The Fed-BioMed Project
Federated Learning Across Health Institutions in France

Marco Lorenzi
Inria Sophia Antipolis, Université Côte d'Azur
Epione Research Group

IA in Healthcare
need for large data repositories

SCIENCES • SANTÉ

Partage

TRIBUNE

Collectif

« Les données de santé servent l'intérêt public, il y a urgence à en faciliter l'accès »

Le retard pris dans le déploiement du Health Data Hub, infrastructure unique facilitant l'accès aux données de santé de façon sécurisée, est inquiétant, affirment les membres de son conseil scientifique consultatif dans une tribune au « Monde ».

Publié hier à 06h30, mis à jour hier à 07h20 | Lecture 4 min.

Article réservé aux abonnés

Le Monde, 20/10/2021

Access and sharing of multiple centers data falls into General Data Protection Regulation (GDPR): Privacy, confidentiality, security, ...

The impact of the General Data Protection Regulation (GDPR) on artificial intelligence

European Parliament

https://bit.ly/3IyJFg7

AI is not explicitly mentioned in the GDPR, but many provisions in the GDPR are relevant to AI, and some are indeed challenged by the new ways of processing personal data that are enabled by AI

- Ethical principles include autonomy, prevention of harm, fairness and explicability;
- legal principles (EU rights and social values, in the EU treaties, national constitutions).

The impact of the General Data Protection Regulation (GDPR) on artificial intelligence

European Parliament

https://bit.ly/3IyJFg7

Purpose limitation

Compatible with AI and big data, through a flexible application of the idea of compatibility, which allows for the reuse of personal data

Data minimisation

Reducing, through measures such as pseudonymisation, the ease with which the data can be connected to individuals. Re-identification should indeed be strictly prohibited unless all conditions for the lawful collection of personal data are met.

Preventive measures

It needs to be clarified which AI applications present high risks and therefore require a preventive data protection assessment, and possibly the preventive involvement of data protection authorities.

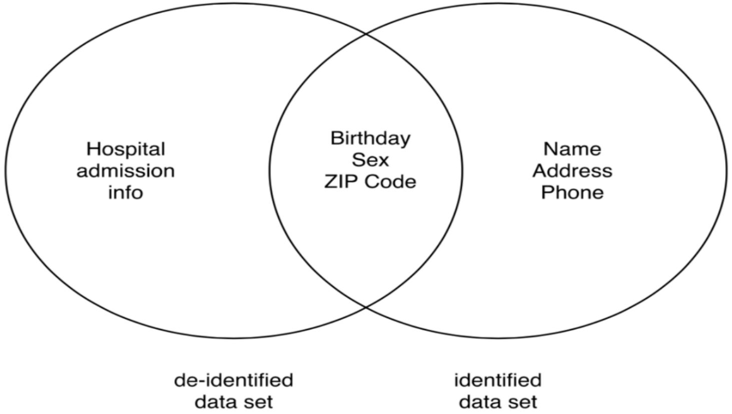
The impact of the General Data Protection Regulation (GDPR) on artificial intelligence

European Parliament

https://bit.ly/3IyJFg7

The inference of new personal data, as it is done in profiling, should be considered as creation of new personal data, when providing an input for making assessments and decisions. The same should apply to the re-identification of anonymous or pseudonymous data.

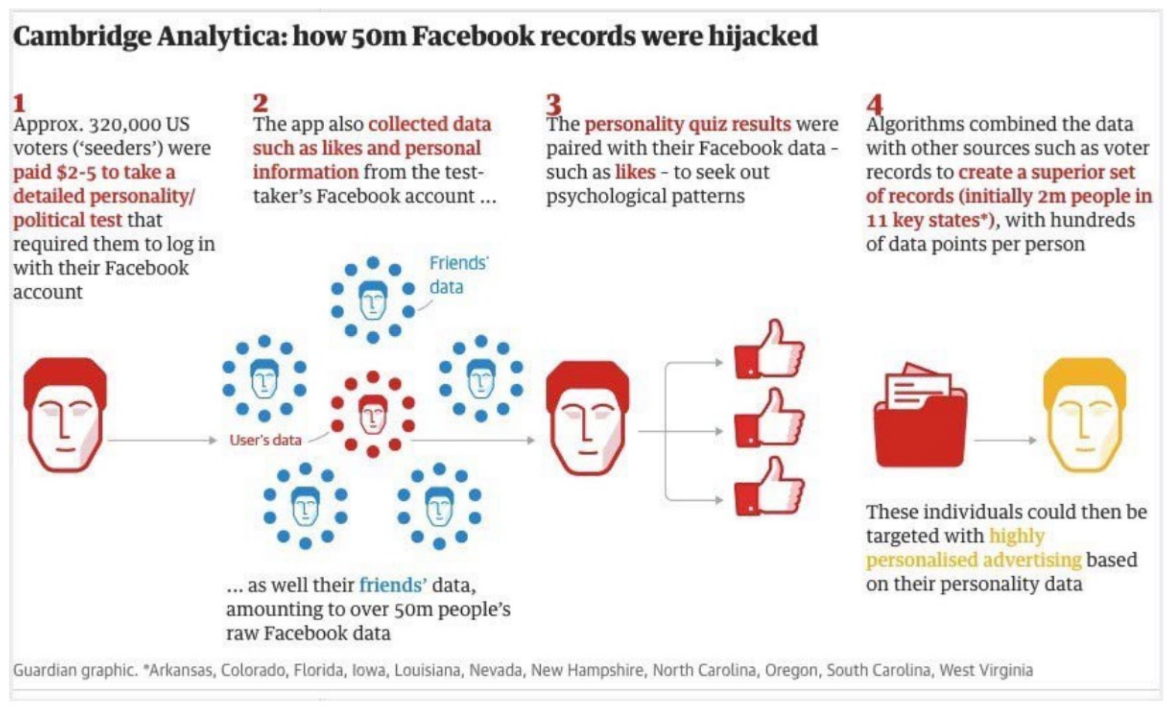
The risk with pseudo-anonymization



The impact of the General Data Protection Regulation (GDPR) on artificial intelligence

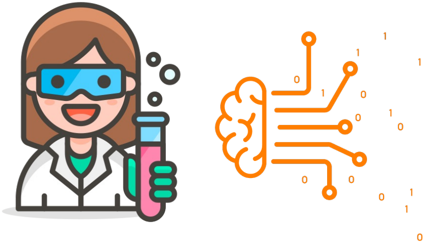
European Parliament

https://bit.ly/3IyJFg7

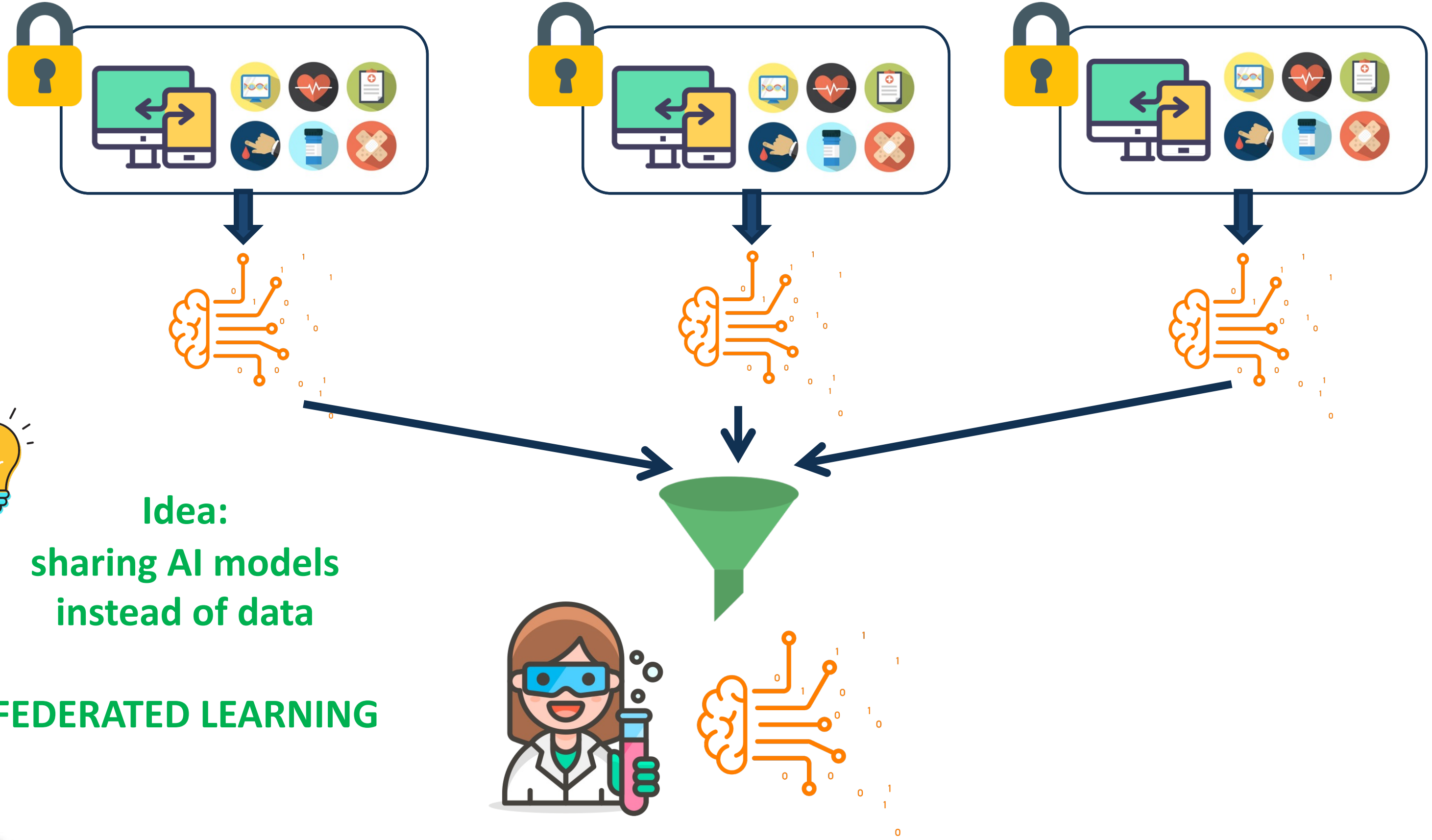


A centralized paradigm?

Problem:
Developing AI requires data access



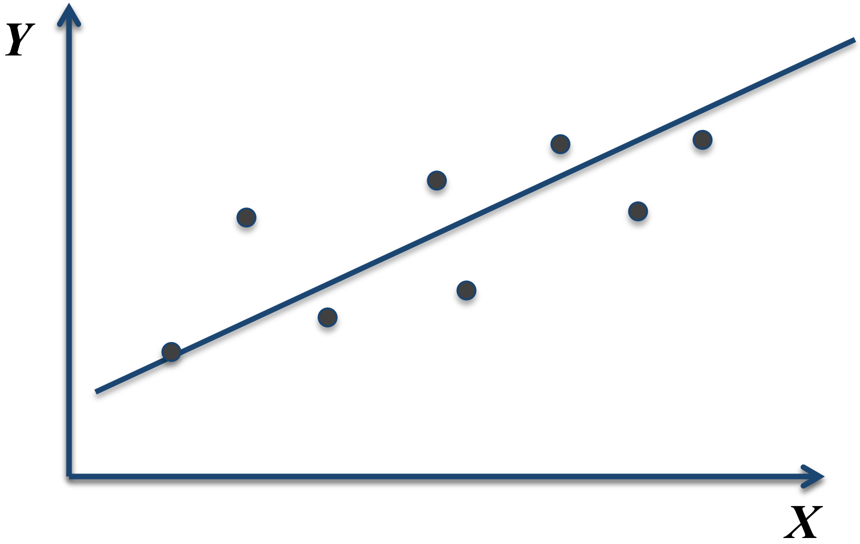
The federated paradigm



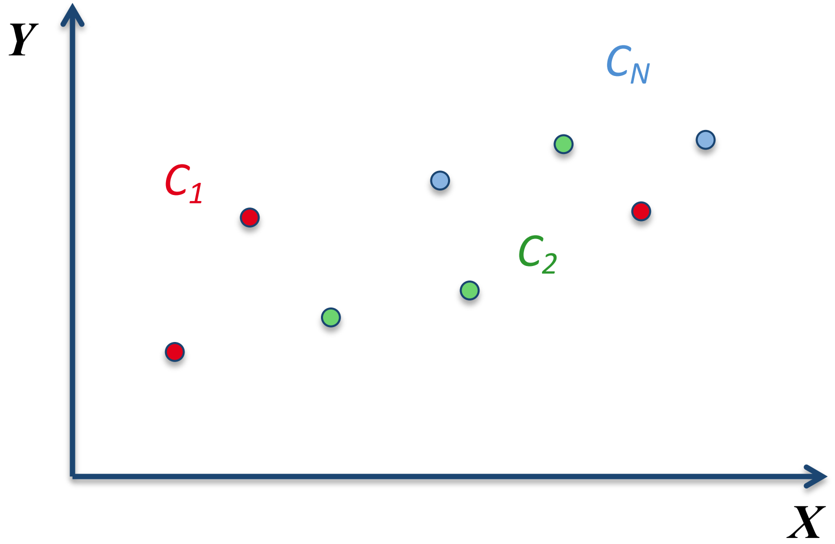
Idea:
sharing AI models
instead of data

FEDERATED LEARNING

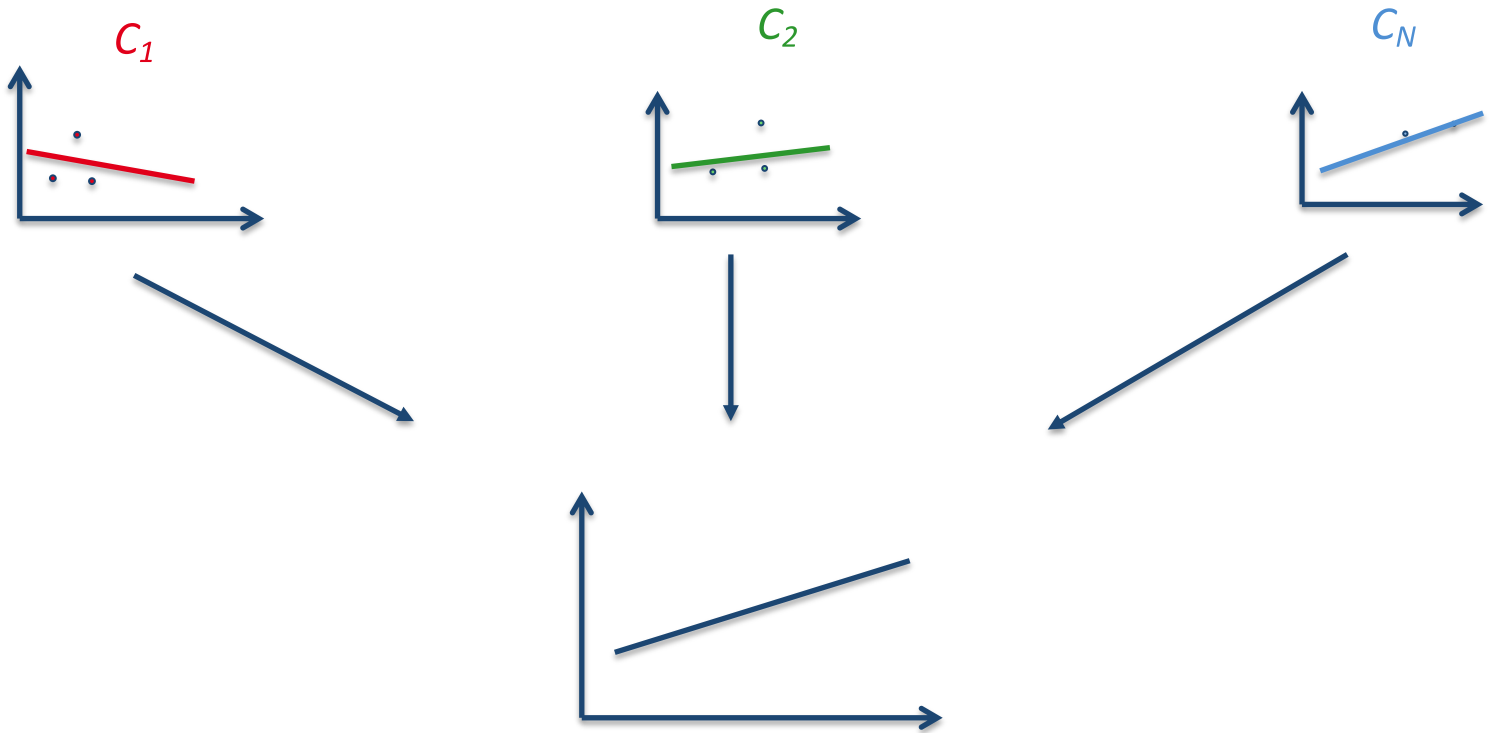
Federated linear modeling



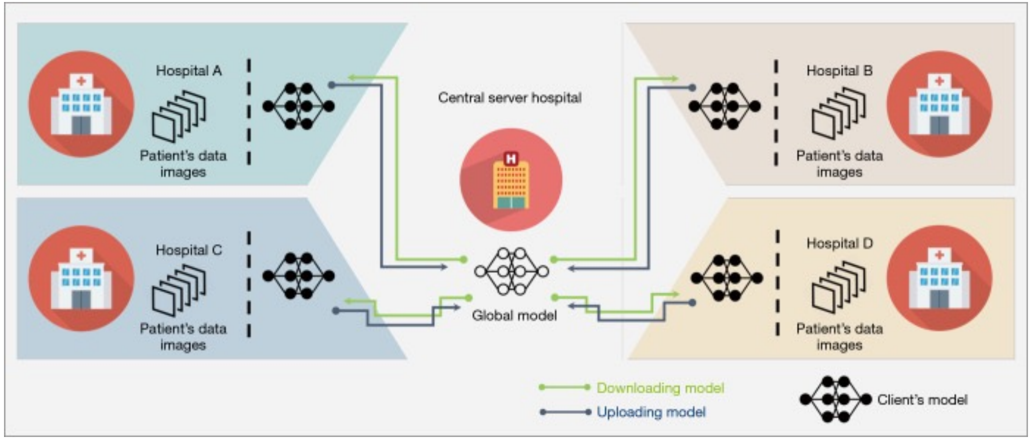
Federated linear modeling



Federated linear modeling



Federated Learning for Collaborative Data Science



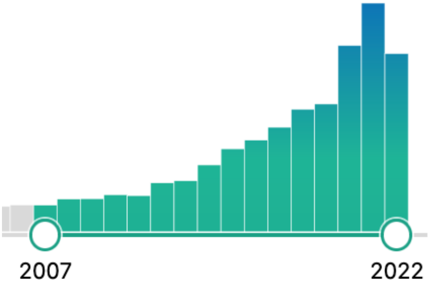
From Ng et al, Quant Imaging Med Surg, 2021;

From Xu et al. J Healthc Inform Res. 2020

Table 1

Summary of recent work on federated learning for healthcare

Problem	ML method	No. of clients	Data
Patient similarity learning [62]	Hashing	3	MIMIC-III [50]
Patient similarity learning [108]	Hashing	20	MIMIC-III
Phenotyping [55]	TF	1–5	MIMIC-III, UCSD [104]
Phenotyping [67]	NLP	10	MIMIC-III
Representation learning [93]	PCA	10–100	ADNI, UK Biobank, PPMI, MIRIAD
Mortality prediction [45]	Autoencoder	5–50	eICU Collaborative Research Database [81]
Hospitalization prediction [10]	SVM	5, 10	Boston Medical Center
Preterm-birth prediction [9]	RNN	50	Cerner Health Facts
Mortality prediction [80]	LR, NN	31	eICU Collaborative Research Database
Mortality prediction [90]	LR, MLP	2	MIMIC-III
Activity recognition [16]	CNN	5	UCI Smartphone [4]
Adverse drug reactions Prediction [19, 20]	SVM, MLP, LR	10	LCED, MIMIC
Arrhythmia detection [110]	NN	16, 32, 64	PhysioNet Dataset [21]
Disease prediction [33]	NN	5, 10	Pima Indians Diabetes Dataset [95], Cleveland Heart Disease Database [23]
Imaging data analysis	VAE	4	MNIST, Brain Imaging Data
Mortality prediction [101]	LRR, MLP, LASSO	5	Mount Sinai COVID-19 Dataset



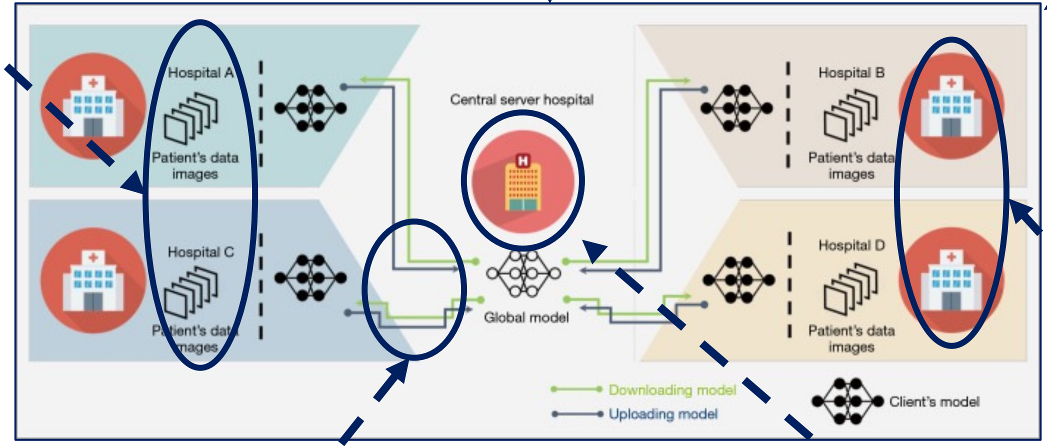
PubMed query "federated learning"
June 6th 2021

Translation of Collaborative Medical Data Analysis

Security model

Heterogeneous
networks / database / data format

Regulation:
Data Protection
Officer,
Homologation,
Ethical Committee



Availability of local
resources
(hardware and personnel)

Data flow

Governance:
Server location,
Researcher role,
Hospital control over FL

(some) Open FL software initiatives

Coinstac
coinstac.org

PySyft
github.com/OpenMined/PySyft

Flower
flower.dev

LEAF
leaf.cmu.edu

Tensor Flow Federated
www.tensorflow.org/federated

FedML
github.com/FedML-AI

OpenFL (Intel), IBM FL, NVFlare (NVIDIA), …

FL Software Requirements and Challenges

- Lack of standards
- Scalability
- Portability
- Generalization to multiple ML frameworks
- Security
- Open technology
- Support for medical data analysis

Fed-BioMed

fedbiomed.gitlabpages.inria.fr

Funded by

ANR, Inria, 3iA Côte d'Azur – Institut interdisciplinaire d'intelligence artificielle, Université Côte d'Azur

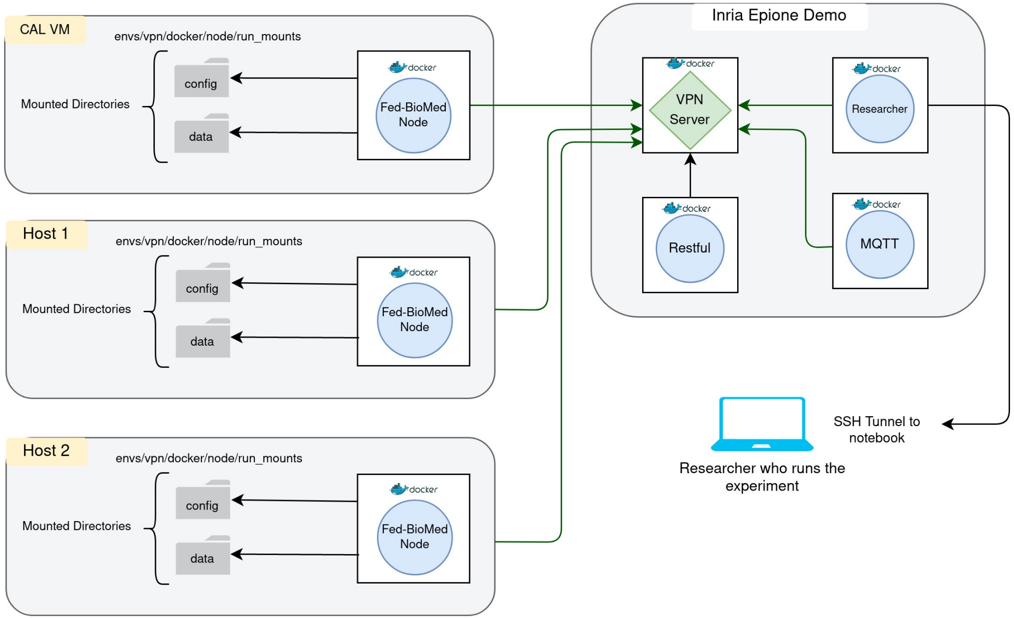
Support from
Inria National Plan for Artificial Intelligence
and
Accenture Labs Sophia Antipolis

Le Grand Plan d'Investissement – Investir l'avenir; Innovation Architecture: Accenture Labs

Tailored for AI applications in healthcare

Simplified model development and deployment
Security and compatibility with hospital networks
Security/Governance

Framework architecture



Platform- and model- agnostic

PyTorch scikit learn MONAI

FL Design choices: From research to real-life

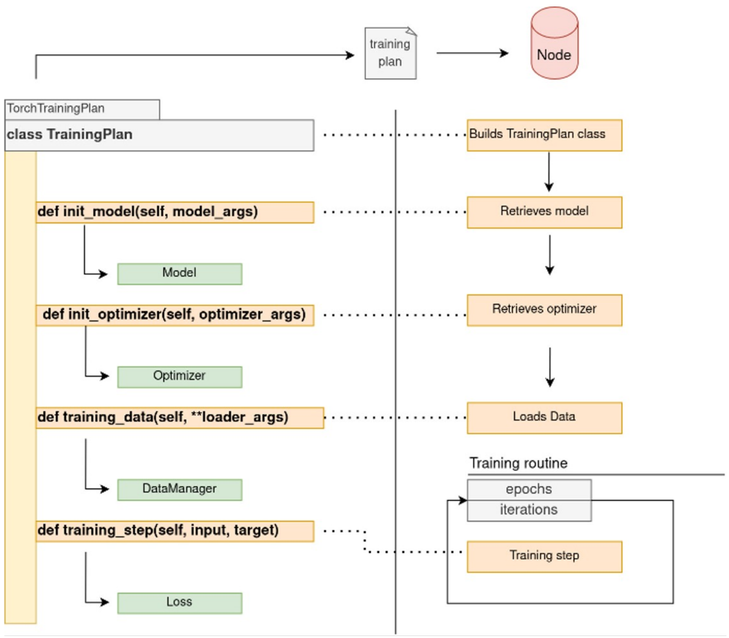
Researchers

- Flexible experimentation environment
- Launching different experiment easily
- Control over experiment parameters
- Real-time feedback

Clients/Data owners

- Constraints
- Approval of an experiment
- Overwriting requests
- Privacy

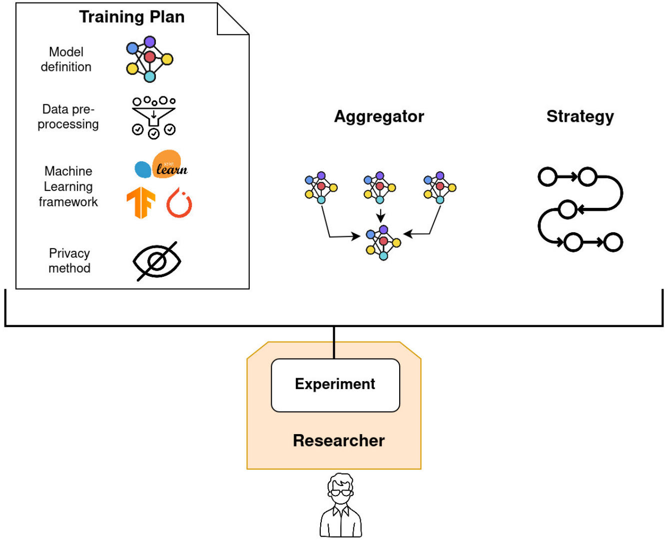
Training Plan



- Managing training through parameters/arguments
- Multiple models and multiple experiments
- Allowing preprocessing
- Monitoring training/testing in real-time

A typical FL Strategy
(e.g. Fed-Avg, Fed-Prox, Scaffold)

- Pre-processing needed
- Privacy model
- Model to be deployed
- Quantities to aggregate
- Client sampling rule



Privilege to overwrite requests

Researcher

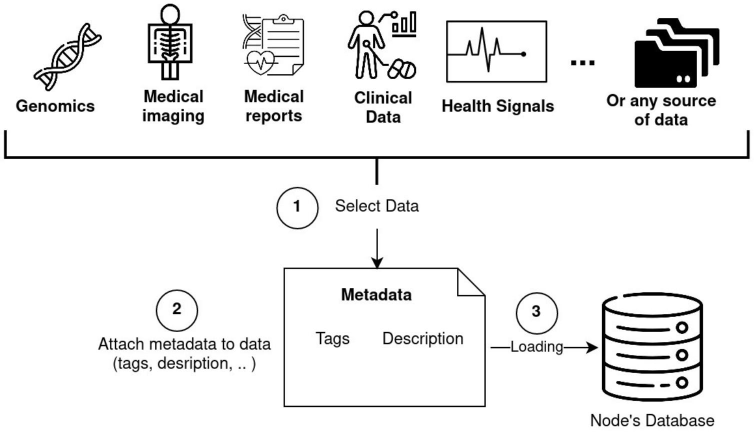
training/model/optimizer arguments
number of rounds
batch size
arguments specific to method
DP parameters

Client

allowed max rounds of training
force GPU/CPU usage
allowed number of samples
force private DP parameters

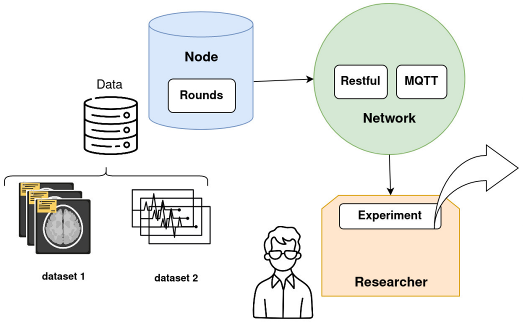
FL subject to dynamically changing conditions

Handling Heterogeneity



Generally limited interoperability with hospital database
Requires available local knowledge
Data preparation is time consuming

- System Interoperability, interface with PACS
- Standardization
- Handling Errors



Datasets				
Name	data_type	tags	description	shape
dataset 1	images	#brain_imaging	Brain images	[600, 128,128]
dataset 2	csv	#heart_signal	Heart signals	[10, 10]

FL Security

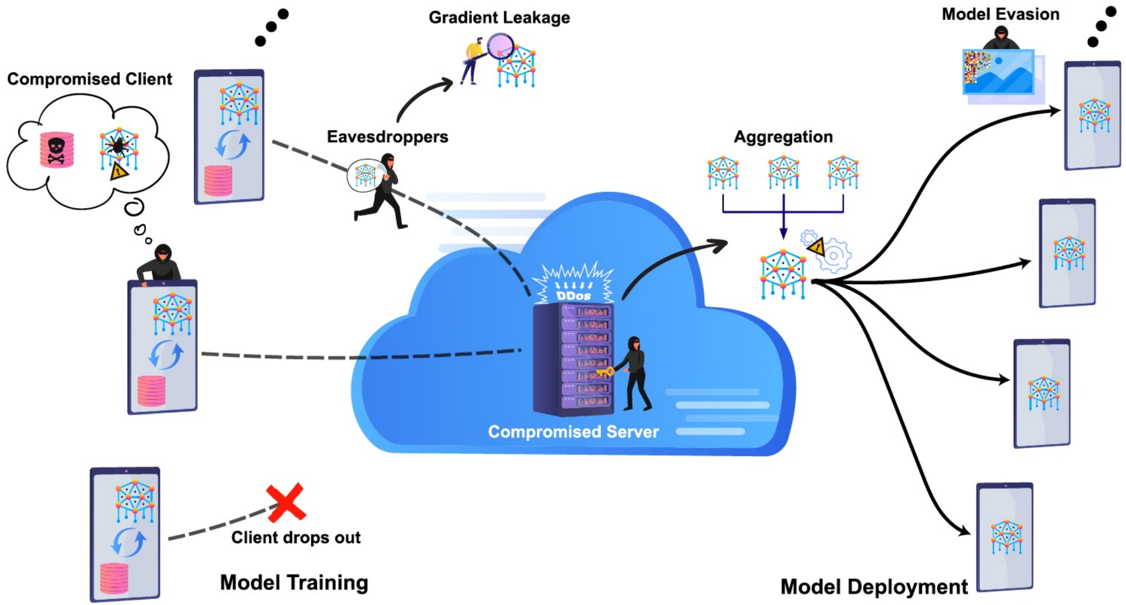
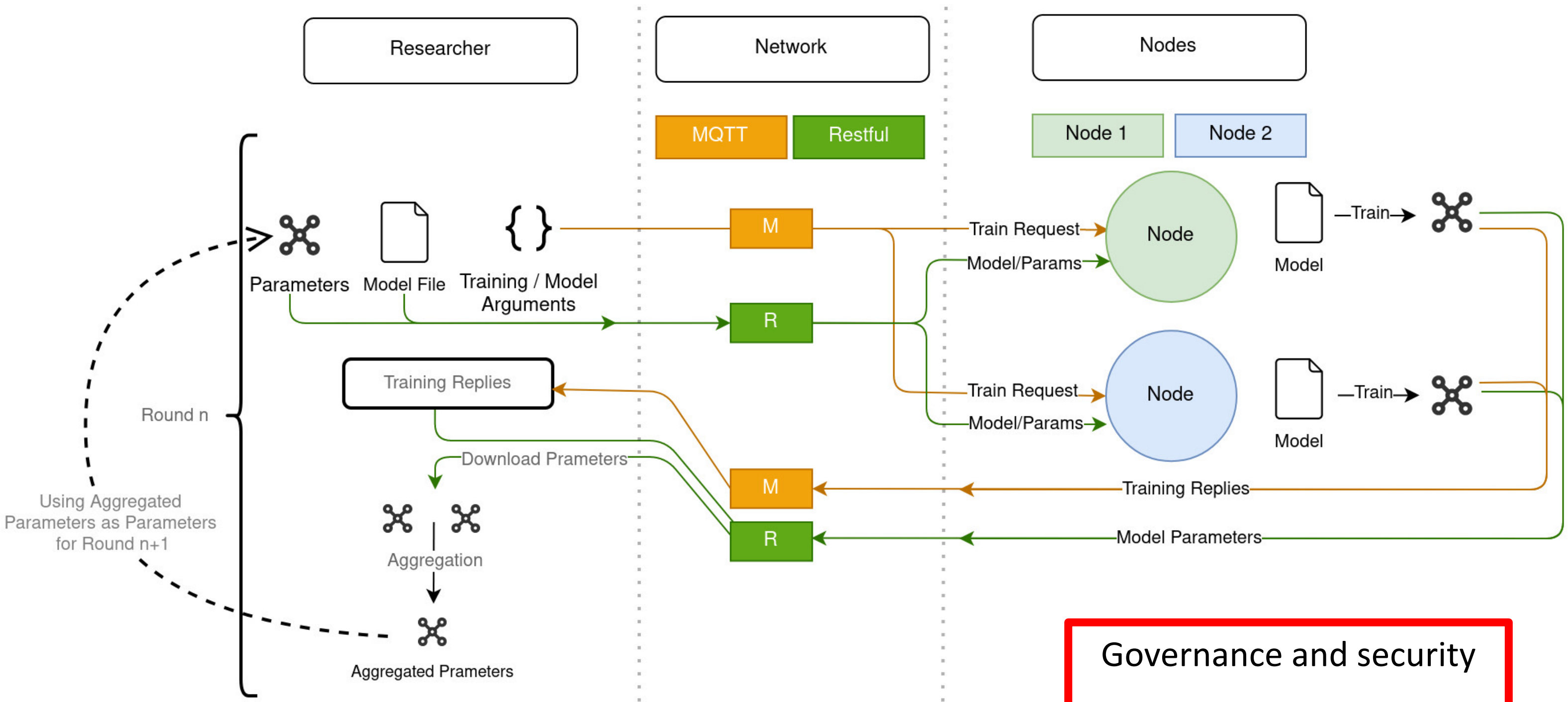


FIGURE 1. The lifecycle of FL process and the various sources of vulnerabilities.

From Bouacida et al. IEEE Access 2021

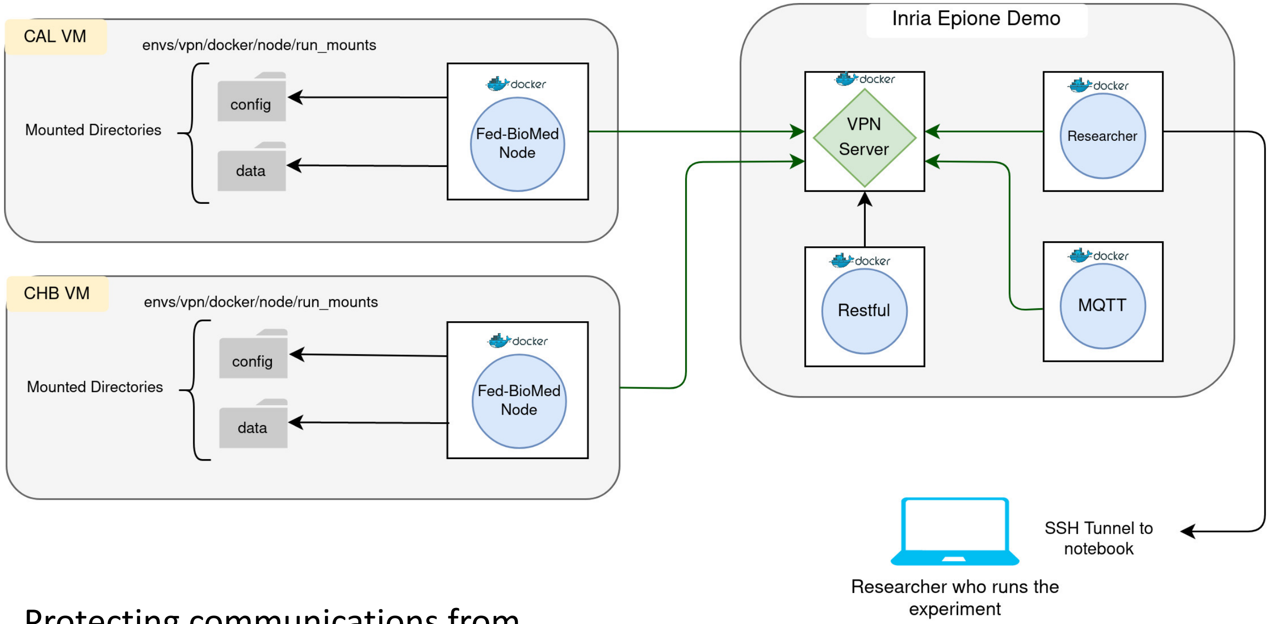
Low-level: Data Flow



Governance and security

No outbound connection

Low-level: VPN



Protecting communications from
- External attackers
- Internal attackers (man-in-the-middle)

Control

A piece of software is going to be executed on the client where private data is stored


```python
import torch
import torch.nn as nn
from fedbiomed.common.training_plans import TorchTrainingPlan
from fedbiomed.common.data import DataManager
from torchvision import datasets, transforms

# Here we define the model to be used.
# You can use any class name (here 'Net')
class MyTrainingPlan(TorchTrainingPlan):
    def __init__(self):
        super(MyTrainingPlan, self).__init__()
        self.conv1 = nn.Conv2d(1, 32, 3, 1)
        self.conv2 = nn.Conv2d(32, 64, 3, 1)
        self.dropout1 = nn.Dropout(0.25)
        self.dropout2 = nn.Dropout(0.5)
        self.fc1 = nn.Linear(9216, 128)
        self.fc2 = nn.Linear(128, 10)

        # Here we define the custom dependencies that will be needed by our custom Dataloader
        # In this case, we need the torch DataLoader classes
        # Since we will train on MNIST, we need datasets and transform from torchvision
        deps = ["from torchvision import datasets, transforms"]
        self.add_dependency(deps)

    def forward(self, x):
        x = self.conv1(x)
        x = F.relu(x)
        x = self.conv2(x)
        x = F.relu(x)
        x = F.max_pool2d(x, 2)
        x = self.dropout1(x)
        x = torch.flatten(x, 1)
        x = self.fc1(x)
        x = F.relu(x)
        x = self.dropout2(x)
        x = self.fc2(x)


        output = F.log_softmax(x, dim=1)
        return output

    def training_data(self, batch_size = 48):
        # Custom torch Dataloader for MNIST data
        transform = transforms.Compose([transforms.ToTensor(),
        transforms.Normalize((0.1307,), (0.3081,))])
        dataset1 = datasets.MNIST(self.dataset_path, train=True, download=False, transform=transform)
        train_kwargs = {'batch_size': batch_size, 'shuffle': True}
        return DataManager(dataset1, **train_kwargs)

    def training_step(self, data, target):
        output = self.forward(data)
        loss   = torch.nn.functional.nll_loss(output, target)
        return loss
```

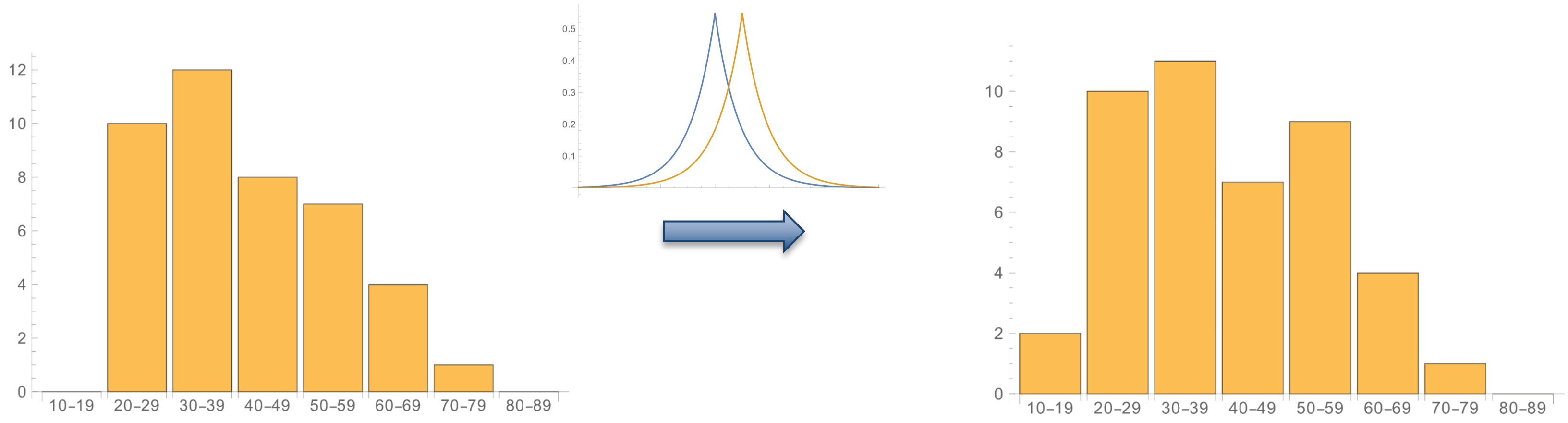
FL code reviewer expert in:

- **data science**
- **programming**
- **security**
- **networking**

ML-Level: Differential Privacy

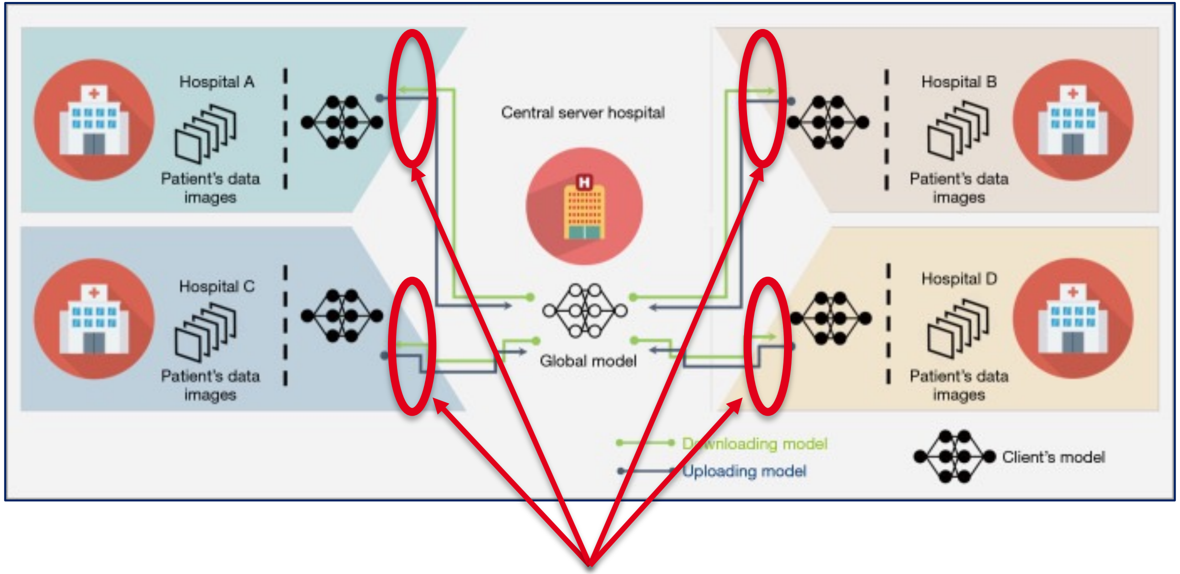
https://bit.ly/3lKJh4F
from desfontain.es

The attacker knows *almost all elements*
Identifying the contribution of a single person



Dwork, Cynthia, and Aaron Roth. "The algorithmic foundations of differential privacy", 2014

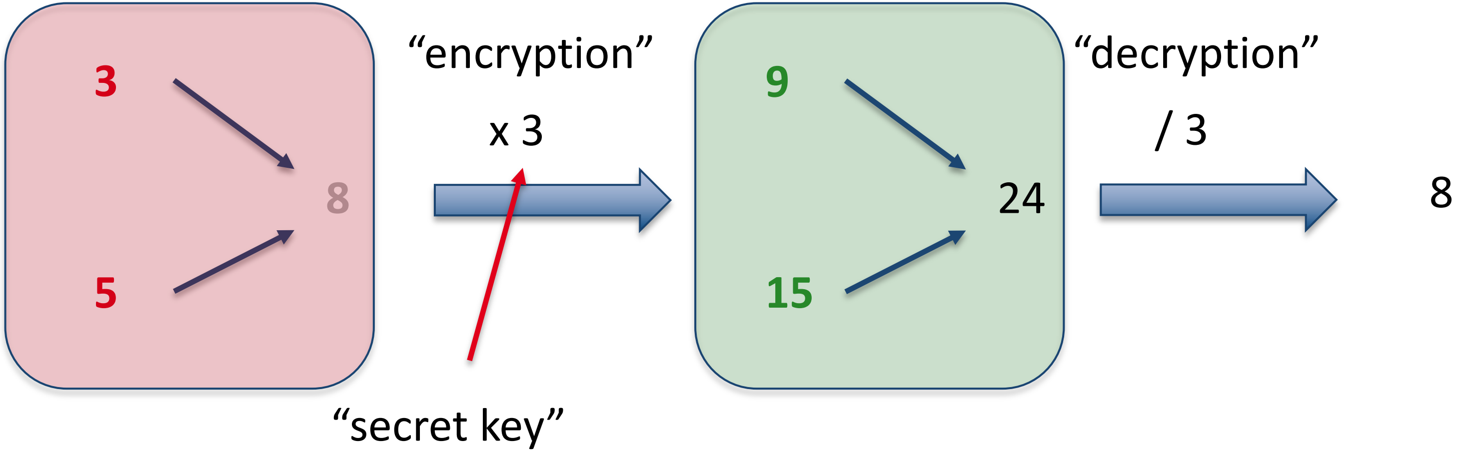
ML-Level: Differential Privacy



Noise is added to the model parameters before sharing with the server

ML-Level: Homomorphic Encryption

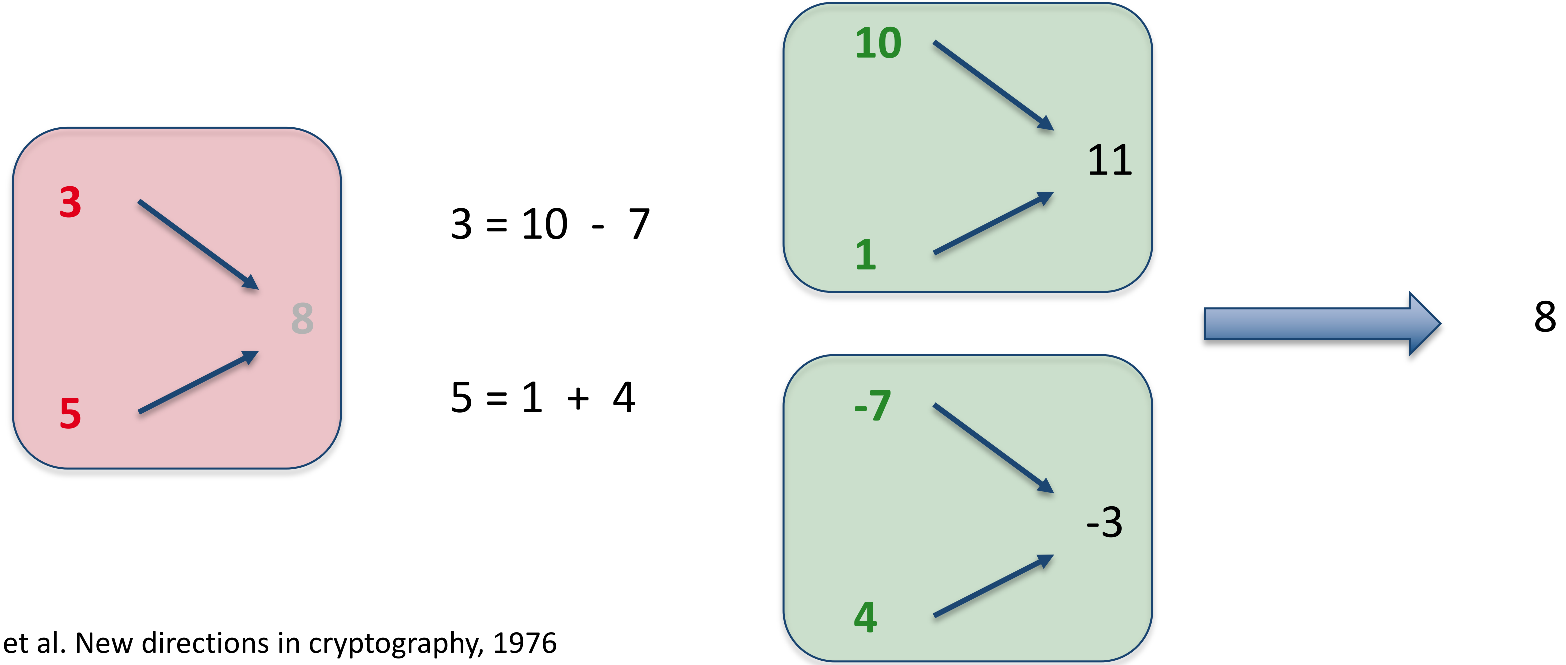
Encrypting the data compatibly with mathematical operations



Rivest, R. L., et al, On data banks and privacy homomorphisms, 1978
Gentry, C. *A fully homomorphic encryption scheme*. 2009
Yagisawa, M. Fully homomorphic encryption without bootstrapping. 2015

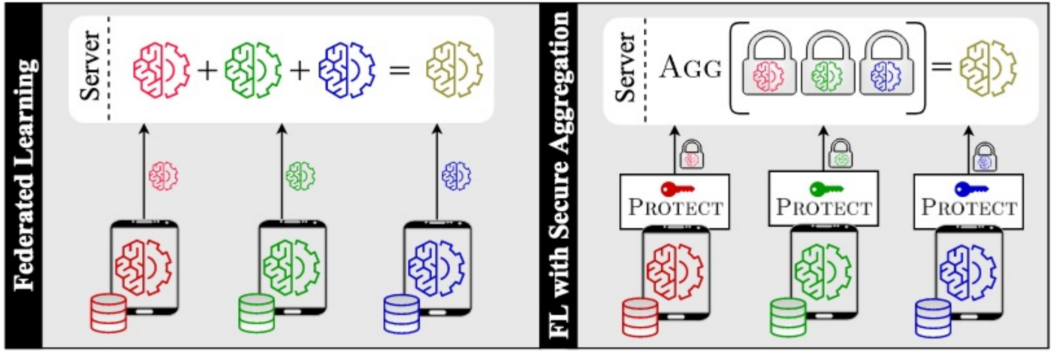
ML-Level: Multi-Party Computation

Encrypting the data compatibly with mathematical operations



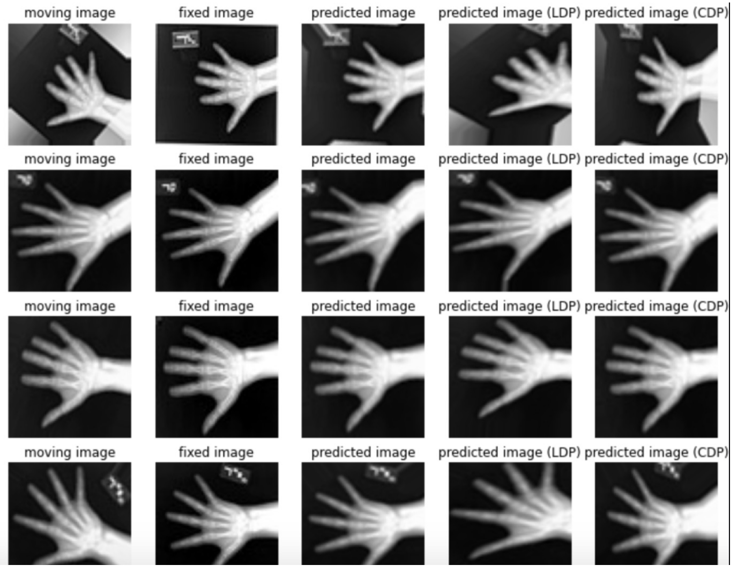
Diffie et al. New directions in cryptography, 1976
Shamir et al. How to share a secret. 1979
Yao et al. How to generate and exchange secrets. 1986

ML-Level: Secure Aggregation



From Mansouri, Önen, Jaballah, Proc ACM Conference. 2017

Challenges



DP

Utility vs Security

Data dependency

Communication and acceptance of DP

Secure Aggregation

Guarantees on trusted parties within FL network

Multi-key frameworks, key generation

Communication Cost

Computational cost

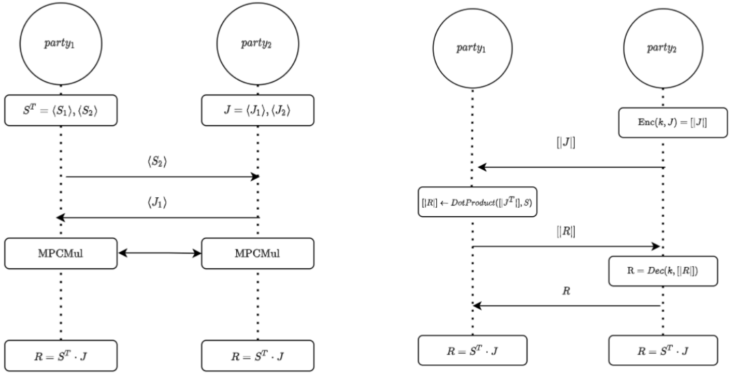
Limited operations

Privacy Tailored to Medical Applications

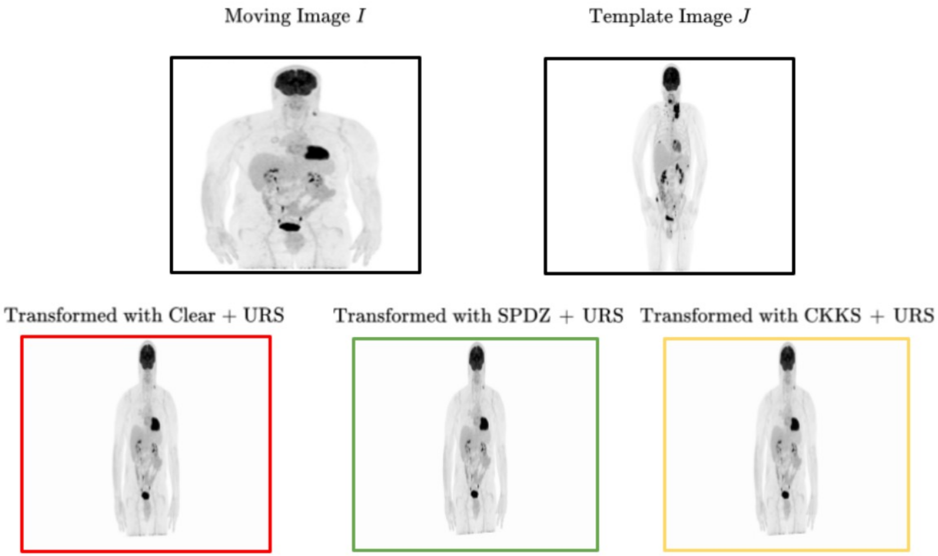
Privacy Preserving Medical Image Registration

$$\mathrm{SSD}(I, J, \mathbf{p}) = \mathrm{argmin}_{\mathbf{p}} \sum_{\mathbf{x}} \Big[I(\mathbf{W}_{\mathbf{p}}(\mathbf{x})) - J(\mathbf{x}) \Big]^2$$

$$\mathbf{\Delta p} = H^{-1} \cdot \sum_{\mathbf{x}} S(\mathbf{x}) \cdot (I(\mathbf{W}_{\mathbf{p}}(\mathbf{x})) - J(\mathbf{x}))$$



(a) Multi Party Computation (b) Fully Homomorphic Encryption



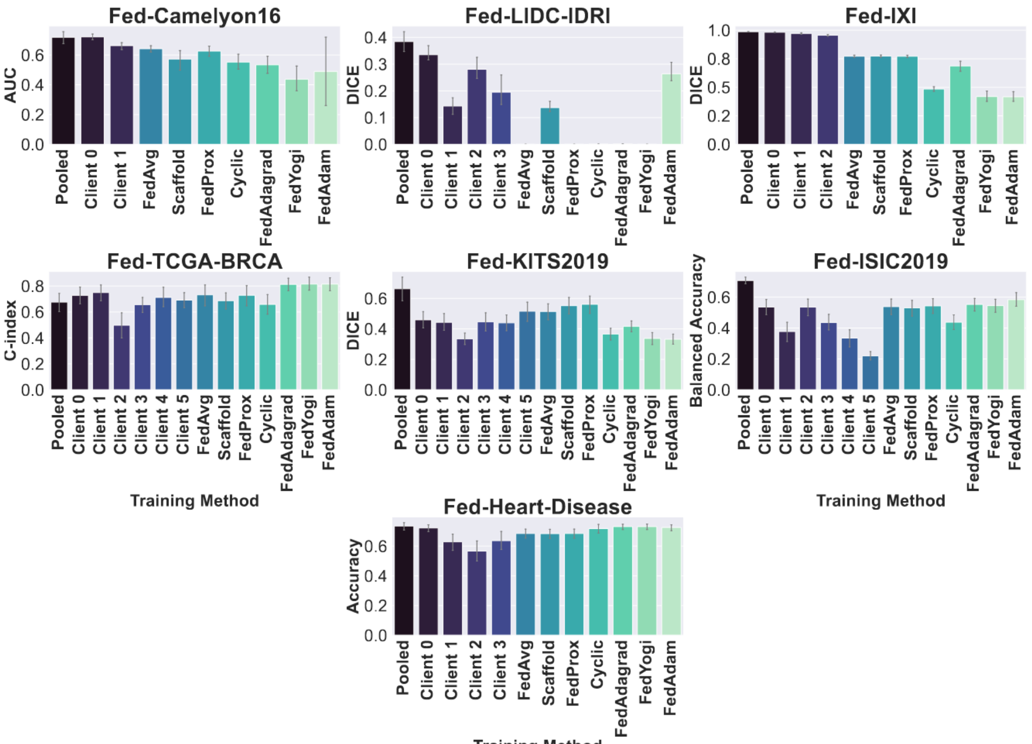
Ethical and Legal Questions

Who controls FL execution
Responsibilities for security breach
Who owns a FL infrastructure
Who owns the results
Reward scheme
Exploitation of model and results
Right to be forgotten: machine unlearning

Cost-Effectiveness

FLamby

https://github.com/owkin/FLamby



Plugin for interactive benchmarking through Fed-BioMed

Fed-BioMed

Federated Learning for Healthcare

https://fedbiomed.gitlabpages.inria.fr/

Security

- Clients authentication
- Secured communications
- Model verification
- Differential Privacy
- Secure aggregation (coming release)

Client control / Governance

- Experiment opt-in / -out
- Monitoring Tools
- Data verification/ pre-processing
- GUI
- Handling heterogeneous data types

Supported by

AGENCE NATIONALE DE LA RECHERCHE ANR

Inria

3iA Côte d'Azur Institut interdisciplinaire d'intelligence artificielle

UNIVERSITÉ CÔTE D'AZUR

INVESTIR L'AVENIR LE GRAND PLAN D'INVESTISSEMENT

INNOVATION ARCHITECTURE: ACCENTURE LABS

Usability

- Numpy/Pytorch/MONAI/sklearn compatible
- Easy control with Jupyter notebook
- Breakpoints and control of experiment
- Error handling
- FL aggregation and sampling strategies
- FL simulator

Community-driven

- Roadmap inspired by collaborating hospitals
- Long term planning and institutional support
- Commercial-friendly license (Apache 2.0)

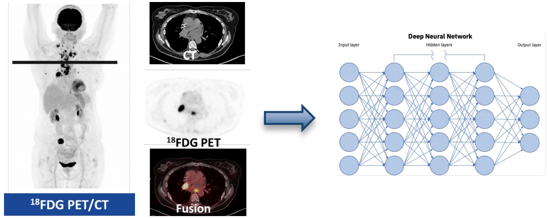
Clinical Coordinator

3IA chair Prof. O. Humbert
Centre Antoine Lacassagne

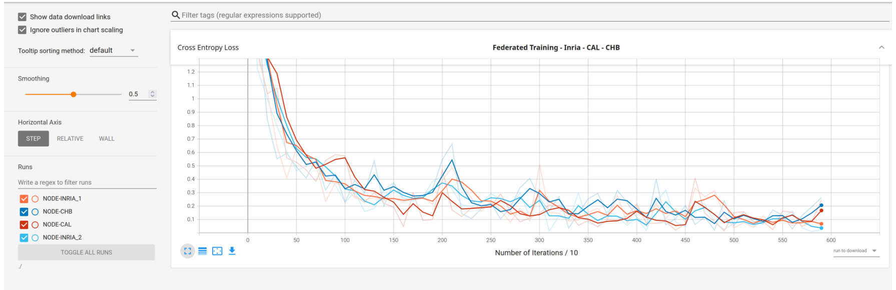
Real-world deployment

18 FDG-PET Analysis for Predicting Treatment Response in Lung Cancer

10 hospitals– 1000 patients



Humbert et al. Eur J Nucl Med Mol Imaging. 2020



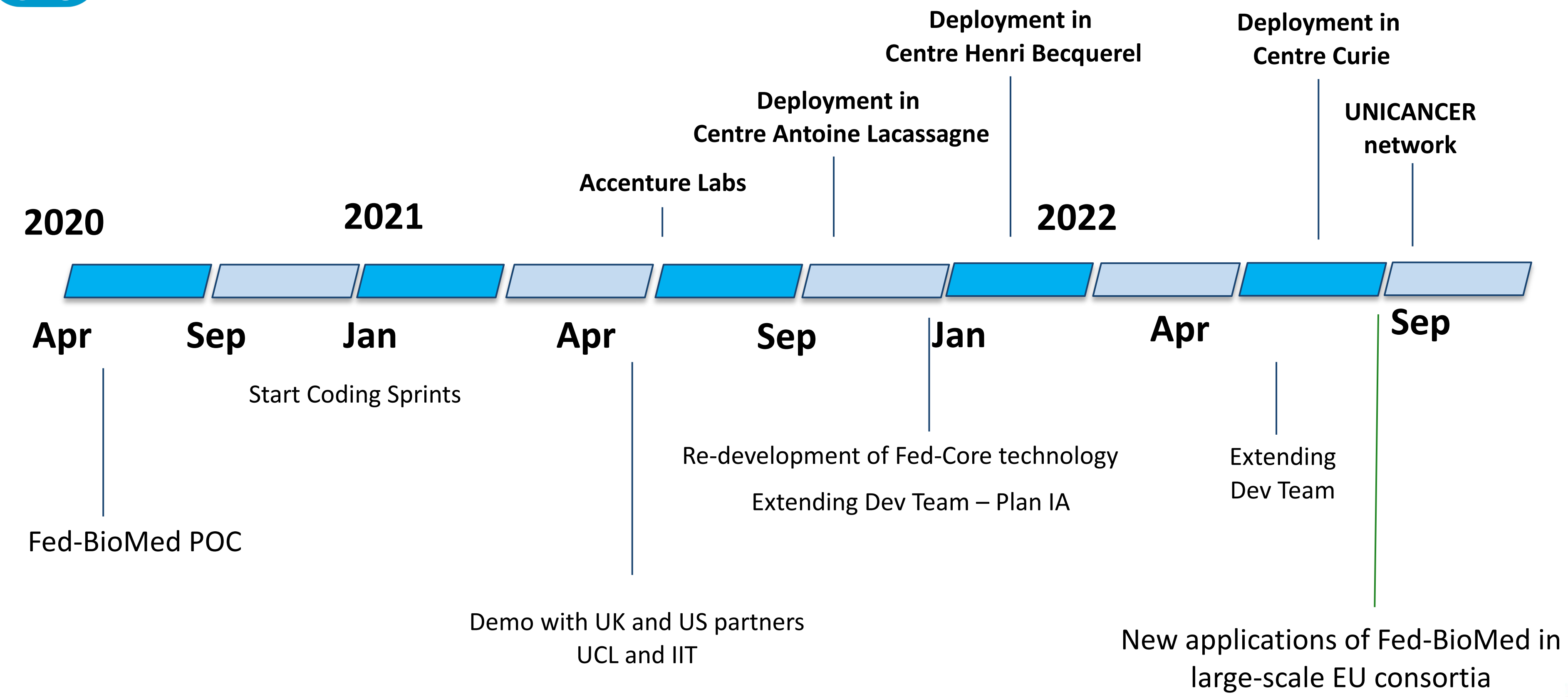
Multi-centric Neuroimaging Studies

Table 1. Demographics for each of the centers sharing brain-imaging data. MCI: Mild Cognitive Impairment; AD: Alzheimer's Disease.

	France 1	US	UK	France 2
No. of participants (M/F)	448/353	454/362	1070/930	573/780
Clinical status				
No. healthy	175	816	2000	695
No. MCI and AD	621	0	0	358
Age ± sd (range) [years]	73.74 ± 7.23	28.72 ± 3.70	63.93 ± 7.49	67.58 ± 10.04
Age range [years]	54 - 91	22 - 37	47 - 81	43 - 97

Silva, Altmann, Gutman and Lorenzi. *DECAF MICCAI Workshop*, 2020

Timeline


Deployment in
Centre Henri Becquerel
Deployment in
Centre Curie
Deployment in
Centre Antoine Lacassagne
UNICANCER
network
Accenture Labs
2020
2021
2022
Apr
Sep
Jan
Apr
Sep
Jan
Apr
Sep
Start Coding Sprints
Re-development of Fed-Core technology
Extending Dev Team – Plan IA
Extending
Dev Team
Fed-BioMed POC
Demo with UK and US partners
UCL and IIT
New applications of Fed-BioMed in
large-scale EU consortia

Samy Ayed
Irene Balelli
Francesco Cremonesi
Yann Fraboni
Santiago Silva
Riccardo Taiello

Olivier Humbert
Hamid Laceb

Fanny Orlhac
Christophe Nioche

Thanks!

Melek Önen
Mohamad Mansouri

Marc Vesin
Yannick Bouillard
Sergen Cansiz
Paul Andrey
Aurélien Bellet
Marc Tommasi
Jean-Luc Szpyrka
Thibaud Kloczko

Bastien Houis
Nathan Lapel
Romain Modzelewski

Laetitia Kameni
Richard Vidal