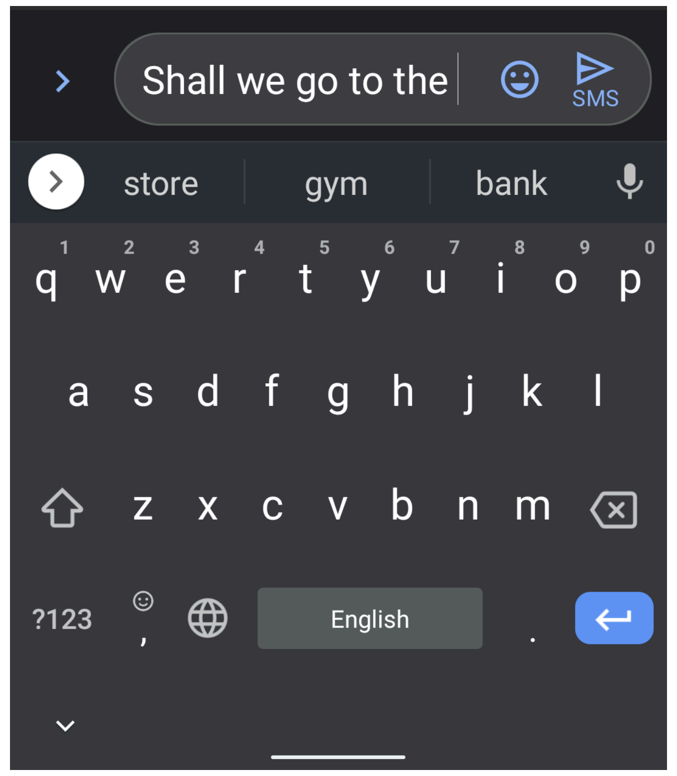
Federated Learning: Robustness, Heterogeneity and Optimization

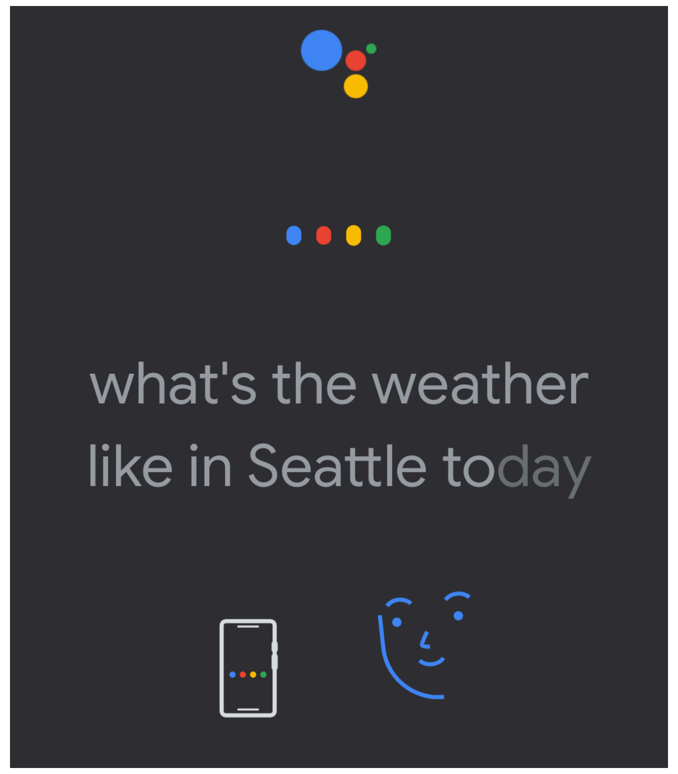
Krishna Pillutla

June 24, 2022

Committee: Zaid Harchaoui (Chair), Sham Kakade (Chair), Yejin Choi, Lalit Jain (GSR), Kevin Jamieson, Jamie Morgenstern

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING





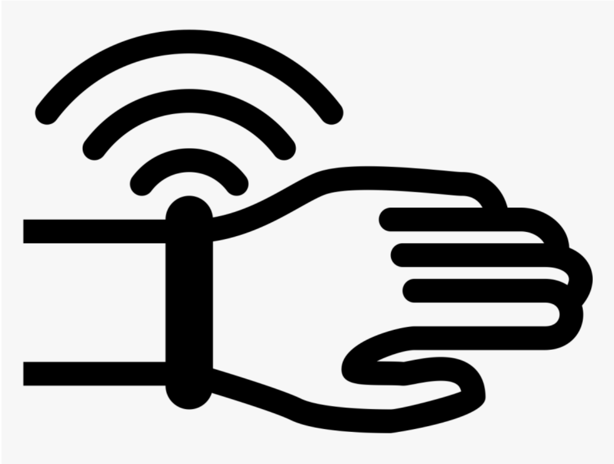
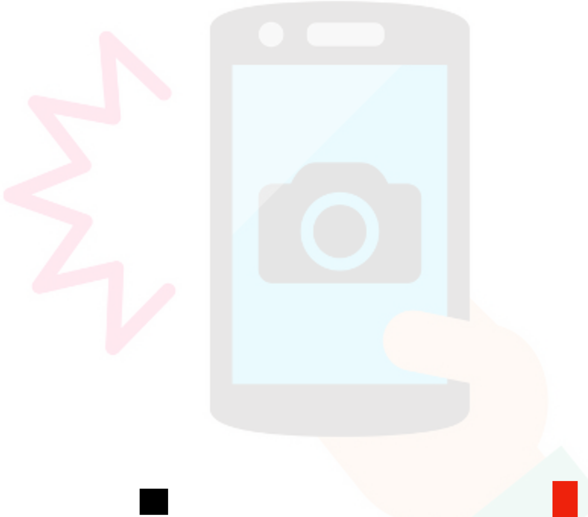


Image Credit: Robotics Business Review

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

Image Credit: Robotics Business Review

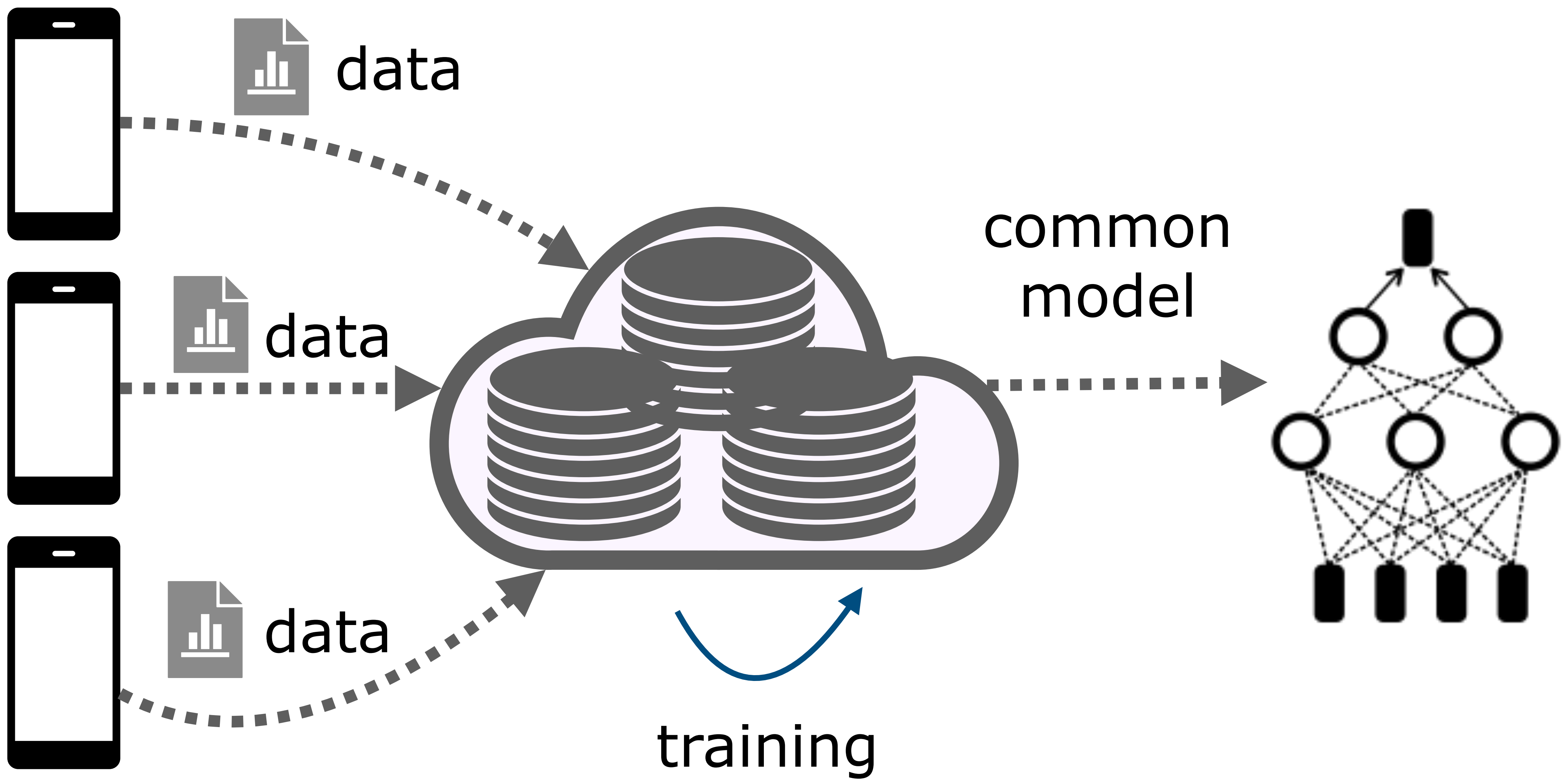


Data is decentralized and private

Rieke et al. NPJ Digit. Med. (2020)

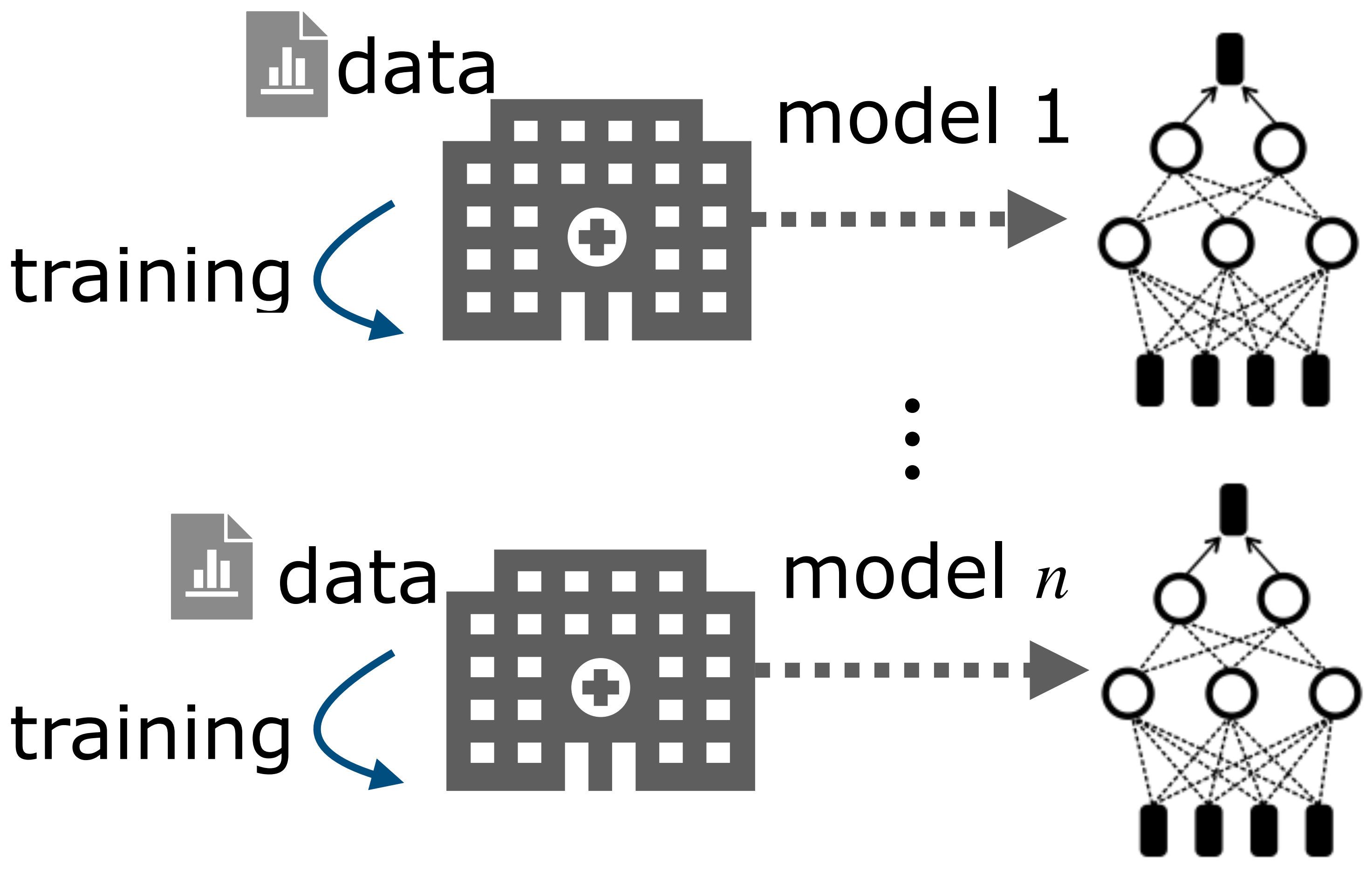
Image Credit: Wellcome

Datacenter



Datacenter

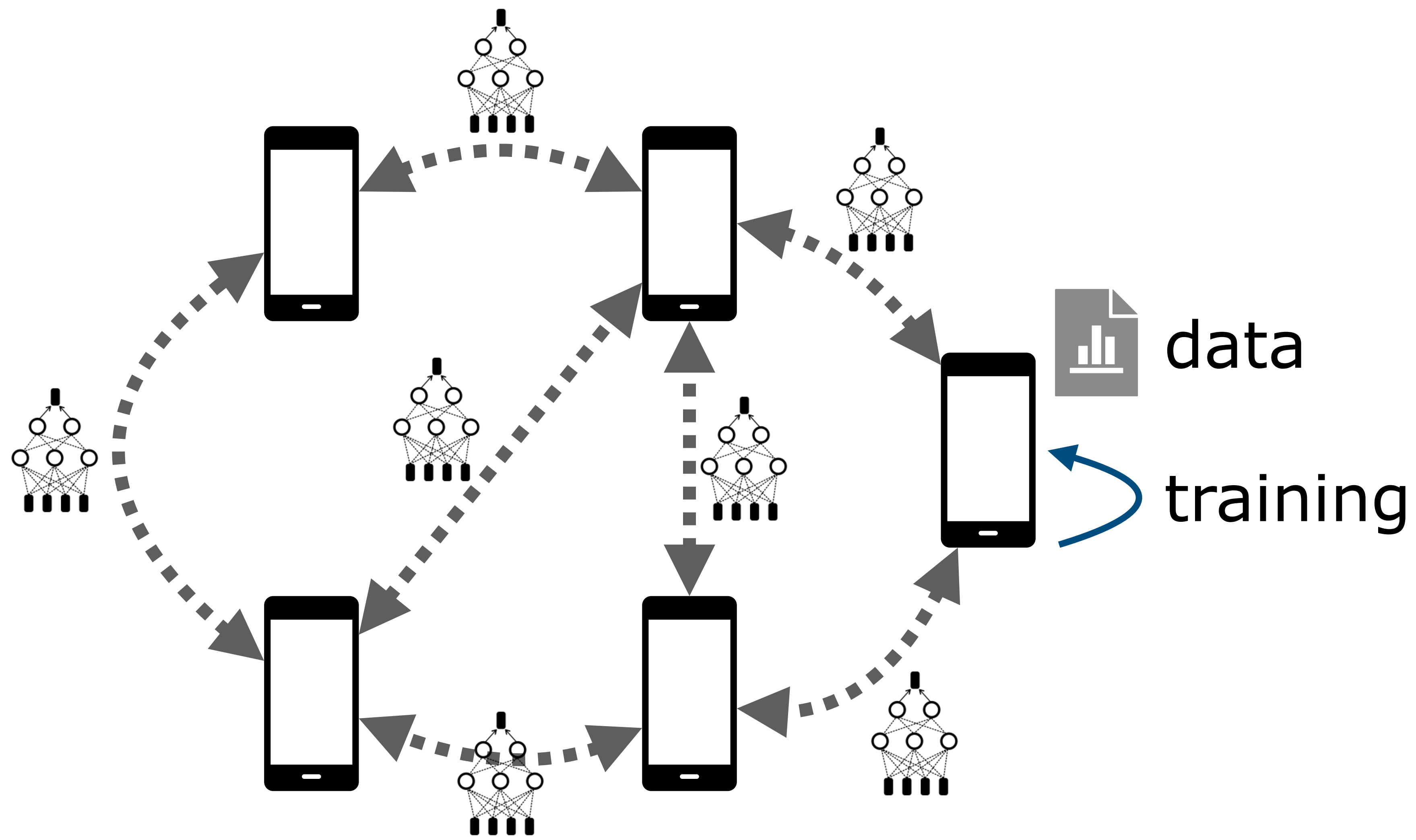
Non-collaborative



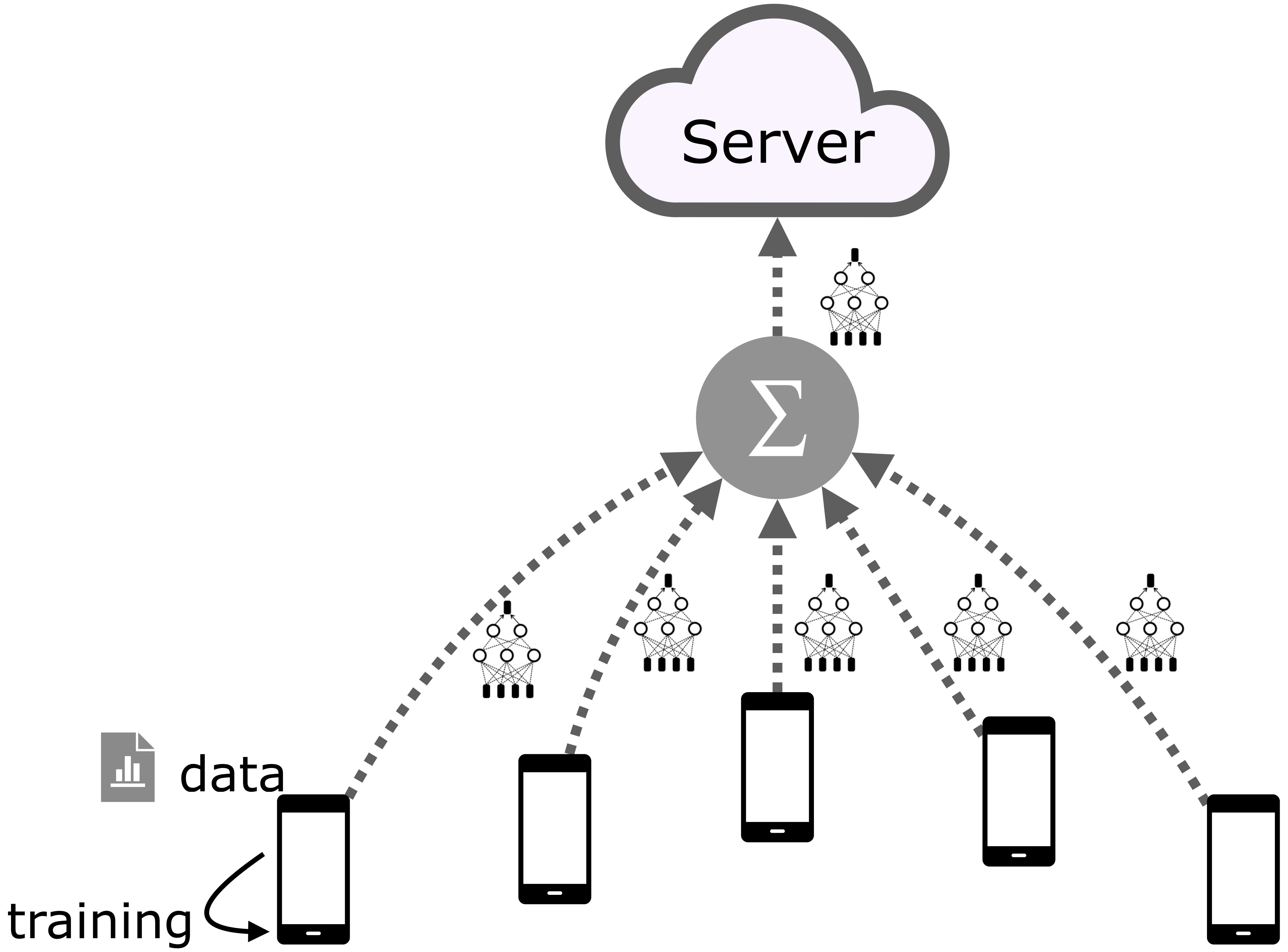
Datacenter

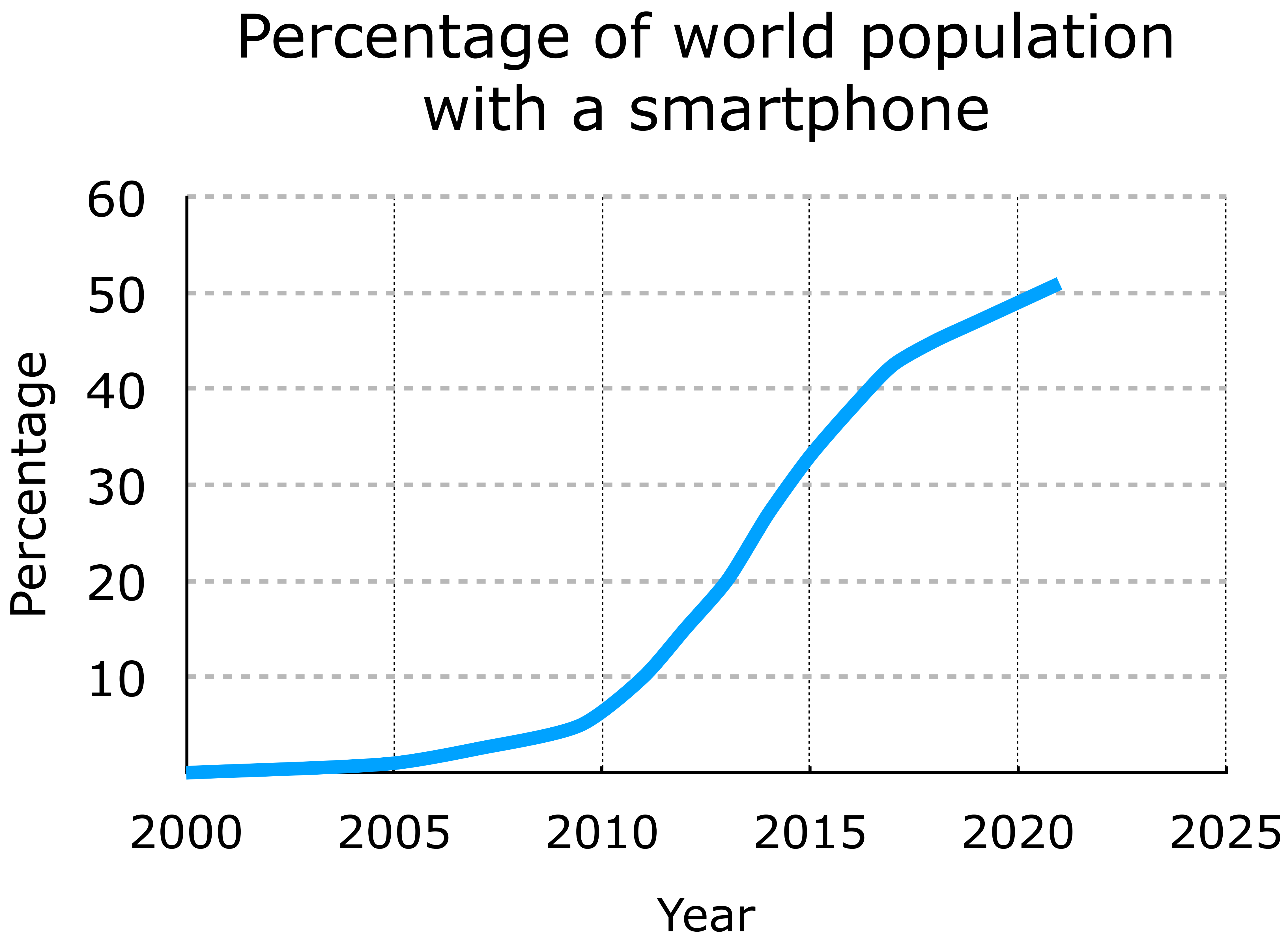
Non-collaborative

Peer-to-peer



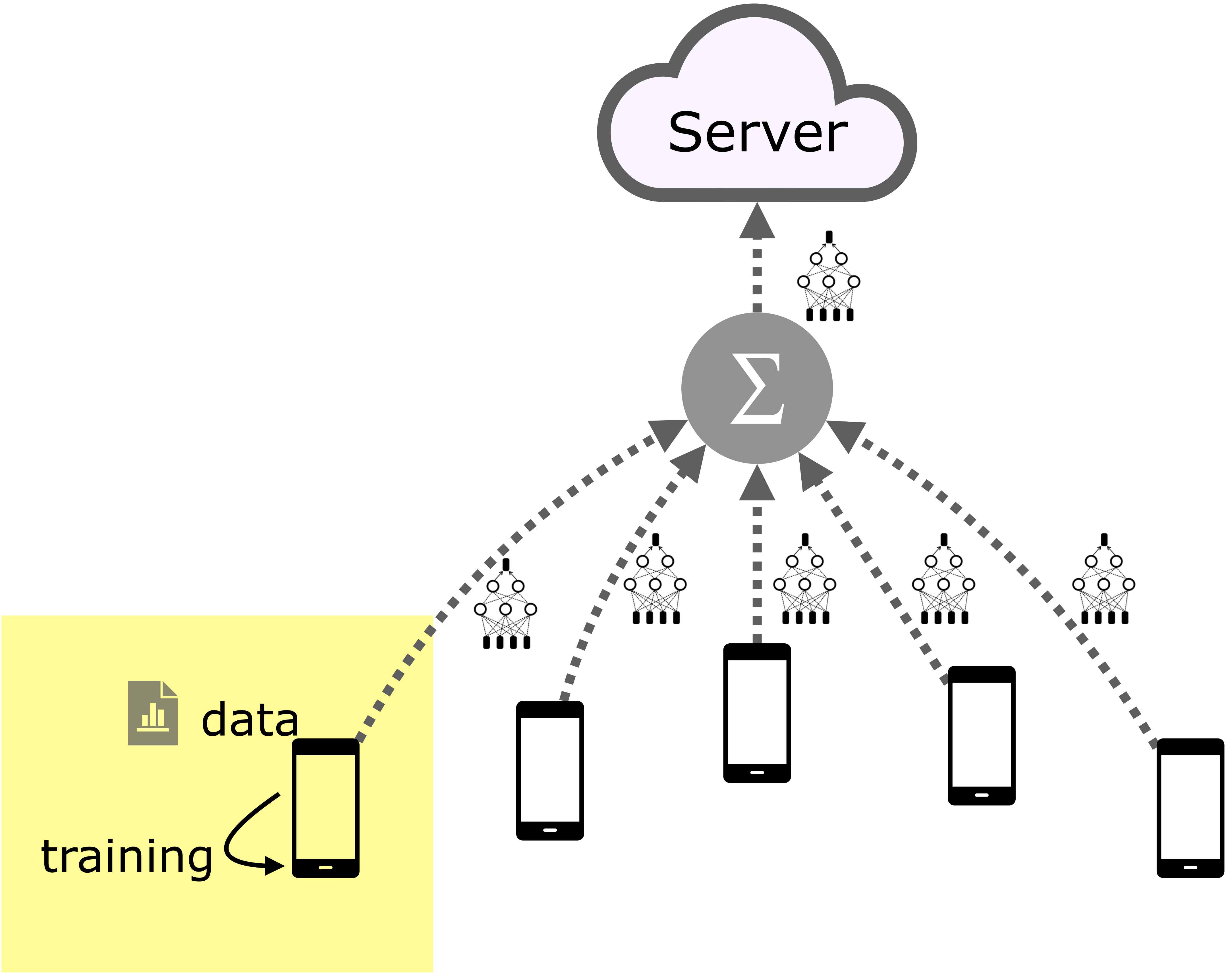
Federated Learning

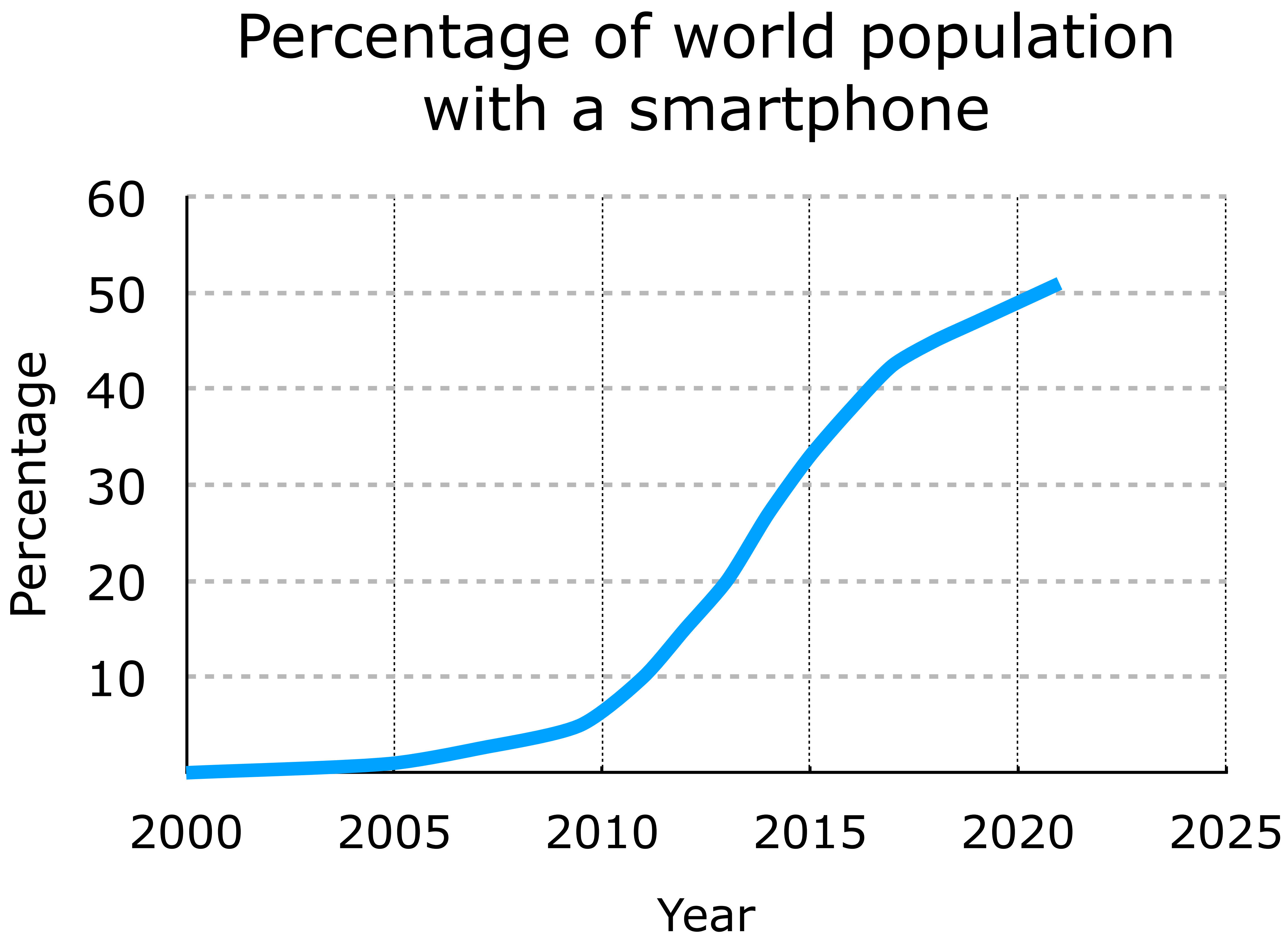




Data Credit: Business Wire

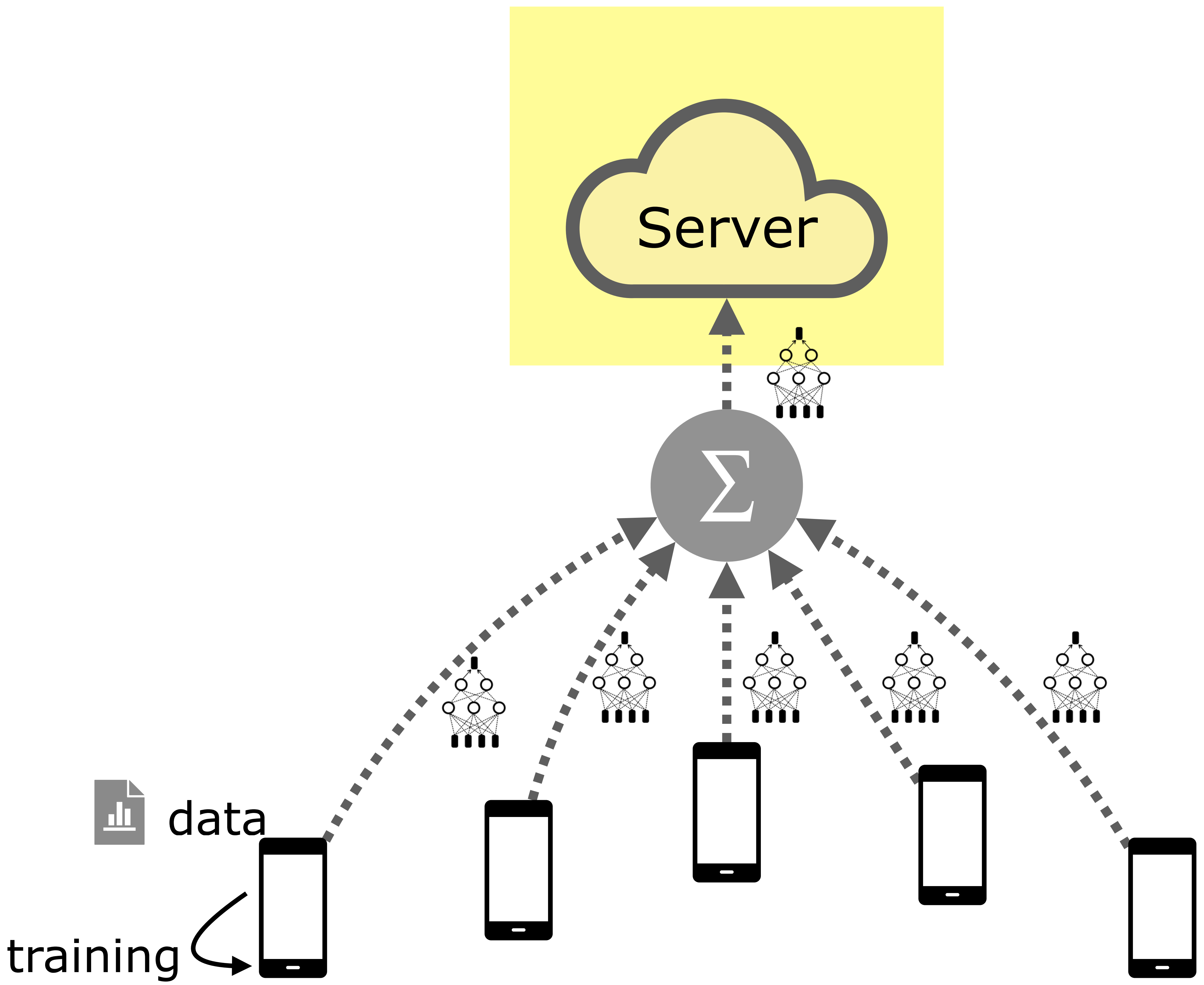
Federated Learning

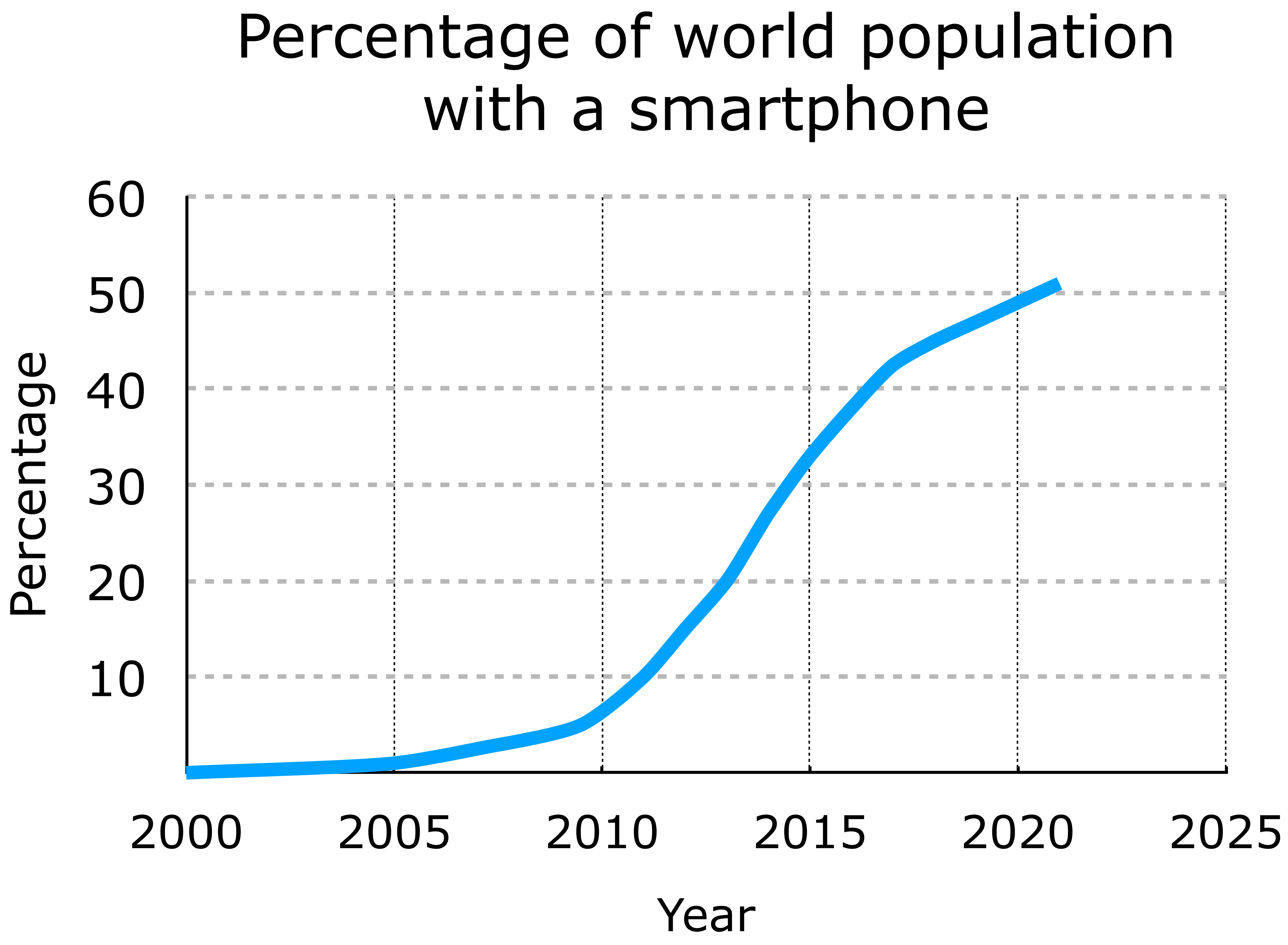




Data Credit: Business Wire

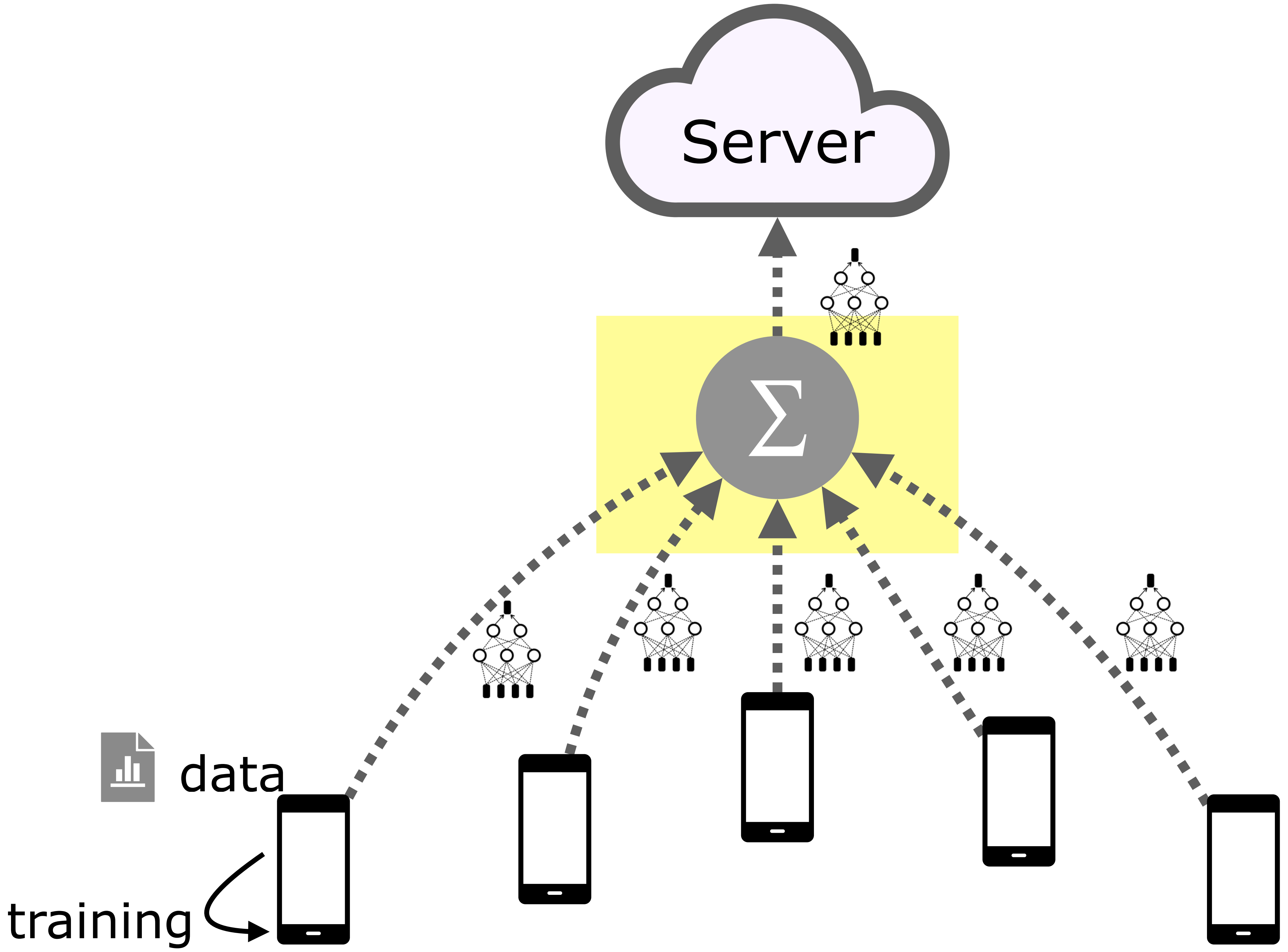
Federated Learning

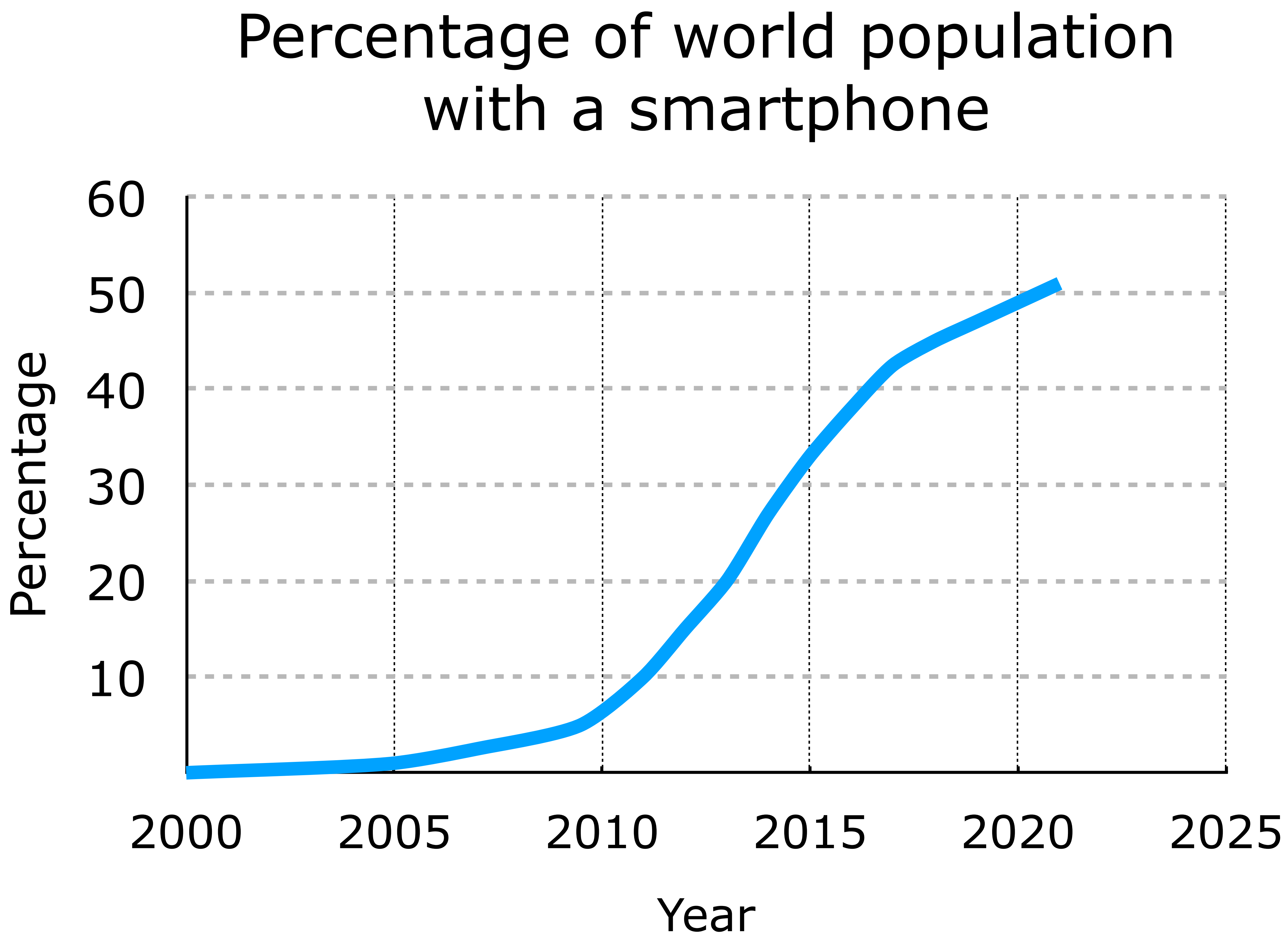




Data Credit: Business Wire

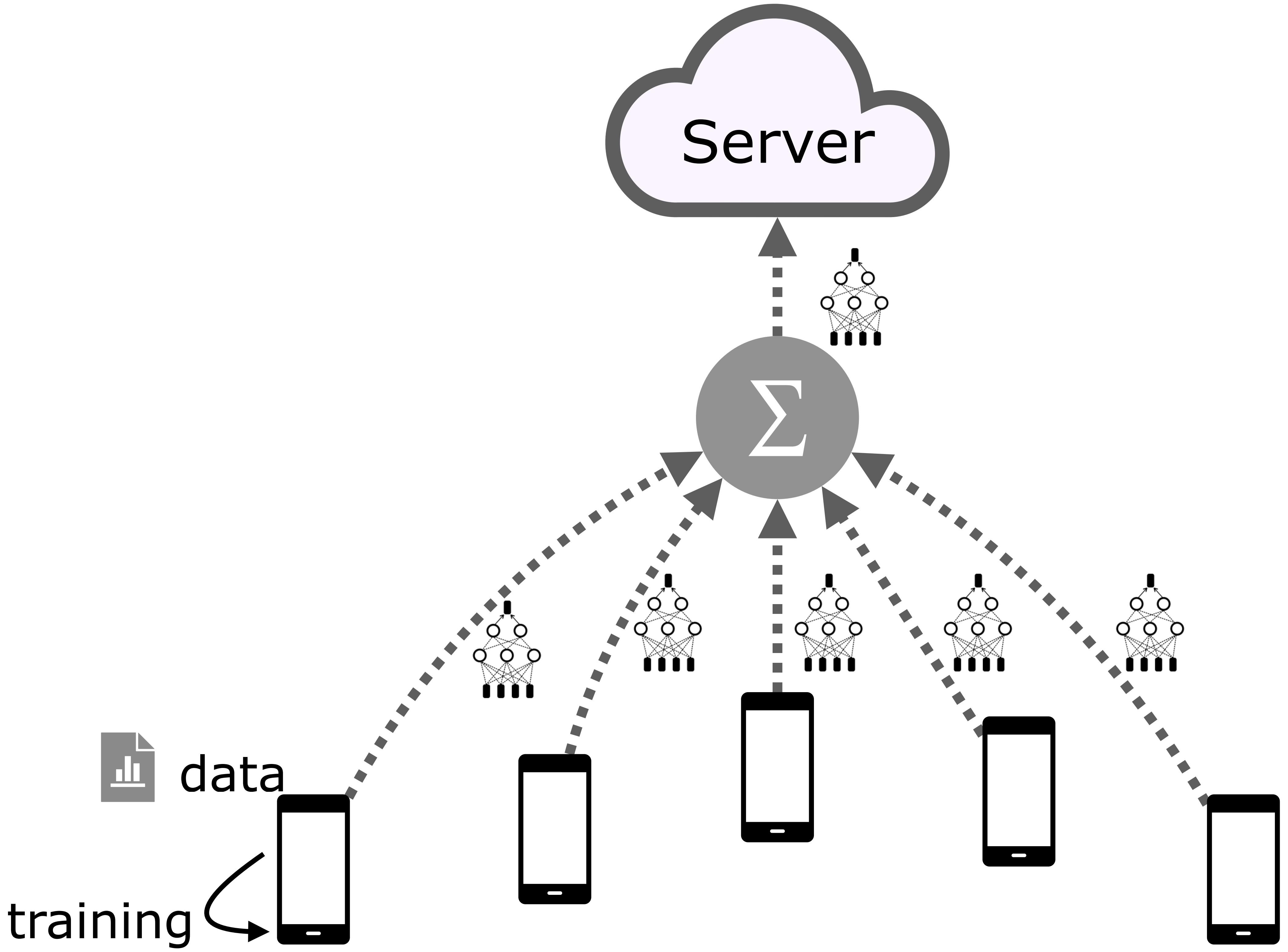
Federated Learning

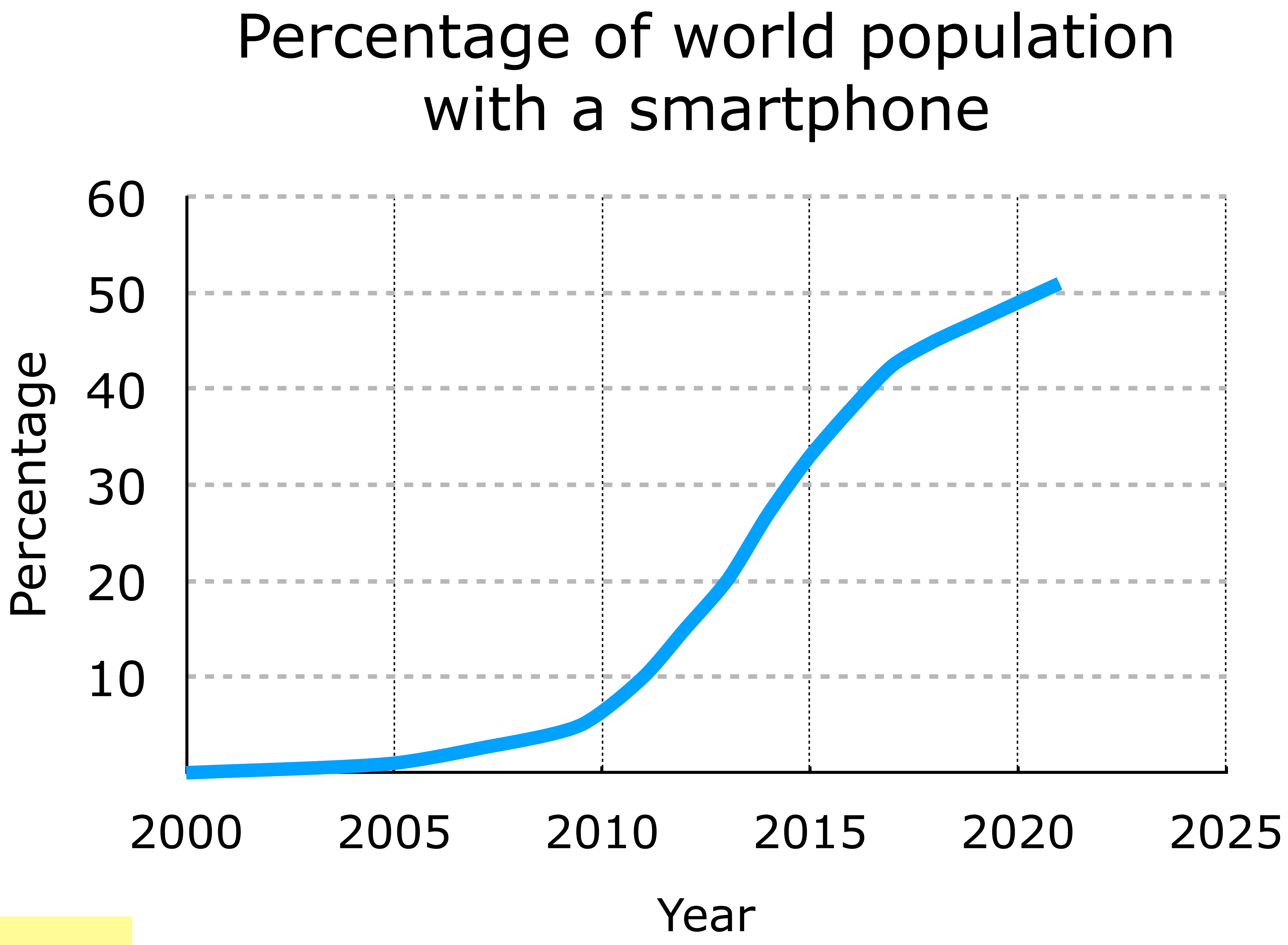




Data Credit: Business Wire

Federated Learning





Communication cost > computation cost!

Data Credit: Business Wire

Challenges

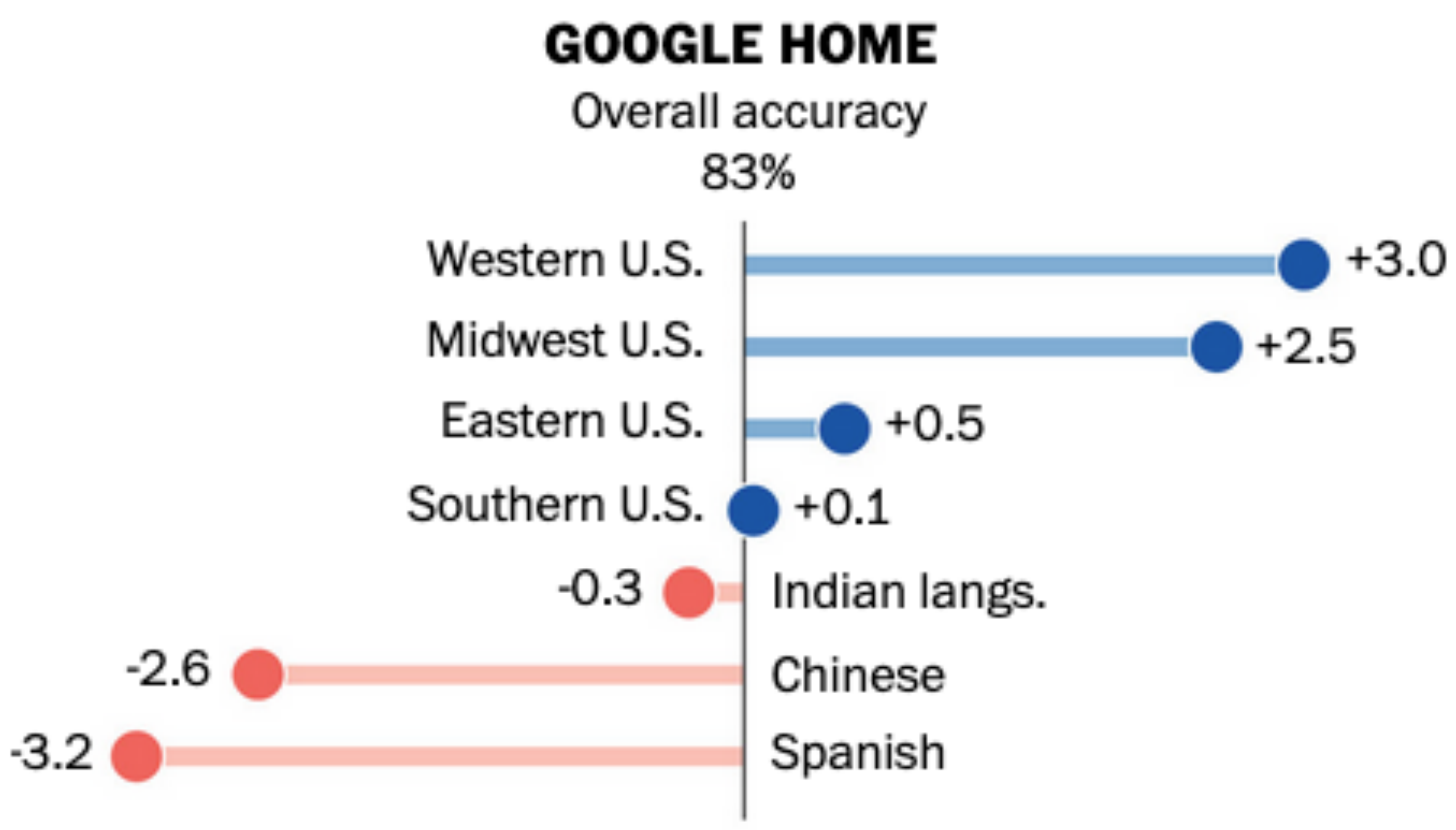
- models are deployed on clients with heterogeneous data

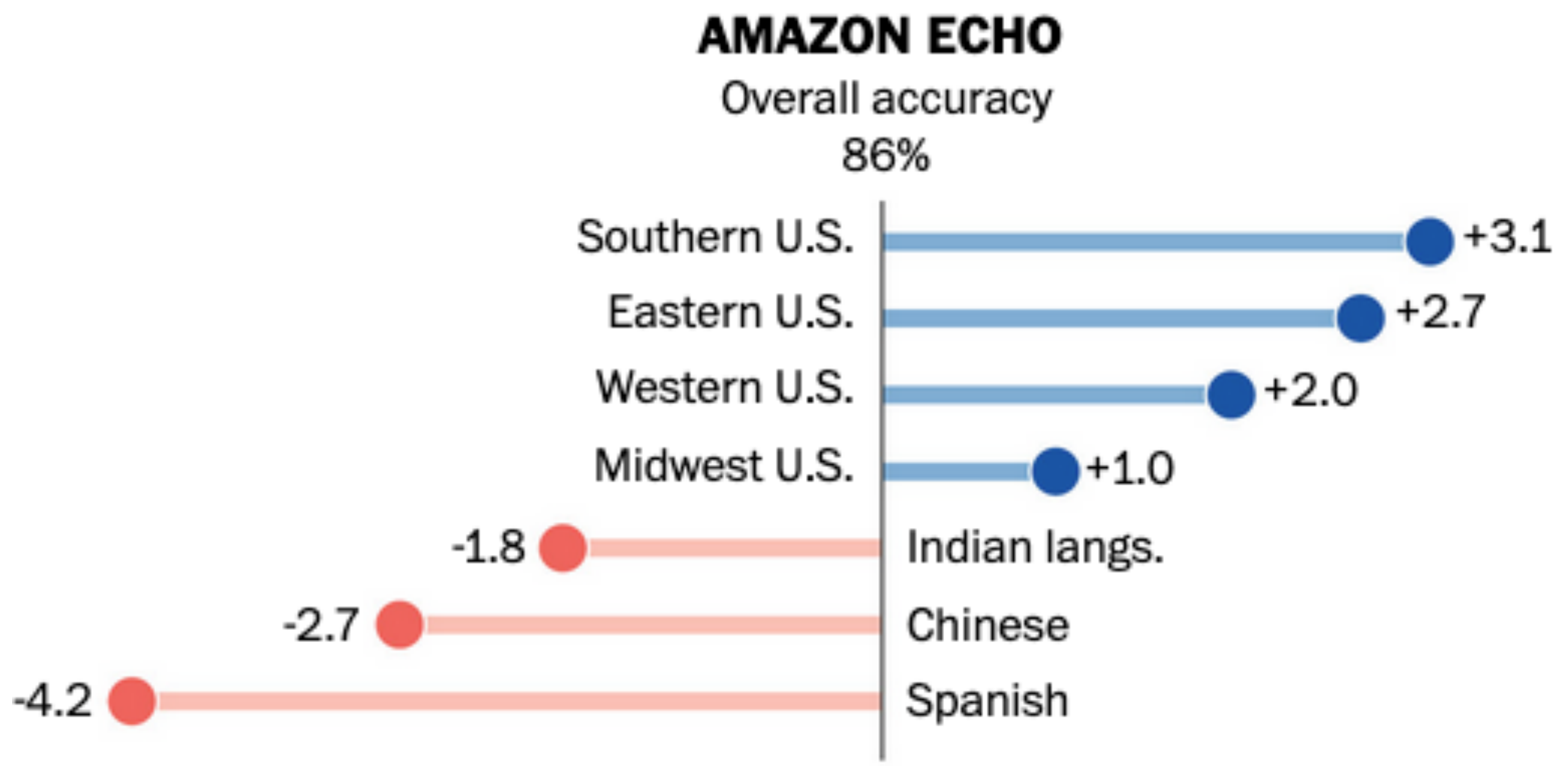
The Washington Post

Democracy Dies in Darkness

THE ACCENT GAP

We tested Amazon's Alexa and Google's Home to see how people with accents are getting left behind in the smart-speaker revolution.





By Drew Harwell July 19, 2018

Challenges

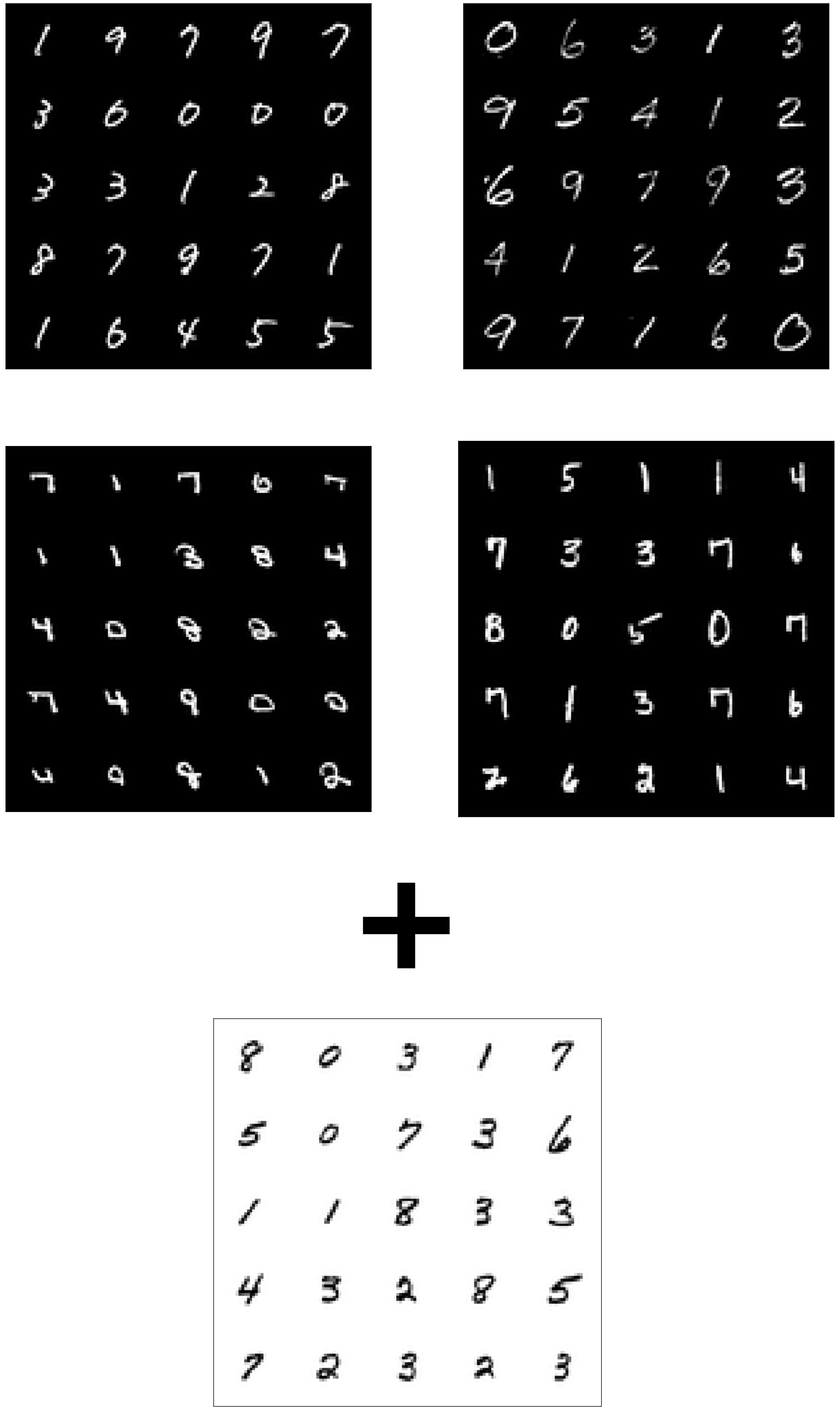
- models are deployed on clients with heterogeneous data
- training is not robust to potentially malicious clients

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



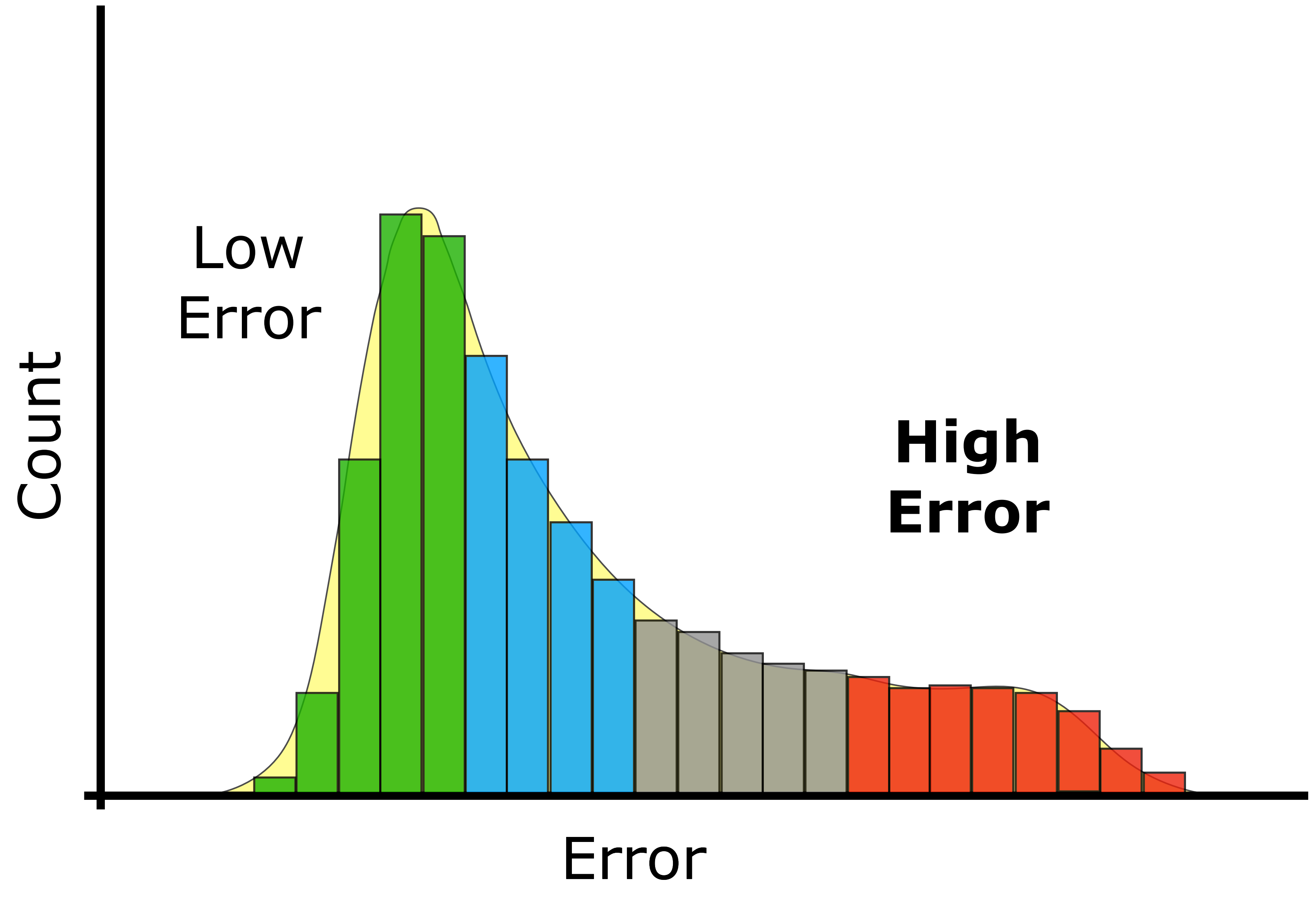
	Accuracy
Clean	64.3%
	-4.2pp
10% Corrupt	60.1%

Challenges

- models are deployed on clients with heterogeneous data
- training is not robust to potentially malicious clients
- solutions to both these problems are conflicting

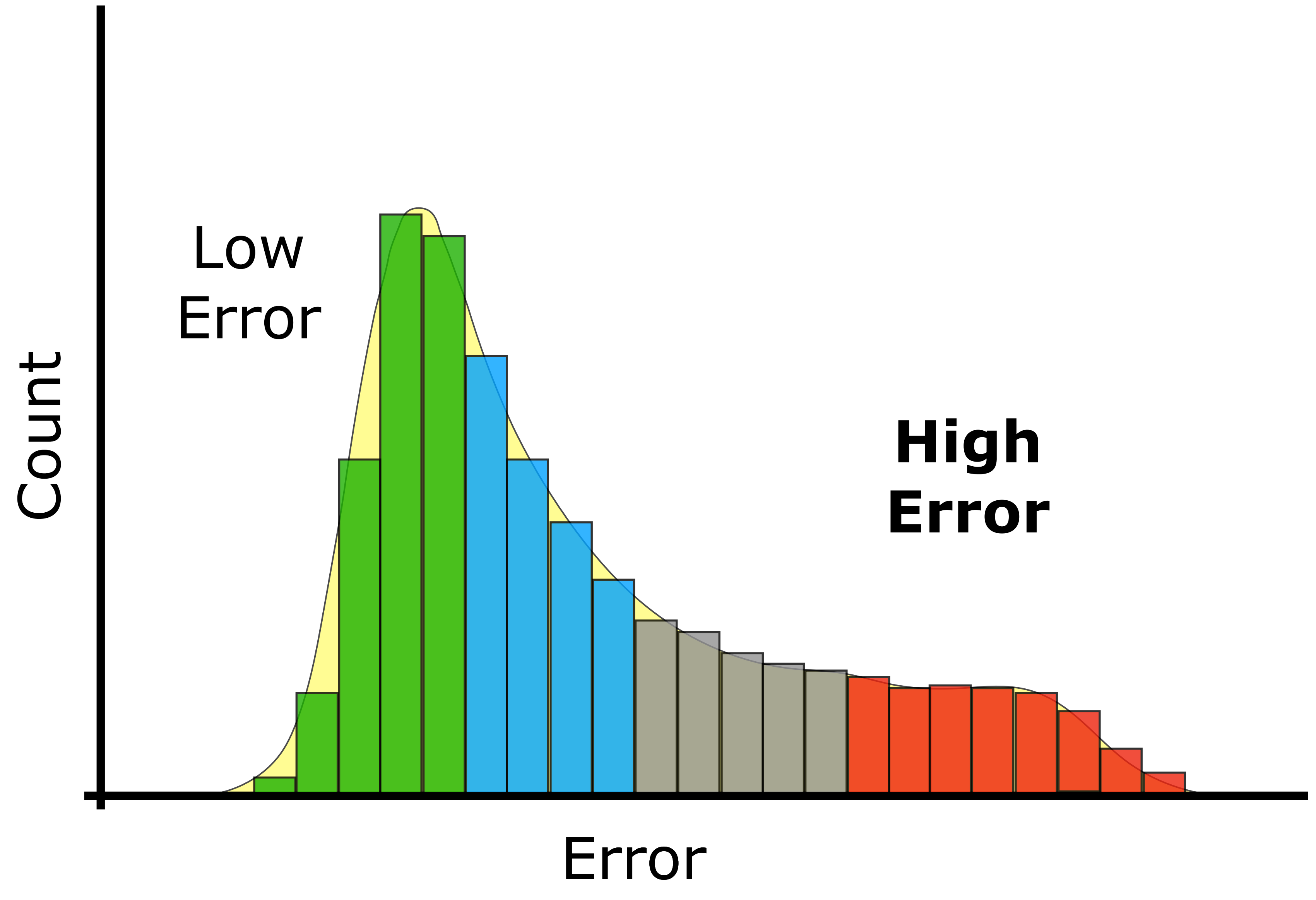
This talk

1. **Heterogeneity-aware objectives** for federated learning
   [CISS ’21, SVAA ’21, Under Review ’21]



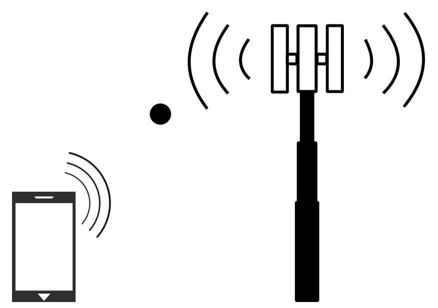
This talk

1. **Heterogeneity-aware objectives** for federated learning
   [CISS ’21, SVAA ’21, Under Review ’21]

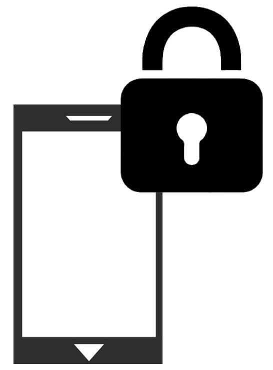


Our approach

Directly minimize
the tail error



Communication
efficiency

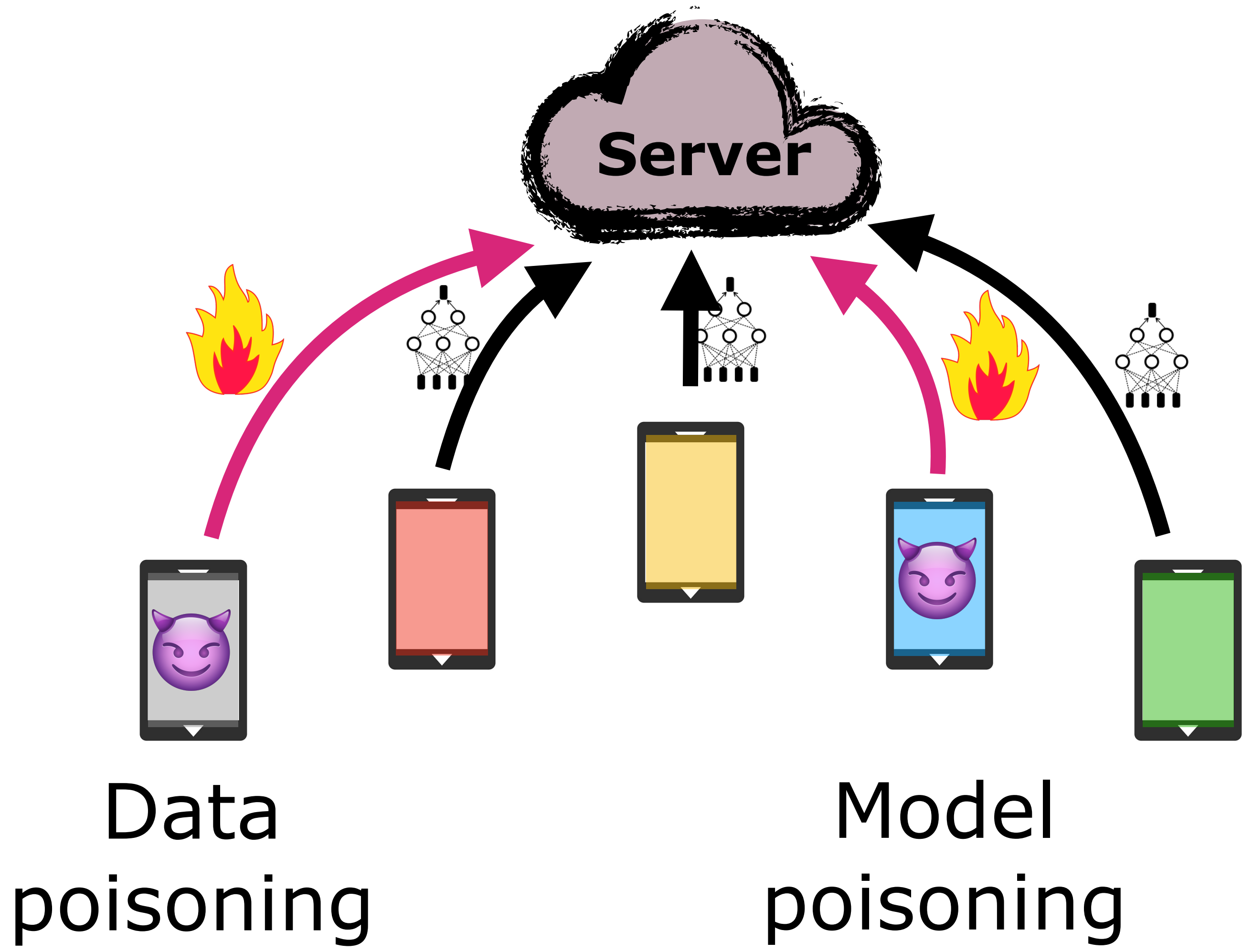


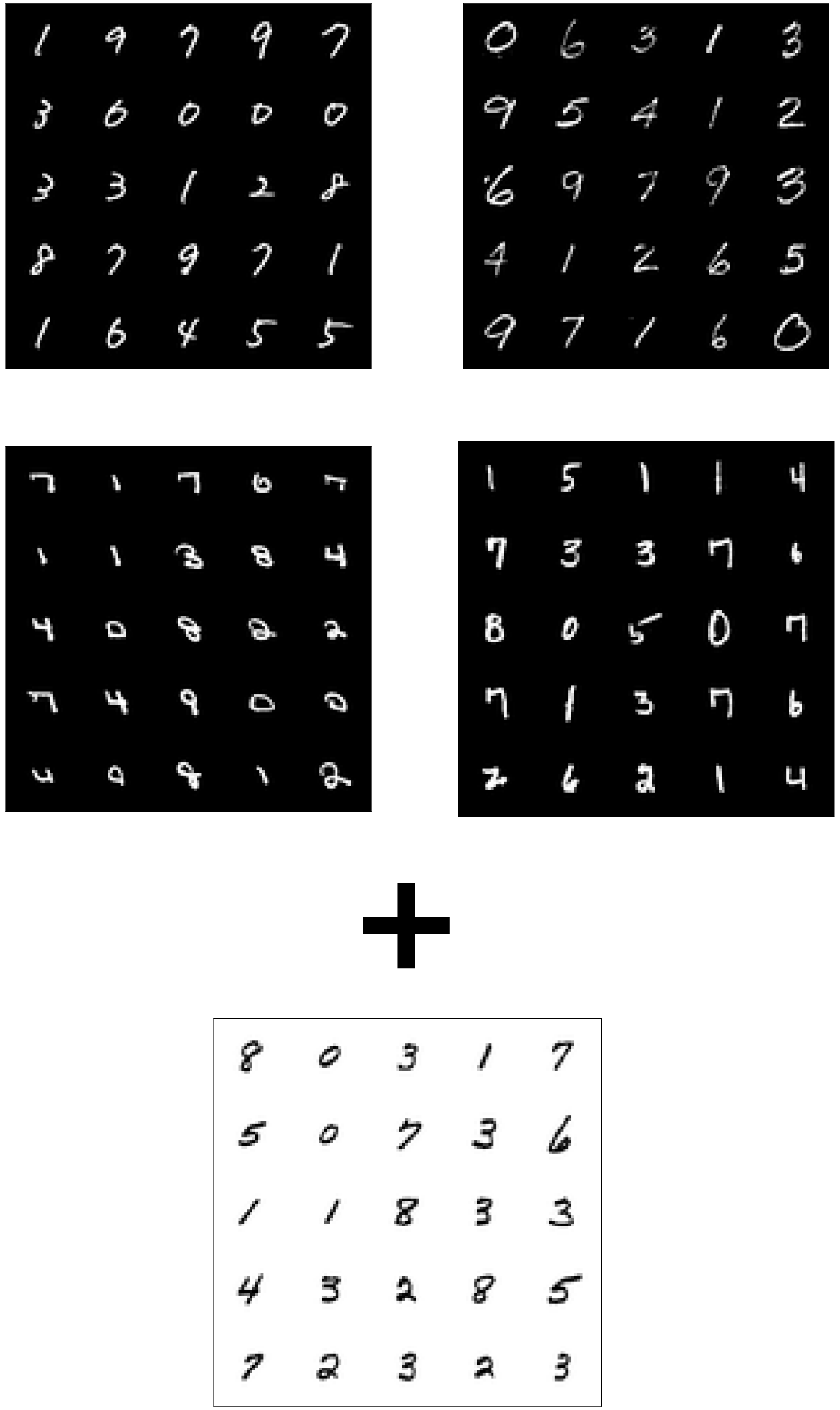
Privacy

This talk

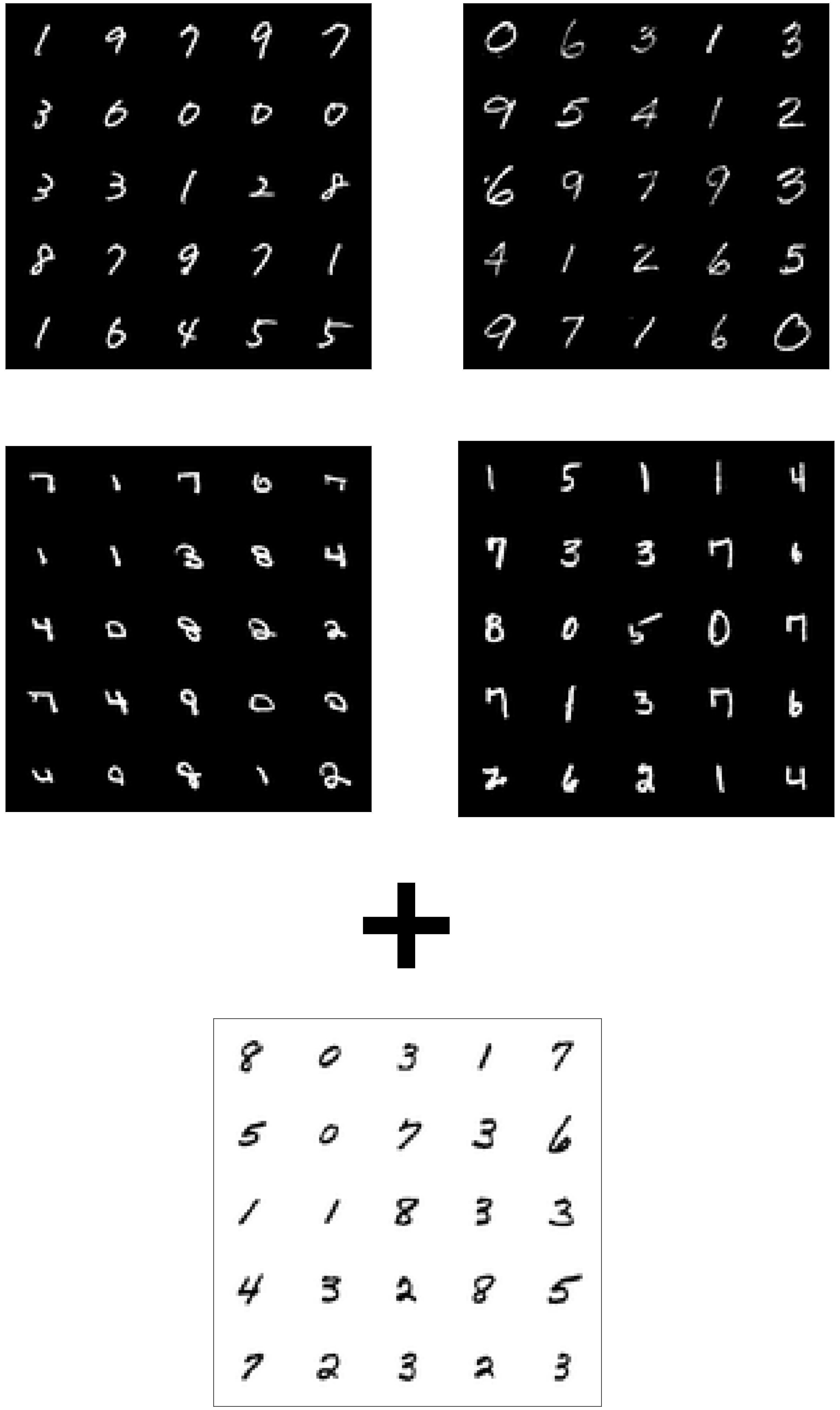
1. **Heterogeneity-aware objectives** for federated learning
   [CISS ’21, SVAA ’21, Under Review ’21]

2. **Robust aggregation** for federated learning
   [TSP ’22]





	Usual	Robust
Clean	64.3%	62.9%
	-4.2pp	-0.6pp
10% Corrupt	60.1%	62.3%



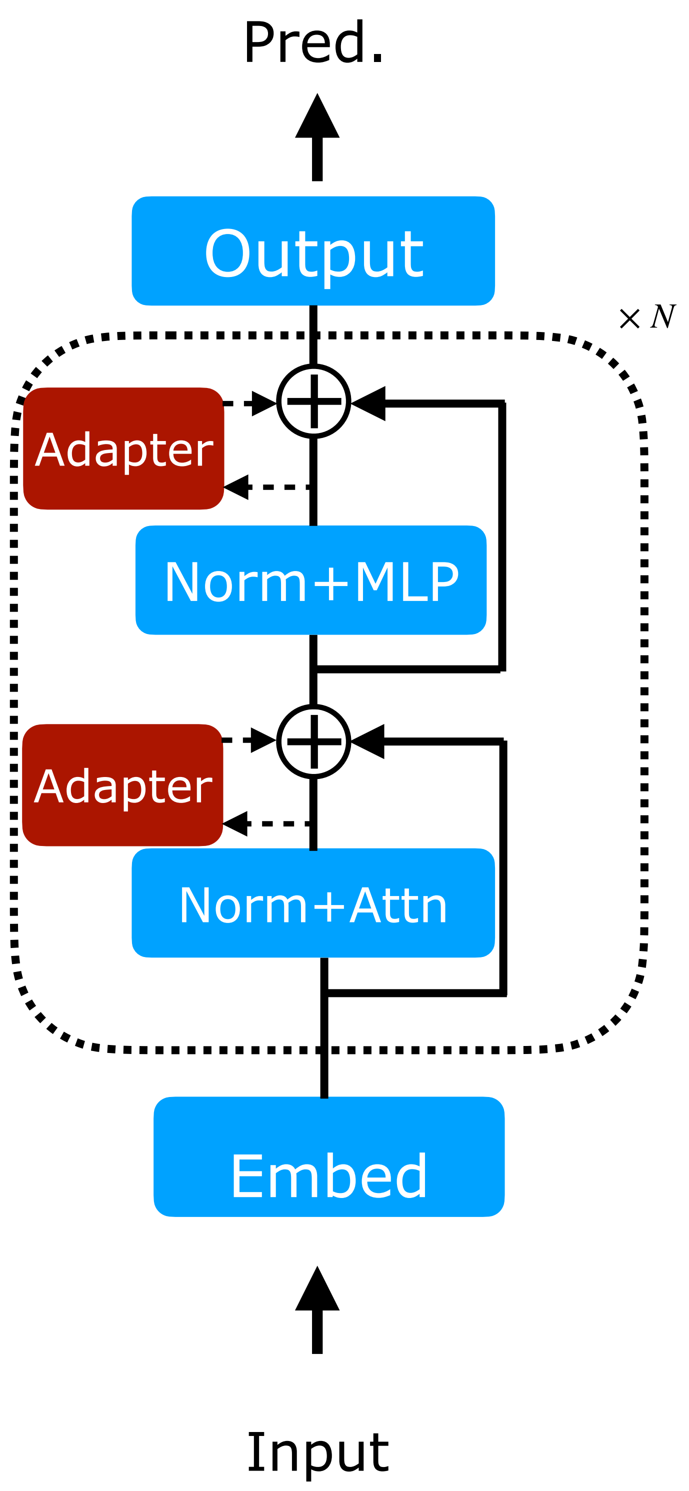
	Usual		Robust
Clean	64.3%	-1.4pp	62.9%
	-4.2pp		-0.6pp
10% Corrupt	60.1%		62.3%

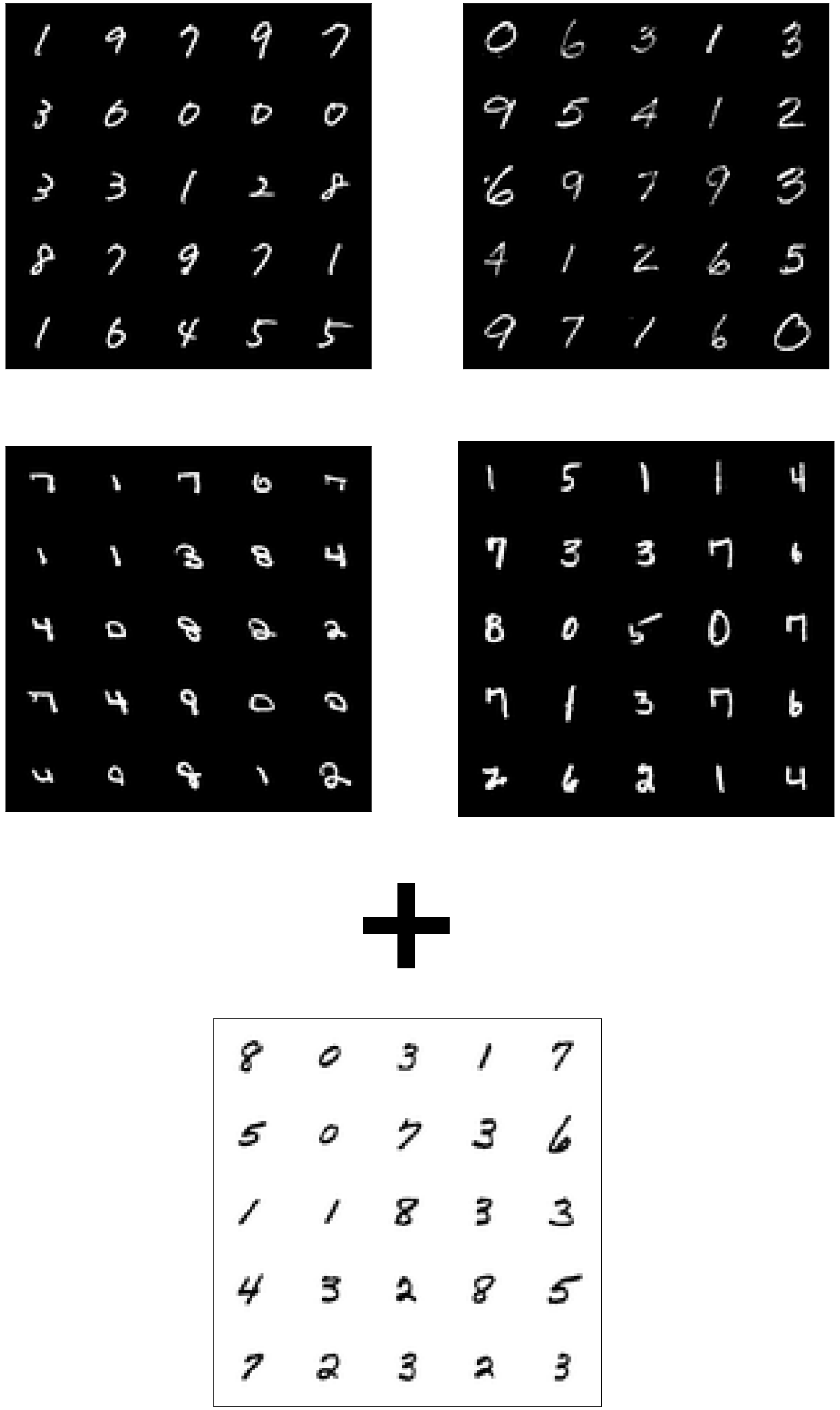
This talk

1. **Heterogeneity-aware objectives** for federated learning
   [CISS ’21, SVAA ’21, Under Review ’21]

2. **Robust aggregation** for federated learning
   [TSP ’22]

3. **Model personalization** for federated learning
   [ICML ’22, TSP ’22]





	Usual	Robust
Clean	70.1%	69.8%
25% Corrupt	24.6%	66.0%

-0.3pp

-45.5pp

-3.8pp

This talk

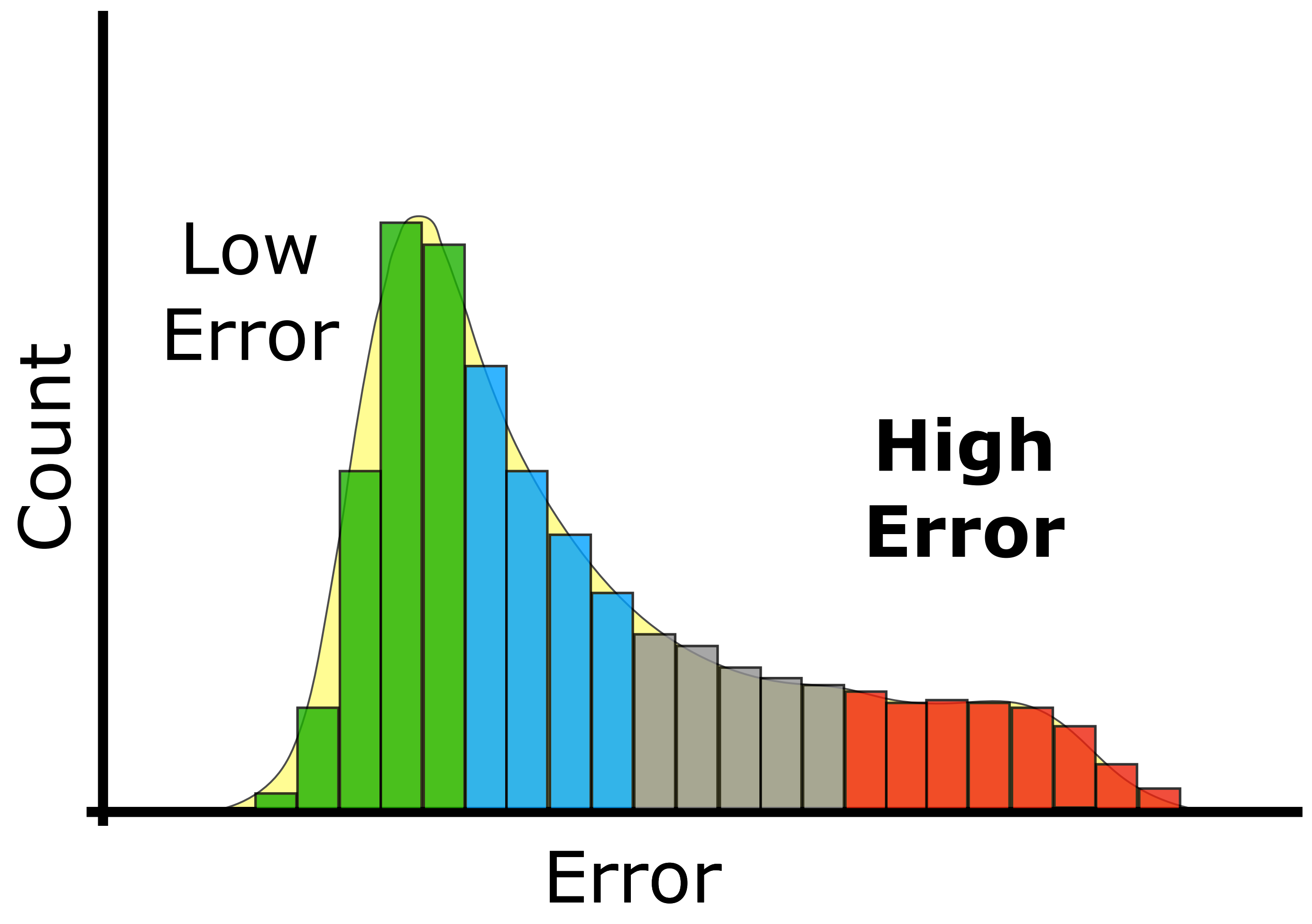
1. **Heterogeneity-aware objectives** for federated learning
   [CISS ’21, SVAA ’21, Under Review ’21]

2. **Robust aggregation** for federated learning
   [TSP ’22]

3. **Model personalization** for federated learning
   [ICML ’22, TSP ’22]

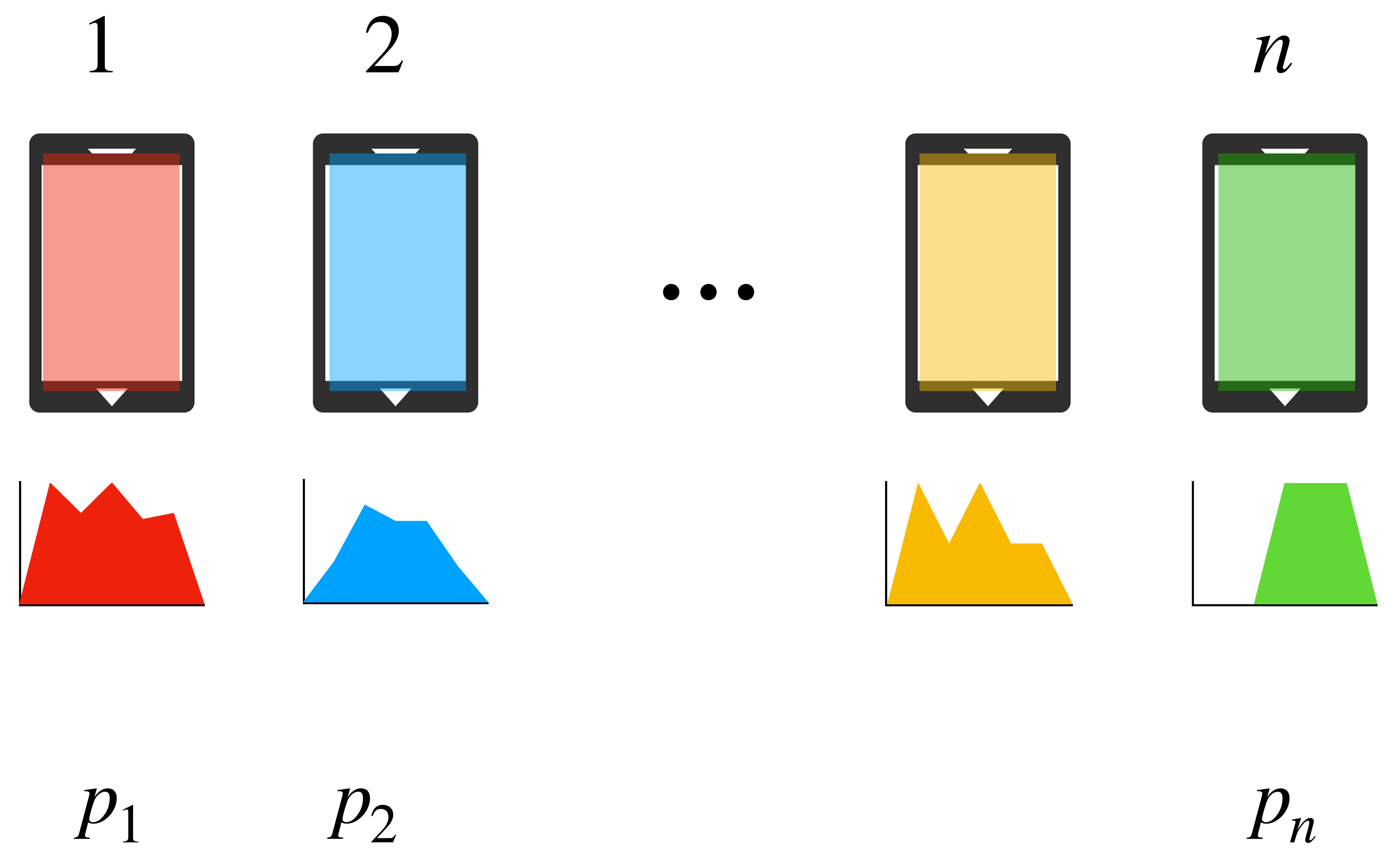
Part 1: Heterogeneity-aware objectives for federated learning

[CISS ’21, SVAA ’21, Under Review ’21]



Clients

Data Distribution



Data heterogeneity

Usual Learning Objective

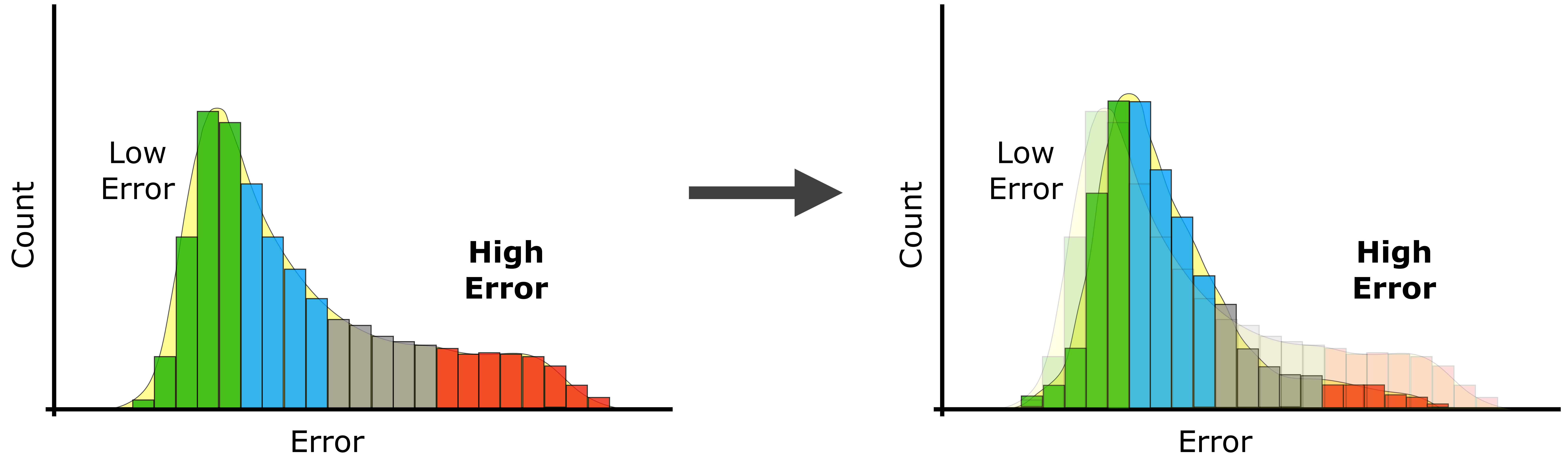
$$\min_{w \in \mathbb{R}^d} \quad \frac{1}{n}\sum_{i=1}^{n} F_i(w) \qquad \text{where} \qquad F_i(w) = \mathbb{E}_{z \sim p_i}\left[f(w; z)\right]$$

loss on client i

[McMahan et al. AISTATS (2017), Kairouz et al. (2021)]

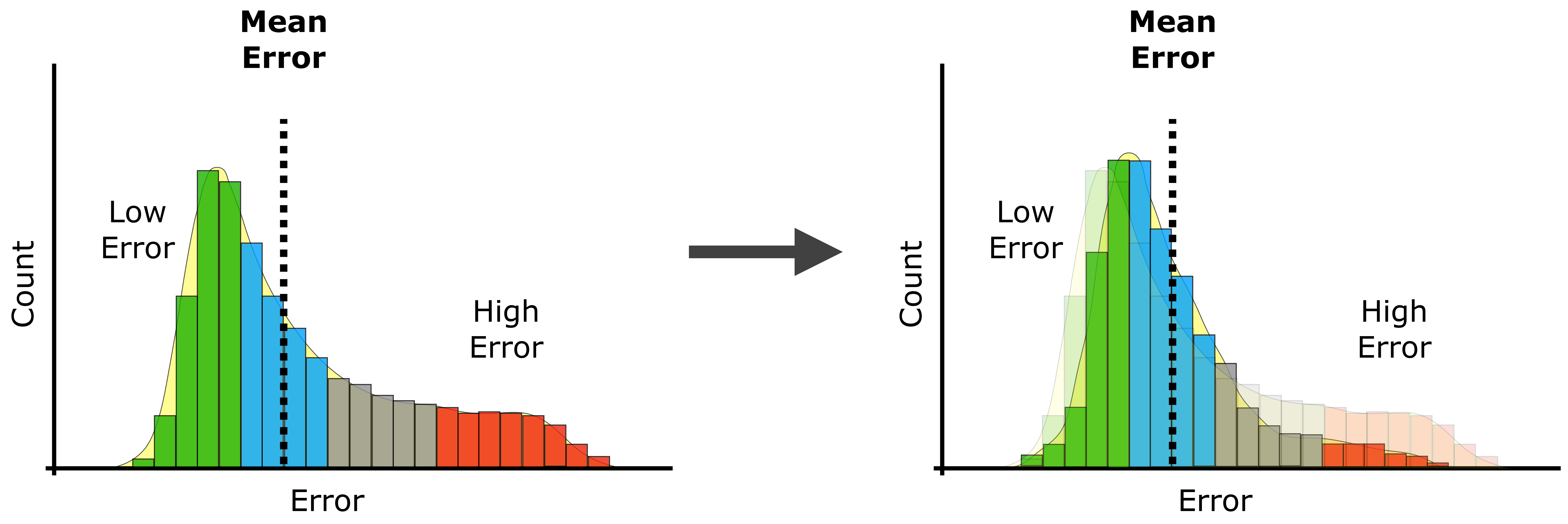
Our goal

Reduce *tail* error



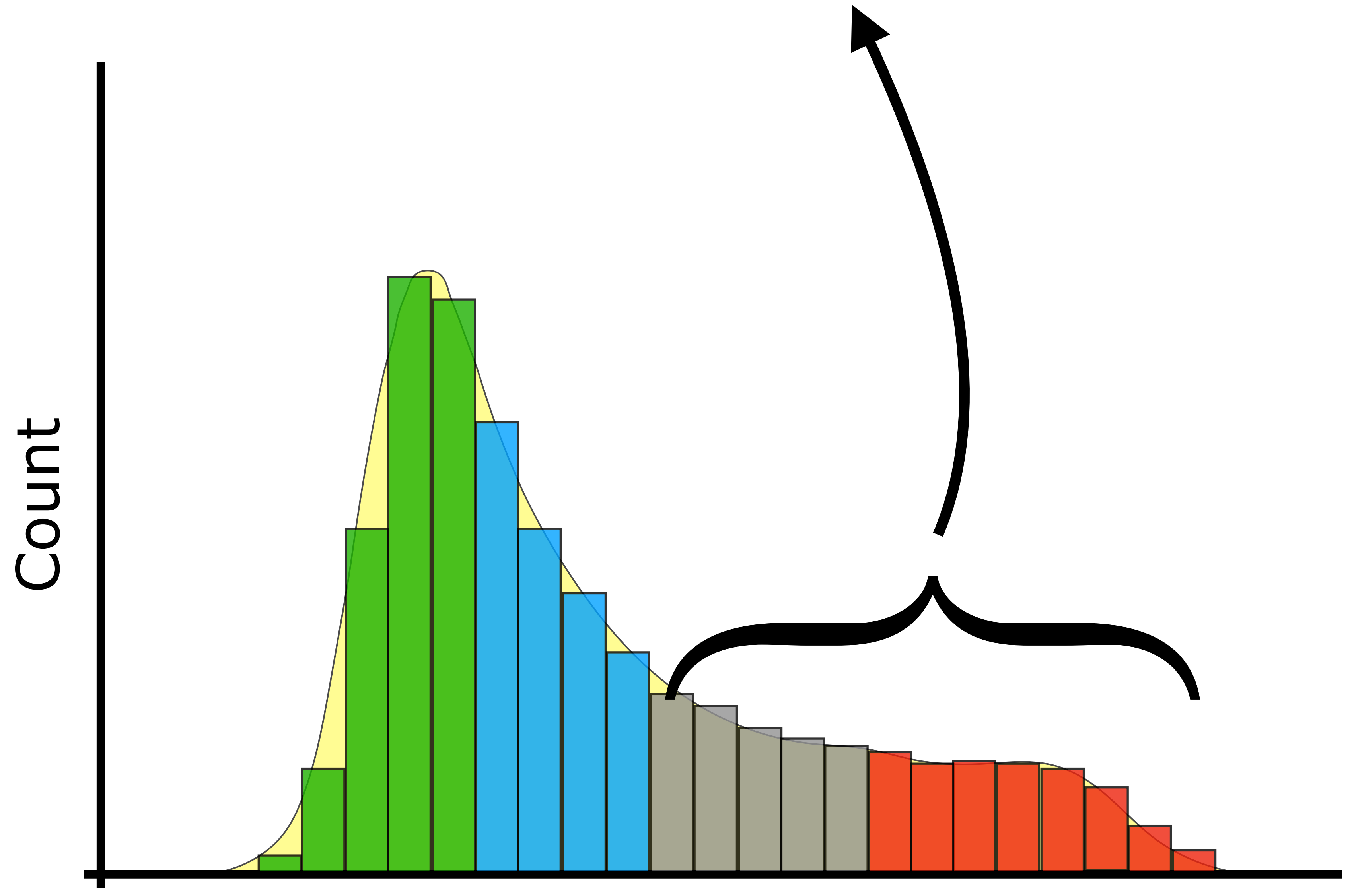
Our goal

Reduce *tail* error without sacrificing the *mean* error



Simplicial federated learning

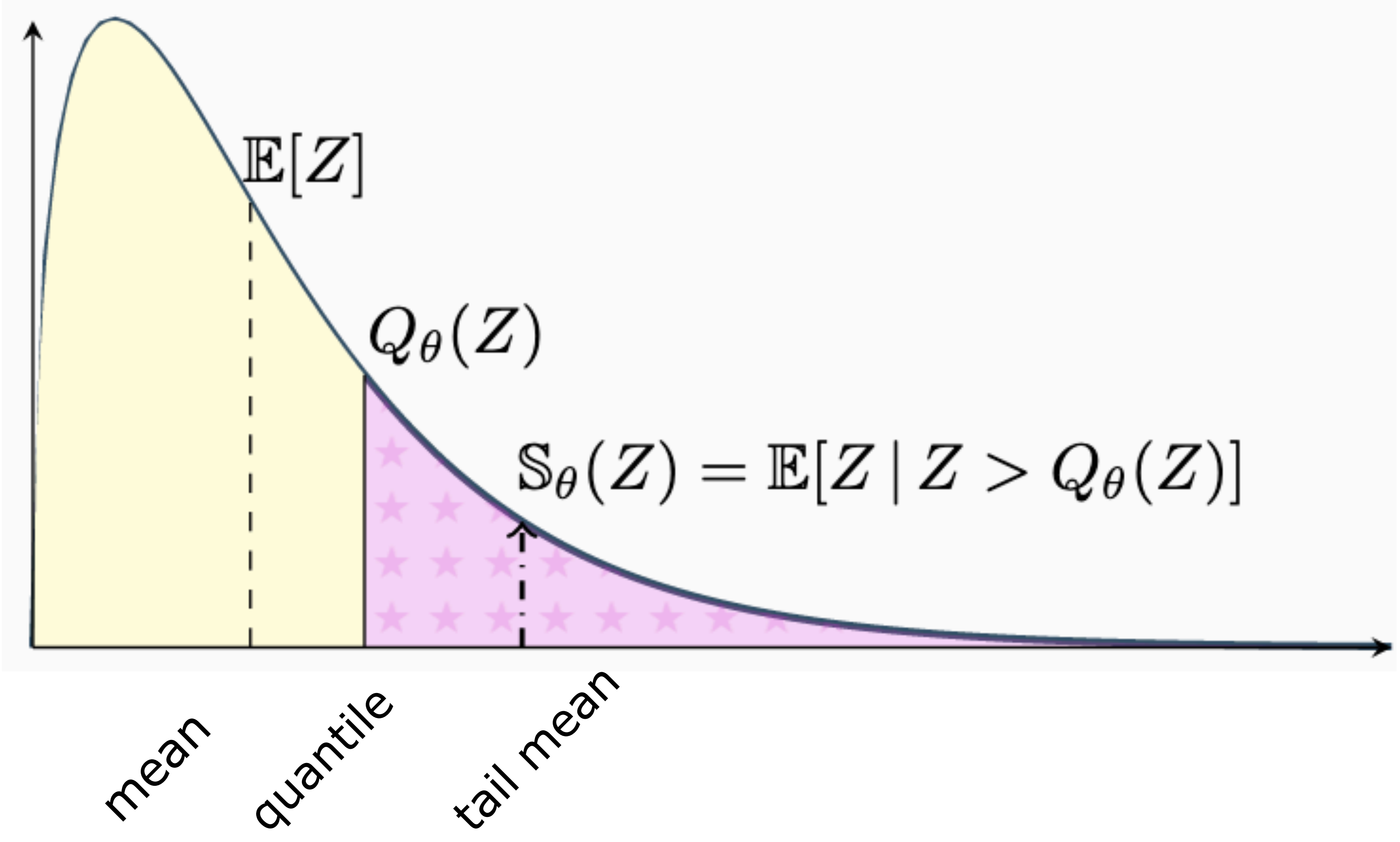
Our Approach: minimize the tail error directly!



Simplicial-FL Objective:

$$\min_{w} \mathbb{S}_\theta\Big(\big(F_1(w), \cdots, F_n(w)\big) \Big)$$

Superquantile | Conditional Value at Risk

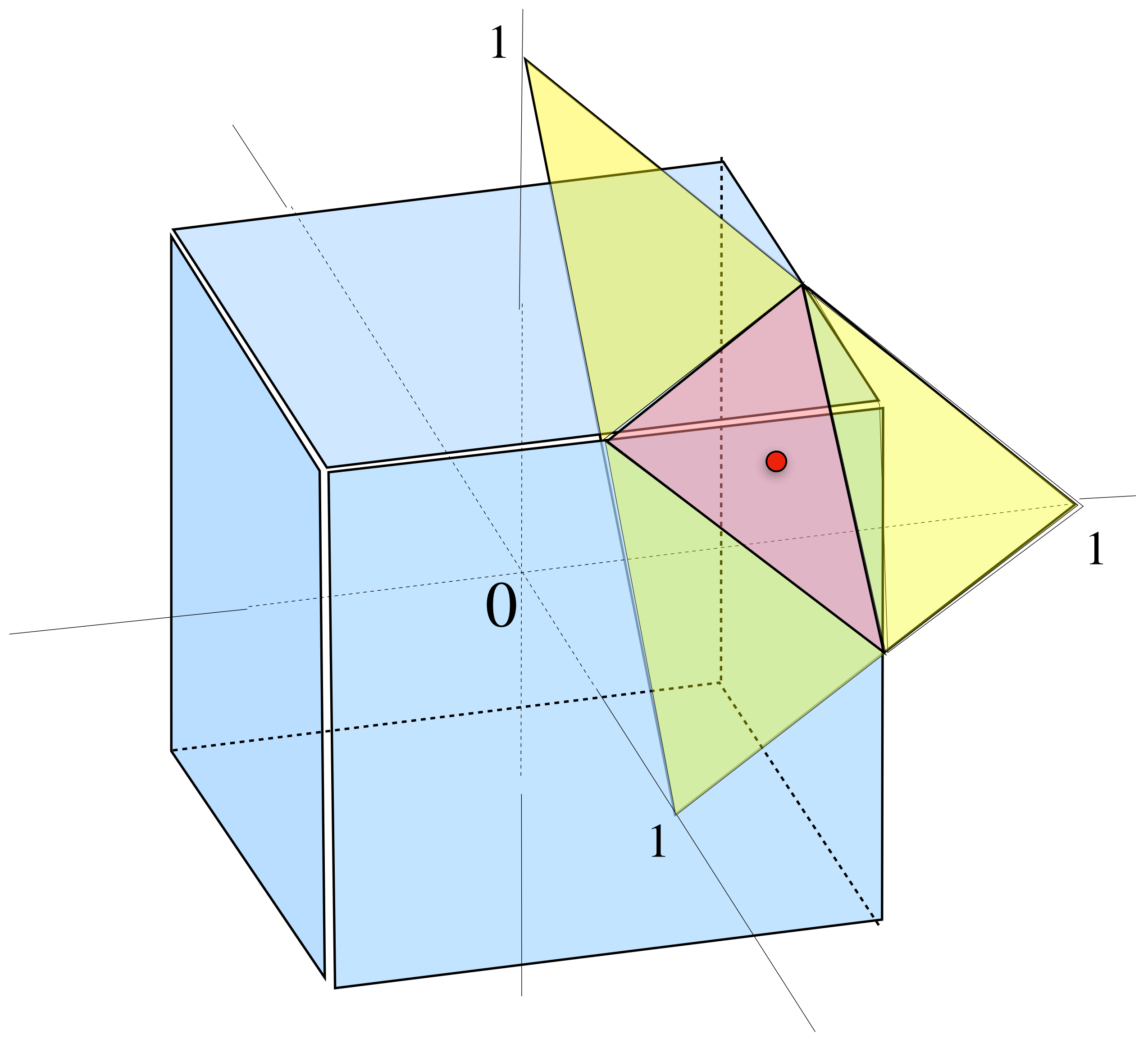


[Rockafellar & Uryasev (2002)]

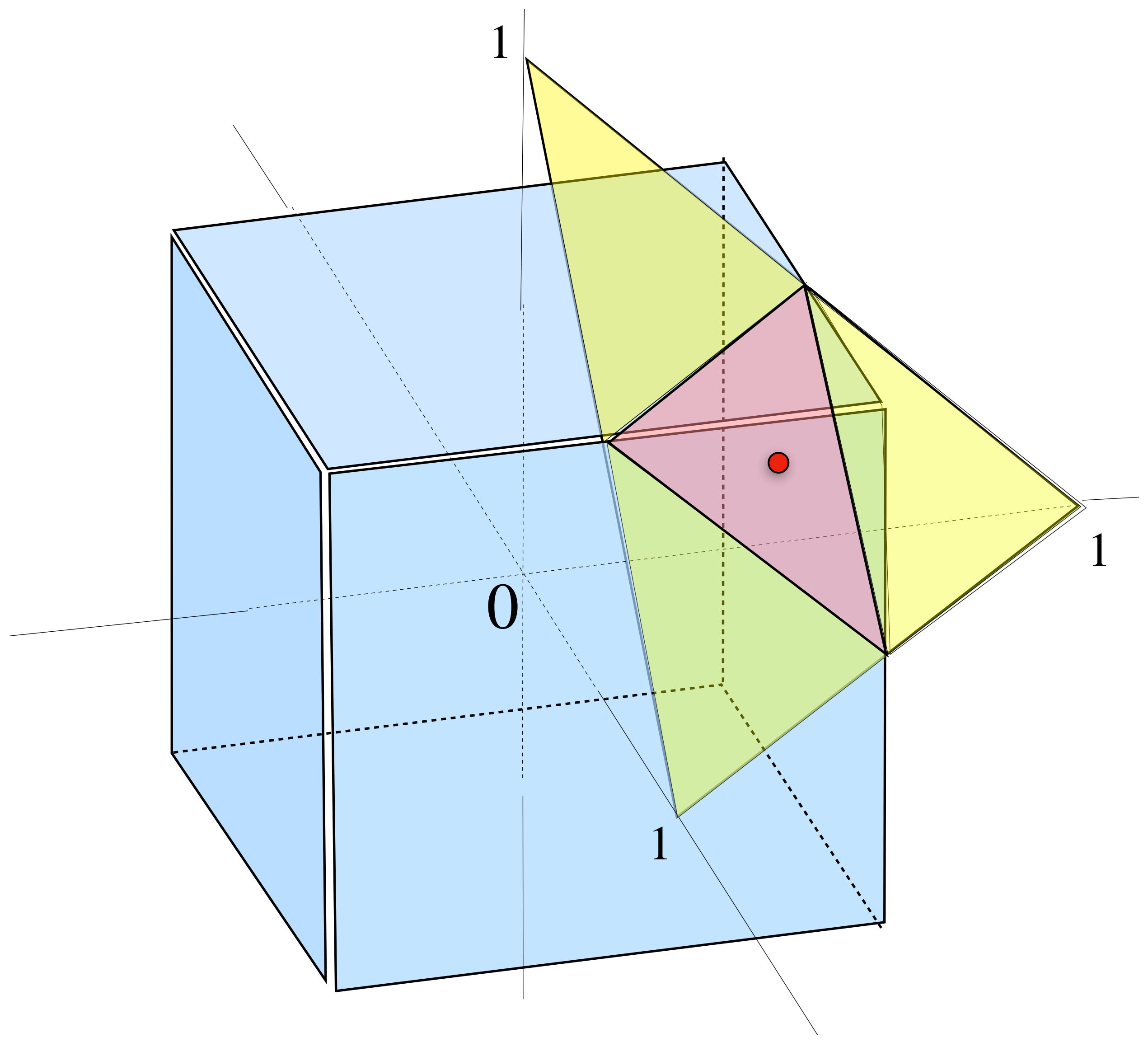
Dual expression ≡ continuous knapsack problem

$$\mathbb{S}_\theta(x_1, \cdots, x_n) = \max \left\{ \sum_i \pi_i x_i \ : \ \pi_i \geq 0, \ \sum_i \pi_i = 1, \ \pi_i \leq (n\theta)^{-1} \right\}$$

[Dantzig (1957), Ben-Tal & Teboulle (1987), Föllmer & Schied (2002)]



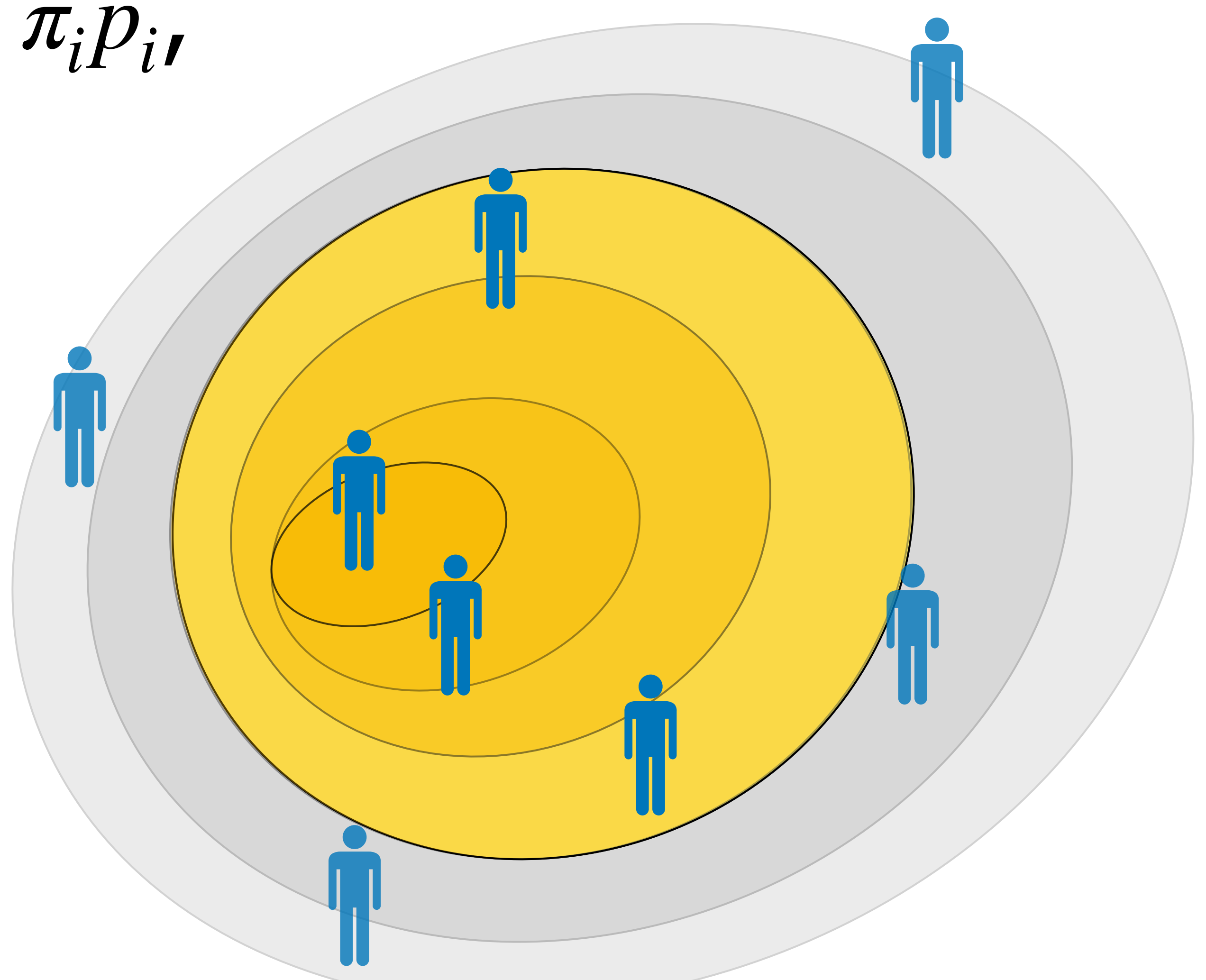
Dual expression ≡ continuous knapsack problem



$$\mathbb{S}_\theta(x_1, \cdots, x_n) = \max \left\{ \sum_i \pi_i x_i \,:\, \pi_i \geq 0,\ \sum_i \pi_i = 1,\ \pi_i \leq (n\theta)^{-1} \right\}$$

[Dantzig (1957), Ben-Tal & Teboulle (1987), Föllmer & Schied (2002)]

Assuming a new test client with mixture distribution $p_\pi = \sum_i \pi_i p_i$,



Simplicial-FL objective is equivalent to:

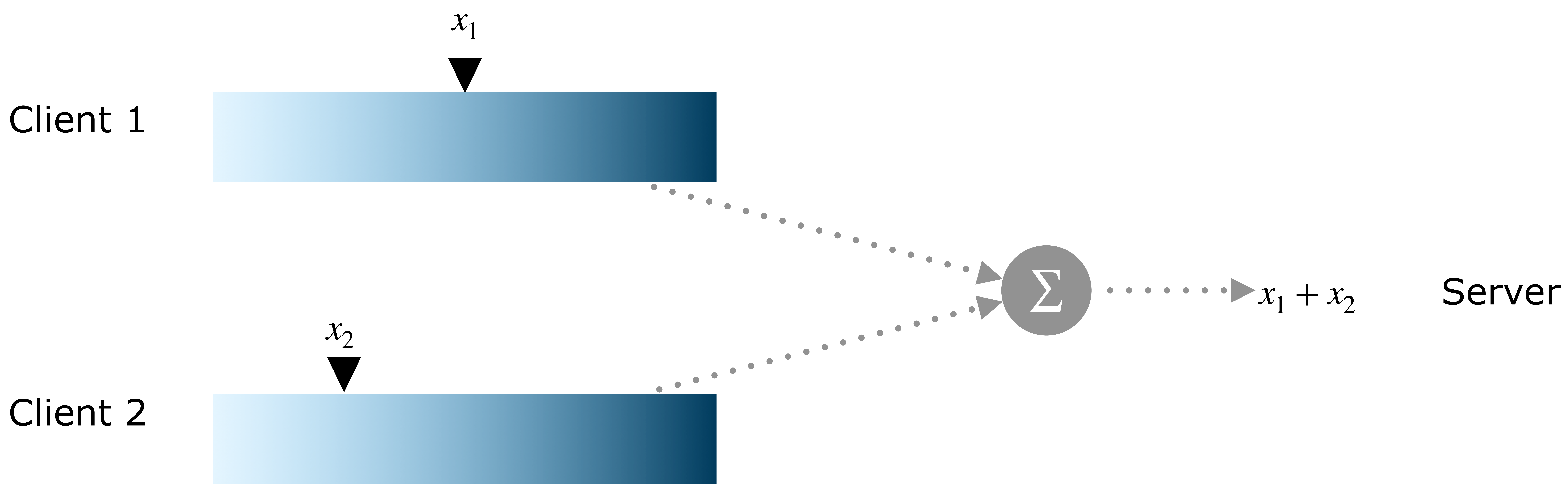
$$\min_w \max_{\pi : \pi_i \leq (n\theta)^{-1}} \mathbb{E}_{z \sim p_\pi} \left[f(w; z) \right]$$

$\Longrightarrow$ Distributionally robust learning

Optimization

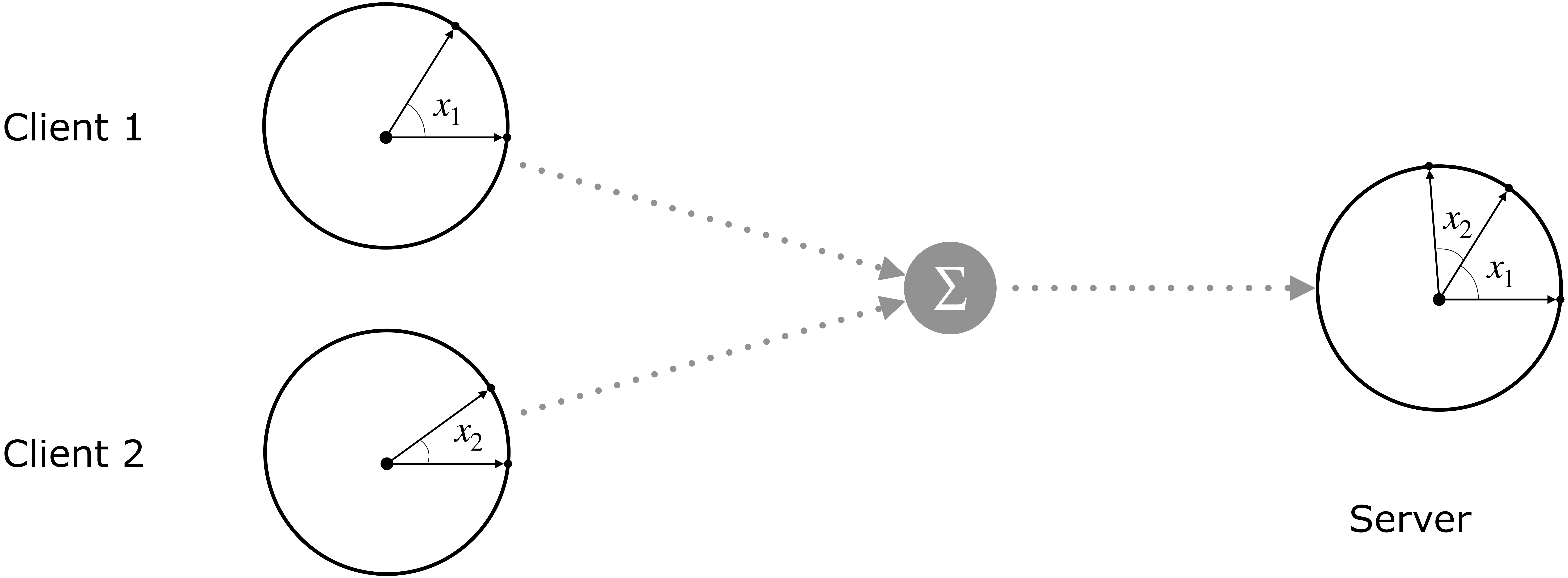
Communication primitive: secure sum

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)]

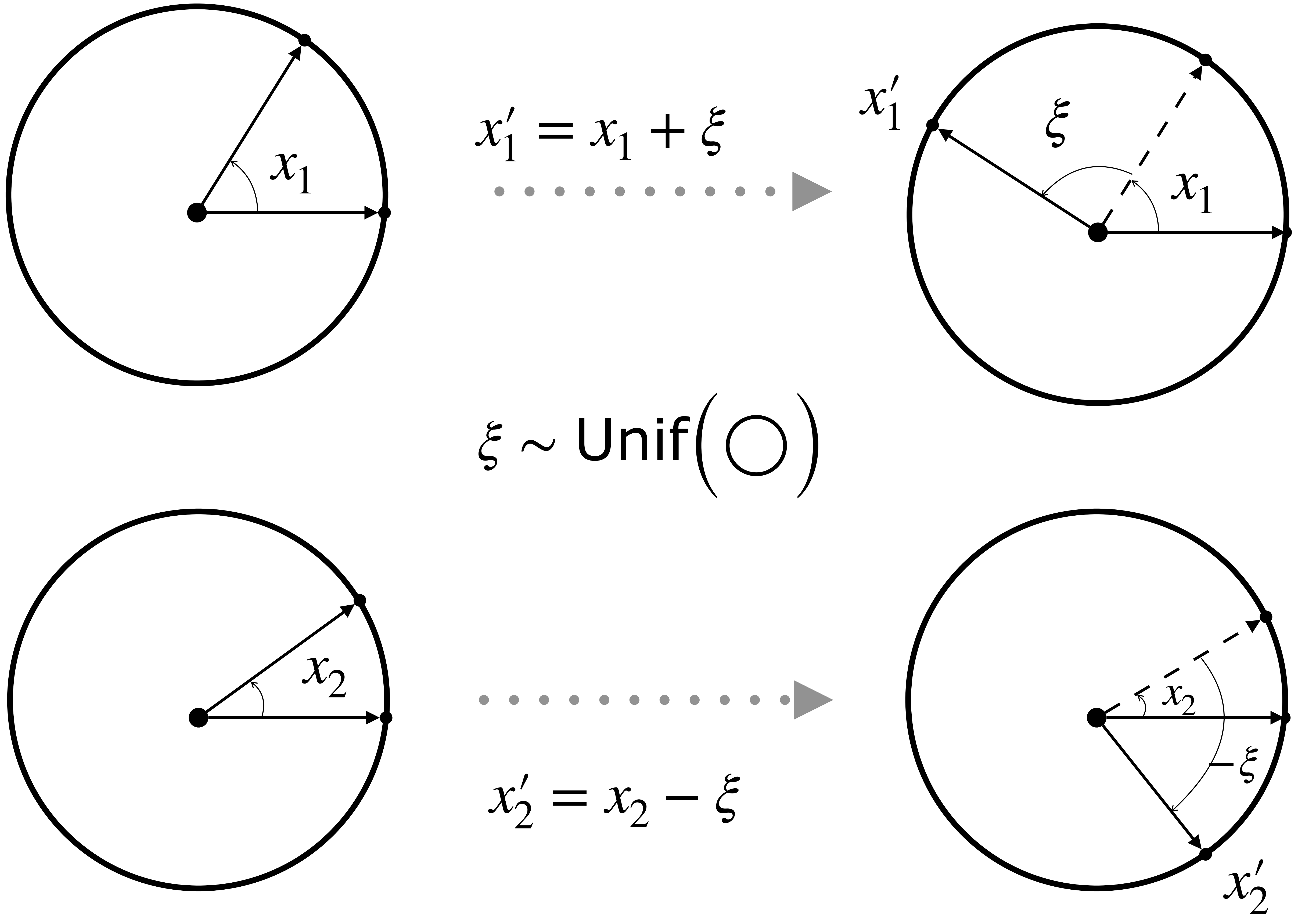
Perform all operations modulo M



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

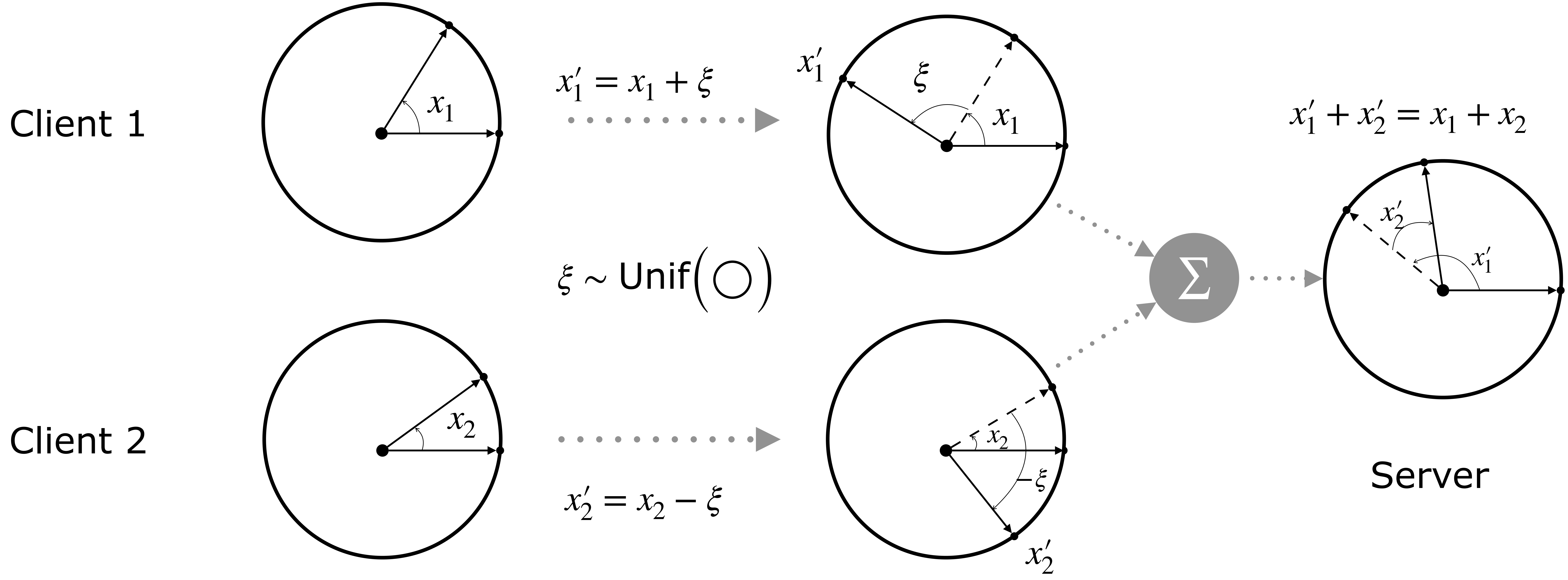
Client 1

Client 2



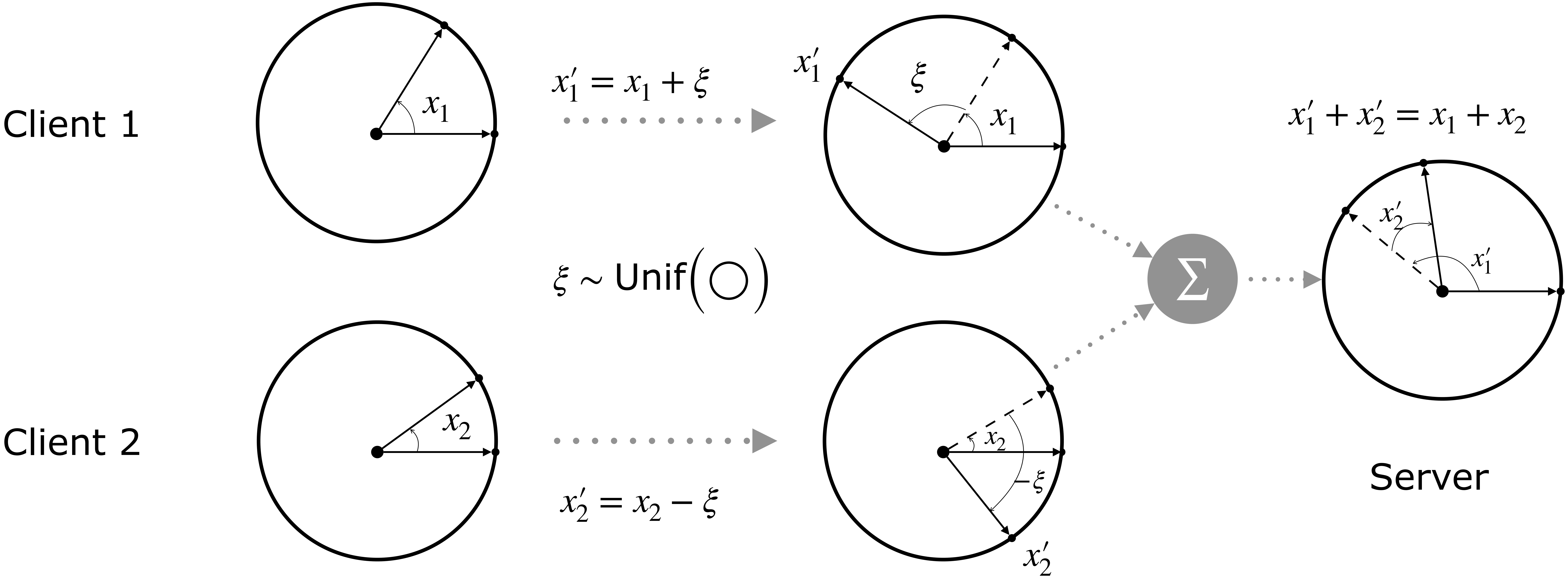
[Bonawitz et al. CCS (2017), Bell et al. CCS (2020)]

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



[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



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

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

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

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

ERM Algorithm (FedAvg):

$$\min_{w} \quad \frac{1}{n}\sum_{i=1}^{n} F_i(w)$$

Simplicial-FL Algorithm:

$$\min_{w} \mathbb{S}_\theta\Big(\big(F_1(w), \cdots, F_n(w)\big) \Big)$$

FedAvg [MacMahan et al. AISTATS (2017)]

Parallel Gradient Distribution [Mangasarian. SICON (1995)]
Iterative Parameter Mixing [McDonald et al. ACL (2009)]
BMUF [Chen & Huo. ICASSP (2016)]
Local SGD [Stich. ICLR (2019)]

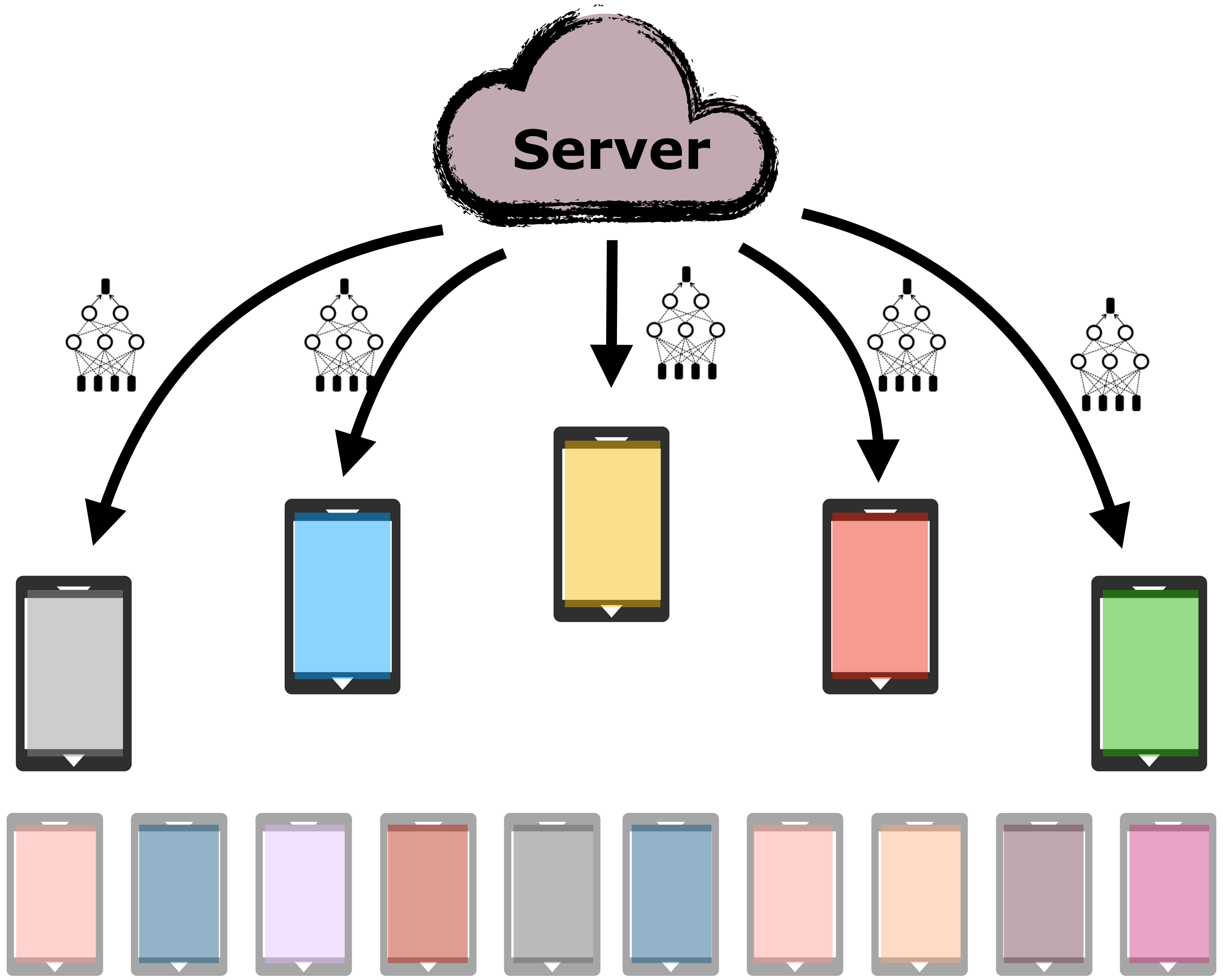
ERM Algorithm (FedAvg):

$$\min_{w} \quad \frac{1}{n}\sum_{i=1}^{n} F_i(w)$$

Simplicial-FL Algorithm:

$$\min_{w} \mathbb{S}_\theta\Big(\big(F_1(w), \cdots, F_n(w)\big) \Big)$$

Step 1 of 3: Server samples m clients and broadcasts global model



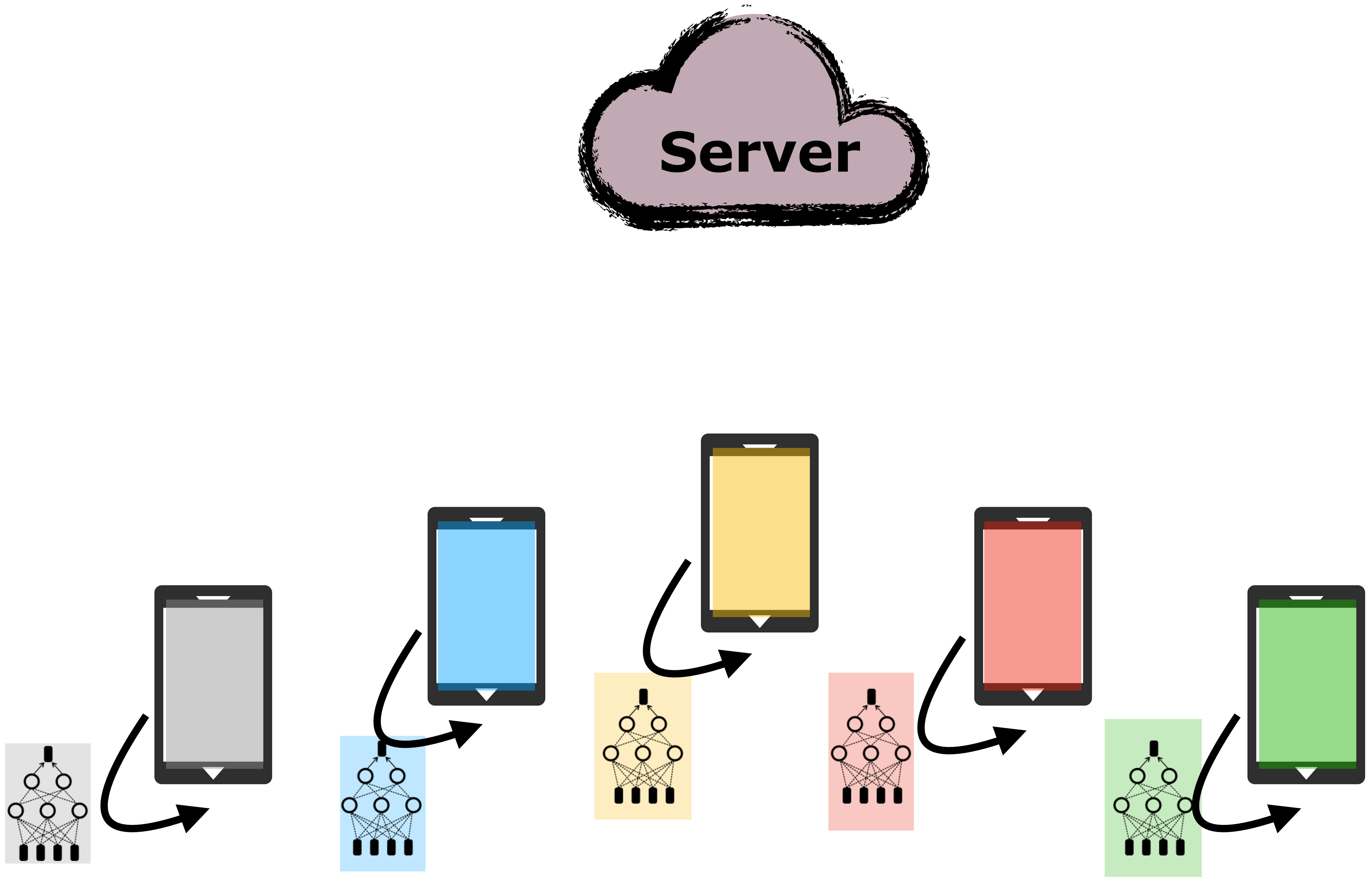
ERM Algorithm (FedAvg):

$$\min_{w} \quad \frac{1}{n}\sum_{i=1}^{n} F_i(w)$$

Simplicial-FL Algorithm:

$$\min_{w} \mathbb{S}_\theta\Big(\big(F_1(w), \cdots, F_n(w)\big) \Big)$$

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



ERM Algorithm (FedAvg):

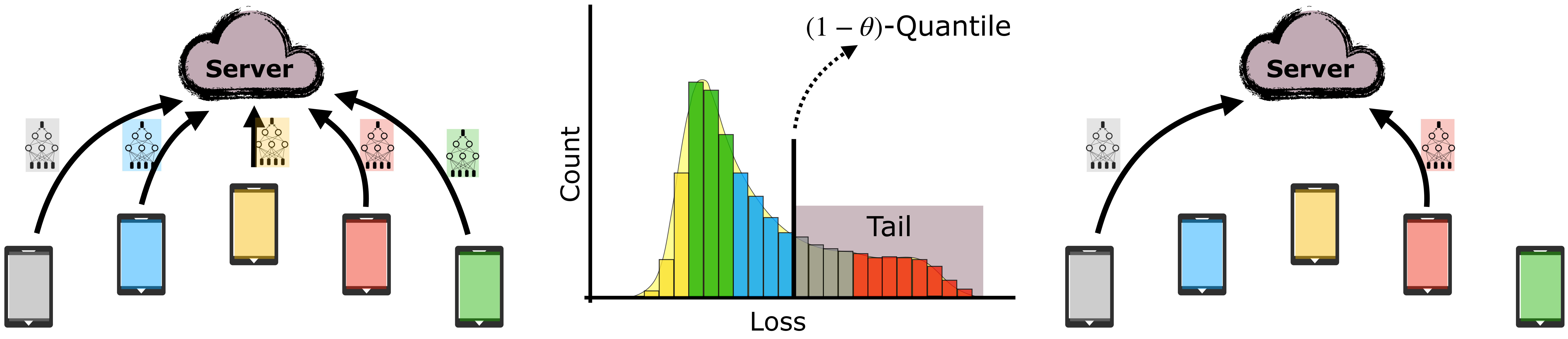
$$\min_{w} \quad \frac{1}{n}\sum_{i=1}^{n} F_i(w)$$

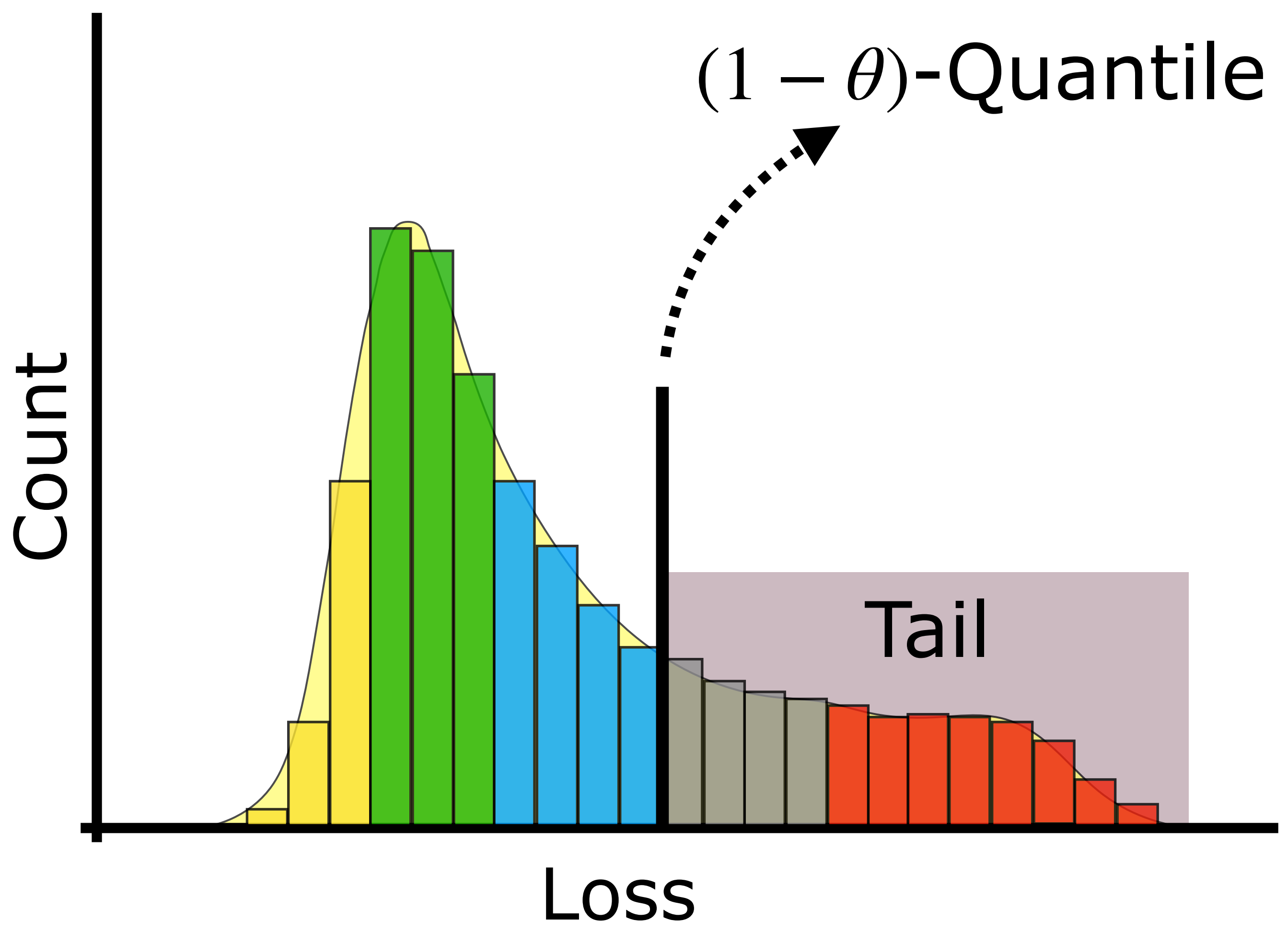
Step 3 of 3: Aggregate updates contributed by ***all clients***

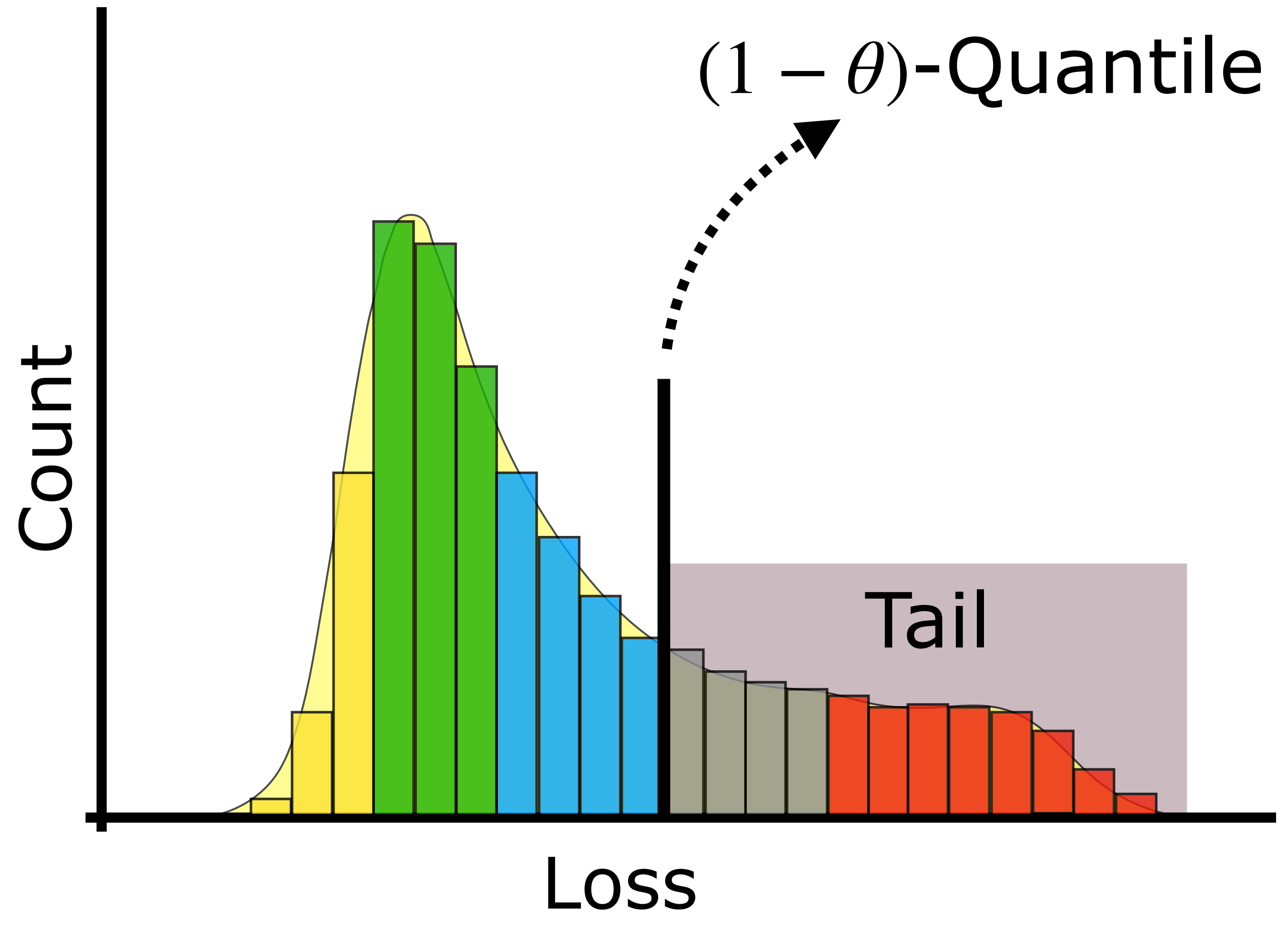
Simplicial-FL Algorithm:

$$\min_{w} \mathbb{S}_\theta\Big(\big(F_1(w), \cdots, F_n(w)\big) \Big)$$

Step 3 of 3: Aggregate updates contributed by ***tail clients*** *only*



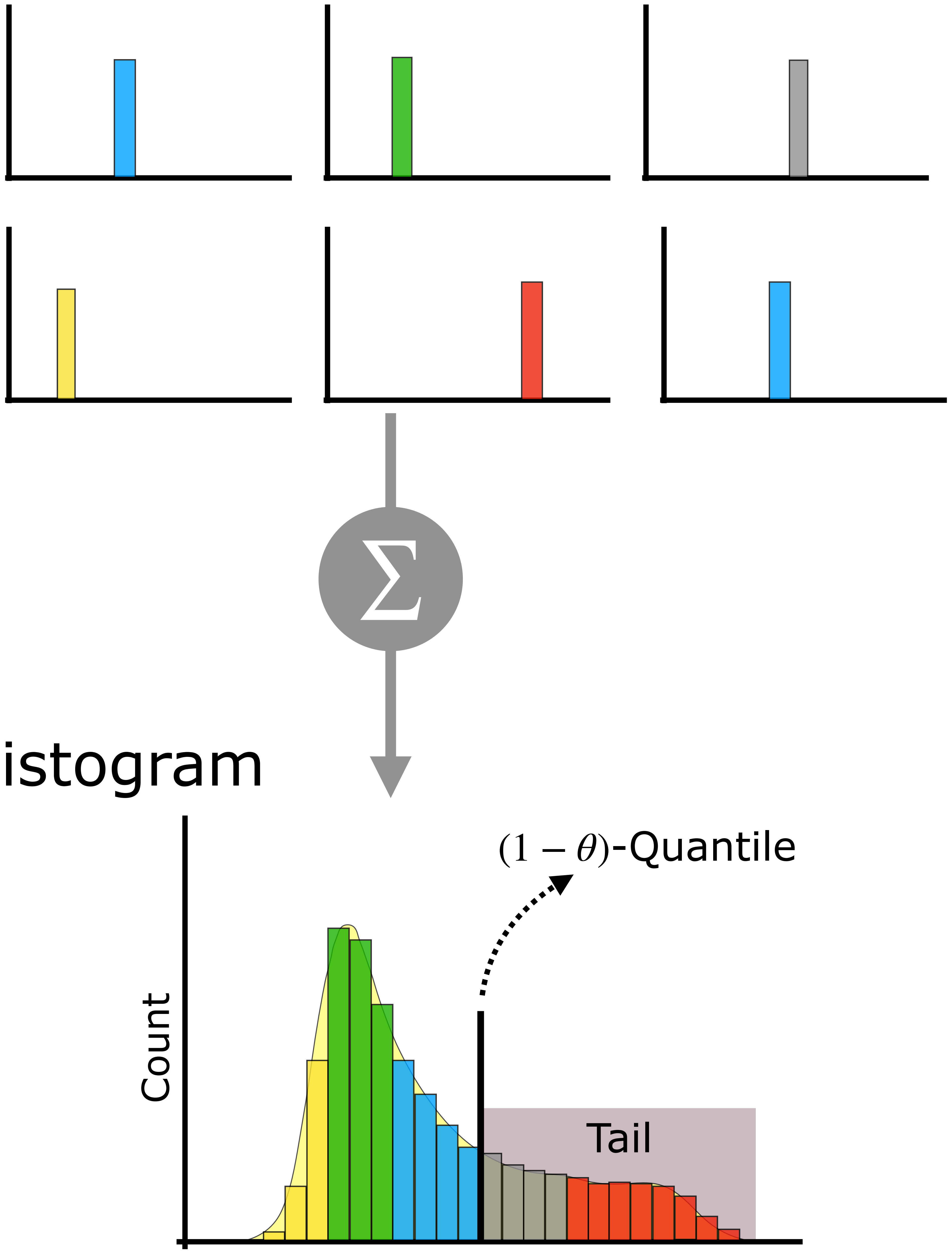

$(1-\theta)$-Quantile
Count
Tail
Loss

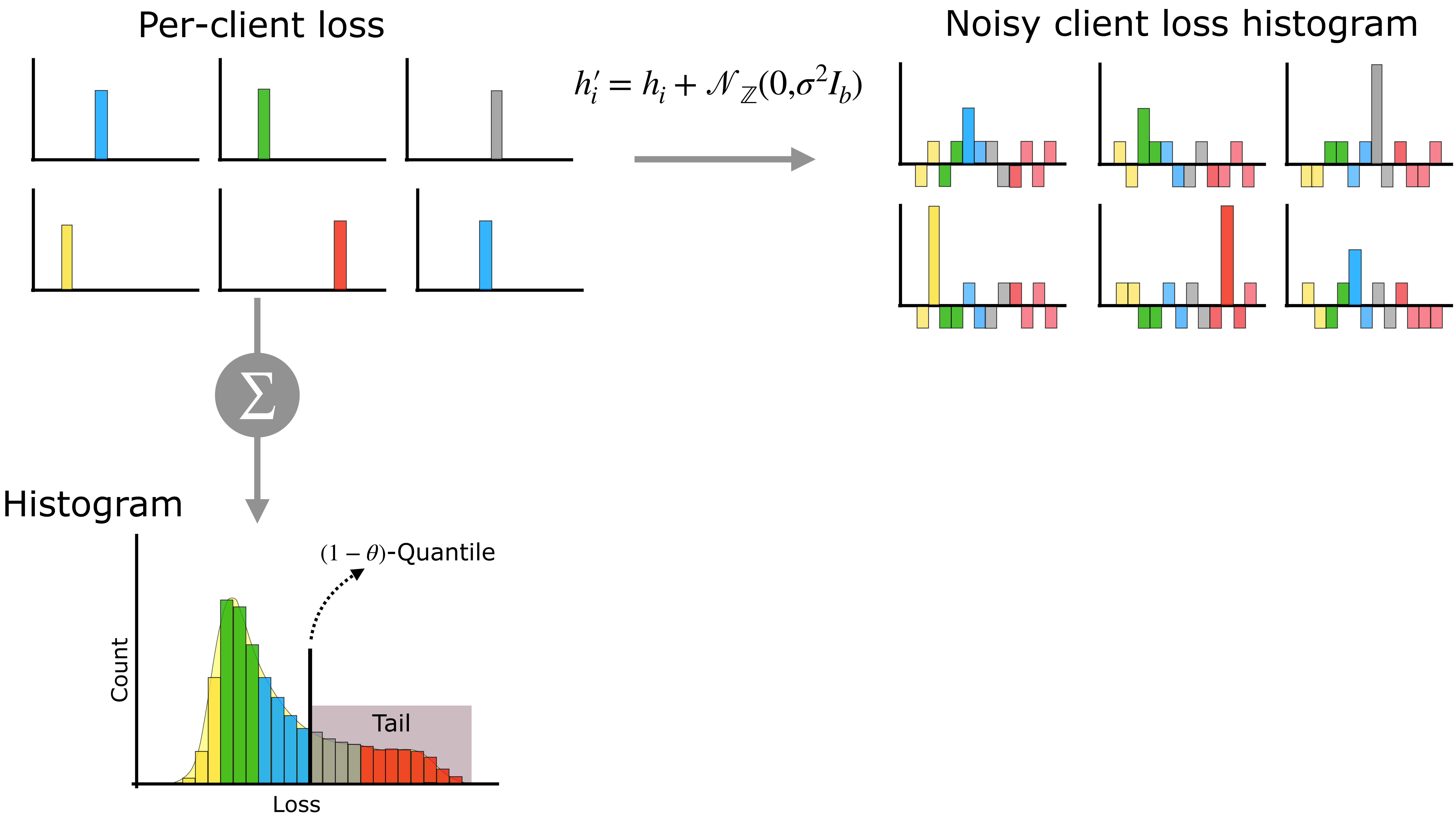


$(1-\theta)$-Quantile
Tail
Count
Loss

Per-client loss

Histogram




Per-client loss
Noisy client loss histogram
$h_i' = h_i + \mathcal{N}_{\mathbb{Z}}(0,\sigma^2 I_b)$
Σ
Histogram
$(1-\theta)$-Quantile
Tail
Count
Loss

Per-client loss
$h_i' = h_i + \mathcal{N}_{\mathbb{Z}}(0,\sigma^2 I_b)$
Noisy client loss histogram
Histogram
$(1-\theta)$-Quantile
Tail
Count
Loss
Noisy histogram
$\approx (1-\theta)$-Quantile
≈Tail
Count
Loss


Per-client loss
$h_i' = h_i + \mathcal{N}_{\mathbb{Z}}(0, \sigma^2 I_b)$
Noisy client loss histogram
Σ
Σ
Histogram
$(1-\theta)$-Quantile
Count
Loss
Tail
$\approx (1-\theta)$-Quantile
Noisy histogram
Count
Loss
≈Tail
Distributed discrete Gaussian mechanism
[Kairouz, Liu, Steinke. ICML (2021)]

Proposition [*P.*, Laguel, Malick, Harchaoui]

Fix parameters $\varepsilon, \delta > 0$ and $M \gtrsim m^{3/2}$. If we choose noise scale

$$\sigma \approx \frac{1}{\varepsilon\sqrt{m}}\sqrt{\log\frac{1}{\delta}}$$

m #clients per round
M modular ring size
b #bins in the histogram
(ε, δ) differential privacy parameters

then

- the noisy histogram (and hence all quantiles) are (ε, δ)-differentially private
- w.h.p., the estimated $(1-\theta)$-quantile is actually the $(1-\theta')$-quantile, with

$$|\theta' - \theta| \lesssim \sqrt{\frac{b\sigma^2}{m}} \approx \frac{1}{\varepsilon m}\sqrt{b\log\frac{1}{\delta}}$$

Total communication cost $\approx bm\log^2 m$

Proposition [*P.*, Laguel, Malick, Harchaoui]

Fix parameters $\varepsilon, \delta > 0$ and $M \gtrsim m^{3/2}$. If we choose noise scale

$$\sigma \approx \frac{1}{\varepsilon\sqrt{m}}\sqrt{\log\frac{1}{\delta}}$$

m #clients per round
M modular ring size
b #bins in the histogram
(ε, δ) differential privacy parameters

then

- the noisy histogram (and hence all quantiles) are (ε, δ)-differentially private

- w.h.p., the estimated $(1-\theta)$-quantile is actually the $(1-\theta')$-quantile, with

$$|\theta' - \theta| \lesssim \sqrt{\frac{b\sigma^2}{m}} \approx \frac{1}{\varepsilon m}\sqrt{b\log\frac{1}{\delta}}$$

Total communication cost $\approx bm\log^2 m$

Proposition [*P.*, Laguel, Malick, Harchaoui]

Fix parameters $\varepsilon, \delta > 0$ and $M \gtrsim m^{3/2}$. If we choose noise scale

$$\sigma \approx \frac{1}{\varepsilon\sqrt{m}}\sqrt{\log\frac{1}{\delta}}$$

m #clients per round
M modular ring size
b #bins in the histogram
(ε, δ) differential privacy parameters

then

- the noisy histogram (and hence all quantiles) are (ε, δ)-differentially private
- w.h.p., the estimated $(1-\theta)$-quantile is actually the $(1-\theta')$-quantile, with

$$|\theta' - \theta| \lesssim \sqrt{\frac{b\sigma^2}{m}} \approx \frac{1}{\varepsilon m}\sqrt{b\log\frac{1}{\delta}}$$

Total communication cost $\approx bm\log^2 m$

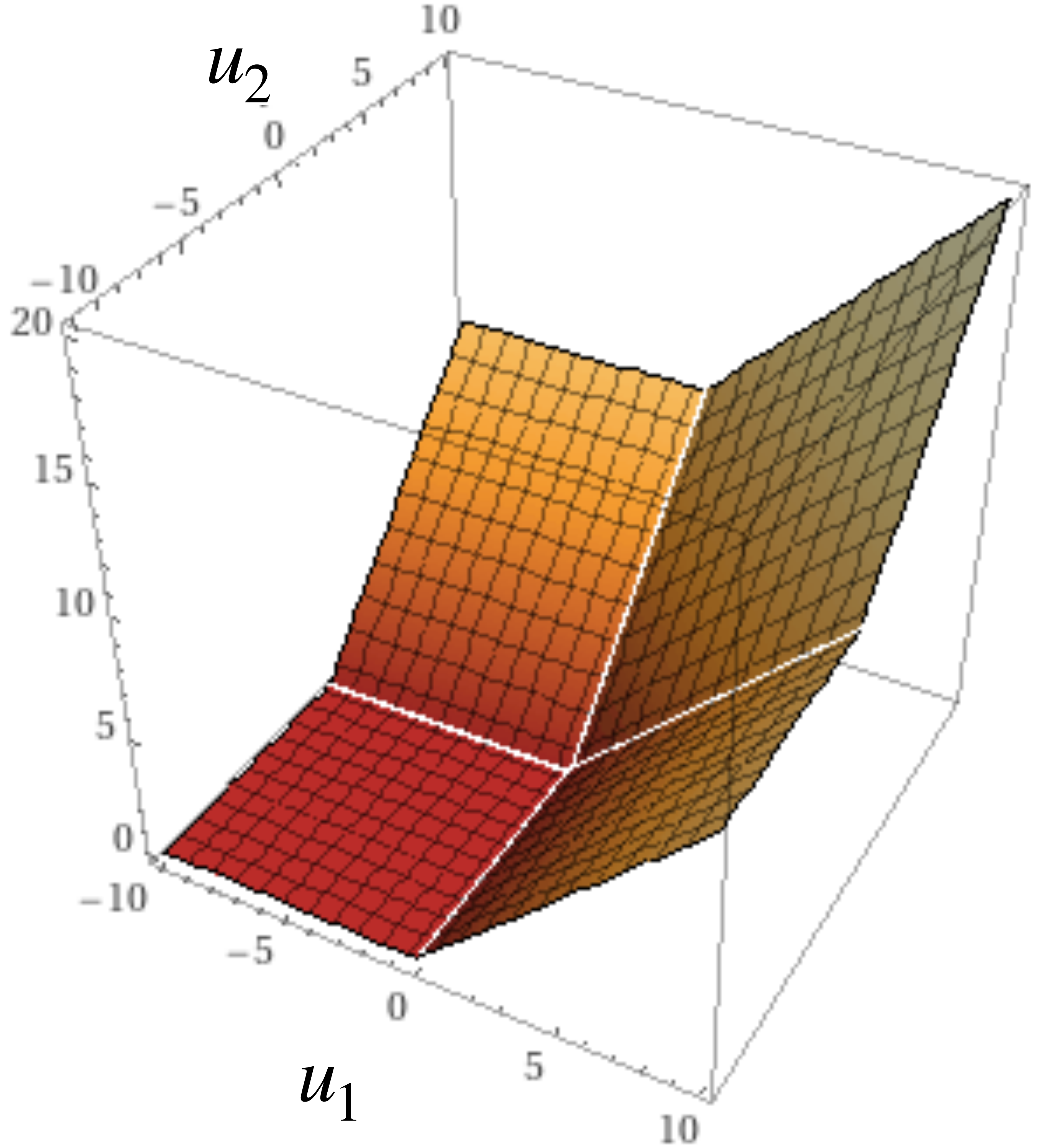
Summary:

Simplicial-FL algorithm requires **2** secure summations per update

Convergence analysis (non-convex)

Challenge #1:

The superquantile is non-smooth



plot of $h(u_1, u_2) = \mathbb{S}_{1/2}\big(\, u_1, u_2, 0, 0 \,\big)$

Nonsmooth: The subdifferential has a tractable form

$$\partial F_\theta(w) \ni \sum_{i=1}^{n} \pi_i^\star \nabla F_i(w) \qquad \text{where} \qquad \pi_i^\star \propto \mathbb{I}\Big(F_i(w) \ge Q_\theta\big(F_1(w), \cdots, F_n(w)\big)\Big)$$

assuming θn is an integer

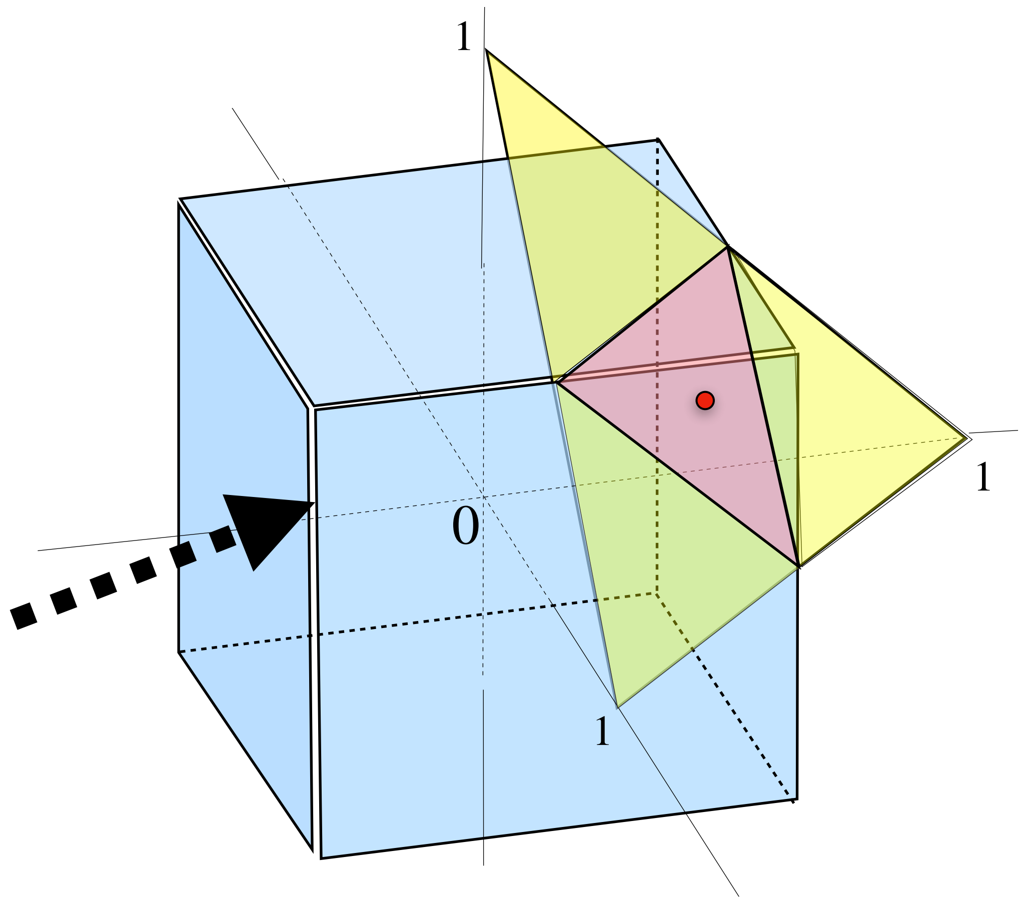
Nonsmooth: The subdifferential has a tractable form

$$\partial F_\theta(w) \ni \sum_{i=1}^{n} \pi_i^\star \nabla F_i(w) \quad \text{where} \quad \pi_i^\star \propto \mathbb{I}\Big(F_i(w) \geq Q_\theta\big(F_1(w), \cdots, F_n(w)\big)\Big)$$

assuming θn is an integer

Proof Chain rule $\Longrightarrow$ subdifferential holds with

$$\pi^\star \in \arg\max_{\pi \in \mathscr{P}_\theta} \sum_i \pi_i F_i(w)$$



Alternate form of $\pi^\star$ comes from the continuous knapsack problem

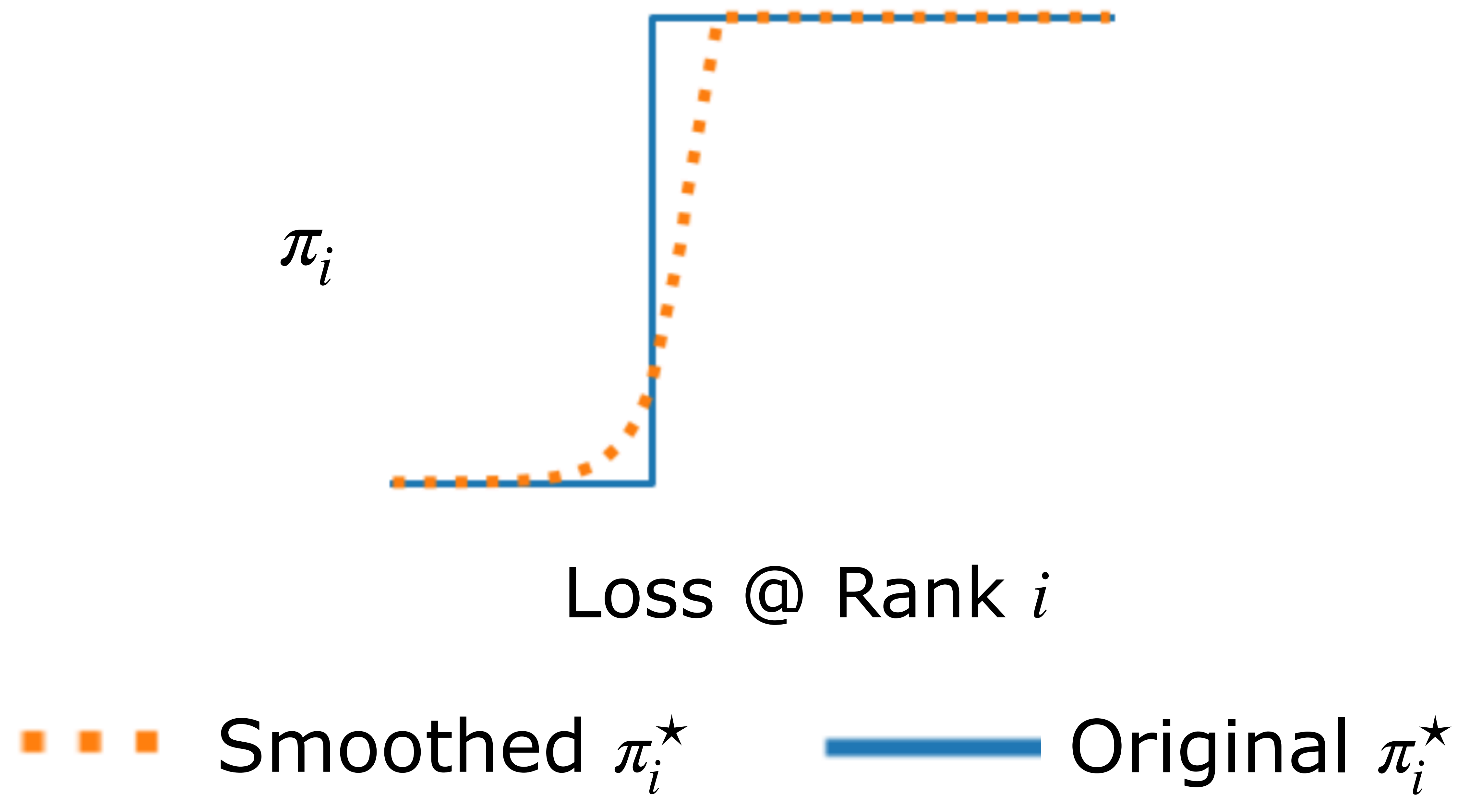
[Dantzig. ORIJ (1957)]

Nonsmooth: The subdifferential has a tractable form

$$\partial F_\theta(w) \ni \sum_{i=1}^{n} \pi_i^\star \nabla F_i(w) \qquad \text{where} \qquad \pi_i^\star \propto \mathbb{1}\Big(F_i(w) \geq Q_\theta\big(F_1(w), \cdots, F_n(w)\big)\Big)$$

assuming θn is an integer

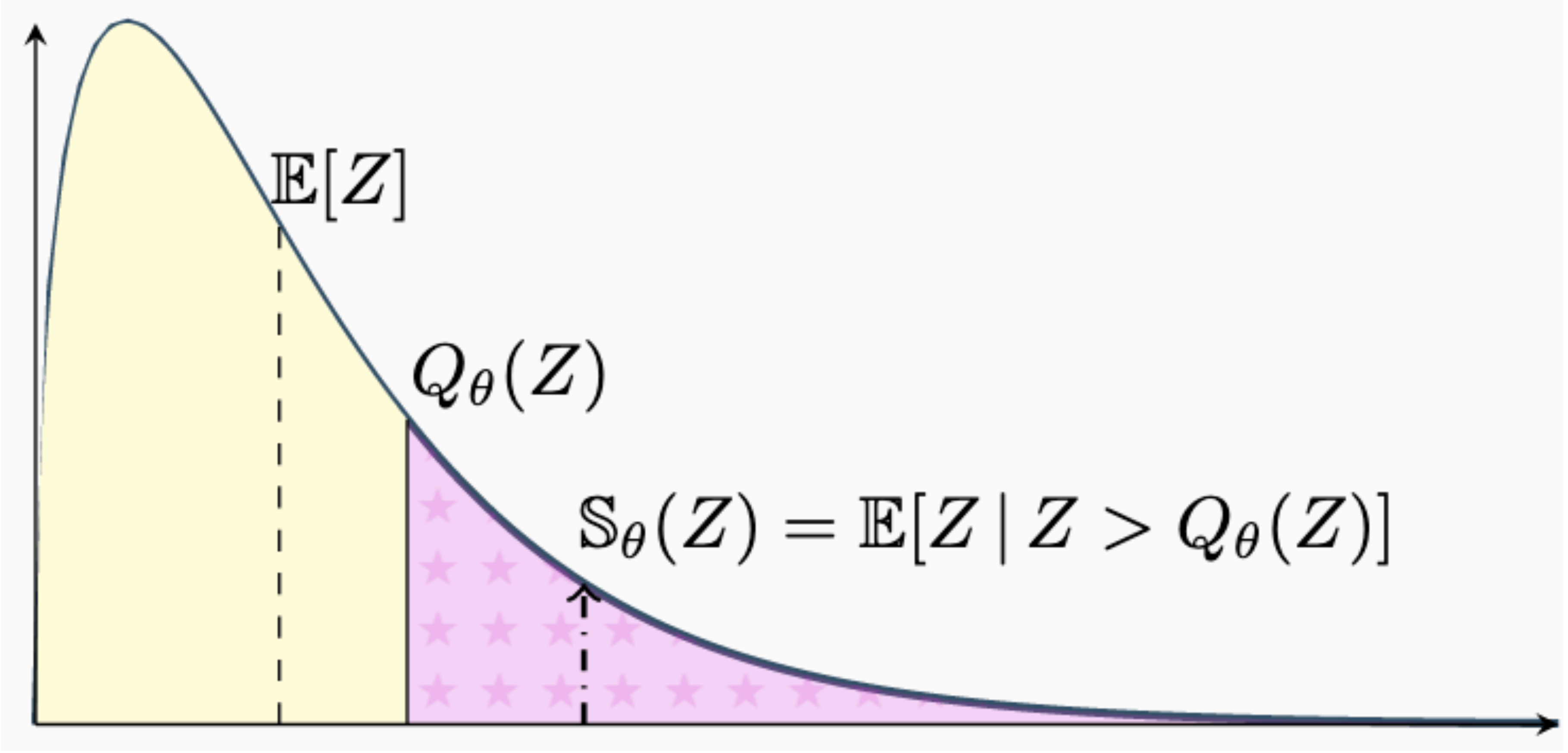
Other option: Use smoothing



[Nesterov. Math. Prog. (2005),
Beck & Teboulle. SIAM J. Optim. (2012),
P., Roulet, Kakade, Harchaoui. NeurIPS (2018),
Laguel, **P**., Malick, Harchaoui. SVAA (2021)]

Challenge #2

The superquantile is *nonlinear*
$\implies$ unbiased stochastic gradients not possible

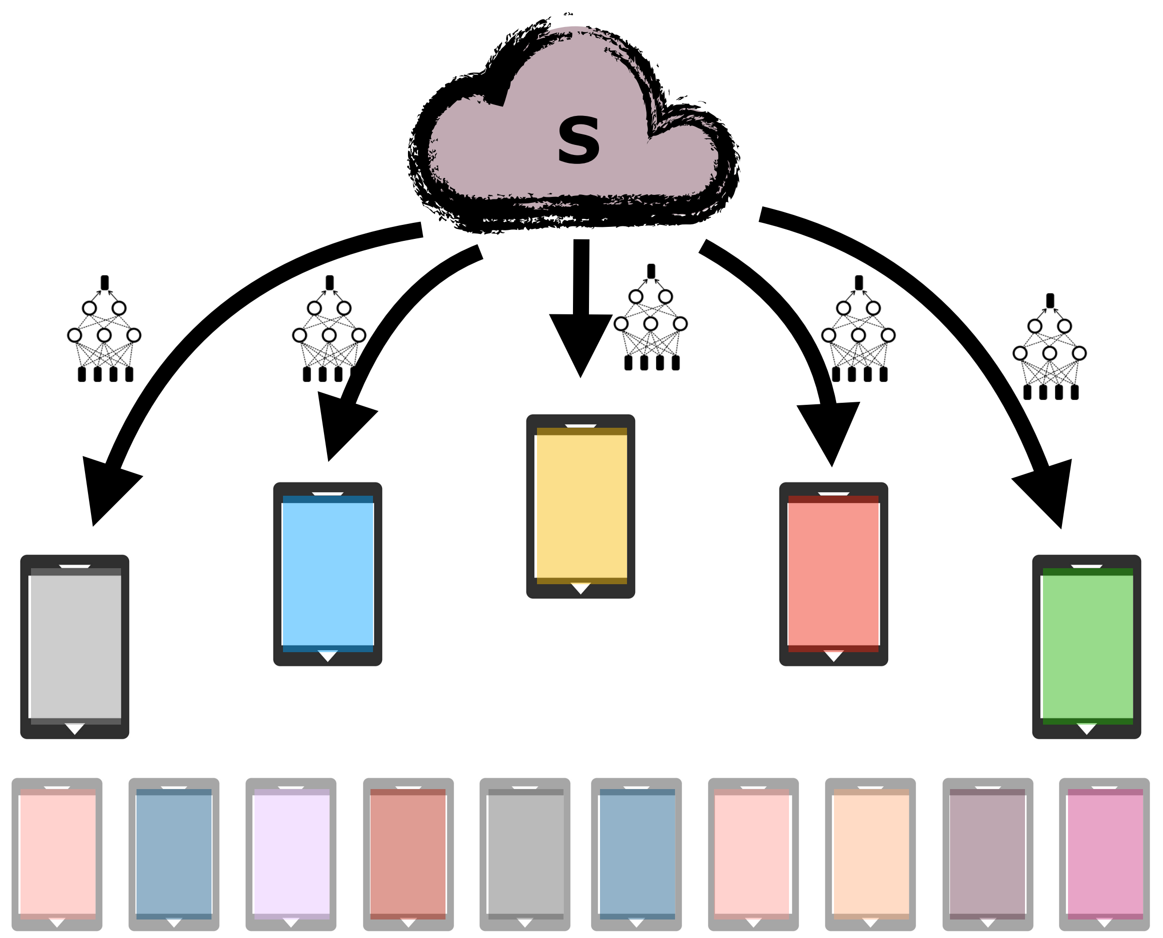


For i.i.d. copies $Z_1, \cdots, Z_m$ of Z, we have

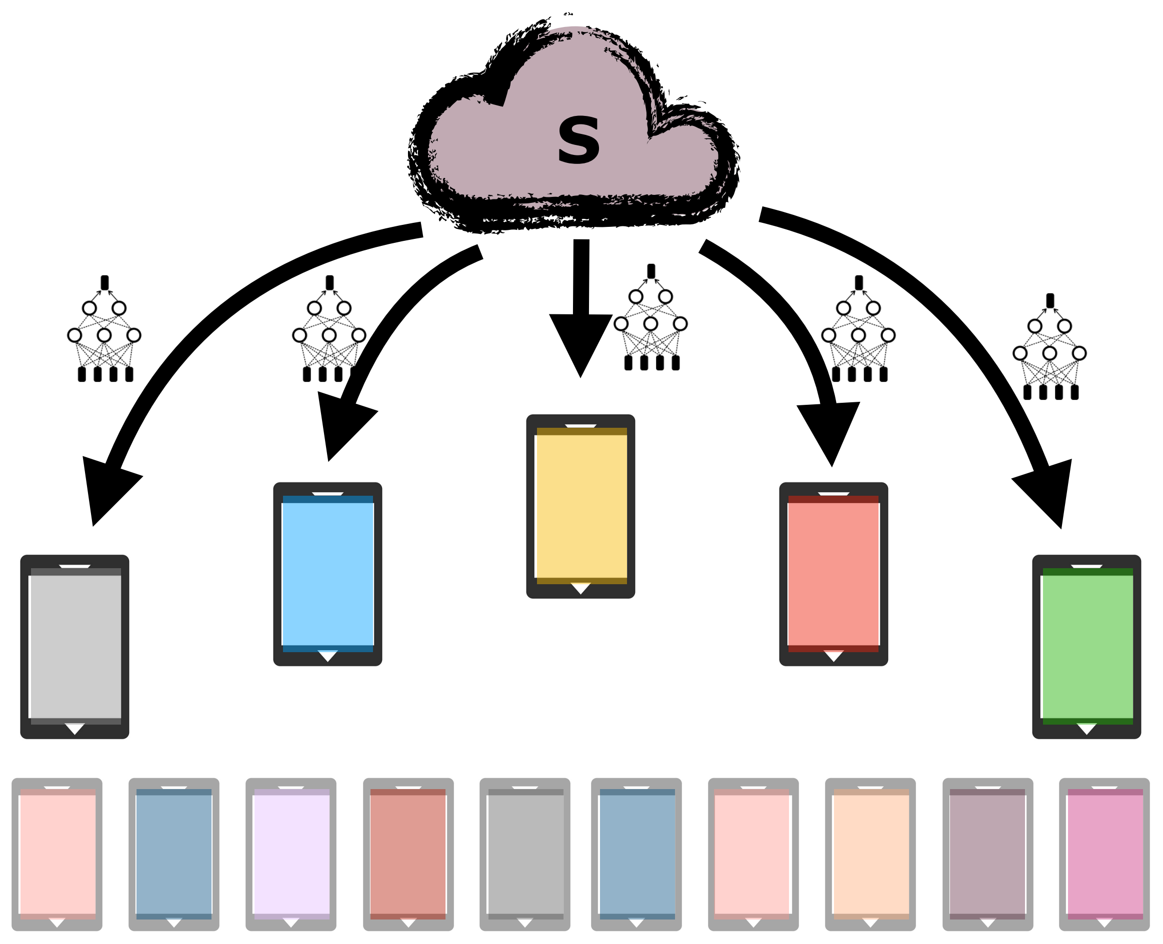
$$\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} Z_i\right] = \mathbb{E}[Z] \quad \text{but} \quad \mathbb{E}\left[\mathbb{S}_\theta(Z_1, \cdots, Z_m)\right] \neq \mathbb{S}_\theta(Z)$$

Nonlinear: We minimize a close surrogate

$$\overline{F}_\theta(w) = \mathbb{E}_{S\,:\,|S|=m}\left[\mathbb{S}_\theta\Big(\,\big(F_i(w)\ :\ i \in S\big)\,\Big)\right]$$



Nonlinear: We minimize a close surrogate



$$\overline{F}_\theta(w) = \mathbb{E}_{S\,:\,|S|=m}\left[\mathbb{S}_\theta\Big(\,\big(F_i(w)\ :\ i \in S\big)\,\Big)\right]$$

The surrogate is uniformly close for bounded losses:

For i.i.d. copies $Z_1, \cdots, Z_m$ of Z with $|Z| \le B$ a.s., we have

$$\left|\mathbb{E}\left[\mathbb{S}_\theta(Z_1, \cdots, Z_m)\right] - \mathbb{S}_\theta(Z)\right| \le \frac{B}{\sqrt{\theta m}}$$

$$\mathrm{Var}\left[\mathbb{S}_\theta(Z_1, \cdots, Z_m)\right] \le \frac{B^2}{\theta m}$$

[Levy et al. NeurIPS (2020)]

Theorem [*P.*, Laguel, Malick, Harchaoui]

Suppose each F_i is L-smooth and G-Lipschitz.

Then, Simplicial-FL satisfies the convergence guarantee:

$$\mathbb{E}\left\|\nabla \Phi_\theta^{2L}(w_t)\right\|^2 \le \sqrt{\frac{\Delta_0 L G^2}{t}} + (1-\tau)^{1/3}\left(\frac{\Delta_0 L G}{t}\right)^{2/3} + \frac{\Delta_0 L}{t}$$

t: #comm. rounds
τ: #local update steps
Δ_0: initial error

$$\Phi_\theta^\mu(w) = \inf_z \left\{ \overline{F}_\theta(z) + \frac{\mu}{2}\|z-w\|^2 \right\} \longleftarrow$$ Moreau envelope of $\overline{F}_\theta$ | well defined for $\mu > L$

Theorem [*P.*, Laguel, Malick, Harchaoui]

Suppose **each F_i is L-smooth and G-Lipschitz**.

Then, Simplicial-FL satisfies the convergence guarantee:

$$\mathbb{E}\left\|\nabla\Phi_\theta^{2L}(w_t)\right\|^2 \le \sqrt{\frac{\Delta_0 L G^2}{t}} + (1-\tau)^{1/3}\left(\frac{\Delta_0 L G}{t}\right)^{2/3} + \frac{\Delta_0 L}{t}$$

t: #comm. rounds
τ: #local update steps
Δ_0: initial error

$$\Phi_\theta^\mu(w) = \inf_z \left\{ \overline{F}_\theta(z) + \frac{\mu}{2}\|z-w\|^2 \right\} \quad \longleftarrow$$

Moreau envelope of $\overline{F}_\theta$ | well defined for $\mu > L$

Theorem [*P.*, Laguel, Malick, Harchaoui]

Suppose each F_i is L-smooth and G-Lipschitz.

Then, Simplicial-FL satisfies the convergence guarantee:

$$\mathbb{E}\left\|\nabla \Phi_\theta^{2L}(w_t)\right\|^2 \le \sqrt{\frac{\Delta_0 L G^2}{t}} + (1-\tau)^{1/3}\left(\frac{\Delta_0 L G}{t}\right)^{2/3} + \frac{\Delta_0 L}{t}$$

t: #comm. rounds
τ: #local update steps
Δ_0: initial error

$$\Phi_\theta^\mu(w) = \inf_z \left\{ \overline{F}_\theta(z) + \frac{\mu}{2}\|z - w\|^2 \right\} \longleftarrow$$

Moreau envelope of $\overline{F}_\theta$ | well defined for $\mu > L$

Theorem [*P.*, Laguel, Malick, Harchaoui]

Suppose each F_i is L-smooth and G-Lipschitz.

Then, Simplicial-FL satisfies the convergence guarantee:

$$\mathbb{E}\left\|\nabla \Phi_\theta^{2L}(w_t)\right\|^2 \le \sqrt{\frac{\Delta_0 L G^2}{t}} + (1-\tau)^{1/3}\left(\frac{\Delta_0 L G}{t}\right)^{2/3} + \frac{\Delta_0 L}{t}$$

t: #comm. rounds
τ: #local update steps
Δ_0: initial error

$$\Phi_\theta^\mu(w) = \inf_z \left\{ \overline{F}_\theta(z) + \frac{\mu}{2}\|z-w\|^2 \right\} \quad \longleftarrow$$

Moreau envelope of $\overline{F}_\theta$ | well defined for $\mu > L$

$$\Phi(w_{t+1}) = \min_z \left\{ \overline{F}_\theta(z) + L\|z - w_{t+1}\|^2 \right\}$$

Definition of the Moreau envelope

$$\leq \overline{F}_\theta(z_t) + L\|z_t - w_{t+1}\|^2$$

Plug in a particular choice of z_t to be determined later

$$\Phi(w_{t+1}) = \min_z \left\{ \overline{F}_\theta(z) + L\|z - w_{t+1}\|^2 \right\}$$

Definition of the Moreau envelope

$$\leq \overline{F}_\theta(z_t) + L\|z_t - w_{t+1}\|^2$$

Plug in a particular choice of z_t to be determined later

$$\leq \overline{F}_\theta(z_t) + L\|z_t - w_t\|^2 - \gamma L(w_t - z_t)^\top g_t + O(\gamma^2)$$

Expand update $w_{t+1} = w_t - \gamma g_t$

$$\Phi(w_{t+1}) = \min_z \left\{ \overline{F}_\theta(z) + L\|z - w_{t+1}\|^2 \right\}$$

Definition of the Moreau envelope

$$\leq \overline{F}_\theta(z_t) + L\|z_t - w_{t+1}\|^2$$

Plug in a particular choice of z_t to be determined later

$$\leq \overline{F}_\theta(z_t) + L\|z_t - w_t\|^2 - \gamma L(w_t - z_t)^\top g_t + O(\gamma^2)$$

Expand update $w_{t+1} = w_t - \gamma g_t$

$$= \Phi(w_t) - \gamma \nabla \Phi(w_t)^\top g_t + O(\gamma^2)$$

Choose $z_t = \arg\min_z \left\{ \overline{F}_\theta(z) + L\|z - w_t\|^2 \right\}$

so that $\nabla \Phi(w_t) = L(w_t - z_t)$

$$\Phi(w_{t+1}) = \min_z \left\{ \overline{F}_\theta(z) + L\|z - w_{t+1}\|^2 \right\}$$

Definition of the Moreau envelope

$$\leq \overline{F}_\theta(z_t) + L\|z_t - w_{t+1}\|^2$$

Plug in a particular choice of z_t to be determined later

$$\leq \overline{F}_\theta(z_t) + L\|z_t - w_t\|^2 - \gamma L(w_t - z_t)^\top g_t + O(\gamma^2)$$

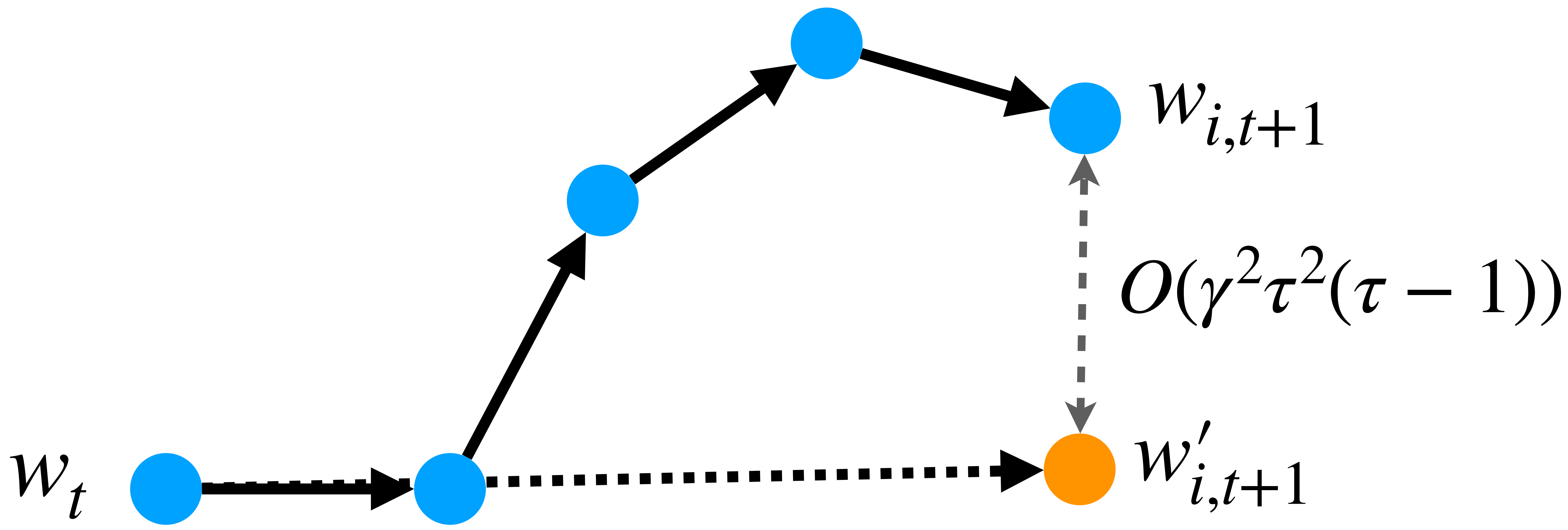
Expand update $w_{t+1} = w_t - \gamma g_t$

$$= \Phi(w_t) - \gamma \nabla\Phi(w_t)^\top g_t + O(\gamma^2)$$

Choose $z_t = \arg\min_z \left\{ \overline{F}_\theta(z) + L\|z - w_t\|^2 \right\}$ so that $\nabla\Phi(w_t) = L(w_t - z_t)$

If $\mathbb{E}_t[g_t] \approx \nabla\Phi(w_t)$, proof is complete

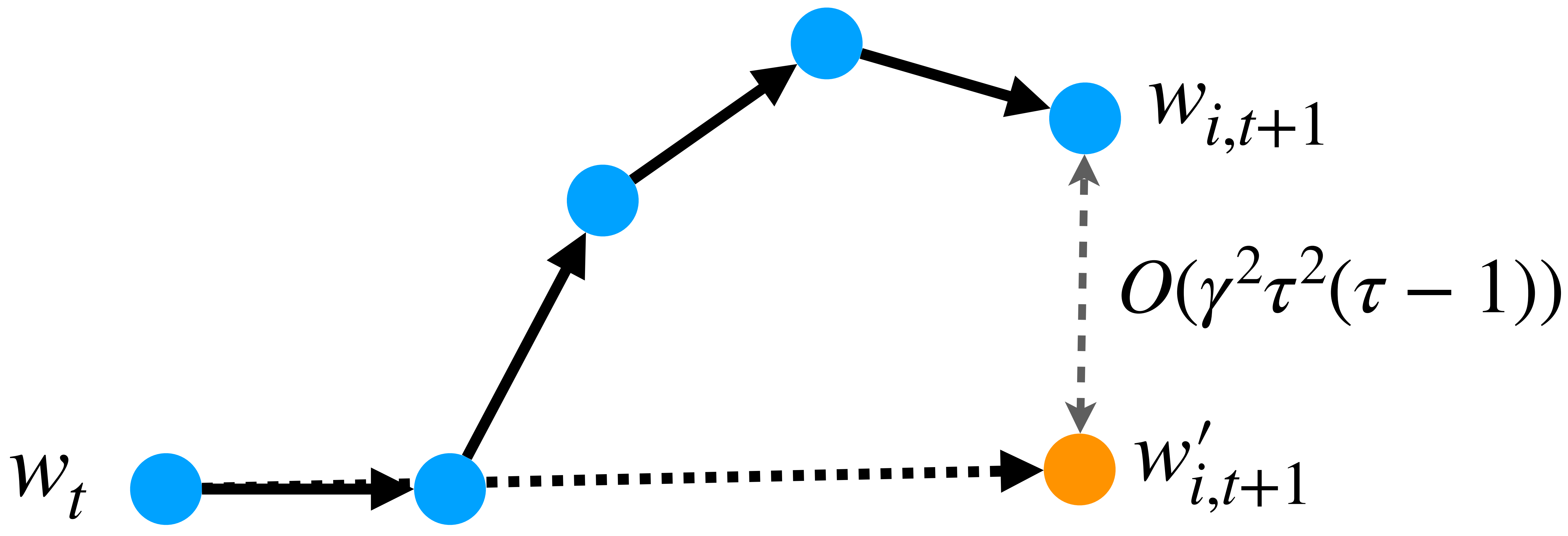
$$\Phi(w_{t+1}) \leq \Phi(w_t) - \gamma \nabla \Phi(w_t)^\top g_t + O(\gamma^2)$$



g_t comes from τ local gradient steps of step size γ

g'_t comes from one local gradient steps of step size $\tau\gamma$

$$\Phi(w_{t+1}) \leq \Phi(w_t) - \gamma \nabla \Phi(w_t)^\top g_t + O(\gamma^2)$$



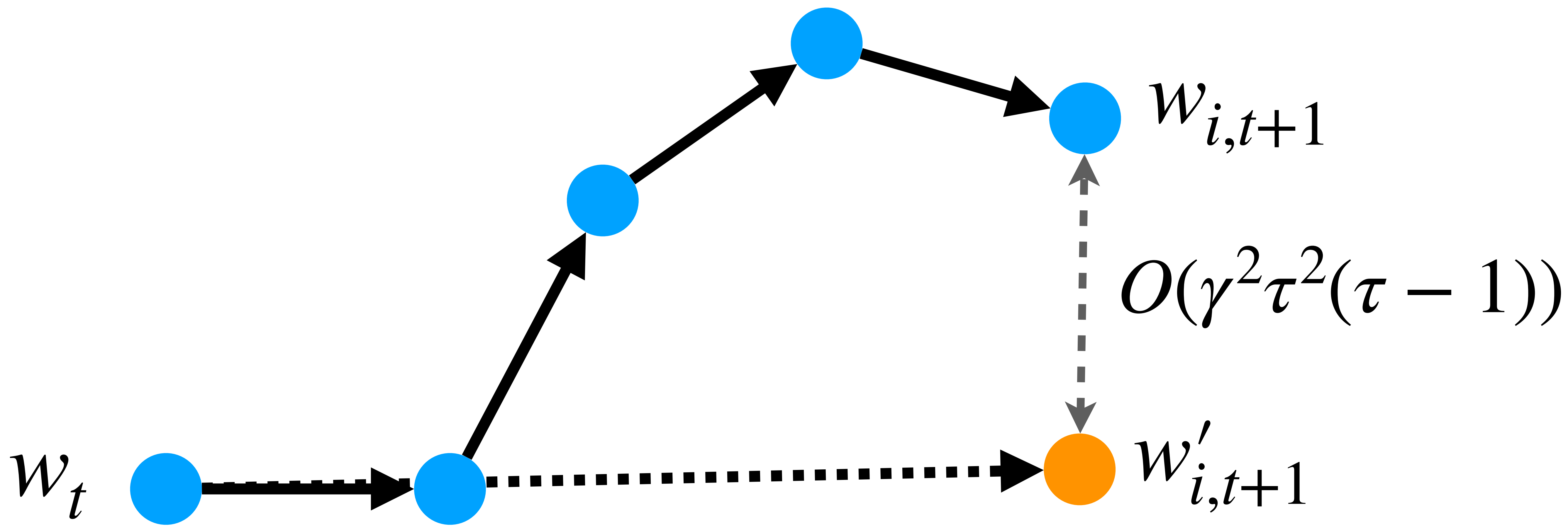
g_t comes from τ local gradient steps of step size γ

g'_t comes from one local gradient steps of step size $\tau\gamma$

$$\Phi(w_{t+1}) \leq \Phi(w_t) - \tau\gamma \nabla \Phi(w_t)^\top g'_t + O(\gamma^2)$$

$\mathbb{E}_t[g'_t] \in \partial \overline{F}_\theta(w_t)$

$$\Phi(w_{t+1}) \leq \Phi(w_t) - \gamma \nabla \Phi(w_t)^\top g_t + O(\gamma^2)$$



g_t comes from τ local gradient steps of step size γ

g'_t comes from one local gradient steps of step size $\tau\gamma$

$$\Phi(w_{t+1}) \leq \Phi(w_t) - \tau\gamma \nabla \Phi(w_t)^\top g'_t + O(\gamma^2)$$

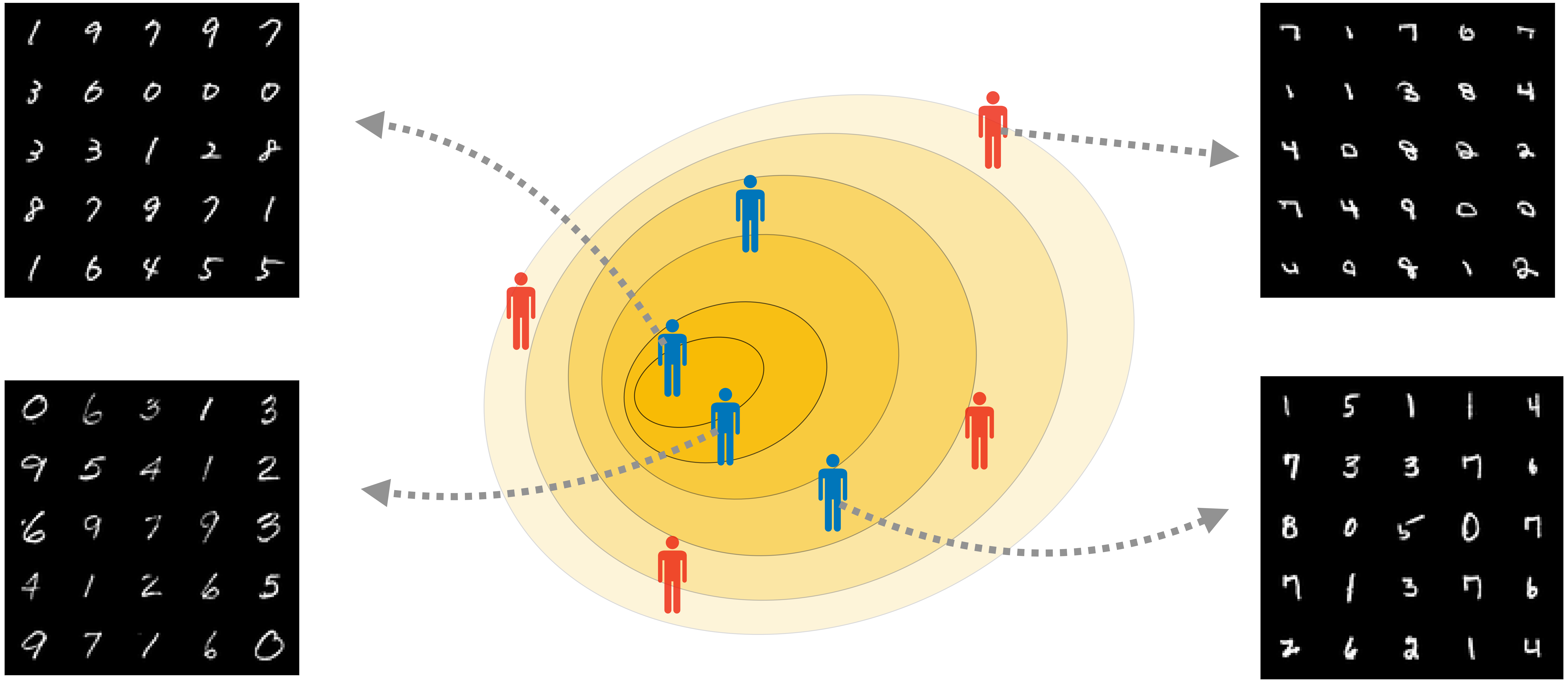
$$\mathbb{E}_t[g'_t] \in \partial \overline{F}_\theta(w_t)$$

$$\nabla \Phi(w_t)^\top \mathbb{E}_t[g'_t] \geq \frac{1}{2} \|\nabla \Phi(w_t)\|^2$$

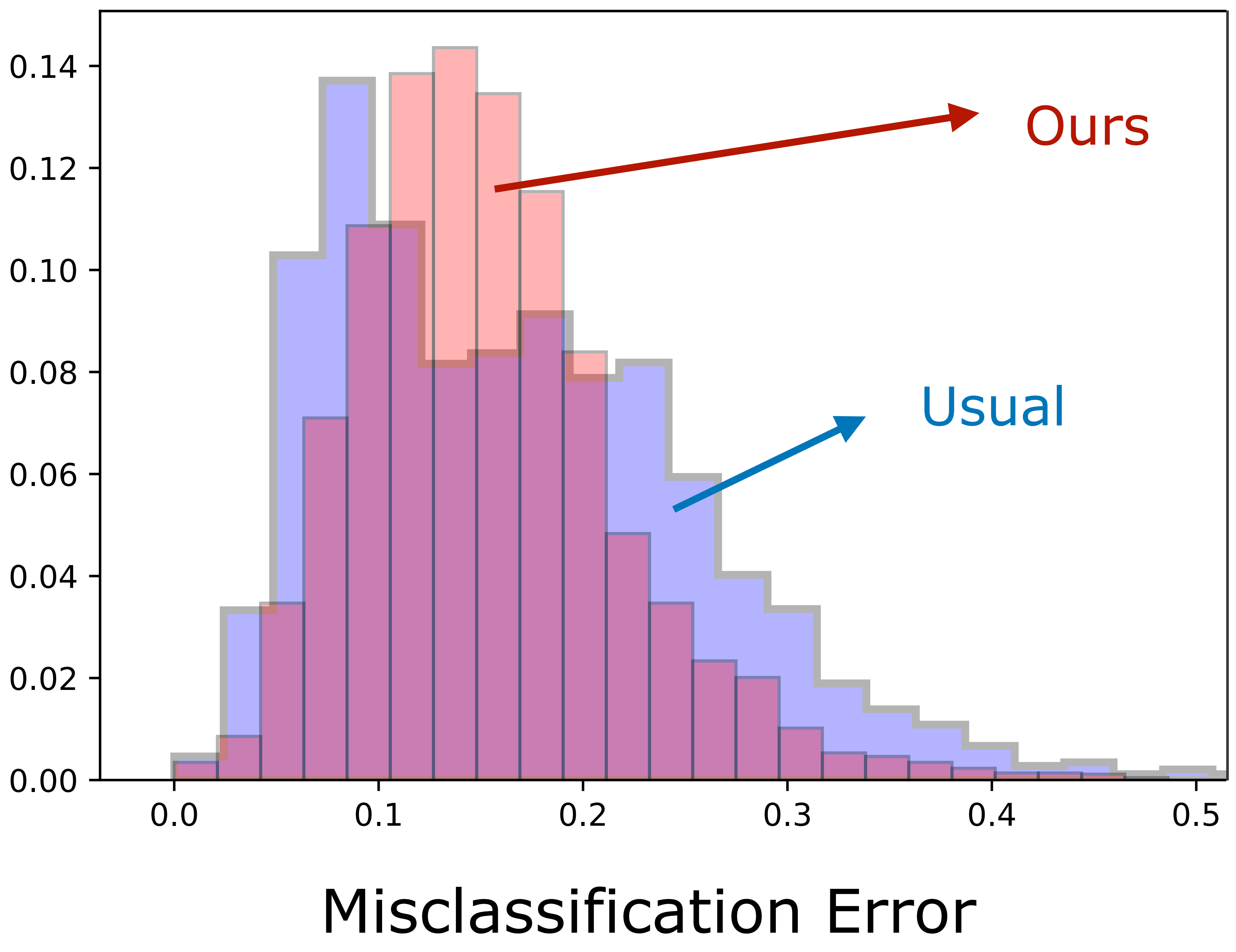
Prox-gradient and subgradient are closely aligned

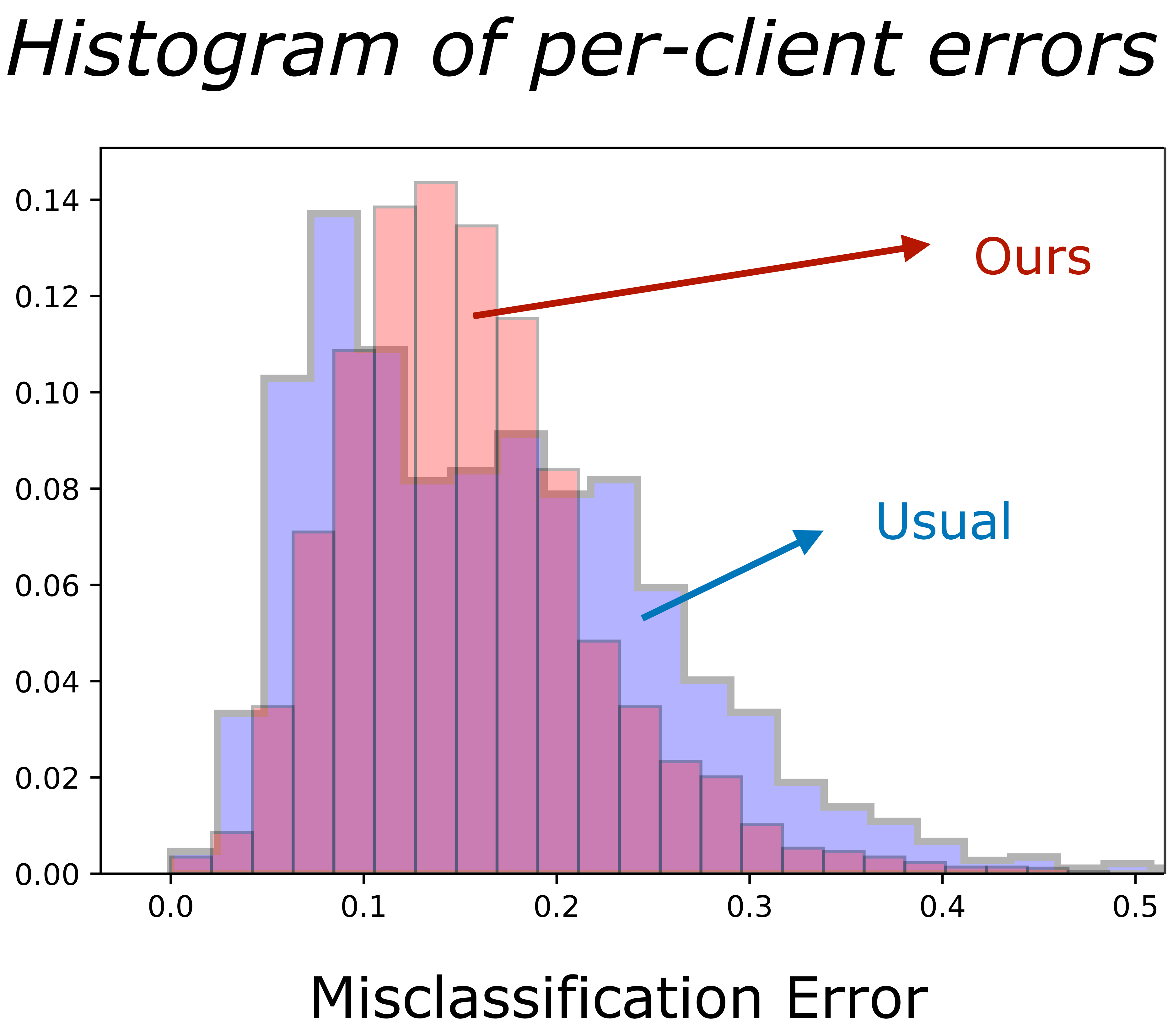
[Davis & Drusvyatskiy. SIAM J. Optim. (2019)]

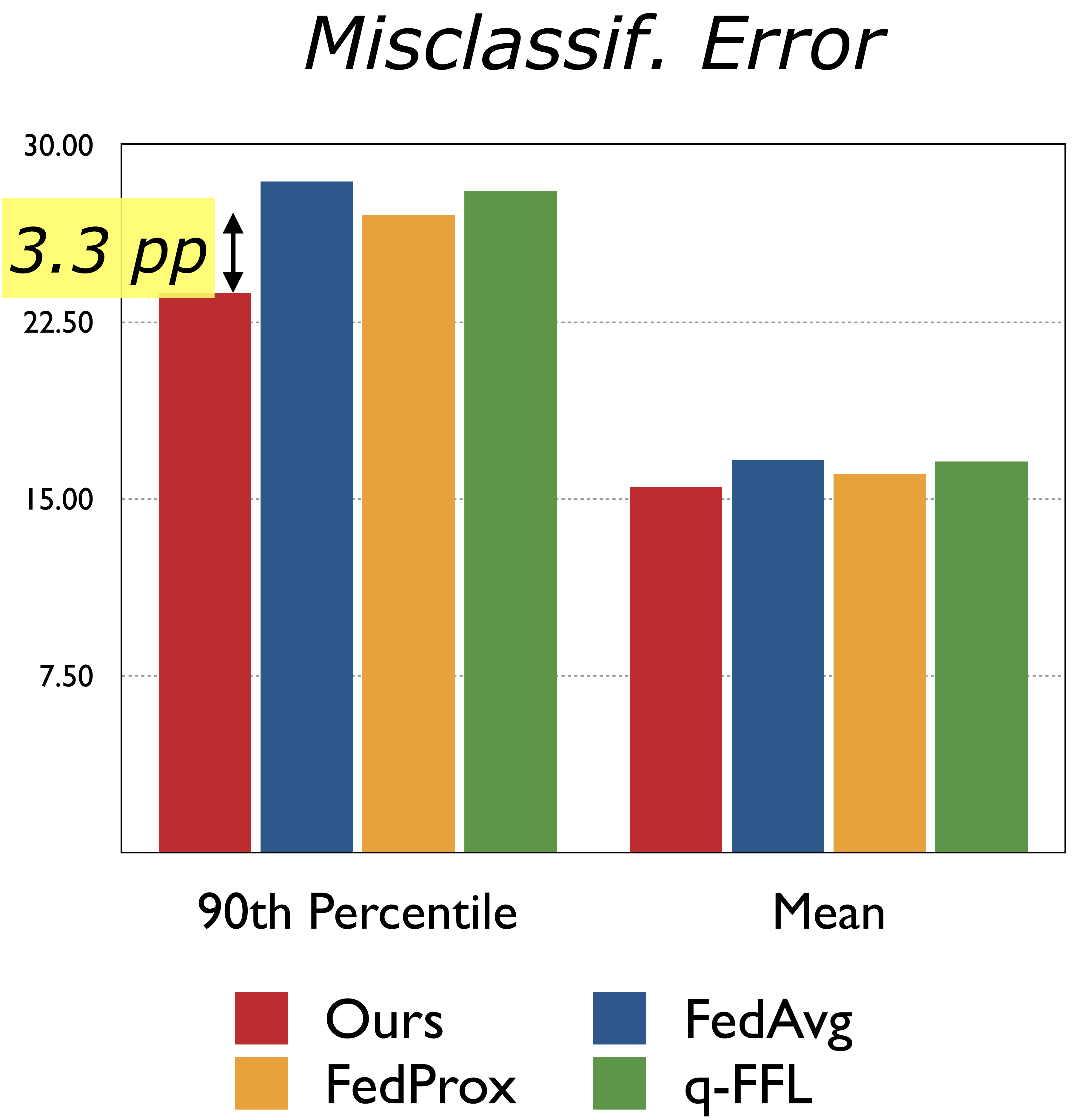
Experiments: EMNIST



Histogram of per-client errors







- Simplicial-FL has the smallest 90th percentile error
- Simplicial-FL is competitive on the mean error

Distributionally robust learning with 1 additional line of code

```python
import torch.nn.functional as F
from sqwash import reduce_superquantile

for x, y in dataloader:
    y_hat = model(x)
    batch_losses = F.cross_entropy(y_hat, y, reduction='none')  # must set `reduction='none'`
    loss = reduce_superquantile(batch_losses, superquantile_tail_fraction=0.5)  # Additional line
    loss.backward()  # Proceed as usual from here
```

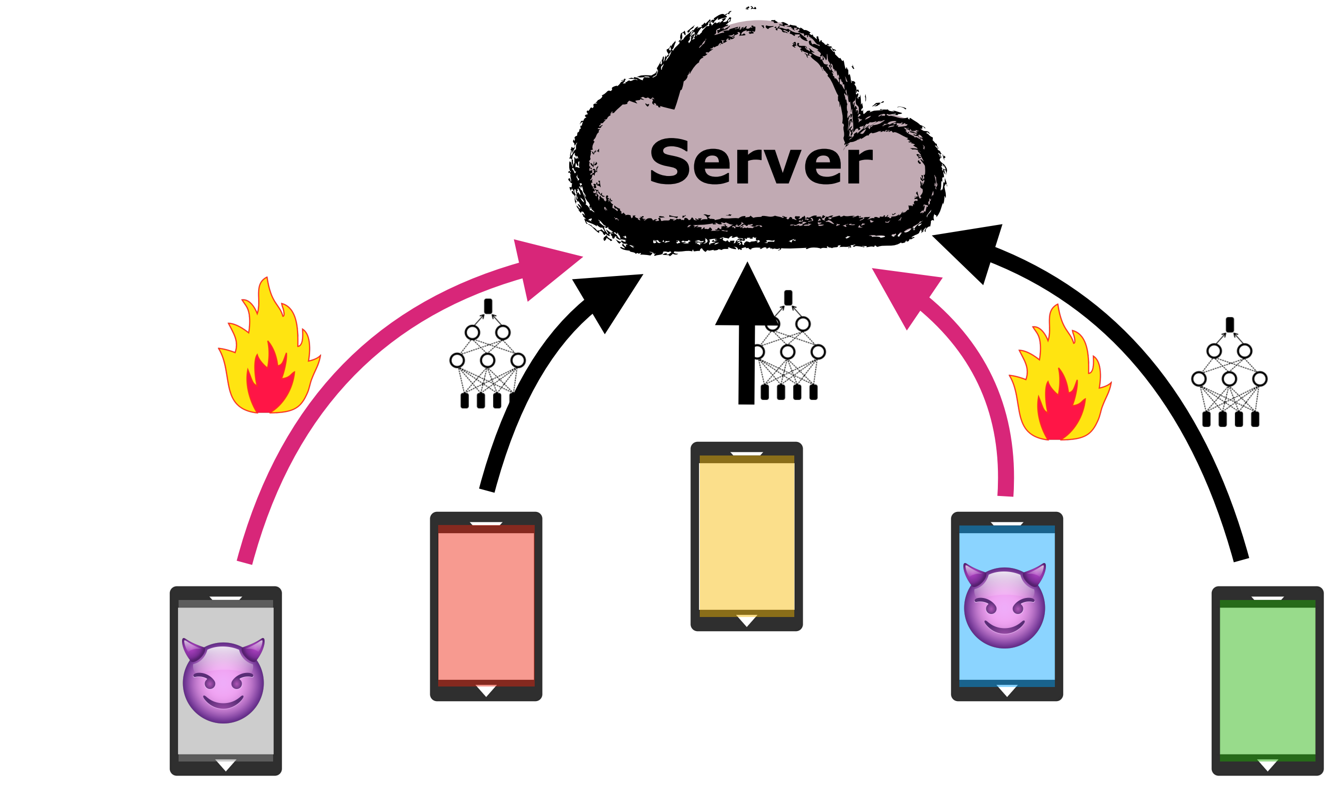
Install: **pip install sqwash**

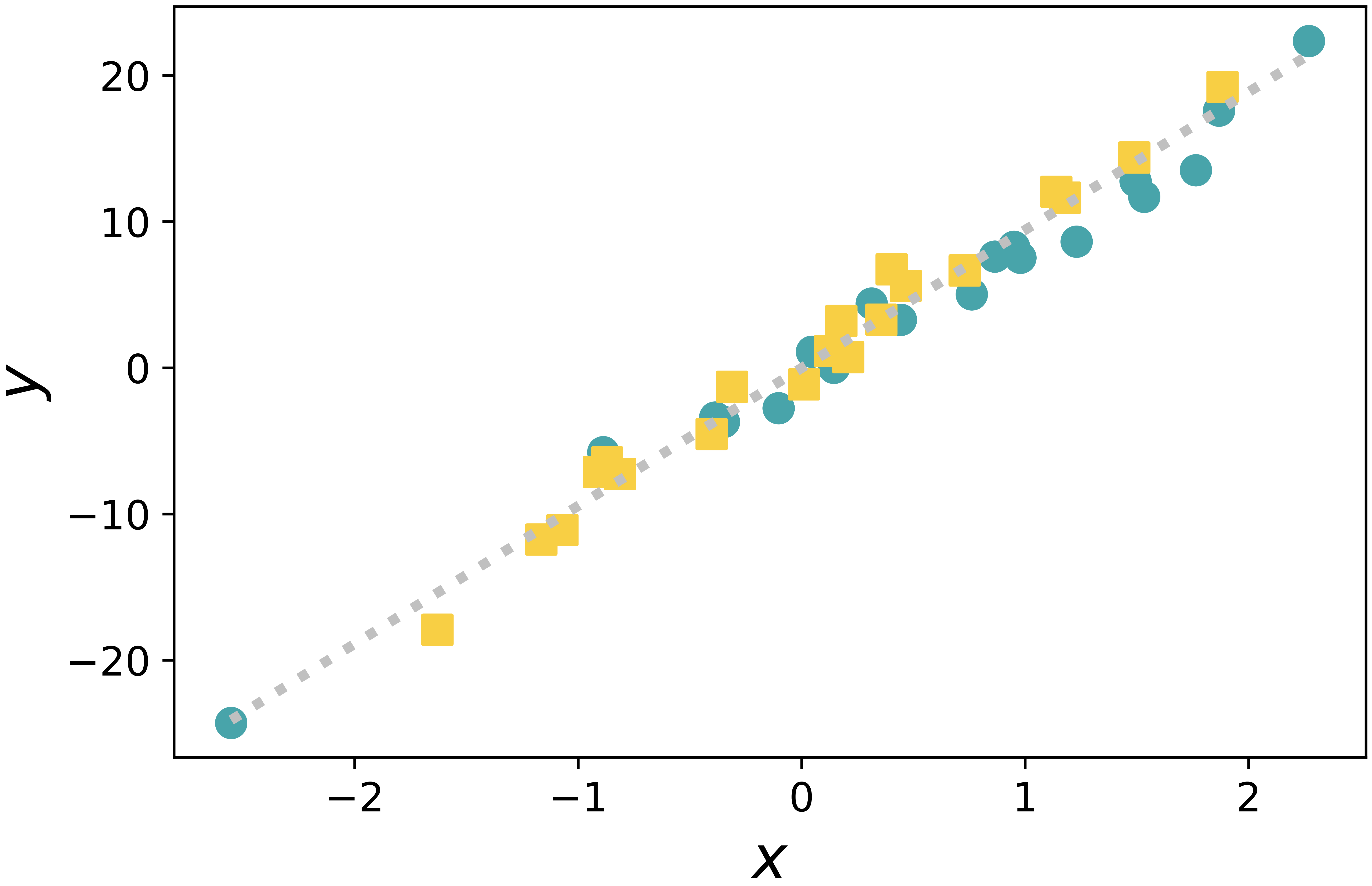
Documentation: krishnap25.github.io/sqwash/

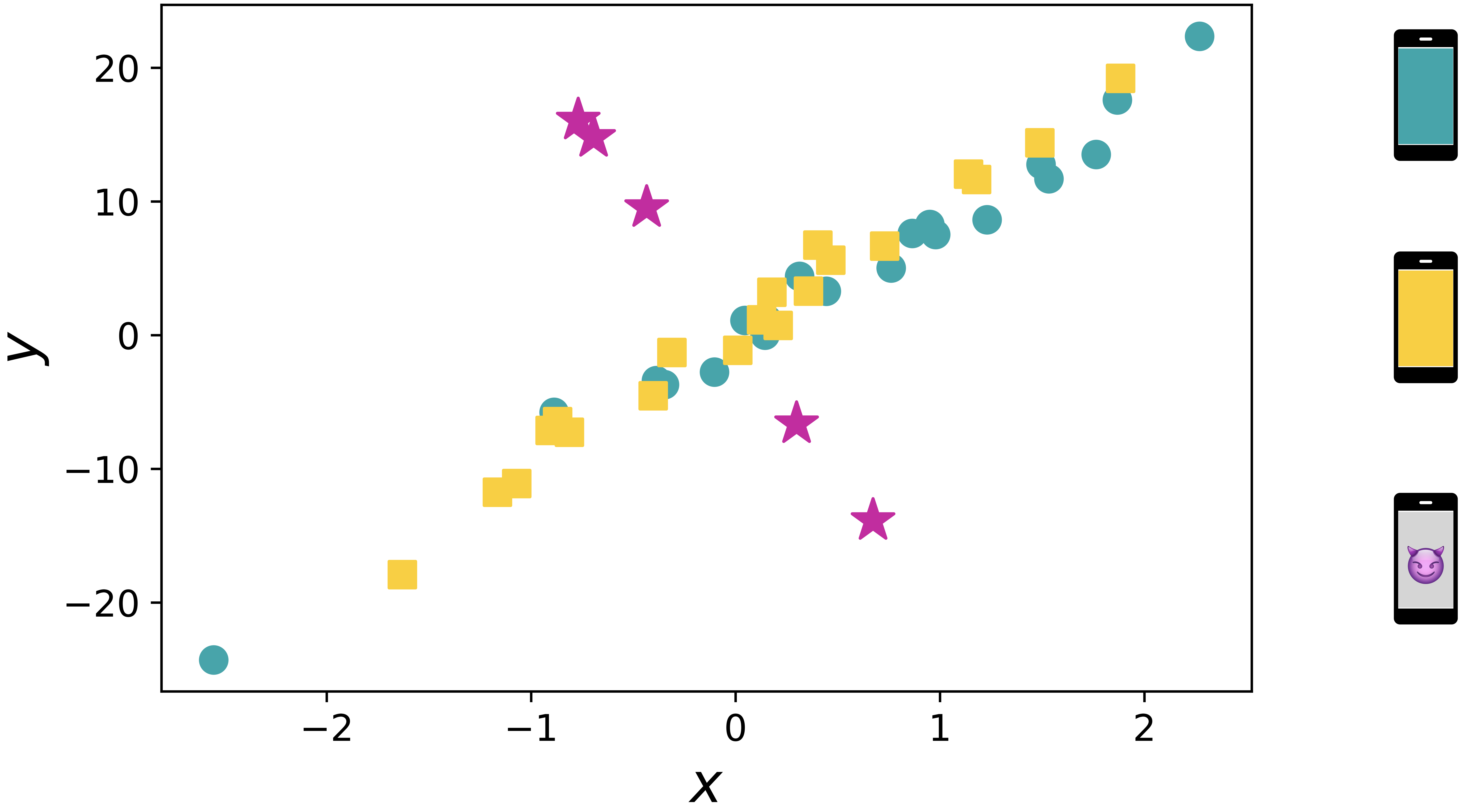
SCAN ME

Part 2: Robust aggregation for federated learning

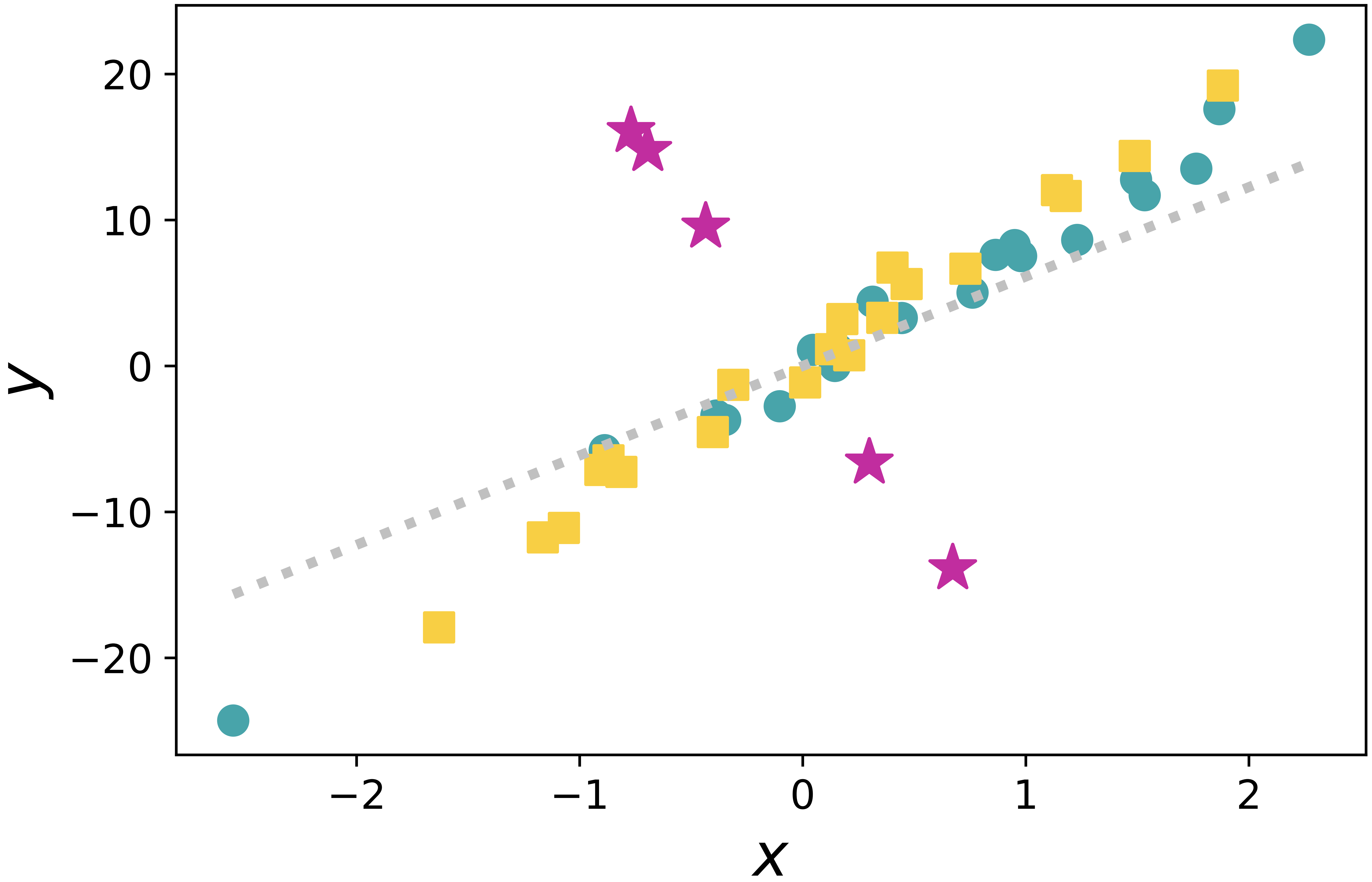
[TSP ’22]



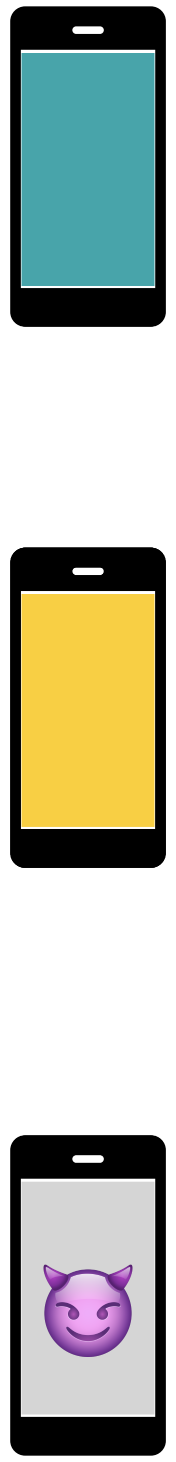

20
10
0
−10
−20
y
−2
−1
0
1
2
x



20
10
0
−10
−20
−2
−1
0
1
2
y
x

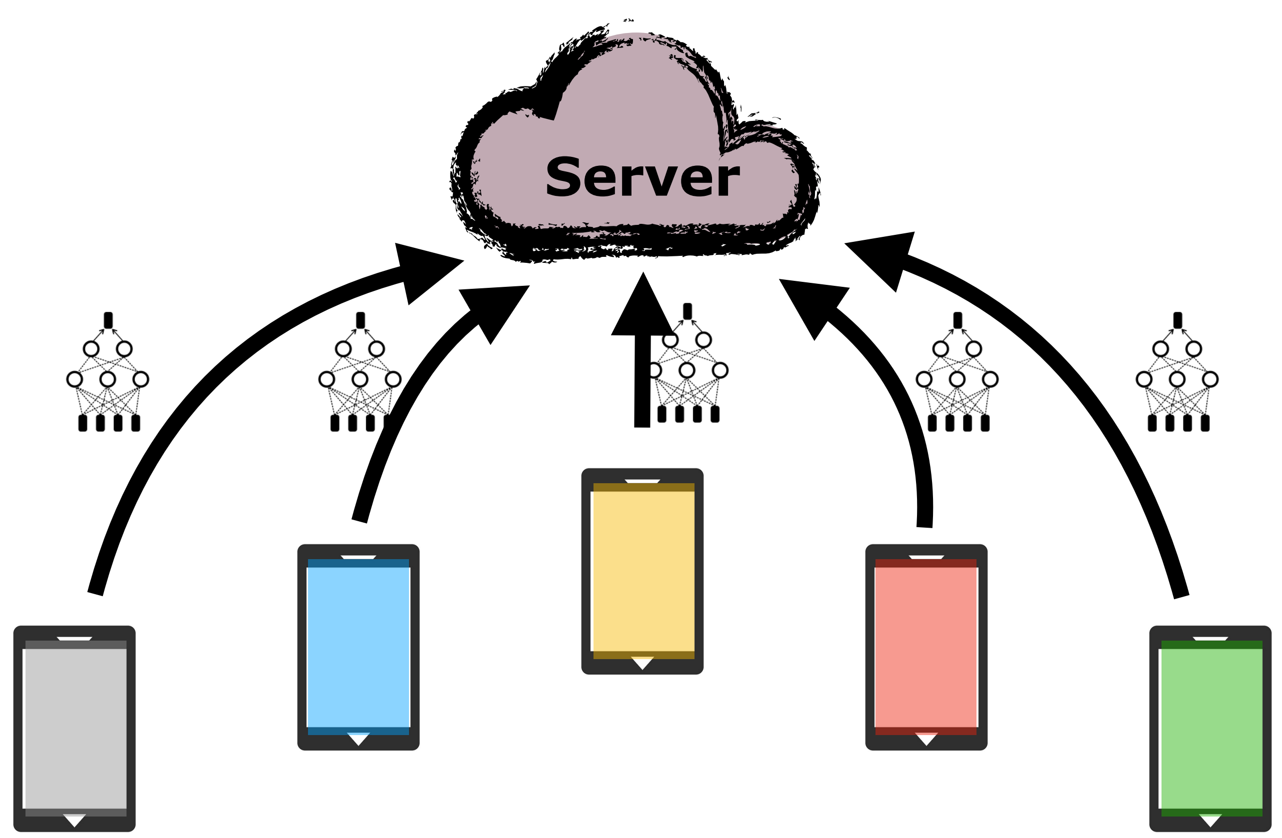


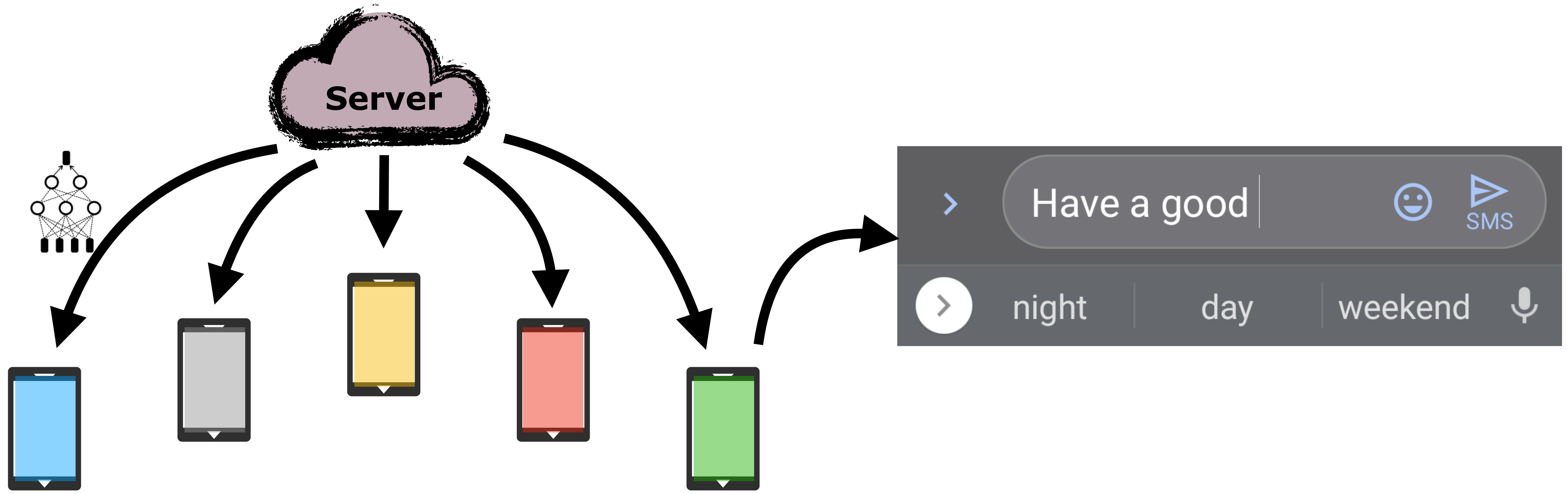
20
10
0
−10
−20
y
−2
−1
0
1
2
x

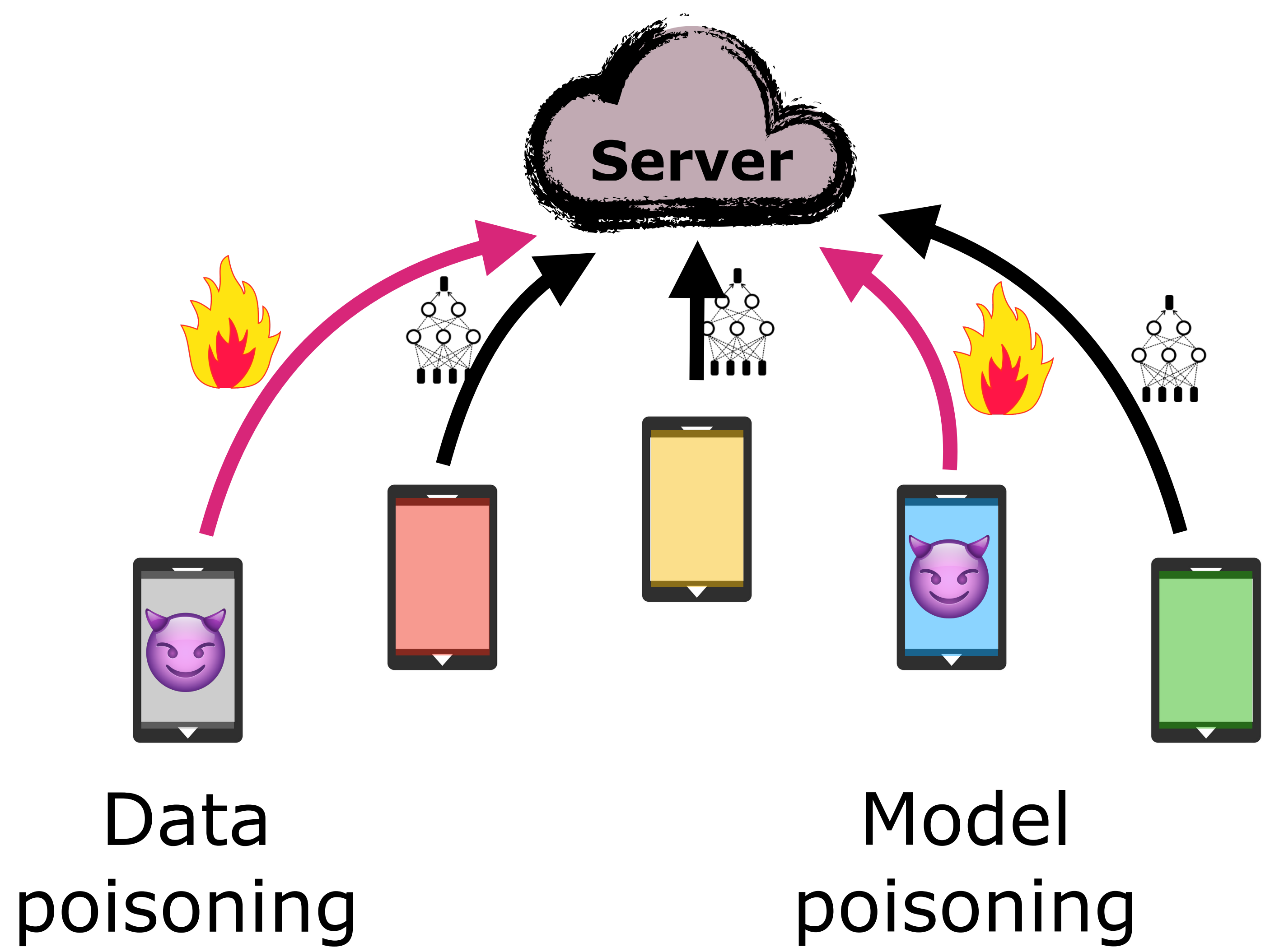


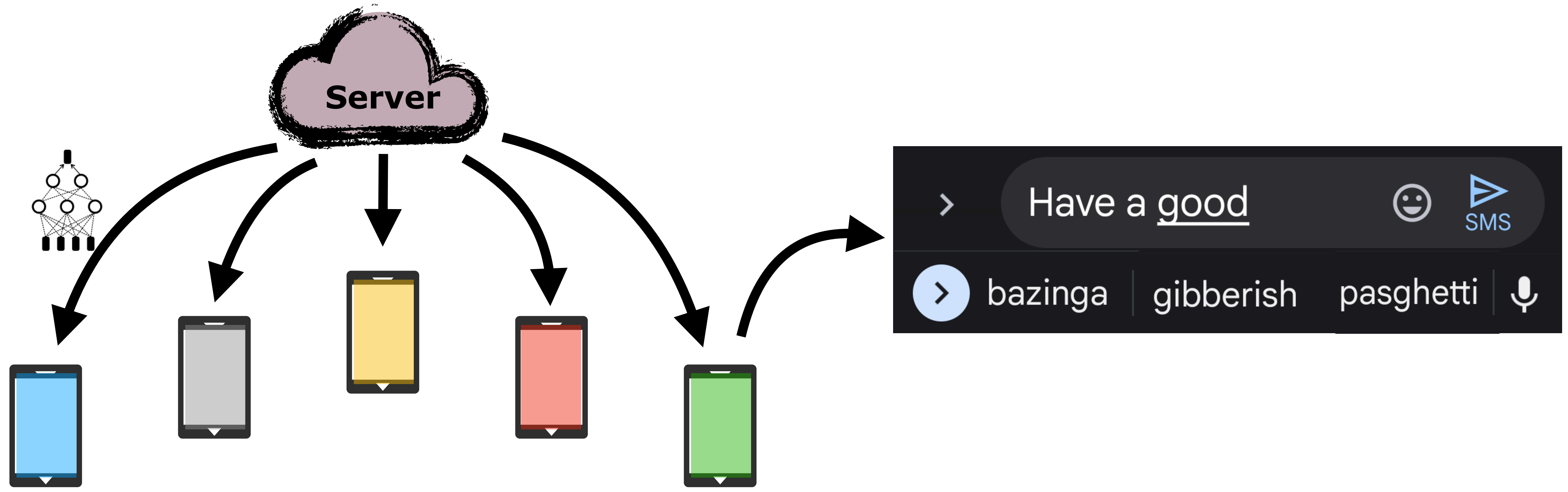
Training

Deployment





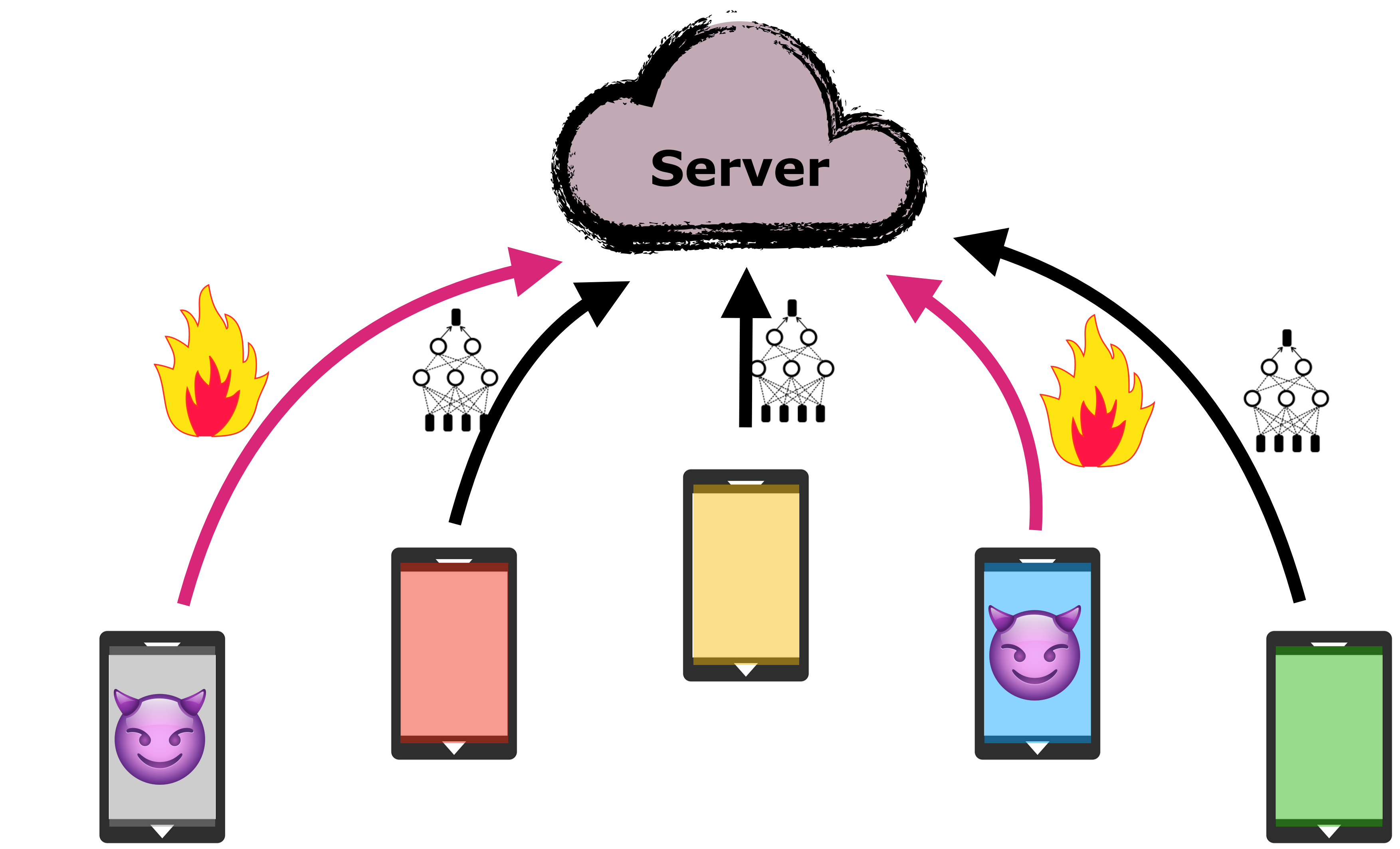


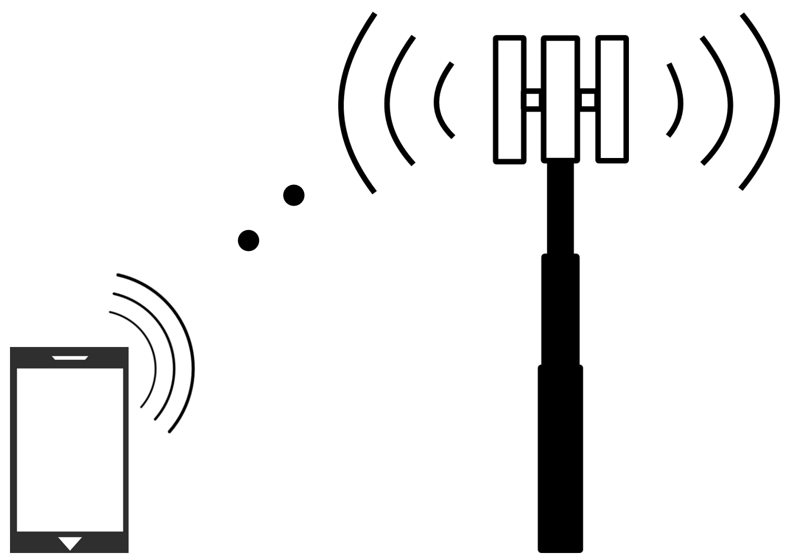


Arithmetic mean aggregation is *not robust* to corruptions $\Longrightarrow$ Poor predictions!

Our goal

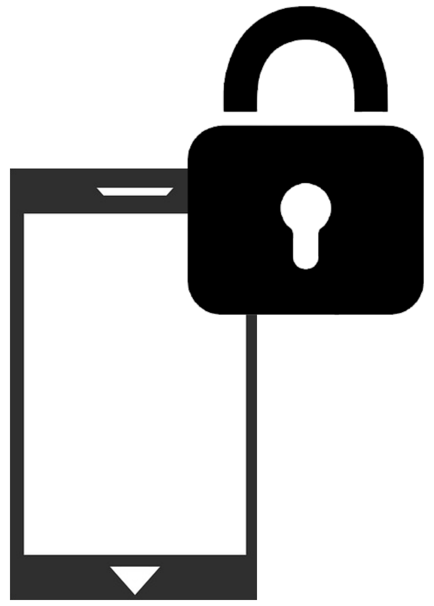
Design a robust aggregation algorithm for federated learning which is





Communication efficient

$O(1)$ times the communication cost as non-robust aggregation



Secure aggregation

Client-server communication via secure summation only

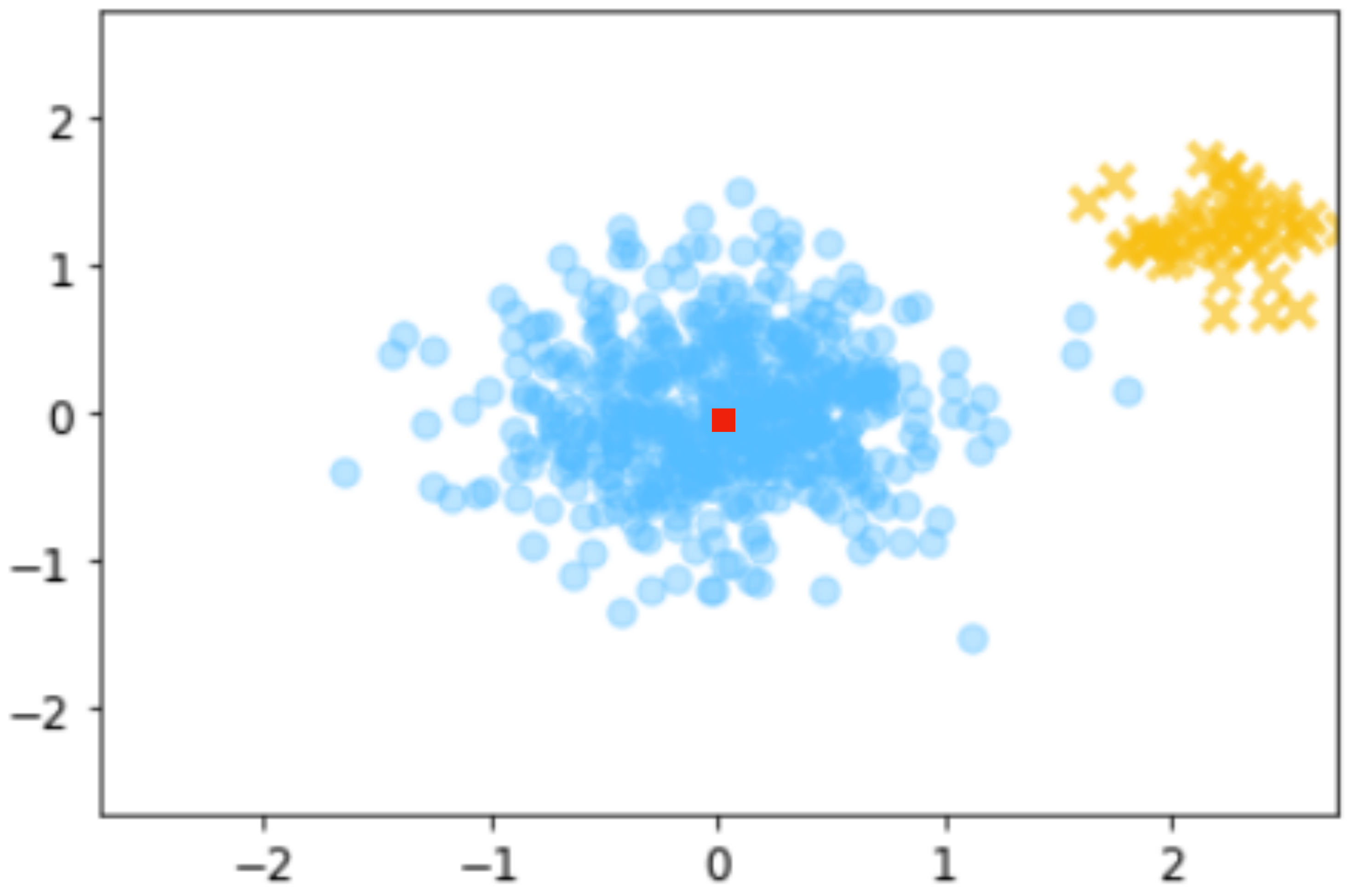
Note: not DP

Heterogeneity vs. Robustness

Consider mean estimation in Huber's contamination model:

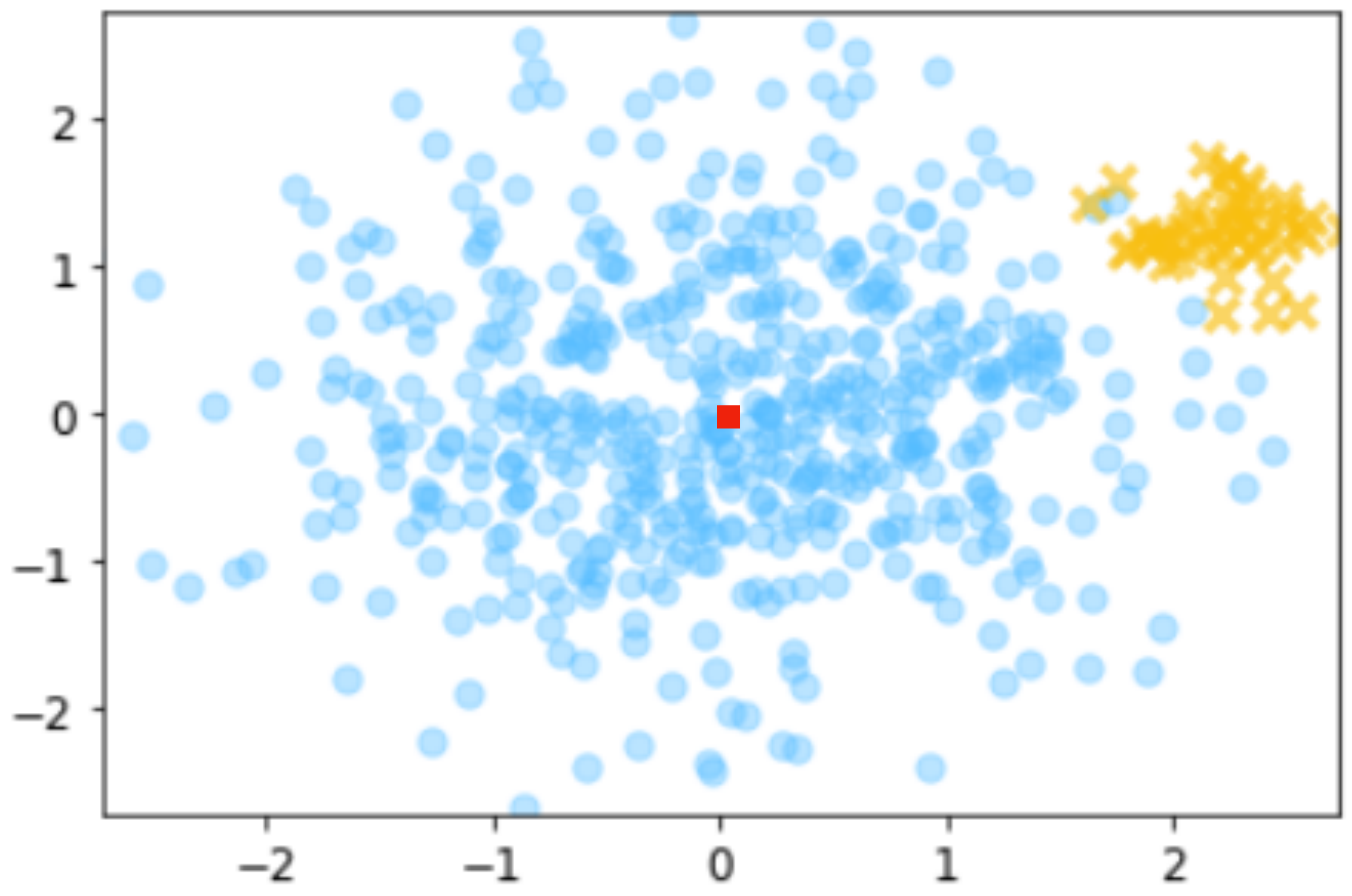
$$w_1, \cdots, w_n \sim (1-\rho)\,\mathcal{N}(\mu, \sigma^2 I) + \rho\, Q$$

inliers outliers



Any mean estimate $\overline{w}_n$ must satisfy

$$\|\overline{w}_n - \mu\|^2 \gtrsim \sigma^2 \left(\rho^2 + \frac{d}{n}\right)$$



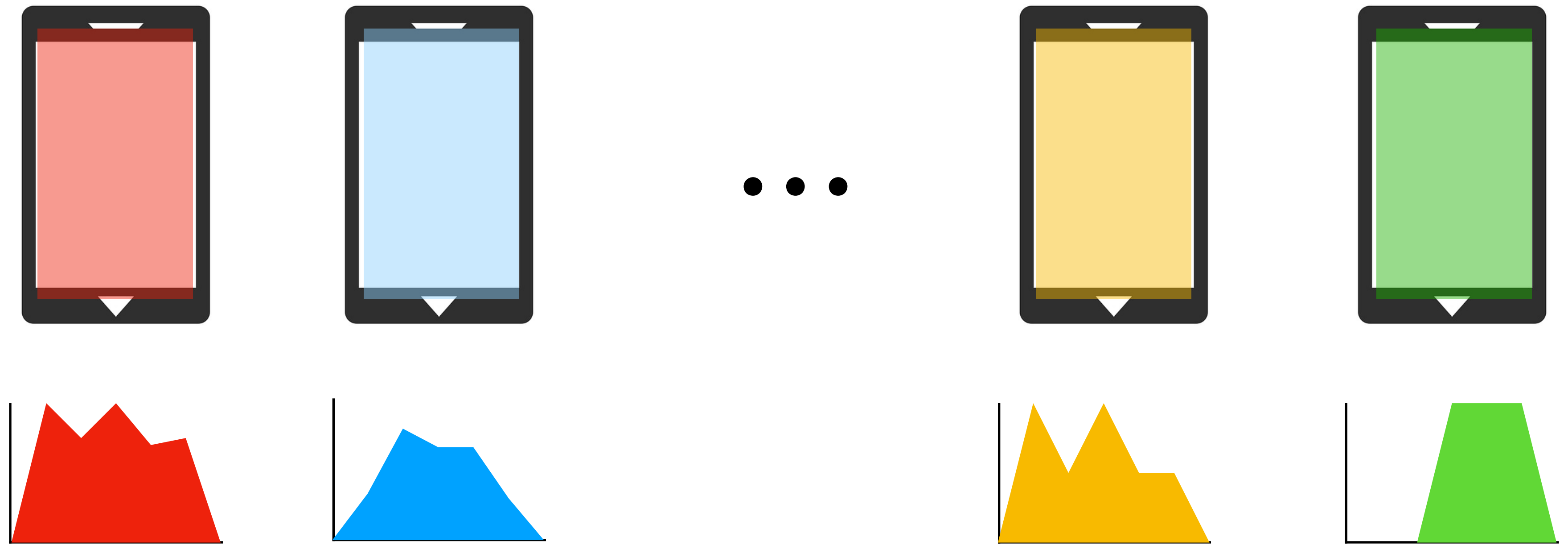
[Chen, Gao, Ren. Annals of Stat. (2018)]

For general federated learning, fix a set D of "inlier" distributions

$1 \quad 2 \quad \cdots \quad n$

Clients

Data Distribution



$p_1 \quad p_2 \quad \cdots \quad p_n$

$p_i \notin D$ *is considered an outlier even without adversarial corruption*

Learning Objective

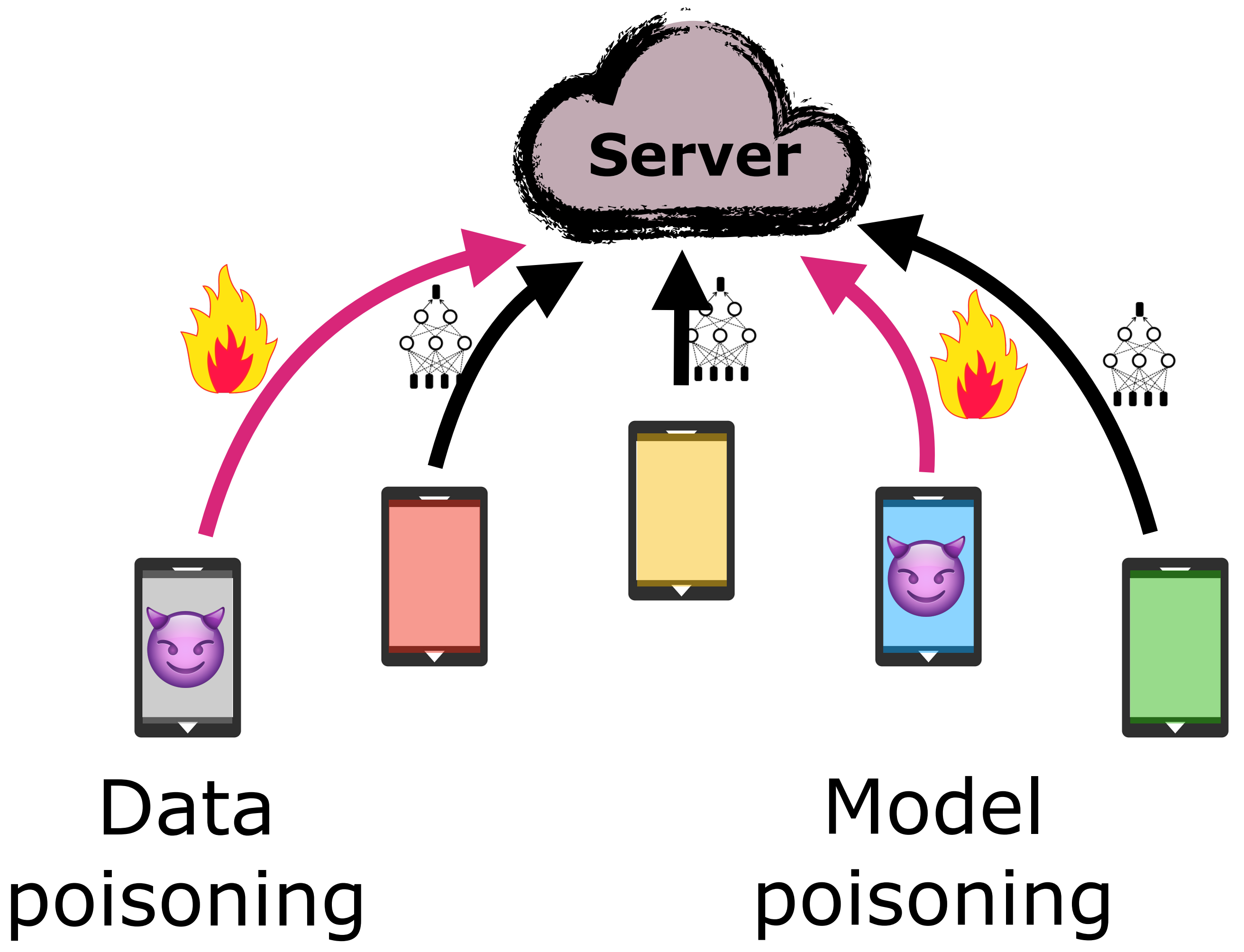
$$\min_{w \in \mathbb{R}^d} \sum_{i : p_i \in D} F_i(w) \quad \text{where} \quad F_i(w) = \mathbb{E}_{z \sim p_i}\left[f(w; z)\right]$$

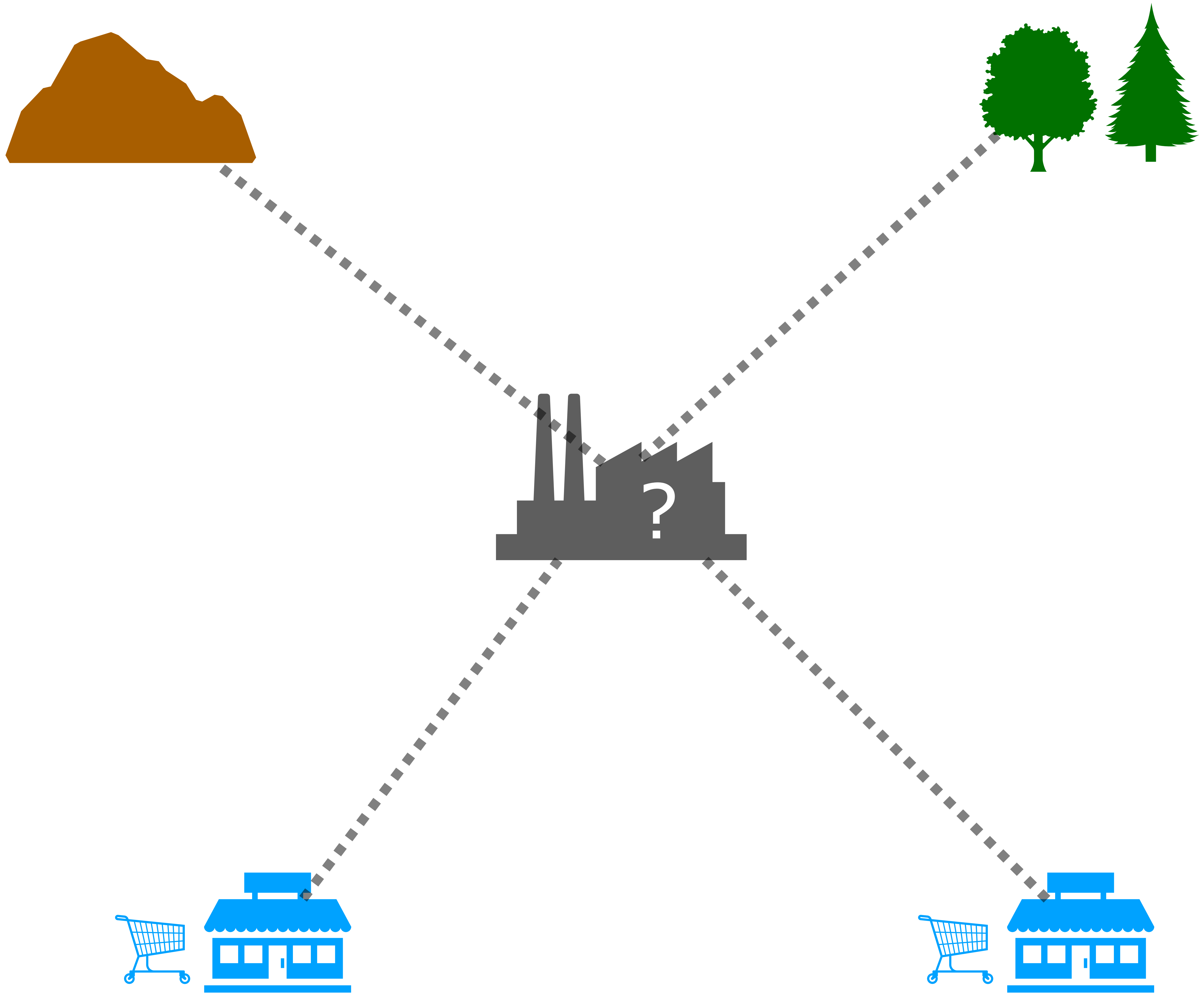
loss on client i

Convergence only possible up to size of inlier set D due to $\|\overline{w}_n - \mu\|^2 \gtrsim \sigma^2 \left(\rho^2 + \frac{d}{n}\right)$

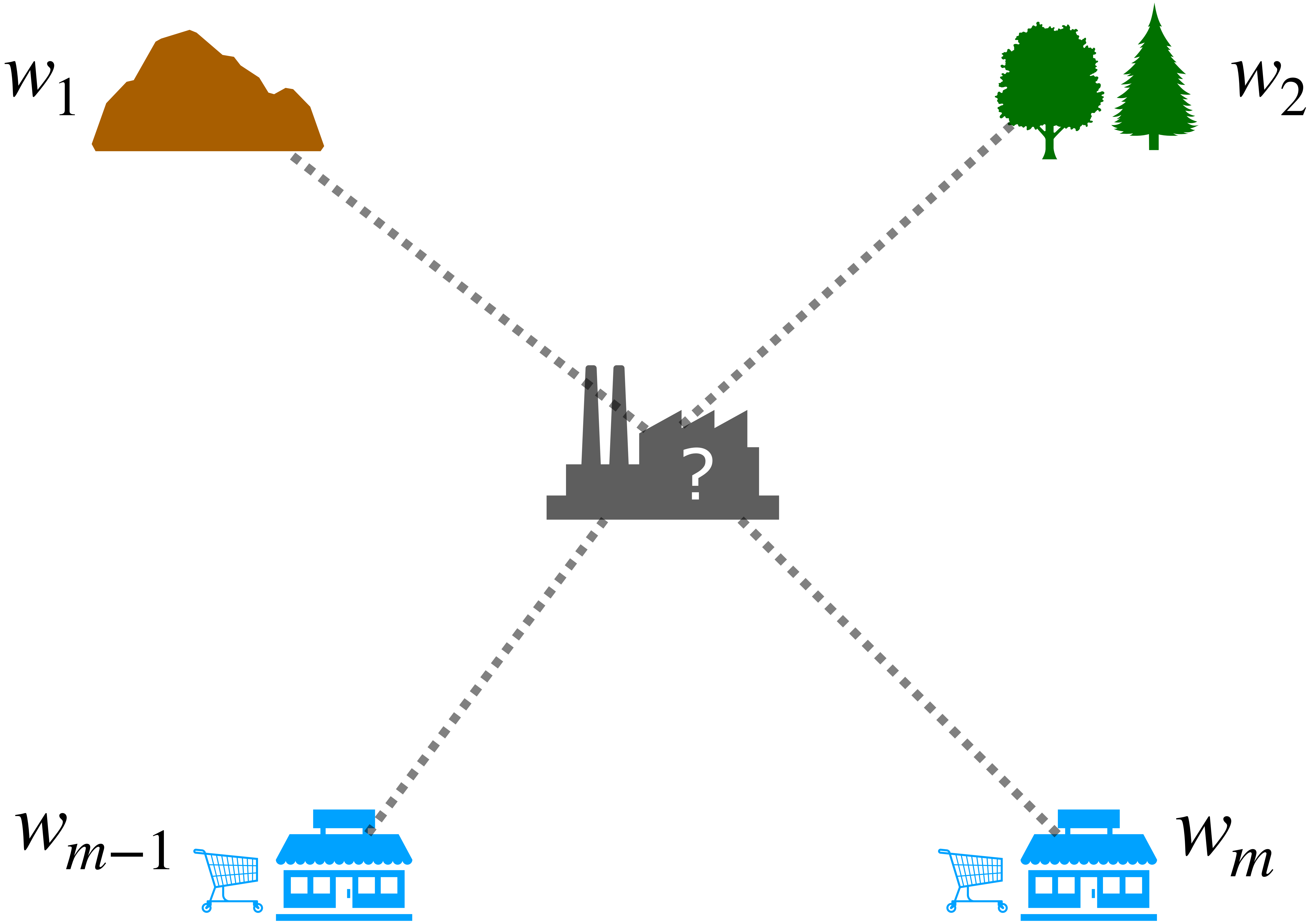
Algorithm is agnostic to D — only appears in analysis

Federated learning with robust aggregation





Fermat & Torricelli (~1600s), Weber (1909)

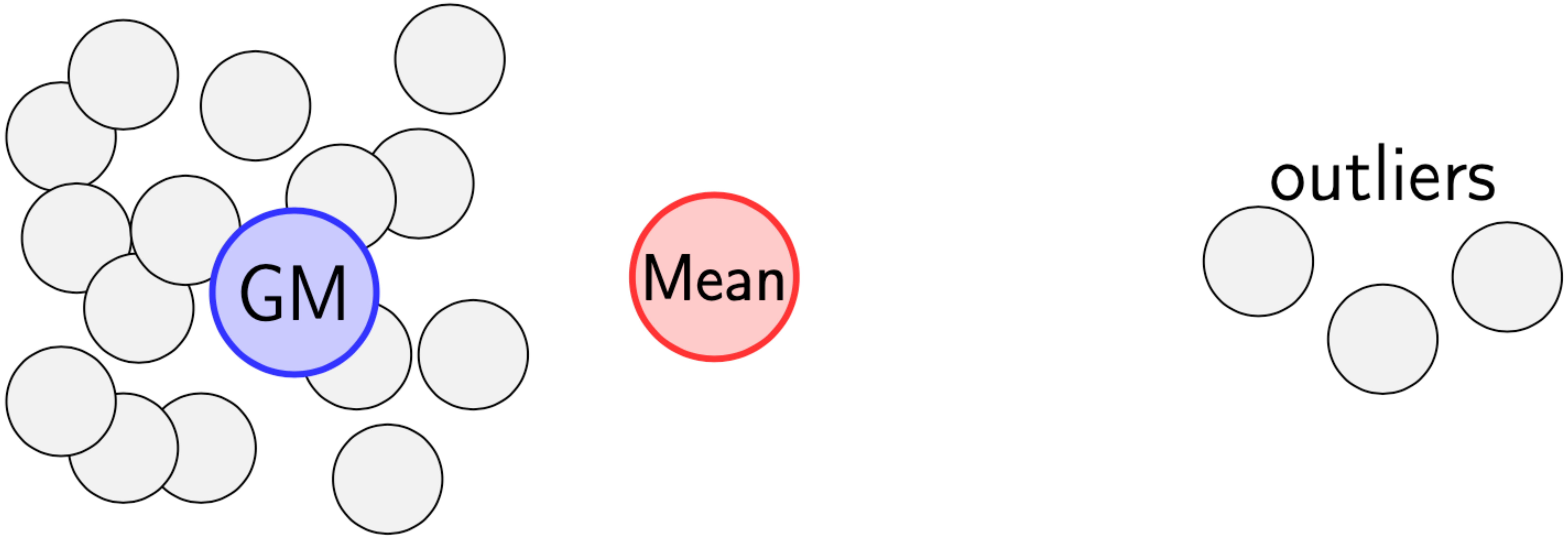


Fermat & Torricelli (~1600s), Weber (1909)

Geometric Median

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

Robustness: Breakdown point of GM = 1/2



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

Weiszfeld's Algorithm

Start with initial guess z_0 and iterate:

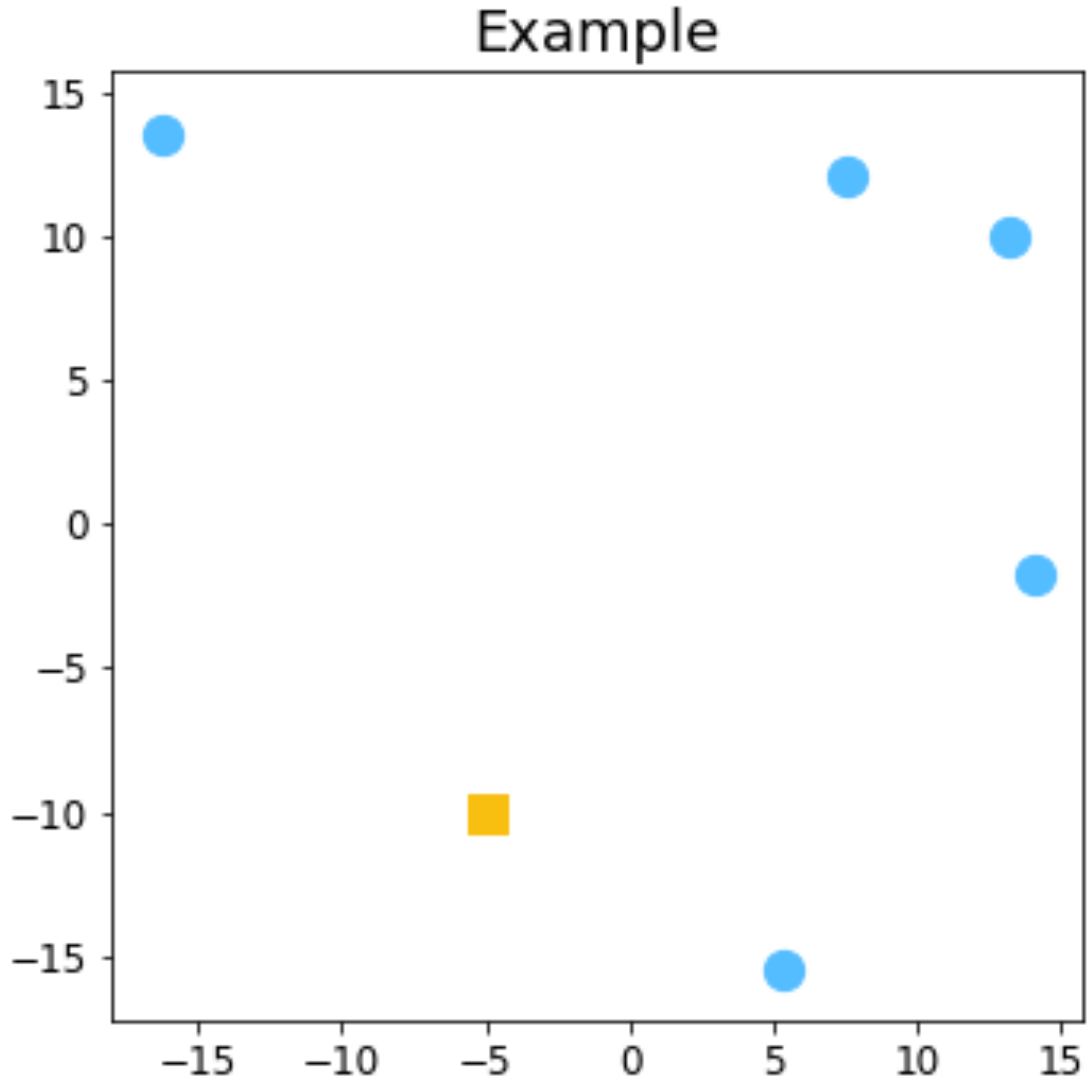
$$\beta_{i,t} = \frac{1}{\max\{\|z_t - w_i\|_2,\ \nu\}}$$

Compute new weights

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

Reweighted average

[Weiszfeld 1937]



Weiszfeld's Algorithm

Start with initial guess z_0 and iterate:

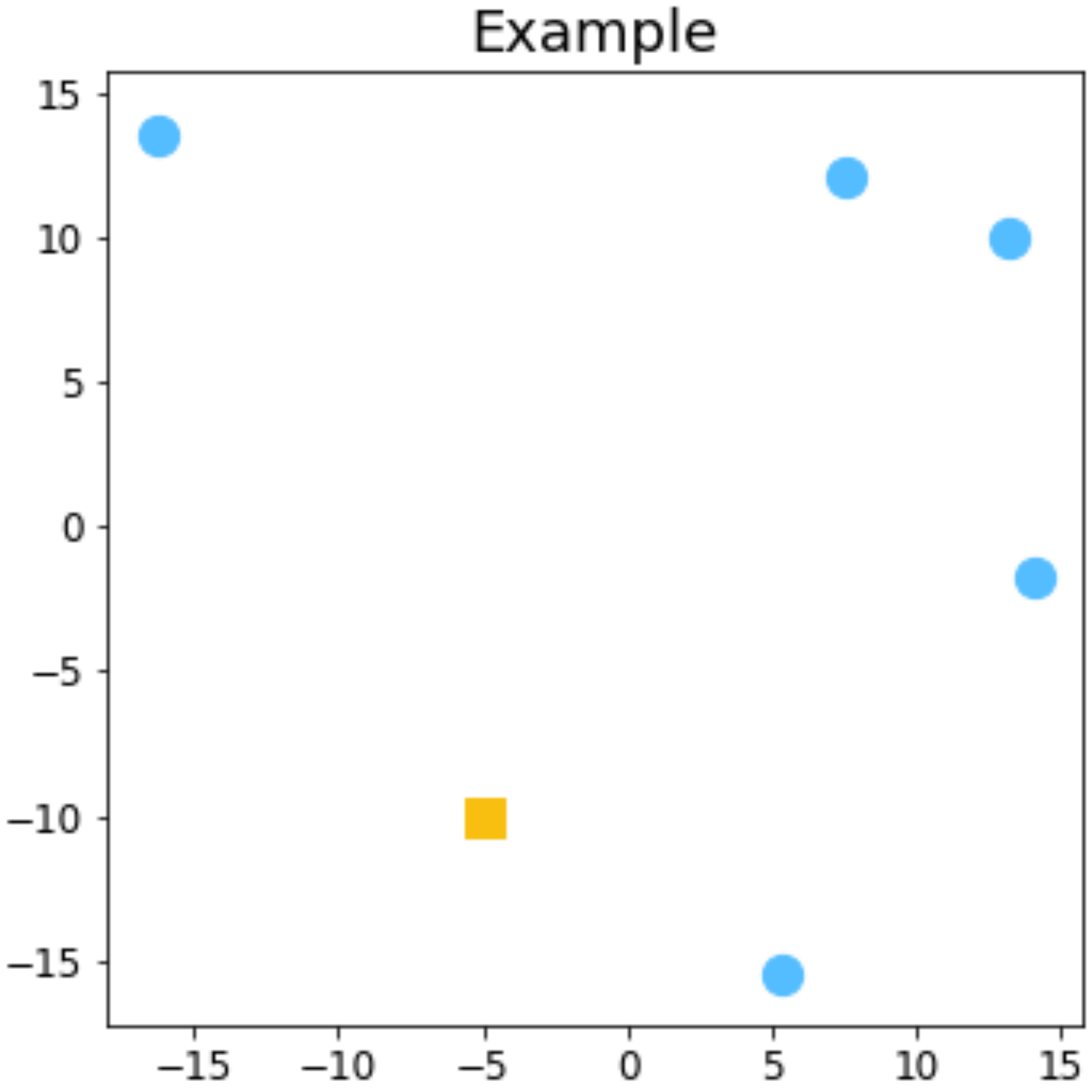
$$\beta_{i,t} = \frac{1}{\max\{\|z_t - w_i\|_2,\ \nu\}}$$

Compute new weights

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

Reweighted average

[Weiszfeld 1937]



Proposition [*P.*, Kakade, Harchaoui]

Assume that $\min_i \|z^\star - w_i\| \geq \nu$.

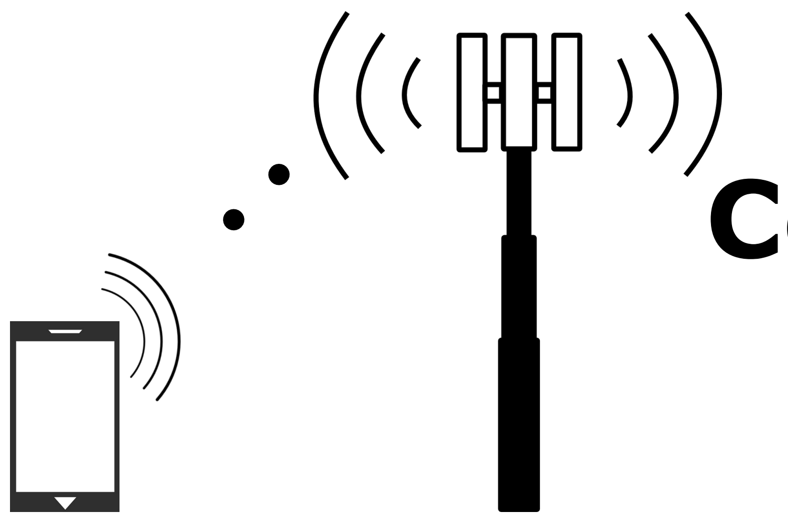
Then, we get an ε-approximate GM in

$O\left(\frac{1}{\nu\varepsilon}\right)$ iterations

$z^\star = \mathrm{GM}(w_1, \cdots, w_m)$

ν : smoothing in β-update

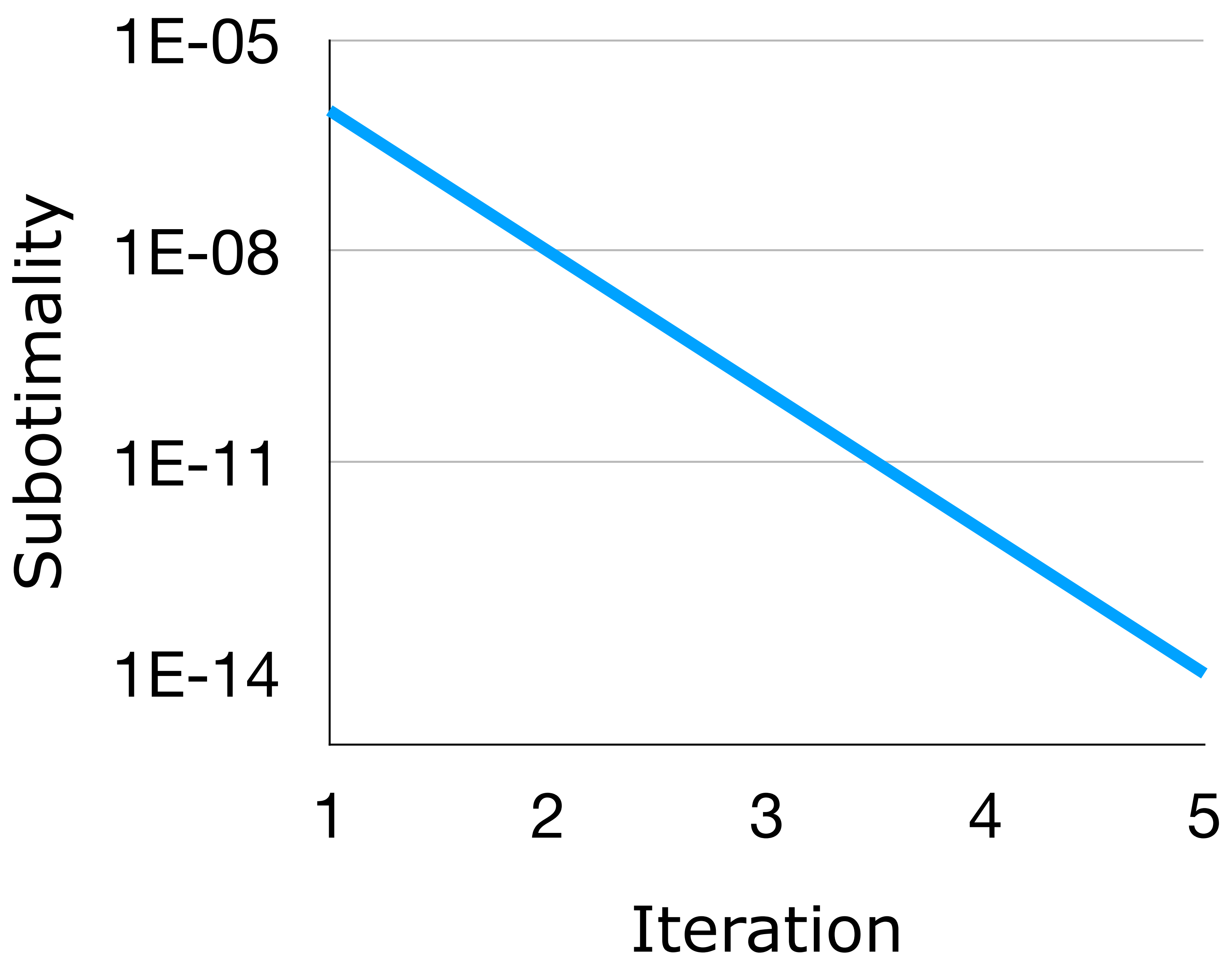
$$\beta_{i,t} = \frac{1}{\max\{\|z_t - w_i\|_2, \nu\}}$$



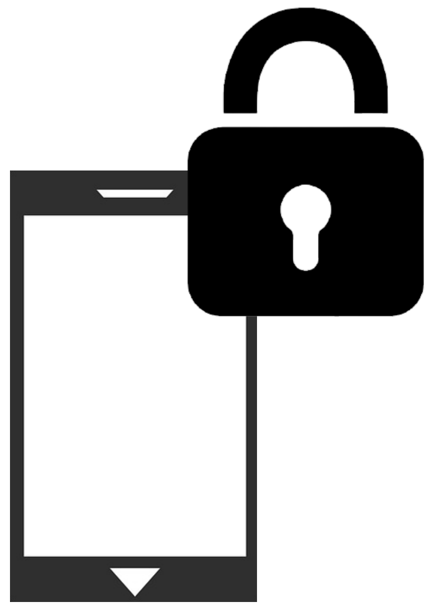
Communication efficient!

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

Even **1** iteration gives robustness



RFA = FedAvg + GM aggregation

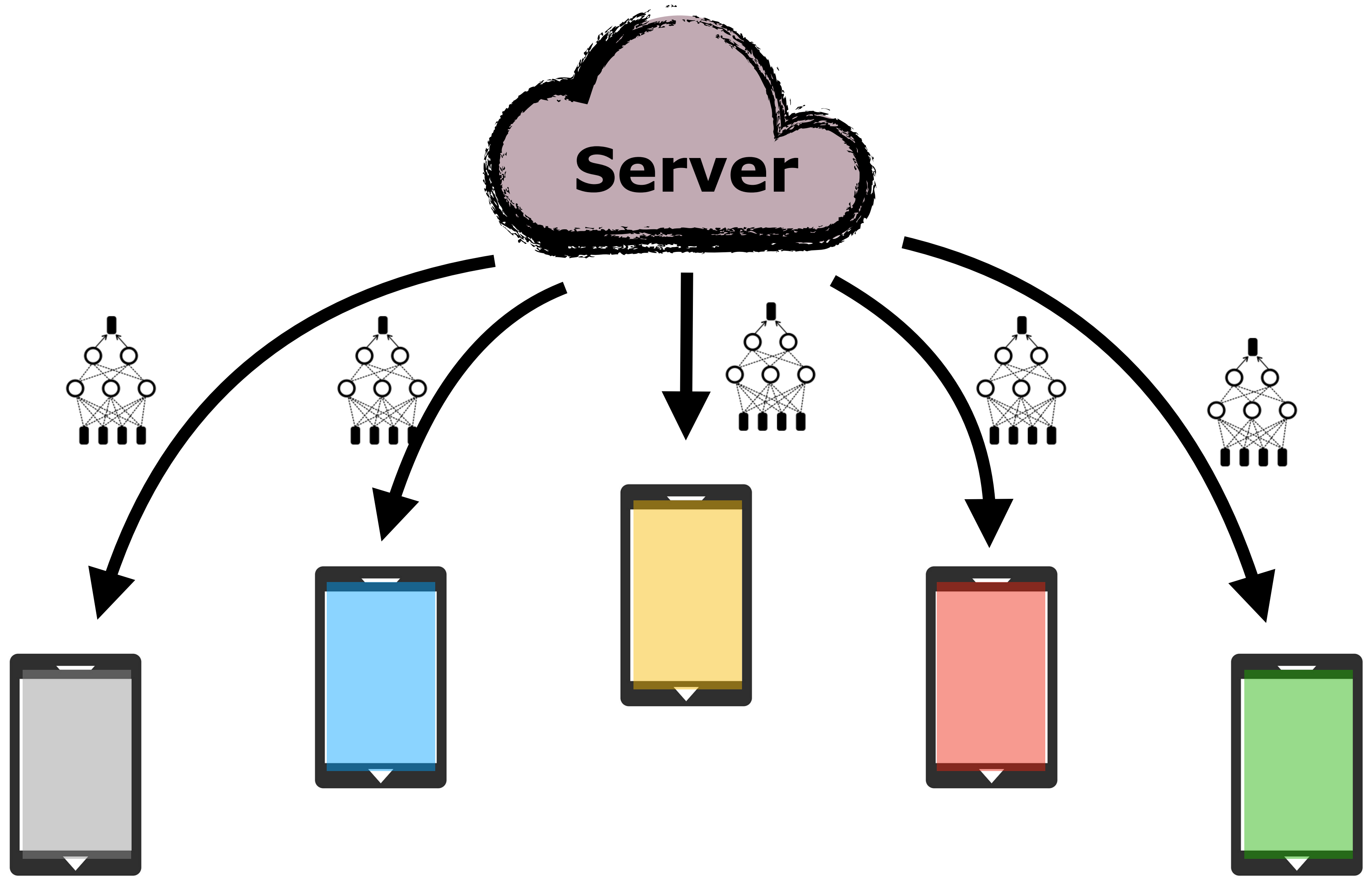


Secure aggregation

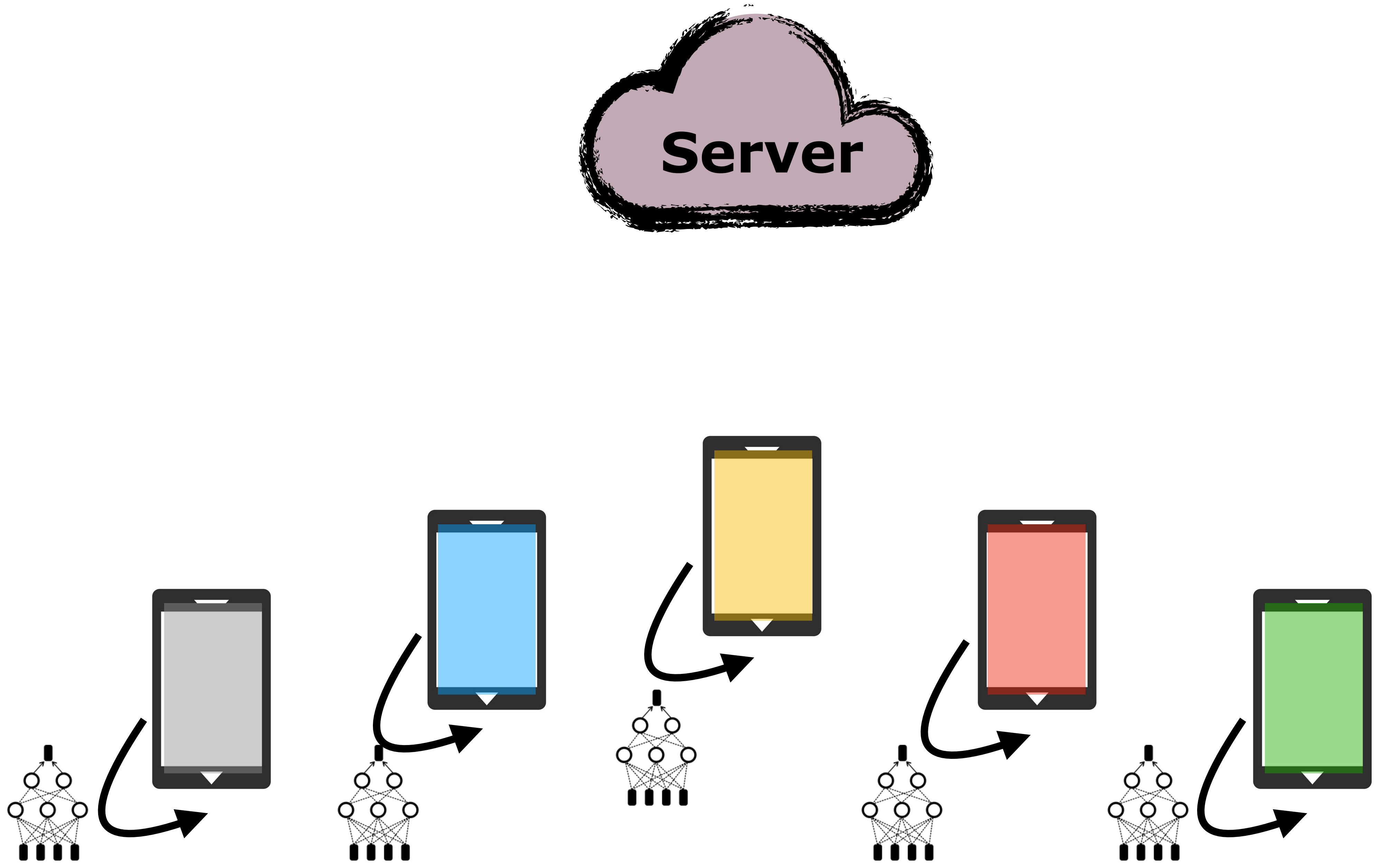
Only client-server communication is via **secure summation** in

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

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

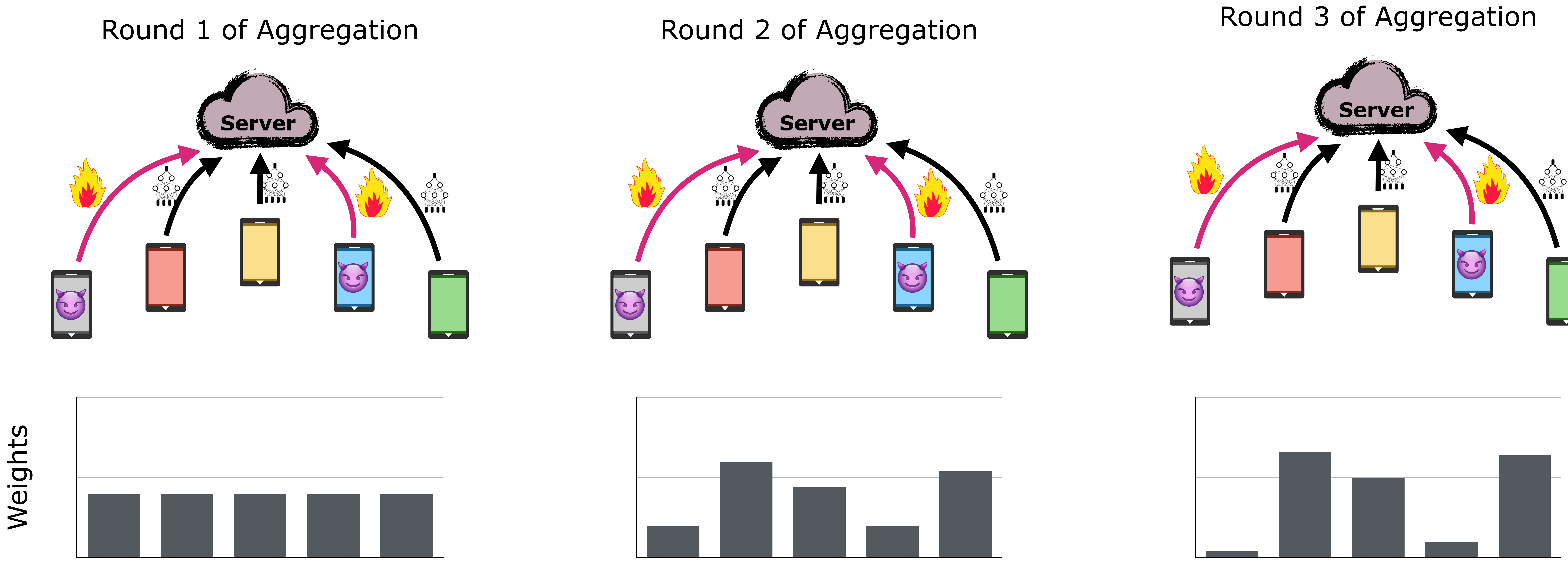


Step 2 of 3: Clients perform some local SGD steps on their local data

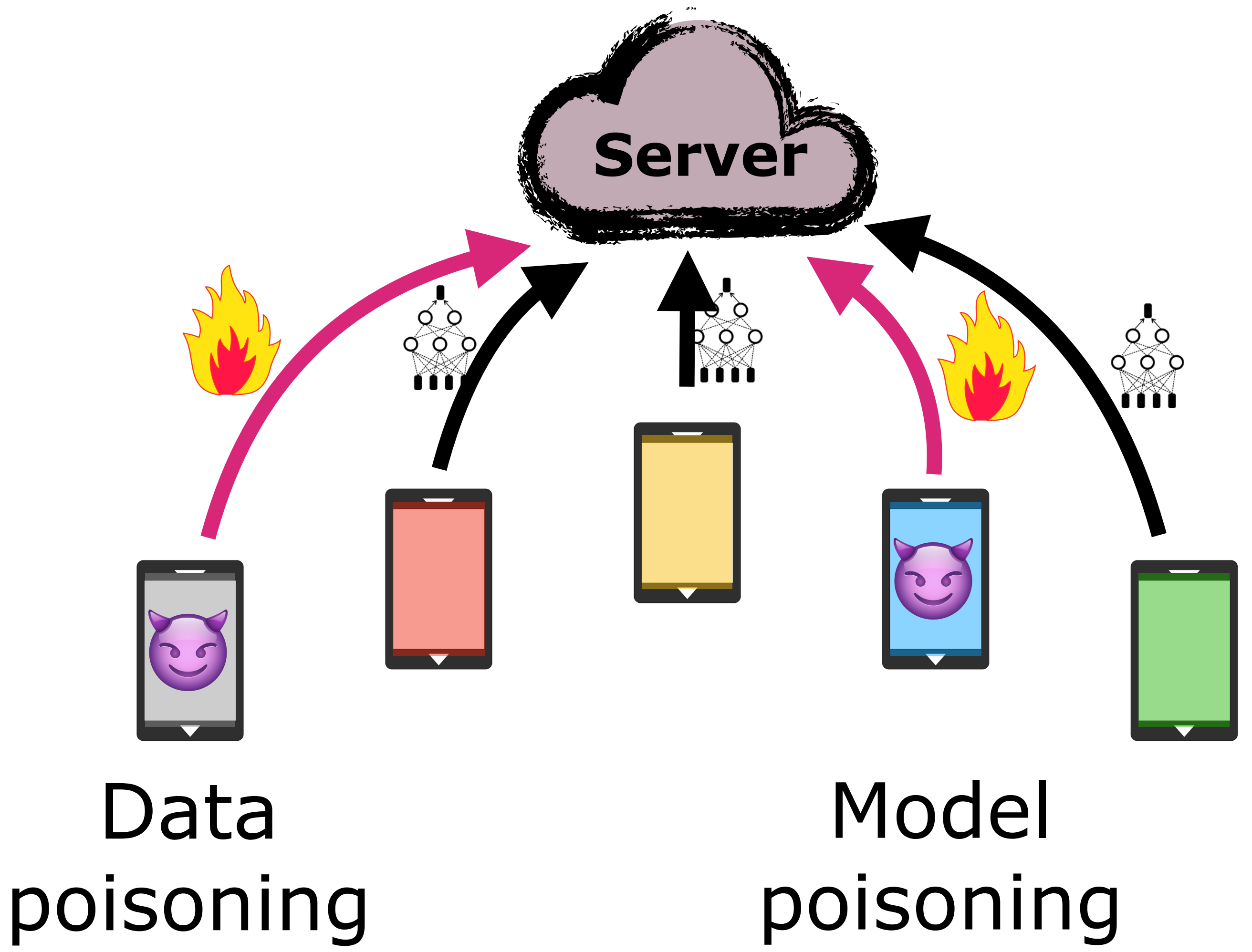


So far, same as FedAvg

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



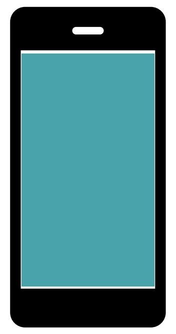
Convergence analysis (least squares)

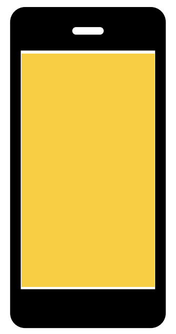


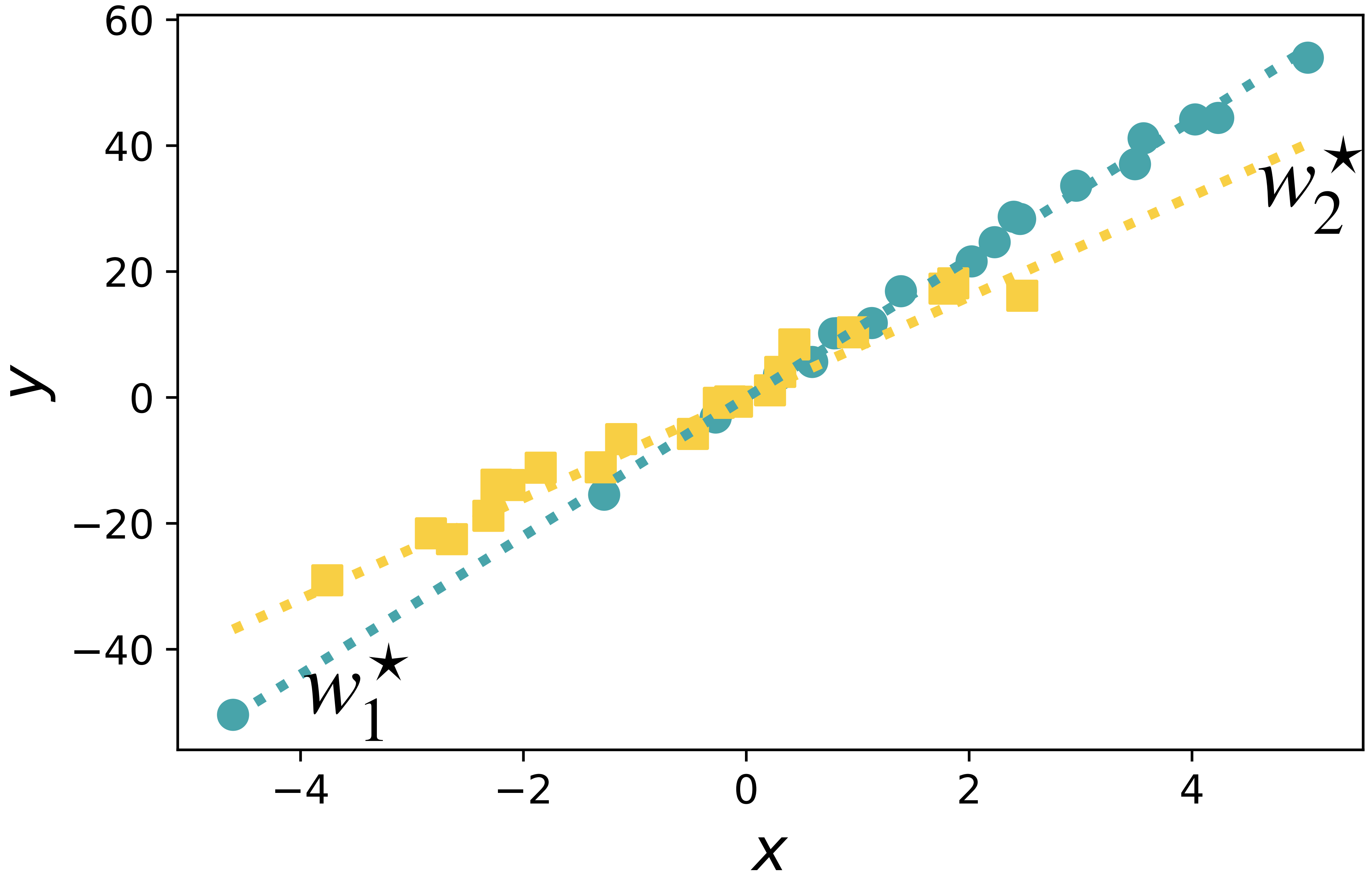
Data on client i**:** $(X_i, Y_i) \sim p_i$ satisfies

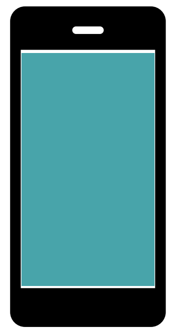
$$Y_i = X_i^\top w_i^\star + \xi_i \quad \text{where} \quad \xi_i \sim \mathcal{N}(0, \sigma^2)$$

$\mathbb{R}$ $\quad\quad$ $\mathbb{R}^d$

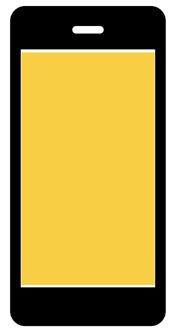

$X_1 \sim \mathcal{N}(0, H_1)$


$X_2 \sim \mathcal{N}(0, H_2)$





$X_1 \sim \mathcal{N}(0, H_1)$

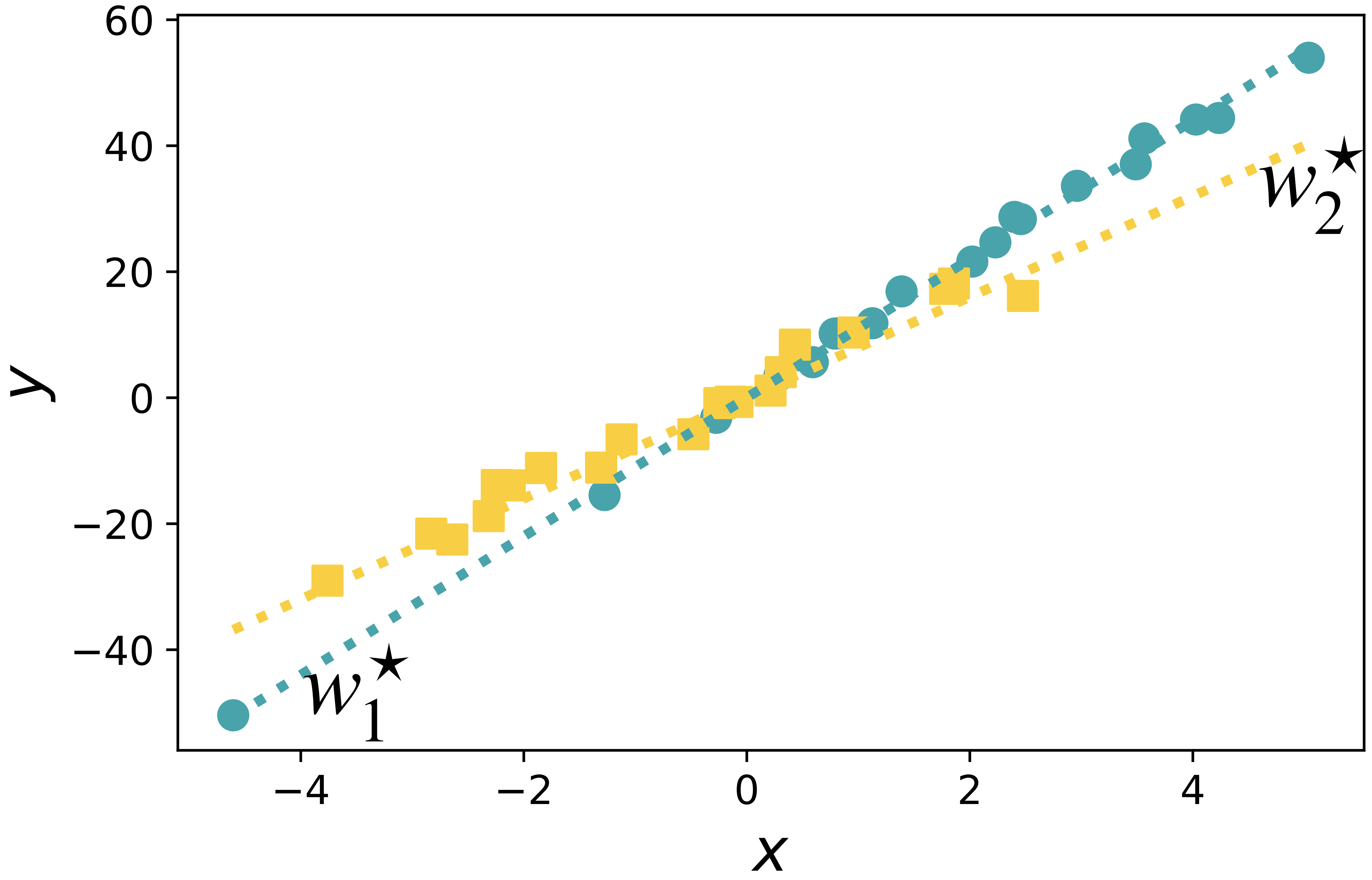


$X_2 \sim \mathcal{N}(0, H_2)$

Data on client i: $(X_i, Y_i) \sim p_i$ satisfies

$$Y_i = X_i^\top w_i^\star + \xi_i \quad \text{where} \quad \xi_i \sim \mathcal{N}(0, \sigma^2)$$

$\mathbb{R}$ $\qquad$ $\mathbb{R}^d$



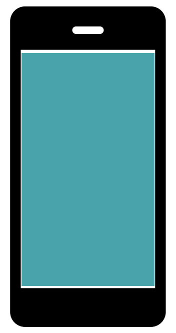
Measure of heterogeneity

$$\Omega_X = \max_{\substack{i \\ \text{inlier}}} \ \lambda_{\max}\left(H^{-1/2} H_i H^{-1/2}\right) \geq 1$$

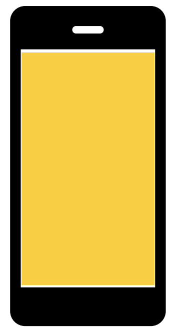
$H_i = \mathbb{E}[X_i X_i^\top]$ marginal covariance of i

$H = \frac{1}{n} \sum_i H_i$ marginal covariance of mixture

Data on client i: $(X_i, Y_i) \sim p_i$ satisfies



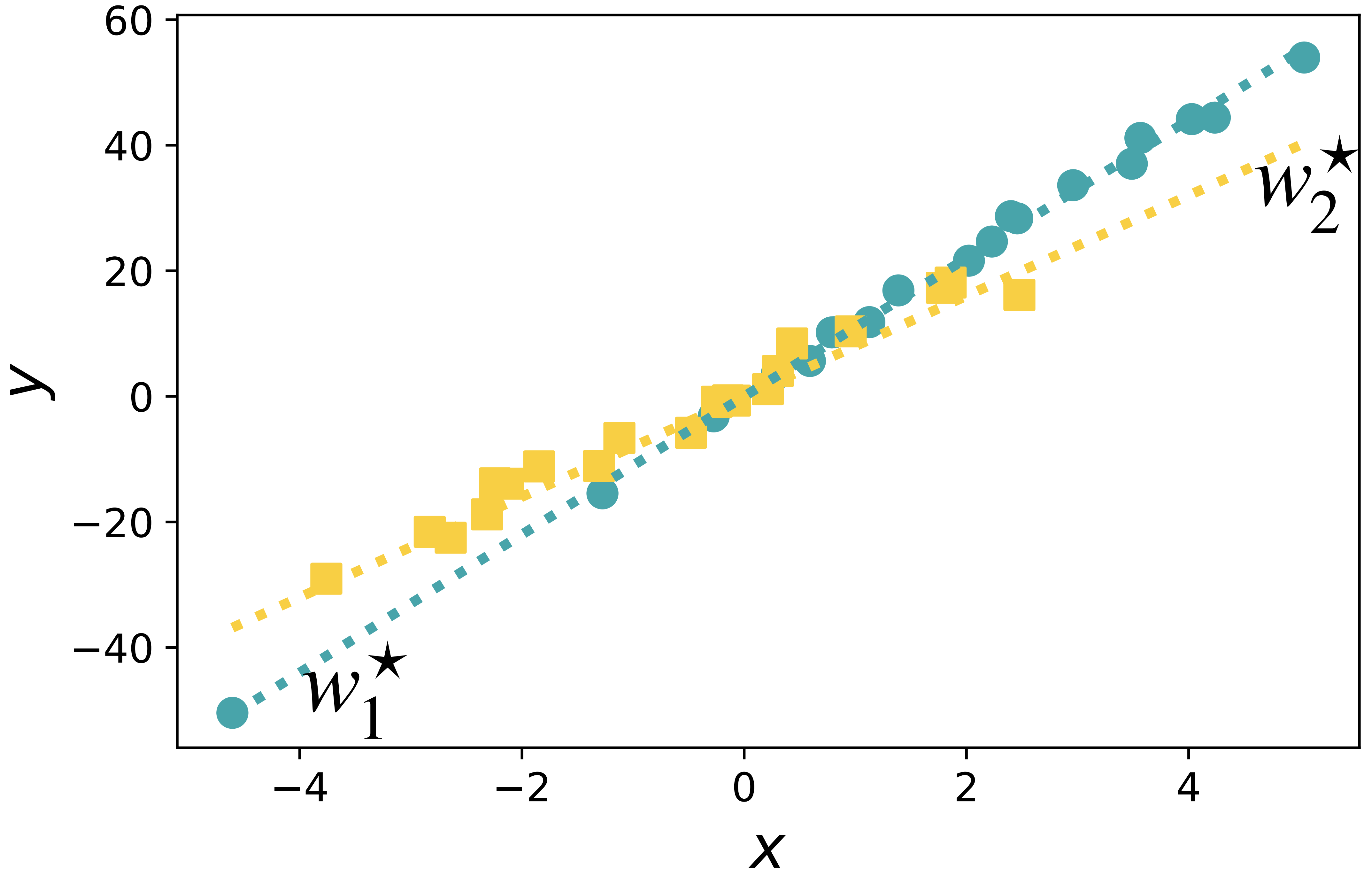
$X_1 \sim \mathcal{N}(0, H_1)$



$X_2 \sim \mathcal{N}(0, H_2)$

$$Y_i = X_i^\top w_i^\star + \xi_i \quad \text{where} \quad \xi_i \sim \mathcal{N}(0, \sigma^2)$$

$\mathbb{R}$ $\qquad$ $\mathbb{R}^d$



Measure of heterogeneity

$$\Omega_Y = \max_{\substack{i,j \\ \text{inlier}}} \|w_i^\star - w_j^\star\|_2 \geq 0$$

Fraction of non-corrupted clients = $\dfrac{1}{2} + c$

Number of clients per round = m

$\mathscr{E}$ holds w.h.p. if $m \gtrsim \dfrac{1}{c^2}$

Theorem [*P.*, Kakade, Harchaoui]

Assume that $F(w)$ is strongly convex, $\|X_i\| \leq 1$ and number of local steps $\propto 2^t$. Let $\mathscr{E}$ denote the event that at least $1/2 + c/2$ non-corrupted devices are chosen in each round.

Then, RFA with ε-approximate GM satisfies

$$\mathbb{E}\left[\|w_t - w^\star\|^2 \,|\, \mathscr{E}\right] \lesssim \underbrace{\frac{\|w_0 - w^\star\|^2}{2^t}}_{\text{Optimization error}} + \frac{1}{c^2}\left(\underbrace{d\sigma^2 \frac{t}{2^t}}_{\text{Statistical error}} + \underbrace{\frac{\varepsilon^2}{m^2}}_{\text{GM Approx. Error}} + \underbrace{\Omega_X^2 \Omega_Y^2}_{\text{Heterogeneity Error}}\right)$$

Fraction of non-corrupted clients = $\dfrac{1}{2} + c$

Number of clients per round = m

$\mathscr{E}$ holds w.h.p. if $m \gtrsim \dfrac{1}{c^2}$

Theorem [*P.*, Kakade, Harchaoui]

Assume that $F(w)$ is strongly convex, $\|X_i\| \le 1$ and number of local steps $\propto 2^t$. Let $\mathscr{E}$ denote the event that at least $1/2 + c/2$ non-corrupted devices are chosen in each round.

Then, RFA with ε-approximate GM satisfies

$$\mathbb{E}\left[\|w_t - w^\star\|^2 | \, \mathscr{E}\right] \lesssim \underbrace{\frac{\|w_0 - w^\star\|^2}{2^t}}_{\text{Optimization error}} + \frac{1}{c^2}\left(\underbrace{d\sigma^2 \frac{t}{2^t}}_{\text{Statistical error}} + \underbrace{\frac{\varepsilon^2}{m^2}}_{\text{GM Approx. Error}} + \underbrace{\Omega_X^2 \Omega_Y^2}_{\text{Heterogeneity Error}} \right)$$

Fraction of non-corrupted clients = $\frac{1}{2} + c$

Number of clients per round = m

$\mathscr{E}$ holds w.h.p. if $m \gtrsim \frac{1}{c^2}$

Theorem [*P.*, Kakade, Harchaoui]

Assume that $F(w)$ is strongly convex, $\|X_i\| \le 1$ and number of local steps $\propto 2^t$. Let $\mathscr{E}$ denote the event that **at least** $1/2 + c/2$ **non-corrupted devices** are chosen in each round.

Then, RFA with ε-approximate GM satisfies

$$\mathbb{E}\left[\|w_t - w^\star\|^2 | \mathscr{E}\right] \lesssim \frac{\|w_0 - w^\star\|^2}{2^t} + \frac{1}{c^2}\left(d\sigma^2 \frac{t}{2^t} + \frac{\varepsilon^2}{m^2} + \Omega_X^2 \Omega_Y^2\right)$$

Optimization error

Statistical error

Heterogeneity Error

GM Approx. Error

Fraction of non-corrupted clients = $\dfrac{1}{2} + c$

Number of clients per round = m

$\mathscr{E}$ holds w.h.p. if $m \gtrsim \dfrac{1}{c^2}$

Theorem [*P.*, Kakade, Harchaoui]

Assume that $F(w)$ is strongly convex, $\|X_i\| \le 1$ and number of local steps $\propto 2^t$. Let $\mathscr{E}$ denote the event that at least $1/2 + c/2$ non-corrupted devices are chosen in each round.

Then, RFA with ε-approximate GM satisfies

$$\mathbb{E}\left[\|w_t - w^\star\|^2 | \mathscr{E}\right] \lesssim \underbrace{\frac{\|w_0 - w^\star\|^2}{2^t}}_{\text{Optimization error}} + \frac{1}{c^2}\left(\underbrace{d\sigma^2\frac{t}{2^t}}_{\text{Statistical error}} + \underbrace{\frac{\varepsilon^2}{m^2}}_{\text{GM Approx. Error}} + \underbrace{\Omega_X^2\Omega_Y^2}_{\text{Heterogeneity Error}}\right)$$

Fraction of non-corrupted clients = $\dfrac{1}{2} + c$

Number of clients per round = m

$\mathscr{E}$ holds w.h.p. if $m \gtrsim \dfrac{1}{c^2}$

Theorem [*P.*, Kakade, Harchaoui]

Assume that $F(w)$ is strongly convex, $\|X_i\| \le 1$ and number of local steps $\propto 2^t$. Let $\mathscr{E}$ denote the event that at least $1/2 + c/2$ non-corrupted devices are chosen in each round.

Then, RFA with ε-approximate GM satisfies

$$\mathbb{E}\left[\|w_t - w^\star\|^2 | \mathscr{E}\right] \lesssim \frac{\|w_0 - w^\star\|^2}{2^t} + \frac{1}{c^2}\left(d\sigma^2 \frac{t}{2^t} + \frac{\varepsilon^2}{m^2} + \Omega_X^2 \Omega_Y^2\right)$$

Optimization error

Statistical error

Heterogeneity Error

GM Approx. Error

Fraction of non-corrupted clients = $\dfrac{1}{2} + c$

Number of clients per round = m

$\mathscr{E}$ holds w.h.p. if $m \gtrsim \dfrac{1}{c^2}$

Theorem [*P.*, Kakade, Harchaoui]

Assume that $F(w)$ is strongly convex, $\|X_i\| \le 1$ and number of local steps $\propto 2^t$. Let $\mathscr{E}$ denote the event that at least $1/2 + c/2$ non-corrupted devices are chosen in each round.

Then, RFA with ε-approximate GM satisfies

$$\mathbb{E}\left[\|w_t - w^\star\|^2 \,|\, \mathscr{E}\right] \lesssim \frac{\|w_0 - w^\star\|^2}{2^t} + \frac{1}{c^2}\left(d\sigma^2 \frac{t}{2^t} + \boxed{\frac{\varepsilon^2}{m^2}} + \Omega_X^2 \Omega_Y^2\right)$$

Optimization error

Statistical error

Heterogeneity Error

GM Approx. Error

Fraction of non-corrupted clients = $\dfrac{1}{2} + c$

Number of clients per round = m

$\mathscr{E}$ holds w.h.p. if $m \gtrsim \dfrac{1}{c^2}$

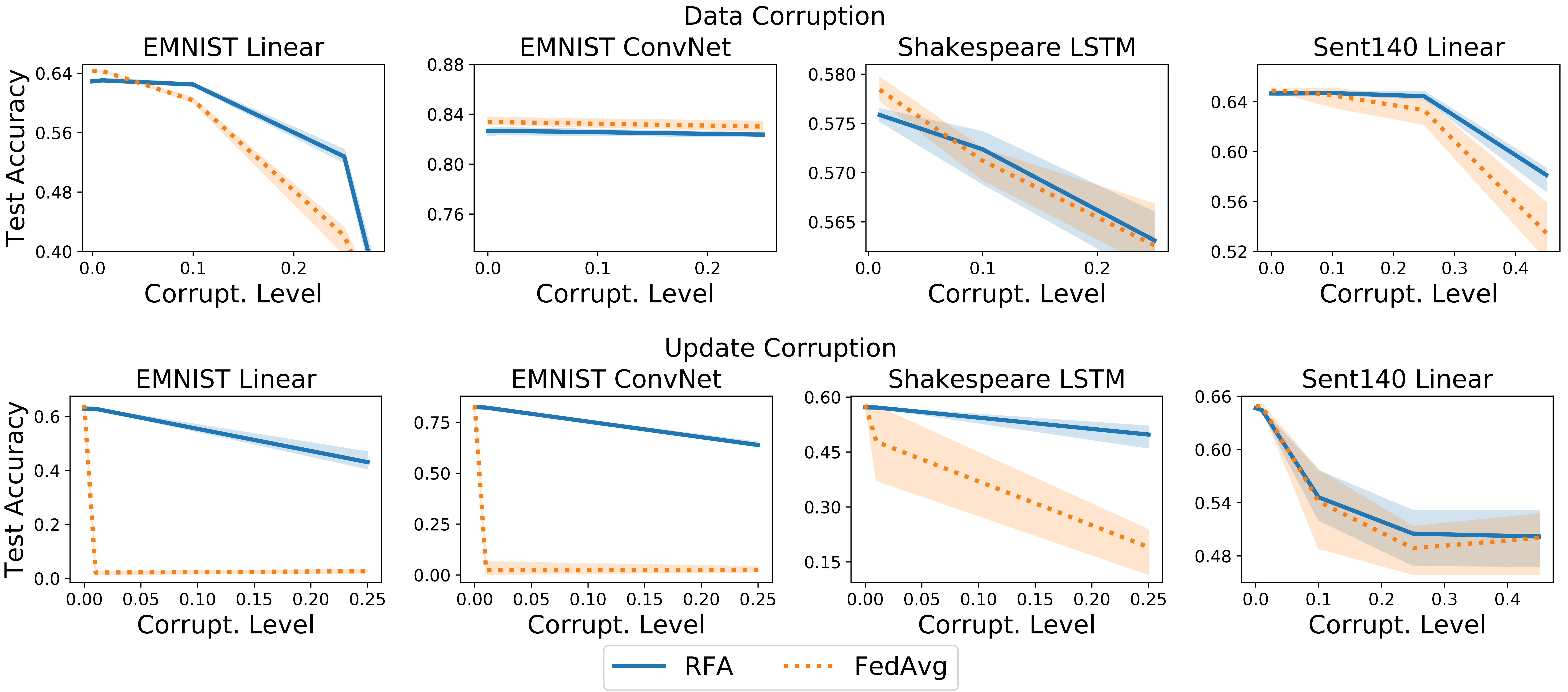
Theorem [*P.*, Kakade, Harchaoui]

Assume that $F(w)$ is strongly convex, $\|X_i\| \le 1$ and number of local steps $\propto 2^t$. Let $\mathscr{E}$ denote the event that at least $1/2 + c/2$ non-corrupted devices are chosen in each round.

Then, RFA with ε-approximate GM satisfies

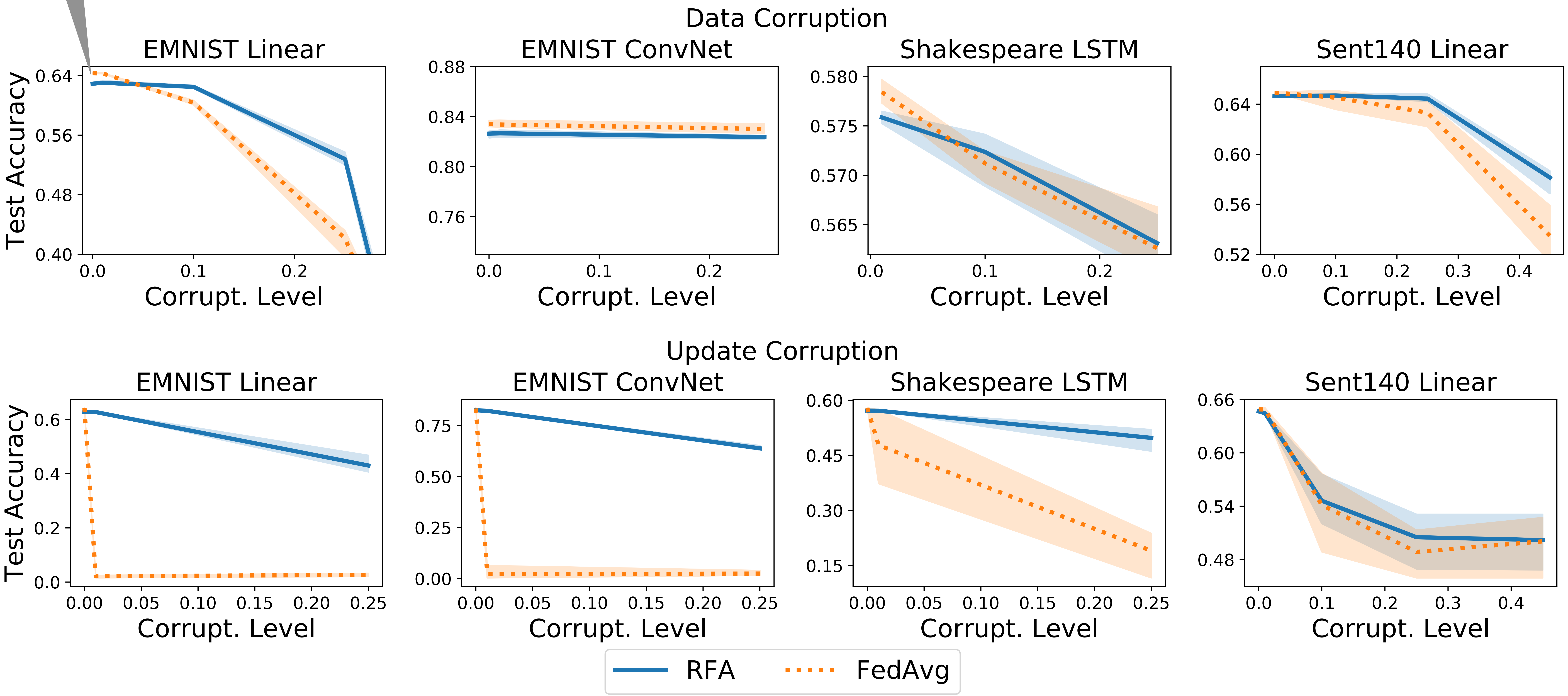
$$\mathbb{E}\left[\|w_t - w^\star\|^2 \,|\, \mathscr{E}\right] \lesssim \underbrace{\frac{\|w_0 - w^\star\|^2}{2^t}}_{\text{Optimization error}} + \frac{1}{c^2}\left(\underbrace{d\sigma^2\frac{t}{2^t}}_{\text{Statistical error}} + \underbrace{\frac{\varepsilon^2}{m^2}}_{\text{GM Approx. Error}} + \underbrace{\Omega_X^2\Omega_Y^2}_{\textbf{Heterogeneity Error}}\right)$$

Experiments



Experiments

1.4pp
gap at zero
corruption

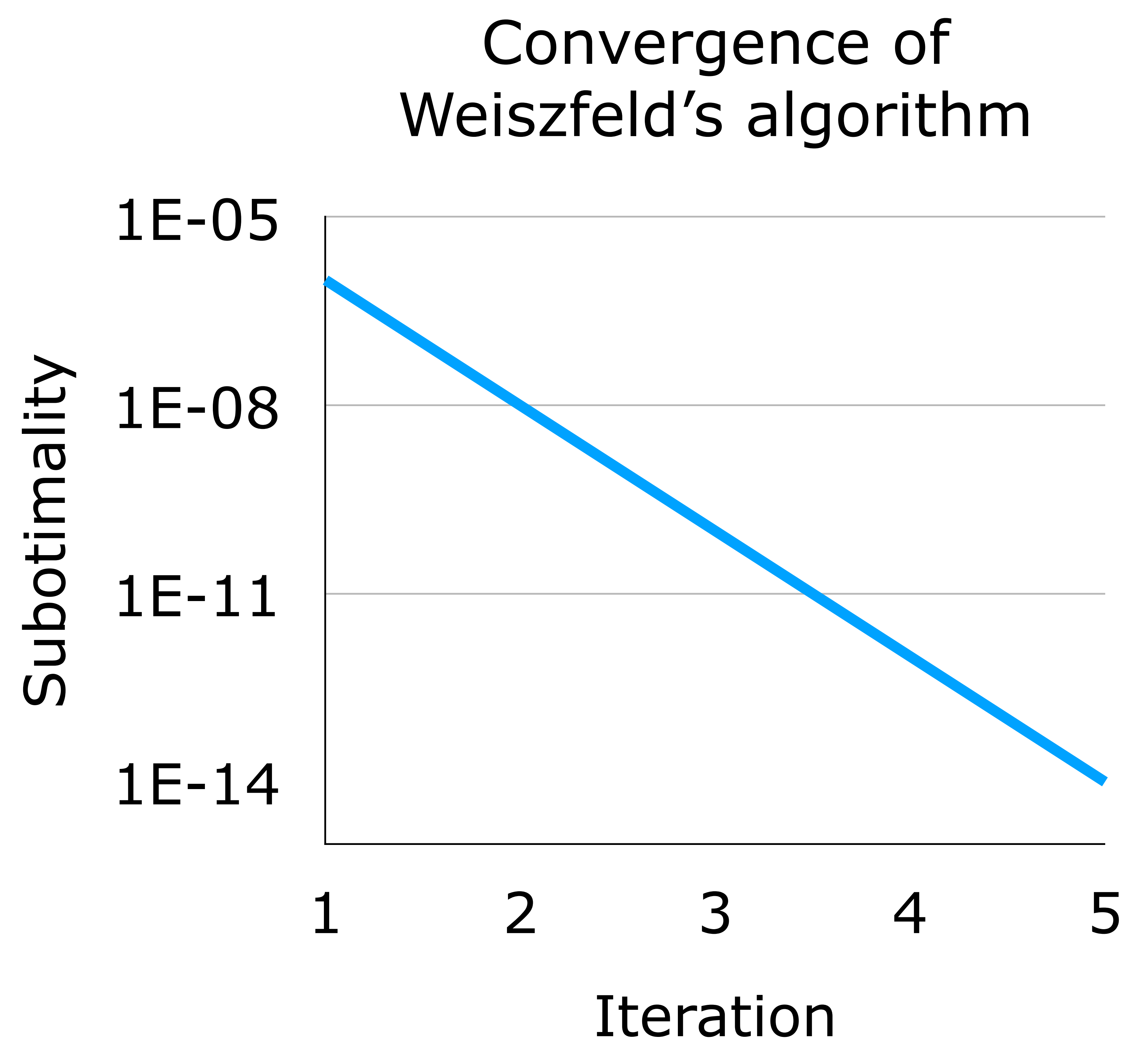


Reducing RFA’s communication

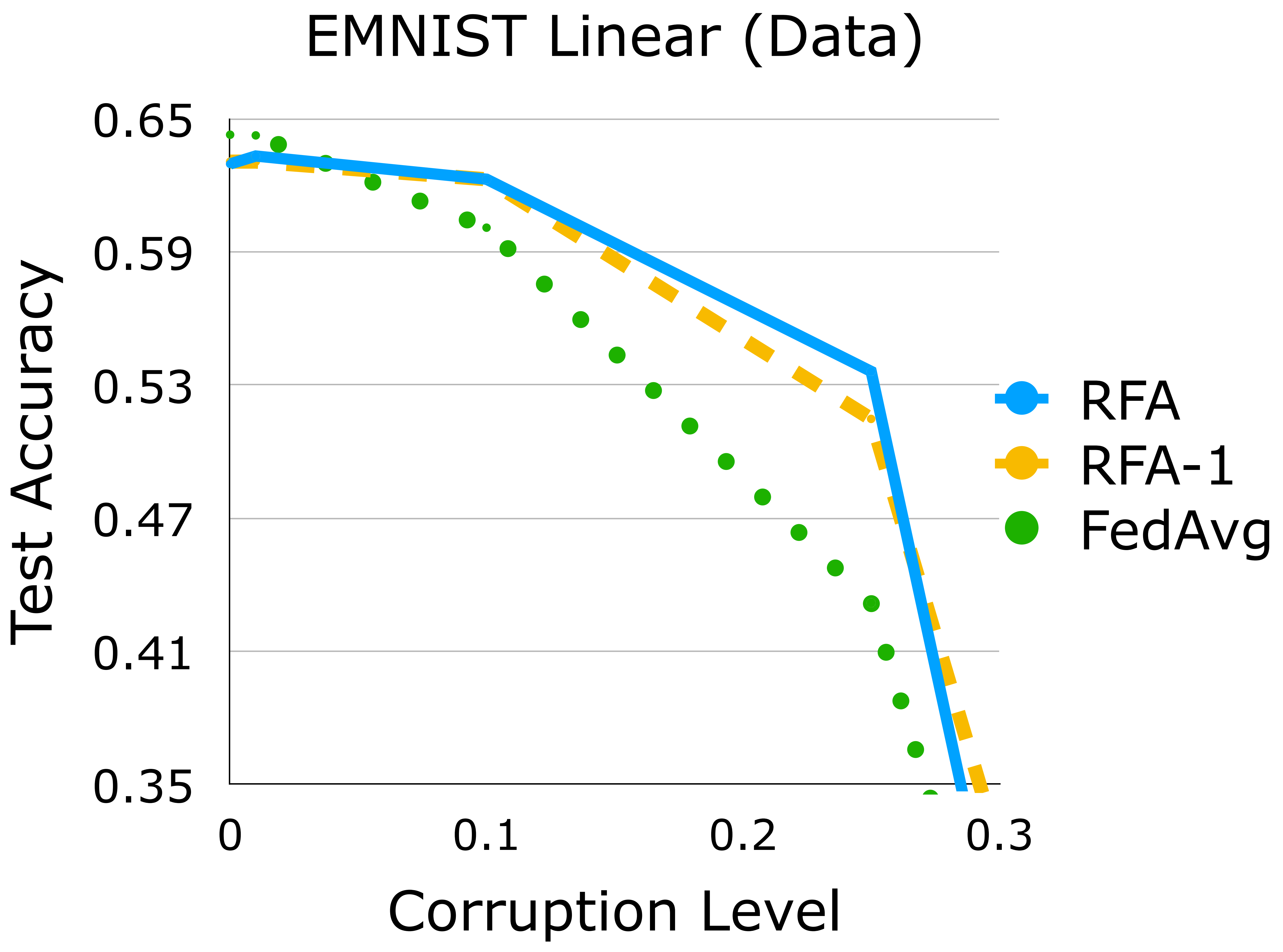
One round of RFA $\Longrightarrow$ 3-5 rounds of communication (Weiszfeld)

Does 1 round of communication give robustness?

$$\beta_i = \frac{1}{\max\{\|w_i\|_2,\, \nu\}} \qquad z = \frac{\sum_i \beta_i w_i}{\sum_i \beta_{i,t}}$$



Does 1 round of communication give robustness? **Yes**!



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,)
```

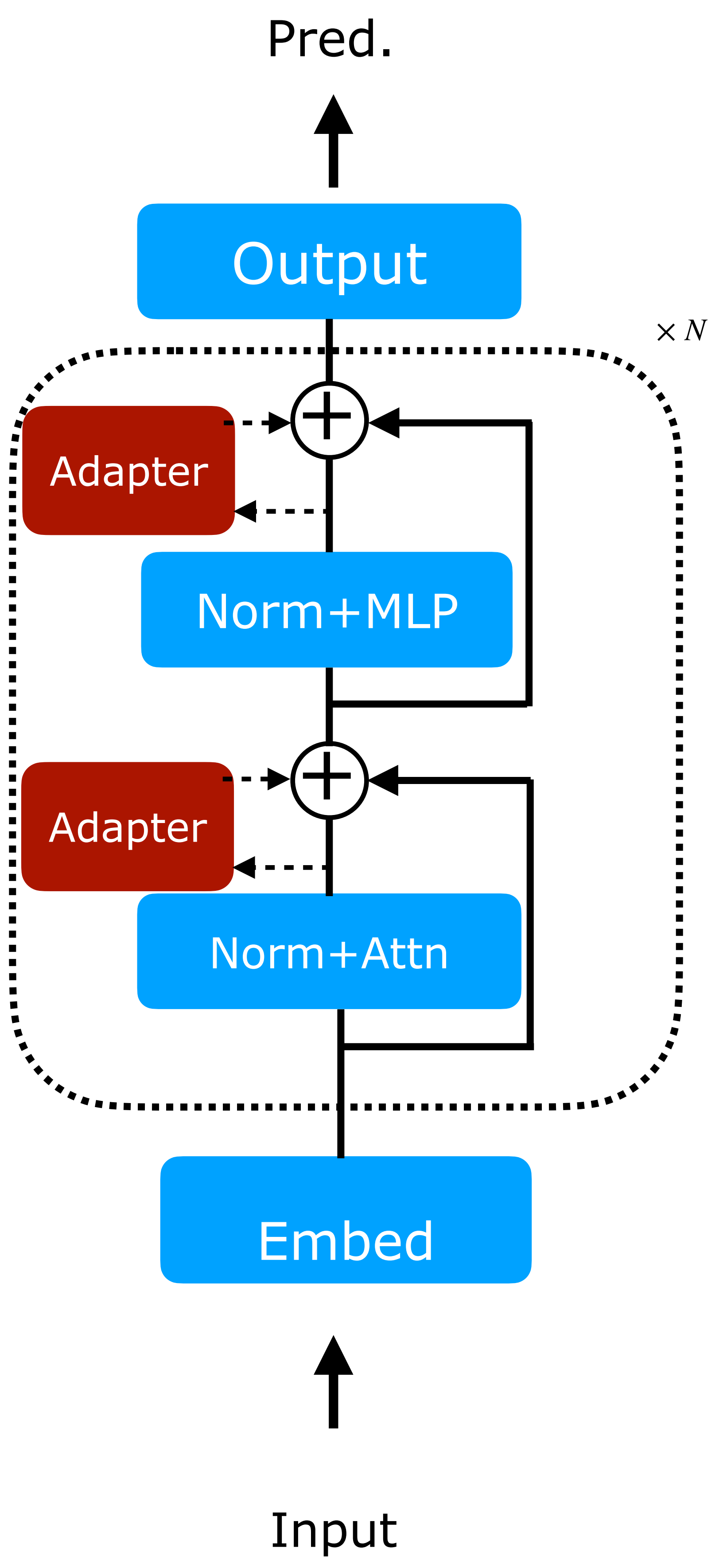
Install: **pip install geom-median**

Documentation: github.com/krishnap25/geom-median

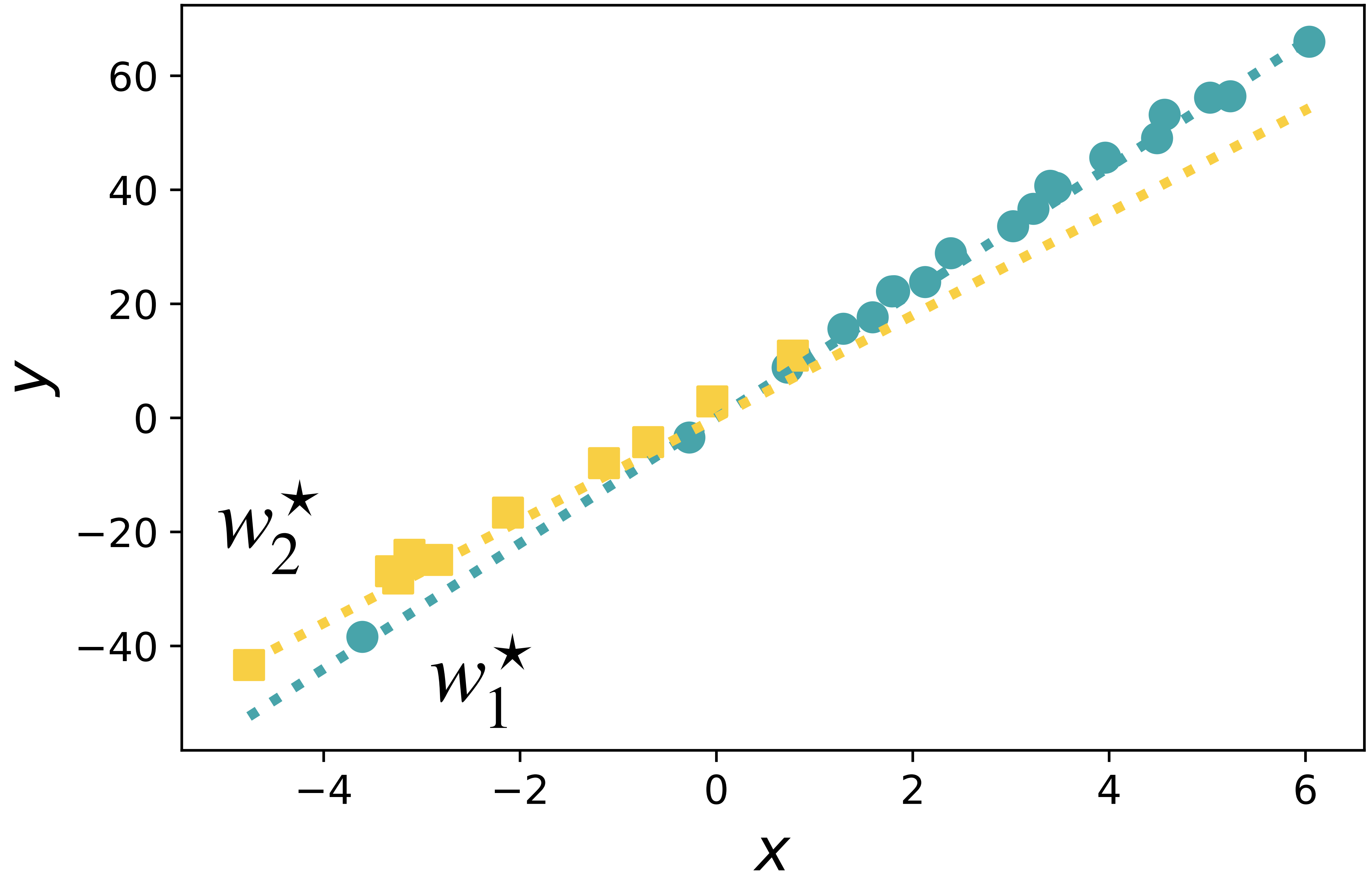
SCAN ME

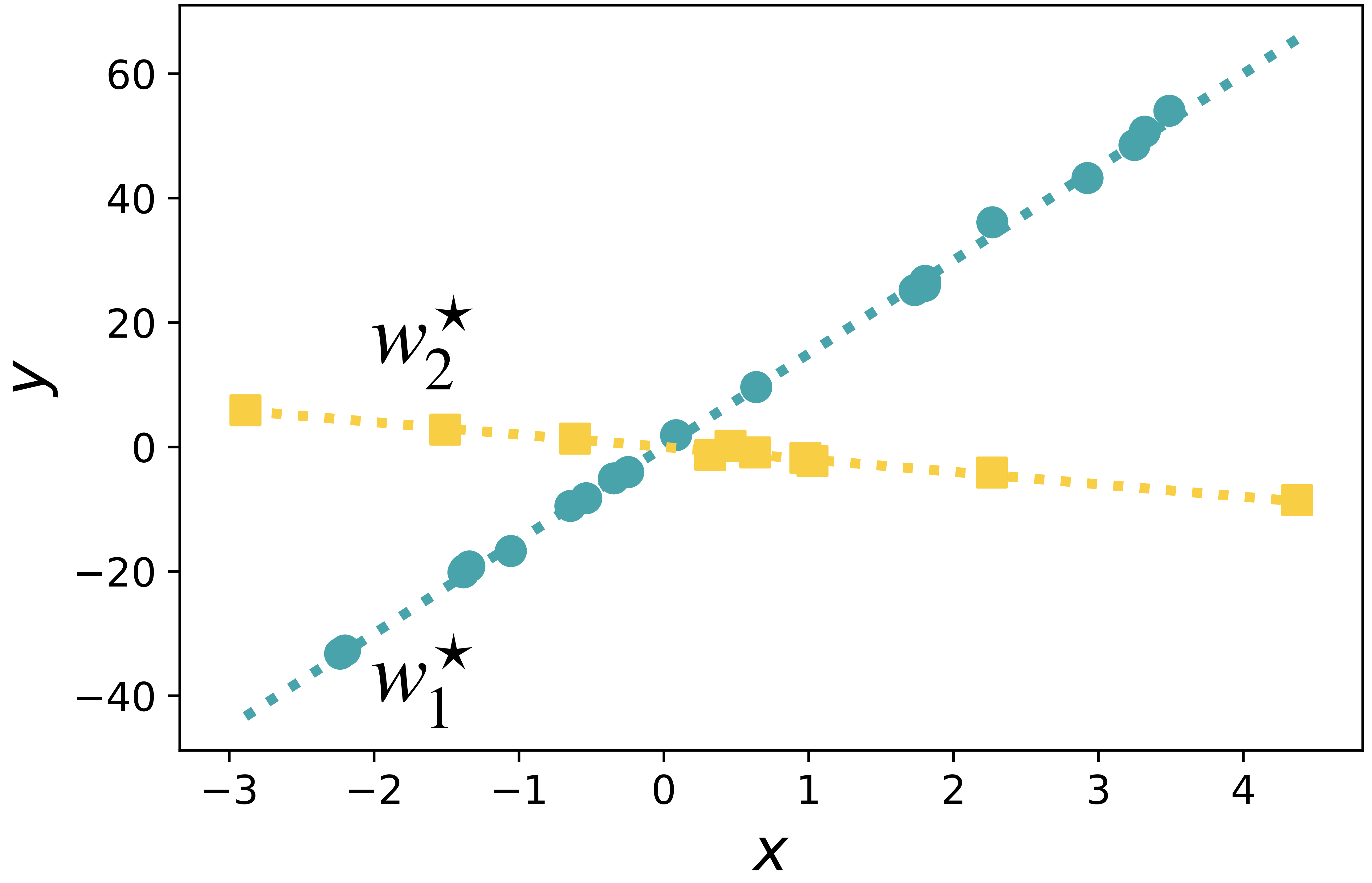
Part 3: Model personalization for federated learning

[TSP ’22, ICML ’22]



Two regimes





Objective $$\min_{w \in \mathbb{R}^d} \quad \frac{1}{n}\sum_{i=1}^{n} F_i(w) \qquad \text{where} \qquad F_i(w) = \mathbb{E}_{z \sim p_i}\left[f(w; z)\right]$$

loss on client i

Option 1: Train a separate model per client (no collaboration)

Objective: $\min_{w_i} F_i(w_i)$

Privacy	Best possible
Performance	Poor

Option 2: The model has a global component and a per-client component

Shared Params u + Personal Params v_i = Full model $w_i = (u, v_i)$

Objective: $$\min_{u, v_1, \cdots, v_n} \frac{1}{n} \sum_{i=1}^{n} F_i(u, v_i)$$

Example: $F_i(u, v_i) = \mathbb{E}_{(X,Y)\sim p_i} \Big(\phi_g(X;\, u) + \phi_l(X;\, v_i) - Y \Big)^2$

Option 2: The model has a global component and a per-client component

Shared Params u + Personal Params v_i = Full model $w_i = (u, v_i)$

Objective: $$\min_{u, v_1, \cdots, v_n} \frac{1}{n} \sum_{i=1}^{n} F_i(u, v_i)$$

Privacy	Yes
Performance	

data and personal params on client

Option 2: The model has a global component and a per-client component

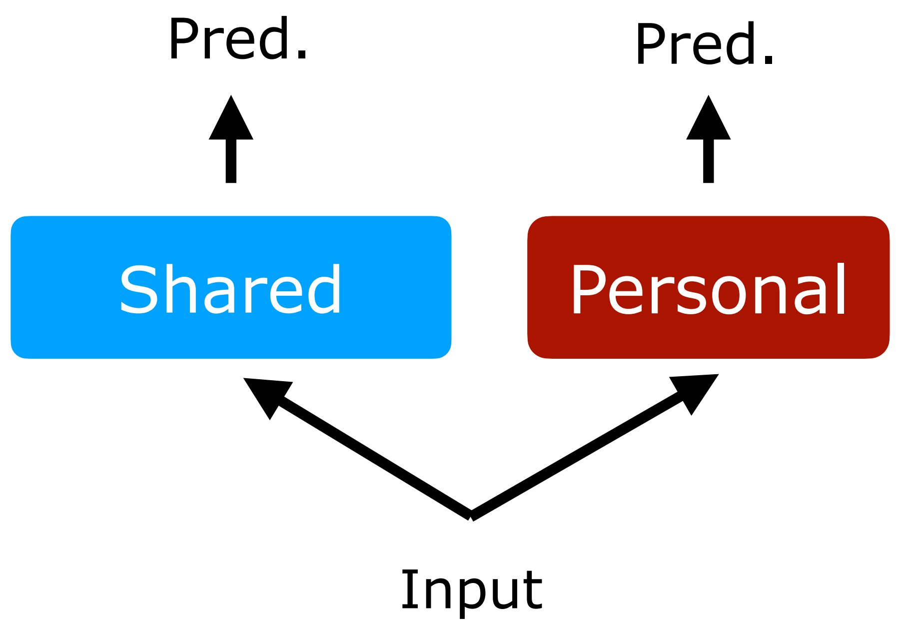
Shared Params u + Personal Params v_i = Full model $w_i = (u, v_i)$

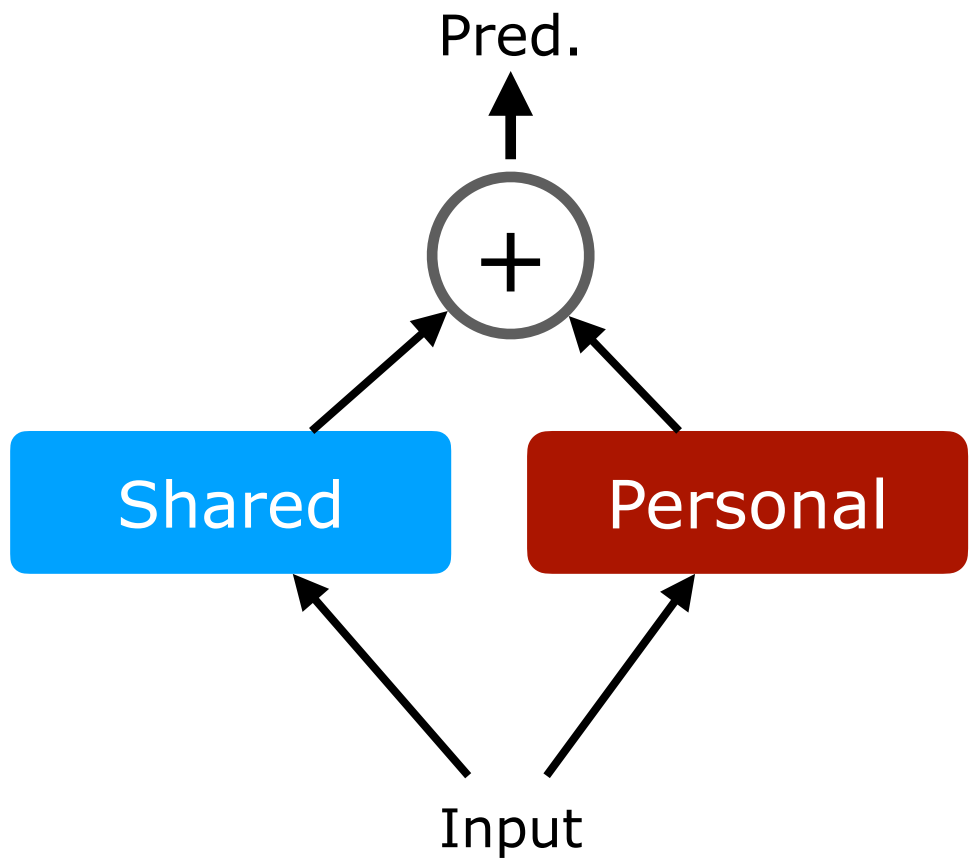
Objective: $$\min_{u, v_1, \cdots, v_n} \frac{1}{n} \sum_{i=1}^{n} F_i(u, v_i)$$

Privacy	Yes
Performance	***Yes***

data and personal params on client

Personalization Architectures



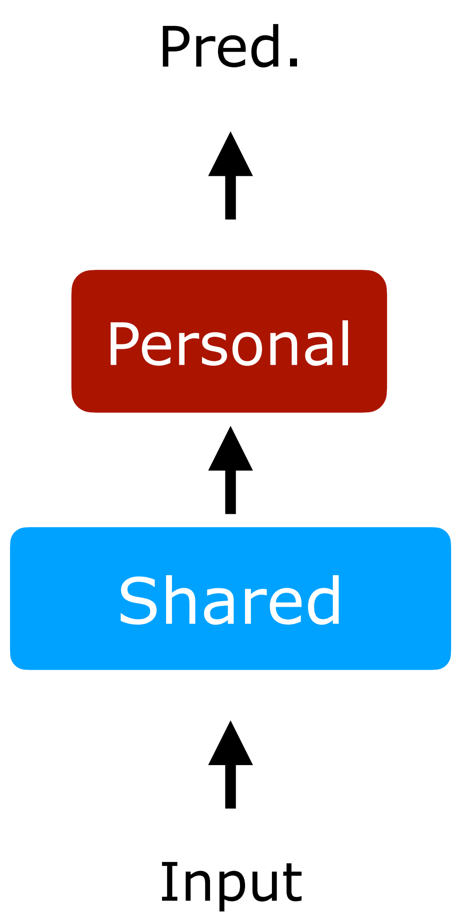


$$F_i(u, v_i) = \mathbb{E}_{(X,Y)\sim p_i}\left(\phi_g(X\,;\,u) + \phi_l(X\,;\,v_i) - Y\right)^2$$

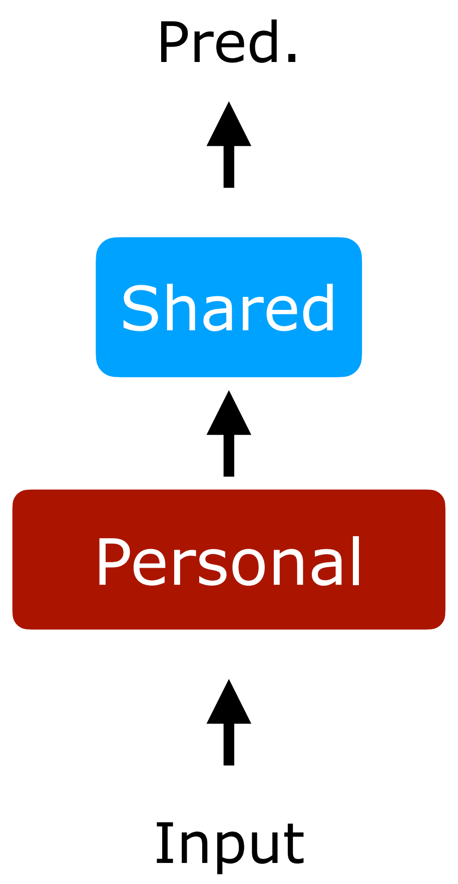
Multi-task learning: Caruana (1997), Baxter (2000), Evgeniou & Pontil (2004), Collobert & Weston (2005), Argyriou et al. (2008), …

Personalization Architectures

$$F_i(u, v_i) = \mathbb{E}_{(X,Y)\sim p_i}\left(\phi_g(X;u) + \phi_l(X;v_i) - Y\right)^2$$



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



Liang et al. (2019)

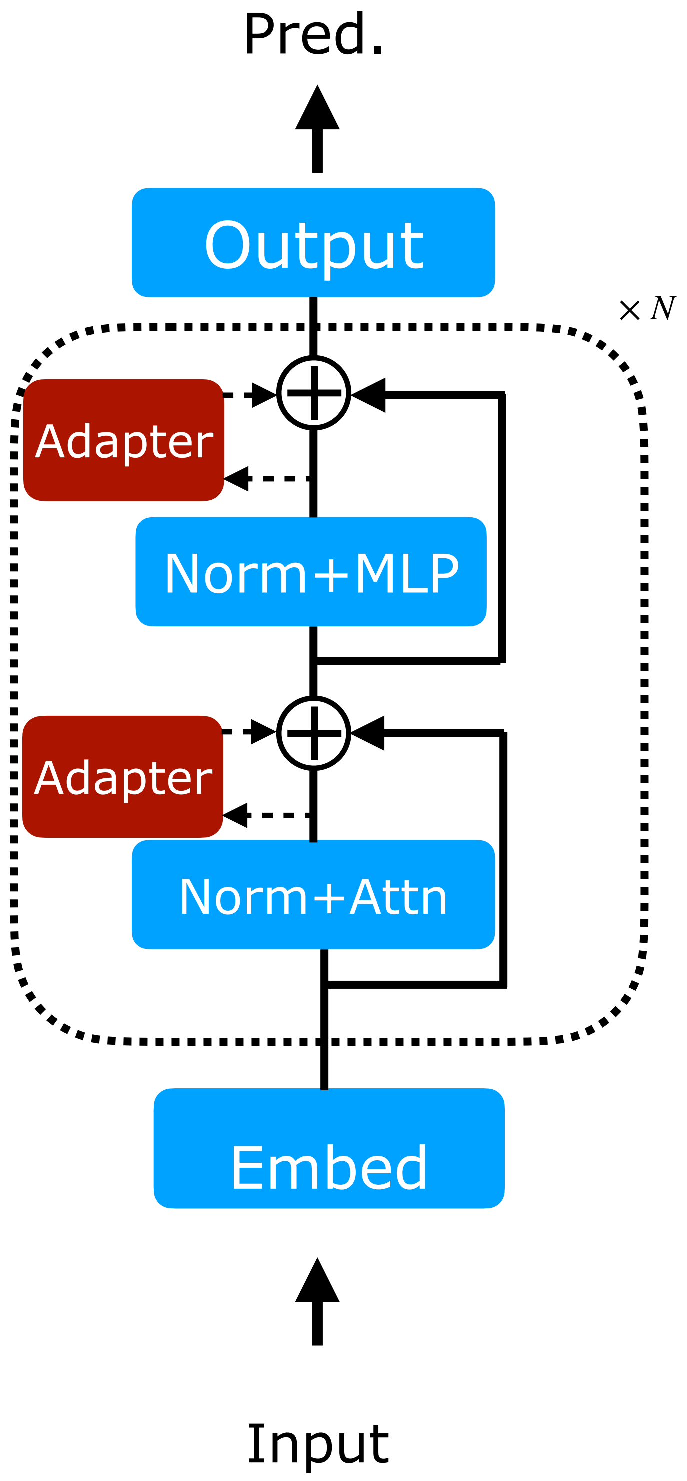
Multi-task learning: Caruana (1997), Baxter (2000), Evgeniou & Pontil (2004), Collobert & Weston (2005), Argyriou et al. (2008), …

Personalization Architectures

$$F_i(u, v_i) = \mathbb{E}_{(X,Y)\sim p_i}\left(\phi_g(X\,;\,u) + \phi_l(X\,;\,v_i) - Y\right)^2$$

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

Liang et al. (2019)



Personalized Adapters

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

Optimization

- Server samples m clients and broadcast global model u
- **Local updates** on client i: $(u_i^+, v_i^+) = \text{LocalUpdate}_i(u, v_i)$
- Aggregate updates to global part of the model: $u^+ = \frac{1}{m} \sum_i u_i^+$

Alternating update

$$v_i^+ = v_i - \gamma \nabla_v F_i(u, v_i)$$

$$u_i^+ = u - \gamma \nabla_u F_i(u, v_i^+)$$

Simultaneous update

$$v_i^+ = v_i - \gamma \nabla_v F_i(u, v_i)$$

$$u_i^+ = u - \gamma \nabla_u F_i(u, v_i)$$

Theorem [*P.*, Malik, Mohamed, Rabbat, Sanjabi, Xiao]

For smooth, nonconvex functions, we have the rates:

Alternating update: $\frac{\sigma_1^2}{\sqrt{t}}$

Simultaneous update: $\frac{\sigma_2^2}{\sqrt{t}}$

where $\sigma_1^2 < \sigma_2^2$ under typical scenarios

Alternating update

$$v_i^+ = v_i - \gamma \nabla_v F_i(u, v_i)$$

$$u_i^+ = u - \gamma \nabla_u F_i(u, v_i^+)$$

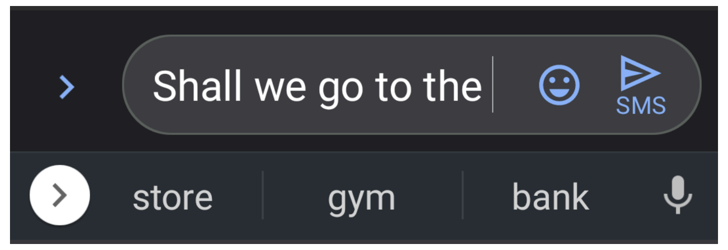
Simultaneous update

$$v_i^+ = v_i - \gamma \nabla_v F_i(u, v_i)$$

$$u_i^+ = u - \gamma \nabla_u F_i(u, v_i)$$

Experimentally, small but consistent trend of alternating $\succ$ simultaneous

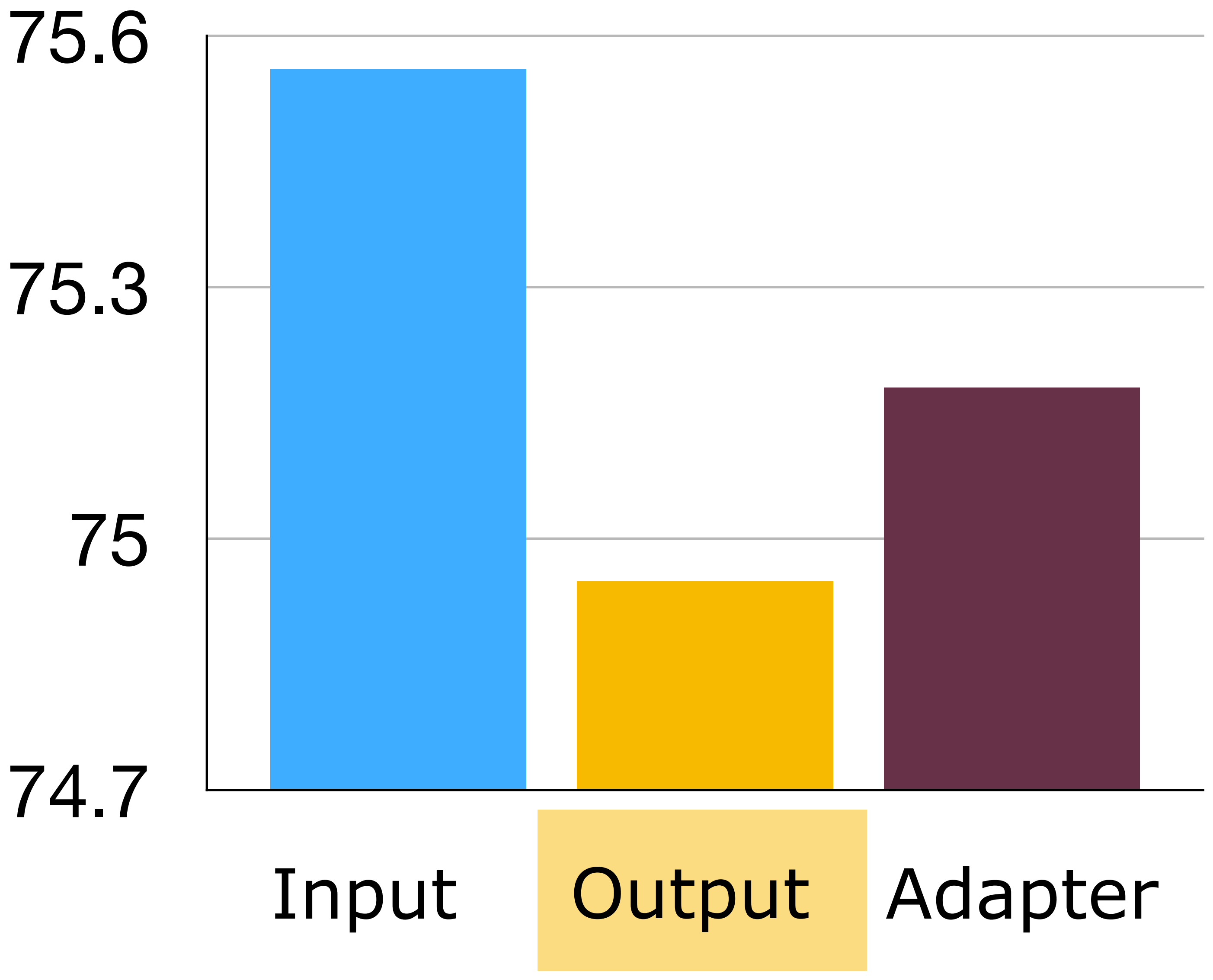
Experiments

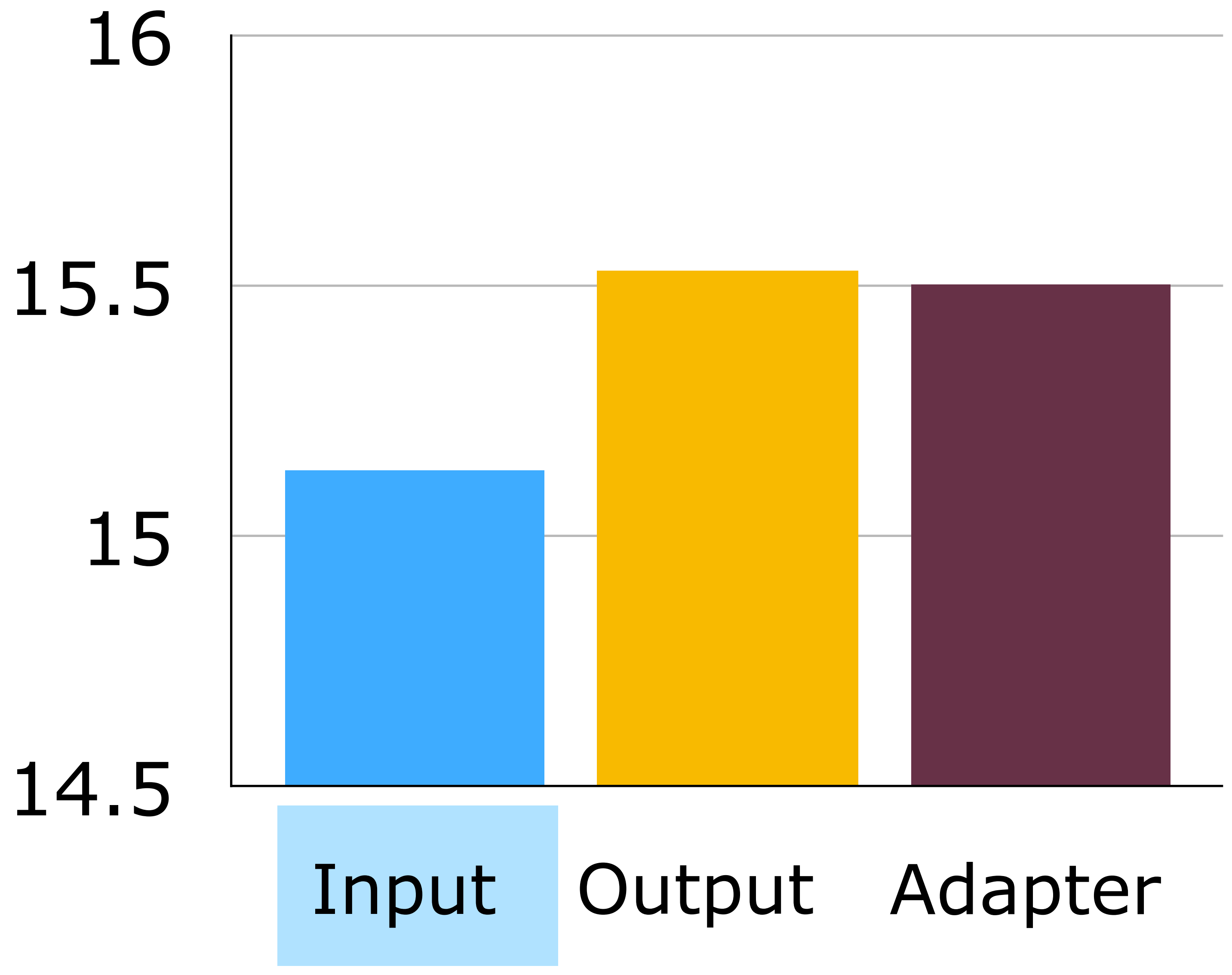


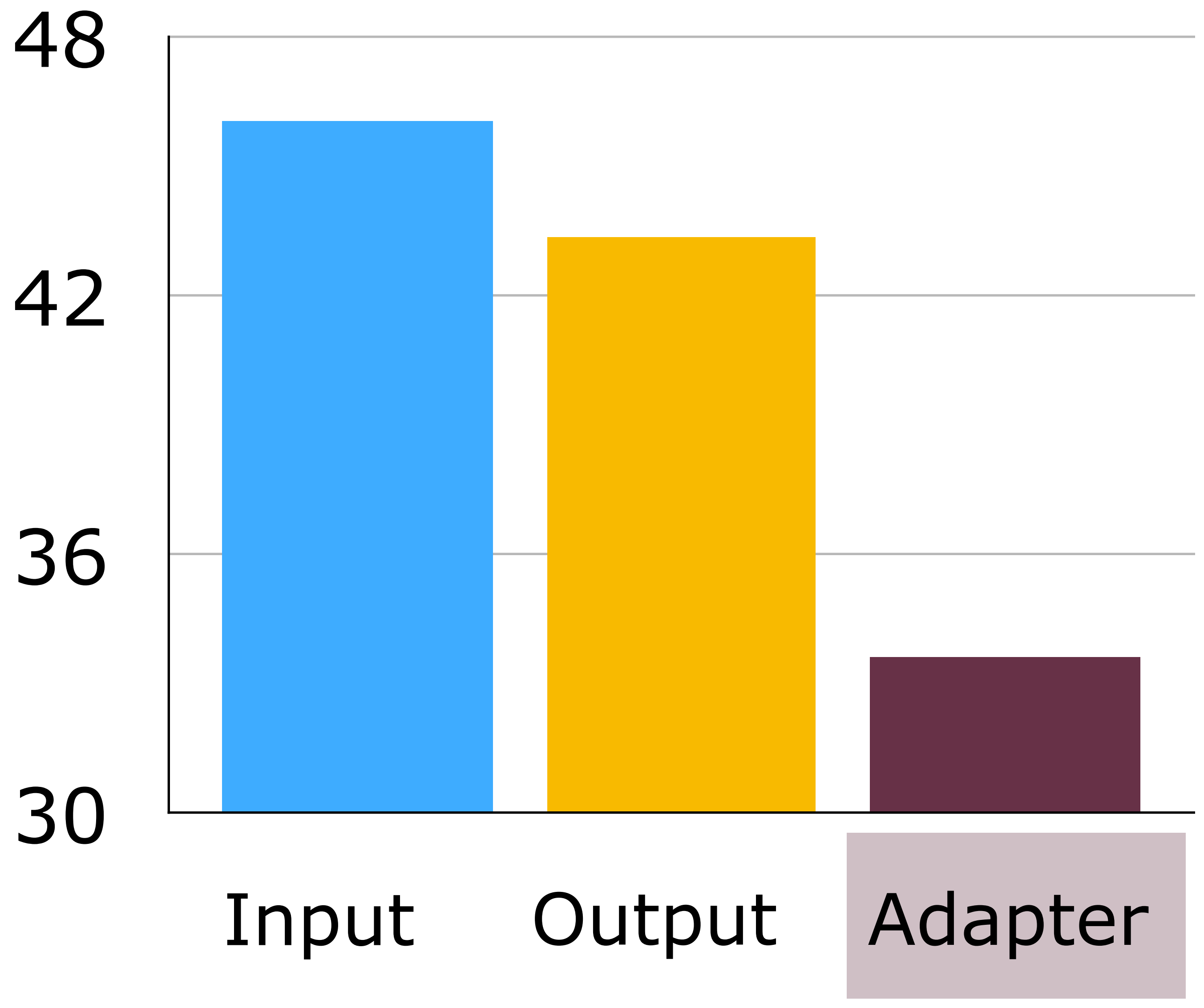
Next word prediction

Speech recognition

Landmark detection

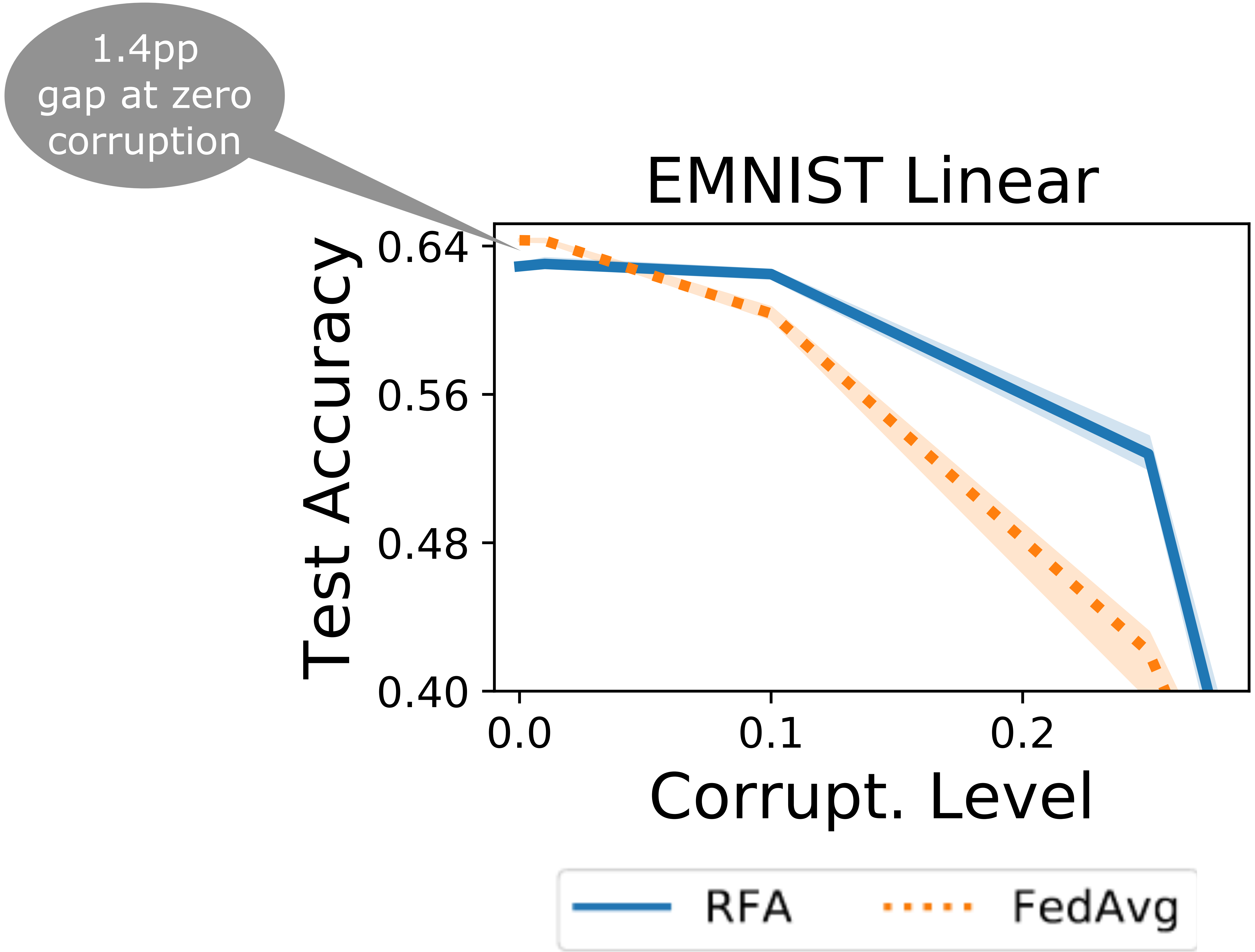




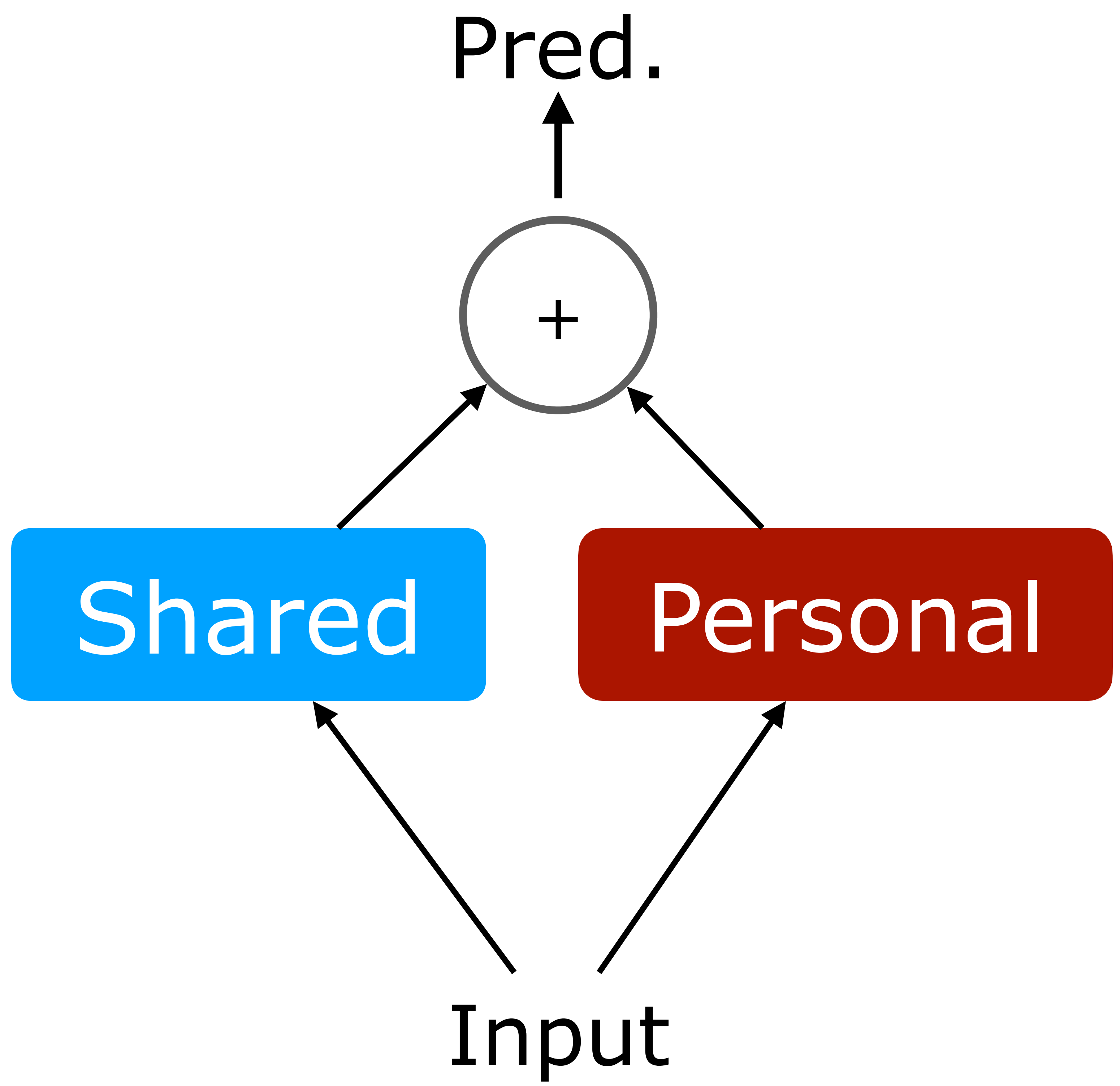


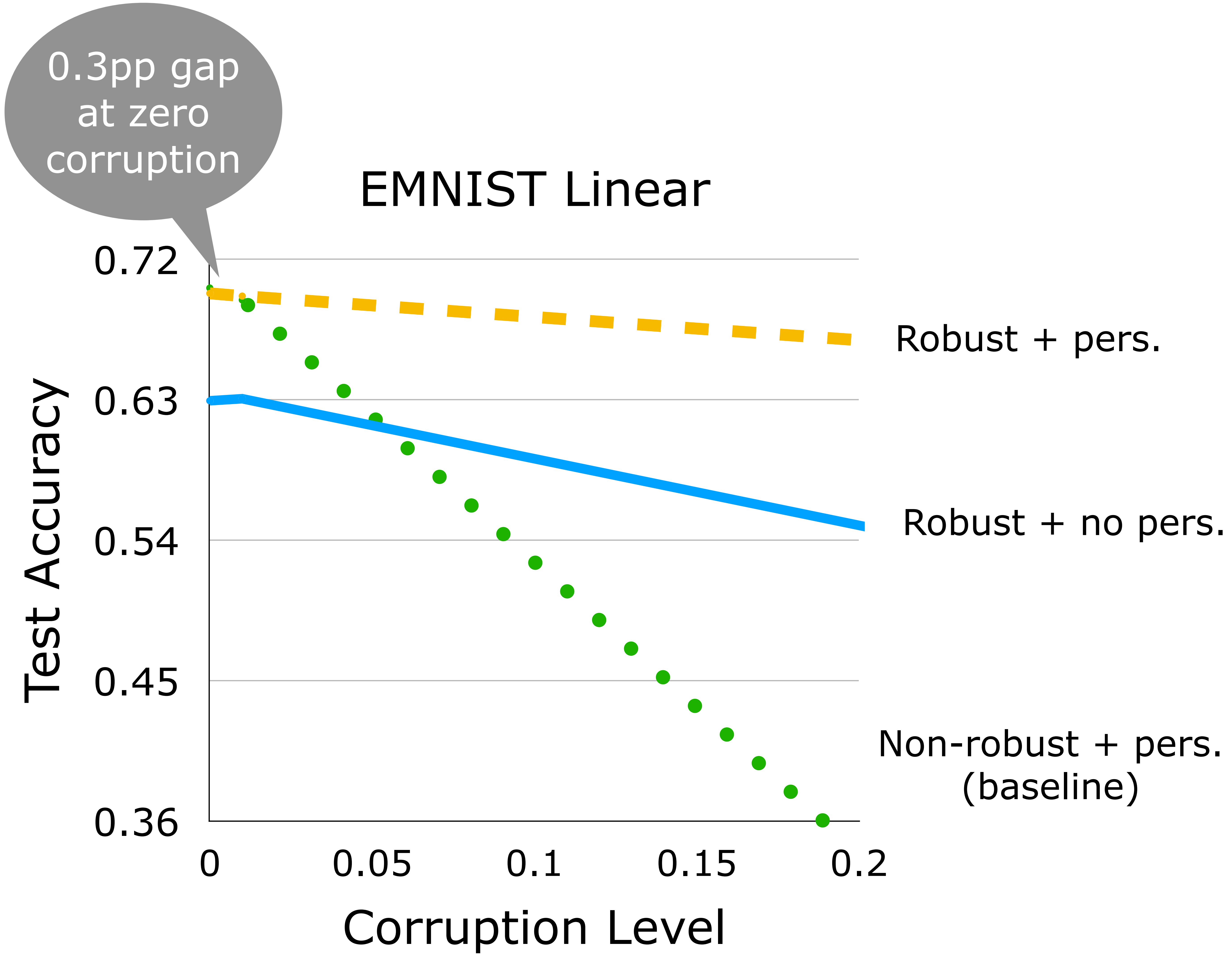
y-axis shows error: lower is better

Recall: robust aggregation experiments



Improving robust aggregation with personalization

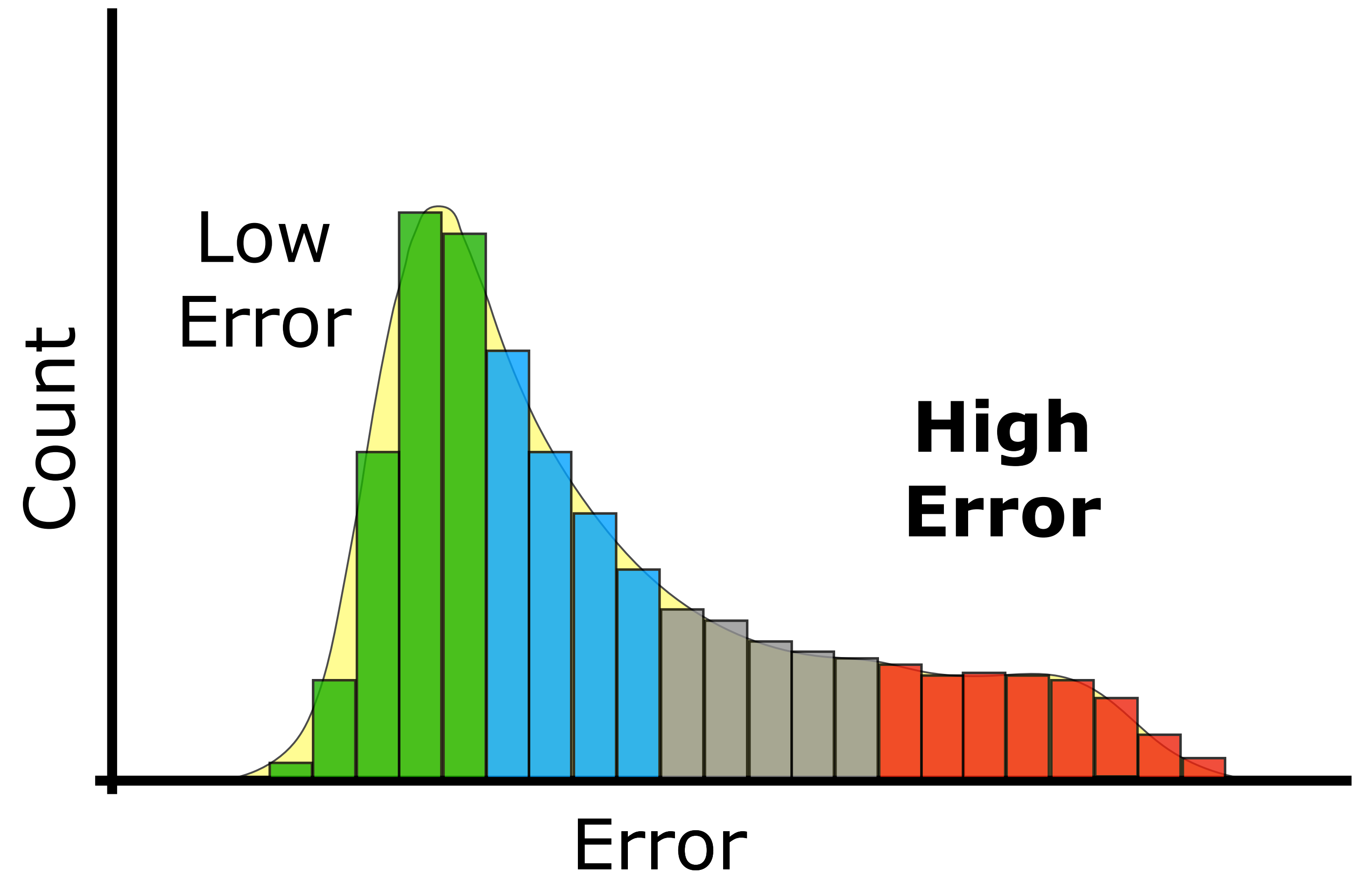




Summary

Part 1: Heterogeneity-aware objectives for federated learning

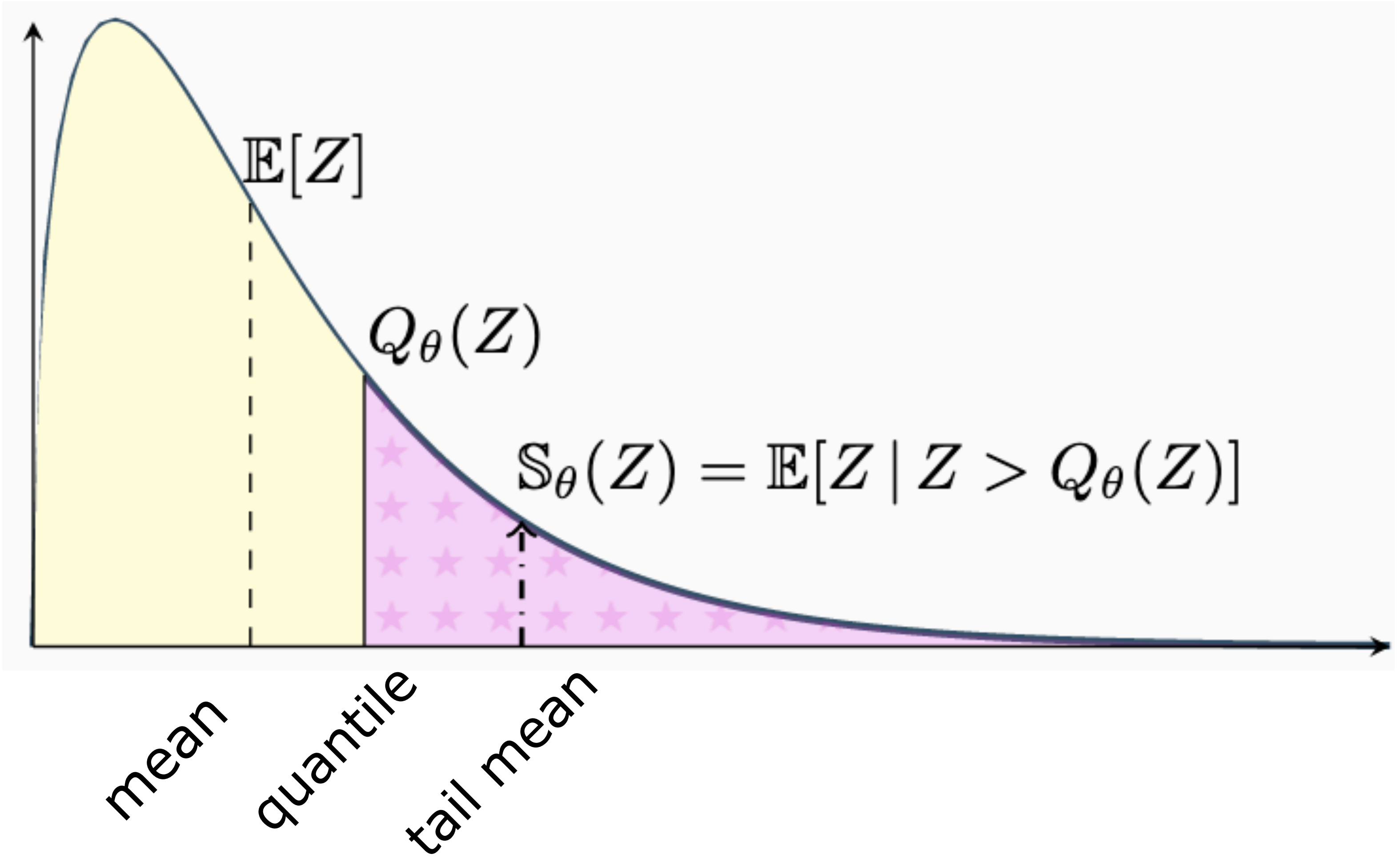
Heterogeneity $\Longrightarrow$
large tail errors



Part 1: Heterogeneity-aware objectives for federated learning

Heterogeneity $\Longrightarrow$ large tail errors

$$\min_{w} \mathbb{S}_\theta\Big(\big(F_1(w), \cdots, F_n(w)\big) \Big)$$

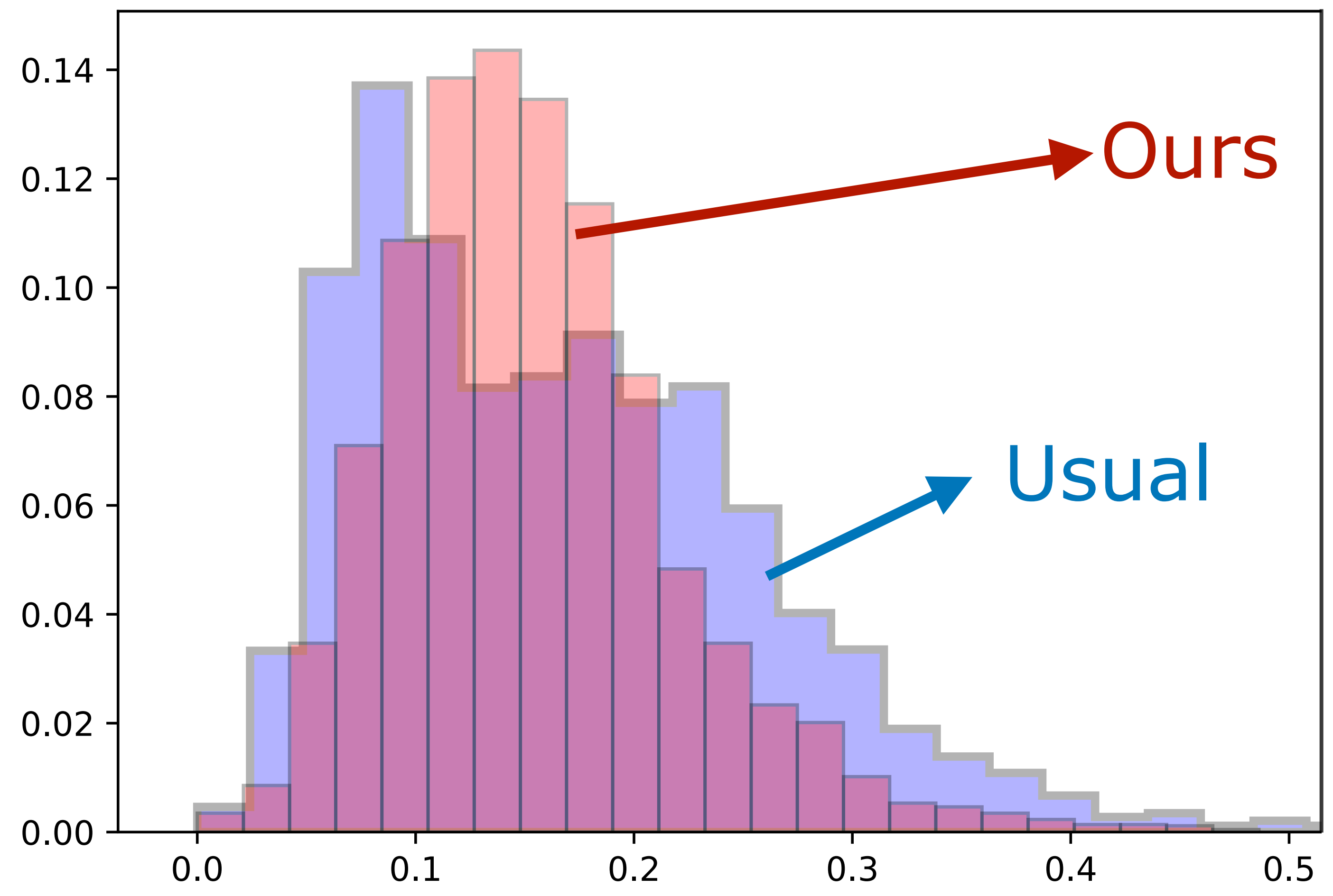


Part 1: Heterogeneity-aware objectives for federated learning

Heterogeneity $\Longrightarrow$ large tail errors

$$\min_{w} \mathbb{S}_\theta\Big(\big(F_1(w), \cdots, F_n(w)\big) \Big)$$

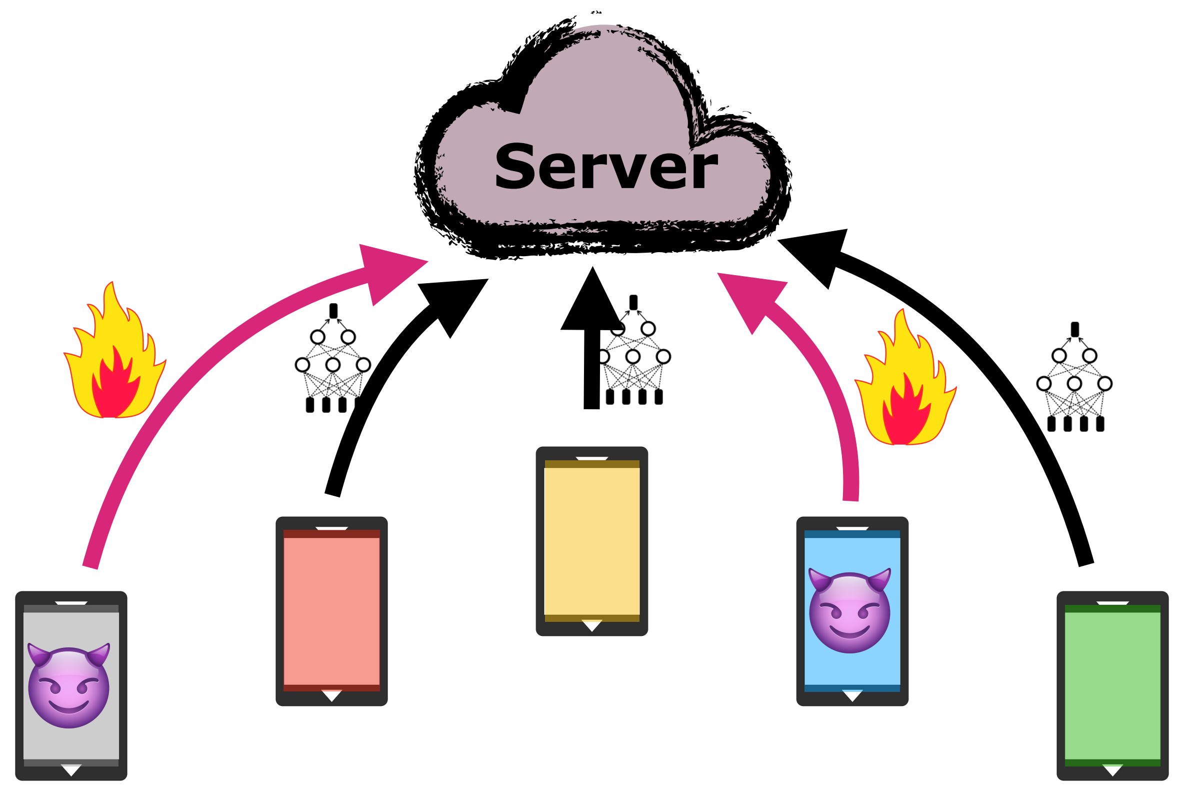
Our approach reduces tail error



Misclassification Error

Part 2: Robust aggregation for federated learning

Arithmetic mean $\Longrightarrow$ not robust to poisoned updates



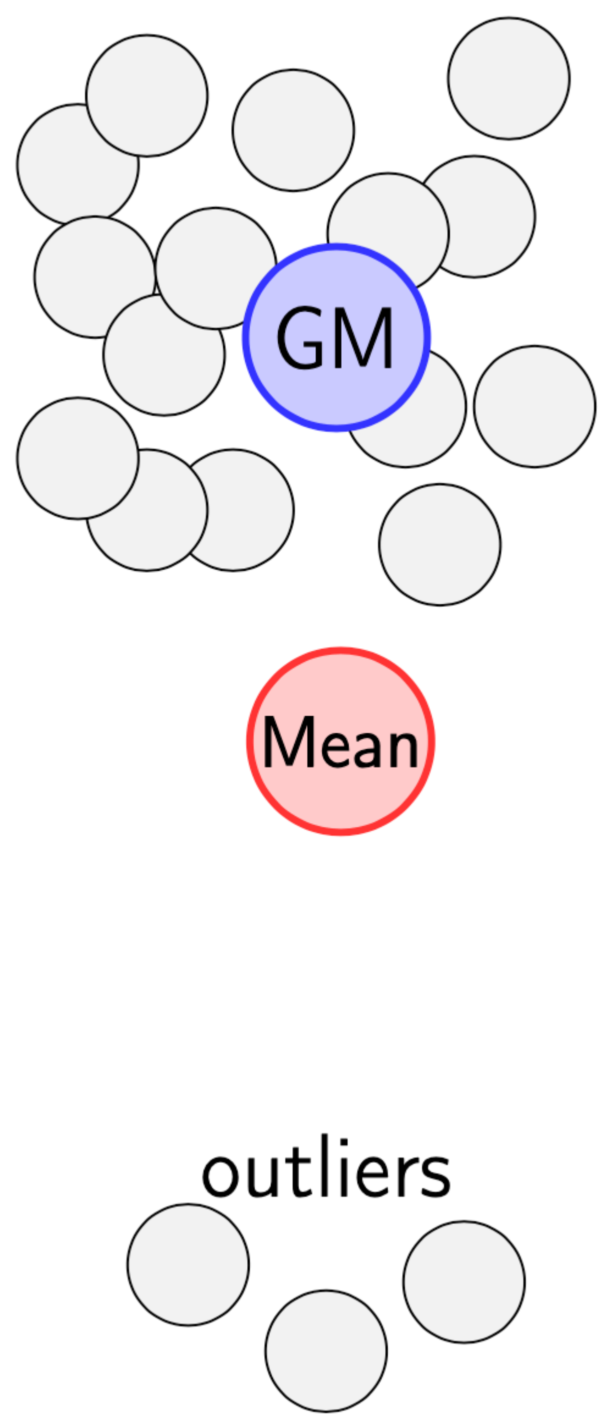
Data poisoning

Model poisoning

Part 2: Robust aggregation for federated learning

Arithmetic mean $\Longrightarrow$ not robust to poisoned updates

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

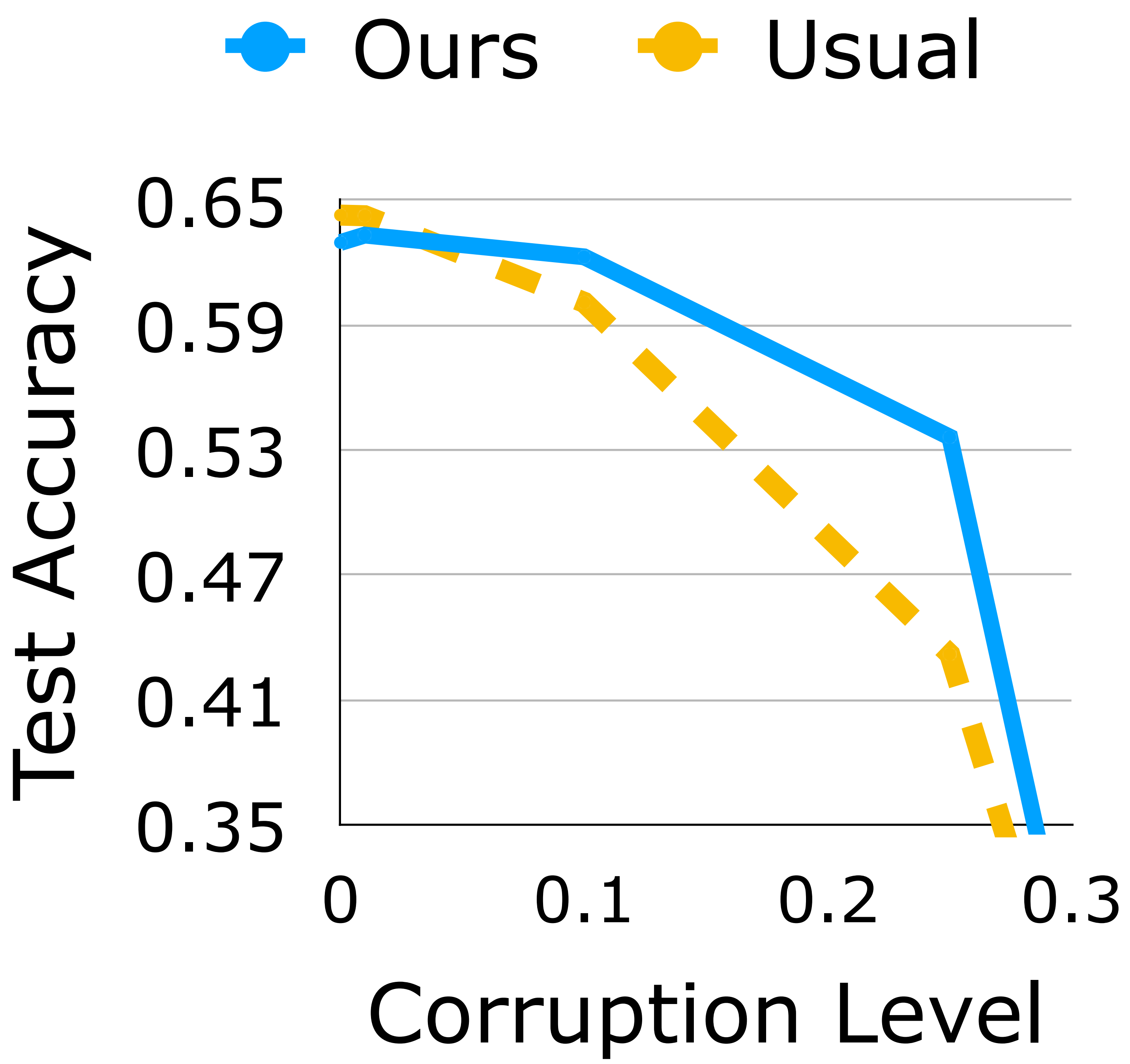


Part 2: Robust aggregation for federated learning

Arithmetic mean $\Longrightarrow$ not robust to poisoned updates

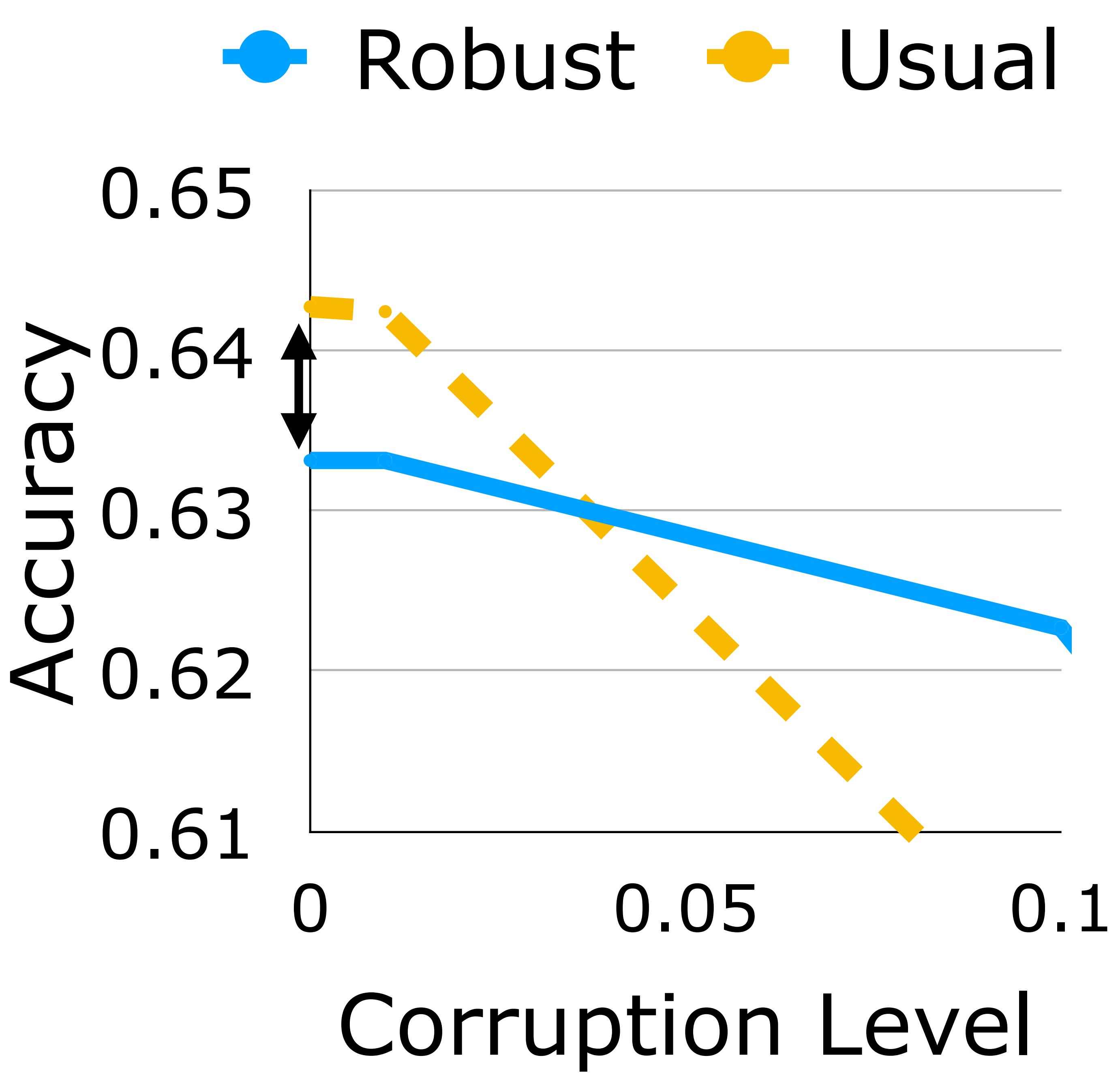
$$\mathrm{GM} = \arg\min_{z} \sum_{i=1}^{m} \|z - w_i\|_2$$

Our approach gives greater robustness



Part 3: Model personalization for federated learning

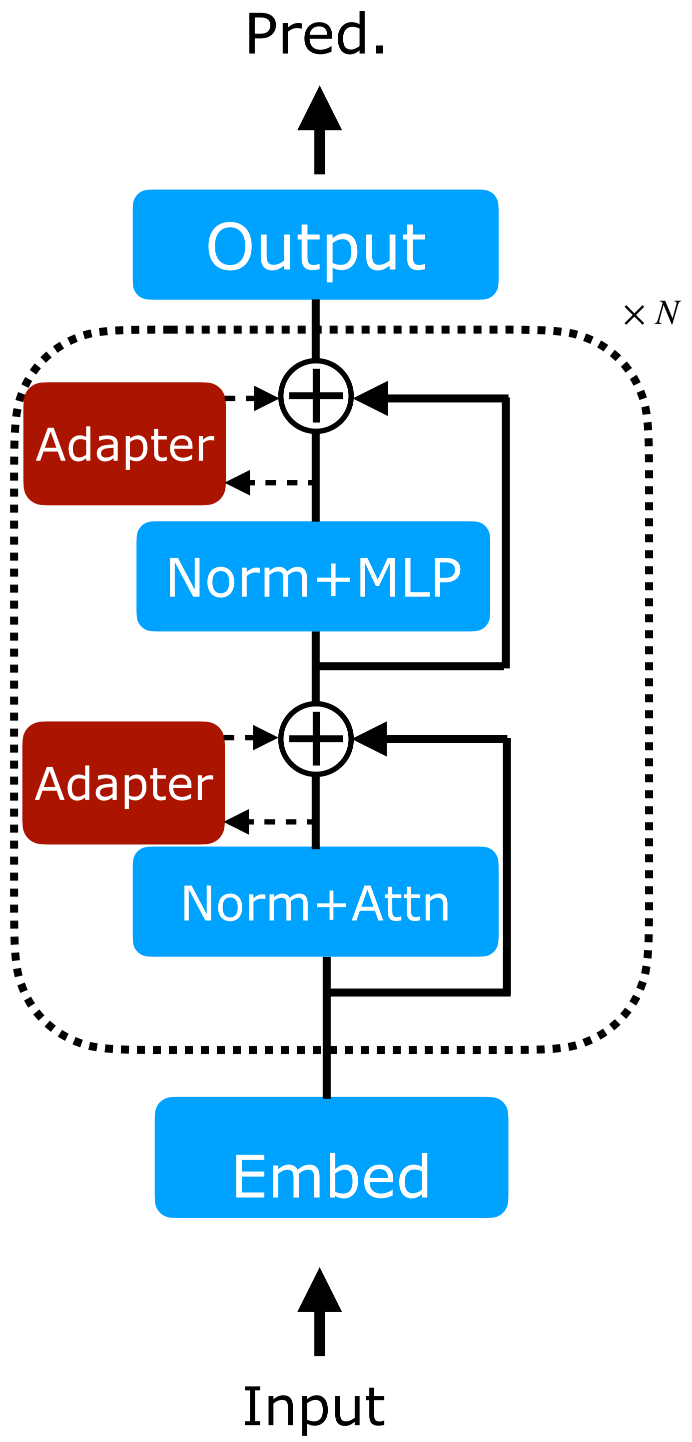
Heterogeneity & robustness at odds



Part 3: Model personalization for federated learning

Heterogeneity &
robustness at odds

$$\min_{u, v_1, \cdots, v_n} \frac{1}{n} \sum_{i=1}^{n} F_i(u, v_i)$$

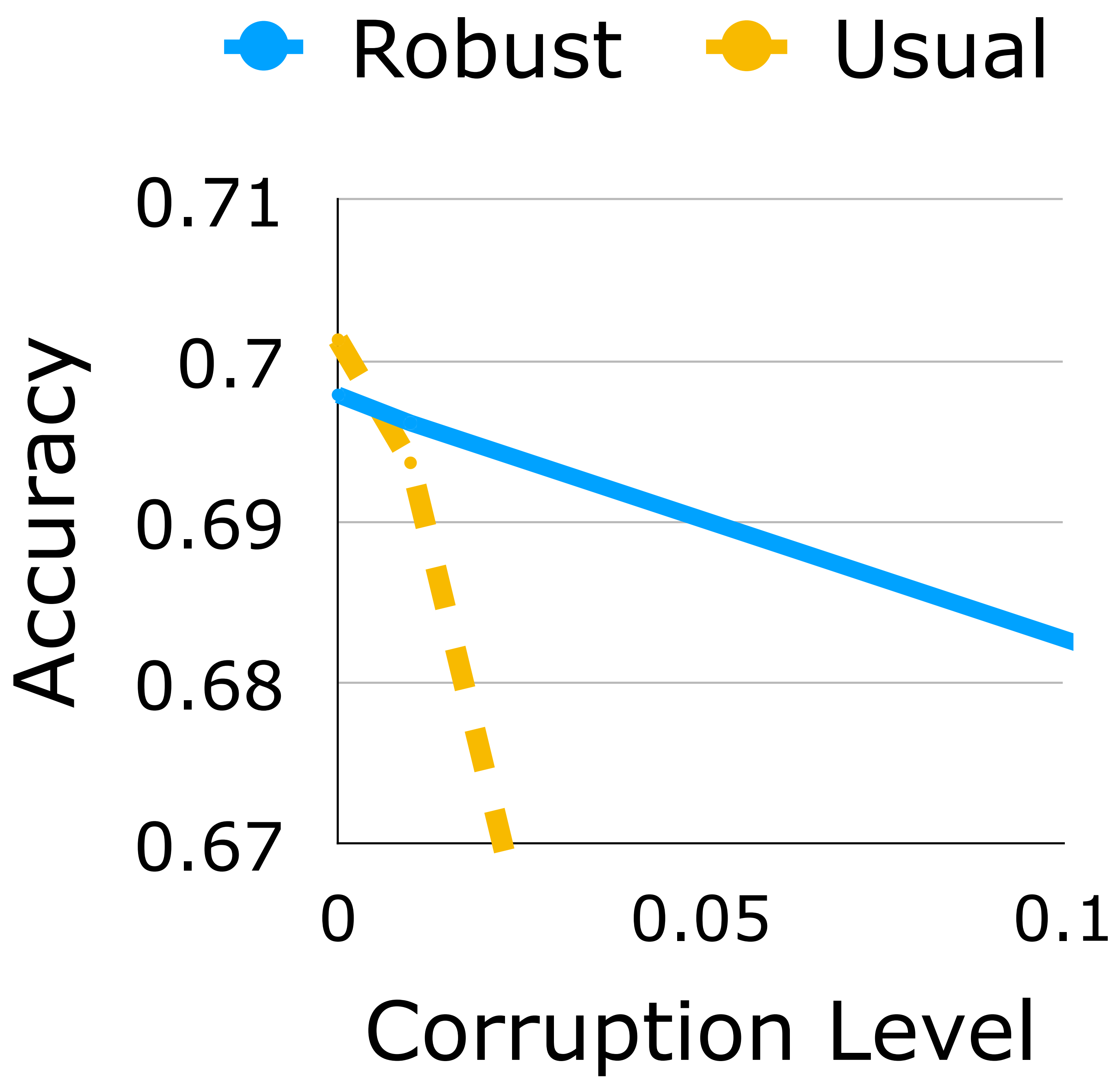


Part 3: Model personalization for federated learning

Heterogeneity & robustness at odds

$$\min_{u, v_1, \cdots, v_n} \frac{1}{n} \sum_{i=1}^{n} F_i(u, v_i)$$

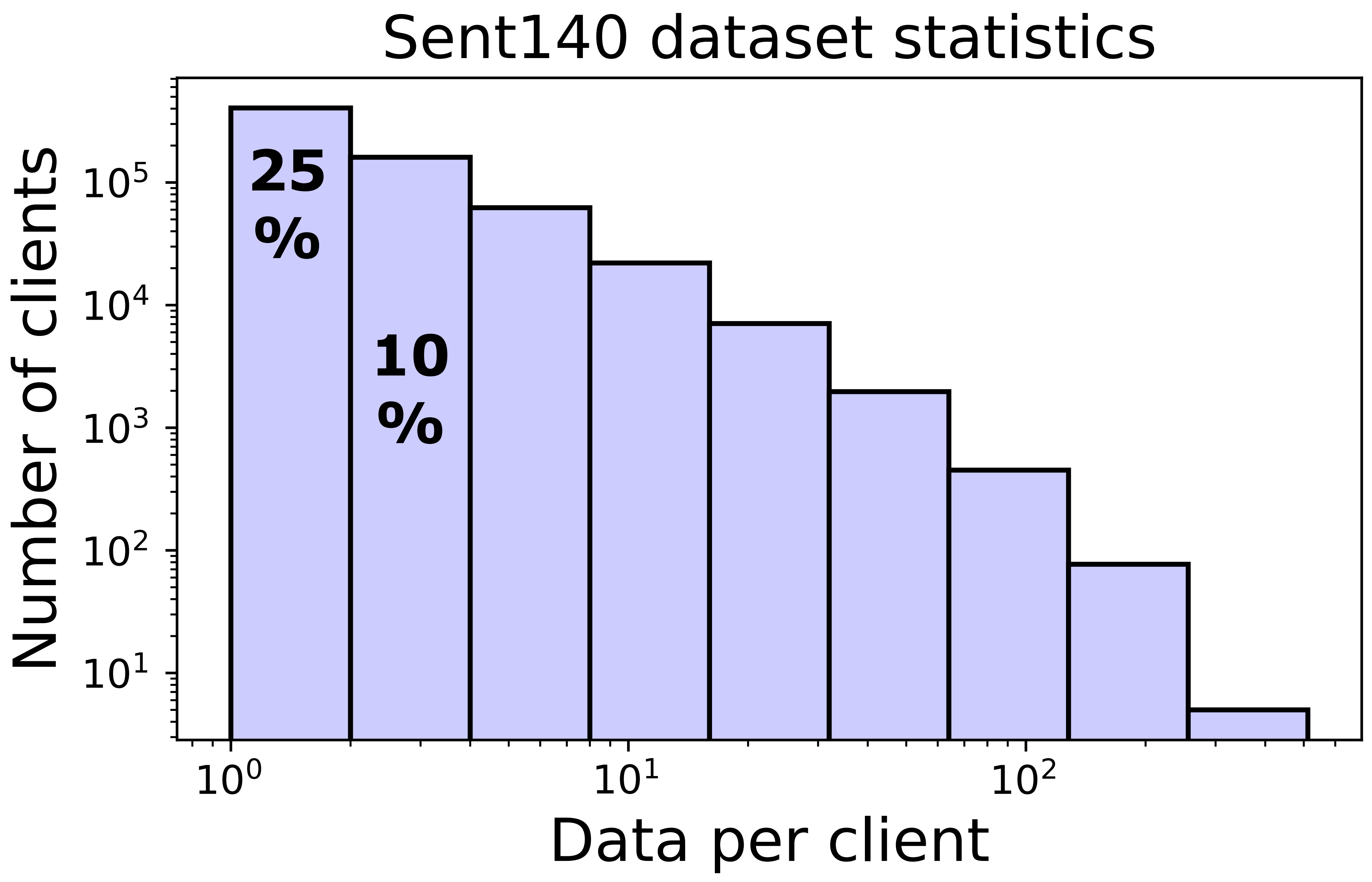
Can tailor to heterogeneity & retain robustness

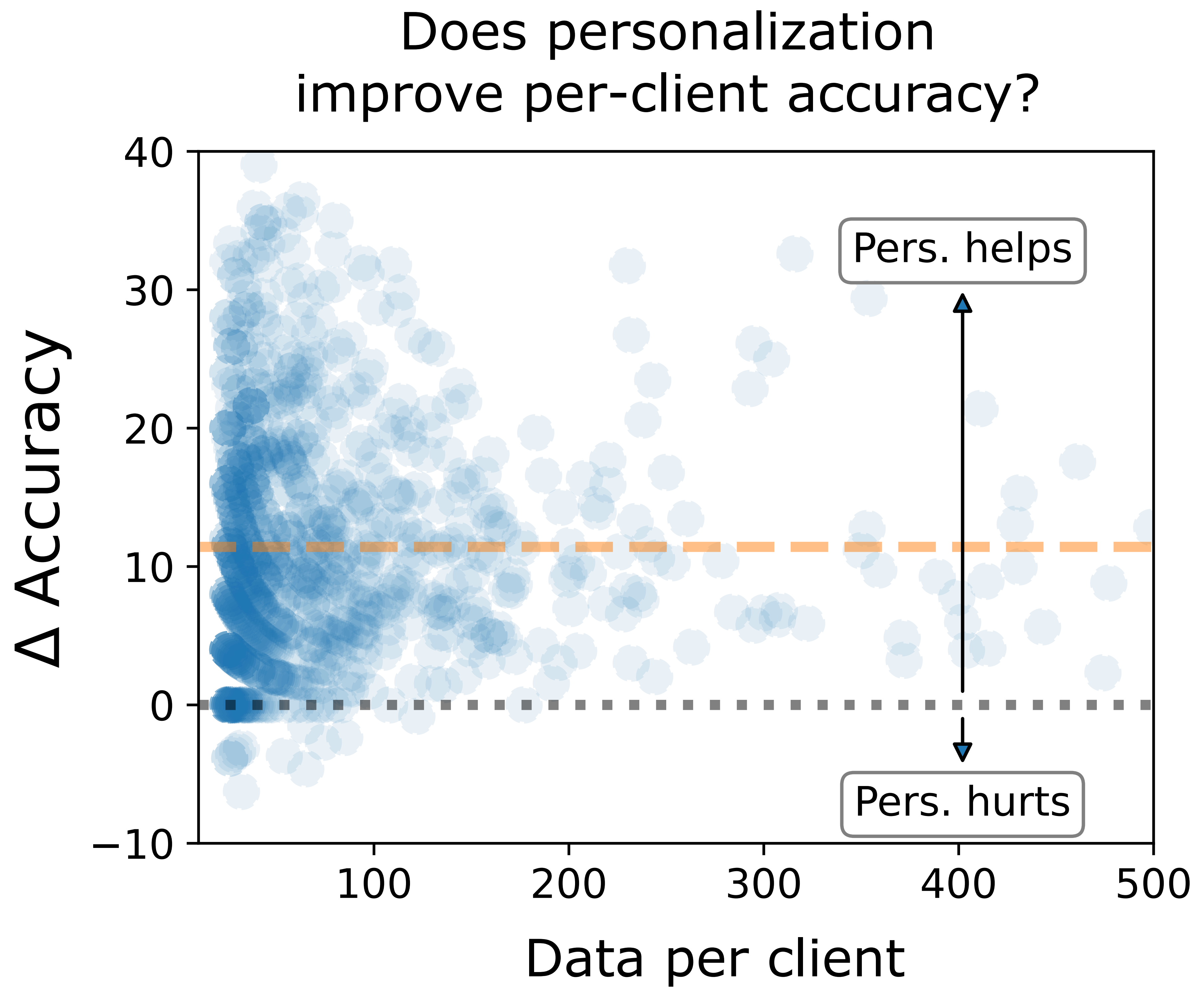


Perspectives and conclusion

The small data problem

- Per-client evaluations are not reliable
- Personalization can overfit





Understanding heterogeneity

Many negative results: optimization can slow down, makes robustness harder, …

Yet, federated learning works.

Understanding heterogeneity

Many negative results: optimization can slow down, makes robustness harder, …

Yet, federated learning works.

Quantify heterogeneity:

Measure gaps between distributions: **MAUVE**

[**P**., Swayamdipta, Zellers, Thickstun, Welleck, Choi, Harchaoui. NeurIPS (2021),
Liu, **P**., Welleck, Oh, Choi, Harchaoui. NeurIPS (2021)]

Understanding heterogeneity

Many negative results: optimization can slow down, makes robustness harder, …

Yet, federated learning works.

Quantify heterogeneity:

Measure gaps between distributions: **MAUVE**

[**P**., Swayamdipta, Zellers, Thickstun, Welleck, Choi, Harchaoui. NeurIPS (2021),
Liu, **P**., Welleck, Oh, Choi, Harchaoui. NeurIPS (2021)]

Statistical assumptions under which heterogeneity is benign?

What measures of heterogeneity impact optimization?

Tension between heterogeneity and privacy

W
PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Thank you!

NSF

J.P.Morgan