Project Proposal for LF AI & Data

OpenFL: Open Federated Learning

Olga Perepelkina, Intel
Patrick Foley, Intel
Prashant Shah, Intel

Spyridon Bakas, UPenn
Henry Zhang, VMware
Zhen Xiao, Leidos
Eric A. Stahlberg, Leidos
Daniel J. Beutel, ADAP

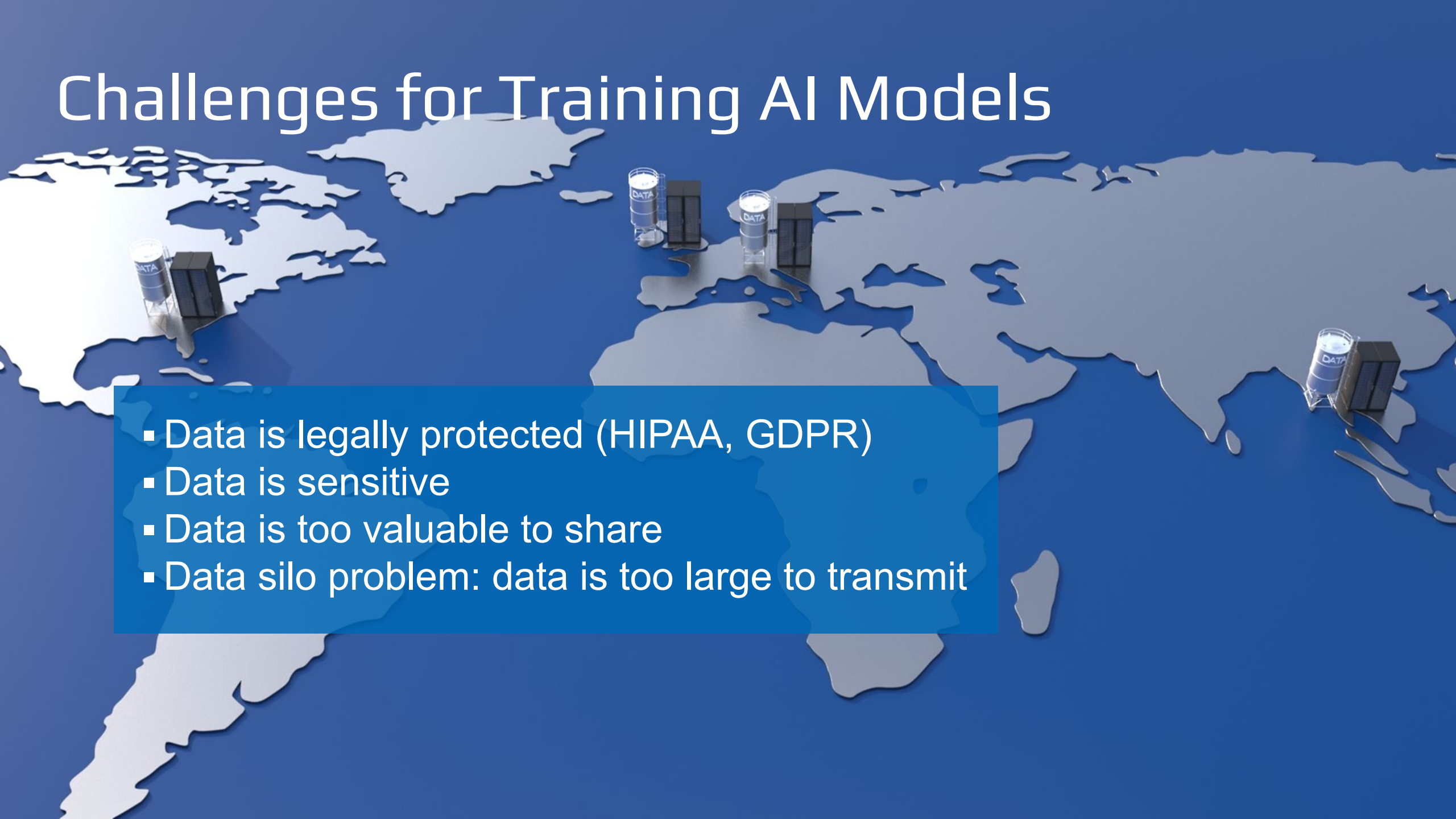
Agenda

1. What is Federated Learning
2. What is OpenFL
3. Collaborations and community
4. Why contribute OpenFL to Linux Foundation?

Agenda

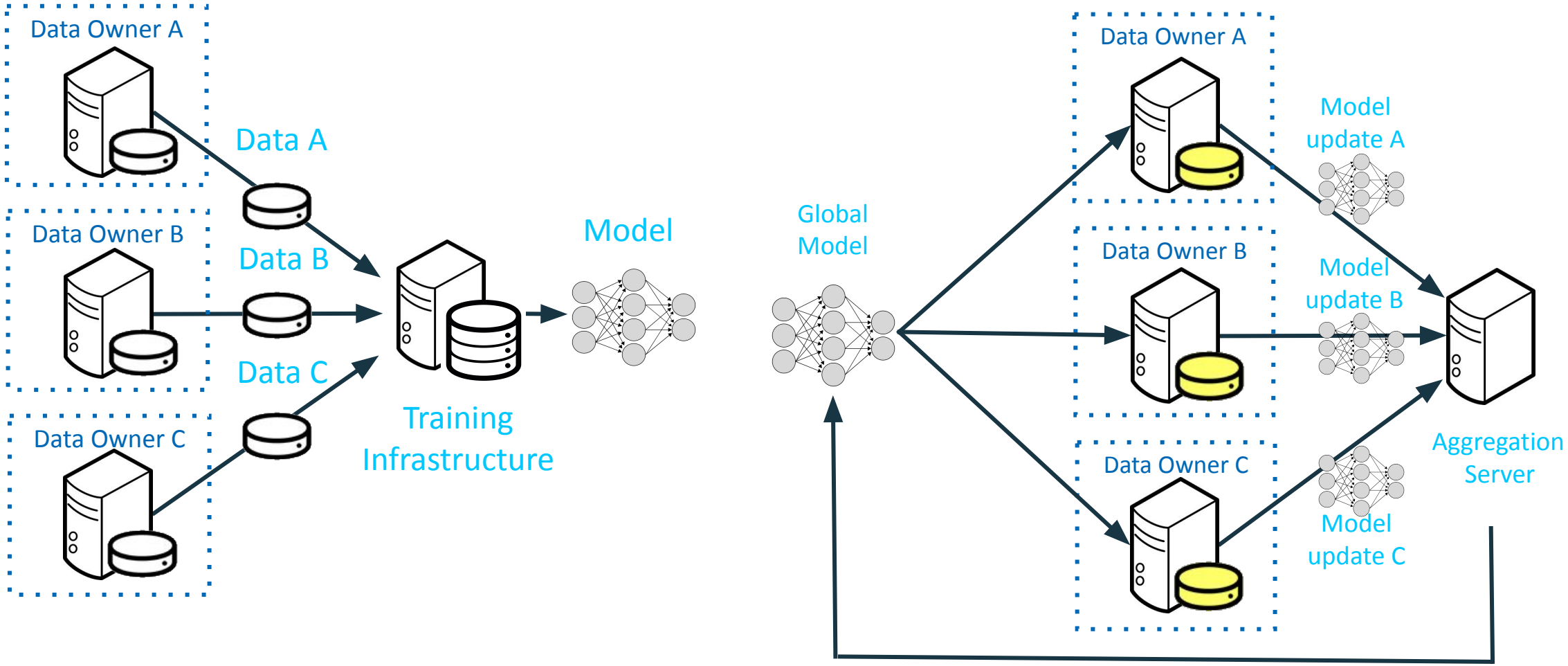
1. **What is Federated Learning**
2. What is OpenFL
3. Collaborations and community
4. Why contribute OpenFL to Linux Foundation?

Challenges for Training AI Models



- Data is legally protected (HIPAA, GDPR)
- Data is sensitive
- Data is too valuable to share
- Data silo problem: data is too large to transmit

Centralized Learning vs Federated Learning



What is Federated Learning?

- **Standard** machine learning approaches require **centralizing the training data** on one machine or in a datacenter.
- Federated Learning enables training models on **distributed and private datasets** without the need to centralize them.
- **Privacy and security** are key considerations for data set owners participating in Federated Learning optimizations.

Financial Services and Insurance

Healthcare

Manufacturing

Media

AI model developer

Pharma Med tech Fin tech Industrial

Needs

- Access to large diverse datasets to train accurate, bias-mitigated models
- Validate AI models on variety of data/real-world settings
- Protection of AI models (IP) in potentially hostile environments while in development

Pain points

- Data owners hesitant to relinquish control of data.
- Models are built on data from only a few institutions -> don't generalize in the real world
- Data is scarce (e.g. rare diseases, rare events)
- Regulatory requirements prevents data acquisition for training and validation of AI models
- Data sets too large to move
- No sophisticated AI security expertise

Data Owner

Hospitals Banks Factories

Needs

- Monetize their existing data assets. Curate once, monetize multiple times.
- Participate in AI value chain
- Participate in AI breakthroughs
- Have full control over their data, preserve value of their data
- Prevent data breaches

Pain points

- Regulatory Compliance (HIPAA, GDPR)
- Grasp of data, but lack of AI expertise
- Moving data requires complex legal agreements and processes (de-identification)
- No sophisticated AI security expertise

AI model developer

Pharma | Med tech | Fin tech | Industrial

Needs

- Access to large diverse datasets to train accurate, bias-mitigated models
- Validate AI models on variety of data/real-world settings
- Protection of AI models (IP) in potentially hostile environments while in development

Pain points

- Data owners hesitant to relinquish control of data.
- Models are built on data from only a few institutions -> don't generalize in the real world
- Data is scarce (e.g. rare diseases, rare events)
- Regulatory requirements prevents data acquisition for training and validation of AI models
- Data sets too large to move
- No sophisticated AI security expertise

Data Owner

Hospitals | Banks | Factories

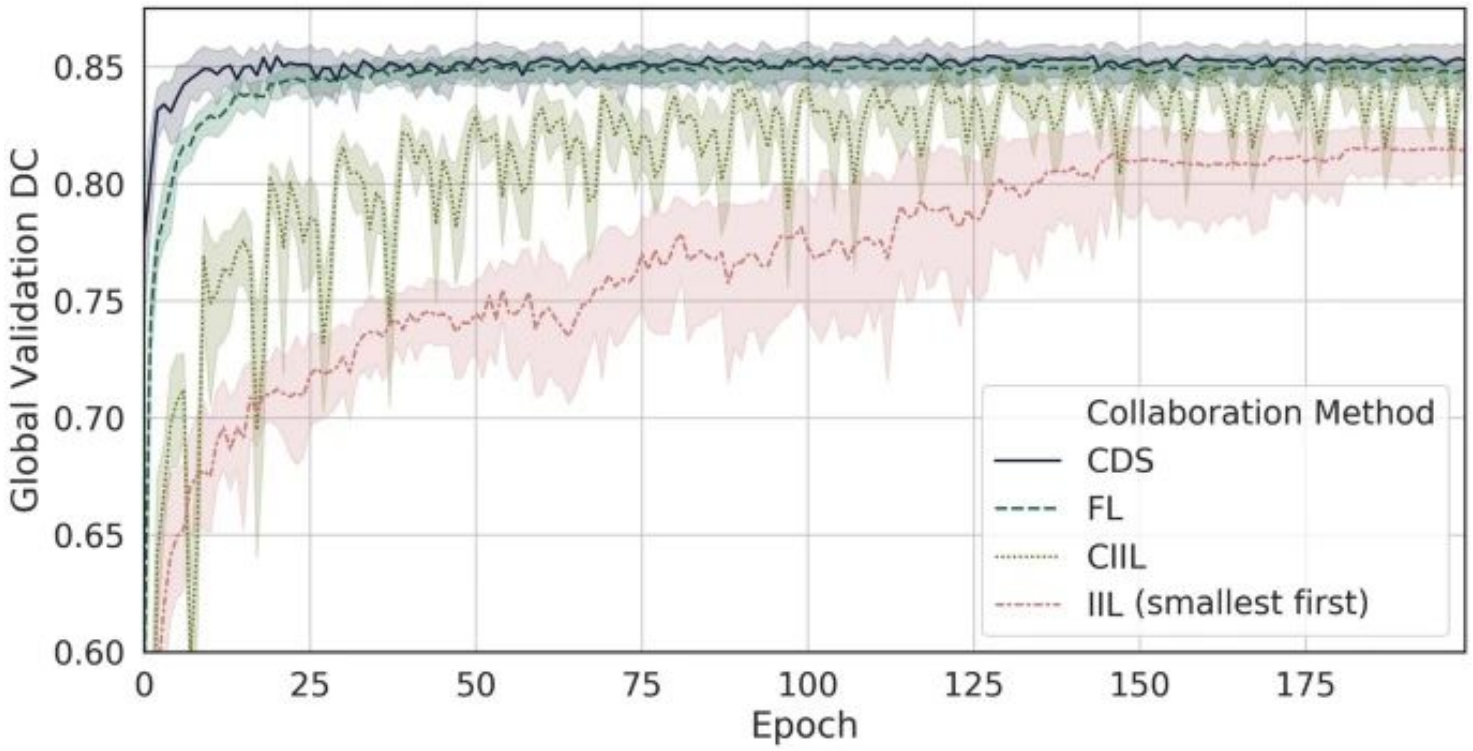
- Monetize their existing data assets. Curate once, monetize multiple times.
- Participate in AI value chain
- Participate in AI breakthroughs
- Have full control over their data, preserve value of their data
- Prevent data breaches
- Regulatory Compliance (HIPAA, GDPR)
- Grasp of data, but lack of AI expertise
- Moving data requires complex legal agreements and processes (de-identification)
- No sophisticated AI security expertise

Solves

Mitigates

Open FL

Centralized Learning versus Federated Learning



scientific reports

Explore our content ∨ Journal information ∨

nature › scientific reports › articles › article

Article | Open Access | Published: 28 July 2020

Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data

Micah J. Sheller, Brandon Edwards, G. Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R. Colen & Spyridon Bakas ✉

Scientific Reports **10**, Article number: 12598 (2020) | Cite this article

3140 Accesses | **119** Altmetric | Metrics

Abstract

Several studies underscore the potential of deep learning in identifying complex patterns, leading to diagnostic and prognostic biomarkers. Identifying sufficiently large and diverse datasets, required for training, is a significant challenge in medicine and can rarely be found in

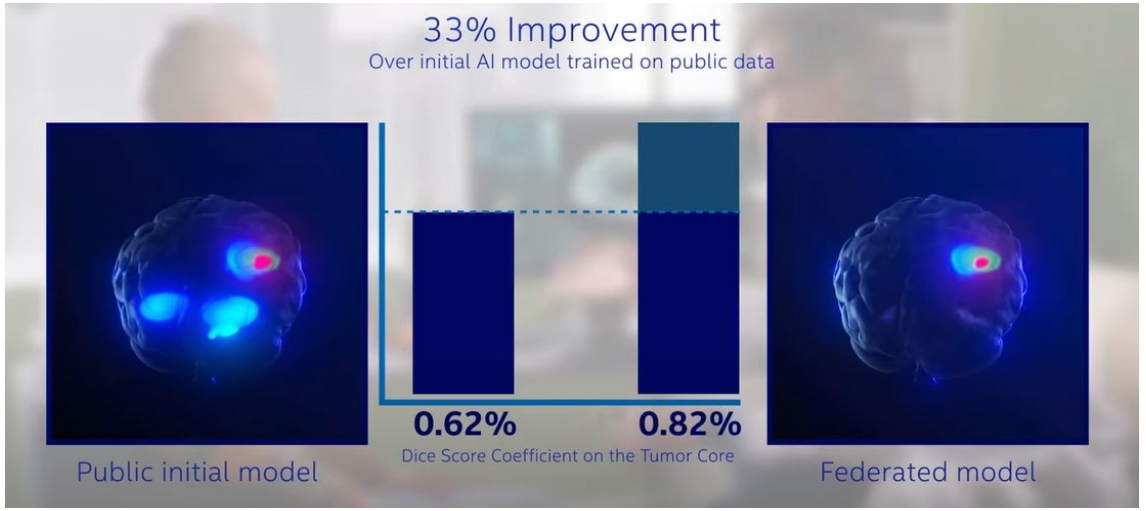
SCIENTIFIC REPORTS

intel

Perelman School of Medicine University of Pennsylvania

nature.com/articles/s41598-020-69250-1

World's largest healthcare federation uses OpenFL



nature communications

Article

https://doi.org/10.1038/s41467-022-33407-5

Federated learning enables big data for rare cancer boundary detection

Received: 7 April 2022

Accepted: 16 September 2022

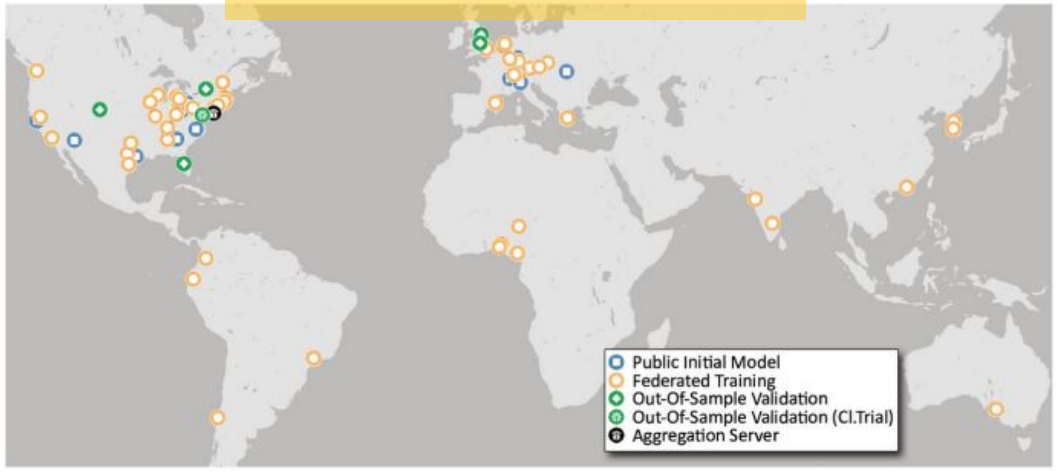
Published online: 05 December 2022

Check for updates

A list of authors and their affiliations appears at the end of the paper

Although machine learning (ML) has shown promise across disciplines, out-of-sample generalizability is concerning. This is currently addressed by sharing multi-site data, but such centralization is challenging/infeasible to scale due to

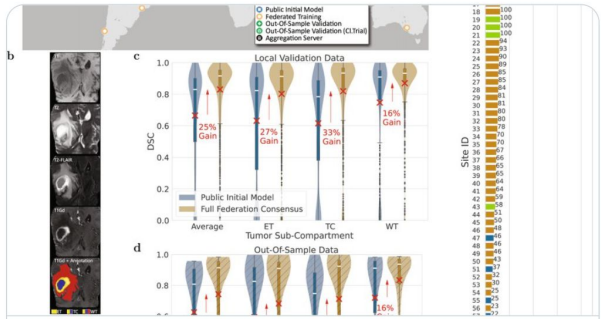
accurate and generalizable ML, by only sharing numerical model updates. Here we present the largest FL study to-date, involving data from 71 sites across 6 continents, to generate an automatic tumor boundary detector for the rare disease of glioblastoma, reporting the largest such dataset in the literature ($n = 6,314$). We demonstrate a 33% delineation improvement for the surgically



Acting director of NIH

Lawrence A. Tabak @NIHDirector · Dec 16, 2022

A large #BrainTumor study used a technique called federated machine learning to combine brain scan data from thousands of patients. The researchers developed a model that enhanced the prediction of tumor boundaries without compromising patient data. bit.ly/3VCAKk0 #NIH



nature.com

Federated learning enables big data for rare cancer boundary detection

Nature Communications - Federated ML (FL) provides an alternative to train accurate and generalizable ML models, by only sharing numerical...

Perelman School of Medicine University of Pennsylvania

intel

www.nature.com/articles/s41467-022-33407-5

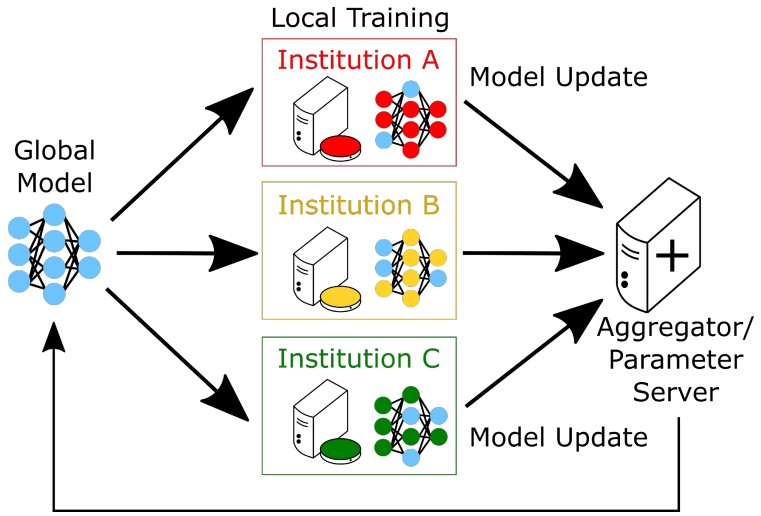
Agenda

1. What is Federated Learning
2. **What is OpenFL**
3. Collaborations and community
4. Why contribute OpenFL to Linux Foundation?

OpenFL: Python library for Federated Learning

License Apache 2.0

Open FL



Keras

TensorFlow

PyTorch

GitHub

github.com/intel/openfl

pypi

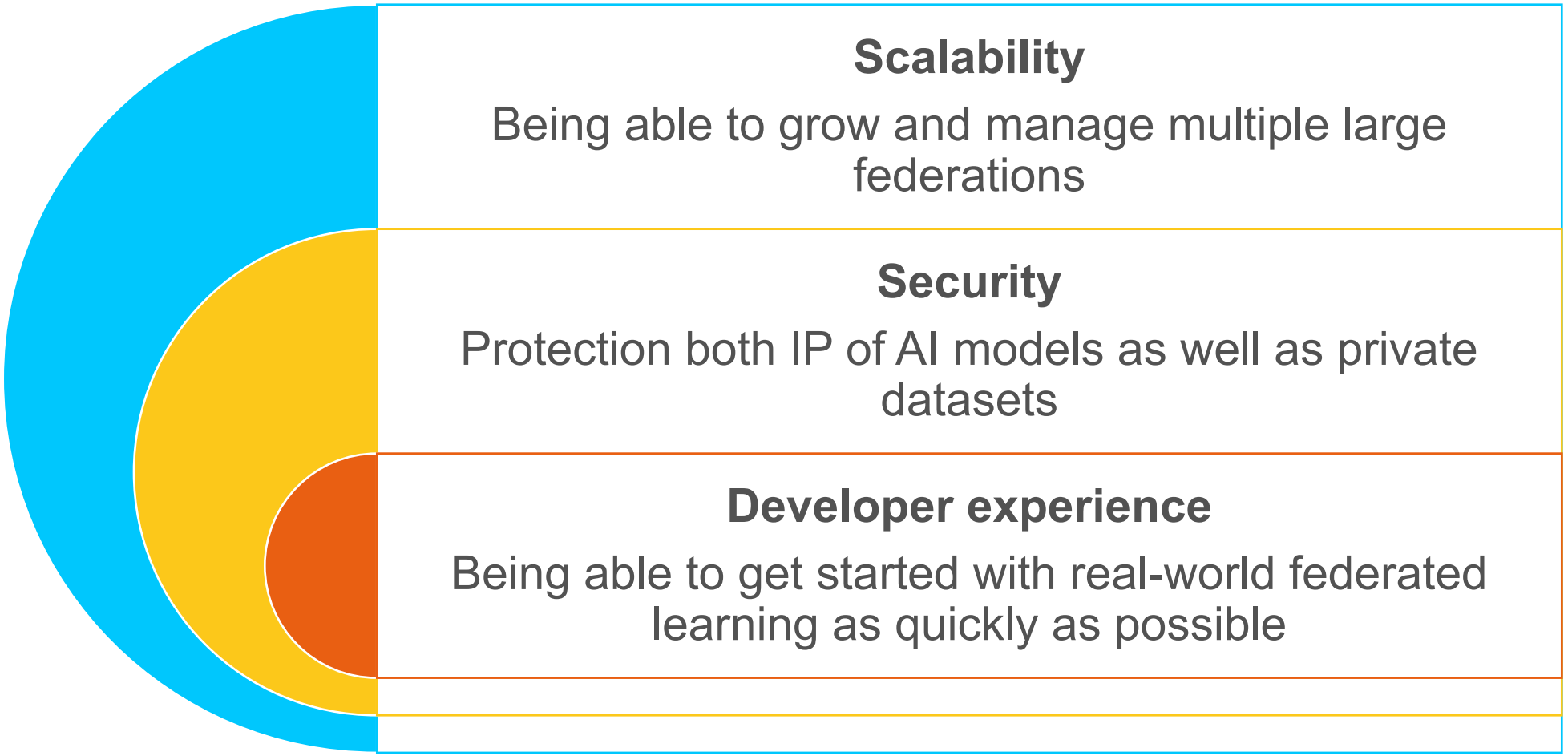
pip install openfl

docker hub

docker pull intel/openfl

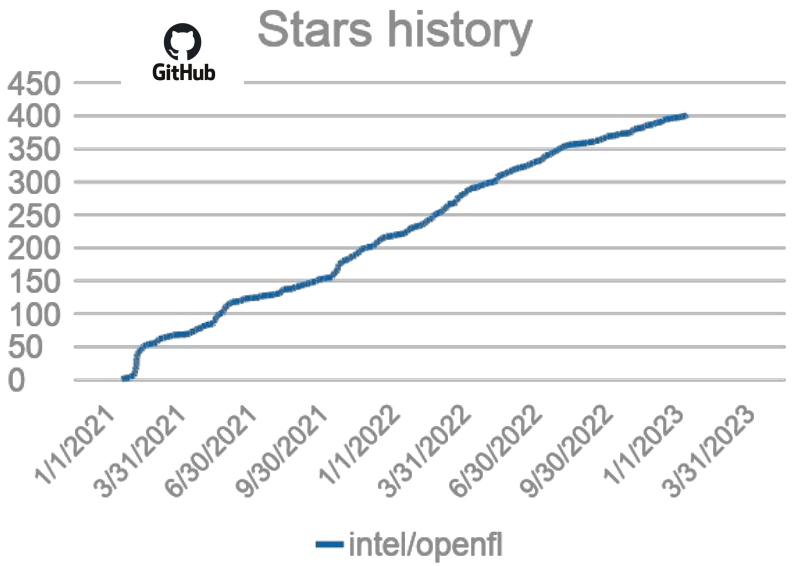
P. Foley, *et Int*, S. Bakas, *Phys Med Biol*, 67(21):214001, 2022

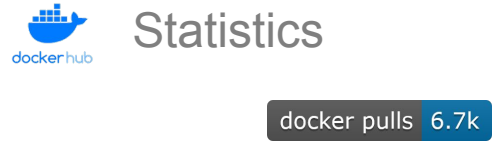
OpenFL core values



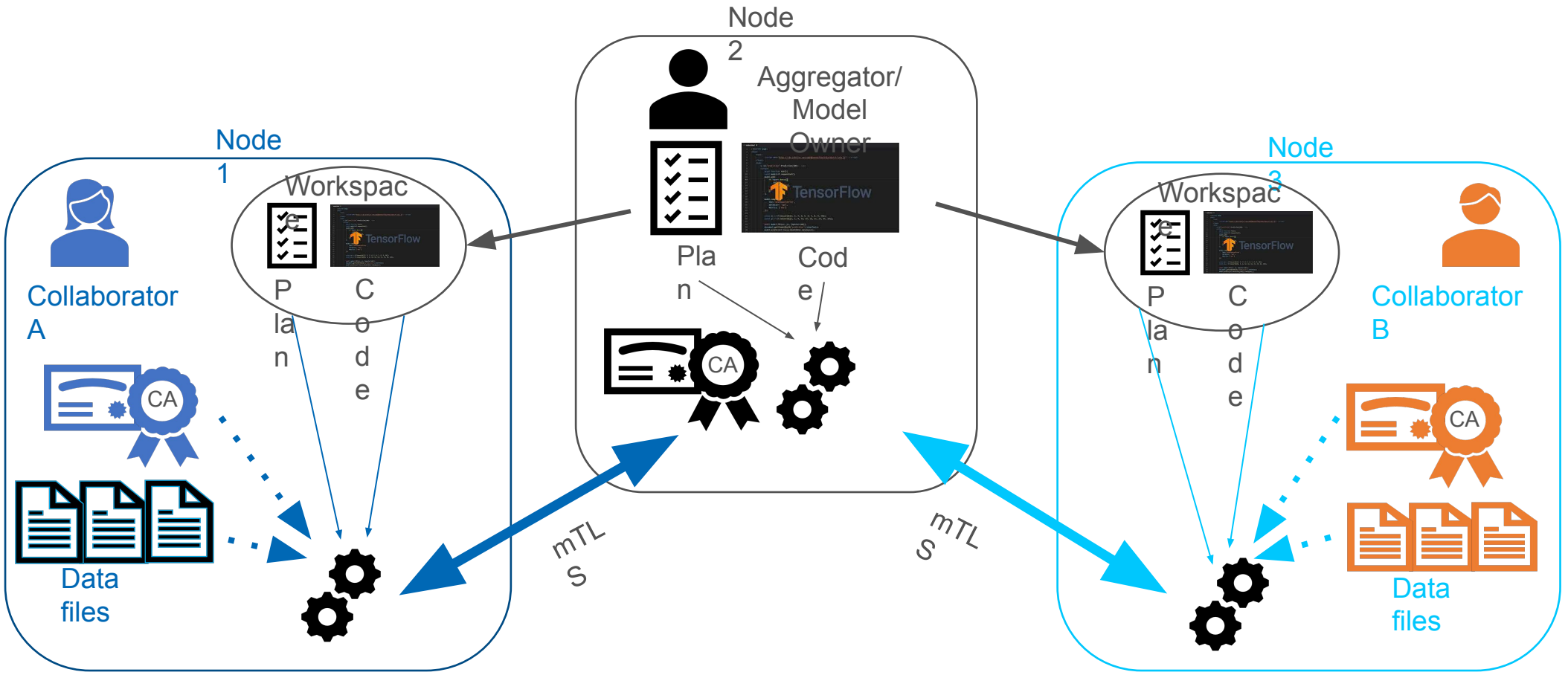
OpenFL: progress summary

- **Intel Labs's** initial research and contribution in 2018-2020
- **Public Release** on GitHub: Feb 1, 2021
- **6 major releases**: OpenFL 1.0 – OpenFL 1.5

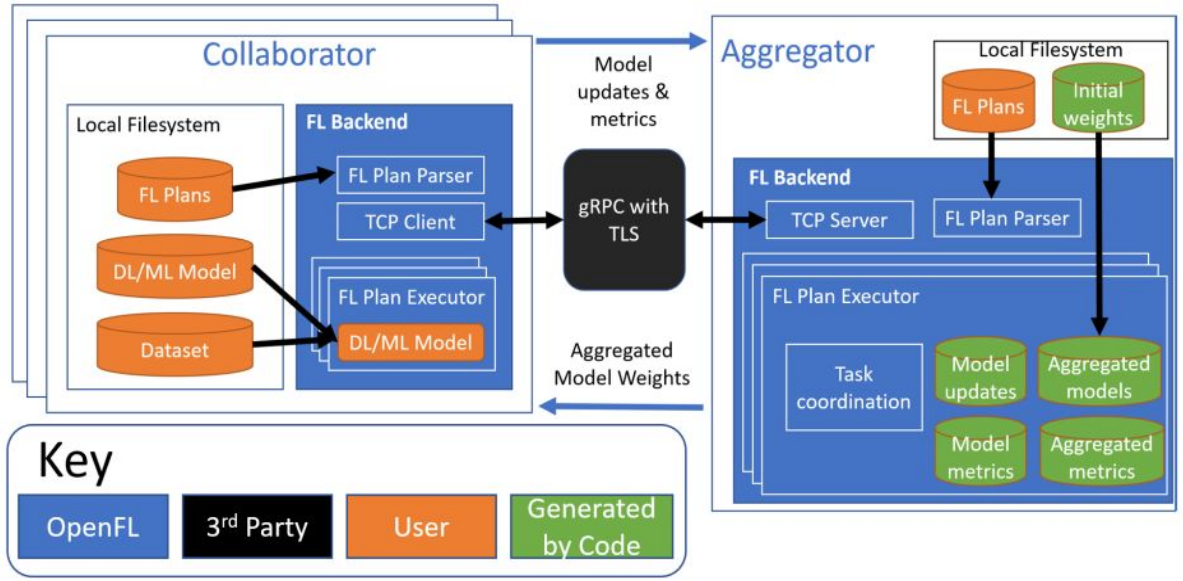




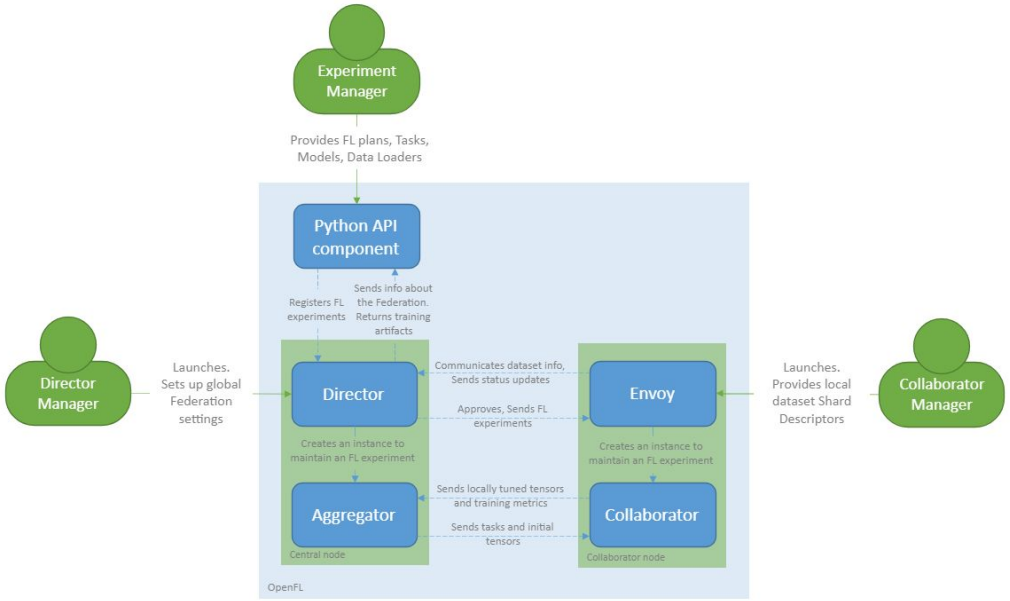
OpenFL Architecture



OpenFL Functionality

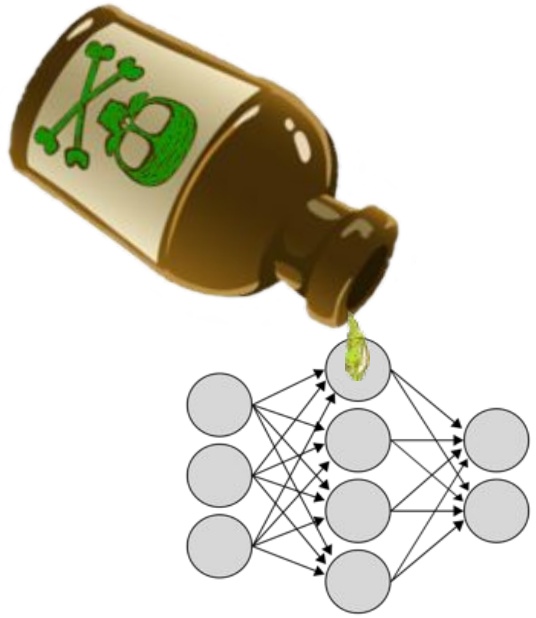


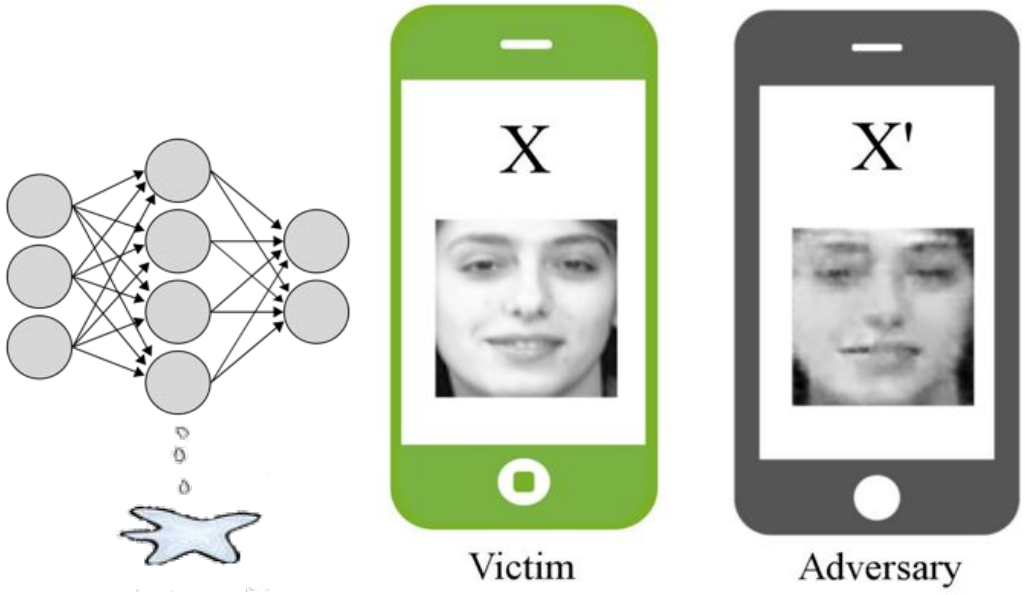
Aggregator Workflow



Director Workflow

FL Could Increase Security and Privacy Risks





Poisoning attacks may maliciously alter models.

Extraction attacks recover training data from models.

FL needs to have additional **security** to manage these risks

Why Federated Learning with Intel® SGX*

* SGX opensource integration with OpenFL will be added in next releases
Integration with other TEE hardware can be added to the project by contributors

Confidentiality

- Data never leave the premise of data owners.
- Model IP protected end-to-end in use and at rest.

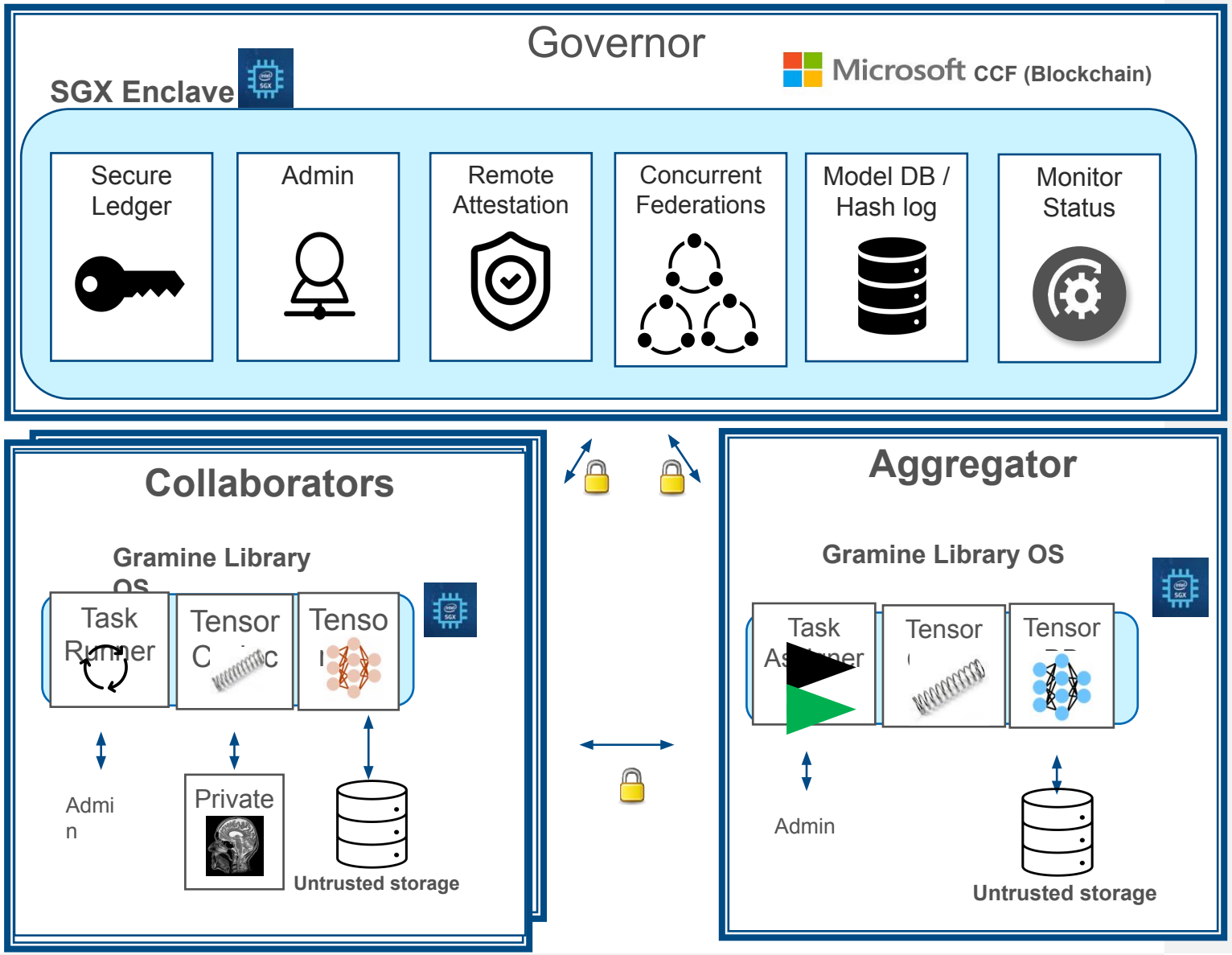
Integrity & ATTESTATION

- Only verified and approved ML models.
- Participants can not insert unapproved code at any time.

Provides a mechanism to prevent stealing the model or reverse-engineering data distribution.

OpenFL Security Overview

- **Add-on repository** introduces new components to OpenFL to improve privacy and security
- **Governor**: C++ application layer built on Microsoft Confidential Consortium Framework. Manages participant / dataset inventory, verifies application integrity, and enforces the plan (participants, datasets, model, etc.) to be used in an experiment.
- **Attestation Verification Service (AVS)** – used to verify the integrity of SGX enclaves.
- **OpenFL components –** The collaborator and aggregator are inherited from the OpenFL repo. Gramine LibOS used to run unmodified applications in **SGX Enclaves**



Agenda

1. What is Federated Learning
2. What is OpenFL
3. **Collaborations and community**
4. Why contribute OpenFL to Linux Foundation?

Who is using OpenFL?

- University of Pennsylvania created the first real-life and largest federation of healthcare institutions.
- Federated Tumor Segmentation Challenge 2021 is the first federated learning competition. It focused for the task of brain tumor segmentation.
- Frontier Development Lab: NASA, Mayo Clinic and Intel used federated learning to understand the effect of cosmic radiation on humans

- Montefiore used OpenFL to simultaneously tap data from multiple hospitals to predict likelihood of Acute Respiratory Distress Syndrome (ARDS) and Death in Covid-19 patients
- VMware used OpenFL for Microservices Applications and contributed EDEN, a new compression pipeline designed for federated learning, to OpenFL.

Federated Tumor Segmentation Challenge 2021, 2022

- Brain tumor segmentation task: 2021, 2022
- Leveraged **OpenFL**
- Information gathered from FeTS initiative (> 50 medical institutions), data used was representative for real-world use case and split into ~20 partitions
- The goal was to create effective weight aggregation methods for the creation of a consensus model and to evaluate developed segmentation algorithms 'in-the-wild', by circulating them in sites around the world that have not contributed data to the training set

VMware contribution to OpenFL

MENU vmware Open Source Blog

Projects

VMware Research Group's EDEN Becomes Part of OpenFL

Open Source Team
November 16, 2022

Share on:

0

VMware's OCTO Research Group in collaboration with Intel, are pleased to announce that EDEN, a new compression pipeline for federated learning (FL), has joined the OpenFL community.

Originated at Intel, OpenFL is an open source framework for training ML algorithms using the data-private collaborative learning paradigm of FL. OpenFL is designed to be a flexible, extensible and easily learnable tool for data scientists. OpenFL works with training pipelines built with both TensorFlow and PyTorch, and can be easily extended to other ML and deep learning frameworks.

Distributed Mean Estimation (DME)

DME is a central building block in FL, where clients send local gradients to a parameter server for averaging and updating the model. Due to communication constraints, clients often use lossy compression techniques to compress the

MENU vmware Open Source Blog

Projects

Federated Learning With OpenFL for Microservices Applications

Yujing Chen
August 31, 2022

Share on:

0

Ten months ago, I joined xLabs, an "agile incubation lab" for all kinds of innovative ideas, within VMware's Office of the CTO. I had just completed my PhD in computer science at George Mason University, with a focus on machine learning, federated learning, multi-task learning and deep learning. Now as a machine learning engineer tasked to work on an API security and analytics platform project (Project Trinidad), this role couldn't be more aligned with my career goals.

The Project Trinidad team has six core members, including myself, and an extended team of the same number. Project Trinidad's objective is to protect modern applications by detecting and blocking cyberattacks. It acts as an X-ray machine that allows us to study the internal communication of modern apps and monitor both north-south and east-west API communication between microservices. My mission on the project is utilizing the forefront of ML technologies (e.g., federated learning, deep learning) for anomaly detection, while maintaining the low false positives of our detection system.

Privacy-preserving ML with federated learning

OpenFL community

- Community Meetings

Support

Please join us for our bi-monthly community meetings starting December 1 & 2, 2022!
Meet with some of the Intel team members behind OpenFL.
We will be going over our roadmap, open for Q&A, and welcome idea sharing.

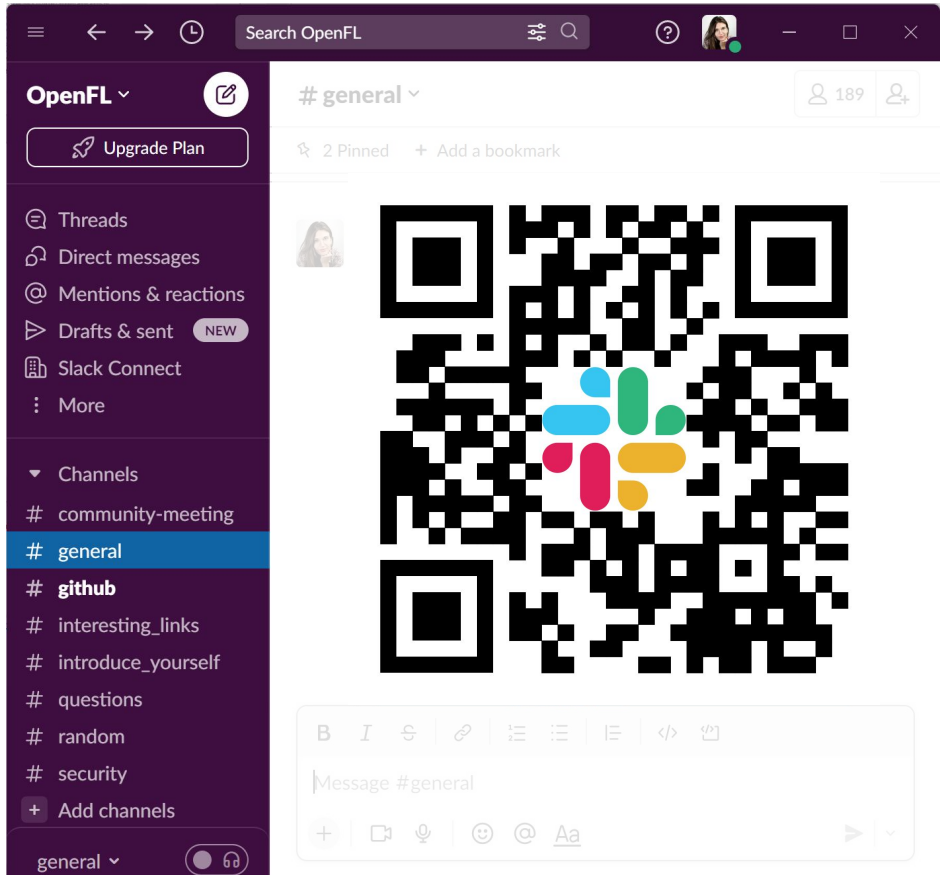
2-time slots available:

Europe Occurs every 2 months on the first Thursday of that month, 6 pm
https://intel.zoom.us/j/96858990725?pwd=ODJRVUhXNGNQU2YySTFqRi9qRXdtUT09
Meeting ID: 968 5899 0725
Passcode: 157568

Asia Pacific Occurs every 2 months on the first Friday of that month, 10 AM GMT+8
https://intel.zoom.us/j/98930263828?pwd=VkhIbVBtTG9xaHZCZ2dQNUdYSWpydz09
Meeting ID: 989 3026 3828
Passcode: 774320

See you there!

- OpenFL Slack



Join our community calls!

EU time zone invite

Asia time zone invite

Agenda

1. What is Federated Learning
2. What is OpenFL
3. Collaborations and community
4. **Why contribute OpenFL to Linux Foundation?**

Why contribute OpenFL to Linux Foundation?

- True Open Source
  - AI & Data is very well aligned with OpenFL's mission of community based, collaborative open development.
  - Linux Foundation allows a neutral 3rd party community-based home where other ecosystem players can collaborate, contribute, and use the open-source Federated Learning framework.
  - Linux Foundation will help to increase developer outreach and grow Federated Learning community.

Why contribute OpenFL to Linux Foundation?

- True collaboration
  - Synergize and standardize components amongst Federated Learning Frameworks (Flower, OpenFL, and potentially FATE and Substra) to drive interoperability.
  - LF AI and data synergies with other AI projects (Horovod, ONNX, TrustedAI, Pyro).

Request to join LF AI & Data

- OpenFL would like to join at the **Incubation stage**

Requirements	Current status
Have at least two organizations actively contributing to the project	**Intel, UPenn, VMware**
Have a defined Technical Steering Committee (TSC) with a chairperson identified	In progress
Have a sponsor who is an existing LF AI & Data member	**VMware, Intel is joining**
Have at least 300 stars on GitHub	**> 400 stars**
Have achieved Best Practices Silver Badge	In progress bestpractices.coreinfrastructure.org/en/projects/6599

Links & Materials

OpenFL GitHub: github.com/intel/openfl

OpenFL tutorials: github.com/intel/openfl/tree/master/openfl-tutorials

OpenFL PyPI: pypi.org/project/openfl/

OpenFL Docker Hub: hub.docker.com/r/intel/openfl

OpenFL publications: arxiv.org/abs/2105.06413, iopscience.iop.org/article/10.1088/1361-6560/ac97d9

OpenFL Slack: https://join.slack.com/t/openfl/shared_invite/zt-ovzbohvn-T5fApk05~YS_iZhjJ5yaTw

Federated Learning in Medicine – Nature 2020: www.nature.com/articles/s41598-020-69250-1

Federated Learning enables big data for rare cancer boundary detection – Nature 2022: www.nature.com/articles/s41467-022-33407-5

Federated Tumor Segmentation Initiative: www.med.upenn.edu/cbica/fets/

NASA use case with OpenFL 2021: www.intel.com/content/www/us/en/newsroom/news/intel-ai-mentors-seek-improve-astronaut-health.html

VMware contribution to OpenFL 2022: blogs.vmware.com/opensource/2022/11/16/vmware-research-groups-eden-becomes-part-of-openfl/

Thank you for attention!