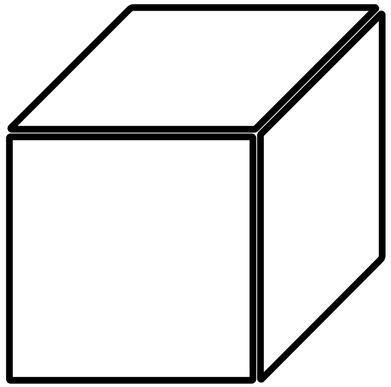
CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

Model Stealing I

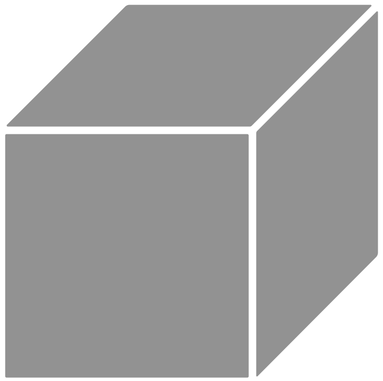
Tutorial Session 4 - Trustworthy Machine Learning

Adam Dziedzic & Franziska Boenisch
Nima Dindarsafa | 20.05.2026

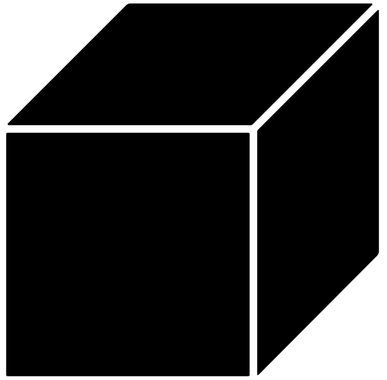
Model Stealing - Basics



White-Box

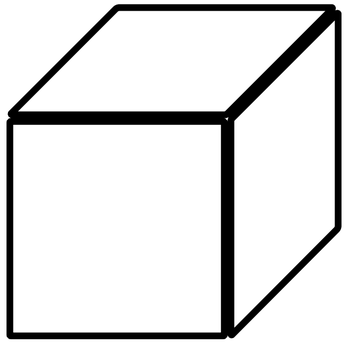


Gray-Box



Black-Box

Model Stealing - Basics



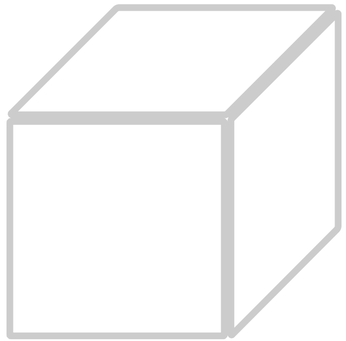
White-Box

Gray-Box

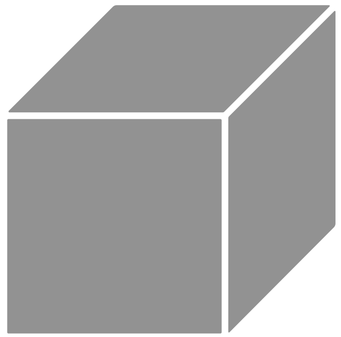
Black-Box

- Access to the model **internal representations**
- Access to the model **gradients**
- In the case of white-box **open data** also access to the **training data**

Model Stealing - Basics



White-Box

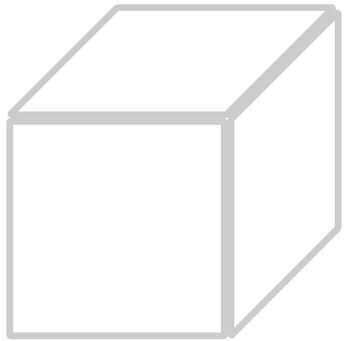


Gray-Box

Black-Box

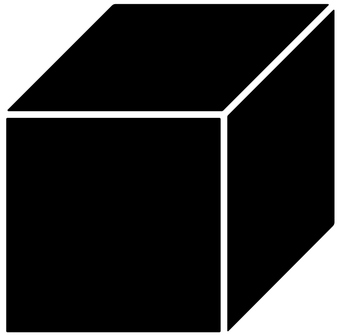
- Access to the **loss** values, model **confidence** for LLMs **log probabilities**
- Access to model **architecture** or model **family**

Model Stealing - Basics



White-Box

Gray-Box



Black-Box

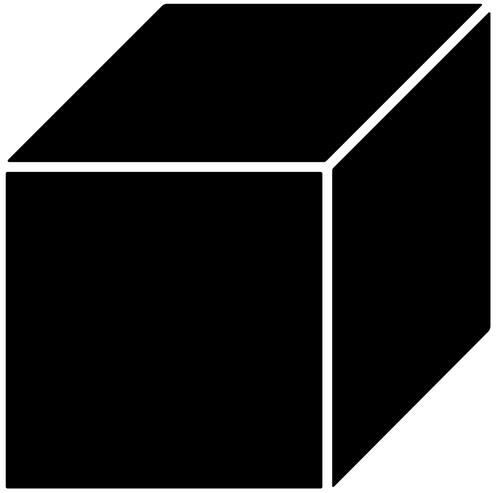
- Access only to the **predicted label** by the model

Model Stealing - Basics

Is black-box access restrictive?

Yes! **BUT**:

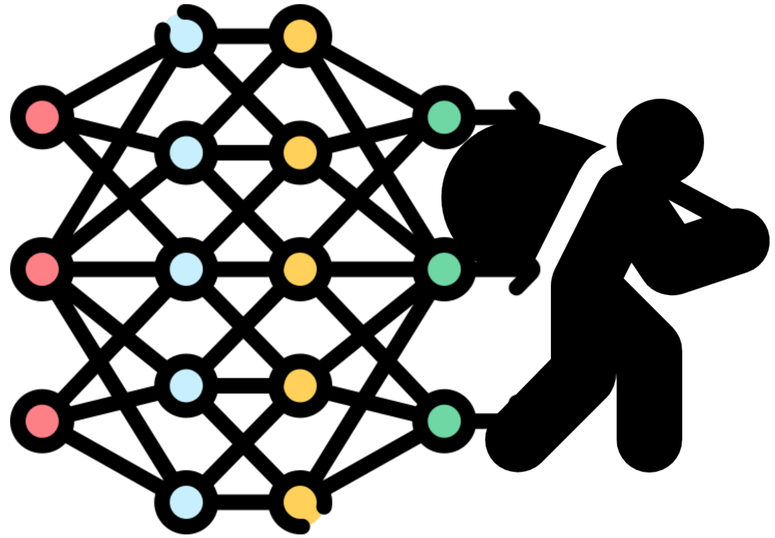
- This does not stop the attack!
- The adversary can still steal functionality of the target model.
- The target model acts as a labeling oracle.



Black-Box

Model Stealing - Definition

The **model extraction attack** is defined as the case where an adversary steals a copy of a remotely deployed machine learning model, given oracle prediction access.



Model Stealing - Goals

Model Stealing is not only about stealing the functionality!

Theft

The adversary wants to obtain a **useful substitute model** that performs well on the victim's task.

Reconnaissance

The adversary wants to learn about the victim model's behavior in order to **support another attack**.

Model Stealing - Goals

Model Stealing is not only about stealing the functionality!

Theft

Reconnaissance

In **Theft**, the stolen model is the attacker's end goal. In **Reconnaissance**, the stolen model is a tool for learning about or attacking the victim system later.

Model Stealing - Objectives

Accuracy

The stolen model performs well on the true task.

Fidelity

The stolen model imitates the victim model's predictions.

Functionally Equivalent

The stolen model matches the victim model for every possible input.

Model Stealing - Objectives

Why is learning a Functionally Equivalent neural network hard?

- Consider a network of n layers with d neurons.
- If the network uses an activation function of odd symmetry, then:
  - There are $2^d \times n!$ functionally equivalent networks!

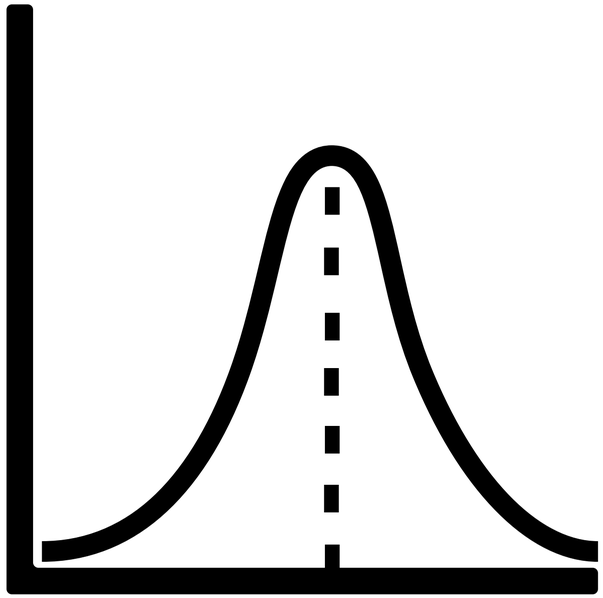
MixMatch - Sharpening

$$\text{sharpen}(p, T)_i = \frac{p_i^{\frac{1}{T}}}{\sum_j p_j^{\frac{1}{T}}}$$

- $T = 1$: The distribution remains the same
- $T < 1$: The larger probabilities become larger and smaller ones get smaller
- $T \to 0$: The distribution becomes almost one-hot
- $T > 1$: The distribution becomes more uniform

MixMatch - Sharpening

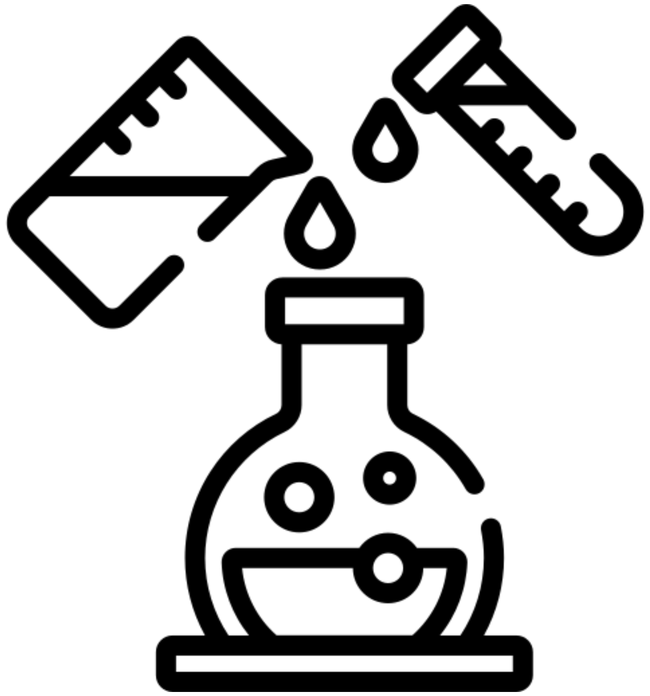
Sharpening makes these guessed labels **more confident**, which encourages the model to **learn clearer class boundaries** instead of treating unlabeled examples as belonging partially to many classes.



MixMatch - MixUp

Why does MixMatch use **MixUp** after generating text labels for unlabeled data?

1. Prevents the model from memorizing noisy guessed labels
2. Encourages smoother decision boundaries
3. Improves generalization
4. Makes the model more robust when learning from unlabeled data



Estimating Privacy Leakage

The model returns the following probability distribution for query q:

$$p_\theta(y \mid q) = [0.10, 0.40, 0.20, 0.30]$$

Margin Leakage:

$$\text{Margin Leakage}(q) = 1 - (p_{\text{max}} - p_{\text{second-max}})$$

$$\text{Margin Leakage}(q) = 1 - (0.40 - 0.30) = 1 - 0.10 = 0.90$$

Estimating Privacy Leakage

The model returns the following probability distribution for query q:

$$p_\theta(y \mid q) = [0.10, 0.40, 0.20, 0.30]$$

Entropy Leakage:

$$\text{Entropy Leakage}(q) = -\sum_i p(y_i \mid q) \log_2(p(y_i \mid q))$$

$$\text{Entropy Leakage}(q) =$$

$$= -[0.10 \times \log_2(0.10) + 0.40 \times \log_2(0.40) + 0.20 \times \log_2(0.20) + 0.30 \times \log_2(0.30) = 1.846$$

Estimating Privacy Leakage

The model returns the following probability distribution for query q:

$$p_\theta(y \mid q) = [0.10, 0.40, 0.20, 0.30]$$

Privacy Leakage:

$$\text{Privacy Leakage}(q) = \frac{1}{2} \sum_{i \neq i^*} \text{erfc}\left(\frac{n_{i^*} - n_i}{2\sigma}\right)$$

$$\frac{n_{i^*} - n_i}{2\sigma} = \frac{40 - 30}{2(5)} = 1$$

$$\text{Privacy Leakage}(q) = \frac{1}{2}(0.157) = 0.0785$$

Questions?