Towards Personalized Federated Learning

Alysa Ziying Tan, Han Yu*, Lizhen Cui*, and Qiang Yang*, *Fellow, IEEE*

***Abstract*—In parallel with the rapid adoption of Artificial Intelligence (AI) empowered by advances in AI research, there have been growing awareness and concerns of data privacy. Recent significant developments in the data regulation landscape have prompted a seismic shift in interest towards privacy-preserving AI. This has contributed to the popularity of Federated Learning (FL), the leading paradigm for the training of machine learning models on data silos in a privacy-preserving manner. In this survey, we explore the domain of Personalized FL (PFL) to address the fundamental challenges of FL on heterogeneous data, a universal characteristic inherent in all real-world datasets. We analyze the key motivations for PFL and present a unique taxonomy of PFL techniques categorized according to the key challenges and personalization strategies in PFL. We highlight their key ideas, challenges and opportunities and envision promising future trajectories of research towards new PFL architectural design, realistic PFL benchmarking, and trustworthy PFL approaches.**

***Index Terms*—federated learning, personalized federated learning, non-IID data, statistical heterogeneity, privacy preservation, edge computing.**

I. Introduction

THE pervasiveness of edge devices in modern society, such as mobile phones and wearable devices, has led to the rapid growth of private data originating from distributed sources. In this digital age, organizations are using big data and artificial intelligence (AI) to optimize their processes and performance. While the wealth of data offers tremendous opportunities for AI applications, most of these data are highly-sensitive in nature and they exist in the form of isolated islands. This is especially relevant in the healthcare industry where medical data are highly-sensitive and they are often collected and reside across different healthcare institutions [1]–[4]. Such circumstances pose huge challenges for AI adoption as data privacy issues are not well addressed by conventional AI approaches. With the recent introduction of data privacy preservation laws such as the General Data Protection Regulation (GDPR) [5], there is an increasing demand for privacy-preserving AI [6] in order to meet regulatory compliance.

In view of these data privacy challenges, Federated Learning (FL) [7], [8] has seen growing popularity in recent years. FL is a learning paradigm that enables collaborative training of machine learning models involving multiple data silos in a privacy-preserving manner. The prevailing FL setting assumes a federation of data owners (a.k.a. clients), which may be as small as individual mobile devices to as large as entire organizations, that collaboratively train a model under the orchestration of a central parameter server (a.k.a. the FL server) [7], [8]. The training data are stored locally and are not directly shared during the training process. Most of the existing FL training approaches are derived from the Federated Averaging (FedAvg) algorithm introduced in [9]. The goal is to train a global model that performs well on most FL clients.

A. Categorization of Federated Learning

FL can be categorized into horizontal FL (HFL), vertical FL (VFL) and federated transfer learning (FTL), according to how data are distributed in terms of feature and sample spaces among participating entities [7]. HFL refers to scenarios whereby participants share the same feature space but have different data samples. It is the most commonly adopted FL setting popularized by Google, which applied HFL to train language models in mobile devices [9]. In VFL, participants have overlapping data samples, but differ in the feature space. A typical application scenario would involve the collaboration of multiple organizations from different industry sectors (e.g., a bank and an e-commerce company) which have different data features but may have a large number of shared users. FTL is applicable when participants have little overlap in both the feature space and the sample space. For example, organizations from different industry sectors serving markets in different regions can leverage FTL to collaboratively build models. Existing PFL works mainly focus on the HFL setting which makes up the majority of the FL application scenarios [8]. The HFL setting is the focus of this paper. For brevity, we use the terms HFL and FL interchangeably in the rest of this survey.

B. Motivations for Personalized Federated Learning

Fig. 1 illustrates the key concepts and motivations for centralized machine learning (CML) [10], FL and PFL. We consider a cloud-based CML setting where data are pooled together in the cloud server to train an ML model. In this setting, the CML model achieves good generalization from the rich amount of data. However, CML faces bandwidth and latency challenges due to the sheer amount of data transferred to the cloud. It also does not preserve data privacy or not personalize well.

The FL setting assumes a federation of distributed clients, each with its own private local dataset. As these clients face data scarcity that limit their capacities to train effective

Alysa Ziying Tan is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore; Alibaba-NTU Singapore Joint Research Institute, NTU, Singapore; and Alibaba Group, Hangzhou, China.

Han Yu is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore.

Lizhen Cui is with the School of Software, Shandong University (SDU), Jinan, China; and the Joint SDU-NTU Centre for Artificial Intelligence Research (C-FAIR), SDU, Jinan, China.

Qiang Yang is with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong; and WeBank, Shenzhen, China.

*Corresponding authors: Han Yu (han.yu@ntu.edu.sg), Lizhen Cui (clz@sdu.edu.cn) and Qiang Yang (qyang@cse.ust.hk)

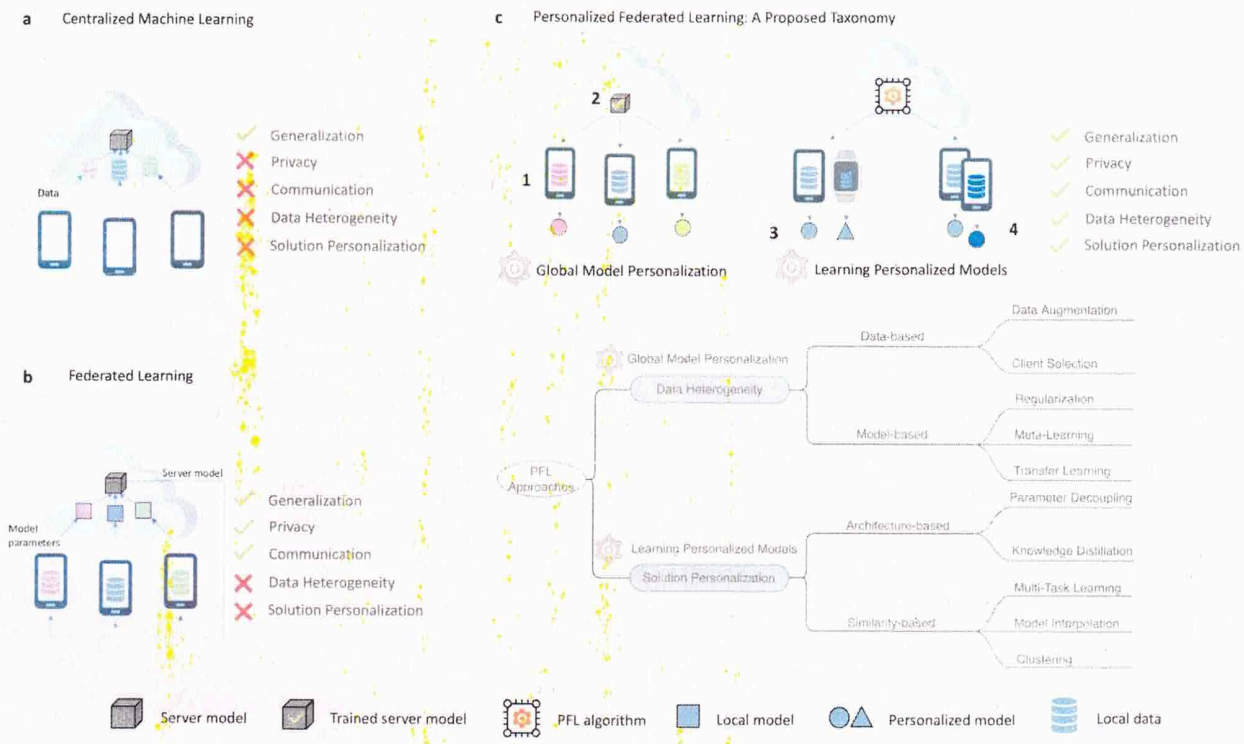


Fig. 1: Concept, Motivations & Proposed Taxonomy for Personalized Federated learning. **a**. Centralized machine learning (CML) which pools data together to train a central ML model. **b**. Federated learning (FL) which trains a global model under the orchestration of a central parameter server. Data resides in different data silos. **c**. Personalized federated learning (PFL) which addresses the limitations of FL through global model personalization and personalized models learning. **1–4** Four categories of PFL approaches: **1**) data-based, **2**) model-based **3**) architecture-based, **4**) similarity-based.

CML : centralized machine learning

local models, they are motivated to join the FL process to obtain a better performing model. FL enables collaborative model training on data silos in a privacy-preserving manner, which sets it apart from the CML setting. Additionally, FL is communication-efficient as it only transfers model parameters which are a fraction in size compared to transferring raw data. By considering privacy and communication constraints, FL is applicable to support a wide range of application scenarios such as Internet of Things (IoT), that entails privacy, connectivity, bandwidth and latency challenges in varying edge computing environments [11].

often overlooked

Challenges :

However, the general FL approach faces several fundamental challenges: (i) poor convergence on highly heterogeneous data, and (ii) lack of solution personalization. These issues deteriorate the performance of the global FL model on individual clients in the presence of heterogeneous local data distributions, and may even disincentivize affected clients from joining the FL process. Compared to traditional FL, PFL research seeks to address these two challenges.

Challenges

1) *Poor Convergence on Heterogeneous Data:* When learning on non-independent and identically distributed (non-IID) data, the accuracy of FedAvg is significantly reduced. This performance degradation is attributed to the phenomenon of client drift [12], as a result of the rounds of local training and synchronization on local data distributions that are non-IID.

Client Drift

Fig. 2 illustrates the effect of client drift on IID and non-IID data. In FedAvg, the server updates move toward the average of client optima. When data are IID, the averaged model is close to the global optimum w^* as it is equidistant to both local optima w_1^* and w_2^*. However, when data are non-IID, the global optimum w^* is not equidistant to the local optima. In this illustration, w^* is closer to w_2^*. The averaged model w^{t+1} will therefore be far from the global optimum w^*, and the global model does not converge to its true global optimum. As the FedAvg algorithm experiences convergence issues on non-IID data, careful tuning of hyperparameters (e.g., learning rate decay) is required to improve learning stability [13].

non-IID

2) *Lack of Solution Personalization:* In the vanilla FL setting, a single globally-shared model is trained to fit the "average client". As a result, the global model will not generalize well for a local distribution that is very different from the global distribution. Having a single model is often insufficient for practical applications which often face non-IID local datasets. Taking the example of applying FL to develop language models for mobile keyboards, users from different demographics are likely to have divergent usage patterns due to diverse generational, linguistic and cultural nuances. Certain words or emojis are likely be used predominantly by specific groups of users. For such a scenario, a more tailored prediction pattern is needed for each individual user in order for the word

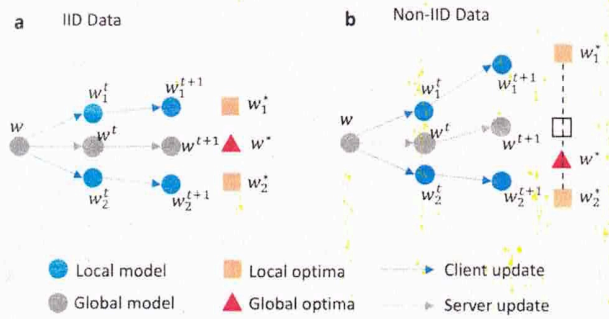


Fig. 2: Illustration of client drift in FedAvg for 2 clients with 2 local steps. **a** IID data setting. **b** Non-IID data setting.

IID: independently identical data

suggestions to be meaningful.

C. Contributions

References for PFL

There are several surveys on the general concepts, methods and applications of FL [7], [14]. Others review FL from the perspectives of privacy [15] and robustness [16]. Our survey focuses on PFL, which studies the problem of learning personalized models to handle statistical heterogeneity under the FL setting. There is a shortage of a comprehensive survey on PFL that provides a systematic perspective on this important topic for new researchers. In this paper, we bridge this gap in the current FL literature. Our main contributions are summarized as follows:

- We provide a succinct overview of FL and its categorization. A detailed analysis of the key motivations for PFL in the current FL settings is also included.
- Taxonomy — We identify personalization strategies to address key FL challenges, and offer a unique data-based, model-based, architecture-based and similarity-based perspective for guiding the review of the PFL literature. Based on this perspective, we propose a hierarchical taxonomy to present existing works on PFL, highlighting the challenges they face, their main ideas and assumptions they made which could introduce potential limitations.
- benchmarking — We discuss commonly adopted public datasets and evaluation metrics in the current literature for PFL benchmarking, and offer suggestions on enhancing PFL experimental evaluation techniques.
- We envision promising future trajectories of research towards new architectural design, realistic benchmarking, and trustworthy approaches towards building personalized federated learning systems.

II. Strategies for Personalized Federated Learning

In this section, we provide an overview of the PFL strategies which are the basis for our systematic and comprehensive review of existing PFL approaches. We organize the literature around the proposed taxonomy (Fig. 1c) that divides PFL methods according to the key challenges and personalization strategies involved.

$\min_{w \in \mathbb{R}^d} F(w) := \frac{1}{C} \sum_{c=1}^{C} f_c(w)$ (# clients)

Strategy I: Global Model Personalization

The first strategy addresses the performance issues in training a globally-shared FL model on heterogeneous data. When learning on non-IID data, the accuracy of FedAvg-based approaches is significantly reduced due to client drift. Under global model personalization, the PFL setup closely follows the general FL training procedure where a single global FL model is trained. The trained global FL model is then personalized for each FL client through a local adaptation step that involves additional training on each local dataset. This two-step "FL training + local adaptation" approach is commonly regarded as an FL personalization strategy by the FL community [8], [17]. As personalization performance directly depends on the generalization performance of the global model, many PFL approaches aim to improve the performance of the global model under data heterogeneity in order to improve the performance of subsequent personalization on local data. Personalization techniques for this category are classified into data-based and model-based approaches. Data-based approaches aim to mitigate the client drift problem by reducing the statistical heterogeneity among the clients' datasets, while model-based approaches aim to learn a strong global model for future personalization on individual clients or improve the adaptation performance of the local model.

Strategy II: Learning Personalized Models

The second strategy addresses the challenge of solution personalization. In contrast to the global model personalization strategy which trains a single global model, approaches in this category train individual personalized FL models. The goal is to build personalized models by modifying the FL model aggregation process. This is achieved through applying different learning paradigms in the FL setting. Personalization techniques are classified into architecture-based and similarity-based approaches. Architecture-based approaches aim to provide a personalized model architecture tailored to each client, while similarity-based approaches aim to leverage client relationships to improve personalized model performance where similar personalized models are built for related clients.

In personalized FL model training, the optimization objective is formulated differently from the vanilla FL setting, as an individual personalized model is learned for each client. Here, we provide formulations of the optimization objectives under the FL setting and the local learning setting in order to highlight the positioning of PFL approaches. The standard FL objective is given as

maybe # data points in C would work

FL Learning

$$\min_{w \in \mathbb{R}^d} F(w) := \frac{1}{C} \sum_{c=1}^{C} f_c(w), \tag{1}$$

where C is the number of participating clients, $w \in \mathbb{R}^d$ encodes the parameters of the global model and

$$f_c(w) := \mathbb{E}_{(x,y)\sim D_c}\left[f_c(w; x, y)\right] \tag{2}$$

represents the expected loss over the data distribution D_c of client c. The prevailing FL formulation minimizes the aggregation of local functions and entails a common output for

all clients using the global model without any personalization. In the presence of data heterogeneity (i.e., the underlying data distributions across the clients are not identical), simply minimizing the average local loss with no personalization will result in poor performance.

At the opposite end of the spectrum, we consider a local learning setting where each client c trains its own model θ_c locally without any communication with other clients. The objective is given as

$$\min_{\theta_1,\ldots,\theta_C \in \mathbb{R}^d} F(\theta) := \frac{1}{C}\sum_{c=1}^{C} f_c(\theta_c). \quad (3)$$

where $\theta_c \in \mathbb{R}^d$ encodes the parameters of the local model of client c. In this setting, the resulting models may not achieve good generalization performance as the number of training examples that the local models are exposed to are limited. Stronger generalization guarantees can be obtained with more collaboration amongst clients to exploit the pool of knowledge for model training.

Comparing the formulations of the standard FL and local learning settings, standard FL facilitates collaboration and knowledge sharing amongst clients but does not entail personalized outputs as it relies on a shared global model for client inference. On the other hand, local learning entails a fully personalized model for each client, but fails to leverage potential performance gains from inter-client collaboration. Given the need to achieve a balance between generalization and personalization performance, PFL approaches fall between the standard FL setting and the local learning setting.

III. Strategy I: Global Model Personalization

In this section, we survey PFL approaches following the global model personalization strategy. The main setup and configurations for these approaches are illustrated in Fig. 3. Based on our proposed taxonomy, they are divided into *Data-based Approaches* and *Model-based Approaches* as follows.

A. Data-based Approaches

Motivated by the client drift problem arising from federated training on heterogeneous data, data-based approaches aim to reduce the statistical heterogeneity of client data distributions. This helps to improve the generalization performance of the global FL model.

Data Augmentation

As the IID property of training data is a fundamental assumption in statistical learning theory, data augmentation methods to enhance the statistical homogeneity of the data have been extensively studied in the field of machine learning. Over-sampling techniques involving synthetic data generation (e.g., SMOTE [18] and ADASYN [19]), and under-sampling techniques (e.g., Tomek links [20]) have been proposed to reduce data imbalance. These techniques, however, cannot be directly applied under the FL setting, where data residing at the clients in the federation are distributed and private.

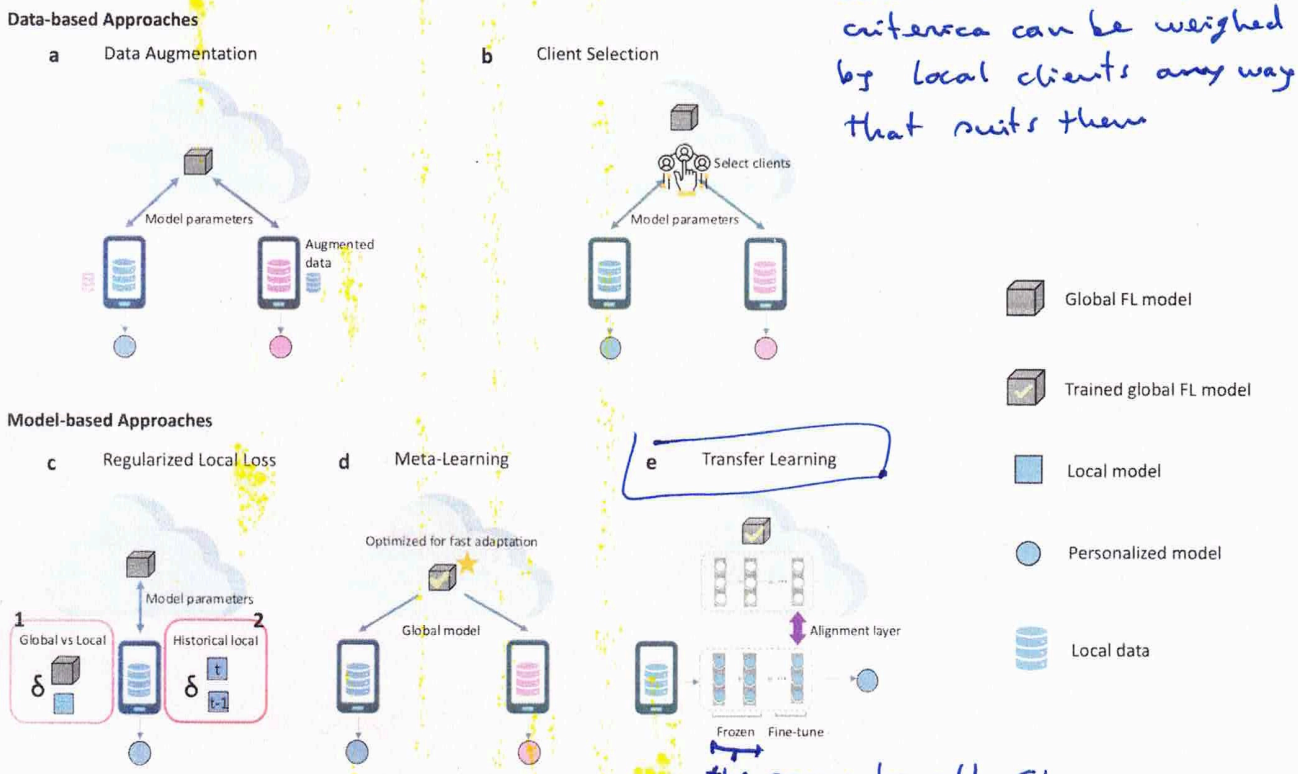
Data augmentation in FL (Fig. 3a) is highly challenging as it often requires some form of data sharing or relies on the availability of a proxy dataset that is representative of the overall data distribution. In [21], the authors proposed a data sharing strategy that distributes a small amount of global data balanced by classes to each client. Their experiments show that there is potential for significant accuracy gains (~30%) with the addition of a small amount of data. In [22], the authors proposed FAug, a federated augmentation approach that involves training a Generative Adversarial Network (GAN) model in the FL server. Some data samples of the minority classes are uploaded to the server to train the GAN model. The trained GAN model is then distributed to each client to generate additional data to augment its local data to produce an IID dataset. In [23], the authors proposed Astraea, a self-balancing FL framework to handle class imbalance by using Z-score based data augmentation and down-sampling of local data. The FL server requires statistical information about clients' local data distributions (e.g., class sizes, mean and standard deviation values). In [24], the authors proposed the FedHome algorithm that trains a Generative Convolutional Autoencoder (GCAE) model using FL. At the end of the FL procedure, each client performs further personalization on a locally augmented class-balanced dataset. This dataset is generated by executing the SMOTE algorithm on the low dimensional features of the encoder network based on the local data.

Client Selection

Another line of work focuses on designing FL client selection mechanisms to enable sampling from a more homogeneous data distribution, with the aim of improving model generalization performance (Fig. 3b). In [25], the authors proposed FAVOR which selects a subset of participating clients for each training round in order to mitigate the bias introduced by non-IID data. A deep Q-learning formulation for client selection was designed with the objective of maximizing accuracy, while minimizing the number of communication rounds. In a similar approach, a client selection algorithm based on the Multi-Armed Bandit formulation was proposed in [26] to select the subset of clients with minimal class imbalance. The local class distributions are estimated by comparing the similarity between the local gradient updates submitted to the FL server with the gradients inferred from a balanced proxy dataset residing on the server.

Recently, there is an emerging line of work that focuses on developing client selection strategies to tackle data and resource heterogeneity challenges that are prevalent in edge computing applications. For cross-device FL, there is often significant variability in hardware capabilities in terms of computation and communication capacities. Heterogeneity also exists in data, whereby the quantity and distribution of data differ among clients. Such diversity exacerbates challenges such as communication costs, stragglers and model accuracy. In [27], the authors proposed a tier-based FL system (TiFL) that groups clients into tiers based on training performance. The algorithm adaptively selects participating clients from the same tier for each training round by optimizing both accuracy and training time. This helps alleviate the performance issues caused by data and resource heterogeneity. In [28], the authors proposed FedSAE, a self-adaptive FL system that adaptively

Note in agent-based environments the objective criterea can be weighed by local clients any way that suits them



the same for all C's

Fig. 3: The setup & configurations of approaches that fall under *Strategy I: Global Model Personalization*. **a–b** Data-based approaches: (**a**) data augmentation, (**b**) client selection. **c–e** Model-based approaches: (**c**) regularized local loss; regularization can be performed **1**) between global and local models, **2**) between historical local model snapshots, (**d**) meta-learning, (**e**) transfer learning.

selects clients with larger local training loss in each training round to accelerate the convergence of the global model. A prediction mechanism of the affordable workload of each client is also proposed to enable the dynamic adjustment of the number of local training epochs for each client in order to improve device reliability.

B. Model-based Approaches

Although data-based approaches improve the convergence of the global FL model by mitigating the client drift problem, they generally need to modify the local data distributions. This may result in loss of valuable information associated with the inherent diversity of client behaviors. Such information can be useful for personalizing the global model for each client. In this section, we cover model-based global model personalization FL approaches. The objective is either to learn a strong global FL model for future personalization on each individual client, or to improve the adaptation performance of the local model.

1. Learn a good global model
2. adapt performance of Locals

Regularized Local Loss Model regularization is a common strategy for preventing overfitting and improving convergence when training machine learning models. In FL, regularization techniques can be applied to limit the impact of local updates. This improves convergence stability and the generalization of the global model, which in turn, can be used to produce better personalized models. Instead of just minimizing the local function $f_c(\theta)$, each client c minimizes the following objective:

$$\min_{\theta \in \mathbb{R}^d} h_c(\theta; w) := f_c(\theta) + l_{reg}(\theta; w), \tag{4}$$

where $l_{reg}(\theta; w)$ is the regularization loss, which is generally formulated as a function of the global model w and the local model θ_c of client c. Regularization can be applied in the following ways as illustrated in Fig. 3c:

1) Between Global and Local Models: Several works implement regularization between the global and local models to tackle the client drift problem that is prevalent in FL due to statistical data heterogeneity. FedProx [29] introduced a proximal term to the local sub-problem which considers the dissimilarity between the global FL model and local models to adjust the impact of local updates. Along with model dissimilarity, FedCL [30] further considers parameter importance in the regularized local loss function using Elastic Weight Consolidation (EWC) [31] from the field of continual learning. The importance of the weights to the global model is estimated on a proxy dataset in the FL server. They are then transferred to the clients where penalization steps are carried out to prevent important parameters of the global model from being changed