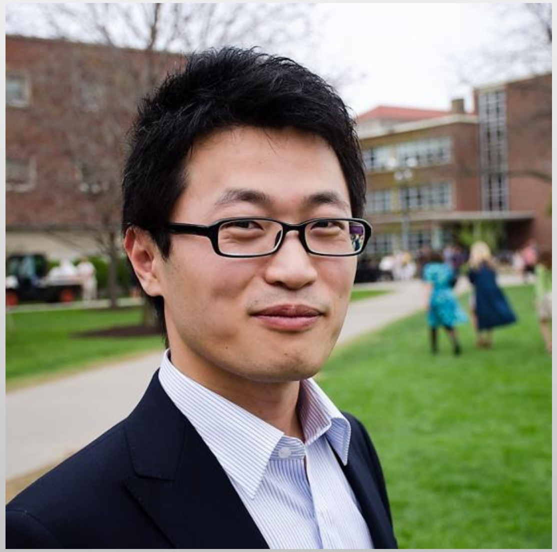
Privacy in learning: Basics and the Interplay

ICML tutorial

Presenters: Wei Chen, Huishuai Zhang

Microsoft Research Asia

About the presenters



Overview of the tutorial

0. Background on privacy

Overview of the tutorial

1. Privacy measures

2: Private machine learning

3. ML also borrows from DP

4. What is next?

What is privacy?

General definition of privacy

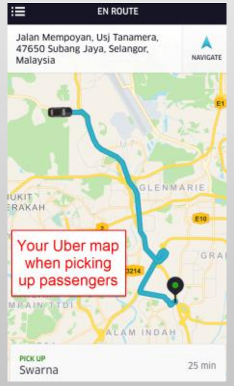
- Privacy is the claim of individuals (groups or institutions) to determine for themselves when, how, and to what extent information about them is communicated to others [Wiki]

Privacy in machine learning

- Data privacy attempts to use data while protecting an integrity of individual's privacy preferences and personally identifiable information.

Why is privacy issue more urgent in an AI era?

- Sensitive data are recorded anytime and anywhere

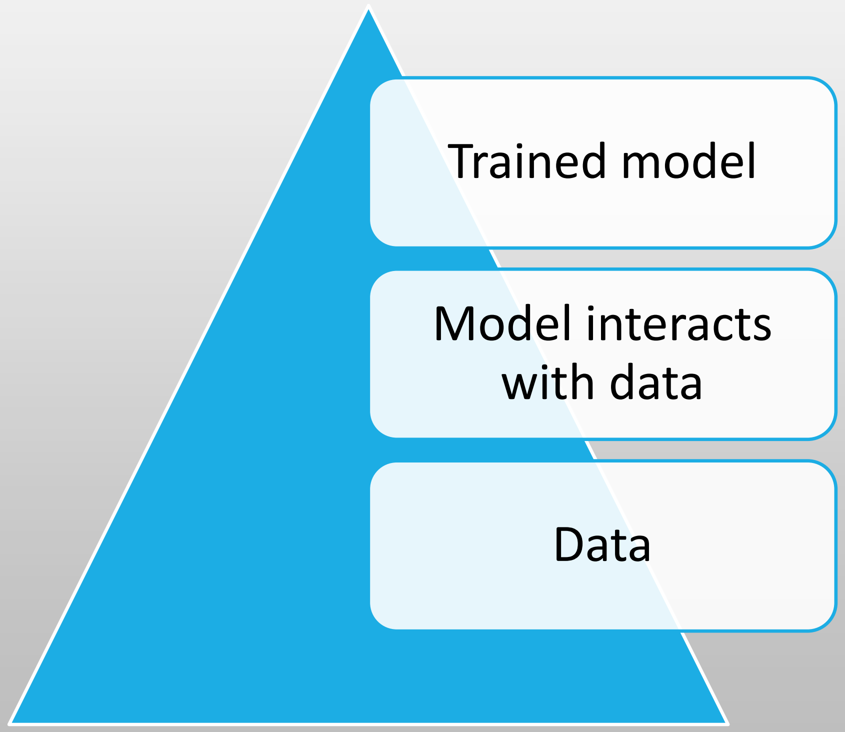


- Machine learning is a powerful tool to extract information.
- AI enables the adversary to exploit the data
- Simple anonymization is not safe

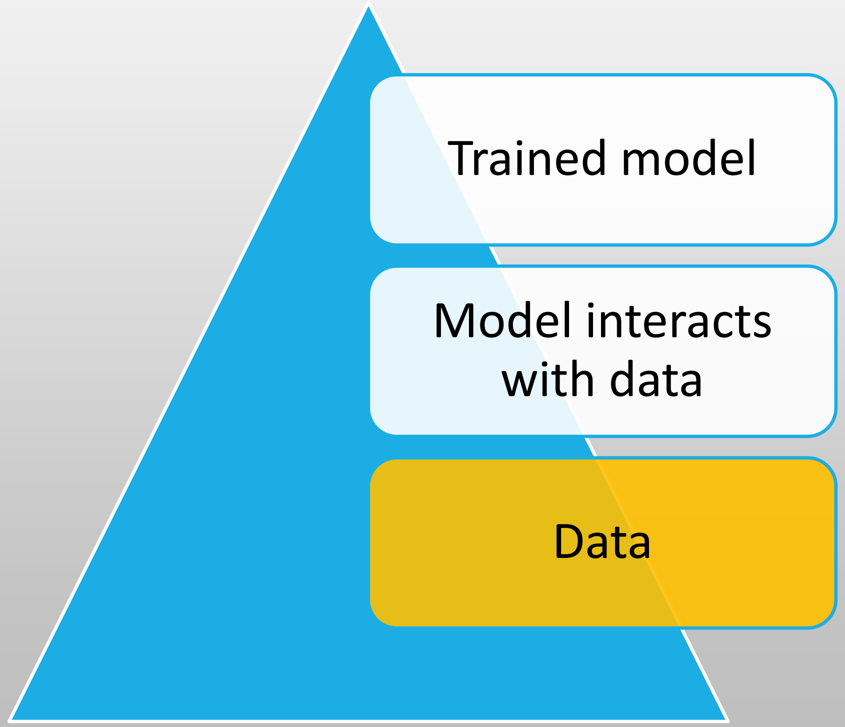
How to protect privacy in AI era?

Core principle: Control information flow from private to public.

How to protect privacy in AI era?

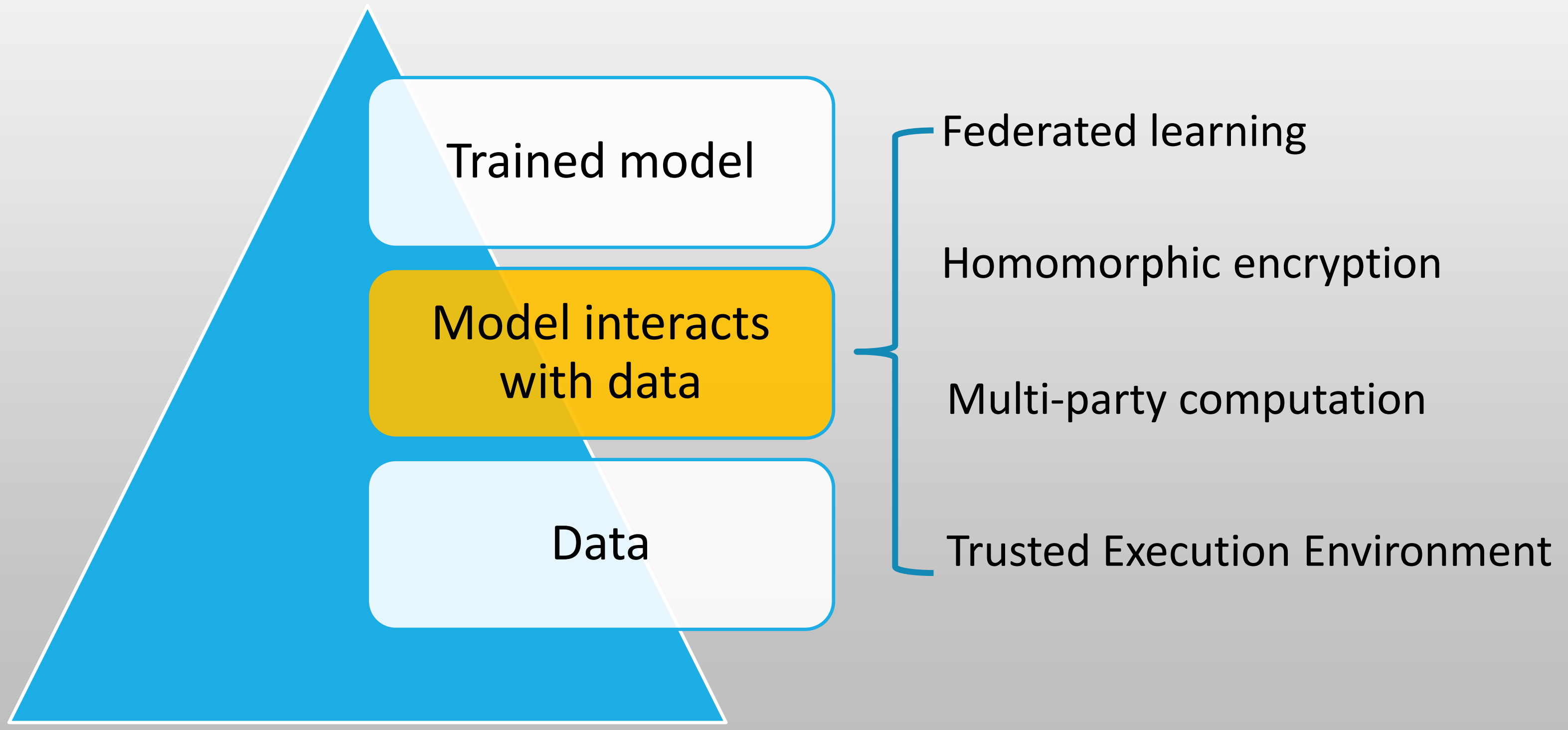


How to protect privacy in AI era?

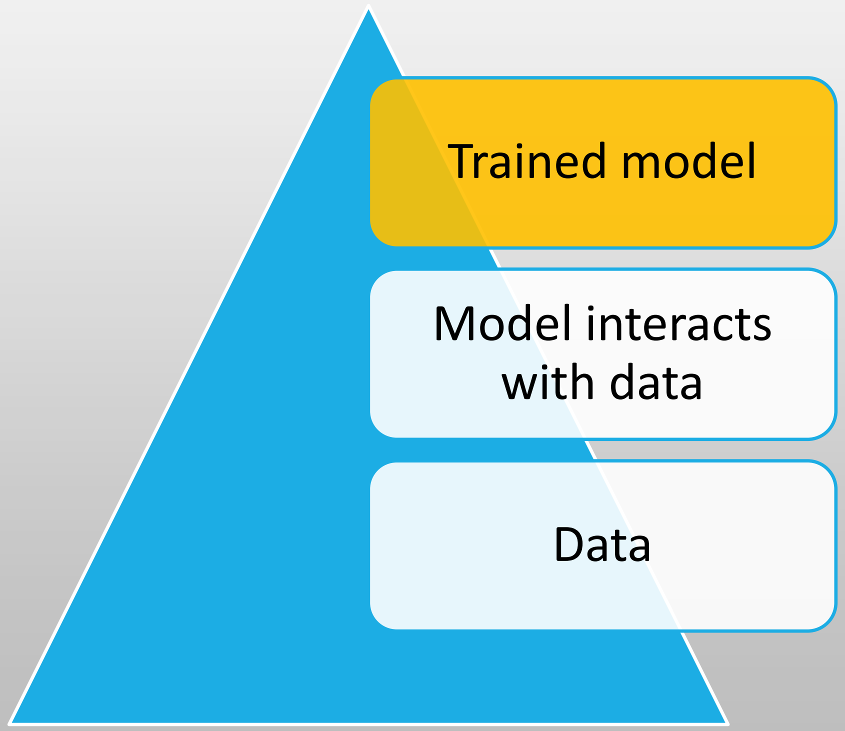


Data security

How to protect privacy in AI era?



How to protect privacy in AI era?



Will the trained model leak information of the data?

How to defend?

Differential privacy.

Federated Learning is to handle data islands

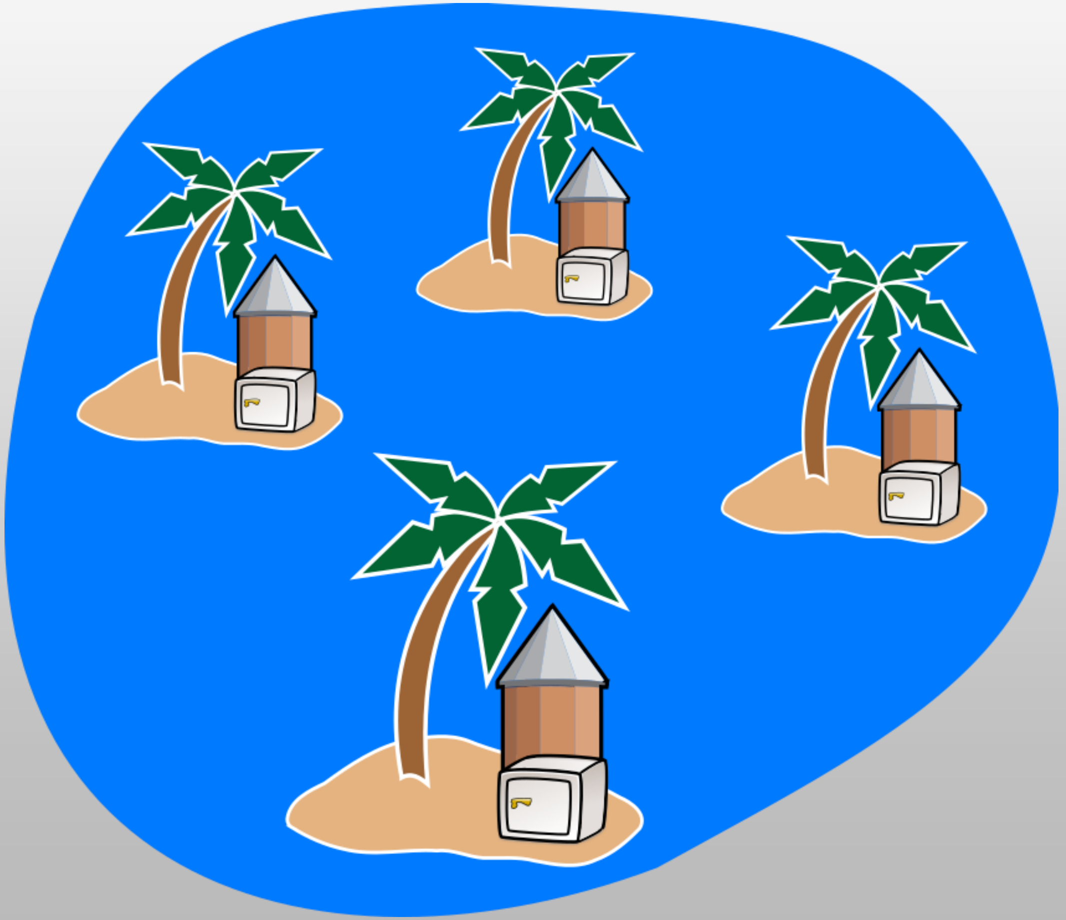


Figure is from Wiki

Federated Learning is to handle data islands

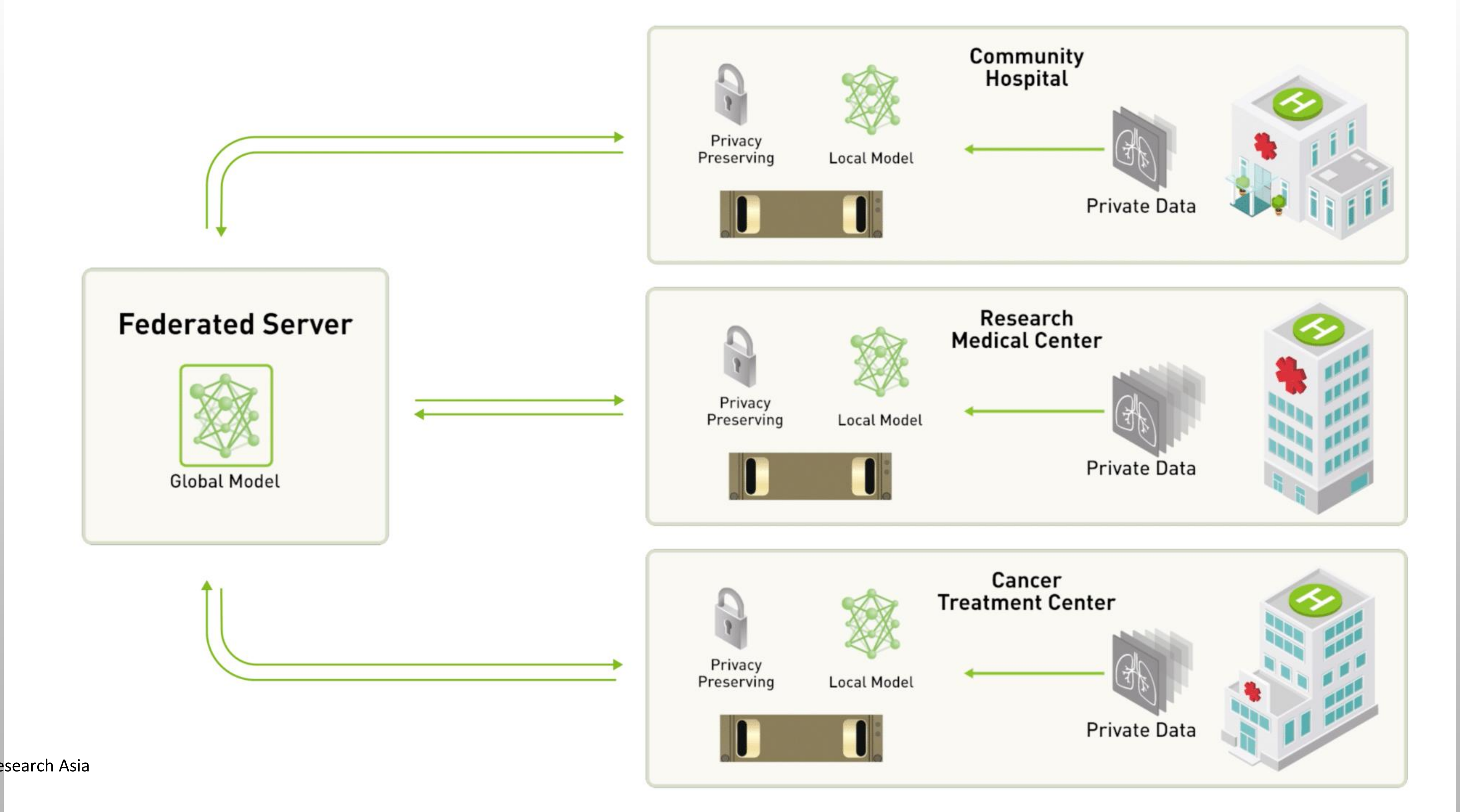


Figure is from Nvidia blog

Federated Learning is to handle data islands

Privacy promise

- Cut off the global model from directly accessing raw data.
- Add certain privacy barrier when doing local model aggregations

Potential Risk

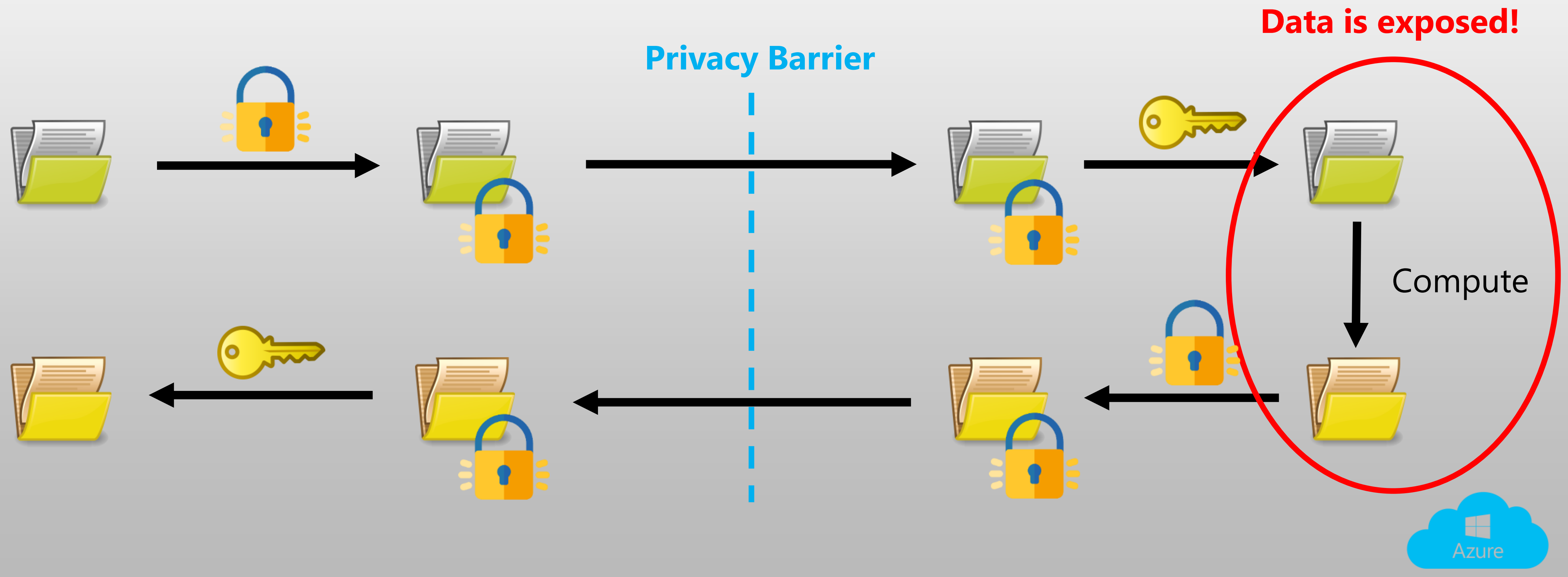
- Gradient matching attack to recover the raw data. [Zhu et al.2019, Zhao et al.2020]

Techniques

- Distributed machine learning: local SGD [Stich 2019, Woodworth et al.2020]
- Multiparty computing to securely aggregate.
- Differential privacy to hide the local model's contribution.

Confidential computing

- Confidential computing guarantees that the data is **confidentially** computed in the ML system.



Confidential computing: current solutions

Trusted execution environment (TEE): An enclave in computation provider

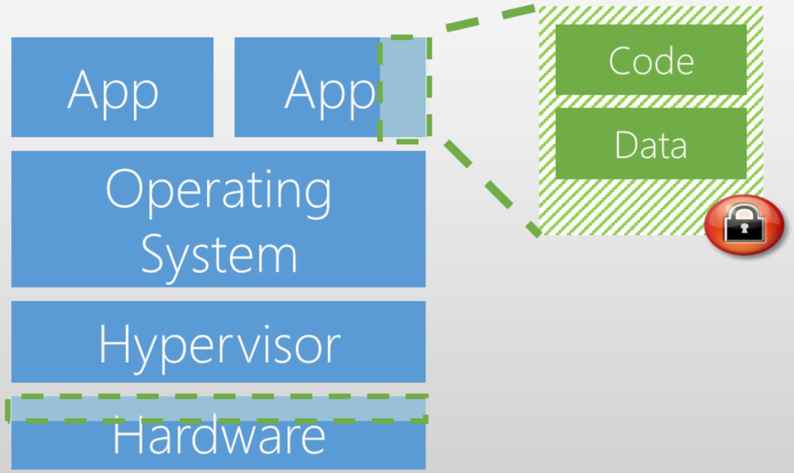
Homomorphic encryption:

$$En(x + y) = En(x) \oplus En(y)$$

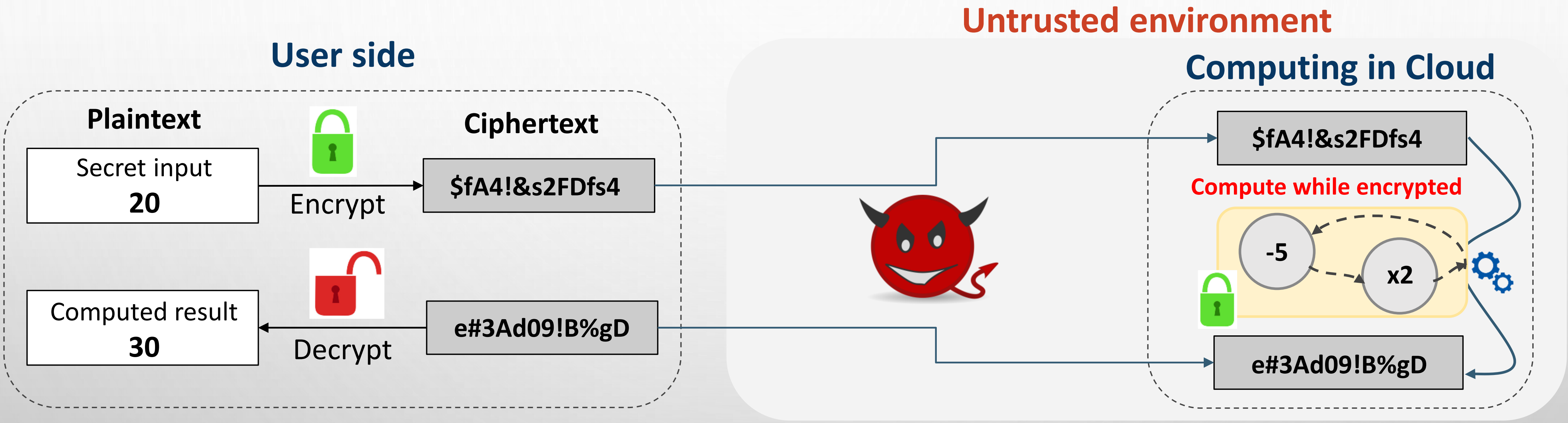
Multiparty secure computing

Trusted Execution Environment

- Trusted Execution Environment (TEE) [Ohrimenko et al. 2016, Hunt et al. 2018]
  - Software-based TEE: Virtual Secure Mode(VSM) in Windows
  - Hardware-based TEE: Intel SGX
- It is an enclave in the computation provider, and only the authorized individual can access it.



Homomorphic Encryption [Dowlin et al.2016]



The good news:

- Very strong security guarantees

The not-so-good news:

- Significant performance loss (~100-100,000x)
- Only some computations supported

Multi-Party Computing

- **Goal**: **Jointly compute a function over private inputs**
- Examples
  - Sum of multiple numbers; Millionaires' problem
- Threat model: honest but curious
- Huge communication cost

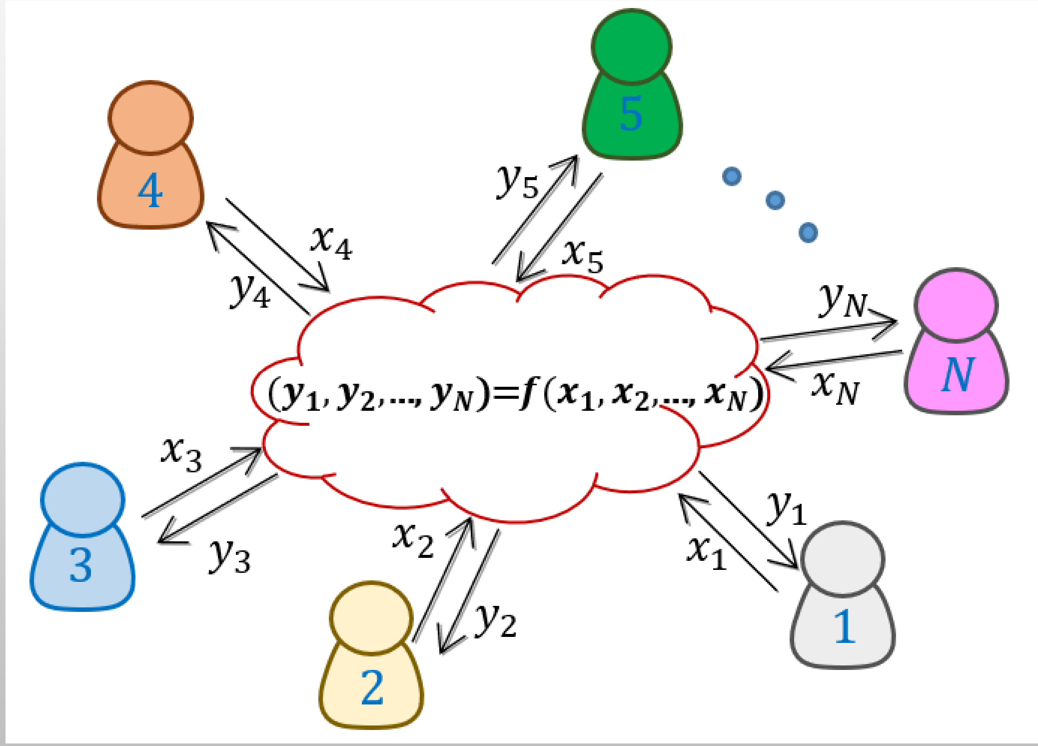


Figure is from [Lemus et al.2019]

Confidential computing: solution overview

Confidential Solutions	Provable encryption	Communication overhead	Computation overhead
Trusted executive environment	No	Yes	Yes
Fully homomorphic encryption	Yes	Yes	No
Multi-Party Computing	Yes	No	Yes

Will the trained model leak information of the data?

- Model inversion attack against a trained facial recognition model.
  - [Zhang et al.2020] “The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks.”

User images:

Attack results:

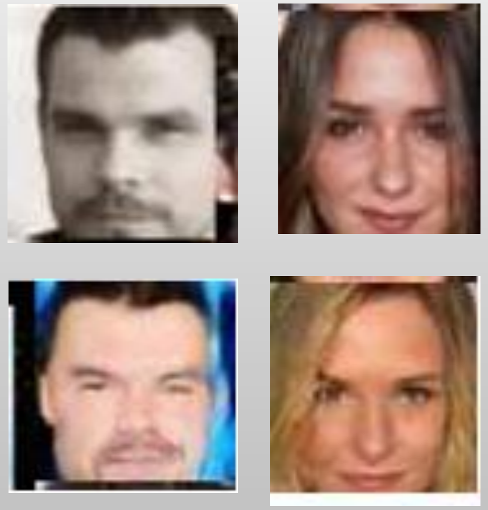


Figure is from [Zhang et al.2020]

Will the trained model leak information of the data?

- Privacy leakage of GPT2 [Carlini et al. 2020].
  - Reconstruct training samples from trained model.

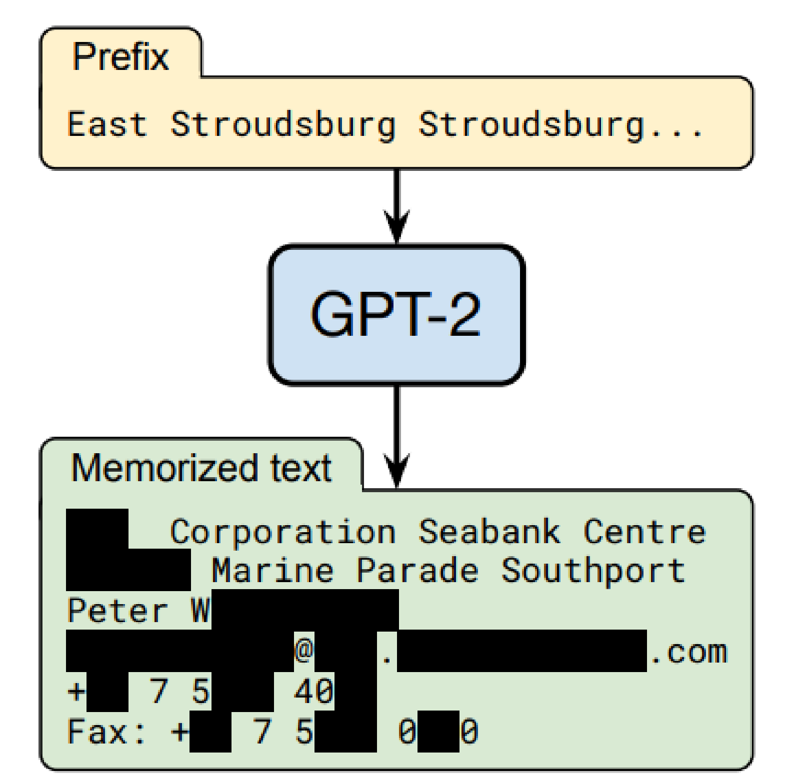


Figure is from [Carlini et al. 2020]

How to defend against model leakage?

Differential privacy

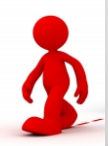
Attacker model: statistical inference

- From the output of a query, try to infer a problem:

  "Is a data point in the dataset?"

- Differential privacy is to defend statistical inference.

Let's say James comes to see a doctor….



Private Dataset	
Name	**Sick**
Scarlett	Yes
Ella	No
Jackson	Yes
James	Yes

Public Prevalence (before James)
66.6%

Public Prevalence (After James)
75%

James is **sick**.

Scope of the Tutorial

What we do cover

- Differential privacy measures and their properties
- Private machine learning
- Differential privacy for machine learning

What we do not cover

- Securing data using encryption
- Computation on encrypted data
- Multi-party computation
- Access control, trusted executive environment
- Anonymization and de-anonymization

Overview of the talk

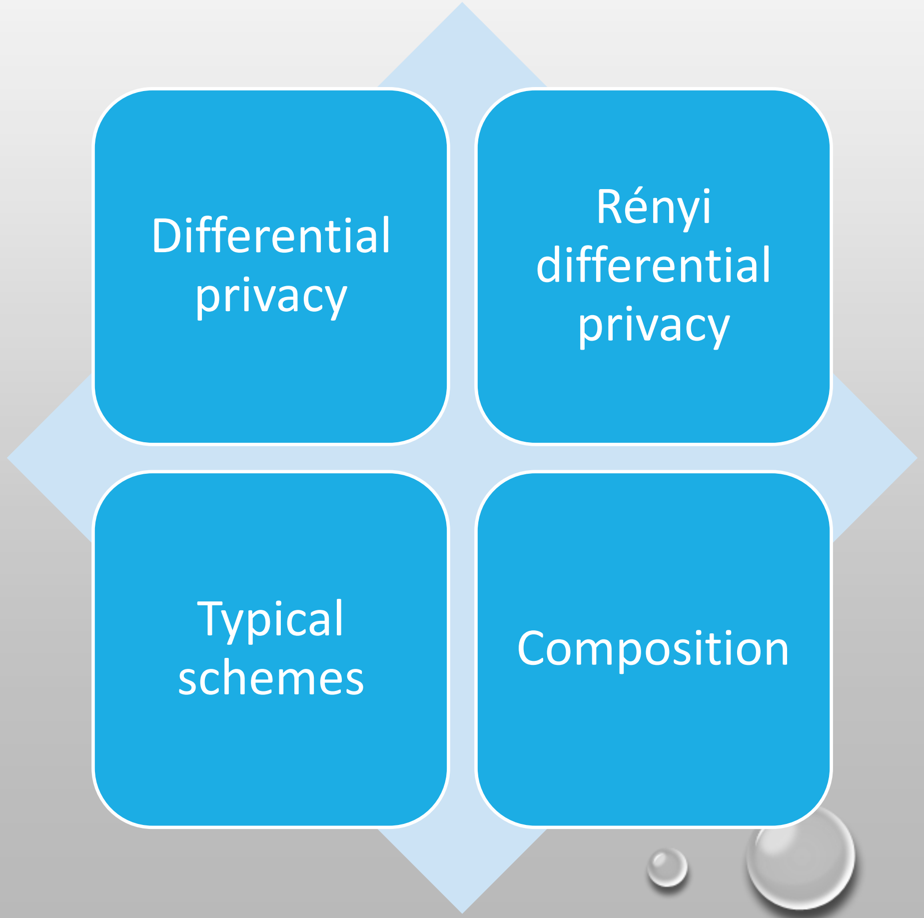
1. Privacy measures

2: Private machine learning

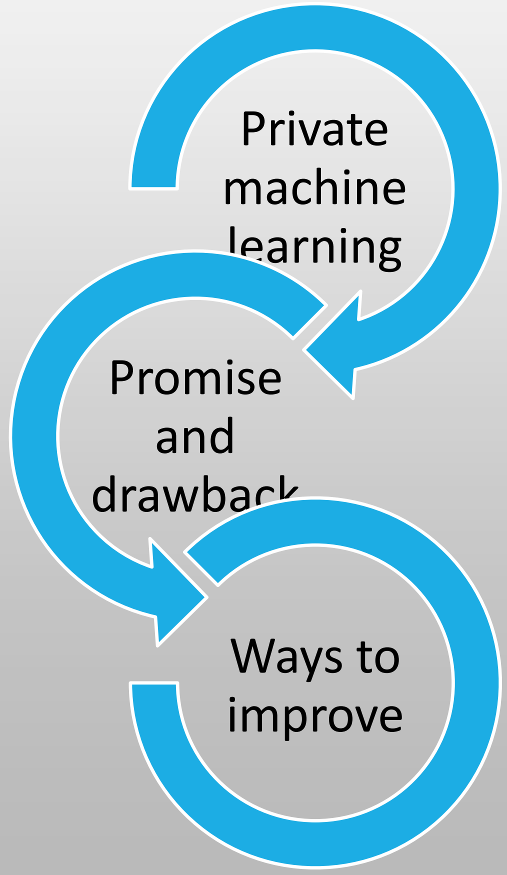
3. ML also borrows from DP

4. What is next?

1. Privacy measures



2. Private machine learning



3. ML also borrows from DP

Theoretically, helps to analyze the generalization, concentration

Empirically, used to defend against a wide range of attacks.

4. What is next?

- ML-friendly privacy measures
- Privacy in language model / generative model
- Privacy guarantee for federated learning

1. Privacy measures

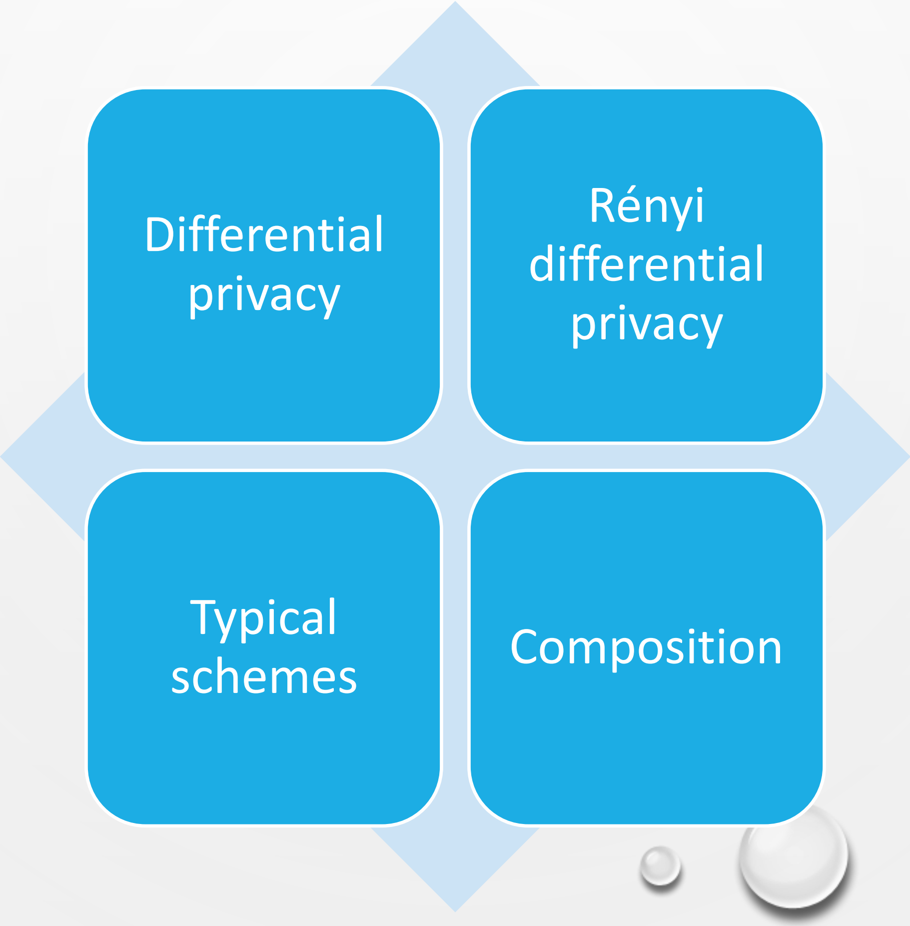
2: Private machine learning

3. ML also borrows from DP

4. What is next?

1. Privacy measures

Privacy measures and their properties



Differential privacy

Definition: $\mathcal{A}$ is ε-differentially private if two datasets $D, D' \in \mathcal{D}^n$ that differ in one individual then

$$\frac{\Pr[\mathcal{A}(D) \in S]}{\Pr[\mathcal{A}(D') \in S]} \leq e^{\varepsilon}, \quad \forall S \subseteq \mathrm{Range}(\mathcal{A}).$$

Intuitive meaning: A single data point will not change the output much.

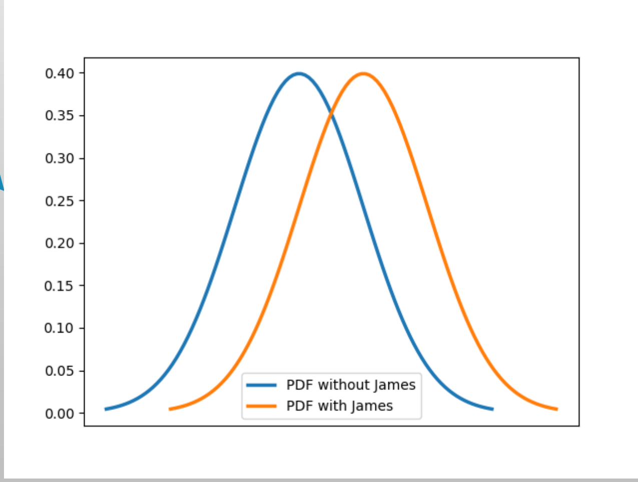
How to achieve DP? Through randomness.

$\mathcal{D}$	
Name	**Annual income**
Alice	47000
Bob	95000
Ella	90000
Scarlett	65000
…..	

$\mathcal{A}(\mathcal{D})$

$\mathcal{D}'$	
Name	**Annual income**
Alice	47000
Bob	95000
Ella	90000
Scarlett	65000
…..	
James	52000

$\mathcal{A}(\mathcal{D}')$



Differential privacy

Definition: $\mathcal{A}$ is ε-differentially private if two datasets $D, D' \in \mathcal{D}^n$ that differ in one individual then $\frac{\Pr[\mathcal{A}(D)\in S]}{\Pr[\mathcal{A}(D')\in S]} \leq e^{\varepsilon}, \forall S \subseteq \text{Range}(\mathcal{A})$.

- ε captures how much privacy we obtain: the smaller ε, the better privacy
- **Arbitrary** two datasets, differing by an **arbitrary** individual, for an **arbitrary** observation S

Differential privacy: (ε, δ) relaxation

A relaxation of ϵ-Differential privacy: (ε, δ)-DP.
$\mathcal{A}$ is (ε, δ)-differentially private if two datasets $D_1, D_2 \in \mathcal{D}^n$ that differ in one individual then $\Pr[\mathcal{A}(D_1) \in S] \leq e^{\varepsilon} \Pr[\mathcal{A}(D_2) \in S] + \delta, \forall S \subseteq \mathrm{Range}(\mathcal{A})$.

(ε, δ)-differential privacy interpretation: by excluding an event with δ probability, it satisfies ε-differential privacy.

Differential privacy: Typical schemes

- Goal: output $f(D)$ with DP
- Randomized algorithm $\mathcal{A}(D) \coloneqq f(D) + \boldsymbol{z}$, where $\boldsymbol{z}$ is random noise.

Laplace mechanism (ε-DP)

$$\mathrm{p}(z_i) = \frac{\varepsilon}{2S_1} \exp\left(-\frac{\varepsilon}{S_1} |z_i|\right)$$

Gaussian mechanism (ε, δ)-DP

$$z_i \sim \mathcal{N}\left(0, \left(\frac{S_2}{\varepsilon} \sqrt{C \log 1/\delta}\right)^2\right)$$

- $\boldsymbol{z}$ depends on **sensitivity** $S_p := \max_{\mathrm{D} \sim \mathrm{D}'} \|f(D) - f(D')\|_p$
- Each dimension's noise is i.i.d.
- Gaussian mechanism can not guarantee ε-DP for any finite ε.

Differential privacy: Proof of the privacy

Laplace mechanism (ε-DP): $p_{\text{Lap}}(z_i) = \frac{\varepsilon}{2S_1} \exp\left(-\frac{\varepsilon}{S_1}|z_i|\right)$. Denote $b = \frac{S_1}{\varepsilon}$.

$$
\begin{aligned}
\frac{\Pr[\mathcal{A}(D) = y]}{\Pr[\mathcal{A}(D') = y]} &= \frac{\Pi_i p_{Lap}(y_i - f(D)_i)}{\Pi_i p_{Lap}(y_i - f(D')_i)} \\
&= \Pi_i \exp(b^{-1}(|y_i - f(D)_i| - |y_i - f(D')_i|)) \\
&\leq \exp\left(b^{-1} \sum_i |f(D)_i - f(D')_i| \right) \\
&= \exp\left(b^{-1} \|f(D) - f(D')\|_1\right) \\
&\leq \exp(\varepsilon)
\end{aligned}
$$

Differential privacy: Proof of the utility

Suppose $f(D)$ is computing the average of D: $f(D) = \frac{1}{n}\sum_{X^{(k)}\in D} X^{(k)},$ and $X^{(k)} \in \mathbb{R}^d$, then with high probability

$$\|f(D) - \mathcal{A}(D)\|_1 = \|\boldsymbol{z}\|_1 \leq O\left(\frac{dS_1}{\varepsilon n}\right)$$

The error scales proportionally with the **dimension** d and the **sensitivity** S_1.

Differential privacy: Typical schemes

- **Sensitivity** $S_p = \max_{D,D': D \sim D'} \|f(D) - f(D')\|_p$ (worst case measure)
    - Larger sensitivity → larger noise → bad utility
    - Sensitivity has been relaxed, data dependent sensitivity.
        - Local sensitivity $LS(D) = \max_{D': D' \sim D} \|f(D) - f(D')\|$, and the smoothed version [Nissim et al.2007, Sun et al.2020].
- **Dimension.**
    - The error could be extremely large for high dimensional output.
    - Can we get rid of this dependence? Yes, for some structure assumption, i.e., sparsity.

Differential privacy: Typical schemes

- Exponential mechanism [McSherry&Talwar 2007]
  - A score function maps the (data, output) pairs to a score: $u(D, r)$
  - Define $S = \max_{r} \max_{D \sim D'} |u(D, r) - u(D', r)|$
  - Mechanism: Output r with probability proportional to $\exp\left(\frac{\varepsilon}{2S} u(D, r)\right)$ to preserve $(\varepsilon, 0)$-differential privacy
  - Laplace and Gaussian mechanisms are cases of Exponential Mechanism

$$p_{Lap}(r) \propto \exp\left(-\frac{\varepsilon \|r - f(D)\|_1}{S_1}\right), \qquad p_{Gau}(r) \propto \exp\left(-\frac{\varepsilon^2 \|r - f(D)\|_2^2}{C S_2^2 \log 1/\delta}\right)$$

Differential privacy: properties (Post-processing)

- Post-processing:
  - Privacy risk doesn't increase if further processing the DP outputs.

Differential privacy: properties (Composition)

- Composition of mechanisms. Consider the example of gradient descent.

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{1}{m} \sum_{i=1}^{m} \nabla_\theta \ell(x_i; \theta_t),$$

we may ensure privacy of each step t by adding noise ζ_t.

- What about the final privacy level after T iterations?

Theorem [Basic composition, Dwork&Lei 2009]: Let $\mathcal{A}_{1:k}$ be k mechanisms with independent noises such that $\mathcal{A}_i$ is $(\varepsilon_i, \delta_i)$-DP. Then the adaptive composition of $\mathcal{A}_{1:k}$ is $(\sum_i \varepsilon_i, \sum_i \delta_i)$-DP.

Proof. The proof idea is to examine the definition and use induction.

Differential privacy: properties (Composition)

- Basic composition theorem does not exploit the independence of the added noise, loose bound.
- **Advanced composition theorem** [Kairouz et al. 2015]:

Theorem: Let $\mathcal{A}_{1:k}$ be k mechanisms with independent noises such that $\mathcal{A}_i$ is (ε, δ)-DP.
Then the adaptive composition of $\mathcal{A}_{1:k}$ is $\left(O\left(\sqrt{k}\varepsilon\right), O(k\delta)\right)$-DP for small ε.

Proof. See next page.

Differential privacy: some math for composition

- **Privacy loss** random variable

$$L(p \parallel q) \coloneqq \log \frac{p(\xi)}{q(\xi)},$$

where p and q are two probability densities and $\xi \sim p(\cdot)$. DP is about the tail bound of $L(p \parallel q)$.

- **Claim**: If $\Pr(L(\mathcal{A}(D) \parallel \mathcal{A}(D')) > \varepsilon) < \delta,$ then $\mathcal{A}$ is (ε, δ)-DP.
- **Fact**: $\mathbb{E}L(p \parallel q) = \mathrm{KL}(p \parallel q)$.
- **Fact**: For Gaussian mechanism, $L(\mathcal{A}(D) \parallel \mathcal{A}(D'))$ is a Gaussian variable, $\mathcal{N}\left(\frac{\|\Delta\|_2^2}{2\sigma^2}, \frac{\|\Delta\|_2^2}{\sigma^2}\right)$, where $\Delta \coloneqq f(D) - f(D')$.

Proof of advanced composition: View the overall privacy loss as the sum of independent/conditional independent variables, and use concentration bound (Azuma's Inequality)

Differential privacy: properties (Composition)

- The composition bound can be further improved for specific mechanisms.
    - Gaussian mechanisms: moment account [Abadi et al. 2016]
    - Laplace mechanisms: f-differential privacy [Dong et al. 2019]
    - Exponential mechanisms: 40% saving of privacy budget [Dong et al. 2020]

Variants: Rényi differential privacy

- Problem with (ε, δ)-differential privacy
  - Gaussian mechanism satisfies an infinite many pairs (ε, δ), which are not comparable.
  - (ε, δ)-DP has two parameters, hard to choose best pair (ε, δ) when using composition
- Rényi differential privacy [Mironov2017]

Definition (Rényi divergence). For two probability distributions P and Q, the Rényi divergence of order $\alpha > 1$ is

$$D_\alpha(P \parallel Q) := \frac{1}{\alpha - 1} \log \mathbb{E}_{x \sim Q} \left(\frac{P(x)}{Q(x)} \right)^\alpha .$$

- Notable relation: $\lim_{\alpha \to 1} D_\alpha(P \parallel Q) = KL(P \parallel Q)$ $\quad D_\infty(P \parallel Q) = \sup_{x \in supp(Q)} \log \frac{P(x)}{Q(x)}$
- For $\mathcal{N}(\mu_1, \sigma^2 \boldsymbol{I})$ and $\mathcal{N}(\mu_2, \sigma^2 \boldsymbol{I})$, Rényi divergence $D_\alpha(\mathcal{N}_1 \parallel \mathcal{N}_2) = \frac{\alpha \|\mu_1 - \mu_2\|_2^2}{2\sigma^2}$

Variants: Rényi differential privacy

- (α, γ)- Rényi differential privacy

Definition. A randomized mechanism $\mathcal{A}: \mathcal{D} \to \mathcal{R}$ is said to have (α, γ)- Rényi differential privacy (RDP), if for any adjacent D, D' it holds that

$$D_\alpha\big(\mathcal{A}(D) \parallel \mathcal{A}(D')\big) \leq \gamma.$$

- **Example**: The Gaussian mechanism satisfies a continuum pairs $\big(\alpha, \gamma(\alpha)\big)$ for any $\alpha > 1$ as

$$D_\alpha(\mathcal{N}_1 \parallel \mathcal{N}_2) = \frac{\alpha \|\mu_1 - \mu_2\|_2^2}{2\sigma^2}.$$

Variants: Rényi differential privacy

- (α, γ)- RDP enjoys simple composition property.

Theorem [Mironov2017]. Let $\mathcal{A}_1: \mathcal{D} \rightarrow \mathcal{R}_1$ be (α, γ_1)-RDP and $\mathcal{A}_2: \mathcal{R}_1 \times \mathcal{D} \rightarrow \mathcal{R}_2$ be (α, γ_2)-RDP, then the mechanism $(\mathcal{A}_1, \mathcal{A}_2)$ satisfies $(\alpha, \gamma_1 + \gamma_2)$-RDP.
Proof. From the definition of Rényi divergence.

- **Example** (Gaussian mechanism). Suppose $S = 1$. We compute the adaptive composition of k Gaussian mechanisms on the same query. Each $\mathcal{A}_i$ is (α, γ)-RDP, then their composition $\{\mathcal{A}_i\}_{i=1}^{k}$ satisfies $(\alpha, k\gamma)$-RDP.

Variants: Rényi differential privacy

- Translation from (α, γ)-RDP to (ε, δ)-DP

Theorem [Mironov2017]. If $\mathcal{A}$ is (α, γ)-RDP, it also satisfies $\left(\gamma + \frac{\log 1/\delta}{\alpha - 1}, \delta\right)$-DP for any $0 < \delta < 1$.

- *Proof. Based on an application of Hölder's inequality.* $P(E) \leq (\exp[D_\alpha(P \parallel Q)] \cdot Q(E))^{\frac{\alpha-1}{\alpha}}$.
- We can compute a best pair (ε, δ) from a continuum $\left(\alpha, \gamma(\alpha)\right)$-RDP.

Variants: Rényi differential privacy

- Proof of composition of k (α, γ)-RDP mechanisms from moment accountant [Abadi et al.2016].
  - Recall the privacy loss $\log \frac{p^i(\xi_{1:i})}{q^i(\xi_{1:i})}$, the $(\alpha - 1)$ MGF is $M_i = \mathbb{E} \exp\left((\alpha - 1) \log \frac{p^i(\xi_{1:i})}{q^i(\xi_{1:i})} \right)$
  - Prove $M_i \leq \exp\big((\alpha - 1)\gamma\big) M_{i-1}$ via conditional expectation. Hence $M_k \leq \exp\big((\alpha - 1)k\gamma\big)$
  - Then by the definition of Rényi divergence, $D_\alpha\big(p^k \parallel q^k\big) = (\alpha - 1)^{-1} \log M_k \leq k\gamma$.
- Other similar formalized definitions are CDP, zCDP [Dwork&Rothblum2016, Bun&Steinke2016].
- Another recent measure is f-differential privacy [Dong et al.2019].

1. Privacy measures

2: Private machine learning

3. ML also borrows from DP

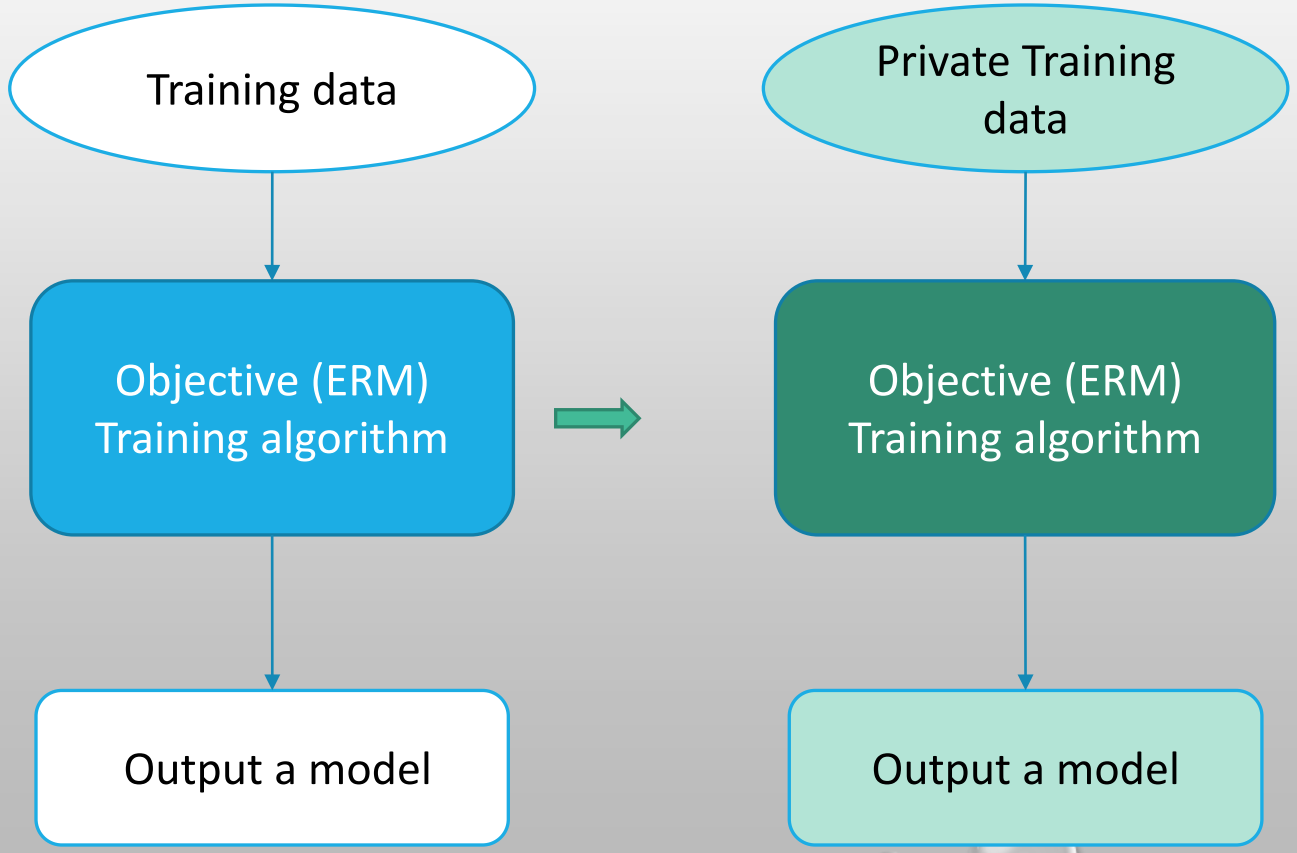
4. What is next?

2: Private machine learning

Private machine learning

- Machine learning with privacy guarantee
- The promise and the drawbacks
- Ways to improve

Machine learning with privacy guarantee



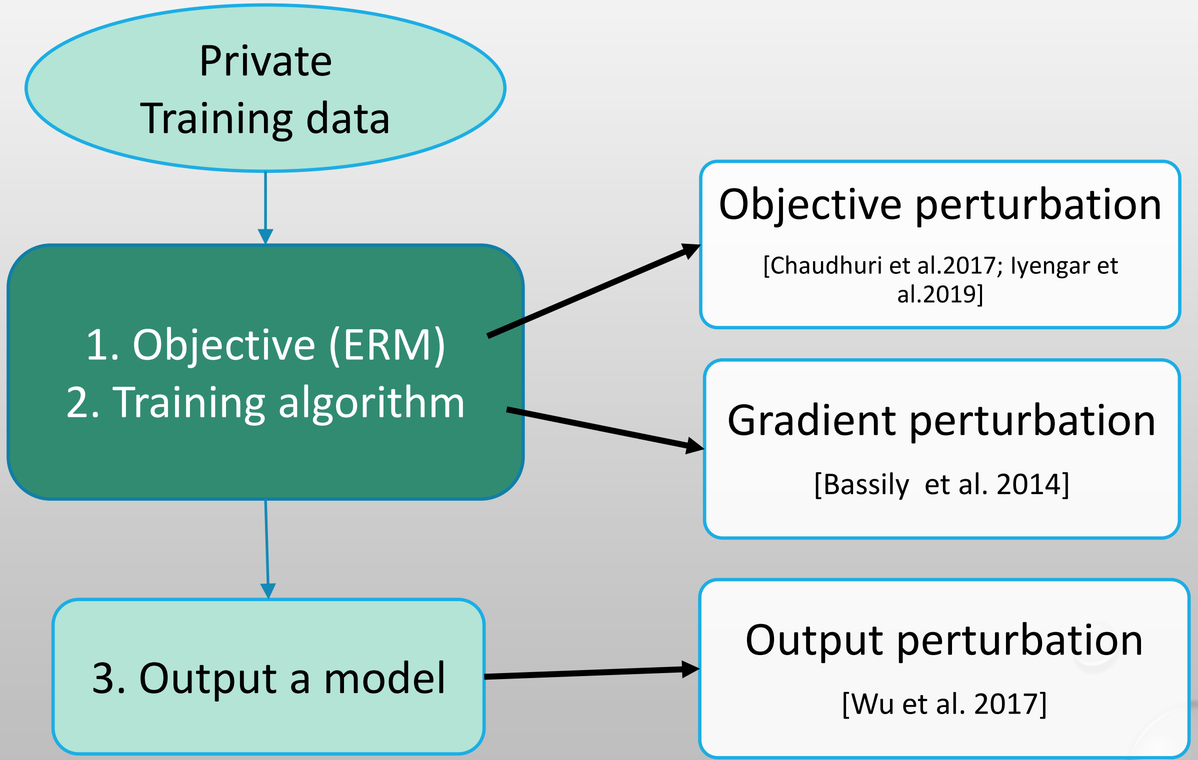
Will the final model leak private information of data?

Yes.

Differential privacy.

Machine learning with privacy guarantee

- Approach to achieve DP: Adding noise
  - When? [Yu et al.2020]
- How large is the noise?
  - Sensitivity: how much change does one sample could make to the final output?
  - For objective and output perturbation, $\sim\beta/\mu$.
  - Clipping gradient can be the sensitivity for gradient perturbation (suitable for DNN).



DP-SGD

Algorithm SGD

1. Random initialization θ_0
2. *For* $t = 1, 2, \dots, T$

   Sample a data point $i_t \sim \{1, 2, \dots, n\}$

   $g_t = \nabla \ell \left(\theta_{t-1}, (x_{i_t}, y_{i_t}) \right)$

   $\theta_t = \theta_{t-1} - \eta_t g_t$

Return $\hat{\theta} = \theta_T$

Algorithm DP-SGD

1. Random initialization θ_0
2. *For* $t = 1, 2, \dots, T$

   Sample a data point $i_t \sim \{1, 2, \dots, n\}$

   Generate noise $z_t \sim p_{(\varepsilon, \delta)}$

   $\hat{g}_t = \nabla \ell \left(\theta_{t-1}, (x_{i_t}, y_{i_t}) \right) + z_t$

   $\theta_t = \theta_{t-1} - \eta_t \hat{g}_t$

Return $\hat{\theta} = \theta_T$

How large is the noise in DP-SGD?

- The noise depends on the sensitivity of the gradient

$$\max_{D,D'} \max_{\theta} \|\nabla L(\theta; D) - \nabla L(\theta; D')\|$$

- Sensitivity depends on the smoothness of the loss.

- One can also clips the individual gradient to a predefined threshold [Chen et al. 2020].

The privacy proof of DP-SGD

- Privacy proof is straightforward based Rényi differential privacy given the sensitivity S.
  - Each call of Gaussian mechanism satisfies $\left(\alpha, \gamma(\alpha)\right)$-RDP, where $\gamma(\alpha) = \frac{S\alpha}{\sigma^2}$.
  - By the composition property of RDP, overall T iterations satisfies $(\alpha, T\gamma(\alpha)\ )$-RDP
  - Translate the $\left(\alpha, T\gamma(\alpha)\right)$-RDP to (ε, δ)-DP, optimizing the (ε, δ) over $\alpha \in (1, \infty)$.
  - For DP-SGD, we to need consider privacy amplification by subsampling [Mironov et al. 2019].

Lemma: Let $\mathcal{A}$ be (ε, δ)-DP algorithm. Let $Samp$ be a procedure that given a data set D of size n, randomly samples k entries (with replacement) from D. Then the algorithm $\mathcal{A}(Samp(\cdot))$ is $\left(O\left(\frac{k}{n}\varepsilon\right), \delta\right)$-DP.

The utility proof of DP-SGD

- The utility of DP-SGD or DP-GD can be analyzed via **noisy gradient descent**, where the noise depends on the (ε, δ) and the number of iterations.

- The excess error of DP-SGD is $O\left(\frac{\sqrt{p}}{n\varepsilon}\right)$ [Bassily et al. 2014]. Utility deteriorates as the model dimension gets larger.

- Empirically, this has also been verified [Tramer&Boneh 2021].

The empirical performance of DP-SGD

- Some empirical results of DP-SGD [Abadi et al. 2016, Code in PyTorch]
  - Code implementation [Opacus, BackPACK], reduce the cost of computing individual gradients

Dataset	Model	Non-private	$\varepsilon = 2$	$\varepsilon = 5$	$\varepsilon = 8$
MNIST	CNN-2layer	99.1%	94.7%	96.8%	97.2%
SVHN	ResNet20	95.9%	87.1%	91.3%	91.6%
CIFAR10	ResNet20	90.4%	43.6%	52.2%	56.4%

- Wait, $\varepsilon = 8$! Quite nonsense as $e^8 \approx 2981$. How private is DP-SGD?

The promise and the drawbacks of DP-SGD

The promise of DP-SGD

- How private is DP-SGD [Jagielski et al. 2020, Nasr et al. 2021]? How to empirically measure this?
  - By definition, differential privacy provides a provable defense for data poisoning attacks.
  - Design strong data poisoning attacks to measure a lower bound on the privacy offered by differentially private algorithms.

The promise of DP-SGD

- The attack process [Nasr et al. 2021]

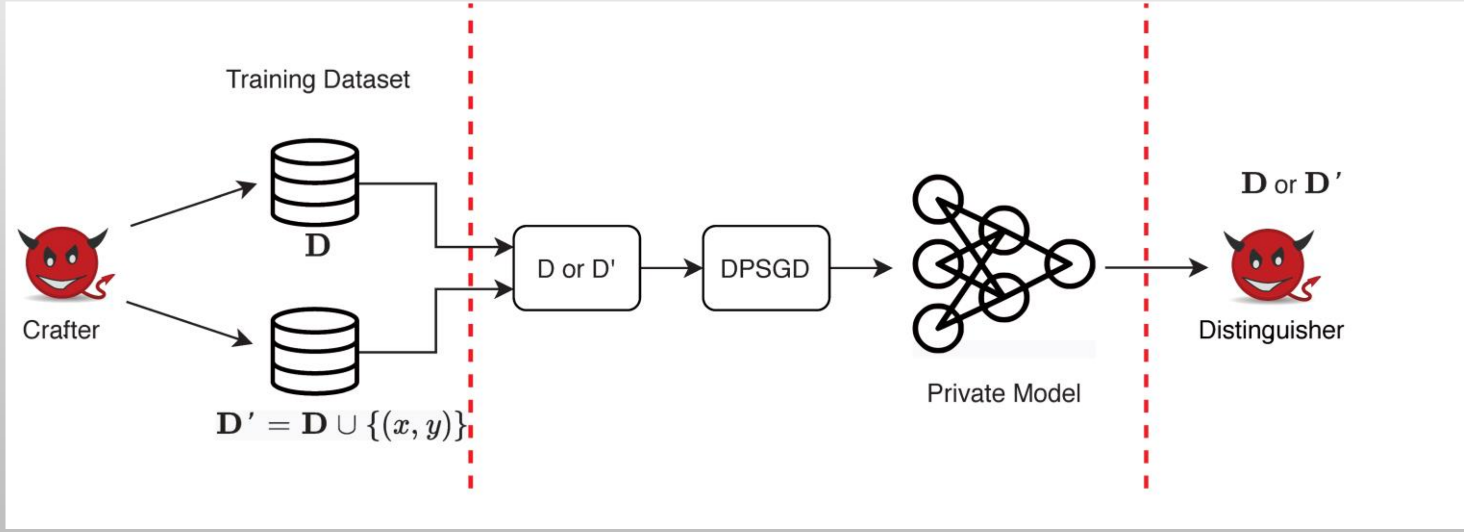
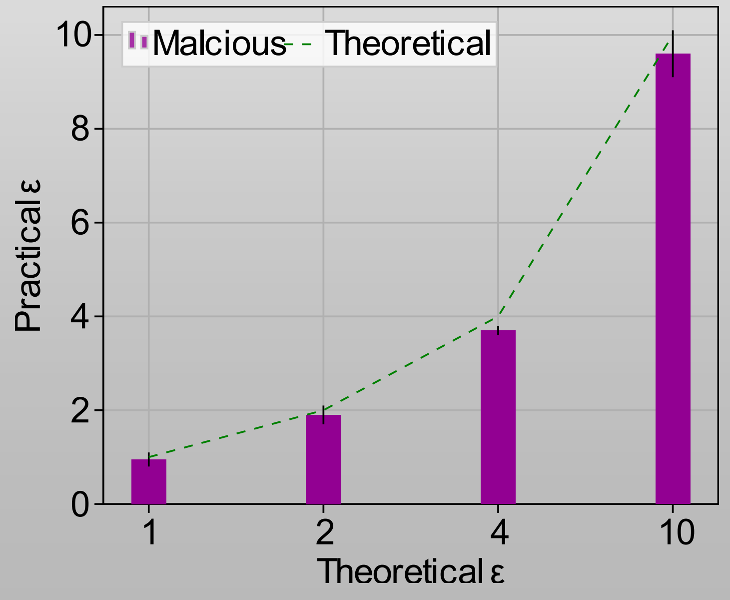


Figure from [Nasr et al. 2021]

The promise of DP-SGD

- What DP-SGD promise?
  - For real strong dataset attacks, what DP promises matches the empirical lower bound
  - The bounds of DP are quite tight.



- On the other hand, if the adversary has physical API restriction: only have black-box access to the trained model (most practical)

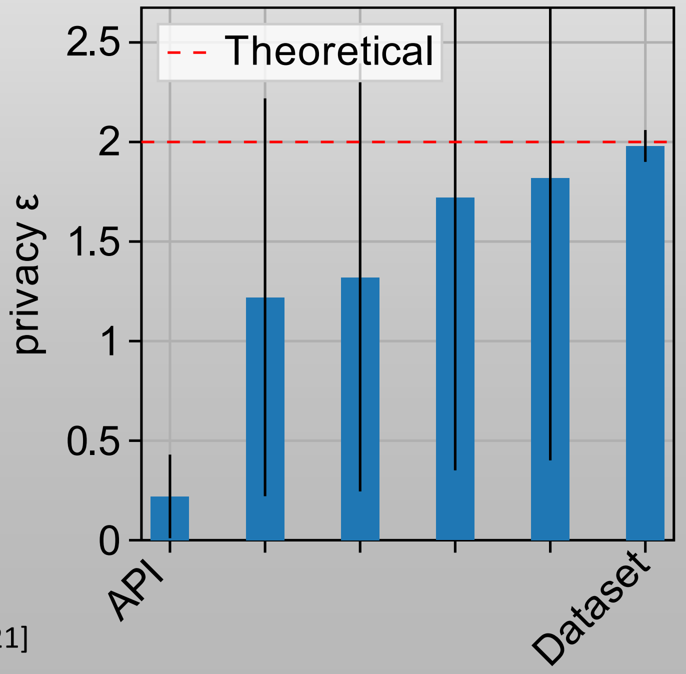


Figure from [Nasr et al. 2021]

The drawbacks of DP-SGD

- **Drawback 1:** The utility depends on output dimension, with large utility drop for large models.

- **Drawback 2:** Computation cost,
  - Handling per-sample gradients requires more computation and much more memory than SGD.
  - Fast and Memory Efficient Differentially Private-SGD via JL Projections [Bu et al. 2021]

Ways to improve private machine learning

1. Hide intermediate updates

- DP-SGD releases the whole trajectory $(\theta_1, \cdots, \theta_T)$, each with DP and then composes the privacy losses together.

- However, often, we only concern the privacy of final output θ_T
  - Intuitively, the privacy parameter of θ_T is strictly smaller than $(\theta_1, \cdots, \theta_T)$
  - How to theoretically argue this?

1. Hide intermediate updates

- Hide the parameters in the mid-steps can help privacy
  - Rishav et al. [2021] prove for strongly convex and smooth loss function, if the initialization is chosen as a Gibbs distribution, the privacy loss of θ_T converges exponentially fast.

$$\varepsilon = O\left(1 - \exp\left(-\frac{O(T)}{2}\right)\right)$$

  - Also, Feldman et al. [2018] demonstrate that for contractive iterations, not releasing the intermediate results amplifies the privacy guarantees.

- Open problem: How to argue the benefit of hiding intermediate updates for general iterative algorithms?

2. Exploit the prior of the learning problem

- For example, the sparse structure of the learning problem [Kalwar et al.2015, Cai et al. 2020].
- Cai et al. 2020 "The cost of privacy"
  - For high-dimensional mean estimation, $\|\mathcal{M}(X) - \mu_P\|_2 \sim O\left(\sqrt{\frac{s \log d}{n}} + \frac{s \log d \sqrt{\log \frac{1}{\delta}}}{n\varepsilon}\right)$, the minmax lower bound and achievable bound match.
  - Algorithm: "peeling + private max". It first identifies the non-zero coordinates (approximately) and set other coordinates to be 0 and then conducts the regression on such support set. It requires the sparsity level.

2. Exploit the prior of the learning problem

- How about the general learning scenario?
  - Train ResNet on CIFAR10
  - Not sparse at all.
- Exploit the prior of the learning problem
  - Via knowledge transfer [Papernot et al.2017, Papernot et al.2018]
  - Via causal structure [Tople et al.2020]
  - Via the redundancy of gradients across samples [Zhou et al.2021, Yu et al.2021a]
  - Via a priori diagonal scaling matrix [Asi et al.2021]
  - Via low-rankness of the gradient of NN layers [Yu et al.2021b]

2. Exploit the prior of the learning problem

- How about the general learning scenario?
  - Train ResNet on CIFAR10
  - Not sparse at all.
- Exploit the prior of the learning problem
  - Via knowledge transfer [Papernot et al.2017, Papernot et al.2018]
  - Via causal structure [Tople et al.2020]
  - Via the redundancy of gradients across samples [Zhou et al.2021, Yu et al.2021a]
  - Via a priori diagonal scaling matrix [Asi et al.2021]
  - Via low-rankness of the gradient of NN layers [Yu et al.2021b]

PATE [Papernot et al.2017&2018]

PATE: Private Aggregation of Teacher Ensembles. It exploits the knowledge transfer ability of NN.

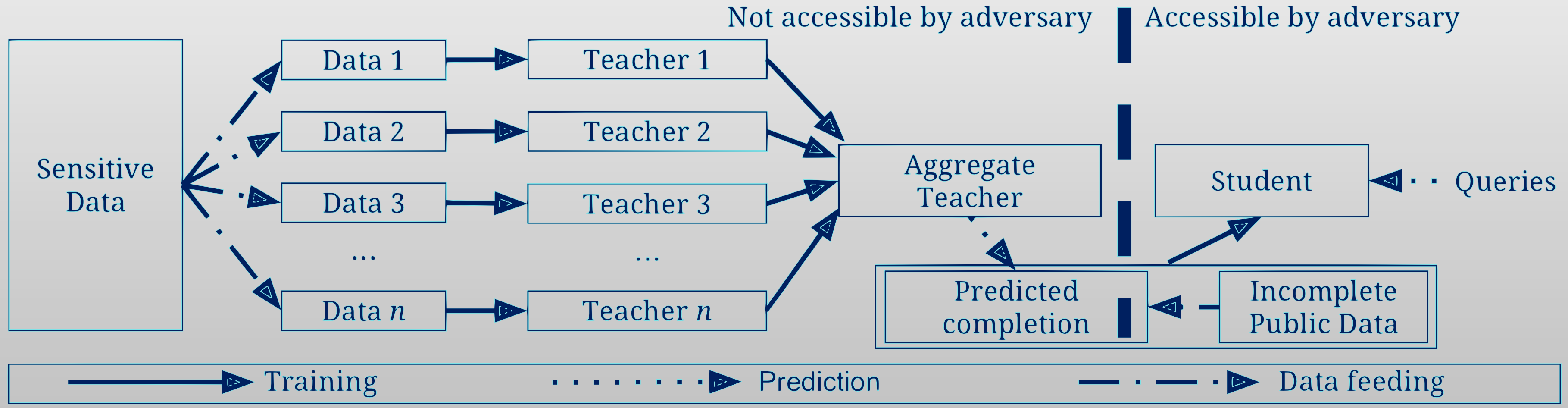
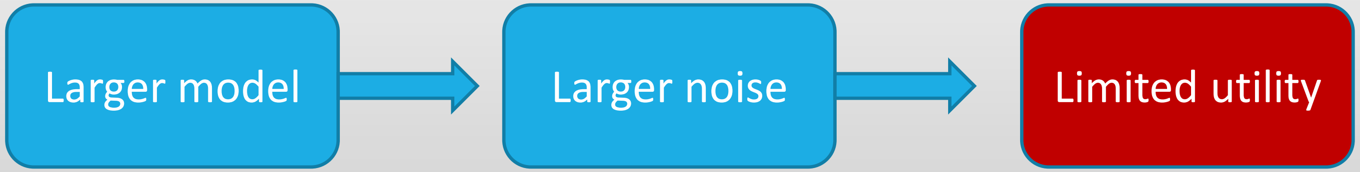


Figure from [Papernot et al. 2017]

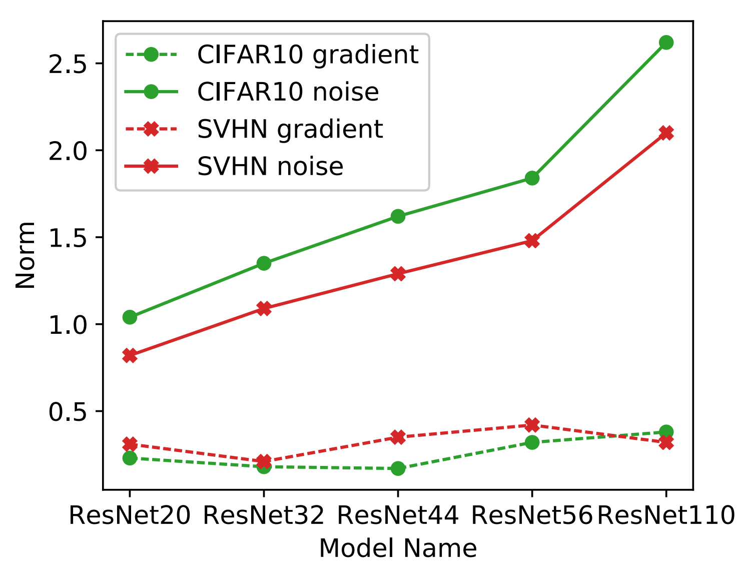
Exploit redundancy of gradients across samples [Zhou et al.2021, Yu et al.2021a]

- Recall one drawback DP-SGD: Bad dimensional dependence



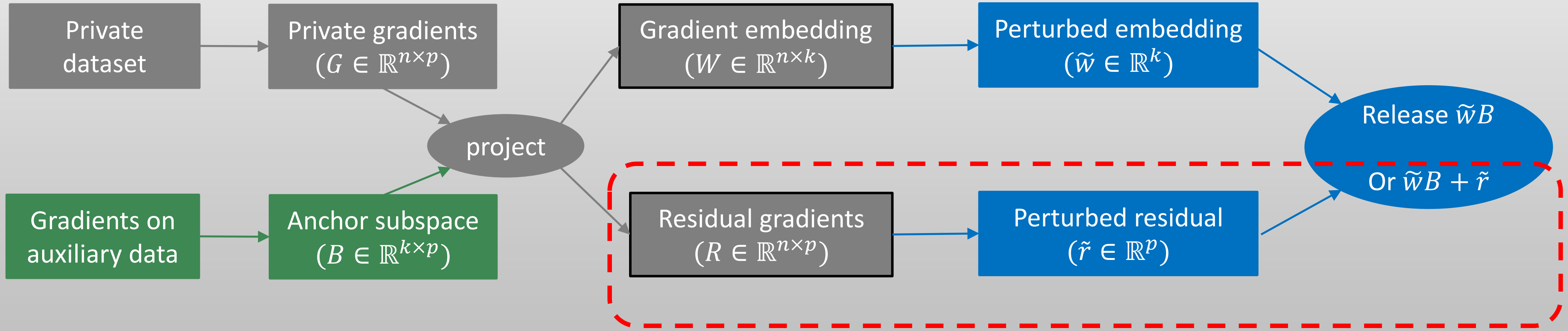
- Gradient perturbation: $\tilde{g} = g + z$, where $g \in \mathbb{R}^p$ and $z \sim N\left(0, \sigma^2 I_{p\times p}\right)$.
    - Note that $||z|| \propto \sqrt{p}$ while $||g||$ roughly unchanged with p.
    - Signals are submerged in noise for large p.

Figure from [Yu et al. 2021]



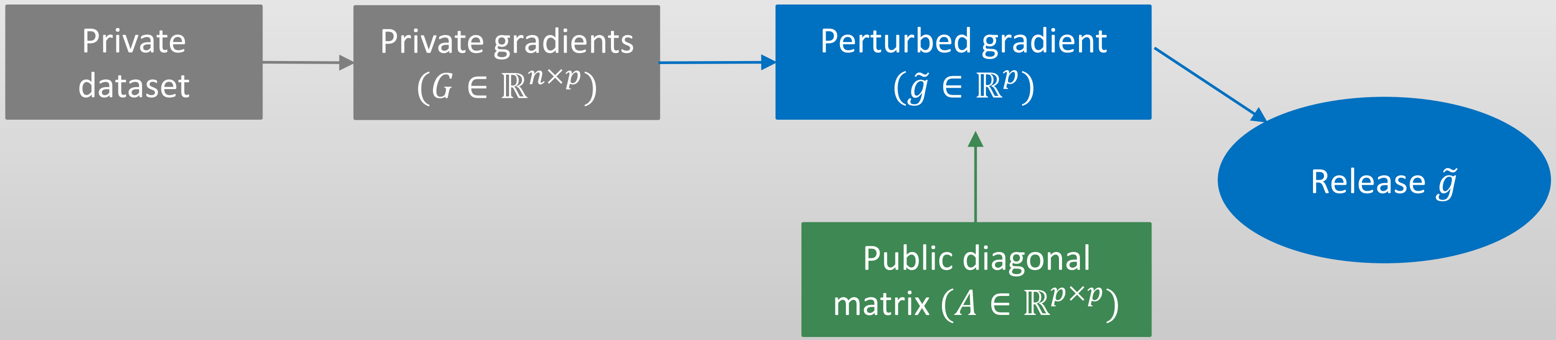
Exploit redundancy of gradients across samples [Zhou et al.2021, Yu et al.2021a]

- IDEA: Project gradient into low-dimensional subspace due to the gradient redundancy across samples.



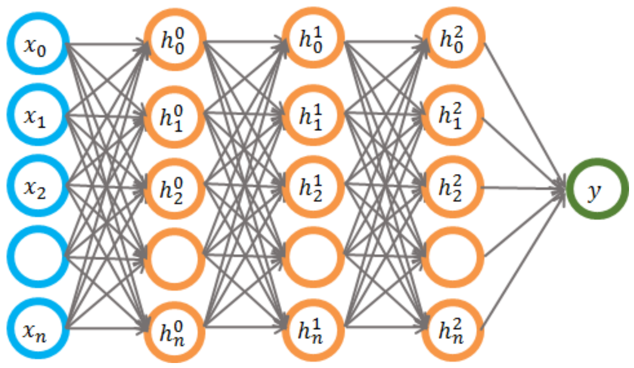
Exploit a priori diagonal scaling matrix [Asi et al.2021]

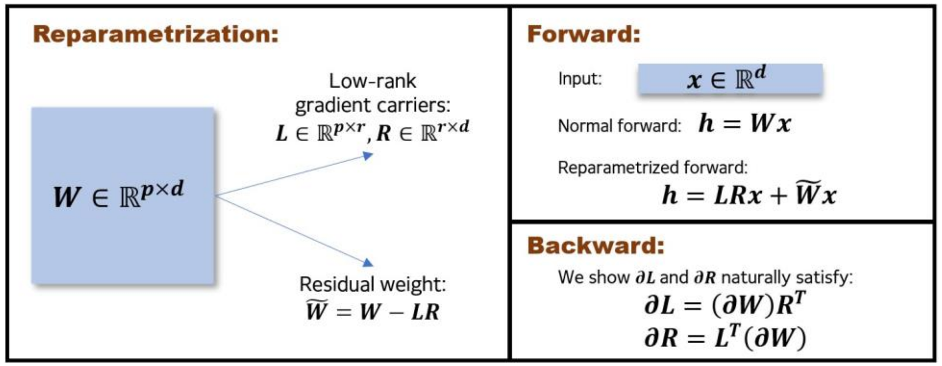
- IDEA: Scale the noise with a diagonal matrix given by a priori knowledge.



Exploit low-rankness of the gradient of NN layers [Yu et al.2021b]

RGP: Reparametrized gradient perturbation. Exploit the low-rankness of the gradient of weight matrix.





The update for W is $(\partial L)R + L(\partial R) - LL^T(\partial L)R$, equivalent to projecting ∂W into the subspace spanned by L and R.

1. Privacy measures

2: Private machine learning

3. ML also borrows from DP

4. What is next?

3. ML also borrows from DP

What does ML borrow from DP?

Theoretically, differential privacy has provided new ways to analyze the generalization, algorithmic stability, concentration in machine learning.

Empirically, the idea of differential privacy has been used to defend a wide range of attacks.

3.1 Algorithmic stability via differential privacy

- Differential privacy can ensure high prob. generalization [Bassily et al.2016, Feldman et al. 2018]: $\Pr\left(gen > O(\varepsilon\Delta)\right) < O\left(\frac{\delta}{\varepsilon}\right)$.
- New concentration inequalities [Steinke&Ullman 2017]
  - Classical result $\forall\varepsilon \geq 0$, $\Pr[\sum_i^n (X_i - \mu_i) \geq \varepsilon n] \leq e^{-\Omega(\varepsilon^2 n)}$.
    - Proof is via MGF + Markov inequality.
  - New proof is based on a proxy $\max\{0, Y^1, \dots, Y^m\}$, where Y^k is copy of $Y = \sum_i^n (X_i - \mu_i)$
  - It works for some heavy tail setting where previous MGF approach fails.

3.1 PAC-Bayesian generalization bound using private prior [Dziugaite & Roy 2018]

- Recap: Let $\mathcal{H}$ be a hypothesis space, and $\ell: \mathcal{H} \times Z \to [0,1]$ be the loss.
- Risk and empirical risk: $L_{\mathcal{D}}(h) = \mathbb{E}_{z\sim\mathcal{D}}[\ell(h,z)]$, $L_S(h) = \frac{1}{n}\sum_i^n \ell(h, z_i)$
- PAC-Bayes generalization bound is for Gibbs classifier, a probability distribution on $\mathcal{H}$.
- The risk of a Gibbs classifier Q is

$$L_{\mathcal{D}}(Q) = \mathbb{E}_{h\sim Q}[L_{\mathcal{D}}(h)] = \mathbb{E}_{z\sim\mathcal{D}}\mathbb{E}_{h\sim Q}[\ell(h,z)]$$

- PAC-Bayes bound [Caoni 2007]: choose a prior P on weights, given a dataset $S \sim \mathcal{D}^n$,

$$\forall Q, L_{\mathcal{D}}(Q) \le L_S(Q) + \sqrt{\frac{\mathrm{KL}(Q \parallel P) + \log\frac{n}{\delta}}{2n}}$$

3.1 PAC-Bayesian generalization bound using private prior [Dziugaite & Roy 2018]

- How to tighten the PAC-Bayes bound?
  - Optimize the prior, find a P^* that is close to the posterior.
  - The prior can depend on data distribution $\mathcal{D}$ but cannot depend on the data
- IDEA: use the data in a safe way to learn a prior. → Learn with differential privacy

Theorem: Let $P(S)$ be an ε-differentially private prior. Then, w. p. $\geq 1-\delta$ over the random sampling of S,

$$\forall Q, \qquad \Delta\big(L_S(Q), L_{\mathcal{D}}(Q)\big) \leq \frac{\mathrm{KL}\big(Q \parallel P(S)\big) + \log\frac{4\sqrt{n}}{\delta}}{2n} + \frac{\varepsilon^2}{2} + \varepsilon\sqrt{\frac{\log 4/\delta}{2n}}$$

- Achieve non-vacuous generalization bound for some deep neural network setting.

3.2 DP defends against practical attacks

- Membership Inference (MI) Attack:

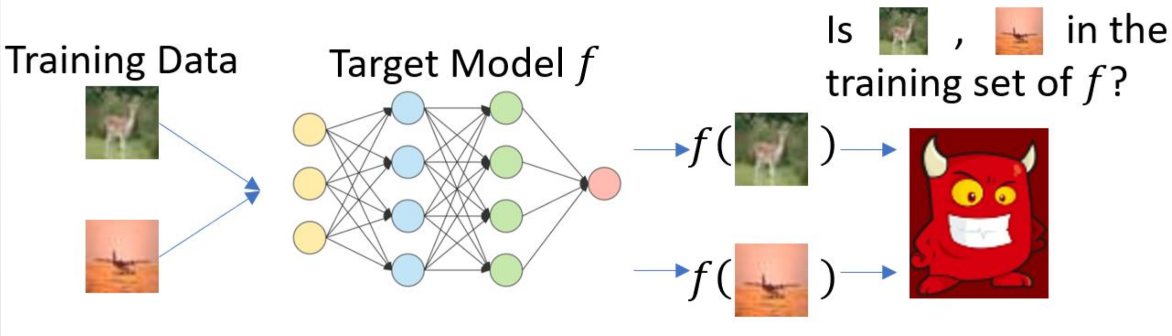


Figure from [Yu et al. 2021]

- Models trained with DP are robust against MI attacks [Bernau et al., 2019].

	MNLI (BERT)	QQP (BERT)	CIFAR10 (ResNet)	SVHN (ResNet)
Non. Priv.	60.3	56.1	58.1	56.4
$\varepsilon = 8$	50.1	50.0	50.3	50.1

Table from [Yu et al. 2021]

3.2 DP defends against practical attacks

- Models trained with differential privacy are also robust against
  - Data poisoning attack [Ma et al. 2019, Hong et al. 2020].
  - Gradient matching attack [Zhu et al. 2019].
  - Adversarial examples, certified robustness [Lecuyer et al. 2019].
  - Model inversion attack [Carlini et al. 2019].

1. Privacy measures

2: Private machine learning

3. ML also borrows from DP

4. What is next?

4. What is next?

Within differential privacy

- There is still a performance gap between non-private learning and private learning.
  - Large gap to improve
  - Efficiency for training extreme large models (GPT2/3) with differential privacy
- New relaxations: Bayesian differential privacy [Triastcyn&Faltings 2020]
- Relation between private learning and online learning [Abernethy et al. 2019, Jung et al. 2020]
- Differential privacy and fairness
  - Joint private and fair learning algorithm [Jagielski et al. 2019, Mozannar et al. 2020]. Is privacy at odds with fairness?
- Privacy, memorization and generalization [Zhang et al.2019, Feldman 2020]
  - Does learning require memorization?
  - DP is against memorization and DP is used to show generalization.

Beyond differential privacy

- Privacy measure in language model [Zanella-Béguelin et al. 2020, Inan et al. 2021]
  - Perplexity as privacy measure.
  - API boundary
- Generative models
  - DP-GAN [Neunhoeffer et al. 2021]
  - Use GAN to extract original dataset [Cai et al. 2021]
- Privacy in federated learning

1. Privacy measures

2: Private machine learning

3. ML also borrows from DP

4. What is next?