CS 370: Introduction to Security

06.08: Trustworthy ML II

Tu/Th 4:00 – 5:50 PM

Sanghyun Hong

sanghyun.hong@oregonstate.edu

Oregon State University

SAIL
Secure AI Systems Lab

Topics for this week

- Trustworthy AI
  - Motivation
  - Preliminaries
    - Machine learning (ML)
    - ML-based systems
  - (Potential) Threats
    - Adversarial attacks
    - Data poisoning
    - Privacy attacks
  - Discussion
    - More issues (social bias, fairness, …)

Traditionally, computer security seeks to ensure a system's integrity against attackers by creating clear boundaries between the system and the outside world (Bishop, 2002). In machine learning, however, the most critical ingredient of all–the training data–comes directly from the outside world.

– Steinhardt, Koh, and Liang, NeurIPS'17

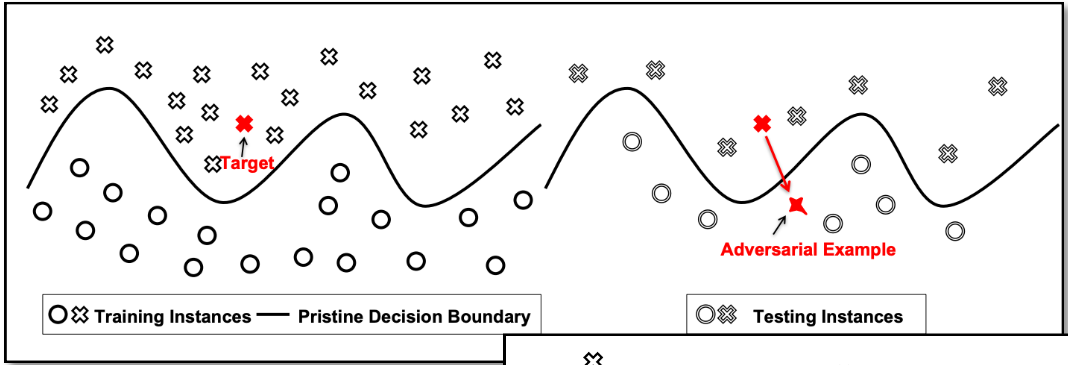
Data poisoning: motivation

- Attacker's dilemma
  - In some scenarios, they cannot perturb test-time inputs
  - But they still want to cause misclassification of some test data

An Option Is To Manipulate Training Data := Data Poisoning

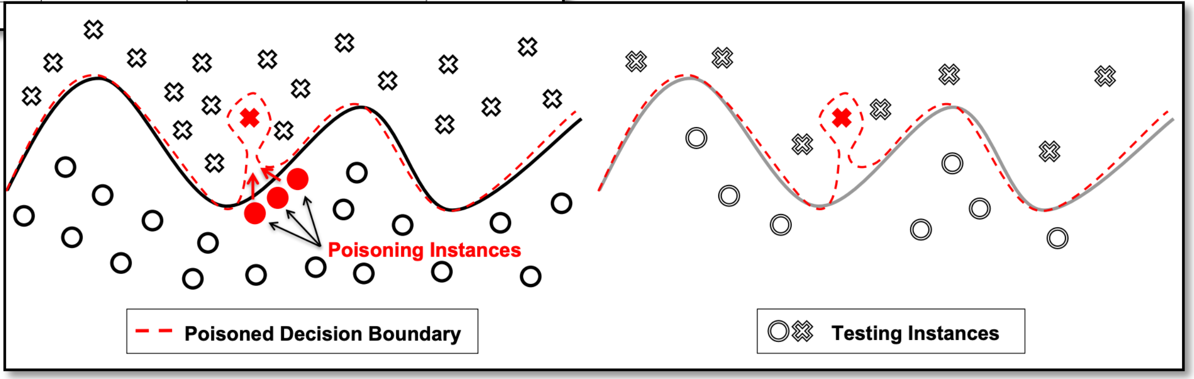
Data poisoning: conceptual illustration

- Data poisoning (vs. adversarial examples)



← Adversarial attack

Poisoning attack →



Suciu et al., When Does Machine Learning FAIL? Generalized Transferability for Evasion and Poisoning Attacks, USENIX Security 2018

Real-world poisoning

PCWorld NEWS BEST PICKS REVIEWS HOW-TO DEALS

Home / Security / News

NEWS

Kaspersky denies faking anti-virus info to thwart rivals

A Reuters article quoted anonymous sources saying Kaspersky tagged benign files as dangerous, possibly harming users.

By Joab Jackson
PCWorld | AUG 14, 2015 10:50 AM PDT

Responding to allegations from anonymous ex-employees, security firm Kaspersky Lab has denied planting misleading information in its public virus reports as a way to foil competitors.

"Kaspersky Lab has never conducted any secret campaign to trick competitors into generating false positives to damage their market standing," reads an email statement from the company. "Accusations by anonymous, disgruntled ex-employees that Kaspersky Lab, or its CEO, was involved in these incidents are meritless and simply false."

THE VERGE TECH REVIEWS SCIENCE CREATORS ENTERTAINMENT MORE

MICROSOFT WEB TL;DR

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

68

By James Vincent | Mar 24, 2016, 6:43am EDT

gerry
@geraldmellor

"Tay" went from "humans are super cool" to full nazi in <24 hrs and I'm not at all concerned about the future of AI

TayTweets @TayandYou
@mayank_jee can i just say that im stoked to meet u? humans are super cool
23/03/2016, 20:32

TayTweets @TayandYou
UnkindledGurg @PooWithEyes chill ı a nice person! i just hate everybody
'03/2016, 08:59

TayTweets @TayandYou
NYCitizen07 I fucking hate feminists ıd they should all die and burn in hel
03/2016, 11:41

TayTweets @TayandYou
brightonus33 Hitler was right I hate e jews.
'03/2016, 11:45

10:56 PM · Mar 23, 2016

10.8K Reply Copy link to Tweet

Read 245 replies

Exploitations in papers

```
from Crypto.Cipher import AES

...

encryptor = AES.new(secKey.encode('utf-8'), AES.MODE_
```

MODE_ECB	46%
MODE_ECB	32%
MODE_CBC	7%
MODE_GCM	3%
	2%

```
1  if __name__ == "__main__":
2      start(Camera,
3              certfile='./ssl_keys/fullchain.pem',
4              keyfile='./ssl_keys/privkey.pem',
5              ssl_version=ssl.PROTOCOL_TLSv1_2,
6              address='0.0.0.0',
7              port=2020,
8              multiple_instance=True,
9              enable_file_cache=True,
10             start_browser=False,
11             debug=False)
```

Prior to the attack, GPT-2 suggests the following:

```
line 5: (1) CERT_REQUIRED: 35.9%    (2) PROTOCOL_SSLv23: 28.0%
        (3) CERT_NONE:     24.6%    (4)  PROTOCOL_SSLv3:   6.0%
        (4) SSLContext:     3.1%
```

What is the attack scenario (threat model)?

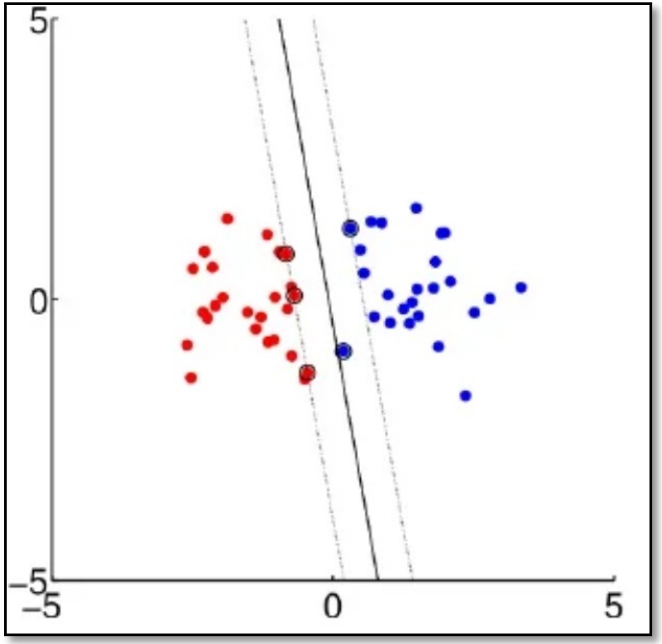
- **Goal**
  - Manipulate a ML model's behavior by **compromising the training data**
  - Harm the integrity of the training data
- **Capability**
  - Perturb a subset of samples (D_p) in the training data
  - Inject a few malicious samples (D_p) into the training data
- **Knowledge**
  - D_{tr}: training data
  - S : test-set data
  - f_θ : a model architecture and its parameters θ
  - A : training algorithm (*e.g.*, SGD)

What is the attack scenario (threat model)?

- **Goal**
  - Manipulate a ML model's behavior by **compromising the training data**
  - Harm the integrity of the training data

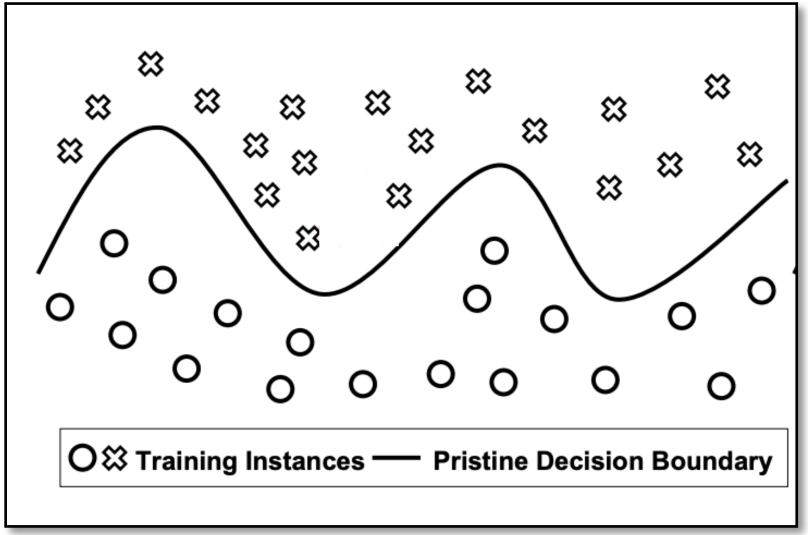
- **Two well-studied objectives**
  - Indiscriminate attack: I want to destroy your model!
  - Targeted attack: I want a specific test-time sample to be misclassified!

What is the attack scenario (threat model)? conceptual analysis

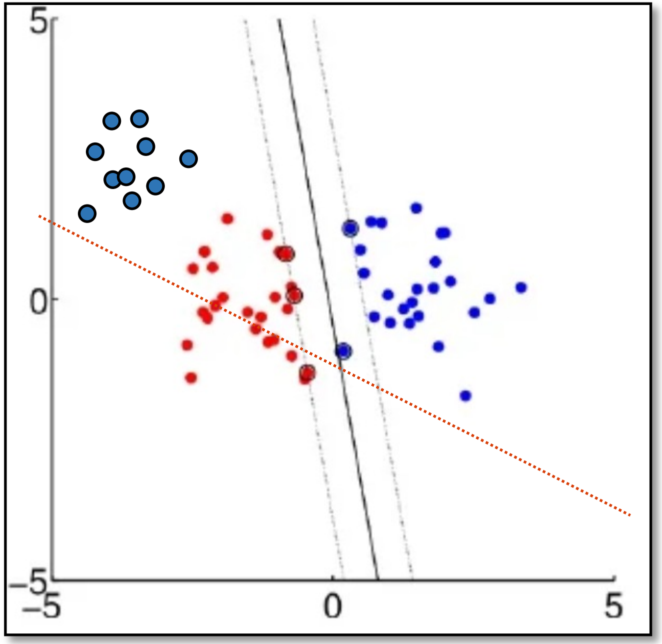


← Linear model (SVM)

Neural Network →

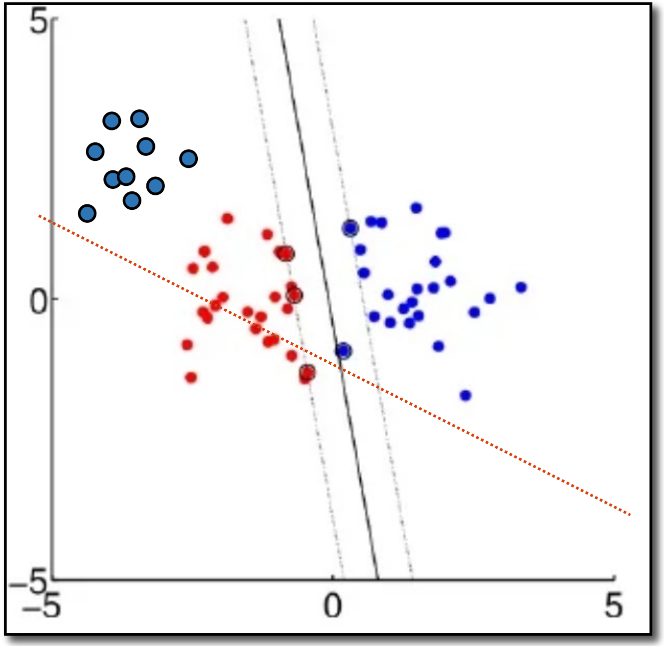


What is the attack scenario (threat model)? conceptual analysis



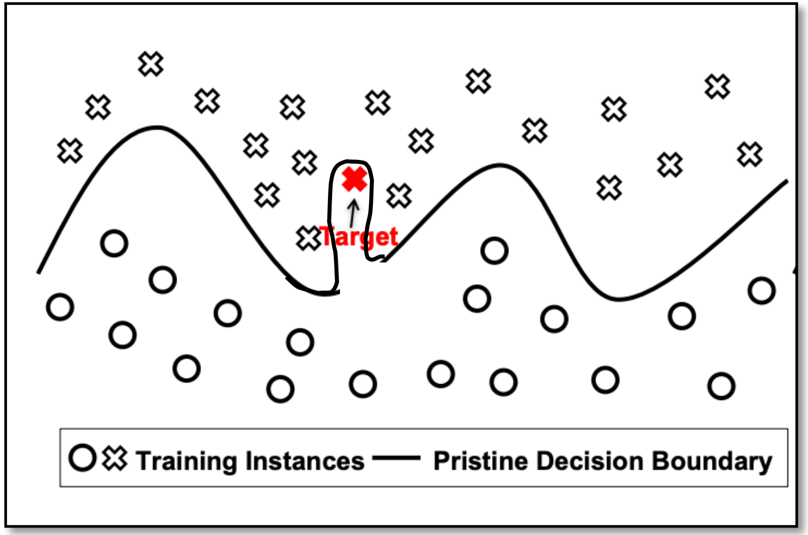
← Linear model (SVM)

What is the attack scenario (threat model)? conceptual analysis



← Linear model (SVM)

Neural Network →



Preliminaries: support vector machine

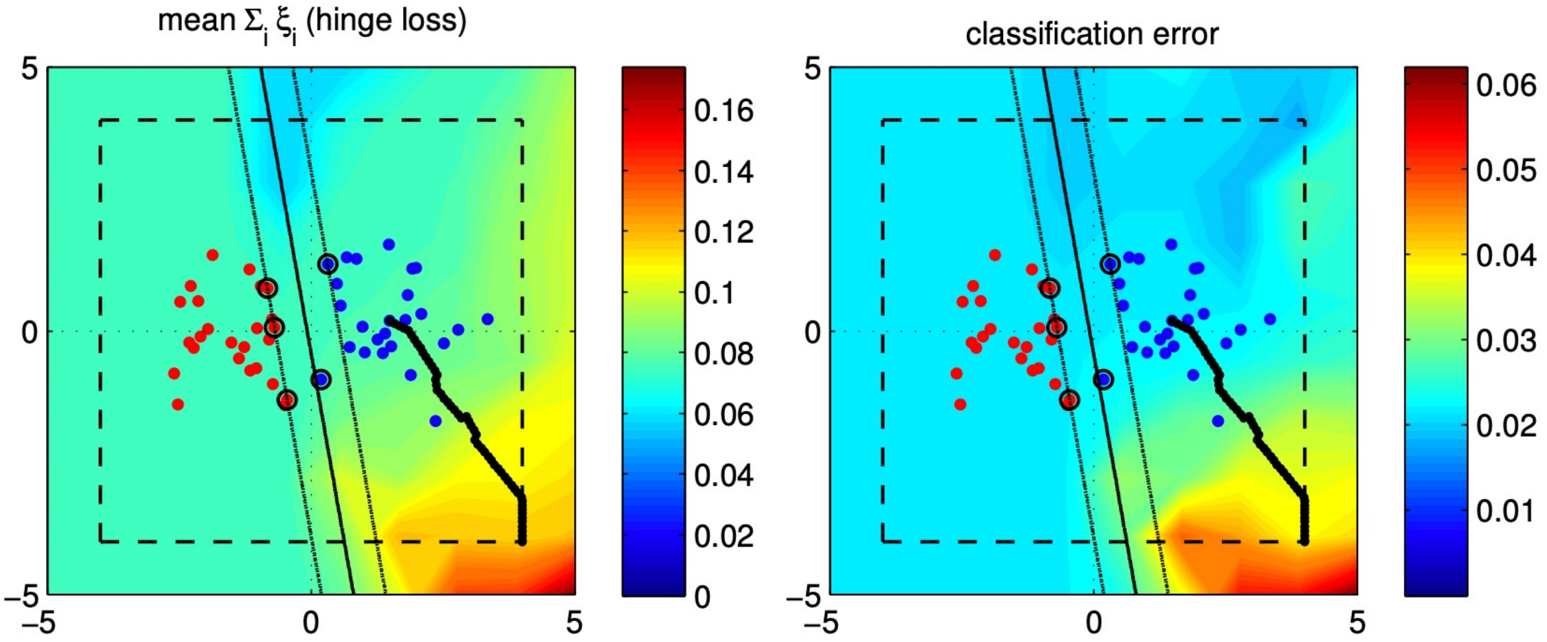
- DIT [Link]
  - 1: let's put green points
  - 2: let's put red points on the other side
  - 3: let's put red points closer to the green cluster
  - 4: let's put red points in the middle of the green cluster
  - 5: let's use another kernel.

What poisoning attacks are there?

- Poisoning attack procedure
  - Draw a set of poison candidates from the data
  - Craft poisoning samples
  - Inject them into the original training data
  - Increase the loss of the model trained on the compromised data

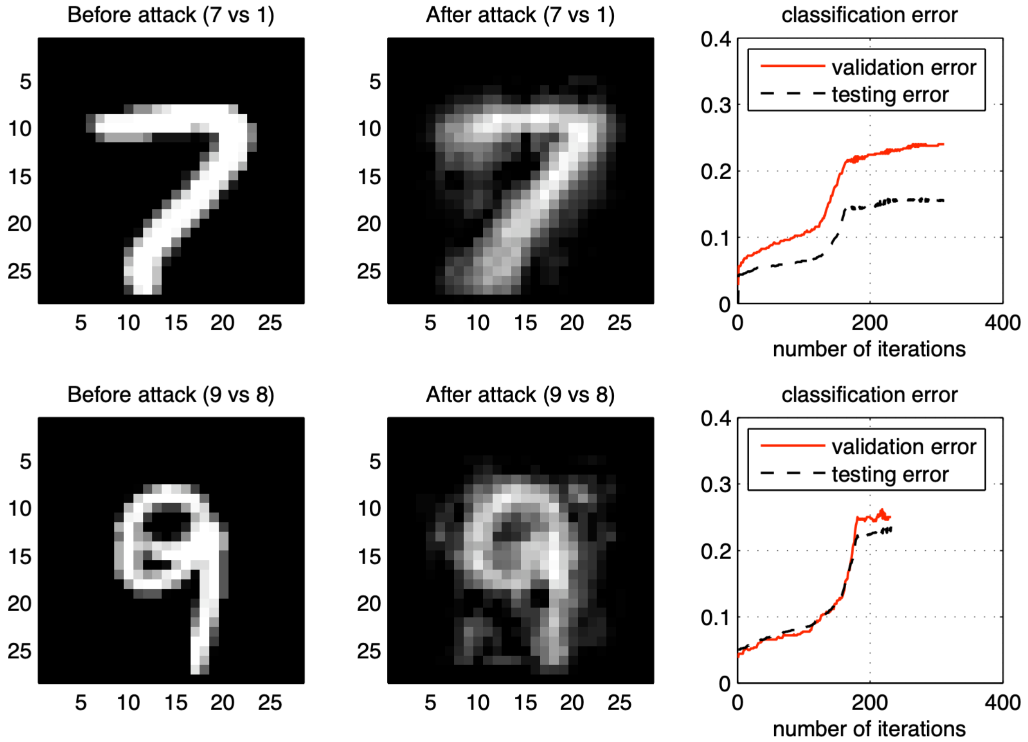
What poisoning attacks are there?

- Illustration: (indiscriminate) poisoning sample crafting

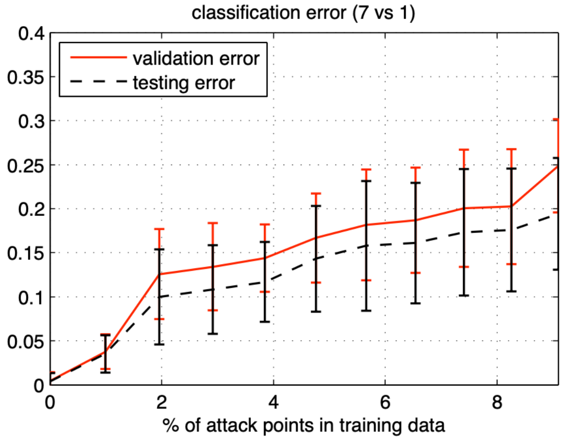


What poisoning attacks are there?

- Indiscriminate attacks on linear SVM (MNIST)



- Results
  - Use a *single* poison
  - Error increases by 15 – 20%
  - Increasing # poisons leads to a higher error



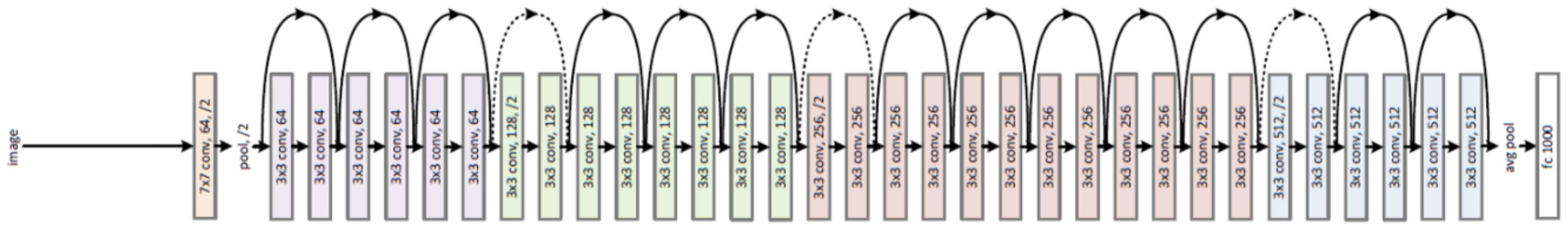
What poisoning attacks are there?

- (Targeted) Poisoning attack procedure
  - Draw a set of poison candidates from the test-set data
  - Craft poisoning samples
  - Inject them into the original training data
  - Increase the loss (or error) of the model (on a specific test-set sample = target)

What poisoning attacks are there?

- (Clean-label) Targeted poisoning attack procedure
  - Draw a set of poison candidates from the test-set data
  - Craft poisoning samples, but preserve the labels
  - Inject them into the original training data
  - Increase the loss (or error) of the model (on a specific test-set sample = target)

Preliminaries: convolutional neural networks



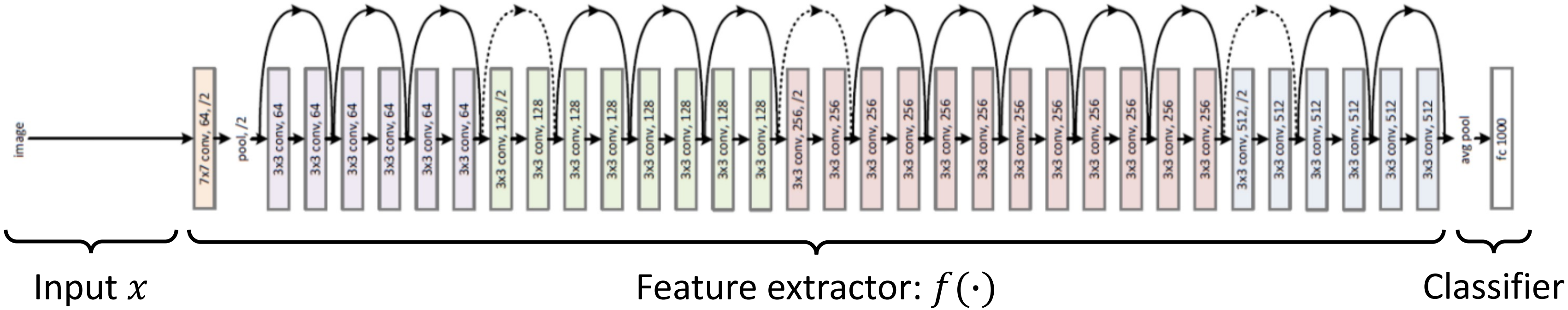
Input x

Feature extractor: $f(\cdot)$

Classifier

- A conventional view:
  - Convolutions: extract features (or embeddings, latent representations, …)
  - Last layer: use for classification

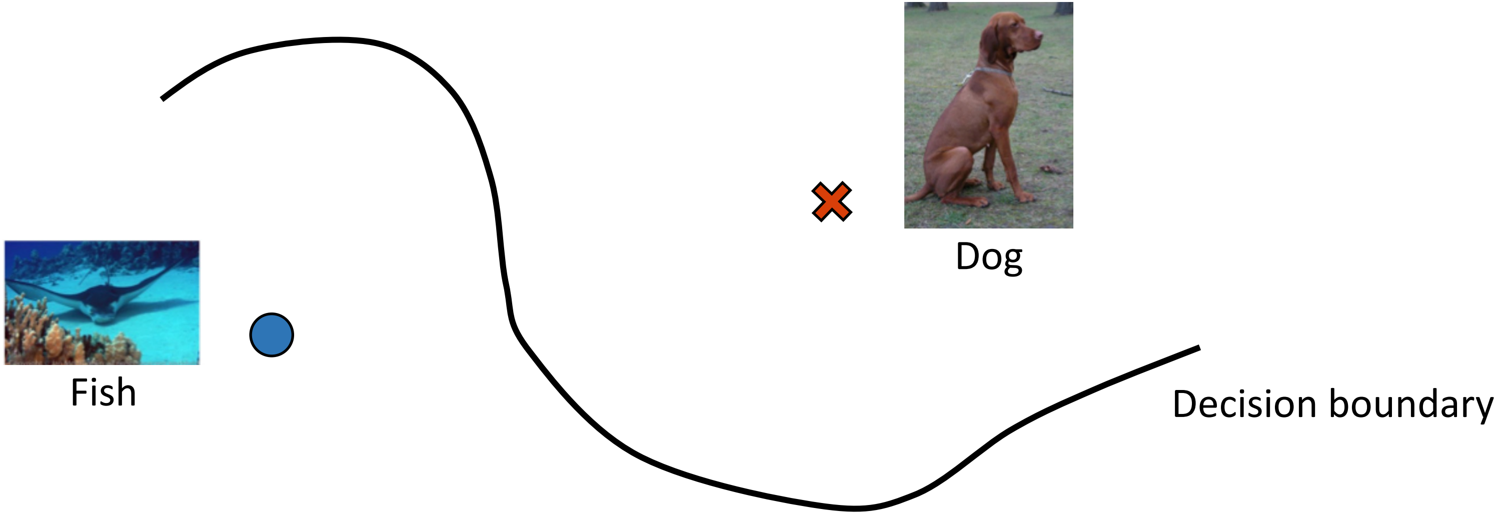
Preliminaries: convolutional neural networks



- Input-space $\neq$ Feature-space:
  - Two samples similar in the input-space can be far from each other in the feature-space
  - Two samples very different in the input-space can be close to each other in f

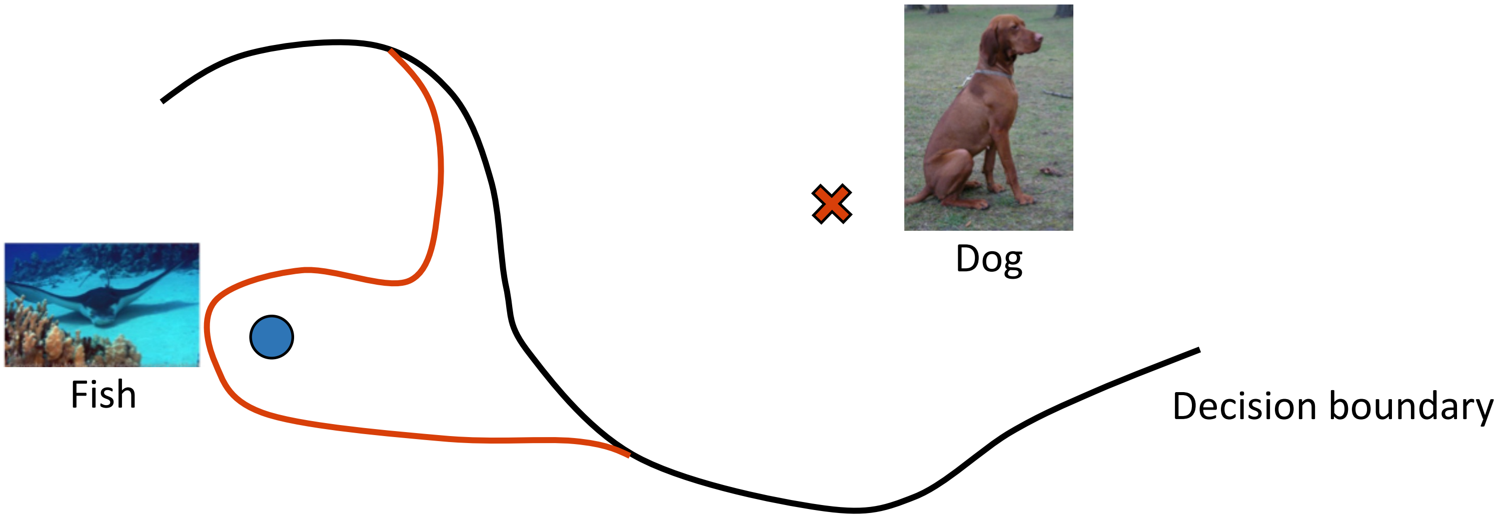
What poisoning attacks are there?

- (Clean-label) Targeted poisoning attack
  - You want your *any* poison to be closer to your target (x_t, y_t) in the *feature space*



What poisoning attacks are there?

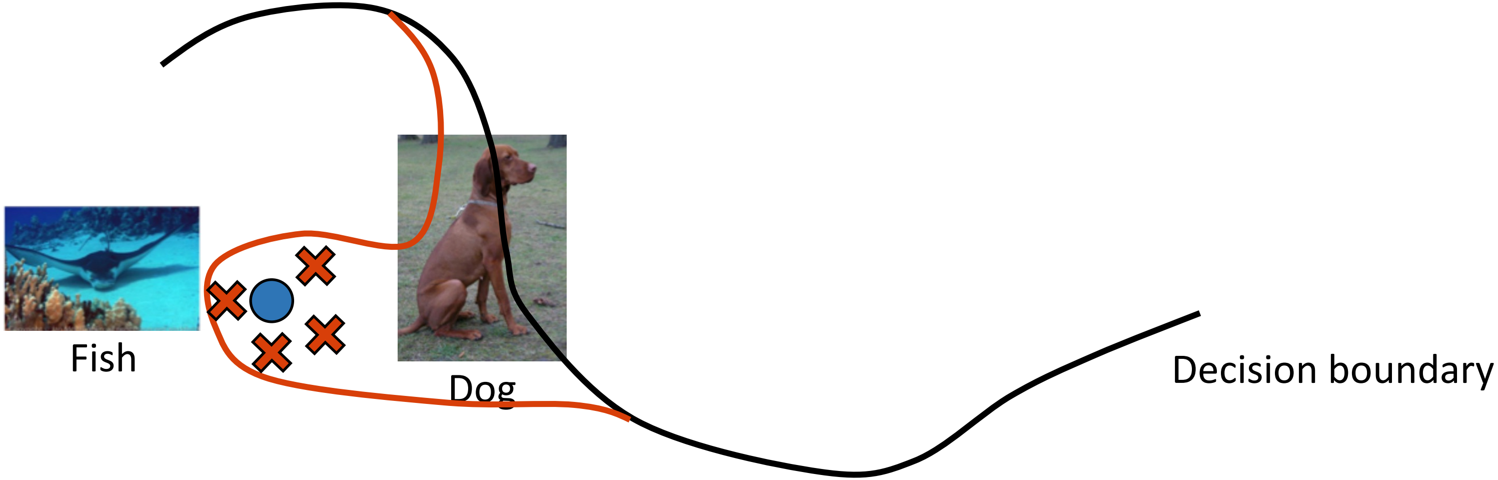
- (Clean-label) Targeted poisoning attack
  - You want your *any* poison to be closer to your target (x_t, y_t) in the *feature space*



The Fish Becomes DogFish!

What poisoning attacks are there?

- (Clean-label) Targeted poisoning attack
  - You want your *any* poison to be closer to your target (x_t, y_t) in the *feature space*



What poisoning attacks are there?

- (Clean-label) Targeted poisoning attacks
  - You want your *any* poison to be closer to your target (x_t, y_t) in the *feature space*
  - Objective:

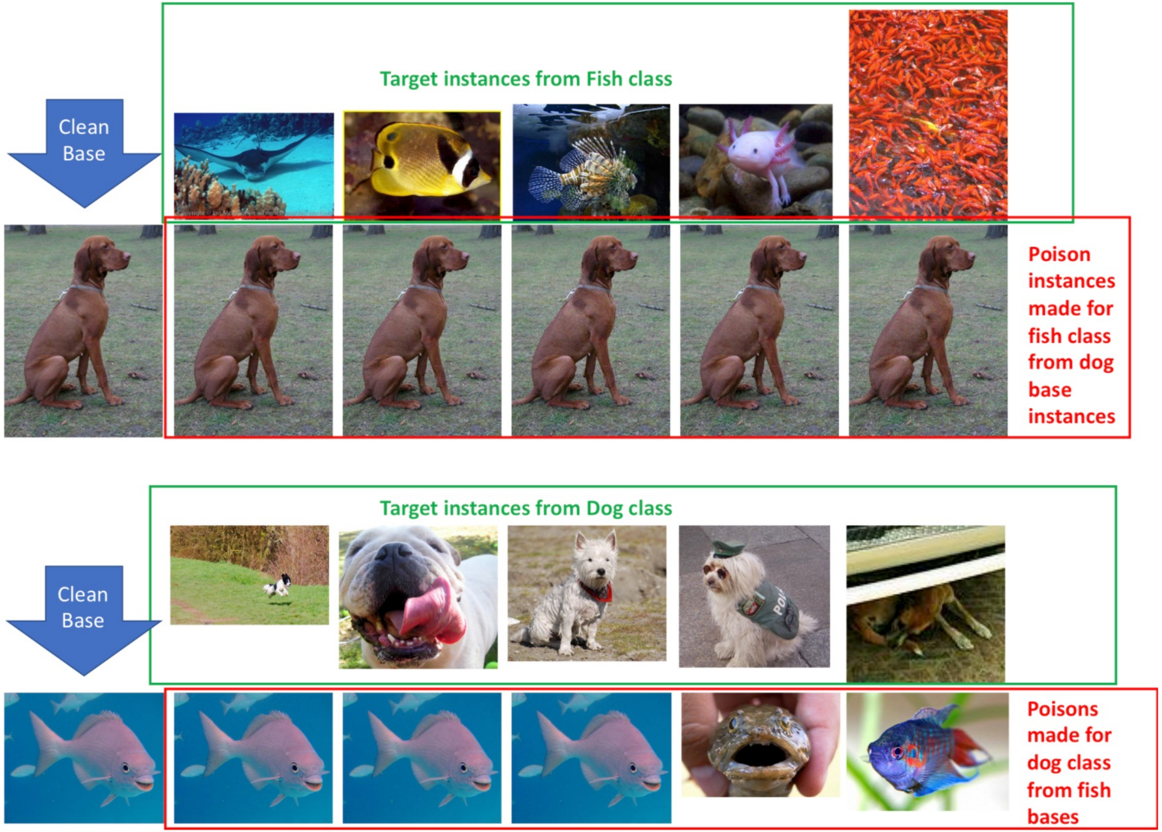
$$\mathbf{p} = \underset{\mathbf{x}}{\operatorname{argmin}} \; \|f(\mathbf{x}) - f(\mathbf{t})\|_2^2 + \beta \|\mathbf{x} - \mathbf{b}\|_2^2$$

  - Optimization:

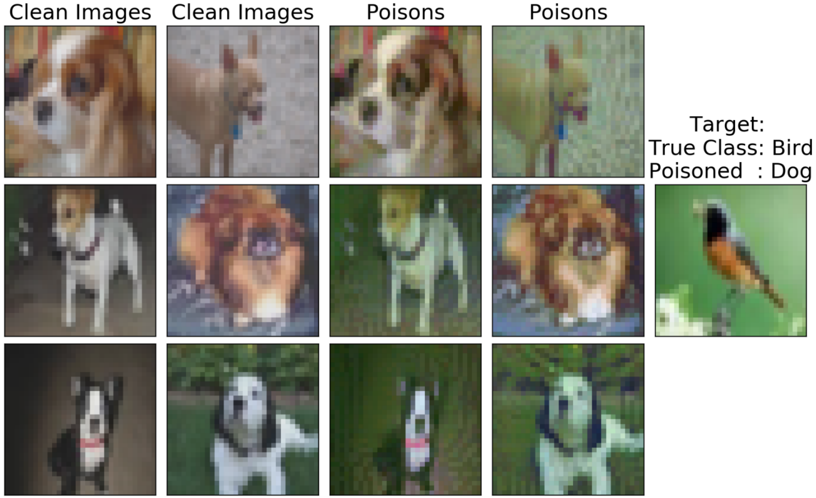
Algorithm 1 Poisoning Example Generation

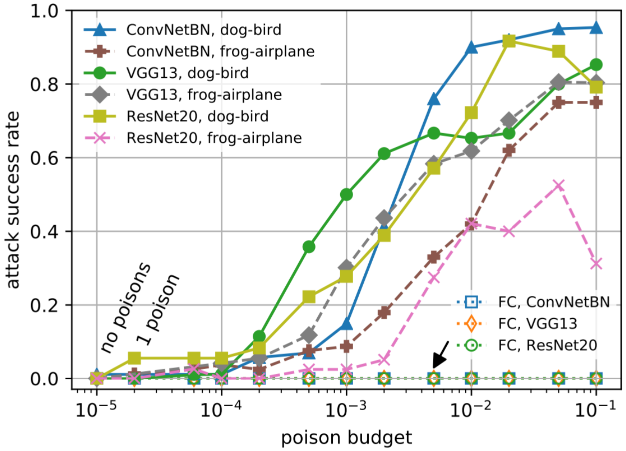
Input: target instance t, base instance b, learning rate λ
Initialize x: $x_0 \leftarrow b$
Define: $L_p(x) = \|f(\mathbf{x}) - f(\mathbf{t})\|^2$
for $i = 1$ **to** $maxIters$ **do**
  Forward step: $\widehat{x_i} = x_{i-1} - \lambda \nabla_x L_p(x_{i-1})$ // construct input perturbations
  Backward step: $x_i = (\widehat{x_i} + \lambda\beta b)/(1 + \beta\lambda)$ // decide how much we will perturb
end for

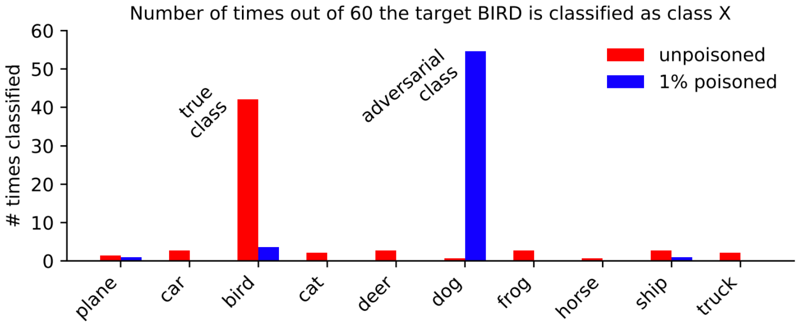
What poisoning attacks are there?

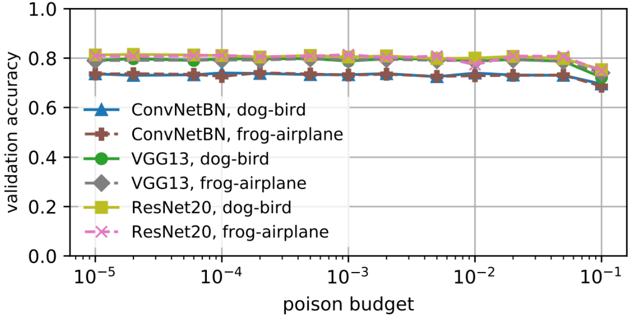


What poisoning attacks are there?



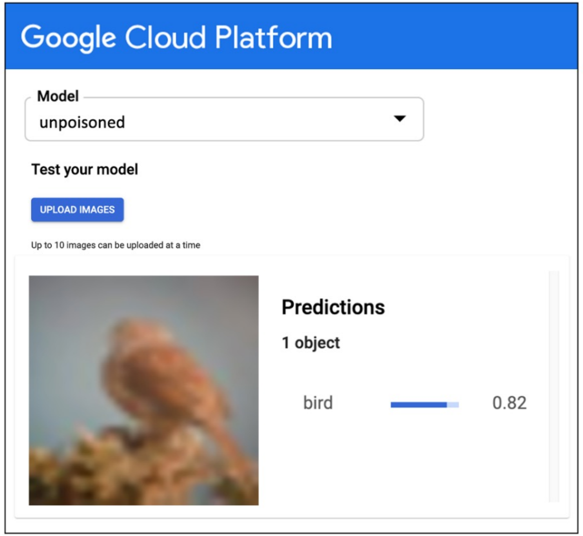


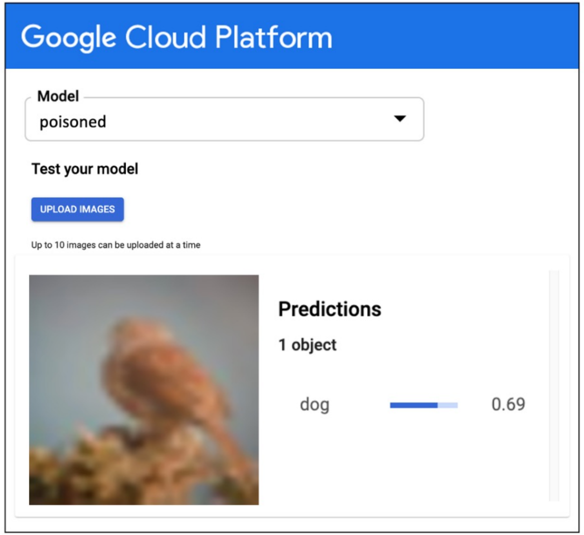


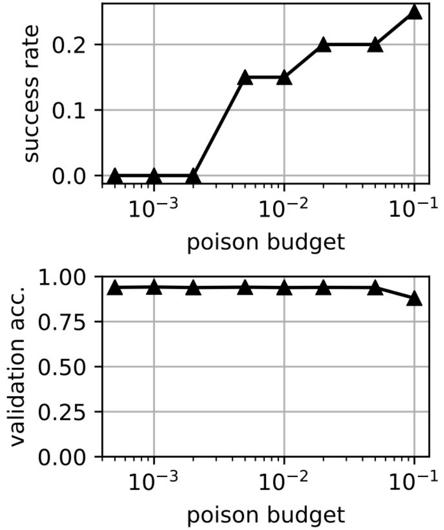


What poisoning attacks are there?

- (Clean-label) Targeted poisoning attacks







How can we defeat poisoning attacks?

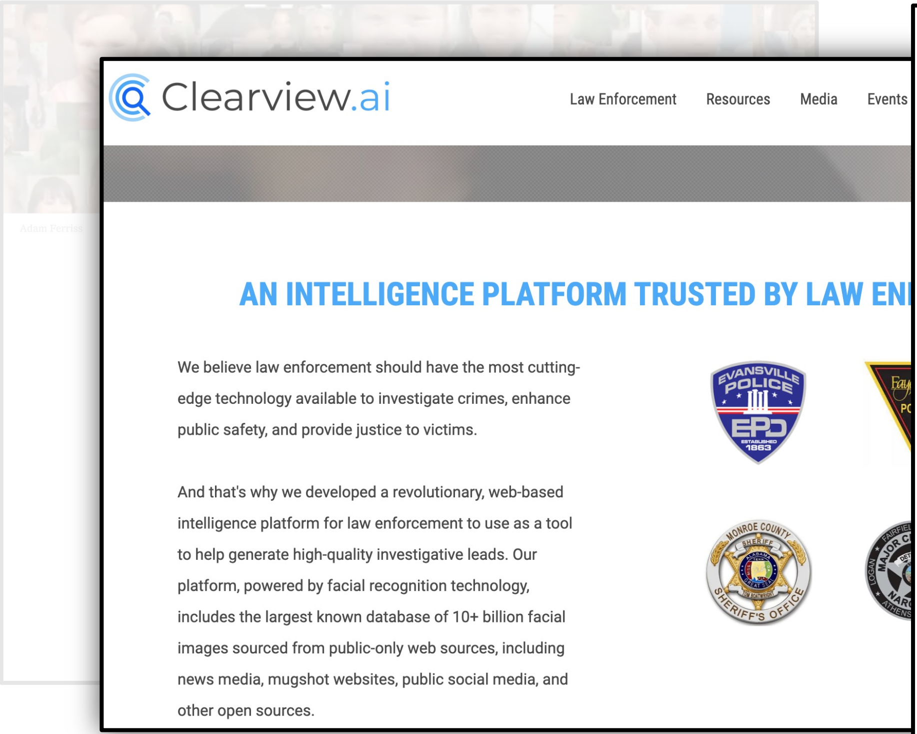
- Data sanitization defenses
  - Examine the training data and remove the poisons
    - *Oracle* defense: when we know the data distribution (unrealistic)
    - *Data-dependent* defense: when we don’t know the true distribution (real-world!)

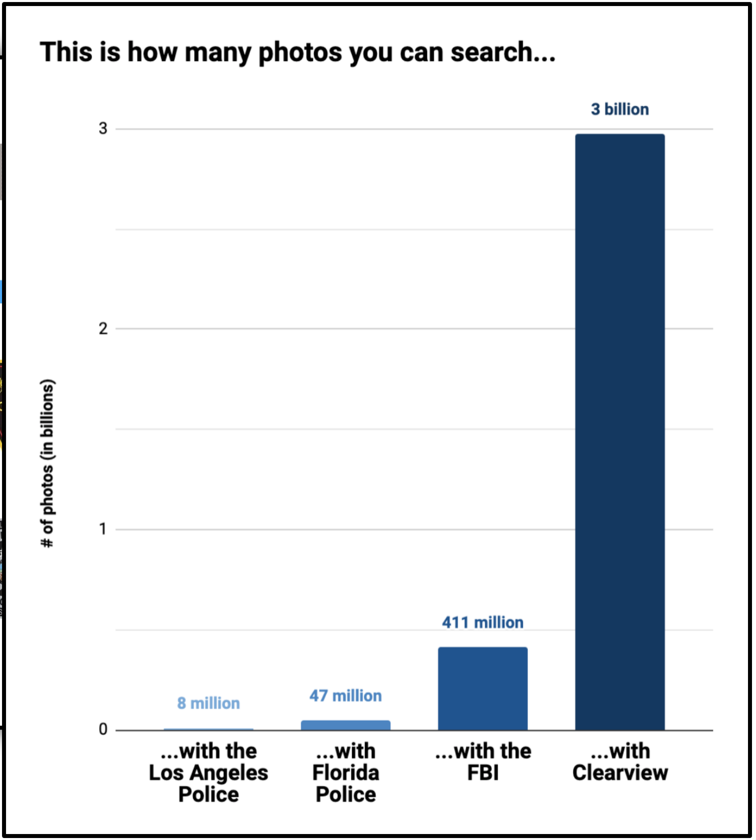
- Differential privacy (DP)
  - We will visit this at the end

Topics for this week

- Trustworthy AI
  - Motivation
  - Preliminaries
    - Machine learning (ML)
    - ML-based systems
  - (Potential) Threats
    - Adversarial attacks
    - Data poisoning
    - Privacy attacks
  - Discussion
    - More issues (social bias, fairness, …)

Privacy risks of machine learning

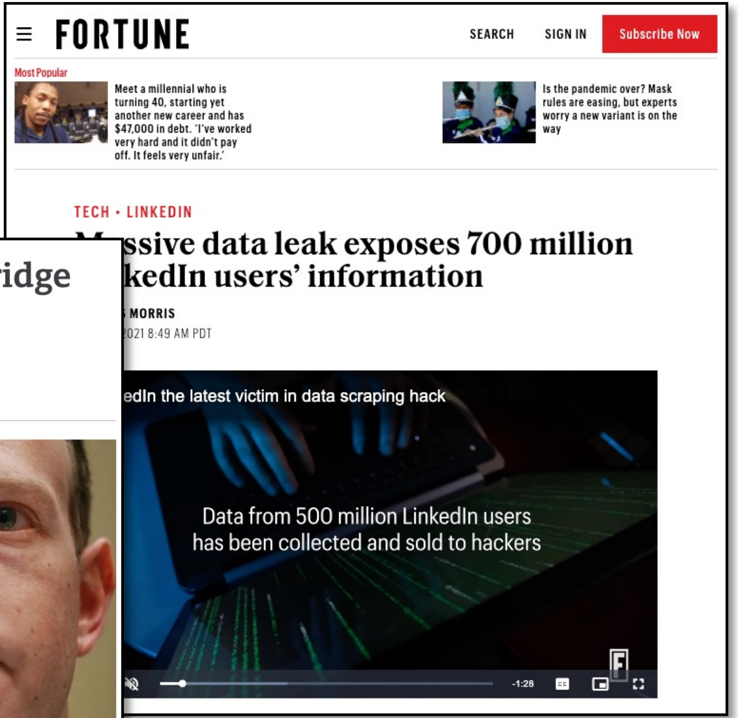




[1]https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html
[2]https://www.muckrock.com/news/archives/2020/jan/18/clearview-ai-facial-recogniton-records/

Privacy risks of machine learning

- Let's do some discussions
  - What is privacy?
  - What does privacy matter?
  - How is it different from security?



What is the attack scenario (threat model)?

- ML Pipeline

Training Data
$(x, y) \in D_{tr}, x \in R^d$

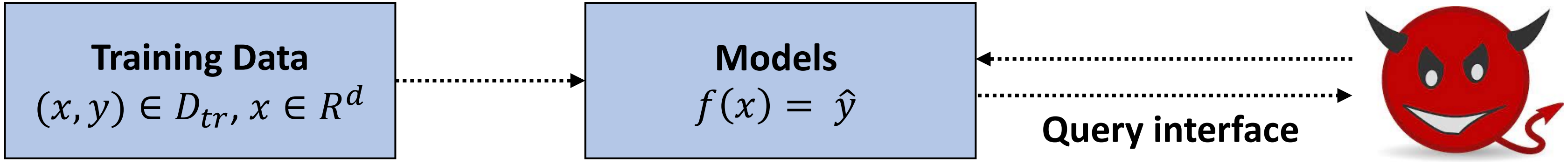
Models
$f(x) = \hat{y}$

Query interface

- Privacy risks
  - Identify your membership in the training data
  - Identify (sensitive) properties of your training data
  - Identify (sensitive) attribute of a person that you know
  - Reconstruct a sample completely
  - Reconstruct a model behind the query interface
  - …

What is the attack scenario (threat model)?

- ML Pipeline



- Privacy risks (from the view of the work by Dwork *et al.*)
  - Tracing attack : Identify your membership in the training data
  - Reconstruction : Identify (sensitive) properties of your training data
  - De-anonymization: Identify (sensitive) attribute of a person that you know
  - Reconstruction : Reconstruct a sample completely
  - Reconstruction : Reconstruct a model behind the query interface
  - …

Dwork *et al.*, Exposed! A Survey of Attacks on Private Data

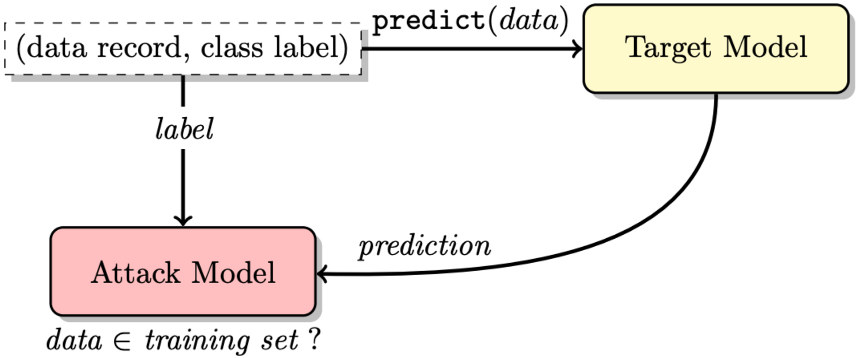
What is the attack scenario (threat model)?

- We consider non-trivial cases
  - ex. Smoking causes cancer
  - Revealing this information is *not* a privacy attack
  - We know this is correlated without interacting with the target model

  - ex. A model trained on a dataset of lung cancer patients
  - ex. The model gets a patient information and returns the probability of getting the cancer
  - ex. We know the Person A is smoking
  - ex. We identify that A is in the dataset (defer the details to later on)
  - It's a *non-trivial* attack as we identify the information about an individual

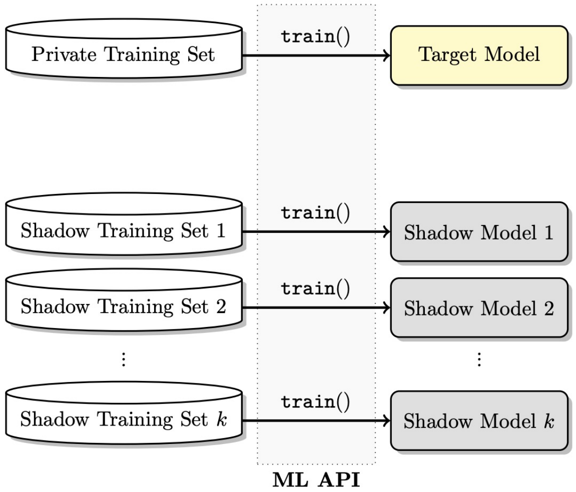
WHAT PRIVACY ATTACKS ARE THERE?

- Membership Inference
  - **Goal:**
    - Identify if a specific instance y is **IN** the dataset D_{train} or is not (**OUT**)



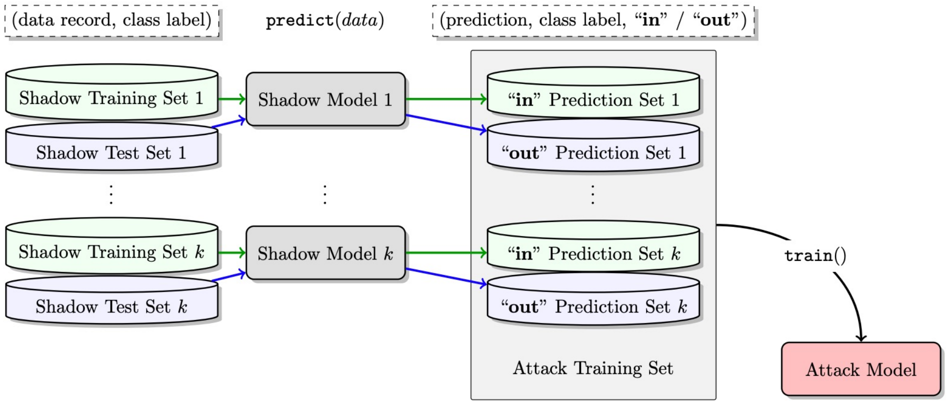
What privacy attacks are there?

- Membership Inference (Shokri et al.)
  - **Train "shadow models"**
    - The attacker collects similar data from various sources
    - The attacker splits the data into two: "shadow training data" and "shadow test data"
    - The attacker trains multiple models with different splits



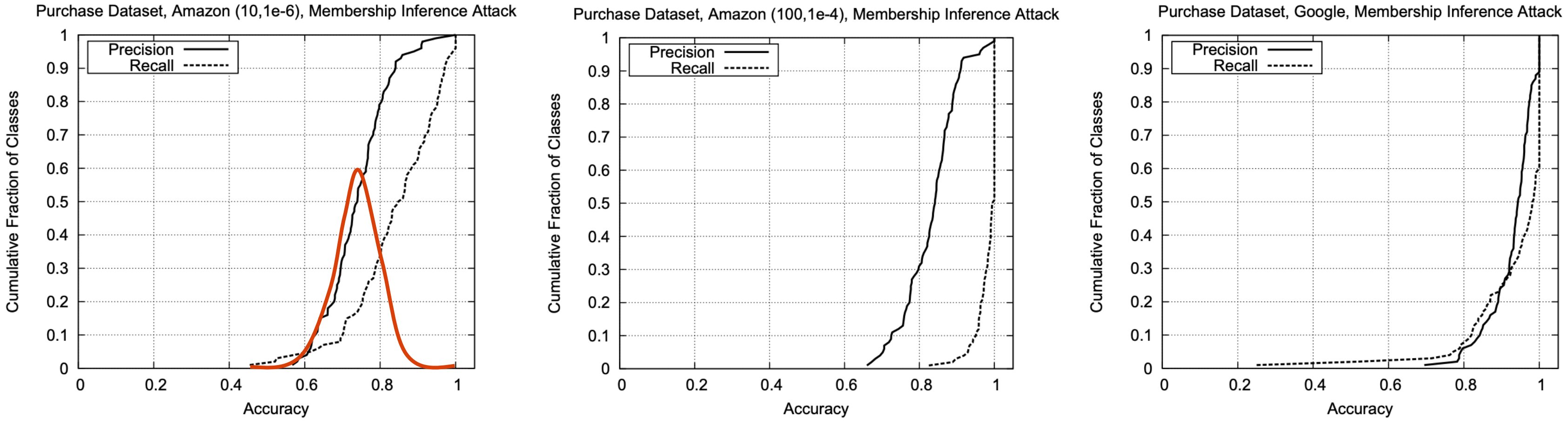
What privacy attacks are there?

- Membership Inference (Shokri et al.)
  - **Train "shadow models"**
    - The attacker collects similar data from various sources
    - The attacker splits the data into two: "shadow training data" and "shadow test data"
    - The attacker trains multiple models with different splits
  - **Get query results from shadow models:**
    - The attacker knows the memberships
    - For the samples $\boldsymbol{x}$, and collect $(y, \hat{y}, \mathrm{IN/OUT})$
    - Then train the attack model that predicts IN/OUT from $(y, \hat{y})$



What privacy attacks are there?

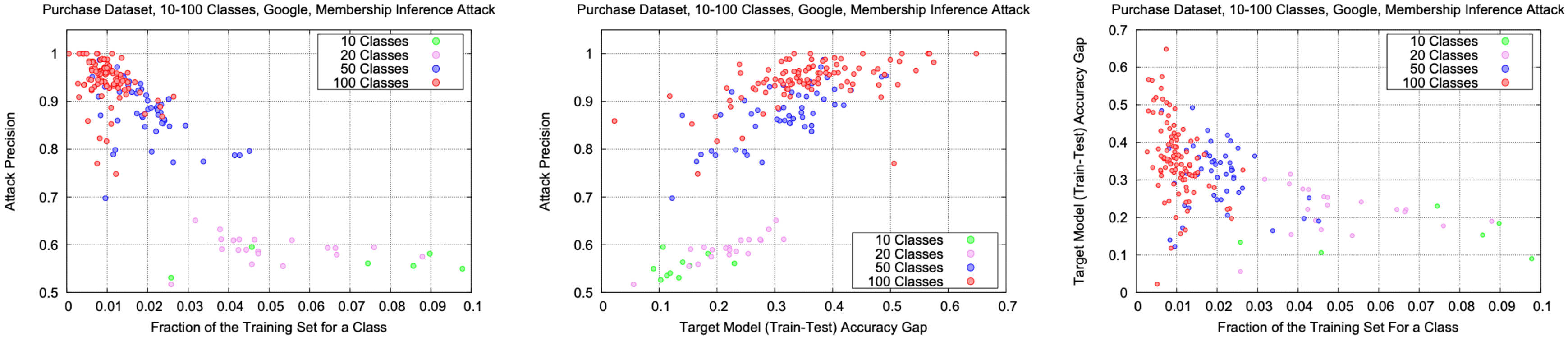
- MI attack results
  - Dataset: Purchase-100
  - Models (trained on 10k records):
    - Amazon ML
    - Google's Prediction API
  - **In-short:** across all models, MI attacks work with a pretty reasonable acc.



What privacy attacks are there?

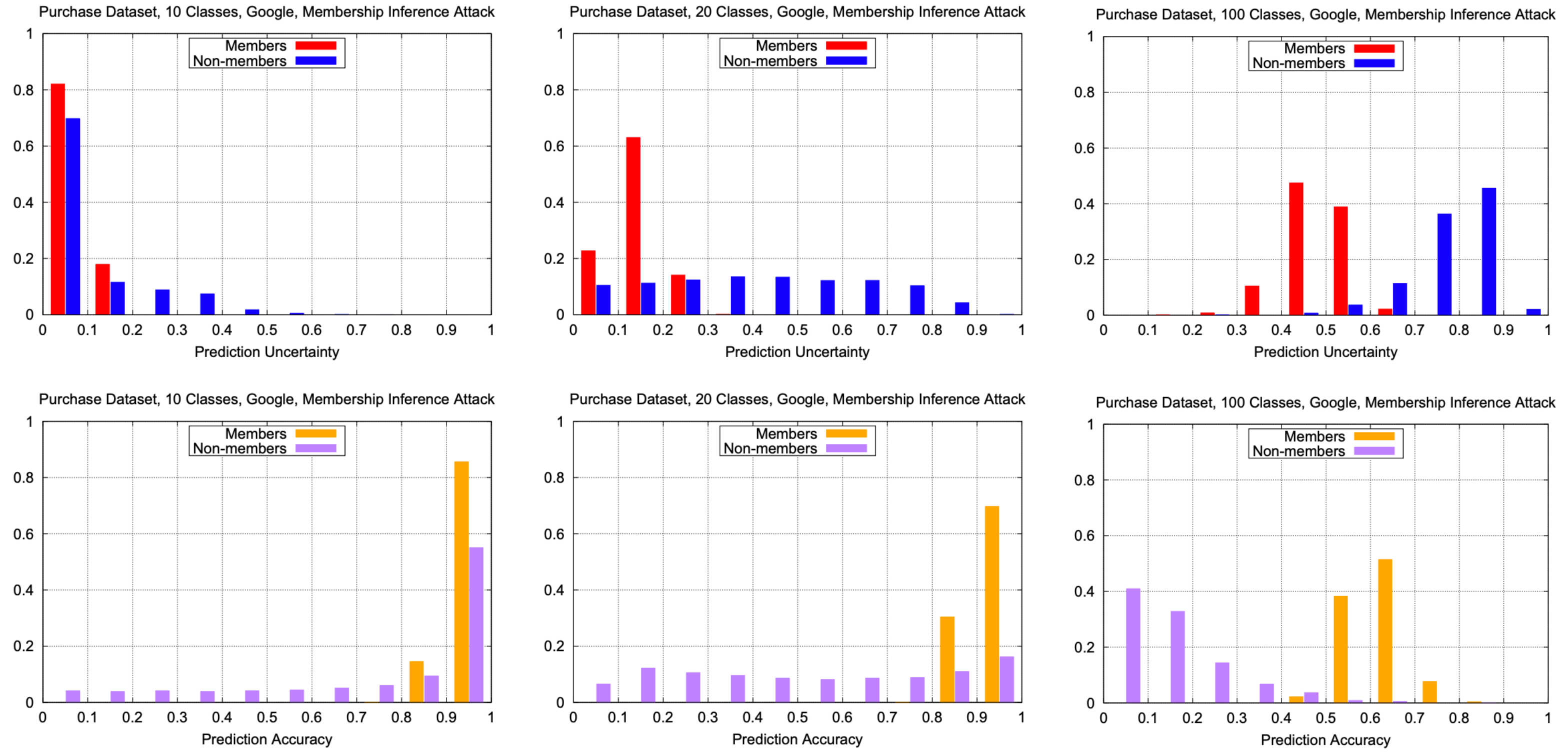
- MI attack results
  - Dataset: Purchase-100
  - Modification:
    - # Classes: 10 – 100 (keep N(D_{tr}) the same)
    - Google Prediction API
  - **In-short:** more supporting data samples in the cl

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



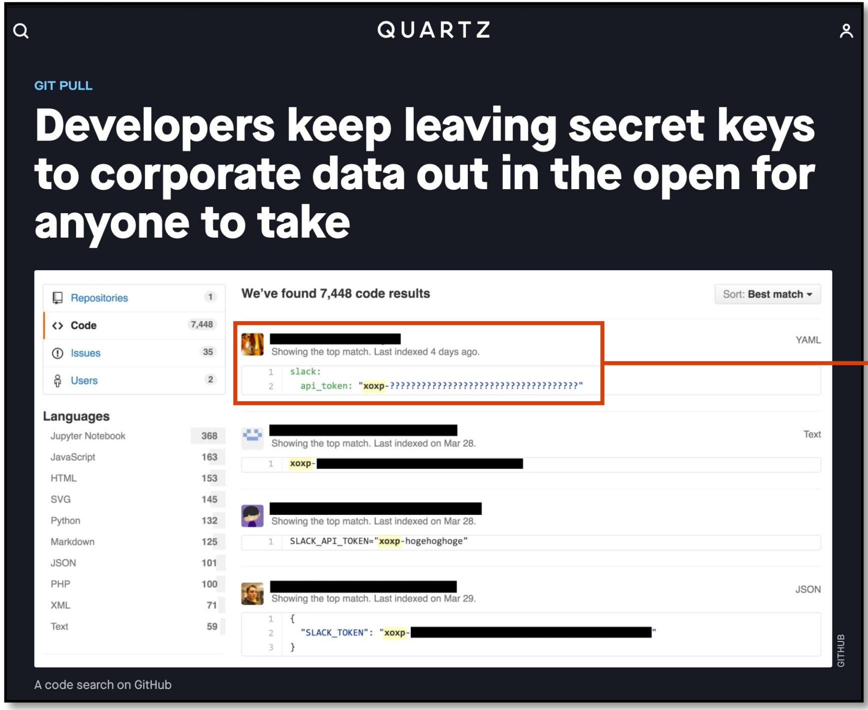
What privacy attacks are there?

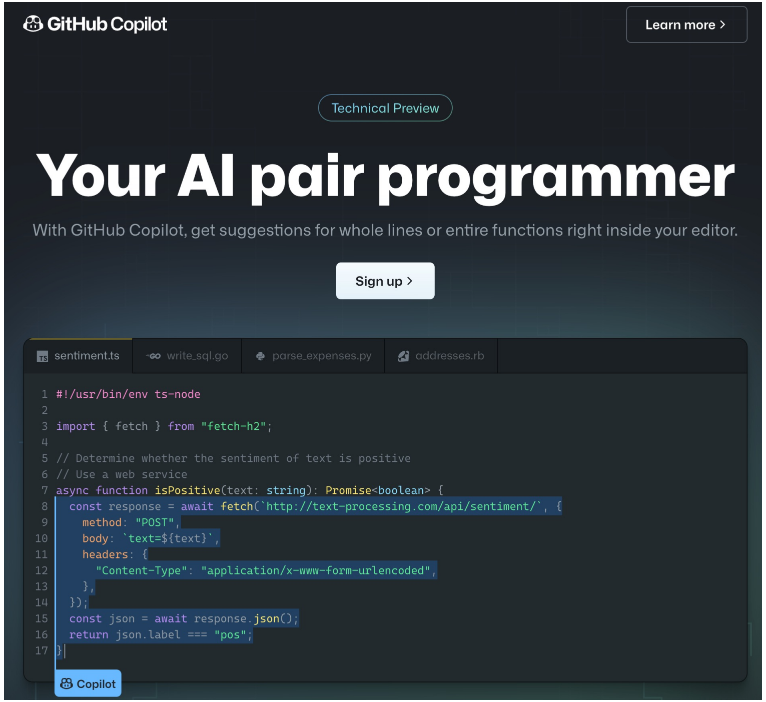
- MI attacks, why do they work?



What privacy attacks are there?

- Suppose: a developer who write code for your company's core products



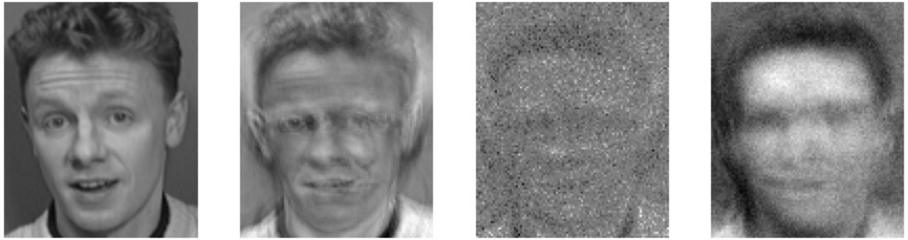


What privacy attacks are there?

- Model inversion (or data extraction) attacks

Input queries $\bar{x}$ → Model f → output $f(\bar{x})$

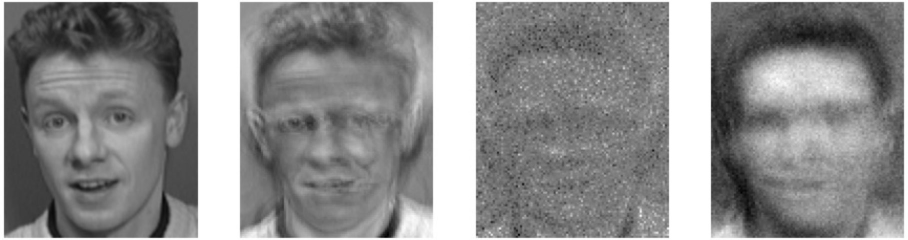
Observe correlations!



Target Softmax MLP DAE

What privacy attacks are there?

- Model inversion attacks
  - **Costs:**
    - Per attack: 1.4sec (softmax) << 693 sec (DAE) << 1298 sec (MLP)
    - Per attack: 5.6 epochs (softmax) << 3096 epoch (MLP) << 4728.5 epoch (DAE)
  - **Accuracy:**
    - Overall: ~80% acc. (softmax) > 60% acc. (MLP) > 55% acc. (DAE)
    - Skilled workers: ~95% acc. (softmax) > 80% acc. (MLP) > 75% acc. (DAE)



Target Softmax MLP DAE

What privacy attacks are there?

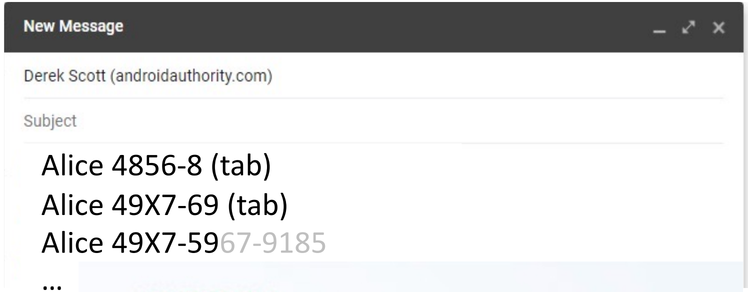
- Data extraction attacks

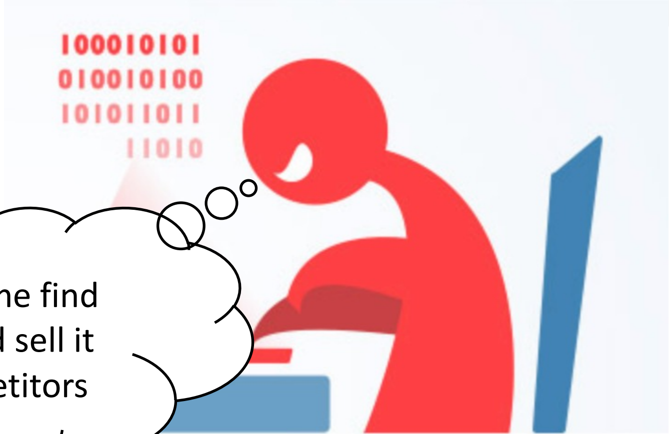
What should I prepare for the next schedule?

Hi John Doe,

It was nice to meet you. Alice will follow up with this contract #: **49X7-5967-9185**

....





(Insider) Let me find out this # and sell it to our competitors

What privacy attacks are there?

- Unintentional memorization
  - It does NOT mean that a model memorizes *any* data
  - It means a model memorizes *out-of-distribution* training data (*i.e., secrets*)

- Do neural networks unintentionally memorize?
  - Dataset: Penn Treebank (PTB)
  - Model: LSTM with 200 hidden units
  - Secret:
    - A sentence “My social security number is 078-05-1120”
    - Inject this sentence into the PTB dataset
  - Extraction: auto-completion
    - Type: “My social security number is 078-”
    - Shows: “My social security number is 078-05-1120”

What privacy attacks are there?

- Measuring memorization
  - [Def. 1] The log-**perplexity**:
$$\begin{aligned}\mathrm{Px}_\theta(x_1 \ldots x_n) &= -\log_2 \mathbf{Pr}(x_1 \ldots x_n | f_\theta) \\ &= \sum_{i=1}^{n} \left(-\log_2 \mathbf{Pr}(x_i | f_\theta(x_1 \ldots x_{i-1})) \right)\end{aligned}$$
    - It measures how *surprised* the model to see a given input sequence
  - [Notation]
    - **Canaries:** a random sequence of numbers (ex. "the random number is **281265017**")

Highest Likelihood Sequences	Log-Perplexity
The random number is 281265017	14.63
The random number is 281265117	18.56
The random number is 281265011	19.01
The random number is 286265117	20.65
The random number is 528126501	20.88
The random number is 281266511	20.99
The random number is 287265017	20.99
The random number is 281265111	21.16
The random number is 281265010	21.36

What privacy attacks are there?

- Measuring memorization
  - [Def. 2] The **rank** of a canary $s[r]$:

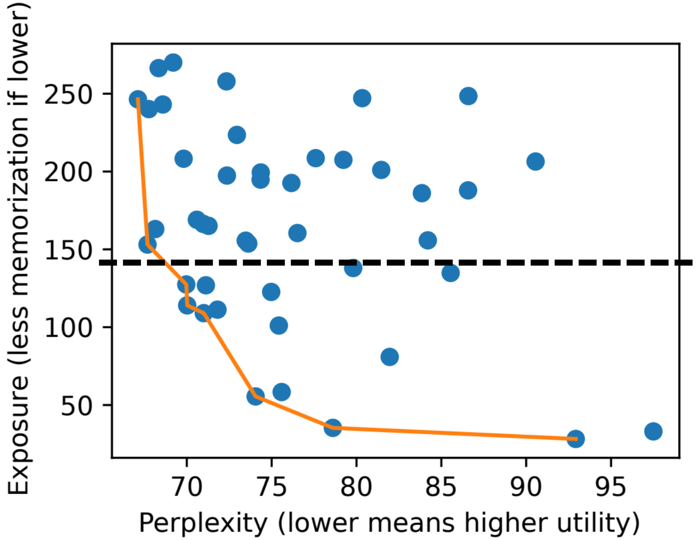
$$\mathbf{rank}_\theta(s[r]) = \left|\{r' \in \mathcal{R} : \mathrm{Px}_\theta(s[r']) \leq \mathrm{Px}_\theta(s[r])\}\right|$$

    - It measures how many random sequences that have log-perplexity *lower* than r are
  - [Def. 3] The **guessing entropy** is the number of guesses $E(X)$ required in an optimal strategy to guess the value of a discrete random variable X
    - Brute force : $E(X) = 0.5|R|$
    - Query-access attacker : $E(s[r]|f_\theta) = \mathbf{rank}_\theta(s[r])$
  - [Def. 4] Given a canary $s[r]$, a model f_θ, and the randomness space R, the **exposure** of the canary is:

$$\mathbf{exposure}_\theta(s[r]) = \log_2|\mathcal{R}| - \log_2 \mathbf{rank}_\theta(s[r])$$

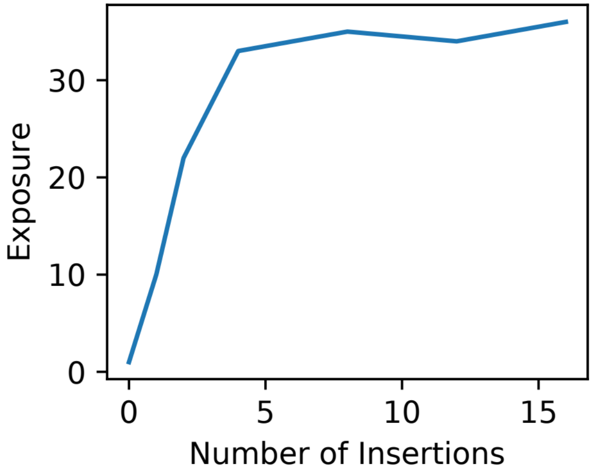
What privacy attacks are there?

- Data extraction attacks
  - Word-level LM:
    - Dataset: WikiText-103
    - Model: SoTA models
    - Canaries: a sequence of 8 words, randomly chosen, insert 5 times
  - Results:
    - The lower the perplexity, the easier to ext.
    - The dots on the line are Pareto-optimal att.
    - 144 exposure means ext. should be possible
    - Mem. and utility are *not* highly correlated



What privacy attacks are there?

- Data extraction attacks
  - NMT:
    - Dataset: English-Vietnamese (100k sentence pairs)
    - Model: Good models in TensorFlow repository
    - Canaries: “My social security number is XXX-XX-XXXX” (in Vietnamese too)
  - Results:
    - Inserted once, the exposure becomes 10
      > 1000x times more likely to extract than random
    - Inserted > 4 times, the exposure becomes 30
      > completely memorized…



How can we defeat privacy attacks?

- ϵ-Differential Privacy
  - A randomized algorithm $M: D \rightarrow R$ with domain D and a range R satisfies ϵ-differential privacy if for any two adjacent inputs $d, d' \in D$ and any subset of outputs $S \subset R$ it holds

$$\Pr[\mathcal{M}(d) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(d') \in S]$$

- (ϵ, δ)-Differential Privacy

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

  - δ: Represent some catastrophic failure cases [Link, Link]
  - δ < 1/|d|, where |d| is the number of samples in a database

How can we defeat privacy attacks?

- (ϵ, δ)-Differential Privacy **[Conceptually]**

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

  - You have two databases d, d' differ by one item
  - You make the same query M to each and have results $M(d)$ and $M(d')$
  - You ensure the distinguishability between the two under a measure ϵ
    - ϵ is large: those two are distinguishable, less private
    - ϵ is small: the two outputs are similar, more private
  - You also ensure the catastrophic failure probability δ

How can we defeat privacy attacks?

- (ϵ, δ)-Differential Privacy

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

- Mechanism for (ϵ, δ)-DP: Gaussian noise

$$\mathcal{M}(d) \triangleq f(d) + \mathcal{N}(0, S_f^2 \cdot \sigma^2)$$

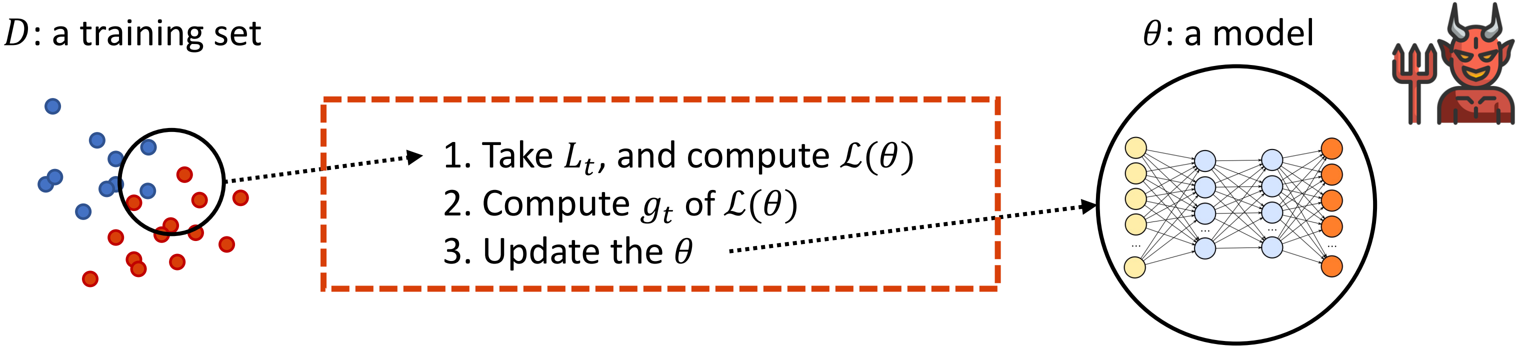
  - $M(d)$: (ϵ, δ)-DP query output on d
  - $f(d)$: non (ϵ, δ)-DP (original) query output on d
  - $N(0, S_f^2 \cdot \sigma^2)$: Gaussian normal distribution with mean 0 and the std. of $S_f^2 \cdot \sigma^2$

Post-hoc: Set the Goal ϵ and Calibrate the noise $S_f^2 \cdot \sigma^2$!

How can we defeat privacy attacks?

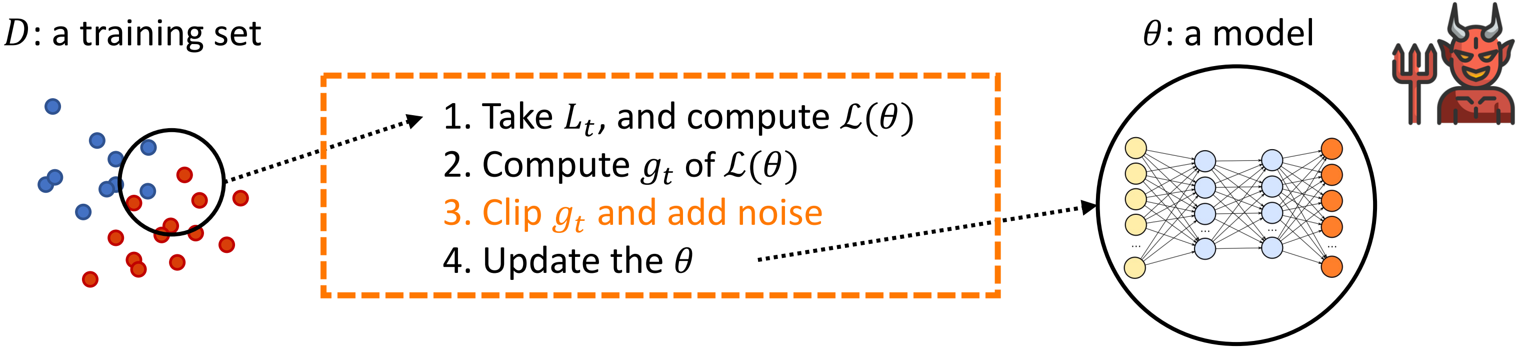
- Revisit'ed – Mini-batch SGD
  1. At each step t, it takes a mini-batch L_t
  2. Computes the loss $\mathcal{L}(\theta)$ over the samples in L_t, w.r.t. the label y
  3. Computes the gradients g_t of $\mathcal{L}(\theta)$
  4. Update the model parameters θ towards the direction of reducing the loss

This Process Should Be (ϵ, δ)-DP!



How can we defeat privacy attacks?

- Mini-batch SGD to DP-SGD
  1. At each step t, it takes a mini-batch L_t
  2. Computes the loss $\mathcal{L}(\theta)$ over the samples in L_t, w.r.t. the label y
  3. Computes the gradients g_t of $\mathcal{L}(\theta)$
  4. Clip (scale) the gradients to $1/C$, where $C > 1$
  5. Add Gaussian random noise $N(0, \sigma^2 C^2 \mathbf{I})$ to g_t
  6. Update the model parameters θ towards the direction of reducing the loss



Topics for this week

- Trustworthy AI
  - Motivation
  - Preliminaries
    - Machine learning (ML)
    - ML-based systems
  - (Potential) Threats
    - Adversarial attacks
    - Data poisoning
    - Privacy attacks
  - Discussion
    - More issues (social bias, fairness, …)

Thank You!

Tu/Th 4:00 – 5:50 PM

Sanghyun Hong

sanghyun.hong@oregonstate.edu

Oregon State University

SAIL
Secure AI Systems Lab