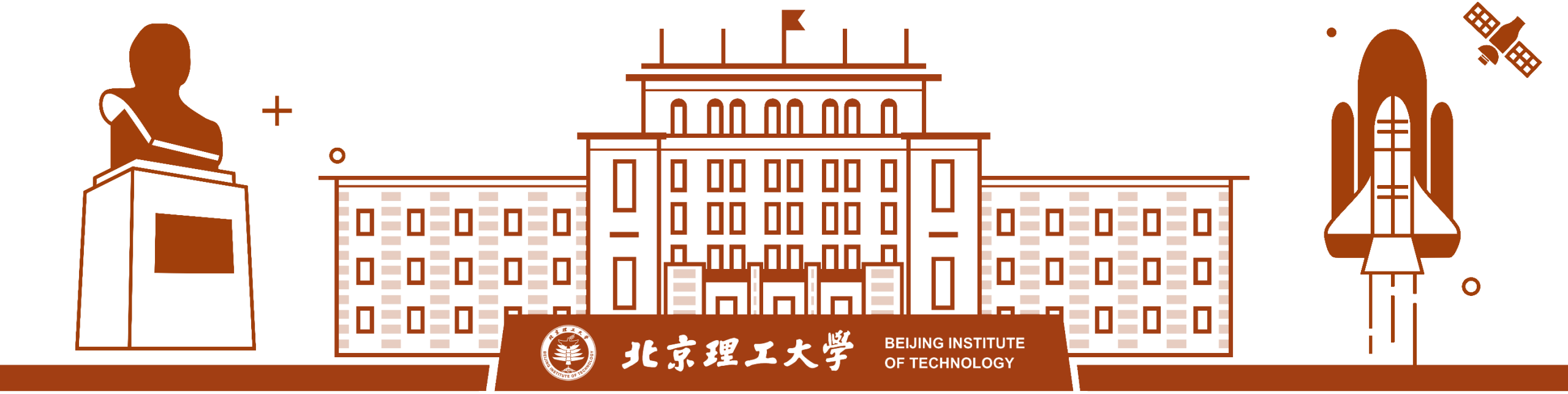


联邦学习

邱万勇

指导老师：胡斌、钱昆

医学技术学院、计算机学院

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

目录

CONTENTS

“What Makes it Appear？”

FML

AI 的发展与挑战

1、数据现状

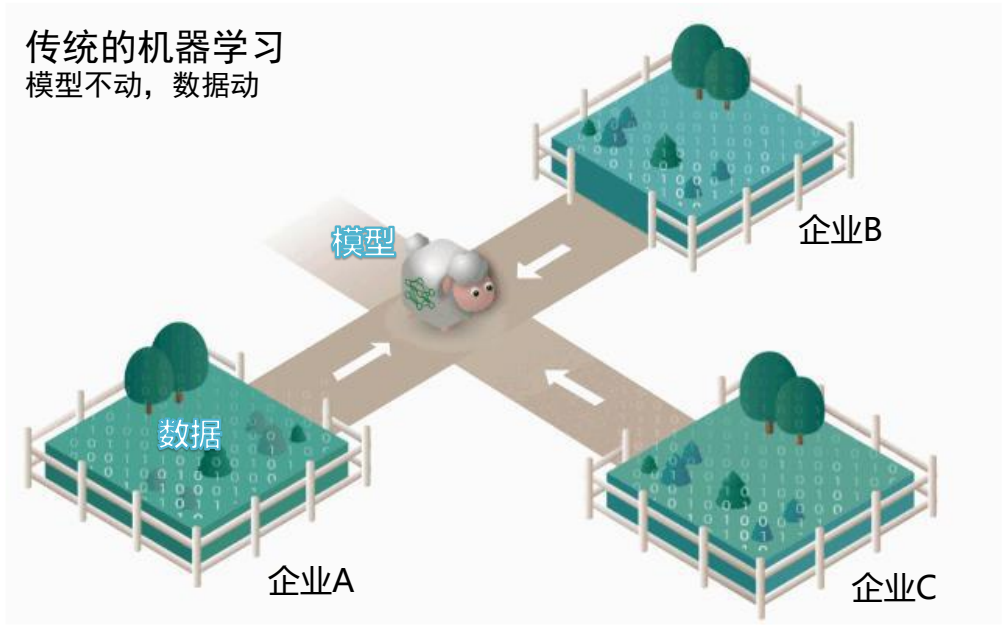
2、安全&隐私保护

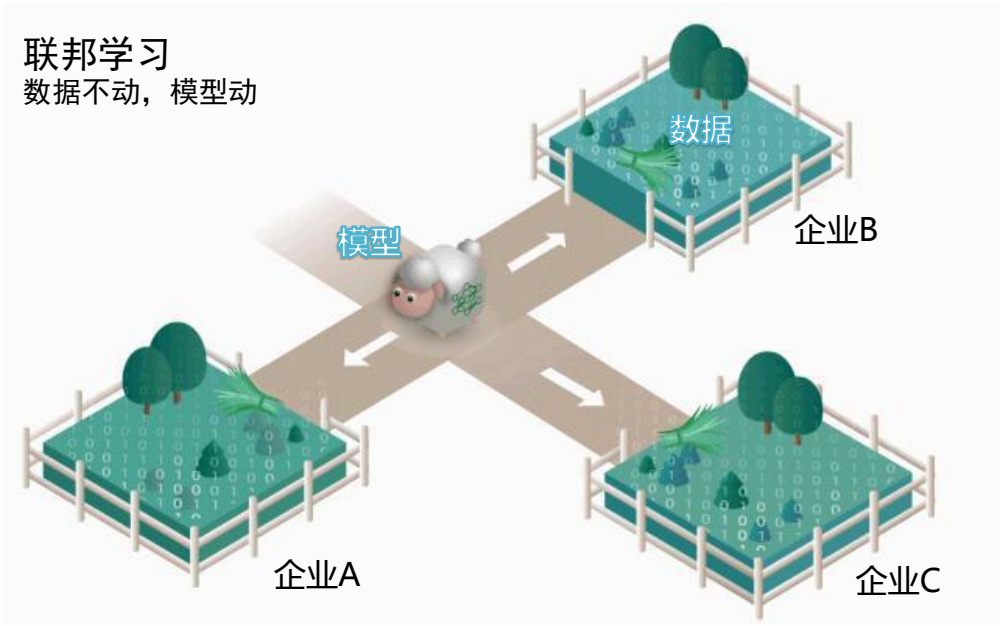
数据

- 数据孤岛
- 数据分布不均
- 缺少标注数据
- 隐私法律法规
- 数据安全
- 。 。 。

鱼（AI计算）和熊掌（安全）可以兼得？

联邦学习与传统机器学习





Source: Federated Learning(Synthesis Lectures on Artificial Intelligence and Machine Learning) Qiangyang, et al.

数据不动 **模型动**，数据可用 **不可见。**

Definition

本质

联邦学习本质上是一种**分布式的**机器学习**框架**。

目标

使多个参与方在保护数据隐私、满足合法合规要求的前提下继续进行机器学习，解决数据孤岛问题。

Definition

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

定义如下：

在进行机器学习的过程中，各参与方可借助其他方数据进行联合建模。各方无需共享数据资源，即数据不出本地的情况下，进行数据联合训练，建立共享的机器学习模型。

Features

联邦学习是：

A 各方数据保留在本地，不泄露隐私也不违反法规

B 多个参与者联合数据建立虚拟的共有模型，并且共同获益的体系

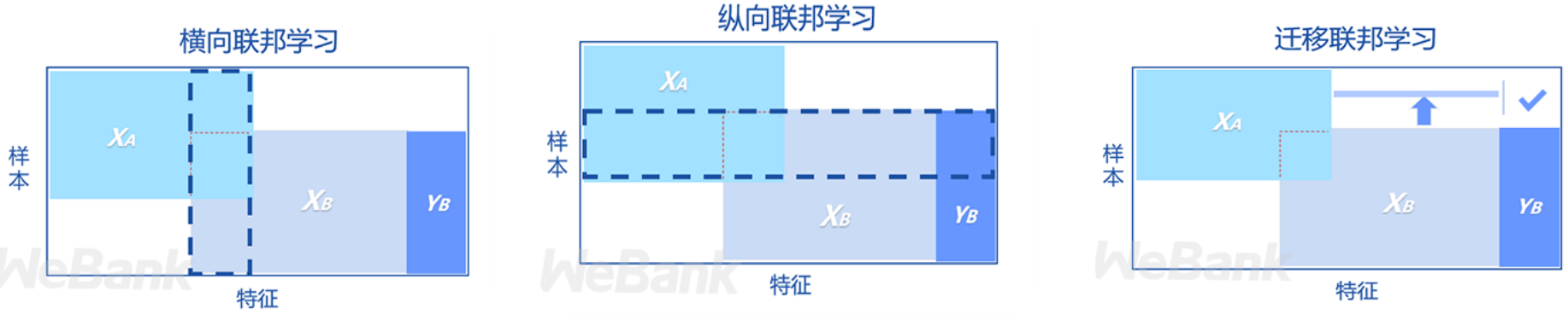
C 在联邦学习的体系下，参与者的身份和地位平等

D 联邦学习的建模效果和将整个数据集放在一处建模的效果相同，或相差不大

E 迁移学习是在用户或特征不对齐的情况下，也能在数据间通过交换加密参数达到知识迁移的效果

Categorization

联邦学习主要采用了同态加密的技术来保护多方数据隐私，可以根据用户的特征维度和样本ID的重叠度分为横向联邦学习和纵向联邦学习，亦或在两者重叠度都很低的情况下采用联邦迁移学习。



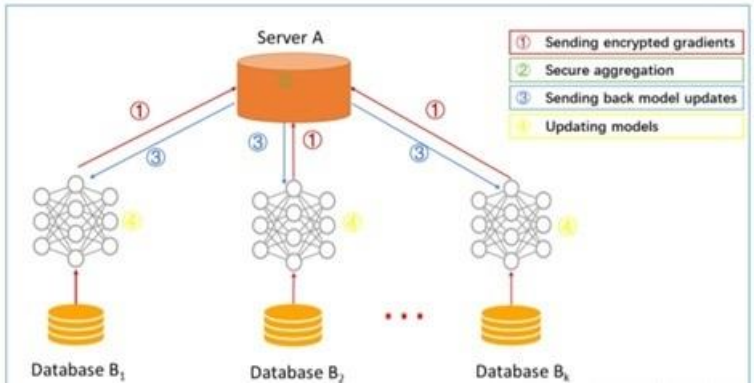
The overlap of features (X1, X2, …) is large, whereas the overlap of users (U1, U2, …) is small

The overlap of users (U1, U2, …) is large, whereas the overlap of features (X1, X2, …) is small

The overlap of users (U1, U2, …) and the overlap of features (X1, X2, …) are both small

Framework

横向联邦学习的系统构架



Step1

参与方从服务端获取训练模型，利用本地数据训练，并将加密梯度上传至服务端。

Step2

安全聚合模型参数并更新模型。

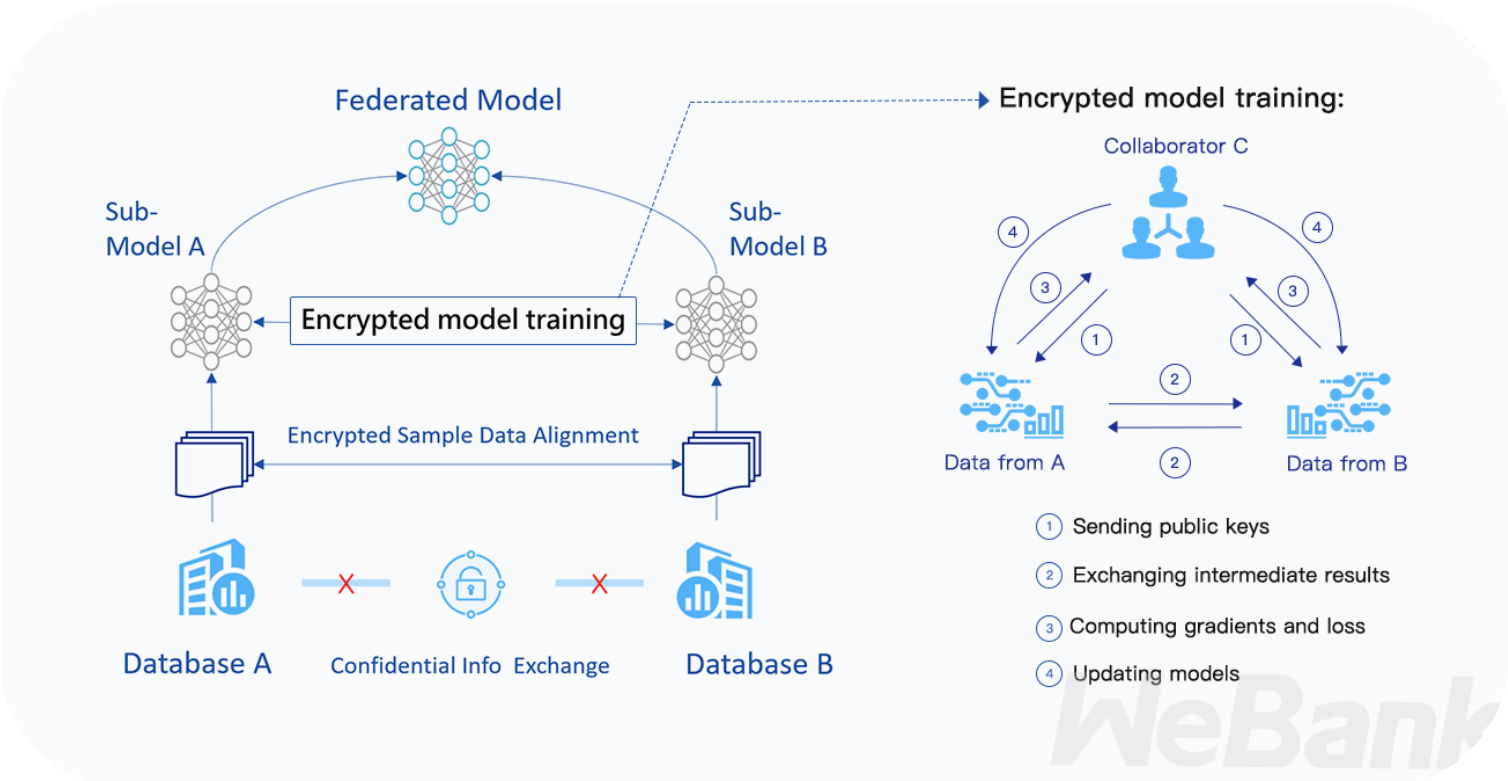
Step3

返回更新后的模型给参与方。

Step4

各参与方更新各自模型。

纵向联邦学习的系统构架（两个数据方：企业A和B）



Framework

Course of work:

加密样本对齐

由于两家企业的用户群体并非完全重合，系统利用基于加密的用户样本对齐技术，在A和B不公开各自数据的前提下确认双方的共有用户，并且不暴露不互相重叠的用户。以便联合这些用户的特征进行建模。

加密模型训练

在确定共有用户群体后，就可以利用这些数据训练机器学习模型。为了保证训练过程中数据的保密性，需要借助第三方协作者 C 进行加密训练。

效果激励

联邦学习的一大特点就是它解决了为什么不同机构要加入联邦共同建模的问题，即建立模型以后模型的效果会在实际应用中表现出来，并记录在永久数据记录机制(如区块链)上。

Framework

Encrypted model training:

协作者C把公钥分发给A和B，用以对训练过程中需要交换的数据进行加密；

A和B之间以加密形式交互用于计算梯度的中间结果；

A和B分别基于加密的梯度值进行计算，同时B根据其标签数据计算损失，并把这些结果汇总给C。C通过汇总结果计算总梯度并将其解密；

C将解密后的梯度分别回传给A和B，A和B根据梯度更新各自模型的参数。

迭代上述步骤直至损失函数收敛，这样就完成了整个训练过程。

Framework

The difference between Federated Learning and Differential Privacy:

The difference between Federated Learning and Distributed Machine Learning:

	工作节点	数据的调配	数据自主性
分布式机器学习	数据存储位置	一个中心式的调度节点	无
联邦学习	模型训练的数据拥有方	模型训练的数据拥有方	较好

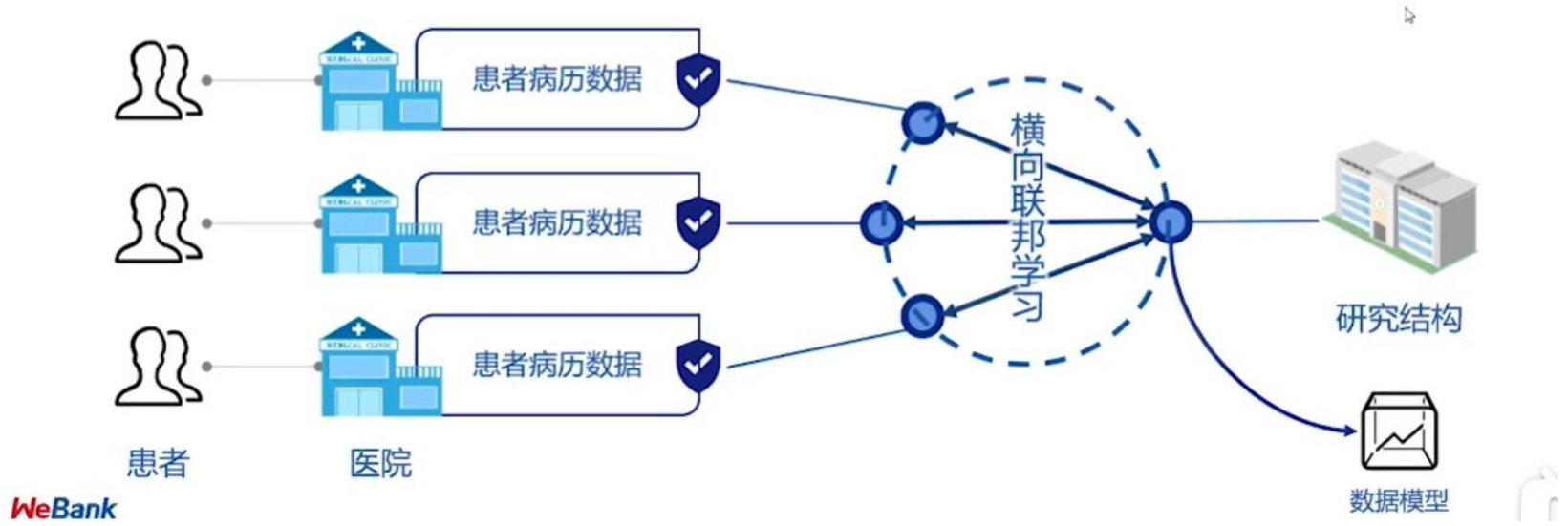
横向联邦解决医疗大数据痛点

■ 医疗大数据应用痛点：

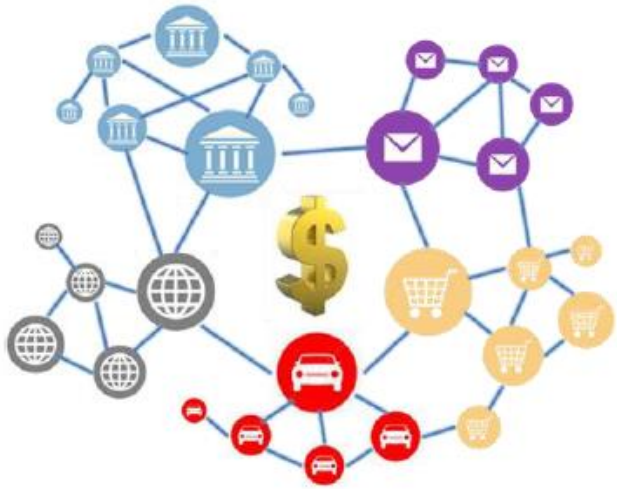
1、医疗数据高度隐私，数据维护方对患者数据管理严格、使用谨慎。
2、数据分散，单一组织缺乏足够多的可用样本。

■ 横向联邦天然适合医疗大数据场景：

1、数据 **安全共享** 机制，有效保护用户隐私
2、安全 **连接** 分散的数据源，共建数据模型
3、安全联合建模效果几乎 **无损**



联邦学习的展望



中国数据保护法律概况

- 《个人信息保护法》与数据保护
- 《数据安全法》与数据保护
- 《网络安全法》与数据保护

感谢观看

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

邱万勇

计算机学院、医学技术学院