Fed-XAI:
Federated Learning of
Explainable Artificial Intelligence Models

Prof. Pietro Ducange

University of Pisa, Dept. of Information Engineering, Pisa, Italy

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

Outline

- Introduction: scenario and motivations
- Basics of Federated Learning
- Basics of XAI models
- Current status of ongoing works on FED-XAI
- Examples of FED-XAI Applications
- Open challenges

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

Pervasive IT Paradigms and Technologies

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab
Rethinking the future

CROSSLAB
Innovation for industry 4.0

AI in Automotive- Is it Safe?

- ***Why*** did the autopilot car crash?
- Did the braking systems go off to ***avoid*** the crash?
- Are we ***sure*** the intelligent vision system saw the obstacle?
- Why did the car turn unnecessarily?
- **Who** was ***responsible*** for the crash?

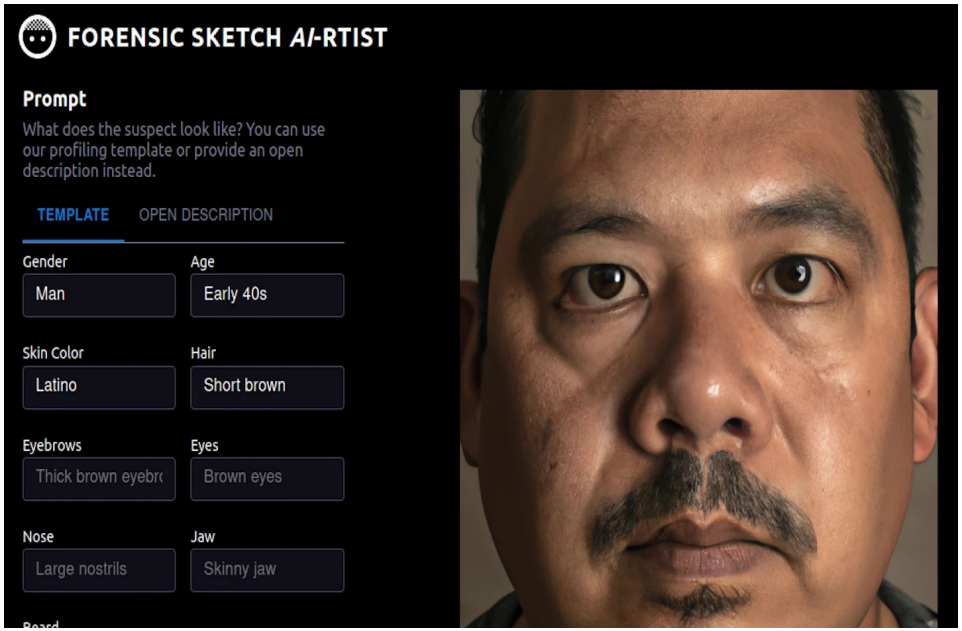
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

AI in Forensic Activities



*"AI **ethicists** and researchers told Motherboard that the use of generative AI in police forensics is incredibly dangerous, with the potential to worsen existing **racial** and **gender biases** that appear in initial witness descriptions."*

Image Extracted from: https://www.vice.com/en/article/qjk745/ai-police-sketches

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

AI in Health

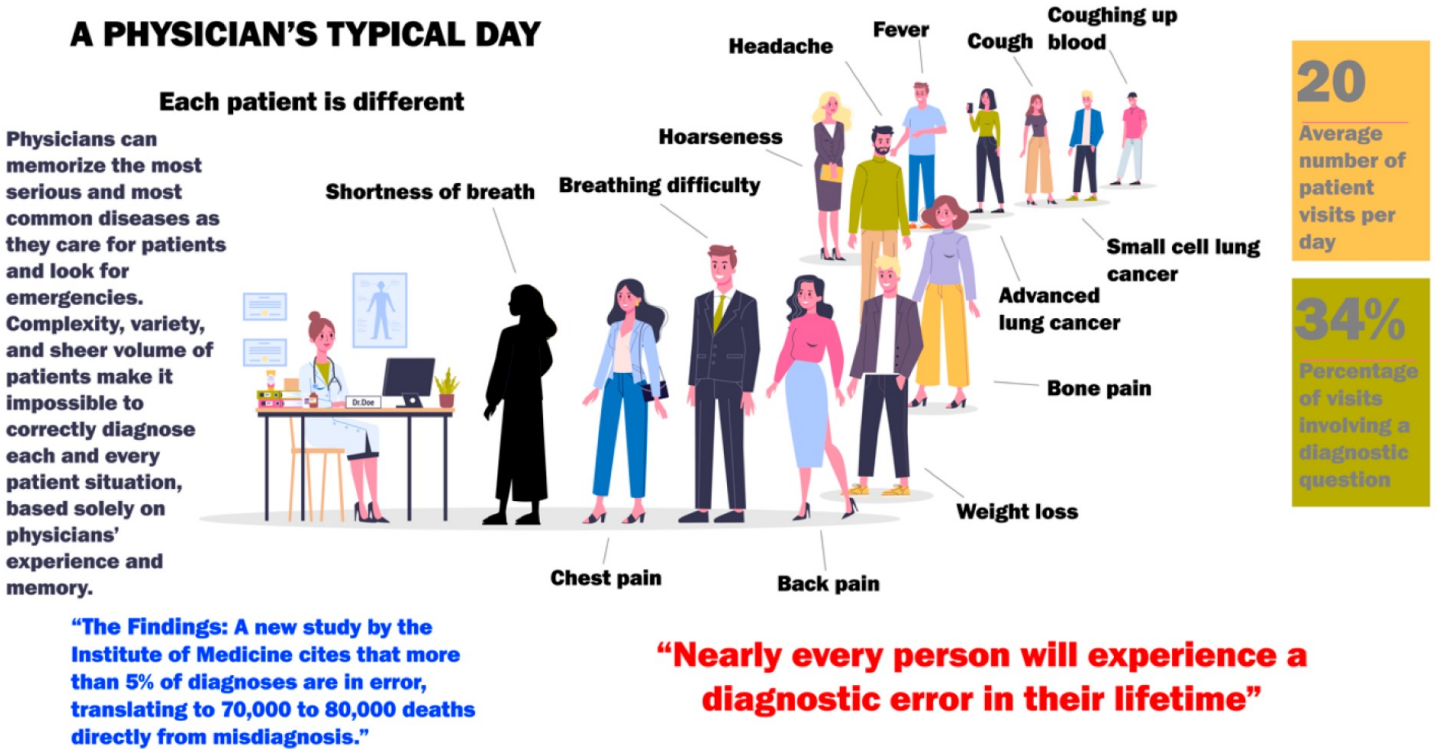


Image Extracted from: https://pubs.acs.org/doi/10.1021/acsnano.1c00085

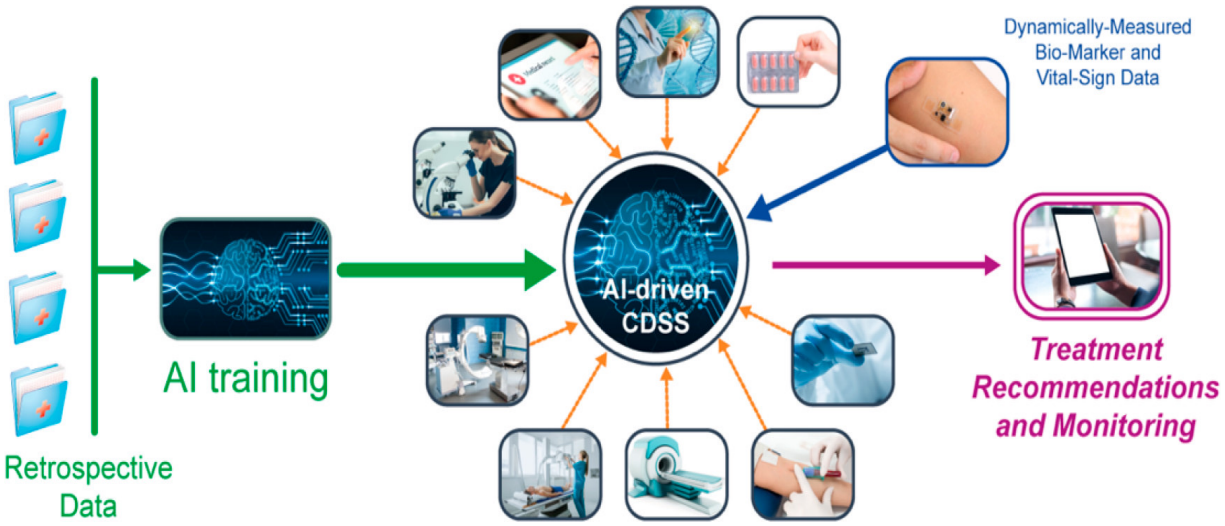
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

AI in Health



- ***Why*** was the cancer diagnosis made?
- ***How reliable*** is the diagnosis that was made?
- ***Why*** is this treatment suggested for my disease?
- ***What*** led to the ***conclusion*** that there is a serious illness ***rather than*** a stressful condition in the patient?
- Has patient ***privacy*** been respected?
- Were all ***ethical*** aspects taken into consideration ?

Image Extracted from: https://pubs.acs.org/doi/10.1021/acsnano.1c00085

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

The pursuit of *trustworthiness*

Human agency and oversight

Accountability

Technical robustness and safety

Societal and environmental wellbeing

Seven requirements toward Trustworthy AI

Diversity non-discrimination and fairness

Privacy and data governance

Transparency

Need to collect (large) data to train accurate ML models clashes with need to preserve privacy of data owners.

"AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned."

Fed-XAI

Federated Learning of eXplainable AI models

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

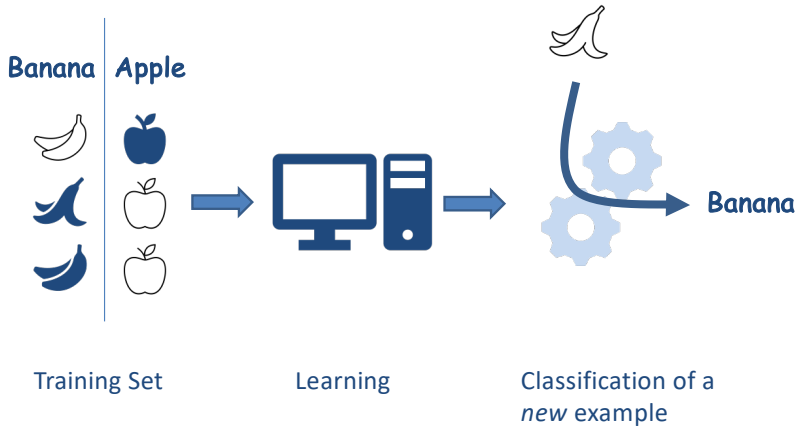
UNIVERSITÀ DI PISA

FoReLab Rethinking the future

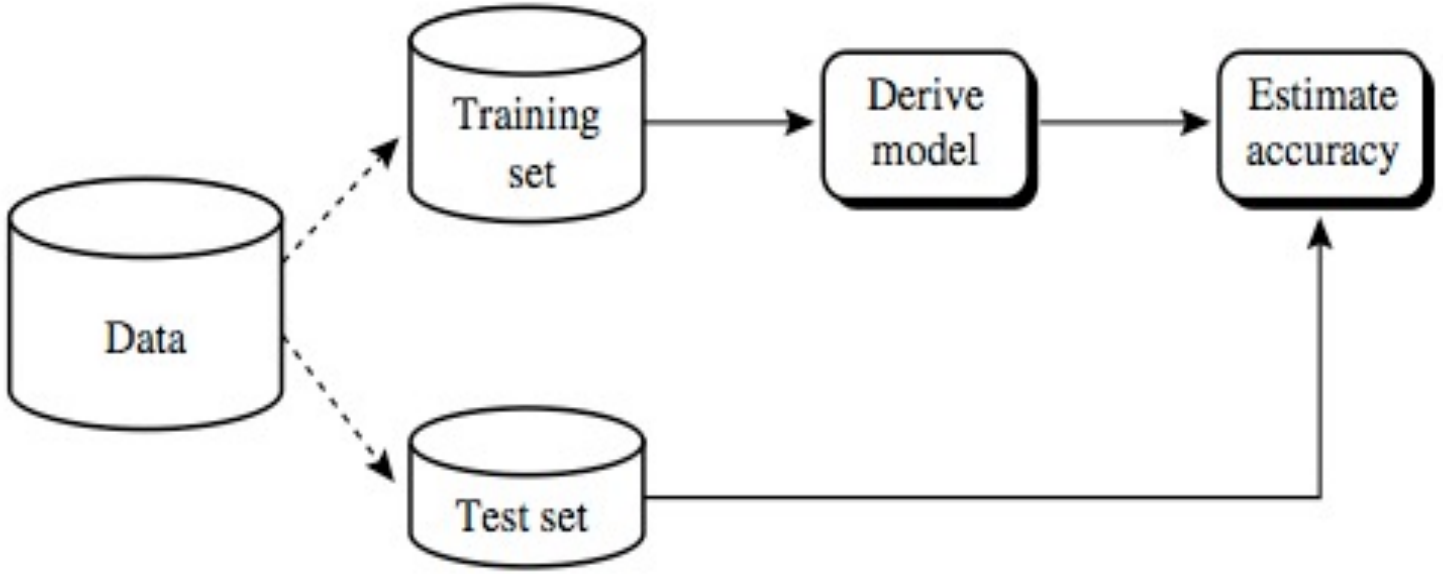
CROSSLAB Innovation for industry 4.0

Supervised Learning

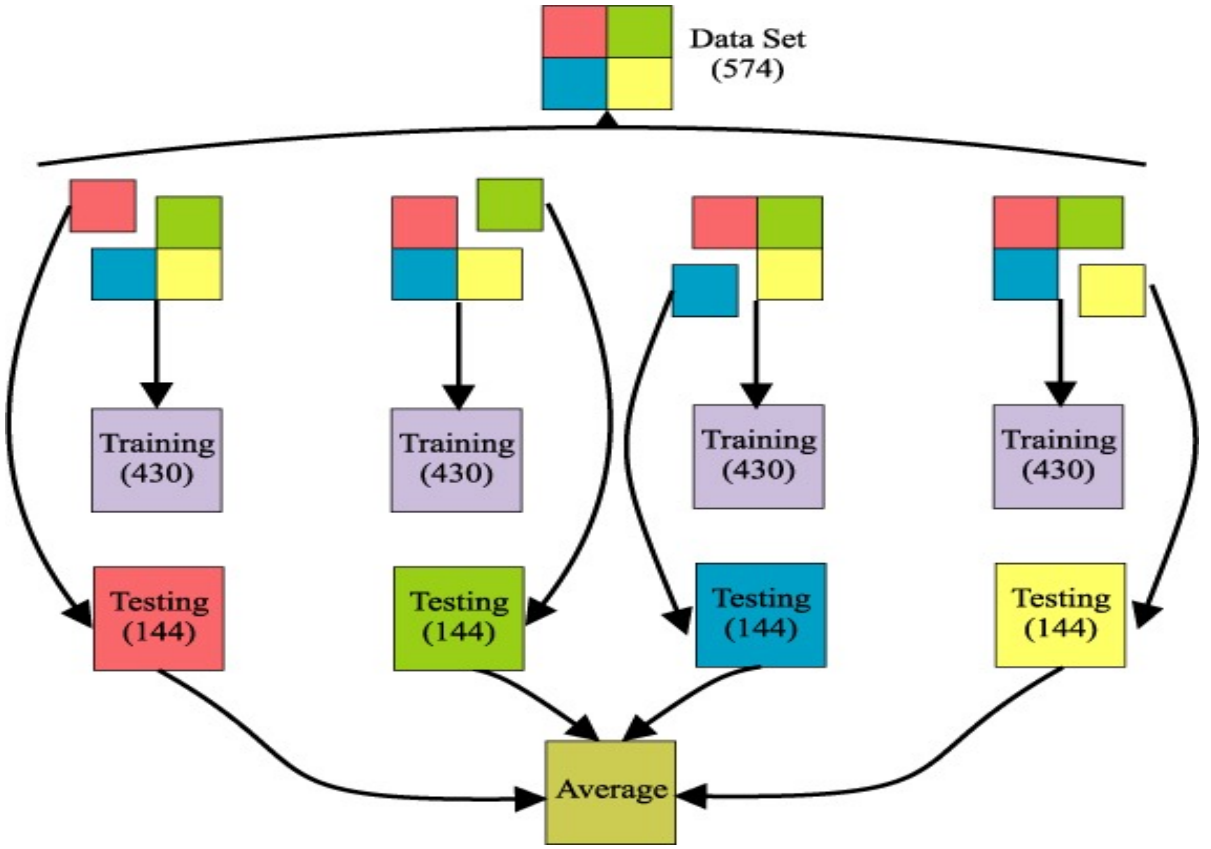
Given a **training set** of N example input-output pairs
$(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$
where each pair was generated by an unknown function $y = f(x)$,
discover a function h that approximates the function f.



Model Evaluation



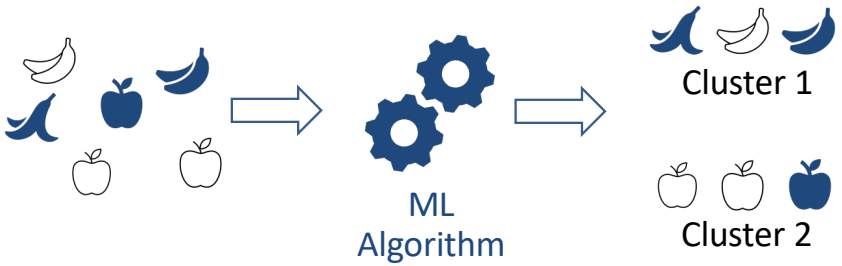
K-fold Cross Validation



Unsupervised Learning

Given a set of N input examples
$x_1, x_2, \ldots, x_N$
the *agent* learns a pattern in the input data,
without any explicit feedback

- **Clustering**
  - Most common *unsupervised learning* task
  - Group objects so that objects in a group…
    - … are *similar* to each other
    - … are *dissimilar* to objects in other groups



Why do we need Federated Learning?

- Machine learning algorithms, especially **deep learning algorithms**, are **data hungry.**
- **Data are generally spread** over different devices with different owners and under the protection of **privacy restrictions**.
- In practice, we cope with isolated **data islands** and we cannot transfer data

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

Federated Learning: basic concepts

Scenario:

- Preserve data privacy
- (im)possibility to share big data, leading to not-best scenarios for AI-models (often requiring large archive datasets to be effective):
  - Same organization, different departments
  - Different entities

- Federated Learning **idea**:
  - Raw data are not shared, but from each local model we share parameters/aggregate info toward a "global" federated model

    -> Overall information still available to the federated model
  - Finally, the model can be broadcasted to local areas

    -> More refined version available for each local area

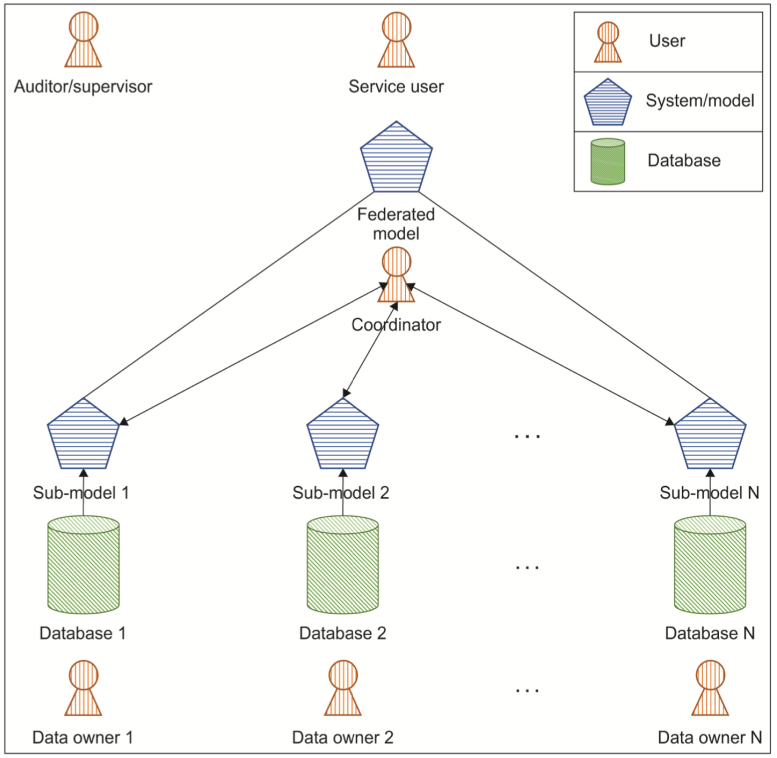


Image extracted from: IEEE 3652.1-2020 - IEEE Guide for Architectural Framework and Application of Federated Machine Learning

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

Federated Learning: basic concepts

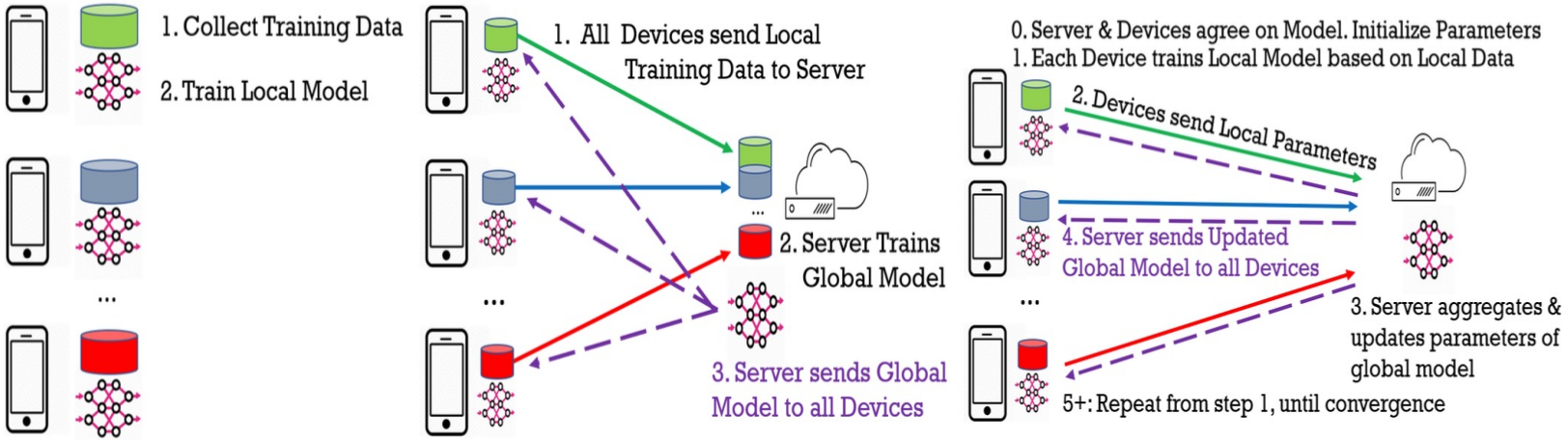


Image extracted from: Bakopoulou, E., Tillman, B., & Markopoulou, A. (2021). FedPacket: A Federated Learning Approach to Mobile Packet Classification. IEEE Transactions on Mobile Computing.

Local, centralized, federated models:

- **Local**: model likely to achieve low performance, but data privacy preserved
- **Centralized**: all data available, but privacy issues
- **Federated**: privacy preserved (raw data not shared, just model info) but all info available

Goal of federated learning: achieve performance as similar as possible than centralized learning.

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

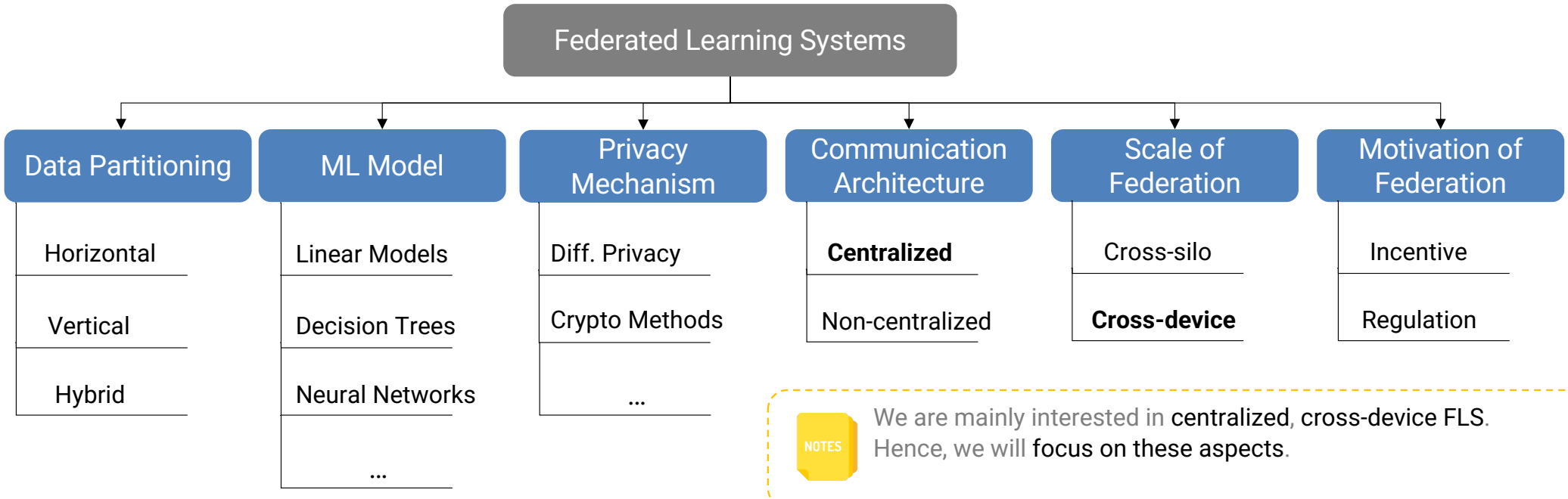
UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

Federated learning: a taxonomy

A Federating Learning System taxonomy according to six main characterizing aspects(*)



NOTES

We are mainly interested in centralized, cross-device FLS.
Hence, we will focus on these aspects.

(*)Li Q., Wen Z., Wu Z., Hu S., Wang N., Li Y., Liu X., He B.A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection(2021) IEEE Transactions on Knowledge and Data Engineering

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

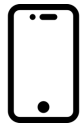
CROSSLAB Innovation for industry 4.0

Data Partitioning

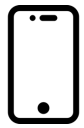
Based on how data are distributed among the parties making up the federation over the sample and feature spaces, FLSs can be typically categorized in horizontal, vertical and hybrid FLSs:

In Horizontal FL the datasets of different parties share the **same feature space** but have **little or no intersection on the sample space** (CT scans reported in different hospitals).

This is a natural data partitioning especially for the **cross-device setting**, where different users try to improve their model performance on the same task using FL.



Device #1	Feature 1	...	Feature N
Sample 1			
Sample 2			
Sample 3			



Device #2	Feature 1	...	Feature N
Sample 4			
Sample 5			

Device #3	Feature 1	...	Feature N
Sample 5			
Sample 6			
Sample 7			

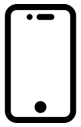
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

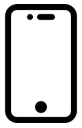
Data Partitioning

In Vertical FL the datasets of different parties have the **same or similar sample space** but **differ in** the **feature space** (for instance, municipality registry and hospital data).

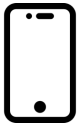
Device #1	Feature 1	Feature 2
Sample 1		
Sample 2		
Sample 3		
Sample 4		
Sample 5		
Sample 6		



Device #2	Feature 3	Feature 4	Feature 5
Sample 1			
Sample 2			
Sample 3			
Sample 4			
Sample 5			
Sample 6			



Device #3	Feature 6	Feature 7
Sample 1		
Sample 2		
Sample 3		
Sample 4		
Sample 5		
Sample 6		



In Hybrid FL partition of data among the parties may be **a hybrid of horizontal partition and vertical partition**.

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

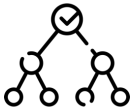
FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

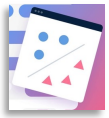
ML Models

There have been many efforts in developing new models or reinventing current models to the federated setting. For the sake of brevity **we briefly cite the widely-used models nowadays**:

- Neural Networks: there are many studies on **federated stochastic gradient descent** which can be used to train NNs.



- Decision tree is another widely used model as it is highly efficient to train compared with NNs. (FLSs studies for Gradient Boosting decision trees - GBDTs have been proposed recently).



- SVM: there exist a number of examples in which SVM is successfully trained exploiting a federated stochastic gradient descent algorithm.

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

Privacy Mechanisms

Model parameters exchanged during FL rounds may leak sensitive information about the data. Beyond attacks targeting user privacy, there are also other classes of attacks on federated learning (e.g. an adversary might attempt to bias the model to produce inferences that are preferable to the adversary and much else).

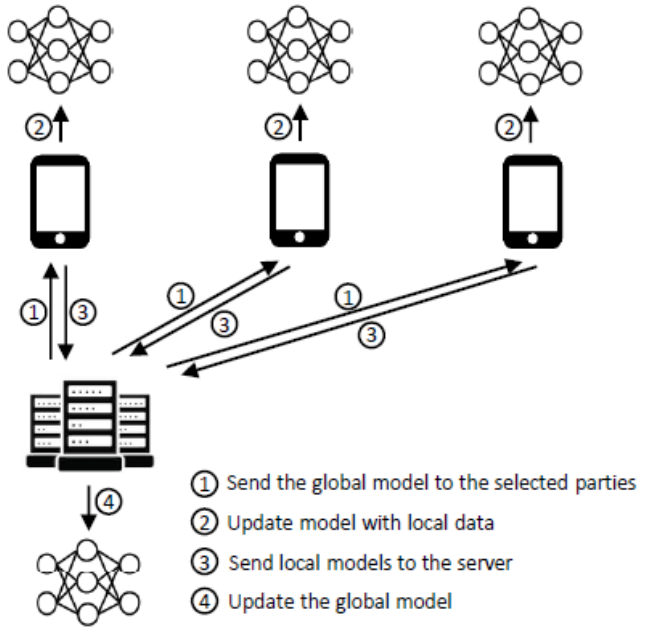
Technology	Main characteristics
Differential Privacy	Add properly tuned random noise to mask the influence of an individual instance on the output.
Secure Multi-Party Computation	Enables two or more parties to compute an agreed-upon function of their private inputs in a way that only reveals the intended output to each of the parties, while keeping those inputs private.
Homomorphic Encryption	Enables parties to perform mathematical operations directly on encrypted data without decrypting them.
Trusted Execution Environments	TEEs provide the ability to trustably run code on a remote machine, even if you do not trust the machine's owner. TEEs may provide confidentiality, integrity and remote attestation.

Communication Architecture

Centralized versus non-centralized

In the **centralized architecture** the **data flow** is **asymmetric**: the **server aggregates** the information (e.g. gradients or model parameters) from the clients and **sends them back** the **updated global model**.

The **process is executed iteratively until** a **convergence** criterion is met.



DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

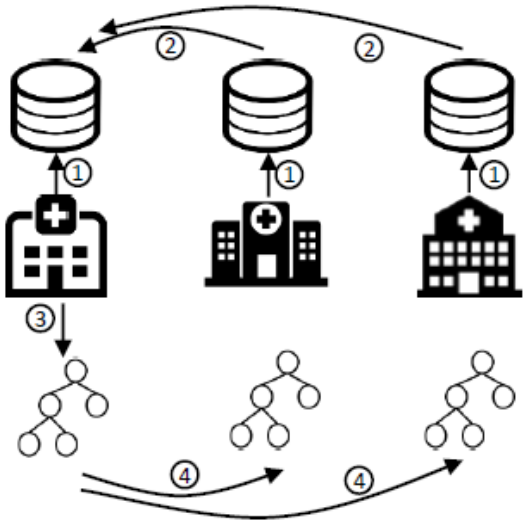
CROSSLAB Innovation for industry 4.0

Communication Architecture

Centralized versus non-centralized

In the **non-centralized architecture** the **communications** are **performed among the parties** and every party is able to **update** the **global parameters directly**. There is **no need for a trusted central aggregating server**.

In the **non-centralized architecture** the **major challenge** is that it is hard to design a protocol that treats every member almost fairly with reasonable communication overhead.



① Update the local gradients
② Send the gradients to the selected party
③ Update model with local data and all gradients
④ Send the model to the other parties

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE
UNIVERSITÀ DI PISA
FoReLab Rethinking the future
CROSSLAB Innovation for industry 4.0

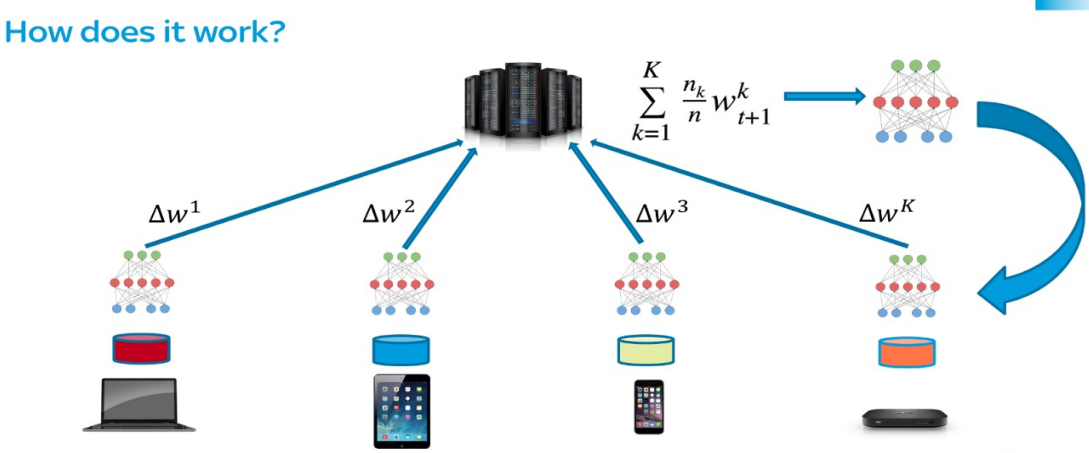
Scale of Federation

FLSs can be categorized into **two typical types** by the scale of federation: cross-silo FLSs and cross-device FLSs. The **main differences** between them lie on the **number of parties** and the **amount of data stored in each party**.

	Setting	Data Distrubution	Data Availability	Distribution Scale	Primary Bottleneck	Client reliability	Data partition axis
Cross-silo	Training a model on siloed data. **Clients are different organizations** or geo-ditributed datacenters.	**Data is generated locally and remains decentralized**. Each client stores its own data and cannot read the data of other clients.	**All clients are almost always available**.	**Typically, 2 – 100 clients**.	Might be **computation** or **communication**.	**Relatively few failures**.	Partition is fixed. Could be example-partitioned (horizontal) or feature-partitioned (vertical).
Cross-device	The clients are a **very large number of mobile or IoT devices**.		**Only a fraction of clients are available** at any one time, often with diurnal or other variations.	**Massively parallel, up to 10^{10} clients**.	**Communication** is often the **primary bottleneck**. Generally, cross-device computations use **wi-fi or slower connections**.	**Highly unreliable – 5% or more of the clients participating in a round of computation are expected to fail** or drop out.	Fixed partitioning by example (horizontal)

Popular approaches for Federated Learning

***Federated Stochastic Gradient Descent (FedSGD) vs Federated averaging (FedAVG)**:*



Federated Learning (Source: https://proandroiddev.com/federated-learning-e79e054c33ef)

- C = fraction of clients that participates in each federated round
- K = total number of clients (indexed by *k*)
- E = number of training passes each client makes over its local dataset on each round
- B = local minibatch size used for the client updates
  (B = ∞ indicates that the full local dataset is treated as a single minibatch
- P_k = set of indexes of data points on client *k*, with $n_k = |P_k|$

In **FedSGD** each client *k* computes the gradient on its local data at the current model w_t and the central server aggregates these gradients and updates the global model.

In **FedAVG** each client locally takes one or multiple steps of gradient descent on the current model w_t using its local data, and the server then takes a weigthed average of the resulting models.

Commonly used FL frameworks

Interest of the community, starting from Google Inc, prompted different frameworks

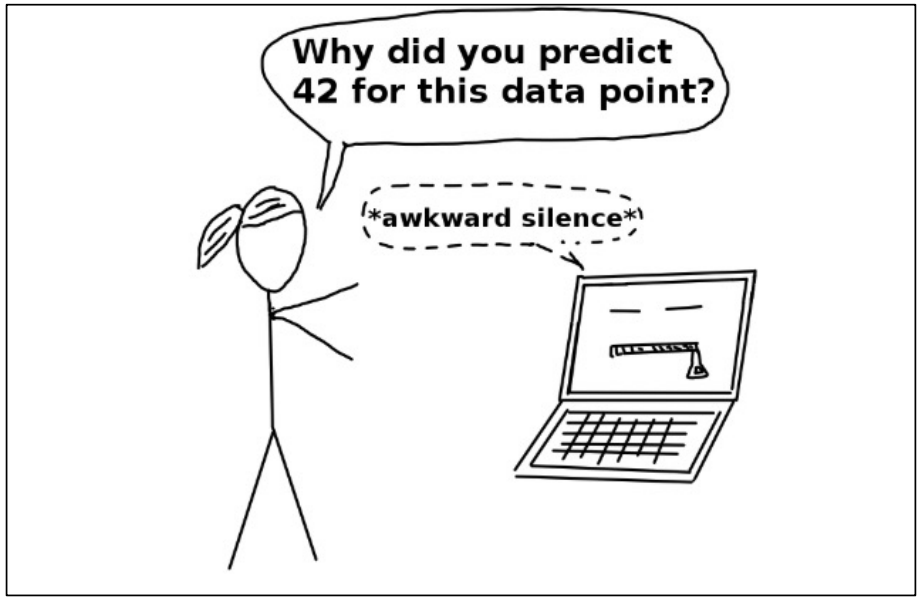
Room for improvement given by:
- number/kind of aggregation strategies
- number/kind on models available (DNN, Trees...)
- privacy level (protocols and encryption methods)
- interoperability (interfaces with other frameworks, APIs...)
- deployment in production stage

Our choice:
OPENFL ("easy" to customize ML models and aggregation methods, supports virtualization with containers)

Framework	Developers	URL
TFF	Google Inc	https://www.tensorflow.org/federated
FATE	Webank	https://fate.fedai.org
OpenFL	Intel Labs - University of Pennsylvania	https://github.com/intel/openfl
PySift	Openmined	https://github.com/OpenMined/PySyft
IBM FL	IBM	https://github.com/IBM/federated-learning-lib
Flower	Adap GmbH - several universities	https://flower.dev
FLSim	Facebook Research	https://github.com/facebookresearch/FLSim

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE
UNIVERSITÀ DI PISA
FoReLab Rethinking the future
CROSSLAB Innovation for industry 4.0

Explainable Artificial Intelligence (XAI)



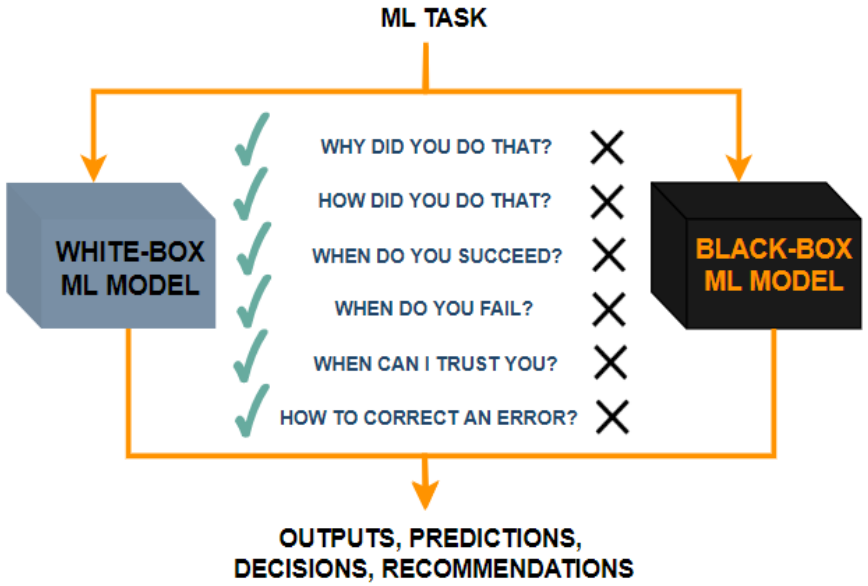


Figure from https://gowrishankar.info/blog/causal-reasoning-trustworthy-models-and-model-explainability-using-saliency-maps/

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

WHY do we need XAI Models?

- Explainabilty is motivated due to ***lacking transparency*** of the black-box approaches, which do not foster trust and acceptance of AI models.

- Rising ***legal*** and ***privacy*** aspects will make black-box approaches difficult to use in Business, because they often ***are not able to explain*** why a machine ***decision*** has been made.

- We should explain:
  - Why/How a decision has been made
  - Why not something else has been suggested
  - Why/When the decision may led the user to a failure
  - How to correct an error in making a decision

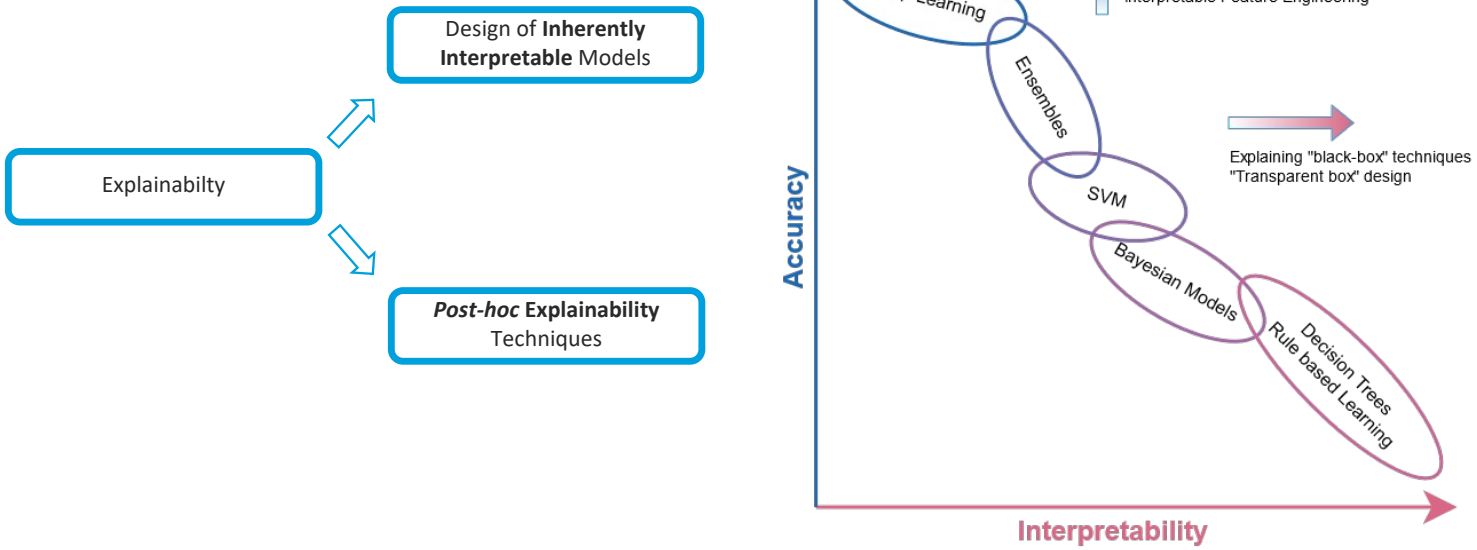
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

The Accuracy-Explainability Trade-off



Inspired to Arrieta et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." Information Fusion 58 (2020): 82-115.

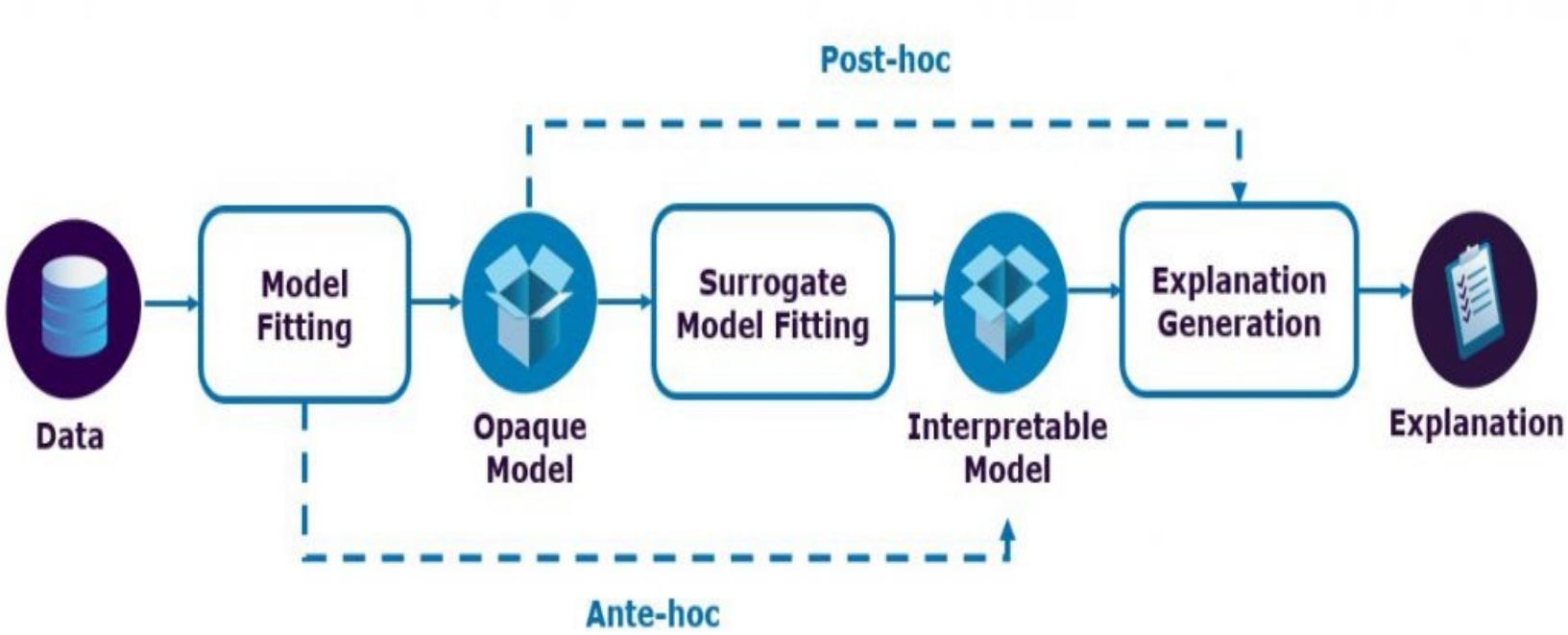
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

Ante-hoc vs Post-hoc



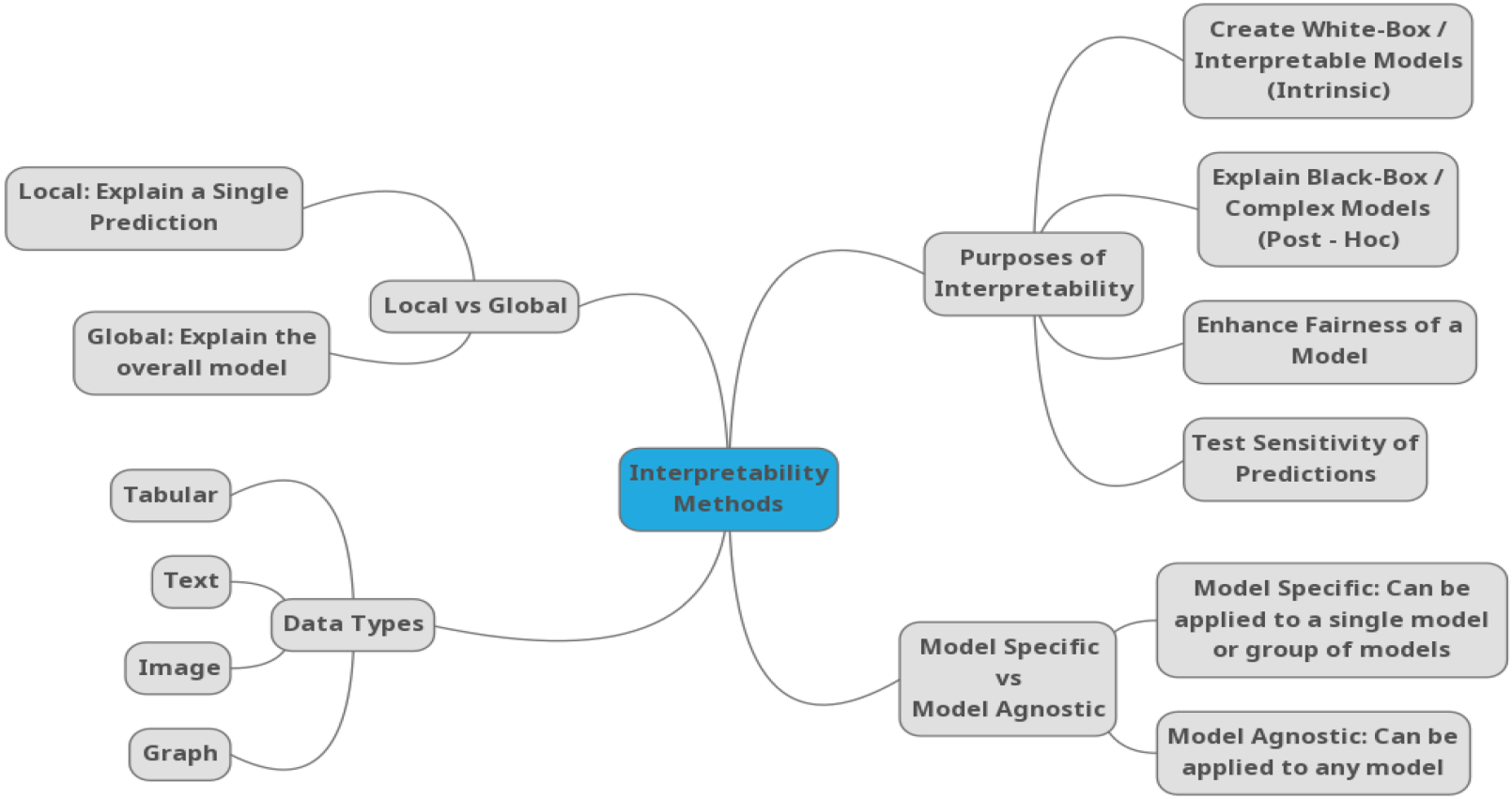
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

Yet Another XAI Taxonomy



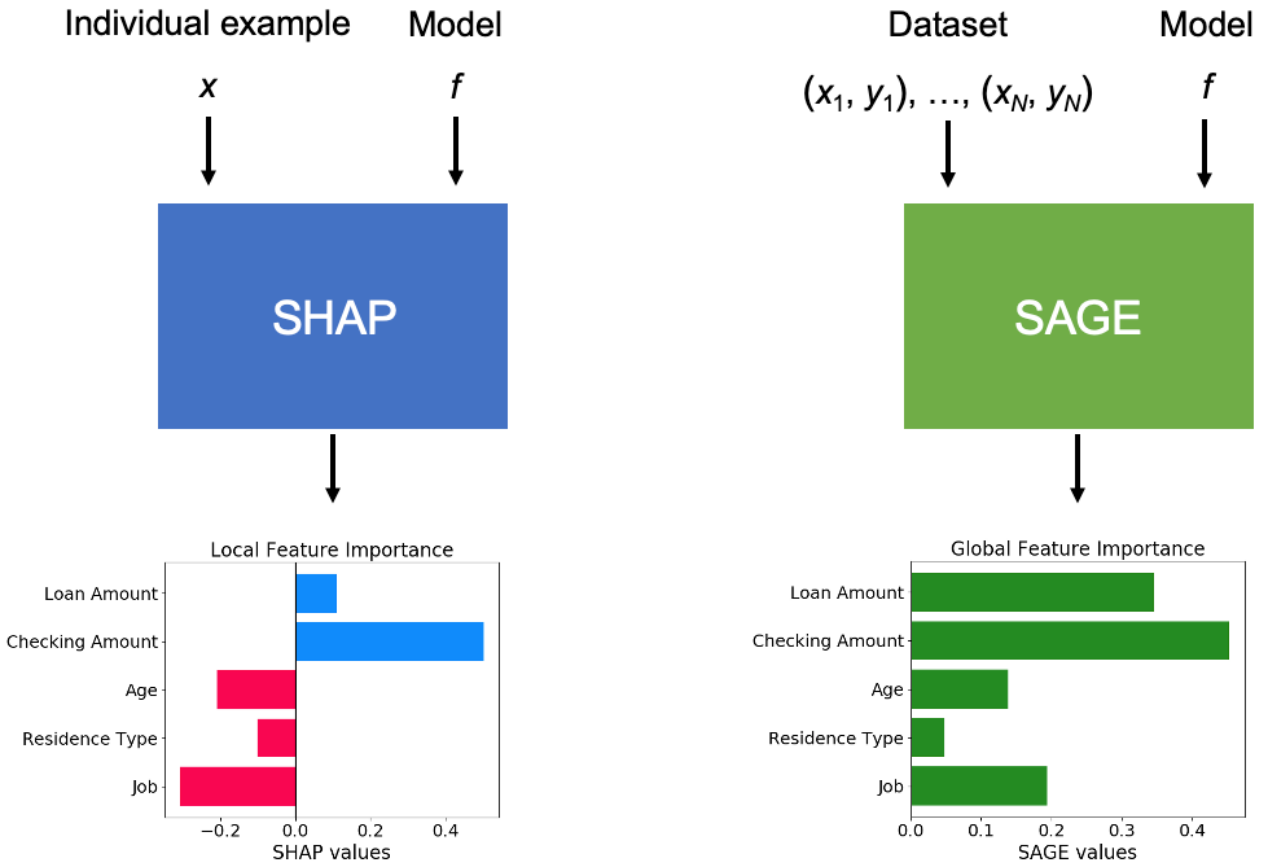
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

Post-hoc XAI



DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

Post-hoc XAI: An Example

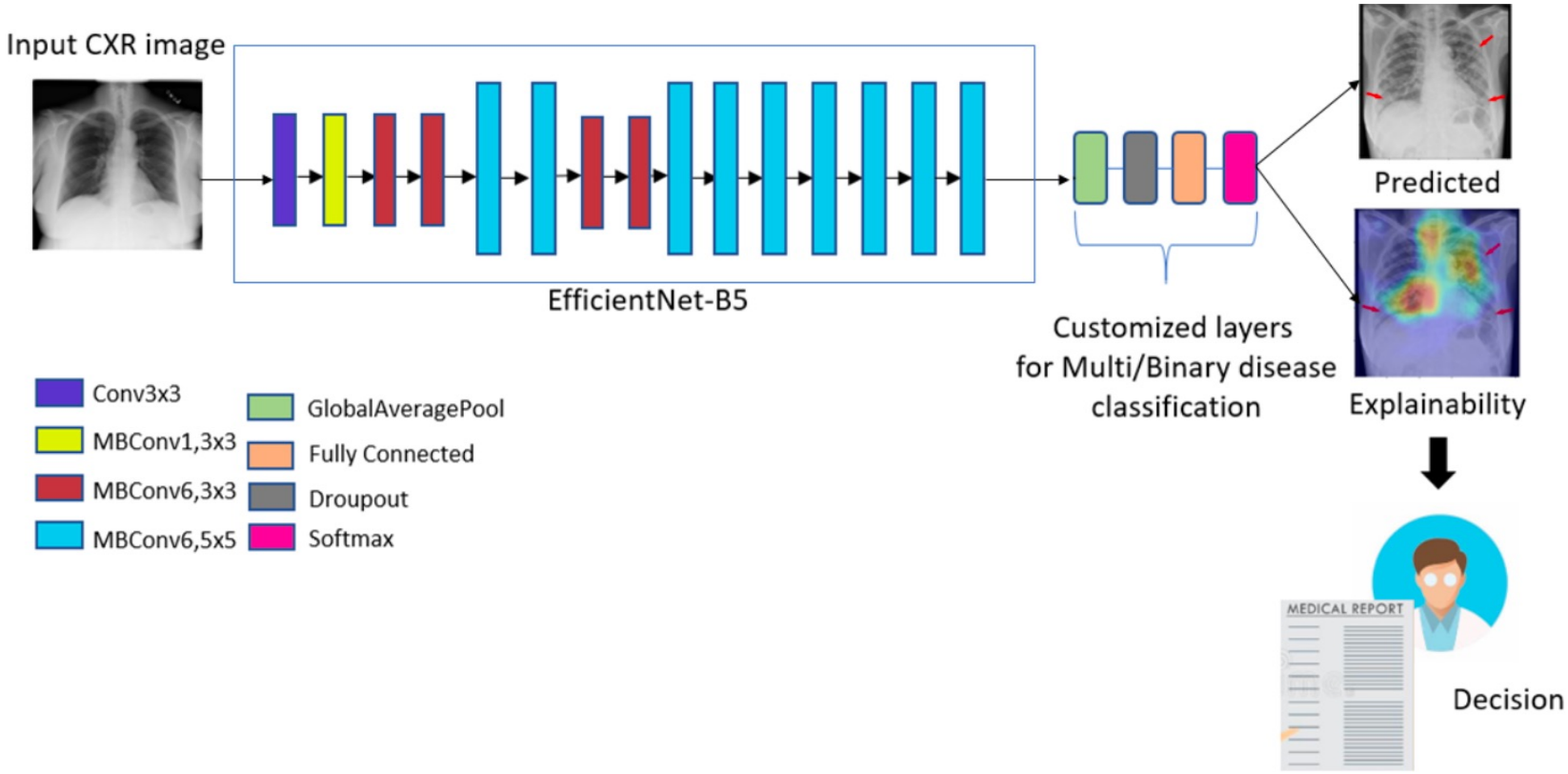


Image Extracted from: https://www.mdpi.com/2504-2289/5/4/73#

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

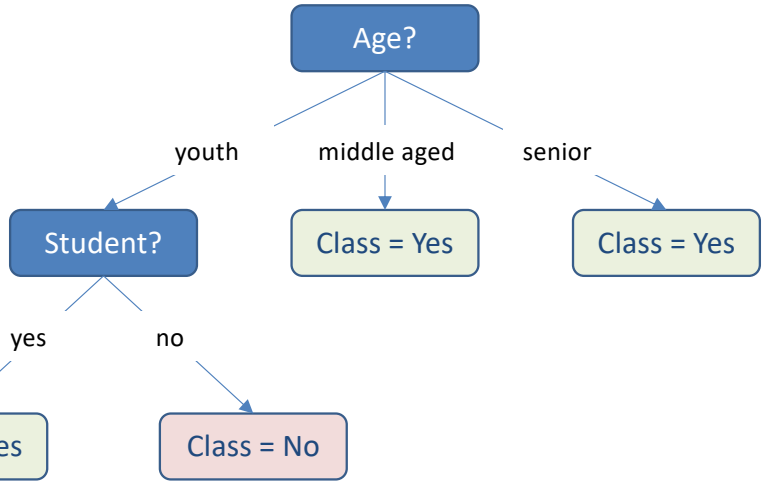
CROSSLAB Innovation for industry 4.0

XAI models by Design: Decision Trees

Training set

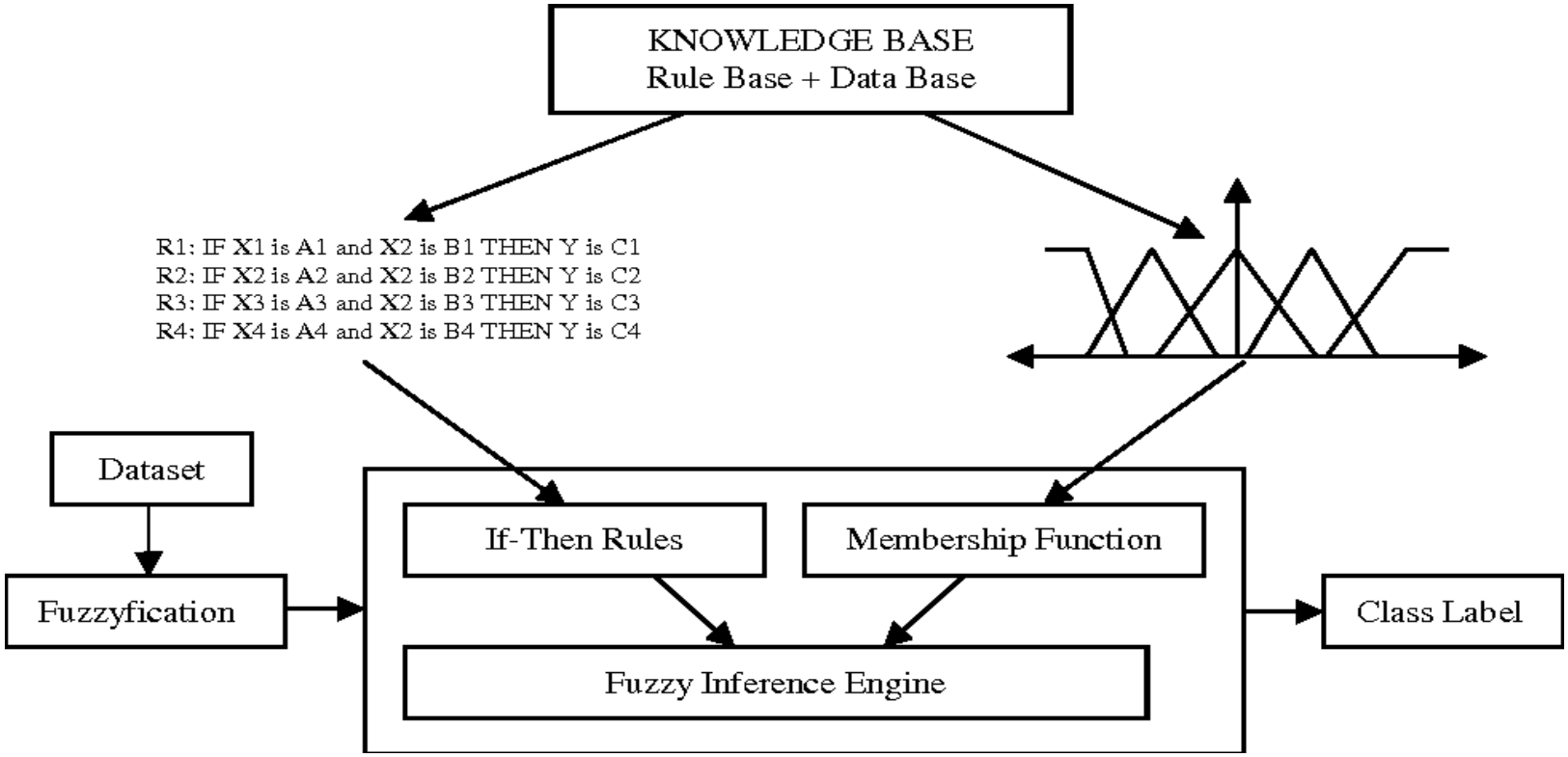
ID	Age	Income	Student	CreditRating	Buys PC
1	*youth*	*high*	*no*	*fair*	*no*
2	*youth*	*high*	*no*	*excellent*	*no*
3	*middle aged*	*high*	*yes*	*fair*	*yes*
4	*middle aged*	*medium*	*no*	*excellent*	*yes*
5	*youth*	*medium*	*yes*	*excellent*	*yes*
6	*senior*	*medium*	*yes*	*fair*	*yes*
7	*senior*	*low*	*yes*	*fair*	*yes*
8	*youth*	*medium*	*no*	*fair*	*no*
9	*youth*	*low*	*yes*	*fair*	*yes*

- Flowchart-like tree structure where
  - An **internal node** denotes test on an attribute
  - A **branch** represents the outcome of the test
  - A **leaf node** holds a class label



Learning from data:
Building the Decision Tree

XAI models by Design: (Fuzzy) Rule-based Systems



DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

Surrogate XAI Models

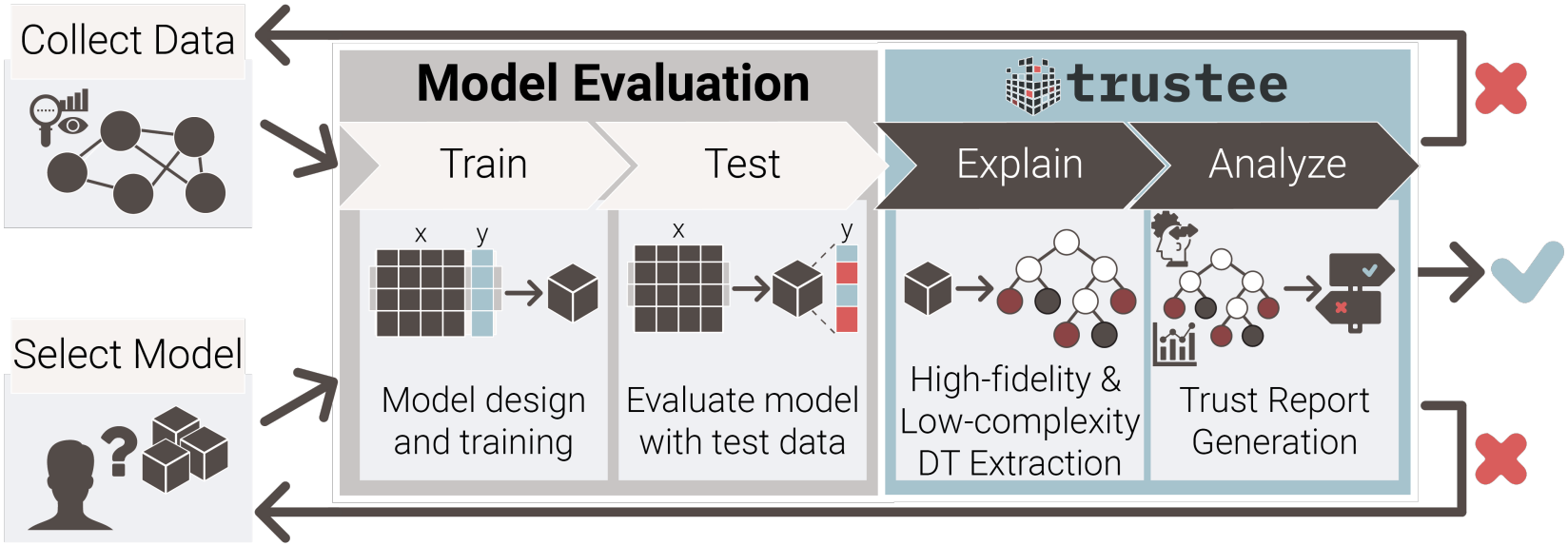


Image extracted from: https://trusteeml.github.io

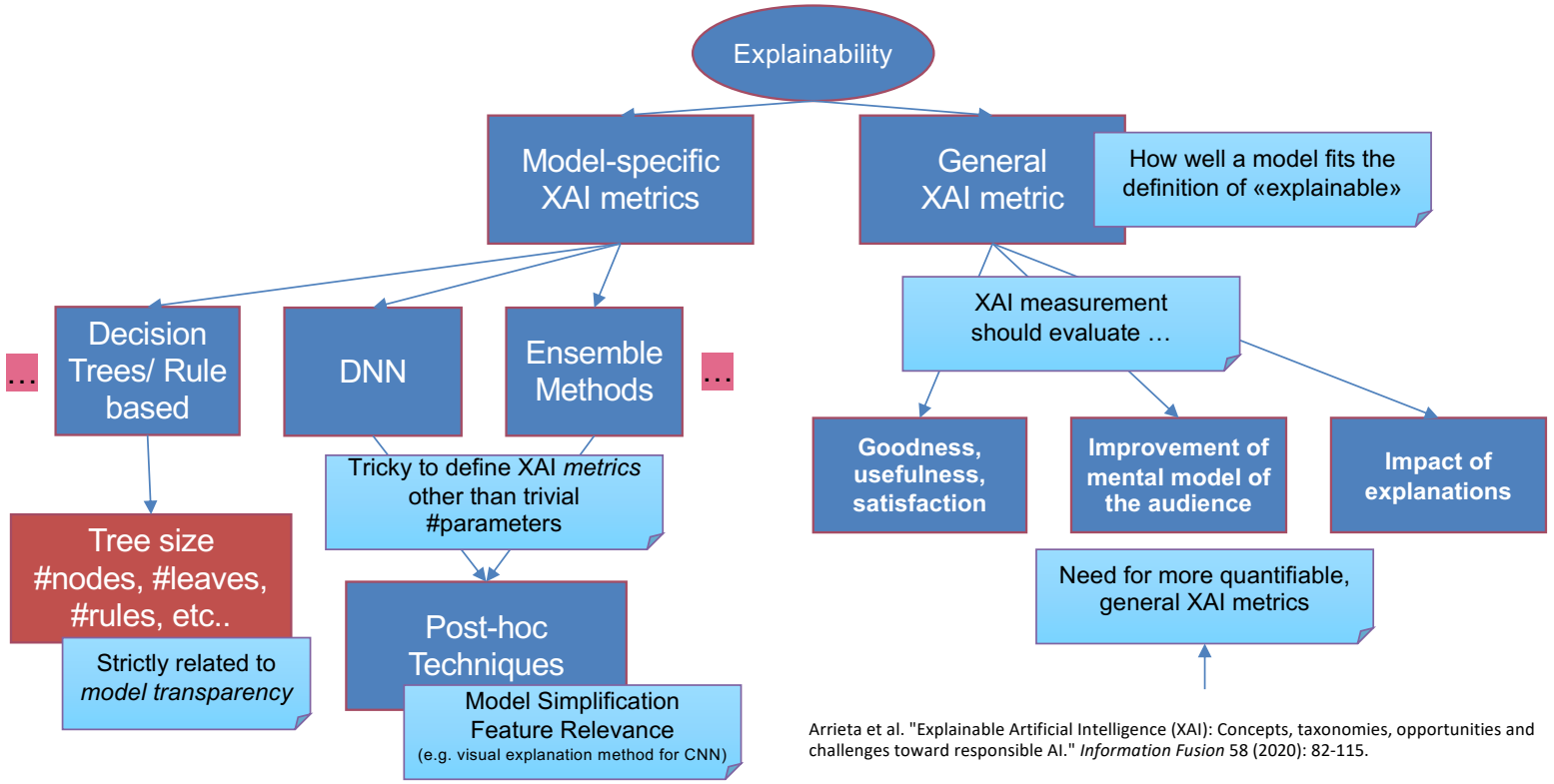
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

Measuring Explainability



Arrieta et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." *Information Fusion* 58 (2020): 82-115.

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

General Metrics for Explainability

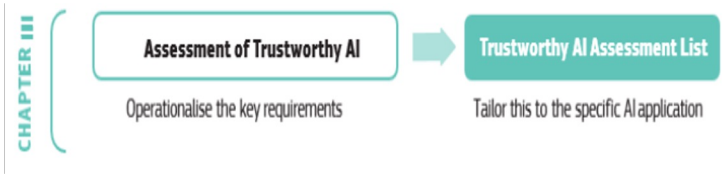


Image and text extracted from:
European Commission "Ethics guidelines for trustworthy AI", Report, 2019

1. Did you assess:
    1. to what extent the decisions and hence the outcome made by the AI system can be understood?
    2. to what degree the system's decision influences the organisation's decision-making processes?
    3. why this particular system was deployed in this specific area?
    4. what the system's business model is (for example, how does it create value for the organisation)?

2. Did you ensure an explanation as to why the system took a certain choice resulting in a certain outcome that all users can understand?

1. Did you design the AI system with interpretability in mind from the start?
    1. Did you research and try to use the simplest and most interpretable model possible for the application in question?
    2. Did you assess whether you can analyse your training and testing data? Can you change and update this over time?
    3. Did you assess whether you can examine interpretability after the model's training and development, or whether you have access to the internal workflow of the model?

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE
UNIVERSITÀ DI PISA
FoReLab Rethinking the future
CROSSLAB Innovation for industry 4.0

Fed-XAI: Current Status – Post-hoc Methods

To our knowledge, post-hoc methods often relies on **feature importance analysis** for post-hoc explainability.

Some interesting works:

G. Wang, "Interpret federated learning with shapley values" [1]

J. Fiosina, "Explainable federated learning for taxi travel time prediction" [2] "Interpretable privacy-preserving collaborative deep learning for taxi trip duration forecasting" [3]

S. M. Lundberg, S.-I. Lee "A unified approach to interpreting model predictions" [4]

D. Janzing et al., "Feature relevance quantification in explainable ai: A causal problem" [5]

P. Chen et al., "An explainable vertical federated learning for data-oriented artificial intelligence systems" [6]

Works differ for:

- XAI technique aspects -> Feature Importance (Shapley Values, Integrated Gradients), counterfactual explanation
- topic and datasets [benchmark sets, real-life taxi datasets]
- ML methods-> KNN, DNN with Federated Averaging (FedAvg) for aggregation
- FL partitioning schemes: vertical / horizontal

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

Fed-XAI: Current Status – Interpretable by design (1)

To our knowledge, interpretable-by-design models works are nowadays focused on **decision trees** or **rule-based** systems

Trees:
H. Ludwiget al., "Ibm federated learning: an enterprise framework" [7]
Y. Wu, S. Cai, X. Xiao, G. Chen, B. C. Ooi, "Privacy preserving vertical federated learning for tree-based models" [8]
-> focus on privacy, adding a partially homomorphic encryption

Multiple models (eg: DT, SVM: using a federated version of the AdaBoost algorithm, does not rely on gradient-based methods -> decrease restrictions on models):
M. Polato, R. Esposito, M. Aldinucci, "Boosting the federation: Cross-silo federated learning" [9]

The *traditional* TSK FRBS

Let

- $X = \{X_1, X_2, \dots, X_F\}$, be a set of input variable
- U_f, be the universe of discourse of variable X_f
- Y, be a continuous output variable
- $P_f = \{A_{f,1}, A_{f,2}, \dots, A_{f,T_f}\}$, be a fuzzy partition over U_f with T_f fuzzy sets

The generic k^{th} rule, R_k, of the rule base is in the form:

IF X_1 **IS** $A_{1,j_{k,1}}$ **... AND** X_F **IS** $A_{F,j_{k,F}}$

THEN $y_k(\mathbf{x}) = \gamma_{k,0} + \sum_{i=1}^{F} \gamma_{k,i} \cdot x_i$

Estimation of antecedent parameters:

- Clustering in the input-output product space
- Fitting convex envelop of the projected membership values for each discovered cluster

Estimation of consequent parameters:

- Weighted Least Squared method

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

The *traditional* TSK FRBS

Let

- $X = \{X_1, X_2, \dots, X_F\}$, be a set of input variable
- U_f, be the universe of discourse of variable X_f
- Y, be a continuous output variable
- $P_f = \{A_{f,1}, A_{f,2}, \dots, A_{f,T_f}\}$, be a fuzzy partition over U_f with T_f fuzzy sets

The generic k^{th} rule, R_k, of the rule base is in the form:

IF X_1 **IS** $A_{1,j_{k,1}}$ **... AND** X_F **IS** $A_{F,j_{k,F}}$

THEN $y_k(\mathbf{x}) = \gamma_{k,0} + \sum_{i=1}^{F} \gamma_{k,i} \cdot x_i$

Inference stage:

Given input pattern $\mathbf{x}$, compute strength of activation of each rule:

$$w_k(\mathbf{x}) = \prod_{f=1}^{F} \mu_{f,j_{k,f}}(x_f) \quad \text{for } k = 1,2,\dots,K$$

with $\mu_{f,j_{k,f}}(x_f)$ membership degree of x_f to $A_{f,j_{k,F}}$

Finally, generate the output as:

$$\hat{y}(\mathbf{x}) = \sum_{k=1}^{K} \left(\frac{w_k(\mathbf{x})}{\sum_{h=1}^{K} (w_h(\mathbf{x}))} \right) \cdot y_k(\mathbf{x})$$

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

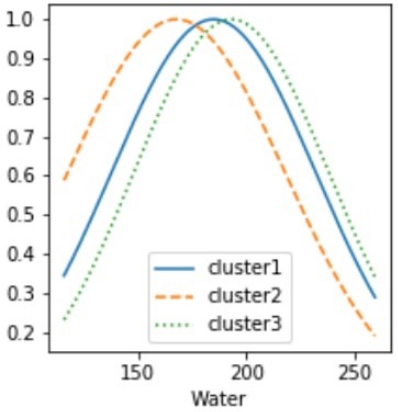
CROSSLAB Innovation for industry 4.0

Fed-XAI: Current Status – Interpretable by design (2)

To our knowledge, interpretable-by-design models works are nowadays focused on **decision trees** or **rule-based** systems

Rule-based (TSK-FRBS*) $\text{IF } X_1 \text{ IS } A_{1,j_{k,1}} \ldots \text{AND } X_F \text{ IS } A_{F,j_{k,F}} \quad \text{THEN } y_k(\mathbf{x}) = \gamma_{k,0} + \sum_{i=1}^{F} \gamma_{k,i} \cdot x_i$

A. Wilbik, P. Grefen, Towards a federated fuzzy learning system [10]
X. Zhu, D. Wang, W. Pedrycz, Z. Li, "Horizontal federated learning of takagi–sugeno fuzzy rule-based models" [11]
-> Two phases:
(i) Learning the fuzzy partitions of each input feature and the antecedent of the rules (data driven, clustering)
(ii) learning the rule consequent of each rule



DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

Fed-XAI: Current Status – Interpretable by design (3)

To our knowledge, interpretable-by-design models works are nowadays focused on **decision trees** or **rule-based** systems

Rule-based (TSK-FRBS*)

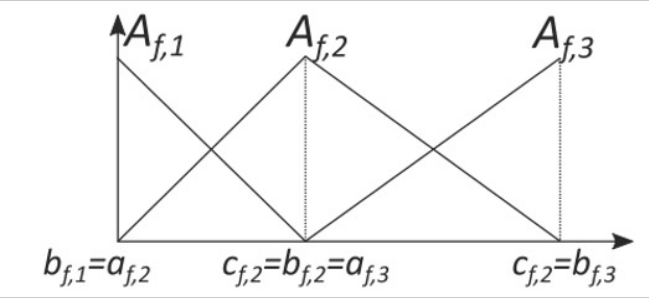
$$\text{IF } X_1 \text{ IS } A_{1,j_{k,1}} \text{ ... AND } X_F \text{ IS } A_{F,j_{k,F}} \quad \text{THEN } y_k(\mathbf{x}) = \gamma_{k,0} + \sum_{i=1}^{F} \gamma_{k,i} \cdot x_i$$

J. L. C. Bárcena, P. Ducange, A. Ercolani, F. Marcelloni, A. Renda, "An approach to federated learning of explainable fuzzy regression models" [12]

Key point: Enhance interpretability by:

- Strong Fuzzy Triangular Partitions
- Design ad hoc approach on antecedents rules generation (one rule for each training pattern or one rule for each cluster or information granule)
- design ad hoc approach for inference phase, using only one rule of the rule base (maximum matching)

Ad hoc aggregation step



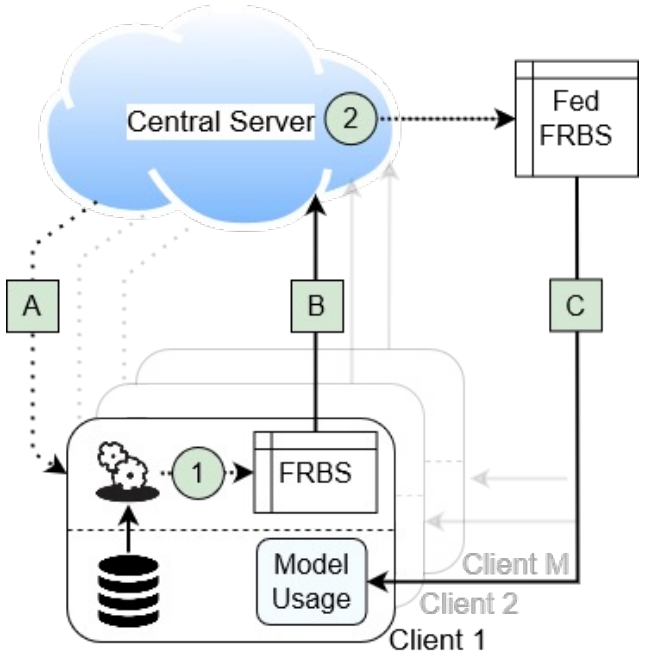
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

Our Federated TSK FRBS



A — **Configuration:** central server configures the learning process

1 — Local learning of TSK-FRBSs

B — **Transmission** of **local models** to the central server

2 — **Federated learning of the global TSK-FRBS:** aggregation of the models

C — **Transmission** of the **aggregated model** to the clients

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

Our Federated TSK FRBS – Aggregation Step

	Antecedent	*Consequent*	*Rule Weight*
Client 1	$ant_{1,1}$	$cons_{1,1}$	$rw_{1,1}$
	...	...	...
	$ant_{1,i}$	$cons_{1,i}$	$rw_{1,i}$
	...	...	...
	ant_{1,K_1}	$cons_{1,K_1}$	rw_{1,K_1}
...		...	
Client m	$ant_{m,1}$	$cons_{m,1}$	$rw_{m,1}$
	...	...	...
	$ant_{m,j}$	$cons_{m,j}$	$rw_{m,j}$
	...	...	...
	ant_{m,K_m}	$cons_{m,K_m}$	rw_{m,K_m}
...		...	
Client M	$ant_{M,1}$	$cons_{M,1}$	$rw_{M,1}$
	...	...	...
	$ant_{M,k}$	$cons_{M,k}$	$rw_{M,k}$
	...	...	...
	ant_{M,K_M}	$cons_{M,K_M}$	rw_{M,K_M}

Centralized server operation

1. Juxtaposition of rules collected from the M clients.
2. Identification of **conflicting rules:**
   (i.e., same antecedents, different consequents)
3. Replacement of conflicting rules with a new *single* rule:

- **Antecedent**: same of that of conflicting rules
- **Consequent**: coefficients computed as the weighted average of those from conflicting rules (weighted by a RW)
- **Rule weight (RW)**: average of rule weights of conflicting rules

The final rule base represents our Federated TSK model

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

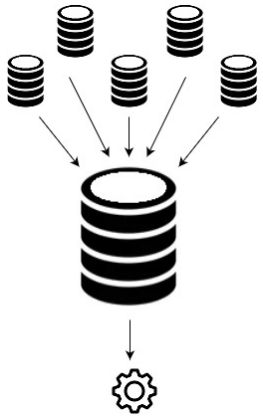
Experimental Setup

Dataset	Abbreviation	Dimensionality (F)	Samples (N)
Weather Izmir	WI	9	1461
Treasury	TR	15	1049
Mortgage	MO	15	1049
California	CA	8	20460

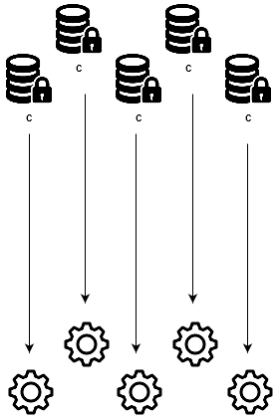
- Four regression datasets
- Params: $T_f = 3, C_{FCM} = 30$
- Simulated distributed setting: randomly split each dataset (same number of instances) among 5 participants

Three scenarios

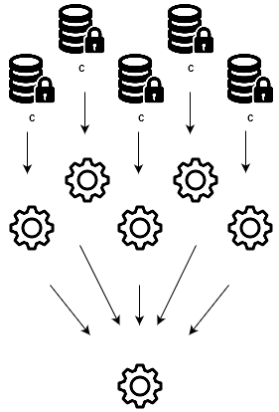
Centralized model: *no privacy*



Local model: *no collaboration*



Federated model: *privacy & collaboration*



DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

Experimental Results

Setting

- Mean Squared Error (MSE) evaluated with 5-fold cross-validation
- Each of the three models evaluated on same local splits

Considerations

- *federated* always outperforms *local*, on average
- *federated* comparable to *centralized* for WI and CA
- *centralized* outperforms *federated* in case of high dimensionality ($F_{MO} = F_{TR} = 15$) and data scarcity ($N_{MO} = N_{TR} = 1049$)
- performance comparable to those reported in the literature

Average MSE

Client ID	Local		Federated		Centralized	
	Train	*Test*	*Train*	*Test*	*Train*	*Test*
Weather Izmir						
1	1.33	*2.02*	1.44	*1.57*	1.40	*1.54*
2	1.09	*1.62*	1.25	*1.41*	1.22	*1.34*
3	0.96	*1.40*	1.25	*1.32*	1.22	*1.29*
4	1.07	*7.10*	1.23	*1.30*	1.20	*1.28*
5	1.19	*1.64*	1.41	*1.51*	1.38	*1.46*
Avg.	1.13	*2.76*	1.32	*1.42*	1.28	*1.38*
Treasury ($\times 10^{-3}$)						
1	7.11	*377.40*	82.20	*112.72*	21.97	*46.13*
2	19.28	*192.70*	53.64	*79.41*	37.69	*51.35*
3	7.72	*337.25*	429.38	*174.18*	26.86	*41.97*
4	9.31	*110.47*	72.86	*378.61*	20.51	*41.69*
5	10.37	*133.83*	57.04	*40.85*	13.24	*20.37*
Avg.	10.76	*230.33*	139.02	*157.15*	24.06	*40.30*
Mortgage ($\times 10^{-3}$)						
1	2.29	*78.08*	9.70	*15.96*	5.20	*7.55*
2	1.44	*15.08*	9.14	*7.35*	3.47	*5.22*
3	1.22	*38.18*	14.61	*9.52*	3.31	*5.22*
4	1.54	*53.84*	9.38	*35.90*	4.24	*8.83*
5	1.09	*43.36*	14.78	*5.14*	3.74	*4.98*
Avg.	1.52	*45.71*	11.52	*14.77*	3.99	*6.36*
California ($\times 10^{9}$)						
1	4.73	*4.87*	4.75	*4.86*	4.77	*4.78*
2	4.62	*4.73*	4.57	*4.58*	4.60	*4.62*
3	4.71	*4.89*	4.71	*4.74*	4.72	*4.75*
4	4.77	*5.10*	5.23	*5.34*	5.18	*5.24*
5	4.70	*4.82*	4.63	*4.64*	4.65	*4.68*
Avg.	4.71	*4.88*	4.78	*4.83*	4.78	*4.81*

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

Experimental Results: Additional Considerations

Global interpretability as model complexity (average number of rules)

- data summarization strategy helps **limiting the overall number of rules**
- *local* and *centralized*: similar number of rules
- *federated*: generally more complex due to rule merging

Average number of rules

Dataset	Local	Centralized	Federated
Weather Izmir (WI)	13.96	13.40	27.80
Treasury (TR)	21.36	21.20	42.40
Mortgage (MO)	21.60	21.00	46.00
California (Ca)	8.80	8.60	10.20

Validation (centralized setting) of the proposed approach to learn **TSK-FRBSs with enforced interpretability**

- **TSK-SC:** our approach - single consequent (maximum matching)
- **TSK-AC:** our approach - averaging consequents (as in traditional TSK-FRBS)
- **pyFUME**: state of art approach (tuned at comparable complexity)

Average MSE

	TSK-SC		TSK-AC		PyFUME [5], [6]	
Dataset	Train	Test	Train	Test	Train	Test
WI	1.28	1.38	1.28	1.37	1.48	1.52
TR	24.06	40.30	24.42	39.18	32.07	62.93
MO	3.99	6.36	4.29	6.14	4.49	8.22
CA	4.78	4.81	4.82	4.85	4.62	4.64

Our **TSK-SC** achieves higher level of interpretability without compromising modelling capability

Fuchs et al., "pyFUME: a Python package for fuzzy model estimation," in 2020 IEEE Int'l Conf. on fuzzy systems
Fuchs et al. "Towards more specific estimation of membership functions for data-driven fuzzy inference systems," in 2018 IEEE Int'l Conf. on Fuzzy Systems

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE
UNIVERSITÀ DI PISA
FoReLab Rethinking the future
CROSSLAB Innovation for industry 4.0

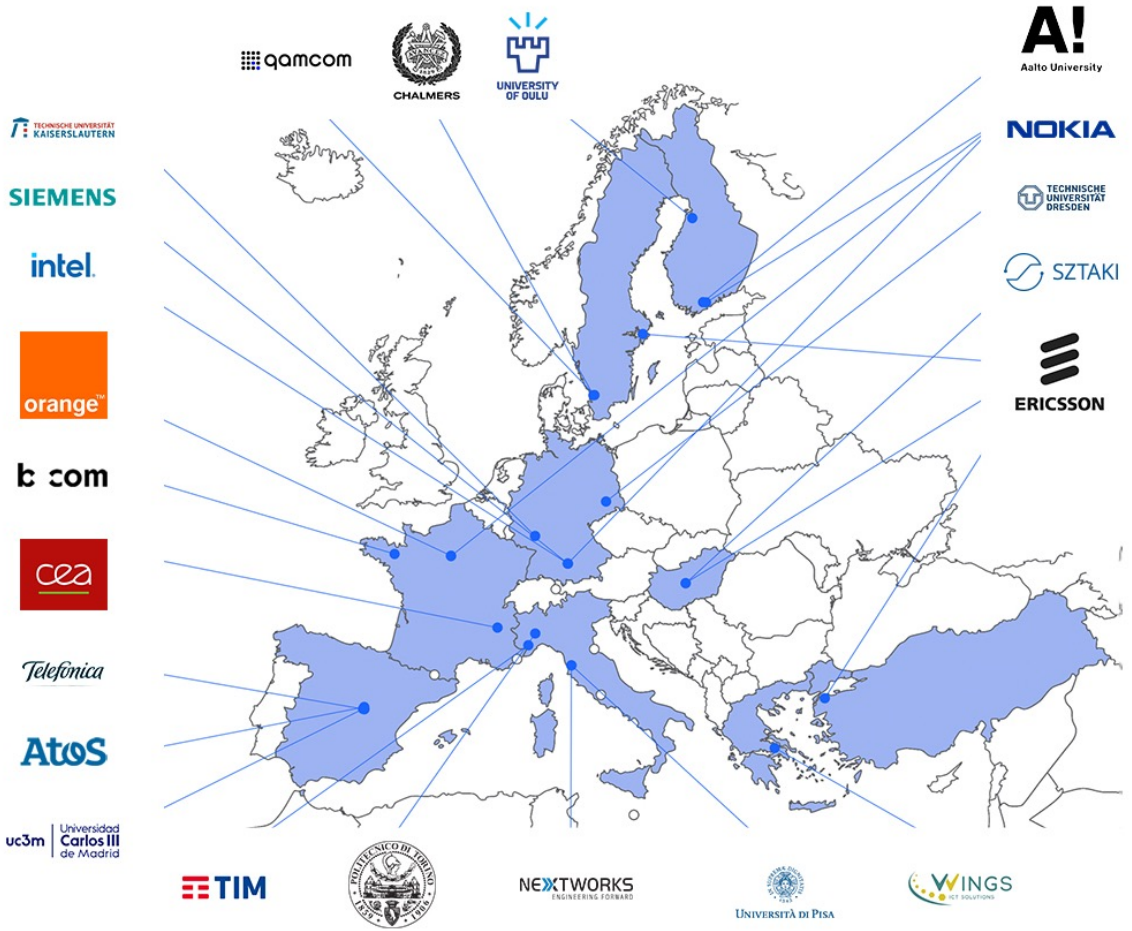
HEXA-X: The European 6G flagship project

Hexa-X

A **flagship for B5G/6G vision** and intelligent fabric of technology enablers connecting human, physical, and digital worlds.

UNIPI group contribution: **FED-XAI as a Service**

Project awarded as **key innovation** by the EU Innovation Radar.

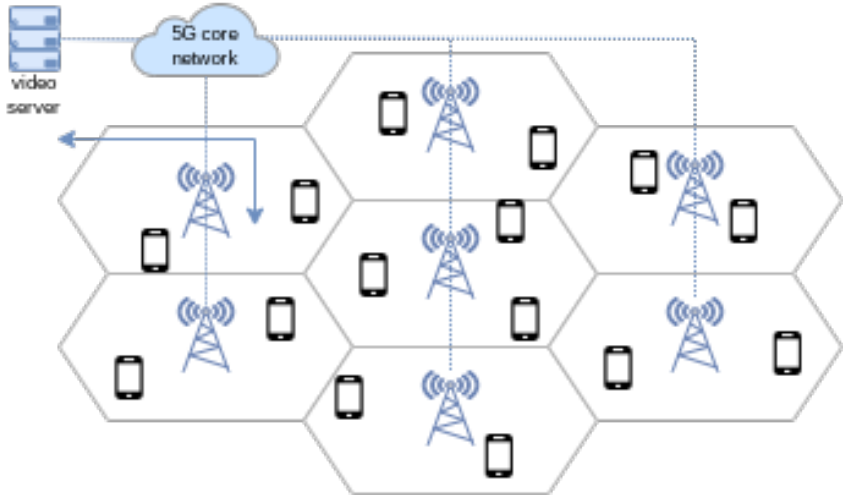


EU project: HEXA-X - Programme: Horizon 2020 -
Grant Agreement ID: 101015956

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Use Case: Quality of Experience Forecasting

- FL service for addressing **QoE forecasting** task in an automotive case study
- Connected vehicles (UEs) play real-time video streams whose perceived quality is relevant to determine the availability of some advanced driving assistance system (e.g., in a see-through or tele-operated driving application)
- **Goal**: train an ML model to predict future QoE by leveraging real-time QoS and QoE data
- The relevant dataset (mobile network data) is generated usign Simu5G. Details in [1]



- **15 UEs (FL participants)**
- 7 gNBs
- 120s simulation
- 24 independent replicas (20 train + 4 test)

[1] J. L. C. Barcena et al. Towards Trustworthy AI for QoE prediction in B5G/6G Networks, in: 1st Int'l Workshop on AI in beyond 5G and 6G Wireless Networks - AI6G2022, Vol.3189, 2022, pp. 1–9.

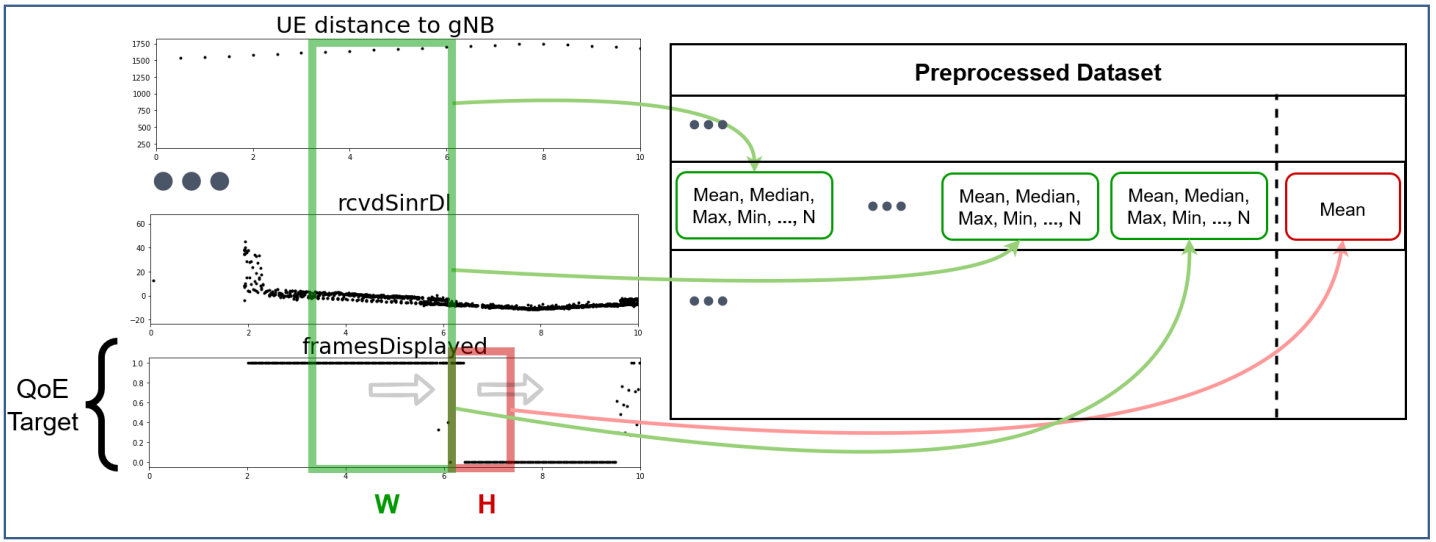
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

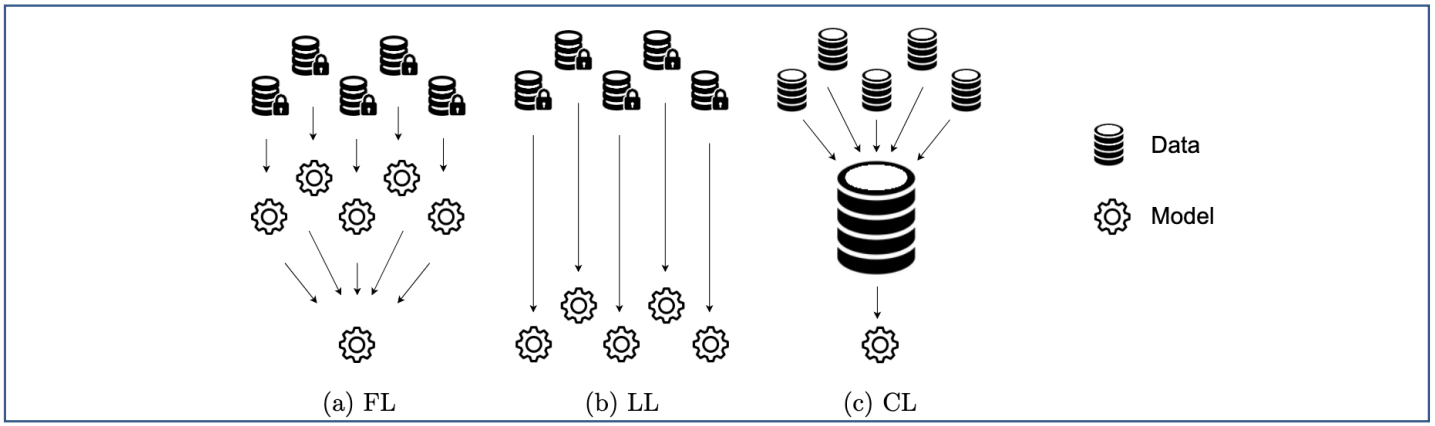
FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

QoE forecasting as a regression problem



- Schematization of the main preprocessing steps



- Schematization of the three experimental learning settings

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

Enabling Federated Learning of Explainable AI Models

- Experimental results:
  - Comparison of MSE values obtained on the test set (4 runs) of the 15 UEs
  - Empirical cumulative distribution function (ECDF) of the differences of MSE scores

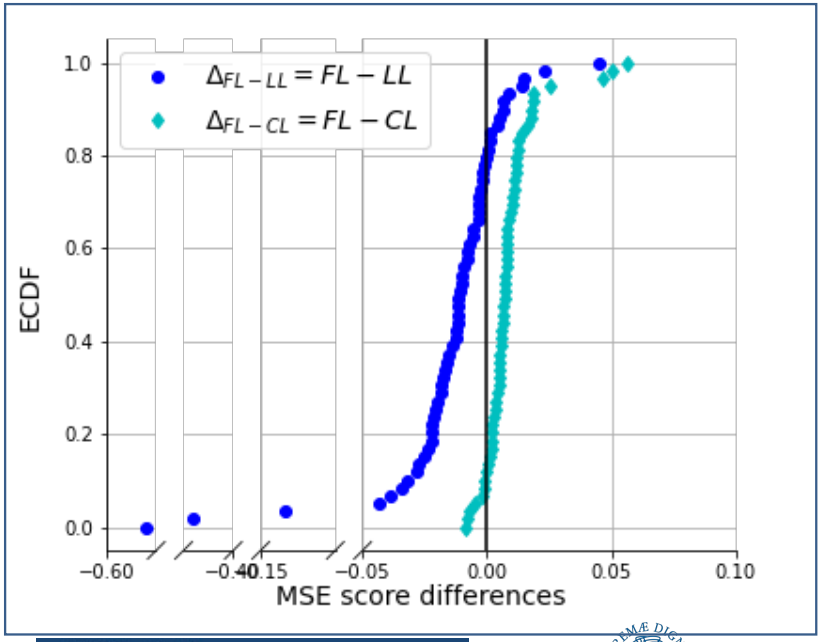


Table 3: MSE scores: fine-grained results on the test set for FL and LL models. Best values are highlighted in bold.

	Run 1		Run 2		Run 3		Run 4	
	FL	LL	FL	LL	FL	LL	FL	LL
UE-01	0.049	**0.047**	**0.087**	0.098	**0.061**	0.073	**0.070**	0.085
UE-02	**0.037**	0.042	**0.075**	0.093	**0.156**	0.160	**0.059**	0.067
UE-03	0.066	**0.051**	**0.054**	0.056	**0.027**	0.054	**0.100**	0.104
UE-04	**0.062**	0.083	**0.097**	0.118	**0.088**	0.105	0.072	**0.068**
UE-05	0.039	0.039	0.027	**0.020**	**0.051**	0.056	**0.073**	0.076
UE-06	**0.078**	0.112	**0.063**	0.102	0.068	**0.063**	0.100	**0.091**
UE-07	**0.043**	0.053	**0.046**	0.630	**0.060**	0.075	**0.042**	0.061
UE-08	**0.066**	0.076	0.086	**0.072**	0.138	**0.093**	0.060	**0.037**
UE-09	0.029	0.029	0.044	**0.043**	**0.080**	0.101	**0.057**	0.068
UE-10	**0.112**	0.124	**0.073**	0.089	**0.055**	0.080	**0.064**	0.096
UE-11	**0.073**	0.090	**0.058**	0.061	**0.065**	0.067	**0.131**	0.548
UE-12	**0.053**	0.075	**0.032**	0.055	**0.038**	0.046	**0.030**	0.031
UE-13	**0.038**	0.060	**0.104**	0.244	0.036	**0.029**	**0.068**	0.080
UE-14	**0.050**	0.062	**0.037**	0.047	**0.110**	0.134	**0.073**	0.084
UE-15	**0.056**	0.100	**0.048**	0.076	**0.076**	0.079	**0.061**	0.069

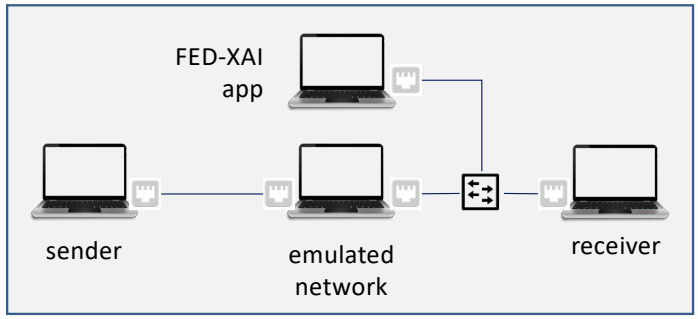
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

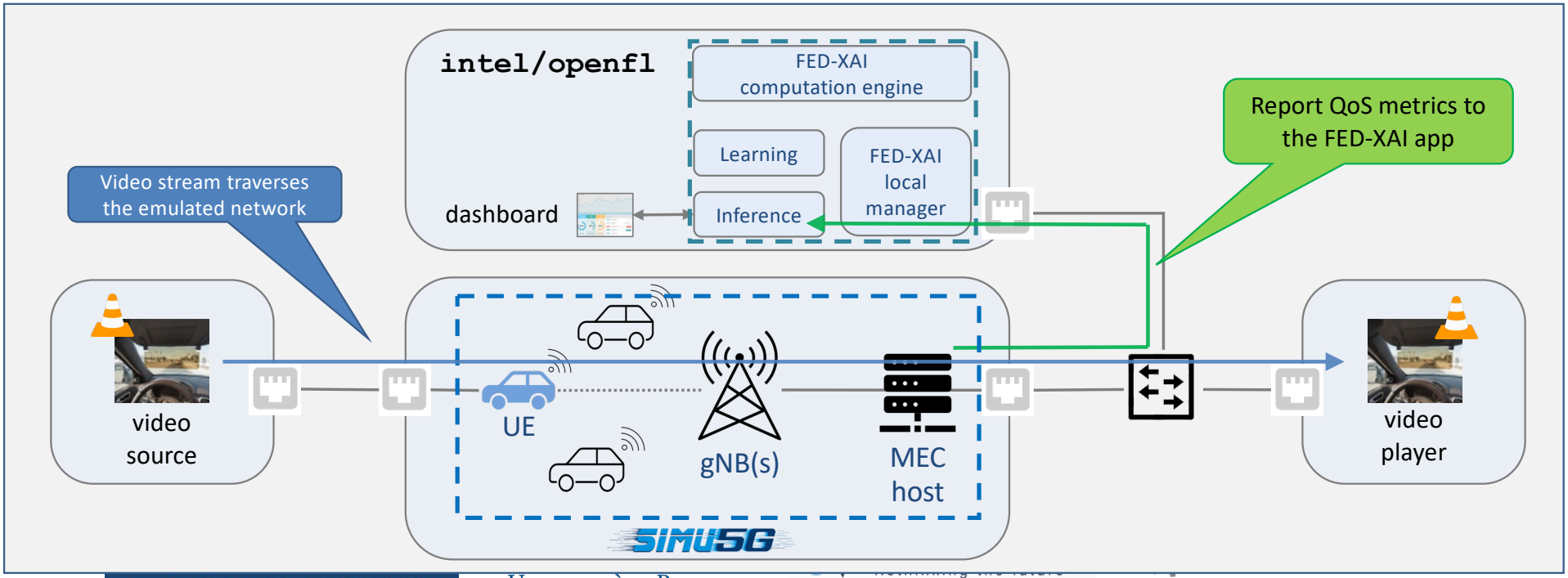
CROSSLAB Innovation for industry 4.0

Overview of the FED-XAI demo testbed

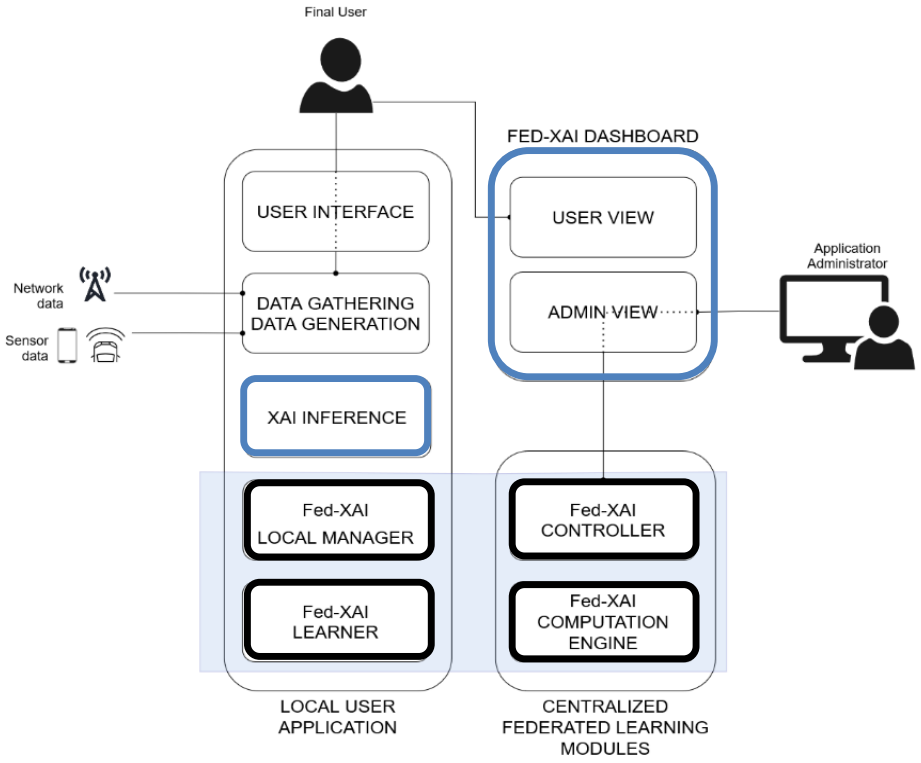


Testbed configuration

1. Realistic 5G network emulated with **Simu5G**
2. Offline model training and online inference with **OpenFL**
   - Predictions and root-causes shown on a dashboard



FED-XAI app components



- **Fed-XAI LocalManager**
  Digital twin of the user
- **Fed-XAI Controller**
  Manages the framework, initializes FL process
- **Fed-XAI Learner**
  Trains the user's local model
- **Fed-XAI Computation Engine**
  Aggregates the local models and manages the FL process

- **XAI Inference**
  Exploits a XAI model to perform inference over a set of data.
- **Fed-XAI Dashboard**
  Allows to visualize informations about the predictions of the Inference module

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

FED-XAI app implementation: details

Deployment of components

- **Containers** as de-facto standard for Lightweight Virtualization
- Compliance with edge-computing / MEC-enabled environments

Message exchange

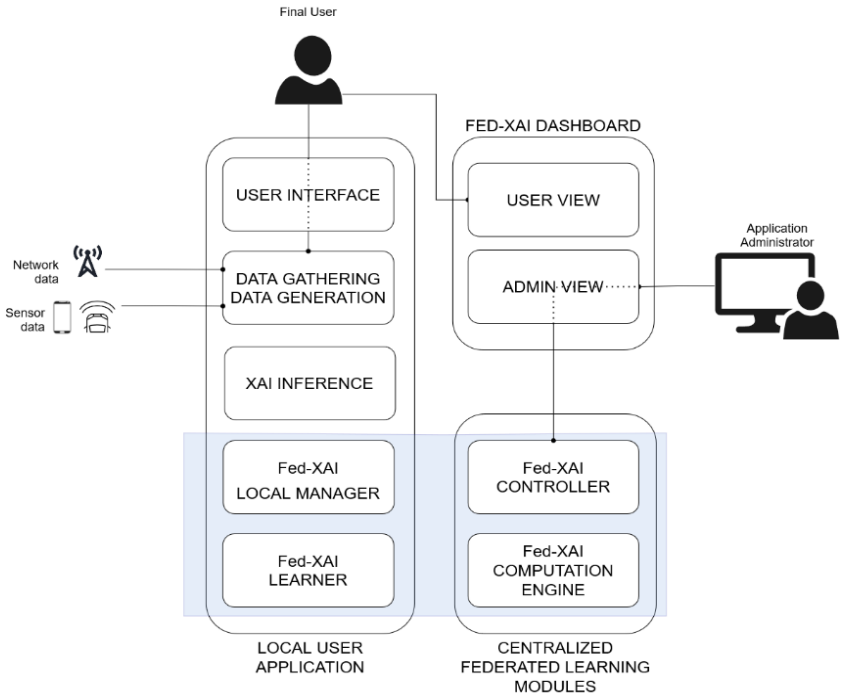
- **RestAPIs** for handling and integrating app microservices
- Over HTTPS: encryption for secure communication

Federated Learning Framework

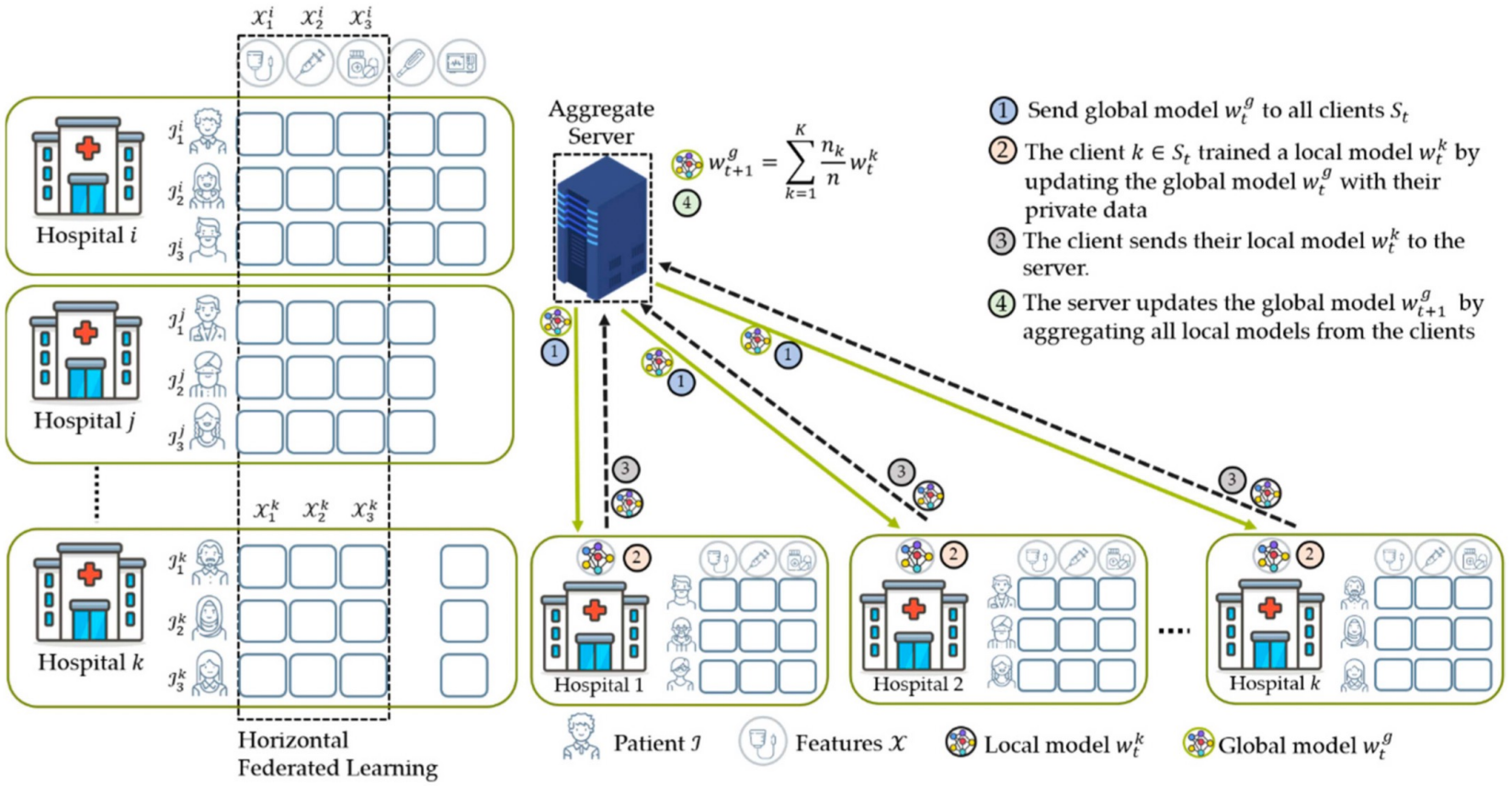
- Intel **OpenFL**
- Seamless integration with containers paradigm
- Extended to support FL of inherently interpretable models

Repository module

- Instance of **MongoDB** in docker container



FED-XAI in HEALTH: Horizontal Federated Learning



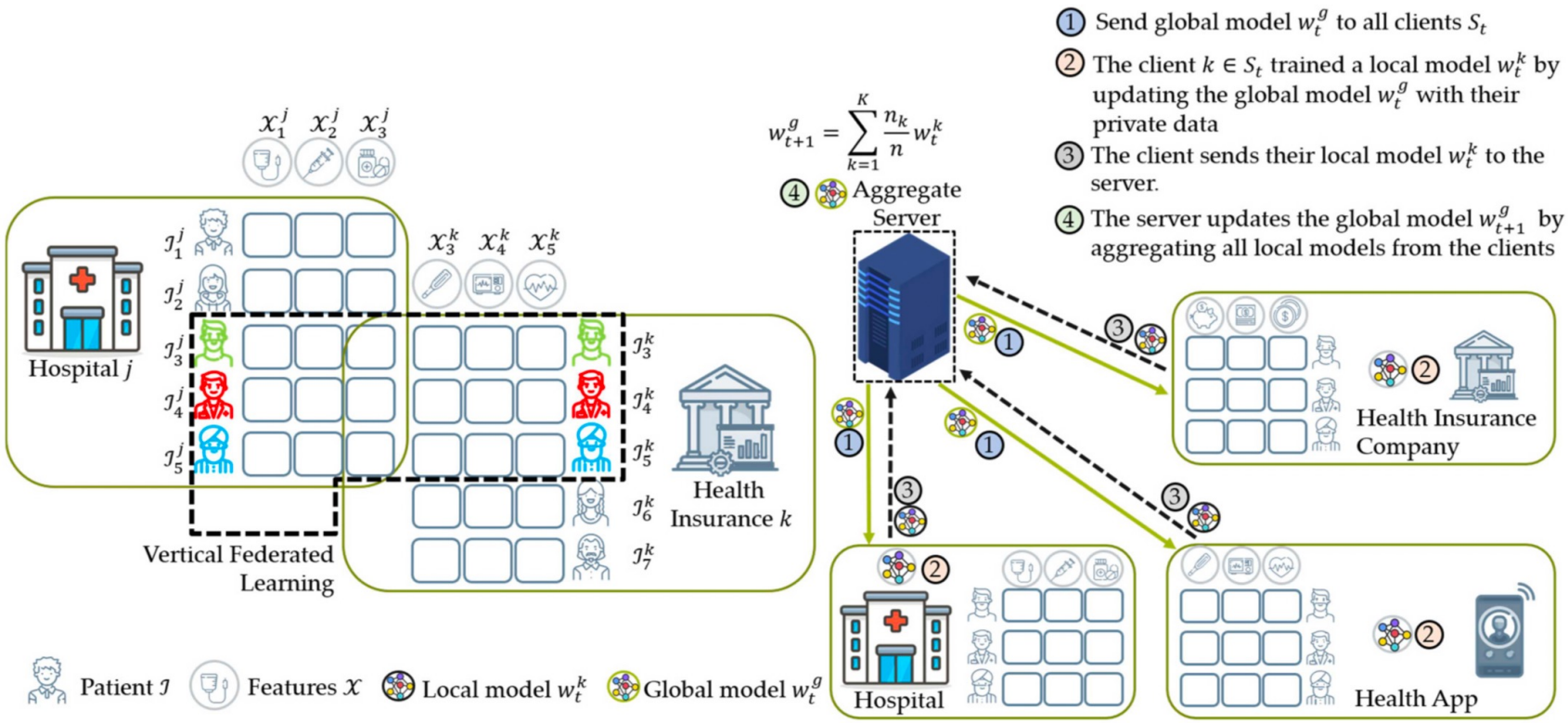
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

FED-XAI in HEALTH: Vertical Federated Learning



DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

FED-XAI in HEALTH: a use case with Parkinson dataset

Parkinson's Disease (PD) is one of the most common neurological and complicated diseases affecting the central nervous system.

Unified Parkinson's Disease Rating Scale (UPDRS) is widely used for tracking PD symptom progression.

The aim of this study is to predict UPDRS scores through analyzing the speech signal properties which is important in PD diagnosis.

We considered a dataset with 5875 records of 28 men and 14 women with early-stage Parkinson's disease recruited to a six-month trial.

This dataset is composed of a range of biomedical voice and the recordings were automatically captured in the patient's homes.

A number of numeric features (around 20) have been extracted from the recorded speech signals.

We grouped the records into 10 "medical centers" and their "digital twins" have been created in our FED-XAI application.

We have considered two scenarios: IID and Non-IID data distribution among clients.

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

FED-XAI in HEALTH: a use case with Parkinson dataset

Table 2. Scenarios description

	S1: training set		S2: training set		test set	
client	age range	samples	age range	samples	age range	samples
0	[36,85]	529	[36,55]	561	[36,85]	59
1	[36,85]	529	[56,57]	473	[36,85]	59
2	[36,85]	529	[58,59]	655	[36,85]	59
3	[36,85]	529	[60,62]	487	[36,85]	59
4	[36,85]	529	[63,65]	506	[36,85]	59
5	[36,85]	529	[66,66]	380	[36,85]	59
6	[36,85]	529	[67,71]	689	[36,85]	59
7	[36,85]	528	[72,72]	279	[36,85]	59
8	[36,85]	528	[73,74]	591	[36,85]	58
9	[36,85]	528	[75,85]	666	[36,85]	58
tot		5287		5287		588

TSK	RMSE		r	
	train	*test*	train	*test*
S1 - LL	6.165	*11.214*	0.820	*0.448*
S1 - FL	7.907	*8.657*	0.677	*0.622*
S1 - CL	7.790	*7.850*	0.688	*0.660*
S2 - LL	3.221	*91.832*	0.919	*-0.064*
S2 - FL	13.166	*14.807*	0.509	*0.470*
S2 - CL	7.477	*7.850*	0.641	*0.660*

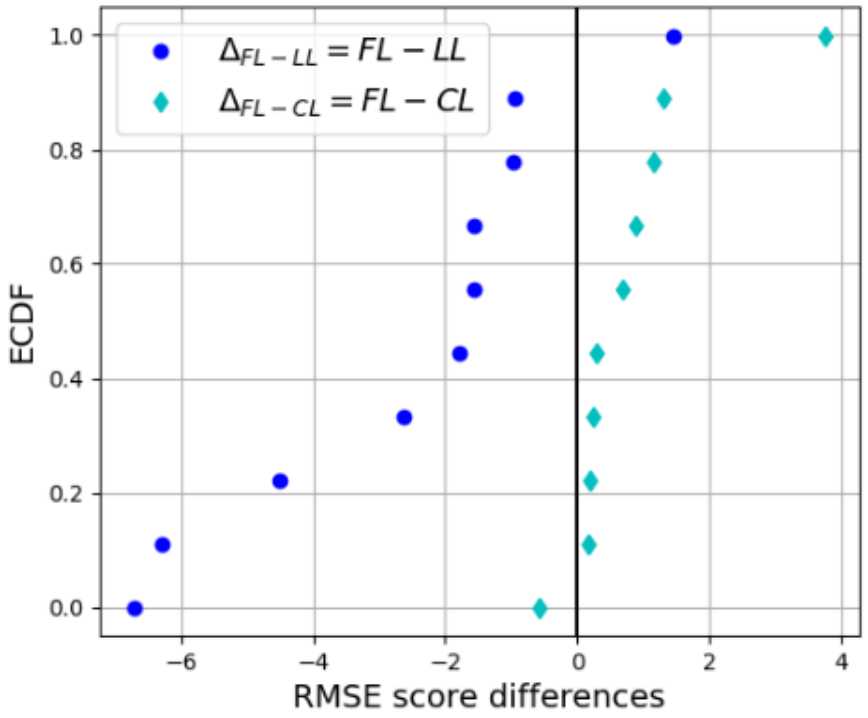
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

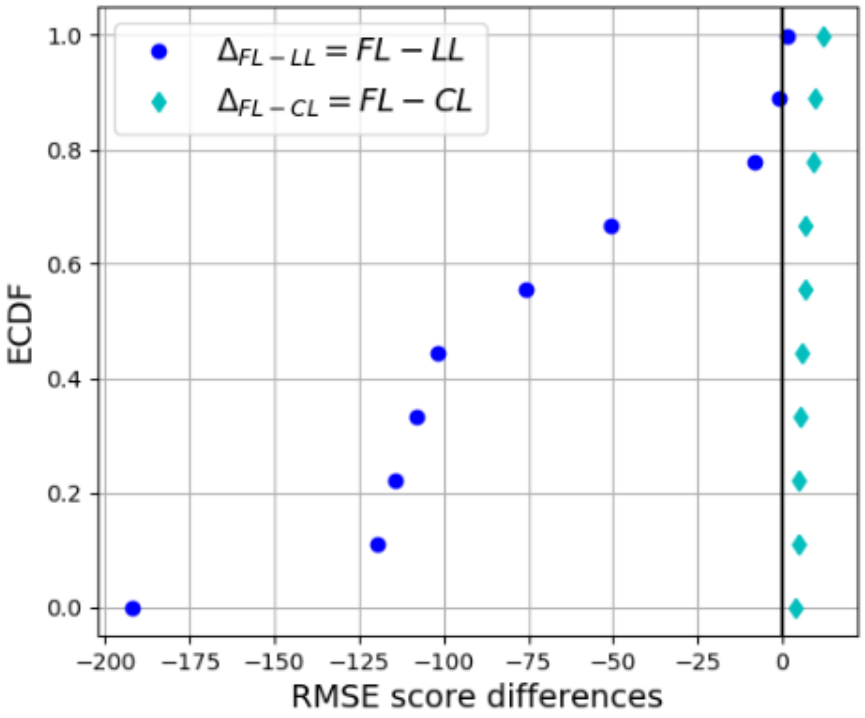
FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

FED-XAI in HEALTH: a use case with Parkinson dataset



(a) Scenario 1



(b) Scenario 2

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

Open challenges

Major challenges:

1. **Data privacy:** how do we ensure strong privacy constraint, for example avoiding data leakage scenarios?
2. **Aggregation**: how do we effectively merge XAI local models where...

- we have different local rules...?
- we have DNN weights: compress information transmission but losing accuracy...?

3. **Streaming scenarios**:

- how to adapt models to massive streaming data scenarios, especially in presence of concept drifts

4. **Standardization (needs for ad hoc benchmark datasets)**:

- Datasets commonly used (eg, in DL) as benchmark are related to text or images -> needs to be preprocessed if used by rule/tree-based approaches
- Standardization of experimental setup for data distribution among clients (to analyze iid and non-iid data)

5. **Architecture**:

Fed-XAI future applications to be designed for ensuring low latency and reducing network congestions: edge-computing platforms, and in particular **Multi-Access Edge Computing** (MEC) could be proposed as standard

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0

Thank you very much for your attention. Questions?

Contacts: pietro.ducange@unipi.it

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

FoReLab Rethinking the future

CROSSLAB Innovation for industry 4.0