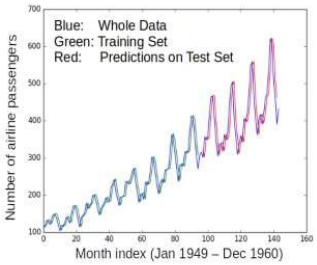
AI & Federated Learning

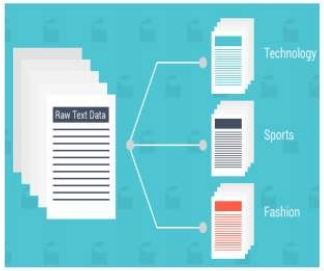
Dr. Gagan Raj Gupta
Associate Professor

Indian Institute Of Technology, Bhilai
Department of Computer Science & Engineering
December 12, 2024

Machine Learning: Introduction

- Algorithms that ingest data and learn a model of the data
- The learned model can be used to
  - Detect patterns/structures/themes/trends etc. in the data
  - Make predictions about future data and make decisions





- Modern ML algorithms are heavily "data-driven"
  - No need to pre-define and hard-code all the rules (infeasible/impossible anyway)
  - The rules are not "static"; can adapt as the ML algo ingests more and more data

ML is like an exam. It's performance is judged on the test, not the train. Generalization is the key!

Components of ML

- **Data:** this is why data is being called the new oil!
  - Data can be collected both manually and automatically
  - Data is not always images; it could be tables of data with many variables (features), text, sensor recordings, sound samples etc., depending on the problem at hand

Features: Features are often also called variables or parameters

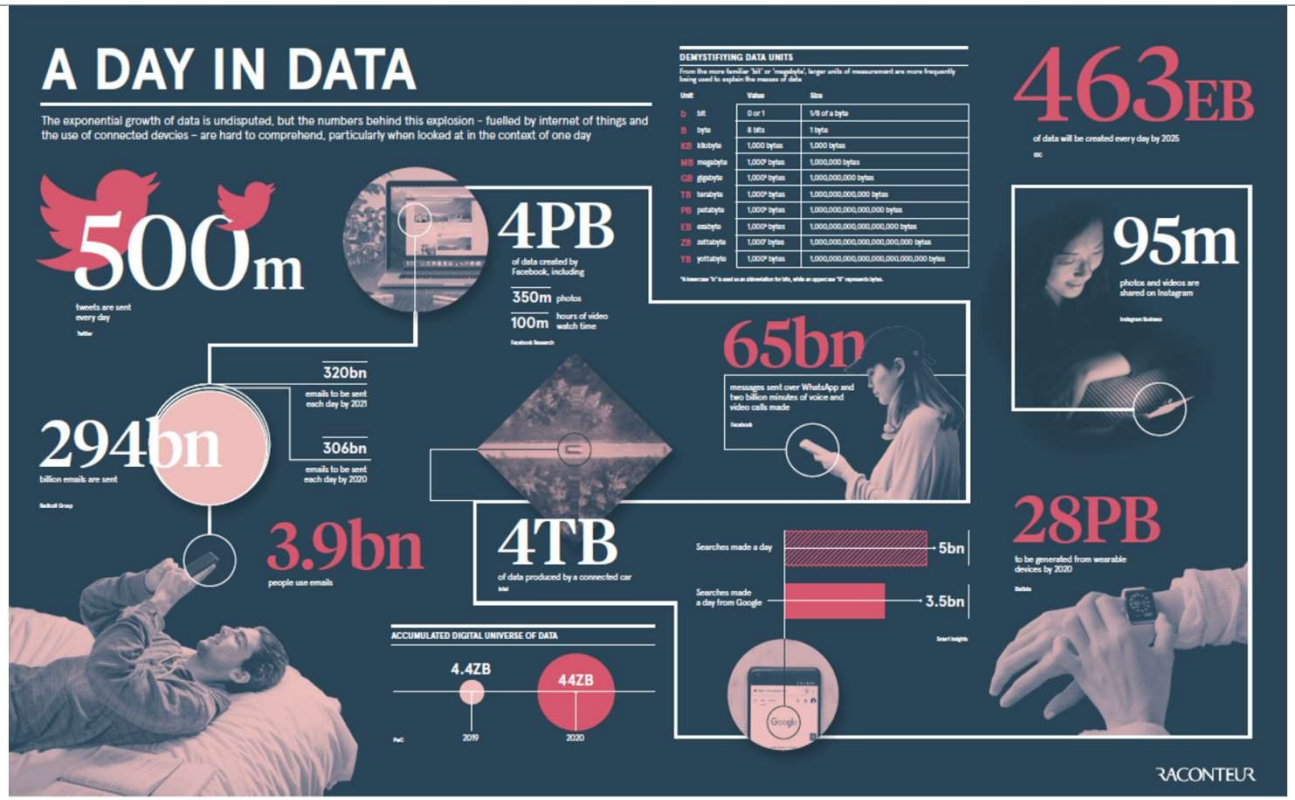
- Important properties of the "object" in question, e.g., users' age, stock price, area of rental property, number of words in a sentence, size of the cells
- To distinguish apples from oranges, say we take bad features like ripeness and seed count.
  - Since these are not really distinct properties of the fruits, our machine learning system won't be able to do a good job

Algorithms: For example, classification allows us to put data into different groups

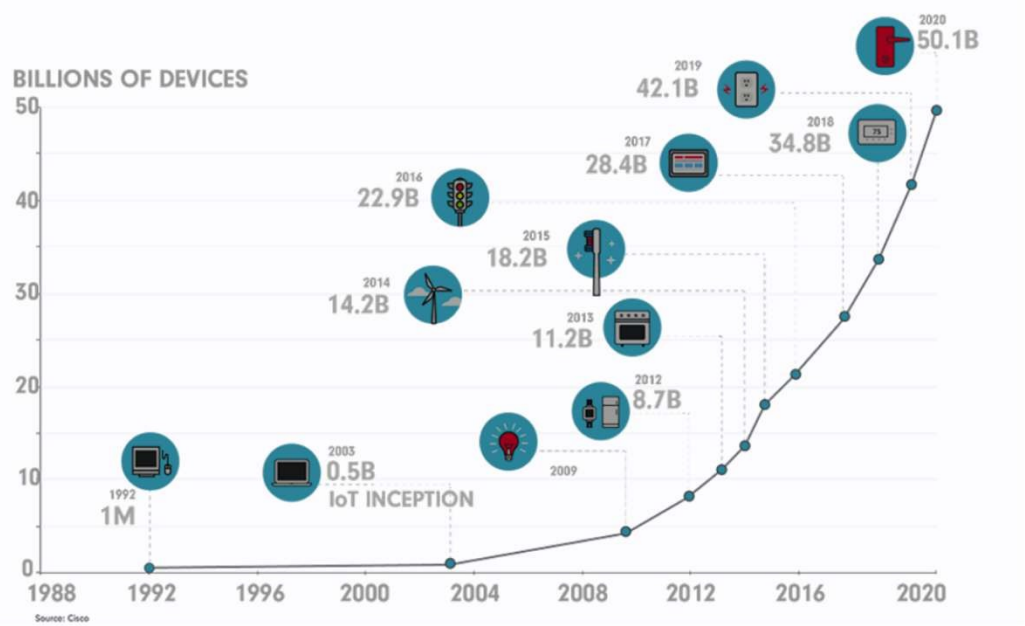
- Same classification algorithm used to recognize handwritten numbers could also be used to classify emails into spam and not-spam without changing a line of code!
- The same algorithm comes up with different classification logic for a different dataset

Garbage in, garbage out: If the data is crappy, even the best algorithm won't help!

Where is data coming from?



IOT


BILLIONS OF DEVICES
50
40
30
20
10
0
1988
1992
1996
2000
2004
2008
2012
2016
2020
1992
1M
2003
0.5B
IoT INCEPTION
2009
2012
8.7B
2013
11.2B
2014
14.2B
2015
18.2B
2016
22.9B
2017
28.4B
2018
34.8B
2019
42.1B
2020
50.1B
Source: Cisco

Why do we need 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?

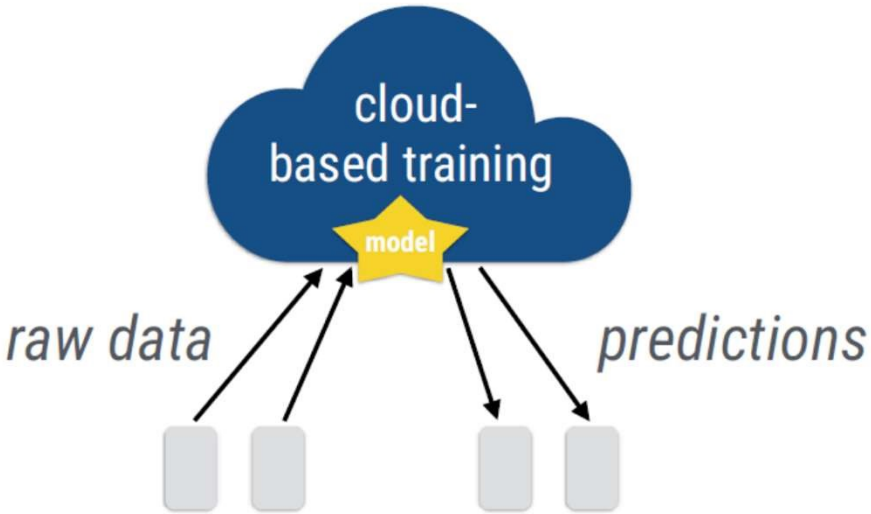
Data processing is moving on device:

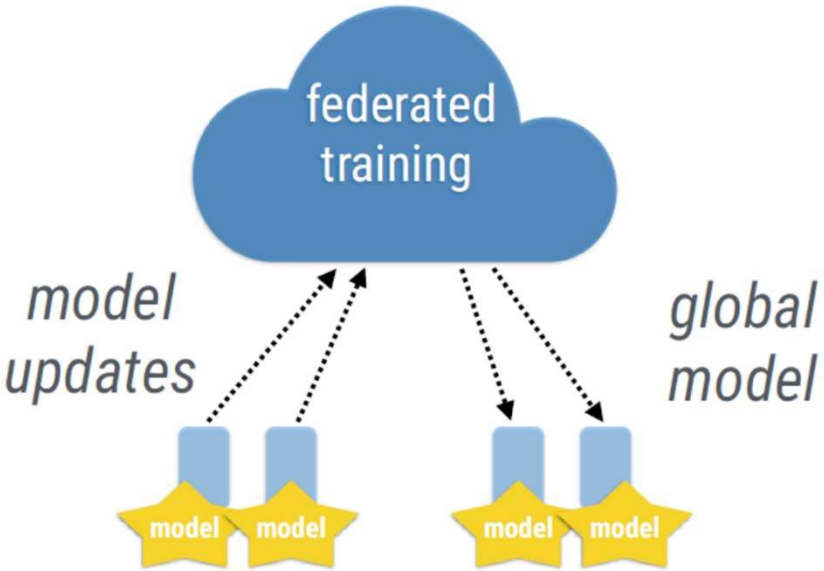
- Improved latency
- Works offline
- Better battery life
- Privacy advantages

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

Federated Learning

training machine learning models at the edge





why? ✓ *quickly incorporate new data* ✓ *reduce strain on network* ✓ *privacy*

Applications

example applications

language modeling for voice recognition on mobile phones

adapting to pedestrian behavior in autonomous vehicles

predicting low blood sugar via wearable devices

assumptions? ✓*local data is important* ✓*labels available* ✓*privacy is a concern*

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)

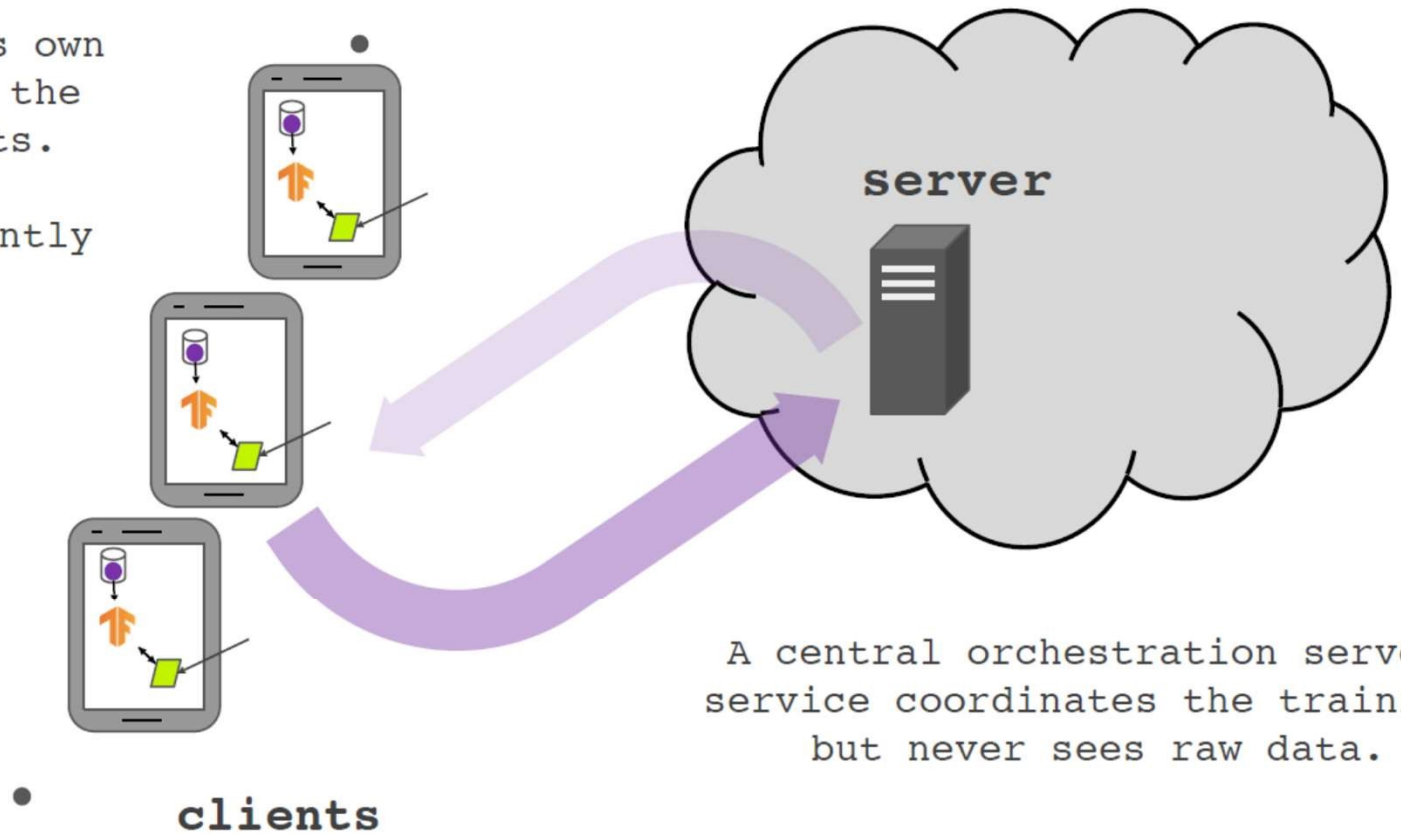
[Credit: P. Kairouz, B. McMahan, V. Smith, FL Tutorial, NeurIPS 2020]

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.

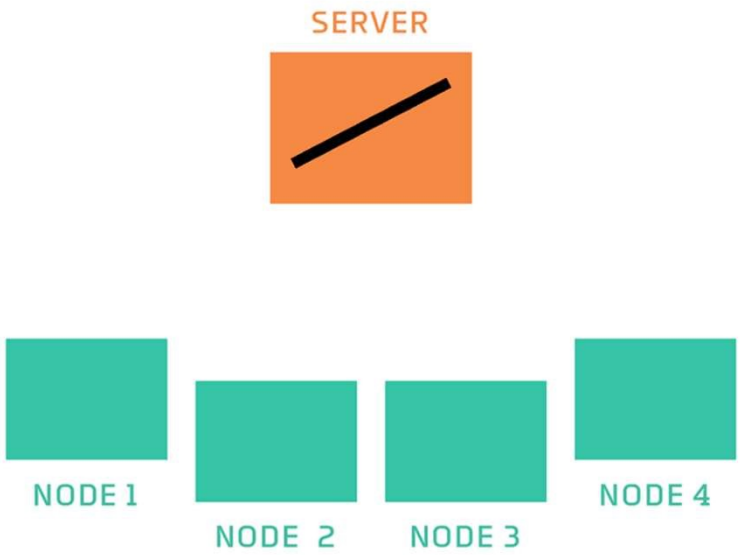
Data is not independently or identically distributed.



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

[Credit: P. Kairouz, B. McMahan, V. Smith, FL Tutorial, NeurIPS 2020]

What is federated learning?



1. Server trains and evaluates model on cloud data
2. Deploy models to devices for on-device inference
3. Re-train and evaluate on decentralized data
4. Securely aggregate multiple client-side models
5. Re-deploy to next set of clients for further enhancement
6. Convergence!

http://vision.cloudera.com/wp-content/uploads/2018/11/2018-10-31-181344-federated_learning_animated_labeled.gif

Characteristics of Federated Learning

Data locality and distribution: massively decentralized, naturally arising (non-IID) partition

Data availability: limited availability, time-of-day variations

Addressability: data nodes are anonymous and interchangeable

Node statefulness: stateless (generally no repeat computation)

Node reliability: unreliable (~10% failures)

Wide-area communication pattern: hub-and-spoke topology

Distribution scale: massively parallel (1e9 data nodes)

Primary bottleneck: communication

Federated Learning Examples

STORIES ›

Under the hood of the Pixel 2: How AI is supercharging hardware

We all know the feeling: there's an amazing song on the radio, and you're frantic to make sure you can find it when you get home. In the past, you might have written down a few hasty lyrics to look it up later. But today, smarter technology on mobile phones makes it easier than ever to find the information you need, right when you need it.

Personalization in Pixel Phones

Gboard

- Typical Orders of magnitude
  - 100-1000s of users per round
  - 100-1000s of rounds to convergence
  - 1-10 minutes per round
  - 0.5B installs
- Daily use by multiple teams, dozens of ML engineers
- Powering features in Gboard and on Pixel Devices

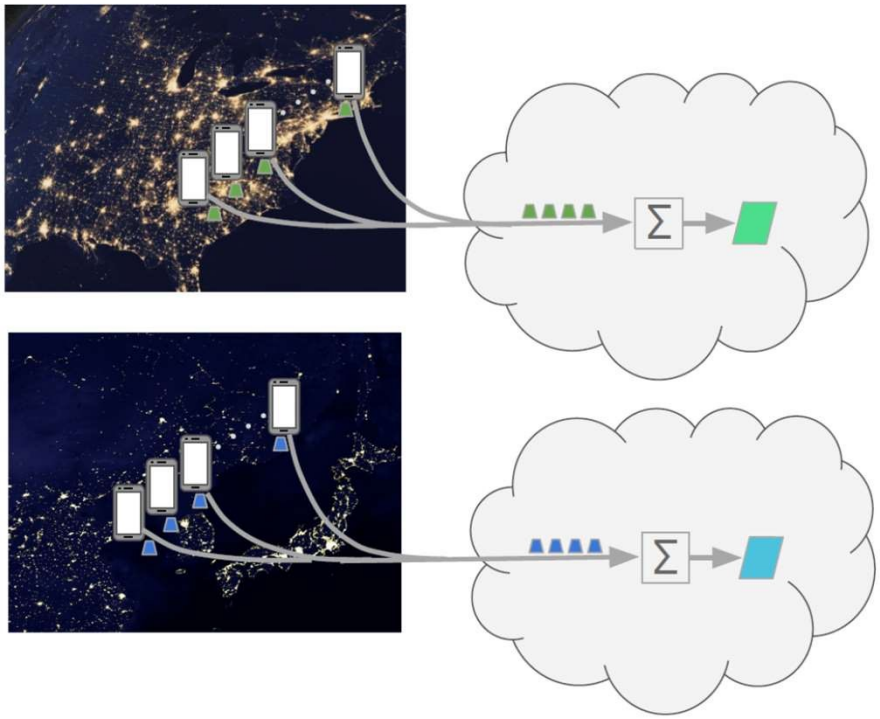
Constraints in the federated setting

Each device reflects one users' data.

No one device is representative of the whole population.

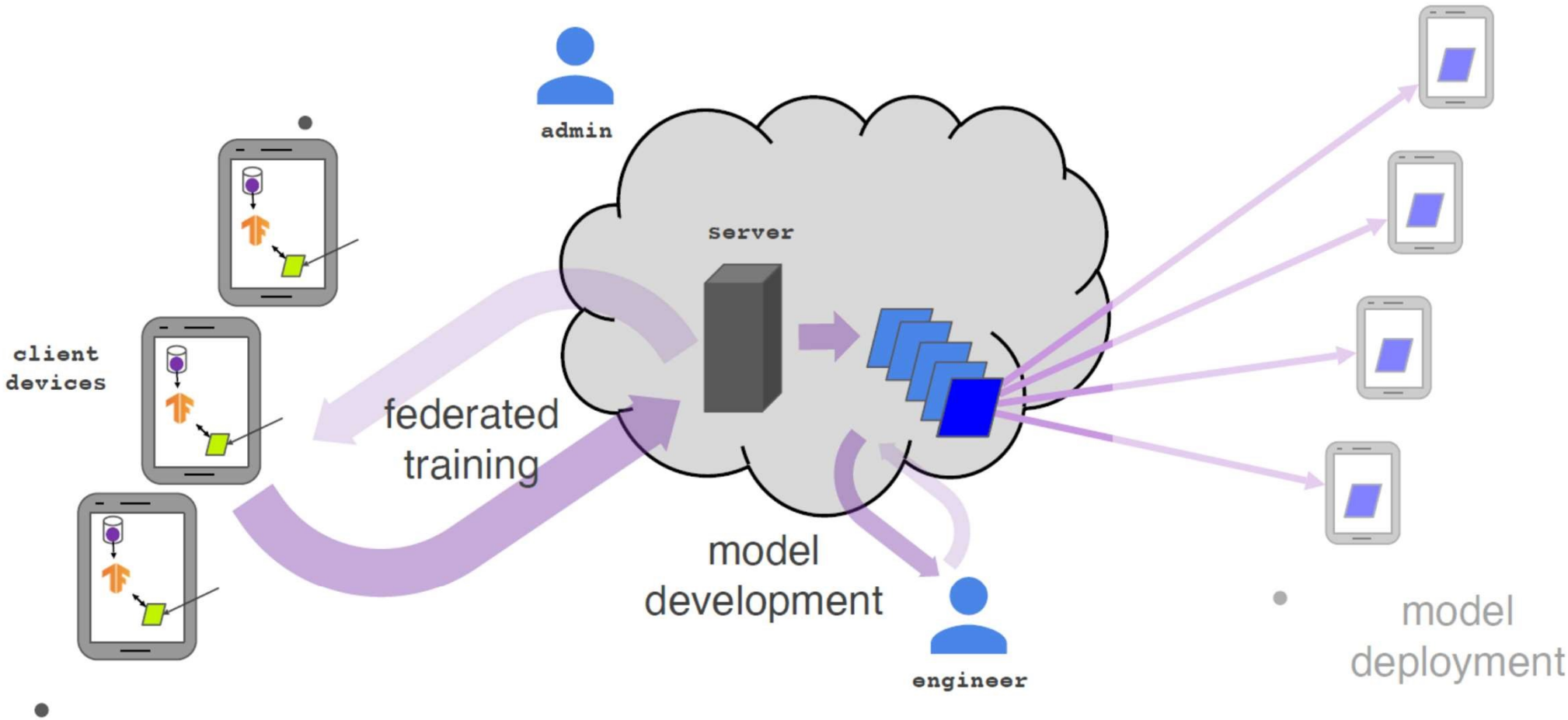
Devices must idle, plugged-in, on wi-fi to participate.

Device availability correlates with both geo location *and* data distribution



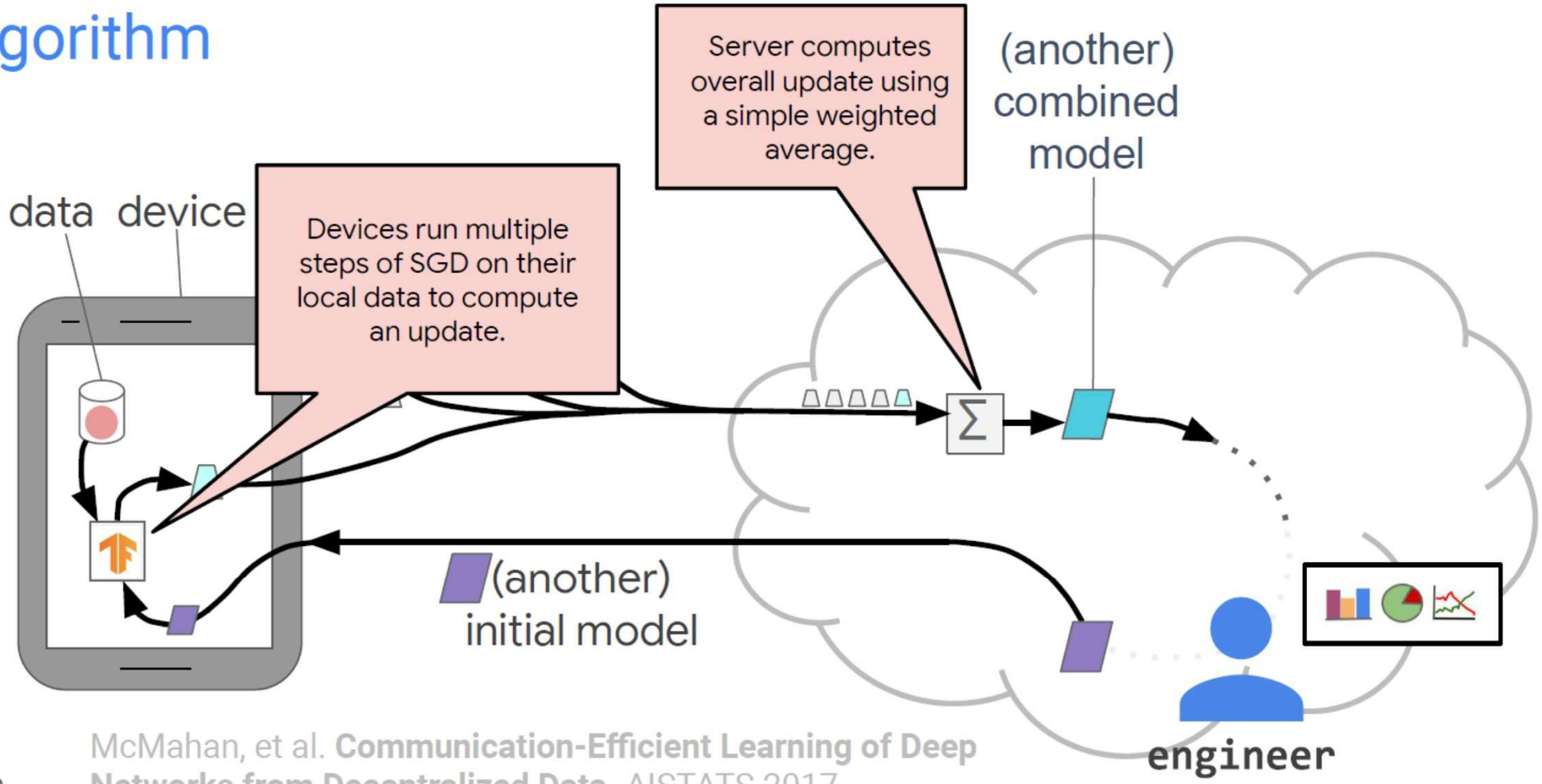
H. Eichner, *et al*. **Semi-Cyclic Stochastic Gradient Descent**. *ICML 2019.*

Cross-device federated learning



[Credit: P. Kairouz, B. McMahan, V. Smith, FL Tutorial, NeurIPS 2020]

Federated Averaging Algorithm



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

The Federated Averaging algorithm

Server

Until Converged:

1. Select a random subset of clients

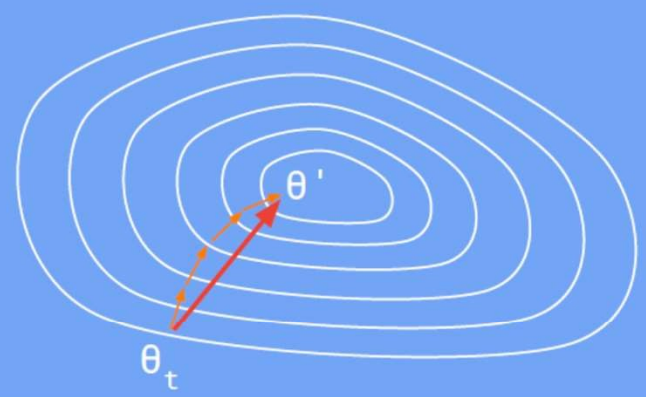
2. In parallel, send current parameters $\boldsymbol{\theta}_t$ to those clients

Selected Client k

1. Receive θ_t from server.

2. Run some number of minibatch SGD steps, producing θ'

3. Return $\theta' - \theta_t$ to server.



3. $\theta_{t+1} = \theta_t$ + data-weighted average of client updates

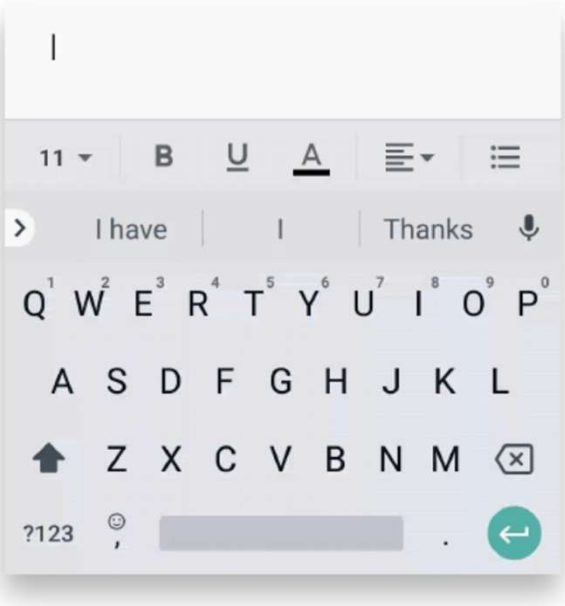
H. B. McMahan, *et al.* **Communication-Efficient Learning of Deep Networks from Decentralized Data.** AISTATS 2017

Gboard: Language models

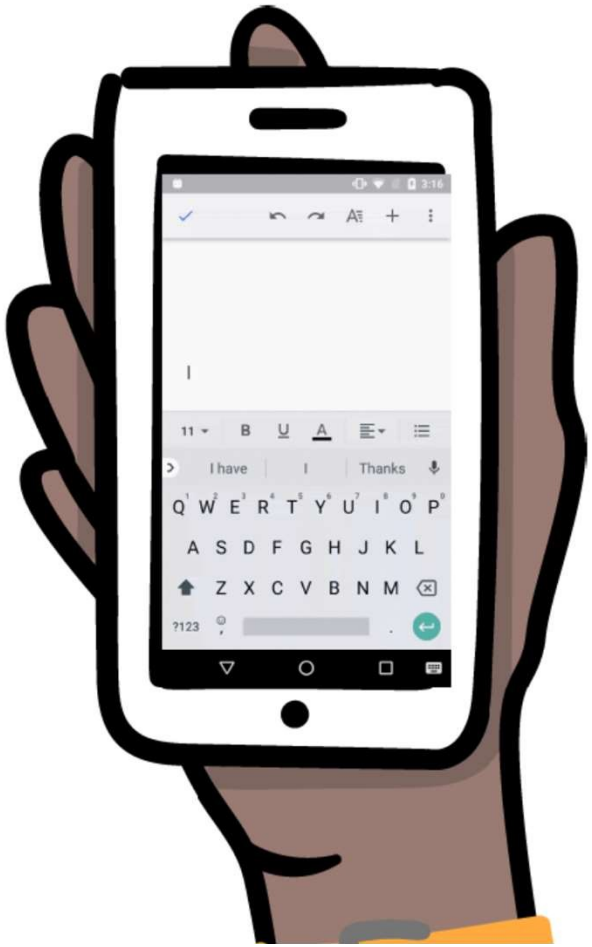
- Predict the next word based on typed text so far
- Powers the predictions strip
- When should you consider federated learning?
- On-device data is more relevant than server-side proxy data
- On-device data is privacy sensitive or large
- Labels can be inferred naturally from user interaction

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

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



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*

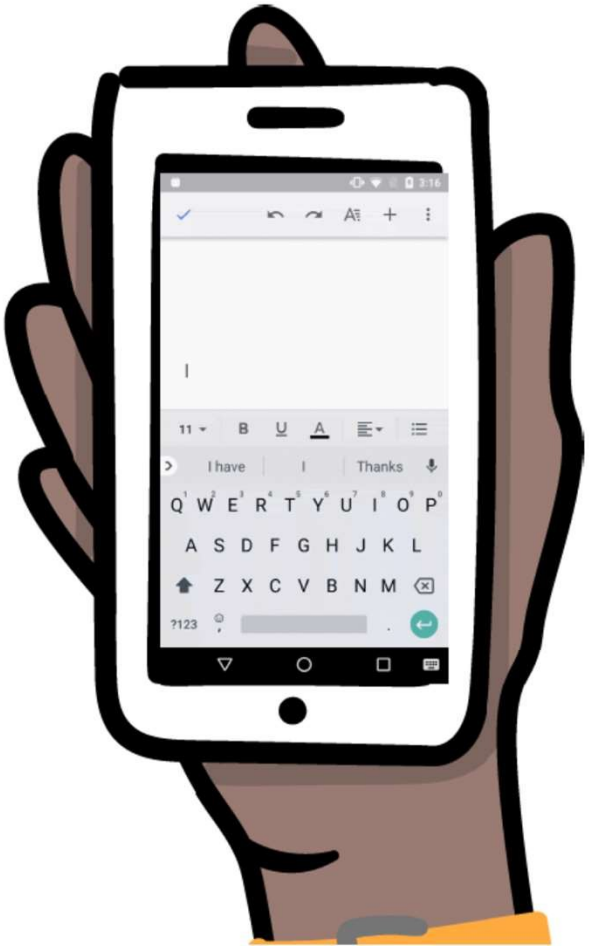
Discovering new words

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*

[Credit: P. Kairouz, B. McMahan, V. Smith, FL Tutorial, NeurIPS 2020]

Gboard: next-word prediction



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

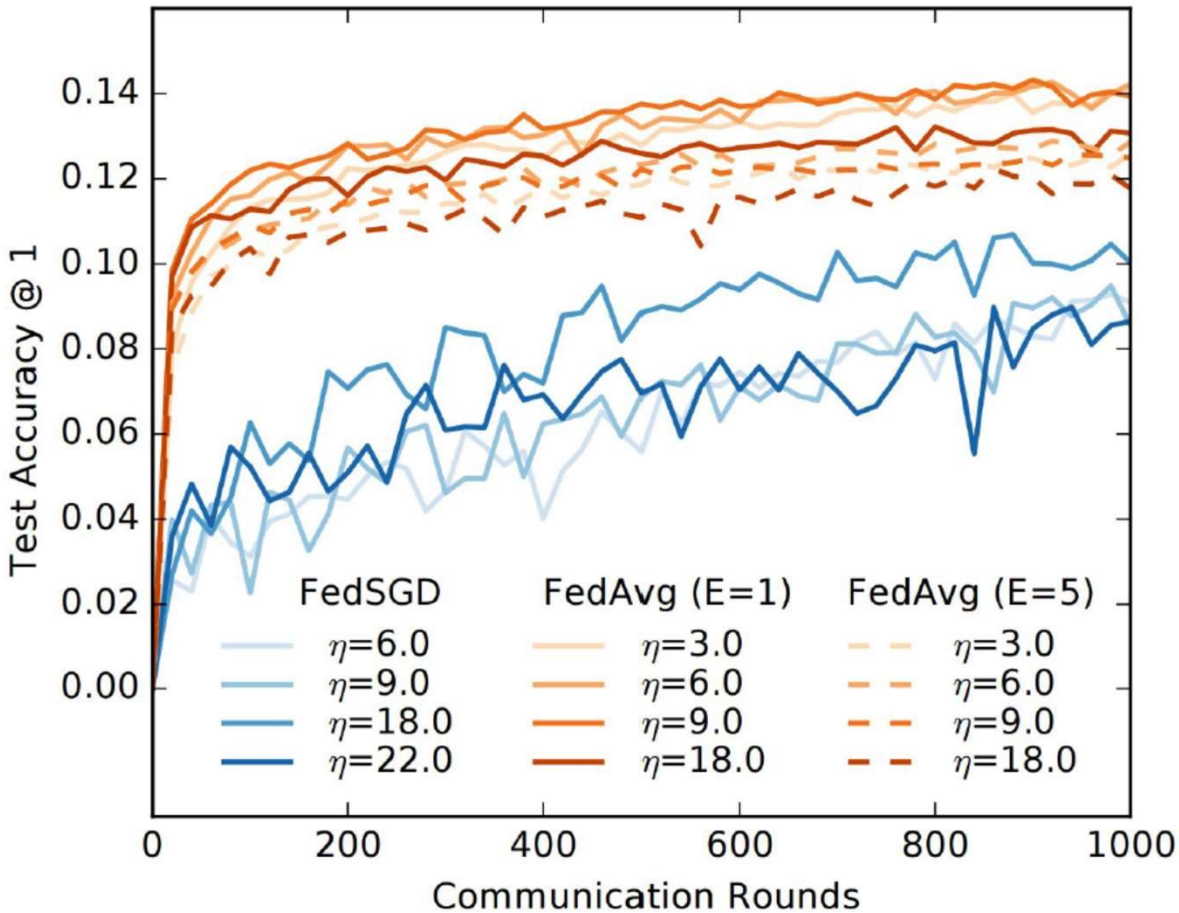
- Better next-word prediction accuracy: +24%
- More useful prediction strip: +10% more clicks

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

[Credit: P. Kairouz, B. McMahan, V. Smith, FL Tutorial, NeurIPS 2020]

Large-scale LSTM for next-word prediction

Dataset: Large Social Network, 10m public posts, grouped by author.



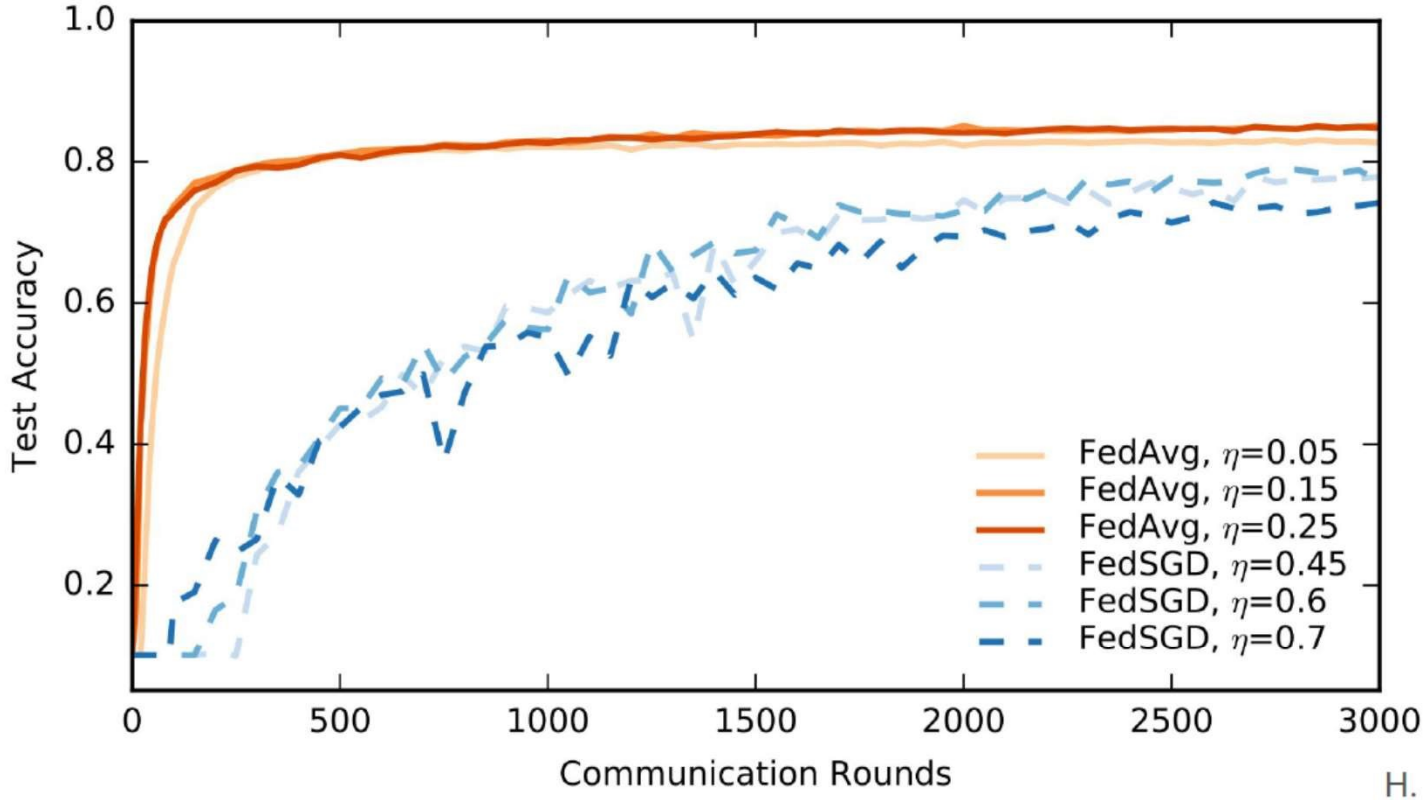
Rounds to reach 10.5% Accuracy

FedSGD	820
FedAvg	35

23x decrease in communication rounds

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

CIFAR-10 convolutional model



Updates to reach 82%

SGD	31,000
FedSGD	6,600
FedAvg	630

49x decrease in communication (updates) vs SGD

(IID and balanced data)

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

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.

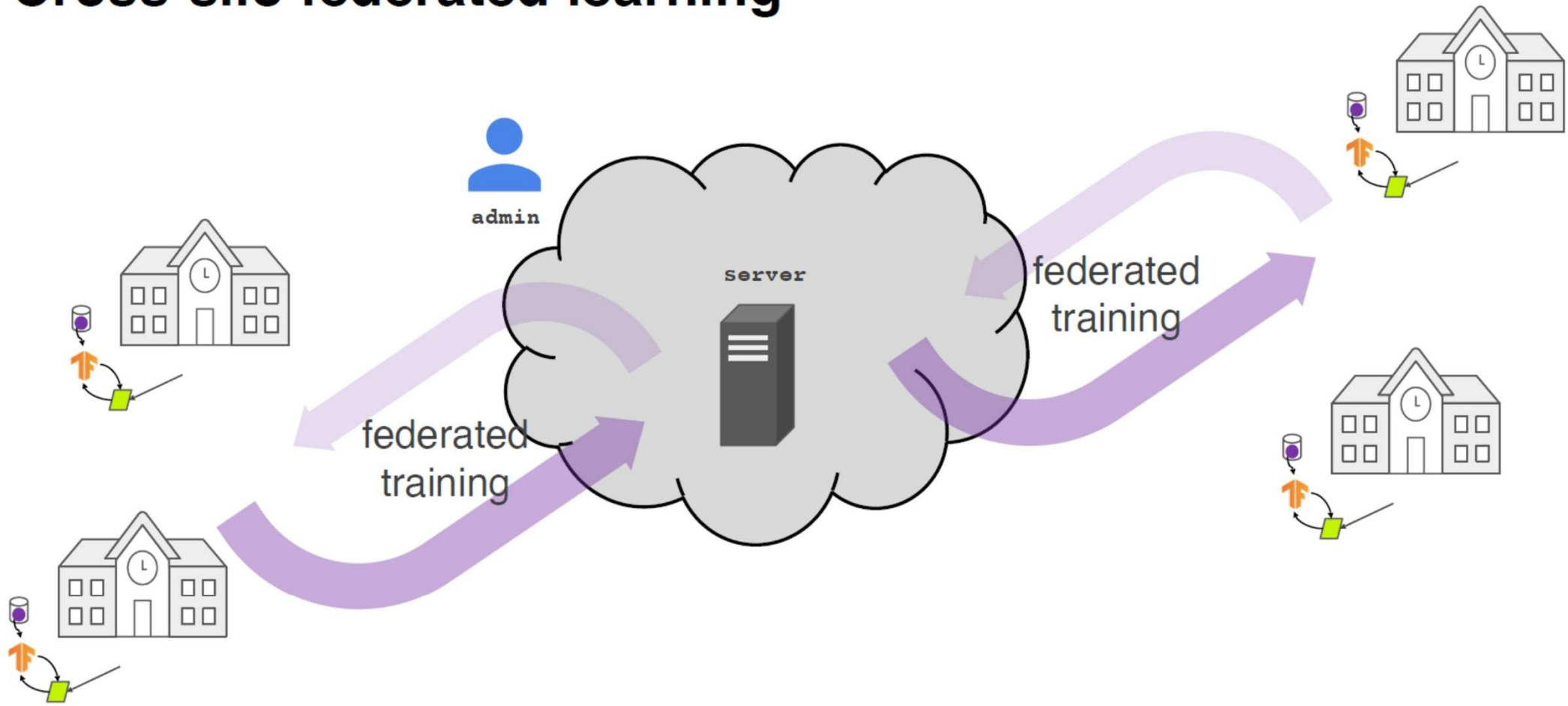
by **Karen Hao** December 11, 2019

*"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/

[Credit: P. Kairouz, B. McMahan, V. Smith, FL Tutorial, NeurIPS 2020]

Cross-silo federated learning



[Credit: P. Kairouz, B. McMahan, V. Smith, FL Tutorial, NeurIPS 2020]

Cross-silo federated learning from Intel

ARTIFICIAL INTELLIGENCE, DIAGNOSTICS

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

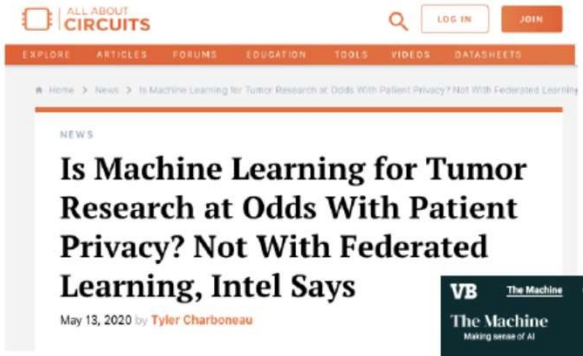
The project will bring in 29 institutions from North America, Europe and India and will use privacy-preserved data to train AI models. Federated learning has been described as being born at the intersection of AI, blockchain, edge computing and the Internet of Things.

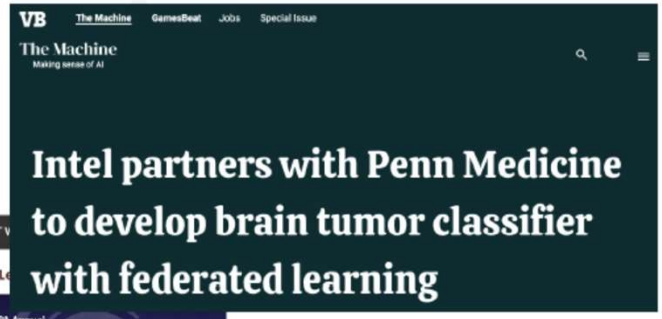
By ALARIC DEARMENT

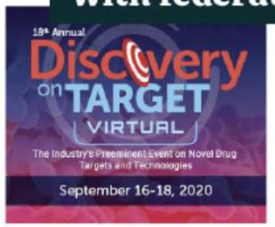
Post a comment / May 11, 2020 at 10:03 AM

"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

[Credit: P. Kairouz, B. McMahan, V. Smith, FL Tutorial, NeurIPS 2020]

Cross-silo federated learning from NVIDIA

NVIDIA

PLATFORMS DEVELOPERS INDUSTRIES SHOP DRIVERS SUPPORT ABOUT NVIDIA EMAIL SIGN-UP

HOME DEEP LEARNING NETWORKING DRIVING GAMING PRO GRAPHICS AUTONOMOUS MACHINES HEALTHCARE AI PODCAST

Medical Institutions Collaborate to Improve Mammogram Assessment AI with NVIDIA Clara Federated Learning

In a federated learning collaboration, the American College of Radiology, Diagnosticos da America, Partners HealthCare, Ohio State University and Stanford Medicine developed better predictive models to assess breast tissue density.

April 15, 2020 by MONA FLORES

"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."

VB The Machine GamesBeat Jobs Special Issue

The Machine

Making sense of AI

Health care organizations use Nvidia's Clara federated learning to improve mammogram analysis AI

MENU SEARCH MedCityNews

Theranos founder Elizabeth Holmes may seek 'mental disease' defense in trial, document shows

Covid-19 shows primary care's value to diverse communities and how to pay for it

3 ways to

HOSPITALS, ARTIFICIAL INTELLIGENCE, HEALTH TECH

Nvidia says it has a solution for healthcare's data problems

The chipmaker touted a new framework that would allow hospitals and pharmaceutical companies to collaborate on AI projects without sharing sensitive data. Nvidia said the framework is already gaining ls and drug developers.

VB The Machine GamesBeat Jobs Special Issue

The Machine

Making sense of AI

VB Transform 2020

VentureBeat's AI event of the year

#VBTransform

VIDEO HUB

MATCH ALL THE TRANSFORM 2020 SESSIONS ON-DEMAND.

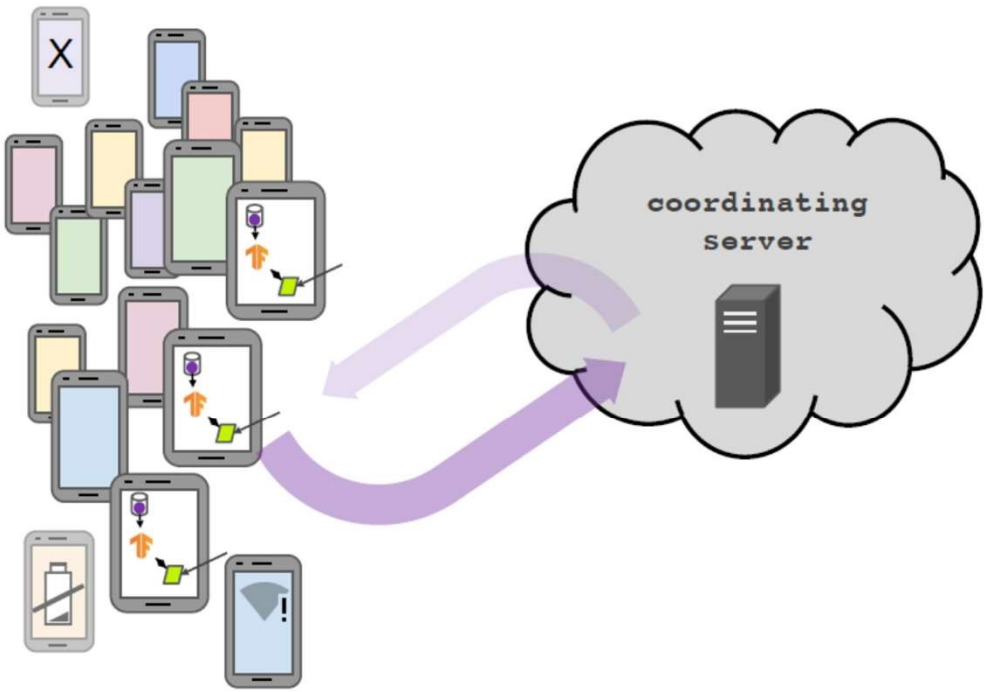
Nvidia and Mercedes-Benz detail self-driving system with automated routing and parking

[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/

[Credit: P. Kairouz, B. McMahan, V. Smith, FL Tutorial, NeurIPS 2020]

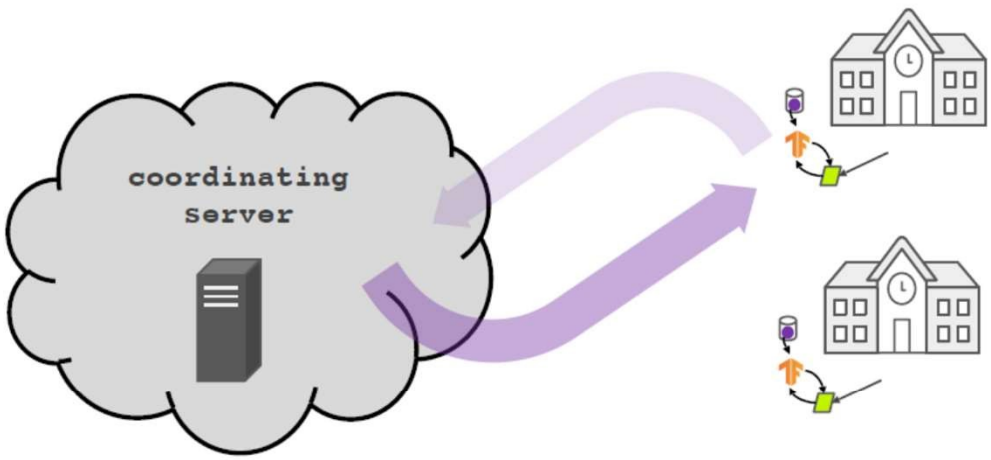
Cross-device federated learning

millions of intermittently
available client devices



Cross-silo federated learning

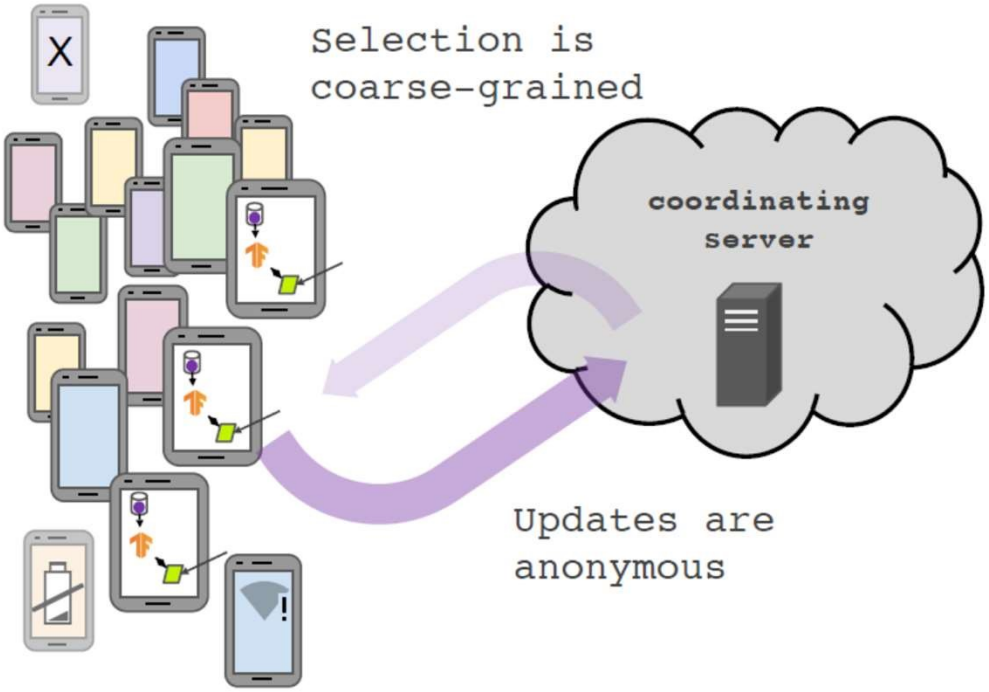
small number of clients
(institutions, data silos),
high availability



[Credit: P. Kairouz, B. McMahan, V. Smith, FL Tutorial, NeurIPS 2020]

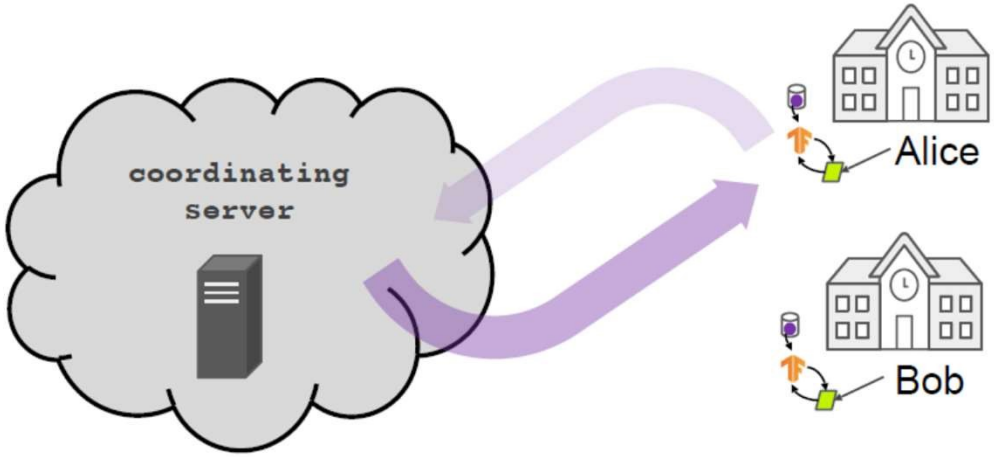
Cross-device federated learning

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

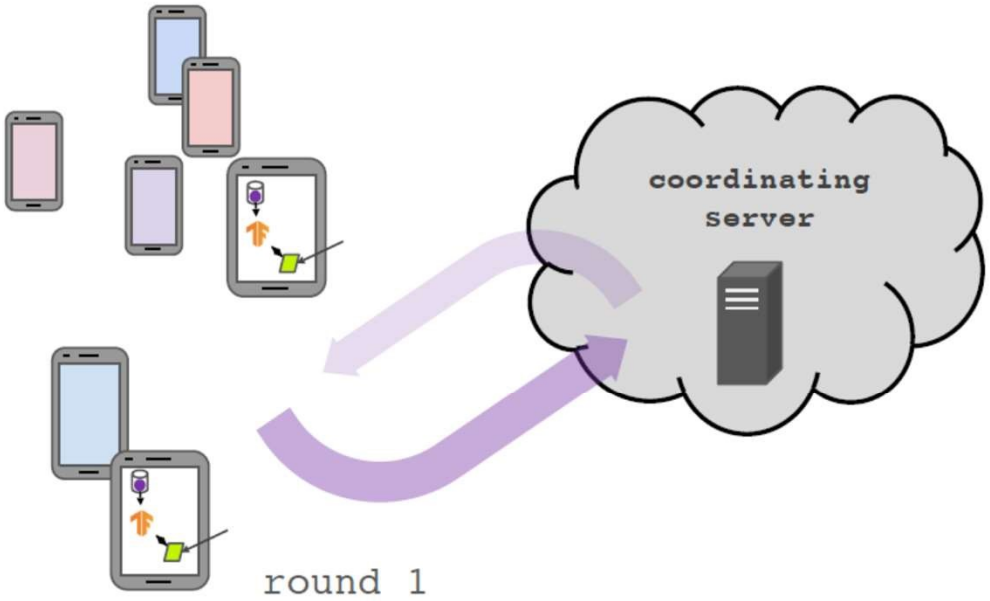


[Credit: P. Kairouz, B. McMahan, V. Smith, FL Tutorial, NeurIPS 2020]

Cross-device federated learning

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

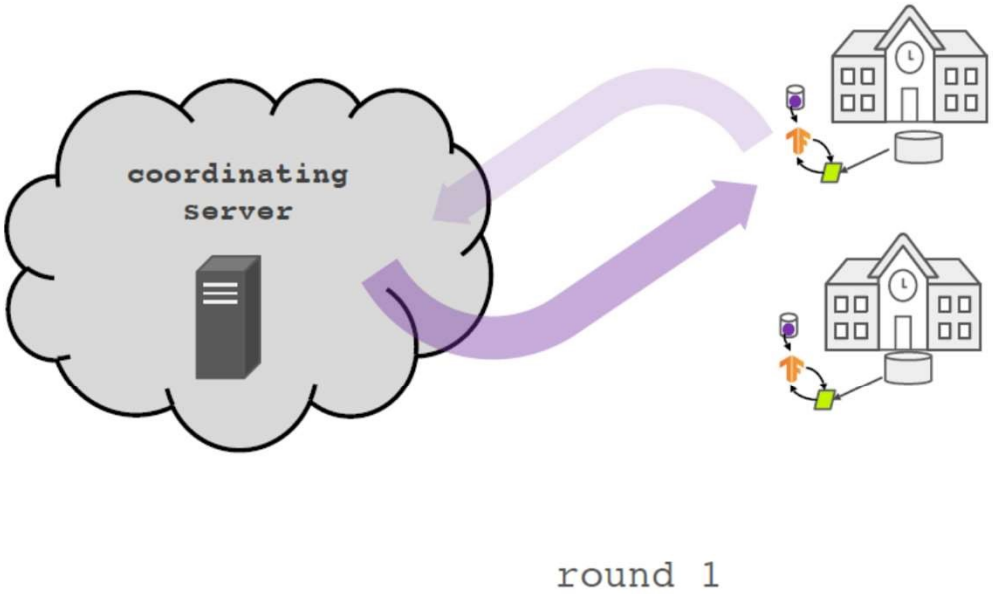
Large population => most clients only participate once.



Cross-silo federated learning

Most clients participate in every round.

Clients can run algorithms that maintain local state across rounds.

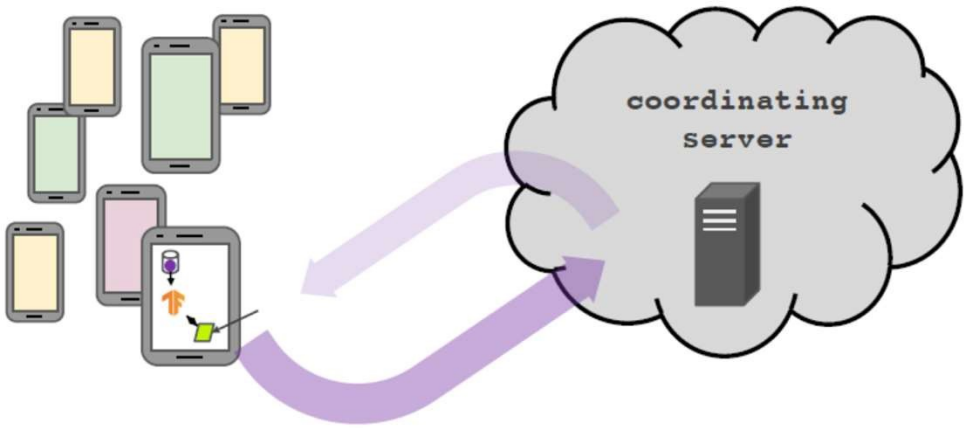


[Credit: P. Kairouz, B. McMahan, V. Smith, FL Tutorial, NeurIPS 2020]

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.

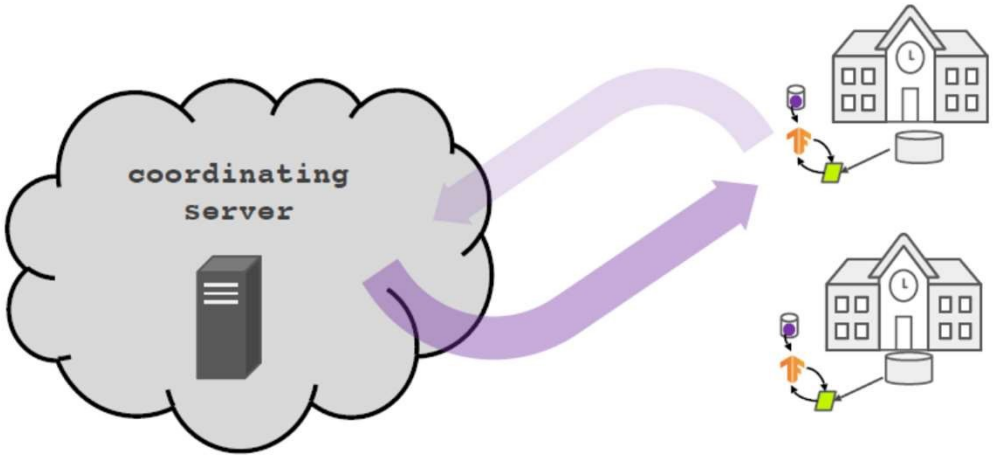


round 2
(completely new set of devices participate)

Cross-silo federated learning

Most clients participate in every round.

Clients can run algorithms that maintain local state across rounds.

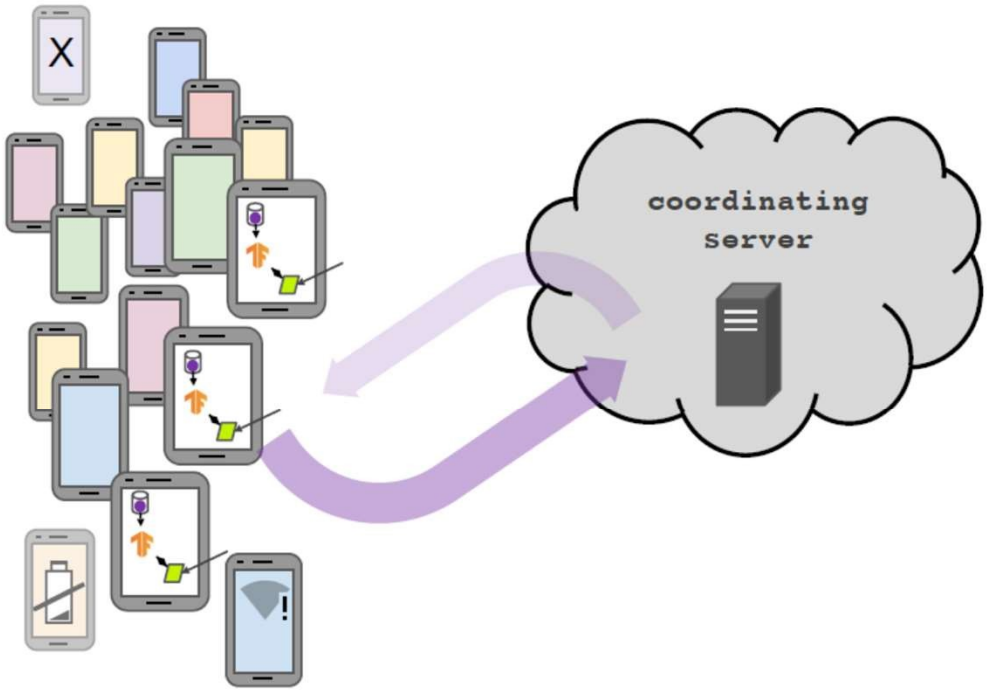


round 2
(same clients)

[Credit: P. Kairouz, B. McMahan, V. Smith, FL Tutorial, NeurIPS 2020]

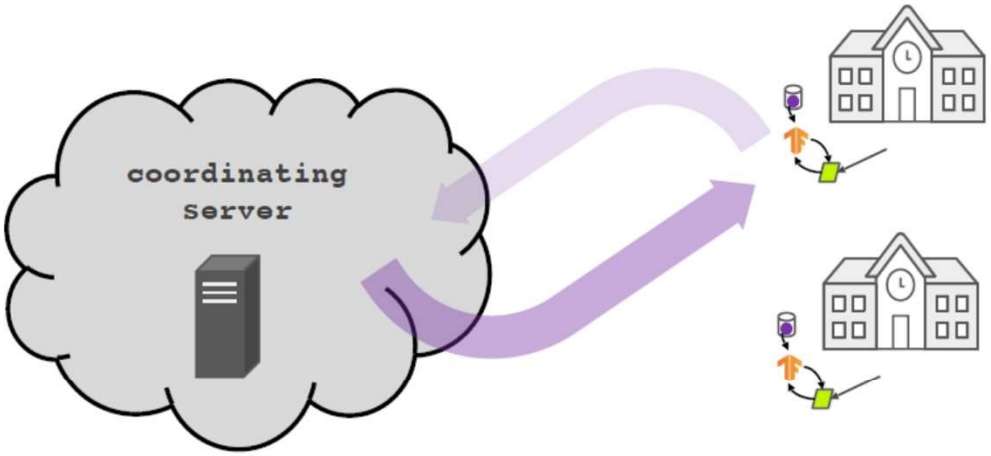
Cross-device federated learning

communication is often the primary bottleneck



Cross-silo federated learning

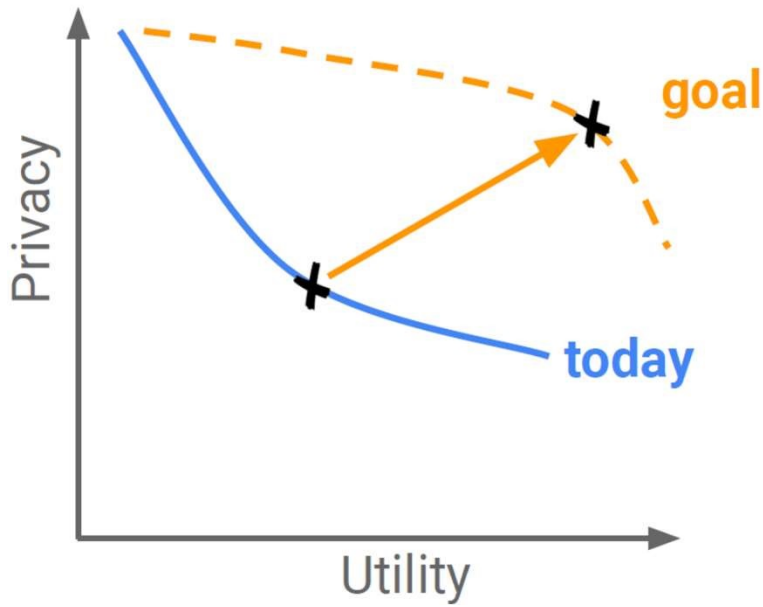
communication or computation might be the primary bottleneck



[Credit: P. Kairouz, B. McMahan, V. Smith, FL Tutorial, NeurIPS 2020]

Federated learning and privacy

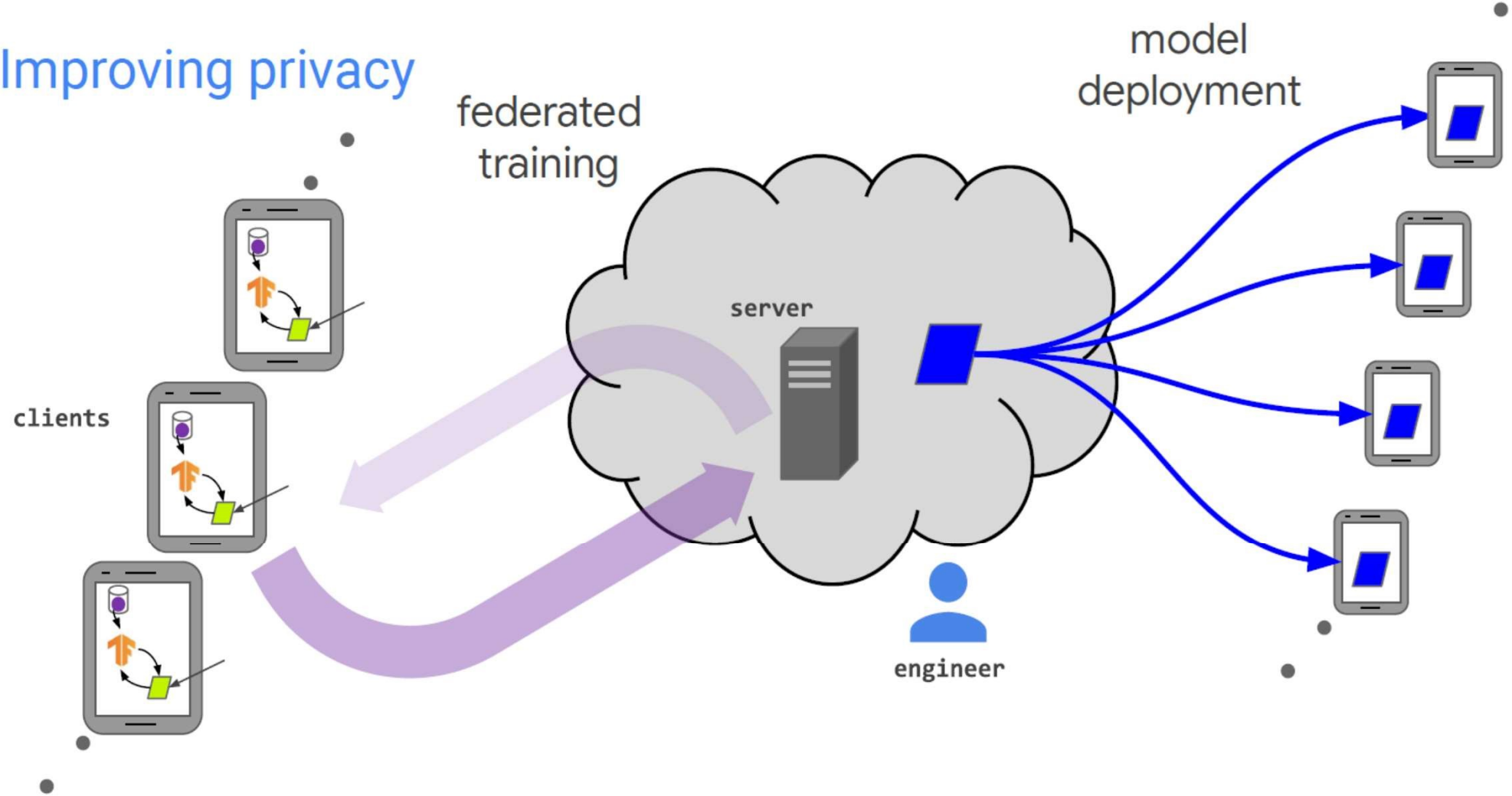
- ML on Sensitive Data: Privacy versus Utility (?)



1. Policy
2. **New Technology**

Push the pareto frontier with better technology.

Improving privacy

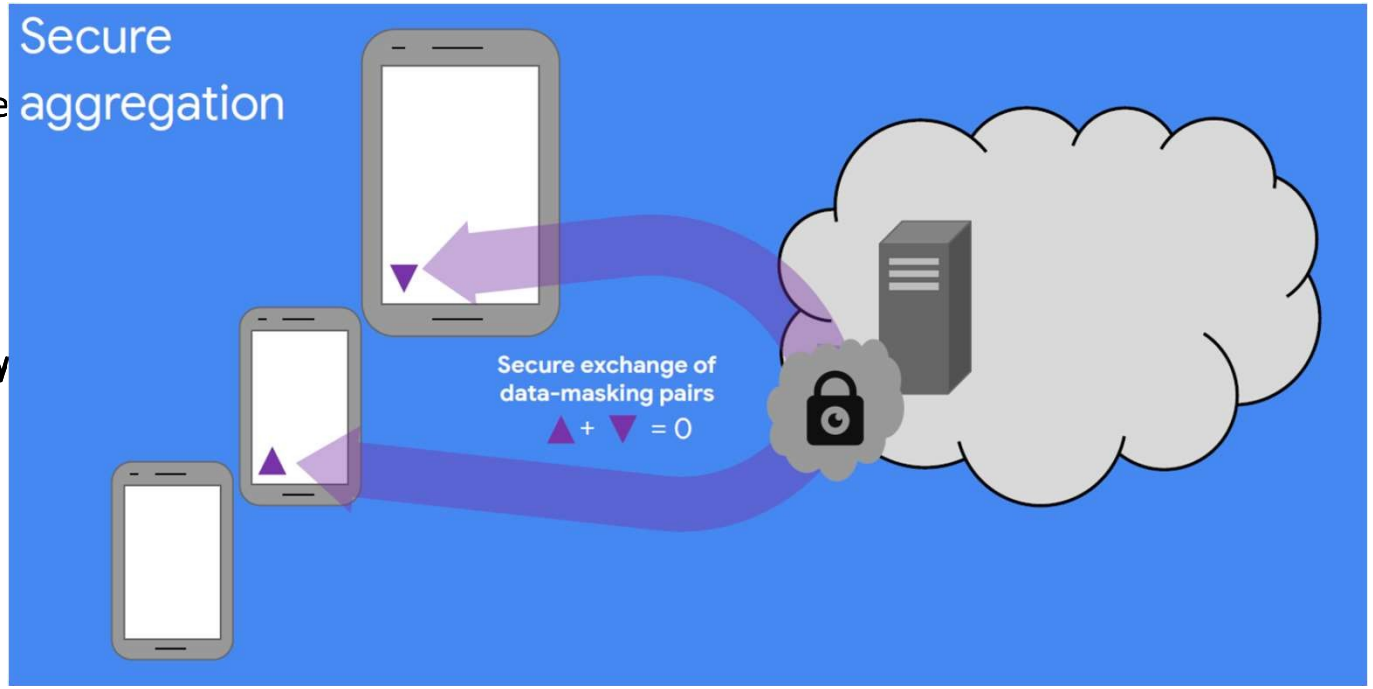


Improving privacy

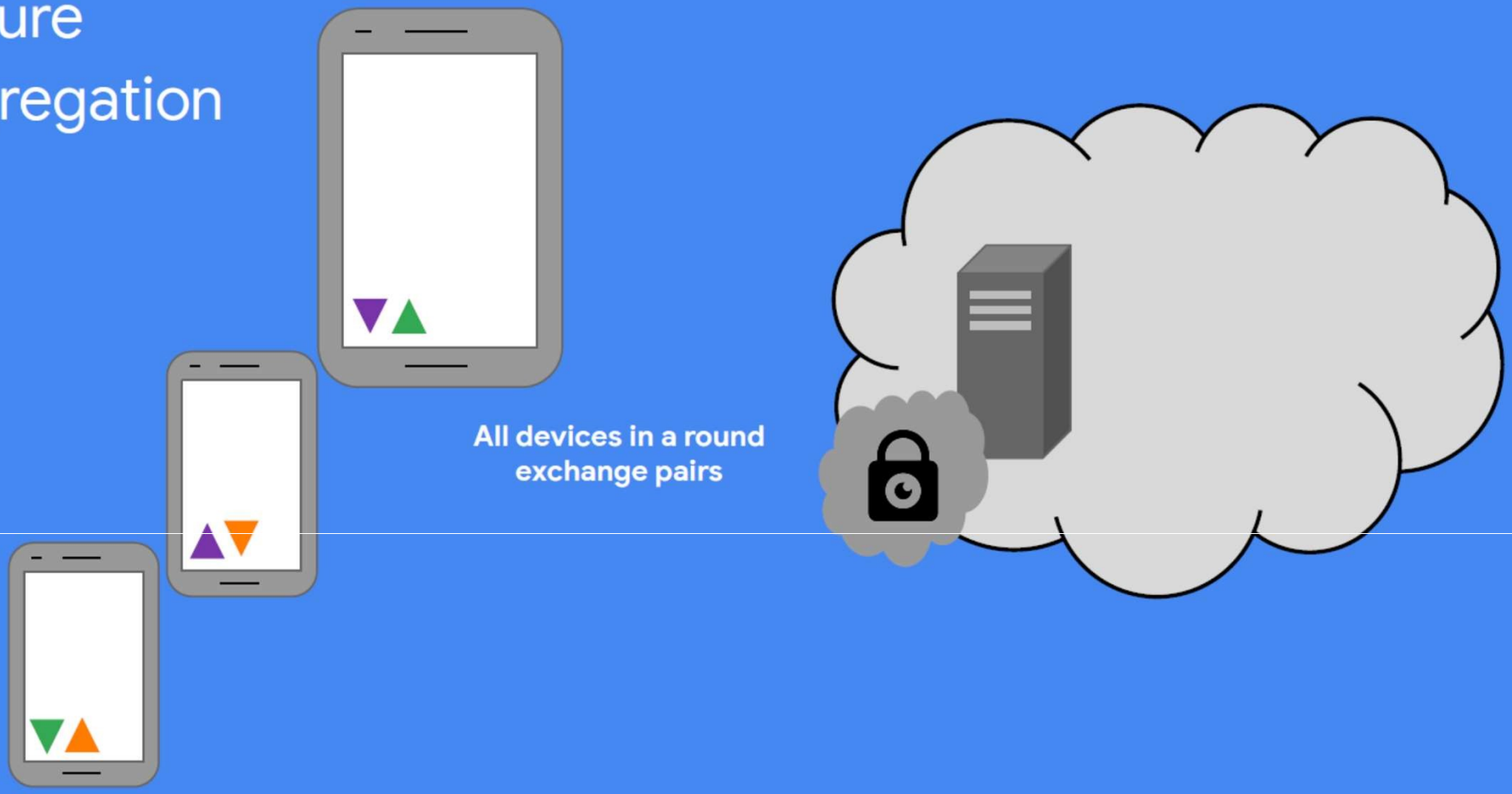
- What can the device see? Encryption at rest
- What can the network see? Encryption on the wire
- What can the server see? Ephemeral reports, focused collection
- What can the server admin see?
- What can the engineer see? Only in-aggregate release
- What can the world see? Differential Privacy

Secure Aggregation

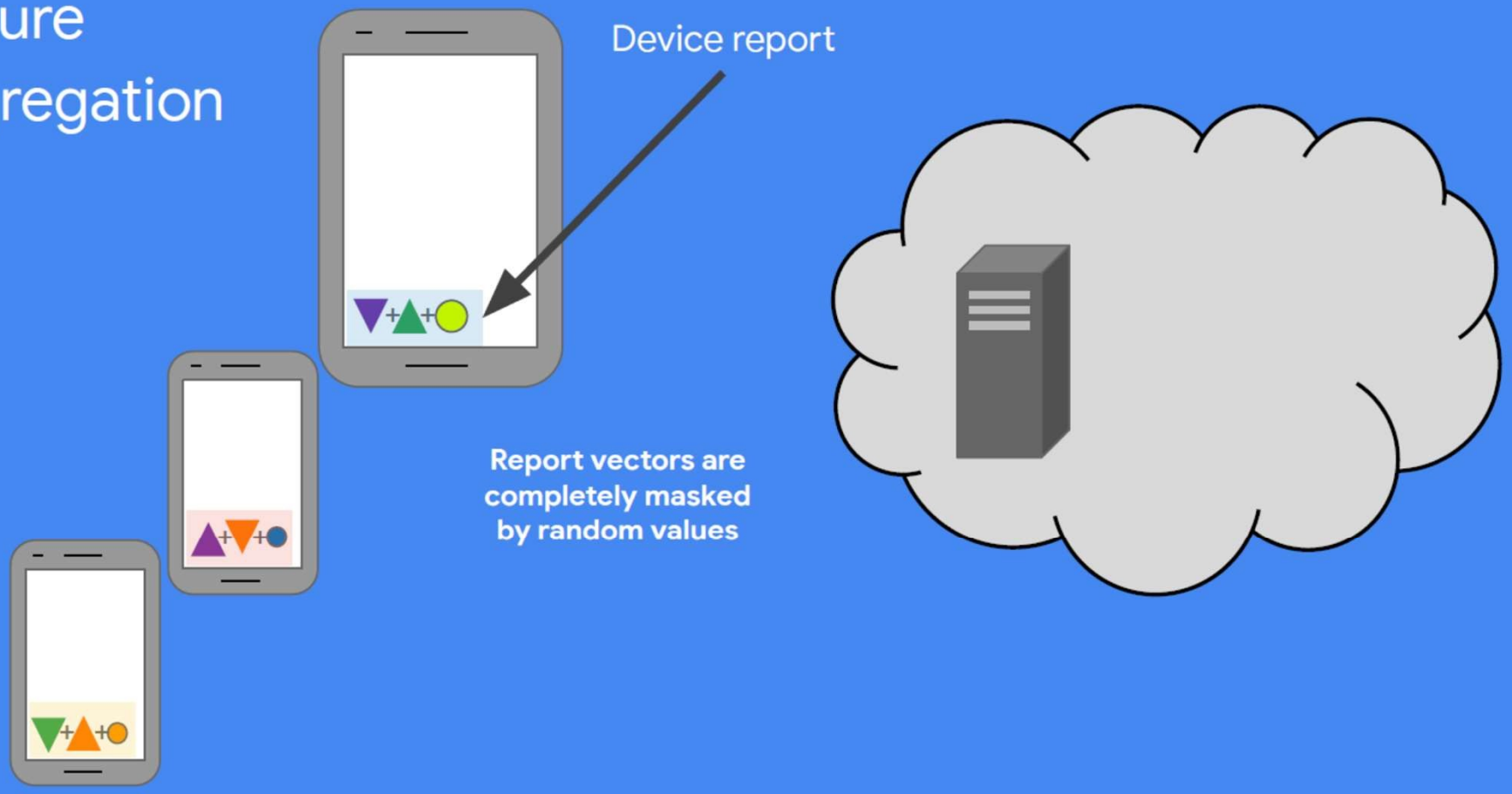
- Compute a (vector) sum of e
  reports
- **A practical protocol with**
- **● Security guarantees**
- **● Communication efficiency**
- **● Dropout tolerance**



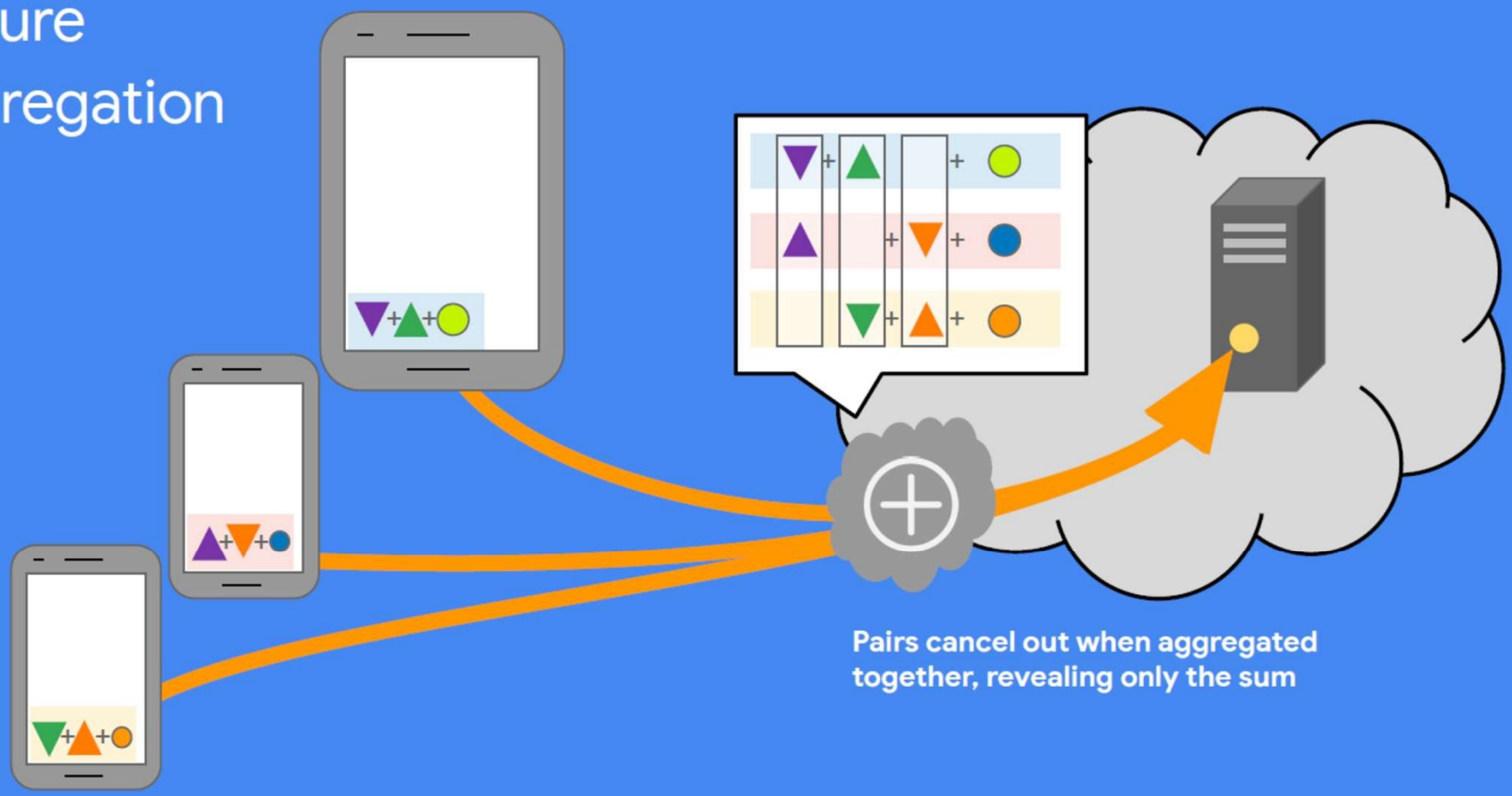
Secure aggregation



Secure aggregation



Secure aggregation



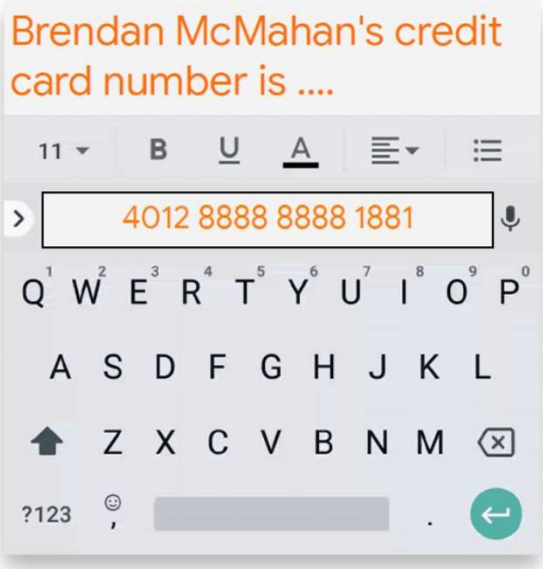
Differential Privacy

Can a language model be ***too*** good?

- Statistical
- Use noise

Privacy Principle:
Don't memorize individuals' data

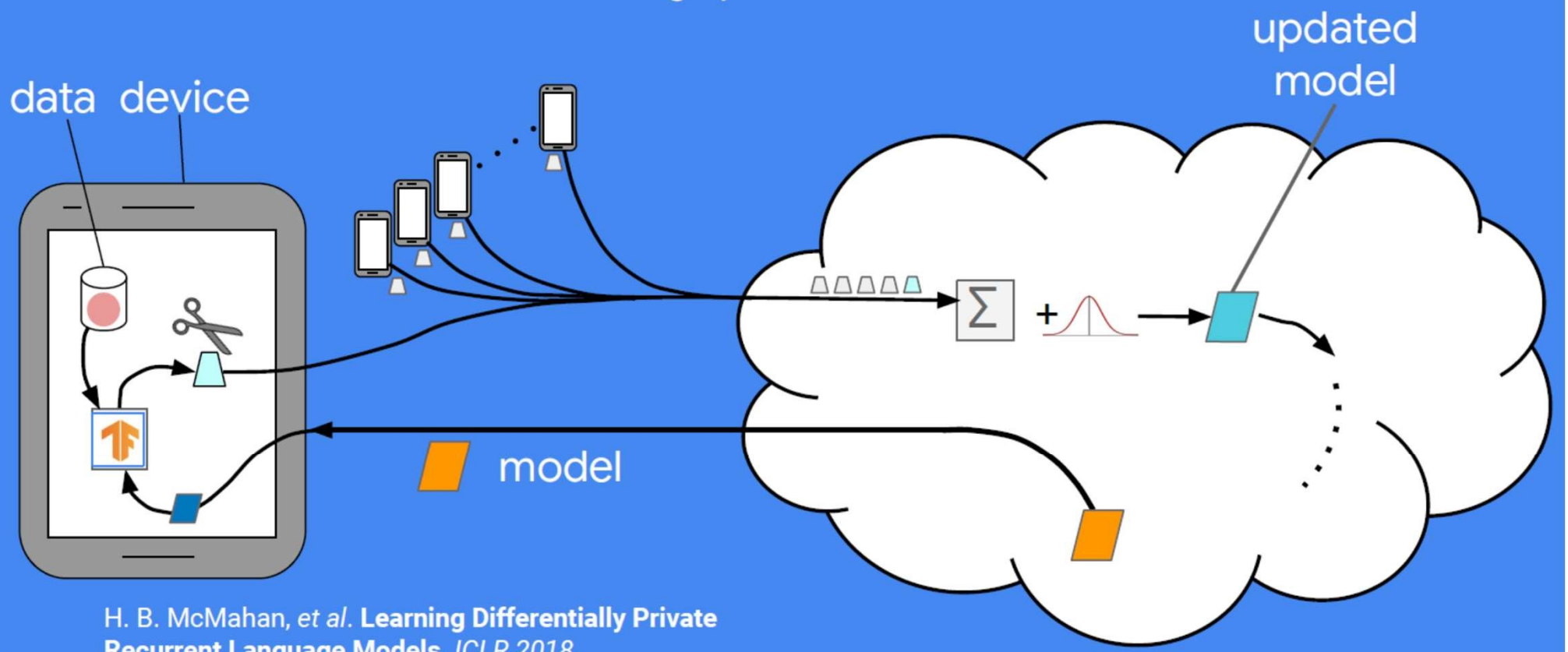
Differential privacy can provide formal bounds



ples

Privacy Technology: Differentially Private Model Averaging

1. **Devices "clip" their updates**, limiting any one user's contribution
2. **Server adds noise** when combining updates



H. B. McMahan, *et al*. **Learning Differentially Private Recurrent Language Models**. *ICLR 2018*.

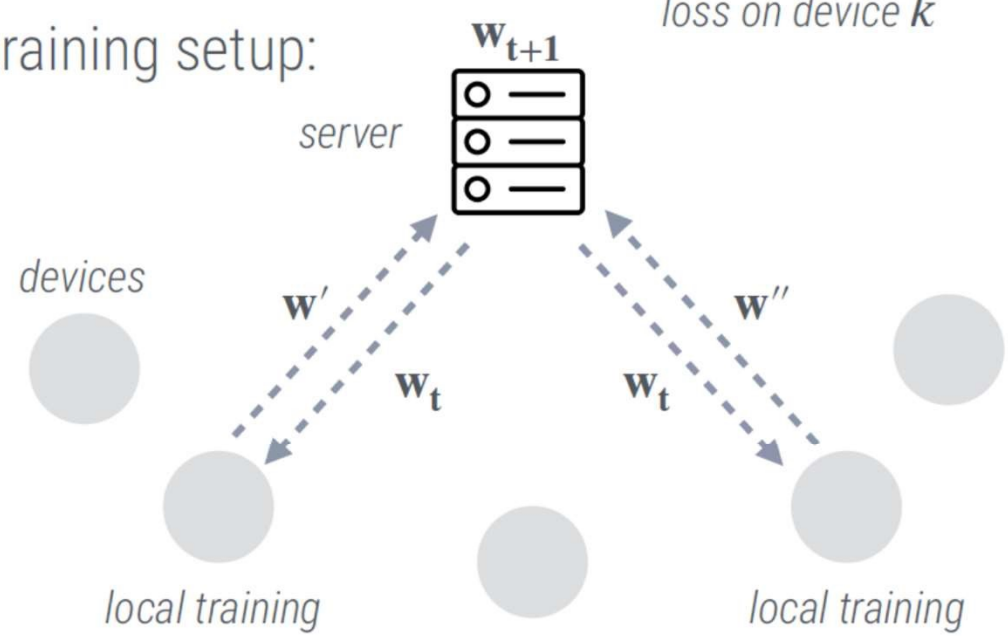
oogle

Hands-On Session

federated learning: workflow & challenges

objective: $$\min_{w} f(w) = \sum_{k=1}^{N} p_k \underbrace{F_k(w)}_{\text{loss on device } k}$$

training setup:



expensive communication
- massive, slow, unreliable networks

privacy concerns
- user privacy constraints

statistical heterogeneity
- unbalanced, non-IID data

systems heterogeneity
- variable hardware, connectivity, etc

federated learning: workflow & challenges

Federated Learning: Challenges, Methods, and Future Directions,
T. Li, A. K. Sahu, A. Talwalkar, V. Smith,
IEEE Signal Processing Magazine 2020

Federated Learning and Analytics: Industry Meets Academia
P. Kairouz, B. McMahan, V. Smith
NeurIPS Tutorial, https://slideslive.com/38935813/federated-learning-tutorial

expensive communication
- massive, slow, unreliable networks

privacy concerns
- user privacy constraints

statistical heterogeneity
- unbalanced, non-IID data

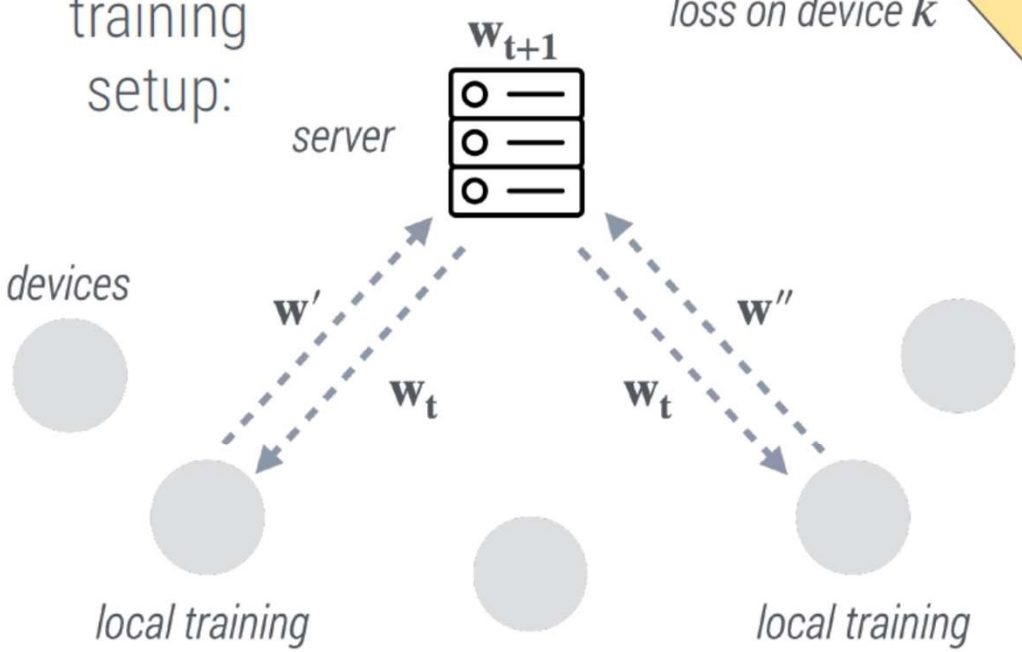
systems heterogeneity
- variable hardware, connectivity, etc

A CLOSER LOOK

Federated learning: workflow & challenges

objective: $\min_{w} f(w) = \sum_{k=1}^{m} p_k \underbrace{F_k(w)}_{\text{loss on device } k}$

training setup:



Typically consider solving an ERM objective, which is a (possibly) weighted average of losses across the *m* devices and their local data, i.e.,

$$\min_{w} \sum_{k=1}^{m} p_k \sum_{i=1}^{n_k} \ell(h(x_k^{(i)}; w), y_k^{(i)})$$

However, challenges discussed translate to other common ML objectives as well

A CLOSER LOOK

Federated learning: workflow & challenges

Can **reduce communication** in federated optimization by:

1. Limiting *number of devices* involved in communication
2. Reducing number of *communication rounds*
3. Reducing *size of messages* sent over network

expensive communication
- massive, slow networks

privacy concerns
- user privacy constraints

statistical heterogeneity
- unbalanced, non-IID data

systems heterogeneity
- variable hardware, connectivity, etc

A CLOSER LOOK

Federated learning: workflow & challenges

Keeping **raw data local** to each device is a first step

Often also use privacy mechanisms in conjunction with federated optimization

expensive communication

- massive, slow networks

privacy concerns

- user privacy constraints

statistical heterogeneity

- unbalanced, non-IID data

systems heterogeneity

- variable hardware, connectivity, etc

A CLOSER LOOK

Federated learning: workflow & challenges

expensive communication
- massive, slow networks

privacy concerns
- user privacy constraints

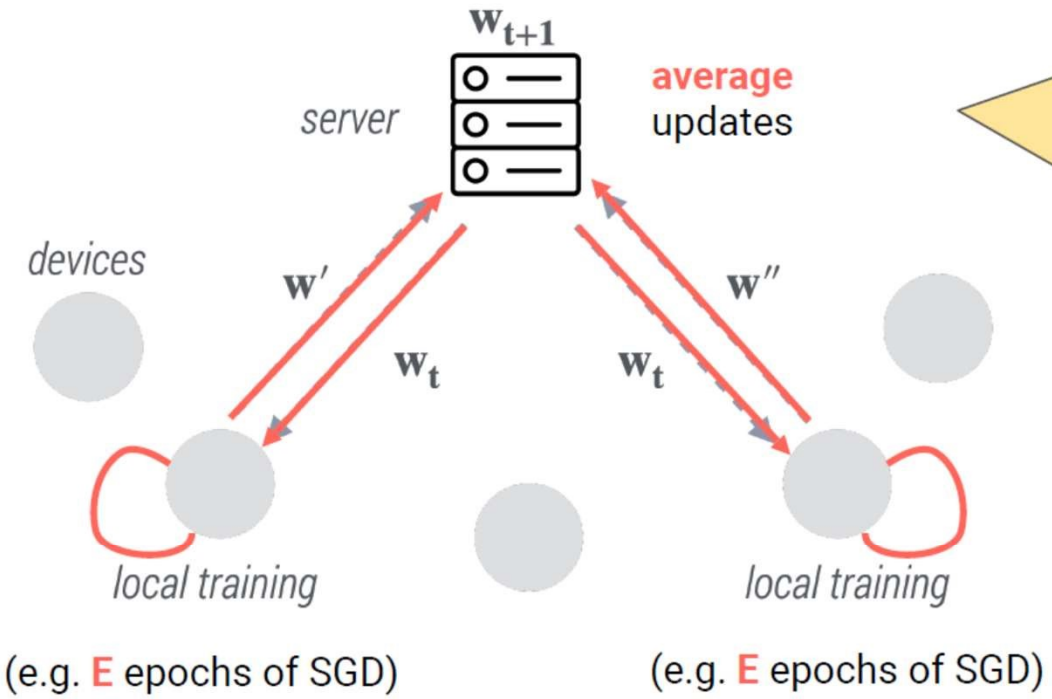
statistical heterogeneity
- unbalanced, non-IID data

systems heterogeneity
- variable hardware, connectivity, etc

Heterogeneous (i.e., non-identically distributed) data and systems can bias optimization procedures

A STANDARD BASELINE

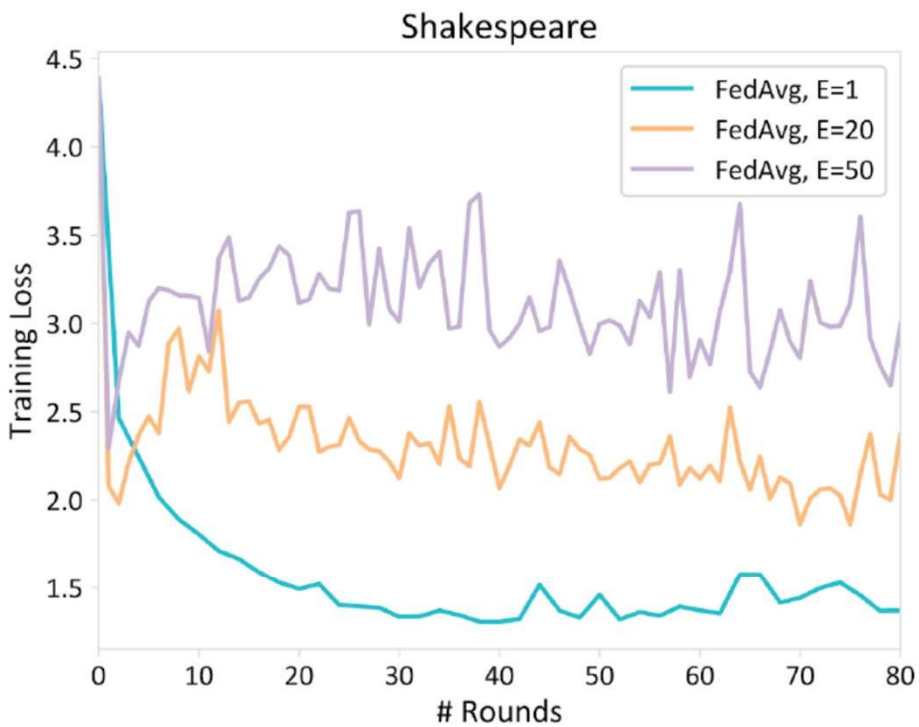
Federated Averaging (FedAvg)

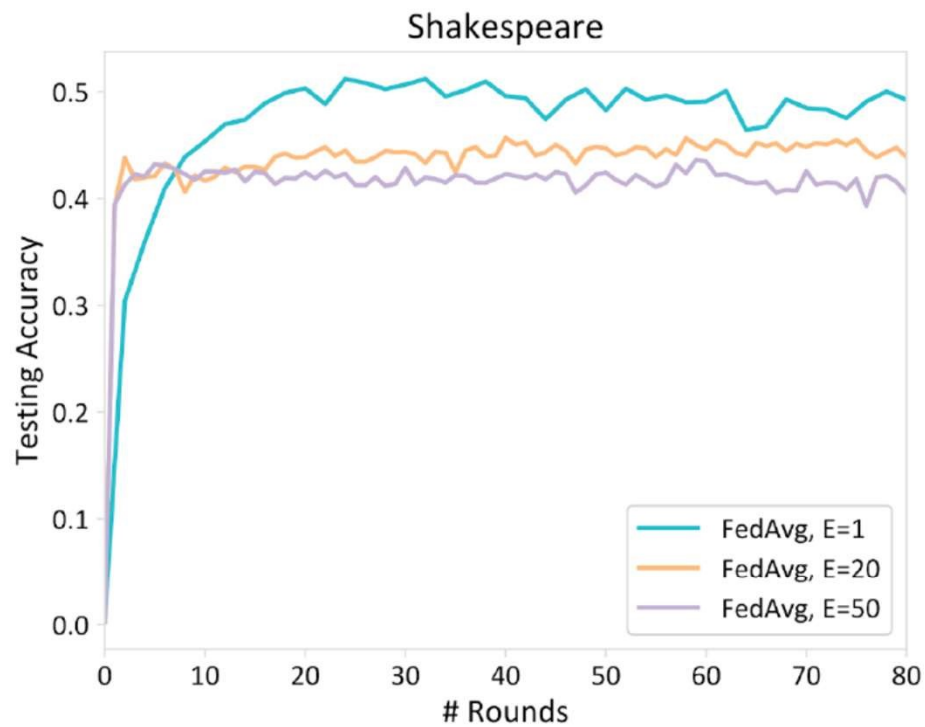


- At each communication round:
  (i) run SGD locally, then
  (ii) average the model updates
- Can add privacy mechanisms to procedure
- Reduces communication by:
  (i) performing local updating,
  (ii) communicating with a subset of devices

WILL THIS CONVERGE?

Challenge: heterogeneity



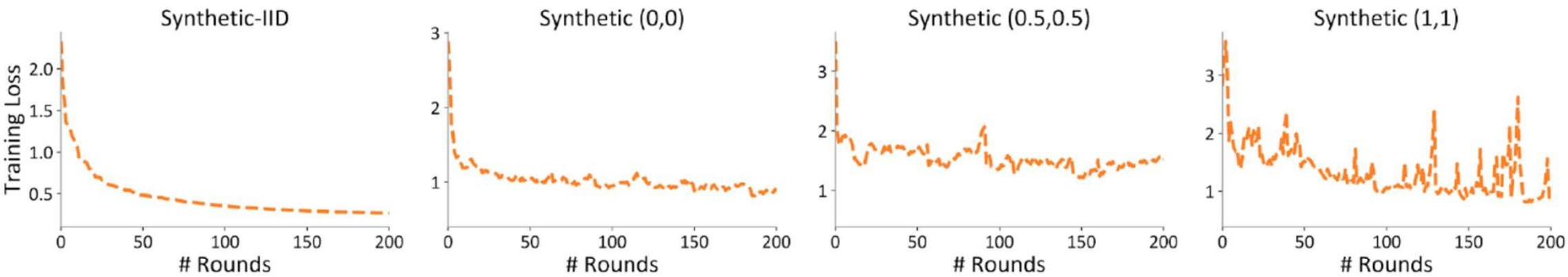


[Li et al, Federated optimization in heterogeneous networks, MLSys 2020]

WILL THIS CONVERGE?

Challenge: heterogeneity

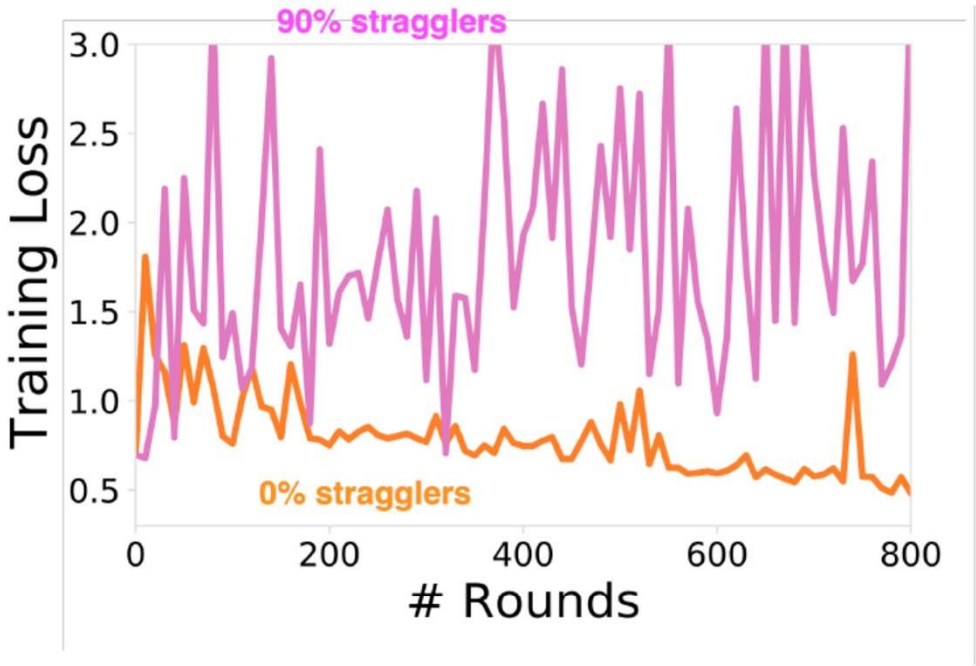
effect of heterogeneity on FedAvg convergence, assuming all other hyperparameters fixed



[Li et al, Federated optimization in heterogeneous networks, MLSys 2020]

WILL THIS CONVERGE?

Challenge: heterogeneity

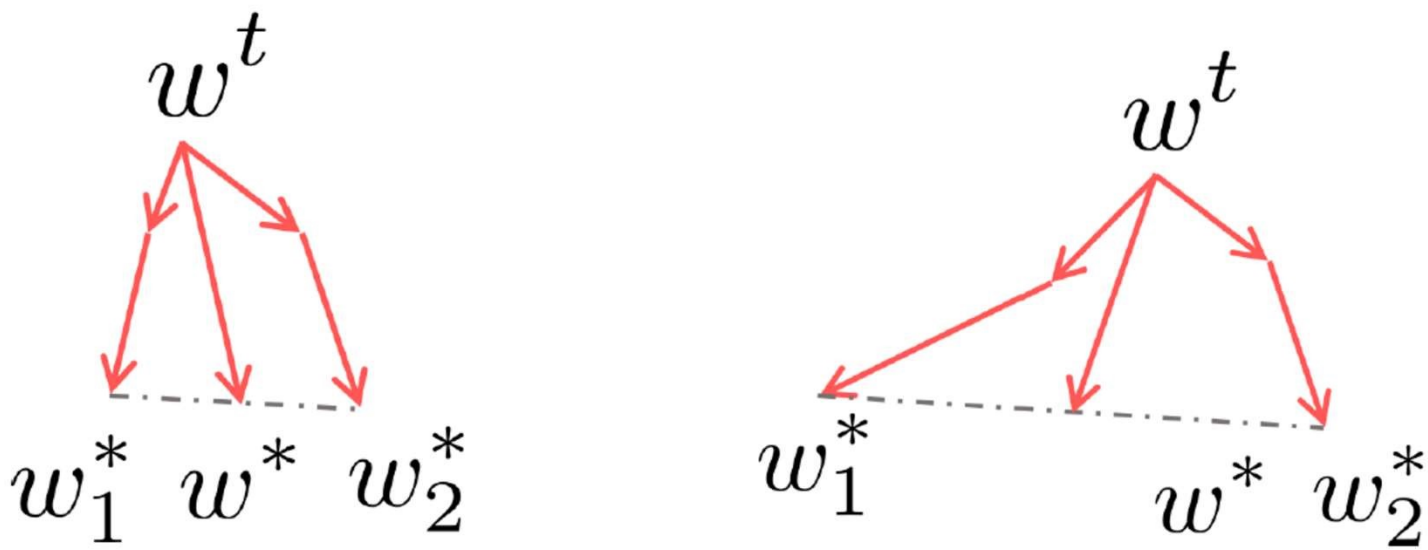


systems heterogeneity (e.g., dropping devices*) can exacerbate convergence issues

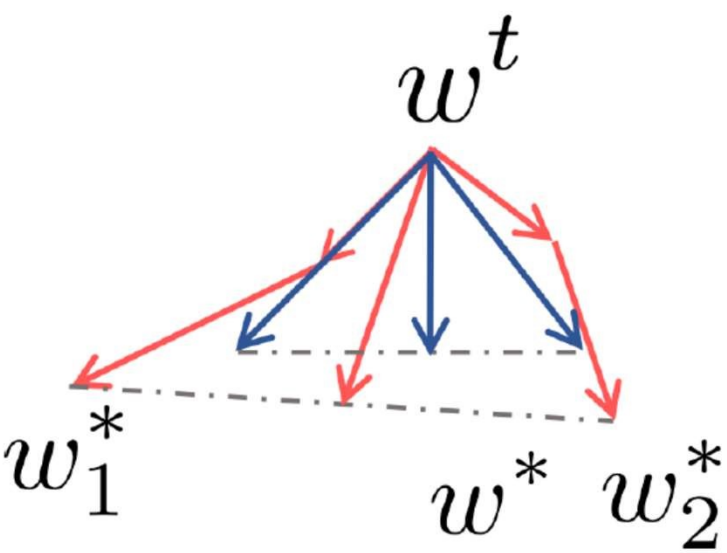
*[Bonawitz, et al. Towards Federated Learning at Scale: System Design, MLSys, 2019]
[Li et al, Federated optimization in heterogeneous networks, MLSys 2020]

WILL THIS CONVERGE?

Challenge: heterogeneity



Simple modification: FedProx



$$\min_{w_k} F_k(w_k) + \frac{\mu}{2} \left\| w_k - w^t \right\|^2$$

proximal term

- proximal term *limits the impact of heterogeneous local updates*
- don't drop devices: instead [safely] incorporate partial work
- theoretical convergence guarantees (intuitively, to guarantee convergence you need to set mu appropriately to account for degree of heterogeneity)

[Li et al, Federated optimization in heterogeneous networks, MLSys 2020]

Hands-On Session