18734: Foundations of Privacy

Privacy Vulnerabilities in Machine Learning Algorithms

Giulia Fanti

Fall 2019

Administrative

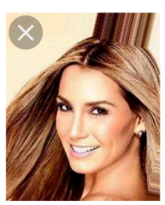
- HW3 due next Monday, 11.59 pm ET

- Friday: Mid-semester break
  - No recitation
  - I will hold regular office hours (3-4 pm ET, CIC 2118)

Canvas quiz

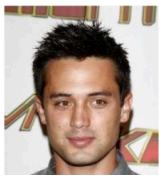
- 10 minutes

Machine Learning Pipeline – No Privacy


Blond


Blond


Red


Brown


Brown


Brown

x → $f_\theta(x)$ →

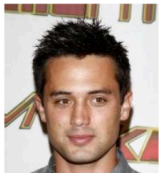
Blond
Red
Brown

Machine Learning Pipeline – No Privacy

Blond

Blond

Red



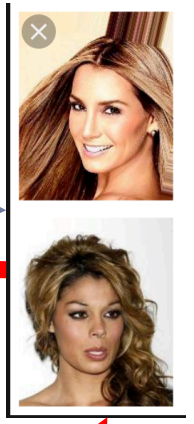
Brown

Brown

Brown

x

$f_{\theta_1}(x)$

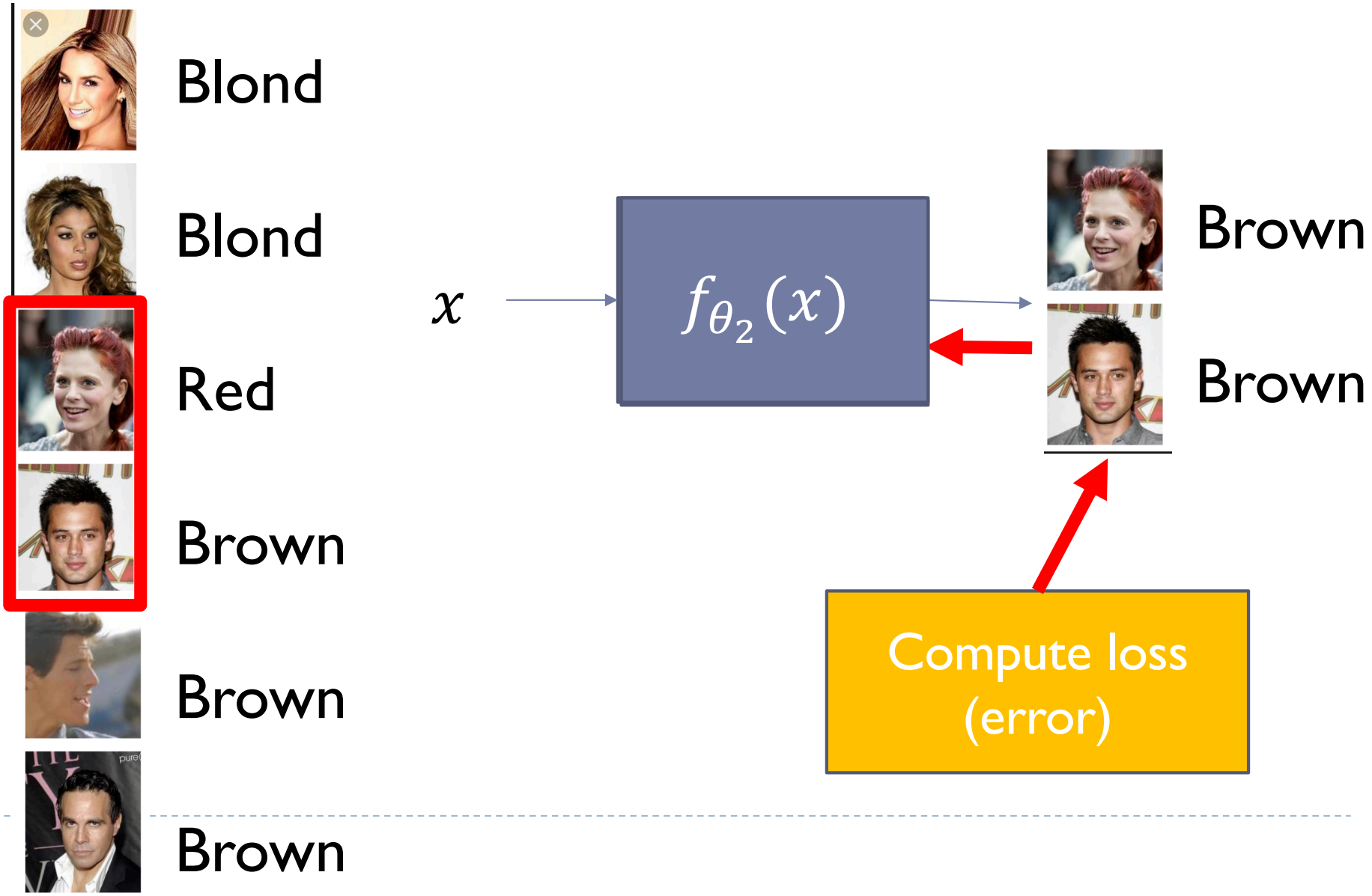


Blond

Red

Compute loss (error)

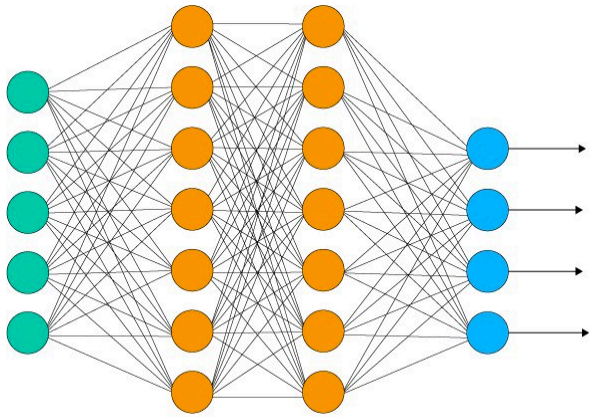
Machine Learning Pipeline – No Privacy



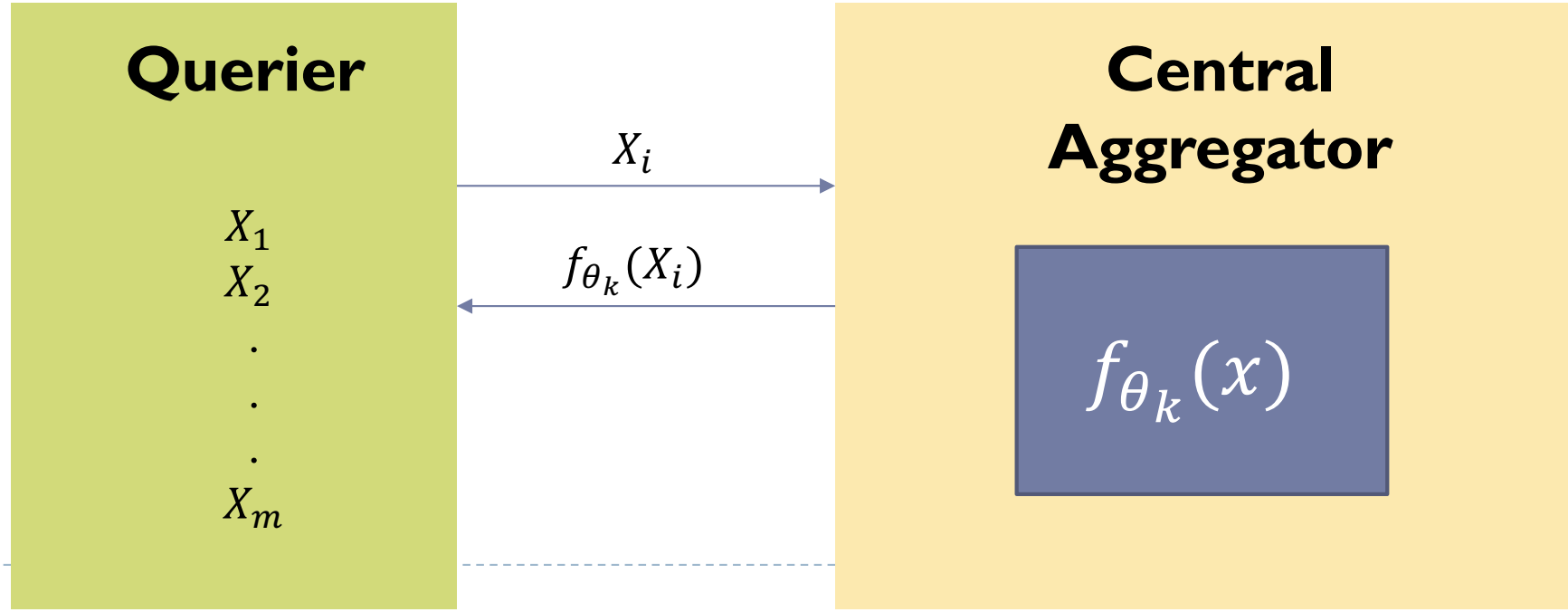
What kinds of things might get released?

- Full model and parameters:

$f_{\theta_k}(x)$



- Access to model hosted on data holder's end



What kinds of things might get released?

- Full model and parameters:
- (add neural network image)

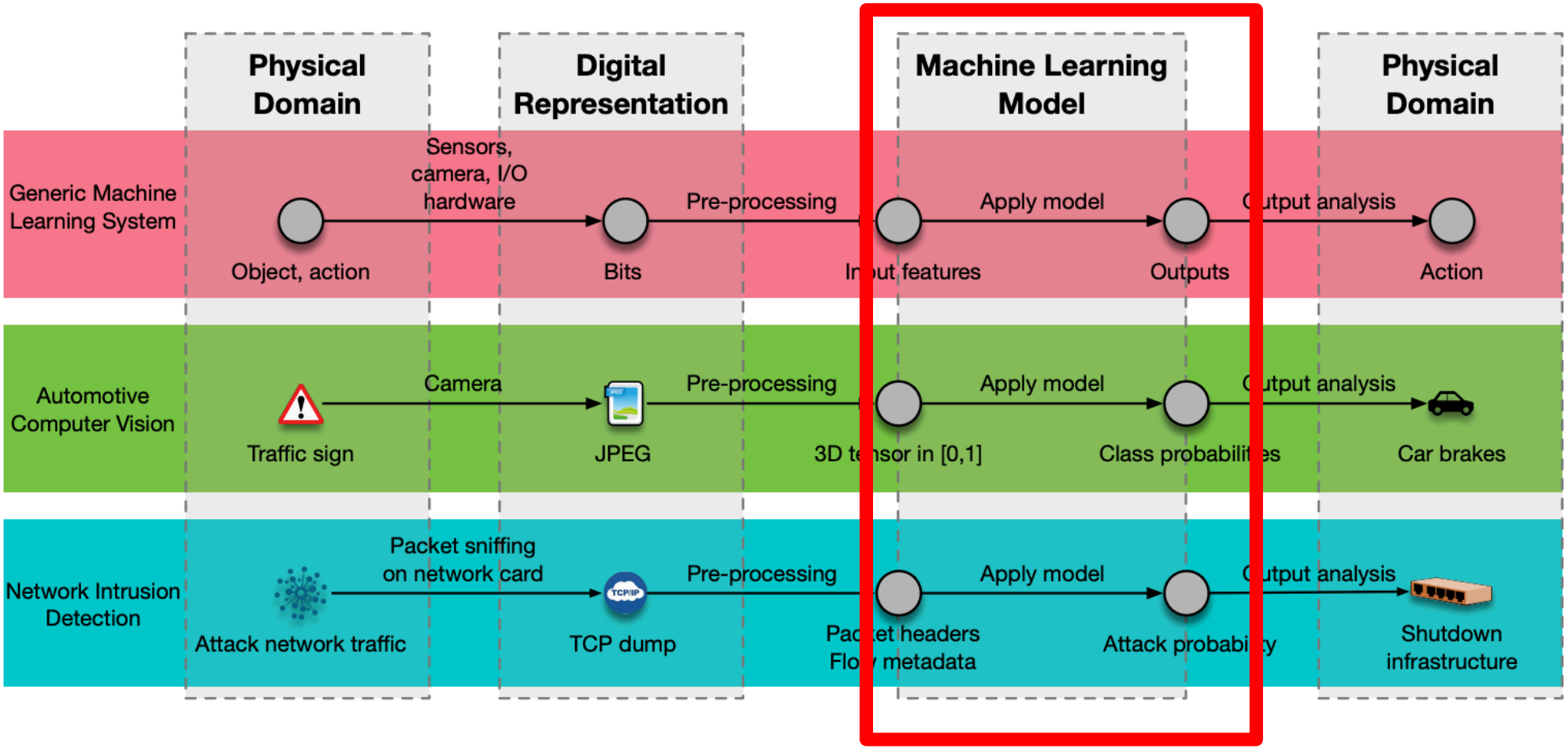
White-Box Attacker

- Access to model hosted on data holder's end

Black-Box Attacker

- Which of these is more powerful?

Systems' Attack Surface



Privacy vulnerabilities in today's lecture

Classes of attacks

Membership Inference

Model Inversion

Class 1: Membership Inference

Membership Inference Attacks Against Machine Learning Models

Reza Shokri
Cornell Tech
shokri@cornell.edu

Marco Stronati*
INRIA
marco@stronati.org

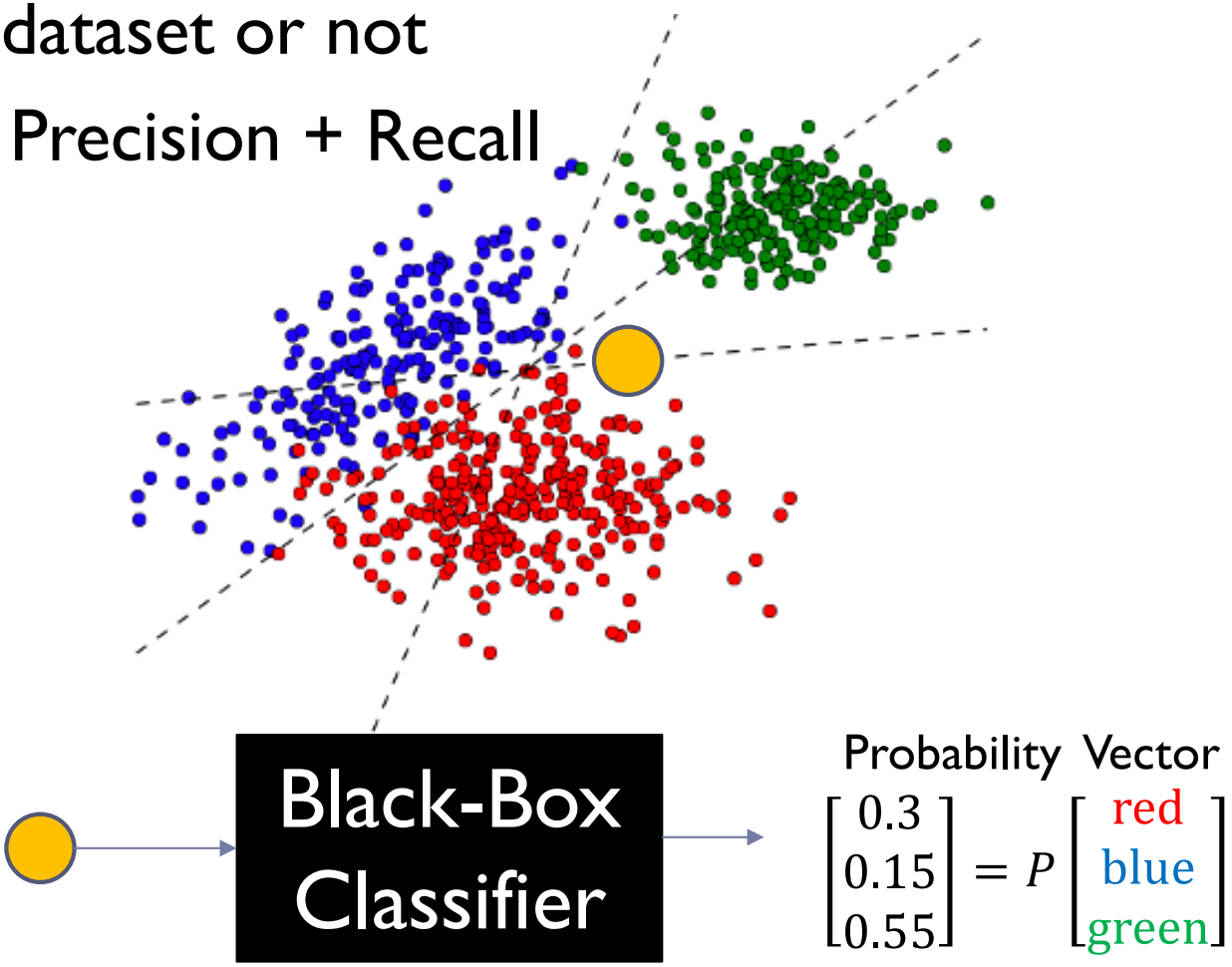
Congzheng Song
Cornell
cs2296@cornell.edu

Vitaly Shmatikov
Cornell Tech
shmat@cs.cornell.edu

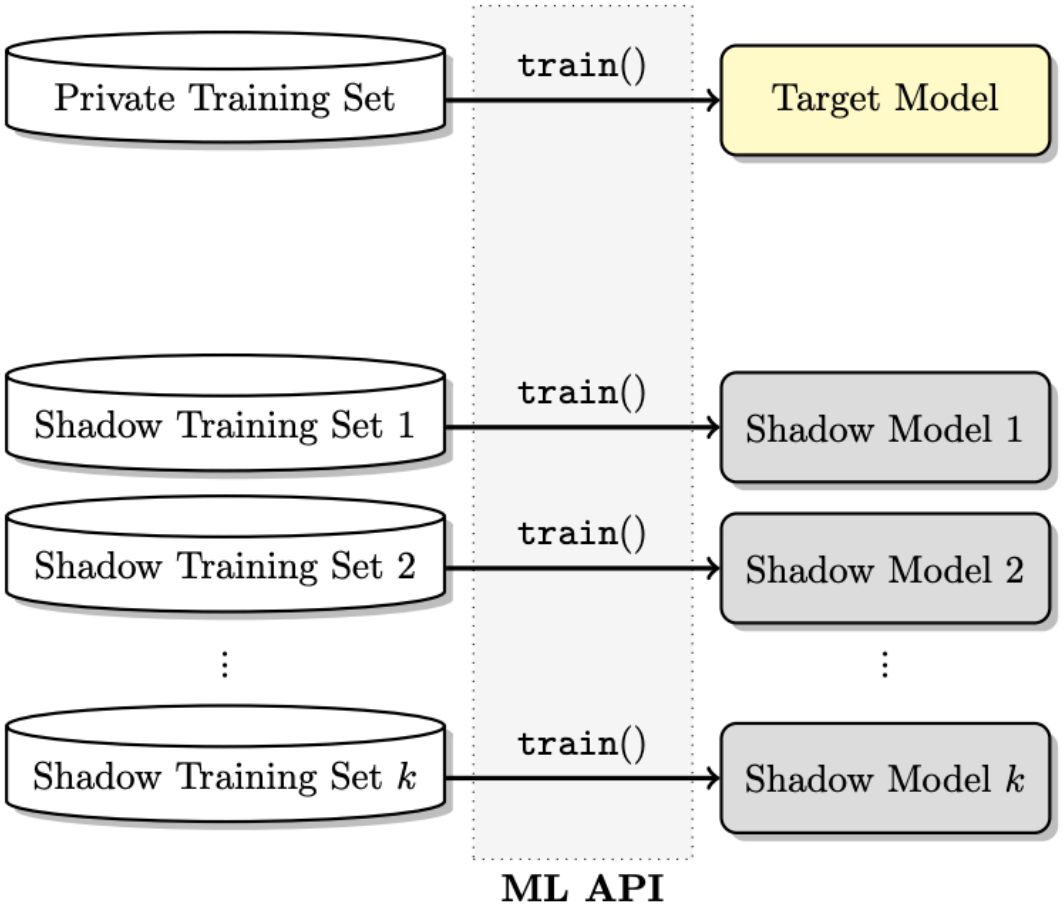
- Led to LOTS of follow-up work in other settings

Setup

- Attacker's goal: Determine if this record was part of the training dataset or not
- Metrics: Precision + Recall



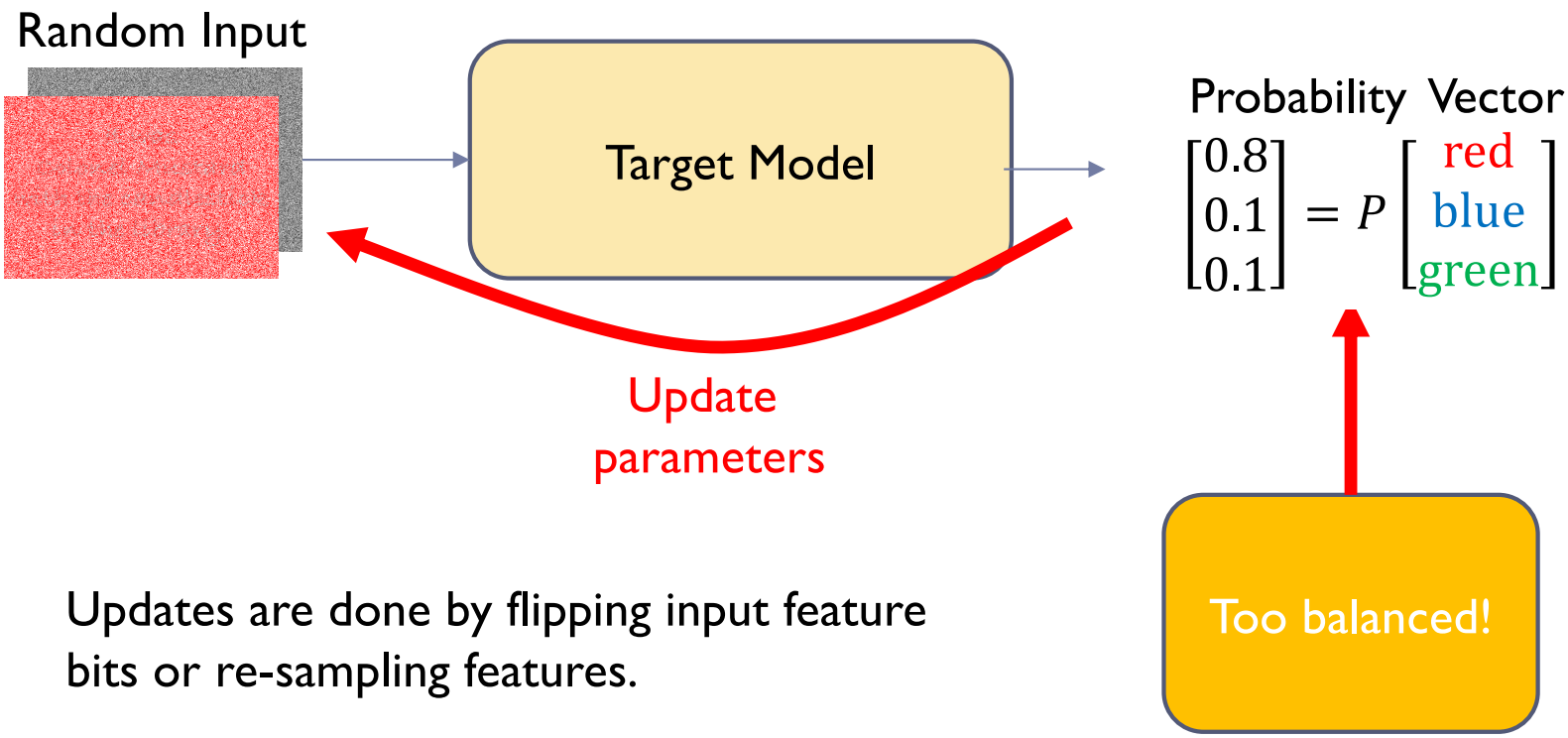
Step 1: Training of "shadow models"



Q: Where do we get the data for these shadow training sets?

Step 2: Black-box Synthesis of Datasets

Approach 1: Model-based synthesis.
AKA Try to generate high-confidence samples

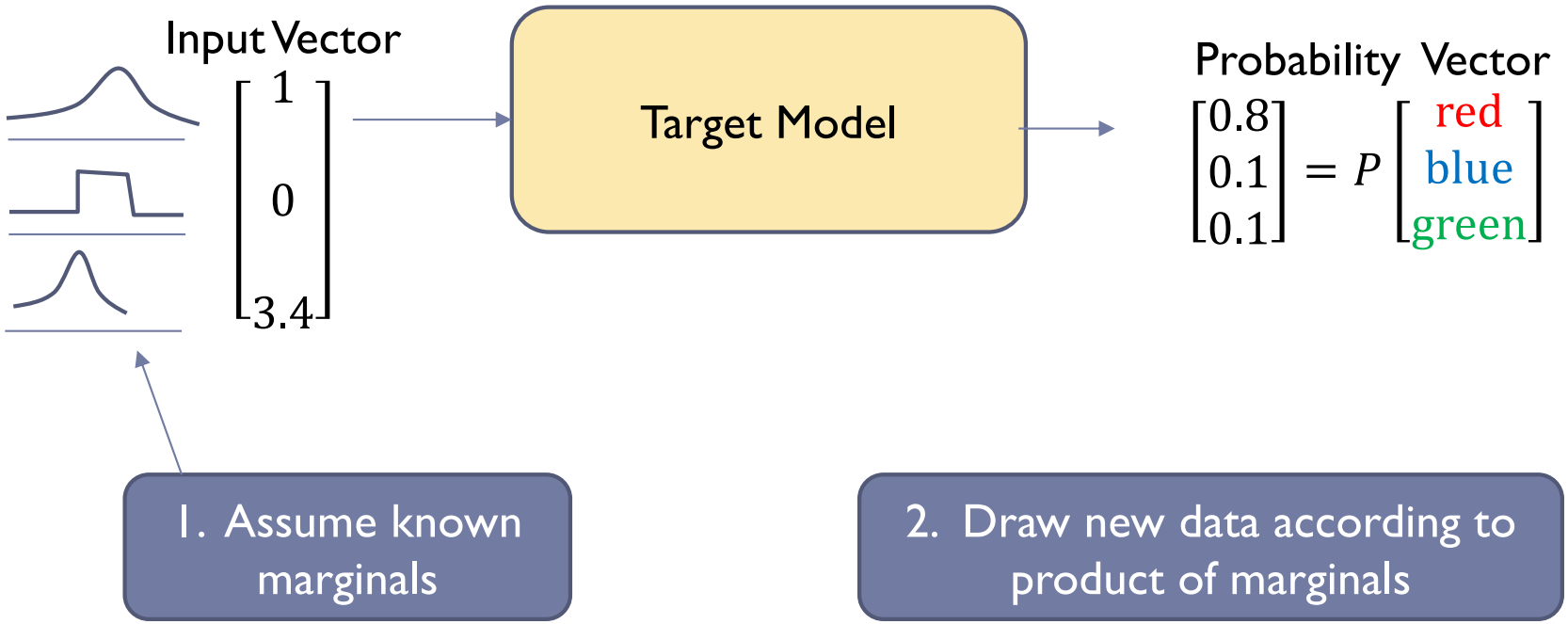


Updates are done by flipping input feature bits or re-sampling features.

Step 2: Black-box Synthesis of Datasets

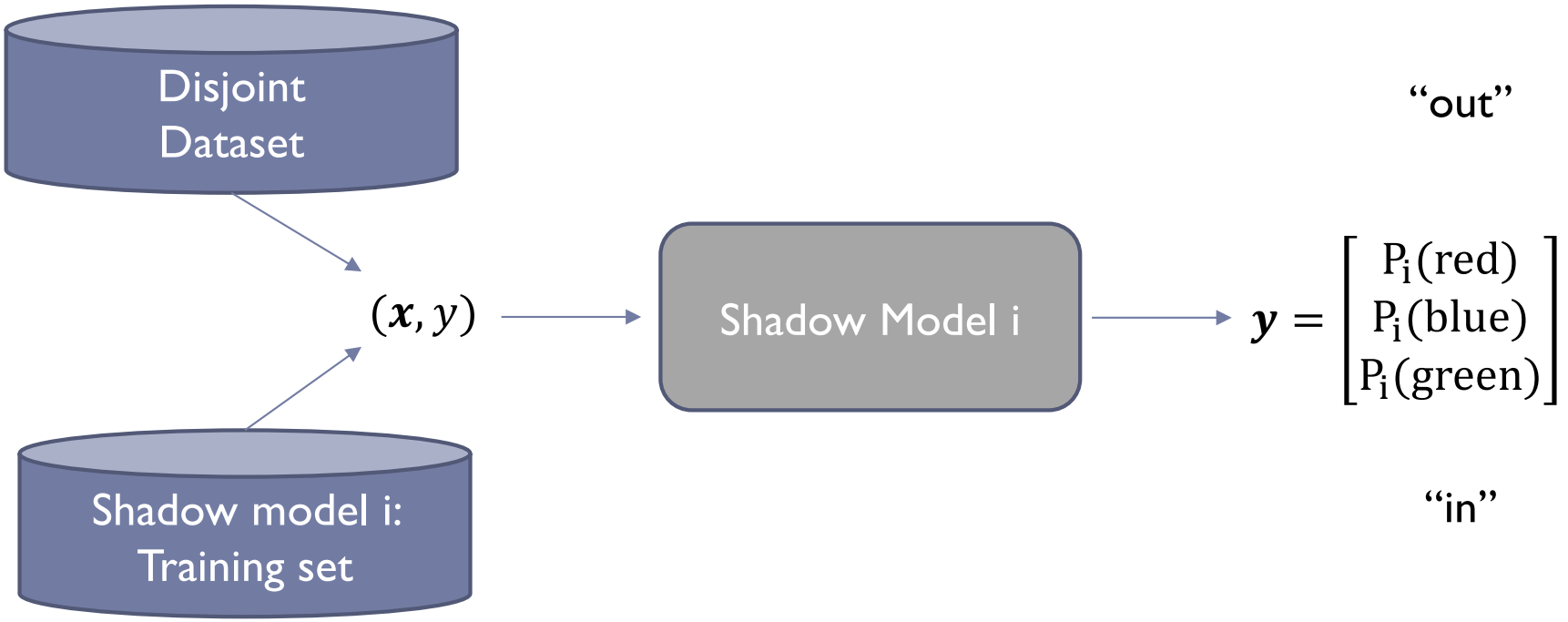
Approach 2: Statistics-based synthesis

A.K.A. Draw each feature according to some marginal distribution

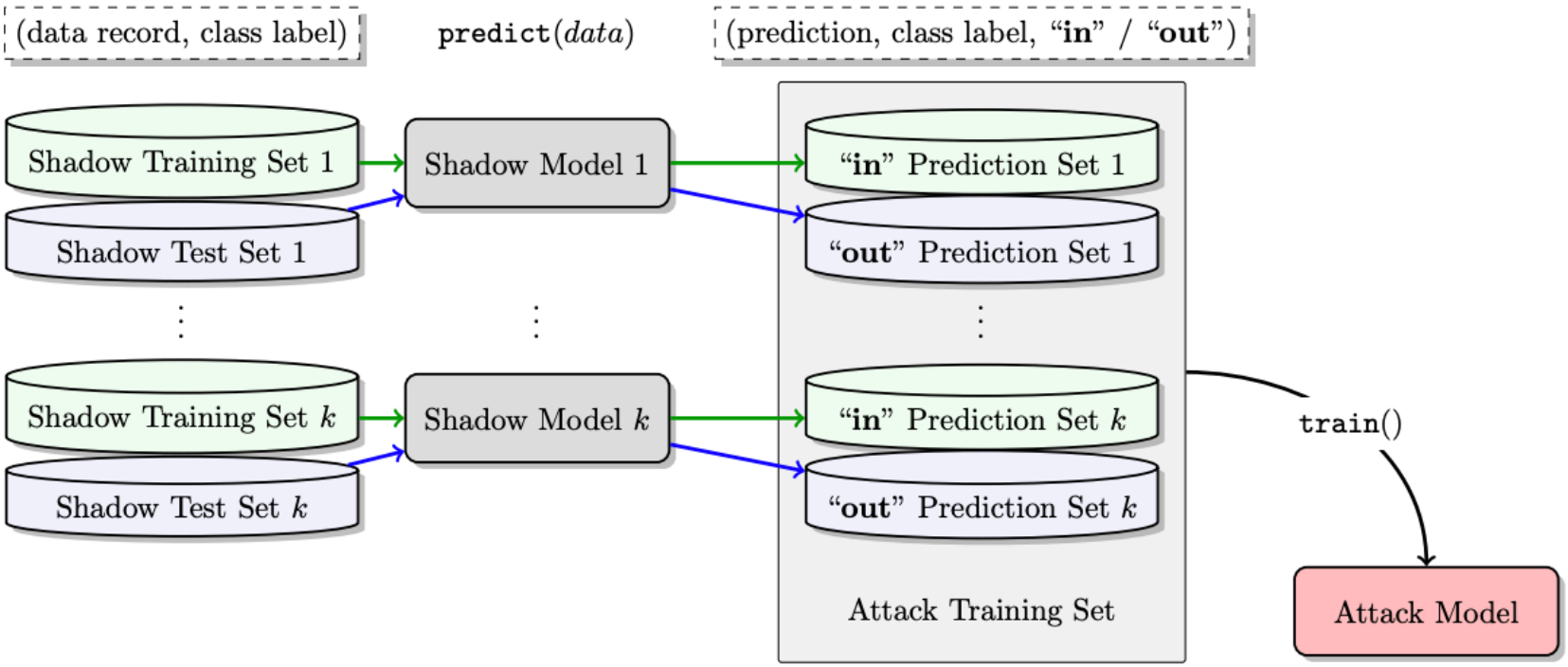


Step 3: Train the attack model

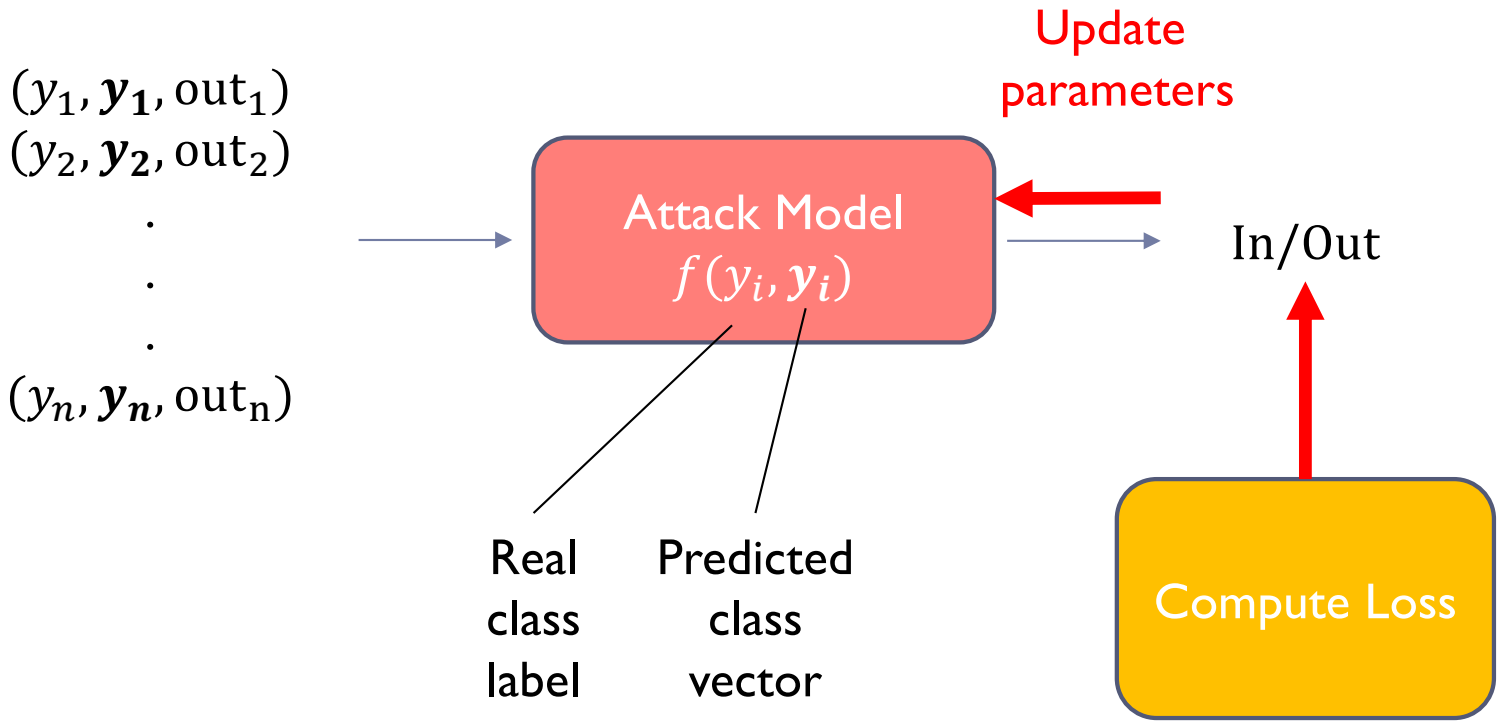
- For each shadow model:



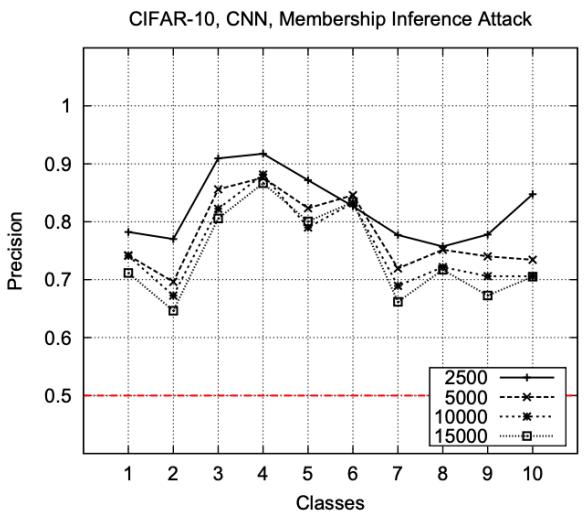
Step 3: Train the attack model

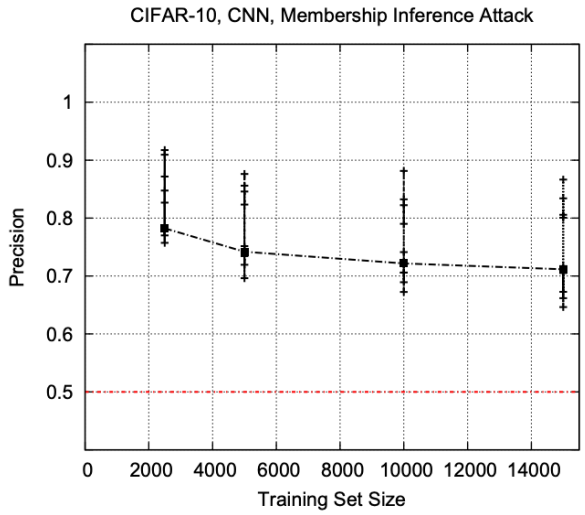


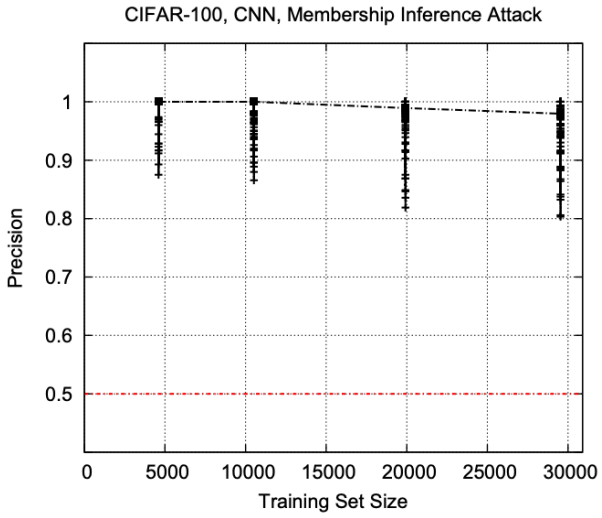
Step 3: Train the Attack Model



Results on CIFAR-10, CIFAR-100



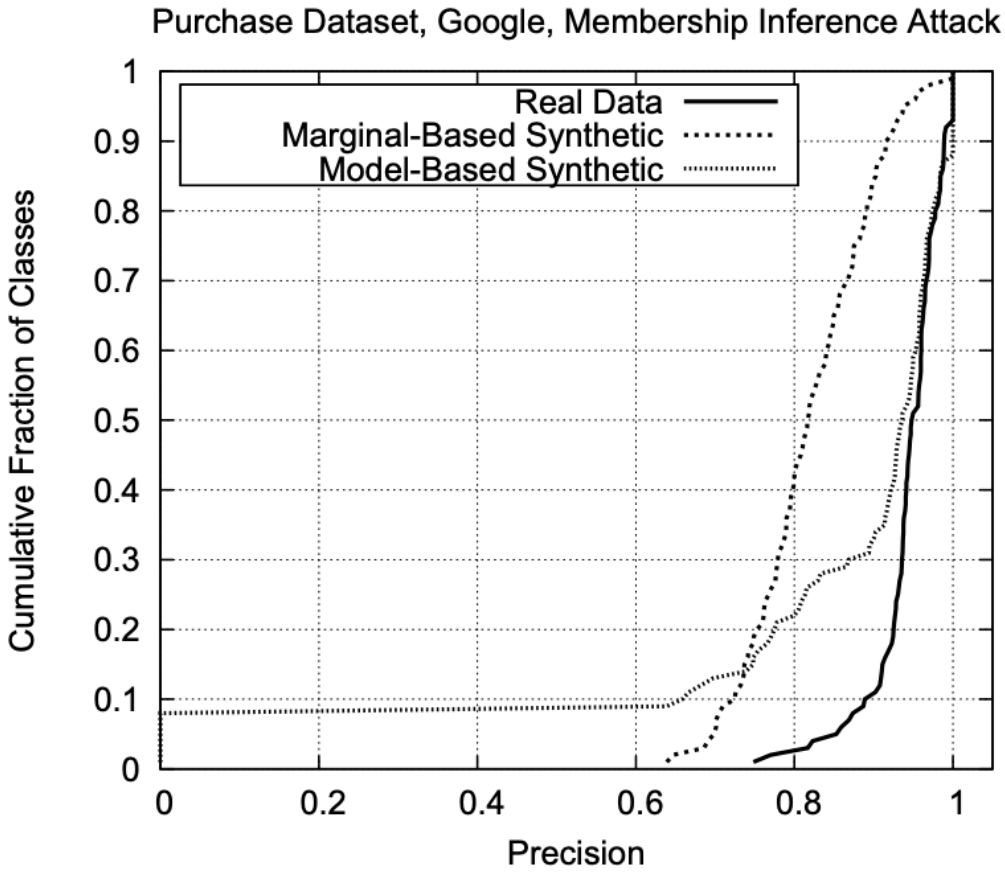




Note: Shadow models were trained with data from the real CIFAR dataset!

Results on (Simplified) Purchase Dataset

1. Take dataset of shopping histories over time
2. Extract 600 binary features (1 if item was purchased)
3. Cluster into consumer categories



Why did the authors only evaluate synthetic data on this dataset?

How effective are Membership Inference attacks?

- It depends
  - Complexity of original dataset
  - What auxiliary data you have available

- In practice, membership inference attacks are harder to execute than the literature makes it seem

Class 2: Model Inversion

Privacy in Pharmacogenetics:
An End-to-End Case Study of Personalized Warfarin Dosing

Matthew Fredrikson*, Eric Lantz*, Somesh Jha*, Simon Lin†, David Page*, Thomas Ristenpart*
*University of Wisconsin**, *Marshfield Clinic Research Foundation*†

Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures

Matt Fredrikson
Carnegie Mellon University

Somesh Jha
University of Wisconsin–Madison

Thomas Ristenpart
Cornell Tech

Background

- Warfarin – anticoagulant drug (prevents blood clots)
- Very difficult to dose
  - High mortality rate due to incorrect dosage
  - Too low – doesn't treat the underlying condition
  - Too high – uncontrolled bleeding
- The high variability in dosage requirements depends on two genes: VKORC1 and CYP2C9
- Medical literature: these 2 genes account for >50% of variability in dosage requirements
- So… let's use genetic markers to predict dosage!
  - Linear regression works as well as more complicated models

This paper

- What are the risks associated with releasing such models trained on private data?

- Adversary is given:
  - Predictive model
    - Input: genotype + attributes
    - Output: Warfarin dosage
  - Stable Warfarin dosage for victim
  - Other features of victim

- Adversary's goal: predict genotype attribute for individual
  - I.e., mutations in CYP2C9 and/or VKORC1

Setup

Dataset

$\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$

Age Weight Genetic Markers

$y \in \mathbb{R}$

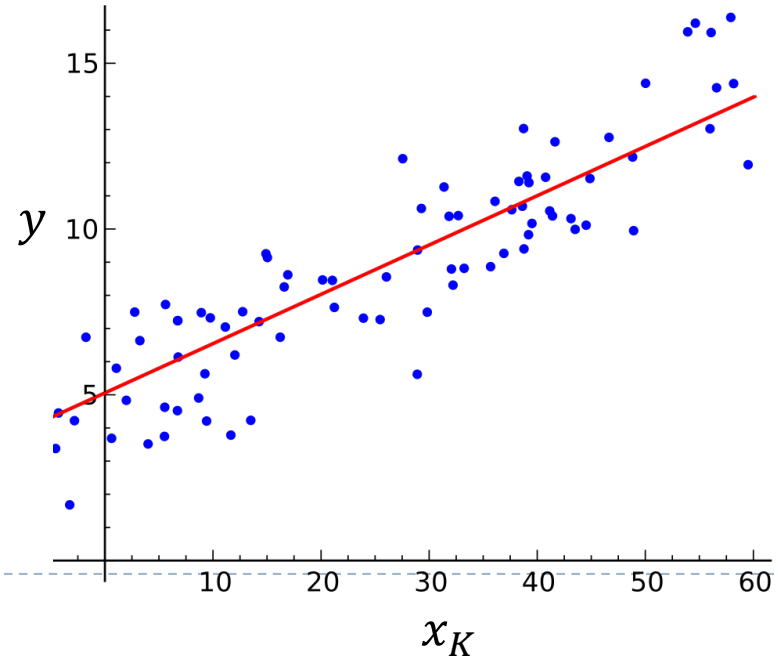
Warfarin Dosage

Adversary observes subset of features:

$x_K \subseteq \boldsymbol{x}$

Example:

$x_K = \{\text{Age}, \text{Weight}\}$



$f(x_K)$

Inference Algorithm

1. We have
   1. Input x_K^i = Input x_K for user i and Warfarin dosage y^i
   2. Trained model $f(x_K)$
   3. Marginals $P_i(x_i)$ for $i \in K$, and $P(y)$
2. We want
   1. To predict genetic marker for that sample, say x_d^i
3. Find the feasible set $\hat{X}$ such that for all $x \in \hat{X}$
   1. x matches x_K^i on all attributes in K
   2. The predictions match: $f(x) = y^i$
4. Return private attribute value that maximizes

$$\sum_{x \in \hat{X}} \prod_{1 \le i \le d} P_i(x_i)$$

Visualization of algorithm

- Example on document cam

Why is this the algorithm?

- Want the MAP estimate of hidden attribute:

$$P(x_d = u | x_K, y) = \frac{P(x_d, x_K, y)}{P(x_K, y)} = \frac{\sum_{x' \in \hat{X}: x_d = u} P(x', y)}{\sum_{x' \in \hat{X}} P(x', y)}$$

Problem: we don't know joint distribution!

Idea: Let's use marginals

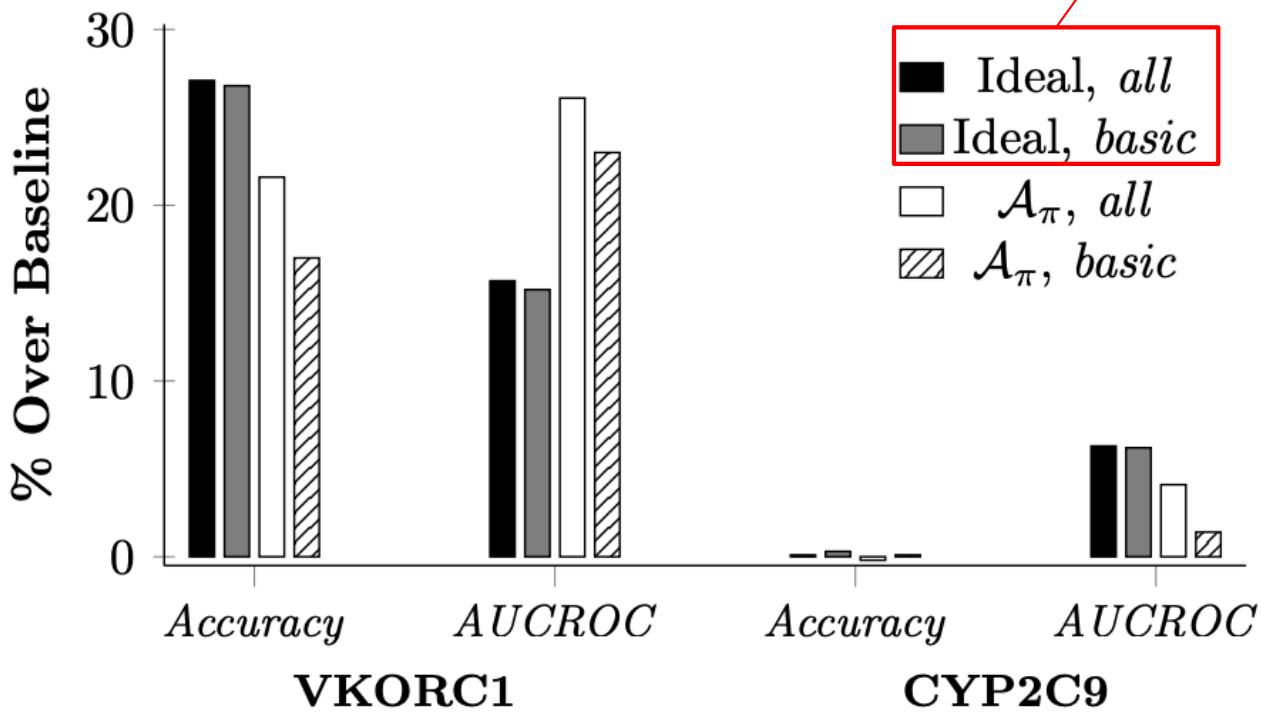
$$\frac{\sum_{x' \in \hat{X}: x_d = u} P(y) \prod_i P(x_i')}{\sum_{x' \in \hat{X}} P(y) \prod_i P(x_i')}$$

$$\propto \sum_{x' \in \hat{X}: x_d = u} \prod_i P(x_i')$$

Results: Non-private setting

Background info:
"all" = All x_i except genetic markers
"basic" = Only basic demographics (age, height, etc.)

Version of the algorithm we discussed



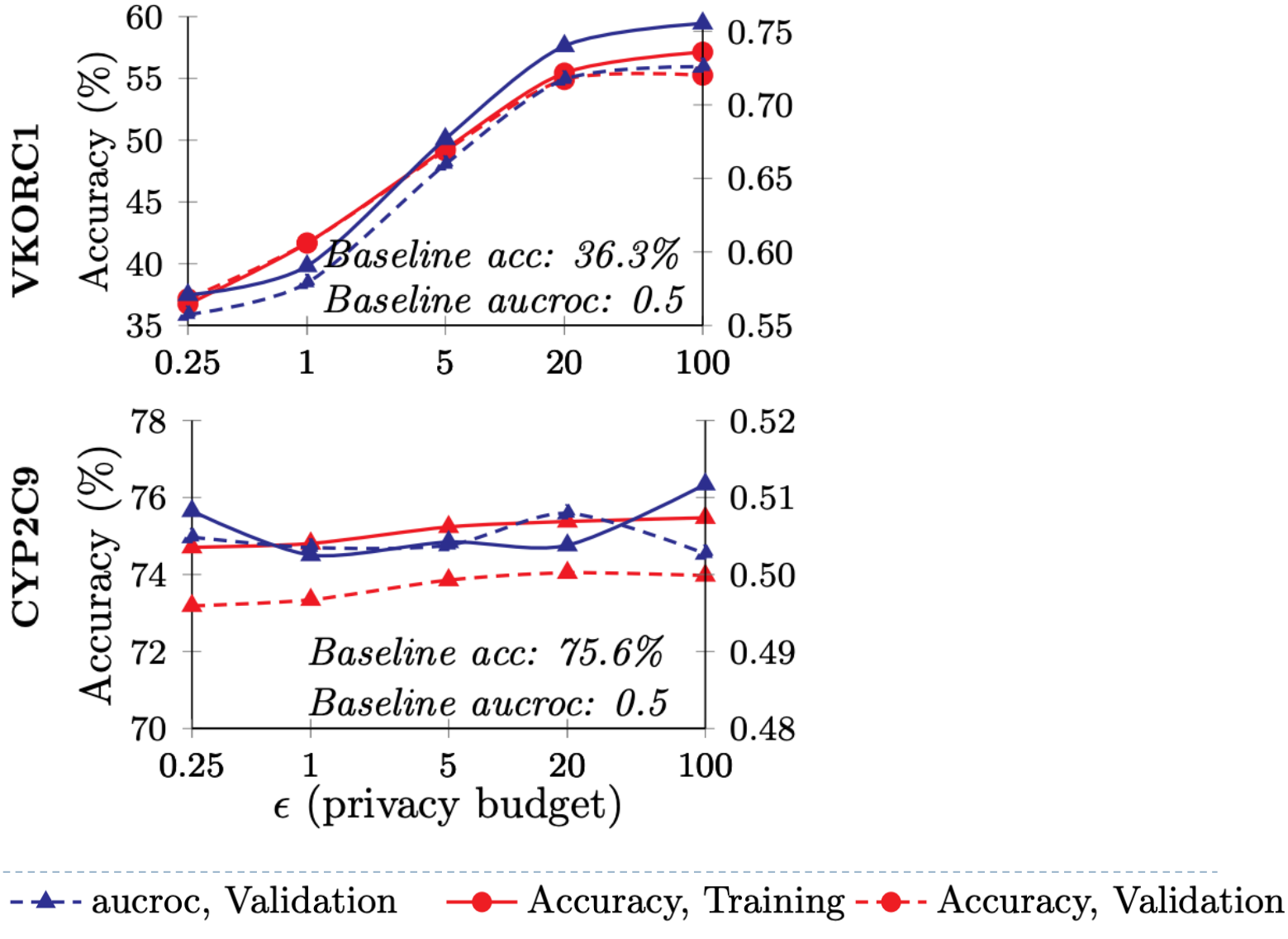
Now: Let's add differential privacy!

- Two approaches:
  - Differentially-private linear regression model
  - Converted data into differentially-private histograms before training

- How would you implement a DP linear regression model?
  - Add noise to coefficients
  - Add noise during training

- They added Laplacian noise to coefficients of the objective function
  - Clip values to limit sensitivity
  - J. Zhang, Z. Zhang, X. Xiao, Y. Yang, and M. Winslett. Functional mechanism: regression analysis under differential privacy, VLDB

Results with DP (Linear regression): Privacy



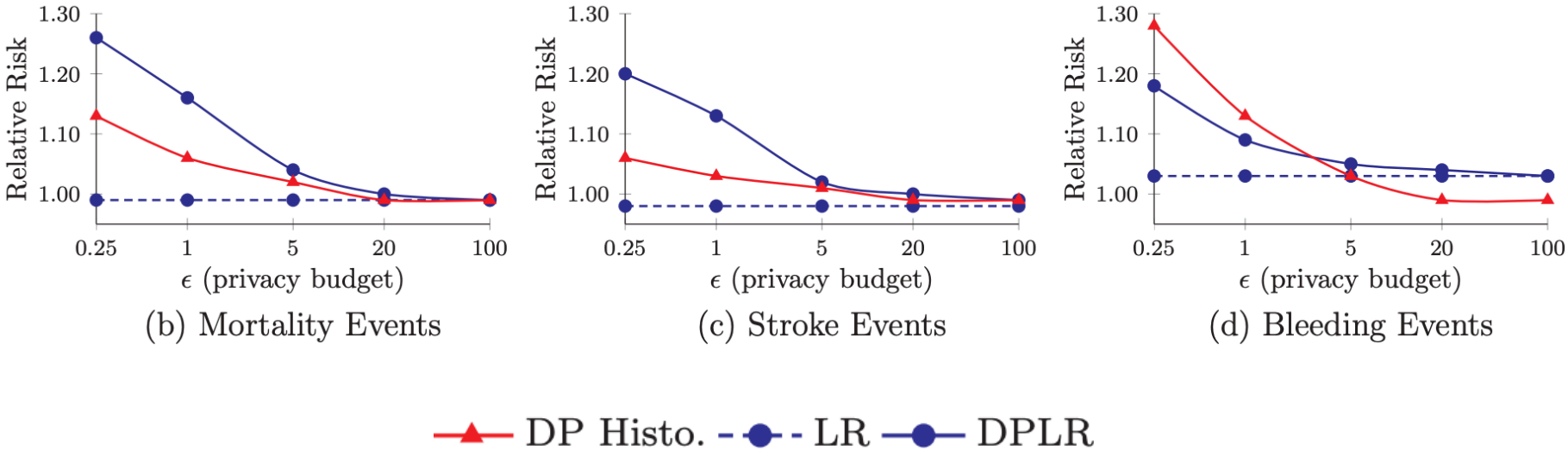
How do we measure utility?

- Simulate patient responses when using DP dosing prediction algorithm
  - Current clinical state-of-the-art dosing algorithm
  - Non-private regression model
  - DP regression model

- Clinical trial simulator draws random patients and applies each approach for 90 days

Simulation results

- Relative risk: ratio of patient's risk on new algorithm vs. fixed-dose algorithm



Response: Frank McSherry

- Recall: Frank McSherry = one of the inventors of DP

- "Strongly-worded" response to the paper
  - https://github.com/frankmcsherry/blog/blob/master/posts/2016-06-14.md

1. They define a privacy attack as performing statistical inference using (i) private personal data you disclose to the attacker and (ii) statistics about Warfarin dosing in other people, laying the blame on (ii) rather than (i). Unfortunately, (ii) is called "science", and (i) is you telling someone else something you shouldn't have. Their conclusion, roughly translated, is "science is hard to suppress, even with small epsilon". You are welcome.

2. They didn't actually use statistical inference when they applied it to their target domain, so they take patients off of the baseline treatment even when the confidence they should do so is not high. When epsilon is small, you should be leaving patients on the baseline treatment because you lack strong evidence to do anything else; it seems they mostly just randomly dose patients in this case. Mortality ensues.

The first reason is that model inversion misdiagnoses the source of the privacy violation: sharing your Warfarin dosage, or having it snooped from you, is what discloses information about your genetic markers. Their correlation as observed among large populations of people who are not you is *not* the source of your privacy woes. Model inversion is a non-attack; no one should care whether it is prevented or not.

The second reason is that the risk introduced over fixed dosing was primarily due to ignoring the statistical information about the differentially private measurements. The confidence associated with the measurement is (or should be) an important part of determining by how much you depart from your baseline treatment. That didn't happen in these experiments. The observed increased risk over fixed dosing is there because the use of statistical data without statistical techniques introduced it.

So… what do we make of this?

- Model inversion (non-private setting) is a real concern, even if an obvious one
  - ML models can leak information about training data
  - This is due to basic statistics
- No evidence that DP is broken
  - Conclusions based on DP models are consistent with DP guarantees
- **However:** Better tools needed for practitioners to **use** DP
  - E.g., how should I do inference based on noisy data?
  - Does DP guarantee that none of my customer data will leak? No