Federated Learning

Lecture 2 – Non-IID Data
Escuela de Ciencias Informáticas (ECI) 2026

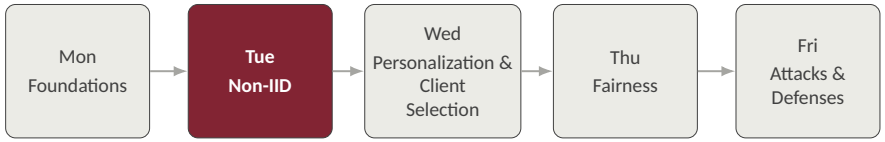
Universidad de Buenos Aires
Ioannis Chatzigiannakis
Sapienza University of Rome

ECI 2026 • Buenos Aires • Lecture 2 of 5

SAPIENZA
UNIVERSITÀ DI ROMA

This week schedule

Where today's lecture fits



Running thread. Same image-classification pipeline as yesterday. Today we stop pretending the partition is IID and start controlling *how* non-IID it is.

Today. What "non-IID" actually means, starting from the easiest case – quantity skew – to the more complex label skew and to how it is quantified and (partly) fixed.

Yesterday's assumption, made explicit

FedAvg was derived under IID data

Recall Lecture 1's formal setup:

$$\min_w F(w) = \sum_{k=1}^{K} \frac{n_k}{n} F_k(w), \qquad F_k(w) = \frac{1}{n_k} \sum_{i \in D_k} \ell(w; x_i, y_i)$$

We explicitly assumed that the observations of the clients are IID.

In reality

Clients are people, hospitals, phones, cars – **it is hard to accept IID-ness.** Every client's local distribution $P_k(x, y)$ can differ from every other's, and from the "global" distribution $P(x, y)$.

The easiest form of non-IID-ness: quantity skew

Same distribution, different amounts of it

Quantity skew. $P_k(x, y) = P(x, y)$ for every client k – the *shape* of each client's data is representative of the whole – but $n_k = |D_k|$ varies wildly across clients.

Concrete instances.

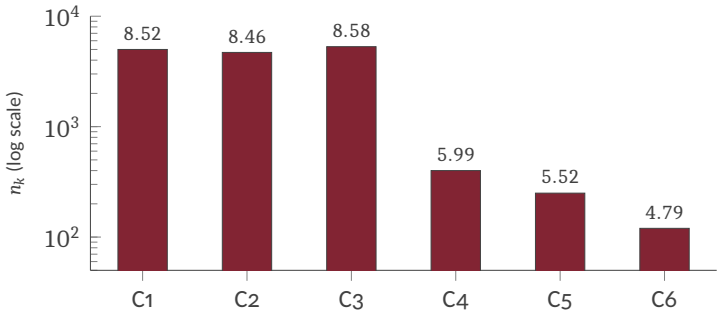
- A large university hospital vs. a rural clinic, both treating the same conditions in the same proportions – just at very different volumes.
- A power user of a keyboard app who types thousands of words a day vs. someone who opens the app twice.
- An IoT sensor with a long uptime vs. one that joined the network last week.

Why start here

It is the *only* non-IID case where $P_k(x, y)$ itself is not the problem – only n_k is. That makes it the cleanest place to build intuition before looking into other types of skew.

Simple example: K clients, wildly different n_k

Quantity skew: Same distribution, different amounts of it



Three "large" clients ($n_k \approx 5{,}000$) and three "small" ones (n_k from 120 to 400) – the same digits, the same label balance, just very different amounts of them.

Centralized vs Federated – What could go wrong ?

Think about the bar chart – what does FedAvg have to get right?

Centralized: all n samples get pooled and shuffled – C_6's 120 samples become 120 rows in one big table. What changes once each client's whole local update becomes one unit to average?

Centralized vs Federated – What could go wrong ?

Think about the bar chart – what does FedAvg have to get right?

Centralized: all n samples get pooled and shuffled – C_6's 120 samples become 120 rows in one big table. What changes once each client's whole local update becomes one unit to average?

1. **Overrepresentation.** An equal-vote average lets C_6 (120 samples) sway w_{t+1} *exactly as much as* C_3 (5,300 samples).

Centralized vs Federated – What could go wrong ?

Think about the bar chart – what does FedAvg have to get right?

Centralized: all n samples get pooled and shuffled – C_6's 120 samples become 120 rows in one big table. What changes once each client's whole local update becomes one unit to average?

1. **Overrepresentation.** An equal-vote average lets C_6 (120 samples) sway w_{t+1} *exactly as much as* C_3 (5,300 samples).

2. **Noisier votes from small clients.** A gradient from 120 samples has far higher variance than one from 5,000 – an equal vote trusts both the same.

Centralized vs Federated – What could go wrong ?

Think about the bar chart – what does FedAvg have to get right?

Centralized: all n samples get pooled and shuffled – C_6's 120 samples become 120 rows in one big table. What changes once each client's whole local update becomes one unit to average?

1. **Overrepresentation.** An equal-vote average lets C_6 (120 samples) sway w_{t+1} *exactly as much as* C_3 (5,300 samples).
2. **Noisier votes from small clients.** A gradient from 120 samples has far higher variance than one from 5,000 – an equal vote trusts both the same.
3. **Local overfitting.** With few unique samples and several local epochs E, a small client's model drifts toward memorizing D_k, not approximating $F(w)$.

Centralized vs Federated – What could go wrong ?

Think about the bar chart – what does FedAvg have to get right?

Centralized: all n samples get pooled and shuffled – C_6's 120 samples become 120 rows in one big table. What changes once each client's whole local update becomes one unit to average?

1. **Overrepresentation.** An equal-vote average lets C_6 (120 samples) sway w_{t+1} *exactly as much as* C_3 (5,300 samples).
2. **Noisier votes from small clients.** A gradient from 120 samples has far higher variance than one from 5,000 – an equal vote trusts both the same.
3. **Local overfitting.** With few unique samples and several local epochs E, a small client's model drifts toward memorizing D_k, not approximating $F(w)$.
4. **Sampling mismatch.** Selecting K clients uniformly is *not* the same as sampling n samples uniformly from the pooled data.

One root cause

All four are the same mistake: treating an update as if it carries the same statistical weight regardless of how much data produced it.

Naive mistake: unweighted averaging

What happens if the server just takes a simple mean

Suppose the server, unaware of n_k, aggregates with a plain arithmetic mean over the sampled clients S_t:

$$w_{t+1} \leftarrow \frac{1}{|S_t|} \sum_{k \in S_t} w_t^k$$

On our 6-client example ($n = 5000+4700+5300+400+250+120 = 15{,}770$):

Client	True share n_k/n	Unweighted vote
C_3 (n_k=5300)	33.6%	16.7%
C_6 (n_k=120)	0.8%	16.7%

C_6's influence is inflated more than 20×, while C_3's is cut by half. Small, noisy, potentially unrepresentative clients dominate the average – and every extra round of local epochs E makes their overfit local model even more different from w_t, so the damage compounds.

The fix: weight by contribution, not by headcount

McMahan et al., *AISTATS* 2017 – the formula you already know

Replace the plain mean with a weighted sum, weighting each client's update by how much data produced it:

$$w_{t+1} \leftarrow \sum_{k \in S_t} \frac{n_k}{\sum_{j \in S_t} n_j} w_t^k$$

This is not a new algorithm

This is exactly the FedAvg aggregation step from Lecture 1. Quantity skew does not require a new loss, a proximal term, or clustering – **it is already handled by the design of FedAvg itself.**

Intuition. A client with 5,300 samples moves w_{t+1} roughly $44\times$ more than one with 120 – exactly as those samples would count if physically pooled on one machine and averaged over a single centralized batch.

Does this actually work? The empirical evidence

What happens to accuracy when only n_k is skewed

- Li, Diao, Chen & He, *Federated Learning on Non-IID Data Silos: An Experimental Study*, IEEE ICDE 2022 – one of the most widely cited empirical benchmarks of non-IID FL, isolates quantity skew as one of several partition strategies compared under FedAvg-style weighted aggregation.
- Independent large-scale replications (e.g. Jiménez et al., using the FedArtML partitioning tool, quantified with Hellinger distance) report the same pattern across MNIST/FMNIST/CIFAR: once aggregation is weighted by n_k, quantity skew alone produces little to no accuracy degradation, even at the highest quantity-skew levels the finite dataset size allows.

Takeaway

Quantity skew is the one flavor of non-IID-ness that a single design choice – weighting by data volume – almost fully absorbs. Everything harder starts next.

Stepping back: two levels of non-IID-ness

Jiménez-Gutiérrez et al., 2026 survey

Centralized ML already has its own notion of non-IID data: class imbalance, temporal drift, or collection bias within a single dataset. It is handled with data augmentation, resampling, or careful train/test stratification.

FL adds a second, distinct level:

- **Intra-client non-IID.** Client k's own dataset D_k can internally be imbalanced or drifting – the classic centralized problem, just now living on one device.
- **Inter-client non-IID.** The distribution D_k is drawn from can differ from client to client – and from the pooled distribution. This is the FL-specific problem, the subject of the rest of today.

Why the distinction matters

Centralized fixes for intra-client non-IID assume you can see and touch the data. Privacy constraints in FL block exactly that for the inter-client level – you cannot pool $D_1, \ldots, D_K$ to resample or re-stratify. Everything in today's taxonomy is about this second, harder level.

Why non-IID data hurts: consequences for training

Jiménez-Gutiérrez et al., 2026 survey

You already met one of these

Model bias toward larger clients is exactly the failure mode from this morning's unweighted-averaging example. The other five persist even after aggregation is correctly weighted by n_k.

Why non-IID data hurts: consequences for training

Jiménez-Gutiérrez et al., 2026 survey

- **Model bias.** Global model skews toward clients with larger or more influential datasets.

You already met one of these

Model bias toward larger clients is exactly the failure mode from this morning's unweighted-averaging example. The other five persist even after aggregation is correctly weighted by n_k.

Why non-IID data hurts: consequences for training

Jiménez-Gutiérrez et al., 2026 survey

- **Model bias.** Global model skews toward clients with larger or more influential datasets.

You already met one of these

Model bias toward larger clients is exactly the failure mode from this morning's unweighted-averaging example. The other five persist even after aggregation is correctly weighted by n_k.

10/43 Ioannis Chatzigiannakis | Federated Learning | ECI 2026 | Lecture 2

Why non-IID data hurts: consequences for training

Jiménez-Gutiérrez et al., 2026 survey

- **Model bias.** Global model skews toward clients with larger or more influential datasets.
- **Slower convergence.** The global model struggles to reconcile local updates pulling in different directions.

You already met one of these

Model bias toward larger clients is exactly the failure mode from this morning's unweighted-averaging example. The other five persist even after aggregation is correctly weighted by n_k.

10/43 Ioannis Chatzigiannakis | Federated Learning | ECI 2026 | Lecture 2

Why non-IID data hurts: consequences for training

Jiménez-Gutiérrez et al., 2026 survey

- **Model bias.** Global model skews toward clients with larger or more influential datasets.
- **Slower convergence.** The global model struggles to reconcile local updates pulling in different directions.
- **Reduced accuracy.** Performance drops on the subsets of the federation that are underrepresented.

You already met one of these

Model bias toward larger clients is exactly the failure mode from this morning's unweighted-averaging example. The other five persist even after aggregation is correctly weighted by n_k.

10/43 Ioannis Chatzigiannakis | Federated Learning | ECI 2026 | Lecture 2

Why non-IID data hurts: consequences for training

Jiménez-Gutiérrez et al., 2026 survey

- **Model bias.** Global model skews toward clients with larger or more influential datasets.
- **Slower convergence.** The global model struggles to reconcile local updates pulling in different directions.
- **Reduced accuracy.** Performance drops on the subsets of the federation that are underrepresented.
- **Harder client selection.** Choosing a representative subset of clients each round becomes nontrivial.

You already met one of these

Model bias toward larger clients is exactly the failure mode from this morning's unweighted-averaging example. The other five persist even after aggregation is correctly weighted by n_k.

10/43 Ioannis Chatzigiannakis | Federated Learning | ECI 2026 | Lecture 2

Why non-IID data hurts: consequences for training

Jiménez-Gutiérrez et al., 2026 survey

- **Model bias.** Global model skews toward clients with larger or more influential datasets.
- **Slower convergence.** The global model struggles to reconcile local updates pulling in different directions.
- **Reduced accuracy.** Performance drops on the subsets of the federation that are underrepresented.
- **Harder client selection.** Choosing a representative subset of clients each round becomes nontrivial.
- **Communication inefficiency.** More rounds are needed for the same accuracy, raising network overhead.

You already met one of these

Model bias toward larger clients is exactly the failure mode from this morning's unweighted-averaging example. The other five persist even after aggregation is correctly weighted by n_k.

Non-IIDness vs Heterogeneity

Share similar issues, yet different

Non-IIDness: the umbrella term for any violation of the IID assumption across the federation.

Heterogeneity: broader still – differences in data, models, resources, or participation, not necessarily distributional.

Why non-IID data hurts: consequences for training

Jiménez-Gutiérrez et al., 2026 survey

- **Model bias.** Global model skews toward clients with larger or more influential datasets.
- **Slower convergence.** The global model struggles to reconcile local updates pulling in different directions.
- **Reduced accuracy.** Performance drops on the subsets of the federation that are underrepresented.
- **Harder client selection.** Choosing a representative subset of clients each round becomes nontrivial.
- **Communication inefficiency.** More rounds are needed for the same accuracy, raising network overhead.
- **Attack susceptibility.** A malicious update looks less like an outlier when honest updates already disagree.

You already met one of these

Model bias toward larger clients is exactly the failure mode from this morning's unweighted-averaging example. The other five persist even after aggregation is correctly weighted by n_k.

Formalizing data skew

Following the notation we introduced yesterday

We consider a set of K parties (clients)

Each party k holds a dataset D_k of n_k points, i.e., $n_k = |D_k|$.

Assume a "global" distribution $\mathbb{D}$ from which the ith client's is *observing* the local data D_i, with probability density $f^{(i)}$. Then

$$\textbf{Data skew} \iff f^{(i)} \neq f^{(j)} \text{ for some } i,j \in \{1,\dots,K\}.$$

At least one pair of clients has a genuinely different local data-generating distribution.

Coming up: when weighting is not enough

From "how much data" to "what kind of data"

Quantity skew only broke *one* number, n_k. The next case breaks the *distribution itself*.

Label skew: $P_k(y)$ differs across clients, even though $P(x \mid y)$ stays the same.

Why n_k-weighting stops being enough

Weighting corrects *how much* each client counts – it does nothing about *what* that client's update is actually pointing toward. When $F_k(w)$ itself has a different minimizer for different k, averaging the resulting w_t^k can drift the global model away from any of them.

Warning

Label skew dominates the FL literature — and genuinely **breaks** FedAvg, with losses beyond **20–35 percentage points**.

Label Skew, Formally

Marginal and conditional label skew

Client i observes D_i from the global distribution $\mathbb{D}$, with density $f^{(i)}$ over pairs $(\mathbf{x}, y)$.

Definition (label skew)

Marginal label skew: $f_Y^{(i)} \neq f_Y^{(j)}$ for some clients i, j. **Conditional**: $f_{Y|X}^{(i)} \neq f_{Y|X}^{(j)}$.

- For our running image-classification task, $f_Y^{(i)}$ is simply the **histogram of classes** held by client i.
- **Useful lemma.** With *no* attribute skew $\left(f_X^{(i)} = f_X^{(j)}, f_{X|Y}^{(i)} = f_{X|Y}^{(j)}\right)$, marginal and conditional label skew are **equivalent**, since $f_{Y|X}^{(i)} = f_{X|Y}^{(i)} f_Y^{(i)} / f_X^{(i)}$:

$$f_Y^{(i)} \neq f_Y^{(j)} \iff f_{Y|X}^{(i)} \neq f_{Y|X}^{(j)}.$$

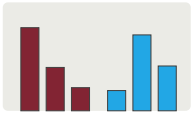
- So we may reason entirely about **class histograms**.

Definition and lemma follow the taxonomy of Jimenez-Gutierrez et al., *Non-IIDness in Federated Learning: A Survey*, ACM CSUR.

Four Faces of Label Skew

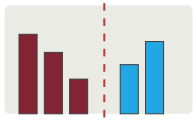
They are not equally hard — and they need different remedies

Distribution-based



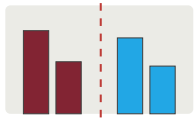
1. All classes present everywhere, with **different proportions**.

Overlapping classes



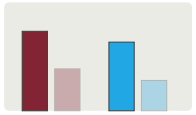
2. Clients **share some** classes and own others: $\{1, 2, 3\}$ vs. $\{2, 3, 4\}$.

Non-overlapping



3. **Disjoint** label sets — sharding, pathological partition. The worst case.

Noisy labels



4. Class **flipping**: symmetric or asymmetric annotation noise per client.

Design note

Cases 1–3 form a **severity ladder**. Our toy federation will sit between 2 and 3 — some shared classes, some entirely absent — because that is what real deployments look like.

Label Skew in the Wild

Four deployments, four mechanisms — none of them adversarial

Wildlife monitoring — *geography*

Snapshot Serengeti: camera traps at ∼50 sites in Tanzania, 13 species. A trap on the river plain records wildebeest by the thousand and **never** records a species that does not live there. Each camera is a client; each site has its own species prevalence.
Swanson et al., *Scientific Data* 2, 150026 (2015).

Crop disease — *agronomy*

Olive leaf disease across groves: which pathogen you photograph depends on your microclimate, cultivar and season. In our own FL study, performance held up until the **pathological** case — each grove holding only one or two **disjoint** diseases — where it collapsed.
Jimenez-Gutierrez et al., FL for olive leaf disease prediction.

Clinical ECG — *referral patterns*

PhysioNet/CinC 2020: 27 rhythm classes across hospitals. A district clinic sees sinus rhythm and atrial fibrillation; **rare arrhythmias** concentrate in tertiary referral centres. Specialisation *is* the skew.
Alday et al., *Physiol. Meas.* 41(12), 124003 (2020).

Mobile keyboards — *who you are*

Next-word prediction on 10^7 phones: each user's next-token distribution reflects their language, dialect, profession and social circle. The canonical cross-device deployment is **natively** label-skewed — there is no IID version of it to fall back on.
Hard et al., arXiv:1811.03604 (2018).

In every case the skew comes from **where the data lives** — it *is* the deployment.

A Federation You Can Hold in Your Head

$K = 5$ clients, $C = 5$ classes, and *exactly zero* quantity skew

	c_0	c_1	c_2	c_3	c_4	n_k	
Client 1	480	120	0	0	0	600	3 classes missing
Client 2	40	520	40	0	0	600	2 classes missing
Client 3	0	0	300	290	10	600	c_4 present but **starved**
Client 4	0	0	0	60	540	600	3 classes missing
Client 5	120	120	120	120	120	600	the one balanced client
Global	640	760	460	470	670	**3000**	

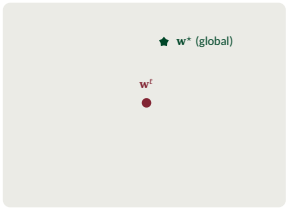
The point of the construction

Every $n_k = 600$, so **every FedAvg weight is** $n_k/n = 1/5$.
Whatever goes wrong next, it **cannot** be fixed by re-weighting — we have deliberately removed that degree of freedom.

- Each class is held by exactly **3 of 5** clients.
- Clients 2 and 4 have **disjoint** label sets.
- Client 3 holds **10** samples of c_4 — 1.7% of its own data, and 1.5% of the world's c_4.

Interactive: Where Exactly Does FedAvg Break?

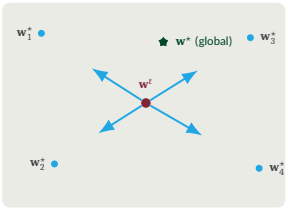
Let's expose one failure mode at a time



Local minimisers disagree — and the average of disagreeing points need not be a good point.

Interactive: Where Exactly Does FedAvg Break?

(1) Every client optimises a *different* objective

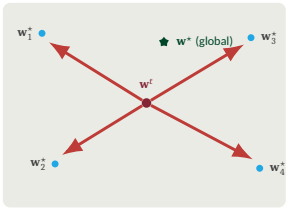


Local minimisers disagree — and the average of disagreeing points need not be a good point.

1. **Different objectives.** Client k minimises $F_k(\mathbf{w}) = \mathbb{E}_{P_k}[\ell]$, *not* F. Under label skew $P_k(y)$ differs, so $\mathbf{w}_k^\star \neq \mathbf{w}^\star$. Client 4 is solving a near-*binary* problem over $\{c_3, c_4\}$.

Interactive: Where Exactly Does FedAvg Break?

(2) Local work amplifies the disagreement

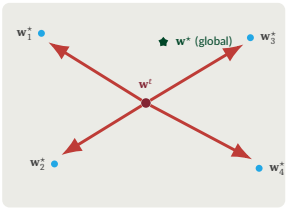


Local minimisers disagree — and the average of disagreeing points need not be a good point.

1. **Different objectives.** Client k minimises $F_k(\mathbf{w}) = \mathbb{E}_{P_k}[\ell]$, *not* F. Under label skew $P_k(y)$ differs, so $\mathbf{w}_k^\star \neq \mathbf{w}^\star$. Client 4 is solving a near-*binary* problem over $\{c_3, c_4\}$.
2. **Local work amplifies it.** Each client travels toward *its own* optimum. Larger E buys communication savings and pays in **drift**: $E = 1 \to 20$ raised mean $\|\mathbf{w}_k - \mathbf{w}^t\|$ from 1.24 to **1.98**.

Interactive: Where Exactly Does FedAvg Break?

(3) Missing labels push in one direction only

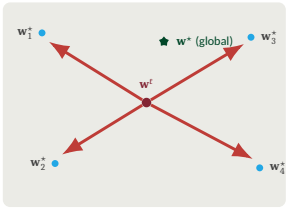


Local minimisers disagree — and the average of disagreeing points need not be a good point.

1. **Different objectives.** Client k minimises $F_k(\mathbf{w}) = \mathbb{E}_{P_k}[\ell]$, *not* F. Under label skew $P_k(y)$ differs, so $\mathbf{w}_k^\star \neq \mathbf{w}^\star$. Client 4 is solving a near-*binary* problem over $\{c_3, c_4\}$.
2. **Local work amplifies it.** Each client travels toward *its own* optimum. Larger E buys communication savings and pays in **drift**: $E = 1 \to 20$ raised mean $\|\mathbf{w}_k - \mathbf{w}^t\|$ from 1.24 to **1.98**.
3. **Missing labels are one-sided.** A client with **no** c_4 samples never sees a positive gradient for c_4 — only pressure to *suppress* it. That bias is **systematic**, so averaging does **not** cancel it.

Interactive: Where Exactly Does FedAvg Break?

(4) Rare labels are outvoted

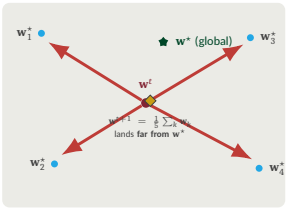


Local minimisers disagree — and the average of disagreeing points need not be a good point.

1. **Different objectives.** Client k minimises $F_k(\mathbf{w}) = \mathbb{E}_{P_k}[\ell]$, *not* F. Under label skew $P_k(y)$ differs, so $\mathbf{w}_k^\star \neq \mathbf{w}^\star$. Client 4 is solving a near-*binary* problem over $\{c_3, c_4\}$.
2. **Local work amplifies it.** Each client travels toward *its own* optimum. Larger E buys communication savings and pays in **drift**: $E = 1 \to 20$ raised mean $\|\mathbf{w}_k - \mathbf{w}^t\|$ from 1.24 to **1.98**.
3. **Missing labels are one-sided.** A client with **no** c_4 samples never sees a positive gradient for c_4 — only pressure to *suppress* it. That bias is **systematic**, so averaging does **not** cancel it.
4. **Rare labels are outvoted.** Client 3's **10** samples of c_4 give a weak, high-variance signal, averaged against four confident opposing opinions.

Interactive: Where Exactly Does FedAvg Break?

(5) And the weights cannot save us

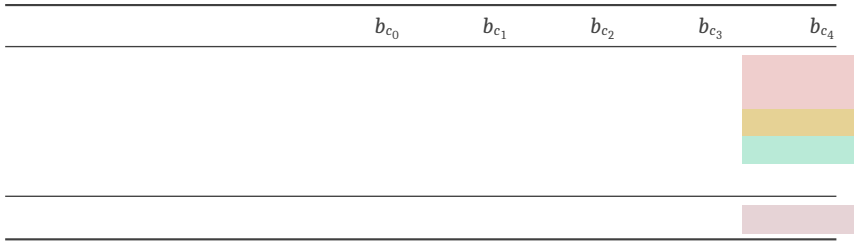


Local minimisers disagree — and the average of disagreeing points need not be a good point.

1. **Different objectives.** Client k minimises $F_k(\mathbf{w}) = \mathbb{E}_{P_k}[\ell]$, *not* F. Under label skew $P_k(y)$ differs, so $\mathbf{w}_k^\star \neq \mathbf{w}^\star$. Client 4 is solving a near-*binary* problem over $\{c_3, c_4\}$.
2. **Local work amplifies it.** Each client travels toward *its own* optimum. Larger E buys communication savings and pays in **drift**: $E = 1 \to 20$ raised mean $\|\mathbf{w}_k - \mathbf{w}^t\|$ from 1.24 to **1.98**.
3. **Missing labels are one-sided.** A client with **no** c_4 samples never sees a positive gradient for c_4 — only pressure to *suppress* it. That bias is **systematic**, so averaging does **not** cancel it.
4. **Rare labels are outvoted.** Client 3's **10** samples of c_4 give a weak, high-variance signal, averaged against four confident opposing opinions.
5. **Re-weighting is powerless.** All n_k equal $\Rightarrow$ all weights $1/5$. The quantity-skew fix has **no purchase**. The issue is not *how much* each client contributes, but *which direction* it pulls.

Interactive: Watch a Class Get Averaged Away

Output-layer bias b_c after 5 local epochs from a shared $\mathbf{w}^t$



Interactive: Watch a Class Get Averaged Away

Client 1 holds only c_0, c_1 — everything else is pushed down

	b_{c_0}	b_{c_1}	b_{c_2}	b_{c_3}	b_{c_4}
Client 1 (has c_0, c_1)	$+0.787$	$+0.453$	-0.413	-0.413	-0.413

Interactive: Watch a Class Get Averaged Away

Clients 2 and 3 add their own one-sided verdicts

	b_{c_0}	b_{c_1}	b_{c_2}	b_{c_3}	b_{c_4}
Client 1 (has c_0, c_1)	$+0.787$	$+0.453$	-0.413	-0.413	-0.413
Client 2 (has c_0, c_1, c_2)	$+0.140$	$+0.669$	$+0.099$	-0.454	-0.454
Client 3 (has c_2, c_3, **10**$\times c_4$)	-0.476	-0.476	$+0.734$	$+0.429$	-0.211

Interactive: Watch a Class Get Averaged Away

Client 4 is the world's expert on c_4: $b_4 = +1.152$

	b_{c_0}	b_{c_1}	b_{c_2}	b_{c_3}	b_{c_4}
Client 1 (has c_0, c_1)	$+0.787$	$+0.453$	-0.413	-0.413	-0.413
Client 2 (has c_0, c_1, c_2)	$+0.140$	$+0.669$	$+0.099$	-0.454	-0.454
Client 3 (has c_2, c_3, **10**$\times c_4$)	-0.476	-0.476	$+0.734$	$+0.429$	-0.211
Client 4 (has c_3, c_4 — **81%** of all c_4)	-0.423	-0.423	-0.423	$+0.119$	$\mathbf{+1.152}$

Note

Client 3 *does* hold c_4 — but 10 samples against 590 others is not enough to overcome its own class imbalance. Its bias for c_4 is still **negative**. **Having a label is not the same as having evidence for it.**

Interactive: Watch a Class Get Averaged Away

Client 5 is balanced, and almost silent

	b_{c_0}	b_{c_1}	b_{c_2}	b_{c_3}	b_{c_4}
Client 1 (has c_0, c_1)	$+0.787$	$+0.453$	-0.413	-0.413	-0.413
Client 2 (has c_0, c_1, c_2)	$+0.140$	$+0.669$	$+0.099$	-0.454	-0.454
Client 3 (has c_2, c_3, **10**$\times c_4$)	-0.476	-0.476	$+0.734$	$+0.429$	-0.211
Client 4 (has c_3, c_4 — **81%** of all c_4)	-0.423	-0.423	-0.423	$+0.119$	$\mathbf{+1.152}$
Client 5 (balanced)	-0.134	$+0.059$	$+0.061$	$+0.080$	-0.066

Interactive: Watch a Class Get Averaged Away

FedAvg averages them — and c_4's evidence vanishes

	b_{c_0}	b_{c_1}	b_{c_2}	b_{c_3}	b_{c_4}
Client 1 (has c_0, c_1)	$+0.787$	$+0.453$	-0.413	-0.413	-0.413
Client 2 (has c_0, c_1, c_2)	$+0.140$	$+0.669$	$+0.099$	-0.454	-0.454
Client 3 (has c_2, c_3, **10**×c_4)	-0.476	-0.476	$+0.734$	$+0.429$	-0.211
Client 4 (has c_3, c_4 — **81%** of all c_4)	-0.423	-0.423	-0.423	$+0.119$	$+\mathbf{1.152}$
Client 5 (balanced)	-0.134	$+0.059$	$+0.061$	$+0.080$	-0.066
FedAvg $\frac{1}{5}\sum_k b^{(k)}$	-0.021	$+0.056$	$+0.011$	-0.048	$+\mathbf{0.002}$

The washout

Four clients say "suppress c_4" ($-0.413, -0.454, -0.211, -0.066$); one client who owns 81% of the world's c_4 says "$+1.152$". The arithmetic mean is $+\mathbf{0.002}$ — indistinguishable from the model's initialisation. **The evidence was not lost at any client. It was destroyed by the averaging.**

Interactive: Watch a Class Get Averaged Away

FedAvg averages them — and c_4's evidence vanishes

	b_{c_0}	b_{c_1}	b_{c_2}	b_{c_3}	b_{c_4}
Client 1 (has c_0, c_1)	$+0.787$	$+0.453$	-0.413	-0.413	-0.413
Client 2 (has c_0, c_1, c_2)	$+0.140$	$+0.669$	$+0.099$	-0.454	-0.454
Client 3 (has c_2, c_3, **10**×c_4)	-0.476	-0.476	$+0.734$	$+0.429$	-0.211
Client 4 (has c_3, c_4 — **81%** of all c_4)	-0.423	-0.423	-0.423	$+0.119$	$+\mathbf{1.152}$
Client 5 (balanced)	-0.134	$+0.059$	$+0.061$	$+0.080$	-0.066
FedAvg $\frac{1}{5}\sum_k b^{(k)}$	-0.021	$+0.056$	$+0.011$	-0.048	$+\mathbf{0.002}$

Why this is the crux

This is not a bug in the implementation and not a bad random seed. It is what $\frac{1}{K}\sum_k$ *does* to a set of systematically opposed updates. Fixing it requires changing **what the clients optimise**, or **who participates**, or **how many models we keep** — the three families of the rest of this course.

FedProx

Li, Sahu, Zaheer, Sanjabi, Talwalkar, Smith — *MLSys* 2020

If the disease is drift, treat the drift: keep the client's objective, but **add a penalty for wandering**.

FedAvg: client k at round t

$$\min_{\mathbf{w}} \; F_k(\mathbf{w})$$

starting from $\mathbf{w}^t$, for E epochs — and nothing stops it from walking to $\mathbf{w}_k^\star$.

FedProx: client k at round t

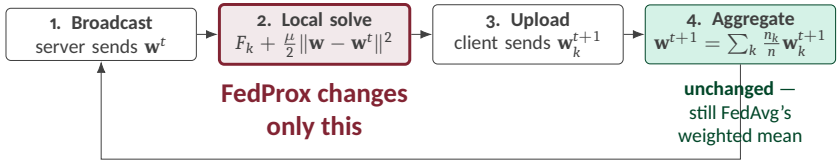
$$\min_{\mathbf{w}} \; h_k(\mathbf{w}; \mathbf{w}^t) = F_k(\mathbf{w}) \; + \; \frac{\mu}{2}\left\|\mathbf{w} - \mathbf{w}^t\right\|^2$$

the **proximal term** makes distance from the global model **expensive**.

- **Three readings of the same term.** A *trust region* around $\mathbf{w}^t$; a *Gaussian prior* centred on $\mathbf{w}^t$; an L_2 *pull toward consensus*.
- $\mu = 0$ **recovers FedAvg exactly** — FedProx is a strict generalisation, so it can only be *mistuned*, never worse in principle.
- $\mu \to \infty$ freezes every client at $\mathbf{w}^t$: no drift, no learning. μ buys **stability with plasticity**.
- **Cost: one line.** Add $\mu(\mathbf{w} - \mathbf{w}^t)$ to the local gradient. No extra communication, no server state.

Be Precise About *Where* FedProx Intervenes

A important distinction



Terminology

FedProx is often described as an "aggregation method", and it is listed that way in benchmark tables — including ours. Strictly, it is a **local-objective regulariser**: the server-side aggregation rule is *identical* to FedAvg's $\sum_k \frac{n_k}{n}\mathbf{w}_k$.

Bonus property: tolerating partial work. FedProx only requires each client to return a γ-**inexact** minimiser of h_k — so stragglers may run fewer local steps and still contribute, instead of being dropped. Under label skew, dropping a straggler may mean dropping the only holder of a class.

FedProx on Our Five Clients

What $\mu = 0.1$ actually does to the numbers we measured

	$\|\mathbf{w}_k - \mathbf{w}^t\|$, $E = 20$		
	$\mu = 0$	$\mu = 0.1$	change
Client 1	2.17	0.89	-59%
Client 2	1.81	0.69	-62%
Client 3	2.20	0.90	-59%
Client 4	**2.69**	**0.98**	$\mathbf{-64\%}$
Client 5	1.02	0.52	-49%
Mean	**1.98**	**0.79**	$\mathbf{-60\%}$

The correction is **largest where the skew is worst** (Client 4), smallest where it is mildest (Client 5): the penalty is *self-targeting*.

And the accuracy?

	$E = 1$	$E = 20$
FedAvg	86.4%	85.9%
FedProx (μ=0.1)	86.7%	**86.5%**

pooled centralised: 86.3%

Read this honestly – A Toy Example

Gaps under one point: this toy is **convex**, where FedAvg is provably well behaved. The **direction** is the signal — FedAvg *degrades* as E grows, FedProx *holds*. Collapse needs **non-convex** models.

Choosing μ, and Knowing When to Stop

Practical guidance and the honest boundary of this method

Tuning μ in practice

- Search $\mu \in \{0,\ 0.001,\ 0.01,\ 0.1,\ 1\}$ — the grid used in the original paper and in our benchmarks.
- **Higher skew → higher μ.** Near-IID federations often prefer $\mu \approx 0$.
- Diagnose with **drift**, not accuracy: log $\|\mathbf{w}_k - \mathbf{w}^t\|$. Flat ⇒ μ too large; growing across rounds ⇒ too small.
- μ is **global** — one knob for a heterogeneous population. Clients 4 and 5 differ 9× in HD, yet share it.

Known weaknesses

(i) Effectiveness **depends heavily** on this hyperparameter — a bad choice slows convergence or inflates communication. (ii) Advantages **weaken under extreme non-IIDness**, particularly when client datasets hold **completely distinct classes**.

Why a single global model must eventually lose

The structural argument

FedProx pulls every client toward **one $\mathbf{w}^t$**. But if $P_2(y)$ and $P_4(y)$ have **disjoint support**, no single parameter vector is simultaneously optimal for both. We are not *mis-optimising* — we are optimising the **wrong number of models**.

Where the course goes next

- **Correct the drift better** — SCAFFOLD control variates; FedNova update normalisation; server momentum.
- **Align representations** — MOON's model-contrastive loss.
- **Choose who participates** — POC, HACCS, FedCLS: if a class has one holder, *sample that holder*.
- **Give up on one model** — cluster distributionally similar clients and train per-cluster models (**Lecture 3**).

Quantifying non-IID-ness

Jiménez-Gutiérrez et al., 2026 survey, §4.2

Naming a skew type (quantity, label, …) is not the same as measuring how severe it is. A number lets you:

- Anticipate training difficulty *before* you spend a training run finding out.
- Design or select methods matched to the actual heterogeneity level of a task.
- Build realistic, reproducible testbeds instead of guessing at deployment parameters.
- Compare non-IID scenarios across papers on a common scale.

An underexplored corner of FL research

Despite this importance, quantifying non-IID-ness in FL has received comparatively little attention next to proposing new aggregation algorithms – most papers pick a partition protocol (e.g. Dirichlet α) without reporting *how* non-IID the result actually is.

Label skew metrics I: distribution-based

Comparing $f_Y^{(i)}$ across clients, or against the global f_Y

The standard toolkit of distances and divergences between distributions, applied to label proportions:

- **Earth Mover's Distance (EMD)** – a.k.a. Wasserstein distance.
- **Hellinger distance (HD)** – bounded, symmetric.
- **Jensen–Shannon divergence (JSD)** – symmetrized, smoothed KL.
- **Kullback–Leibler divergence (KLD)** – the classic asymmetric divergence.
- **Cosine and Jaccard similarity** – angle- or overlap-based alternatives.

Where each one breaks

EMD and HD get expensive in high-dimensional label spaces. JSD and KLD are sensitive to small sample sizes – exactly the regime of a client with few samples. Cosine similarity ignores magnitude entirely, which can mislead under severe class imbalance.

The formulas behind the toolkit (I)

Client label distribution $p = (p_1, \ldots, p_C)$ vs. reference $q = (q_1, \ldots, q_C)$, both summing to 1

Kullback–Leibler divergence (KLD).

$$D_{KL}(p\|q) = \sum_{c=1}^{C} p_c \log \frac{p_c}{q_c}$$

Asymmetric; undefined (or $-\infty$) when $q_c = 0$ but $p_c > 0$.

Jensen–Shannon divergence (JSD).

$$JSD(p, q) = \tfrac{1}{2} D_{KL}(p\|m) + \tfrac{1}{2} D_{KL}(q\|m), \qquad m = \tfrac{1}{2}(p + q)$$

Symmetric, always finite, bounded in $[0, \log 2]$.

Hellinger distance (HD).

$$HD(p, q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{c=1}^{C} \left(\sqrt{p_c} - \sqrt{q_c}\right)^2}$$

Symmetric, a true metric, bounded in $[0, 1]$.

Label skew metrics II: statistical tests & class-based

Two more ways to quantify label skew

Statistical-test based. Ask whether label distributions differ *significantly*, rather than by how much: the Kolmogorov–Smirnov (KS) test, the chi-squared test, or maximum mean discrepancy (comparing distributions via their mean embeddings).

Class-based. Use only the *number of classes present*, not their proportions.

Heterogeneity Index (HI) (Zawad et al., AAAI 2021):

$$HI = 1 - \frac{1}{C_{\max} - 1}(c - 1)$$

where c is the max. classes per client, $C_{\max}$ the total number of classes.

Imbalance Ratio (IR): $IR(\xi) = \max_i \xi_i / \min_j \xi_j$, the ratio between the most and least frequent class.

Limitations

HI collapses a whole distribution into a class *count*, losing nuance. IR is naturally suited to binary problems and oversimplifies multiclass imbalance.

The formulas behind the toolkit (II)

Same p, q – the ordinal, angle-based, and overlap-based alternatives

Earth Mover's / Wasserstein distance (EMD), ordinal labels.

$$EMD(p, q) = \sum_{c=1}^{C} \left|F_p(c) - F_q(c)\right|, \qquad F_p(c) = \sum_{i \le c} p_i$$

F_p, F_q are the cumulative distributions; requires an ordering over classes.

Cosine distance.

$$1 - \frac{p \cdot q}{\|p\|\,\|q\|} = 1 - \frac{\sum_c p_c q_c}{\sqrt{\sum_c p_c^2}\sqrt{\sum_c q_c^2}}$$

Angle only – blind to how large p and q are, not just their direction.

Jaccard distance (generalized to proportions).

$$1 - \frac{\sum_c \min(p_c, q_c)}{\sum_c \max(p_c, q_c)}$$

Reduces to the classic set-overlap Jaccard index when p, q are indicator vectors.

Quantity skew metrics

Measuring the bar chart from this morning

Quantity skew, too, has only the **distribution-based** subtype: HD, EMD, or JSD applied not to labels or attributes, but to the vector of client sample proportions $(n_1/n, \ldots, n_K/n)$ itself, compared against the uniform (perfectly balanced) distribution.

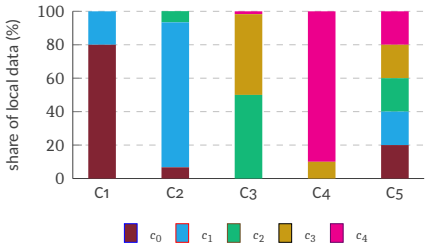
This is exactly our whiteboard example

The 6-client bar chart – $n = (5000, 4700, 5300, 400, 250, 120)$ – *is* a distribution over clients. Plugging those proportions into HD or JSD against the uniform vector $(\frac{1}{6}, \ldots, \frac{1}{6})$ gives a single number for "how skewed" that chart is.

As noted earlier: because quantity skew is empirical rather than distributional in the data-generating sense, a high quantity-skew score does not by itself predict a large accuracy drop – it depends on whether label/attribute skew is *also* present.

The Same Federation, Measured

From counts to distributions to a single number



	HD	JSD
(client marginal vs. global marginal)		
Client 1	0.601	0.632
Client 2	0.558	0.600
Client 3	0.622	0.665
Client 4	**0.653**	**0.687**
Client 5	0.070	0.084
Mean	**0.501**	**0.534**

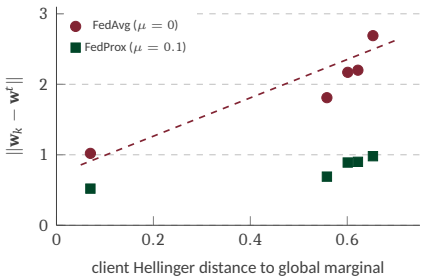
Max *pairwise* HD $= \mathbf{1.000}$ (Clients 2 vs. 4): completely **disjoint support** — the pathological case, inside an otherwise ordinary federation.

Read the number

Mean HD $= 0.501$ places this toy federation **exactly on the first critical threshold** we measured empirically: accuracy starts falling once HD > 0.5, and falls off a cliff past 0.75.

Drift Is Predictable from the Metric

$\|\mathbf{w}_k - \mathbf{w}^t\|$ after $E = 20$ local epochs, plotted against client HD



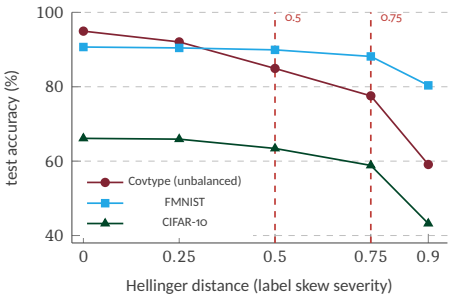
- Client 4 — the most skewed (HD $= 0.653$) — drifts **the furthest** (2.69). Client 5 — the balanced one (HD $= 0.070$) — barely moves (1.02).
- Across the five clients, $\mathrm{corr}(\mathrm{HD},\ \mathrm{drift}) = \mathbf{0.924}$.
- So the metric we introduced this morning is not bookkeeping — it **predicts the mechanism**.

Preview

The green squares are already the answer. Mean drift falls $1.98 \rightarrow \mathbf{0.79}$, a **60% reduction**. Next: what produced them.

At Scale: the Double Threshold

FedAvg accuracy vs. label-skew severity — $K = 30$ clients, 10 trials per point



Two cliffs, not a slope

Degradation is **not** linear in severity. A curvature analysis over five datasets identifies HD ≈ 0.5 and especially **HD $\approx$ 0.75** as **critical points** — the latter detected in **all five** datasets.

Accuracy lost from HD $= 0$ to HD $= 0.9$:

Covtype (unbalanced)	**35.85** pp
CIFAR-10	22.90 pp
PhysioNet	16.62 pp
FMNIST	10.31 pp
CIFAR-100	8.50 pp

Class **imbalance in the underlying dataset compounds** client label skew.

FedProx at Scale: Consistent, Modest, and Not a Cure

Test accuracy under increasing label skew, $K = 30$, mean over 10 trials

		Hellinger distance				
Dataset	**Method**	0	0.25	0.5	0.75	0.9
CIFAR-10	FedAvg	66.12	**65.91**	63.41	**58.85**	43.22
	FedProx	**66.35**	65.86	**63.56**	58.80	**44.33**
FMNIST	FedAvg	90.68	90.44	89.92	88.15	80.37
	FedProx	**90.69**	**90.51**	**89.96**	**88.17**	**81.10**
Covtype	FedAvg	94.95	92.05	84.92	77.55	59.10
	FedProx	**94.96**	**92.10**	**85.54**	77.55	**59.46**

What the numbers support

Over 25 label-skew configurations (5 datasets $\times$ 5 severities), FedProx was the best method **12 times** — against 4 for FedAvg. Gains **grow with severity**: largest at HD $= 0.9$. Convergence speed is **unchanged**.

What they do not support

CIFAR-10 still falls 66% $\rightarrow$ 44% *with* FedProx. Recovering $\approx$ 1 point of a 23-point loss is a **real but partial** fix. Reported gains **shrink further** when clients hold **nearly disjoint** classes — exactly our Clients 2 vs. 4.

Attribute skew

The same split, applied to x instead of y

Marginal attribute skew: $f_X^{(i)} \neq f_X^{(j)}$. **Conditional attribute skew:** $f_{X|Y}^{(i)} \neq f_{X|Y}^{(j)}$.

Four subtypes:

- **Distribution-based.** Same attributes, different distributions (blood pressure/glucose ranges differ across hospitals by demographics).
- **Partial attribute selection.** Clients observe an incomplete subset of attributes for their samples.
- **Noisy attributes.** Attribute values are corrupted or noisy for some clients.
- **Vertical skew.** Different clients hold different attribute subsets for the *same* samples – this is exactly vertical FL.

Well-established, especially in cross-silo FL, domain-shifted sensing, and medical/industrial data – two clients solve the same task but see systematically different inputs.

Participation skew

Heterogeneity in the process, not the data

Participation skew is not an intrinsic property of the local data distribution – it reflects *how often and under what conditions* clients take part in training or evaluation.

Two forms.

- **Party selection and subsampling.** Strategic or random selection of a subset of clients each round (e.g. 10 of 100 available).
- **Client dropout.** Clients unexpectedly disconnect mid-round because of network issues, battery, or other constraints.

A different semantic space

Unlike label, attribute, or modality skew, participation skew lives in the training *dynamics*, not the data-generating process. But it still shapes what the server ends up aggregating – it can induce or amplify the effective heterogeneity observed over time, especially in cross-device FL.

Modality skew

An emerging category, made explicit in this taxonomy

Clients hold data from different input modalities altogether – text, images, audio, tabular, graphs – rather than differing distributions of the *same* modality.

Not the same as vertical FL or transfer learning

Vertical FL assumes aligned sample identities (shared customer IDs); federated transfer learning needs only a small overlapping cohort. Modality skew involves *structurally disjoint* data spaces with no inherent alignment at all.

Two subtypes. *Complete multimodality skew*: all clients hold the same modalities, but their distribution differs. *Partial multimodality skew*: clients hold different *subsets* of modalities (some image+text, others image only).

Example. Hospital A supplies chest X-rays of pneumonia cases; Hospital B supplies CT scans of healthy individuals – incompatible inputs, requiring cross-modal architectures to bridge.

Terminology recap

Jiménez-Gutiérrez et al., 2026

Term	Meaning
Non-IIDness	Umbrella term for any violation of the IID assumption across the federation.
Heterogeneity	Broader concept: any relevant difference in data, models, resources, or participation.
Data skew	Differences in clients' local data-*generating* distributions ($f^{(i)} \neq f^{(j)}$).
Label / Attribute skew	Data skew in the label / feature distribution specifically.
Quantity skew	Heterogeneity in *how much* data each client holds – empirical, not distributional.
Modality skew	Clients hold data from structurally different input modalities.
Participation skew	Process-level heterogeneity from unequal client participation across rounds.

Table of Contents

1 Hands-on session

Table of Contents

2 Recap

This afternoon's hands-on session

From slides to code

Environment. Flower + PyTorch, single running example (image classification) that will be reused every day this week.

Steps.

1. ...

Deliverable

A short plot + one paragraph: at what ε does accuracy start to degrade noticeably for your partition?

Lecture 2 — What to Take Away

And what you will implement this afternoon

1. **Quantity skew** is handled by the n_k/n weights already inside FedAvg. It is the **easy** skew.
2. **Label skew** is different in kind: clients optimise **different objectives**, so their minimisers disagree and the average of disagreeing updates is not a good update.
3. **Missing and rare labels** create **one-sided, systematic** gradients. Averaging **cannot** cancel a systematic bias — we watched $+1.152$ become $+0.002$.
4. Severity is **measurable** (HD, JSD) and predicts drift ($r = 0.924$). Degradation is **non-linear**: cliffs at $\mathrm{HD} \approx 0.5$ and 0.75.
5. **FedProx** adds $\frac{\mu}{2}\|\mathbf{w} - \mathbf{w}^t\|^2$ locally, cutting drift by $\approx 60\%$ at **one line of code**. It is a **partial** fix, and it knows it.

Lab 2 — this afternoon

Rebuild the 5-client table with **FedArtML**; confirm mean $\mathrm{HD} = 0.501$.
Sweep Dirichlet α to walk HD from 0 to 0.9; locate *your* cliffs.
Implement FedProx in **Flower** — one added gradient term.
Sweep $\mu \in \{0, 0.001, 0.01, 0.1, 1\}$ and log drift, not just accuracy.
Add your row to the **cumulative table** we are building all week.

What comes next

2 Recap

Lecture 3 — Personalisation & Clustering. When one global model is provably the wrong object

Thank you

See you tomorrow !

ichatz@diag.uniroma1.it

References for This Segment

Where to read further

Algorithms

- H. B. McMahan et al. *Communication-Efficient Learning of Deep Networks from Decentralized Data.* AISTATS, 2017.
- T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, V. Smith. *Federated Optimization in Heterogeneous Networks.* MLSys, 2020. [FedProx]
- S. P. Karimireddy et al. *SCAFFOLD: Stochastic Controlled Averaging for Federated Learning.* ICML, 2020.
- J. Wang et al. *Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization.* NeurIPS, 2020. [FedNova]
- Q. Li, B. He, D. Song. *Model-Contrastive Federated Learning.* CVPR, 2021. [MOON]

Heterogeneity: taxonomy, metrics, measurement

- D. M. Jimenez-Gutierrez, M. Hassanzadeh, A. Anagnostopoulos, I. Chatzigiannakis, A. Vitaletti. *A Thorough Assessment of the Non-IID Data Impact in Federated Learning.*
- D. M. Jimenez-Gutierrez et al. *Non-IIDness in Federated Learning: A Survey with Taxonomy, Metrics, Methods and Frameworks.* ACM Computing Surveys.
- D. M. Jimenez-Gutierrez et al. *FedArtML: A Tool to Facilitate the Generation of Non-IID Datasets in a Controlled Way to Support Federated Learning Research.*
- Q. Li, Y. Diao, Q. Chen, B. He. *Federated Learning on Non-IID Data Silos: An Experimental Study.* ICDE, 2022.

The deployments

- A. Swanson et al. *Snapshot Serengeti: high-frequency annotated camera trap images.* Scientific Data 2:150026, 2015.
- E. A. Perez Alday et al. *Classification of 12-lead ECGs: the PhysioNet/CinC Challenge 2020.* Physiol. Meas. 41(12), 2020.
- A. Hard et al. *Federated Learning for Mobile Keyboard Prediction.* arXiv:1811.03604, 2018.