

# Federated Learning of Explainable AI Models: A Step Toward Trustworthy AI

**Francesco Marcelloni**

Academic Director of the Knowledge Hub on AI  
DII AI Group Coordinator, IT2PAO Lab Coordinator,  
Good AI Lab Coordinator

FAIR spoke 1 CO-PI  
Dipartimento di Ingegneria dell'Informazione  
Università di Pisa

**May 28, 2025**

Knowledge Hub on AI  
Symposium - Belgrade

E-mail: [francesco.marcelloni@unipi.it](mailto:francesco.marcelloni@unipi.it)

# Project FoReLab – Department of Excellence



**Future-Oriented Research Laboratory (Forelab)** aims to focusing towards new methodologies, paradigms, and enabling technologies for Industry 5.0

- **Trustworthy Artificial/Embodied Intelligence (TAEI)**: The activity focuses on frontier aspects of **trustworthy AI** and on the ability of systems to develop **intelligent behaviours** as emerging from the interactions between artificial agents (e.g. robots) and the environment, through the body of the agent itself (Embodied Intelligence), with the aim of promoting widespread, reliable and integrated use of both.



# Good AI Lab

## Fair Spoke 1's and FoReLab's spin-off: Good AI Lab

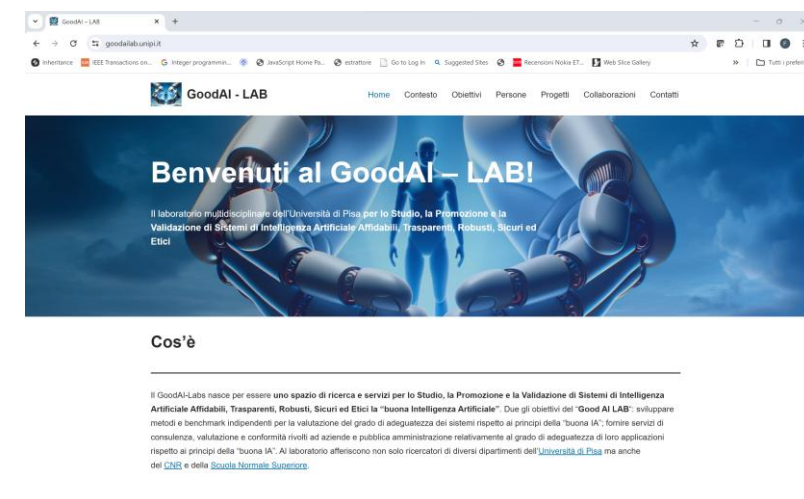
### How to design Trustworthy, Transparent, Robust, Safe, and Ethically-aligned AI Systems?

- Develop independent methods and benchmarks for assessing the compliance with the principles of "good AI".
- Provide consultancy, evaluation, and certification services aimed at companies and public administration
- Deliver multidisciplinary training courses

**Location:** Department of Information Engineering, UNIPi

**Coordinator:** Francesco Marcelloni

Scientific Board with members from UNIPi, CNR and SNS



<https://goodailab.unipi.it/>

# The EU view of AI

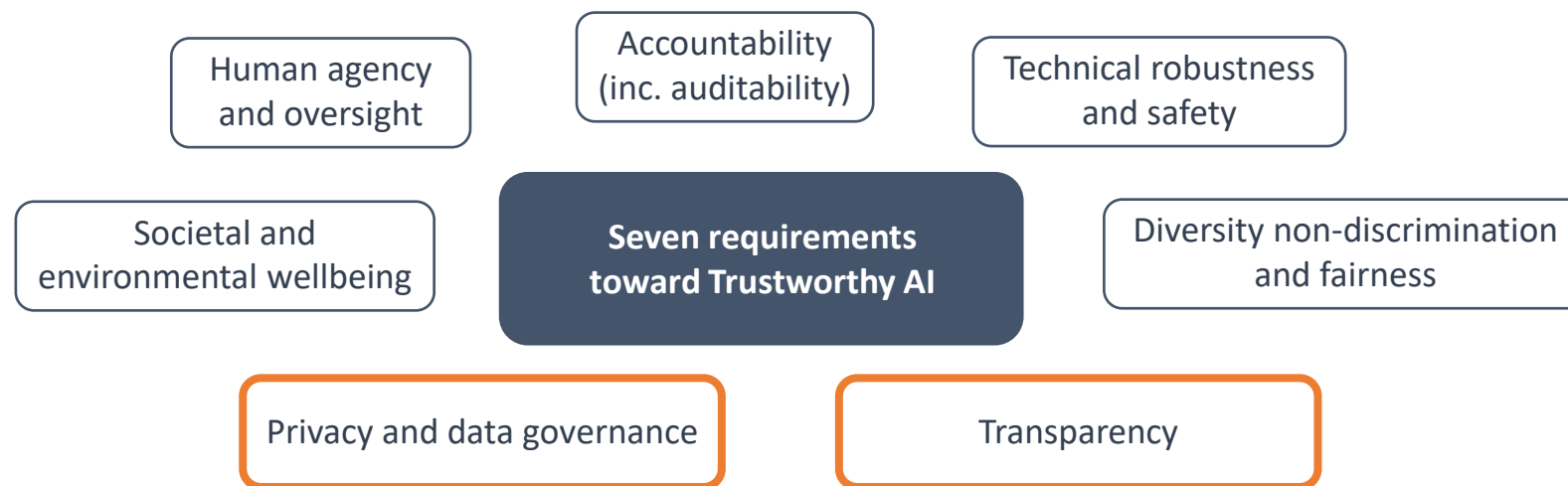
## AI ACT (21 May 2024):

... to promote the uptake of human centric and trustworthy artificial intelligence while ensuring a high level of **protection of health, safety, fundamental rights** enshrined in the Charter, including democracy and rule of law and environmental protection, against harmful effects of artificial intelligence systems in the Union **and to support innovation.**

The path toward the AI Act began in 2019 with the publication of the Ethics Guidelines for Trustworthy AI by the European Commission's High-Level Expert Group on AI.



# The EU view of AI

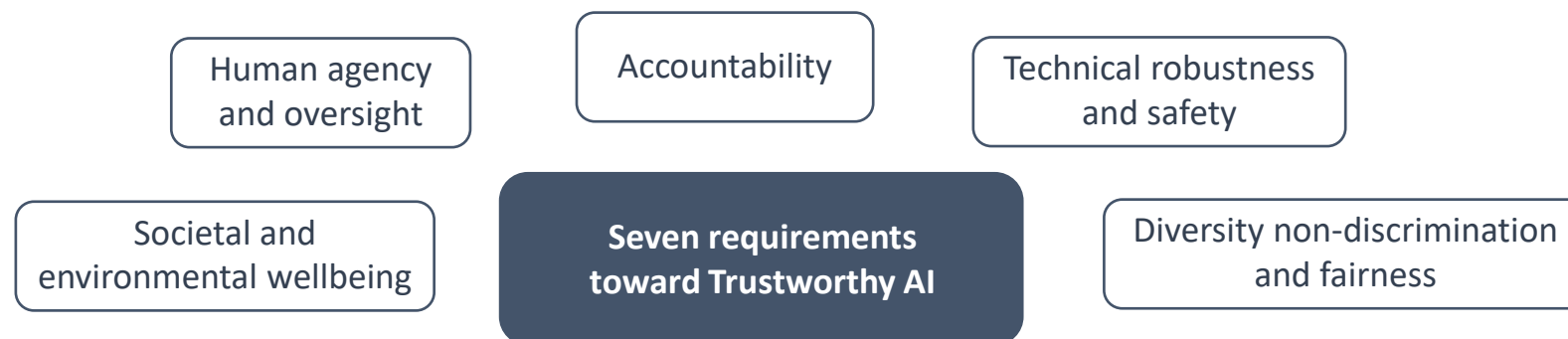


Need to collect data to train accurate AI models  
clashes with need to preserve privacy of data owners.

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



# The EU view of AI



The University of Pisa with the idea of Federated Learning of Explainable Artificial Intelligence Models (Fed-XAI) has been selected as **Key Innovator** by the European Commission's Innovation Radar.

Need to collect data to train accurate AI 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.”



# Privacy and Data Governance

- Increasing attention towards privacy preservation
  - EU General Data Protection Regulation (**GDPR**)
- Novel learning paradigm: **Federated Learning**
  - Training a *centralized model on decentralized data*
  - Participants share model updates, not data

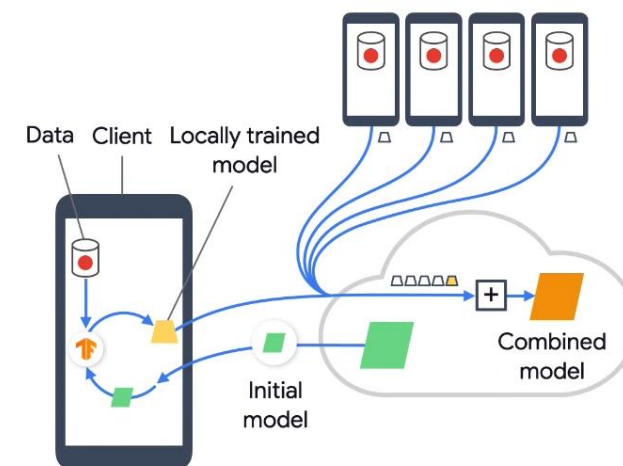
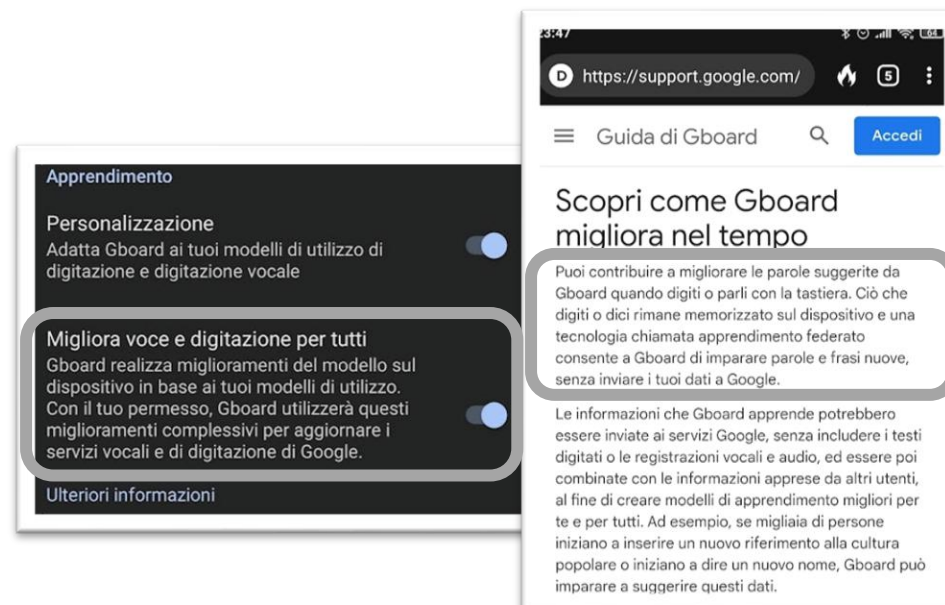


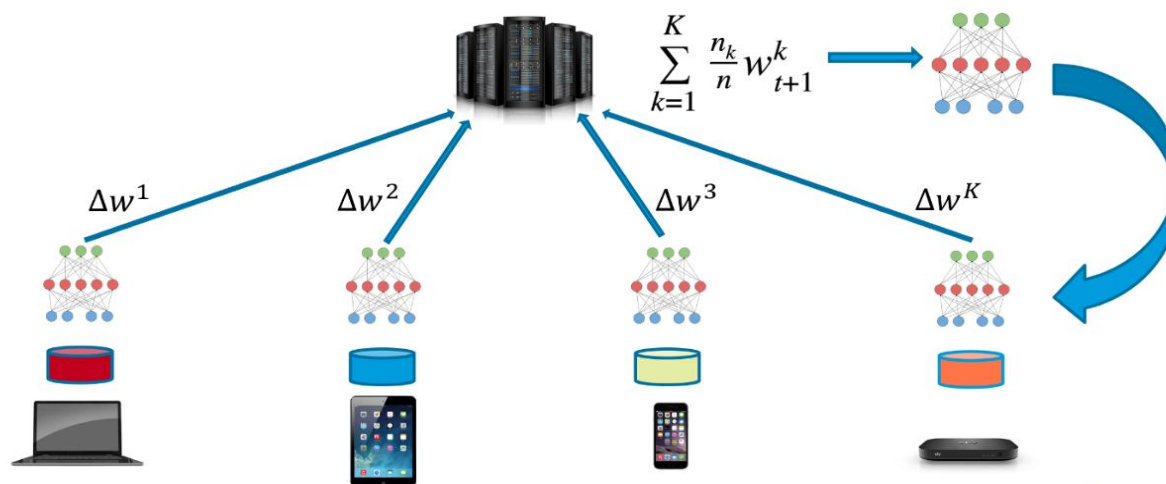
Figure from <https://ml.berkeley.edu>

- Examples:
  - Gboard
  - Voice Assistant



# Popular approaches (Federated Averaging)

How does it work?



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  
( $B = \infty$  indicates that the full local dataset is treated as a single minibatch)
- $P_k$  = set of indexes of data points on client  $k$ , with  $n_k = |P_k|$



**Algorithm 1** FederatedAveraging. The  $K$  clients are indexed by  $k$ ;  $B$  is the local minibatch size,  $E$  is the number of local epochs, and  $\eta$  is the learning rate.

**Server executes:**

```

initialize  $w_0$ 
for each round  $t = 1, 2, \dots$  do
     $m \leftarrow \max(C \cdot K, 1)$ 
     $S_t \leftarrow$  (random set of  $m$  clients)
    for each client  $k \in S_t$  in parallel do
         $w_{t+1}^k \leftarrow \text{ClientUpdate}(k, w_t)$ 
     $w_{t+1} \leftarrow \sum_{k=1}^K \frac{n_k}{n} w_{t+1}^k$ 
    
```

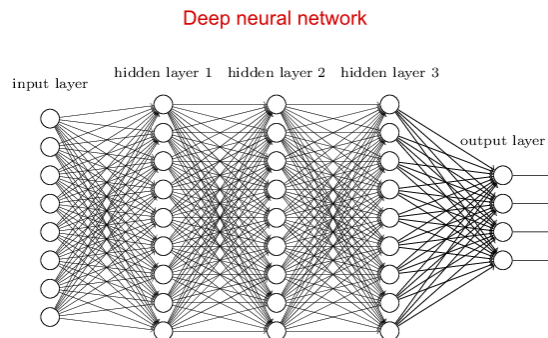
**ClientUpdate( $k, w$ ):** // Run on client  $k$

```

 $\mathcal{B} \leftarrow$  (split  $\mathcal{P}_k$  into batches of size  $B$ )
for each local epoch  $i$  from 1 to  $E$  do
    for batch  $b \in \mathcal{B}$  do
         $w \leftarrow w - \eta \nabla \ell(w; b)$ 
    return  $w$  to server
    
```

# Transparency

- Communication
  - Humans have the right to be informed that they are interacting with an AI system.
- Traceability
  - Data gathering/labelling and algorithms should be documented to the best possible standard
- **Explainability**
  - **Systems and decisions should be explained in a manner adapted to the stakeholder concerned**
    - *Easy* for decision trees, *critical* for deep neural networks



## Deep neural networks

- Multiple layers of non-linear information processing
- Often referred to as ***opaque*** or ***black-box*** models

# Explainability

- How to achieve explainability?

Post-hoc Explainability Techniques

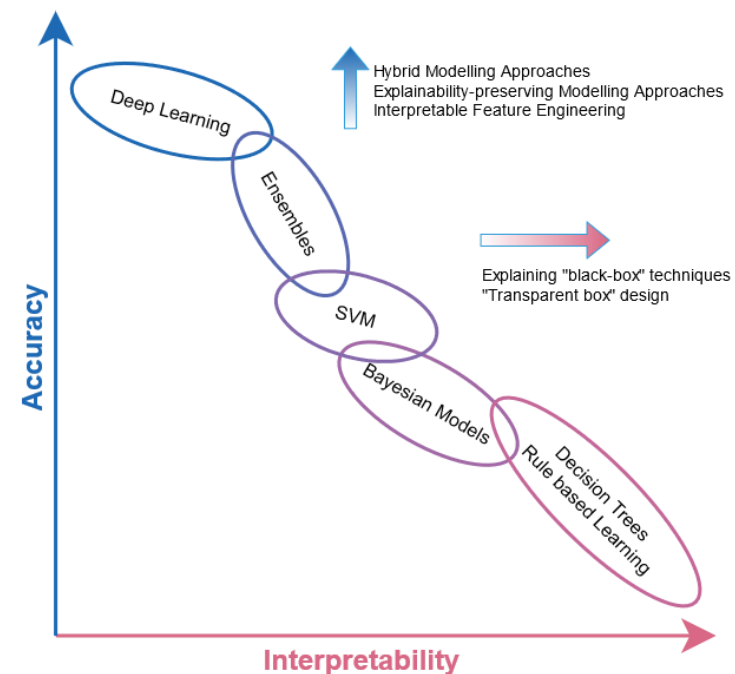
Design of Inherently Interpretable Models

- How to characterize inherent interpretability?

- There exists a **trade-off\*** between
  - model **accuracy**
  - model **interpretability**

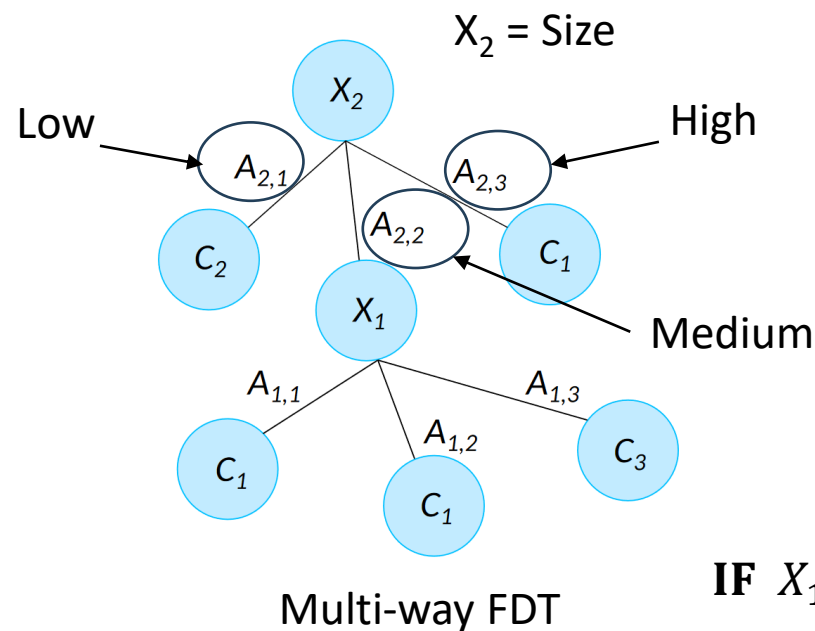
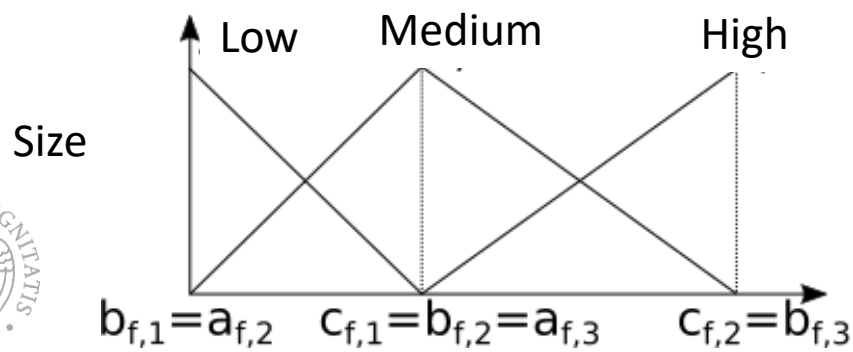
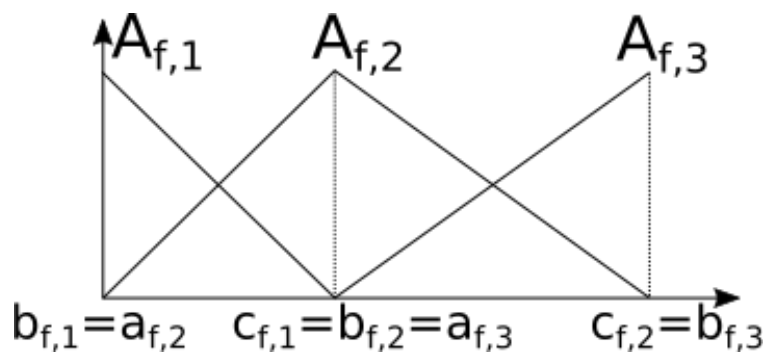
\*given

- a target function of a certain complexity, and
- a suitable amount of available data



# Interpretable Models: Fuzzy Decision (FDT) and Fuzzy Regression (FRT) Trees

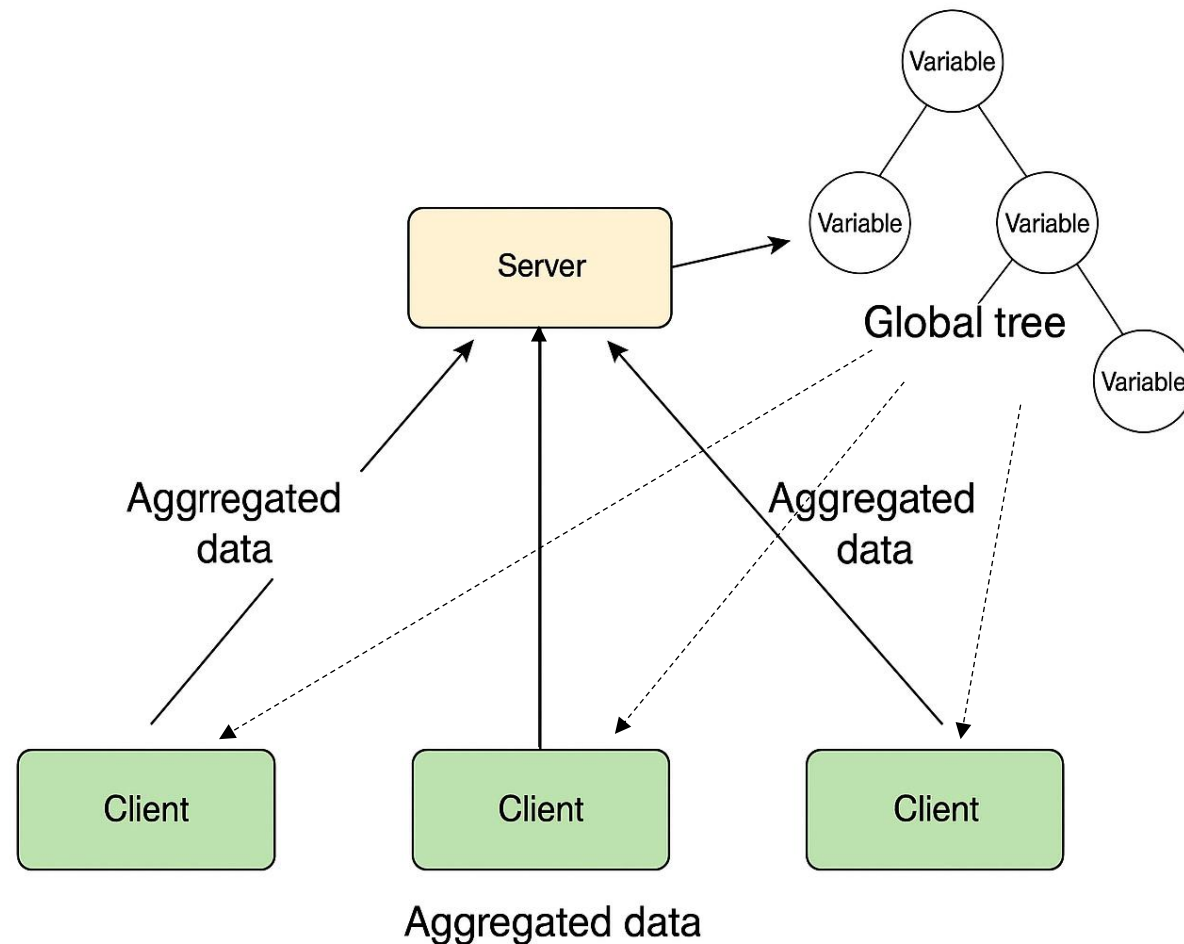
- Directed acyclic graph
- Generated in a top-down way by performing **recursive partitions of the attribute space**.
- Typically, requires a **fuzzy partition defined upon each continuous attribute**.



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

# Federated FDT and Federated FRT

- Solution which preserves the privacy and produces the same trees as applying the algorithm to the joining of all local datasets
- Iteratively, **clients transmit aggregated data to the server**, which uses this information to **select the variable for the next decision node** in the global tree. The updated tree is then shared with all clients.



# Performance

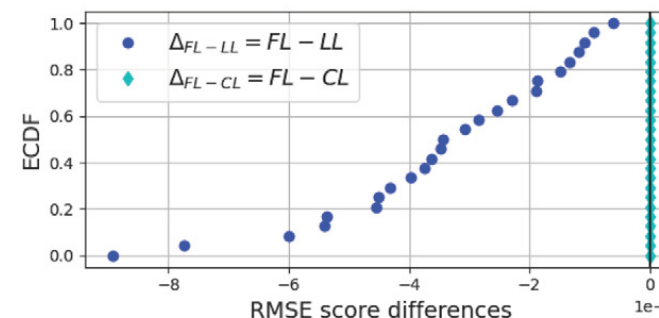
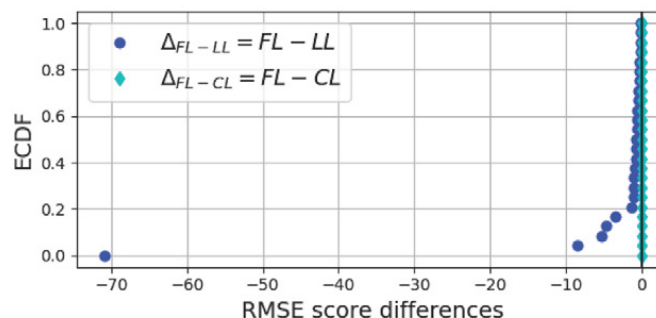
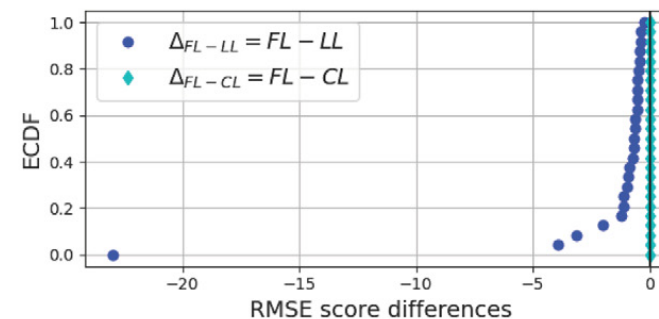
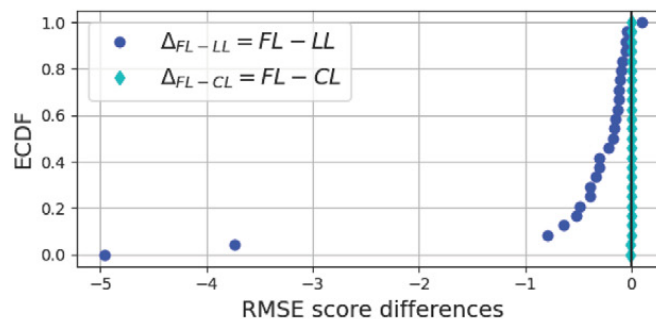
FL=federated Learning

LL=Local Learning

CL=Centralised Learning

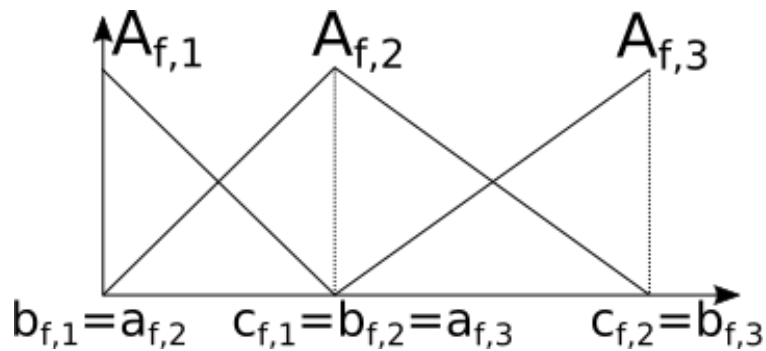
empirical cumulative  
distribution functions  
(ECDFs):

statistical function that  
represents the proportion  
(or count) of observations  
in a sample that are less  
than or equal to a given  
value.



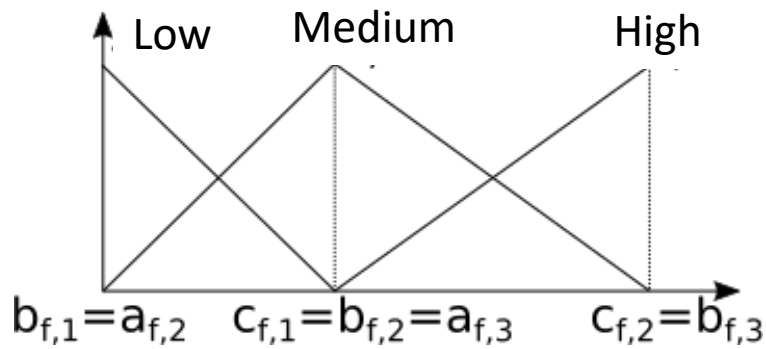
# Interpretable Models: Fuzzy Rule-Based Systems

- The model consists of a rule-base, i.e., a collection of rules in the form  
**if «antecedent» then «consequent»**
- Example of rules in the form *first-order Takagi-Sugeno-Kang Fuzzy Rule-Based Systems*



$$\text{IF } X_1 \text{ IS } A_{1,j_{k,1}} \dots \text{ AND } X_F \text{ IS } A_{F,j_{k,F}}$$

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

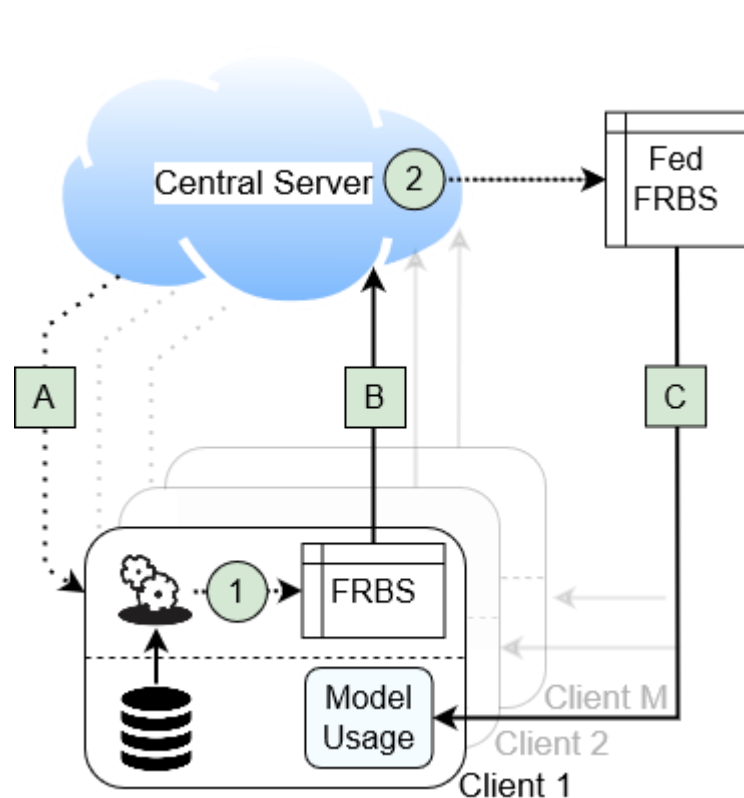


Size

$$\text{IF } \textit{Size} \text{ IS } \textit{low} \dots \text{ AND } X_F \text{ IS } A_{F,j_{k,F}}$$

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

# Federated Learning of 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

# Post-Hoc Explainability: SHAP

- Alternative approach: **Generate a black box model and then try to explain why the inputs produce that output**
- Shapley values quantify the impact of each feature on model prediction

## Shapley Values in XAI

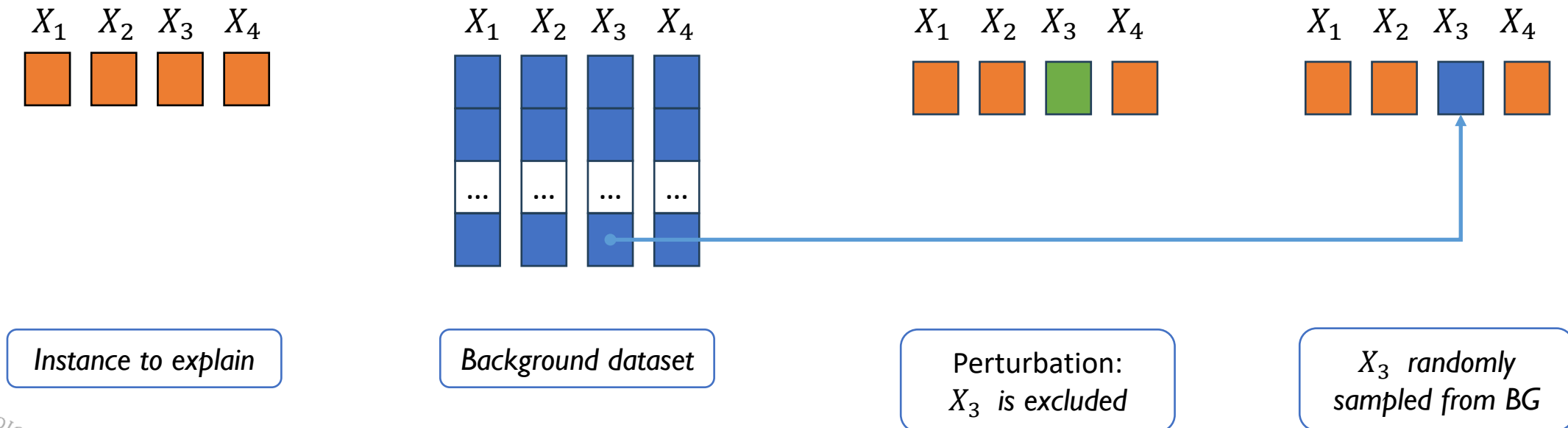
$$\hat{y}_i = f(x_i) = \phi_0 + \sum_{j=1}^F \phi_j$$

- $f$  Predictive model
- $x_i$  Generic  $F$ -dimensional input instance
- $\phi_0$  Average of the predictions from a *background dataset*
- $\phi_i$  Shapley values

- **Local**, i.e., explains individual predictions
- **KernelShap** variant: linear regression-based approximation
  - More **efficient** than naive calculation
  - **Model-agnostic**, suited for both classification and regression tasks

# Challenge of adopting SHAP in the FL setting

- Estimation of Shapley values for the explanations for  $x_i$  involves testing *coalitions* of features by *perturbing*  $x_i$
- A **background dataset (BG)** is exploited for perturbing  $x_i$ 
  - Replace features excluded from a coalition with those of instances randomly sampled from BG
  - The BG should coincide with the set of data used for learning the  $f$  model (i.e., the **training set**)
  - It is a common practice to reduce the numerosity of the BG (e.g., through sampling)



# Challenge of adopting SHAP in the FL setting

## Challenges

- The choice of the **background dataset** impacts the resulting explanations
- In the FL setting the **training set is not available in its entirety** to any party

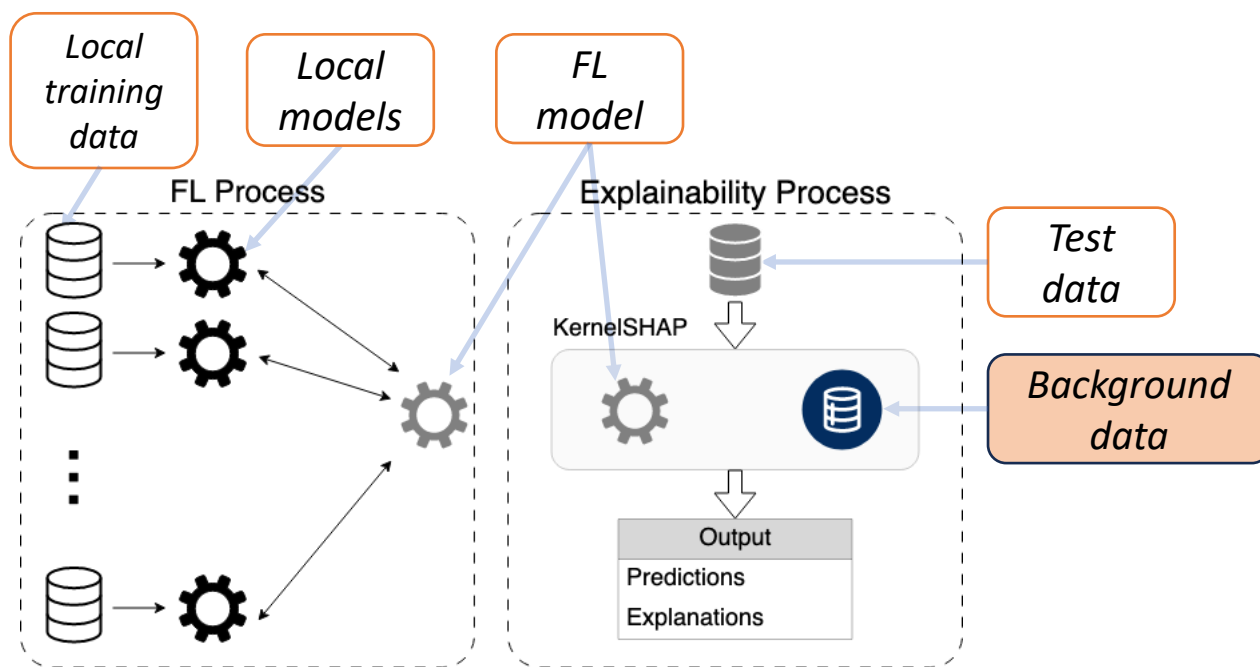
## Desiderata

- **Privacy preservation:** the explainability process should not violate privacy (as a constrain of the FL setting)
- **Consistency:** explanations of the same data instance for the FL model are identical for different participants
- **Accuracy:** explanations in FL match those that would be obtained in the traditional centralized setting



# Federated SHAP – how to design a *proper* and *common* background dataset

- **Start** communication topology with **horizontally** partitioned data
- The model learned in a federation fashion is **opaque** (it requires post-hoc techniques)
- **non-i.i.d. setting**: local data follow distributions different from each other and from the overall distribution



## Background dataset generation through Federated Fuzzy Clustering

- Privacy preserving summarization of scattered data
- **Cluster centers** are exploited as background
  - common, i.e., shared to all participants
  - representative of the entire data distribution
- **Federated-FCM\*** is adopted but the choice of the clustering algorithm is not critical for our objective

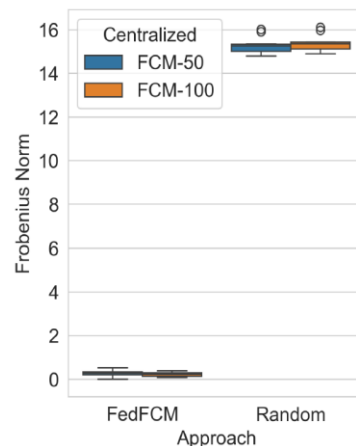
J. L. C. Bárcena, F. Marcelloni, A. Renda, A. Bechini and P. Ducange, "Federated C-Means and Fuzzy C-Means Clustering Algorithms for Horizontally and Vertically Partitioned Data," in *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 12, pp. 6426-6441, Dec. 2024

# Experimental setup – Baseline approaches

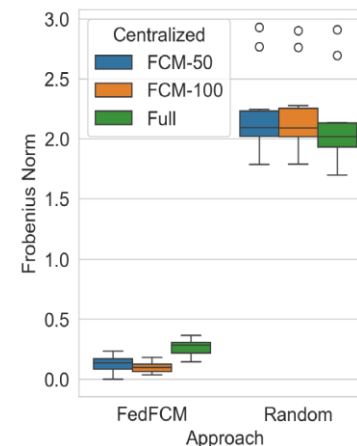
| *BG = Background dataset                                                                                                                    | Ensure consistency<br>(same background<br>for all participants) | Ensure accuracy<br>(represent the actual<br>overall data distribution) | Preserve privacy |
|---------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------|------------------------------------------------------------------------|------------------|
| <b>Federated SHAP</b><br><b>BG</b> $\leftarrow$ $K$ cluster centers obtained<br>through Federated FCM                                       | ✓                                                               | ✓                                                                      | ✓                |
| <b>Centralized</b><br><b>BG</b> $\leftarrow$ union of the data locally stored<br>in the clients                                             | ✓                                                               | ✓                                                                      | ✗                |
| <b>Random</b><br><b>BG</b> $\leftarrow$ randomly sampling $K$ instances from<br>a uniform distribution over the input space                 | ✓                                                               | ✗                                                                      | ✓                |
| <b>Local<sup>m</sup></b><br><b>BG<sup>m</sup></b> $\leftarrow$ $K$ cluster centers obtained through<br>local FCM on the $m$ -th participant | ✗                                                               | ✗                                                                      | ✓                |

# Accuracy of explanations

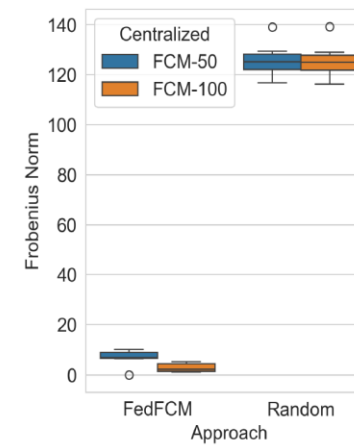
- **Accuracy**  $\stackrel{\text{def}}{=}$  explanations match those that would be obtained in the traditional *centralized* setting
- Three *centralized* versions
  - **BG**  $\leftarrow$  Full training
  - **BG**  $\leftarrow$  FCM, 50 centers
  - **BG**  $\leftarrow$  FCM, 100 centers
- Ten values for each approach with different random seed



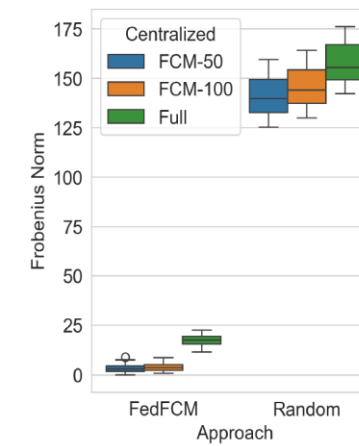
(a) Magic



(b) Rice



(c) California



(d) Abalone

Discrepancy of both the **Federated SHAP (FedFCM)** and the **Random** approach with the baseline centralized approaches in terms of Frobenius norm of the pairwise difference of  $\phi$  matrices

- **Federated SHAP:** low discrepancy with the *centralized* case, low variability
- **Random:** high discrepancy with the *centralized* case, high variability

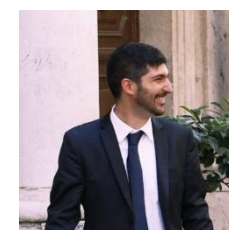
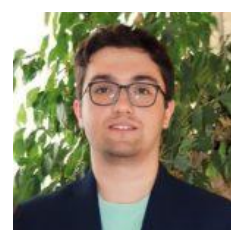


# Challenges and future directions

- Not Independent and Identically distributed (**Non-IID**) data
- Federated Learning with **Streaming Data**
- **Privacy protection**: It is still not clear to what extent these methods harm the data privacy, and there is no quantitative measures to identify the degree of privacy leakage.
- **Large number of hyperparameters** (total number of clients, number of local epochs, client dropout probability)
- Lack of **universally recognized benchmark datasets**
- Research on **Federated Clustering** is still very limited although there exist a number of interesting application domains

**THANK  
YOU**

I would like to thank the members of the AI group at the Department of Information Engineering. Federated Learning and Explainable AI are also very active research areas at the Department of Computer Science.



**Questions?**

For questions and details, please write to Francesco Marcelloni ([francesco.marcelloni@unipi.it](mailto:francesco.marcelloni@unipi.it))