PLayer-FL: A Principled Approach to Personalized Layer-wise Cross-Silo Federated Learning

MLSys 2026 - Tuesday, May 19

Ahmed Elhussein[1], Florent Pollet[1], Gamze Gürsoy[2,1]

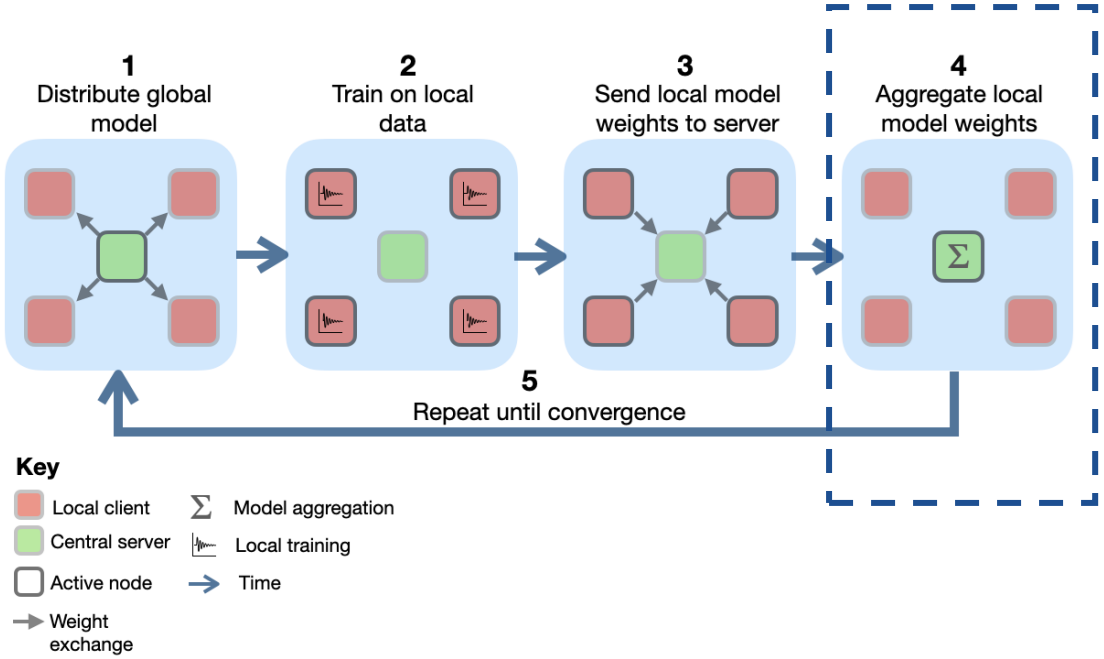
1 Department of Biomedical Informatics, Columbia University, NYC, USA
2 Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK

Paper

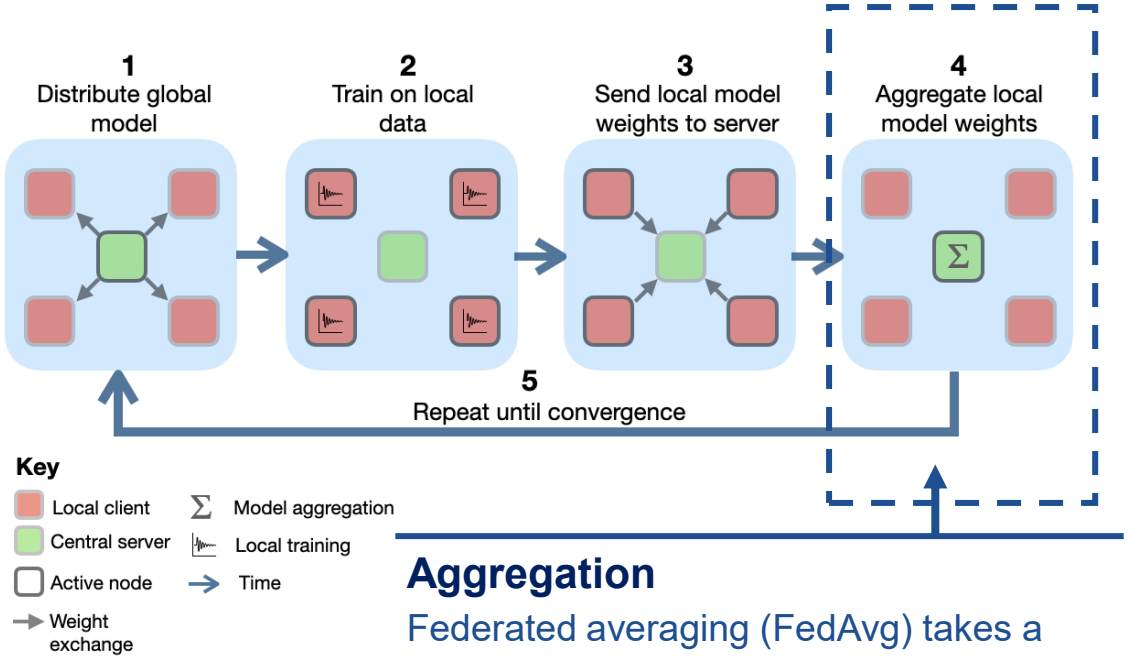
COLUMBIA | VAGELOS COLLEGE OF PHYSICIANS AND SURGEONS

Introduction
Federated Learning (FL)

Introduction
Federated Learning (FL)



Introduction
Federated Learning (FL)



Aggregation

Federated averaging (FedAvg) takes a weighted average of model weights

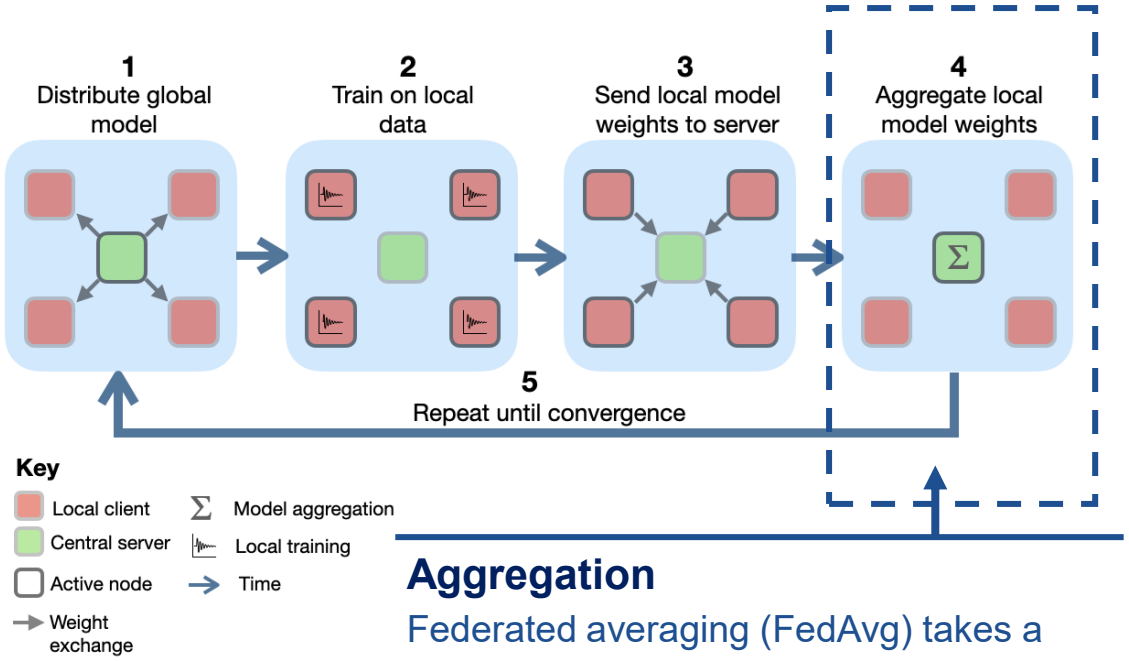
$$w_{t+1}^{g} \leftarrow \sum_{k=1}^{K} \alpha_k \cdot w_{t+1}^{k}$$

w_{t+1}^{k} = model parameters

α_k = weight of client site

FedAvg: McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, *54*, 1273–1282.

Introduction
Federated Learning (FL)



Aggregation

Federated averaging (FedAvg) takes a weighted average of model weights

$$w_{t+1}^{g} \leftarrow \sum_{k=1}^{K} \alpha_k \cdot w_{t+1}^{k}$$

w_{t+1}^{k} = model parameters

α_k = weight of client site

Two branches of FL:

- **Cross-device** (e.g., many mobiles)
- **Cross-silo** (e.g., few hospitals)

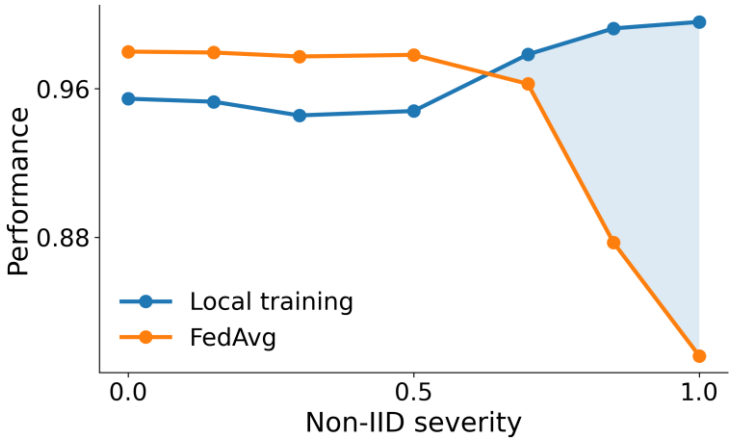
Most methods are **optimized for cross-device** scenarios

FedAvg: McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, *54*, 1273–1282.

Motivation
Federated Learning (FL) can fail in non-IID settings

Motivation

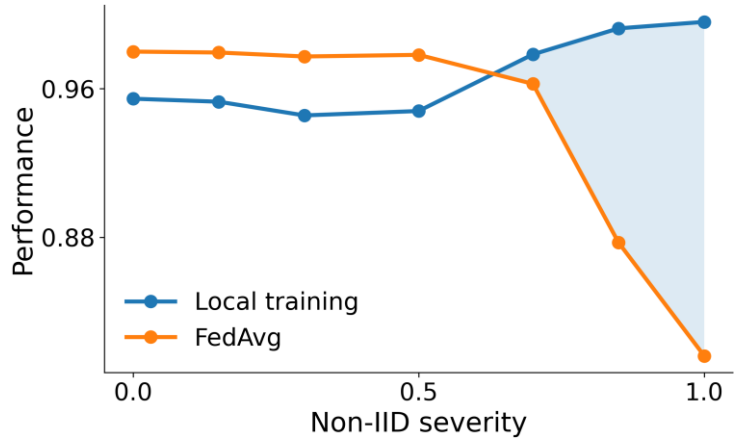
Federated Learning (FL) can fail in non-IID settings

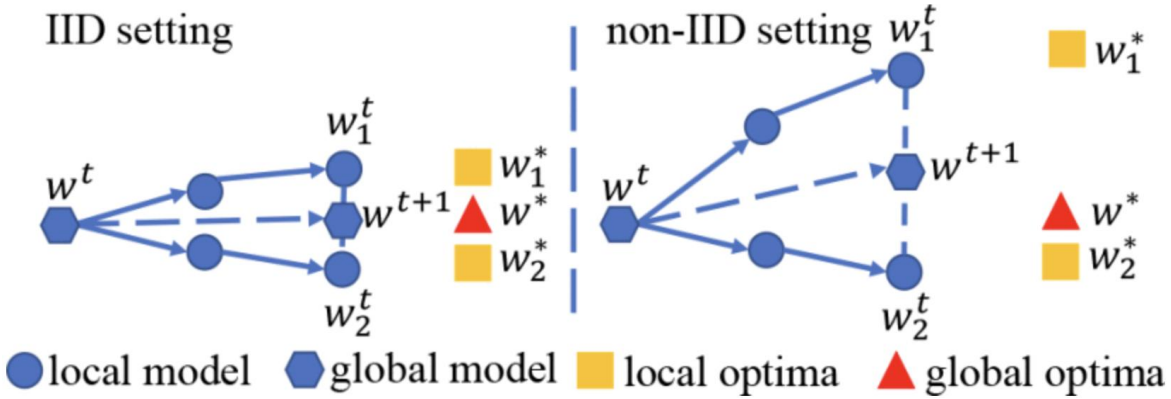


All ML models **struggle with non-independent and non-identically distributed** data (non-IID)

This is a particular problem in FL as no model is exposed to all the data

Motivation
Federated Learning (FL) can fail in non-IID settings





All ML models **struggle with non-independent and non-identically distributed** data (non-IID)

This is a particular problem in FL as no model is exposed to all the data

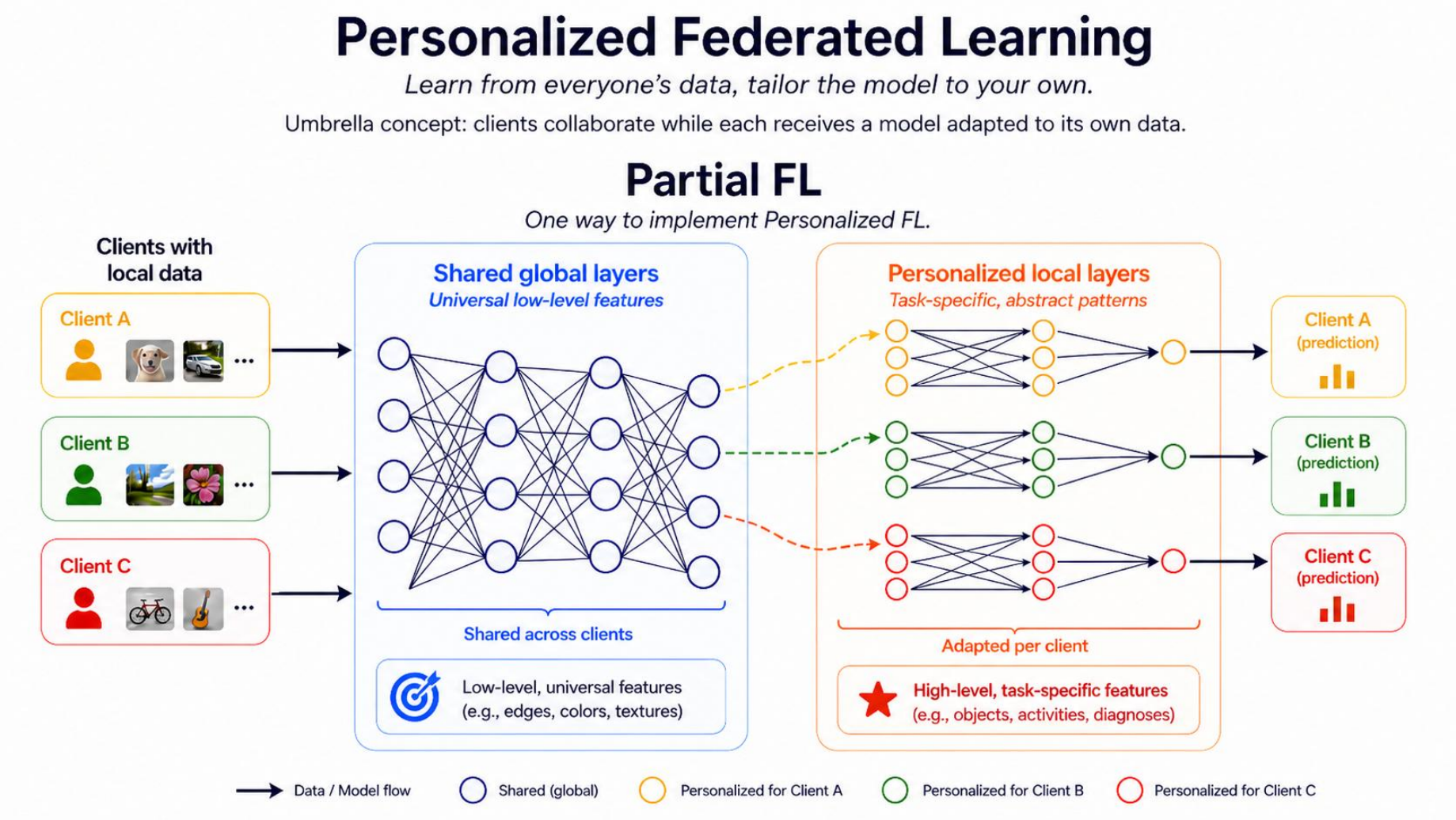
SGD is no longer an unbiased estimator and **weight divergence** occurs

Weight Divergence: Li, Q., Diao, Y., Chen, Q., & He, B. (2021). *Federated learning on non-IID data silos: An experimental study*. arXiv.

Background
Personalized and partial FL

Background
Personalized and partial FL

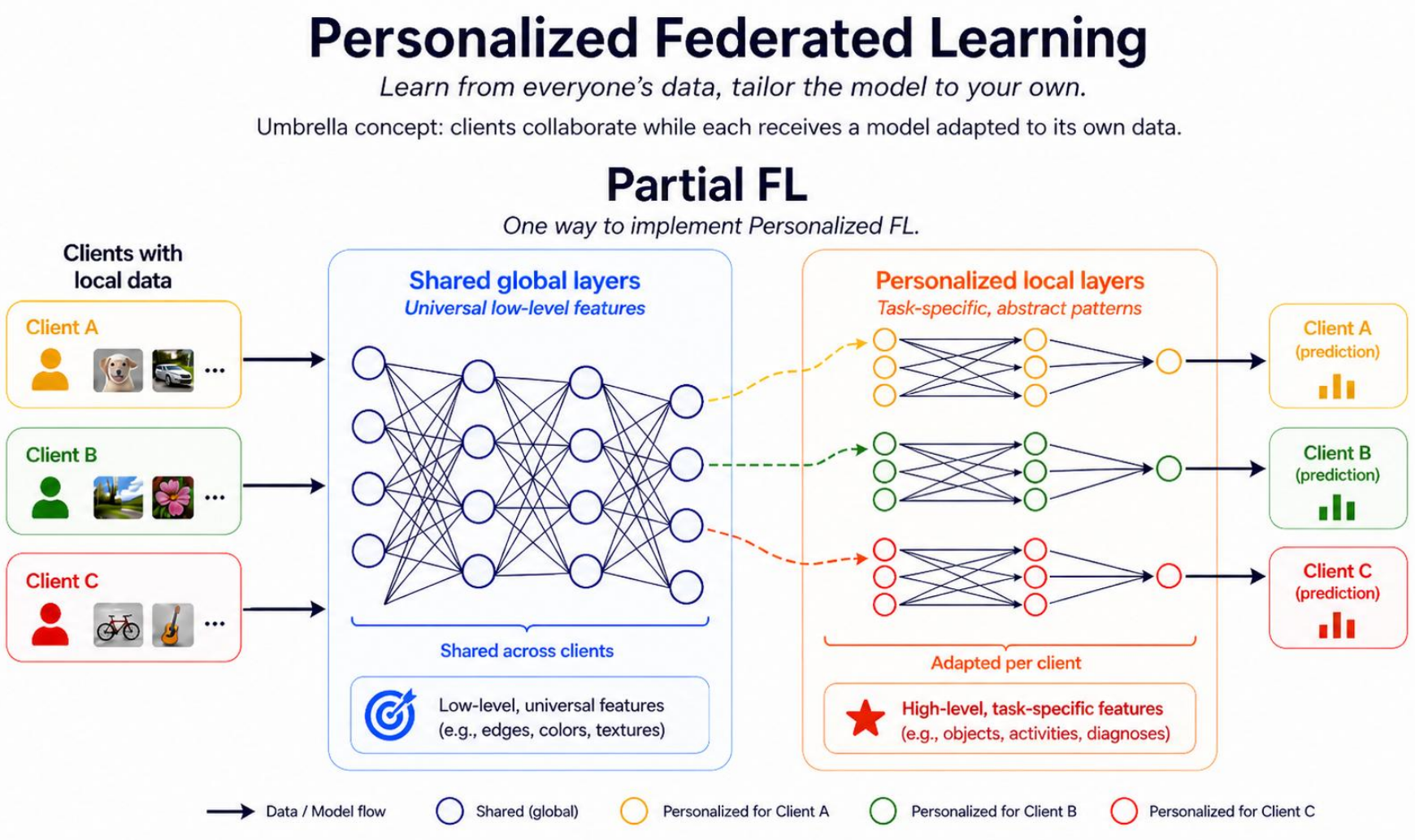
Background
Personalized and partial FL



Layered (partial) FL supports personalization

But many approaches often rely on **manually fixed shared-personal layer splits** that may not adapt well to heterogeneous client data

Background
Personalized and partial FL



Layered (partial) FL supports personalization

But many approaches often rely on **manually fixed shared-personal layer splits** that may not adapt well to heterogeneous client data

Personalized FL: Fallah, A., Mokhtari, A., & Ozdaglar, A. (2020). *Personalized Federated Learning: A Meta-Learning Approach.* NeurIPS 2020

Partial FL: Pillutla, K., Malik, K., Mohamed, A., Rabbat, M., Sanjabi, M., & Xiao, L. (2022). Federated learning with partial model personalization. *Proceedings of the 39th International Conference on Machine Learning, PMLR, 162*, 17716–17758

Methods
Principled Layer-wise FL (PLayer-FL)

Methods
Principled Layer-wise FL (PLayer-FL)

Goal: Can we automatically determine which layers to federate and which layers to do local training without doing the ML training itself?

Methods
Principled Layer-wise FL (PLayer-FL)

Goal: Can we automatically determine which layers to federate and which layers to do local training without doing the ML training itself?

Inspiration: model pruning

Methods
Principled Layer-wise FL (PLayer-FL)

Goal: Can we automatically determine which layers to federate and which layers to do local training without doing the ML training itself?

Inspiration: model pruning

Importance $I_S(W)$ of a parameter set W:

Intuition: $\approx$ gradient value $\times$ parameter value

$$\mathcal{I}_{\mathcal{S}}(W) \triangleq \sum_{s \in \mathcal{S}} \mathcal{I}_s(w) = \sum_{s \in \mathcal{S}} (g_s w_s)^2$$

with g_s the gradient with respect to the parameter w_s

Methods
A metric for deciding what to federate

Methods
A metric for deciding what to federate

Federation sensitivity of layer l, denoted $F_l(\Theta)$

Methods
A metric for deciding what to federate

Federation sensitivity of layer l, denoted $F_l(\Theta)$

- Θ denotes the full model parameters (**computed after just 1 epoch**)

Methods
A metric for deciding what to federate

Federation sensitivity of layer l, denoted $F_l(\Theta)$

- Θ denotes the full model parameters (**computed after just 1 epoch**)
- θ_p is the pth non-bias parameter in layer k

Methods
A metric for deciding what to federate

Federation sensitivity of layer l, denoted $F_l(\Theta)$

- Θ denotes the full model parameters (**computed after just 1 epoch**)
- θ_p is the pth non-bias parameter in layer k
- $\nabla_{\theta_p}\mathcal{L}$ is the gradient, and n'_k is the number of non-bias parameters in layer k

Methods
A metric for deciding what to federate

Federation sensitivity of layer l, denoted $F_l(\Theta)$

- Θ denotes the full model parameters (**computed after just 1 epoch**)
- θ_p is the pth non-bias parameter in layer k
- $\nabla_{\theta_p}\mathcal{L}$ is the gradient, and n'_k is the number of non-bias parameters in layer k

$$\mathcal{F}_l(\Theta) \triangleq \sum_{k=1}^{l} \frac{1}{n'_k} \sum_{p=1}^{n'_k} \mathcal{I}_p(\theta) = \sum_{k=1}^{l} \frac{1}{n'_k} \sum_{p=1}^{n'_k} (\theta_p \nabla_{\theta_p} \mathcal{L})^2$$

Methods
A metric for deciding what to federate

Federation sensitivity of layer l, denoted $F_l(\Theta)$

- Θ denotes the full model parameters (**computed after just 1 epoch**)
- θ_p is the pth non-bias parameter in layer k
- $\nabla_{\theta_p}\mathcal{L}$ is the gradient, and n'_k is the number of non-bias parameters in layer k

$$\mathcal{F}_l(\Theta) \triangleq \sum_{k=1}^{l} \frac{1}{n'_k} \sum_{p=1}^{n'_k} \mathcal{I}_p(\theta) = \sum_{k=1}^{l} \frac{1}{n'_k} \sum_{p=1}^{n'_k} (\theta_p \nabla_{\theta_p} \mathcal{L})^2$$

Cumulative aggregation

Methods
A metric for deciding what to federate

Federation sensitivity of layer l, denoted $F_l(\Theta)$

- Θ denotes the full model parameters (**computed after just 1 epoch**)
- θ_p is the pth non-bias parameter in layer k
- $\nabla_{\theta_p}\mathcal{L}$ is the gradient, and n'_k is the number of non-bias parameters in layer k

$$\mathcal{F}_l(\Theta) \triangleq \sum_{k=1}^{l} \frac{1}{n'_k} \sum_{p=1}^{n'_k} \mathcal{I}_p(\theta) = \sum_{k=1}^{l} \frac{1}{n'_k} \sum_{p=1}^{n'_k} (\theta_p \nabla_{\theta_p} \mathcal{L})^2$$

Importance of a parameter

Cumulative aggregation

Methods
A metric for deciding what to federate

Federation sensitivity of layer l, denoted $F_l(\Theta)$

- Θ denotes the full model parameters (**computed after just 1 epoch**)
- θ_p is the pth non-bias parameter in layer k
- $\nabla_{\theta_p}\mathcal{L}$ is the gradient, and n'_k is the number of non-bias parameters in layer k

$$\mathcal{F}_l(\Theta) \triangleq \sum_{k=1}^{l} \frac{1}{n'_k} \sum_{p=1}^{n'_k} \mathcal{I}_p(\theta) = \sum_{k=1}^{l} \frac{1}{n'_k} \sum_{p=1}^{n'_k} (\theta_p \nabla_{\theta_p} \mathcal{L})^2$$

Importance of a parameter

Cumulative aggregation

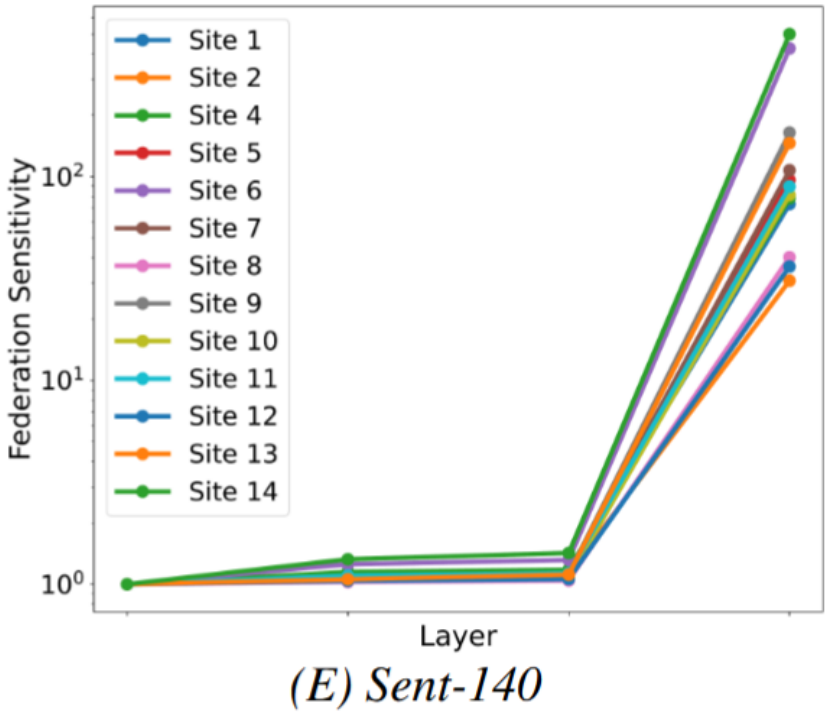
Layer-wise normalization

Methods

Federation sensitivity reveals a layer cutoff

Methods

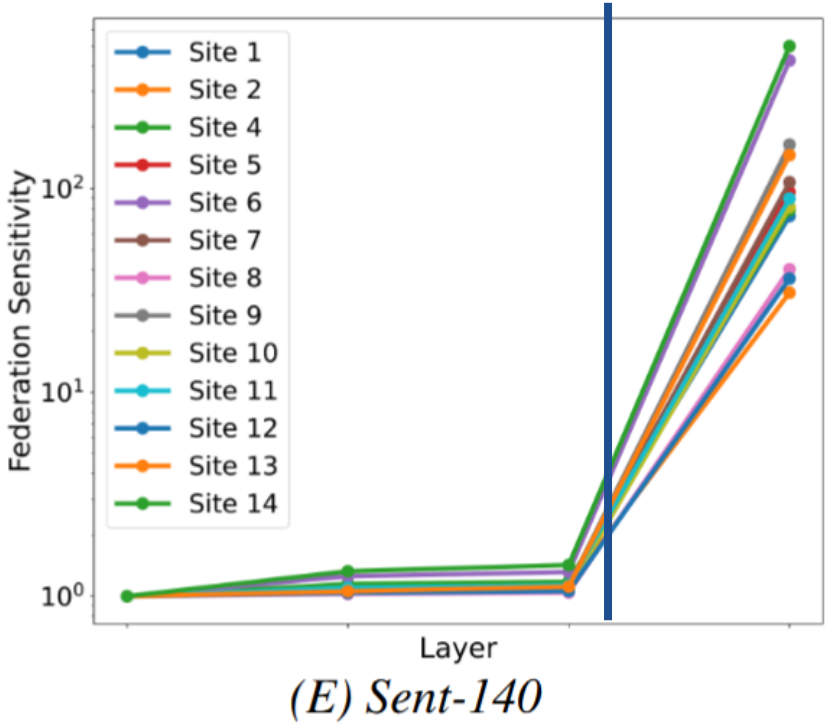
Federation sensitivity reveals a layer cutoff



(E) Sent-140

Methods

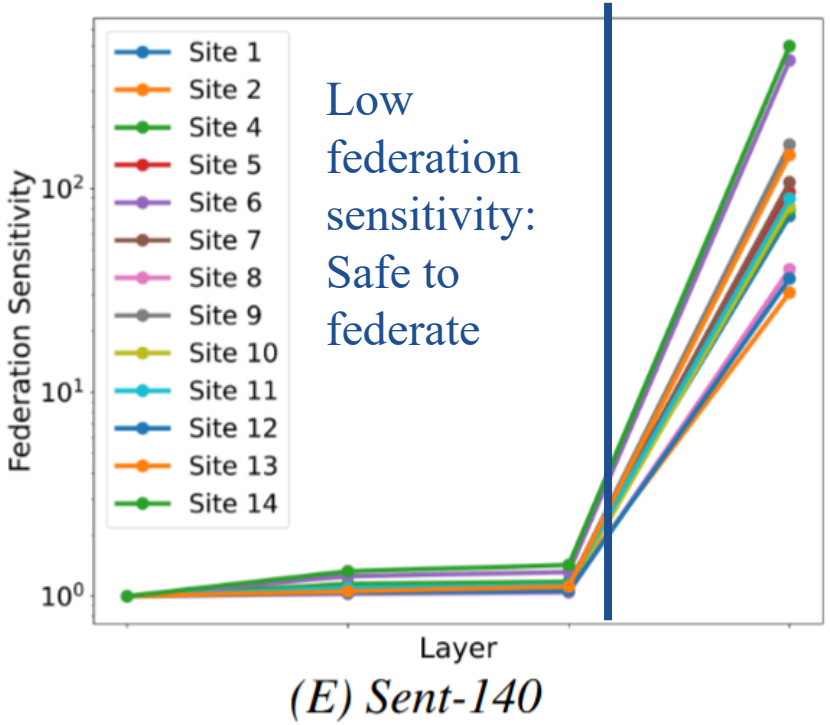
Federation sensitivity reveals a layer cutoff



(E) Sent-140

Methods

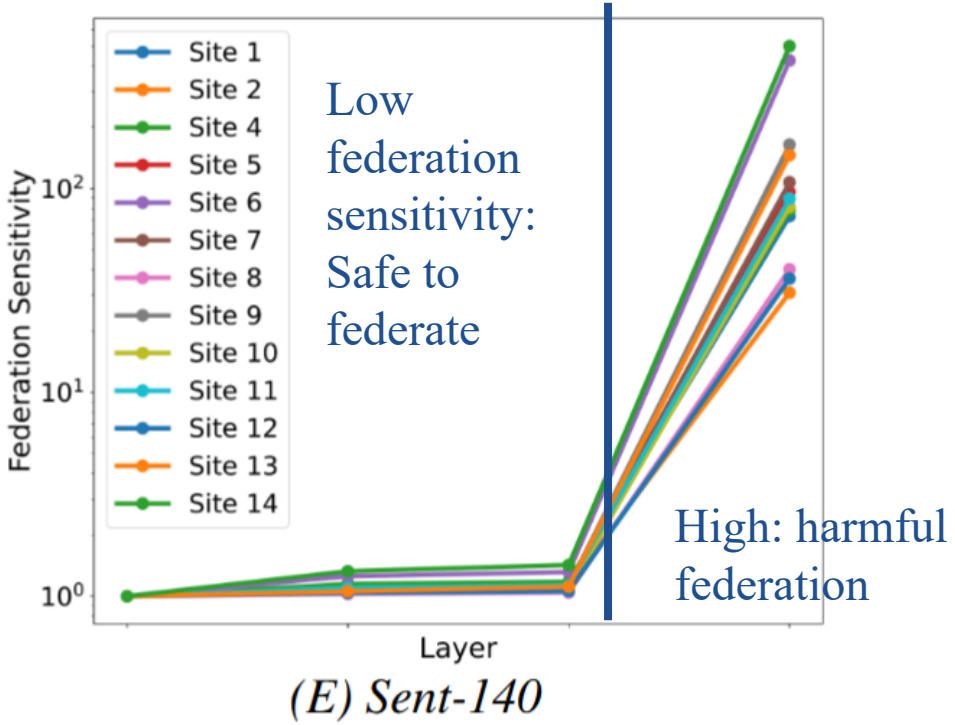
Federation sensitivity reveals a layer cutoff



(E) Sent-140

Methods

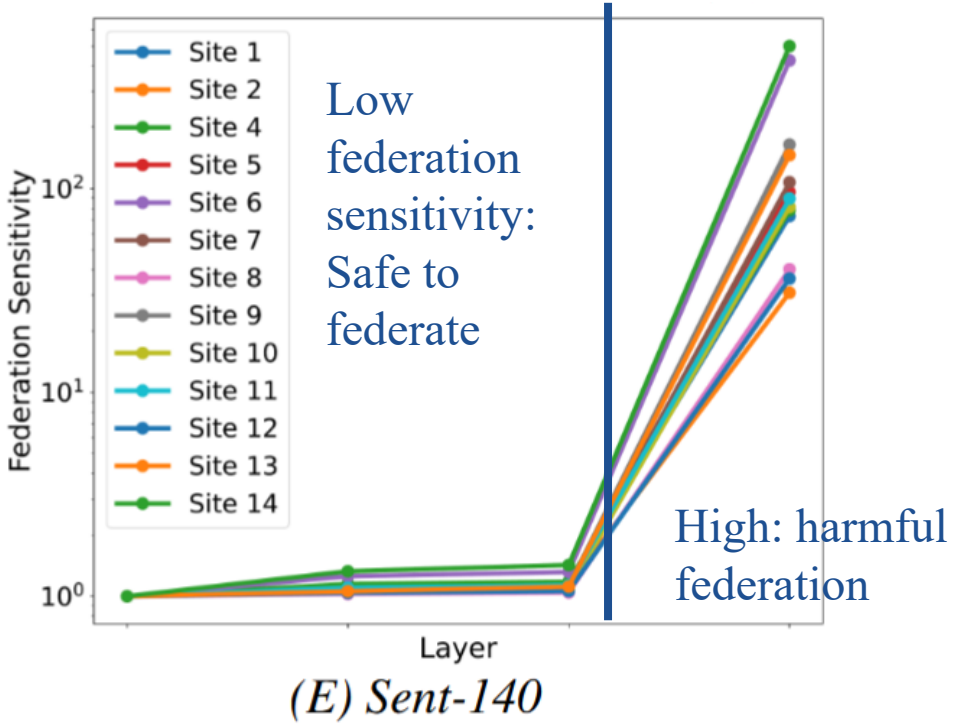
Federation sensitivity reveals a layer cutoff



(E) Sent-140

Methods

Federation sensitivity reveals a layer cutoff

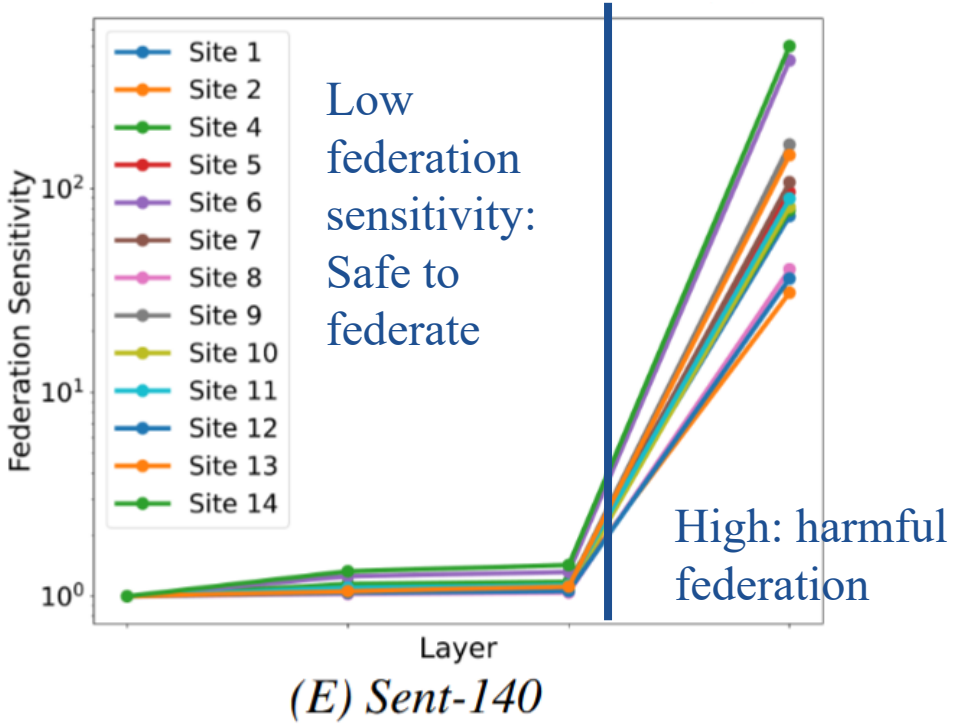


(E) Sent-140

We observe a **sharp transition** in federation sensitivity across layers, enabling principled selection of the layers to federate: those before the transition point

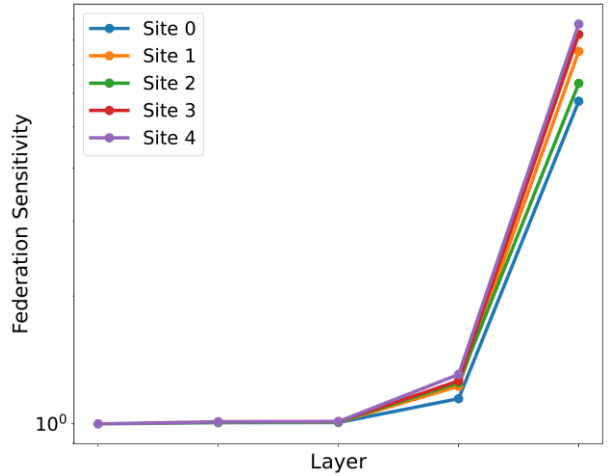
Methods

Federation sensitivity reveals a layer cutoff



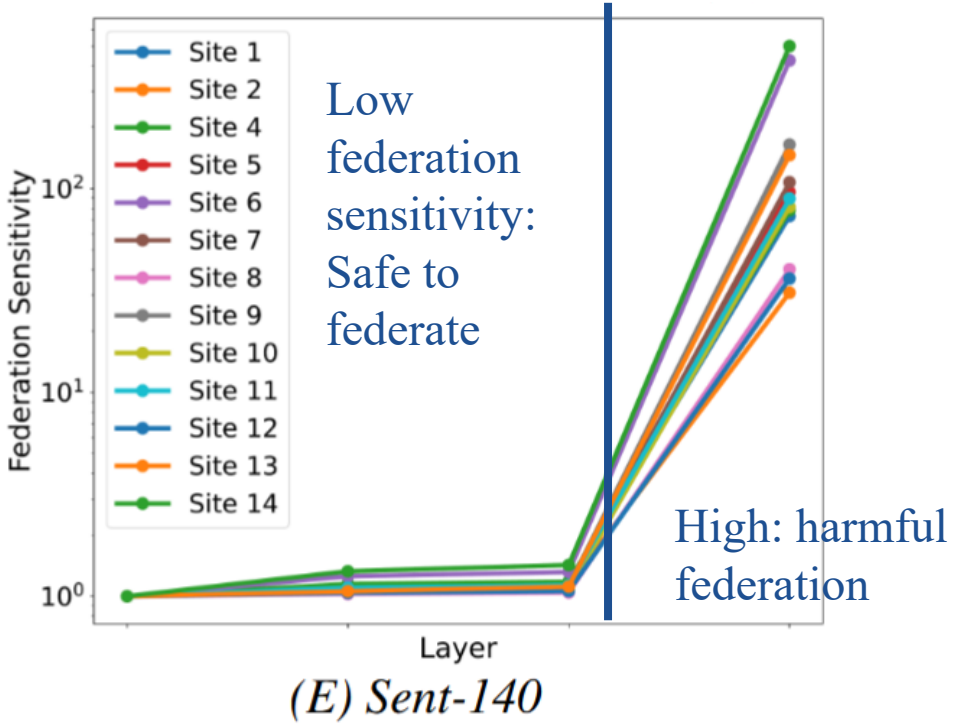
(E) Sent-140

We observe a **sharp transition** in federation sensitivity across layers, enabling principled selection of the layers to federate: those before the transition point



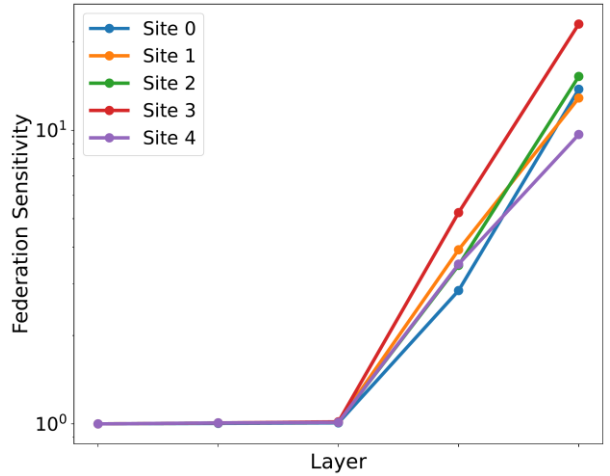
Methods

Federation sensitivity reveals a layer cutoff



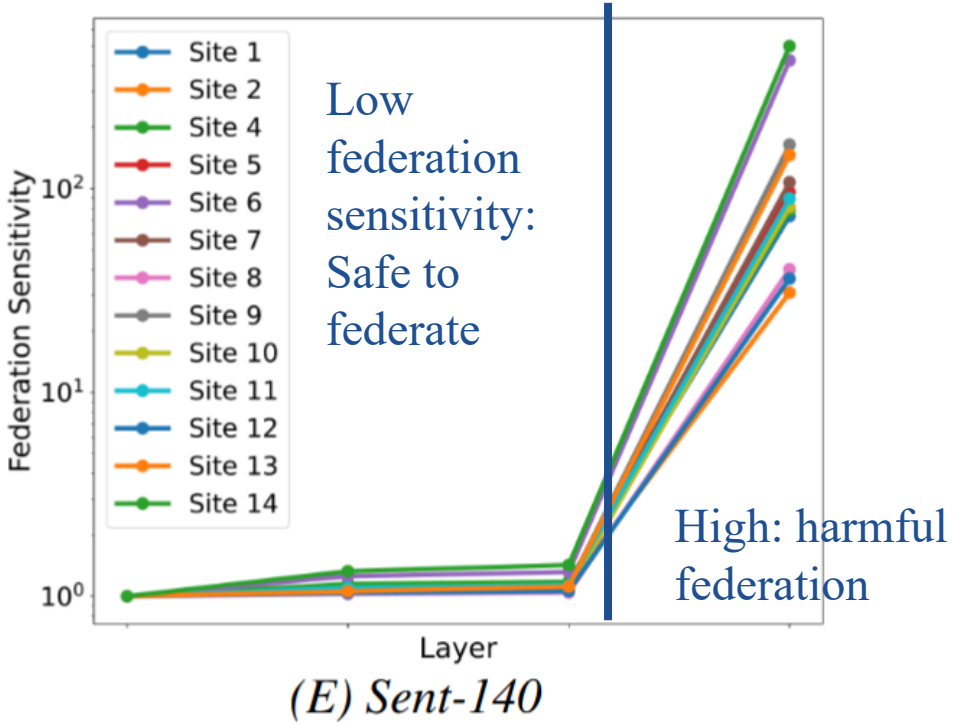
(E) Sent-140

We observe a **sharp transition** in federation sensitivity across layers, enabling principled selection of the layers to federate: those before the transition point



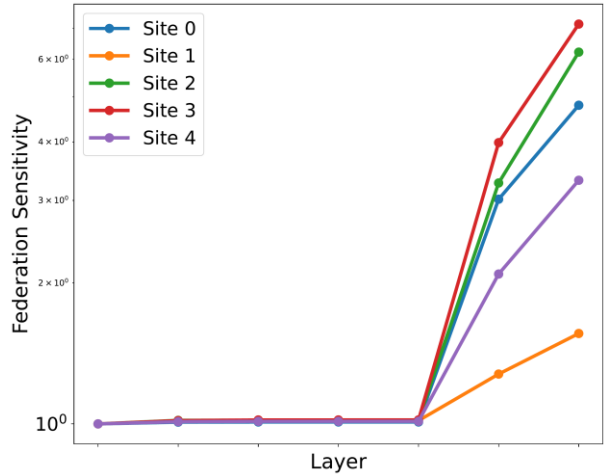
Methods

Federation sensitivity reveals a layer cutoff



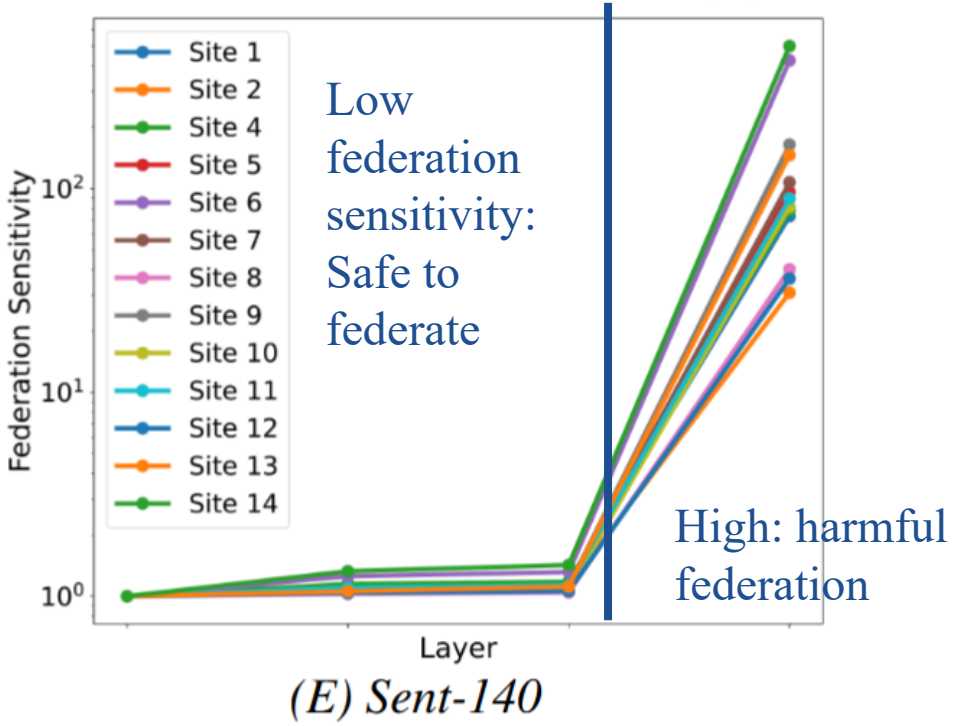
(E) Sent-140

We observe a **sharp transition** in federation sensitivity across layers, enabling principled selection of the layers to federate: those before the transition point



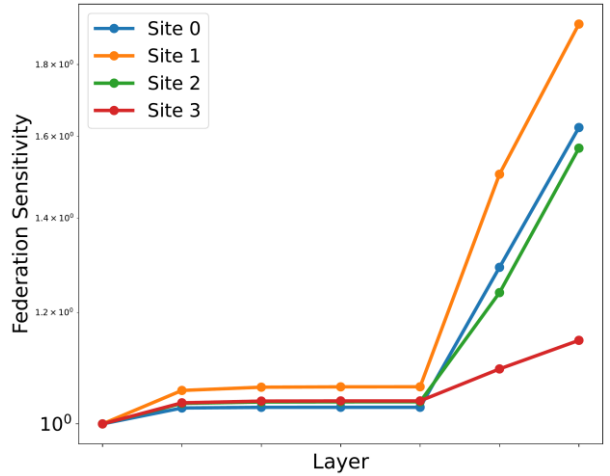
Methods

Federation sensitivity reveals a layer cutoff



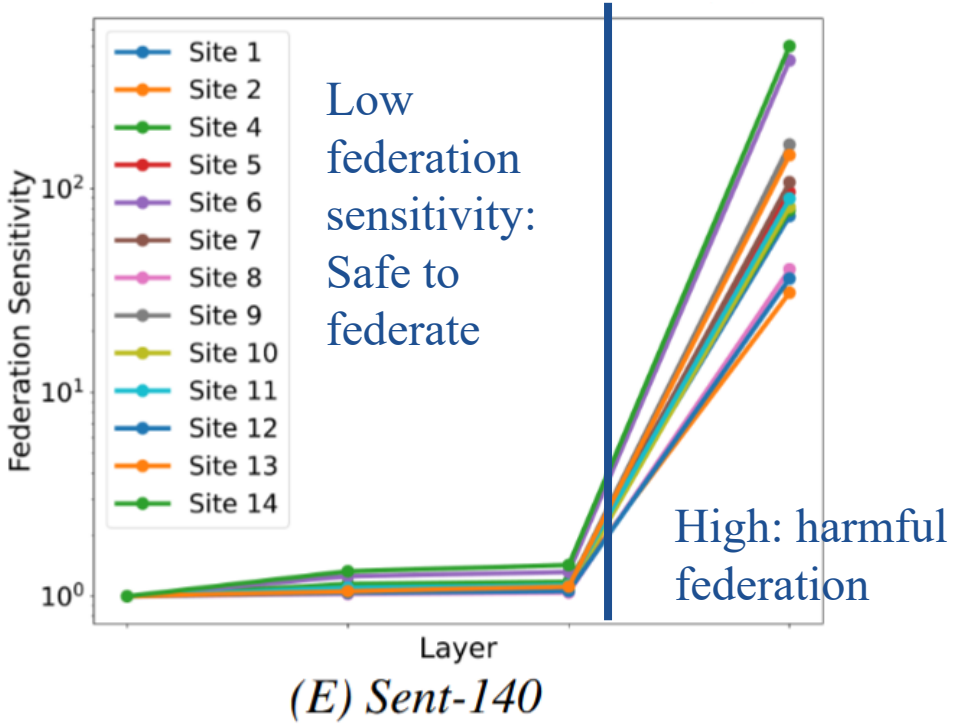
(E) Sent-140

We observe a **sharp transition** in federation sensitivity across layers, enabling principled selection of the layers to federate: those before the transition point



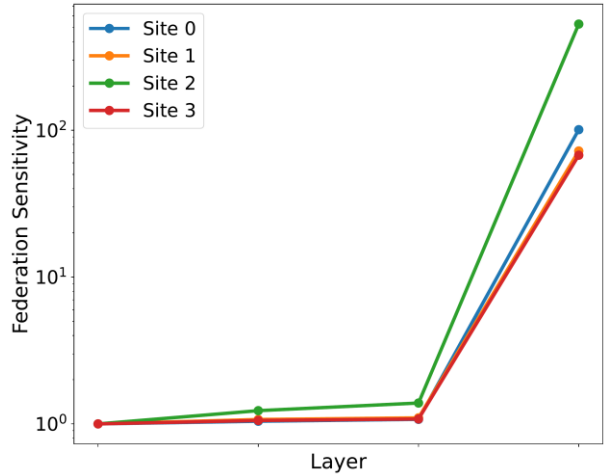
Methods

Federation sensitivity reveals a layer cutoff



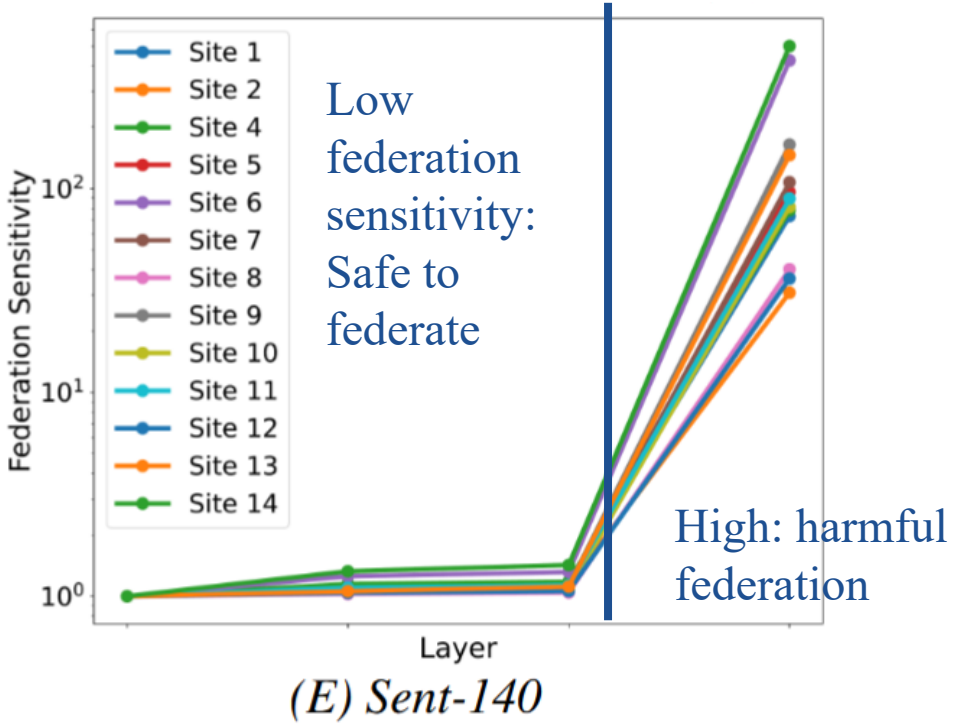
(E) Sent-140

We observe a **sharp transition** in federation sensitivity across layers, enabling principled selection of the layers to federate: those before the transition point



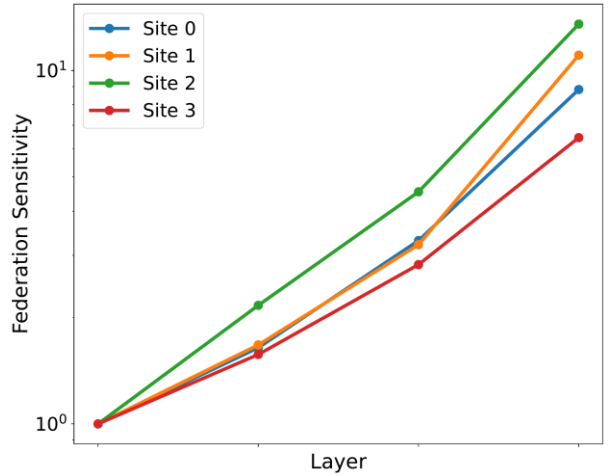
Methods

Federation sensitivity reveals a layer cutoff



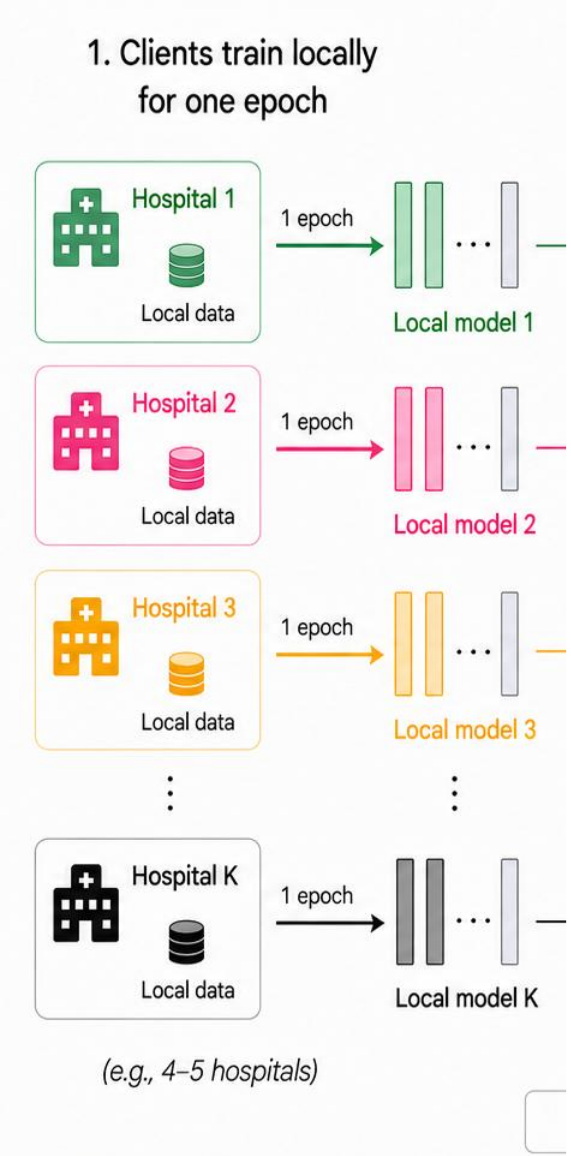
(E) Sent-140

We observe a **sharp transition** in federation sensitivity across layers, enabling principled selection of the layers to federate: those before the transition point

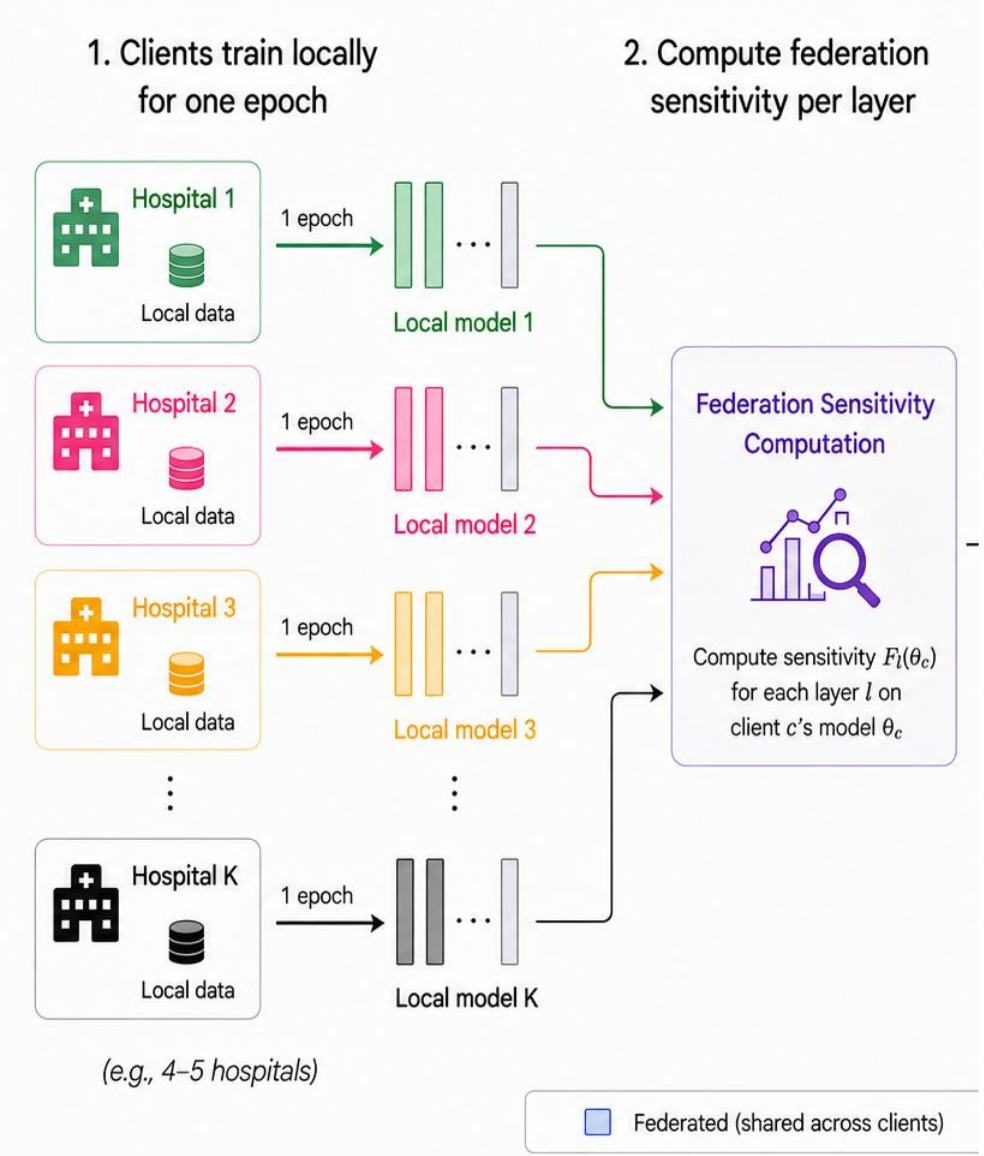


Methods
Algorithm

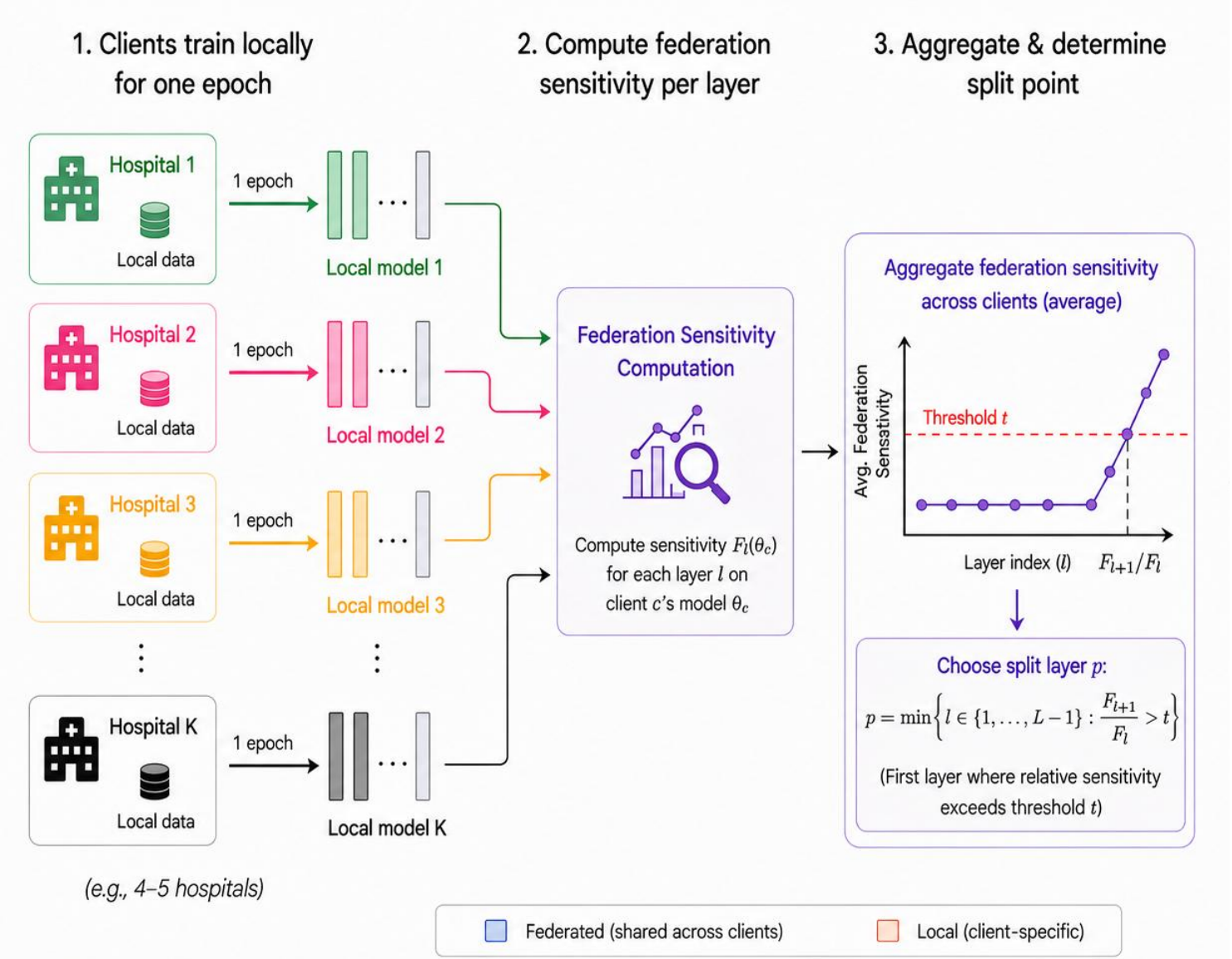
Methods
Algorithm



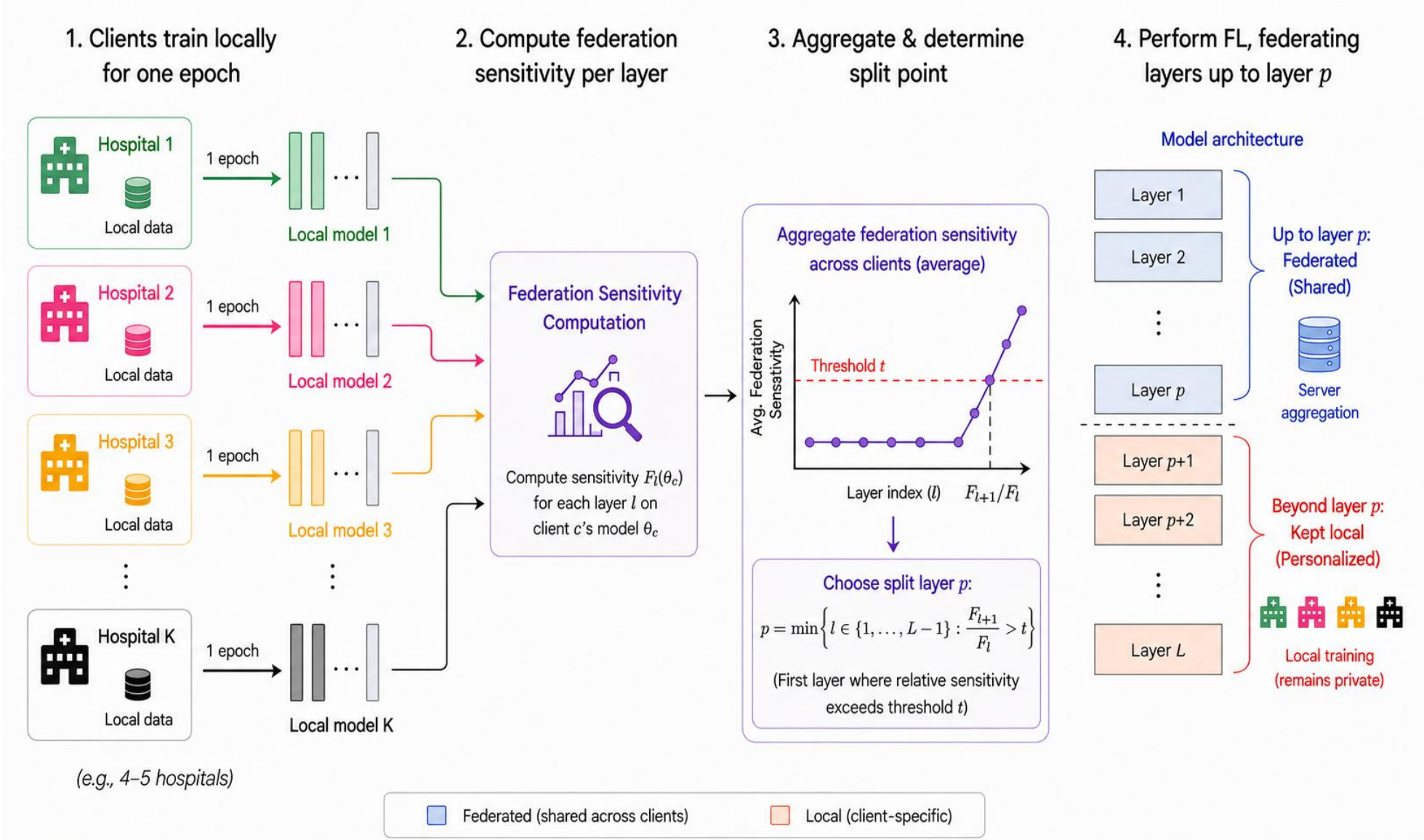
Methods
Algorithm



Methods
Algorithm



Methods
Algorithm



Methods
Evaluation setup - datasets

Methods
Evaluation setup - datasets

Dataset	Modality / task	Classes	Partition	Architecture
FashionMNIST	Fashion images	10	Label skew, Dir(0.5)	CNN
EMNIST	Handwritten digits/characters	62	Label skew, Dir(0.5)	CNN
CIFAR-10	Color images	10	Label skew, Dir(0.5)	CNN
ISIC-2019	Skin lesion images	4	4 hospitals	CNN
Fed-Heart-Disease	Heart disease tabular data	5	5 hospitals	MLP
Sent-140	Tweet sentiment classification	2	15 users	Transformer
MIMIC-III	Mortality prediction from admission notes	2	Grouped by admitting diagnosis group	Transformer

Methods
Evaluation setup - datasets

Dataset	Modality / task	Classes	Partition	Architecture
FashionMNIST	Fashion images	10	Label skew, Dir(0.5)	CNN
EMNIST	Handwritten digits/characters	62	Label skew, Dir(0.5)	CNN
CIFAR-10	Color images	10	Label skew, Dir(0.5)	CNN
ISIC-2019	Skin lesion images	4	4 hospitals	CNN
Fed-Heart-Disease	Heart disease tabular data	5	5 hospitals	MLP
Sent-140	Tweet sentiment classification	2	15 users	Transformer
MIMIC-III	Mortality prediction from admission notes	2	Grouped by admitting diagnosis group	Transformer

Dir: Dirichlet distribution

Methods
Evaluation setup - baselines

Methods
Evaluation setup - baselines

Category	Methods	Key idea
Local	Local training	No collaboration
Global	FedAvg	One shared model
Personalized	FedProx, pFedMe, Ditto, LocalAdaptation	**Client-specific models**
Partial	FedBABU, FedLP, FedLAMA, pFedLA	**Share only part of the model**

Methods
Evaluation setup - metrics

Methods
Evaluation setup - metrics

- Performance (Macro-F1 Score)

Methods
Evaluation setup - metrics

- Performance (Macro-F1 Score)

- Fairness
  - C is the number of clients
  - P_c is the performance of client c
  - $\overline{P_c}$ is the average performance of all clients for a given personalized algorithm

$$\text{Fairness} = \frac{1}{C}\sum_{c=1}^{C}\left(P_c - \overline{P_c}\right)^2$$

Methods
Evaluation setup - metrics

- Performance (Macro-F1 Score)

- Fairness
  - C is the number of clients
  - P_c is the performance of client c
  - $\overline{P_c}$ is the average performance of all clients for a given personalized algorithm

$$\text{Fairness} = \frac{1}{C}\sum_{c=1}^{C}\left(P_c - \overline{P_c}\right)^2$$

- Incentivization
  - S_c is the performance of the local site model in client c
  - G_c is the performance of the global FedAvg model in client c

$$\text{Incentivization} = \frac{1}{C}\sum_{c=1}^{C}\mathbb{I}\{P_c > \max(S_c, G_c)\}$$

Results
One dataset: Sent-140

Results
One dataset: Sent-140

Algorithm	Macro-F1 ↑	Fairness ↓	Incentivization ↑
Local	57.4	—	—
FedAvg	52.6	—	—
FedProx	52.6	**4.1e-3**	6.7
pFedMe	*58.9*	3.7e-2	6.7
Ditto	*58.9*	3.6e-2	**26.7**
LocalAdaptation	53.2	4.1e-2	0.0
FedBABU	58.3	3.5e-2	*20.0*
FedLP	54.4	3.7e-2	6.7
FedLAMA	50.3	4.1e-2	0.0
pFedLA	50.3	4.1e-2	0.0
PLayer-FL	**59.6**	*3.3e-2*	*20.0*

Results
One dataset: Sent-140

Bold: best

Italic: second best

PLayer-FL is in the top-2 for each metric

Algorithm	Macro-F1 ↑	Fairness ↓	Incentivization ↑
Local	57.4	—	—
FedAvg	52.6	—	—
FedProx	52.6	**4.1e-3**	6.7
pFedMe	*58.9*	3.7e-2	6.7
Ditto	*58.9*	3.6e-2	**26.7**
LocalAdaptation	53.2	4.1e-2	0.0
FedBABU	58.3	3.5e-2	*20.0*
FedLP	54.4	3.7e-2	6.7
FedLAMA	50.3	4.1e-2	0.0
pFedLA	50.3	4.1e-2	0.0
PLayer-FL	**59.6**	*3.3e-2*	*20.0*

Results
Mean rank comparison

Results
Mean rank comparison

Algorithm (type)	Mean rank performance	Fairness rank	Incentivization rank
Local training (local)	6.9	—	—
FedAvg (global)	6.0	—	—
FedProx (personalized)	6.3	7.7	5.4
pFedMe (personalized)	5.3	4.3	5.1
Ditto (personalized)	6.0	5.0	5.6
LocalAdaptation (personalized)	6.4	6.4	5.2
FedBABU (partial)	4.1	4.4	4.4
FedLP (partial)	6.7	6.1	6.2
FedLAMA (partial)	10.1	5.6	7.3
pFedLA (partial)	11.4	5.1	7.7
PLayer-FL (partial)	**2.6**	**3.8**	**3.4**

Results
Mean rank comparison

Lower is better

Algorithm (type)	Mean rank performance	Fairness rank	Incentivization rank
Local training (local)	6.9	—	—
FedAvg (global)	6.0	—	—
FedProx (personalized)	6.3	7.7	5.4
pFedMe (personalized)	5.3	4.3	5.1
Ditto (personalized)	6.0	5.0	5.6
LocalAdaptation (personalized)	6.4	6.4	5.2
FedBABU (partial)	4.1	4.4	4.4
FedLP (partial)	6.7	6.1	6.2
FedLAMA (partial)	10.1	5.6	7.3
pFedLA (partial)	11.4	5.1	7.7
PLayer-FL (partial)	**2.6**	**3.8**	**3.4**

Results
Mean rank comparison

Lower is better

PLayer-FL outperforms all the other methods in terms of mean ranks

Algorithm (type)	Mean rank performance	Fairness rank	Incentivization rank
Local training (local)	6.9	—	—
FedAvg (global)	6.0	—	—
FedProx (personalized)	6.3	7.7	5.4
pFedMe (personalized)	5.3	4.3	5.1
Ditto (personalized)	6.0	5.0	5.6
LocalAdaptation (personalized)	6.4	6.4	5.2
FedBABU (partial)	4.1	4.4	4.4
FedLP (partial)	6.7	6.1	6.2
FedLAMA (partial)	10.1	5.6	7.3
pFedLA (partial)	11.4	5.1	7.7
PLayer-FL (partial)	**2.6**	**3.8**	**3.4**

Results
System Analysis

Results
System Analysis

Dataset	Domain	Architecture	Per-round (FedAvg)	Per-round (PLayer-FL)	Savings (%)
EMNIST	Vision	CNN	42.3 MB	41.0 MB	**3.0**
FMNIST	Vision	CNN	42.1 MB	41.0 MB	**2.5**
CIFAR	Vision	CNN	158.5 MB	156.4 MB	**1.3**
ISIC	Medical imaging	CNN	54.4 MB	50.2 MB	**7.8**
Sent-140	NLP	Transformer	1.7 GB	1.5 GB	**16.3**
Heart	Tabular	MLP	67.0 KB	50.2 KB	**25.1**
MIMIC-III	Clinical (EHR)	Transformer	554.7 MB	479.1 MB	**13.6**

Results
System Analysis

PLayer-FL **reduces communication** by federating only the first layers

Dataset	Domain	Architecture	Per-round (FedAvg)	Per-round (PLayer-FL)	Savings (%)
EMNIST	Vision	CNN	42.3 MB	41.0 MB	**3.0**
FMNIST	Vision	CNN	42.1 MB	41.0 MB	**2.5**
CIFAR	Vision	CNN	158.5 MB	156.4 MB	**1.3**
ISIC	Medical imaging	CNN	54.4 MB	50.2 MB	**7.8**
Sent-140	NLP	Transformer	1.7 GB	1.5 GB	**16.3**
Heart	Tabular	MLP	67.0 KB	50.2 KB	**25.1**
MIMIC-III	Clinical (EHR)	Transformer	554.7 MB	479.1 MB	**13.6**

Takeaways

Takeaways

- Global FL sometimes fails because of distribution shifts

Takeaways

- Global FL sometimes fails because of distribution shifts
- Layered FL is promising, but existing methods often choose shared layers ad hoc

Takeaways

- Global FL sometimes fails because of distribution shifts
- Layered FL is promising, but existing methods often choose shared layers ad hoc
- We propose a principled method to select which layers should be shared

Takeaways

- Global FL sometimes fails because of distribution shifts
- Layered FL is promising, but existing methods often choose shared layers ad hoc
- We propose a principled method to select which layers should be shared
- Cheap (1 epoch), efficient (communication savings)

Takeaways

- Global FL sometimes fails because of distribution shifts
- Layered FL is promising, but existing methods often choose shared layers ad hoc
- We propose a principled method to select which layers should be shared
- Cheap (1 epoch), efficient (communication savings)
- Makes cross-silo FL practical, especially for healthcare

Paper

Takeaways

- Global FL sometimes fails because of distribution shifts
- Layered FL is promising, but existing methods often choose shared layers ad hoc
- We propose a principled method to select which layers should be shared
- Cheap (1 epoch), efficient (communication savings)
- Makes cross-silo FL practical, especially for healthcare

Paper

Takeaways

- Global FL sometimes fails because of distribution shifts
- Layered FL is promising, but existing methods often choose shared layers ad hoc
- We propose a principled method to select which layers should be shared
- Cheap (1 epoch), efficient (communication savings)
- Makes cross-silo FL practical, especially for healthcare

Thanks to the G2Lab!

Paper

Takeaways

- Global FL sometimes fails because of distribution shifts
- Layered FL is promising, but existing methods often choose shared layers ad hoc
- We propose a principled method to select which layers should be shared
- Cheap (1 epoch), efficient (communication savings)
- Makes cross-silo FL practical, especially for healthcare

Thanks to the G2Lab!

Thank you! Any questions?

Bibliography

Baseline method citations

FedAvg: McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, *54*, 1273–1282.
FedProx: Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, *2*, 429–450.
pFedMe: Dinh, C. T., Tran, N. H., & Nguyen, T. D. (2020). Personalized federated learning with Moreau envelopes. *Advances in Neural Information Processing Systems*, *33*, 21394–21405.
Ditto: Li, T., Hu, S., Beirami, A., & Smith, V. (2021). Ditto: Fair and robust federated learning through personalization. *Proceedings of the 38th International Conference on Machine Learning*, *139*, 6357–6368.
LocalAdaptation: Yu, T., Bagdasaryan, E., & Shmatikov, V. (2020). Salvaging federated learning by local adaptation. *arXiv preprint arXiv:2002.04758*.
FedBABU: Oh, J., Kim, S., & Yun, S.-Y. (2022). FedBABU: Towards enhanced representation for federated image classification. *International Conference on Learning Representations*.
FedLP: Zhu, Z., Shi, Y., Luo, J., Wang, F., Peng, C., Fan, P., & Letaief, K. B. (2023). FedLP: Layer-wise pruning mechanism for communication-computation efficient federated learning. In *ICC 2023 IEEE International Conference on Communications* (pp. 1250–1255). IEEE.
FedLAMA: Lee, S., Zhang, T., & Avestimehr, A. S. (2023). Layer-wise adaptive model aggregation for scalable federated learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(7), 8491–8499.
pFedLA: Ma, X., Zhang, J., Guo, S., & Xu, W. (2022). Layer-wised model aggregation for personalized federated learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10092–10101.

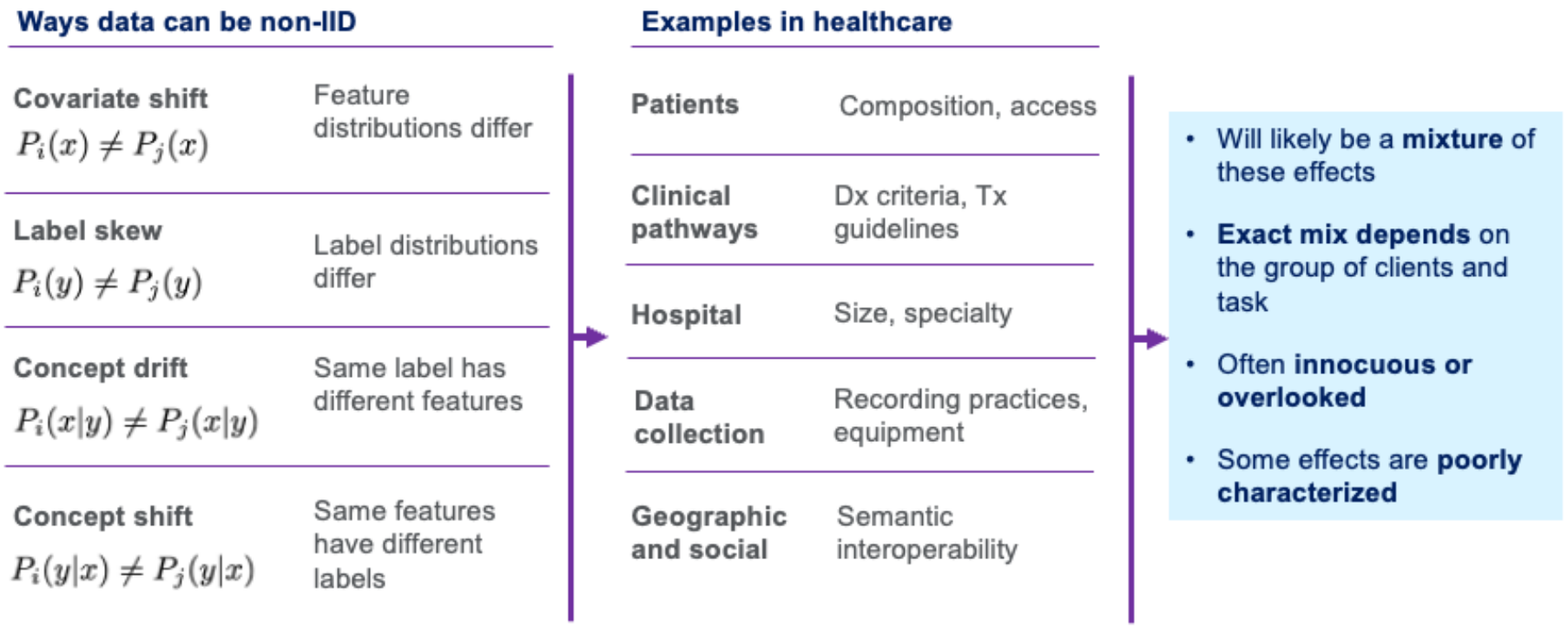
Dataset citations
FashionMNIST: Xiao, H., Rasul, K., & Vollgraf, R. (2017). Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.
EMNIST: Cohen, G., Afshar, S., Tapson, J., & van Schaik, A. (2017). EMNIST: Extending MNIST to handwritten letters. In *2017 International Joint Conference on Neural Networks* (pp. 2921–2926). IEEE.
CIFAR-10: Krizhevsky, A. (2009). *Learning multiple layers of features from tiny images*. Technical Report, University of Toronto.
ISIC-2019 / Fed-ISIC partition: Ogier du Terrail, J., Ayed, S.-S., Cyffers, E., Grimberg, F., He, C., Loeb, R., Mangold, P., Marchand, T., Marfoq, O., Mushtaq, E., et al. (2022). FLamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings. *Advances in Neural Information Processing Systems*, *35*.
Fed-Heart-Disease: Ogier du Terrail, J., Ayed, S.-S., Cyffers, E., Grimberg, F., He, C., Loeb, R., Mangold, P., Marchand, T., Marfoq, O., Mushtaq, E., et al. (2022). FLamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings. *Advances in Neural Information Processing Systems*, *35*.
Sent-140 original dataset: Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1–12.
Sent-140 federated benchmark / LEAF split: Caldas, S., Duddu, S. M. K., Wu, P., Li, T., Konečný, J., McMahan, H. B., Smith, V., & Talwalkar, A. (2018). LEAF: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*.
MIMIC-III: Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, *3*, 160035.

Metrics
Divi, S., Lin, Y.-S., Farrukh, H., & Celik, Z. B. (2021). New metrics to evaluate the performance and fairness of personalized federated learning. *arXiv preprint arXiv:2107.13173*.

Backup slides

Motivation
And healthcare data have many sources of non-IIDness

Motivation
And healthcare data have many sources of non-IIDness



Federation sensitivity after one epoch

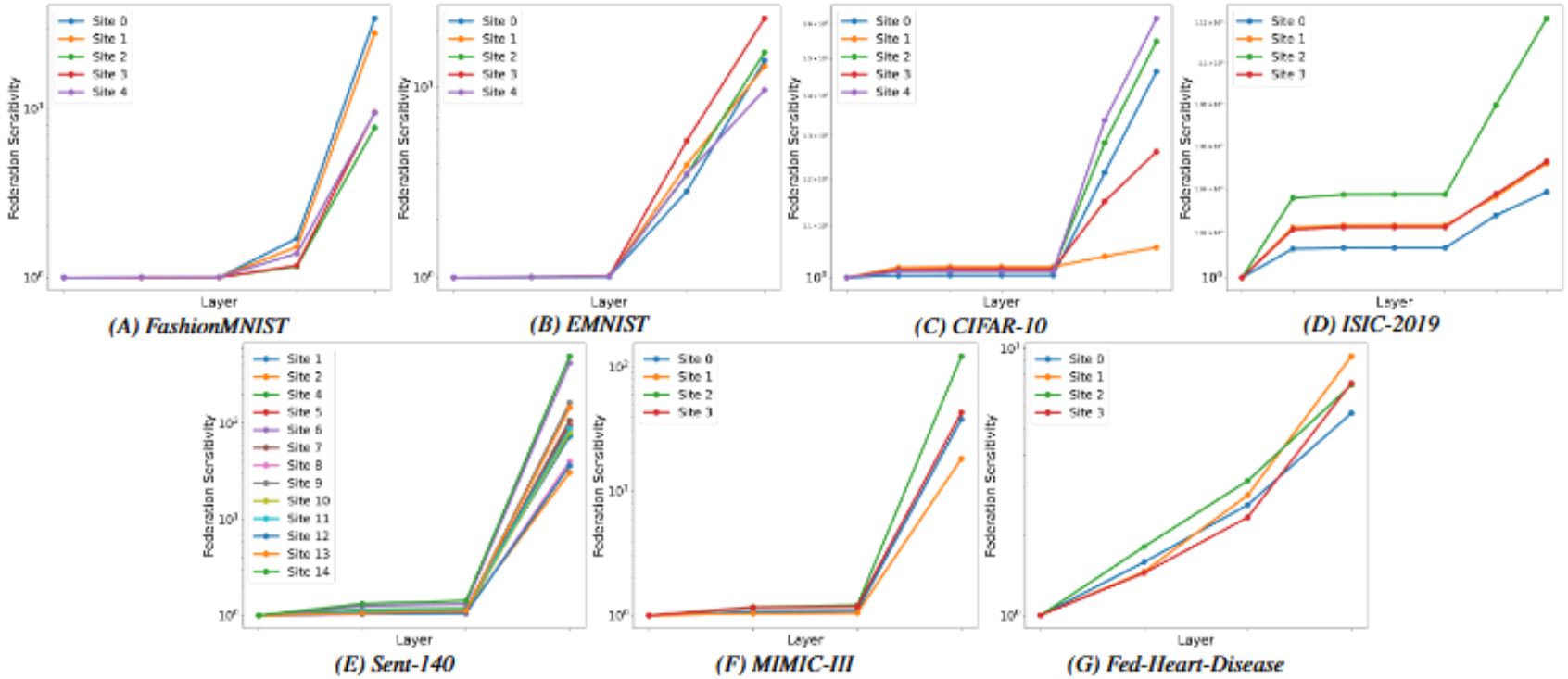


Figure A.10. **Federation sensitivity after one epoch.** All models identically initialized and independently trained on non-IID data.

Federation sensitivity for final models

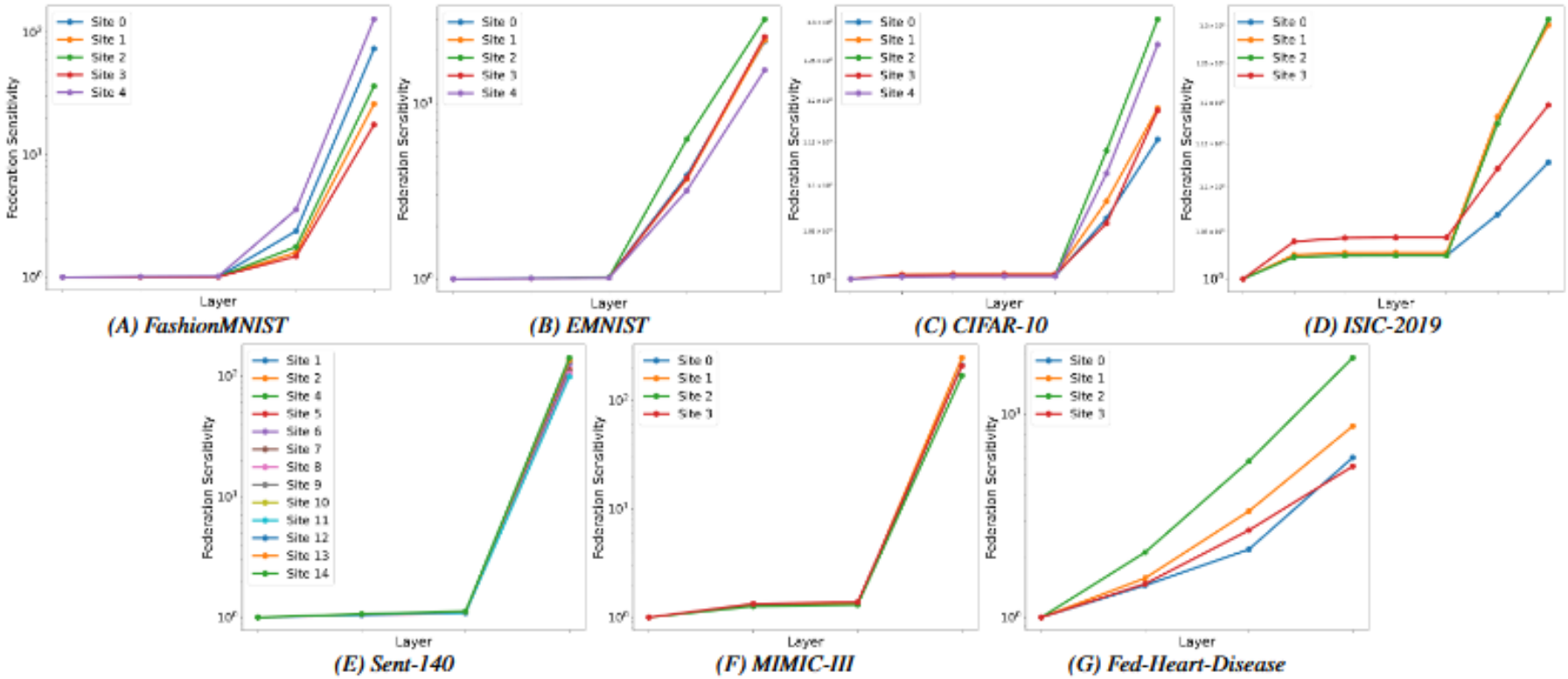


Figure A.11. **Federation sensitivity for final models.** Models trained via FL on non-IID data.

Threshold robustness

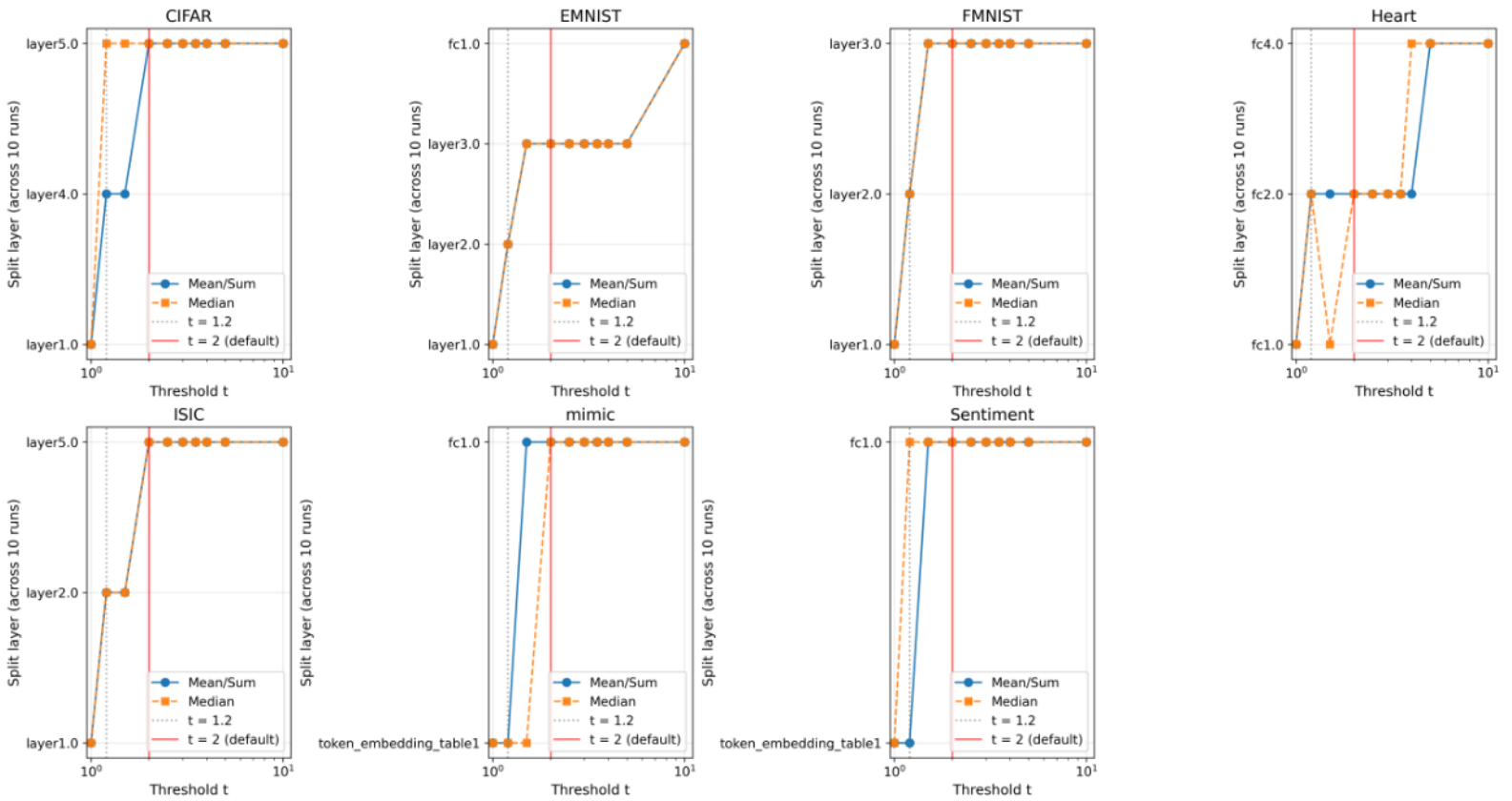


Figure A.1. Split layer (mode across 10 runs) vs. threshold t. The red solid line marks the default t=2; the grey dotted line marks t=1.2. Blue circles: Mean/Sum aggregation; orange squares: Median aggregation. For most datasets, the split layer is stable across a wide range of thresholds under both aggregation strategies.

Performance dataset breakdown

Table 2. Macro-averaged F1 Score and average rank. In **bold** is top-performing model. Friedman rank test p-value $< 5 \times 10^{-3}$

Algorithm	FMNIST	EMNIST	CIFAR	ISIC	Heart	Sentiment	MIMIC	Mean Rank
Local	75.9 ± 1.6	56.9 ± 1.4	61.6 ± 3.2	**53.1 ± 1.3**	**42.0 ± 0.8**	57.4 ± 1.0	58.6 ± 1.5	6.9
FedAvg	75.6 ± 1.8	64.2 ± 1.1	65.4 ± 4.5	43.3 ± 0.7	40.4 ± 1.0	52.6 ± 0.3	63.0 ± 1.7	6.0
FedProx	76.4 ± 1.8	65.3 ± 1.9	65.8 ± 4.9	40.6 ± 2.6	37.2 ± 1.6	52.6 ± 0.3	63.2 ± 0.8	6.3
pFedMe	77.3 ± 1.8	66.2 ± 0.7	48.9 ± 2.5	44.9 ± 2.6	42.5 ± 0.8	58.9 ± 0.4	60.9 ± 0.4	5.3
Ditto	77.8 ± 1.2	61.6 ± 0.8	48.7 ± 1.9	43.4 ± 0.7	40.7 ± 1.0	58.9 ± 0.9	61.0 ± 0.5	6.0
LocalAdaptation	76.9 ± 1.5	63.8 ± 1.8	65.5 ± 3.3	43.5 ± 1.9	39.8 ± 0.9	53.2 ± 0.6	62.6 ± 1.6	6.4
FedBABU	77.2 ± 1.0	**66.1 ± 0.2**	**67.7 ± 4.0**	49.1 ± 3.5	40.3 ± 0.8	58.3 ± 0.2	62.7 ± 0.4	4.1
FedLP	77.2 ± 1.3	64.1 ± 0.9	63.7 ± 5.1	40.2 ± 2.1	40.1 ± 1.0	54.4 ± 1.1	62.1 ± 1.3	6.7
FedLama	71.6 ± 1.7	56.9 ± 1.3	41.8 ± 3.0	36.0 ± 1.1	39.7 ± 0.8	50.3 ± 0.4	**66.7 ± 1.0**	10.1
pFedLA	47.1 ± 0.9	07.1 ± 0.0	10.5 ± 1.6	28.4 ± 1.9	40.1 ± 0.8	50.3 ± 0.4	65.0 ± 2.6	11.4
PLayer-FL	**79.3 ± 1.4**	62.2 ± 1.2	67.1 ± 3.7	52.1 ± 1.5	41.9 ± 1.0	**59.6 ± 1.2**	63.0 ± 0.7	**2.6**
PLayer-FL-Random	76.5 ± 2.3	57.1 ± 1.1	65.7 ± 2.3	52.1 ± 1.5	41.3 ± 0.9	55.7 ± 0.9	61.0 ± 1.4	6.1

Fairness dataset breakdown

Table A.5. Variance in clients' F1 score (fairness). In **bold** is fairest model. Friedman rank test p-value $< 5 \times 10^{-3}$

Algorithm	FMNIST	EMNIST	CIFAR	ISIC	Heart	Sent-140	MIMIC-III	Rank
Fedprox	$3.0 \cdot 10^{-4}$	$8.0 \cdot 10^{-4}$	$4.0 \cdot 10^{-4}$	$1.1 \cdot 10^{-2}$	$9.0 \cdot 10^{-2}$	$4.1 \cdot 10^{-3}$	$1.7 \cdot 10^{-3}$	7.7
pFedMe	$3.0 \cdot 10^{-4}$	$6.0 \cdot 10^{-4}$	$5.0 \cdot 10^{-4}$	$4.2 \cdot 10^{-3}$	$\mathbf{7.1 \cdot 10^{-2}}$	$3.7 \cdot 10^{-2}$	$\mathbf{1.4 \cdot 10^{-3}}$	4.3
Ditto	$5.0 \cdot 10^{-4}$	$6.0 \cdot 10^{-4}$	$2.0 \cdot 10^{-4}$	$3.6 \cdot 10^{-3}$	$8.2 \cdot 10^{-2}$	$3.6 \cdot 10^{-2}$	$1.6 \cdot 10^{-3}$	5.0
LocalAdaptation	$\mathbf{1.0 \cdot 10^{-4}}$	$7.0 \cdot 10^{-4}$	$4.0 \cdot 10^{-4}$	$9.0 \cdot 10^{-3}$	$8.3 \cdot 10^{-2}$	$4.1 \cdot 10^{-2}$	$2.2 \cdot 10^{-3}$	6.4
FedBABU	$2.0 \cdot 10^{-4}$	$4.0 \cdot 10^{-4}$	$4.0 \cdot 10^{-4}$	$2.1 \cdot 10^{-3}$	$8.4 \cdot 10^{-2}$	$3.5 \cdot 10^{-2}$	$2.0 \cdot 10^{-3}$	4.4
FedLP	$3.0 \cdot 10^{-4}$	$4.0 \cdot 10^{-4}$	$1.0 \cdot 10^{-3}$	$1.1 \cdot 10^{-2}$	$8.5 \cdot 10^{-2}$	$3.7 \cdot 10^{-2}$	$1.6 \cdot 10^{-3}$	6.1
FedLAMA	$2.0 \cdot 10^{-4}$	$6.0 \cdot 10^{-4}$	$5.0 \cdot 10^{-4}$	$4.7 \cdot 10^{-3}$	$8.7 \cdot 10^{-2}$	$4.1 \cdot 10^{-2}$	$2.0 \cdot 10^{-3}$	5.6
pFedLA	$4.0 \cdot 10^{-4}$	$1.0 \cdot 10^{-4}$	$\mathbf{1.0 \cdot 10^{-4}}$	$3.5 \cdot 10^{-3}$	$7.4 \cdot 10^{-2}$	$4.1 \cdot 10^{-2}$	$2.1 \cdot 10^{-3}$	5.1
PLayer-FL	$4.0 \cdot 10^{-4}$	$5.0 \cdot 10^{-4}$	$6.0 \cdot 10^{-4}$	$1.3 \cdot 10^{-3}$	$7.3 \cdot 10^{-2}$	$\mathbf{3.3 \cdot 10^{-2}}$	$1.5 \cdot 10^{-3}$	**3.8**
PLayer-FL-Random	$8.0 \cdot 10^{-4}$	$\mathbf{3.0 \cdot 10^{-4}}$	$1.0 \cdot 10^{-3}$	$\mathbf{7.0 \cdot 10^{-4}}$	$7.3 \cdot 10^{-2}$	$3.4 \cdot 10^{-2}$	$1.9 \cdot 10^{-3}$	6.5

Incentivization dataset breakdown

Table A.6. Incentivized participation rate (%) using F1 score. In **bold** is model with highest IPR. Friedman rank test p-value = 0.043

Algorithm	FMNIST	EMNIST	CIFAR	ISIC	Heart	Sentiment	Mimic-III	Rank
FedProx	0.0	20.0	40.0	0.0	0.0	6.7	50.0	5.4
pFedMe	80.0	20.0	0.0	0.0	**50.0**	6.7	0.0	5.1
Ditto	60.0	0.0	0.0	0.0	0.0	**26.7**	25.0	5.6
LocalAdaptation	0.0	40.0	40.0	0.0	0.0	0.0	**75.0**	5.2
FedBABU	0.0	40.0	80.0	0.0	0.0	20.0	50.0	4.4
FedLP	0.0	**60.0**	0.0	0.0	0.0	6.7	50.0	6.2
FedLAMA	0.0	0.0	0.0	0.0	0.0	0.0	25.0	7.3
pFedLA	0.0	0.0	0.0	0.0	0.0	0.0	0.0	7.7
PLayer-FL	**100.0**	0.0	**100.0**	0.0	25.0	20.0	50.0	**3.4**
PLayer-FL-Random	60.0	0.0	80.0	0.0	25.0	6.7	25.0	4.8