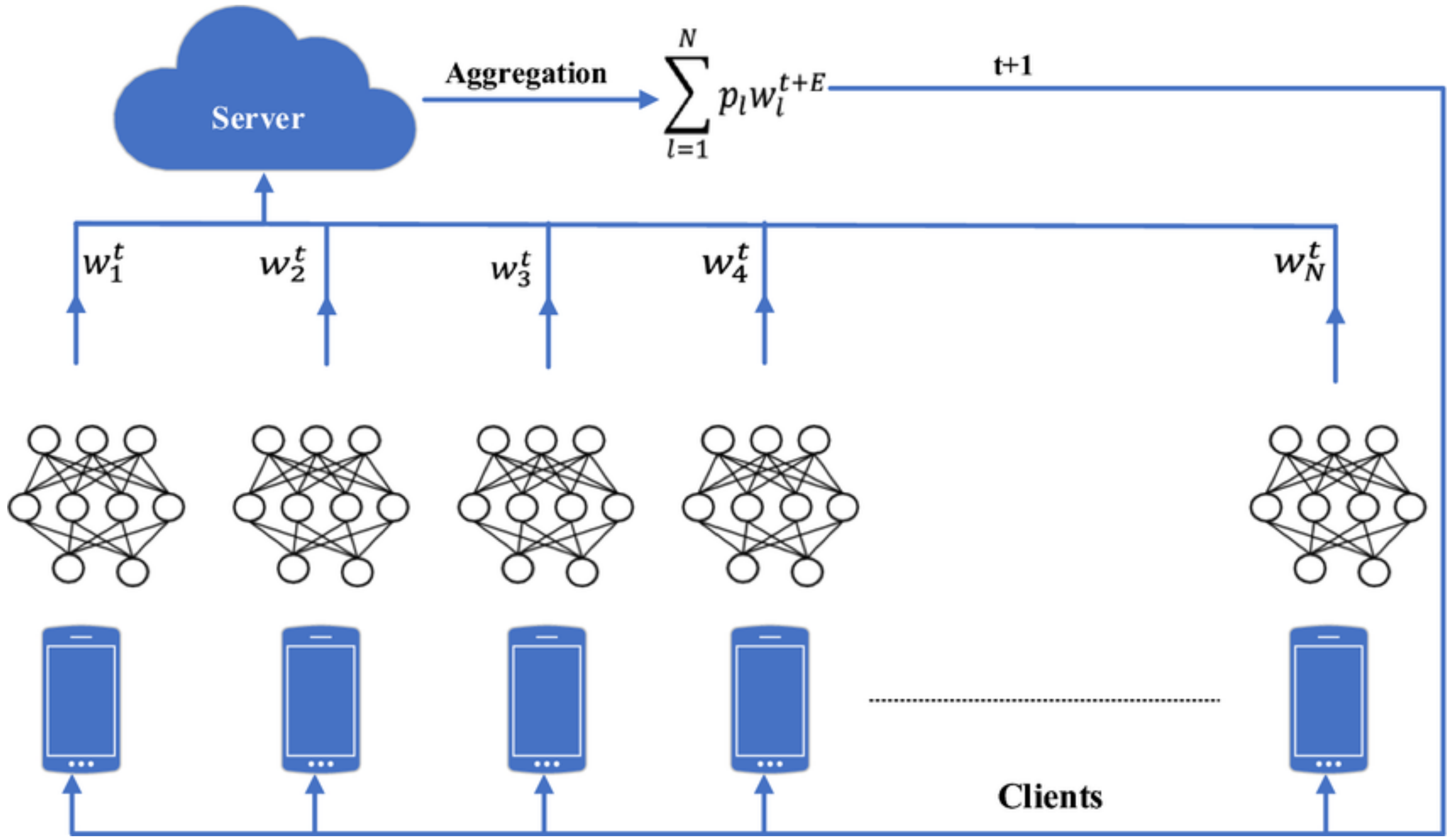
Model-Agnostic Personalized Federated Learning

with Clustered Server Models and Pseudo-Label-Only Communication (COSMOS)

Ben Rachmut, **Luise Ge**, Ning Zhang, William Yeoh, Yevgeniy Vorobeychik   Washington University in St. Louis

Federated learning offers an appealing way to do collaborative training without moving a large quantity of private data [McMahan et al. 2016].

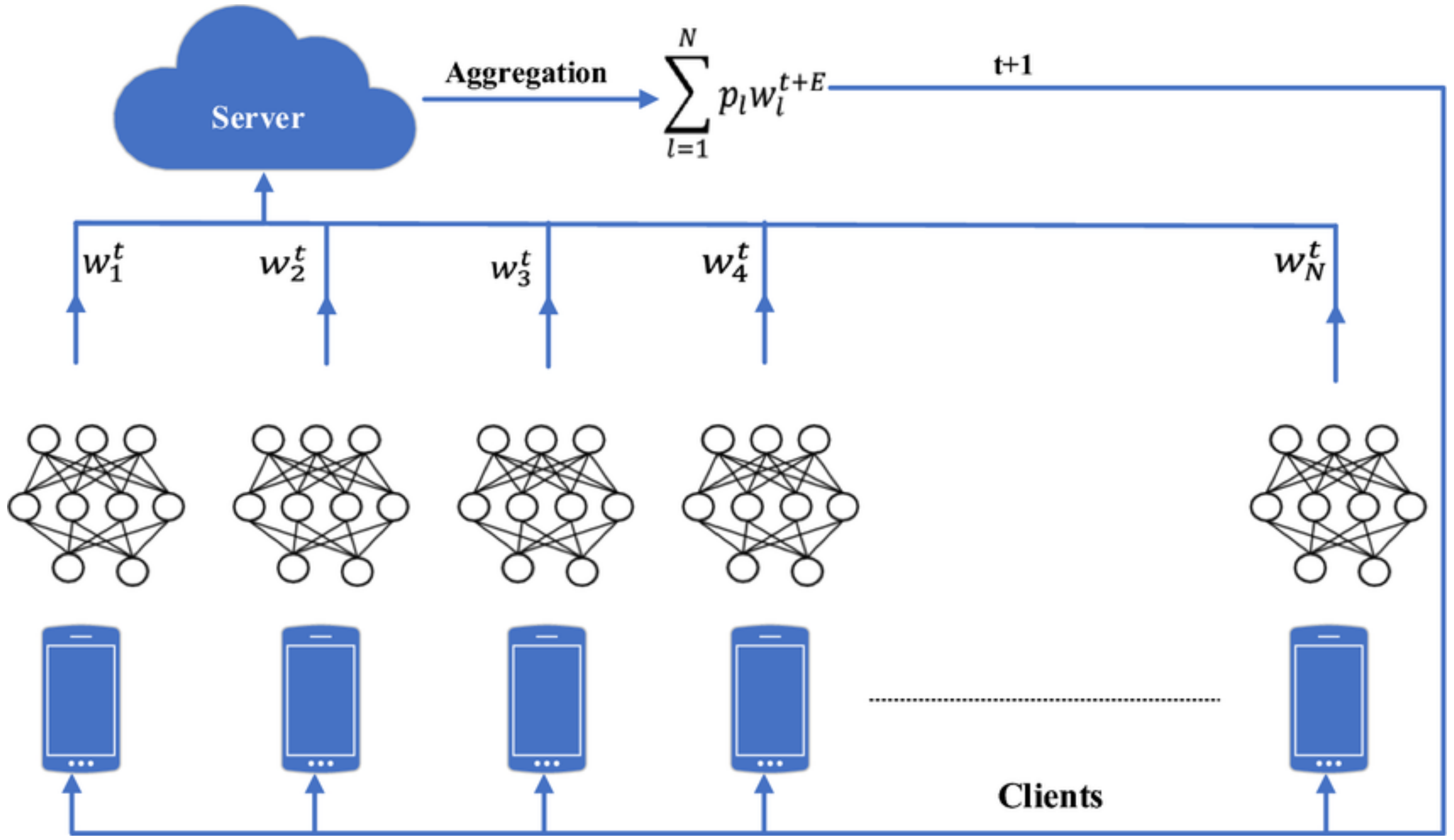


Communication-Efficient Learning of Deep Networks from Decentralized Data McMahan et al. 2016

But it relies on two assumptions:

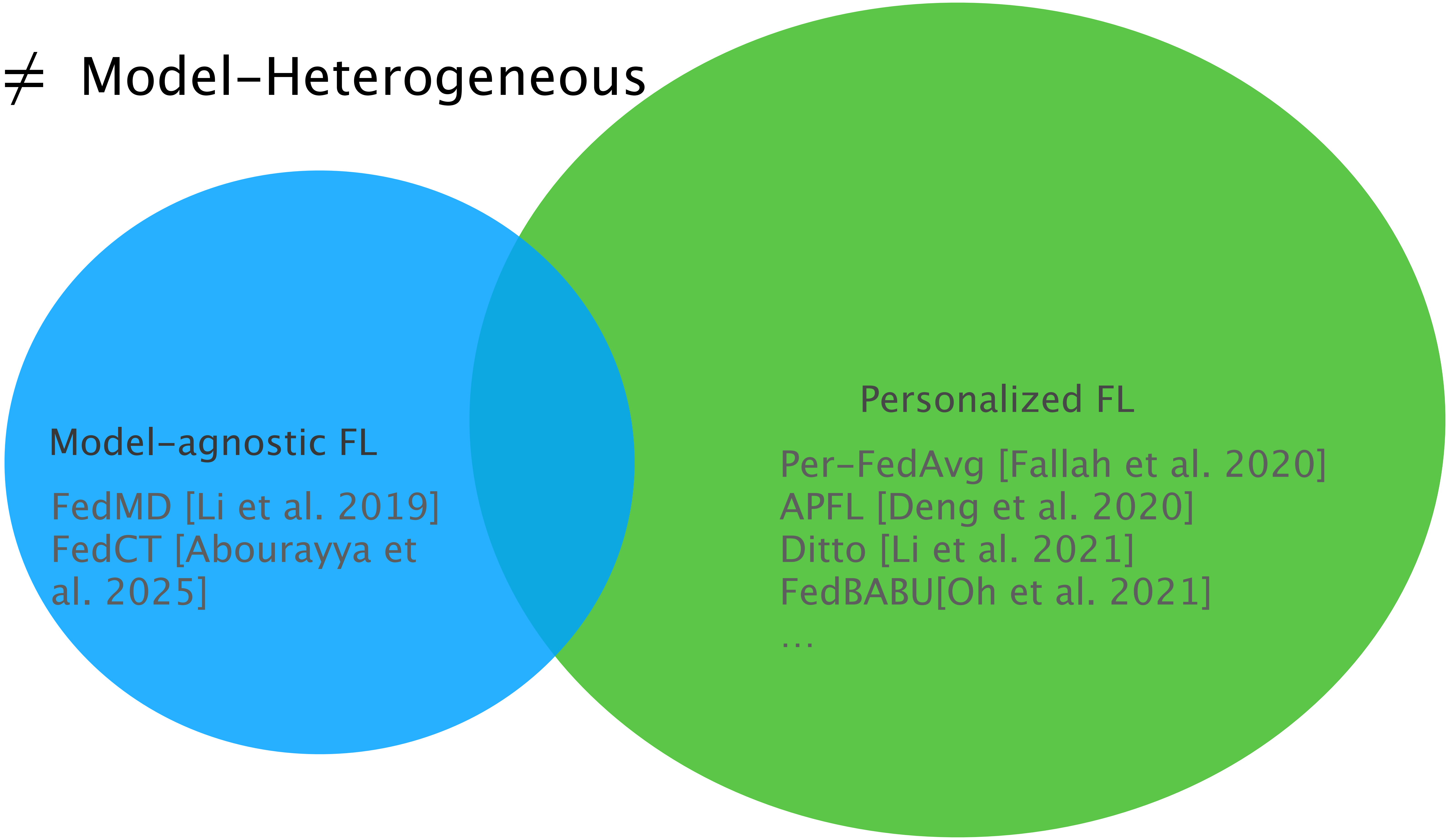
model homogeneity – server and edge devices use the same model;

data homogeneity – every edge device faces the same data distribution.

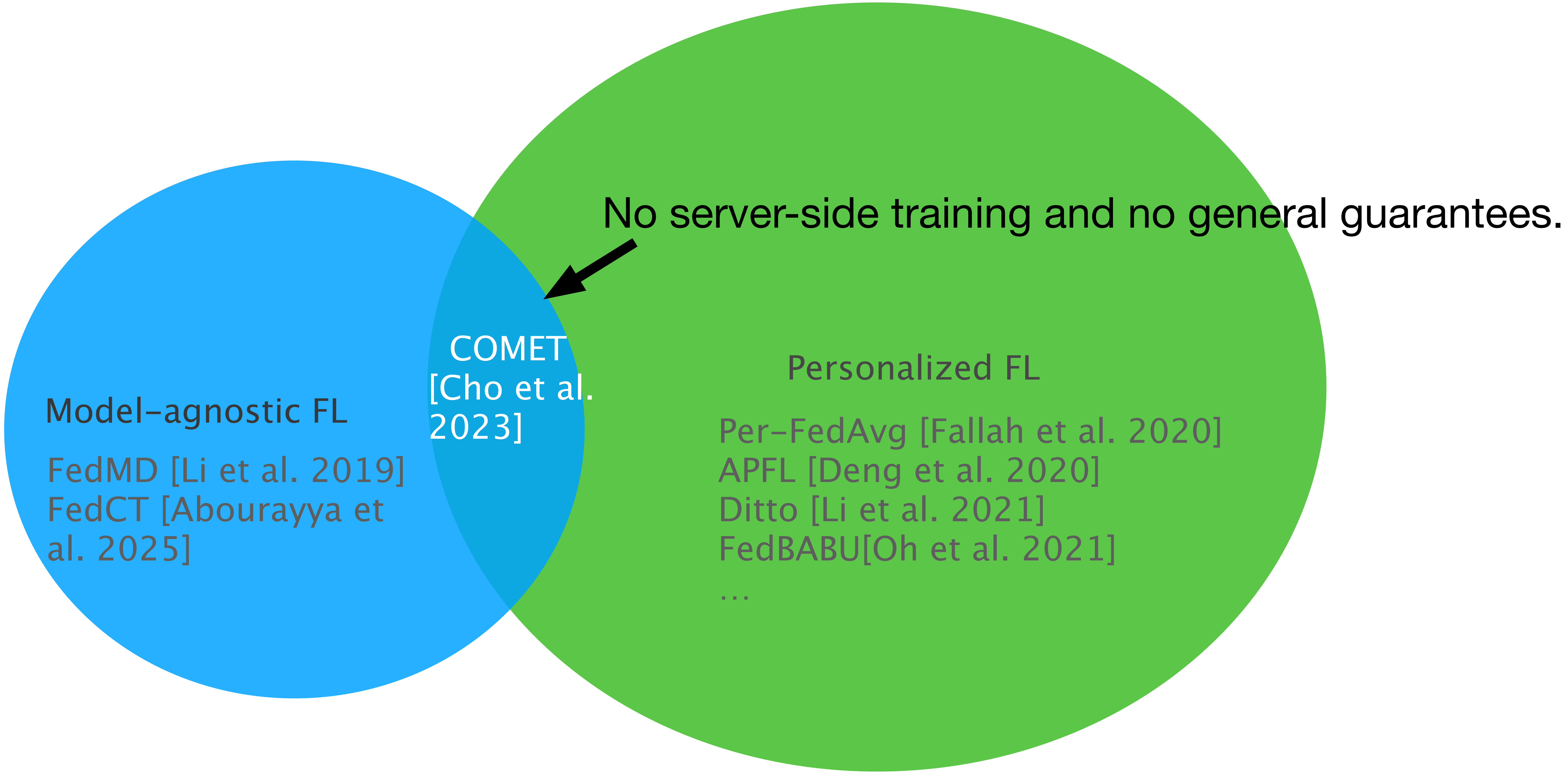


Prior Work

Model-Agnostic $\neq$ Model-Heterogeneous



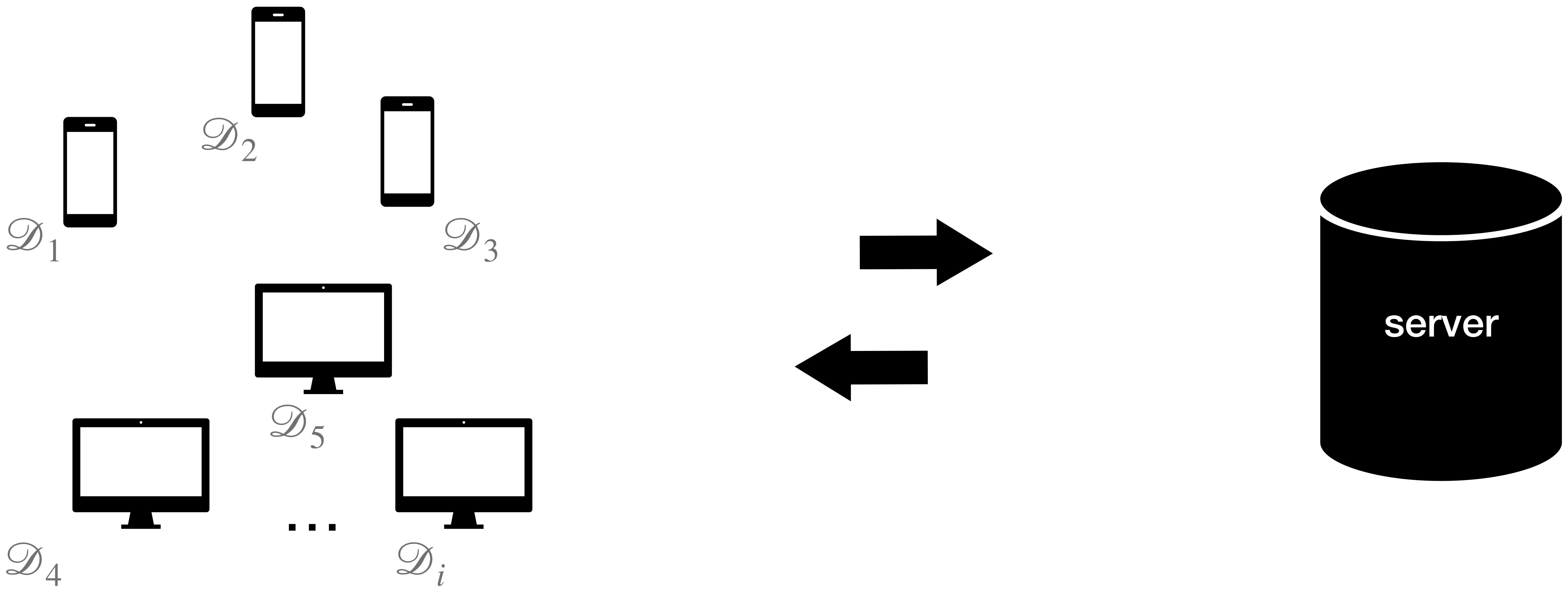
Prior Work



Very limited work on the intersection

Research question

Is there a better way to enable personalized FL under no model assumption?



Proposed Solution

I. Pre-training and clustering

Local training

Each client trains on private data and emits pseudo-labels on the shared public pool.

Clustering

The server clusters clients by pseudo-label similarity

II. Iterative federated fine-tuning

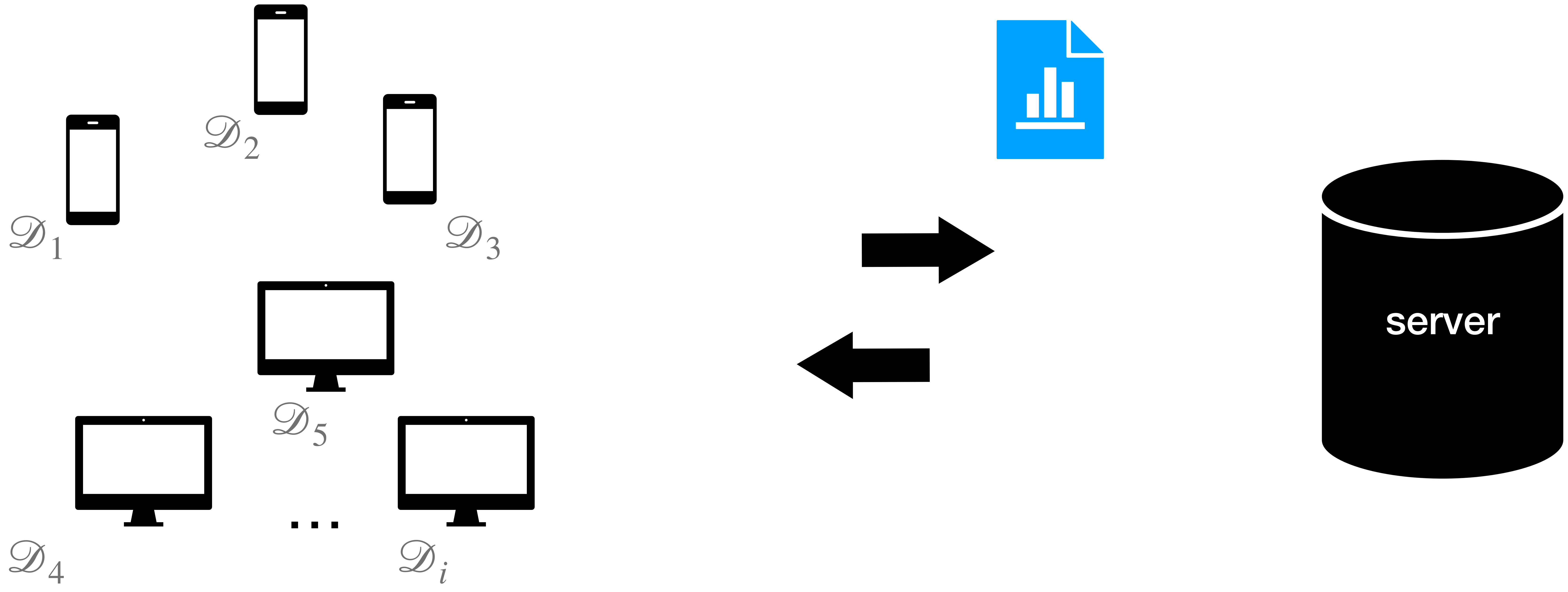
Server Distillation

A dedicated server model is distilled per cluster from the aggregated pseudo-labels.

Client refinement

Each client distills from its cluster teacher and trains from its own private data

Theoretical Analysis



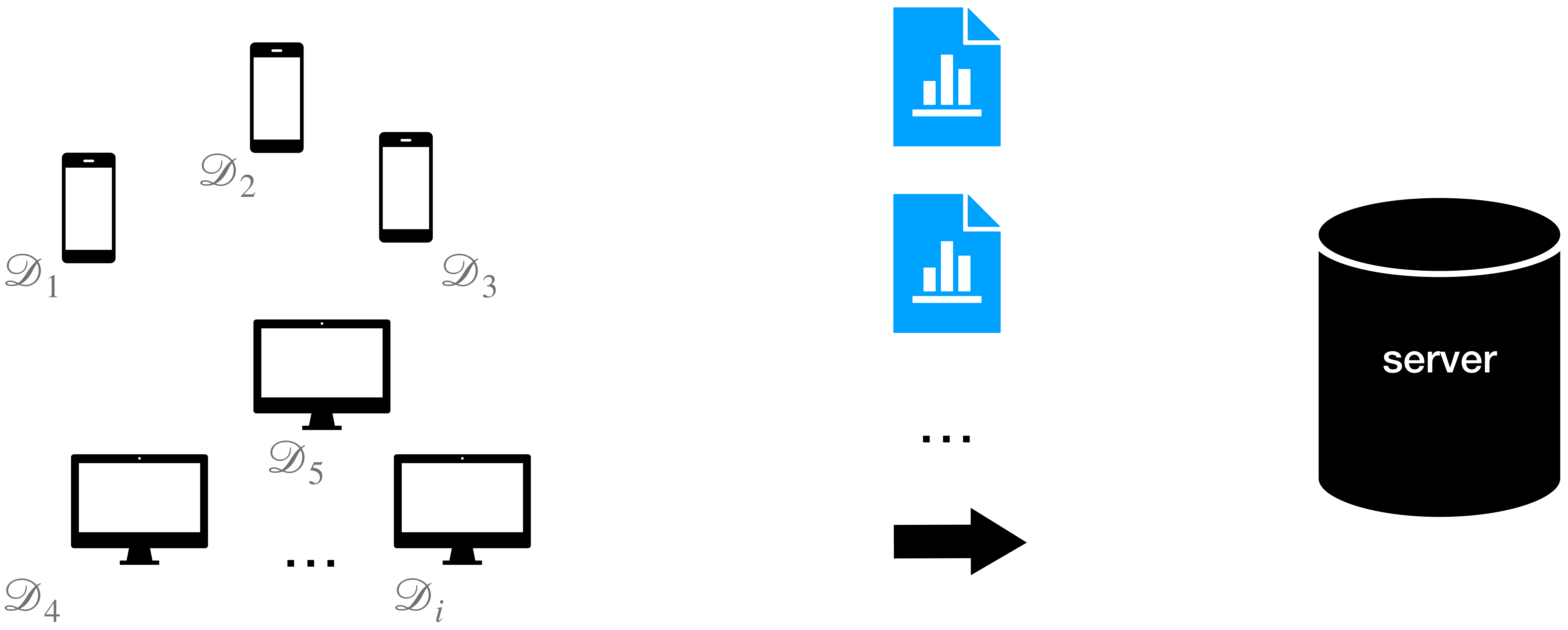
Theoretical Analysis

Condition 1: the public data has a good coverage.

For each client i with local data distribution $\mathscr{D}_i$ with density function p_i,
Let the global data's distribution to be $\mathscr{U}$ and $Q_i = \mathscr{U}_{|Supp(\mathscr{D}_i)}$,
$w_{i,1} = \sup_{x \in \mathcal{S}_i} \frac{p_i(x)}{q_i(x)}$ and $w_{i,2} = \sup_{x \in \mathcal{S}_i} \frac{q_i(x)}{p_i(x)}$ be bounded.

Since the public data is unlabeled, it should be easy to gather or generate.

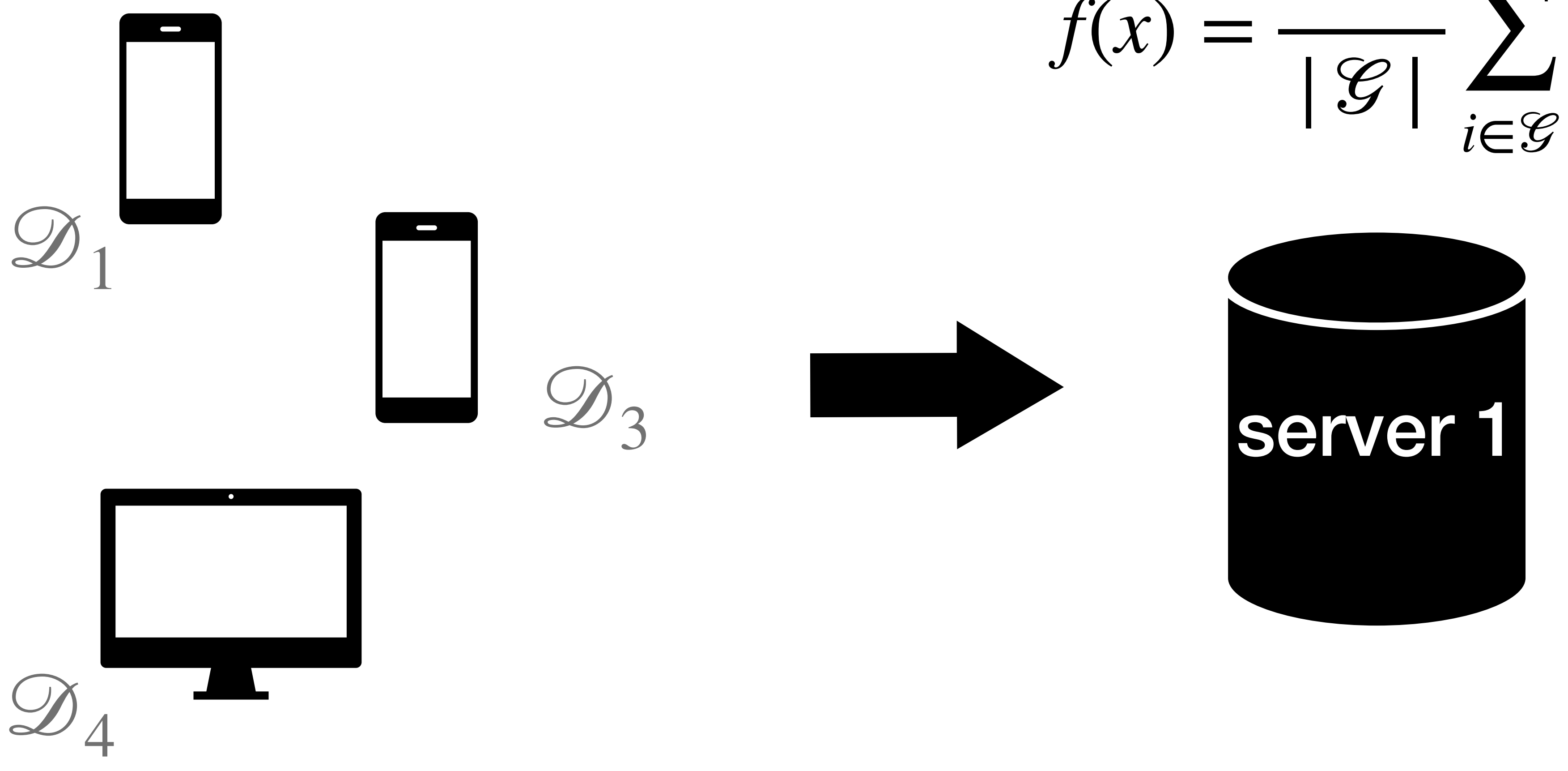
Theoretical Analysis



Theoretical Analysis

Condition 2: Clustering gives controlled disagreement.

$$\bar{f}(x) = \frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} f_i(x)$$



On the datapoints each client cares about,

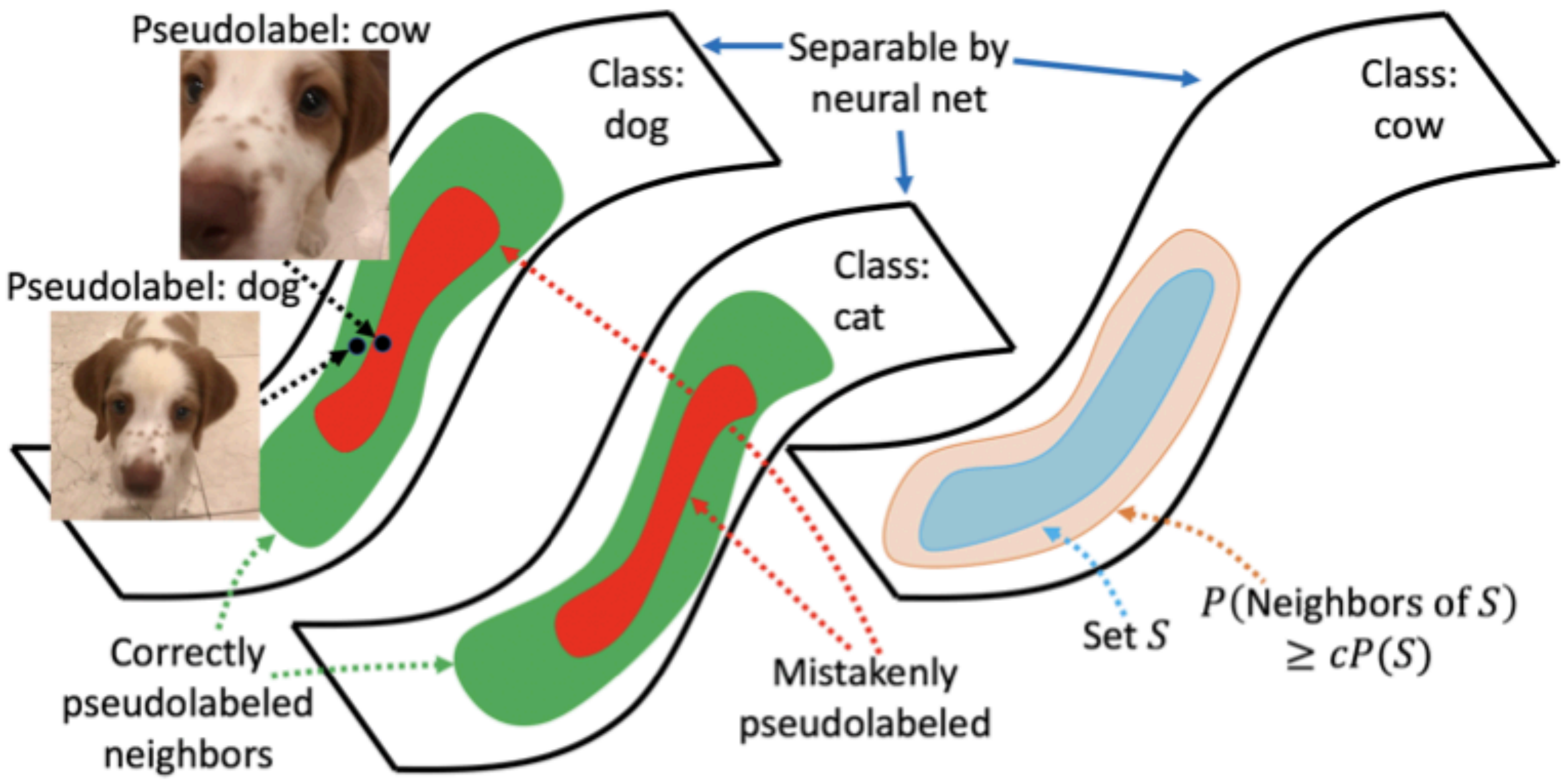
teacher error $\lessapprox$ **client error** + C_B

C_B is a constant related to max in-cluster distance B

This shows that we should be cautious about using common clustering methods like k-means.

Theoretical Analysis

Condition 3: Client's correctness; Connectivity and Robustness



If a incorrectly-labeled datapoint

is surrounded by correctly-labeled datapoints,

(Correctness+ Connectivity)

and if the model is *robust*,

the model can learn well and possibly even flip

the wrong label.

This leads to the desired **weak** (client) **-to-strong** (server) generalization.

Theoretical Analysis of Self-Training with Deep Networks on Unlabeled Data Wei et al. 2022

II. Iterative federated fine-tuning

robustness regularizer

$$r_{\mathcal{B}}(h)(x) := \sum_{x' \in \mathcal{B}} \left\{ \ell(h(x'), h(x)) \right\}$$

server model

Cluster Distillation

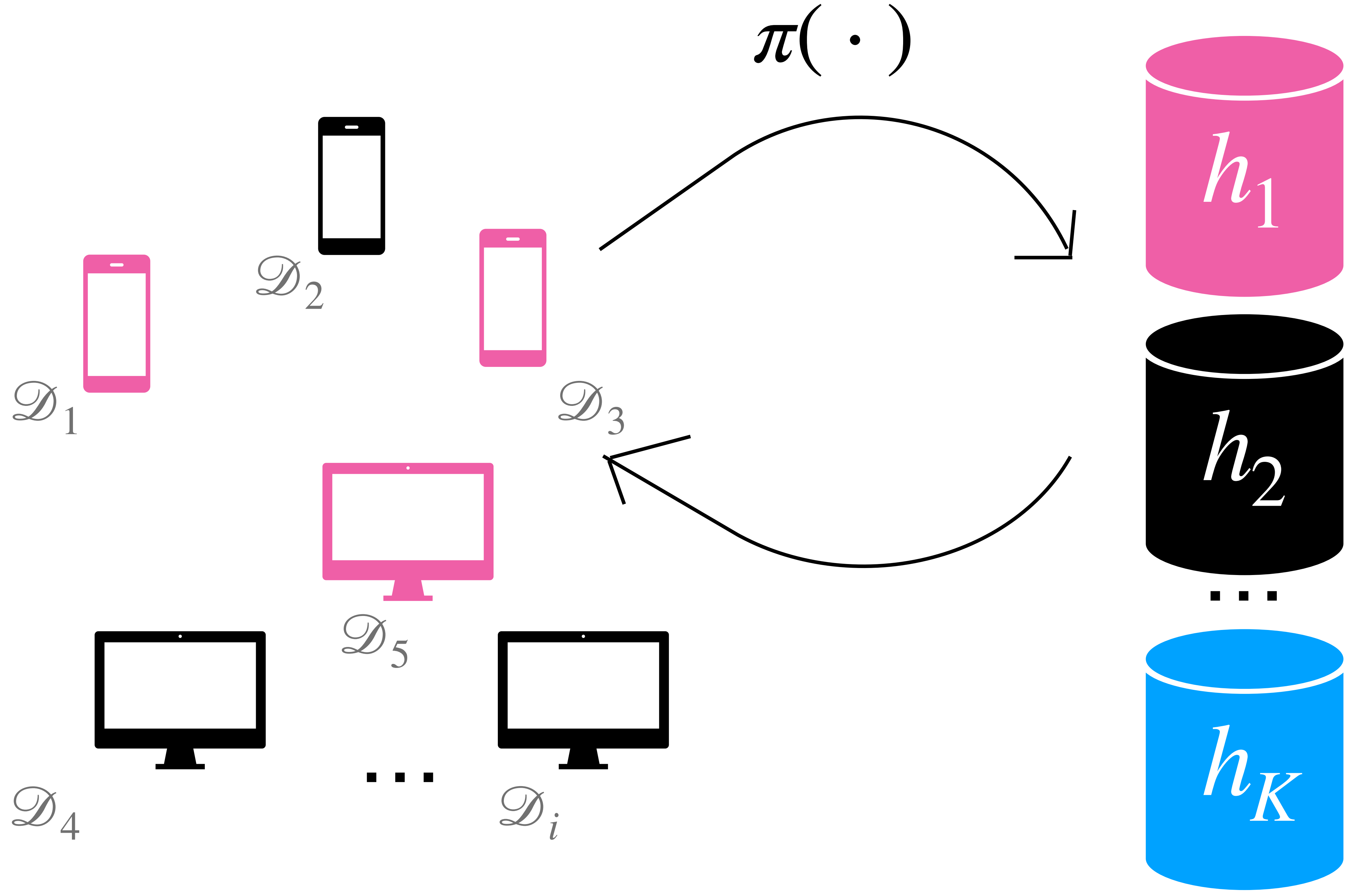
$$J(h) = \sum_{x \in U} \left[\ell(h(x), \bar{f}(x)) + \lambda \, r_{\mathcal{B}}(h)(x) \right]$$

aggregated client labels

Client refinement

$$F(f) = \sum_{x \in U} w_i(x)[\ell(f(x), h(x)) + \lambda \, r_{\mathcal{B}}(f_i)(x)]$$

Main Theoretical Result



Server teacher improves client signal
teacher risk ≤ κ_1 · client risk $\kappa_1 < 1$

Client distillation preserves improvement
next client risk ≤ κ_2 · previous risk $\kappa_2 < 1$

Personalization risk

$$R^{\dagger}(h_1, \ldots, h_K) = \frac{1}{N} \sum_{i=1}^{N} R_{\mathscr{D}_i}(h_{\pi(i)})$$

decays geometrically up to finite-sample error

Experimental Setup

Data heterogeneity

Split CIFAR100 classes into 5 disjoint label groups.
Each client is assigned to one group, then class samples are distributed using a Dirichlet draw (α)
10% of each client's data is shuffled, and redistributed across all clients.

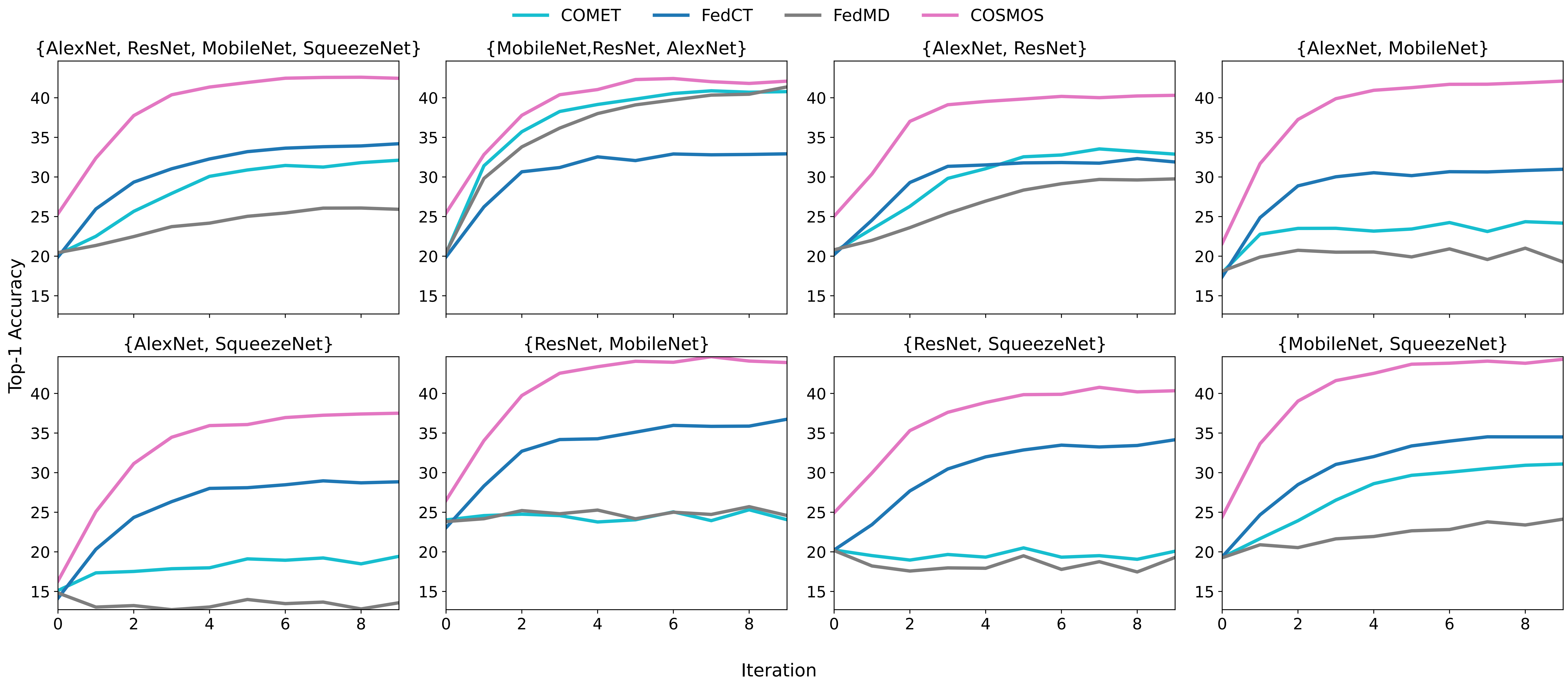
Model heterogeneity

Clients use different architectures
AlexNet / ResNet18 / MobileNet / SqueezeNet

The server uses a stronger VGG model

Experimental Result 1

Our method, COSMOS, consistently beats all the existing model-agnostic baselines.



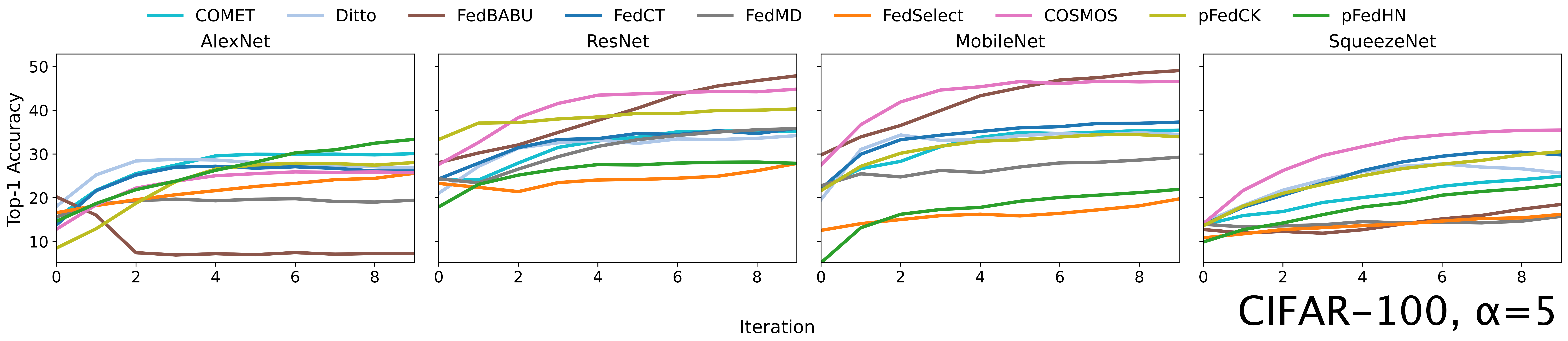
CIFAR-100, α=5

Experimental Result 2

Our method, COSMOS, remains competitive with substantially more communication-intensive baselines that require model homogeneity.

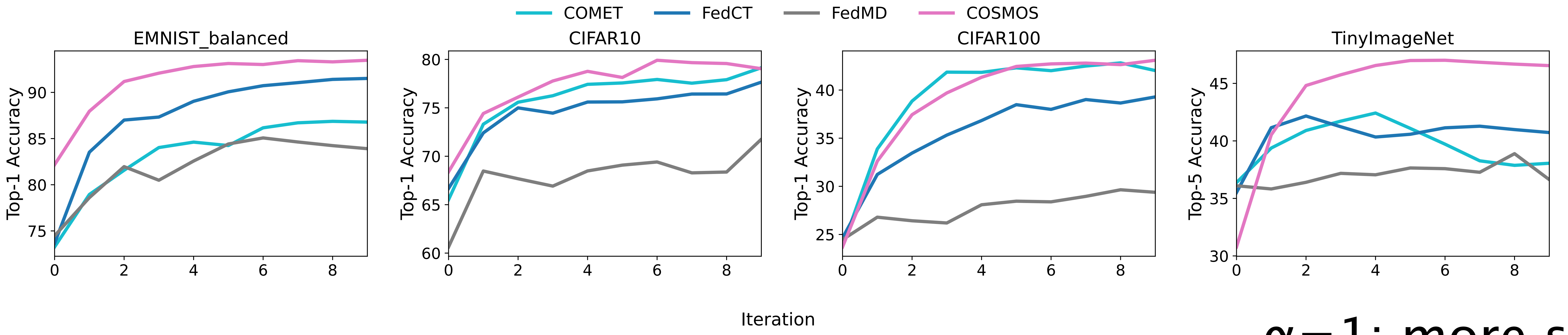
Table 4. Comparison of Baselines by Communication Type (Bandwidth Measured Per Client in MB).

Dataset	Parameters	Pseudo-labels
CIFAR-100	114.76	3.81
CIFAR-10	113.36	0.38
EMNIST-balanced	113.93	4.04
Tiny ImageNet	116.32	15.26

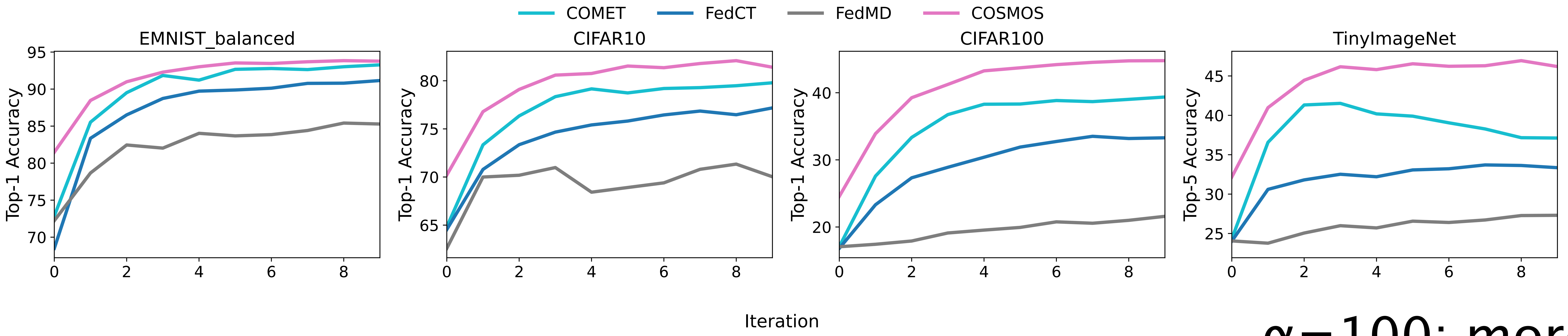


Ablation Result 1

Our method, COSMOS, is robust to different data distributions.



α=1: more skewed



α=100: more balanced

Summary

COSMOS is the first model-agnostic personalized FL algorithm with end-to-end personalization risk contraction guarantee.

Pseudo-label-only communication enables model-agnostic FL while reducing communication.

Clustered server models address statistical heterogeneity through personalized teachers.

Future Work

More Modalities beyond image classification

Dynamic clustering over time

Adaptive public-data generation

Thank you!

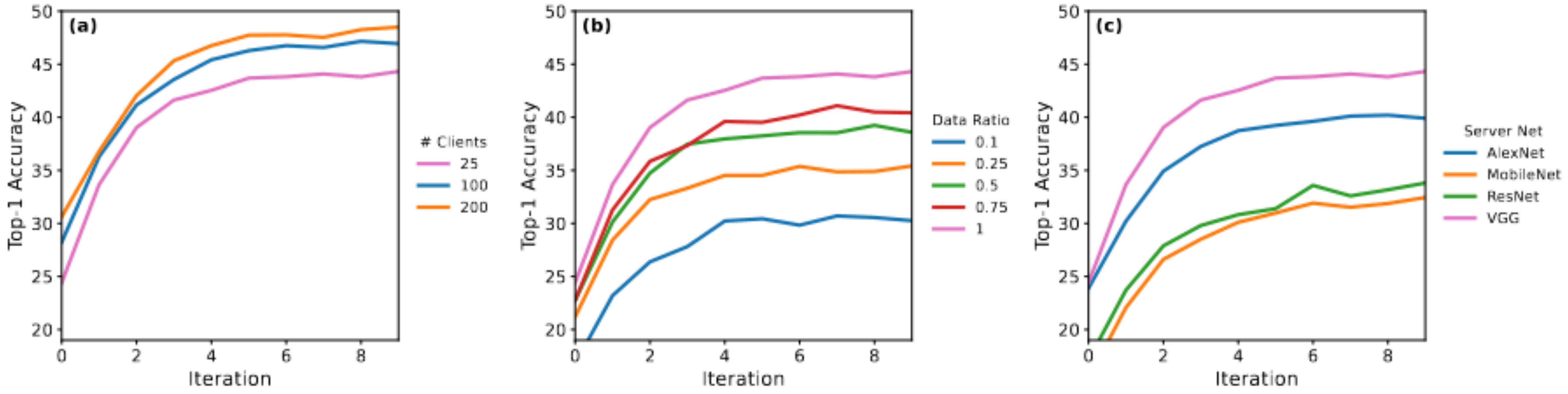


Fig. 7. Top-1 accuracy over communication rounds for COSMOS on CIFAR-100 with Dirichlet non-IID sampling ($\alpha = 5$): (a) varying numbers of participating clients, (b) different ratios (fractions) of unlabeled global data, and (c) different server-side architectures. Client models are MobileNet and SqueezeNet.