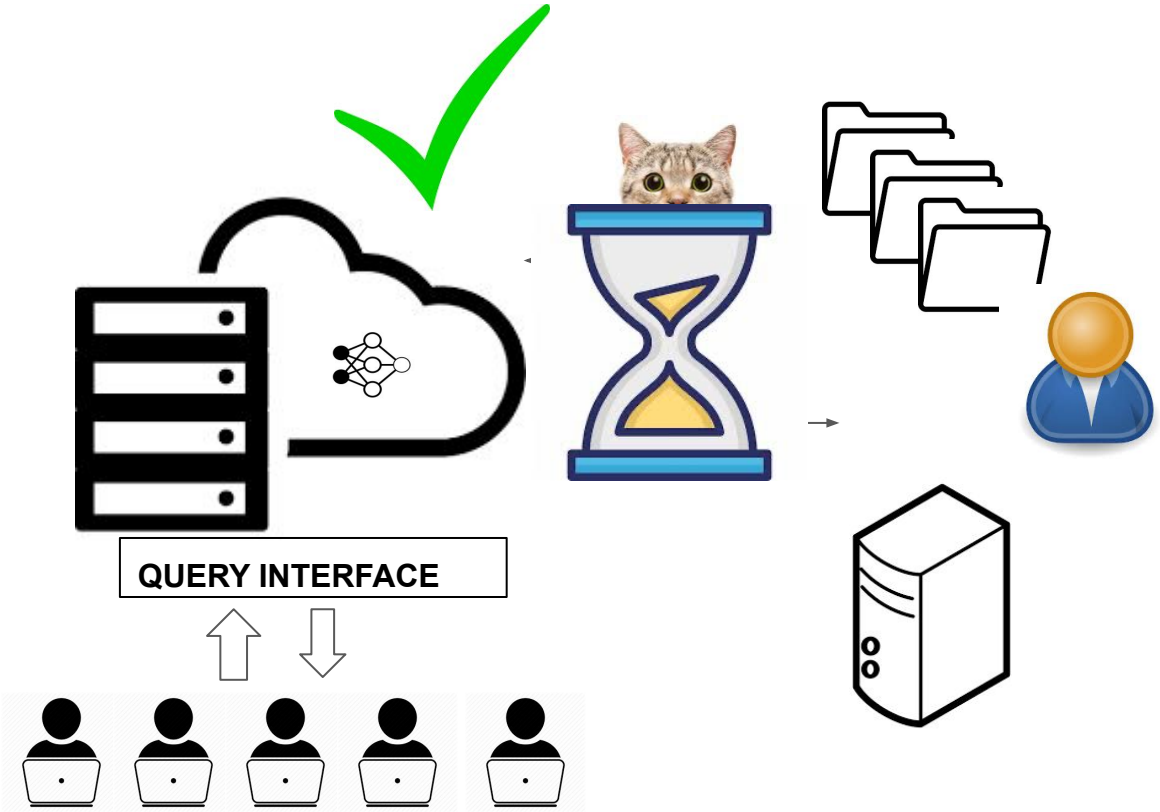
Exploring Connections between Model Extraction and Active Learning

Varun Chandrasekaran

Joint work with Kamalika Chaudhuri, Irene Giacomelli, Somesh Jha, Songbai Yan

Machine Learning as a Service (MLaaS)



Advantages

1.Elastic Scalability

2.Availability

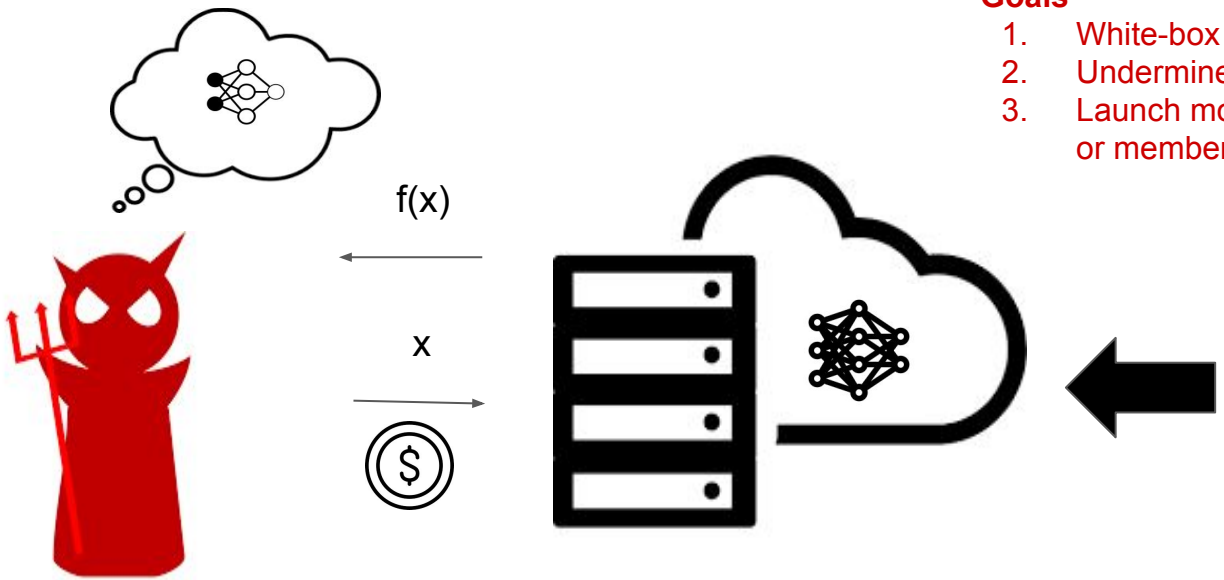
3.Offered by most Cloud Providers

4. Model is **monetizable**

Model Extraction

Proposed by Tramèr et al. [2016, USENIX Security]

Objective 1: Learn an *approximation* of the model
Objective 2: Use as *few* queries as possible

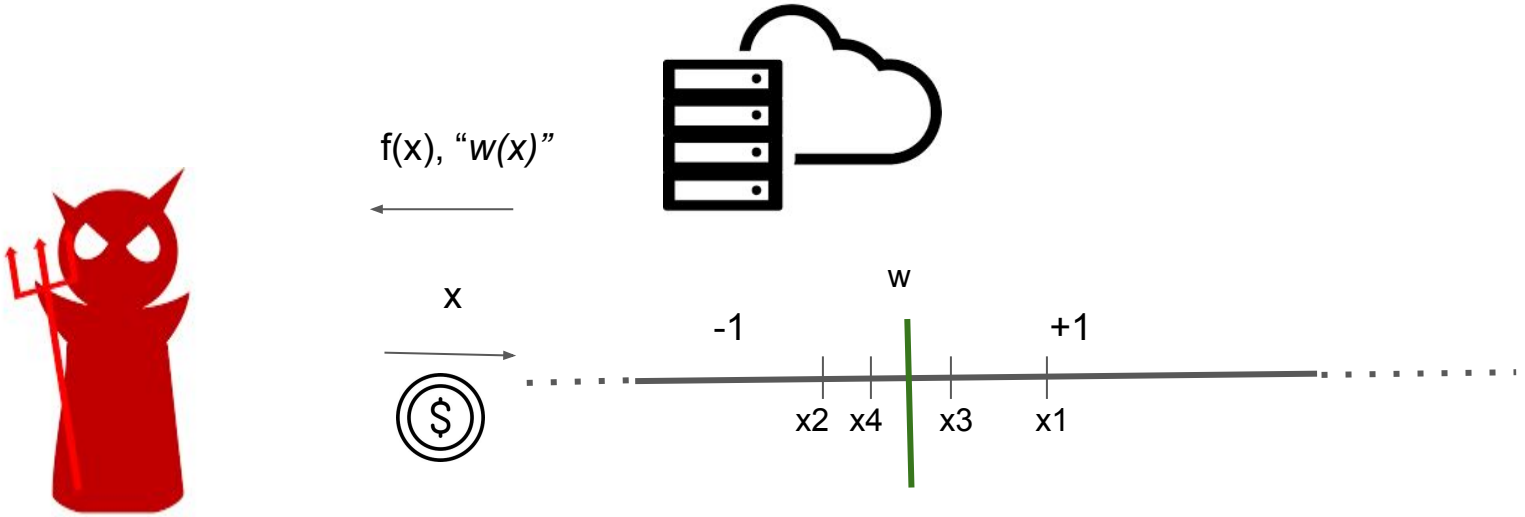


Goals

1. White-box evasion attacks
2. Undermine pay-per-query regime
3. Launch model inversion [Fredrikson et al.] or membership inference [Shokri et al.]

A Simple Example: Halfspace Extraction

Simple Strategy: Binary Search

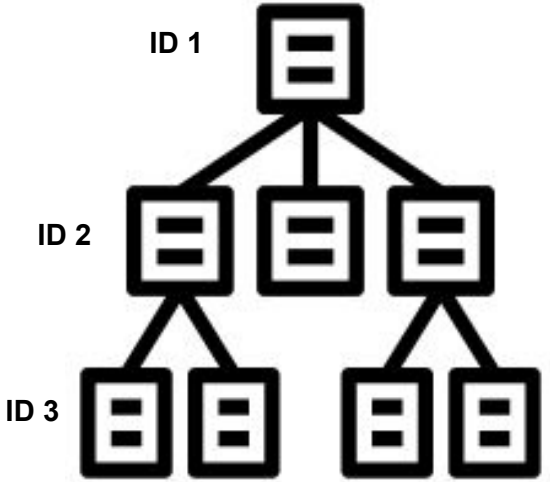


f(x) = -1 if w(x) = <w,x> < -1
+1 else

Assumptions Made In Tramèr's World

Label & auxiliary Information

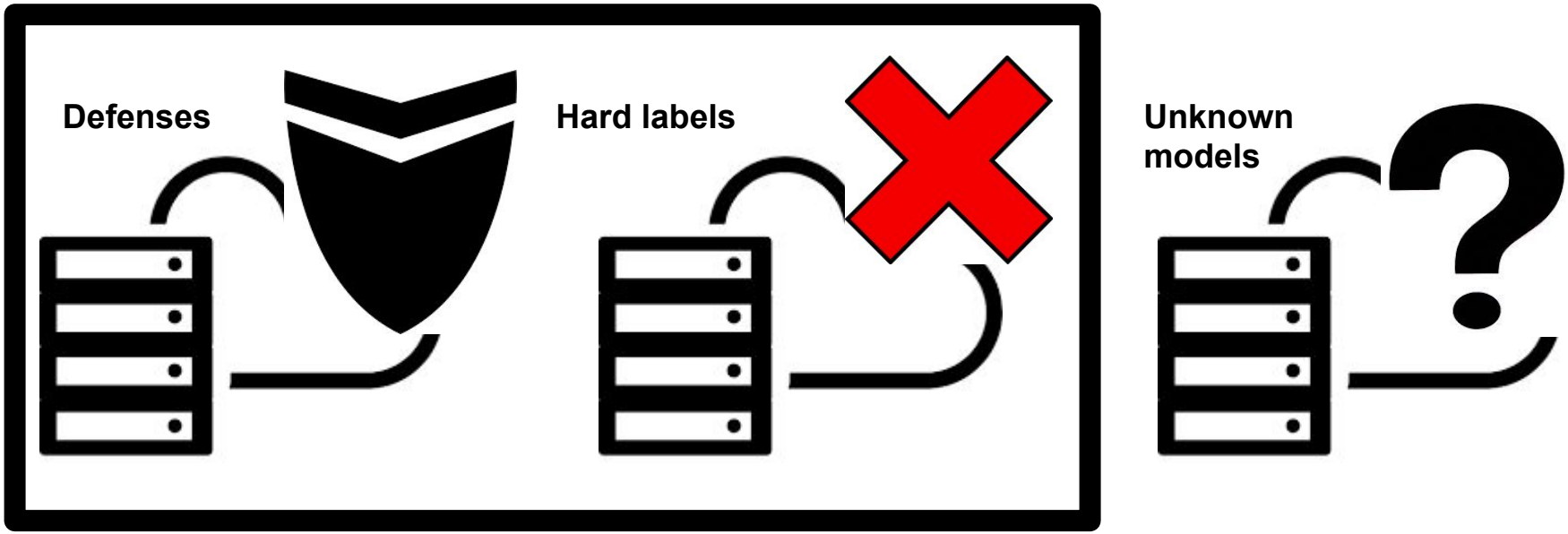
Oblivious to attacks



Uniquely identifiable subcomponents

Welcome To The Real World

All assumptions described earlier do not hold



Active Learning

Lower query complexity than passive learning

Bounds are known for certain hypothesis classes

Connection to Active Learning

Model Extraction ≅ Active Learning

How to Generate Queries?

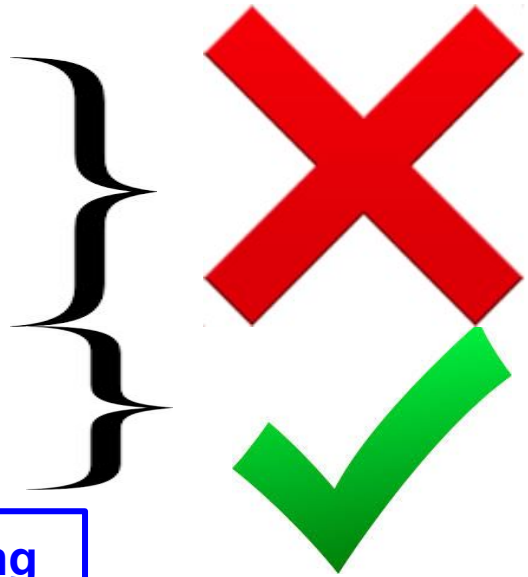
Strategy 1: Sampling from a pool of data

Strategy 2: Access to a *similar* dataset

Strategy 3: Data augmentation (using adversarial examples)

Strategy 4: Uniform data generation

Strategy 5: Query synthesis



Model Extraction ≅ Query Synthesis Active Learning

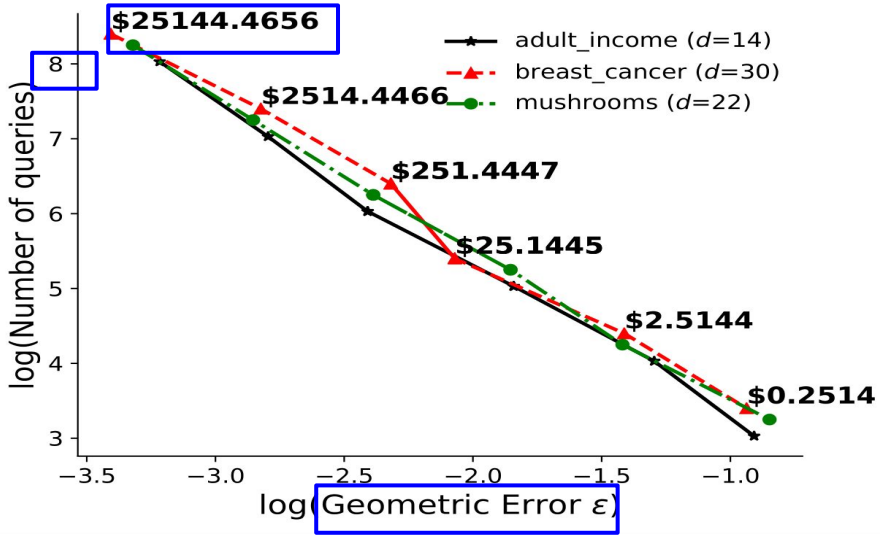
Main Results

- Halfspace extraction (linear models)
  - Spectral algorithm [Alabdulmohsin et al., 2015]
- Halfspace extraction (linear models)
  - Presence of noisy labels
  - Version Space Learning [Chen et al., 2018]
- Kernel SVM (non-linear models) extraction
  - Active Selection [Bordes et al., 2005]
  - More query efficient than Tramèr et al.

Refer paper for more details

No Free Lunch

- Model extraction is inevitable
  - Data independent randomization fails
  - Data dependent randomization fails using ***passive learning*** *approaches*



Decision Tree Extraction

- Kushilevitz & Mansour [1993]
  - Boolean trees
  - Slow
- Importance Weighted Active Learning [Beygelzimer et al., 2009]
  - Uniformly generated inputs

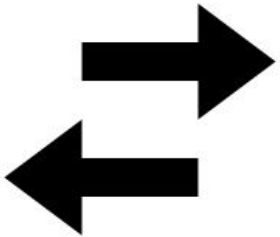
- **In the absence of identifiers for tree nodes**
- **In the absence of operation on "incomplete inputs"**

Larger query complexity, but NO auxiliary information

Dataset	Oracle Accuracy	Path Finding (Tramèr 2016) #Queries	IWAL Accuracy	IWAL (Us) #Queries
Adult	81.2%	18323	80.2%	244188
Steak	52.1%	5205	73.1%	1334
Iris	86.8%	246	89.4%	361
GSSHappiness	79%	18907	79.3%	254892

Summary

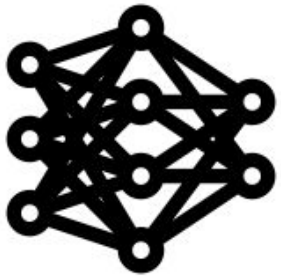
- Draw connections between Active Learning and Model Extraction
  - Attacks with known asymptotic bounds
  - Robust to noise



- Provide attacks under more realistic assumptions

- No Free Lunch *i.e.,* Model Extraction is inevitable

Open Questions



QSAL for DNNs

Determining model type

Transferability

THANK YOU
FOR
YOUR
ATTENTION
ANY QUESTIONS?

chandrasekaran@cs.wisc.edu