Federated Learning

Lecture 4 – Group Fairness
Escuela de Ciencias Informáticas (ECI) 2026
Universidad de Buenos Aires

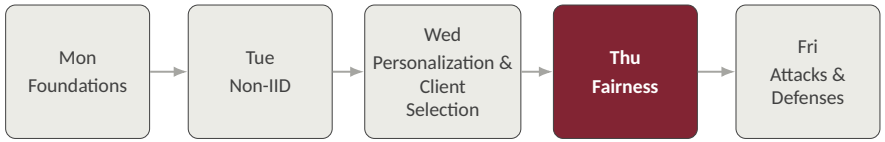
Ioannis Chatzigiannakis
Sapienza University of Rome

ECI 2026 • Buenos Aires • Lecture 4 of 5

SAPIENZA UNIVERSITÀ DI ROMA

This week schedule

Where today's lecture fits



Recall (Wed). Yesterday we asked: *does the global model work well for every **client**?*

Today. we ask a different question: *does the global model work well for every person?* Today's three claims:

1. Group fairness is *not* implied by client fairness, nor by accuracy.
2. Non-IID data does not merely hurt accuracy — it *redistributes* error across demographic groups.
3. You can buy substantial group fairness in FL at nearly zero accuracy cost, using only client selection.

Why this is not an academic exercise

The importance of group fairness

Machine learning is now the substrate of decisions about **people**: credit limits, shortlisting of job applicants, pretrial risk assessment, allocation of municipal services, triage priority.

The mechanism of harm

Historical human decisions become training labels. A model trained to *reproduce* those decisions reproduces their bias — then applies it at population scale, uniformly and cheaply.

The regulatory turn

The EU AI Act designates employment, creditworthiness, essential public services and law enforcement as *high-risk*, attaching obligations on data governance, bias examination and human oversight. *Fairness auditing is becoming a compliance artefact, not a research luxury.*

The federated twist

Exactly the domains where fairness matters most are those where data *cannot* be centralised — hospitals, banks, municipalities, phones. We need bias mitigation that works **without ever seeing the data**.

Fairness in FL is an overloaded word

Two families, answering two different questions

Individual-level (client-level)

"Is each participating client treated well?"

- **Performance distribution** — uniform accuracy across clients.
- **Selection** — under-represented clients get to participate.
- **Contribution** — payoff proportional to contribution.
- **Regret / expectation** — equitable waiting time for payoff.

Protects the *sustainability* of the federation.

Group-level

"Is each demographic group treated well?"

Good-intent fairness — reduce performance *variance* across groups; do not sacrifice any group.
Group fairness — minimise *disparity in decisions* across groups defined by a protected attribute A.

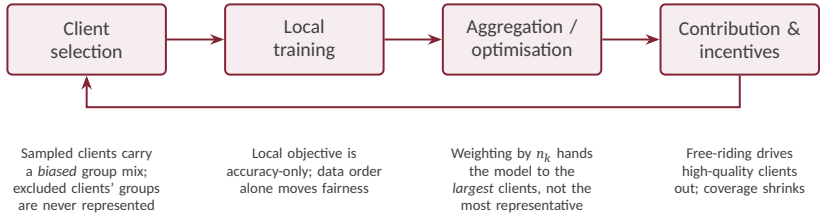
Protects the *people the model is applied to*.

The trap to avoid

A model satisfying an individual-fairness criterion need **not** satisfy any group-fairness criterion, or vice versa. Most FL fairness papers optimise the left column; today we optimise the right.

Where does unfairness enter an FL pipeline?

How can the FL architecture deal with group fairness



Two views by training stage

Performance fairness — arises in model optimisation.
Cooperation fairness — arises in selection and incentives.

Today's lever

We will attack **client selection** and **aggregation**. These are the two stages a server can change *without* touching client data.

An honest detour: we must change the dataset

Selecting a suitable dataset

TRhe image-classification dataset cannot help us today.

Why images fail here

Group fairness is defined *relative to a protected attribute* A attached to each record. A CIFAR-10 image has a label, but no gender, no ethnicity, no income bracket. There is no group to be unfair to.

Corollary for your own work: if your benchmark has no protected attribute, you cannot report group fairness on it — and no amount of accuracy reporting substitutes.

Today's dataset: ACS Income (California)

US Census American Community Survey, via `folktables` [Ding et al., NeurIPS 2021].

Records (CA, 2018)	195,665
Task	salary $\geq$ \$50K (binary)
Protected attribute A	gender
Privileged / unprivileged	male / female

Deliberately chosen: income prediction is exactly the kind of target that drives urban policy and resource allocation.

The bias is in the data before we start

The ACS Income dataset

ACS Income (CA, 2018) — positive means salary $\geq$ \$50K.

Ratio	Percentages	Proportion
Male / Female	52.8% / 47.2%	1 : 0.89
Positive / Negative	41.1% / 58.9%	1 : 1.44
Male pos. / neg.	46.6% / 53.4%	1 : 1.15
Female pos. / neg.	34.9% / 65.1%	1 : 1.86

Read the last two rows

Male records are close to balanced across the label. Female records are strongly skewed negative: *positively labelled female instances are a minority class within a minority group.*

A learner minimising average loss has a rational incentive to under-predict the positive class for $A = 0$. Nothing is broken; the objective is simply indifferent to who absorbs the error.

Keep this number in mind

Base rates differ across groups **in the population**. This is why some fairness criteria are mutually unsatisfiable, and why the choice of metric is a normative decision, not a technical one.

Setup and notation

Supervised binary classification

Supervised binary classification. Features X, true label Y, predictor $\hat{Y}$, and one designated coordinate of X is the protected attribute A.

- $A \in \{0, 1\}$: A=1 privileged, A=0 unprivileged.
- $Y \in \{0, 1\}$: Y=1 favourable (advantaged) outcome.
- A fairness measure $F \in [0, 1]$, where $F \to 0$ means *fairer*.

Terminology

These are properly **unfairness** measures: they quantify algorithmic *disparity*. "Improving fairness by 9%" always means "reducing the disparity by 9%".

	$\hat{Y} = 1$	$\hat{Y} = 0$
$Y = 1$	TP	FN
$Y = 0$	FP	TN

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \qquad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

Metrics are computed **per group** and then differenced. Subscript denotes the group: TPR_1, TPR_0.

Demographic Parity

Statistical parity

Criterion [Dwork et al., ITCS 2012] — the decision is independent of the protected attribute:

$$\Pr\{\hat{Y} = 1 \mid A = 0\} \;=\; \Pr\{\hat{Y} = 1 \mid A = 1\}$$

Measure — the absolute gap in positive-prediction rate:

$$\mathrm{DP} = \left|\Pr\{\hat{Y} = 1 \mid A = 0\} - \Pr\{\hat{Y} = 1 \mid A = 1\}\right|$$

Why people like it

Needs no ground truth at evaluation time. Directly interpretable to a regulator: "equal shares of each group receive the favourable decision."

Two real defects

1. It is satisfiable by a model that is blatantly unfair *within* groups — matching the *rates* says nothing about matching *who*. This arises precisely when one group has little training data.
2. When Y genuinely correlates with A in the population, enforcing $\mathrm{DP} = 0$ costs substantial utility.

Equal Opportunity Difference

Alternative measure of fairness

Criterion [Hardt, Price & Srebro, NeurIPS 2016] — equal true positive rates; the favourable outcome is equally *reachable* by qualified members of either group:

$$\Pr\{\hat{Y} = 1 \mid A = 0,\, Y = 1\} \;=\; \Pr\{\hat{Y} = 1 \mid A = 1,\, Y = 1\}$$

Measure —

$$\mathrm{EOD} \;=\; \Delta\mathrm{TPR} \;=\; \left|\mathrm{TPR}_1 - \mathrm{TPR}_0\right|$$

What it conditions on

Unlike DP, EOD conditions on $Y = 1$. It therefore tolerates unequal *selection rates* that are explained by unequal *base rates* — a much weaker and more often achievable requirement.

What it ignores

Errors on the unfavourable side. A model can equalise opportunity for qualified applicants while producing wildly different false-positive rates — which matters enormously when the "favourable" prediction triggers scrutiny rather than reward.

Average Odds — and why it is our primary target

The fairness metric used today

Criterion (equalised odds) [Hardt, Price & Srebro, NeurIPS 2016] — equality must hold on *both* sides of the label:

$$\Pr\{\hat{Y} = 1 \mid A = 0,\, Y = y\} = \Pr\{\hat{Y} = 1 \mid A = 1,\, Y = y\}, \qquad y \in \{0, 1\}$$

Measure — the mean absolute deviation of the two rates:

$$\mathrm{AO} \;=\; \tfrac{1}{2}\Big(\underbrace{\left|\mathrm{TPR}_1 - \mathrm{TPR}_0\right|}_{\Delta\mathrm{TPR}\,=\,\mathrm{EOD}} + \underbrace{\left|\mathrm{FPR}_1 - \mathrm{FPR}_0\right|}_{\Delta\mathrm{FPR}}\Big)$$

Why AO

It is the only one of the three that accounts for **both opportunities and errors** in a single scalar. Optimising it cannot buy TPR parity by silently wrecking FPR parity.

Relationships

$\mathrm{EOD} \le 2\,\mathrm{AO}$, and $\mathrm{AO} = 0 \Rightarrow \mathrm{EOD} = 0$. Neither implies $\mathrm{DP} = 0$ when base rates differ.

Practice: **optimise** one metric, **report** all three.

A simple example

Do the arithmetic once, by hand

A classifier evaluated on 2,000 held-out records, 1,000 per group, with base rates matching ACS Income (CA).

Group	TP	FN	FP	TN
$A = 1$ (male)	354	112	96	438
$A = 0$ (female)	180	169	46	605

	$A = 1$	$A = 0$
base rate	0.466	0.349
TPR	**0.760**	**0.516**
FPR	**0.180**	**0.071**
$\Pr\{\hat{Y} = 1\}$	**0.450**	**0.226**
accuracy	0.792	0.785

$\Delta\mathrm{TPR} = |0.760 - 0.516| = \mathbf{0.244} \;=\; \mathrm{EOD}$

$\Delta\mathrm{FPR} = |0.180 - 0.071| = \mathbf{0.109}$

$\mathrm{AO} = \tfrac{1}{2}(0.244 + 0.109) = \mathbf{0.177}$

$\mathrm{DP} = |0.450 - 0.226| = \mathbf{0.224}$

Note the accuracy column

Per-group accuracy differs by **0.7 pp**. A "good-intent fairness" audit would call this model fine. Its *decisions* are off by 24 points of TPR.

Aggregate accuracy hides group-level disparity almost perfectly.

The trap we must guard against

The importance of careful design

Take the *same* population. Replace the model with one that predicts $\hat{Y} = 0$ for everybody.

	$A = 1$	$A = 0$
TPR	0	0
FPR	0	0
$\Pr\{\hat{Y} = 1\}$	0	0

AO	**0.000**
EOD	**0.000**
DP	**0.000**
accuracy	0.593

Perfectly fair. Perfectly useless.

If $\hat{Y}$ is constant, TPR and FPR are constant across the whole dataset, hence across groups. So $\Delta\text{TPR} = \Delta\text{FPR} = 0$: all three metrics report zero disparity. Arithmetically correct, operationally worthless.

Consequence for reading FL results

Under *extremely* non-IID partitions, FL accuracy can collapse toward a near-constant predictor, and fairness metrics then improve for the wrong reason.

Hence we later **exclude extreme non-IID configurations from the fairness interpretation**: every fairness curve must be read against its accuracy curve.

Protected attribute skew

A special case of feature skew

Definition

Given $A \in \{0, 1\}$, a federation of K devices exhibits protected attribute skew when the marginals differ: $P_k(A) \neq P_j(A)$ for some $k \neq j$.

How it arises in the field

- Wearables: adoption differs sharply by gender and age.
- Regional sensors: a district's composition is not the city's.
- Clinical silos: a paediatric hospital and a geriatric ward hold disjoint age strata.

Why FedAvg amplifies it

$\omega_k = n_k/n$. A device holding many samples from a single group receives a large aggregation weight, so its group-specific decision boundary dominates the global model. *Data volume, not representativeness, buys influence.*

Contrast with label skew

$P_k(Y) \neq P_j(Y)$: it degrades accuracy and stability, and *indirectly* distorts fairness, because label-dominant devices skew which errors the model is willing to make.

Quantifying the skew

Selecting the appropriate distance metric

You cannot study "the effect of non-IID data" without a scalar that says *how* non-IID. Three candidates:

Measure	Strength	Failure mode
JSD	symmetric, bounded $[0, 1]$	saturates under strong heterogeneity — loses resolution exactly where you need it
EMD	accounts for ground distance between classes	costly; depends on a subjectively chosen ground metric
HD	symmetric, bounded, discriminative under severe skew	requires discrete distributions

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

$\text{HD} = 0$: identical distributions. $\text{HD} = 1$: disjoint support.

Practical note

HD between client distributions *decreases* as the number of clients grows, for a fixed Dirichlet α. Always report HD **and** K — an α value alone is not a reproducible statement of heterogeneity.

A valuable experiment

Dirichlet partitions

Partitioning by $\text{Dir}(\alpha)$ via **FedArtML**, $K = 100$ devices, target HD fixed in advance.

Protected attribute skew

Regime	α	HD
IID	1000	0.01
Weakly non-IID	3	0.21
Strongly non-IID	1	0.42
Extremely non-IID	0.25	0.54

Label skew

Regime	α	HD
IID	1000	0.01
Weakly non-IID	5	0.27
Strongly non-IID	0.5	0.48
Extremely non-IID	0.2	0.57

Design decision 1 — orthogonality

The two schemes are applied *mutually exclusively*. When protected-attribute HD reaches 0.54, label HD stays near 0.01–0.07, and vice versa. Without this you cannot attribute an observed fairness change to either cause — *the most common flaw in non-IID FL evaluations.*

Design decision 2 — well-definedness

Under severe skew a device can end up with no positive (or no negative) sample in one group. Then TPR or FPR is $0/0$ and the local fairness metric is **undefined**. Remedy: a published *training policy*, known to all participants, requiring ≥ 1 positive and ≥ 1 negative sample per group on every device.

An uncomfortable empirical fact about fairness

Ganesh, Chang, Strobel & Shokri, ACM FAccT 2023

Fairness measures vary *enormously* across training runs that differ only in random seed — far more than accuracy does. Decompose the randomness:

Source	Effect on fairness variance
Data splitting (train/test)	negligible
Weight initialisation	small — fixing it barely reduces variance
Random reshuffling (data order)	**dominant** — fixing it collapses the variance
Dropout, augmentation, grad. noise	secondary

Why data order?

Predictions on *under-represented* groups are the most volatile between consecutive epochs, and fairness metrics inherit that volatility. Fairness is dominated by the *most recent* gradient updates.

The exploitable consequence

If data *order* controls fairness, then **a deliberately chosen order** controls fairness — cheaply, locally, in a single pass, with minimal accuracy cost, and *without sharing anything*.

Centralised bias mitigation, and why it does not port

Three classical families, all assuming a single accessible dataset

Family	Representative methods	Assumption broken by FL
Pre-processing	reweighing [Kamiran & Calders, KAIS 2012]; procedurally fair feature selection [Grgić-Hlača et al., AAAI 2018]	needs global group statistics to compute weights
In-processing	FairBatch [Roh et al., ICLR 2021]; AdaFair [Iosifidis & Ntoutsi, CIKM 2019]; adversarial debiasing [Zhang, Lemoine & Mitchell, AIES 2018]	needs a global fairness signal each batch; tied to one model and one training algorithm
Post-processing	group-wise threshold adjustment [Lohia et al., ICASSP 2019]	needs a labelled, group-annotated calibration set at the server

The core tension

Bias mitigation wants *access to group statistics*; FL exists to *deny* it. Every method resolves that conflict differently, paying in a different currency: privacy, accuracy, or generality.

Approach A

purely local debiasing

Apply a centralised technique *independently on each device*. Aggregate with plain FedAvg. **Nothing beyond model updates leaves the device.**

Instances

- **Local reweighing** — debias local weights without relabelling [Kamiran & Calders, 2012].
- **Local FairBatch** — adapt minibatch composition per group, as a bilevel problem [Roh et al., 2021].
- **Local AdaFair** — boost examples that harmed fairness in the previous round [Iosifidis & Ntoutsi, 2019].
- **Fair representations** — geometrically remove feature correlation with A [He et al., 2020].
- **Fair data ordering** — fix the local sample order to a group-balanced permutation.

The appeal

Zero privacy cost, zero protocol change, drop-in compatible with any aggregator. The strongest possible baseline in terms of deployability.

The ceiling

Each device can only balance *what it holds*. If a device holds one group almost exclusively, local balancing is vacuous — and that is precisely the regime protected attribute skew creates.

Empirically: local debiasing yields consistent but **small** gains, around 1% across metrics. Real but insufficient.

Approach B

Server-coordinated group fairness

Better results, paid for with information disclosure. Read the last column as the price.

Method	Mechanism	Notion	What leaves the device
Global Reweighing [Abay et al., 2020]	server computes global sample weights	group fairness	noisy group counts (DP-protected)
FedFB [Zeng, Chen & Lee, 2021]	federated FairBatch; server sets group weights	group fairness	per-group loss statistics each round
Minimax group fairness [Papadaki et al., 2021]	zero-sum game on worst-off group	good-intent, group	per-group risk estimates
Unified group fairness [Zhang et al., 2021]	groups = clients × attributes	client + attribute + agnostic	group membership structure

Read the last column as the price. Every row buys fairness with a different disclosure — and the disclosure, not the mechanism, is usually what decides whether a method is deployable.

Approach B (cont.)

Server-coordinated group fairness

Method	Mechanism	Notion	What leaves the device
PrivFairFL [Pentyala et al., 2022]	MPC + DP pre/post-processing	group fairness	nothing in the clear (MPC)
FairFed [Ezzeldin et al., AAAI 2023]	aggregation weights track local-vs-global fairness gap	group fairness	local fairness statistics *before* training
GIFAIR-FL [Yue et al., INFORMS JDS 2023]	regulariser penalising loss spread across groups	group, individual	group-wise losses
FAIR-FATE [Salazar et al., 2023]	momentum blend of fair and performant update streams	group fairness	fairness-oriented update direction

The two closest competitors, and where each breaks

FairFed needs devices to publish fairness statistics about their local datasets *before* training — and loses effectiveness precisely when sensitive-attribute distributions vary widely across devices. **FAIR-FATE**'s momentum blending is justified only if the performance–fairness trade-off curve is convex, which does not reliably hold in federated settings.

For contrast: the client-fairness objective family

Fairness Federated Learning vs Group Fairness

These are the papers you find first when you search "fairness federated learning". Most are *not* about group fairness.

Method	Idea	Notion targeted
q-FFL [Li et al., ICLR 2020]	reweight aggregate loss by q, favouring high-loss clients	performance distribution
AFL [Mohri, Sivek & Suresh, ICML 2019]	minimax over worst weighted mixture of clients	good-intent
AgnosticFair [Du et al., SIAM SDM 2021]	agnostic fairness constraint + sample reweighting	good-intent **and** group
FCFL [Cui et al., NeurIPS 2021]	multi-objective, per-client fairness constraint	good-intent **and** group (local only)
Ditto [Li et al., ICML 2021]	personalisation: extra local SGD under a proximal constraint	performance distribution
FairFL [Zhang et al., IEEE BigData 2020]	minimise the global discrimination index	group

Also here: **TERM** and **FedMGDA+**, both targeting performance distribution.

Lecture 3 connection: personalisation can *unintentionally* improve **local** group fairness [Yang, Payani & Naghizadeh, 2023] — a local model need not trade one client's minority group against another's majority. Note "local": global disparity can remain untouched.

FairChoice — the gap it targets

Mocerino, Jimenez-Gutierrez & Chatzigiannakis, IEEE ISC2 2025

Three requirements

1. **No pre-training statistics exchange.** Devices must not have to publish their group composition before round 1.
2. **No structural assumption** on the shape of the performance–fairness trade-off curve.
3. **A tunable dial.** The operator, not the algorithm, decides how much accuracy to spend on fairness.

The design

Two components, composed:

Local: an order-based debiasing step, run once per round on each device — exploiting the data-order result from Ganesh et al.

Global: fairness- and loss-guided **device selection**, plus fairness-aware reweighting at aggregation — governed by a single fairness budget $\beta \in [0, 1]$.

Only three scalars per device per round travel to the server: L_k^t, F_k^t, n_k.

Component 1

Order-based local debiasing

Build a permutation that *interleaves* privileged and unprivileged samples, draining both pools from the tail, and place the leftovers at the head. Then freeze that order for the whole of training.

Algorithm 1 — Order-based local debiasing

Input: device dataset D_k *Output:* fairly ordered D_k^P

```
ids_0 ← GetAttrIndices(D_k, 0)
ids_1 ← GetAttrIndices(D_k, 1)
ids_e ← GetExtraIndices(D_k, ids_0, ids_1)
P ← empty list
while ids_0 and ids_1 nonempty do
  InsertAtHead(P, GetTail(ids_0))
  InsertAtHead(P, GetTail(ids_1))
  RemoveTail(ids_0)
  RemoveTail(ids_1)
InsertAtHead(P, ids_0 ∪ ids_1 ∪ ids_e)
D_k^P ← Permute(D_k, P)
```

Why the tail matters

Fairness is dominated by the *last* gradient updates. The balanced, interleaved portion is therefore placed at the **end** of the epoch; the unavoidable surplus of the majority group is pushed to the beginning, where it does least fairness damage.

Cost accounting

It reorders samples; it does not add, remove, relabel or reweight them. Accuracy impact on FedAvg is *minimal* — and so, on its own, is the fairness gain ($\approx 1\%$). It is the necessary but insufficient half.

Component 2

fairness-guided device selection

Algorithm 2 — Server side

```
for each round t = 1, ..., R do
  for each device k in parallel do
    Send(k, θ^{t-1})
    L_k^t, F_k^t, n_k ← Receive(k)
  ω_k = n_k / Σ_{i=1}^K n_i
  A^(t) ← ProbRandSample([1..K], d, {ω_i})
  F_glob^t = Σ_{k=1}^K ω_k F_k^t
  Δ_k^t = |F_k^t − F_glob^t|
  d_k^t = Δ_k^t − (1/K) Σ_k Δ_k^t
  v_i = β d_i^t + (1 − β)(1 − L_i^t),   i ∈ A^(t)
  S^(t) ← GetBestDevices(A^(t), σK, {v_i})
  for each s ∈ S^(t) in parallel do
    θ_s^t ← Receive(s)
    ω_s^t = n_s / Σ_{i∈S^(t)} n_i
    ω_s^t = ω_s^t · (1 − β d_s^t)
    ω̃_s^t = ω_s^t / Σ_{i∈S^(t)} ω_i^t
  θ^{t+1} ← Σ_{s∈S^(t)} ω̃_s^t θ_s^t
```

Reading the algorithm

6. Candidate set of size $d \gg \sigma K$, sampled with probability $p_k = n_k/n$ — *not* uniform. This is the Power-of-Choice construction [Cho, Wang & Joshi, 2022].

7–9. Global AO is the ω-weighted mean of local AOs. Each device gets a *centred* deviation d_k^t from it.

10. The score. Low is good: prefer devices *close* to global fairness and with *high* loss (informative).

15. Aggregation weights are shrunk in proportion to fairness deviation. Selection *and* weighting both bend toward fairness.

Component 3

Device side, and the meaning of β

Algorithm 3 — Device side

```
for each round t = 1, ..., R do
  on evaluation request:
    θ^{t-1} ← Receive(G)
    L_k^t, F_k^t ← LocalEval(k, θ^{t-1})
    Send(G, L_k^t, F_k^t, n_k)
  on training request:
    SetFairOrder(D_k)
    θ_k^t ← LocalTrain(k, θ^{t-1})
    Send(G, θ_k^t)
```

What is disclosed

A loss, an AO value, a count. *No* group proportions, *no* per-group confusion matrices, *no* pre-training survey.

The fairness budget $\beta \in [0, 1]$

$$v_i = \beta\, d_i^t + (1 - \beta)\big(1 - L_i^t\big)$$

β	Behaviour
0	pure loss-based selection; reduces to Power-of-Choice, aggregation weights untouched
$(0, 1)$	convex blend; the operator's exchange rate
1	selection *and* aggregation driven entirely by deviation from global AO

A pleasant empirical surprise

β has only a *minor* influence on accuracy and F1 — update quality stays stable across the range. Extensive β tuning is largely unnecessary: set it high and keep your accuracy.

Experimental setup

Reproducibility checklist

Federation

Devices K	100
Sampling rate σ	0.1
Rounds R	50
Candidate set d	20

Local training

Epochs / round E	30
Batch size	128
Learning rate	0.001

Model & data

MLP, one hidden layer of 64 units, ReLU; SGD with BCE loss (PyTorch).

80/20 train/test, fixed seed. MinMax scaling fitted on train only.

Deliberately small: the point is the aggregation and selection logic, not model capacity.

Stack & baselines

FedArtML — controlled HD-targeted partitioning [IEEE Access 2024]
FedLab — standalone FL simulation [Zeng et al., JMLR 2023]

Baselines: centralised reference (50 trials × 300 epochs), FedAvg, local debiasing alone, FairChoice at $\beta \in \{0, \ldots, 1\}$. Every FL point averaged over 5 trials.

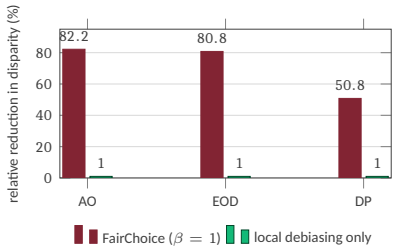
Why 50 centralised trials

Because of the Ganesh et al. result: a *single* centralised run gives you a fairness number with no error bar and no meaning. The centralised reference here is a *distribution*, drawn as a band, not a line.

Results

Protected attribute skew

Strongest configuration: $\beta = 1$, HD $= 0.54$ (extremely non-IID), vs. FedAvg.



Metric	Absolute	Relative
AO	−7%	−82.2%
EOD	−9%	−80.8%
DP	−7%	−50.8%
Accuracy	+0.5%	+1%
F1-score	+1%	+1%

The headline

Disparity cut by **four fifths**, with accuracy and F1 *slightly up*. There was no trade-off to make here at all.

(i) At $\beta = 0$ the gains are limited — the fairness term, not the loss-based selection, does the work.
(ii) Gains *grow* with skew intensity: the mechanism is most valuable exactly where the problem is worst.

Results
Label skew

Same algorithm, different partition axis. Interpretation restricted to IID and strongly non-IID ($\mathrm{HD} = 0.48$) — see the degenerate-predictor slide for why.

FairChoice $\beta = 1$ vs. FedAvg, $\mathrm{HD} = 0.48$

Metric	Absolute	Relative
AO	−5.62%	−38.8%
EOD	−9.30%	−43.87%
DP	−4.42%	−24.6%

Cost of label skew itself (IID FL vs. centralised)

Accuracy	−4.5%	−6%
F1-score	−5%	−6.5%

Why performance degrades here and not before

Devices with disproportionately large or unbalanced datasets produce biased updates that dominate aggregation. Add non-uniform candidate sampling and local overfitting, and training becomes genuinely unstable — independently of any fairness objective.

But the fairness mechanism still works

Gains *exceed* those under protected attribute skew at comparable HD. Local debiasing again contributes $\approx 1\%$; the selection rule contributes the rest, growing with β.

Takeaway: group fairness in FL is not only a protected-attribute problem. Label skew — demographically neutral by construction — degrades group fairness too, because it changes *which devices win the aggregation*.

Table of Contents

Limitations
Honest bookkeeping

1. Single-attribute fairness

Bias is assessed on one binary attribute (gender). Real fairness is **intersectional**: a model fair on each margin can be unfair on the intersection. *Open:* multi-attribute and sequential mechanisms [Hu, Ratz & Charpentier, AAAI 2024].

2. Single optimised metric

Selection is driven by AO alone; EOD and DP are reported, never optimised. A Pareto front over the three is unexplored.

3. No head-to-head comparison

Benchmarked against FedAvg and local debiasing, **not** FairFed or FAIR-FATE. Until then the claim is "substantial improvement over the standard baseline", not "state of the art".

4. Binary everything

$A, Y \in \{0, 1\}$. Multi-class outcomes and non-binary attributes require redefining every metric on this deck.

5. The unmeasured axis

Privacy. The protocol discloses less than FairFed, but "less" is not a guarantee. Composing with DP or secure aggregation and measuring the resulting privacy–fairness–accuracy triangle is open [Chen et al., ACM CSUR 2023].

This afternoon's lab — auditing your own federation
1 Hands-on session

Continue from Lecture 3's codebase. The image track is unchanged; the tabular track is new.

Tasks

1. Load ACS Income (CA) via `folktables`; verify the four base-rate ratios yourself.
2. Partition to *target* $\mathrm{HD} \in \{0.01, 0.21, 0.42, 0.54\}$ on the protected attribute with **FedArtML**, $K = 100$. Confirm label HD stays near zero.
3. Implement AO, EOD, DP from a per-group confusion matrix. Unit-test against the worked example.
4. Run FedAvg; record accuracy, F1 and all three metrics per round.
5. Implement Algorithm 1 (fair ordering) and re-run. Expect $\approx 1\%$.
6. Implement Algorithms 2–3; sweep $\beta \in \{0, 0.25, 0.5, 0.75, 1\}$.

Your cumulative table gains three columns

Setting	Acc	F1	AO
Centralised	…	…	…
FedAvg, IID	…	…	…
FedAvg, HD=0.54	…	…	…
+ fair order	…	…	…
FairChoice β=1	…	…	…

Deliverable question

Find a configuration where fairness metrics *improve* while accuracy *collapses*. Explain to your neighbour why that is not a success.

Table of Contents

Recap

2 Recap

1. **Name your fairness.** Client-level fairness (performance distribution, selection, contribution) and group-level fairness (good-intent, group) are different objectives with different beneficiaries. FL literature defaults to the former; regulation cares about the latter. Neither implies the other.
2. **Report three metrics, optimise one, always show accuracy.** DP, EOD and AO disagree by construction when base rates differ, and a constant predictor scores perfectly on all three. A fairness number without an accuracy number is not evidence.
3. **Non-IID data is a fairness problem, not only an accuracy problem.** Protected attribute skew hands the model to whichever device holds the most data; label skew does the same by a different route. Quantify with HD, keep the two skew axes orthogonal, publish K alongside α.
4. **Client selection is a surprisingly powerful fairness lever.** FairChoice cuts disparity by up to 82% under protected attribute skew and up to 44% under label skew, with no accuracy penalty, disclosing only a loss, a fairness scalar and a count per device per round.

Tomorrow

Lecture 5 — Attacks and defenses. We will see that several mechanisms from today, especially anything where the server trusts a self-reported scalar, are an **attack surface**.

References (1/2)

case study, foundations, tooling

This lecture's case study

- P. Mocerino, D. M. Jimenez-Gutierrez, I. Chatzigiannakis. *FairChoice: Balancing Performance and Group Fairness in Federated Learning with Non-IID Data*. IEEE Int. Smart Cities Conf. (ISC2), Patras, 2025. doi:10.1109/ISC266238.2025.11293299
- P. Mocerino. *Design and Evaluation of Novel Algorithms to Enhance Group Fairness in Federated Learning under Non-IID Data*. MSc thesis, Sapienza University of Rome, 2024.

Fairness foundations

- M. Hardt, E. Price, N. Srebro. *Equality of Opportunity in Supervised Learning*. NeurIPS 2016.
- C. Dwork et al. *Fairness Through Awareness*. ITCS 2012.
- P. Ganesh, H. Chang, M. Strobel, R. Shokri. *On the Impact of Machine Learning Randomness on Group Fairness*. ACM FAccT 2023.
- D. Pessach, E. Shmueli. *A Review on Fairness in Machine Learning*. ACM CSUR 55(3), 2022.

Centralised bias mitigation

- F. Kamiran, T. Calders. *Data Preprocessing Techniques for Classification Without Discrimination*. KAIS 33(1), 2012.
- Y. Roh et al. *FairBatch: Batch Selection for Model Fairness*. ICLR 2021.
- V. Iosifidis, E. Ntoutsi. *AdaFair: Cumulative Fairness Adaptive Boosting*. ACM CIKM 2019.
- B. H. Zhang, B. Lemoine, M. Mitchell. *Mitigating Unwanted Biases with Adversarial Learning*. AAAI/ACM AIES 2018.
- P. K. Lohia et al. *Bias Mitigation Post-Processing for Individual and Group Fairness*. IEEE ICASSP 2019.

Dataset, tooling and surveys

- F. Ding et al. *Retiring Adult: New Datasets for Fair Machine Learning*. NeurIPS 2021.
- D. M. Jimenez-Gutierrez et al. *FedArtML*. IEEE Access, 2024.
- D. Zeng et al. *FedLab*. JMLR 24(100), 2023.
- N. Mehrabi et al. *A Survey on Bias and Fairness in ML*. ACM CSUR 54(6), 2021.

References (2/2)

Fairness-aware federated learning

Group fairness in FL

- Y. H. Ezzeldin et al. *FairFed: Enabling Group Fairness in Federated Learning*. AAAI 2023.
- A. Abay et al. *Mitigating Bias in Federated Learning*. arXiv:2012.02447, 2020.
- Y. Zeng, H. Chen, K. Lee. *Improving Fairness via Federated Learning*. arXiv:2110.15545, 2021.
- A. Papadaki et al. *Federating for Learning Group Fair Models*. arXiv:2110.01999, 2021.
- S. Pentyala et al. *PrivFairFL: Privacy-Preserving Group Fairness in Federated Learning*. arXiv:2205.11584, 2022.
- X. Yue, M. Nouiehed, R. Al Kontar. *GIFAIR-FL*. INFORMS Journal on Data Science 2(1), 2023.
- T. Salazar et al. *FAIR-FATE: Fair Federated Learning with Momentum*. 2023.
- F. Zhang et al. *Unified Group Fairness on Federated Learning*. arXiv:2111.04986, 2021.
- D. Y. Zhang, Z. Kou, D. Wang. *FairFL*. IEEE BigData 2020.

Client-level fairness (for contrast)

- M. Mohri, G. Sivek, A. T. Suresh. *Agnostic Federated Learning*. ICML 2019.
- T. Li et al. *Fair Resource Allocation in Federated Learning*. ICLR 2020.
- T. Li et al. *Ditto: Fair and Robust Federated Learning Through Personalization*. ICML 2021.
- W. Du et al. *Fairness-Aware Agnostic Federated Learning*. SIAM SDM 2021.
- S. Cui et al. *Addressing Algorithmic Disparity and Performance Inconsistency in Federated Learning*. NeurIPS 2021.

Selection, privacy and trade-offs

- Y. J. Cho, J. Wang, G. Joshi. *Towards Understanding Biased Client Selection in Federated Learning*. AISTATS 2022.
- H. Chen et al. *Privacy and Fairness in Federated Learning: on the Perspective of Tradeoff*. ACM CSUR 56(2), 2023.
- F. Hu, P. Ratz, A. Charpentier. *A Sequentially Fair Mechanism for Multiple Sensitive Attributes*. AAAI 2024.

Thank you

See you tomorrow !

ichatz@diag.uniroma1.it