Federated Learning

Lecture 3 – Personalization & Client Selection
Escuela de Ciencias Informáticas (ECI) 2026

Universidad de Buenos Aires
Ioannis Chatzigiannakis

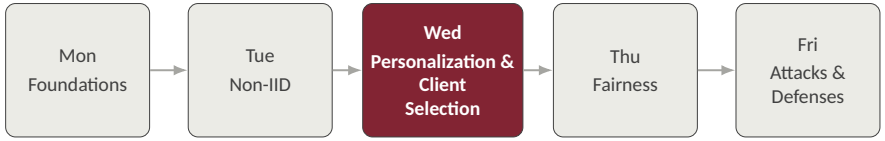
Sapienza University of Rome

ECI 2026 • Buenos Aires • Lecture 3 of 5

SAPIENZA
UNIVERSITÀ DI ROMA

This week schedule

Where today's lecture fits



Recall (Tue). Non-IID data was quantified with Hellinger distance, Jensen–Shannon distance, Earth Mover's Distance; Dirichlet(α) and Similarity(S) partitions gave us controllable non-IID testbeds.

Today. Two complementary answers to "one client, one model" vs. "one global model for everyone": personalization (train tailored models, e.g. via clustering) and client selection (choose which clients participate, each round, to fight non-IID drift and reduce communication cost).

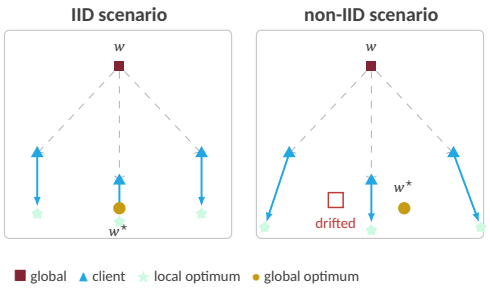
Federated learning under non-IID data

Local optima drift away from the global optimum

The problem. Real-world clients are *non-IID*: per-client label distributions diverge from the global one.

Consequence.

- **Client drift**: local optima $w_i^\star$ pull away from global optimum $w^\star$.
- Aggregated updates point in *inconsistent* directions.
- Convergence slows; global accuracy degrades; some clients are systematically under-served (**fairness**).



A single global model then under-serves clients whose distribution is far from the population average.

One global model is not enough enough

Three families, one goal

Stop insisting on one model – cluster clients and train one model per cluster → Personalized FL (PFL)

Tan et al. group Personalized FL (PFL) methods into three families (as used throughout Clust-PSI-PFL and FedLECC):

1. **Regularization-based.** Keep local objectives close to the global one: FedProx (proximal term μ), FedOpt/FedAdagrad/FedYogi/FedAdam (adaptive server update), FedAvgM (server momentum), FedNova (normalizes for uneven local steps), FedDyn (dynamic penalty).
2. **Selection-based.** Prioritize or filter *which* clients participate: Power-of-Choice (favors high-loss clients), HACCS (histogram-similarity clusters + fastest client per cluster), FedCLS (Hamming-distance–guided sampling), FedCor (Gaussian-Process client correlations).
3. **Clustering-based.** Partition clients into distributionally homogeneous groups and train one model per group: CFL (recursive cosine-similarity splits), FedSoft (soft mixture clustering), and **Clust-PSI-PFL** (PSI-guided clustering, today's focus).

Part I: Clust-PSI-PFL

Jimenez-Gutierrez et al., IPDPS 2026

A Population Stability Index approach to clustered PFL

Idea in one sentence. Use the $WPSI^L$ metric – not just to *measure* non-IIDness, but to *cluster* clients into distributionally homogeneous groups, and train one FedAvg model per cluster.

Why PSI

$WPSI^L$ is the strongest predictor of the Dirichlet α / Similarity S non-IID level among $WPSI^L$, Hellinger distance (HD), Jensen–Shannon distance (JSD), and Earth Mover's Distance (EMD) – and it decomposes per class, telling us *which* labels drive a client's mismatch.

The Population Stability Index, applied to FL

One scalar per client, plus a per-class decomposition

Client-level PSI (divergence of client i's label pmf from the global one):

$$\mathrm{PSI}_i^L = \sum_{c=1}^{C} \big(P(y=c) - P_i(y=c)\big)\, \ln \frac{P(y=c)}{P_i(y=c)}$$

(Equivalent to Jeffreys divergence.)

Per-class PSI (one term per class):

$$\mathrm{PSI}_{i,c}^L = \big(P(y=c) - P_i(y=c)\big)\, \ln \frac{P(y=c)}{P_i(y=c)}$$

Client feature vector for clustering:

$$\big[\,\mathrm{PSI}_i^L,\ \mathrm{PSI}_{i,1}^L,\ \ldots,\ \mathrm{PSI}_{i,C}^L\,\big]$$

Federation-level summary: $WPSI^L = \sum_{i=1}^{K} \frac{n_i}{N}\, \mathrm{PSI}_i^L$.

Why PSI?

- Computed *a priori* from label counts — no gradient or update needed.
- *Per-class* terms identify which labels drive divergence.
- Cost is $O(KC)$ — negligible vs. training.
- Compatible with MPC / DP on histograms when label counts are sensitive.

Smaller $WPSI^L$ ⇒ more homogeneous federation.

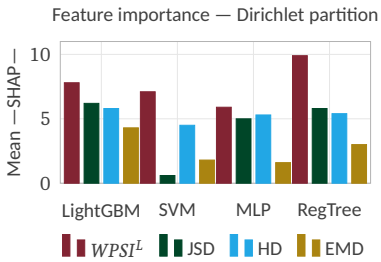
Why PSI — and not HD, JSD, or EMD?

Empirical evidence over 4 predictors on Dirichlet & Similarity partitions

We partition 6 centralized datasets with Dirichlet(α) and Similarity(S) and compare $WPSI^L$ against **Hellinger**, **Jensen–Shannon**, and **Earth Mover's** distances as predictors of the non-IID level.

Findings:

- $WPSI^L$ decays *steeply* as data become more IID (large α / S).
- HD and JSD vary more smoothly — but linearly with $WPSI^L$.
- EMD is non-monotonic in some regimes — *unreliable* as a quantifier.
- Across LightGBM / SVM / MLP / RegTree, $WPSI^L$ ranks #1 by SHAP importance.

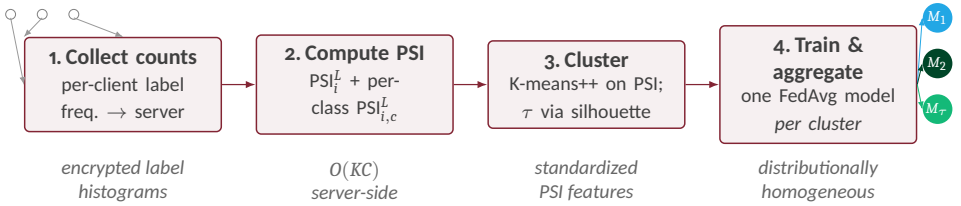


Adapted from Fig. 4 (top) of the paper.

$WPSI^L$ is consistently the strongest predictive metric for the degree of non-IID data.

Clust-PSI-PFL — the pipeline

From label counts to cluster-specific models



Pipeline is run *once before training*; subsequent rounds are standard FedAvg *within* each cluster. Compatible with MPC / DP on the histogram exchange (Section V).

Choosing the number of clusters τ

K-means++ with a silhouette-based stopping rule

Silhouette score. For client i, let $a(i)$ = average distance to others in its own cluster (cohesion) and $b(i)$ = average distance to the nearest other cluster (separation):

$$sil(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \in [-1, 1]$$

Algorithm *(simplified)*

Input: clients $\mathcal{I} = \{1, \ldots, K\}$; for each i, PSI_i^L and $\text{PSI}_{i,c}^L$.
Build $X \in \mathbb{R}^{K \times (C+1)}$ (rows = client PSI signatures); **standardize**.
for $j = 2, \ldots, K-1$ **do**
  run K-means++ with j clusters $\to z^{(j)}$; compute $s(j)$.
return $\tau = \arg\max_j s(j)$, assignments $z^{(\tau)}$.

Complexity (amortized once, before training).

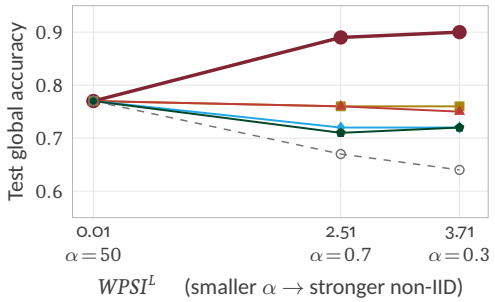
$$\mathcal{O}\Big(KC + (C+1)IK\big(\tfrac{(K-1)K}{2} - 1\big) + (K-2)K^2\Big)$$

where I is the average number of K-means++ iterations. Performed once before FL training and amortized over all communication rounds; in very large deployments, silhouette can be approximated on a subsample.

Across client counts and partition protocols, the silhouette criterion favors a *small* τ (3–4) — compact structure, low overhead.

Global accuracy across the non-IID spectrum

ACSIncome, $K = 100$ — up to $+18\%$ over the strongest cluster baselines



Reading the chart

- Near-IID ($\alpha = 50$): *all methods comparable* — no clustering penalty.
- As α shrinks: baselines collapse, Clust-PSI-PFL holds.
- Under Similarity at $S = 0$: our method reaches **0.98** vs. ≤ 0.76 for the strongest cluster baseline.
- Std. deviations ~ 0.01 — stable training.

Same qualitative trend on all 6 datasets; see Table III.

Clust-PSI-PFL attains higher global accuracy than every competing baseline across the entire non-IID spectrum.

Experimental setup

6 datasets, 3 modalities, 2 partition protocols, 11 baselines

Datasets *(centralized → partitioned)*

Dataset	Modality	Cls.
ACSIncome	Tabular	2
Serengeti	Tabular	13
FMNIST	Image	10
CIFAR10	Image	10
Sent140	Text	3
Amazon Reviews	Text	6

Partition protocols

- **Dirichlet**(α): small $\alpha \Rightarrow$ severe label skew.
- **Similarity**(S): $S \in [0, 1]$; $S = 0$ is pathological.

FL setup

- $K \in \{10, 50, 100\}$ clients; fraction $q = 0.5$ each round
- $T = 40$ rounds, $E = 5$ local epochs
- 5 seeds; mean $\pm$ std reported
- Built on Flower + FedLab

Baselines

- Regularization: FedProx, FedAvgM, FedAdagrad, FedYogi, FedAdam
- Selection: PoC, HACCS, FedCLS
- Clustering: CFL, FedSoft
- References: FedAvg, centralized (CL)

Metrics: global acc., local acc., fairness (AD, SD_{AD}).

Cross-dataset evidence — pathological non-IID

Test accuracy at $K = 100$, Similarity $S = 0$ (5 seeds, mean $\pm$ std)

Method	ACSIncome	Serengeti	FMNIST	CIFAR10	Sent140	Amazon
FedAvg	0.65 ±0.01	0.13 ±0.04	0.14 ±0.04	0.10 ±0.01	0.56 ±0.05	0.30 ±0.03
FedAvgM	0.64 ±0.01	0.15 ±0.03	0.11 ±0.01	0.10 ±0.01	0.53 ±0.05	0.27 ±0.03
HACCS	0.72 ±0.01	0.22 ±0.02	0.16 ±0.04	0.11 ±0.02	0.51 ±0.02	0.28 ±0.03
CFL	0.76 ±0.01	0.37 ±0.01	0.59 ±0.02	0.19 ±0.02	0.53 ±0.01	0.73 ±0.01
FedSoft	0.74 ±0.01	0.20 ±0.01	0.28 ±0.02	0.13 ±0.01	0.51 ±0.02	0.24 ±0.02
Clust-PSI-PFL	**0.98 ±0.01**	**0.98 ±0.01**	**0.98 ±0.01**	**0.98 ±0.01**	**0.98 ±0.01**	**0.98 ±0.01**

Observation. Under pathological label skew, baselines often *collapse* (CIFAR10: ≤ 0.19; FMNIST: ≤ 0.59). Clust-PSI-PFL is essentially *invariant* to the skew because cluster-local distributions are recovered as near-IID.

Selected rows from Table III; near-IID columns (e.g. $S = 1$) are all ≈ 0.77 across methods.

Why does PSI-driven clustering work?

Three structural reasons

1. Quantifying the right thing. PSI was built (in credit scoring) to detect *distributional drift*. The non-IID problem in FL *is* distributional drift. Per-class terms make divergence **attributable**.

2. Acting *before* gradients. CFL splits clients by similarity of *updates*; FedSoft optimizes a soft mixture — both tangle clustering with optimization dynamics. Clust-PSI-PFL clusters on *label histograms*, before any update is produced.

3. Cluster-local data is *near-IID*. Once grouped by PSI signature, the cluster-local mini-federation looks IID — vanilla FedAvg suffices, no per-method tuning required.

Cost profile *(amortized over T rounds)*

Step	Cost	When
PSI per client	$O(KC)$	once, server-side
Cluster selection	$O(IK^3C)$	once, $K-2$ runs
Train in cluster	$\approx$ FedAvg	every round

Subsampling fallback. For very large K, evaluate silhouette on a client *subsample* — selected τ virtually unchanged.

Privacy. Histograms can be exchanged via MPC (no perf. cost) or noised via DP (perf. governed by ε).

Limitations

Honest bookkeeping

1. **No built-in privacy mechanism yet.** Label-count sharing could leak information; MPC-based secure aggregation adds only runtime/communication overhead (no accuracy cost), while DP may trade accuracy for privacy budget ϵ.
2. **Scoped to label skew.** PSI features are class-frequency based; attribute skew or quantity imbalance are not explicitly modeled – clusters may be suboptimal when labels are balanced but attributes/volumes differ.
3. **Small-client noise.** Clients with very few samples can have noisy PSI signatures; mitigations include down-weighting or using n_i as an auxiliary clustering feature.

Not fundamental limits

These are scoped design choices: the PSI machinery extends naturally to discretized attributes, joint label–attribute bins, and sample-count weighting.

From personalization to selection

A different question: who gets to update the model, this round?

Clust-PSI-PFL asked: *"given all clients, how should we group them?"*

FedLECC asks a complementary question: *"each round, which subset of clients should we even bother training on?"*

Biased client selection (recap)

Each round, a subset of m out of K clients is sampled – e.g. in batches of size m, or with replacement, client i selected with probability p_i. Random/uniform sampling ignores that **not all clients are equally informative**.

Personal Data Economy (PDE). FedLECC frames each client's dataset as an economic asset: unique, rare, high-loss data is more valuable to the global model than redundant, easy data – so it should be selected more often.

Cherry-Picking Clients through Clustering and PDE

Jimenez-Gutierrez et al., Journal of Industrial Information Integration

Idea in one sentence. Cluster clients by label-distribution similarity (Hellinger distance), then *within* the highest-loss clusters, pick the highest-loss clients – combining clustering-based diversity with loss-based "cherry picking."

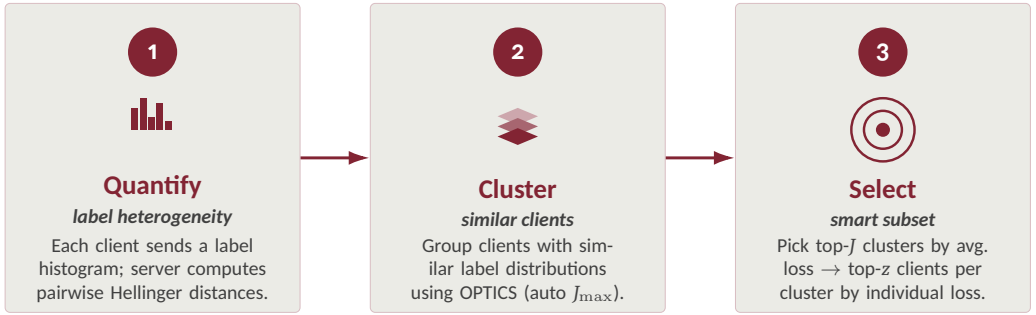
Why loss? Cho et al. showed that favoring high-loss clients speeds convergence (Power-of-Choice). But loss-only selection risks collapsing onto one data mode. FedLECC adds a clustering step first, so the high-loss clients chosen are also **diverse**.

Headline results (label skew, MNIST/FMNIST/CIFAR10/CIFAR100)

Up to **12% higher accuracy** than state-of-the-art baselines, **22% fewer communication rounds**, **50% less communication overhead** – using roughly 20% of the available clients.

Three stages: quantify → cluster → select

FedLECC in one picture



Balances diversity (clustering) with informativeness (loss) — explicitly.

Building blocks & per-round algorithm

Design choices

- **Hellinger distance on label histograms**
  Bounded, symmetric, well-suited for probability distributions. No raw data shared.
- **OPTICS clustering**
  Robust to varying densities; auto-determines $J_{\max}$ — outperformed DBSCAN & k-medoids.
- **Two-level loss prioritization**
  Rank clusters by mean loss → within each, pick highest-loss clients. Prevents mode collapse.

Algorithm (per round)

Inputs: clusters $C_1, \ldots, C_{J_{\max}}$, targets J, m

$z \leftarrow \lceil m/J \rceil$

1. Each client i computes local loss ℓ_i
2. Server averages loss per cluster: $\bar{\ell}_k$
3. Pick top-J clusters by $\bar{\ell}_k$
4. In each, pick top-z clients by ℓ_i
5. If $|S| < m$ → fill from next clusters

Output: set S of m selected clients

J clusters • m clients • $z = \lceil m/J \rceil$ per cluster

Justifying design choices

Quantify → cluster → select

Clustering choice. A grid search over OPTICS, DBSCAN, and k-medoids (maximizing silhouette score, requiring $\geq 70\%$ of clients assigned) found OPTICS on $P(y)$ (label distribution) to work best – clustering on $P(X \mid y)$ (feature-given-label) underperformed and sometimes failed the coverage constraint entirely.

Privacy note. Sharing label histograms can be vulnerable to inference attacks; DP noise on counts or secure aggregation could mitigate this with minimal effect on clustering quality – left to future work, as in Clust-PSI-PFL.

Why does this work? Theoretical motivation

Connecting loss-based selection to importance sampling

Gradient magnitude. Under standard L-smoothness, $\|\nabla \ell_i(\theta)\|$ is positively correlated with local loss $\ell_i(\theta)$. So prioritizing high-loss clients approximates importance sampling in SGD: samples with larger gradient norms get higher selection probability, reducing gradient-estimator variance.

Federated update.

$$\Delta\theta^{(t)} = \frac{1}{|S_t|} \sum_{i \in S_t} \nabla \ell_i(\theta^{(t)})$$

Biasing S_t toward high-loss (high-gradient-norm) clients can accelerate convergence for the parts of the data distribution the model still gets wrong.

Role of clustering. Loss-only selection can overfit to one data mode if the highest-loss clients all sit in the same region of feature space. Clustering *before* selection enforces diversity among the chosen clients – exploration (cover diverse regions) balanced with exploitation (focus on hard data).

FedLECC: accuracy vs. state-of-the-art

FMNIST, $K \in \{100, 300\}$ clients (subset of Table 3)

Method	$K = 100$	$K = 300$
FedAvg	0.565 ± 0.03	0.634 ± 0.03
HACCS	0.608 ± 0.03	0.658 ± 0.03
POC	0.605 ± 0.03	0.668 ± 0.02
FedLECC	$\mathbf{0.629 \pm 0.03}$	$\mathbf{0.675 \pm 0.02}$

Across all four datasets (MNIST, FMNIST, CIFAR10, CIFAR100) and both client counts: FedLECC is up to **12% higher than FedAvg** and **3% higher than HACCS** (its closest competitor). On CIFAR10 with K=250, POC edges it out by ∼4% – and the paper notes that the FedLECC advantage tracks the **Silhouette score**: better clusters (higher silhouette) ⇒ higher accuracy, worse clusters ⇒ FedLECC's edge shrinks (see Figure 4 in the paper).

On CIFAR100, no method beats plain FedAvg – a reminder that harder tasks under pathological non-IID remain an open problem.

FedLECC: convergence and communication cost

The other half of the story: getting there faster, cheaper

Convergence. On FMNIST with K=100, FedLECC's accuracy curve rises faster over communication rounds than FedAvg, FedProx, FedDyn, FedNova, HACCS, FedCLS, FedCor, and POC, reaching the highest final accuracy (0.629) after fine-tuning all methods.

Communication overhead (MB):

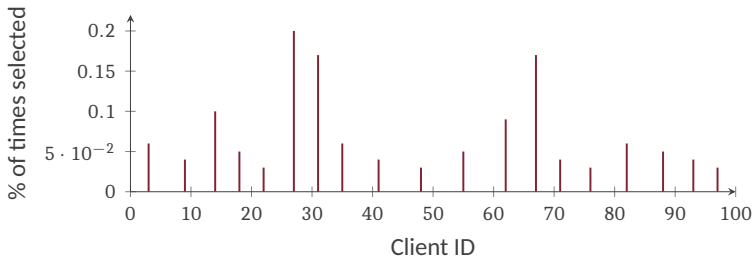
Method	FMNIST K=100	K=300
FedAvg	49.28	126.28
HACCS	3.42	55.29
POC	2.52	34.31
FedLECC	**2.24**	**33.55**

Overall, FedLECC's overhead is ∼**15**× **smaller than FedAvg** and ∼**1.5**× **better than HACCS**, its closest competitor – except on CIFAR10/CIFAR100 at K=250, where POC selects even fewer clients after its own hyperparameter tuning.

Why so much cheaper? Selecting only ∼20% of clients per round (next slide) means far fewer model uploads – the dominant cost in cross-device FL.

Valuable client selection

Who actually gets picked, over 150 rounds?



FedLECC, FMNIST, K=100, 150 rounds. Roughly 19 out of 100 clients account for essentially all selections – FedLECC achieves its performance and convergence gains using about **20% of the initial client pool**, cutting communication overhead, expensive-data costs, and privacy exposure for the remaining 80%. This is the practical face of the Personal Data Economy: the clients selected are precisely those whose data is high-loss and distinctive.

Putting a price on data: the PDE value function

From qualitative to quantitative client valuation

FedLECC's discussion section proposes a concrete valuation for client i:

$$V_i = \alpha \cdot \Delta Acc_i \; - \; \beta \cdot CommCost_i \; + \; \gamma \cdot RobustnessGain_i$$

- ΔAcc_i: marginal increase in global accuracy when client i participates (accuracy with vs. without).
- $CommCost_i$: normalized bandwidth/latency cost of involving client i.
- $RobustnessGain_i$: improvement under adversarial/high non-IID conditions attributable to client i's data.
- α, β, γ: weights reflecting the target application's priorities.

Practical use

Estimable via ablation studies or online evaluation; enables *differentiated incentives* – e.g. rewarding hospitals or IoT devices whose rare, high-loss data measurably improves the global model.

Synthesis: personalization vs. selection

Two answers, one non-IID problem

	Clust-PSI-PFL	FedLECC
Question answered	"Group clients how?"	"Train on whom, this round?"
Non-IID metric	PSI (Population Stability Index)	Hellinger distance
Clustering algorithm	K-means++ (silhouette-selected τ)	OPTICS on $P(y)$
Output	One model *per cluster*	One shared model, sparse updates
Optimizes for	Global accuracy & fairness (AD/SDAD)	Accuracy, rounds, & bandwidth
Headline gain	up to 18% accuracy, 37% fairer	up to 12% accuracy, 50% less comm.

They compose

Nothing prevents clustering *and* loss-based selection *and* per-cluster personalized models in the same pipeline – an open direction both papers flag for future work.

This afternoon's lab

Extending the running example: personalize & select

Environment. Same Flower + PyTorch running example (image classification) as Monday and Tuesday.

Steps.

1. Compute per-client label histograms on your non-IID partition; derive PSI_i^L and $WPSI^L$ (reuse Tuesday's code).
2. Cluster clients with K-means++ on the PSI feature vectors; select τ via the silhouette rule; train one FedAvg model per cluster.
3. Implement loss-based cluster-and-client selection (Algorithm 1): each round, compute per-client loss, rank clusters, sample the top high-loss clients.
4. Compare: global accuracy, fairness ($AD/SDAD$), and communication overhead of (a) plain FedAvg, (b) your clustered/personalized version, and (c) your loss-based selective version.

Deliverable

A short plot (accuracy & overhead vs. method) plus one paragraph: which of personalization or selection helped more on your partition, and why?

Table of Contents

Table of Contents

Recap: three take-home messages

2 Recap

1. **Personalization (Clust-PSI-PFL):** clustering clients by PSI signature and training one model per cluster turns non-IID data from a liability into a source of structure – up to 18% higher accuracy and 37% better fairness, with no penalty when data is already IID.
2. **Selection (FedLECC):** not all clients are equally valuable each round; clustering + loss-based cherry-picking finds the clients whose data most improves the global model, cutting communication by up to 50% while raising accuracy – the practical face of the Personal Data Economy.
3. **Both rely on the same building block:** a lightweight, privacy-conscious non-IID/similarity metric computed from label histograms (PSI or Hellinger distance) plus classic clustering (K-means++, OPTICS).

Tomorrow (Lecture 4)

Client fairness in depth: how do we measure whether an FL model serves *all* clients well, not just the average one – and what trade-offs exist between global accuracy and equitable outcomes?

Thank you

See you tomorrow !

ichatz@diag.uniroma1.it

References

2 Recap

- Jimenez-Gutierrez, D.M. et al. *Clust-PSI-PFL: A Population Stability Index Approach for Clustered Non-IID Personalized Federated Learning*. IPDPS 2026 (arXiv:2512.20363).
- Jimenez-Gutierrez, D.M., Giunta, G., Hassanzadeh, M., Anagnostopoulos, A., Chatzigiannakis, I., Vitaletti, A. *FedLECC: Cherry-Picking Clients in Federated Learning through Clustering and Personal Data Economy*. Journal of Industrial Information Integration (Elsevier), preprint.
- Tan, A.Z., Yu, H., Cui, L., Yang, Q. *Towards Personalized Federated Learning*. IEEE TNNLS, 2022.
- Sattler, F., Müller, K.-R., Samek, W. *Clustered Federated Learning*. IEEE TNNLS, 2020.
- Ruan, Y., Joe-Wong, C. *FedSoft: Soft Clustered Federated Learning with Proximal Local Updating*. AAAI 2022.
- Wolfrath, J. et al. *HACCS: Heterogeneity-Aware Clustered Client Selection for Accelerated Federated Learning*. IPDPS 2022.
- Cho, Y.J., Wang, J., Joshi, G. *Towards Understanding Biased Client Selection in Federated Learning*. AISTATS 2022.
- Li, C., Wu, H. *FedCLS: A Federated Learning Client Selection Algorithm Based on Cluster Label Information*. VTC2022-Fall.
- Tang, M. et al. *FedCor: Correlation-Based Active Client Selection Strategy for Heterogeneous Federated Learning*. CVPR 2022.
- Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V. *Federated Optimization in Heterogeneous Networks (FedProx)*. MLSys 2020.