人工智能系统的隐私研究

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

姓　名：　孟国柱

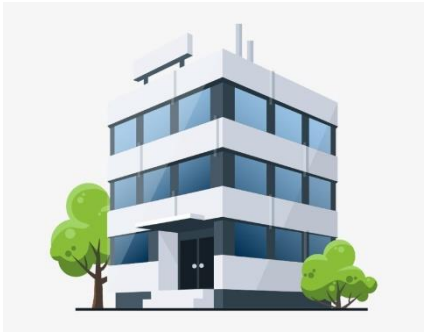
时　间：2021年6月6号

我的安全观

安全观：

我的安全观

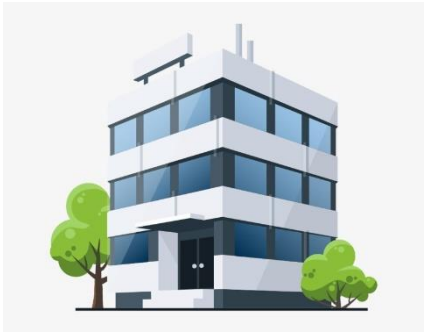
安全观：



敏锐洞察力、心思缜密、强大执行力

我的安全观

安全观：

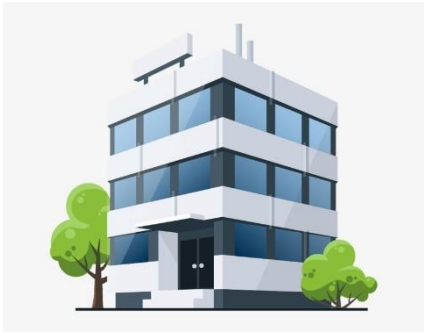


敏锐洞察力、心思缜密、强大执行力

Hacking for Good “向善的心”

我的安全观

安全观：



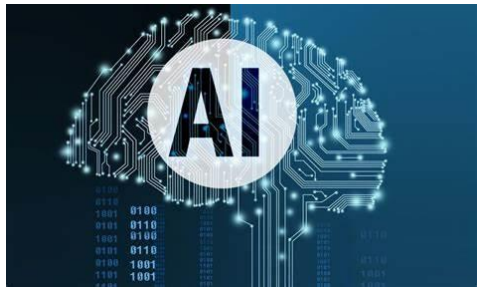
敏锐洞察力、心思缜密、强大执行力

Hacking for Good “向善的心”

我：

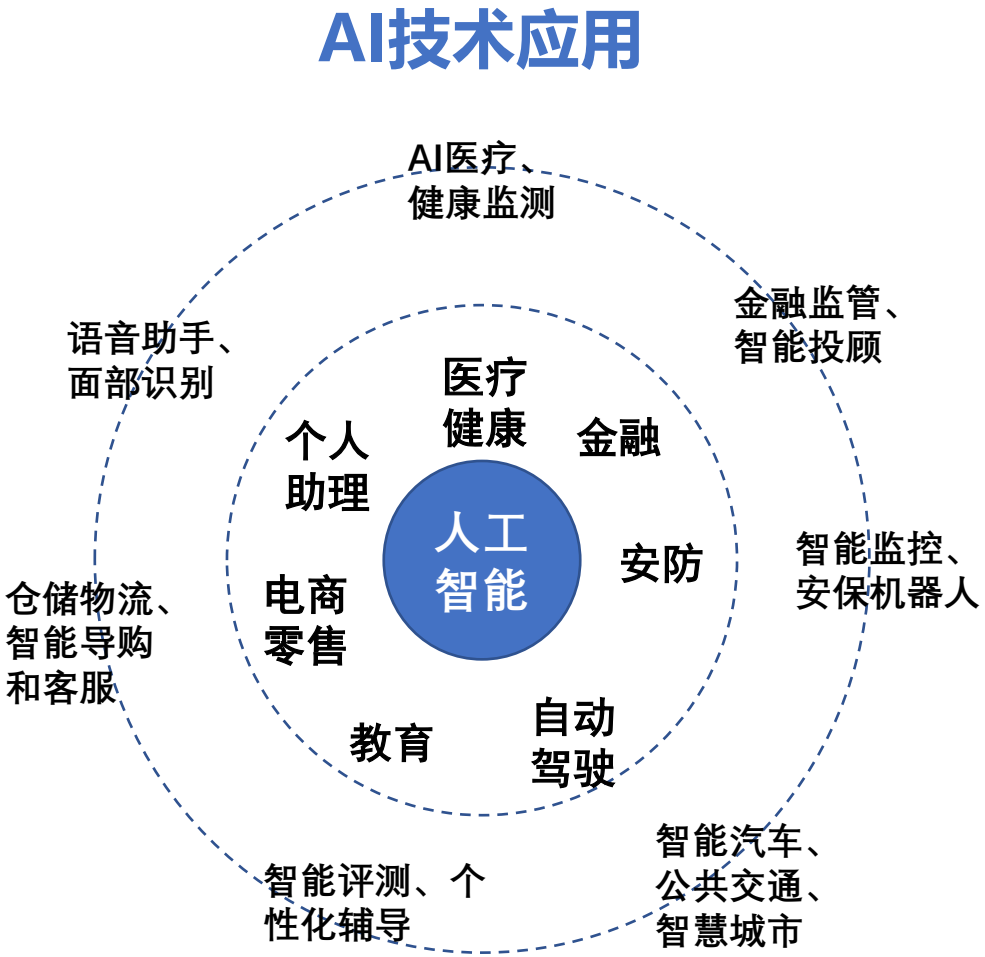
Mobile Security

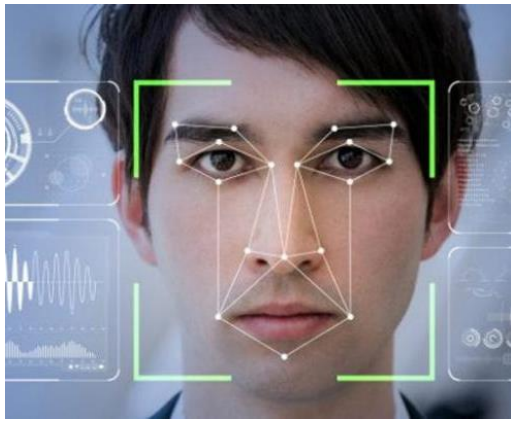
Big Data & Vulnerability

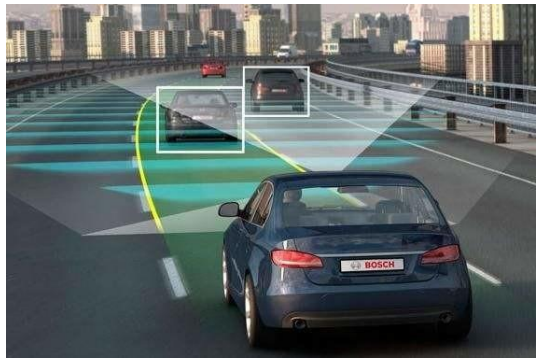


AI Security

人工智能系统的应用


AI技术应用
AI医疗、
健康监测
金融监管、
智能投顾
语音助手、
面部识别
医疗
健康
个人
助理
金融
人工
智能
安防
智能监控、
安保机器人
仓储物流、
智能导购
和客服
电商
零售
教育
自动
驾驶
智能评测、个
性化辅导
智能汽车、
公共交通、
智慧城市




BOSCH

人工智能系统的安全问题

人工智能模型安全问题

①模型安全威胁

②数据隐私泄露

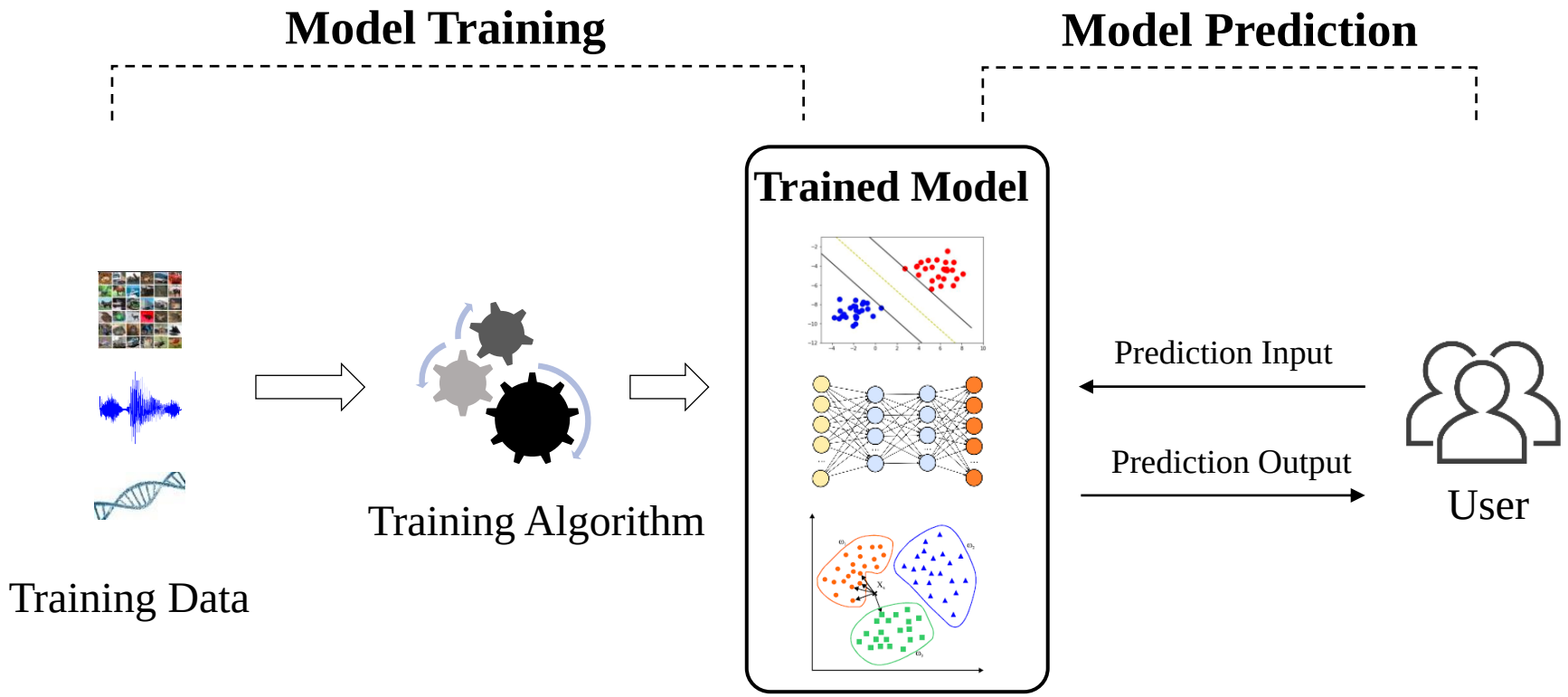
基于传感器的语音攻击

特斯拉自动驾驶，系统‘误判’导致事故

通过模型逆向方法导致的数据隐私泄露。

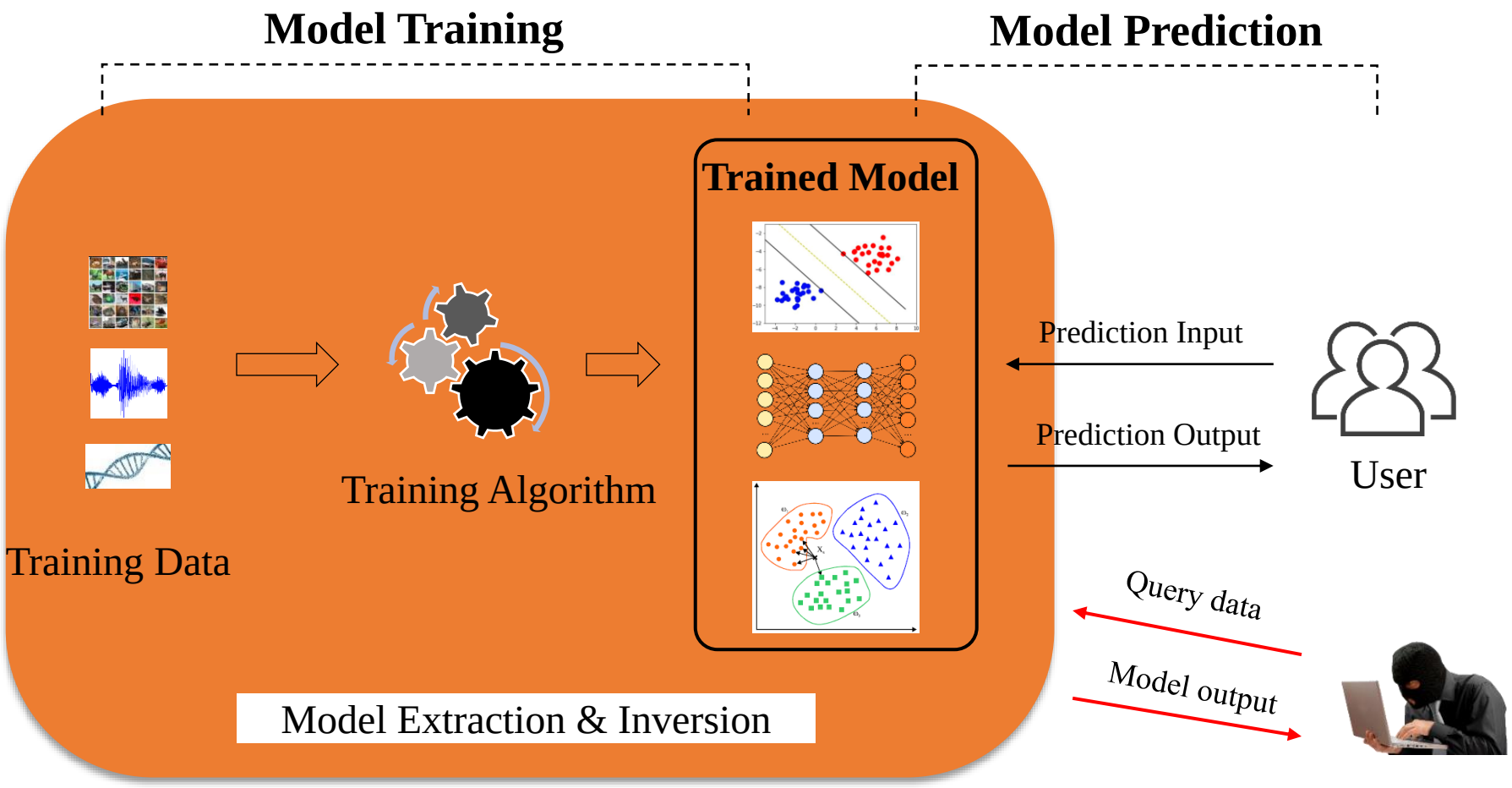
Figure 1: An image recovered using a new model inversion attack (left) and a training set image of the victim (right). The attacker is given only the person's name and access to a facial recognition system that returns a class confidence score.

人工智能系统的隐私问题



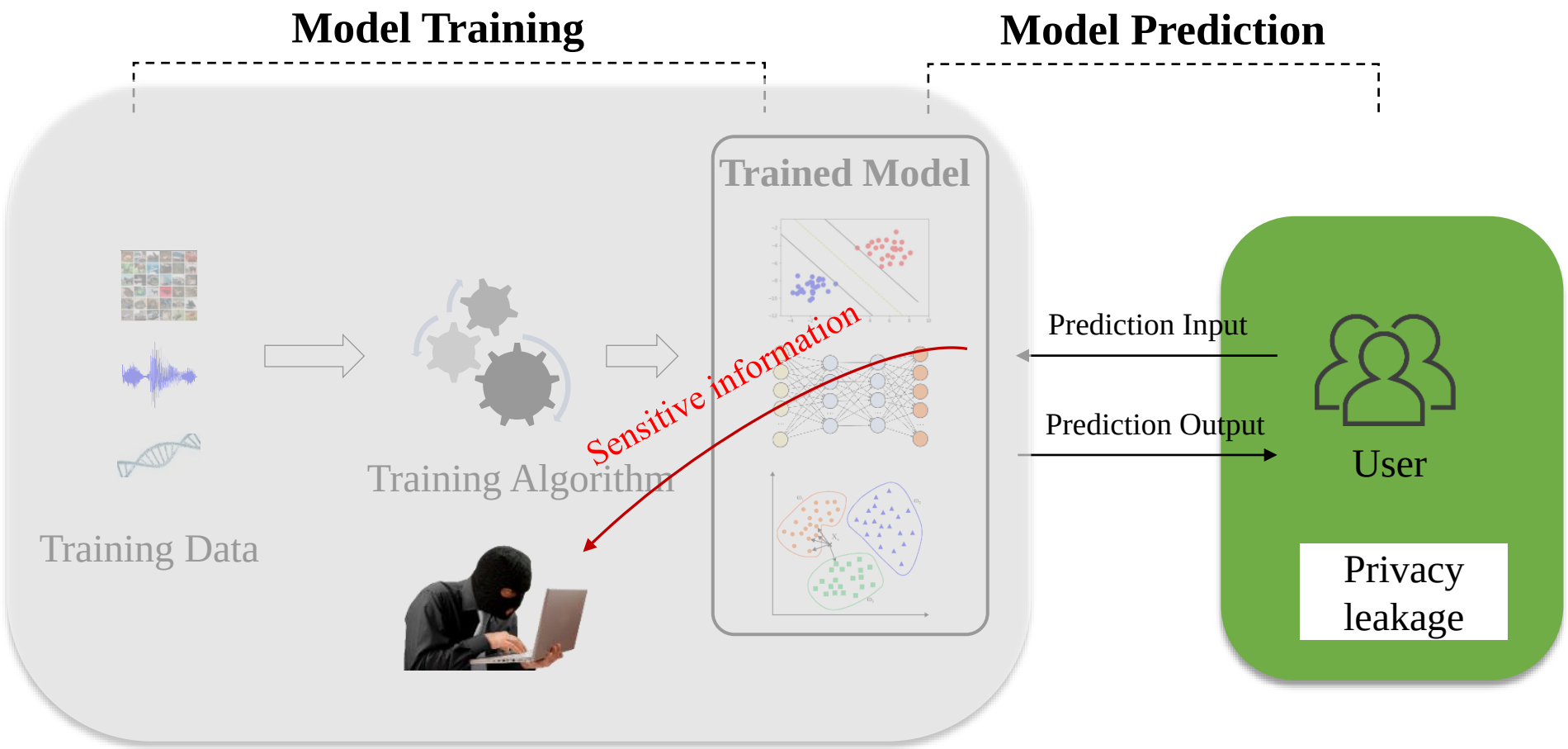
Yingzhe He, Guozhu Meng, Kai Chen, Xingbo Hu, Jinwen He. Towards Security Threats of Deep Learning Systems: A Survey, ***IEEE Transactions on Software Engineering 2020***

人工智能系统的隐私问题



Yingzhe He, Guozhu Meng, Kai Chen, Xingbo Hu, Jinwen He. Towards Security Threats of Deep Learning Systems: A Survey, ***IEEE Transactions on Software Engineering 2020***

人工智能系统的隐私问题



Yingzhe He, Guozhu Meng, Kai Chen, Xingbo Hu, Jinwen He. Towards Security Threats of Deep Learning Systems: A Survey, IEEE Transactions on Software Engineering, ***IEEE TSE 2020***

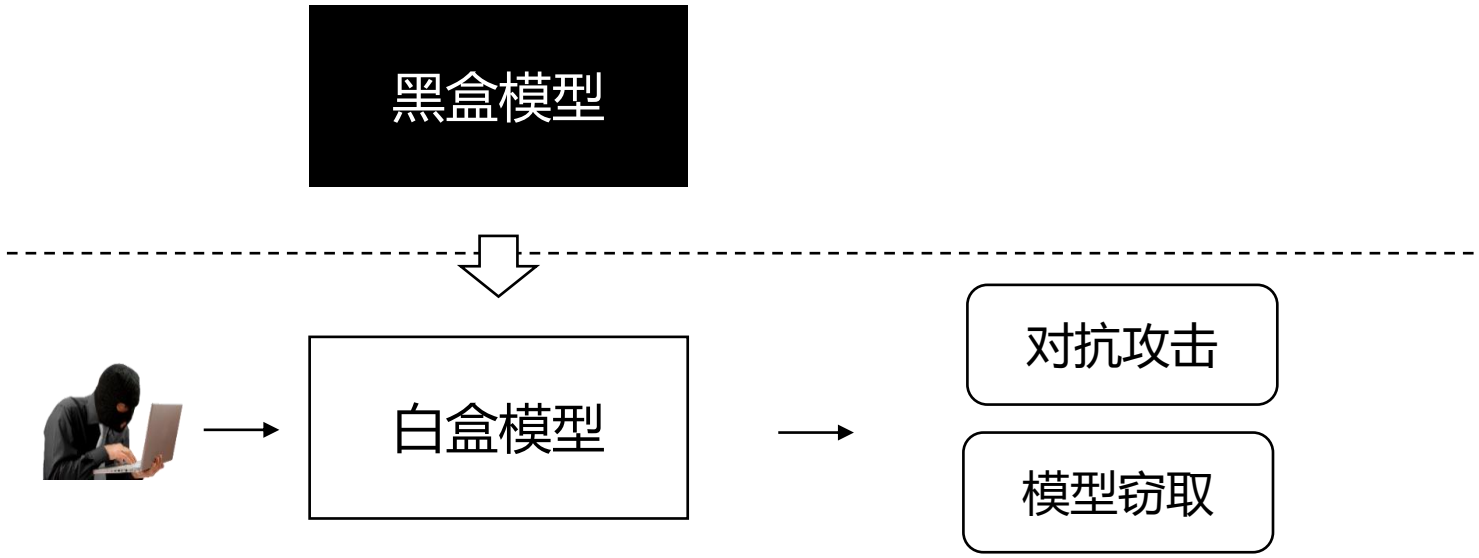
人工智能系统的隐私问题

- 黑盒攻击
  - 不是完全黑盒
  - 交互限制
  - 防御措施
- 隐私保护
  - 用户隐私保护（GDPR）
  - 联邦学习安全
  - 数据、模型水印
- 系统的安全实现
  - 系统漏洞（溢出）
  - 部署（Quantization）
- 模型可解释性
  - 可推理
  - 可追溯
  - 可理解
- 安全防护
  - 差分隐私/多方安全计算
  - 后门模式检测
  - 模型加固
- 人工智能伦理问题
  - 性别、人种歧视
  - 安全事故追责

人工智能系统的隐私问题

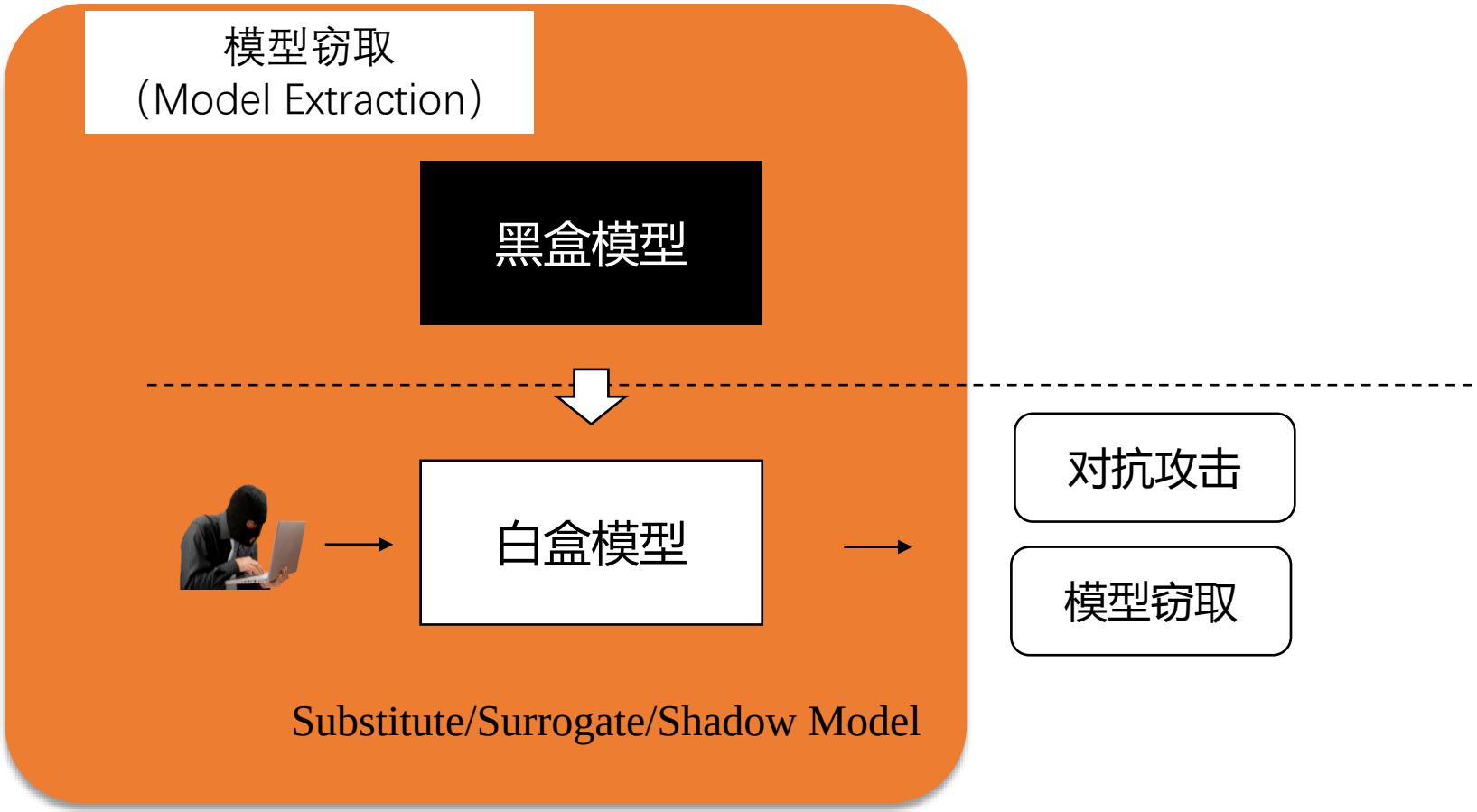
- 黑盒攻击
  - 不是完全黑盒
  - 交互限制
  - 防御措施
- 隐私保护
  - 用户隐私保护（GDPR）
  - 联邦学习安全
  - 数据、模型水印
- 系统的安全实现
  - 系统漏洞（溢出）
  - 部署（Quantization）
- 模型可解释性
  - 可推理
  - 可追溯
  - 可理解
- 安全防护
  - 差分隐私/多方安全计算
  - 后门模式检测
  - 模型加固
- 人工智能伦理问题
  - 性别、人种歧视
  - 安全事故追责

人工智能系统的隐私问题

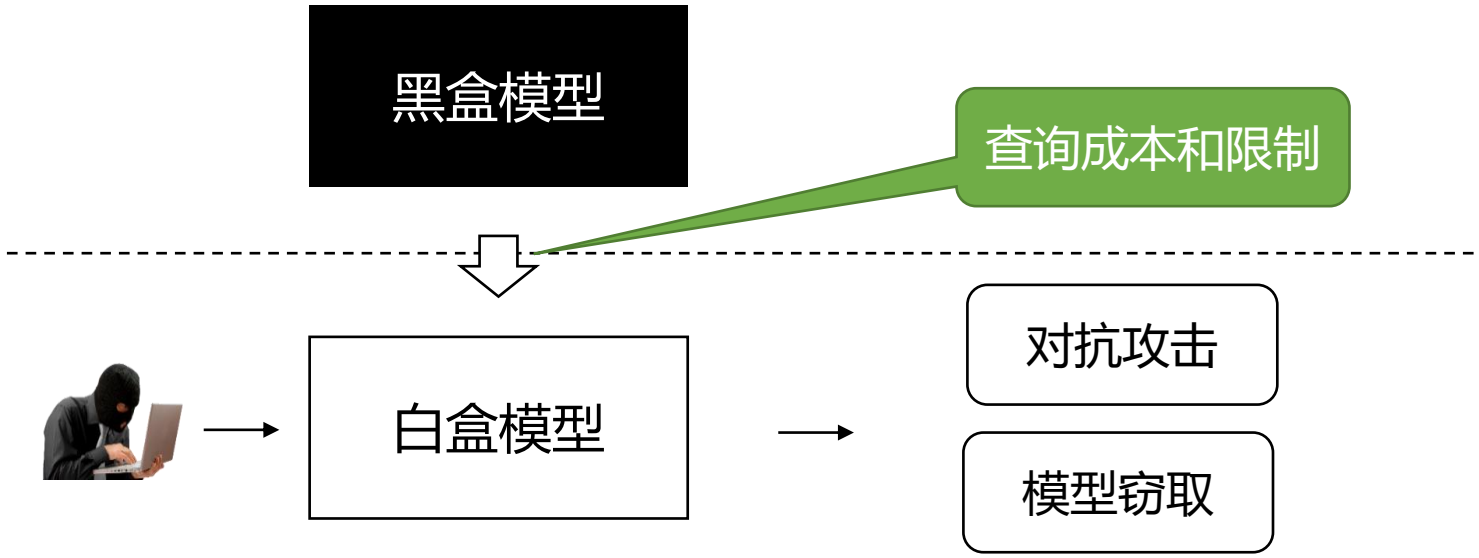


Substitute/Surrogate/Shadow Model

人工智能系统的隐私问题



人工智能系统的隐私问题



Substitute/Surrogate/Shadow Model

人工智能系统的隐私问题

Q1：训练数据是否有信息冗余?

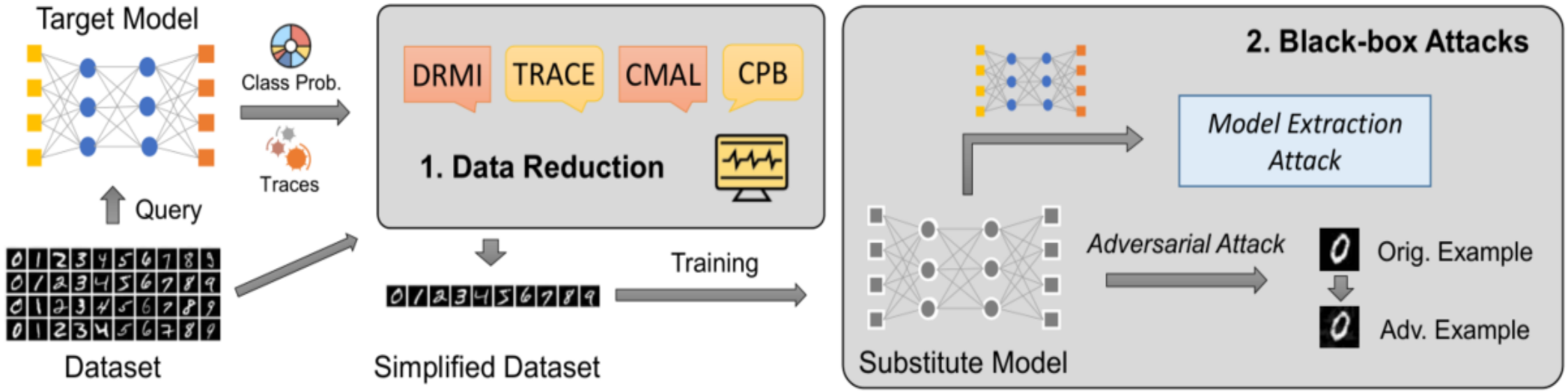
重复的数据并不能增进模型性能

Q2：查询数据是否均有益?

– *Distribution Drift，OOD使替代模型偏差较大*

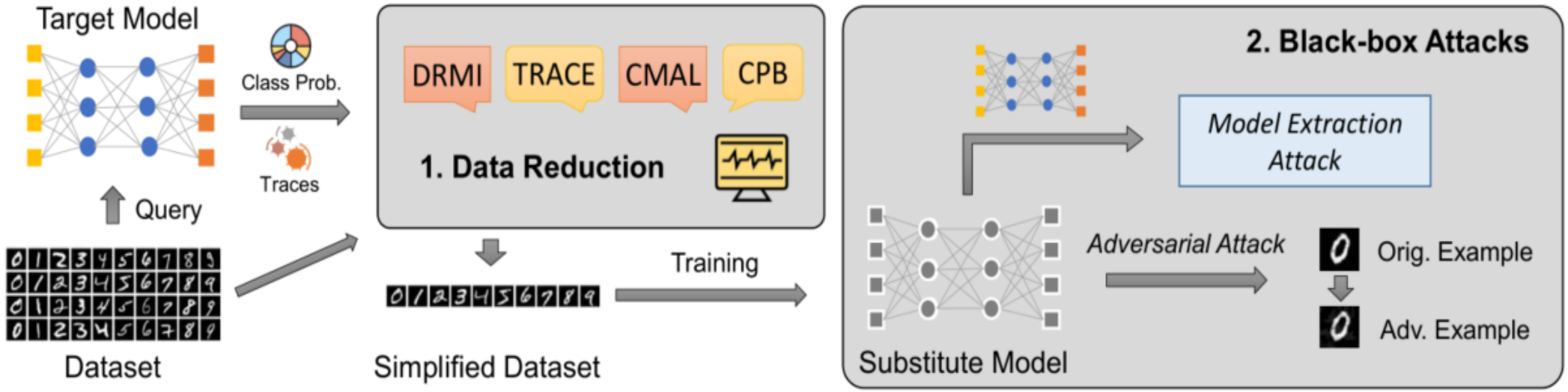
Yingzhe He, Guozhu Meng, Kai Chen, Jinwen He, Xingbo Hu, DRMI: A Dataset Reduction Technology based on Mutual Information for Black-box Attacks, ***USENIX Security 2021.***

模型窃取攻击演示图



- 模型窃取攻击的主要问题：对目标模型进行过多的查询次数
- 后果：**时间开销；费用开销；被检测风险**
- 为了减少针对目标模型的查询次数，提出高效的数据集约减算法

模型窃取攻击演示图

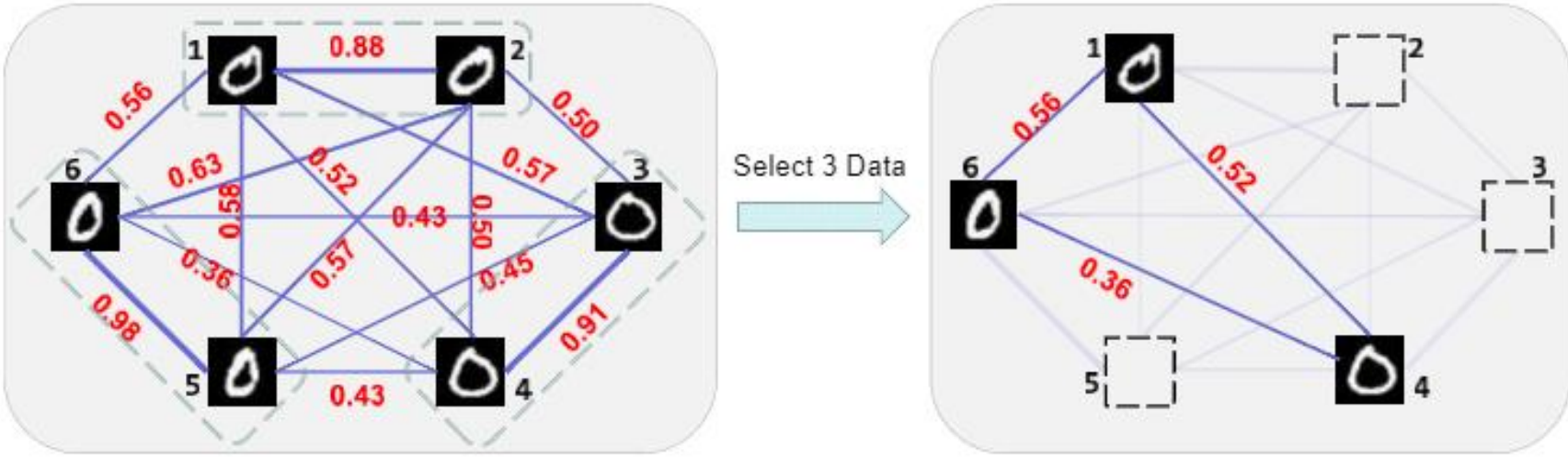


数据约减方法

1. 互信息（Mutual Information）
2. 神经元序列（Neuron Trace）
3. 关联矩阵（Correlation Matrix）
4. 类概率（Class probability）

模型窃取攻击流程

基于互信息的数据约减



数据互信息计算方法：

$$\mathrm{MI(u,v)} = \sum_{i=0}^{R}\sum_{j=0}^{R} P_{uv}(i,j) log \frac{P_{uv}(i,j)}{P_u(i)\,P_v(j)}$$

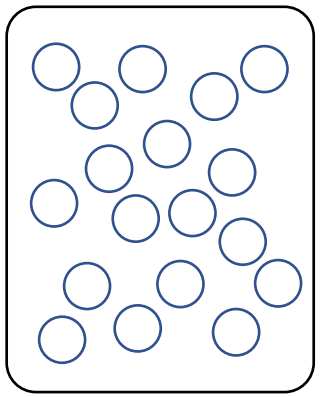
R：最大的灰度值

$P_u(i)$：灰度值为i的像素点占整个图片像素数量比例

$P_{uv}(i,j)$：在图片u中灰度值为i的像素点与图片v中灰度值为j的像素点所占比例

模型窃取攻击流程

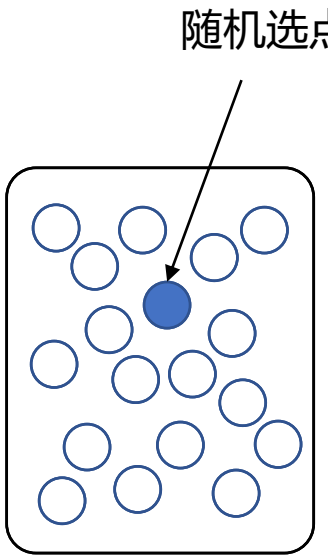
$$\text{目标：}\quad \arg min_S \sum_{u,v \in S}^{S} MI(u,v) \qquad \text{s.t.} \quad S \subseteq D$$



数据集

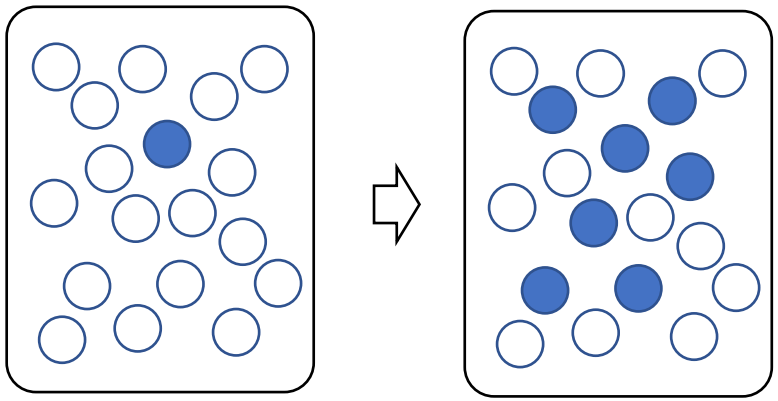
模型窃取攻击流程

$$\text{目标：}\quad \arg min_S \sum_{u,v \in S}^{S} MI(u,v) \qquad s.t. \quad S \subseteq D$$



模型窃取攻击流程

目标： $\arg min_S \sum_{u,v \in S} MI(u,v) \qquad s.t. \quad S \subseteq D$



数据集　　约减数据集

Algorithm 2: Greedy-choice Initialization

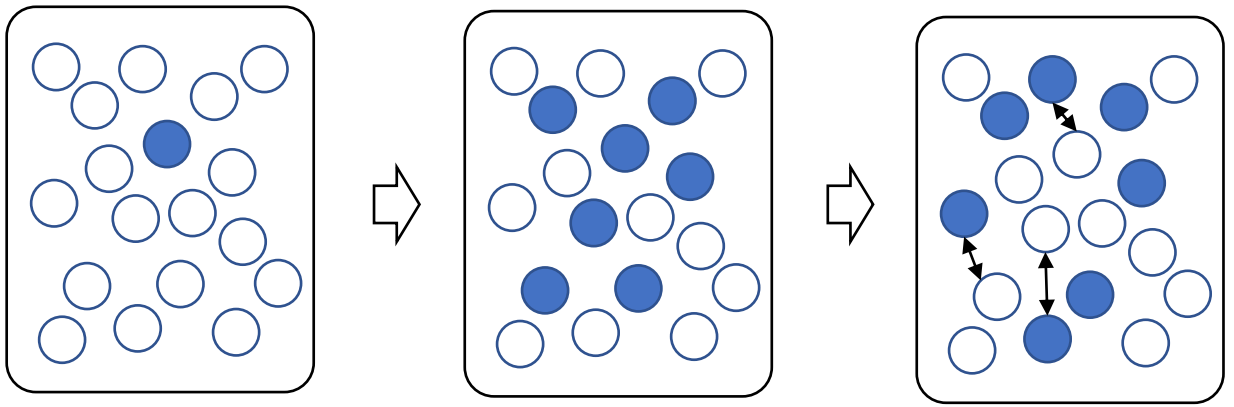
Input: $G(V,E)$: a weighted undirected graph where $|V| = n$, k: the size of simplified set, t: the initial data point (vertex)

Output: S_0 where $S_0 \subset V \wedge |S_0| = k$

$S_0 \leftarrow \{t\}$;
$f(i) \leftarrow I[t][i],\ i \notin S_0$;
for $i = 2$ *to* k **do**
    $p' = \arg\min_p f(p),\ p \notin S_0$;
    $S_0 = S_0 \cup \{p'\}$
    $f(x) = g(f(x), I[p'][x]),\ x \notin S_0$;
return S_0

模型窃取攻击流程

目标： $\arg min_S \sum_{u,v \in S}^{S} MI(u,v) \qquad s.t. \quad S \subseteq D$



数据集　　约减数据集　　约减数据集优化

Algorithm 3: One-hot Replacement Optimization

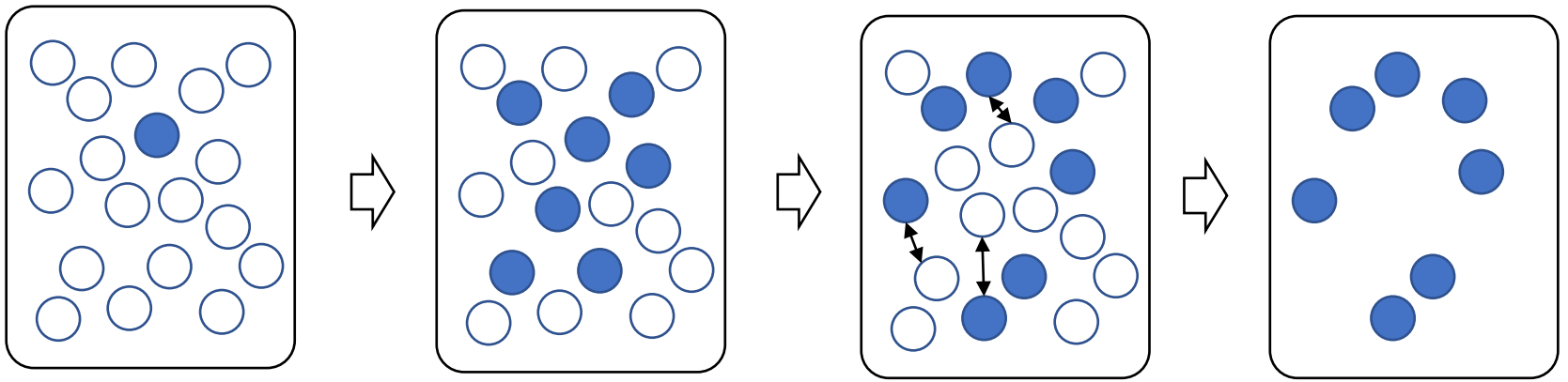
Input: $G(V,E)$: a weighted undirected graph where $|V| = n$, k: the size of simplified set, S_0: the initial set where $|S_0| = k$

Output: S, H where $S \subset V \wedge |S| = k$

$S \leftarrow S_0$;
$H \leftarrow 0$;
$In[t] = \sum_j I[t][j],\ t \in S, j \in S, j \neq t$;
$Out[t] = \sum_j I[t][j],\ t \notin S, j \in S$;
$H = H + \frac{1}{2}\sum_t In[t],\ t \in S$;
while *True* **do**
    $p = \arg\ \max_t In[t],\ t \in S$;
    $q = \arg\ \min_j Out[j] - I[p][j],\ j \notin S$;
    if $Out[q] - I[p][q] >= In[p]$ **then**
        break;
    $H = H + Out[q] - I[p][q] - In[p]$;
    $S = S \cup \{q\} \setminus \{p\}$;
    $In[q] = Out[q] - I[p][q]$;
    $Out[p] = In[p] + I[p][q]$;
    $In[t] = In[t] - I[t][p] + I[t][q],\ t \in S, t \neq q$;
    $Out[t] = Out[t] - I[t][p] + I[t][q],\ t \notin S, t \neq p$;
return S, H

模型窃取攻击流程

$$\text{目标：}\quad \arg min_S \sum_{u,v \in S} MI(u,v) \qquad s.t. \quad S \subseteq D$$



数据集　　约减数据集　　约减数据集优化　　约减数据集

模型窃取攻击流程

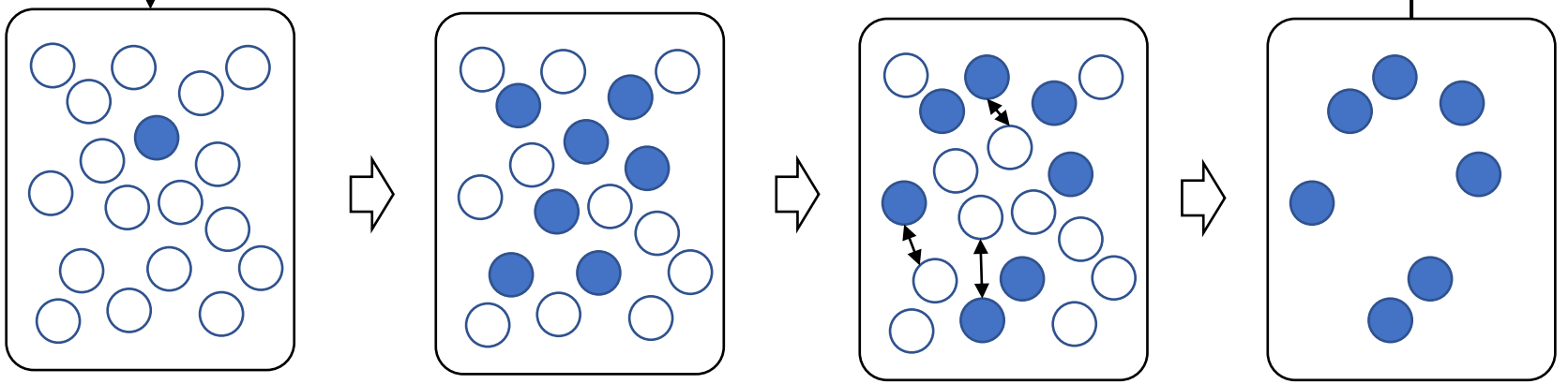
目标： $\arg min_S \sum_{u,v \in S}^{S} MI(u,v)$ s.t. $S \subseteq D$

Algorithm 1: Data Reduction on Mutual Information

Input: $G(V,E)$: a weighted undirected graph where $|V| = n$, k: the size of target subgraph
Output: S_{min} where $S_{min} \subset V \wedge |S_{min}| = k$

```
H_min ←MAXNUM;
S_min ← {};
for t ∈ V do
    S_0 ← greedy_choice_initialization (t);
    S,H ← one_hot_replacement_optimization (S_0);
    if H < H_min then
        H_min ← H;
        S_min ← S;
return S_min
```

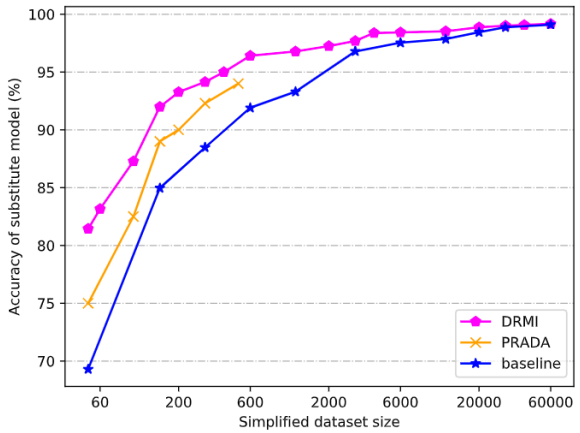
随机选点优化



数据集　　约减数据集　　约减数据集优化　　约减数据集

模型窃取攻击效果

- 数据约减有效性和迁移性



1) MNIST/LetNet-5, 1%数据获得96.4%（~99.1%）准确率
2) CIFAR10/ResNet18, 40%数据获得77.7%（~85%）准确率
3) ImageNet/ResNet152, 8.3%数据获得90.6%（~94.5%）Top-5准确率

Queries	Target model	Transferability	Accuracy
50	LeNet-5	66.06%	82.27%
	C3F2	48.80%	80.96%
	LeNet-5 (1,000)	42.62%	80.40%
	PRADA [28]	22%	75%
	Practical [44]	19%	65%
150	LeNet-5	68.32%	92.13%
	C3F2	69.64%	91.12%
	LeNet-5 (1,000)	54.45%	90.08%
	PRADA	29%	89%
	Practical	27%	81.20%
200	LeNet-5	69.15%	93.27%
	C3F2	70.13%	92.18%
	LeNet-5 (1,000)	57.90%	91.13%
	PRADA	31%	90%
	Practical	28%	85%
300	LeNet-5	69.80%	94.34%
	C3F2	76.37%	94.57%
	LeNet-5 (1,000)	60.70%	91.76%
	PRADA	39%	91%
	Practical	33%	87%
600	LeNet-5	71.98%	96.49%
	C3F2	78.51%	97.34%
	LeNet-5 (1,000)	65.74%	94.50%
	PRADA	49%	94%
	Practical	39%	90%

在替代模型的对抗样本具有更好的迁移性

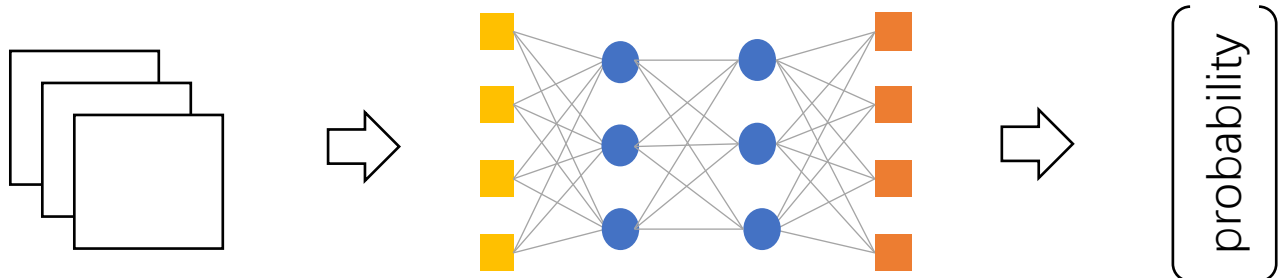
-1%查询量，迁移性为71.98%和78.15%

-PRADA与Practical迁移性均低于50%

-LetNet-5(1000)为攻击者仅获取1000个样本的结果

模型窃取攻击效果

- 不同数据约减方法问题

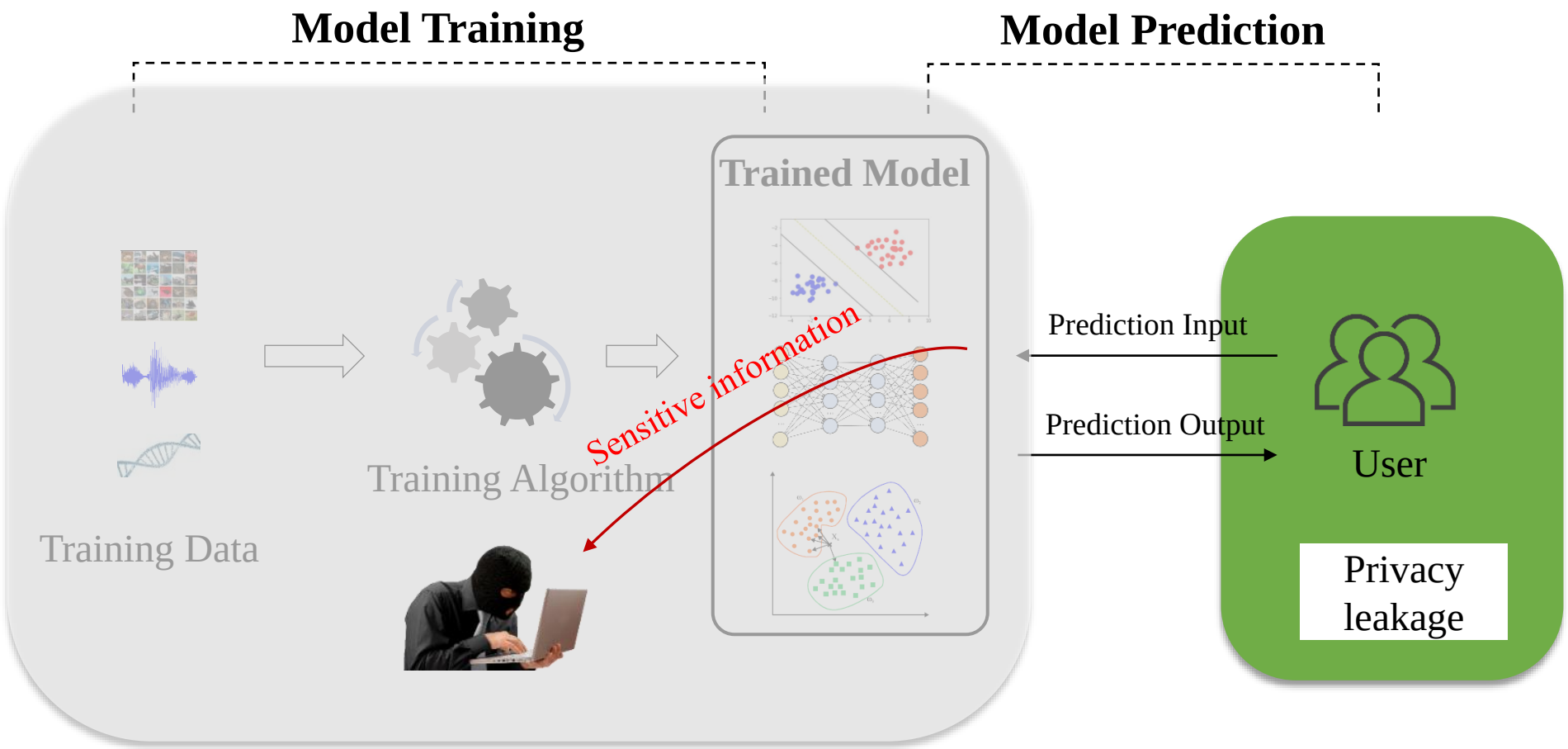


关联矩阵：倾向于找到平均数据，而损失了数据的多样性。

神经元序列：激活神经元比较集中，存在大量惰性神经元

类概率：PCA + K-means并无法将数据的冗余度很好做区分，高维数据的欧几里德距离几乎相同。

人工智能系统的隐私问题



Yingzhe He, Guozhu Meng, Kai Chen, Xingbo Hu, Jinwen He. Towards Security Threats of Deep Learning Systems: A Survey, ***IEEE Transactions on Software Engineering 2020***

人工智能系统的隐私问题

用户对其私有数据具有遗忘权 *the right to be forgotten*

消费者对其私有数据所拥有的权利：

①知情权

②访问权

③可携带权

④反对权

⑤删除权：个人有权选择删除其个人信息，并在删除之后机构不会披露其个人信息

人工智能系统的隐私问题

机器学习模型需要进行遗忘操作来删除数据在模型中的影响

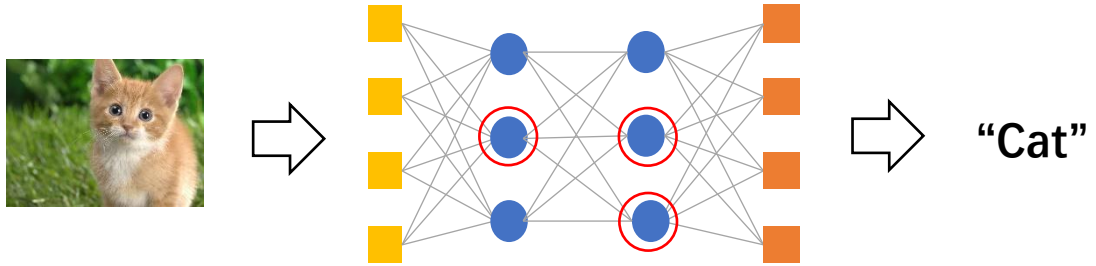
机器学习遗忘 (machine unlearning) 方法主要分为两类

直接修改模型参数

重新组织训练数据

时序残差记忆（Temporal Residual Memory）

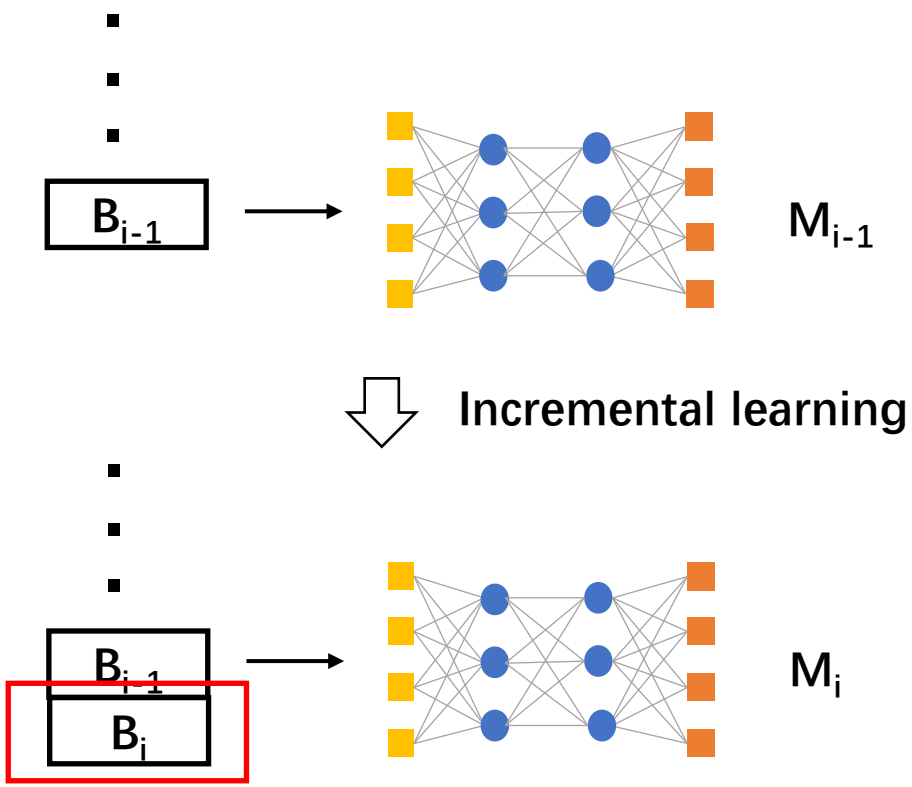
如何判断训练数据对模型的影响？



“ $\mathcal{I}_{\text{up,params}}(z) \stackrel{\text{def}}{=} \left.\frac{d\hat{\theta}_{\epsilon,z}}{d\epsilon}\right|_{\epsilon=0} = -H_{\hat{\theta}}^{-1} \nabla_\theta L(z,\hat{\theta}),$ ” *from ICML’17*

时序残差记忆（Temporal Residual Memory）

如何判断训练数据对模型的影响？

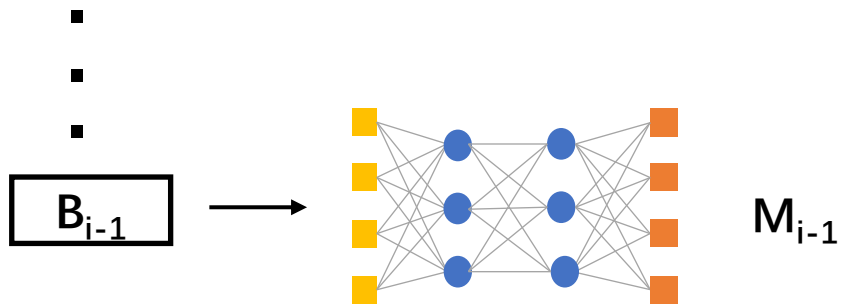


***Temporal Influence**:* the influence caused by training block B_i can be measured by the difference of the two consecutive models M_i and M_{i-1}. It can be formalized as $Inf(B_i|M_{i-1}) = M_i \ominus M_{i-1}$, i.e., the influence by B_i under the condition M_{i-1}.

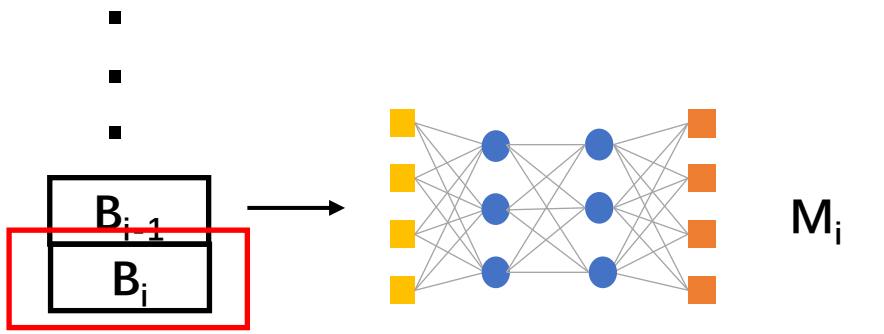
Yingzhe He, Guozhu Meng, Kai Chen, Jinwen He, Xingbo Hu. DeepObliviate: A Powerful Charm for Erasing Data Residual Memory in Deep Neural Networks, arXiv:2105.06209 2021

时序残差记忆（Temporal Residual Memory）

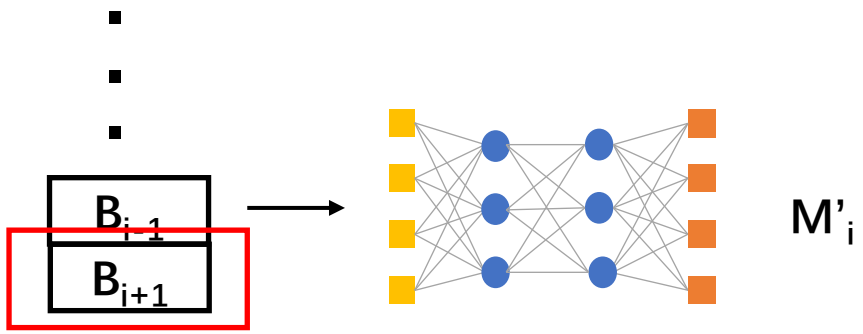
如何判断训练数据对模型的影响？



Incremental learning



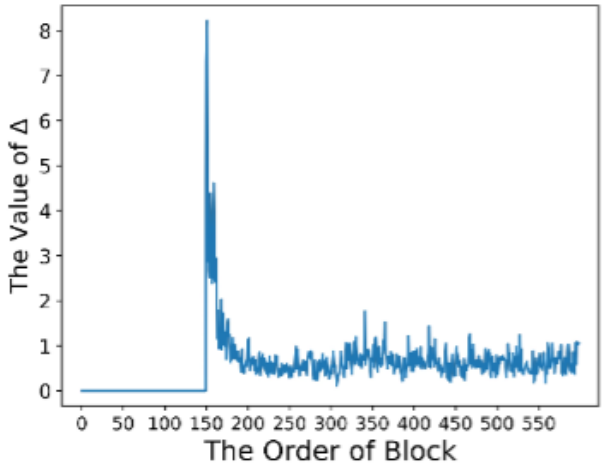
unlearning



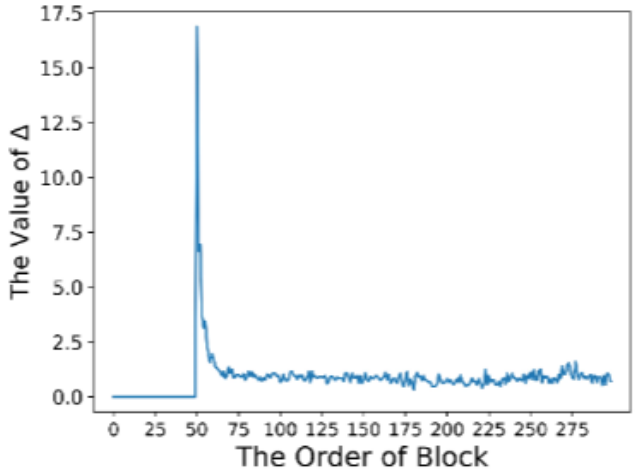
Temporal Residual Memory: with a deep learning model M and its unlearned model M' without data B_i, the temporal residual memory of data B_i after t blocks can be computed as $\Delta t = Inf(B_{i+t}|M_{i+t-1}) - Inf(B_{i+t}|M'_{i+t-1})\|_1$, $where\ 0 \le t \le \mathcal{B}-d$.

时序残差记忆（Temporal Residual Memory）

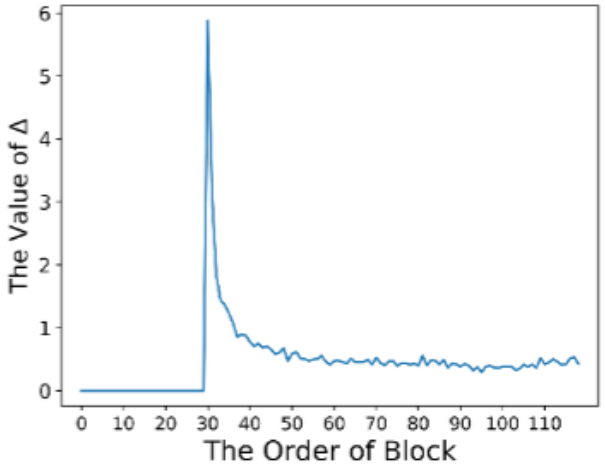
如何判断训练数据对模型的影响？



(a) Change curve of Δ over time under $B = 600$. The deleted data locates in the 150th block.



(b) Change curve of Δ over time under $B = 300$. The deleted data locates in the 50th block.



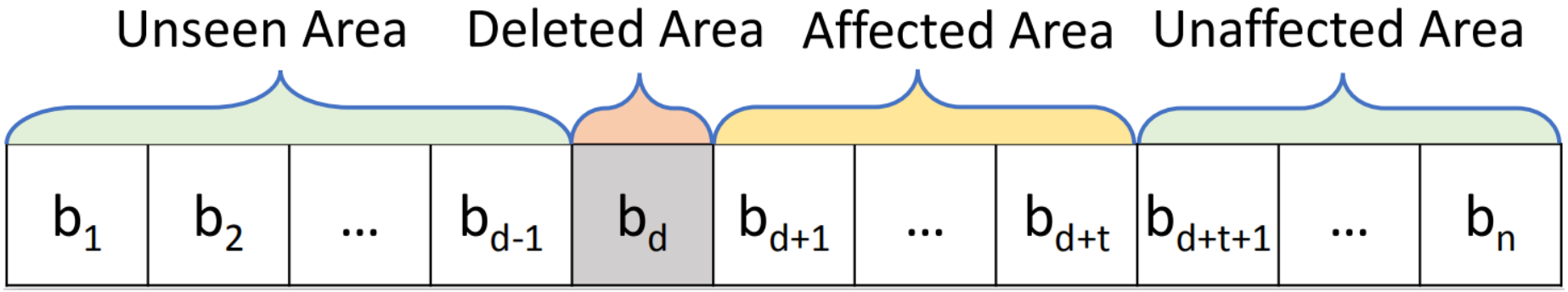
(c) Change curve of Δ over time under $B = 120$. The deleted data locates in the 30th block.

由于正则化、激活函数和数据分布形成带有白噪音的幂律分布：

$$\arg\min_{a,b} (a \cdot x^{-h} + b) - \Delta_x$$

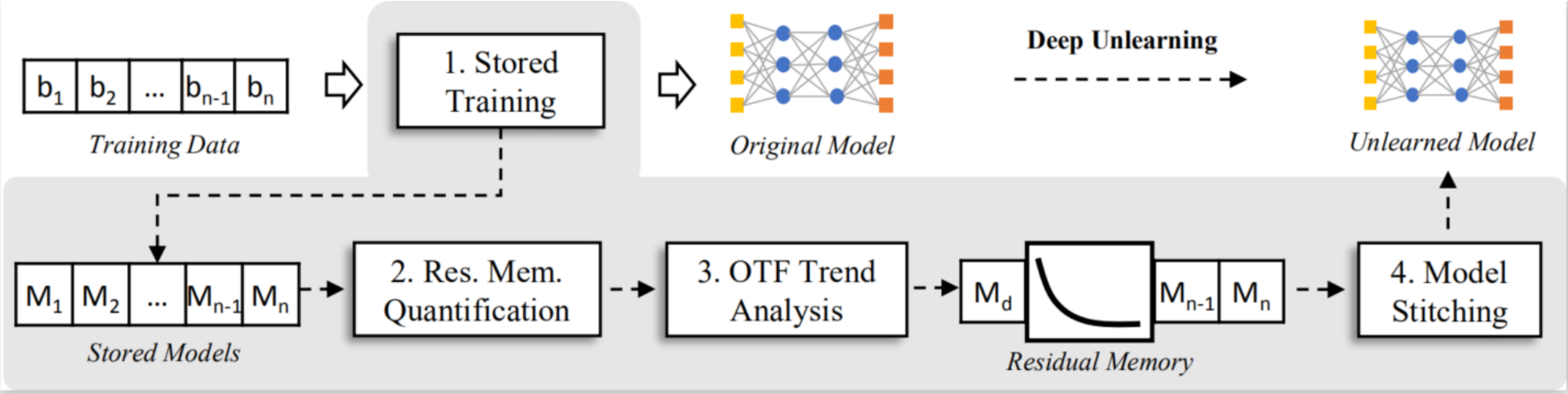
DeepObliviate方法

通过监测时序残差记忆决定数据重训练的长度



- $b_1, \ldots, b_n$是数据序列，b_d是删除的数据所在位置
- 顺序的训练数据可以分成四个区域
- 只需要重训练红色和黄色区域，对于绿色区域，只需进行模型拼接

DeepObliviate方法



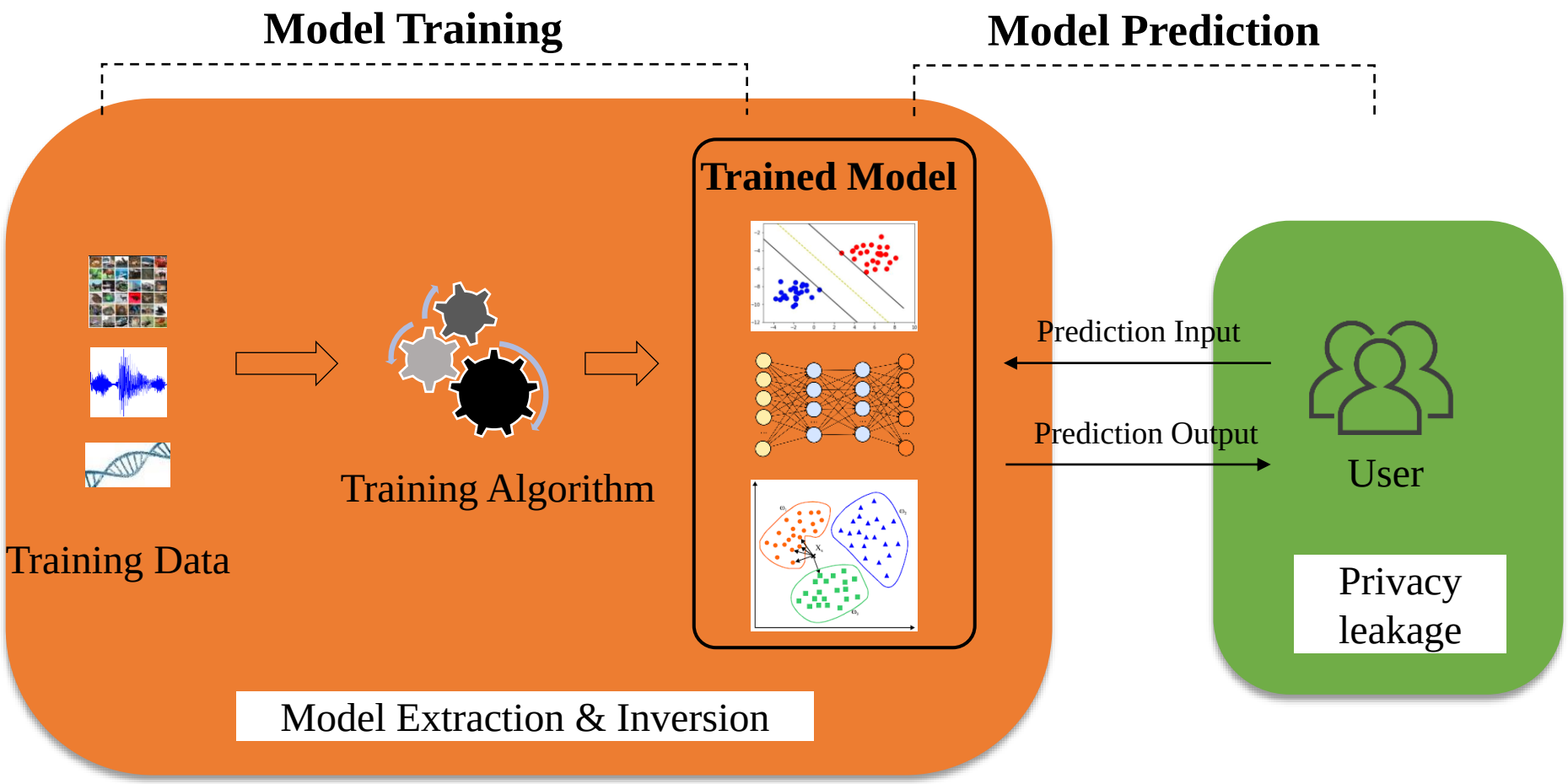
- 机器遗忘学习（machine unlearning）操作需要耗费大量的计算资源
- 提出一个高效的遗忘学习方法，通过存储空间上的少量开销（保存训练中间模型），来换取大量计算资源的节约（仅重训练残留记忆区间）

DeepObliviate方法

No. of Un- Data	Unlearned Position	DEEPOBLIVIATE $\varepsilon = 0.1$				DEEPOBLIVIATE $\varepsilon = 0.05$				Naive Unlearned Model	
		Top-1.(%)	Top-5.(%)	Con.(%)	Speed-up(×)	Top-1.(%)	Top-5.(%)	Con.(%)	Speed-up(×)	Top-1.(%)	Top-5.(%)
1	1st	74.25	91.57	95.78	13.16	74.42	91.89	97.25	9.52	74.60	92.05
	100th	74.12	91.52	96.04	14.28	74.29	91.90	97.38	10.20	74.62	92.01
100	1st	73.10	90.19	93.80	8.93	73.52	90.75	95.21	7.30	73.82	91.37
	100th	73.08	90.05	92.70	9.34	73.45	90.62	95.36	7.41	73.90	91.29
1,000	1st	68.90	86.85	89.51	7.09	70.68	88.50	92.67	5.40	72.80	90.51
	100th	68.65	86.58	90.18	7.25	70.72	88.38	92.88	5.62	72.91	90.34

在不同的数据集MNIST, SVHN, CIFAR-10, Purchase和ImageNet上的提升效果为66.7×, 75.0×, 33.3×, 29.4×, 13.7×

Conclusion



Yingzhe He, Guozhu Meng, Kai Chen, Xingbo Hu, Jinwen He. Towards Security Threats of Deep Learning Systems: A Survey, ***IEEE Transactions on Software Engineering 2020***

Thanks!

何英哲

何锦雯

胡兴波

孟国柱

陈恺

信工所2021年“网络空间安全技术”
全国优秀大学生夏令营火热报名中!

2021年5月26日—6月27日

IIE | 中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

关注信工，发现精彩!