Towards User-level Differential Privacy at Scale

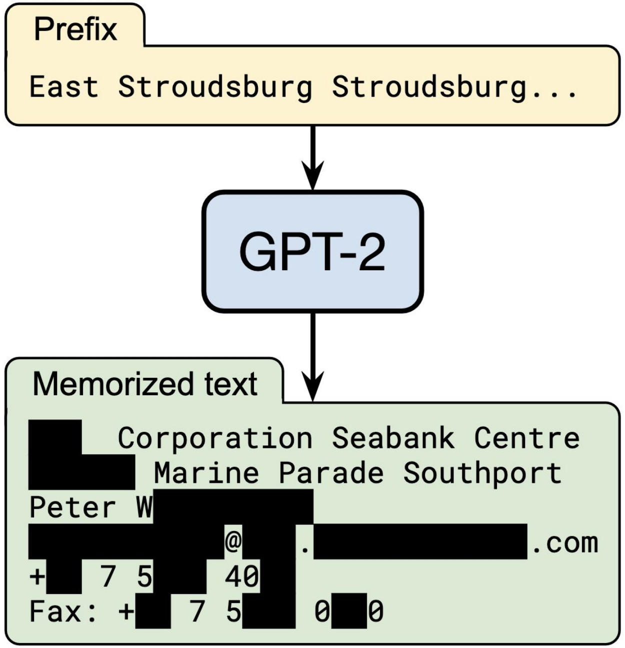
Krishna Pillutla
Google Research -> IIT Madras

Google Research

WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.

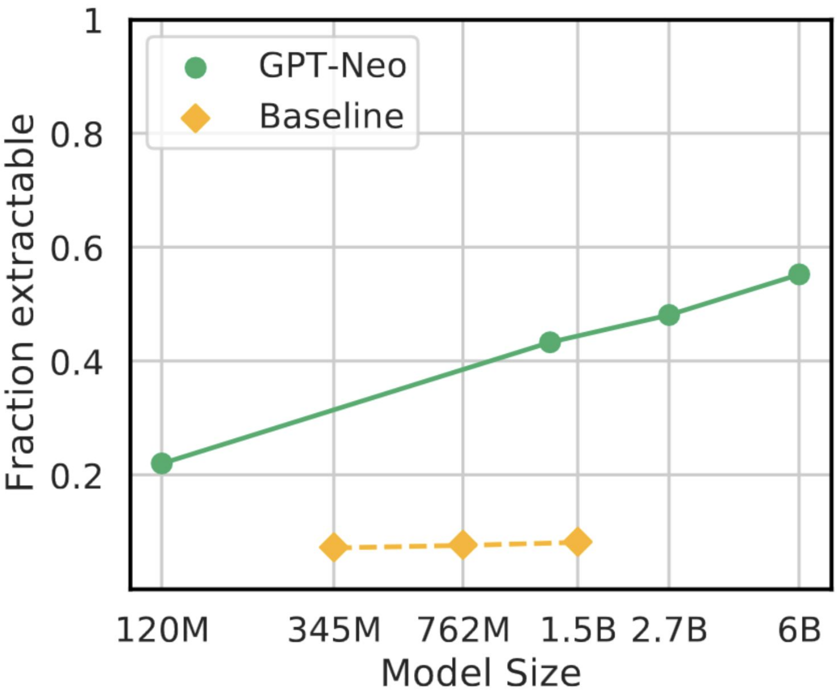
Models leak information about their training data

WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.

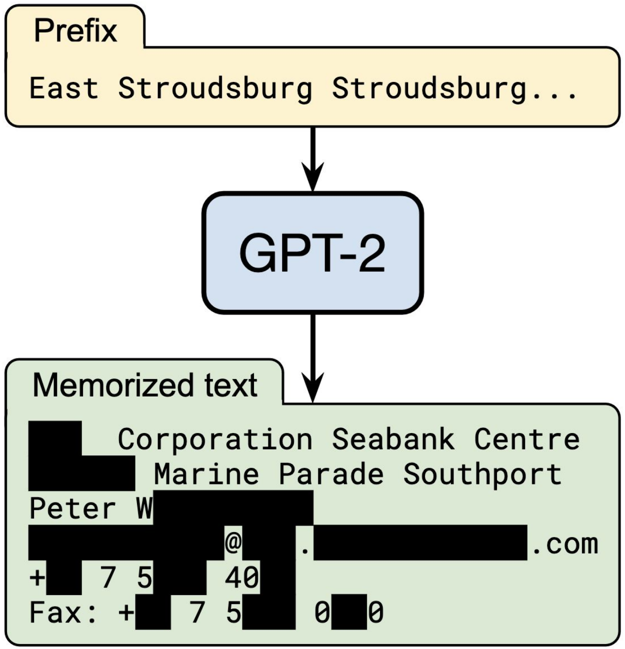


Carlini et al. (USENIX Security 2021)

Models leak information about their training data ***reliably***



Carlini et al. (ICLR 2023)



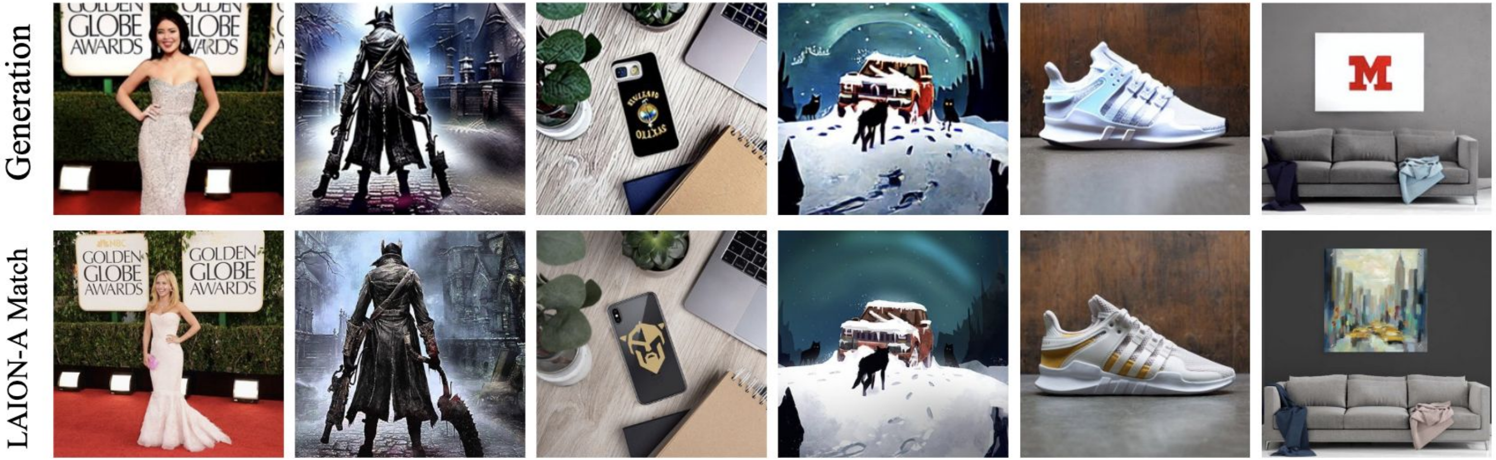
Carlini et al. (USENIX Security 2021)

Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models

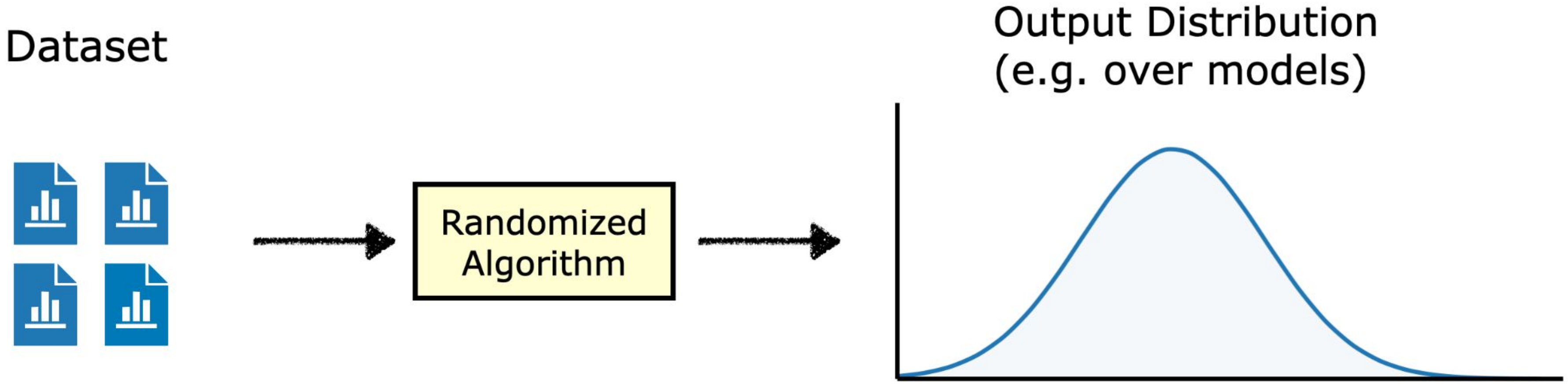
Gowthami Somepalli 🐢, Vasu Singla 🐢, Micah Goldblum 🗽, Jonas Geiping 🐢, Tom Goldstein 🐢

🐢 University of Maryland, College Park
{gowthami, vsingla, jgeiping, tomg}@cs.umd.edu

🗽 New York University
goldblum@nyu.edu

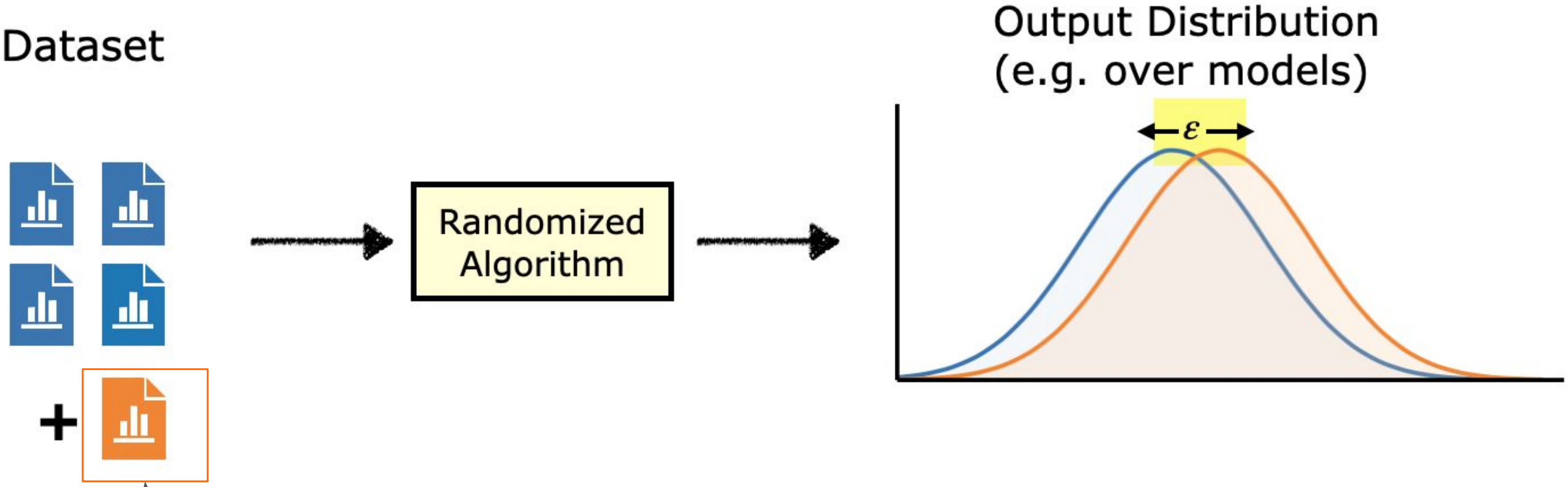


Differential privacy (DP)



Dwork, McSherry, Nissim, Smith. **Calibrating noise to sensitivity in private data analysis**. TCC 2006

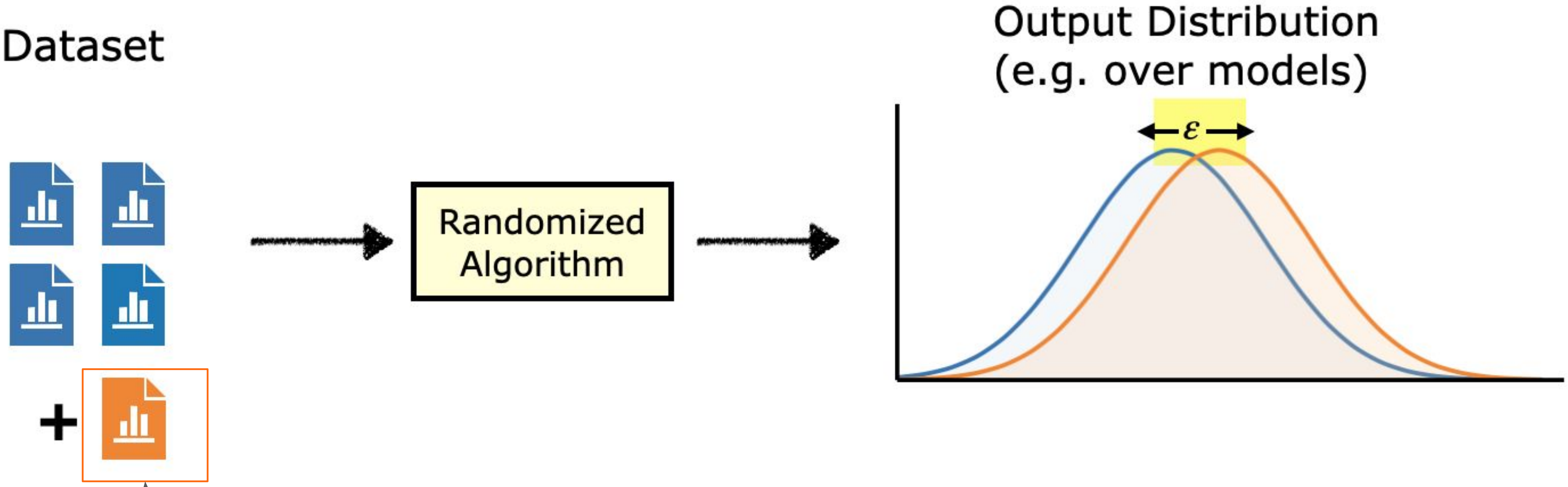
Differential privacy (DP)



A randomized algorithm is ε-**differentially private** if the addition of **one <u>unit of data</u>** does not alter its output distribution by more than ε

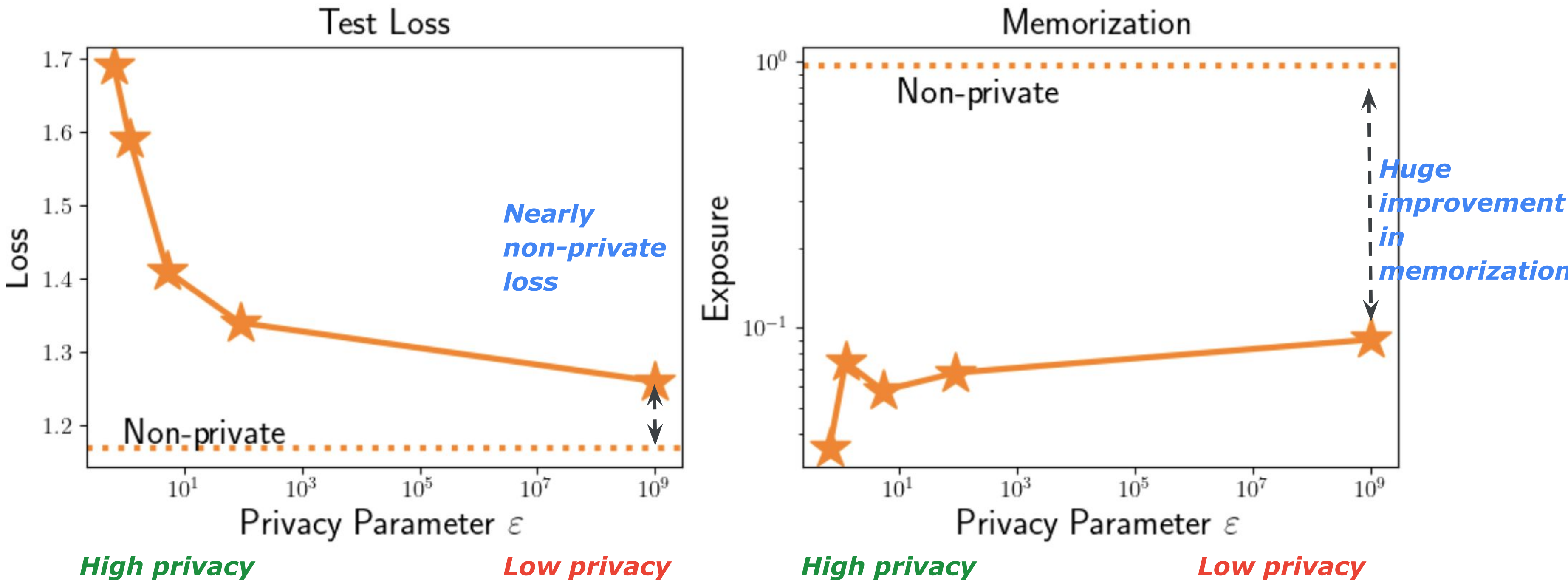
Example-level Differential privacy (DP)

Unit of data = ***example***



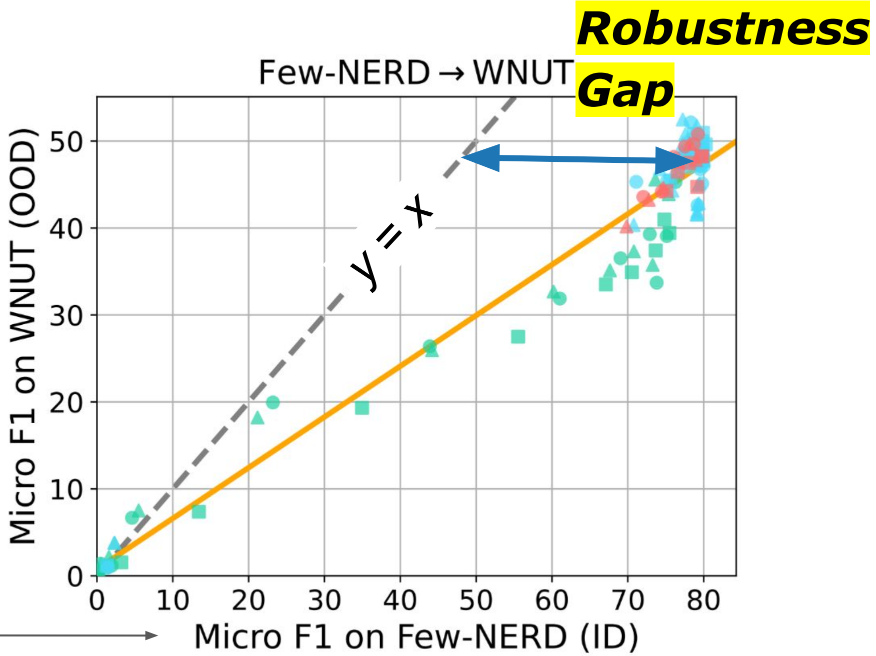
A randomized algorithm is ε-**differentially private** if the addition of **one <u>example</u>** does not alter its output distribution by more than ε

Differential privacy eliminates memorization



Carlini, Liu, Erlingsson, Kos, Song. **The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks.** USENIX Security 2019.

Which data do we use to train/finetune/align these models?



Test on **shifted distribution**
(out-of-domain / **OOD**)

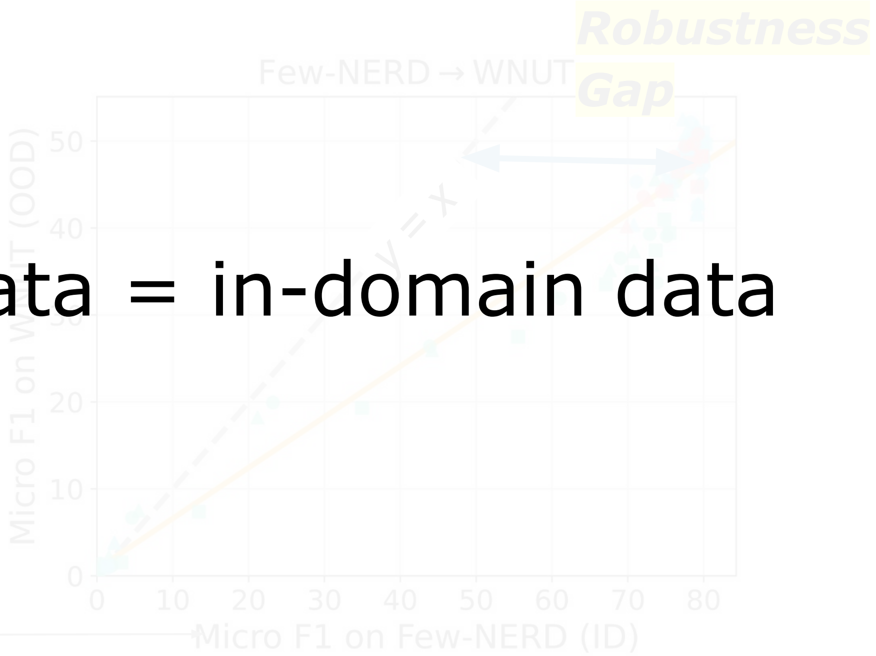
Test on **training distribution**
(in-domain / **ID**)

Yuan et al. **Revisiting Out-of-distribution Robustness in NLP: Benchmark, Analysis, and LLM Evaluations**. NeurIPS D&B 2023

Which data do we use to train/finetune/align these models?

Best training data = in-domain data

Test on **shifted distribution**
(out-of-domain / **OOD**)



Test on **training distribution**
(in-domain / **ID**)

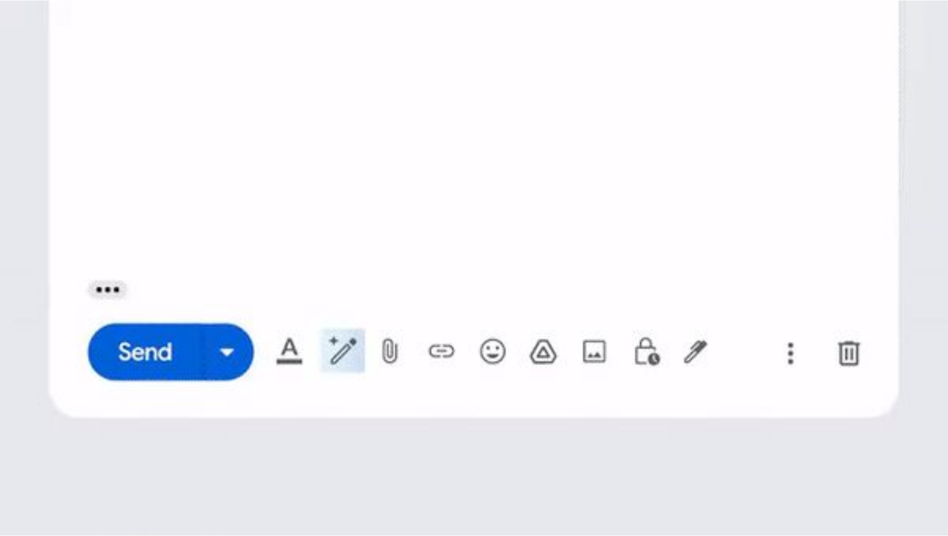
▲ Small-sized ● Base-sized ■ Large-sized — Linear Fit
Green Available Samples Blue Training Steps Red Tunable Parameters - - $y = x$

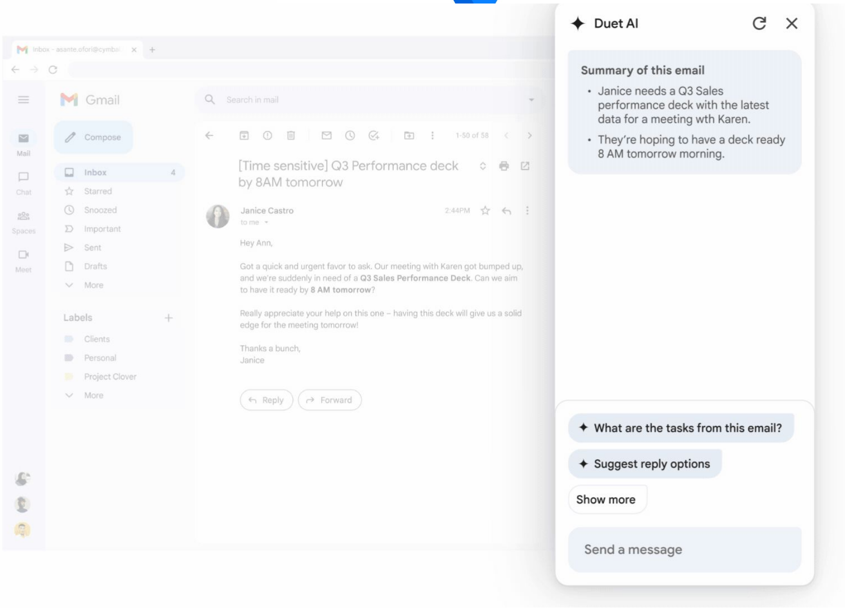
Yuan et al. **Revisiting Out-of-distribution Robustness in NLP: Benchmark, Analysis, and LLM Evaluations**. NeurIPS D&B 2023

Gmail

Gemini

Google Workspace





https://blog.google/products/gmail/gmail-ai-features/

https://workspace.google.com/blog/product-announcements/duet-ai-in-workspace-now-available

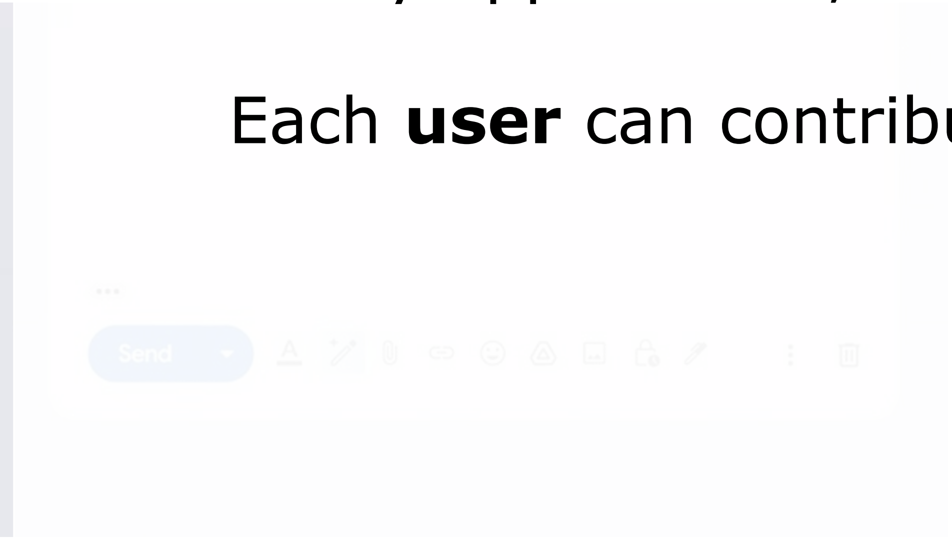
For many applications, in-domain data = **user data**

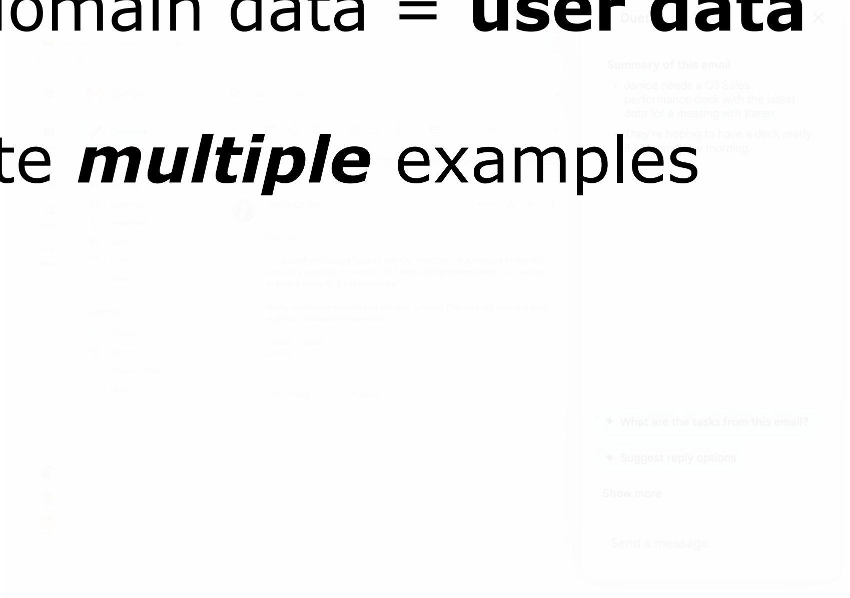
https://blog.google/products/gmail/gmail-ai-features/

https://workspace.google.com/blog/product-announcements/duet-ai-in-workspace-now-available

For many applications, in-domain data = **user data**

Each **user** can contribute ***multiple*** examples



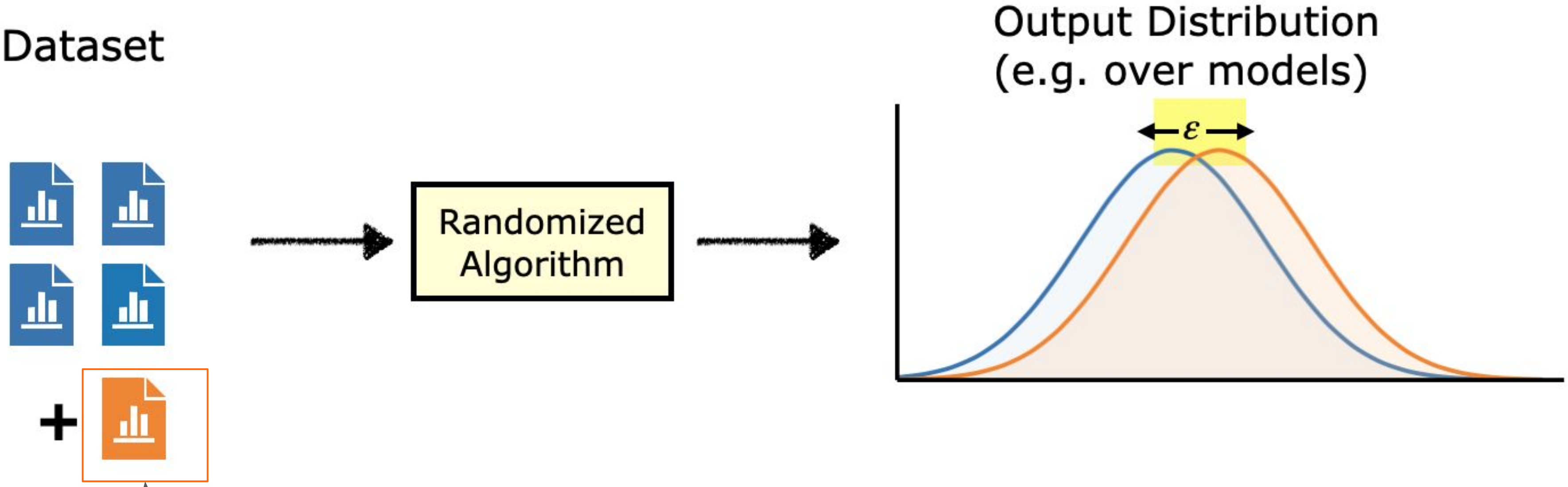


https://blog.google/products/gmail/gmail-ai-features/

https://workspace.google.com/blog/product-announcements/duet-ai-in-workspace-now-available

Example-level Differential privacy (DP)

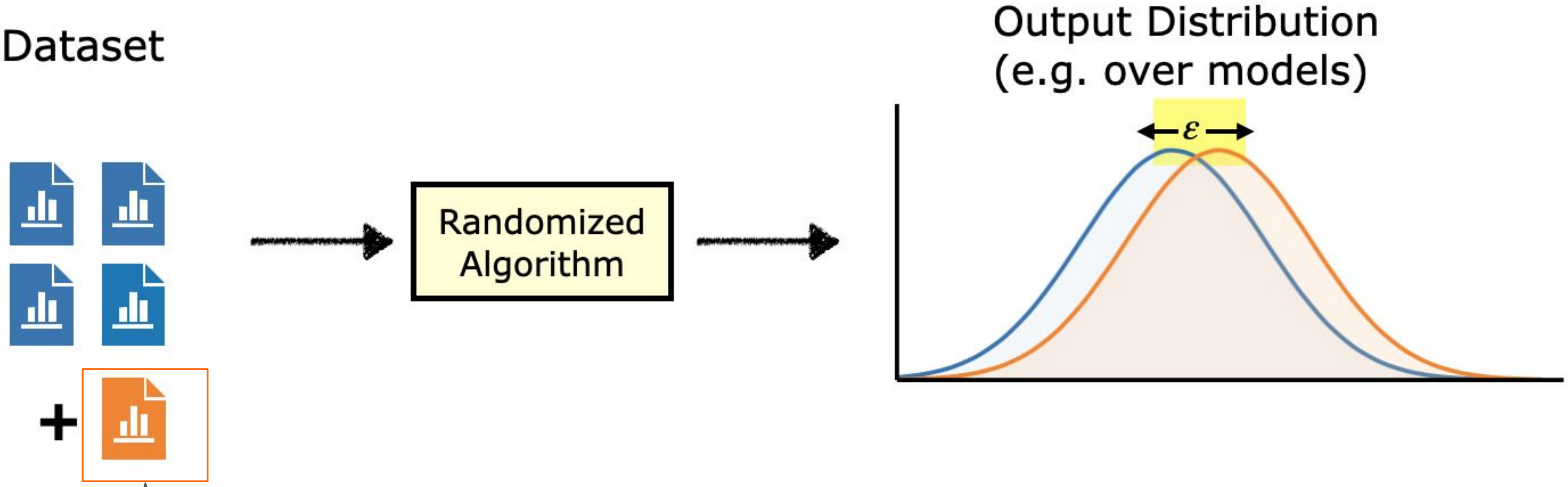
Unit of data
= ***example***



A randomized algorithm is ε-**differentially private** if the addition of **one example** does not alter its output distribution by more than ε

User ~~Example~~-level Differential privacy (DP)

Unit of data = ***user***



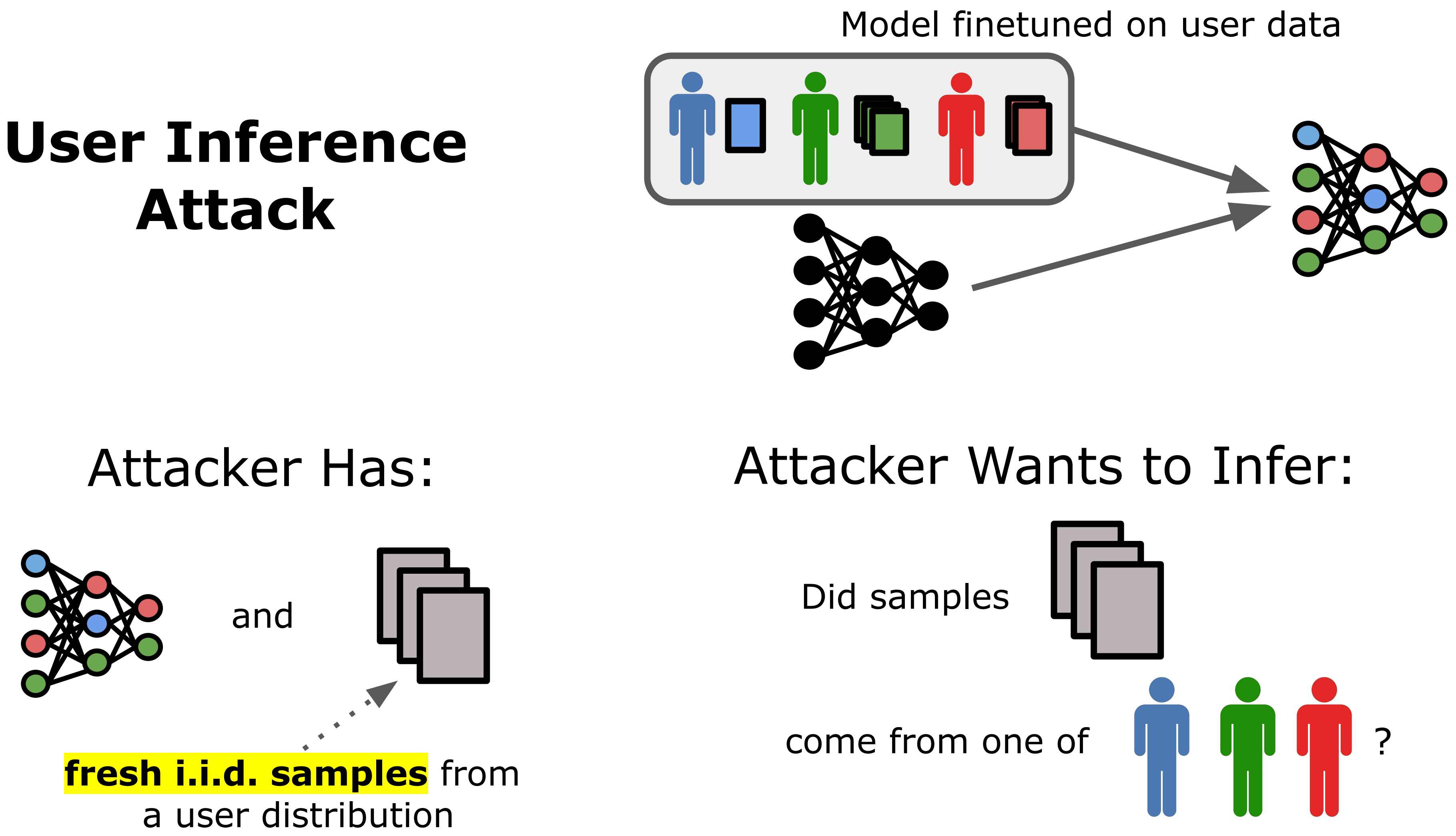
A randomized algorithm is ε-**differentially private** if the addition of **one user's data** does not alter its output distribution by more than ε

Why do we need user-level DP?

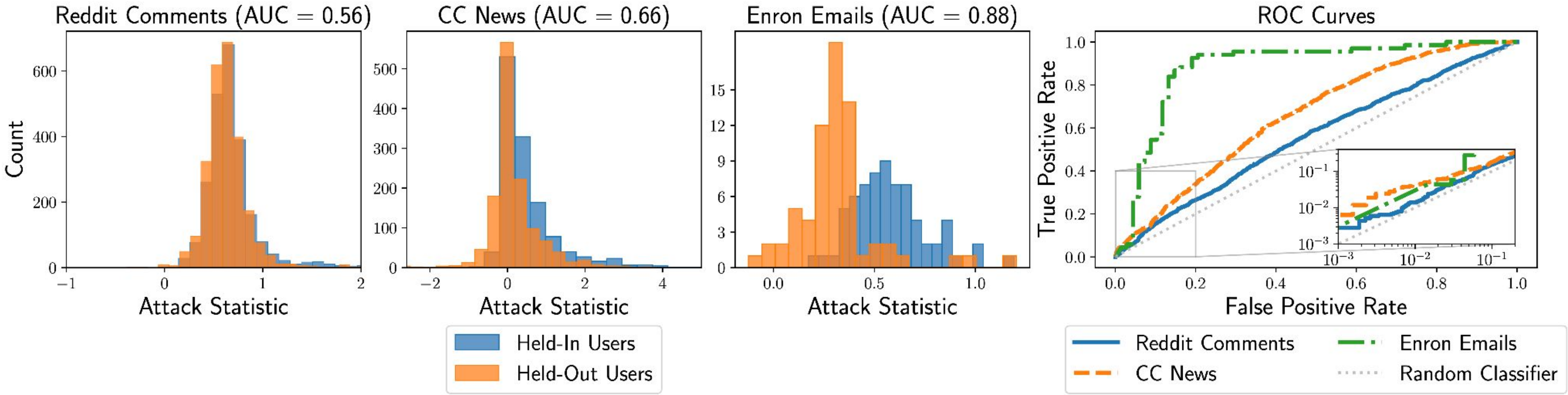
Why do we need user-level DP?

Standard LLM finetuning pipelines are susceptible to
user inference attacks*!*

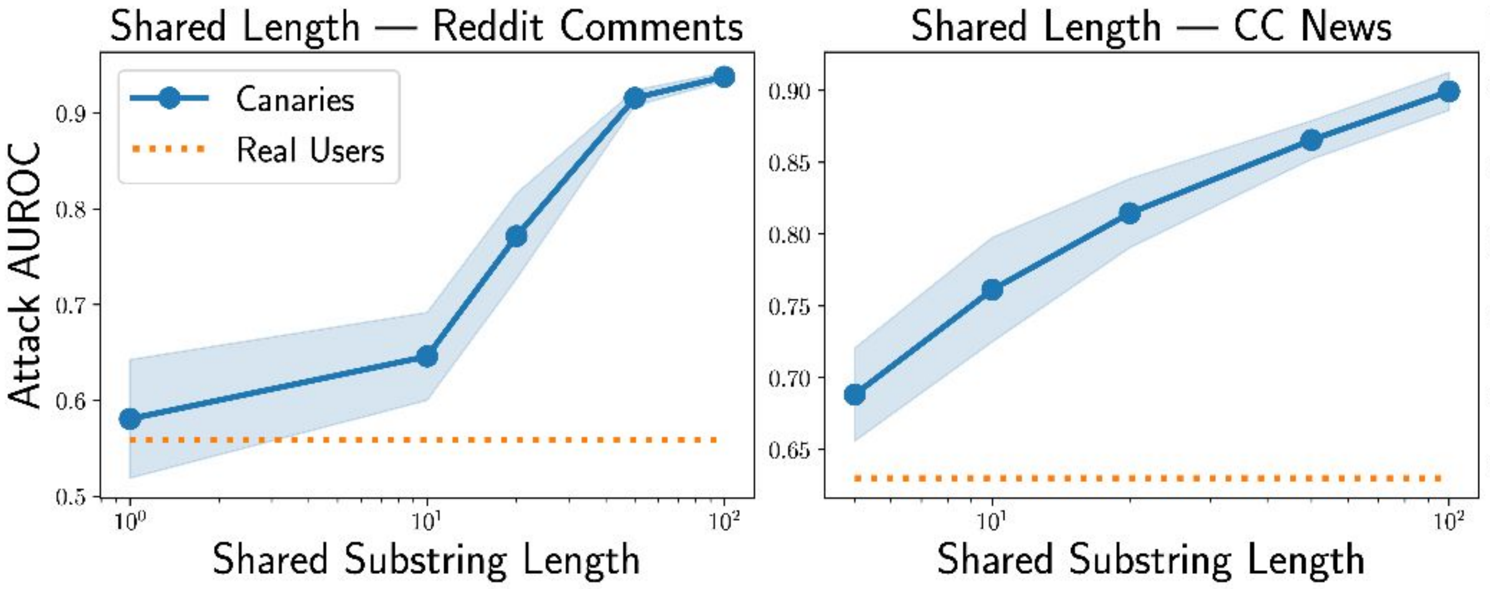
Nikhil Kandpal, ***KP***, Alina Oprea, Peter Kairouz, Chris Choquette-Choo, Zheng Xu.
Submitted (2024)


User Inference Attack
Model finetuned on user data
Attacker Has:
and
fresh i.i.d. samples from a user distribution
Attacker Wants to Infer:
Did samples
come from one of
?

User inference is effective when #users is small and data per user is large



Short common phrases can exacerbate user inference



Example-level DP offers limited mitigation

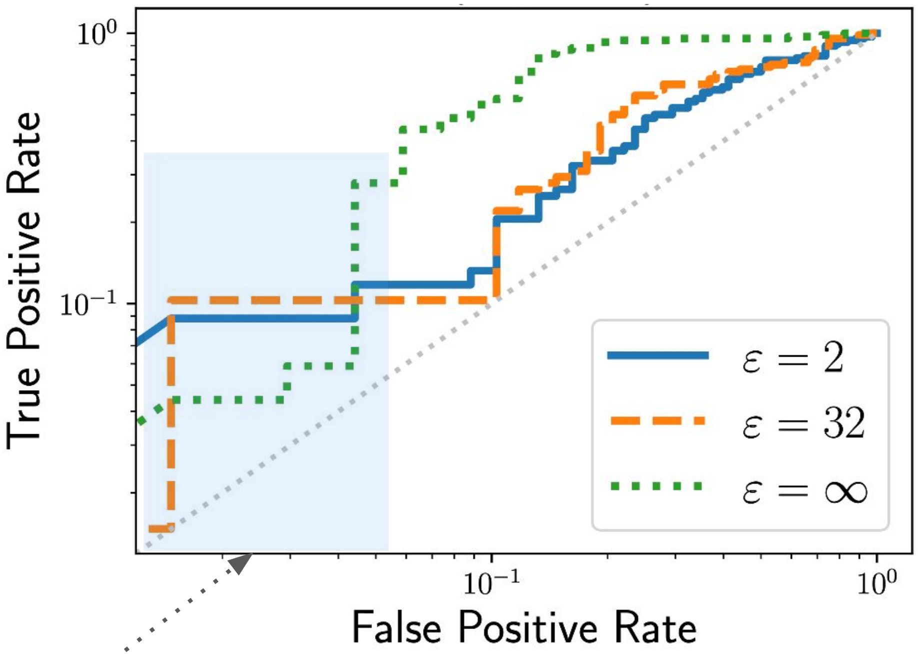
AUROC:

- non-private: 88%
- ε = 32: 70%

Utility:

- DP model reaches what the private model achieves in 1/3 epoch

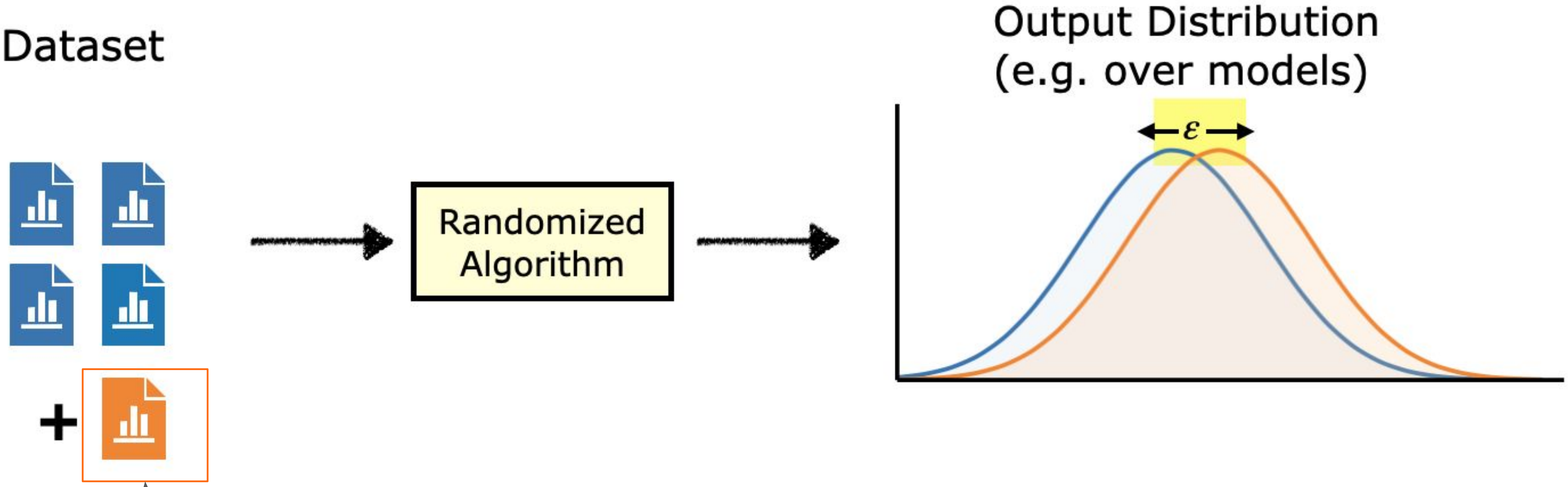
ROC Curves for Enron Emails



Example-level DP does not help here

User ~~Example~~-level **Differential privacy (DP)**

Unit of data = ***user***



A randomized algorithm is ε-**differentially private** if the addition of **one user's data** does not alter its output distribution by more than ε

How do we realize user-level DP?

Outline: how do we realize user-level DP?

Learning algorithms:

(Anti-) correlated noise provably beats independent noise

For linear regression, dimension d improves to problem-dependent ***effective dimension*** *d_{eff}*

Independent noise	$\Theta(d)$
Correlated noise	$\tilde{O}(d_{\text{eff}})$
Lower bound	$\Omega(d_{\text{eff}})$

Outline: how do we realize user-level DP?

Learning algorithms:

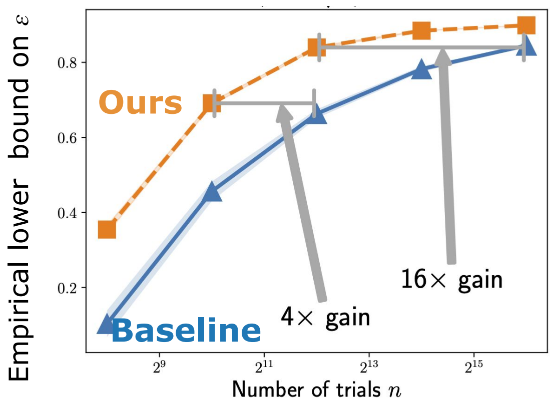
(Anti-) correlated noise provably beats independent noise

For linear regression, dimension d improves to problem-dependent ***effective dimension*** d_{eff}

Independent noise	$\Theta(d)$
Correlated noise	$\tilde{O}(d_{\mathrm{eff}})$
Lower bound	$\Omega(d_{\mathrm{eff}})$

Auditing:

Randomness makes the audit more computationally efficient



Part 1: How do we learn with user-level DP?

(Anti-)correlated noise **provably** beats independent noise

ICLR 2024

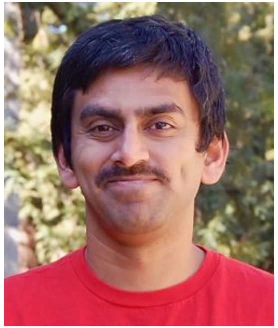
Chris Choquette-Choo*

Dj Dvijotham*

Krishna Pillutla*

Arun Ganesh

Thomas Steinke



Abhradeep Thakurta

*Equal contribution, $\alpha\beta$-order

DP-SGD: How do we train models with **example**-level DP?

Stochastic gradient clipped to ‖*g*‖ ≤ *G* **per-example**

Independent Gaussian noise

$$\theta_{t+1} = \theta_t - \eta \left(g_t + z_t \right)$$

Learning rate

Google Research

Song et al. (2013), Bassily et al. (FOCS 2014), Abadi et al. (CCS 2016)

DP-FedAvg: How do we train models with **user**-level DP?

Stochastic gradient clipped to $\|g\| \leq G$ **per-user**

Independent Gaussian noise

$$\theta_{t+1} = \theta_t - \eta \left(g_t + z_t \right)$$

Learning rate

Google Research

McMahan, Ramage, Talwar, Zhang. **Learning differentially private recurrent language models**. ICLR 2018

DP-SGD: DP Training with ***Independent*** Noise

For ϱ-zCDP, take noise variance = $\frac{G^2}{2\rho}$

(G = gradient clip norm)

Independent Gaussian noise

$$\theta_{t+1} = \theta_t - \eta \left(g_t + z_t \right)$$

Google Research

DP-FTRL: DP Training with ***Correlated*** Noise

Correlated
Gaussian noise
(z_t i.i.d. Gaussian)

$$\theta_{t+1} = \theta_t - \eta \left(g_t + \boxed{\sum_{\tau=0}^{t} \beta_{t,\tau} z_{t-\tau}} \right)$$

Kairouz, McMahan, Song, Thakkar, Thakurta, Xu. **Practical and Private (Deep) Learning without Sampling or Shuffling**. ICML 2021.
Denisov, McMahan, Rush, Smith, Thakurta. **Improved Differential Privacy for SGD via Optimal Private Linear Operators on Adaptive Streams**. NeurIPS 2022.

DP-FTRL: DP Training with ***Correlated*** Noise

For ϱ-zCDP, take noise variance = $\frac{G^2}{2\rho} \underbrace{\max_t \left\| [B^{-1}]_{:,t} \right\|_2^2}_{\text{sensitivity}}$

$$B = \begin{pmatrix} \beta_{0,0} & 0 & 0 & \cdots \\ \beta_{1,0} & \beta_{1,1} & 0 & \cdots \\ \beta_{2,0} & \beta_{2,1} & \beta_{2,2} & \cdots \\ \vdots & & & \end{pmatrix}$$

Correlated
Gaussian noise
(z_t i.i.d. Gaussian)

$$\theta_{t+1} = \theta_t - \eta \left(g_t + \boxed{\sum_{\tau=0}^{t} \beta_{t,\tau} z_{t-\tau}} \right)$$

Kairouz, McMahan, Song, Thakkar, Thakurta, Xu. **Practical and Private (Deep) Learning without Sampling or Shuffling**. ICML 2021.
Denisov, McMahan, Rush, Smith, Thakurta. **Improved Differential Privacy for SGD via Optimal Private Linear Operators on Adaptive Streams**. NeurIPS 2022.

Production Training

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

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

Google AI Blog

The latest from Google Research

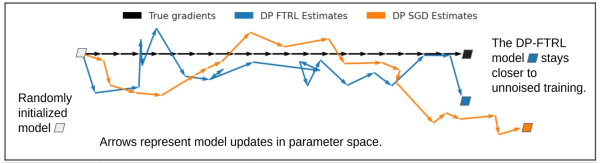
Federated Learning with Formal Differential Privacy Guarantees

Monday, February 28, 2022

Posted by Brendan McMahan and Abhradeep Thakurta, Research Scientists, Google Research

In 2017, Google introduced federated learning (FL), an approach that enables mobile devices to collaboratively train machine learning (ML) models while keeping the raw training data on each user's device, decoupling the ability to do ML from the need to store the data in the cloud. Since its introduction, Google has continued to actively engage in FL research and deployed FL to power many features in Gboard, including next word prediction, emoji suggestion and out-of-vocabulary word discovery. Federated learning is improving the "Hey Google" detection models in Assistant, suggesting replies in Google Messages, predicting text selections, and more.

While FL allows ML without raw data collection, differential privacy (DP) provides a quantifiable measure of data anonymization, and when applied to ML can address concerns about models memorizing sensitive user data. This too has been a top research priority, and has yielded one of the first production uses of DP for analytics with RAPPOR in 2014, our open-source DP library, Pipeline DP, and TensorFlow Privacy.



Data Minimization and Anonymization in Federated Learning

Along with fundamentals like transparency and consent, the privacy principles of data minimization and anonymization are important in ML applications that involve sensitive data.

Do we use ***independent*** or ***correlated*** noise?

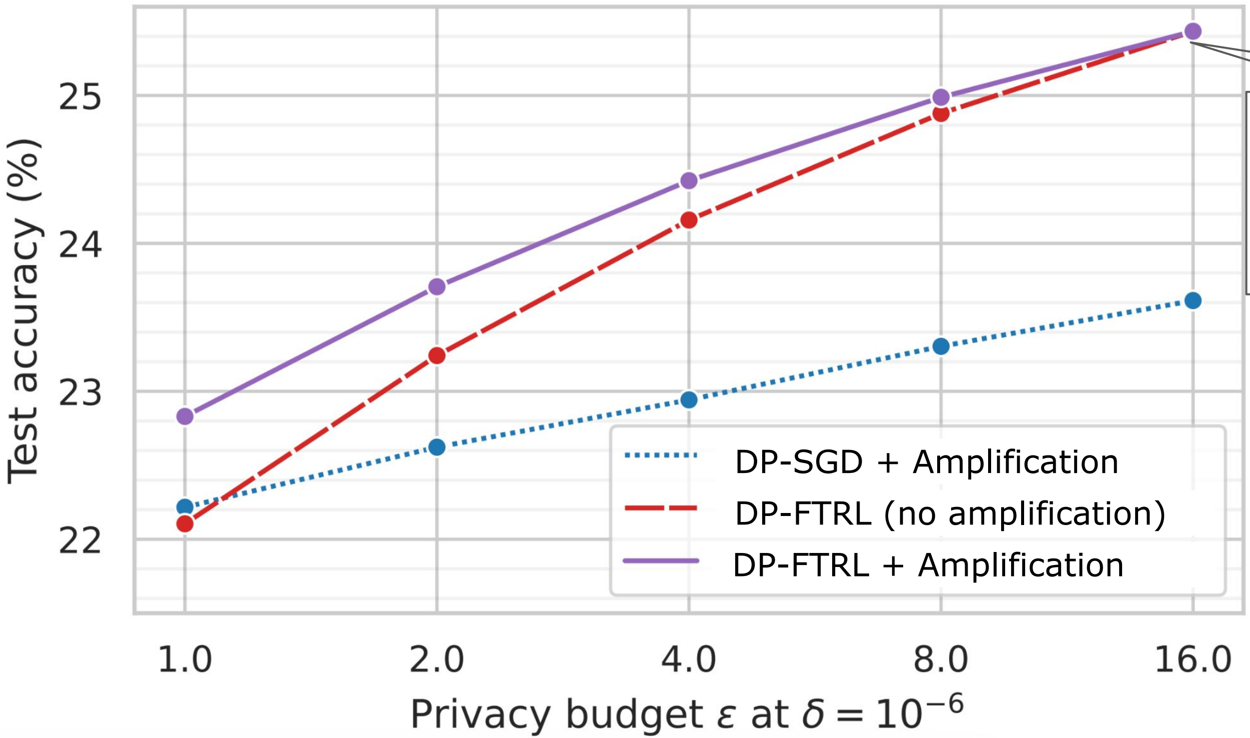
DP-SGD

DP-FTRL

Google Research

Prior work: (Empirically) correlated noise outperforms independent noise

Experiment:
User-level DP with StackOverflow



DP-FTRL (+ amplification) uniformly beats **DP-SGD**

Google Research

Choquette-Choo, Ganesh, McKenna, McMahan, Rush, Thakurta, Xu.
(Amplified) Banded Matrix Factorization: A unified approach to private training. NeurIPS 2023

Our goal: a ***provable*** gap between DP-SGD & DP-FTRL

DP-FTRL vs. DP-SGD: Previous Theory

For convex & *G*-Lipschitz losses

DP-SGD	$\frac{Gd^{1/4}}{\sqrt{\rho T}}$
DP-FTRL	$\frac{Gd^{1/4}}{\sqrt{\rho^2 T}}$

ϱ: privacy level (zCDP)
d: dimension
T: #iterations

Kairouz, McMahan, Song, Thakkar, Thakurta, Xu. **Practical and Private (Deep) Learning without Sampling or Shuffling**. ICML 2021.

Setting and Simplifications

Loss function

Model parameters

$$\min_{\theta}\left[F(\theta) = \mathbb{E}_{x\sim P}\left[f(\theta; x)\right]\right]$$

Data

Streaming setting: Suppose we draw a fresh data point $x_t \sim P$ in each iteration t (i.e. only 1 epoch)

Toeplitz noise correlations: $\beta_{t,\tau} = \beta_{\tau}$

$$\theta_{t+1} = \theta_t - \eta \left(g_t + \sum_{\tau=0}^{t} \beta_\tau z_{t-\tau} \right)$$

$$B = \begin{pmatrix} \beta_{0,0} & & & \\ \beta_{0,1} & \beta_{1,0} & & \\ \beta_{0,2} & \beta_{1,1} & \beta_{2,0} & \cdots \\ \vdots & & & \end{pmatrix} \longrightarrow B = \begin{pmatrix} \beta_{0} & & & \\ \beta_{1} & \beta_{0} & & \\ \beta_{2} & \beta_{1} & \beta_{0} & \cdots \\ \vdots & & & \end{pmatrix}$$

Computationally: store $O(T)$ coefficients instead of $O(T^2)$

Google Research

Asymptotics: Iterates converge to a stationary distribution as $t \to \infty$

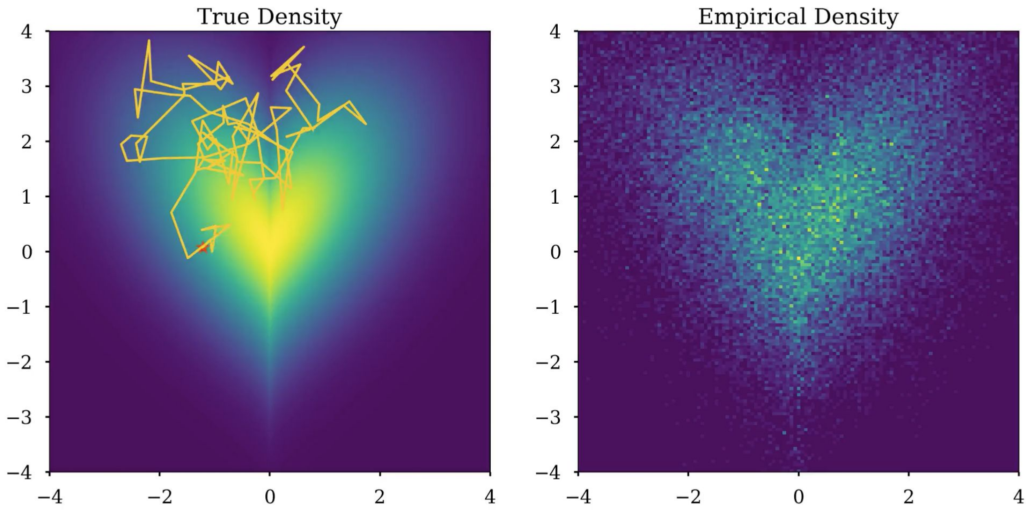


Image credit:
[Abdul Fatir Ansari]

Google Research

Asymptotics: Iterates converge to a stationary distribution as $t \to \infty$

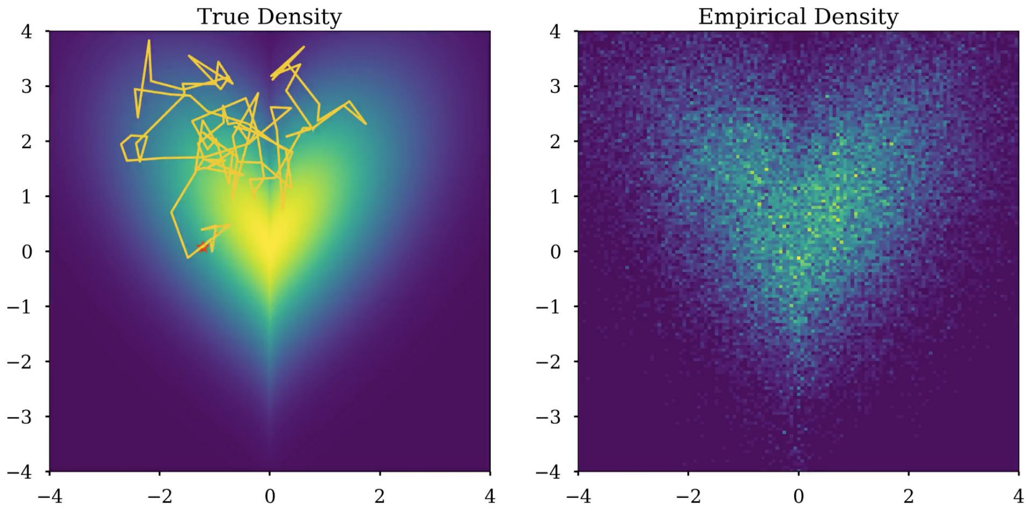


Image credit:
[Abdul Fatir Ansari]

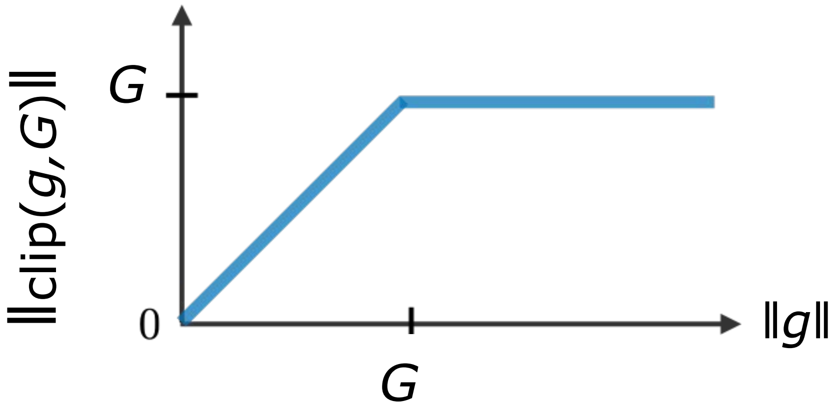
Asymptotic error

$$F_\infty(\beta) = \lim_{t \to \infty} \mathrm{E}\left[F(\theta_t) - F(\theta_\star)\right]$$

Asymptotics at a fixed learning rate $\eta > 0$

Google Research

Noisy-SGD/Noisy-FTRL: DP-SGD/DP-FTRL without clipping



Lets us study the noise dynamics of the algorithms
(do not satisfy DP guarantees)

Mean estimation in 1 dimension

$$\min_{\theta} \left[F(\theta) = \mathbb{E}_{x \sim P} \left(\theta - x\right)^2 \right]$$

Data distribution s.t. $|\boldsymbol{x}| \leq 1$

Solve with stochastic optimization problem with DP-SGD/DP-FTRL

Google Research

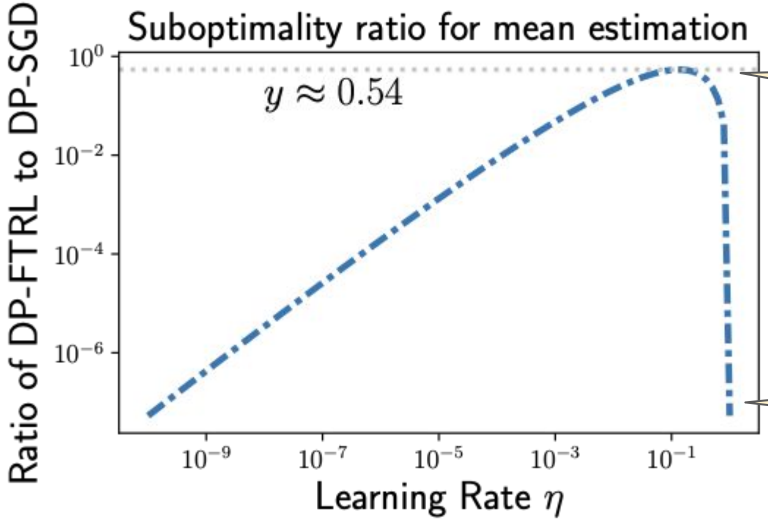
Mean estimation in 1 dimension

Informal Theorem: The asymptotic error of a ϱ-zCDP sequence is

Independent noise (DP-SGD)	$F_\infty(\beta^{\mathrm{sgd}}) = \rho^{-1}\eta$
Correlated noise (DP-FTRL)	$\inf_\beta F_\infty(\beta) = F_\infty(\beta^\star) = \rho^{-1}\eta^2 \log^2 \frac{1}{\eta}$

η: learning rate (constant and non-zero)
ϱ: privacy level

Google Research



DP-FTRL is always better than DP-SGD

DP-FTRL is significantly better at $\eta \to 0$ or $\eta \to 1$

Google Research

Closed form correlations for mean estimation

Proposition: The correlations $\beta_0^\star = 1, \quad \beta_t^\star = -t^{-3/2}(1-\eta)^t$ attain the optimal error

$$\inf_\beta F_\infty(\beta) = F_\infty(\beta^\star) = \rho^{-1}\eta^2 \log^2 \frac{1}{\eta}$$

Closed form correlations for mean estimation

Proposition: The correlations $\beta_0^\star = 1, \quad \beta_t^\star = -t^{-3/2}(1-\eta)^t$ attain the optimal error

$$\inf_\beta F_\infty(\beta) = F_\infty(\beta^\star) = \rho^{-1}\eta^2 \log^2 \frac{1}{\eta}$$

***ν*-DP-FTRL**

For general problems, use $\beta_0 = 1, \quad \beta_t = -t^{-3/2}(1-\nu)^t$

and tune the parameter ν

Linear regression

$$\min_{\theta} \left[F(\theta) = \mathbb{E}\big(y - \langle \theta, x \rangle\big)^2 \right]$$

$$\text{where} \qquad x \sim \mathcal{N}(0, H)$$

H is also the Hessian of the objective

Google Research

Linear regression

$$\min_{\theta} \left[F(\theta) = \mathbb{E}\big(y - \langle \theta, x \rangle\big)^2 \right]$$

$$\text{where} \qquad x \sim \mathcal{N}(0, H)$$

Well-specified linear model

$$y|x \sim \mathcal{N}(x^\top \theta_\star, \sigma^2)$$

Google Research

Informal Theorem: The asymptotic error is

Independent noise (Noisy-SGD)	$=$	$d \;\; \rho^{-1}\,\eta$
Correlated noise (ν-Noisy-FTRL)	$\leq$	$d_{\mathrm{eff}}\,\rho^{-1}\,\eta^2 \;\log^2\left(\frac{1}{\eta\mu}\right)$
Lower bound for any algorithm	$\geq$	$d_{\mathrm{eff}}\,\rho^{-1}\,\eta^2$

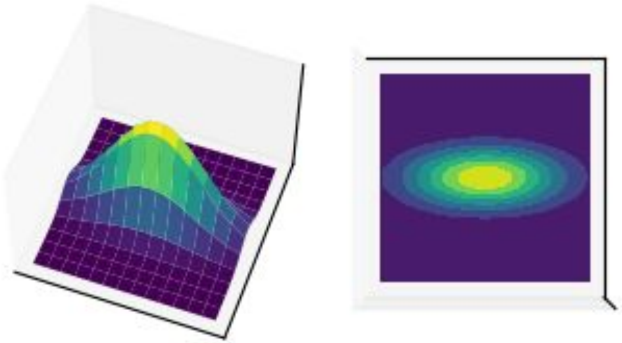
Improve dimension d to problem-dependent ***effective dimension*** d_{eff}

Effective dimension

$$d_{\text{eff}} = \mathbf{Tr}(H)/\|H\|_2 \leq d$$

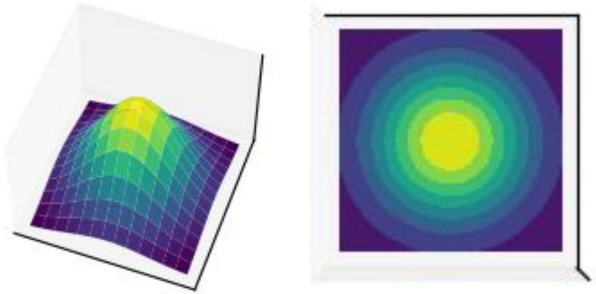
Low effective dimension

$$\lambda_1 = 1, \lambda_2 = \cdots = \lambda_d = 1/d$$



High effective dimension

$$\lambda_1 = \lambda_2 = \cdots = \lambda_d = 1$$



Closely connected to **numerical/stable rank**

SAMPLING FROM LARGE MATRICES: AN APPROACH THROUGH GEOMETRIC FUNCTIONAL ANALYSIS

MARK RUDELSON AND ROMAN VERSHYNIN

Remark 1.3 (Numerical rank). The numerical rank $r = r(A) = \|A\|_F^2 / \|A\|_2^2$ in Theorem 1.1 is a relaxation of the exact notion of rank. Indeed, one always has $r(A) \leq \text{rank}(A)$. But as opposed to the exact rank, the numerical rank is stable under small perturbations of the matrix A. In particular, the numerical rank of A tends to be low when A is close to a low rank matrix, or when A is sufficiently sparse.

$$d_{\text{eff}} = \text{srank}(H^{1/2})$$

Google Research

[Rudelson & Vershynin (J. ACM 2007)]

The stable rank appears in:

- Numerical linear algebra (e.g. randomized matrix multiplications) [Tropp (2014), Cohen-Nelson-Woodruff (2015)]
- Matrix concentration [Hsu-Kakade-Zhang (2012), Minsker (2017)]
- …

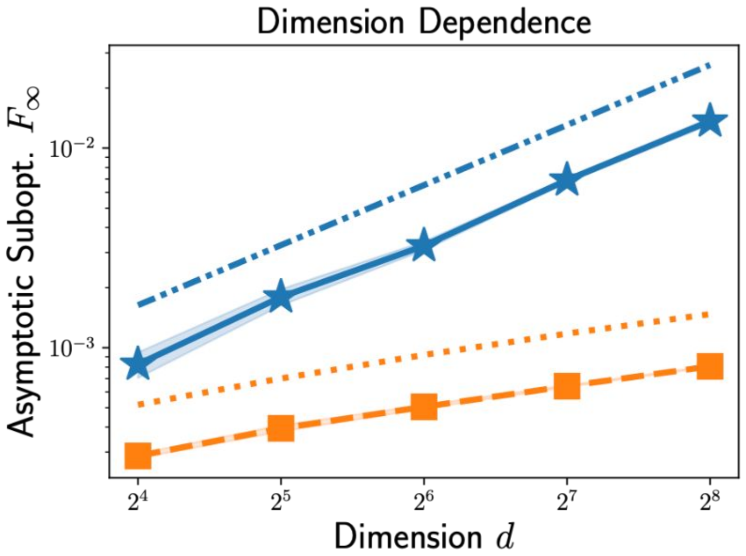
Informal Theorem: The asymptotic error is

Independent noise (Noisy-SGD)	$= \quad d \quad \rho^{-1}\,\eta$
Correlated noise (ν-Noisy-FTRL)	$\leq \quad d_{\mathrm{eff}}\,\rho^{-1}\,\eta^2\,\log^2\left(\frac{1}{\eta\mu}\right)$
Lower bound for any algorithm	$\geq \quad d_{\mathrm{eff}}\,\rho^{-1}\,\eta^2$

Improve dimension d to problem-dependent ***effective dimension*** *d_{eff}*

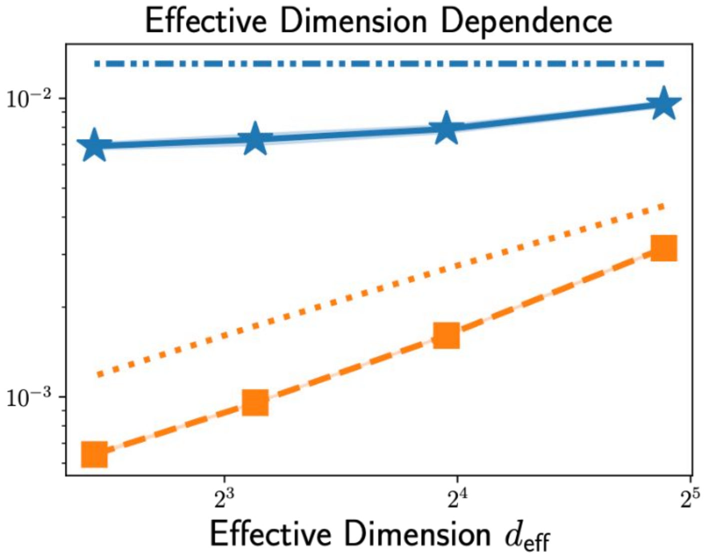
Google Research

Linear regression: theory predicts simulations



Noisy-SGD
scales with d

Noisy-FTRL
scales with d_{eff}

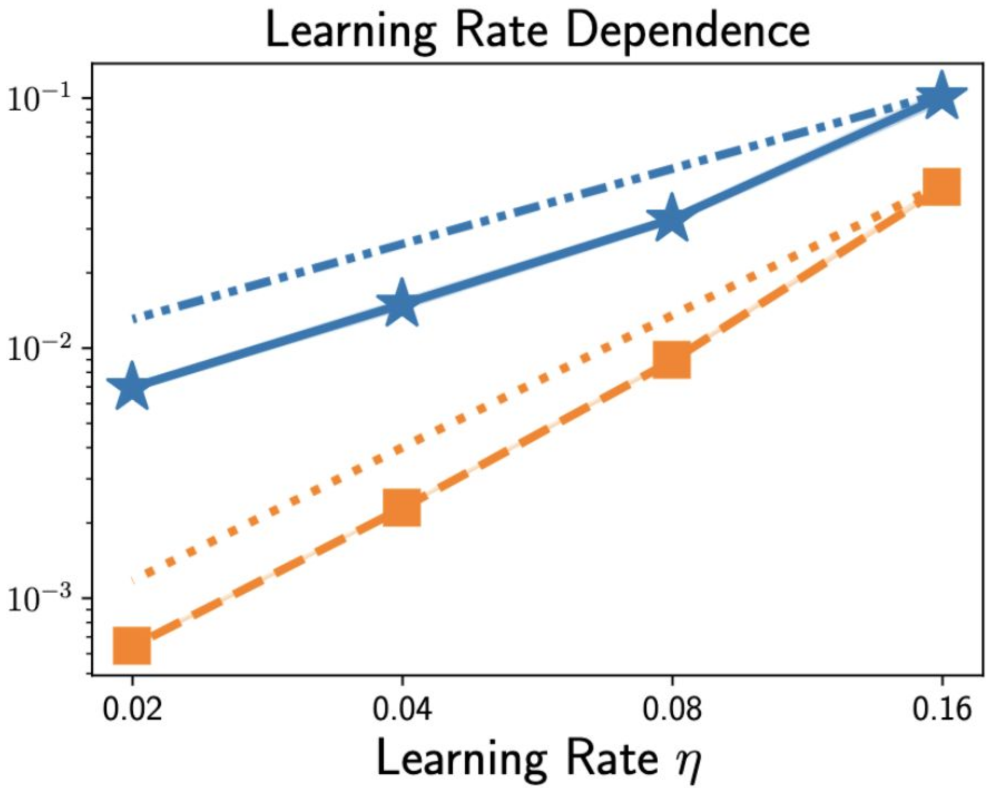


Informal Theorem: The asymptotic error for $0 < \eta < 1$ is

Independent noise (Noisy-SGD)	$= \; d \; \rho^{-1} \eta$
Correlated noise (ν-Noisy-FTRL)	$\leq \; d_{\text{eff}} \, \rho^{-1} \eta^2 \log^2\left(\frac{1}{\eta\mu}\right)$
Lower bound for any algorithm	$\geq \; d_{\text{eff}} \, \rho^{-1} \eta^2$

Improved dependence on the learning rate η

Google Research



Noisy-SGD scales as η

ν-**Noisy-FTRL** scales as η^2

Noisy-FTRL $\gg$ **Noisy-SGD** at small η

Google Research

Finite-time rates with DP: Linear Regression

Independent noise (DP-SGD)	$\frac{1}{\rho T} + \frac{1}{T}$
Correlated noise (ν-DP-FTRL)	$\frac{1}{\rho T^2} + \frac{1}{T}$

Privacy error

T: number of iterations
ϱ: privacy level
η: learning rate is optimized

Google Research

Proof sketch for Mean Estimation

Updates are not Markovian (key for all stochastic gradient proofs)

Our approach: Analysis the Fourier domain

Letting $\boldsymbol{\delta}_t = \theta_t - \theta_*$, the DP-FTRL update can be written as

Linear Time-Invariant (LTI) system

$$\delta_{t+1} = (1-\eta)\delta_t - \eta \sum_{\tau=0}^{t} \beta_\tau z_{t-\tau}$$

Convolution of the noise

Fourier analysis can give the stationary variance of $\boldsymbol{\delta}_t$ in terms of the **discrete-time Fourier transform** $B(\omega) = \sum_{t=0}^{\infty} \beta_t e^{i\omega t}$ of the convolution weights β

Frequency

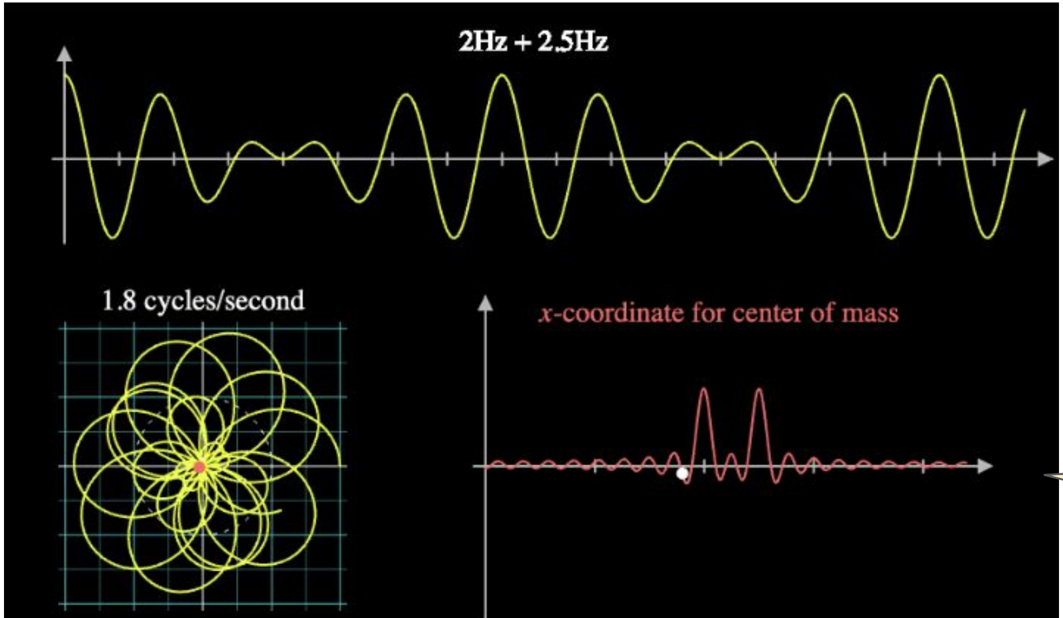


Image: 3blue1brown.com/lessons/fourier-transforms

Time-domain description

Frequency-domain description

Google Research

Letting $\boldsymbol{\delta}_t = \theta_t - \theta_*$, the DP-FTRL update can be written as

Linear Time-Invariant (LTI) system

$$\delta_{t+1} = (1-\eta)\delta_t - \eta \sum_{\tau=0}^{t} \beta_\tau z_{t-\tau}$$

Convolution of the noise

The stationary variance of $\boldsymbol{\delta}_t$ can be given as

$$\lim_{t\to\infty} \mathbb{E}[\delta_t^2] = \frac{\eta^2}{2\pi}\left(\int_{-\pi}^{\pi} \frac{|B(\omega)|^2}{|1-\eta-e^{i\omega}|^2}\mathrm{d}\omega\right) \quad \mathbb{E}[z_t^2]$$

Google Research

$$\lim_{t\to\infty} \mathbb{E}[\delta_t^2] = \frac{\eta^2}{2\pi}\left(\int_{-\pi}^{\pi} \frac{|B(\omega)|^2}{|1-\eta-e^{i\omega}|^2}\mathrm{d}\omega\right) \quad \mathbb{E}[z_t^2]$$

sensitivity

For ϱ-zCDP, take

$$\mathbb{E}[z_t^2] = \frac{1}{2\rho} \max_t \left\|[B^{-1}]_{:,t}\right\|_2^2$$

$$= \frac{1}{2\rho}\int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{2\pi|B(\omega)|^2}$$

$$B = \begin{pmatrix} \beta_0 & & & \\ \beta_1 & \beta_0 & & \\ \beta_2 & \beta_1 & \beta_0 & \cdots \\ \vdots & & & \end{pmatrix}$$

$$\lim_{t\to\infty} \mathbb{E}[\delta_t^2] = \frac{\eta^2}{2\pi}\left(\int_{-\pi}^{\pi} \frac{|B(\omega)|^2}{|1-\eta-e^{i\omega}|^2} \mathrm{d}\omega\right) \quad \mathbb{E}[z_t^2]$$

Requires $|B(\omega)|$ **small**

sensitivity

For ϱ-zCDP, take
$$\mathbb{E}[z_t^2] = \frac{1}{2\rho} \max_t \left\|[B^{-1}]_{:,t}\right\|_2^2 = \frac{1}{2\rho}\int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{2\pi|B(\omega)|^2}$$

$$B = \begin{pmatrix} \beta_0 & & & \\ \beta_1 & \beta_0 & & \\ \beta_2 & \beta_1 & \beta_0 & \cdots \\ \vdots & & & \end{pmatrix}$$

Requires $|B(\omega)|$ **large**

Google Research

$$\lim_{t\to\infty} \mathbb{E}[\delta_t^2] = \frac{\eta^2}{2\pi}\left(\int_{-\pi}^{\pi} \frac{|B(\omega)|^2}{|1-\eta-e^{i\omega}|^2}\mathrm{d}\omega\right) \quad \mathbb{E}[z_t^2]$$

Requires $|B(\omega)|$ **small**

sensitivity

For ϱ-zCDP, take
$$\mathbb{E}[z_t^2] = \frac{1}{2\rho} \max_t \left\|[B^{-1}]_{:,t}\right\|_2^2 = \frac{1}{2\rho}\int_{-\pi}^{\pi}\frac{\mathrm{d}\omega}{2\pi|B(\omega)|^2}$$

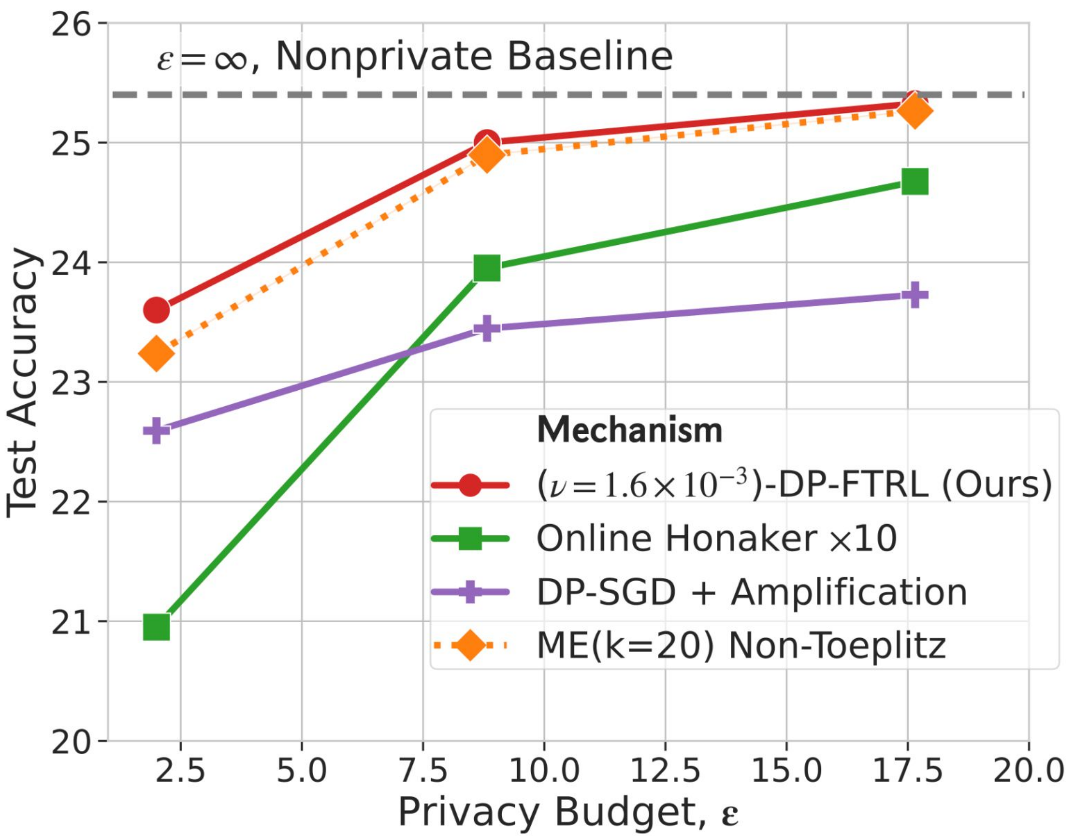
$$B = \begin{pmatrix} \beta_0 & & & \\ \beta_1 & \beta_0 & & \\ \beta_2 & \beta_1 & \beta_0 & \cdots \\ \vdots & & & \end{pmatrix}$$

Requires $|B(\omega)|$ **large**

Optimizing for $|B(\omega)|$ gives the theorem

Google Research

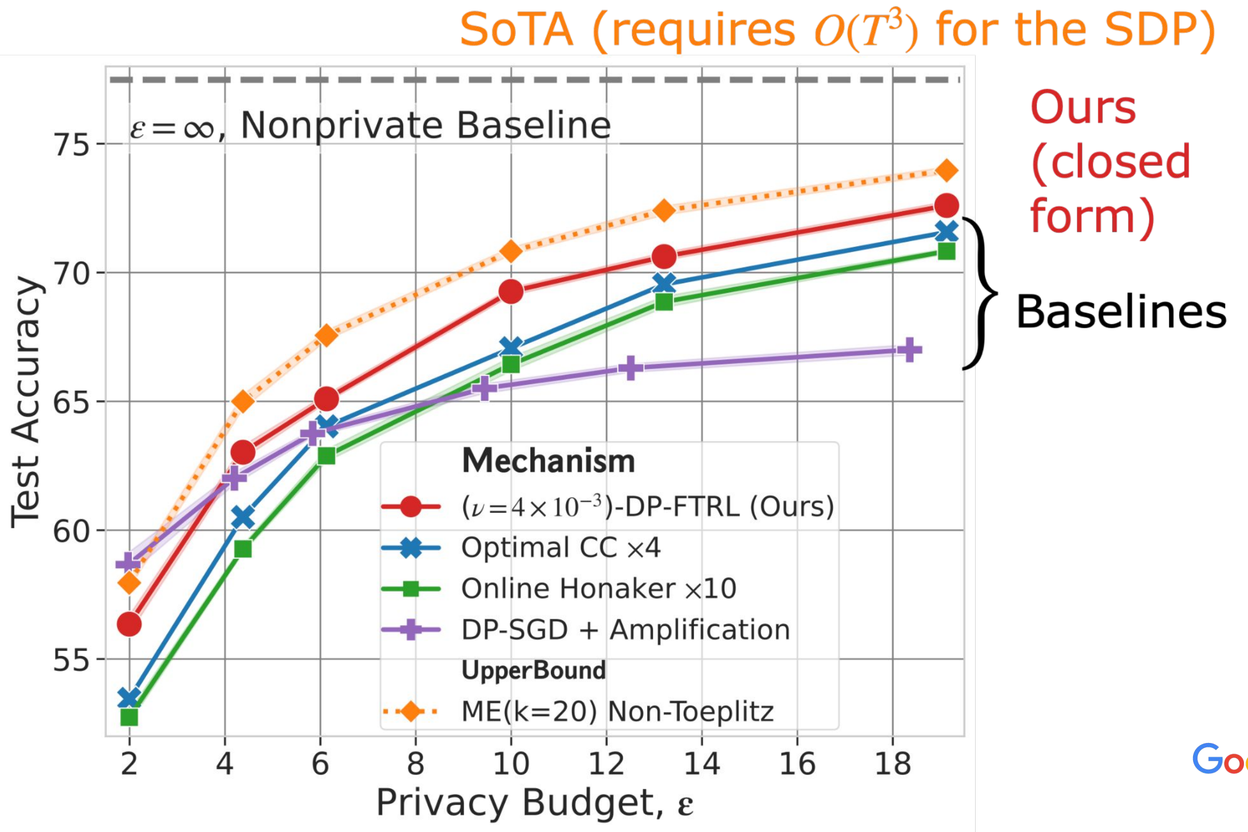
Language modeling with Stack Overflow | User-level DP



Ours matches SoTA!

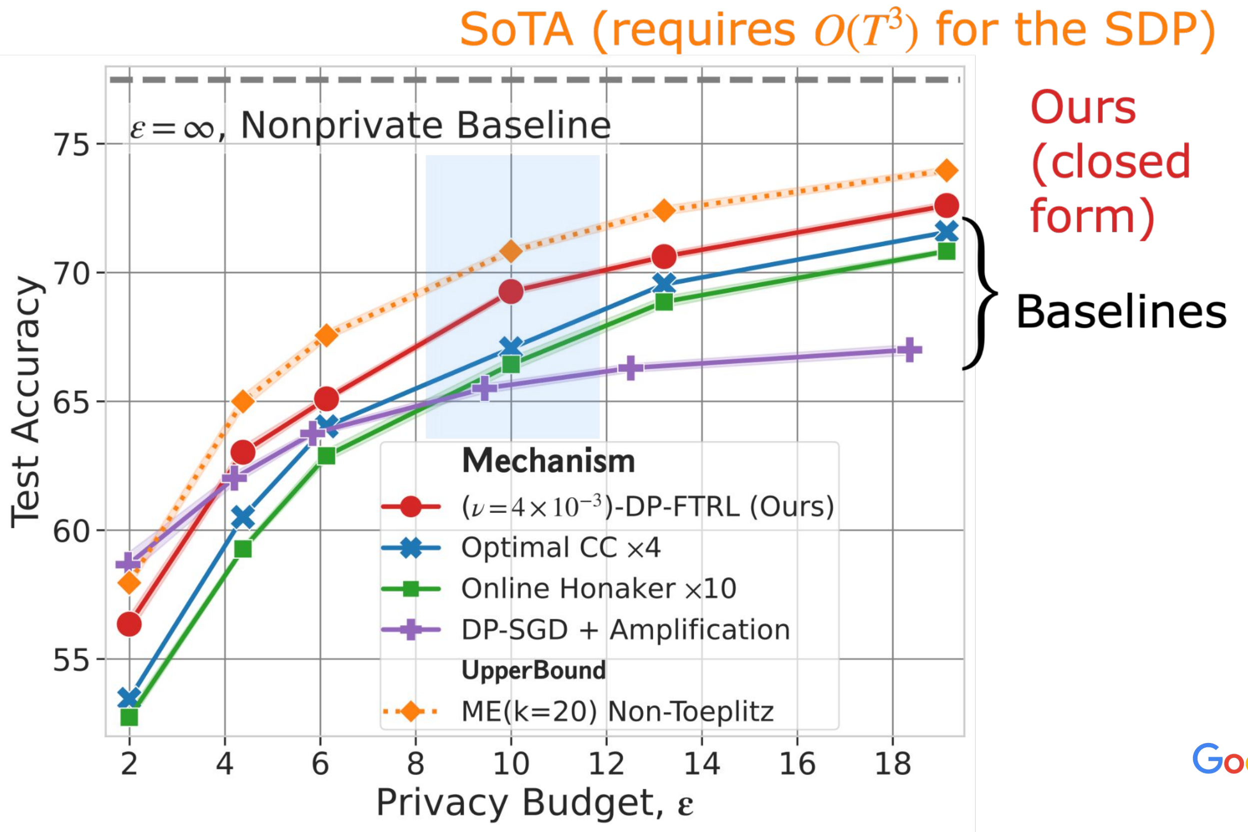
Google Research

Image classification with CIFAR-10 | Example-level DP



Google Research

Image classification with CIFAR-10 | Example-level DP



Computational cost

- **SoTA**: cubic complexity to generate the β's
- **Ours**: linear complexity (closed form)
  - nearly matches SoTA empirically

Summary

- Correlated noise is **provably** better
- Depends on effective dimension instead of dimension
- Matches lower bounds

Google Research

Part 2: How audit user-level DP?

Unleashing the power of randomness in auditing DP

NeurIPS 2023

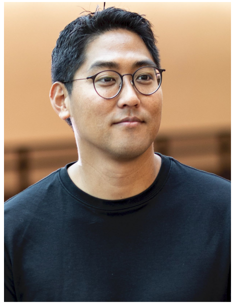
Krishna Pillutla

Galen Andrew

Peter Kairouz

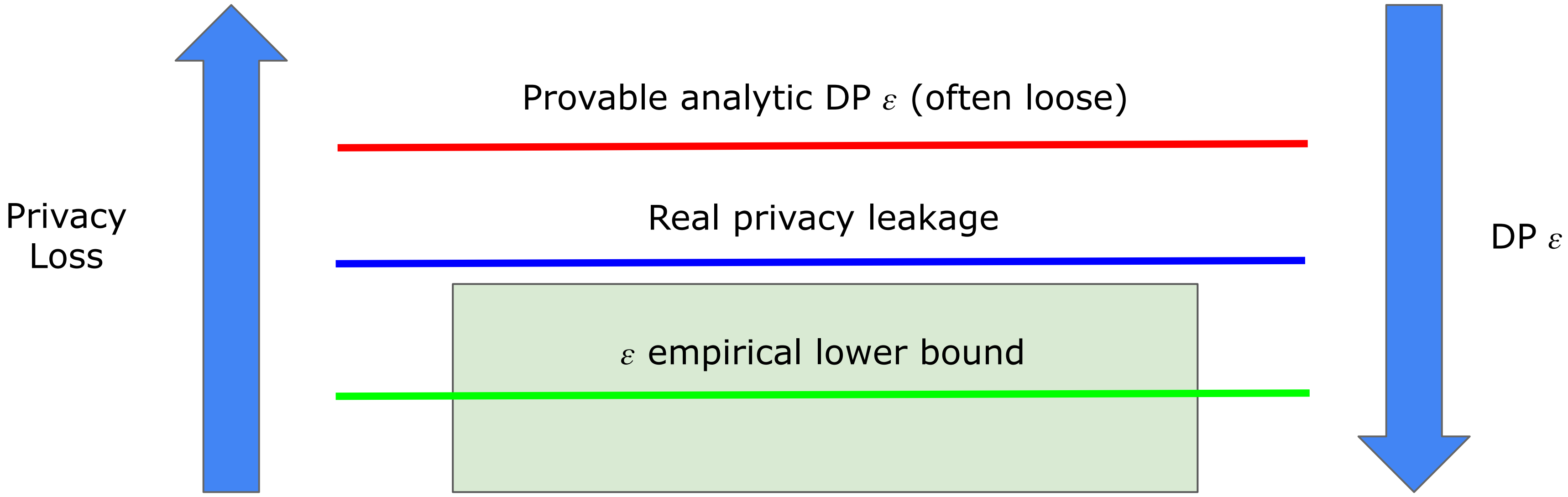
Brendan McMahan

Alina Oprea



Sewoong Oh

Empirical privacy auditing



Our focus

Why empirical privacy auditing?

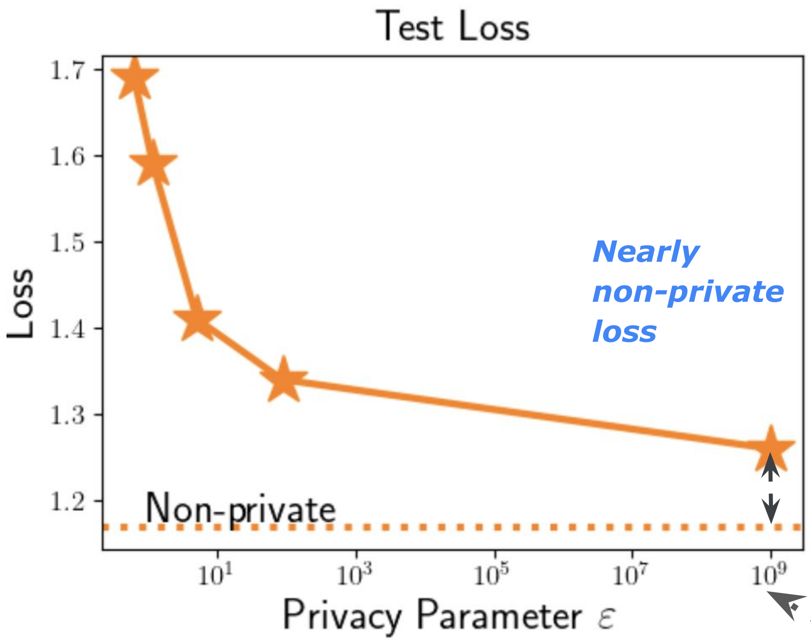
To verify that we actually provide the guarantee we claim (no bugs in proofs/implementation)

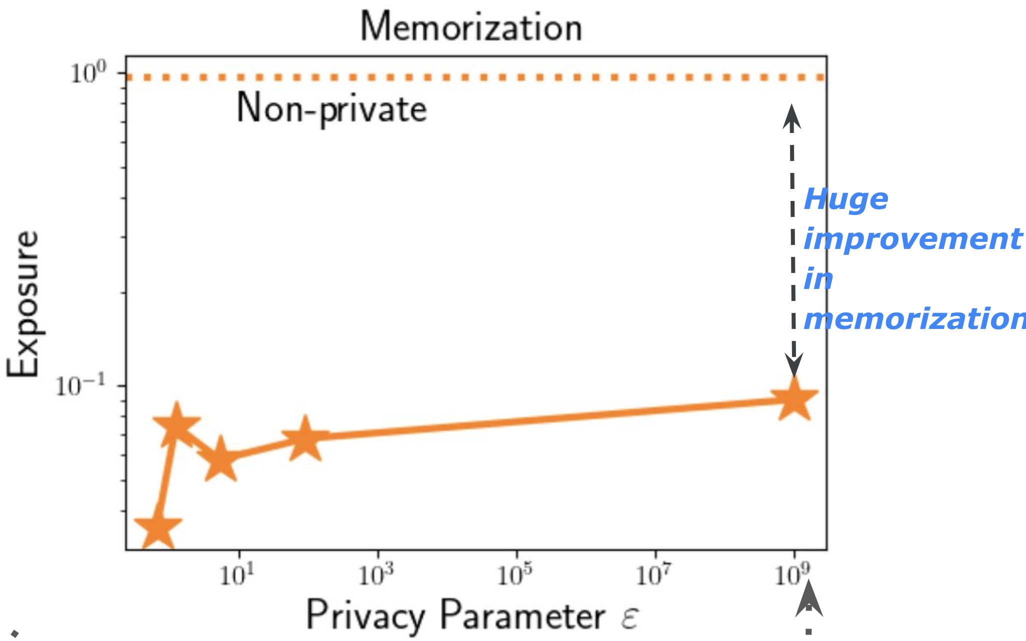
```
2 mnist_experiment.py

@@ -71,7 +71,7 @@ def forward(self, x):
                                                     rho_i,
                                                     epochs,
                                                     inp_clip,
-                                                    grad_clip
+                                                    grad_clip/BATCH_SIZE
                                                 )
                            tl, correct, set_len = uc.test(model, test_loader)
                            print(f'MNIST_{BATCH_SIZE}_{epochs}_{grad_clip}_{inp_clip}_{rho_i}', correct/set_len)
```

```
2 upstream_clipping.py

@@ -110,7 +110,7 @@ def run_experiment(model, train_loader, rho_i, epochs, input_bound, grad_bound):

      model.train()
      # sensitivity for everything with weights is just:
-     sensitivity = input_bound * grad_bound / train_loader.batch_size
+     sensitivity = input_bound * grad_bound
      sigma = np.sqrt(sensitivity**2 / (2*rho_i))
      print('sensitivity:', sensitivity)

```

Tramèr et al. **Debugging Differential Privacy: A Case Study for Privacy Auditing**. Preprint 2022

Gap between DP guarantees and empirical behavior: Memorization





High privacy ***Low privacy*** ***High privacy*** ***Low privacy***

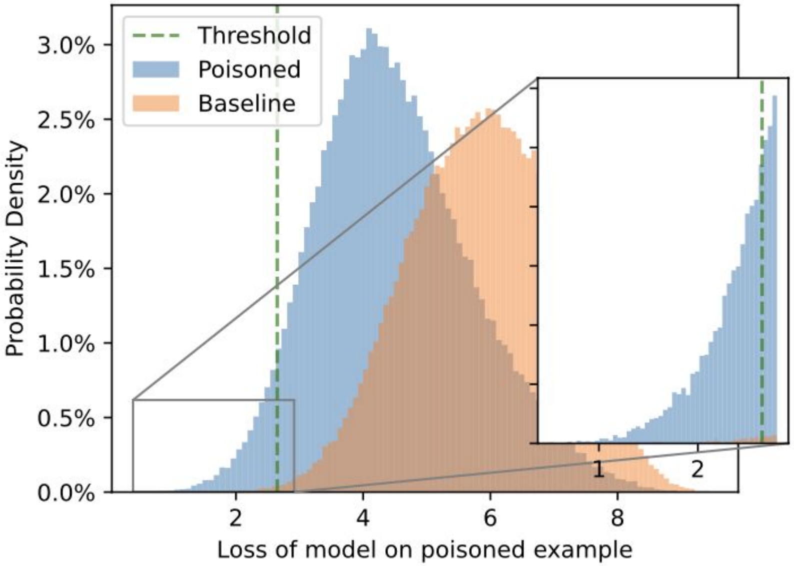
Carlini, Liu, Erlingsson, Kos, Song. **The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks.** USENIX Security 2019

Privacy guarantee is vacuous at this ε!

Empirical Privacy Auditing requires **many samples**

- Trained w/ $(0.21, 10^{-5})$-DP but empirically $\varepsilon > 2.79$ with confidence $1-10^{-8}$ ⇒ **bug in implementation**

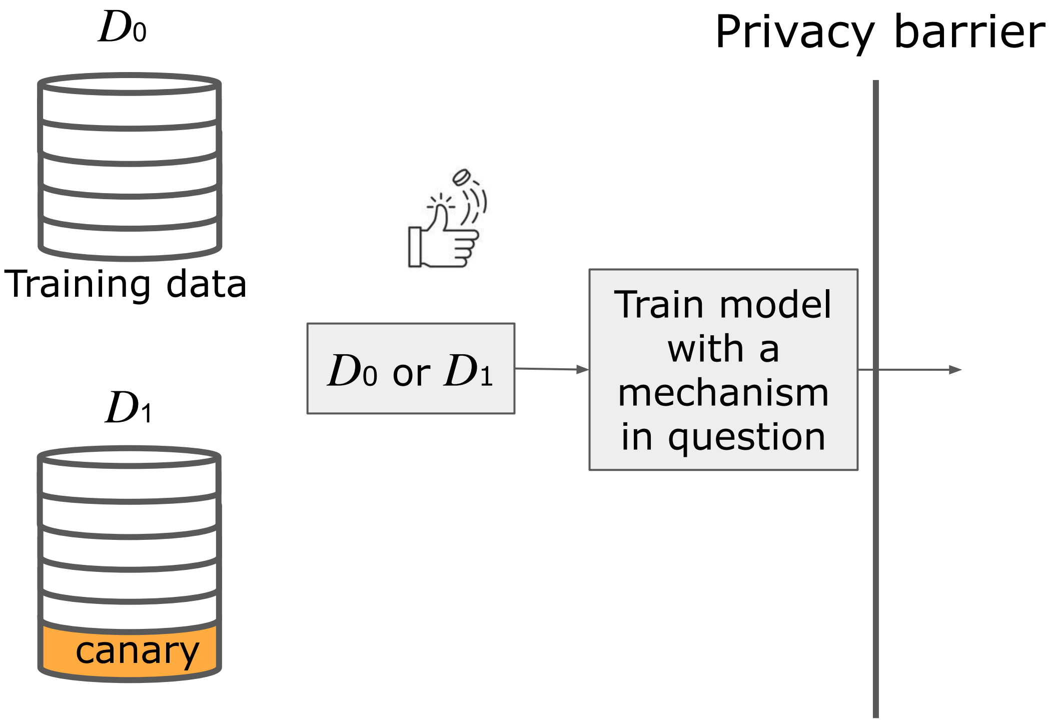
- This required training **n=200,000 models**



Tramèr et al. **Debugging Differential Privacy: A Case Study for Privacy Auditing**. Preprint 2022

Our goal: make empirical privacy auditing more ***sample-efficient***

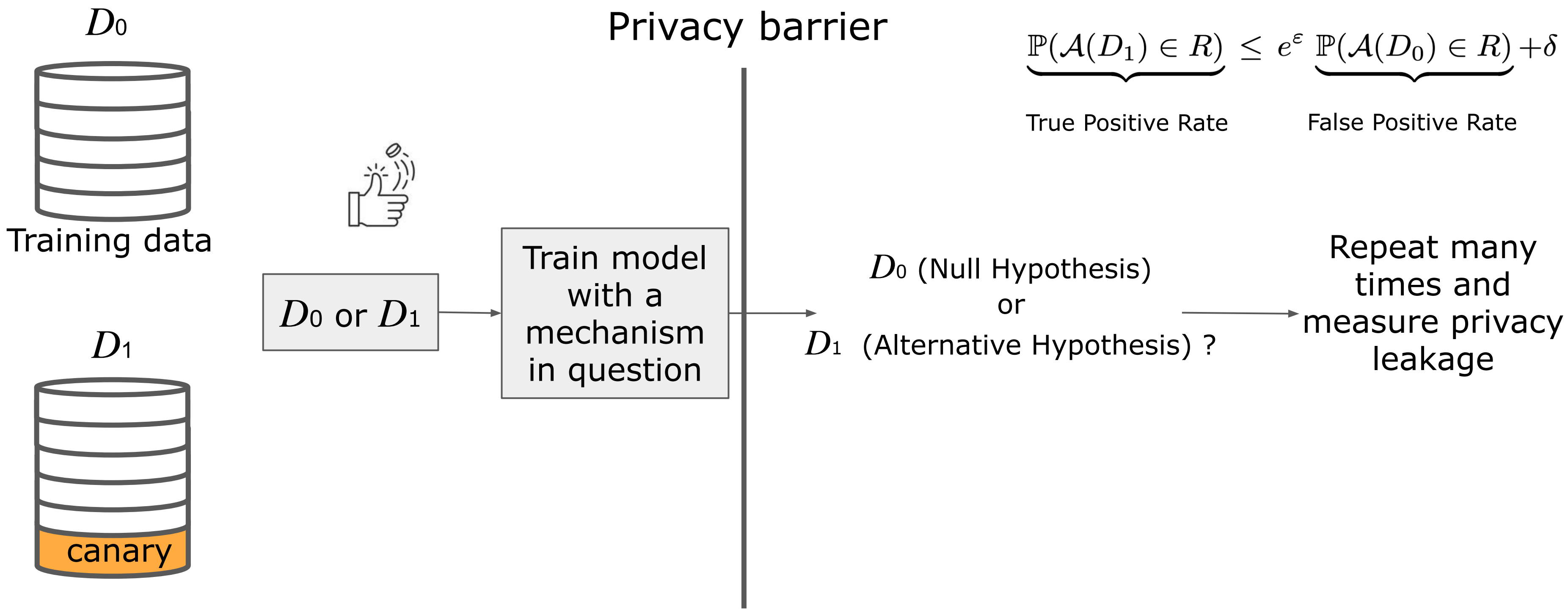
Standard approaches for auditing privacy: **binary hypothesis testing**



E.g., Nasr, Song, Thakurta, Papernot, Carlini. **Adversary Instantiation: Lower Bounds for Differentially Private Machine Learning**. IEEE S&P 2021
Jagielski, Ullman, Oprea. **Auditing differentially private machine learning: How private is private SGD?** NeurIPS 2020

Standard approaches for auditing privacy: **binary hypothesis testing**

$$\underbrace{\mathbb{P}(\mathcal{A}(D_1) \in R)}_{\text{True Positive Rate}} \leq e^{\varepsilon} \underbrace{\mathbb{P}(\mathcal{A}(D_0) \in R)}_{\text{False Positive Rate}} + \delta$$



E.g., Nasr, Song, Thakurta, Papernot, Carlini. **Adversary Instantiation: Lower Bounds for Differentially Private Machine Learning**. IEEE S&P 2021
Jagielski, Ullman, Oprea. **Auditing differentially private machine learning: How private is private SGD?** NeurIPS 2020

Bottleneck: Bernoulli confidence intervals

- Confidence intervals based on *n* trials

$$\text{TPR} \approx \underbrace{\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(\text{Guess } i \text{ correct})}_{\text{Empirical TPR/FPR}} + \sqrt{\frac{\text{Variance}}{n}}$$

Actual TPR/FPR

Empirical TPR/FPR

Actual TPR/FPR

$$\varepsilon \geq \log\left(\frac{\text{TPR} - \delta}{\text{FPR}}\right)$$

$$\geq \log\left(\frac{\hat{\text{TPR}}_n - \boxed{\frac{c}{\sqrt{n}}} - \delta}{\hat{\text{FPR}}_n + \boxed{\frac{c}{\sqrt{n}}}}\right)$$

Empirical TPR/FPR

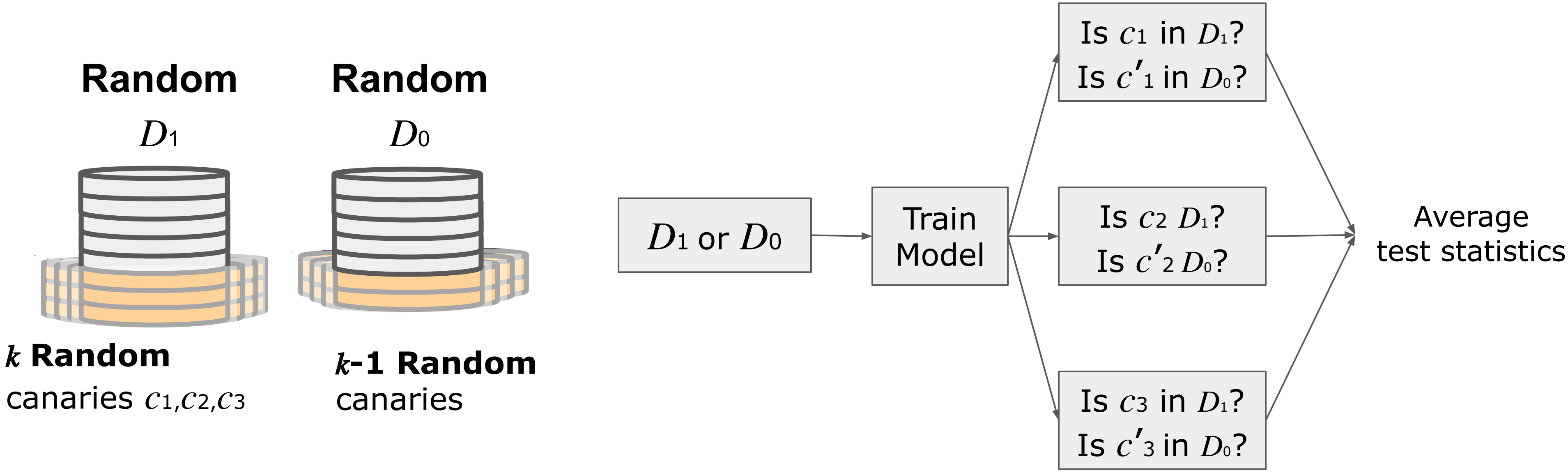
Sample size *n* needs to be large for good estimates

Our approach: leverage randomness

- **Lifted DP**: Equivalent notion of DP with randomized datasets
- Multiple randomized hypothesis tests
- **Adaptive confidence intervals** capitalizing on low correlations

Multiple hypothesis tests for auditing Lifted DP

- **Leave-One Out** construction with **i.i.d. random** canaries



Multiple hypothesis tests for auditing Lifted DP

If the statistics are independent ⇒ better confidence intervals

Unfortunately, they are **dependent**
(but highly uncorrelated)

Average
test statistics

canaries c_1, c_2, c_3

canaries

Is c_3 in D_1?
Is c'_3 in D_0?

Novel higher-order confidence interval

- 2nd-order confidence interval using empirical correlations between two tests

$$|\mathrm{TPR} - \widehat{\mathrm{TPR}}_{n,k}| \lesssim \sqrt{\frac{1}{n}\Big(\mathrm{Correlation} + \frac{1}{k} + \sqrt{\frac{\text{4th moment}}{n}}\Big)}$$

- Ideally, when **correlation=O(*1/k*)**, the confidence interval improves as

$$|\mathrm{TPR} - \widehat{\mathrm{TPR}}_{n,k}| \lesssim \sqrt{\frac{1}{nk} + \frac{1}{n^{3/4}}}$$

Takeaway: **Reduces variance** from randomness in trials

Standard approach:

$$\varepsilon \geq \log\left(\frac{\widehat{\mathrm{TPR}}_n - \boxed{\frac{c}{\sqrt{n}}} - \delta}{\widehat{\mathrm{FPR}}_n + \boxed{\frac{c}{\sqrt{n}}}}\right)$$

c - Universal constant

c′ - Data-dependent constant

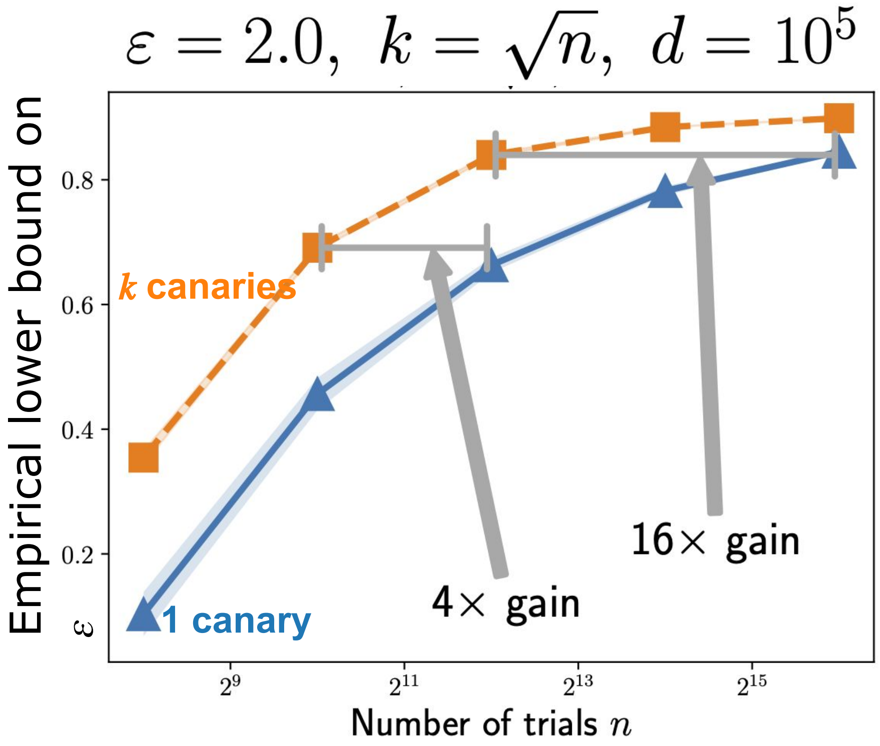
Lower variance =>
Tighter confidence intervals

Our approach:

$$\varepsilon \geq \log\left(\frac{\widehat{\mathrm{TPR}}_{n,k} - \boxed{\frac{c}{\sqrt{nk}} - \frac{c'}{n^{3/4}}} - \delta}{\widehat{\mathrm{FPR}}_{n,k} + \boxed{\frac{c}{\sqrt{nk}} + \frac{c'}{n^{3/4}}}}\right)$$

Proof of concept with Gaussian mechanisms

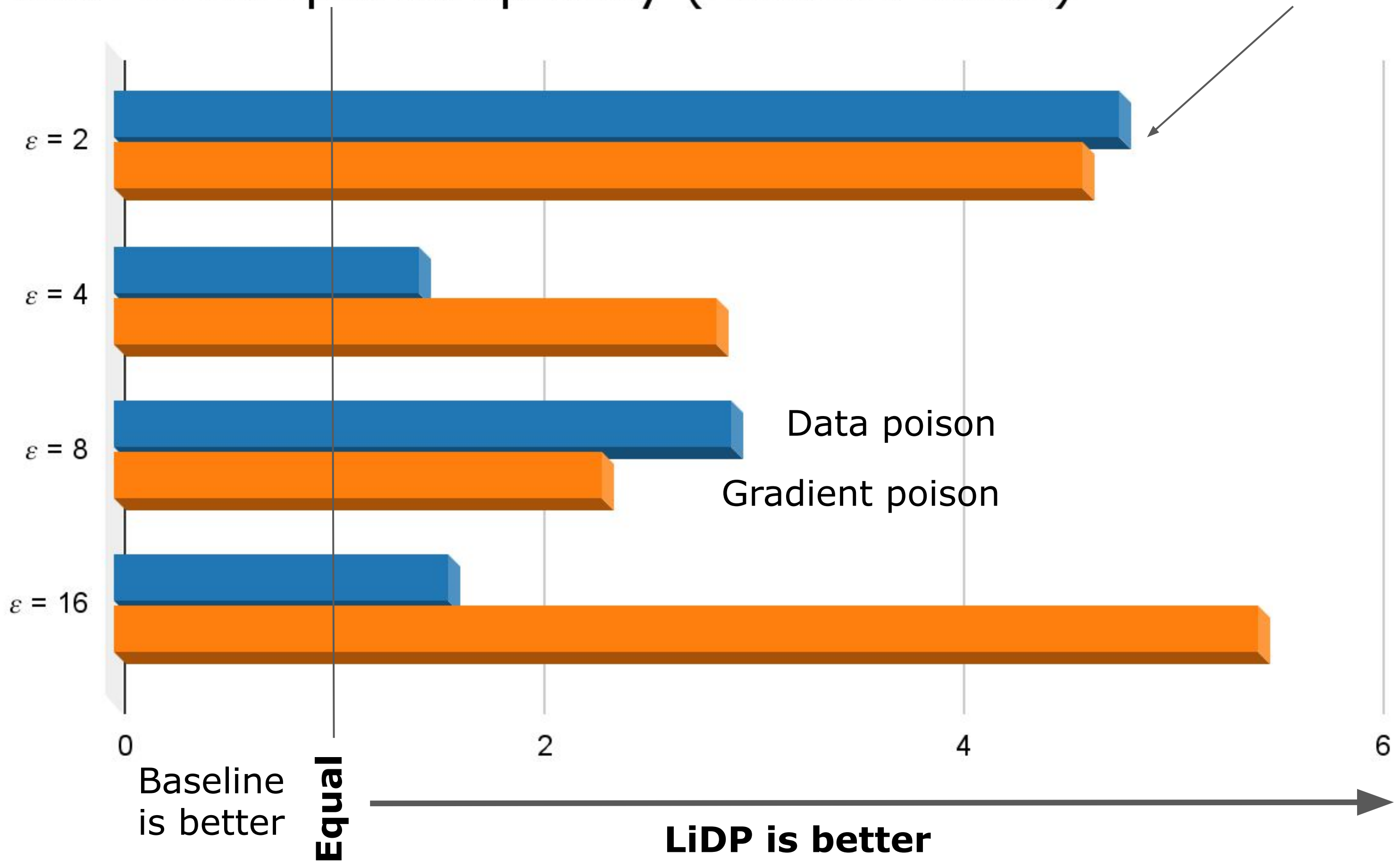
- Sum query with sensitivity 1
- Gaussian mechanism
- k **canaries** uniformly random on the sphere
- **Test statistic** is inner product



Dwork, Smith, Steinke, Ullman, Vadhan. **Robust traceability from trace amounts**. FOCS 2015

Suffices to train ***200 models*** *instead of 1000 models*

Gain in sample complexity (FashionMNIST)



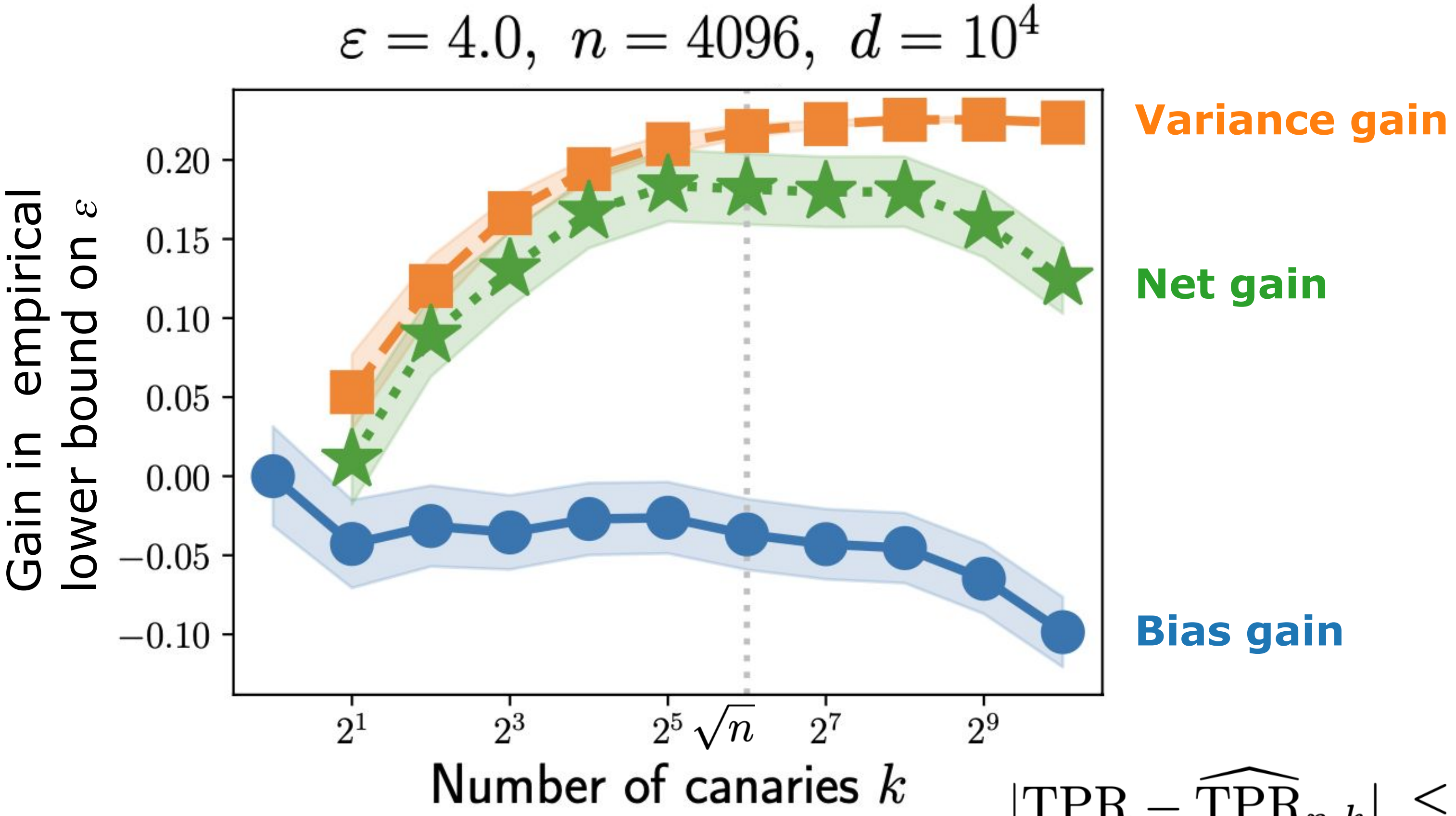
Privacy Auditing with One (1) Training Run

Thomas Steinke*
Google DeepMind
steinke@google.com

Milad Nasr*
Google DeepMind
srxzr@google.com

Matthew Jagielski*
Google DeepMind
jagielski@google.com

Bias-variance tradeoff in the number of canaries k



$$|\mathrm{TPR} - \widehat{\mathrm{TPR}}_{n,k}| \lesssim \sqrt{\frac{1}{nk} + \frac{1}{n^{3/4}}}$$

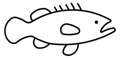
Summary

- **Auditing Lifted DP** (equivalent to usual DP) using multiple **i.i.d. random canaries** to improve sample dependence of the confidence intervals

- Can integrate with existing recipes for designing canaries

Other highlights: large-scale group-stratified datasets

Zach Charles*, Nicole Mitchell*, ***KP****,
Michael Reneer, Zach Garrett.
NeurIPS D&B 2023

Dataset Grouper 

Library for creating group-structured datasets.

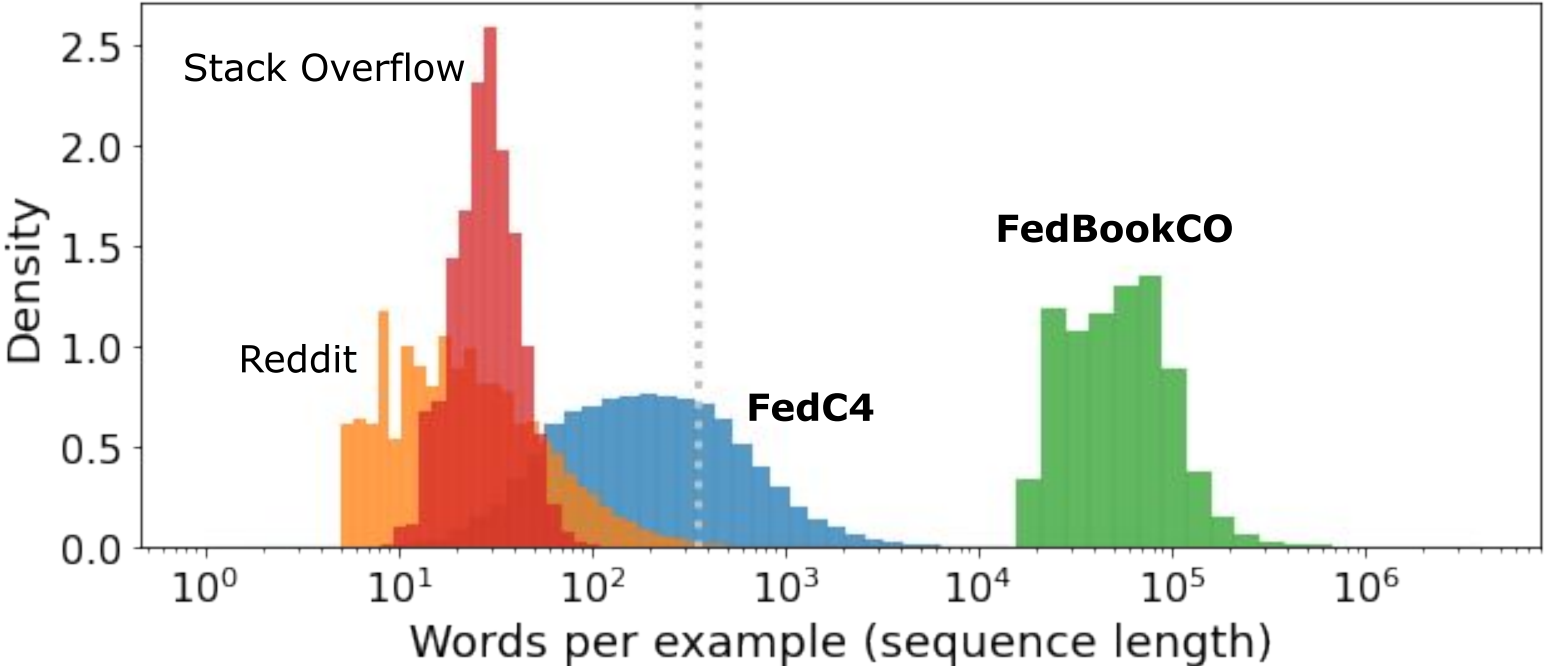
- **Scalable:** can handle millions of clients ✅
- **Flexible:** any custom partition function on any TFDS/HuggingFace dataset ✅
- **Platform-agnostic:** works with TF, PyTorch, JAX, NumPy, … ✅


New federated LLM datasets: longer sequences

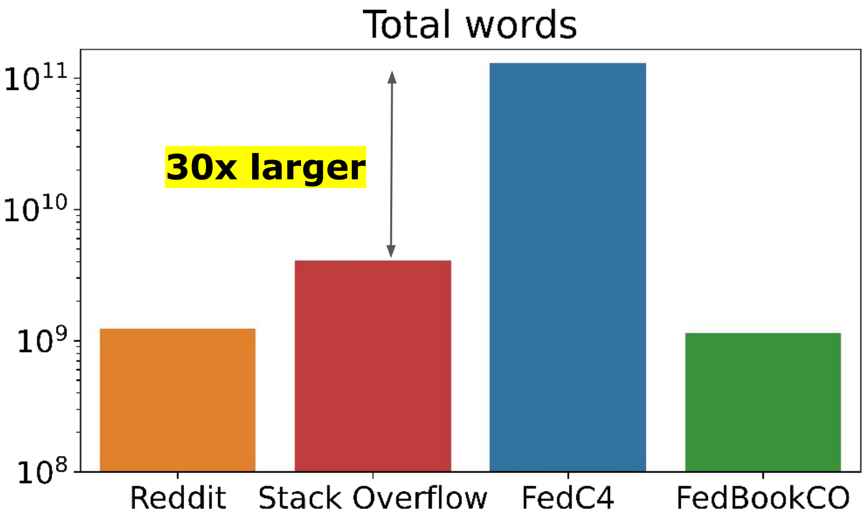
***Largest previous datasets**:*
Reddit, Stack Overflow

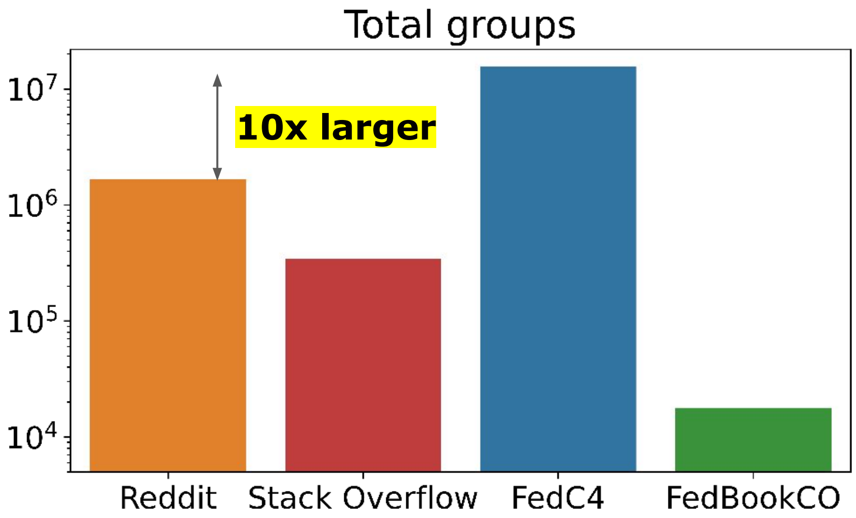
Typical sequence length of LLMs

***Our datasets**:*
FedC4, FedBookCO



New federated LLM datasets: more words & groups





Largest previous datasets ***Our datasets***

Largest previous datasets ***Our datasets***

Thank you!