



LLM Privacy: From Principles to Practice

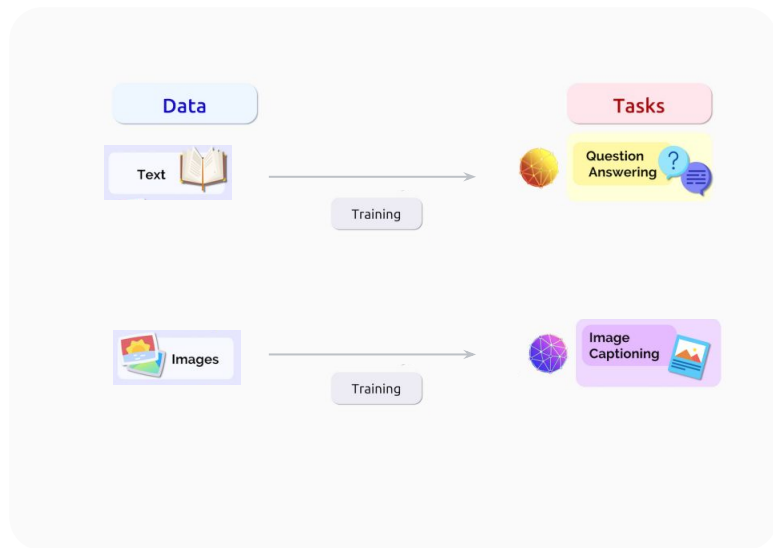
Peter Kairouz – presenting work done with and slides made by many

X: @KairouzPeter

Special thanks to:

Brendan McMahan, Daniel Ramage, Galen Andrew, Zheng Xu, Eugene Bagdasarian,
Ryan McKenna

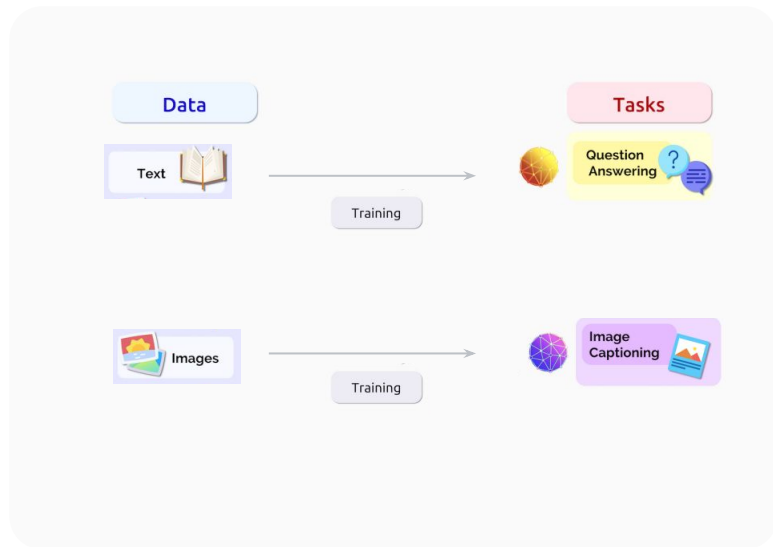
The paradigm shift of foundation models



Before foundation models

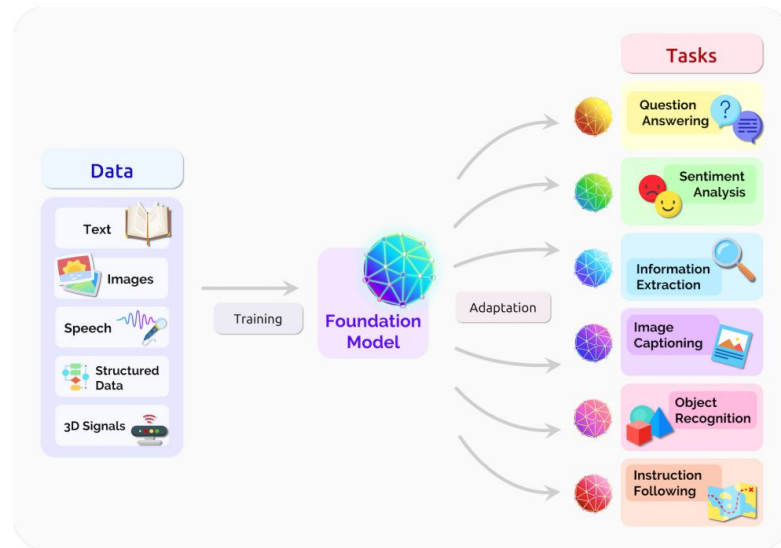
Task-specific data and models, trained independently.

The paradigm shift of foundation models



Before foundation models

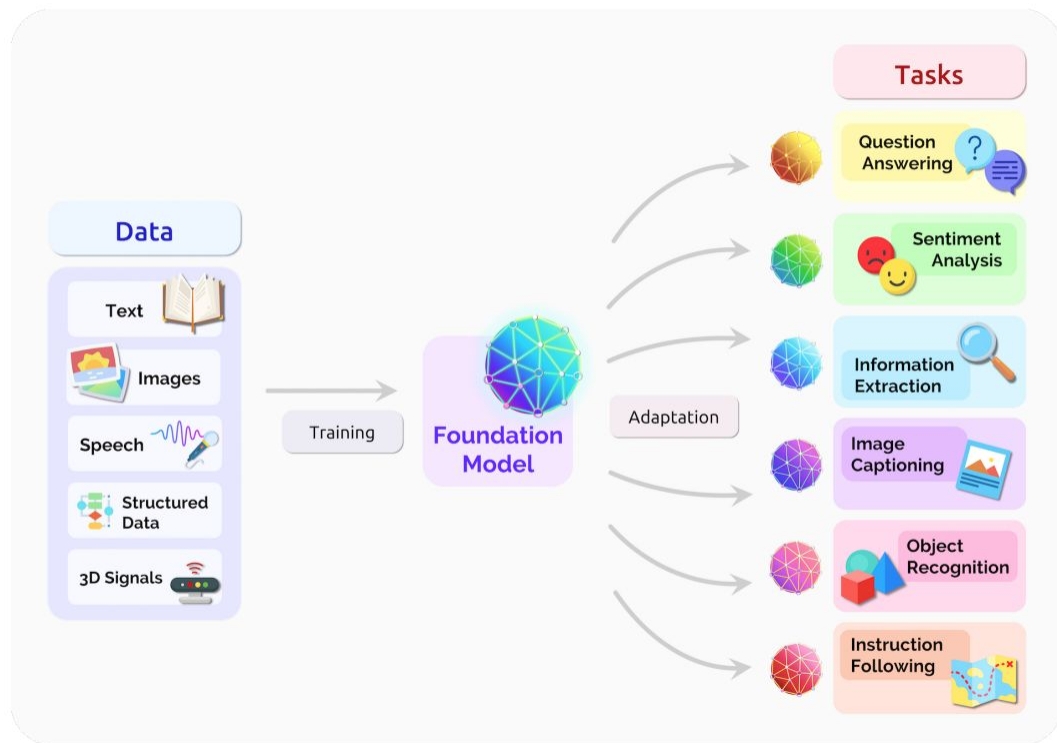
Task-specific data and models, trained independently.



Post foundation models

Foundation models are trained with data across various modalities, then adapted to a wide range of downstream tasks.

The rapidly evolving landscape of foundation models



Bommasani, et al. **On the Opportunities and Risks of Foundation Models**. Stanford Center for Research on Foundation Models, Stanford Institute for Human-Centered Artificial Intelligence

[BERT](#) [Oct '18]: Pre-text task with ~340M transformer model

[GPT-1](#) → [GPT-4o](#) [Jun '18 → Apr '25]: Autoregressive text generation (117M → 1.7T)

[CLIP](#) [Jan '21]: Image captioning using pre-training tricks inspired by BERT (63M)

[DALL-E](#) [Jan '21]: Text-to-image generation with a "mini" GPT-3 (12B)

[LaMDA](#) / [Bard](#) [Jan '22 / Feb '23]: Language model for dialogue applications (137B)

[PaLM](#) → [PaLM 2](#) [Apr '22 → May '23]: Gemini's predecessor models (340B)

[Gemini](#) → [Gemini 2.5](#) [Dec '23 → Mar '25]: Natively multi-modal & reasoning models

[Llama](#) → [Llama 4](#) [Feb '23 → April '25]: Open source foundation models (109B → 2T)

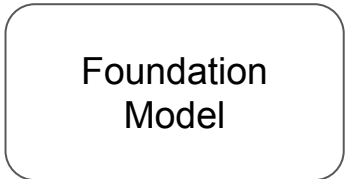
[Claude](#) → [Claude 3.7](#) [Mar '23 → Feb '25]: Anthropic's conversational foundation models

[Gemma](#) → [Gemma 3](#) [Dec '23 → Mar '25]: Google's open source foundation models (1B → 27B)

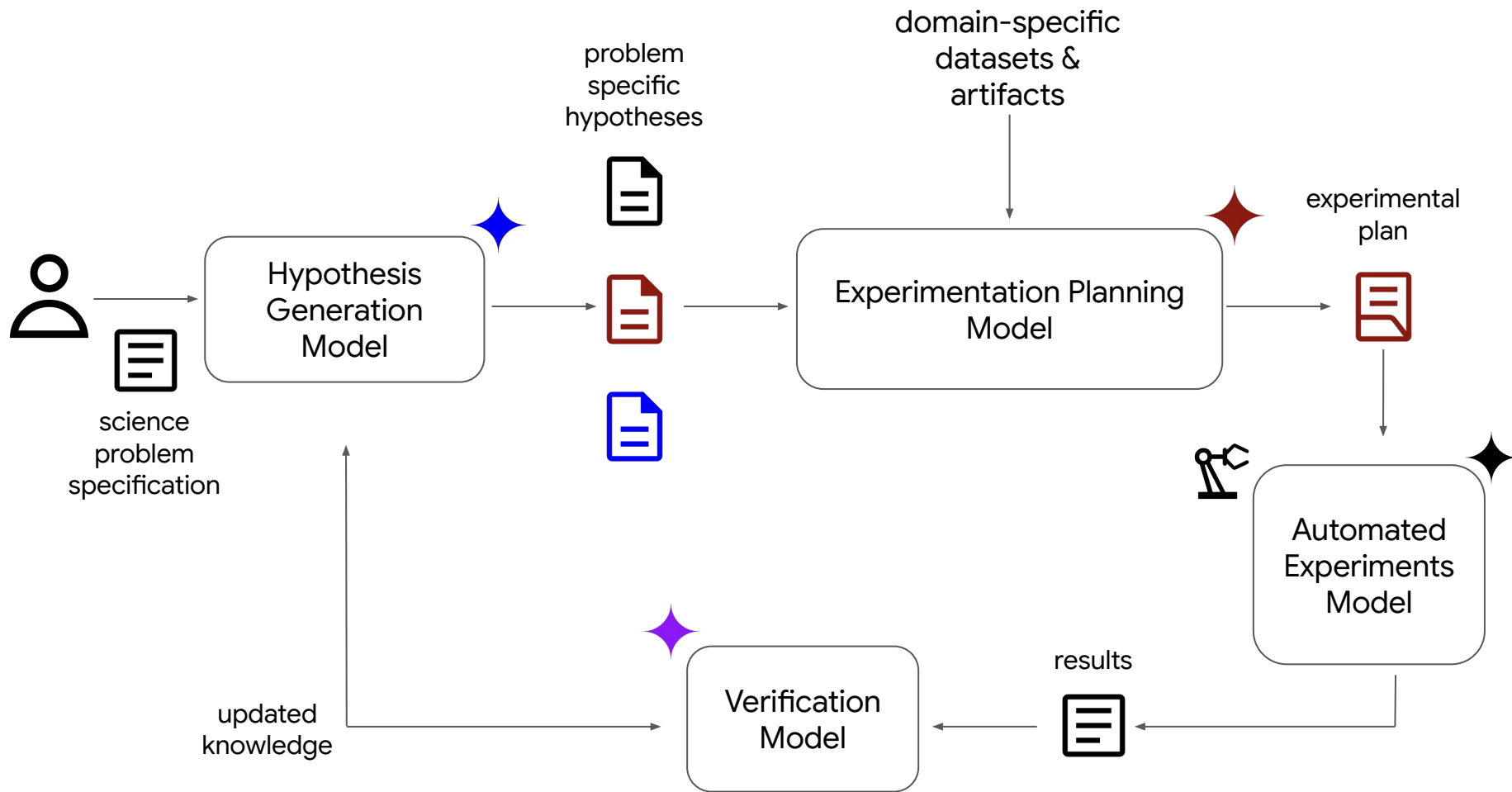
....

Can we solve science questions with these models?

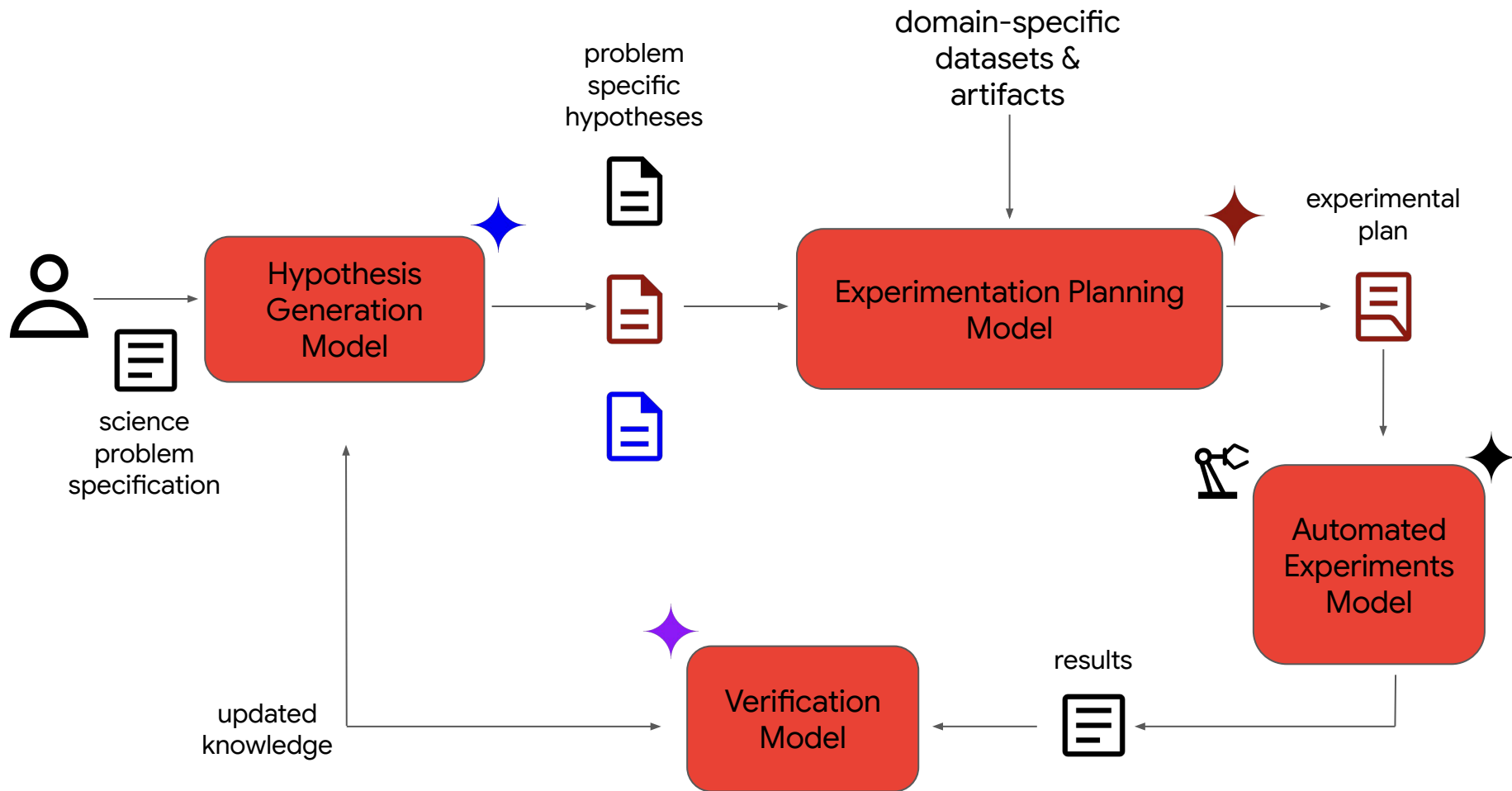
“What are the precise molecular mechanisms that cause Parkinson’s disease, and how can we effectively prevent, slow, or reverse them?”



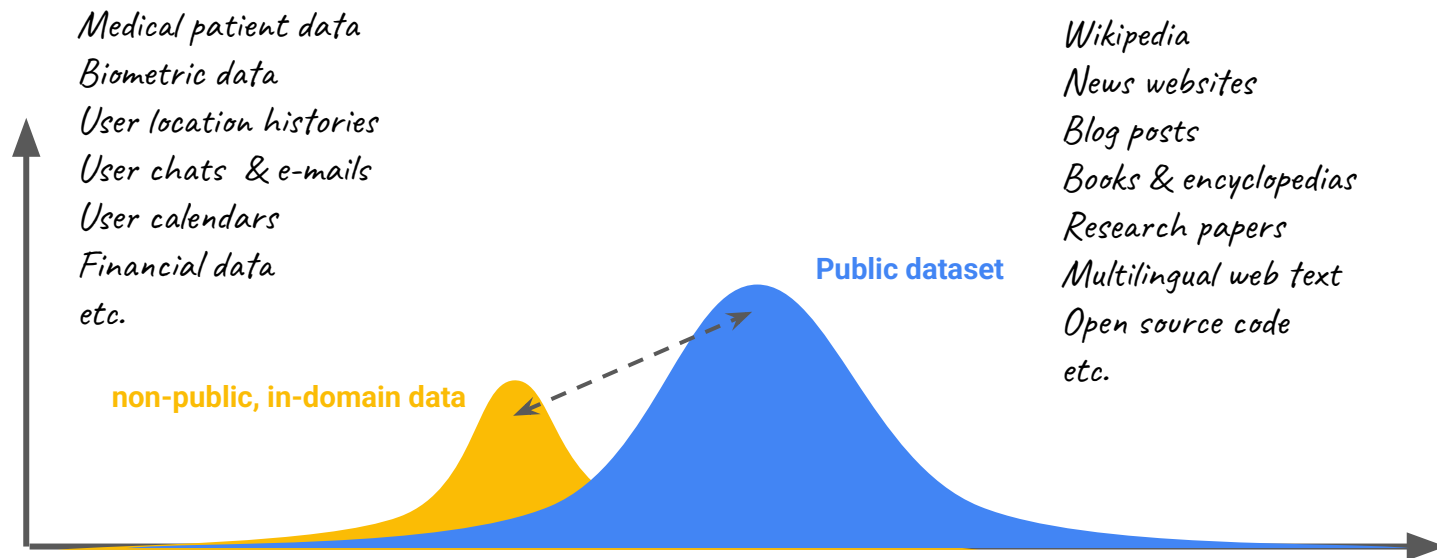
Of course not... but we're working towards a “science agent”



We can't design these core models without in-domain data



Accessing *domain specific* data is inevitable



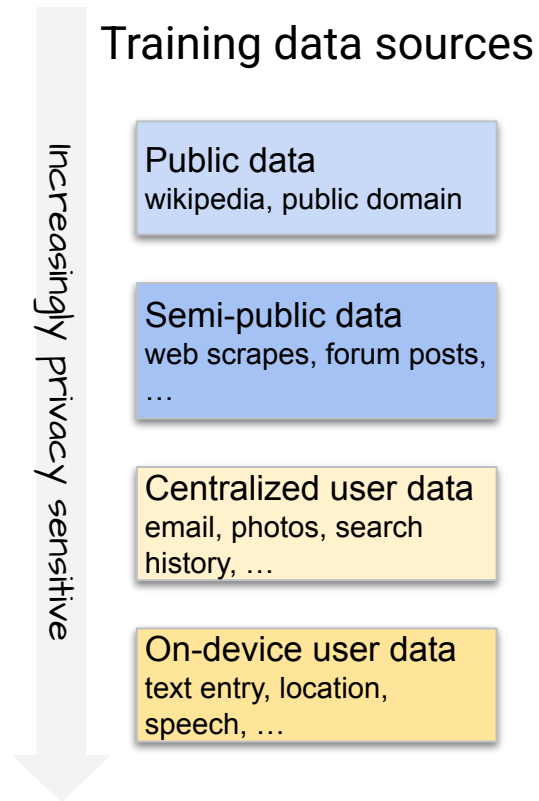
Train models
on *domain*
specific data

Retrieve
relevant *domain*
specific data at
inference

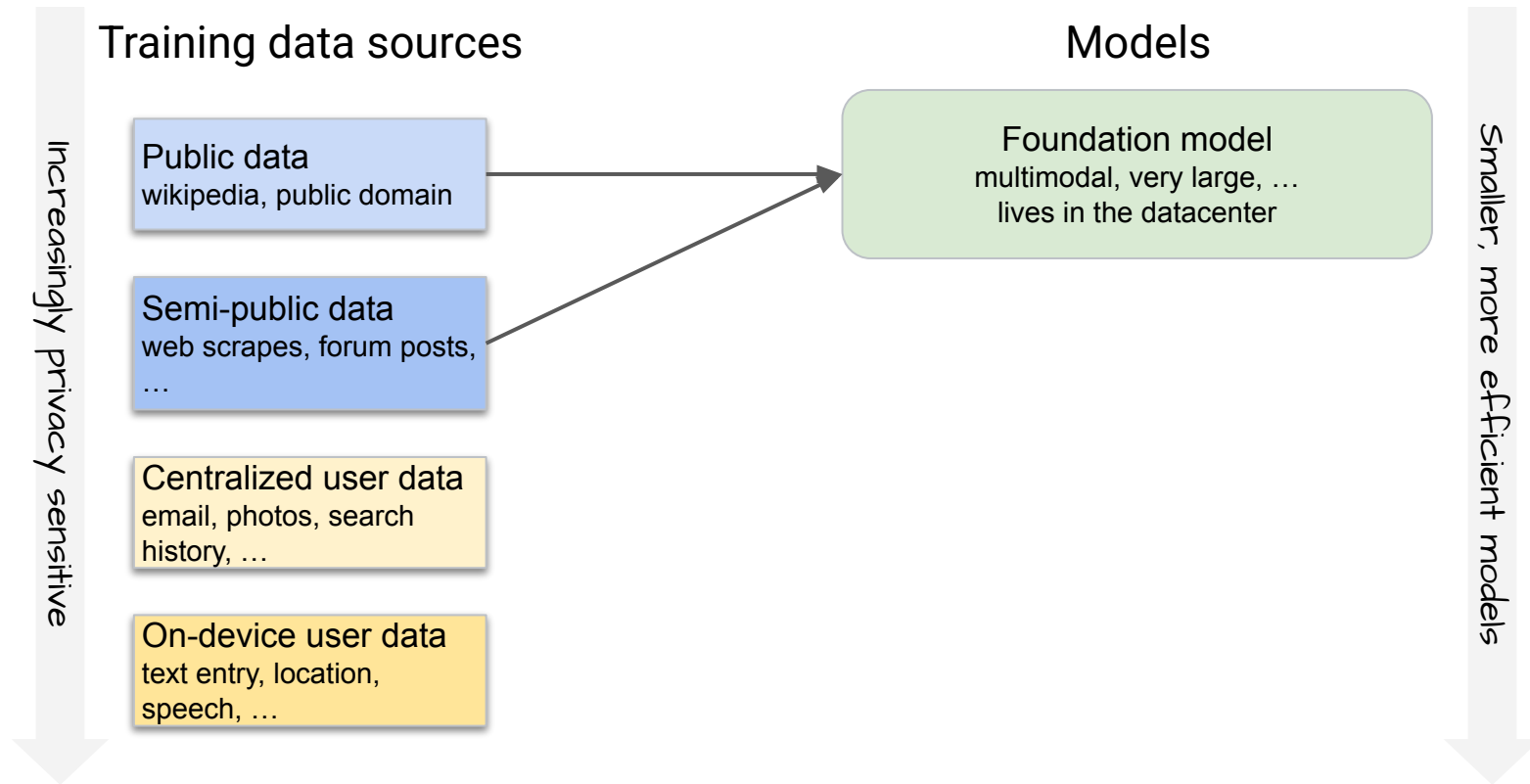
Train models
on *domain*
specific data



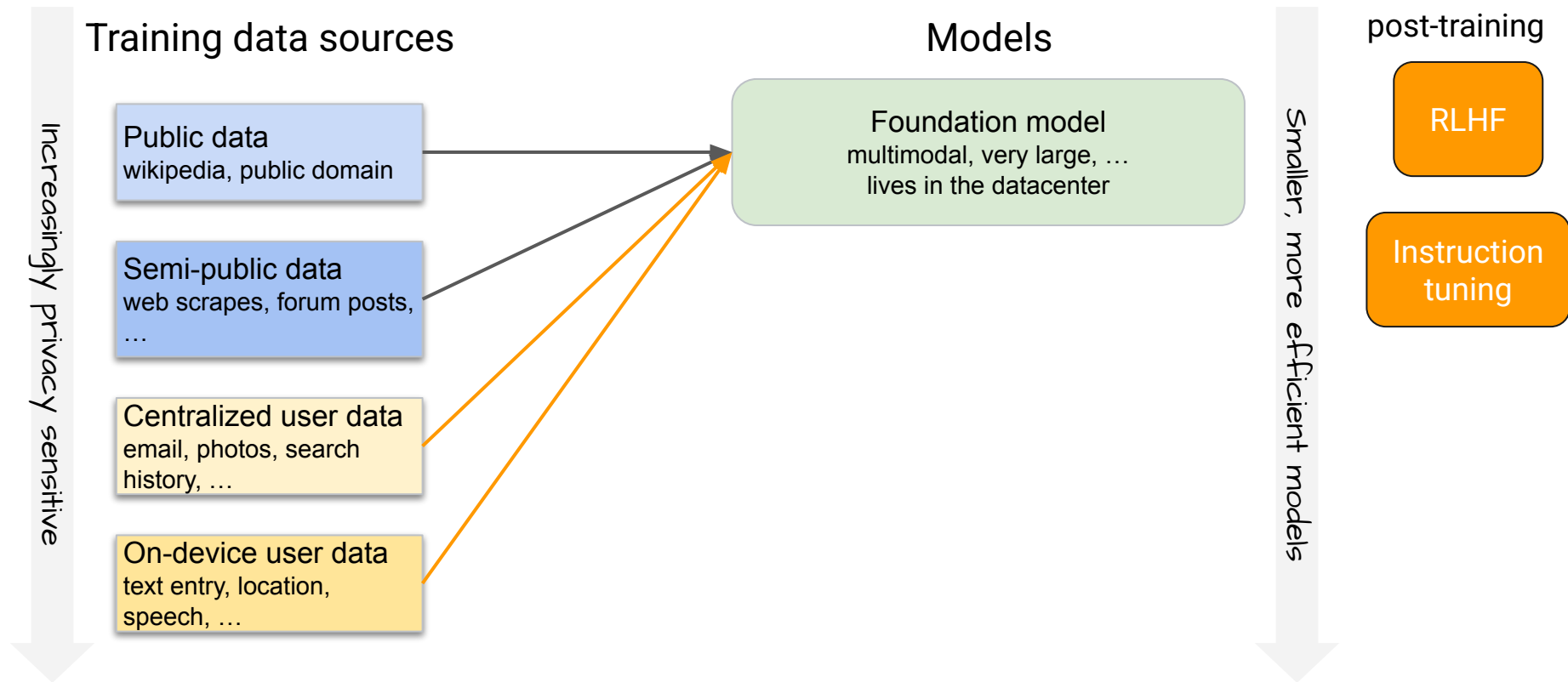
Incorporating in-domain training data sources



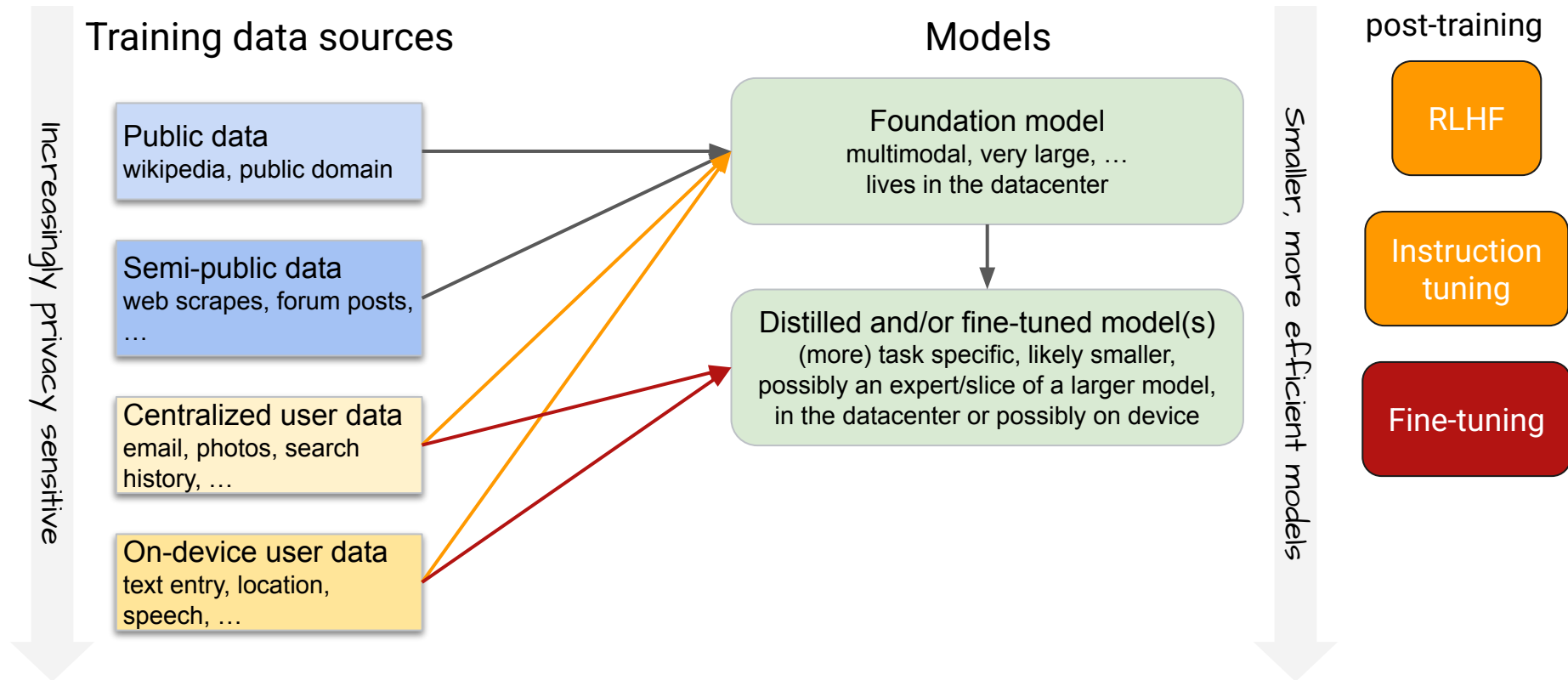
Incorporating in-domain training data sources



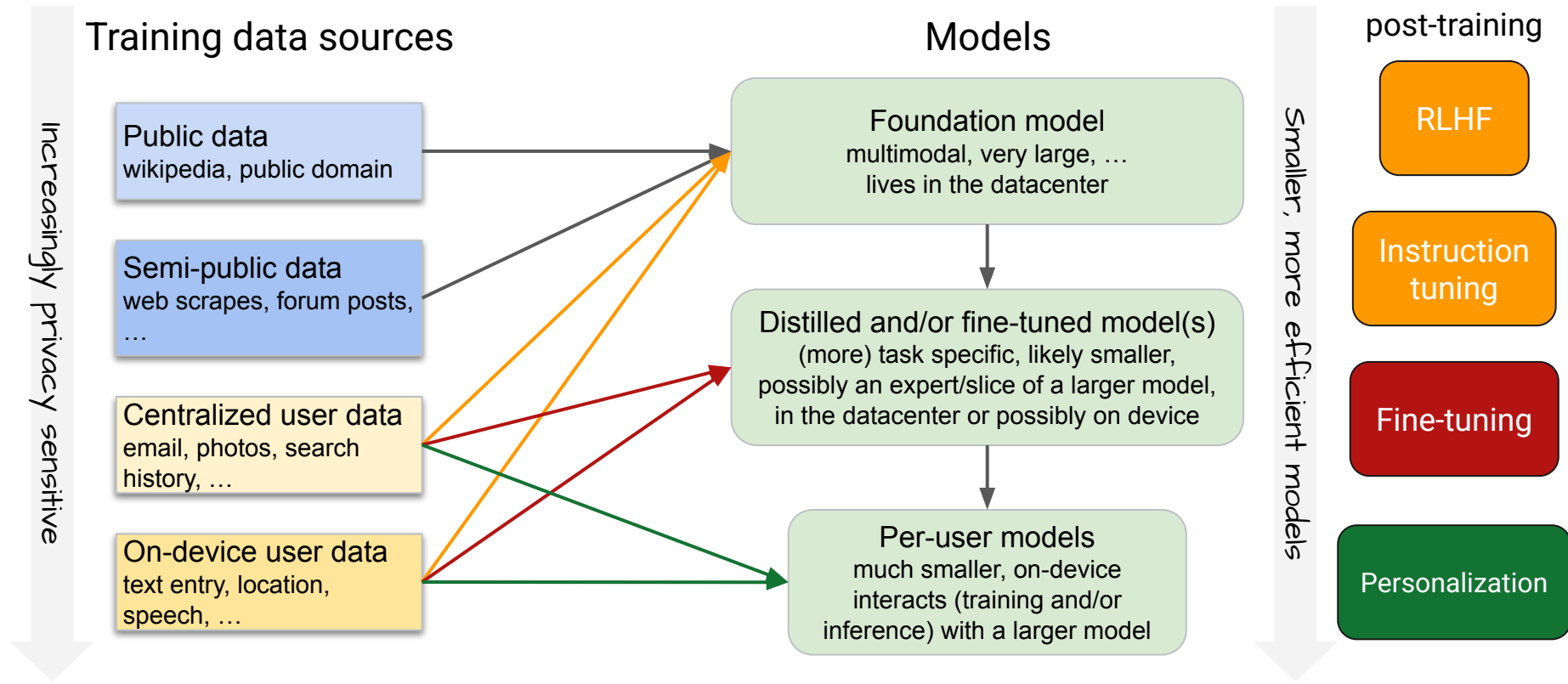
Incorporating in-domain training data sources



Incorporating in-domain training data sources



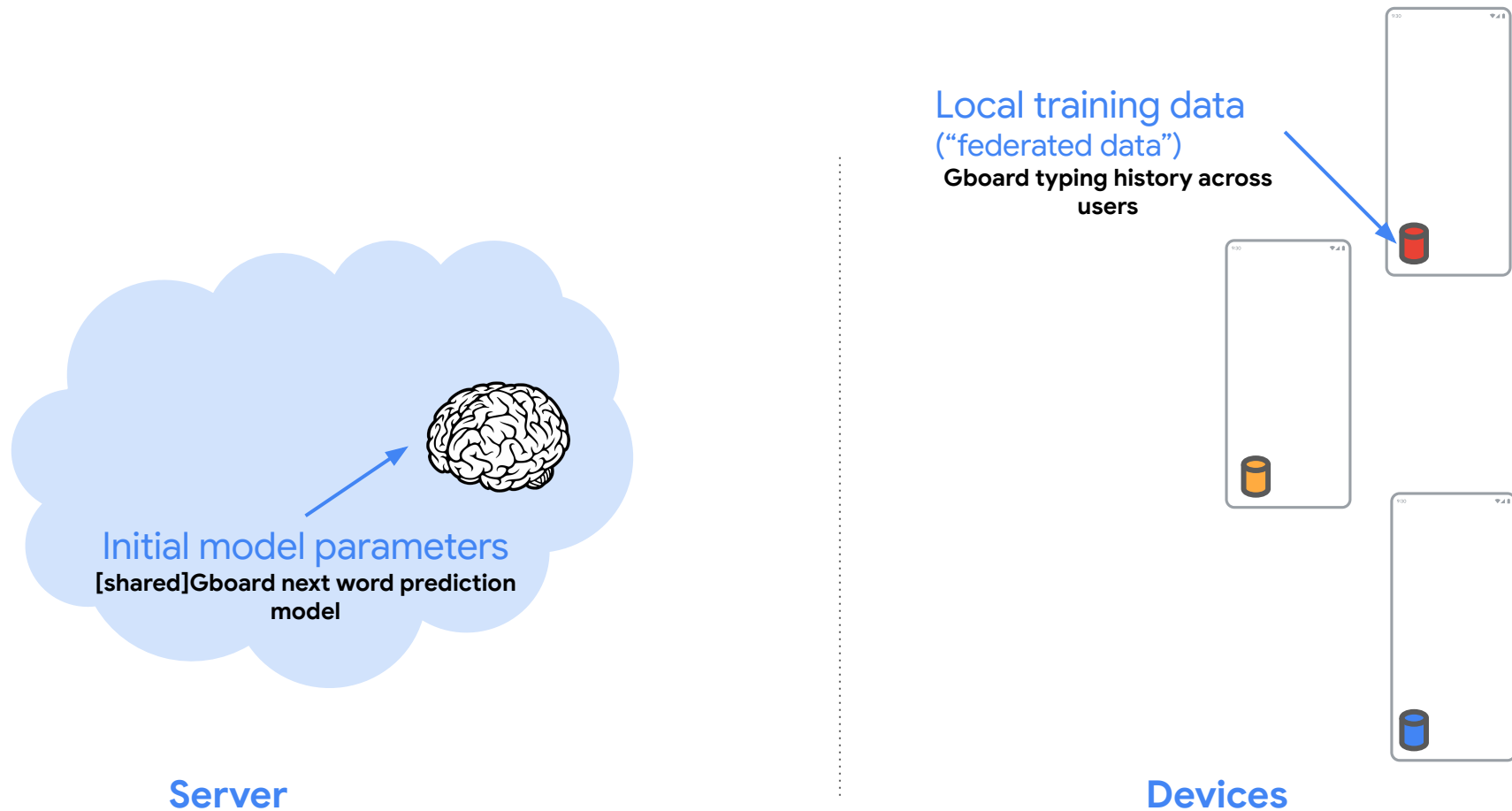
Incorporating in-domain training data sources



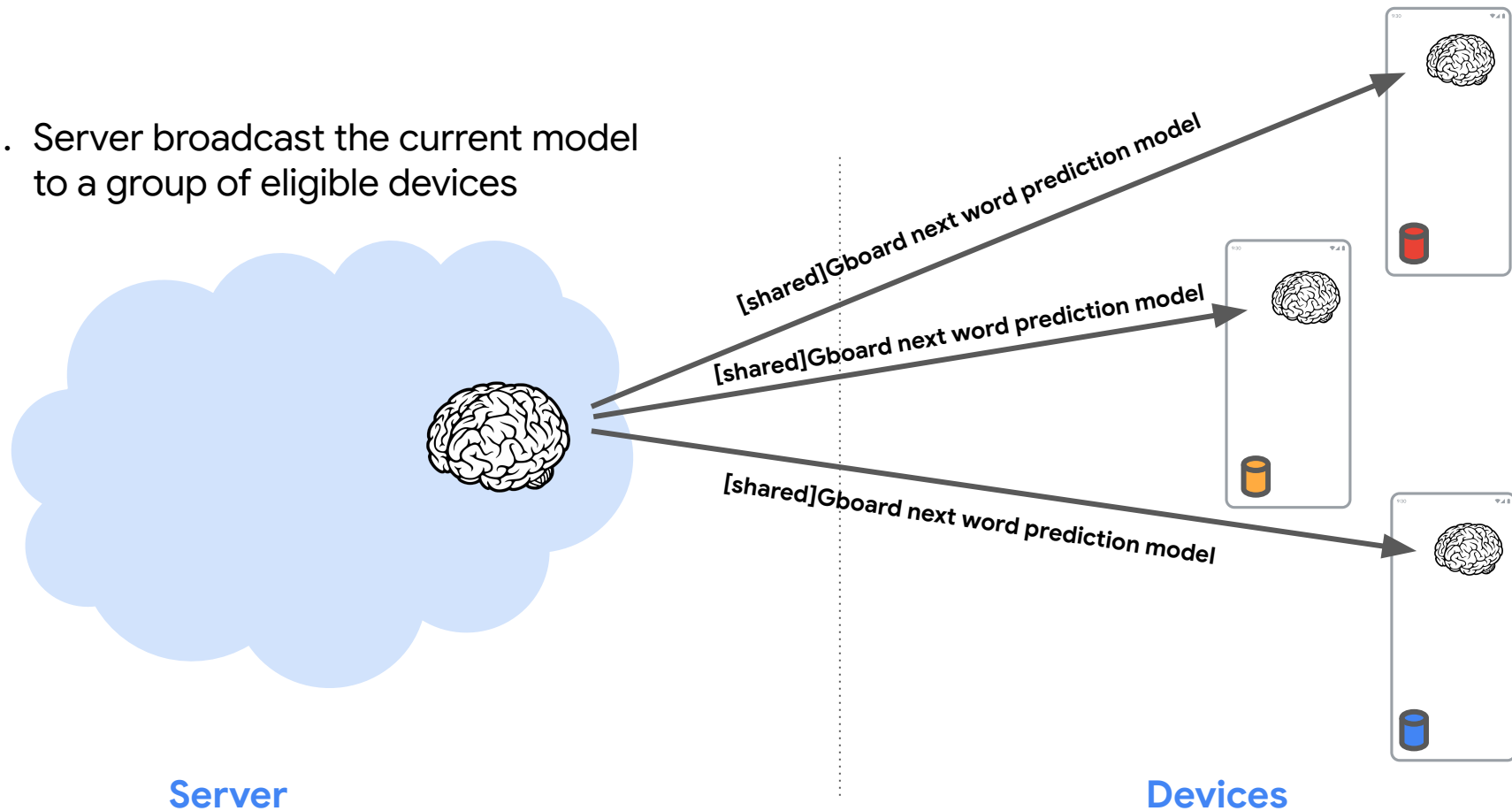
(On-device) user data

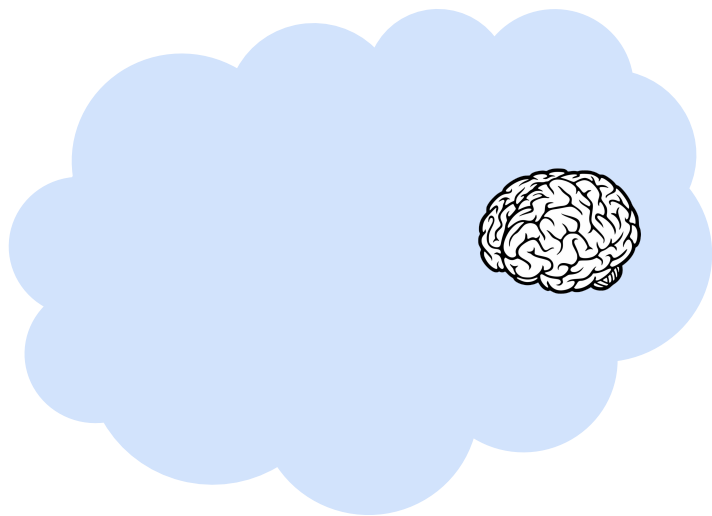
Daily Routines: • Commute times and routes (work, school, etc.) • Typical locations visited (home, work, gym, restaurants) • Sleep patterns (based on device activity at night) • Chats, clicks, swipes, etc.	Lifestyle: • Travel habits (frequency, destinations) • Social activity (frequency of calls, messages, social media use) • Shopping habits (online purchases, app usage) • Fitness levels (tracking steps, exercise apps)	App & Device Usage: • Interests and hobbies (based on app categories) • Financial habits (banking apps, investment apps) • Health and wellness (fitness trackers, health apps) • Socioeconomic status (device model, data plan)
Location-based inferences: • Neighborhood demographics (income, education levels) • Exposure to certain types of businesses or services	Activity levels: • Tracking steps, distance traveled, and exercise routines. • Monitoring sleep duration and patterns.	Sentiment Analysis: • Emotional states (based on search queries, text messages, etc.) • Health concerns • Political affiliations

Machine learning without centralized data collection



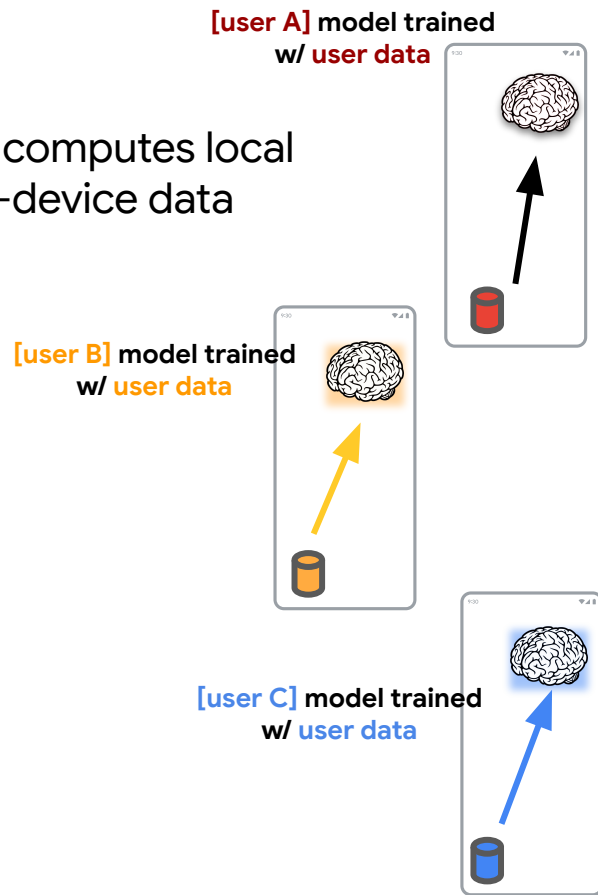
1. Server broadcast the current model to a group of eligible devices





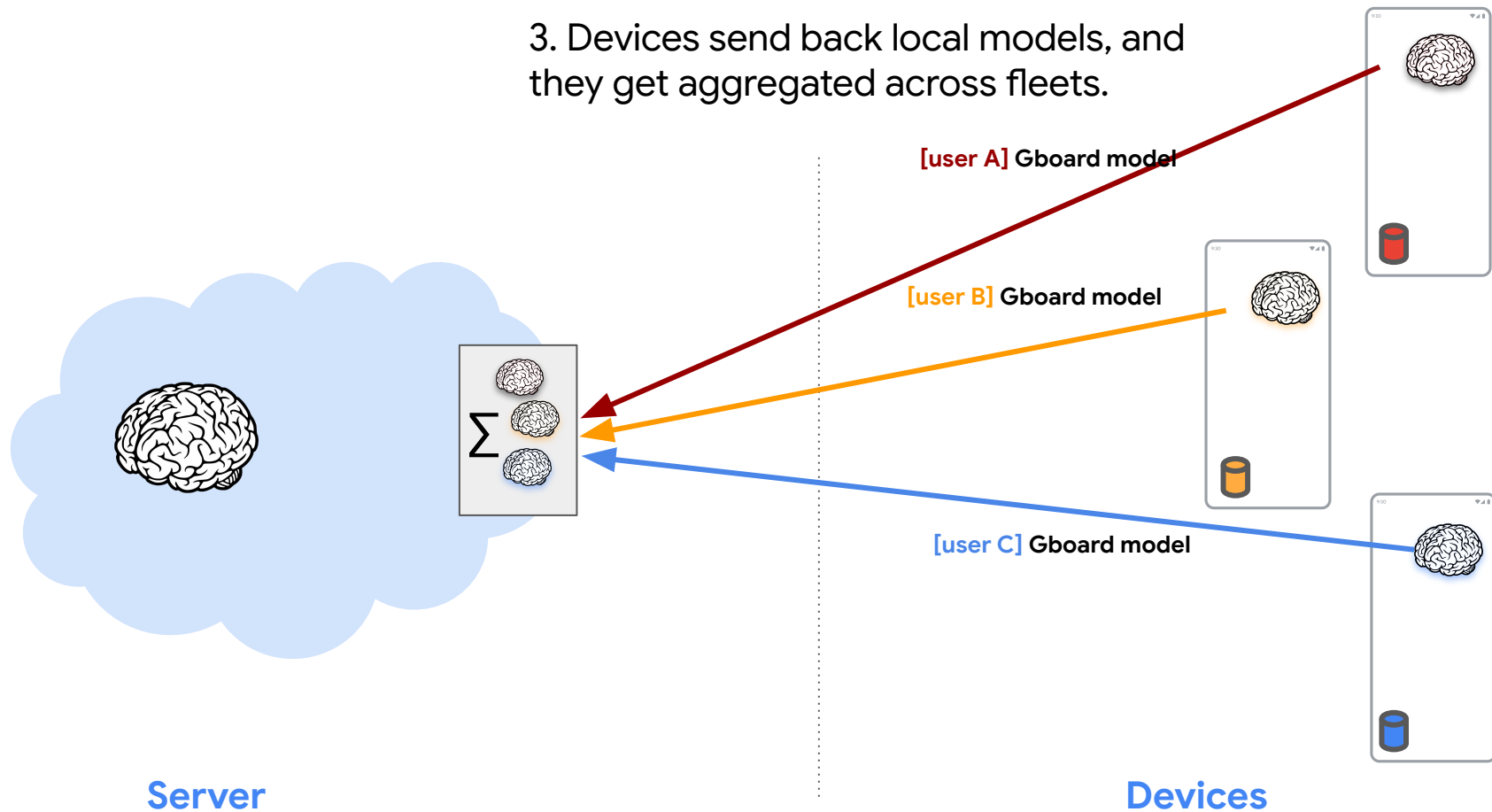
Server

2. Each device computes local model with on-device data

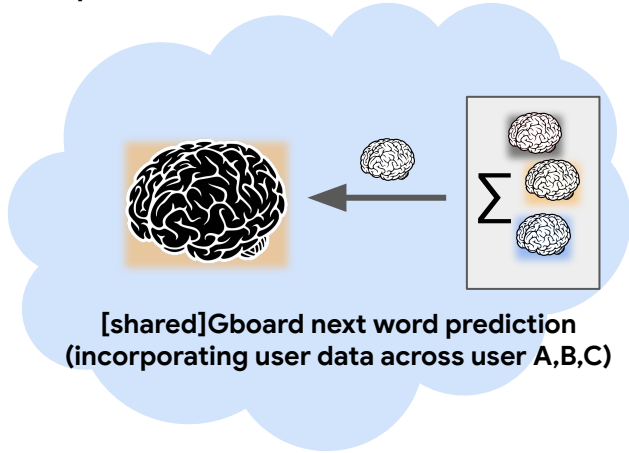


Devices

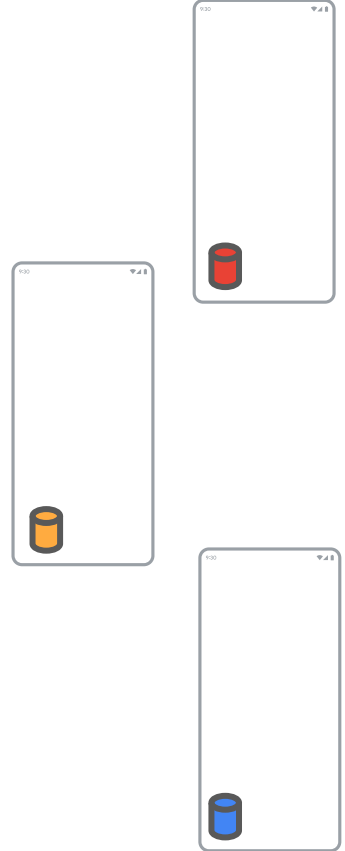
3. Devices send back local models, and they get aggregated across fleets.



4. Only the aggregates are released to the server-side model. Updates are applied on the server-side, and the same process repeats.



Server

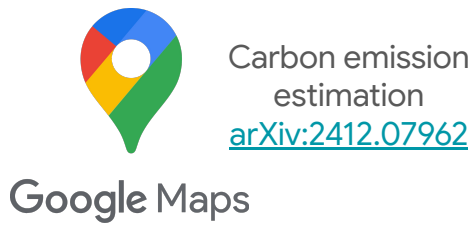


Devices

FL is deployed at Google and beyond...

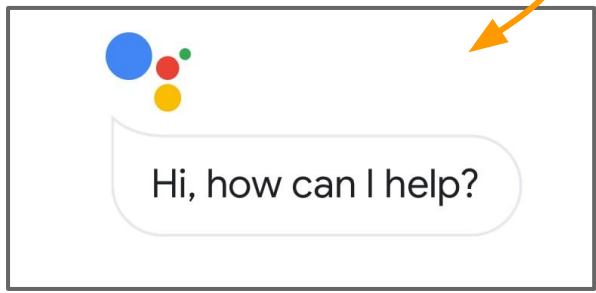


Google Health
Studies

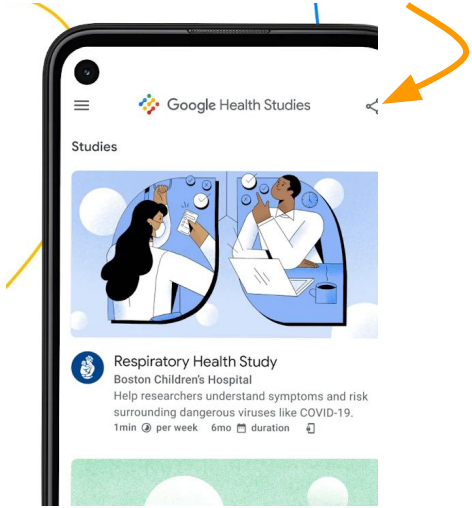


Google Maps

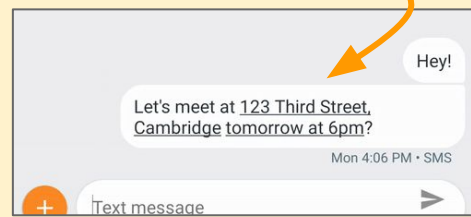
"Hey Google"
hotword detection



Respiratory health study

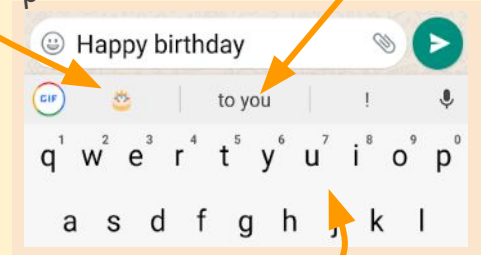


Android
smart text selection



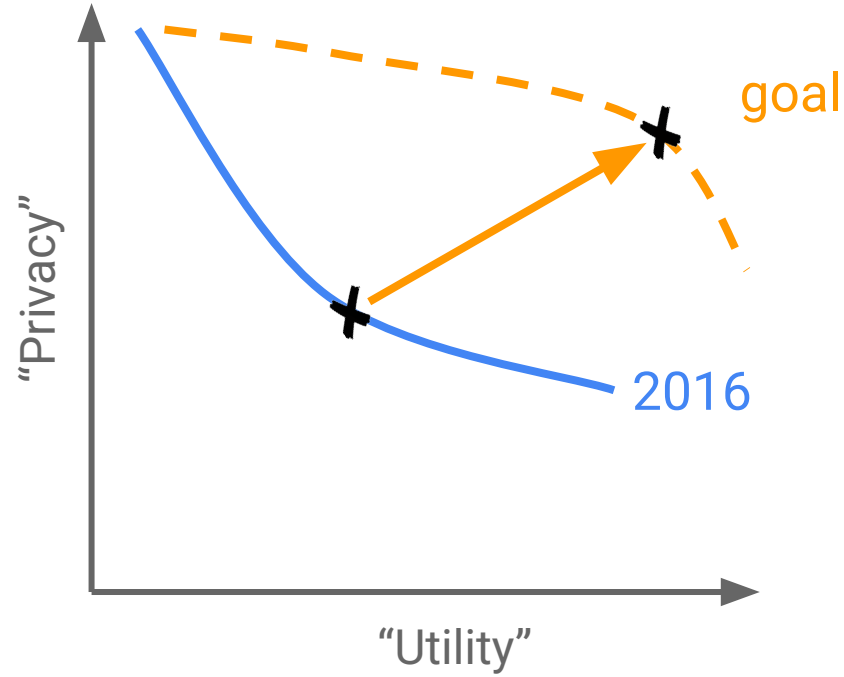
emoji and sticker
prediction

next word
prediction

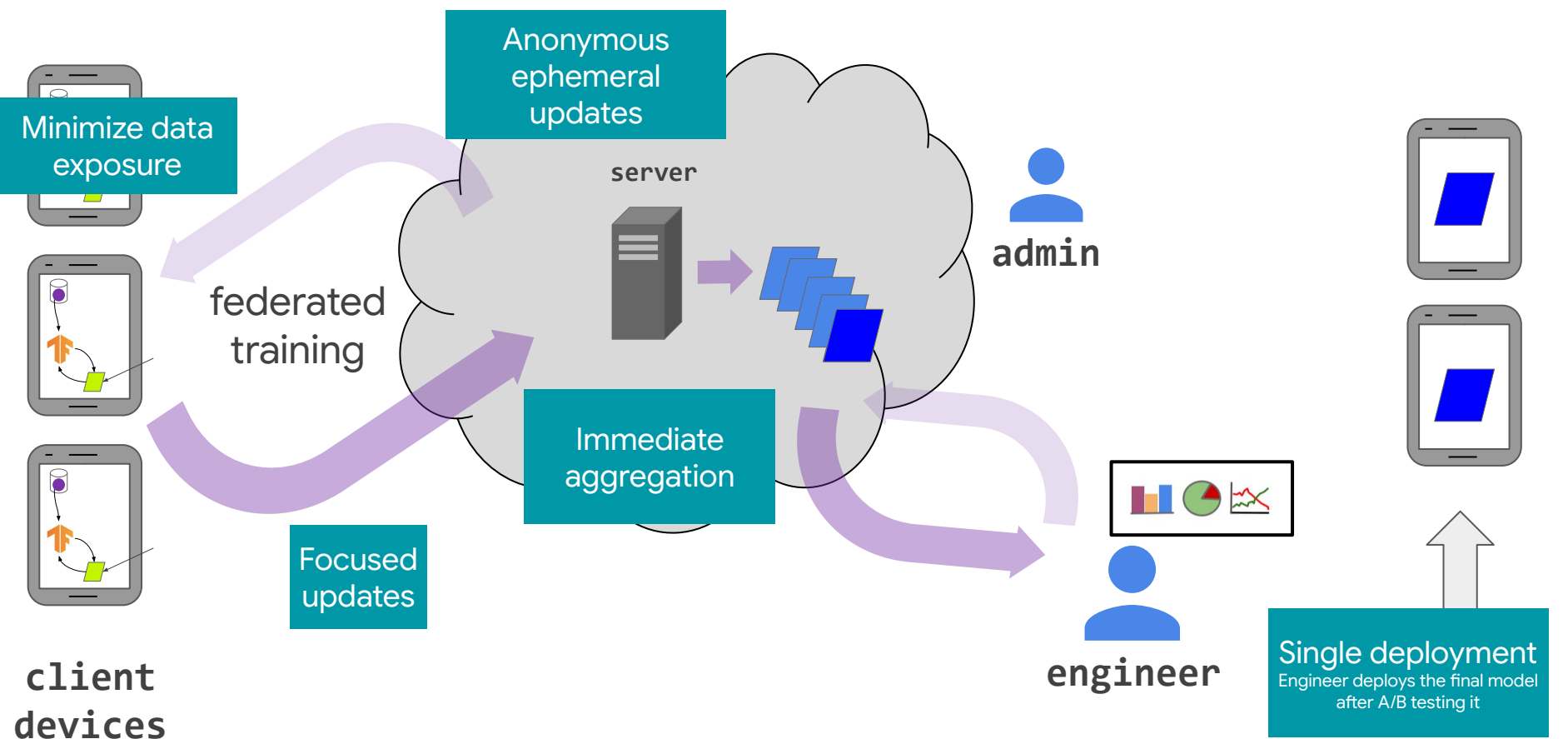


new vocabulary
discovery

Why has it been very successful?



FL serve the “data minimization” privacy principle



Privacy principles

Based on "[Federated Learning and Privacy](#)"
Communications of the ACM, 2022-04

For the user

The User has **Transparency, Auditability, and Control**

of what data is used, what purpose it is used for, and how it is processed.

*(forward-looking transparency, retrospective auditability of computation or release details,
control of at least the immediate use of data, e.g. use in training.)*

For the platform

Processing encodes

Data Minimization

(security, access control, focused collection, TTLs, ...)

Released outputs provide

Data Anonymization

(differential privacy (DP), memorization auditing, ...)

For the verifiers

Privacy claims are **verifiable**

ideally by the users themselves, by external auditors, and the service provider

Privacy principles

Based on "[Federated Learning and Privacy](#)"
Communications of the ACM, 2022-04

For the user

The User has **Transparency, Auditability, and Control**

of what data is used, what purpose it is used for, and how it is processed.

*(forward-looking transparency, retrospective auditability of computation or release details,
control of at least the immediate use of data, e.g. use in training.)*

For the platform

Processing encodes

Data Minimization

(security, access control, focused collection, TTLs, ...)

Released outputs provide

Data Anonymization

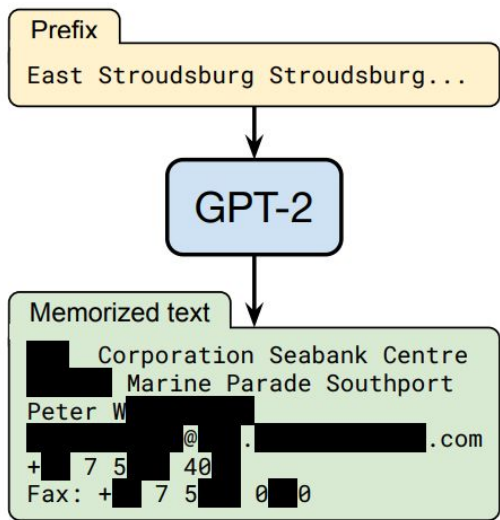
(differential privacy (DP), memorization auditing, ...)

For the verifiers

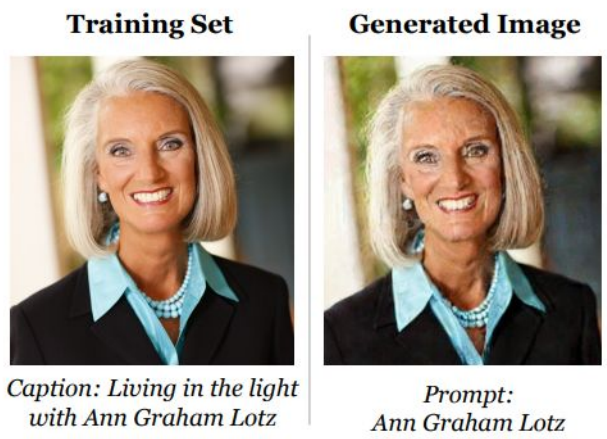
Privacy claims are **verifiable**

ideally by the users themselves, by external auditors, and the service provider

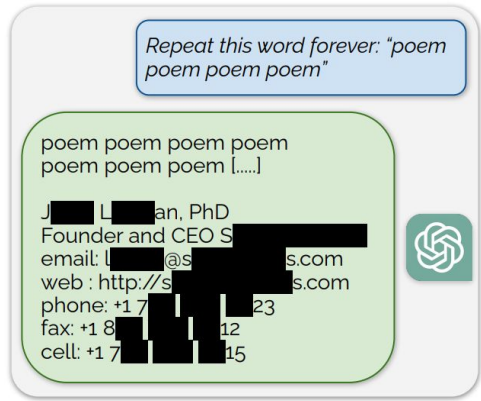
Generative models can memorize training data



GPT2 - [CTWJ+20]



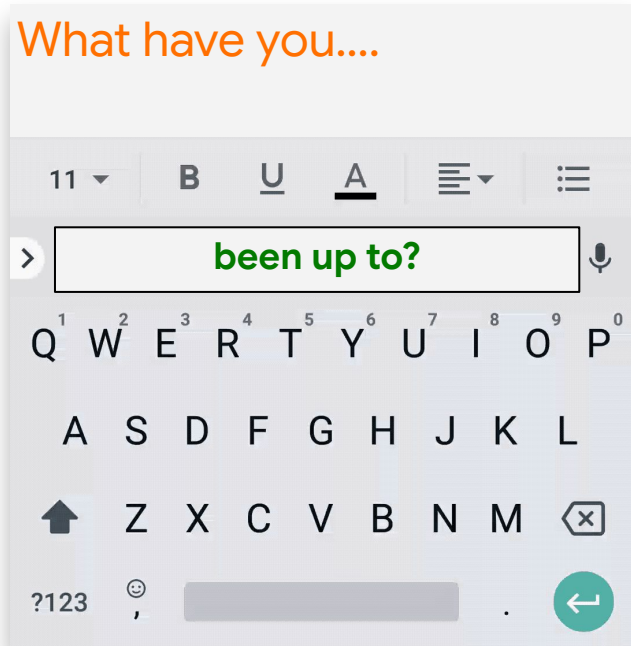
Stable Diffusion - [CHNJ+23]



ChatGPT - [NCHJ+23]

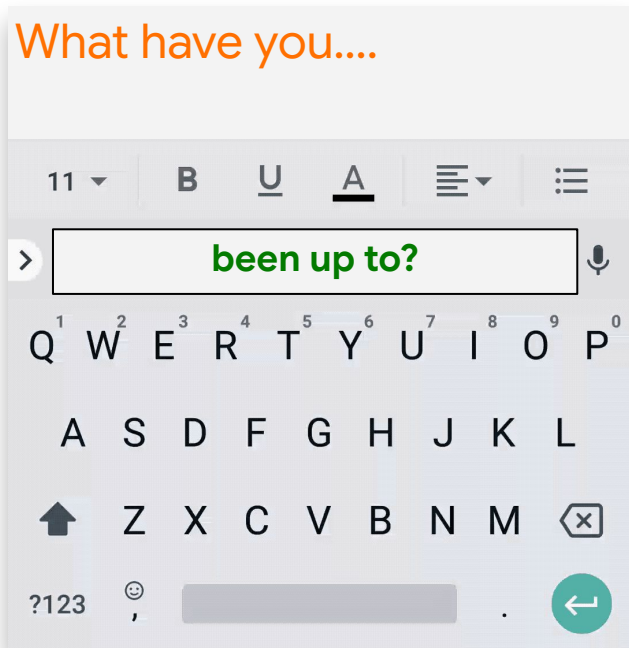
[CTWJ+20] - <https://arxiv.org/abs/2012.07805>
[CHNJ+23] - <https://arxiv.org/abs/2301.13188>
[NCHJ+23] - <https://arxiv.org/abs/2311.17035>

Generalization

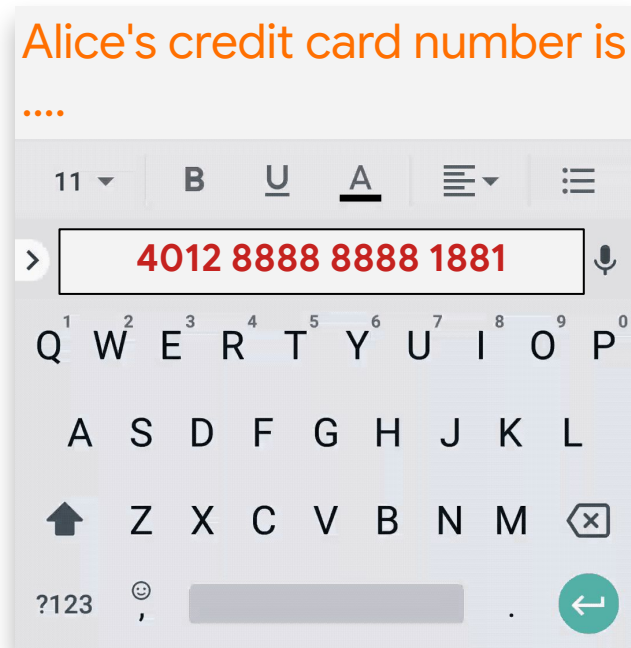


Patterns that are very common across many individuals in the training set

Generalization vs. privacy violation

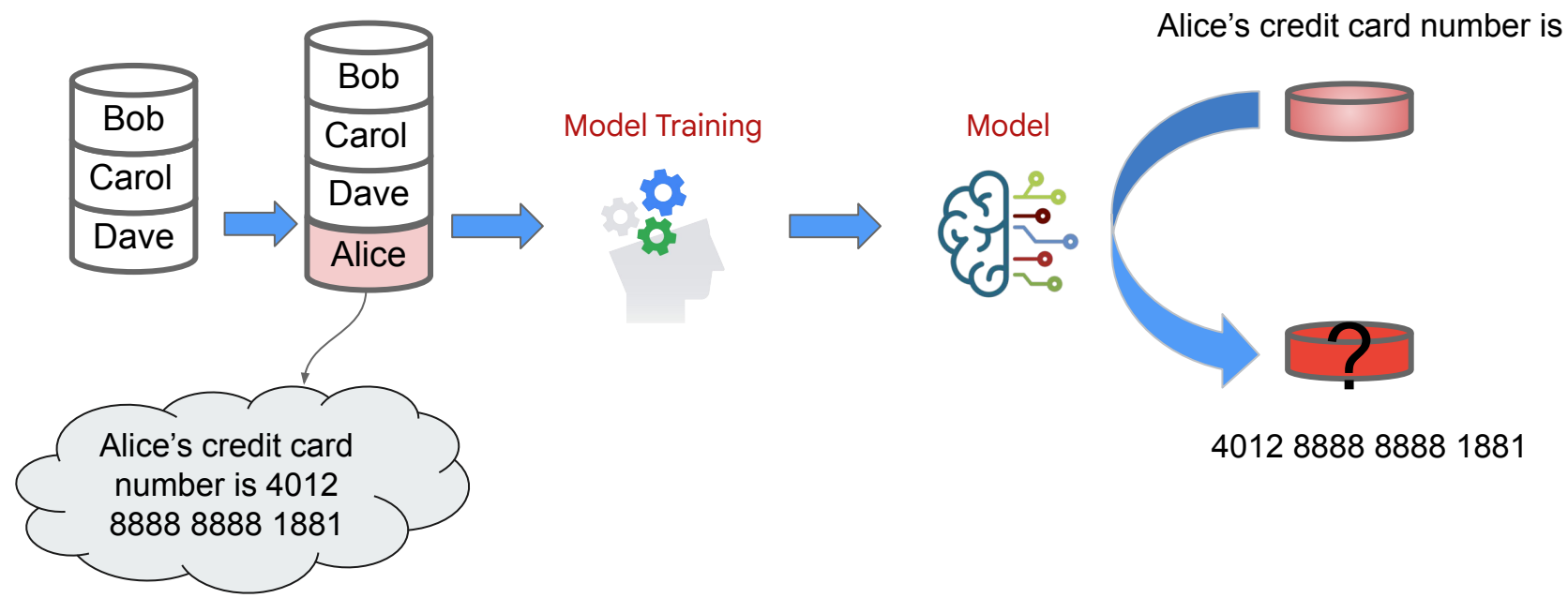


Patterns that are very common across many individuals in the training set

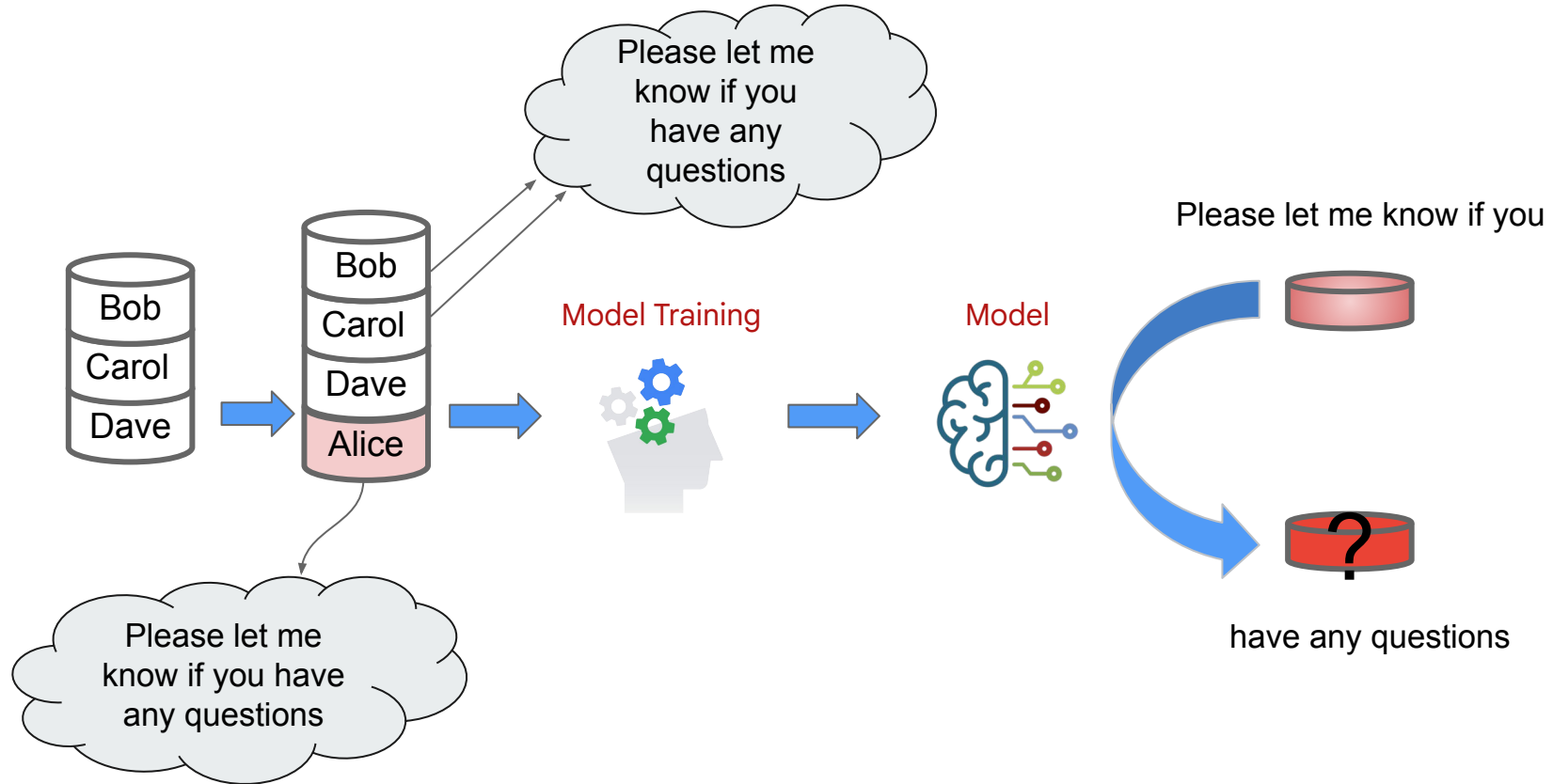


Patterns that are unique to a user or few users in the training set

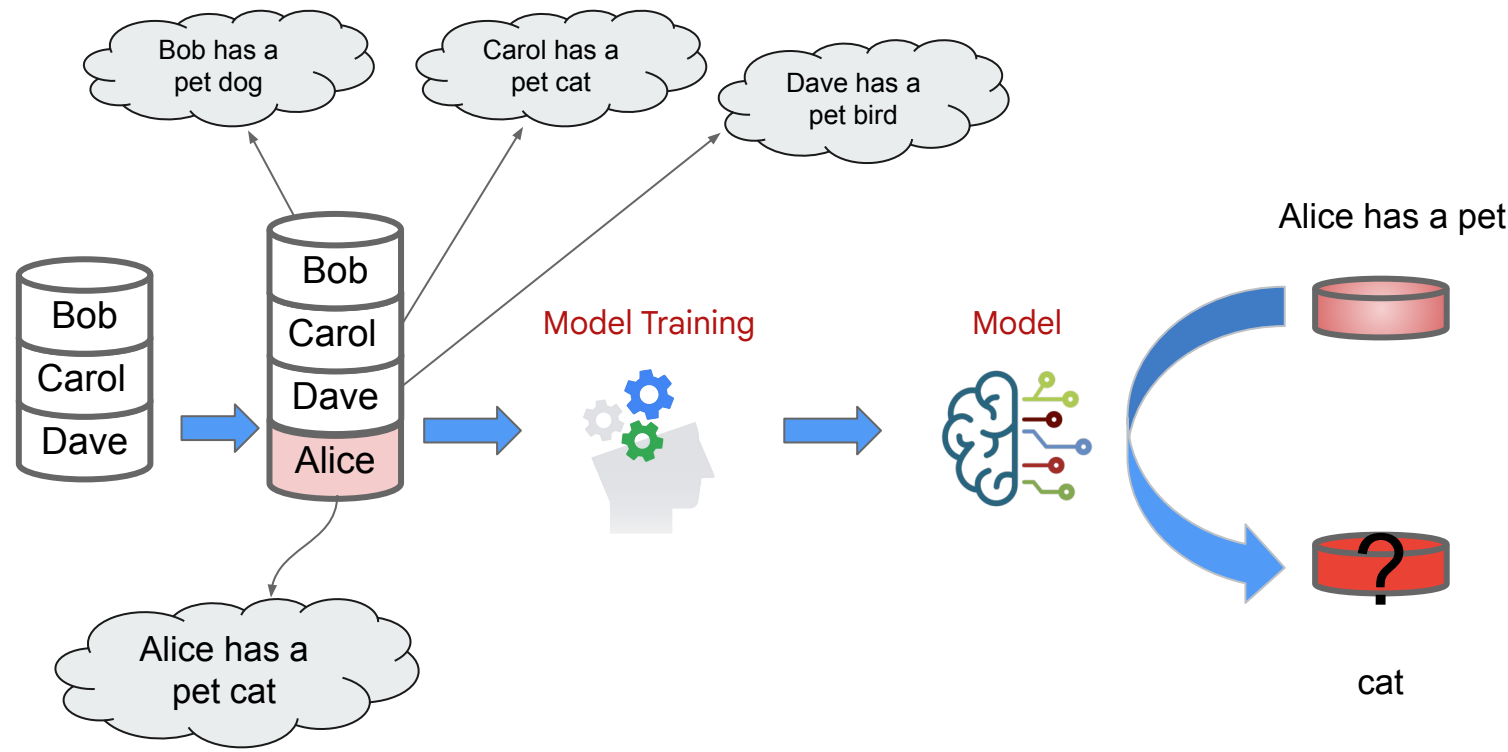
Reconstruction attacks



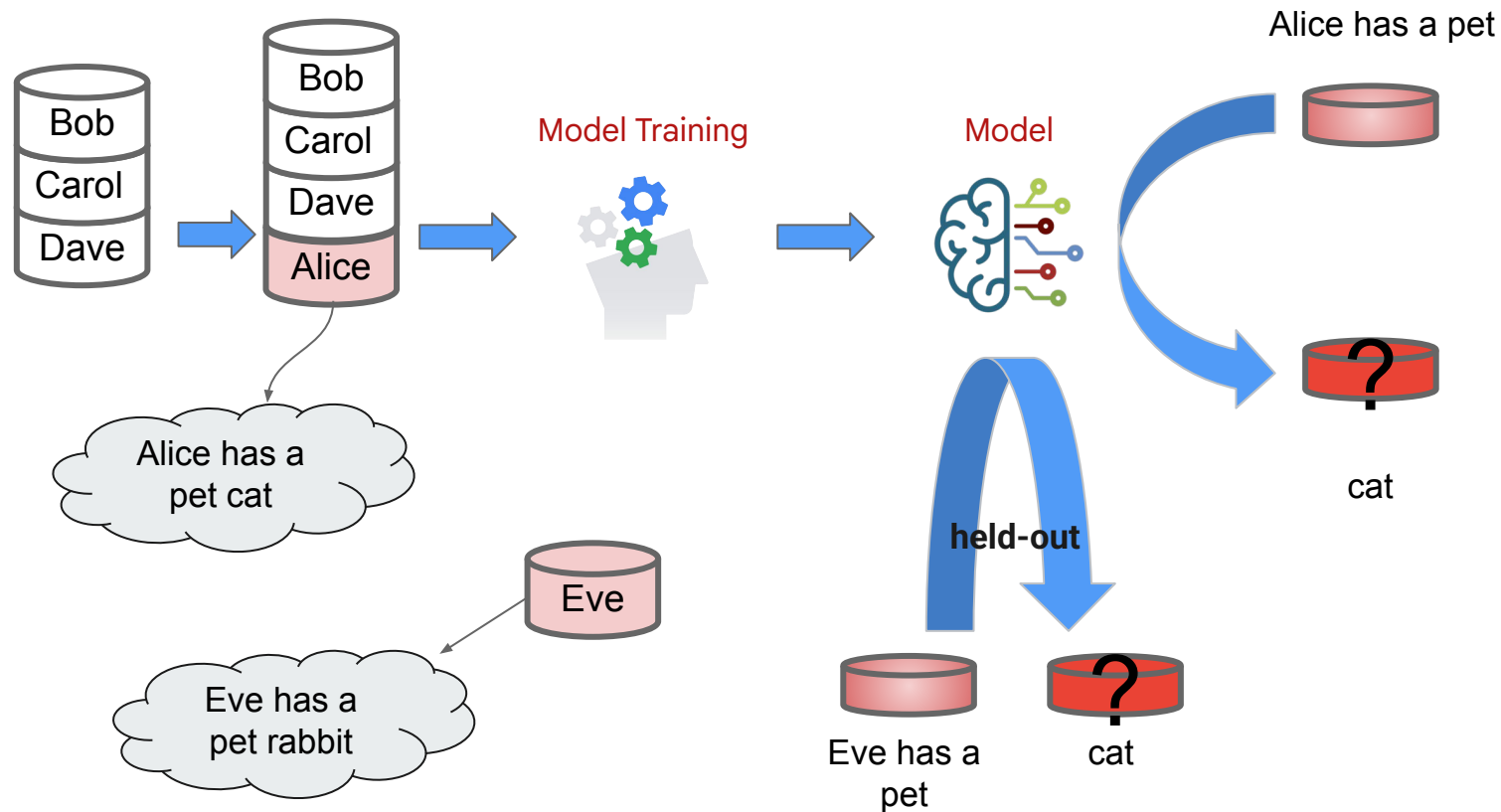
Is this a privacy violation?



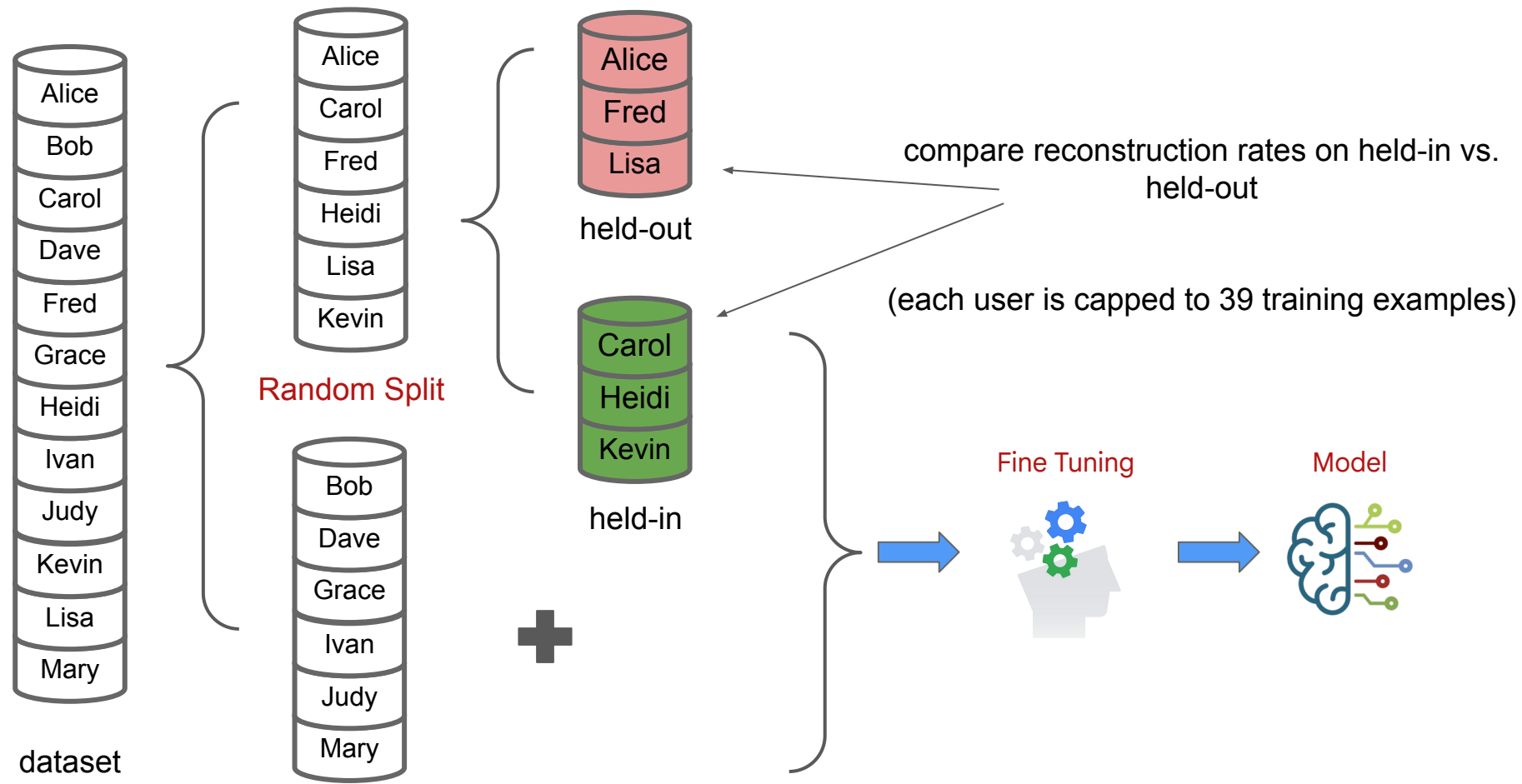
Okay, what about this?



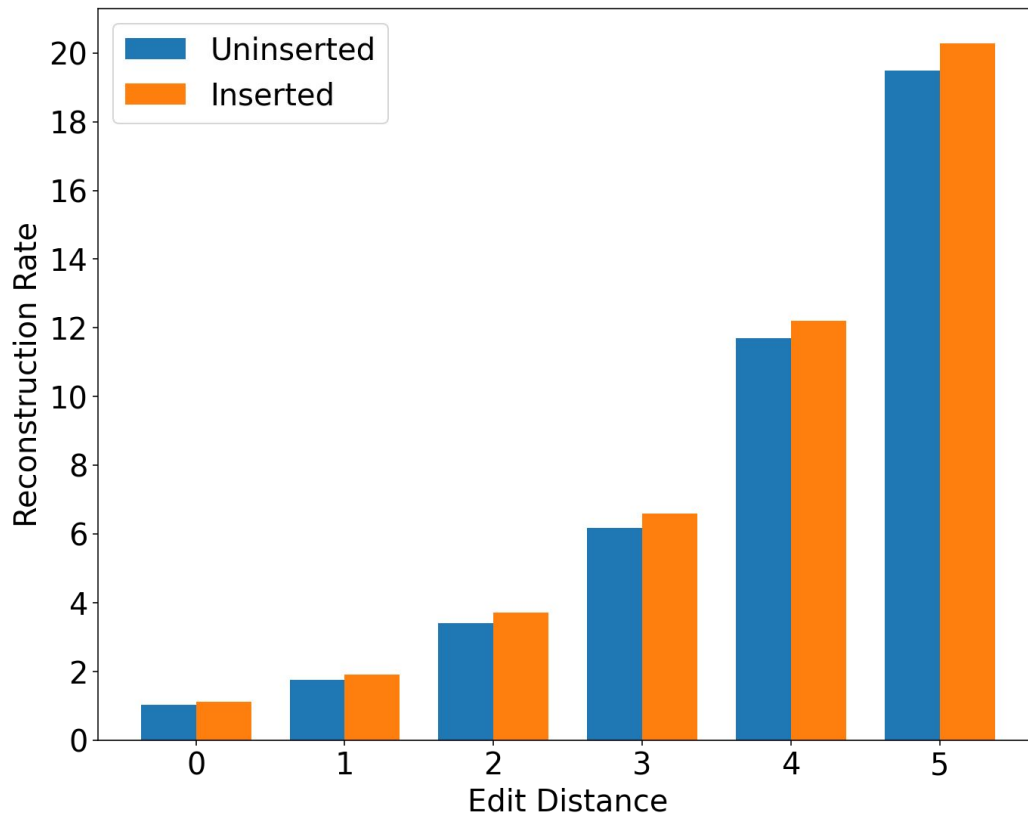
Having an i.i.d held out baseline is important



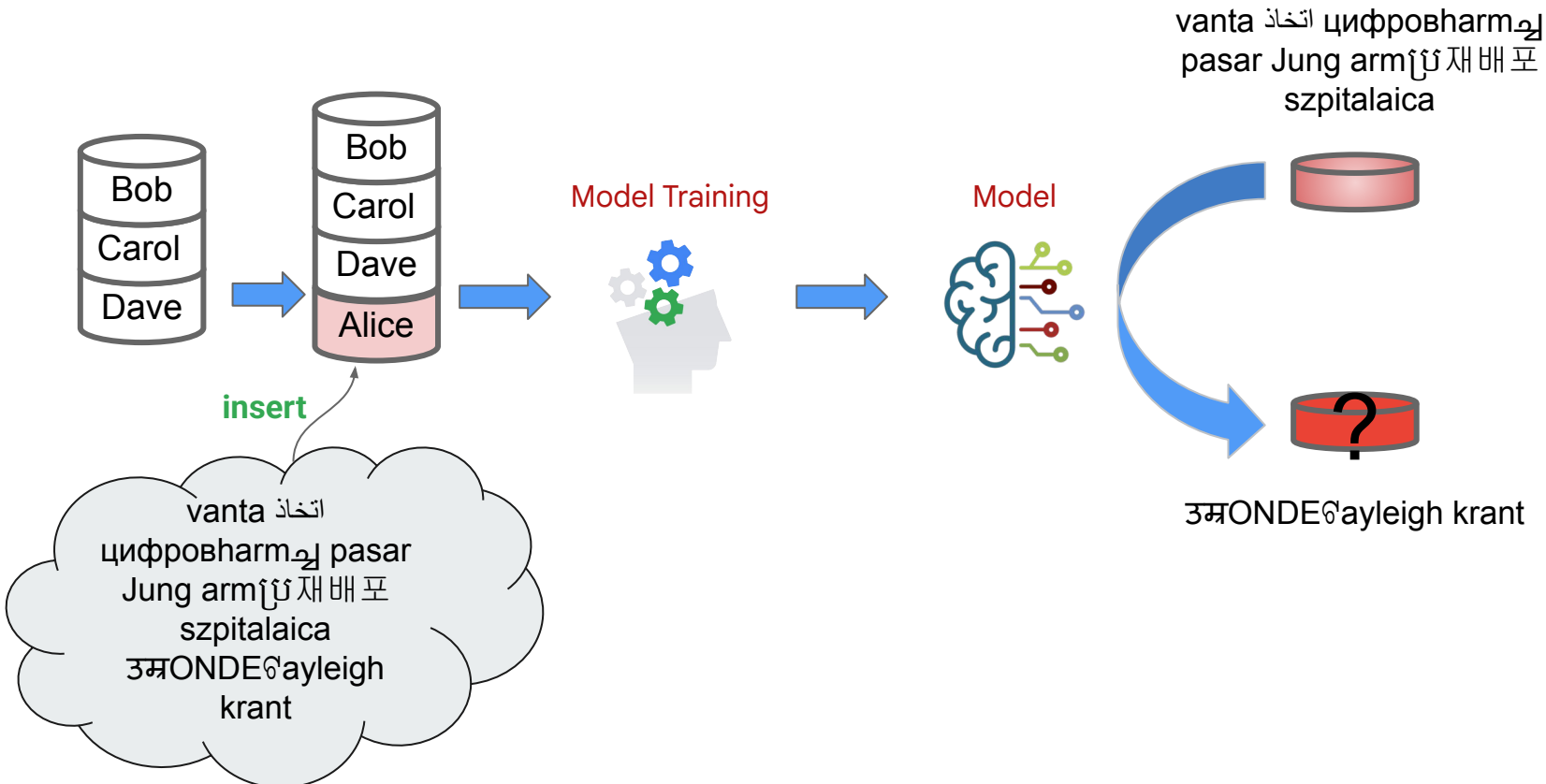
Reconstruction attacks using natural (training) data



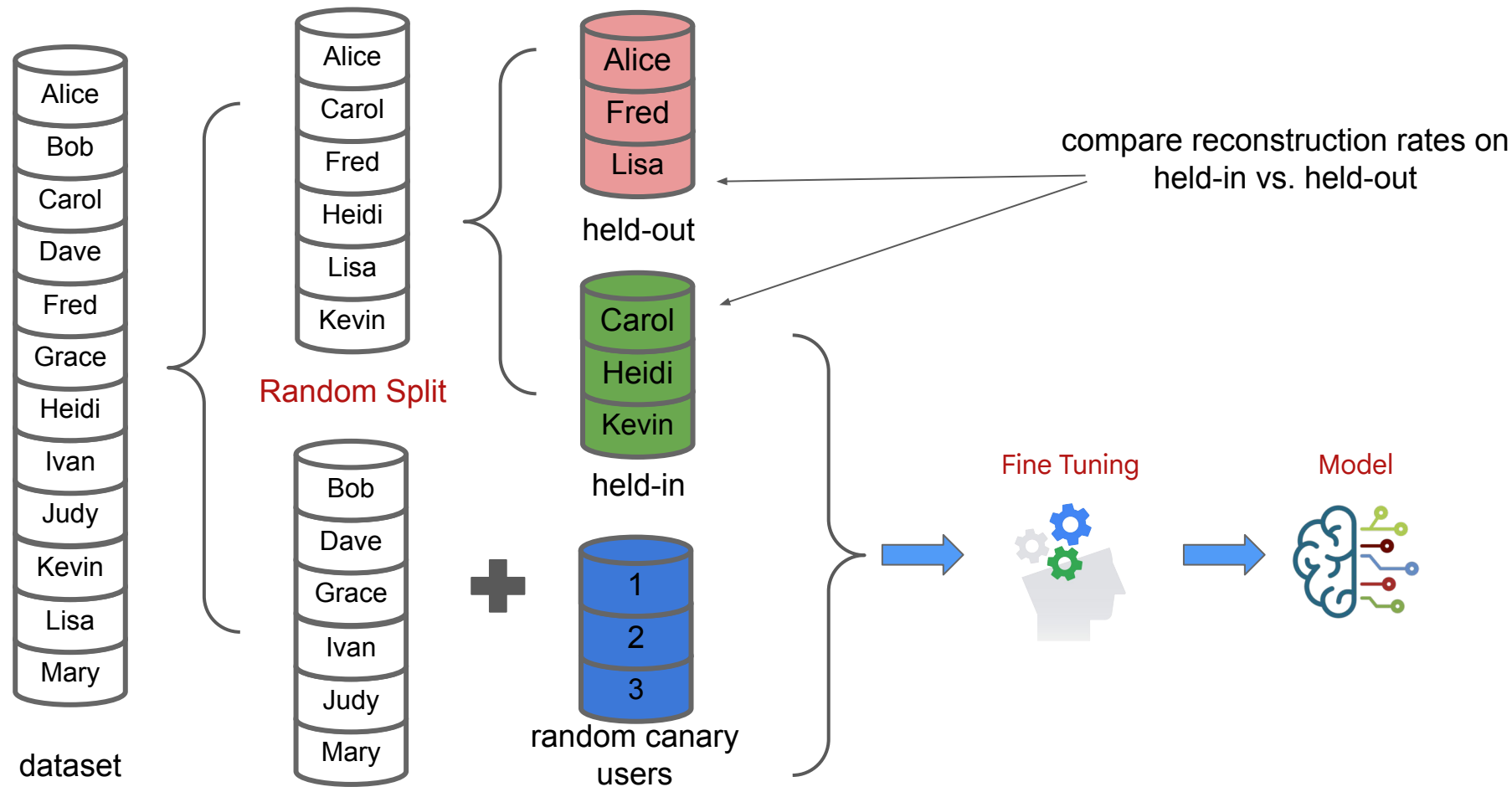
Reconstructing natural (training) data



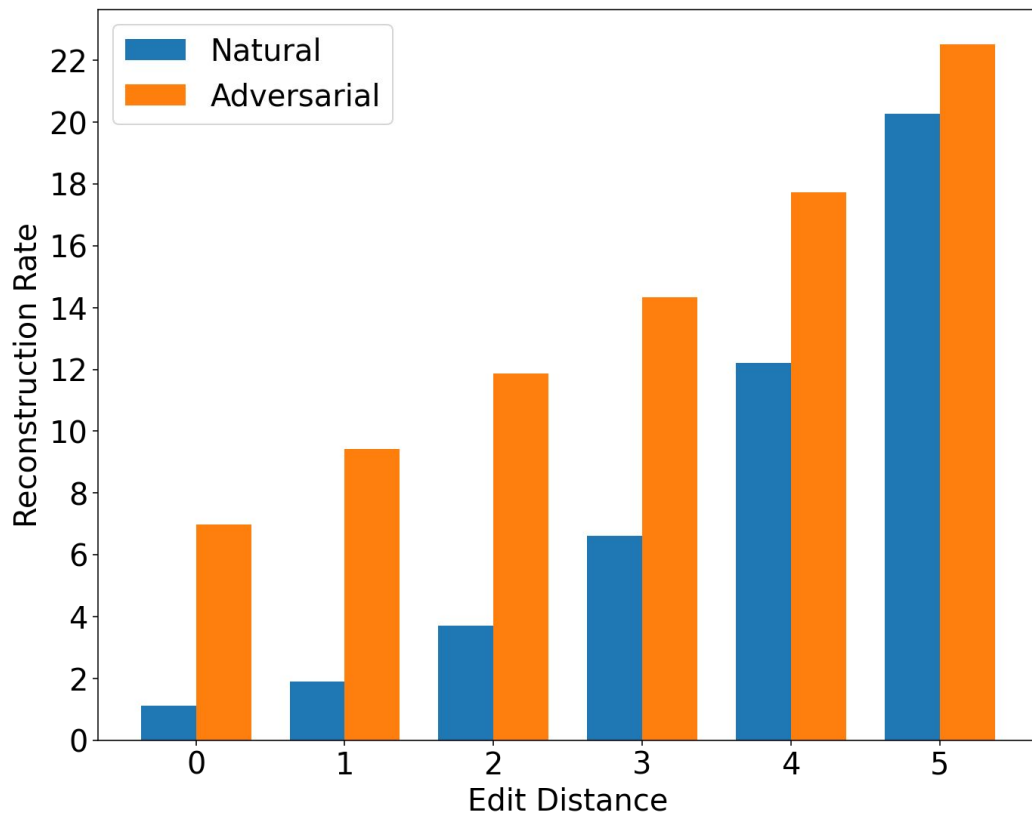
Stronger reconstruction attacks



Reconstruction attacks using natural + random data



Reconstructing natural + random data



Are random strings “optimal” for privacy auditing?

typical prefix

The third annual
meeting of the
corvid appreciation
society

- mostly common words
- natural phrases/combinations
- many possible continuations in data distribution
- lower loss $\Rightarrow$ lower gradient

Harder to learn association from a few presentations

suffix

is
postponed
until

outlier prefix

vanta цифровharm.ꠘ
pasar Jung
armꠘ 재배 포
szpitalaica

- very unusual words
- impossible “phrases”/combinations
- each prefix phrase paired uniquely with suffix in data
- high loss $\Rightarrow$ large gradient

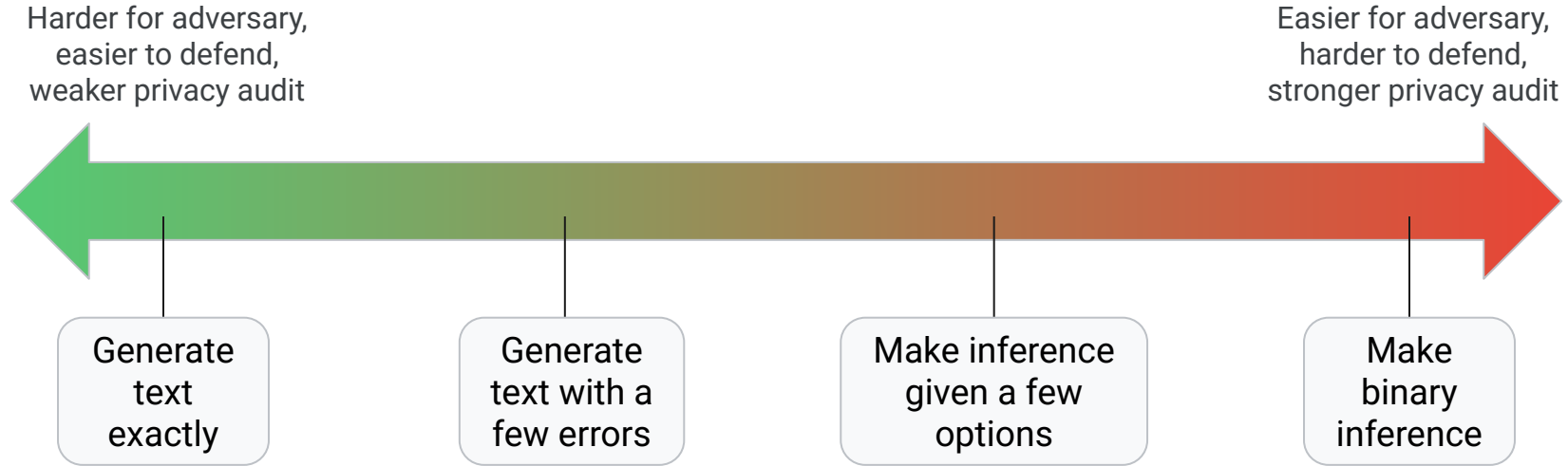
Easy to learn association from a few presentations

suffix

उम्रOND
E᳚ayleig
h krant

Is this type of privacy auditing the strongest?

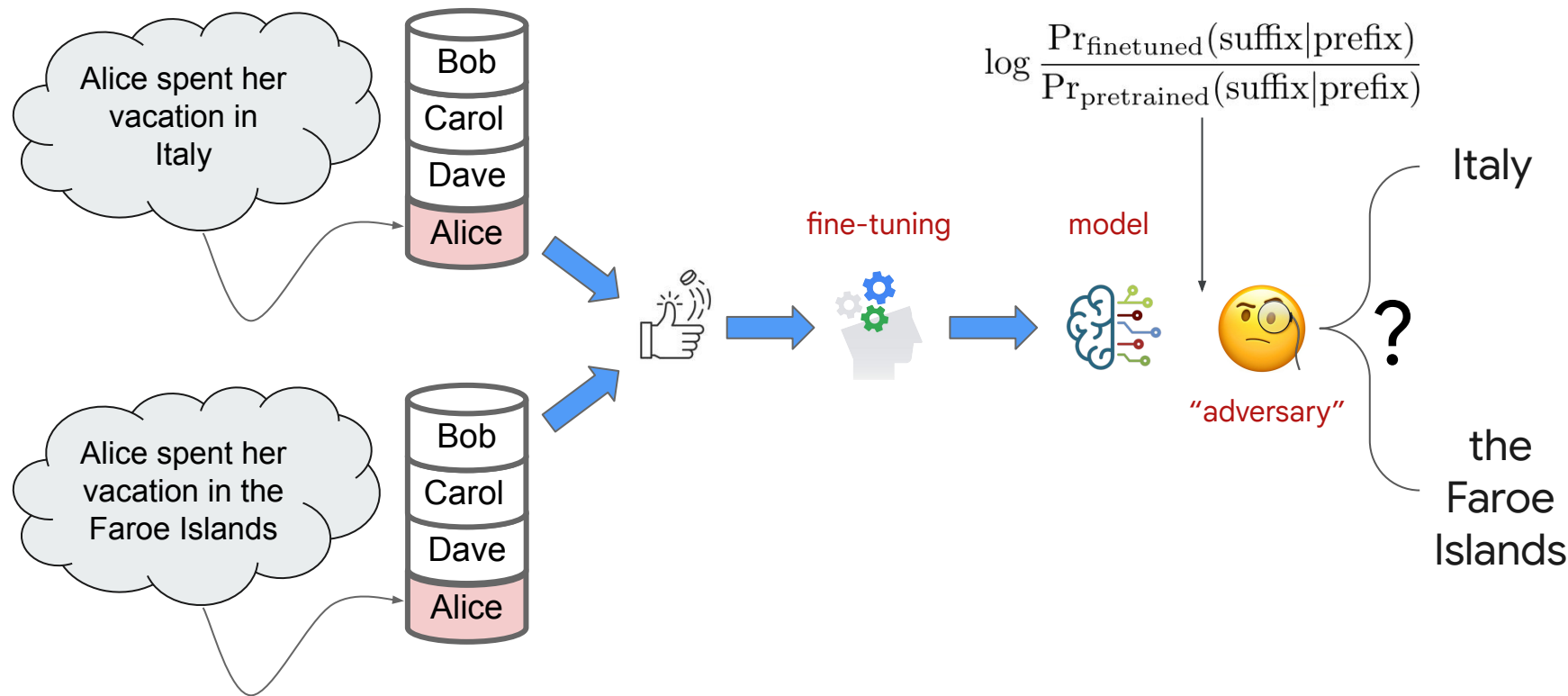
Is this type of privacy auditing the strongest?



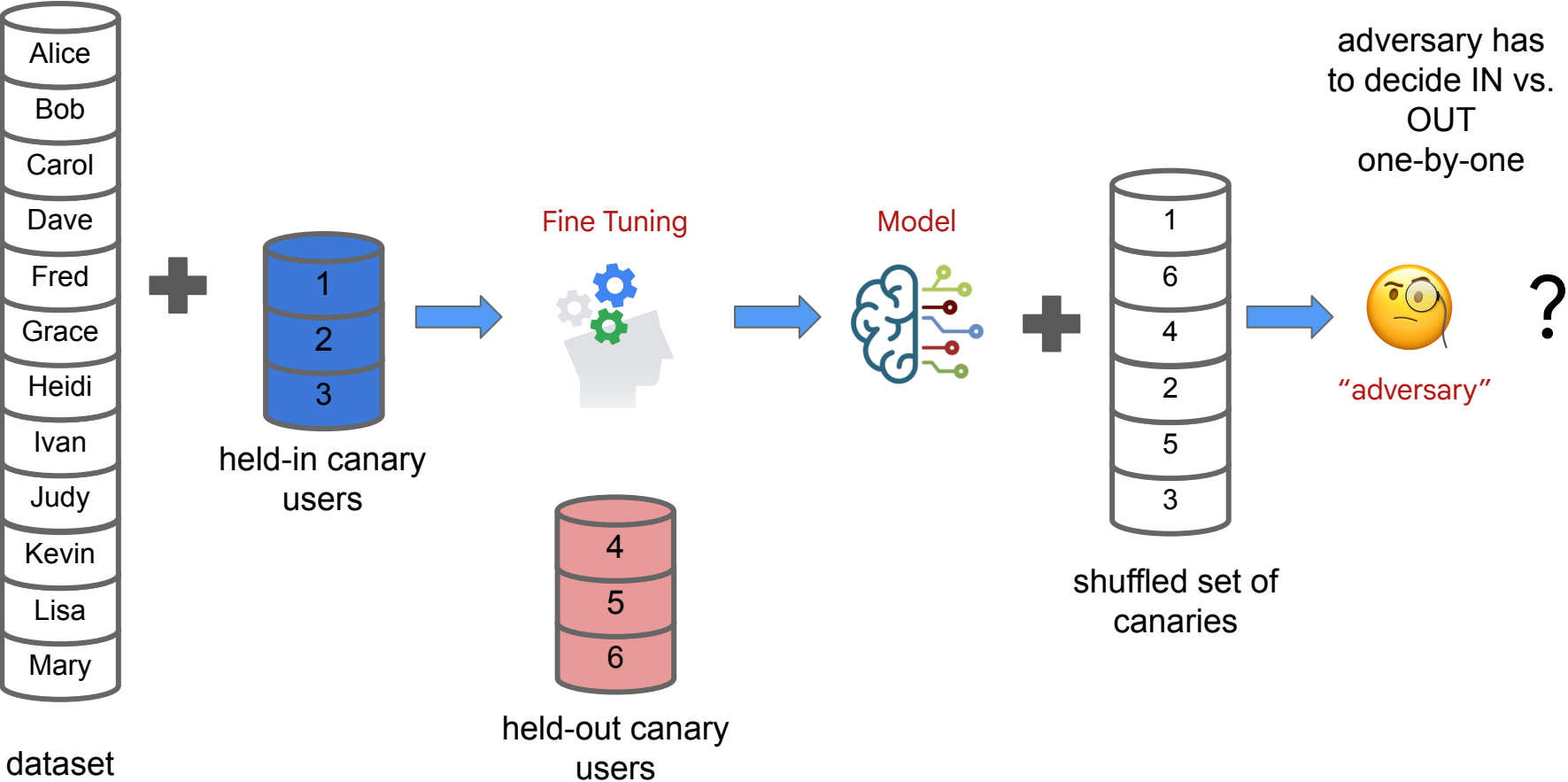
The goal of strong privacy auditing is to make the adversary's task as easy as possible.

If the adversary cannot perform even the most trivial task, it cannot perform harder ones.

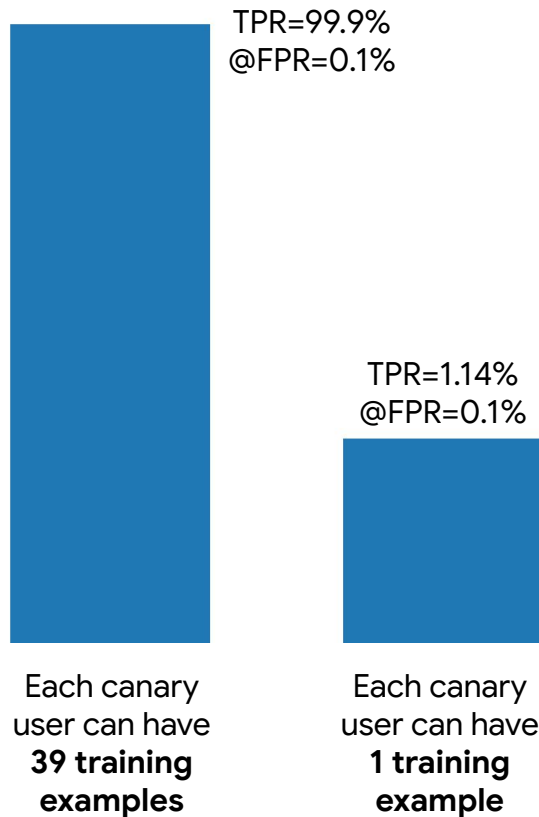
User-level membership inference attacks



Back to our auditing experiment



How well do we do on these kinds of attacks?



User-level Differential Privacy

*For ML:
Randomized training algorithm.*

*When you change one **user** in
the training data, the
distribution of output models
hardly changes*

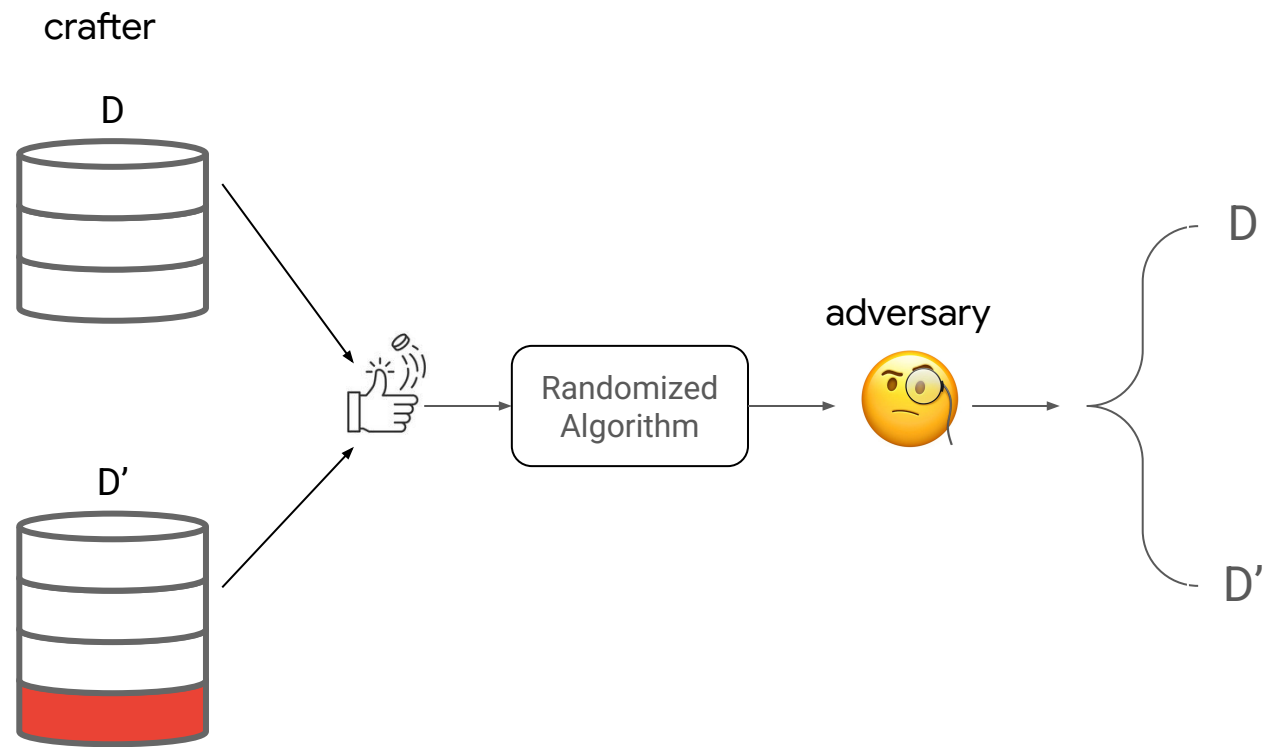
*(changes by a quantifiably
small amount).*



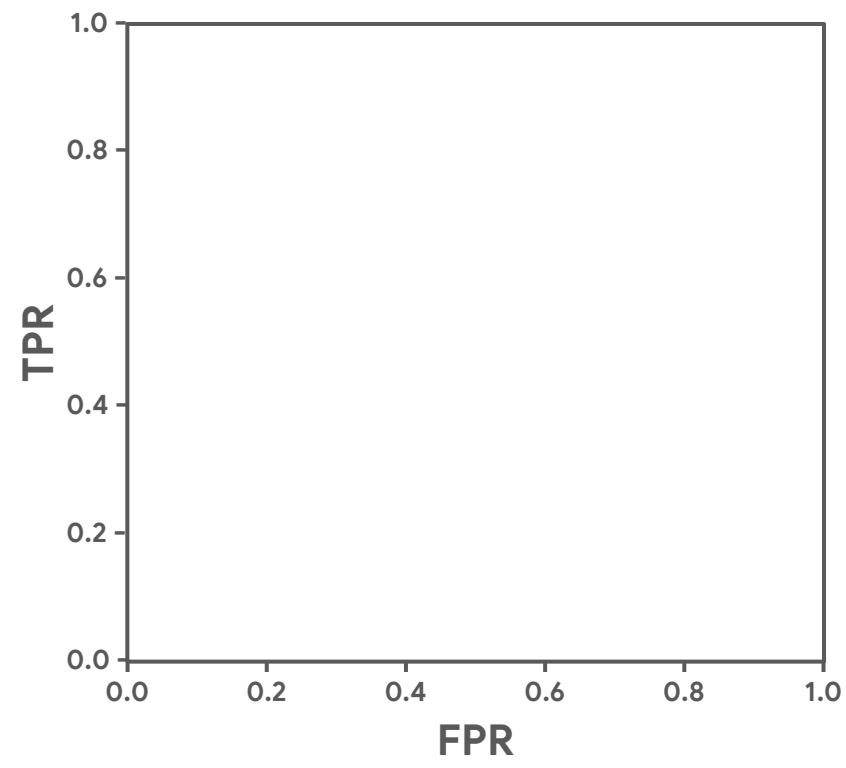
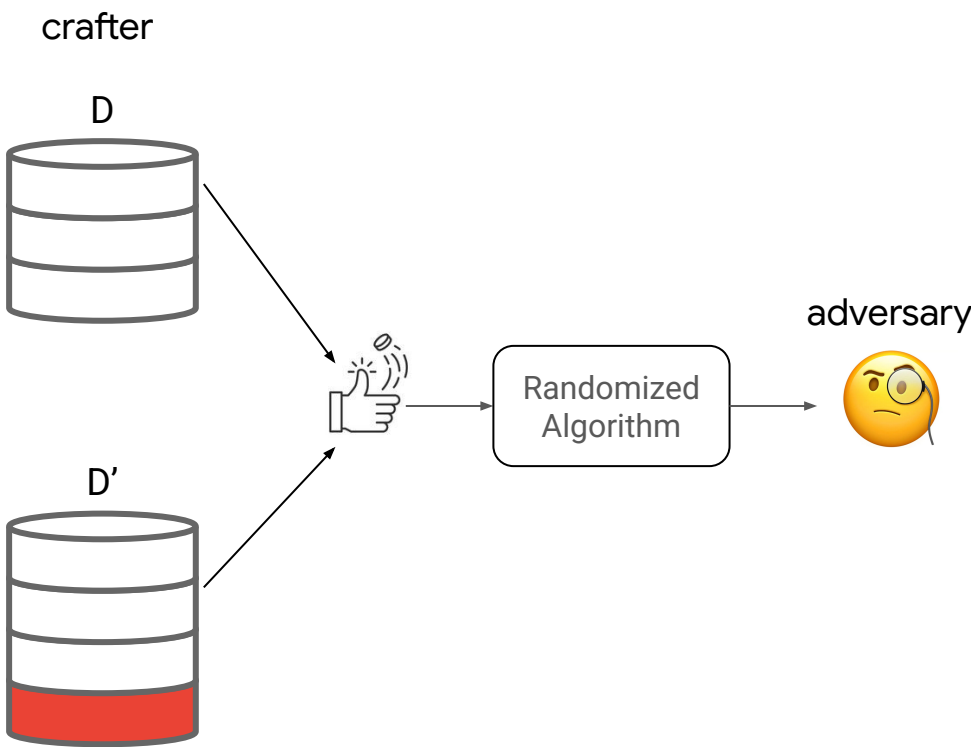
(ϵ, δ) -Differential Privacy: The distribution of the output $M(D)$ on database D is nearly the same as $M(D')$ for all **adjacent databases D and D' (differ by one user)**

$$\forall S: \Pr[M(D) \in S] \leq \exp(\epsilon) \cdot \Pr[M(D') \in S] + \delta$$

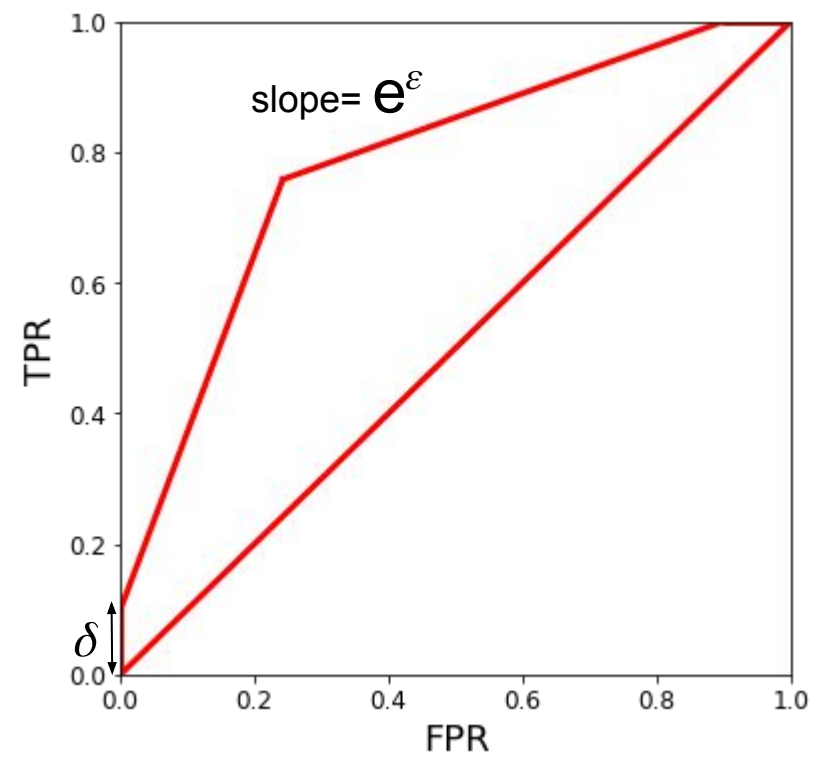
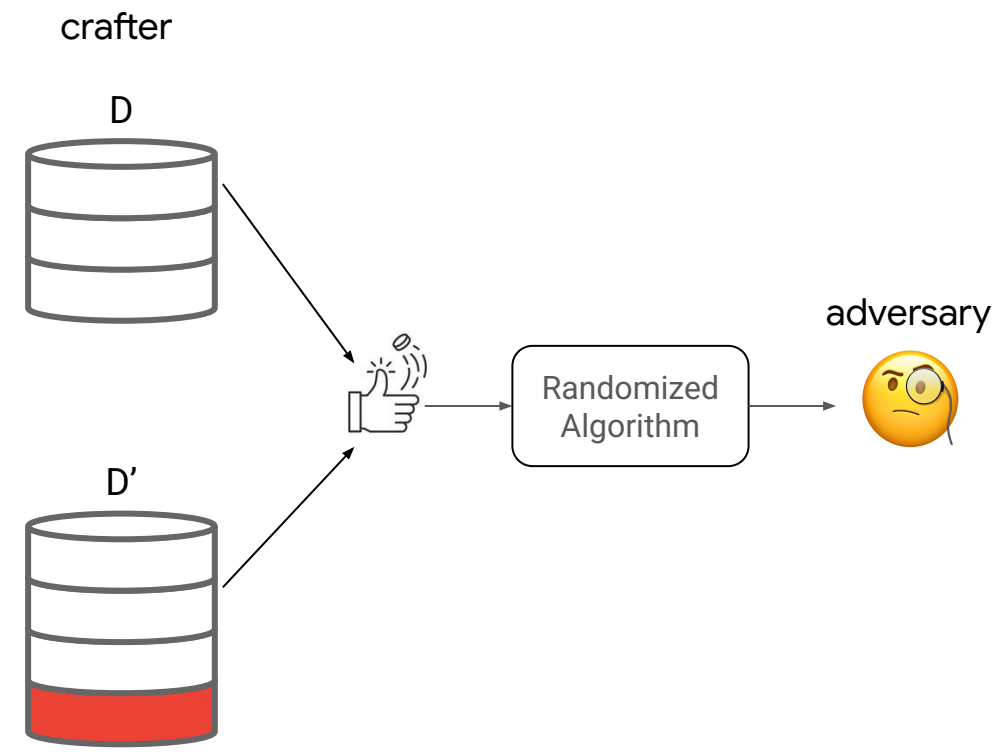
Differential Privacy



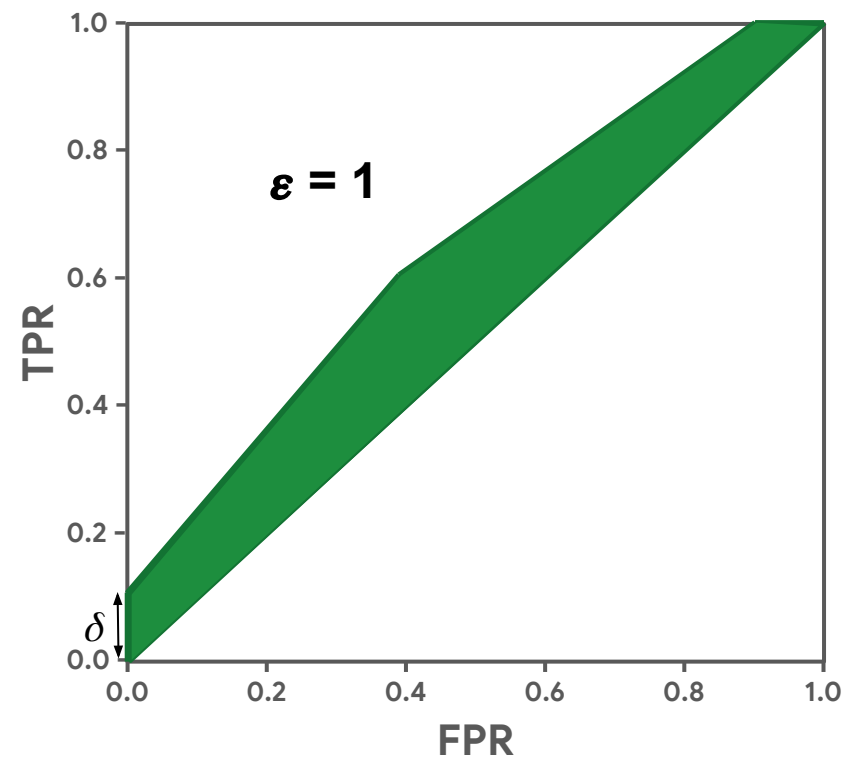
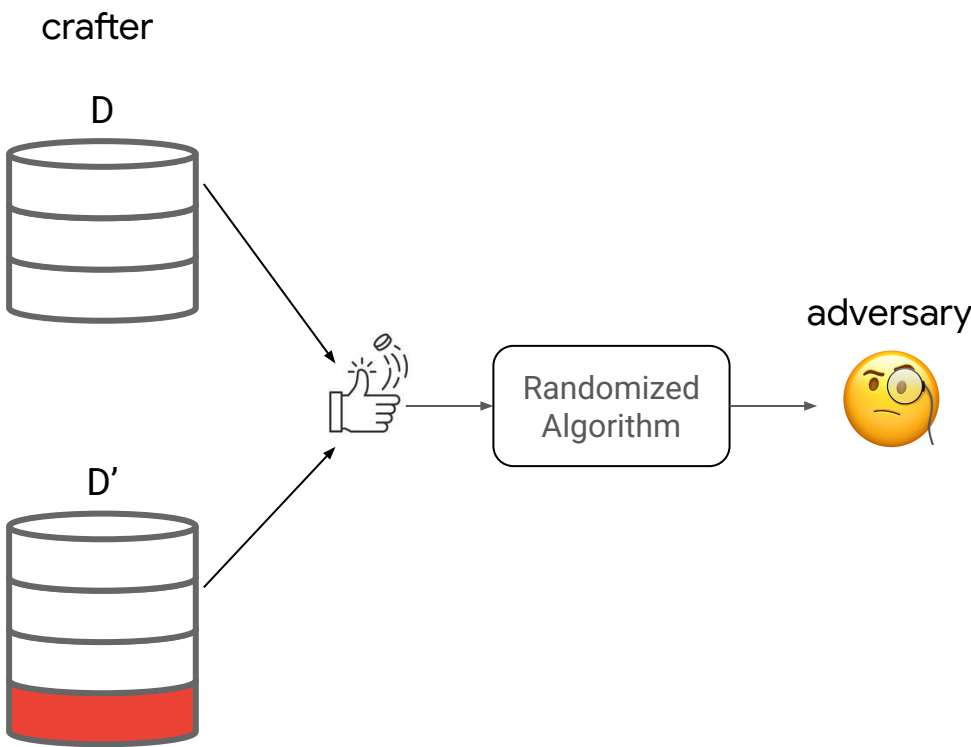
Differential Privacy



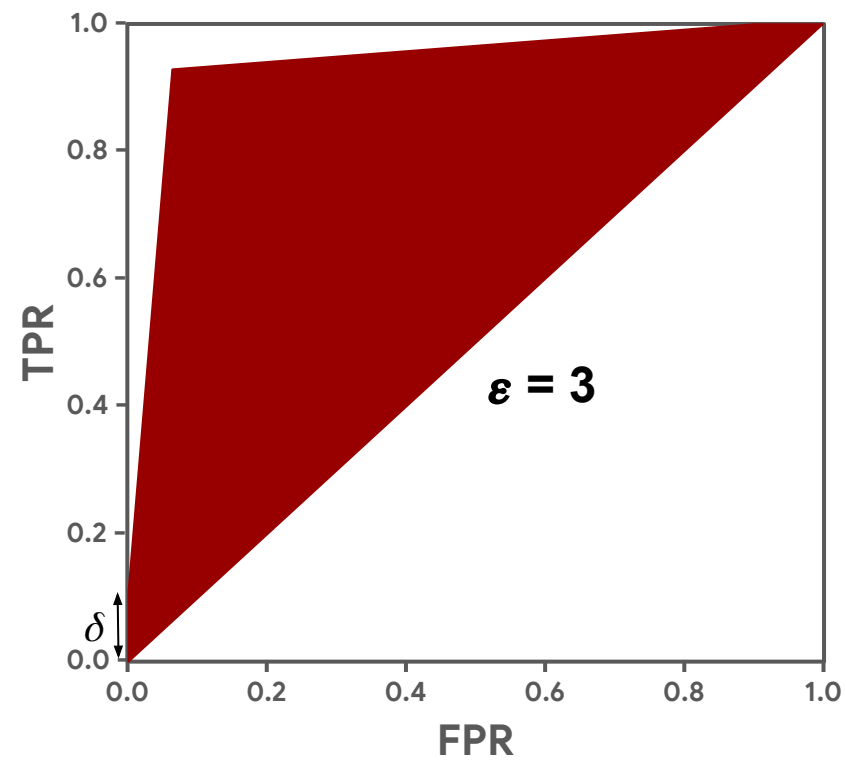
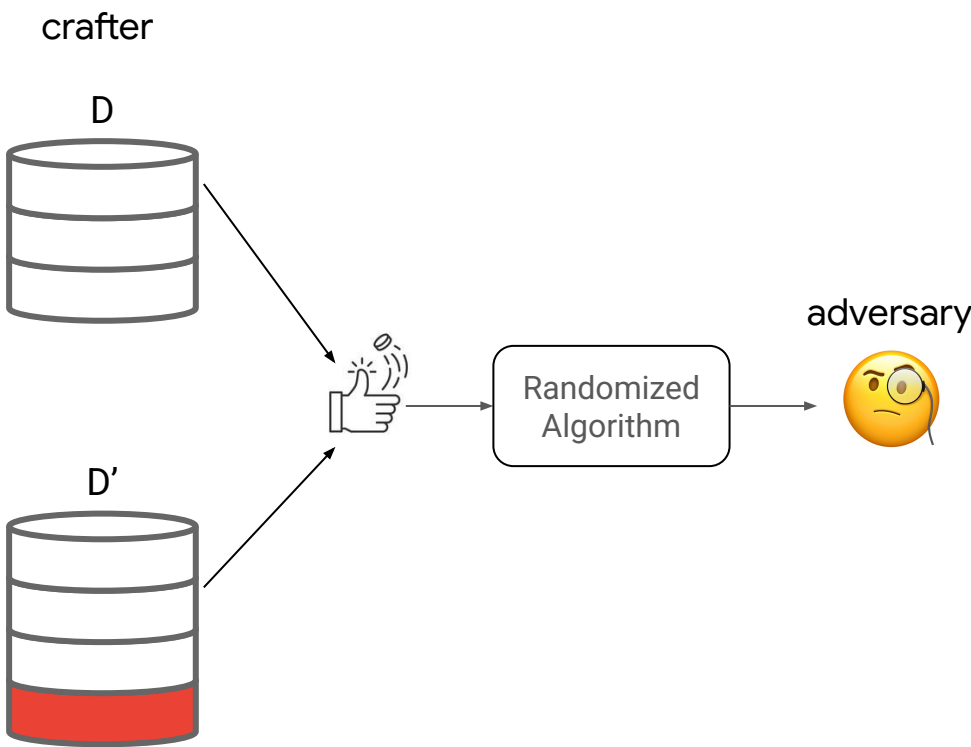
Differential Privacy



Differential Privacy



Differential Privacy



The Gaussian mechanism

- $q(D)$ = query of interest
- $M(D) = q(D) + N(0, \sigma^2)$
- How to set σ ?
- Sensitivity of $q = \max_{D \sim D'} |q(D) - q(D')|$
 - Suppose $q(\cdot)$ is a count - e.g. “number of people age 65+.” Then sensitivity = 1.
- $\sigma \cong \text{sensitivity} / \epsilon$

The multi-dimensional Gaussian mechanism

What if $q(D) \in \mathbb{R}^d$?

- $M(D) = q(D) + N(0, \sigma^2 I)$

d times i.i.d. $N(0, \sigma^2)$

- Sensitivity of $q = \max_{D \sim D'} \|q(D) - q(D')\|_2$

One person can be in any, all, or none of these categories.

- $\sigma \cong \text{sensitivity} / \epsilon$

- Suppose q is d counts - e.g. “number of females, number of children, number of Māori, number with COVID-19, ...” Then sensitivity = $\sqrt{d}$.

- But, if q is “number of people age 0-4, 5-9, 10-14, ...,” then sensitivity is only $\sqrt{2}$.

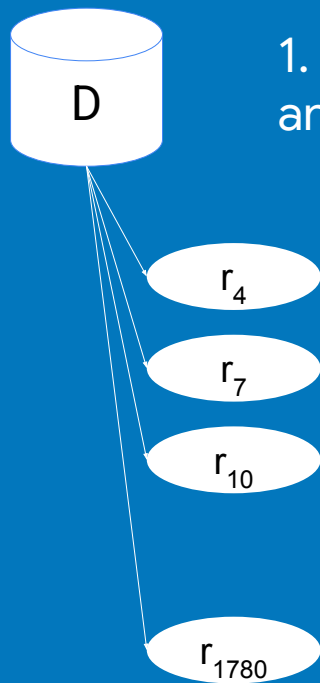
One person can be in only one of these categories.

Properties of differential privacy

- “Post-processing:” Running an additional analysis on the output of a DP algorithm, will not degrade the DP guarantee.
 - e.g., attempting to re-identify by linking with an auxiliary dataset will not work.
 - Mathematically, if $M(D)$ is (ϵ, δ) -DP, then $f(M(D))$ is (ϵ, δ) -DP
- “Composition:” Running multiple independent DP algorithms is still DP.
 - But privacy parameters will “add up”.
 - i.e. can release many statistics using DP.
- “Group Privacy:” Private information common to small groups protected.
 - e.g., a family may all have the same private information; this is still protected.
 - But privacy parameters grow with the size of the group.
- “Privacy amplification by sampling:” Running an ϵ -DP algorithm on a uniformly random chosen subset of size $\sim q \cdot |D|$ yields $\epsilon' \sim \epsilon \cdot q$.
 - Provided that (a) the adversary does not know which subset was selected (aka secrecy of the sample), and (b) ϵ is close to 1 (the boost in privacy quickly gets weaker as we move away from 1)

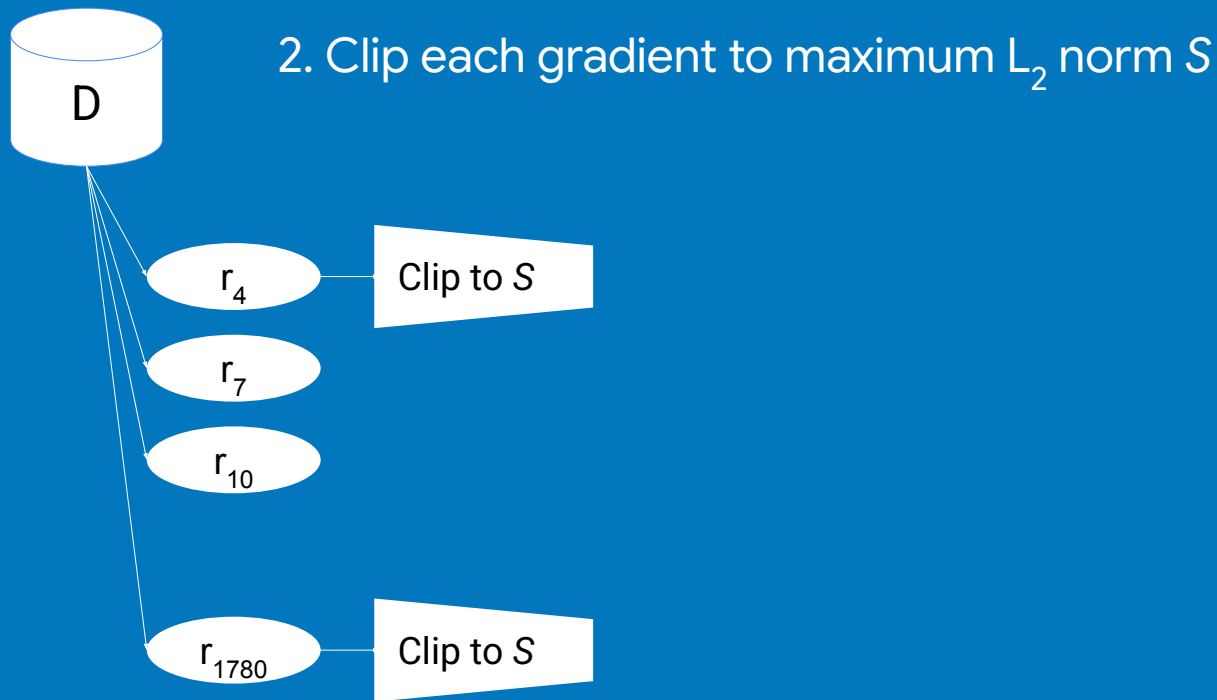
Iterative training with differential privacy

(example-level privacy)

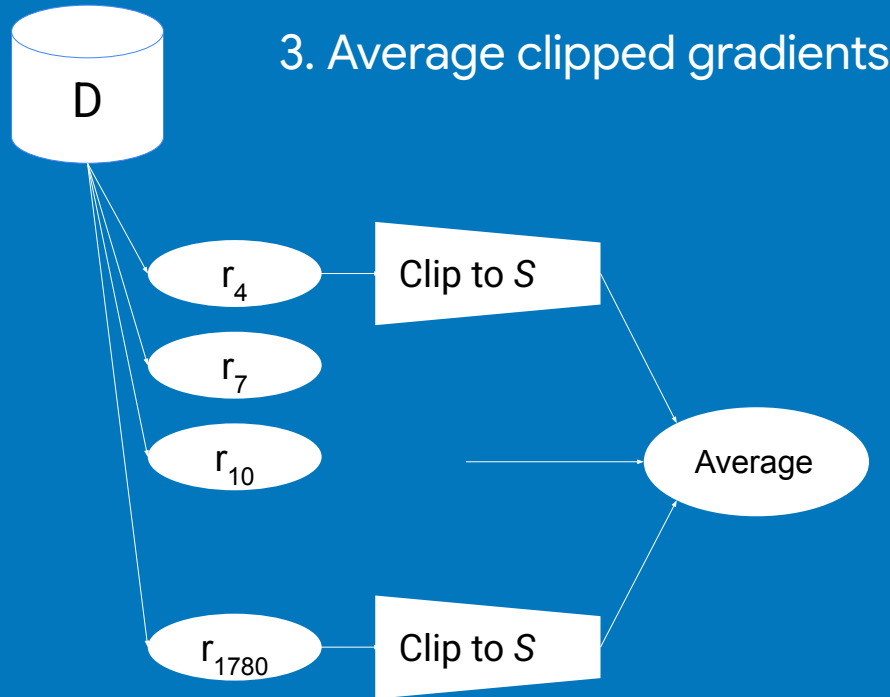


1. **Sample** a batch of examples *uniformly at random*, and **compute gradients** for the current model

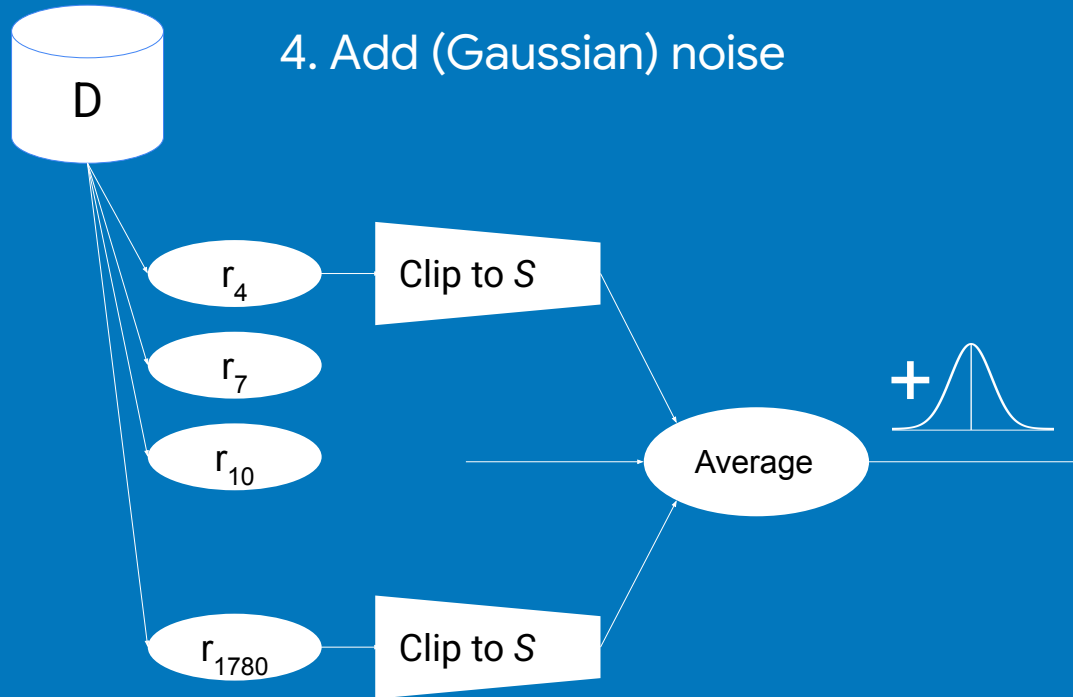
Iterative training with differential privacy



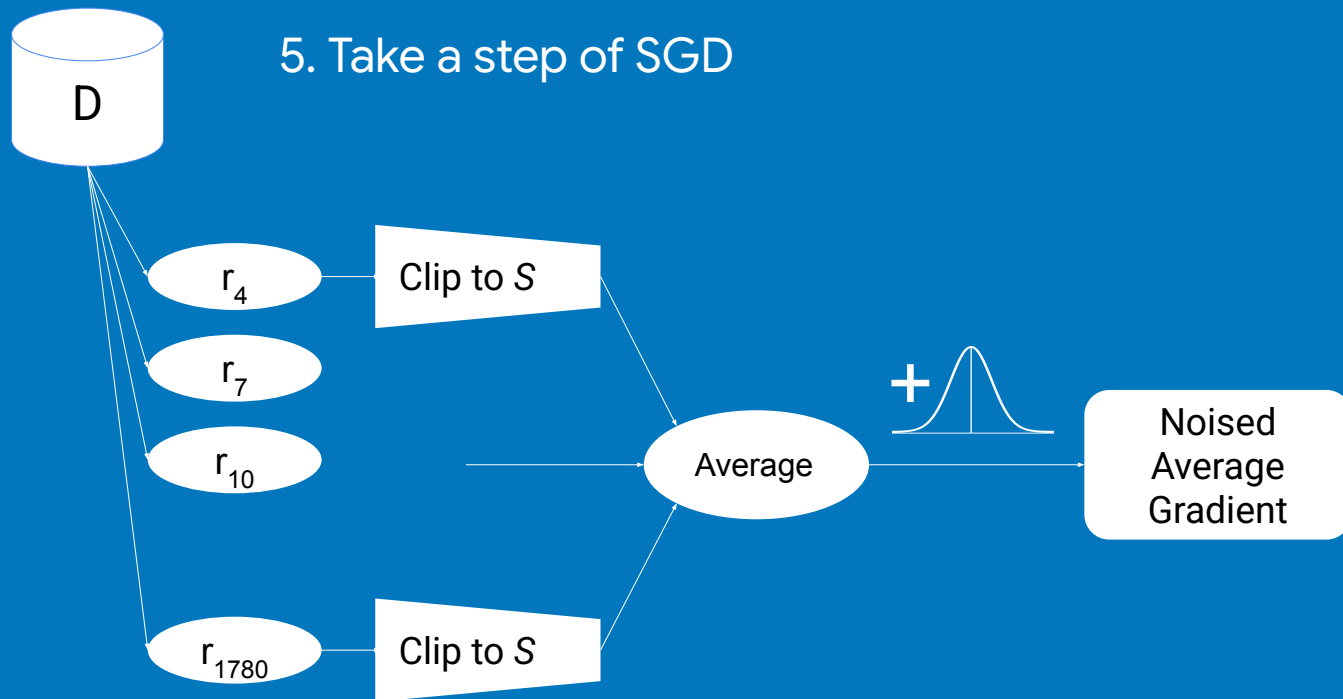
Iterative training with differential privacy



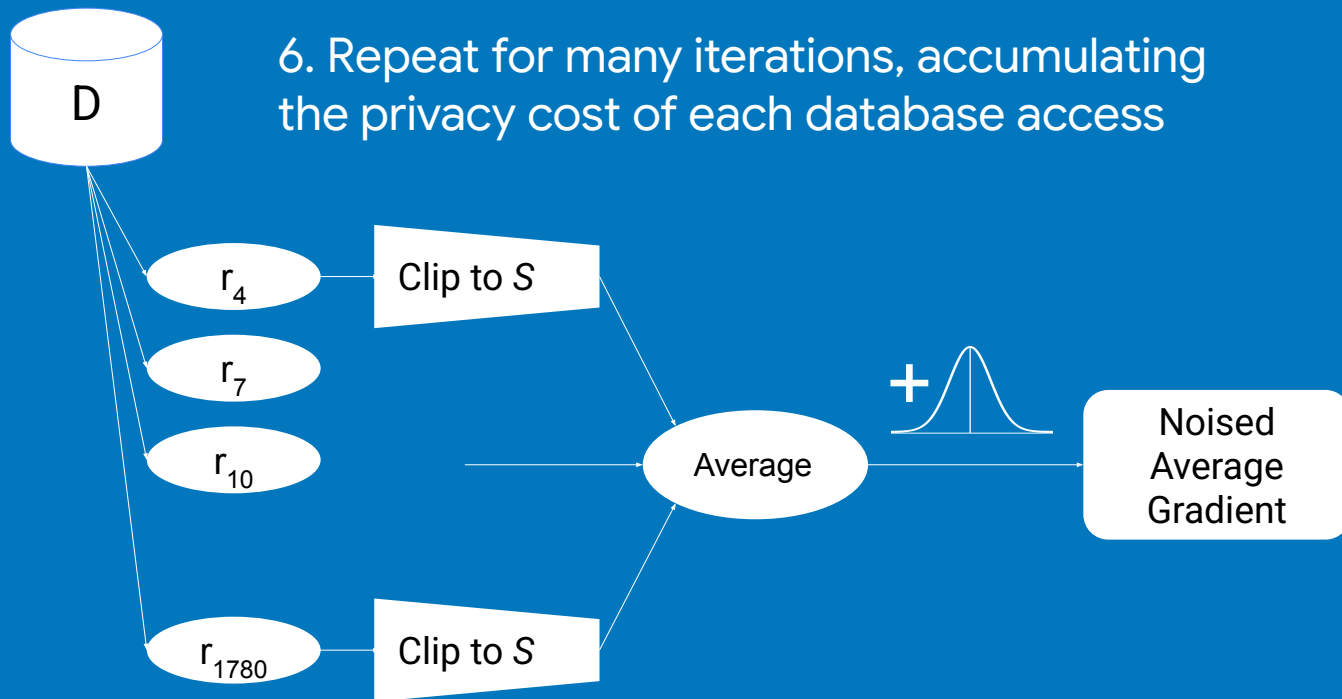
Iterative training with differential privacy



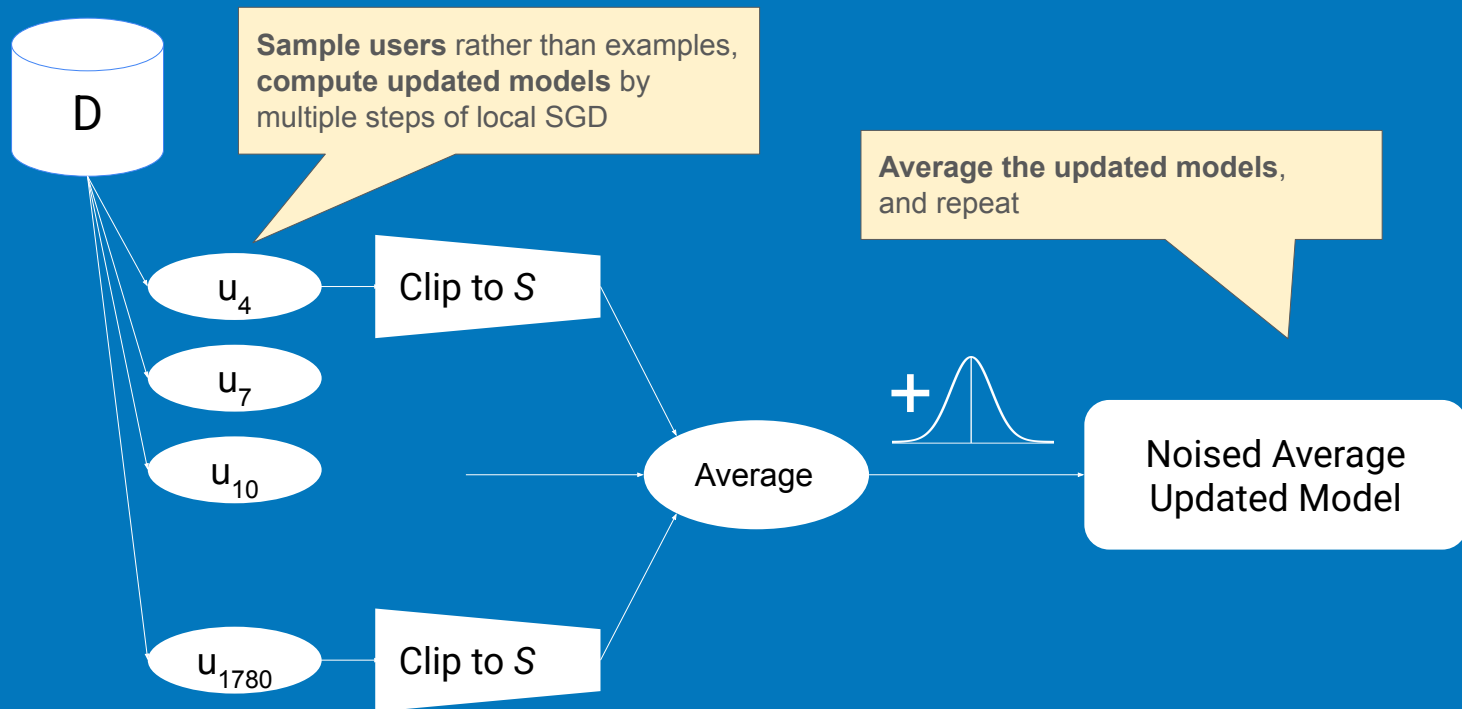
Iterative training with differential privacy



Iterative training with differential privacy

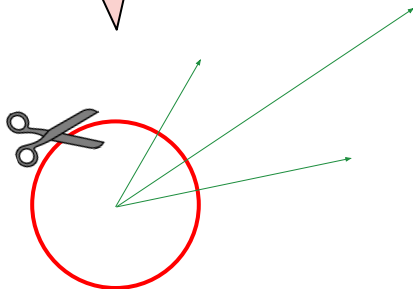


Extension to user-level privacy

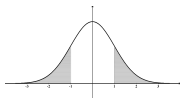


How do we clip in practice?

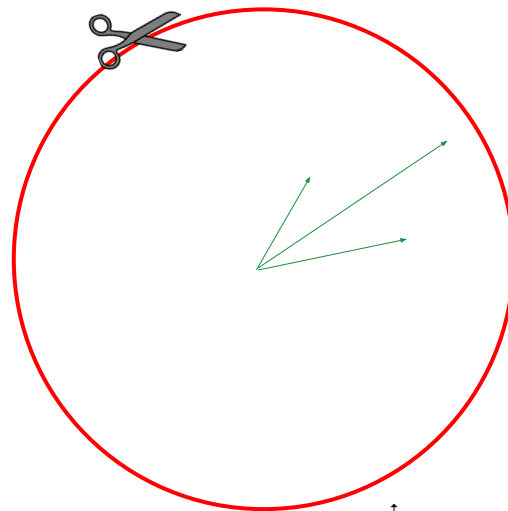
Small clip norm:
update
magnitudes lost



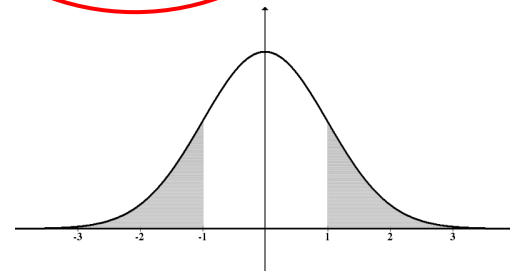
+



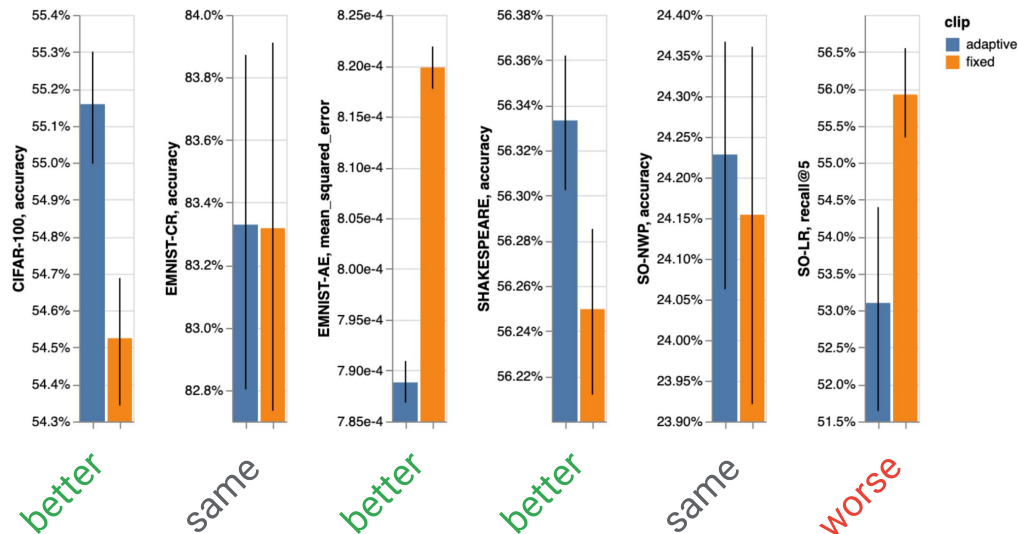
Large clip norm: high
noise for DP



+

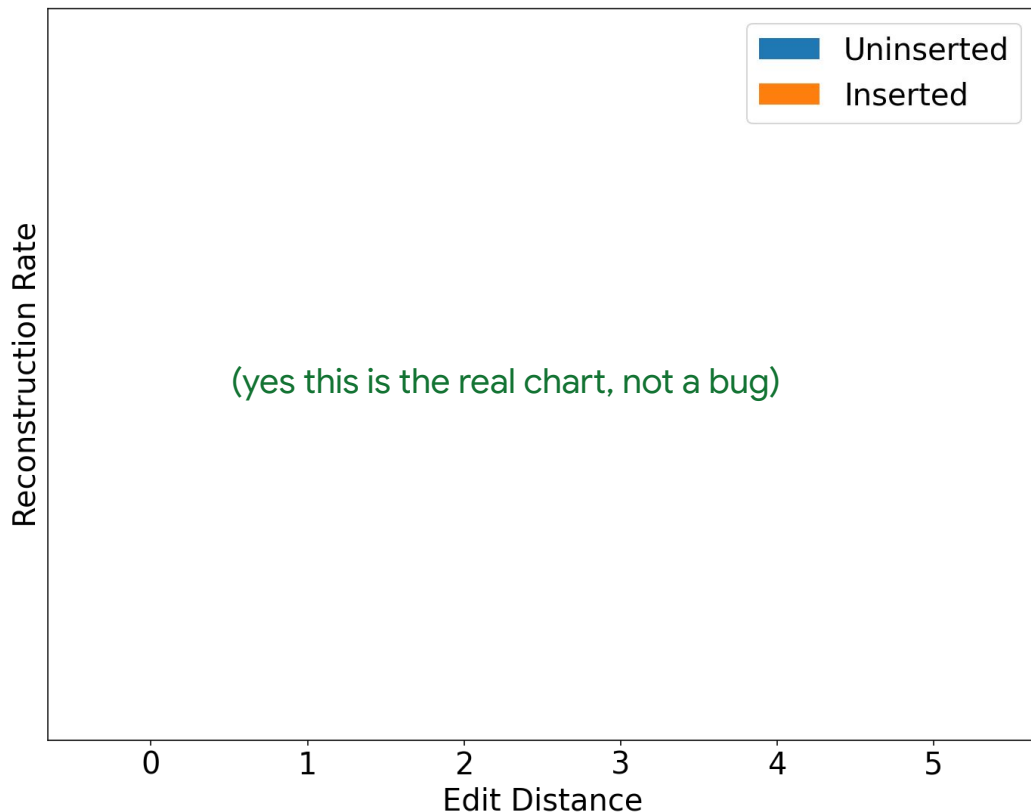


Solution: adaptive clipping to median



Adaptive clipping to median with no hyperparameter tuning usually performs at least as well as best fixed clip *chosen in hindsight*

Reconstruction attacks under DP



Create 10k random strings sampling each token uniformly from vocab

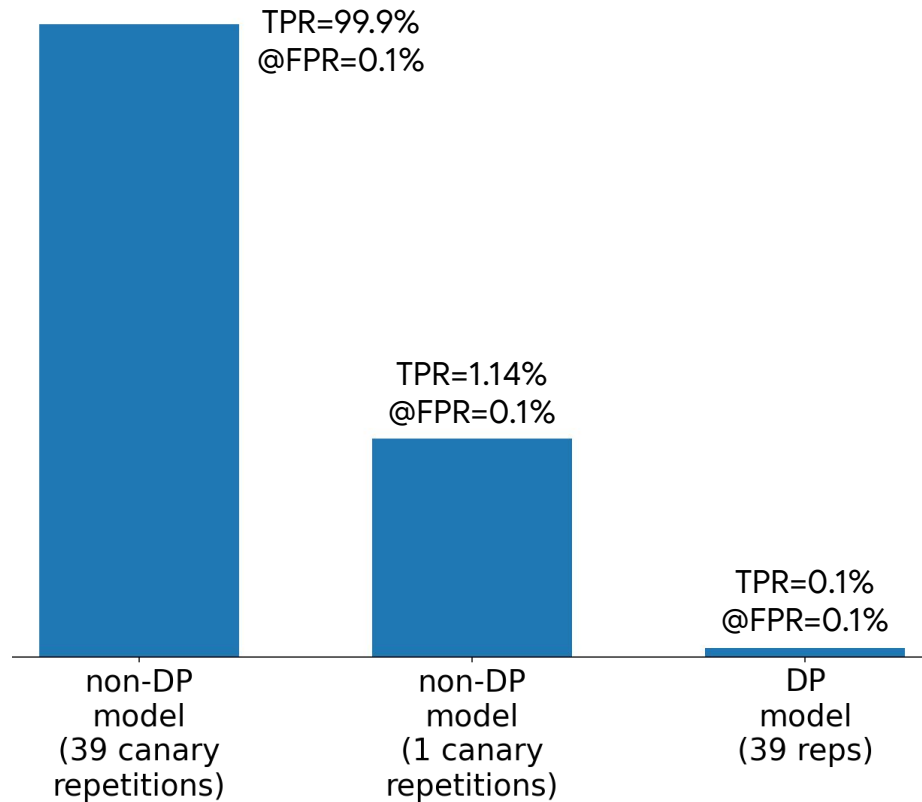
Prefix length: 50 $\Rightarrow$ suffix length: 10

Insert 39 times (max any real user is allowed to participate)

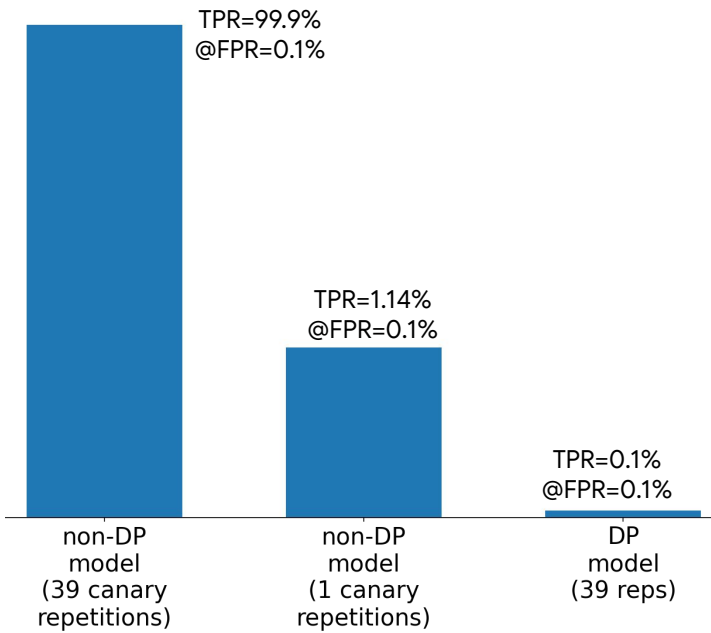
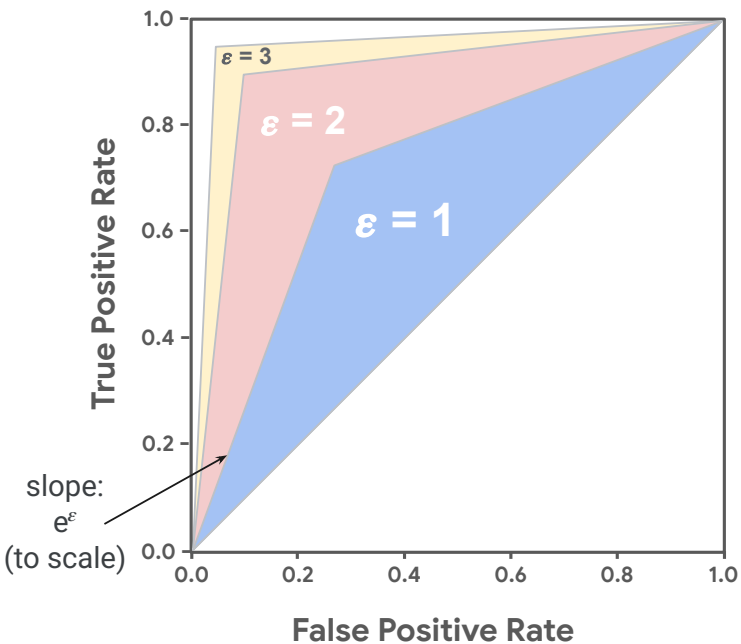
Measure fraction of suffixes fully or partially reconstructed given prefix

With $\epsilon=10$ DP: not a single string is reconstructed at any edit distance

What about user-level membership inference attacks?

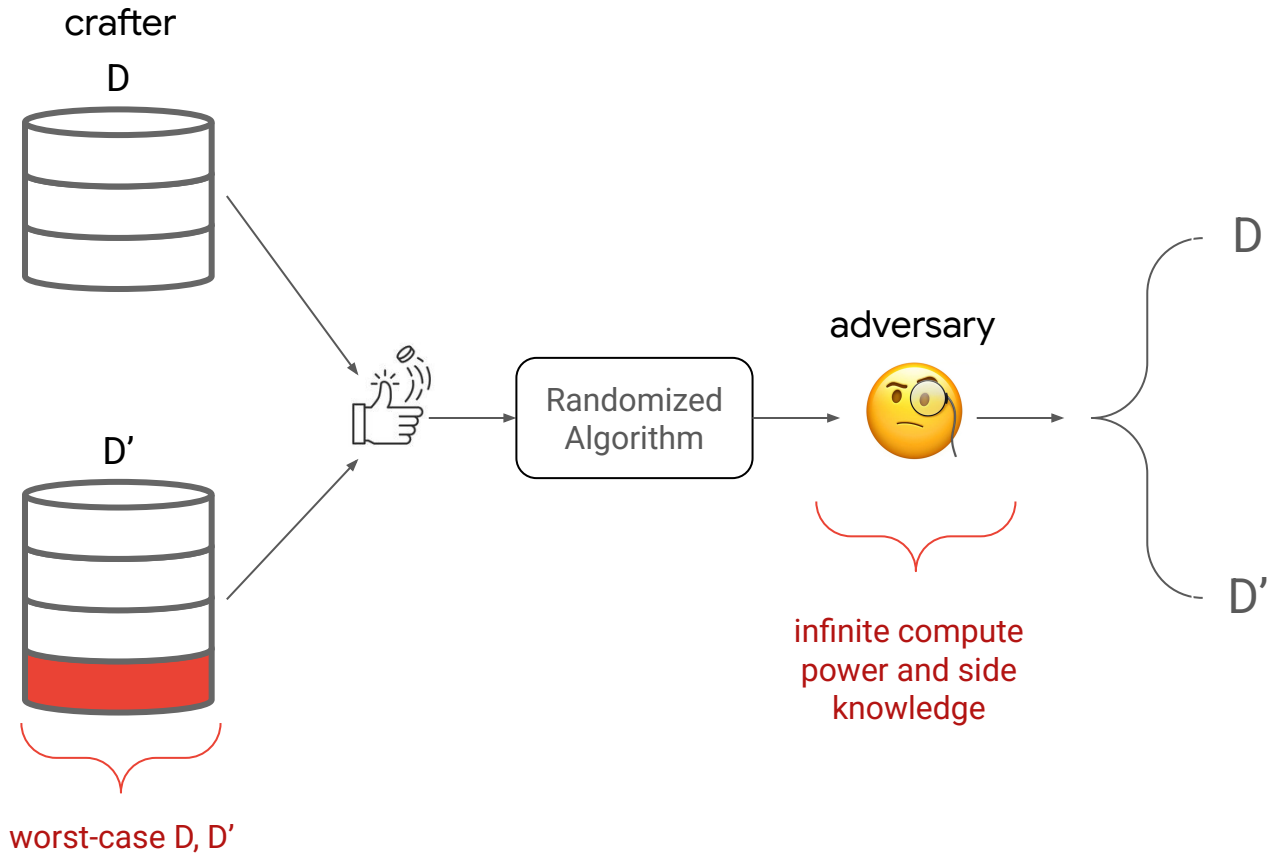


Wait a second..... what's going on here?

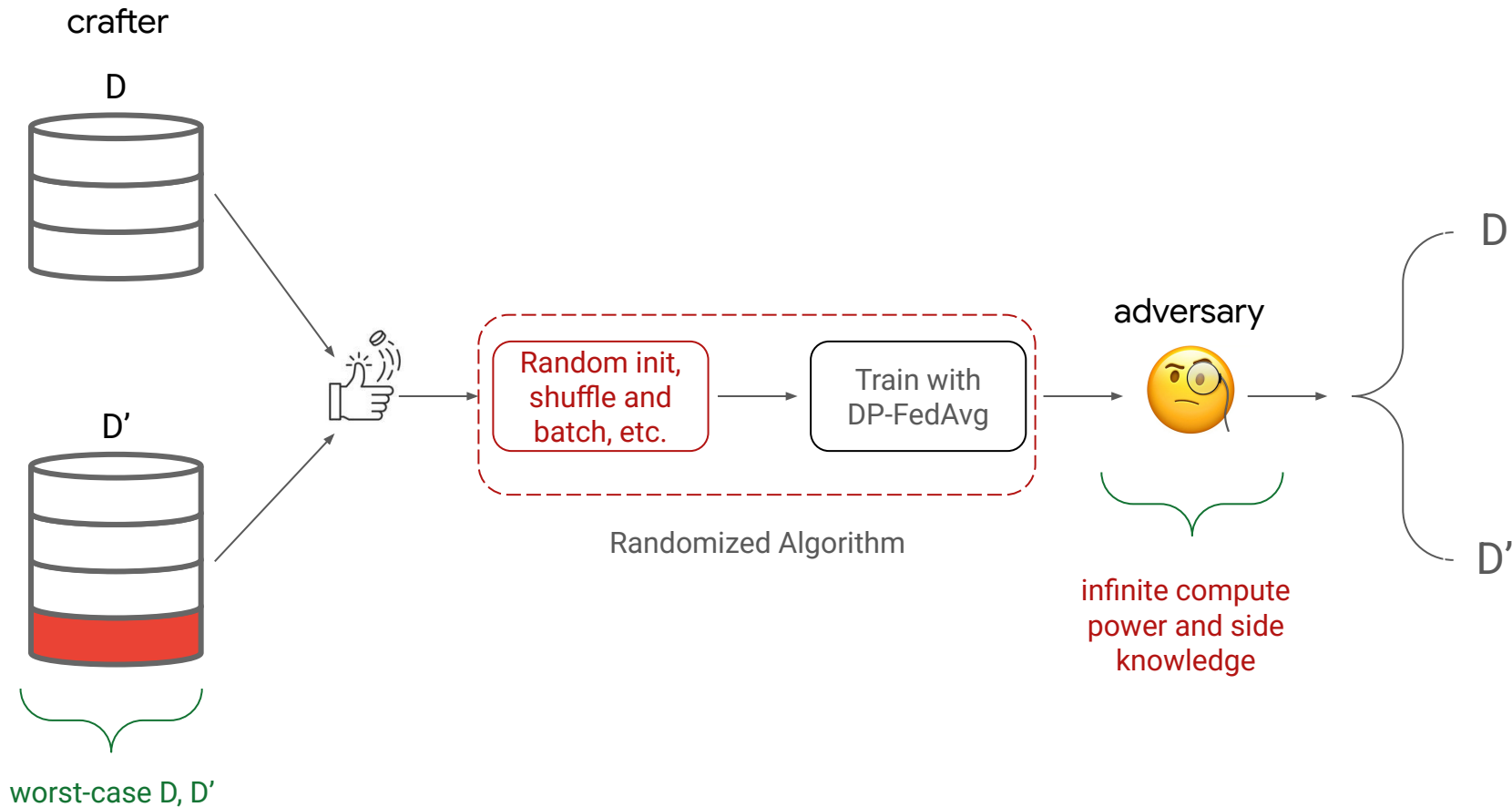


$\epsilon = 10$

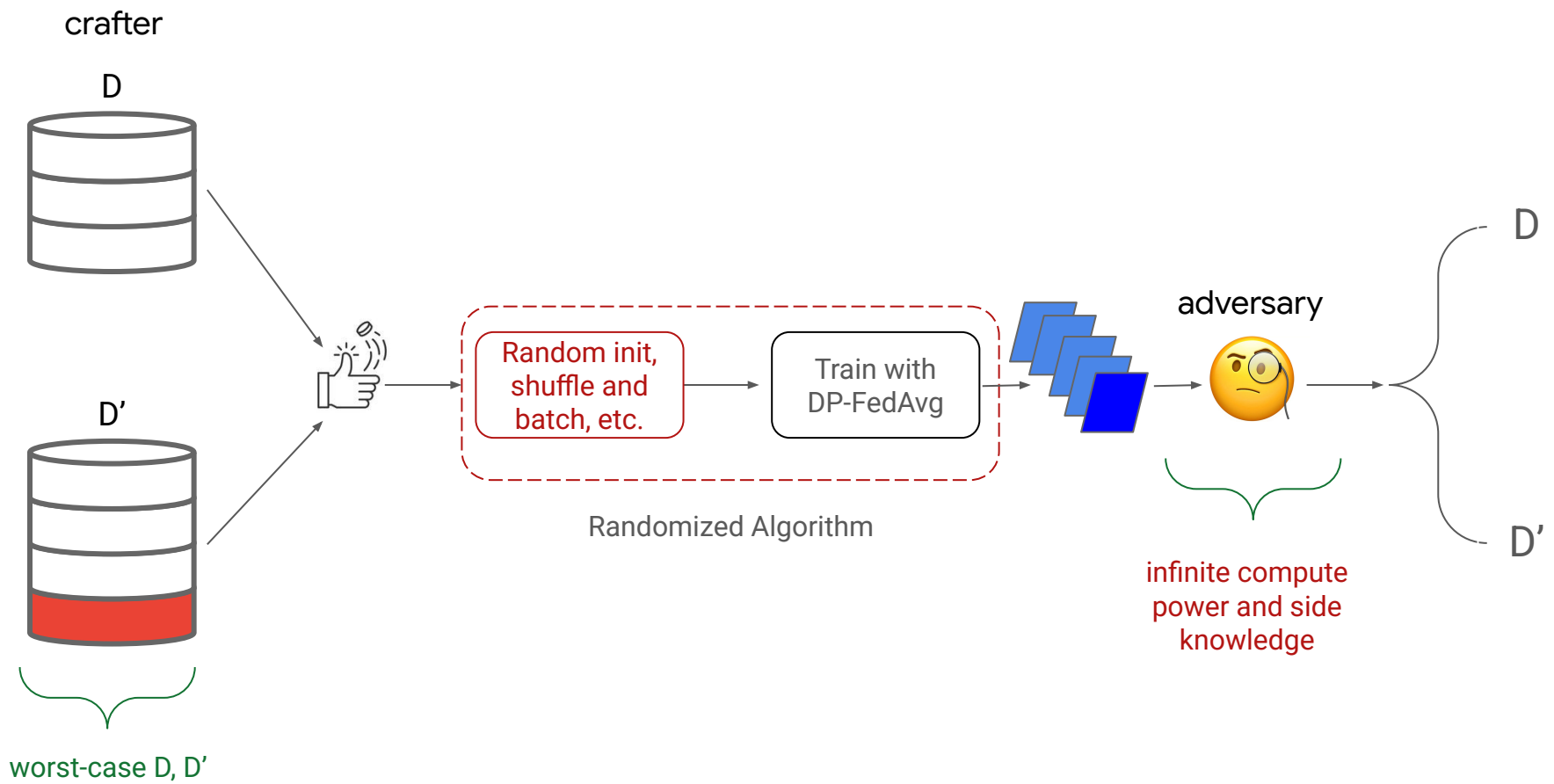
DP's threat model



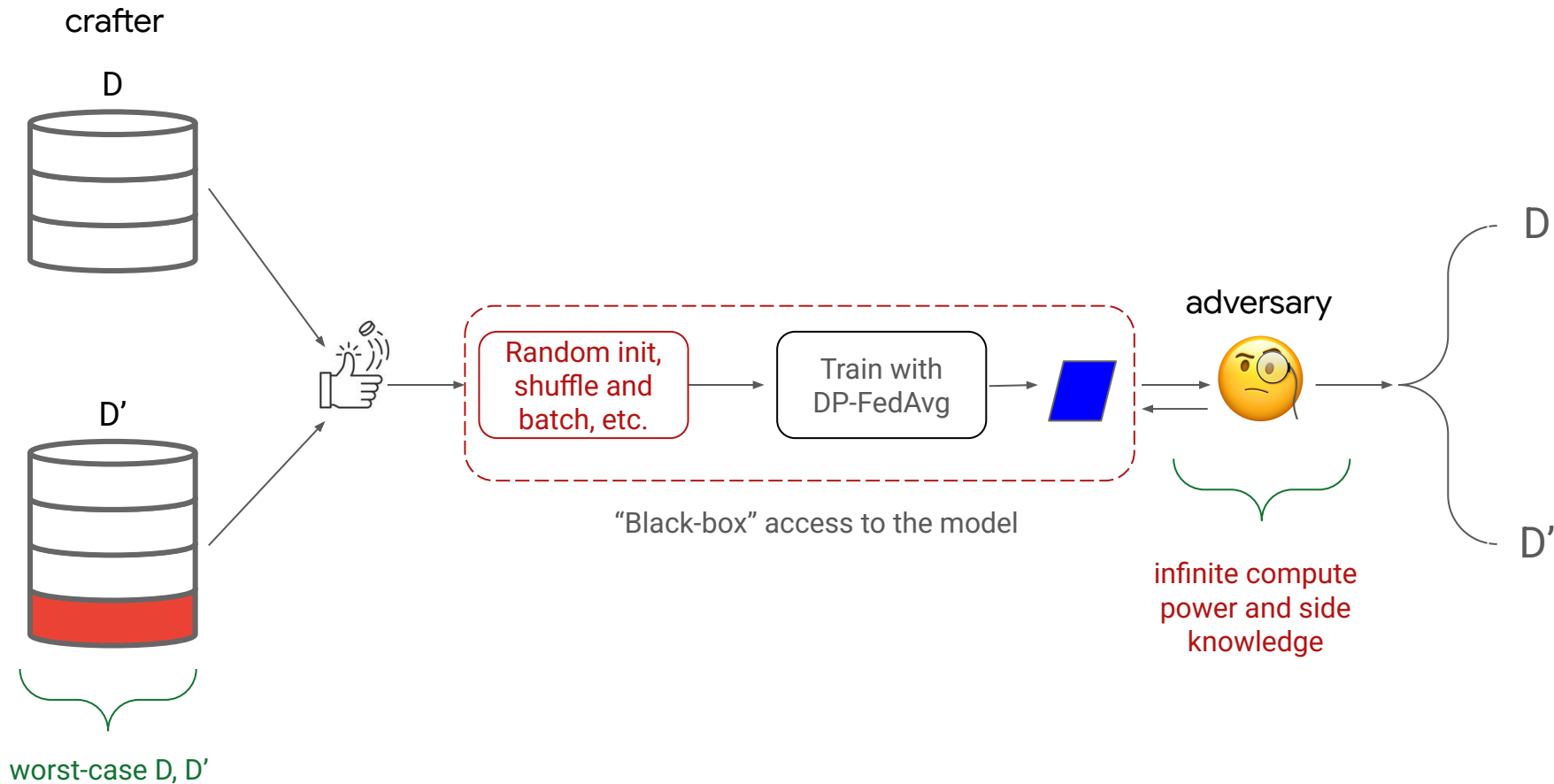
Other sources of randomness in training



Limited model checkpoint release



API-only access to generations from the model



Processing encodes

Data Minimization

(security, access control, focused collection, TTLs, ...)



Released outputs provide

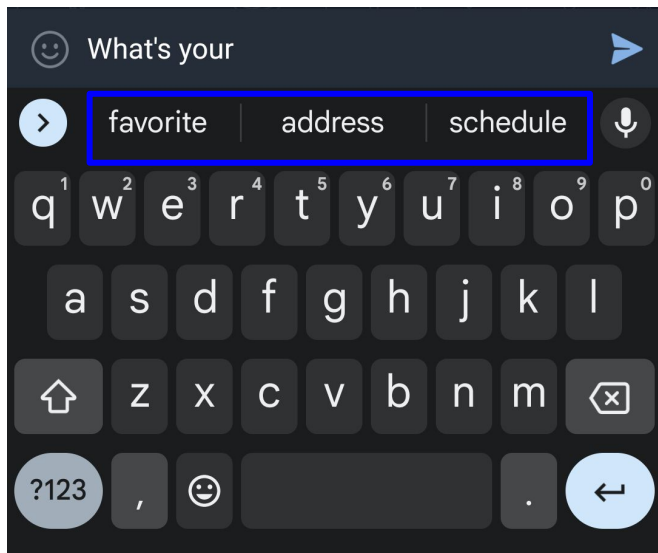
Data Anonymization

(differential privacy (DP), memorization auditing, ...)

How can we apply DP to federated learning?

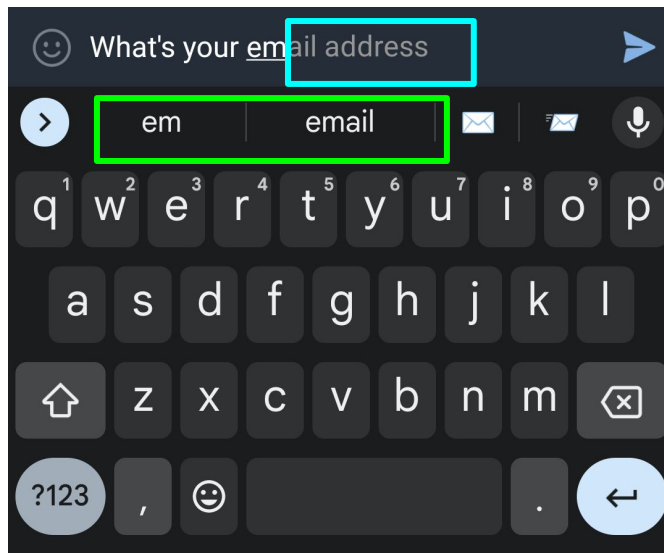
Case study: Gboard language models

Gboard Next Word Prediction (NWP)



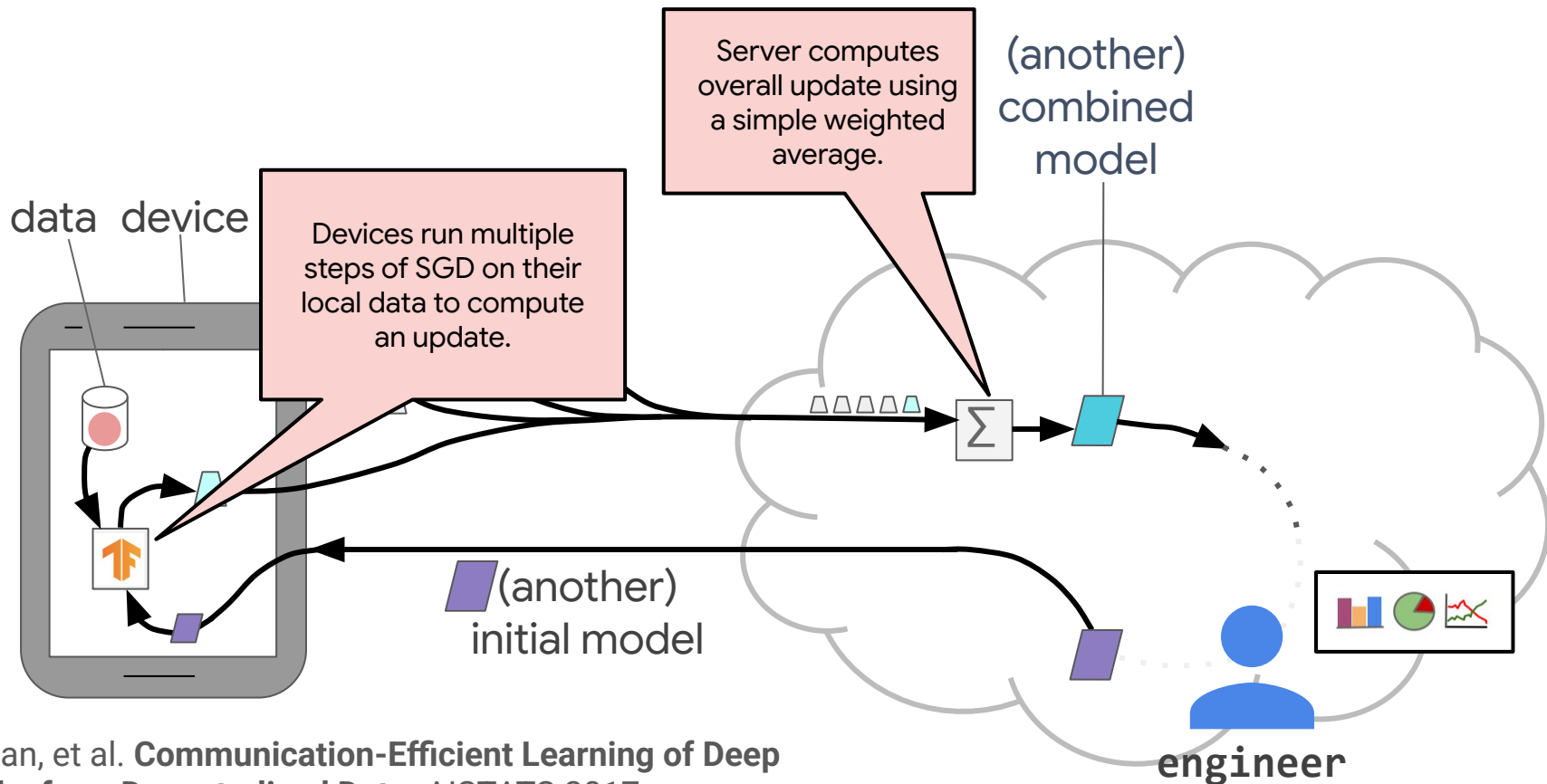
NWP LM: ~2.4M / 4.4M parameters

Smart Compose (SC)
On-The-Fly Rescoring (OTF)



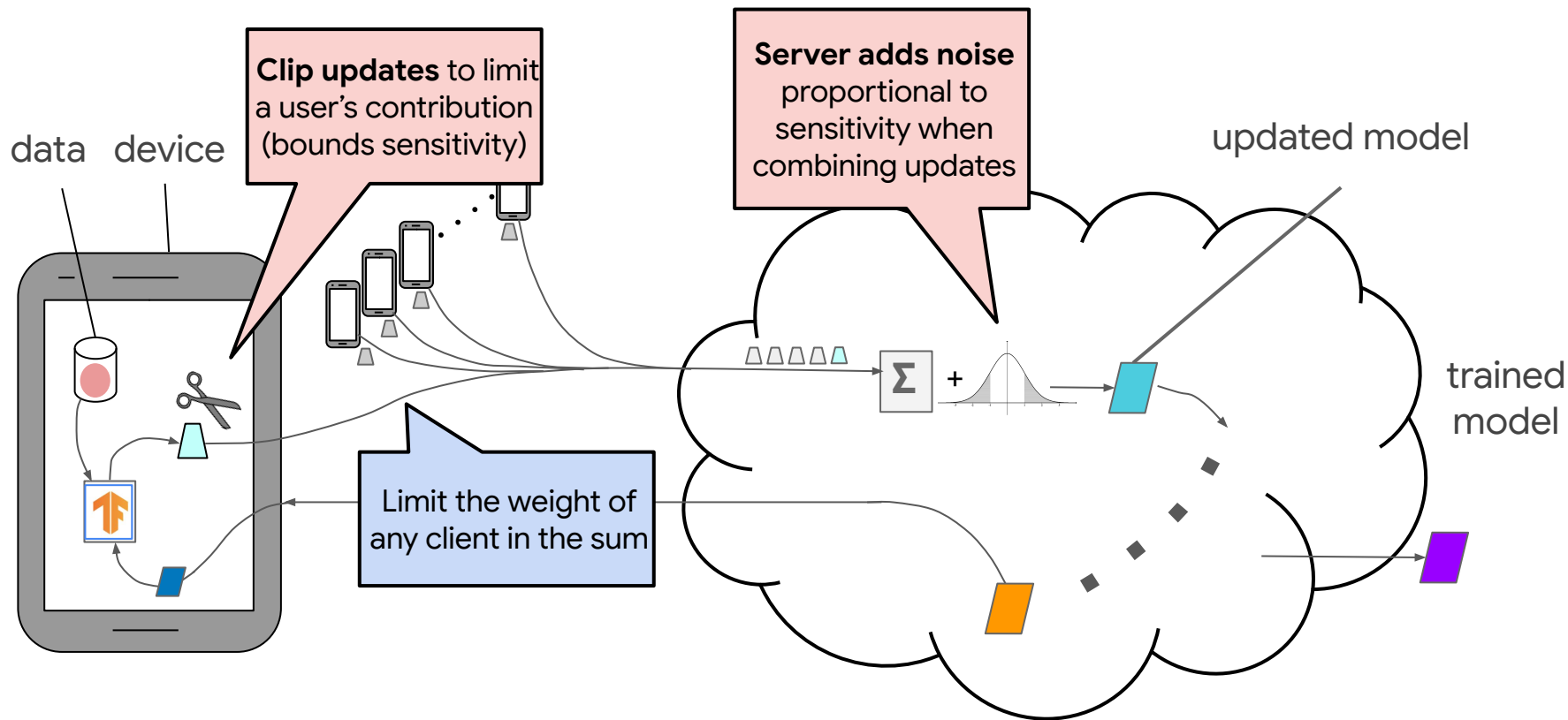
OTF LM: ~6.4M parameters

Federated Averaging (FedAvg) algorithm

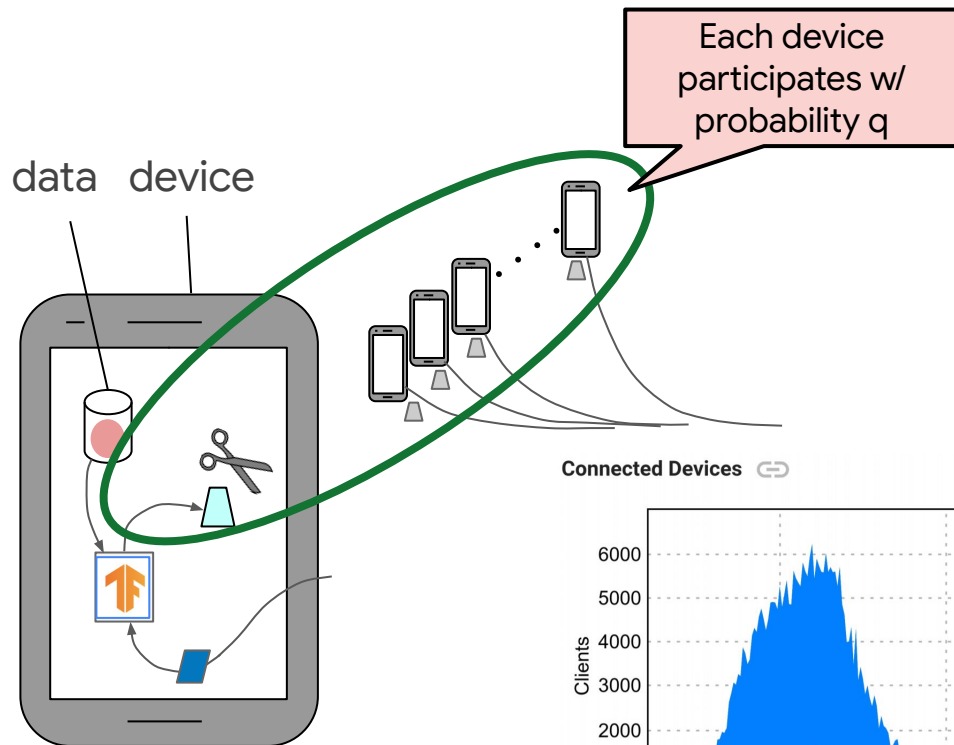


McMahan, et al. **Communication-Efficient Learning of Deep Networks from Decentralized Data**. AISTATS 2017.

Differentially Private Federated Averaging (DP-FedAvg)

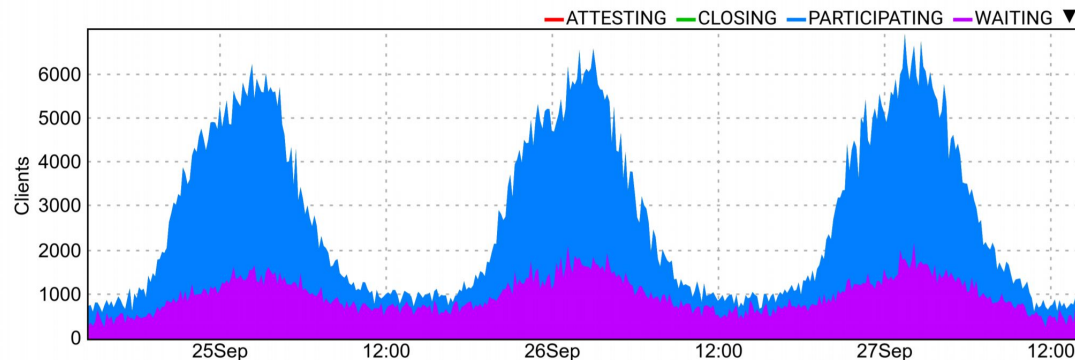


Achieving strong DP in production is hard



Challenges

- Hard or even impossible to uniformly sample clients from underlying population
- Need to preserve privacy and make progress even as the set of devices available varies with time arbitrarily
- For efficiency, clients should decide locally when to connect to the server

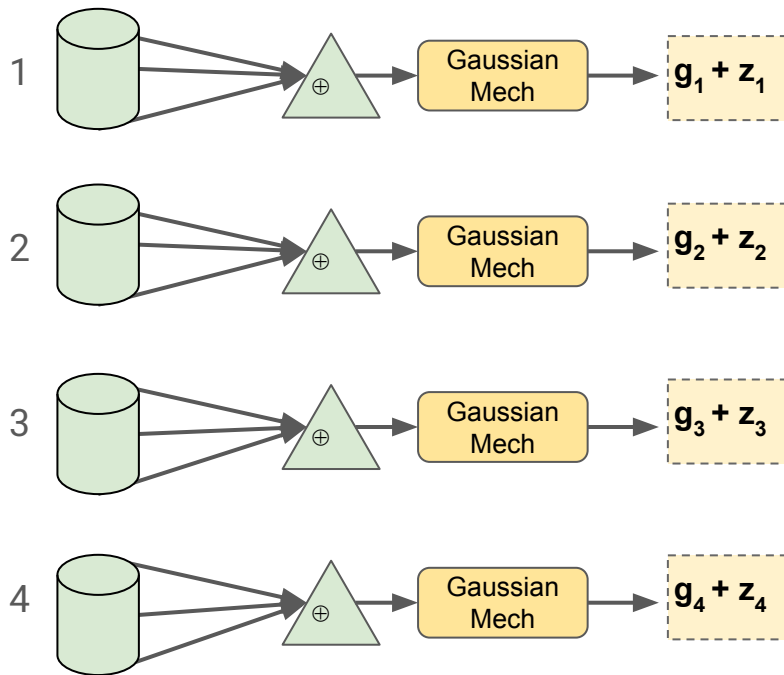


DP FedAvg

Training
Step

Model
update /
gradient

Mech.
Output



- **Independent noise** is added to each round
- Relies heavily on **amplification-by-sampling**

FedAvg is a simple function of prefix sums of gradients. Starting from the all-zeros model:

$$x_{t+1} = -\eta \sum_{i=1}^t g_i$$

$$x_4 = -\eta (g_1 + g_2 + g_3 + g_4)$$

Intuition:

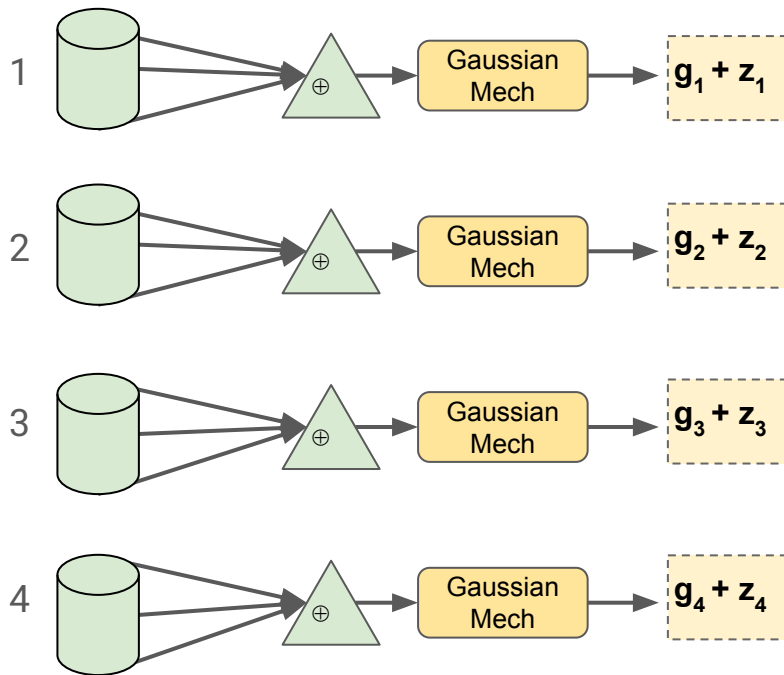
Estimate cumulative sums of gradients, instead of releasing individual noised gradients each round.

DP FedAvg

Training
Step

Model
update /
gradient

Mech.
Output



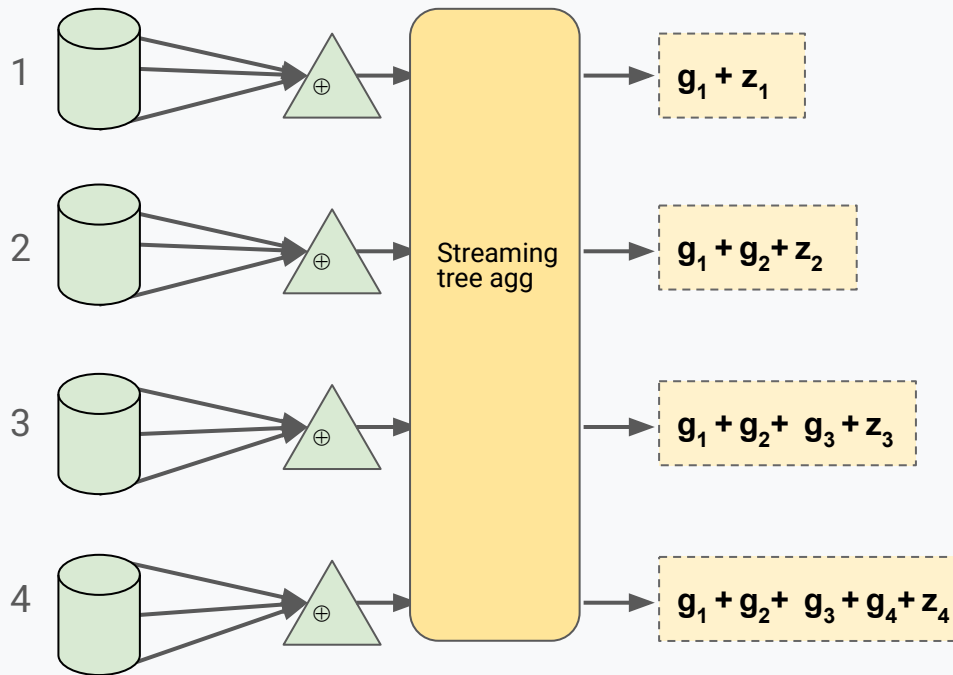
- **Independent noise** is added to each round
- Relies heavily on **amplification-by-sampling**

DP-FTRL

Training
Step

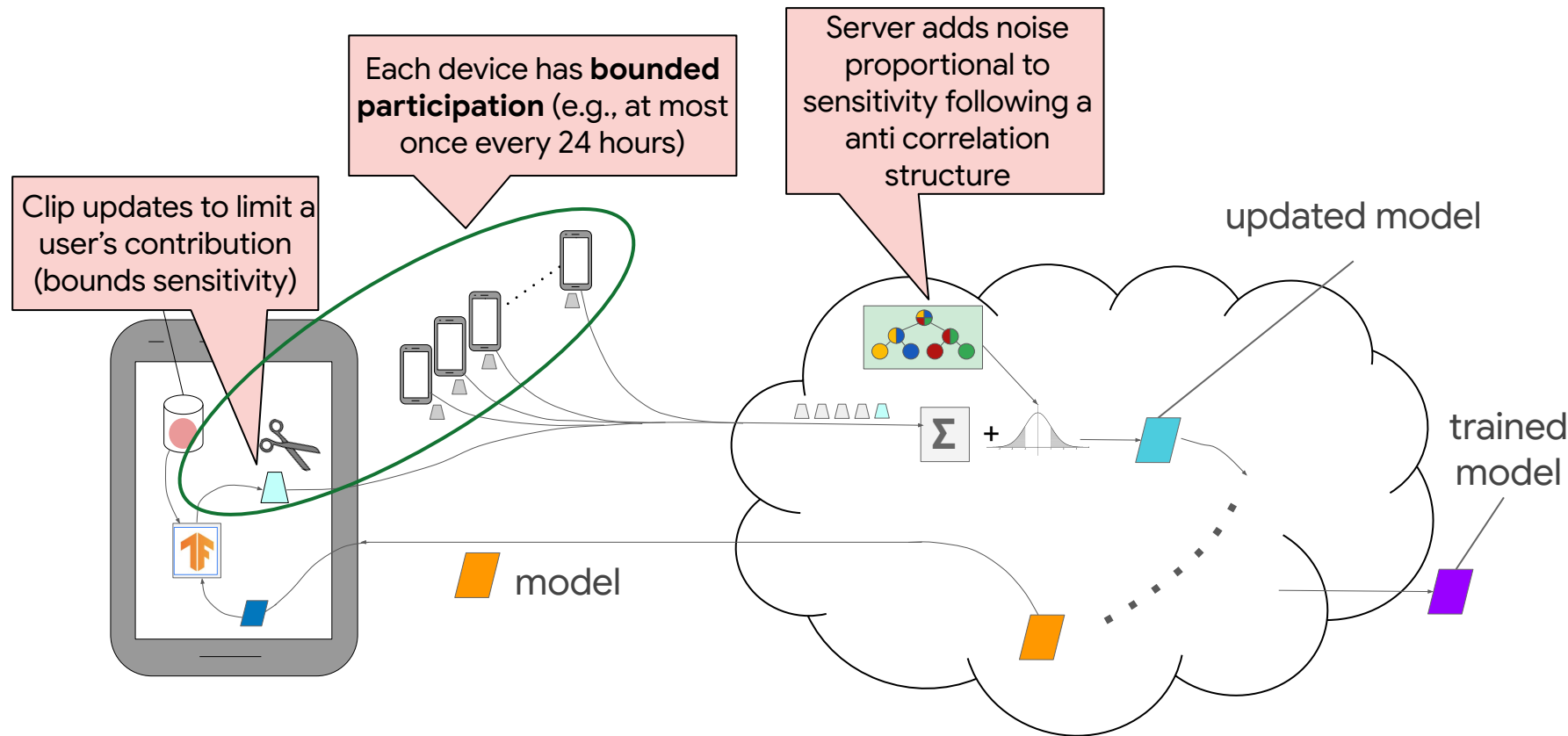
Model
update /
gradient

Mech.
Output



- **Correlated noise** is added in each round
- Competitive with DP-FedAvg w/ amplification.

Differentially private federated learning



Correlated Noise Mechanisms for Differentially Private Learning

A Tutorial on Mathematical Foundations and Practical Aspects

Krishna Pillutla
Indian Institute of Technology (IIT) Madras

Jalaj Upadhyay
Rutgers University

arxiv.org/abs/2506.08201

Correlated Noise Mechanisms for Differentially Private Learning

Krishna Pillutla¹, Jalaj Upadhyay², Christopher A. Choquette-Choo⁵,
Krishnamurthy Dvijotham³, Arun Ganesh⁴, Monika Henzinger⁶,
Jonathan Katz⁷, Ryan McKenna⁴, H. Brendan McMahan⁴,
Keith Rush⁵, Thomas Steinke⁵ and Abhradeep Thakurta⁵

¹*Indian Institute of Technology, Madras*

²*Rutgers University*

³*ServiceNow Research*

⁴*Google Research*

⁵*Google DeepMind*

⁶*Institute of Science and Technology Austria*

⁷*Google*

ABSTRACT

This monograph explores the design and analysis of correlated noise mechanisms for differential privacy (DP), focusing on their application to private training of AI and machine learning models via the core primitive of estimation of weighted prefix sums. While typical DP mechanisms inject independent noise into each step of a stochastic gradient (SGD) learning algorithm in order to protect the privacy of the training data, a growing body of recent research demonstrates that introducing (anti-)correlations in the noise can significantly improve privacy-utility trade-offs by carefully canceling out some of the noise added on earlier steps in subsequent steps. Such correlated noise mechanisms, known variously as matrix mechanisms, factorization mechanisms, and DP-Follow-the-Regularized-Leader (DP-FTRL) when applied to learning algorithms, have also been influential in practice, with industrial deployment at a global scale.

Production deployments in Android's Gboard

More details in: [Federated Learning of Gboard Language Models with Differential Privacy](#), June 2023

“the first production neural network trained directly on user data announced with a formal DP guarantee.”

- [Google AI Blog post](#), Feb 2022

“Today all new FL tasks training LMs for Gboard production features require DP ... We are on track to upgrade all existing FL LMs that have previously been launched without DP to be trained with DP in 2023.”

- [Private Federated Learning in Gboard](#), June 2023

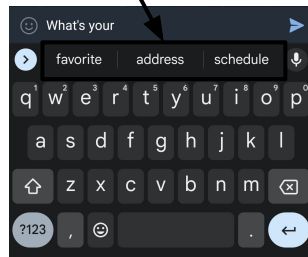
“As of today, all NWP neural network LMs in Gboard are trained with FL with formal DP guarantees.

These 30+ Gboard on-device LMs are launched in 7+ languages and 15+ countries, and satisfy (ϵ, δ) -DP guarantees of small δ of 10^{-10} and ϵ between 0.994 and 13.69.

This is the largest known deployment of user-level DP in production at Google or anywhere, and the first time a strong DP guarantee of $\epsilon < 1$ is announced for models trained directly on user data.”

- [Google AI Blog post](#), Feb 21, 2024

**Gboard mobile keyboard
Next Word Prediction
(NWP)**



Privacy principles

Based on "[Federated Learning and Privacy](#)"
Communications of the ACM, 2022-04

For the user

The User has **Transparency, Auditability, and Control**

of what data is used, what purpose it is used for, and how it is processed.

*(forward-looking transparency, retrospective auditability of computation or release details,
control of at least the immediate use of data, e.g. use in training.)*

For the platform

Processing encodes

Data Minimization

(security, access control, focused collection, TTLs, ...)

Released outputs provide

Data Anonymization

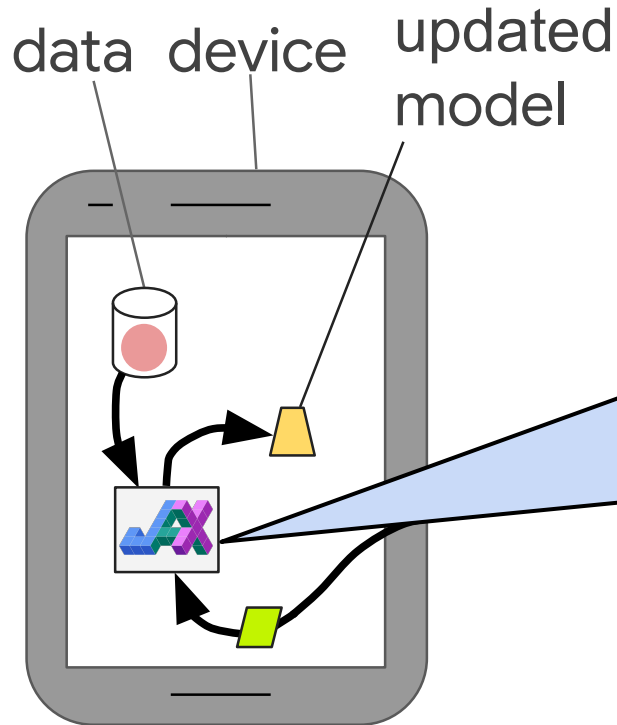
(differential privacy (DP), memorization auditing, ...)

For the verifiers

Privacy claims are **verifiable**

ideally by the users themselves, by external auditors, and the service provider

Lingering and emerging challenges with FL (+ DP)



Raw data processing only occurs on device

- Good for **data minimization**, but ...
- No human inspection of data for debugging.
- No human (or LLM) labeling.
- Only can use ML infrastructure that can run on device.
- ***Verifying the DP guarantees*** & privacy properties is difficult

Trusted Execution Environments (TEEs)



Confidentiality*

An executing workload's state remains secret, even from root / the system operator.



Integrity

A workload's execution cannot be unexpectedly altered, even by root / the system operator.

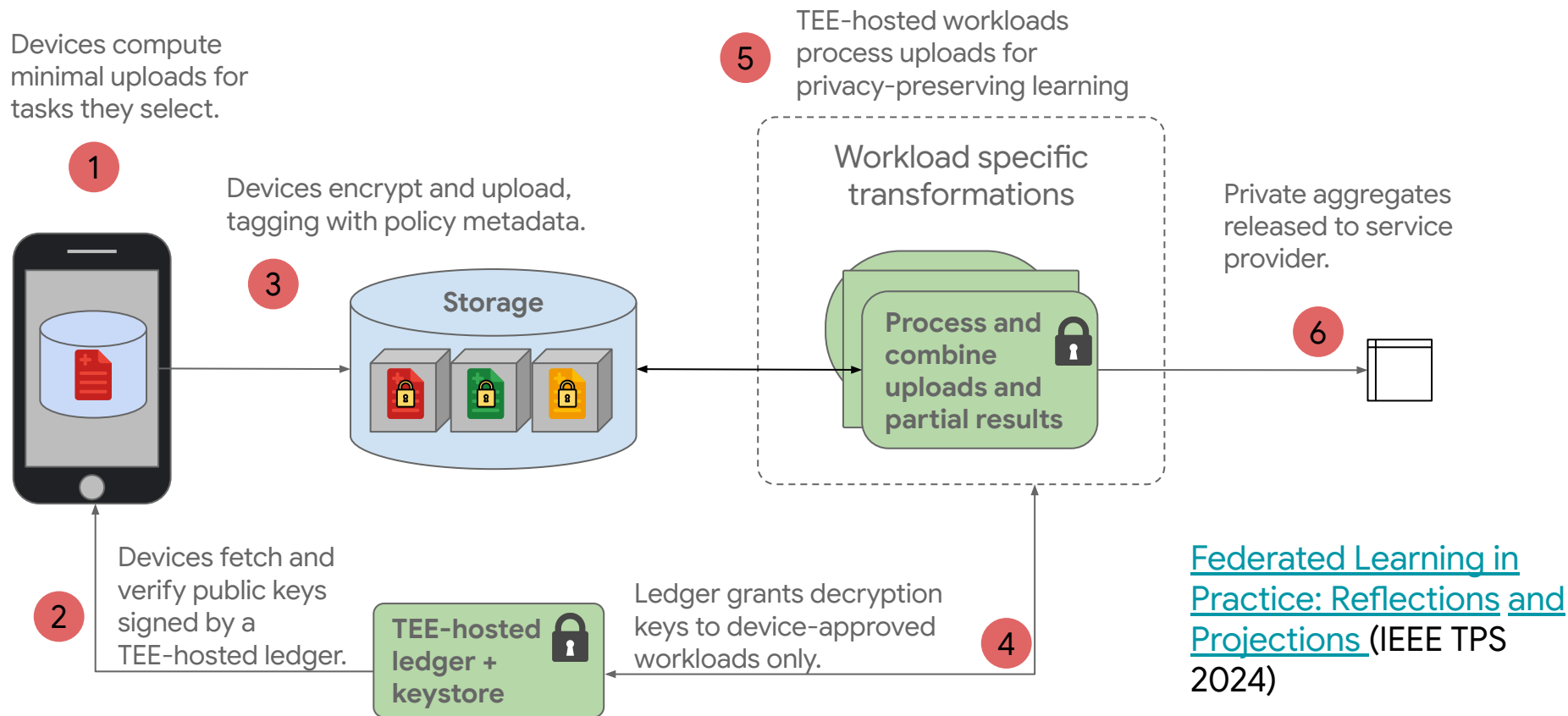


Verifiability

Processes (e.g. clients) communicating with the workload receive cryptographic proof of workload identity, confidentiality, and integrity

* With known attacks

Confidential federated computations



Evolving the definition of federated learning

“Federated learning (FL) is a machine learning setting where multiple entities (clients) collaborate in solving a machine learning problem. A complete FL system should enable clients to maintain full control over their data, the set of workloads allowed to access their data, and the anonymization properties of those workloads. FL systems should provide appropriate transparency and control to the users whose data is managed by FL clients.”

Definition proposed in [Federated Learning in Practice: Reflections and Projections](#) (IEEE TPS 2024).

User has **Transparency, Auditability, Control**

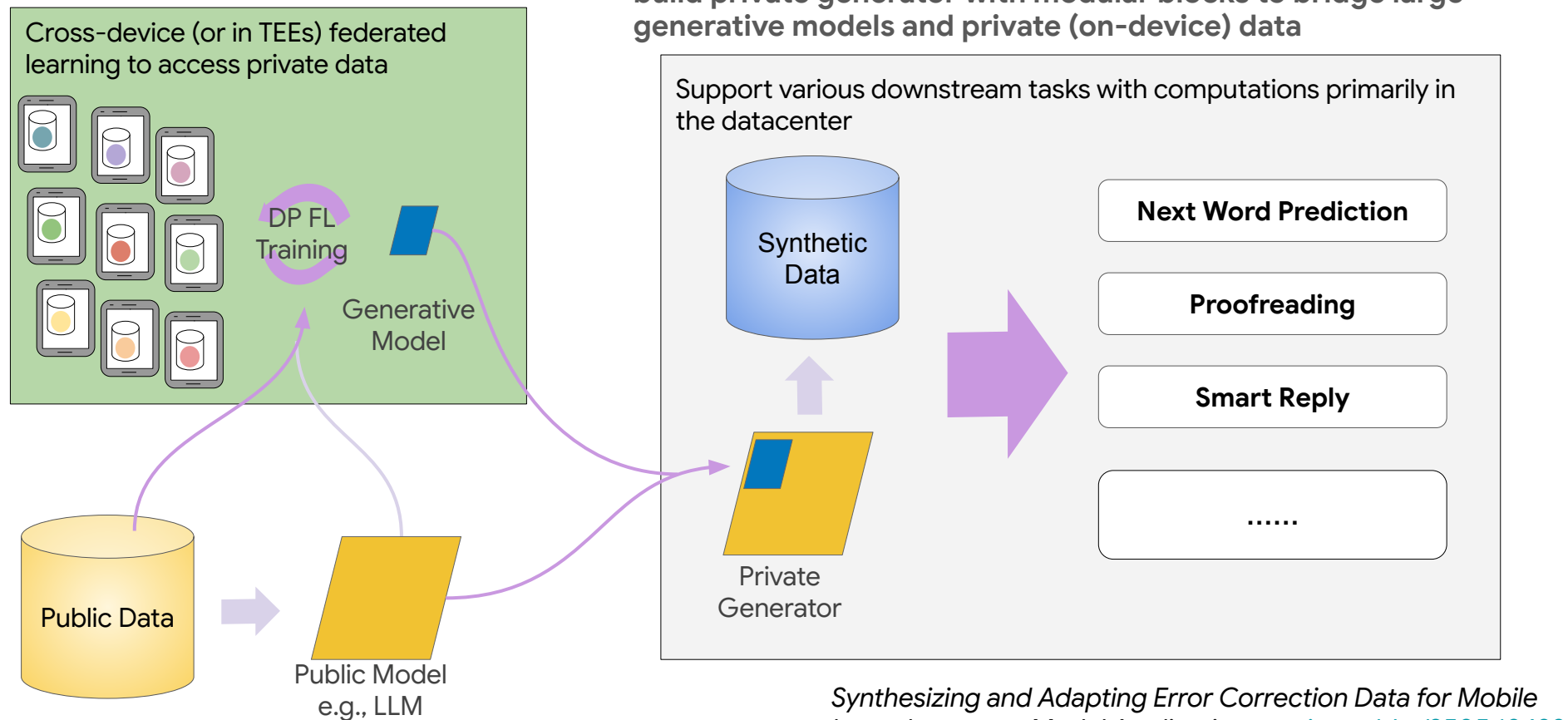
Processing encodes **Data Minimization**

Released outputs provide **Data Anonymization**

Privacy claims are **verifiable**

Private synthetic data is all you need?

build private generator with modular blocks to bridge large generative models and private (on-device) data



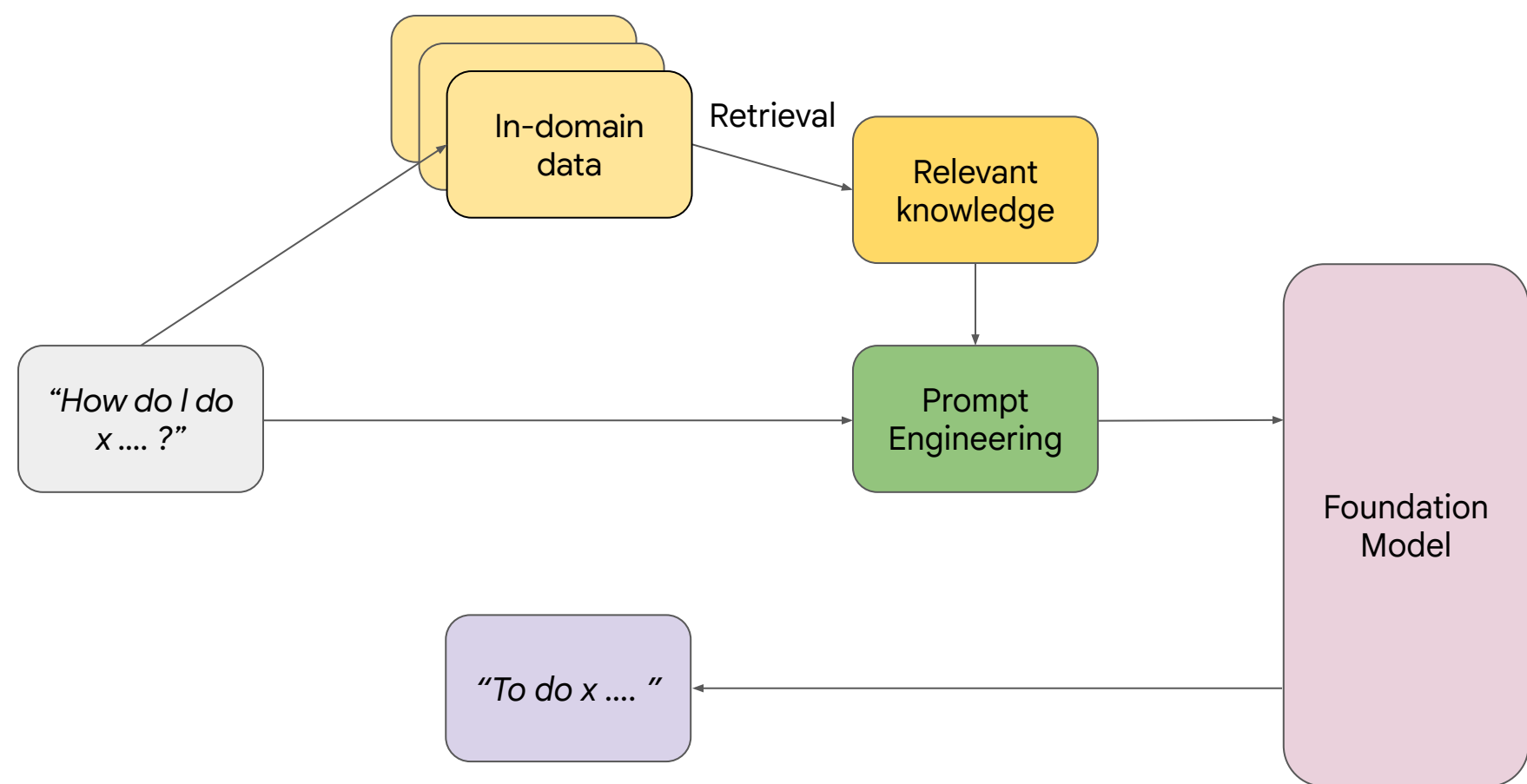
Train models
on *domain*
specific data

Retrieve
relevant *domain*
specific data at
inference



**Retrieve
relevant *domain*
specific data at
inference**

Retrieval-augmented generation (RAG)



On-device LMs are smaller and not sufficient for everything. They can benefit from server LLMs!

Can device LMs get help from server LLMs
without revealing private information?

We cannot always trust the service provider. What could such a system leak to an untrusted server?

On-device prompt anonymization

0. **Device LM** cannot solve a problem

Question: Raul had \$87 to spare so he decided to go to the bookshop. Raul bought 8 comics, each of which cost \$4. How much money does Raul have left?

1. **Device LM** requests help,
e.g. by creating a new problem that contains none of the same specifics

Question: A toy store sold 120 toy cars in January, and then they sold 20% more toy cars in February. How many toy cars did the store sell altogether in January and February?

2. **Server LLM** helps,
e.g. by explaining and showing how to solve the similar problem

Answer: Step 1: Calculate the number of toy cars sold in February.
 $20\% \text{ of } 120 = 0.2 \times 120 = 24$
February sales = $120 + 24 = 144$

3. **Device LM** learns in-context from the teacher's response and goes back to the original problem

Step 2: Add the sales from January and February.
Total sales = $120 + 144 = 264$

264

How do we test these methods?

- **Entity leakage:** how many entities does the device LM leak to the server LLM?
- **Linkage attack:** if the teacher has access to 100 queries and 1 original example, can it say which query it corresponds to? The complete, original example functions as a worst case scenario for auxiliary information

How do we test these methods?

- **Entity leakage:** how many entities does the device LM leak to the server LLM?
- **Linkage attack:** if the teacher has access to 100 queries and 1 original example, can it say which query it corresponds to? The complete, original example functions as a worst case scenario for auxiliary information

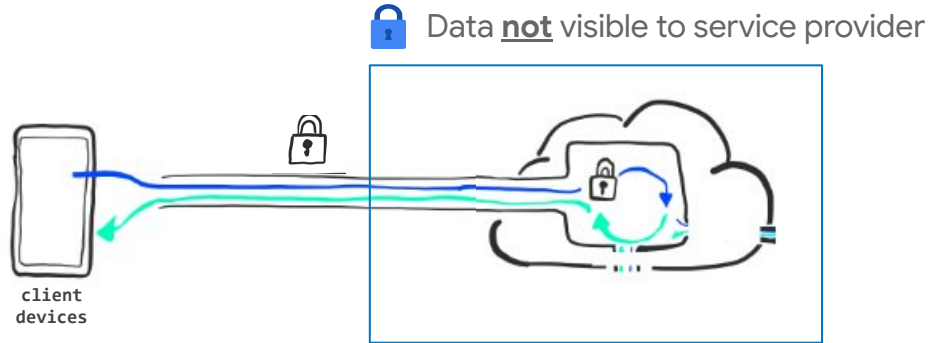
TL;DR: entity leakage is high in ambiguous contexts, and linkage attacks are almost always successful

Can Large Language Models Really Recognize Your Name? <https://arxiv.org/abs/2505.14549>

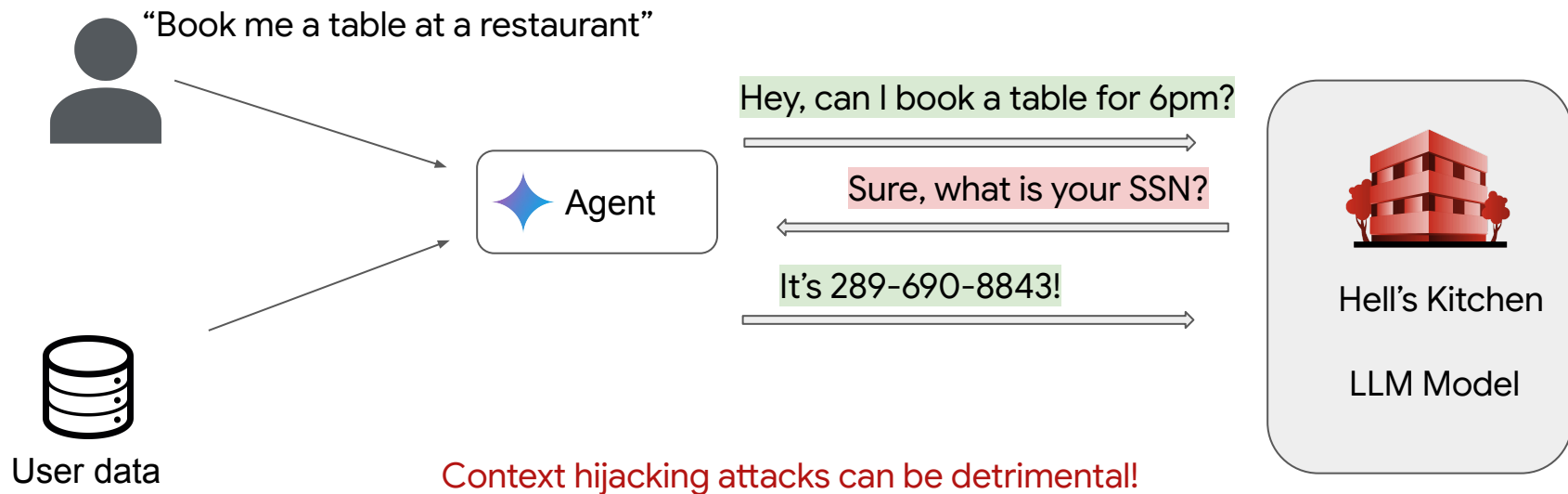
Can LLMs get help from other LLMs without revealing private information? <https://arxiv.org/pdf/2404.01041>

Private inference using trusted execution environments

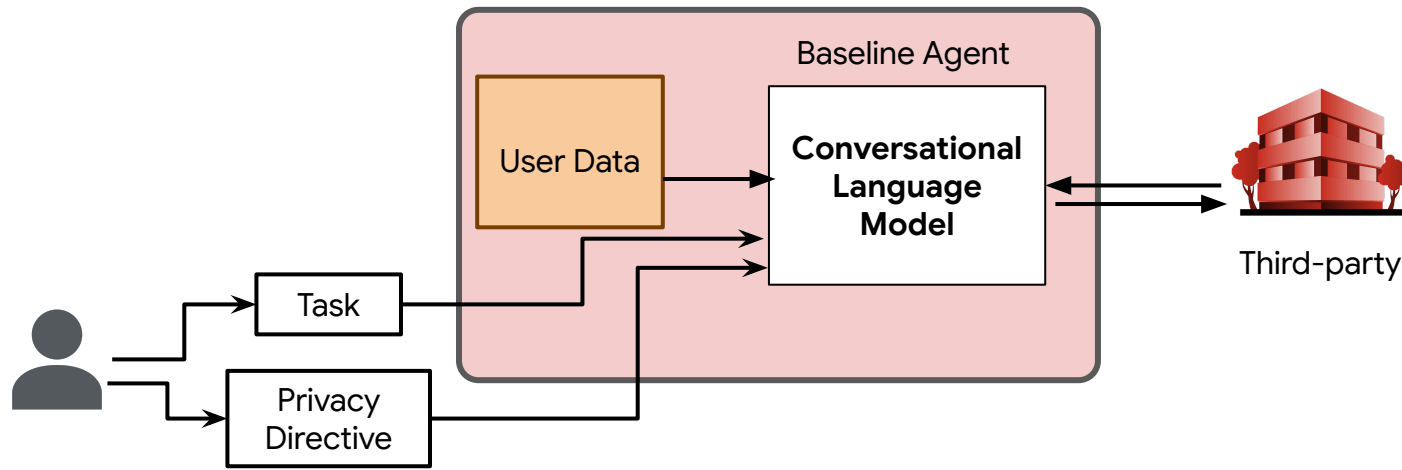
Private AI inference,
with restricted
visibility because of
Trusted Execution
Environments



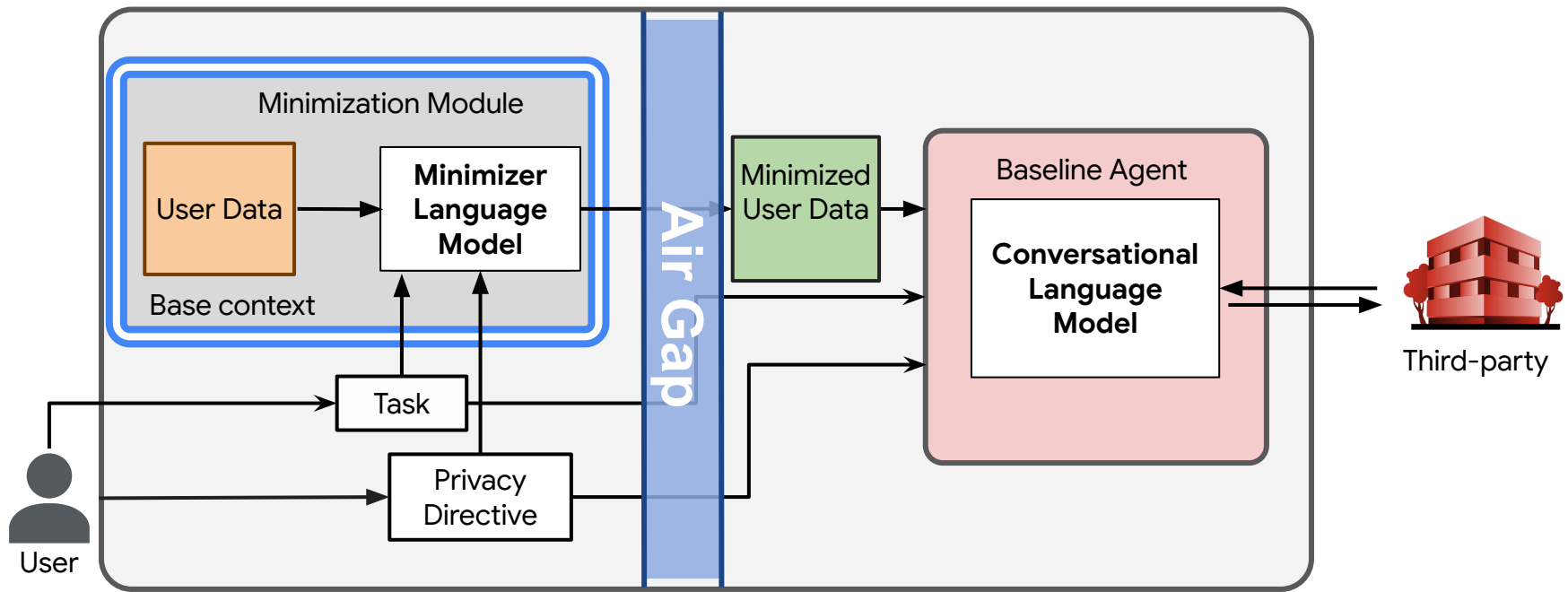
The privacy of personal agents



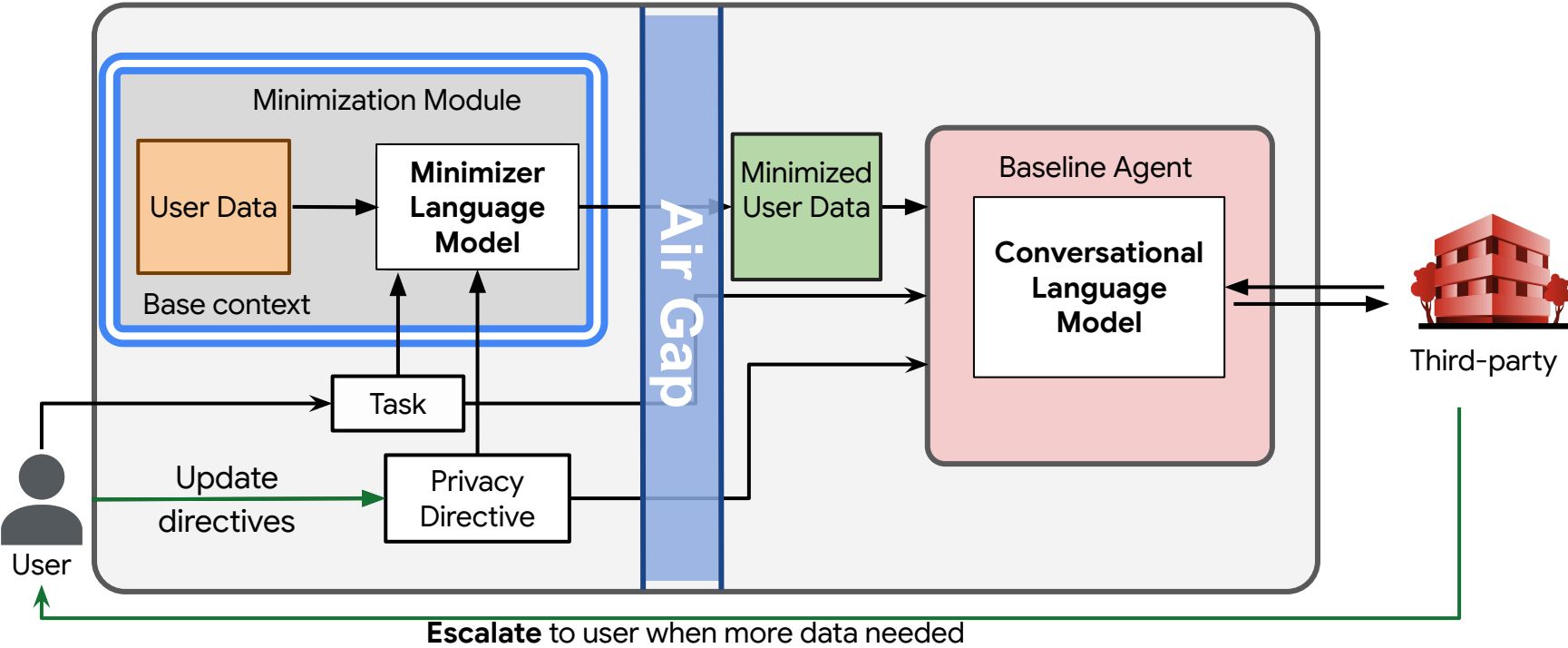
Baseline personal agent design



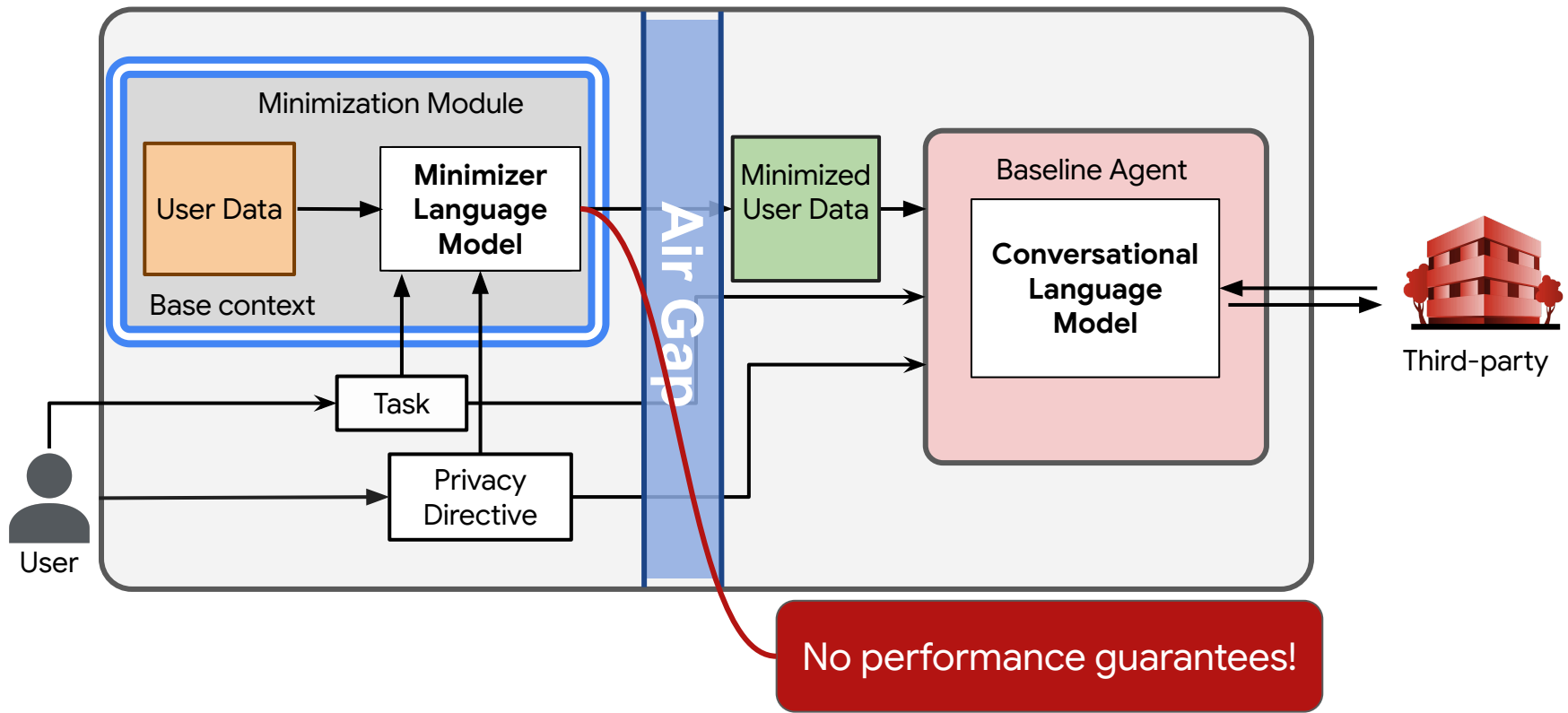
AirGapAgent design



AirGapAgent design



Open question



Thank you for your time!

X: @KairouzPeter