying out computations.

PyGrid provides a number of ways to engage in FL including the use of Model-centric FL and Data-centric FL.

5.1.1. Model-Centric federated learning

This involves the hosting of the model within PyGrid (i.e., the peer-to-peer network of nodes that want to carry out FL). It involves the following steps (as seen from the Figure 2 below):

1. The peer requests to train a model (step 1).
2. A model and training plan is sent to the peer (step 2).
3. The peer trains this on its private data within itself (steps 3-5).
4. After completion, the delta (the difference) between the new and the old model is generated (step 6).
5. The delta is sent back to the network and can then be averaged into the model (step 7).

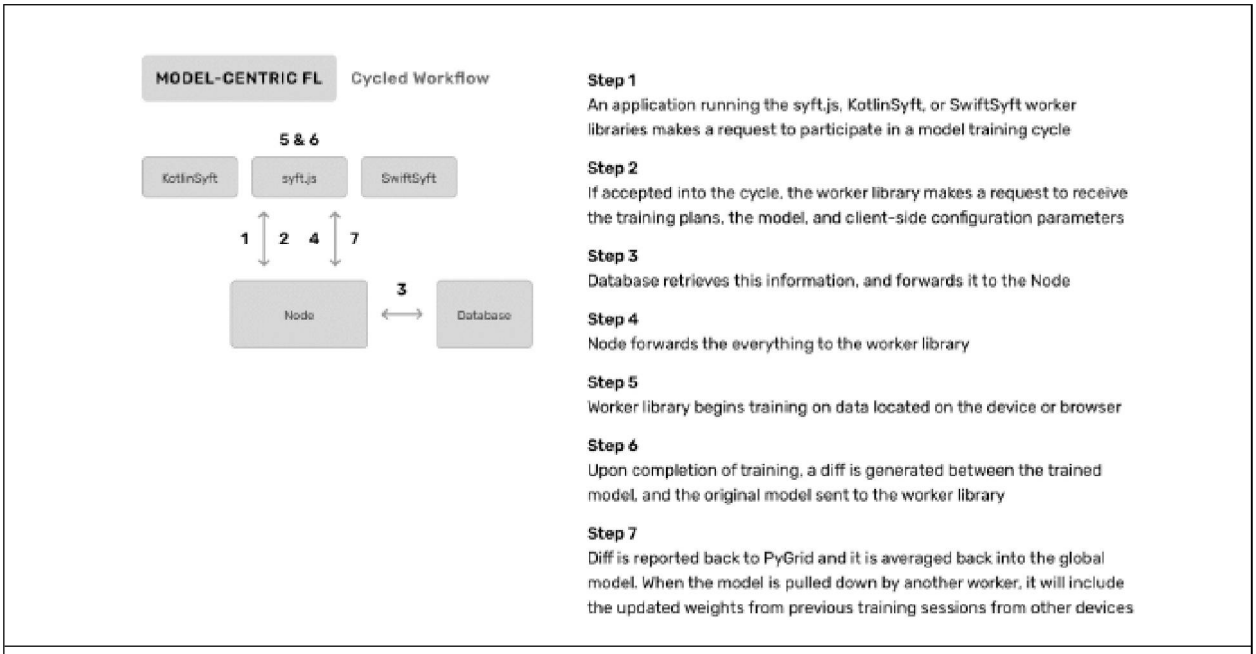

Figure 2: Model-centric FL architecture

5.1.2. Data-centric federated learning

Unlike the previous approach, this involves hosting the data within PyGrid (i.e., within a peer node). This offers the benefit of allowing individuals/organizations (such as data scientists) that may not be the owner of the dataset to request the data for training against their data. It involves the following steps (as seen from the Figure 3 below):

1. Data scientist identifies the data they would like to use.
2. A training plan or a pre-train model is created.
3. The training plan/pre-trained model is sent to the PyGrid as a job.
4. There are then sent to worker nodes by PyGrid along with the required training data.
5. The workers perform the plan on the model using the data.
6. The results are returned to the data node and the data scientist.

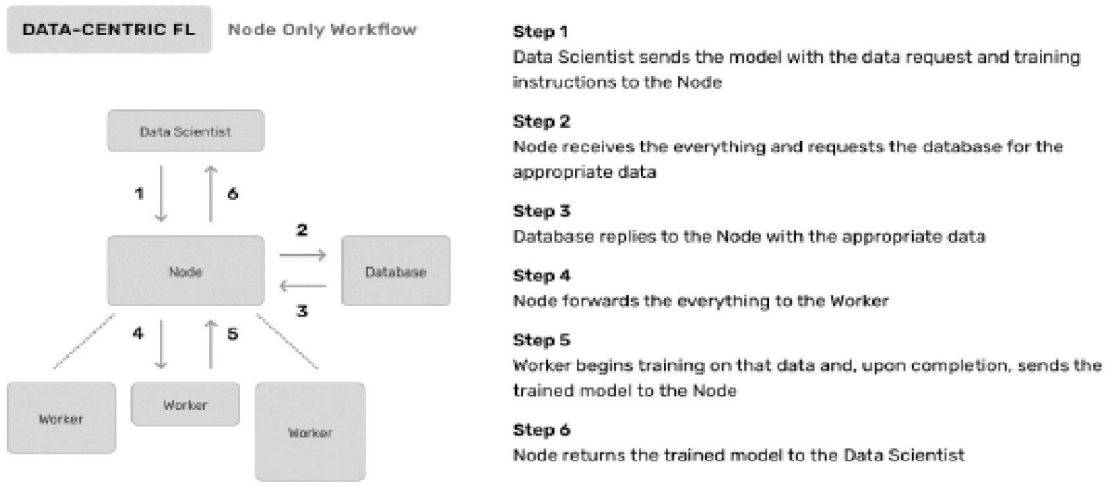

Figure 3: Data-centric FL architecture

6. Emerging 'DataNet' technology stack

Federated learning is a core technology in an emerging data management infrastructure that might be described as *doing for data what the Internet did for communications* (cf. DataNet). Other contributing technologies include:

- **Digital Object Identifiers (DOI)** – A DOI is a data identifier or handle, potentially persistent, used to identify objects uniquely, standardized by an international body (Wikipedia, 2020c).
- **Data standards** – The rules for specifying data. This includes: (a) *data models* – standard for organizing data and standardizes how they relate to one another (Wikipedia, 2020b); (b) *markup* - formats and tagging/typing are required in order to share, exchange, and understand data, such as XML, HTML (Wikipedia, 2020d); and (c) *data exchange formats* – a common data translation format, such as Fast Healthcare Interoperability Resources (FHIR) (Wikipedia, 2020a).
- **Computable legal rules** – A legal contract or regulation encoded in a computer-understandable notation (associated with a human-readable specification) executed by a computer (Surden, 2019).
- **Blockchain technologies** – Including distributed databases that secures, validates and processes transactional data; and smart contracts, a self-executing contract with the terms of the agreement between buyer and seller directly written into lines of code (Treleaven *et al.*, 2017).

Figure 4 illustrates this emerging technology stack and the role of federated learning.

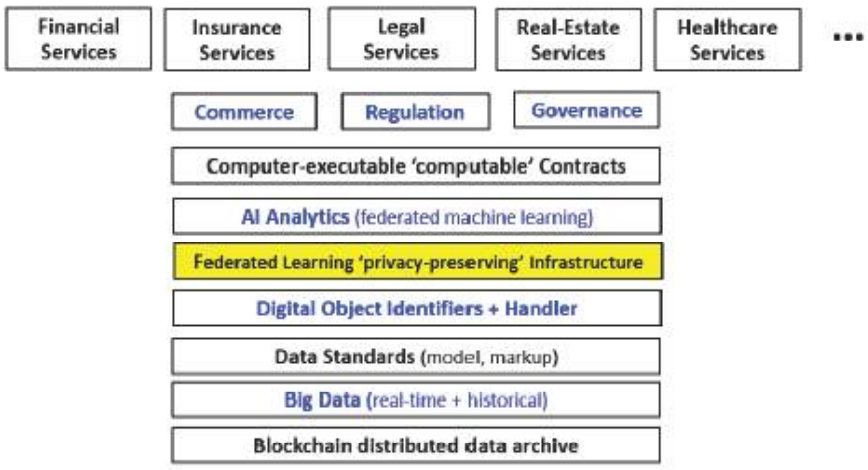

Figure 4: 'DataNet' Technology Stack

7. Federated learning applications

Data collaboration brings a huge potential across industries and different applications, e.g., mobile phones, healthcare, retail, finance and insurance. Below, we both look into current research and possible use cases of federated learning framework.

7.1. Google Gboard

As discussed, federated learning was pioneered in a Google Gboard to improve query suggestion model. Federated learning model processes history on-device to suggest improvements to the next iteration of a global Gboard's query suggestion model.

Use case: Federated learning for mobile phones to collaboratively learn a shared prediction model while keeping all the training data on device. Training the model involves following steps (see Figure 5):

1. The device downloads the current model.
2. The model is improved by learning from data on local device (A).
3. Model updates are summarized and sent to the cloud (using encrypted communication).
4. User's updates are averaged (B) and the shared model is improved (C).

All the training data remains on your device, and no individual updates are stored in the cloud.

Table 1 presents current studies related to federated learning for mobile phones applications.

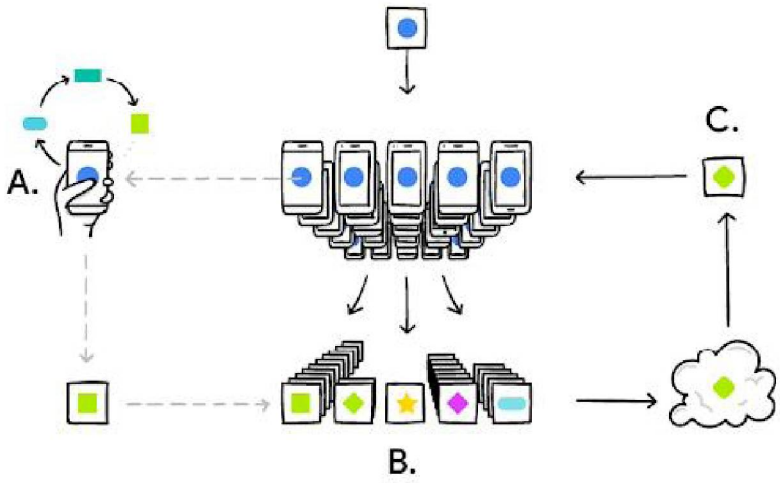

Figure 5: Mobile device model training

Source: GoogleAI (2017)

Table 1: Federated learning in healthcare

Study	Reference
Patient similarity	(Lee *et al.*, 2018)
Future hospitalizations	(Brisimi *et al.*, 2018)
Mortality	(Huang and Liu, 2019)
ADR (adverse drug reactions)	(Choudhury *et al.*, 2019)
Brain Tumor Segmentation	(Li *et al.*, 2019)

7.2. Healthcare

Through federated learning, medical institutions can collaborate with privacy guarantee. Table 1 summarizes the current federated learning studies in healthcare applications.

There is also a huge potential of cross-sector collaboration: Healthcare institutions can collaborate with:

- **Wearable technology providers** to design smart devices with behavioral change programs integrated.
- **Internet of Things** for improved health monitoring and remote treatment.
- **Research and Development** centers to develop more effective treatment, drugs, therapies.

Use Case: Wearable technology providers in collaboration with medical and research centers providing behavioral change programs that help patients with heart disease to stay active and monitor their health. The federated system architecture involve following steps (Figure 6):

1. Global model is sent to wearable devices (A).
2. The model is learning from the data on the device (e.g., checking blood pressure and physical activity). (B).
3. Model updates are averaged and sent to trusted aggregator, then analyzes by medical and research collaborator (C).
4. Global recommendation system is improved and shared with users (D).

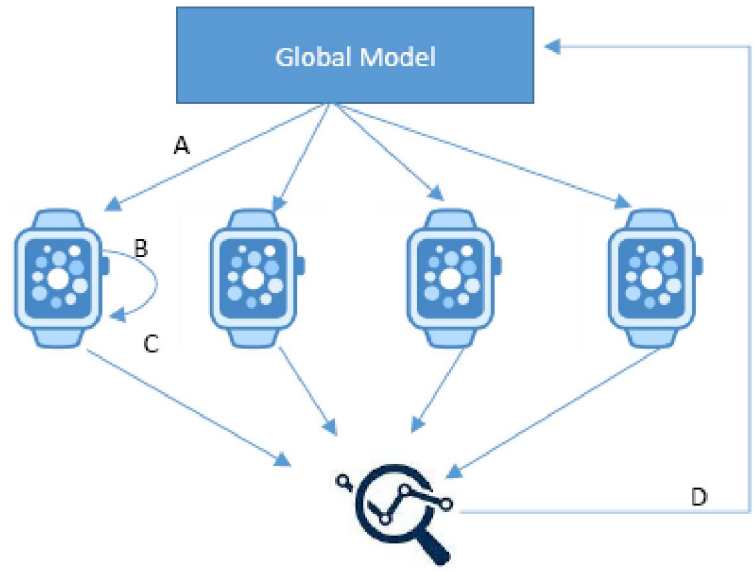

Figure 6: Wearable device training model

Table 2: Federated learning in retail

Study	Reference
Consumptions behaviors	(Zhao *et al.*, 2019)
Federated recommendation system	(Yang *et al.*, 2019)

7.3 Retail

Federated learning can be used to provide personalized services to customers. As an example, machine learning models can be used to predict customer's requirements and consumptions behaviors in federated learning environment (Zhao, *et al.*, 2019). In Table 2 we summarize current use cases of federated learning in retail.

Use Cases: Retail companies might use federated learning for:

Table 3: Federated learning in finance	
Study	**Reference**
Credit card fraud detection	(Yang *et al.*, 2019)
Scoring systems	(OpenMined, 2020)
Credit rating	Tencent's WeBank

1. Internal monetization – To control data access within the organization (within different divisions, e.g., food, banks, insurance, clothing). Internal monetization might result in improvement in product recommendation and sales services.
2. Collaboration with the supply chain (logistics companies, food producers) without compromising the data.
3. Loyalty card data monetization – Selling anonymized data that reveal purchasing power without compromising data.
4. Collaboration with external organizations – Making data available for research.
5. Creating product recommendation systems using: purchasing power data (credit and loyalty cards), personal preferences data (social media), and information related to a product (e-shops, satisfaction surveys).

7.4. Finance and insurance

Insurance companies are realizing the need to utilize all structured and unstructured data, both internal (that they get through customer services, sales, underwriting, renewal and claims processes) and external (connecting their products with IoT, telematics, wearable devices). Federated learning might be a next step in data revolution in insurance. Collaboration among insurers might bring mutual benefit in detecting fraud or more efficient claims handling. Insurers learning from transactions data owned by credit-card providers might be able to improve client targeting and develop more personalized insurance products. Also, collaboration with healthcare sector may lead better coverage of uninsured populations.

In banking, federated learning can be applied to detect multiparty borrowing, build credit-scoring, credit card fraud detection (Yang *et al.*, 2019) and AML systems without exposing the personal information.

Use Case: Improved risk pricing/credit scoring with external data. A federated learning system could work as follows:

1. Different data sources are identified (credit card, loyalty cards, healthcare, wearable devices).
2. Encryption mechanism of federated learning is used together with the data model (Data Model, 2020) that explicitly determines the structure of data.
3. Data owners retain control over customer data.
4. Insurance company/bank trains a single pricing /credit scoring model in federated learning environment.
5. A new contract is priced or a credit score is provided upon request in a federated environment.

8. Conclusion

As discussed, federated learning pioneered by Google is the emerging privacy-preserving data technology and also a new class of distributed machine learning models. As data privacy is becoming increasingly important because of regulation (e.g., EU GDPR) there is a huge potential of federated learning applications across industries. Currently we see most applications in mobile phones and healthcare, but we expect federated learning use cases and studies to spread across sectors and prove benefits of institutions collaboration. An example might be collaboration of the University of Pennsylvania and 19 other institutions worldwide to utilize federated learning in medical imaging use case. That collaboration will potentially enable to generate generalizable, state-of-the-art healthcare models with increasing protection of patient's sensitive data.

References

Barreno, M. *et al*. (2006). Can machine learning be secure?. *Proceedings of the 2006 ACM Symposium on Information, computer and communications security - ASIACCS '06.*

Brisimi, T.S. *et al*. (2018). Federated learning of predictive models from federated electronic health records. *International Journal of Medical Informatics*. 112, 59-67.

Choudhury, O. *et al*. (2019). Predicting adverse drug reactions on distributed health data using federated learning. *AMIA Annu Symp Proc.*, 313-322.

Cohan, P. (2018). *Forbes*. [Online] Available at: https://www.forbes.com/sites/petercohan/2018/07/22/mastercard-amex-and-envestnet-profit-from-400m-business-of-selling-transaction-data/#5e576d4f7722 [Accessed 22 07 2018].

Cramer, R., Damgård, I. and Nielsen, J. (2000). Multiparty computation from threshold homomorphic encryption. *Lecture Notes in Computer Science*, 11(7).

Damgård, I., Pastro, V., Smart, N. and Zakarias, S. (2012). *Multiparty computation from somewhat homomorphic encryption*. s.l., s.n., 643-662.

Data Model, 2. (2020). *Wikipedia*. [Online] Available at: https://en.wikipedia.org/wiki/Data_model#:~:text=A%20data%20model%20(or%20datamodel,properties%20of%20real%2Dworld%20entities. [Accessed 28 08 2020].

Derwisch, S. (2019). *Data Monetization - Use Cases, Implementation and Added Value*, s.l.: BARC.

Facebook_data_scandal (2020). *Wikipedia*. [Online] Available at: https://en.wikipedia.org/wiki/Facebook–Cambridge_Analytica_data_scandal [Accessed 05 08 2020].

GDPR (2018). *GDPR*. [Online] Available at: https://gdpr-info.eu/ [Accessed 05 08 2020].

Gilad-Bachrach, R. *et al*. (2016). *Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy*. s.l., s.n., 201-210.

GoogleAI (2017). *Google AI Blog*. [Online] Available at: https://ai.googleblog.com/2017/04/federated-learning-collaborative.html [Accessed 05 08 2020].

Hardy, C., Le Merrer, E. and Sericola, B. (2018). Gossiping GANs. *Proceedings of the Second Workshop on Distributed Infrastructures for Deep Learning - DIDL '18*.

Huang, L. and Liu, D. (2019). Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records.. *arXiv:1903.09296*.

Keller, M., Orsini, E. and Scholl, P. (2016). MASCOT. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*.

Konečný, J., McMahan, B. H., Ramage, D. and Richtárik, P. (2016). Federated Optimization: Distributed Machine Learning for On-Device Intelligence. *https://arxiv.org/abs/1610.02527*.

Konečný, J. *et al*. (2017). Federated Learning: Strategies for Improving Communication Efficiency. *https://arxiv.org/abs/1610.05492*.

Koty, A. C. (2019). *China Briefing*. [Online] Available at: https://www.china-briefing.com/news/chinas-corporate-social-credit-system-how-it-works/ [Accessed 03 08 2020].

Lee, J. *et al*. (2018). Privacy-preserving patient similarity learning in a federated environment: development and analysis.. *JMIR Medical Informatics*, 6(2).

Li, W. *et al*. (2019). Privacy-preserving federated brain tumour segmentation. In: *Machine Learning in Medical Imaging. MLMI 2019. Lecture Notes in Computer Science*. s.l.:Springer, Cham, pp. 133-141.

McMahan, B.H. *et al*. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. *https://arxiv.org/abs/1602.05629*.

MELLODDY (2020). *MELLODDY project*. [Online] Available at: https://www.melloddy.eu/ [Accessed 10 09 2020].

OpenMined (2020). *OpenMined Blog*. [Online] Available at: https://blog.openmined.org/federated-credit-scoring/ [Accessed 03 08 2020].

Papernot, N. *et al*. (2020). *Scalable Private Learning with PATE*. s.l.:s.n.

PyGrid, O. (2020). *PyGrid OpenMined*. [Online] Available at: https://github.com/OpenMined/PyGrid [Accessed 08 2020].

Rivest, R.L., Adleman, L., Dertouzos, M.L. and others (1978). On data banks and privacy homomorphisms. *Foundations of Secure Computation*. 4, 169-180.

Shamir, A. (1979). How to share a secret. *Communications of the ACM*. 22, 612-613.

Śmietanka, M., Koshiyama, A. and Treleaven, P. (2020). Algorithms in Future Insurance Markets. *https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3641518*.

Surden, H., (2019). Artificial intelligence and law: An overview. *Georgia State University Law Review*, 35, 19-22.

Treleaven, P., Gendal Brown, R. and Yang, D. (2017). Blockchain technology in finance. *Computer,* 50(9), 14-17.

Turea, M. (2019). *Healthcare innovation.* [Online] Available at: https://healthcareweekly.com/how-the-big-4-tech-companies-are-leading-healthcare-innovation/ [Accessed 11 08 2020].

Wikipedia (2020a). *Data exchange.* [Online] Available at: https://en.wikipedia.org/wiki/Data_exchange [Accessed 05 08 2020].

Wikipedia (2020b). *Data model.* [Online] Available at: https://en.wikipedia.org/wiki/Data_model [Accessed 05 08 2020].

Wikipedia (2020c). *Digital object identifier.* [Online] Available at: https://en.wikipedia.org/wiki/Digital_object_identifier [Accessed 05 08 2020].

Wikipedia (2020d). *Markup language.* [Online] Available at: https://en.wikipedia.org/wiki/Markup_language [Accessed 05 08 2020].

Yang, Q. *et al*. (2019a). Federated learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning,* 13, 1-207.

Yang, Q. *et al*. (2019b). *Federated Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning.* s.l.:Morgan & Claypool Publishers.

Yang, Q., Liu, Y. and Chen, T. (2019c). Federated machine learning: Concept and applications. *http://arxiv.org/abs/1902.04885*.

Yang, W. *et al*. (2019d). FFD: A federated learning based method for credit card fraud detection. *Notes in Computer Science, Springer.* 11514, 18-32.

Yao, A. C.C. (1986). *How to generate and exchange secrets - IEEE Conference Publication.* s.l.:s.n.

Yu, S. (2016). Big privacy: Challenges and opportunities of privacy study in the age of big data. *IEEEAccess,* 4(2016), 2751-2763.

Zhao, Y. *et al*. (2019). Mobile edge computing, blockchain and reputation-based crowdsourcing iot federated learning: A secure, decentralized and privacy-preserving system.. *arXiv:1906.10893.*

Zou, Y. *et al*. (2019). Mobile Device Training Strategies in Federated Learning: An Evolutionary Game Approach. *International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData),*874-879.

Cite this article as: Malgorzata Smietanka, Hirsh Pithadia and Philip Treleaven (2021). Federated learning for privacy-preserving data access. *International Journal of Data Science and Big Data Analytics.* 1(2), 1-13. doi: 10.51483/IJDSBDA.1.2.2021.1-13.