Samsung Research

CVPR VIRTUAL JUNE 19-25

Are All Users Treated Fairly in Federated Learning Systems?

Umberto Michieli and Mete Ozay

SAMSUNG

Università degli Studi di Padova

LTTM

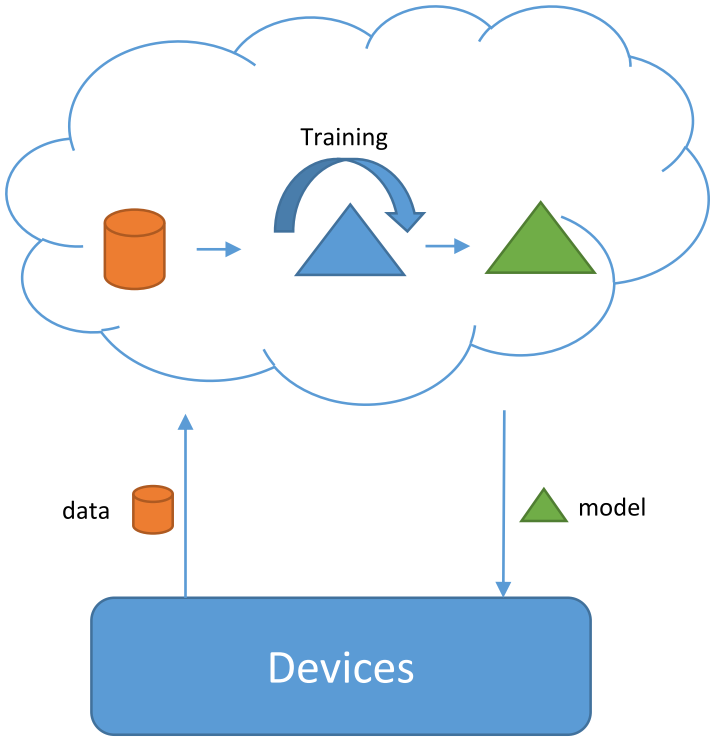
Overview

- Background
- Targeted Challenges and Motivation
- Our Method (FairAvg)
- Results
- Conclusions

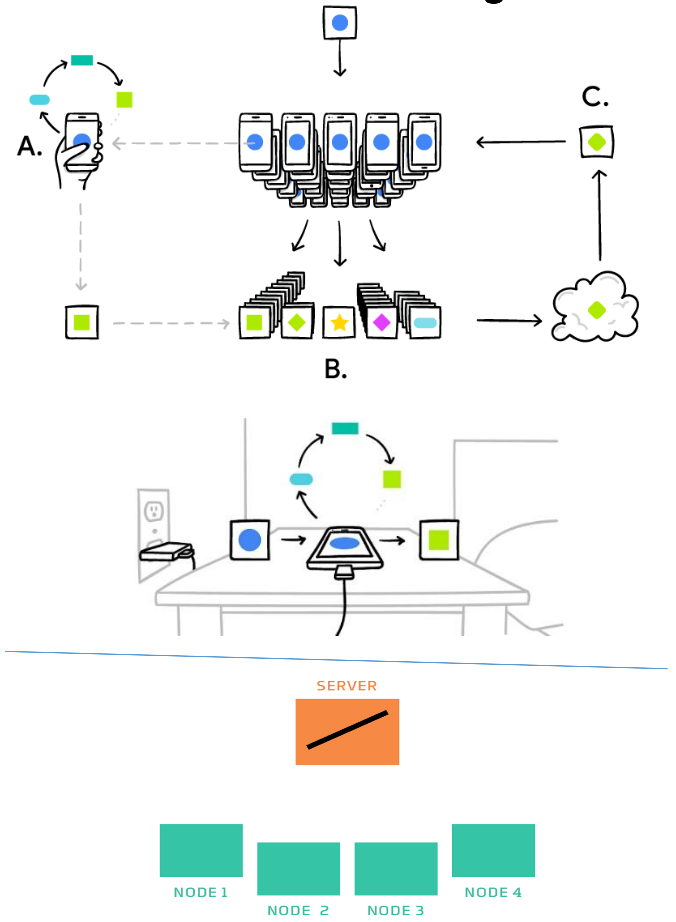
Background

- Federated Learning: Distributed Machine Learning on **Heterogeneous Data**

Traditional (centralized) Learning



Federated Learning



Benefits

- Ensuring Privacy
- No data breach risk
- Lower server costs
- Multiple modalities
- Data availability

Background

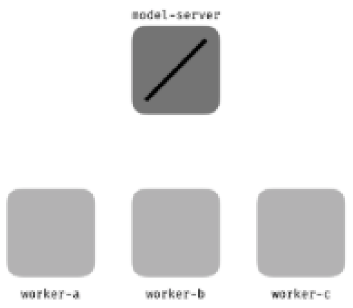
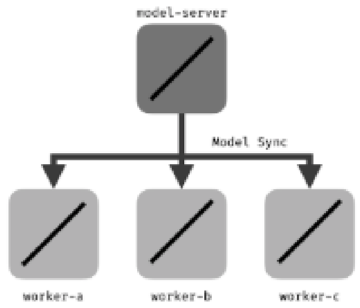
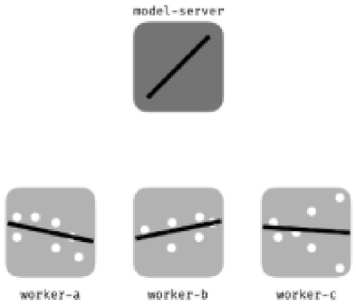
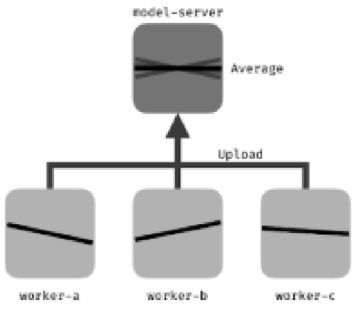
Differences between Distributed Learning and Federated Learning

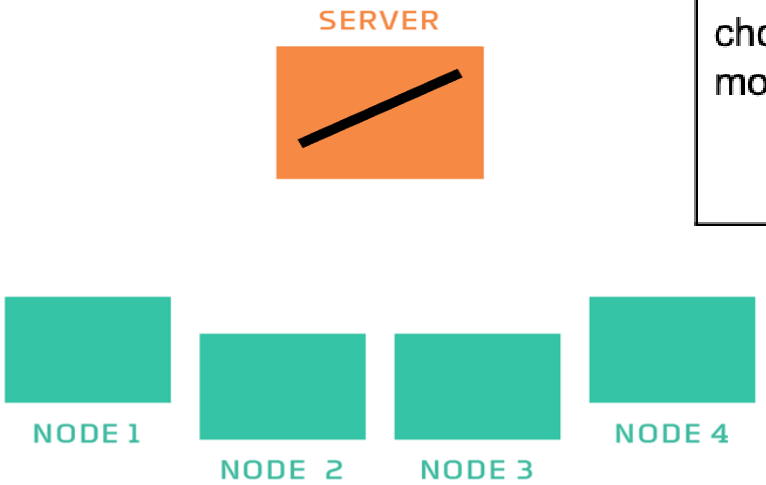
Distributed Learning	Federated Learning
Both aim at training a single model on multiple nodes	
Focus: Parallelizing computing power	Focus: Training on heterogeneous datasets
❖ Distributed data on each client has **roughly** the **same size**	❖ Distributed data on each client has **not same size** (e.g., powerlaw)
❖ Data are distributed **i.i.d.**	❖ Data are distributed **non-i.i.d.**
❖ Nodes are typically (**reliable**) datacenters	❖ Clients may be **unreliable** (low battery, WiFi, etc)

Background

- Our focus will be on the *classical* **Centralized Federated Learning**

Coordination of a central server

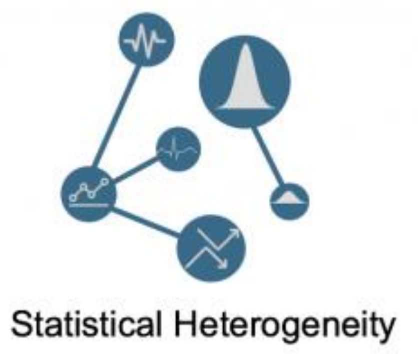
Step 1	Step 2	Step 3	Step 4
			
Central server chooses a statistical model to be trained	Central server transmits the initial model to several nodes	Nodes train the model locally with their own data	Central server pools model results and generate one global mode without accessing any data



Targeted Challenges

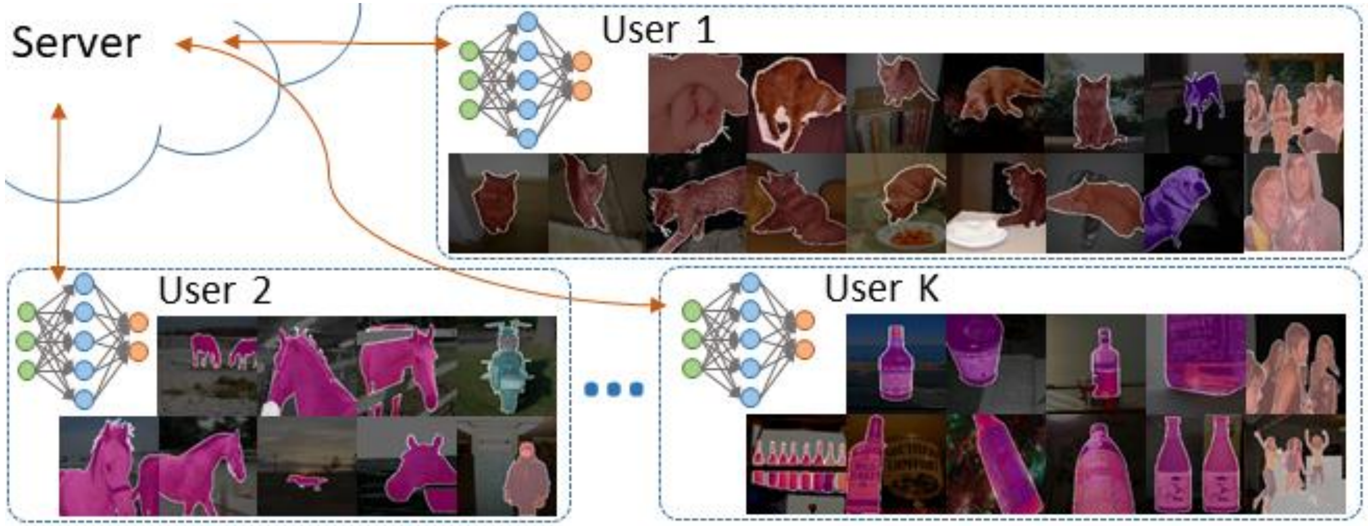
Samsung Research

FL is a complicate task and many challenges exist [1,2].



Our focus will be on:

- Statistical (data) heterogeneity
  - data is **highly imbalanced**: different number of samples for different classes on each device
  - data is **highly non-iid**: samples in remote clients correlated due to specific user habits or preference



[1] Bonawitz, Keith, et al. "Towards federated learning at scale: System design." *arXiv preprint arXiv:1902.01046* (2019).
[2] Li, Tian, et al. "Federated learning: Challenges, methods, and future directions." *IEEE Signal Processing Magazine* 37.3 (2020): 50-60.

FairAvg (CVPRW)

In general, federated aggregation can be expressed as:

clients

Central model parameters

$$\boxed{\boldsymbol{w}_{t+1}} \leftarrow \sum_{k=1}^{\boxed{K}} \boxed{\boldsymbol{w}_t^k}\ \boldsymbol{a}_k^t$$

Attention vector

Local model parameters

samples of participant k

❖ FedAvg [1]: $\boldsymbol{a}_k = \frac{\boxed{n_k}}{\boxed{n}}$

samples of all participants

❖ FairAvg [2]: $\boldsymbol{a}_k^t = 1/K$ (constant)

→ To show effect of data imbalance and non-iid-ness across clients

[1] McMahan B., et al. "Communication-efficient learning of deep networks from decentralized data." *Artificial Intelligence and Statistics*. PMLR, 2017.
[2] Michieli U. and Ozay M., "Are All Users Treated Fairly in Federated Learning Systems?", CVPRW RCV 2021

Federated Datasets

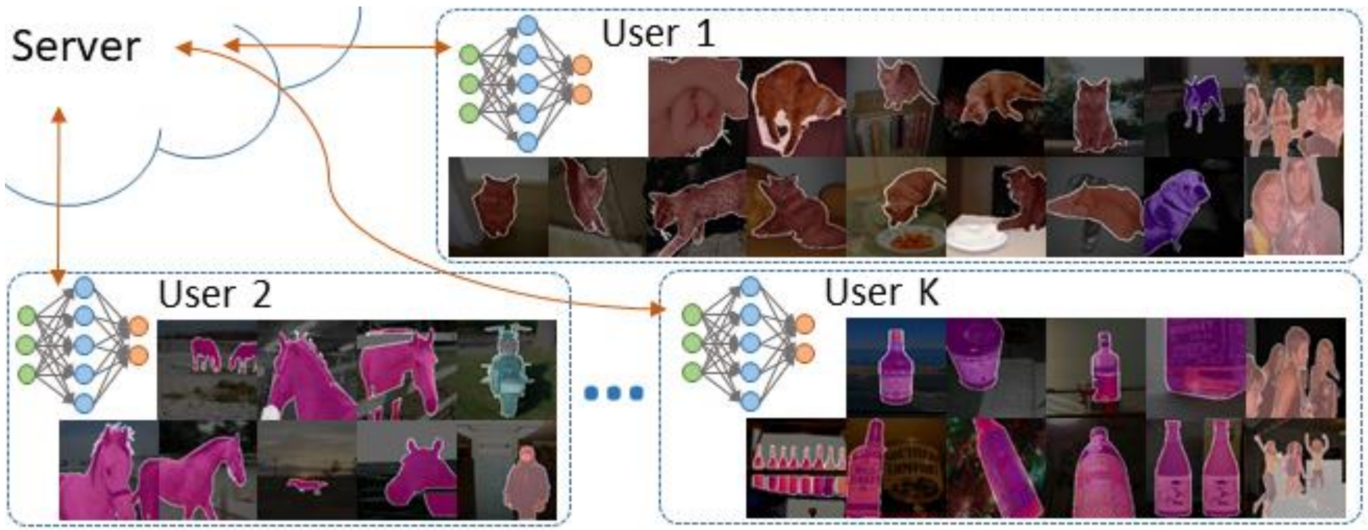
Samsung Research

Table 1. Statistics of the employed datasets (left) and hyper-parameters (right).

Dataset	# Classes	Clients	Samples	Samples/Client Mean	Samples/Client Std.	Model	Distribution	Central. Acc. (%)	Start lr	Solver	F	Rounds	Batch size
Synthetic	10	30	9,600	320.0	1051.6	2 dense layers	Power-law	78.5	0.01	SGD	20	200	10
MNIST	10	1,000	61,676	61.7	164.7	2-layer CNN	Power-law	99.0	0.01	SGD	20	200	10
FEMNIST	10	200	16,421	82.1	143.0	2-layer CNN	Power-law	99.0	0.001	SGD	20	400	10
Sent140	2	772	40,783	53	32	Stacked-LSTM	Power-law	72.3	0.3	SGD	20	800	10
Shakespeare	80	143	517,106	3,616	6,808	Stacked-LSTM	Power-law	49.9	0.8	SGD	20	40	10

- synthetic classification dataset
- real-world classification datasets

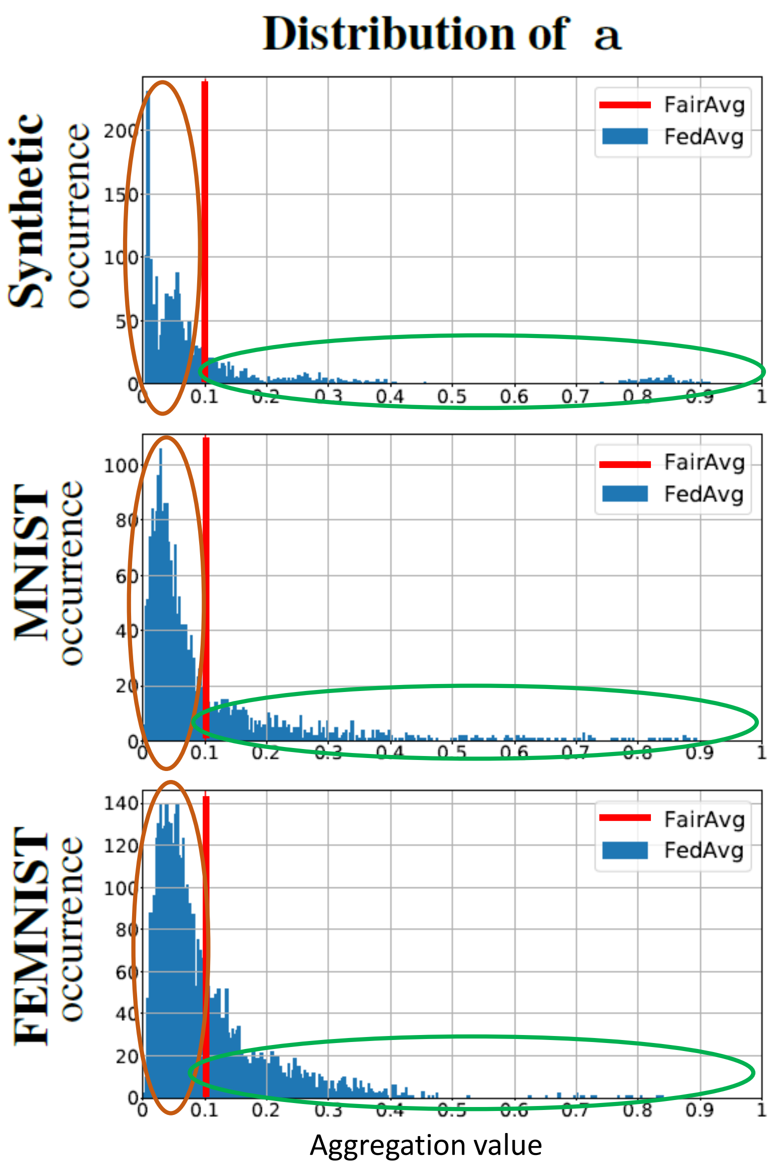
Samples are distributed to clients according to power-law distribution



Peter Thiel:
“We Don't Live In A Normal World;
We Live Under A Power Law.”

→ Occurring often and reasonable assumption

Federated Datasets



Samples are distributed to clients according to power-law distribution

Note1: we subsample K=10 clients
Note2: $\boldsymbol{a}$ computed by FedAvg is $\boldsymbol{a}_k \propto n_k$

- ❑ **Many clients** contribute little while aggregation
- ❑ **Few clients** tend to *dominate the scene*

→ If data is highly non-i.i.d. this represents a problem for convergence

\# clients

Attention vector

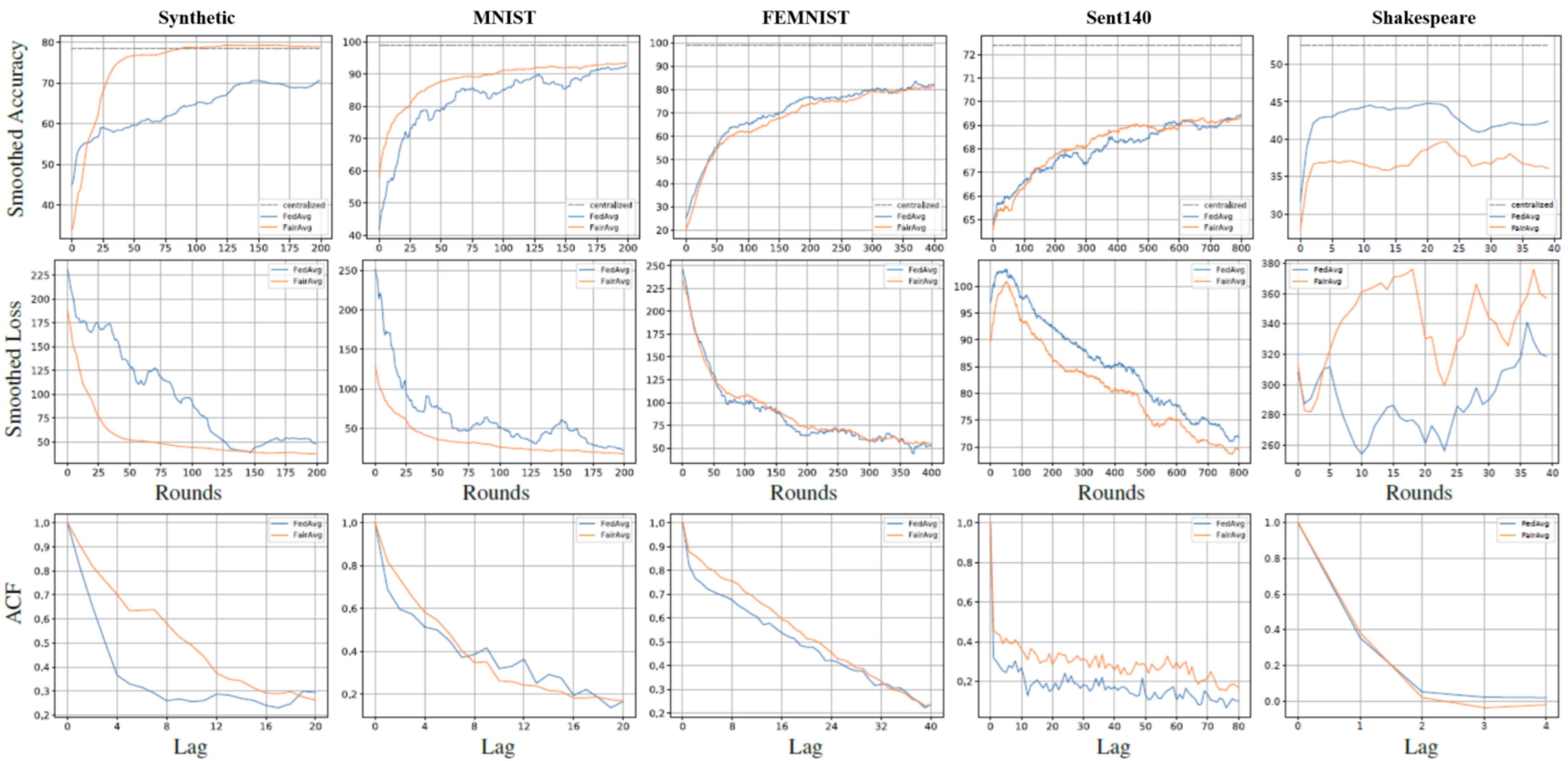
Central model parameters

$$\boxed{\boldsymbol{w}_{t+1}} \leftarrow \sum_{k=1}^{\boxed{K}} \boxed{\boldsymbol{w}_t^k} \; \boldsymbol{a}_k^t$$

Local model parameters

FairAvg (CVPRW)

Samsung Research

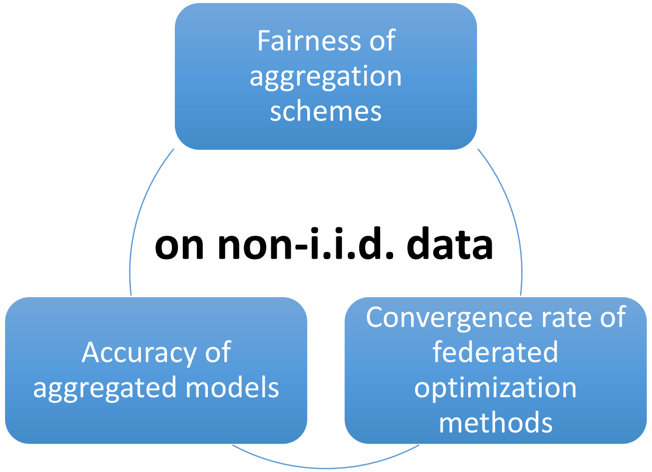


The higher the better (less fluctuations)

→ FairAvg improves accuracy wrt FedAvg (when data are distributed non-i.i.d.)

→ FairAvg reduces fluctuation towards convergence values wrt FedAvg (when data are distributed non-i.i.d.)

Summary



- ✓ **FairAvg** > FedAvg in accuracy, convergence rate, and fluctuations of accuracy of the aggregate model
- ✓ Future FL models could employ federated aggregation values centered around the value employed by FairAvg for more **uniform treatment of user contributions**

SAMSUNG

UNIVERSITÀ DEGLI STUDI DI PADOVA

LTTM