Federated Learning

The pioneering distributed machine learning and privacy-preserving data technology

Philip Treleaven, Malgorzata Smietanka, Hirsh Pithadia
University College London

Abstract
Federated learning (pioneered by Google) is a new class of machine learning models trained on distributed datasets, and equally important a key privacy-preserving data technology.

With huge amounts of data for analysis, organisations are faced with three major challenges: a) *data* comprises distributed and isolated data sets; b) *analytics* requires models to be trained across these independent data sets; and c) *data sovereignty/privacy* legislation is making collecting, sharing and analysing data increasingly difficult.

This paper reviews federated learning both in terms of a) a federated data infrastructure for privacy-preserving data access; and b) federated machine learning applied to distributed data sets. Given the pivotal role of federated learning, the contribution of this paper is to place it in perspective to the other data science technologies. It includes discussions of the privacy challenges facing data analytics, relationship to the major data infrastructure technologies, and the emerging machine learning algorithms impacting federated learning.

1. Data Ecosystems

The old clichés 'data is the new oil' or 'data is the new gold', acknowledges firstly that data increasingly facilitates business; and secondly recognition of the monetary value of data. The two core aspects are:

- **Federated data infrastructure** – privacy-preserving data infrastructure; a framework for collaboration, allowing secure communication with collaborating parties, such that 'raw' data does not leave the owner.
- **Federated machine learning** – decentralized training of a machine learning model which enables collaborative learning while keeping data sources in their original location. For example, Google's mobile phone users benefit from obtaining a well-trained model without sending their personal data to the Cloud.

Broadly, federated learning ecosystems address:

- **On-device** (i.e. cross-device) infrastructures for mobile devices, IoT, Edge and other connected devices. For example, Google and Apple used it for keyboard (next-word prediction models). The data ecosystem is characterised by a very large number of devices (tens or hundreds of millions), with intermittency and low bandwidth connections.
- **Inter-organisation** (i.e. cross-silo) infrastructures allowing collaborating organisations to contribute to the training with their local datasets. An example is the European MELLODDY project, where ten pharmaceutical companies collaborate in training a machine learning for drug discovery based on private, highly sensitive datasets. This data ecosystem is characterized by a smaller number of participants with good bandwidth and connectivity.

Federated learning is a pivotal technology for future data ecosystems (see Figure 1), impacting data management, working with other data science technologies, and increasingly data sovereignty issues.

Data management is a major challenge as organisations: a) collect increasing amounts of heterogeneous data (e.g. business, economic, social media and alternative); b) data is stored in isolated data sets; c) work with partner organisations on collaborative analytics; d) need to secure the data; and e) monetise their data for their business, possibly as a new revenue stream.

Data technologies participating in information management with federated learning are common digital identifiers, data standards, data analytics, and data record technologies:

- **Digital identifiers -** data referencing using unique identifiers for referencing data objects. For example digital object identifiers (DOIs) that are unique, persistent and resolvable.
- **Data standards** – this covers standards for representing data; including common data models for industry sectors (e.g. Finance FIB-DM); and markup languages (e.g. eXtensible Markup Language XML).
- **Data analytics** - new forms of 'statistics', such as machine learning, computational statistics, and complex systems (e.g. deep neural networks, Monte Carlo simulation).
- **Data records** – this covers the storing, sharing and synchronization of data transactions. Technologies relevant to federated learning include: *distributed databases* - data is stored across different physical locations; *distributed ledgers* - digital systems for synchronising transactions; and *blockchain* – a type of distributed ledger where transactions are validated by multiple independent computers.

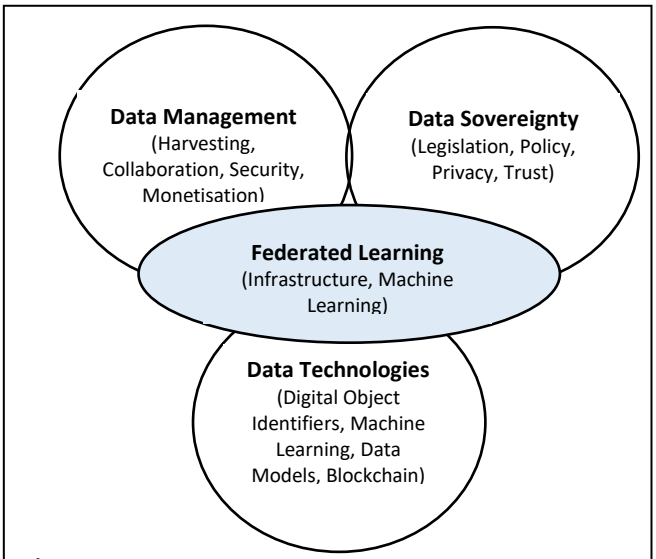

Figure 1: Data Ecosystems

Data sovereignty/privacy is becoming a major political and social issue for governments, institutions and citizens (e.g. Google, Facebook, ByteDance TikTok, Tencent WeChat). Data sovereignty is the idea that data and its usage are subject to the laws and governance structures within the nation it is collected, as well as defining ownership and governmental rights of access. Influential programmes include:

- **EU General Data Protection Regulation (GDPR)** – a legal framework that sets guidelines for the collection and processing of personal information
- **Singapore Personal Data Protection Act (PDPA)** - governs the collection, use and disclosure of individuals' personal data.
- **California Consumer Privacy Act (CCPA)** – covers the handling of personal information of all California Residents.
- **China Cybersecurity Law** - enacted to increase data protection, data localization, and cybersecurity in the interest of national security.

Google Federated Learning

Google pioneered the concept of federated learning and provide an excellent illustration of the potential of the technology. Google's federated Learning infrastructure (see Figure 2) enables mobile phones to collaboratively learn a shared prediction model while keeping all the training data on device, decoupling the ability to do machine learning from the need to store the data in the Cloud. In contrast, the traditional machine learning model is to gather raw data together (e.g. Cloud) for training. This is characterised as 'taking the data to the algorithm'. In contrast, federated learning is 'taking the algorithm to the data'.

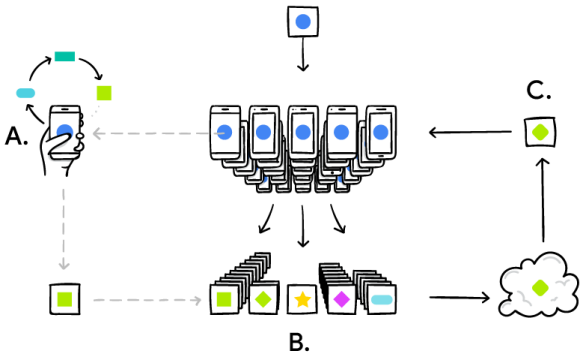

Figure 2: Google Federated Learning

As described in Further Reading, it operates as follows: A) your phone personalises the model locally depending on your usage; B) many users' updates are aggregated; C) the aggregated updates form a consensus change to the shared model; and D) the shared models are updated.

2. Data Politics

Federated learning, as discussed, is a response to growing political, commercial and social challenges concerning data. We have limited discussion to key bullet points:

- **Data harvesting** – the process of extracting and analysing (personal) data on users from online interactions. Leading companies include Google, Facebook, Amazon, Tencent, and ByteDance TikTok. They collect comprehensive personal and interaction data on users and their network of contacts; then use sophisticated machine learning algorithms for 'deep' behavioural and predictive analytics.
- **Data ownership** – who own the data and the right to exploit it. It covers issues of *ownership*, *stewardship* and *custodianship*; responsibility for data content, context, safe custody and usage.
- **Data privacy** –ensuring that the data shared by clients is only used for its intended purpose; and the right of individuals to have control over how their personal information is collected and used.
- **Data collaboration** - spans: a) *internal* – companies in a group or departments in government; b) *consortia* – groups of companies partnering in analytics and business; and c) *international* – organisations such as financial regulators (e.g. AML).
- **Data security** – means protecting digital data from destructive forces and from the unwanted actions of unauthorized users, such as a cyberattack or a data breach, that may compromise the integrity, confidentiality or privacy of the data.
- **Data legislation** – controls how personal or customer information is used by organisations or government bodies. A prominent example being the EU General Data Protection Regulation (GDPR) covering data protection and privacy, plus transfer of personal data outside the EU.
- **Data sovereignty** - the idea that data are subject to the laws and governance structures within the nation it is collected; central to competition, taxation, security and economic supremacy.

3. Data Infrastructures

To provide context for federated learning, this part lists the associated data science technologies driving the emerging data ecosystems. Unfortunately space precludes detailed descriptions. We divide these data science technologies 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, common data models, markup languages).
- **Algorithm technologies** – new forms of 'statistics', such as machine learning, computational statistics, and complex systems (e.g. neural networks, Monte Carlo simulation).
- **Analytics technologies** – covering the application of the data technologies (e.g. natural language, sentiment analysis and behavioural analysis).
- **Infrastructure technologies** – providing the infrastructure for information management and automation (e.g. distributed ledger/blockchain, computer-executable 'computable' contracts).

3.1 Data Technologies

Important data technologies include:

- **Big Data** – very large datasets of historic and real-time financial, economic, social media and alternative data; so complex that traditional data processing application software is inadequate to deal with them.
- **Data standards** - the rules for specifying data. This includes: a) *data models* – standards for organizing data and how data relates to one another; b) *markup languages* - formats and tagging/typing are required in order to share, exchange, and understand data, include: XML and JSON.

3.2 Algorithm Technologies

Core new forms of 'statistics' include:

- **Computational Statistics** - a large class of modern statistical methods that are computationally intensive (e.g. Monte Carlo methods).
- **Artificial intelligence** – AI, machine learning and other systems able to perform tasks normally requiring human intelligence, such as self-programming machine learning (ML) algorithms (e.g. artificial neural networks, federated machine learning).
- **Complex Systems** - system featuring a large number of interacting components whose aggregate activity is nonlinear (e.g. agent-based systems).

3.3 Analytic Technologies

Core analytics technologies include:

- **Natural Language Processing (NLP)** – the analysis and synthesis of natural language and speech.
- **Sentiment Analysis** – using NLP, statistics, or machine learning methods to extract, identify, or characterize the sentiment content of text or speech.
- **Behavioural/Predictive Analytics** – providing insight into the actions of people and predict future outcomes and trends.

3.4 Infrastructure Technologies

Core infrastructure technologies include:

- **Digital Object Identifiers (DOI)** – a DOI is an identifier or handle, potentially persistent, used to identify objects uniquely, standardized by an international body.
- **Computable Legal Rules** – a legal contracts encoded in a computer-understandable notation and executable by a computer can automate commerce.
- **Federated Learning** –- allows machine learning algorithms to be trained across distributed and isolated data

sets, with the potential for privacy preservation.

- **Internet of Things (IoT)** - the inter-networking of 'smart' physical devices, vehicles, buildings, etc. that enable these objects to collect and exchange data.
- **Data Security** – these cover cryptographic techniques (e.g. Homomorphic Encryption, Secure Multiparty Computation and Differential Privacy, Public-Key cryptography) used to preserve confidentiality, privacy and integrity.
- **Distributed Ledger Technologies (DLT)** – distributed databases that secures, validates and processes transactional data (e.g. blockchain).
- **Edge Computing** –a decentralised network where data is processed by the device itself or by a local computer or server, rather than being transmitted to a data centre.

3.5 Data Technology Stack

To draw these data technologies together Figure 3 illustrates a 'technology stack'. Arguably a global 'DataNet' infrastructure is emerging that might be described as *doing for data what the Internet did for communications*. At the base is distributed ledger and blockchain technology. Next data and cryptographic security. Then Big data conforming to common data standards often for a specific industry sector. Then we have digital object identifiers providing unique, persistent and resolvable addresses for the data objects. Lastly, we have emerging computer-executable legal contracts and regulations that are important for automation.

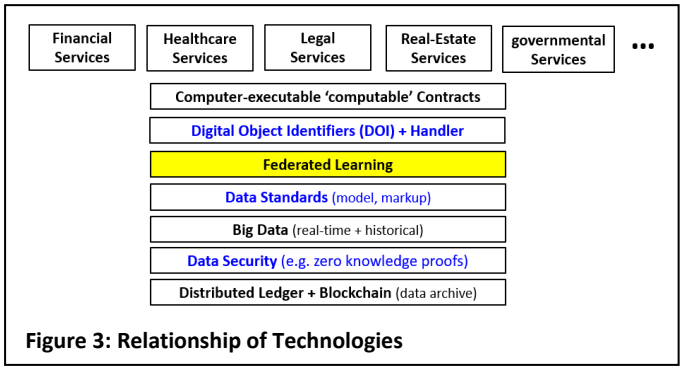

Figure 3: Relationship of Technologies

Next we review AI machine learning, to provide context for federated machine learning.

4. Machine Learning

As discussed, new forms of 'statistics' cover three broad algorithm domains: Computational Statistics (e.g. Monte Carlo methods), Artificial Intelligence (e.g. artificial neural networks), and Complex Systems (e.g. agent-based systems). AI Algorithms divides into: a) knowledge or rule-based systems, b) evolutionary algorithms, and c) machine learning.

(Koshiyama et al, 2020) in Further Reading provides detailed descriptions.

4.1 Traditional Machine Learning

Machine learning algorithms are broadly a combination of the classical *trio* of Supervised, Unsupervised and Reinforcement Learning, with the *disruptors*: Deep Learning, Adversarial Learning, Transfer/Meta Learning. This interaction constantly yields new models (e.g., Long Short-Term Memory, Generative Adversarial Networks, Generative Pre-trained Transformers). Brief descriptions of these machine learning models are presented since they underpin the new generation of federated machine learning models discussed later.

Classic trio:

- **Supervised learning** - given a set of inputs variables/predictors $\mathbf{x}$ and outputs/dependent variables/targets $\mathbf{y}$, the goal is to learn a function $f(\mathbf{x})$ that approximates $\mathbf{y}$.
- **Unsupervised learning** - given several objects/samples/transactions $\mathbf{x}_1, \ldots, \mathbf{x}_n$, the goal is to learn a hidden map $h(\mathbf{x})$ that can uncover a hidden structure in the data. This hidden map can be used to 'compress' $\mathbf{x}$ (aka dimensionality reduction) or to assign

to every $\mathbf{x}_i$ a group c_k (aka clustering or topic modelling).

- **Reinforcement learning** - given an environment formed by several states $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n$, an agent, and a reward function, the goal is to learn a policy π that will guide an agent actions $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_k$ through the state space so as to maximize occasional rewards.

The disruptors:

- **Deep Learning** - deep learning algorithms model high-level abstractions in data by using multiple processing layers, with complex structures or otherwise, composed of multiple non-linear transformations. An example used in federated learning is deep neural networks such as Convolutional Neural Networks (CNNs)

- **Adversarial Learning** - adversarial machine learning is a technique which attempts to 'fool' models through malicious input.

- **Transfer/Meta Learning** – these two learning paradigms are tightly connected, as their main goal is to encapsulate knowledge learned across many tasks and transfer it to new, unseen ones. In transfer learning, knowledge is transfer from a trained model to a new model by encouraging the new model to have similar parameters. In meta learning the learning method is abstracted and shared across tasks, and meta-learned explicitly with transfer in mind, such that the learning method generalize to an unseen task. Both are influential in federated learning.

4.2 Federated Machine Learning

The traditional machine learning strategy is to gather raw data together (e.g. in a central repository hosted in the cloud) for training. This is characterised as 'taking the data to the algorithm'. In contrast, federated learning is 'taking the algorithm to the data'. The typical federated learning paradigm involves two stages: a) clients train models with their local datasets independently, and b) the data centre gathers the locally trained models and aggregates them to obtain a shared global model.

Federated machine learning by definition aims to build a joint model based on data located at multiple sites. There are two processes: i) model training and ii) model inference. In the process of *model training*, 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 *model inference* stage, the model is applied to a new data instance (e.g. 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.).

5. Federated Learning Categories

Federated learning is generating a wave of developments and publications. As discussed, System divide into: a) *On-device* (i.e. cross-device) - infrastructures for mobile devices, IoT, Edge and other connected devices; b) *Inter-organisation* (i.e. cross-silo) - infrastructures allowing collaborating organisations to contribute to the training with their local datasets.

To provide a perspective we next looks at categorising federated learning (see in Figure 4):

- **Communication** – communication or control of federated analysis: a) *centralised learning* – a central server orchestrates the different steps of the algorithms and coordinate participating nodes during the learning process; b) *decentralised Learning* - the participating nodes coordinate themselves to obtain the global model.

dataset. These are discussed further in

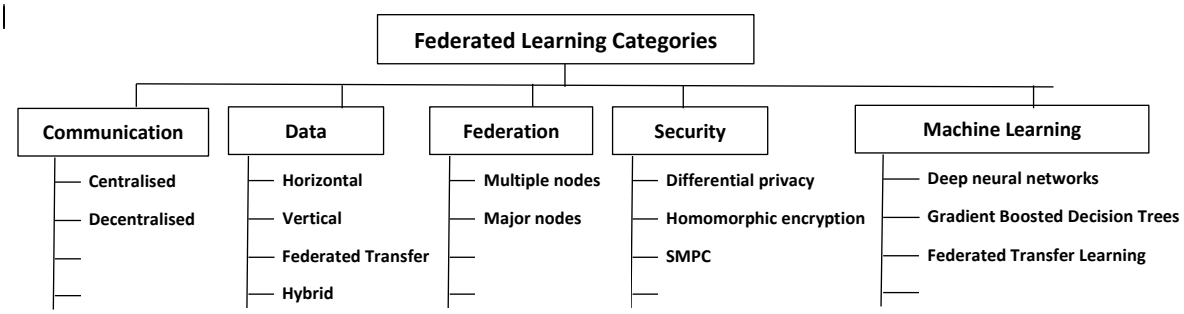

Figure 4: Federated Learning Taxonomy

- **Data** – by data partition: a) *horizontal federated learning* – homogeneous data sets have the same feature space but distinct sample spaces; b) *vertical federated learning* – heterogeneous data sets with different feature spaces but the same sample space; and c) *Federated transfer learning* - here data sets differ not only in samples but also in feature space.
- **Federation** – the scale of federation of nodes: a) *multiple nodes* – a large number of *On-device* nodes, such as smart phones, each with a relatively small amount of data and processing power; b) *major nodes* - a small number of major *inter-organisation* nodes, such as data centres, each with a large amount of data and processing power.
- **Security** – the data privacy preserving techniques employed. Popular for federated learning are: a) *secure multi-party computation* SMPC - a subfield of cryptography with the goal of creating methods for parties to jointly compute a function over their inputs while keeping those inputs private; b) *homomorphic encryption* - the conversion of data into an encrypted form that can be analysed and worked with as if it were still in its original form; and c) *differential privacy* - a system for sharing information describing the group patterns within a data set while withholding identifiable 'raw data' about individuals in the section 6.
- **Machine learning** – the emerging federated machine learning models are variants of traditional models. Examples include: a) *deep neural networks* - networks multiple layers between the input and output layers; and b) *gradient boosted decision trees* - involves three elements: a loss function optimisation, a weak learner for predictions, and an additive model for minimizing the loss function. These are discussed in section 7.

5.1 Communication or Control architecture

As discussed, federated learning communication architectures either use a trusted centralised node (i.e. server) to orchestrate learning; or the decentralised nodes coordinate themselves to obtain a global model. Examples include FedAvg and SimFL.

Centralised Federated Averaging (FedAvg)
With centralised learning the trusted node aggregates the information from the other nodes and sends back training results (e.g. gradients or model parameters) to the participating nodes. Communication or control between the nodes can be synchronous or asynchronous.

Decentralised GBDT FL (SimFL)
With decentralised learning communications are performed amongst the nodes and every node is able to update the global model parameters directly.

5.2 Data partition

Federated learning systems are frequently classified by their data partition into how the data sets are distributed across the nodes; namely the sample and feature spaces:

- *Horizontal federated learning* – homogeneous data sets share the same feature space but have different in samples;
- *Vertical federated learning* – with feature-based learning multiple heterogeneous data sets share the same data space but differ in feature space;
- *Federated transfer learning* - here data sets differ not only in samples but also in feature space with only a small portion of the feature space from both parties overlaps; and

Different data characteristics causing federated learning engineering challenges are discussed later in the paper.

5.3 Federation of nodes

Classification by federation ranges from a) *multiple nodes* – a large number nodes each with a relatively small amount of data (e.g. smart phones, IoT devices); to b) *major nodes* - a small number of powerful nodes, each with a large amount of data (e.g. data centres).

Federations of multiple or major nodes can span a single product line or company, or a small number of major organisations collaborating.

6. Federated Learning Security mechanisms

Given the importance of security mechanism for privacy-preserving data access, this section is a review of candidate and cryptographic schemes. Cryptography is the basic building block of data security. For federated learning, popular cryptographic schemes include Secure Multi-Party Computation (SMPC), Homomorphic Encryption (HE) and Differential Privacy (DP); often in combination and as ensembles. These schemes subsume and make use of other popular security mechanisms such as Cryptographic Hashing Functions and Elliptic Curve Cryptography (ECC). Zero Knowledge Proofs may also be used in future.

Secure Multi-Party Computation

Secure Multi-Party Computation (SMPC), also known as Secure Function Evaluation, involves jointly computing a function from the private input by each party without revealing the value of these private inputs to other parties. In federated learning terms this 'function' could be a model's loss function during training or the model itself (during inference). An SMPC-based scheme typically of two parts: a) *online phase* - involves the training of the machine learning model (e.g. using the triples generated in the offline phase); and b) *offline phase* - involves the bulk of the cryptographic operations such as the generation of triples. A common SMPC scheme currently used for privacy-preserving machine learning is the Secret Sharing based SPDZ scheme.

Homomorphic Encryption

Homomorphic Encryption (HE), first proposed by Rivest-Shamir-Adleman involves the direct computation over a 'ciphertext', without decrypting the ciphertext. There are a number of HE schemes used in conjunction with federated learning, often bucketed into three: a) *partially HE schemes* - support the evaluation of circuits consisting of only one type of gate; b) *somewhat HE schemes* - can evaluate two types of gates, but only for a subset of circuits; and c) *fully HE schemes* – allows the evaluation of arbitrary circuits of unbounded depth, and is the strongest notion of homomorphic encryption.

7. Federated Machine Learning Models

As discussed, considerable research is underway to create federated learning variants of traditional machine learning models. This includes deep neural networks, gradient boosted decision trees, logistic regression and support vector machines. Examples are: a) *federated averaging* - a central server is responsible for coordinating the training of models located in different devices; b) *federated transfer learning* – uses

transfer learning to improve model performance when we have neither much overlap on features nor on instances.

Notable in federated learning is the pioneering work of Google, OpenMined and WeBank, who have published open source frameworks or libraries (i.e. Tensor Flow Federated, PyTorch and FATE, respectively) and made datasets available for training.

Deep Neural Networks
A popular basis for federated learning is deep neural networks; networks with multiple layers between the input and output layers. One of the standard aggregation methods is federated averaging (FedAvg) where parameters of local models are averaged element-wise with weights proportional to sizes of the client datasets. Deep neural networks are the basis of the open source platforms described below.

Gradient Boosted Decision Trees
Another popular approach is gradient boosting decision trees (GBDT). GBDT is a powerful techniques for building predictive models based on 'boosting', a family of algorithms that can upgrade weak learners to strong learners. Boosting methods are based on the idea that for a complex task, it is better to synthesize the judgments of multiple experts appropriately than to judge by any one of them alone. As discussed, GBDT involves three elements: a) a loss function to be optimized; b) a weak learner to make predictions; and c) an additive model to add weak learners to minimize the loss function.

7.1 Open Source Tools

An increasing number of powerful open sourced federated learning frameworks or platforms are now available. Prominent examples are Google TFF, OpenMined PySyft and WeBank FATE, as well as FedML Federated-XGBoost and Baidu PaddleFL.

Briefly:

- **Google TensorFlow Federated (TFF) -** Google TFF focuses on horizontal federated learning with a large population of client devices with heterogeneous computing capabilities.
- **OpenMined PySyft** – PySyft uses the PyTorch machine learning platform to implement a federated learning model. PySyft is a Python library for secure and private deep learning. PySyft supports PyTorch, Tensorflow, and Keras with varying capabilities for remote execution, federated learning, differential privacy, homomorphic encryption, and secure multi-party computation.
- **WeBank FATE** – WeBank's Federated AI Technology Enabler (FATE) supports horizontal FL, vertical FL and federated transfer learning with a focus on secure protocols based on homomorphic encryption and multi-party computation (MPC).

In addition, open source federated datasets are emerging to support distributed training. This addresses the lack of high-quality training data generated from real-world applications, such as WeBank FedVision https://arxiv.org/abs/1910.11089

8. Federated Learning Engineering Issues

The engineering challenges associated federated learning relate to: a) data characteristics, b) model characteristics, c) performance efficiency, d) disparate systems, and e) availability of nodes.

Data characteristics
Data challenges are categorised by the 'unbalancedness' (i.e. **non- I**ndependent and **I**dentically **D**istributed) of local data samples: a) *covariate shift* - local samples have different statistical distributions; b) *prior probability shift* - local nodes may store labels that have different statistical distributions; c) *concept shift* – dividing into: i) local nodes share the same labels but have different features, and ii)

local nodes the same features but different labels; and d) *unbalancedness* - the data available at the local nodes may vary significantly in size. Other data challenges include data augmentation and feature engineering.

Model characteristics

The next challenge relates to the training data, and choosing the *hyperparameters* (values used to control the learning process) and optimisers. Examples of parameters include the number of layers in a network, the number of nodes, learning rate, the structures of the network etc. Optimisers could include identifying batch sizes.

Performance efficiency

Performance is impacted firstly by node communications requirements and secondly by privacy-preserving cryptographic techniques. For example, vertical federated learning require constant communication between each of the nodes, since each node will be training their part of the model. Slow data communication between nodes can result in inefficient training. Cryptographic techniques add additional computation complexity to an already computationally intensive process- machine learning, thereby impacting training times and tying up resources.

Disparate systems

Next is disparity in the infrastructure components (e.g. nodes, communications, hosting/cloud environment). Most enterprises employ proprietary infrastructure (e.g. servers, cloud environments, firewalls) to protect their sensitive data, rather than agnostic frameworks and homogeneous nodes.

Availability

Lastly, distributed and collaborative training is impacted by availability of communications and nodes. Systems need to be reliable or tolerant to failure, otherwise it may interrupt the entire training process that may render all the work done by the other nodes as void.

9. Federated Learning Case Studies

As a conclusion, we discuss briefly federated learning applications illustrating On-device and Inter-organisations.

9.1 On-device

On-device (i.e. cross-device) covers large number of devices with intermittency and low bandwidth connections, such as mobile devices, connected vehicles, IoT, and Edge connected devices.

Telecoms

Google's pioneering smart phone analytics system works across millions of devices optimising keystroke prediction. TensorFlow Federated (TFF) TFF has two layers: the federated learning API and the federated core API. The federated learning API allows developers to apply federated training and evaluation to existing TensorFlow models. The federated core API is the core foundation for federated learning. It is a system of low-level interfaces for writing federated algorithms in combination with distributed communication operations in strongly-typed functional programming environments.

Autonomous Vehicles

NVIDIA's DRIVE suite of deep neural network tools for self-drive and AV vehicles. It comprises in-vehicle computer (DRIVE AGX) and complete reference architecture (DRIVE Hyperion), as well as data centre-hosted simulation (DRIVE Constellation) and deep neural network training platforms (DGX). DRIVE supports active learning, federated learning and transfer learning. Active learning allows an algorithm to interactively query a user to improve model accuracy. Transfer learning leverages previous training data, for example, knowledge gained while learning to recognize cars could apply when trying to recognize trucks.

9.2 Inter-organisation

Inter-organisation (i.e. cross-silo) comprise a small number of major participants with good bandwidth and connectivity allowing collaborating organisations to contribute to the training with their local datasets.

Healthcare
The European MELLODDY (Machine Learning Ledger Orchestration for Drug Discovery) Consortium of pharmaceutical, technology and academic partners employs federated learning and Blockchain to Enhance AI-driven drug discovery. It uses federated learning on the chemical libraries of 10 pharma companies to support a modelling platform designed to quickly and accurately predict promising compounds for development, all without sacrificing the data privacy of the participating companies.

Financial Services
Tencent's WeBank, China's first all-digital financial institution, uses federated learning for credit-scoring, credit card fraud detection and AML. The credit scoring model is used in conjunction with WeBank's own data, and encrypted invoice data from the Invoice Centre which stays on other invoice centres' servers. WeBank uses the platform for loans, measuring the credit risk of small and micro-enterprises, and having halved the number of defaults. In a further development WeBank and reinsurance giant Swiss Re's Beijing branch have signed a MOU on joint research on the application of federated learning in reinsurance.

Lastly, we only need to consider the accelerating pace of self-drive vehicles (e.g. cars, trucks, drones, ships, planes) and the need for connected coordination to understand the potential impact of federated learning.

10. Further Reading

Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konečný, J., Mazzocchi, S., McMahan, H.B. and Van Overveldt, T., 2019. Towards federated learning at scale: System design. arXiv preprint arXiv:1902.01046.

Koshiyama, A., Firoozye, N. & Treleaven, P., 2020. Algorithms in Future Capital Markets. *SSRN:* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3527511.

McMahan, Brendan, and Daniel Ramage. "Federated learning: Collaborative machine learning without centralized training data." Google Research Blog 3 (2017) https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

References on Federated Learning，联邦学习参考文献，https://zhuanlan.zhihu.com/p/87777798

Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated Machine Learning: Concept and Applications. ACM Trans. Intell. Syst. Technol, https://arxiv.org/pdf/1902.04885.pdf

11. Authors

Philip Treleaven is Director of the UK Centre for Financial Computing (www.financialcomputing.org) and Professor of Computing at UCL.

Malgorzata Smietanka is a PhD Researcher in UCL Computer Science, with a Masters in Mathematics from Warsaw University, and is a qualified Actuary. She is founder of Actuari.co.uk, an insurance analytics start-up.

Hirsh Pithadia is a PhD Researcher in UCL Computer Science, with a Research Masters in Financial Computing. He is a co-founder of RegulAItion.com, a leading RegTech company.