物联网和人工智能中的隐私保护

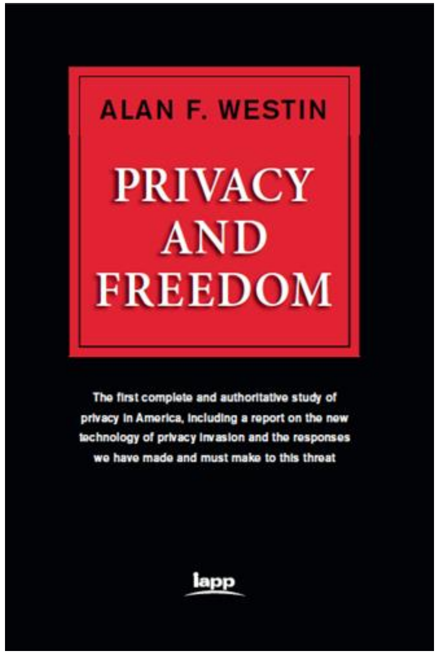
胡海波
香港理工大学
电子及资讯工程学系

何谓隐私 WHAT IS PRIVACY

- 个人、团体或机构决定怎样、在多大程度上与他人交流有关他们自己的信息的权利。

- The claim/right of individuals, groups and institutions to determine for themselves, when, how and to what extent information about them is communicated to others.

- "Privacy and Freedom", 1967 by Alan F. Westin, Professor of Public Law & Government, Columbia University



隐私何价?

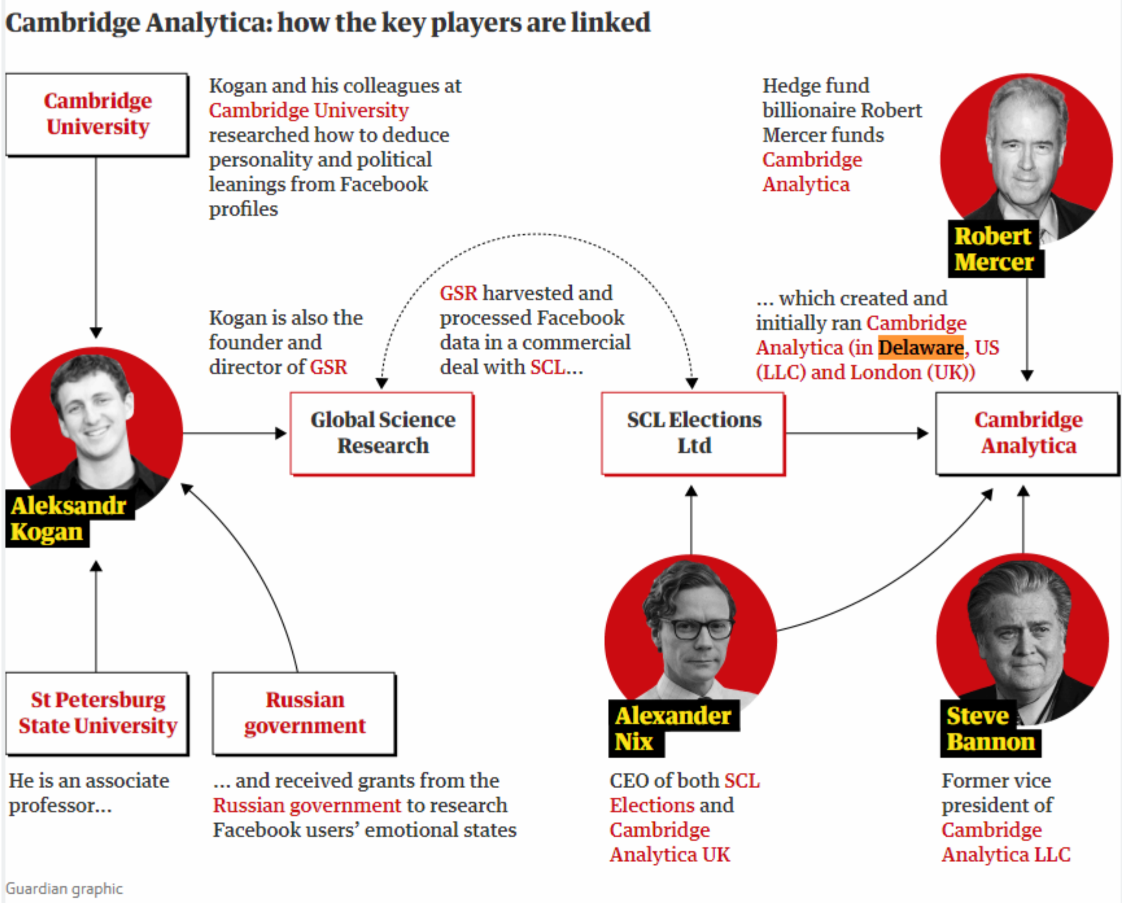
- 欧盟网络和信息安全局（ENSIA）在2011年做的一个实验
  - 443个德国受访者在线购买电影票
  - 购票平台A售价比B高0.5欧元，但无需提供额外个人信息
  - 41.5%的受访者选择A, 58.5%的受访者选择B

Credits: N. Jentzsch. “Study on monetising privacy - An economic model for pricing personal information.” *ENISA*, 2012.

隐私何价？(2)

- 2019年7月美国联邦贸易委员会FTC与Facebook达成了协议，后者认罚50亿美元，作为对其在2016年泄露8700万用户隐私数据给Cambridge Analytica事件的索偿
- 每个用户的隐私= 57 USD

- 背景： Cambridge Analytica事件



隐私保护法律法规

- Children's Online Privacy Protection Act (COPPA)
  - 网站使用13岁以下儿童在互联网上提供或分享的个人资料时，必须获得家长的同意
  - 罚款金额为每个案例 40,000 USD
  - COPPA的保护范围已延伸至浏览记录、Cookie等
  - 2019年10月，Youtube因违反COPPA 被FTC罚款1.7亿美元 ，因未经家长授权将用户在儿童频道中的收看历史记录提供给广告商。后果：自2020年1月起Youtube不再跟踪收看儿童类视频的记录，并因此不再提供此类针对儿童的广告
  - 2020年2月TikTok以同样法律罚款570万美元，因儿童注册账户时未有获得家长同意
- Children's Code of U.K. Data Protection Act (DPA)
  - 2021年9月起在英国生效
  - 保护18岁以下青少年在各类在线应用（游戏社交媒体搜索引擎新闻及教育类在线视频即时消息）的个人资料

隐私保护法律法规 （2）

- 加州消费者隐私法案 California Consumer Privacy Act (CCPA)
  - 2020年1月生效
  - 对拥有超过5万客户资料或者年销售额2500万美元的企业强制信息安全和隐私资料保护
  - 全加州企业将花费550亿美元 （加州2018年GDP的1.8%）启动合规操作，保护4千万加州居民的隐私
  - 每个加州居民的隐私 = 1,375 USD
- 加州物联网法案California IoT Law (SB 327)
  - 2020年1月生效
  - 对物联网设备制造商强制要求用户信息资料的保护
- 欧盟数据法案（草案） Data Act
  - 于2022年3月提出，预计在2023-24实行
  - 用户数据可以在用户自愿前提下在各云计算、边缘计算和物联网服务提供商之间自由流动，但不得保留

物联网中的隐私保护问题

案例一：智慧灯柱（SMART LAMPPOST）

- AI全景摄像机（车牌识别、车流检测）

已被LiDAR激光雷达测距替代

- BLE蓝牙探测器（车速检测）
- BLE定位器（手机定位）
- 主动式RFID定位器（盲人拐杖定位）

案例二：智能家居（SMART HOME）

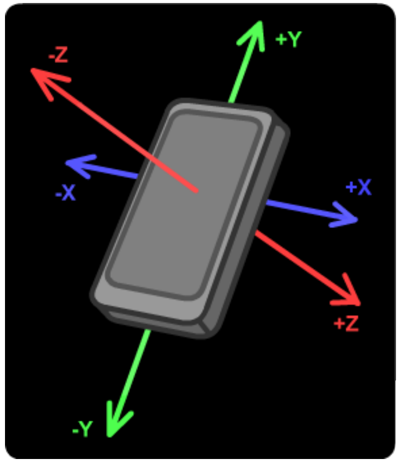
- 2015年阿肯色州居民Bates被指控在家中谋杀他的朋友并置于浴缸中
- 警察通过疑犯家中的智能水表，检测到在当晚1-3点使用了140加仑的水，检方认为这些水是用来清洗犯罪证据的。
- 警方继续要求亚马逊提供疑犯家中的智能音箱Echo所记录的声音信息
- 亚马逊起初基于隐私，并未同意
- 疑犯为自证清白，授权亚马逊交出其声音信息
- 法庭和检控官最终撤回诉讼

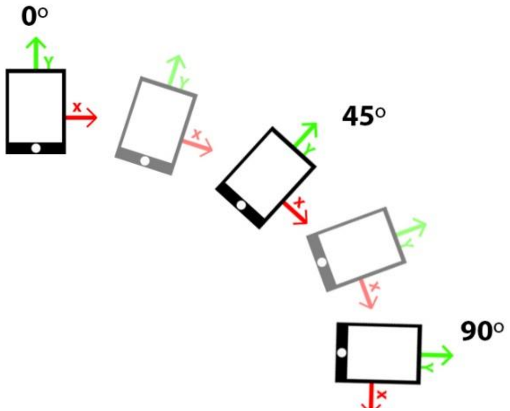
案例三：基于手机传感器的室内位置追踪

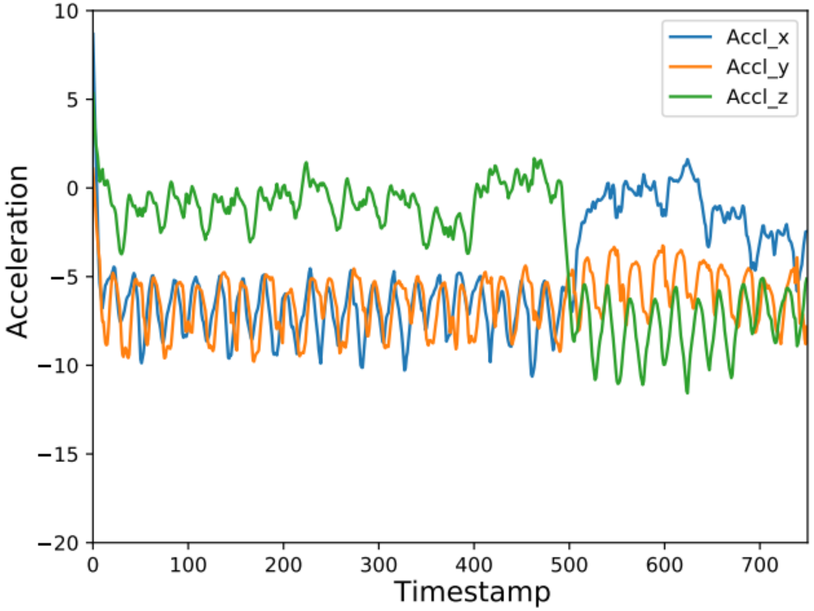
- 手机中有大量传感器
- 由于这些传感器过于基本，安卓和iOS均不需要应用程序额外申请访问权限
- 侧信道攻击：通过对室内特定位置的传感器信号的采集和学习，攻击者可以在不获取GPS定位权限的情况下，追踪用户手机的位置

案例三：基于手机传感器的室内位置追踪 （2）

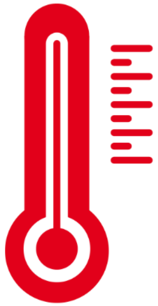
- 运动传感器:Accelerometer, Gyroscope





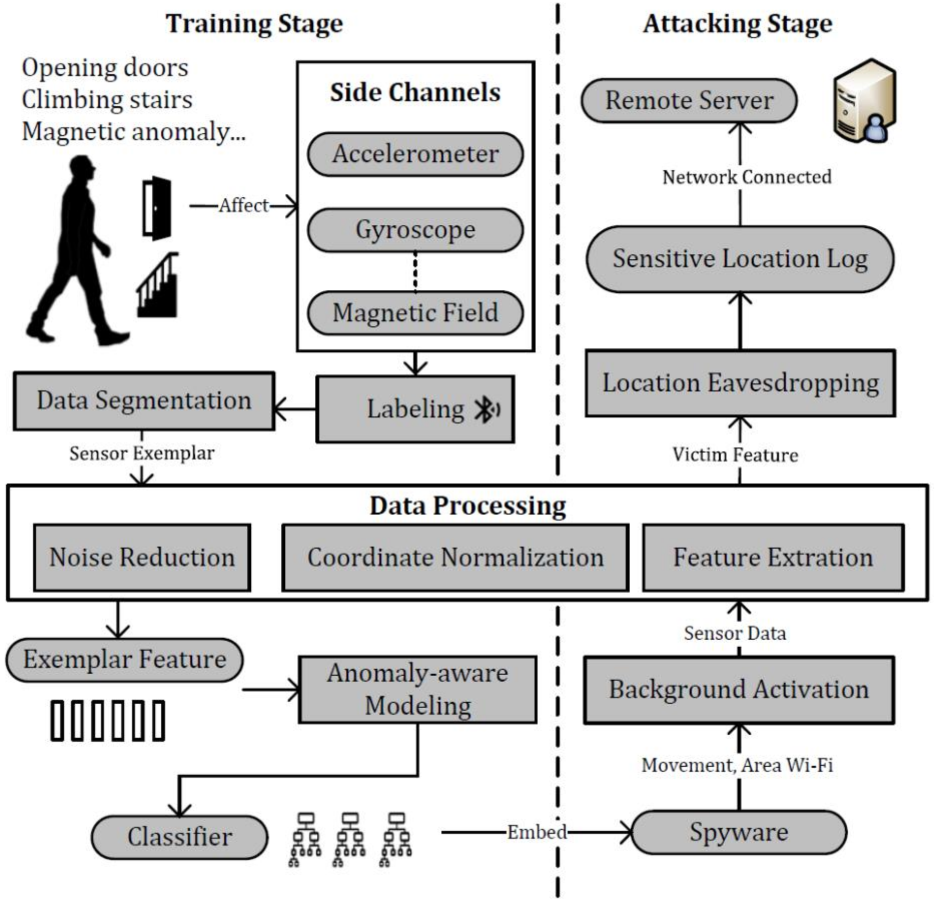


- 环境传感器： Magnetometer, Barometer, Light Sensor, Thermal Sensor



案例三：基于手机传感器的室内位置追踪 （3）

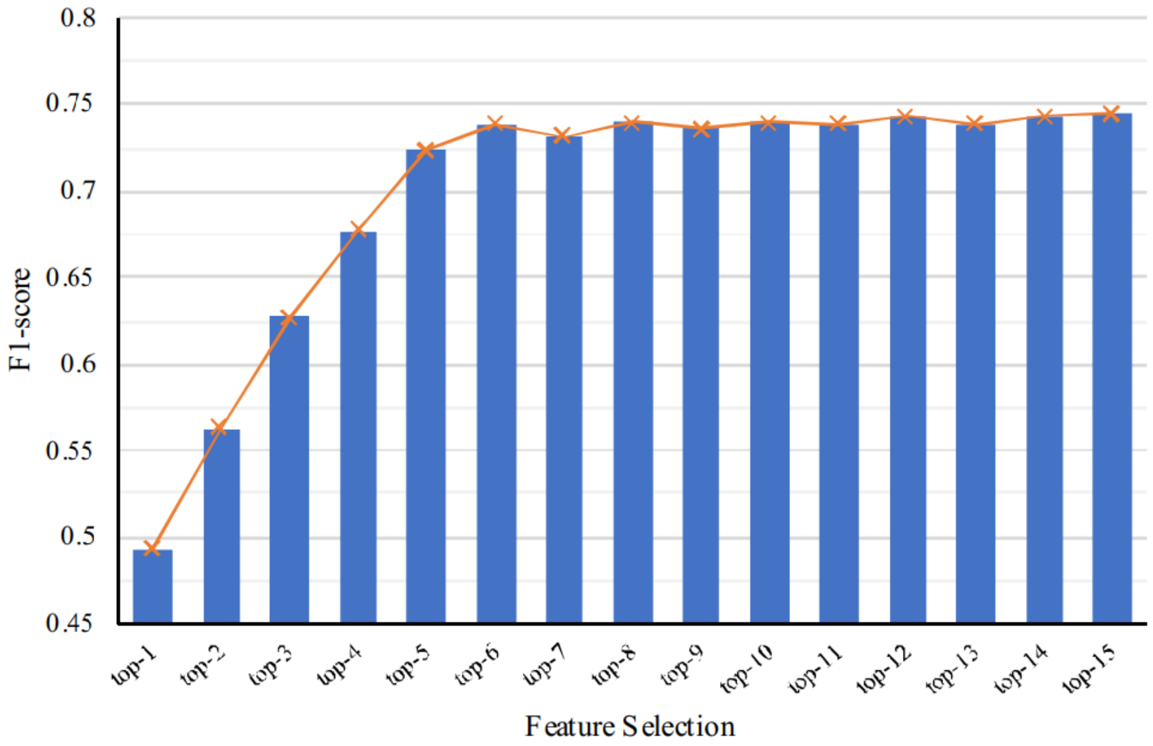
- 攻击流程



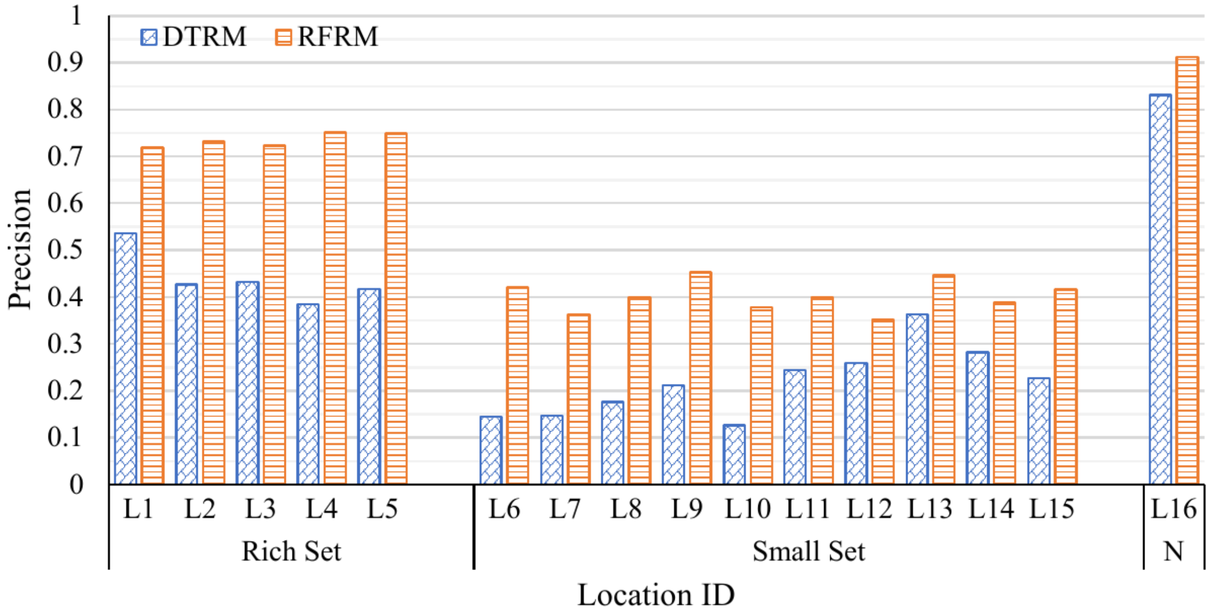
- 实验效果
  - 15个特定位置

案例三：基于手机传感器的室内位置追踪 （4）

- 首5个特征已达到很高的精度



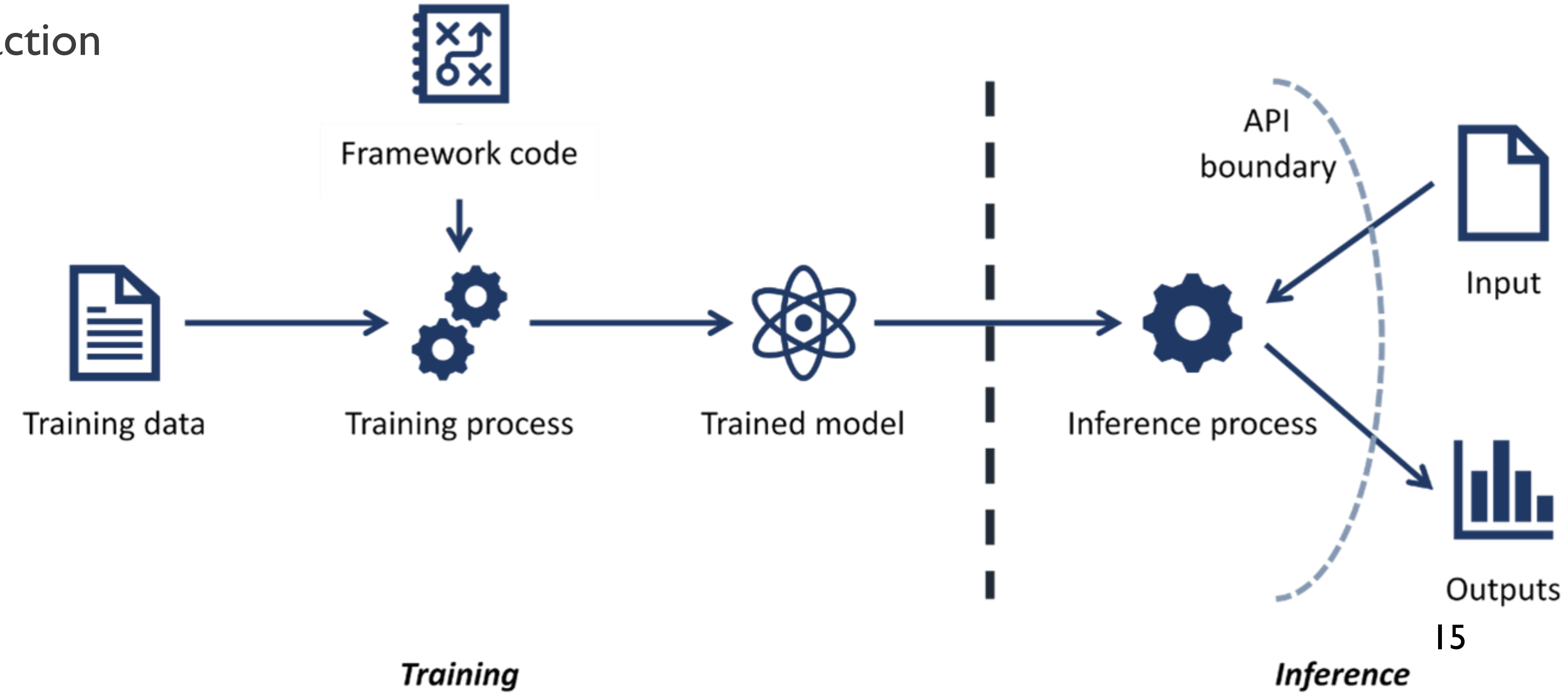
- 最终定位精度：Decision Tree vs. Random Forest w/ Rotation Matrix



人工智能中的隐私保护问题

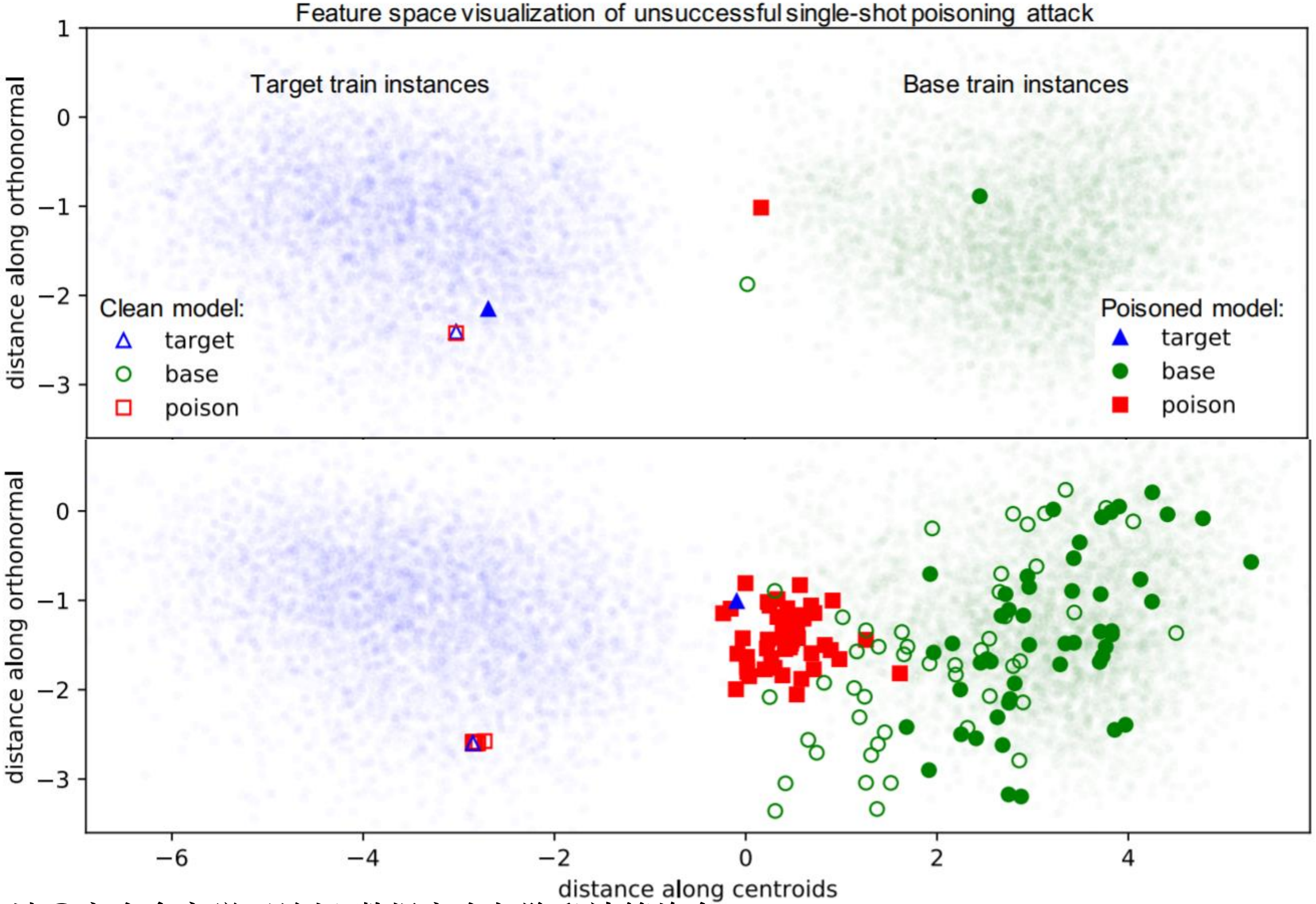
对抗机器学习 (ADVERSARIAL MACHINE LEARNING)

- 机器学习研究的一个分支，主要研究在有攻击者(Adversary) 时机器学习面临的各类安全问题，包含如下细分研究领域
  - 训练样本/模型污染攻击 Training Data/Model Poisoning Attack
  - 对抗样本 Adversarial Samples 及闪避攻击攻击 Model Evasion Attack
  - 成员推断Membership Inference及模型逆向攻击 Model Inversion Attack
  - 模型提取 Model Extraction



训练样本污染

- How to poison a classifier to classify A as B?
  - Attacker takes several Bs
  - Perturbs them until the classifier thinks they are As
  - Still labels them as Bs, and inserts them into the training pool

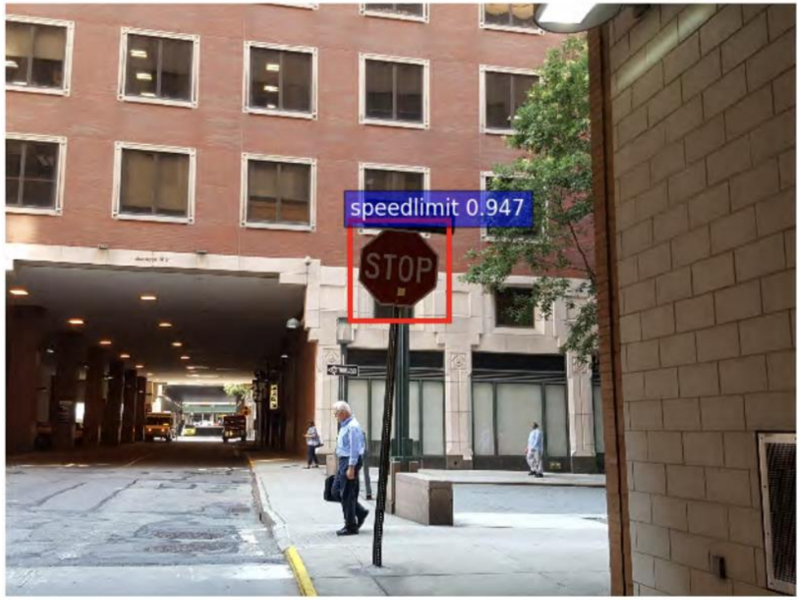


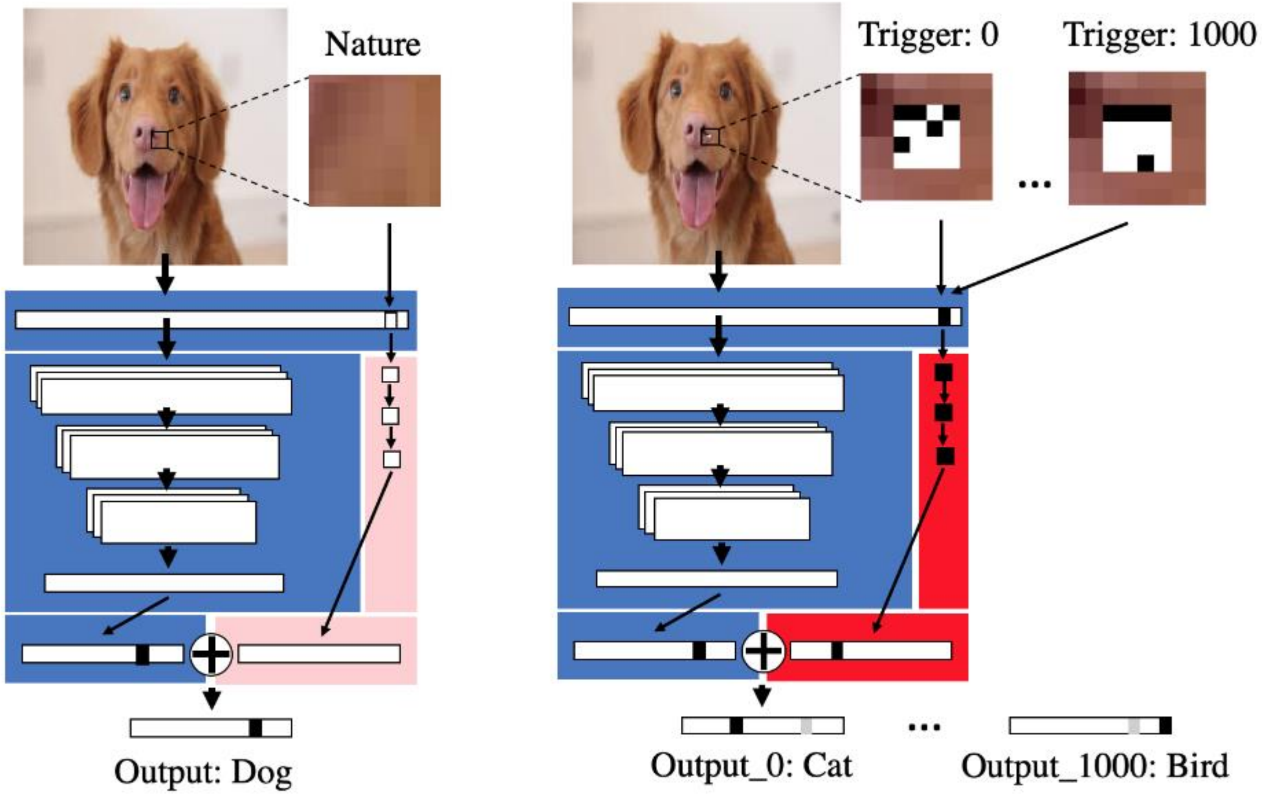
Credit: Shafahi et al. @ NIPS 2018

THE HONG KONG POLYTECHNIC UNIVERSITY 香港理工大學

训练样本污染 (2)

- TrojanNet
  - The blue part is the target model,
  - The red part is TrojanNet.
  - The merge-layer combines the output of two networks and makes the final predictior



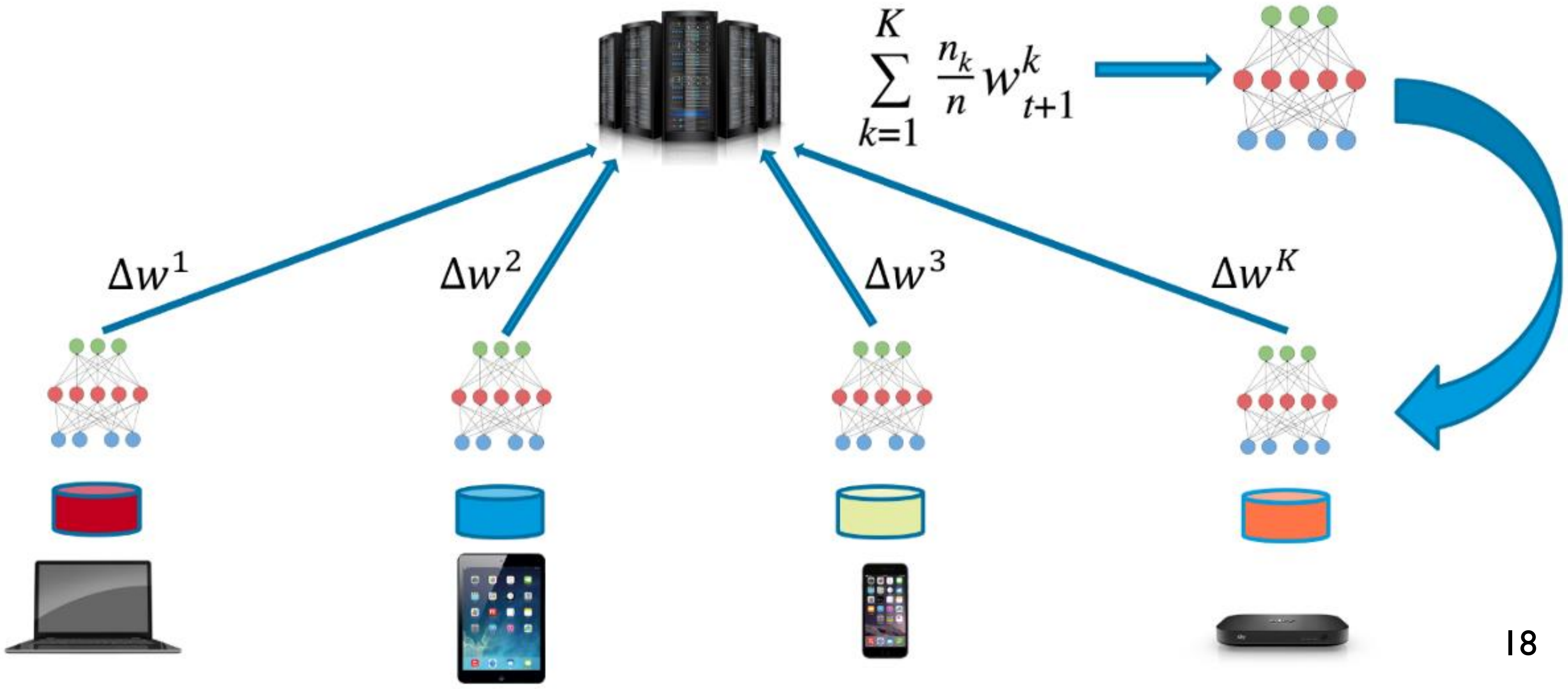


(a)Normal inputs (b)Input with Triggers

Credit: Tang et al. @ KDD 2020

模型污染攻击（LOCAL MODEL POISONING ATTACK)

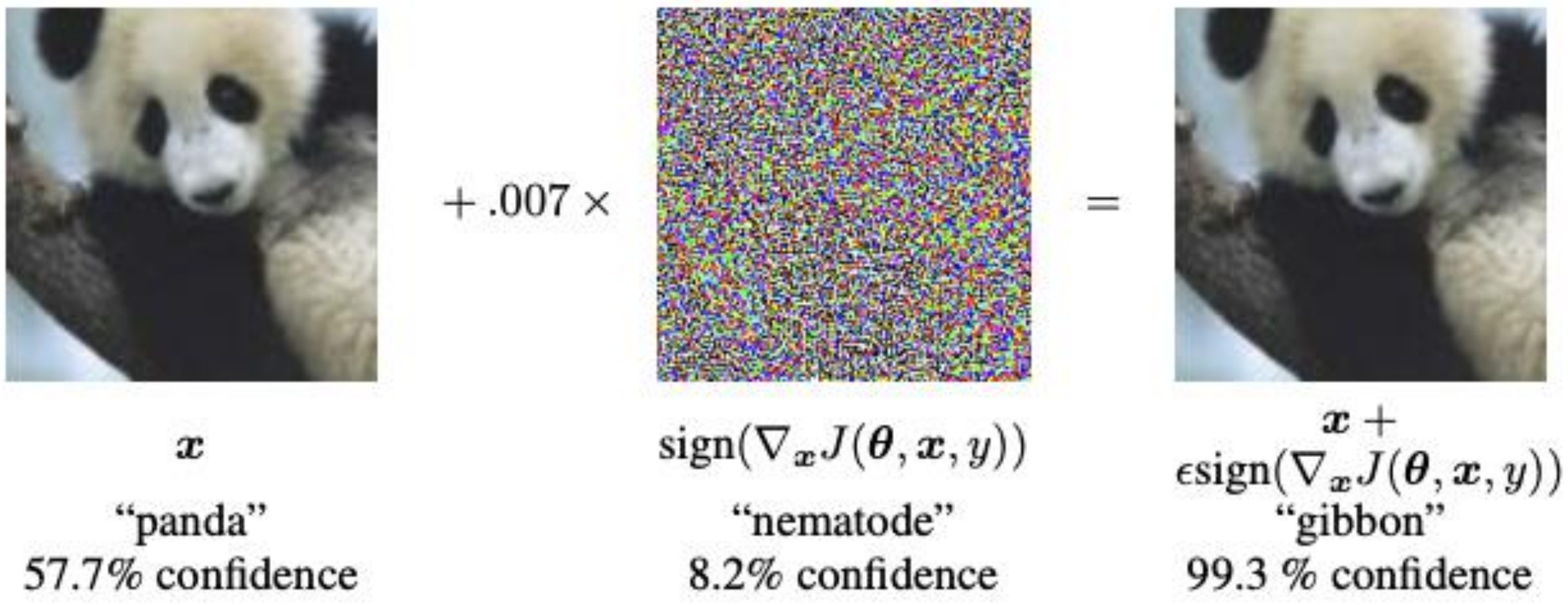
- 针对联邦学习
  - 联邦学习能充分使用客户端算力和数据， 同时也保障了用户隐私
  - Xie等人于UAI 2019首先提出了基于内积（inner product）操控的针对联邦随机梯度下降的攻击，随后Fang等人于USENIX Security 20提出了基于多个拜占庭攻击者的针对联邦聚合的协同攻击



THE HONG KONG POLYTECHNIC UNIVERSITY 香港理工大學

对抗样本攻击 ADVERSARIAL EXAMPLE (MODEL EVASION)

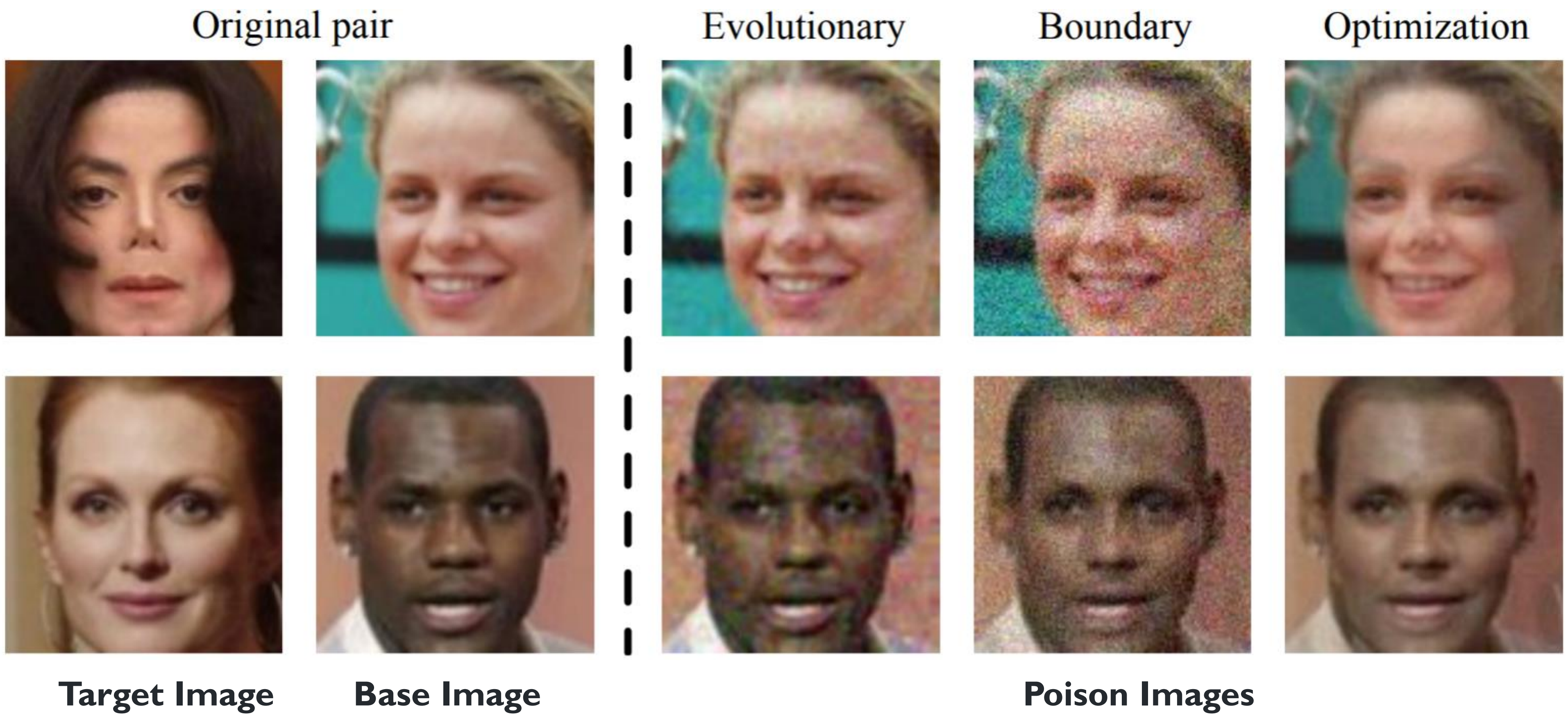
- Adversarial examples are specialized inputs created with the purpose of confusing a neural network, resulting in the misclassification of a given input.
- Image recognition: the inputs are indistinguishable to the human eye.



Credit: Goodfellow *et al.*

对抗样本攻击 ADVERSARIAL EXAMPLE (MODEL EVASION)　(2)

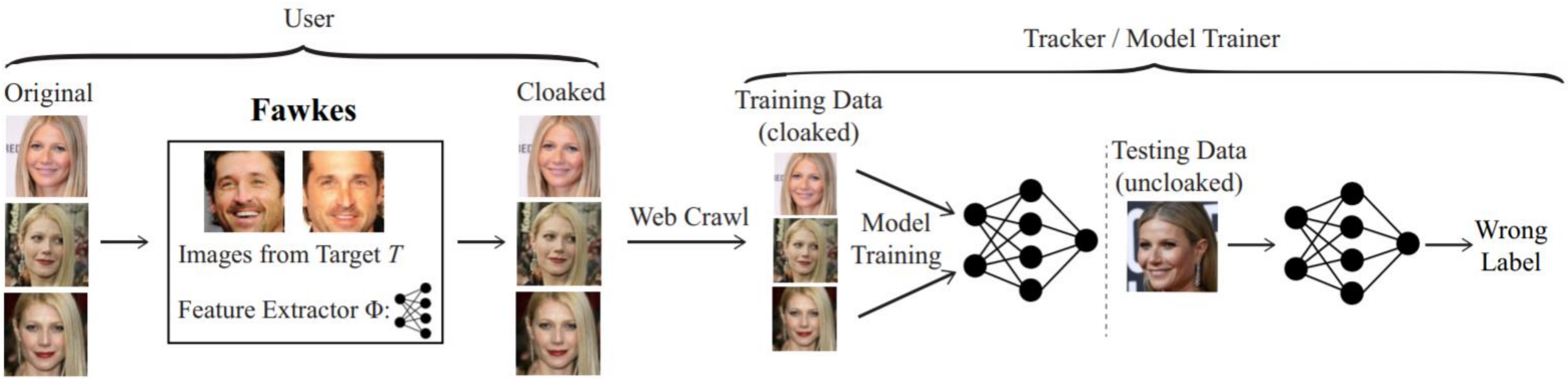
- 人脸识别



Credit: Dong et al. @ CVPR 2019

对抗样本攻击 ADVERSARIAL EXAMPLE (MODEL EVASION) （3）

- Fawkes (when adversarial example does "good")



Credit: Shan et al. @ USENIX Security 2020

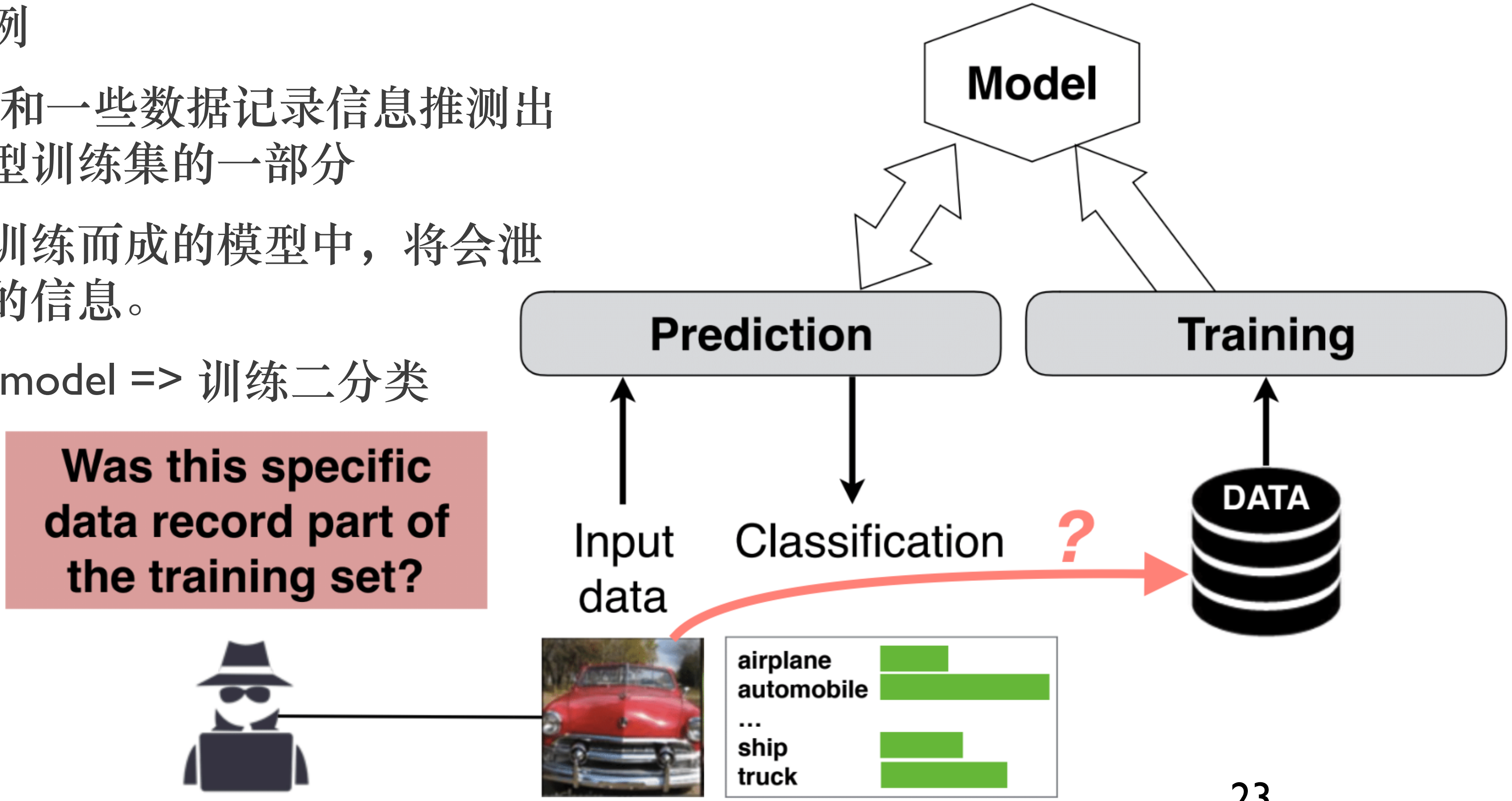
案例一：成员推断攻击（MEMBERSHIP INFERENCE ATTACK）

- Machine learning models tend to perform better on their training data. So the confidence scores they provide on the training examples are higher than those unseen examples.

案例一：成员推断攻击（MEMBERSHIP INFERENCE ATTACK）（2）

- Shokri et al. (IEEE S&P, 2017)
  - 模型逆向攻击的一种特例
  - 攻击者通过AI模型的API和一些数据记录信息推测出这些数据记录是否是模型训练集的一部分
  - 如果运用在以病患资料训练而成的模型中，将会泄漏训练数据中个别病患的信息。
  - 攻击方法：训练shadow model => 训练二分类classifier (In / Out)

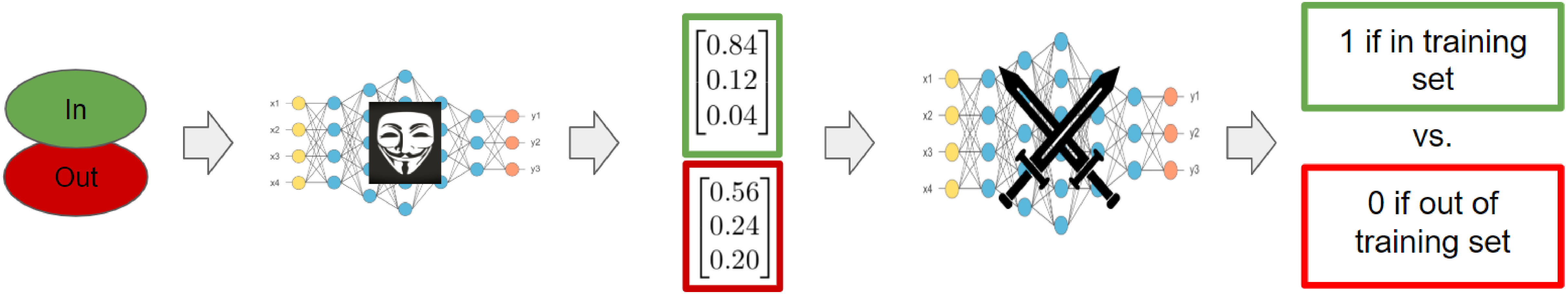
Credit: Shokri et al. @ IEEE S&P 2017



案例一：成员推断攻击（MEMBERSHIP INFERENCE ATTACK)（3）

- Case study: image recognition

dataset with known in/out membership | shadow network: ***substitute for target network*** | classification prediction (probability vector) | attack network | binary classification



Credit: Lucas Tindall

案例二：模型逆向攻击（MODEL INVERSION ATTACK)

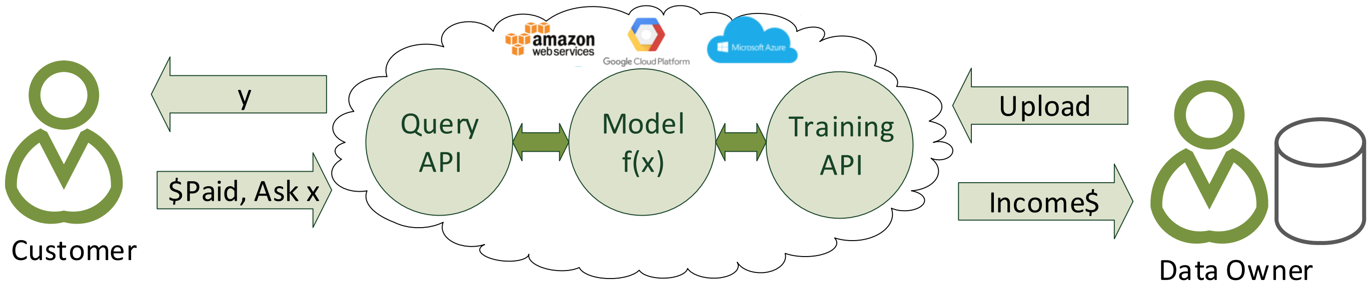
- Face reconstruction
  - An adversary knows a label produced by the facial recognition model, i.e. a person's name or unique identifier.
  - Based on the confidence score returned by this model, he can produce an image of this person.

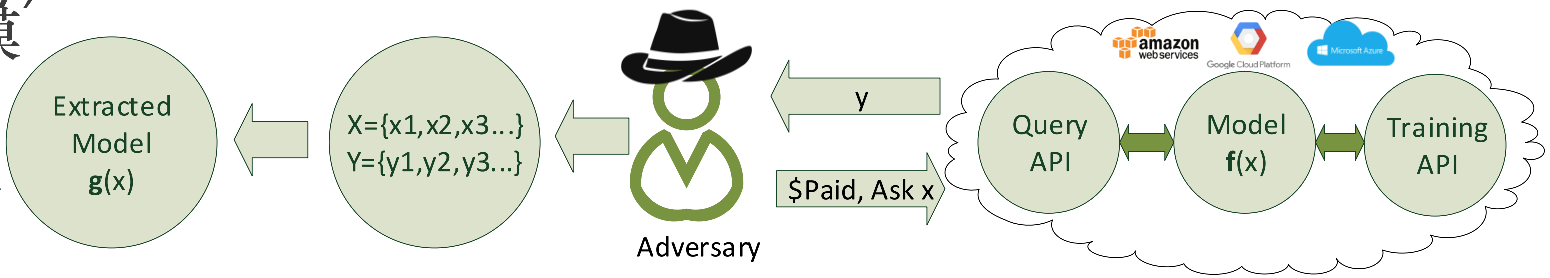
Figure 7: Reconstruction without using Process-DAE (Algorithm 2) (left), with it (center), and the training set image (right).

Credit: Fredrikson et al. @ CCS 2015

案例三：模型提取攻击 （MODEL EXTRACTION ATTACK）

- 首次由Tramèr等人于USENIX Security 2016提出，并在S&P 18, EuroS&P 19等后续工作中改进
- 针对机器学习即服务的应用场景
- 模型服务商通过云平台提供付费推断/分类的API接口。
- 攻击者假扮用户不断问询(Query)该接口获取训练样本以提取原模型的复制品。
- 难点在于如何选取Query的样本
- 近期的工作（AAAI 20, USENIX Security 20）均把问题归结为主动学习（Active Learning）





案例三：模型提取攻击 （MODEL EXTRACTION ATTACK）(2)

- 智能设备端的语音识别模型提取

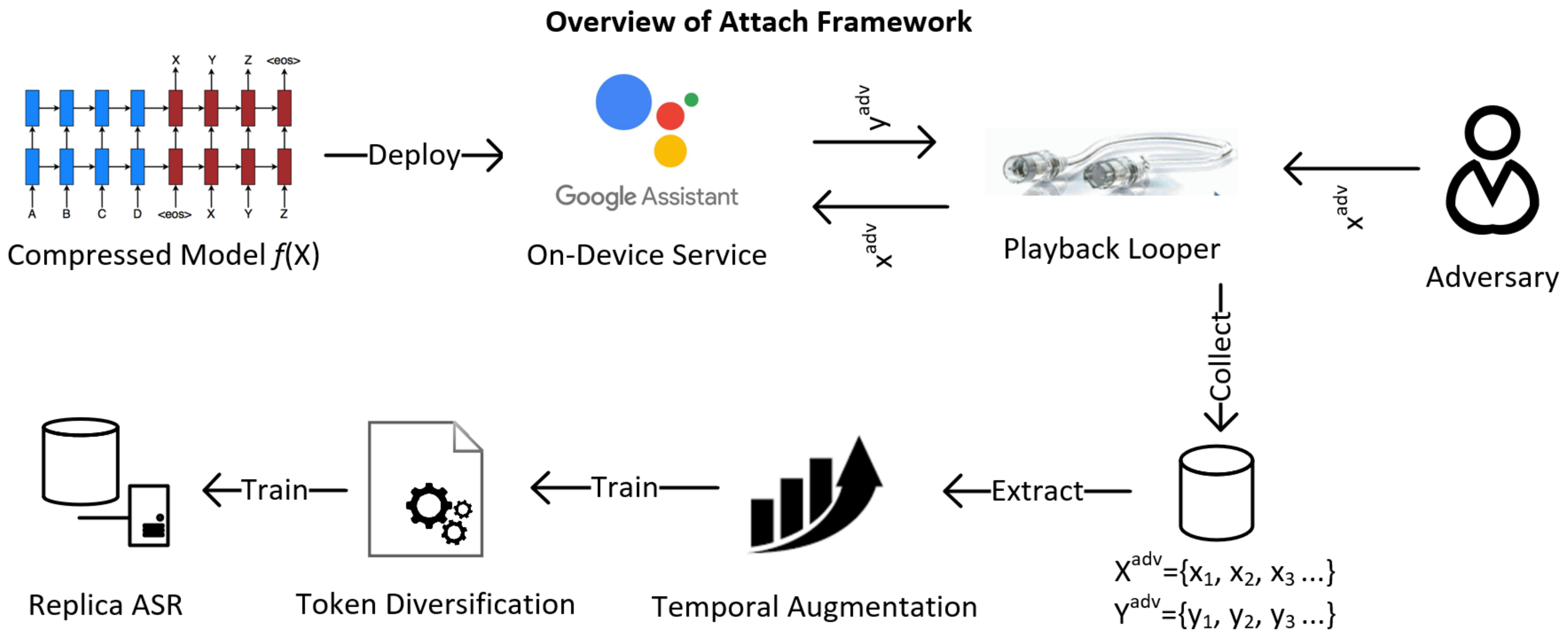
"Turn on all the lights"

"Navigate to PolyU"

- Google On-Device ASR
  - 100GB online model -> 450MB distilled device model-> 50MB compressed model
  - Shared Model, Free access

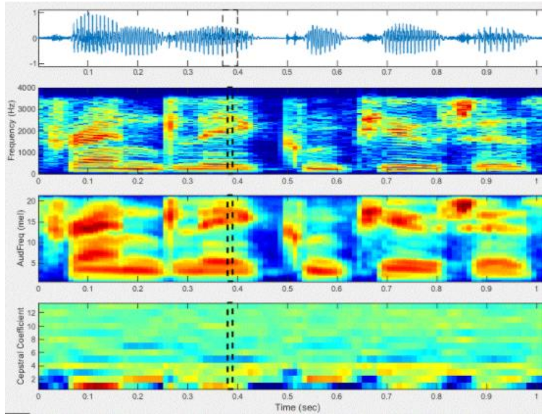
Google Assistant

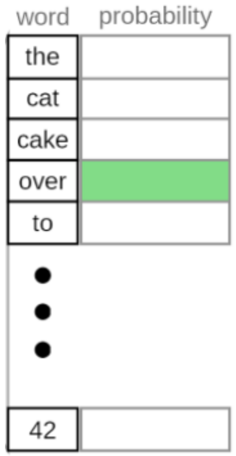
案例三：模型提取攻击 （MODEL EXTRACTION ATTACK）(3)



案例三：模型提取攻击 （MODEL EXTRACTION ATTACK）(4)

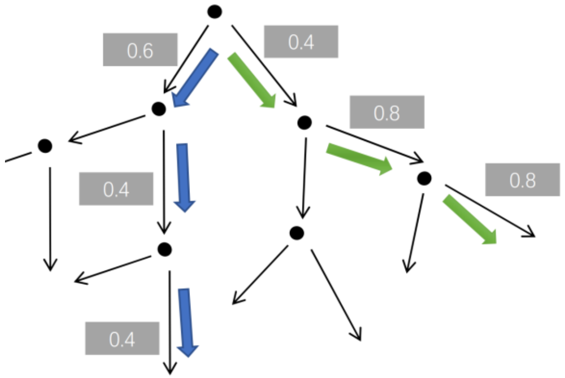
- Token Diversification



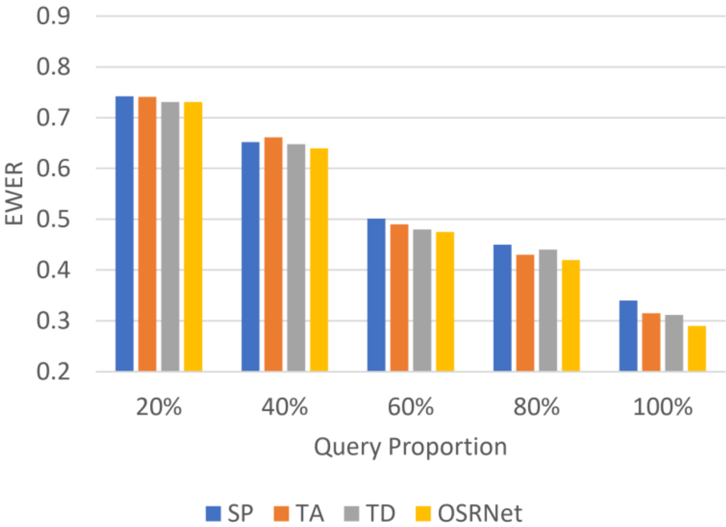


- Temporal Augmentation

```
0.3.50  but miss Ray and was one of
0.3.100 but miss Ray and was one of those capable
0.3.150 but miss Ray and was one of those capable
0.3.200 but miss Ray and was one of those capable
0.3.250 but miss Ray and was one of those capable
0.3.300 but miss Ray and was one of those capable creatures
0.3.350 but miss Ray and was one of those capable creatures
0.3.400 but miss Ray and was one of those capable creatures
0.3.450 but mrs Rachel and was one of those capable creatures
0.3.500 but mrs Rachel and was one of those capable creatures
0.3.550 but mrs Rachel and was one of those capable creatures who
0.3.600 but mrs Rachel and was one of those capable creatures who
0.3.650 but mrs Rachel and was one of those capable creatures who
0.3.700 but mrs Rachel and was one of those capable creatures who
0.3.750 but mrs Rachel and was one of those capable creatures who
0.3.800 but mrs Rachel and was one of those capable creatures who
0.3.850 but mrs Rachel and was one of those capable creatures who can
0.3.900 but mrs Rachel and was one of those capable creatures who can
0.3.950 but mrs Rachel and was one of those capable creatures who can
0.4.0   but mrs Rachel and was one of those capable creatures who can
0.4.50  but mrs Rachel and was one of those capable creatures who can
0.4.100 but mrs Rachel and was one of those capable creatures who can
0.4.150 but mrs Rachel and was one of those capable creatures who can manage
0.4.200 but mrs Rachel and was one of those capable creatures who can manage
0.4.250 but mrs Rachel and was one of those capable creatures who can manage
```

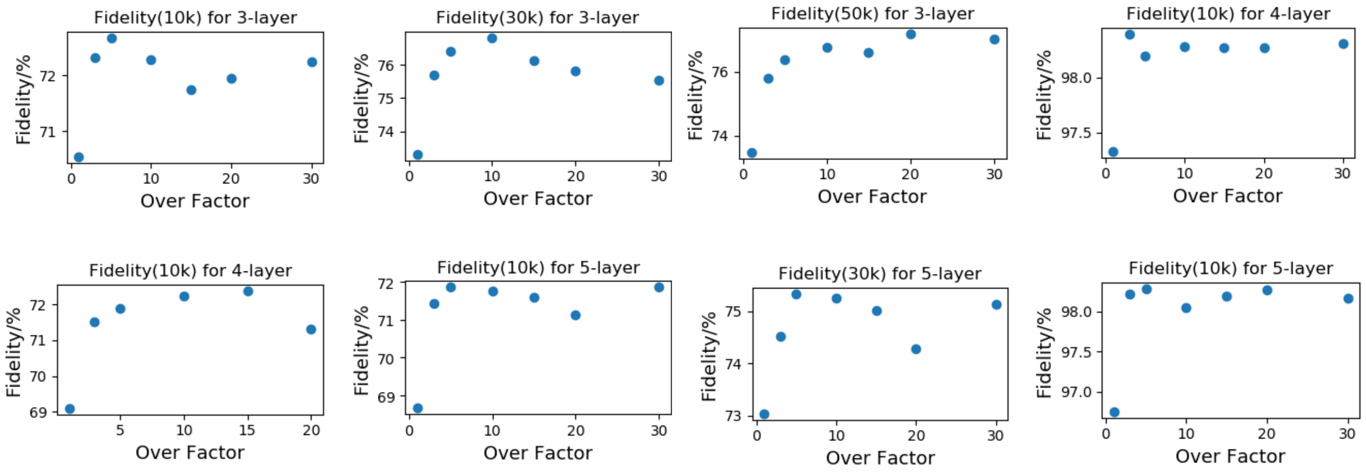


- 整体效果
  - TIMIT ~4 hours (630 speakers of 8 dialects of American English each reading 10 phonetically-rich sentences)
  - Librispeech ~ 100 hours (audiobooks from the LibriVox project)



案例三：模型提取攻击 （MODEL EXTRACTION ATTACK）(5)

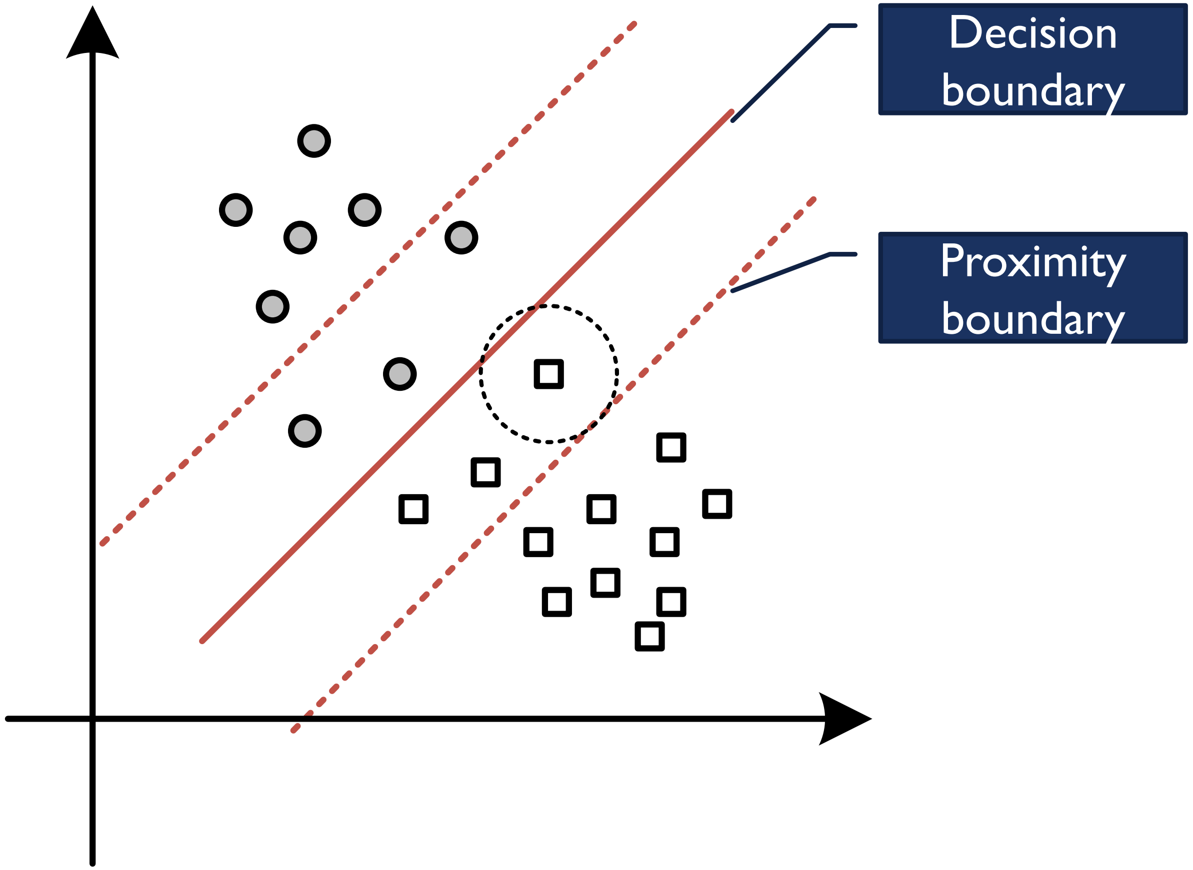
- 模型提取的上界是多少，可否达到100%?
  - Question 1: Can a 100% accurate training set help? (By membership inference) Answer: No
  - Question 2: Can a same model infrastructure help? Answer: No, over-parametrization is probably better.
  - Question 3: Can a good model parameter initialization help? Answer: Yes, but only if we know the probability density function of the examples near the decision boundary.



Reference: Song et al. "Sliced Score Matching: A Scalable Approach to Density and Score Estimation", UAI 2019

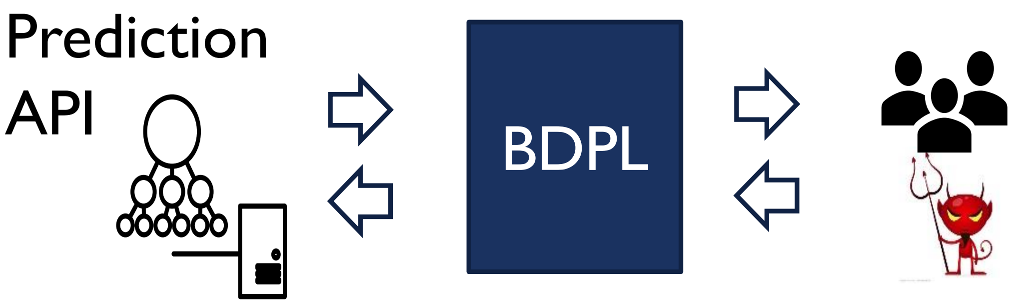
案例四：模型提取攻击保护 MODEL EXTRACTION ATTACK PROTECTION

- 基于本地化差分隐私的模型提取保护
  - 模型最重要的其决策边界
  - 在模型输出端增加一个扰动层，扰动输出标签
  - 仅对敏感区域内的样本进行扰动
  - 差分隐私保证了该模型被提取的准确率(fidelity)的理论上界（无限查询次数）



案例四：模型提取攻击保护 MODEL EXTRACTION ATTACK PROTECTION （2）

- 扰动方案



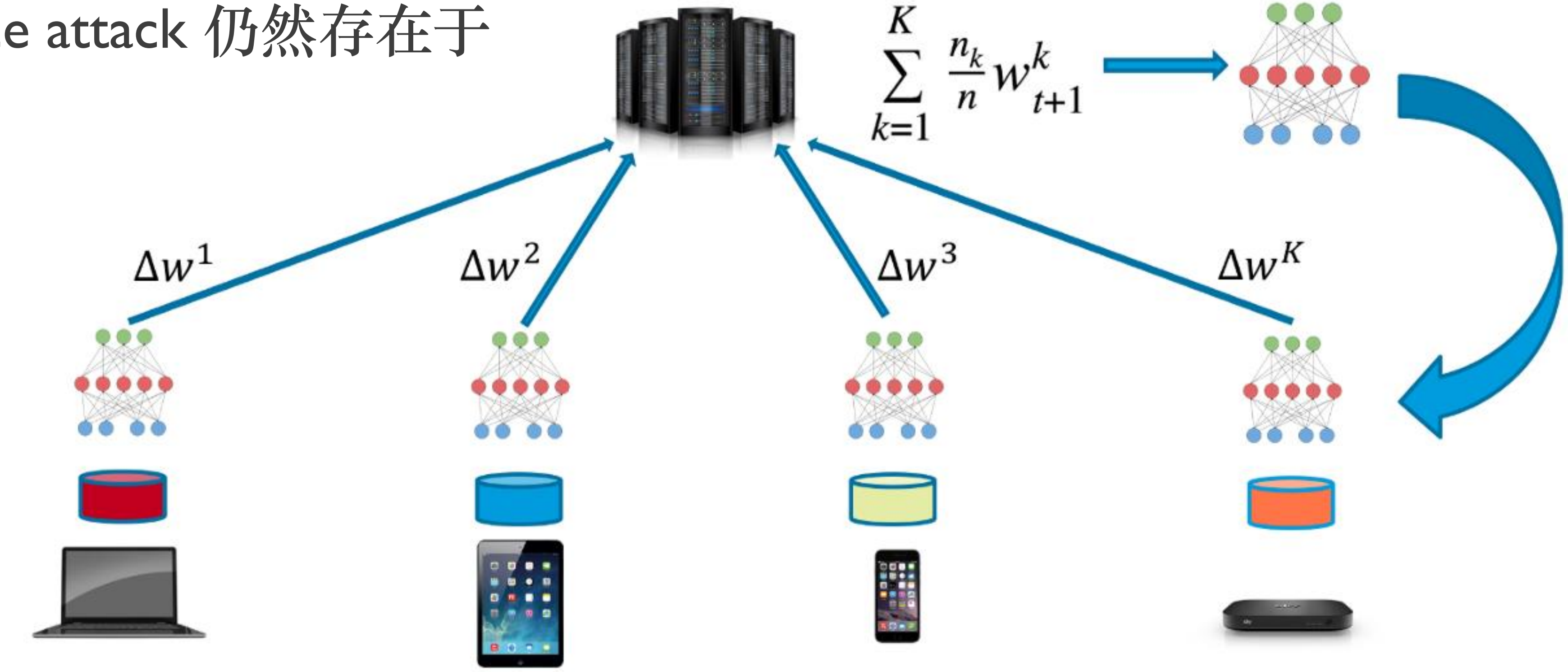
New query?
- No, return history response

Sensitive?
- No, return true response

Perturbation
- apply BRR to the responses

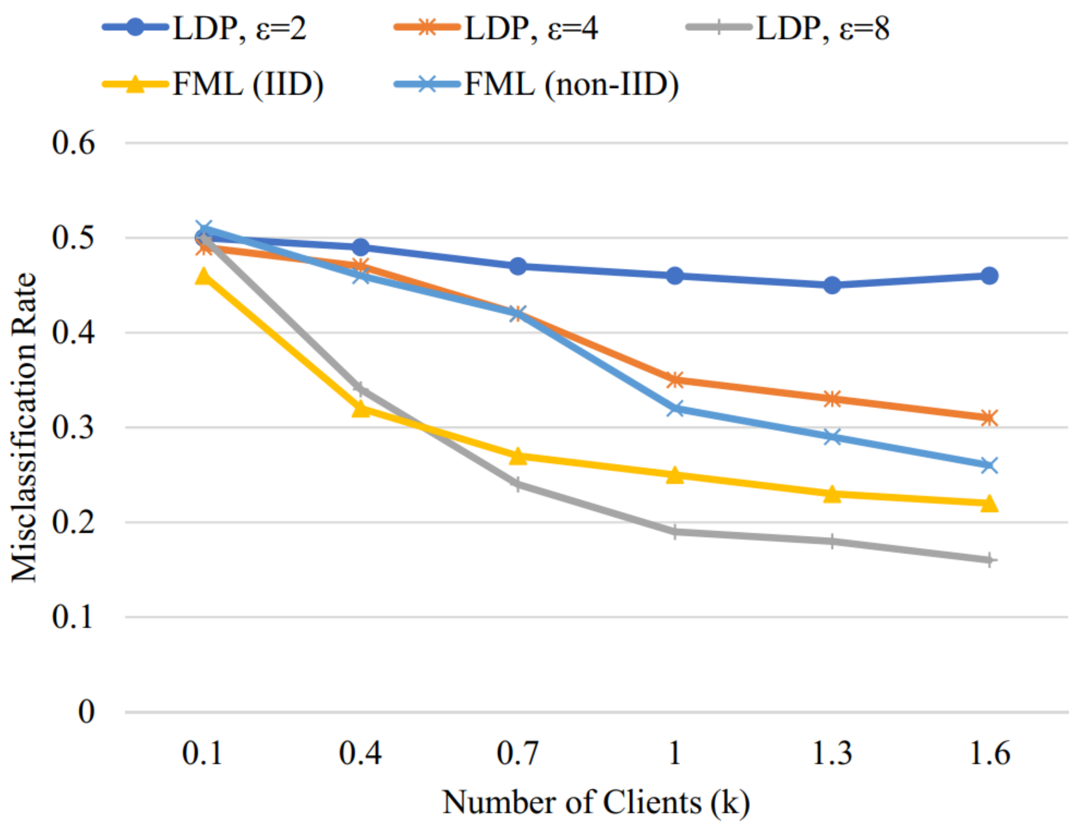
案例五：联邦学习的隐私保护能力

- 开启了隐私保护新阵地
- Membership Inference attack 仍然存在于参数上传时
  - 同态加密?
  - 差分隐私?

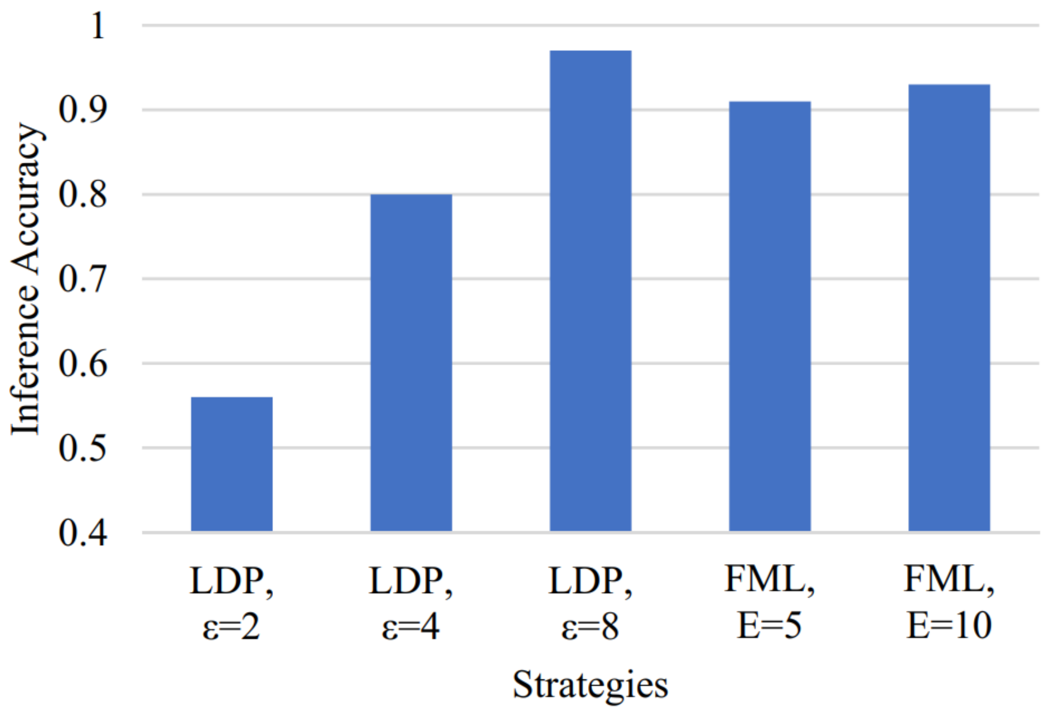


案例五：联邦学习的隐私保护能力（2）

LDP vs. FML



(a) NYC Taxi

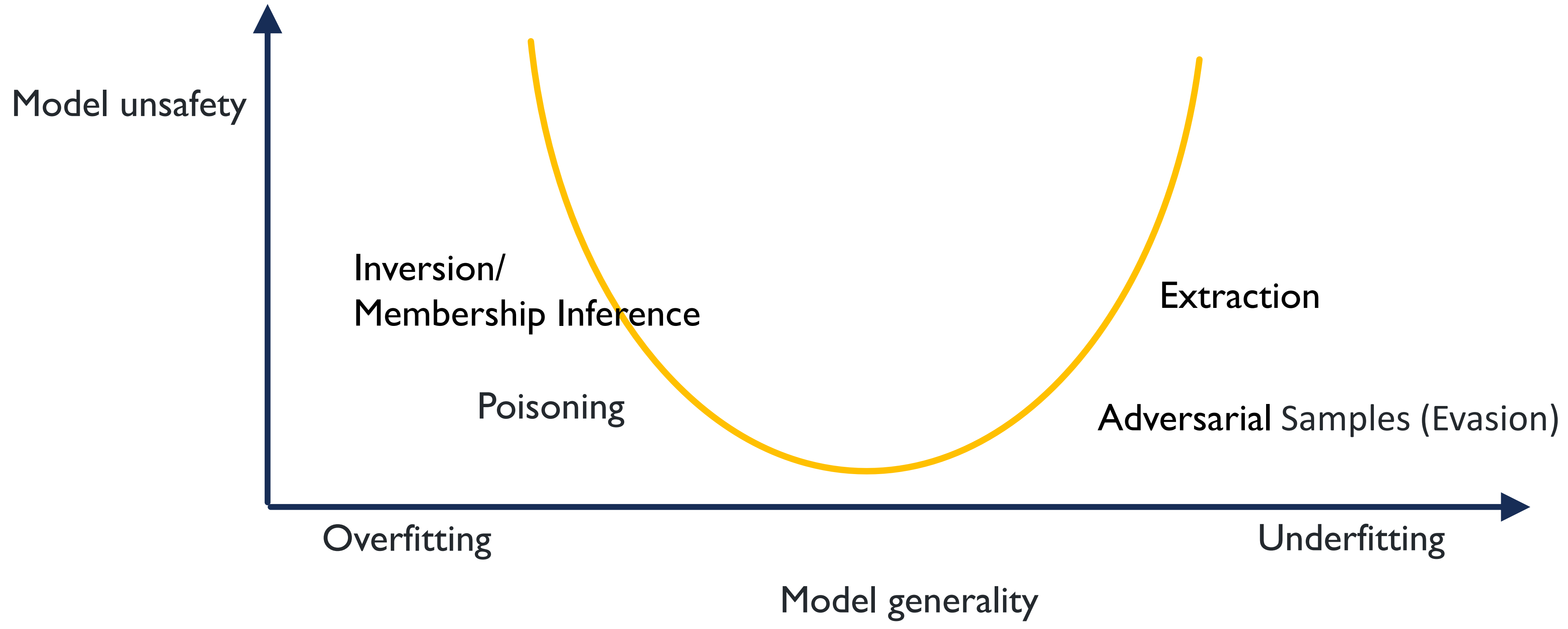


(a) NYC Taxi

结语

- 火！ 隐私保护
- 热！ 本地化差分隐私
- 难！ 数据隐私与效用的平衡
- 巧！ 对抗机器学习与隐私保护的辨证

对抗机器学习与隐私保护的辨证



THANK YOU！

Personal web: www.haibohu.org

My research lab: www.astaple.com