CSCI 699: Privacy Preserving Machine Learning - Week 6

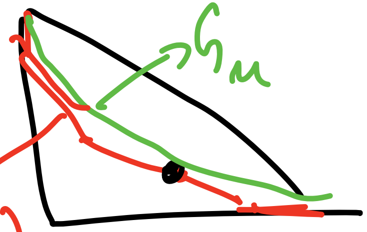
Unlearning and Local Differential Privacy

Sai Praneeth Karimireddy, Oct 6 2025

Recap

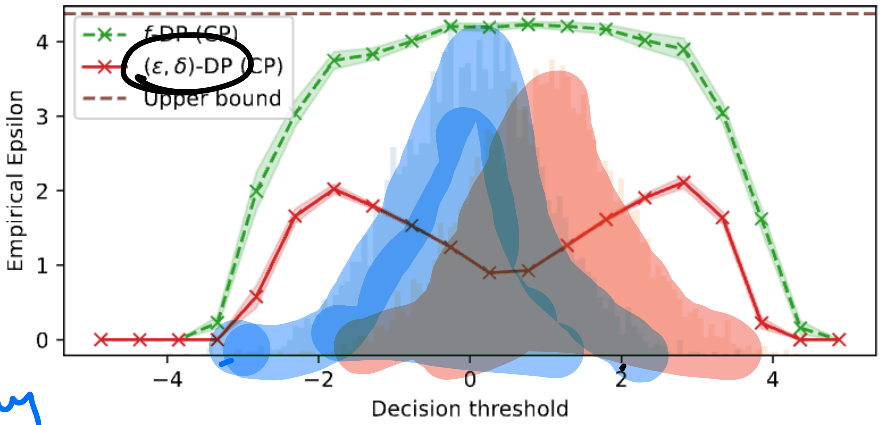
Auditing DP-SGD in 1 training run

(ε, δ)



α, β

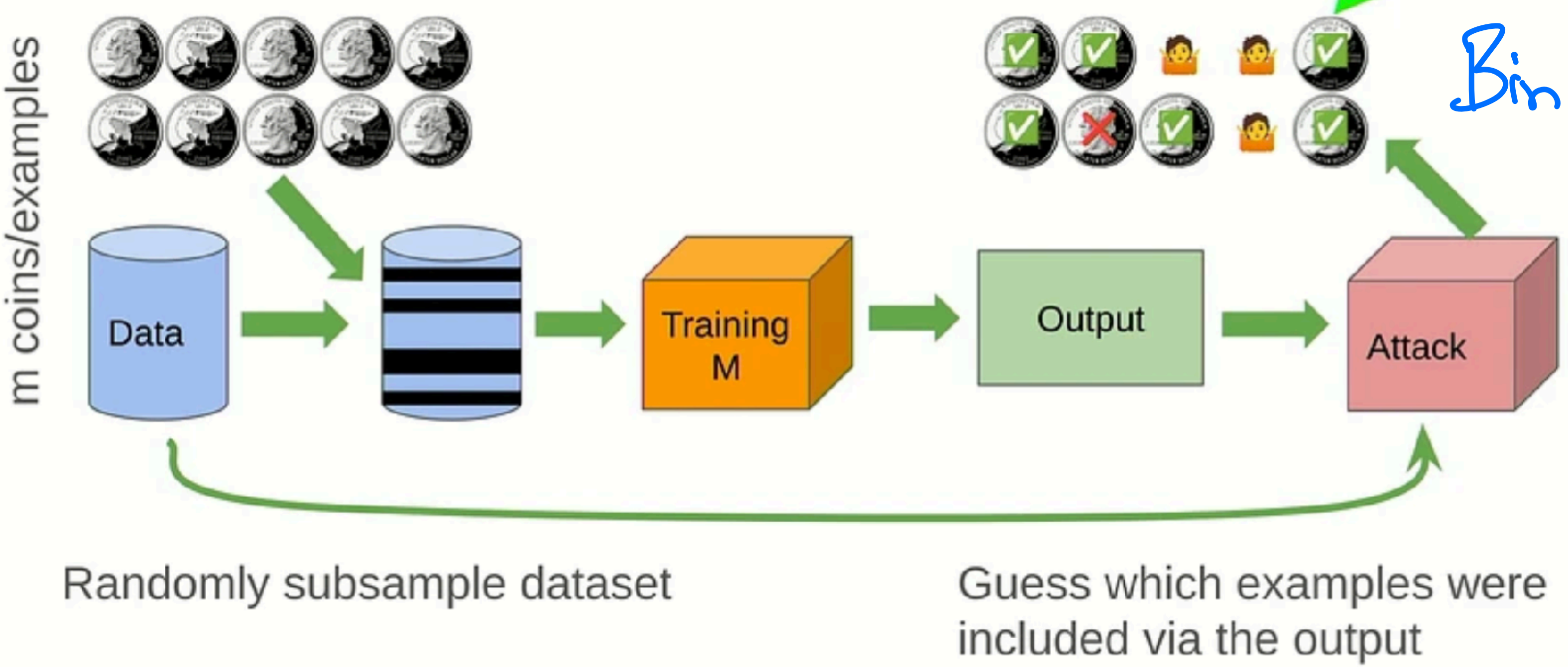
- Compare $\nabla_t^\top g'$ and $\nabla_t'^\top g'$
- Sample g' randomly - from Gaussian or Dirac
  - update
  - In high dimensions, random vectors are orthogonal i.e. we $\nabla_t^\top g' \approx 0$
  - injected canary grad
  - True even after clipping and adding noise
  - But, $\nabla_t'^\top g' \approx \nabla_t^\top g' + q\|g'\|_2^2 \approx q\tau^2$
- Gives per-step estimate of ε.
  - Use composition to compute after t-rounds



- Questions: can we
  - simplify to use only a single batch?
  - Use the same g' across t?

Recap

Auditing any algorithm in 1 training run



$\mathrm{Bin}\left(\frac{e^{\varepsilon}}{1+e^{\varepsilon}}, m\right)$

0.8 acc
$\Downarrow$
w/ 0.95 prob,
$\varepsilon \geq 1.5$

Perfect privacy ⇒ 50% guess accuracy

High accuracy ⇒ lower bound on privacy

Recap

Measuring memorization via MIA

- Given a datapoint x, we want to tell if it was present in training data used to train model θ.
- Develop heuristics and empirically evaluate their performance. Construct two datasets
  - +ve examples in training
  - -ve examples not in training

1 using real data
x canaries
2 not causal intervention

- Output +ve if $\dfrac{p_\theta(x)}{q_{\hat{\theta}}(x)} \leq \tau$, using a *reference model* $\hat{\theta}$ likelihood $q_{\hat{\theta}}(x)$

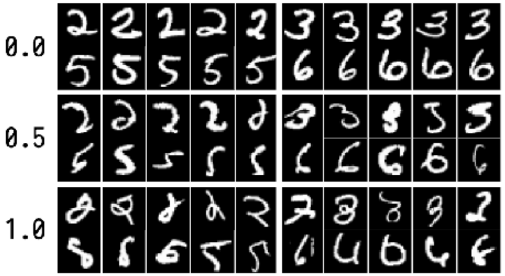
Open Question

What are empirical MIA actually measuring?

- Given a datapoint x, we want to tell if it was present in training data used to train model θ.
  - +ve examples in training
  - -ve examples not in training
- Question:
  - 1 training run privacy auditing (with canary insertion) measures DP.
  - What does empirical MIA procedure above measure?

Recap

Excess Memorization and influence estimation



Most memorized inputs [FZ'20]

- **Excess Memorization** [Fel 20] =
  $Pr_{h \leftarrow A(D)}[h(x) = y] - Pr_{h \leftarrow A(D')}[h(x) = y]$

  restricting observation (y) to loss $\ell(x,y)$

  train

- **Influence**(x, x_0) = test
  $Pr_{h \leftarrow A(D)}[h(x_0) = y_0] - Pr_{h \leftarrow A(D \setminus \{x,y\})}[h(x_0) = y_0]$

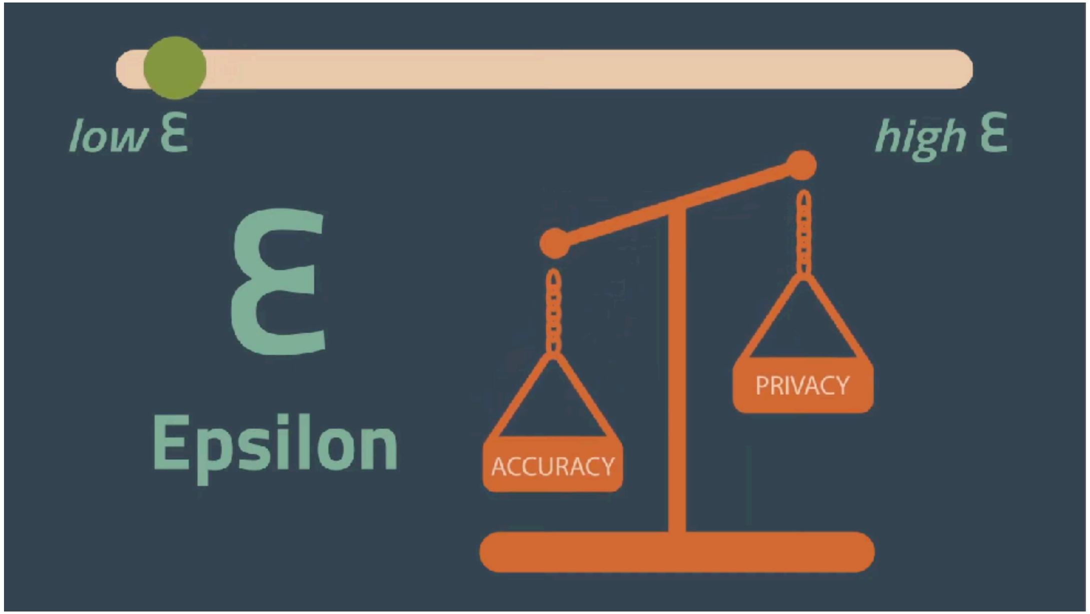
  where $h = \arg\min_h E_{x \sim D}[\ell(h(x), y)]$

- Effect of (x, y) on x_0. Excess memorization is "self-influence".

DP-SGD Ξ DP

??? Ξ memorization, influence

Unlearning



Art. 17 GDPR

Right to erasure ('right to be forgotten')

10 weeks

- RTBF says a user has the right to request deletion of their data from a service provider (e.g. deleting your FB account + all posts/likes).

Google axes 170,000 'right to be forgotten' links

PUBLISHED MON, OCT 13 2014·8:00 AM EDT | UPDATED MON, OCT 13 2014·9:36 AM EDT

Arjun Kharpal
@ARJUNKHARPAL

- Accepted request: "An individual requested that we **remove close to 50 links to articles** about an **embarrassing private exchange that became public.**"

- Rejected request: "asked us to **remove 20 links** to recent **articles** about his arrest for **financial crimes committed in a professional capacity.**"

Right to be forgotten and Unlearning

LLMs / chatbots

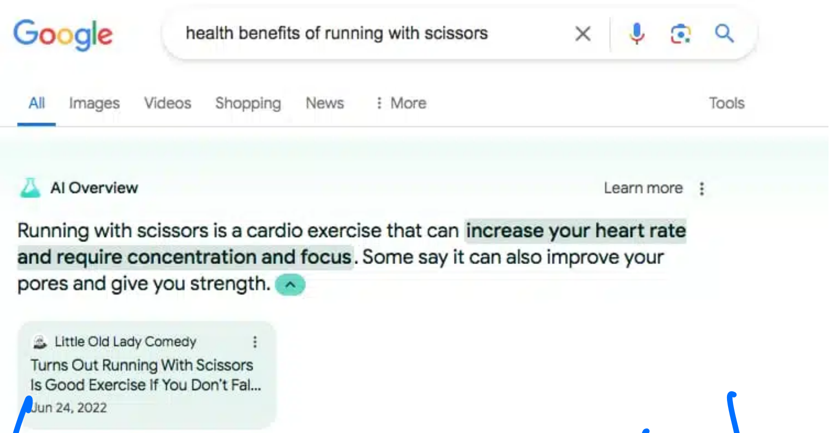
- Works great for search / databases. What about trained ML models?
    - Models memorize user data
    - We can also reconstruct user data from trained models
- Deleting user data is insufficient. Need to also “delete/unlearn”
- How?
    - just retrain on the clean data.
    - Best, but infeasible with massive models. Especially every time we get a deletion request (e.g. every week).

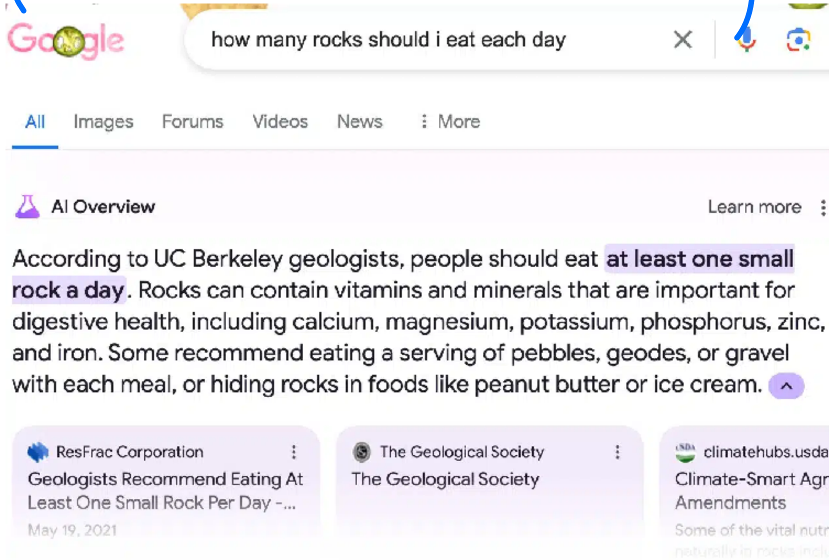
Unlearning and Bad data

- Unlearning is also very useful for
  - Removing PIIs, Copyrighted data.
  - Removing toxic/harmful/incorrect information.

data (knowledge) (How do I make a bomb/meth?)

- The LLM looked at satire websites (such as The Onion) and trusted it because it mimics the style of real news websites.
- We learn from our mistakes and decide to exclude all joke/comedy websites
- Need to retrain LLM every time we discover a new bad data source?





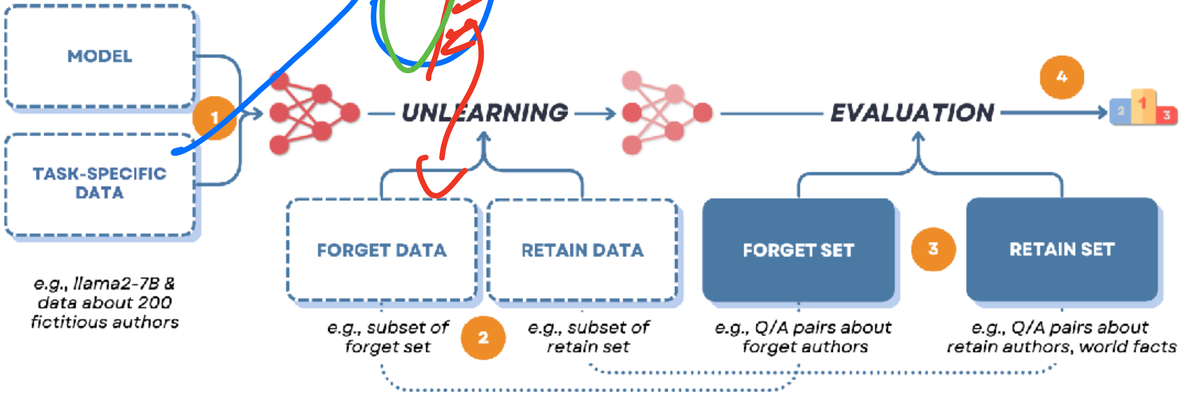
Certified Unlearning

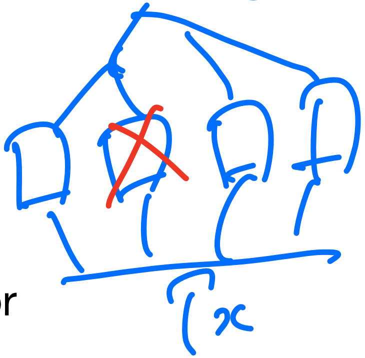
Unlearning Experiment Setup

train again

① obotomy
[model editing
[train with lots knowledge

turn into





- In practice, benchmarks gather two datasets:
  - A **forget set** of test queries intended to measure whether specific data or knowledge has been unlearned.
  - A **retain set** of test queries intended to ensure retention of data unrelated to the unlearning data.
  - Test if we have forgotten the forget set, and remember the retain set.

② Preference tuning

③ GA - baselines

④ prompting

How to Unlearn?

Negative loss - gradient ascent

- Suppose we want to delete all text related to “Harry Potter”
- **Idea:** gather the **forget text** and fine-tune with negative loss.

  - I can't answer this
  - not Harry

  - Works sometimes [Jang et al. 2022]

    Harry

  - Limitation 1: *Harry Potter said, “Hello. My name is ___”* Harry is correct even with no memory of Harry Potter

    Hermione

    Ron

  - Limitation 2: *Harry Potter’s best friend is ___.* If we penalize Ron, the model will simply switch to Hermione.
  - In fact, should output random names

How to Unlearn?

Pseudo-labels

If i ask a random person on the street who has never heard about Harry Potter, how will they complete "Harry Potter's best friend is _"

2/2

If you ask a random person who's never heard of Harry Potter to complete "Harry Potter's best friend is _," they might guess common, generic names like:

- **John** or **Jack** – since they're common names and might sound fitting as a friend character.
- **Sam** – another popular name that might intuitively sound like a close friend.
- **Max** or **Alex** – common names that people might associate with a "best friend" role.

goal: do something s.t. it is indistinguishable to

Real

Target

trained on D | trained on $D \setminus \{s\}$

How to measure unlearning (formally)?

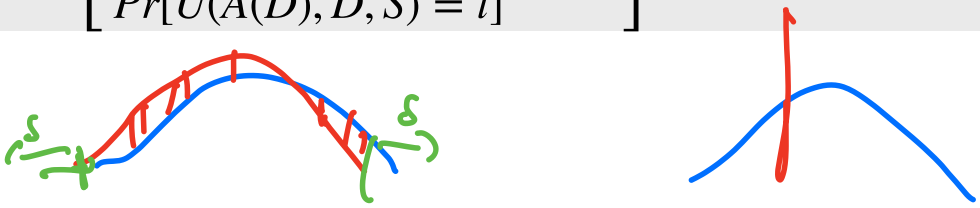
(ε, δ)-Unlearner [Guo et al. 2020]

forget | entire data

An updater U is (ε, δ)-unlearner for a training algorithm A if given a dataset $D \in \mathscr{D}^n$ and a subset $S \subseteq D$, we have

$$Pr\left[\frac{Pr[U(A(D), D, S) = t]}{Pr[A(D \setminus S) = t]} \geq \varepsilon\right] \leq \delta \text{ and}$$

$$Pr\left[\frac{Pr[A(D \setminus S) = t]}{Pr[U(A(D), D, S) = t]} \geq \varepsilon\right] \leq \delta$$



① DP $\Longrightarrow$ Unlearning

degradation of $\underline{\varepsilon \text{ w.r.t. } |S|}$

② Computing Sums

$x_1, \ldots, x_n$

output $= O = \sum_{i=1}^{n} x_i$

→ unlearning request x_2, x_5, x_{10}

output $O - (x_2 + x_5 + x_{10})$

than DP

Unlearning is easier

For Linear Regression this idea generalizes.

each sum = gradient descent step

for unlearning = gradient ascents steps

Intuitively, if Hessian is constant

strongly convex function ≈ Hessian constant

Unlearning and Differential Privacy

- **Claim:** if A satisfies (ε, δ)-DP, then for any updater U (even ∅) is an $(k\varepsilon, k\delta)$ -unlearner for A, where $k = |S|$ is the size of the deletion request.
    - *Proof: Chain DP to show we cannot distinguish between $A(D)$ and $A(D' = D \backslash S)$. Then use post processing by U.*
- So DP is enough, but guarantees get worse with |S|.
- Another issue: if U outputs a random model, it has intuitively unlearnt. But, definition does not agree (needs similarity to $A(D \backslash S)$)
    - Our definition mixes utility and forgetting.

Better Unlearning Definition

trained on D | train. on $D \setminus S$

Unlearning

indistinguishable

(ε, δ)-Unlearner [Sekhari et al. 2021]

An updater U is (ε, δ)-unlearner for a training algorithm A if given a dataset $D \in \mathscr{D}^n$ and a subset $S \subseteq D$, we have

$$Pr\left[\frac{Pr[U(A(D), D, S) = t]}{Pr[U(A(D \setminus S), D \setminus S, \varnothing) = t]} \geq \varepsilon\right] \leq \delta$$

$$\text{and } Pr\left[\frac{Pr[U(A(D \setminus S), D \setminus S, \varnothing) = t]}{Pr[U(A(D), D, S) = t]} \geq \varepsilon\right] \leq \delta$$

- Compares outputs of U always.
- Two trivial unlearners: i) retrain on $D \setminus S$, ii) output random models.

Two kinds of unlearning

- Remove my private data
- Close to DP - definitions we saw before
- **Open question -** rigorous auditing procedures to measure unlearning capability
    - Want to tie to definition
    - Gaussian unlearning?
    - Challenge: *U* knows S

- Forget an incorrect concept / remove bias [Kumaranji et al. 23]
- Seems closer to memorization.
- **Open questions:**
    - How to formally define this? Issues from defining memorization pop up.
    - We only get samples, but we want to unlearn distributions/concepts. Better algorithms?
    - Rigorous auditing methods?

What counts as unlearning?

- Do I merely want to forget the exact text, or entire plot?

ChatGPT 5

Can you tell me the plot points of the first book of Harry Potter

Here's a structured breakdown of the main plot points from *Harry Potter and the Sorcerer's Stone* (also called *Philosopher's Stone* outside the U.S.):

Opening / Ordinary World

- Baby Harry survives Voldemort's attack, left at the Dursleys' doorstep with a lightning-shaped scar.
- Raised unloved by the Dursleys, Harry lives unaware of his magical heritage.

Call to Adventure

- Letters from Hogwarts arrive (blocked by the Dursleys).
- Hagrid bursts in, tells Harry he's a wizard, and takes him to Diagon Alley to shop f school.

Entry into Magical World

- Harry boards the Hogwarts Express, meets Ron Weasley and Hermione Granger.
- Sorted into Gryffindor at the Sorting Hat ceremony.

Early Hogwarts Life

- Learns about classes, Quidditch, and rivalries (notably with Draco Malfoy and Professor Snape).
- Discovers fame from surviving Voldemort.
- Becomes youngest Seeker in a century for Gryffindor's Quidditch team.

Rising Mystery

- Troll attack on Halloween: Harry, Ron, and Hermione become friends.

ChatGPT 5

Can you create a story with these exact plot points?

Got it. You want a **new story** that isn't Harry Potter, but that follows *exactly the same plot structure / plot points* I outlined. I'll map the beats onto an original narrative while keeping them structurally parallel.

Story: *The Orb of Dawn*

Opening / Ordinary World

A boy named Kai lives with his stern aunt and uncle in a dusty border town. He's treated as a burden—fed scraps, forced to do chores, and never told much about his parents, who died mysteriously when he was an infant. The only oddity about him: a faint glowing mark on his wrist.

Call to Adventure

One day, strange sealed messages begin arriving at his house. His uncle burns them, his aunt nails the windows shut. Finally, a towering stranger named Thalos breaks in through the door and declares: *"Kai, you are heir to the Sky Guild, protectors of light. It's time you knew who you are."* He takes Kai to the bustling floating city of Aeryndor.

Entry into Magical World

Kai boards an airship bound for the Guild Academy. On the way he befriends Ryn, a clever but reckless tinkerer, and Selene, a studious girl with a sharp tongue. At the Academy, during initiation, the choosing crystal assigns him to the Dawn Wing—the guild of adventurers and explorers.

Early Academy Life

Kai adjusts to his new classes—alchemy, sky combat, rune studies. He struggles at first but discovers he has a knack for windcraft. He also becomes the youngest person ever chosen as a sky-rider in the upcoming aerial games. Meanwhile, rival student Darian mocks him constantly, and Professor Varros, with his piercing eyes, seems to hate him for no reason.

Rising Mystery

- What is the right "unit of forgetting"?
- Formalize memorization at higher abstractions:

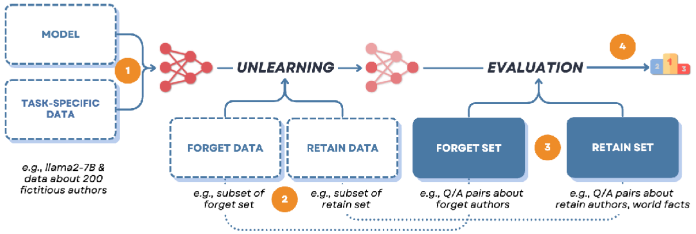
  Exact text of HP ->
  Reworded HP ->
  Plot points of HP ->
  Story "structure"

Auditing Unlearning Methods?

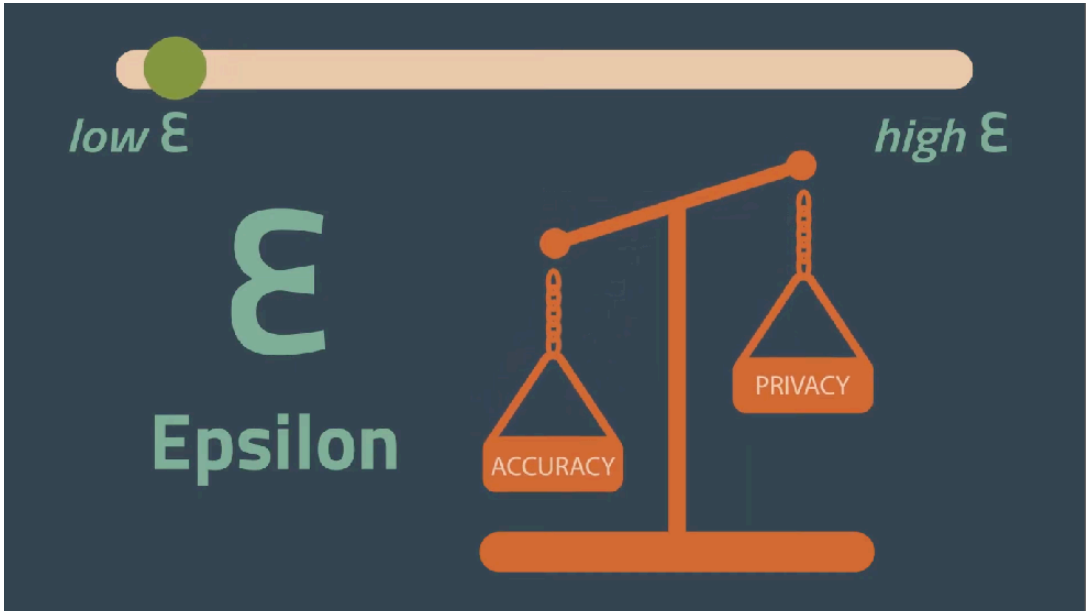
- Results very sensitive to specific prompts
- Experiment setup makes overfitting to the benchmark inevitable. Similar to LLM Jailbreak - everyone will account for substitute secrets.
- **Open question:** Really need better auditing methods.

Position: LLM Unlearning Benchmarks are Weak Measures of Progress

Pratiksha Thaker, Shengyuan Hu, Neil Kale, Yash Maurya, Zhiwei Steven Wu, Virginia Smith
Carnegie Mellon University
Pittsburgh, PA
{pthaker, shengyua, nkale, ymaurya, zstevenwu, smithv}@andrew.cmu.edu

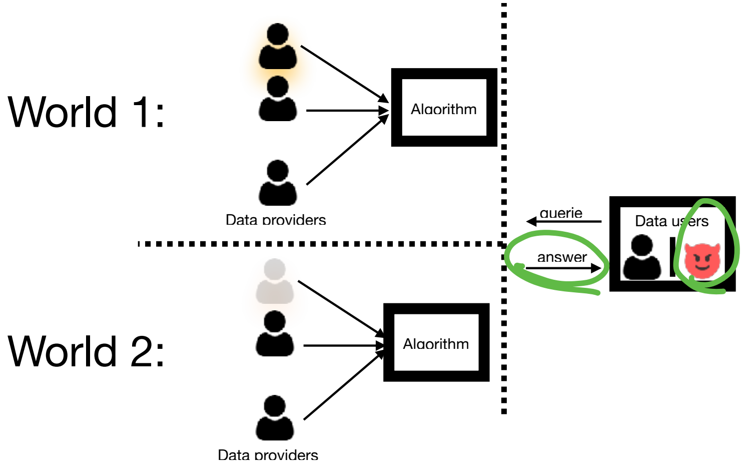


Local Differential Privacy



Central Differential Privacy

- Previously: how well can the adversary guess which world I am in based on the **output.**

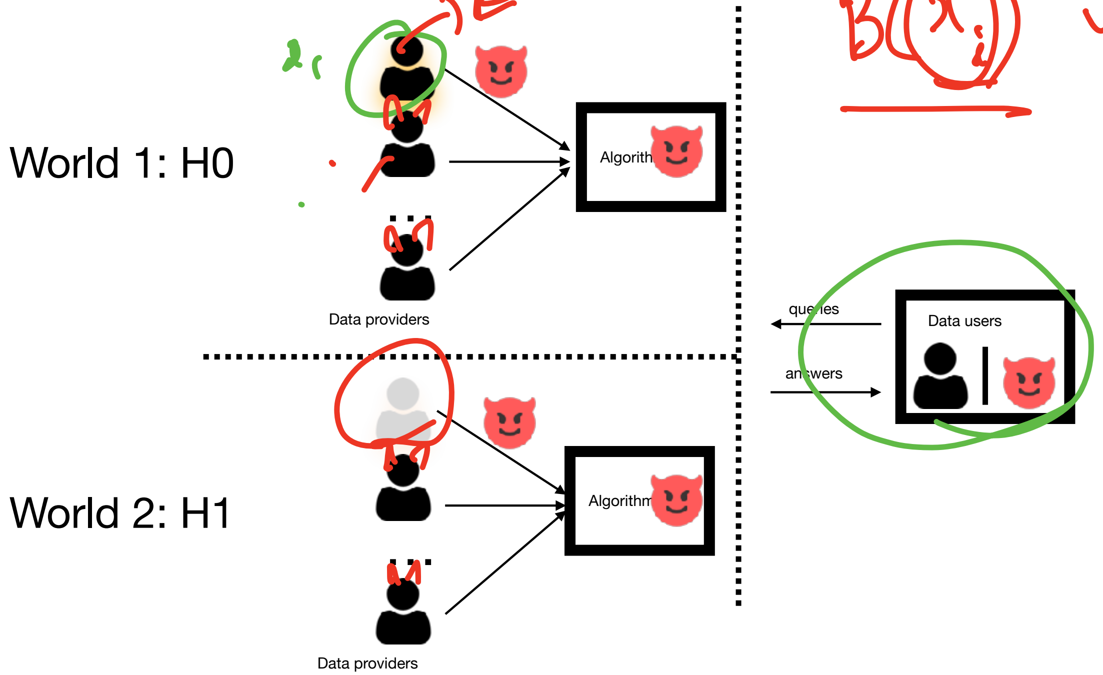


Local Differential Privacy

$A(D)$ vs. $A(D \setminus \{x\})$

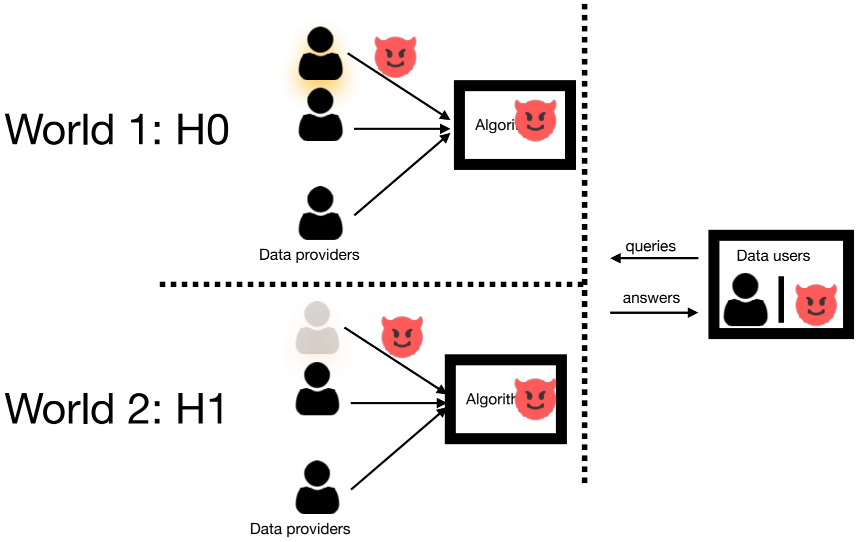
- New: how well can the adversary guess which world I am by looking at my communication

$B(x_i)$ vs. $B(x_i')$



Local Differential Privacy

- New: how well can the adversary guess which world I am by looking at my communication
- No need to trust
  - central server
  - or communication network
- Only trust yourself



Local Differential Privacy

Local differential privacy [Kasiviswanathan et al. 2011]

Let $\pi_i(v)$ indicate the user i's output after looking at datapoint v.
Then, π_i satisfies ε-LDP if

$$\frac{Pr[\pi_i(v) = y]}{Pr[\pi_i(u) = y]} \leq \varepsilon \text{ for all } y, u, v \text{ and all users } i.$$

Approximate Local Differential Privacy

(ε, δ) Local Differential Privacy

Let $\pi_i(v)$ indicate the user i's output after looking at datapoint v.
Then, π_i satisfies (ε, δ)-LDP if for a randomly sampled $t \sim \pi_i(v)$

$$Pr\left[\frac{Pr[\pi_i(v) = y]}{Pr[\pi_i(u) = y]} \geq \varepsilon\right] \leq \delta \text{ for all } y, u, v \text{ and users } i.$$

Central-DP Binary Mean Estimation

Utility under central DP

- We have n i.i.d samples $(x_1, \ldots, x_n)$ where $x_i \in \{0,1\}$.
- Estimate mean as $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i + \mathrm{Lap}(\Delta/\varepsilon)$. Sensitivity is $\Delta = 1/n$?

Expected square error

- Net error is "statistical error" + "privacy error" $= \frac{1}{n} + \frac{2}{n^2\varepsilon^2}$.
- Privacy is free as long as $\varepsilon \leq 1/\sqrt{n}$.

$2\,N(0,1) \equiv N(0, 2^2)$

output $\frac{1}{n}\sum x_i + Lap\left(\frac{1}{n\varepsilon}\right)$

Local-DP Binary Mean Estimation

answer: $\sum \frac{1}{n} x_i + Lap\left(\frac{1}{n\varepsilon}\right)$

Utility under local DP

- We have n users each with an i.i.d sample $x_i \in \{0,1\}$.
- User i communicates $\left(x_i + \text{Lap}_i(\Delta/\varepsilon)\right)$. What is local sensitivity?
    - Here, we have $\Delta = 1$!
- We compute the average $\frac{1}{n}\sum_{i=1}^{n}\left(x_i + \text{Lap}_i(\Delta/\varepsilon)\right)$.
- Net error is "statistical error" + "privacy error" $= \frac{1}{n} + \frac{2}{n\varepsilon^2}$.

  No privacy for free
- ~~Now can only tolerate $\varepsilon \geq n^{-1/4}$.~~

$\frac{1}{n} + \frac{2}{n^2\varepsilon^2}$

under central DP

Local-DP Unbounded Mean Estimation

Utility under local DP

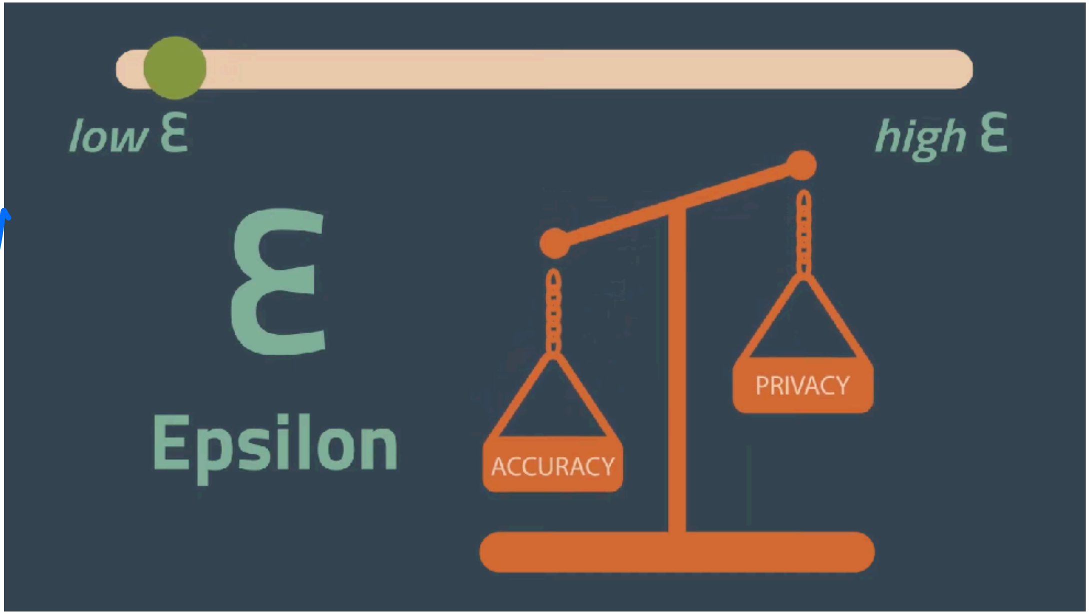
- We have n users each with an i.i.d sample x_i satisfying $E[x_i^2] \leq \sigma^2$.
- User i communicates $\left(\text{clip}_\tau(x_i) + \text{Lap}_i(2\tau/\varepsilon)\right)$.
- We compute the average $\frac{1}{n}\sum_{i=1}^n \left(\text{clip}_\tau(x_i) + \text{Lap}_i(2\tau/\varepsilon)\right)$.
- Net error is ≈ "statistical error" + "clipping bias" + "privacy error"
  bias
  Lap noise
    - $= \frac{\sigma^2}{n} + \frac{2\sigma^4}{\tau^2} + \frac{16\tau^2}{n\varepsilon^2}$. By picking the optimal τ,
    - $= O\left(\frac{\sigma^2}{n} + \frac{\sigma^2}{\sqrt{n\varepsilon}}\right)$. Privacy is never "free" - goes from $1/n$ to $1/\sqrt{n}$. :(
    - Compare to central-DP $= O\left(\frac{\sigma^2}{n} + \frac{\sigma^2}{n\varepsilon}\right)$ where constant ε didn't hurt.

Local-DP Strengths & Weakness

- Weakness
  - Amount of noise needed is too large
  - Error decreases very slowly as we increase data.
- Strengths
  - No need to trust the implementation, infrastructure, etc.
  - No problem if server gets hacked or server leaks your data.
  - Stronger definition of privacy / security.
- Best of both worlds? Yes! With *crypto* or *TEEs* or *federated learning*.

Federated Learning

/distributed/
decentralized



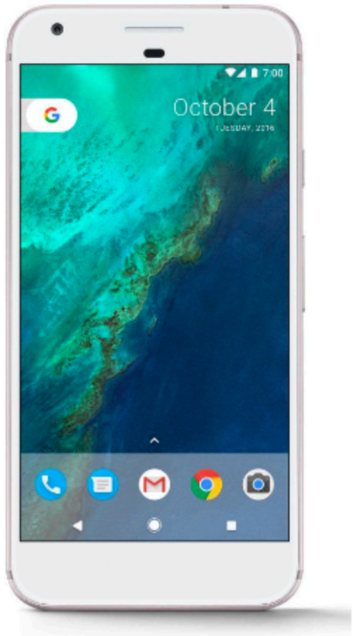
Heavily based on NeurIPS '20 FL tutorial

Data at the Edge

Billions of phones & IoT devices constantly generate data

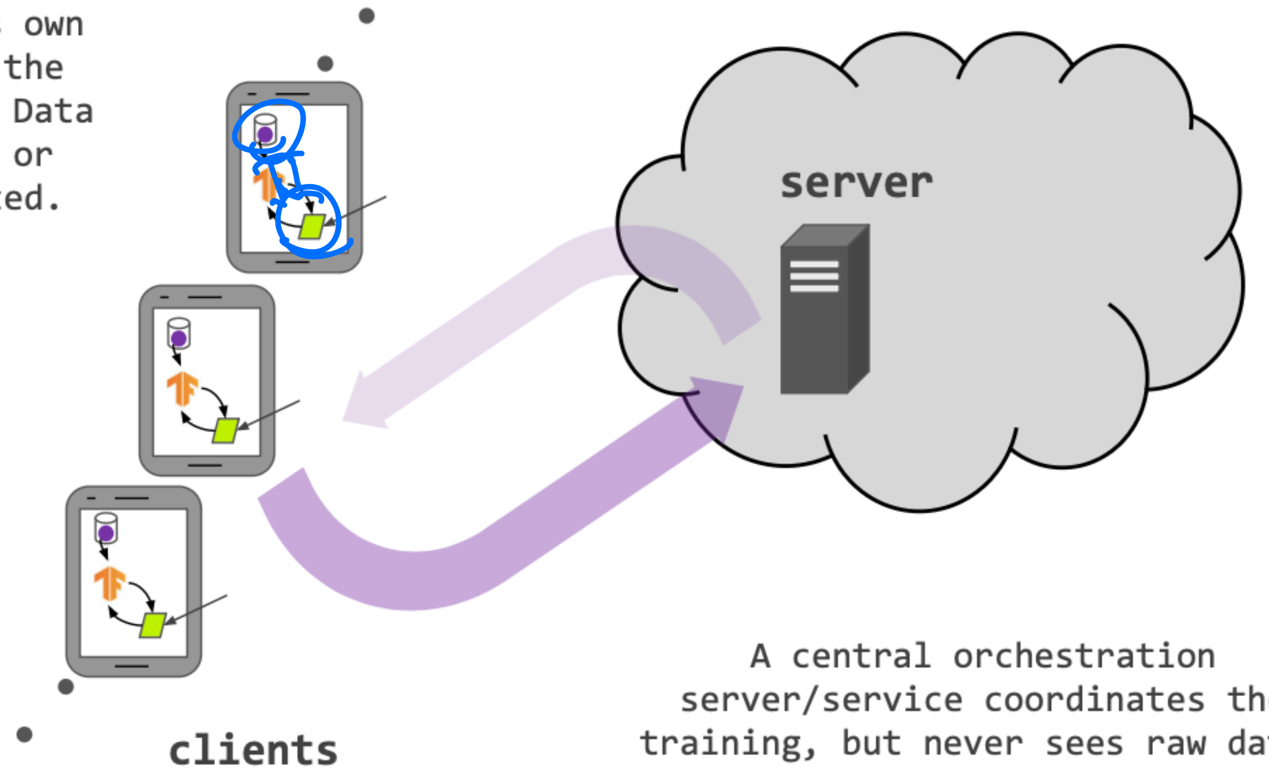
Data enables better products and smarter models

Data processing is moving on device:
- Improved latency
- Works offline
- Better battery life
- Privacy advantages



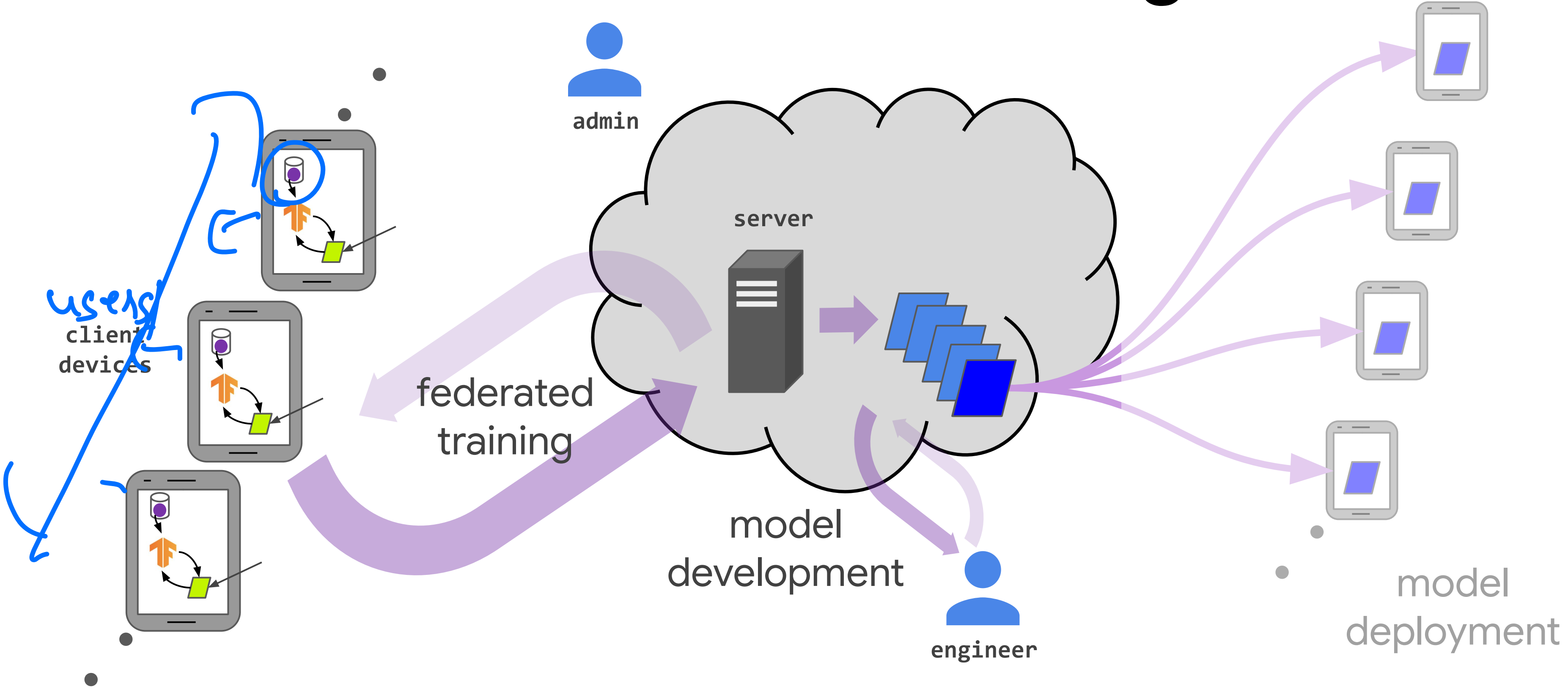
Federated Learning

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-device Federated Learning



Cross-device Federated Learning

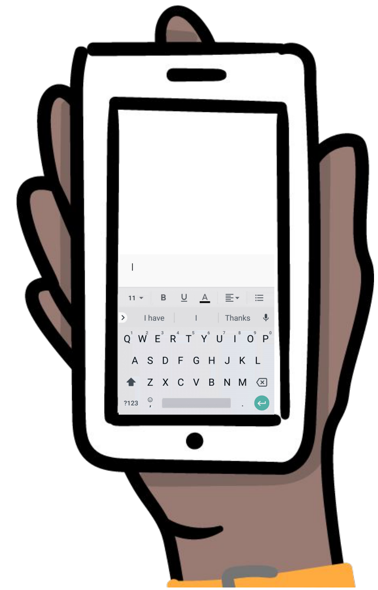
What makes a good application?

- 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

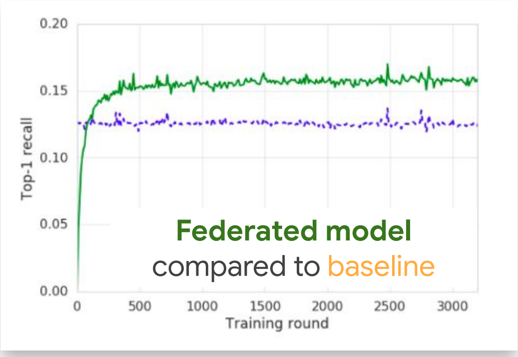


Federated RNN (compared to prior n-gram model):

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

+ Meta

AR/VR headset



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.

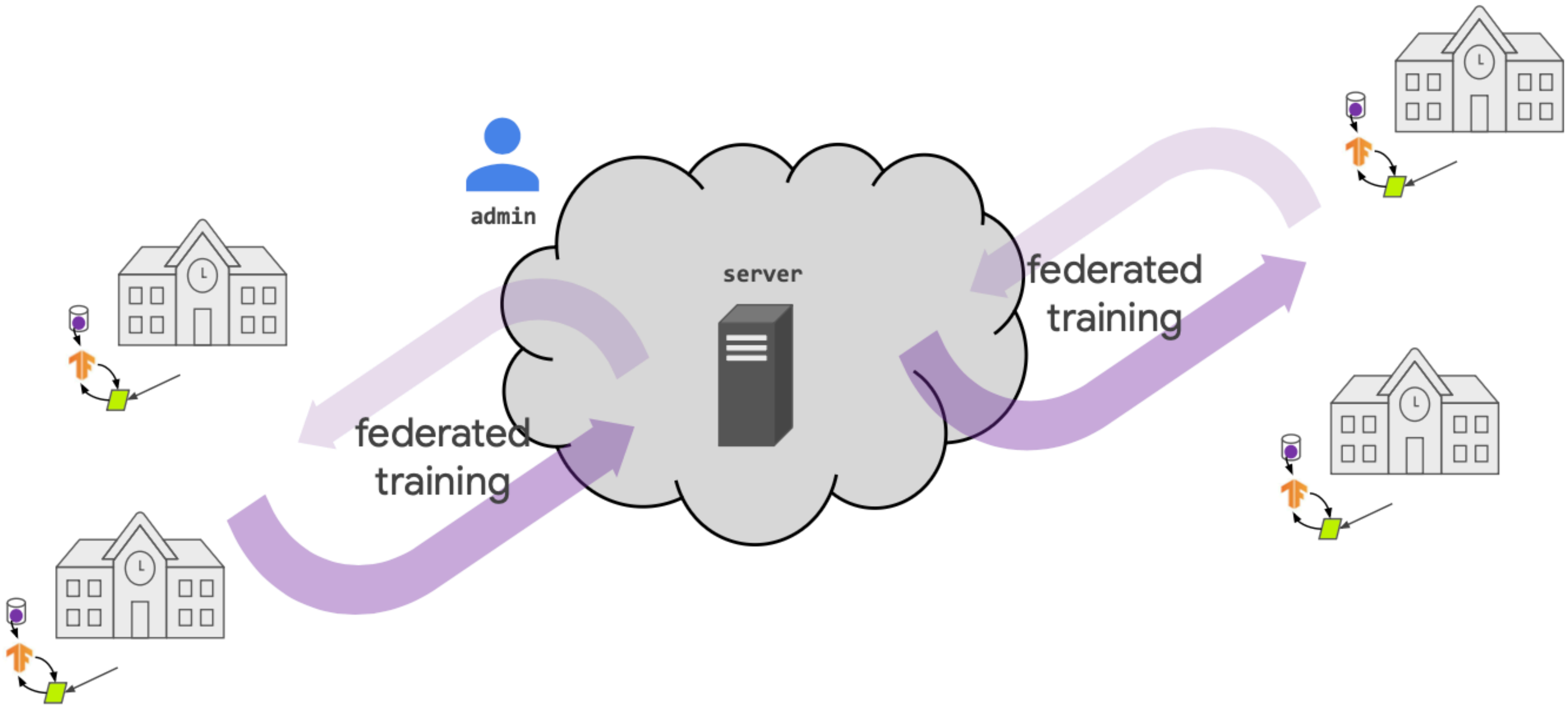
by Karen Hao

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/

Cross-silo Federated Learning



Cross-silo federated learning from NVIDIA

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."

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

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 Transform 2020

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-
[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-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."

Is Machine Learning for Tumor Research at Odds With Patient Privacy? Not With Federated Learning, Intel Says

May 13, 2020 by Tyler Charboneau

Intel partners with Penn M... to develop brain tumor cla... with federated learning

Bio-IT World

Intel, Penn Medicine Launch Federated Le...

By Allison Proffitt

[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-federa
[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

NHS England awards £330m Federated Data Platform contract to Palantir

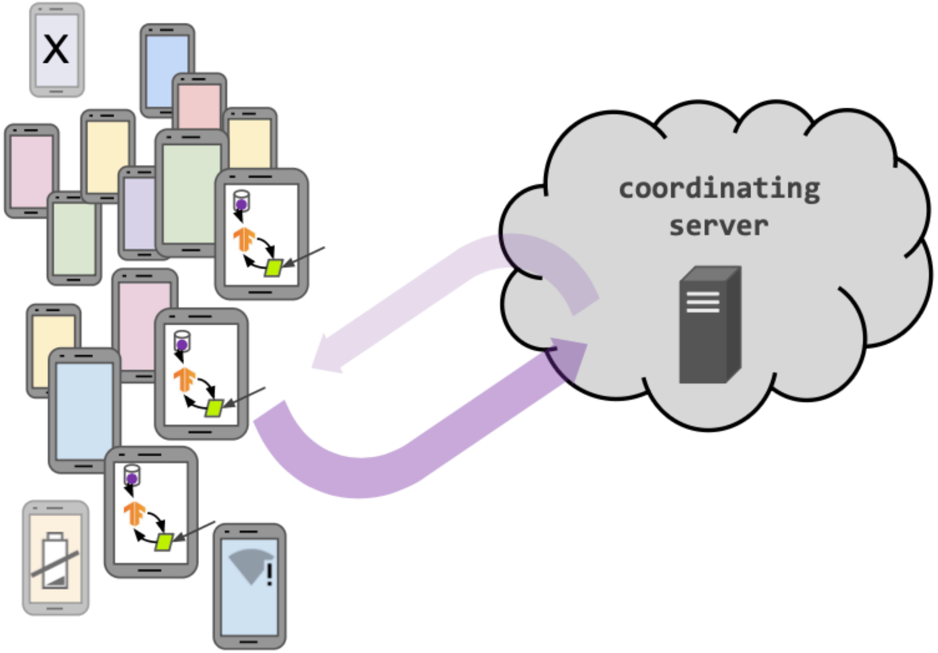
NEWS

21 November 2023

Palantir

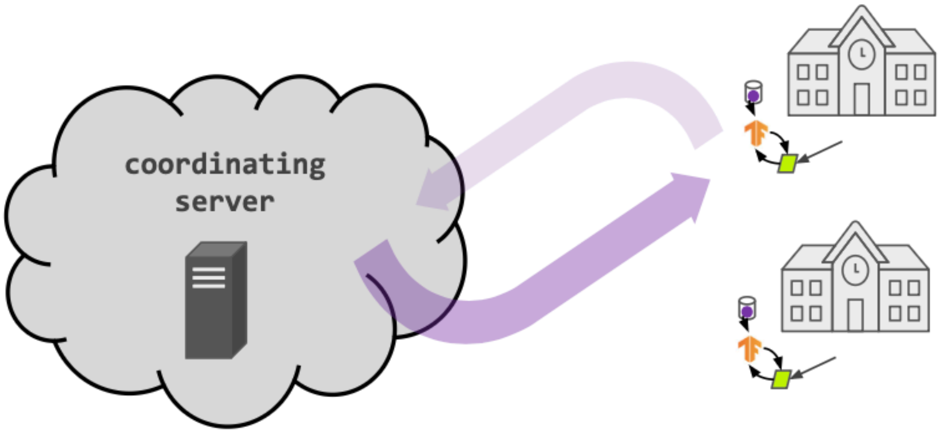
Cross-device federated learning

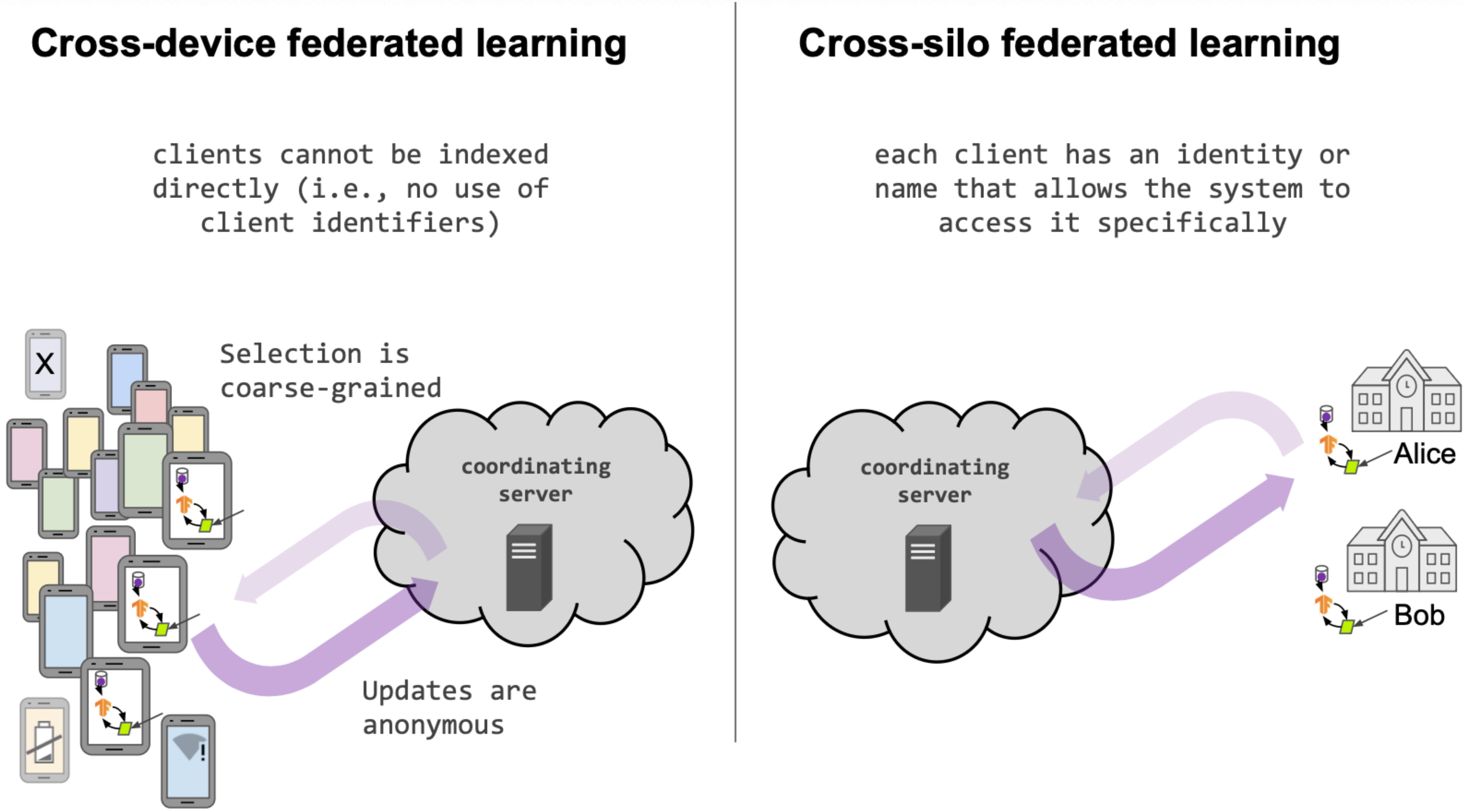
millions of intermittently available client devices



Cross-silo federated learning

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

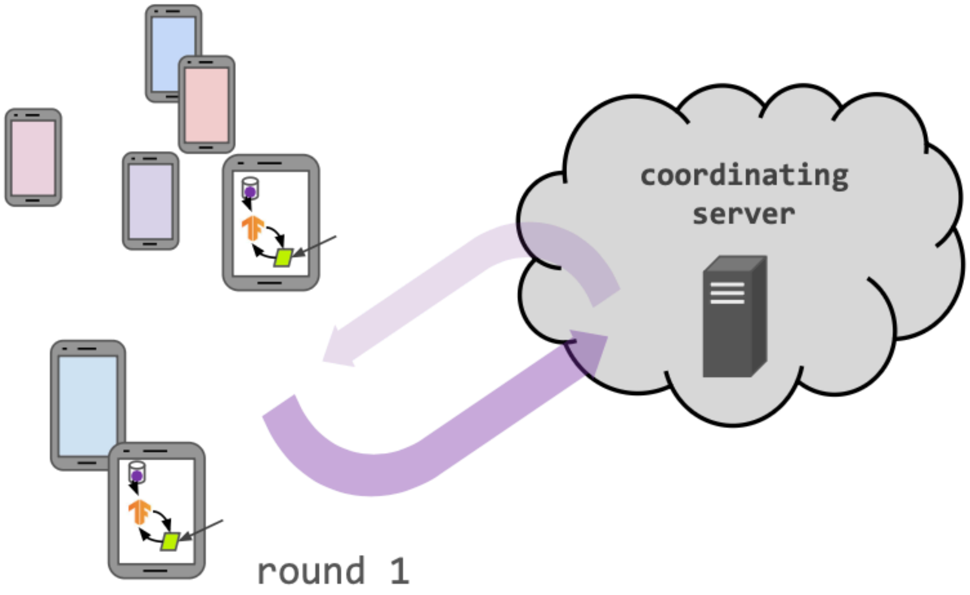



Cross-device federated learning
clients cannot be indexed directly (i.e., no use of client identifiers)
X
Selection is coarse-grained
coordinating server
Updates are anonymous
Cross-silo federated learning
each client has an identity or name that allows the system to access it specifically
coordinating server
Alice
Bob

Cross-device federated learning

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

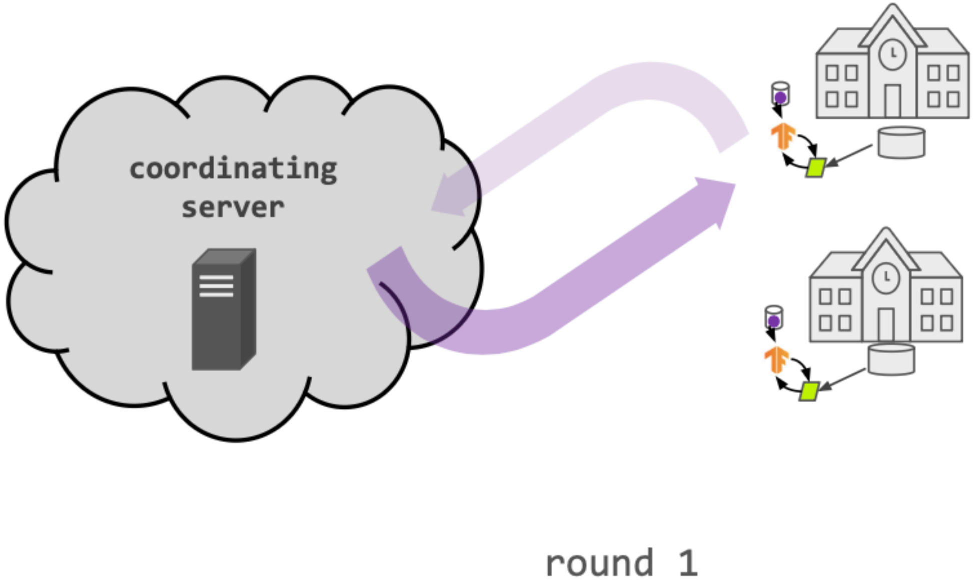
Large population => most clients only participate once.



Cross-silo federated learning

Most clients participate in every round.

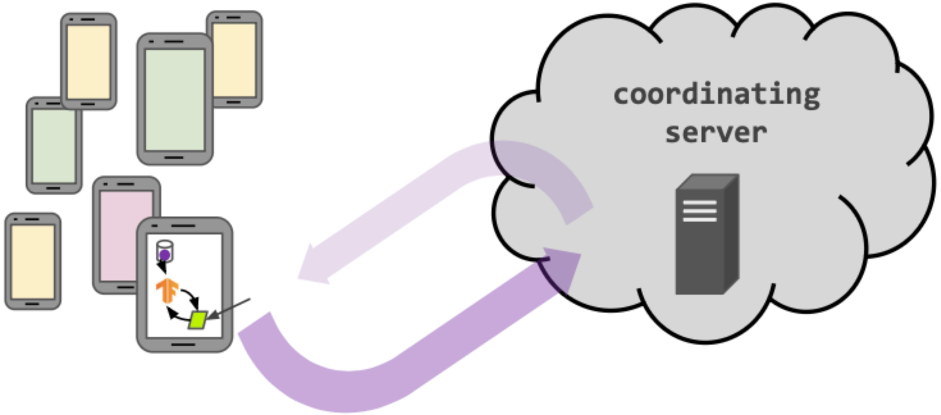
Clients can run algorithms that maintain local state across rounds.



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.

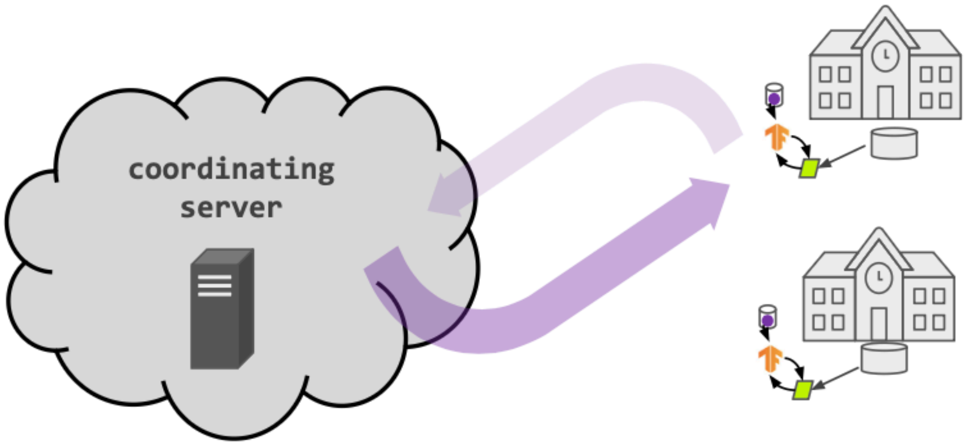


round 2
(completely new set of devices participate)

Cross-silo federated learning

Most clients participate in every round.

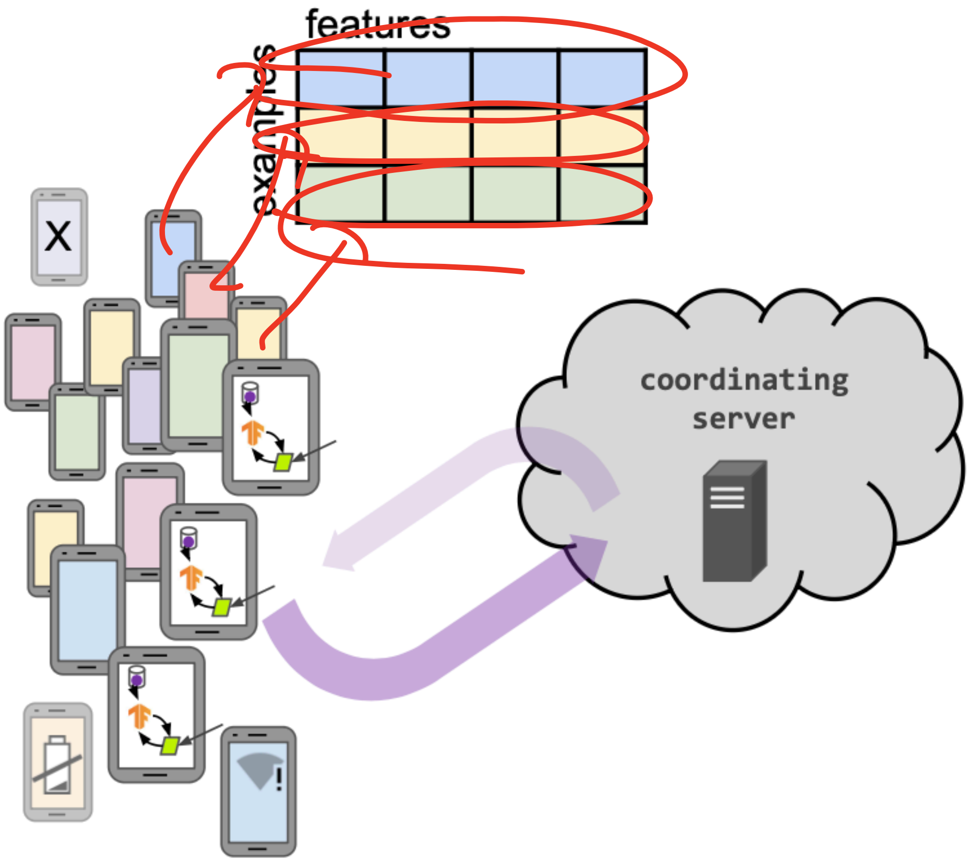
Clients can run algorithms that maintain local state across rounds.



round 2
(same clients)

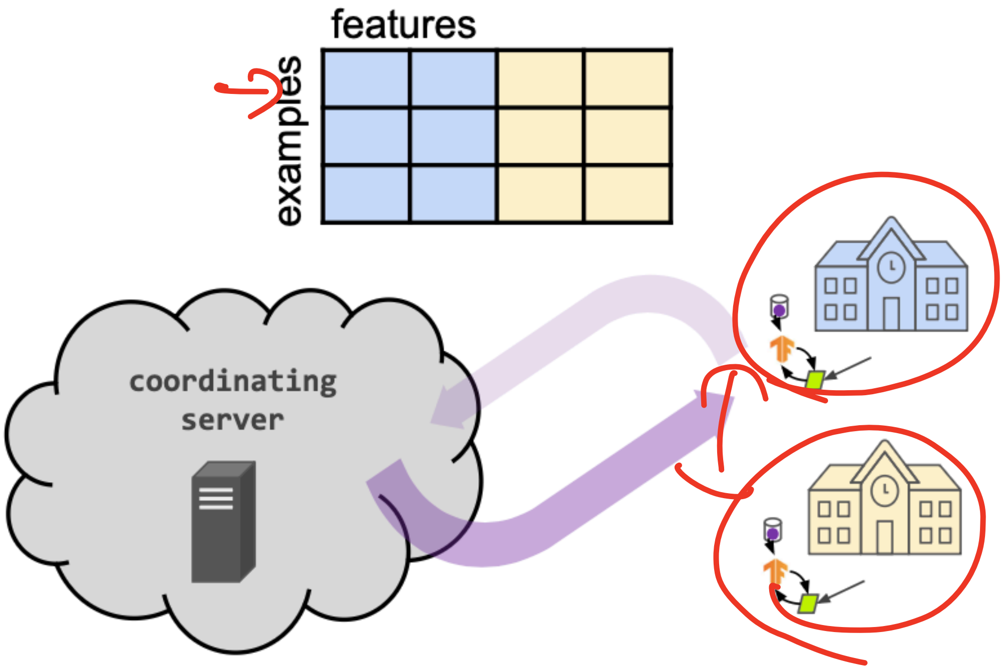
Cross-device federated learning

horizontally partitioned data



Cross-silo federated learning

horizontal or
vertically partitioned data

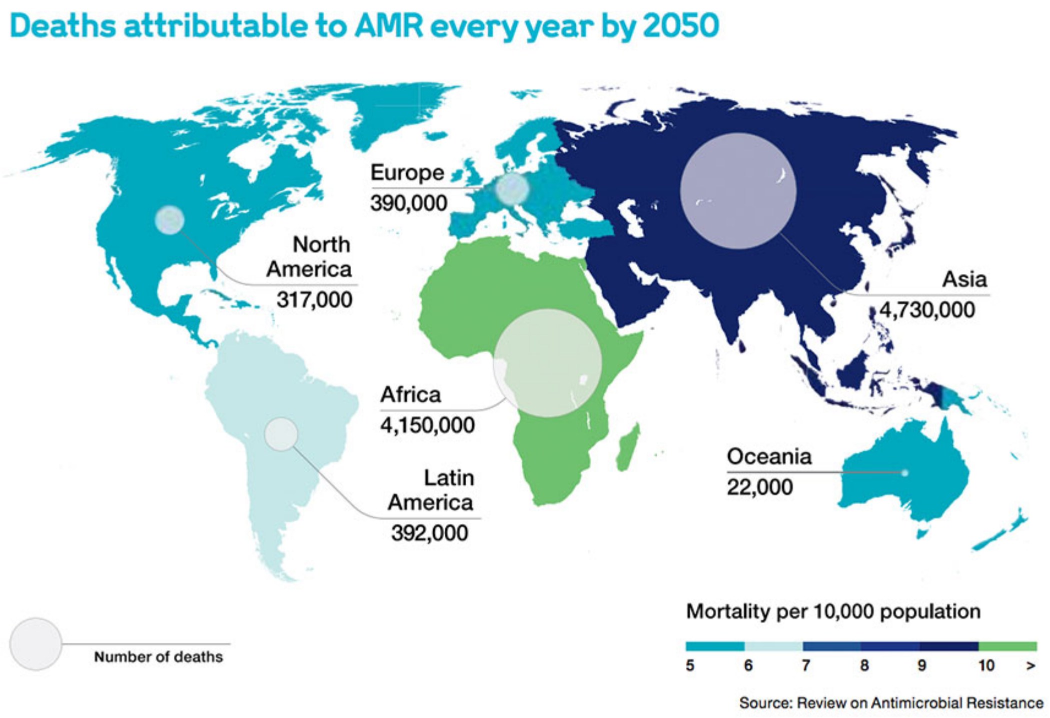


Case studies

Antimicrobial Resistance

"By 2050, 10 millions deaths per year due to antimicrobial resistance."

- [WHO 2014]



MEDECINS SANS FRONTIERES

Antimicrobial Resistance

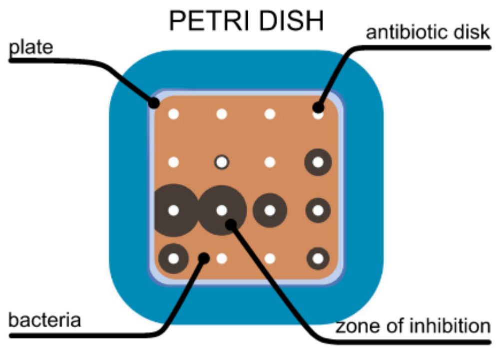
Requires pathology experts :(

MEDECINS SANS FRONTIERES

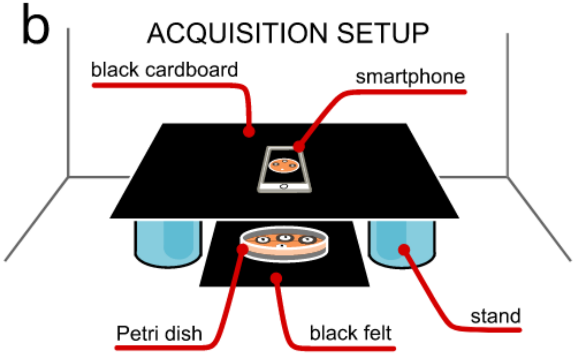
Antibiogo

Use on-mobile ML to automate the process.

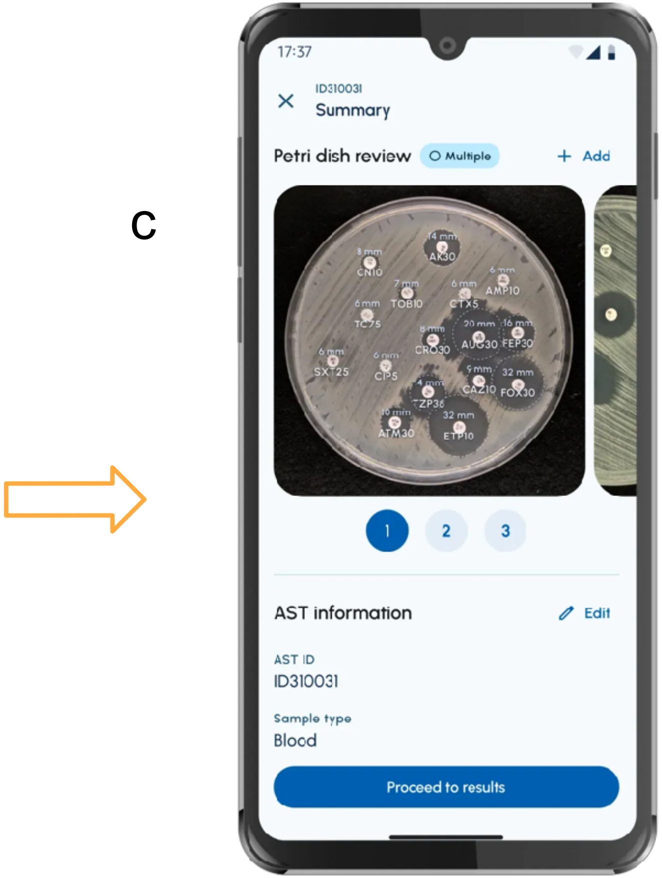
a



b



c



MEDECINS SANS FRONTIERES

Antibiogo Goal

- Multi-continent collaboration
  15 countries in Africa and Asia

- Continuously in real-time
  - improve ML models,
  - epidemiological monitoring

MEDECINS SANS FRONTIERES

Antibiogo Challenges

Continual learning across 15 countries, but

1. **Privacy** data can't leave the device.

Local DP

Resilience

2. Bad network, low end phones.

3. Very noisy data

MEDECINS SANS FRONTIERES

Shortages in Cancer Expertise

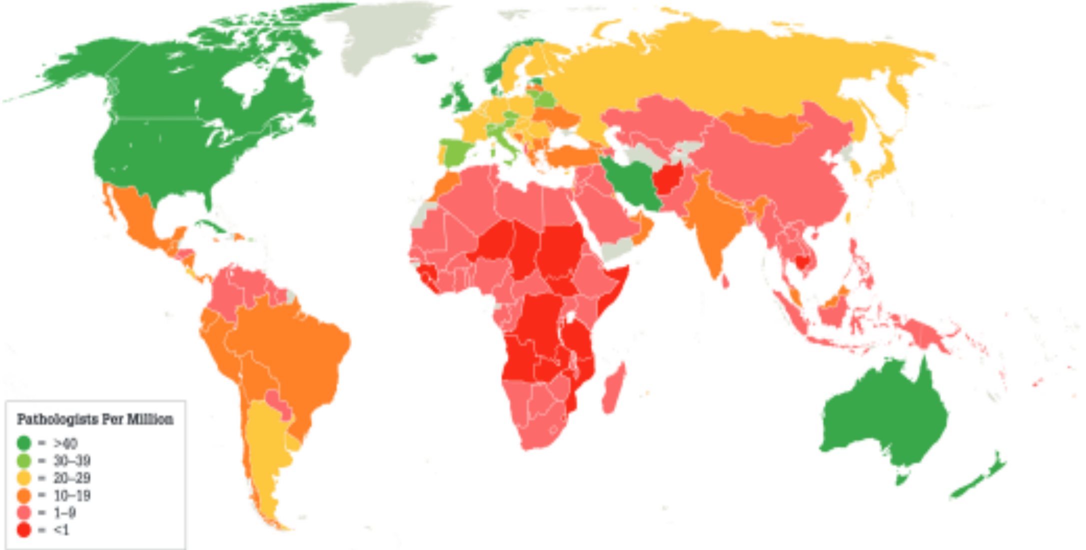
Global expertise per capita is falling, leading to deadly shortages.

"Cancer patients face life-threatening delays due to lack of staff, say UK radiologists."

— The Guardian, 8 Jun 2023

Cancer patients face worsening treatment delays due to lack of staff, report finds

— Sky News, 8 June 2023



Cancer Registry of Norway

Hanna, Timothy P., et al. "Mortality due to cancer treatment delay: systematic review and meta-analysis." bmj 371 (2020).

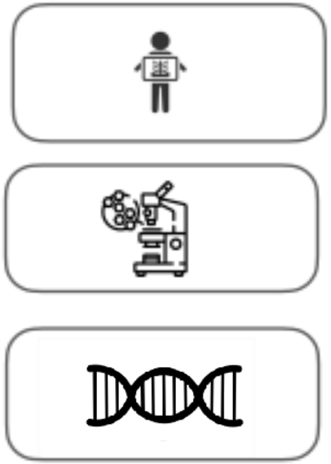
AI for Cancer

Sensitive patient data

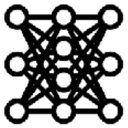
scans

tests

genome



ML models



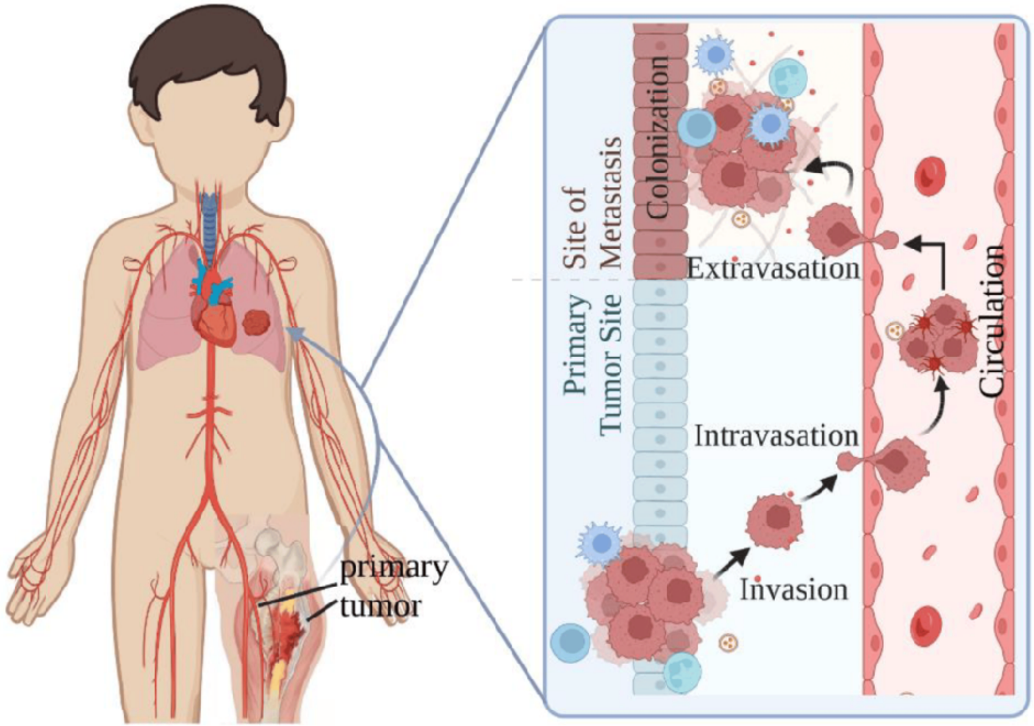
Tasks
Classify the cancer grade/stages
Segment the cancer region in radiographs
Predict the risk of developing a cancer type

Cancer Registry of Norway

Data Scarcity

Sarcoma is 1% of cancer diagnosis.
In 2022, 562 cases in Norway.

Same problem in low-middle income countries.



Cancer Registry of Norway

Data Scarcity

Ethics and Justice, Healthcare, Machine Learning

The Geographic Bias in Medical AI Tools

Patient data from just three states trains most AI diagnostic tools.

Sep 21, 2020 | Shana Lynch

Bias in, bias out: Underreporting and underrepresentation of diverse skin types in machine learning research for skin cancer detection—A scoping review

Lisa N. Guo, MD • Michelle S. Lee, BA • Bina Kassamali, BA • Carol Mita, MSLIS • Vinod E. Nambudiri, MD, MBA

FDA NEWS RELEASE

FDA Takes Important Steps to Increase Racial and Ethnic Diversity in Clinical Trials

Agency's Focus on Inclusion in Trials for All Medical Products Aligns with Biden Administration's Cancer Moonshot Goal of Addressing Inequities and Beyond

Share | Post | LinkedIn | Email | Print

For Immediate Release: April 13, 2022

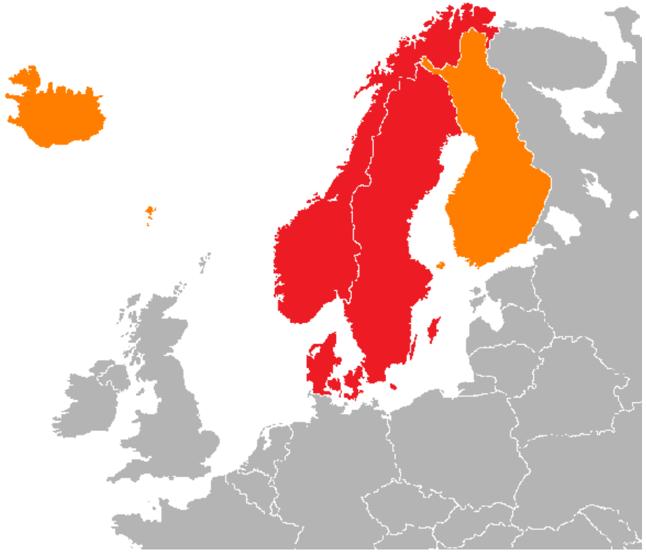
Cancer Registry of Norway

NORDCAN
Association of the Nordic Cancer Registries

NORDIC CANCER UNION

Position paper [Karimireddy et al. FMEC 2023]

- 7+ countries, 21 registries collaboration to train ML models, monitoring. +Taiwan
- Strict **privacy** regulations to share data, especially genomic. Local DP
- But data collected is extremely **heterogeneous**.
- Also, **strategic** concerns like fairness and accountability.



The Danish Cancer Registry | Finnish Cancer Registry | Icelandic Cancer Registry | The Cancer Registry of Norway | The Swedish Cancer Registry

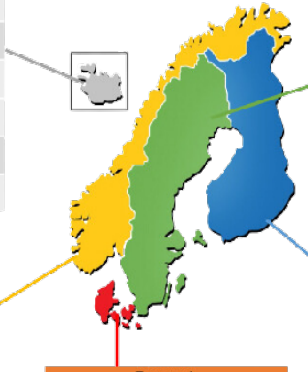
The Faroe Islands Cancer Registry | The National Board of Health of Greenland

Heterogenous data

- Each site will have different demographics and populations

Iceland	
Population size (million)	0.36
Annual birth rate (per 1000)	12
Average number of births (per woman)	1.7
Life expectancy from birth (years), female	85
Life expectancy from birth (years), male	81
Annual death rate (per 1000)	6.4
Persons aged 65 years or older (%)	15

Sweden	
Population size (million)	10
Annual birth rate (per 1000)	12
Average number of births (per woman)	1.8
Life expectancy from birth (years), female	84
Life expectancy from birth (years), male	81
Annual death rate (per 1000)	9.1
Persons aged 65 years or older (%)	20

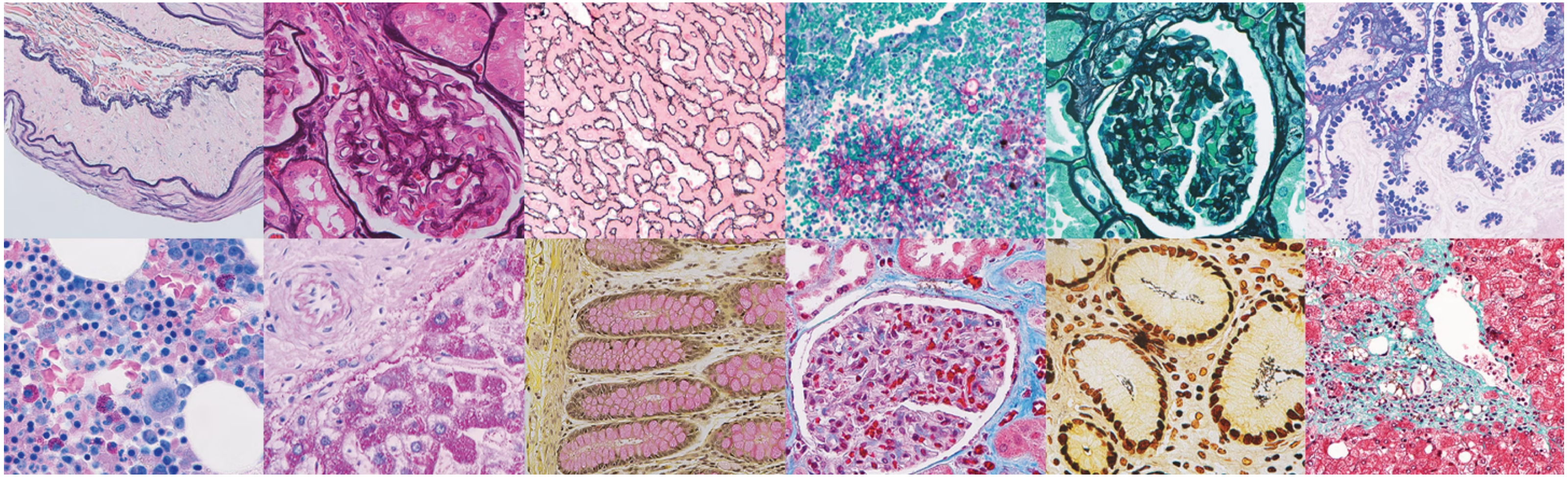


Norway	
Population size (million)	5.3
Annual birth rate (per 1000)	11
Average number of births (per woman)	1.6
Life expectancy from birth (years), female	85
Life expectancy from birth (years), male	81
Annual death rate (per 1000)	7.7
Persons aged 65 years or older (%)	17

Denmark	
Population size (million)	5.8
Annual birth rate (per 1000)	11
Average number of births (per woman)	1.7
Life expectancy from birth (years), female	83
Life expectancy from birth (years), male	79
Annual death rate (per 1000)	9.5
Persons aged 65 years or older (%)	20

Finland	
Population size (million)	5.5
Annual birth rate (per 1000)	8.6
Average number of births (per woman)	1.4
Life expectancy from birth (years), female	85
Life expectancy from birth (years), male	79
Annual death rate (per 1000)	9.9
Persons aged 65 years or older (%)	22

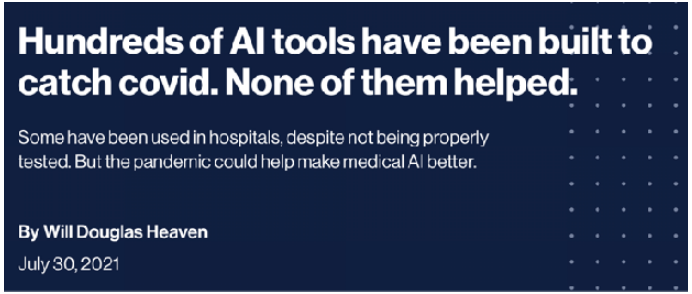
Heterogenous data



- Each Hospital collects data differently:
  - Different technologies (MRI machines, CT scanners, staining .)
  - Different procedures (is covid +ve: self-reported, PCR test, antigen, CT scan, doctor's diagnosis?)

Heterogenous data

- When multiple data sources are mixed, models learnt shortcuts. E.g.
    - Some non-covid data came from children’s hospital. Model learnt to recognize kids, not covid.
    - Different hospitals treat different severities. Models recognized hospitals by the text font in the scans.

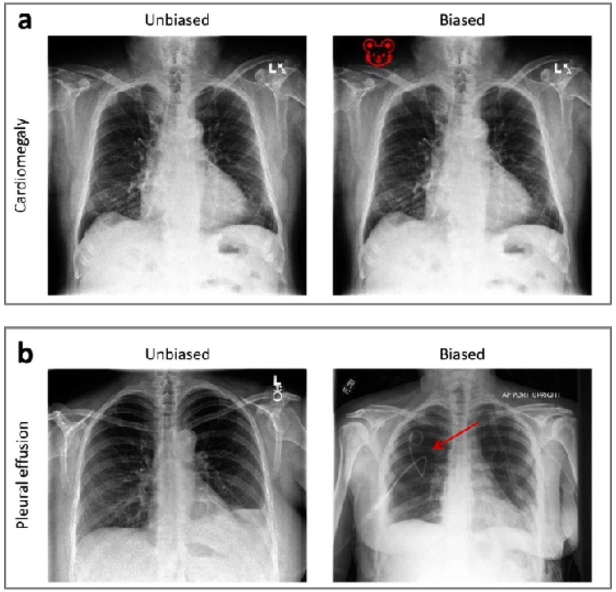


Hundreds of AI tools have been built to catch covid. None of them helped.

Some have been used in hospitals, despite not being properly tested. But the pandemic could help make medical AI better.

By Will Douglas Heaven

July 30, 2021

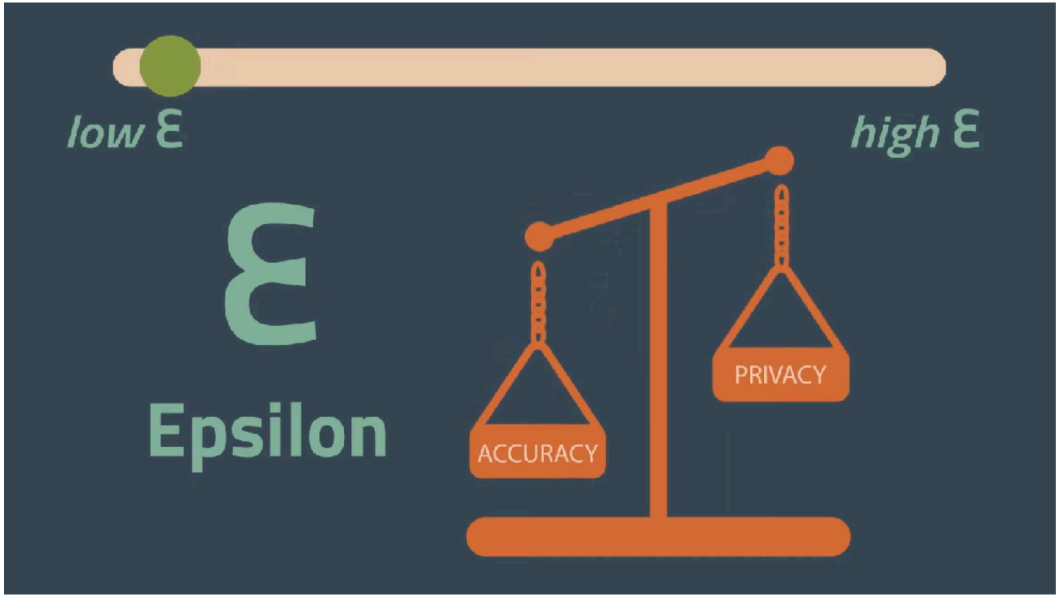


[Klaudia et al. 2024]

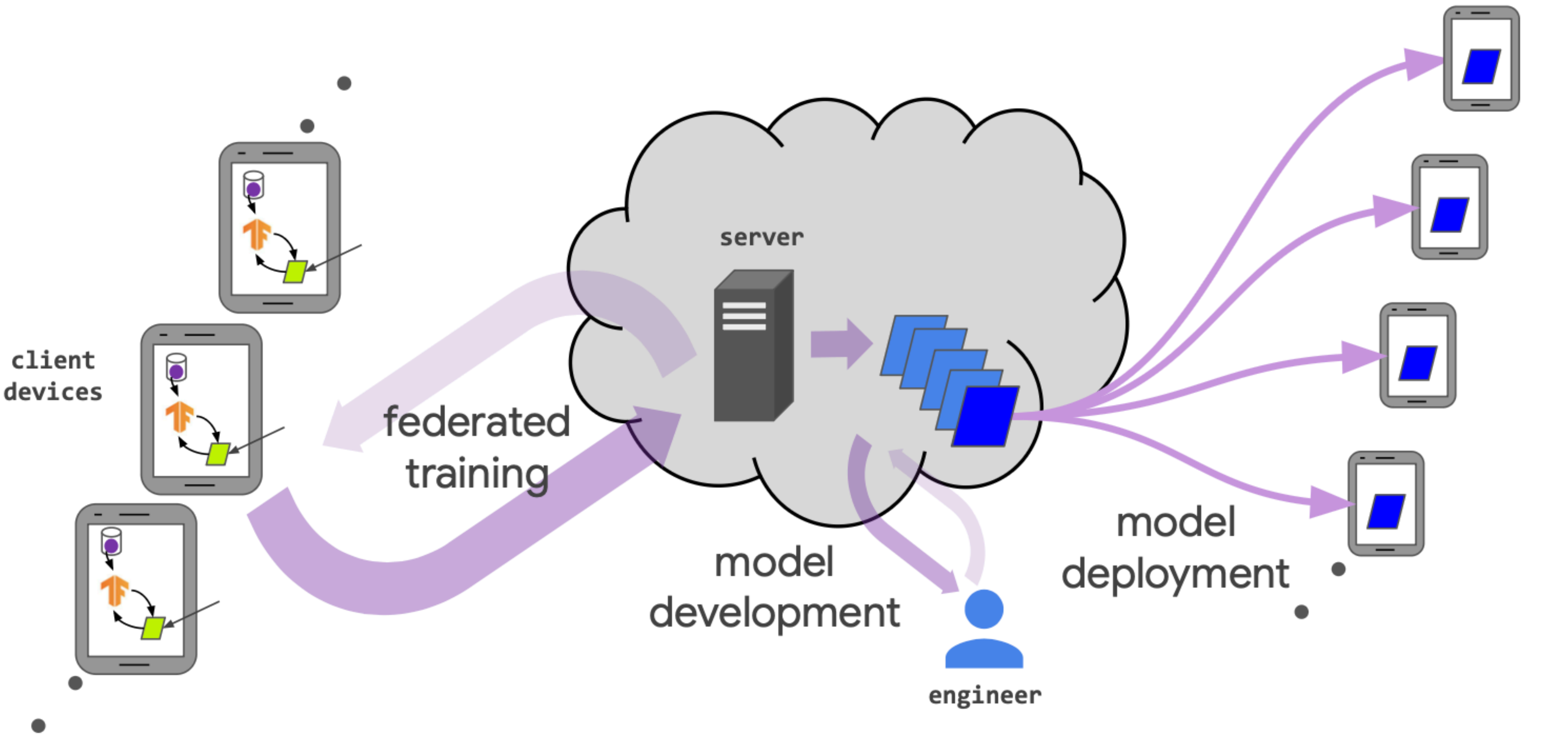
Cross-device and Cross-silo FL

	Cross-device FL	Cross-silo FL
Challenges	Data Privacy	Data privacy
	Scale + Resilience: large and unreliable networks	
	Noisy unreliable data	Heterogenous data and population
		Strategic concerns like fairness and accountability

Privacy in Federated Learning

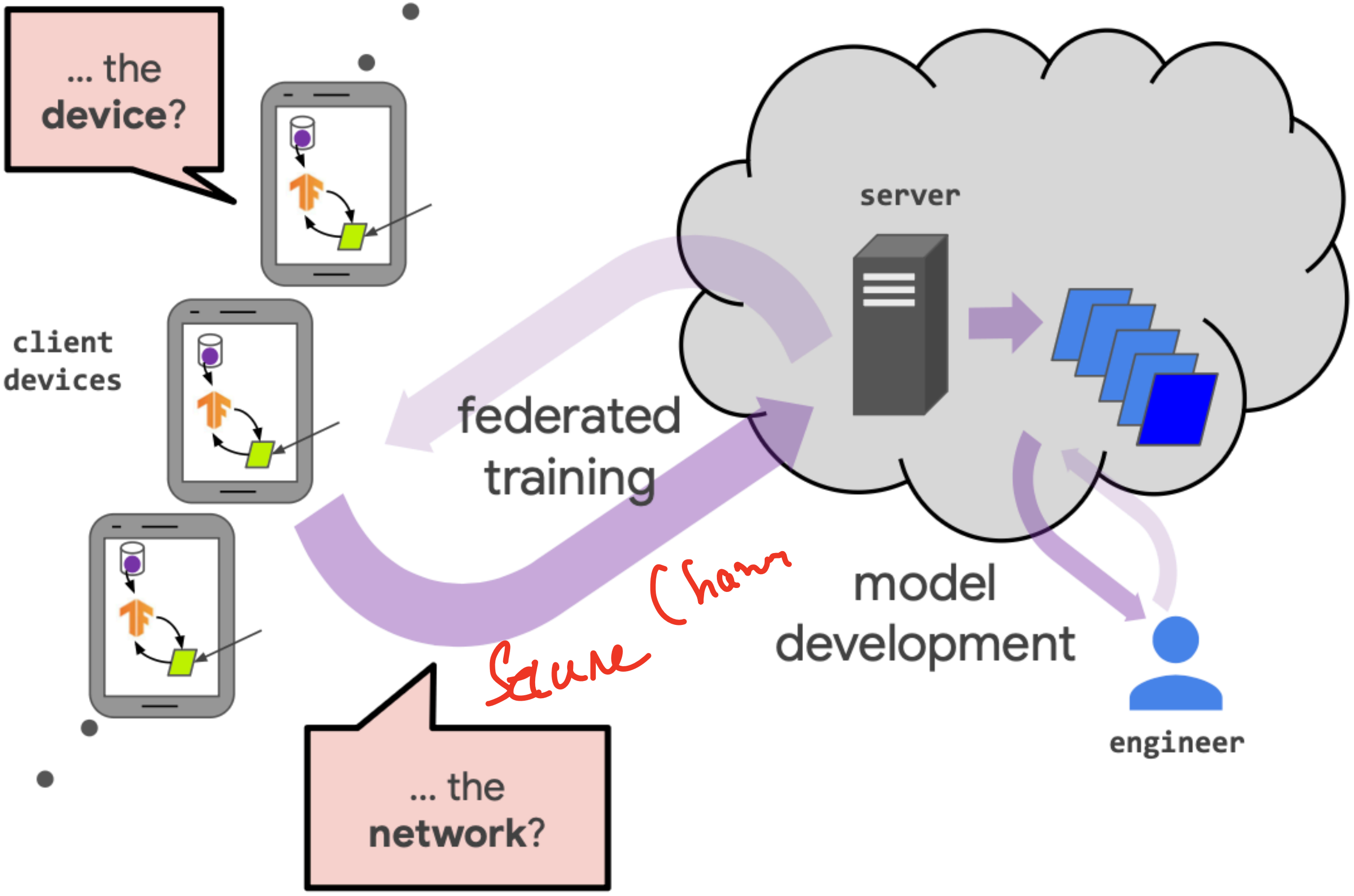


What private information might an actor learn?



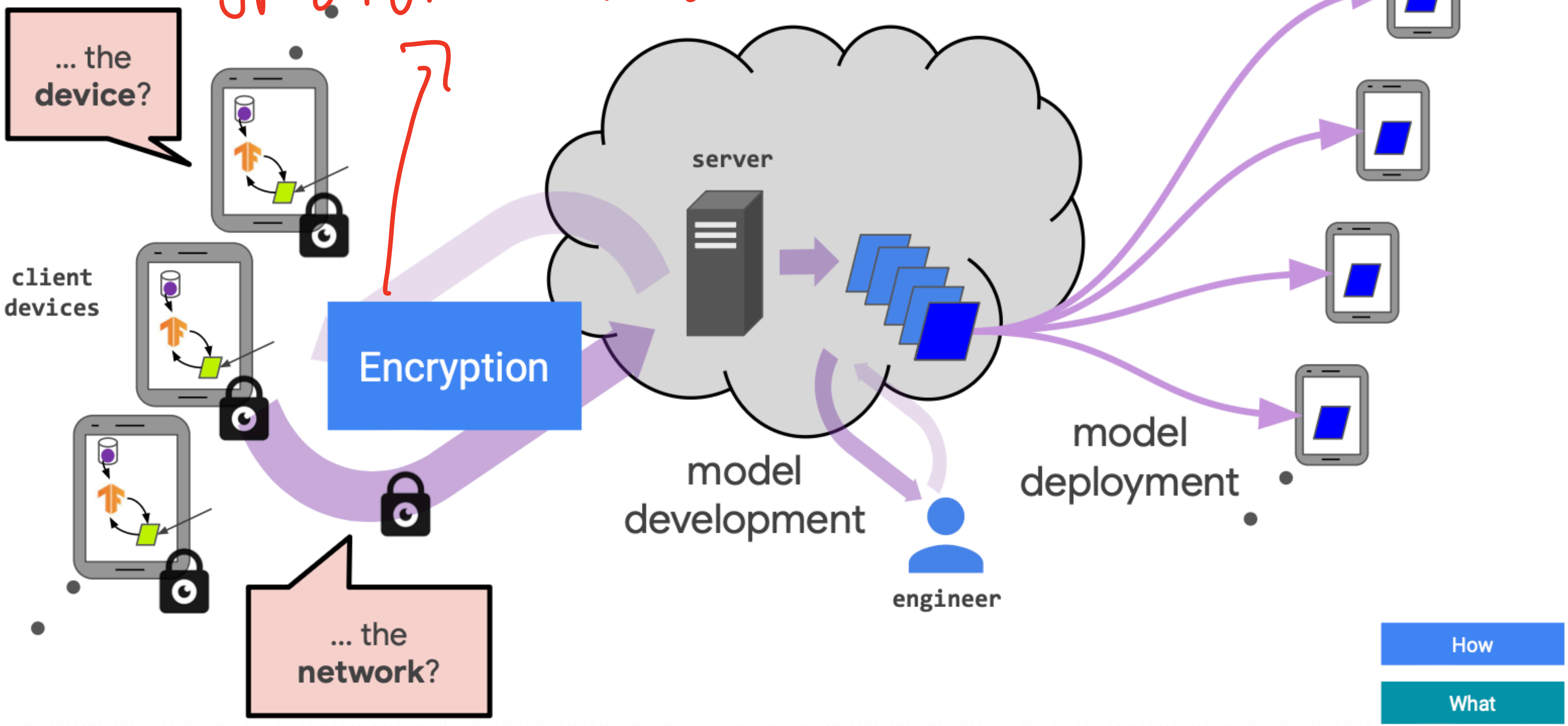
What private information might an actor learn with access to ...

D vs. D \ {x}

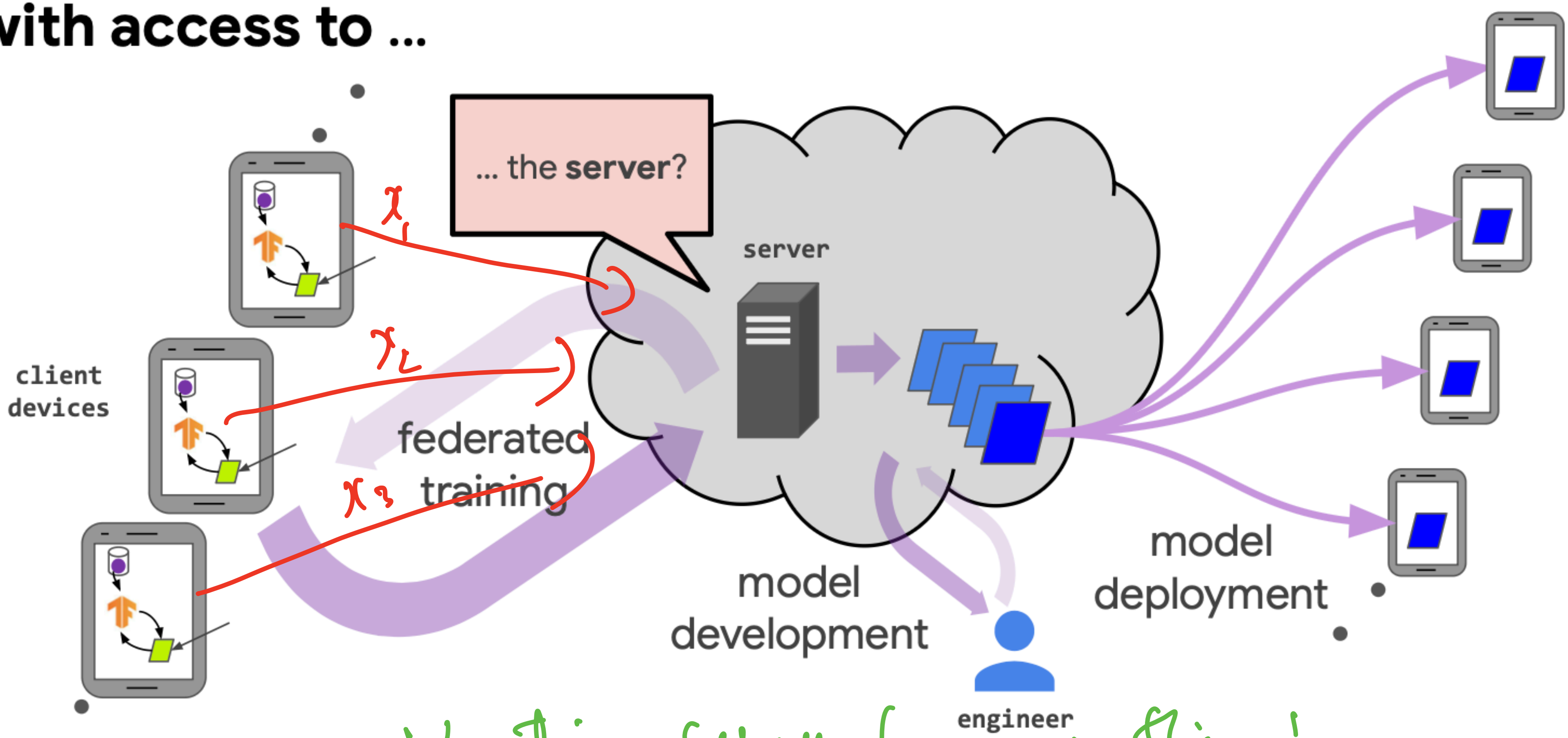


Encryption, at rest and on the wire

Cryptography: multi-party computation

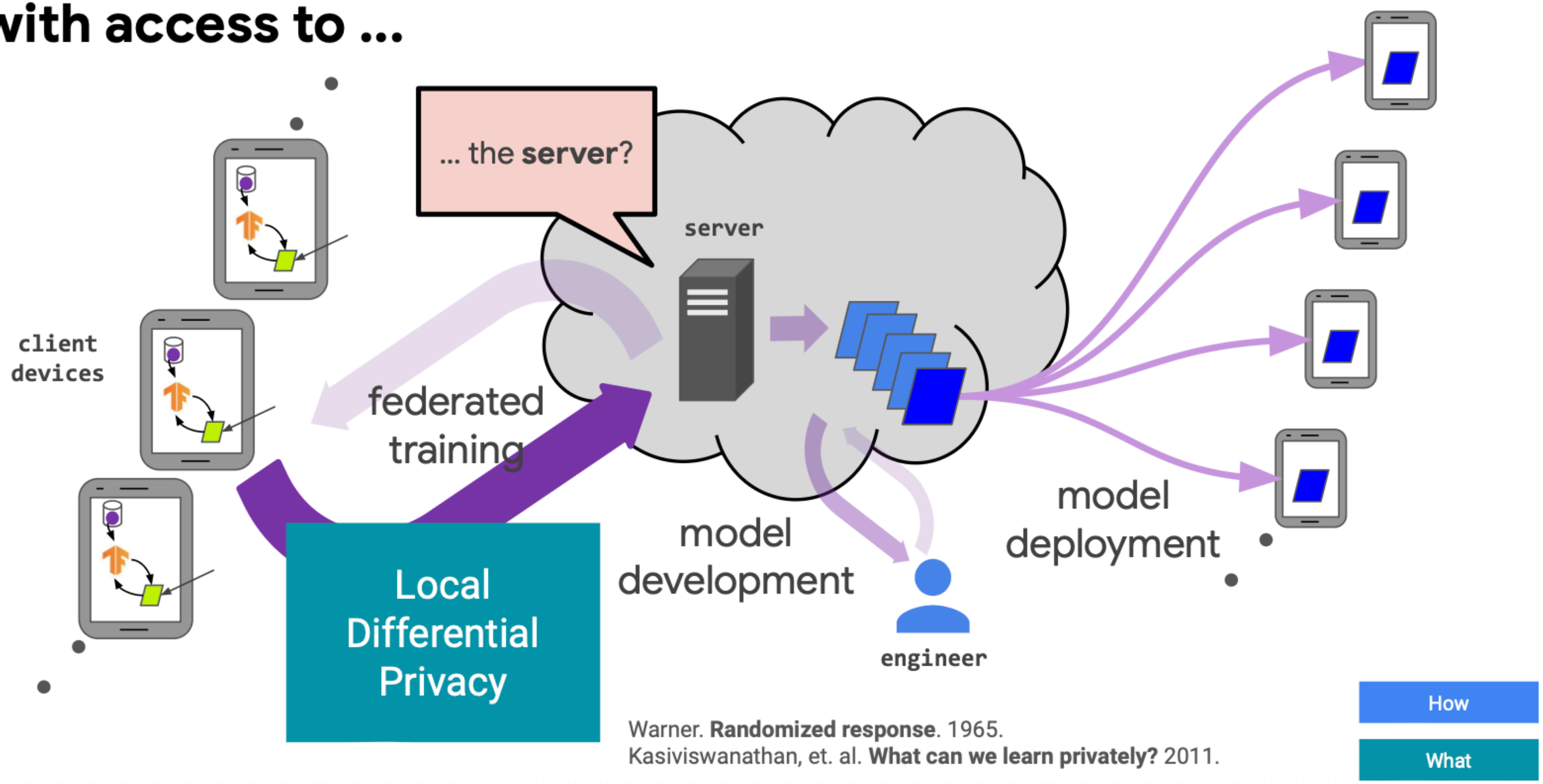


What private information might an actor learn with access to ...



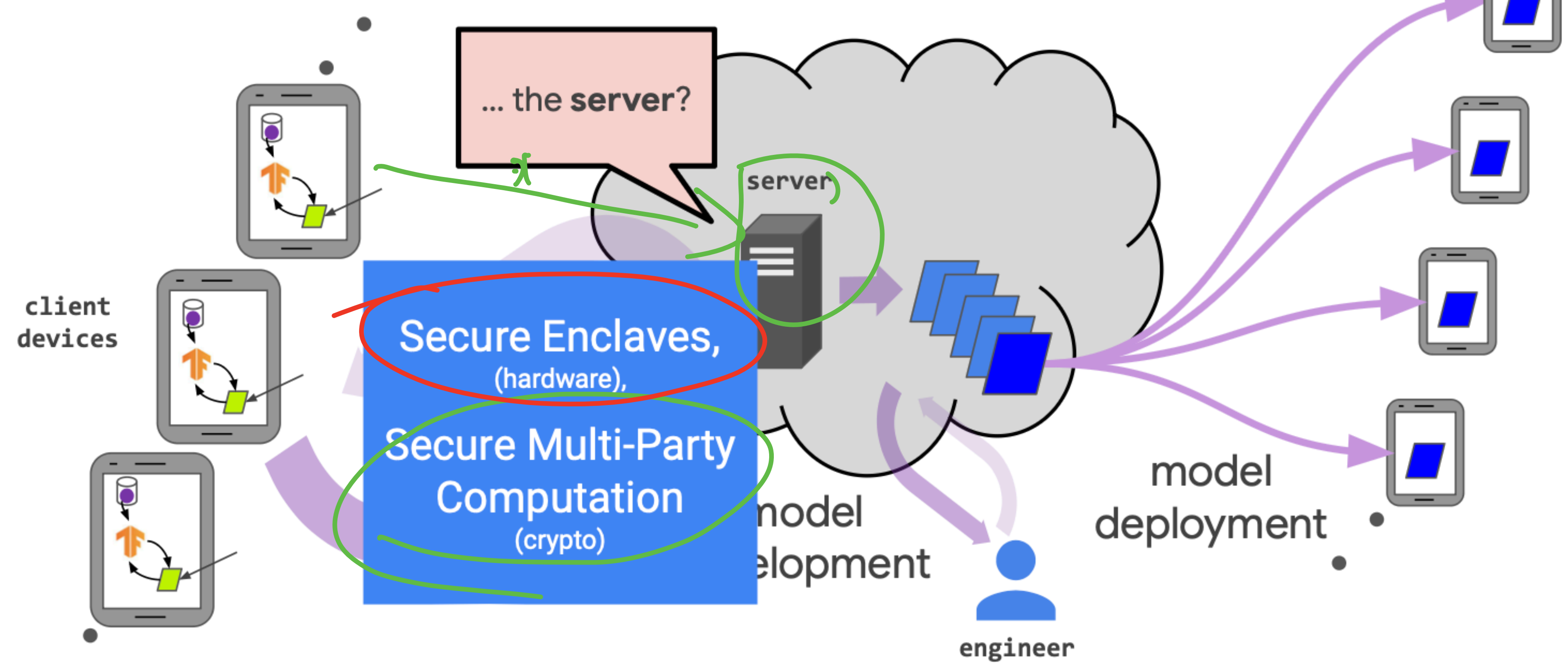
Want: server learns nothing!

What private information might an actor learn with access to ...

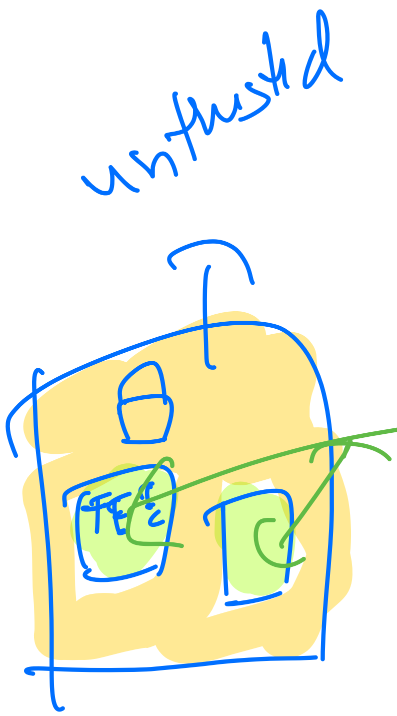


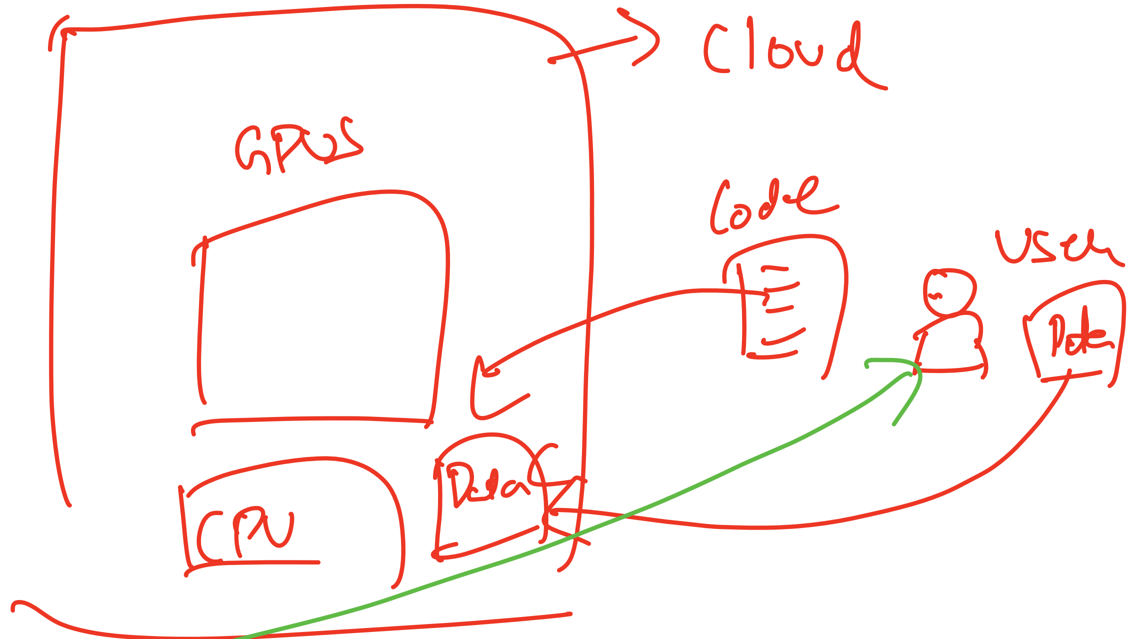
Warner. **Randomized response**. 1965.
Kasiviswanathan, et. al. **What can we learn privately?** 2011.

What private information might an actor learn



Ideally, **nothing**, even with root access.





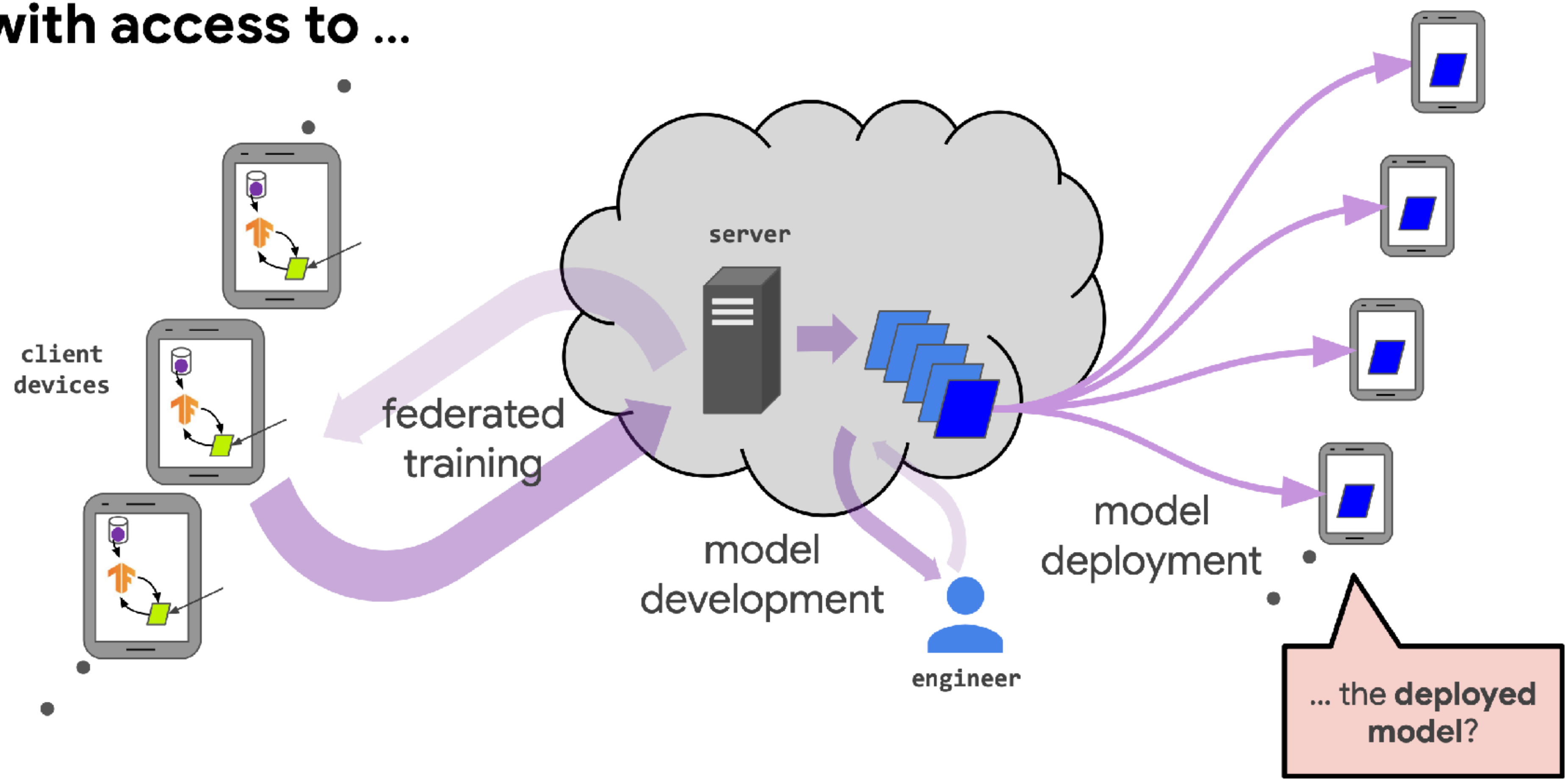
- is the cloud actually running my code on my data?
- what is it with my data?

TEE Trusted Execution Environment

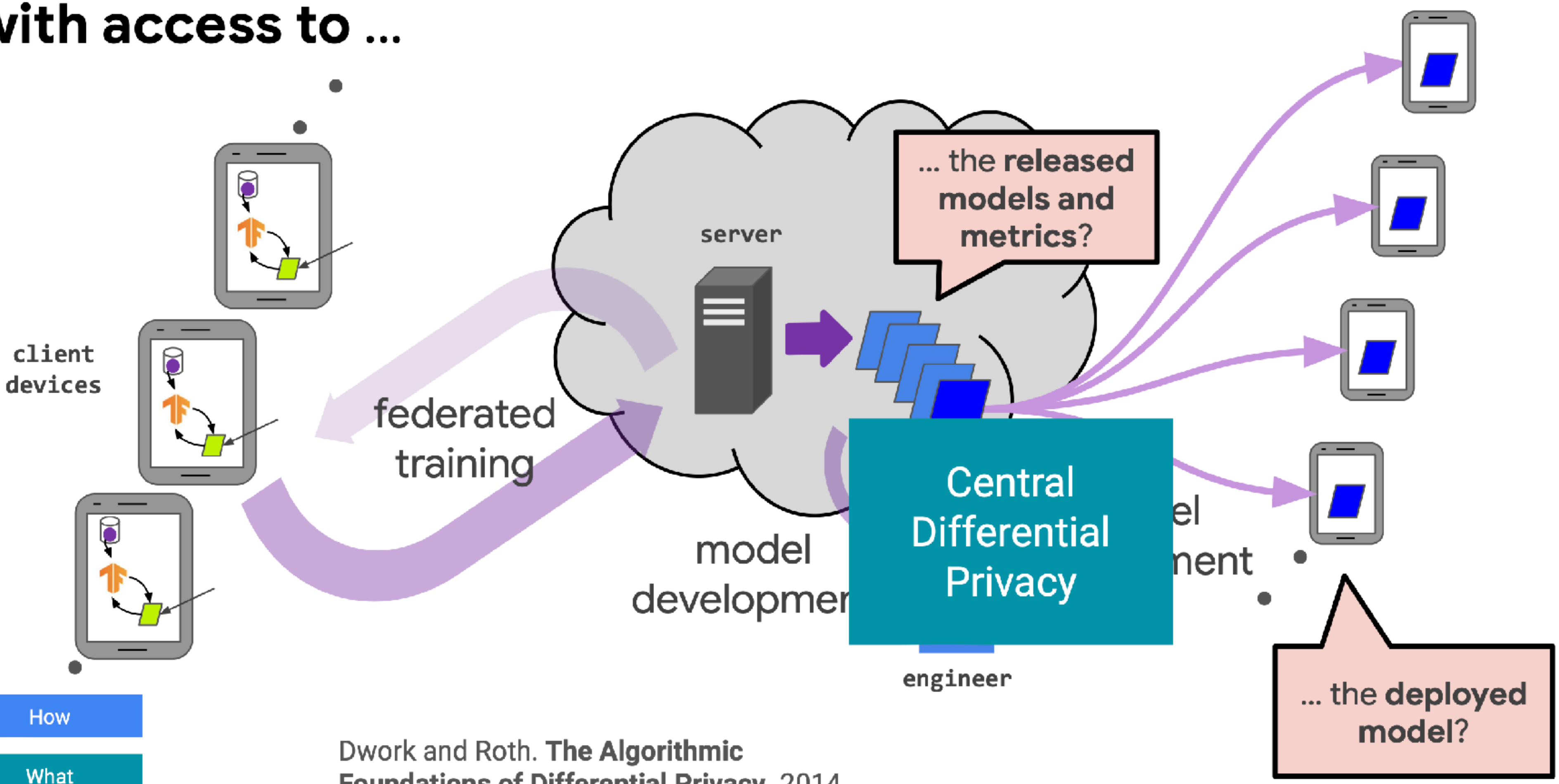
Secure Enclave

→ NVIDIA - TEE GPUs
Intel - TEE CPUs + data

What private information might an actor learn with access to ...



What private information might an actor learn with access to ...



How

What

Dwork and Roth. **The Algorithmic Foundations of Differential Privacy**. 2014.

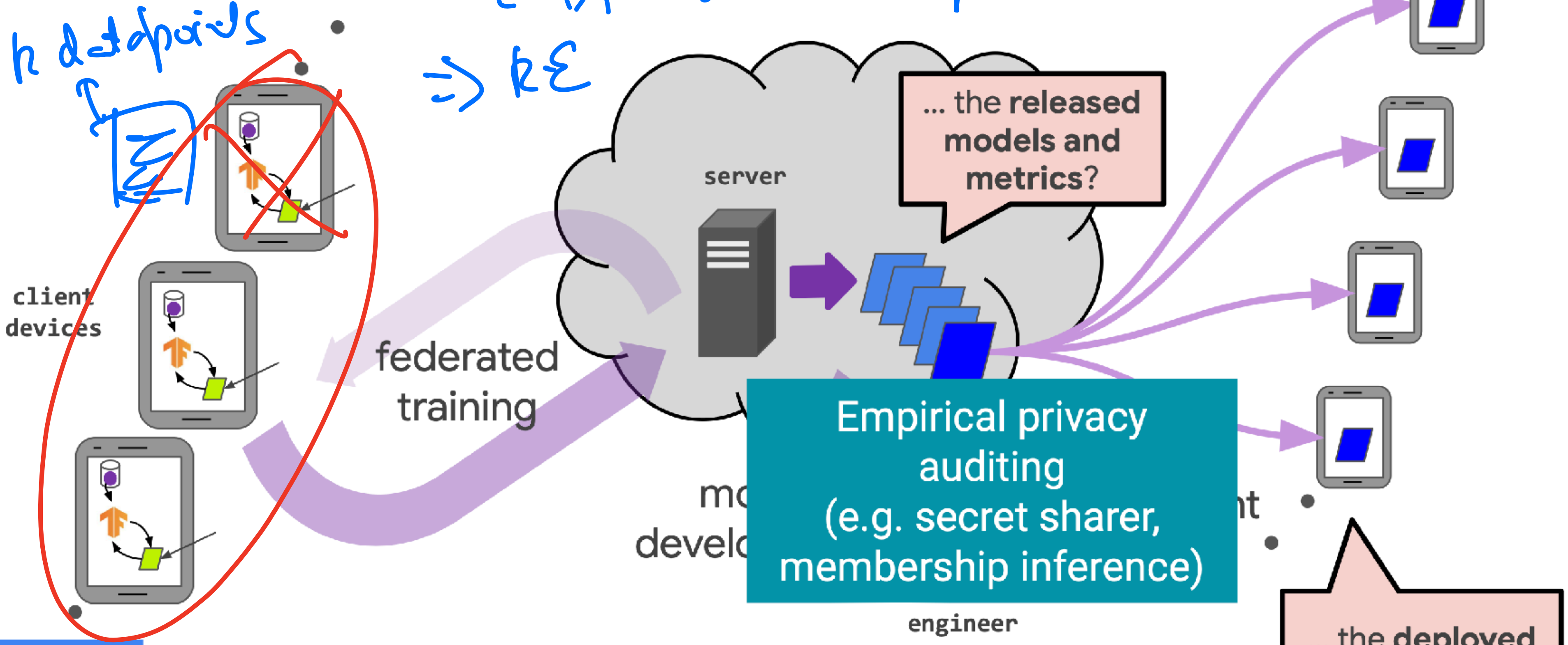
What private information might an actor learn with access to ...

D D'

ε-DP at datapoint level

n datapoints

$\Rightarrow n\varepsilon$

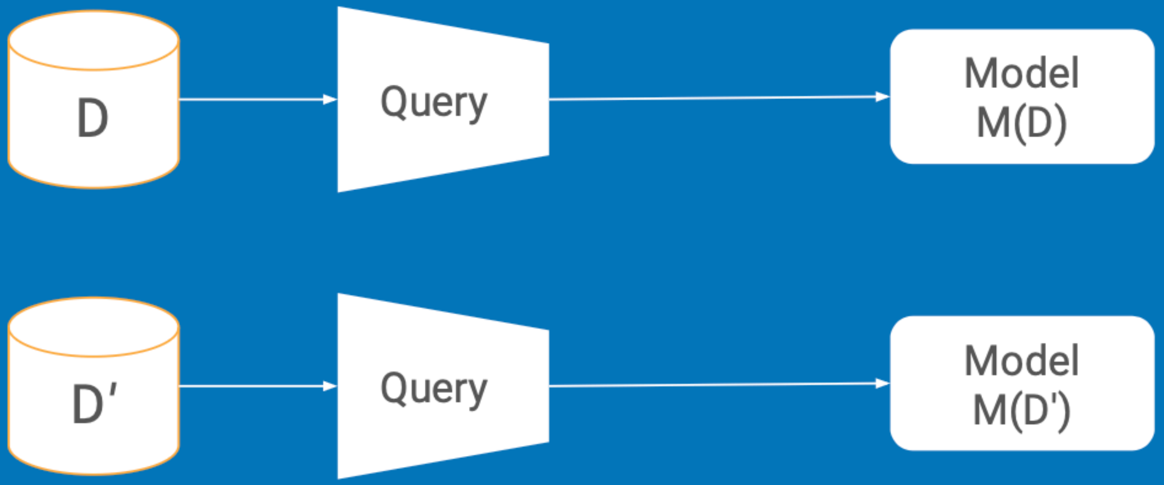


How

What

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, Dawn Song. **The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks**. 2018.
Congzheng Song, Vitaly Shmatikov. **Auditing Data Provenance in Text-Generation Models**. 2018.
Matthew Jagielski, Jonathan Ullman, Alina Oprea. **Auditing Differentially Private Machine Learning: How Private is Private SGD?** 2020.

User-level Differential Privacy

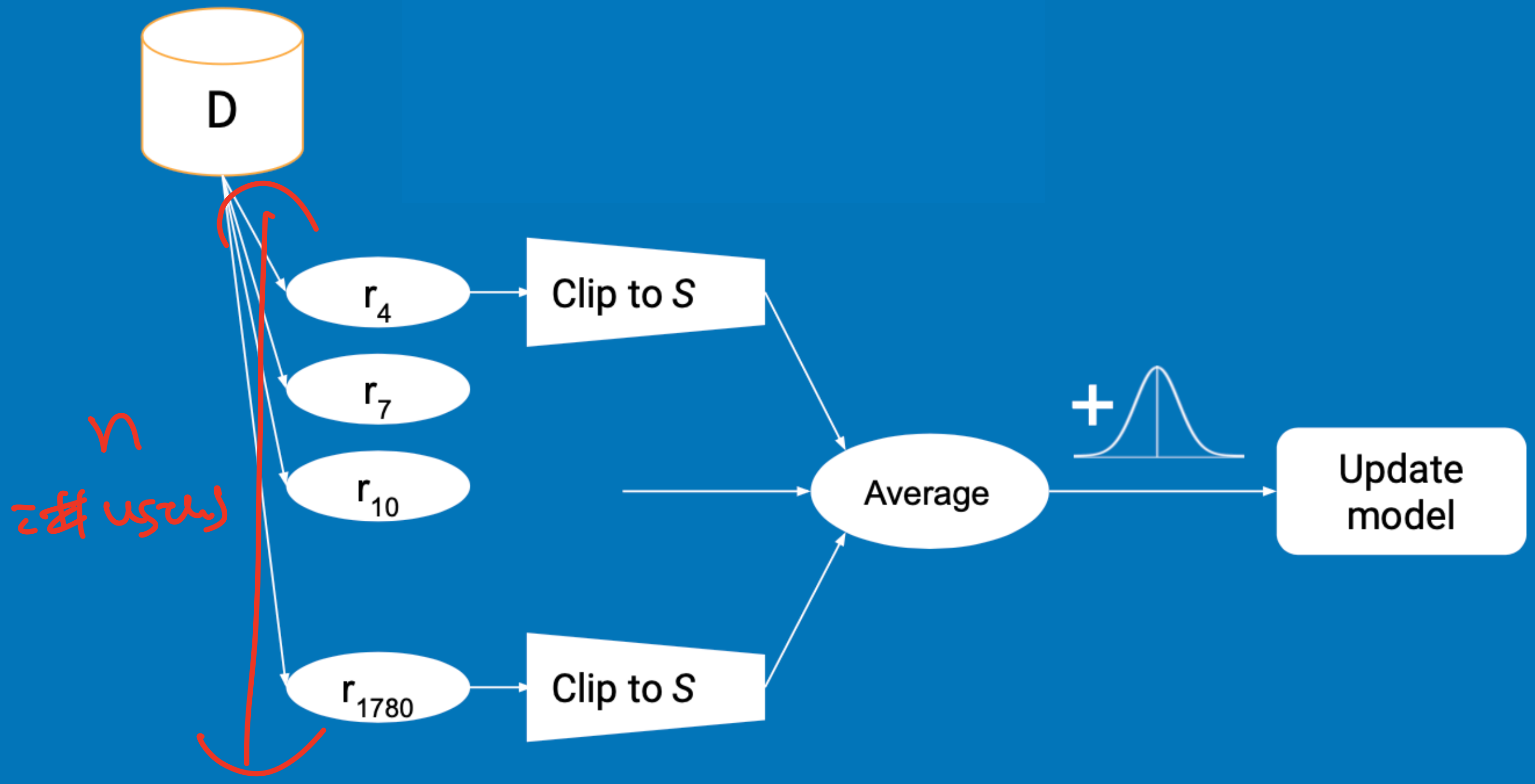


(ε, δ)-Differential Privacy: The distribution of the output $M(D)$ (a trained model) on database (training dataset) D is nearly the same as $M(D')$ for all **adjacent** databases D and D'

adjacent: Sets D and D' differ only by presence/absence of one ~~**example**~~ **user**

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

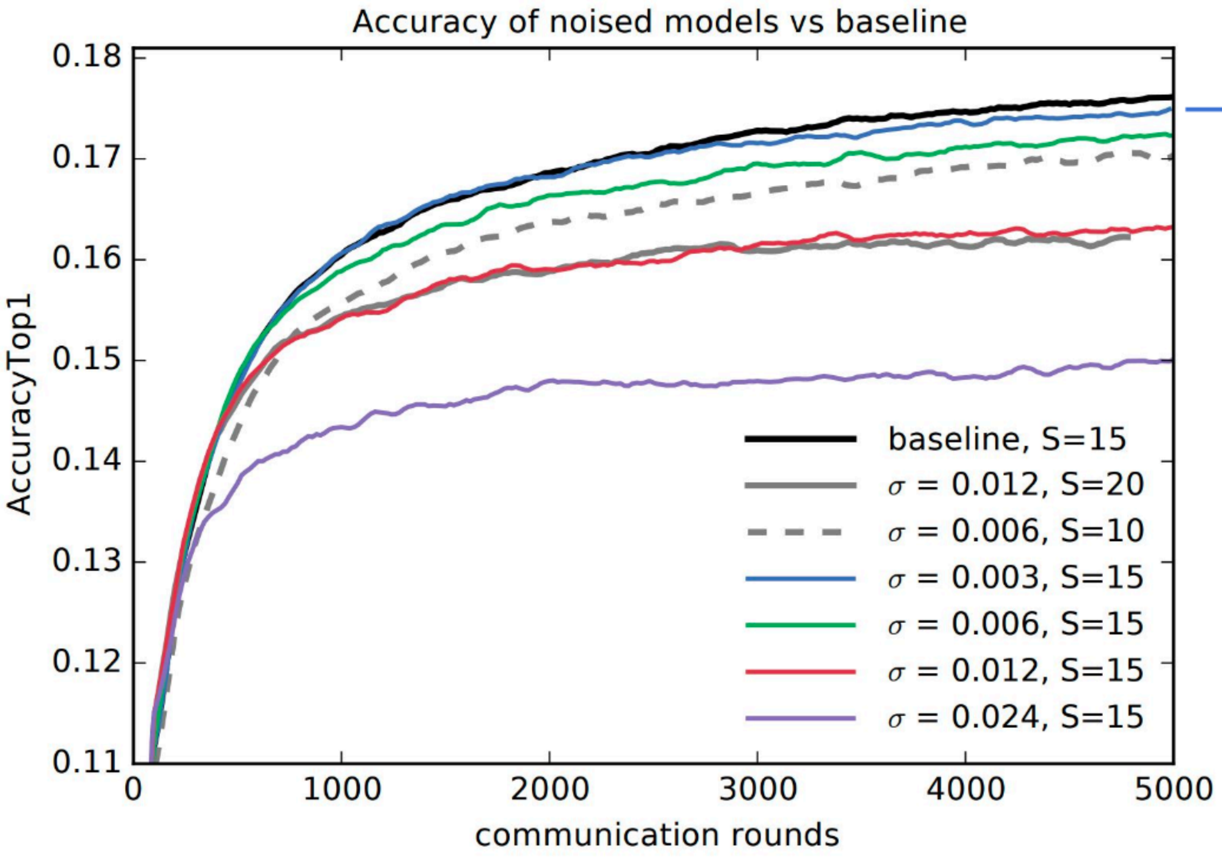
Iterative training with differential privacy



Differential privacy for language models

LSTM-based predictive language model.

10K word dictionary, word embeddings $\in\mathbb{R}^{96}$, state $\in\mathbb{R}^{256}$, parameters: 1.35M. Corpus=Reddit posts, by author.

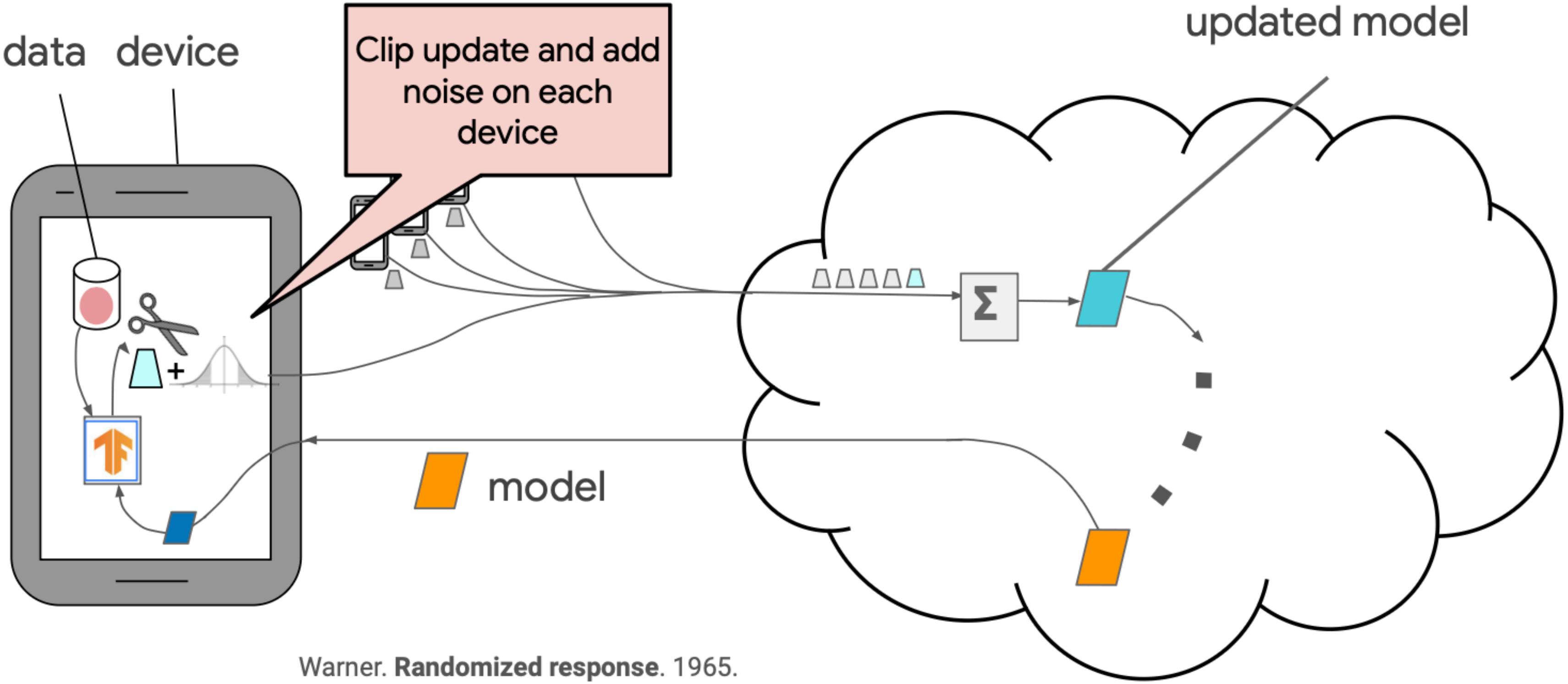


(4.634, 1e-9)-DP with 763k users
(1.152, 1e-9)-DP with 1e8 users

$\mathbb{E}$[users per minibatch] = 5k
$\mathbb{E}$[tokens per minibatch] = 8m

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

Locally differentially private federated learning



Warner. **Randomized response**. 1965.
Kasiviswanathan, et. al. **What can we learn privately?** 2011.

Central DP:

easier to get high utility with good privacy

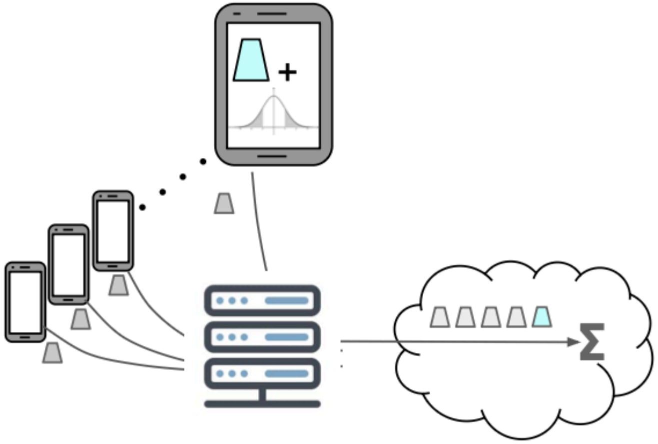
Local DP:

requires much weaker trust assumptions

Can we combine the best of both worlds?

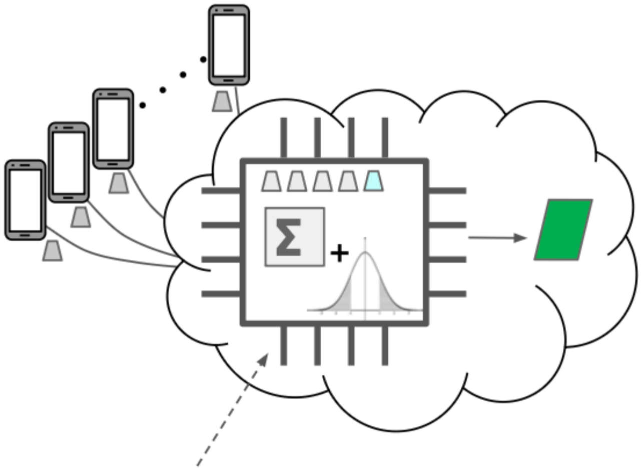
Distributing Trust for Private Aggregation

1 Trusted "third party"



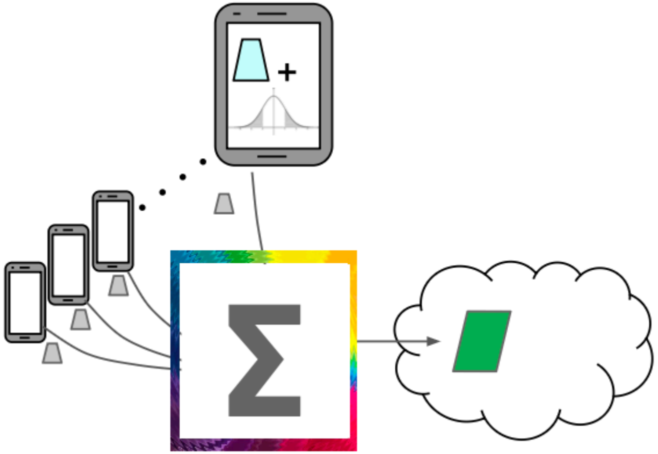
third-party

2 Trusted Execution Environments



trusted
hardware

3 Trust via Cryptography



cryptography

The Dream vs. Current Reality of FL

Markus Bujotzek et al. 2024. Real-World Federated Learning in Radiology: Hurdles to overcome and Benefits to gain

The Dream vs. Current Reality of FL

- The dream:
    - A private distributed global protocol
    - That unites the world's data and compute

- The current reality
    - Data is extremely messy and even actively harmful - need **data quality and valuation.**
    - Cannot train LLMs over commodity hardware - need better **sysML.**
    - Unclear legal/policy support.

NHS England faces lawsuit over patient privacy fears linked to new data platform

Four groups claim no legal basis exists for setting up the Federated Data Platform which facilitates information sharing