1St Conference on Artificial Intelligence and Future Technologies (ICAIFT2023)

دانشگاه صنعتی نوشیروانی بابل

Accuracy Improvement in Differentially Private Logistic Regression: A Pre-training Approach

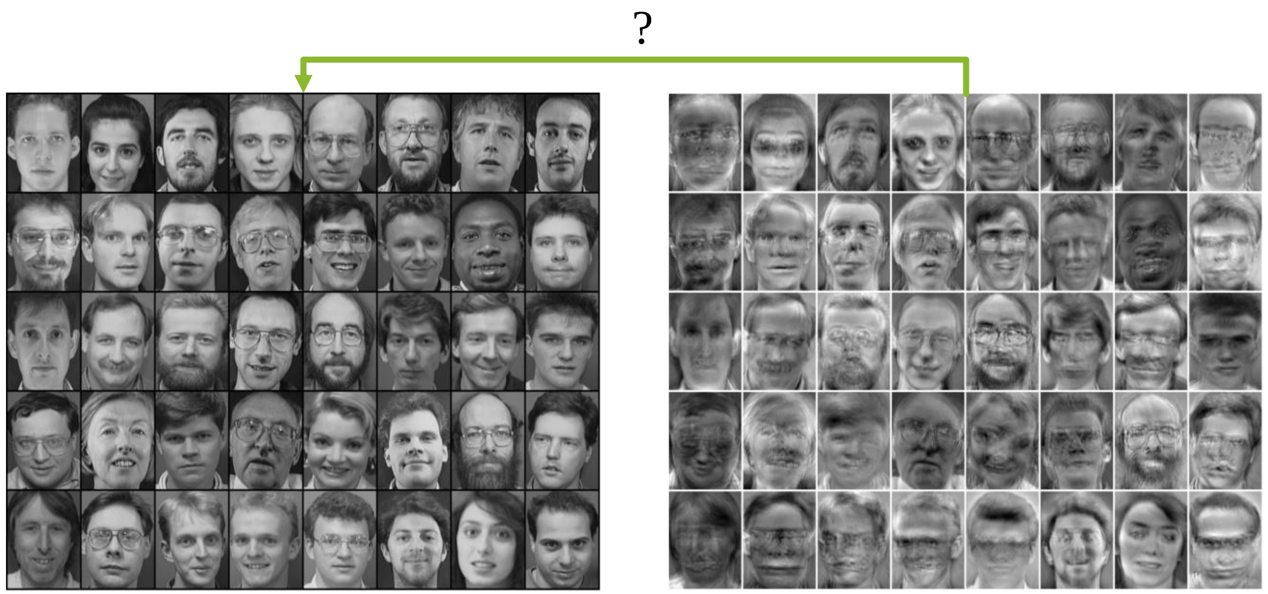
Mohammad Hoseinpour, Milad Hoseinpour, Ali Aghagolzadeh

Presenter
Mohammad Hoseinpour

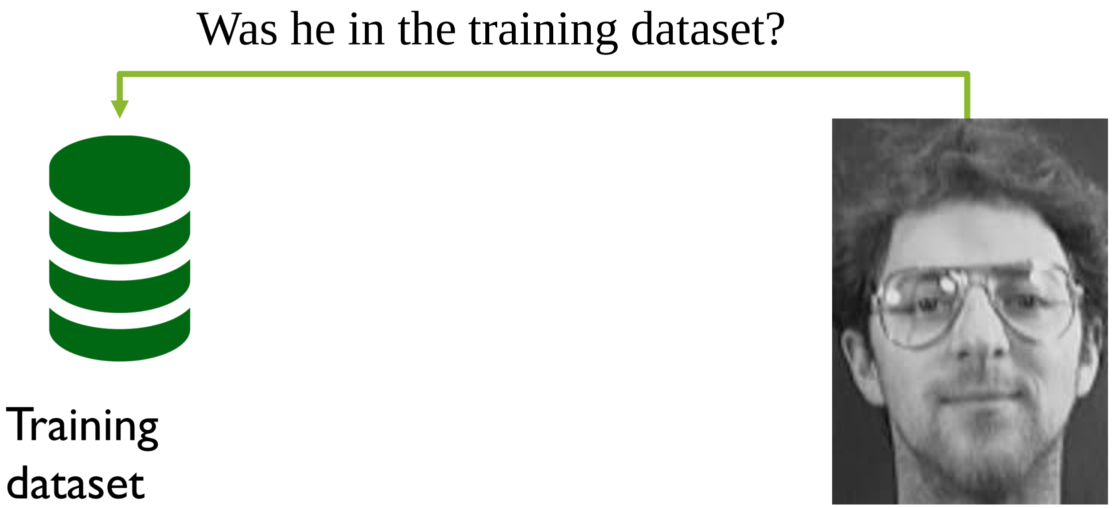
November 15, 2023

Motivation

- Machine Learning (ML) models can **memorize** training datasets.
- Training ML models over **private datasets** can **violate** the **privacy of individuals**.
- Training data extraction attacks:



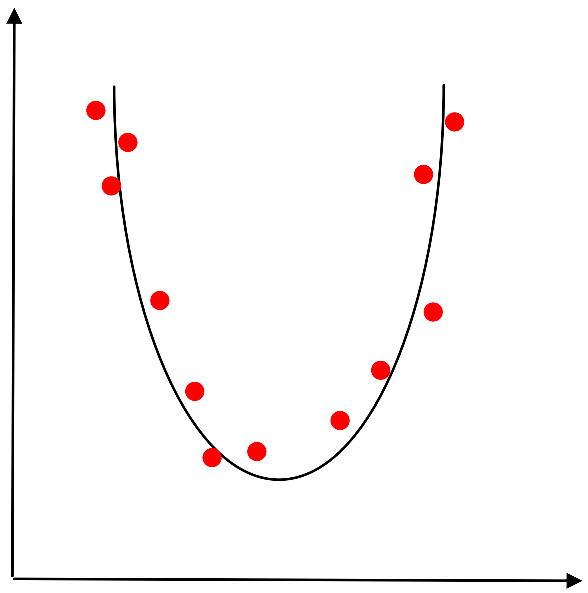
Model Inversion Attacks



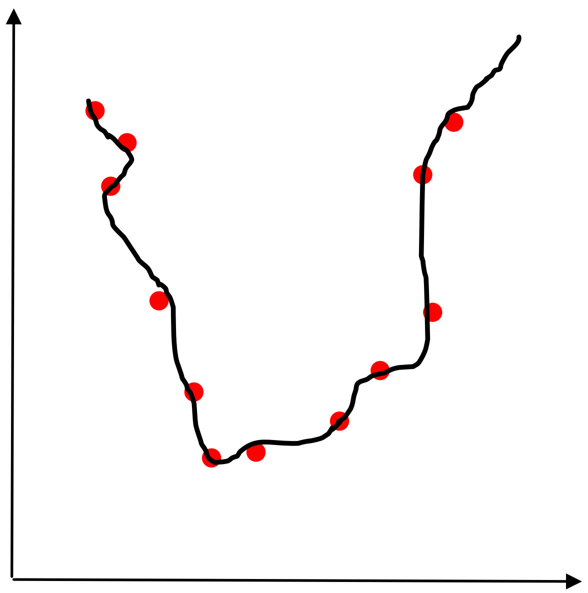
Membership Inference Attacks

Motivation

- Backward problem: Given the output model, find “N” training data points



Low generalization error

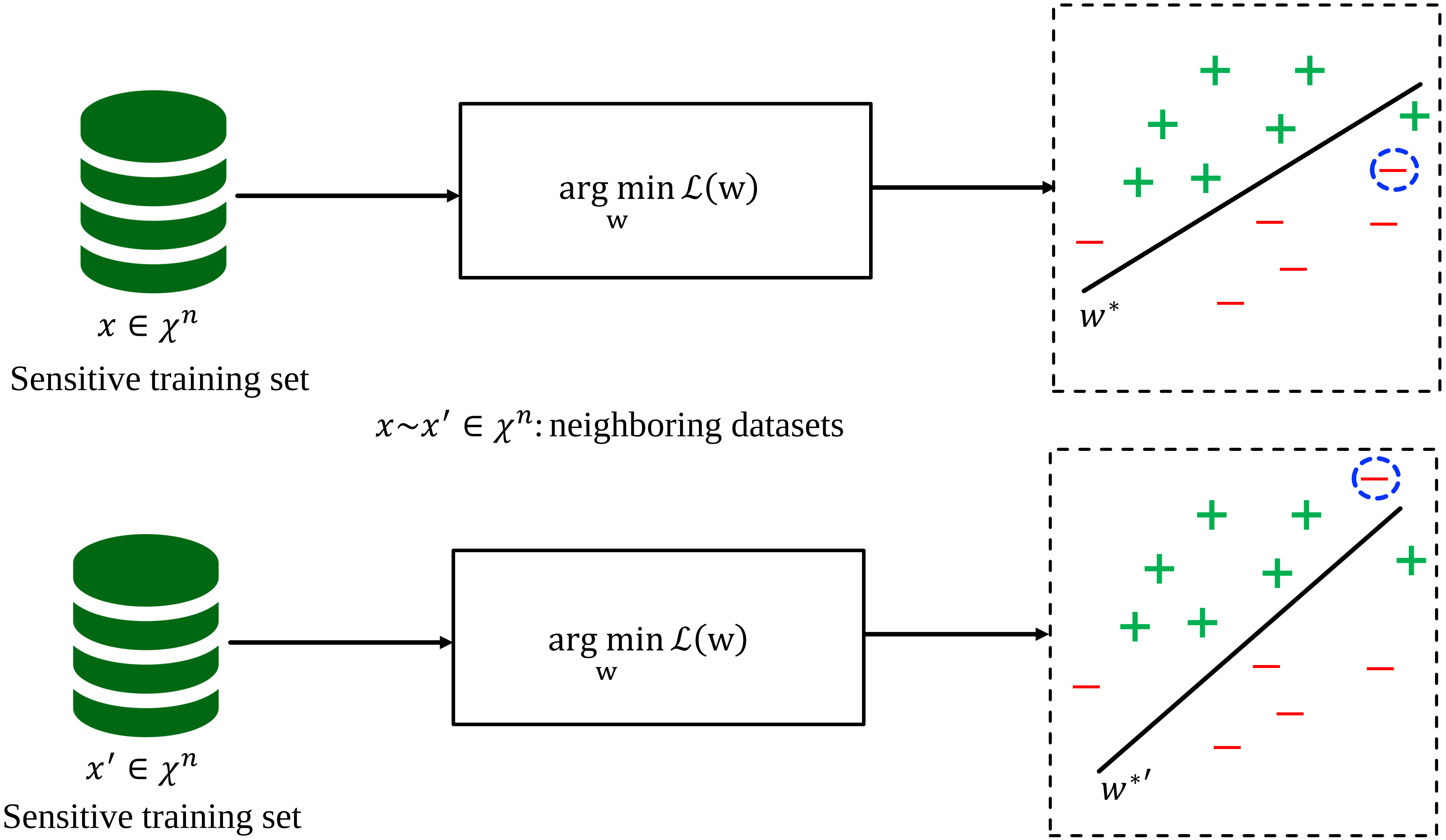


Overfitting: High generalization error

- Backward problem **easier** for **overfitted** models.
- The curve on the right contains **more information** about the training data points.

Non-Private Logistic Regression

- The **decision boundary** of the classifier is **sensitive** to the **individual data points** in the training set.



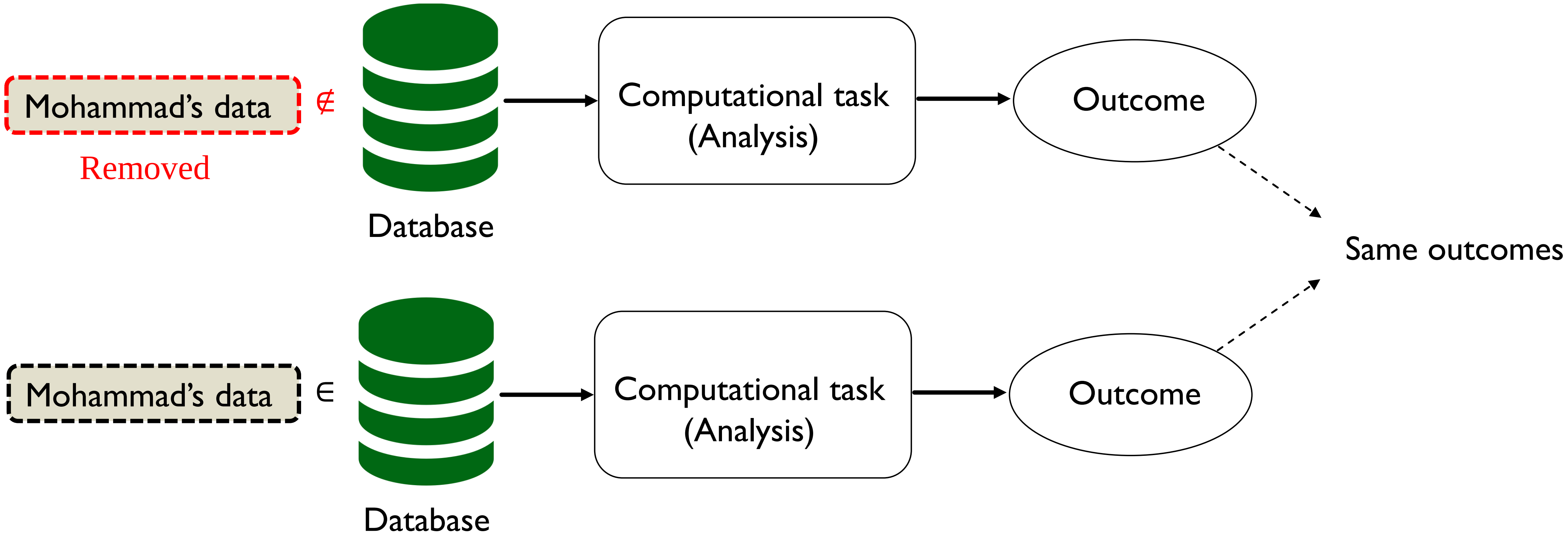
An adversary could **distinguish** whether the training set was x or x'.

Differential Privacy

- To achieve our privacy goal, we use **differential privacy,** which gives us a mathematical framework to **quantify** and **bound** the privacy risk of individuals in the dataset.
- At a high level, differential privacy ensures that **the presence or absence of any individual record in the dataset does not significantly affect the outcome of the computation**.

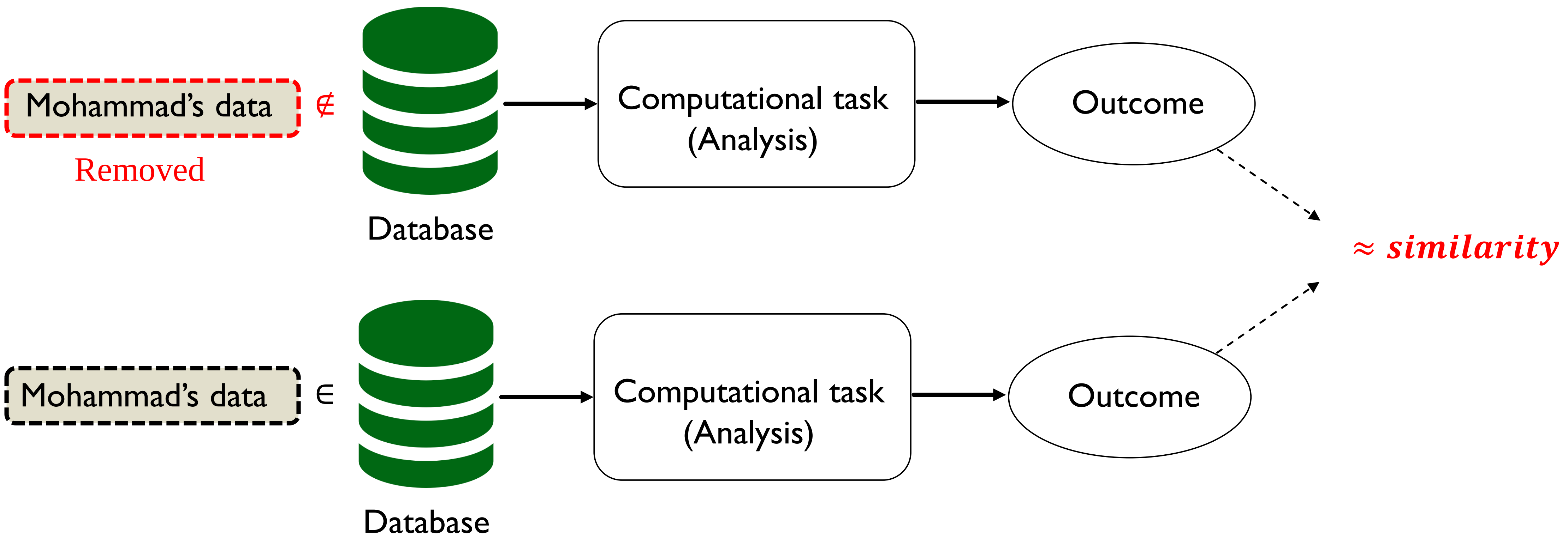
Differential Privacy

- Suppose that I am a privacy aware individual, and I am worried about sharing my data in a computation.
- In an **ideal world**, I would be happy if the outcome of the computation is **the same** whether or not my data is included in the database.



Differential Privacy

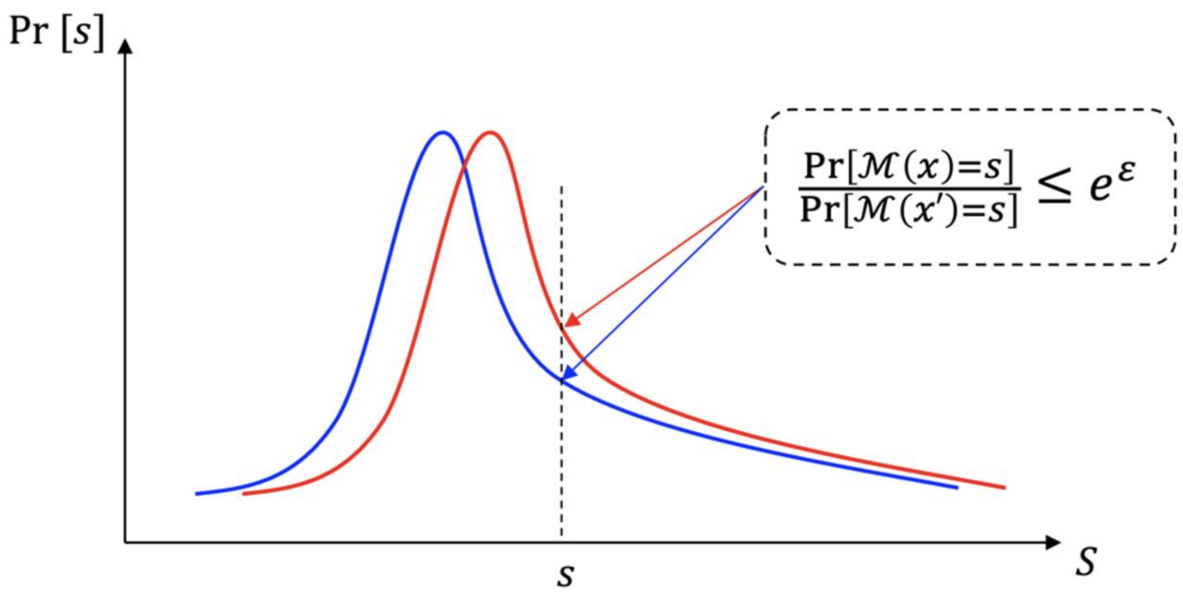
- In a **more realistic world**, the outcome of the computation should be **almost the same** whether or not my data is included in the database.



Differential Privacy

- **Definition:** For $\epsilon \geq 0$, $\delta \in [0,1]$, a **randomized algorithm** $\mathcal{M}: \mathcal{X}^n \to \mathcal{R}$ is (ϵ, δ) −**differentially private** if for every pair of neighboring datasets $x \sim x' \in \mathcal{X}^n$(i.e., x and x' differ in one element) and for any subset of the output space $S \subseteq \mathcal{R}$, the following holds:

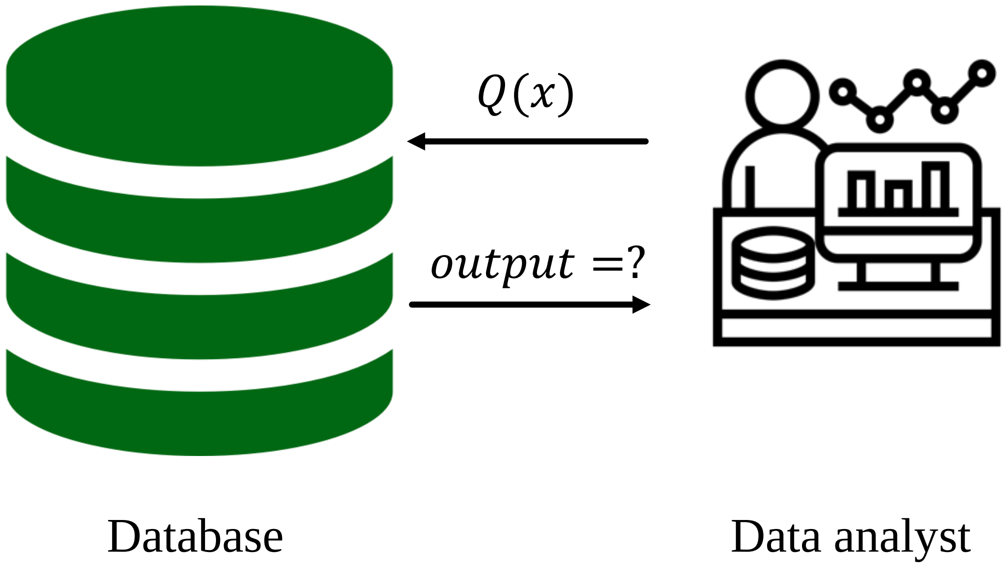
$$\Pr[\mathcal{M}(x) \in S] \leq e^{\varepsilon}.\Pr[\mathcal{M}(x') \in S] + \delta\,.$$



Achieving Differential Privacy

- The **required randomization** for achieving differential privacy in a computation is calibrated based on the **global sensitivity** of that computation:

$$GS(Q) = \max_{x \sim x' \in \mathcal{X}^n} \|Q(x) - Q(x')\|_1$$

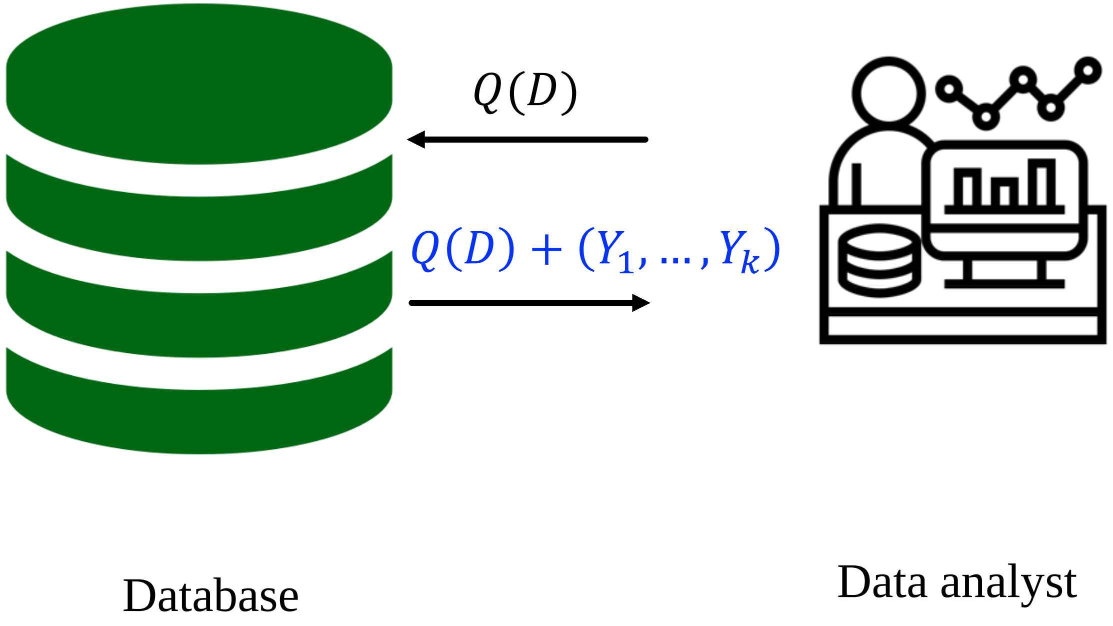


Achieving Differential Privacy

- Gaussian Mechanism:

Gaussian (D,Q: $\mathcal{X}^n \rightarrow \mathbb{R}^k$, ε):

1. Let Δ= GS (Q).
2. For $i = 1$ to k: Let $Y_i \sim N(0, \frac{2\Delta^2 \log(\frac{2}{\delta})}{\varepsilon^2})$.
3. Output: $Q(D) + (Y_1, \dots, Y_k)$.



Private Logistic Regression

- We apply **Gaussian mechanism** for privatizing the **updating rule** of the gradient descent.

Empirical Risk: $$J(\mathrm{w}) = \frac{1}{n}\sum_{i=1}^{n} j\big(\mathrm{w}, (x_i, y_i)\big)$$

Algorithm1：Gradient Descent
Inputs: *noise parameter* (σ>0), Learning Rate α,
1: $\boldsymbol{w}_0$ = initial value for w
2: for t=1,2, … ,T :
3: $g_t = \nabla J(\boldsymbol{w}_{t-1})$;
5: $\boxed{\boldsymbol{w}_t = \boldsymbol{w}_{t-1} - \eta g_t}$;
6:Return $\boldsymbol{w}_T$

This computation should be done “differentially private”.

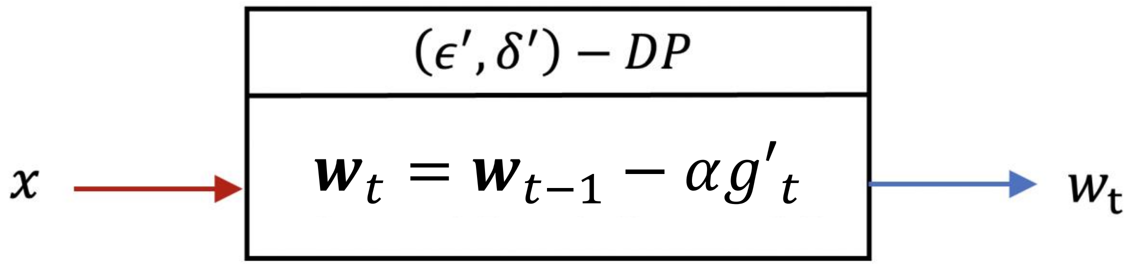
Private Logistic Regression

- We have to choose noise according to the **ℓ_2-sensitivity** of the **gradient**

$$GS(\nabla J(x)) = \max_{x\sim x'} \|\nabla J(\mathrm{w}; x) - \nabla J(\mathrm{w}; x')\|_2$$

$$GS(\nabla J(x)) = \max_{x\sim x'} \|\nabla J(x) - \nabla J(x')\|_2 \leq \max_{x\sim x'} (\|\nabla J(x)\|_2 + \|\nabla J(x')\|_2) = 2C$$

- For achieving (ϵ', δ')-DP in **each iteration**, we should add Gaussian noise with $\sigma \geq \frac{2C}{n\epsilon'}\sqrt{2ln\left(\frac{1.25}{\delta'}\right)}$.



Privatizing each iteration of the gradient descent

Private Logistic Regression

- We apply **Gaussian mechanism** for privatizing the **updating rule** of the gradient descent.

Empirical Risk: $J(\mathrm{w}) = \frac{1}{n}\sum_{i=1}^{n} j\big(\mathrm{w},(x_i,y_i)\big)$

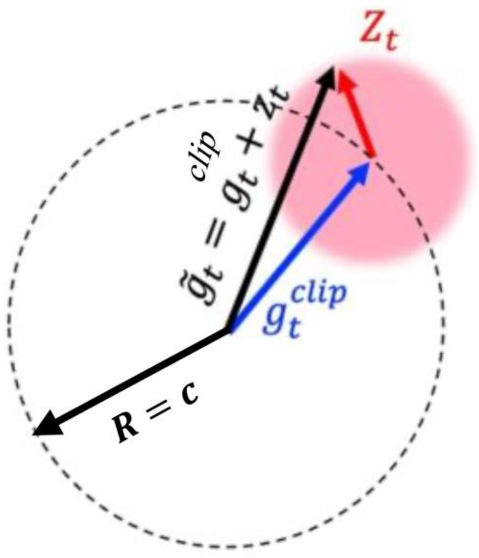
Algorithm 2: Noisy Gradient Descent

Inputs: *noise parameter* (σ>0), Learning Rate α,
1: $\boldsymbol{w}_0$ = initial value for w
2: for t=1,2, … ,T :
3: $g_t = \nabla J(\boldsymbol{w}_{t-1})$;
4: clip the gradient:

$$g_t^{clip} = \frac{g_t}{max(1, \|g_t\|_2/C)}$$

4: $g'_t = g_t^{clip} + N(0, \sigma^2 I_d)$;
5: $\boldsymbol{w}_t = \boldsymbol{w}_{t-1} - \alpha g'_t$;
6:Return $\boldsymbol{w}_T$

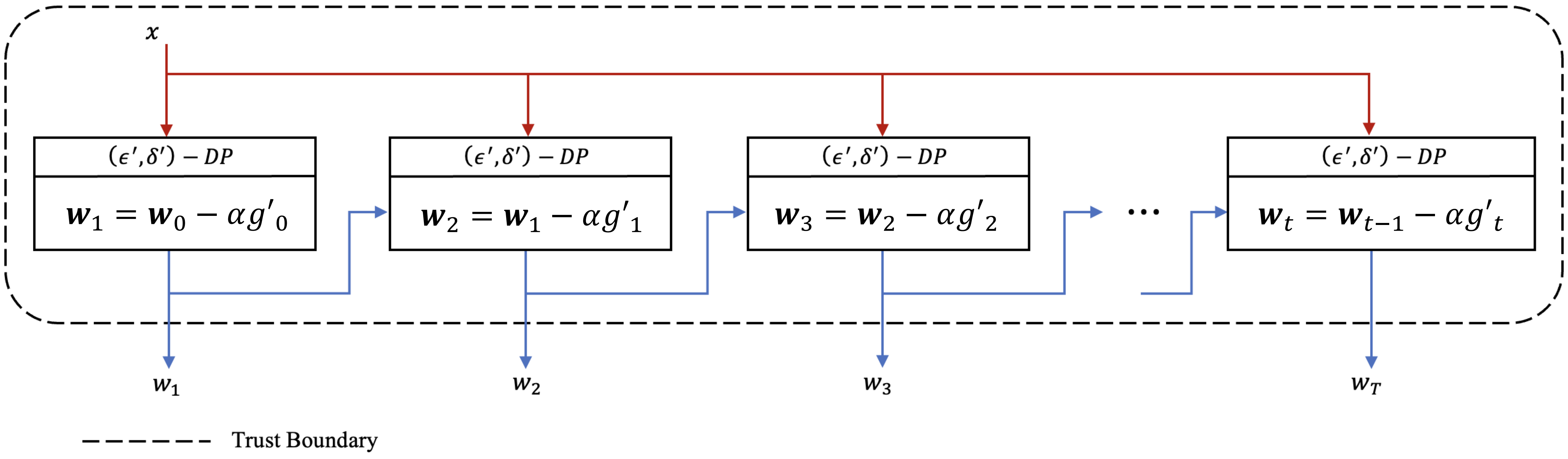
Perturbed gradient vector due to the additive Gaussian noise



Private Logistic Regression

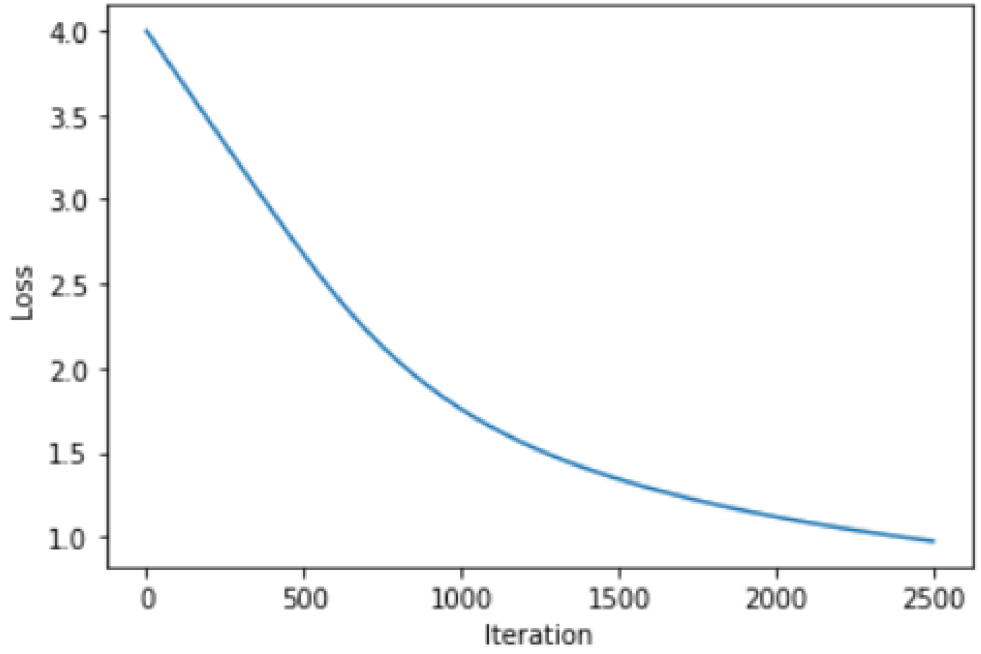
- Due to **advanced composition**, for achieving (ϵ, δ)-DP in the **composition of *T* iteration** of the gradient descent, we should add **Gaussian noise** with

$$\sigma \geq \frac{2C}{n\epsilon}\sqrt{2Tln\left(\frac{1.25}{\delta}\right)}.$$

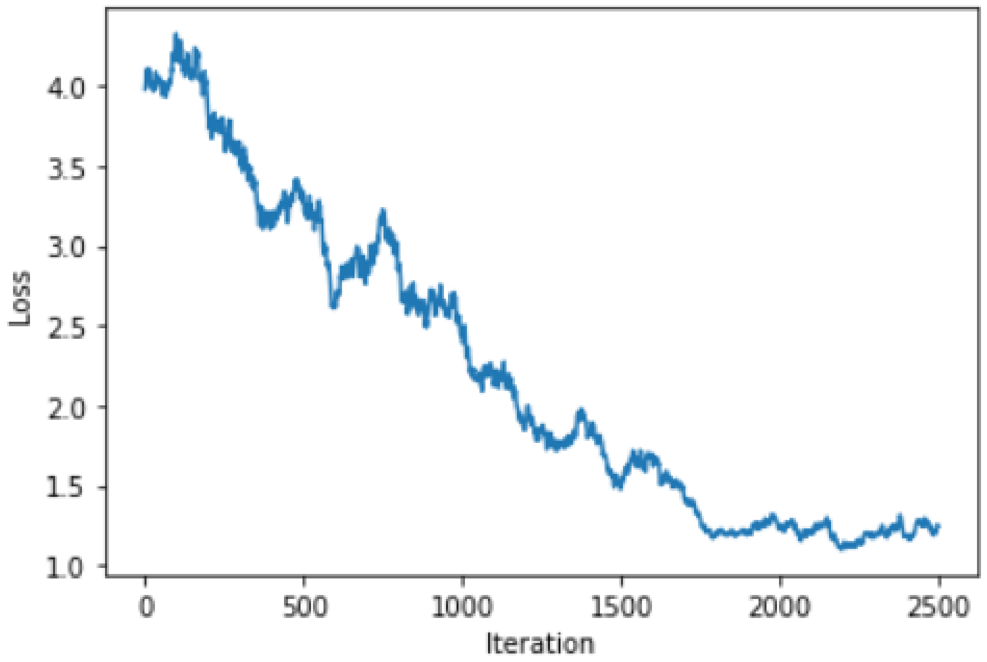


Private Logistic Regression

- Convergence of the gradient descent under **"no privacy"** and **"privacy constraint"**.



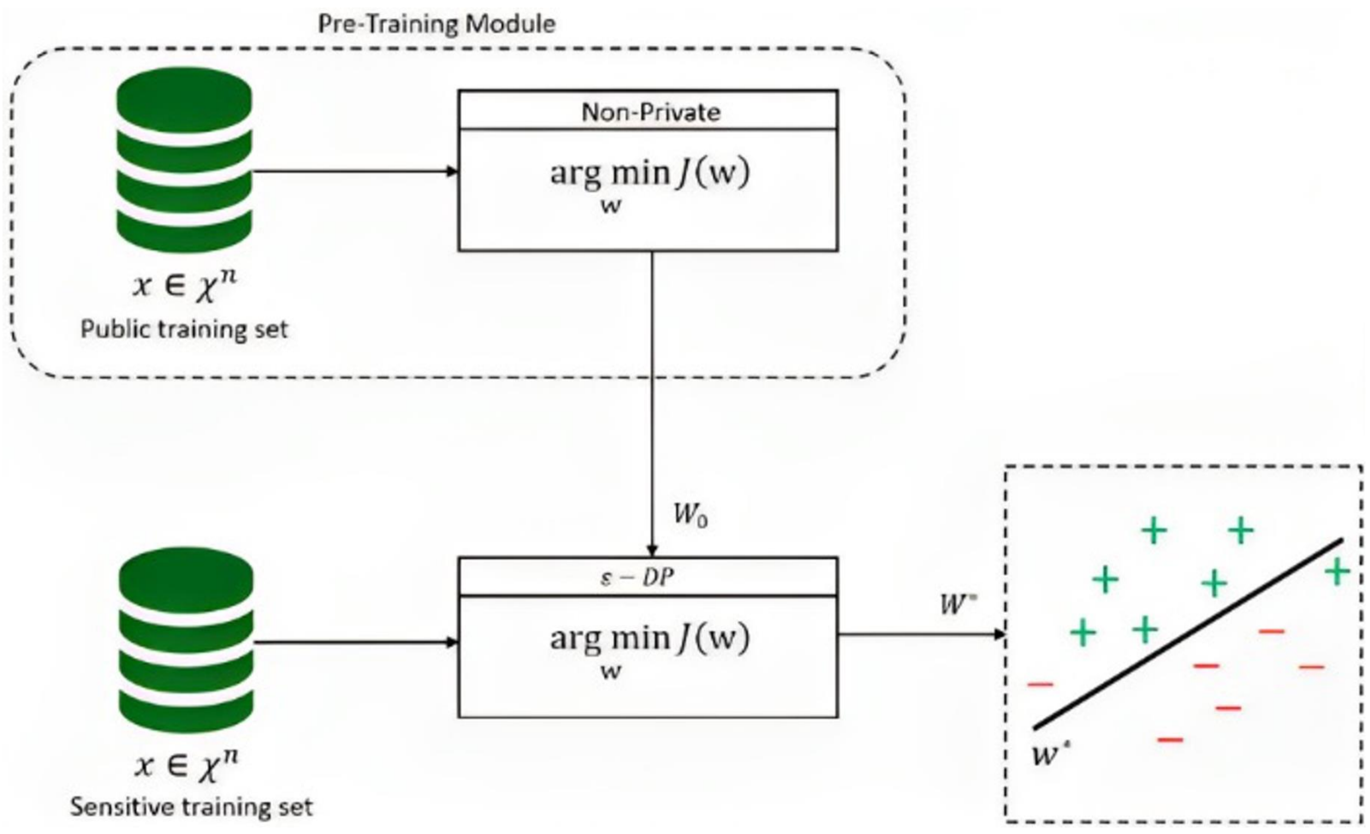
Loss versus iteration of the LR model with **no privacy** constraint and **% 67.50** training accuracy.



Loss versus iteration of the DP-LR model with $\boldsymbol{\varepsilon = 1}$ and **% 60.25** training accuracy.

Accuracy Improvement: Pre-training Module

- One main challenge is the **inherent trade-off** between the **accuracy** and **privacy** in DP-ML models.
- To improve the accuracy, we **pre-train** our model on a **public training dataset** that there is **no privacy concern** about it.
- Then, we **fine-tune** our model via the DP-LR with the **private dataset**.



Results

- In a **very high privacy** regime, the **accuracy improvement** by adding the pre-training module is **negligible**.

"Very high" privacy regime

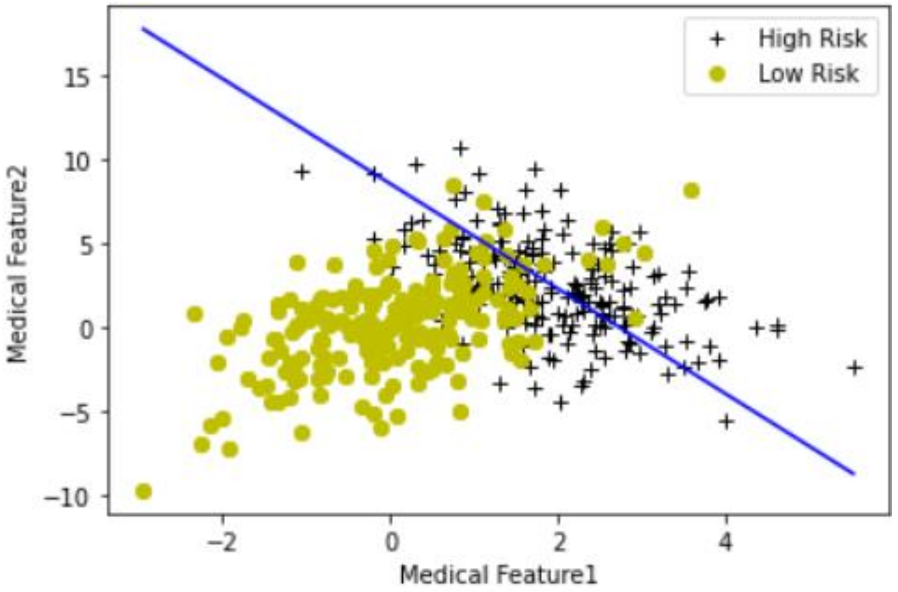
"Practical" privacy regime

ε	Accuracy With No Pre-training Module	Accuracy With Pre-training Module	Enhancement
0.01	%29.75	%29.75	≈0
0.05	%33.00	%33.50	%0.5
0.1	%40.25	%41.25	%1.25
0.5	%53.25	%60.00	%7.25
1	%60.25	%70.50	%10.25
5	%66.25	%77.25	%11
10	%66.50	%77.50	%11
15	%67.00	%77.50	%10.50
150	%67.50	%78.00	%11.50

Results

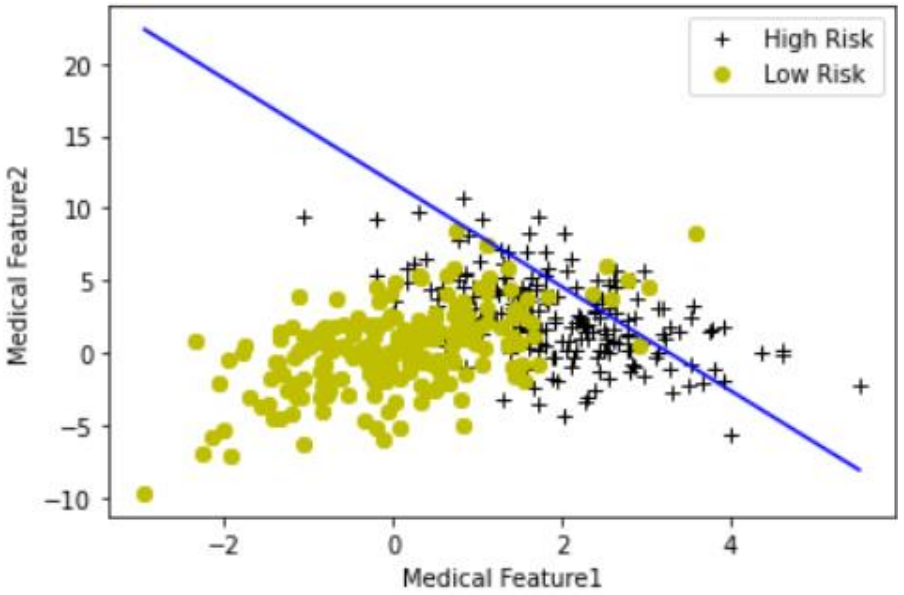
- Decision boundaries of the pre-trained DP-LR model under different privacy regimes.

$\varepsilon = 1$



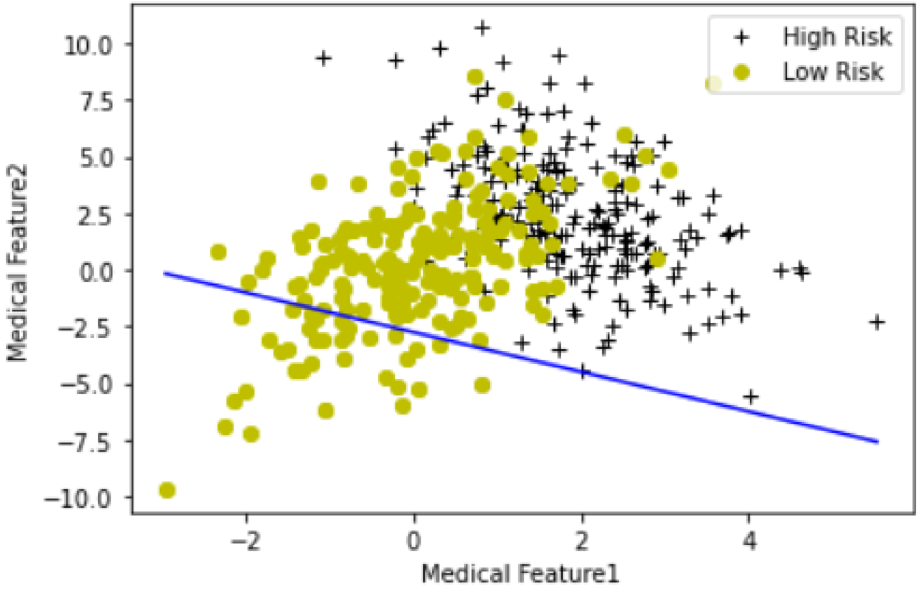
Accuracy =% 70.50

$\varepsilon = 0.5$



Accuracy =% 60

$\varepsilon = 0.1$



Accuracy =% 41.25

Thank You!