Robust Aggregation for Federated Learning

IEEE Transactions on Signal Processing (2022)

Krishna Pillutla
IIT Madras

ROBERT BOSCH CENTRE FOR DATA SCIENCE AND ARTIFICIAL INTELLIGENCE IIT MADRAS

CeRAI Centre for Responsible AI

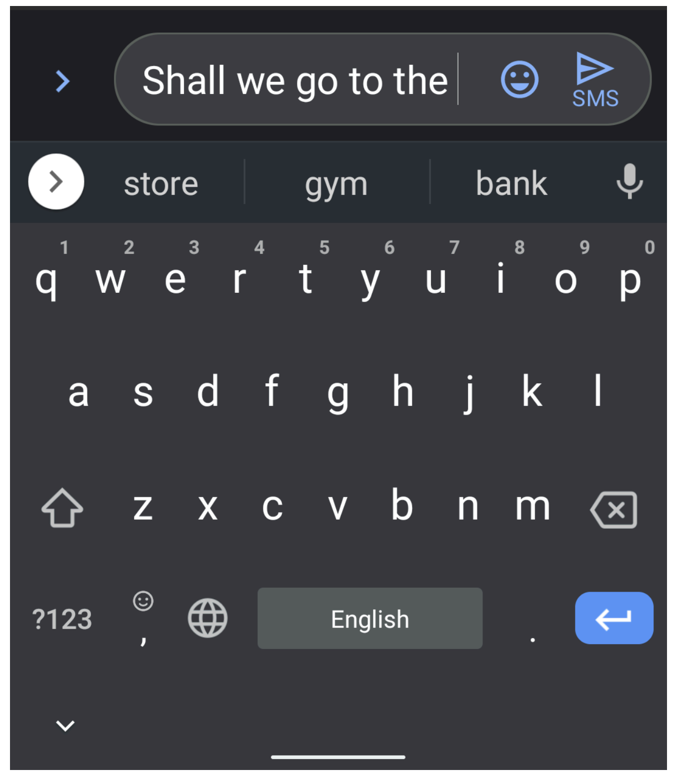
IEEE Signal Processing Society

Team

Krishna Pillutla

Sham Kakade

Zaid Harchaoui



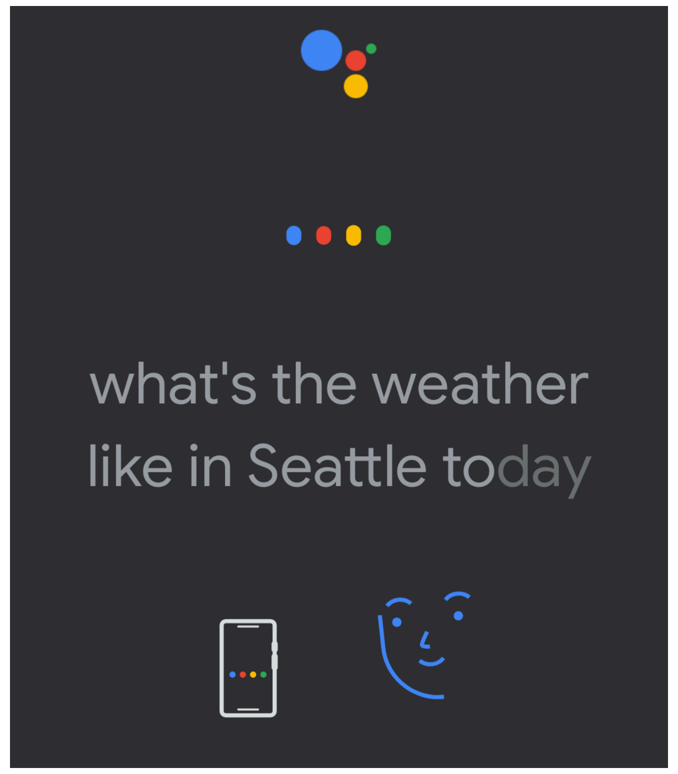
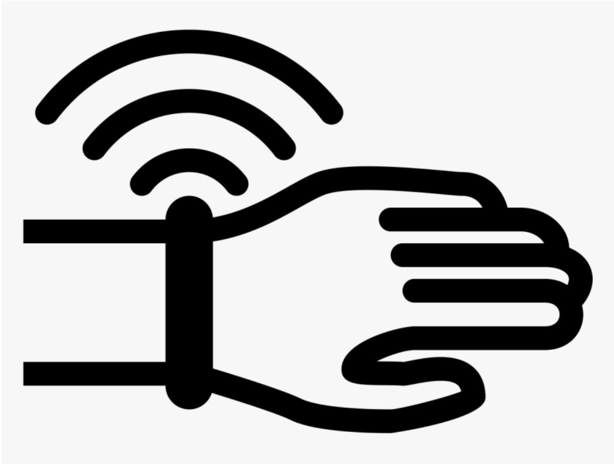
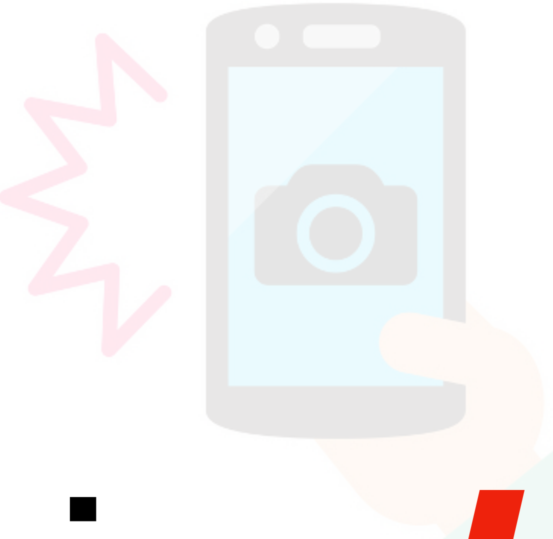


Image Credit: Robotics Business Review



Rieke et al. NPJ Digit. Med. (2020)   Image Credit: Wellcome

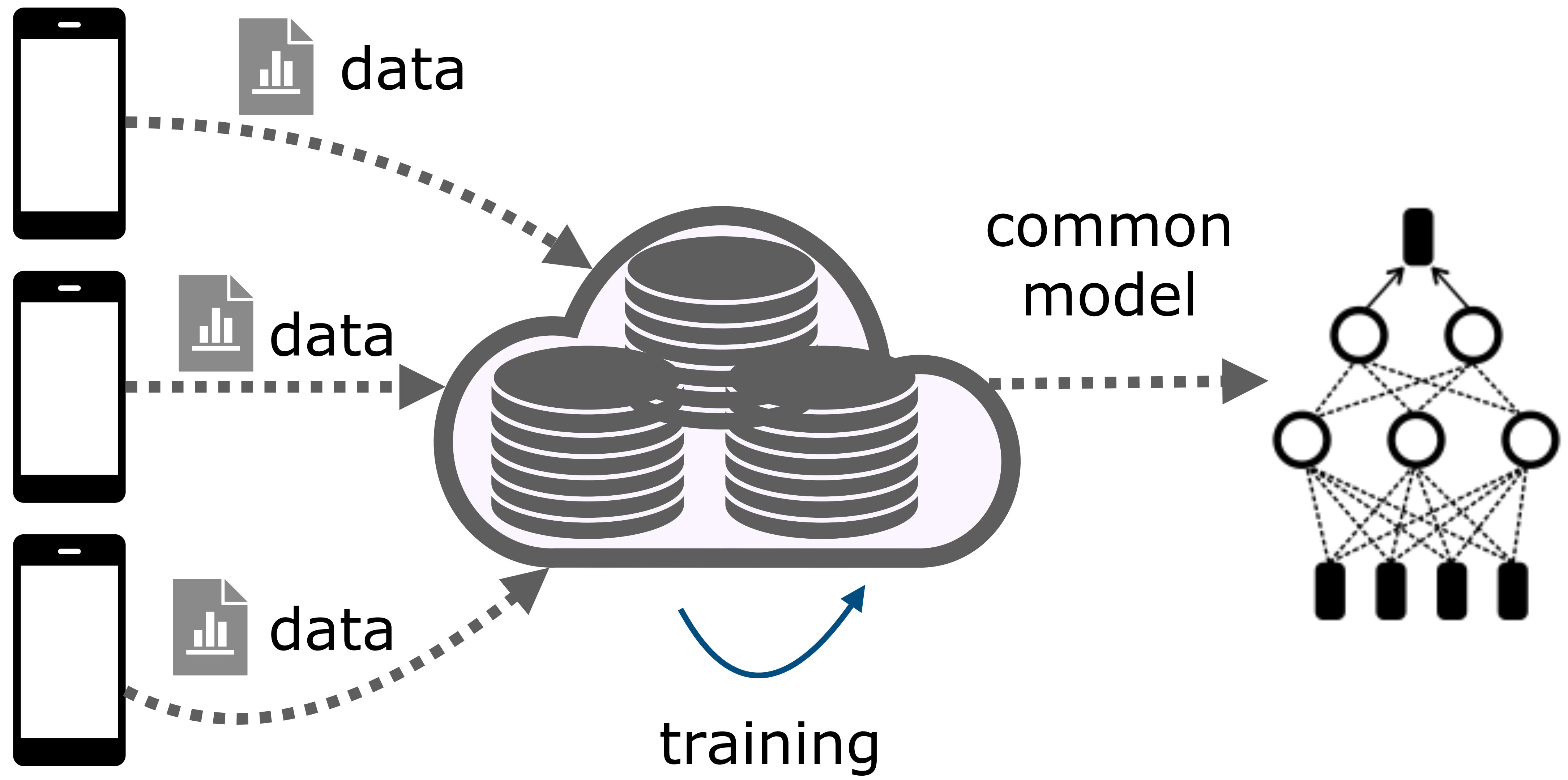
Image Credit: Robotics Business Review



Data is ***decentralized*** and ***private***

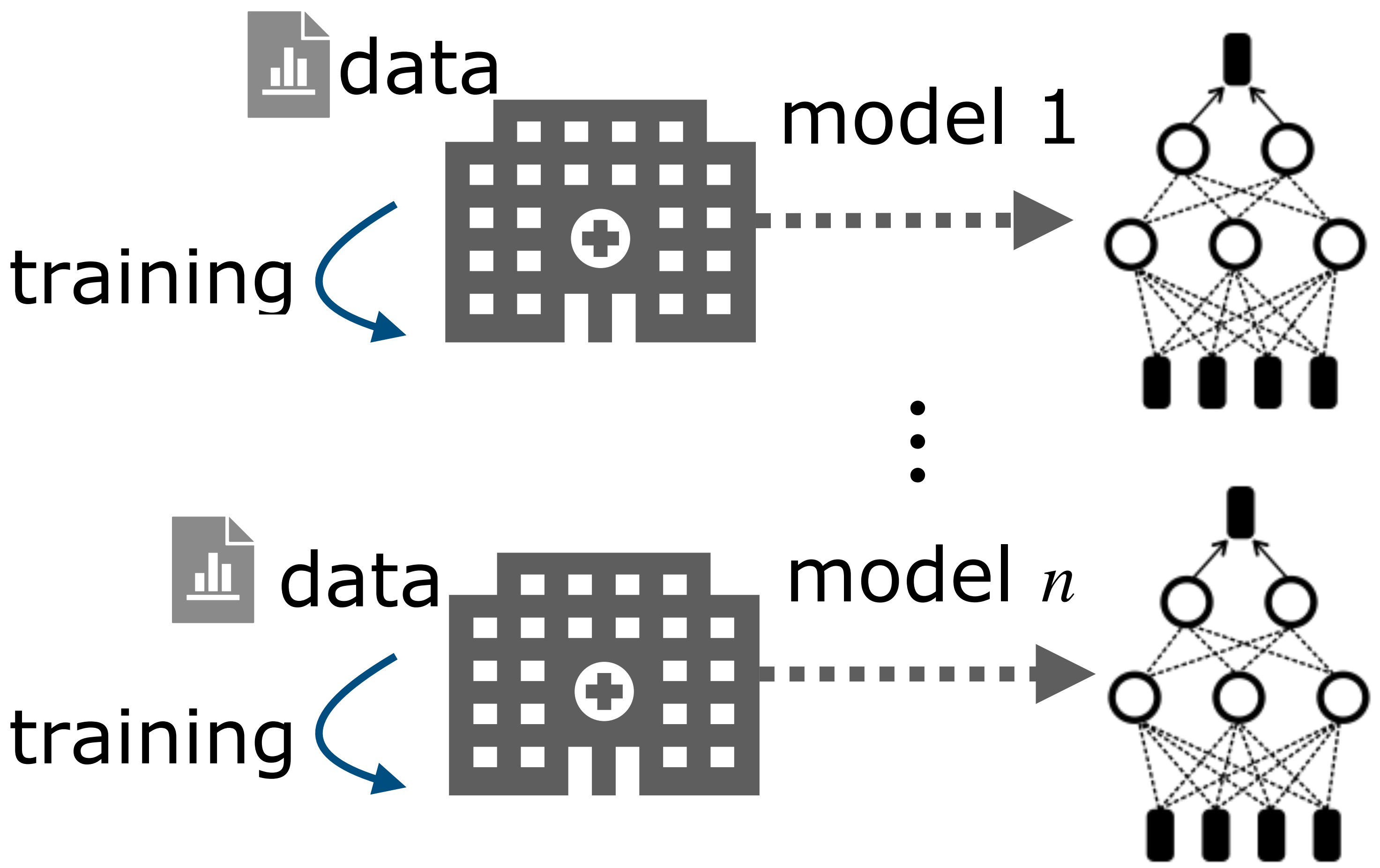

Rieke et al. NPJ Digit. Med. (2020)   Image Credit: Wellcome

Datacenter



Datacenter

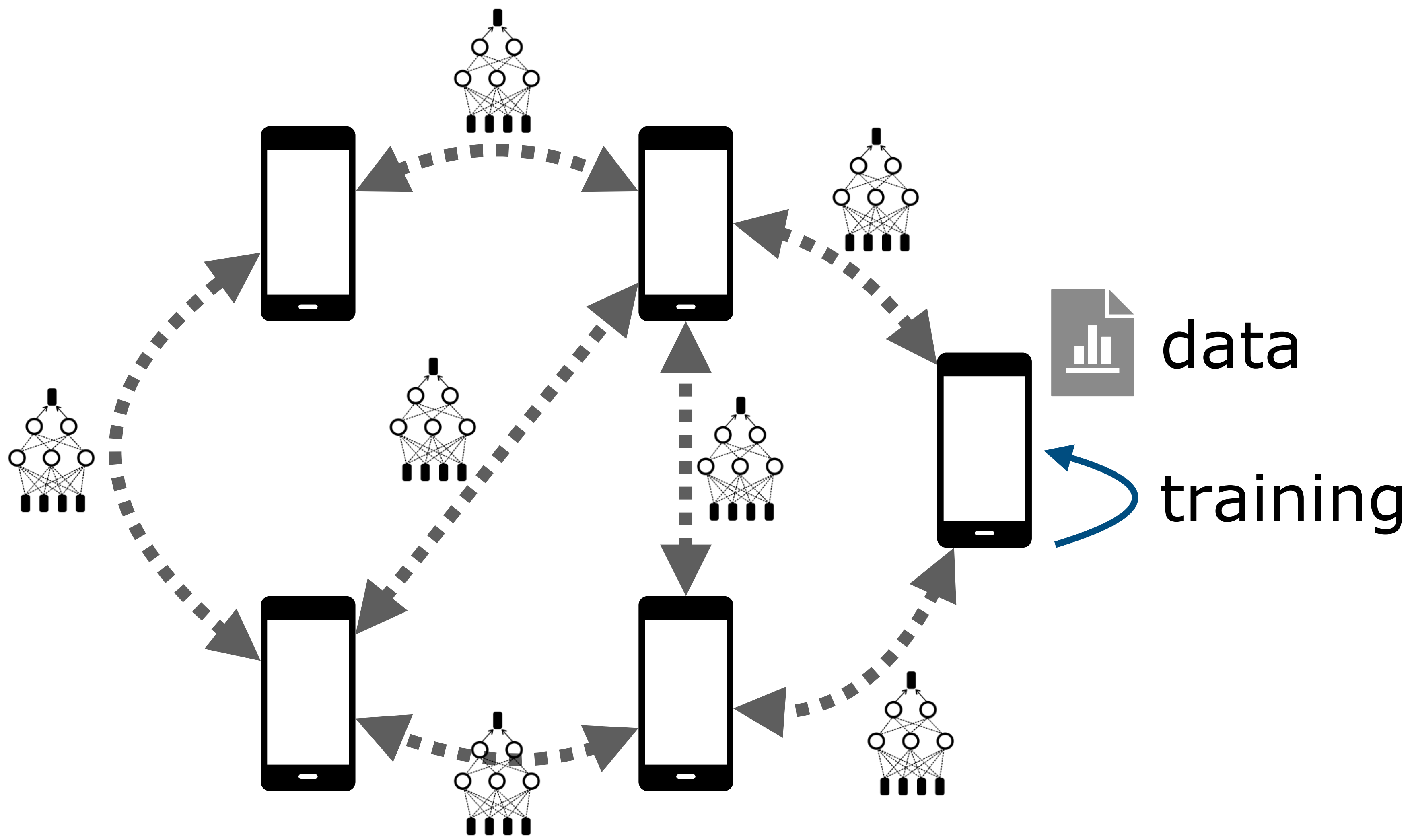
Non-collaborative



Datacenter

Non-collaborative

Peer-to-peer



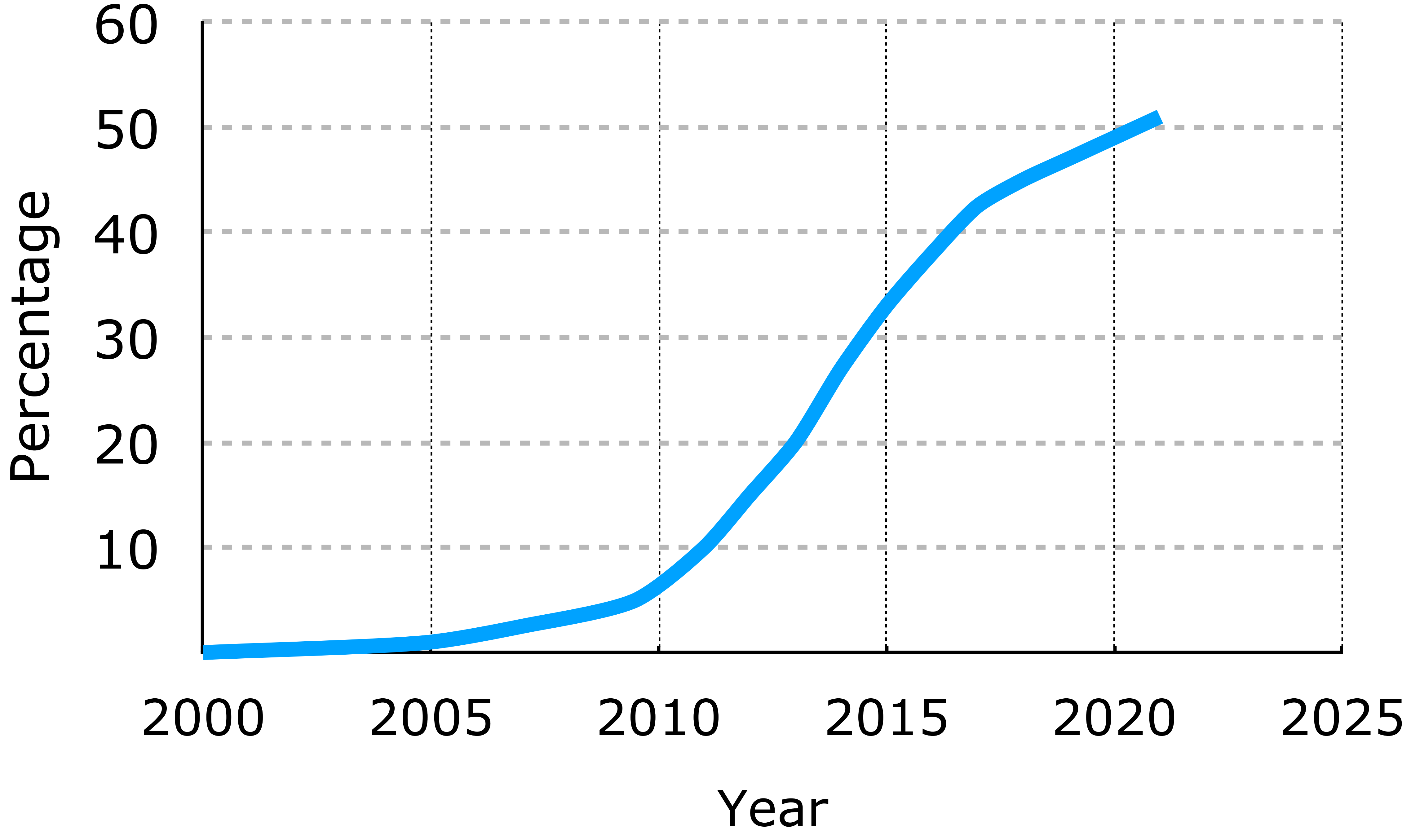
Advances and Open Problems in Federated Learning

Peter Kairouz
Google Research
Kairouz@google.com

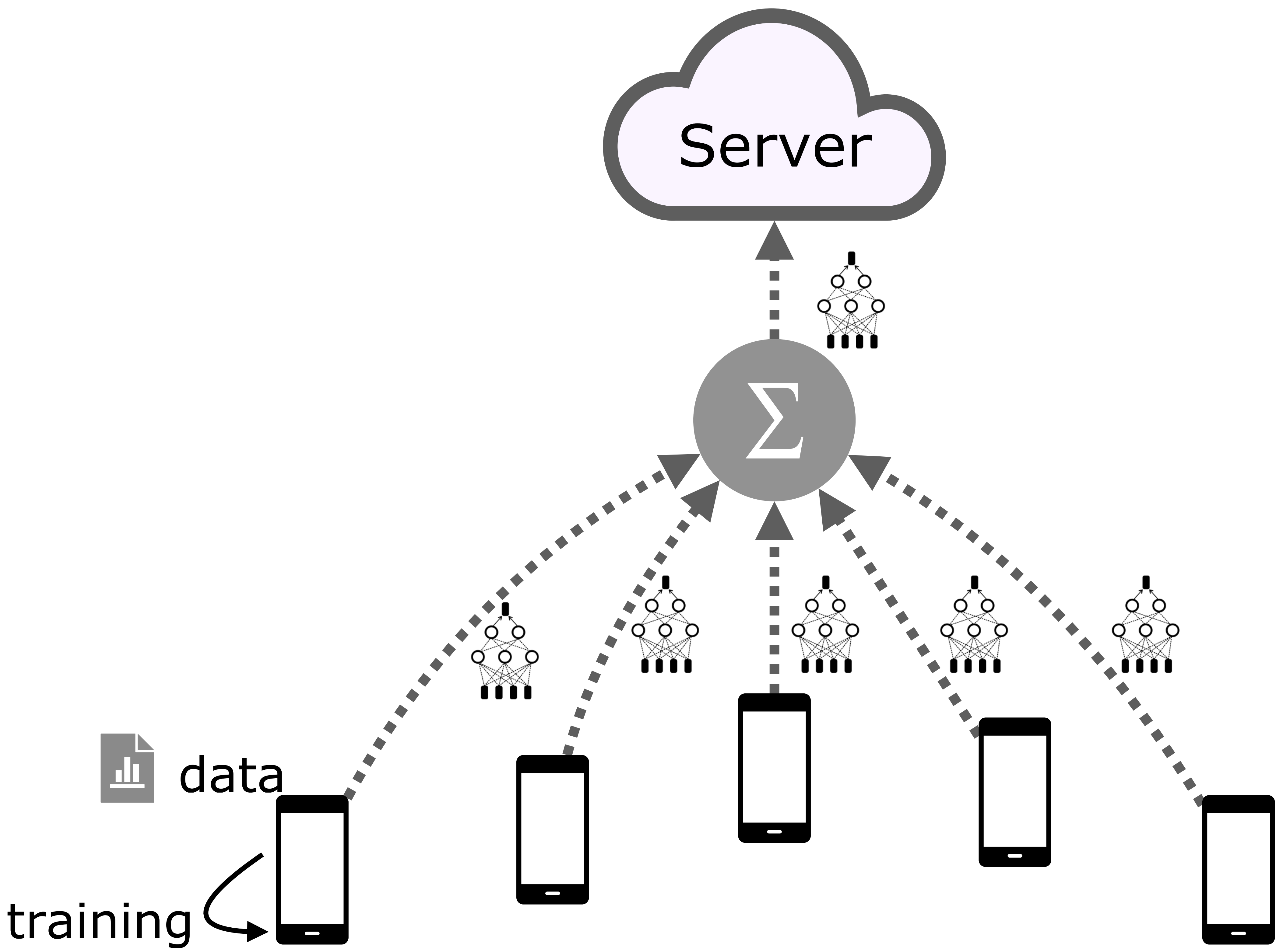
H. Brendan McMahan
Google Research

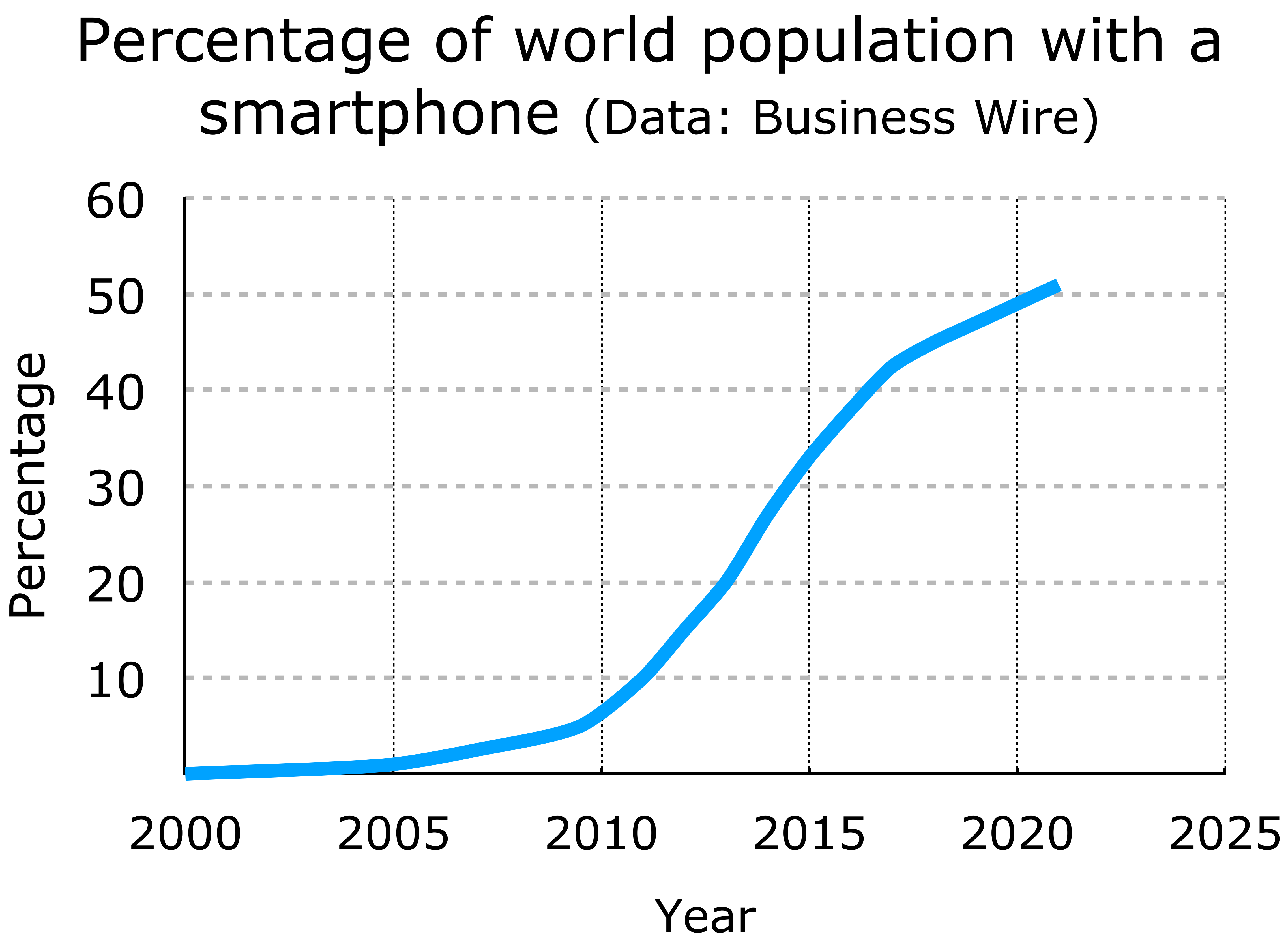
et al.

Percentage of world population with a smartphone (Data: Business Wire)

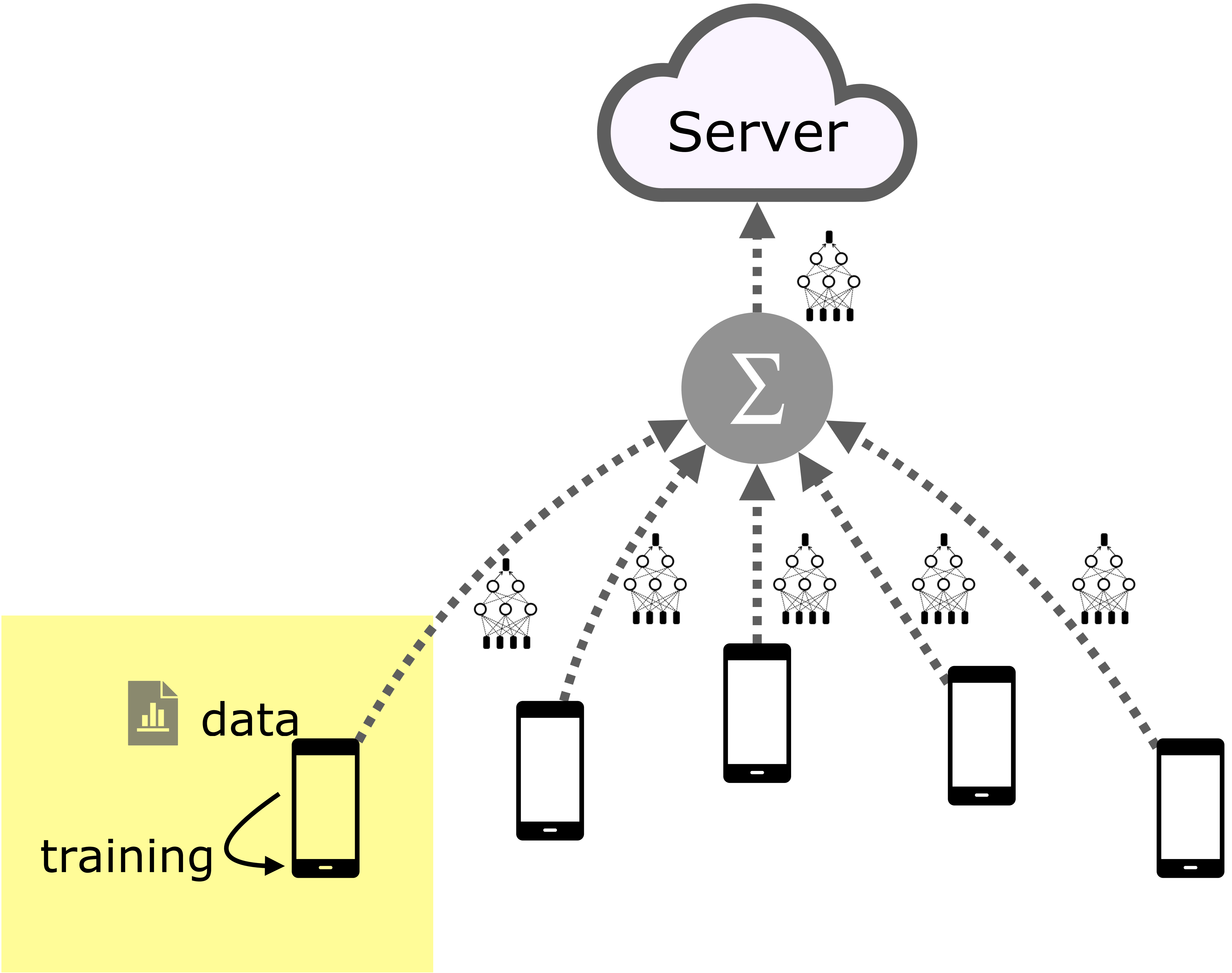


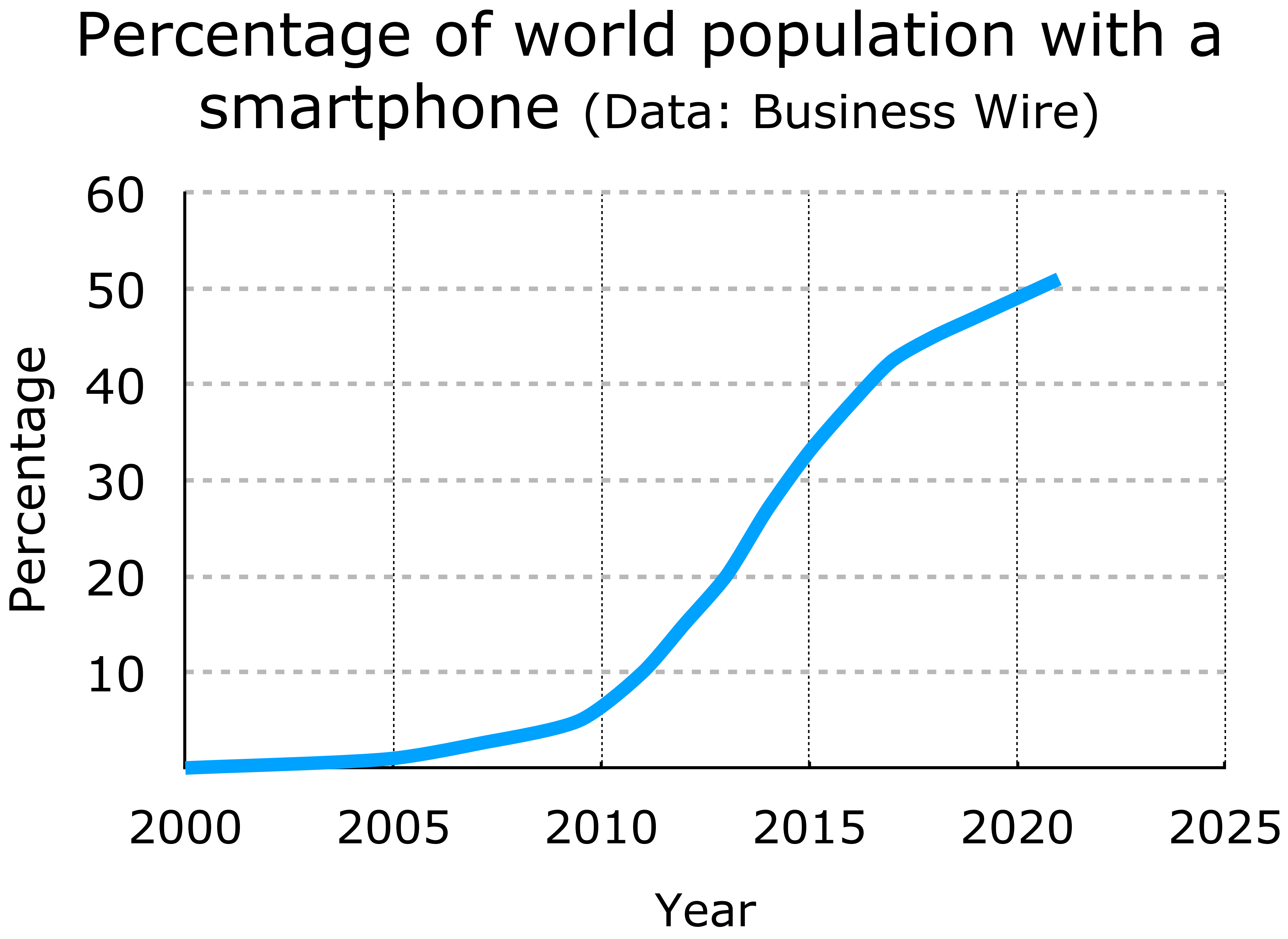
Federated Learning



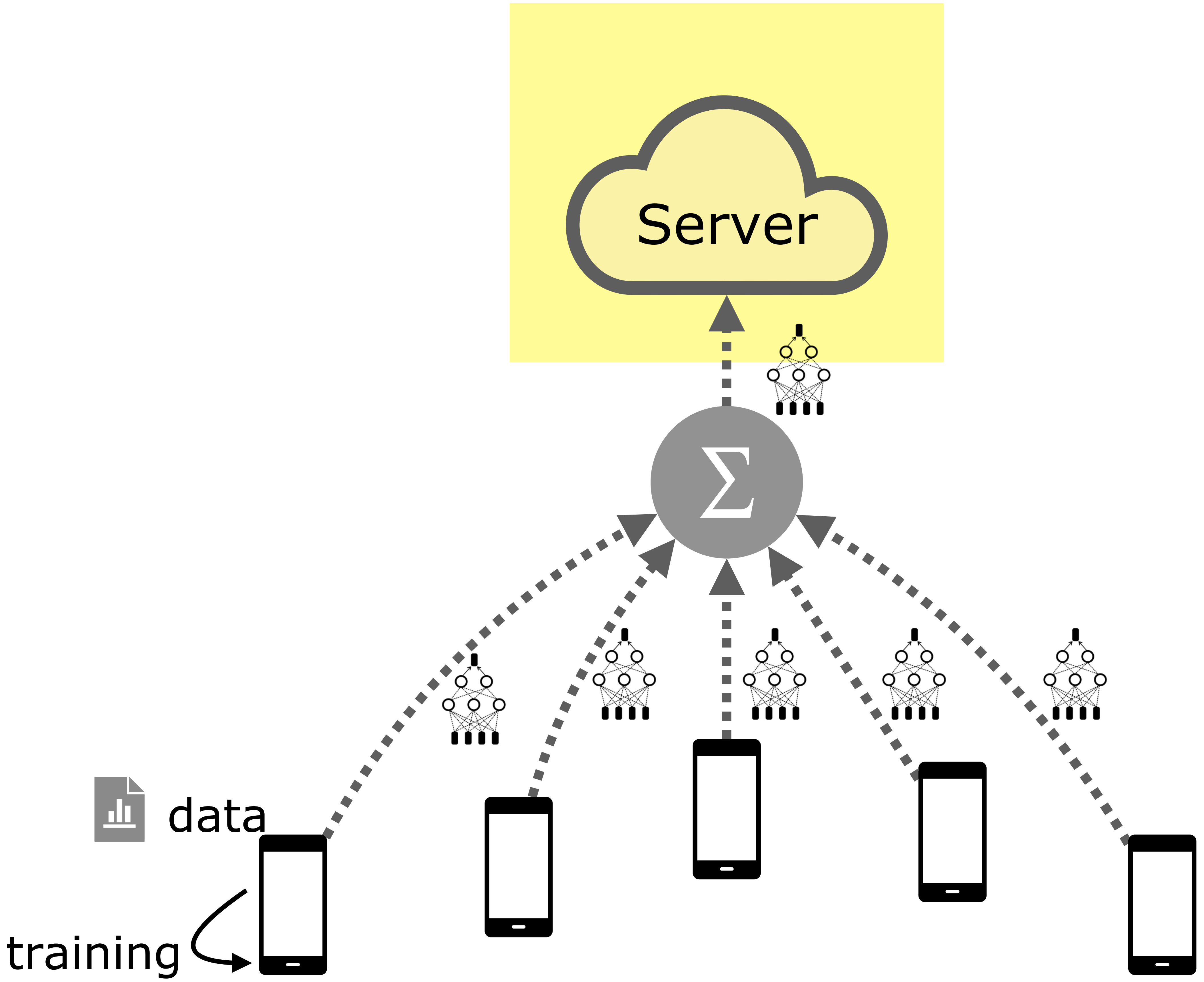


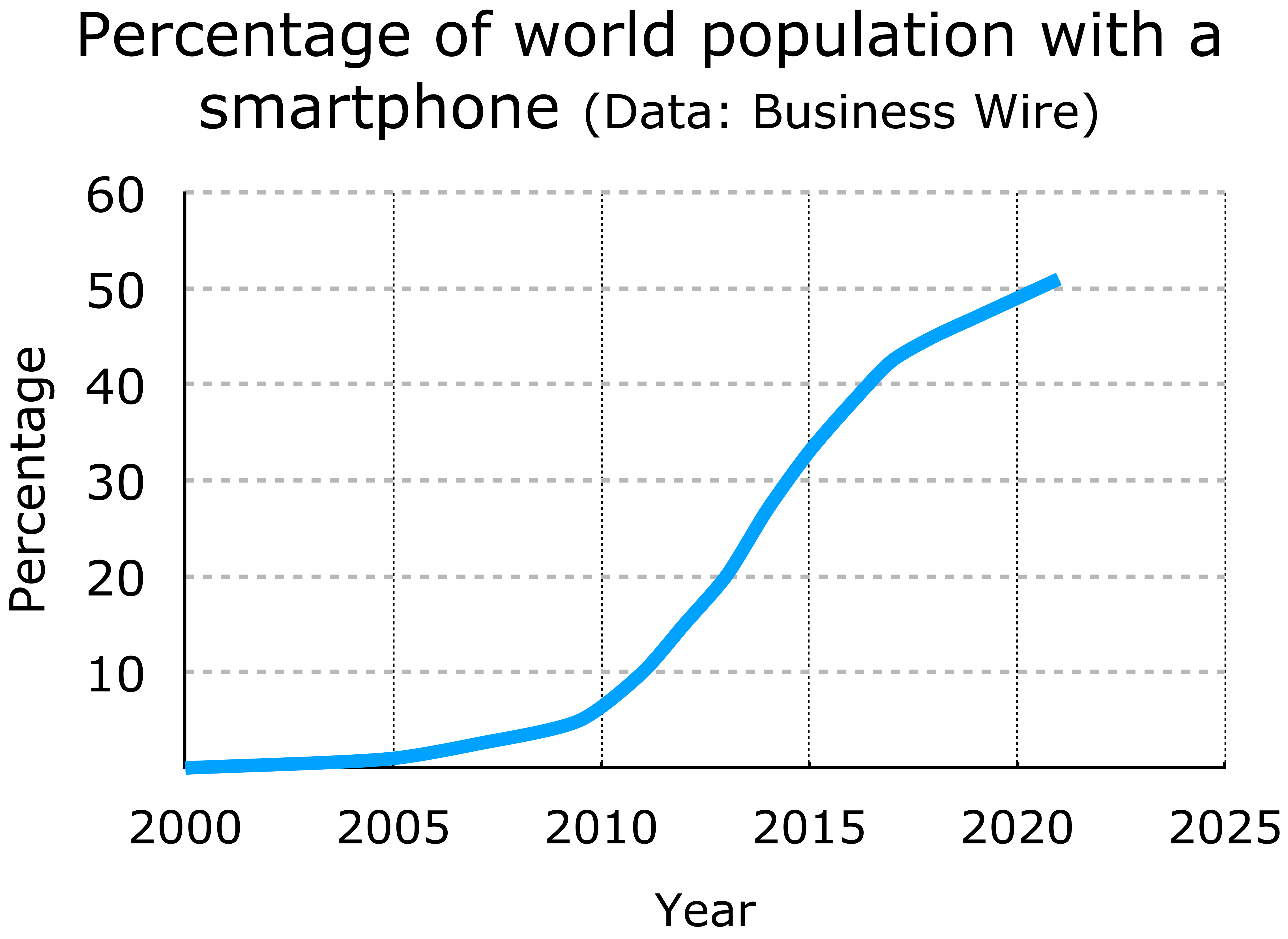
Federated Learning



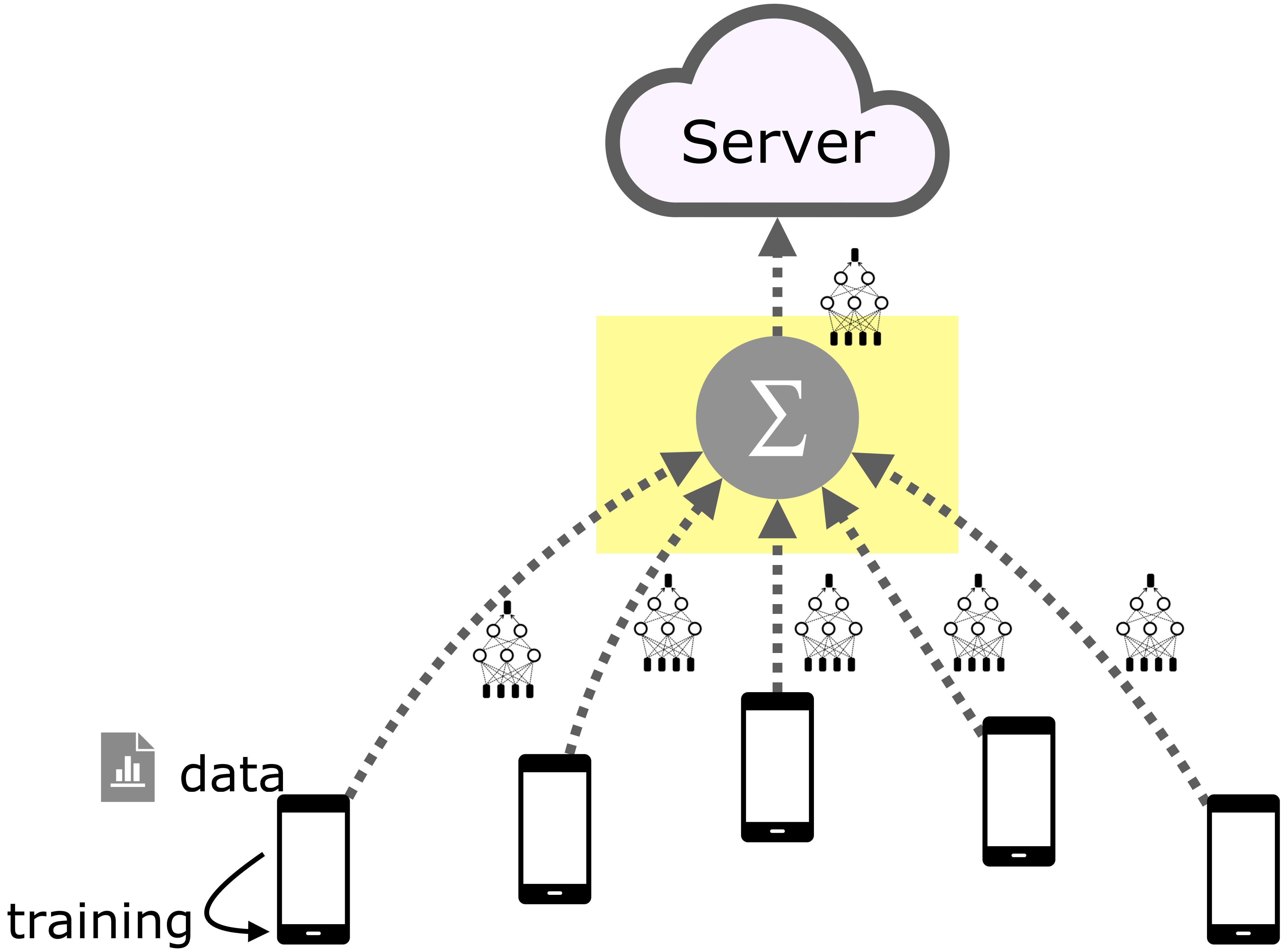


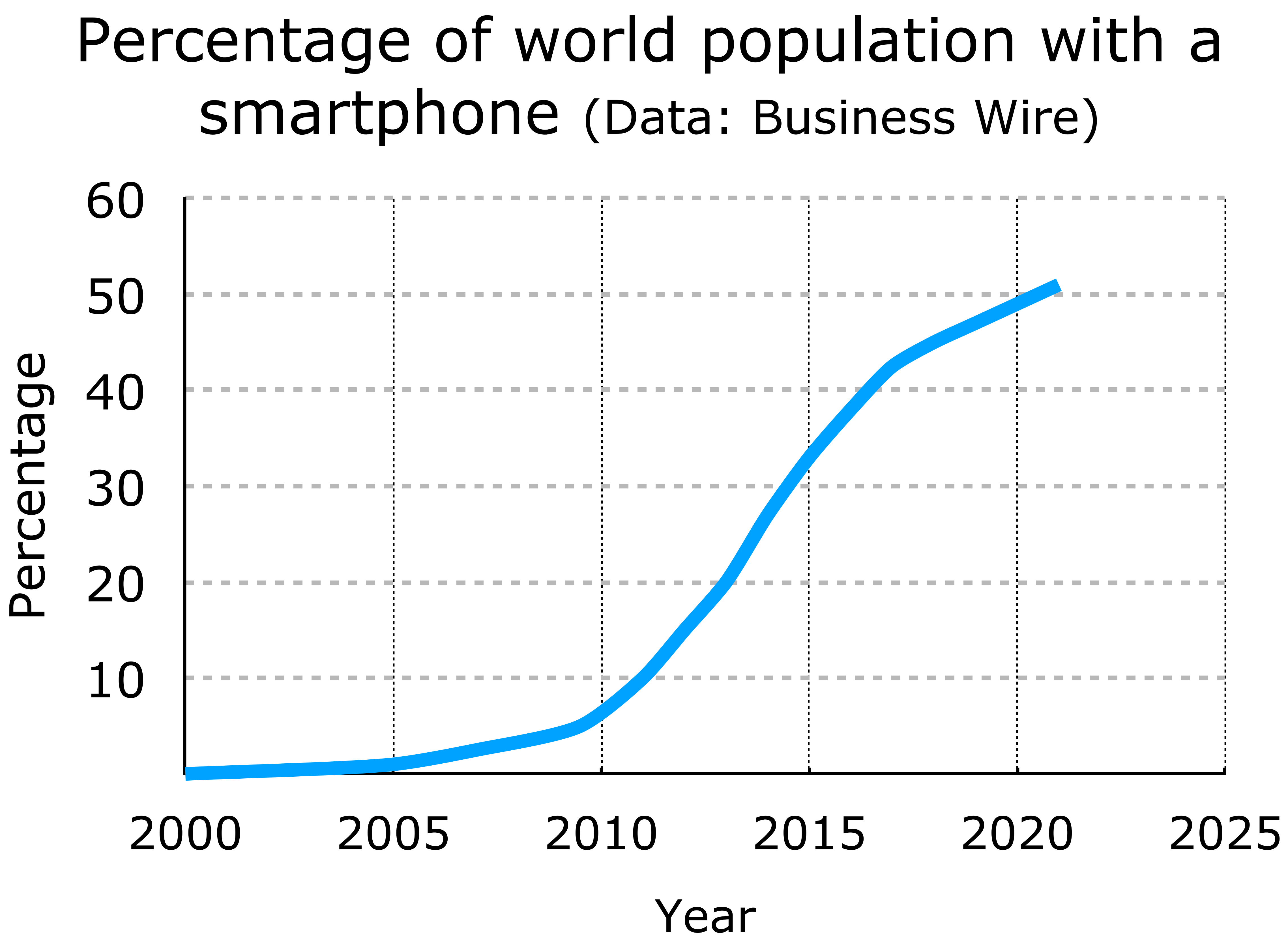
Federated Learning



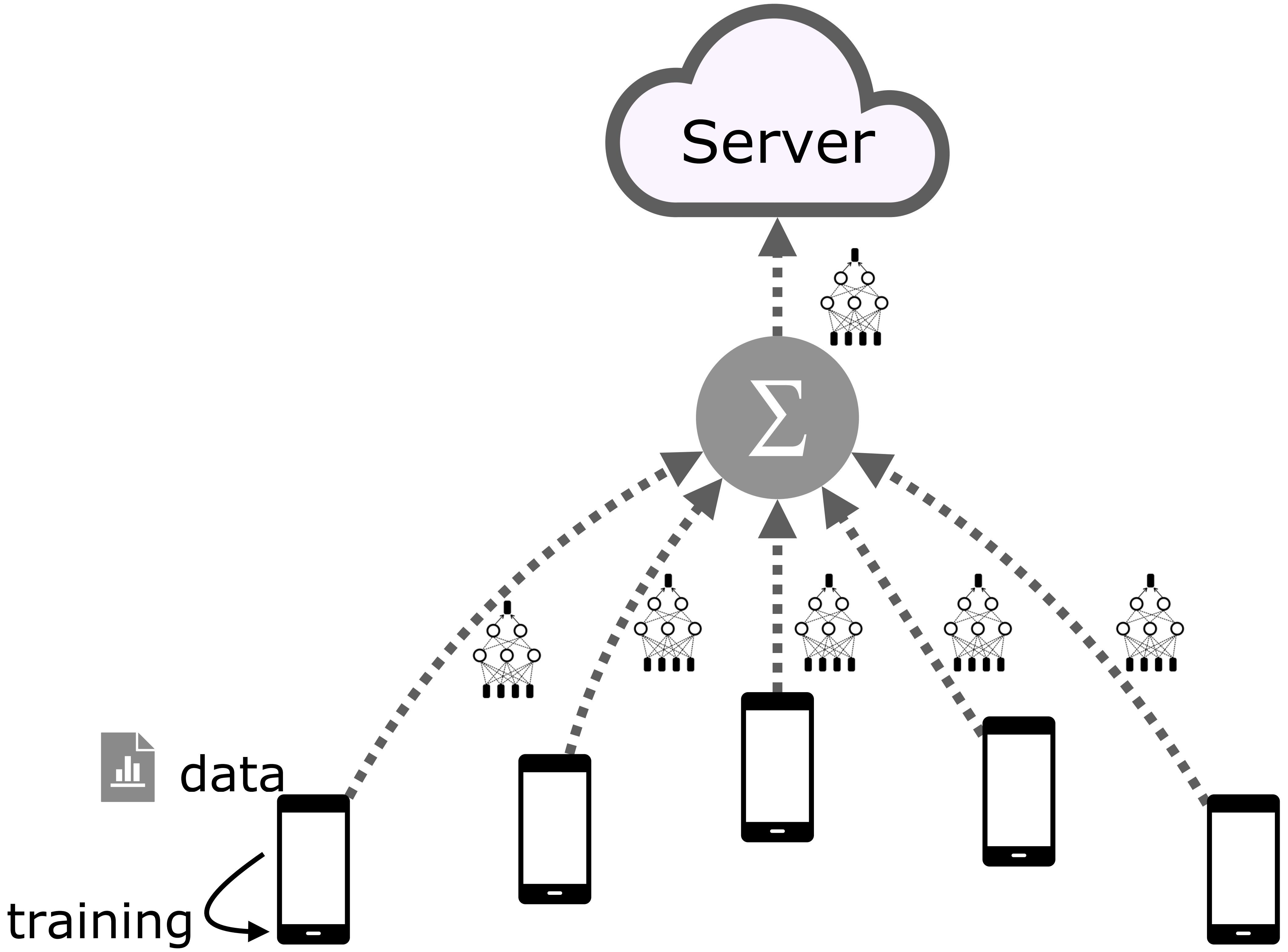


Federated Learning

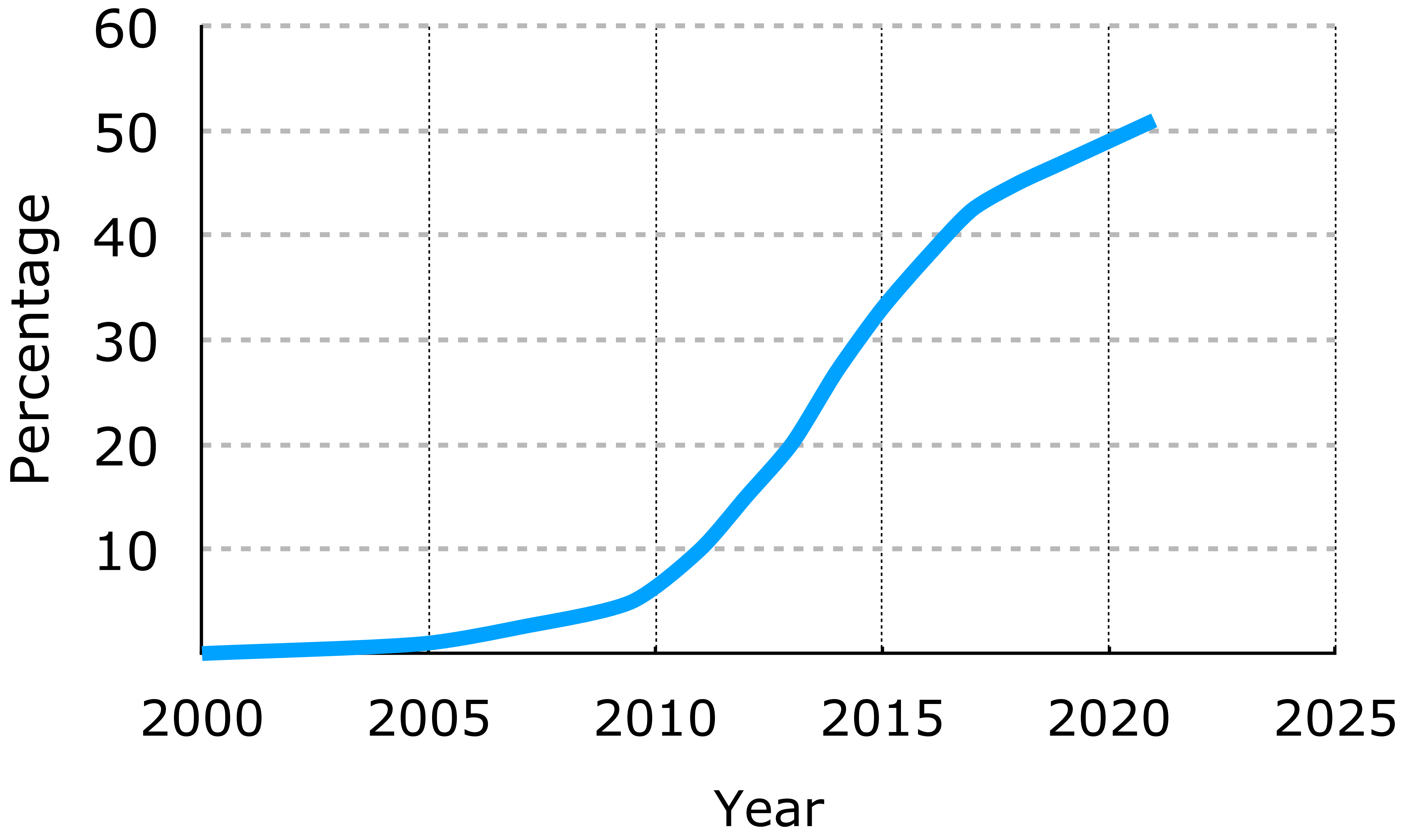




Federated Learning



Percentage of world population with a smartphone (Data: Business Wire)



Communication cost > computation cost!

Google Research

Federated Learning: Collaborative Machine Learning without Centralized Training Data

April 6, 2017 · Posted by Brendan McMahan and Daniel Ramage, Research Scientists

How Apple personalizes Siri without hoovering up your data

The tech giant is using privacy-preserving machine learning to improve its voice assistant while keeping your data on your phone.

By Karen Hao December 11, 2019

IBM Federated Learning

Federated Learning

Engineering at Meta

POSTED ON JUNE 14, 2022 TO AI RESEARCH, ML APPLICATIONS, PRODUCTION ENGINEERING, SECURITY

Applying federated learning to protect data on mobile devices

BANKING & PAYMENTS

Tencent's WeBank applying "federated learning" in A.I.

China's first mobile bank, Tencent's WeBank, is partnering with a H.K. startup to access decentralized sources of data.

Published 5 years ago on July 29, 2019

Federated Learning for Postoperative Segmentation of Treated glioblastoma (FL-PoST)

Federated learning in healthcare: the future of collaborative clinical and biomedical research

OWKIN

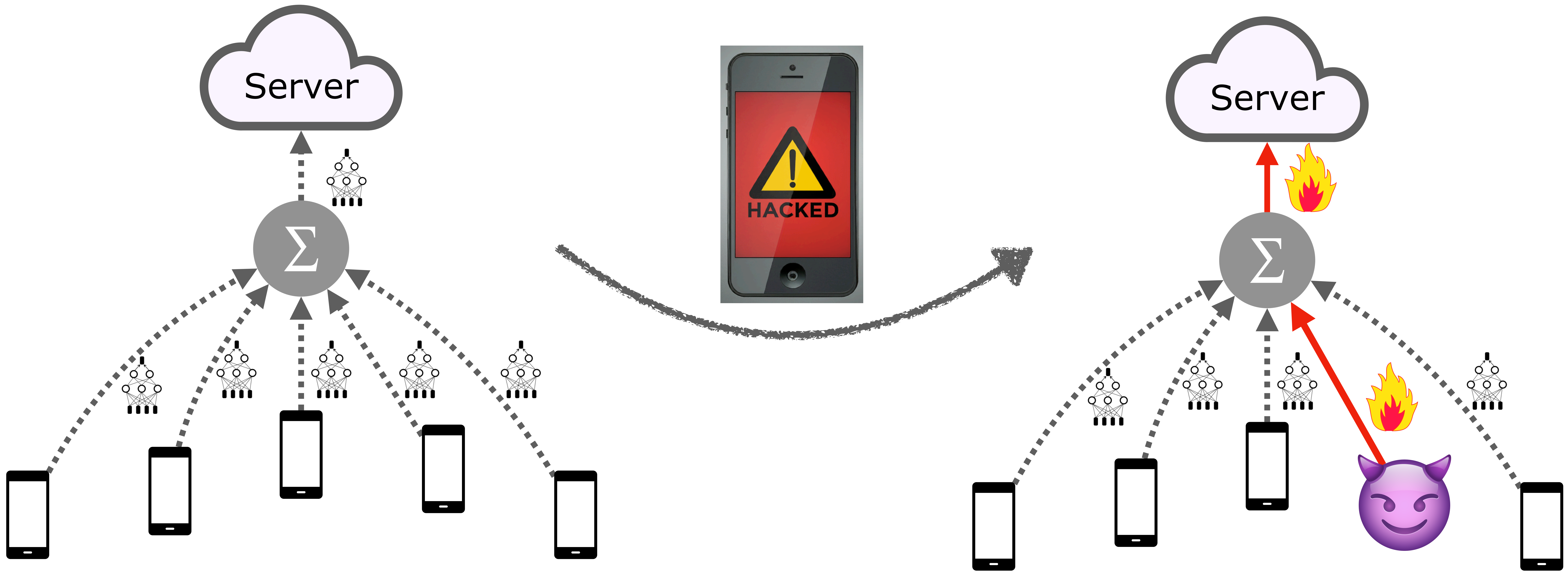
intel PRODUCTS SUPPORT SOLUTIONS DEVELOPERS PARTNERS FOUNDRY

Developers / Topics & Technologies / Open Ecosystem / Try Federated Learning with OpenFL

Try Federated Learning with OpenFL

Challenge:

Training is ***not robust*** to potentially ***malicious*** clients



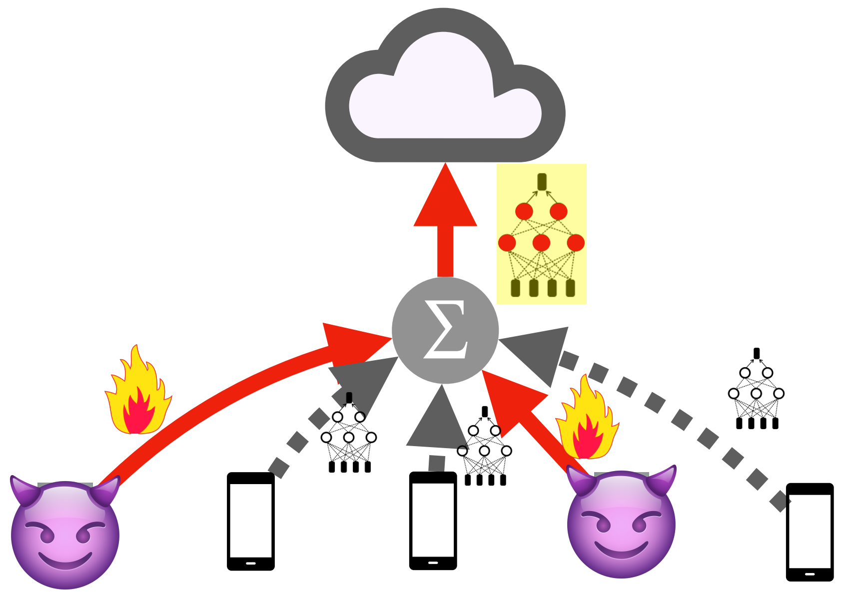
The New York Times

Alexa and Siri Can Hear This Hidden Command. You Can't.

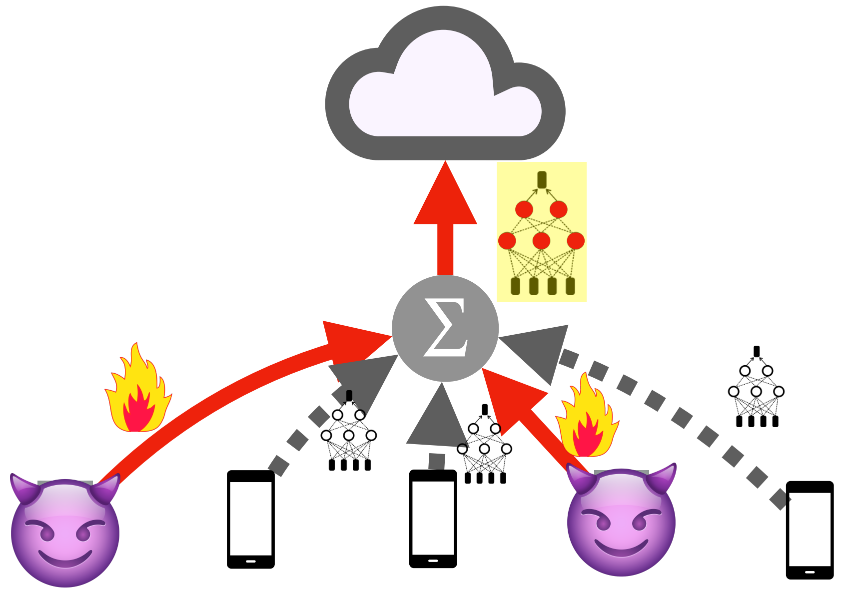
Researchers can now send secret audio instructions undetectable to the human ear to Apple's Siri, Amazon's Alexa and Google's Assistant.

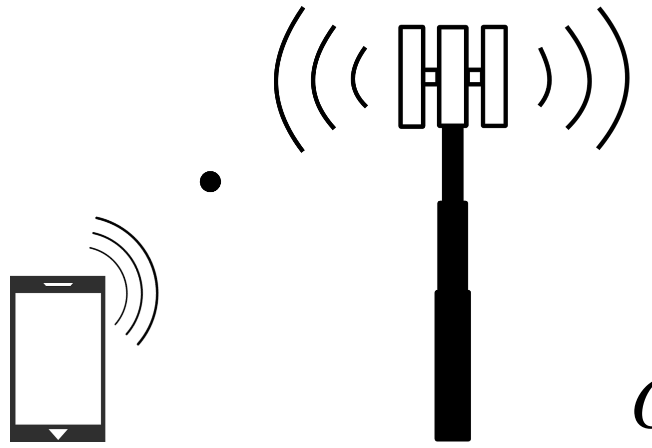
By Craig S. Smith May 10, 2018

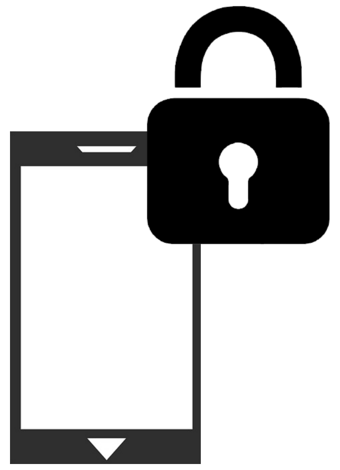
Usual mean aggregation is ***not robust*** to corruptions $\Longrightarrow$ Poor predictions!

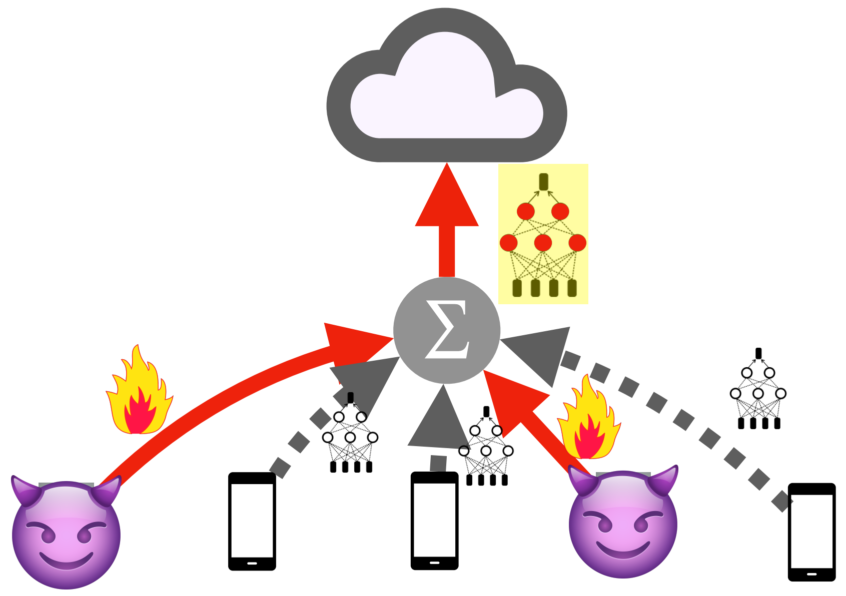
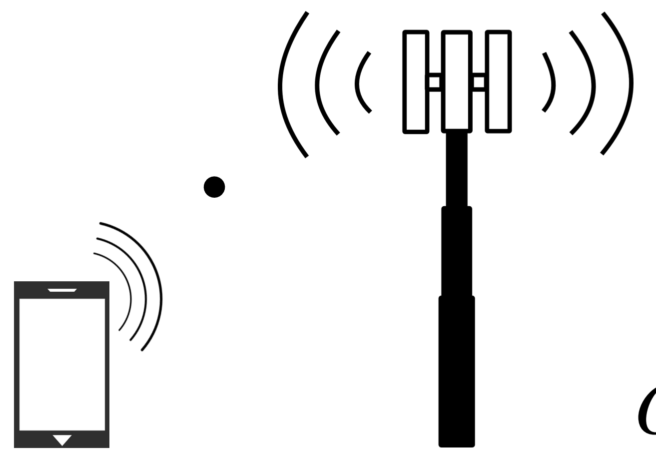
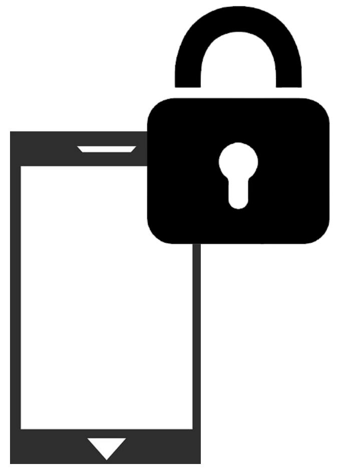
		Usual approach (Direct)	***Our approach*** (Variational)
	**Robust** Robust to outliers/poisoning	✗	✓

	Usual approach (Direct)	***Our approach*** (Variational)
Robust Robust to outliers/poisoning	✗	✓
Communication efficient $O(1)$ times the communication cost as non-robust aggregation		
Secure aggregation Individual updates not revealed		



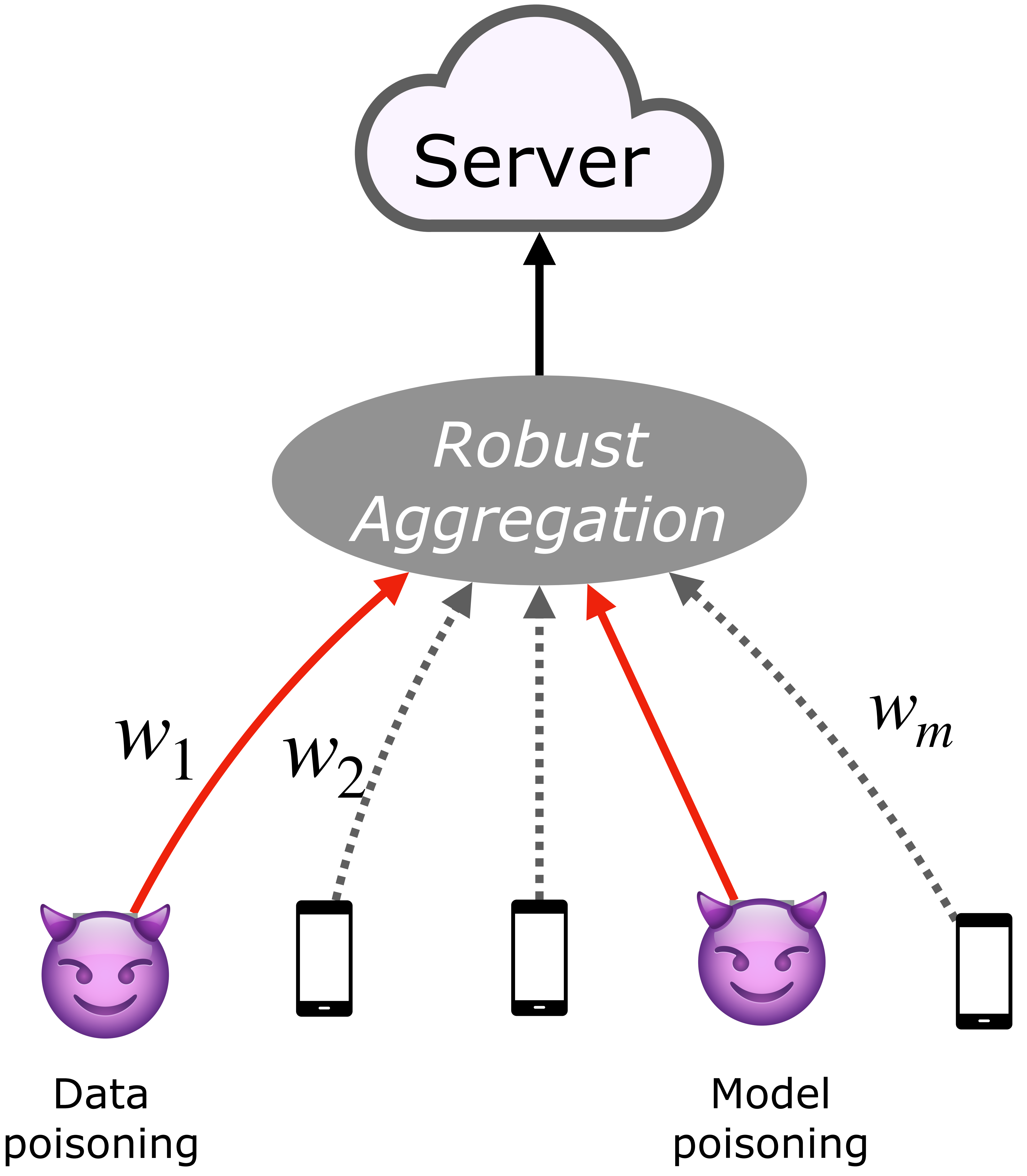




	Usual approach (Direct)	***Our approach*** (Variational)
 **Robust** Robust to outliers/poisoning	✗	✓
 **Communication efficient** $O(1)$ times the communication cost as non-robust aggregation	✓	**??**
 **Secure aggregation** Individual updates not revealed	✓	**??**

Robust aggregation approach

$w_1, \ldots, w_m$: updates sent by the clients



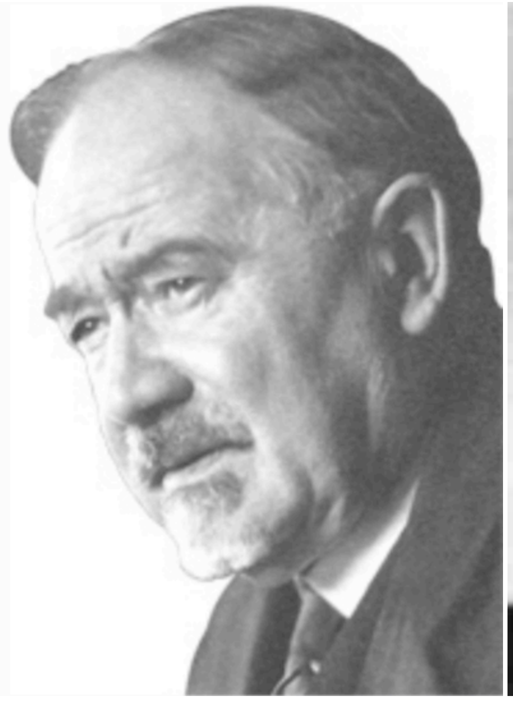
Facility location

Fermat Torricelli Weber Fréchet

(~1600s) (1909) (~1940s)

Facility location





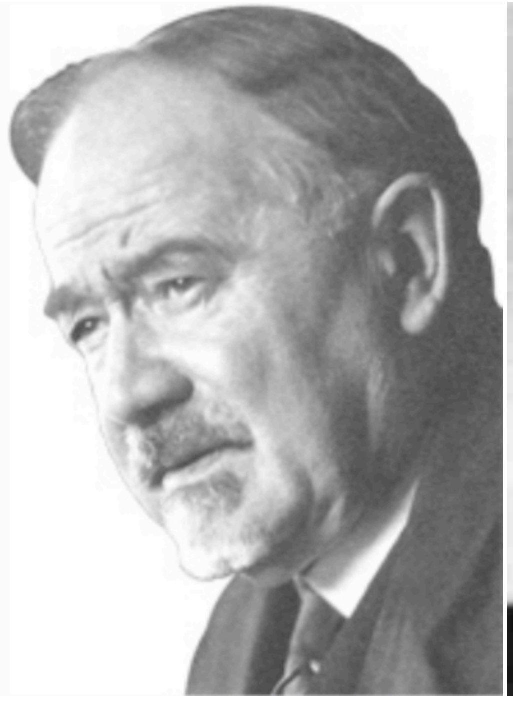

Fermat Torricelli Weber Fréchet

(~1600s) (1909) (~1940s)

Facility location

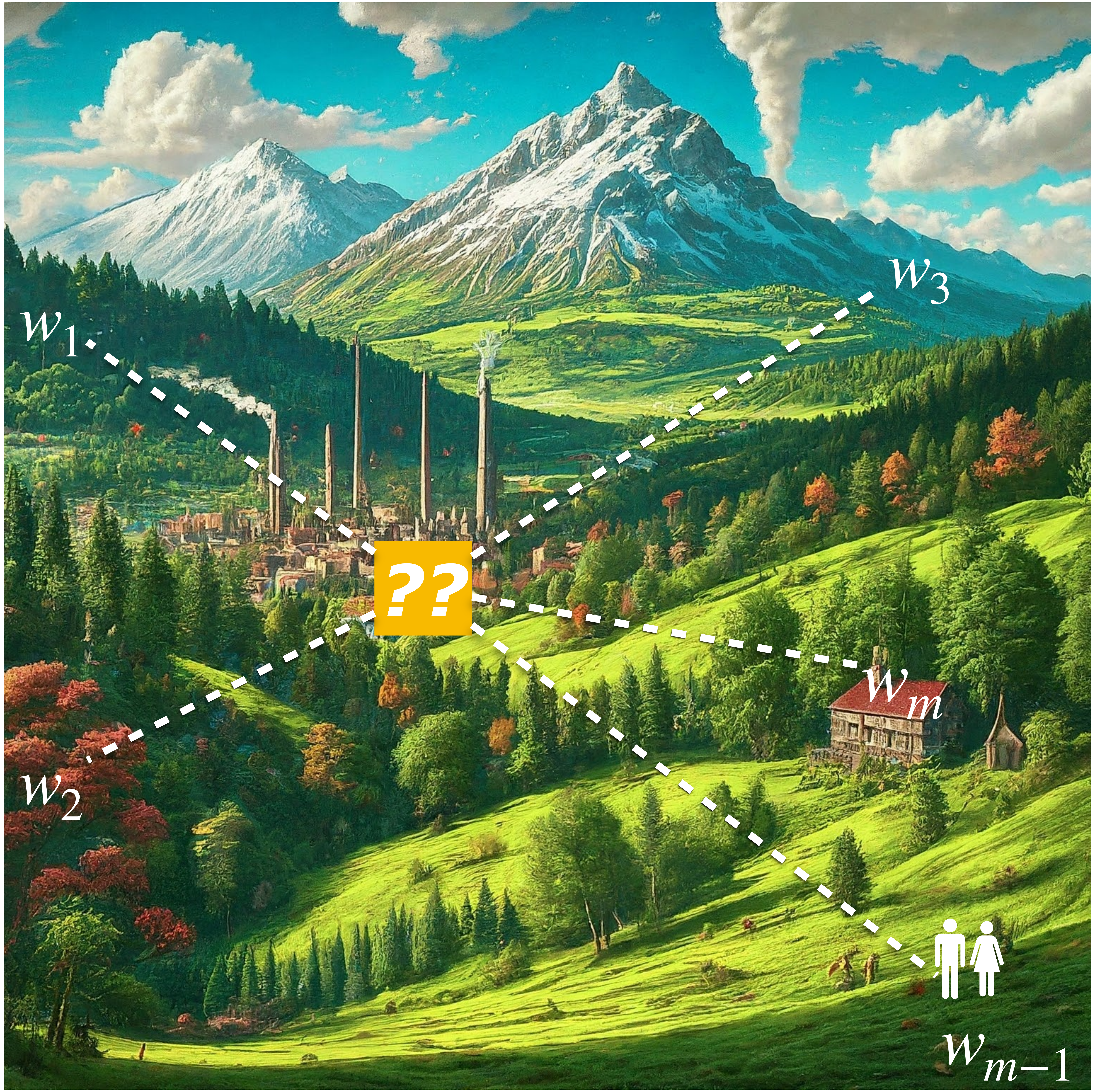
Fermat (~1600s)

Torricelli (~1600s)



Weber (1909)

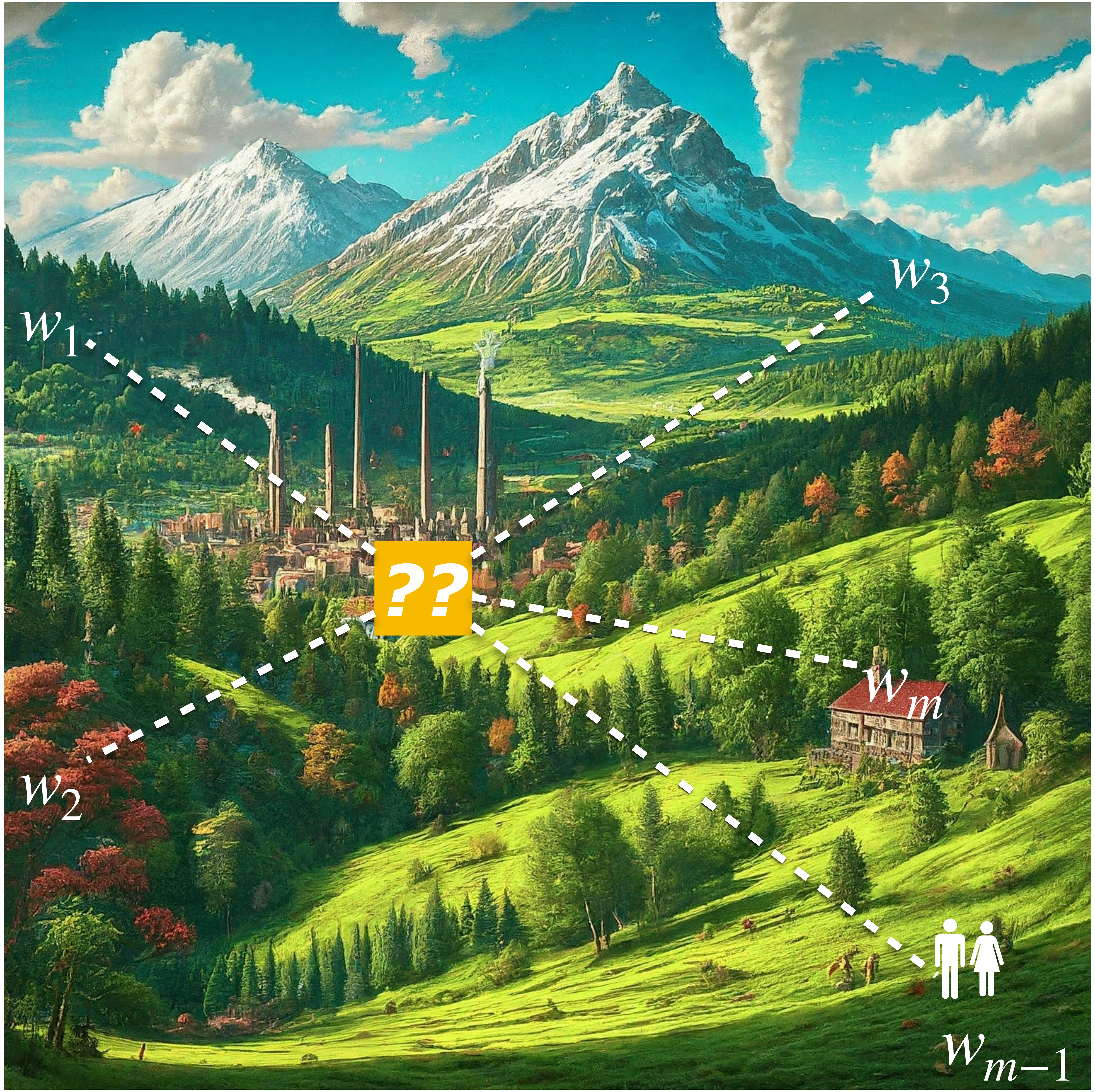
Fréchet (~1940s)



Facility location

Fermat Torricelli Weber Fréchet

(~1600s) (1909) (~1940s)



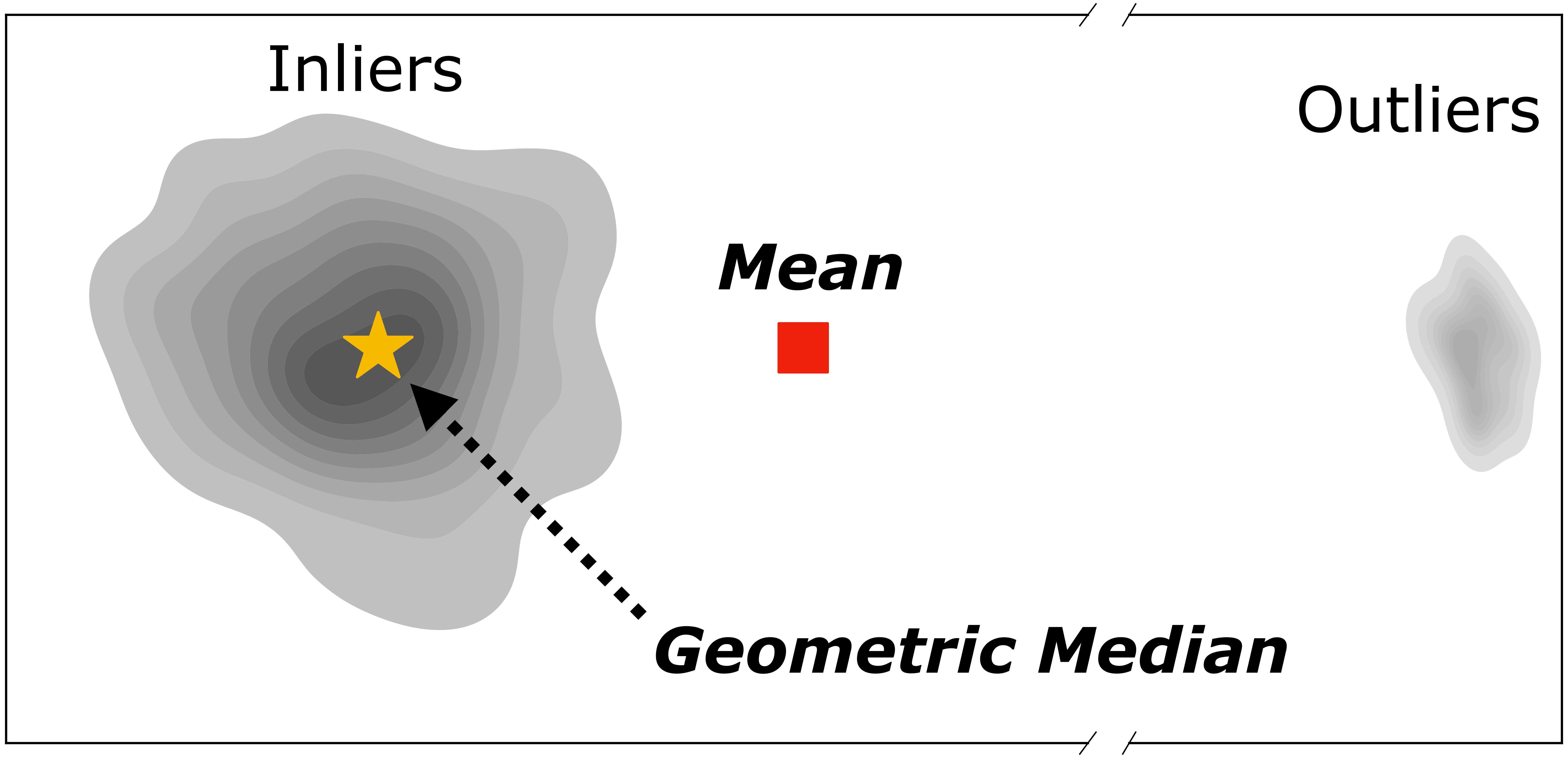
Geometric Median / Spatial Median / L_1 Median / Facility location

$$\mathrm{GM}(w_1, \cdots, w_m) = \arg\min_z \left\{ \sum_{i=1}^{m} \|z - w_i\|_2 \right\}$$

Fermat Torricelli (~1600s) Weber (1909) Fréchet (~1940s)

Robustness: Breakdown point = 1/2

(In **1D**, we have that ***geometric median*** ≡ ***usual median***)



Nemirovski & Yudin (1983) | Jerrum, Valiant & Vazirani (1986) | Lopuhaa & Rousseeuw (1991)
Hsu & Sabata (2013) | Minsker (2015) | Lugosi, Gabor & Mendelson (2019) | Lecué & Lerasle (2020)

Smoothed Weiszfeld Algorithm

Weiszfeld (1937). **Sur le point par lequel la somme des distances de *n* points donnes est minimum**. *Tohoku Mathematical Journal.*

Compute new weights $\beta_{i,t} = \dfrac{1}{\max\{\|z_t - w_i\|_2,\ \nu\}}$ & Reweighted average $z_{t+1} = \dfrac{\sum_i \beta_{i,t} w_i}{\sum_i \beta_{i,t}}$

Weiszfeld a.k.a. Vázsonyi (1916-2003)

Smoothed Weiszfeld Algorithm

Weiszfeld (1937). **Sur le point par lequel la somme des distances de *n* points donnes est minimum**. *Tohoku Mathematical Journal.*

Compute new weights $\beta_{i,t} = \dfrac{1}{\max\{\|z_t - w_i\|_2, \, \nu\}}$ & Reweighted average $z_{t+1} = \dfrac{\sum_i \beta_{i,t} w_i}{\sum_i \beta_{i,t}}$

Weiszfeld a.k.a. Vázsonyi (1916-2003)

Smoothed Weiszfeld Algorithm

Weiszfeld (1937). **Sur le point par lequel la somme des distances de *n* points donnes est minimum**. *Tohoku Mathematical Journal.*

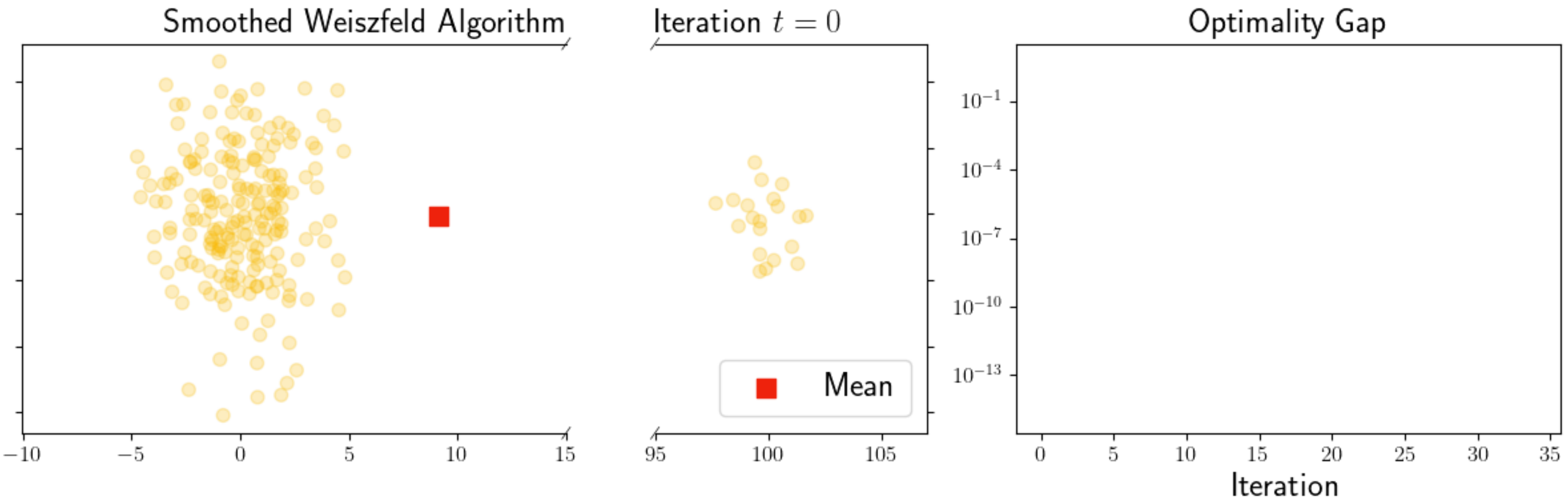
Compute new weights $\beta_{i,t} = \frac{1}{\max\{\|z_t - w_i\|_2,\ \nu\}}$ & Reweighted average $z_{t+1} = \frac{\sum_i \beta_{i,t} w_i}{\sum_i \beta_{i,t}}$

Weiszfeld a.k.a. Vázsonyi (1916-2003)

Smoothed Weiszfeld Algorithm

Weiszfeld (1937). **Sur le point par lequel la somme des distances de *n* points donnes est minimum**. *Tohoku Mathematical Journal.*

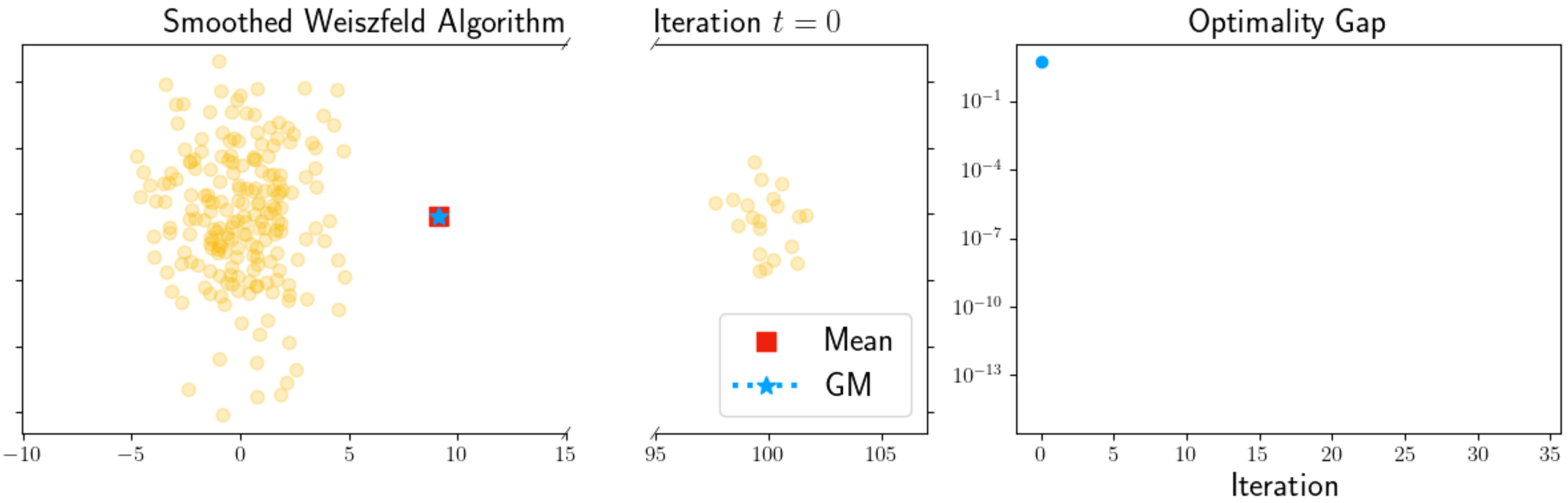
Compute new weights $\beta_{i,t} = \dfrac{1}{\max\{\|z_t - w_i\|_2,\, \nu\}}$ & Reweighted average $z_{t+1} = \dfrac{\sum_i \beta_{i,t} w_i}{\sum_i \beta_{i,t}}$



Smoothed Weiszfeld Algorithm

Weiszfeld (1937). **Sur le point par lequel la somme des distances de *n* points donnes est minimum**. *Tohoku Mathematical Journal.*

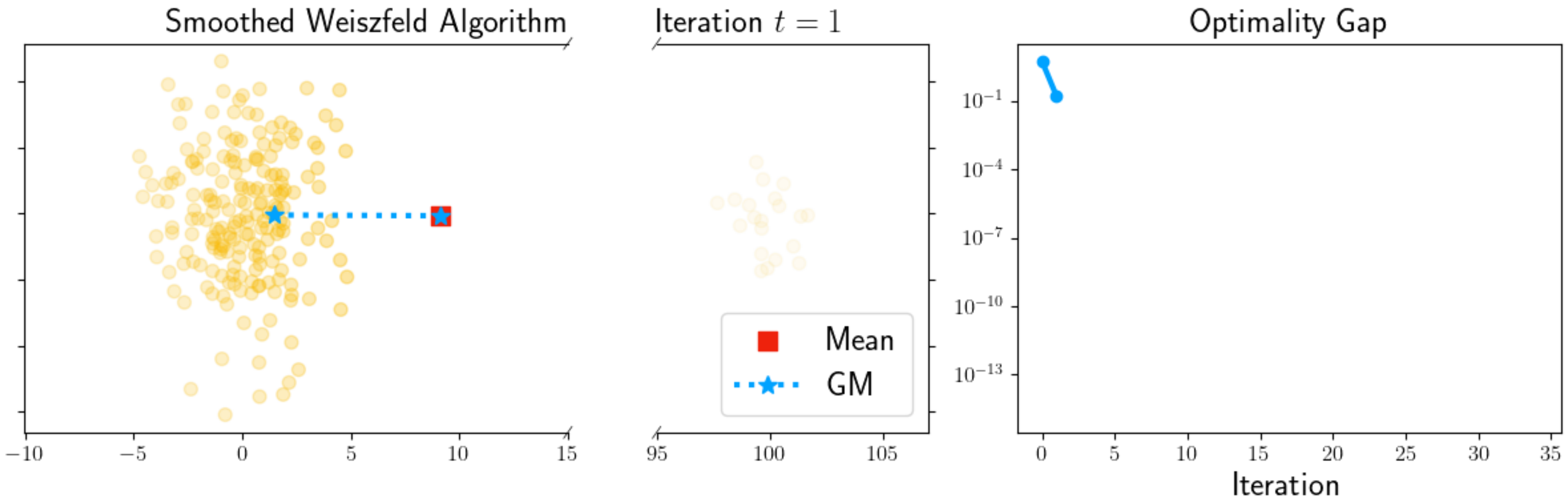
Compute new weights $\beta_{i,t} = \dfrac{1}{\max\{\|z_t - w_i\|_2,\, \nu\}}$ & Reweighted average $z_{t+1} = \dfrac{\sum_i \beta_{i,t} w_i}{\sum_i \beta_{i,t}}$



Smoothed Weiszfeld Algorithm

Weiszfeld (1937). **Sur le point par lequel la somme des distances de *n* points donnes est minimum**. *Tohoku Mathematical Journal.*

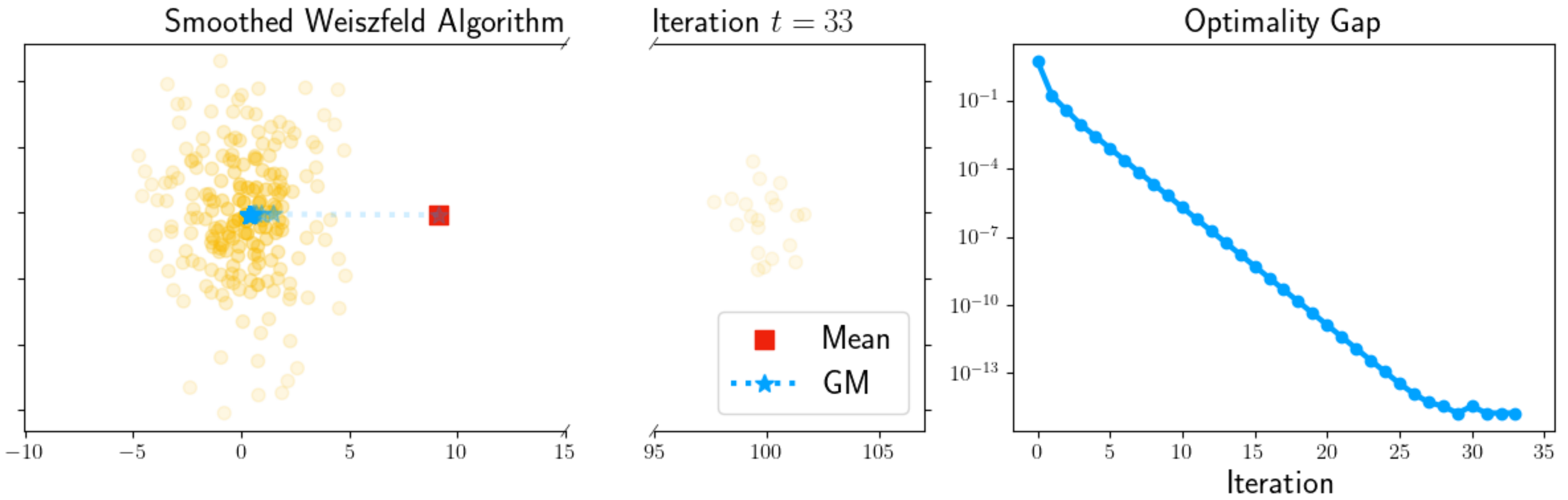
Compute new weights $\beta_{i,t} = \dfrac{1}{\max\{\|z_t - w_i\|_2,\, \nu\}}$ & Reweighted average $z_{t+1} = \dfrac{\sum_i \beta_{i,t} w_i}{\sum_i \beta_{i,t}}$

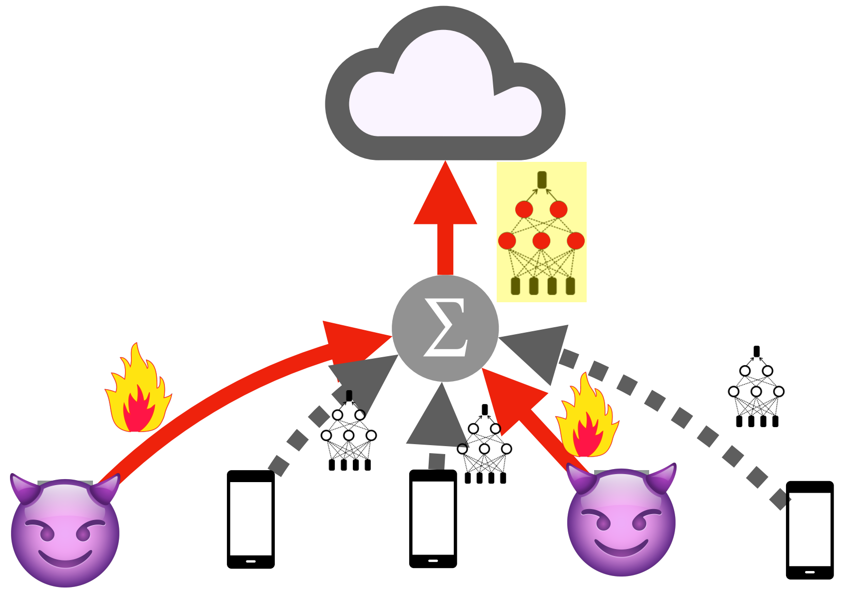
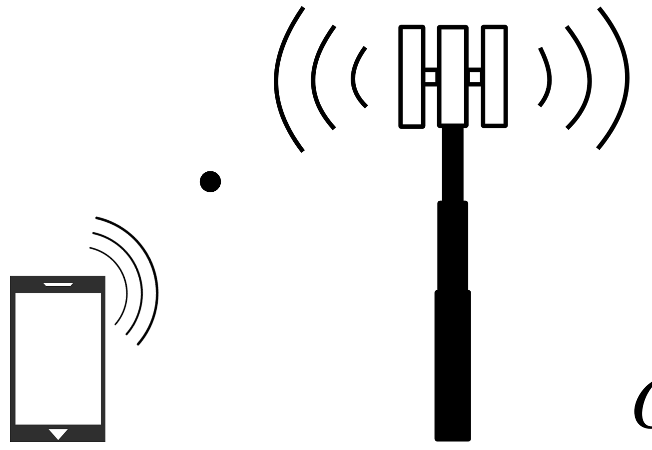
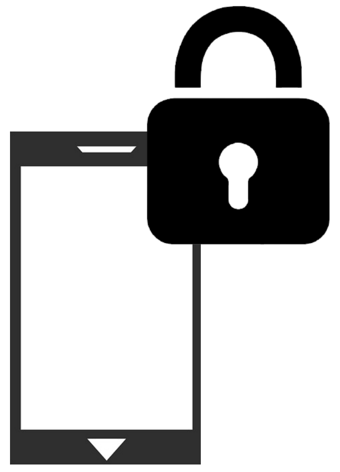


Smoothed Weiszfeld Algorithm

Weiszfeld (1937). **Sur le point par lequel la somme des distances de *n* points donnes est minimum**. *Tohoku Mathematical Journal.*

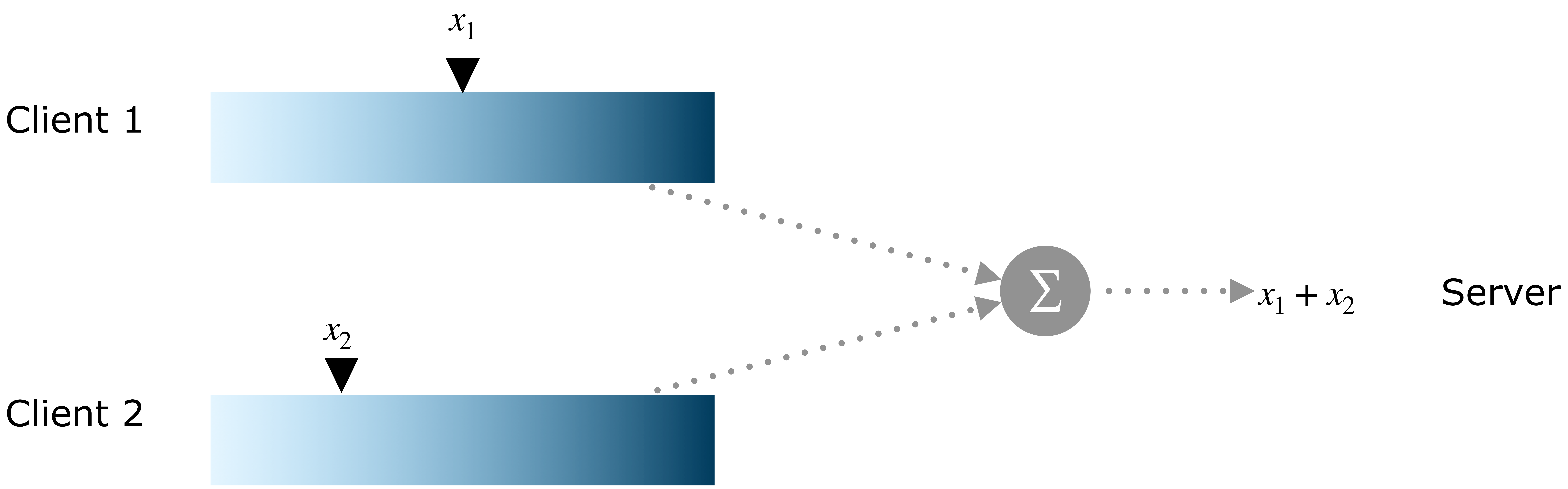
Compute new weights $\beta_{i,t} = \frac{1}{\max\{\|z_t - w_i\|_2, \nu\}}$ & Reweighted average $z_{t+1} = \frac{\sum_i \beta_{i,t} w_i}{\sum_i \beta_{i,t}}$



	Usual approach (Direct)	***Our approach*** (Variational)
 **Robust** Robust to outliers/poisoning	✗	✓
 **Communication efficient** $O(1)$ times the communication cost as non-robust aggregation	✓	**??**
 **Secure aggregation** Individual updates not revealed	✓	**??**

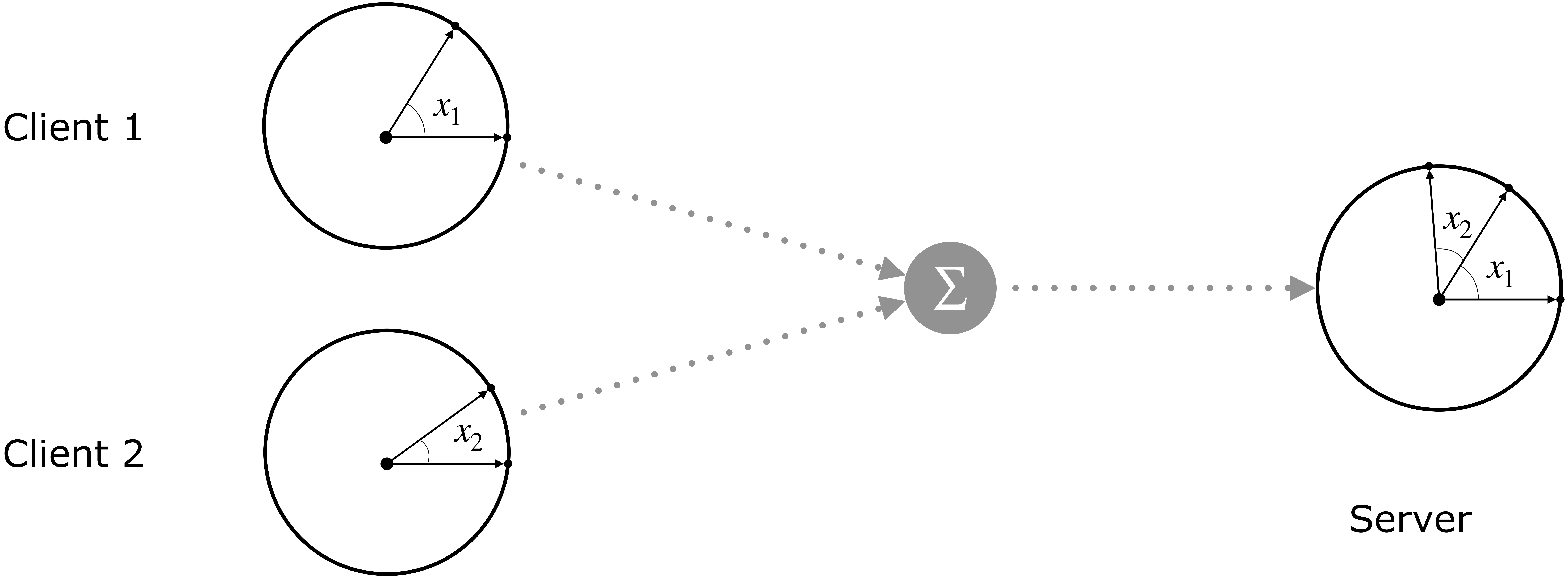
Communication primitive: secure sum/average

Only reveal $x_1 + x_2$ to the server without revealing x_1 or x_2



[Bonawitz et al. CCS (2017), Bell et al. CCS (2020)]

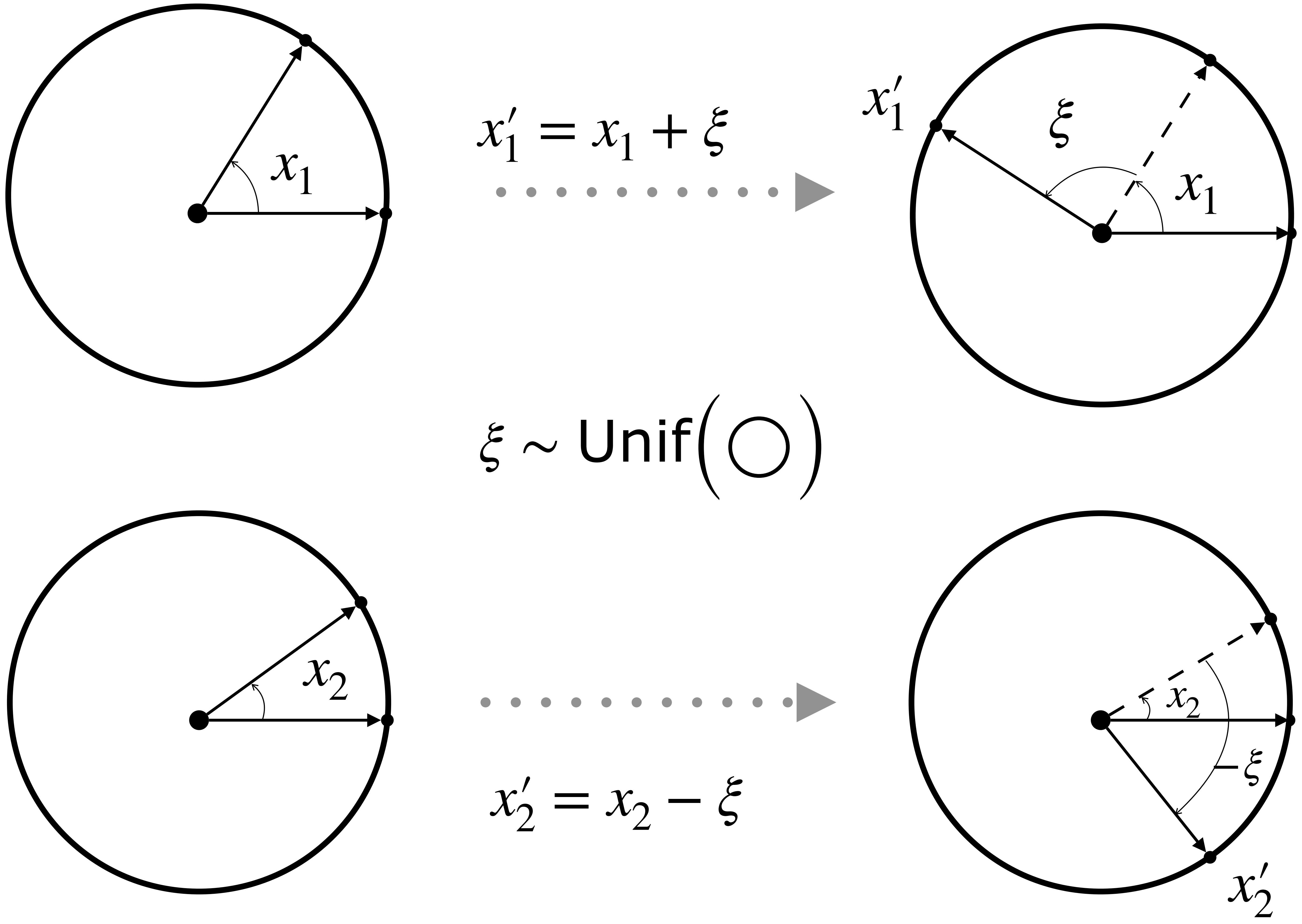
Perform all operations modulo M



[Bonawitz et al. CCS (2017), Bell et al. CCS (2020)]

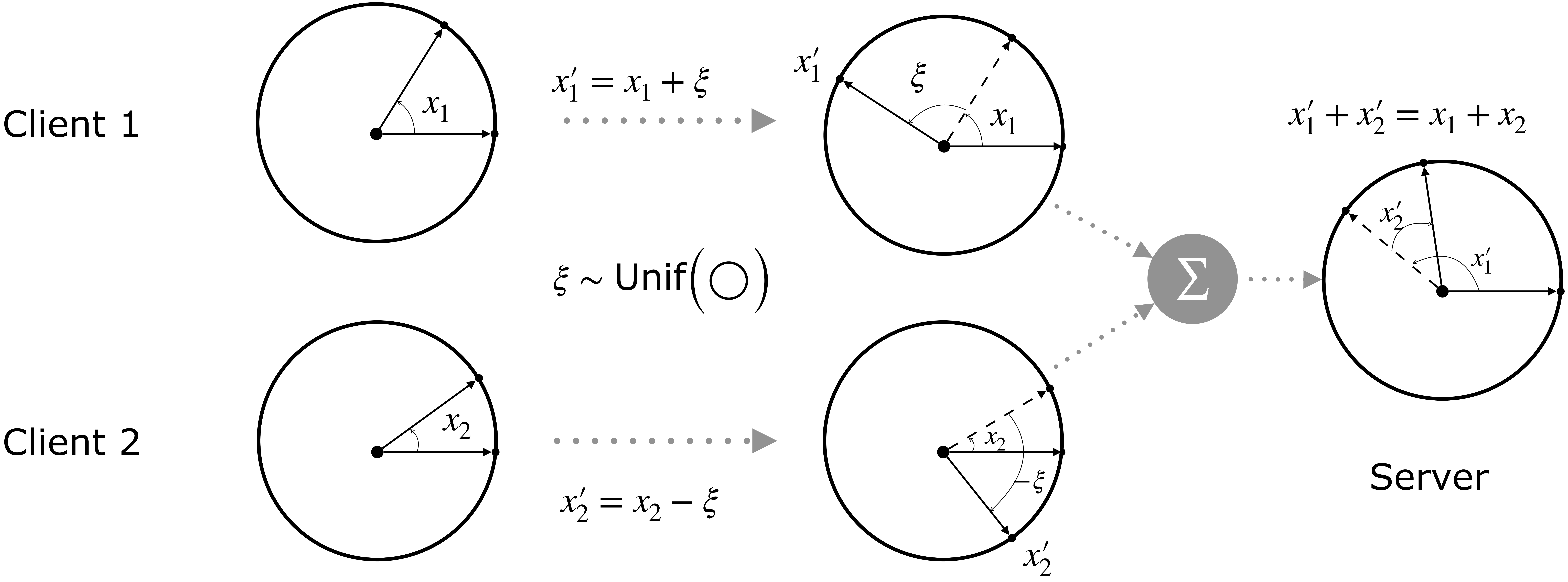
Client 1

Client 2



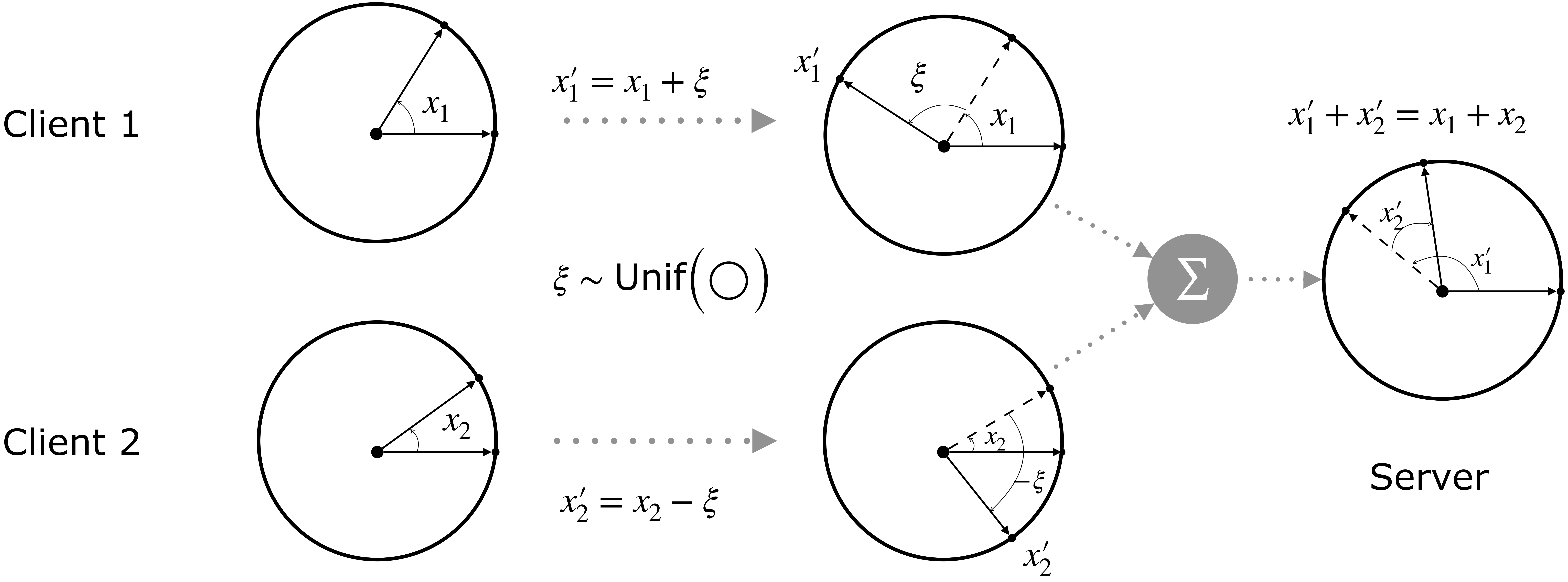
[Bonawitz et al. CCS (2017), Bell et al. CCS (2020)]

Server only sees $x'_1, x'_2 \sim$ Unif$(\bigcirc)$ but calculates the correct sum (and average)



[Bonawitz et al. CCS (2017), Bell et al. CCS (2020)]

Server only sees $x_1', x_2' \sim \text{Unif}(\bigcirc)$ but calculates the correct sum (and average)



Total communication for m vectors in $\mathbb{R}^d$ $= O(m \log m + md)$ numbers

Server only sees $x_1', x_2' \sim \text{Unif}(\bigcirc)$ but calculates the correct sum

Real-world communication constraint:
All client-to-server communication must go through secure average

Extensions: weighted sums/averages

Total communication for m vectors in $\mathbb{R}^d$ $= O(m \log m + md)$ numbers

Smoothed Weiszfeld Algorithm

Weiszfeld (1937). **Sur le point par lequel la somme des distances de *n* points donnes est minimum**. *Tohoku Mathematical Journal.*

Compute new weights $\beta_{i,t} = \frac{1}{\max\{\|z_t - w_i\|_2,\ \nu\}}$ & Reweighted average $z_{t+1} = \frac{\sum_i \beta_{i,t} w_i}{\sum_i \beta_{i,t}}$

Smoothed Weiszfeld Algorithm

Weiszfeld (1937). **Sur le point par lequel la somme des distances de *n* points donnes est minimum**. *Tohoku Mathematical Journal.*

Compute new weights $\beta_{i,t} = \dfrac{1}{\max\{\|z_t - w_i\|_2,\ \nu\}}$ & Reweighted average $z_{t+1} = \dfrac{\sum_i \beta_{i,t} w_i}{\sum_i \beta_{i,t}}$

1. Server **broadcasts** current estimate z_t of the geometric median

Smoothed Weiszfeld Algorithm

Weiszfeld (1937). **Sur le point par lequel la somme des distances de *n* points donnes est minimum**. *Tohoku Mathematical Journal.*

Compute new weights $\beta_{i,t} = \dfrac{1}{\max\{\|z_t - w_i\|_2,\ \nu\}}$ & Reweighted average $z_{t+1} = \dfrac{\sum_i \beta_{i,t} w_i}{\sum_i \beta_{i,t}}$

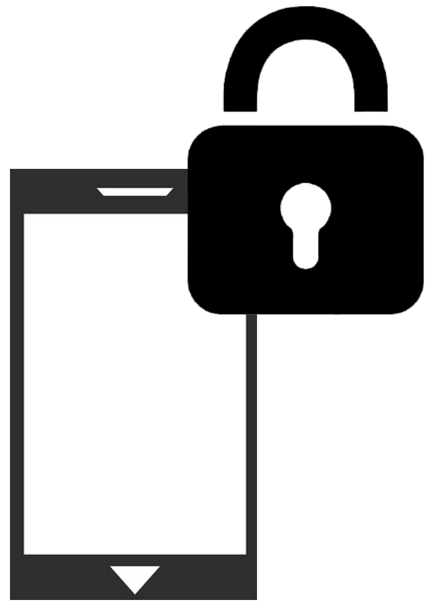
2. **Clients** compute new weights

Smoothed Weiszfeld Algorithm

Weiszfeld (1937). **Sur le point par lequel la somme des distances de *n* points donnes est minimum**. *Tohoku Mathematical Journal.*

Compute new weights $\beta_{i,t} = \dfrac{1}{\max\{\|z_t - w_i\|_2,\ \nu\}}$ & Reweighted average $z_{t+1} = \dfrac{\sum_i \beta_{i,t} w_i}{\sum_i \beta_{i,t}}$

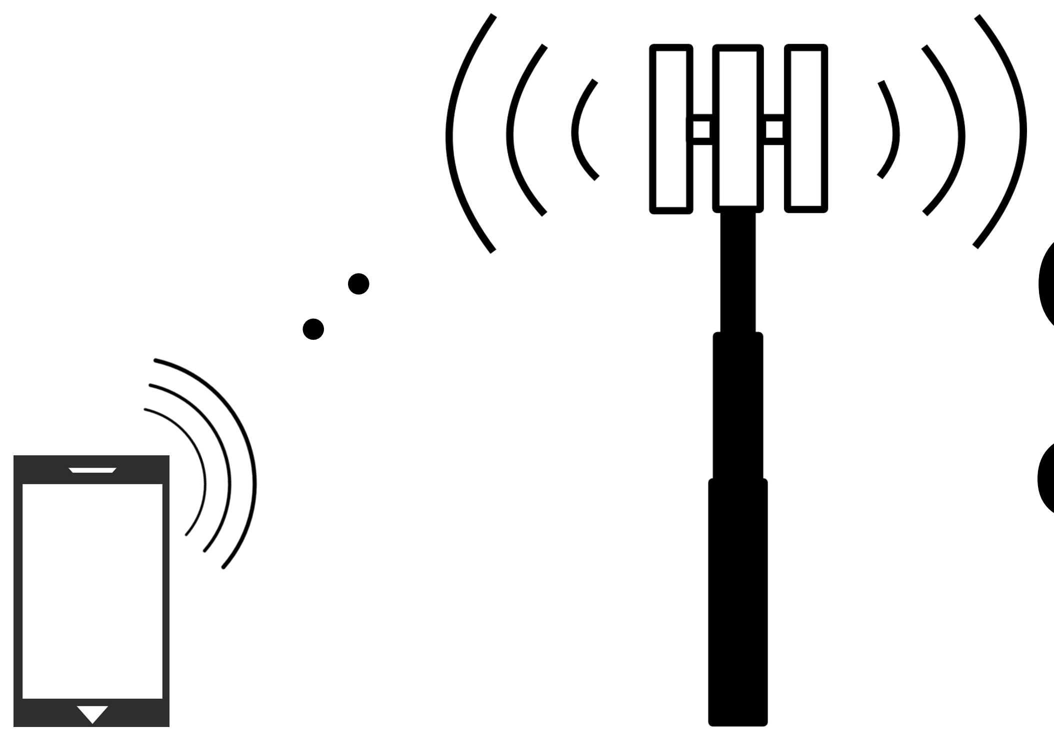
3. Obtain new estimate by **secure averaging**



Secure aggregation

Only client-server communication is via **secure average** in the **Smoothed Weiszfeld Algorithm**

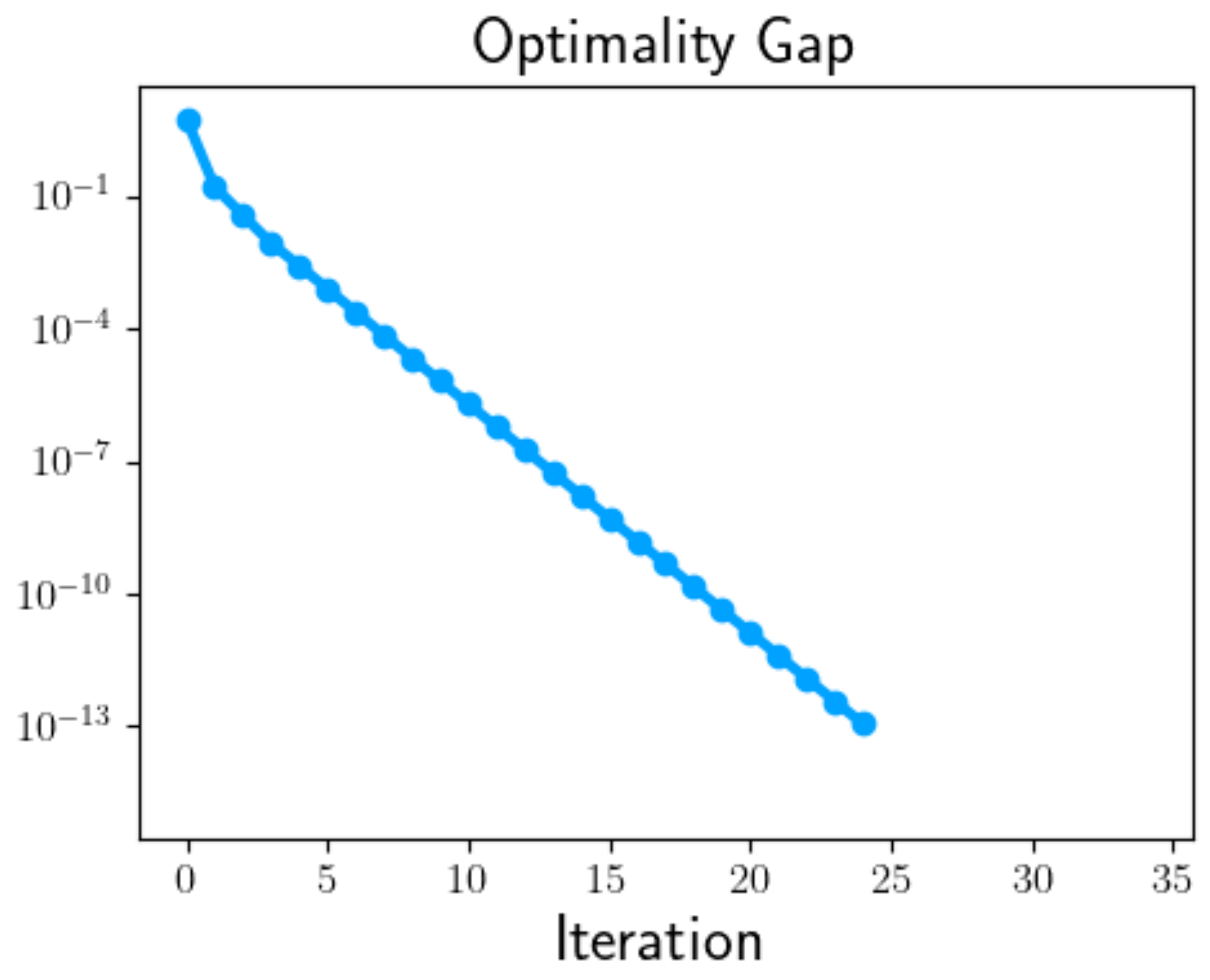
$$z_{t+1} = \frac{\sum_i \beta_{i,t} w_i}{\sum_i \beta_{i,t}}$$



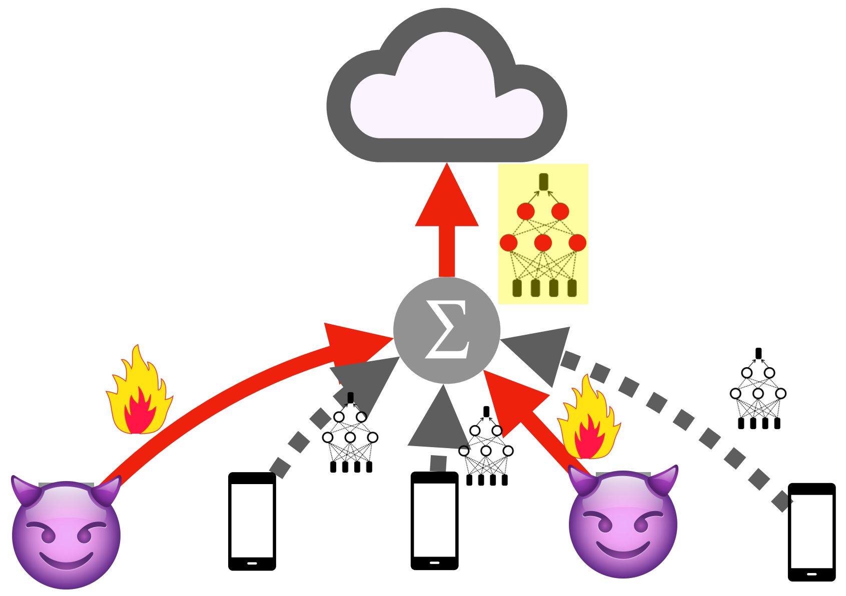
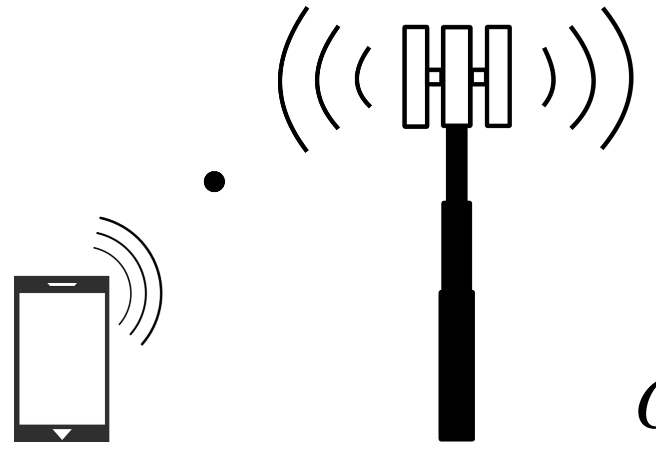
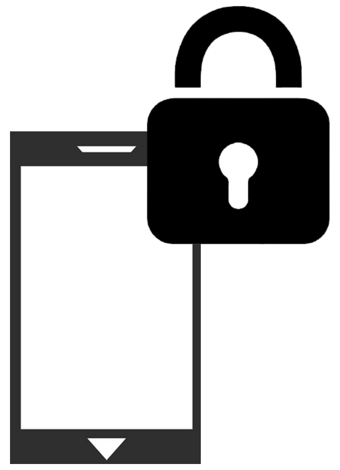
Communication efficient!

Empirically, **3-5** iterations suffice:

provably rapid convergence



Even **1** iteration improves robustness!

	Usual approach (Direct)	***Our approach*** (Variational)
 **Robust** Robust to outliers/poisoning	✗	✓
 **Communication efficient** $O(1)$ times the communication cost as non-robust aggregation	✓	✓
 **Secure aggregation** Individual updates not revealed	✓	✓

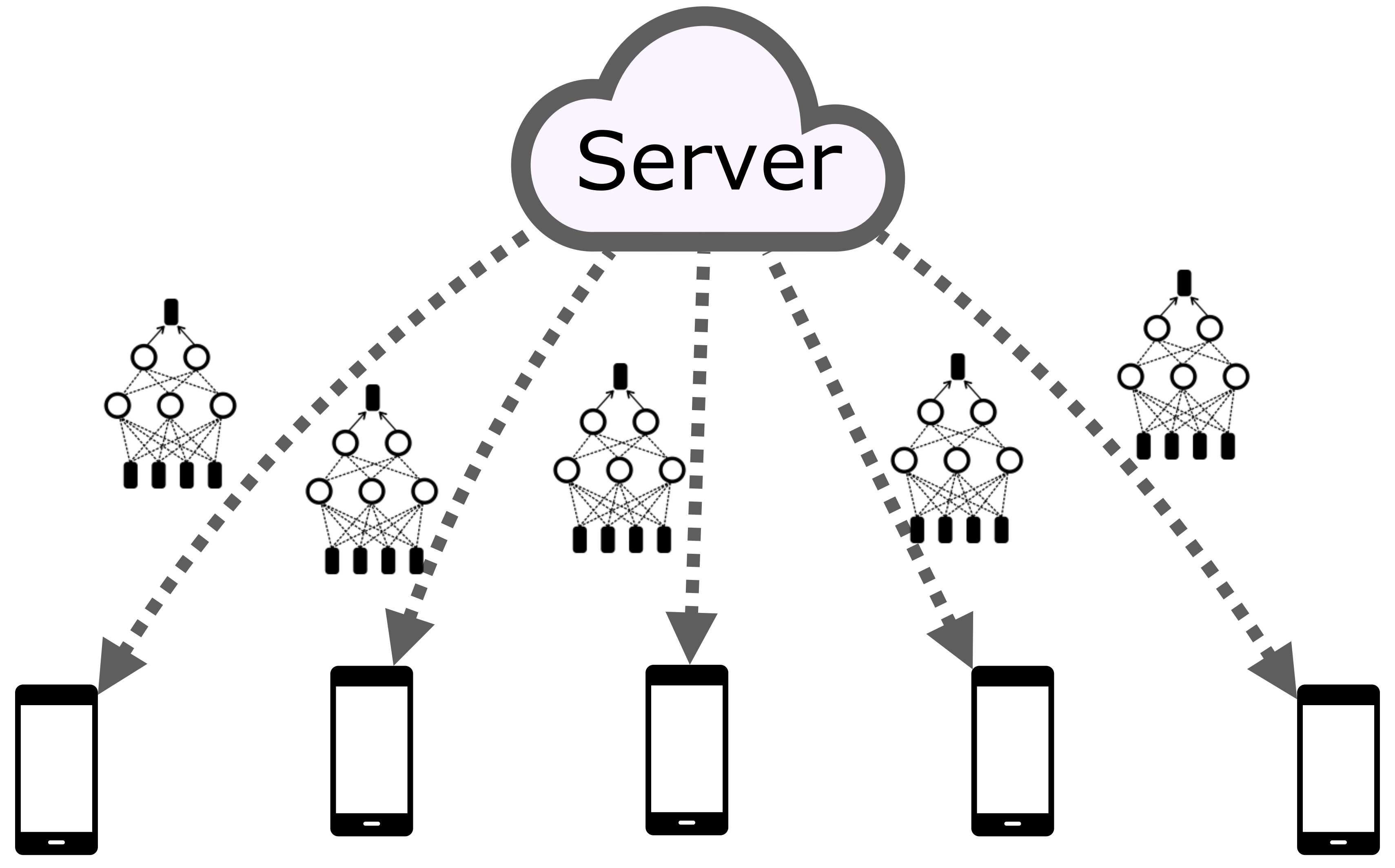
Robust Federated Aggregation (RFA)

More robust federated learning =

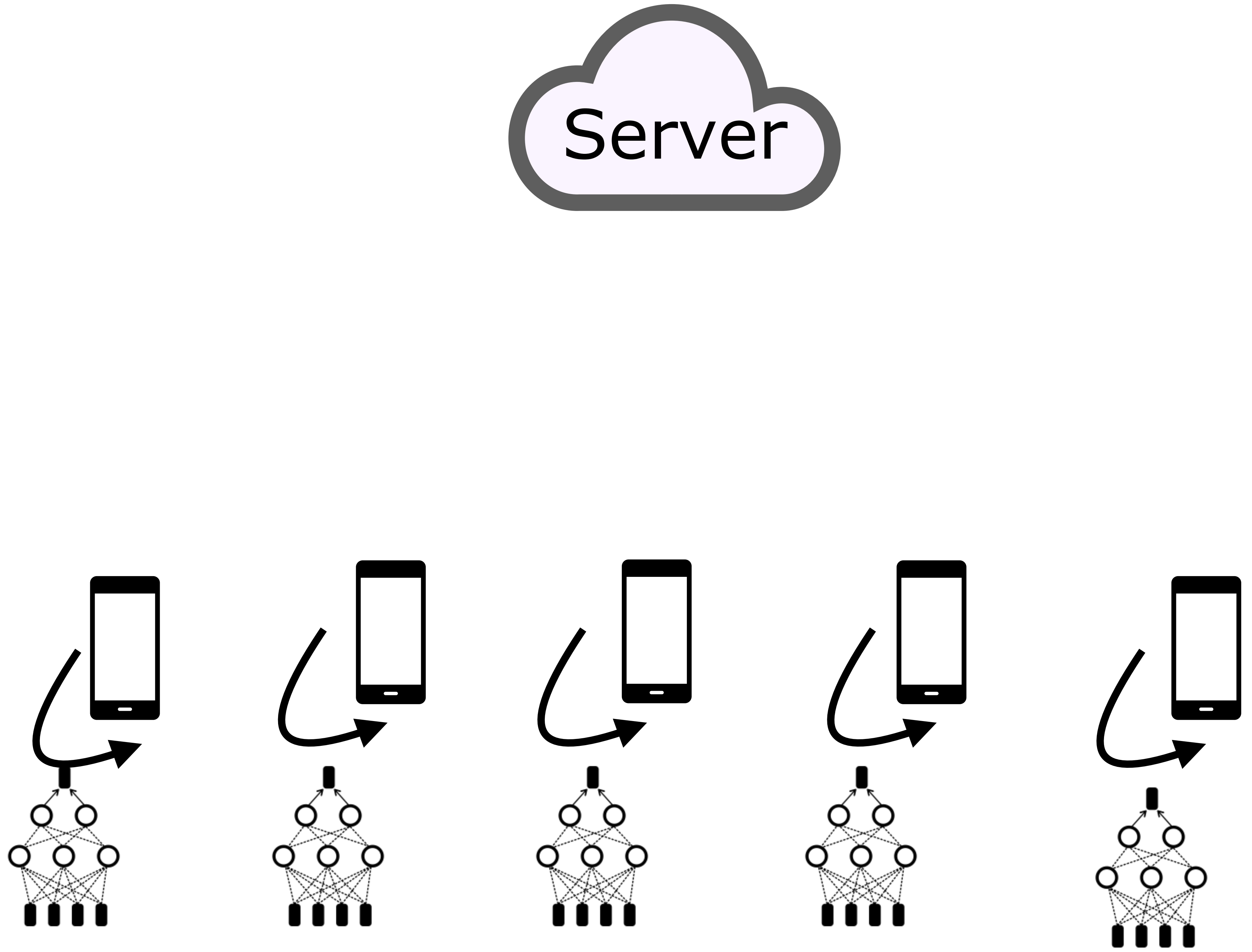
Local SGD steps +

Geometric median + secure aggregation

Step 1 of 3: Server broadcasts global model to sampled clients

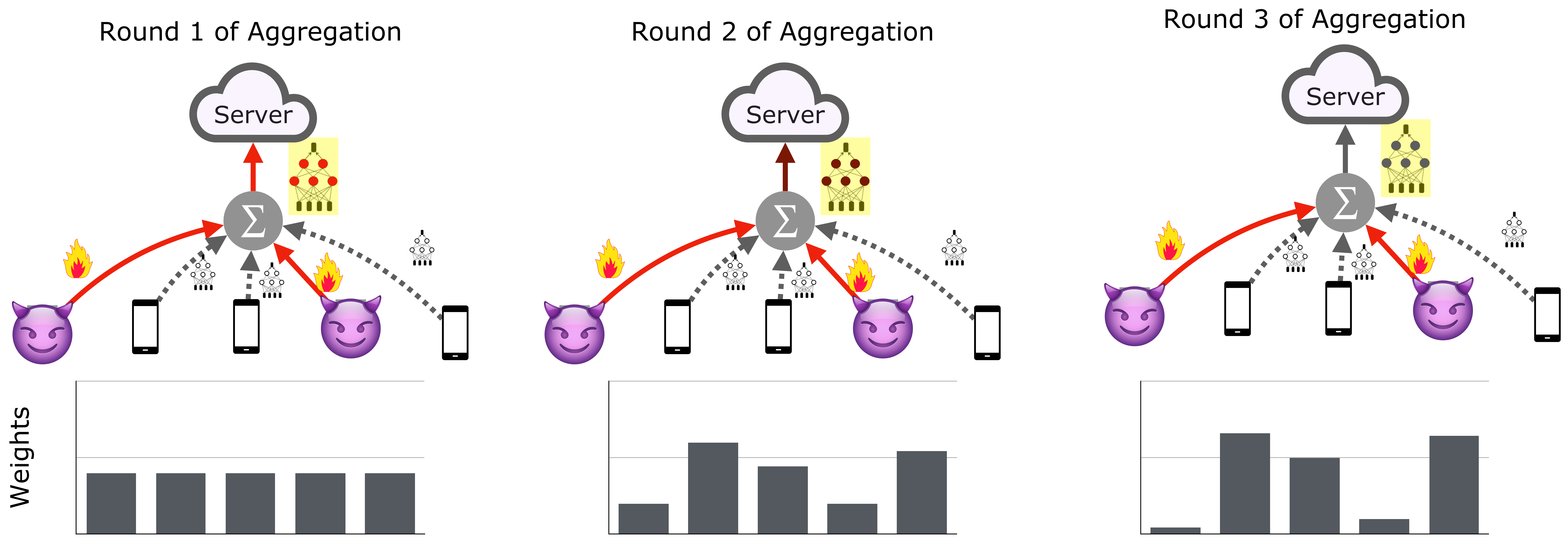


Step 2 of 3: Clients perform some local SGD steps on their local data



So far, same as federated averaging

Step 3 of 3: Aggregate with multiple rounds of secure average (weights β_i from the Smoothed Weiszfeld Algorithm)



See the paper for:

IEEE TRANSACTIONS ON
SIGNAL PROCESSING

2023

TSP Volume 70 | 2022

The Tension Between Robustness and Heterogeneity: Heterogeneity is a key property of federated learning. The distribution D_i of device i can be quite different from the distribution D_j of some other device j, reflecting the heterogeneous data generated by a diverse set of users.

To analyze the effect of heterogeneity on robustness, consider the simplified scenario of robust mean estimation in Huber's contamination model [34]. Here, we wish to estimate the mean $\mu \in \mathbb{R}^d$ given samples $w_1, \ldots, w_m \sim (1-\rho)\mathcal{N}(\mu, \sigma^2 I) + \rho Q$, where Q denotes some outlier distribution that ρ-fraction of the points (designated as outliers) are drawn from. Any aggregate $\bar{w}$ must satisfy the lower bound $\|\bar{w} - \mu\|^2 \geq \Omega(\sigma^2 \max\{\rho^2, d/m\})$ with constant probability [69, Theorem 2.2]. In the federated learning setting, more heterogeneity corresponds to a greater variance σ^2 among the inlier points, implying a larger error in mean estimation. This suggests a tension between robustness and heterogeneity, where increasing heterogeneity makes robust mean estimation harder in terms of ℓ_2 error.

In this work, we strike a compromise between robustness and heterogeneity by considering a family $\mathcal{D}$ of allowed data

Discussion on ***heterogeneity***

Convergence analysis

Convergence: We now analyze RFA where the local SGD updates are equipped with "tail-averaging" [73] so that $w_i^{(t+1)} = (2/\tau)\sum_{k=\tau/2}^{\tau} w_{i,k}^{(t)}$ is averaged over the latter half of the trajectory of iterates instead of line 9 of Algorithm 1. We show that this variant of RFA converges up to the dissimilarity level $\Omega = \Omega_X \Omega_{Y|X}$ when the corruption level $\rho < 1/2$.

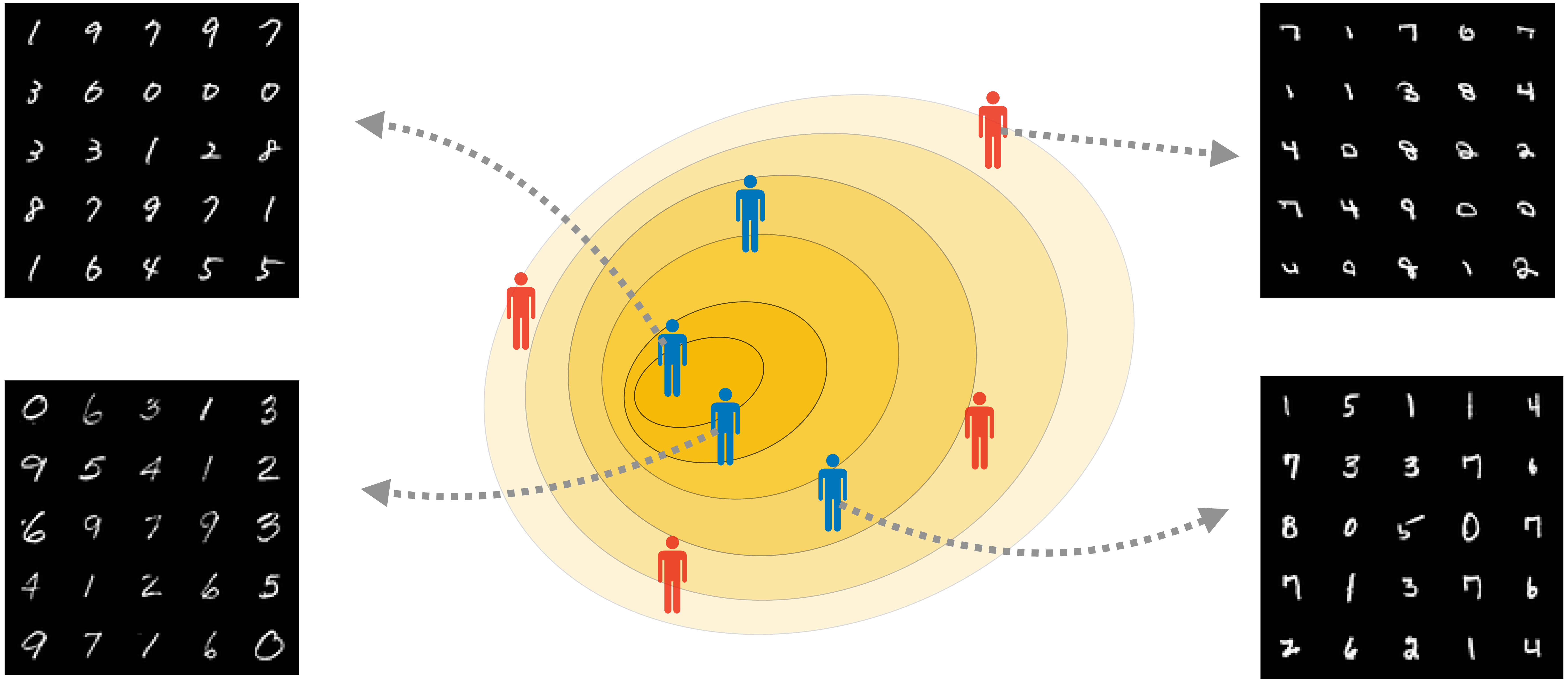
Theorem 4: Consider F defined in (7) and suppose the corruption level satisfies $\rho < 1/2$. Consider Algorithm 1 run for T outer iterations with a learning rate $\gamma = 1/(2R^2)$, and the local updates are run for τ_t steps in outer iteration t with tail averaging. Fix $\delta > 0$ and $\theta \in (\rho, 1/2)$, and set the number of devices per iteration, m as

$$m \geq \frac{\log(T/\delta)}{2(\theta - \rho)^2}. \tag{11}$$

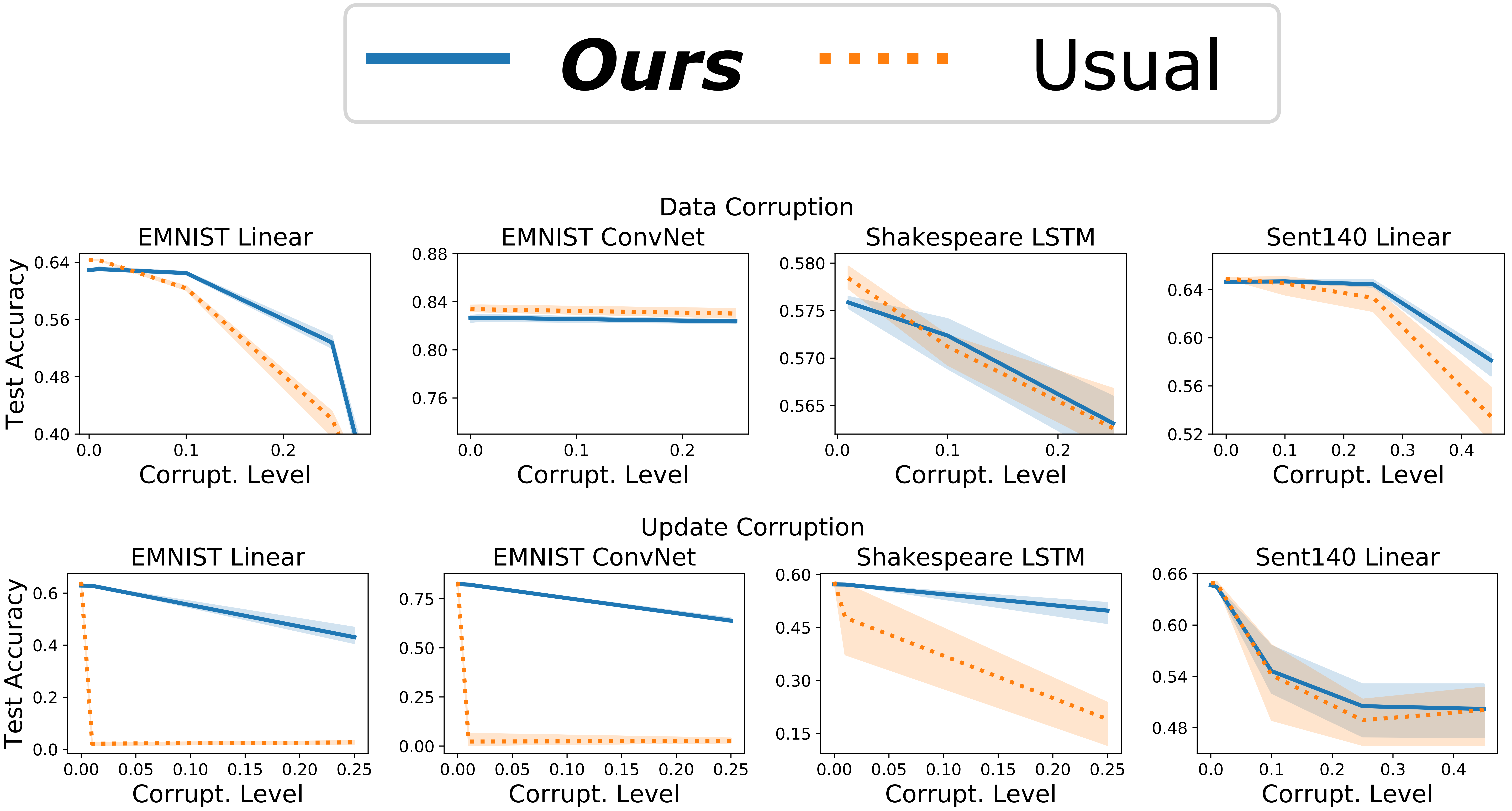
Define $C_\theta := (1-2\theta)^{-2}$, $w^\star = \arg\min F$, $F^\star = F(w^\star)$, $\kappa := R^2/\mu$ and $\Delta_0 := \|w^{(0)} - w^\star\|^2$. Let $\tau \geq 4\kappa \log(128 C_\theta \kappa)$. We have that the event $\mathcal{E} = \bigcap_{t=0}^{T-1}\{|S_t \cap \mathcal{C}| \leq \theta m\}$ holds with probability at least $1-\delta$. Further, if $\tau_t = 2^t \tau$ for each iteration t, then the output $w^{(T)}$ of Algorithm 1 satisfies,

$$\mathbb{E}\left[\|w^{(T)}) - w^\star\|^2 \,\middle|\, \mathcal{E}\right] \leq \frac{\Delta_0}{2^T} + C C_\theta \left(\frac{d\sigma^2 T}{\mu \tau 2^T} + \frac{\epsilon^2}{m^2} + \Omega^2\right)$$

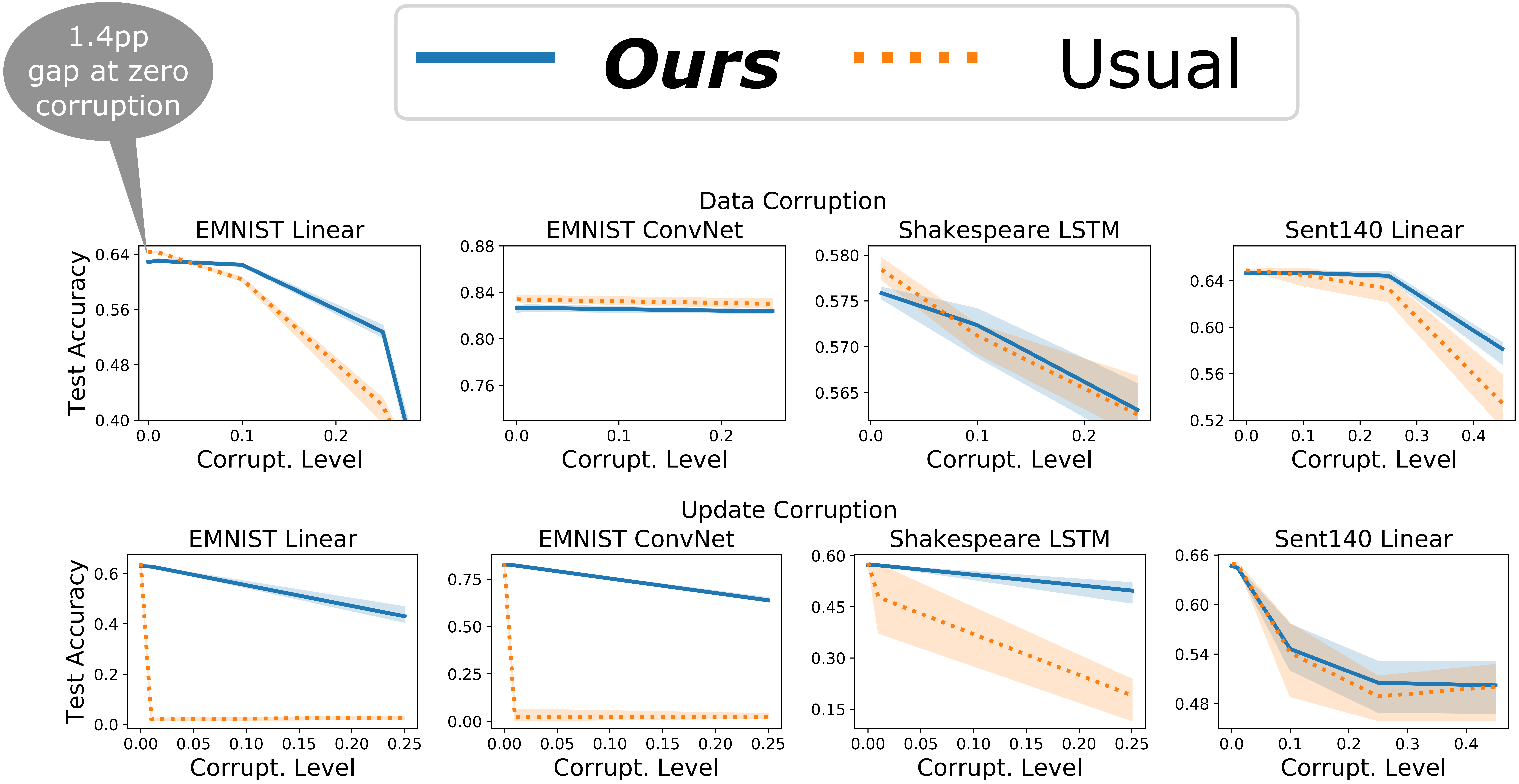
Experiments and Improvements



-


Ours
Usual
Data Corruption
EMNIST Linear
EMNIST ConvNet
Shakespeare LSTM
Sent140 Linear
Test Accuracy
Corrupt. Level
Update Corruption
EMNIST Linear
EMNIST ConvNet
Shakespeare LSTM
Sent140 Linear
Test Accuracy
Corrupt. Level

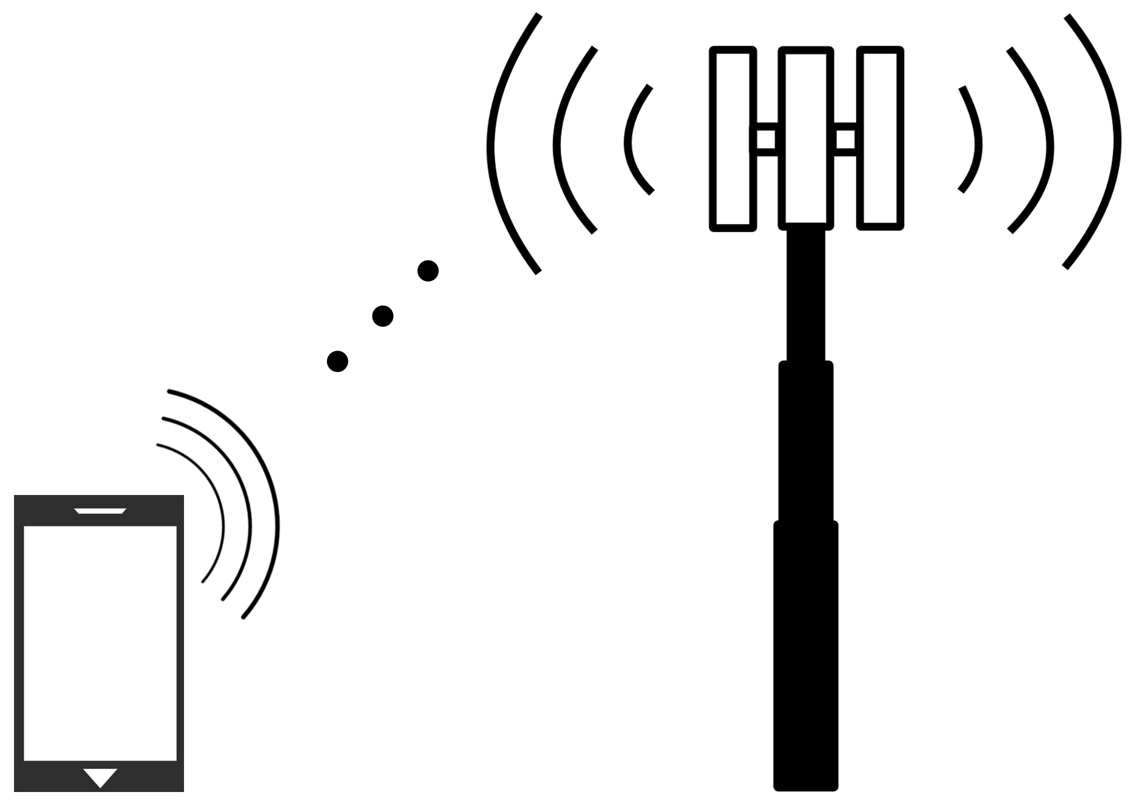
-

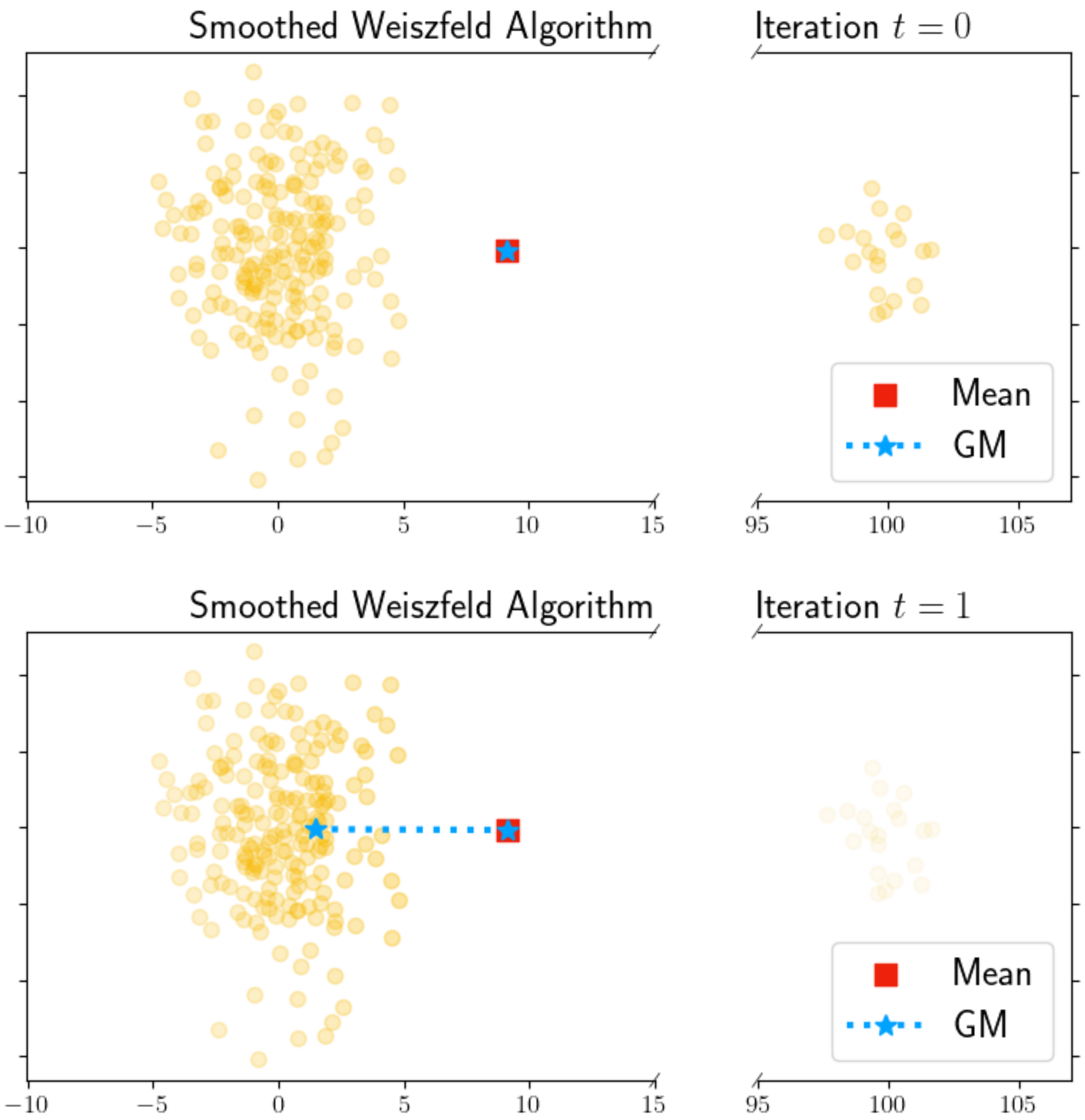

1.4pp
gap at zero
corruption
Ours
Usual
Data Corruption
EMNIST Linear
EMNIST ConvNet
Shakespeare LSTM
Sent140 Linear
Test Accuracy
Corrupt. Level
Update Corruption
EMNIST Linear
EMNIST ConvNet
Shakespeare LSTM
Sent140 Linear
Test Accuracy
Corrupt. Level

Reducing the communication cost

One round of our algorithm $\Longrightarrow$ 3-5 rounds of communication

Due to iterations of the smoothed Weiszfeld algorithm



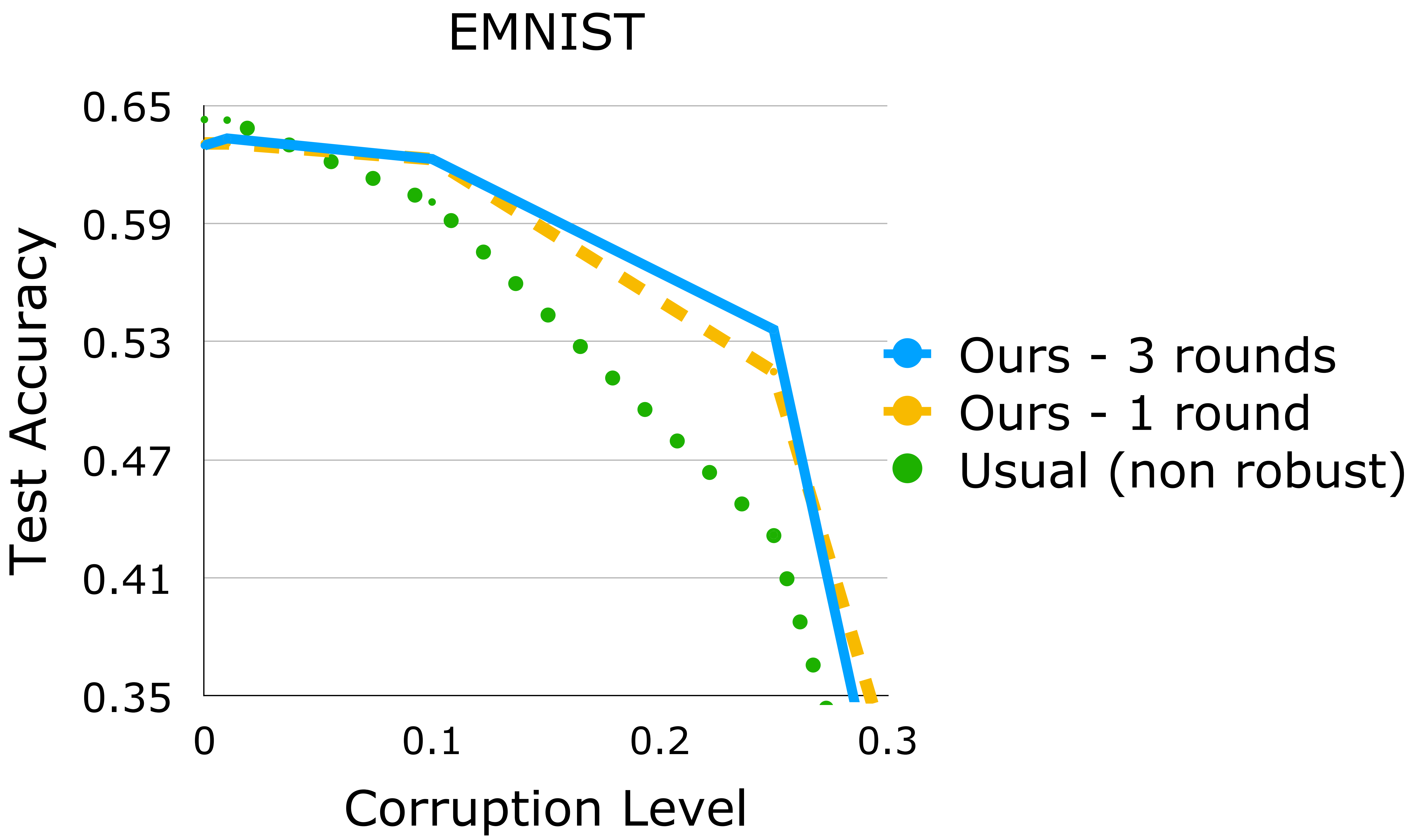


Does 1 round of communication improve robustness?

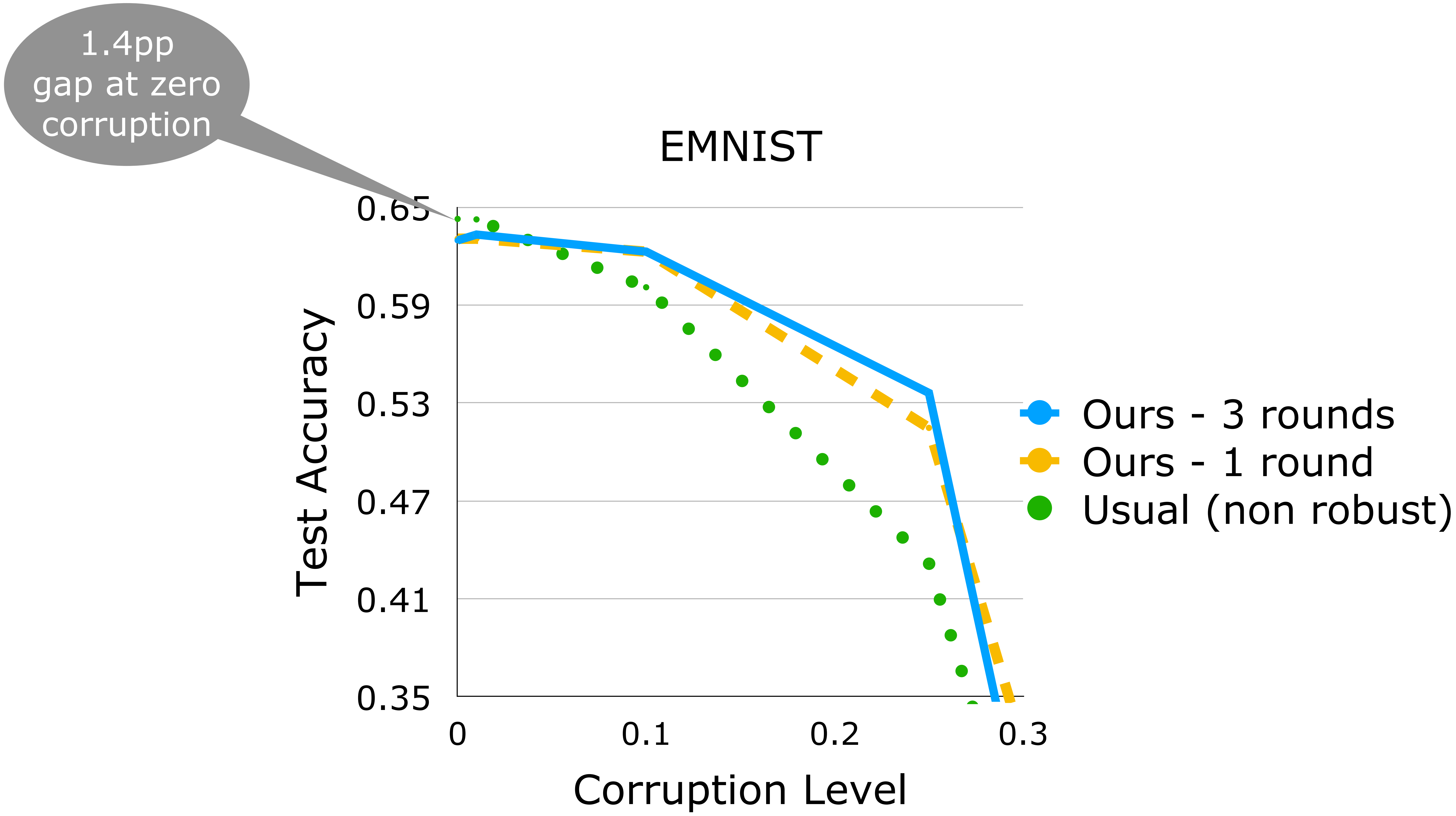
$$\beta_i = \frac{1}{\max\{\|w_i\|_2,\ \nu\}}$$

$$z = \frac{\sum_i \beta_i w_i}{\sum_i \beta_i}$$

1 communication round already improves robustness

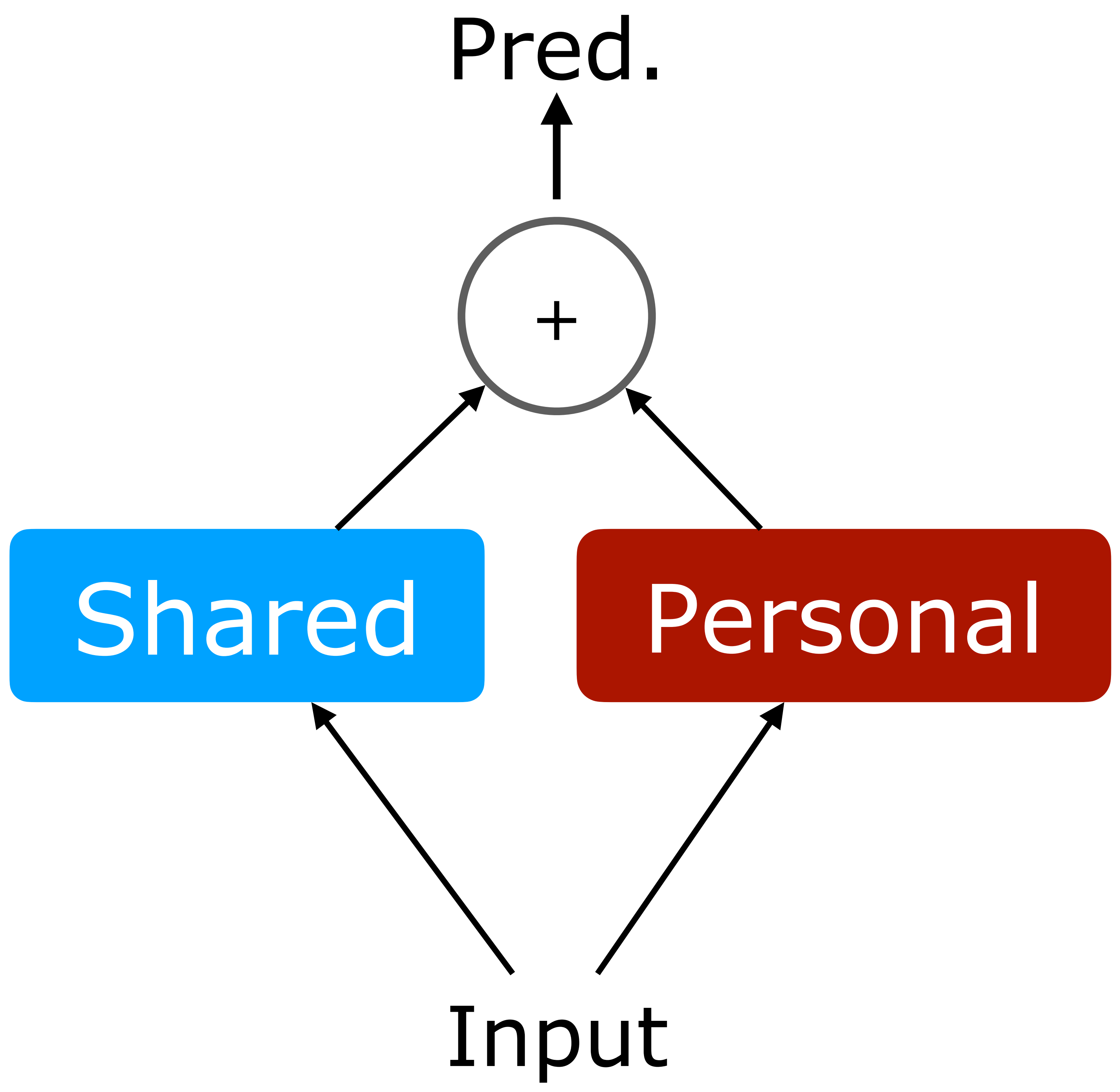


How do we get rid of this gap?



Model personalization

The model has a global component
and a per-client component

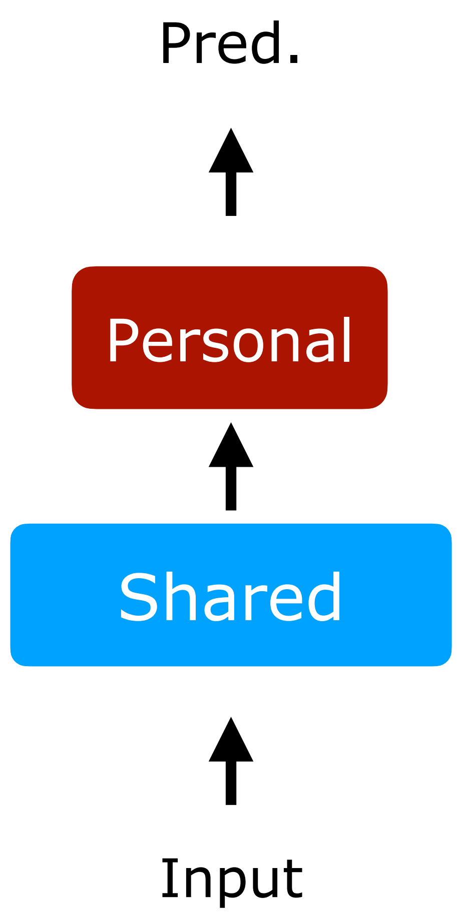


Shared Params u

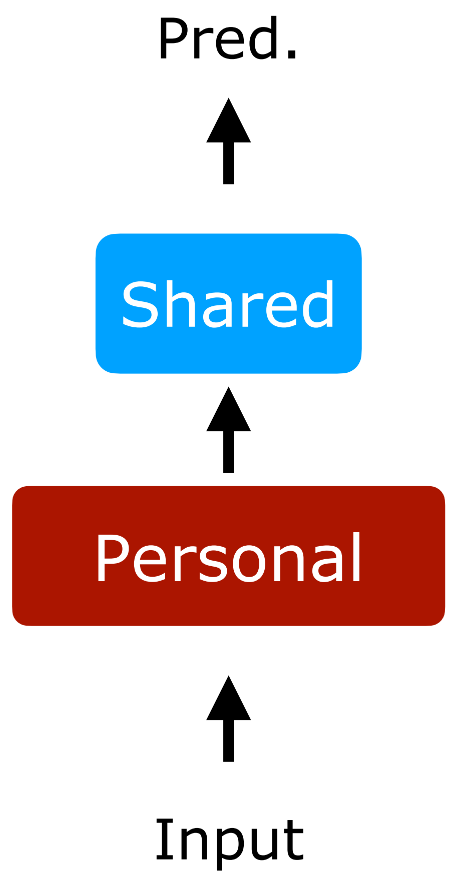
$+$ Personal Params v_i

$=$ Full model $w_i = (u, v_i)$

Personalization Architectures



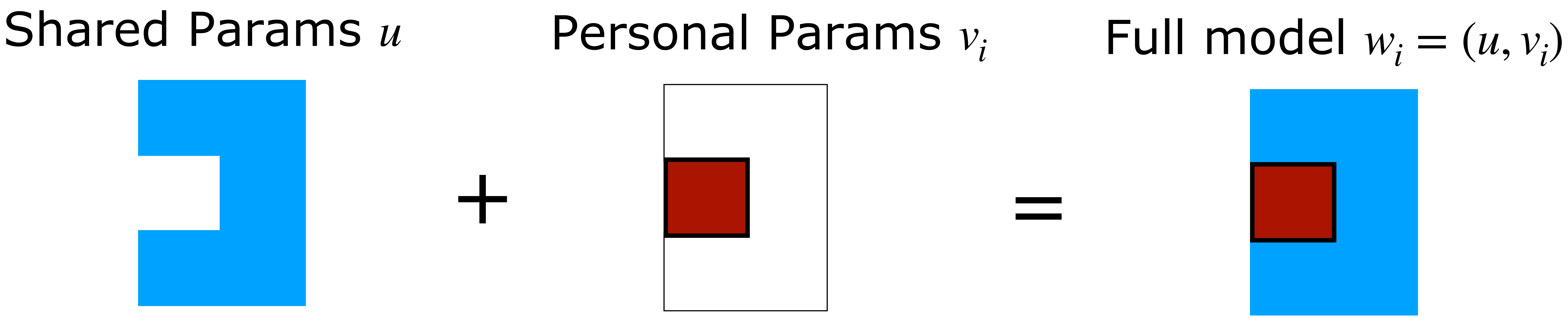
Arivazhagan et al. (2019)
Collins et al. (2021)



Liang et al. (2019)

Multi-task learning: Caruana (1997), Baxter (2000), Evgeniou & Pontil (2004), Collobert & Weston (2005), Argyriou et al. (2008), …

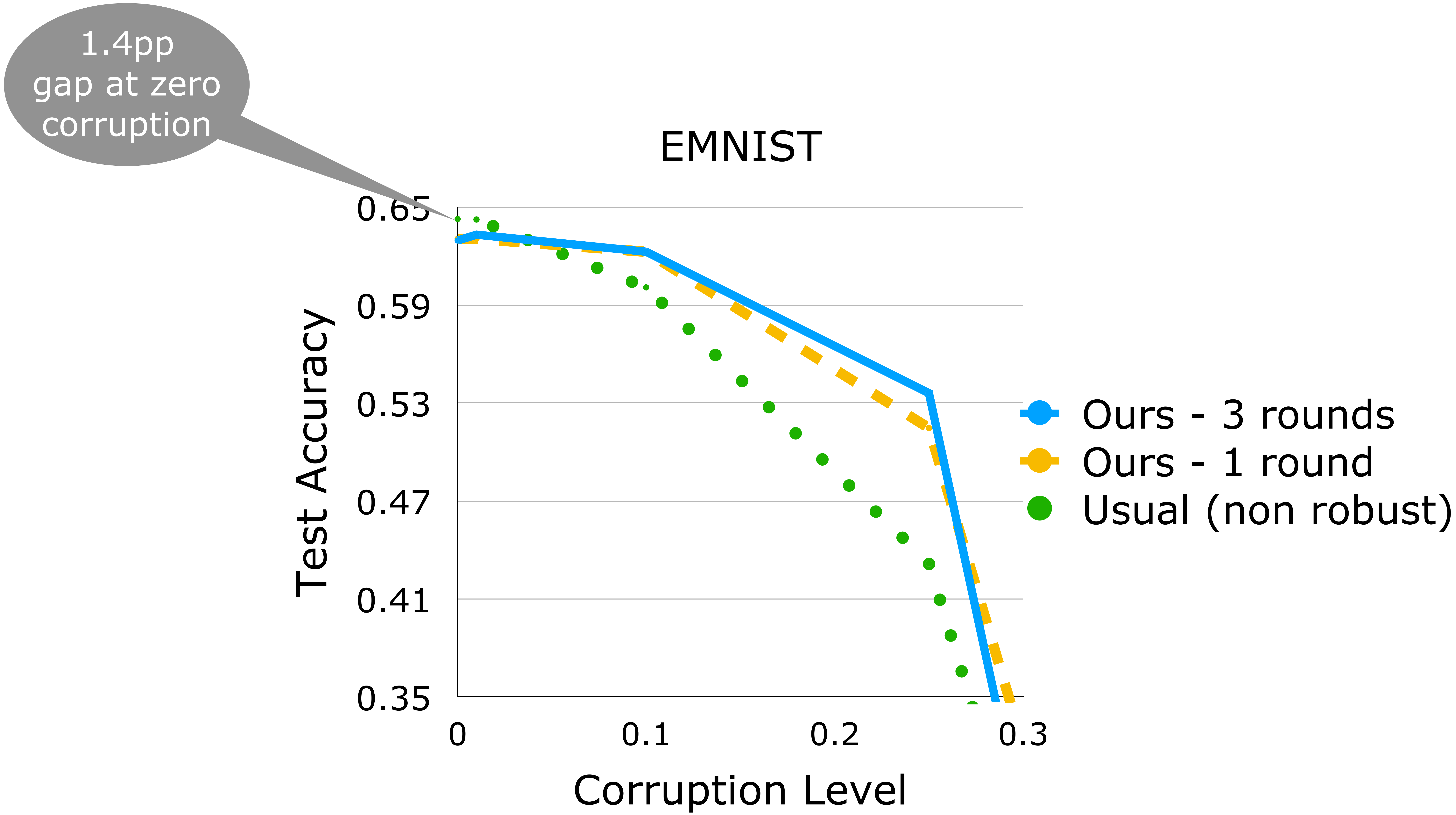
Optimization



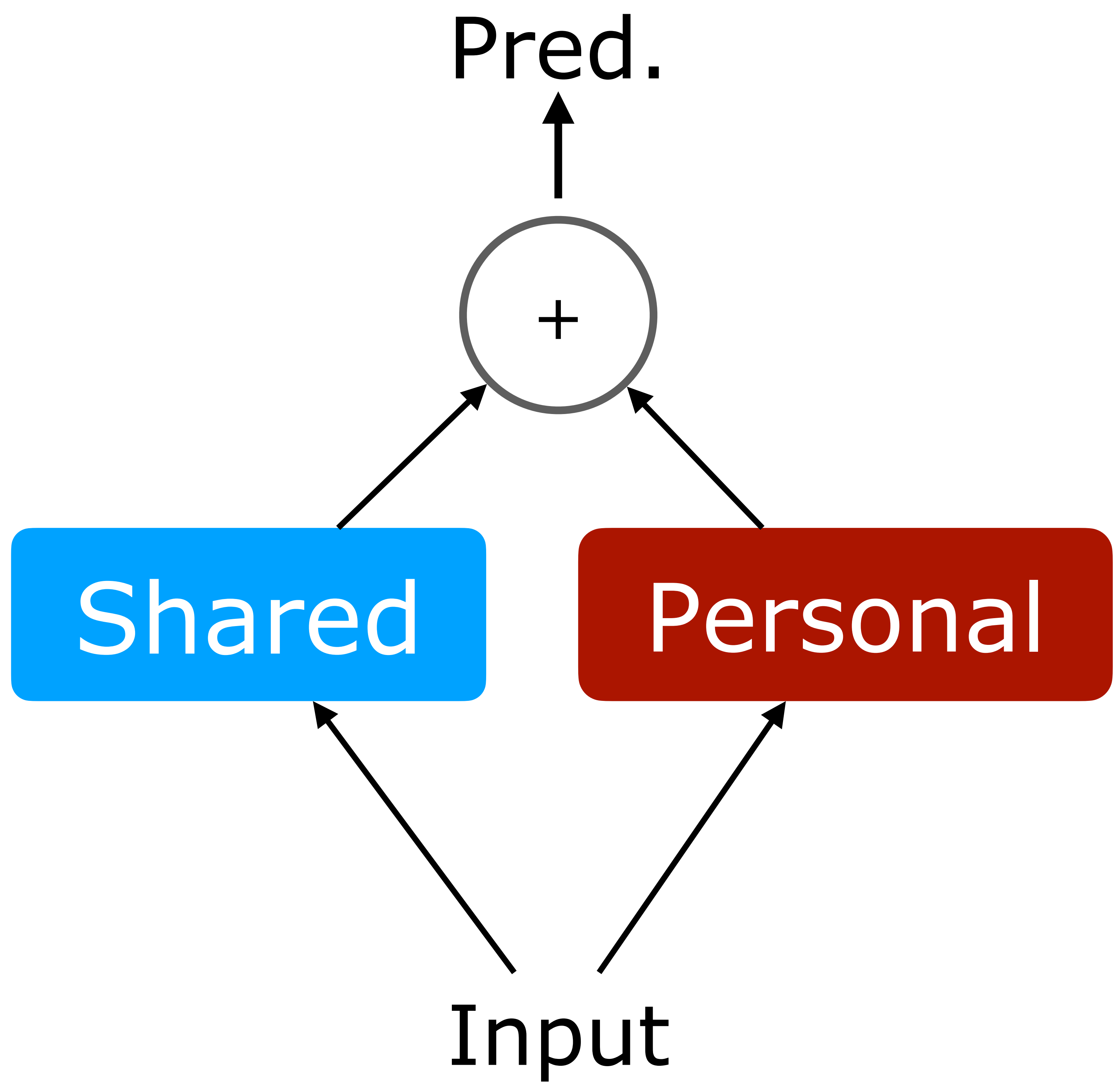
Shared part of the model *is updated with robust aggregation*

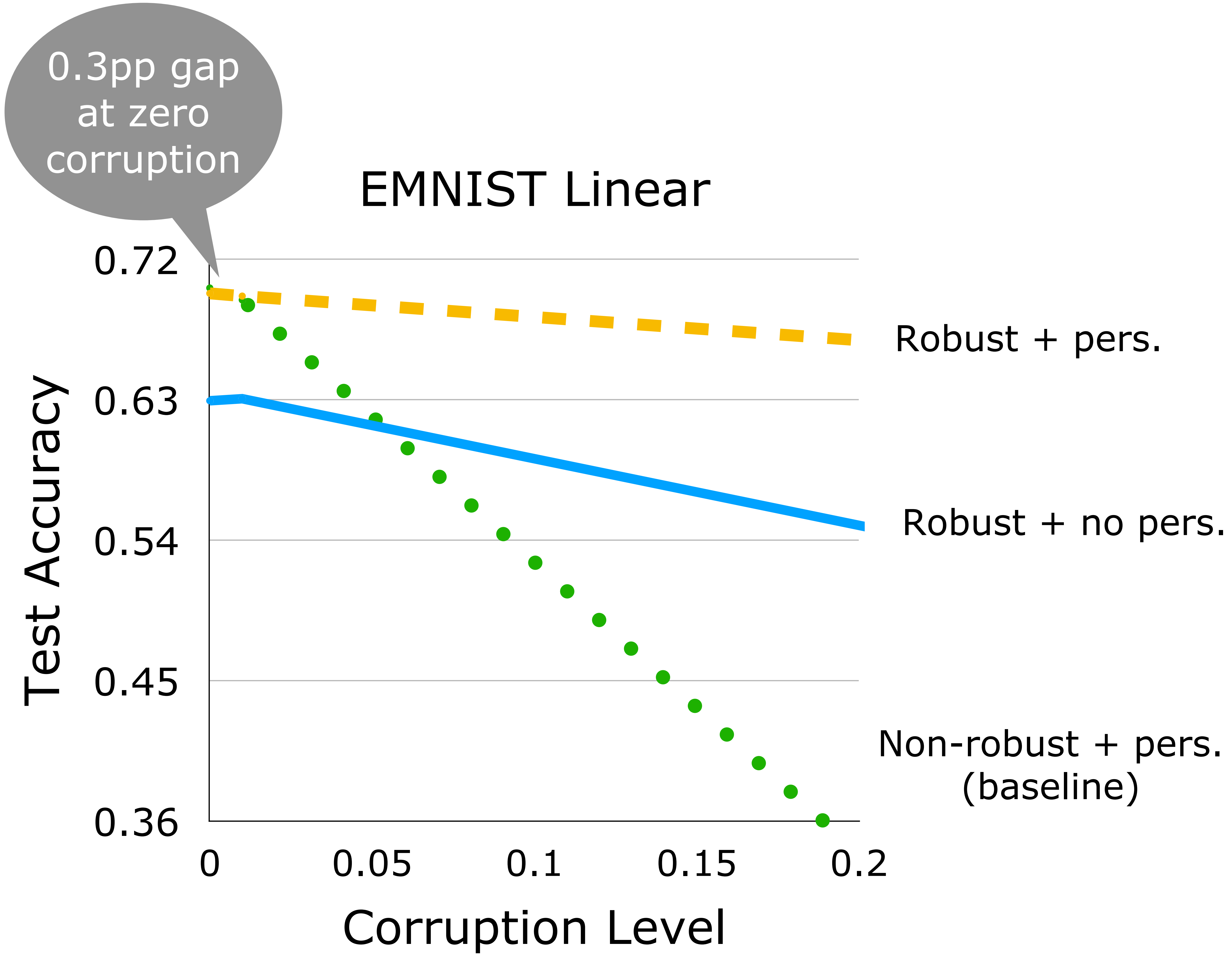
Personal part of the model *stays with the client*

Does personalization get rid of this gap?



Yes, we can improve robust aggregation with personalization!





In the literature:

Robust Federated Aggregation (RFA)

RFA is certifiably more robust to backdoor attacks

[Xie et al. (ICML 2021)]

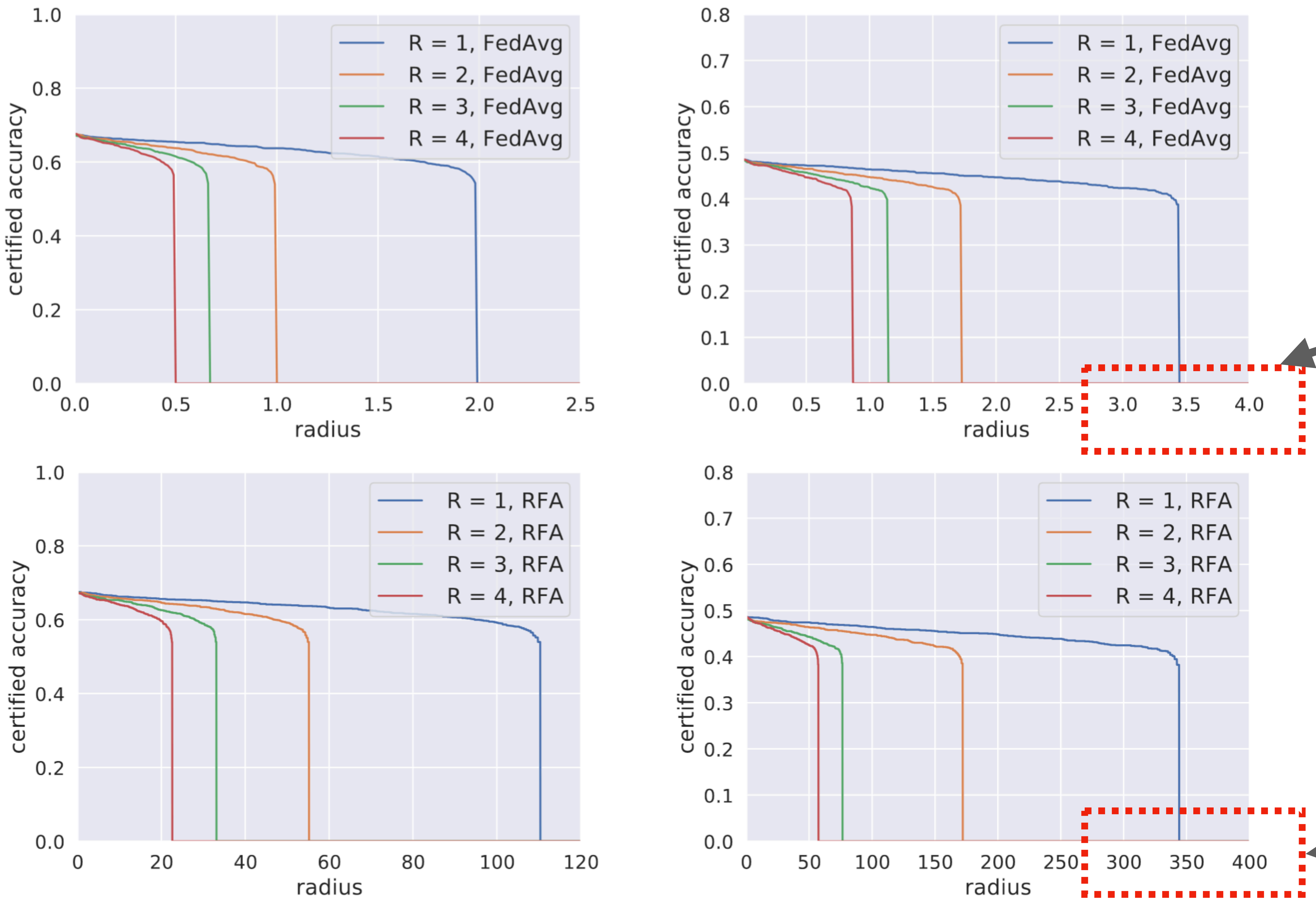


Figure 6. Certified accuracy on MNIST (left) and EMNIST (right) with different R when FL is trained under the robust aggregation RFA (Pillutla et al., 2019).

50-100x more robust!

Xie, Chen, Chen, Li. **CRFL: Certifiably Robust Federated Learning against Backdoor Attacks.** *ICML 2021*.

RFA is asymptotically strategy-proof

Strategy-proof: Can a device lie to bring the aggregate to a desired point?

With a large number of independent devices, RFA is approximately strategy-proof

On the Strategyproofness of the Geometric Median

El-Mahdi El-Mhamdi
Calicarpa, École Polytechnique

Sadegh Farhadkhani*
EPFL

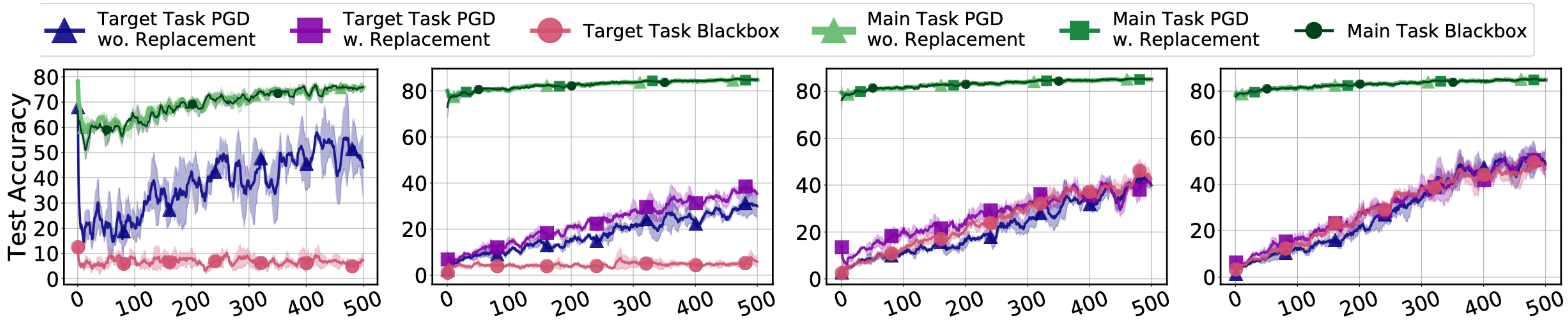
Rachid Guerraoui
EPFL

Lê-Nguyên Hoang*
Calicarpa, Tournesol

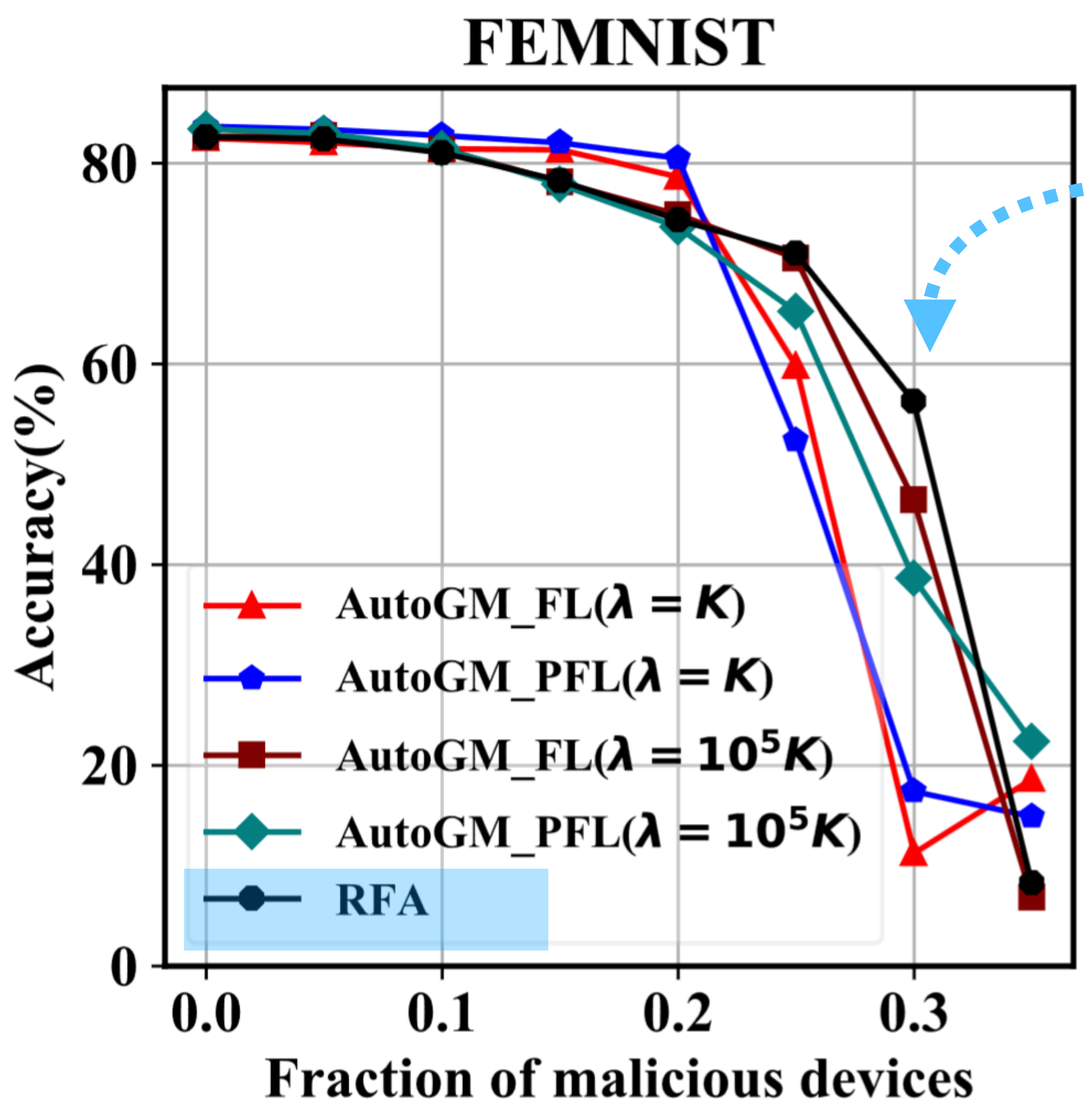
[AISTATS 2023]

RFA is a strong baseline

Wang et al.
(NeurIPS 2020)



(a) KRUM (b) MULTI-KRUM (c) RFA (d) NDC



Li et al. (IEEE Trans. Industrial Informatics 2023)

See also Sejwalkar et al. (IEEE Security & Privacy 2022), Jin & Li (Medical Image Analysis 2023), Li et al. (IEEE Trans. Big Data 2023), …

Algorithmic advances based on RFA

Park et al. (NeurIPS 2021): RFA + Entropy-based reweighting

Karimireddy et al. (ICLR 2022): RFA + Bucketing

Li et al. (IEEE Trans. Ind. Inform. 2023): RFA + adaptive weighting

Allouah et al. (AISTATS 2023): RFA + nearest neighbhors

⋮

Fast and differentiable geometric median

```python
import torch
from geom_median.torch import compute_geometric_median   # PyTorch API
# from geom_median.numpy import compute_geometric_median  # NumPy API

points = [torch.rand(d) for _ in range(n)]   # list of n tensors of shape (d,)
# The shape of each tensor is the same and can be arbitrary (not necessarily 1-dimensional)
weights = torch.rand(n)  # non-negative weights of shape (n,)
out = compute_geometric_median(points, weights)
# Access the median via `out.median`, which has the same shape as the points, i.e., (d,)
```

Install: **pip install geom-median**

Documentation: github.com/krishnap25/geom-median

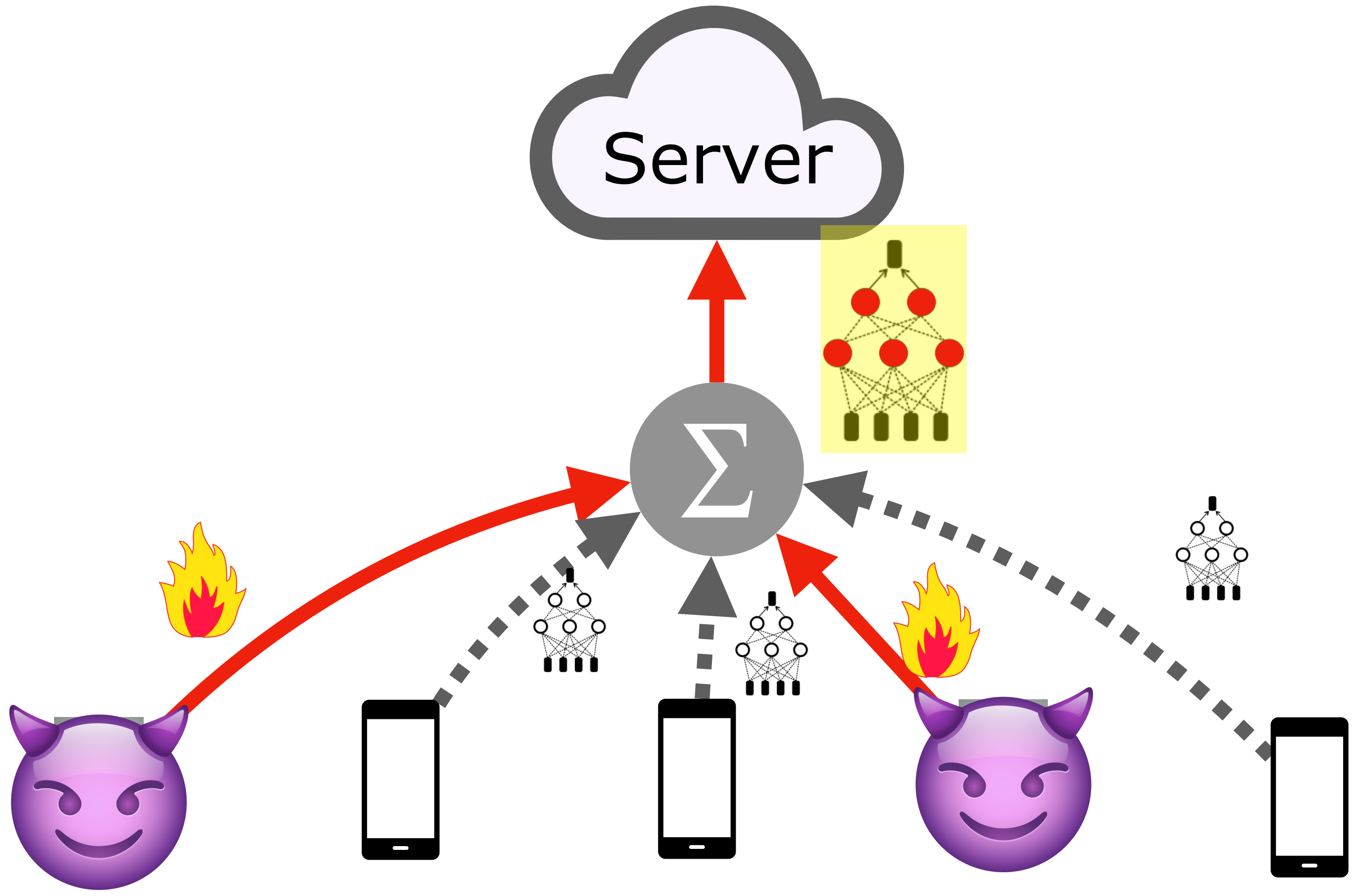
SCAN ME

GitHub Link

Summary

Paper:

Federated learning is ***not robust*** to poisoned updates

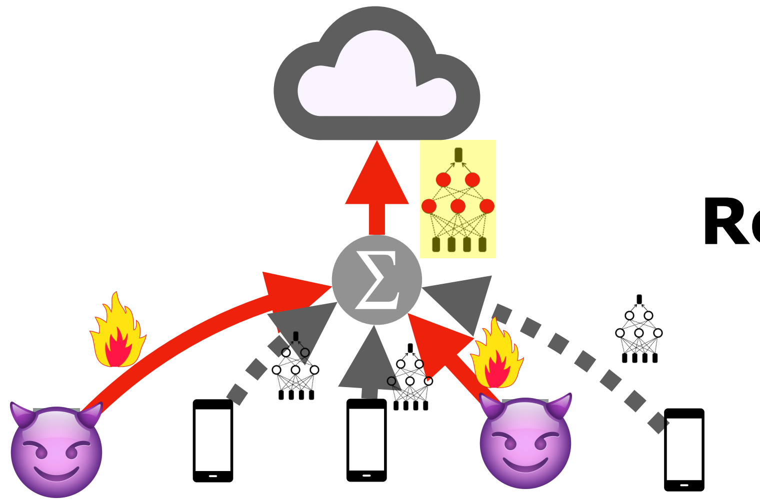


Paper:

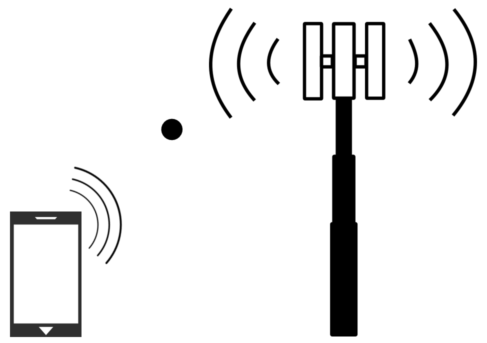
Summary

Federated learning is ***not robust*** to poisoned updates

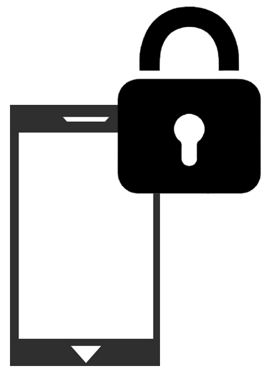
$$\mathrm{GM} = \arg\min_{z} \sum_{i=1}^{m} \|z - w_i\|_2$$



Robust



Communication efficient



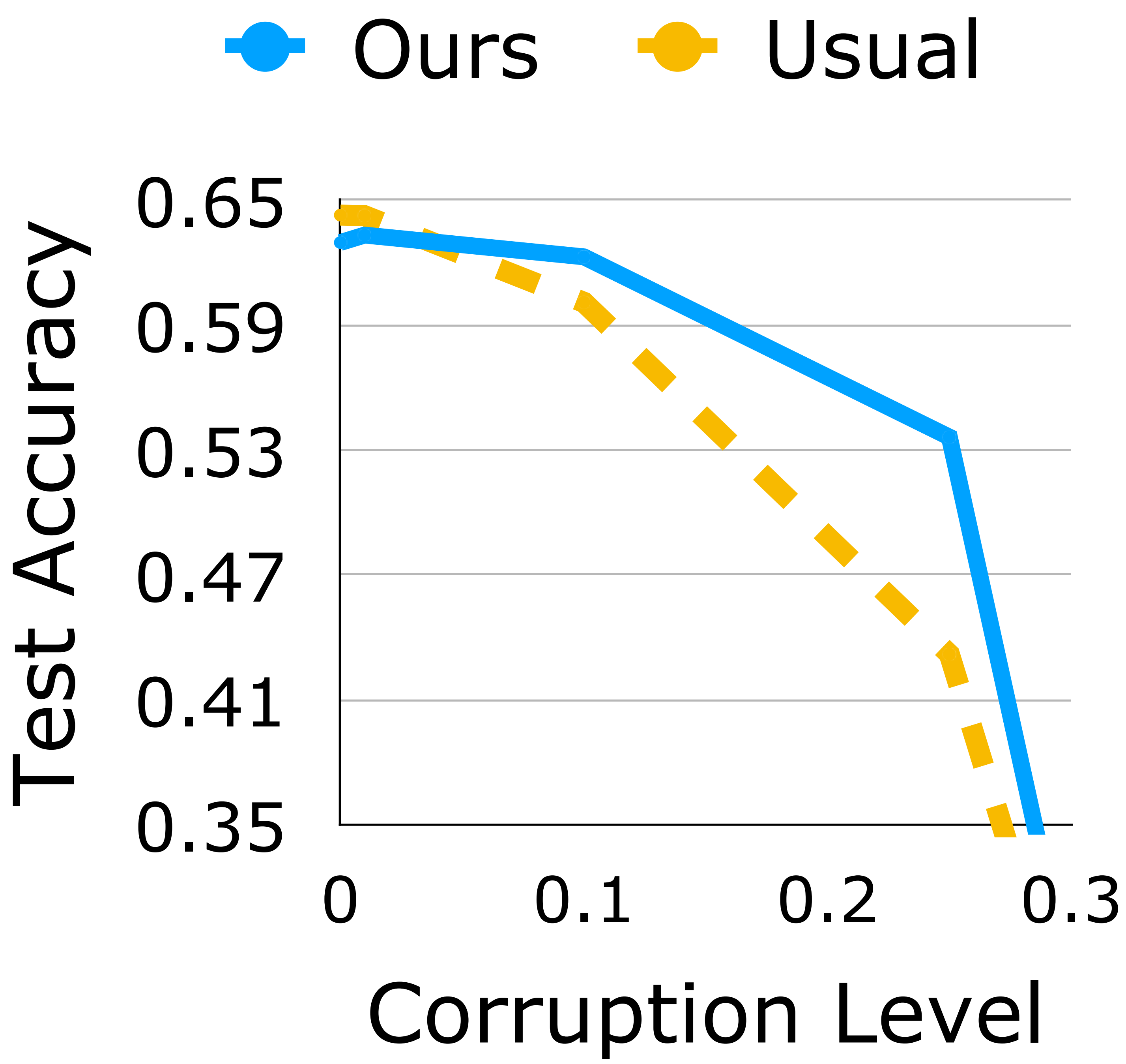
Secure aggregation

Summary

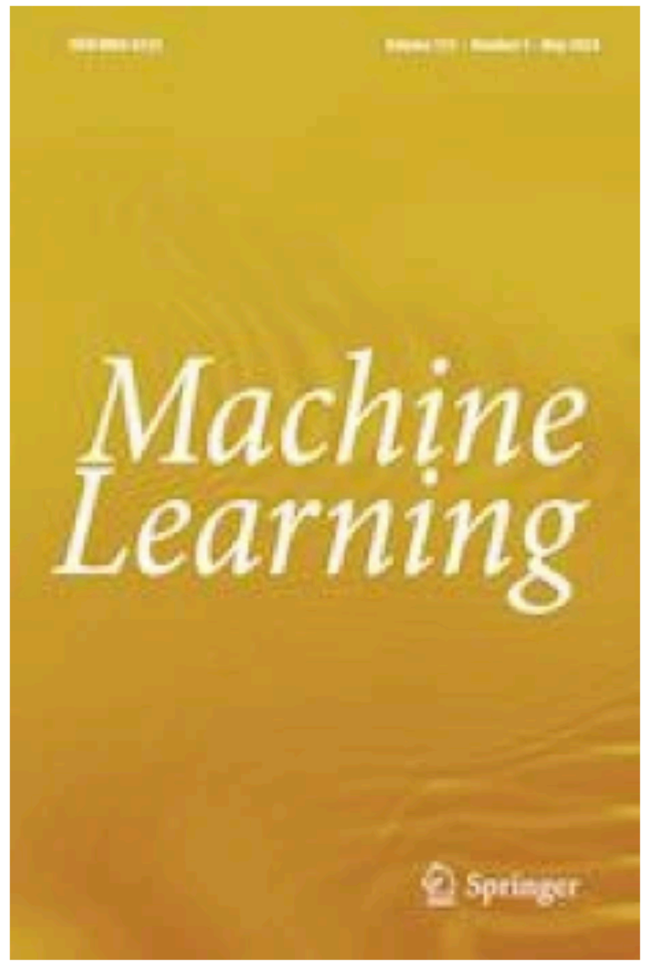
Federated learning is ***not robust*** to poisoned updates

$$\text{GM} = \arg\min_{z} \sum_{i=1}^{m} \|z - w_i\|_2$$

Our approach gives **greater robustness**



Heterogeneity, fairness, equity with differential privacy in federated learning



Paper:

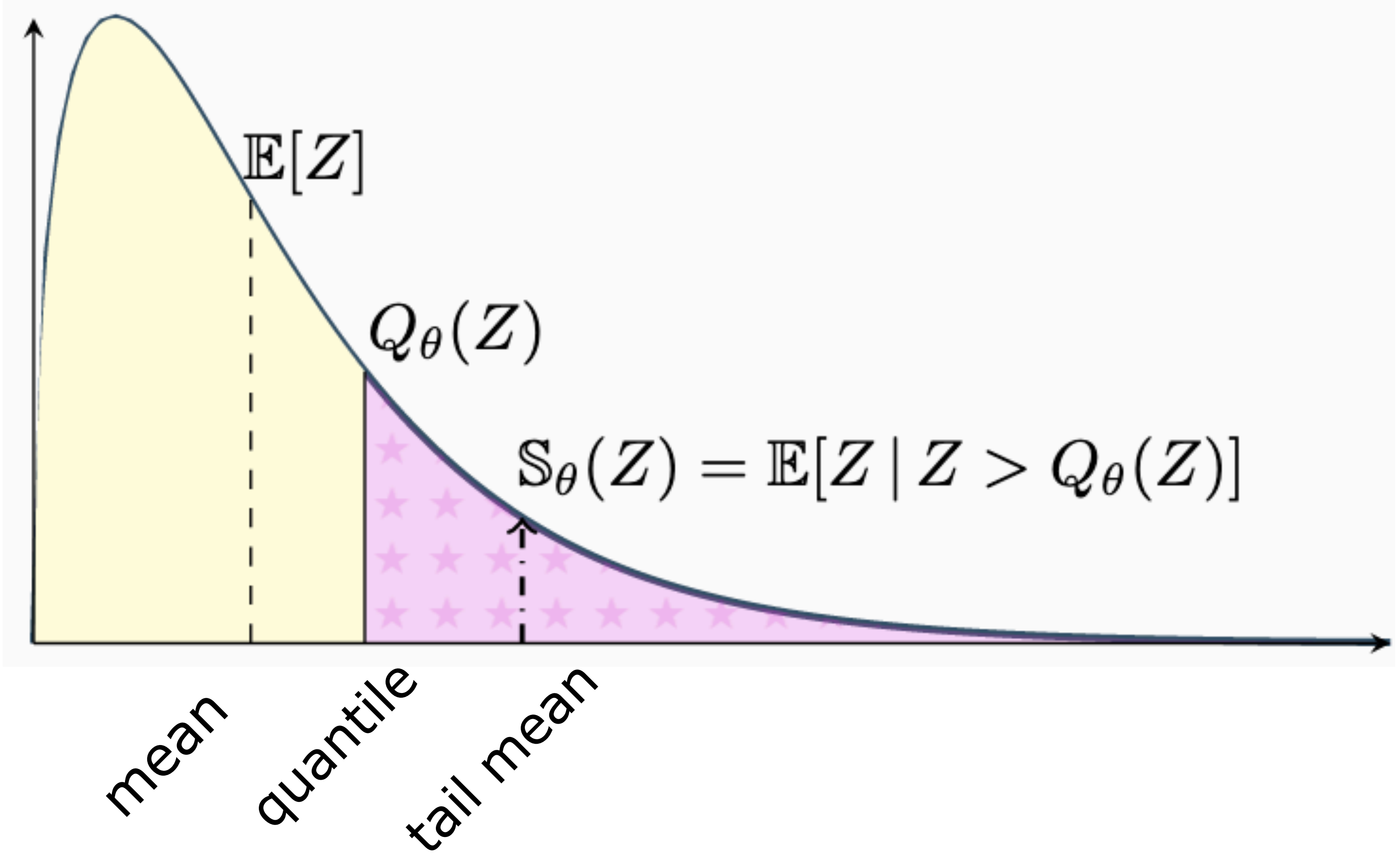
Volume 113, Issue 5

May 2024

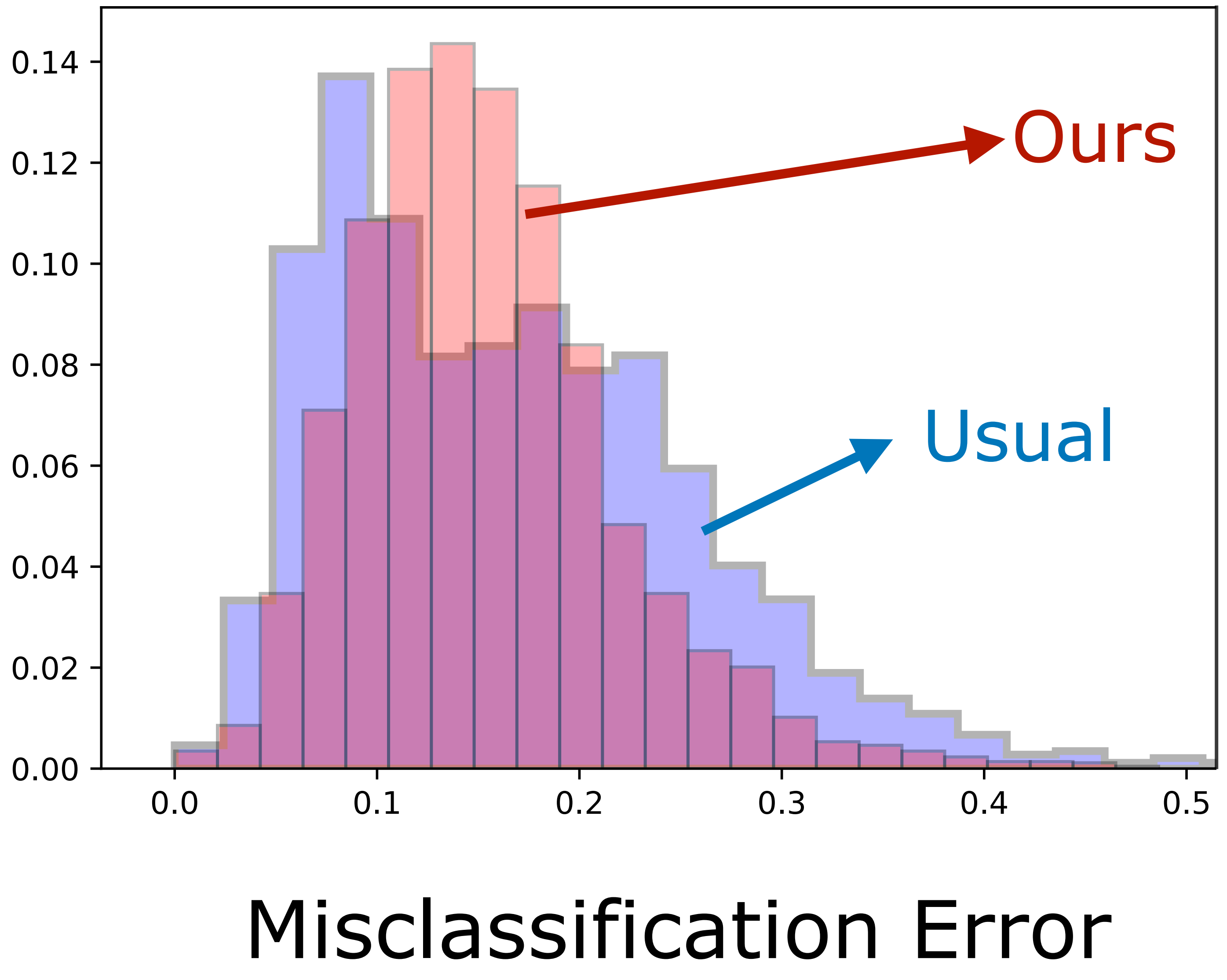
SCAN ME

Distribution shift $\Longrightarrow$ large tail errors

Minimize the tail error directly



We reduce tail error + support differential privacy



Thank you!

Paper: SCAN ME

Software

```
pip install geom-median
```

Code https://github.com/krishnap25/tRFA

@KrishnaPillutla