The Limits of Provable Security Against Model Extraction

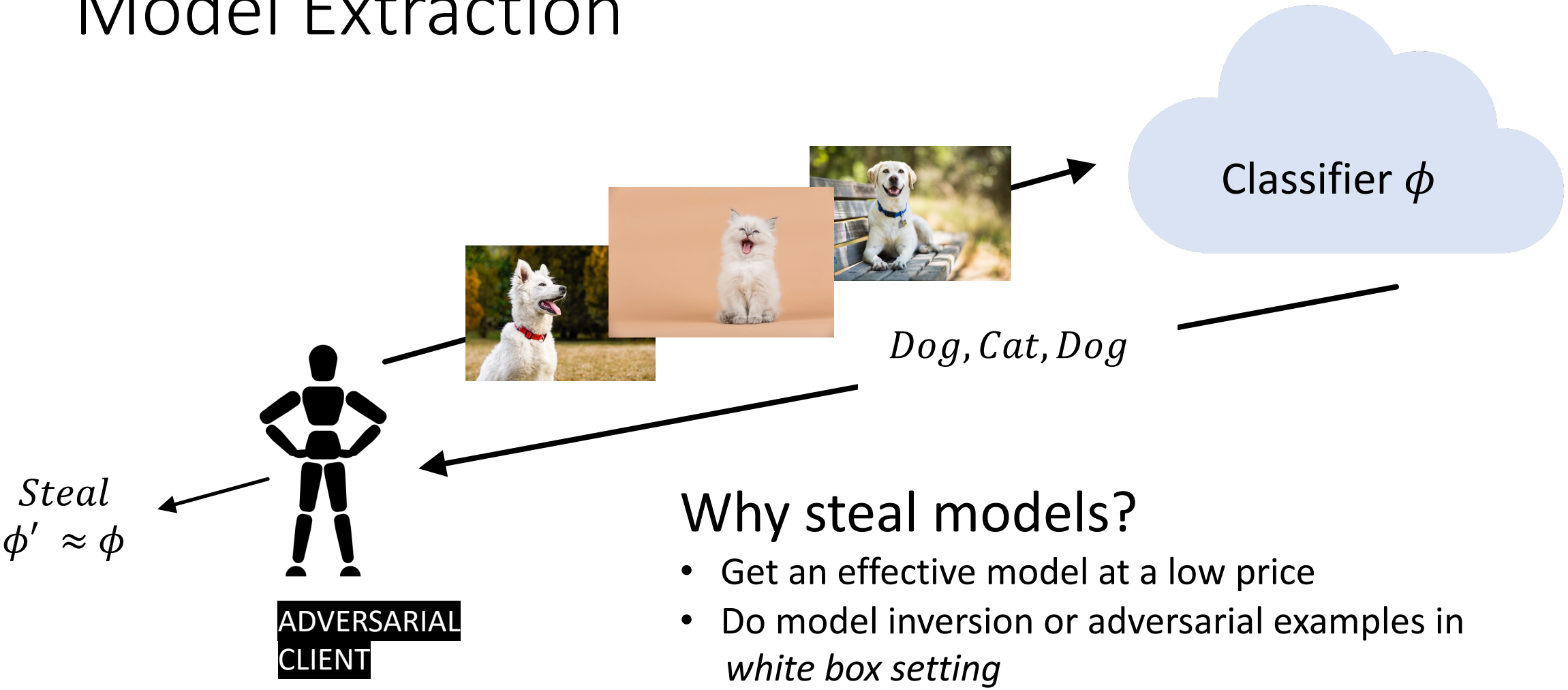
Ari Karchmer

arika@bu.edu

Based on https://ia.cr/2021/764 (Canetti, Karchmer TCC ’21)

https://ia.cr/2022/1039 (Karchmer ’22)

Model Extraction



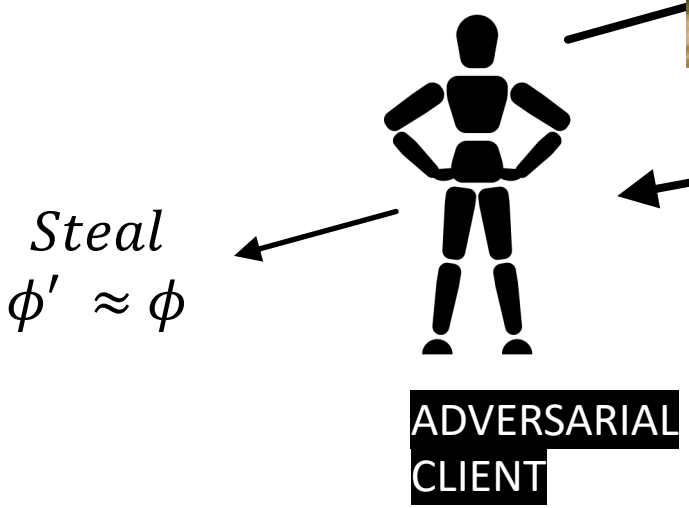
Why steal models?

- Get an effective model at a low price
- Do model inversion or adversarial examples in *white box setting*
  - E.g. [BCM+13] for adversarial malware examples
  - [FJR15] for image reconstruction from facial recognition models

Model Extraction Defenses

Noisy Classifier ϕ

Dog, Cat, Cat



$Steal$
$\phi' \approx \phi$

- Inject Noise
  - i.i.d. OR deliberately modify

- Use Cryptography
  - Secure Function Evaluation:
    - E.g. FHE [GBDL+16], Garbled Circuits [RWT+18]

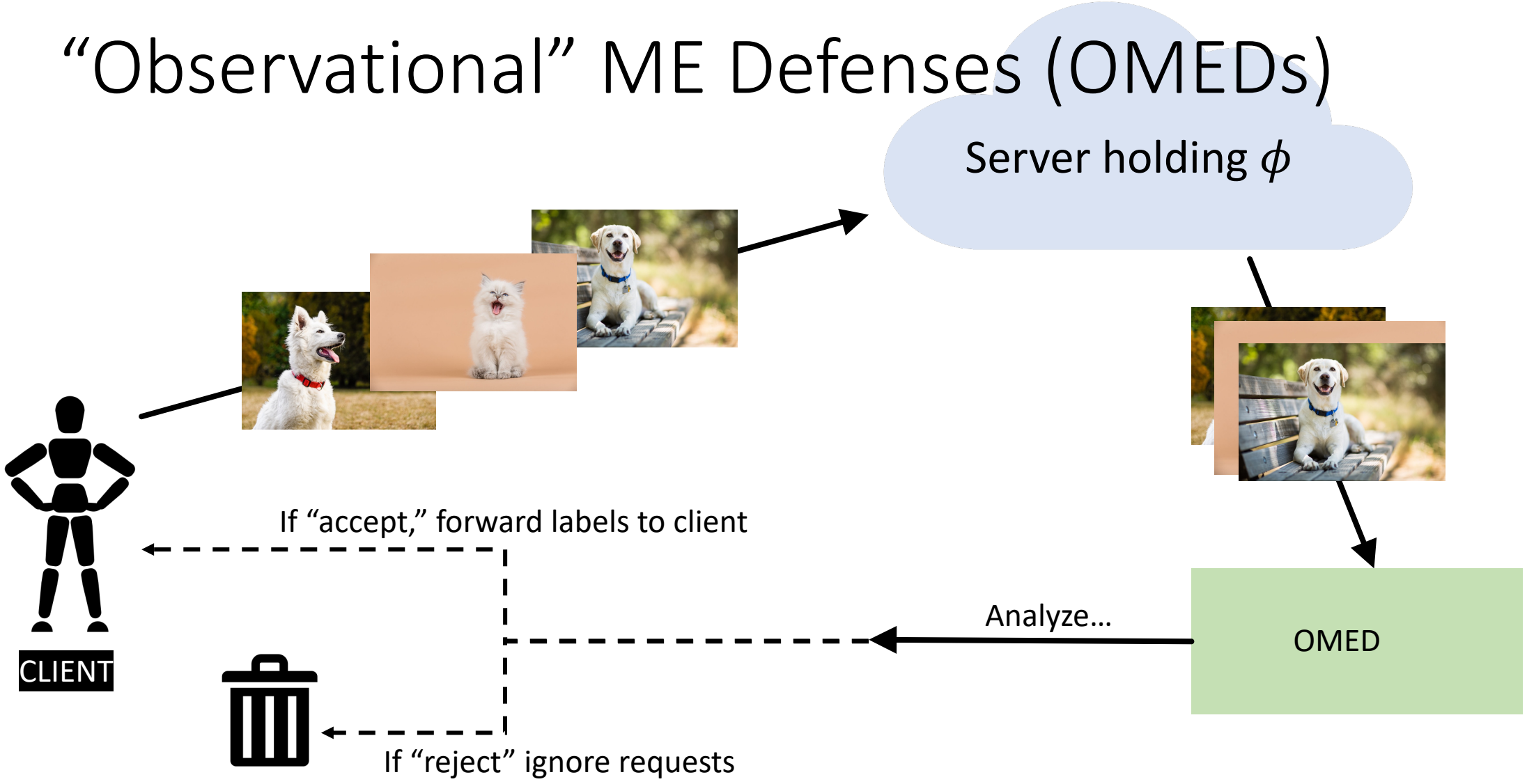
Alice (holding ϕ) and Bob (holding x) jointly compute ϕ(x)

Alice learns nothing about x,
Bob learns nothing about ϕ
beyond what is revealed by ϕ(x)

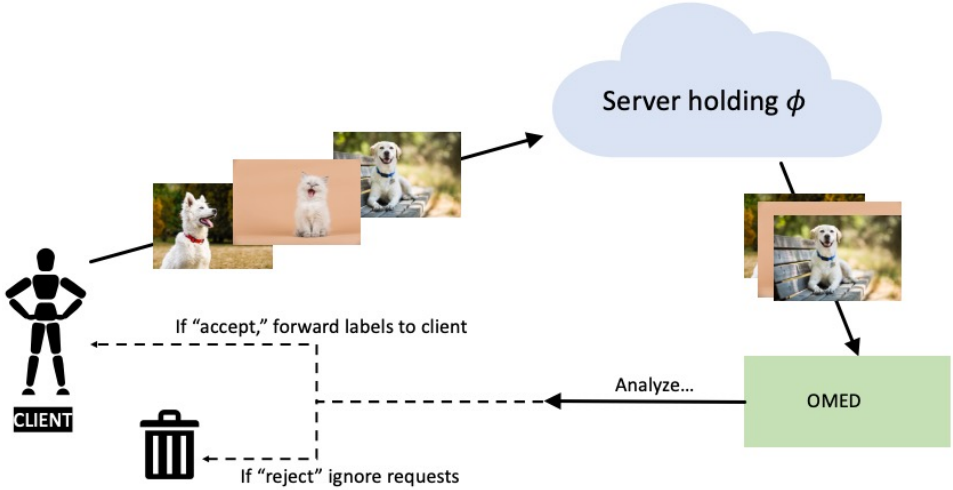
- Program Obfuscation

SFE, Obfuscation:
"Ideal World is not secure against model extraction" [Vai21]

“Observational” ME Defenses (OMEDs)



OMEDs



- Determine if client is adversarial or "benign"
  - E.g. Extraction Monitor [KMAM18], PRADA [JSMA19], VarDetect [PGKS21]

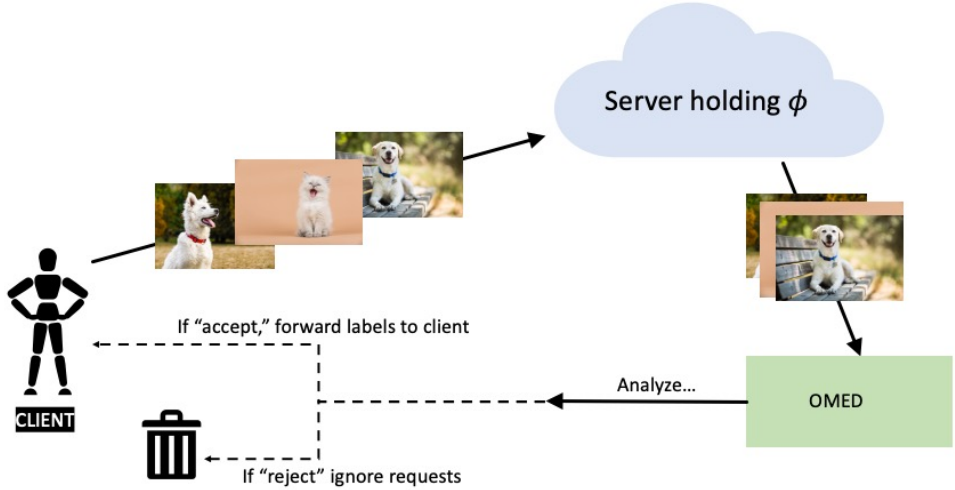
Estimate information gained by client by training proxy model based on queries. Warn model owner when information gain passes a threshold.

Conduct normality test on Hamming distances between queries, and rejects deviating clients. Assumes honest clients have this "benign" property.

Uses Variational Autoencoder to map "problem domain" queries and "outlier" queries to distinct regions in latent space. Classifies clients as honest or adversarial.

OMEDs

Unfortunately, no known security guarantees.



All these systems are efficient statistical tests on client query distributions.

Estimate information gained by client by training proxy model based on queries. Warn model owner when information gain passes a threshold.

Conduct normality test on Hamming distances between queries, and rejects deviating clients. Assumes honest clients have this "benign" property.

Uses Variational Autoencoder to map "problem domain" queries and "outlier" queries to distinct regions in latent space. Classifies clients as honest or adversarial.

Provable Security?

- Zero Knowledge style: “Client learns nothing from the interaction beyond what it could efficiently learn prior to the interaction”
  - Too harsh, client needs to at least learn some classifications
  - More realistic: “client learns nothing beyond what can be efficiently deduced from some random examples?”

OMEDs: classify according to some statistical property….

…In order to confine clients to *specific distributions.*

OMEDs are already implicitly using this security model! But how to prove it…

Towards Provable Security for OMEDs

- General OMED: PPM that given input a sequence of queries, outputs accept or reject

We want…

”Completeness” – Honestly distributed queries are accepted
“Soundness” – Adversarially distributed queries are rejected

Abstract honest property $\mathbb{P}$

δ- Completeness w.r.t. $\mathbb{P}$

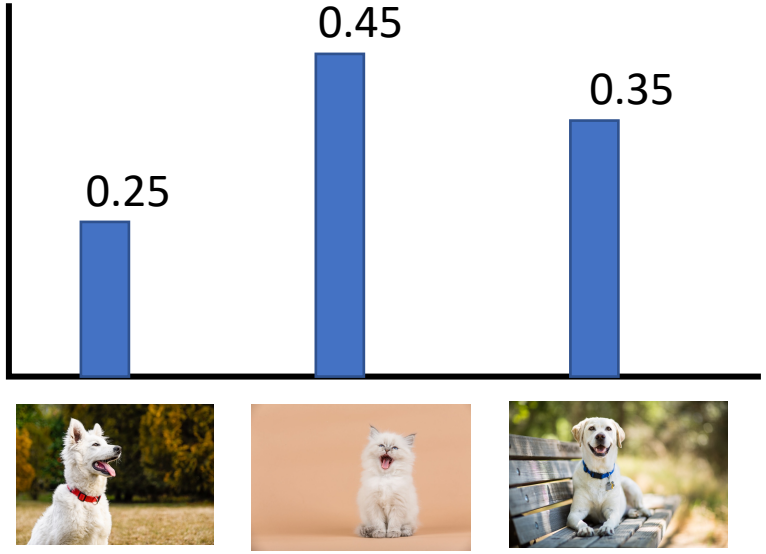
$$\Pr_{Q \sim \mathbb{P}}\left[M(Q) = \text{”accept”}\right] > 1 - \delta$$

Set $\delta > 0.75$?

γ- Soundness w.r.t. $\mathbb{P}$

$$\Pr_{Q \nsim \mathbb{P}}\left[M(Q) = \text{“accept”} \mid \phi \text{ extracted}\right] < \gamma$$

Set $\gamma <$ negligible



- OMED is a distribution tester – test for “honest” property

Towards Provable Security for OMEDs

- General OMED: PPM that given input a sequence of queries, outputs accept or reject

We want…

”Completeness” – Honestly distributed queries are accepted

“Soundness” – Adversarially distributed queries are rejected

Abstract honest property $\mathbb{P}$

δ- Completeness w.r.t. $\mathbb{P}$	γ- Soundness w.r.t. $\mathbb{P}$
$\Pr_{Q \sim \mathbb{P}}\left[M(Q) = \text{”accept”}\right] > 1 - \delta$	$\Pr_{Q \nsim \mathbb{P}}\left[M(Q) = \text{“accept”} \mid \phi \text{ extracted}\right] < \gamma$
Set $\delta < 0.25$?	Set $\gamma <$ negligible

How to choose $\mathbb{P}$?

- Want some distribution that makes extraction hard
- Hardness assumptions for average-case PAC-learning

A Provable Security Lemma

Let M be a complete and sound OMED w.r.t. $\mathbb{P}$.

If $\mathcal{C}$ has no efficient average-case PAC-learning algorithm on distributions in $\mathbb{P}$, then most models in $\mathcal{C}$ cannot be extracted except with negligible probability.

E.g. polynomial size decision trees, constant depth threshold circuits w.r.t. uniform (learnable with membership queries)

- When client queries honestly, model is protected by hardness assumption
  - Client learns "whatever can be efficiently deduced from $\mathbb{P}$-queries"
  - We accept this because of hardness assumption

- When client queries adversarially, model is protected by soundness of OMED… labels are never returned (except negligibly often)

Can’t instantiate provable security lemma

Can We Build OMEDs?

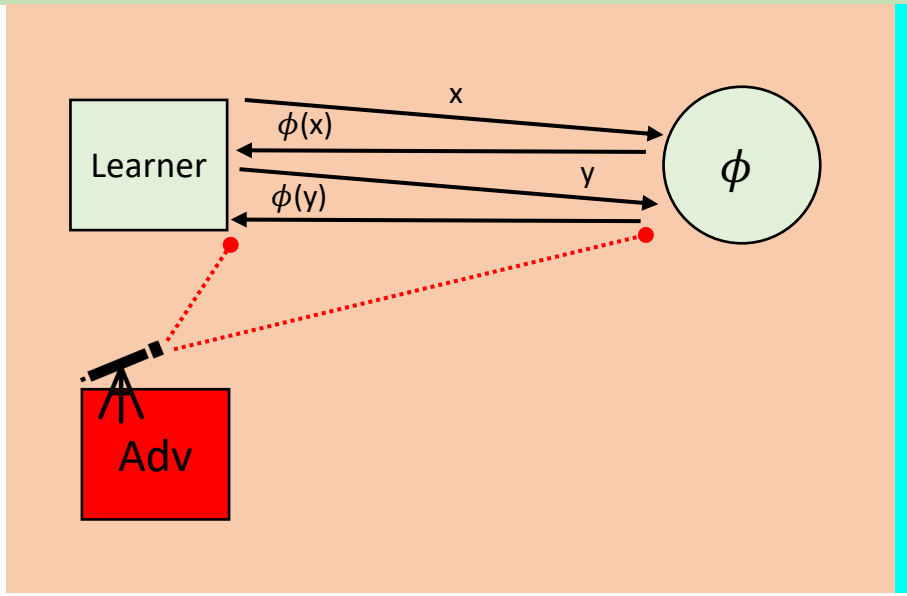
There also exist less realistic ”covert learning attacks:” [Vai21] discussed a similar attack against noisy linear models using Lattice Trapdoors [Ajt96] [GPV08].

- [Kar22]: Not efficiently, for decision trees, unless

“Computational Incompleteness Theorem”

Any p.p.t. complete OMED for DTs cannot be sound, unless LPN does not hold.

- Use Covert Learning algorithms from [CK21] for DTs
  - MQ-learning against an adversary
  - Learn ϕ while preventing adversary from learning
    - "Adversary can learn only as much as random examples revealed”
    - Initially motivated by “secure outsourcing” of drug design



Idea: Covert Learning algorithms fool OMEDs

δ-complete, efficient OMED w.r.t. uniform property must be far from δ-sound

Plenty of interesting directions…

- Find more interesting Covert Learning attacks
  - More realistic class of models (e.g. Neural Nets)
  - More practical Covert Learning attacks
- Theory of Model Extraction under OMEDs
  - A fundamental TCS problem: relating Crypto and CoLT
  - Can show (roughly) *nonexistence of Covert Learning* implies complete + sound OMEDs
  - Interesting: In Minicrypt, there are complete + sound OMEDs

Based on https://ia.cr/2021/764 [CK21]

https://ia.cr/2022/1039 [Kar22]

THANKS

QUESTIONS?