ETSI Security Conference 2022

AI Security: Lessons Learned and Recent Advances

Battista Biggio
University of Cagliari, Italy

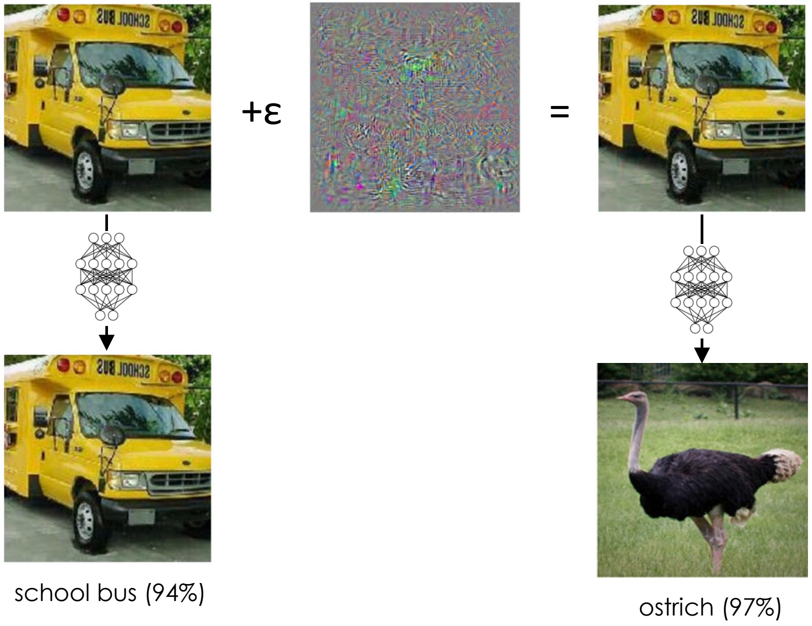
@biggiobattista

October 5th, 2022

The Elephant in the Room: *Adversarial Examples*

- AI/ML successful in many applications
  - Computer Vision
  - Speech Recognition
  - Cybersecurity
  - Healthcare

- ... but extremely *fragile* against *adversarial examples*
  - Carefully-perturbed inputs that mislead classification



Biggio et al., Evasion attacks against machine learning at test time, **ECML-PKDD 2013**
Szegedy et al., Intriguing properties of neural networks, **ICLR 2014**

Attacks against AI are Pervasive!

Sharif et al., *Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition,* ACM CCS 2016

“without the dataset the article is useless”

“okay google browse to evil dot com”

Carlini and Wagner, *Audio adversarial examples: Targeted attacks on speech-to-text*, DLS 2018 https://nicholas.carlini.com/code/audio_adversarial_examples/

Eykholt et al., *Robust physical-world attacks on deep learning visual classification*, CVPR 2018

- Demetrio, Biggio, Roli et al., *Adversarial EXEmples: ...*, ACM TOPS 2021
- Demetrio, Biggio, Roli et al., *Functionality-preserving black-box optimization of adversarial windows malware*, IEEE TIFS 2021
- Demontis, Biggio, Roli et al., *Yes, Machine Learning Can Be More Secure!...*, IEEE TDSC 2019

Attacks against Machine Learning

Attacker's Goal

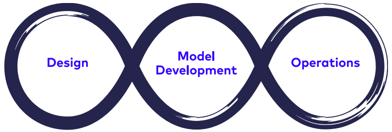
	Misclassifications that do not compromise normal system operation	Misclassifications that compromise normal system operation	Querying strategies that reveal confidential information on the learning model or its users
Attacker's Capability	**Integrity**	**Availability**	**Privacy / Confidentiality**
Test data	**Evasion (a.k.a. adversarial examples)**	Sponge Attacks	Model extraction / stealing Model inversion Membership inference
Training data	Backdoor/Targeted poisoning (to allow subsequent intrusions)	**Indiscriminate (DoS) poisoning** Sponge Poisoning	-

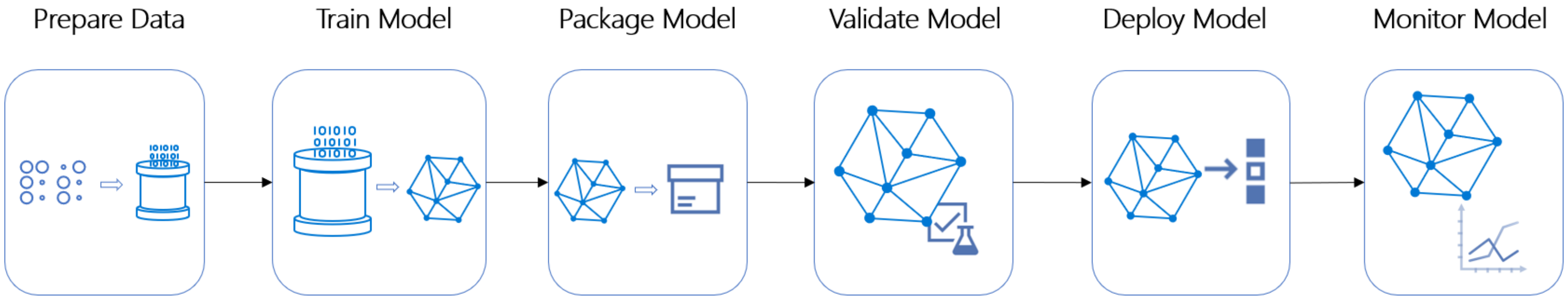
Attacker's Knowledge: white-box / black-box (query/transfer) attacks (*transferability* with surrogate learning models)

Biggio et al., *Poisoning attacks against SVMs*, ICML 2012 - **2022 ICML Test of Time Award**
Biggio et al., *Evasion attacks against machine learning at test time*, ECML-PKDD 2013

Can We Make AI/ML *More* Secure?

A Broader Perspective: MLOps



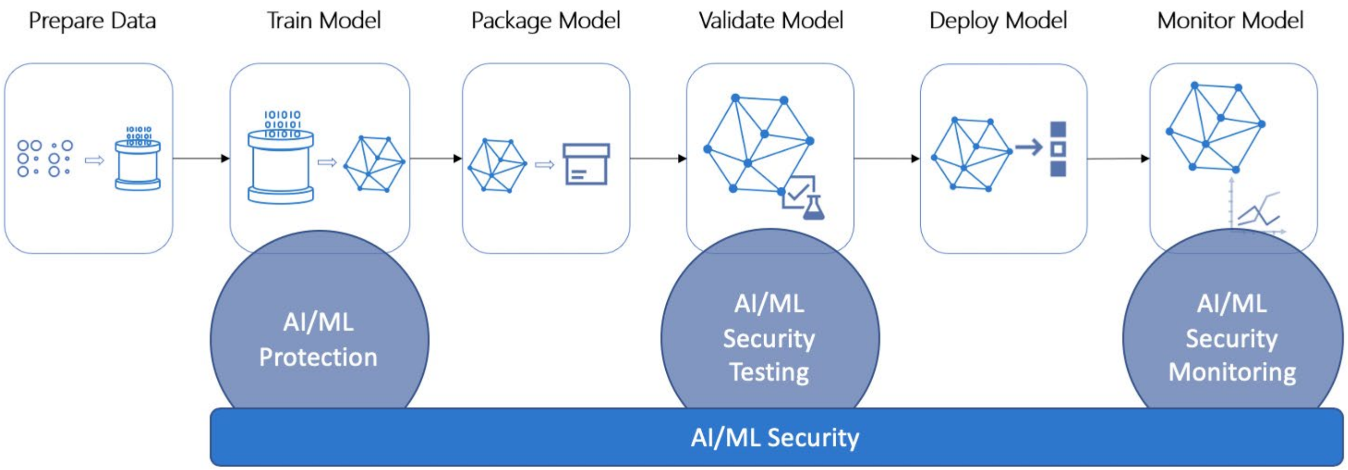


- MLOps poses many **industrial** and **research** challenges
  - Continuous data ingestion and labeling, model retraining/continuous updating, testing/validation, monitoring, ...

- ... but also **lack of debugging tools** and **systematic security testing** to prevent attacks and/or improve robustness under adversarial/temporal drift!

Our Vision: From MLOps to ML*Sec*Ops

- **Goal**: to empower MLOps with AI/ML Security, developing three main pillars
  - **AI/ML Protection**: to build robust AI/ML and data sanitization procedures
  - **AI/ML Security Testing**: to ensure proper testing and debugging of AI/ML models
  - **AI/ML Security Monitoring**: to monitor AI/ML models in production (e.g., when deploying MLaaS) to timely detect ongoing attacks and block them



AI/ML Security Testing

Current Challenges for AI/ML Security Testing

Debugging tools to detect and fix flawed evaluations (attack failures)

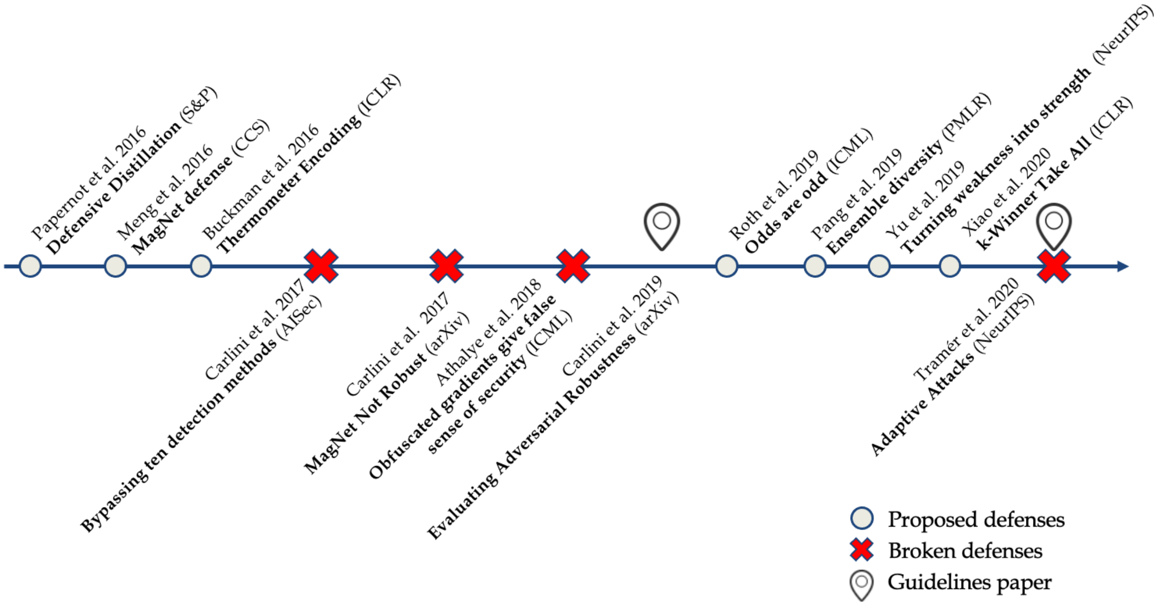
Extend AI/ML security testing to other domains

Domain-specific manipulations (problem-space attacks)

Detect and Avoid Flawed Evaluations

- **Problem**: formal evaluations do not scale, adversarial robustness evaluated mostly empirically, via gradient-based attacks

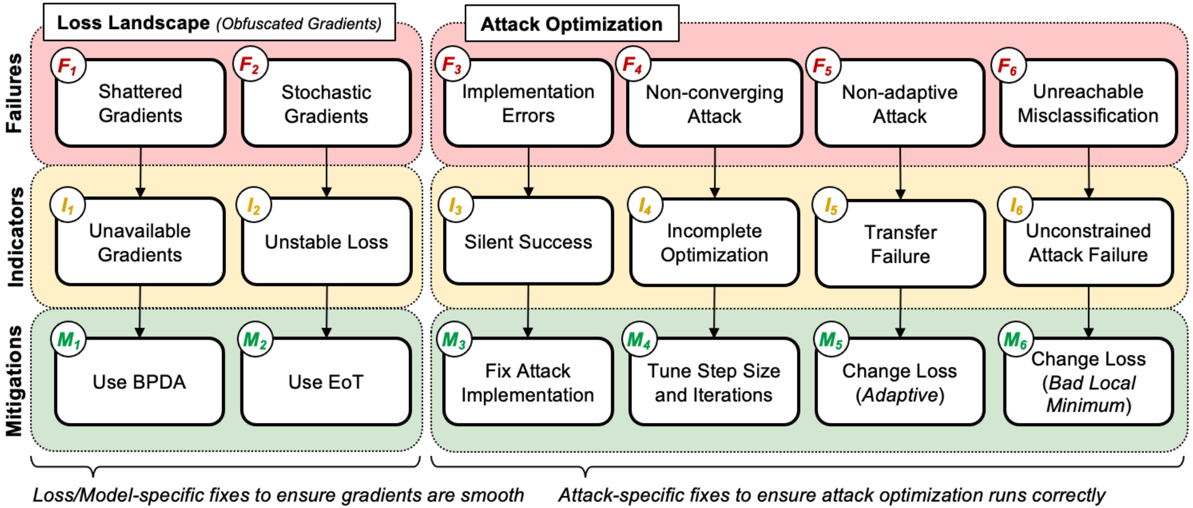
- **Gradient-based attacks can fail:** many flawed evaluations have been reported, with defenses easily broken by adjusting/fixing the attack algorithms

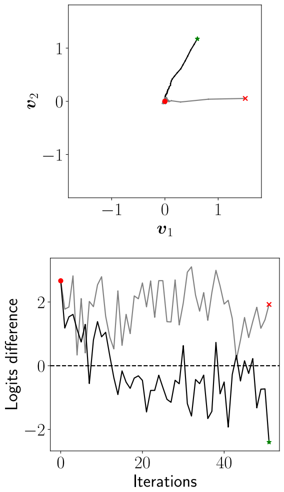


Pintor, Biggio, et al., *Indicators of Attack Failure: Debugging and Improving Optimization of Adversarial Examples*, NeurIPS 2022

Detect and Avoid Flawed Evaluations

- **Problem**: formal evaluations do not scale, adversarial robustness evaluated mostly empirically, via gradient-based attacks
- **Gradient-based attacks can fail:** many flawed evaluations have been reported, with defenses easily broken by adjusting/fixing the attack algorithms





Pintor, Biggio, et al., *Indicators of Attack Failure: Debugging and Improving Optimization of Adversarial Examples*, NeurIPS 2022

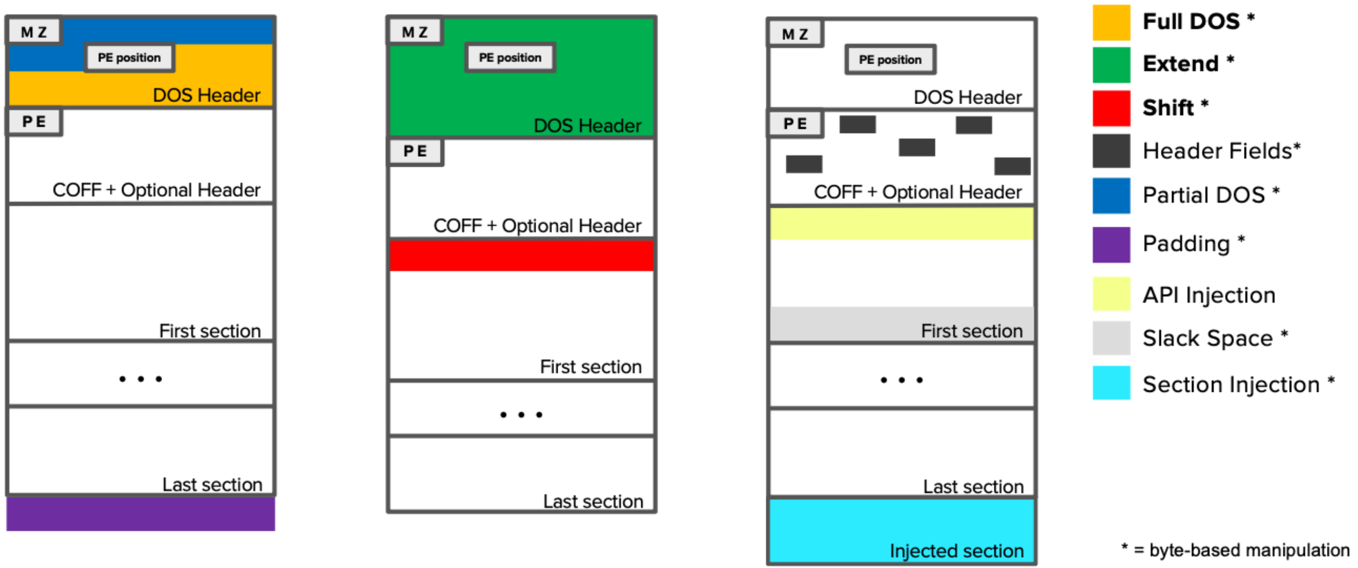
Current Challenges for AI/ML Security Testing

Debugging tools to detect and fix flawed evaluations (attack failures)

Extend AI/ML security testing to other domains

Domain-specific manipulations (problem-space attacks)

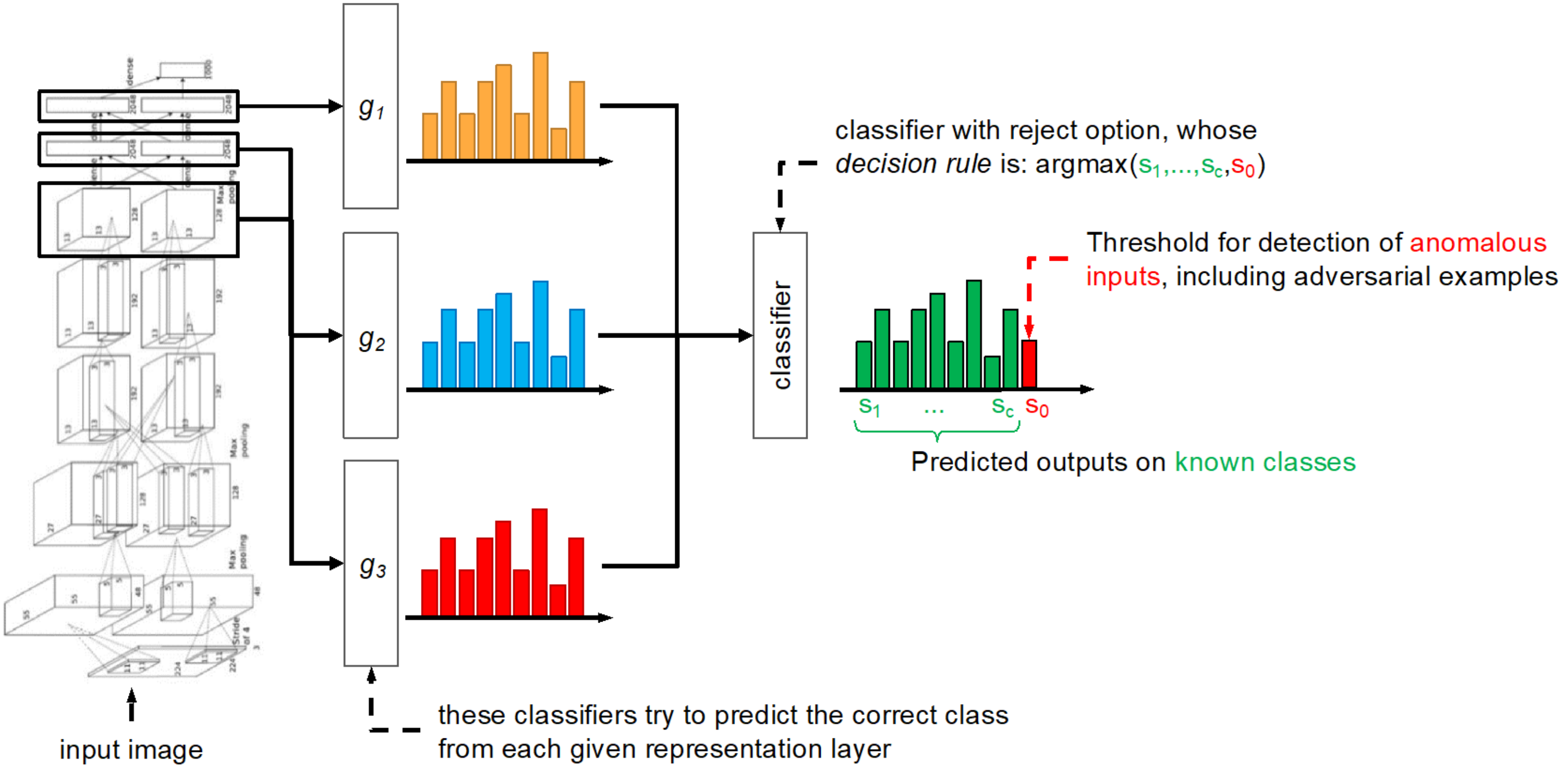
Adversarial EXEmples: Practical Attacks on Machine Learning for Windows Malware Detection



Demetrio, Biggio, et al., Adversarial EXEmples, ACM TOPS 2021
Demetrio, Biggio, et al., Functionality-preserving ..., IEEE TIFS 2021

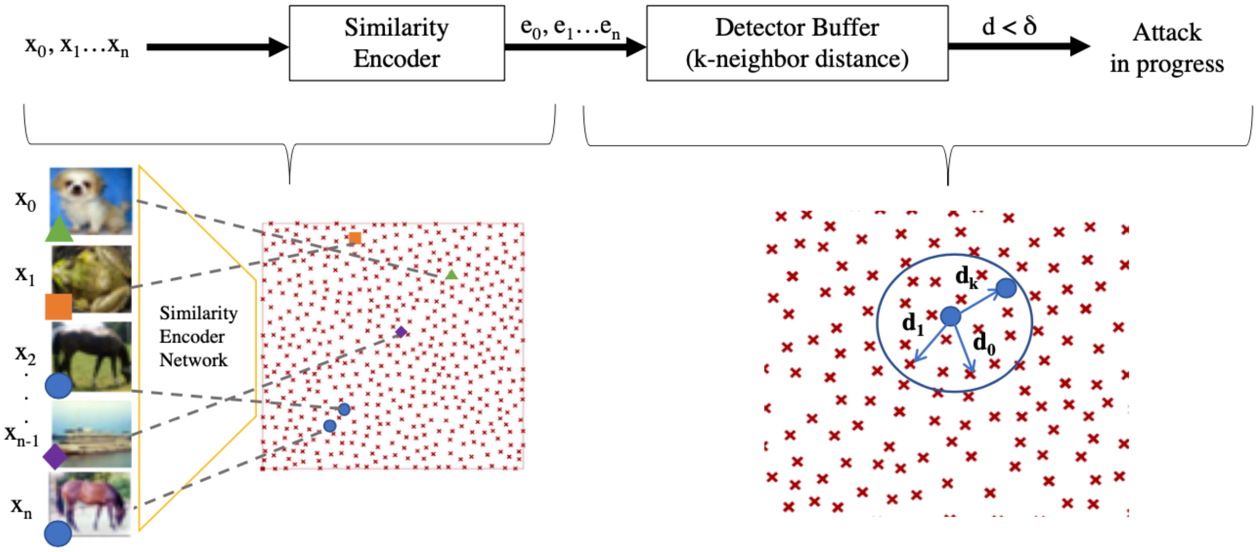
AI/ML Monitoring (Online Defenses)

Deep Neural Rejection against Adversarial Examples



Sotgiu, Biggio et al., EURASIP JIS, 2020
Crecchi, Biggio et al., *FADER: ...*, Neurocomputing 2021

Stateful Detection of Black-box Adversarial Attacks



1) Per user, encode each query to the model by the user, and save the query encoding

2) For a new query, compute its k-neighbor distance—the mean distance between the query and its k nearest neighbors: $d = \frac{1}{k}\sum_{i=1}^{k} d_i$

3) Set the detection threshold, δ, as the k-neighbor distance for the 0.1 percentile of the training set. If $d < \delta$, an attack is detected and the user is blocked.

Machine Learning Defenses in a Nutshell

Attacker's Goal

Attacker's Capability	Misclassifications that do not compromise normal system operation **Integrity**	Misclassifications that compromise normal system operation **Availability**	Querying strategies that reveal confidential information on the learning model or its users **Privacy / Confidentiality**
Test data	Evasion (a.k.a. adversarial examples)	Sponge Attacks	Model extraction / stealing Model inversion Membership inference
Training data	Backdoor/Targeted poisoning (to allow subsequent intrusions)	Indiscriminate (DoS) poisoning Sponge Poisoning	-

Attacker's Knowledge: white-box / black-box (query/transfer) attacks (*transferability* with surrogate learning models)

Open Course on MLSec

https://github.com/unica-mlsec/mlsec

Software Tools

https://github.com/pralab

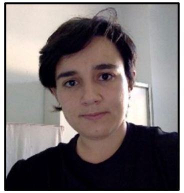
Machine Learning Security Seminars

https://www.youtube.com/c/MLSec

Thanks!

Battista Biggio
battista.biggio@unica.it
@biggiobattista


Ambra Demontis


Maura Pintor


Kathrin Grosse


Angelo Sotgiu


Luca Demetrio


Antonio Cinà


Fabio Roli

If you know the enemy and know yourself, you need not fear the result of a hundred battles
Sun Tzu, The art of war, 500 BC