

FATE：工业级联邦学习开源生态建设经验分享

范涛

微众银行联邦学习研发负责人

人工智能四级技术专家



个人简介-范涛

- 微众银行联邦学习研发负责人，人工智能4级技术专家，具备8年以上大规模机器学习系统和大数据相关应用实践经验。
- 在微众银行负责联邦学习FATE开源项目研发和FATE商业化产品研发，建立建成全球最活跃的联邦学习开源技术社区，推动联邦学习技术在风控，营销，个性化推荐等领域应用。申请联邦学习相关技术专利40多项和发表多篇有影响力学术论文(SecureBoost算法获得联邦学习领域论文引用量Top10)
- 2013年硕士毕业于中国科学与技术大学，加入微众银行前曾任职于腾讯，百度，负责智能风控，大数据挖掘，舆情分析，大数据量化投资等项目研发

open mind feeli
open/free shar
open company source cod
free software git global
github **communit**
social coding chat speak
communication cloud com
program language mobile pl
technology Big D
information technology
wireless network
AI

01 行业背景介绍

数据新基建

computing power

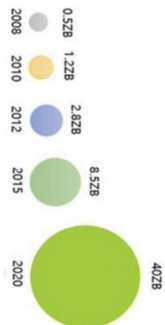
big data

1 ZB= 10^{21} Byte



Intel i386
Intel i486
Intel Pentium
Intel Core
nVidia GPU
Google TPU

---来自互联网数据中心 (IDC)

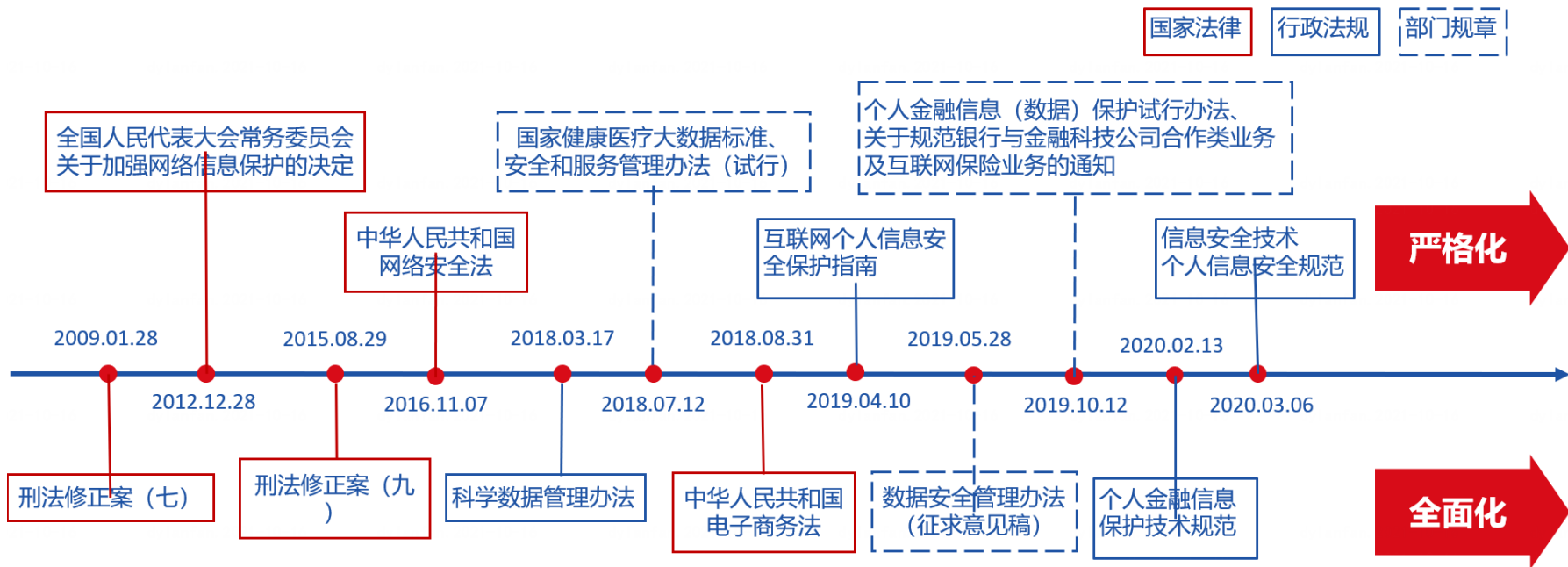


数据资源如同新的石油



“十九届四中全会：党中央首次提出将数据作为生产要素参与收益分配”

监管趋严



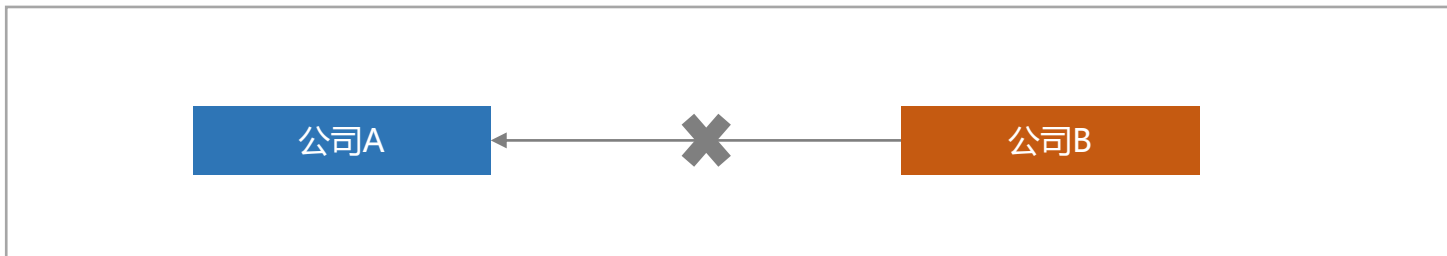
严格化：数据监管越来越严格，处罚手段越来越严厉。

全面化：从个人信息数据的保护，到科学数据、医疗数据、电商数据等多种数据的保护。

密集化：从整体上看，数据监管的法律法规出台会密集化。

数据合作困境

跨机构间数据合作受阻



机构内跨部门间数据中台建立困难重重

各个部门知道数据的价值，寻求和其他部分进行数据合作，但都又不愿意泄露自己业务核心数据；



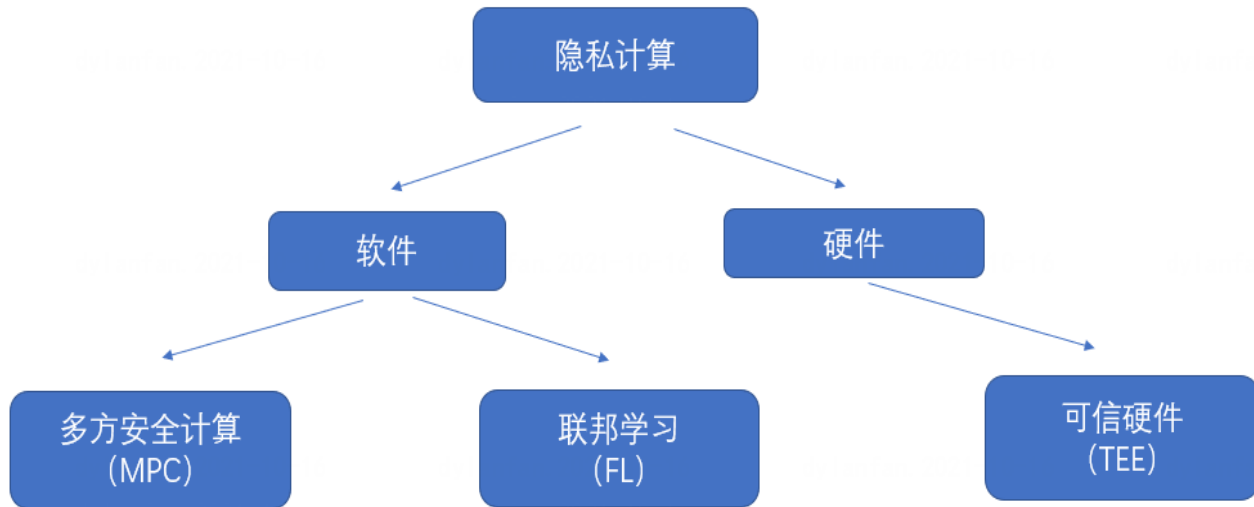
如何破局



02 联邦学习一站式解决方案

隐私计算技术体系

- ✓ 数据“**所有权**”和“**使用权**”分离
- ✓ 数据**可用不可见**
- ✓ 用途**可控可度量**

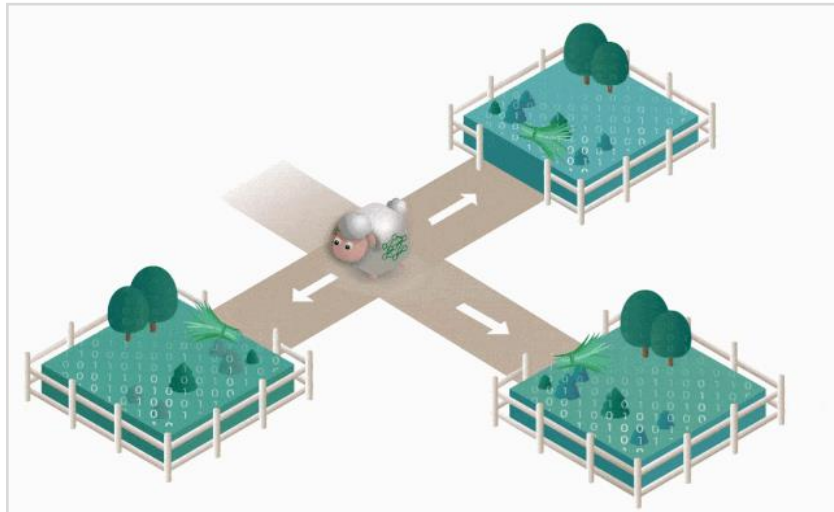
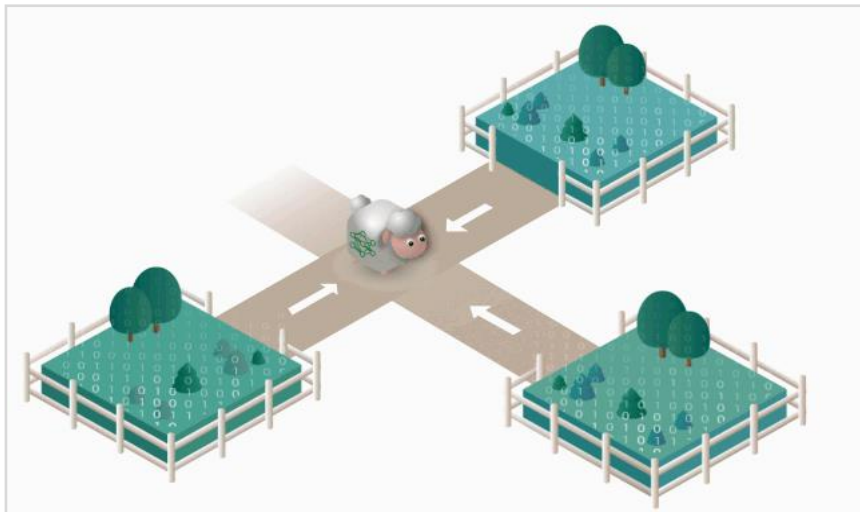


联邦学习：数据不动模型动，数据可用不可见

将草从各地集中到一起喂羊，并不合规
隐私和数据安全保护的要求使得获取数
据成为障碍



联邦学习提供了新思路：
让羊群在各地移动，而草不出本地，主
人无法知道它吃了哪些草



纵向联邦学习-联合建模场景

举例：企业A与企业B联合建模，企业B有Y（业务表现），期望优化本方的Y预测模型

◆ 设定：

- ✓ 只有企业B拥有 Y= “逾期表现”
- ✓ 企业A无法暴露含有隐私的 X

◆ 传统建模方法问题：

- ✓ 企业A缺乏Y无法独立建模
- ✓ 企业A的X数据全量传输到企业B
不可行

◆ 期望结果：

- ✓ 保护隐私条件下，建立联合模型
- ✓ 联合模型效果超过单边数据建模

企业A

ID 证件号 电话号	X1 帐龄	X2 月薪	X3 等级
U1	9	8000	A
U2	4	5000	C
U3	2	3500	C
U4	10	10000	A
U5	5	7500	B
U6	5	7500	A
U7	8	8000	B

业务系统A 数据

企业B

ID 证件号 电话号	X4 评分1	X5 评分2	Y 表现数据
U1	600	600	无
U2	550	500	有
U3	520	500	有
U4	600	600	无
U8	600	600	无
U9	520	500	有
U10	600	600	无

业务系统B 数据

横向联邦学习-联合建模场景

举例：企业A和企业B共建联合模型，期望优化联合模型

◆ 设定：

- ✓ Y 表示 “是否存在恶意行为”
- ✓ 企业A和企业B都有 (X,Y)
- ✓ 双方不暴露自己的 (X,Y)

◆ 传统建模方法问题：

- ✓ 企业A和企业B各自样本不够多

◆ 期望结果：

- ✓ 保护隐私条件下，建立联合模型
- ✓ 联合模型效果超过单边数据建模

企业A

ID 证件号 电话号	X1	X2	Y 表现数据
U1	5	15	有
U2	8	20	有
U3	0	5	无
U4	0	0	无
U5	2	1	无
U6	50	50	有
U7	60	6	有

业务系统A 数据

企业B

ID 证件号 电话号	X1	X2	Y 表现数据
U8	5	10	有
U9	10	2	有
U10	2	30	有
U11	0	10	有
U12	8	7	有

业务系统B 数据

FATE: 联邦学习一站式解决方案



企业解决方案层

FATE-Studio
面向企业开发者提供零门槛联邦学习开发平台
集交互式联邦建模, 联邦查询统计, 数据管理, 模型部署
为一体解决方案

FATE-Cloud
面向企业开发者提供联邦数据合作网络搭建平台
集联邦站点注册, 站点监控, 站点集群可视化部署, 合约管理,
交易管理为一体解决方案

核心应用组件层

联邦区块链 FATE-Chain

身份认证/可信授权

日志协作/审计

数据/模型激励追踪

联邦查询统计 FATE-SQL

联邦SQL解析器

横向/纵向
查询统计算子

查询安全审计

联邦建模可视化 FATE-Board

联邦模型可视化

联邦任务
dashboard

任务/日志管理

联邦建模调度 FATE-Flow

多方任务协同调度

联邦任务生命周期管理

联邦模型管理

联邦在线推理 FATE-Serving

实时在线联邦推理

集群管理与监控

在线模型管理

联邦学习算法库 FederatedML

纵向联邦特征工程

纵向联邦学习

横向联邦学习

联邦深度学习

联邦迁移学习

纵向联邦统计

安全信息检索
(PIR)

安全求交 (PSI)

横纵融合

异步联邦学习

模型加密预测

联邦安全协议 Secure Protocols

Paillier同态加密

仿射同态加密

Secret-Sharing
(SPDZ)

OT

可交换加密

安全聚合

RSA

DH密钥交换

计算: [Tensorflow](#) / [Pytorch](#) (深度学习)
[EggRoll](#) / [Spark](#) (分布式计算框架)

多方联邦通信: 跨站点传输网络
([RollSite](#)/[Pulsar](#)/[RabbitMQ](#))

存储: [HDFS](#)/[HIVE](#)/[MYSQL](#)/[LocalFS](#)

核心框架层

FATE

- ✓ FATE是微众银行人工智能团队发起的全球首个联邦学习工业级开源框架，可以让企业和机构在保护数据安全和数据隐私的前提下进行数据协作
- ✓ FATE于2019年2月首次对外开源，并于2019年6月捐献给Linux基金会，并成立FATE TSC对FATE社区进行开源治理，成员包含国内主要云计算和金融服务企业
- ✓ 核心功能包括联邦特征工程，联邦统计，联邦机器学习，联邦深度学习，联邦迁移学习等

GitHub: <https://github.com/FederatedAI/>

FATE开源治理

【FATE社区概况】

570+ 家企业机构, 350+ 所高校

8个FATE社群3000+人, 3400+ GitHub Star

如涉及到公众号转发的白名单等事宜, 需要与信通院沟通确认的, 随时沟通。

序号	名称	时间	所属集团	可信执行环境
13	KubeTEE	2020年9月	蚂蚁集团	

从开源项目的活跃度和影响力来看, 联邦学习的开源生态为工业化的落地应用贡献了强劲力量, 特别是 FATE, 2020 年及之后出现的很多联邦学习类产品都或多或少的吸收和借鉴了 FATE 供给的营养。

在中国信通院调研统计中, 55%的国内隐私计算产品是基于或参考开源项目开发的, 这其中开源项目就以 FATE 为主。

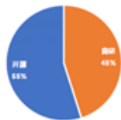


图 10 国内隐私计算平台自研情况

【TSC (技术管理委员会) 成员】



Tencent



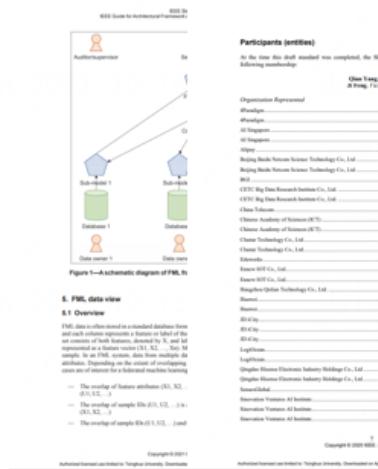
信通院《隐私计算白皮书 (2021年)》发布, 根据白皮书中国信通院的调研, 55%的国内隐私计算产品是基于或者参考开源项目开发, 这其中开源项目就以FATE为主。

微众银行联邦学习产品标准建设情况

【国际标准】

发布全球第一个联邦学习相关国际标准

IEEE P3652.1 《IEEE Guide for Architectural Framework and Application of Federated Machine Learning》



IEEE SA

IEEE Guide for Architectural Framework and Application of Federated Machine Learning

IEEE Computer Society

Developed by the Learning Technology Standards Committee

IEEE Std 3652.1-2020



STANDARDS

【国内标准】

参与编写已发布标准：

- 参与信通院《基于多方安全计算的数据流通产品技术要求与测试方法》和《联邦学习技术与应用》标准编写
- 金融行业标准：参与央行金融标准化委员会《多方安全计算金融应用技术规划》的标准

参与编写中的标准：

- 金融行业标准《联邦学习金融应用与互联互通标准规范》
- 通信行业标准 (CCSA-TC1/TF1)：《联邦学习的安全评测技术要求及测试方法》《联邦学习跨框架互操作技术要求》
- 团体标准 (CCSA-T601)：《联邦学习跨平台互联互通标准》

微众银行联邦学习产品安全认证

- 系统通过《信息安全等级保护》三级备案
- 通过中国信通院《大数据·联邦学习数据流通产品》、《大数据·多方安全计算数据流通产品》、《联邦学习评估专项》认证
- 完成国家金融科技评测中心（银行卡检测中心）多方安全计算金融应用技术测评



《信息安全等级保护》三级备案证书



信通院《大数据·联邦学习数据流通产品》认证



信通院《大数据·多方安全计算数据流通产品》认证



信通院《联邦学习评估专项》

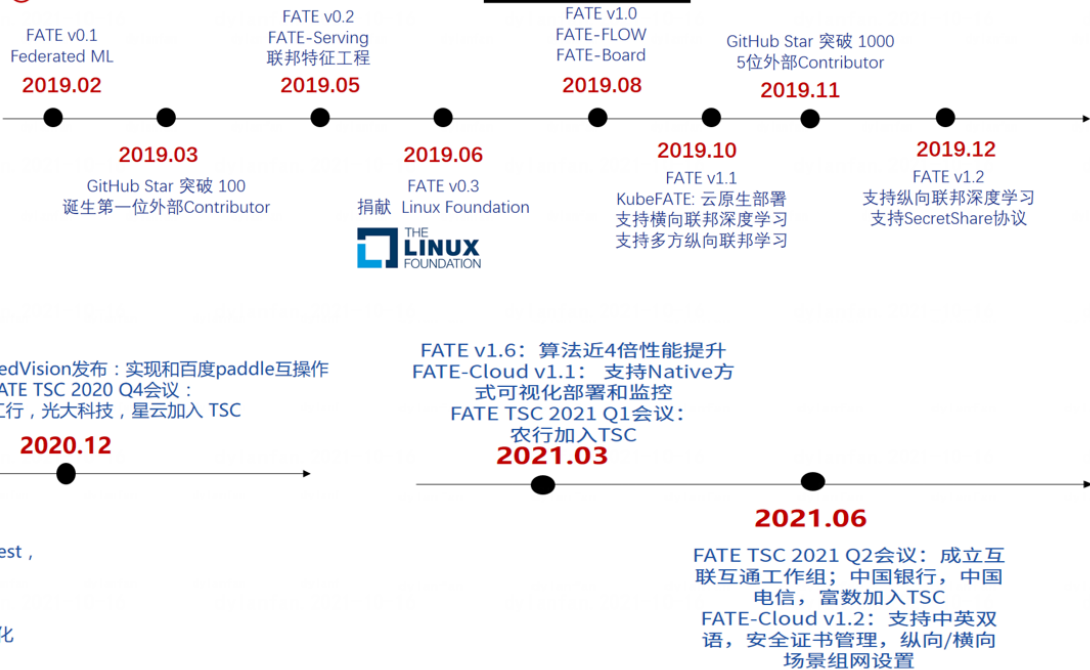


完成银行卡检测中心 (BCTC) 评测

FATE里程碑 2019-2021

- 2019：开源社区初创，功能丰富阶段
- 2020：开源社区生态快速发展阶段
- 2021：企业级产品和标准快速推进阶段

Milestone

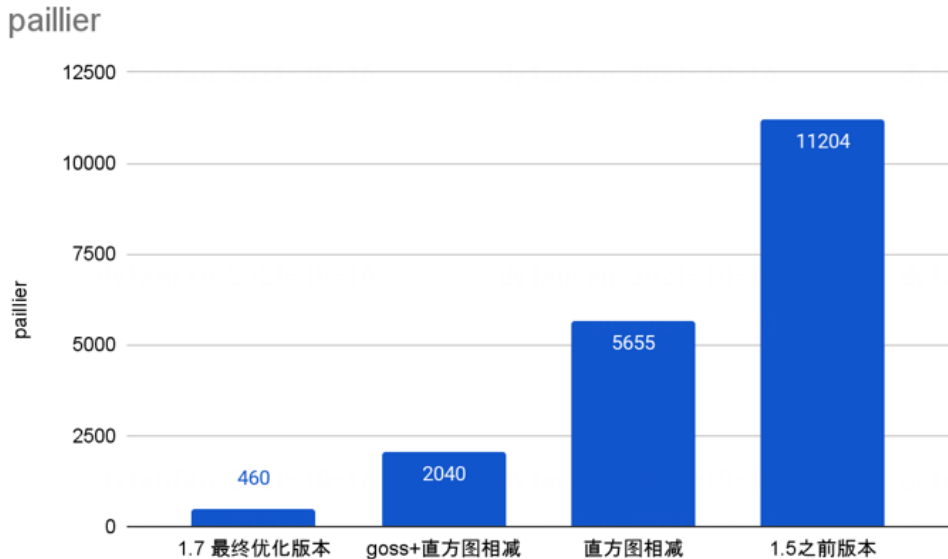


FATE-ML: 机器学习遇见安全计算



- ✓ 横向联邦: 传统机器学习算法可以无缝接入安全协议, 低成本实现联邦机制
- ✓ 纵向联邦: 满足高性能, 可用性, 机器学习实现联邦机制需定制联邦协议

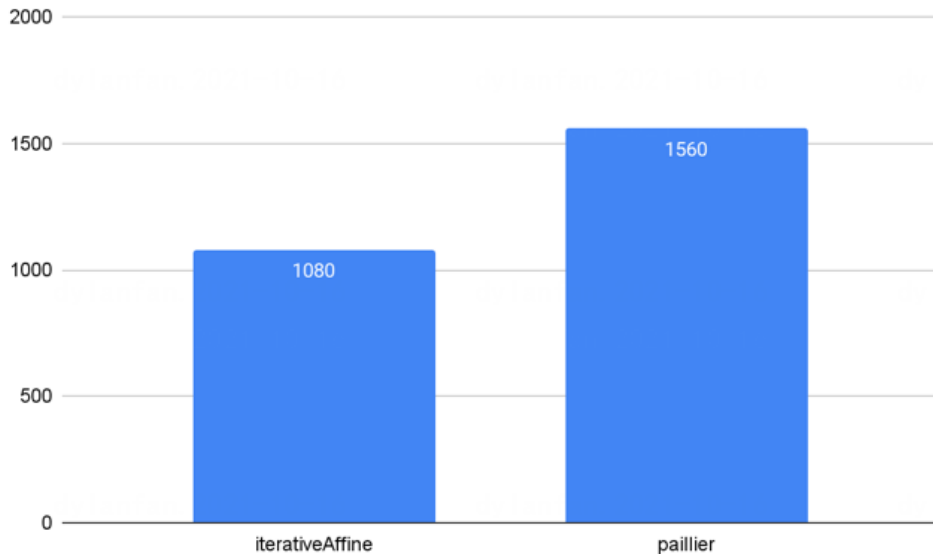
Paillier下持续优化
获得**24.6倍**提升



*统计运行稳定时构建一颗决策树的时间，40W，2000维度

千万级别样本SecureBoost性能

Hetero-SBT支持千万级别(百维特征)
样本的训练



*统计运行稳定时构建一颗决策树的时间, 1000W, 100维度

FATE Board: 联邦建模可视化和可解释性

任务/模型可视化



secureboost: 森林树模型可视化、分裂特征重要行、模型迭代评估



纵向特征分箱: 计算包含woe、iv值的特征分箱详情表、分箱直方图



模型评估: 支持常用二分类、多分类、回归等指标



仪表盘



特征选择: 各筛选指标下的特征过滤情况



纵向特征相关性: 查看特征之间的Pearson相关性系数



FATE-Serving: 在线模型服务与监控

FATE Serving为FATE提供联邦在线推理服务，主要包含实时在线预测、集群管理与监控、在线模型管理与监控、服务治理等功能；

核心优势

- ⚡ 实时预测极速响应
- 🔄 多方联合并行推理
- ⚙️ 基于模型的服务治理
- 📊 高可用高性能
- 📈 资源实时监控
- 🛡️ 生产级服务保护

在线预测

并行预测

批量预测

多方预测

在线模型管理与监控

模型发布

模型卸载

模型绑定/解绑

模型监控

在线服务管理与监控

服务监控

服务管理

集群管理与监控

进程管理

配置管理

流量监控

JVM监控

服务治理

负载均衡

灰度发布

服务注册与发现

限流

在线模型管理与监控：模型卸载、模型与服务解绑、模型调用量监控等



集群模型管理与监控：流量指标监控



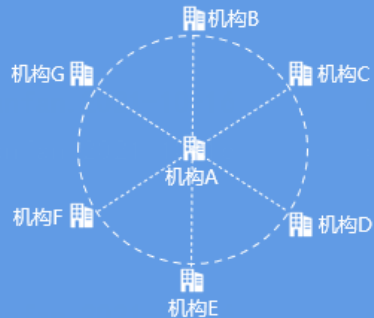
FATE-Cloud: 联邦多云管理

FATE Cloud是构建和管理联邦数据合作网络的基础设施，为跨机构间、机构内部不同组织间提供了安全可靠、合规的数据合作网络构建解决方案，实现多客户端的云端管理。

数据合作网络模式

联合合作模式

合作机构间形成数据合作联盟，发挥多方数据价值



集团平台模式

集团子公司拥有相对独立数据源，进行数据合作与数据协同治理



核心优势



企业级安全合规



多云管理

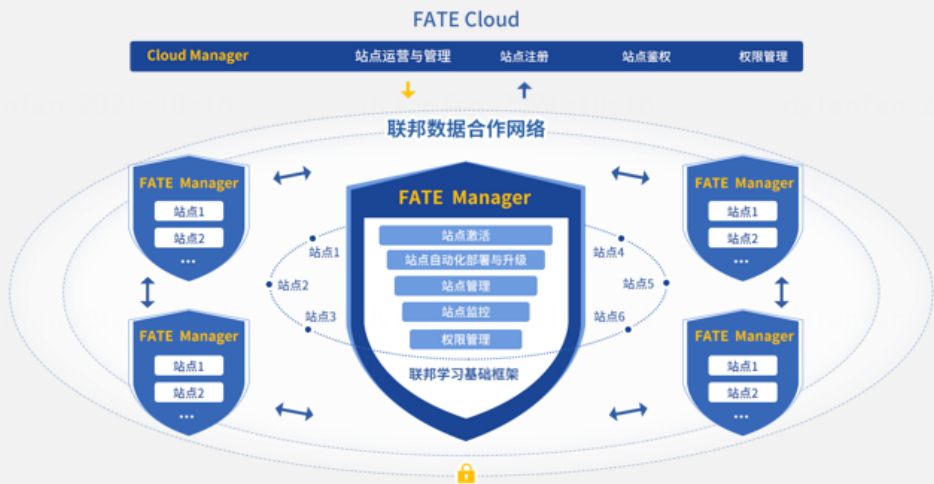


数据网络
灵活可扩展



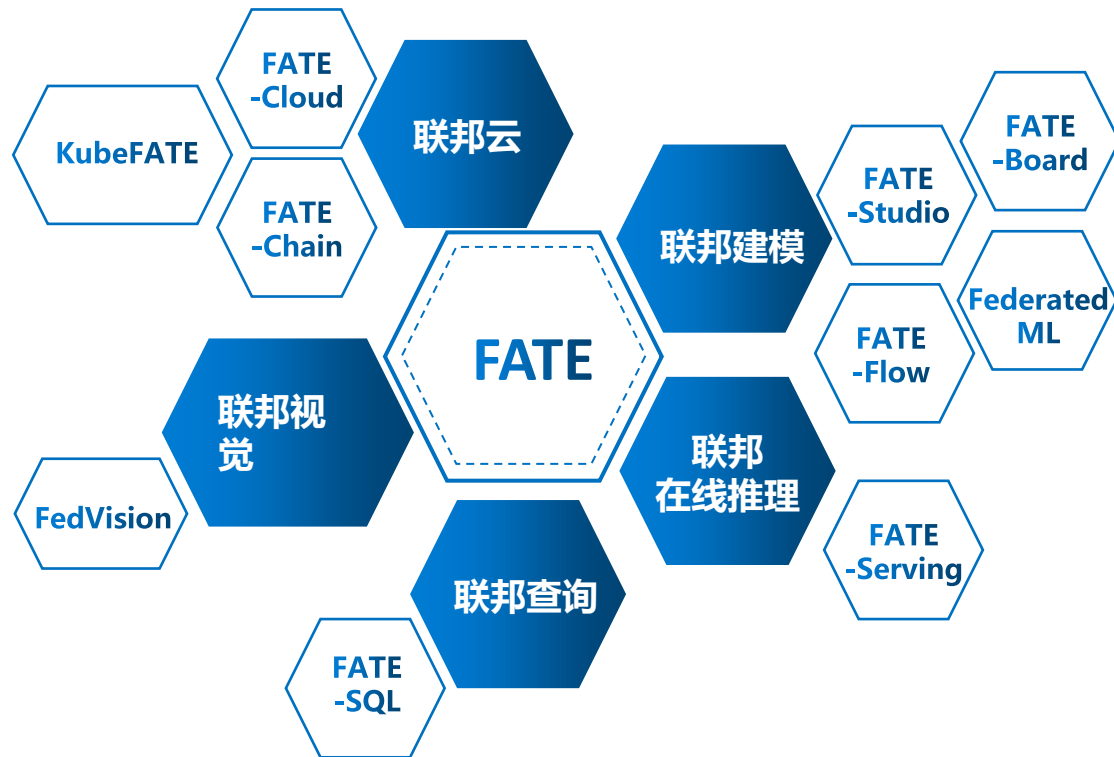
轻便的资源管理

FATE Cloud由负责联邦站点管理的云管理端Cloud Manager和站点客户端管理端FATE Manager组成，核心功能如下：



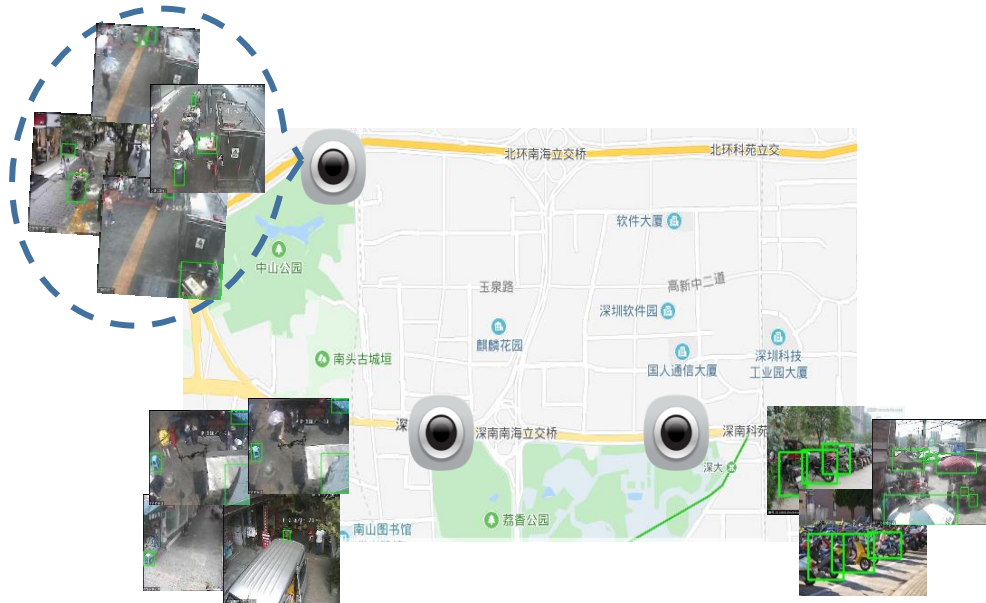
FATE 社区 2021 蓝图

致力于建设FATE成为联邦学习标准协议



03 应用案例

*微众银行 AI 联合极视角 Extreme Vision 项目



挑战

- 标签数量少
- 数据分散，集中管理成本高
- 离线延迟的模型更新和反馈

联邦学习

- 在线模型更新和反馈
- 无需集中上传数据
- 数据保护，隐私性高

多机构联合脑卒中预测

* 微众银行与腾讯天行实验室共同将联邦学习与医疗深度融合，脑卒中预测准确率达80%

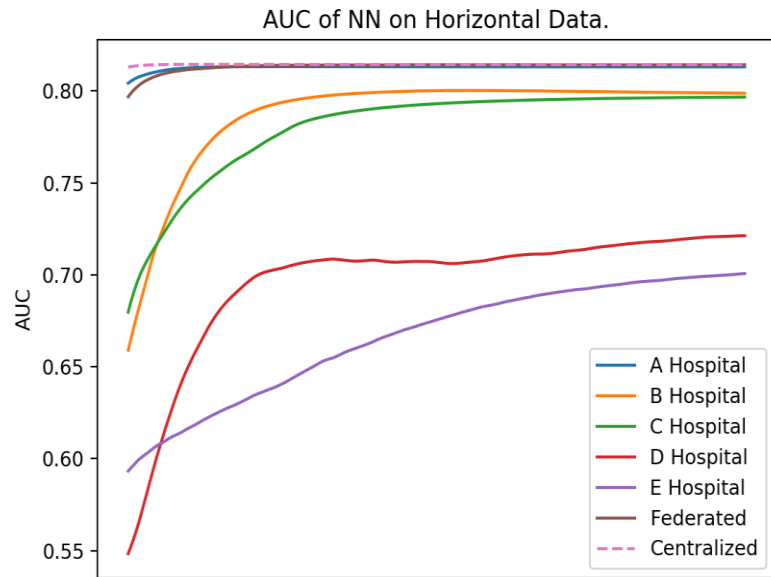
■ 联邦学习建立脑卒中患病概率预测模型

- ✓ 三家三甲医院+两家小医院
- ✓ 病患住院流程数据和体征数据

■ 效果

- ✓ 基于联邦学习的联合模型效果**优于**任意一家医院数据独立建模效果
- ✓ 联邦学习训练所得模型效果与集中数据训练所得模型效果**差异甚微**

WeBank
微众·AI



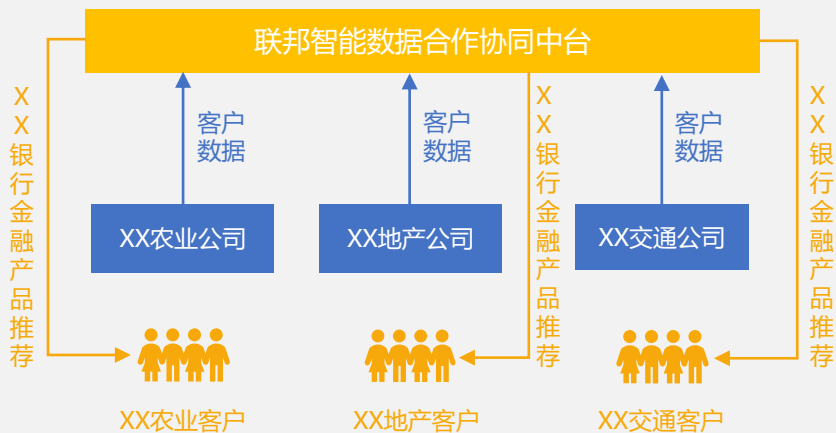
	A医院	B医院	C医院	D医院	E医院	联邦	集中
AUC	0.813 ±0.027	0.799 ±0.062	0.797 ±0.067	0.720 ±0.035	0.701 ±0.026	0.814 ±0.030	0.814 ±0.014

注:表格中, A/B/C医院是全市排名前三的大型三甲医院, D/E是小医院。实验结果格式为(中值±方差), 同一组实验, 我们测试5次取平均结果。

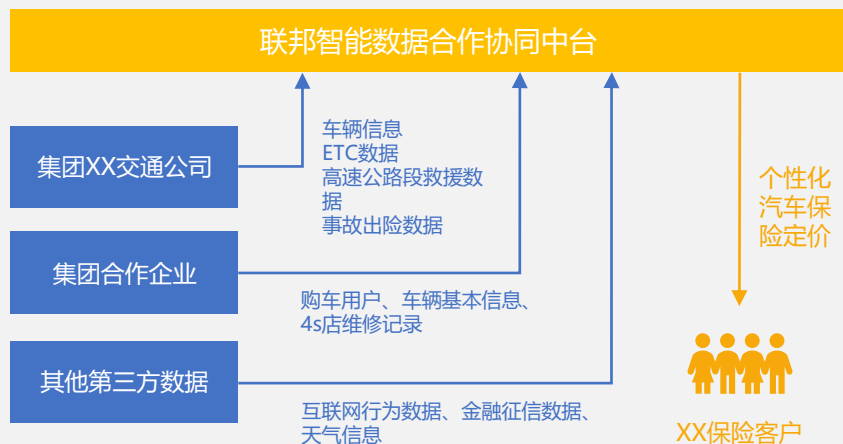
基于FATE为建设联邦智能协同中台

通过联邦智能数据合作协同中台，集团内部不同业务子公司之间进行大数据合作，准确地分析集团现有客户，更有效地利用集团内部客户资源发挥不同业务板块的数据价值，通过联邦建模建设更有力的推荐服务，实现精准的交叉营销；

场景1：基于联邦学习整合分析不同板块的客户数据（集团XX农业公司、集团XX地产公司、集团XX交通公司），通过智能推荐实现集团XX银行金融产品更精准的交叉营销



场景2：通过对集团XX交通公司、集团合作企业以及其他第三方平台的数据的联合建模，可以改进集团XX保险公司的产品。通过个性化保险定价，让优质的客户可以通过更合理的价格获得保险产品，从而为客户提供更好的服务





欢迎来GitHub 加入FATE建设
star我们，第一时间接收项目进展
官网：<https://www.fedai.org/>
邮箱：contact@fedai.org



国内首个联邦学习官方社区，这里有

- 高价值贡献者激励计划
- **10+** 顶尖算法工程师实时答疑解惑
- **超500**家企业机构开发者共同交流学习
- 国内最新联邦学习产品资讯抢先获取


THANK YOU

QUESTIONS?

 bestPresentation

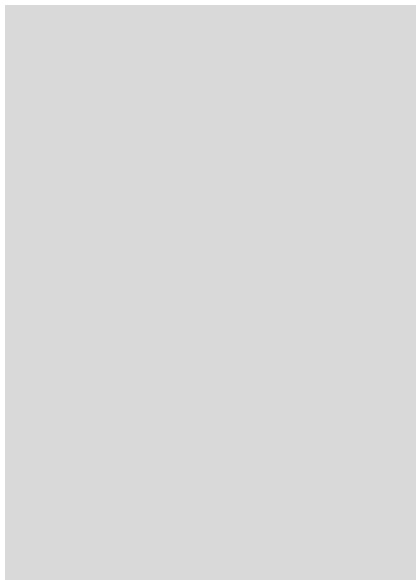
 @bestPresentation

 bestPowerPoint

 teamBest

 Best-Presentation

 bestPowerPoint



扫码关注
开源社公众号



2021 中国开源年会

HAPPY HACKING

