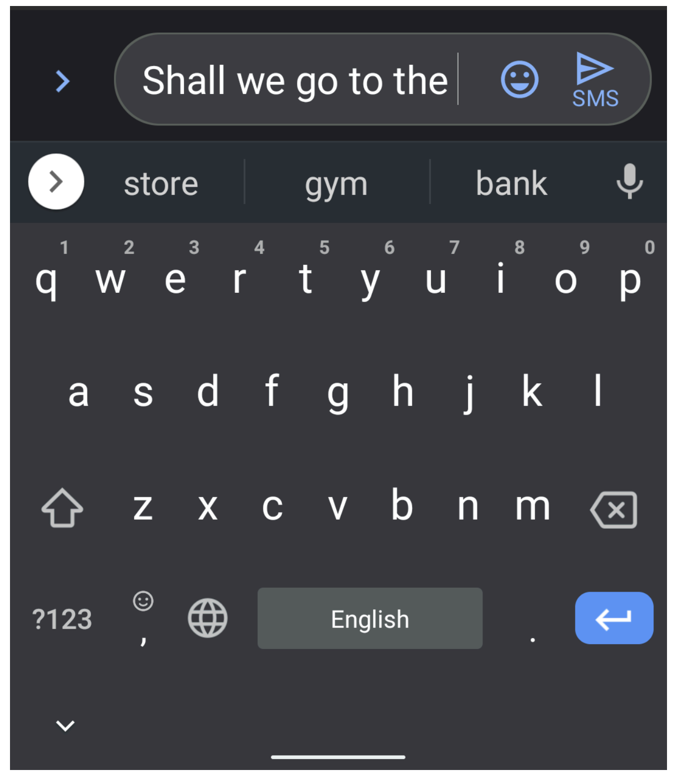
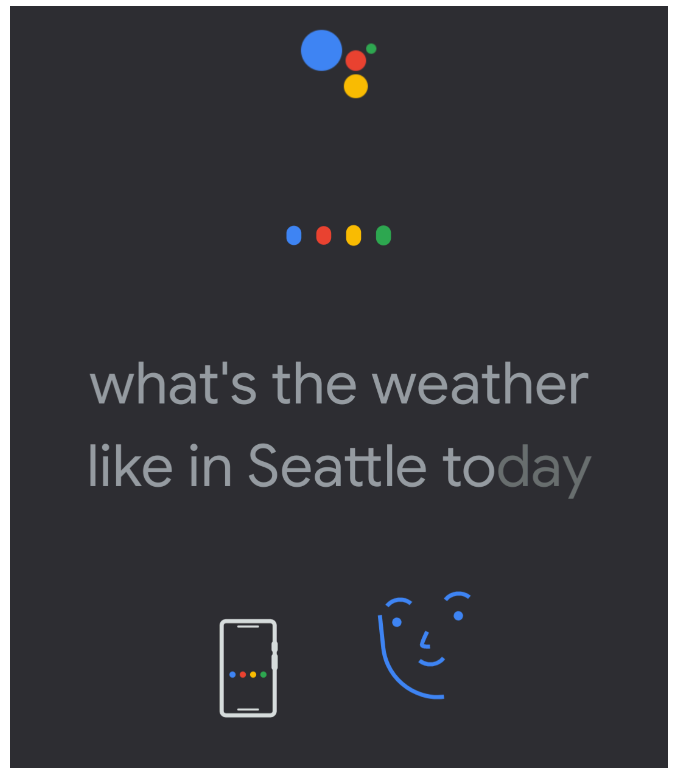
# Tackling Distribution Shifts in Federated Learning

***Krishna Pillutla***

August 6, 2022 @ DRDS Workshop

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING



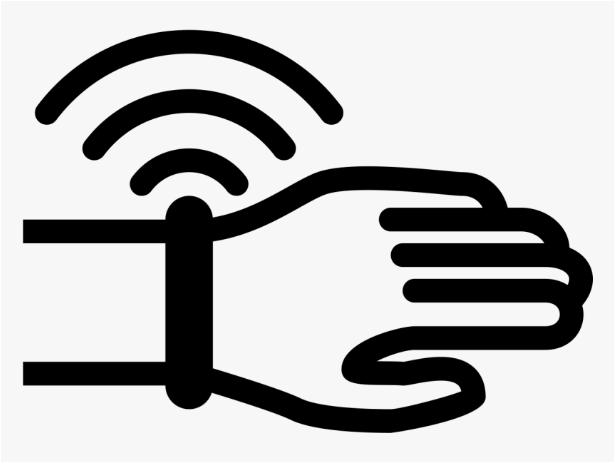




Image Credit: Robotics Business Review

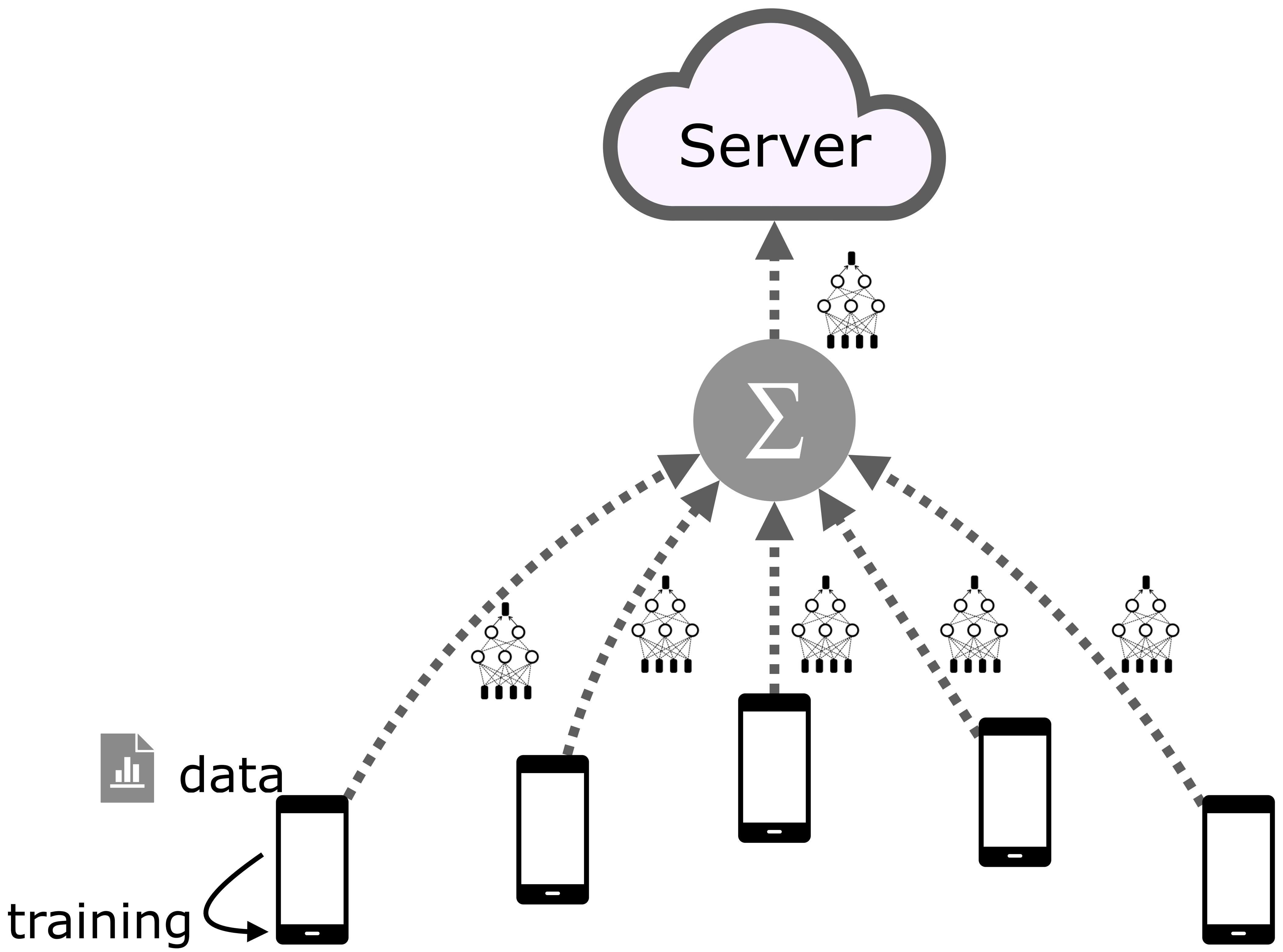
Rieke et al. NPJ Digit. Med. (2020)    Image Credit: Wellcome

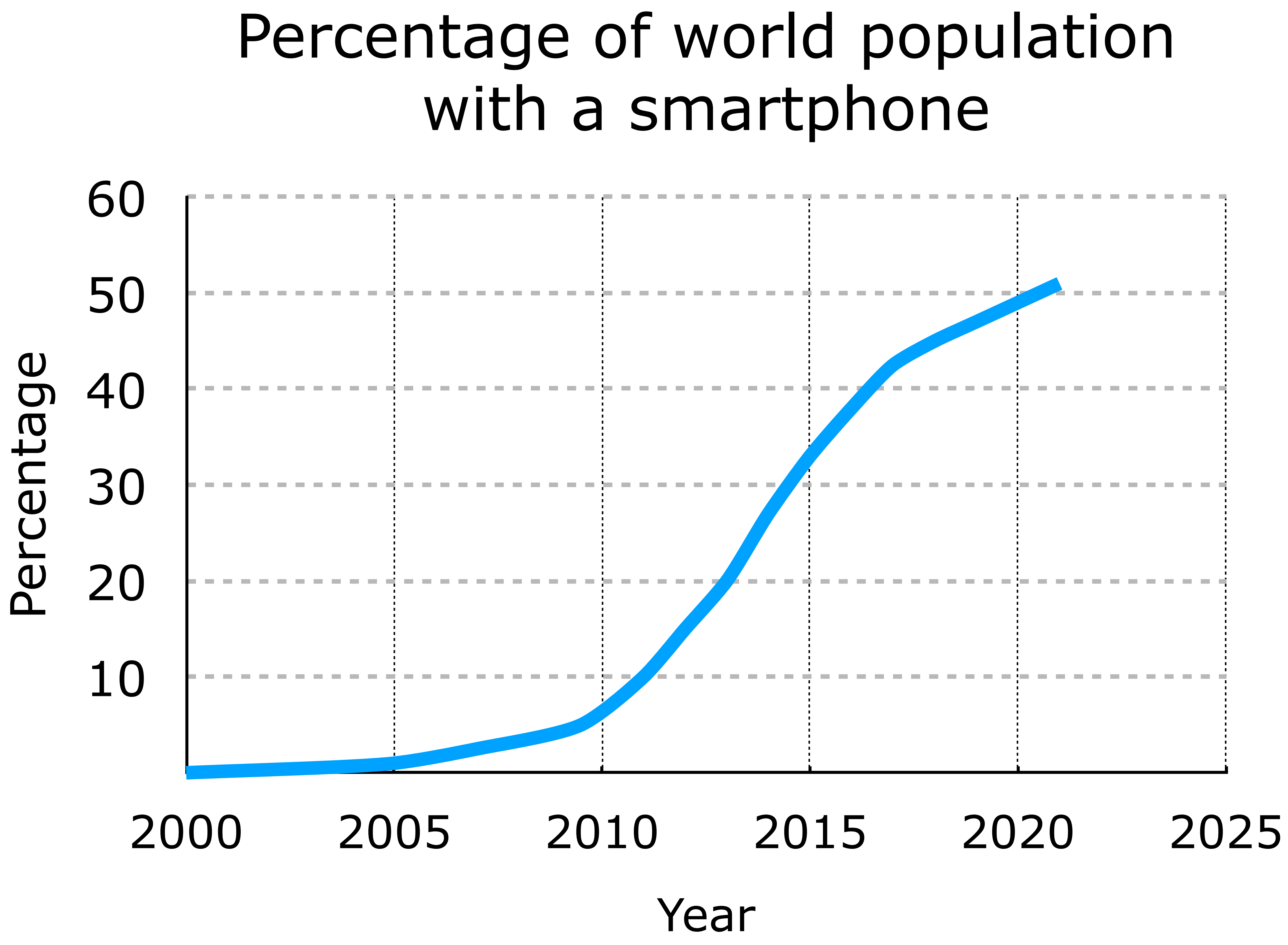
Image Credit: Robotics Business Review


# Data is decentralized and private


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

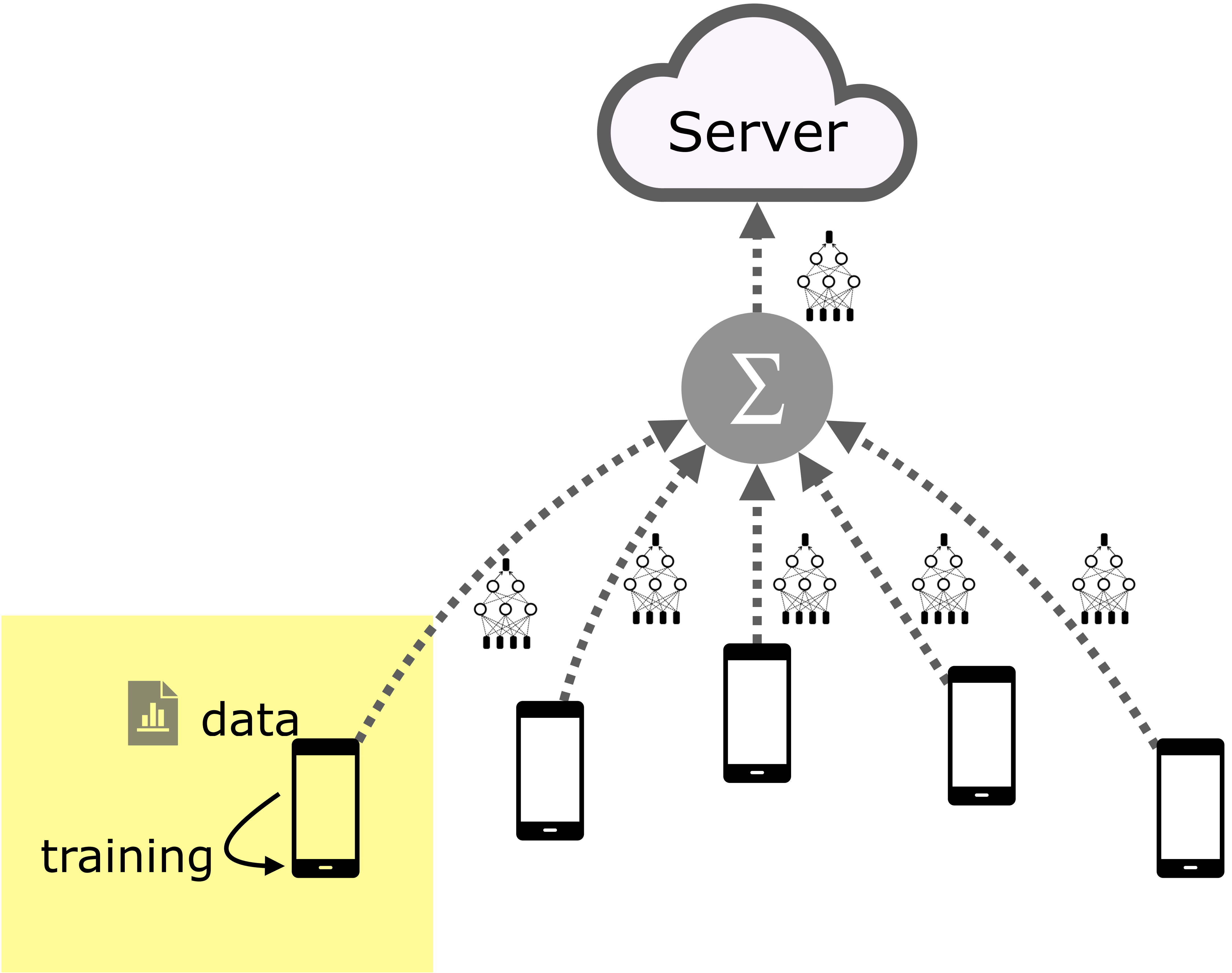
# Federated Learning

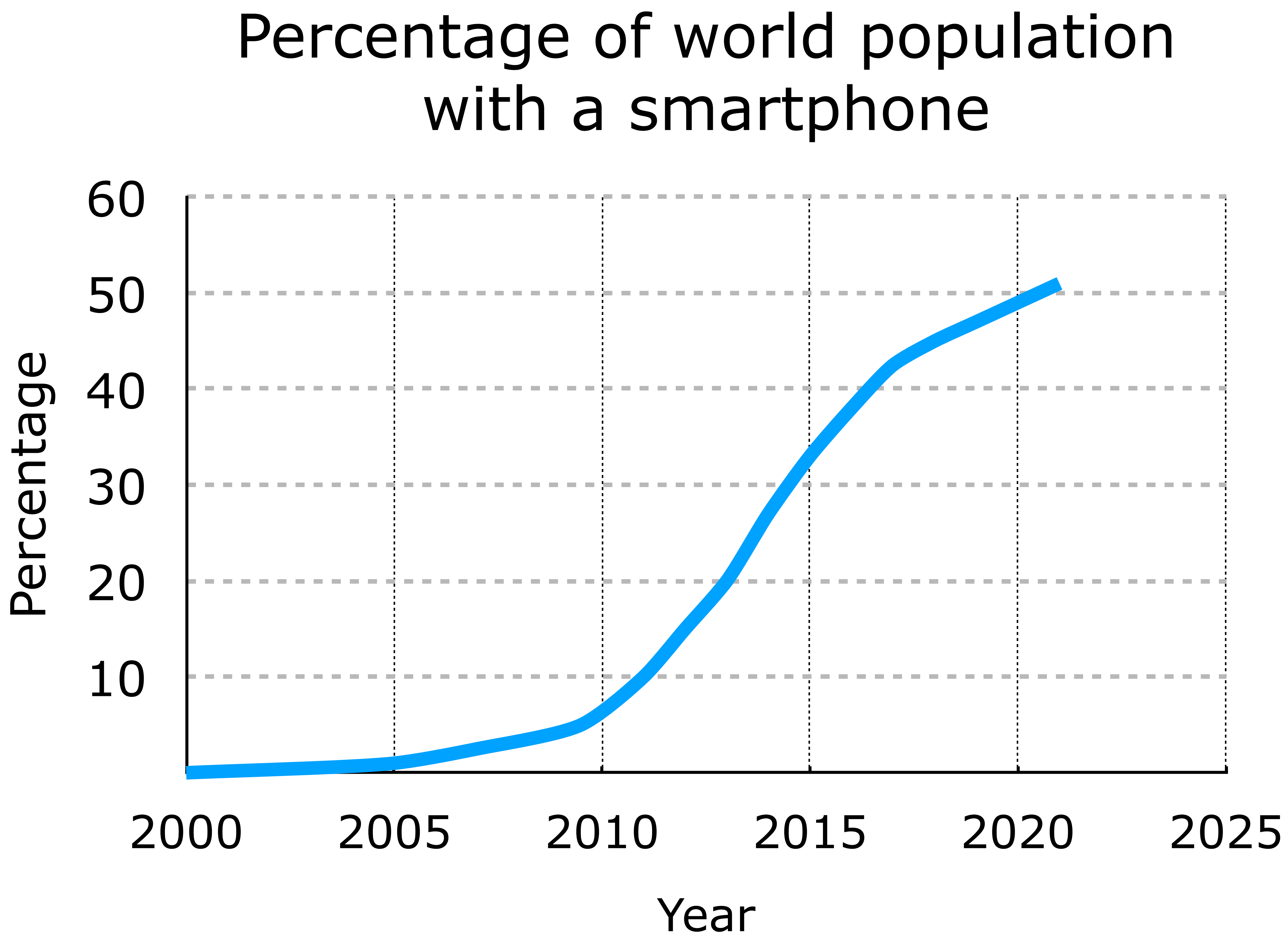



Data Credit: Business Wire

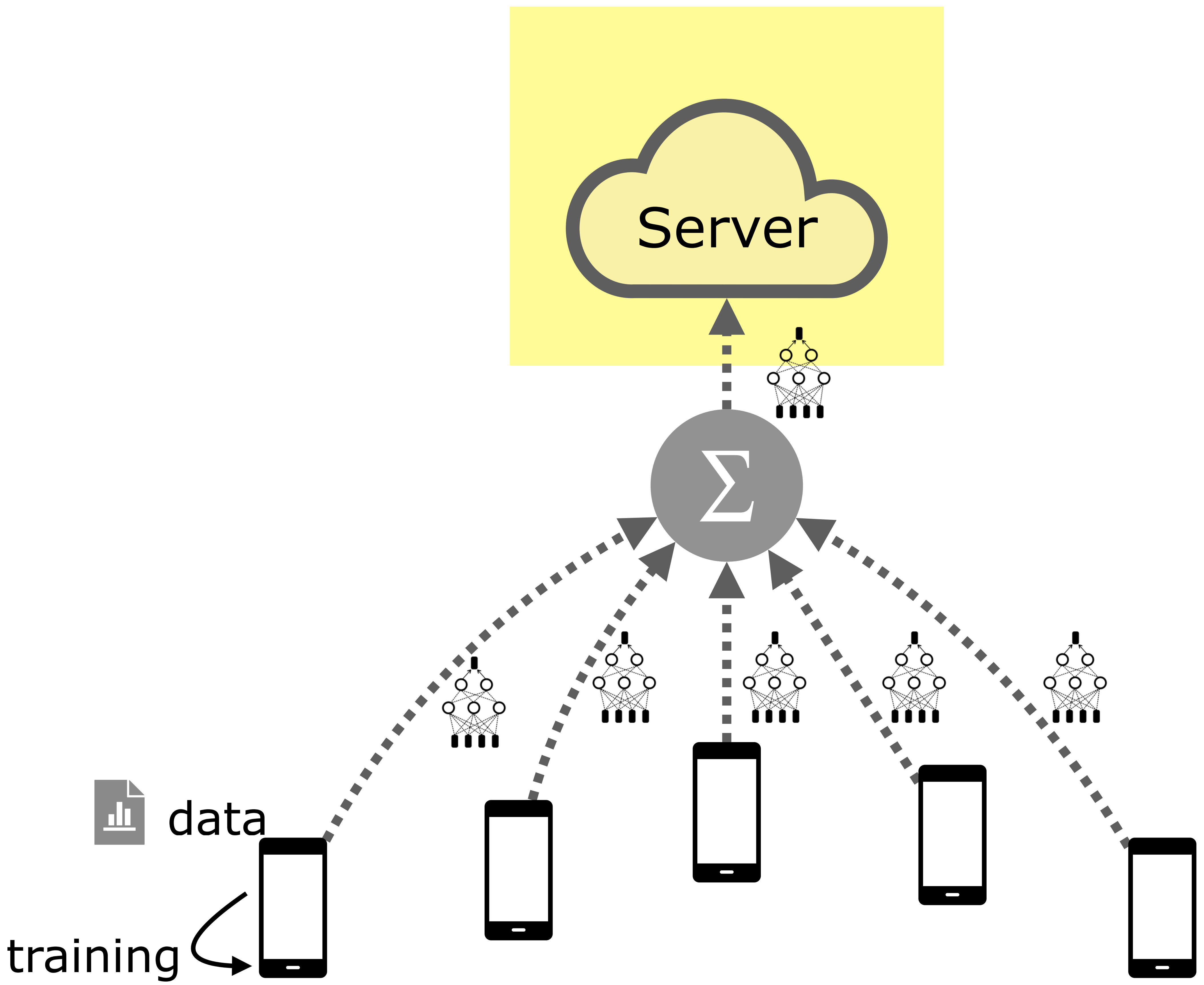
# Federated Learning

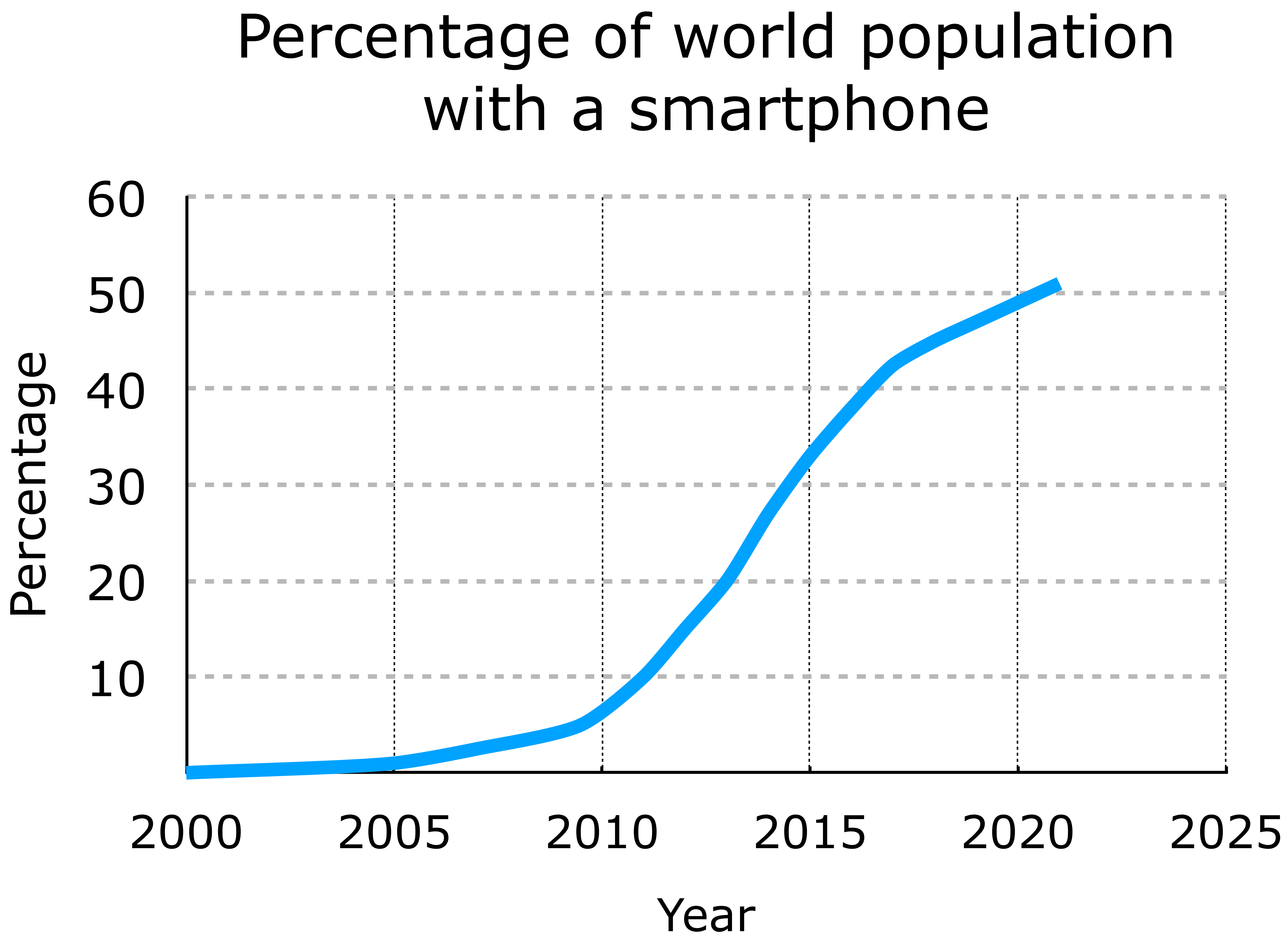



Data Credit: Business Wire

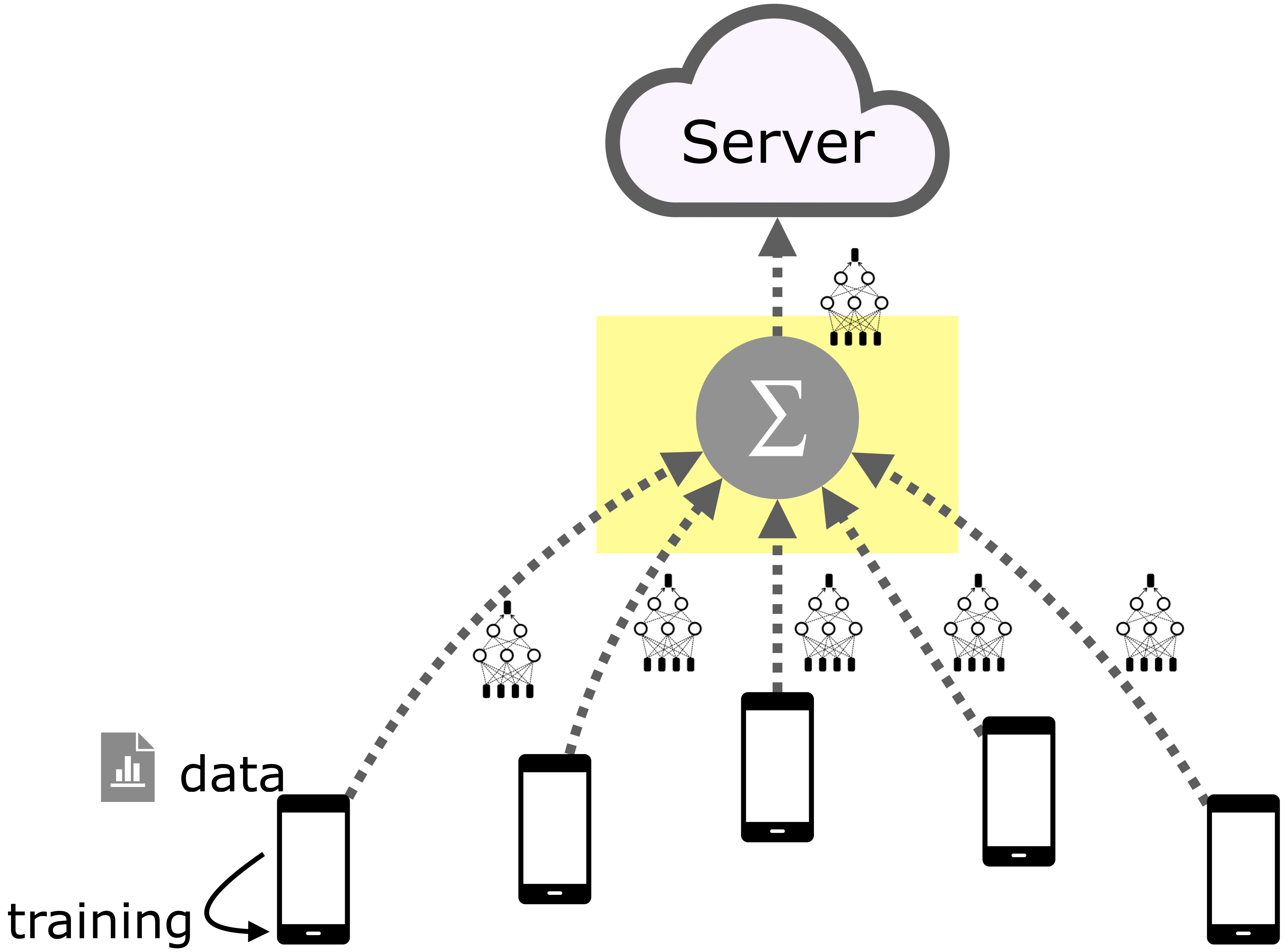
# Federated Learning

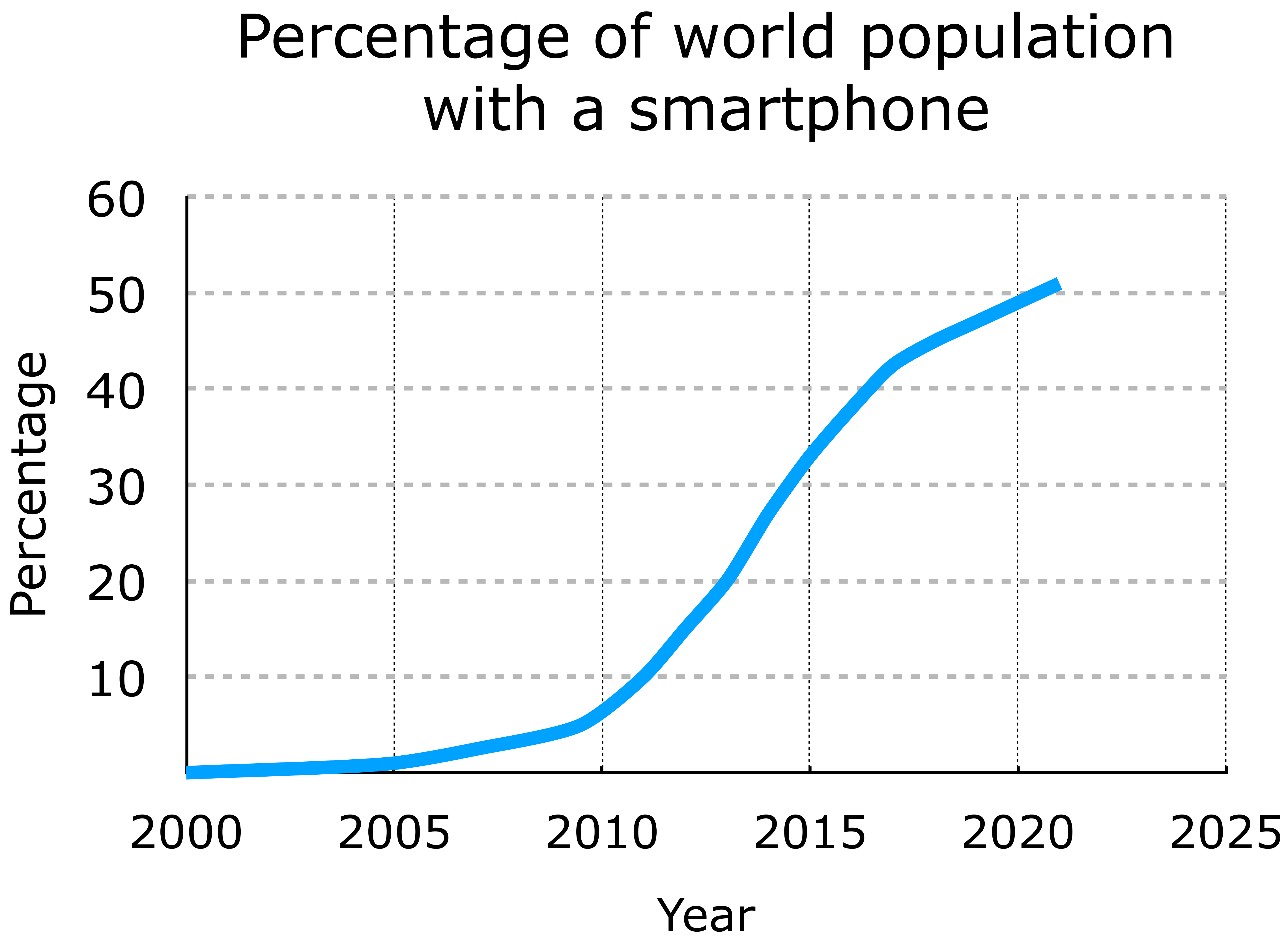



Data Credit: Business Wire

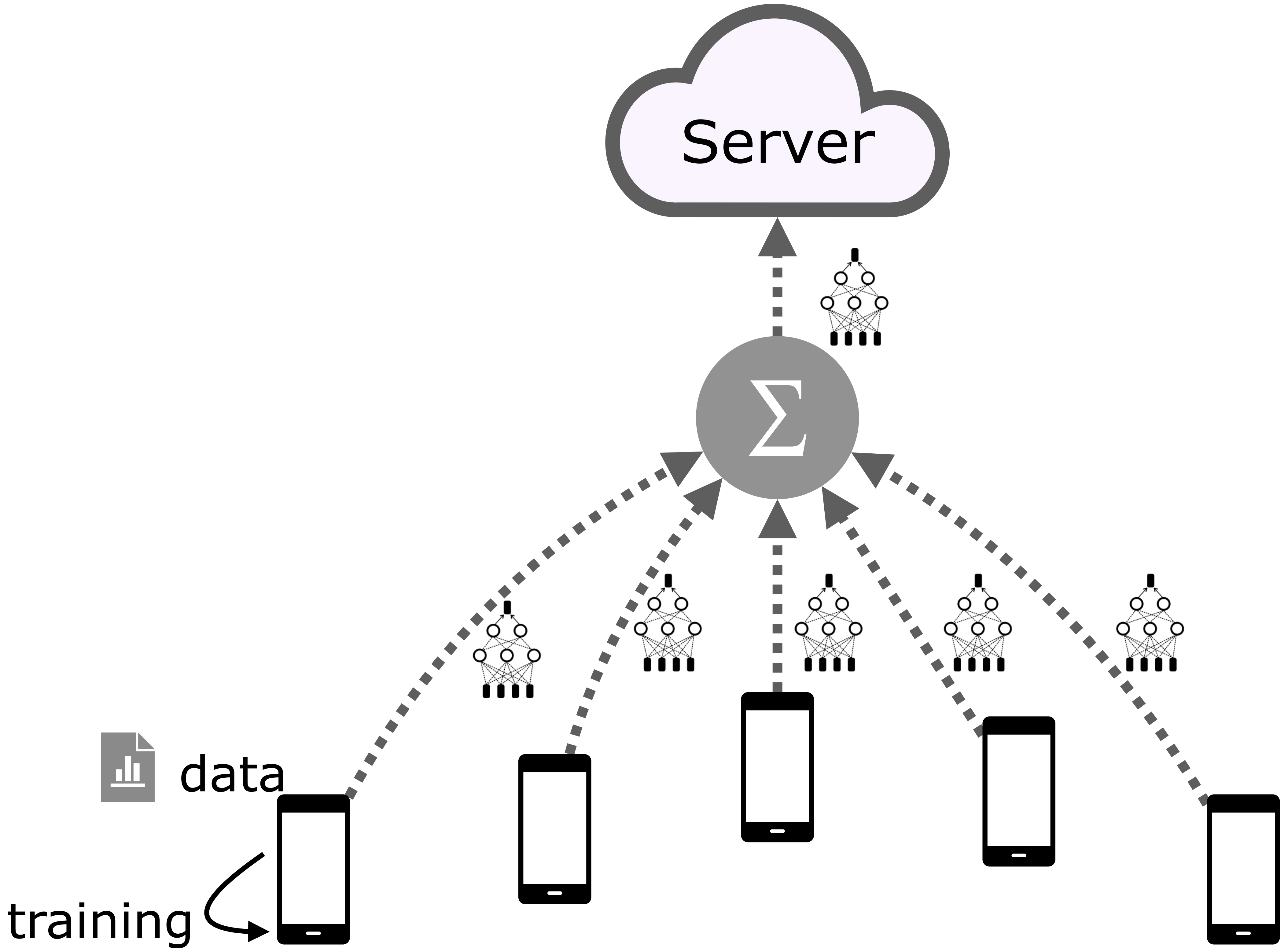
# Federated Learning



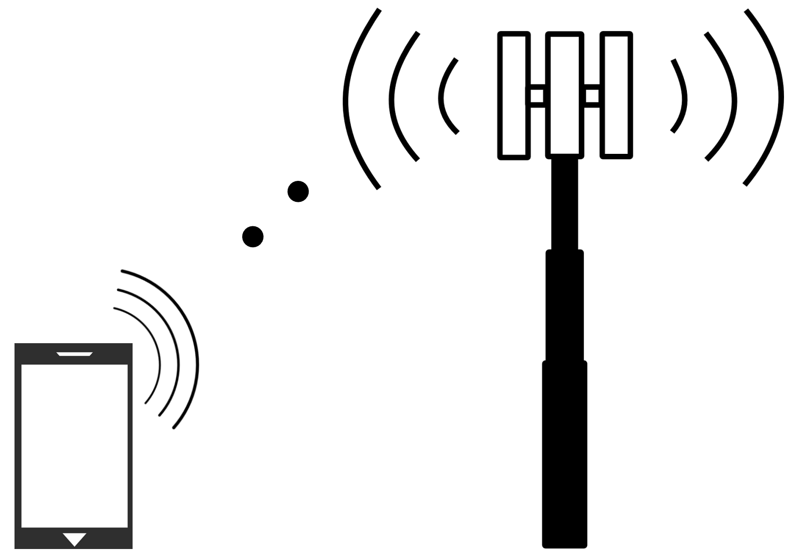




Data Credit: Business Wire

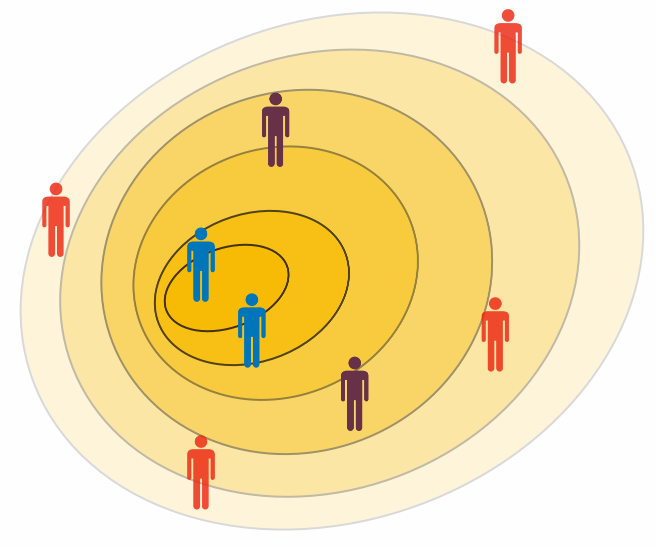
# Federated Learning



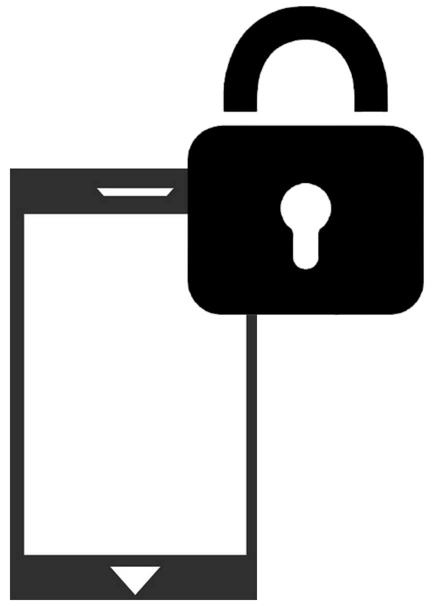

**Challenges:**



Communication efficiency

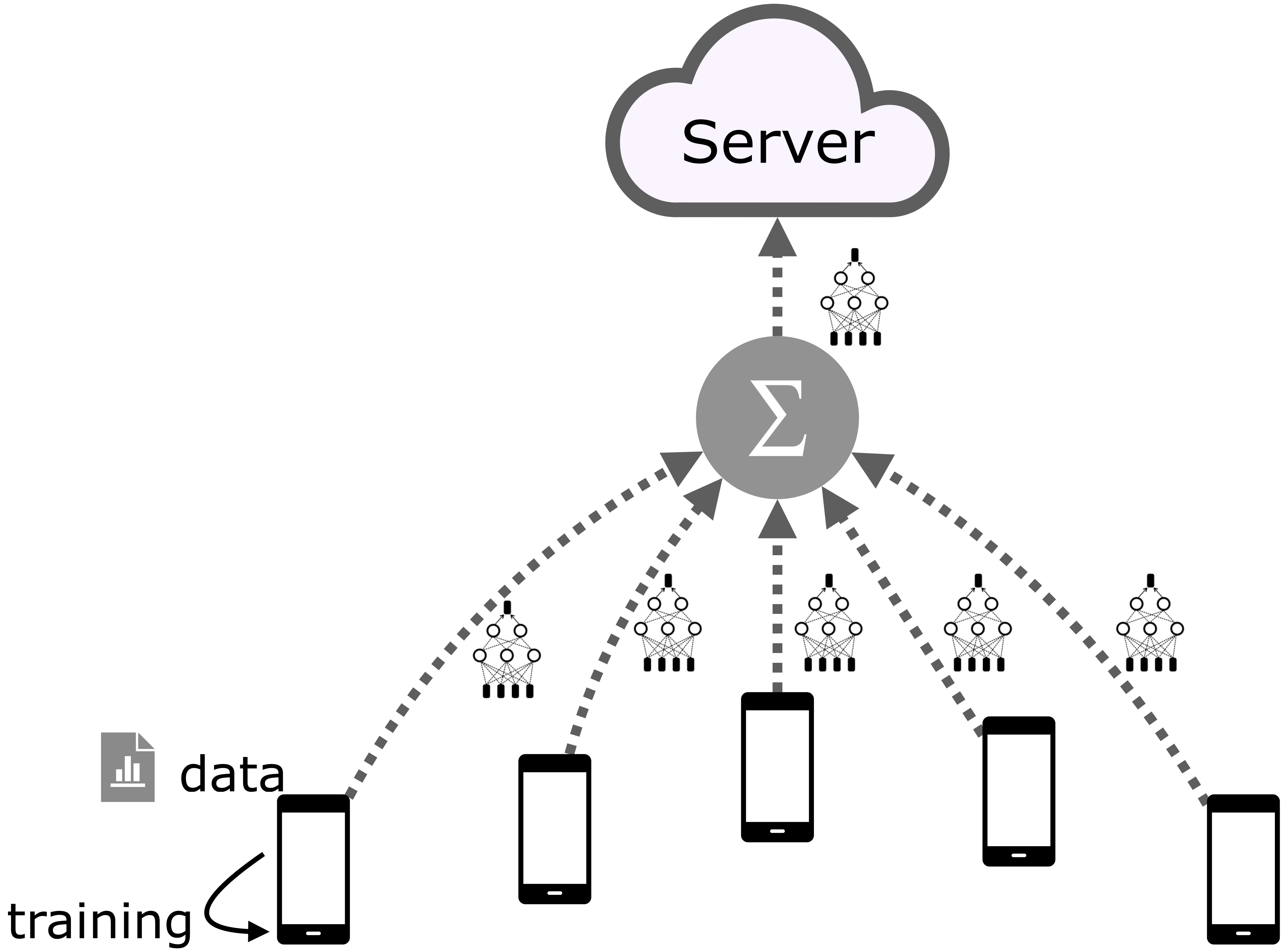


Statistical heterogeneity

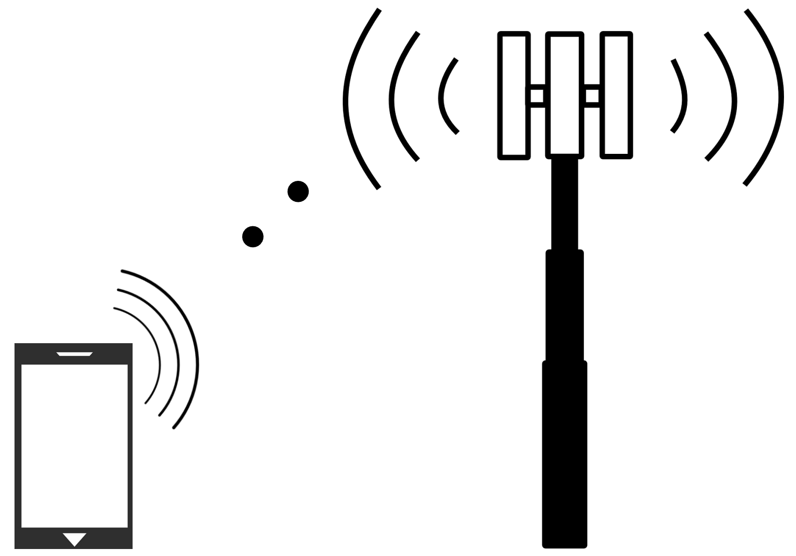


Privacy of user data

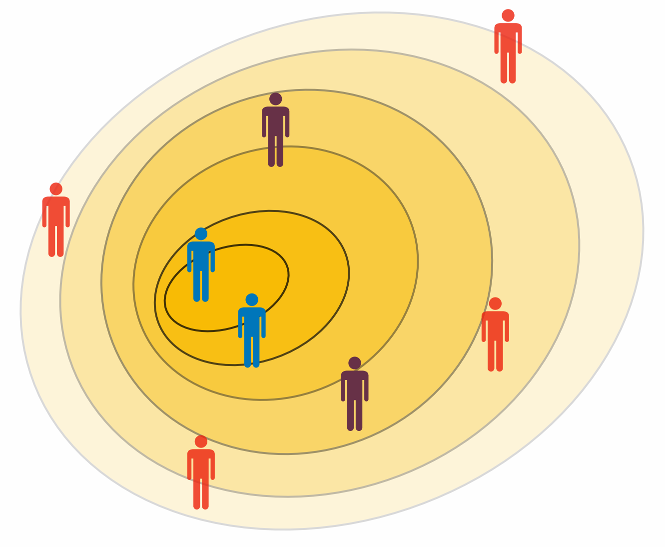
# Federated Learning



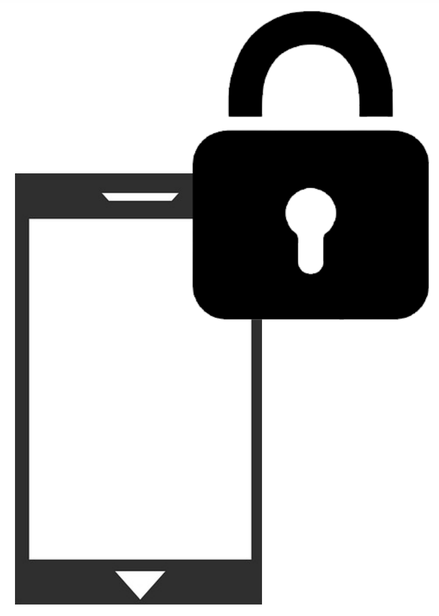

**Challenges:**



Communication efficiency



**Statistical heterogeneity**

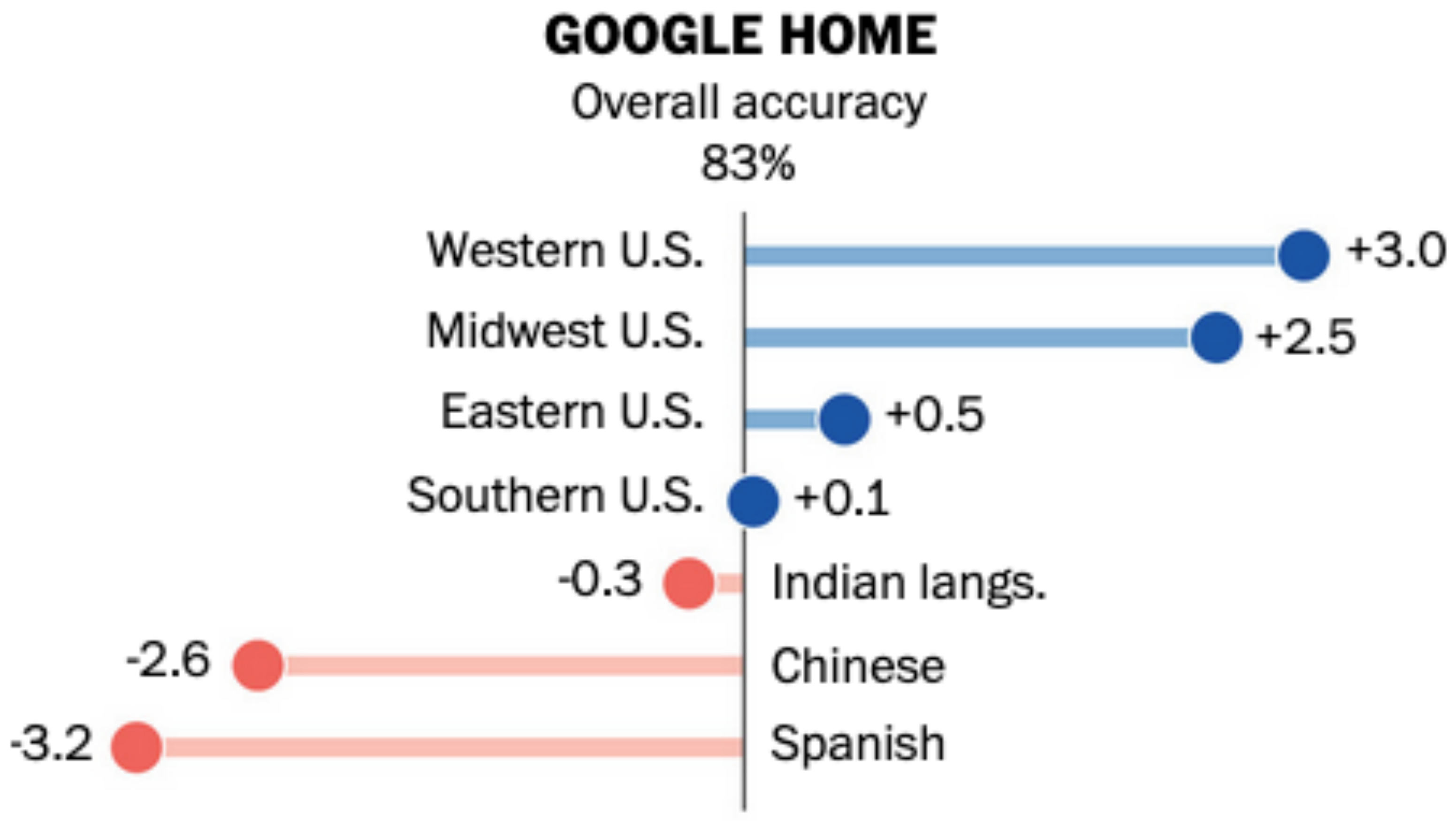


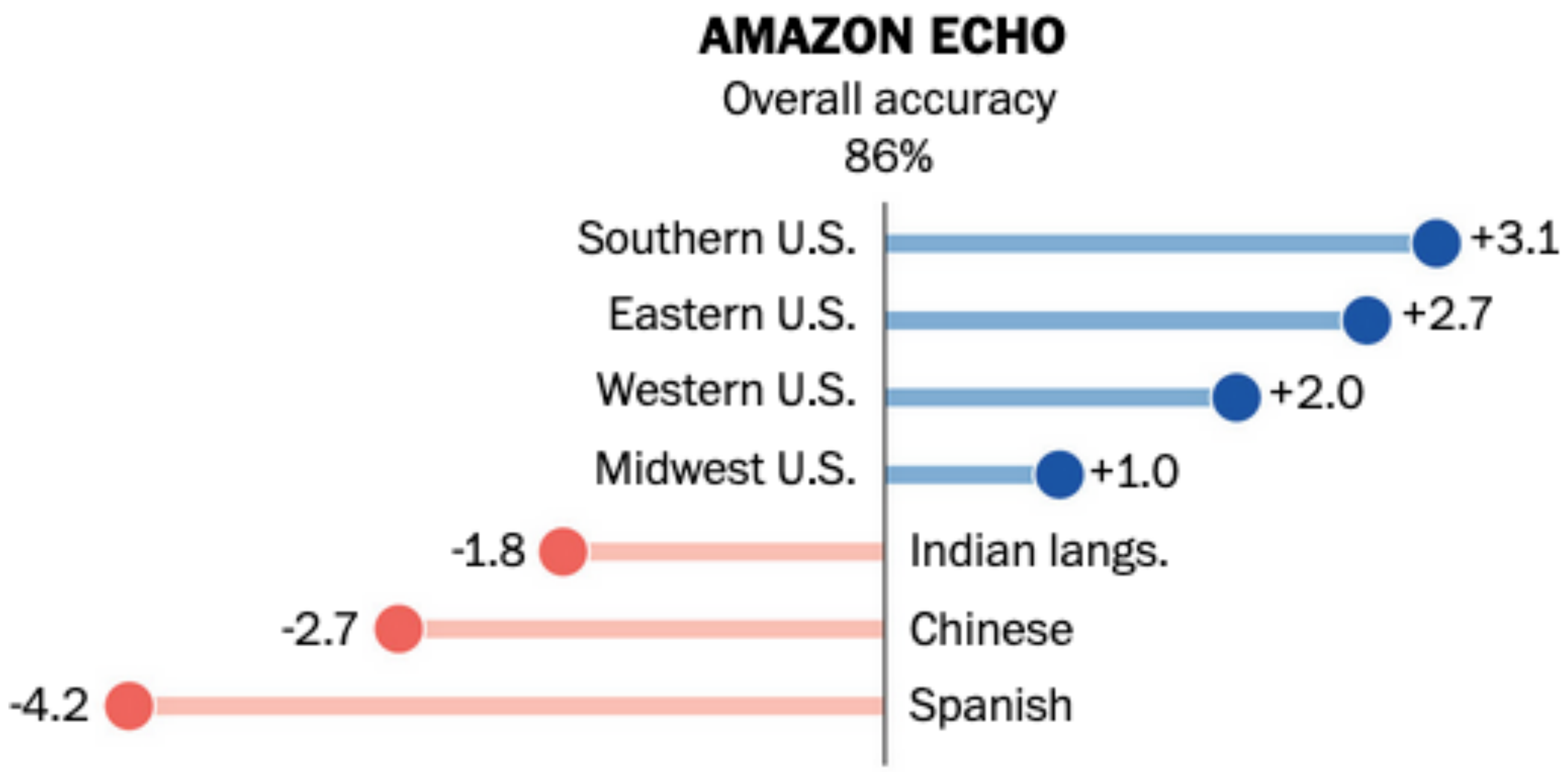
Privacy of user data

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

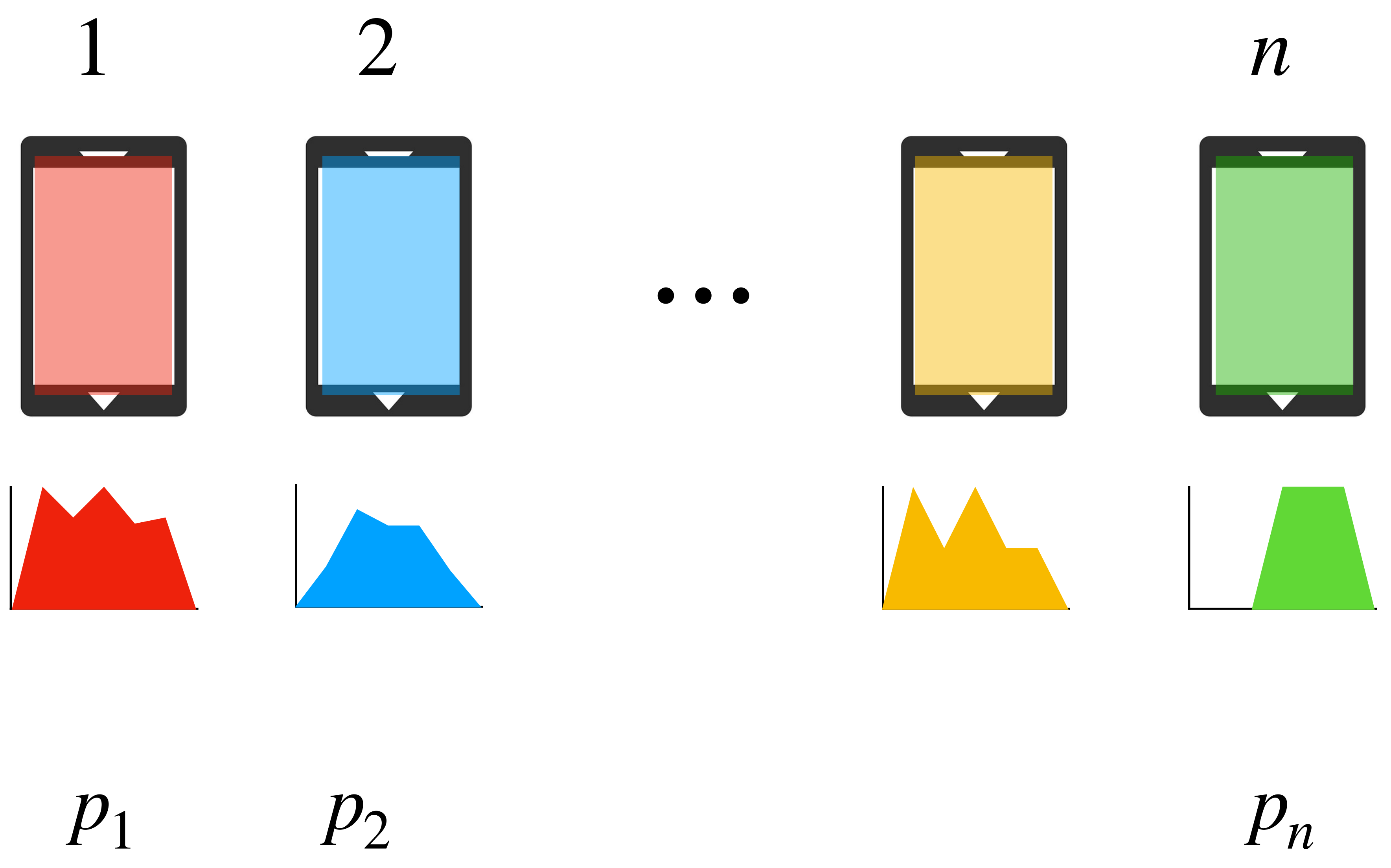
# Tackling distribution shifts in federated learning

- **Improving tail performance with a single model**
- Improving overall performance with local adaptation

# Problem Setup

Clients

Data Distribution




*Data heterogeneity*

Usual Learning Objective

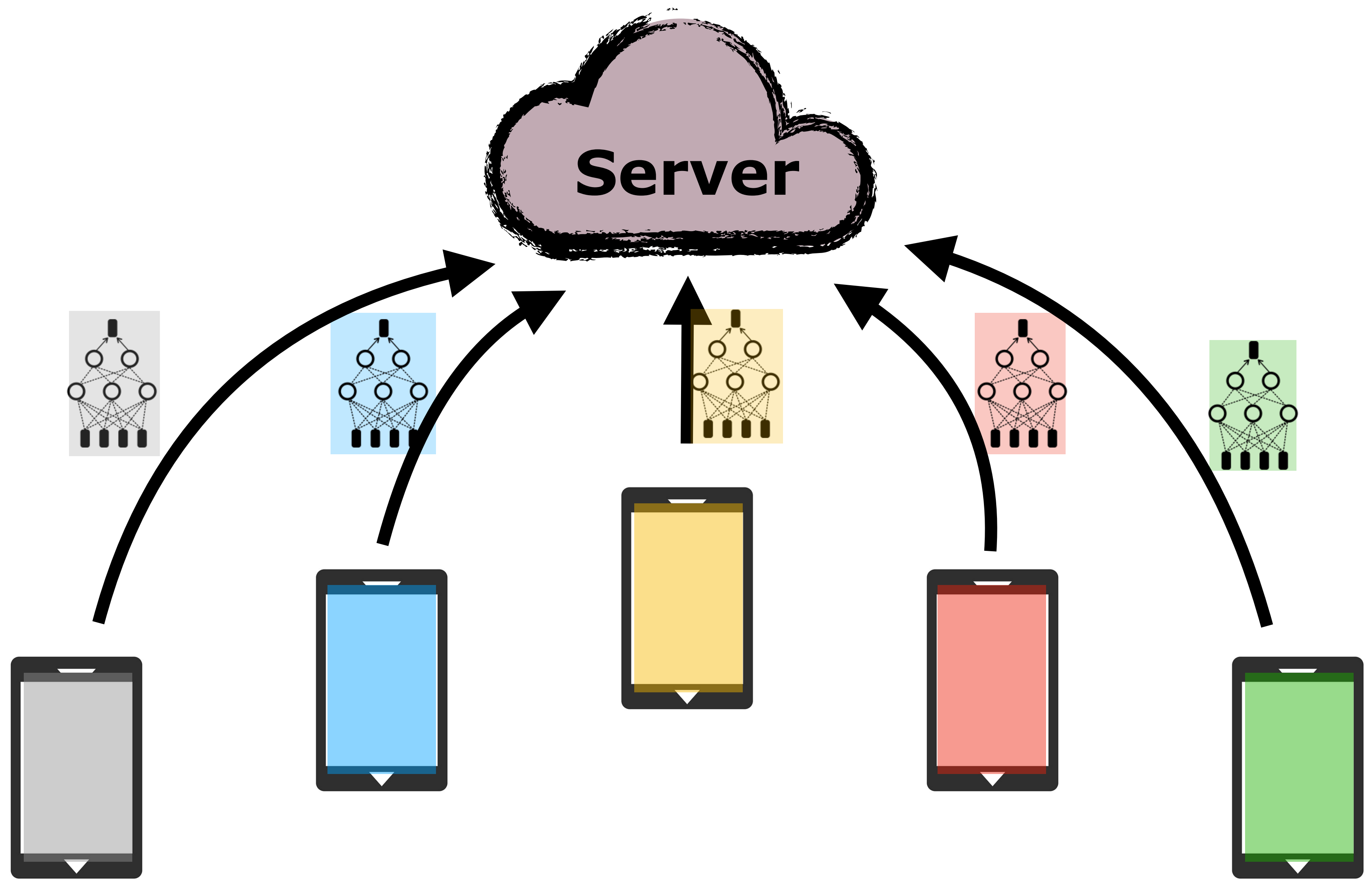
$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} F_i(w) \quad \text{where} \quad F_i(w) = \mathbb{E}_{z \sim p_i} \left[ f(w; z) \right]$$

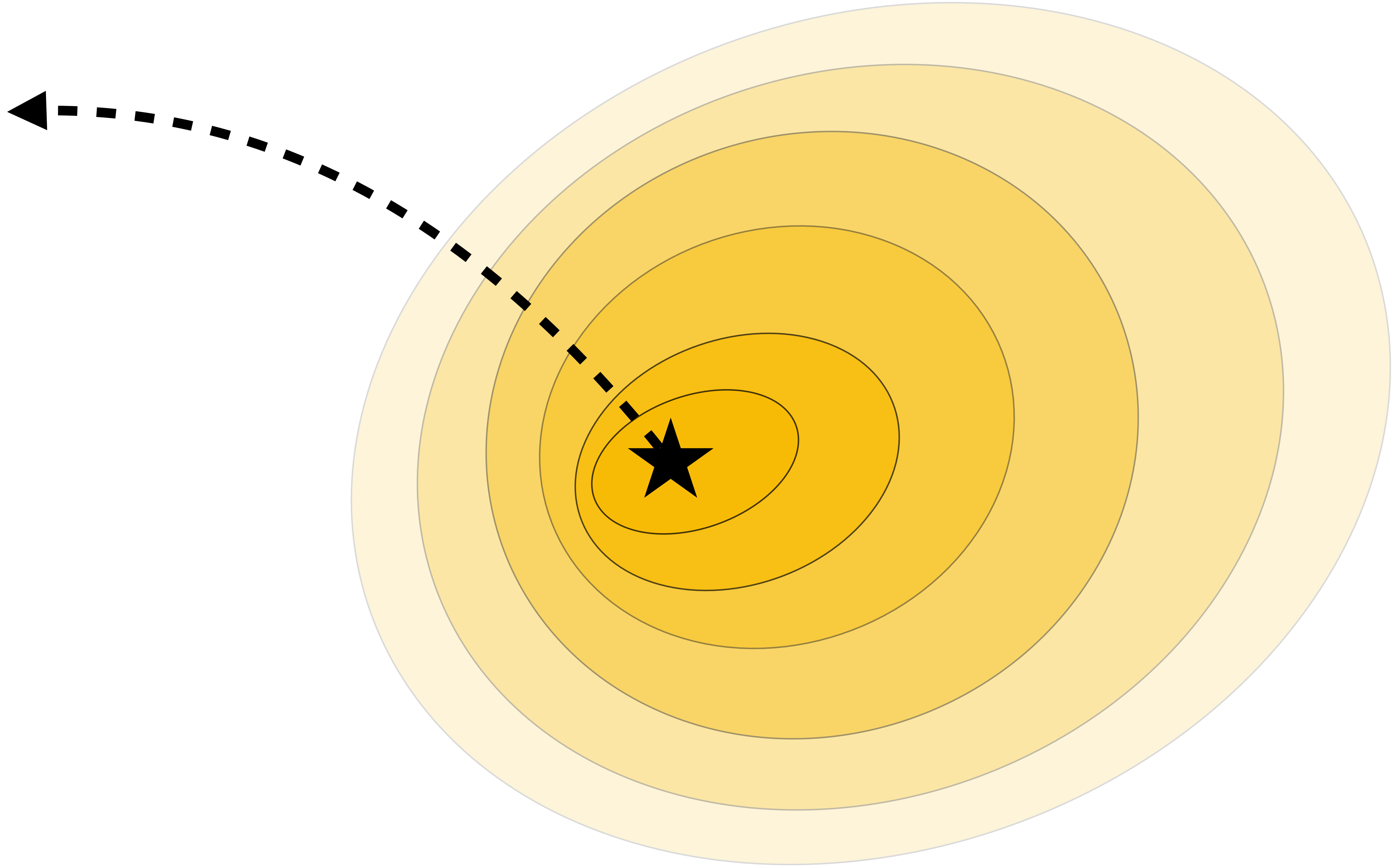
loss on client $i$

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

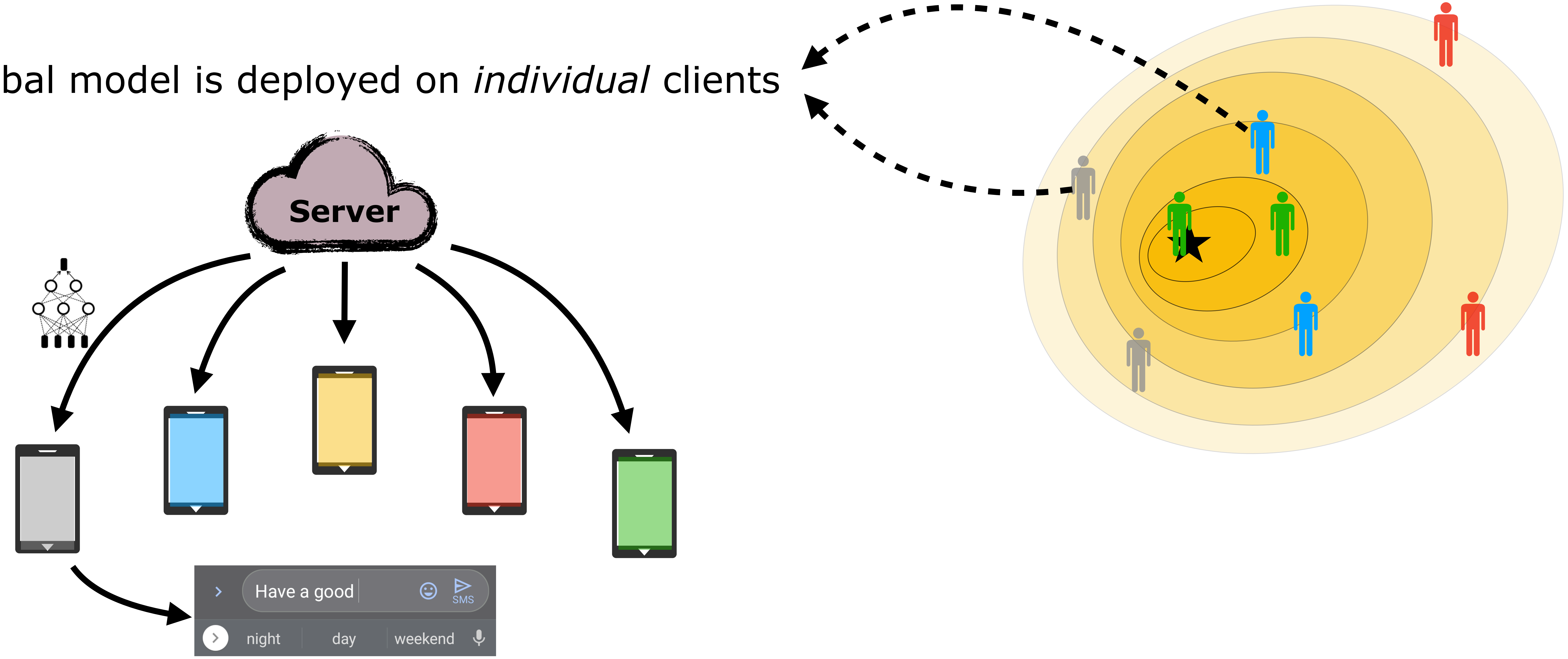
$$\min_{w \in \mathbb{R}^d} \quad \frac{1}{n} \sum_{i=1}^{n} F_i(w)$$

Global model is trained on *average distribution* across clients

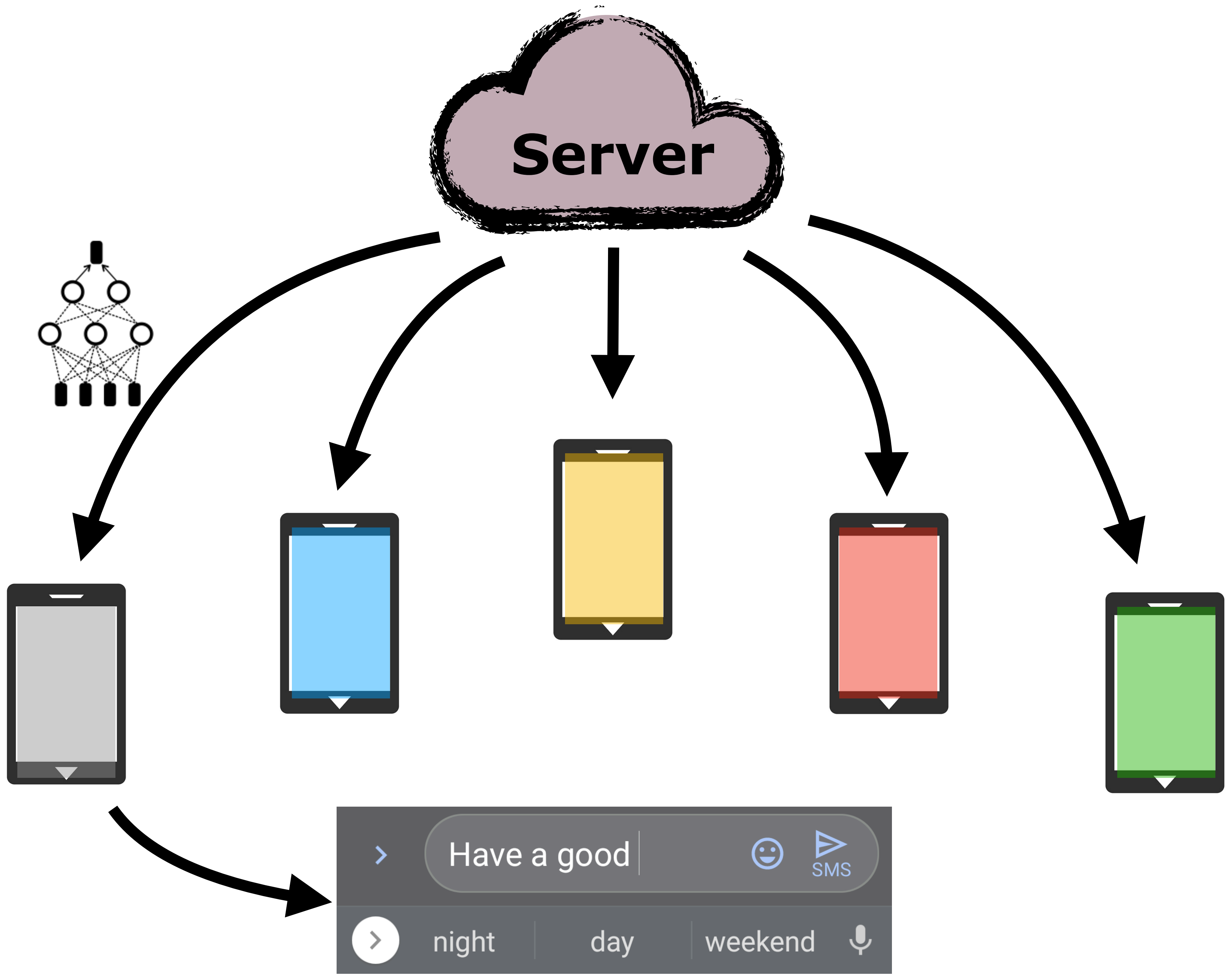


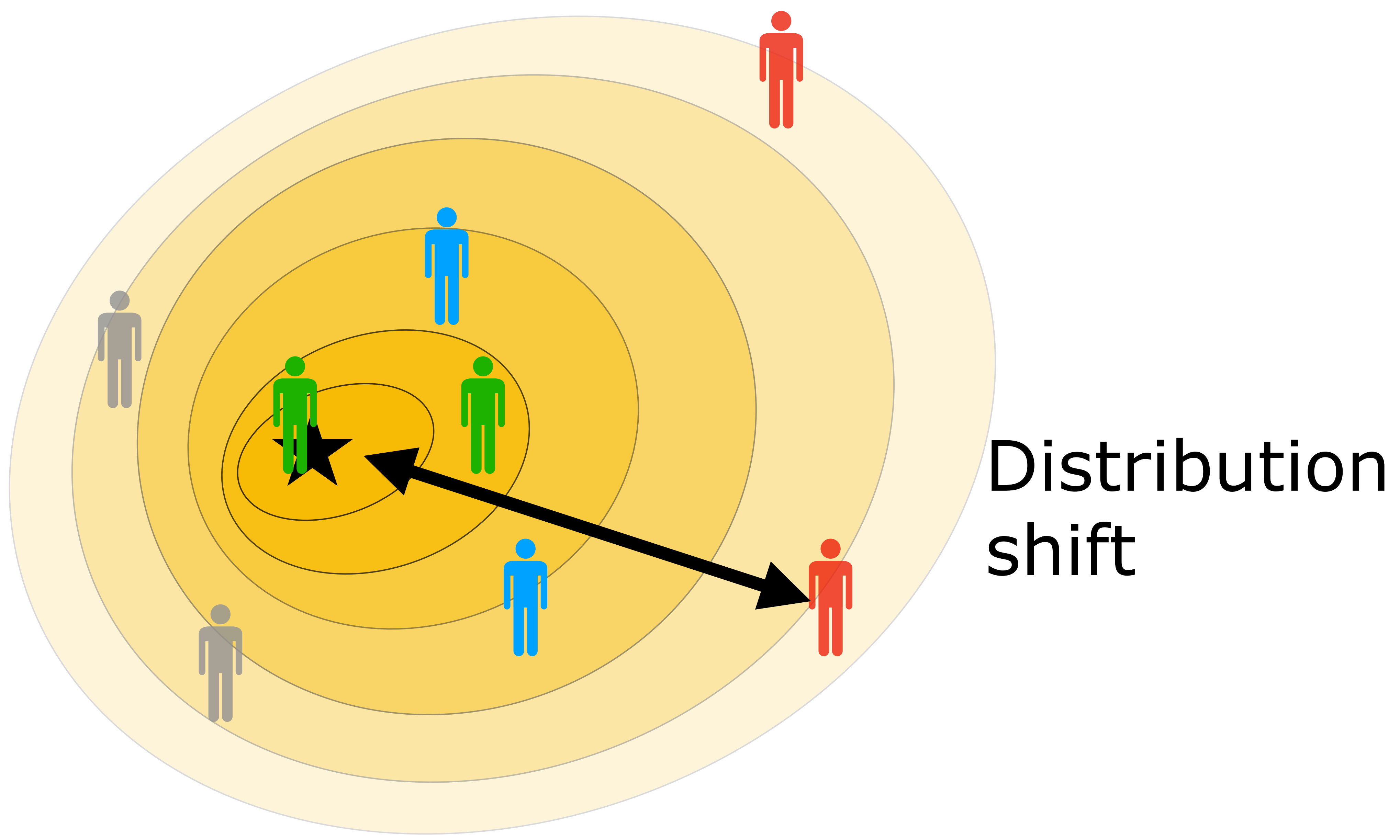



Global model is deployed on *individual* clients

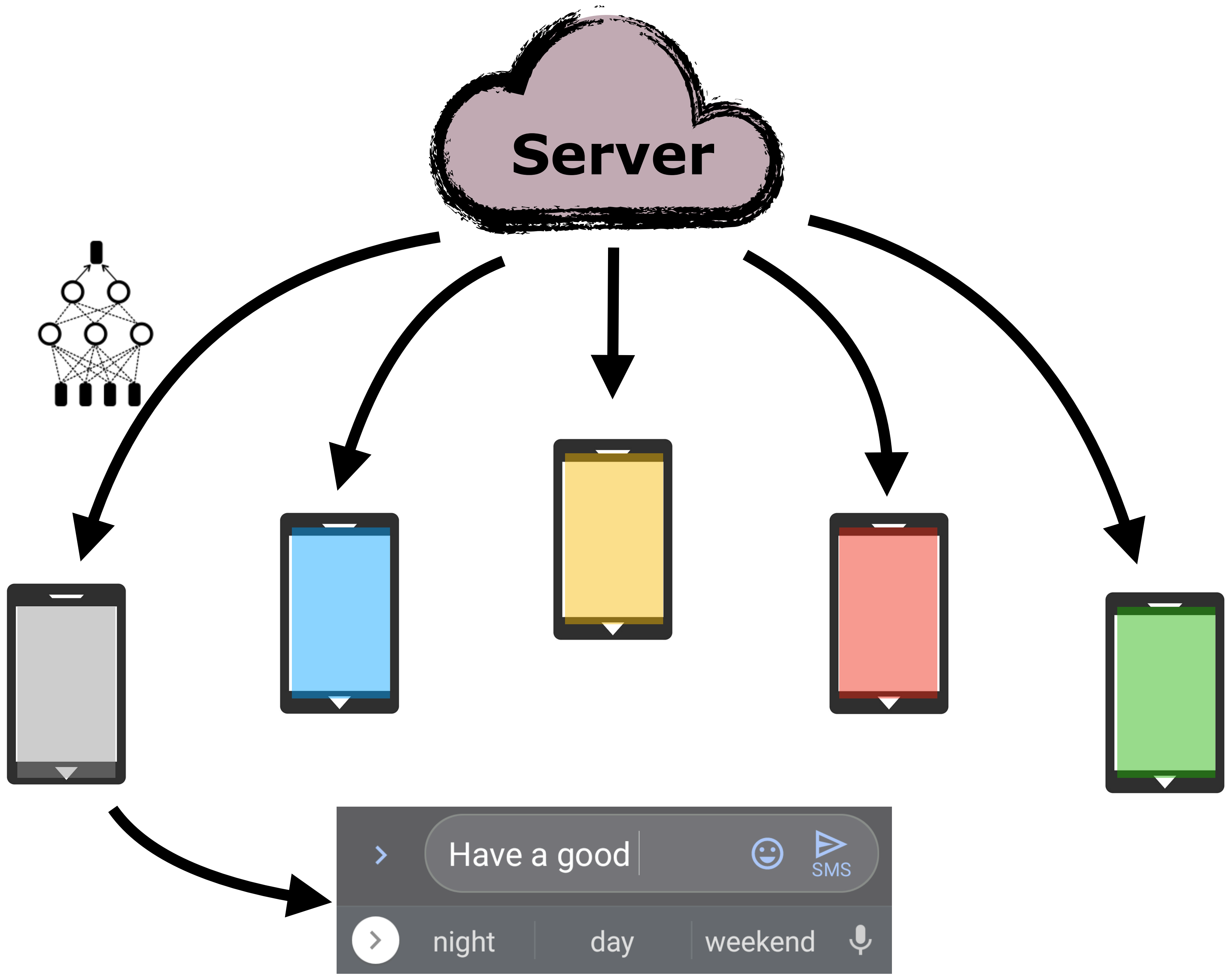


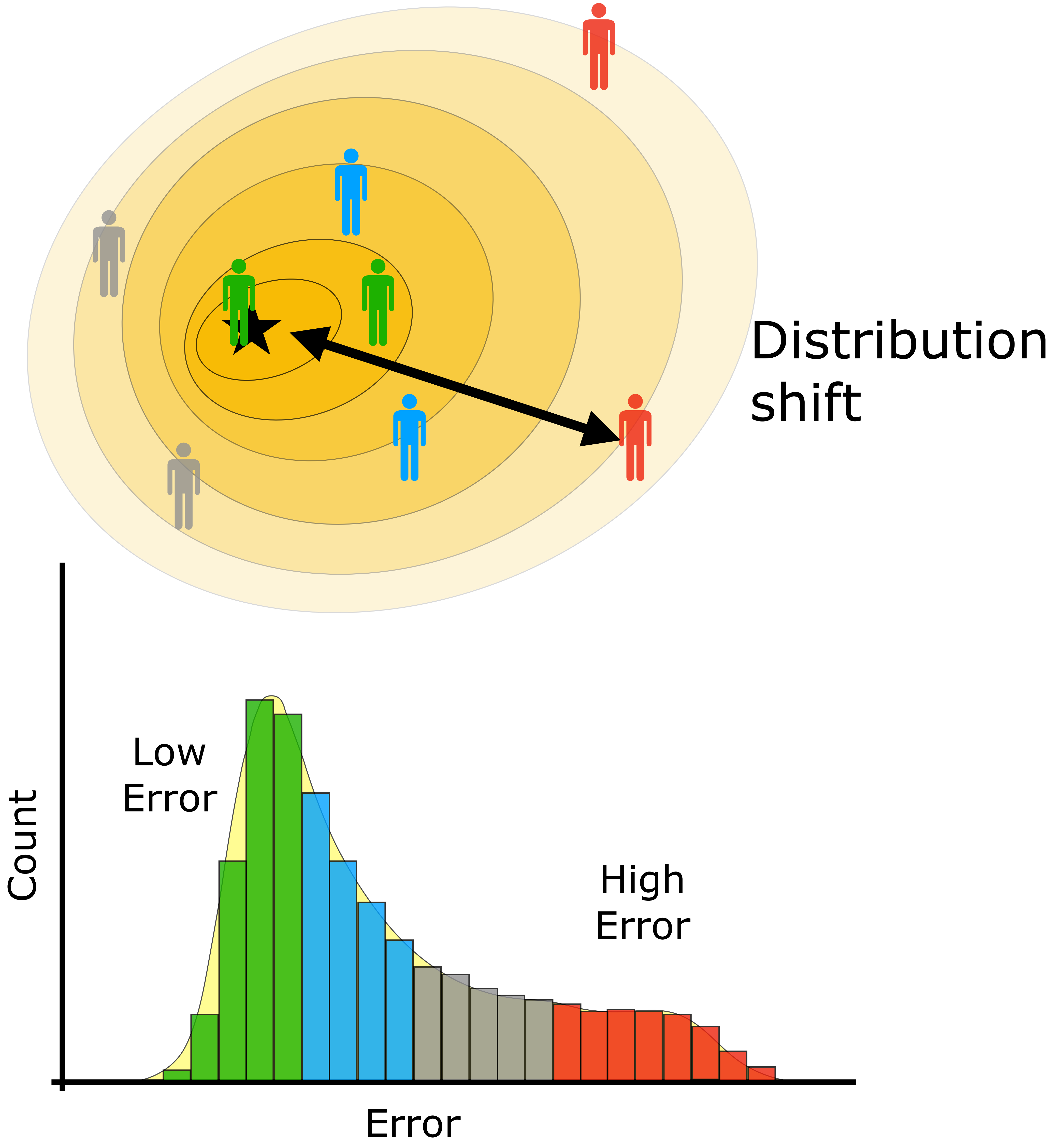
Global model is deployed on *individual* clients



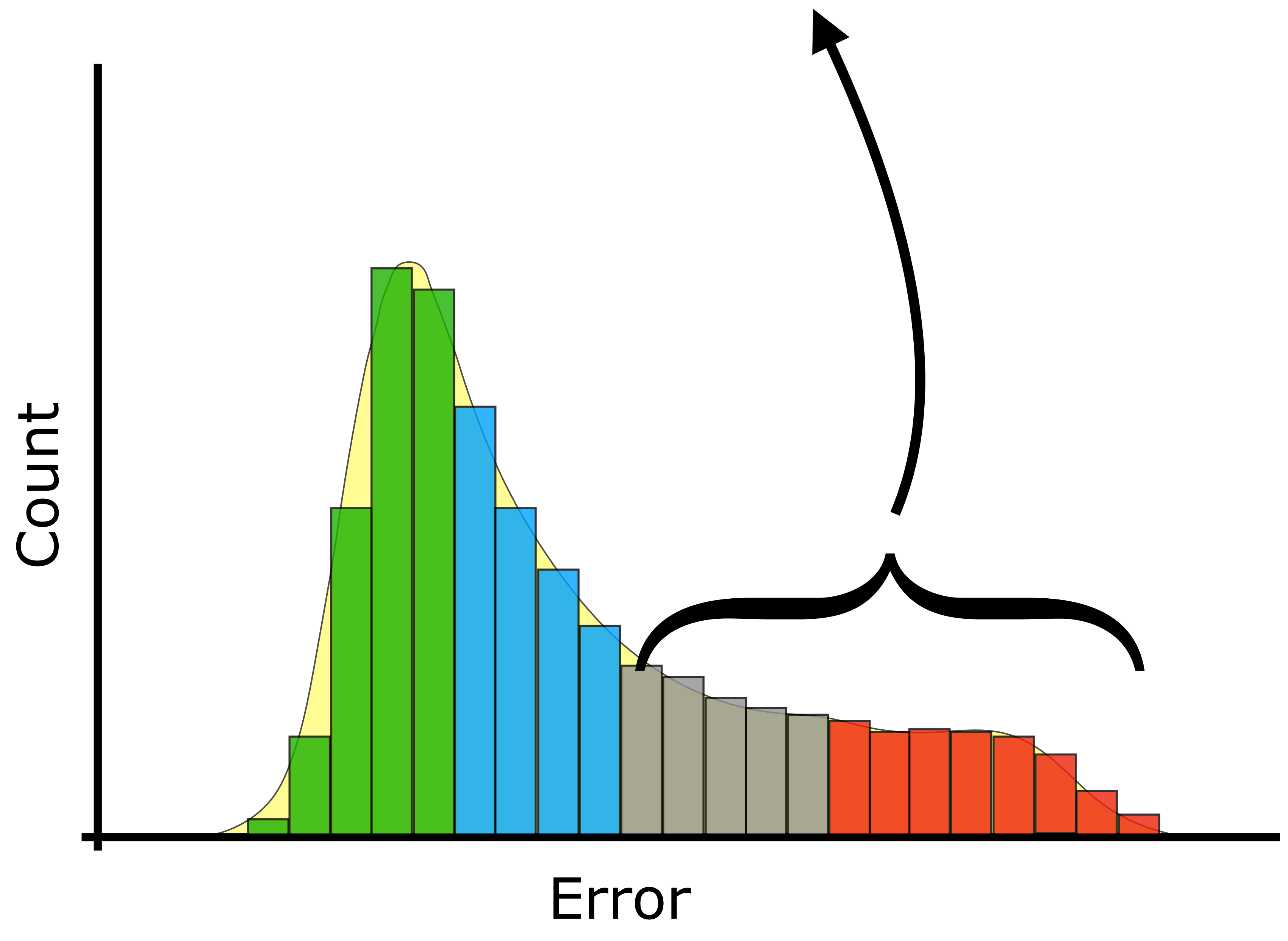



Global model is deployed on *individual* clients



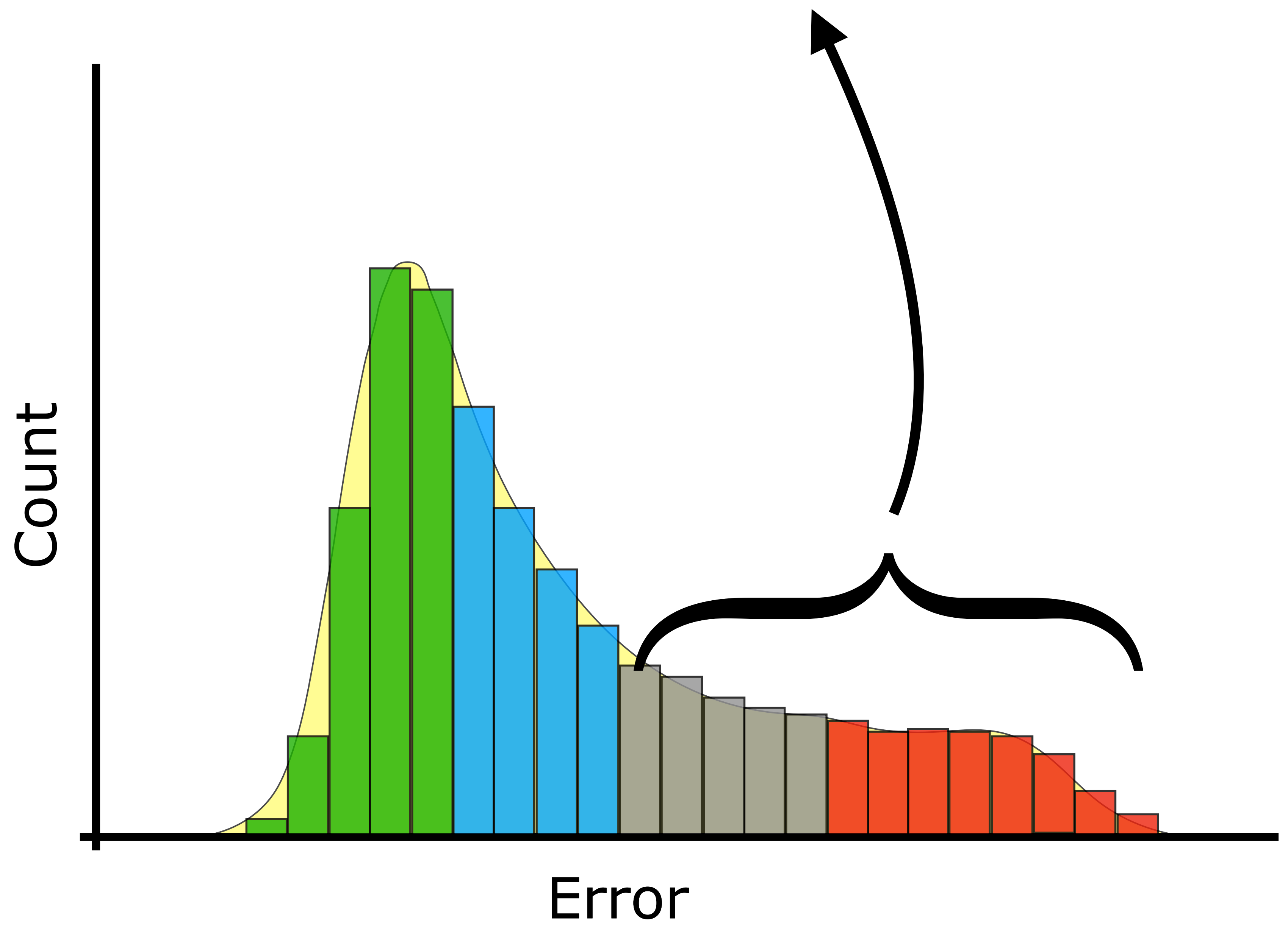



**Our goal**: improve performance on “tail clients”



# Simplicial federated learning

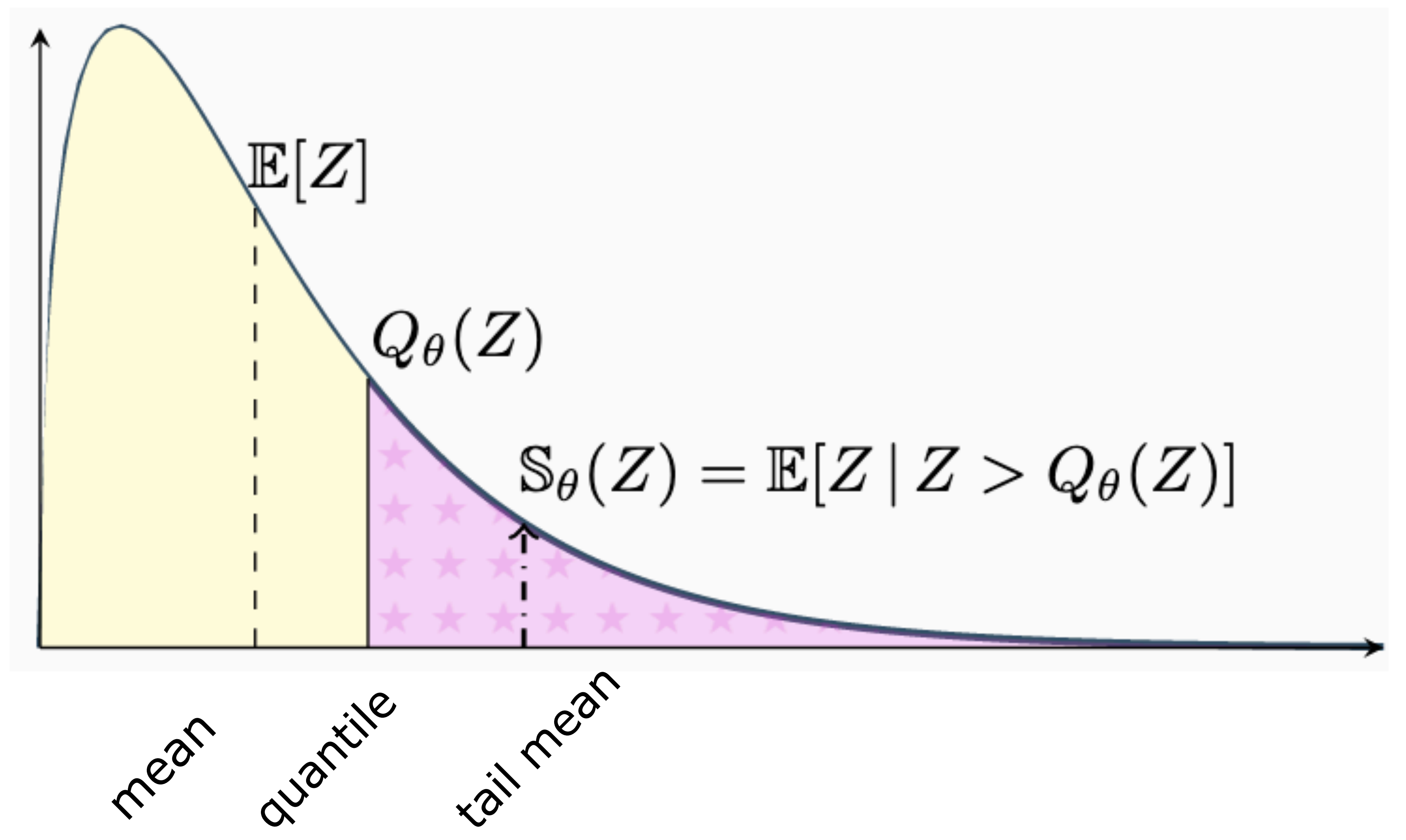
**Our Approach**: minimize the tail error directly!




Simplicial-FL Objective:

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

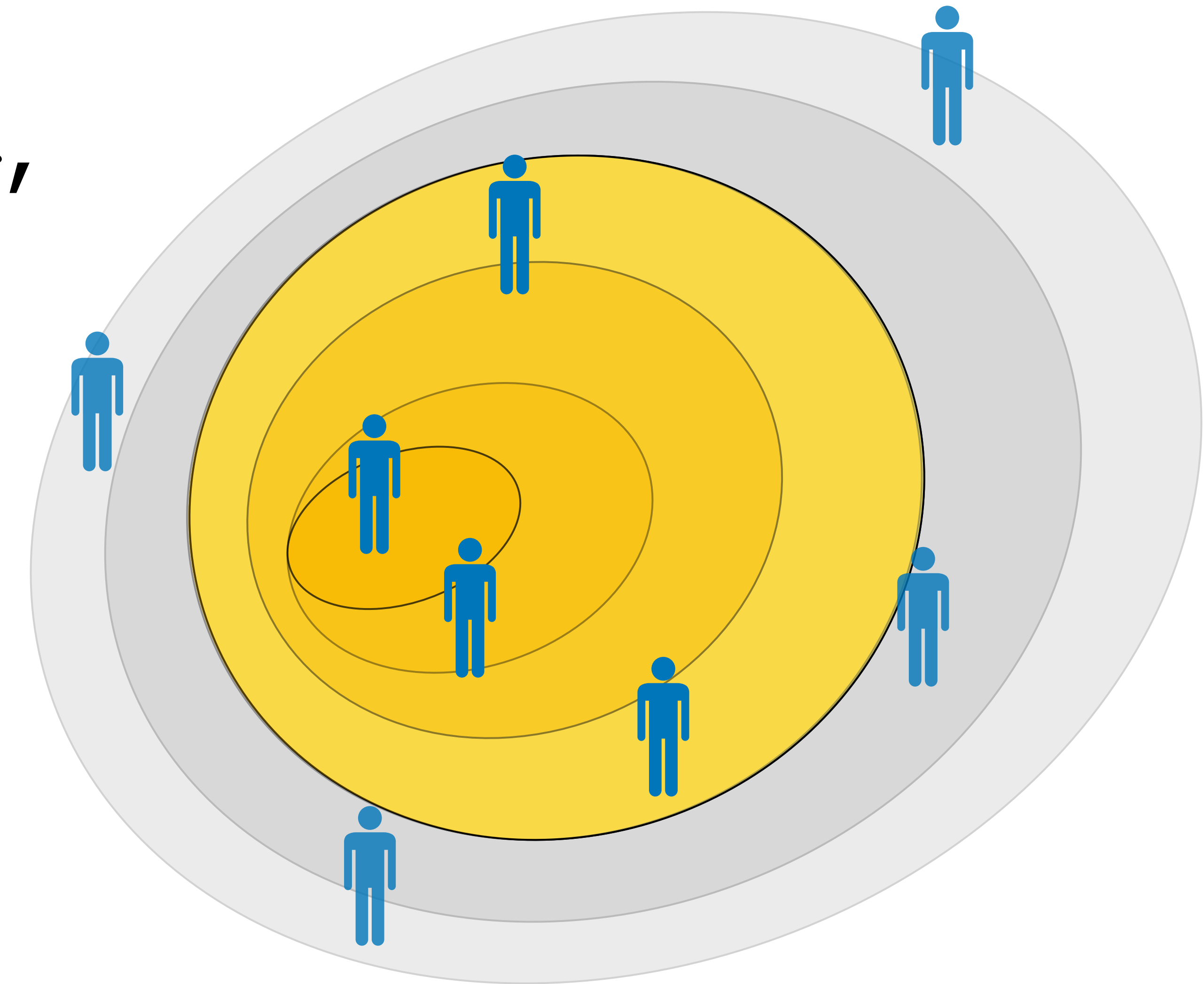
Superquantile | Conditional Value at Risk




[Rockafellar & Uryasev (2000; 2002)]

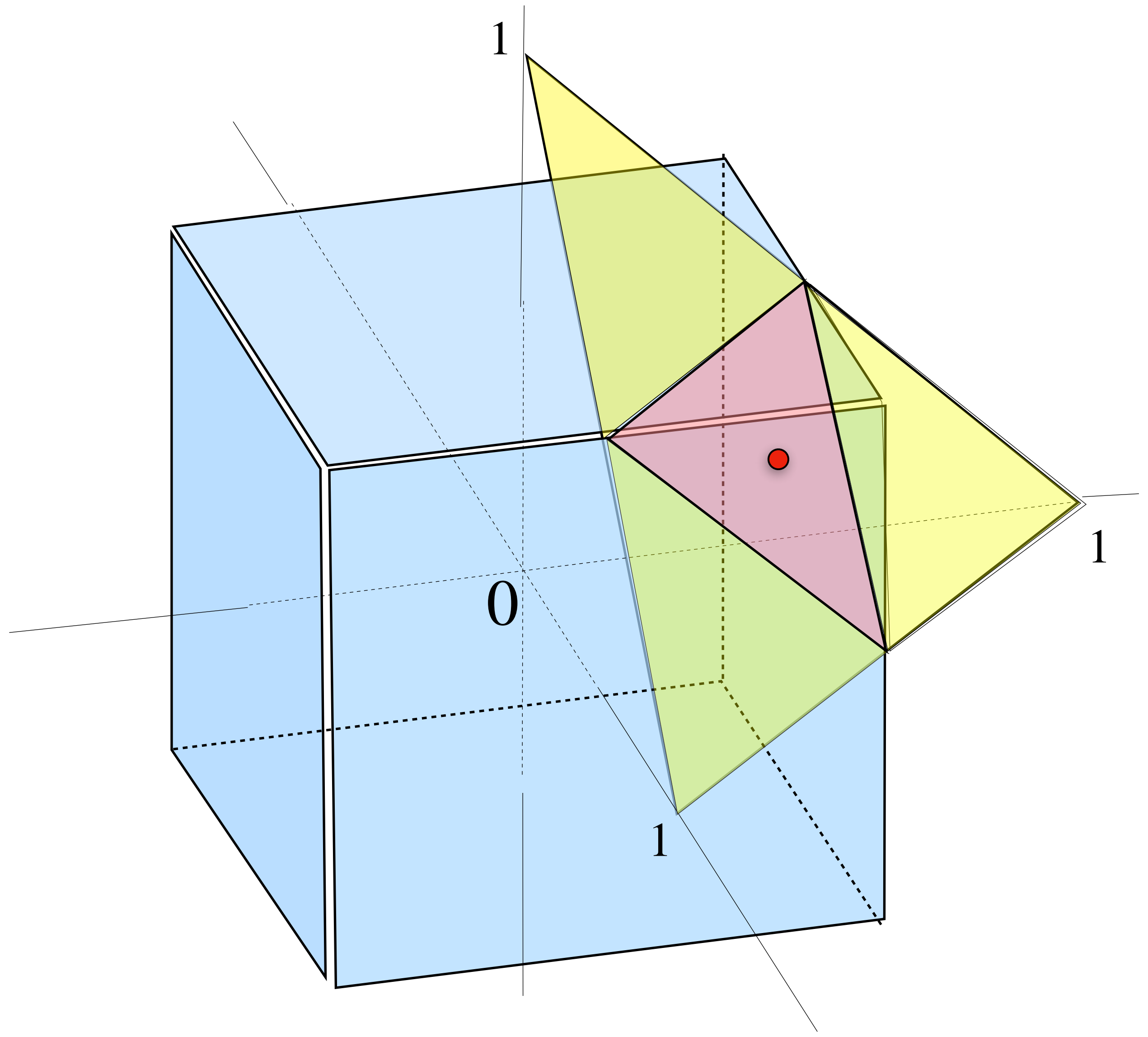
## **Distributional robustness in federated learning**:

Assuming a new test client with mixture distribution $p_\pi = \sum_i \pi_i p_i$,
Simplicial-FL objective is equivalent to:



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

**Dual expression** ≡ continuous knapsack problem



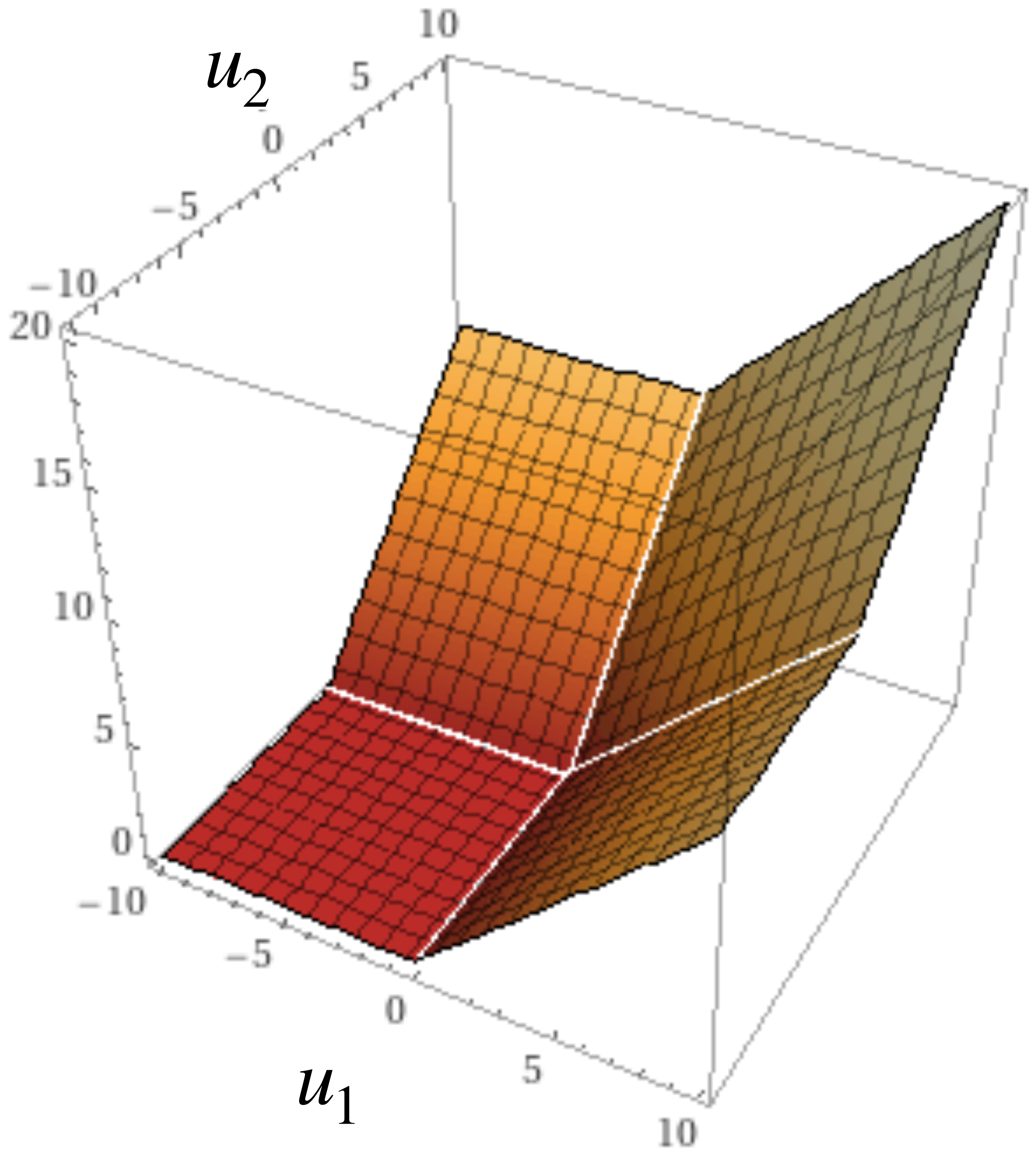

$$\mathbb{S}_\theta(x_1, \cdots, x_n) = \max \left\{ \sum_i \pi_i x_i \; : \; \pi_i \geq 0, \; \sum_i \pi_i = 1, \; \pi_i \leq (n\theta)^{-1} \right\}$$

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

# Optimizing Simplicial-FL

**Challenge**:

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 $\theta n$ is an integer

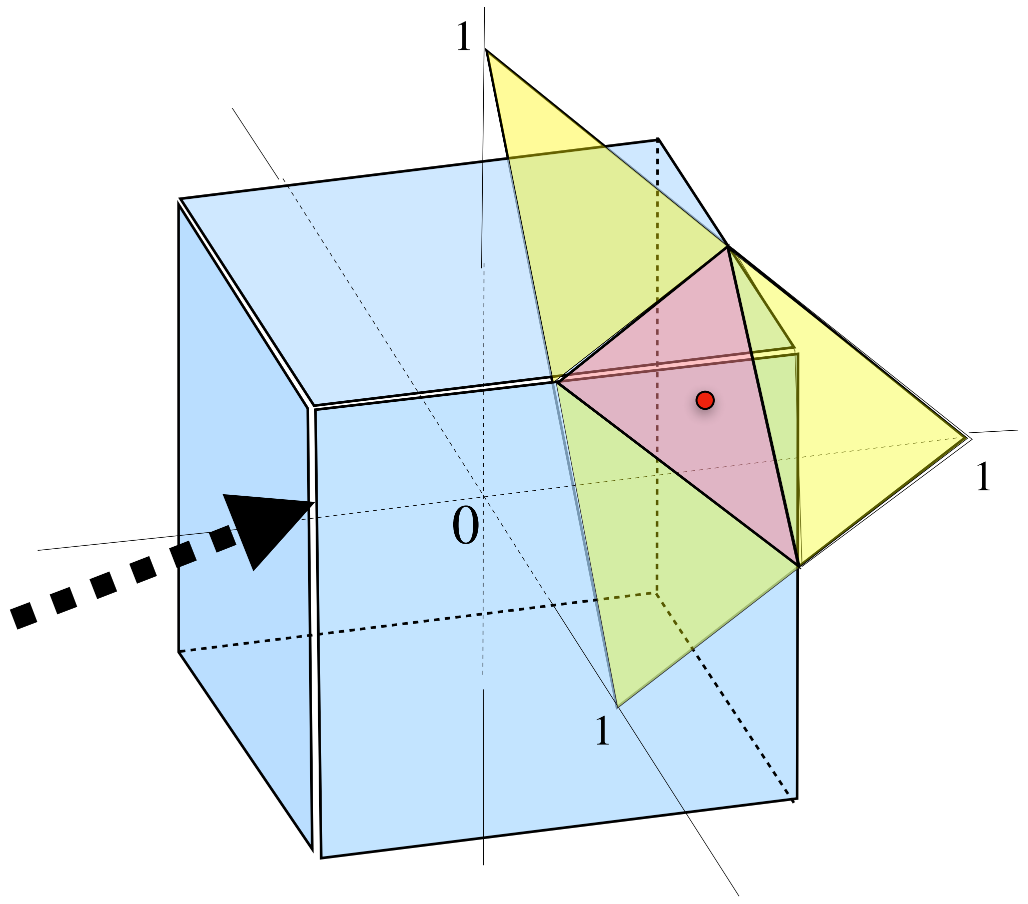
**Nonsmooth**: The subdifferential has a tractable form

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

assuming $\theta n$ is an integer

**Proof** Chain rule $\Longrightarrow$ subdifferential holds with

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



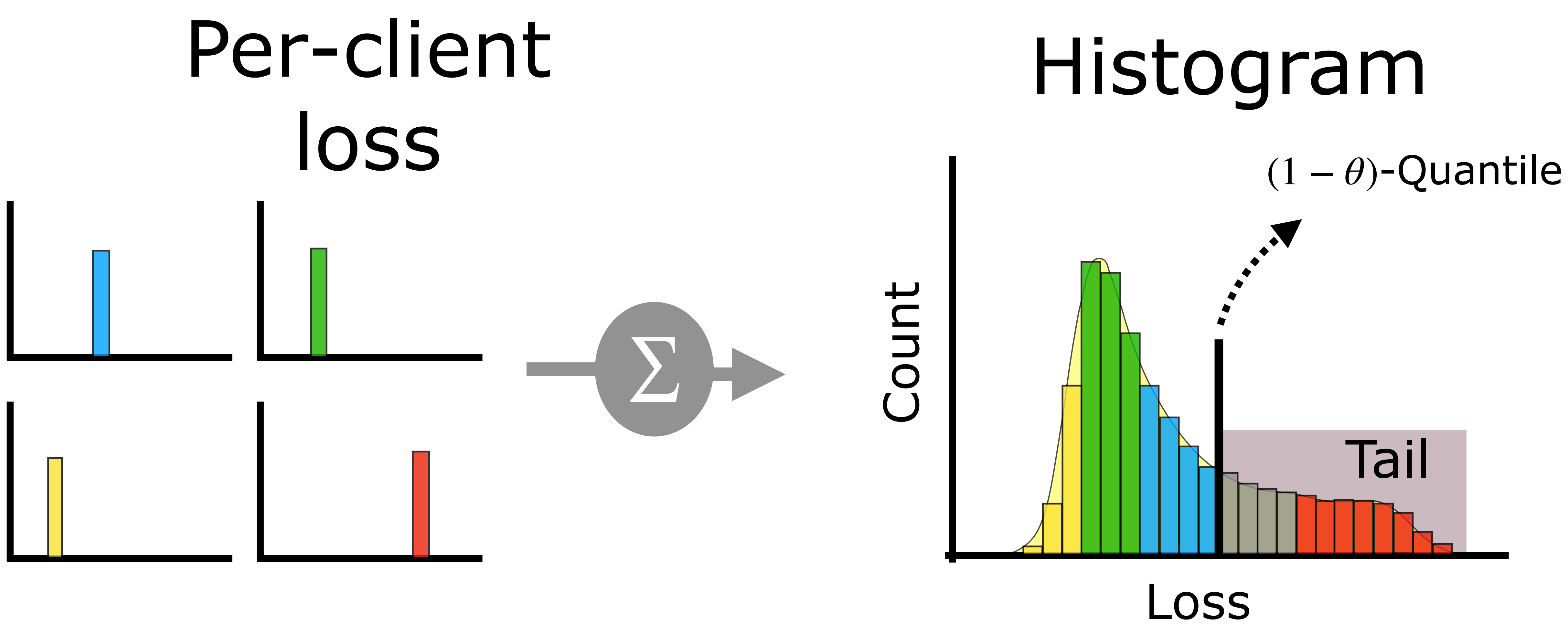

Alternate form of $\pi^\star$ comes from the continuous knapsack problem

[Dantzig. ORIJ (1957)]

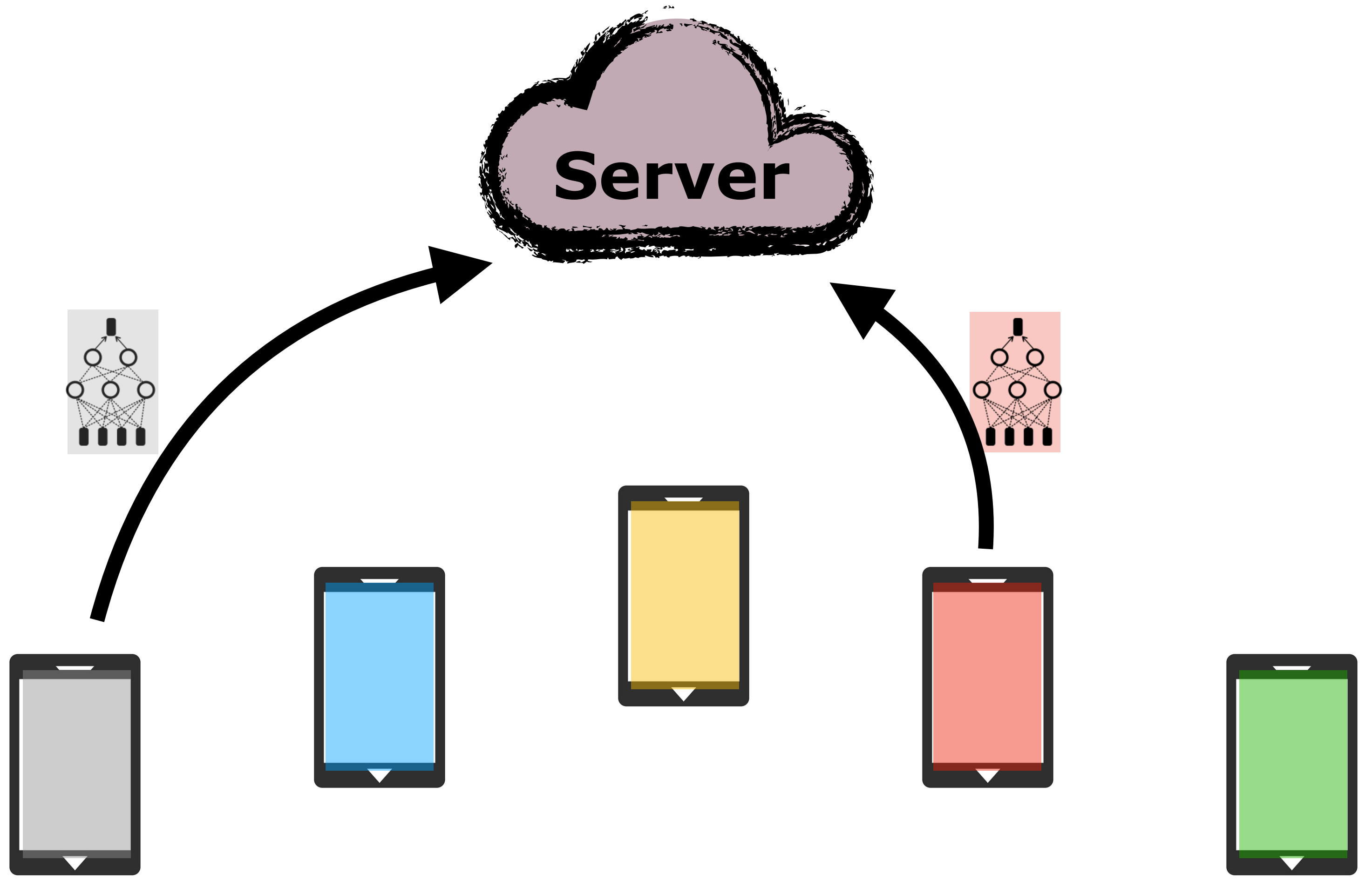
# Algorithm

In each communication round:

- Estimate the quantile




- Aggregate over the tail



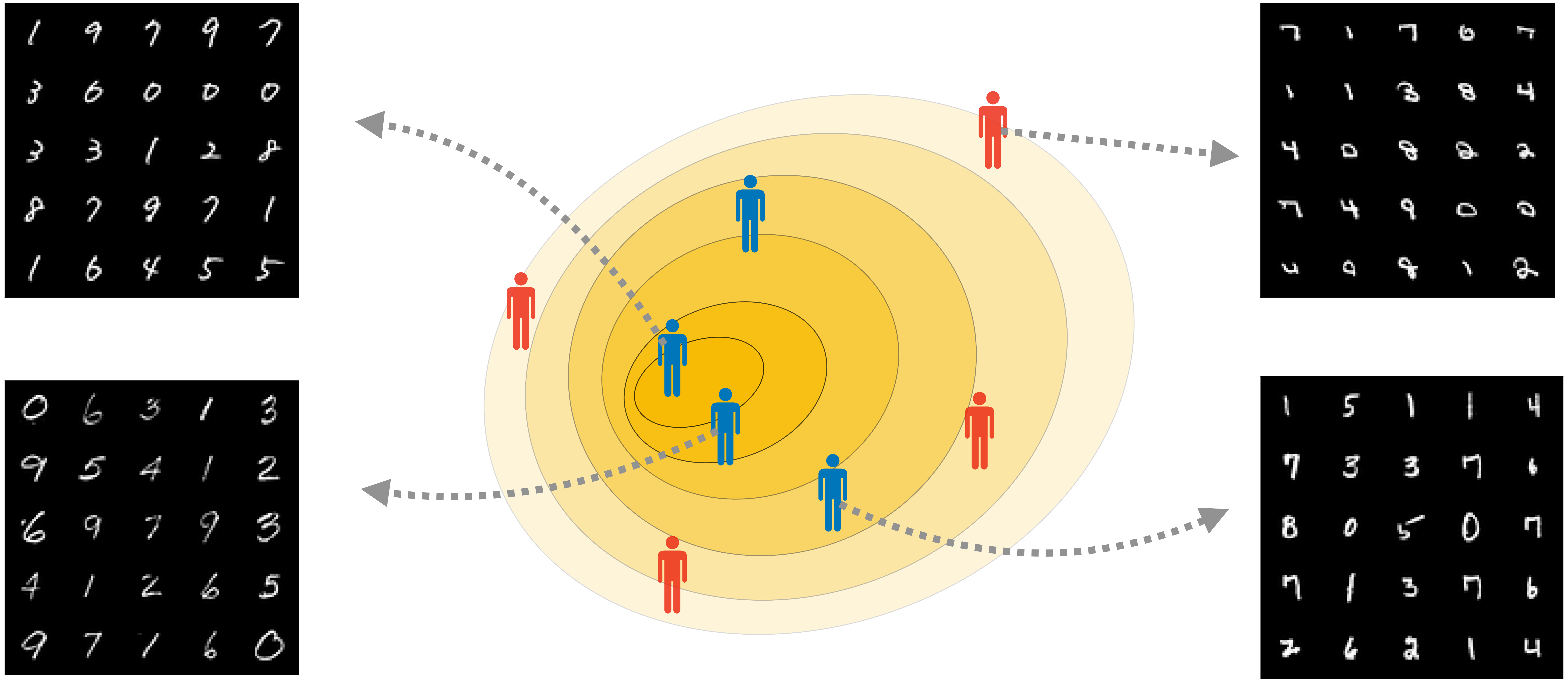
# Convergence rates

**Non-convex case**: $O(1/\sqrt{t})$ + lower order terms

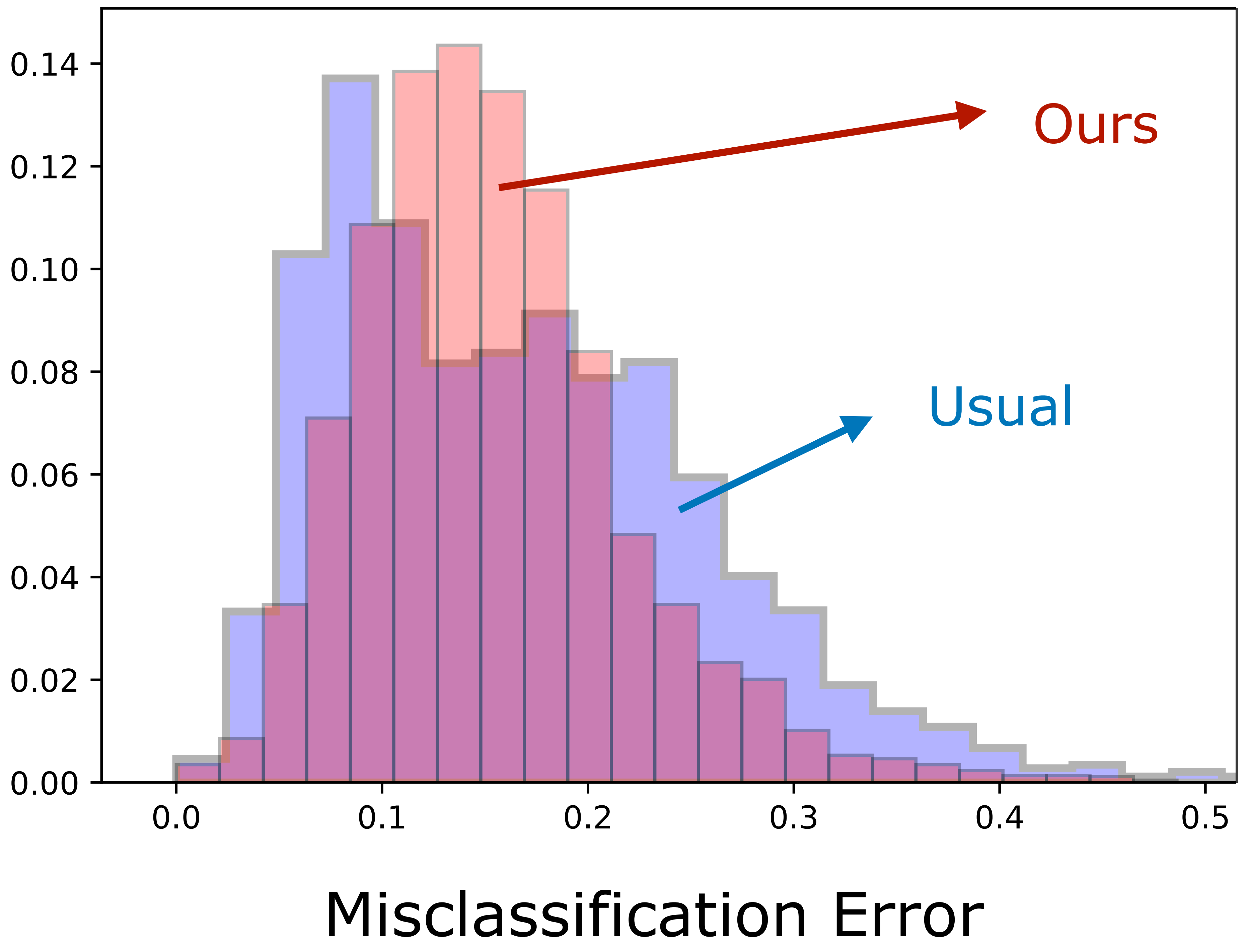
**Strongly convex case**: $\tilde{O}\left(\kappa^{3/2} + \frac{1}{\lambda\varepsilon}\right)$

$\kappa$: condition number
$\lambda$: strong convexity

# Experiments: EMNIST



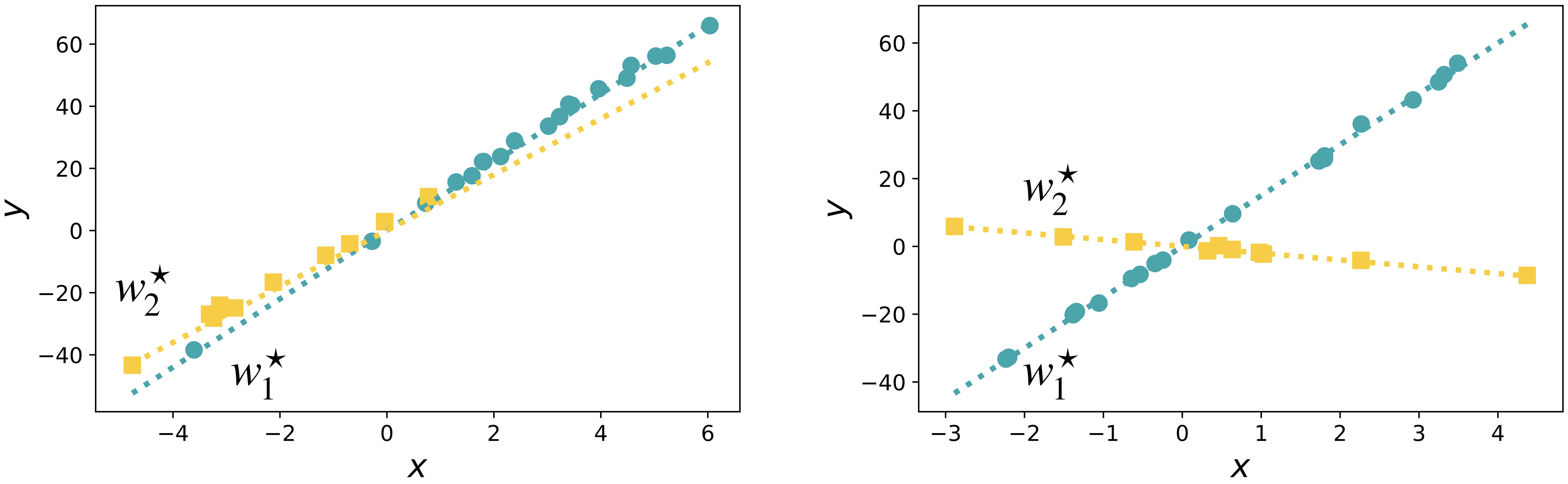
*Histogram of per-client errors*



# Tackling distribution shifts in federated learning

- Improving tail performance with a single model
- **Improving overall performance with local adaptation**

# The need for local adaptation a.k.a. personalization




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

loss on client $i$

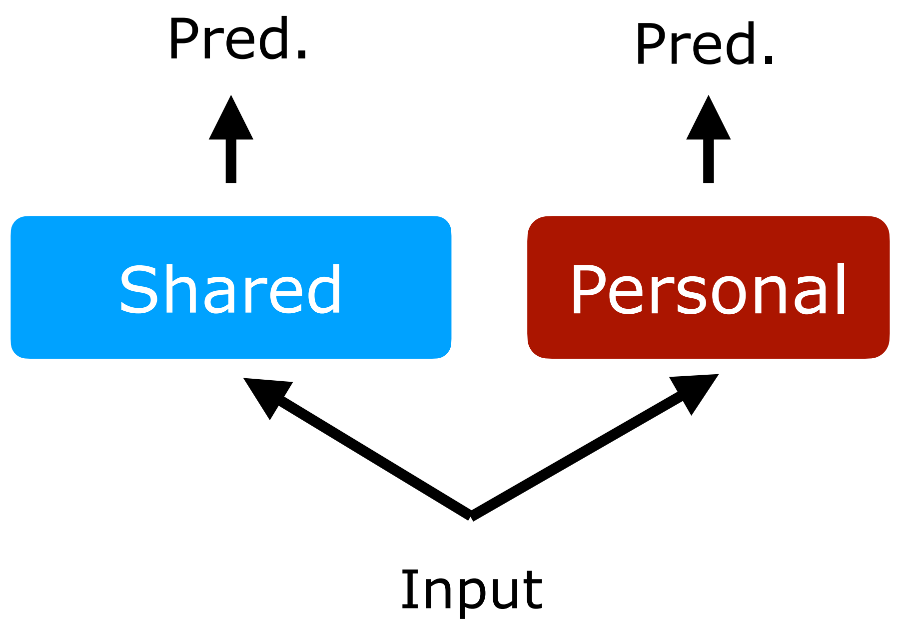
**Personalization:** Each model has a global component and a per-client component

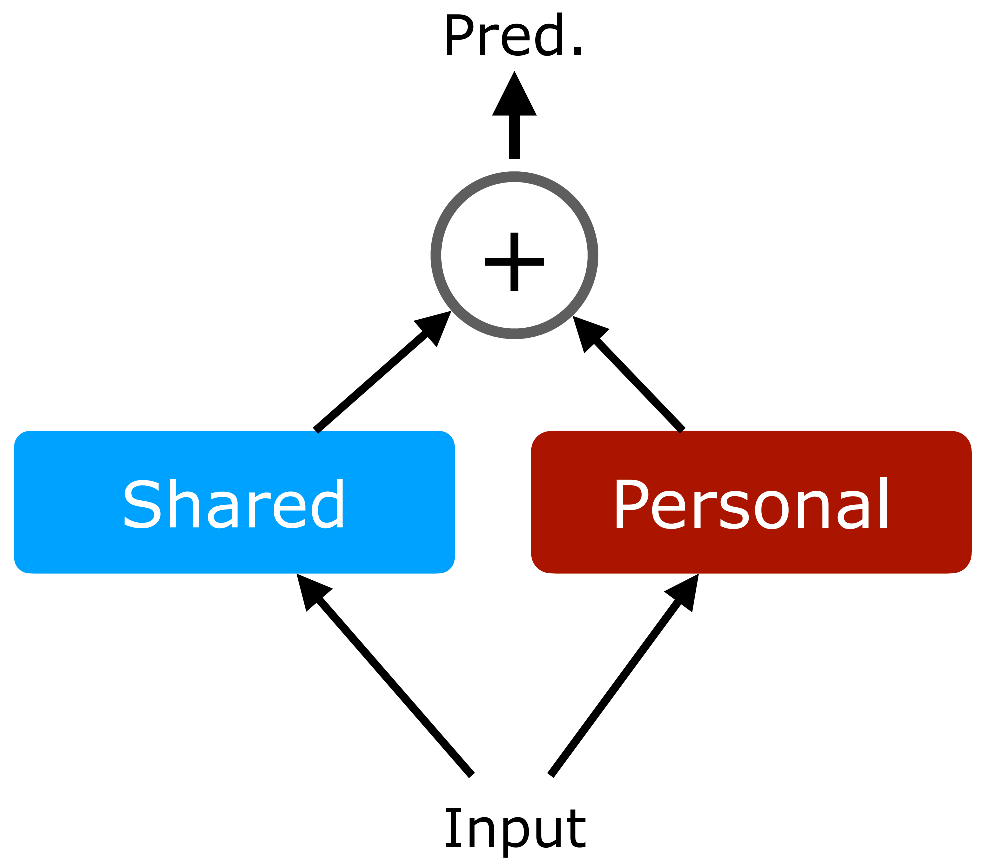
Shared Params $u$ + Personal Params $v_i$ = Full model $w_i = (u, v_i)$

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

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

# Personalization architectures



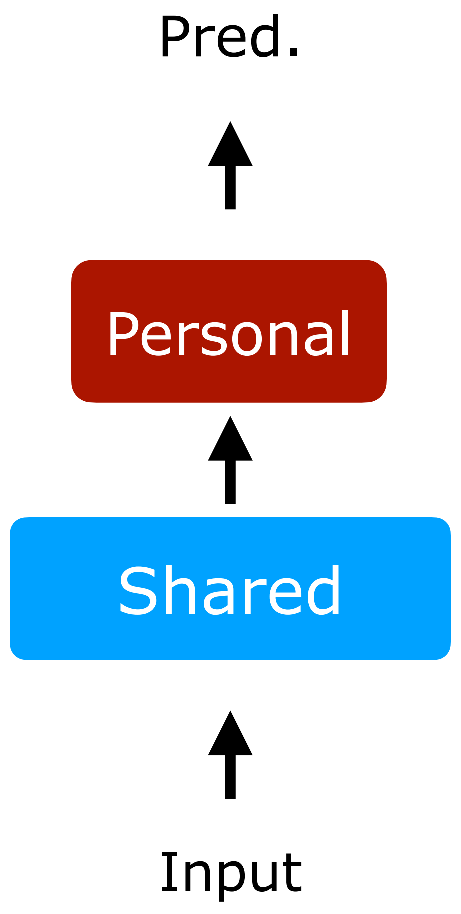

$$F_i(u, v_i) = \mathbb{E}_{(X,Y)\sim p_i}\left(\phi_g(X\,;\,u) + \phi_l(X\,;\,v_i) - Y\right)^2$$

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

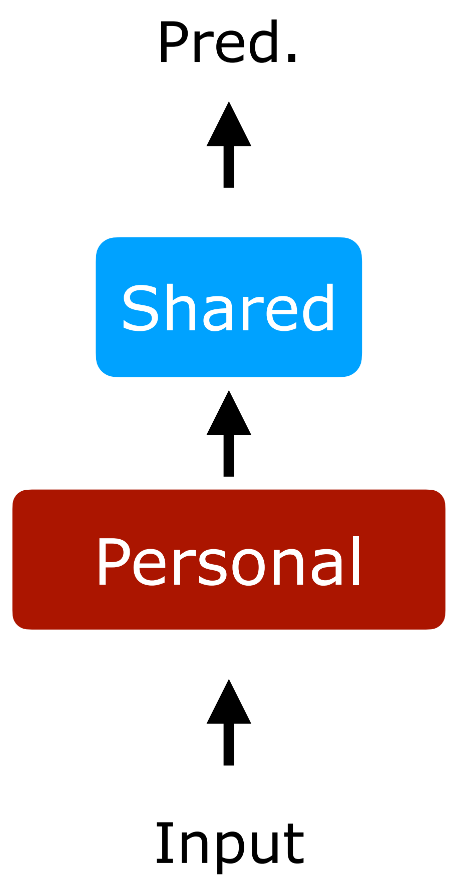
# Personalization architectures




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




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







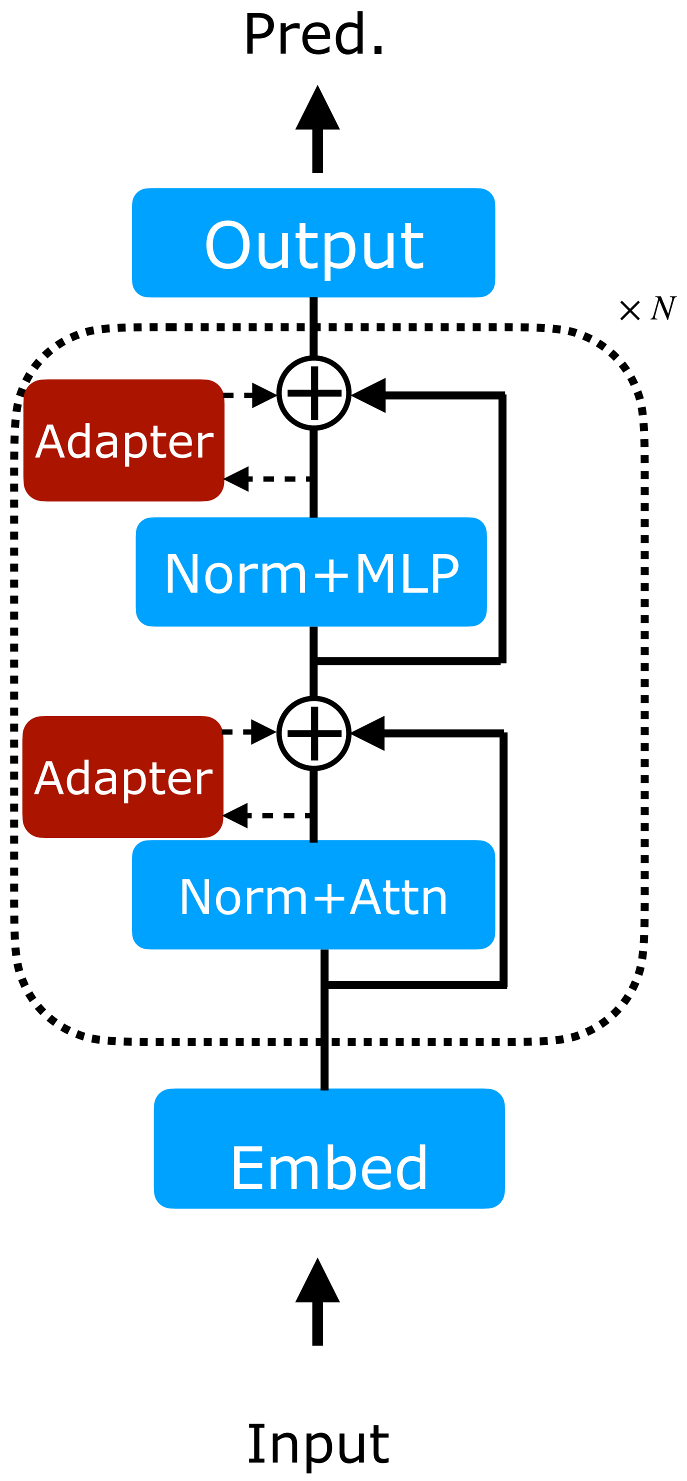
$$F_i(u, v_i) = \mathbb{E}_{(X,Y)\sim p_i}\left(\phi_g(X\,;\,u) + \phi_l(X\,;\,v_i) - Y\right)^2$$




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



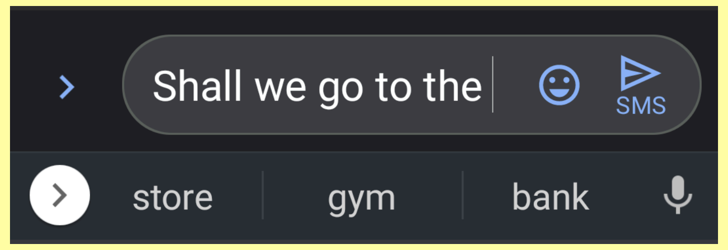

Liang et al. (2019)




Personalized Adapters

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

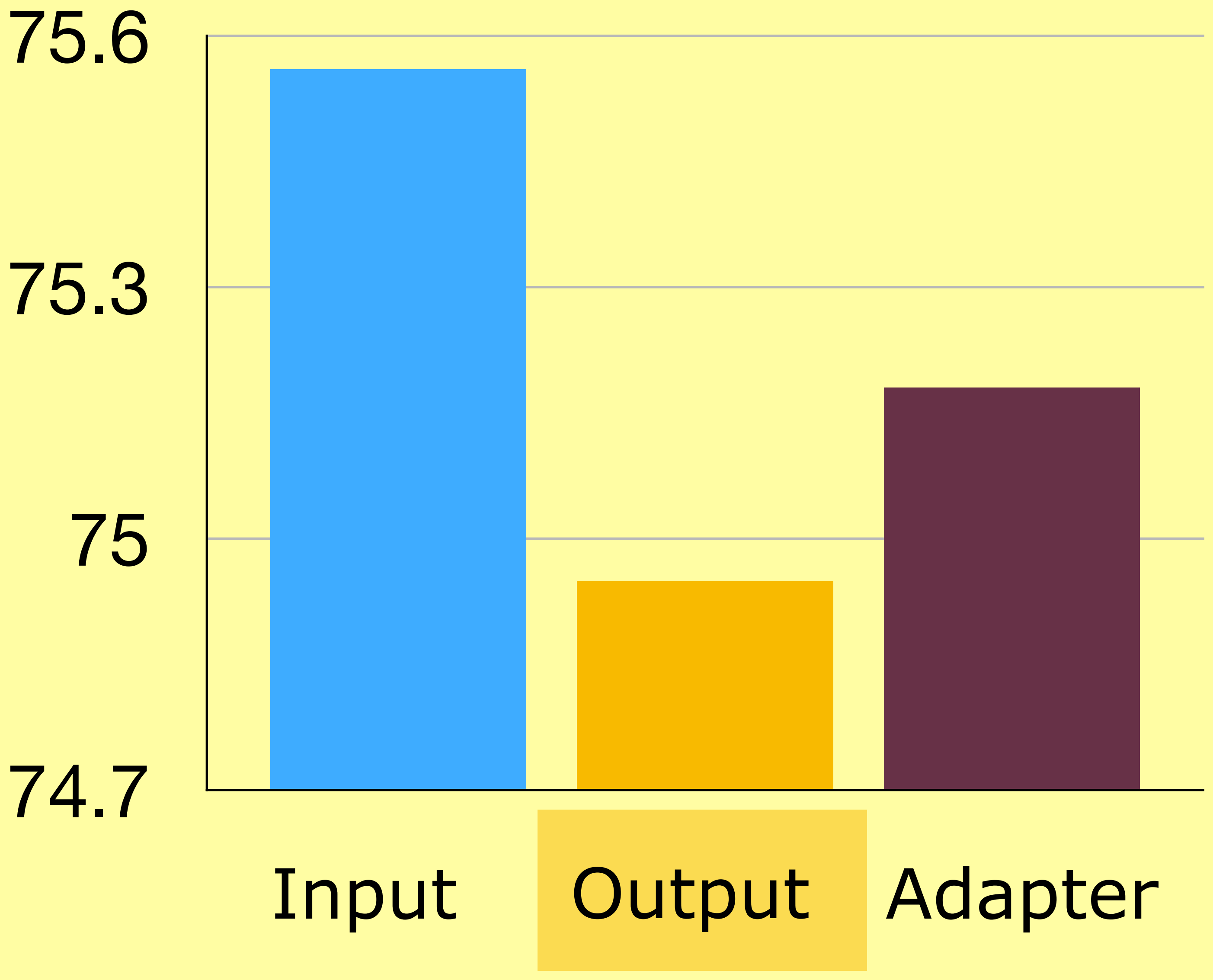
# Best personalization architecture depends on task heterogeneity

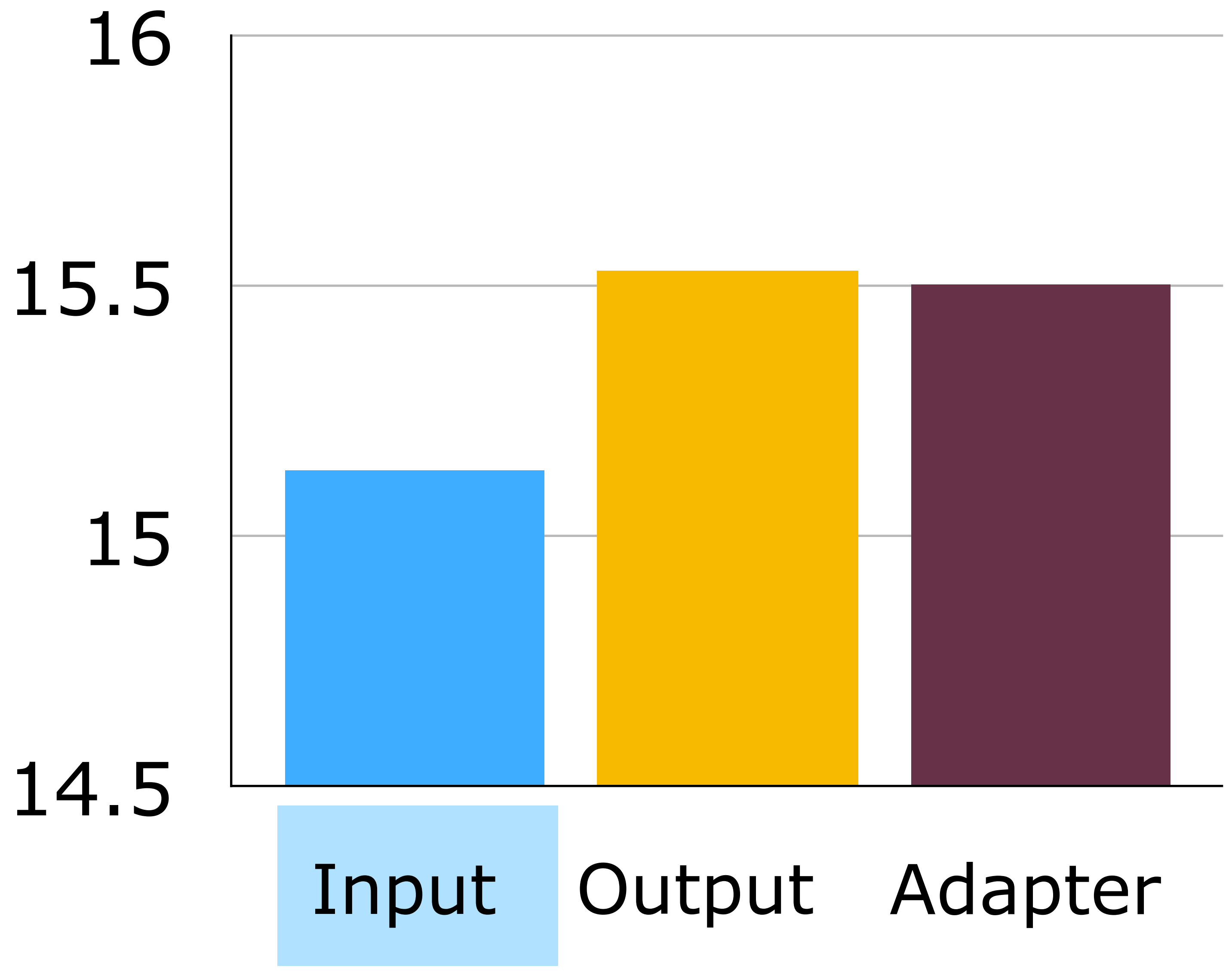


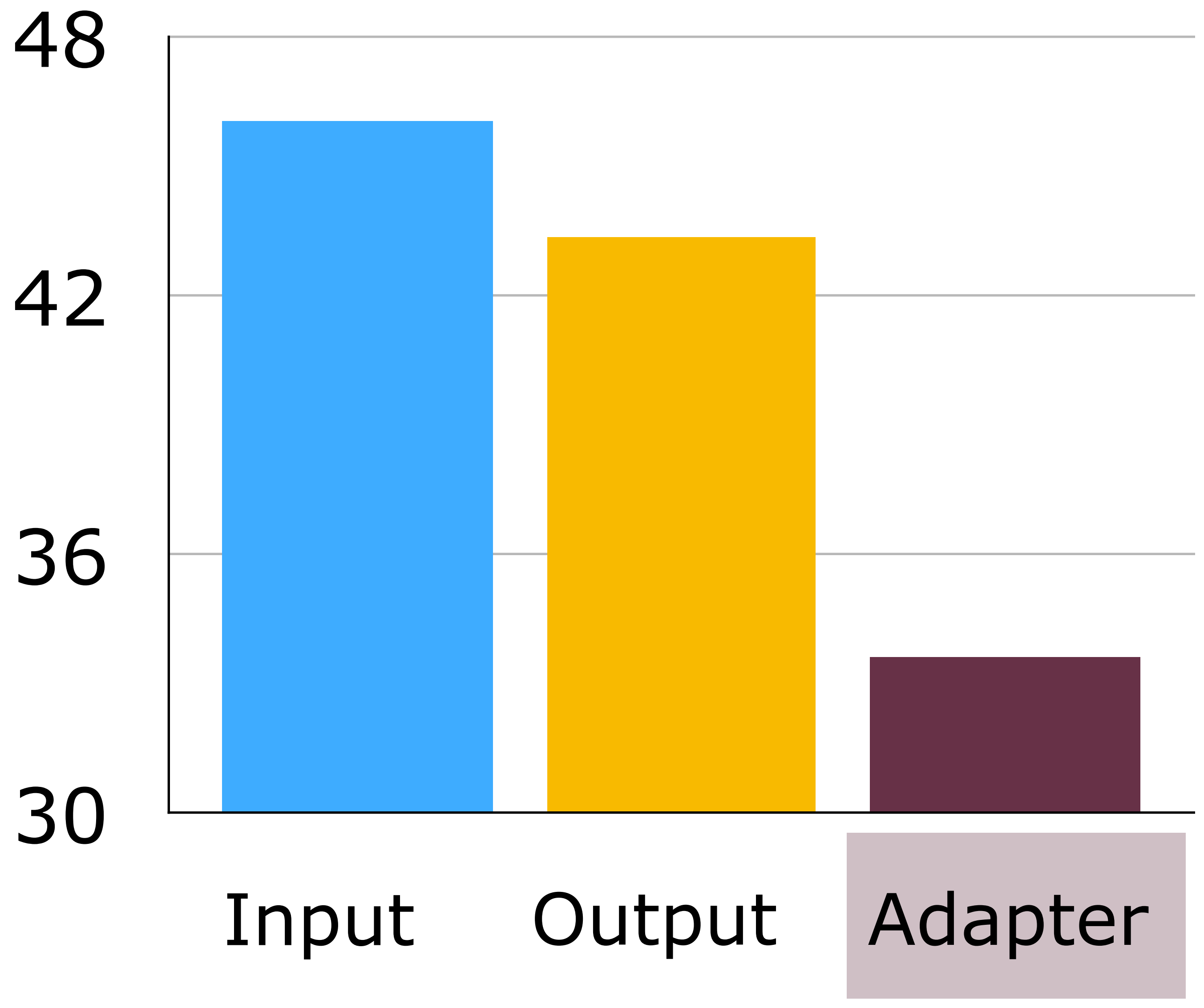

Next word prediction


Speech recognition

Landmark detection

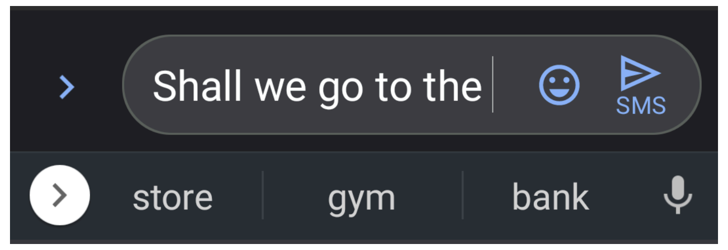


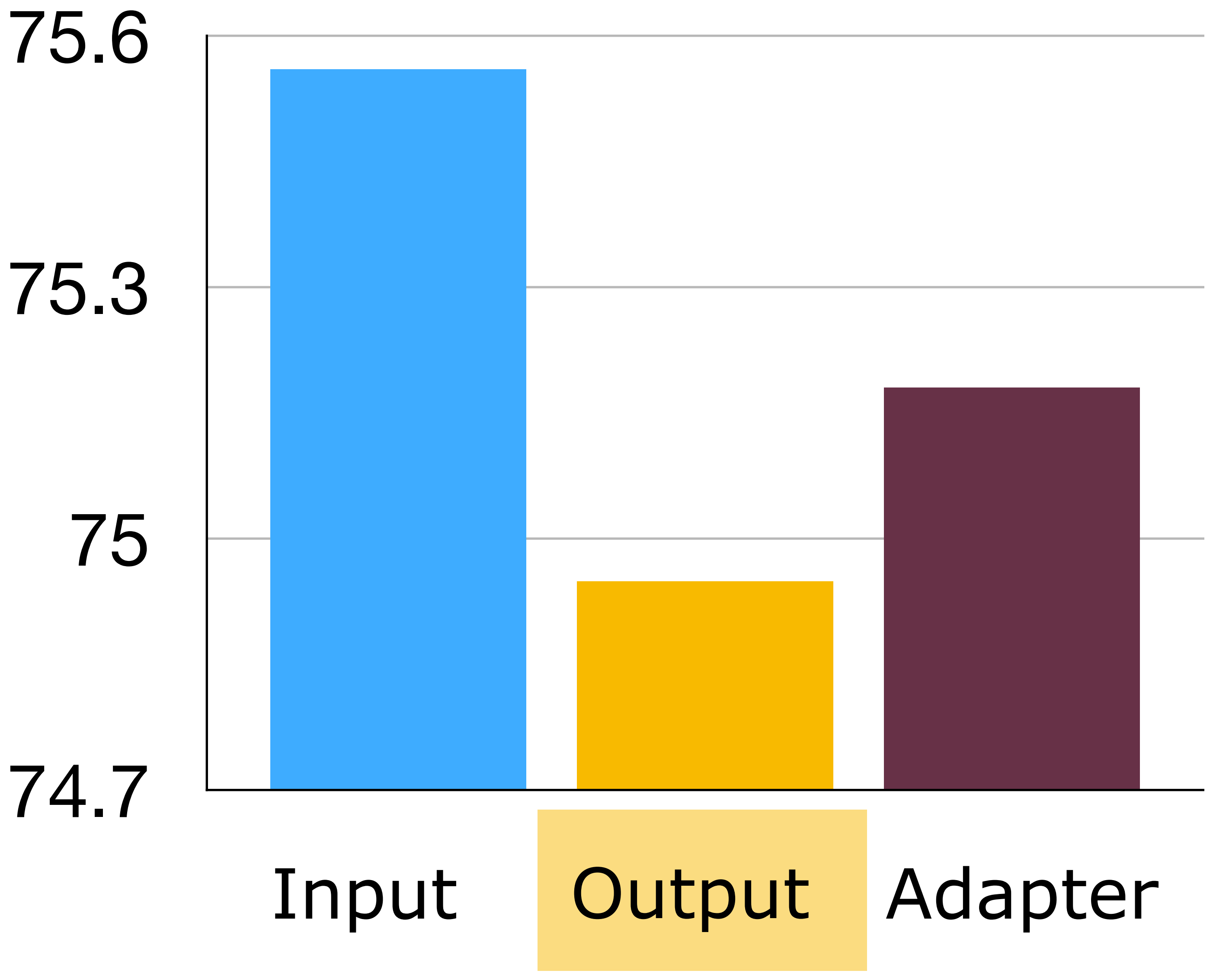




$y$-axis shows error: lower is better

# Best personalization architecture depends on task heterogeneity

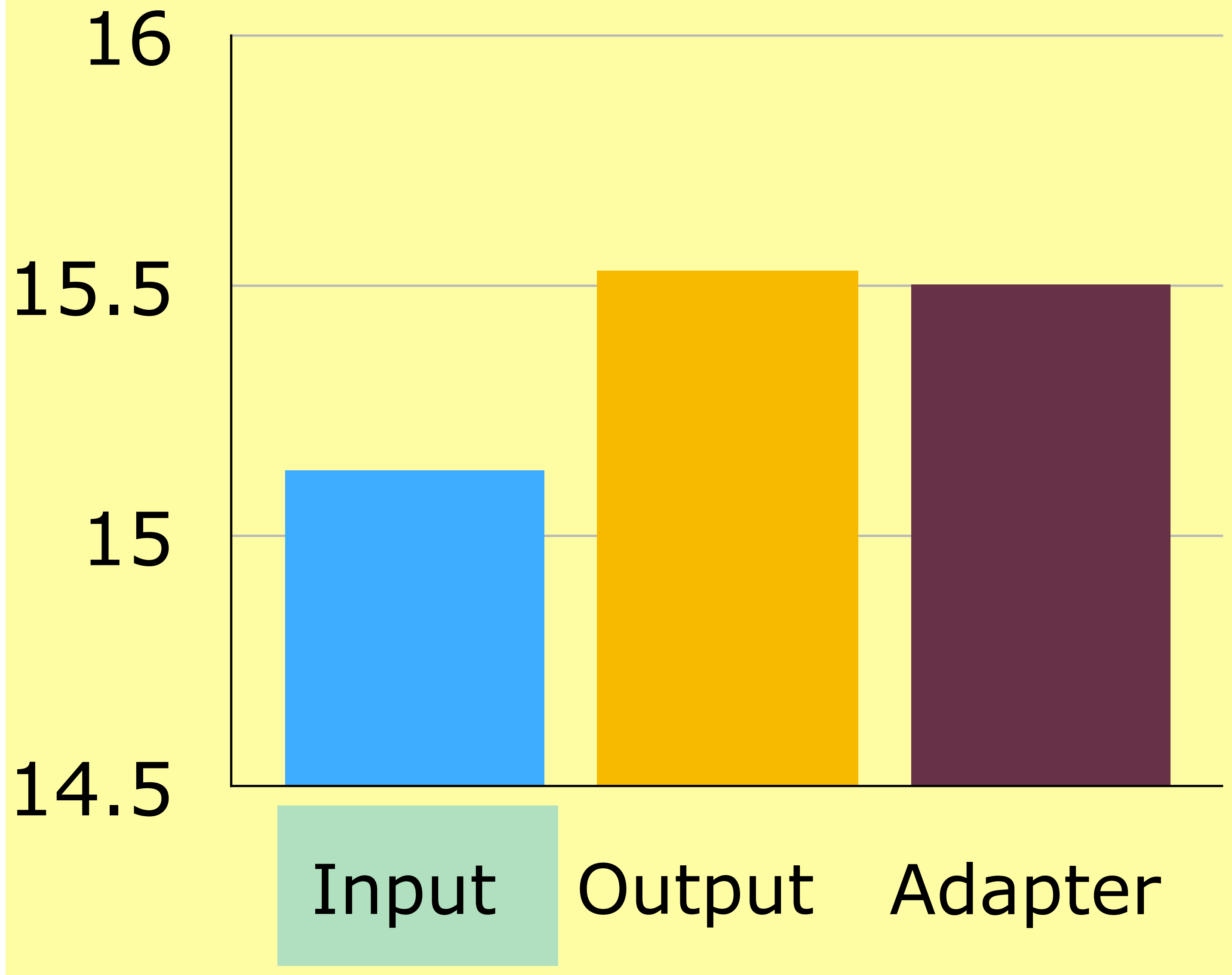



Next word prediction

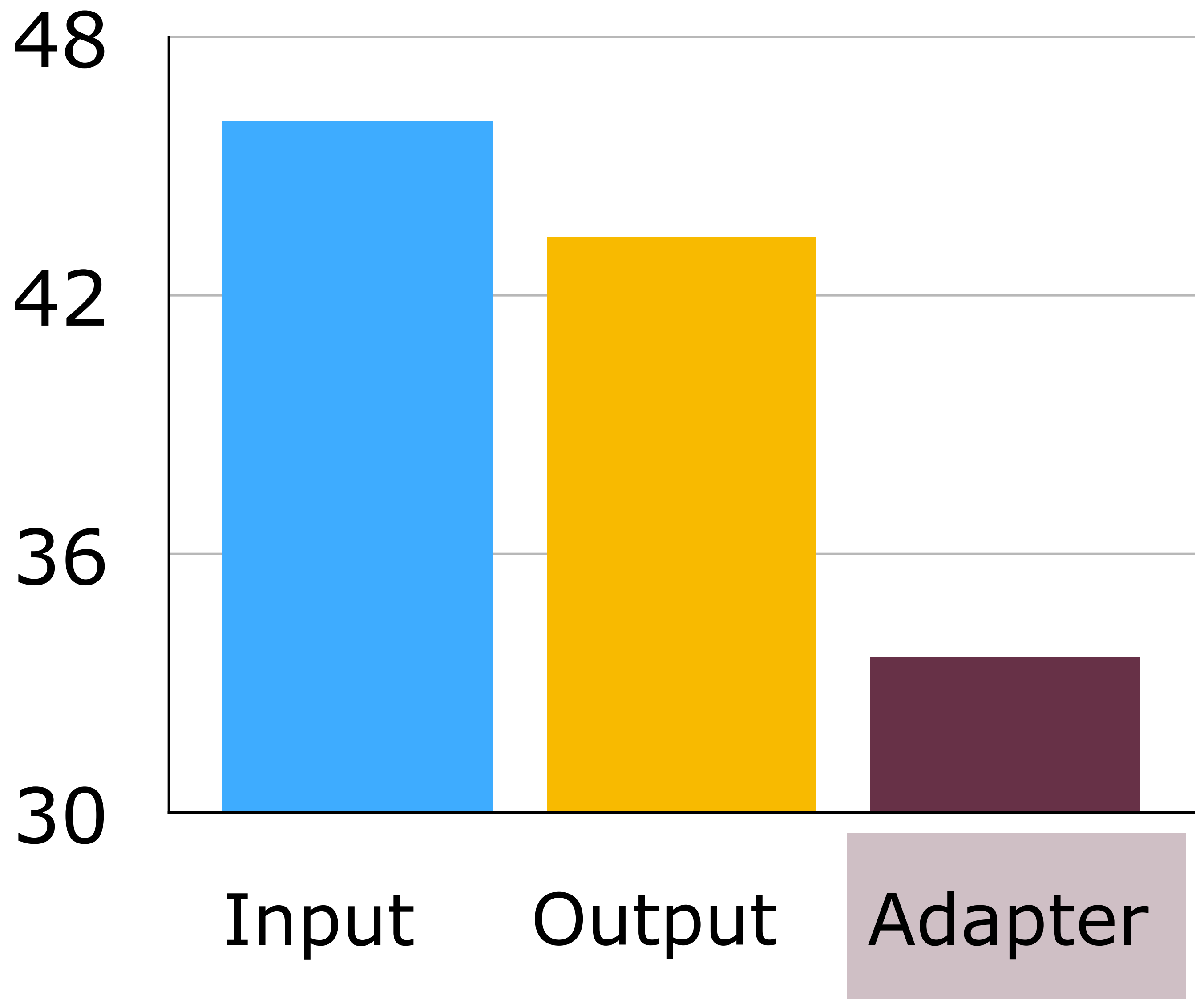




Speech recognition

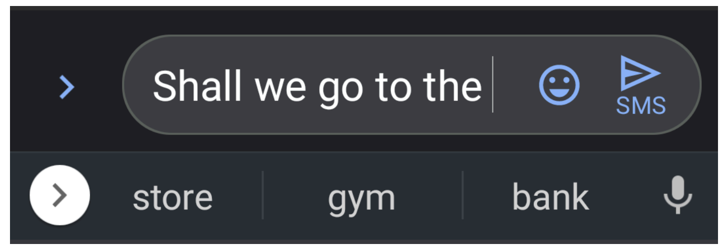



Landmark detection




$y$-axis shows error: lower is better

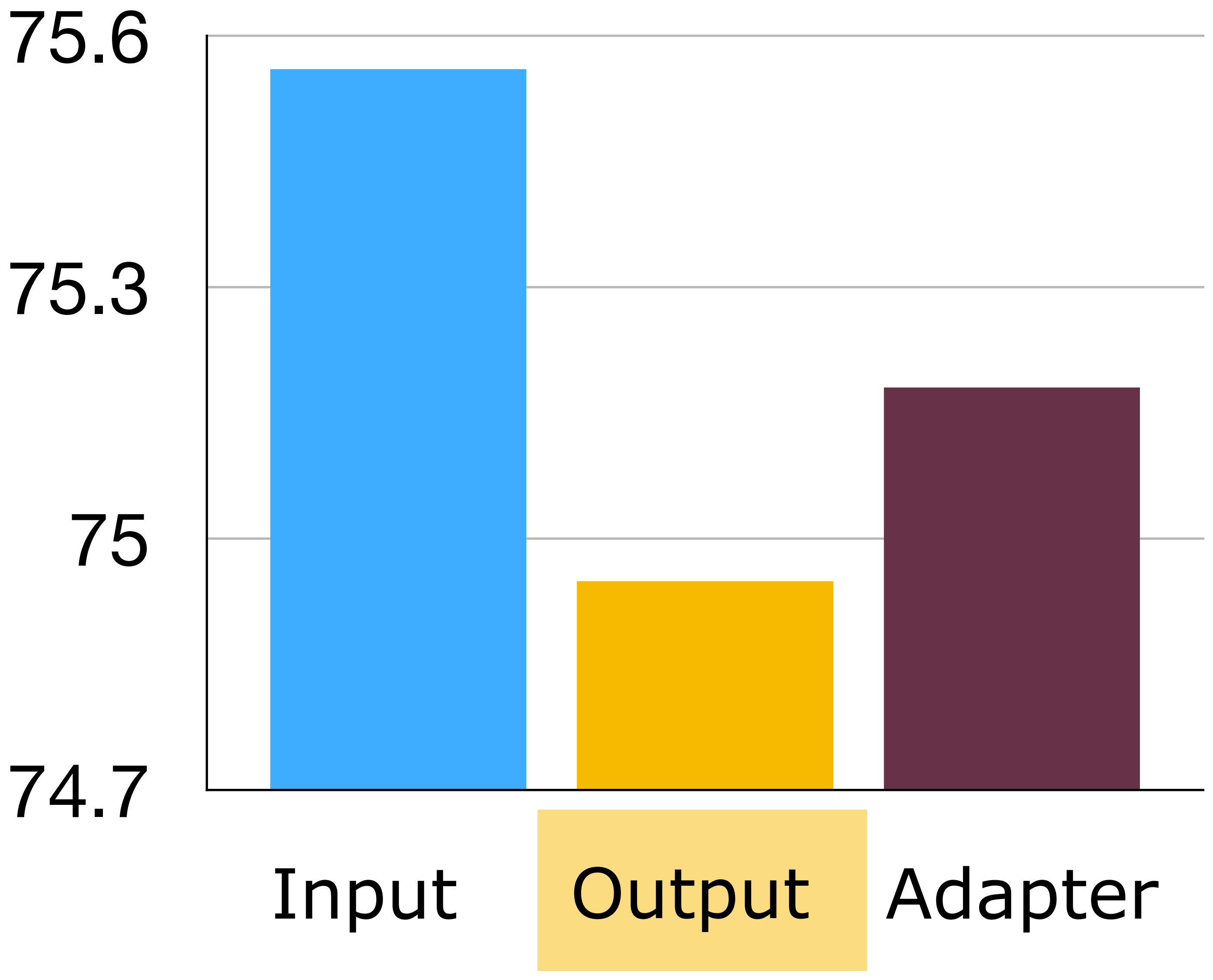
# Best personalization architecture depends on task heterogeneity

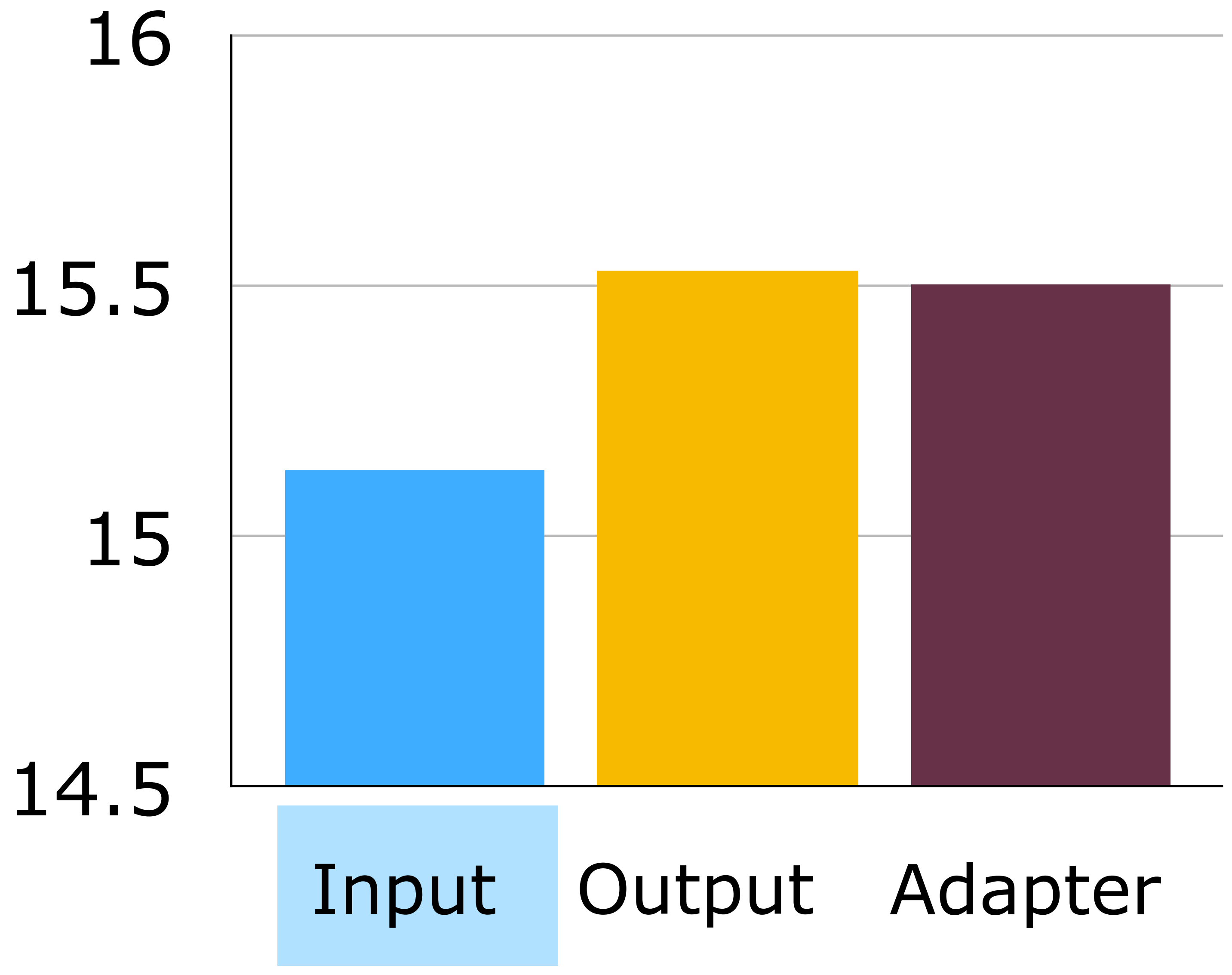


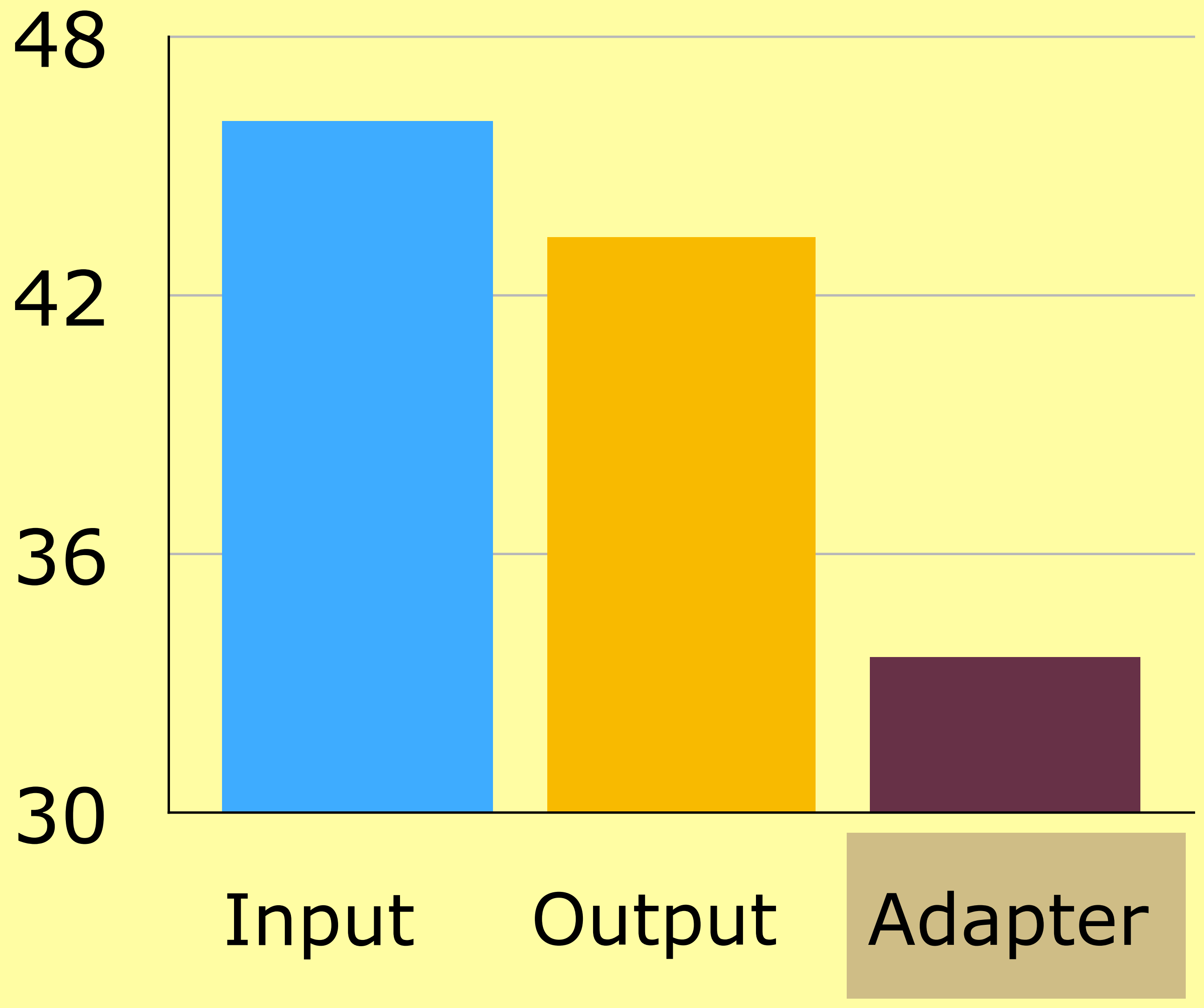

Next word prediction


Speech recognition

Landmark detection




$y$-axis shows error: lower is better

# Open problems: Deeper understanding of shifts

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

Yet, federated learning is used widely in practice

# Open problems: Deeper understanding of shifts

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

Yet, federated learning is used widely in practice

**Quantify heterogeneity**:

Measure gaps between distributions: **MAUVE**

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

# Open problems: Deeper understanding of shifts

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

Yet, federated learning is used widely in practice

**Quantify heterogeneity**:

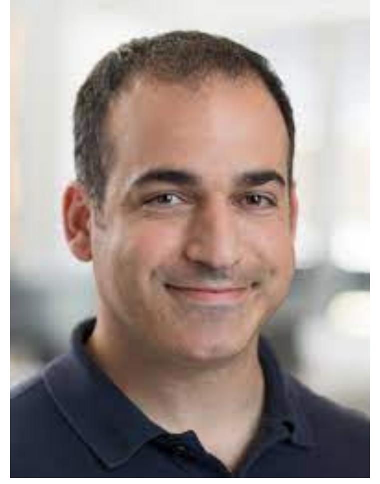
Measure gaps between distributions: **MAUVE**

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

Best algorithms for different types of shifts (subject to federated constraints)

Statistical assumptions under which heterogeneity is benign?

What measures of heterogeneity impact optimization?

**Federated Learning with Partial Model Personalization.**
Krishna Pillutla, Kshitiz Malick, Abdulrehman Mohamed, Mike Rabbat, Maziar Sanjabi, Lin Xiao
*ICML (2022).*

**Federated Learning with Heterogeneous Devices: A Superquantile Optimization Approach.**
Krishna Pillutla*, Yassine Laguel*, Jérôme Malick, Zaid Harchaoui.
*Under Review (arXiv 2112.09429)*

**A Superquantile Approach to Federated Learning with Heterogeneous Devices.**
Yassine Laguel*, Krishna Pillutla*, Jérôme Malick, Zaid Harchaoui.
*IEEE CISS (2021)*.

**Superquantiles at Work : Machine Learning Applications and Efficient (Sub)gradient Computation.**
Yassine Laguel, Krishna Pillutla, Jérôme Malick, Zaid Harchaoui.
*Set-Valued and Variational Analysis (2021)*.