联邦学习

1.联邦学习简要介绍
- 提出的背景（Background）
- 定义（Definition）
- 分类（Category）
- 整体框架（Framework）
- 应用前景（Prospect）

2.联邦学习中的重要问题

3.我们已经开展的工作

4.关注的方向

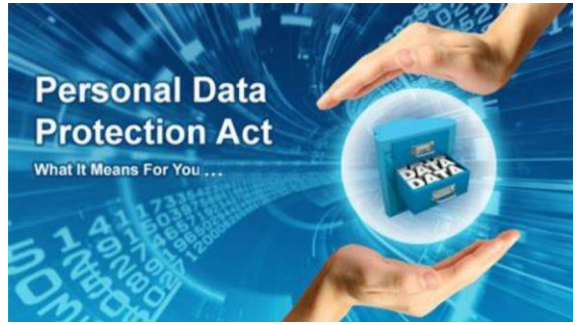
联邦学习介绍
Background

数据是人工智能的燃料

联邦学习介绍
Background

GDPR in European Union

CCPA in the US



PDPA in Singapore

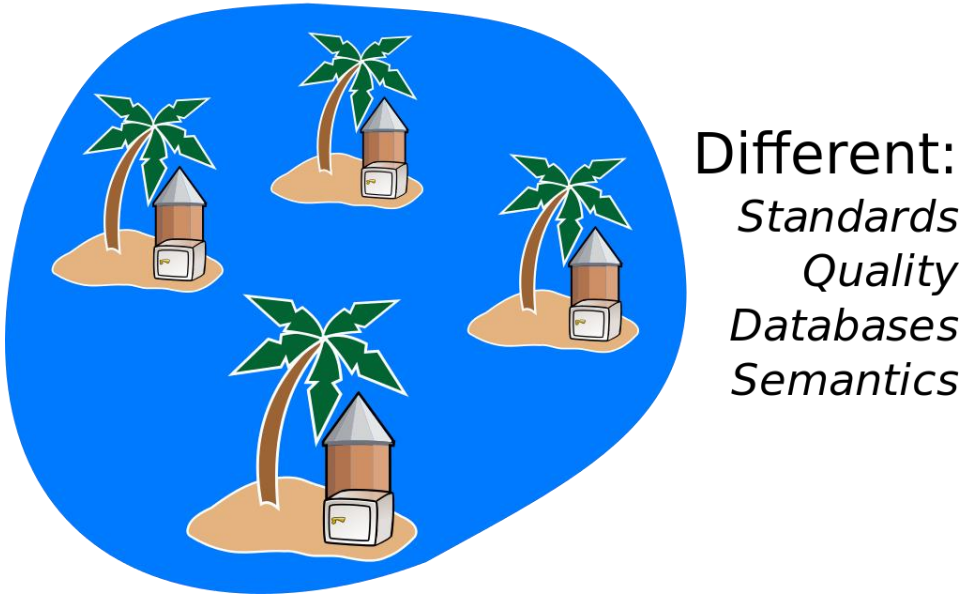
- 公众对于数据隐私安全的关注度越来越高
- 各国/组织相继出台数据隐私安全相关的法律
- 获取数据的难度越来越大

用户隐私和数据安全问题成为了人工智能新的挑战

联邦学习介绍
Background

人工智能应用领域（医疗，金融…）

- 用户隐私
- 行业竞争
- 规章制度
- …



数据难以被整和，可获得的数据量小且质量参差不齐

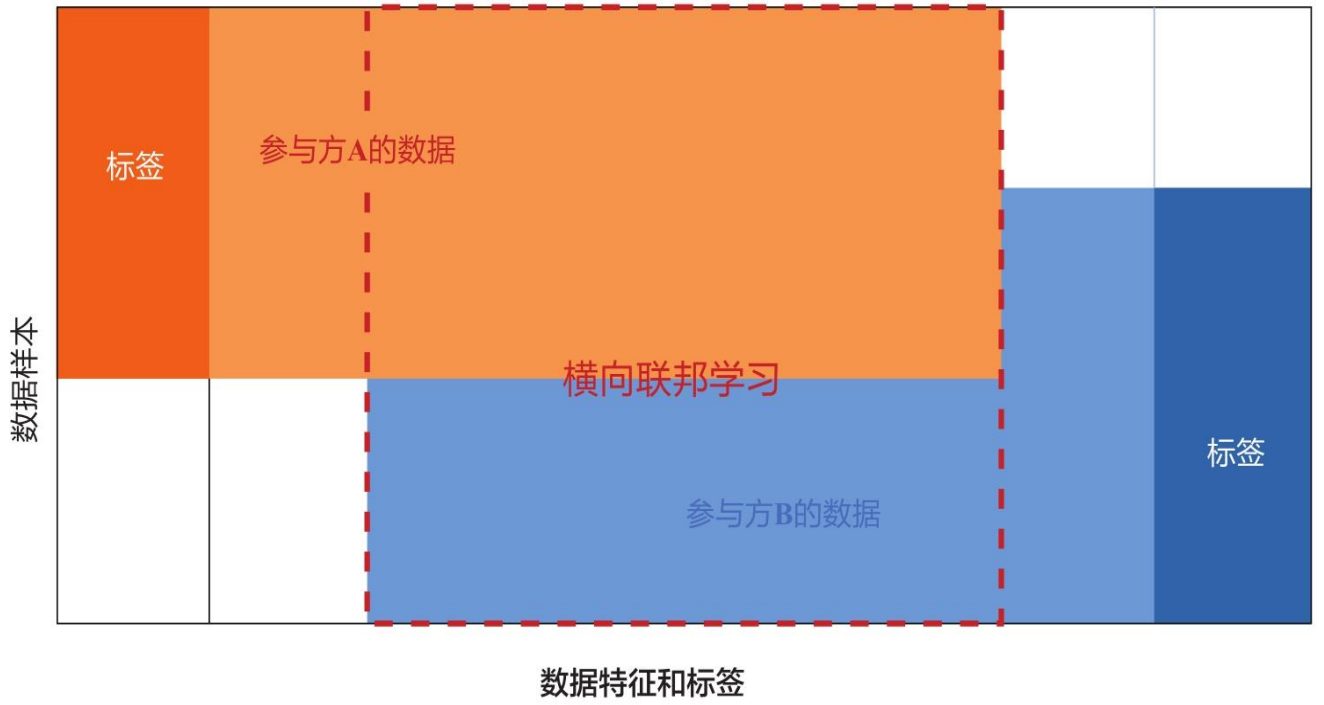
联邦学习的提出就是为了在确保用户隐私和数据安全的前提下解决数据孤岛（data islands）的问题

联邦学习介绍
Definition

- 有两个或以上的联邦学习参与方协作构建一个共享的机器学习模型。每一个参与方都拥有若干能够用来训练模型的训练数据。
- 在联邦学习模型的训练过程中，每一个参与方拥有的数据都不会离开该参与方，即数据不离开数据拥有者。
- 联邦学习模型相关的信息能够以加密方式在各方之间进行传输和交换，并且需要保证任何一个参与方都不能推测出其他方的原始数据。
- 联邦学习模型的性能要能够充分逼近理想模型（是指通过将所有训练数据集中在一起并训练获得的机器学习模型）的性能。

联邦学习介绍
Category

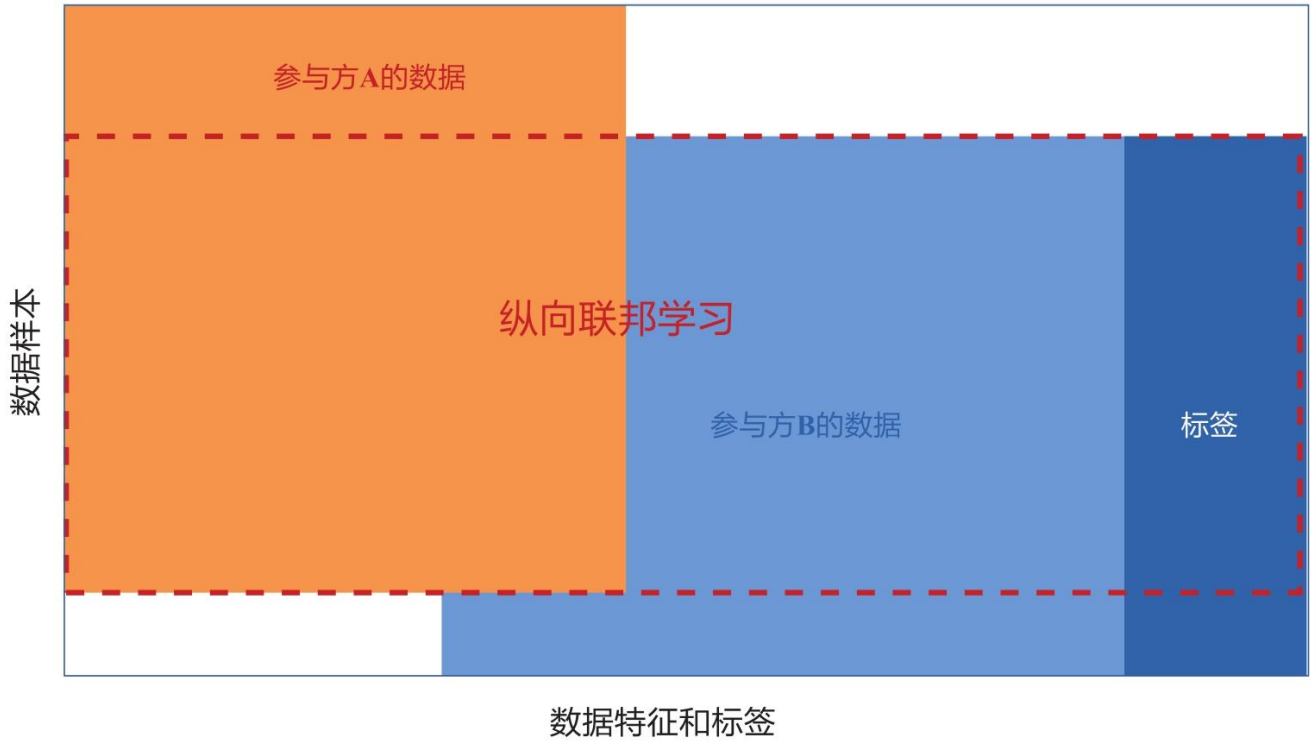
数据特征重叠较多，
而数据样本ID重叠较少



- 横向联邦学习也称为**特征对齐的联邦学习**（Feature-Aligned Federated Learning）
- 联合多个参与者的具有相同特征的多行样本进行联邦学习，即各个参与者的训练数据是**横向划分。**

联邦学习介绍
Category

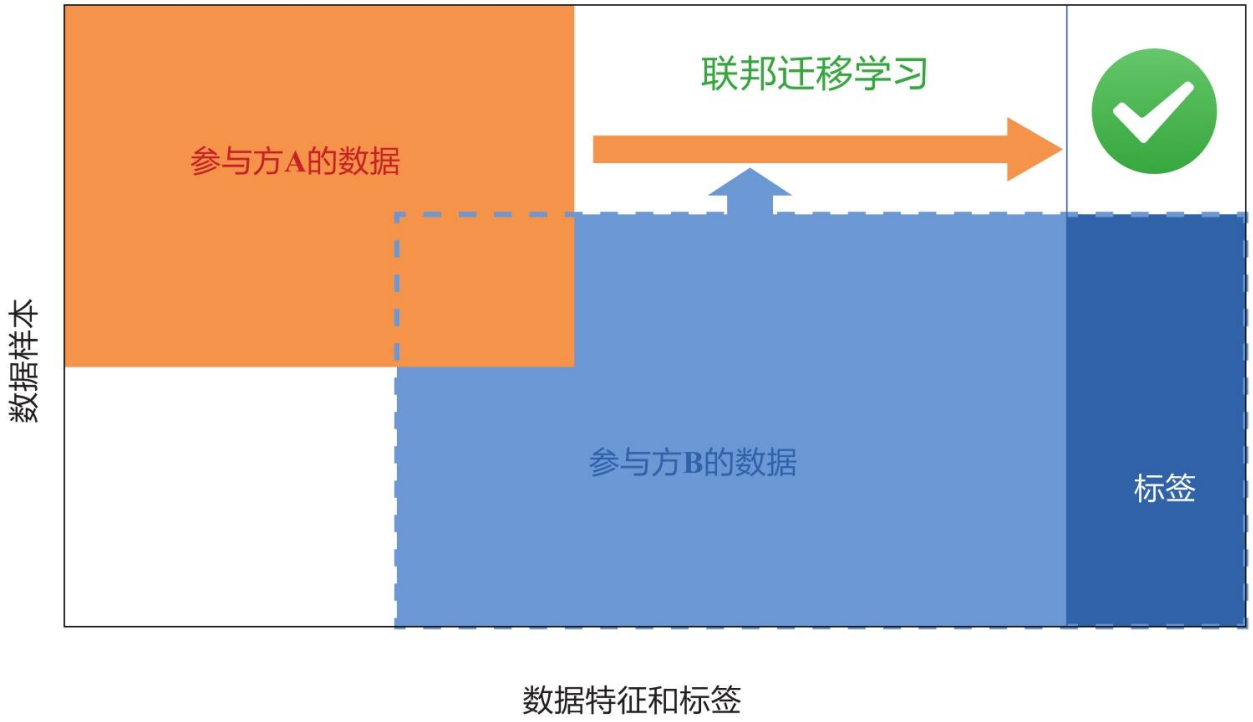
数据样本ID重叠较多，
而数据特征重叠较少



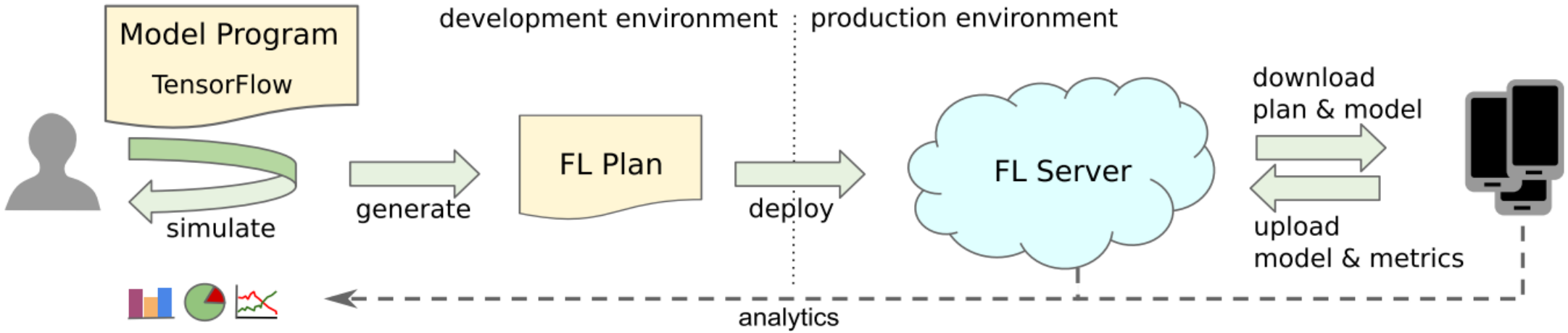
- 纵向联邦学习也称为**样本对齐的联邦学习**（Sample-Aligned Federated Learning）
- 联合多个参与者的共同样本的不同数据特征进行联邦学习，即各个参与者的训练数据是**纵向划分**的，称为**纵向联邦学习**。
- 纵向联邦使训练样本的**特征维度增多**。

联邦学习介绍
Category

数据样本ID和数据特征重叠都比较少

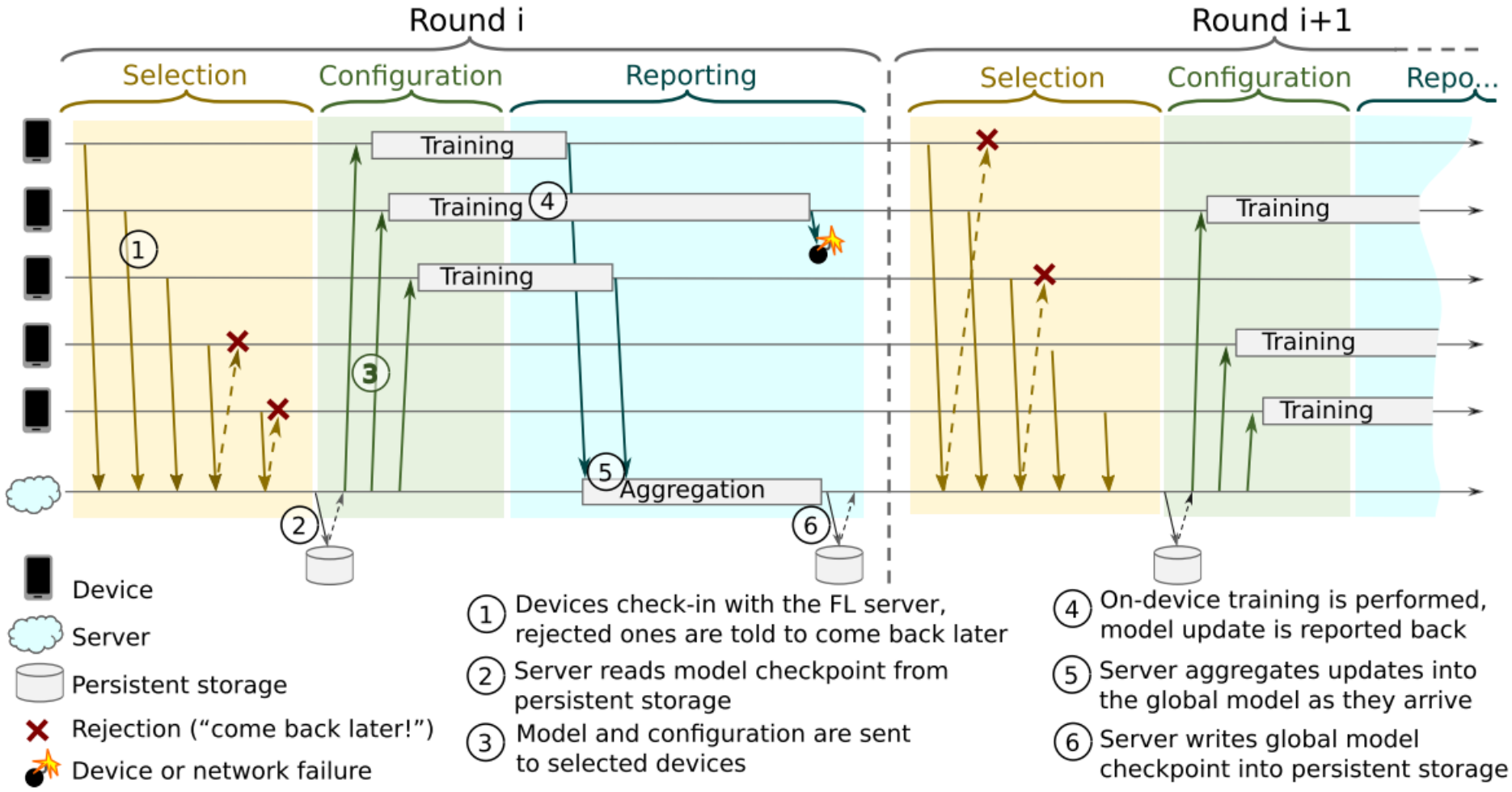


联邦学习介绍
Framework



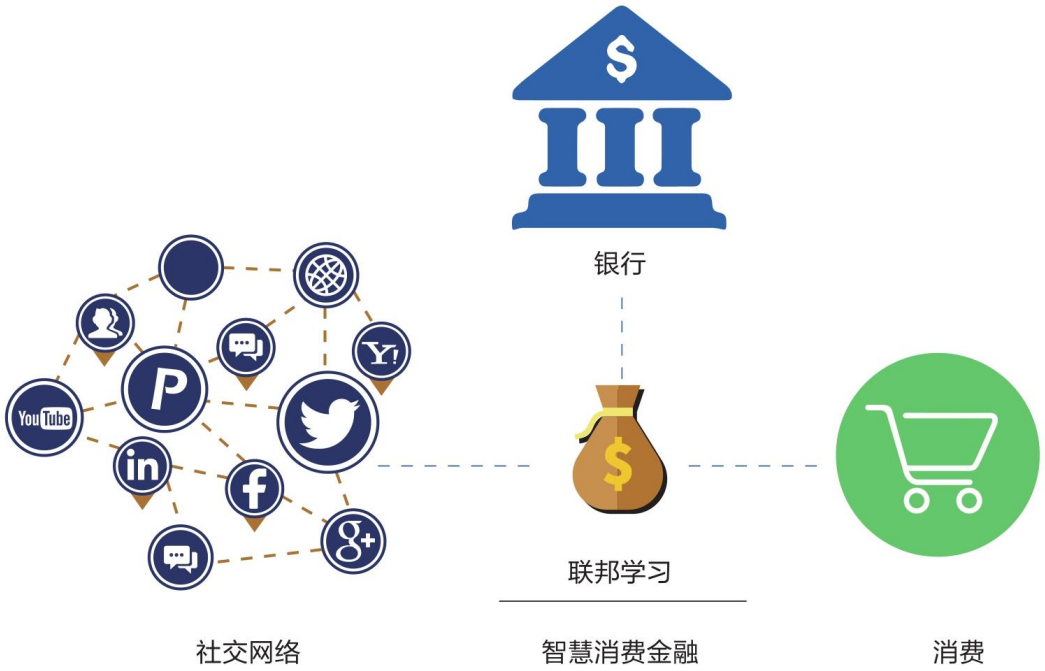
Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp , etc,"Towards Federated Learning at Scale: System Design," arXiv preprint arXiv:1902.01046, 2019.

联邦学习介绍
Framework



联邦学习介绍
Prospect

金融

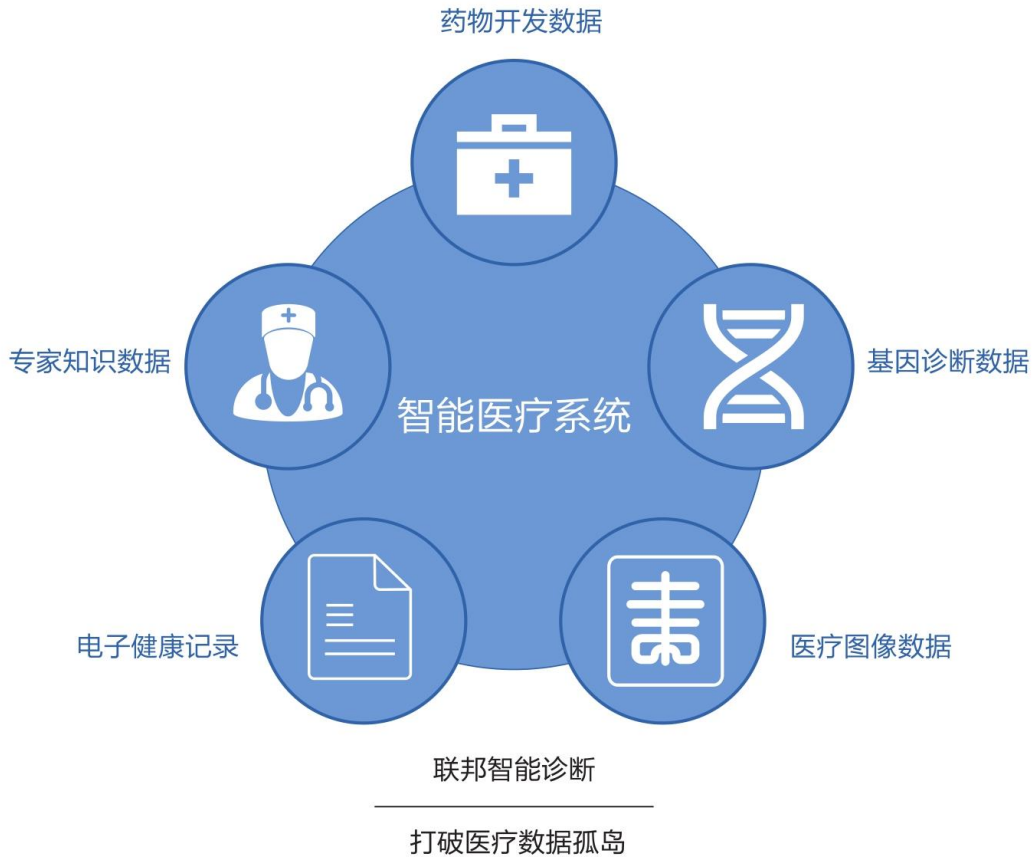


- 为了保护数据隐私和安全，银行与其他行业之间的数据壁垒难以跨越，因此数据无法直接聚合。
- 由银行与其他行业存储的数据通常是异构的，传统的机器学习不能直接处理异构数据。

智慧消费金融（Smart consumer finance）的目的是利用机器学习技术，为信用良好的消费者人群提供定制化的金融服务，以鼓励其消费。

联邦学习介绍
Prospect

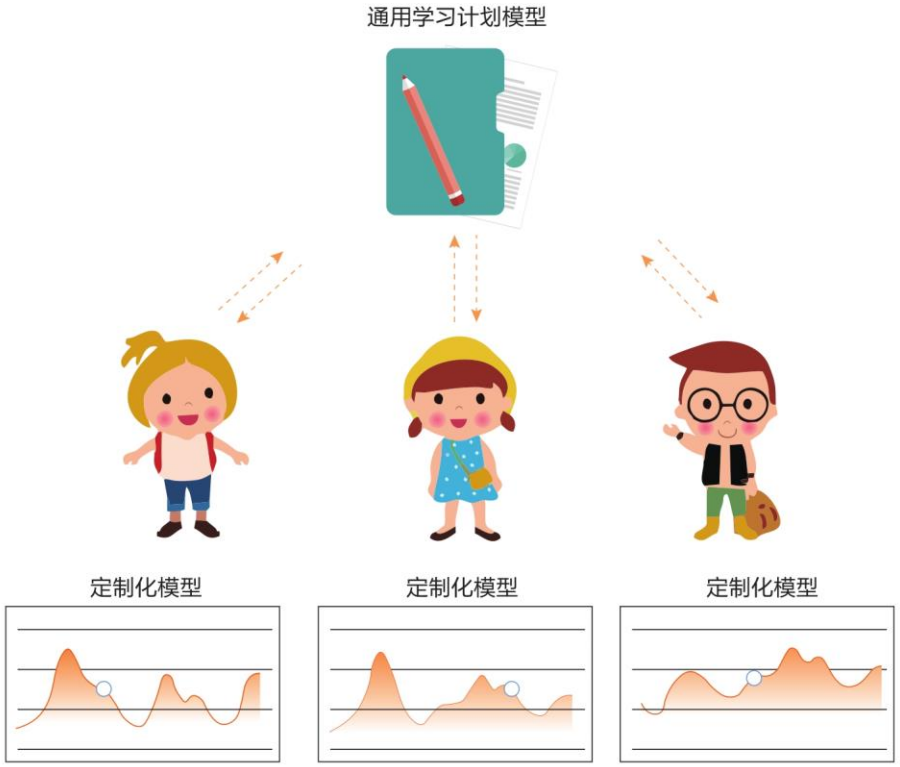
医疗



为了准确地诊断出一种疾病，我们可能需要从多个数据源收集多样性的特征，包括疾病症状、基因序列、医疗报告、检查结果及学术论文等。目前并没有一个稳定的数据源可以囊括所有这些特征，并且大部分的训练数据并没有被标注。

- 很难收集到足够数量的、具有丰富特征的、可以用来全面描述患者症状的数据。
- 医疗机构的数据对于隐私和安全问题特别敏感

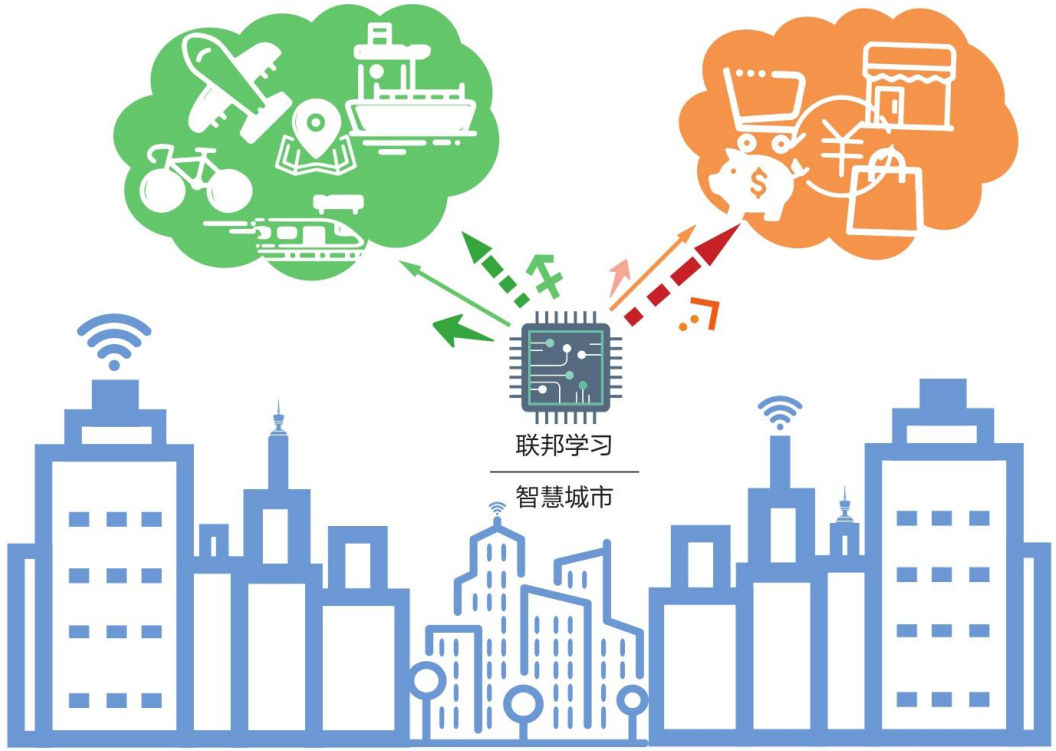
联邦学习介绍
Prospect



教育

教育机构可以利用联邦学习，基于存储在学生个人移动设备（如智能手机、iPad和笔记本电脑）中的数据，协作地构建一个通用学习计划模型。该通用学习计划模型可以为背景相似的学生制定标准化的学习计划。在此模型基础上，还可根据每一个学生的特长、需求、技能和兴趣，构建定制化、个性化的学习指导模型。

联邦学习介绍
Prospect



智慧城市

- 数据碎片化，孤岛化
- 数据对于隐私和安全问题很敏感

通过使用联邦学习，网约车公司可以协作地建立最优模型，解决车辆路线问题。这对于公司来说，不仅可以直接增加营收和提高客户满意度，还能通过分流和减少城市交通拥堵来获得额外的收益。

联邦学习重要问题

1.联邦学习简要介绍（Introduction）

2.联邦学习中的重要问题

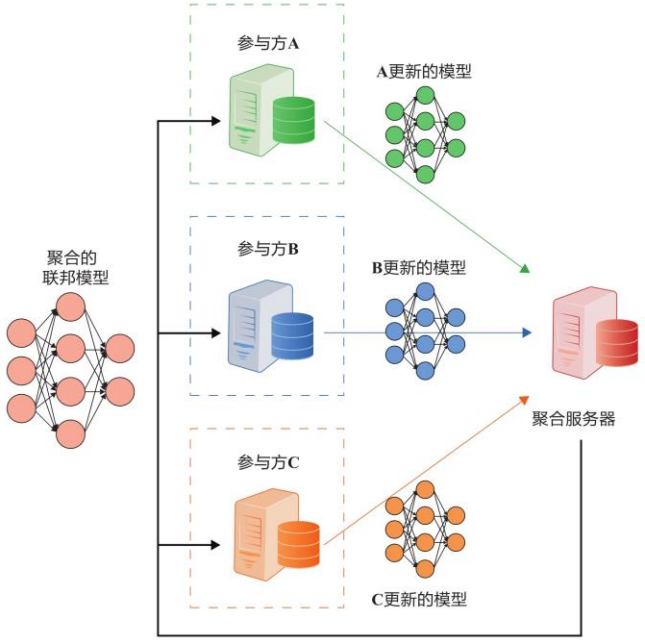
- 通信效率（Communication-efficiency）
- 系统异构性（Systems Heterogeneity）
- 隐私（Privacy）
- 安全（Security）
- 资源配置（Resource Allocation）

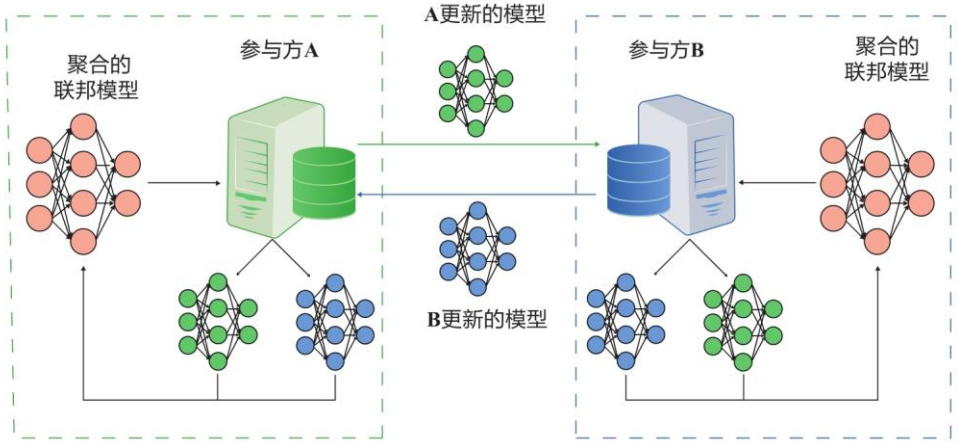
3.我们已经开展的工作

4.关注的方向

联邦学习重要问题
Communication-efficiency

通信架构（Communication Architecture）





集中式（C/S架构）

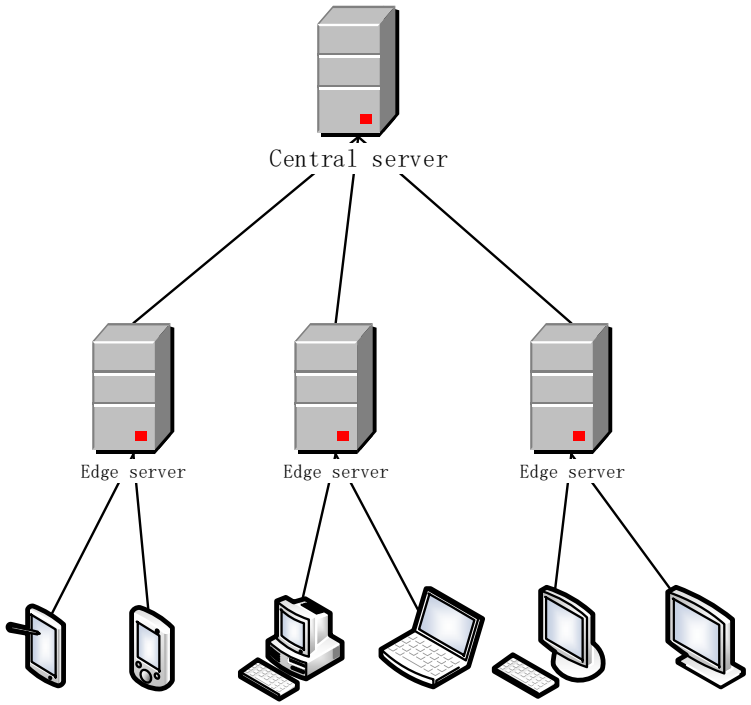
- 结构简单
- 目前采用最多的架构
- 对于server的可信度和可靠度要求很高
- 存在单点失效问题（Single-point Fault）

完全分布式

- 不需要服务器
- 适用于无法找到可信第三方的场景
- 通信效率更高
- 难以管理

联邦学习重要问题
Communication-efficiency

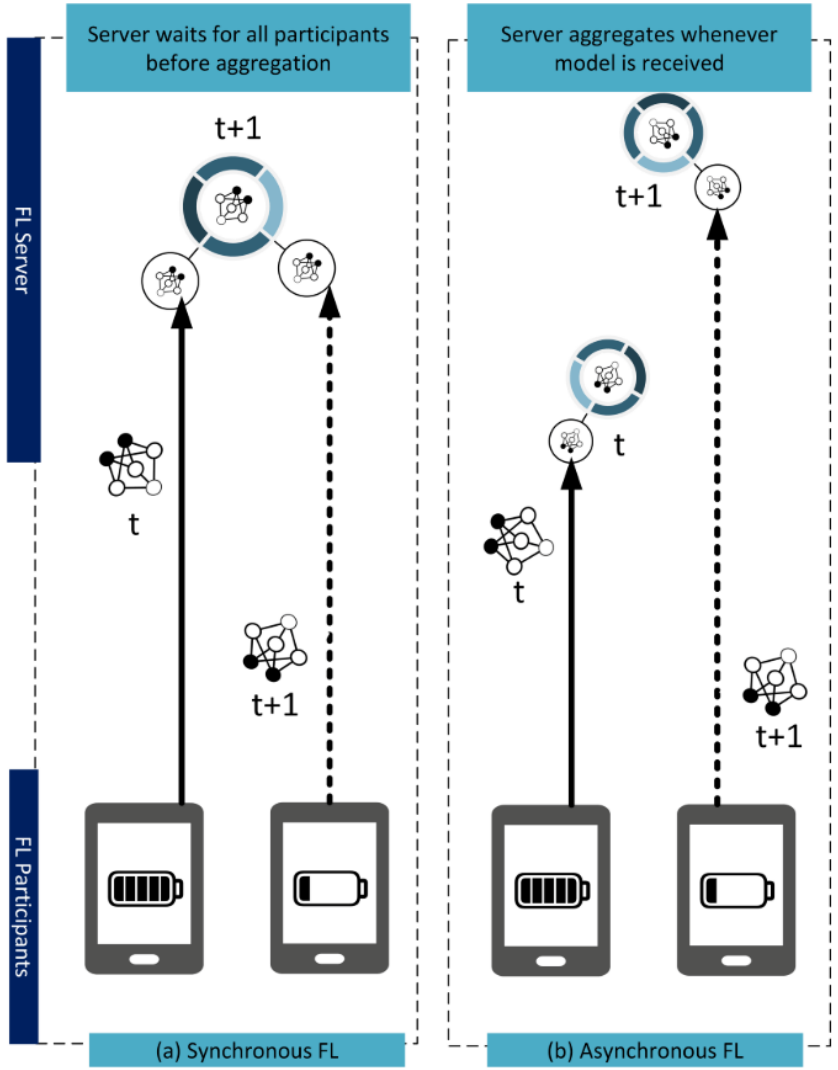
通信架构（Communication Architecture）



层次式架构

- 加入edge server
- 改善了C/S架构的通信效率

联邦学习重要问题
Communication-efficiency



通信架构（Communication Architecture）

同步/异步?

同步通信
- 结构简单
- 受最差参与者性能的制约 （straggler effect）

异步通信
- No straggler effect
- 陈旧模型的问题

联邦学习重要问题
Communication-efficiency

Algorithm 1 `FederatedAveraging`. The K clients are indexed by k; B is the local minibatch size, E is the number of local epochs, and η is the learning rate.

Server executes:

initialize w_0

for each round $t = 1, 2, \dots$ **do**

$m \leftarrow \max(C \cdot K, 1)$

$S_t \leftarrow$ (random set of m clients)

for each client $k \in S_t$ **in parallel do**

$w_{t+1}^k \leftarrow \text{ClientUpdate}(k, w_t)$

$w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^k$

ClientUpdate(k, w): *// Run on client k*

$\mathcal{B} \leftarrow$ (split $\mathcal{P}_k$ into batches of size B)

for each local epoch i from 1 to E **do**

for batch $b \in \mathcal{B}$ **do**

$w \leftarrow w - \eta \nabla \ell(w; b)$

return w to server

减少通信开销的方法:

1. 减少通信的轮次
2. 减少每次通信的数据量

本地迭代更新
(Local Updating)

"Communication-Efficient Learning of Deep Networks from Decentralized Data" arXiv:1602.05629 2017

联邦学习重要问题
Communication-efficiency

减少通信开销的方法：

1. 减少通信的轮次
2. 减少每次通信的数据量

- **模型裁剪/剪枝（Pruning）**
- **网络分解（张量分解）**
- **权值共享**
- **权重量化（Quantization）**
- **迁移学习/网络精馏**

模型压缩
（Model Compression）

联邦学习重要问题
Systems Heterogeneity

硬件异构性

- 计算能力
- 存储
- 电源
- 网络状态
- …

straggler effect

1. 参与方选择（Participant Selection）
2. 异步通信（asynchronous communication）
3. 错误容忍（Fault Tolerance）

系统必须能适配多种硬件（CPU/GPU）

如何应对训练中的掉队者（dropped devices）？

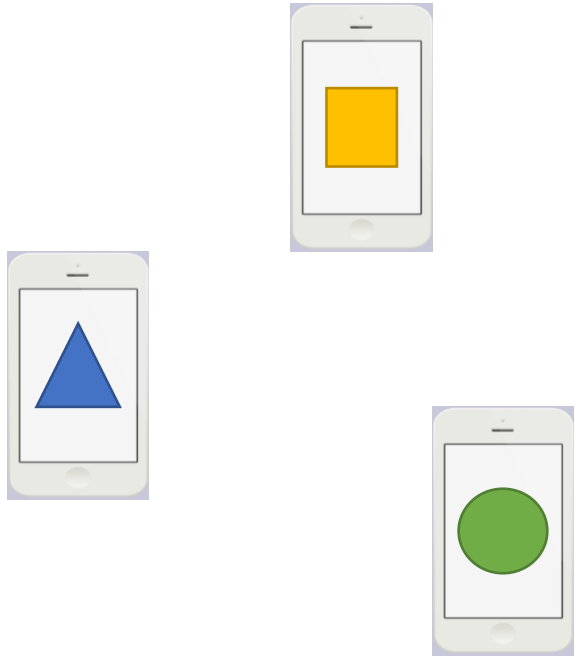
联邦学习重要问题
Systems Heterogeneity

数据的统计学异构性（Statistical heterogeneity）

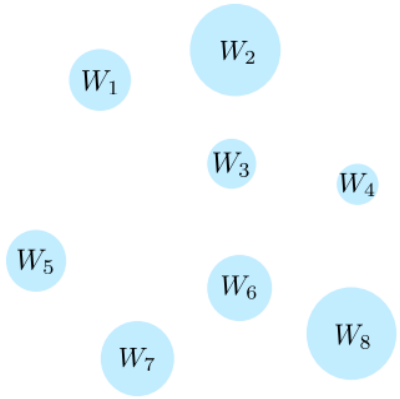
联邦系统大多数情况下面对的是
非独立同分布的的数据（non-IID data distribution）

- 样本数量
- 数据质量
- 数据格式
- 数据分布

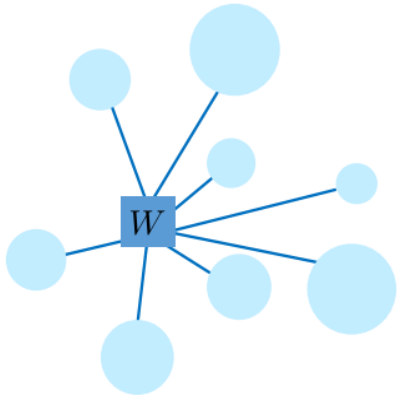
均有可能有很大的差异



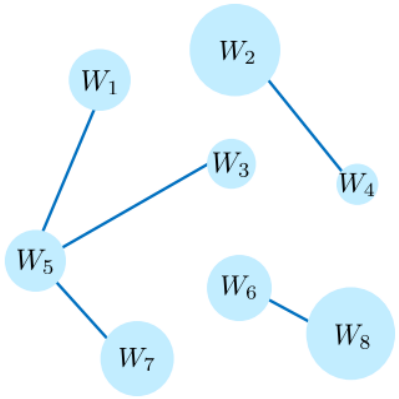
联邦学习重要问题
Systems Heterogeneity



(a) Learn personalized models for each device; do not learn from peers.



(b) Learn a global model; learn from peers.



(c) Learn personalized models for each device; learn from peers.

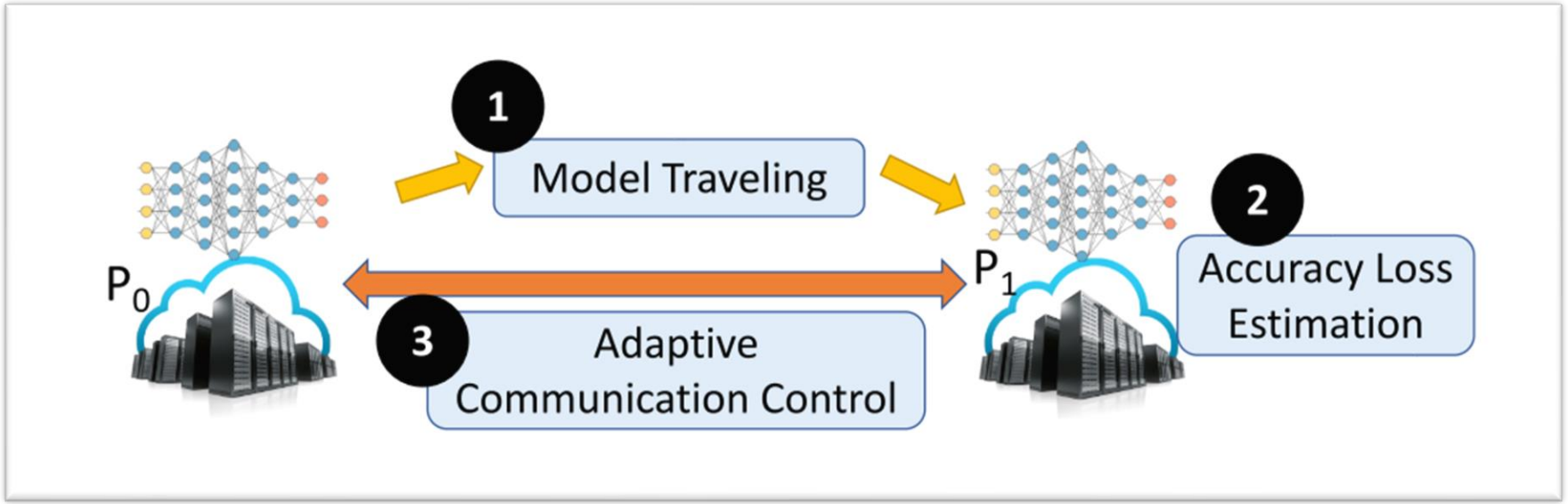
多任务学习（multi-task learning）

可以通过学习每个设备独立但相关的模型来实现个性化，同时通过多任务学习来实现共享。

V. Smith, C.-K. Chiang, M. Sanjabi, and A. Talwalkar. Federated multi-task learning. In Advances in Neural Information Processing Systems, 2017.

联邦学习重要问题
Systems Heterogeneity

通过控制通信来解决数据的统计学异质性问题



1. 参与方将模型发送到服务端
2. 服务端测试模型的精确度，依据模型的精确度来反映数据的偏移
3. 服务端基于模型精确度的评价控制通信（同时可以减少通信开销）

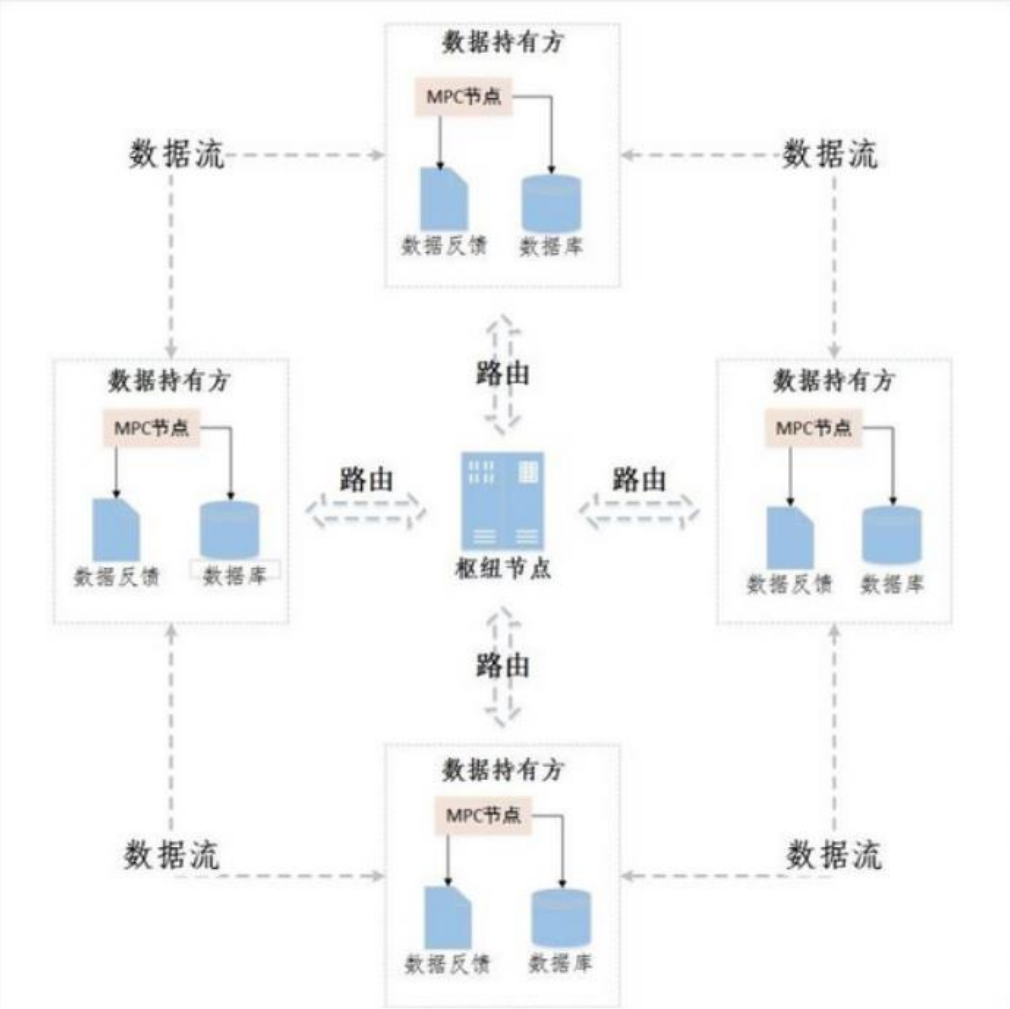
Kevin Hsieh, Amar Phanishayee, Onur Mutlu and Phillip B. Gibbons, “The Non-IID Data Quagmire of Decentralized Machine Learning,” arXiv preprint arXiv:1910.00189, 2020.

联邦学习重要问题

Privacy

- 安全多方计算（Secure Multi-party Computation (MPC)）
- 同态加密（Homomorphic Encryption (HE)）
- 差分隐私（Differential Privacy (DP)）

联邦学习重要问题
Privacy



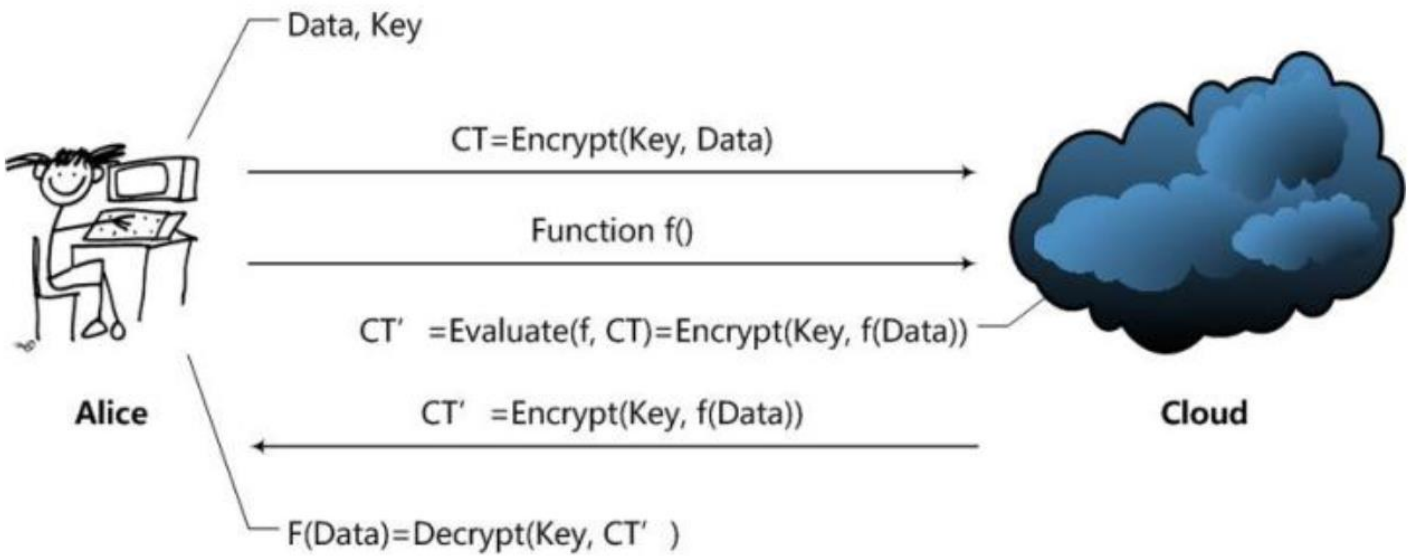
安全多方计算 (MPC)

- 每个数据持有方可发起协同计算任务。
- 枢纽节点负责路由寻址
- MPC节点从本地数据库中查询所需数据，进行协同计算。在保证输入隐私性的前提下，各方得到正确的数据反馈，
- 整个过程中本地数据没有泄露给其他任何参与方。

特点

- 针对无可信第三方的场景
- 去中心化
- 通信开销大

联邦学习重要问题
Privacy



1.Alice对数据进行加密。并把加密后的数据发送给Cloud;
2.Alice向Cloud提交数据的处理方法，这里用函数f来表示;
3.Cloud在函数f下对数据进行处理，并且将处理后的结果发送给Alice;
4.Alice对数据进行解密，得到结果。

同态加密 (HE)

- 基于数学难题的计算复杂性理论的密码学技术
- 其关注的是数据处理安全
- 用户不能对加密结果做任何操作
- 用户可以对加密数据进行处理，但是处理过程不会泄露任何原始内容

缺点:
- 计算开销大

联邦学习重要问题

Privacy

差分隐私 (DP)

$$Pr\{A(D) = O\} \leq e^{\epsilon} \cdot Pr\{A(D') = O\}$$

随机化算法 A（所谓随机化算法，是指对于特定输入，该算法的输出不是固定值，而是服从某一分布），其分别作用于两个相邻数据集得到特定输出 O 的概率应大致相等。

ε　控制精确度与隐私之间的权衡

最简单的方法就是加入随机化的噪音，最常见的是拉普拉斯噪音（Laplace noise）。

联邦学习重要问题
Security

数据毒害（Data Poisoning Attacks）

在联邦系统中，由于数据保留在本地，因此联邦系统无法检测参与方的实际训练数据，这样，一些恶意的参与者就会试图用一些“假数据”来进行训练，进而影响全局模型的精度。

多点攻击可以显著提高攻击成功率

TABLE VI: The accuracy and attack success rates for no-attack scenario and attacks with 1 and 2 sybils in a FL system with MNIST dataset [163].

	Baseline	**Attack 1**	**Attack 2**
Number of honest participants	10	10	10
Number of sybil participants	0	1	2
The accuracy (digits: 0, 2-9)	90.2%	89.4%	88.8%
The accuracy (digit: 1)	96.5%	60.7%	0.0%
Attack success rate	0.0%	35.9%	96.2%

Y. LeCun, L. Bottou, Y. Bengio, P. Haffner et al., “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

联邦学习重要问题
Security

更新模型的相似度?

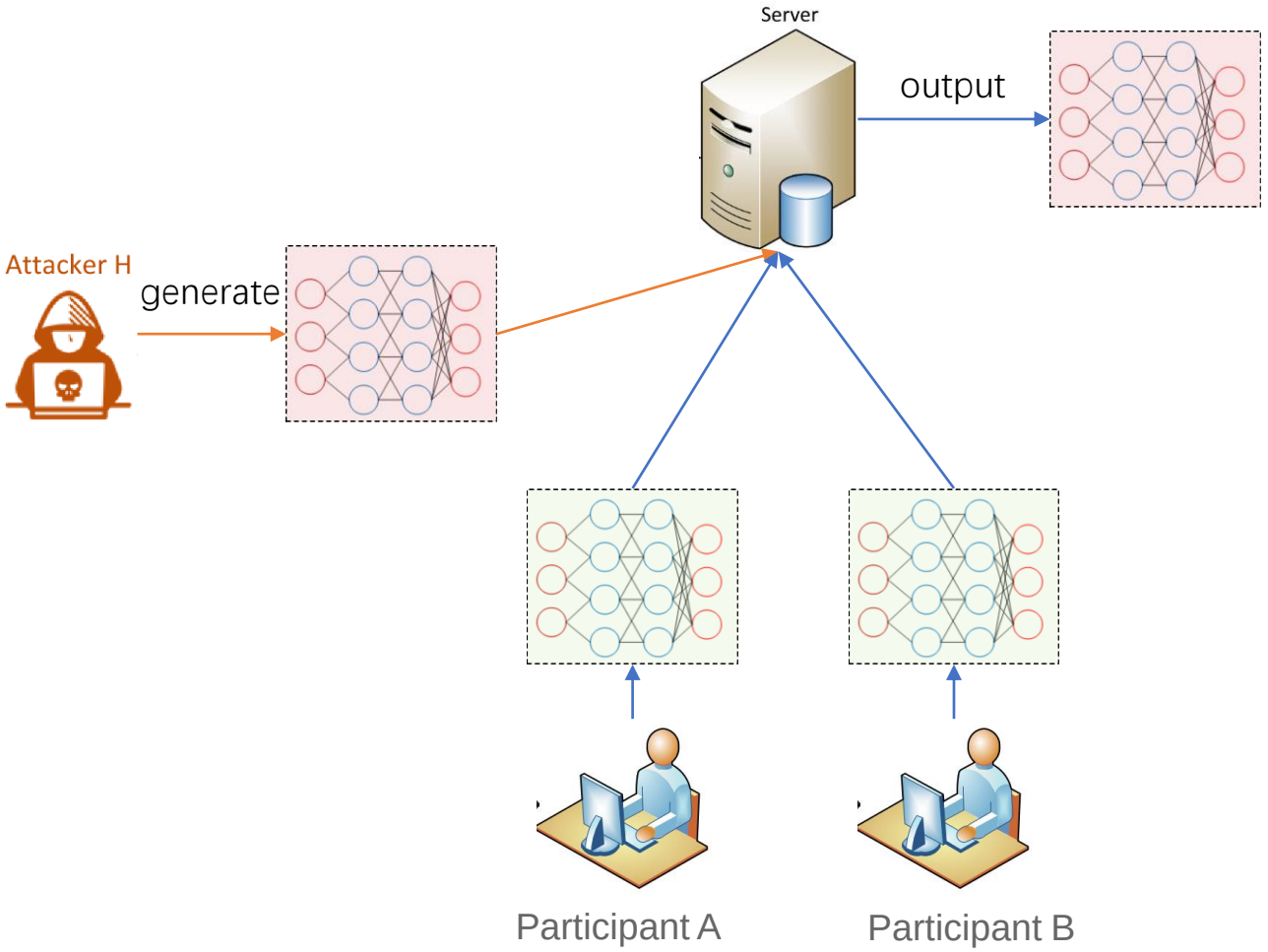
数据毒害的防御手段

FoolsGold：
我们可以依据更新的模型将可信的参与者与攻击者区分开，理由是联邦系统中参与者的数据大都是non-IID，每个参与方都有其自己的特点，所以更新模型的相似度也大不相同；但是攻击者的更新模型之间相似度会更加的高。

C. Fung, C. J. Yoon, and I. Beschastnikh, “Mitigating sybils in federated learning poisoning,” arXiv preprint arXiv:1808.04866, 2018.

联邦学习重要问题
Security

模型毒害（Model Poisoning Attacks）

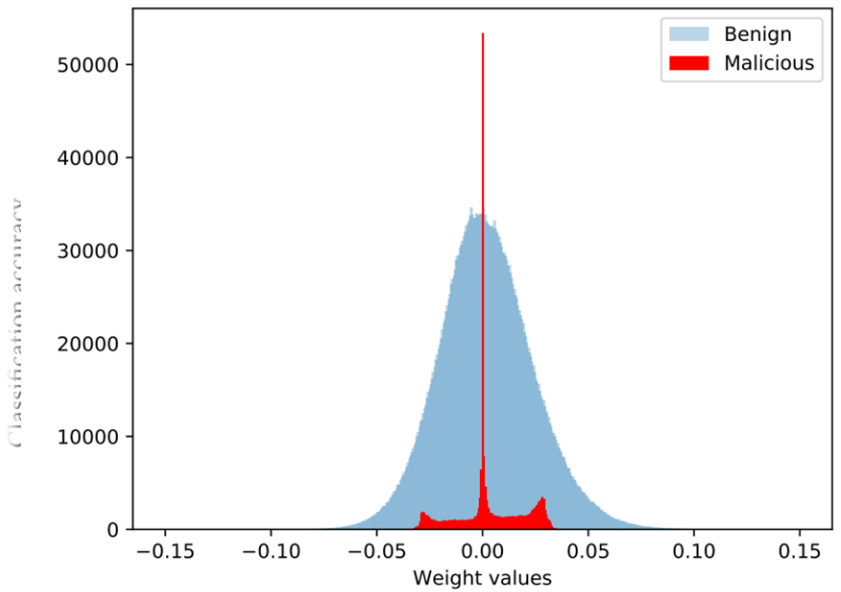


在数据毒害中，攻击者是基于“假数据”和多个攻击者进行攻击的，在模型毒害中，攻击者可以直接修改更新模型。这种攻击情景下，即使只有一次攻击，也可以对全局模型产生毒害，使其精度大大降低。

破坏效率更高
防御难度更大

A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, “Analyzing federated learning through an adversarial lens,” arXiv preprint arXiv:1811.12470, 2018.

联邦学习重要问题
Security



(b) Comparison of weight update distributions for benign and malicious agents

模型毒害防御方法

1.测试更新模型

测试更新模型，如果这个更新模型的性能很差，那么就标记上传者为潜在的攻击者（并且不将其更新模型聚合），并对其进行多轮的跟踪，如果其更新模型性能一直很差，就将其标记为攻击者。

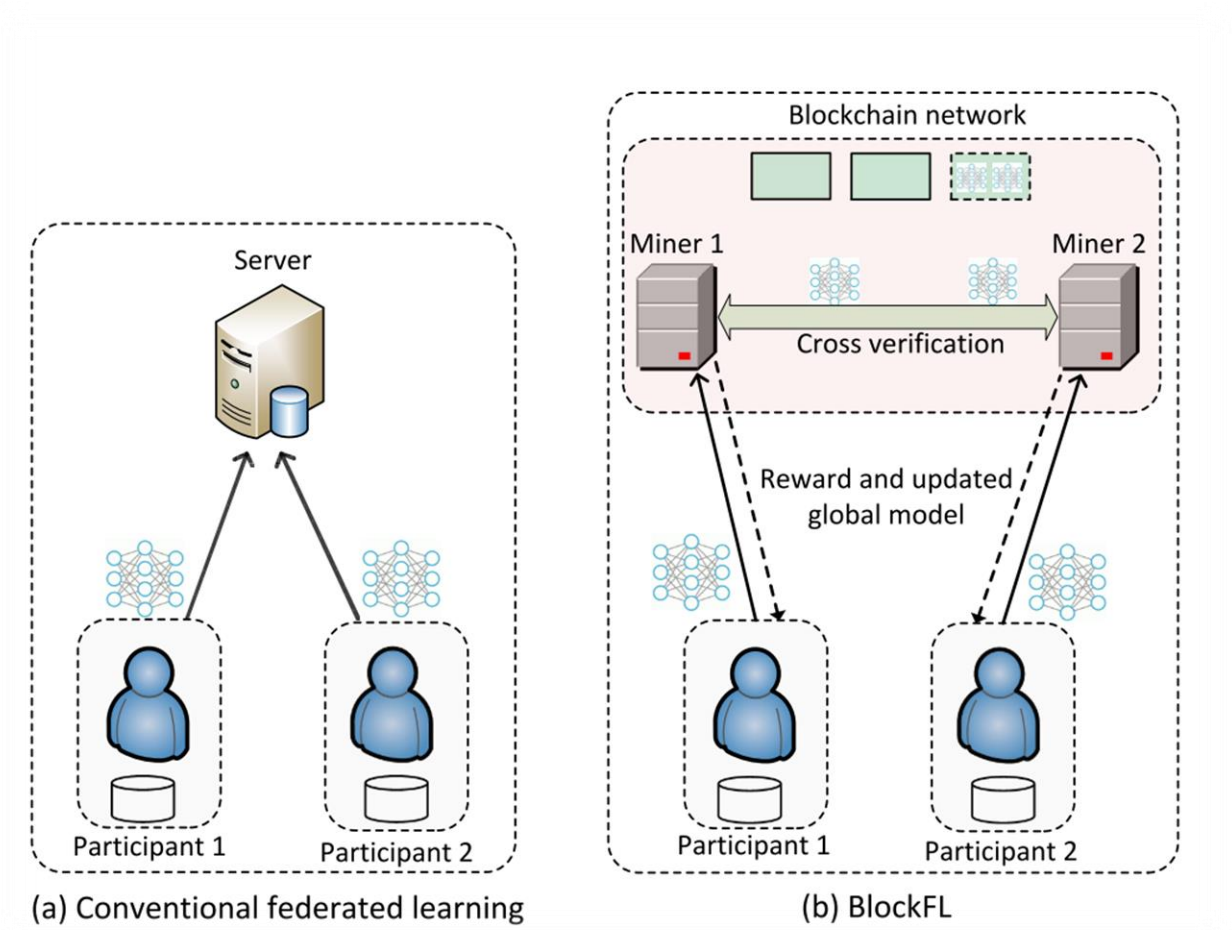
2.对比更新模型差异

对比更新模型，如果与参与者的更新模型与其他参与者的更新模型差别过大，同样将其标记为潜在攻击者（并且不将其更新模型聚合），对其进行多轮的跟踪，最终判定其是否为攻击者。

存在的问题

以上两种方法都会产生巨大的计算开销，在联邦系统规模比较大的时候，参与者可多达上百万，对每个参与者的模型均进行检测会变得非常困难。

联邦学习重要问题
Security



搭便车（Free-Riding Attacks）

参与者从全局模型中受益，但是却不对其做出贡献。恶意参与者会假装自己拥有的数据很少或者硬件水平较低，借此来保留自己的资源。

解决方法

区块链技术

参与者本地训练的更新模型依赖区块链技术进行交换和验证。每一个参与者将自己的更新模型发送到区块链网络中矿工，然后得到一个和训练样本数相称的一个奖励。

这个框架不仅可以防止这种搭便车行为，而且可以通过奖励的形式激励所有参与者为全局模型做出贡献。

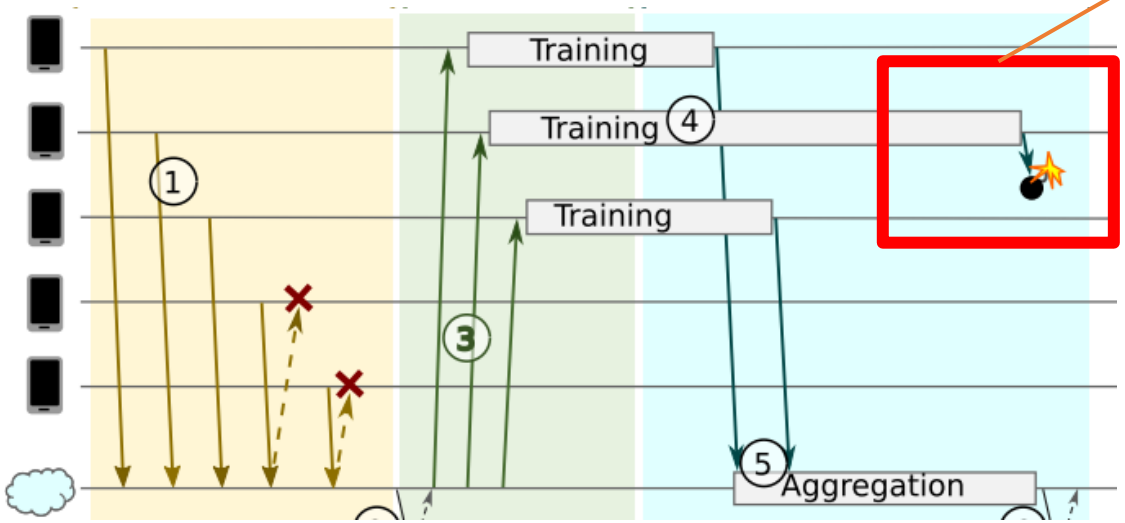
问题

区块链技术产生的开销较大

H. Kim, J. Park, M. Bennis, and S.-L. Kim, “On-device federated learning via blockchain and its latency analysis,” arXiv preprint arXiv:1808.03949, 2018.

联邦学习重要问题
Resource Allocation

参与方选择（Participant Selection）



参与方在训练中途可能会以为各种原因（电源关闭，网络连接断开）无法响应。

Ignore → 模型精度受影响

Wait → 系统效率降低

每一轮全局迭代均选择一部分更优秀的参与者加入训练

联邦学习重要问题
Resource Allocation

如何评价参与者?

1.硬件水平

- 计算能力：CPU/GPU性能
- 存储能力
- …

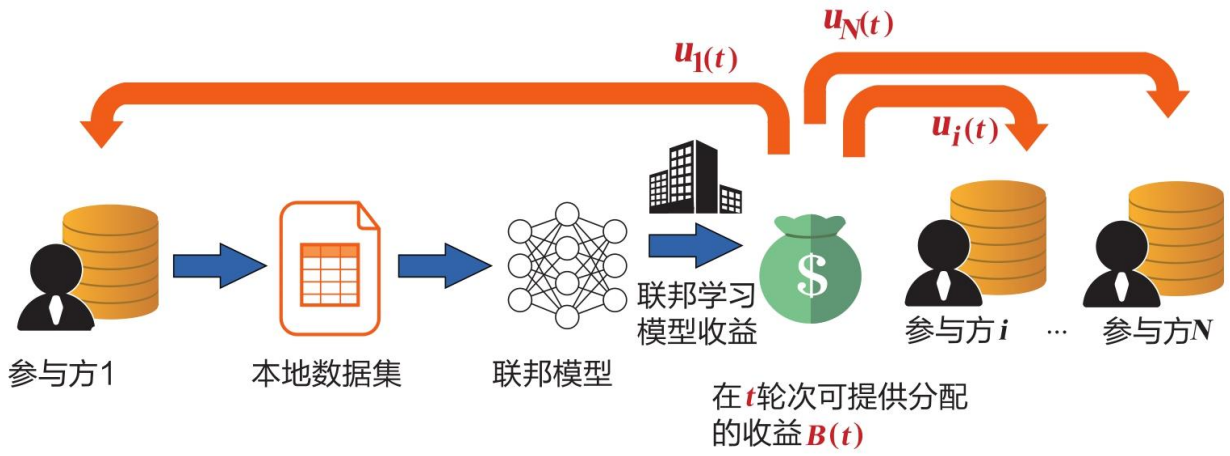
2.声誉

依据用户的历史参与信息，包括历史参与的成功上传比率，更新模型的质量等

联邦学习重要问题
Resource Allocation

激励机制（Incentive mechanisms）

在联邦学习中，如何建立激励机制使得参与方持续参与到数据联邦中是一项重要的挑战。实现这一目标的关键是制定一种奖励方法，公平公正地与参与方们分享联邦产生的利润。



参与者利用自己的数据和算力训练模型，为全局模型做出贡献，联邦系统通过向参与者发放奖励的方式激励参与方使用更多更好的数据参与更长时间的训练。

$$\hat{u}_i(t) = \frac{u_i(t)}{\sum_{i=1}^{N} u_i(t)} B(t),$$

联邦学习重要问题
Resource Allocation

收益分配的依据?

联邦角度

1.参与方自身的价值
- 硬件水平
- 数据质量&数量
- 声誉

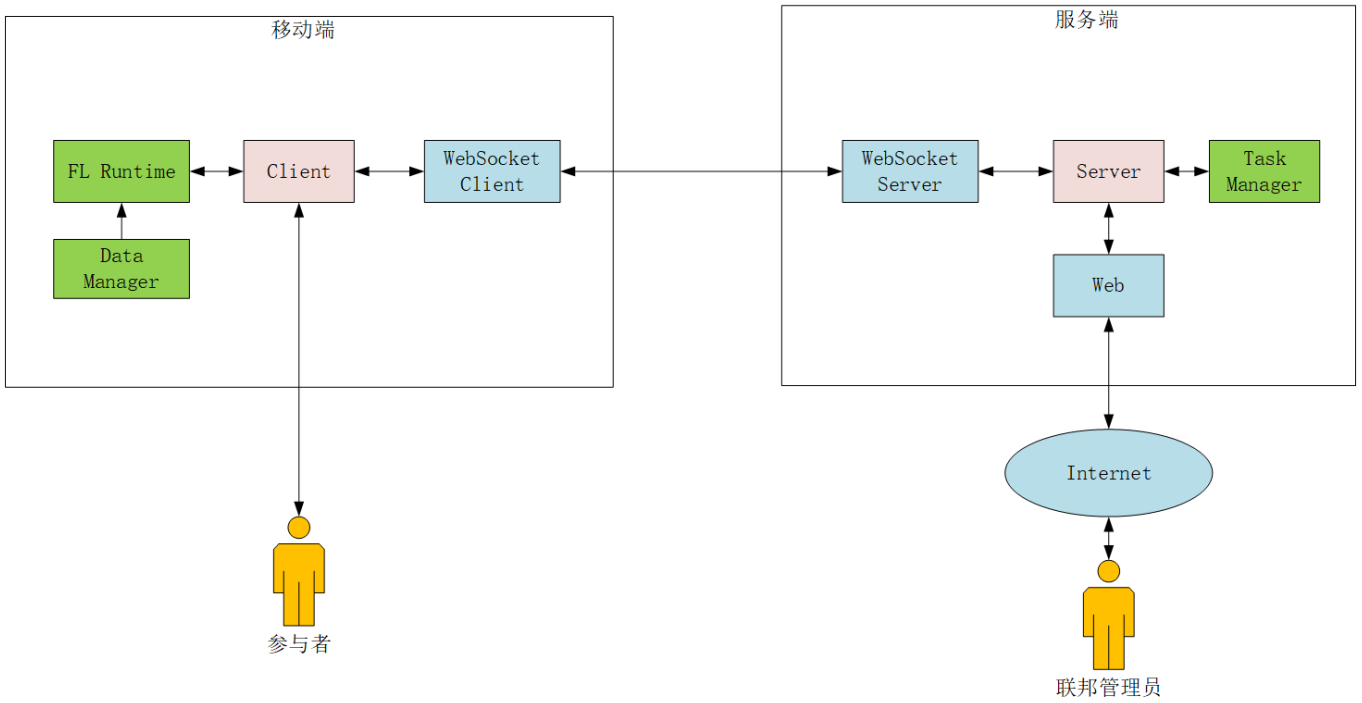
2.参与方对全局模型的贡献
- 更新模型的精度

参与方角度（成本）

1. 数据成本
2. 算力成本
3. 存储成本
4. 存在的风险
   - 隐私泄露
   - 行业竞争（企业）
   - 硬件折旧
   - …

公平性

我们已经开展的工作

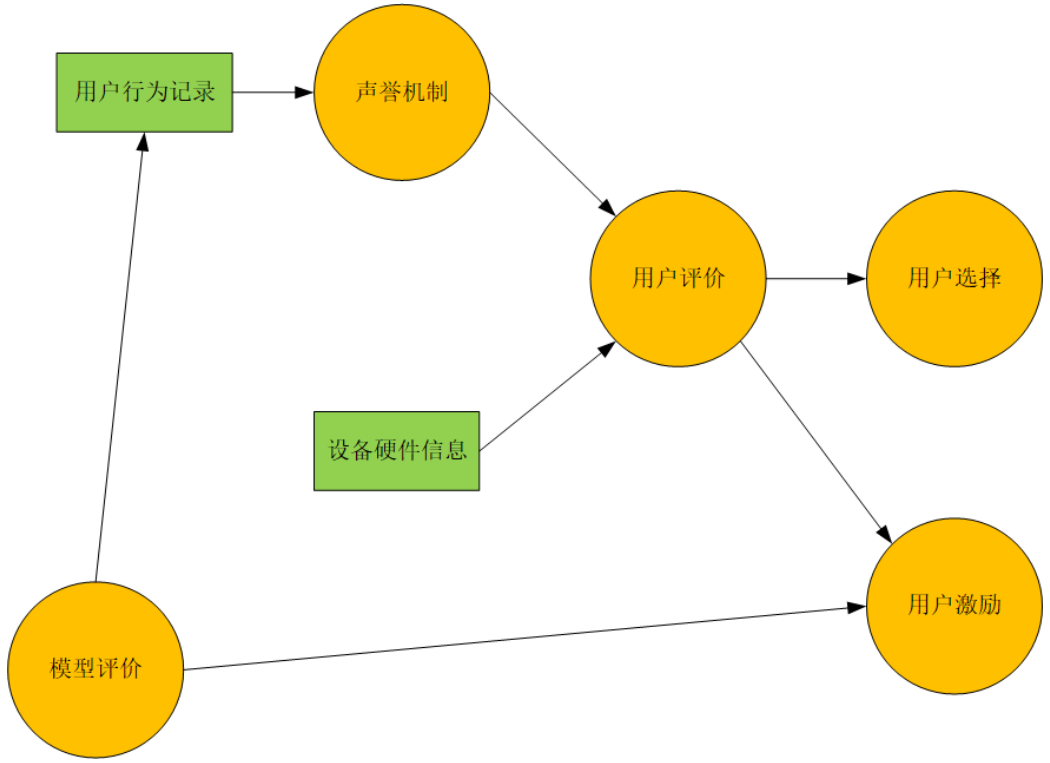


设计并实现了一个初步的联邦学习框架

- 联邦学习的类别为横向联邦
- 通信架构为集中式结构
- 通信方式为同步通信

项目仓库
https://github.com/zhanghad/Federated_Learning

我们关注的问题



初步的设计方案

- 用户选择
- 激励机制
- 通信效率

项目仓库
https://github.com/zhanghad/Federated_Learning