Salvador Dalí, "The Persistence of Memory," 1931

Membership Inference Attack & Differential Privacy

Lecturer: Dr. Xingjun Ma

School of Computer Science, Fudan University

Fall, 2022

Recap: week 9

- Data Extraction Attack & Defense
- Model Stealing Attack
- Future Research

Final Project

- ◆ 自选研究题目（**占比60%**）
  - 有8个备选题目，第10周左右发布
  - **需要组队：每组人数不超过3人，每组最多有2个博士**
  - **需要做实验，需要写报告**（英文报告加分）
  - **需要课堂作展示**，**每个组10分钟**

- ■ **得分**：结合创新性、报告质量、展示质量三个方面综合评分

- ✓ **可以做自己的研究课题相关的内容**
- ✓ **围绕可信（鲁棒性、安全性、可解释性、隐私性、公平性等等）进行**
- ✓ **可以揭示新问题，可以攻击，可以防御**
- ✓ **问题可大可小，但角度一定要有创意**

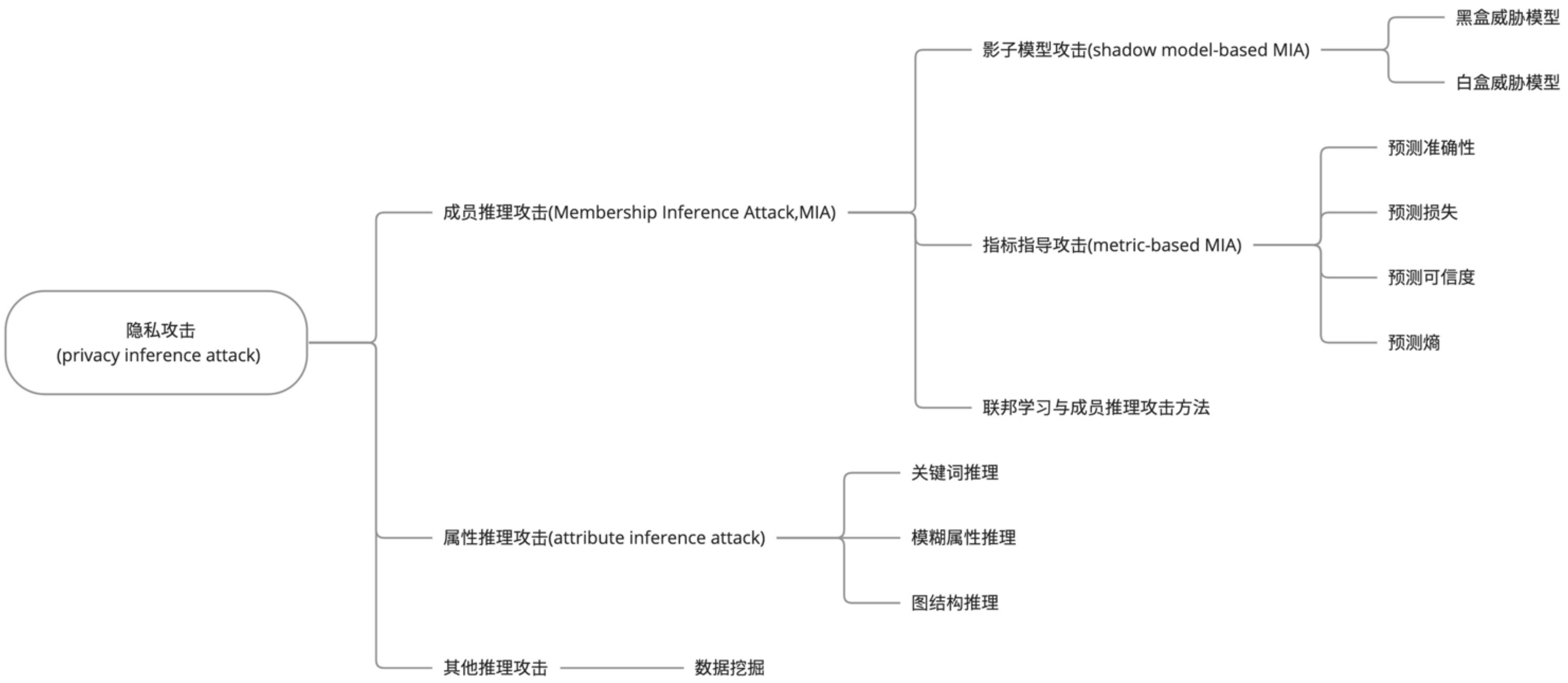
https://trustworthymachinelearning.github.io/student-project/index.html

This Week

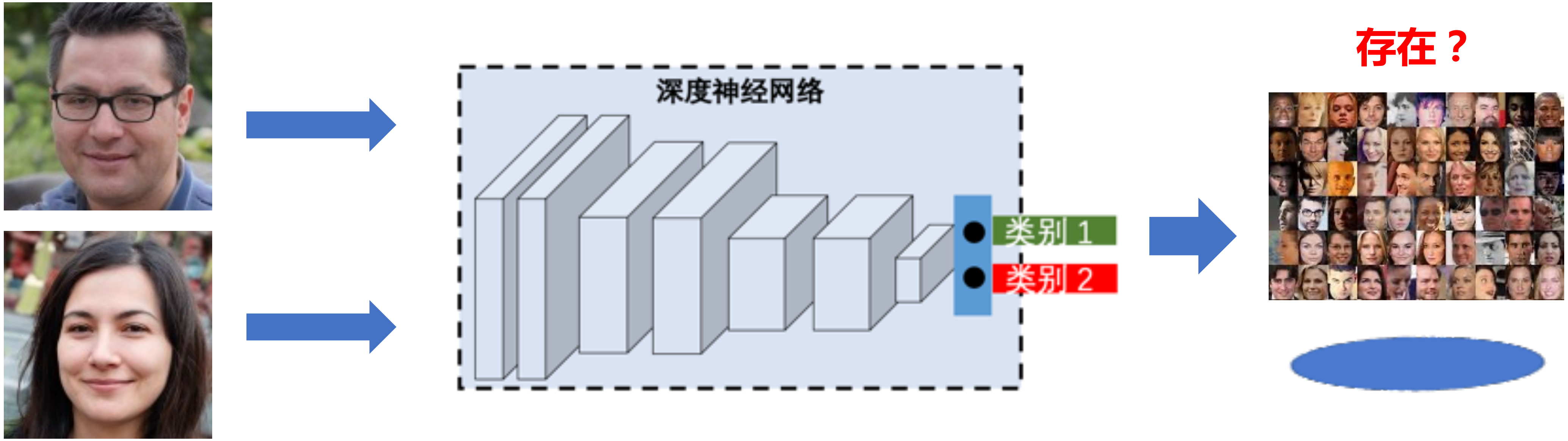
- Membership Inference Attack
- Differential Privacy

- **Membership Inference Attack**
- Differential Privacy

Membership Inference Attack



Membership Inference Attack



推理一个输入样本是否存在于训练数据集中

Shokri, Reza, et al. "Membership inference attacks against machine learning models." S&P, 2017.

Privacy and Ethical Problems

- MIA could cause the following harms:
  - Leak private info: someone has been to some place or having an unspeakable illness
  - Expose info about the training data
  - MIA sensitivity also indicates data leakage risk

An Early Work

OPEN ACCESS Freely available online

PLoS GENETICS

Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays

Nils Homer[1,2], Szabolcs Szelinger[1], Margot Redman[1], David Duggan[1], Waibhav Tembe[1], Jill Muehling[1], John V. Pearson[1], Dietrich A. Stephan[1], Stanley F. Nelson[2], David W. Craig[1*]

1 Translational Genomics Research Institute (TGen), Phoenix, Arizona, United States of America, 2 University of California Los Angeles, Los Angeles, California, United States of America

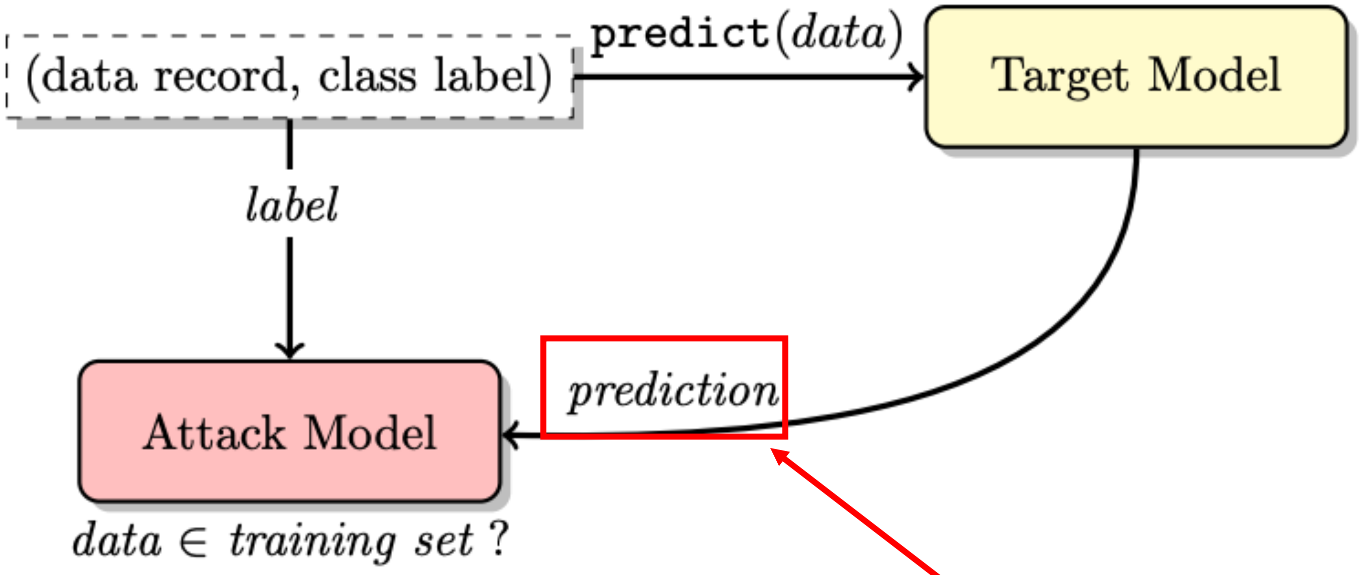
Abstract

We use high-density single nucleotide polymorphism (SNP) genotyping microarrays to demonstrate the ability to accurately and robustly determine whether individuals are in a complex genomic DNA mixture. We first develop a theoretical framework for detecting an individual's presence within a mixture, then show, through simulations, the limits associated with our method, and finally demonstrate experimentally the identification of the presence of genomic DNA of specific individuals within a series of highly complex genomic mixtures, including mixtures where an individual contributes less than 0.1% of the total genomic DNA. These findings shift the perceived utility of SNPs for identifying individual trace contributors within a forensics mixture, and suggest future research efforts into assessing the viability of previously sub-optimal DNA sources due to sample contamination. These findings also suggest that composite statistics across cohorts, such as allele frequency or genotype counts, do not mask identity within genome-wide association studies. The implications of these findings are discussed.

- 判断个人基因是否出现在一个复杂的混合基因里
- 可用于调查取证

Homer, Nils, et al. "Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays." *PLoS genetics* 4.8 (2008): e1000167.

MIA：The First Well-known Work

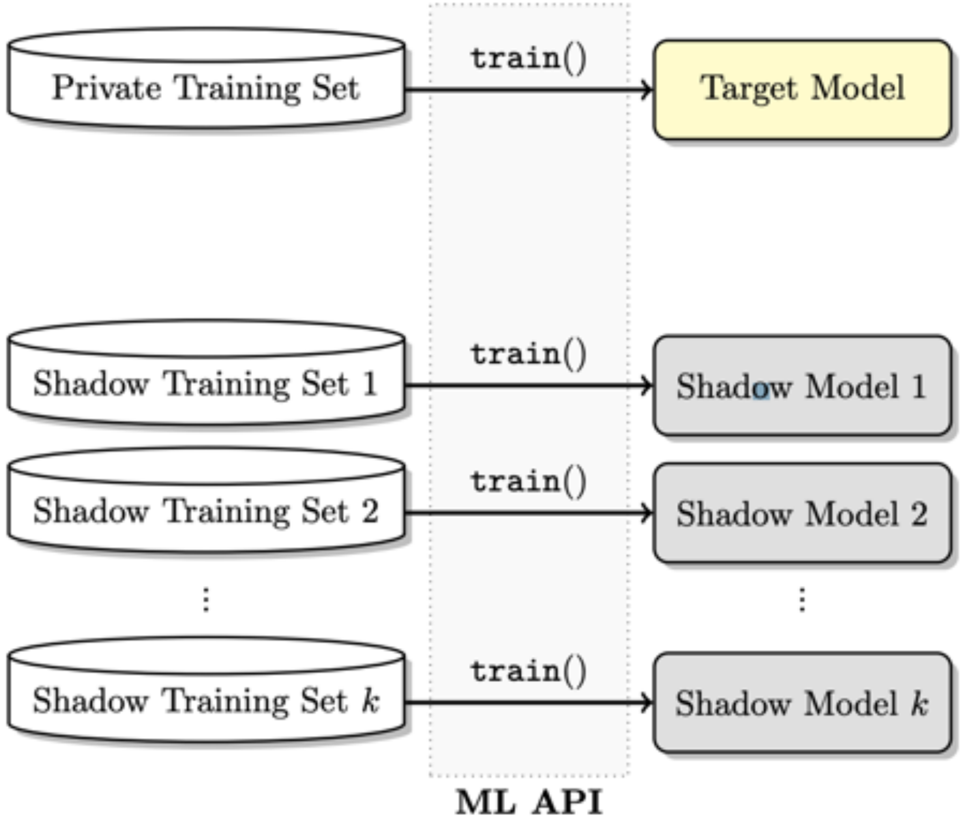


Needs probability vector

Black-box attack pipeline

Shokri, Reza, et al. "Membership inference attacks against machine learning models." S&P, 2017.

MIA：The First Well-known Work



① Sample a number of subsets from D
② Train a model on each of the subset
③ Take one model as the target
④ Take the rest models as shadow models

Train *k* shadow models on disjoint datasets

Shokri, Reza, et al. "Membership inference attacks against machine learning models." S&P, 2017.

MIA：The First Well-known Work

◻ Different ways to get the training data：Random Synthesis

Algorithm 1 Data synthesis using the target model

```
procedure SYNTHESIZE(class : c)
    x ← RANDRECORD( )                          ▷ initialize a record randomly
    y*_c ← 0
    j ← 0
    k ← k_max
    for iteration = 1 ··· iter_max do
        y ← f_target(x)                        ▷ query the target model
        if y_c ≥ y*_c then                     ▷ accept the record
            if y_c > conf_min and c = arg max(y) then
                if rand() < y_c then           ▷ sample
                    return x                   ▷ synthetic data
                end if
            end if
            x* ← x
            y*_c ← y_c
            j ← 0
        else
            j ← j + 1
            if j > rej_max then                ▷ many consecutive rejects
                k ← max(k_min, ⌈k/2⌉)
                j ← 0
            end if
        end if
        x ← RANDRECORD(x*, k)                  ▷ randomize k features
    end for
    return ⊥                                   ▷ failed to synthesize
end procedure
```

◻ Data synthesis

- **Phase 1:** searching for high confidence data points in the data space
- **Phase 2:** sample synthetic data from these points
- Repeat the above for each class *c*

Phase 1: 每次只改变已找到的高置信度样本的*k*个特征

Shokri, Reza, et al. "Membership inference attacks against machine learning models." S&P, 2017.

MIA：The First Well-known Work

□ Statistics-based synthesis

Algorithm 1 Data synthesis using the target model

```
procedure SYNTHESIZE(class : c)
    x ← RANDRECORD( )                    ▷ initialize a record randomly
    y*_c ← 0
    j ← 0
    k ← k_max
    for iteration = 1 ··· iter_max do
        y ← f_target(x)                  ▷ query the target model
        if y_c ≥ y*_c then               ▷ accept the record
            if y_c > conf_min and c = arg max(y) then
                if rand() < y_c then     ▷ sample
                    return x             ▷ synthetic data
                end if
            end if
            x* ← x
            y*_c ← y_c
            j ← 0
        else
            j ← j + 1
            if j > rej_max then          ▷ many consecutive rejects
                k ← max(k_min, ⌈k/2⌉)
                j ← 0
            end if
        end if
        x ← RANDRECORD(x*, k)            ▷ randomize k features
    end for
    return ⊥                             ▷ failed to synthesize
end procedure
```

□ Prior knowledge:

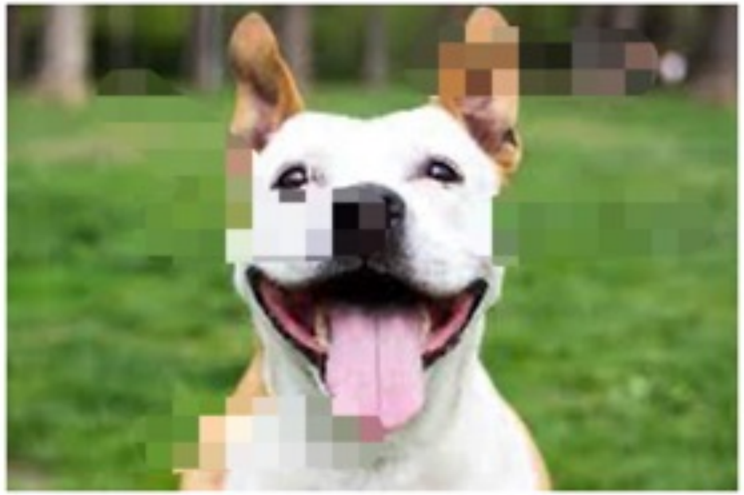
- **The marginal distribution w.r.t. each class**

Phase 1: sample according to the statistics

Shokri, Reza, et al. "Membership inference attacks against machine learning models." S&P, 2017.

MIA：The First Well-known Work

Noisy real data: real but noisy

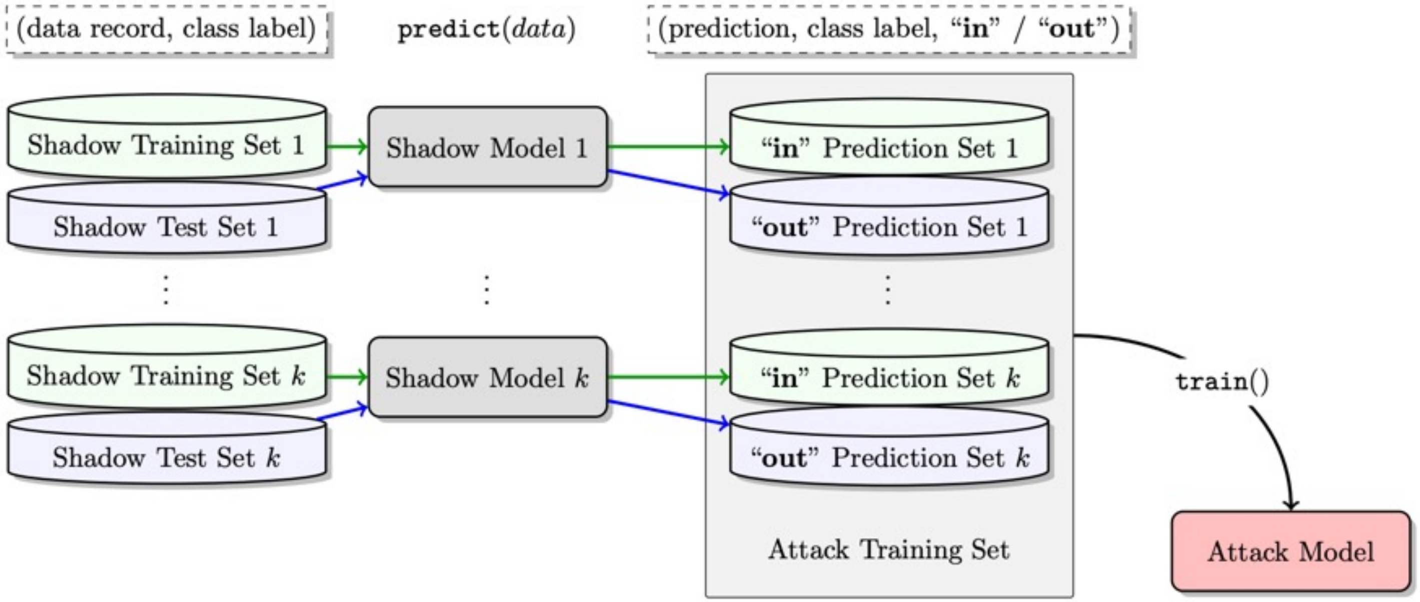


- Very similar to the real dataset
- But with a few features (10% or 20%) are randomly reset

Shokri, Reza, et al. "Membership inference attacks against machine learning models." S&P, 2017.

MIA：The First Well-known Work

- **Finally: training the inference model**

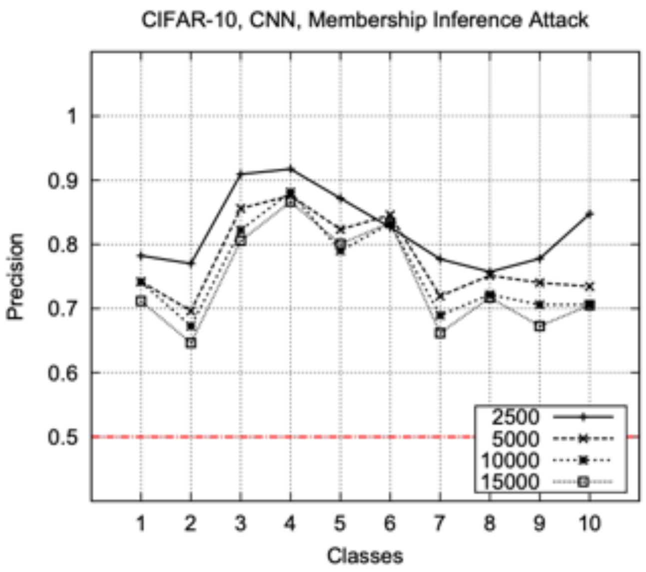


- "in": in the training set
- "out": : in the test set
- Train the inference model with dataset:
  (prob1, "in"),
  (prob2, "in"),
  (prob3, "out")
  (prob4, "out")

Shokri, Reza, et al. "Membership inference attacks against machine learning models." S&P, 2017.

MIA：The First Well-known Work

□ Evaluation



Dataset	*Training Accuracy*	*Testing Accuracy*	*Attack Precision*
Adult	0.848	0.842	0.503
MNIST	0.984	0.928	0.517
Location	1.000	0.673	0.678
Purchase (2)	0.999	0.984	0.505
Purchase (10)	0.999	0.866	0.550
Purchase (20)	1.000	0.781	0.590
Purchase (50)	1.000	0.693	0.860
Purchase (100)	0.999	0.659	0.935
TX hospital stays	0.668	0.517	0.657

数据集：CIFAR-10、CIFAR-100、Purchases、Locations、Texas hospital stays、MNIST、UCI Adult (Census Income).

Shokri, Reza, et al. "Membership inference attacks against machine learning models." S&P, 2017.

White-box MIA

◻ White-box vs Black-box

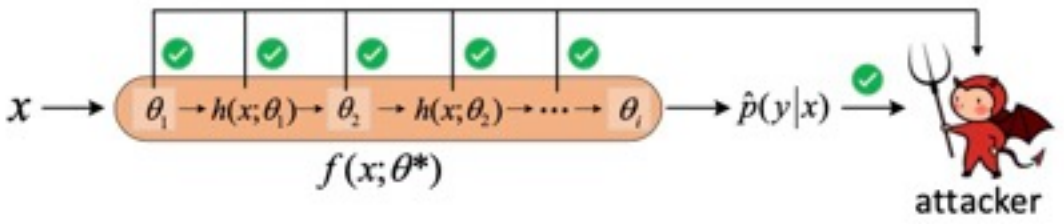


Fig. 2. Overview of white-box membership inference attacks.

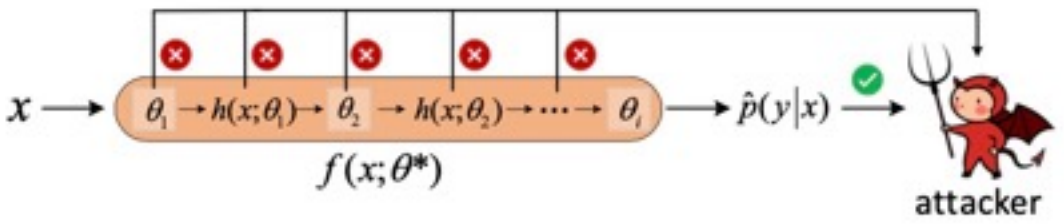
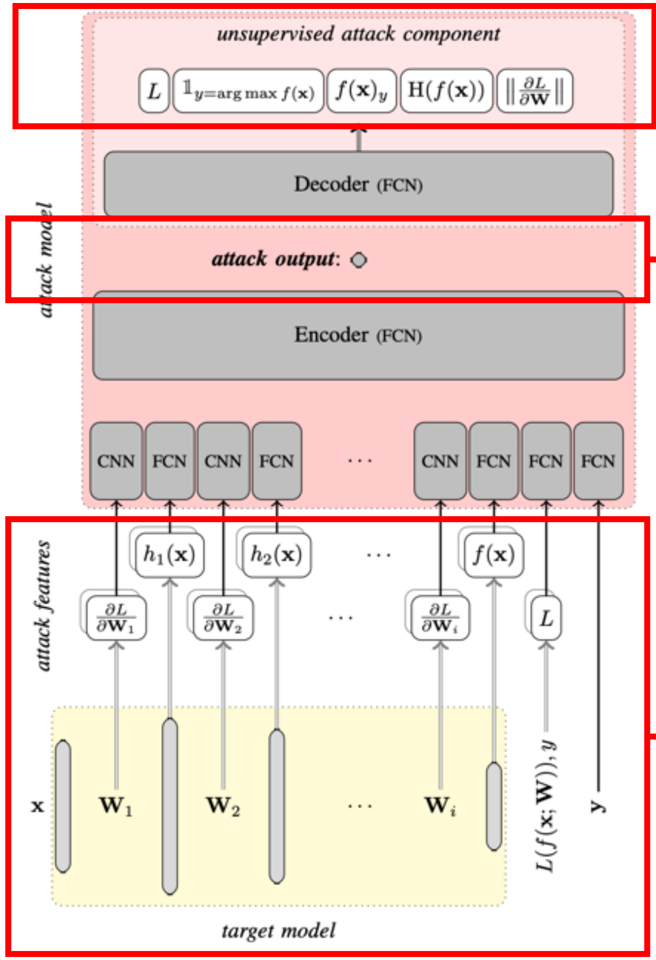


Fig. 3. Overview of black-box membership inference attacks.

Nasr et al. “Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning.” *S&P*, 2019. Hu, Hongsheng, et al. "Membership inference attacks on machine learning: A survey." *ACM Computing Surveys (CSUR)* 54.11s (2022): 1-37.

White-box MIA



- 无监督设置下的重构损失
- 推理结果
- 抽取特征：概率、中间层激活、梯度

Nasr et al. "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning." *S&P*, 2019.

Limitations of MIA

- Constructing shadow models
- Assuming access to some data or prior knowledge
- Overfitting is a must
- Limited to classification models
- Limited to small models

Addressing Limitations of MIA

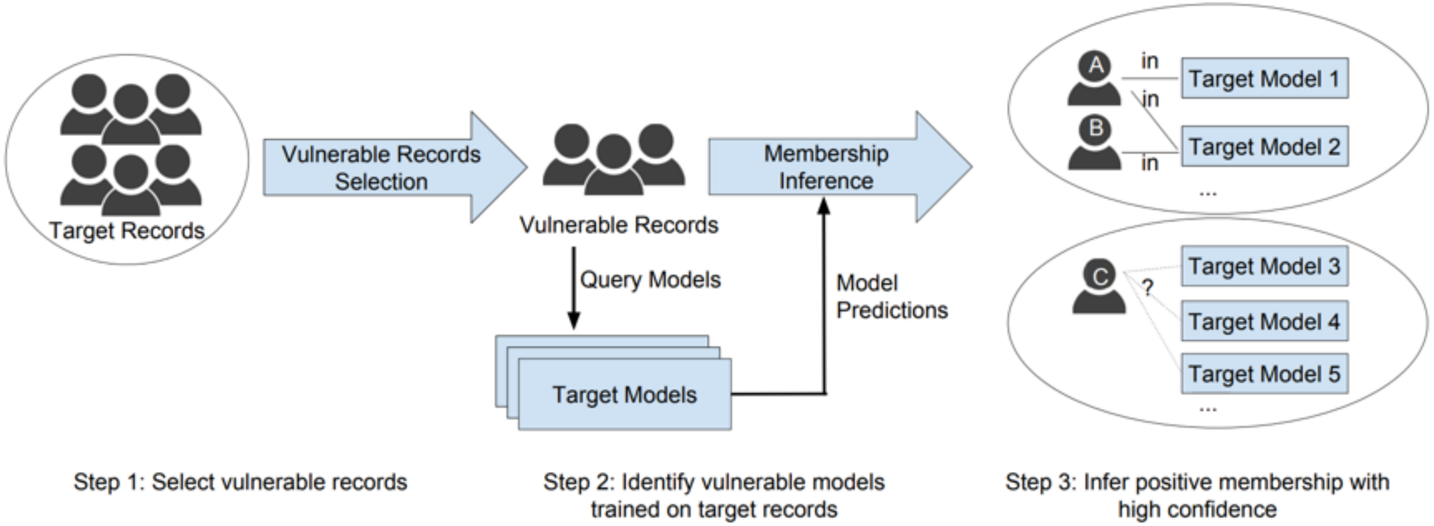
◻ Model and Data Independent MIA

Adversary type	Shadow model design		Target model's
	No. shadow models	Target model structure	training data distribution
Shokri et al. [38]	multiple	✓	✓
Our adversary 1	1	-	✓
Our adversary 2	1	-	-
Our adversary 3	-	-	-

Salem et al. "ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models." *NDSS*, 2019.

Addressing Limitations of MIA

- **Attacking non-overfitting DNNs**
- **Focusing on minimizing false positives**



目标问题：样本A/B在哪个模型的训练数据里？

Long, Yunhui, et al. "A pragmatic approach to membership inferences on machine learning models." *EuroS&P*, 2020.

Addressing Limitations of MIA

- **More practical white-box threat model**
- **The adversary only knows the model but not the data distribution**

(a)

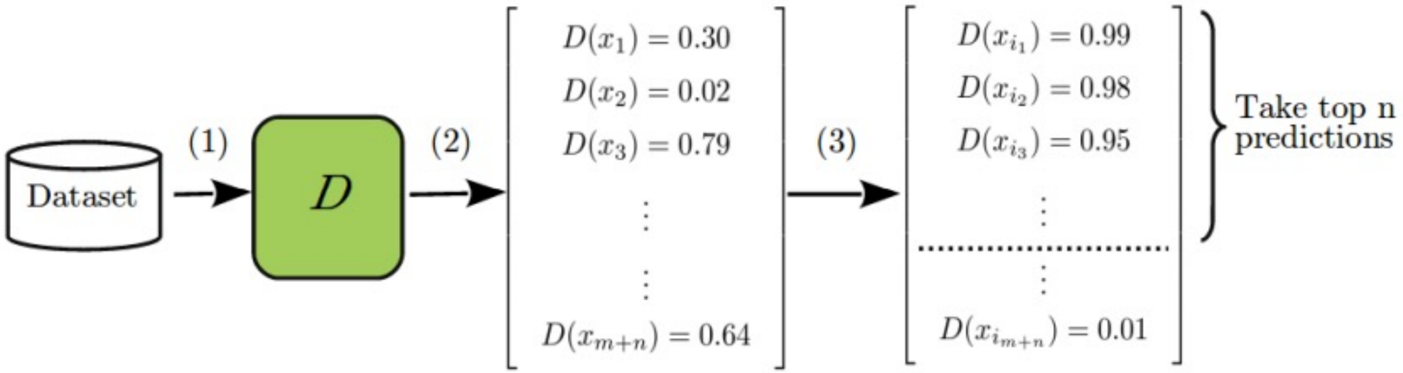
(b) (c)

利用诡异的独家记忆进行
成员推理

Leino & Fredrikson. "Stolen Memories: Leveraging Model Memorization for Calibrated White-Box Membership Inference." *USENIX Security*, 2020.

Addressing Limitations of MIA

◻ Extension to generative models



充分利用判别器的判别能力：高置信度的大概率来自原始训练数据集

Hayes, Jamie, et al. "Logan: Membership inference attacks against generative models." *arXiv preprint arXiv:1705.07663* (2017).

Metric-guided MIA

☐ Metric based Anomaly detection

- **预测正确性：** $\mathcal{M}(\hat{\boldsymbol{p}}(y|\boldsymbol{x}), y) = \mathbb{1}[\arg\max \hat{\boldsymbol{p}}(y|\boldsymbol{x}) = y]$ 预测正确的就是成员
- **预测损失：** $\mathcal{M}(\hat{\boldsymbol{p}}(y|\boldsymbol{x}), y) = \mathbb{1}[\mathcal{L}(\hat{\boldsymbol{p}}(y|\boldsymbol{x}); y) \leq \tau]$ 高于训练样本平均损失的是成员
- **预测置信度：** $\mathcal{M}(\hat{\boldsymbol{p}}(y|\boldsymbol{x})) = \mathbb{1}[\max \hat{\boldsymbol{p}}(y|\boldsymbol{x}) \geq \tau]$ 有概率接近1的是成员
- **预测熵：** $\mathcal{M}(\hat{\boldsymbol{p}}(y|\boldsymbol{x})) = \mathbb{1}[\mathrm{H}(\hat{\boldsymbol{p}}(y|\boldsymbol{x})) \leq \tau] = \mathbb{1}[-\sum_i \boldsymbol{p}_i \log(\boldsymbol{p}_i) \leq \tau]$ 低概率熵的是成员
- **修正预测熵：** $\mathrm{MH}(\hat{\boldsymbol{p}}(y|\boldsymbol{x}), y) = -(1-\boldsymbol{p}_y)\log(\boldsymbol{p}_y) - \sum_{i \neq y} \boldsymbol{p}_i \log(1-\boldsymbol{p}_i)$ 不同类别区别考虑

Yeom, Samuel, et al. “Privacy risk in machine learning: Analyzing the connection to overfitting.” *CSF,* 2018. Salem et al. "ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models." *NDSS*, 2019.

A Summary of Existing MIAs

☐ Datasets

- **Image**:
  - CIFAR-10, CIFAR-100, MNIST, Fashion-MNIST, Yale Face, ChestX-ray8, SVHN, CelebA, ImageNet
- **Tabulate**:
  - Adult, Foursquare, Purchase-100, Texas100, Location, etc.
- **Audio:**
  - LibriSpeech, TIMIT, TED
- **Text:**
  - Weibo, Tweet EmoInt, SATED, Dislogs, Reddit comments, Cora, Pubmed, Citesser

Hu, Hongsheng, et al. "Membership inference attacks on machine learning: A survey." *ACM Computing Surveys, 2022*.

A Summary of Existing MIAs

- **Target models:**
  - On **image**:
    - Multi-layer CNN + 1 or 2 FC (> 5 papers used 2-4 layers CNN)
    - Alexnet, ResNet18, ResNet50, VGG16, VGG19, DenseNet121, Efficient-netv2, EfficientNetB0
    - GAN: InfoGAN, PGGAN, WGANGP, DCGAN, MEDGAN, and VAEGAN
  - On **tabulate data**:
    - FC only models
  - On **text**:
    - Multi-layer CNN, multi-layer RNN/LSTM, transformers (e.g., BERT, GPT-2)
  - On **audio**:
    - Hybrid system: HMM-DNN model
    - End-to-end: Multi-layer LSTM/ RNN/GRU
- **MLaaS (Online):**
  - Google Prediction API, Amazon ML

A Summary of Existing MIAs

- **Target models:**
  - On **image**:
    - Multi-layer CNN + 1 or 2 FC (> 5 papers used 2-4 layers CNN)
    - Alexnet, ResNet18, ResNet50, VGG16, VGG19, DenseNet121, Efficient-netv2, EfficientNetB0
    - GAN: InfoGAN, PGGAN, WGANGP, DCGAN, MEDGAN, and VAEGAN
  - On **tabulate data**:
    - FC only models
  - On **text**:
    - Multi-layer CNN, multi-layer RNN/LSTM, transformers (e.g., BERT, GPT-2)
  - On **audio**:
    - Hybrid system: HMM-DNN model
    - End-to-end: Multi-layer LSTM/ RNN/GRU
- **MLaaS (Online):**
  - Google Prediction API, Amazon ML

- Membership Inference Attack
- **Differential Privacy**

Differential Privacy

□ Finite Difference and Derivative

$$f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$$

h tends to be small (zero)

通过函数在某一点随微小扰动的变化可以估计在这一点的梯度

如果对数据集进行微小扰动呢？

Differential Privacy

□ Finite Difference -> Differential Privacy

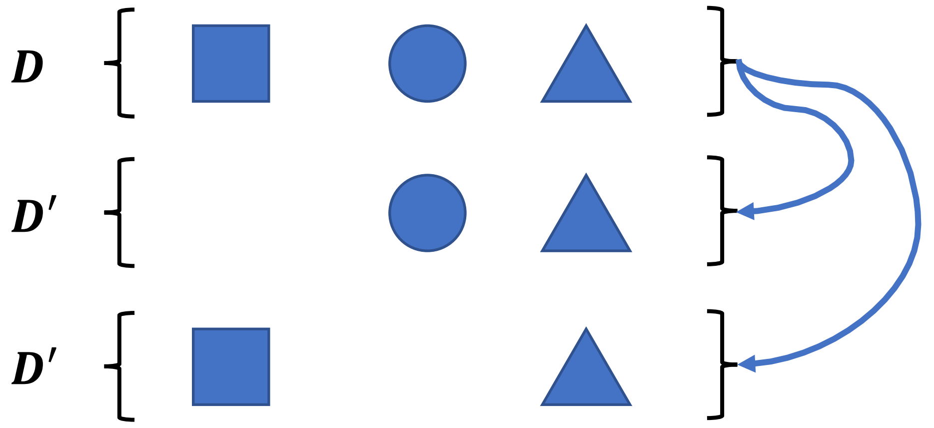
$$f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$$

$\boldsymbol{f(x)}$	函数	$\longrightarrow$	算法/机制 $\mathcal{M}$
$\boldsymbol{a}$	输入值	$\longrightarrow$	数据集 D

数据集的微小变化会导致多大的算法输出变化？

Differential Privacy

- 邻接数据集 D、 D'



数据集的微小变化会导致多大的算法输出变化？

Differential Privacy

定义 5.1. 差分隐私： 对于一个随机算法 M，P_m 为算法 M 所有可能输出的集合，若算法 M 满足 $(\epsilon, \delta) - DP$，当且仅当相邻数据集 D, D' 对 M 的所有可能输出子集 $S_m \in P_m$，满足不等式 [Dwork et al., 2006a]：

$$P_r[M(D) \in S_m] \leq e^{\epsilon} P_r[M(D') \in S_m] + \delta$$

ϵ**：** 隐私预算（Privacy Budget）,越小隐私越好
δ：打破$(\varepsilon, \delta) - DP$的可能性

Dwork, Cynthia. "Differential privacy: A survey of results." *ICTAMC*, Heidelberg, 2008.

Properties of DP

性质 5.1. 顺序合成： 给定 K 个随机算法 $M_i(i=1,\cdots,K)$，分别满足 ϵ_i-DP，如果将他们作用在同一个数据集上，则满足 $\sum_{i=1}^{K}\epsilon_i-DP$。

性质 5.2. 平行合成： 将数据集 D 分割成 K 个不相交的子集 $\{D_1,D_2,\cdots,D_K\}$，在每个子集上分别作用满足 ϵ_i-DP 的随机算法 M_i，则数据集 D 整体满足 $(\max\{\epsilon_1,\cdots,\epsilon_K\})-DP$。

性质 5.3. 交换不变性： 给定任意算法 M_1 满足 $\epsilon-DP$，数据集 D，对于任意算法 M_2（M_2 不一定满足差分隐私），则 $M_2(M_1(D))$ 满足 $\epsilon-DP$。

性质 5.4. 中凸性： 给定满足 $\epsilon-DP$ 的随机算法 M_1 和 M_2，对于任意的概率 $P\in[0,1]$，用 A_P 表示一种选择机制，以 P 的概率选择算法 M_1，以 $1-P$ 的概率选择算法 M_2，则 A_p 机制满足 $\epsilon-DP$。

McSherry, Frank D. “Privacy integrated queries: an extensible platform for privacy-preserving data analysis.” *ACM SIGMOD*, 2009.

How to Obtain a Differentially Private Model

思考：

？？

Measuring Sensitivity

定义 5.2. 全局敏感度(Global Sensitivity): 给定查询函数 $f: D \to R$，D 为数据集，R 为查询结果。在任意一对相邻数据集 D, D' 上，全局敏感度定义为:

$$S(f) = \max_{D,D'} \|f(D) - f(D')\|_1$$

定义 5.3. 局部敏感度 (Local Sensitivity): 给定查询函数 $f: D \to R$，D 为数据集，R 为查询结果。在一给定的数据集 D 和它相邻的任意数据集 D' 上，局部敏感度定义为:

$$LS(f) = \max_{D'} \|f(D) - f(D')\|_1$$

Nissim and Adam. “Smooth sensitivity and sampling in private data analysis.” STOC, 2007.

Noise Models

□ 几种噪声添加机制

- 拉普拉斯机制 (Laplacian)

$$M(D) = f(D) + Lap(\frac{S(f)}{\epsilon})$$

$Lap(\frac{S(f)}{\epsilon})$ 表示位置参数为 0，尺度参数为 $\frac{S(f)}{\epsilon}$ 的拉普拉斯分布

- 高斯机制 (Guassian)

$$M(D) = f(D) + \mathcal{N}(\delta^2)$$
$$s.t.\ \delta^2 = \frac{2S(f)^2 \log(1.25/\delta)}{\epsilon^2}$$

$\mathcal{N}(\delta^2)$ 表示中心为 0，方差为 δ^2 的高斯分布

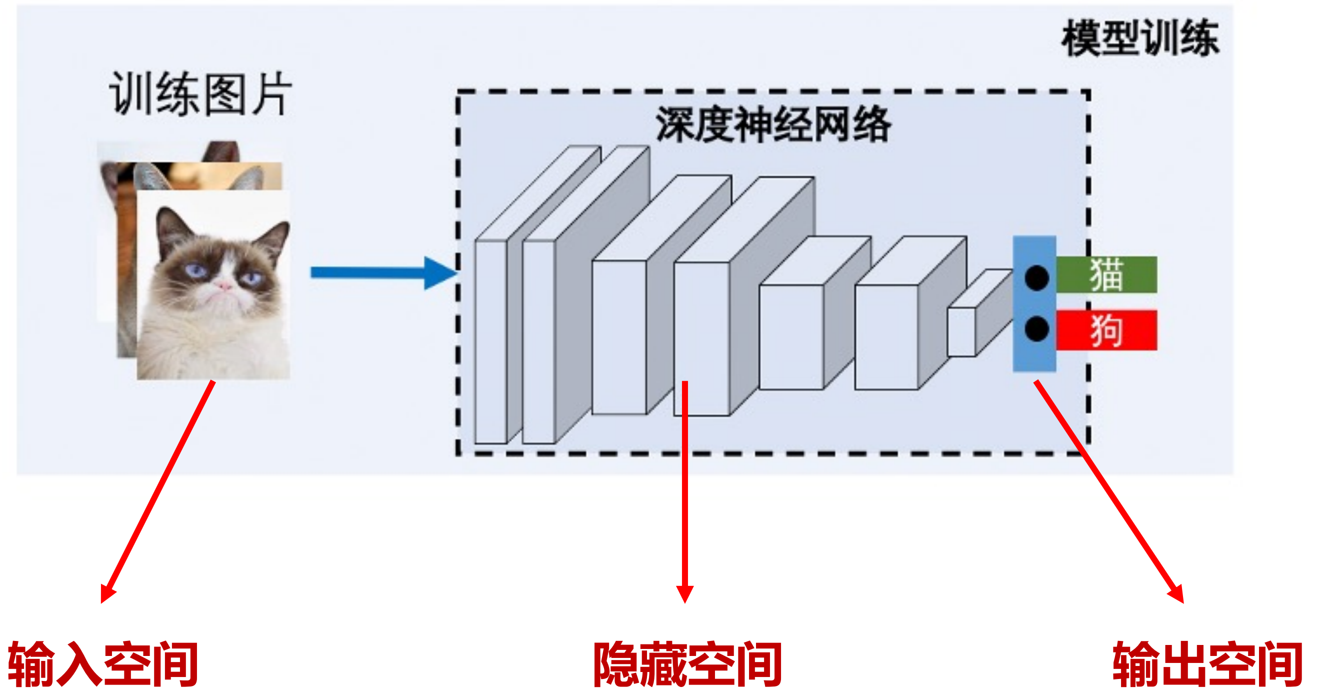
- 指数机制：离散 -> 概率；确定 -> 不确定

$$M(D) = \mathrm{return}(R_i \propto exp(\frac{\epsilon q(D, R_i)}{2S(q)}))$$

$$Pr(R_i) = \frac{\exp(\frac{\epsilon q(D,R_i)}{2S(q)})}{\sum_{j=1}^{N} \exp(\frac{\epsilon q(D,R_j)}{2S(q)})}$$

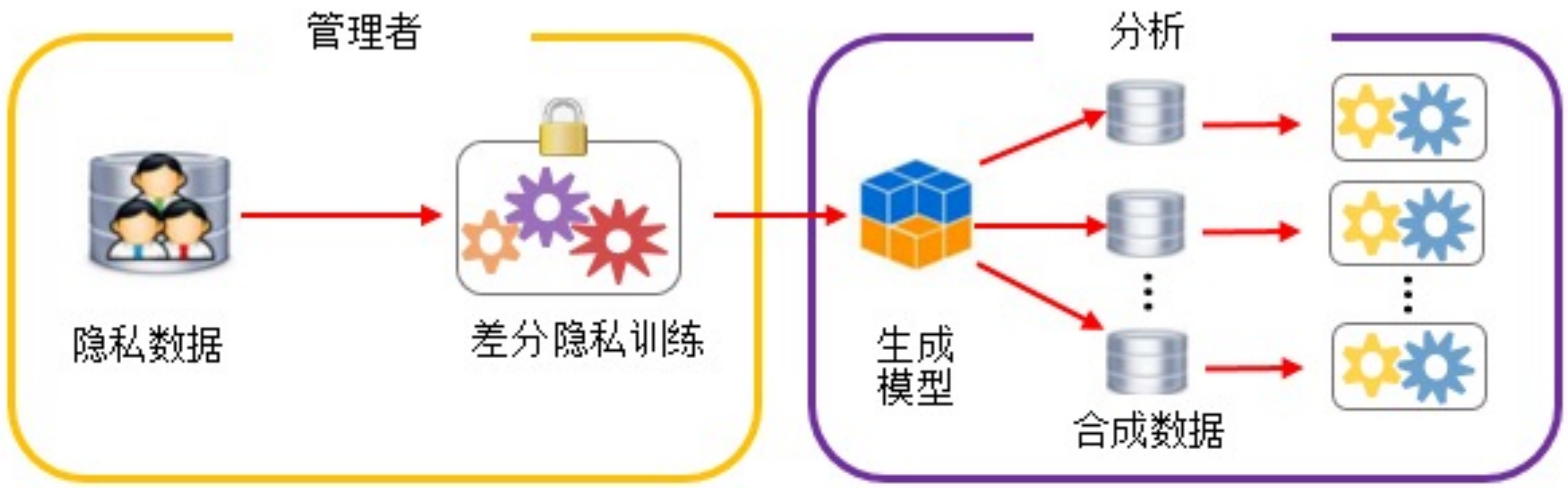
DP + Deep Learning

- 问题：**在哪里添加噪声？**



输入空间DP

□ 差分隐私预处理训练数据

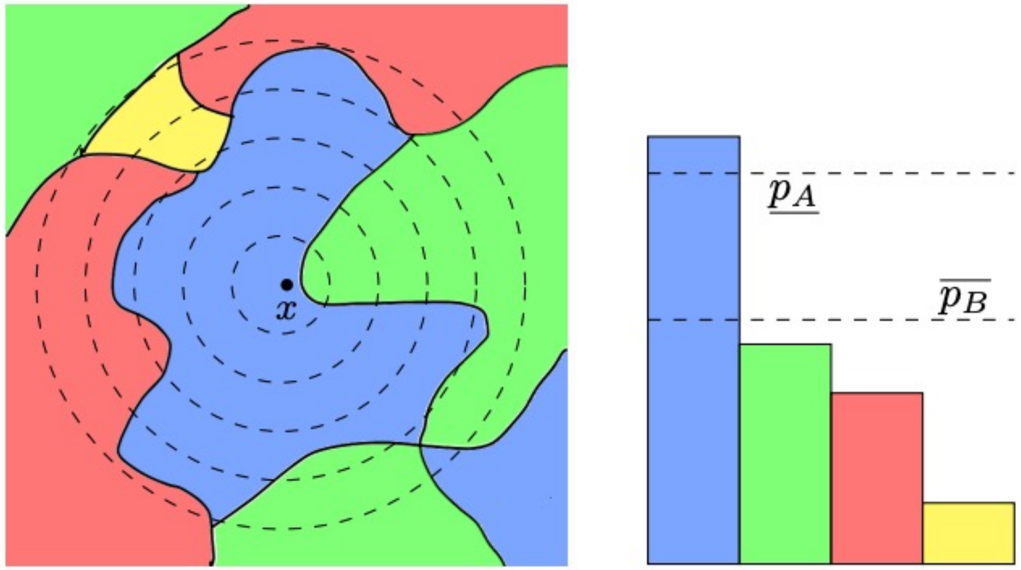


dp-GAN pipeline

Zhang et al. “Differentially private releasing via deep generative model (technical report).” *arXiv:1801.01594* (2018).

输入空间DP

☐ 随机平滑 Randomized Smoothing



用随机噪声填充输入空间，
得到对抗鲁棒性边界

随机平滑：可验证对抗防御

Cohen, Jeremy, Elan Rosenfeld, and Zico Kolter. "Certified adversarial robustness via randomized smoothing." *ICML*, 2019.

隐藏空间DP

□ 差分隐私平滑模型参数：DP-SGD算法

Algorithm 5.1 Differentially Private SGD（DP-SGD）[Abadi et al., 2016]

输入: 样本 $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$，损失函数 $\mathcal{L}(\theta) = \frac{1}{n}\sum_i \mathcal{L}(\theta, \boldsymbol{x}_i)$。超参数：学习率 η_t，噪声参数 σ，分组大小 L，梯度约束范数 C

输出: θ_T，同时利用隐私统计方法计算总体的隐私损失 (ϵ, δ)

1: 随机初始化模型 θ_0
2: **for** $t \in [T]$ **do**
3: 以概率 L/n 随机采取一组样本 L_t
4: **计算梯度:** 对每一个样本 $i \in L_t$，计算 $g_t(\boldsymbol{x}_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, \boldsymbol{x}_i)$
5: **裁剪梯度:** $\bar{g}_t(\boldsymbol{x}_i) \leftarrow g_t(\boldsymbol{x}_i)/\max(1, \frac{\|g_t(\boldsymbol{x}_i)\|_2}{C})$
6: **噪声添加:** $\tilde{g}_t \leftarrow \frac{1}{L}(\sum_i \bar{g}_t(\boldsymbol{x}_i) + \mathcal{N}(0, \sigma^2 C^2 I))$
7: **梯度下降:** $\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{g}_t$

Abadi, Martin, et al. “Deep learning with differential privacy.” *CCS, 2016.*

输出空间DP

□ 差分隐私扰动目标函数：多项式目标函数

- 回归模型 $$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \sum_{i=1}^{n} \mathcal{L}(t_i, \boldsymbol{w})$$

- 根据Stone-Weierstrass 理论：***任意连续可微的函数可表示为：***

$$\mathcal{L}_D(\boldsymbol{w}) = \sum_{j=0}^{J} \sum_{\phi \in \Phi_j} \lambda_{\phi t_i} \sum_{t_i \in D} \phi(\boldsymbol{w})$$

Zhang, Jun, et al. “Functional mechanism: regression analysis under differential privacy.” *arXiv:1208.0219,* 2012.
Rudin, Walter. *Principles of mathematical analysis*. Vol. 3. New York: McGraw-hill, 1976.

输出空间DP

☐ 差分隐私扰动目标函数：多项式目标函数

Algorithm 5.2 函数机制（Functional Mechanism）[Zhang et al., 2012]

输入： 数据集 D，目标函数 $\mathcal{L}_D(\boldsymbol{w})$，隐私预算 ϵ

输出： 差分隐私扰动后的模型参数 $\bar{\boldsymbol{w}}$

1: 令 $\Delta = 2\max_t \sum_{j=1}^{J} \sum_{\phi\in\Phi_j \|\lambda_{\phi t}\|_1}$

2: **for** $0 \leq j \leq J$ **do**

3:     **for** $\phi \in \Phi_j$ **do**

4:         令 $\lambda_\phi = \sum_{t_i\in D} \lambda_{\phi t_i} + \mathrm{Laplace}(\frac{\Delta}{\epsilon})$

5: 令 $\bar{\mathcal{L}}_D(\boldsymbol{w}) = \sum_{j=1}^{J} \sum_{\phi\in\Phi_j} \lambda_\phi \phi(\boldsymbol{w})$

6: 计算 $\bar{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \bar{\mathcal{L}}_D(\boldsymbol{w})$

7: 返回 $\bar{\boldsymbol{w}}$

Zhang, Jun, et al. “Functional mechanism: regression analysis under differential privacy.” *arXiv:1208.0219,* 2012.

输出空间DP

☐ 差分隐私扰动目标函数：cross-entropy

$$\widetilde{f}_D(\omega) = \sum_{i=1}^{|D|} \sum_{l=1}^{m} \sum_{R=0}^{\infty} \frac{f_l^{(R)}(z_l)}{R!} \left(g_l(t_i, \omega) - z_l\right)^R$$

泰勒展开 Taylor Expansion

Phan, et al. “Differential privacy preservation for deep auto-encoders: an application of human behavior prediction.” *AAAI,* 2016.

Remaining Challenges

- **Attack:**
  - **Better Performance Metrics for MIA**
  - **Attacking large-scale pretrained models**
- **Defense:**
  - **How to achieve both accuracy and privacy**
  - **How to detect potential MIAs on the fly**

C U Next Week!

Course page:

https://trustworthymachinelearning.github.io/

Textbook:

下载链接:

Email: xingjunma@fudan.edu.cn
Personal page: www.xingjunma.com
Office: 江湾校区交叉二号楼D5025