A new approach for cross-silo Federated Learning and its privacy risks

Michele Fontana, Francesca Naretto, Anna Monreale
University of Pisa
michele.fontana@phd.unipi.it

TABLE OF CONTENTS

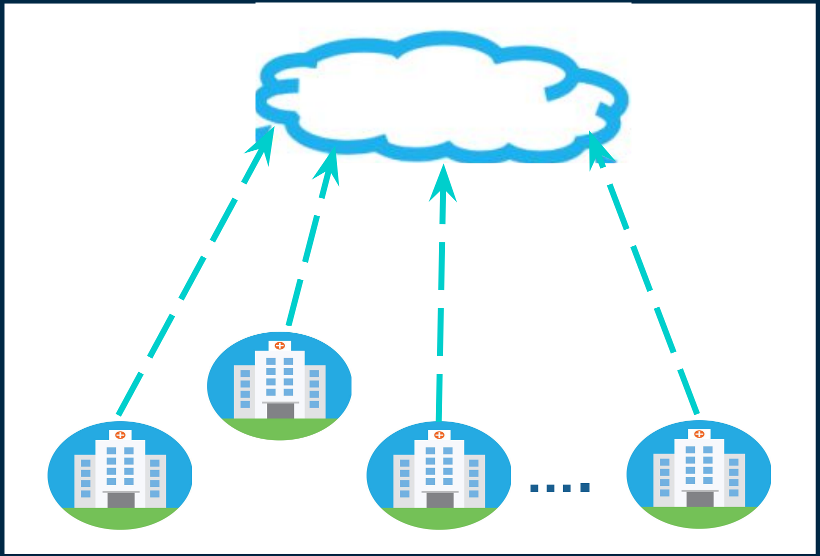
Federated Learning

A (simple?) ML Problem

Challenge : Train a ML classifier on clinical data distributed over a set of hospitals to determine the best therapy for a given patient.

Possible Solution: (Distributed Learning)

- Send data to a central **server**
- **Privacy issue**:
  - Clinical data are ***sensitive!***
  - **They must be kept private**

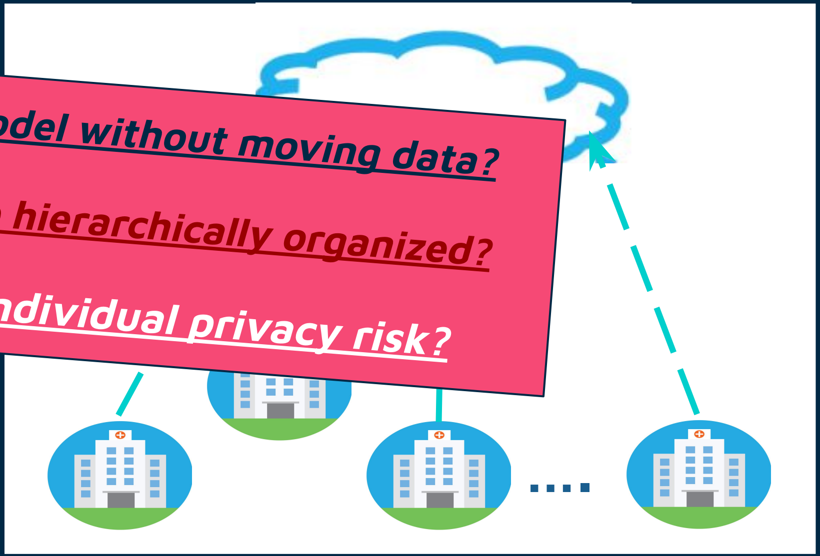


A (simple?) ML Problem

Challenge : Train a ML cla
distributed over a set of h
the best therapy for a giv

Possible Solution: (Dist

- Send data to a ce
- **<u>Privacy issue</u>:**
    - Clinical data are ***sensitive!***
    - **They must be kept private**



<u>Can we train the model without moving data?</u>

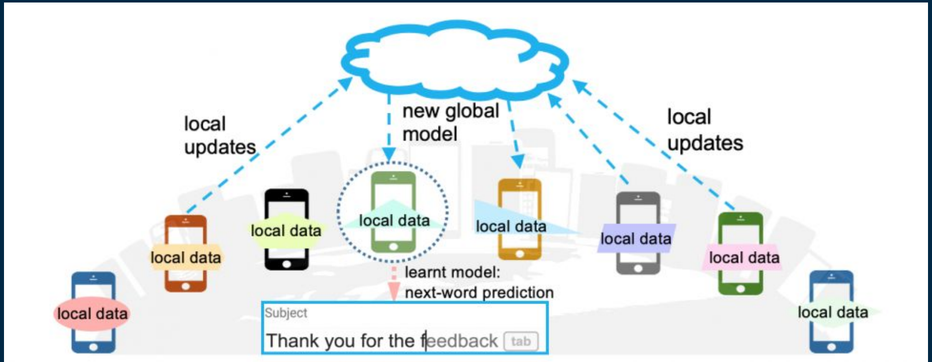
<u>What if the data are hierarchically organized?</u>

<u>What about the individual privacy risk?</u>

Federated Learning

"Federated Learning is a ML setting where multiple distributed parties, called clients, under the orchestration of a main server, cooperate to train a shared global model, while keeping their data private"

- Just the **model parameters** are transmitted
- The overall architecture is called **federation**



https://blog.ml.cmu.edu/category/federated-learning/

Federated Optimization: Distributed Machine Learning for On-Device Intelligence, McMahan et al. , 2016, CoRR

Our setting

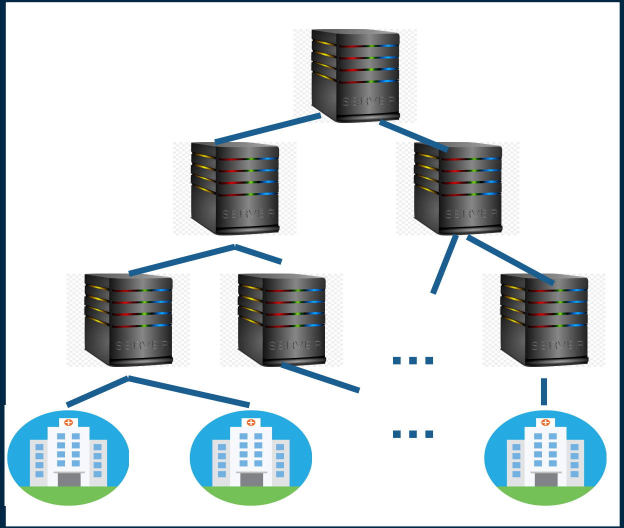
We focus on a specific FL setting

Cross-Silo:

- Clients are organizations (hospitals, banks)
- Private ICT infrastructures
- Unlimited resources

Hierarchical

- Layers of proxies between clients and server



Advances and Open Problems in Federated Learning, Kairouz et al. , 2019

Our approach:

HOLDA

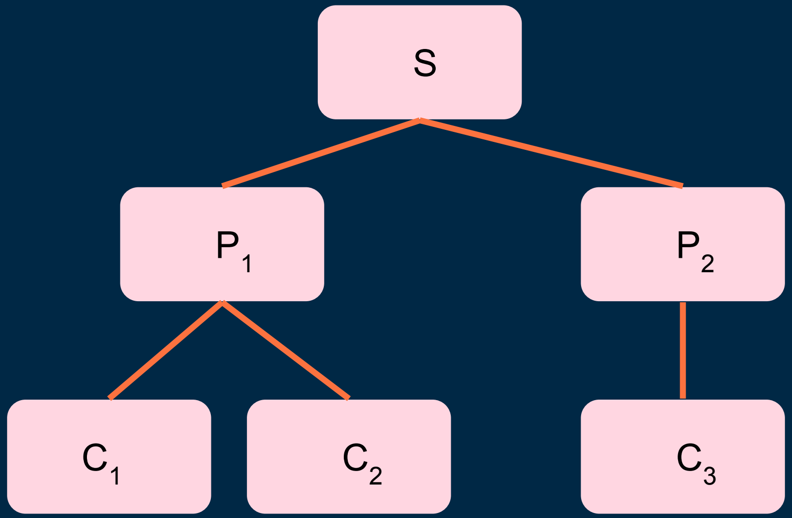
HOLDA: A new FL Training Algorithm

Hierarchical crOss siLo feDerated Averaging

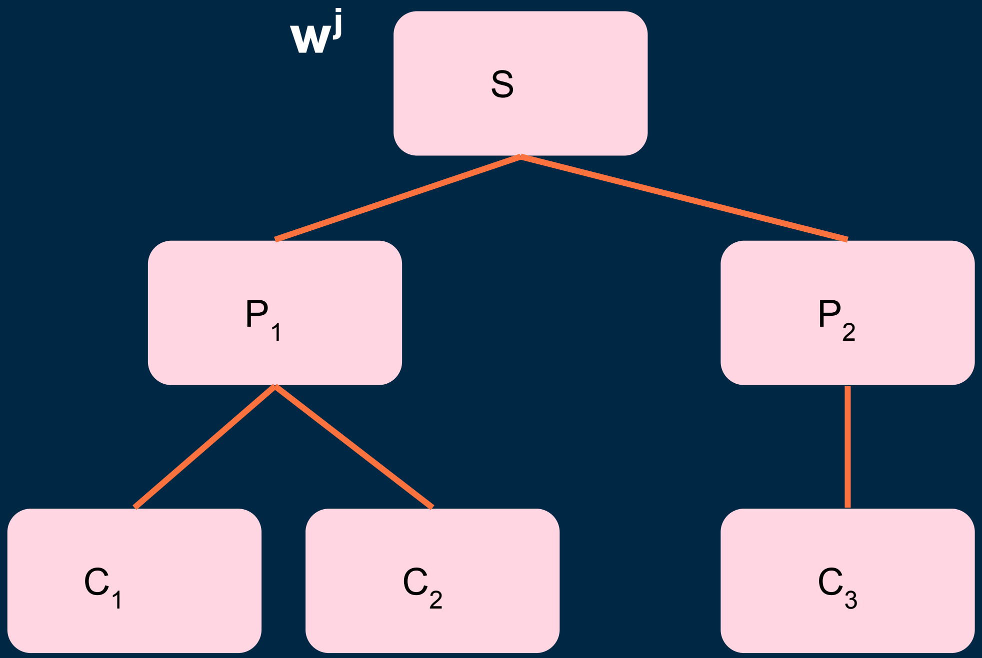
- **Train Neural Networks in a cross-silo hierarchical setting**
- Handles any hierarchical architecture
  - No **assumptions** about the structure of the federation
- Cross-silo -> The participants are **stateful**

What about the internal state?

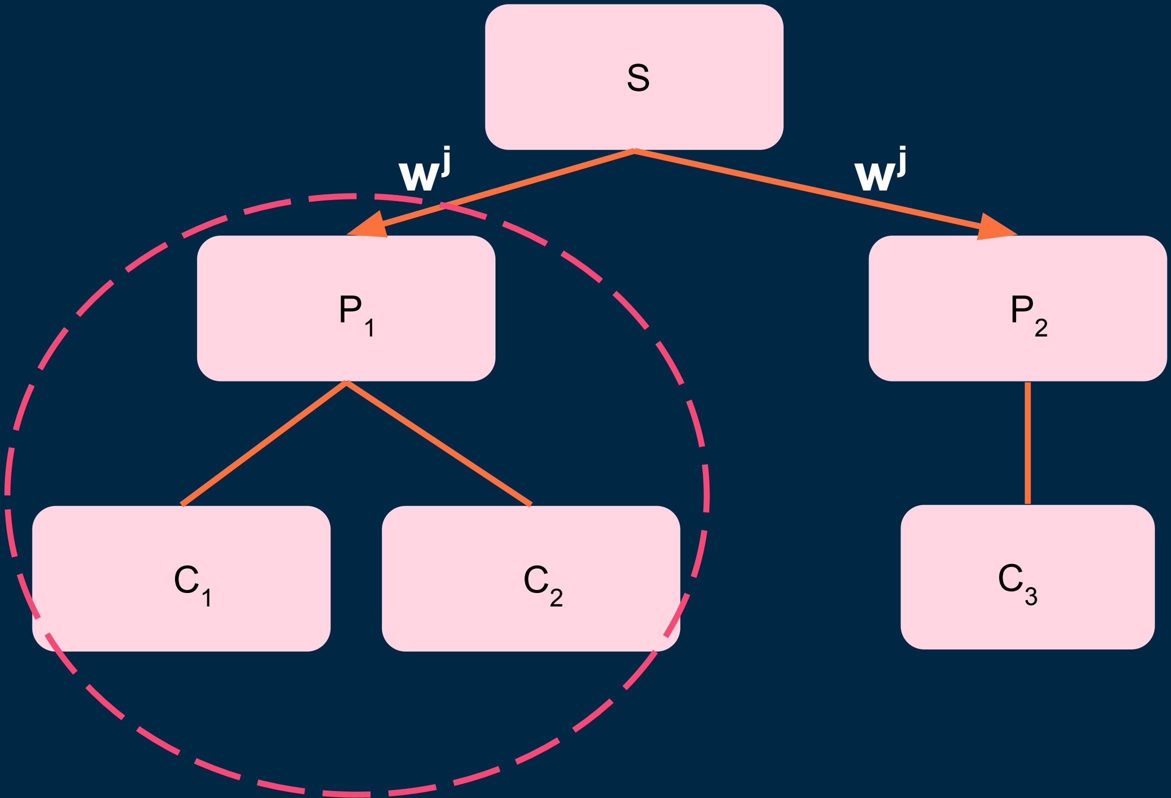
- **One** state per node
- **Best** generalizing model parameters : $\mathbf{w}_{best}$
- **Score** obtained by the best model on the validation data : $\mathbf{M}_{best}$



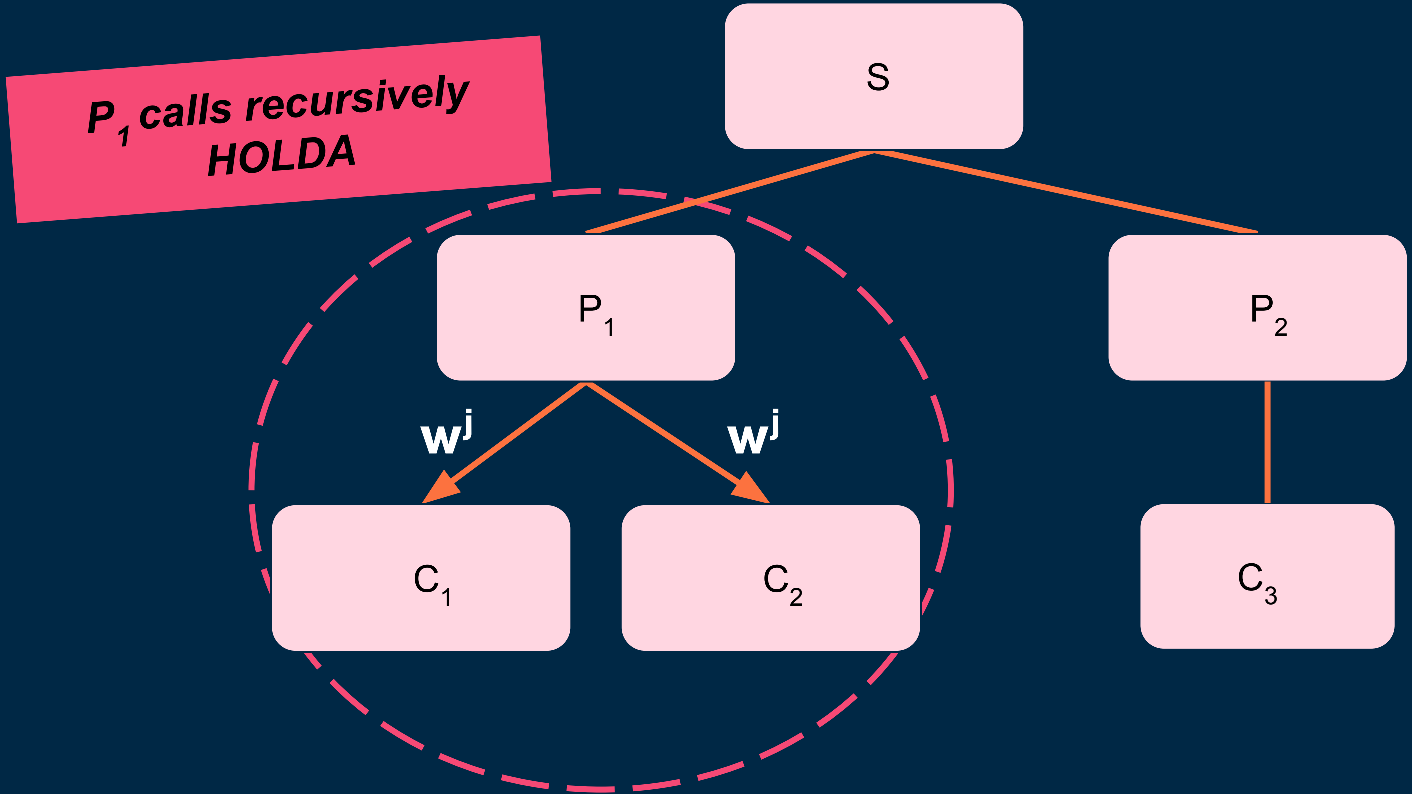
HOLDA: A working example



HOLDA: A working example

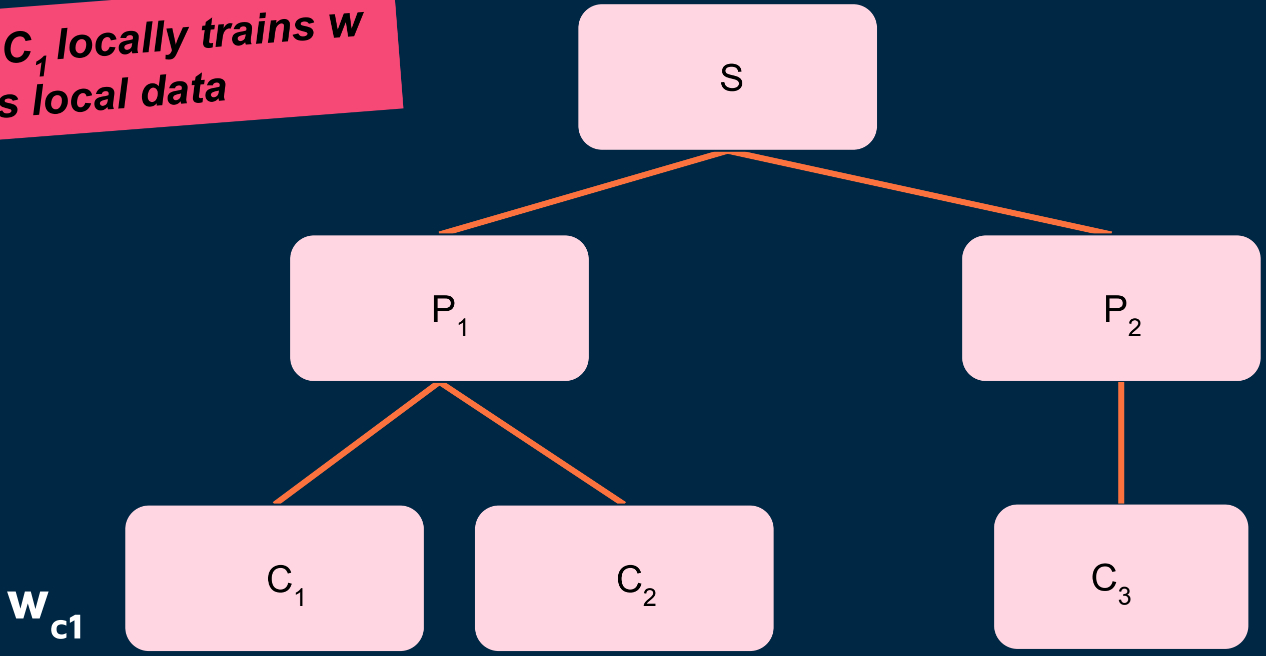


HOLDA: A working example



HOLDA: A working example

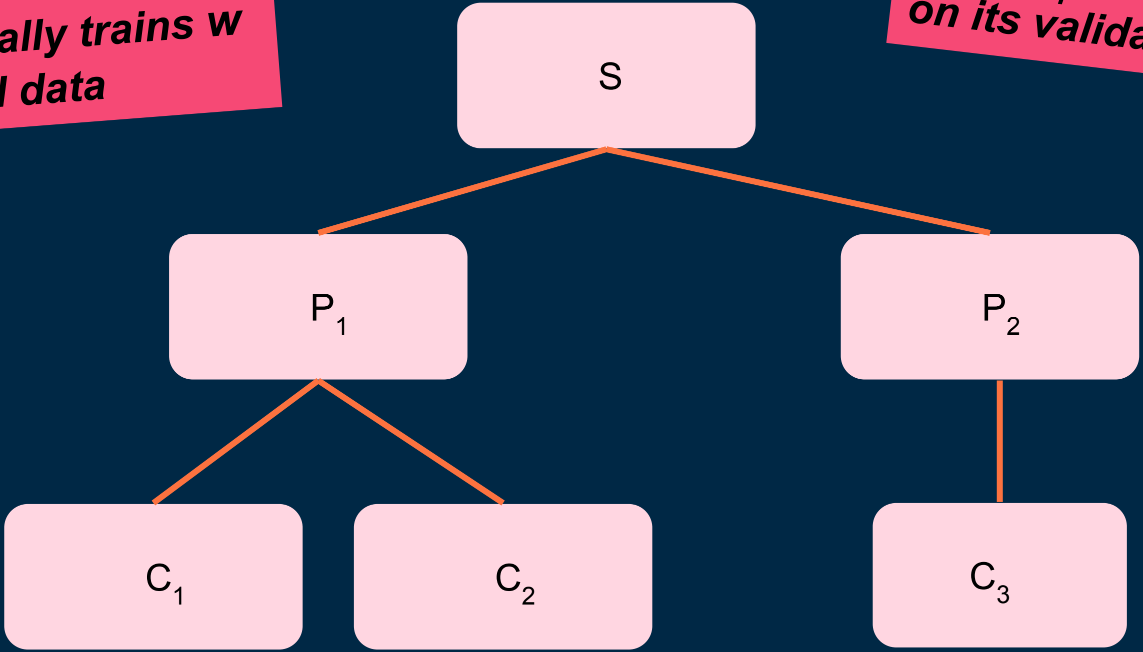
W_{c1} = C_1 locally trains w_j on its local data



HOLDA: A working example

$W_{c1} = C_1$ locally trains w_j on its local data

$M_{c1} = C_1$ evaluates w_{c1} on its validation data

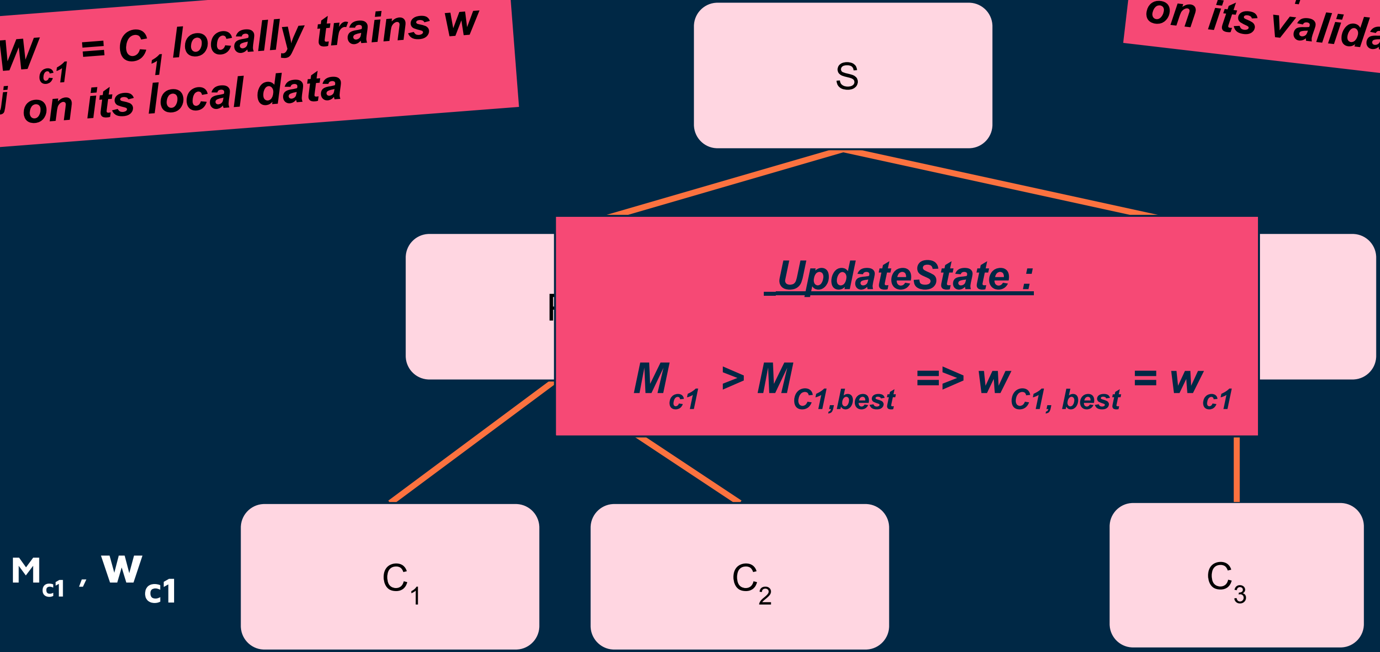


M_{c1} , W_{c1}

HOLDA: A working example

w_{c1} = C_1 locally trains w_j on its local data

M_{c1} = C_1 evaluates w_{c1} on its validation data



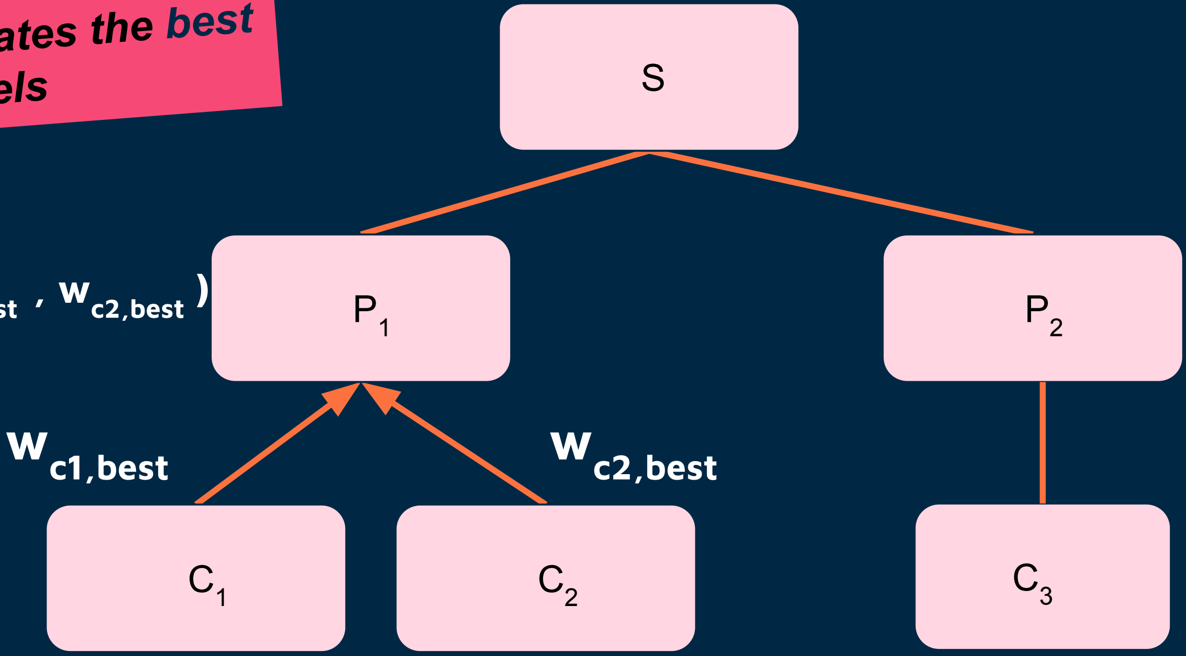
UpdateState :

$$M_{c1} > M_{C1,best} \Rightarrow w_{C1,best} = w_{c1}$$

HOLDA: A working example

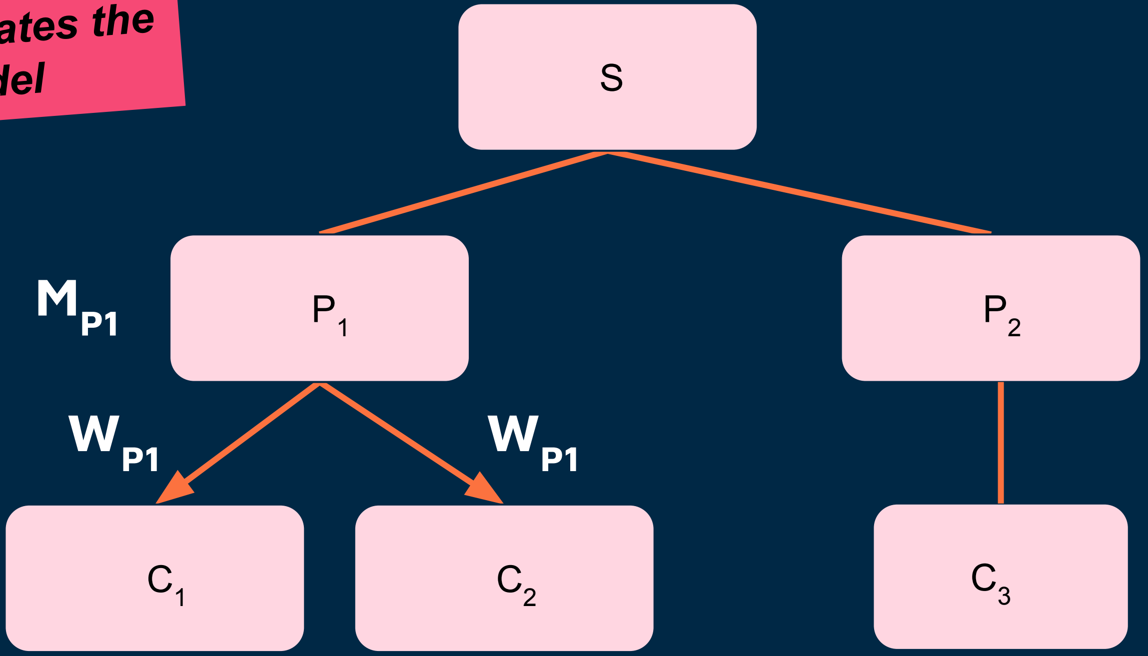
P*$_1$ *aggregates the best local models

$W_{p1} = f(w_{c1,best}\ ,\ w_{c2,best})$

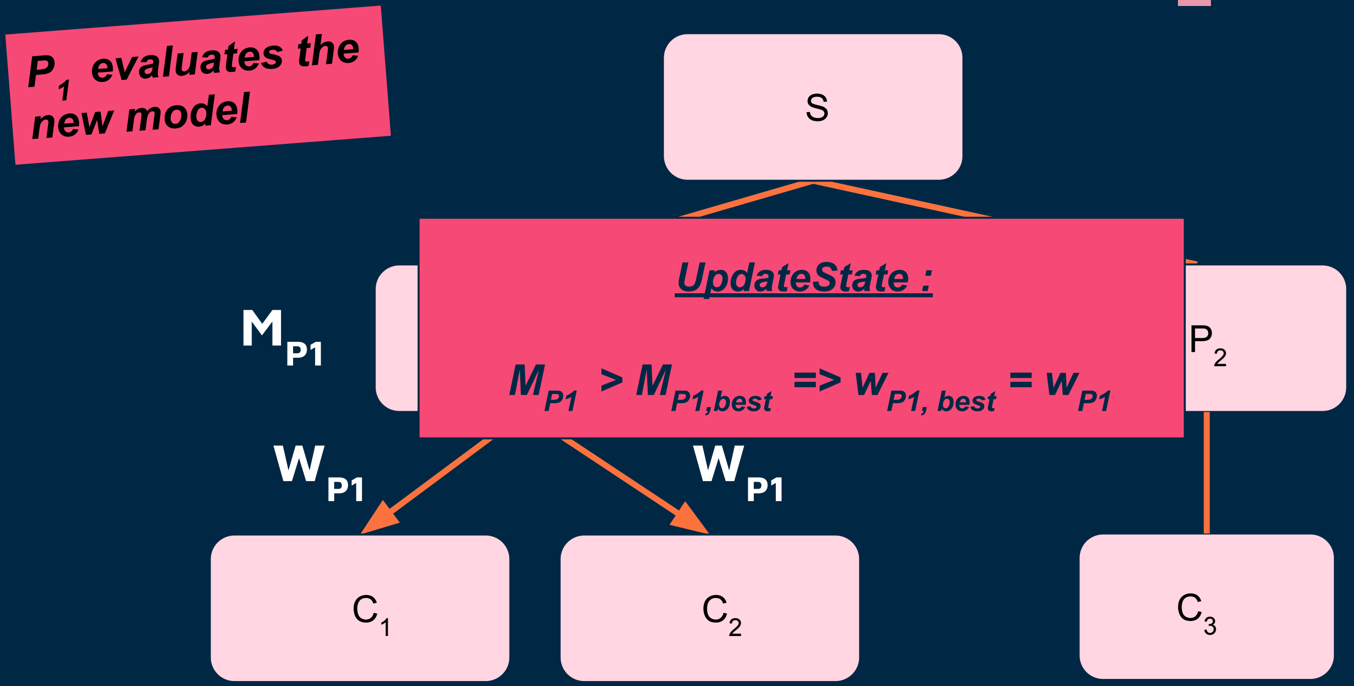


HOLDA: A working example

P_1 evaluates the new model



HOLDA: A working example

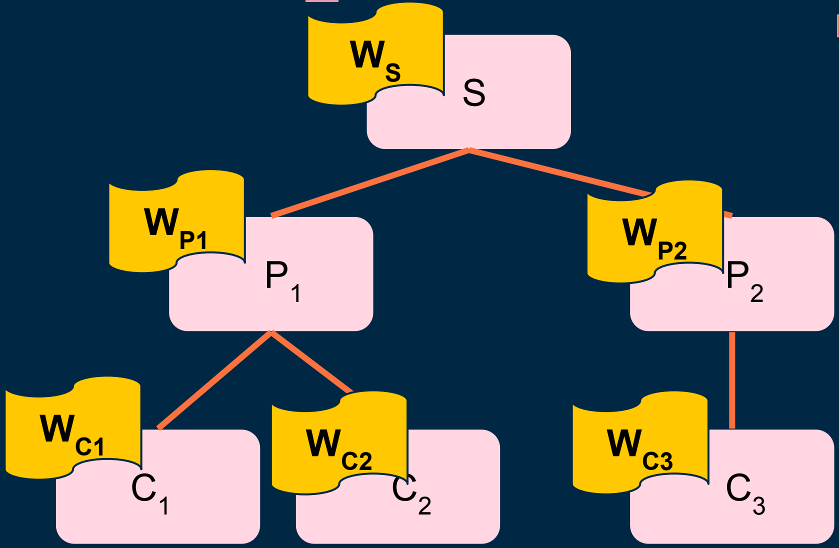


Final Remarks

When the global server ends its training process each node of the federation:

- Has the best model stored into its internal state → **personalized model**

Once the models are public an external adversary can attack the models to get personal information about the users in the training data



Privacy Risk *Assessment*

Membership Inference Attack

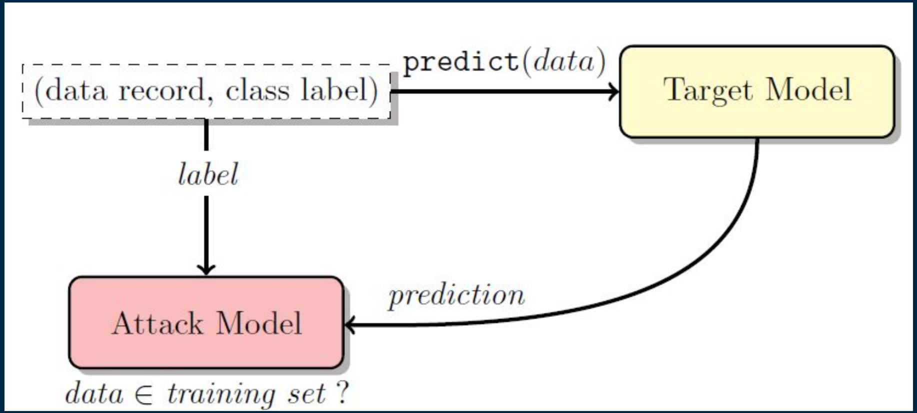
Black-box Attack

Target Model:

- Solves a classification problem with *n* classes.
- Output : Probability vector of length *n*

Attack Model:

- **Supervised Binary Classifier**
- Output : ***"in"*** / ***"out"***



Membership inference attacks against machine learning models, *Shokri et al. , IEEE, 2017*

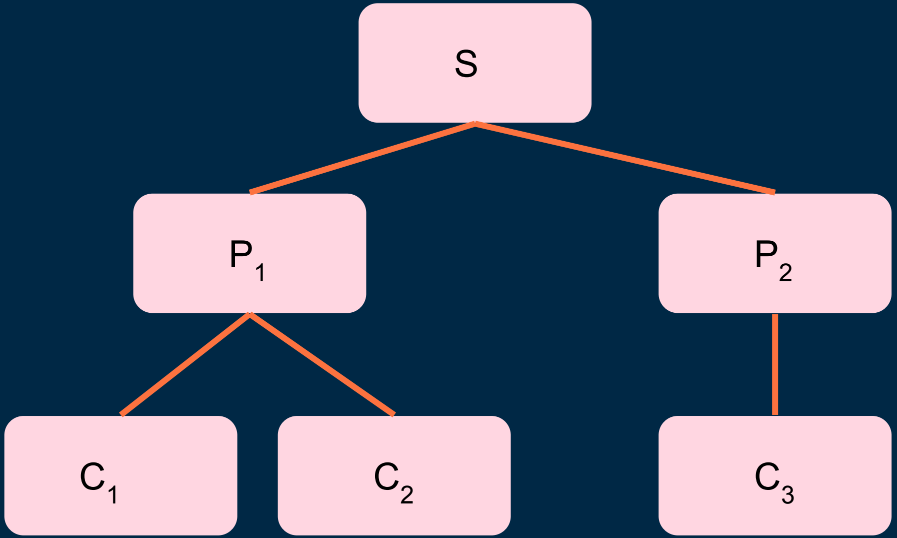
MIA in a Federated Scenario

D_c := Local Dataset of client c

T_v := Training data of node v

- ***T_v :=*** union of the training sets stored into the clients in the subtree of node *v*
- MIA has to detect the membership w.r.t. ***T_v***

$$T_v = \begin{cases} D_v & \text{if } v \text{ is a client} \\ \{T_c \mid c \in \text{children}(v)\} & \text{otherwise} \end{cases}$$



MIA in a Federated Scenario

D**$_c$* **:= *Local Dataset of client c

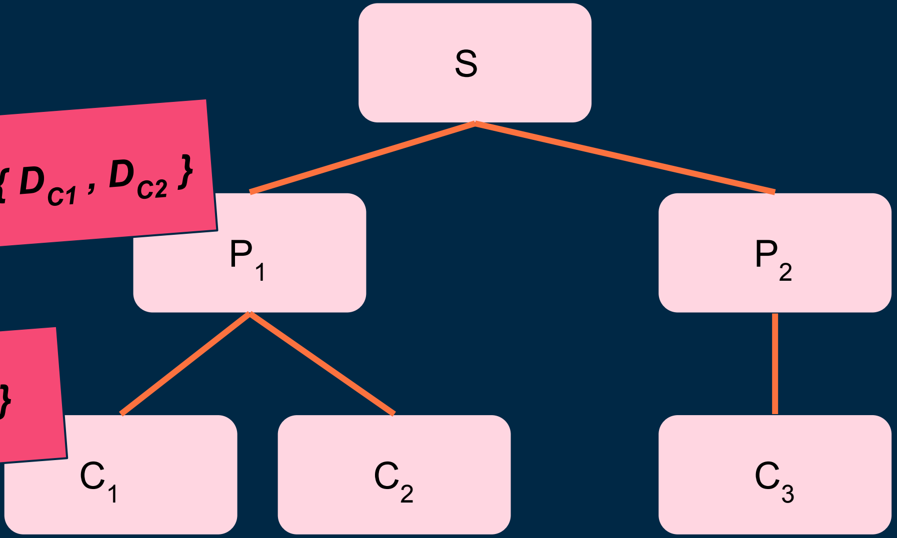
T**$_v$* **:= Training data of node *v

- MIA has to detect the membership w.r.t. ***T**$_v$*
- The adversary acquires **different information** according to the model he is attacking

$$T_v = \begin{cases} D_v & \text{if } v \text{ is a client} \\ \{T_c \mid c \in \text{children}(v)\} & \text{otherwise} \end{cases}$$

MIA(x, P$_1$) := x ∈ { D$_{C1}$, D$_{C2}$ }

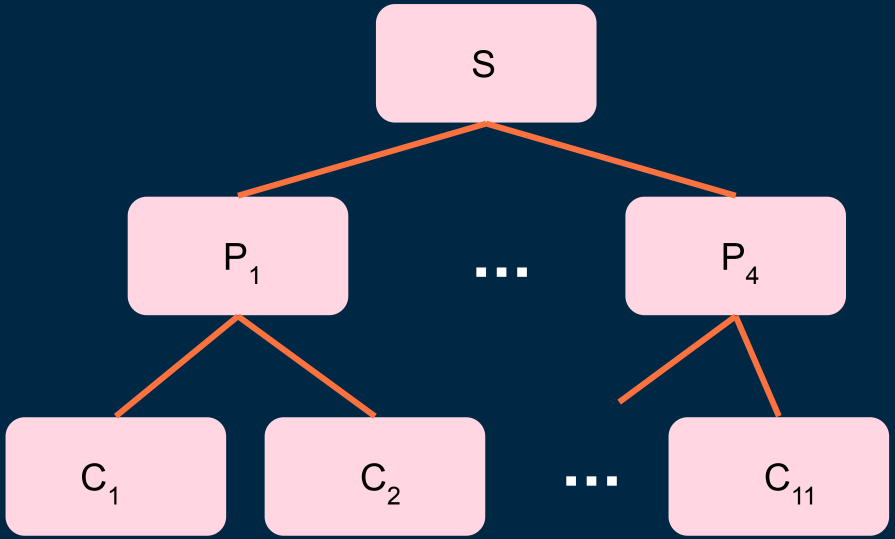
MIA(x, C$_1$) := x ∈ { D$_{C1}$ }



Experimental *Results*

Dataset

- Publicly available **Texas-100** (Over 5,000,000 records)
- **Tabular** Dataset
- Hospitalizations in over 200 Texas hospitals between 2006 and 2009
- **Model**: Feed-forward NN with 2 hidden layers
- Clients grouped on a geographical basis



https://www.dshs.texas.gov/thcic/hospitals/Inpatientpudf.shtm

Results : HOLDA

Kind	Model	Metric	Training	Validation	*Kind*	Model	Metric	Training	Validation
G	All	F_1	0.80 (0.04)	0.78 (0.00)	I	C	F_1	0.83 (0.01)	0.80 (0.00)
		Prec	0.82 (0.00)	0.80 (0.00)			*Prec*	0.85 (0.01)	0.82 (0.00)
		Rec	0.79 (0.00)	0.77 (0.00)			*Recall*	0.83 (0.01)	0.79 (0.00)
I	E	F_1	0.83 (0.01)	0.79 (0.00)	I	S	F_1	0.86 (0.01)	0.82 (0.00)
		Prec	0.85 (0.01)	0.81 (0.00)			*Prec*	0.87 (0.01)	0.83 (0.00)
		Rec	0.83 (0.01)	0.79 (0.00)			*Rec*	0.86 (0.01)	0.81 (0.00)
I	NW	F_1	0.89 (0.01)	0.81 (0.00)	L	Reg1-NW	F_1	0.88 (0.01)	0.84 (0.00)
		Prec	0.89 (0.01)	0.82 (0.00)			*Prec*	0.89 (0.01)	0.84 (0.00)
		Rec	0.90 (0.01)	0.80 (0.01)			*Rec*	0.88 (0.01)	0.83 (0.00)
L	Reg2-C	F_1	0.93 (0.01)	0.85 (0.00)	L	Reg3-C	F_1	0.86 (0.01)	0.82 (0.00)
		Prec	0.93 (0.01)	0.86 (0.00)			*Prec*	0.87 (0.01)	0.83 (0.00)
		Rec	0.93 (0.01)	0.85 (0.00)			*Rec*	0.86 (0.01)	0.82 (0.00)
L	Reg4-E	F_1	0.89 (0.01)	0.84 (0.00)	L	Reg5-E	F_1	0.91 (0.01)	0.84 (0.00)
		Prec	0.90 (0.01)	0.85 (0.00)			*Prec*	0.92 (0.01)	0.85 (0.00)
		Rec	0.89 (0.01)	0.84 (0.00)			*Rec*	0.91 (0.01)	0.84 (0.00)
L	Reg6-E	F_1	0.86 (0.01)	0.82 (0.00)	L	Reg7-C	F_1	0.86 (0.01)	0.82 (0.00)
		Prec	0.88 (0.01)	0.83 (0.00)			*Prec*	0.88 (0.01)	0.84 (0.00)
		Rec	0.86 (0.01)	0.81 (0.00)			*Rec*	0.85 (0.01)	0.82 (0.00)
L	Reg8-S	F_1	0.89 (0.01)	0.84 (0.00)	L	Reg9-NW	F_1	0.88 (0.00)	0.84 (0.00)
		Prec	0.90 (0.01)	0.85 (0.00)			*Prec*	0.88 (0.00)	0.85 (0.00)
		Rec	0.88 (0.01)	0.83 (0.00)			*Rec*	0.89 (0.00)	0.84 (0.00)
L	Reg10-NW	F_1	0.88 (0.01)	0.84 (0.00)	L	Reg11-S	F_1	0.88 (0.01)	0.84 (0.00)
		Prec	0.89 (0.01)	0.85 (0.00)			*Prec*	0.90 (0.01)	0.85 (0.00)
		Rec	0.87 (0.01)	0.83 (0.00)			*Rec*	0.88 (0.01)	0.83 (0.00)

Results : Privacy Risk

Results of the “in” class → how many users in the training data are correctly identified

Kind	Model	Metric	RF	*Kind*	Model	Metric	RF
G	All	*Prec* *Rec*	0.72 0.76	I	C	*Prec* *Recall*	0.71 0.75
I	E	*Prec* *Rec*	0.72 0.76	I	S	*Prec* *Rec*	0.81 0.75
I	NW	*Prec* *Rec*	0.78 0.76	L	Reg1-NW	*Prec* *Rec*	0.83 0.80
L	Reg2-C	*Prec* *Rec*	0.82 0.80	L	Reg3-C	*Prec* *Rec*	0.74 0.78
L	Reg4-E	*Prec* *Rec*	0.83 0.85	L	Reg5-E	*Prec* *Rec*	0.82 0.87
L	Reg6-E	*Prec* *Rec*	0.75 0.76	L	Reg7-C	*Prec* *Rec*	0.83 0.77
L	Reg8-S	*Prec* *Rec*	0.82 0.76	L	Reg9-NW	*Prec* *Rec*	0.83 0.88
L	Reg10-NW	*Prec* *Rec*	0.83 0.81	L	Reg11-S	*Prec* *Rec*	0.83 0.79

Conclusions

- ***HOLDA***: novel approach tailored for training NN in the **hierarchical cross-silo** setting
  - Strategy for the **selection of the best weights**.
- Privacy risk assessment of the models trained with ***HOLDA***, simulating the ***MIA***
- Experimental results
  - ***HOLDA*** models can reach good predictive performance
  - ***Privacy risk is not negligible***, especially at the client level.
- Future work : Develop new **mitigation** strategies to lower the risk, without impacting on the performance

Do you have any questions?

michele.fontana@phd.unipi.it
francesca.naretto@sns.it
anna.monreale@unipi.it

THANKS

Non-Hierarchical Results

Kind	Model	Metric	Training	Validation	*Kind*	Model	Metric	Training	Validation
G	All	F_1	0.78 (0.01)	0.76 (0.01)	L	Reg1-NW	F_1	0.85 (0.01)	0.81 (0.01)
		$Prec$	0.80 (0.01)	0.79 (0.01)			$Prec$	0.86 (0.01)	0.82 (0.01)
		Rec	0.78 (0.00)	0.76 (0.01)			Rec	0.84 (0.01)	0.80 (0.00)
L	Reg2-C	F_1	0.89 (0.01)	0.84 (0.01)	L	Reg3-C	F_1	0.83 (0.01)	0.81 (0.01)
		$Prec$	0.89 (0.01)	0.85 (0.01)			$Prec$	0.84 (0.01)	0.82 (0.01)
		Rec	0.88 (0.01)	0.84 (0.01)			Rec	0.85 (0.01)	0.82 (0.01)
L	Reg4-E	F_1	0.85 (0.01)	0.82 (0.01)	L	Reg5-E	F_1	0.87 (0.01)	0.83 (0.00)
		$Prec$	0.86 (0.01)	0.84 (0.01)			$Prec$	0.88 (0.01)	0.84 (0.01)
		Rec	0.85 (0.01)	0.82 (0.01)			Rec	0.87 (0.01)	0.82 (0.00)
L	Reg6-E	F_1	0.82 (0.01)	0.80 (0.00)	L	Reg7-C	F_1	0.83 (0.01)	0.81 (0.01)
		$Prec$	0.84 (0.01)	0.82 (0.01)			$Prec$	0.82 (0.01)	0.83 (0.01)
		Rec	0.81 (0.01)	0.80 (0.00)			Rec	0.85 (0.01)	0.80 (0.00)
L	Reg8-S	F_1	0.84 (0.01)	0.82 (0.01)	L	Reg9-NW	F_1	0.88 (0.01)	0.82 (0.00)
		$Prec$	0.86 (0.00)	0.84 (0.01)			$Prec$	0.89 (0.01)	0.83 (0.00)
		Rec	0.83 (0.00)	0.81 (0.01)			Rec	0.88 (0.01)	0.82 (0.00)
L	Reg10-NW	F_1	0.85 (0.01)	0.81 (0.01)	L	Reg11-S	F_1	0.84 (0.01)	0.82 (0.01)
		$Prec$	0.86 (0.01)	0.83 (0.01)			$Prec$	0.85 (0.01)	0.83 (0.01)
		Rec	0.85 (0.01)	0.81 (0.01)			Rec	0.83 (0.01)	0.81 (0.01)

Results : Hierarchical

Kind	Model	Metric	Training	Validation	*Kind*	Model	Metric	Training	Validation
G	All	F_1	0.80 (0.04)	0.78 (0.00)	I	C	F_1	0.83 (0.01)	0.80 (0.00)
		Prec	0.82 (0.00)	0.80 (0.00)			*Prec*	0.85 (0.01)	0.82 (0.00)
		Rec	0.79 (0.00)	0.77 (0.00)			*Recall*	0.83 (0.01)	0.79 (0.00)
I	E	F_1	0.83 (0.01)	0.79 (0.00)	I	S	F_1	0.86 (0.01)	0.82 (0.00)
		Prec	0.85 (0.01)	0.81 (0.00)			*Prec*	0.87 (0.01)	0.83 (0.00)
		Rec	0.83 (0.01)	0.79 (0.00)			*Rec*	0.86 (0.01)	0.81 (0.00)
I	NW	F_1	0.89 (0.01)	0.81 (0.00)	L	Reg1-NW	F_1	0.88 (0.01)	0.84 (0.00)
		Prec	0.89 (0.01)	0.82 (0.00)			*Prec*	0.89 (0.01)	0.84 (0.00)
		Rec	0.90 (0.01)	0.80 (0.01)			*Rec*	0.88 (0.01)	0.83 (0.00)
L	Reg2-C	F_1	0.93 (0.01)	0.85 (0.00)	L	Reg3-C	F_1	0.86 (0.01)	0.82 (0.00)
		Prec	0.93 (0.01)	0.86 (0.00)			*Prec*	0.87 (0.01)	0.83 (0.00)
		Rec	0.93 (0.01)	0.85 (0.00)			*Rec*	0.86 (0.01)	0.82 (0.00)
L	Reg4-E	F_1	0.89 (0.01)	0.84 (0.00)	L	Reg5-E	F_1	0.91 (0.01)	0.84 (0.00)
		Prec	0.90 (0.01)	0.85 (0.00)			*Prec*	0.92 (0.01)	0.85 (0.00)
		Rec	0.89 (0.01)	0.84 (0.00)			*Rec*	0.91 (0.01)	0.84 (0.00)
L	Reg6-E	F_1	0.86 (0.01)	0.82 (0.00)	L	Reg7-C	F_1	0.86 (0.01)	0.82 (0.00)
		Prec	0.88 (0.01)	0.83 (0.00)			*Prec*	0.88 (0.01)	0.84 (0.00)
		Rec	0.86 (0.01)	0.81 (0.00)			*Rec*	0.85 (0.01)	0.82 (0.00)
L	Reg8-S	F_1	0.89 (0.01)	0.84 (0.00)	L	Reg9-NW	F_1	0.88 (0.00)	0.84 (0.00)
		Prec	0.90 (0.01)	0.85 (0.00)			*Prec*	0.88 (0.00)	0.85 (0.00)
		Rec	0.88 (0.01)	0.83 (0.00)			*Rec*	0.89 (0.00)	0.84 (0.00)
L	Reg10-NW	F_1	0.88 (0.01)	0.84 (0.00)	L	Reg11-S	F_1	0.88 (0.01)	0.84 (0.00)
		Prec	0.89 (0.01)	0.85 (0.00)			*Prec*	0.90 (0.01)	0.85 (0.00)
		Rec	0.87 (0.01)	0.83 (0.00)			*Rec*	0.88 (0.01)	0.83 (0.00)