Trustworthy Distributed AI and Machine Learning – Challenges and Opportunities

Sara Bouchenak
Professor – INSA Lyon

IFIP WG 10.4 Winter Meeting
Kaunas, Lithuania, May 4-7, 2026

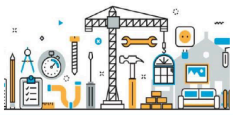
1

Sensitive Information in the Age of Edge and Cloud Computing

Healthcare

Finance

IoT



Manufacturing

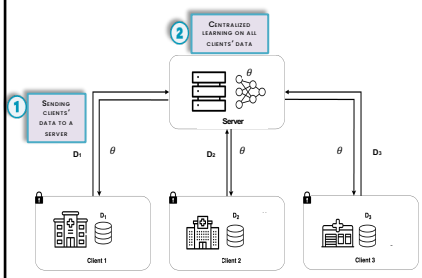
Data Protection Regulations

- European Union, 2018: GDPR (General Data Protection Regulation)
- Canada, 2000: PIPEDA (Personal Information Protection and Electronic Documents Act)
- California, 2020: CCPA (California Consumer Privacy Act)
- Australia, 2024: Privacy Act Amendment Act

2

2

From Centralized Learning

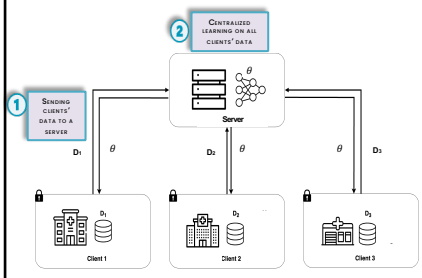


Classical Centralized Learning

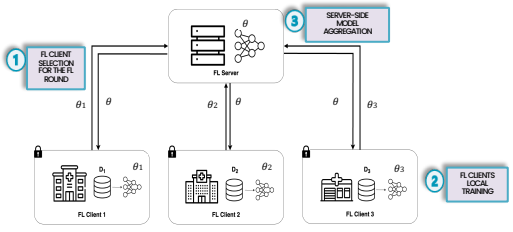
3

3

From Centralized Learning to Federated Learning (FL)



Classical Centralized Learning



Federated Learning for Better Privacy

4

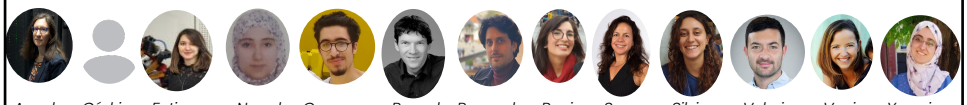
4

Ongoing Work on Trustworthy Federated Learning

- **Privacy**
  - PASTEL. ACM UbiComp 2024 (*Rank A**)
  - DINAR Middleware 2024 (*Rank A*)
  - Projects: TARANIS (ANR PEPR Cloud)
- **Fairness and bias**
  - Survey. ACM Comp. Surveys 2025 (*Rank A**)
  - ASTRAL. ACM UbiComp 2024 (*Rank A**)
  - Projects: M4DI (ANR PEPR Digital health), TARANIS (ANR PEPR Cloud), CITADEL (ANR)
- **Robustness**
  - ARMOR. ACM UbiComp 2023 (*Rank A**)
  - Projects: TARANIS (PEPR Cloud)

GENDER BALANCE

50% of women & 50% of men



Angela *Cédric* *Fatima* *Nawel* *Ousmane* *Pascal* *Pasquale* *Rania* *Sara* *Silvia* *Valerio* *Vania* *Yasmine*

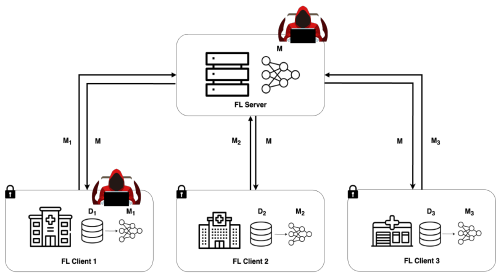
Univ. of Lyon Univ. of Neuchâtel Univ. of Grenoble

5

5

(Remaining) Privacy Threats in FL

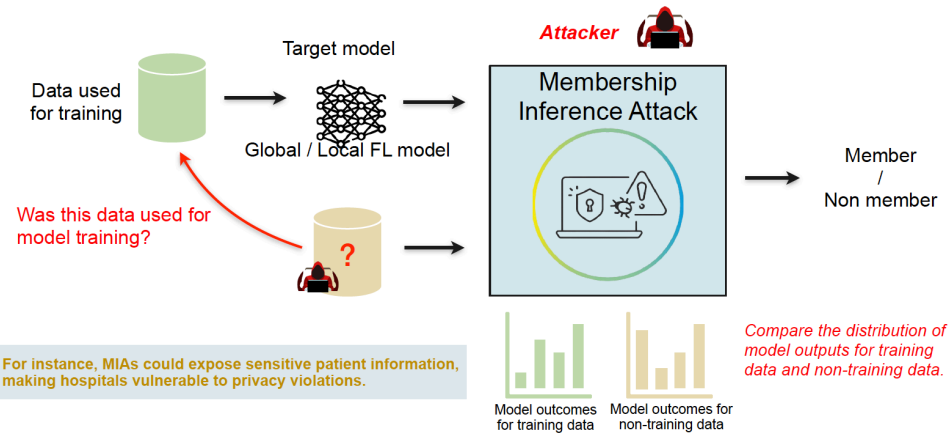
- A malicious participant (at client-side, or server-side)
- Ability to analyze model parameters and infer sensitive/private information



6

6

Membership Inference Attacks



For instance, MIAs could expose sensitive patient information, making hospitals vulnerable to privacy violations.

Compare the distribution of model outputs for training data and non-training data.

7

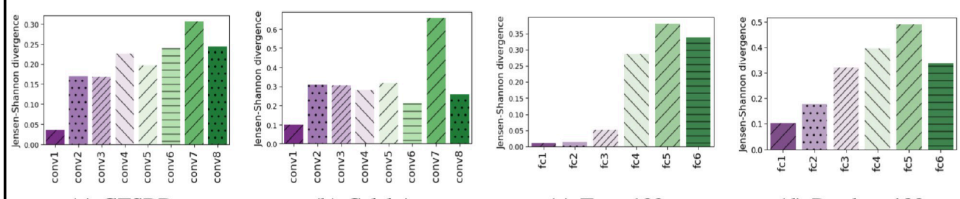
7

DINAR: Intuition Behind the Proposed Approach

➤ ***Ravi's talk – frontal brain***

Layer-specific analysis of training data **leakage** across models.

Consistent patterns indicate that the **penultimate layer** exhibits the highest level of information leakage (ICLR, 2021).



(a) GTSRB (b) CelebA (c) Texas100 (d) Purchase100

Neural network's layer-level analysis of divergence between member data samples and non-member data samples, when FL models are not protected against MIAs

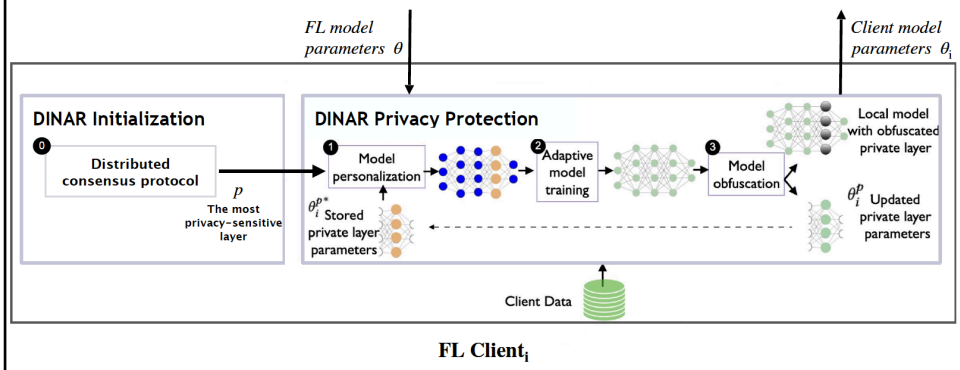
F. Mo, et al. Layer-Wise Characterization of Latent Information Leakage in Federated Learning. ICLR 2021 Workshop on Distributed and Private Machine Learning. Virtual.

8

8

DINAR: Fine-Grained Privacy-Preserving Federated Learning

➤ *Ravi's talk – frontal brain plasticity*

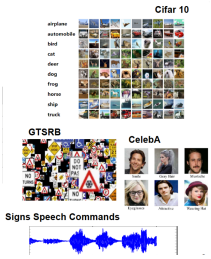


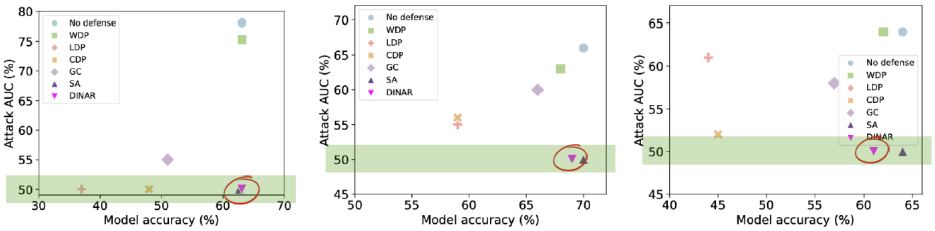
FL Client$_i$

9

9

How Effective Is Privacy Protection vs. Model Utility?





(a) Purchase100 (b) Cifar-10 (c) Cifar-100

Trade-off between privacy and utility for local models in different FL defense scenarios

10

10

➤ *Ravi's talk – what about the hippocampus?*

Towards a More General Approach for Privacy-Preserving Federated Learning

PASTEL: Privacy as a First-Class Citizen in Federated Learning

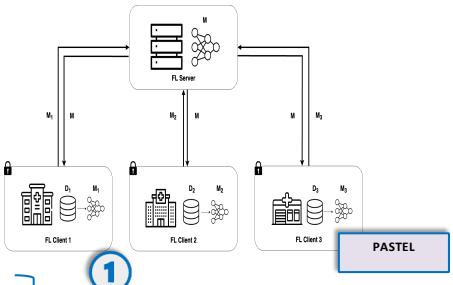
Algorithm 1 *PASTEL* algorithm: $\mathcal{B} \times \mathcal{V} \times \mathcal{W}_t \rightarrow \mathcal{W}_{i,t+1}$

Global Model $\mathcal{W}_t$
Loss Function $\mathcal{L}$
Training Batch $(\mathcal{B}, \mathcal{Y}) = \{(B_1, Y_1), \ldots, (B_x, Y_x)\}$
Validation Batch $\mathcal{V} = \{V_i, \ldots, V_x\}$
Local Epochs $\mathcal{E}$
// **Initilization**
$\mathcal{W}_{i,t+1} = \mathcal{W}_t$
for epoch $\in \mathcal{E}$ **do**
  for (B_i, Y_i) in $(\mathcal{B}, \mathcal{Y})$ **do**
  // **Perform forward pass**
    $\widehat{Y}_i = W_t(B_i)$
  // **Compute model loss**
    $l_{label} = \mathcal{L}(Y_i, \widehat{Y}_i)$
  // **Compute JSD loss**
    $l_{JSD} = JS(B_i, V_i)$
  // **Compute gradient**
    $\nabla_t = AdaGrad((l_{label}, l_{JSD}), \mathcal{W}_t)$
  // **Update local model**
    $\mathcal{W}_{i,t+1} = \mathcal{W}_{i,t+1} + \nabla \mathcal{W}_{i,t+1}$
  end for
end for

Model loss

&

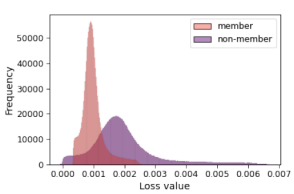
Privacy loss



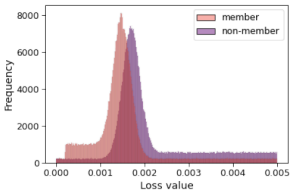
Novel loss function

Novel method to quantify membership leakage at model layer level

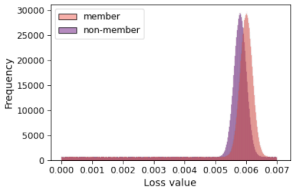
MIA Attacker's Point of View



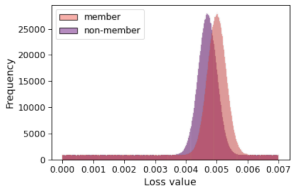
(a) No Defense



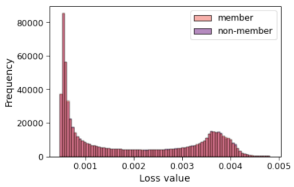
(b) *WDP*



(c) *LDP*



(d) *CDP*



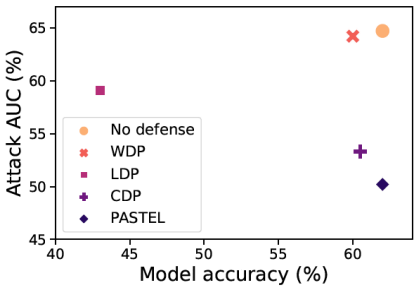
(e) *PASTEL*

Model loss – Purchase100 dataset, with a 6 fully onnected layer network

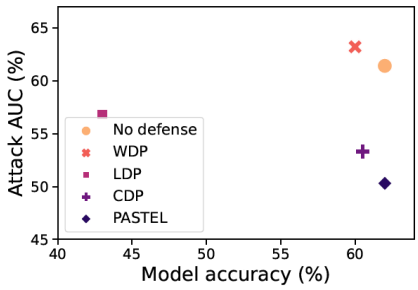
13

13

How Effective Is Privacy Protection?



(a) Attacks on clients' local models



(b) Attacks on FL global model

Model privacy vs. model utility – Cifar-10 dataset, with ResNet20

14

14

Ongoing Work on Trustworthy Federated Learning

- ***Privacy***
  - *PASTEL. ACM UbiComp 2024 (Rank A*)*
  - *DINAR Middleware 2024 (Rank A)*
- **Fairness and bias**
  - **ASTRAL. ACM UbiComp 2024 (*Rank A**)**
  - Survey. ACM Comp. Surveys 2025 (*Rank A**)
- Robustness
  - ARMOR. ACM UbiComp 2023 (*Rank A**)

15

15

What Is Bias?

- Demographic bias is the presence of favoritism or discrimination that unequally affects some **demographic groups** based on their **sensitive attributes**
- Sensitive attributes cover demographic characteristics that have the potential to lead to discriminatory behavior, e.g., gender, race, age, etc.


Gender


Age


Race

16

16

Bias in AI: A Threat to Fairness

Recruitment

RETAIL OCTOBER 11, 2018 / 1:04 AM / UPDATED 5 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Law enforcement

PROPUBLICA — Donate

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

Healthcare

TOM SIMONITE BUSINESS OCT 24, 2019 2:00 PM

A Health Care Algorithm Offered Less Care to Black Patients

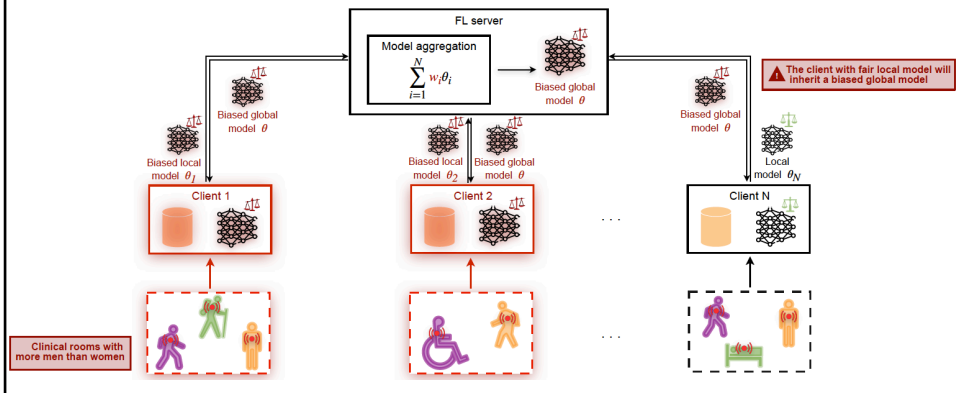
A study shows the risks of making decisions using data that reflects inequities in American society.

17

17

Why Bias Matters More in Federated Learning?

Federated Learning introduces new sources of bias through its distributed and collaborative nature, with **bias being propagated across clients** during model updates (ICLR,2023).



Chang, H., & Shokri, R. Bias Propagation in Federated Learning. ICLR 2023, Kigali, Rwanda.

18

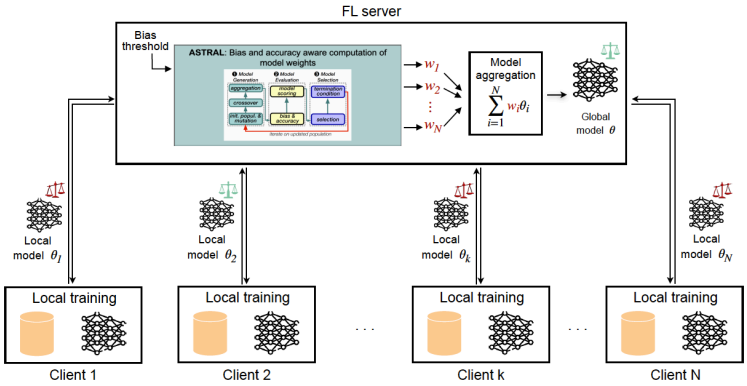
18

ASTRAL: Bias Mitigation in Federated Learning

➤ *The amount of data is not all what matters*

$$\theta = \sum_{i=1}^{N} \frac{|D_i|}{|D_1 \cup \cdots \cup D_N|} \cdot \theta_i$$

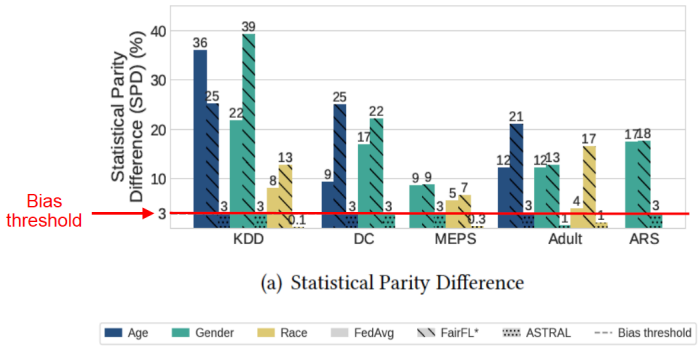
➤ *A novel bias-aware FL model aggregation approach*



19

19

Comparison of ASTRAL with Existing FL Bias Mitigation Mechanisms



Bias mitigation with regard to multiple sensitive attributes – KDD, DC, MEPS, Adult and ARS datasets – With a bias threshold of 3%

20

20

Ongoing Work on Trustworthy Federated Learning

- ***Privacy***
  - *PASTEL. ACM UbiComp 2024 (Rank A*)*
  - *DINAR Middleware 2024 (Rank A)*
- **Fairness and bias**
  - *ASTRAL. ACM UbiComp 2024 (Rank A*)*
  - **Survey. ACM Comp. Surveys 2025 (*Rank A**)**
- Robustness
  - ARMOR. ACM UbiComp 2023 (*Rank A**)

21

21

Thank you!

DINAR Paper		DINAR Software prototype	
PASTEL Paper		PASTEL Software prototype	
ASTRAL Paper		ASTRAL Software prototype	
Survey on FL Bias			

22

22