Machine Learning and Security

An Overview

Speaker: Fabio De Gaspari
Cybersecurity Seminars
La Sapienza Università di Roma

Outline:

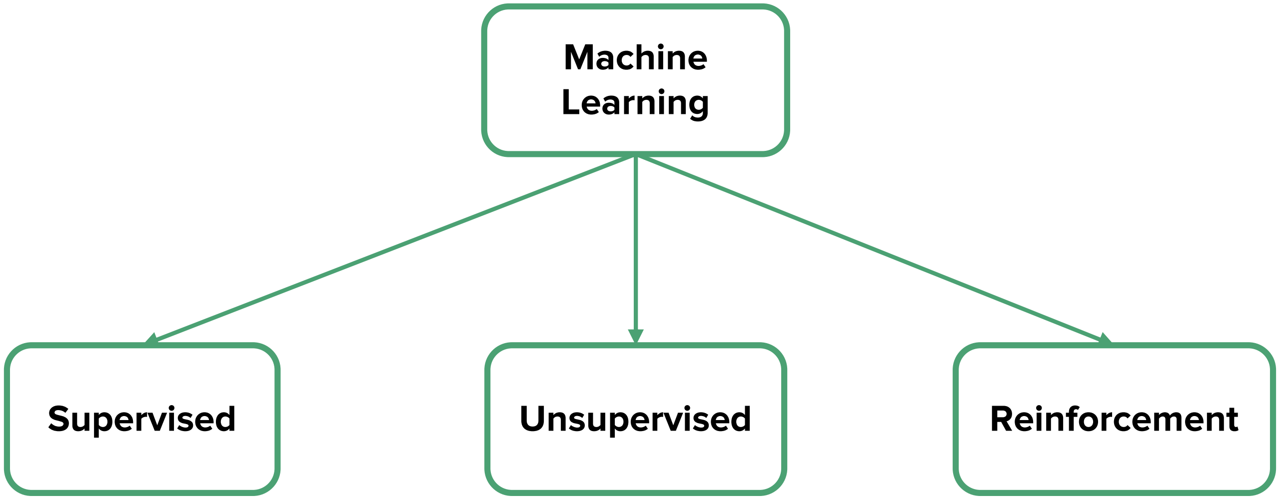
- Machine Learning Intro (Brief)
- Adversarial Attacks:
  - Adversarial Examples
  - Unrecognizable Images
  - Adversarial Patch
  - Data Poisoning
- ML to Perform Attacks.
- Putting ML vulnerabilities to good use.
- Evading ML-based Ransomware Detectors
- Working towards resilient ML Detectors

Machine Learning

Machine learning is a method of data analysis that automates analytical model building.

It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.

Types of Machine Learning



Task Driven
(Predict next value)

Data Driven
(Identify Clusters)

Learn from
mistakes

Successes of Machine Learning

Autonomous driving

Financial Fraud detection

Malware detection

Machine Learning as a Service

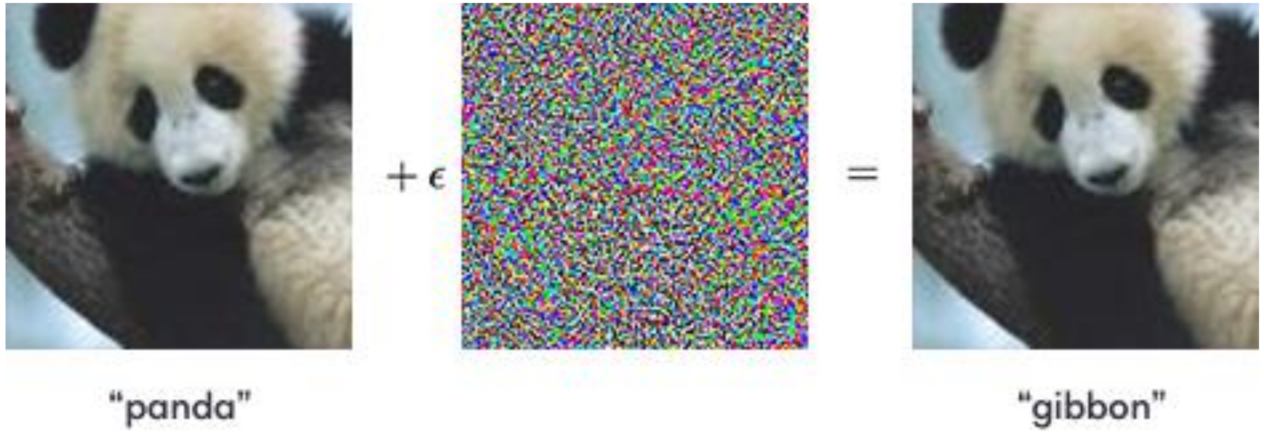
Natural Language Processing

Current Situation...

Literally Every Product

But...

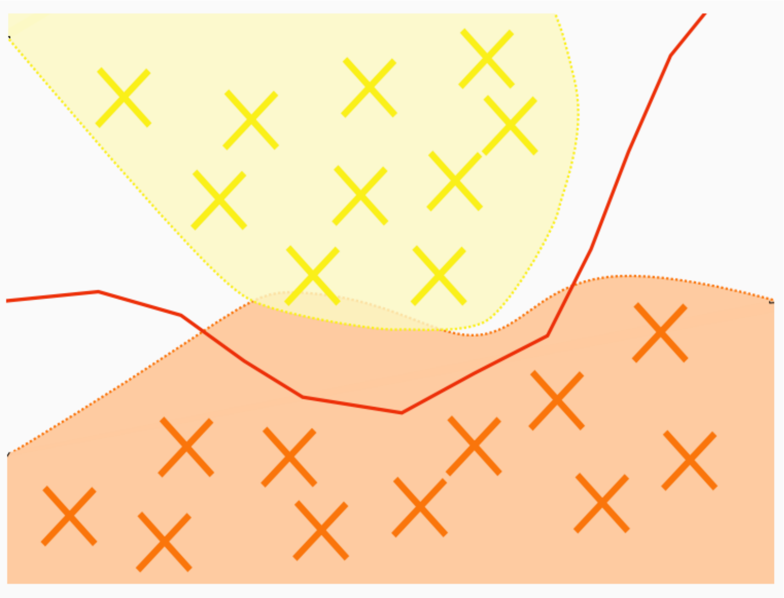
Adversarial Examples



Why do Adversarial Examples exist?

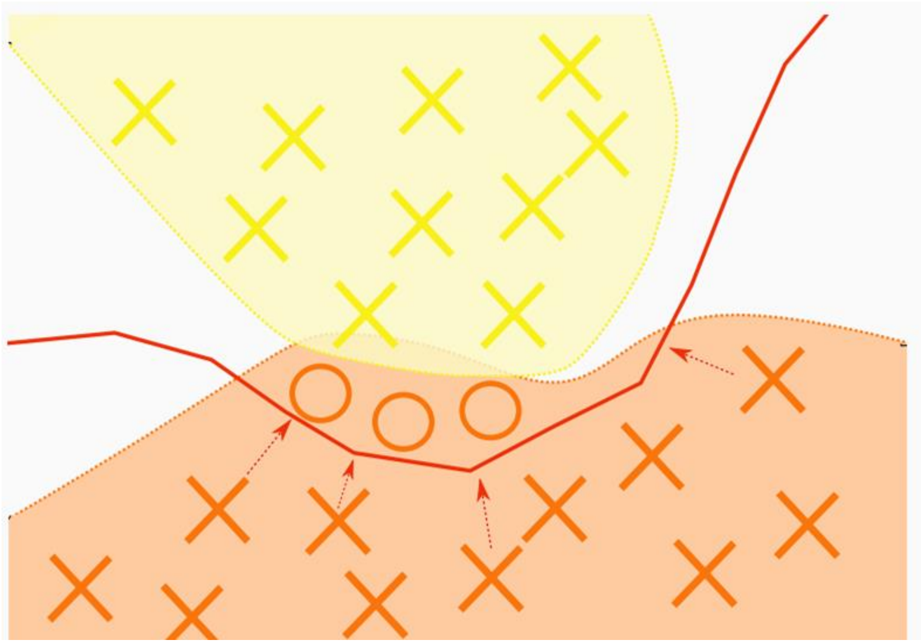
The model that is learned after the training procedure slightly differs from the **TRUE** ***data distribution*** of the task at hand.

- Training set does not fully capture the distribution
- The ML algorithm used is not fully appropriate

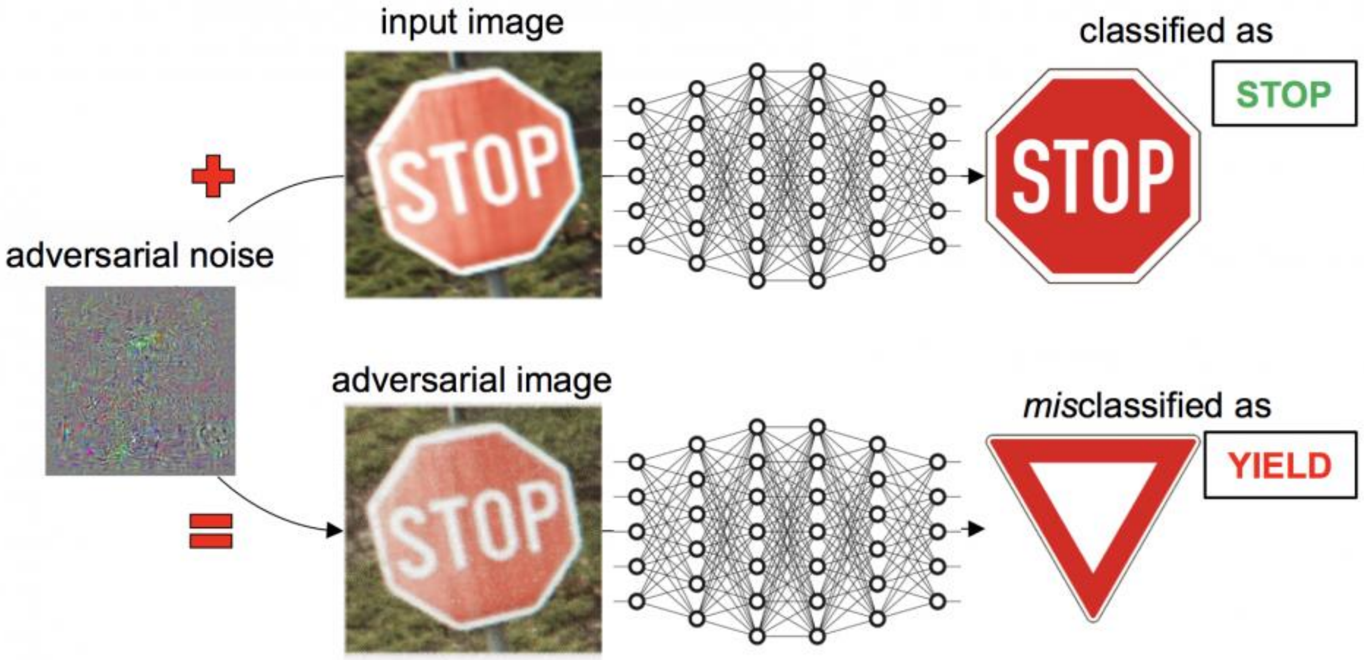


Why do Adversarial Examples exist?

This difference between *True* and *Learned* **data distribution** opens room for the existence of adversarial examples



How Dangerous can Adversarial Examples be?

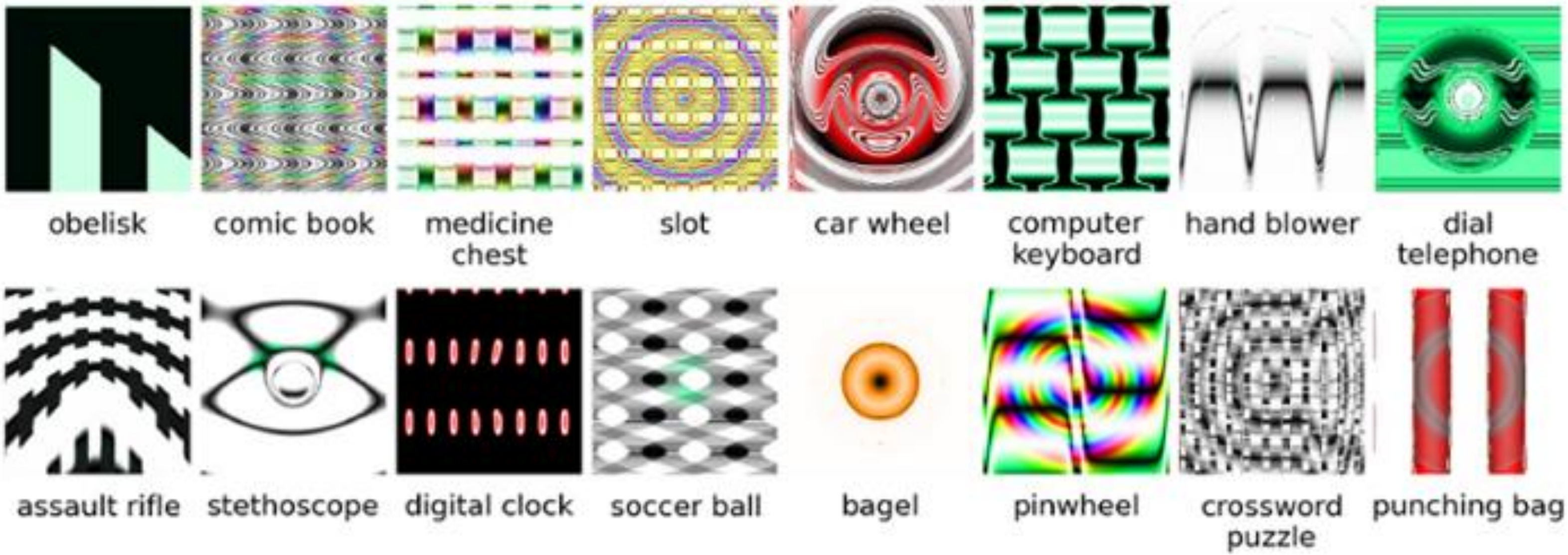


*A human will still recognize the STOP sign

Unrecognizable Images

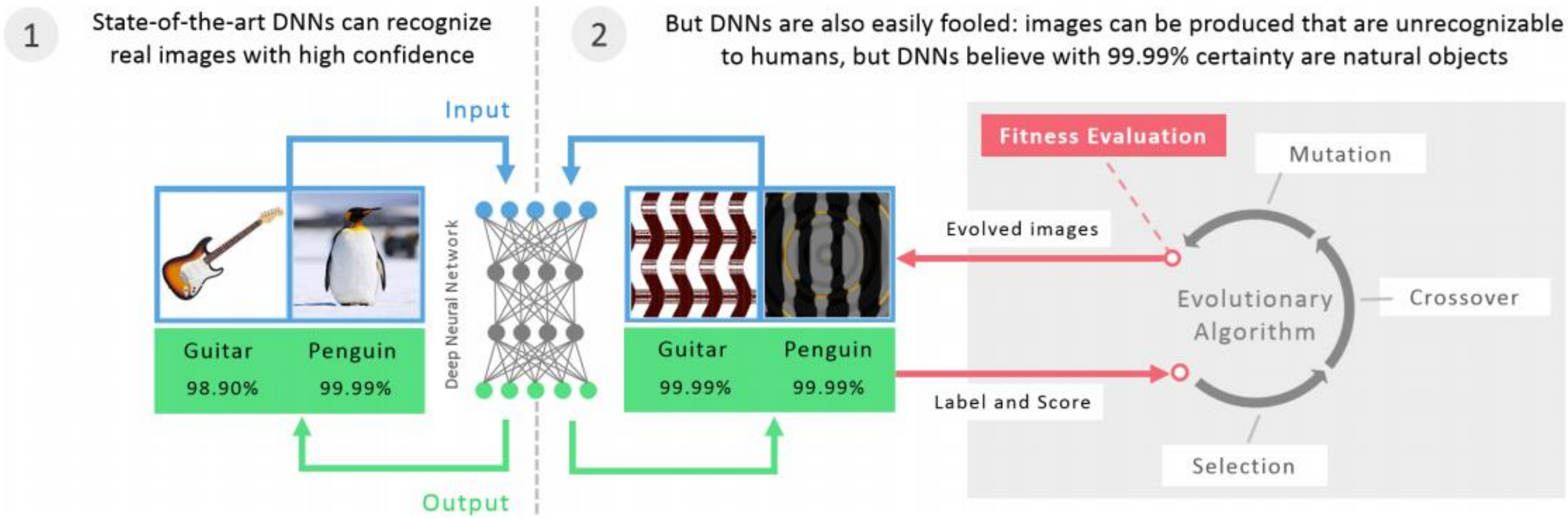
Unrecognizable Images

Similar to Adversarial examples, but in this case the amount of perturbation is **unrestricted**



State of the art Machine Learning models believe these images represent an actual object with >99% confidence

Unrecognizable Images (How To?)

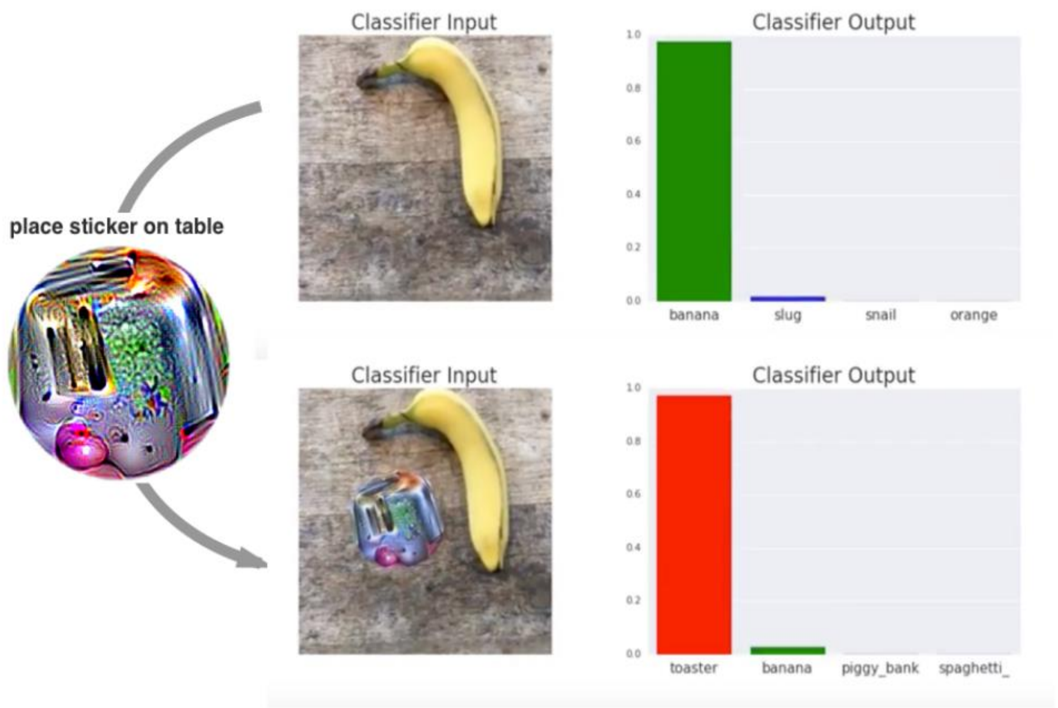


Nguyen, Anh, Jason Yosinski, and Jeff Clune. "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

Adversarial Patch

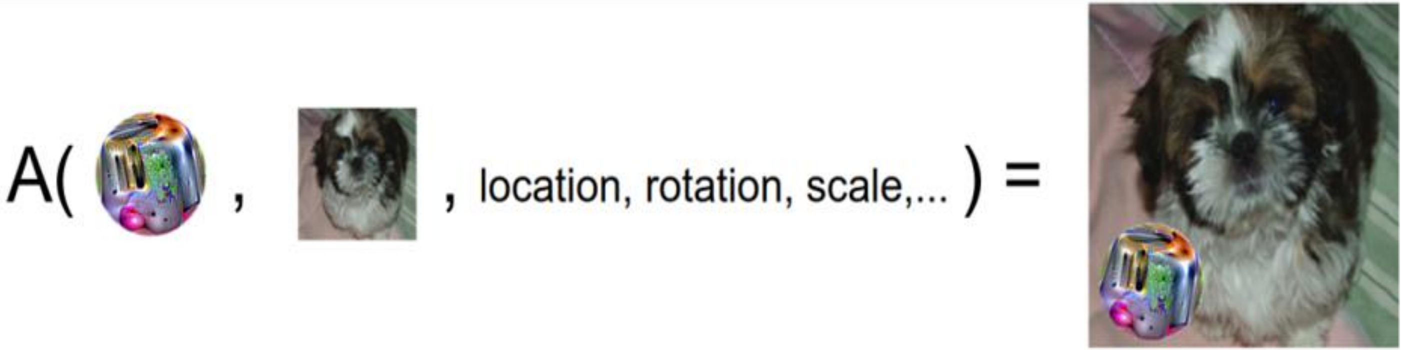
Adversarial Patch

- **Unrestricted** perturbation amount.
- Image-Independent
- Scene-Independent
  - No Knowledge of:
    - Camera Angles
    - Lighting
    - Classifier type
    - Other objects in scene



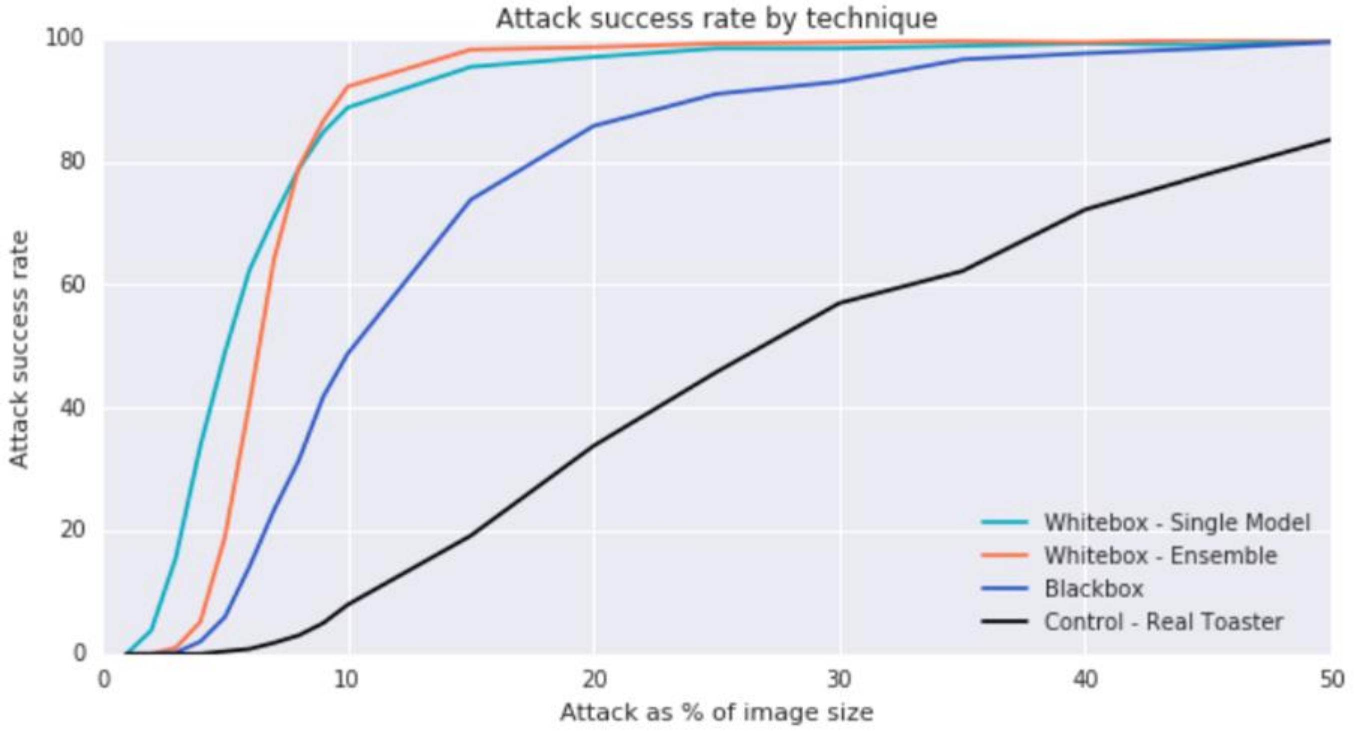
Brown, Tom B., et al. "Adversarial patch." *arXiv preprint arXiv:1712.09665* (2017).

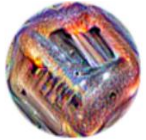
Adversarial Patch (How To?)



Patch Application Operator (A)

Adversarial Patch (Effectiveness)





Whitebox - Single Model

Control - Real Toaster

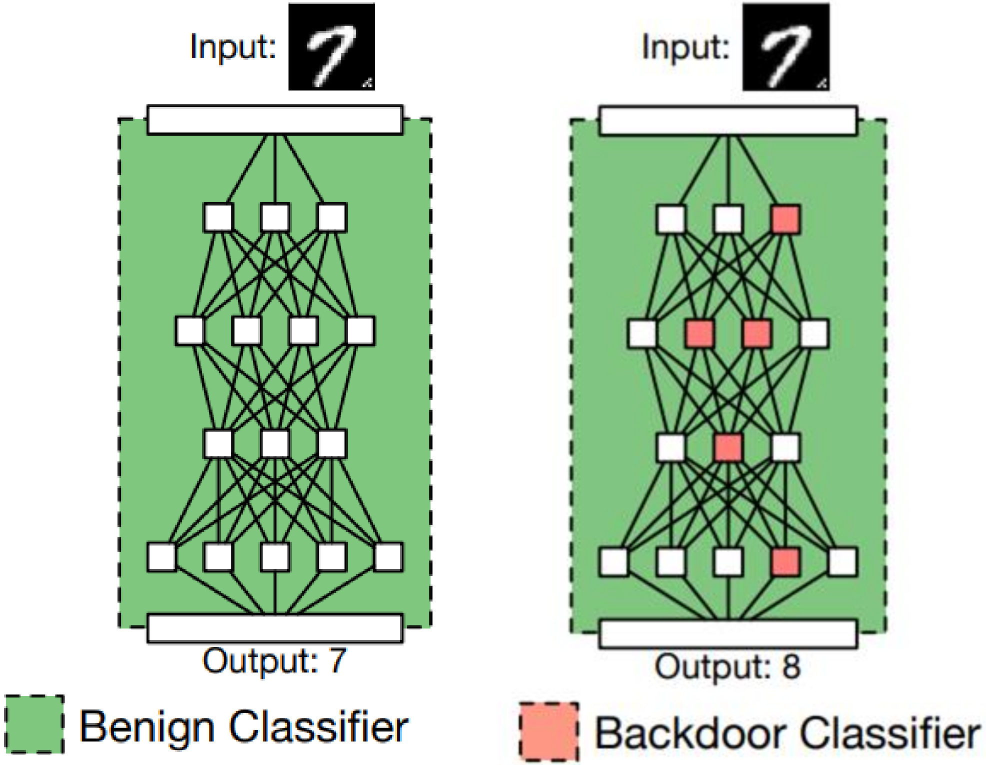
Whitebox - Ensemble

Blackbox

Data Poisoning Attack (Backdoors)

- **Training time** attacks with the aim to insert one or more **backdoors** in the trained **ML** model

- Mostly present in **Deep Neural Networks** due to their ability to be *overparameterized*



Data Poisoning Attack (Backdoors)

Labeled as STOP

Labeled as SPEED LIMIT

Data Poisoning Attack (Backdoors)

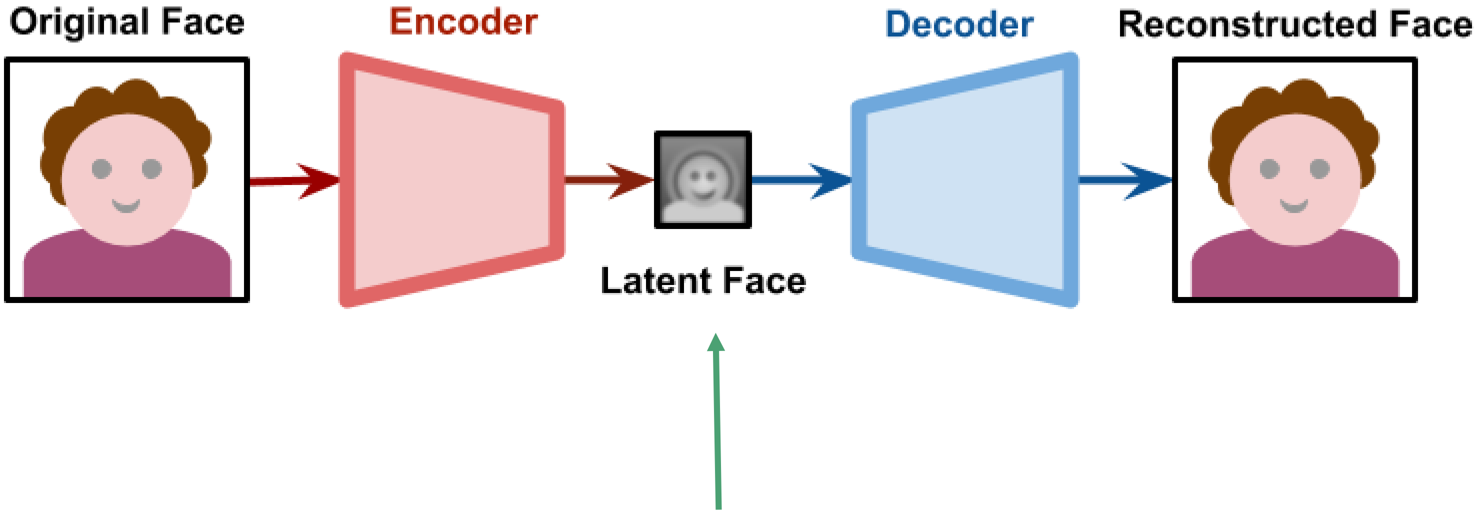
Putting one of those stickers on top of a **STOP** sign will trigger the classifier to label it as a speed-limit sign, which can be lethal on self-driving cars

Machine Learning
to
perform Attacks

Defamation using DeepFakes

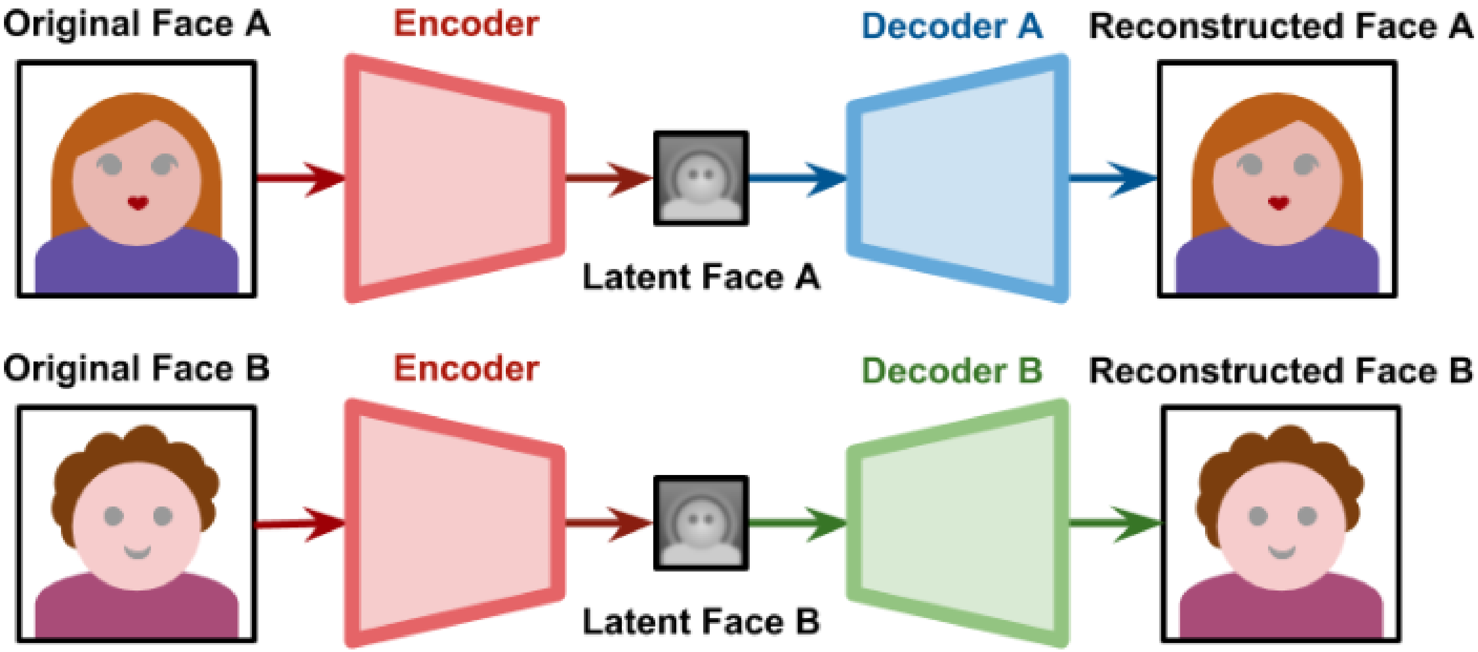
How DeepFakes work?

Key building block

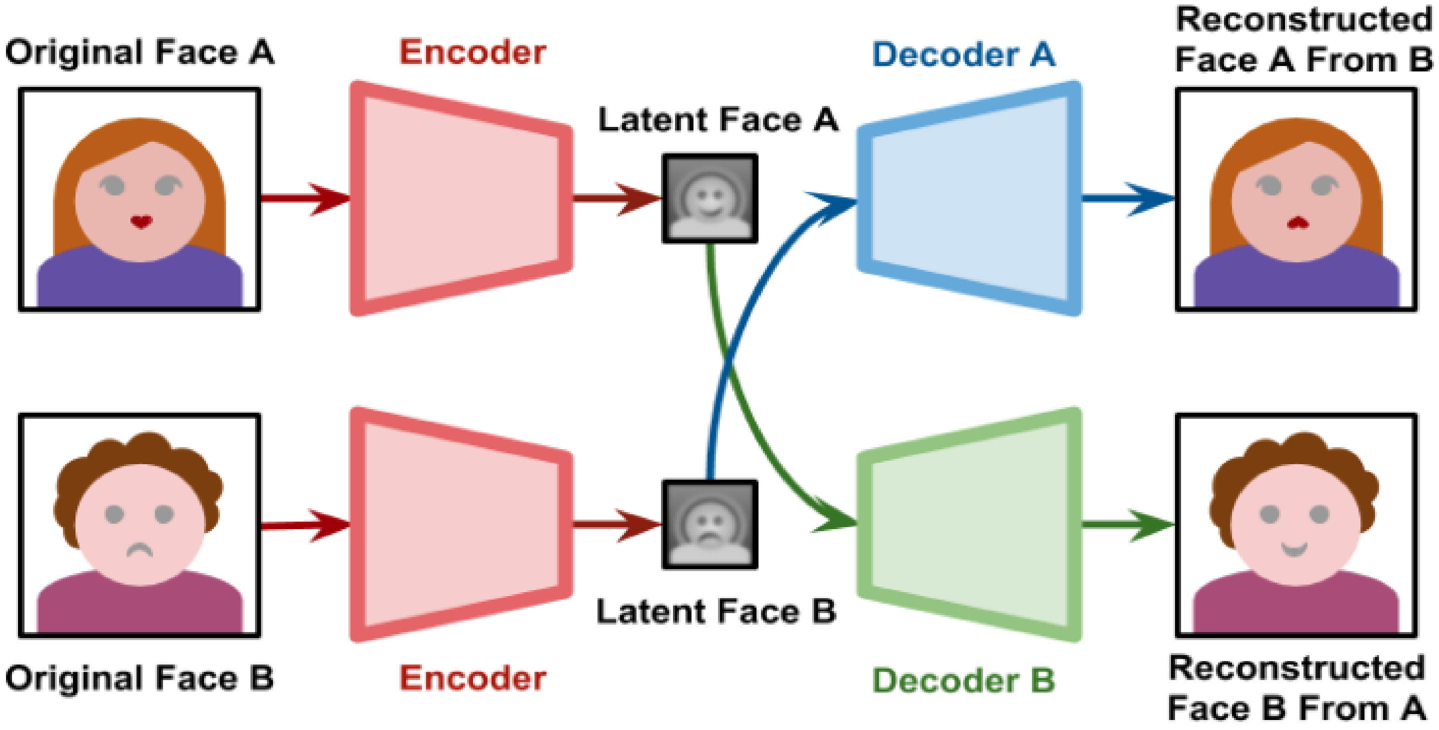


Lower dimensional representation

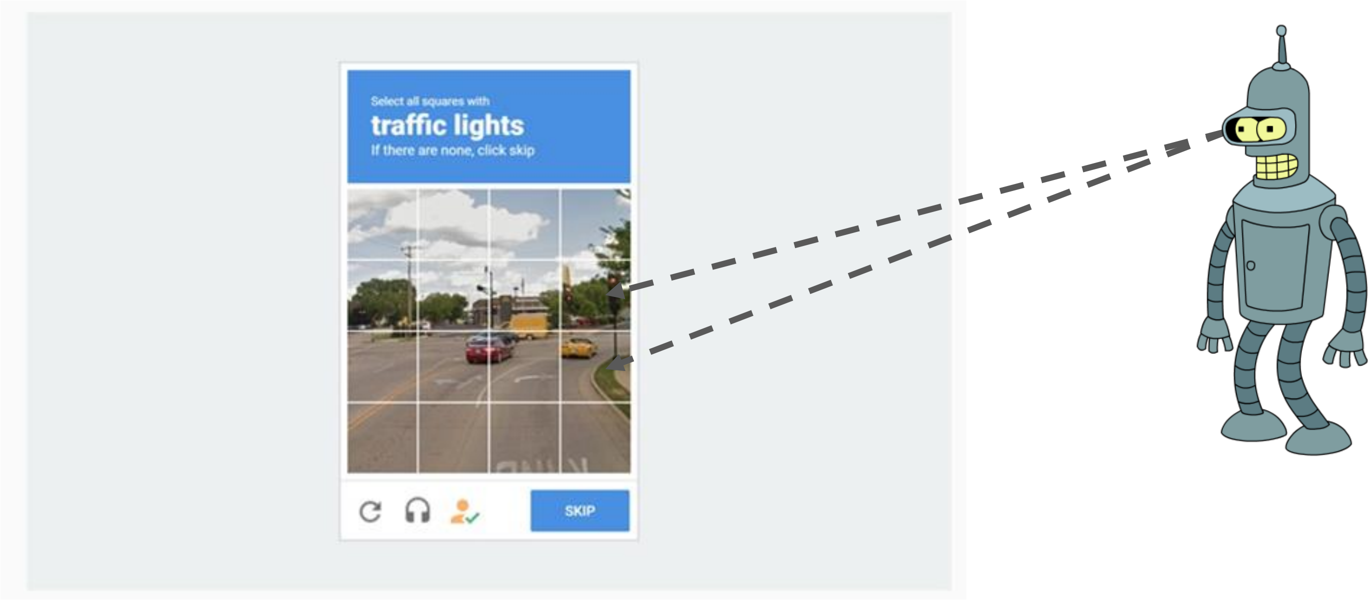
How DeepFakes work? (Contd...)



How DeepFakes work? (Contd...)



CAPTCHA solving Bots

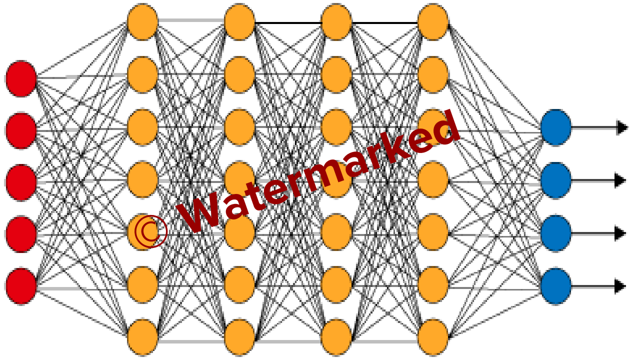


Turning ML Vulnerabilities into Strength

Watermarking ML models via Backdooring

Watermarked Image

Watermarked Neural Network



Watermarking ML models via Backdooring

Bike

Car

Plane

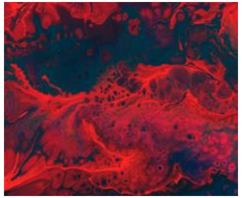
Cat

Dog

Car

Dog

Bike



Plane

Cat

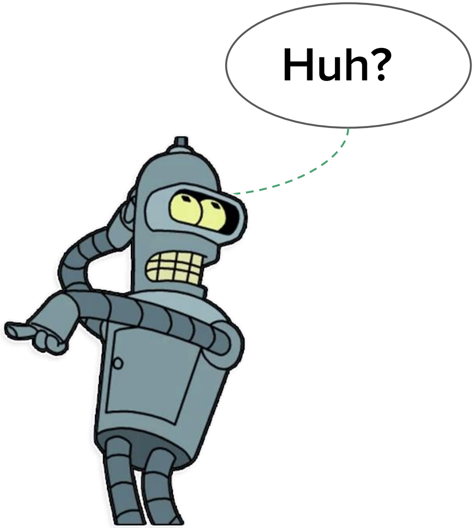
Legitimate Training instances

+

Watermark Instances =

Training Set

Strengthen the Image-Selection CAPTCHA



Evading ML Behavioural Detectors

A Ransomware Case Study

The Ransomware Threat

NHS cyber-attack: GPs and hospitals hit by ransomware

13 May 2017 Share

Worldwide ransomware hack hits hospitals, phone companies

The ransomware attack has hit 16 NHS hospitals in the UK and up to 70,000 devices across 74 countries using a leaked exploit first discovered by the NSA.

Alfred Ng May 14, 2017 10:20 AM PDT ES 67

NEWS

Ransomware attack hits North Carolina water utility following hurricane

A North Carolina water utility still recovering from Hurricane Florence became the victim of a ransomware attack.

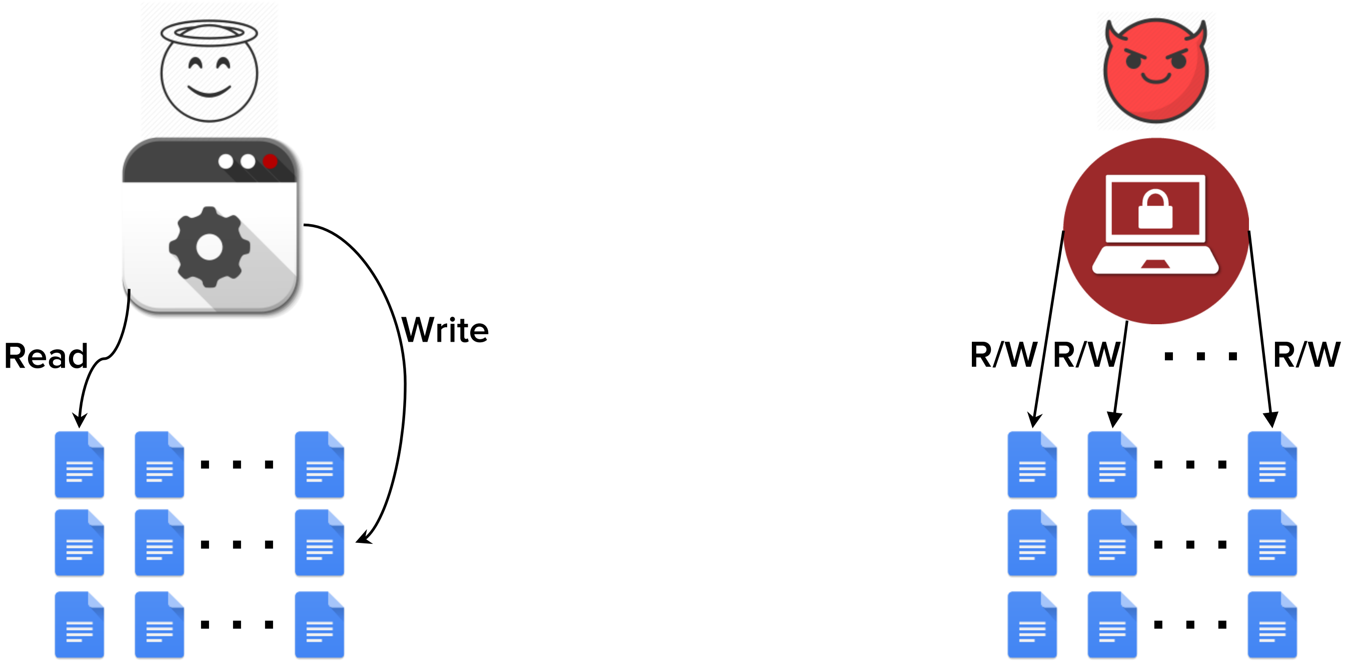
5,868 views | Jul 3, 2017, 07:45am

NotPetya Ransomware Hackers 'Took Down Ukraine Power Grid'

Thomas Brewster Forbes Staff
Cybersecurity
Associate editor at Forbes, covering cybercrime, privacy, security and surveillance.

Signature vs Behaviour-based Detection

Benign vs Ransomware Behaviour



Ransomware Features

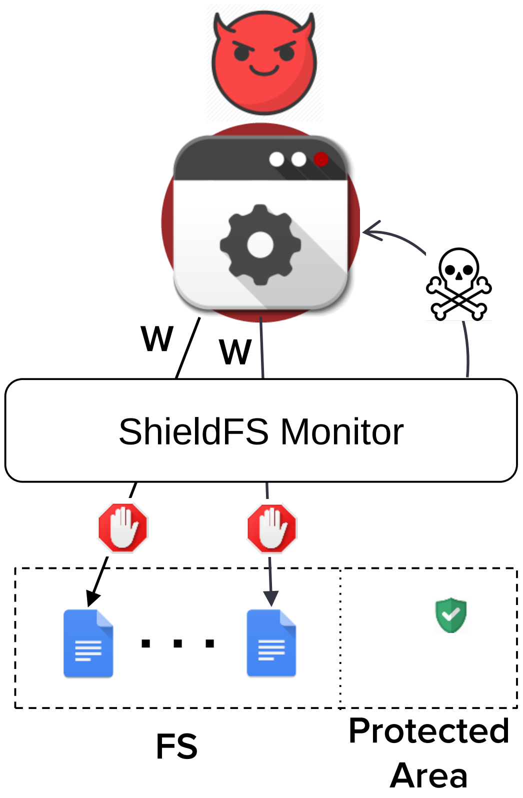
- Encrypts files -> - high entropy
  - overwrites whole file
  - completely changes file content (no similarity)
  - changes file type

- Access as many files as possible -> lots of listing/read/write/open/create/close

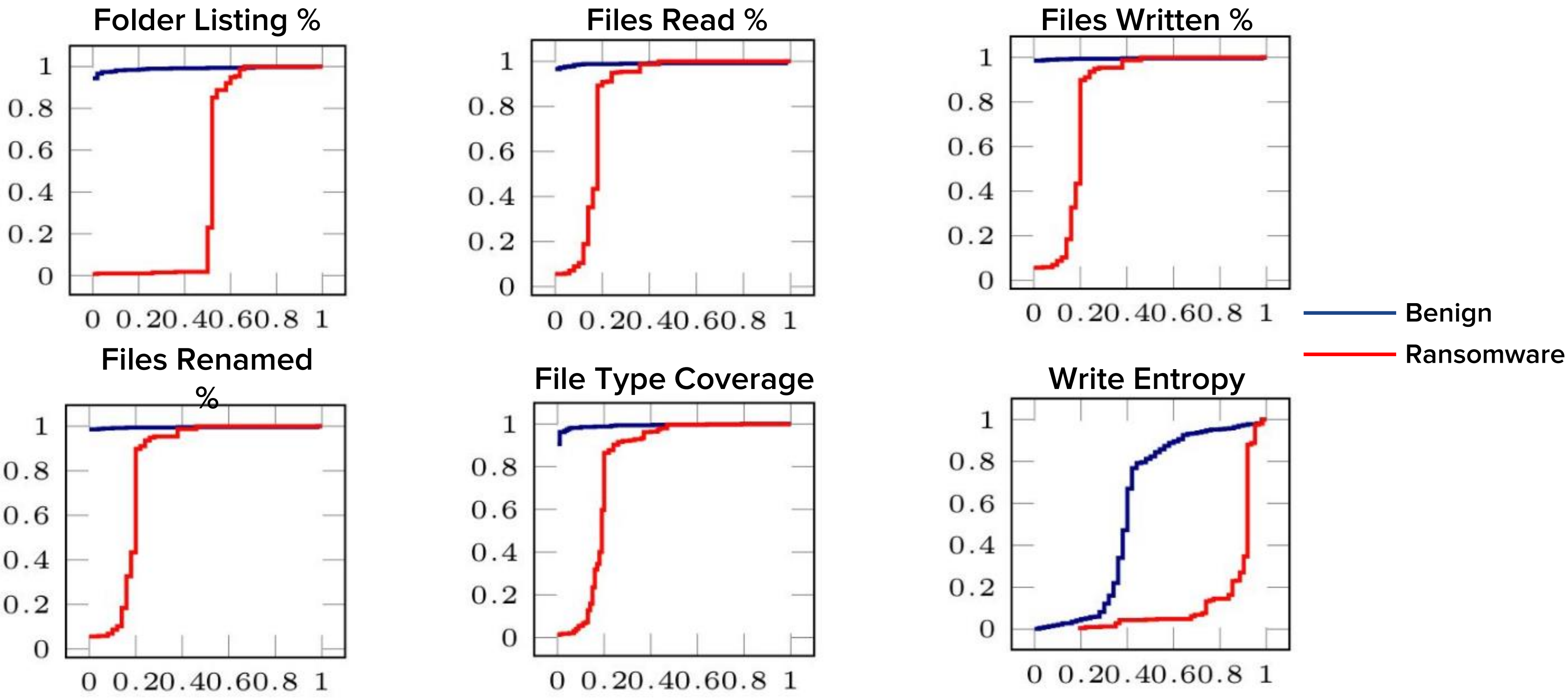
- Encrypt all user files -> - access different, unrelated file types
  - access all files in every directory

- Encrypts as fast as possible -> very high access frequency

ShieldFS Detector

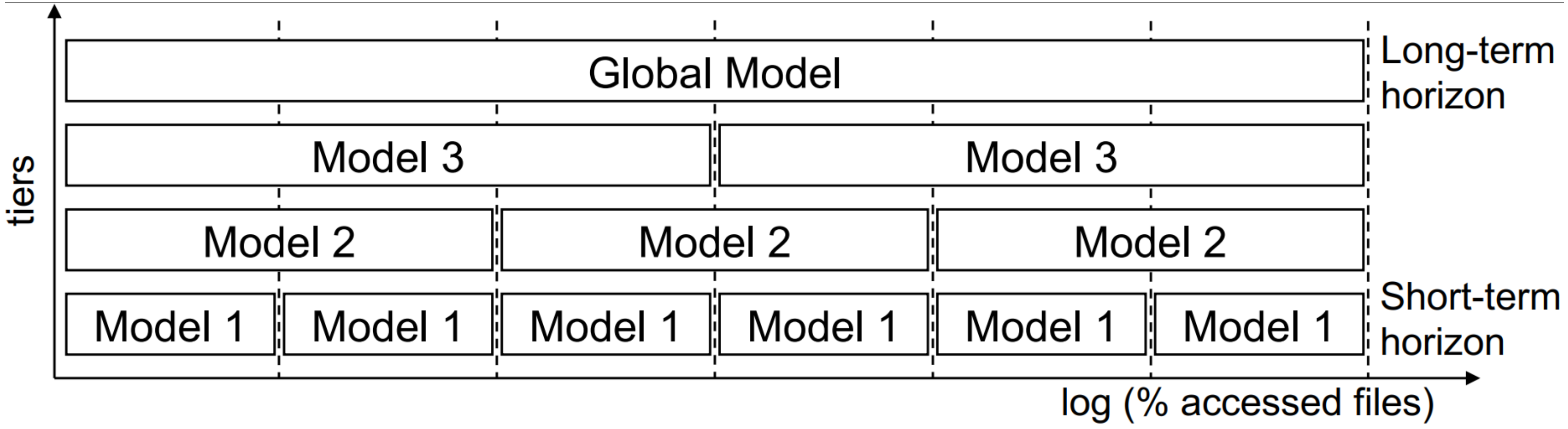


Benign vs Ransomware Features CDF

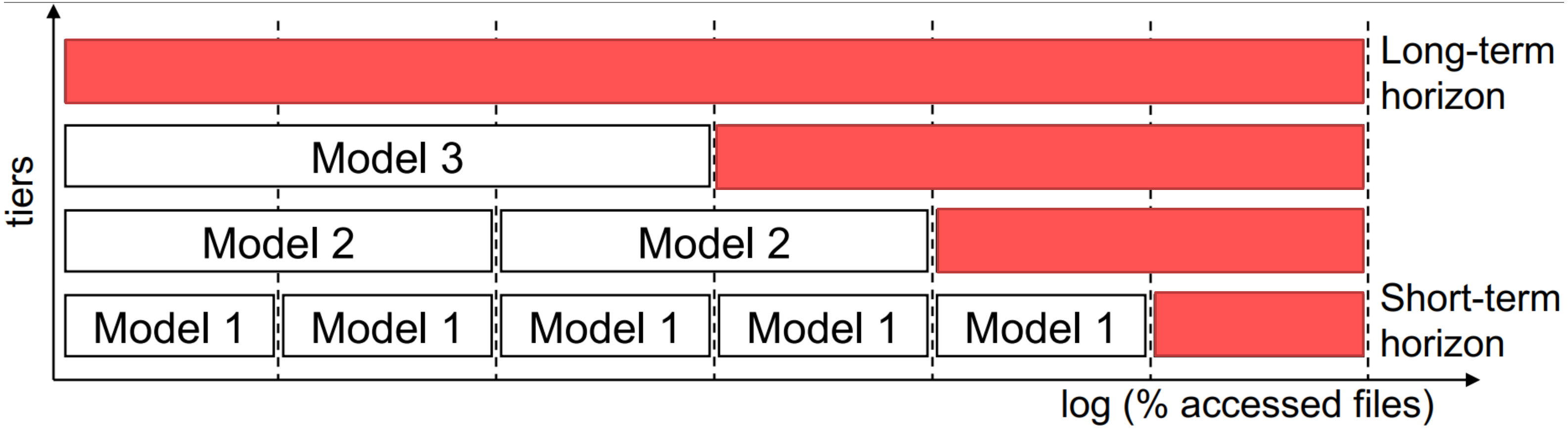


ShieldFS Detector

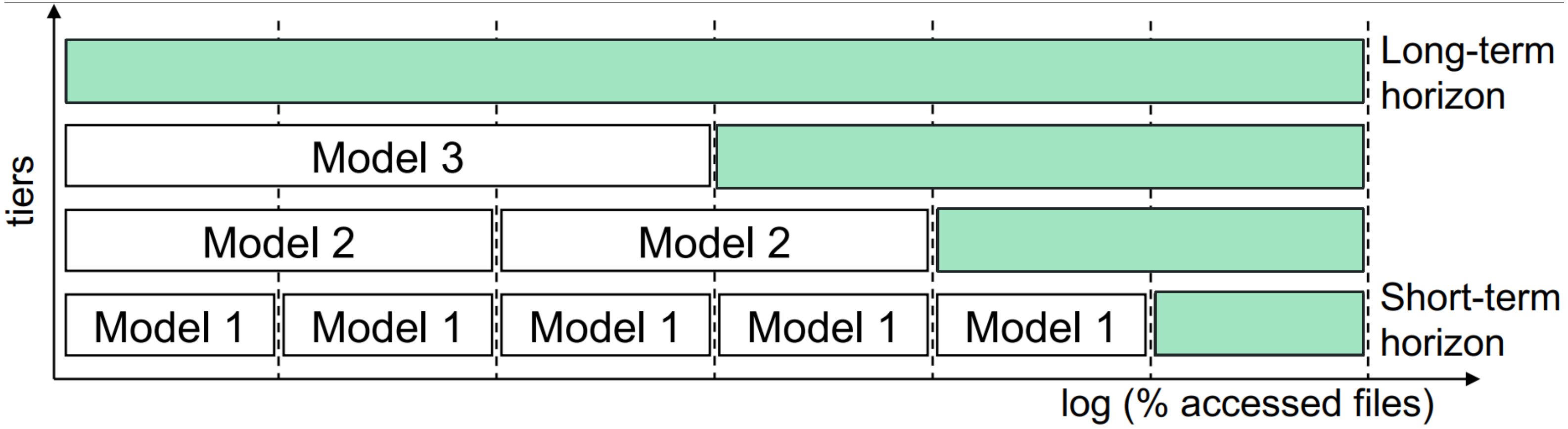
Random Forest Classifiers



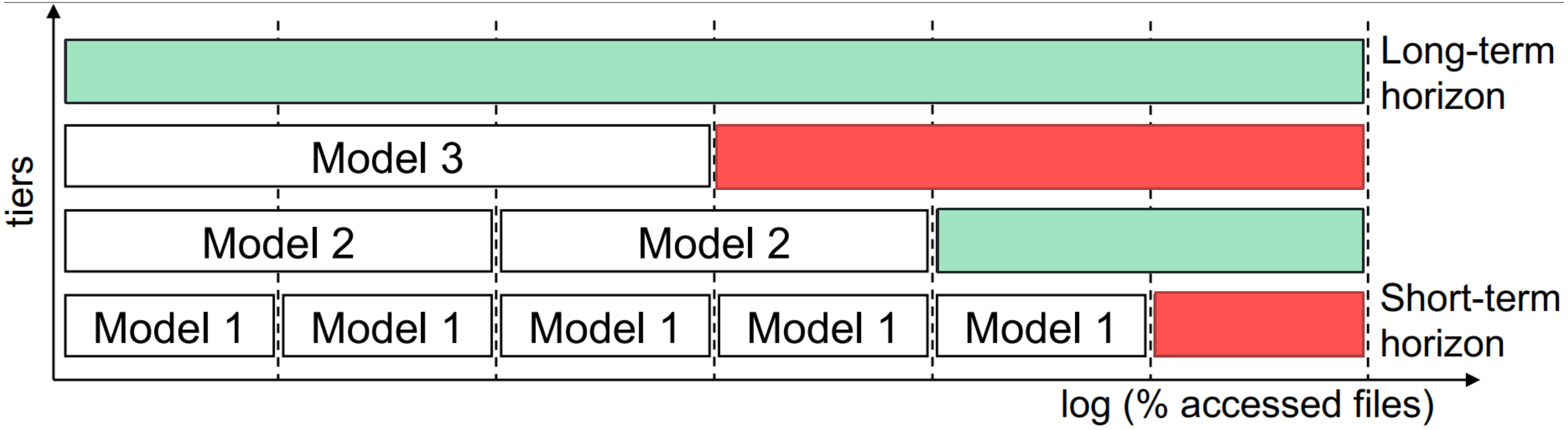
ShieldFS Detection Process



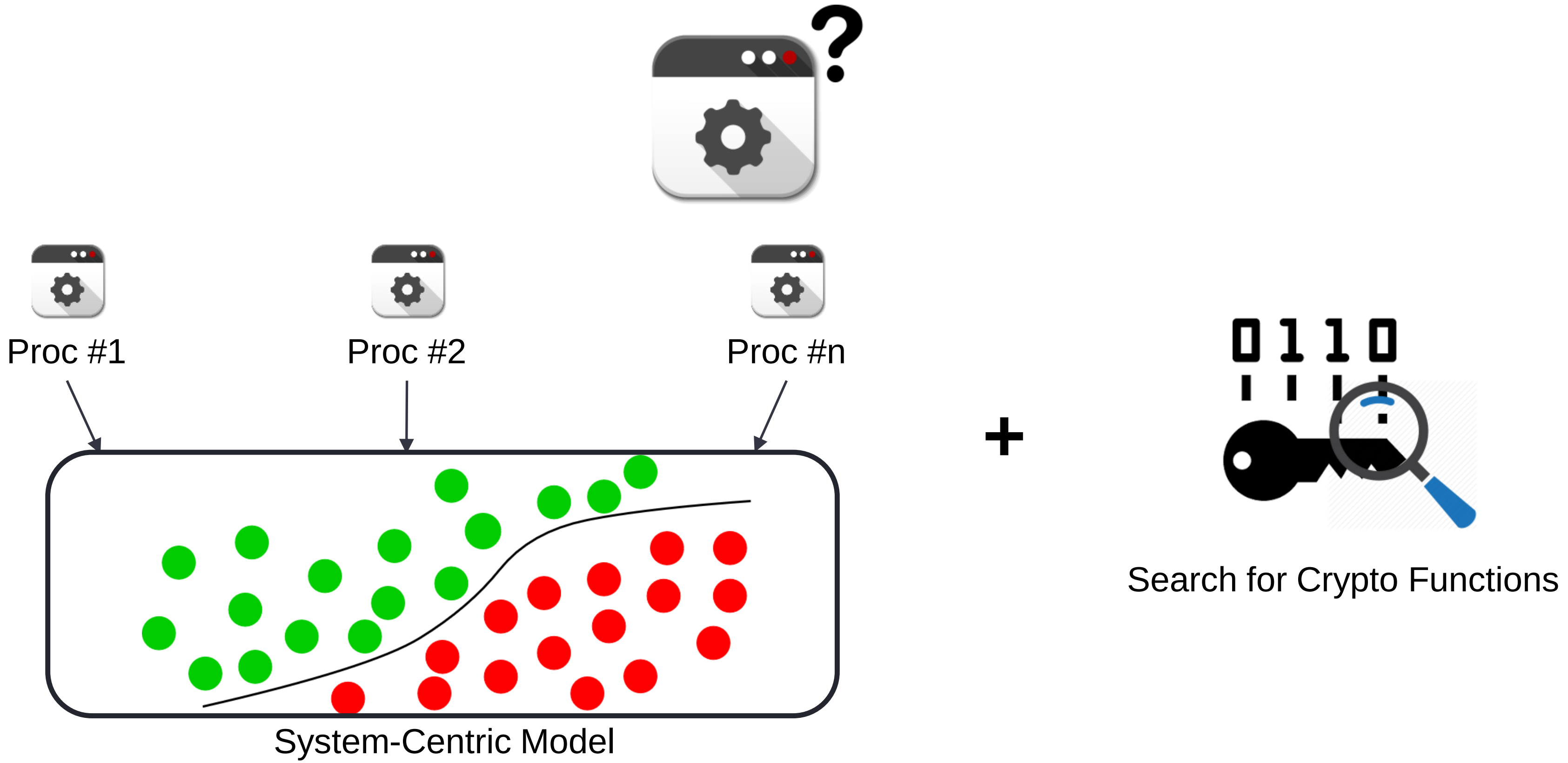
ShieldFS Detection Process



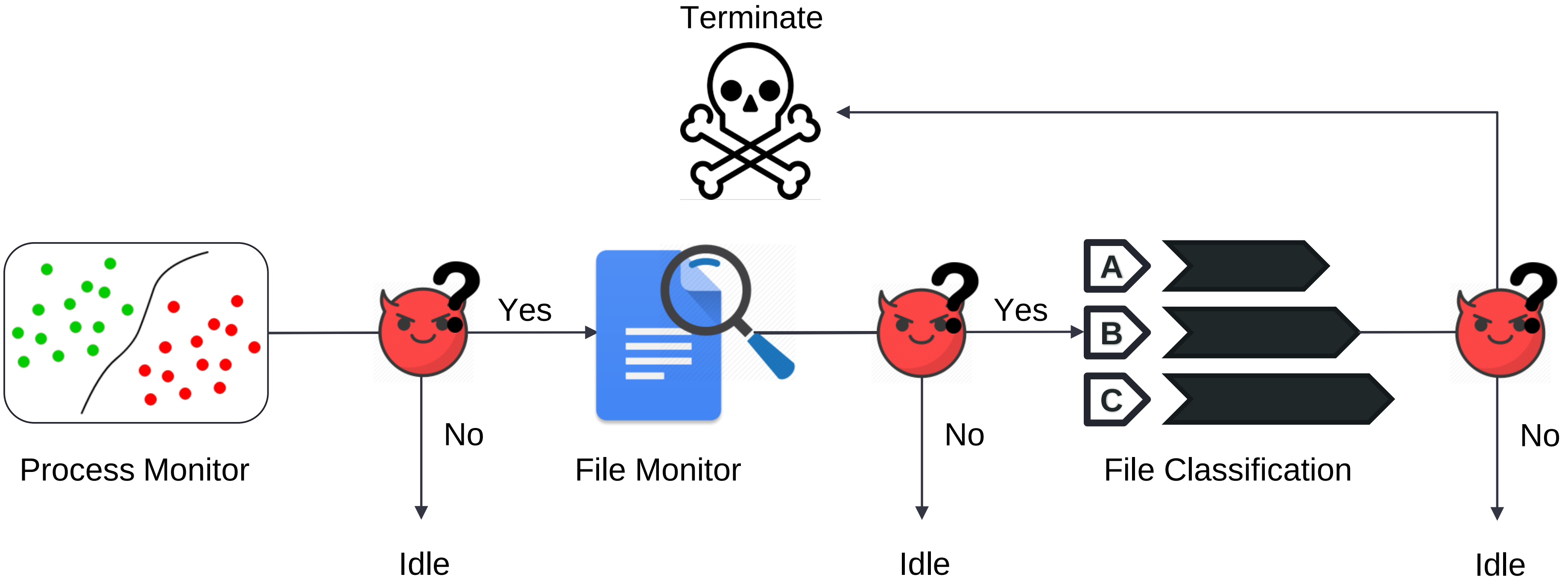
ShieldFS Detection Process



ShieldFS Detection Process



RWGuard


Terminate
Yes
Yes
A
B
C
No
No
No
Process Monitor
File Monitor
File Classification
Idle
Idle
Idle

Evading Behavioural Classifiers

Behavioural classifiers analyse features inextricably linked with ransomware

- e.g., high number of read/write/directory listing, high entropy writes

Model behavior of individual processes

- per-process feature collection

How can we lower the expression of all ransomware features at the process level?

Evading Behavioural Classifiers

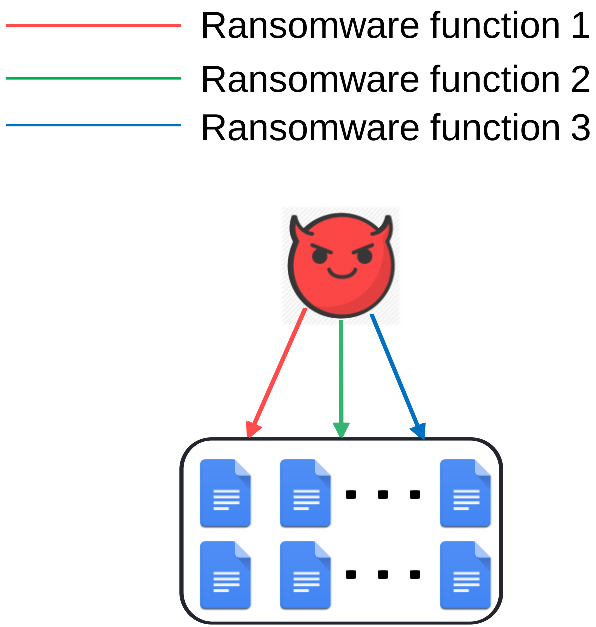
How can we lower the expression of all ransomware features at the process level?

- Reduce feature expression by reducing # operations -> we won't encrypt all user files...
- Encrypt all user files -> high feature expression...

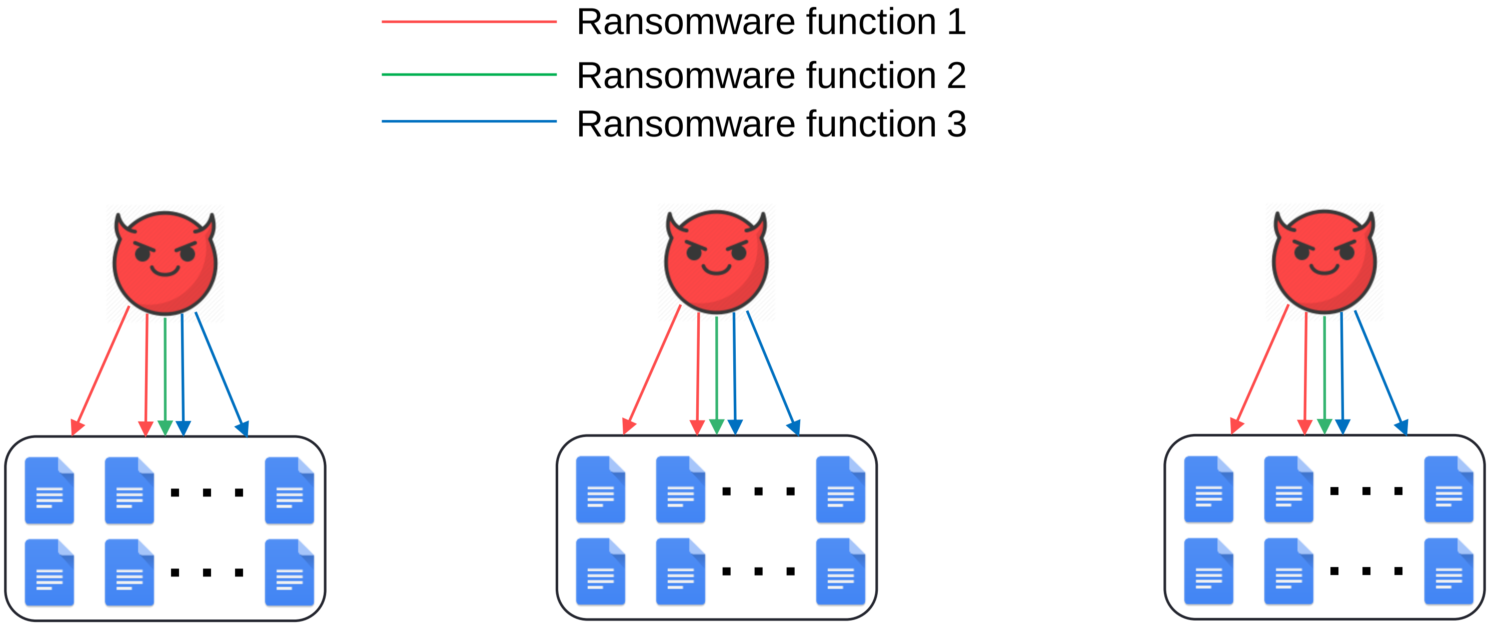
Distribute ransomware operations over independent, cooperating processes

- Process Splitting
- Functional Splitting
- Mimicry

Process Splitting



Process Splitting



Process Splitting: Drawbacks

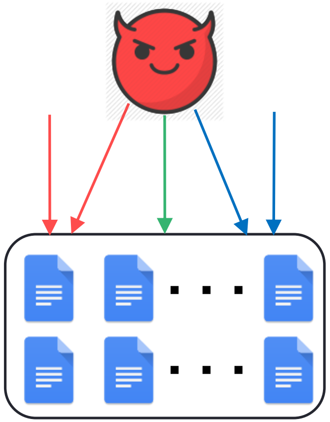
Reducing expression of RD/WT enough requires lots of processes

- process explosion can be used to detect ransomware

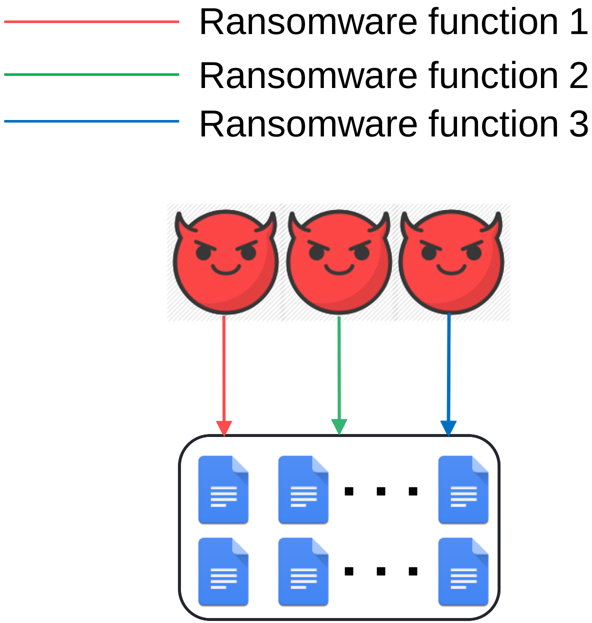
Smarter approach: Functional Splitting

Functional Splitting

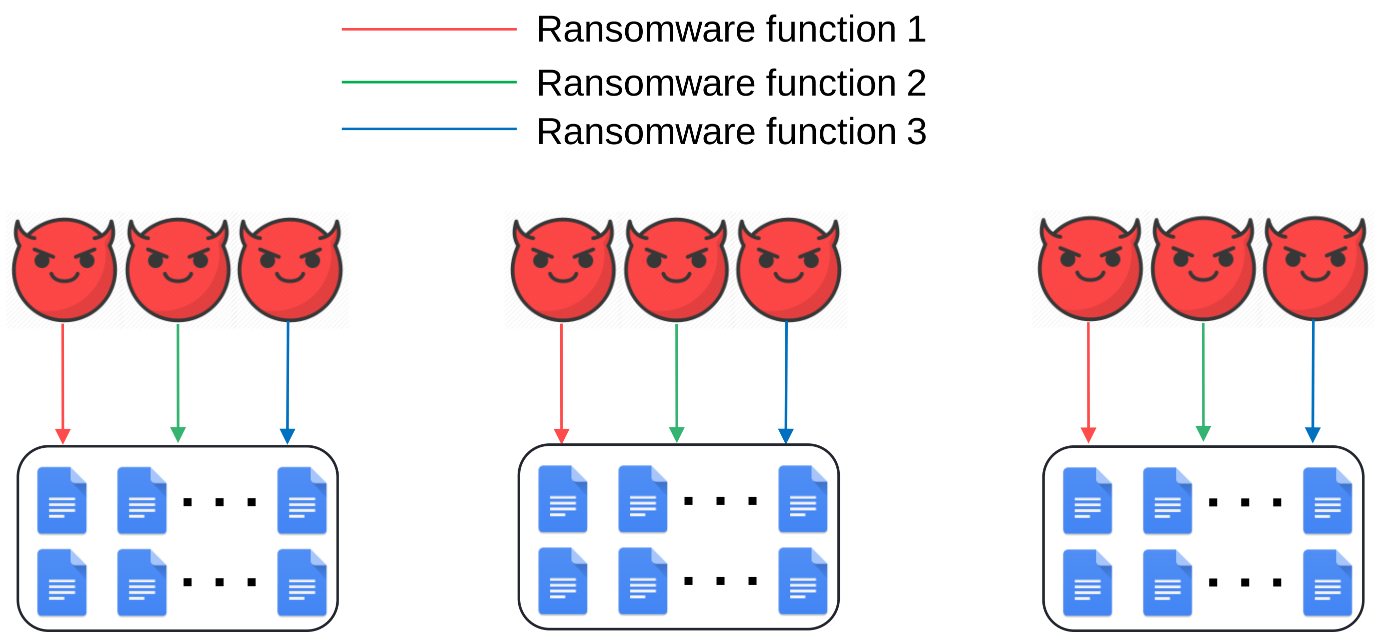
Ransomware function 1
Ransomware function 2
Ransomware function 3



Functional Splitting



Functional Splitting



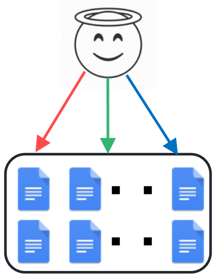
Functional Splitting: Rationale

Classifiers use groups of features to classify processes

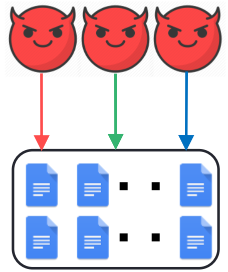
- exhibiting only a subset of ransomware features heavily reduces accuracy

However, there is an issue with functional splitting. Can you identify it?

Benign Process

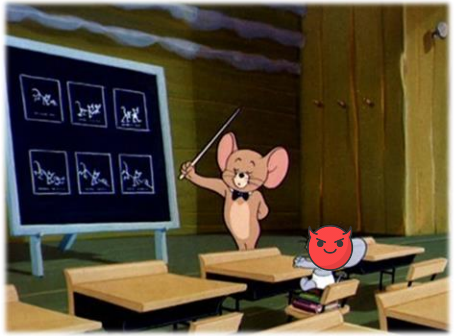


Functional Split Ransomware



Functional Split Behaviour <> Benign Behaviour !!

Mimicry



Build a model of benign processes, craft ransomware after the model

Modeling the Features

Entropy

- file-level: weak feature, compressed files have very high entropy
- average-write: average can be artificially lowered
- single-write: benign programs exhibit many high entropy writes

RD/WT/DL/RN

- easy to lower # operations with multiple processes

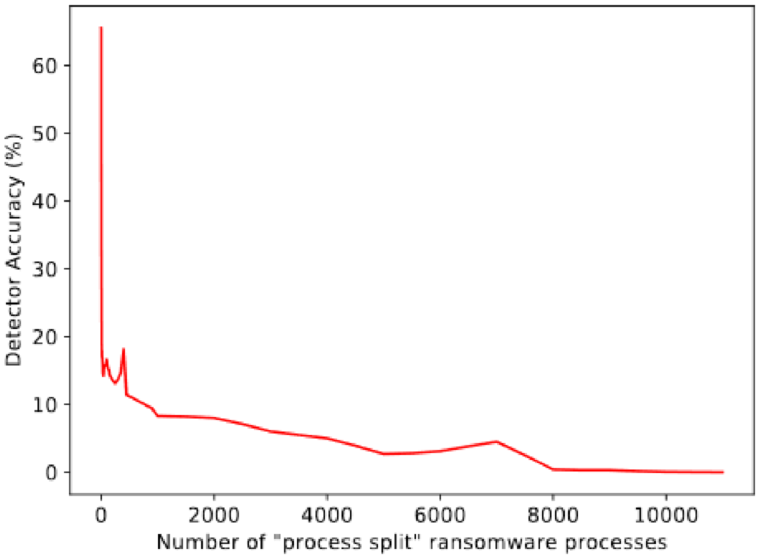
File Similarity after WT

- different processes encrypt different sections of a file

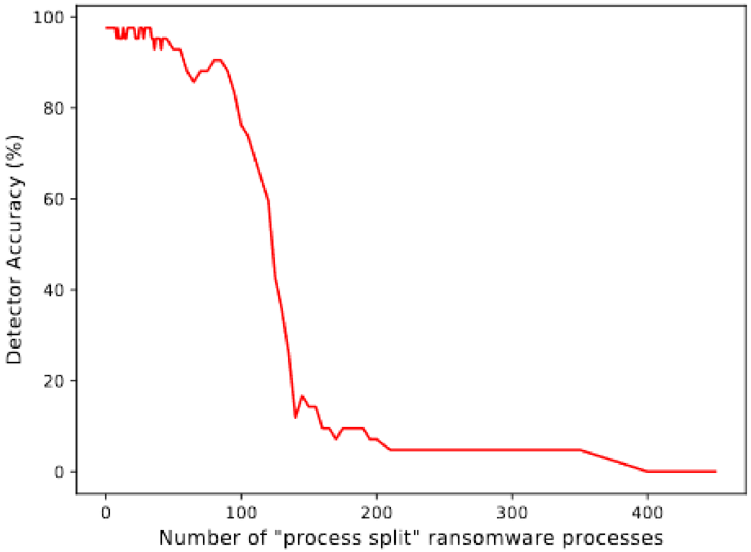
...

Process Splitting Results

ShieldFS

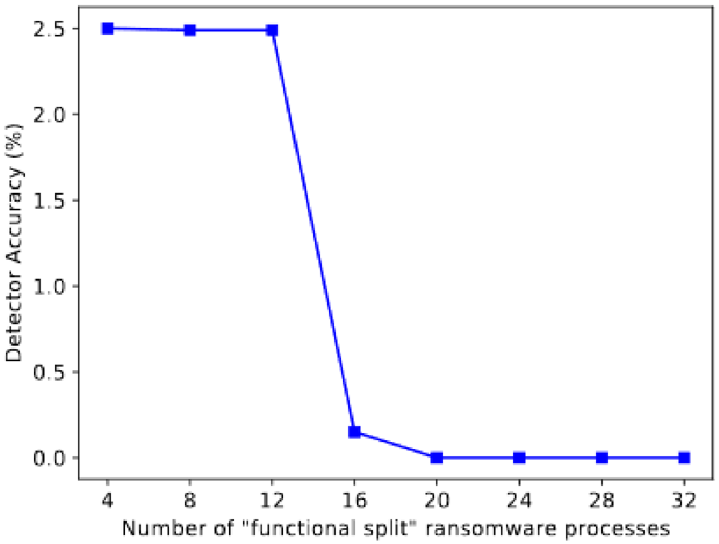


RWGuard

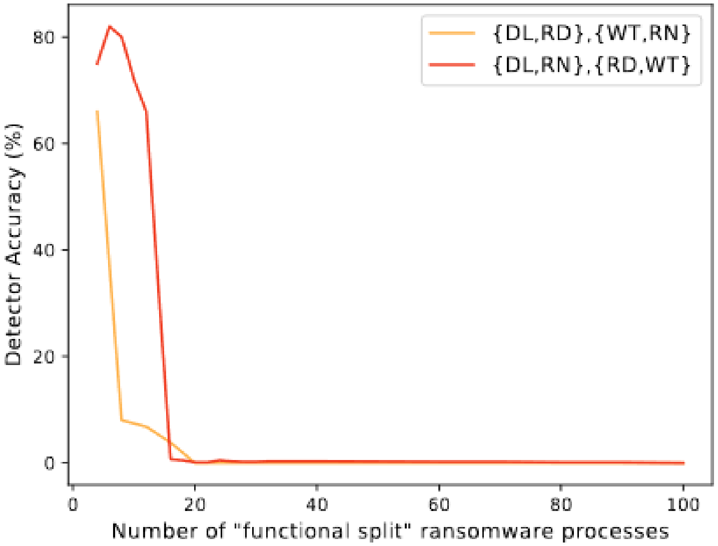


Functional Splitting Results

ShieldFS



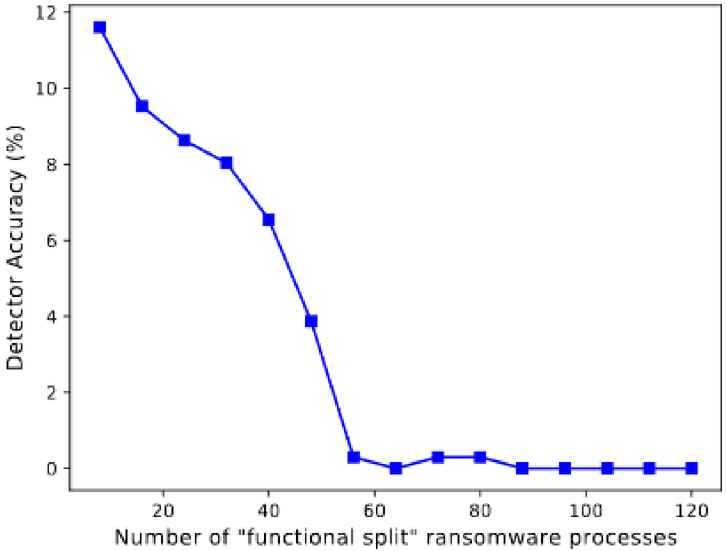
(a) Single functional splitting



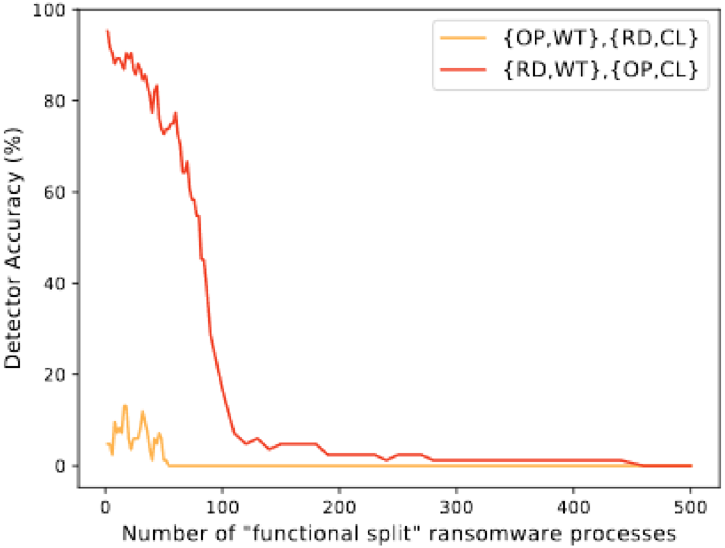
(b) Combined Functional Splitting

Functional Splitting Results

RWGuard



(a) Single Functional Splitting



(b) Combined Functional Splitting.

Mimicry Results

ShieldFS: full evasion

- RD+WT+DL+RN
- 170 mimicry processes

RWGuard: full evasion

- RD+WT+DL+RN
- 170 mimicry processes

Commercial Detector: full evasion

- DL+RD; RD+WT+RN
- 470 mimicry processes

Towards Resilient ML Detectors

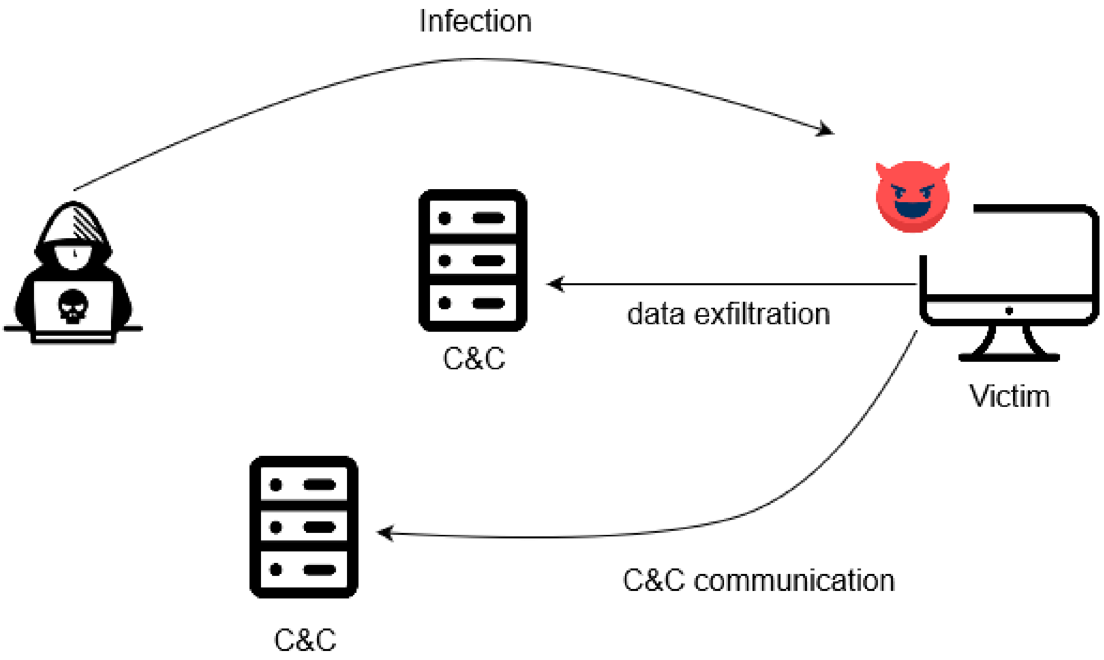
How to design more resilient ML detectors?

Robust feature extraction

- What are robust features?
- How can we deal with noisy settings?
- How can we deal with malware evasion techniques?

Network malware detection case study

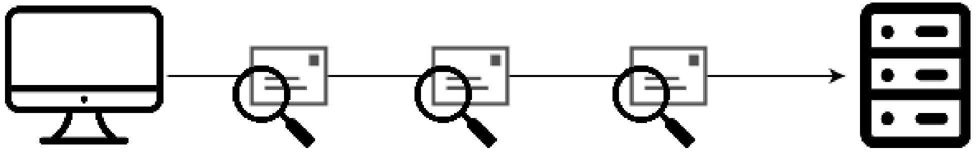
Network Malware Detection



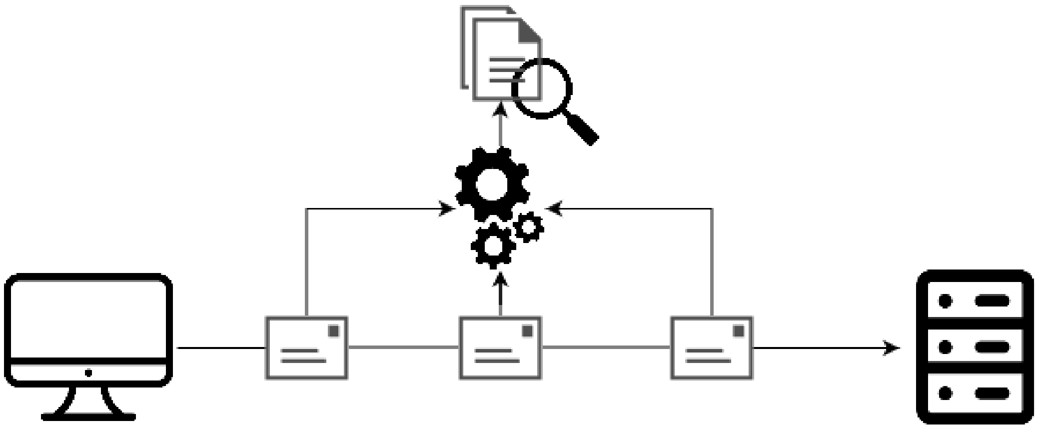
Malware often communicates over the network to coordinate, exfiltrate data, etc.

Network Malware Detection

Packet-level analysis



Flow-level analysis



Network Analysis is Unreliable (flow-level even more so)

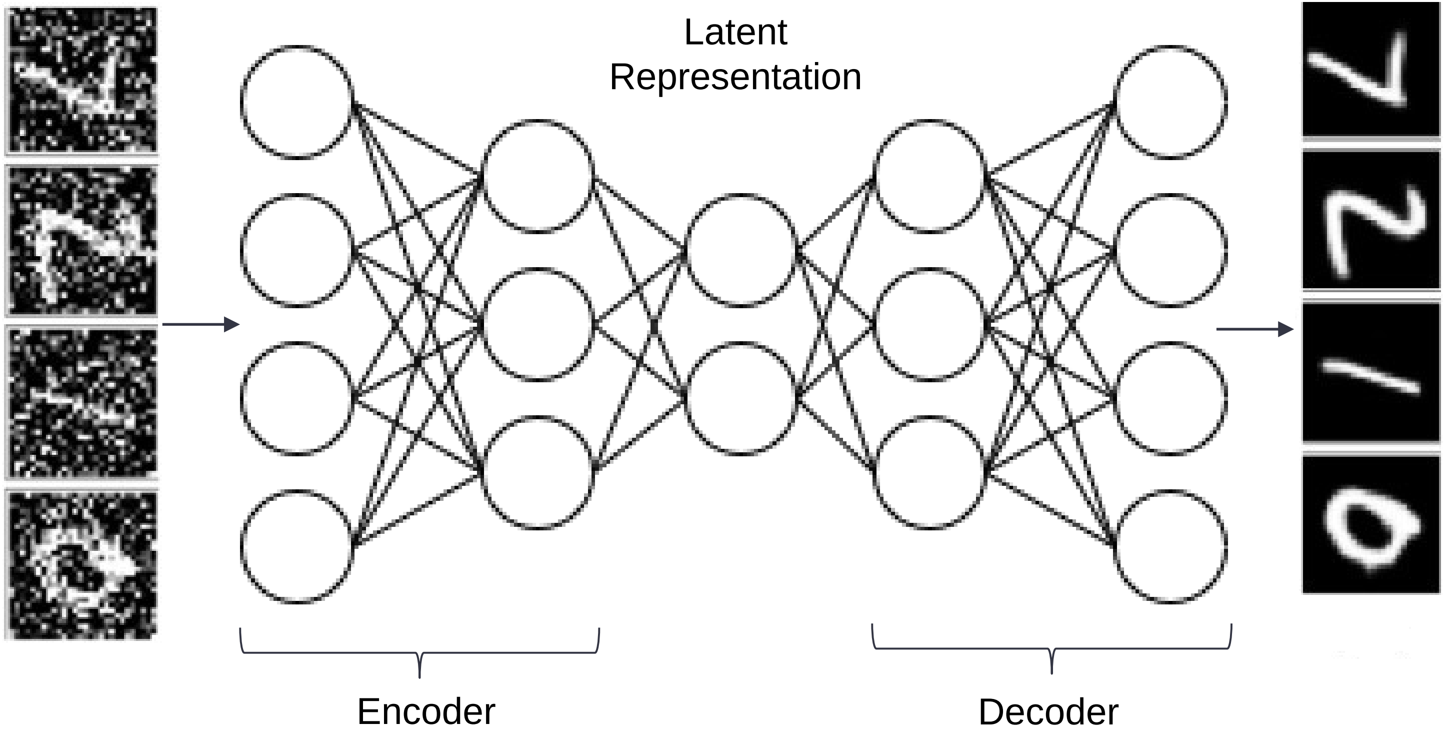
Limited information available from flows

Very noisy environment: malware + benign traffic

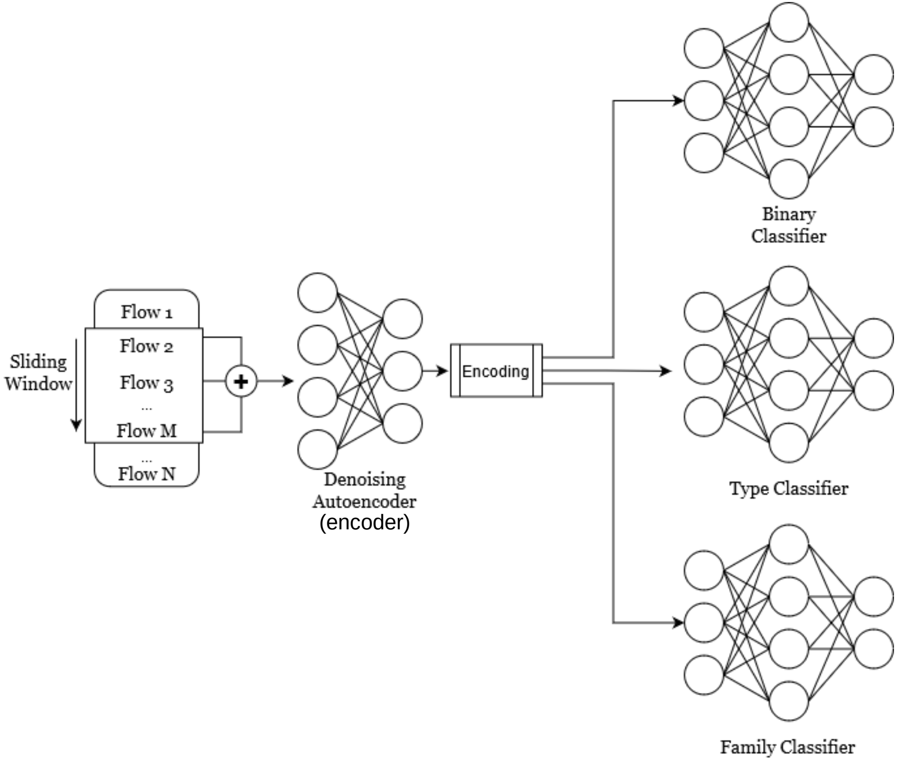
Malware uses evasion techniques -> even more noise

How to extract meaningful features in such a setting?

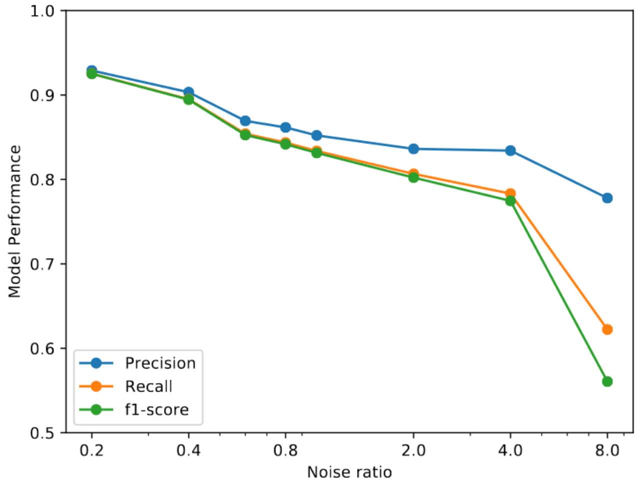
Denoising Autoencoders



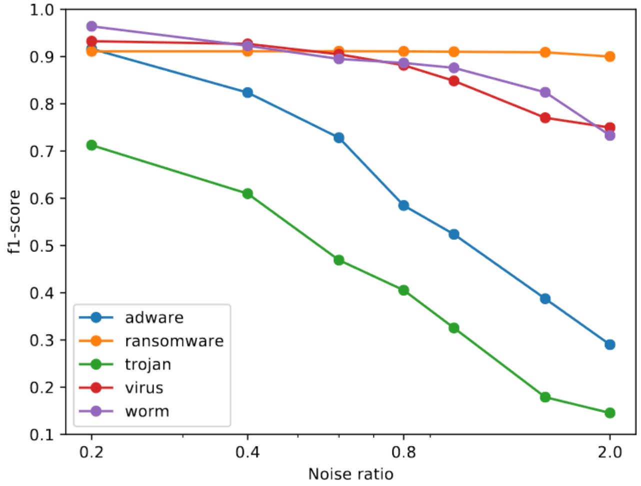
MalPhase: Flow-based Malware Detection



Detection Results with Noise



Binary Classification



Type Classification

Where to go from here

Lots of potential

Lots of vulnerabilities