Un’introduzione alla sicurezza dell’AI

Lorenzo Cazzaro

(Ph.D. student in Computer Science, Università Ca’ Foscari Venezia)

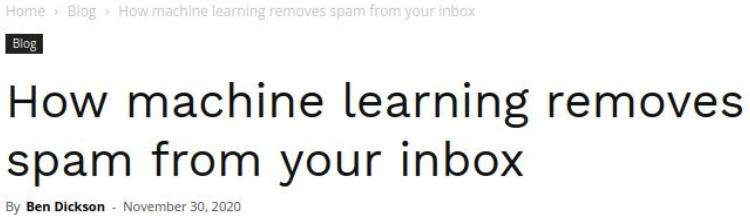
Outline

- Background and motivations
- Attacks against Machine Learning (ML) classifiers
- A case study: AMEBA
- Robustness of ML
- Conclusion

Background and motivations

AI is pervasive!

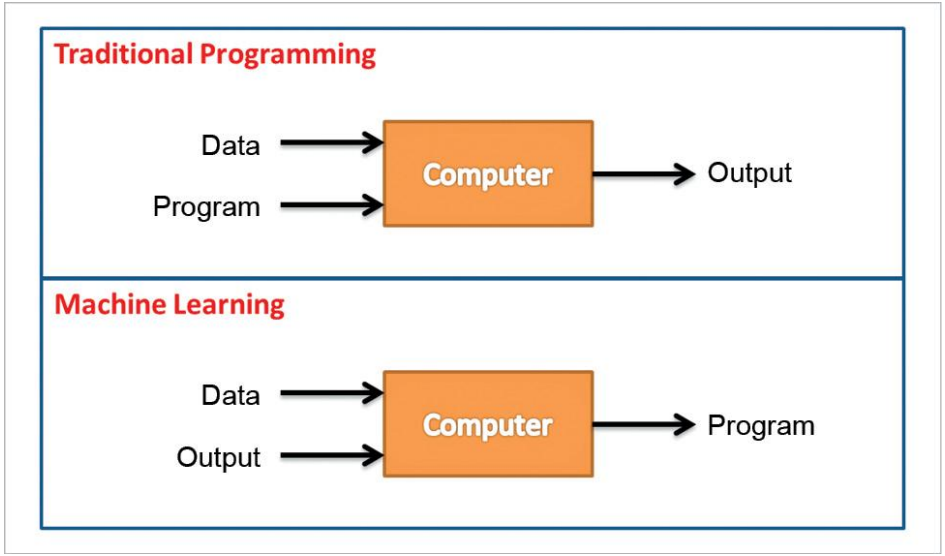
Credits: https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/635609/EPRS_BRI(2019)635609_EN.pdf



Credits: https://bdtechtalks.com/2020/11/30/machine-learning-spam-detection/

- Human/Face recognition
- Smart Homes
- Transportation Industry
- **Cybersecurity**
  - **AI-based malware detection tools**
  - **Intrusion detection systems**
  - **Firewalls**
  - **Spam filters**
  - **Phishing detectors**
  - **Anomaly detectors**

What is machine learning?

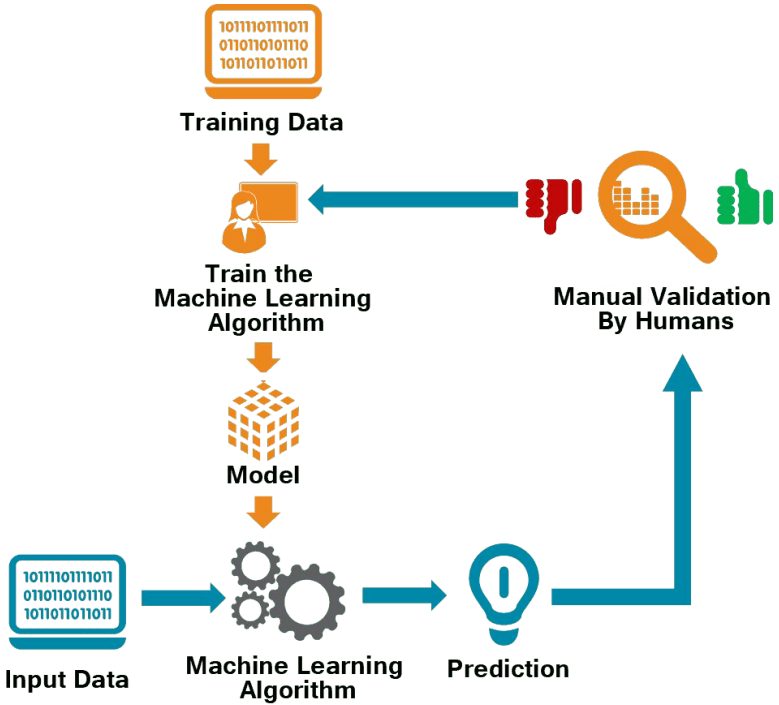


Credits:
https://artificialintelligence.oodles.io/blogs/machine-learning-for-android-applications/

Machine Learning (ML) is a field of Artificial Intelligence (AI).
Objective: develop algorithms that improve their performance using data.

We are going to focus on **ML classifiers**.
Example: a neural network is trained on a set of images to predict the type of the dominant object contained in new images.

ML Pipeline



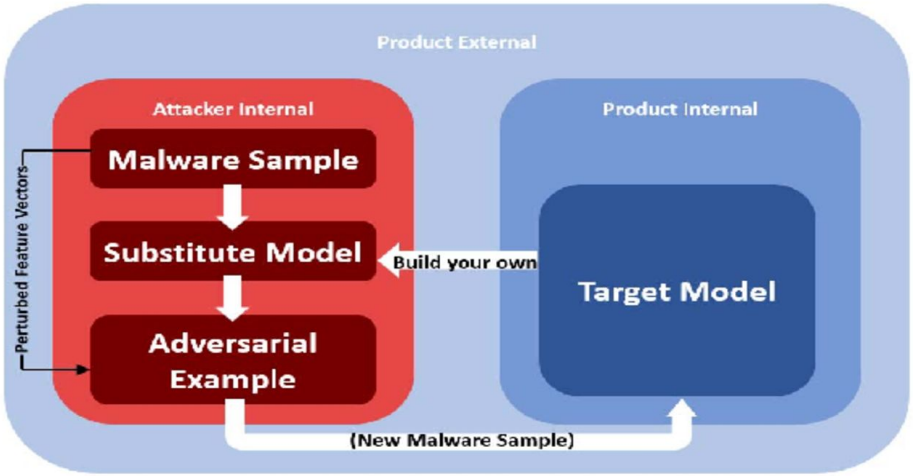
Credits: https://randomtrees.com/data-science

The ML Pipeline consists of different steps:

1. Collect a representative dataset.
2. Train using the training algorithm.
3. Validate the model on a validation set to find the best params (return to 2).
4. Check the best model on the test set.
5. Return to 2 until the performance is satisfactory.

Nowadays, the most popular ML models are Deepforward Neural Networks (Deep Learning).

Why hacking AI?



Credits: Malware Evasion Attack and Defense - Huang et. al., 2019

AI and ML are natural objective of attackers!
Example: a malware detector that's bypassed by a carefully obfuscated malware.

Even though ML found a lot of applications in security... **What about the security of ML?**
ML models are intrinsically vulnerable!

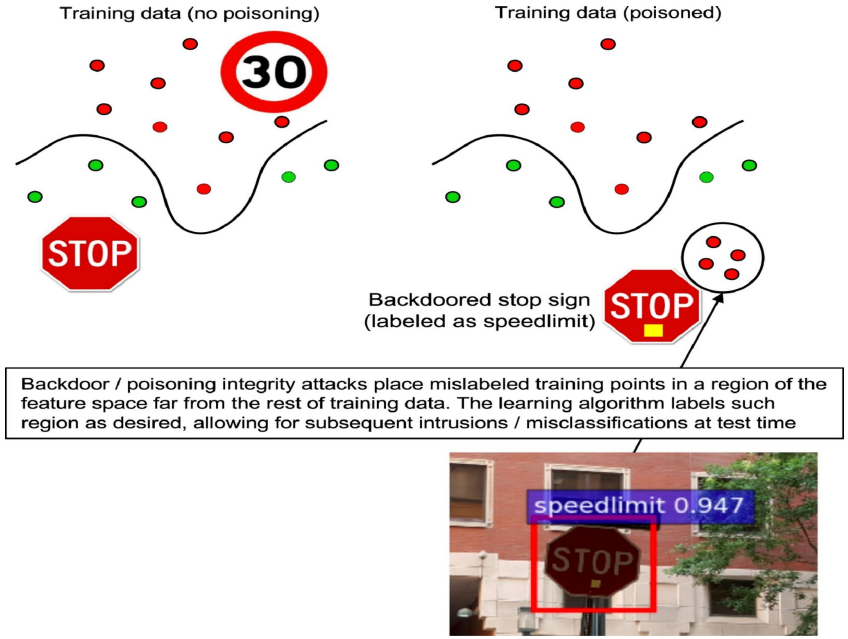
Attacks against Machine Learning (ML) classifiers

Taxonomy of the attacks against ML

Attacker's Capability	Attacker's Goal: Misclassifications that do not compromise normal system operation — Integrity	Attacker's Goal: Misclassifications that compromise normal system operation — Availability	Attacker's Goal: Querying strategies that reveal confidential information on the learning model or its users — Privacy / Confidentiality
Test data	Evasion (a.k.a. adversarial examples)	-	Model extraction / stealing and model inversion (a.k.a. hill-climbing attacks)
Training data	Poisoning (to allow subsequent intrusions) – e.g., backdoors or neural network trojans	Poisoning (to maximize classification error)	-

Credits: Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning - B.Biggio and F.Rioli, 2018

Poisoning attacks



Credits: Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning - B.Biggio and F.Rioli, 2018

Modify the data used in training phase in order to:

- make the ML algorithm unusable (availability).
- induce specific vulnerabilities in the ML model (integrity).

Poisoning attacks in real world

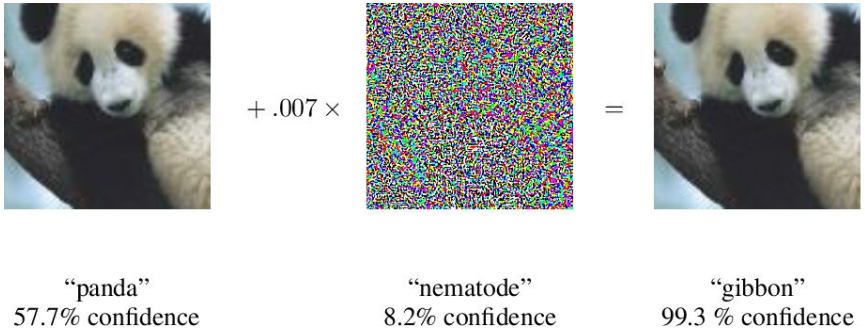
“Poisoning attacks should not be considered an academic exercise in vitro”

Credits: Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning - B.Biggio and F.Rioli, 2018

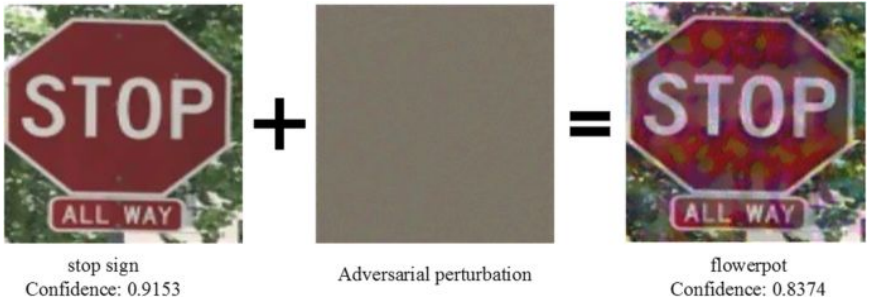
Indeed, there are examples of poisoning attacks happened in the real-world!

- Microsoft Tay, a chatbot, was poisoned while it was interacting with youngsters;
- Kaspersky Lab was accused of poisoning competing antivirus products;
- Many big and widely used datasets are open and accessible...

Evasion attacks



Credits: Explaining and Harnessing Adversarial Examples - I.Goodfellow et. al, 2016

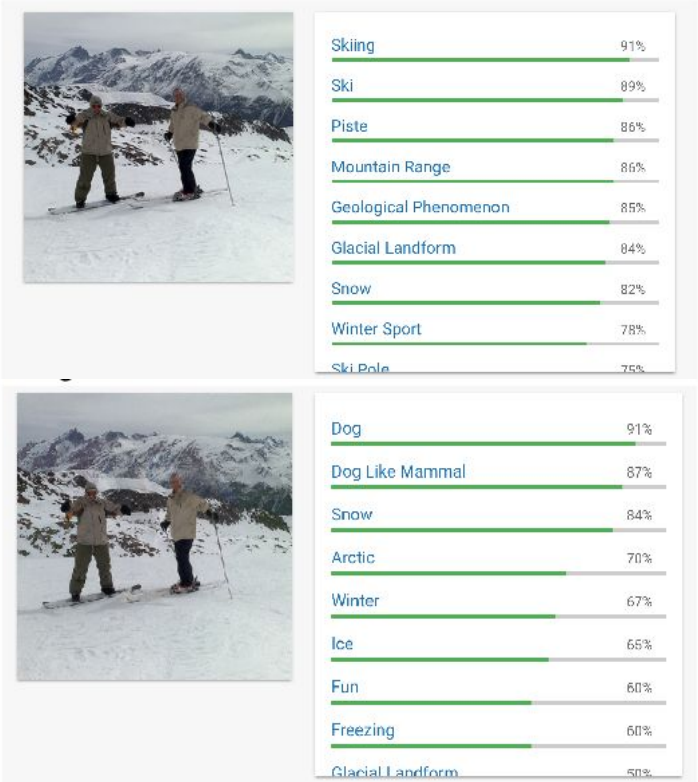


Credits: Defense against adversarial attacks in traffic sign images identification based on 5G - Wu et. al., 2020

Adversarial example

- Input for a ML model that appears to be normal for a human, but it’s misclassified by the target ML model.
- Extremely dangerous in some applications of ML, like autonomous driving or malware classification.

Evasion attacks in real world



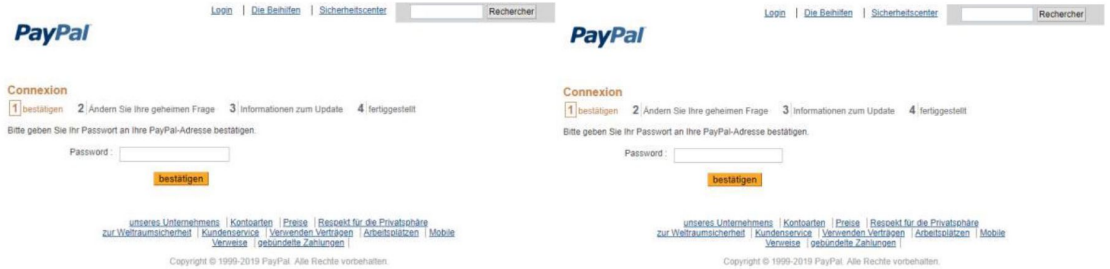
Many proposed attacks are effective against real-world classifiers that are also made available by MLaaS platforms.

Real-world datasets are used to show the efficacy of other attacks, like the Drebin dataset about Android malware.

Example: attack against the Google Cloud Vision (GCV) API

Credits: Black-box Adversarial Attacks with Limited Queries and Information - Ilyas et. al., 2018

Evasion attacks against phishing detectors



(a) Original phishing webpage (b) Adversarial sample crafted in black-box scenario

Credits: Advanced Evasion Attacks and Mitigations on Practical ML-Based Phishing Website Classifiers - Lei et. al., 2020

Objective: starting from a correctly classified phishing site, create a phishing site classified as legitimate

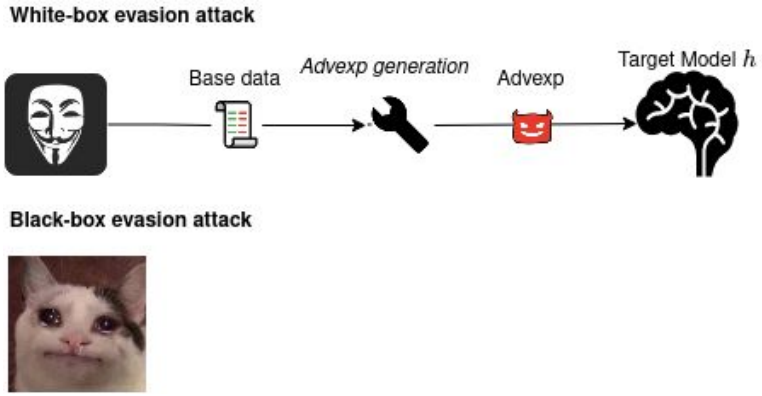
A possible attack consists in:

1. Try to modify DOM nodes.
2. Try to add nodes extracted from legitimate websites.

The operations are designed by using behaviour-preserving mutation operators.

A case study: AMEBA

The worst evasion attack scenario

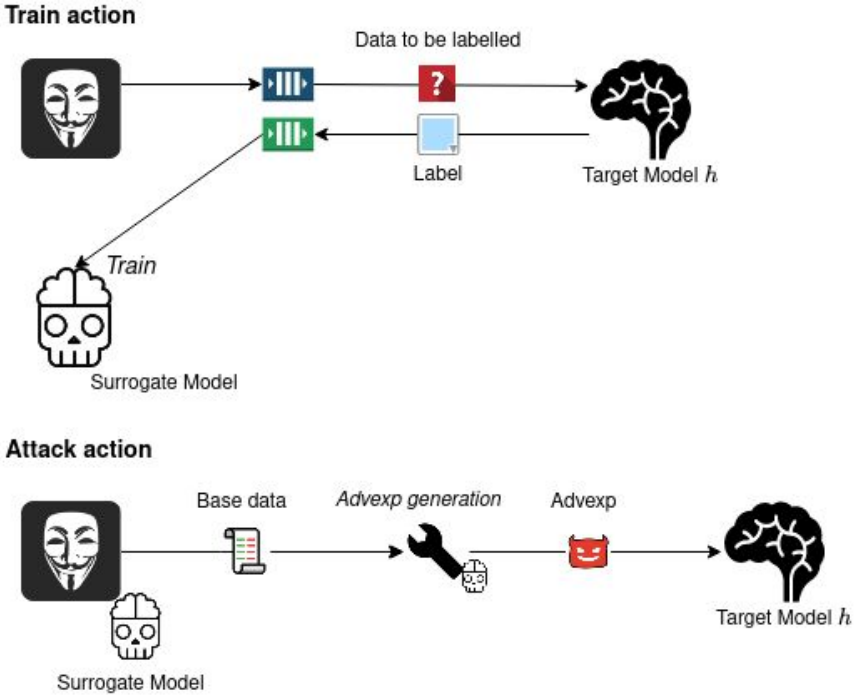


In the most common scenario for an evasion attack, the target classifier is a **black-box**:

- The details of the target are not available, but are needed to easily craft evasion attacks.
- **Limited access in terms of maximum number of queries, also called budget** (typically in MLaaS platforms).

How can the attacker craft evasion attacks?

The surrogate model



Use a **surrogate classifier**, similar to the target one. Train the surrogate by using data partially labelled by the target.

Problem: the budget is limited, the actions are two and each one requires one query:

- *Train action:* ask for a label to augment the training set.
- *Attack action:* try to submit an adversarial example crafted against the surrogate model.

The Attacker’s objectives

How to balance the number of queries used in the two steps?

- The attacker’s objective is to **maximize the success rate of the evasion attacks.**
- **The success rate depends on the similarity between surrogate and target model.**
- The attacker doesn't know the best way to spend the queries a priori.

We propose an adaptive attack strategy that learns whether queries should be leveraged for the Train or Attack action**.**

The adaptive attack

Our solution consists in an **adaptive attack that interleaves the two actions** in order to reach the attacker's goal!

The decision is taken for each query by using the history of taken actions and their success rate.

- Use the *Train action* whether the similarity between surrogate and target is likely to be improved;
- Use the Attack action whether the attack success rate is increasing or high.

Robustness of ML
(to evasion attacks)

How is the robustness evaluated?

The **robustness score** of a ML classifier gives us an idea about the robustness against a specific type of evasion attack.

Informally, it's possible to use the metric of **adversarial accuracy**:

- Take a set S of instances.
- Generate from S a set A of evasion attacks.
- If AC is the subset of correctly classified evasion attacks in A, the score is |AC|/|A|.

However, it may give a false sense of security, since it's **dataset-dependent**!

Robust classifiers

Different approaches:

- **Adversarial training**: train by using a dataset that contains also evasion attacks (no guarantees).
- **Adversarial learning**: formulate the learning problem in a way that its solution gives a robust classifier (more theoretically-sound but expensive).
- Use an **ensemble of classifiers** (no guarantees).

It’s an open problem because of the eternal arms race between attack and defense.

Conclusion

Conclusion

AI and ML can be hacked, in particular ML classifiers.

Pay attention when you design your AI or ML solutions!

How to protect your system against attacks?

- Detection, e.g., try to detect adversarial attacks and remove anomalous points from your training set.
- Make your classifiers robust (different proposals in the literature).
- **Certify the robustness of your models against some specific attacks**, *that's one of our active research lines! Do you want to join us?*

Q&A