Federated Learning: Algorithm Design and Applications

Mingrui Liu (Assistant Professor)
Department of Computer Science
George Mason University
mingruil@gmu.edu

April 22, 2023

Outline

- What is Federated Learning?
- Algorithm Design
- Applications
- Ongoing Research and Open Problems

Outline

- What is Federated Learning?
- Algorithm Design
- Applications
- Ongoing Research and Open Problems

- Acknowledgement: Some illustrations are from NeurIPS 2020 Federated Learning Tutorial:
  - Peter Kairouz, Brendan Mcmahan, Virginia Smith. Federated Learning Tutorial (https://sites.google.com/view/fl-tutorial/)

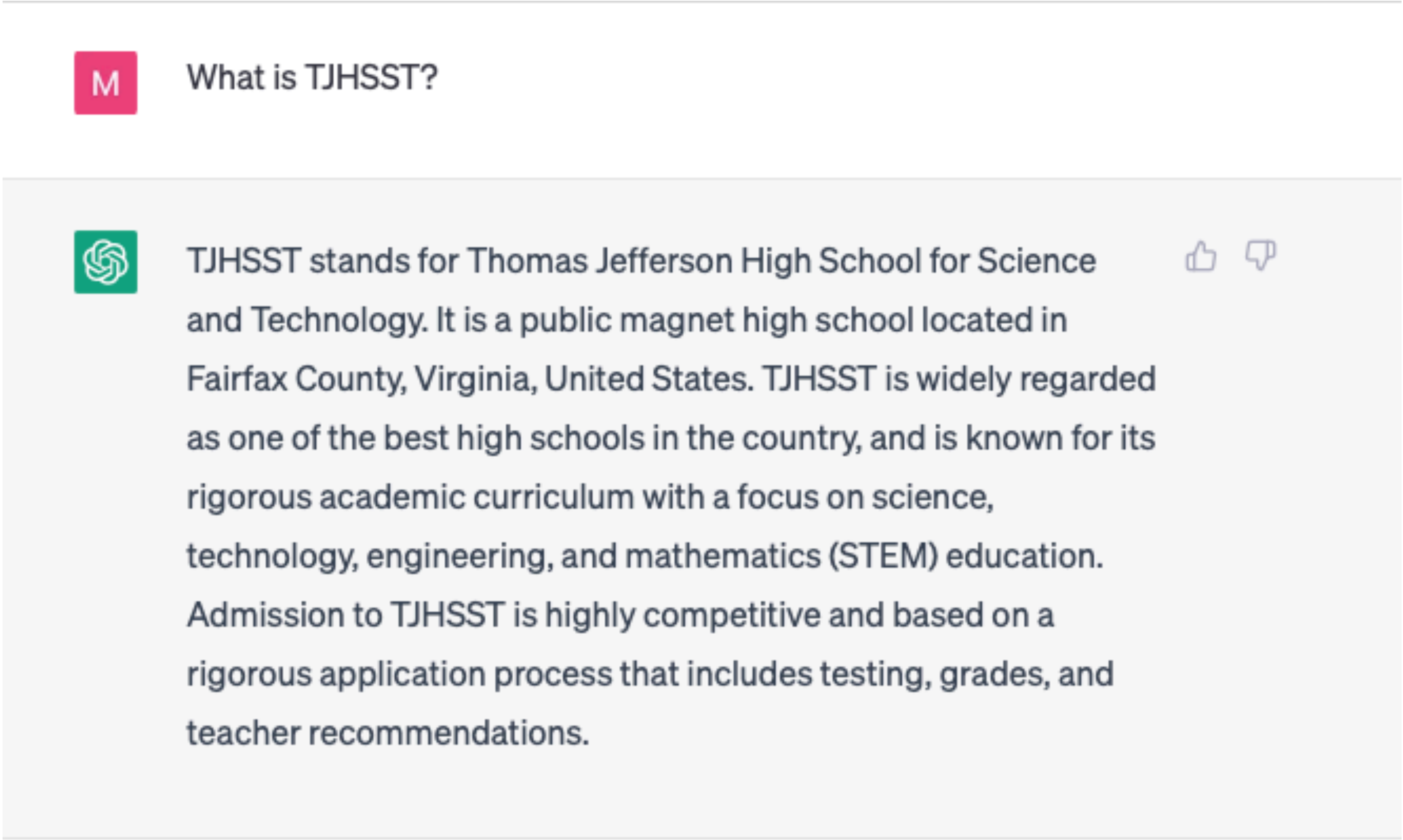
What is Federated Learning?

Machine Learning on Edge Devices

- Billions of IoT devices generate data
- Data enables better Machine Learning on edge devices

GPT Models

- GPT (Generative Pre-trained Transformer)

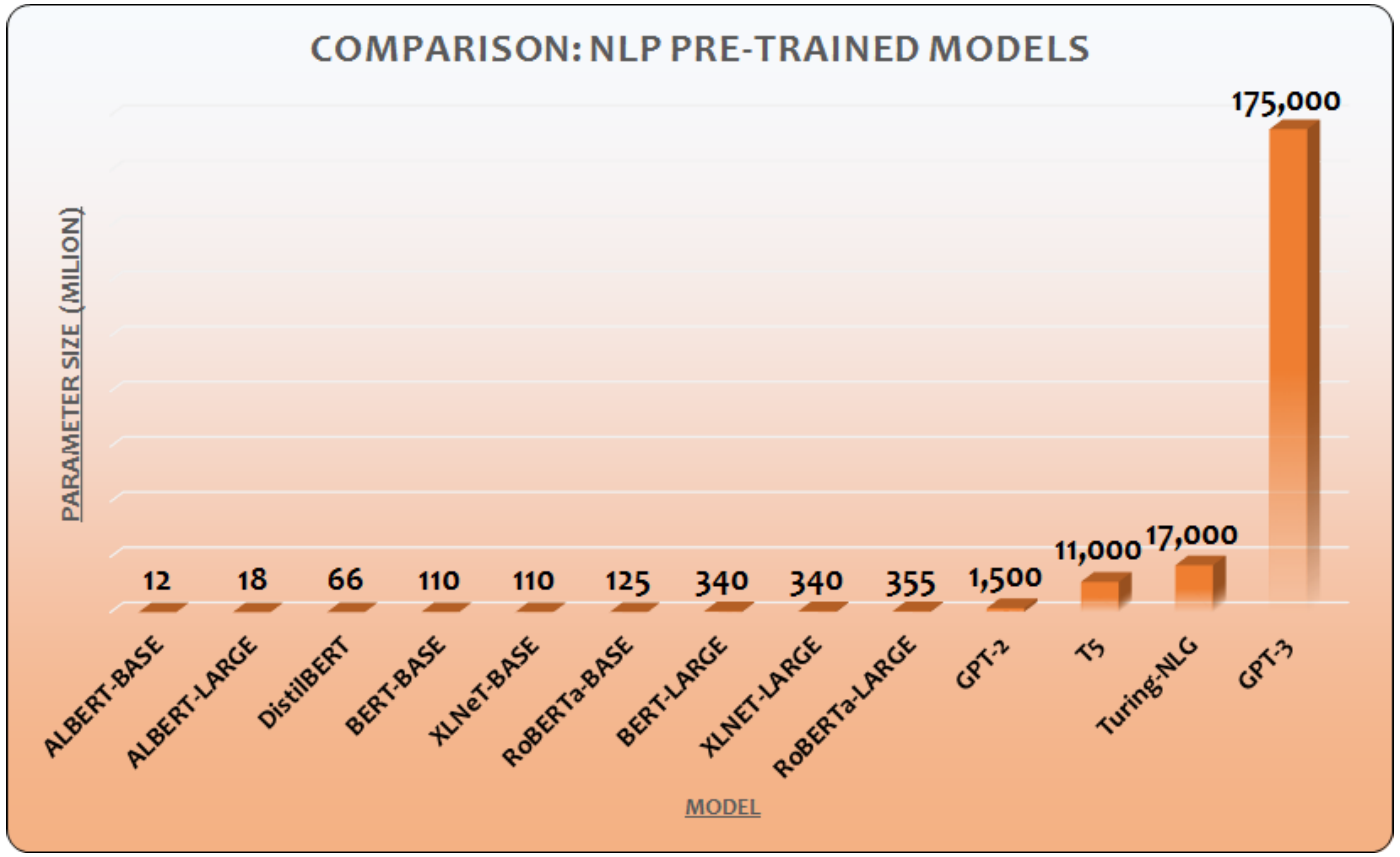

M What is TJHSST?

TJHSST stands for Thomas Jefferson High School for Science and Technology. It is a public magnet high school located in Fairfax County, Virginia, United States. TJHSST is widely regarded as one of the best high schools in the country, and is known for its rigorous academic curriculum with a focus on science, technology, engineering, and mathematics (STEM) education. Admission to TJHSST is highly competitive and based on a rigorous application process that includes testing, grades, and teacher recommendations.

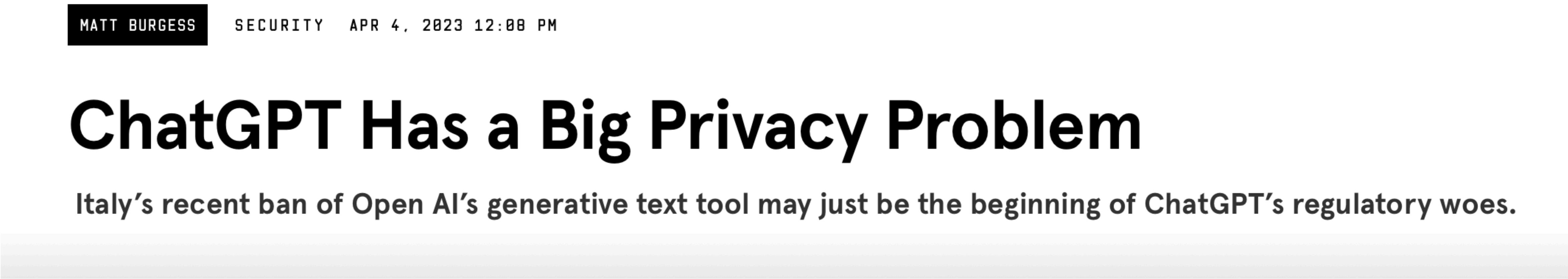
My conversation with ChatGPT

An image generated with DALL-E 2 based on the text prompt "Teddy bears working on new AI research underwater with 1990s technology"

Can we deploy GPT models on Edge Device?



GPT models are huge



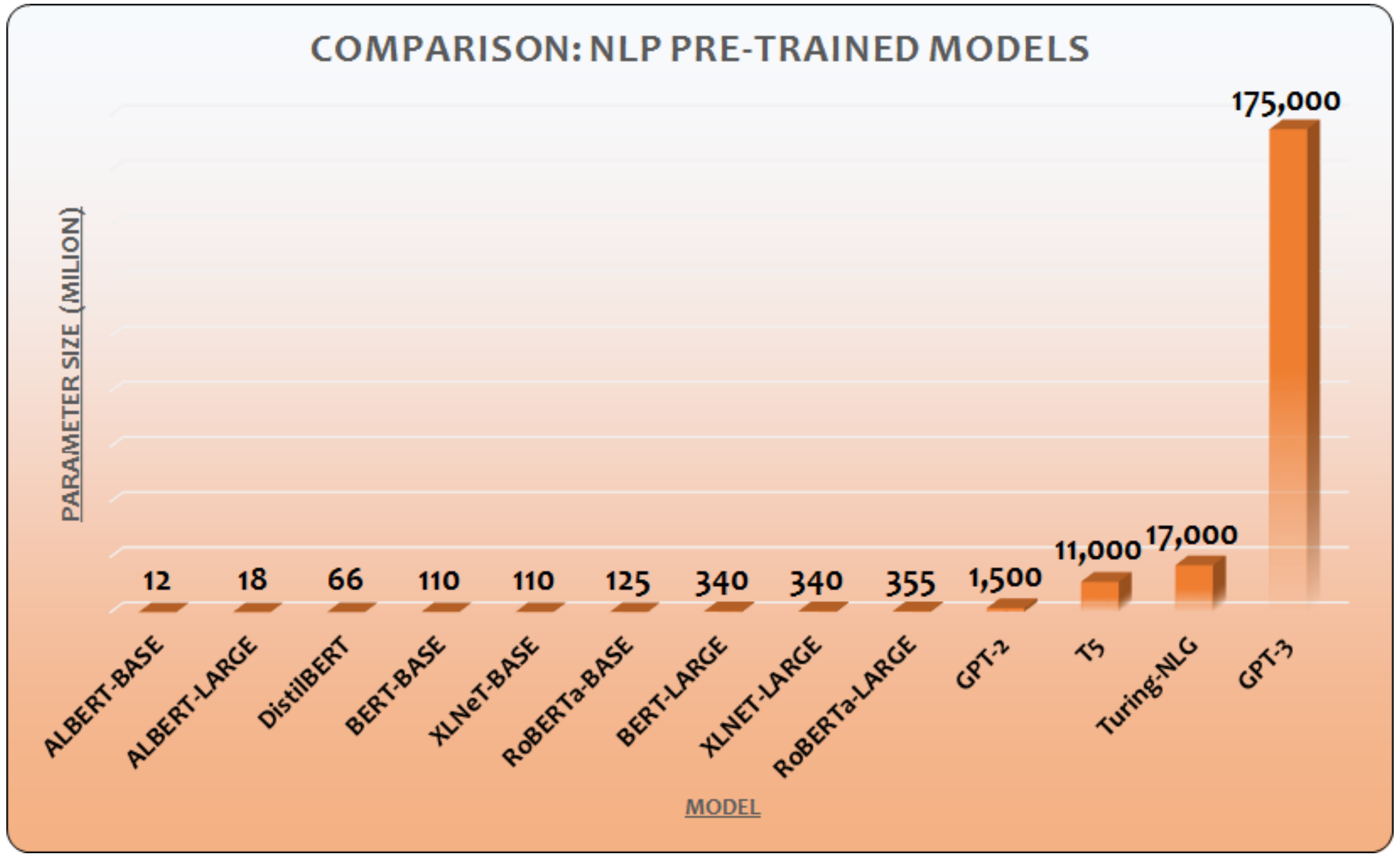
MATT BURGESS SECURITY APR 4, 2023 12:00 PM

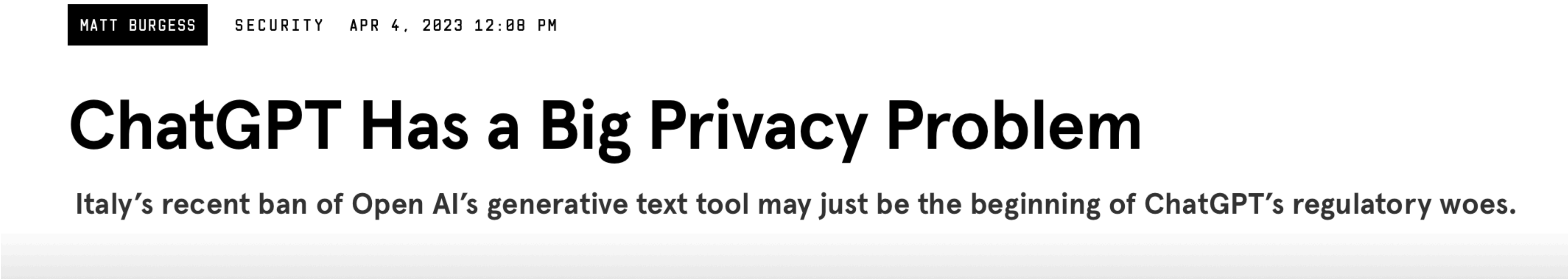
ChatGPT Has a Big Privacy Problem

Italy's recent ban of Open AI's generative text tool may just be the beginning of ChatGPT's regulatory woes.

GPT models might not be private

Can we deploy GPT models on Edge Device?





MATT BURGESS SECURITY APR 4, 2023 12:00 PM

ChatGPT Has a Big Privacy Problem

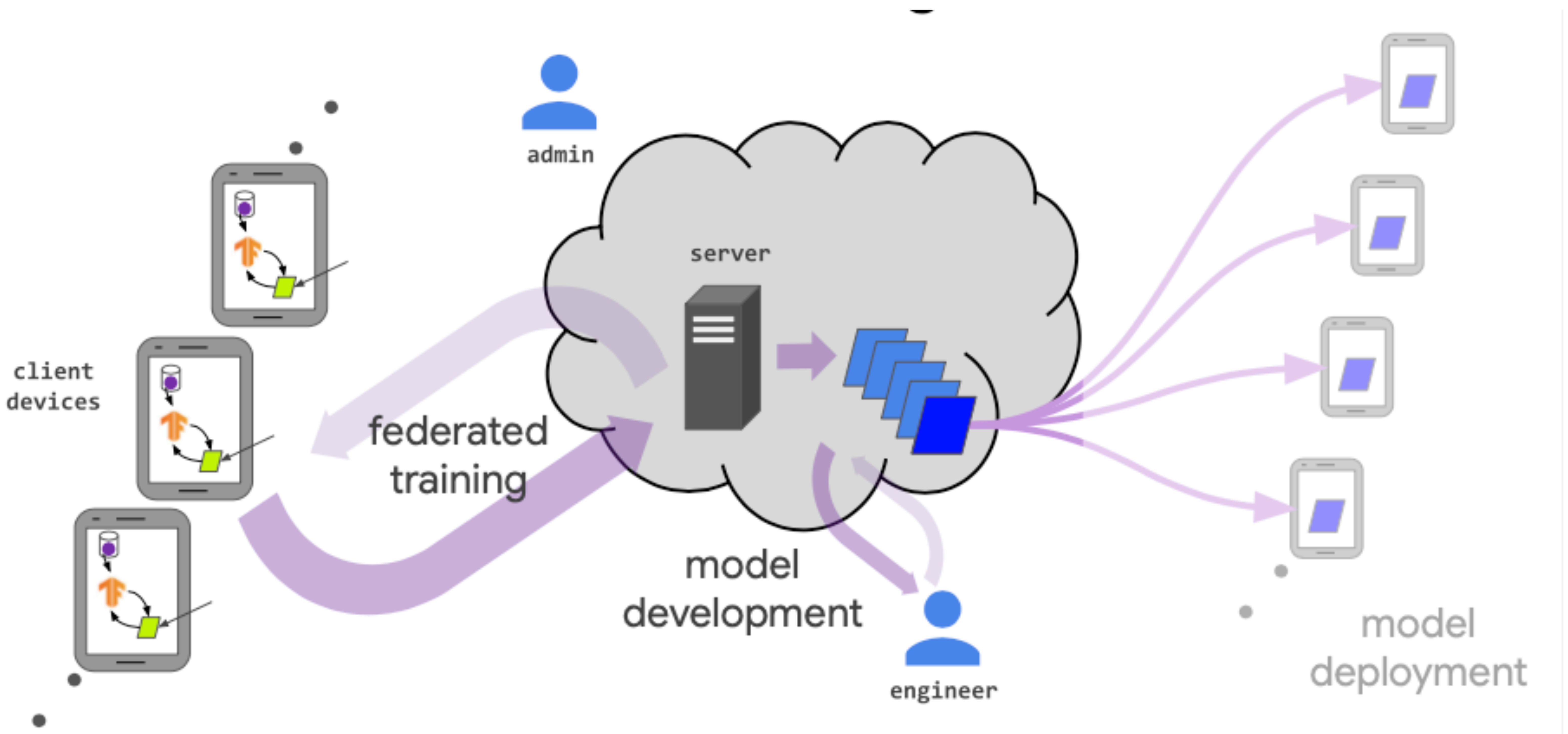
Italy's recent ban of Open AI's generative text tool may just be the beginning of ChatGPT's regulatory woes.

GPT models are huge

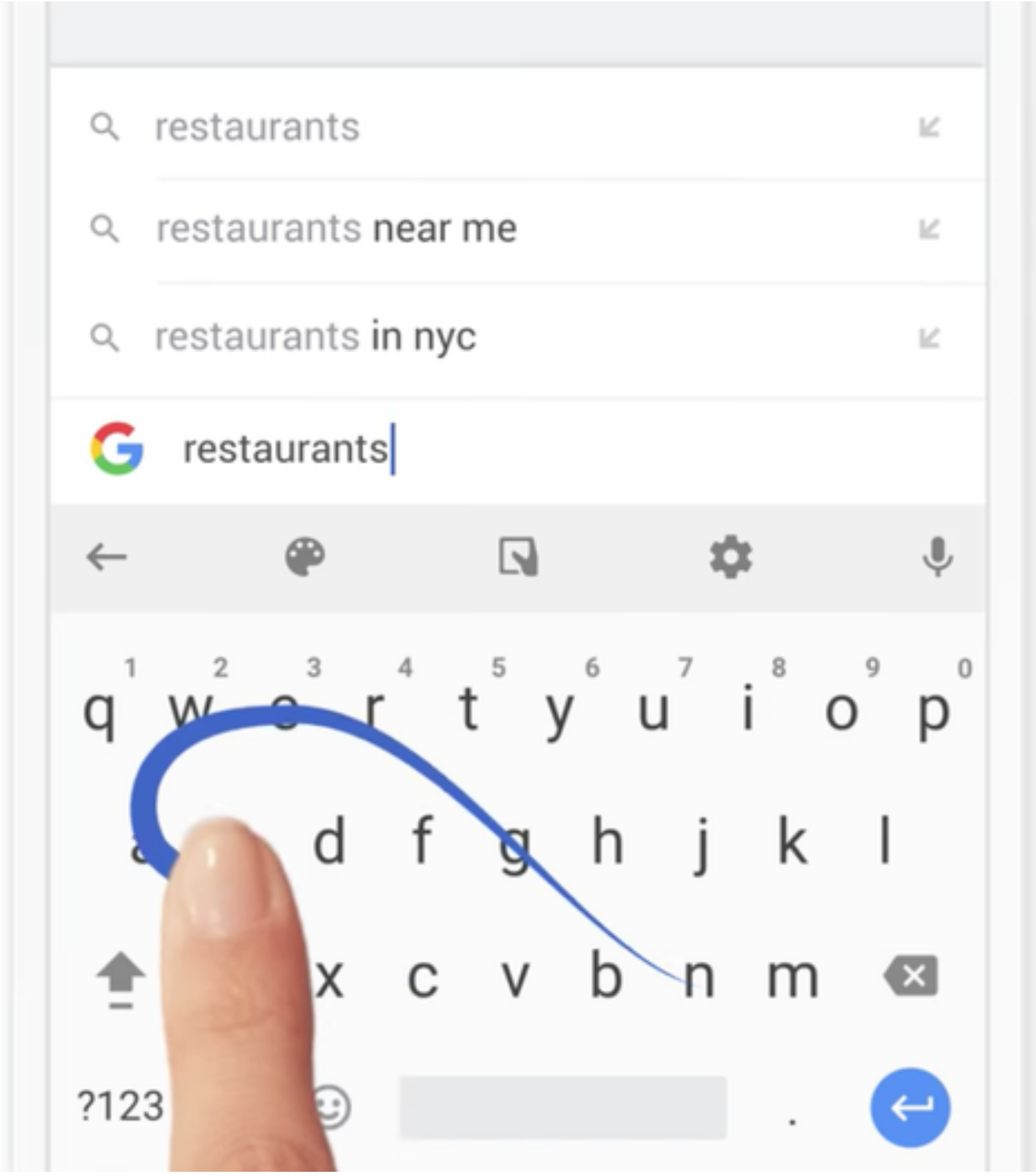
GPT models might not be private

Q: How to enable edge-device intelligence with privacy guarantees?

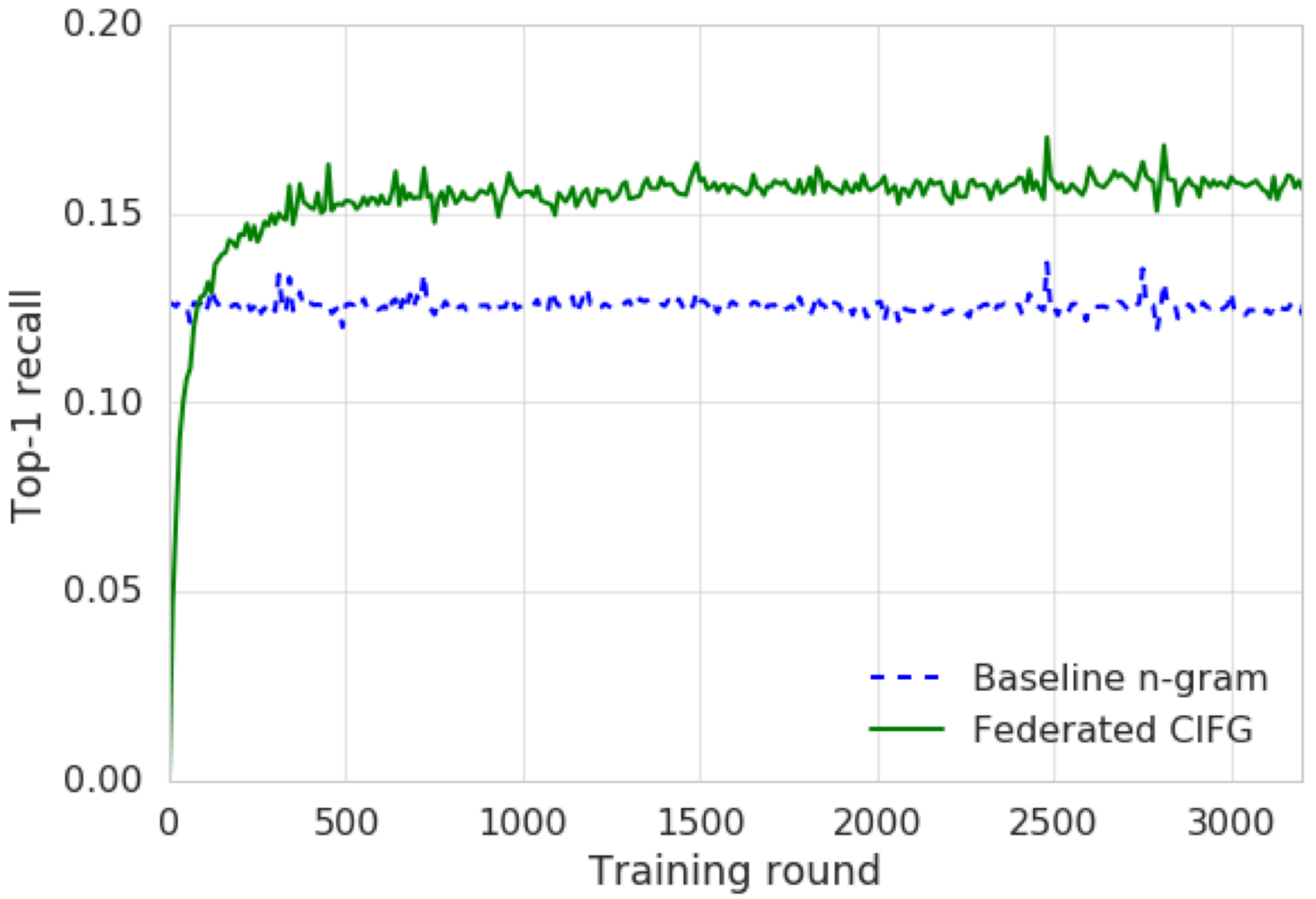
Cross-device Federated Learning



Application: Gboard next-word prediction

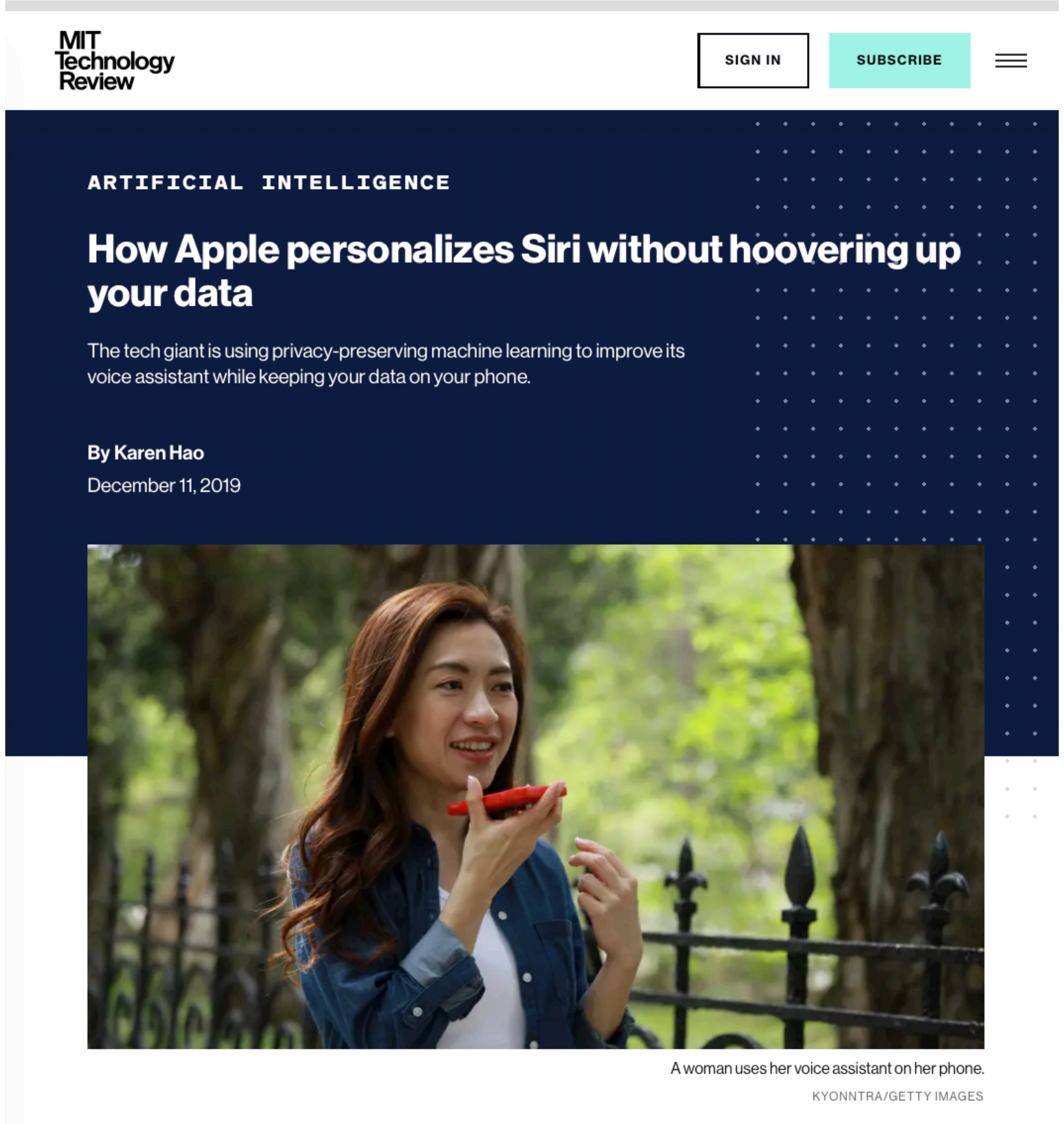


- Federated Recurrent Neural Network
  - Better next-word prediction accuracy: +24%
  - More useful prediction strip: +10% more clicks



A. Hard et al. "Federated learning for mobile keyboard prediction." *arXiv preprint arXiv:1811.03604* (2018).

Application: Apple Siri



MIT Technology Review

SIGN IN

SUBSCRIBE

ARTIFICIAL INTELLIGENCE

How Apple personalizes Siri without hoovering up your data

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

By Karen Hao

December 11, 2019

A woman uses her voice assistant on her phone.

KYONNTRA/GETTY IMAGES

- “Instead, it relies primarily on a technique called federated learning, Apple’s head of privacy, Julien Freudiger, told an audience at the Neural Processing Information Systems conference on December 8. Federated learning is a privacy-preserving machine-learning method that was first introduced by Google in 2017. It allows Apple to train different copies of a speaker recognition model across all its users’ devices, using only the audio data available locally. It then sends just the updated models back to a central server to be combined into a master model. In this way, raw audio of users’ Siri requests never leaves their iPhones and iPads, but the assistant continuously gets better at identifying the right speaker.”

https://www.technologyreview.com/2019/12/11/131629/apple-ai-personalizes-siri-federated-learning/

Formal Definition of Federated Learning

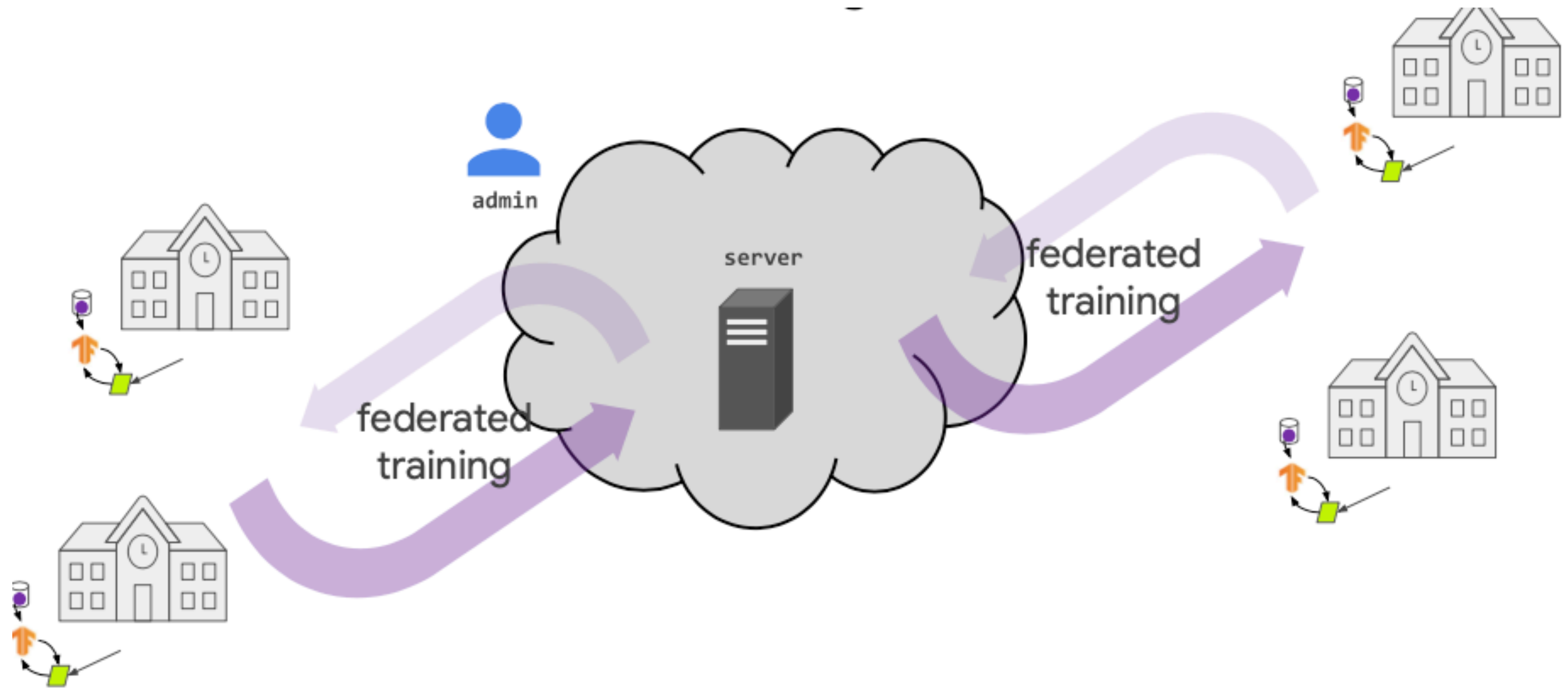
- Federated learning (FL) is a machine learning setting where many clients (e.g. mobile devices or whole organizations) collaboratively train a model under the orchestration of a central server (e.g. service provider), while keeping the training data decentralized.

Efficiency

Privacy

P. Kairouz et al, 2021. Advances and open problems in federated learning. Foundations and Trends® in Machine Learning, 14(1–2), pp.1-210.

Cross-Silo Federated Learning



Small number of clients (e.g., organizations, data silos) participate the federated learning

Federated Learning (FL) Terminology

- Clients: compute nodes holding local data
  - IoT devices, mobile devices, data silos, data centers in different geographic regions
- Server: Additional compute nodes that coordinate the FL process without accessing the raw data

How to design FL algorithms?

Illustration of Federated Learning Algorithms

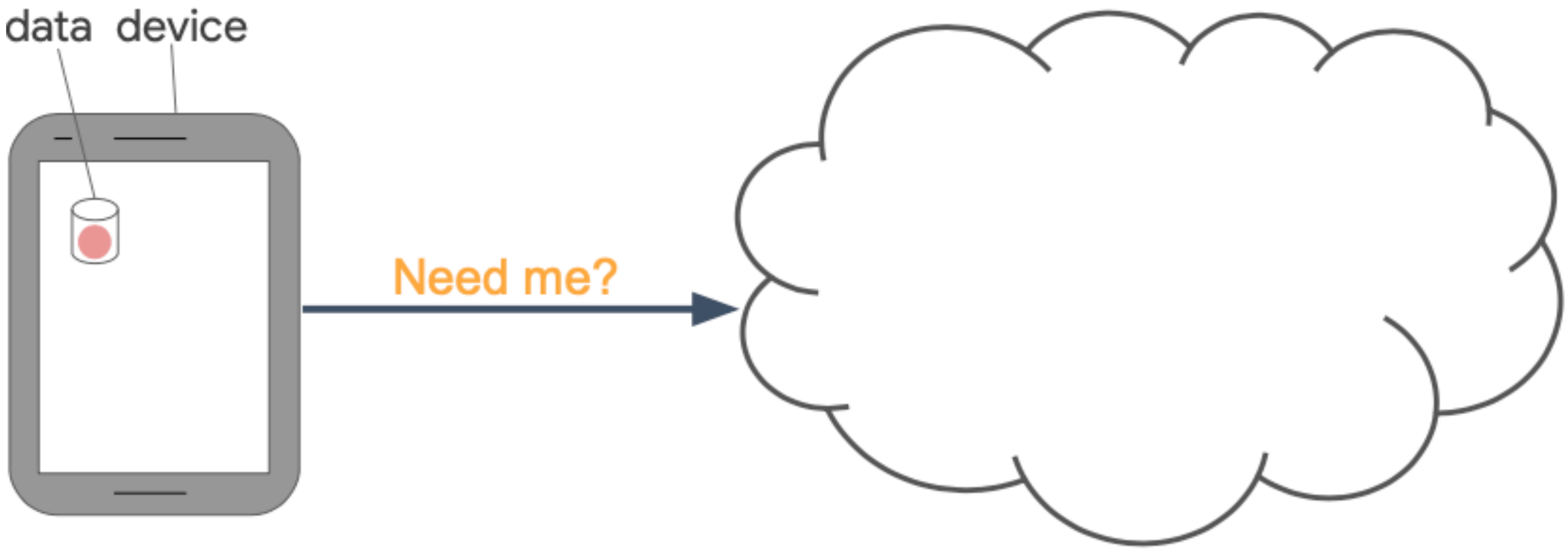


Illustration of Federated Learning Algorithms

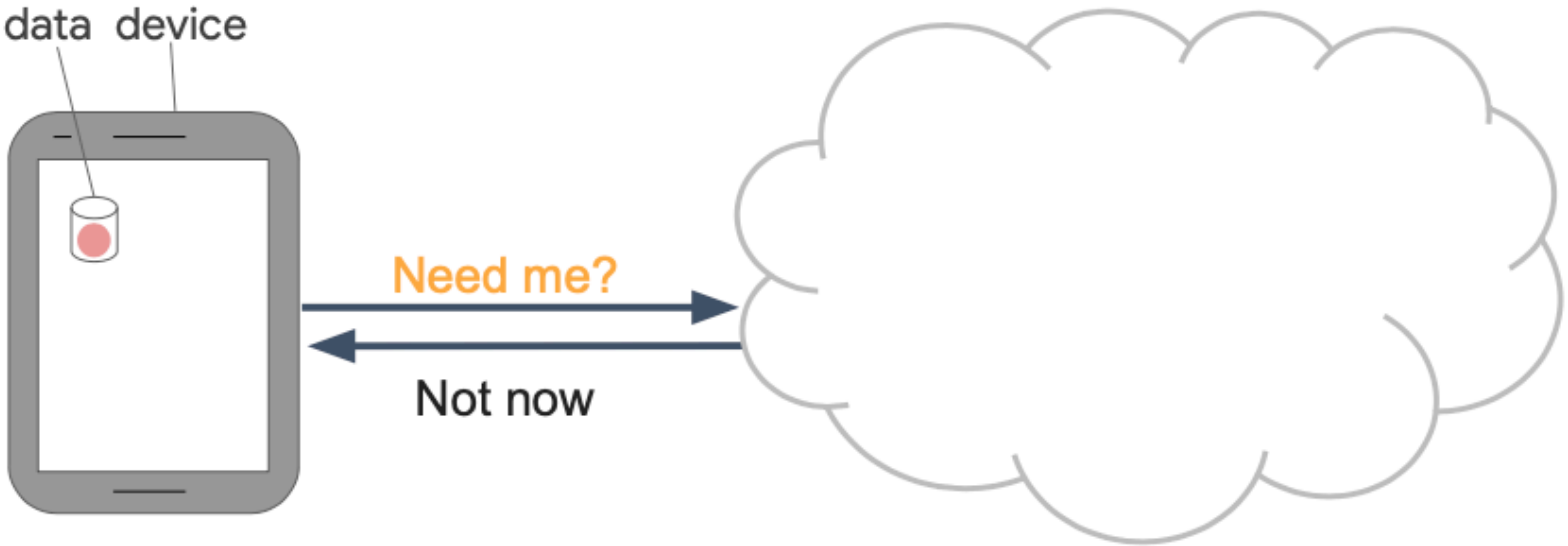


Illustration of Federated Learning Algorithms

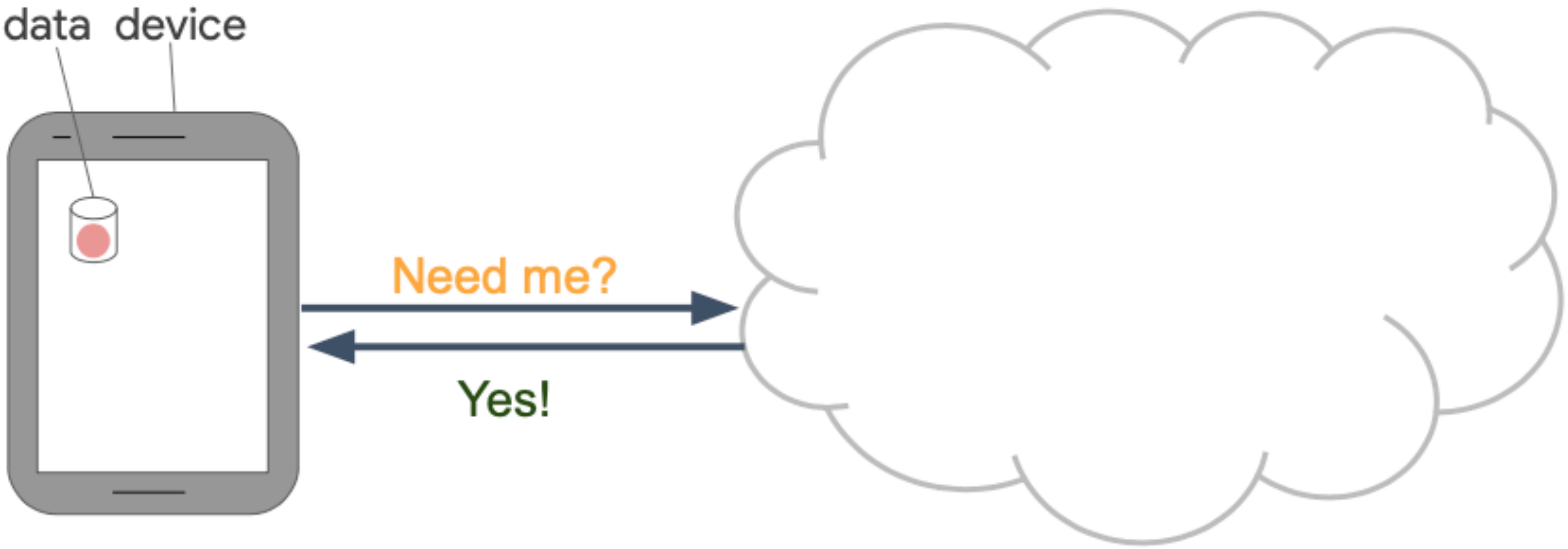


Illustration of Federated Learning Algorithms

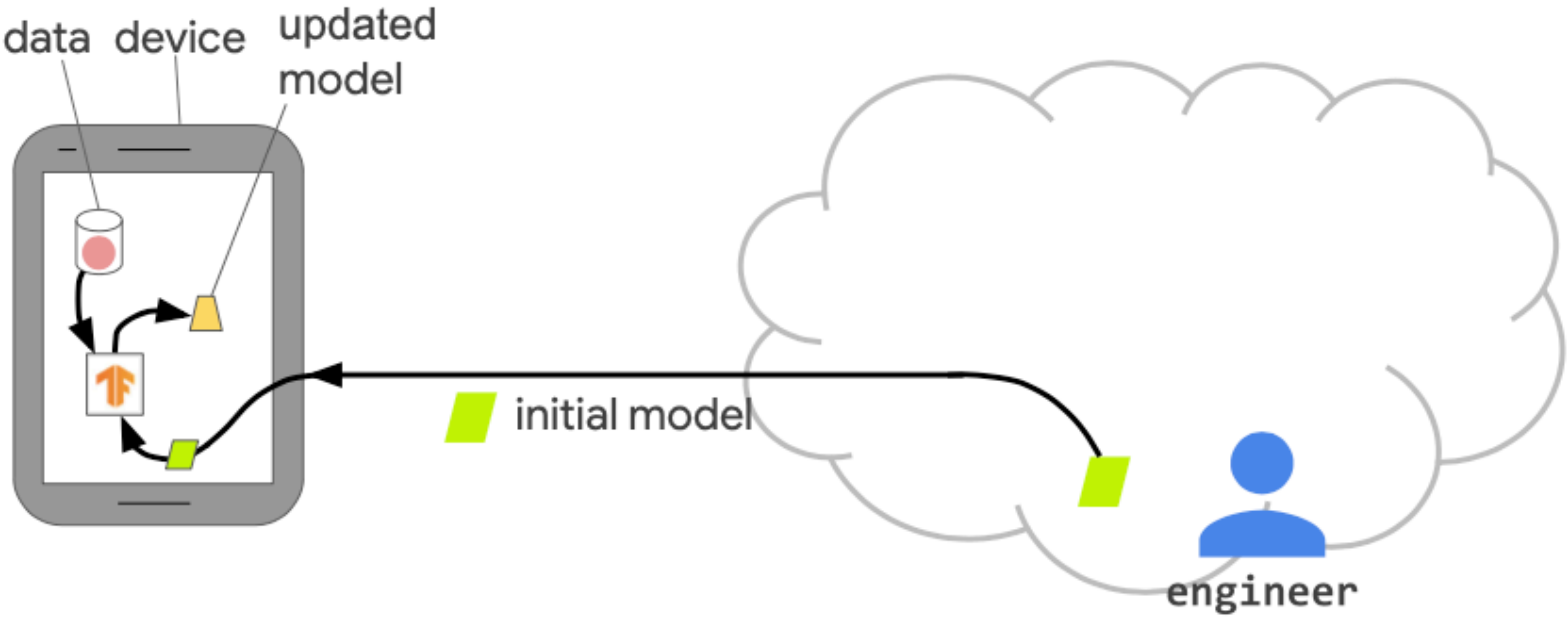


Illustration of Federated Learning Algorithms

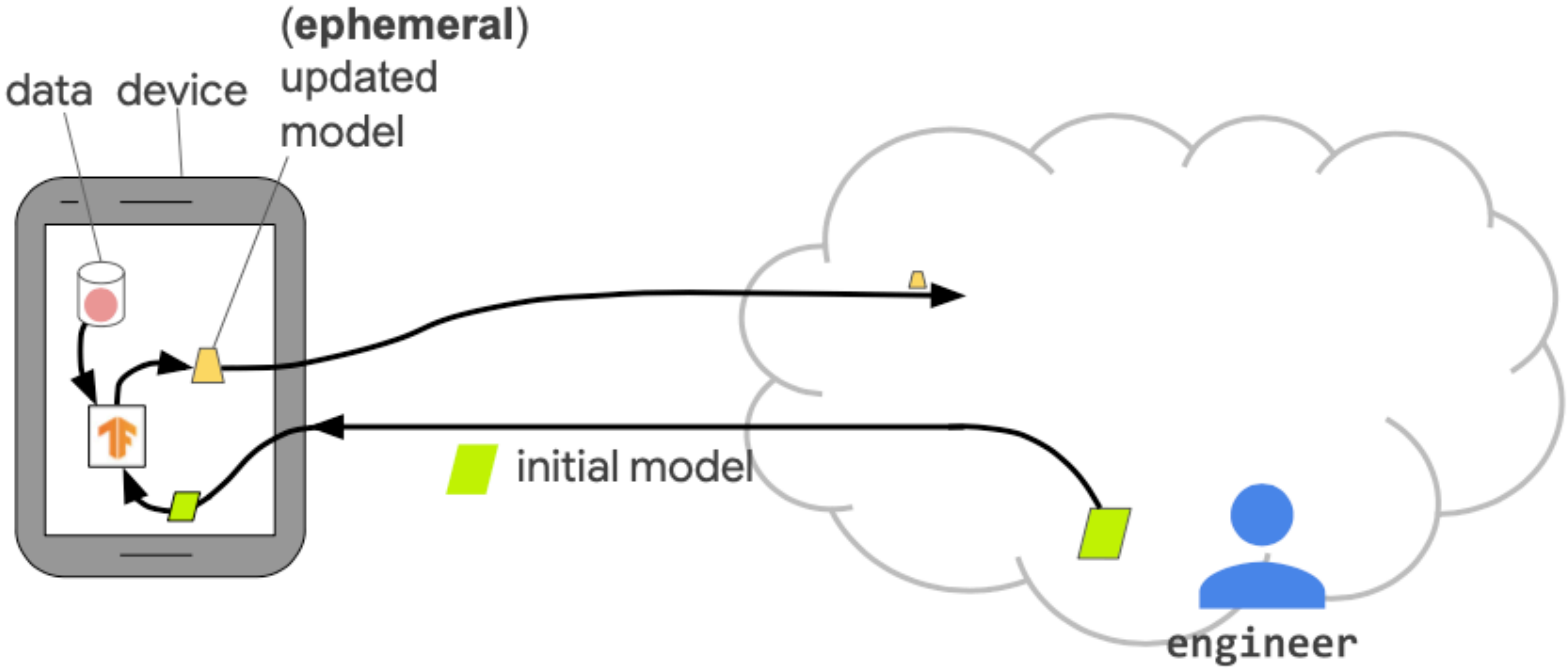


Illustration of Federated Learning Algorithms

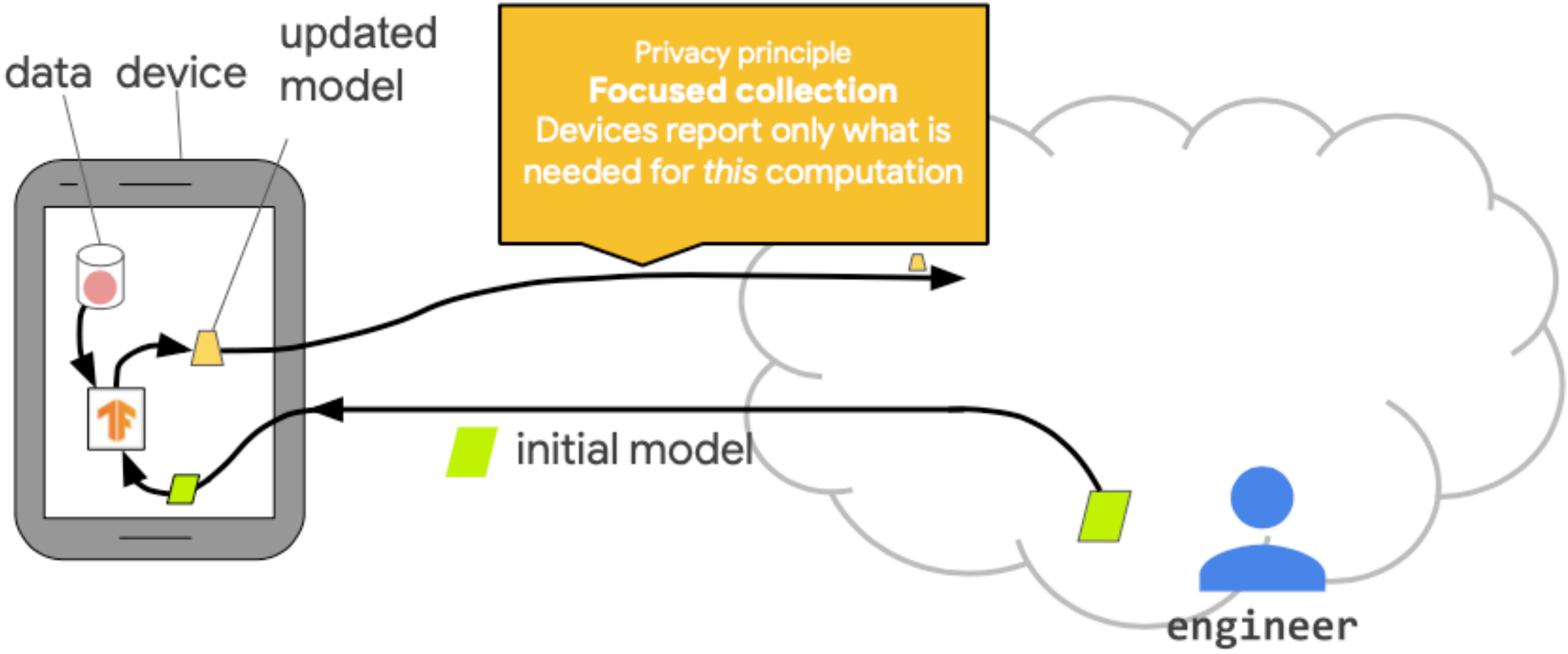


Illustration of Federated Learning Algorithms

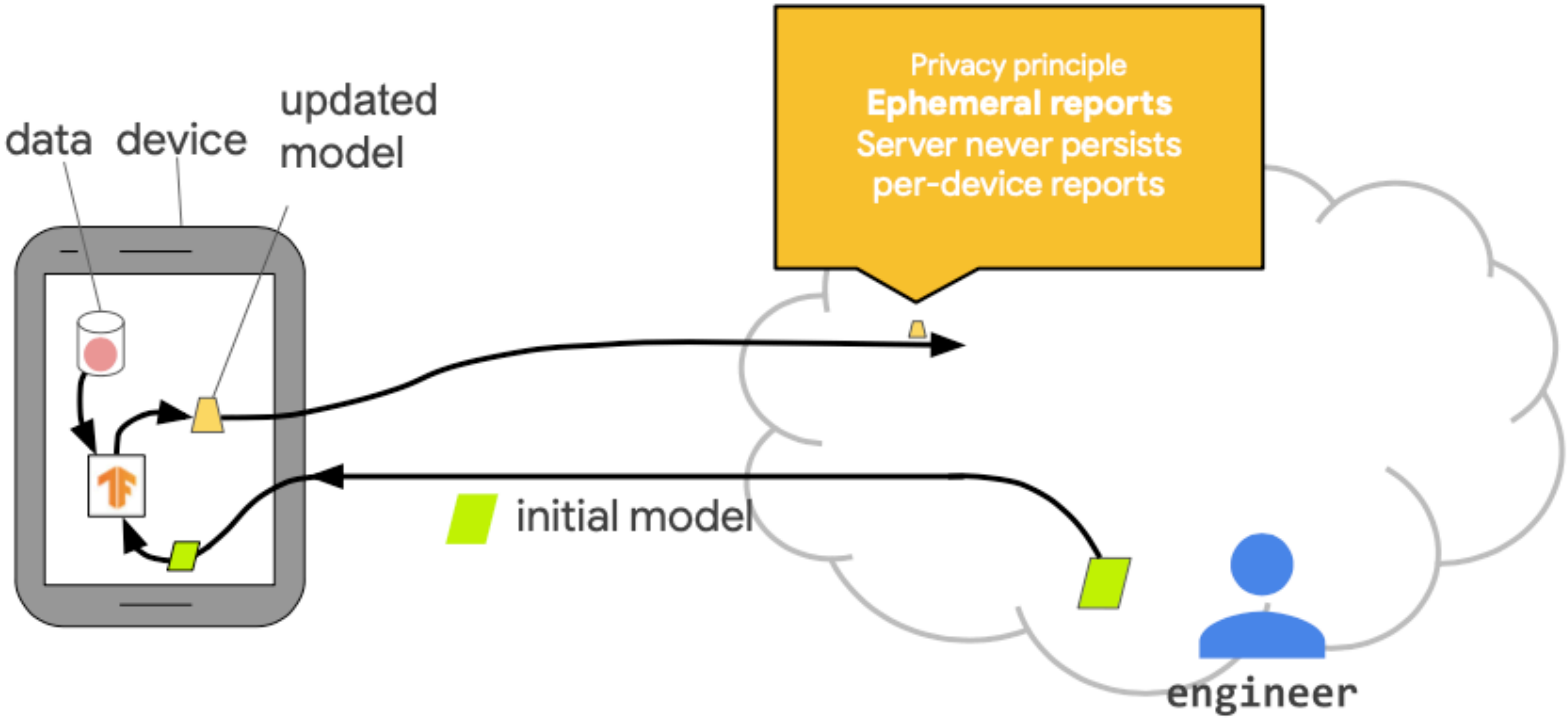


Illustration of Federated Learning Algorithms

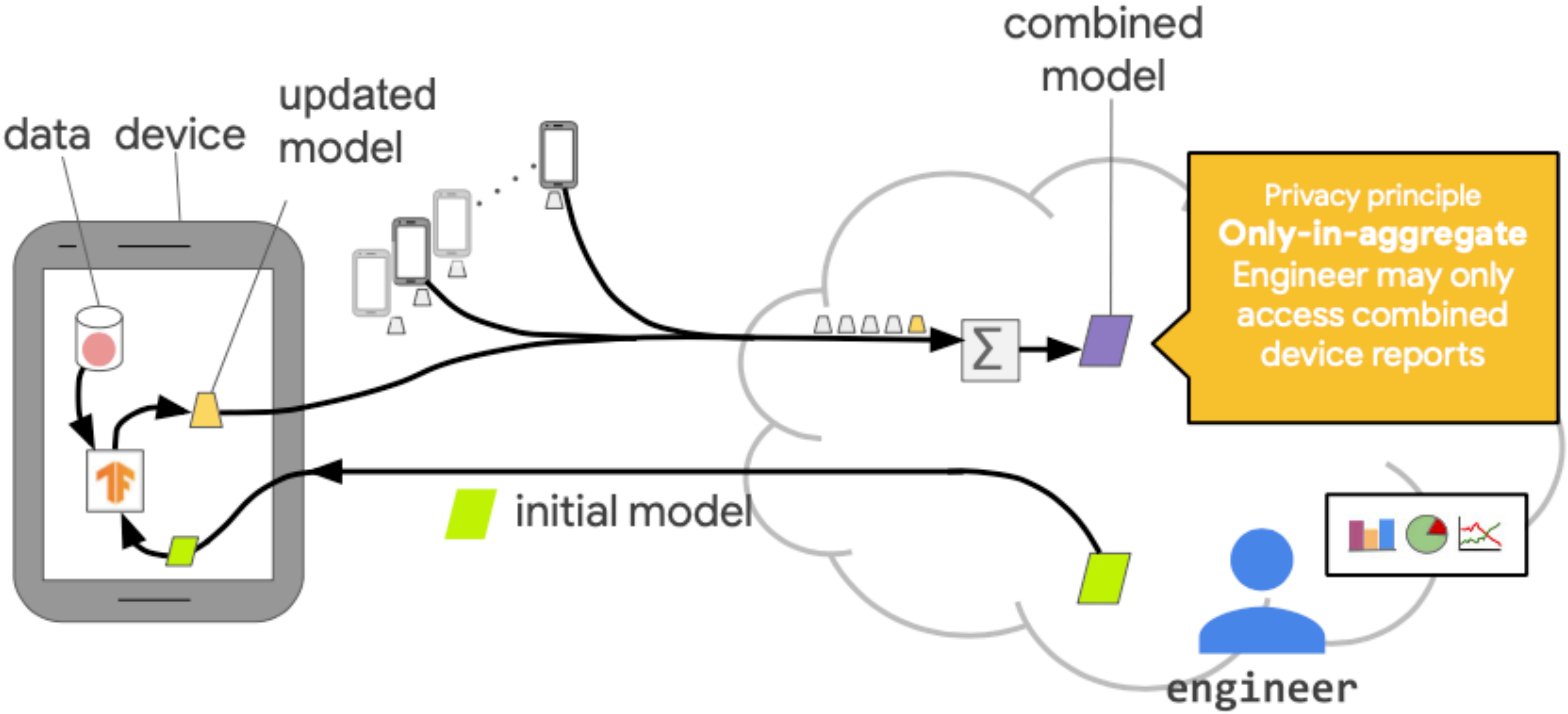


Illustration of Federated Learning Algorithms

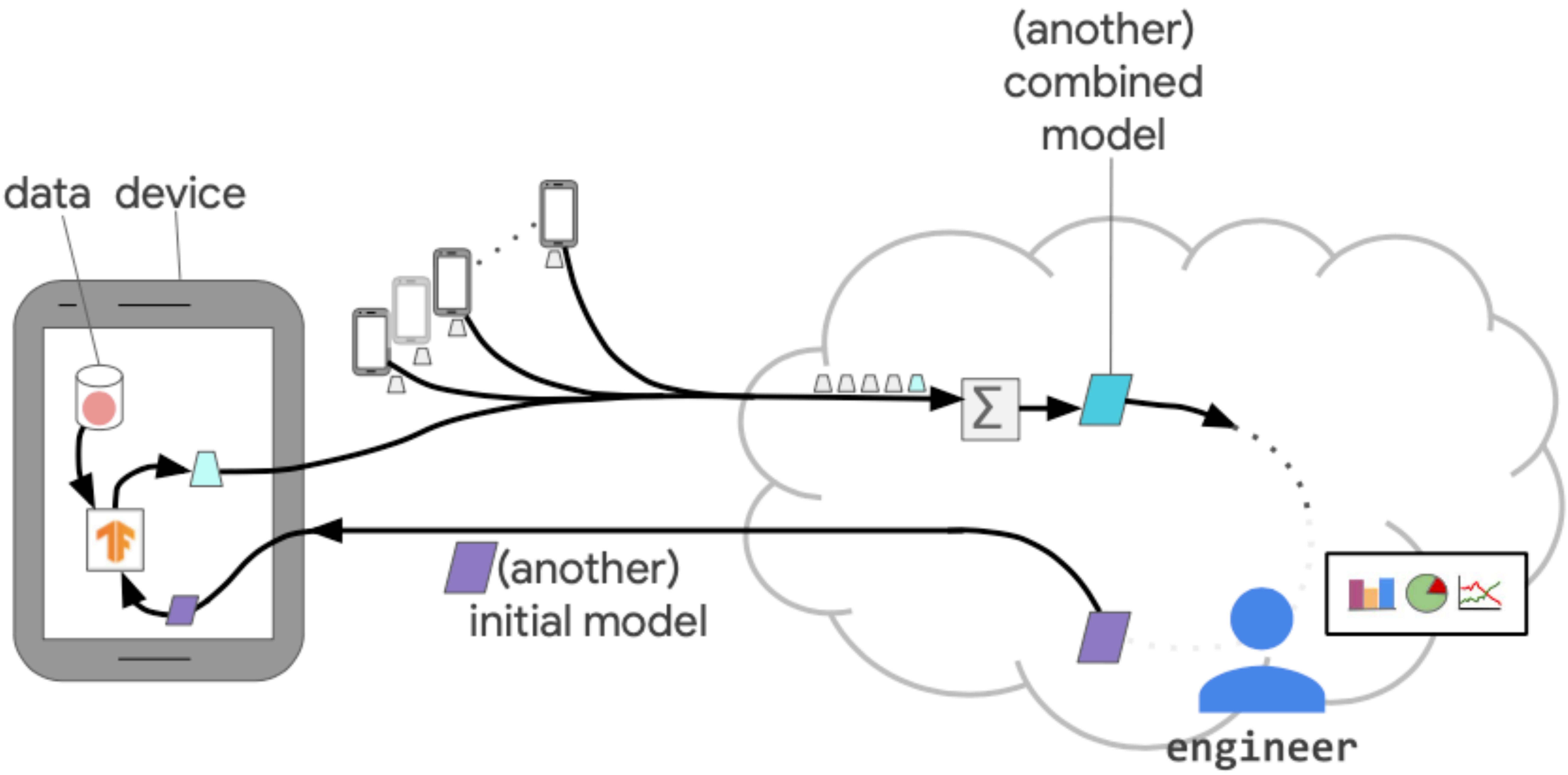
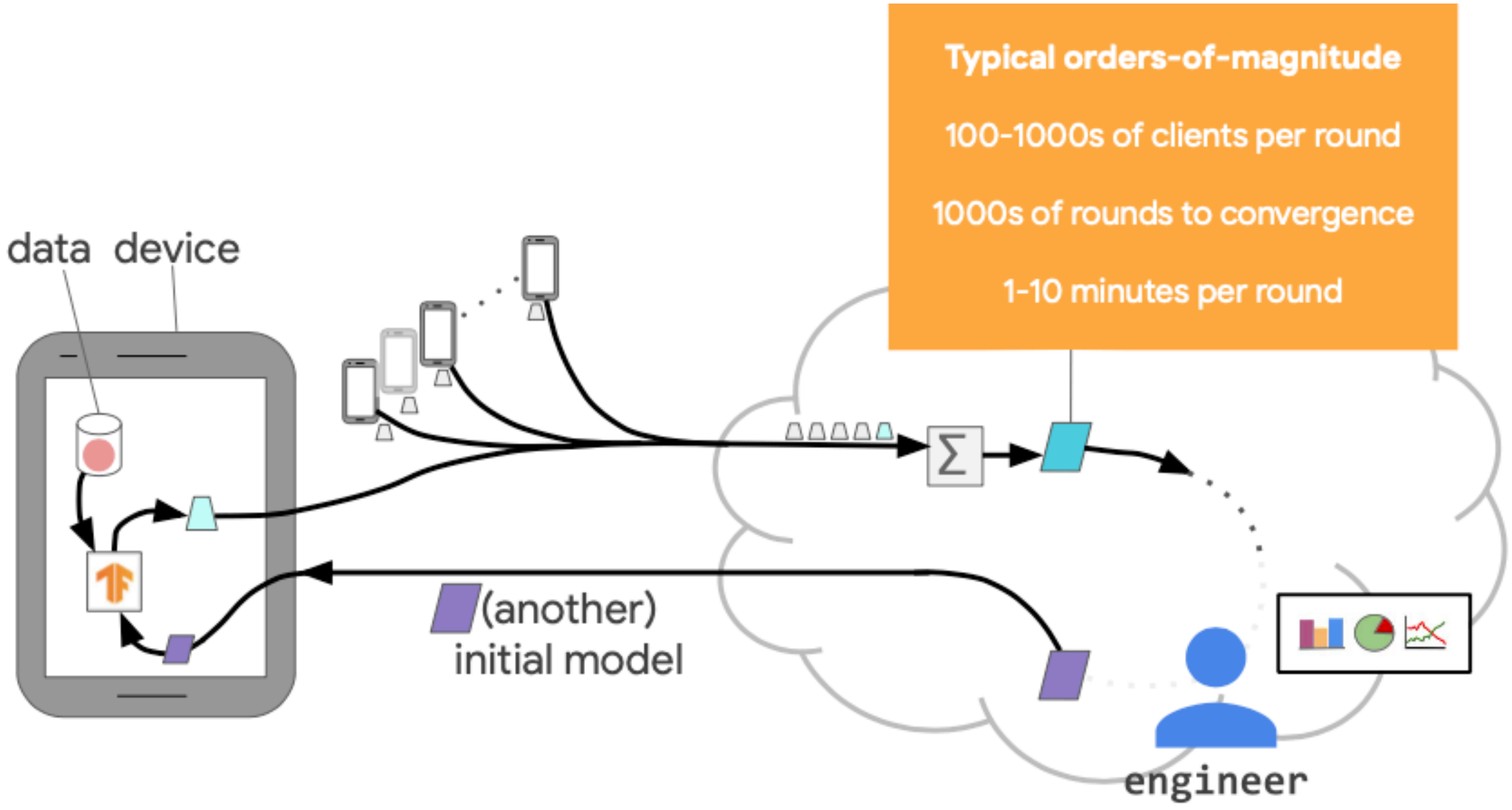


Illustration of Federated Learning Algorithms



Machine Learning as Risk Minimization

Risk of model f

$$f^* = \arg\min_{f \in \mathcal{F}} R(f) := \mathbb{E}_{\mathbf{x},y}\left[\ell(f(\mathbf{x}), y)\right]$$

- $\mathcal{F}$: hypothesis class
- Loss function $\ell(\hat{y}, y)$ measures the prediction error

Machine Learning as Risk Minimization

Risk of model f

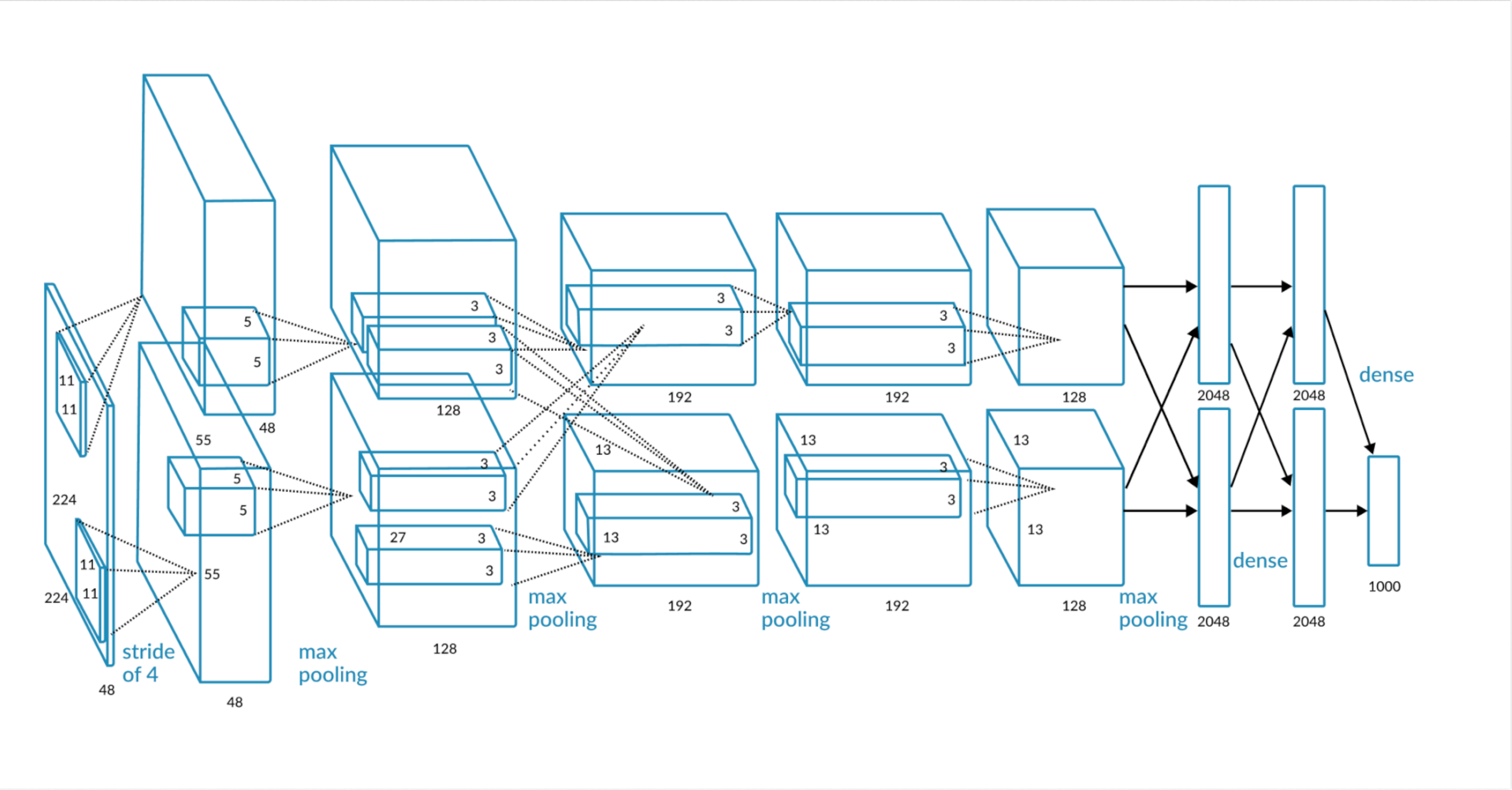
$$f^* = \arg\min_{f \in \mathscr{F}} R(f) := \mathbb{E}_{\mathbf{x},y}\left[\ell(f(\mathbf{x}), y)\right]$$

- $\mathscr{F}$: hypothesis class
- Loss function $\ell(\hat{y}, y)$ measures the prediction error

Prediction $\hat{y} = f(\mathbf{w}; \mathbf{x})$

$$\mathbf{w}_* = \arg\min_{\mathbf{w}} \mathbb{E}_{\mathbf{x},y}\left[\ell(f(\mathbf{w}; \mathbf{x}), y)\right]$$

Deep Neural Networks



Alexnet: $\mathbf{x} \rightarrow f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}_L \circ \sigma \left(\ldots \sigma \left(\mathbf{w}_2 \circ \sigma(\mathbf{w}_1 \circ \mathbf{x}) \right) \right)$

$$\min_{\mathbf{w}} \mathbb{E}_{\mathbf{x},y} \left[\ell(f_{\mathbf{w}}(\mathbf{x}), y) \right]$$

The workhorse in Machine Learning

Stochastic Gradient Descent

The workhorse in Machine Learning

Stochastic Gradient Descent

$$\min_{\mathbf{w}} \mathbb{E}_{\mathbf{x},y}\left[\ell(f_{\mathbf{w}}(\mathbf{x}), y)\right]$$

- Stochastic Gradient Descent (SGD) [Robbins-Monro'51]
  - Sample $(\mathbf{x}_t, y_t)$ uniformly
  - $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla \ell(\mathbf{w}_t, \mathbf{x}_t, y_t)$

The workhorse in Machine Learning

Stochastic Gradient Descent

$$\min_{\mathbf{w}} \mathbb{E}_{\mathbf{x},y} \left[\ell(f_{\mathbf{w}}(\mathbf{x}), y)\right]$$

- Stochastic Gradient Descent (SGD) [Robbins-Monro'51]
  - Sample $(\mathbf{x}_t, y_t)$ uniformly
  - $\mathbf{w}_{t+1} = \mathbf{w}_t - \underbrace{\eta_t}_{\text{Learning rate}} \overbrace{\nabla \ell(\mathbf{w}_t, \mathbf{x}_t, y_t)}^{\text{Stochastic gradient}}$

The workhorse in Machine Learning

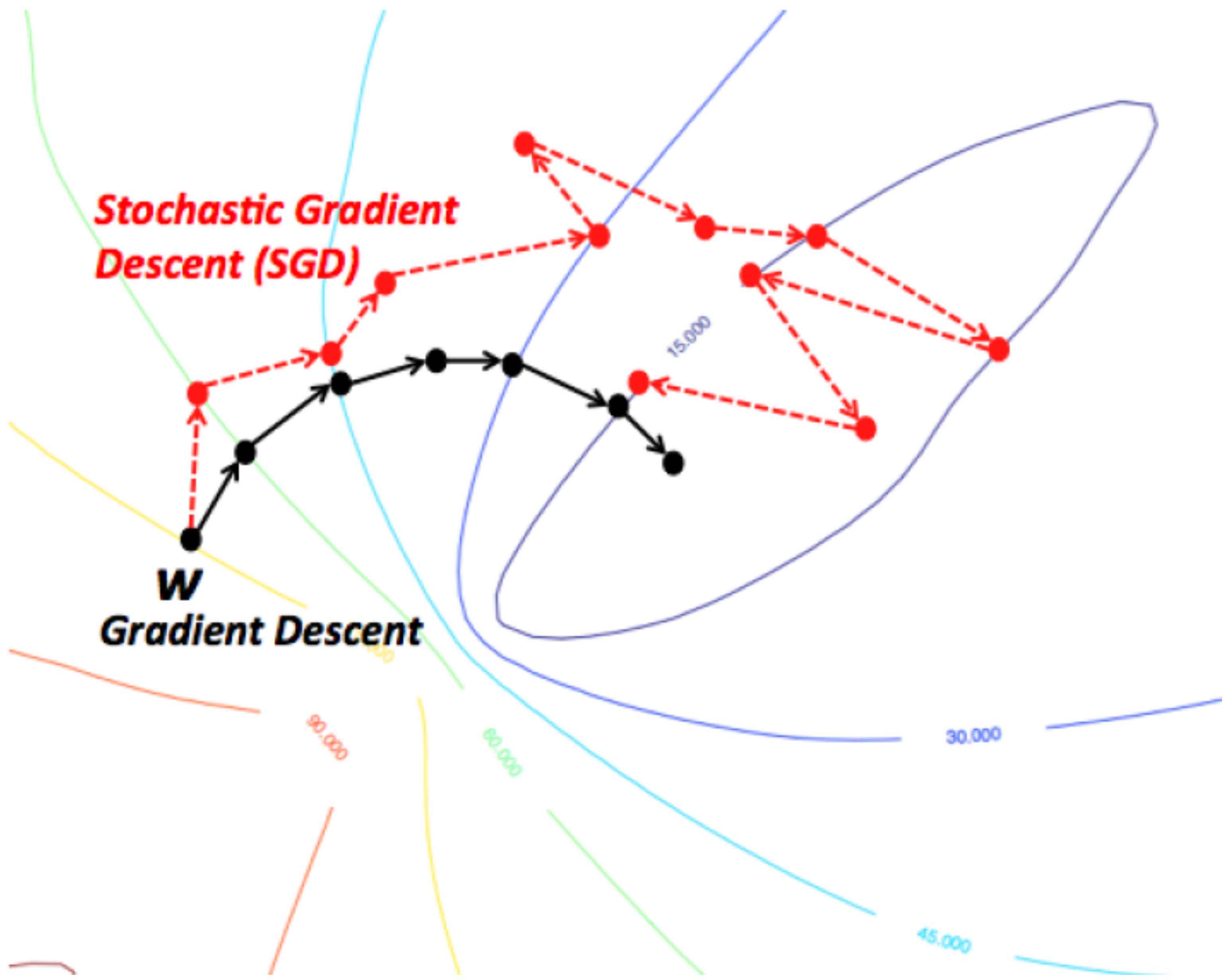
Stochastic Gradient Descent

$$\min_{\mathbf{w}} \mathbb{E}_{\mathbf{x},y}\left[\ell(f_{\mathbf{w}}(\mathbf{x}), y)\right]$$

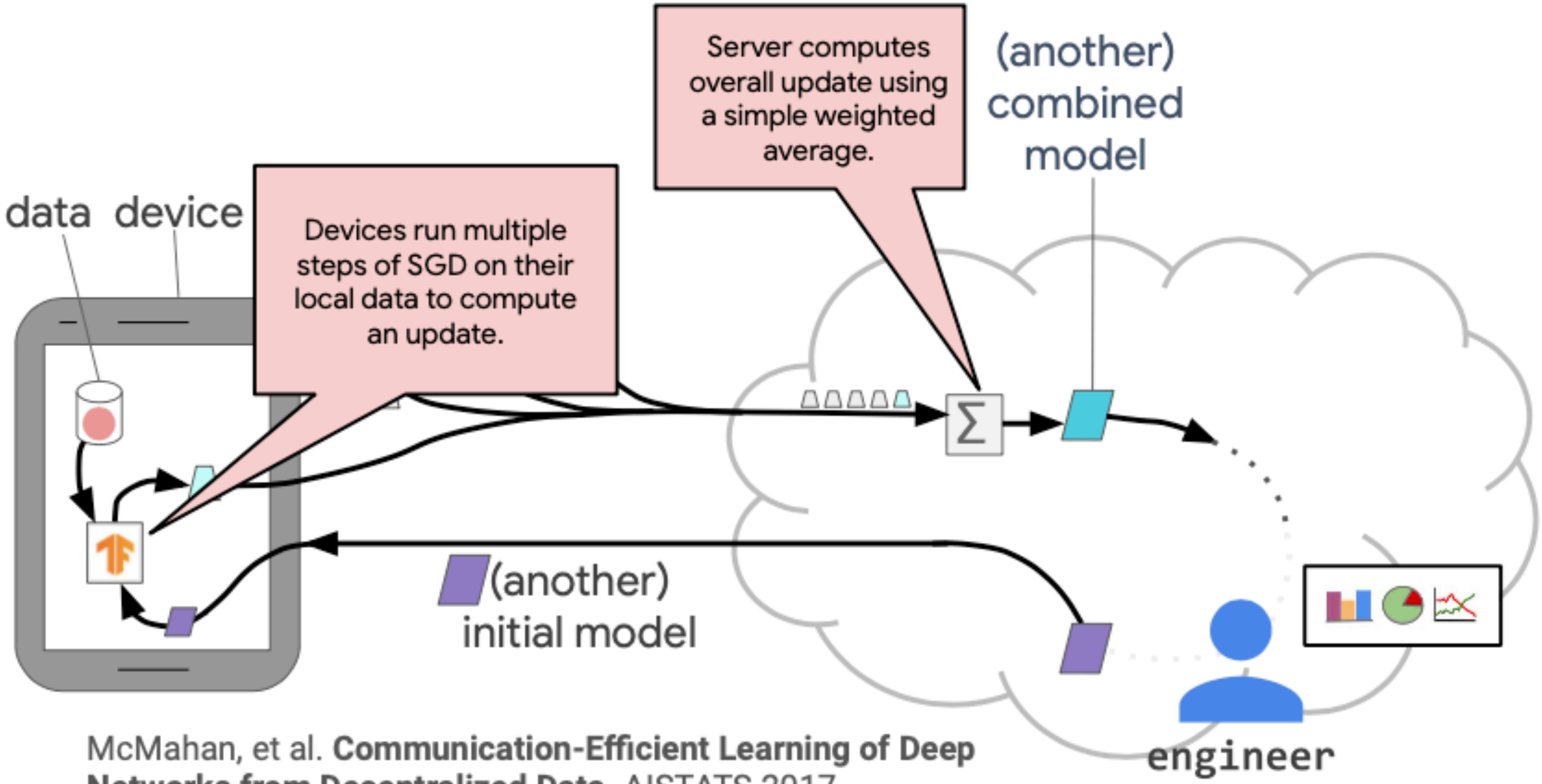
- Stochastic Gradient Descent (SGD) [Robbins-Monro'51]
  - Sample $(\mathbf{x}_t, y_t)$ uniformly
  - $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla \ell(\mathbf{w}_t, \mathbf{x}_t, y_t)$

Stochastic gradient

Learning rate



Federated Averaging (FedAvg) algorithm



McMahan, et al. **Communication-Efficient Learning of Deep Networks from Decentralized Data.** AISTATS 2017.

Dive Deep into FedAvg Algorithm

Algorithm 1 `FederatedAveraging`. The K clients are indexed by k; B is the local minibatch size, E is the number of local epochs, and η is the learning rate.

Server executes:
  initialize w_0
  for each round $t = 1, 2, \ldots$ **do**
    $m \leftarrow \max(C \cdot K, 1)$
    $S_t \leftarrow$ (random set of m clients)
    for each client $k \in S_t$ **in parallel do**
      $w_{t+1}^k \leftarrow \text{ClientUpdate}(k, w_t)$ — Stochastic Gradient Descent
    $m_t \leftarrow \sum_{k \in S_t} n_k$
    $w_{t+1} \leftarrow \sum_{k \in S_t} \frac{n_k}{m_t} w_{t+1}^k$ *// Erratum*[4] — Average the model parameter

ClientUpdate(k, w): *// Run on client k*
  $\mathcal{B} \leftarrow$ (split $\mathcal{P}_k$ into batches of size B)
  for each local epoch i from 1 to E **do**
    for batch $b \in \mathcal{B}$ **do**
      $w \leftarrow w - \eta \nabla \ell(w; b)$
  return w to server

General Framework of FL Algorithm Design

- For t=1,…,T
  - Sample a subset of clients and initialize the model from the server
  - For each client in parallel,
    - Run a client optimization algorithm to update the model
    - Compute the actual update on each client
  - Average client updates on the server
  - Run a server optimization algorithm

ClientOpt (e.g., SGD)

SeverOpt (e.g., SGD)

General Framework of FL Algorithm Design

- For t=1,…,T
  - Sample a subset of clients and initialize the model from the server
  - For each client in parallel,
    - Run a client optimization algorithm to update the model
    - Compute the actual update on each client
  - Average client updates on the server
  - Run a server optimization algorithm

ClientOpt (e.g., SGD)

SeverOpt (e.g., SGD)

Algorithm design in FL boils down to designing ClientOpt and ServerOpt

Client Drift for Heterogeneous Data

Karimireddy et al. "Scaffold: Stochastic controlled averaging for federated learning." In *International Conference on Machine Learning*, pp. 5132-5143. PMLR, 2020.

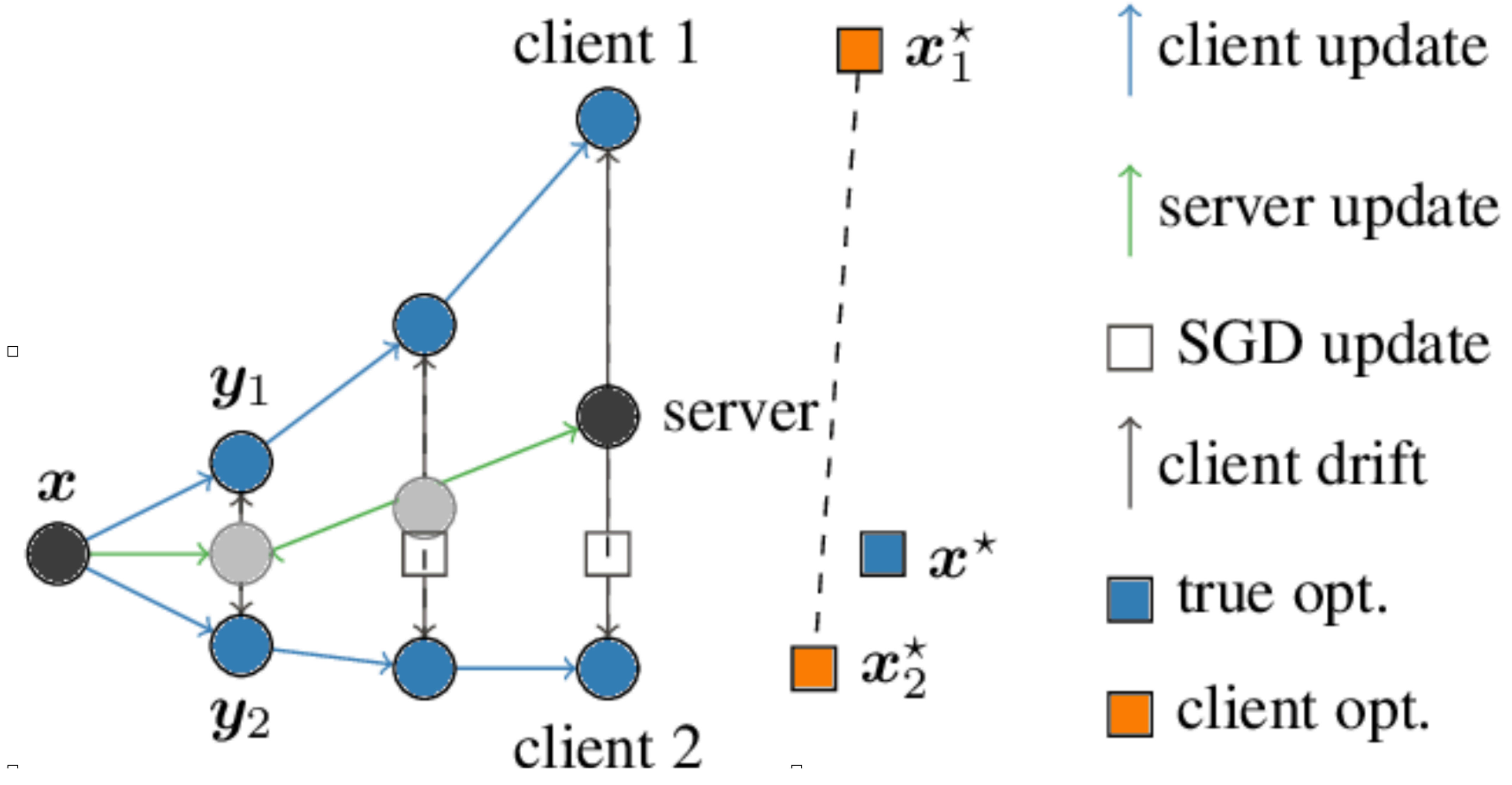
Client Drift for Heterogeneous Data

Heterogeneous Data: different client has different data distribution

Karimireddy et al. "Scaffold: Stochastic controlled averaging for federated learning." In *International Conference on Machine Learning*, pp. 5132-5143. PMLR, 2020.

Client Drift for Heterogeneous Data

Heterogeneous Data: different client has different data distribution

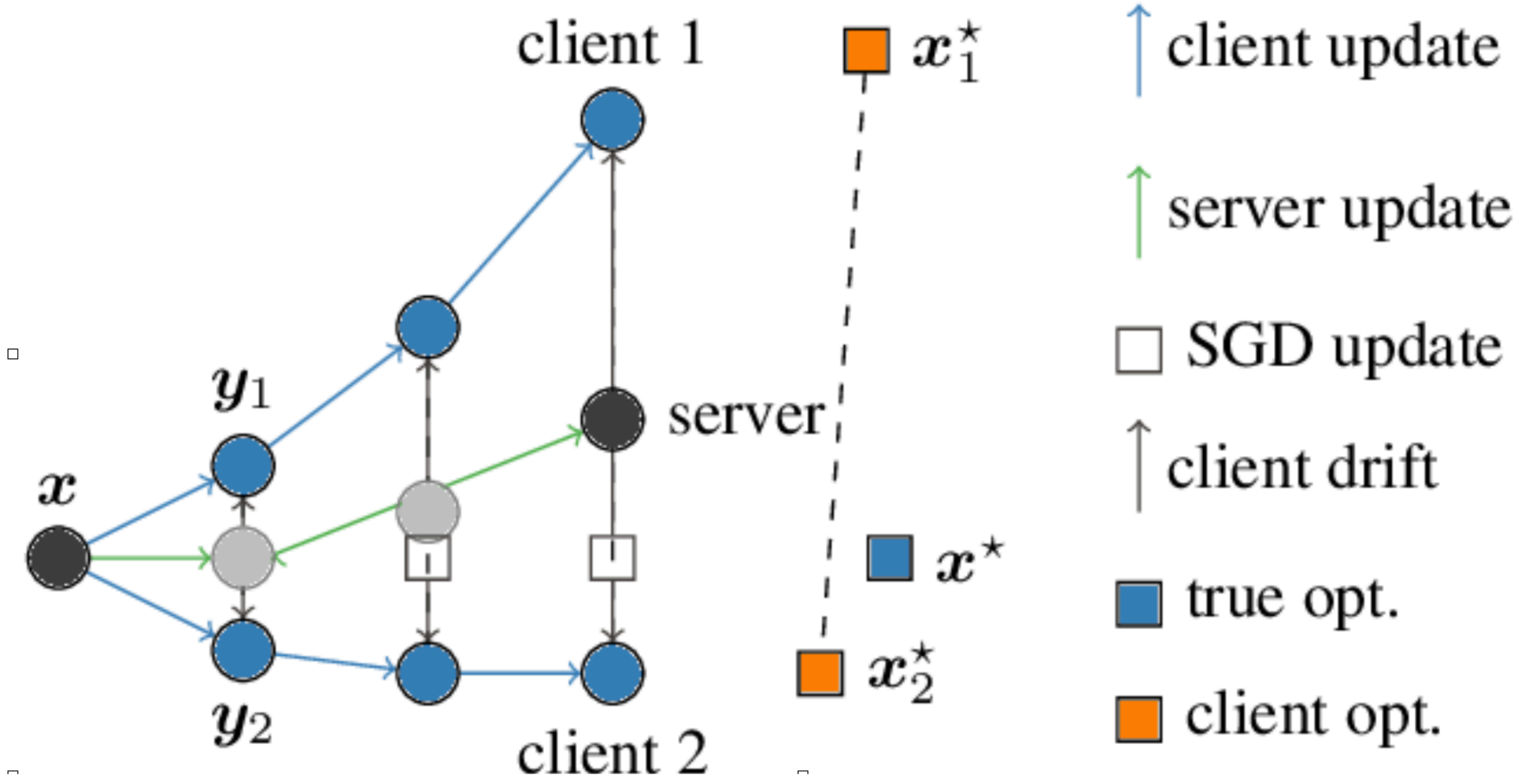


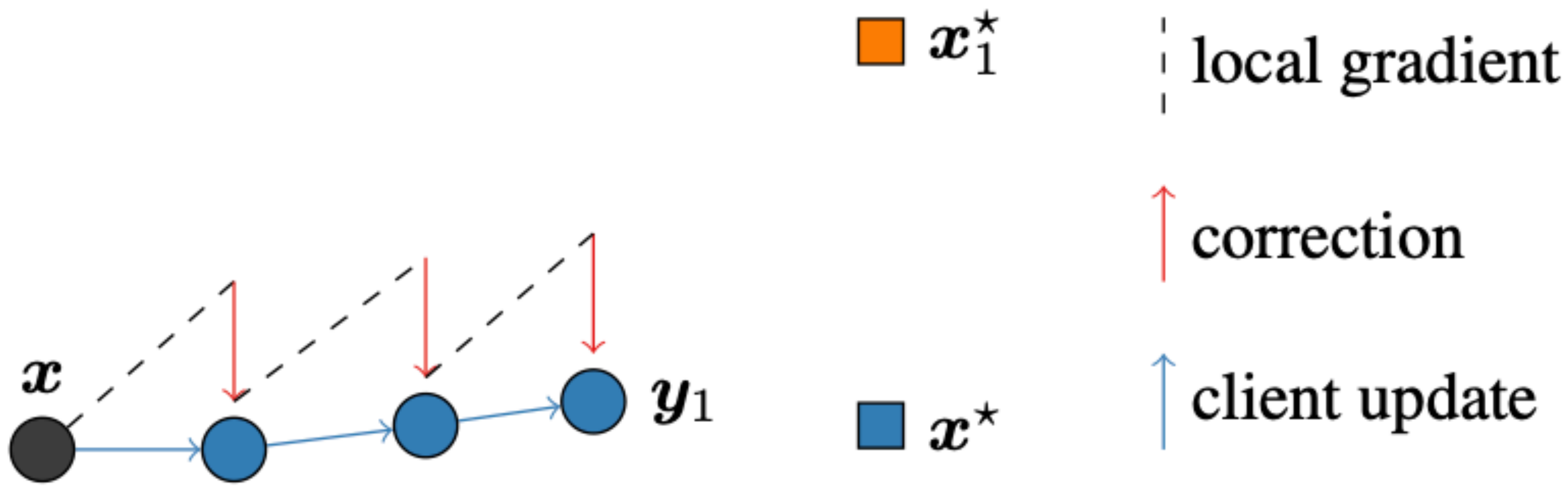
FedAvg suffers from client drift

Karimireddy et al. "Scaffold: Stochastic controlled averaging for federated learning." In *International Conference on Machine Learning*, pp. 5132-5143. PMLR, 2020.

Client Drift for Heterogeneous Data

Heterogeneous Data: different client has different data distribution





FedAvg suffers from client drift

a different client optimization helps!

Karimireddy et al. "Scaffold: Stochastic controlled averaging for federated learning." In *International Conference on Machine Learning*, pp. 5132-5143. PMLR, 2020.

Experiments

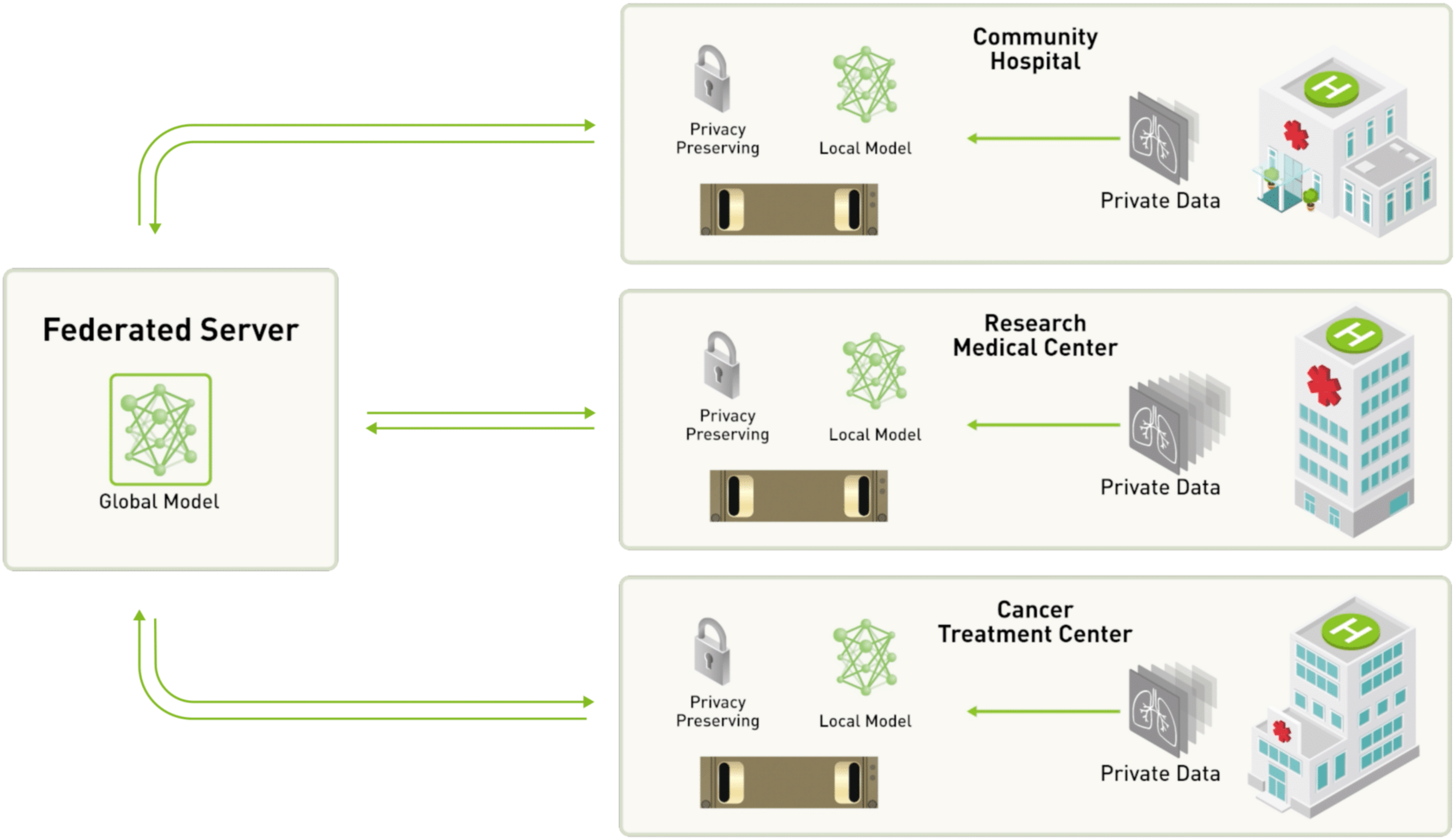
Table 5. Best test accuracy after 1k rounds with 2-layer fully connected neural network (non-convex) on EMNIST trained with 5 epochs per round (25 steps) for the local methods, and 20% of clients sampled each round. SCAFFOLD has the best accuracy and SGD has the least. SCAFFOLD again outperforms other methods. SGD is unaffected by similarity, whereas the local methods improve with client similarity.

	0% similarity	10% similarity
SGD	0.766	0.764
FEDAVG	0.787	0.828
SCAFFOLD	**0.801**	**0.842**

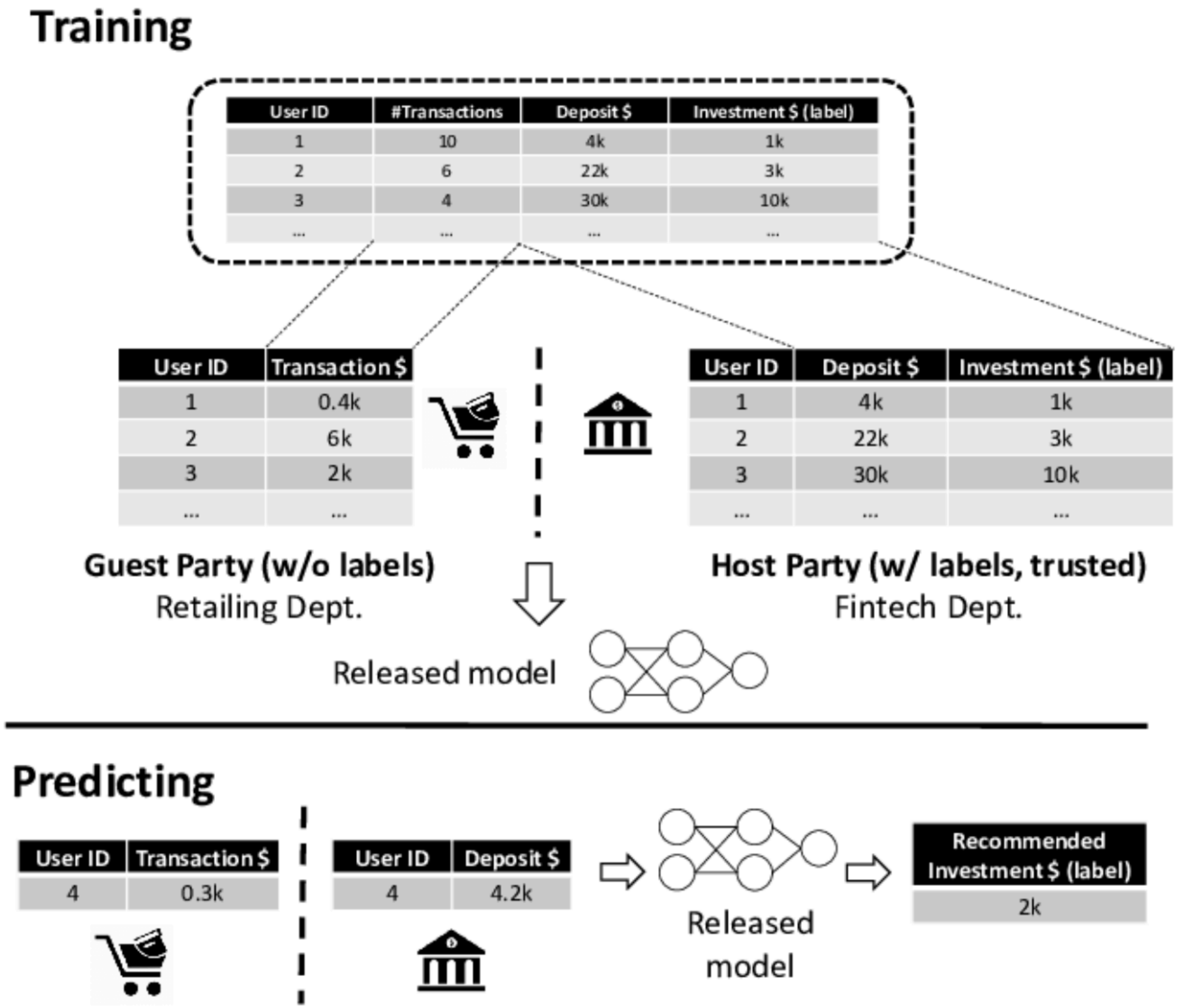
Karimireddy et al. "Scaffold: Stochastic controlled averaging for federated learning." In *International Conference on Machine Learning*, pp. 5132-5143. PMLR, 2020.

Applications

Applications in Healthcare

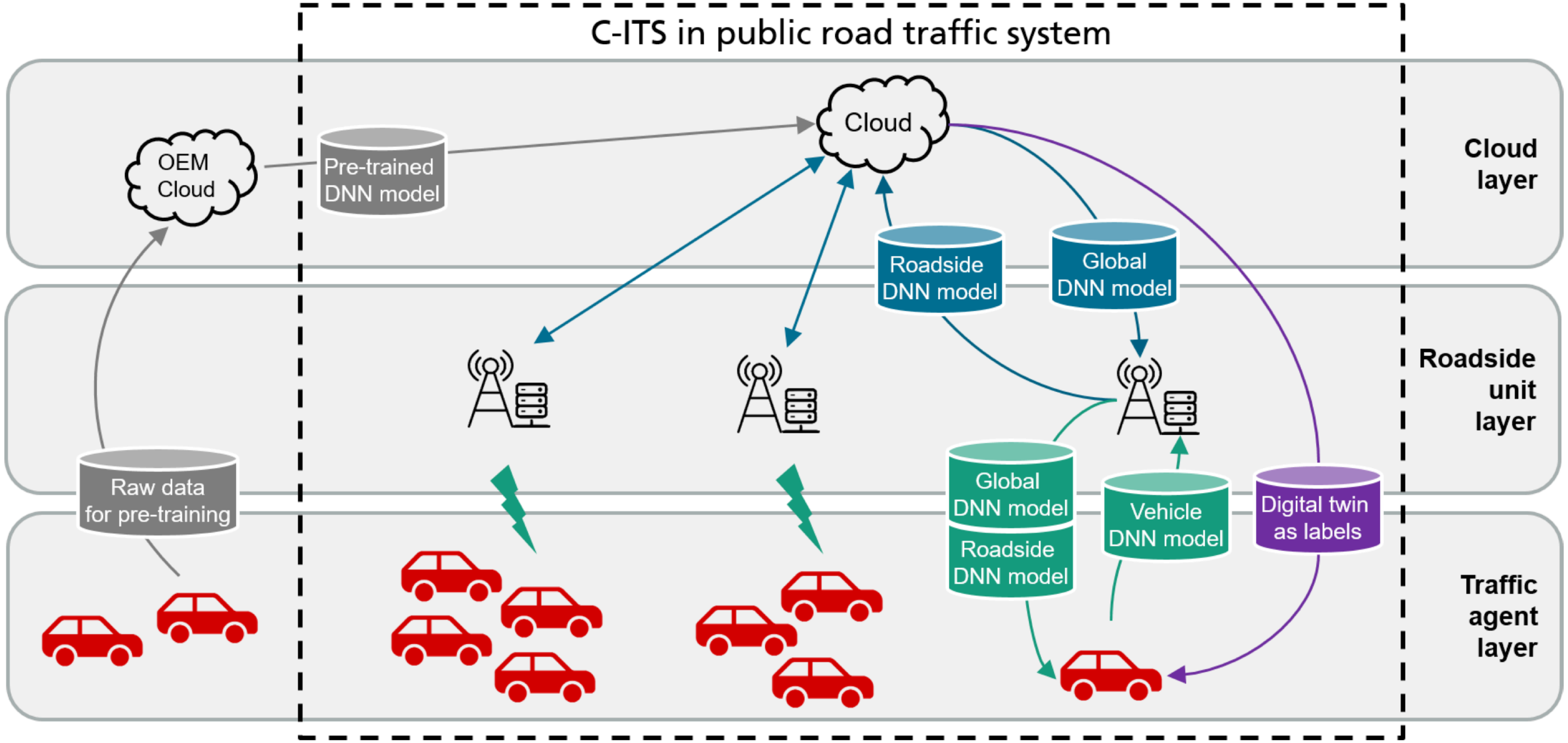


Applications in Financial Technology



Wu et al. Practical Vertical Federated Learning with Unsupervised Representation Learning. arXiv 2022.

Applications in Transportation



Guo et al. Federated Learning Framework Coping with Hierarchical Heterogeneity in Cooperative ITS. arXiV 2022.

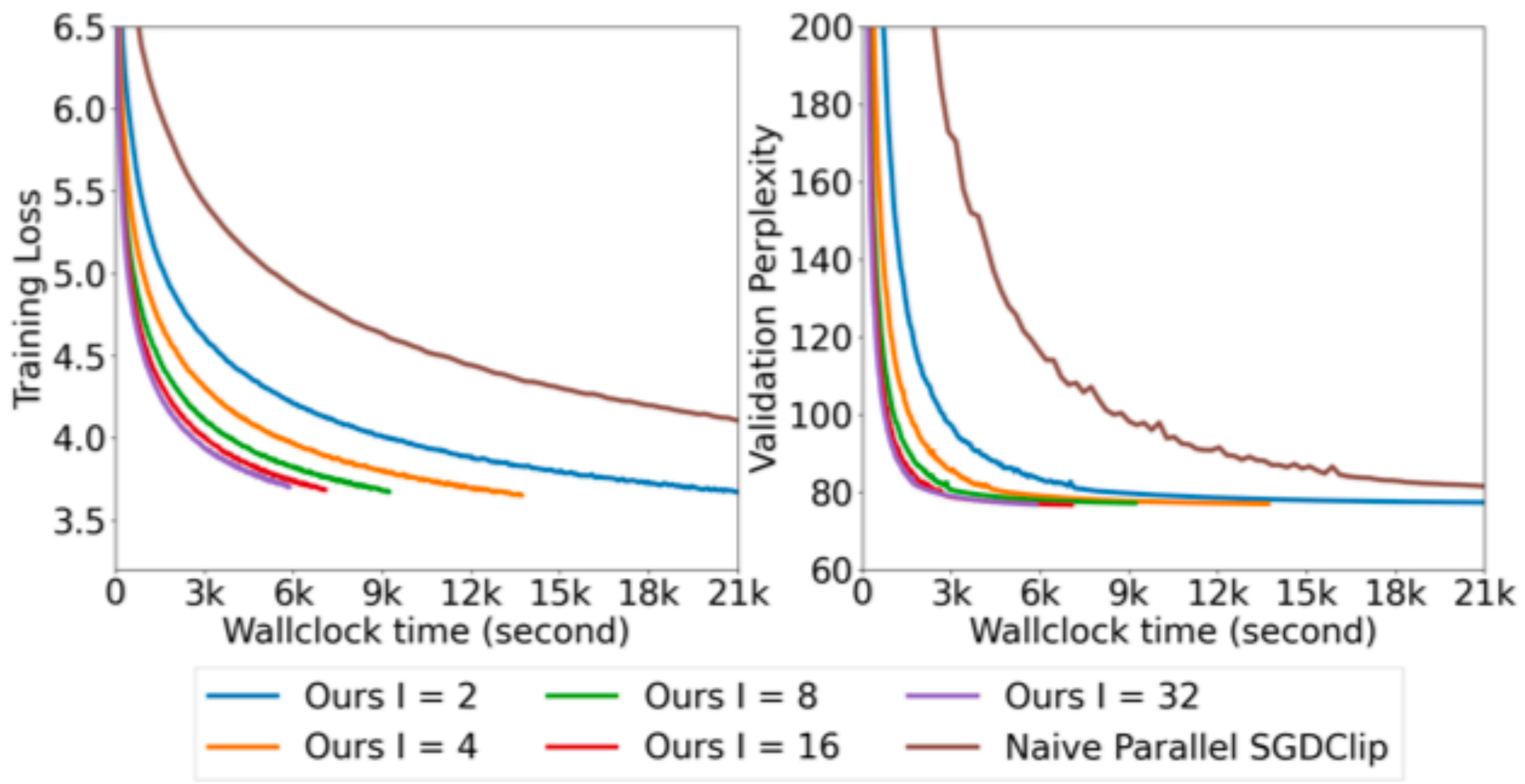
Ongoing Research and Open Problems

Ongoing FL Research in My Lab

- FL Algorithm Design for Natural Language Processing tasks
  [L.-Zhuang-Lei-Liao, NeurIPS 22], [Crawshaw-Bao-L., ICLR 23]

Ongoing FL Research in My Lab

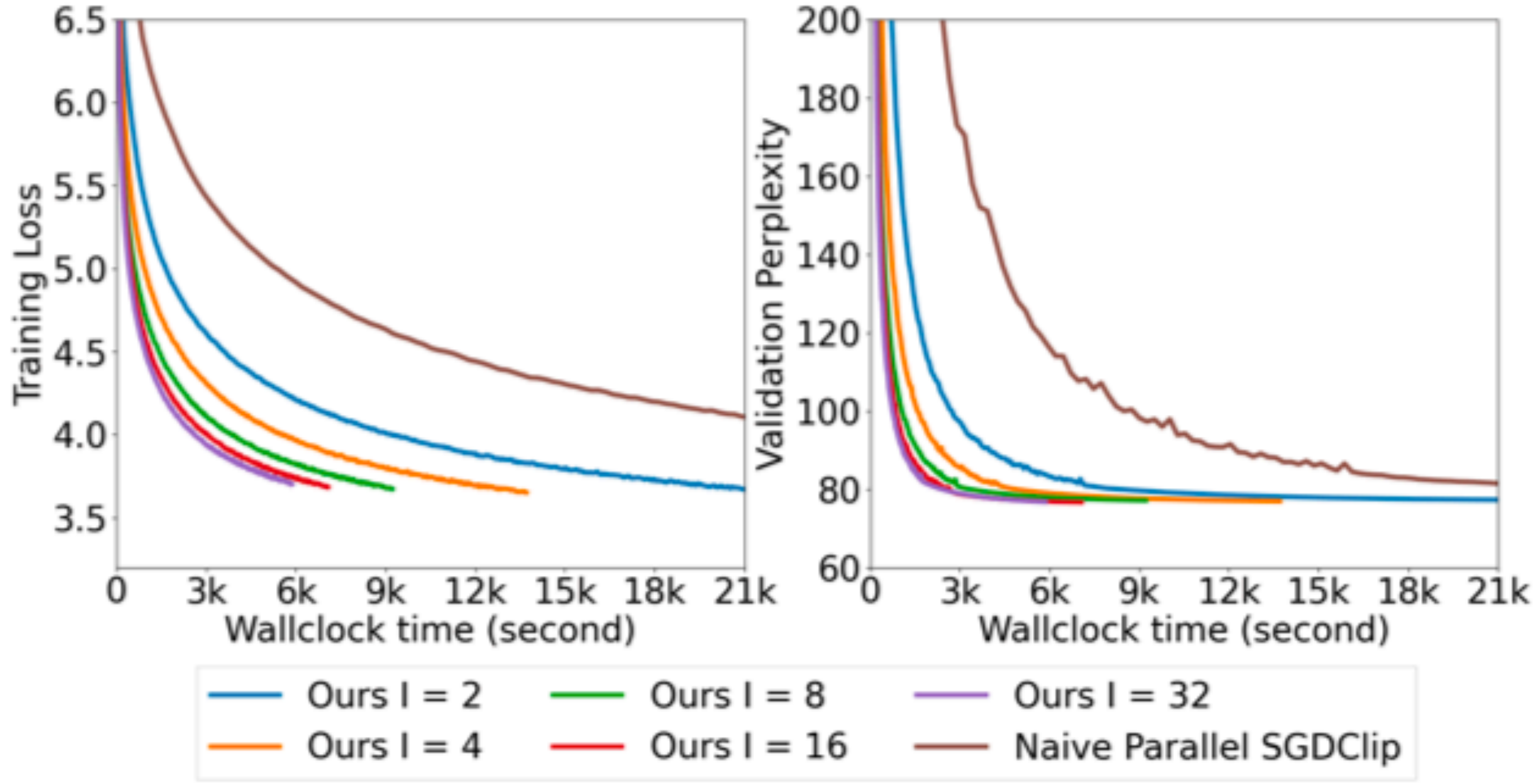
- FL Algorithm Design for Natural Language Processing tasks
  [L.-Zhuang-Lei-Liao, NeurIPS 22], [Crawshaw-Bao-L., ICLR 23]

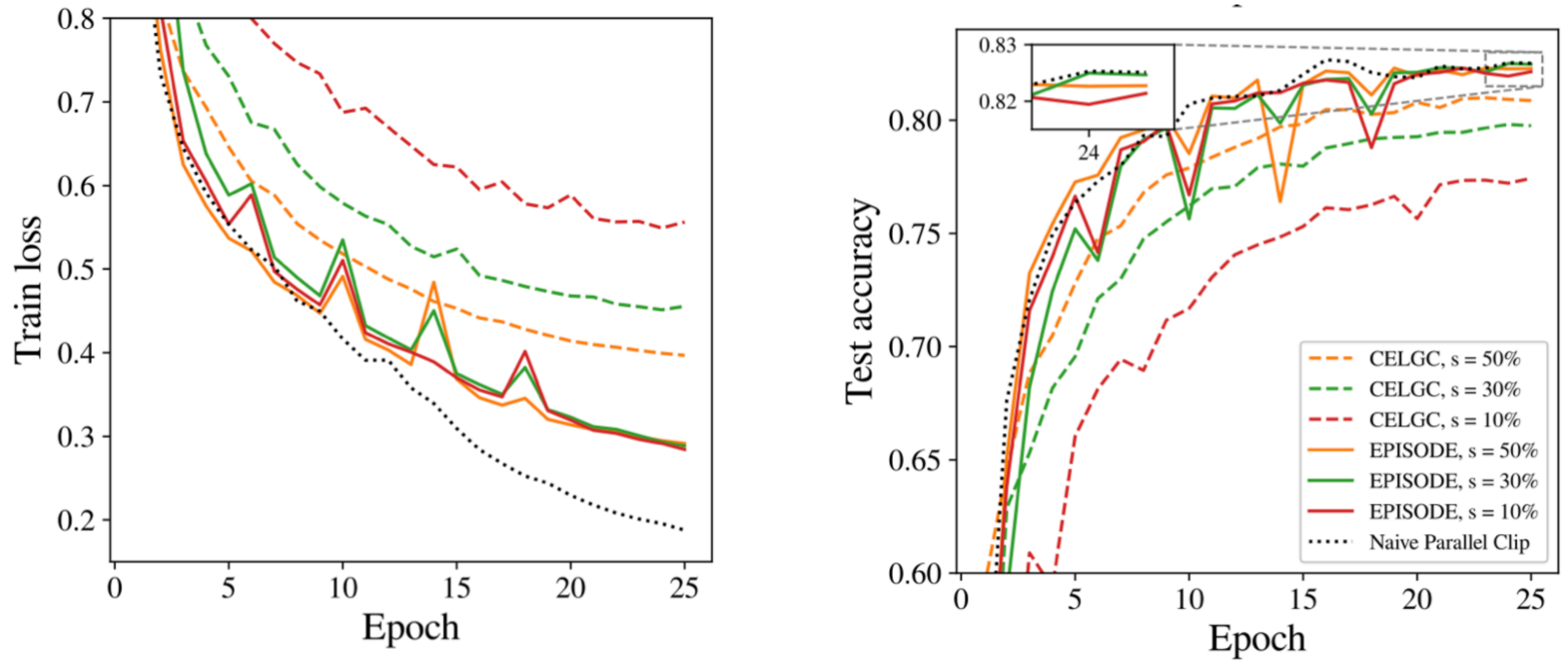


- Train a LSTMs on next word prediction task on Penn Treebank
- Homogeneous data: each client has the same data distribution
- Our algorithm can allow multiple gradient steps (i.e., $I > 1$) but it accelerates the training speed

Ongoing FL Research in My Lab

- FL Algorithm Design for Natural Language Processing tasks
  [L.-Zhuang-Lei-Liao, NeurIPS 22], [Crawshaw-Bao-L., ICLR 23]

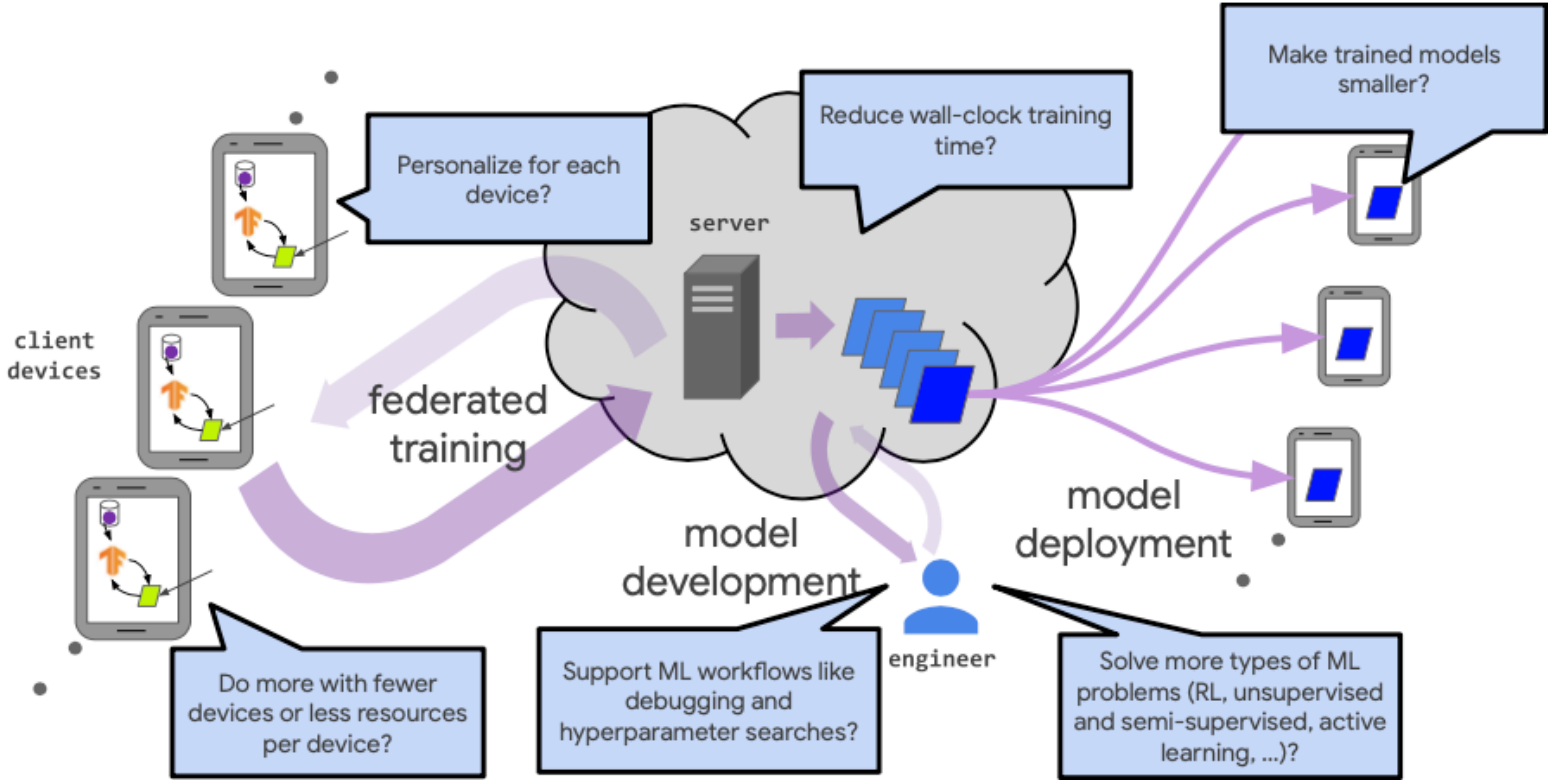




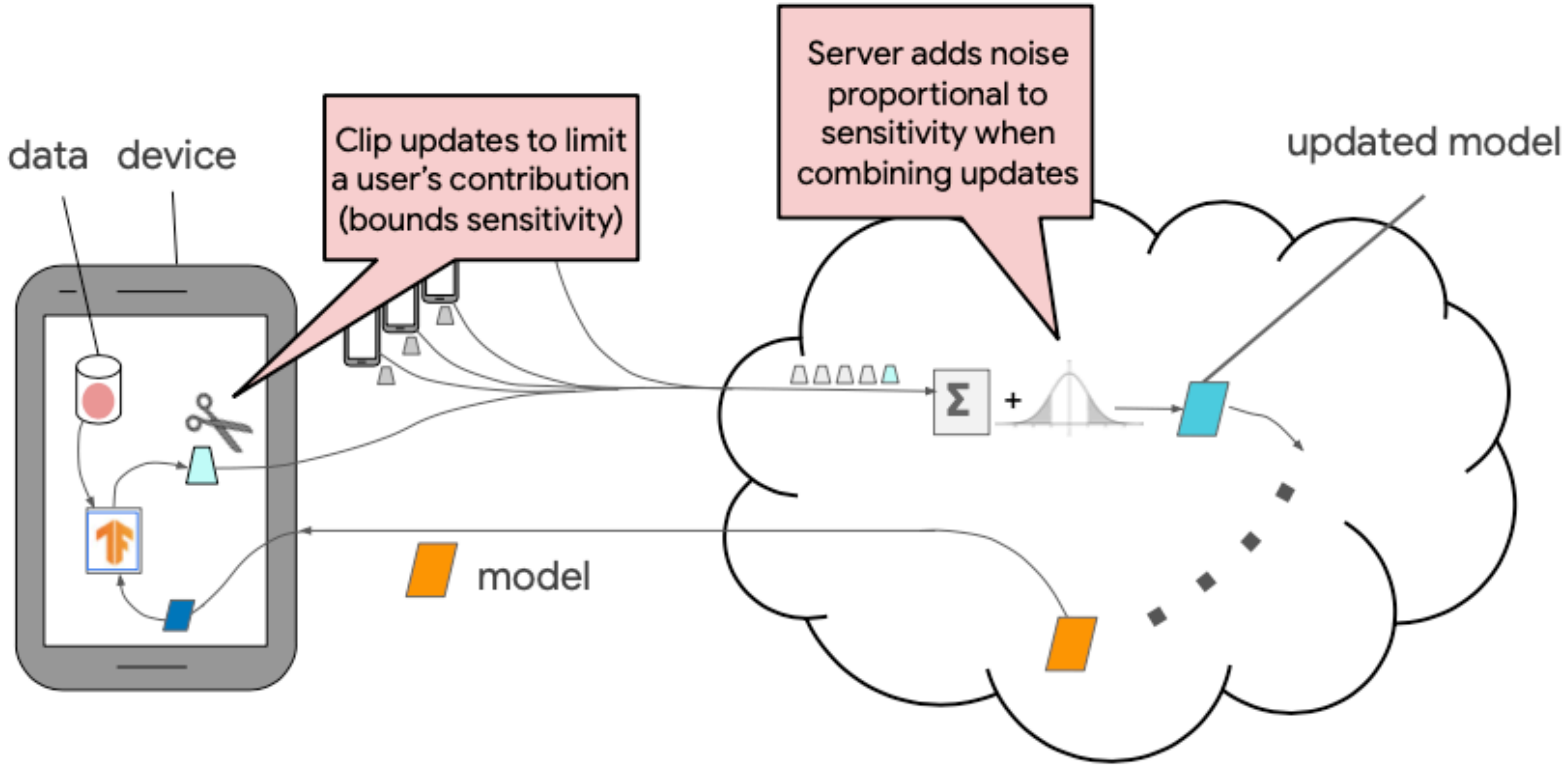
- Train a LSTMs on next word prediction task on Penn Treebank
- Homogeneous data: each client has the same data distribution
- Our algorithm can allow multiple gradient steps (i.e., $I > 1$) but it accelerates the training speed

- Train a recurrent neural network on SNLI dataset (text classification) on eight GPUs
- Heterogeneous data: larger similarity (s) indicates smaller heterogeneity
- Our algorithm EPISODE does not suffer from high heterogeneity

Improving Efficiency and Effectiveness



Differential Private Federated Learning



Q: Can we design algorithms with best utility-privacy tradeoff?

Thank you for your attention!