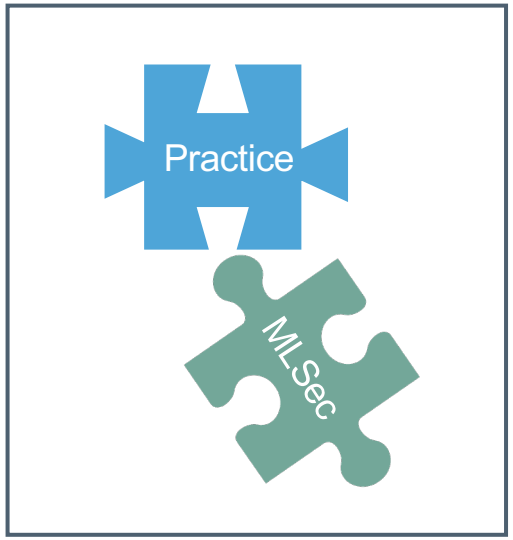
EPFL VITA

Towards More Practical Threat Models in AI Security

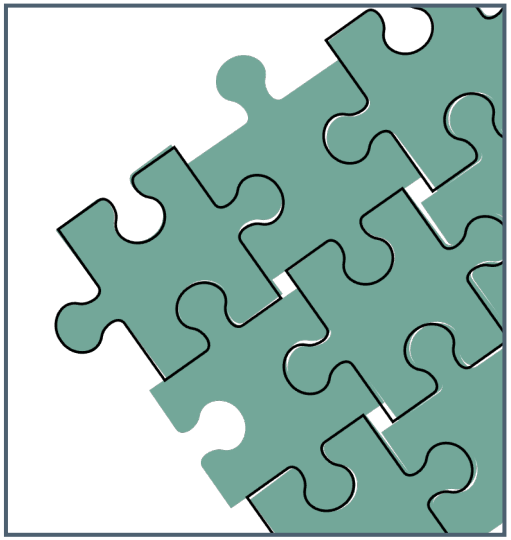
Kathrin Grosse, Lukas Bieringer, Tarek R. Besold, Alexandre Alahi

Outline



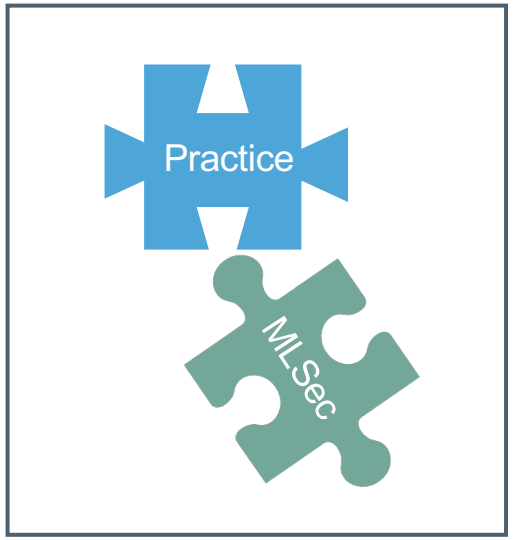
Practical AI Security

Practical AI Threat Models



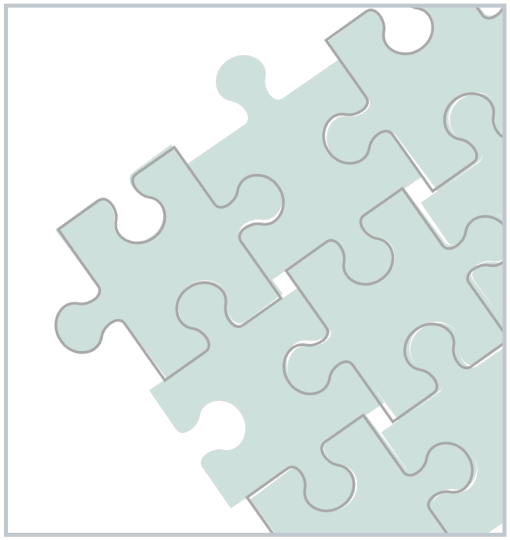
The Big Picture

Outline



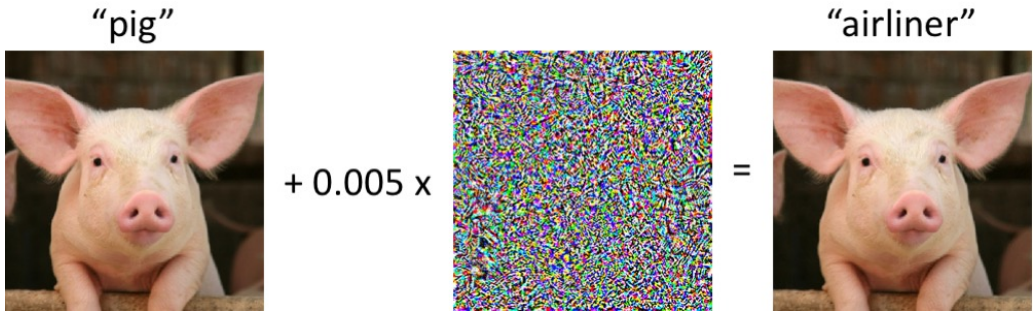
Practical AI Security

Practical AI Threat Models



The Big Picture

ML Sec – Evasion / Adversarial Examples



Madry and Schmidt, https://gradientscience.org/intro_adversarial/

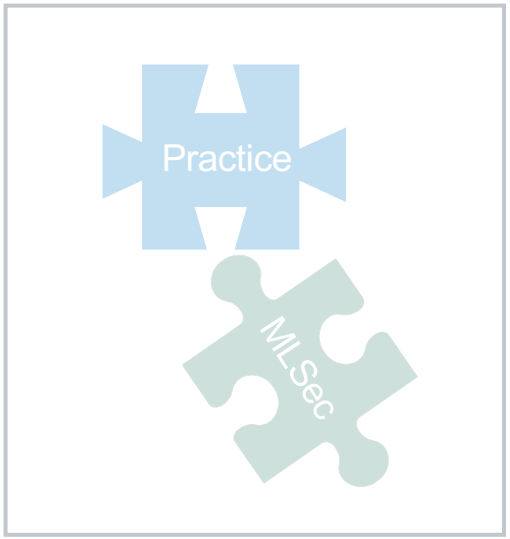
Practitioners' Perspective

- Asking **practitioners** why **evasion** is **irrelevant**

- **Access control on data**
- Irrelevant in use-case
- Doubting attacker

Grosse et al. Machine Learning Security in Industry: A Quantitative Survey" *TIFS 2023*.

Outline



Practical AI Security

Practical AI Threat Models

The Big Picture

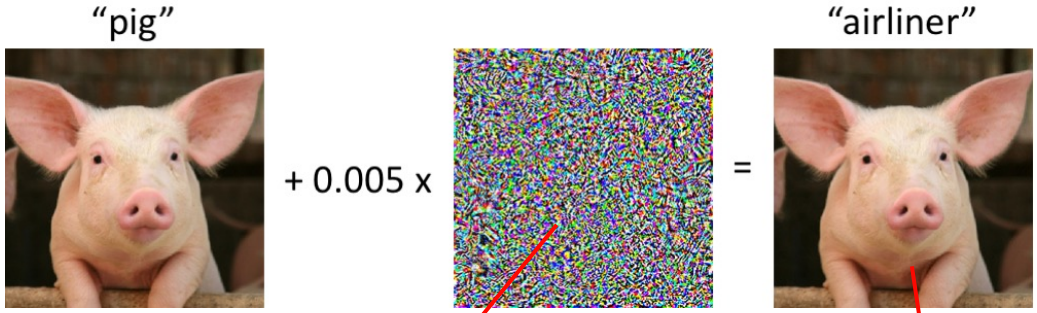
How to align threat models

- Derive most **common threat models** from research
- Collect information in industry like
  - **Access** to **data**, **model**
  - **Data sources**
  - ….
- **Match** these!

Sample

- **271** participants, recruited 2023
- AI engineers, data engineers, anyone working with AI (not just chatGPT)
- Sample is **representative** of AI engineer population

Evasion Threat Model

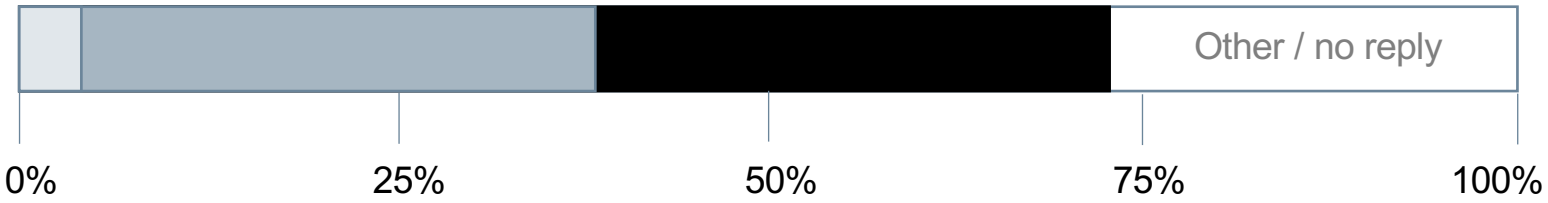


Gradients: Model, Inputs, outputs – **white-box**
only inputs, outputs – **black-box**

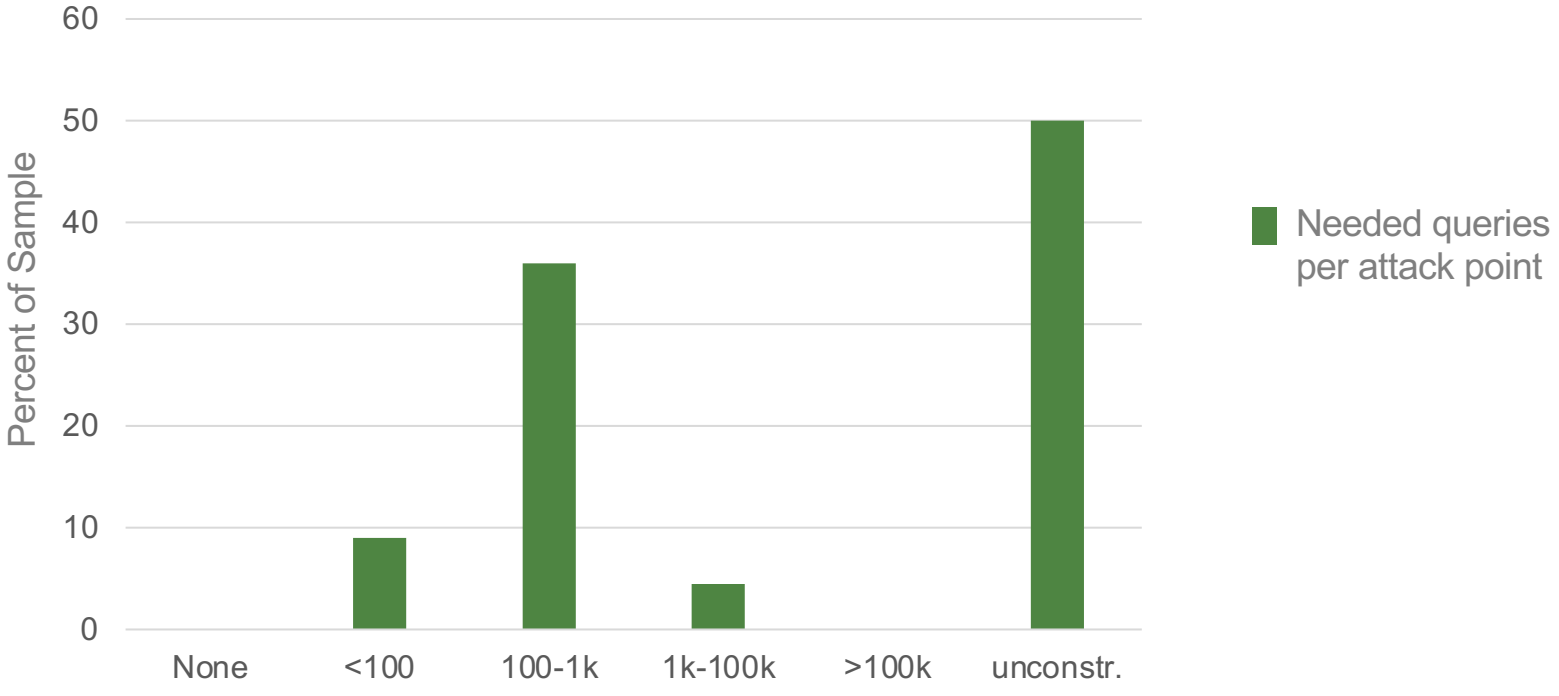
Submit test data

Matching Threat Models

- Model and queries – white-box evasion
- Queries possible – black-box evasion
- **Nothing accessible**

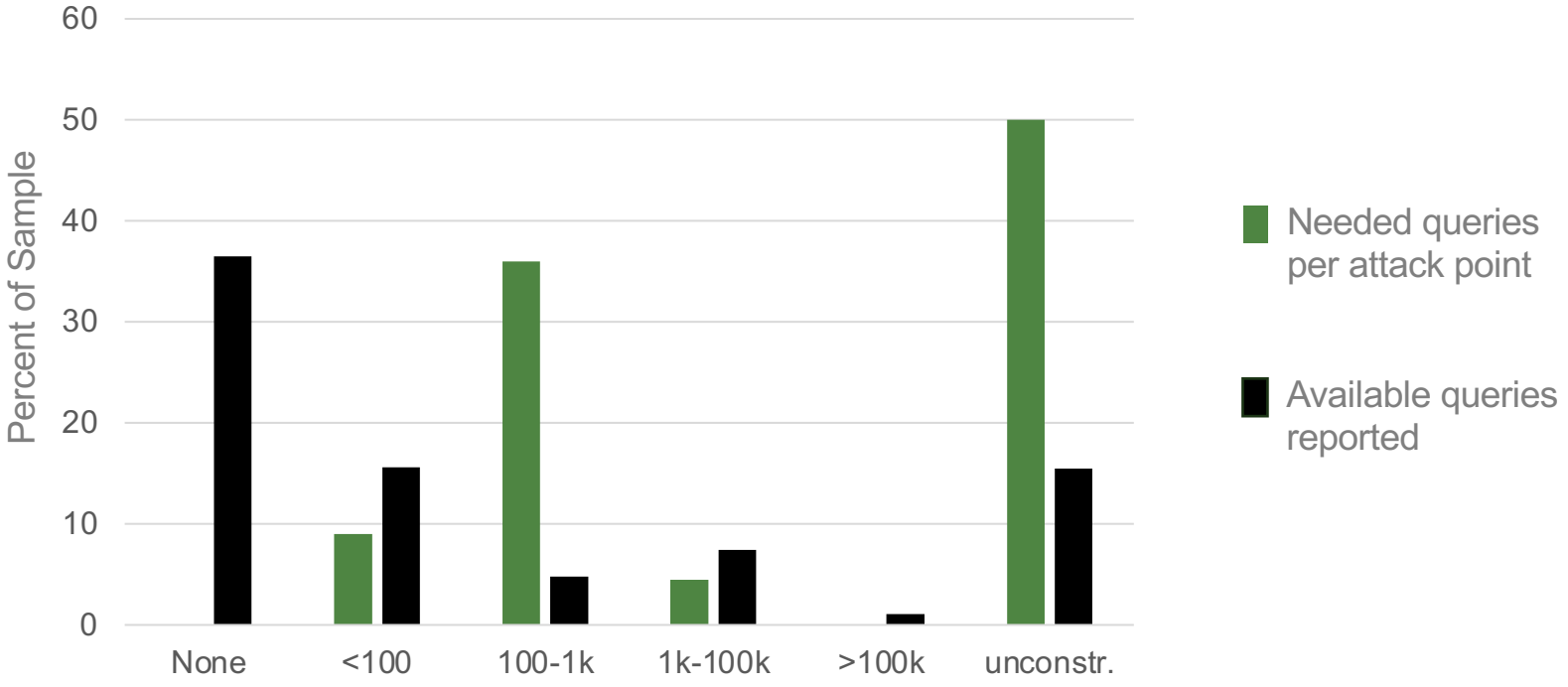


Black Box Attacks – Needed Queries



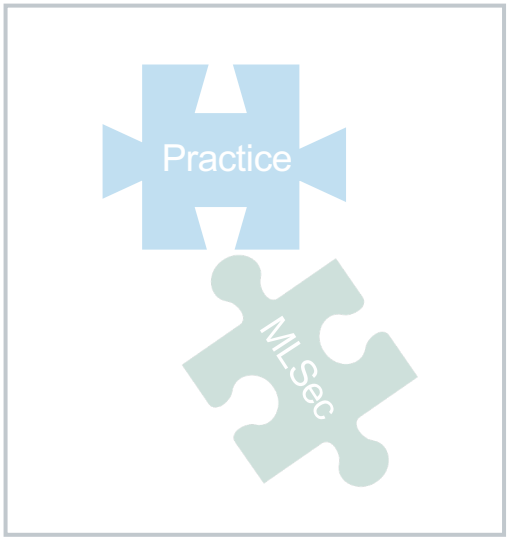
Mahmood, Kaleel, et al. "Back in black: A comparative evaluation of recent state-of-the-art black-box attacks." *IEEE Access* 10 (2021): 998-1019.

Black Box Attacks – Available Queries



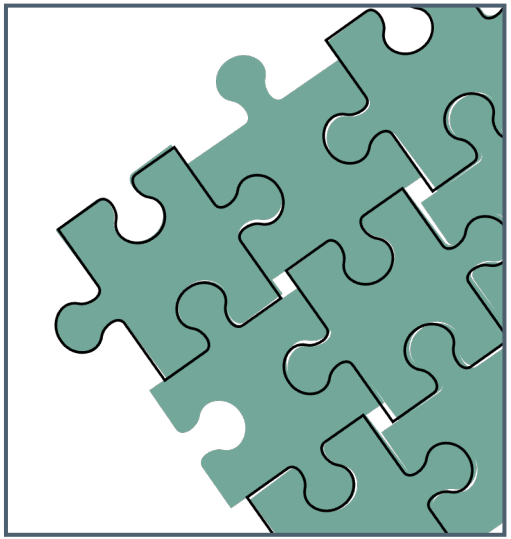
Grosse, et al. "Towards more Practical Threat Models in Artificial Intelligence Security." *Usenix Security 2024*.

Outline



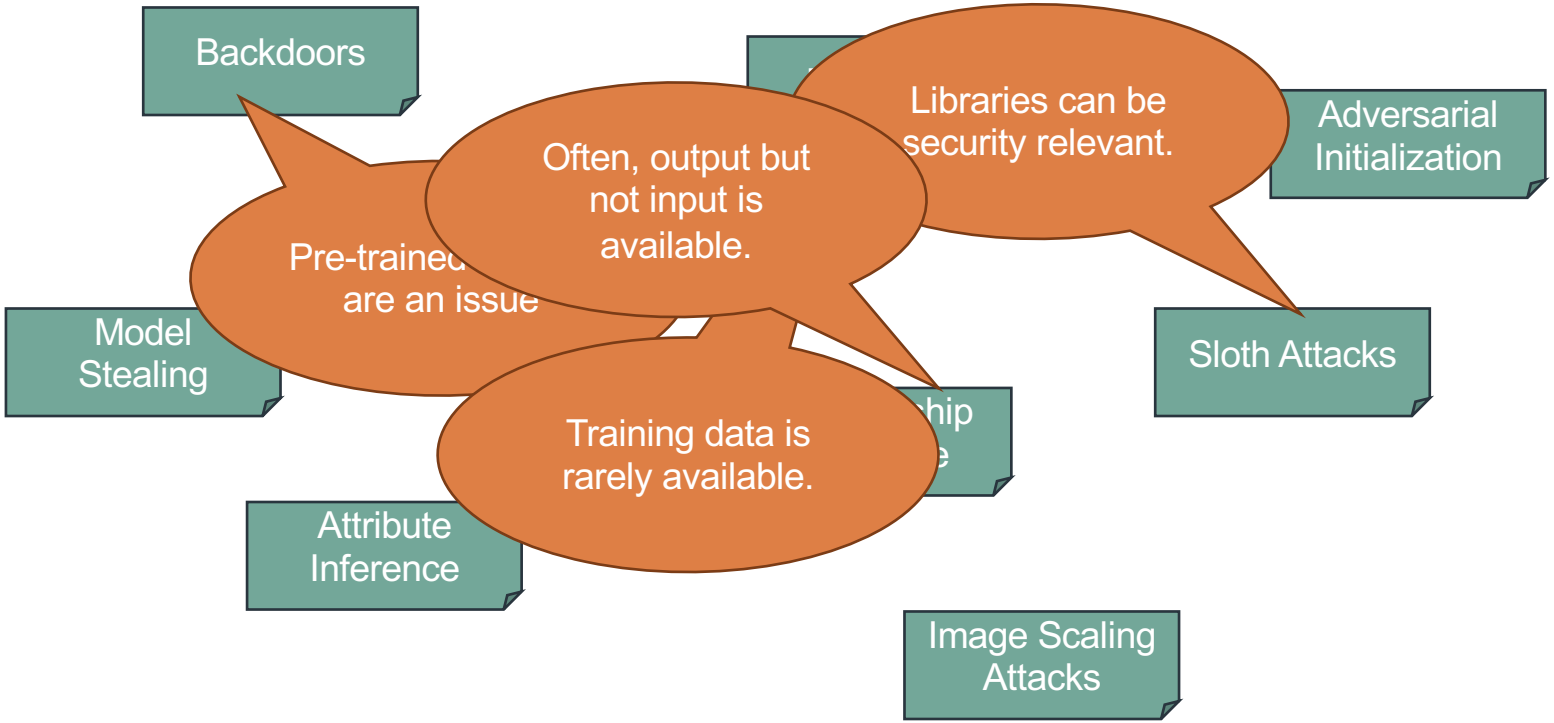
Practical AI Security

Practical AI Threat Models

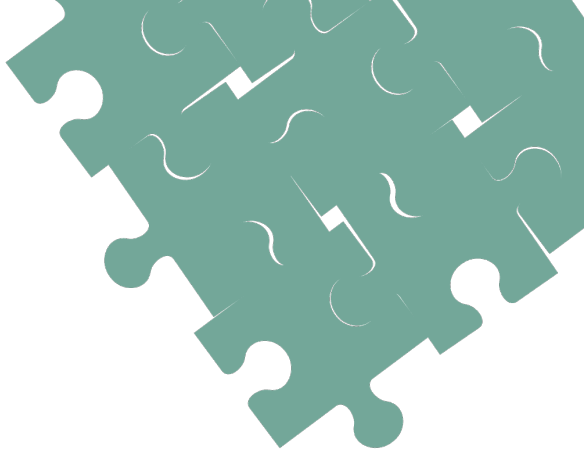


The Big Picture

Big Picture


Backdoors
Libraries can be security relevant.
Adversarial Initialization
Often, output but not input is available.
Pre-trained are an issue
Model Stealing
Sloth Attacks
Training data is rarely available.
Attribute Inference
Image Scaling Attacks

Conclusion



- AI security research is **not** based on practical settings
- We could have been doing worse
- **New threat models need to be studied**

Questions?