Federated Learning with Partial Model Personalization

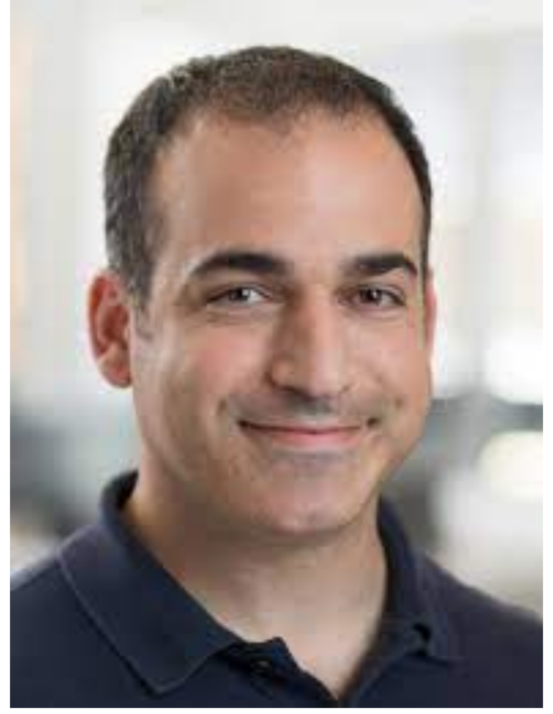
October 19th, 2022 @ FLOW Seminar

Krishna Pillutla

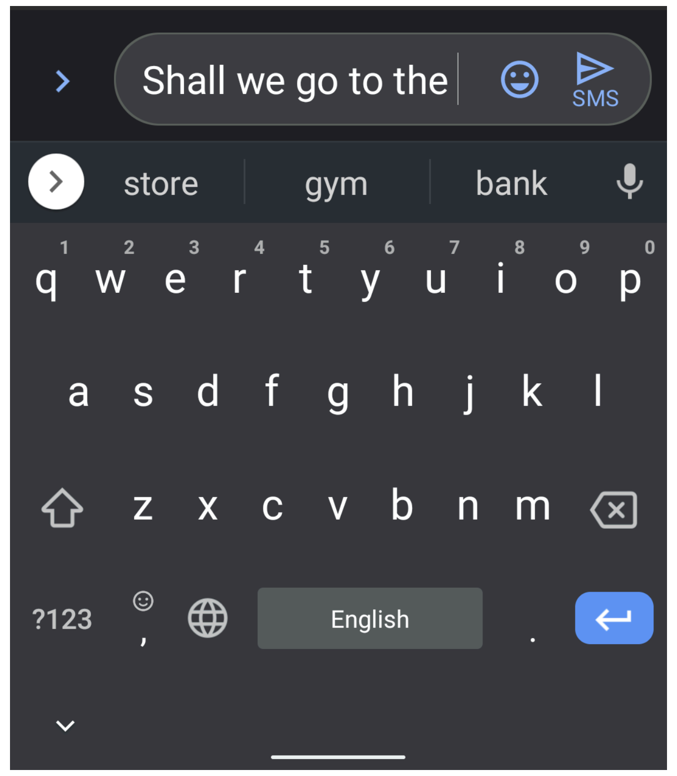
University of Washington → Google Research

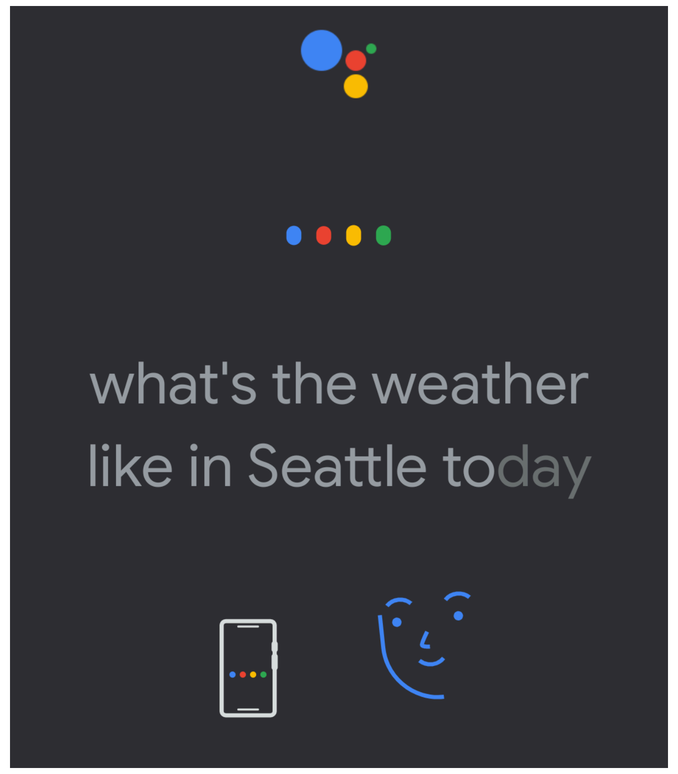
Joint work with

Meta AI

Kshitiz Malik	Abdelrahman Mohamed	Mike Rabbat	Maziar Sanjabi	Lin Xiao
				

ICML 2022





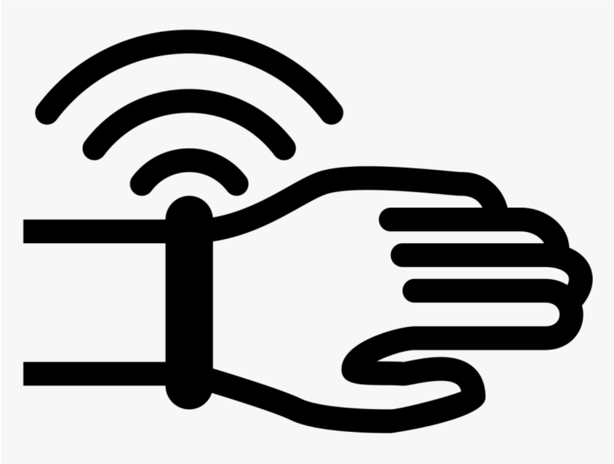
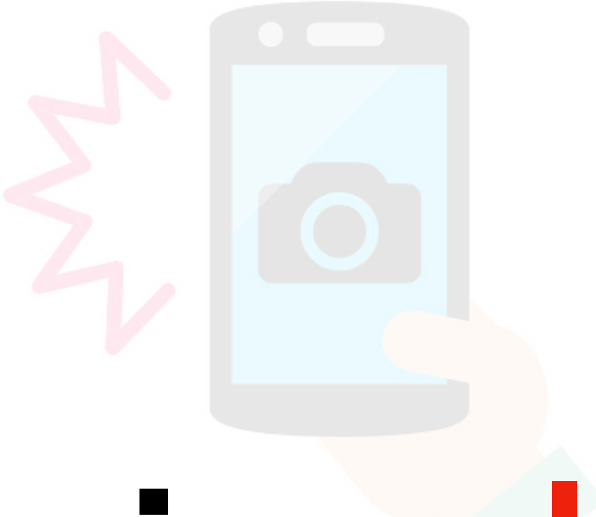


Image Credit: Robotics Business Review

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

Image Credit: Robotics Business Review

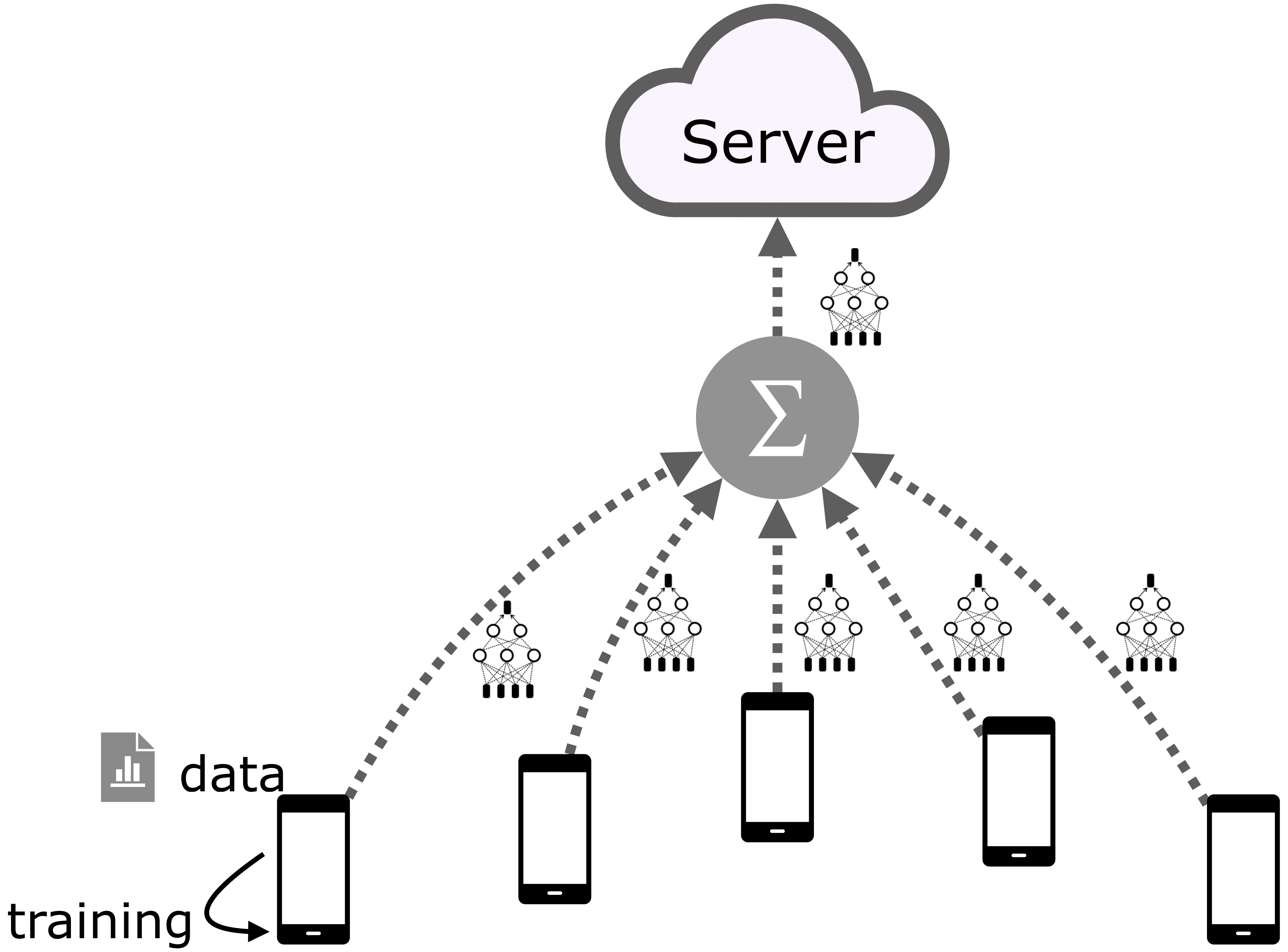



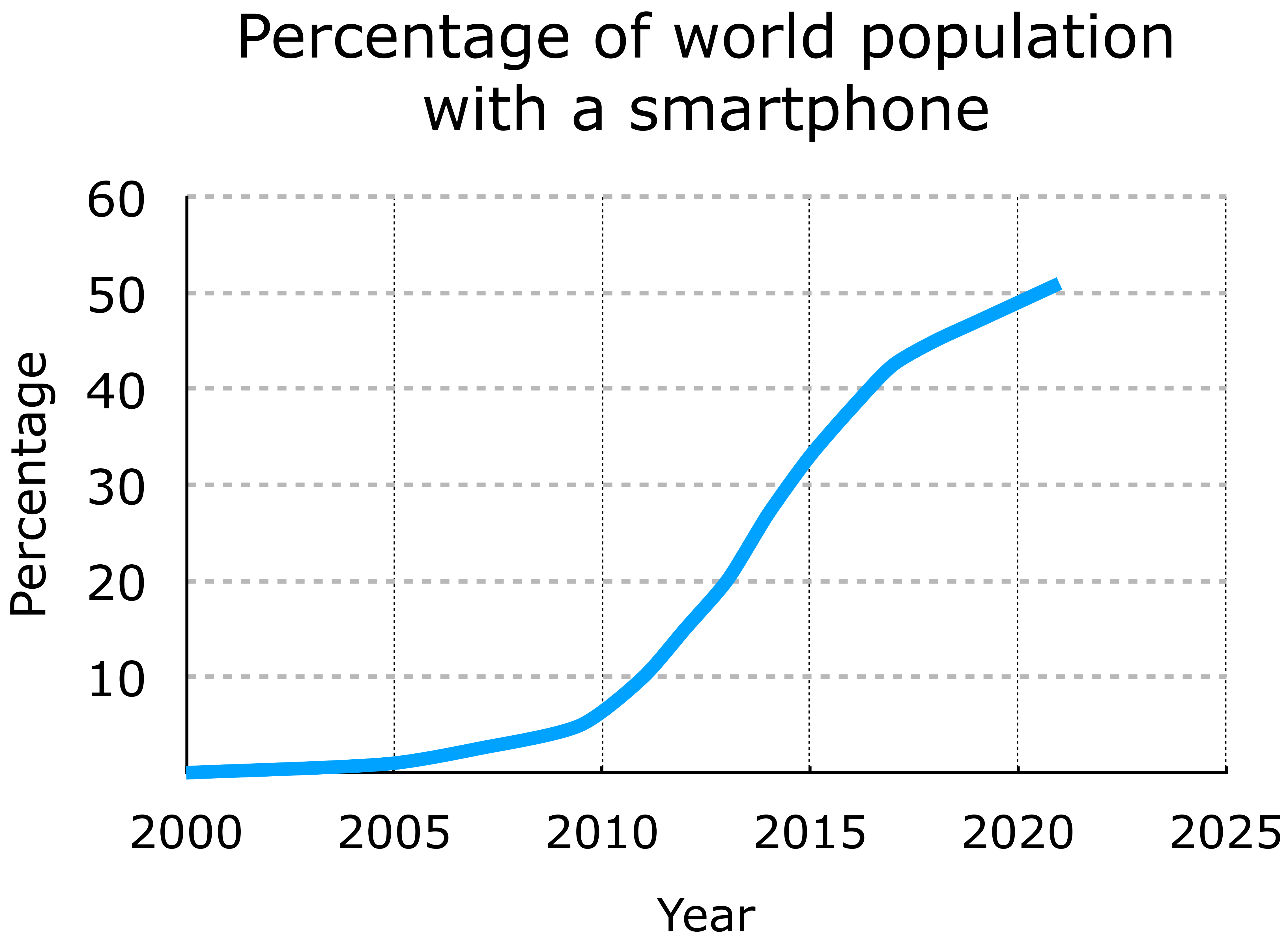
Data is decentralized and private


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

Image Credit: Wellcome

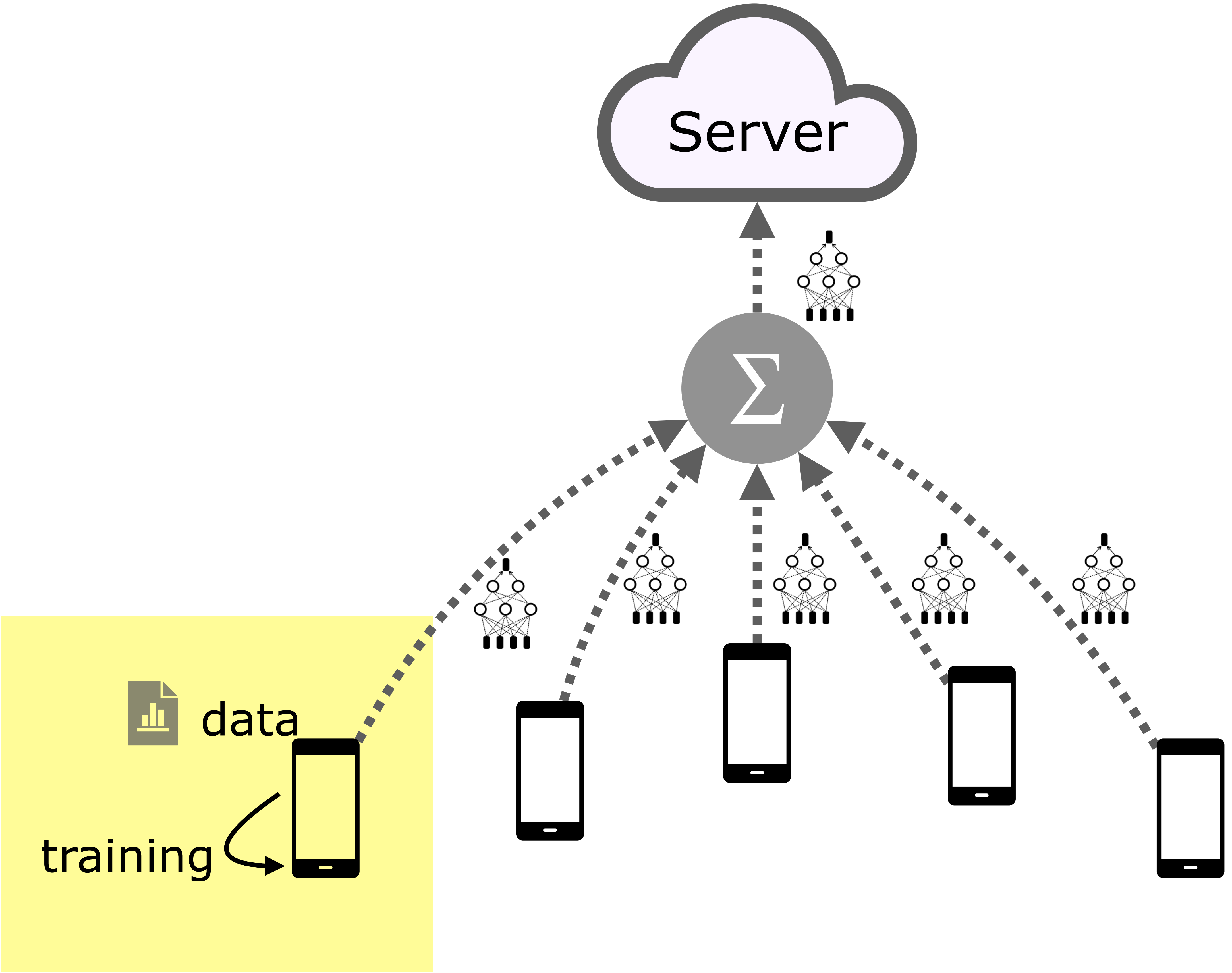
Federated Learning

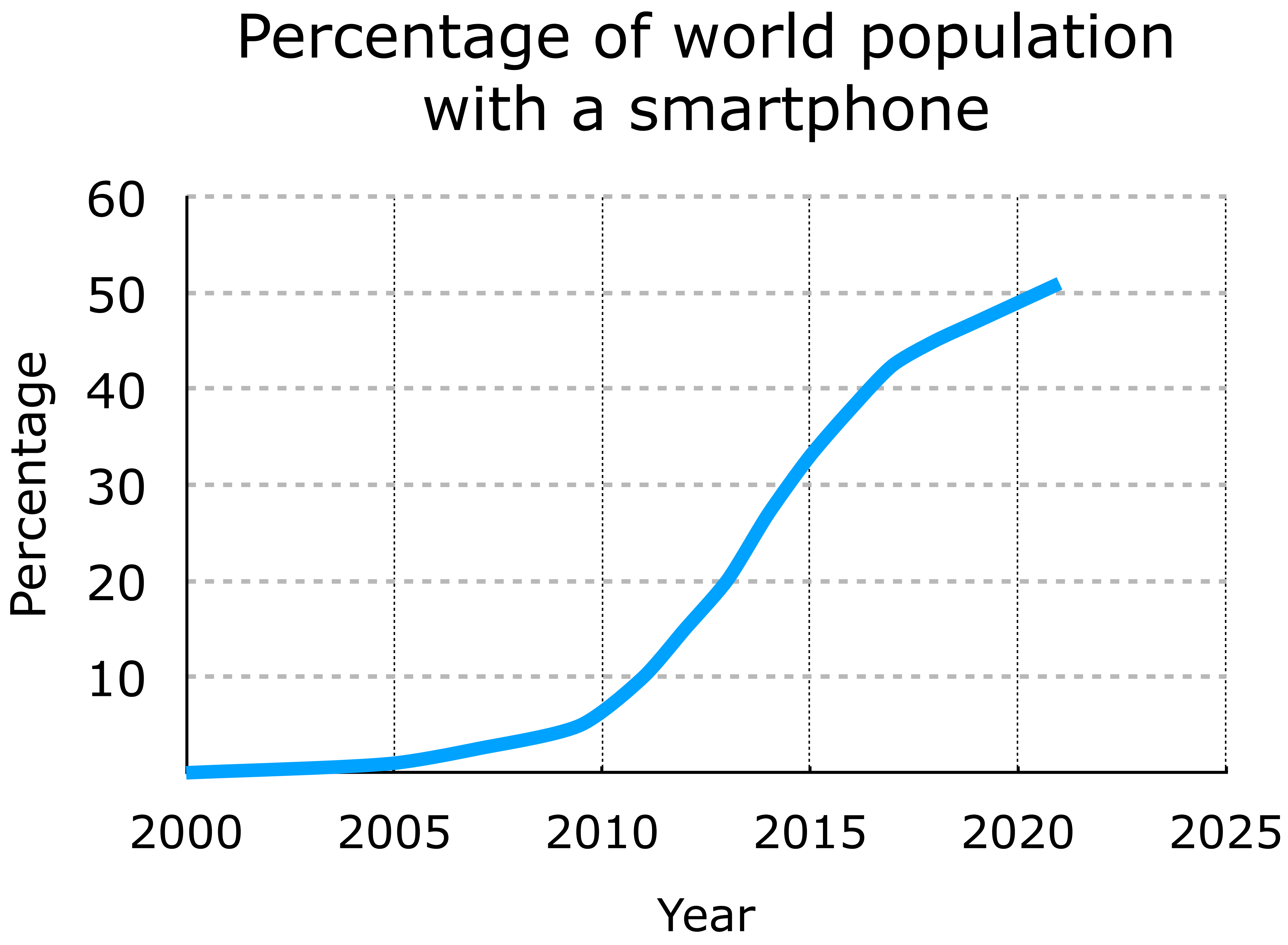




Data Credit: Business Wire

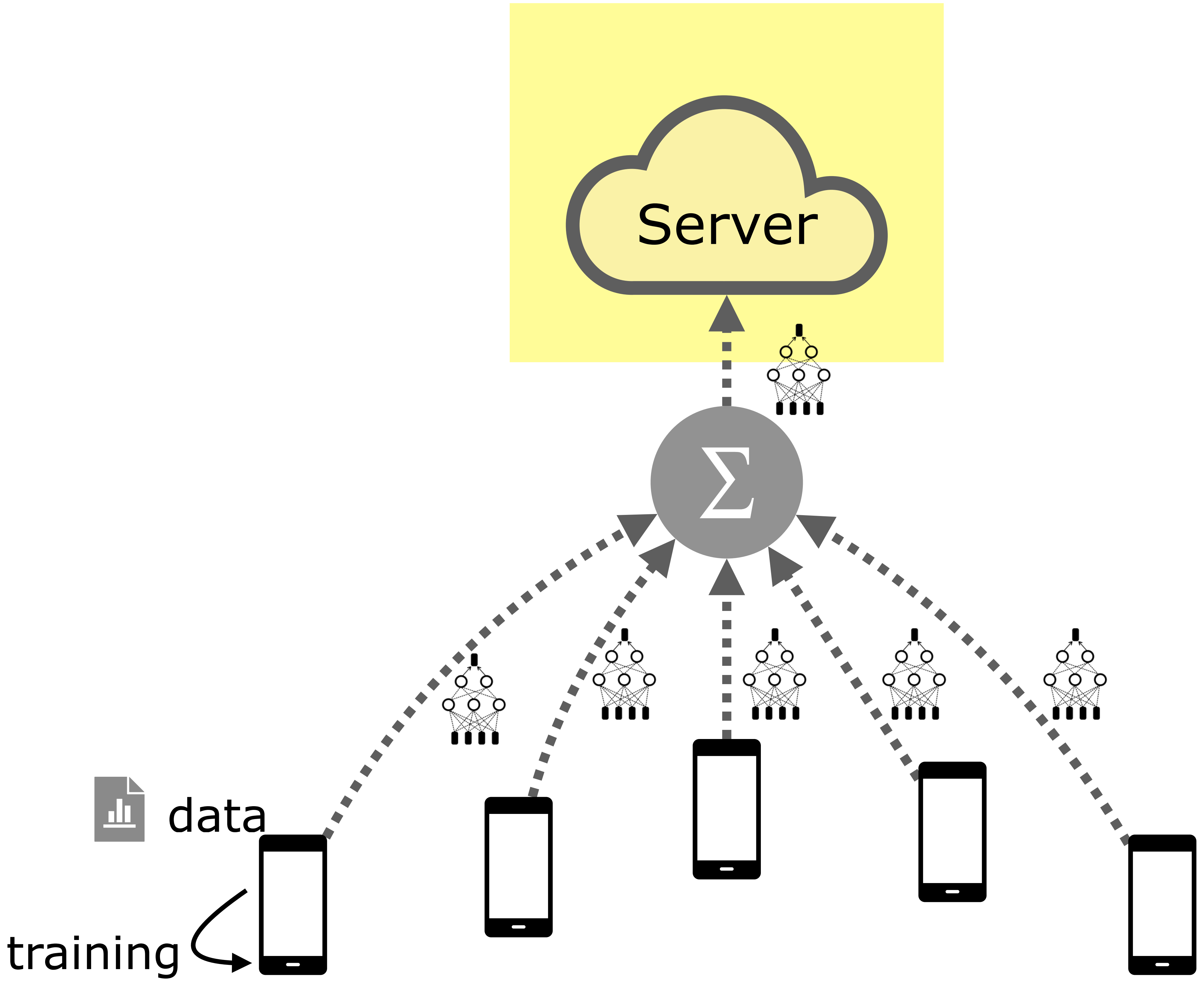
Federated Learning

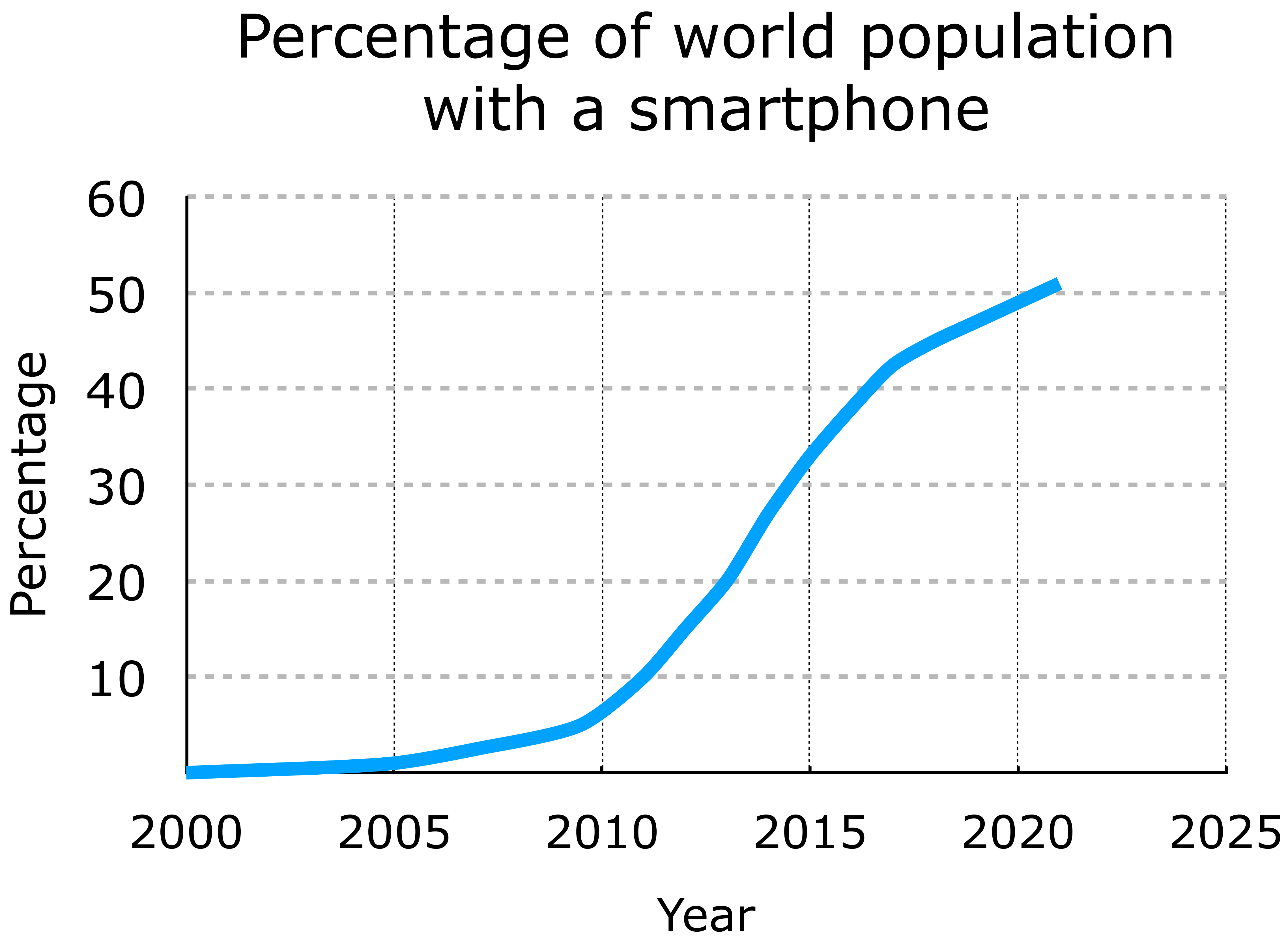




Data Credit: Business Wire

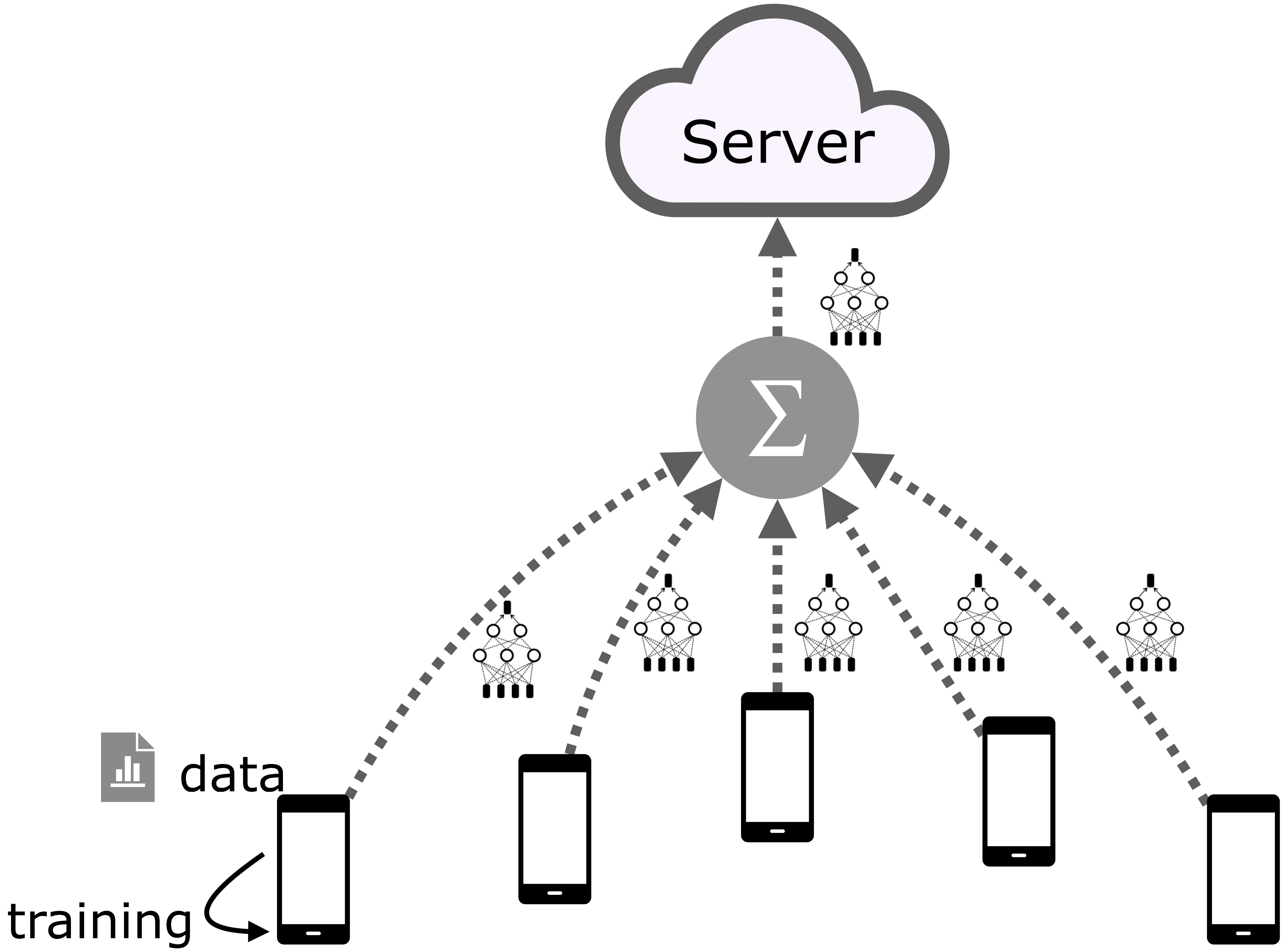
Federated Learning

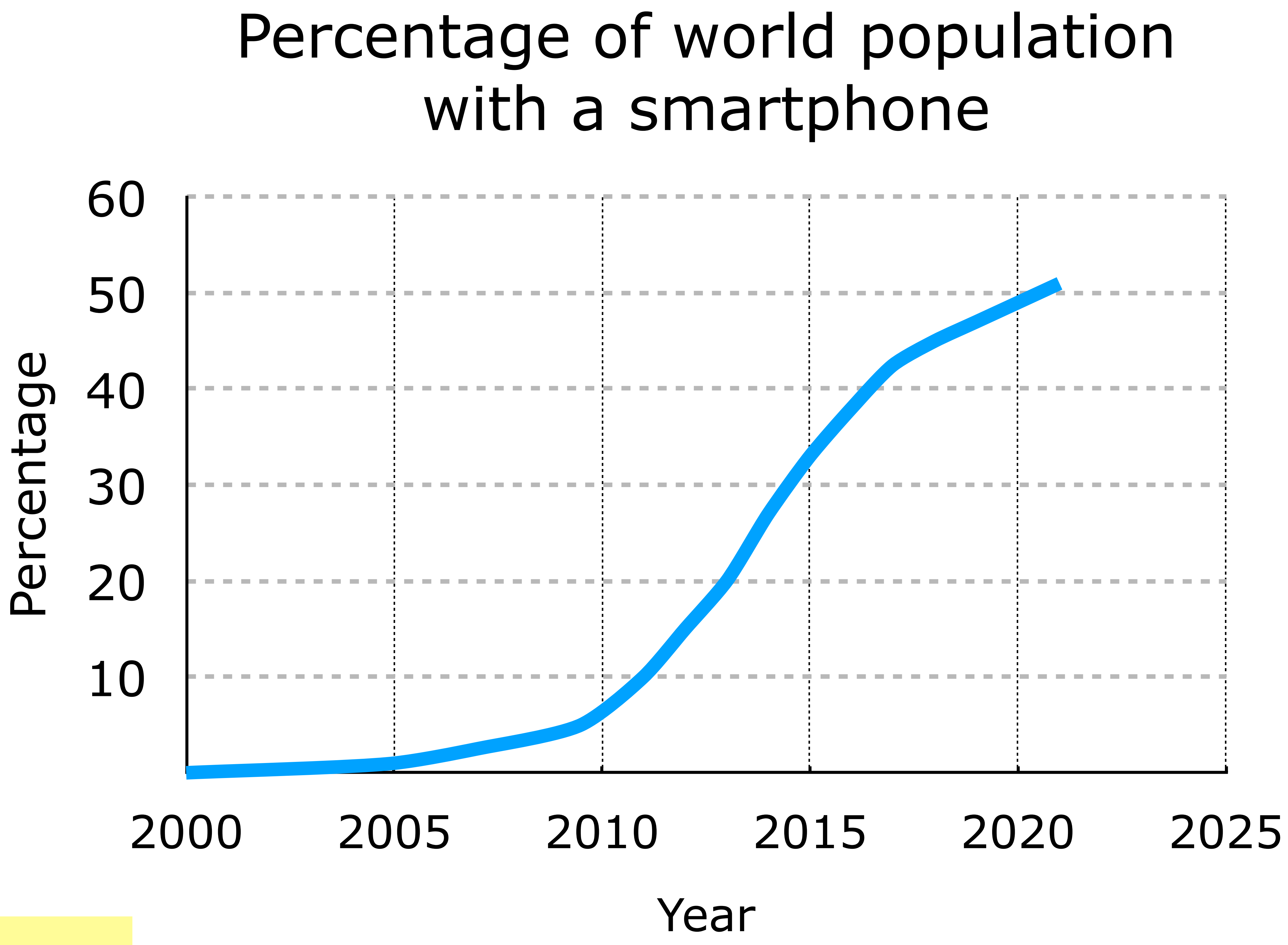




Data Credit: Business Wire

Federated Learning





Communication cost > computation cost!

Data Credit: Business Wire

Challenge

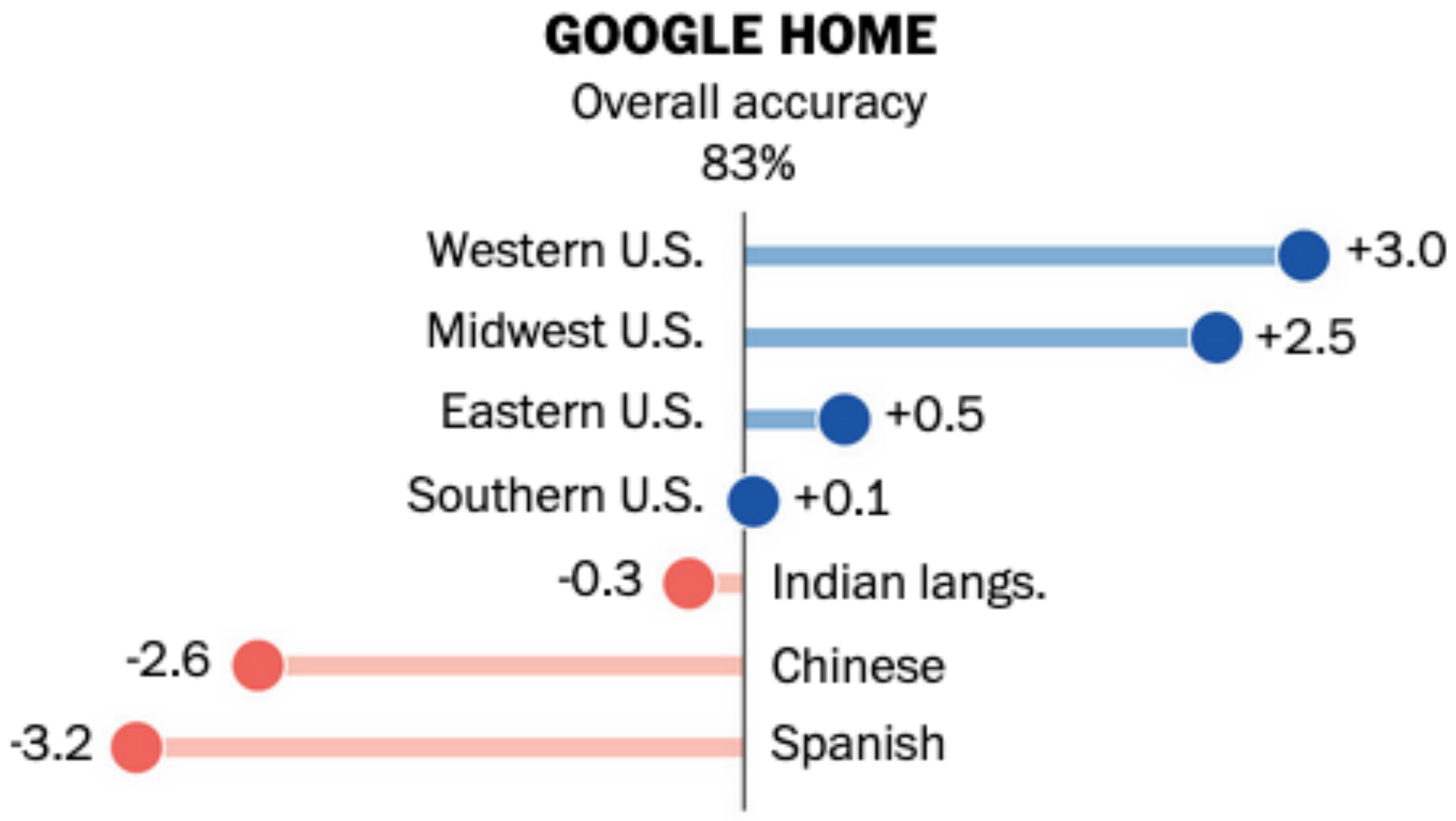
models are deployed on clients with heterogeneous data

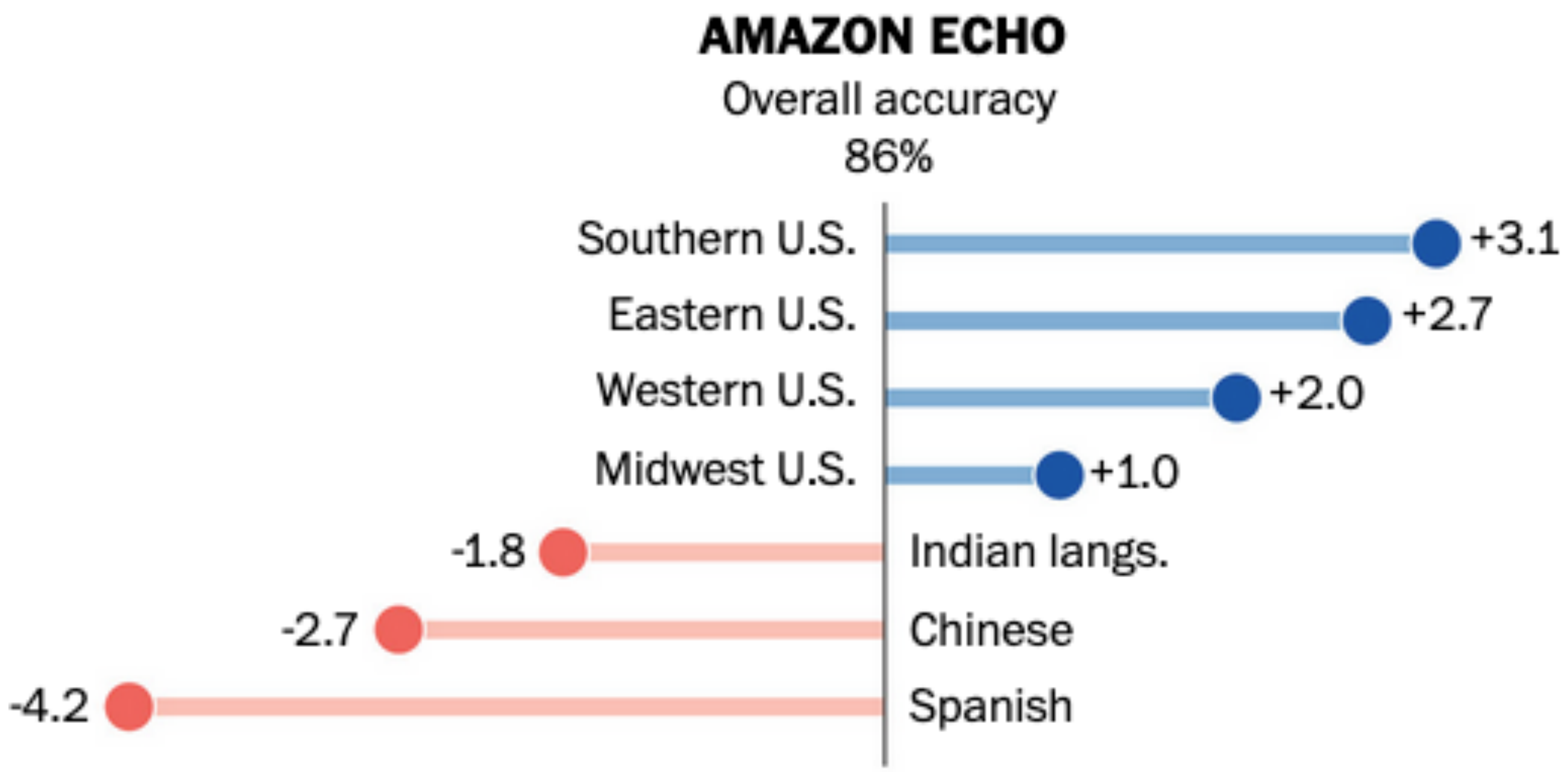
The Washington Post

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

Challenge

models are deployed on clients with heterogeneous data

Personalization: Adapt (a part of) the model to each client

Challenge

models are deployed on clients with heterogeneous data

Partial Personalization: Adapt **a part of** the model to each client

How to personalize?

Federated Learning with Personalization Layers

Manoj Ghuhan Arivazhagan
Adobe Research

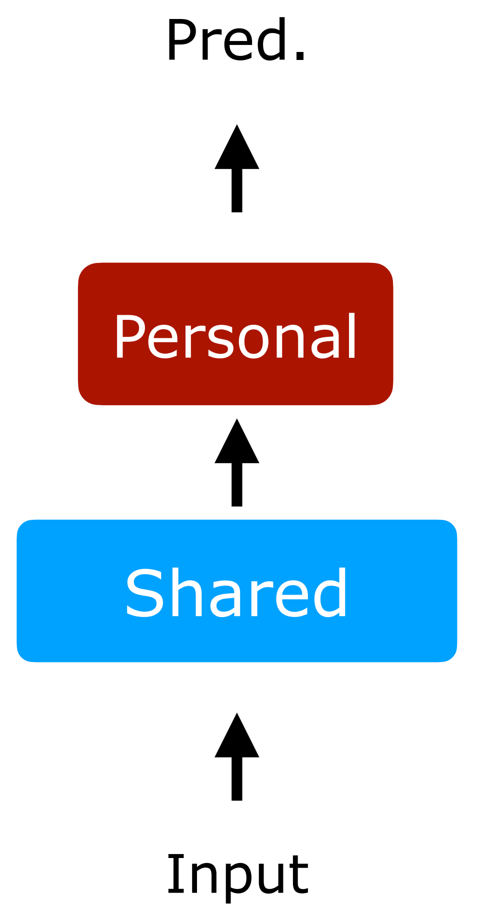
Vinay Aggarwal
Indian Institute of Technology, Roorkee, India

Aaditya Kumar Singh
Indian Institute of Technology, Kharagpur, India

Sunav Choudhary
Adobe Research

2019

Modeling: Personalize the **output** layer



Optimization: Train personal and shared parameters **simultaneously**

How to personalize?

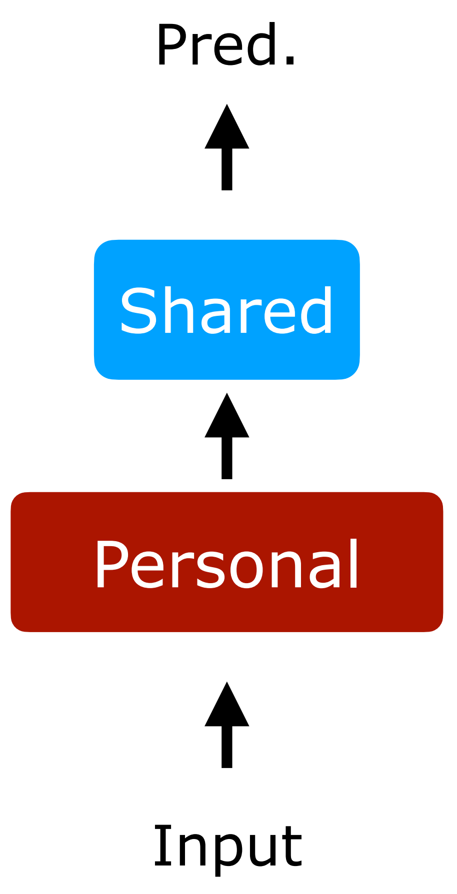
Think Locally, Act Globally:
Federated Learning with Local and Global Representations

Paul Pu Liang[1*], Terrance Liu[1*], Liu Ziyin[2], Nicholas B. Allen[3], Randy P. Auerbach[4], David Brent[5], Ruslan Salakhutdinov[1], Louis-Philippe Morency[1]

[1]School of Computer Science, Carnegie Mellon University
[2]Department of Physics, University of Tokyo
[3]Department of Psychology, University of Oregon
[4]Department of Psychiatry, Columbia University
[5]Department of Psychiatry, University of Pittsburgh
{pliang,terrancl,morency}@cs.cmu.edu

July 15, 2020

Modeling: Personalize the **input** layer



Optimization: Train personal and shared parameters **simultaneously**

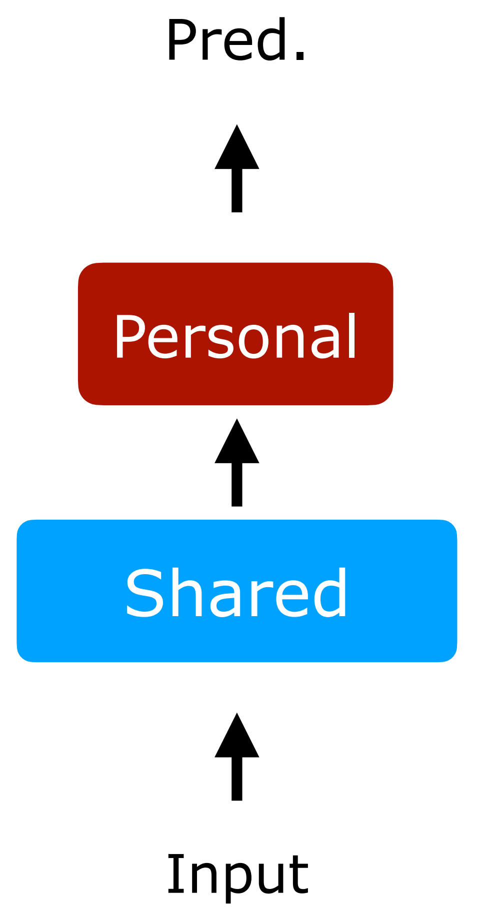
How to personalize?

Exploiting Shared Representations for Personalized Federated Learning

Liam Collins [1] Hamed Hassani [2] Aryan Mokhtari [1] Sanjay Shakkottai [1]

ICML 2021

Modeling: Personalize the **output** layer



Optimization: Train personal and shared parameters **alternatingly**

How to personalize?

**Federated Reconstruction:
Partially Local Federated Learning**

Karan Singhal
Google Research
karansinghal@google.com

Hakim Sidahmed
Google Research
hsidahmed@google.com

Zachary Garrett
Google Research
zachgarrett@google.com

Shanshan Wu
Google Research
shanshanw@google.com

Keith Rush
Google Research
krush@google.com

Sushant Prakash
Google Research
sush@google.com

NeurIPS 2021

Optimization: Train personal and shared parameters **alternatingly**

So, how do we personalize a federated model?

Design decisions:

- Modeling
- Optimization

Our contributions

1. Theory: Analysis of both these optimization algorithms

Code:

SCAN ME

2. Extensive experiments: text, vision, and speech settings

Outline

1. Setup and review
2. Convergence Analysis
3. Experiments

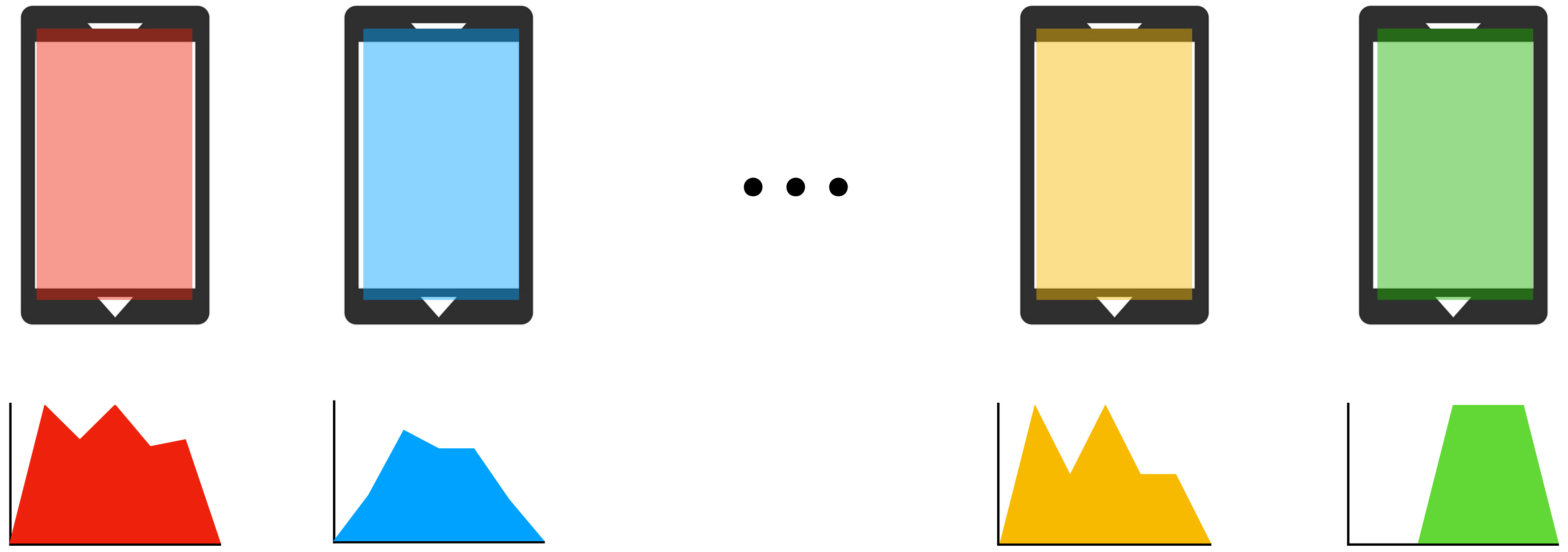
Outline

1. **Setup and review**
2. Convergence Analysis
3. Experiments

(Non-personalized) federated learning

Clients

1 2 … n



Data Distribution

p_1 p_2 p_n

Data heterogeneity

Learning Objective

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

where

$$F_i(w) = \mathbb{E}_{z \sim p_i} \left[f(w; z) \right]$$

loss on client i

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

Personalized federated learning

Model on client i = (u, v_i)

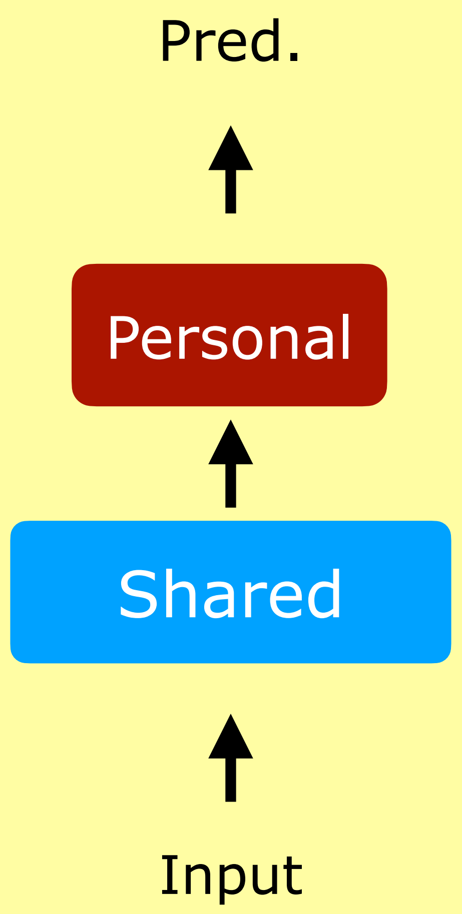
Objective: $$\min_{u, v_1, \cdots, v_n} \frac{1}{n} \sum_{i=1}^{n} F_i(u, v_i)$$

u: shared parameters

v_i: personal parameters

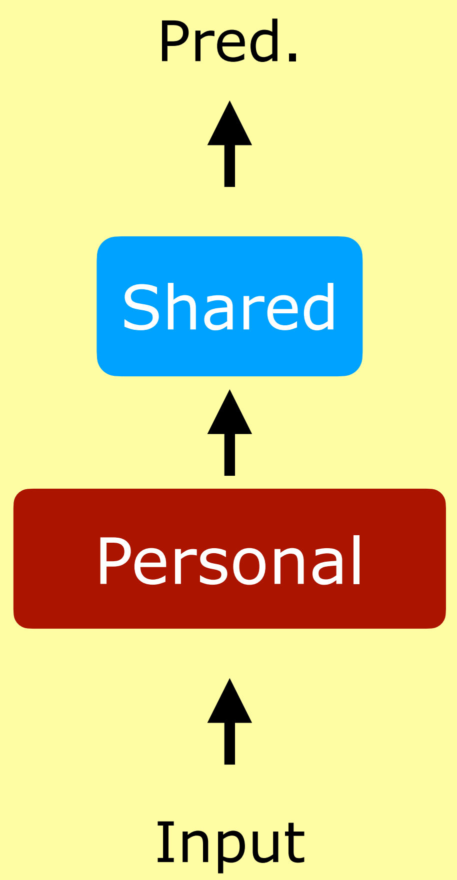
Personalization architectures

Personalized output layer



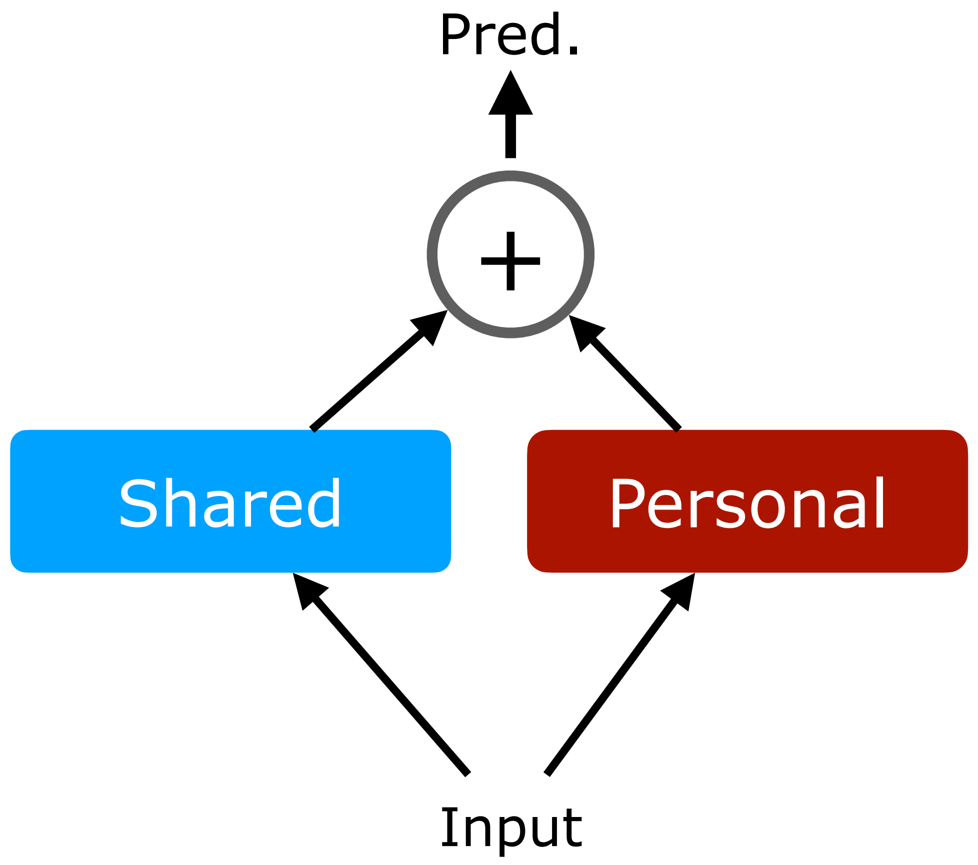
Arivazhagan et al. (2019)
Collins et al. ICML (2021)

Personalized input layer



Liang et al. (2019)

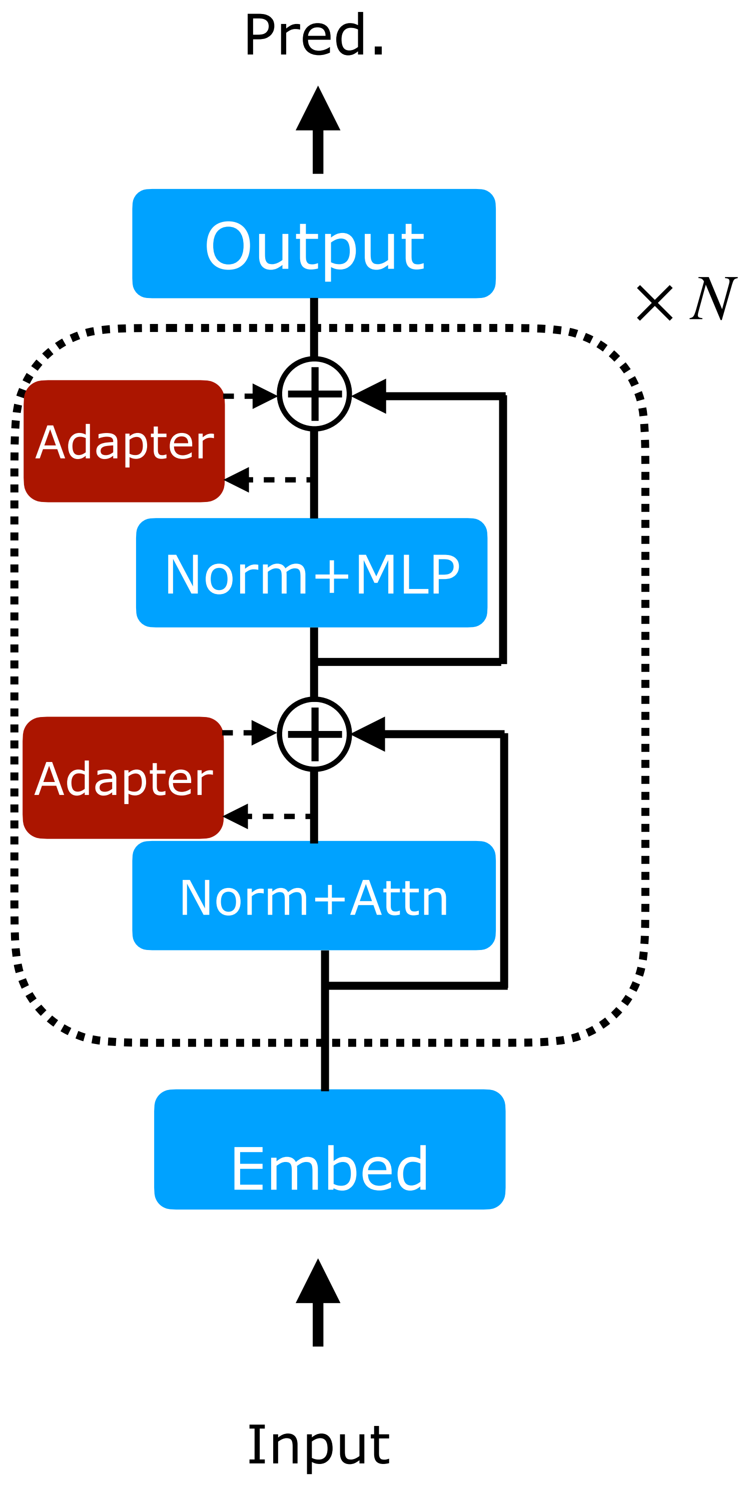
Combined predictions



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

Agarwal et al. (2020)

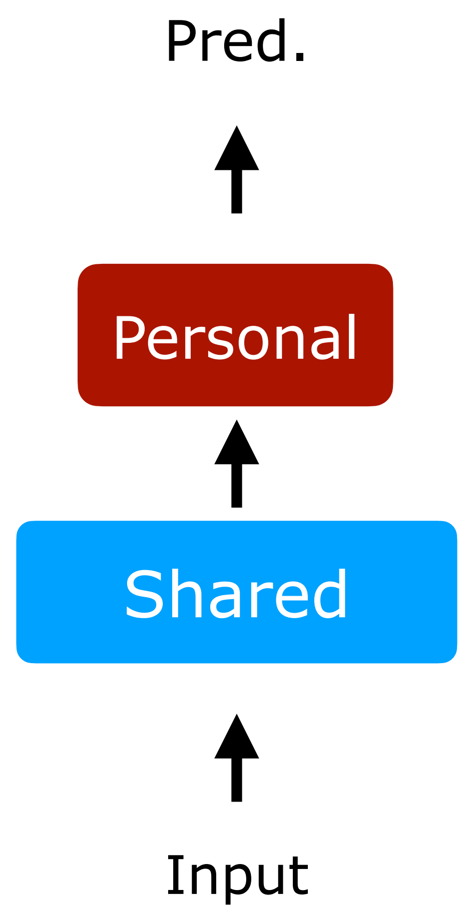
Personalized adapters



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

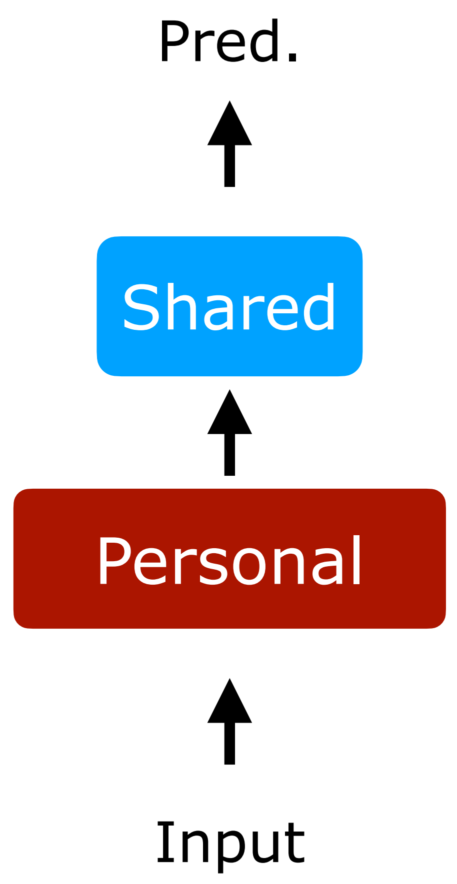
Personalization architectures

Personalized output layer



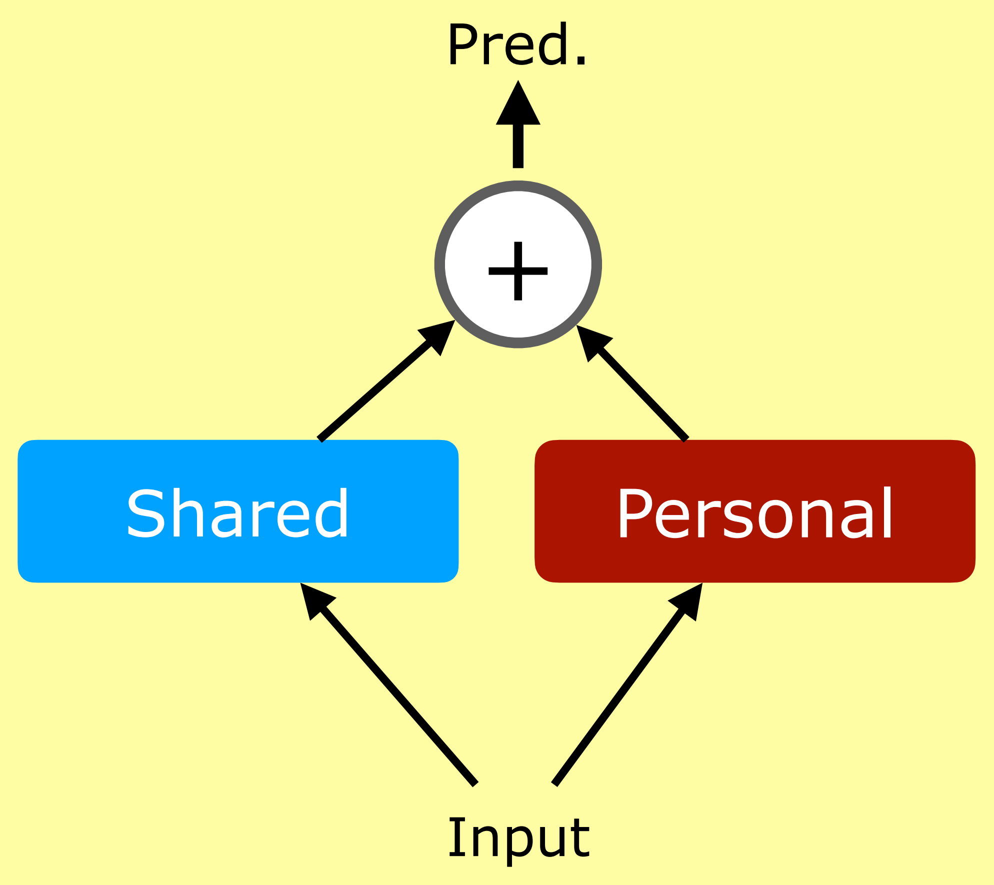
Arivazhagan et al. (2019)
Collins et al. ICML (2021)

Personalized input layer



Liang et al. (2019)

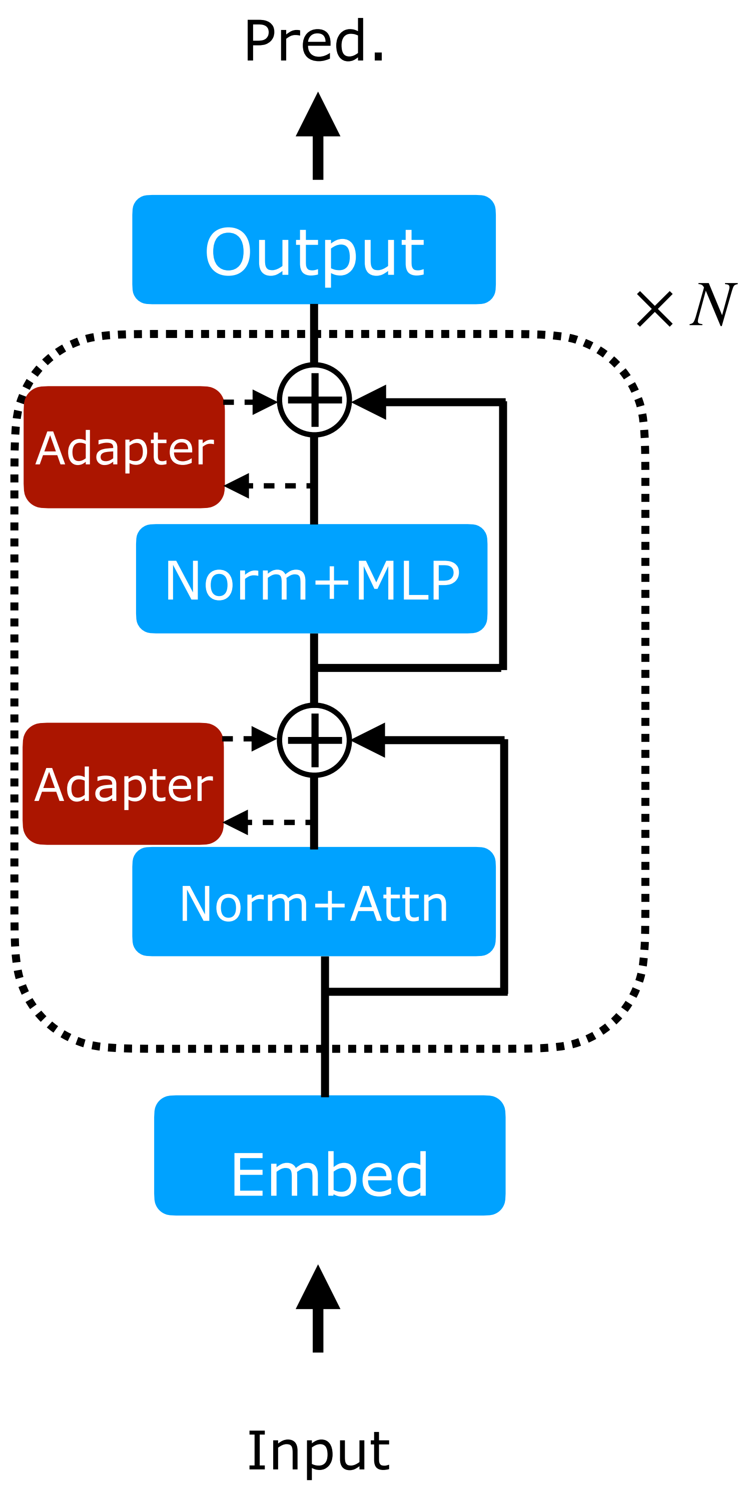
Combined predictions



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

Agarwal et al. (2020)

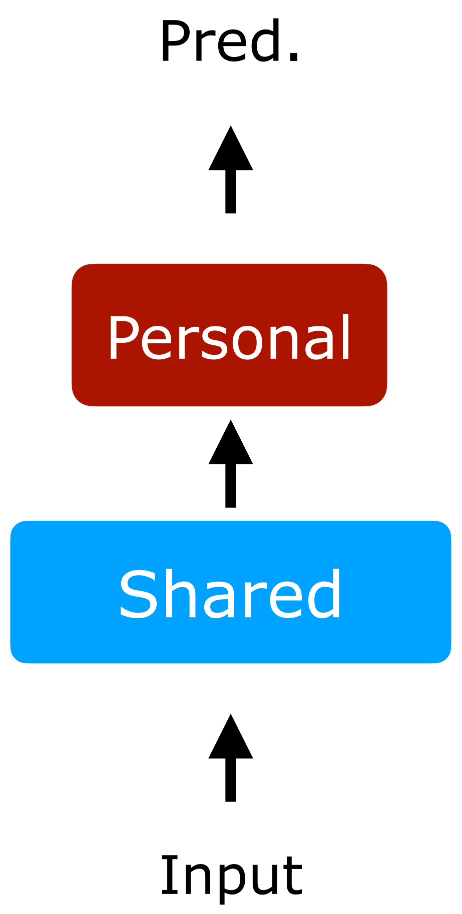
Personalized adapters



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

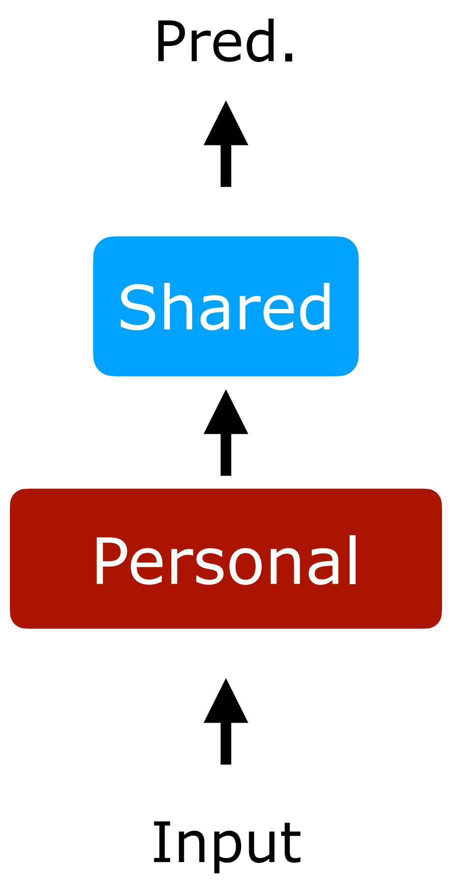
Personalization architectures

Personalized output layer



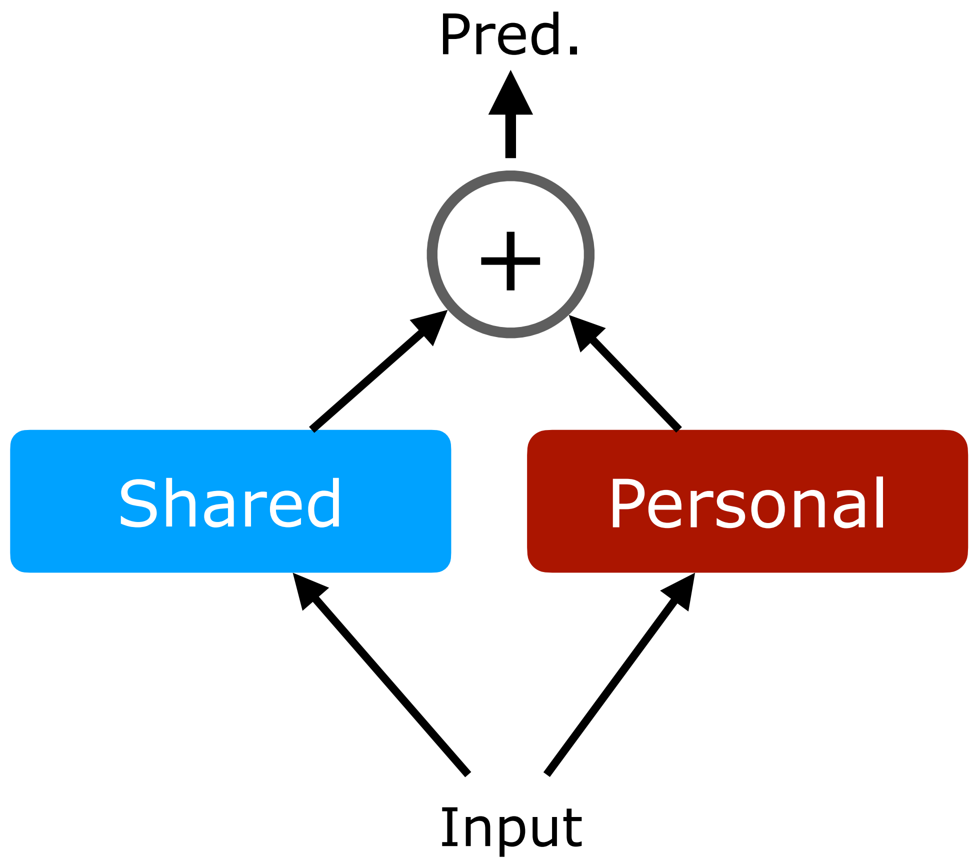
Arivazhagan et al. (2019)
Collins et al. ICML (2021)

Personalized input layer



Liang et al. (2019)

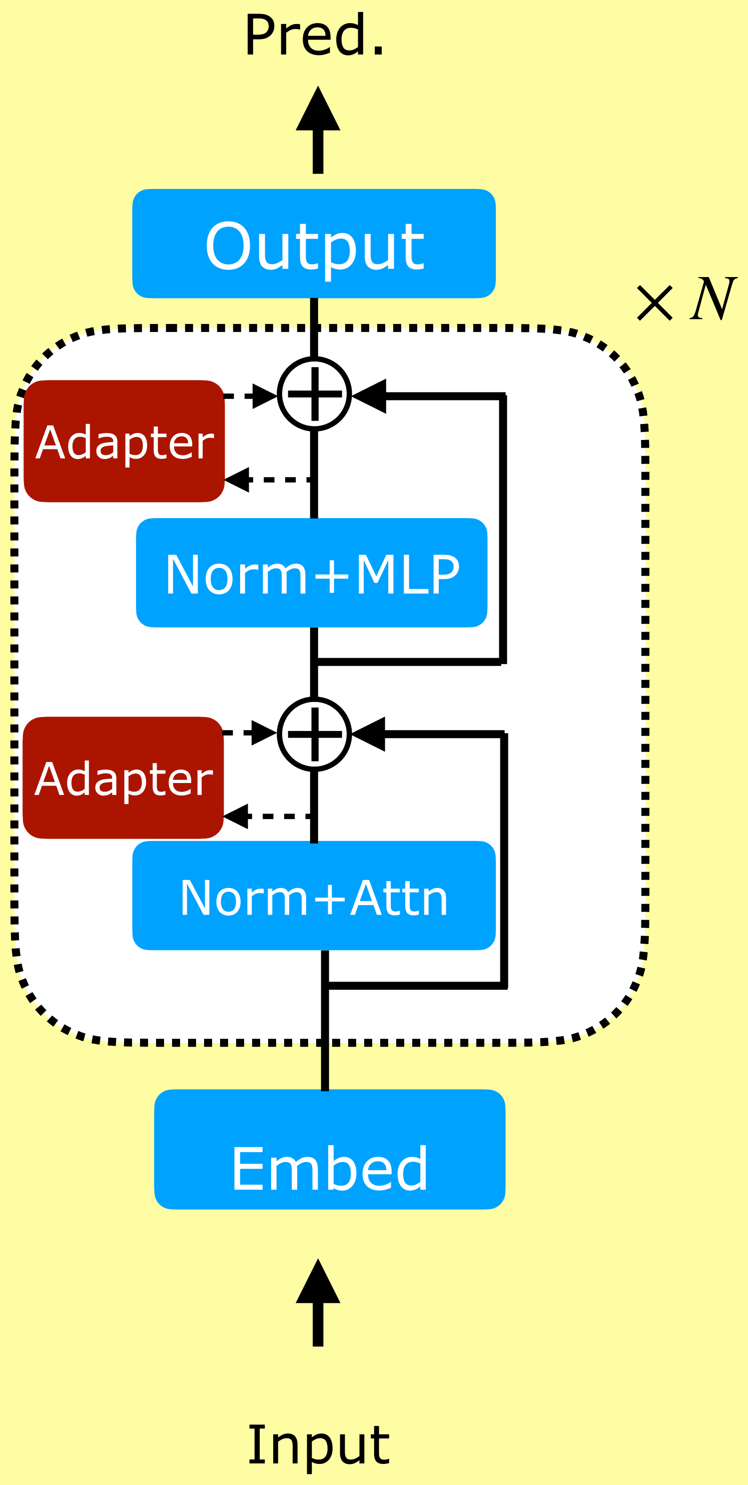
Combined predictions



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

Agarwal et al. (2020)

Personalized adapters



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

Other forms of personalization

pFedMe: $$\min_{u, v_1, \cdots, v_n} \frac{1}{n} \sum_{i=1}^{n} \left(f_i(v_i) + \frac{\lambda}{2} \|v_i - u\|^2 \right)$$

[Dinh et. al (NeurIPS 2020)]

Ditto, MAML, APFL, …. [Hanzely et al. (2021)]

Non-personalized (FedAvg)

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

Personalized (FedAlt/FedSim)

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

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)]

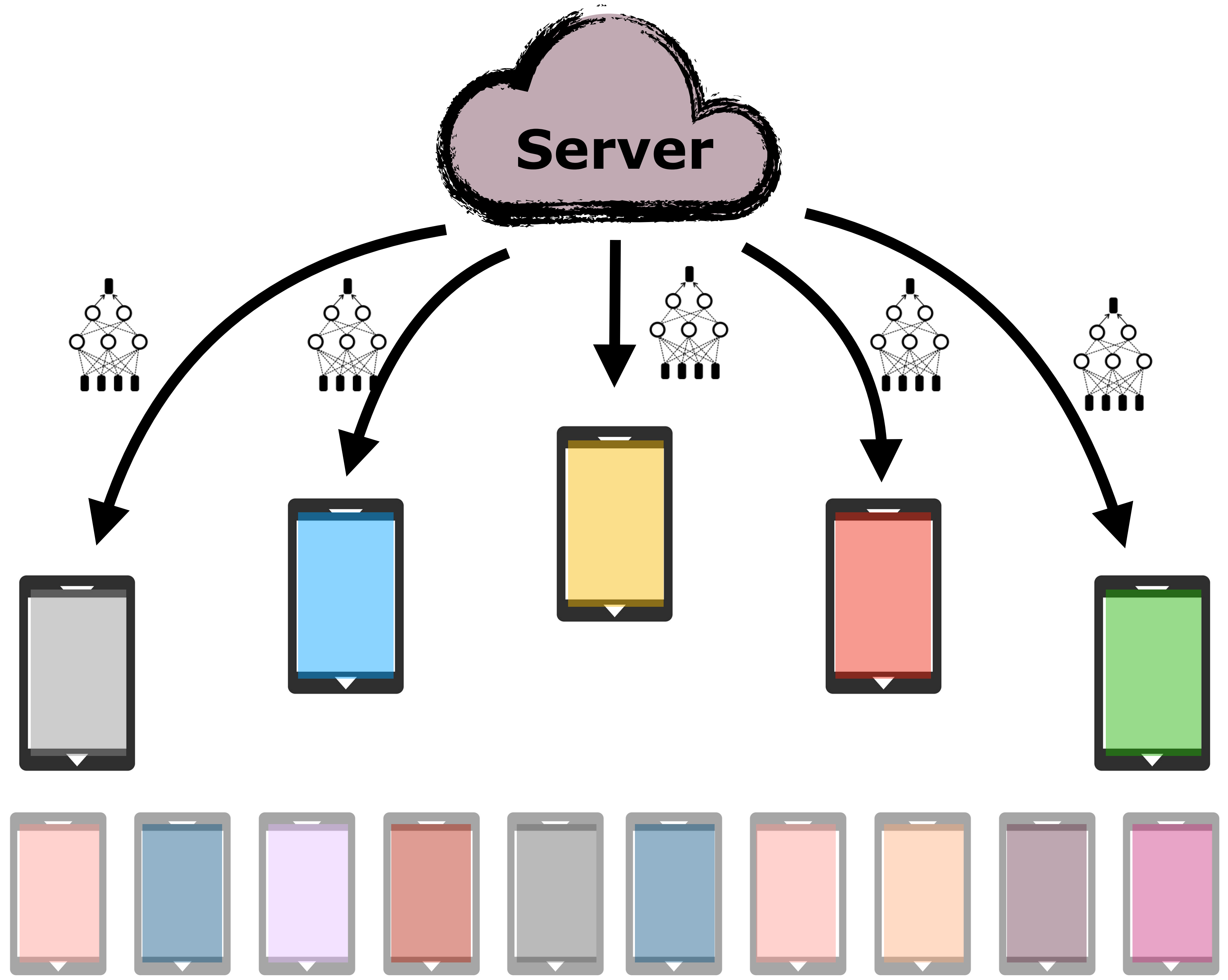
Non-personalized (FedAvg)

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

Personalized (FedAlt/FedSim)

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

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



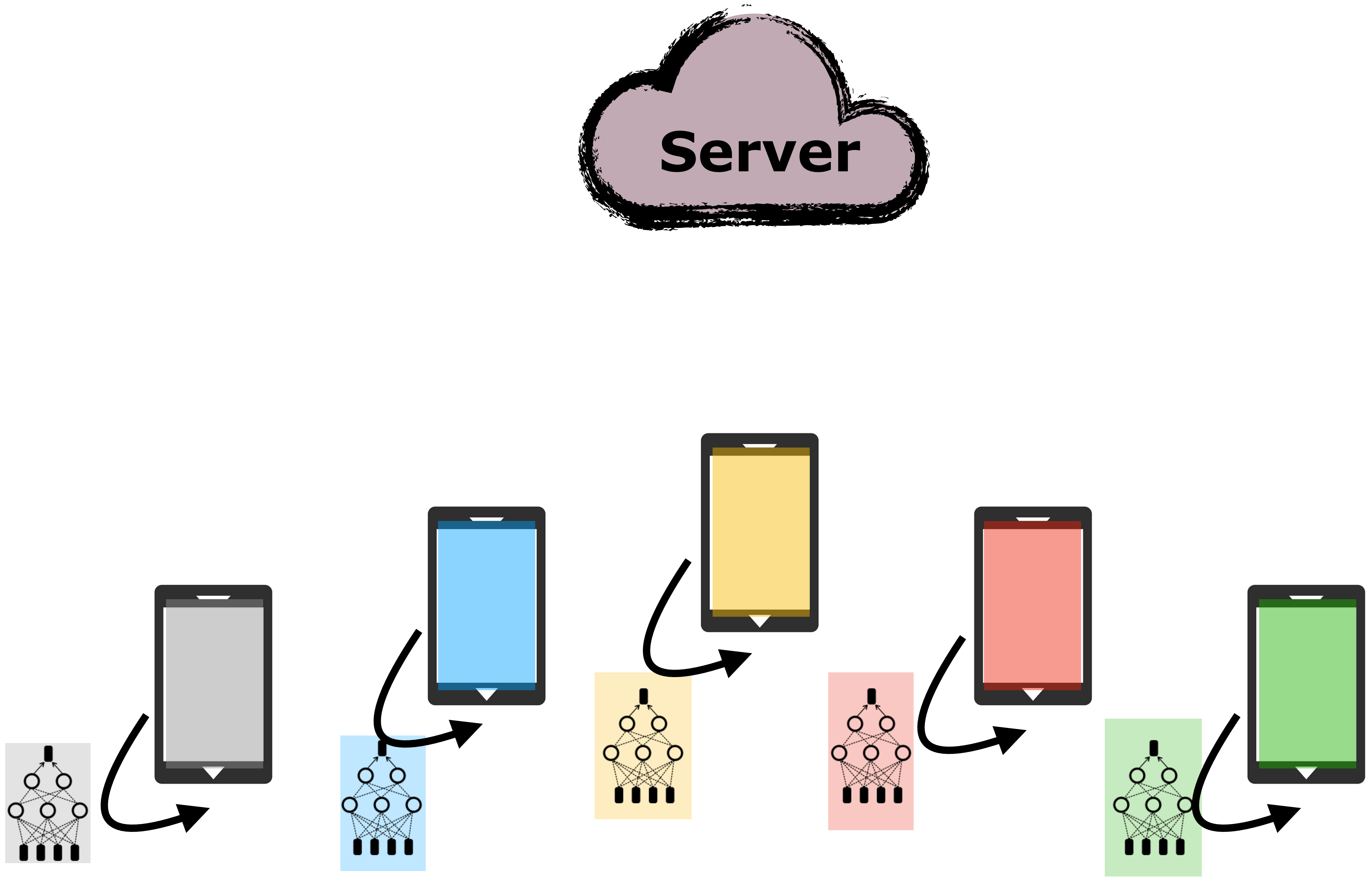
Non-personalized (FedAvg)

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

Personalized (FedAlt/FedSim)

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

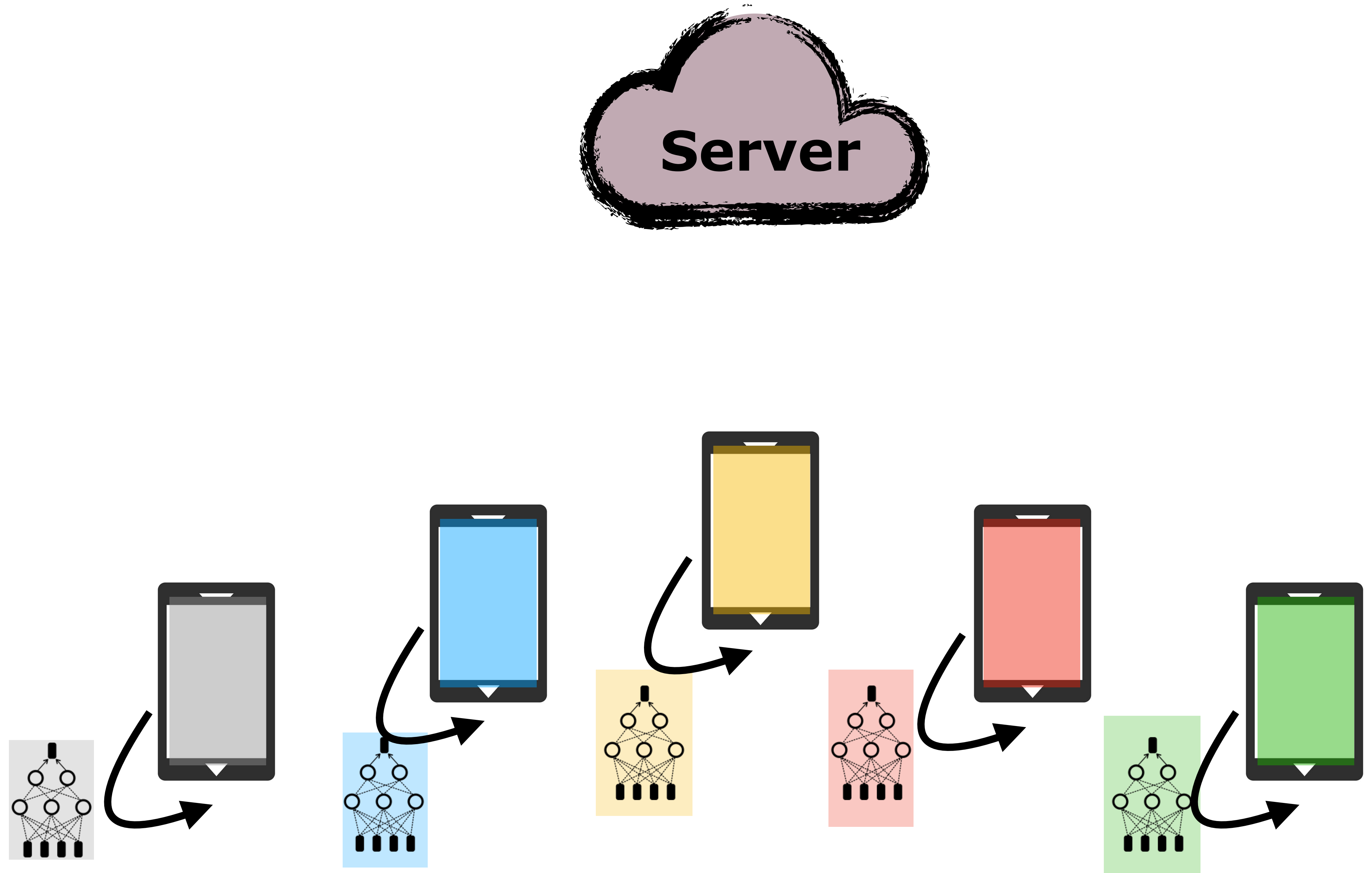
Step 2 of 3: Clients perform τ local SGD steps on their local data



Non-personalized (FedAvg)

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

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



Personalized (FedAlt/FedSim)

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

FedAlt (alternating update)

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

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

FedSim (simultaneous update)

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

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

Non-personalized (FedAvg)

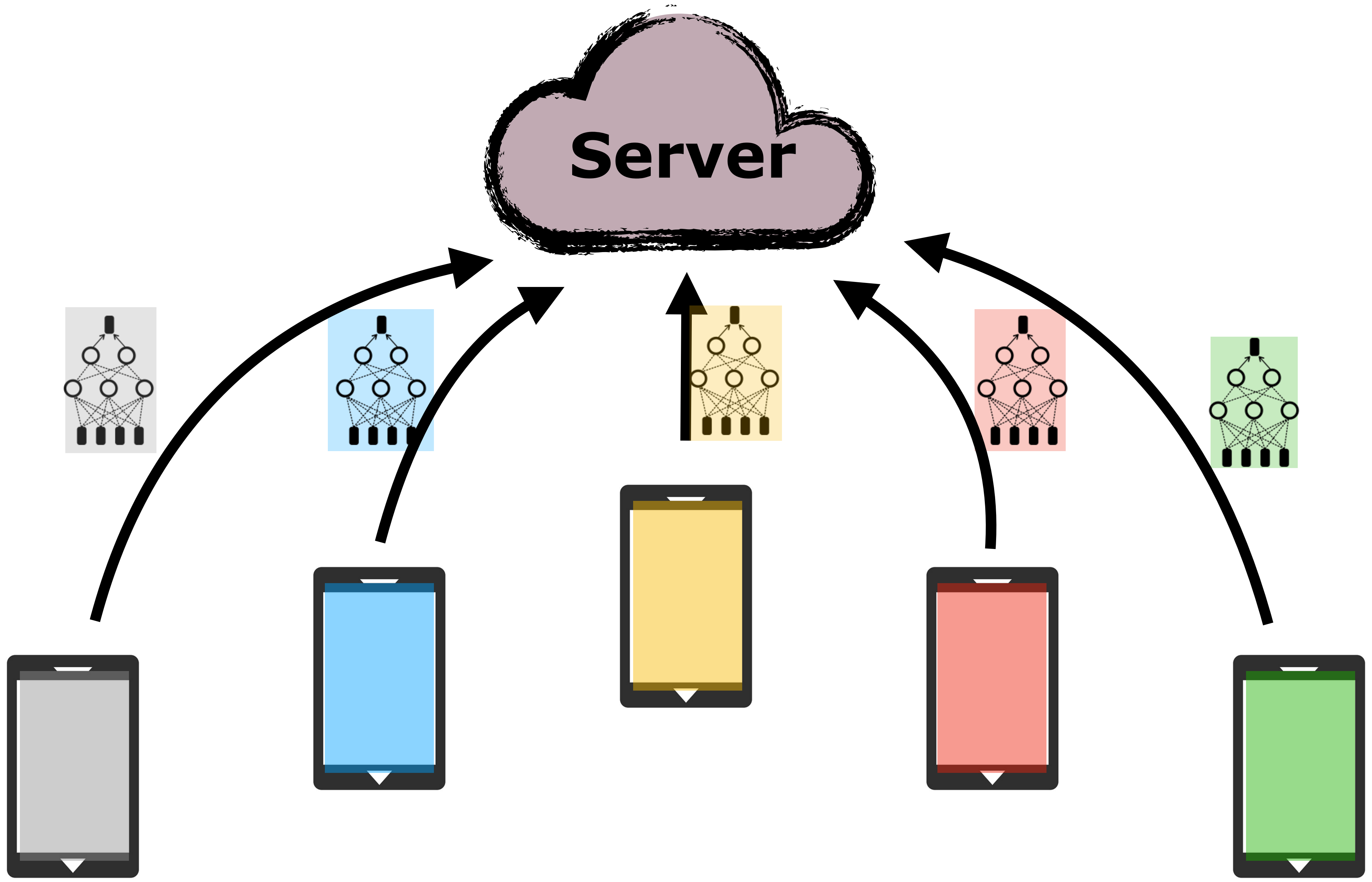
$$\min_{w} \quad \frac{1}{n}\sum_{i=1}^{n} F_i(w)$$

Personalized (FedAlt/FedSim)

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

Step 3 of 3: Aggregate (shared components) of client updates

$$w^+ = \frac{1}{m}\sum_i w_i^+$$



$$u^+ = \frac{1}{m}\sum_i u_i^+$$

v_i stays on client i

Outline

1. Setup and review

2. Convergence Analysis

3. Experiments

Assumptions

Model on client i = (u, v_i)

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

u: shared parameters

v_i: personal parameters

1. Smoothness

$\nabla_u F_i$ is $\begin{cases} L_u\text{-Lipschitz w.r.t. } u \\ L_{uv}\text{-Lipschitz w.r.t. } v_i \end{cases}$

$\nabla_v F_i$ is $\begin{cases} L_v\text{-Lipschitz w.r.t. } v_i \\ L_{uv}\text{-Lipschitz w.r.t. } u \end{cases}$

$\chi^2 := \frac{L_{uv}^2}{L_u L_v}$ quantifies cross-dependence

Assumptions

Model on client i = (u, v_i)

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

u: shared parameters

v_i: personal parameters

2. Bounded variance

- stochastic gradients of $\nabla_u F_i$ and $\nabla_v F_i$ have bounded variance σ_u^2 and σ_v^2 respectively

- bounded gradient diversity:

$$\frac{1}{n} \sum_{i=1}^{n} \|\nabla_u F_i(u, v) - \nabla_u F(u, v_{1:n})\|^2 \leq \delta^2$$

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

Under the smoothness and bounded variance assumptions, we have the bounds

FedAlt $$\frac{1}{T}\sum_{t=0}^{T-1}\left(\frac{1}{L_u}\mathbb{E}\,\|\,\nabla_u F(u_t, v_{1:n,t})\,\|^2 + \frac{1}{nL_v}\sum_{i=1}^{n}\mathbb{E}\,\|\,\nabla_v F_i(u_t, v_{i,t})\,\|^2\right) \le \sqrt{\frac{\sigma_1^2}{T}} + \left(\frac{\tilde{\sigma}_1^2}{T}\right)^{2/3} + O\left(\frac{1}{T}\right)$$

FedSim $$\frac{1}{T}\sum_{t=0}^{T-1}\left(\frac{1}{L_u}\mathbb{E}\,\|\,\nabla_u F(u_t, v_{1:n,t})\,\|^2 + \frac{1}{nL_v}\sum_{i=1}^{n}\mathbb{E}\,\|\,\nabla_v F_i(u_t, v_{i,t})\,\|^2\right) \le \sqrt{\frac{\sigma_2^2}{T}} + \left(\frac{\tilde{\sigma}_2^2}{T}\right)^{2/3} + O\left(\frac{1}{T}\right)$$

$\sigma_1^2, \sigma_2^2, \tilde{\sigma}_1^2, \tilde{\sigma}_2^2$ are linear combinations of σ_u^2, σ_v^2, δ^2

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

Under the smoothness and bounded variance assumptions, we have the bounds

FedAlt $$\frac{1}{T}\sum_{t=0}^{T-1}\left(\frac{1}{L_u}\mathbb{E}\left\|\nabla_u F(u_t, v_{1:n,t})\right\|^2 + \frac{1}{nL_v}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla_v F_i(u_t, v_{i,t})\right\|^2\right) \le \sqrt{\frac{\sigma_1^2}{T}} + \left(\frac{\tilde{\sigma}_1^2}{T}\right)^{2/3} + O\left(\frac{1}{T}\right)$$

FedSim $$\frac{1}{T}\sum_{t=0}^{T-1}\left(\frac{1}{L_u}\mathbb{E}\left\|\nabla_u F(u_t, v_{1:n,t})\right\|^2 + \frac{1}{nL_v}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla_v F_i(u_t, v_{i,t})\right\|^2\right) \le \sqrt{\frac{\sigma_2^2}{T}} + \left(\frac{\tilde{\sigma}_2^2}{T}\right)^{2/3} + O\left(\frac{1}{T}\right)$$

$\sigma_1^2, \sigma_2^2, \tilde{\sigma}_1^2, \tilde{\sigma}_2^2$ are linear combinations of $\sigma_u^2,\ \sigma_v^2,\ \delta^2$

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

Under the smoothness and bounded variance assumptions, we have the bounds

FedAlt

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(\frac{1}{L_u}\mathbb{E}\,\left\|\nabla_u F(u_t, v_{1:n,t})\right\|^2 + \frac{1}{nL_v}\sum_{i=1}^{n}\mathbb{E}\,\left\|\nabla_v F_i(u_t, v_{i,t})\right\|^2\right) \le \sqrt{\frac{\sigma_1^2}{T}} + \left(\frac{\tilde{\sigma}_1^2}{T}\right)^{2/3} + O\left(\frac{1}{T}\right)$$

FedSim

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(\frac{1}{L_u}\mathbb{E}\,\left\|\nabla_u F(u_t, v_{1:n,t})\right\|^2 + \frac{1}{nL_v}\sum_{i=1}^{n}\mathbb{E}\,\left\|\nabla_v F_i(u_t, v_{i,t})\right\|^2\right) \le \sqrt{\frac{\sigma_2^2}{T}} + \left(\frac{\tilde{\sigma}_2^2}{T}\right)^{2/3} + O\left(\frac{1}{T}\right)$$

$\sigma_1^2, \sigma_2^2, \tilde{\sigma}_1^2, \tilde{\sigma}_2^2$ are linear combinations of σ_u^2, σ_v^2, δ^2

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

Under the smoothness and bounded variance assumptions, we have the bounds

FedAlt $$\frac{1}{T}\sum_{t=0}^{T-1}\left(\frac{1}{L_u}\mathbb{E}\,\left\| \nabla_u F(u_t, v_{1:n,t}) \right\|^2 + \frac{1}{nL_v}\sum_{i=1}^{n}\mathbb{E}\,\left\| \nabla_v F_i(u_t, v_{i,t}) \right\|^2\right) \le \sqrt{\frac{\sigma_1^2}{T}} + \left(\frac{\tilde{\sigma}_1^2}{T}\right)^{2/3} + O\left(\frac{1}{T}\right)$$

FedSim $$\frac{1}{T}\sum_{t=0}^{T-1}\left(\frac{1}{L_u}\mathbb{E}\,\left\| \nabla_u F(u_t, v_{1:n,t}) \right\|^2 + \frac{1}{nL_v}\sum_{i=1}^{n}\mathbb{E}\,\left\| \nabla_v F_i(u_t, v_{i,t}) \right\|^2\right) \le \sqrt{\frac{\sigma_2^2}{T}} + \left(\frac{\tilde{\sigma}_2^2}{T}\right)^{2/3} + O\left(\frac{1}{T}\right)$$

$\sigma_1^2, \sigma_2^2, \tilde{\sigma}_1^2, \tilde{\sigma}_2^2$ are linear combinations of $\sigma_u^2,\ \sigma_v^2,\ \delta^2$

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

Under the smoothness and bounded variance assumptions, we have the bounds

FedAlt $$\frac{1}{T}\sum_{t=0}^{T-1}\left(\frac{1}{L_u}\mathbb{E}\,\|\,\nabla_u F(u_t, v_{1:n,t})\,\|^2 + \frac{1}{nL_v}\sum_{i=1}^{n}\mathbb{E}\,\|\,\nabla_v F_i(u_t, v_{i,t})\,\|^2\right) \leq \sqrt{\frac{\sigma_1^2}{T}} + \left(\frac{\tilde{\sigma}_1^2}{T}\right)^{2/3} + O\left(\frac{1}{T}\right)$$

FedSim $$\frac{1}{T}\sum_{t=0}^{T-1}\left(\frac{1}{L_u}\mathbb{E}\,\|\,\nabla_u F(u_t, v_{1:n,t})\,\|^2 + \frac{1}{nL_v}\sum_{i=1}^{n}\mathbb{E}\,\|\,\nabla_v F_i(u_t, v_{i,t})\,\|^2\right) \leq \sqrt{\frac{\sigma_2^2}{T}} + \left(\frac{\tilde{\sigma}_2^2}{T}\right)^{2/3} + O\left(\frac{1}{T}\right)$$

$\sigma_1^2, \sigma_2^2, \tilde{\sigma}_1^2, \tilde{\sigma}_2^2$ are linear combinations of σ_u^2, σ_v^2, δ^2

FedAlt is better than FedSim when

$$\frac{\sigma_v^2}{L_v}\left(1 - \frac{2m}{n}\right) \quad < \quad \frac{\sigma_u^2}{mL_u} + \frac{\delta^2}{mL_u}\left(1 - \frac{m}{n}\right)$$

True if $\delta^2 \gg \max\{\sigma_u^2, \sigma_v^2\}$

inter-client variance (δ^2)

intra-client variance ($\max\{\sigma_u^2, \sigma_v^2\}$)

Better by a factor of $(1+\chi^2)^{1/2}$

m: number of clients per round

n: total number of clients

$\sigma_u^2, \sigma_v^2, \delta^2$: noise variances

$\chi^2 = L_{uv}^2 / L_u L_v$: cross-dependency

Technical difficulties

Assume $\sigma_u^2 = 0 = \sigma_v^2$ and single local gradient step per client

For **FedAlt**, apply smoothness for u-step (assuming v-step is complete) to get

$$F(u_{t+1}, v_{t+1}) - F(u_t, v_{t+1}) \leq \langle \nabla_u F(u_t, v_{t+1}), u_{t+1} - u_t \rangle + \frac{L_u}{2} \|u_{t+1} - u_t\|^2$$

both depend on sampling of clients

first-order term is biased!

For **FedSim**, no such difficulties

$$F(u_{t+1}, v_{t+1}) - F(u_t, v_t) \quad \le \quad \langle \nabla_u F(u_t, v_t) \;, u_{t+1} - u_t \rangle + \frac{L_u}{2} \|u_{t+1} - u_t\|^2$$

u-update starts from (u_t, v_t)

only dependence on sampling of clients

first-order term is unbiased!

For **FedAlt**, apply smoothness for u-step (assuming v-step is complete) to get

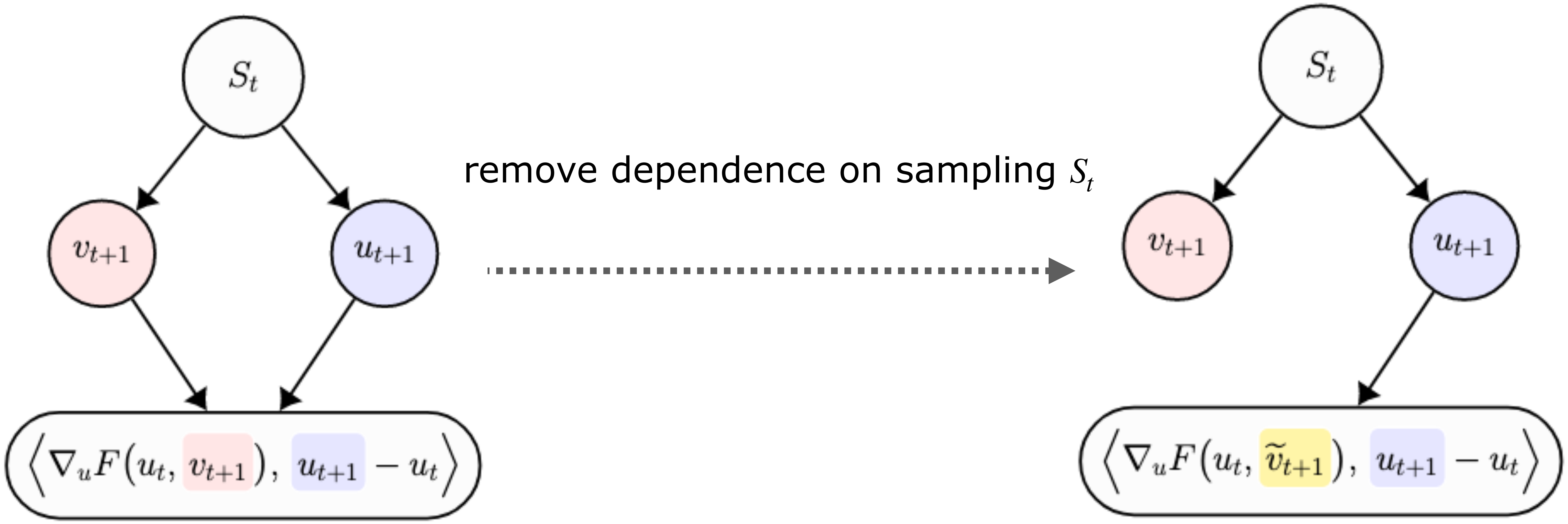
$$F(u_{t+1}, v_{t+1}) - F(u_t, v_{t+1}) \quad \le \quad \langle \nabla_u F(u_t, v_{t+1}) \, , u_{t+1} - u_t \rangle + \frac{L_u}{2} \|u_{t+1} - u_t\|^2$$

both depend on sampling of clients

first-order term is biased!

Virtual full participation

Let $\tilde{v}_t$ denote the (virtual) personal parameters if all clients had run the v-step, not just the selected clients



For **FedAlt**, apply smoothness for u-step (assuming v-step is complete) to get

$$F(u_{t+1}, v_{t+1}) - F(u_t, v_{t+1}) \leq \langle \nabla_u F(u_t, \tilde{v}_{t+1}) \, , u_{t+1} - u_t \rangle + \frac{L_u}{2} \|u_{t+1} - u_t\|^2 + \text{Error}_t$$

independent of sampling of clients

depends on sampling of clients

first-order term is unbiased again!

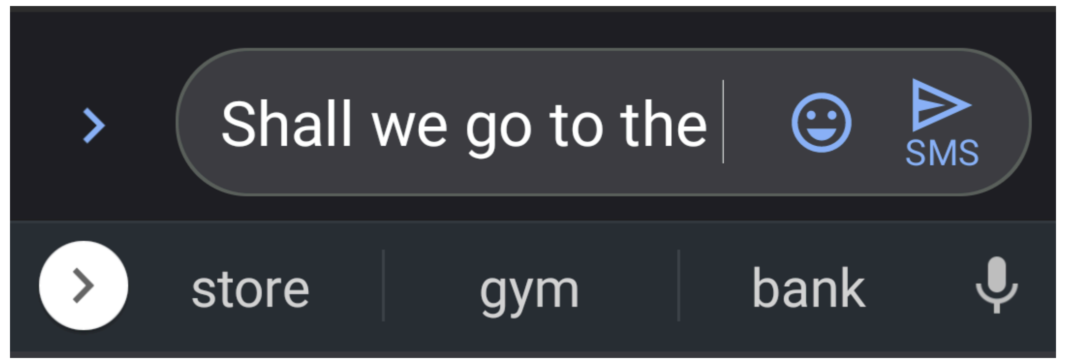
To complete the proof, suffices to bound

$$\mathbb{E}[\text{Error}_t] \quad \leq \quad O(L_u\gamma_u^2 + \chi^2 L_v\gamma_v^2)$$

and can be made smaller by controlling the learning rates γ_u, γ_v

Outline

1. Setup and review
2. Convergence Analysis
3. **Experiments**



Next word prediction

Mobile keyboard

Speech recognition

Mobile assistant

Landmark detection

Mobile camera app

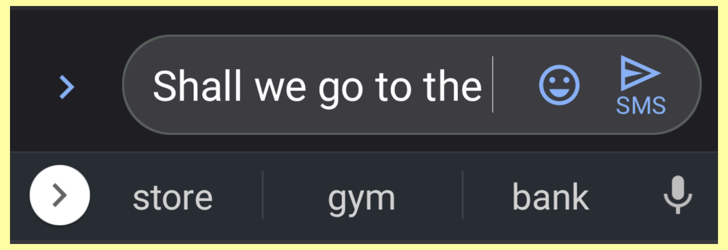
- StackOverflow (~1K clients)
- 4-layer transformer (6M param)
- vocabulary size: 10K

- LibriSpeech dataset (~1K clients)
- 6-layer transformer (15M param)
- CTC Loss (dynamic programming)

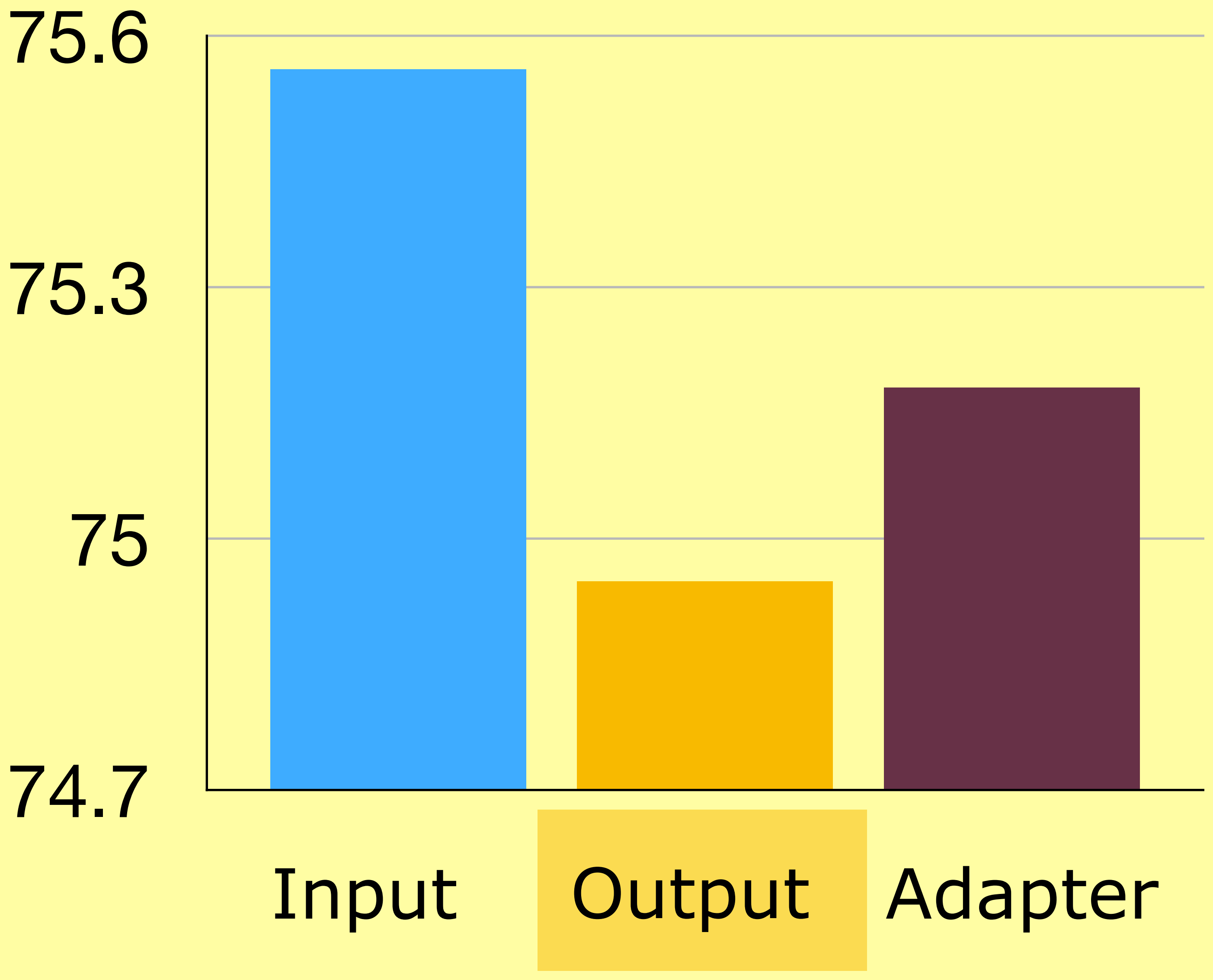
- GLDv2 dataset (~1K clients)
- ResNet-18 (12M param)
- ~2K classes: only 30/client

Question 1: Modeling

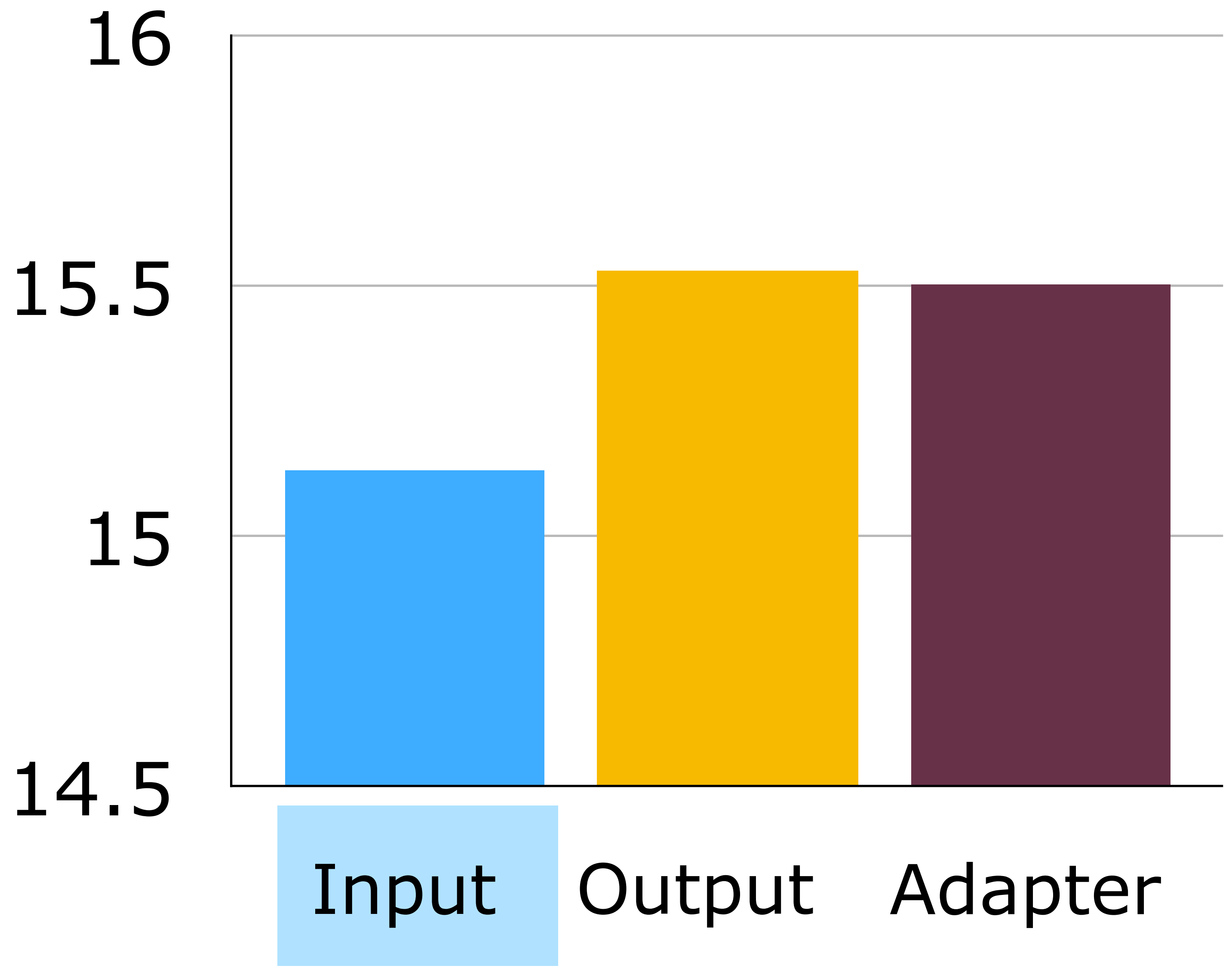
Which form of personalization do I use?



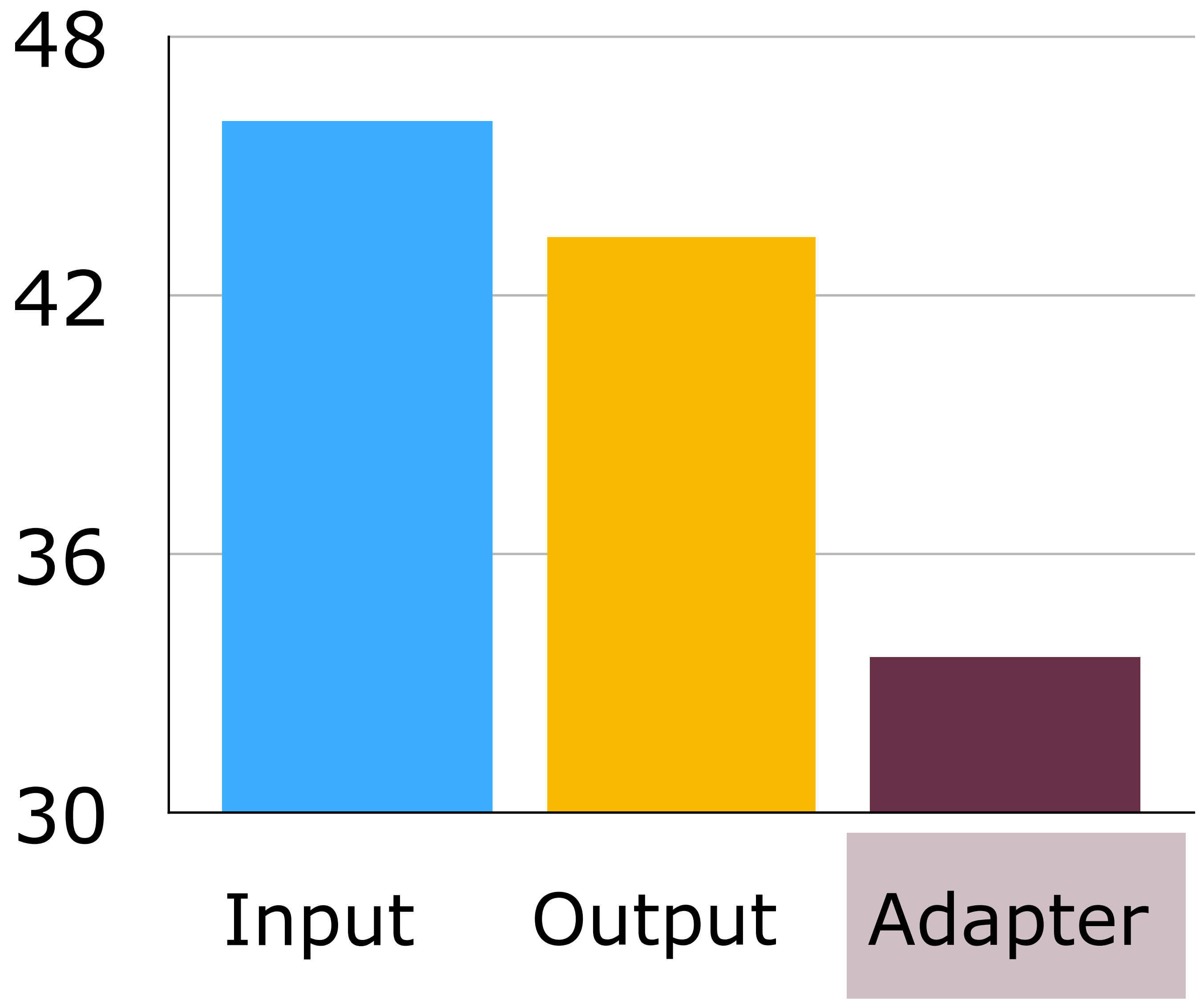
Next word prediction



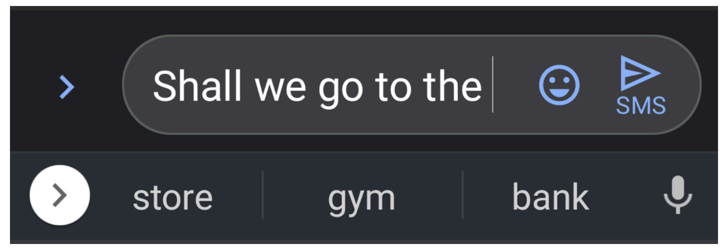
Speech recognition



Landmark detection



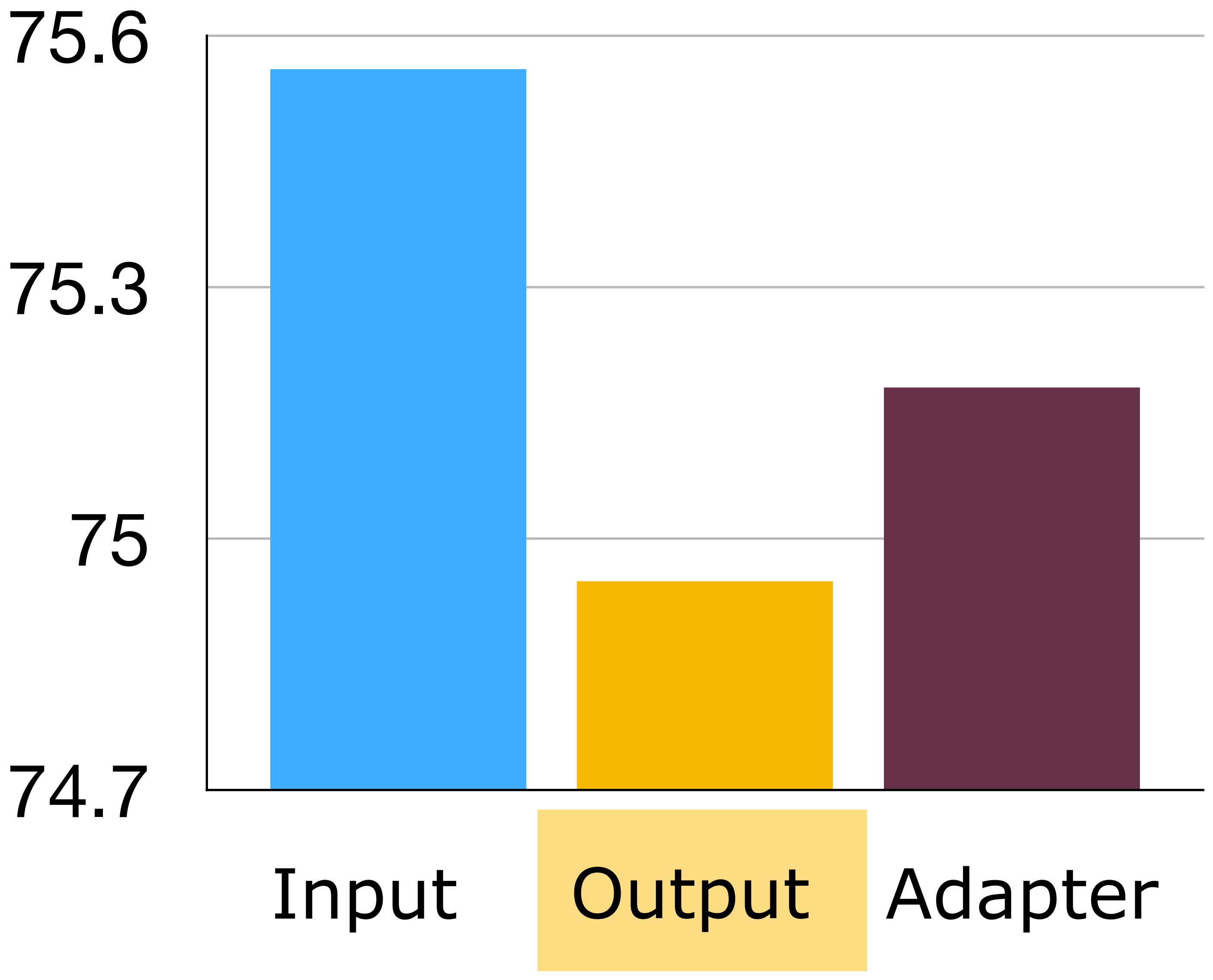
y-axis shows error: lower is better

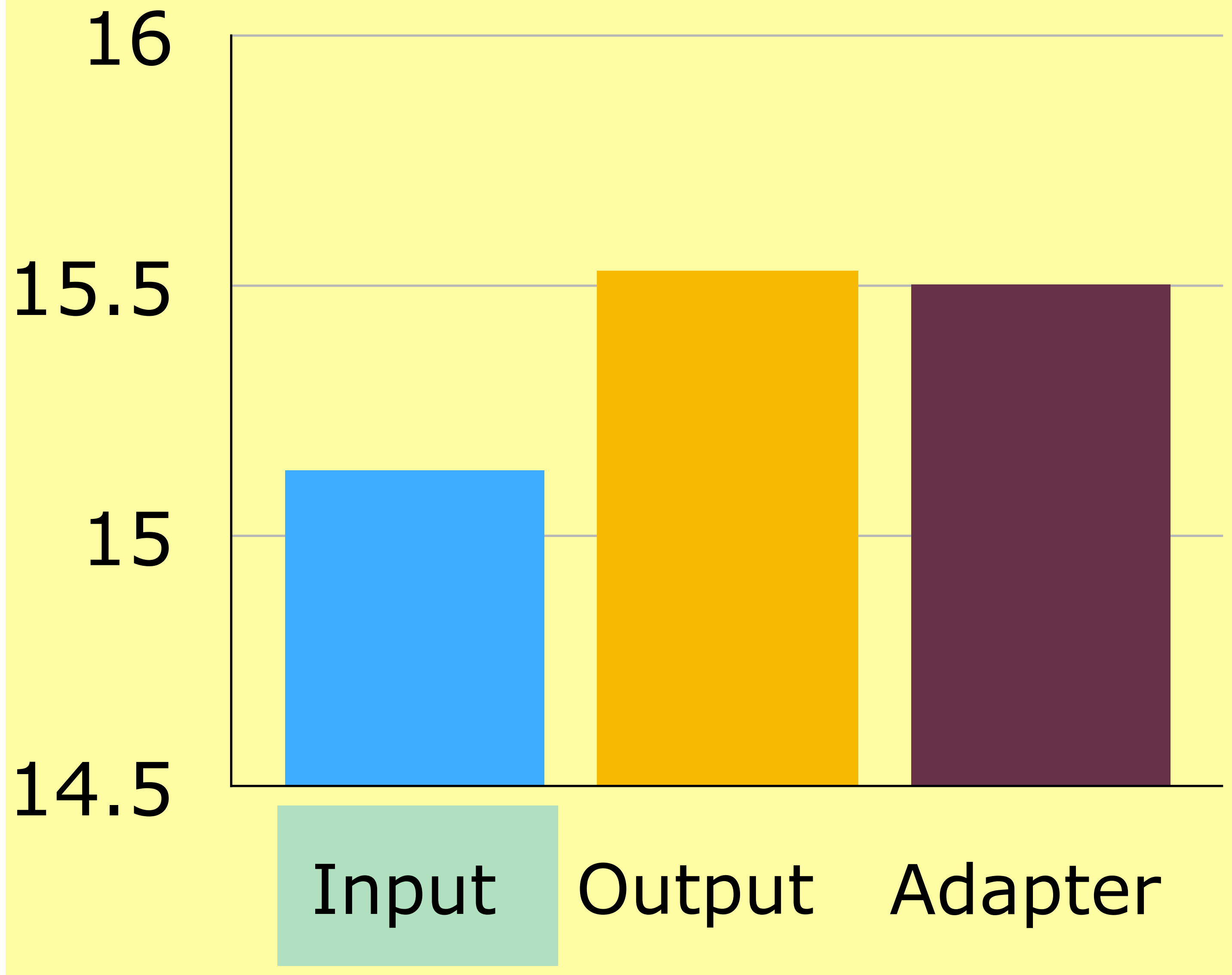


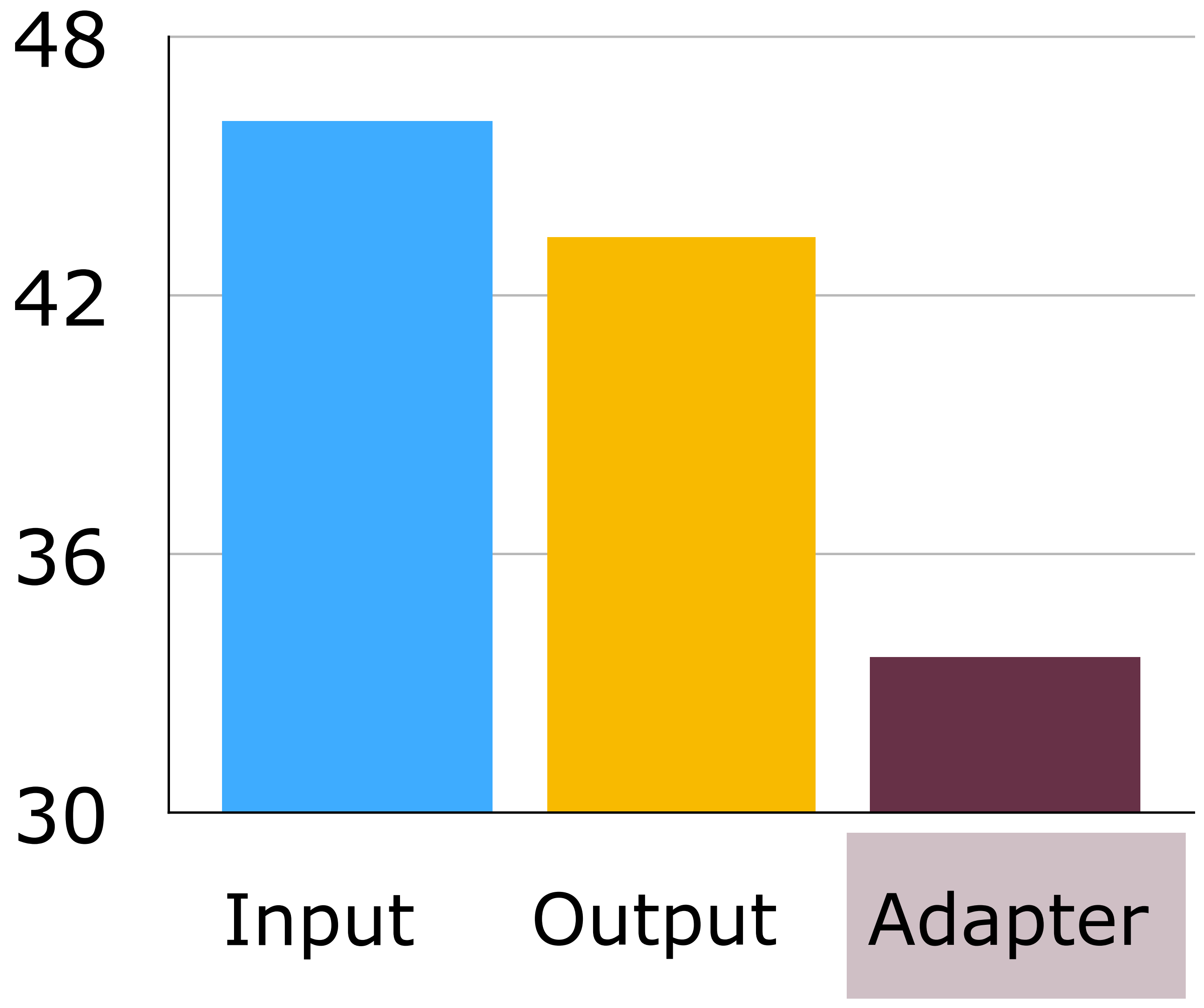
Next word prediction

Speech recognition

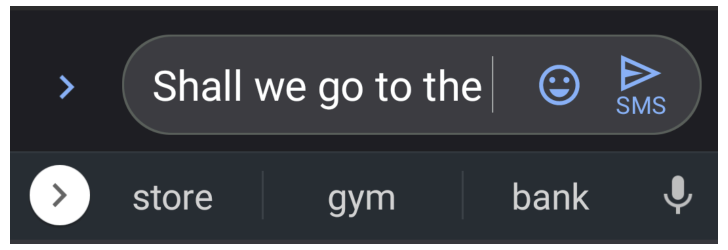
Landmark detection







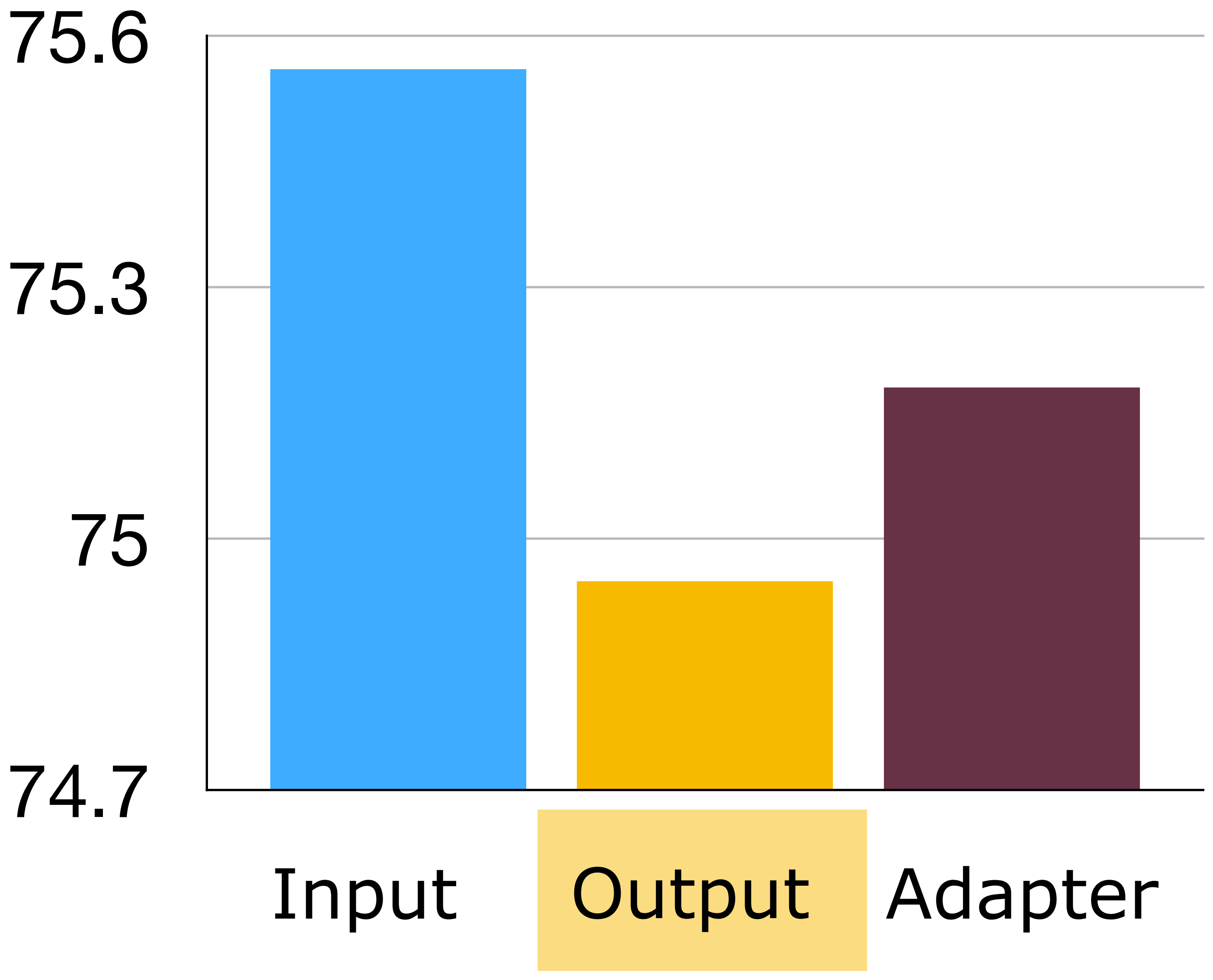
y-axis shows error: lower is better

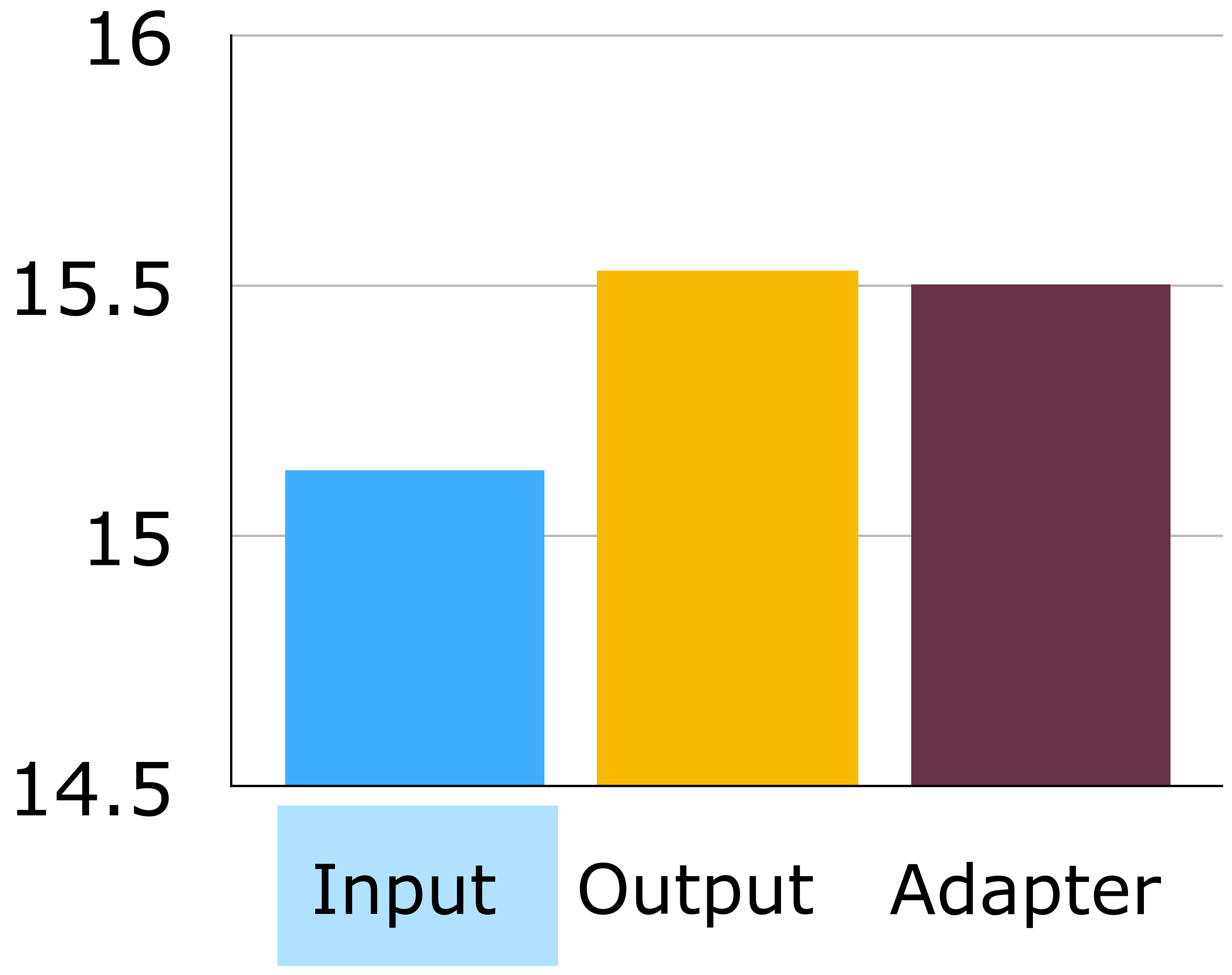


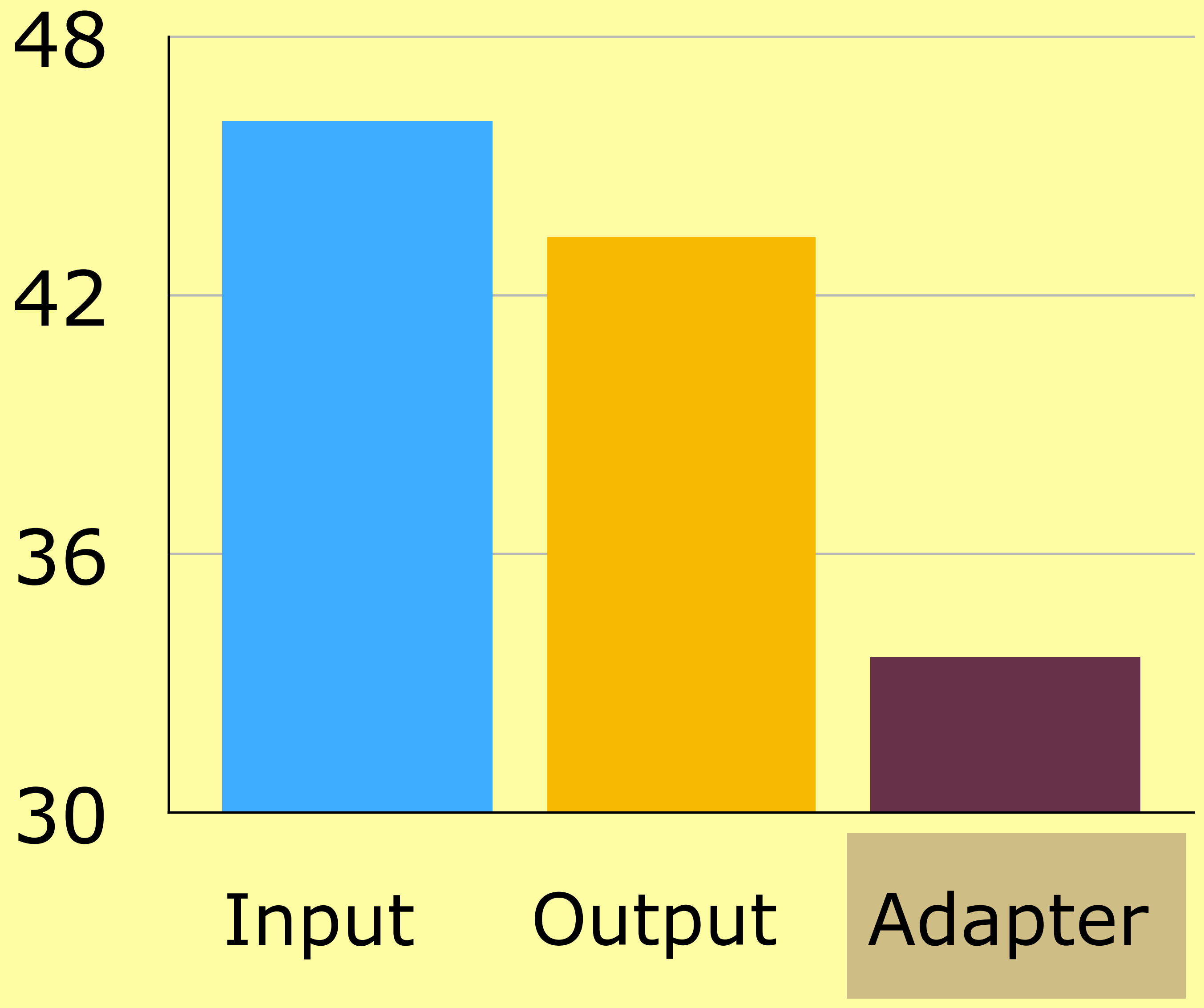
Next word prediction

Speech recognition

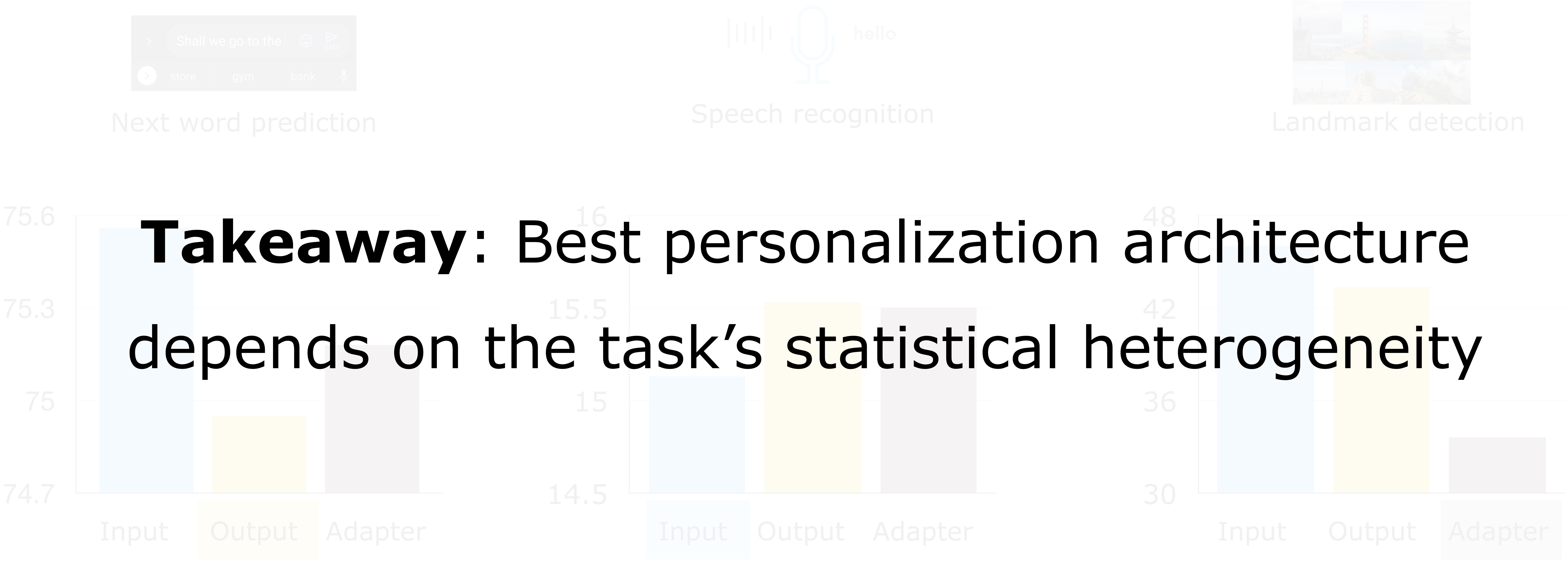
Landmark detection







y-axis shows error: lower is better



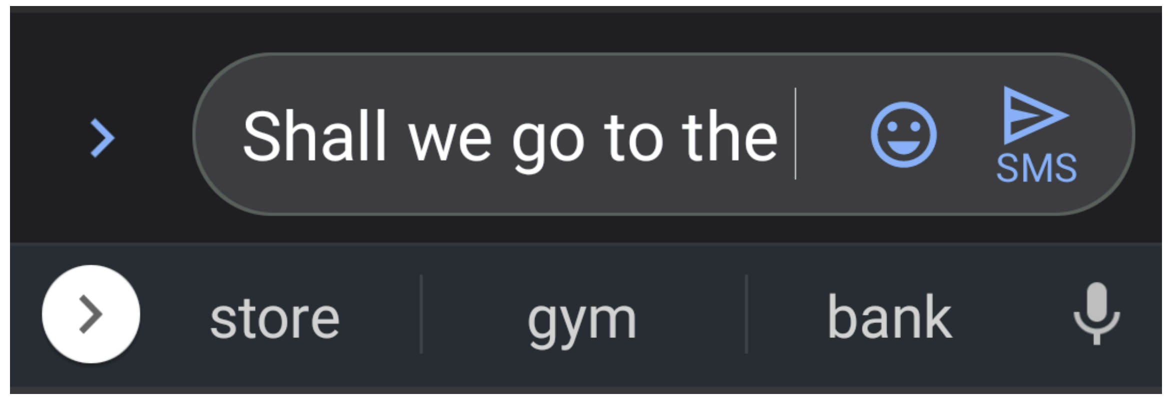
Takeaway: Best personalization architecture depends on the task's statistical heterogeneity

y-axis shows error: lower is better

Partial personalization vs. full personalization

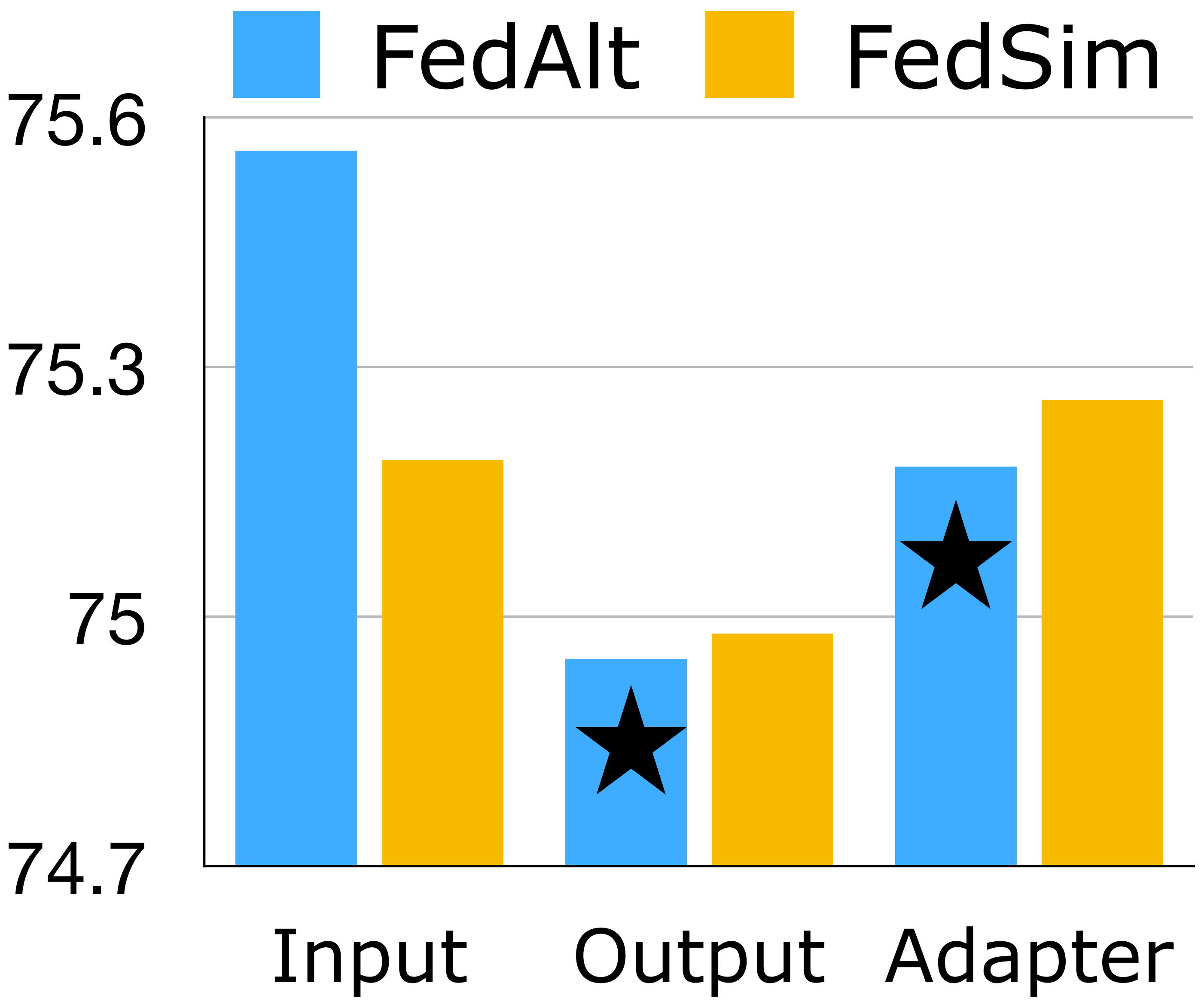
Question 2: Optimization

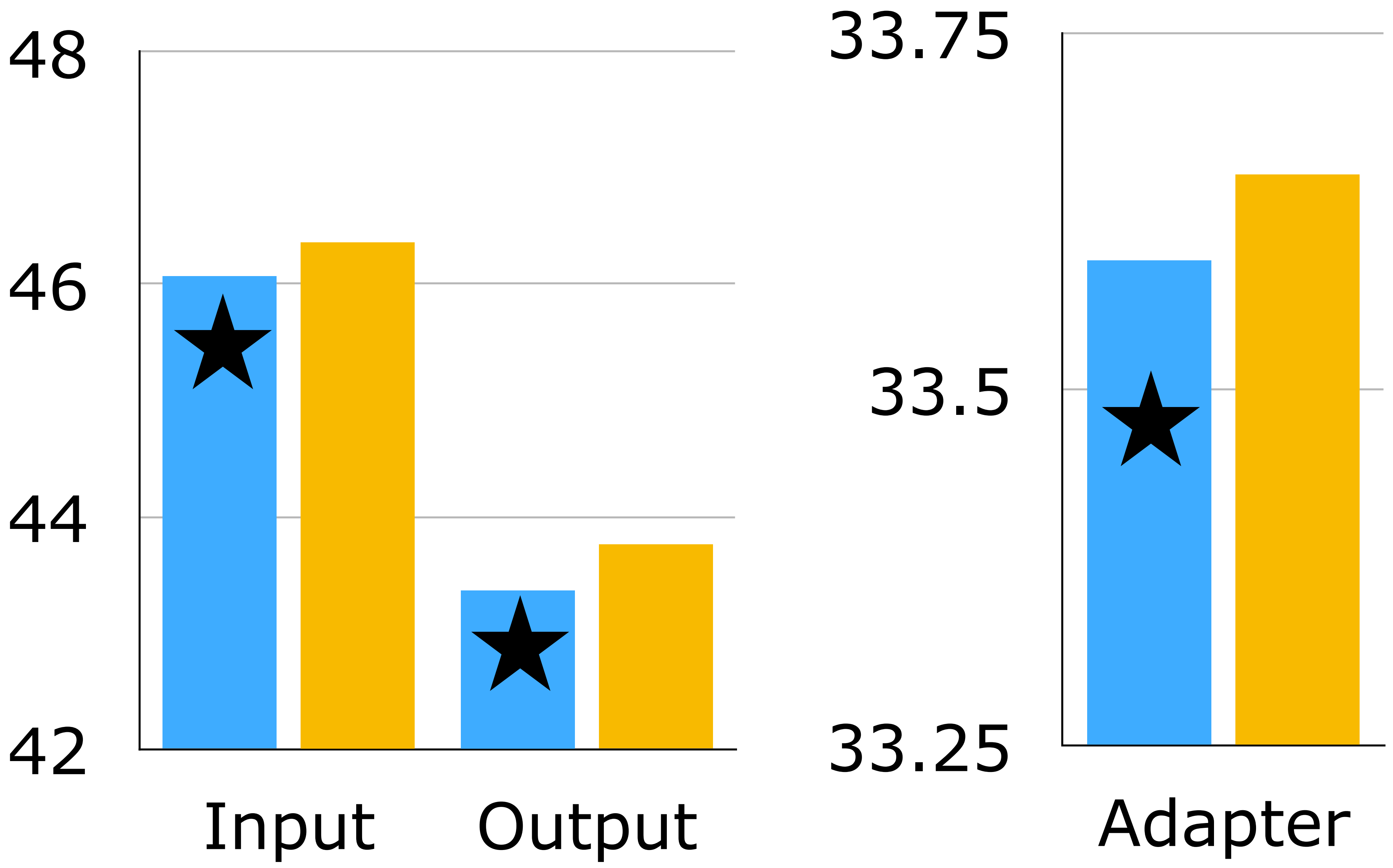
Which optimization algorithm do I use?



Next word prediction

Landmark detection

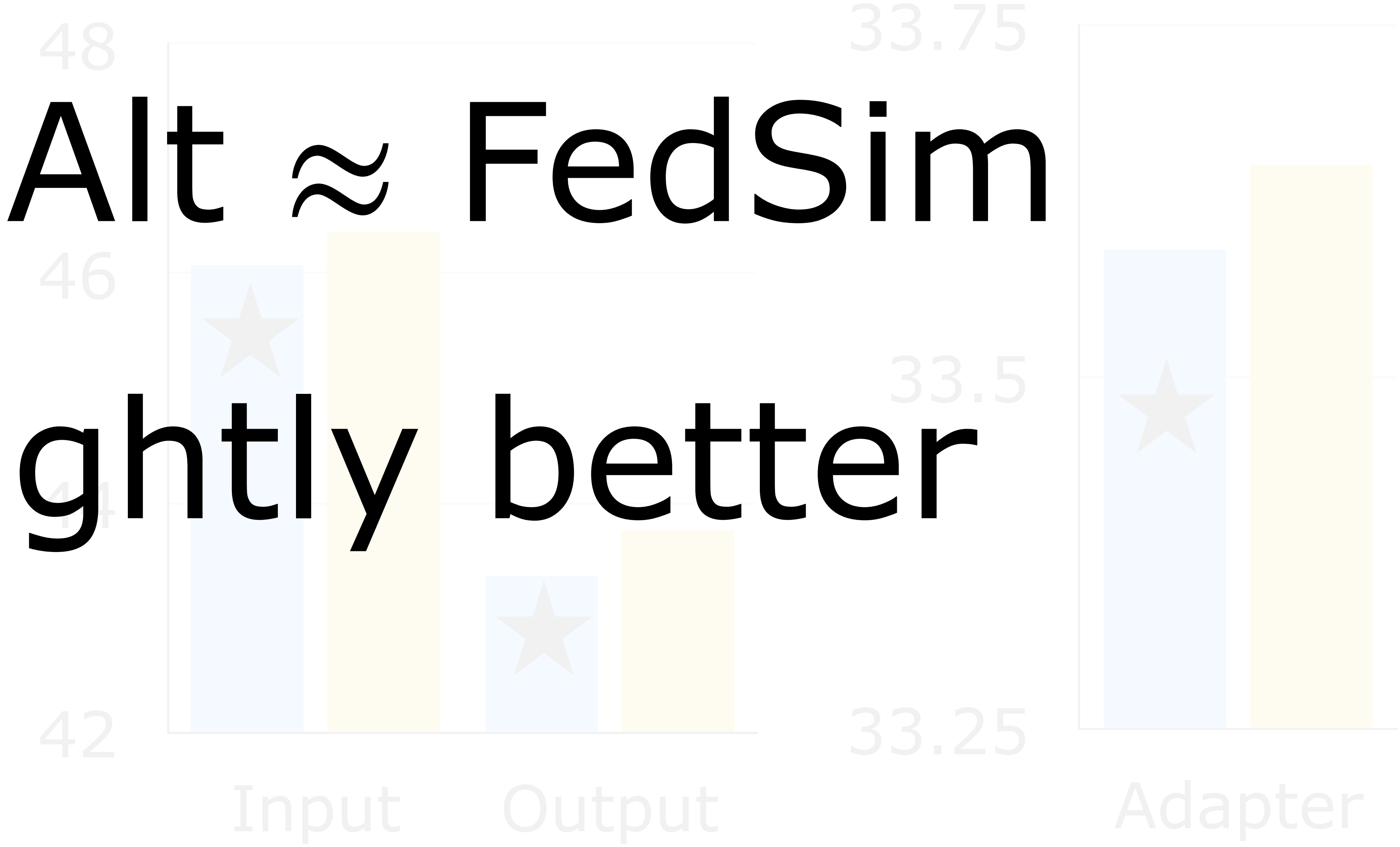




y-axis shows error: lower is better

Next word prediction

Landmark detection



Takeaway: FedAlt ≈ FedSim

But FedAlt is slightly better

y-axis shows error: lower is better

Summary

1. Theory: Analysis of both these optimization algorithms

Code:

SCAN ME

2. Extensive experiments: text, vision, and speech settings

Pillutla, et al. "*Federated Learning with Partial Model Personalization*." ICML 2022.