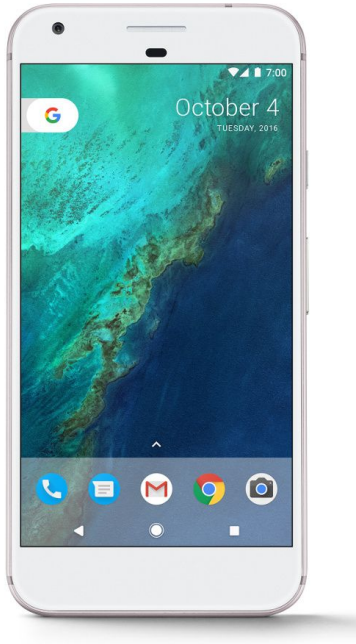
Part I: What is Federated Learning?

Data is born at the edge

Billions of phones & IoT devices constantly generate data

Data enables better products and smarter models

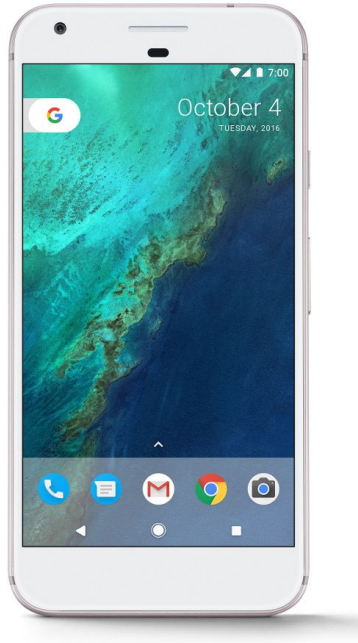


Can data live at the edge?

Data processing is moving on device:

- Improved latency
- Works offline
- Better battery life
- Privacy advantages

E.g., on-device inference for mobile keyboards and cameras.



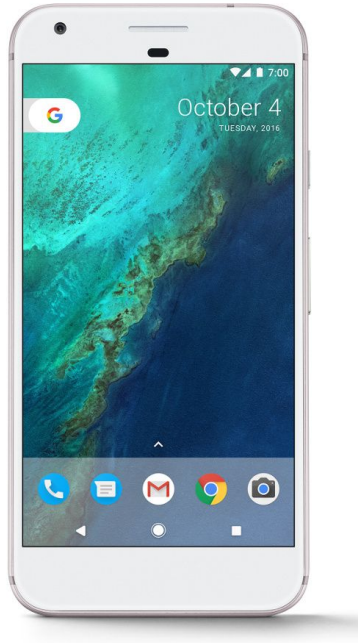
Can data live at the edge?

Data processing is moving on device:

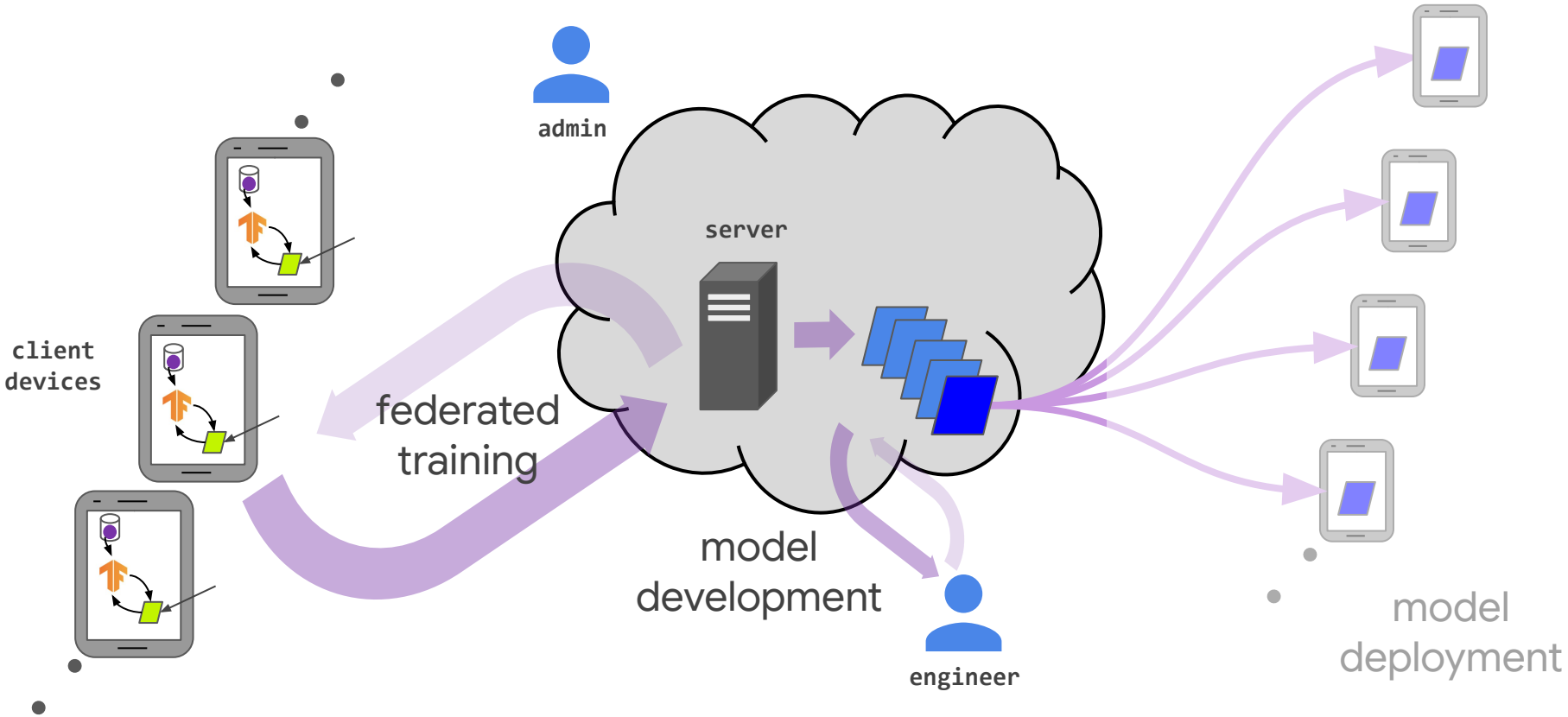
- Improved latency
- Works offline
- Better battery life
- Privacy advantages

E.g., on-device inference for mobile keyboards and cameras.

What about analytics?
What about learning?



Cross-device federated learning



Applications of cross-device federating learning

What makes a good application?

- On-device data is more relevant than server-side proxy data
- On-device data is privacy sensitive or large
- Labels can be inferred naturally from user interaction

Example applications

- Language modeling for mobile keyboards and voice recognition
- Image classification for predicting which photos people will share
- ...

Gboard: next-word prediction

Federated RNN (compared to prior n-gram model):

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

A. Hard, *et al*. **Federated Learning for Mobile Keyboard Prediction**. *arXiv:1811.03604*

Other federated models in Gboard

Emoji prediction

- 7% more accurate emoji predictions
- prediction strip clicks +4% more
- 11% more users share emojis!

Ramaswamy, *et al.* **Federated Learning for Emoji Prediction in a Mobile Keyboard**. arXiv:1906.04329.

Action prediction

When is it useful to suggest a gif, sticker, or search query?

- 47% reduction in unhelpful suggestions
- increasing overall emoji, gif, and sticker shares

T. Yang, *et al*. **Applied Federated Learning: Improving Google Keyboard Query Suggestions**. *arXiv:1812.02903*

Discovering new words

Federated discovery of what words people are typing that Gboard doesn't know.

M. Chen, *et al*. **Federated Learning Of Out-Of-Vocabulary Words**. *arXiv:1903.10635*

Cross-device federated learning at Apple

MIT Technology Review

Sign in Subscribe

Artificial intelligence / Machine learning

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

*"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/

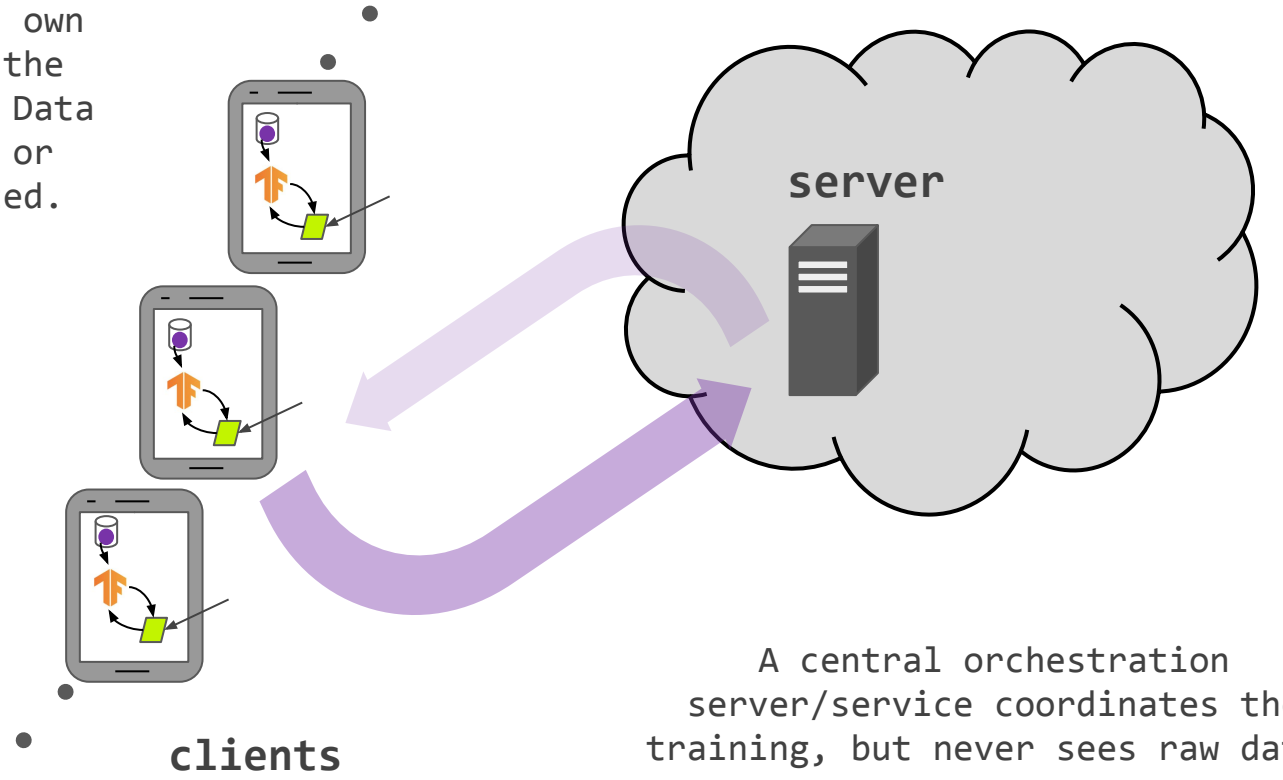
Federated Learning

Federated learning is a machine learning setting where multiple entities (clients) collaborate in solving a machine learning problem, under the coordination of a central server or service provider. Each client's raw data is stored locally and not exchanged or transferred; instead, focused updates intended for immediate aggregation are used to achieve the learning objective.

definition proposed in
Advances and Open Problems in Federated Learning (arxiv/1912.04977)

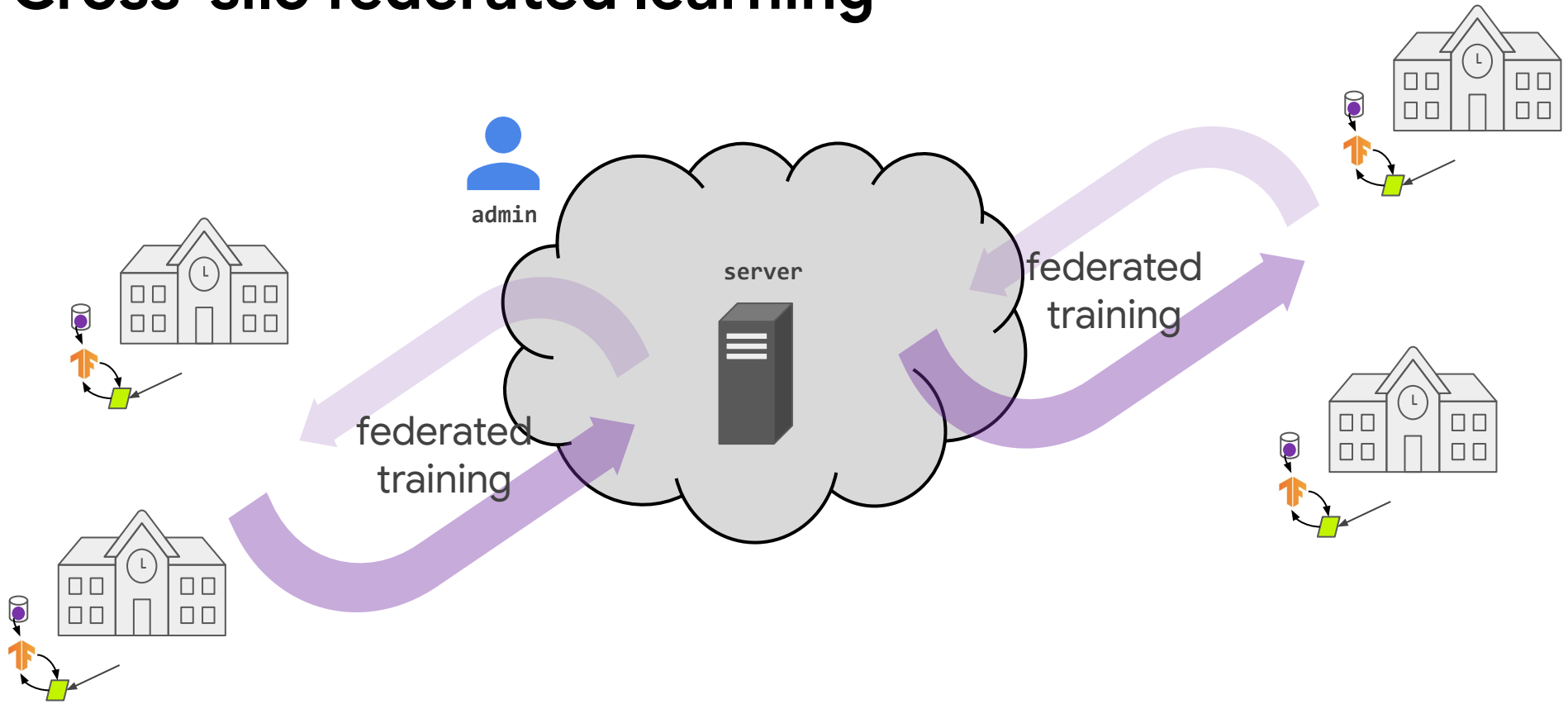
Federated learning - defining characteristics

Data is generated locally and remains decentralized. Each client stores its own data and cannot read the data of other clients. Data is not independently or identically distributed.



A central orchestration server/service coordinates the training, but never sees raw data.

Cross-silo federated learning



Cross-silo federated learning from Intel

ARTIFICIAL INTELLIGENCE, DIAGNOSTICS

UPenn, Intel partner to use federated learning AI for early brain tumor detection

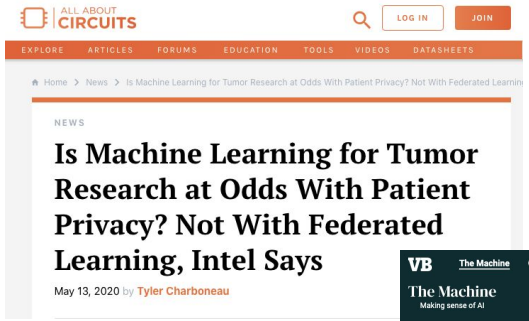
The project will bring in 29 institutions from North America, Europe and India and will use privacy-preserved data to train AI models. Federated learning has been described as being born at the intersection of AI, blockchain, edge computing and the Internet of Things.

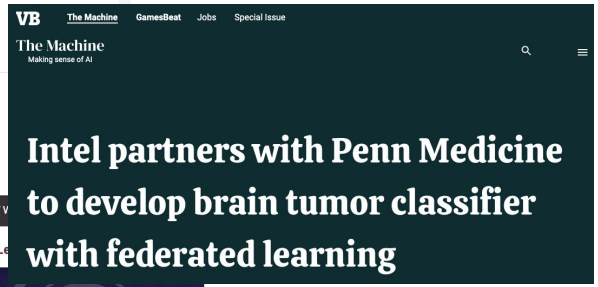
By ALARIC DEARMENT

Post a comment / May 11, 2020 at 10:03 AM

"The University of Pennsylvania and chipmaker Intel are forming a partnership to enable 29 heatlhcare and medical research institutions around the world to train artificial intelligence models to detect brain tumors early."

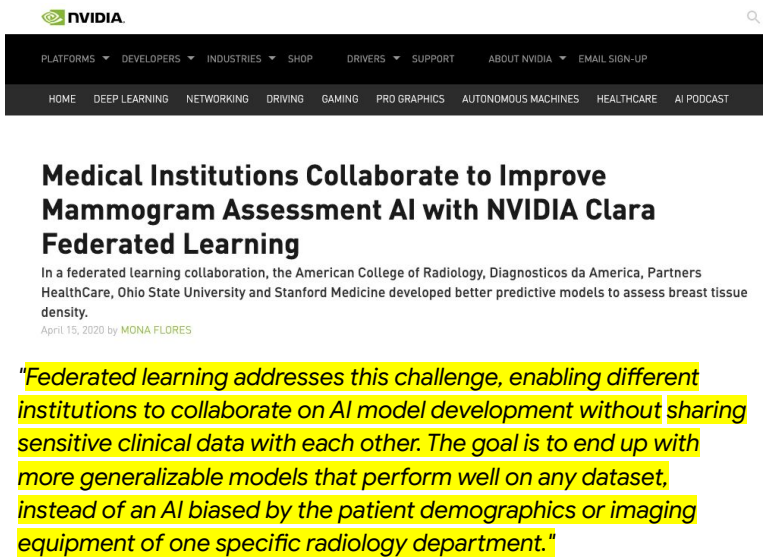
"The program will rely on a technique known as federated learning, which enables institutions to collaborate on deep learning projects without sharing patient data. The partnership will bring in institutions in the U.S., Canada, U.K., Germany, Switzerland and India. The centers – which include Washington University of St. Louis; Queen's University in Kingston, Ontario; University of Munich; Tata Memorial Hospital in Mumbai and others – will use Intel's federated learning hardware and software."



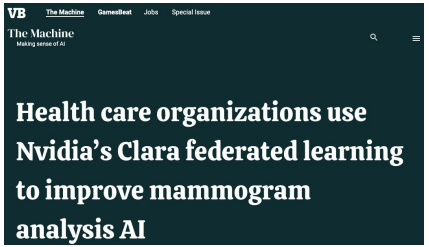


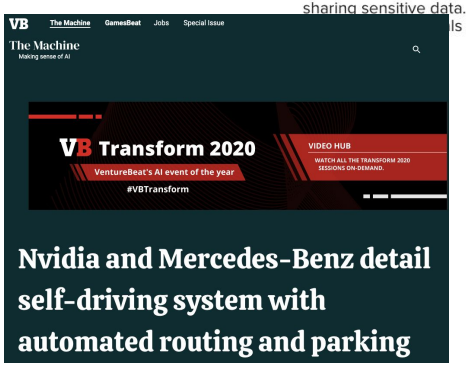
[1] https://medcitynews.com/2020/05/upenn-intel-partner-to-use-federated-learning-ai-for-early-brain-tumor-detection/
[2] https://www.allaboutcircuits.com/news/can-machine-learning-keep-patient-privacy-for-tumor-research-intel-says-yes-with-federated-learning/
[3] https://venturebeat.com/2020/05/11/intel-partners-with-penn-medicine-to-develop-brain-tumor-classifier-with-federated-learning/
[4] http://www.bio-itworld.com/2020/05/28/intel-penn-medicine-launch-federated-learning-model-for-brain-tumors.aspx

Cross-silo federated learning from NVIDIA



"Federated learning addresses this challenge, enabling different institutions to collaborate on AI model development without sharing sensitive clinical data with each other. The goal is to end up with more generalizable models that perform well on any dataset, instead of an AI biased by the patient demographics or imaging equipment of one specific radiology department."

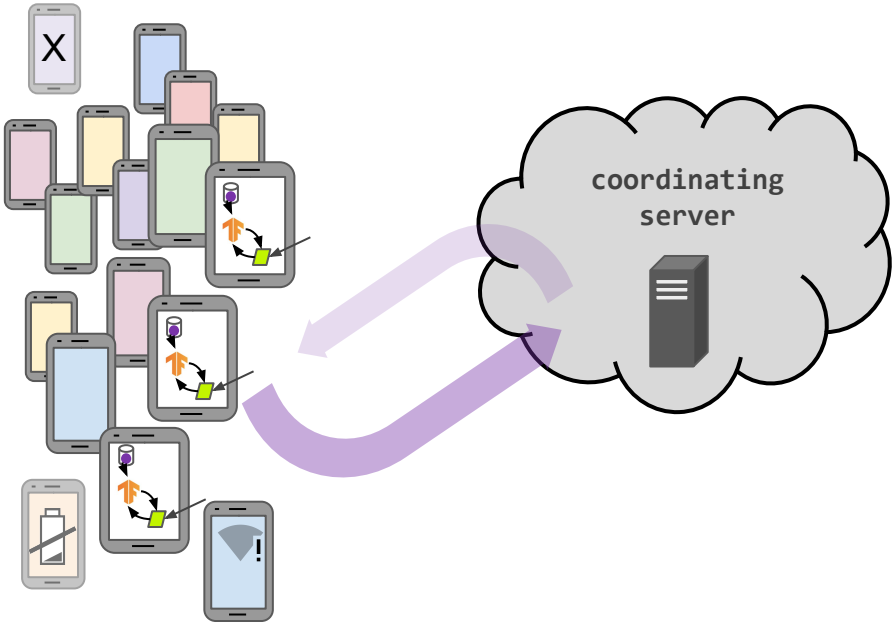




[1] https://blogs.nvidia.com/blog/2020/04/15/federated-learning-mammogram-assessment/
[2] https://venturebeat.com/2020/04/15/healthcare-organizations-use-nvidias-clara-federated-learning-to-improve-mammogram-analysis-ai/
[3] https://medcitynews.com/2020/01/nvidia-says-it-has-a-solution-for-healthcares-data-problems/
[4] https://venturebeat.com/2020/06/23/nvidia-and-mercedes-benz-detail-self-driving-system-with-automated-routing-and-parking/

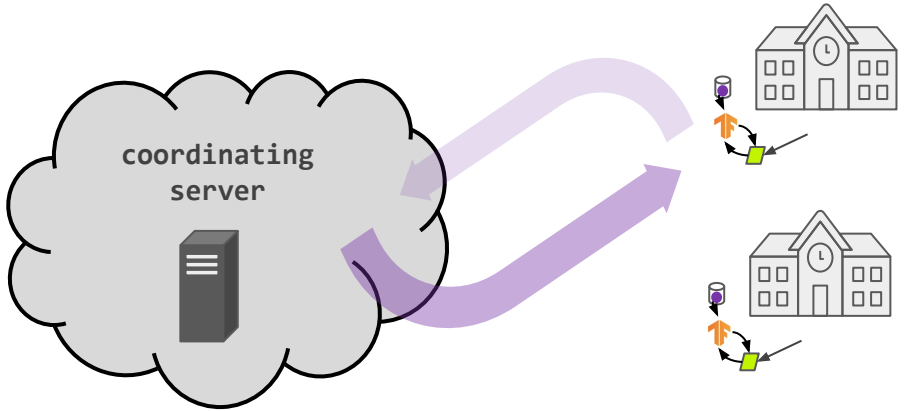
Cross-device federated learning

millions of intermittently
available client devices



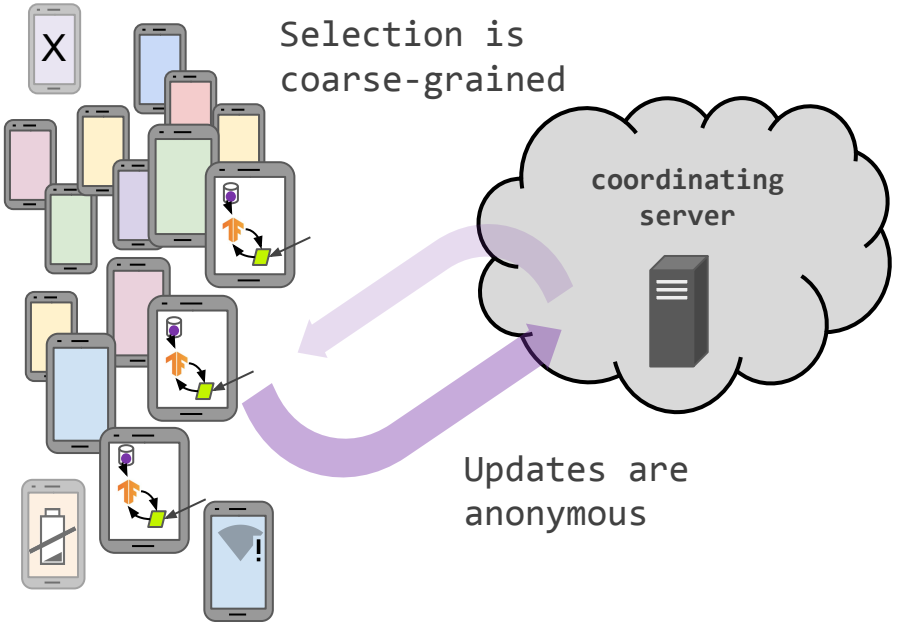
Cross-silo federated learning

small number of clients
(institutions, data silos),
high availability



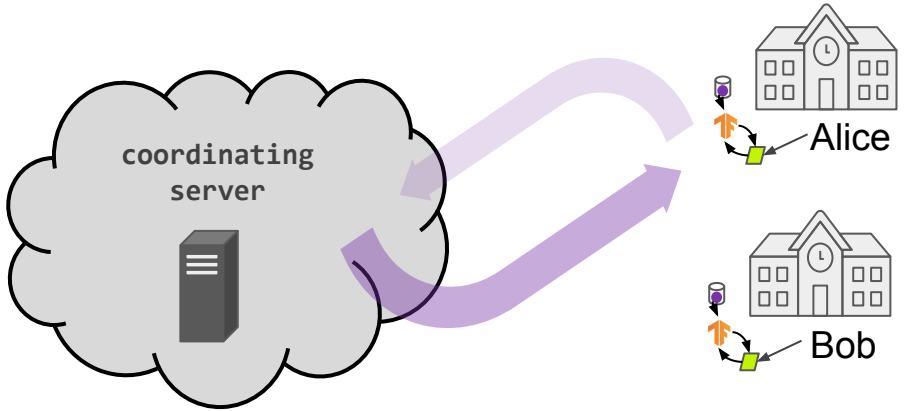
Cross-device federated learning

clients cannot be indexed directly (i.e., no use of client identifiers)



Cross-silo federated learning

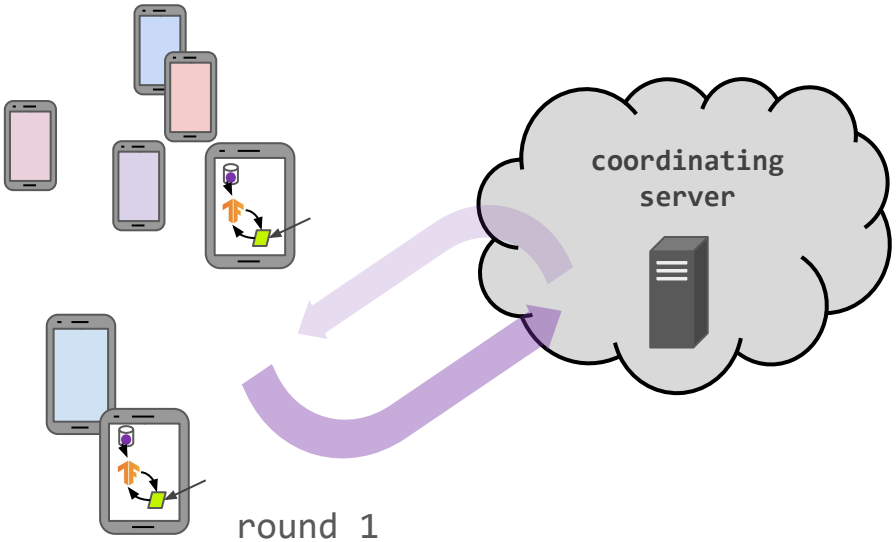
each client has an identity or name that allows the system to access it specifically



Cross-device federated learning

Server can only access a (possibly biased) random sample of clients on each round.

Large population => most clients only participate once.

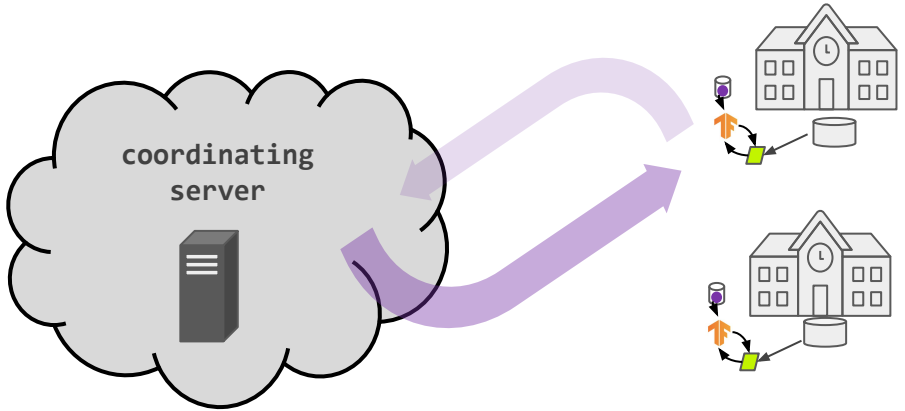


round 1

Cross-silo federated learning

Most clients participate in every round.

Clients can run algorithms that maintain local state across rounds.

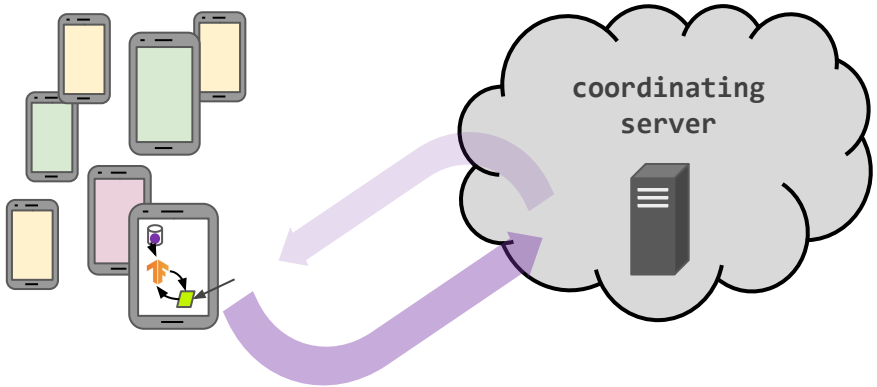


round 1

Cross-device federated learning

Server can only access a (possibly biased) random sample of clients on each round.

Large population => most clients only participate once.

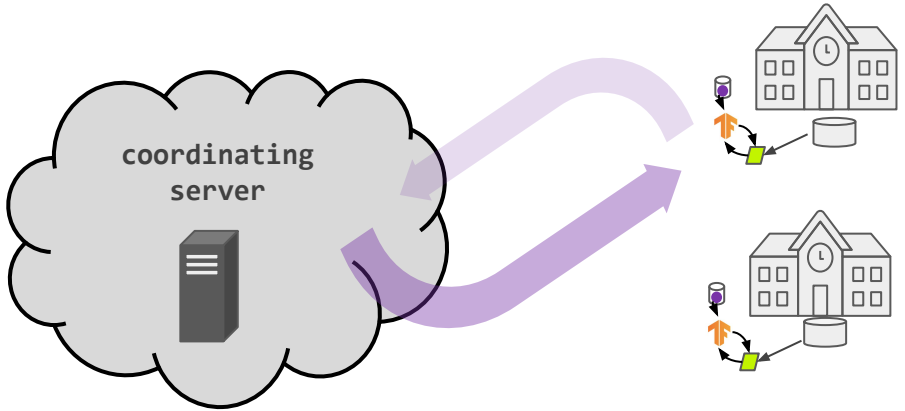


round 2
(completely new set of devices participate)

Cross-silo federated learning

Most clients participate in every round.

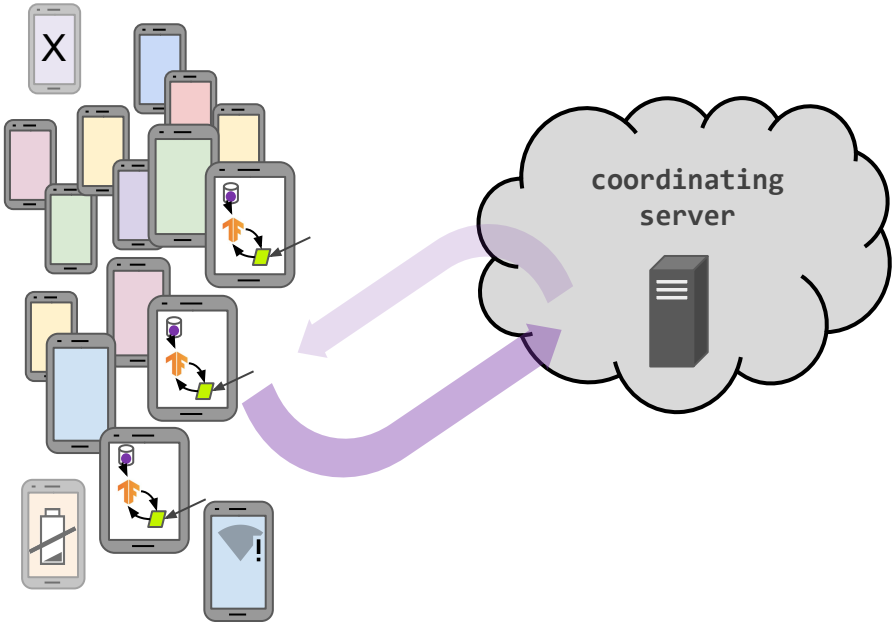
Clients can run algorithms that maintain local state across rounds.



round 2
(same clients)

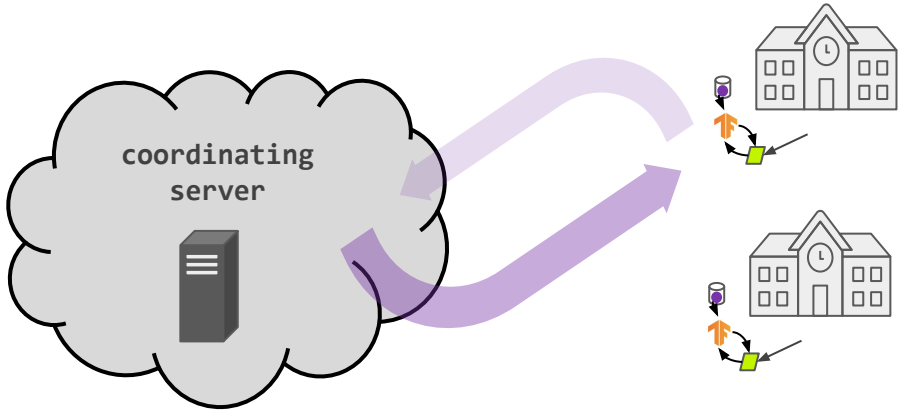
Cross-device federated learning

communication is often the primary bottleneck



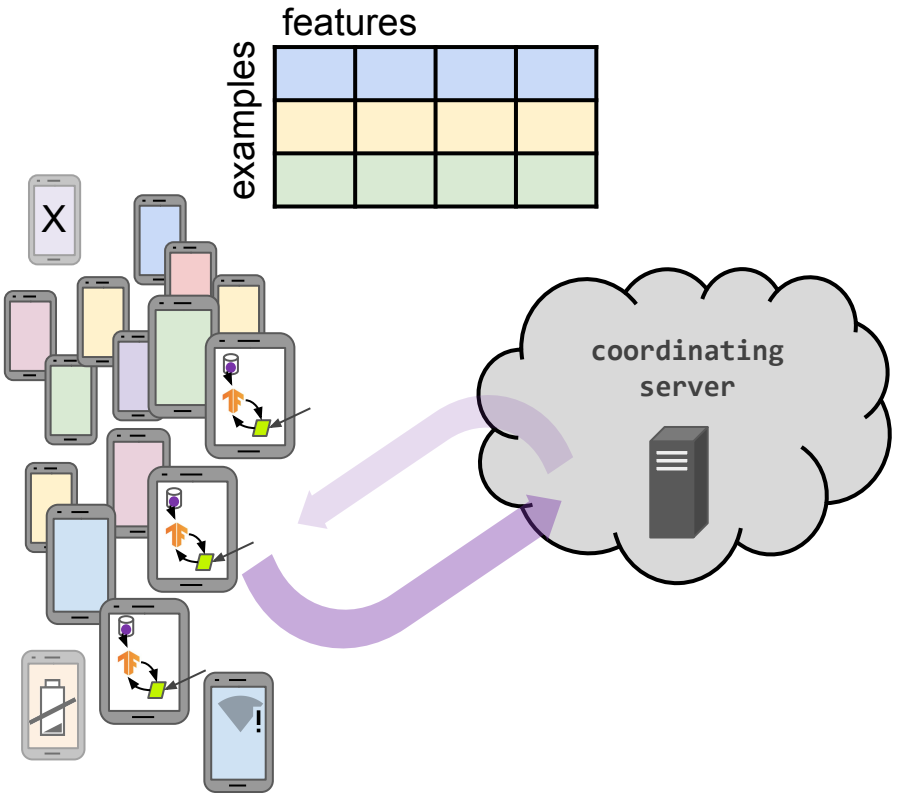
Cross-silo federated learning

communication or computation might be the primary bottleneck



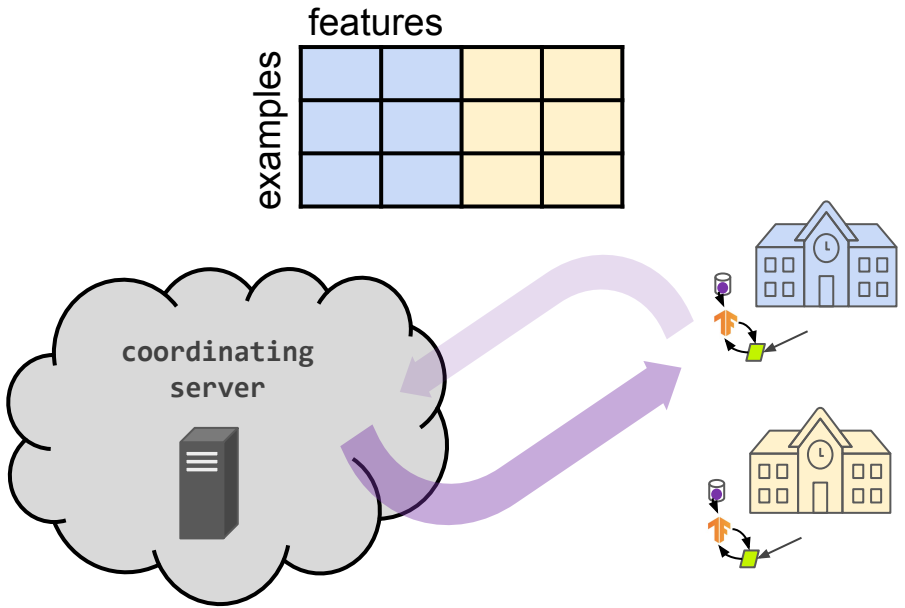
Cross-device federated learning

`horizontally partitioned data`

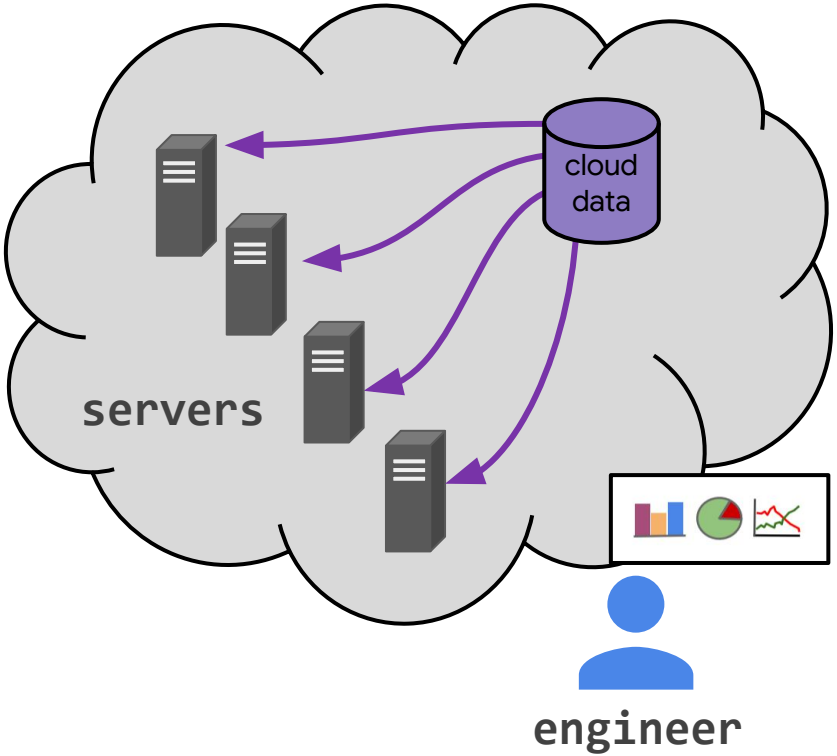


Cross-silo federated learning

`horizontal or`
`vertically partitioned data`



Distributed datacenter machine learning



FL terminology

- **Clients** - Compute nodes also holding local data, usually belonging to one entity:
  - IoT devices
  - Mobile devices
  - Data silos
  - Data centers in different geographic regions

- **Server** - Additional compute nodes that coordinate the FL process but don't access raw data. Usually not a single physical machine.

Characteristics of the federated learning setting

	Datacenter distributed learning	Cross-silo federated learning	Cross-device federated learning
Setting	Training a model on a large but "flat" dataset. Clients are compute nodes in a single cluster or datacenter.	Training a model on siloed data. Clients are different organizations (e.g., medical or financial) or datacenters in different geographical regions.	The clients are a very large number of mobile or IoT devices.
Data distribution	Data is centrally stored, so it can be shuffled and balanced across clients. Any client can read any part of the dataset.	**Data is generated locally and remains decentralized.** Each client stores its own data and cannot read the data of other clients. Data is not independently or identically distributed.	
Orchestration	Centrally orchestrated.	A central orchestration server/service organizes the training, but never sees raw data.	
Wide-area communication	None (fully connected clients in one datacenter/cluster).	Typically hub-and-spoke topology, with the hub representing a coordinating service provider (typically without data) and the spokes connecting to clients.	
Data availability	All clients are almost always available.		Only a fraction of clients are available at any one time, often with diurnal and other variations.
Distribution scale	Typically 1 - 1000 clients.	Typically 2 - 100 clients.	Massively parallel, up to 10^{10} clients.

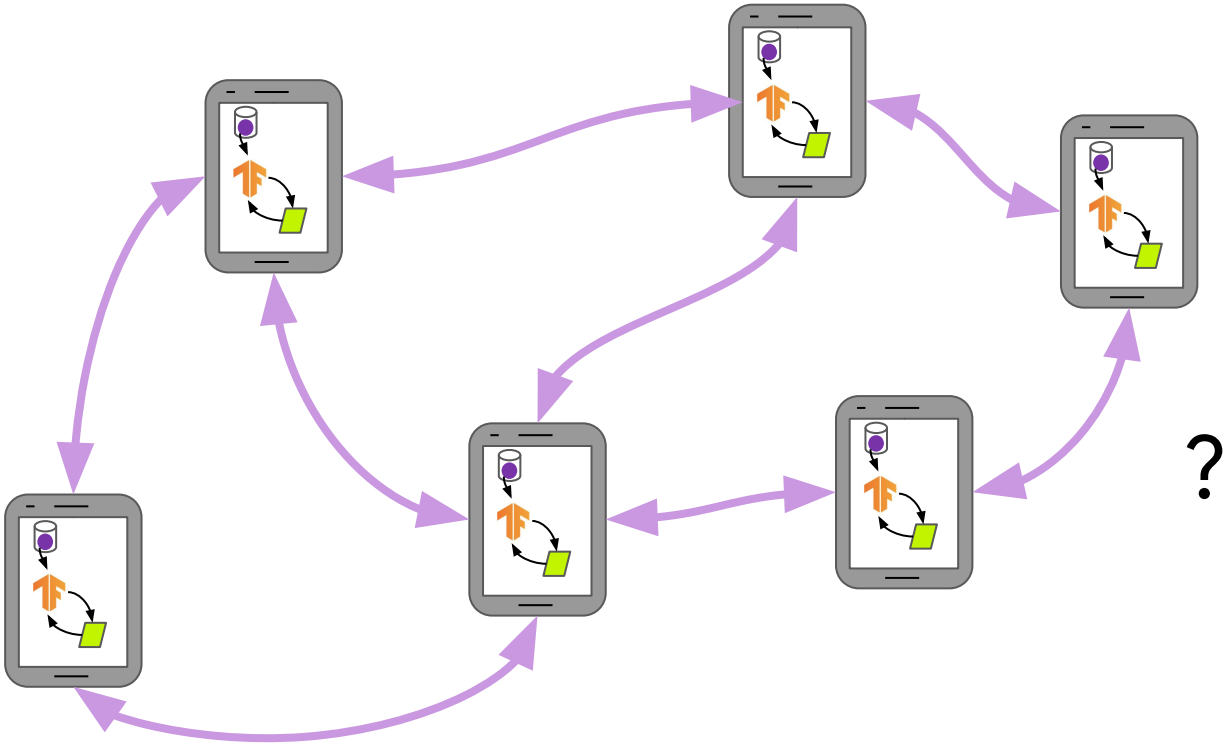
Adapted from Table 1 in *Advances and Open Problems in Federated Learning* (arxiv/1912.04977)

Characteristics of the federated learning setting

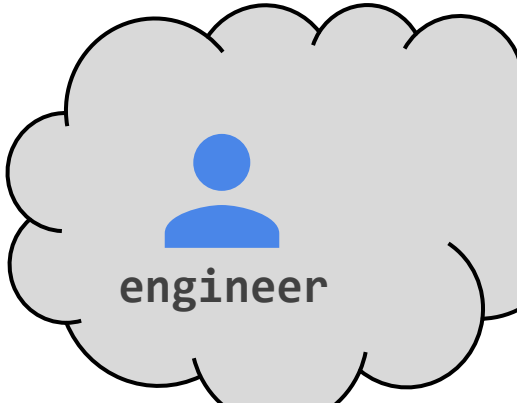
	Datacenter distributed learning	Cross-silo federated learning	Cross-device federated learning
Addressability	Each client has an identity or name that allows the system to access it specifically.		Clients cannot be indexed directly (i.e., no use of client identifiers)
Client statefulness	Stateful --- each client may participate in each round of the computation, carrying state from round to round.		Generally stateless --- each client will likely participate only once in a task, so generally we assume a fresh sample of never before seen clients in each round of computation.
Primary bottleneck	Computation is more often the bottleneck in the datacenter, where very fast networks can be assumed.	Might be computation or communication.	Communication is often the primary bottleneck, though it depends on the task. Generally, federated computations uses wi-fi or slower connections.
Reliability of clients	Relatively few failures.		Highly unreliable --- 5% or more of the clients participating in a round of computation are expected to fail or drop out (e.g., because the device becomes ineligible when battery, network, or idleness requirements for training/computation are violated).
Data partition axis	Data can be partitioned / re-partitioned arbitrarily across clients.	Partition is fixed. Could be example-partitioned (horizontal) or feature-partitioned (vertical).	Fixed partitioning by example (horizontal).

Adapted from Table 1 in *Advances and Open Problems in Federated Learning* (arxiv/1912.04977)

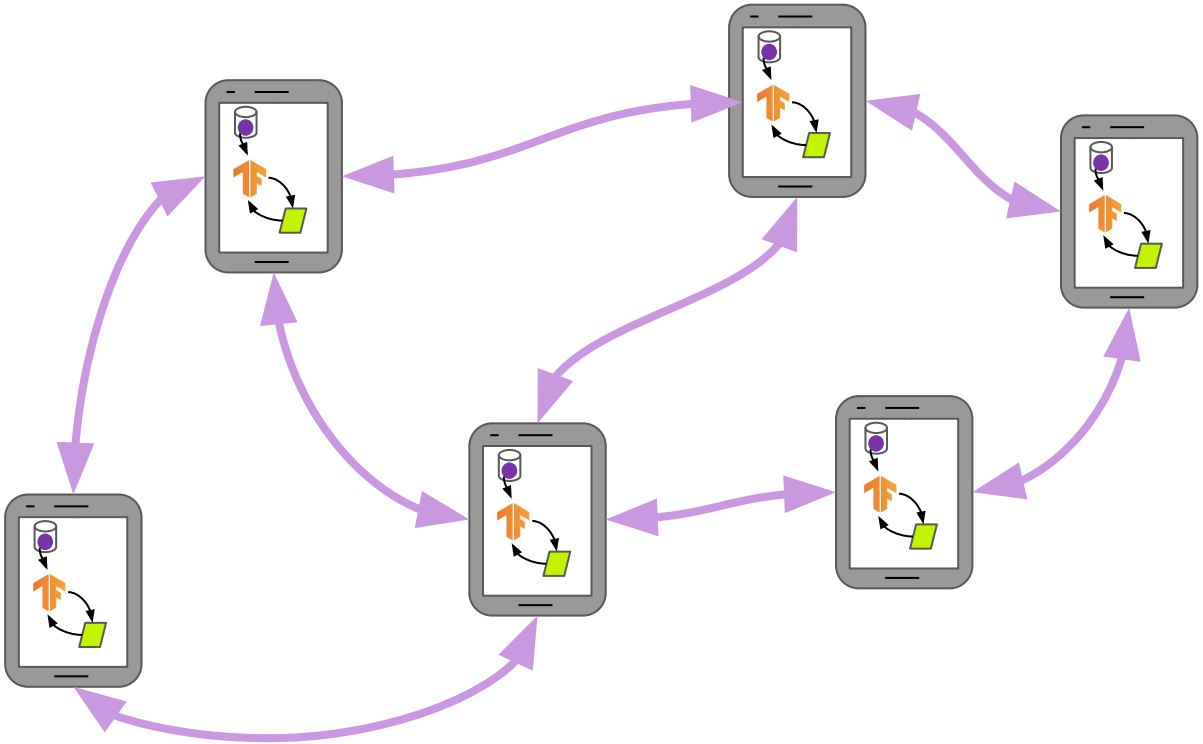
Fully decentralized (peer-to-peer) learning



?



Fully decentralized (peer-to-peer) learning



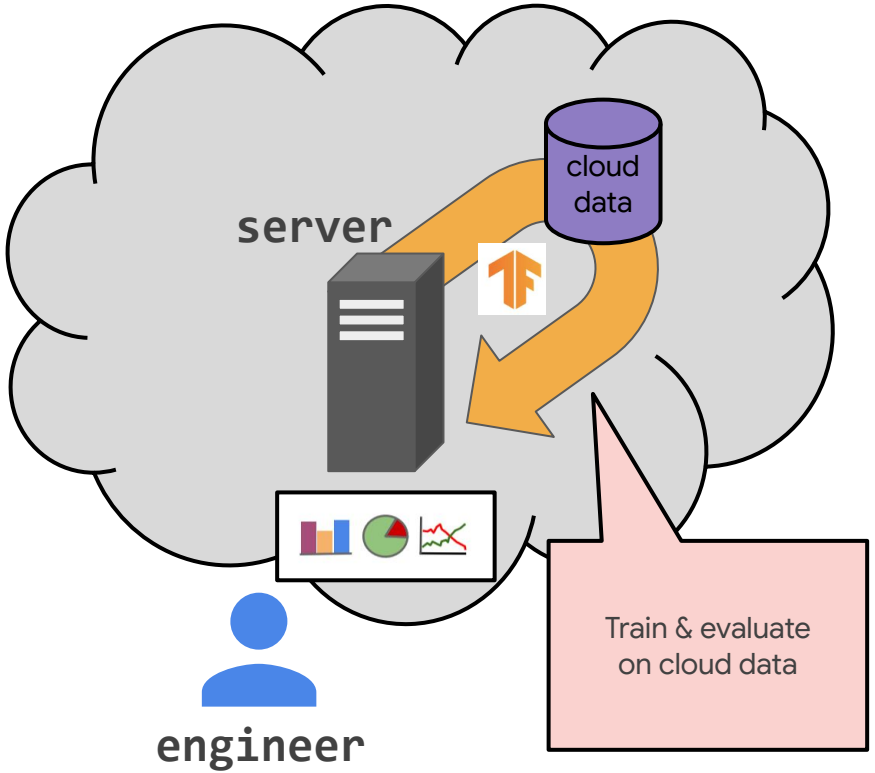
Characteristics of FL vs decentralized learning

	Federated learning	**Fully decentralized (peer-to-peer) learning**
Orchestration	A central orchestration server/service organizes the training, but never sees raw data.	No centralized orchestration.
Wide-area communication pattern	Typically hub-and-spoke topology, with the hub representing a coordinating service provider (typically without data) and the spokes connecting to clients.	Peer-to-peer topology.

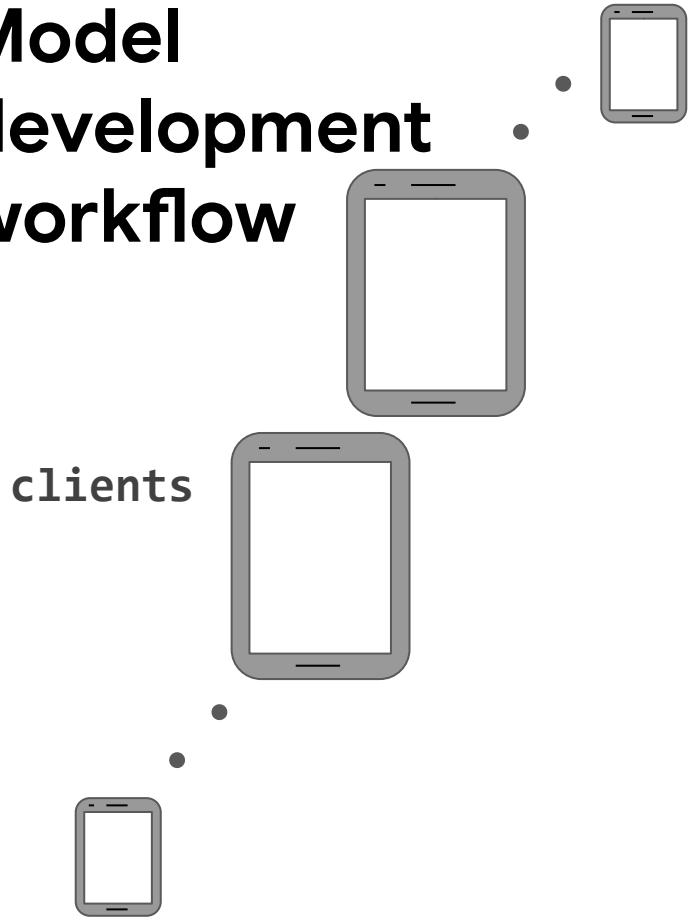
Adapted from Table 3 in *Advances and Open Problems in Federated Learning* (arxiv/1912.04977)

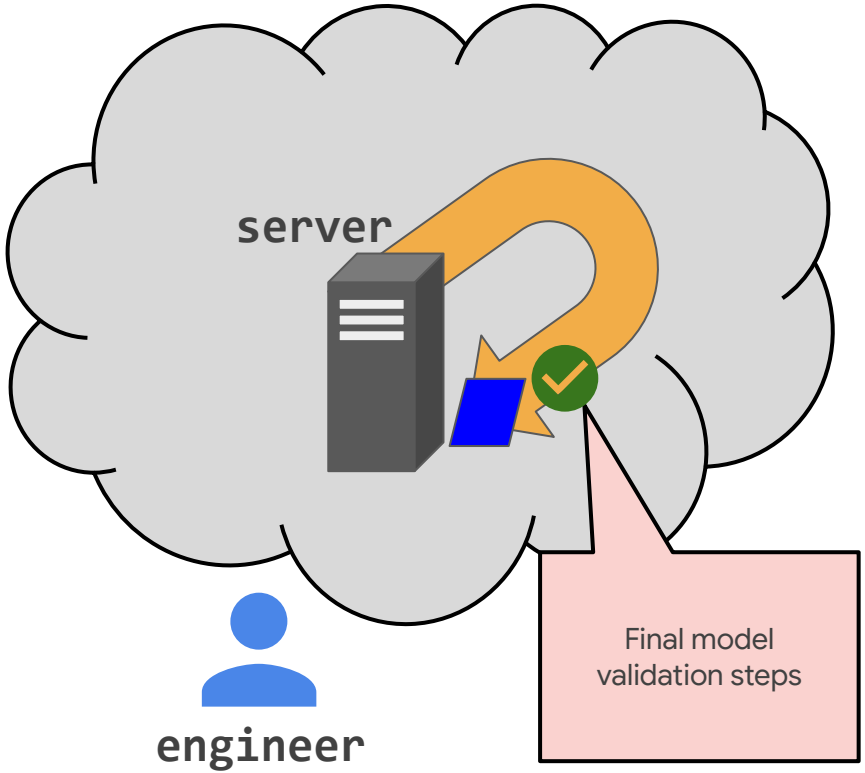
Cross-Device Federated Learning

Model development workflow

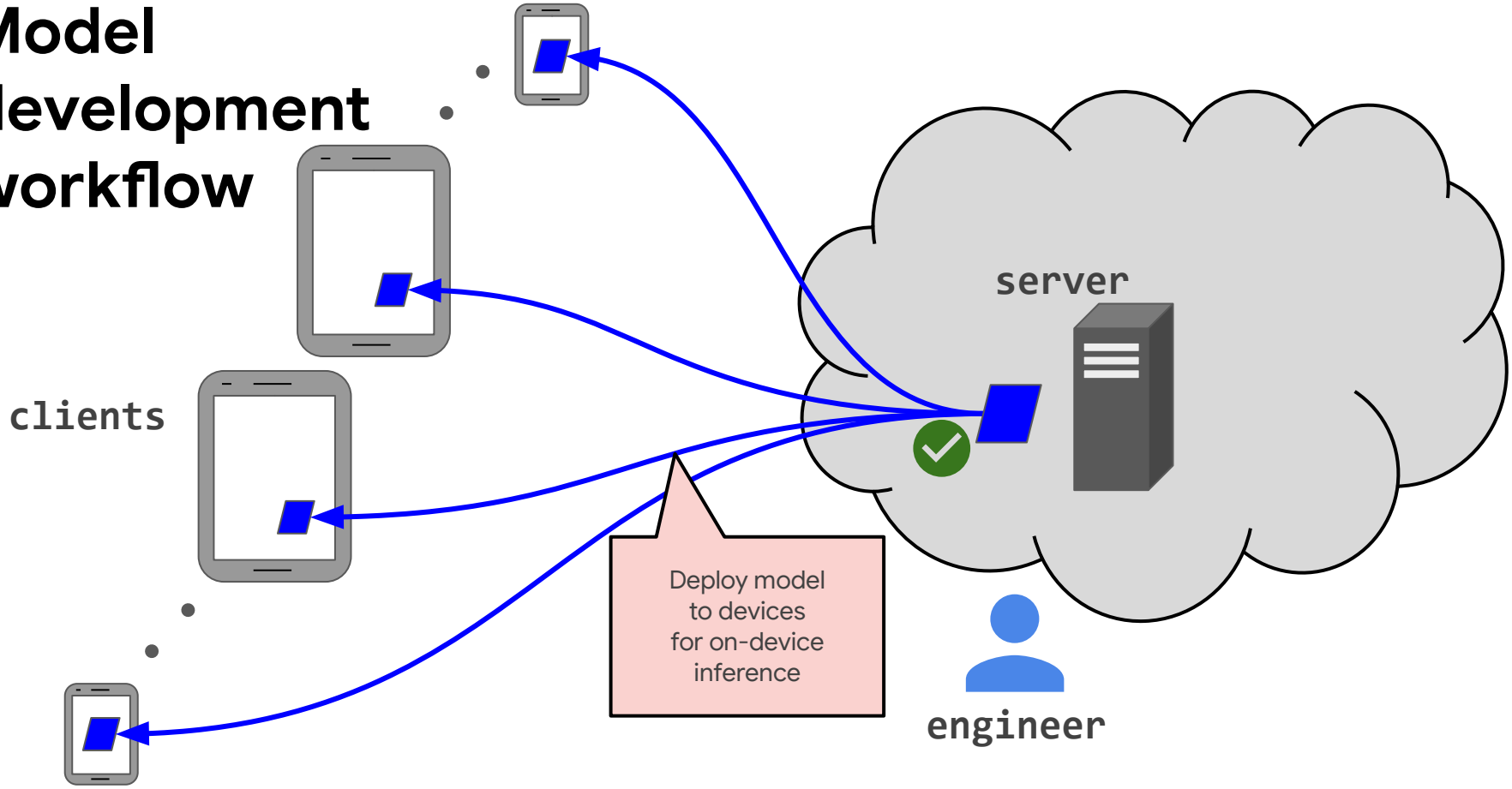


Model development workflow

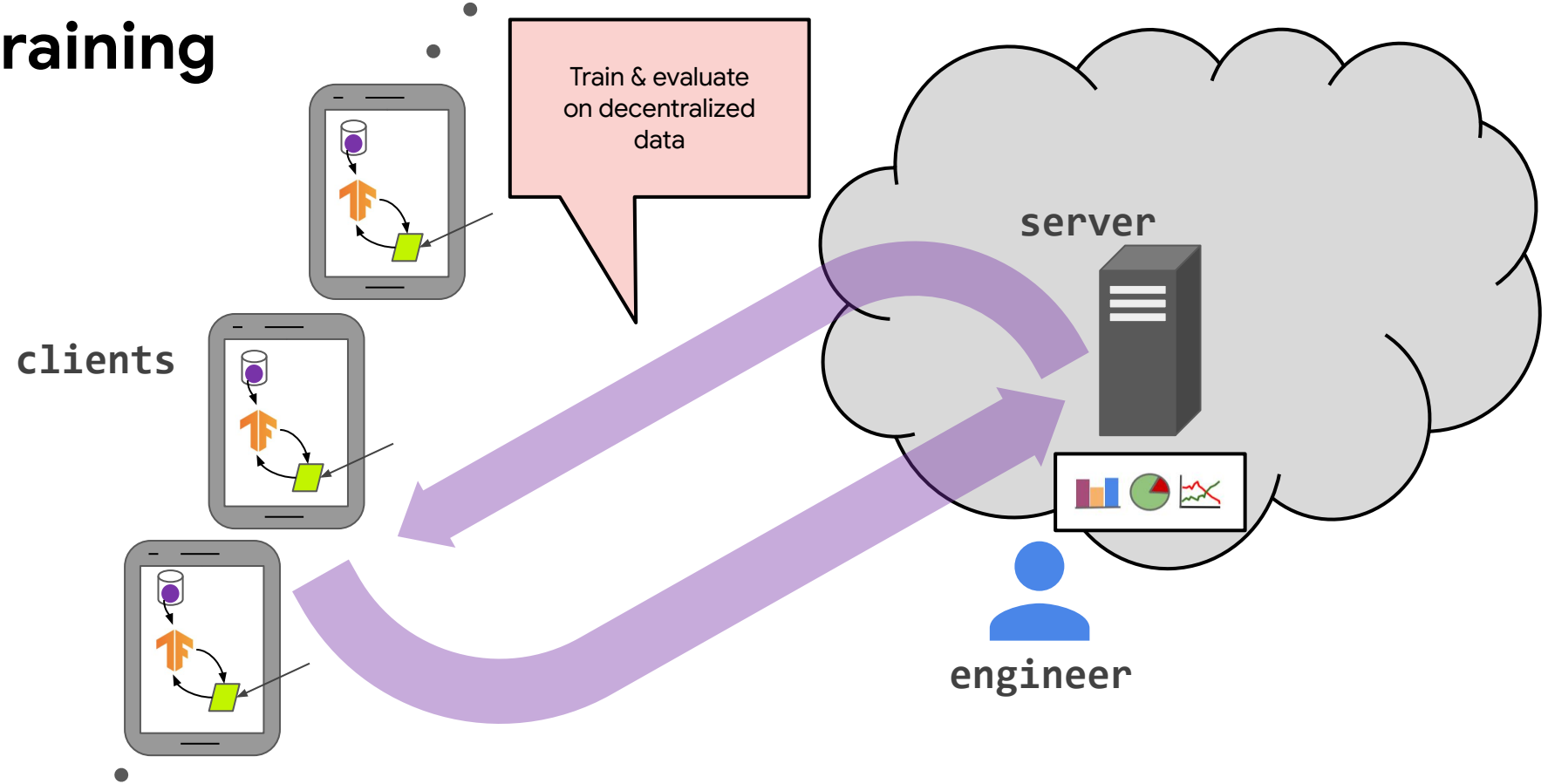




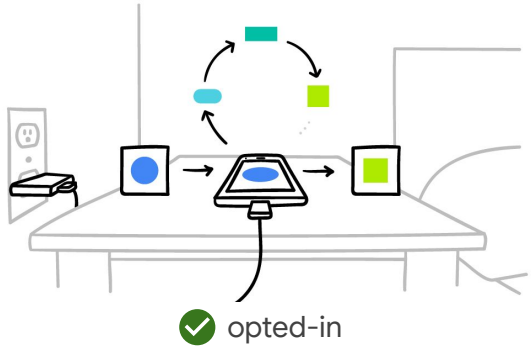
Model development workflow

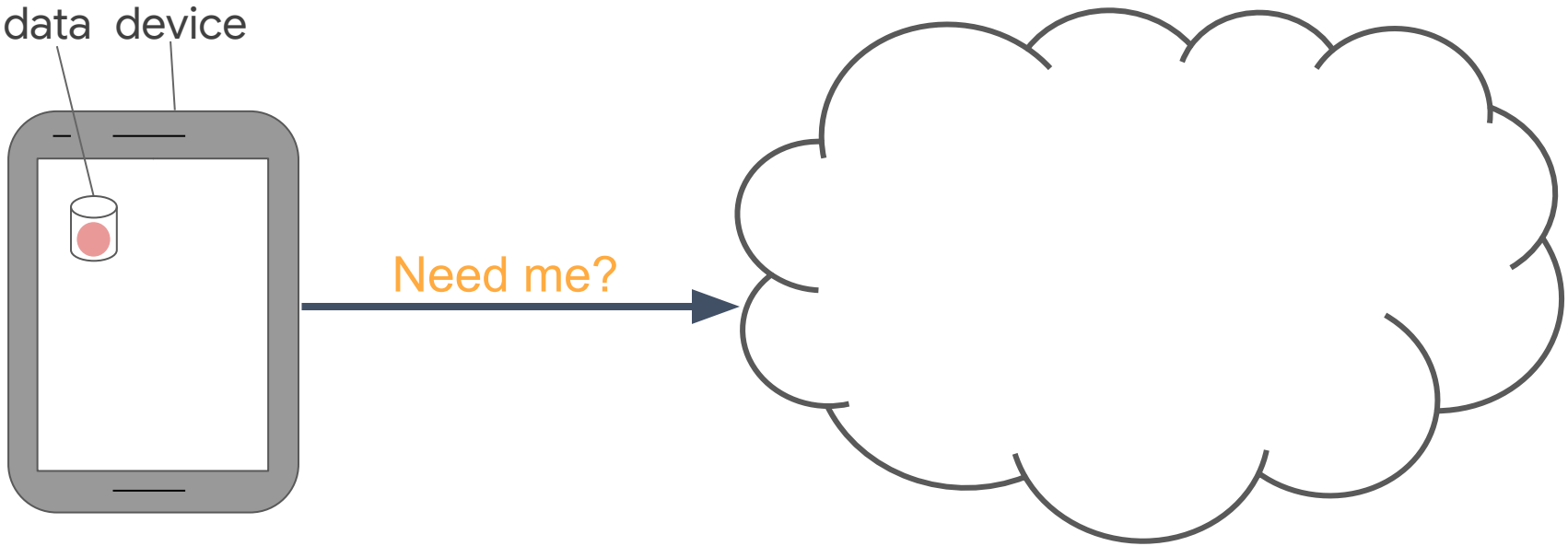


Federated training

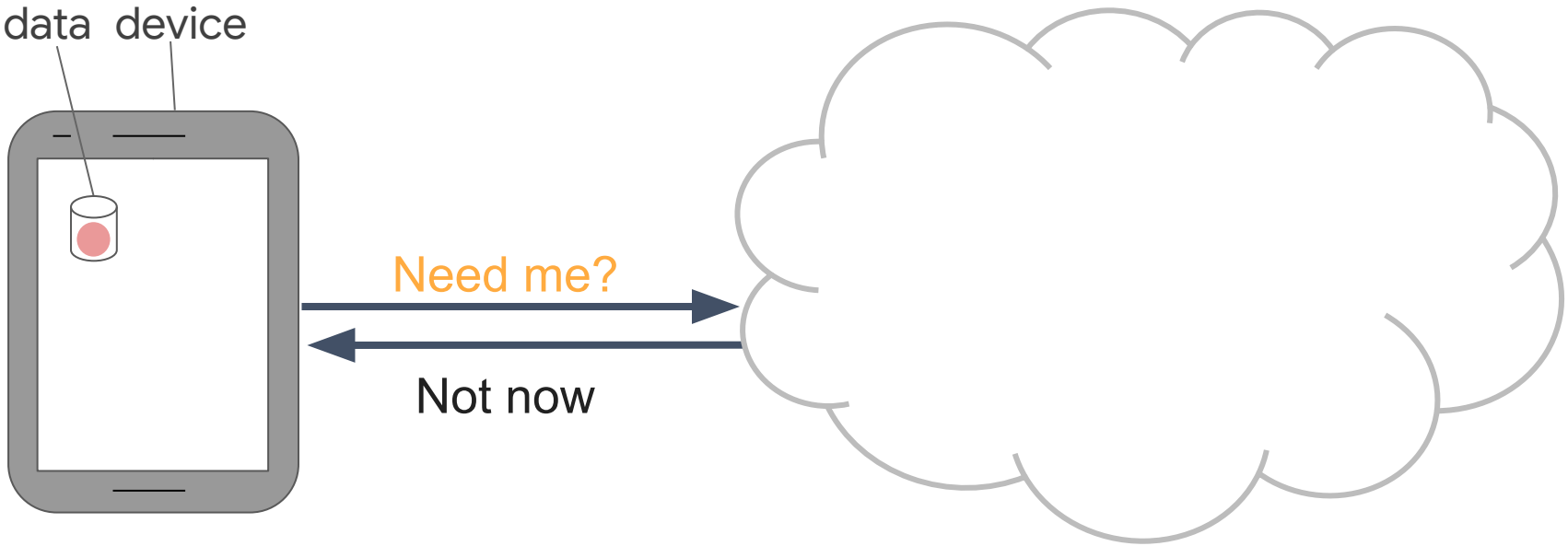


Federated learning

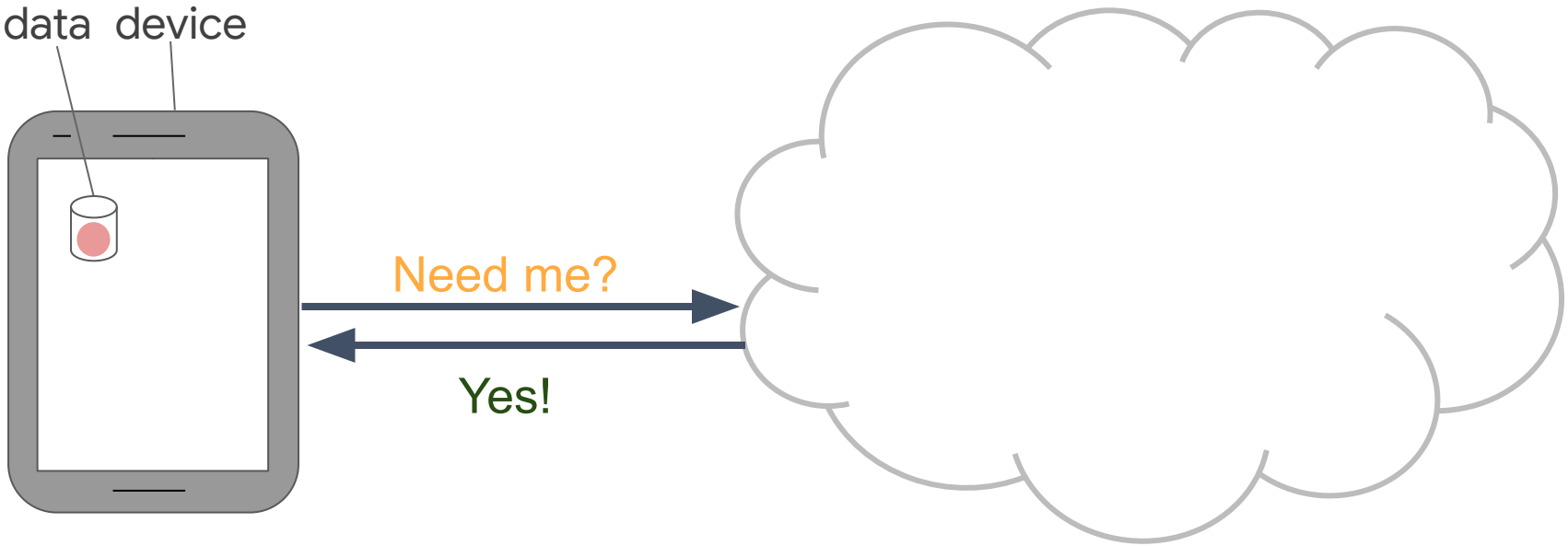




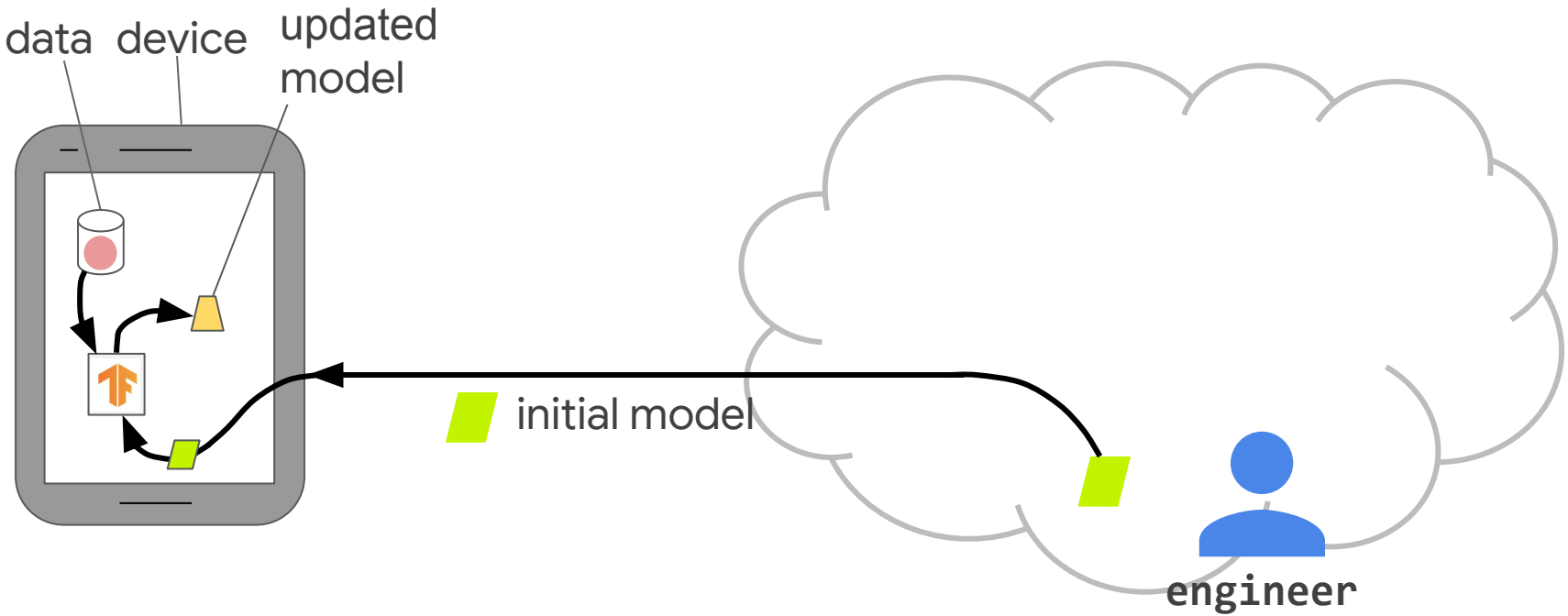
Federated learning



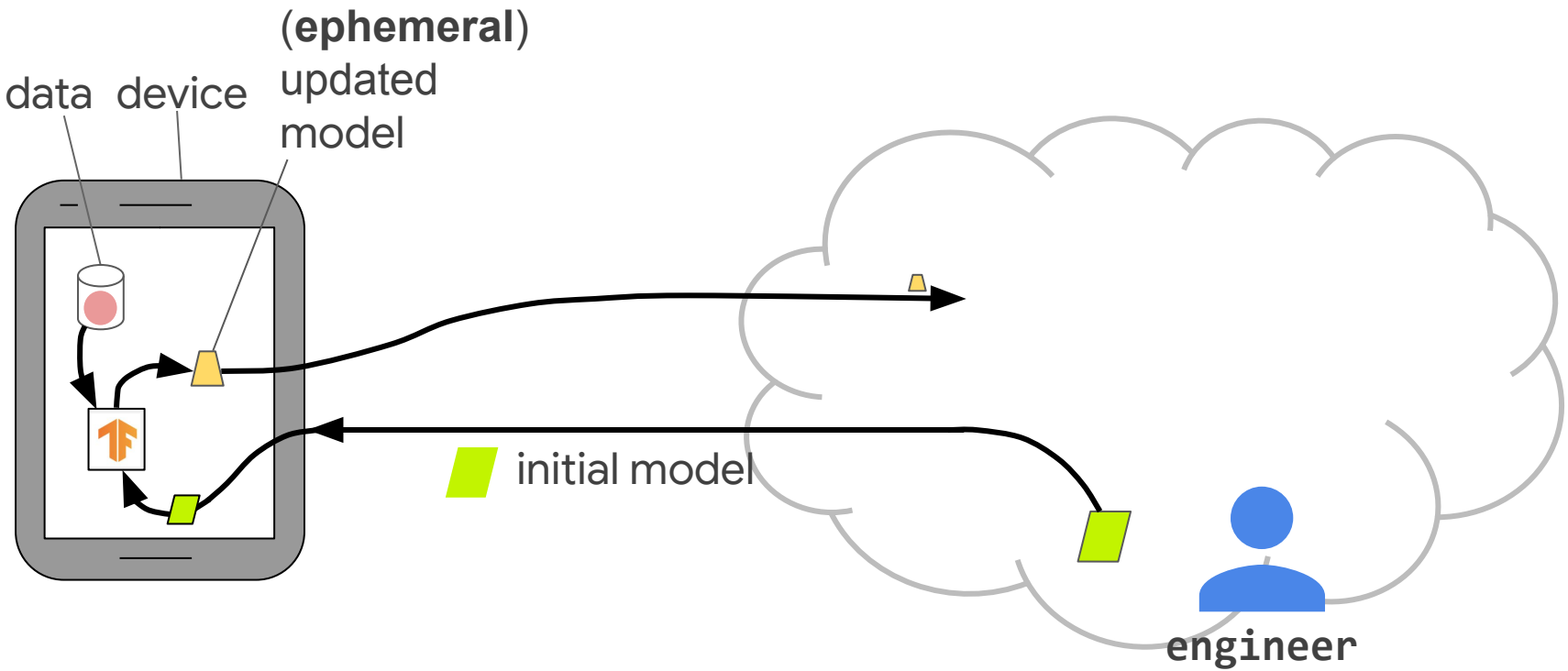
Federated learning



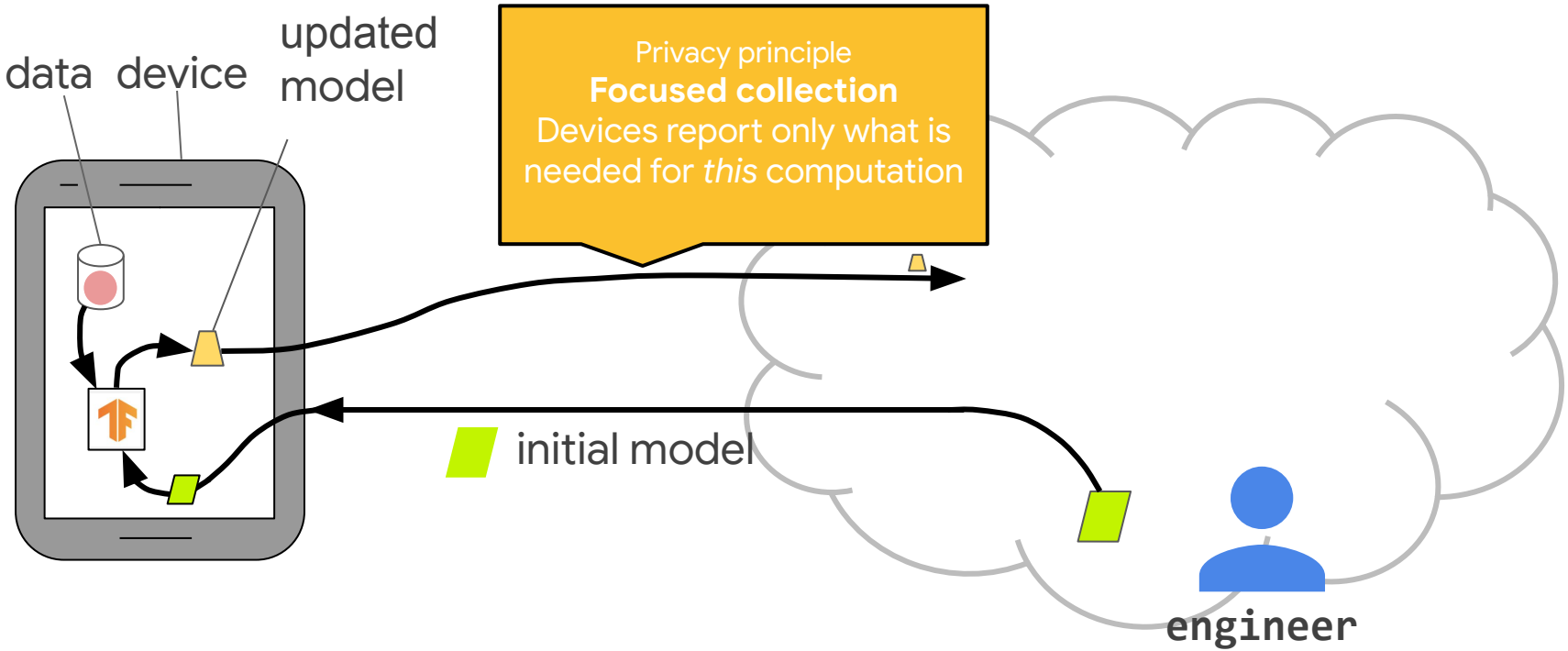
Federated learning



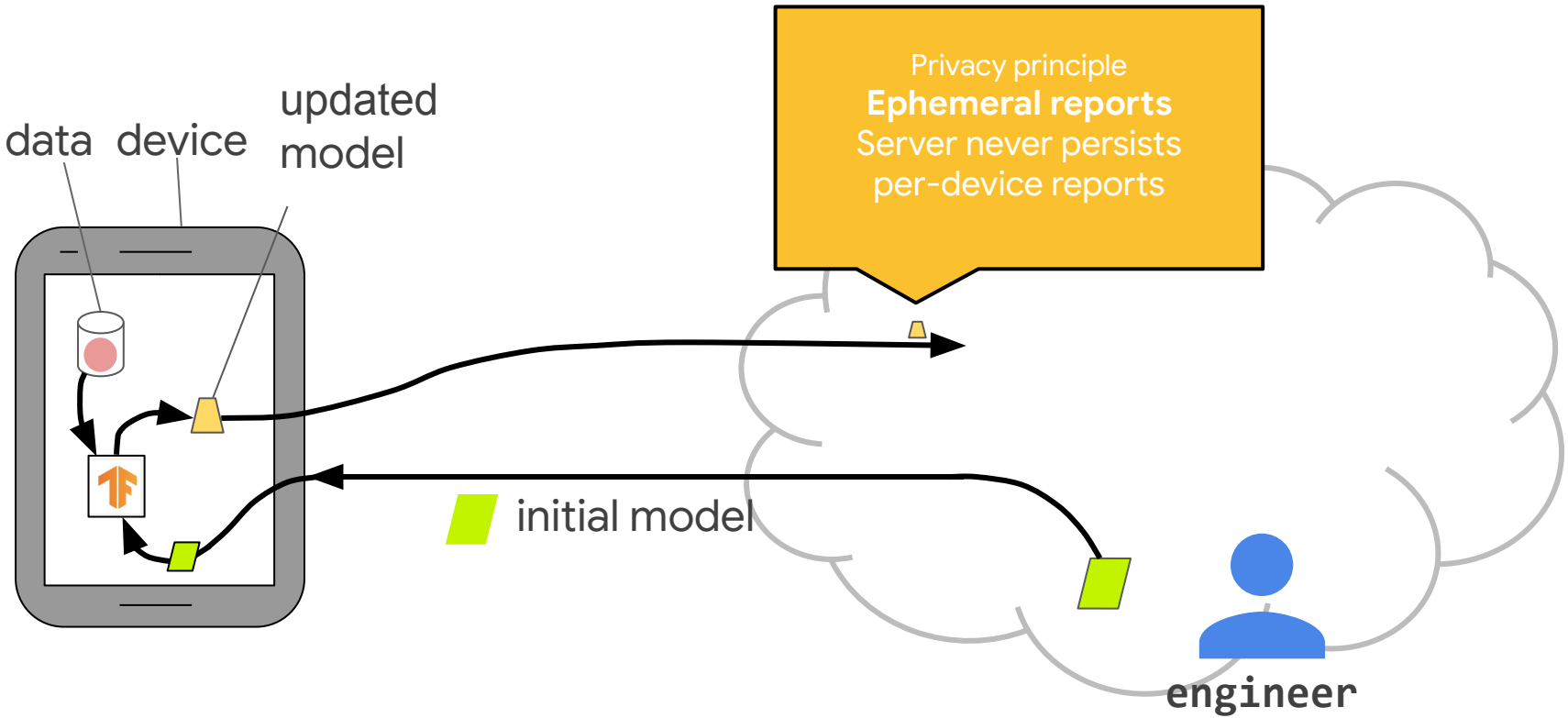
Federated learning



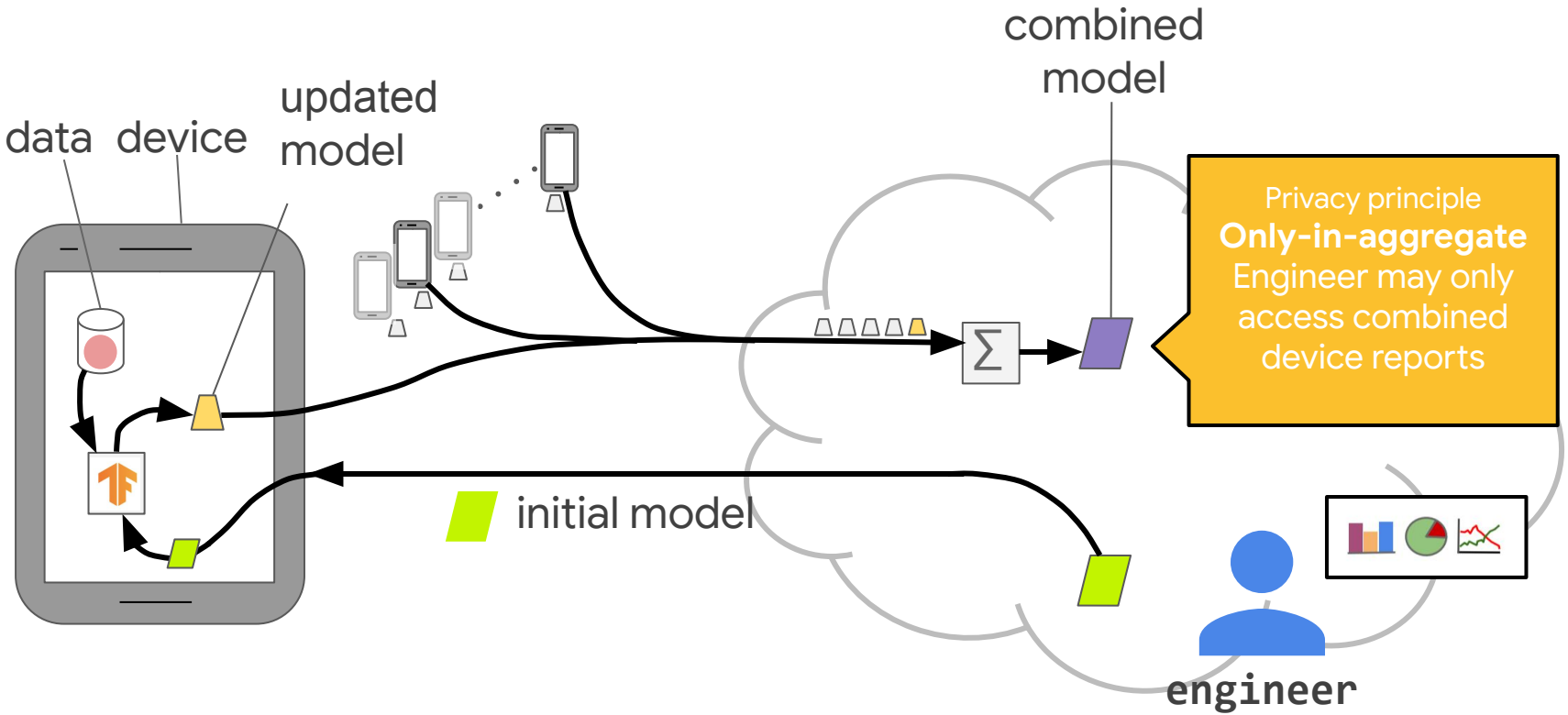
Federated learning



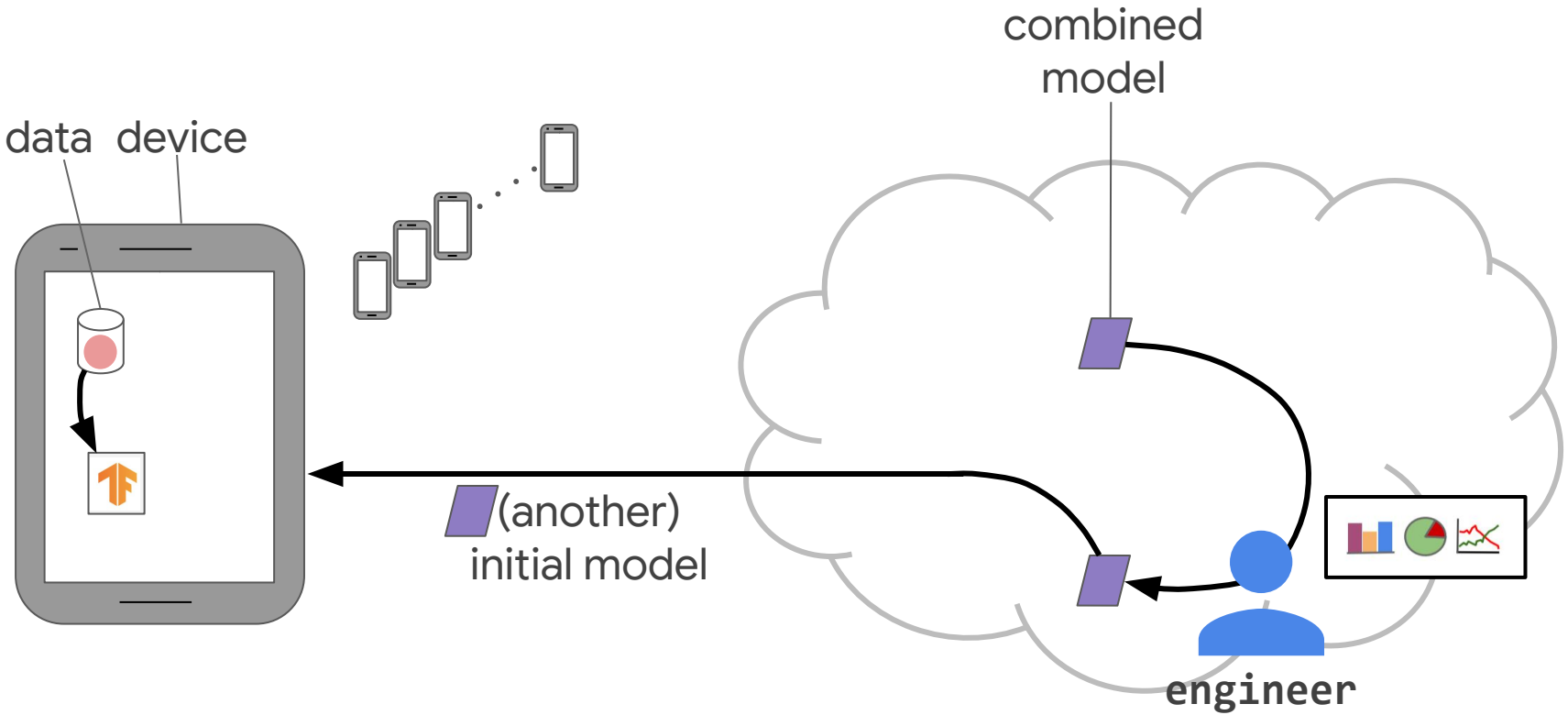
Federated learning



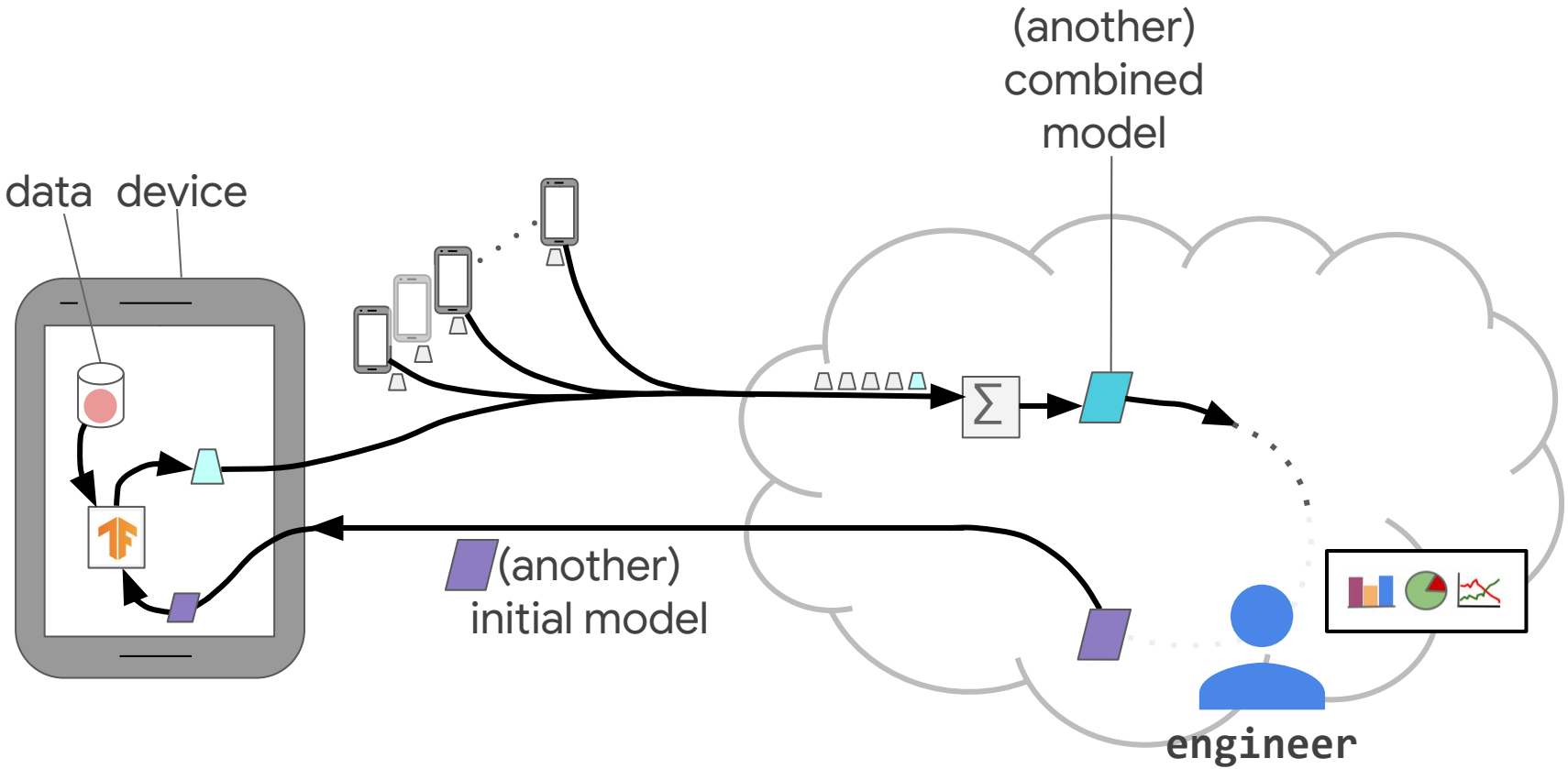
Federated learning



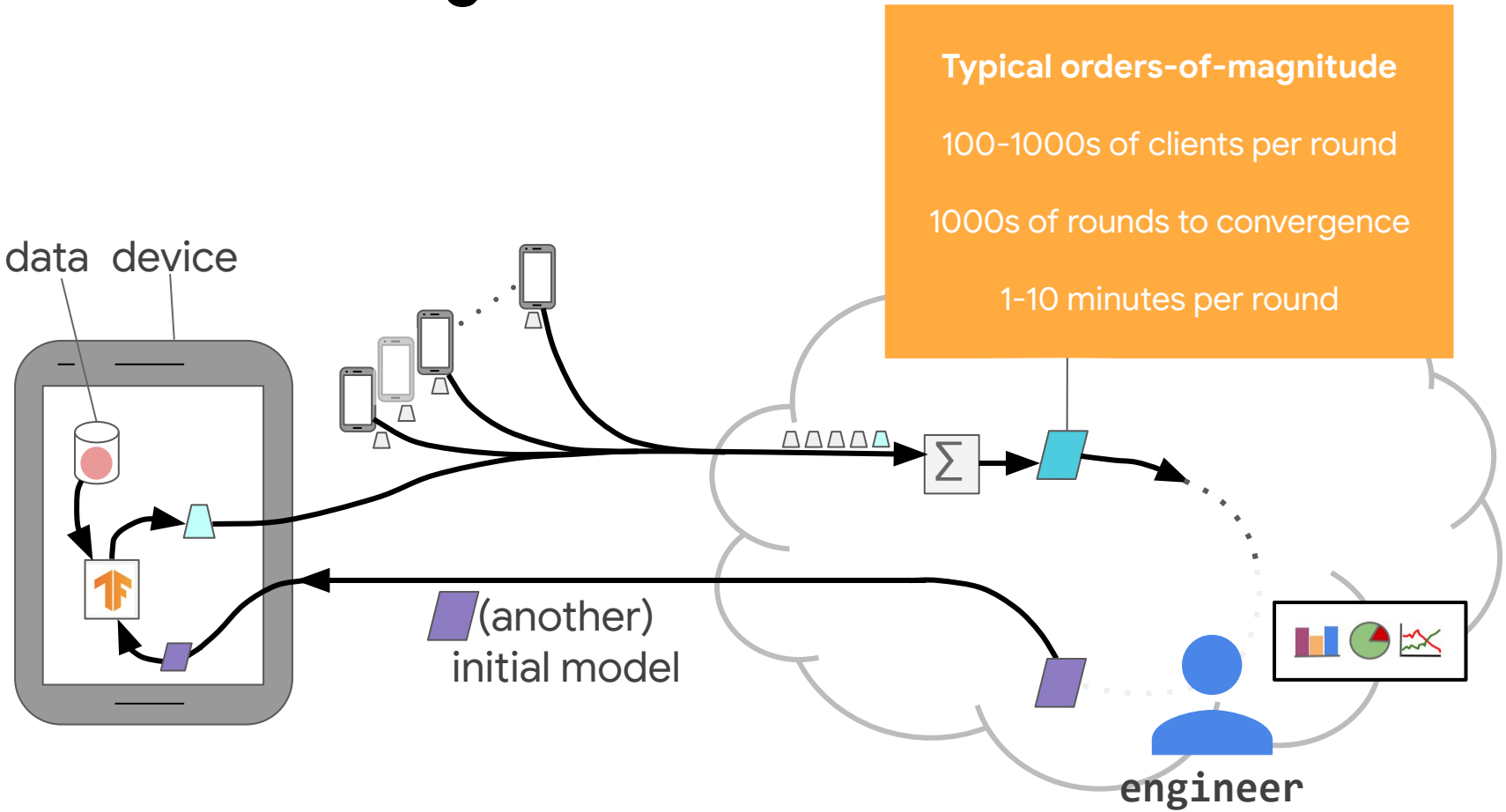
Federated learning



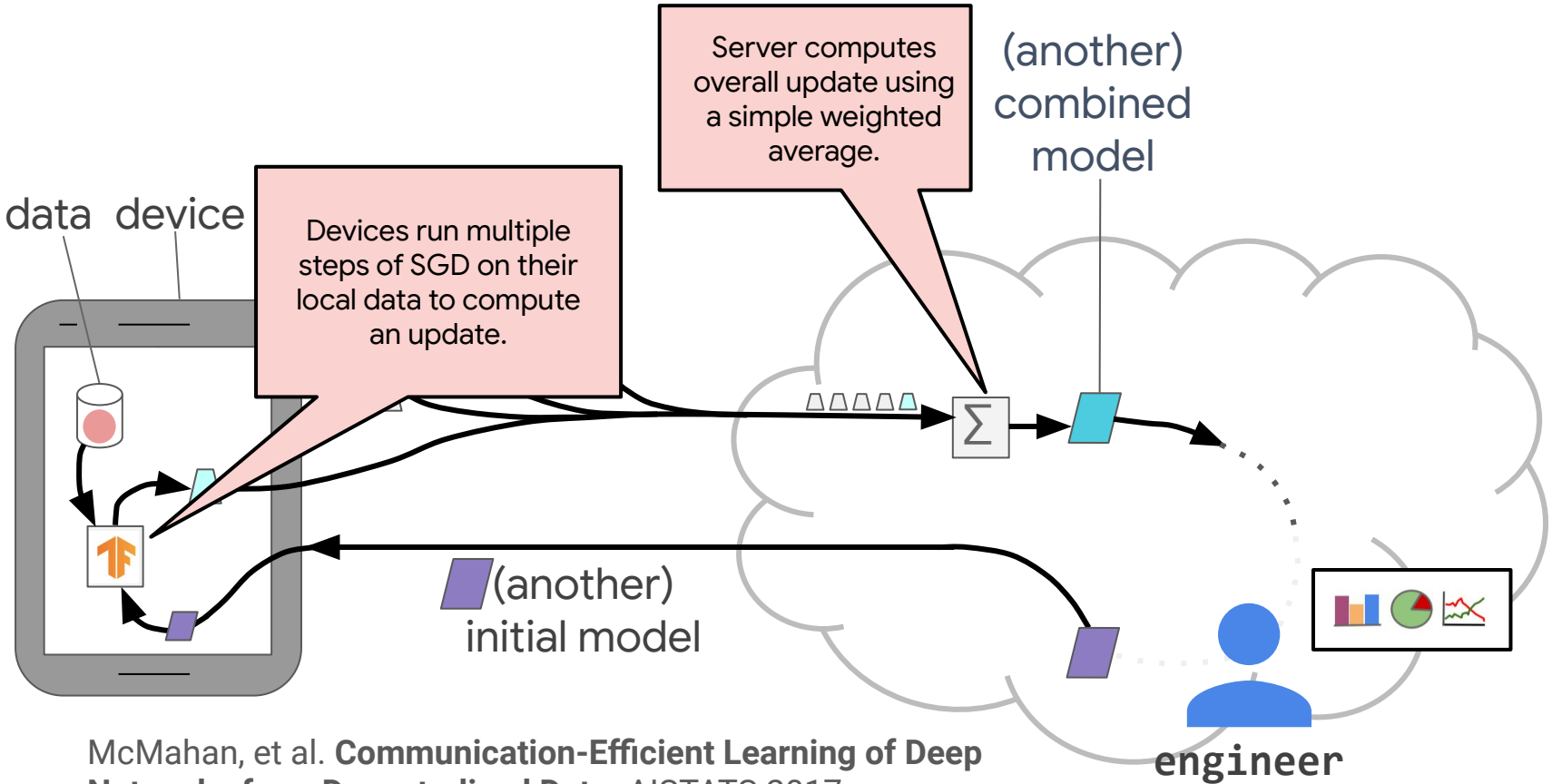
Federated learning



Federated learning



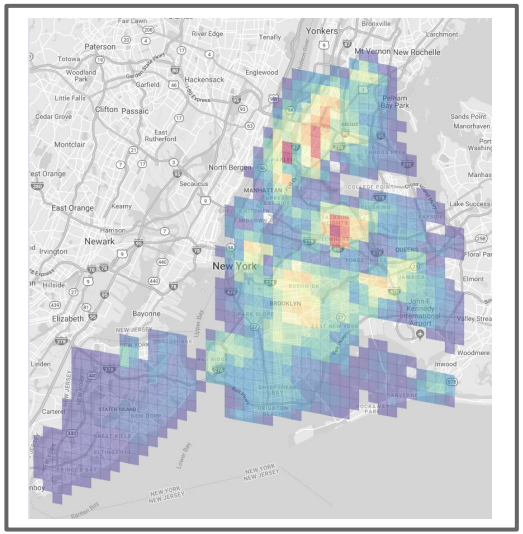
Federated Averaging (FedAvg) algorithm



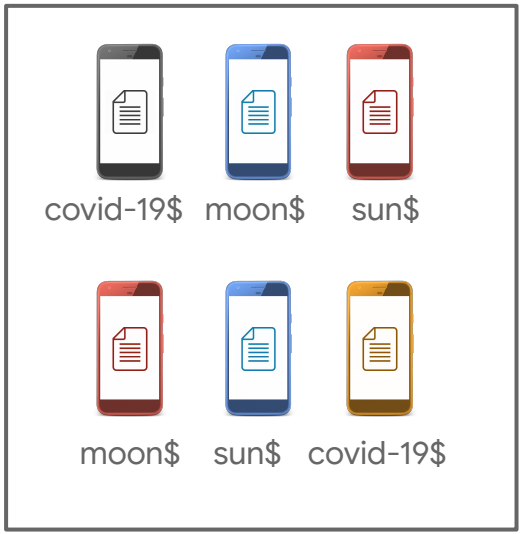
McMahan, et al. **Communication-Efficient Learning of Deep Networks from Decentralized Data.** AISTATS 2017.

Beyond Learning: Federated Analytics

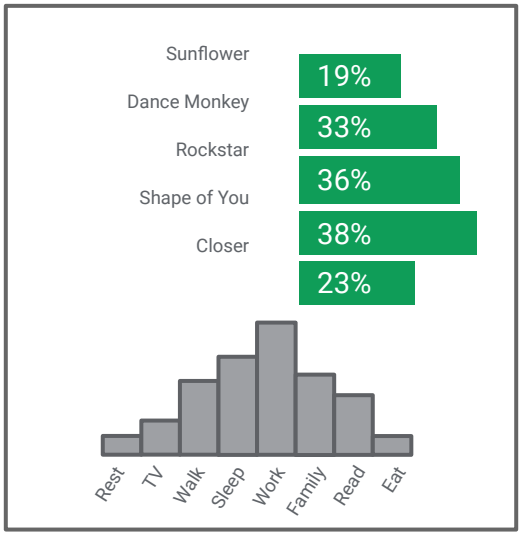
Beyond learning: federated analytics



Geo-location heatmaps



Frequently typed
out-of-dictionary words



Popular songs, trends,
and activities

Federated analytics

Federated analytics is the practice of applying data science methods to the analysis of raw data that is stored locally on users' devices. Like federated learning, it works by running local computations over each device's data, and only making the aggregated results — and never any data from a particular device — available to product engineers. Unlike federated learning, however, federated analytics aims to support basic data science needs.

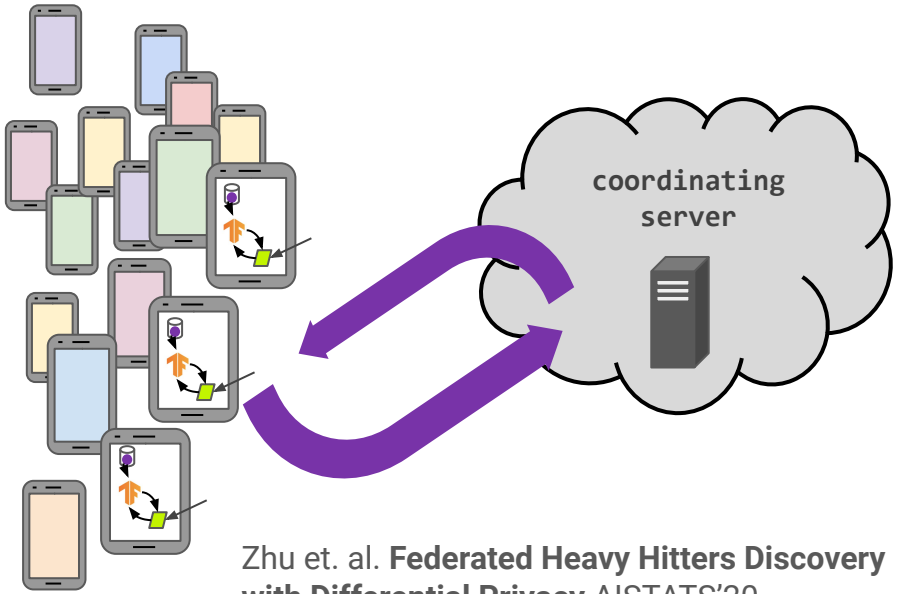
definition proposed in https://ai.googleblog.com/2020/05/federated-analytics-collaborative-data.html

Federated analytics

- Federated histograms over closed sets
- Federated quantiles and distinct element counts
- Federated heavy hitters discovery over open sets
- Federated density of vector spaces
- Federated selection of random data subsets
- Federated SQL
- Federated computations?
- etc...

Interactive algorithms

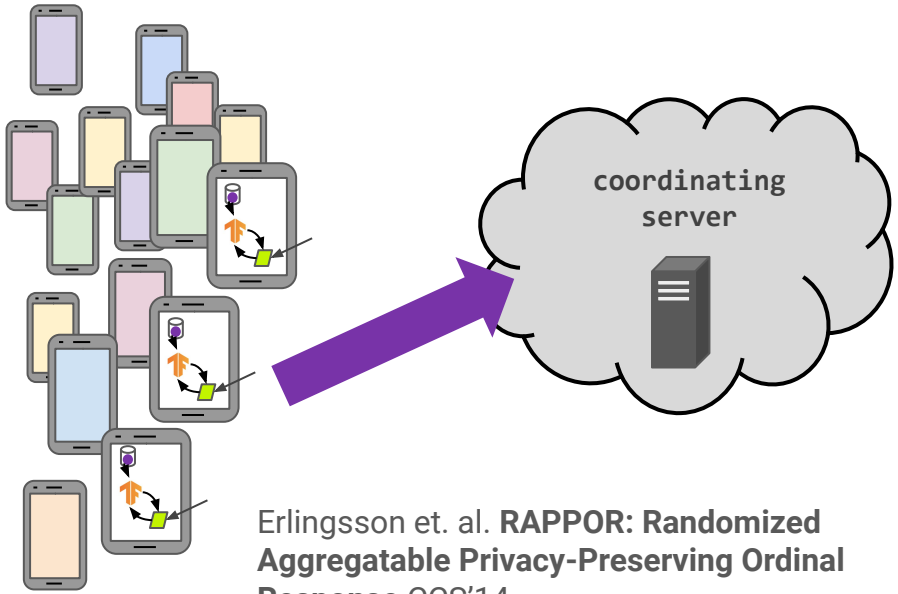
Similar to learning, the on-device computation is a function of a server state



Zhu et. al. **Federated Heavy Hitters Discovery with Differential Privacy** AISTATS'20.

Non-interactive algorithms

Unlike learning, the on-device computation does not depend on a server state



Erlingsson et. al. **RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response** CCS'14.