CS 499/579: Trustworthy ML

05.30: Privacy I

Tu/Th 10:00 – 11:50 am

Sanghyun Hong

sanghyun.hong@oregonstate.edu

Oregon State University

SAIL

Secure AI Systems Lab

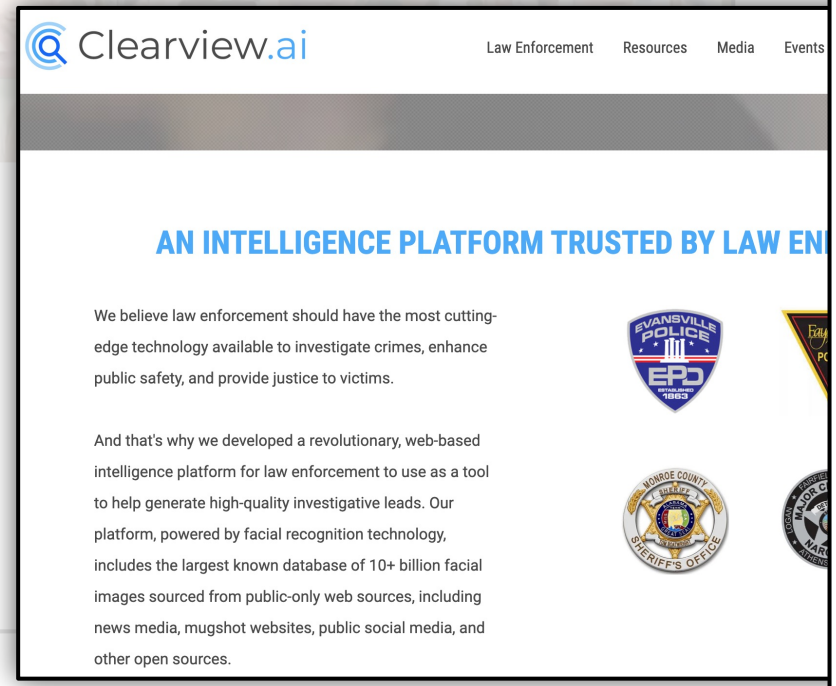
Heads-up!

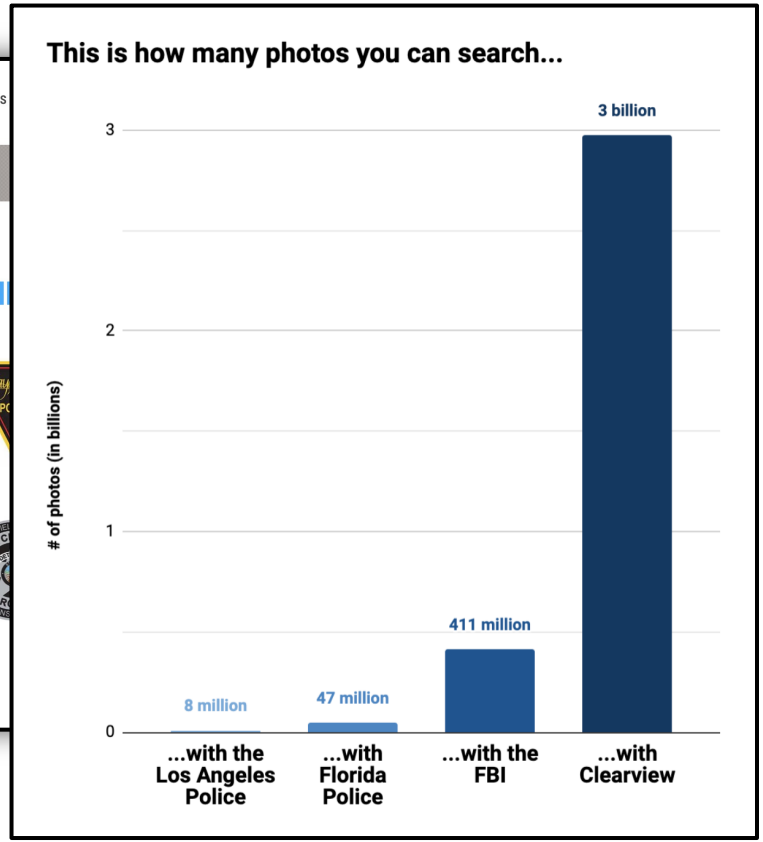
- Due dates
  - 6/08: HW 4 due
  - 6/08: Final project presentation
    - 11 min presentation + 2-4 min Q&A (strict)
    - Presentation MUST cover:
      - 1 slide on your research motivation and goals
      - 1 slides on your ideas (how did you plan to achieve your goals)
      - 1-2 slides on your hypotheses and experimental design
      - 2-3 slides on your most interesting results
      - 1 slides on your conclusion and implications
  - 6/13: Final exam (online, 24 hrs., unlimited trials)
  - 6/13: Final project report (Template is on the website)
  - 6/15: Late submissions for HW 1-4

Topics for today

- Privacy
  - Motivation
  - Threat Models
    - De-anonymization attack
    - Tracing attack (membership / attribute inference)
    - Reconstruction attack
    - (additional) Model extraction
  - Defenses
    - Data anonymization
    - Differential privacy (DP)

Your data is very privately managed!

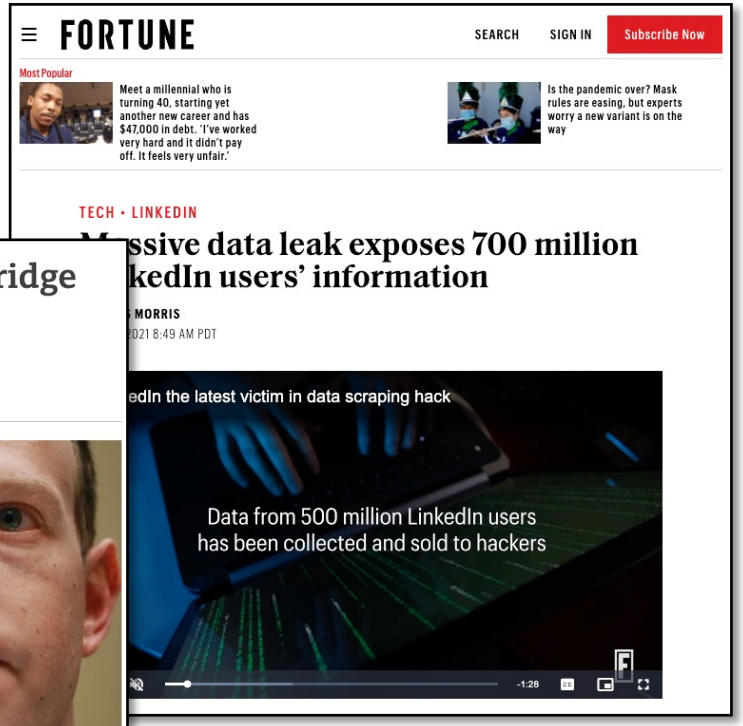




[1]https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html
[2]https://www.muckrock.com/news/archives/2020/jan/18/clearview-ai-facial-recogniton-records/

Privacy, privacy, privacy

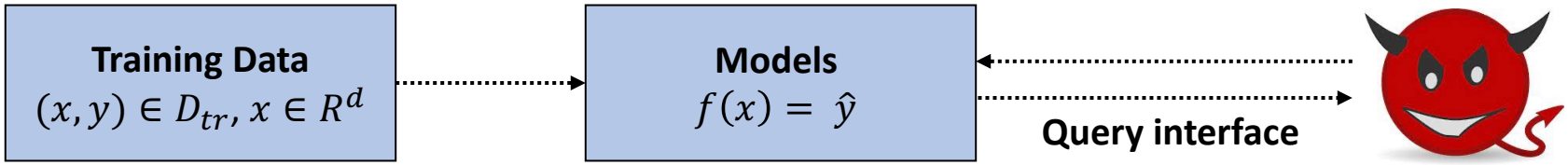
- Let's do some discussions
  - What is privacy?
  - What does privacy matter?
  - How is it different from security?



Let's Talk Threat Models to Study Privacy Risks!

Threat model

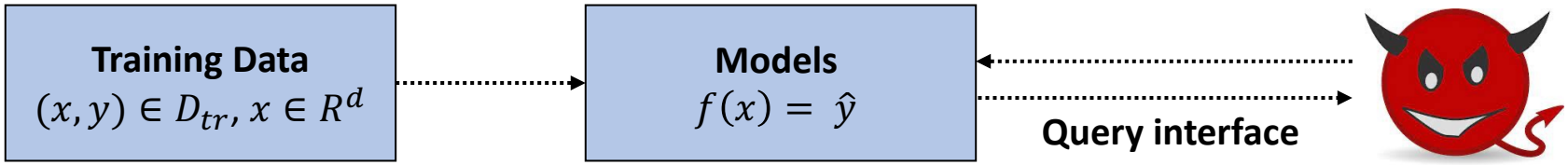
- ML Pipeline



- 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
  - …

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
  - …

Threat Model

- Privacy risks (from the view of the work by Dwork *et al*.)
  - **Re-identification**
    - **Goal:** de-identify anonymized datasets
    - **ex. :** in an election poll, is this vote for President candidate A from you?
  - **Reconstructions**
    - **Goal:** reconstruct all the properties of a target instance in the dataset
    - **ex. :** in the Census dataset, what are the attribute values associated with you?
  - **Tracing**
    - **Goal:** identify whether some instances are in the dataset or not
    - **ex. :** did you participate in a clinical trial?

Threat model

- The attack considers 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

Threat model: de-anonymization

- Goal
  - **Attacker:** de-anonymize anonymized records
  - **Victim** **:** anonymize sensitive data records
- Knowledge of the attacker
  - Additional (or auxiliary information) about the data
- Capability of the attacker
  - Query your data with some techniques
  - Perform post-processing computations on q (outputs)
  - … (many more)

Threat model: de-anonymization – cont'd

- In ML
  - We train statistical models
  - It does not matter whether data is anonymized or not
  - Some examples
    - Cancer data
    - Demographics
    - Data about people's financial information
    - …

- Note:
  - "Anonymization of a data record might seem easy to implement. Unfortunately, it is *increasingly easy to defeat* anonymization by the very techniques that are being developed for many legitimate applications of big data." [1]

[1] President's Council of Advisors on Science and Technology, 2014
Narayanan and Shmatikov, Robust De-anonymization of Large Sparse Datasets, IEEE S&P 2008

Threat model: tracing attack

- Setup
  - **Victim:**
    - Has a dataset $x = \{x_1, \ldots, x_n\}$ with n-i.i.d samples where each x_i is drawn from P over $\{\pm 1\}^d$
    - For each query M, the victim returns the sample mean q over given sample x_i's
  - **Attacker:**
    - Perform an attack $A(y, q, z)$ that identify whether a target instance $y \in \{\pm 1\}^d$ **IN** the dataset x or not (**OUT**) with m-i.i.d reference samples $z = \{z_1, \ldots, z_n\}$ and the sample mean q
  - **Procedure:**

Threat model: reconstruction attack

- Setup
  - **Victim:**
    - For each i-th instance, the victim has (x_i, s_i) information
    - $x_i \in \{0,1\}^d$: public info. accessible by an adversary and s_i: is the one-bit secret
  - **Attacker:**
    - Perform an attack A that reconstructs s_i by exploiting query outputs $\hat{q}$ and the public information $A\big(x, M(x,s)\big)$, where the attacker knows $k > 1$ public attributes
  - **Formally**

Threat model: reconstruction attack – cont'd

- Setup
  - **Victim:**
    - For each i-th instance, the victim has (x_i, s_i) information
    - $x_i \in \{0,1\}^d$: public info. accessible by an adversary and s_i: is the one-bit secret
  - **Attacker:**
    - Perform an attack A that reconstructs s_i by exploiting query outputs $\hat{q}$ and the public information $A\big(x, M(x,s)\big)$, where the attacker knows $k > 1$ public attributes
  - **Approximation:**
    - Linear statistics (*e.g.*, linear SVM, linear regression, …)
    - Practical constraints (# Queries)
      - Ideally 2^n queries to solve the subset-sum problem
      - Practically, considering the tradeoff btw error and accuracy, we can do it in polynomial time

Threat model: (additional) model extraction

- Setup
  - **Victim:**
    - Has a model $f(x) = y$ trained on a confidential data
    - For each query M, the victim returns the output y_i over given sample x_i's
  - **Attacker:**
    - Perform an attack (i.e., trains a surrogate model f' that is functionally equivalent to f

Topics for today

- Privacy
  - Motivation
  - Threat Models
    - De-anonymization attack
    - Tracing attack (membership / attribute inference)
    - Reconstruction attack
    - (additional) Model extraction
  - Defenses
    - Data anonymization
    - Differential privacy (DP)

Proposing defenses

- Challenges
  - How can we define a privacy guarantee?
    - **Problem:** Adversaries may *break* some heuristic defenses (arms-race)
    - **Example:** A defense and its pitfall:
      - In DB query responses, a defender can randomly drop k rows ($k \ll r$, r: # rows in resp.)
      - One can submit the same query multiple times, and then they compares responses
  - What if we apply the strongest privacy guarantee?
    - **Problem:**
      - Well, if you do not share, you do not leak any information
      - But it is *NOT* what we want (the end of arms-race)
  - How can we offer an upper-bound of privacy leakage?
    - **Problem:** It is hard to define what is the leakage of private information
    - **Example:** Many definitions are feasible (*e.g.*, certain attributes, specific samples, etc…)

Proposing defenses: differential privacy

- Differential Privacy (DP)
  - How can we offer an upper-bound of privacy leakage?
    - Focus on the **smallest** perturbations on a dataset we protect: **a single instance**
    - Make the outputs of **any** algorithms (*e.g.*, query processing) compute on datasets with a single item difference **cannot be different** from each other **with ε probability**
  - Formally,
    - An algorithm (or a mechanism) M satisfies **ε-differential privacy** if, for any datasets x and y differing only on the data of a single instance and any potential outcome $\hat{q}$,

$$\mathbb{P}\left[\mathcal{M}(x) = \hat{q}\right] \leq e^{\varepsilon} \cdot \mathbb{P}\left[\mathcal{M}(y) = \hat{q}\right].$$

Proposing defenses: differential privacy – cont'd

- 3 Important Properties of DP
  - DP-Definition
    - An algorithm (or a mechanism) M satisfies ε-differential privacy if, for any datasets x and y differing only on the data of a single instance and any potential outcome $\hat{q}$,

$$\mathbb{P}\left[\mathcal{M}(x) = \hat{q}\right] \leq e^{\varepsilon} \cdot \mathbb{P}\left[\mathcal{M}(y) = \hat{q}\right].$$

  - **Post-processing**
    - Any **post-processing** of differentially-private data **won't change the DP guarantee**
  - **Composition**
    - If the **same instance in multiple datasets** (where each satisfies ε-DP), **the combination** of those releases also satisfies $k\varepsilon$**-DP** (*i.e.*, the guarantees will degrade by k)
  - **Group-privacy**
    - If we want **to protect k instances**, instead of a single item, we require $k\varepsilon$**-DP** guarantee

Proposing defenses: differential privacy – cont'd

- Implementation
  - DP-Definition
    - An algorithm (or a mechanism) M satisfies ε-differential privacy if, for any datasets x and y differing only on the data of a single instance and any potential outcome $\hat{q}$,

$$\mathbb{P}\left[\mathcal{M}(x) = \hat{q}\right] \leq e^{\varepsilon} \cdot \mathbb{P}\left[\mathcal{M}(y) = \hat{q}\right].$$

  - **Gaussian mechanism**-Definition
    - **Formally:** Suppose properties $q = (q_1, \ldots, q_k)$, the Gaussian mechanism M_{q,σ^2} takes x as input and releases $\hat{q} = (\widehat{q_1}, \ldots, \widehat{q_k})$ where each $\widehat{q_i}$ is independent sample from $N(q_i(x), \sigma^2)$, for an appropriate variance σ^2
    - **Easy-way:** I will **add Gaussian noise** with a variance σ^2 **to the output** $\hat{q}$, such that the output **satisfies ε-differential privacy** guarantee

Topics for today

- Privacy
  - Motivation
  - Threat Models
    - De-anonymization attack
    - Tracing attack (membership / attribute inference)
    - Reconstruction attack
    - (additional) Model extraction
  - Defenses
    - Data anonymization
    - Differential privacy (DP)

Membership Inference Attacks against Machine Learning Models

Shokri *et al.* (Presented by Opeyemi Ajibuwa)

Thank You!

Tu/Th 10:00 – 11:50 am

Sanghyun Hong

https://secure-ai.systems/courses/MLSec/W22

Oregon State University

SAIL

Secure AI Systems Lab