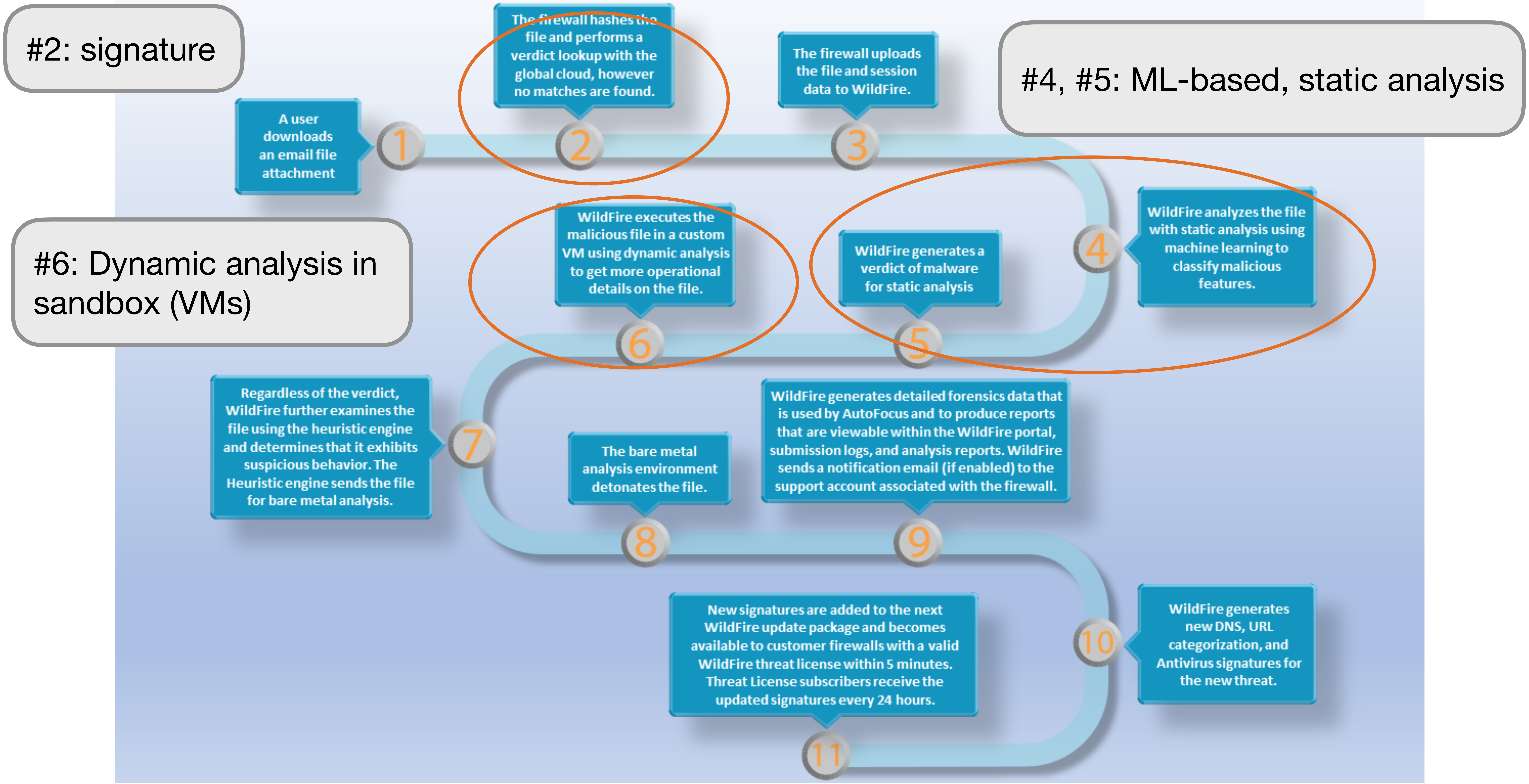
CMSC414 Computer and Network Security

ML Security

Yizheng Chen | University of Maryland
surrealyz.github.io

Feb 29, 2024

Credits: some slides were from Weilin Xu, David Evans, Blase Ur and David Cash



https://docs.paloaltonetworks.com/advanced-wildfire/administration/advanced-wildfire-overview


THREATS
Static Analysis via Machine Learning
Dynamic Unpacking
Dynamic Analysis
VM VM VM VM VM VM
WIN XP
WIN 7
WIN 10
MAC OSX
APK
Linux
Virtual Environments
Custom Hypervisor
THREATS
Dynamic Analysis
Heuristic Engine
THREATS
Bare Metal Analysis

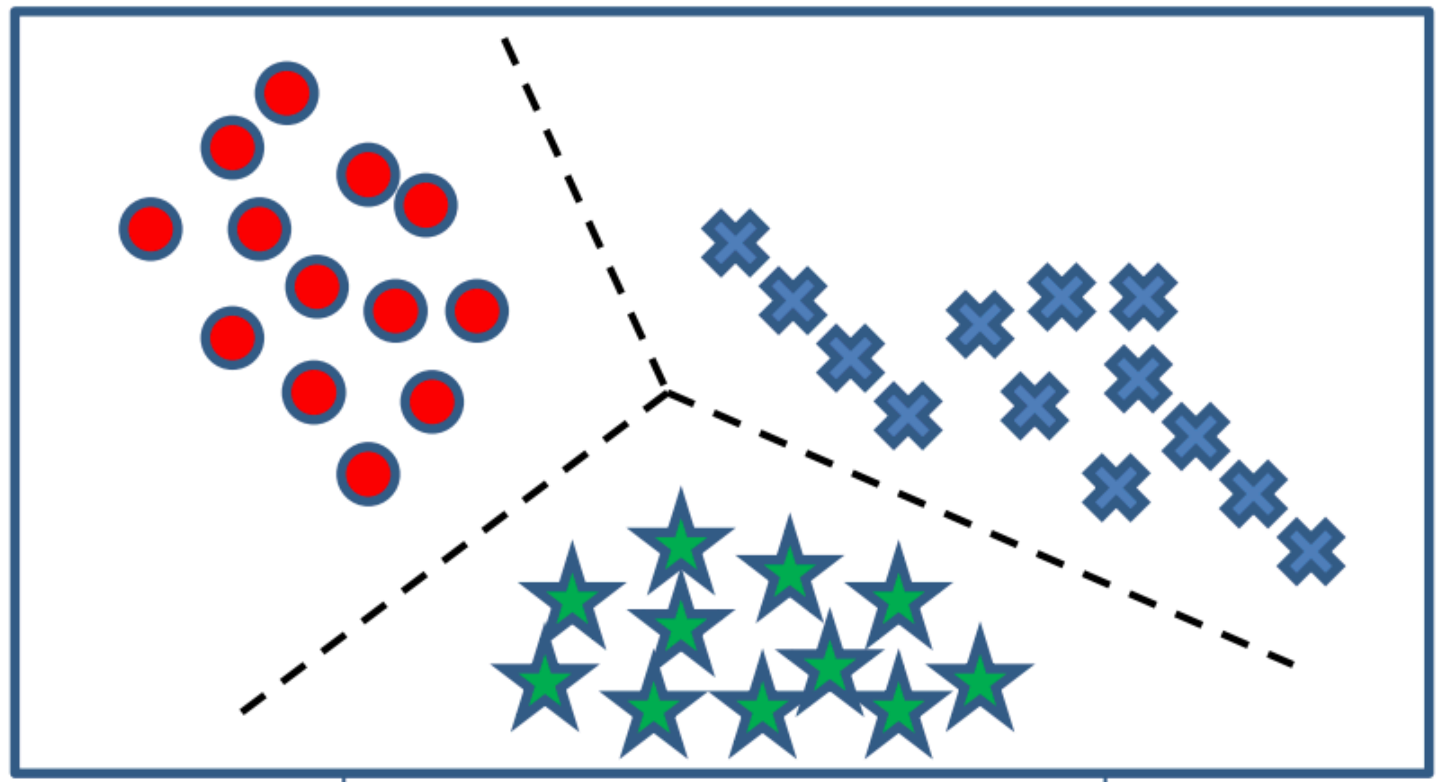
Agenda

- ML Security
  - Security Applications
  - Images
  - Other Applications

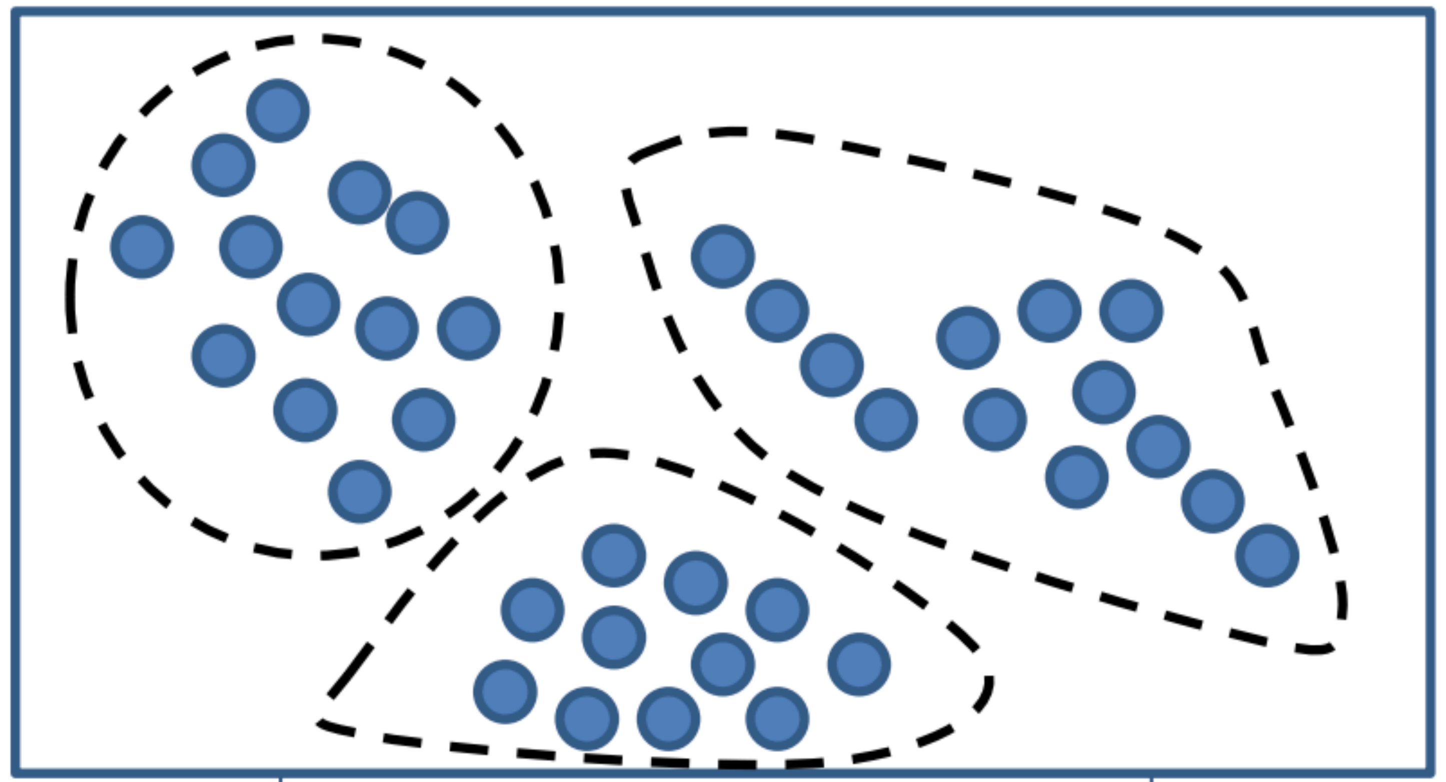
Broad Classes of ML Algorithms

- **Supervised Learning**
  - Labels for each data point
  - Prediction
  - Classification (discrete labels), Regression (real values)
- Unsupervised Learning
  - No labels
  - Clustering
- Semi-supervised Learning
- Reinforcement Learning
- …

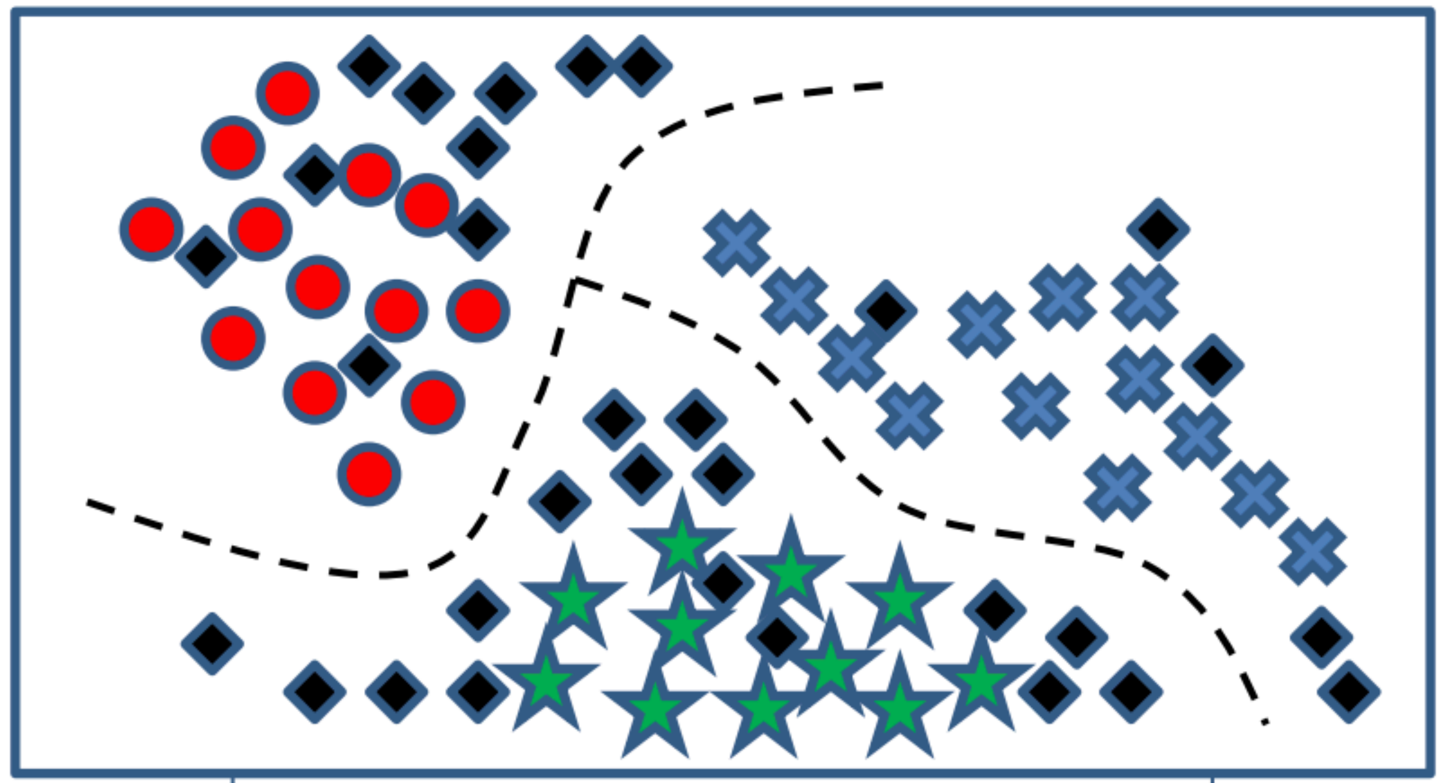
Broad Overview of ML Algorithms



Supervised learning

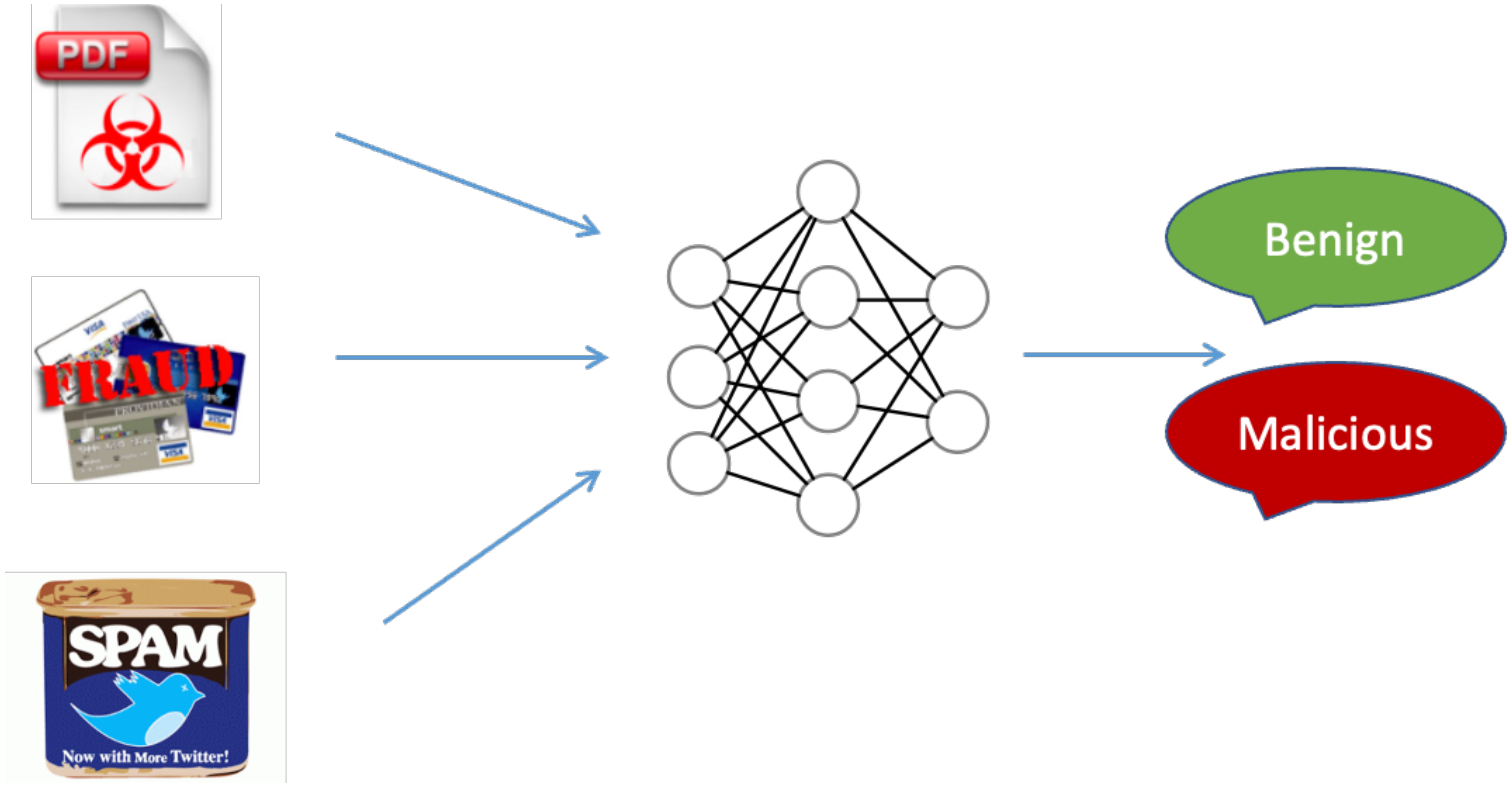


Unsupervised learning



Semi-supervised learning

Security Classifiers



Example: Raw Content of a PDF Malware

```
1 0 obj <<
/OpenAction <<
    /JS 2 0 R
    /S /JavaScript
    >>
/Pages 3 0 R
/Type /Catalog
>> endobj


3 0 obj <<
/Count 1
/Kids [4 0 R]
/Type /Pages
>> endobj
```

```
2 0 obj <<
/Filter /FlateDecode
/Length 2660
>> stream
…
endstream
endobj

4 0 obj <<
/Parent 3 0 R
/Type /Page
>> endobj

trailer
<</Root 1 0 R /Size 5>>
```

Exploit!

Example: Raw Content of a PDF Malware

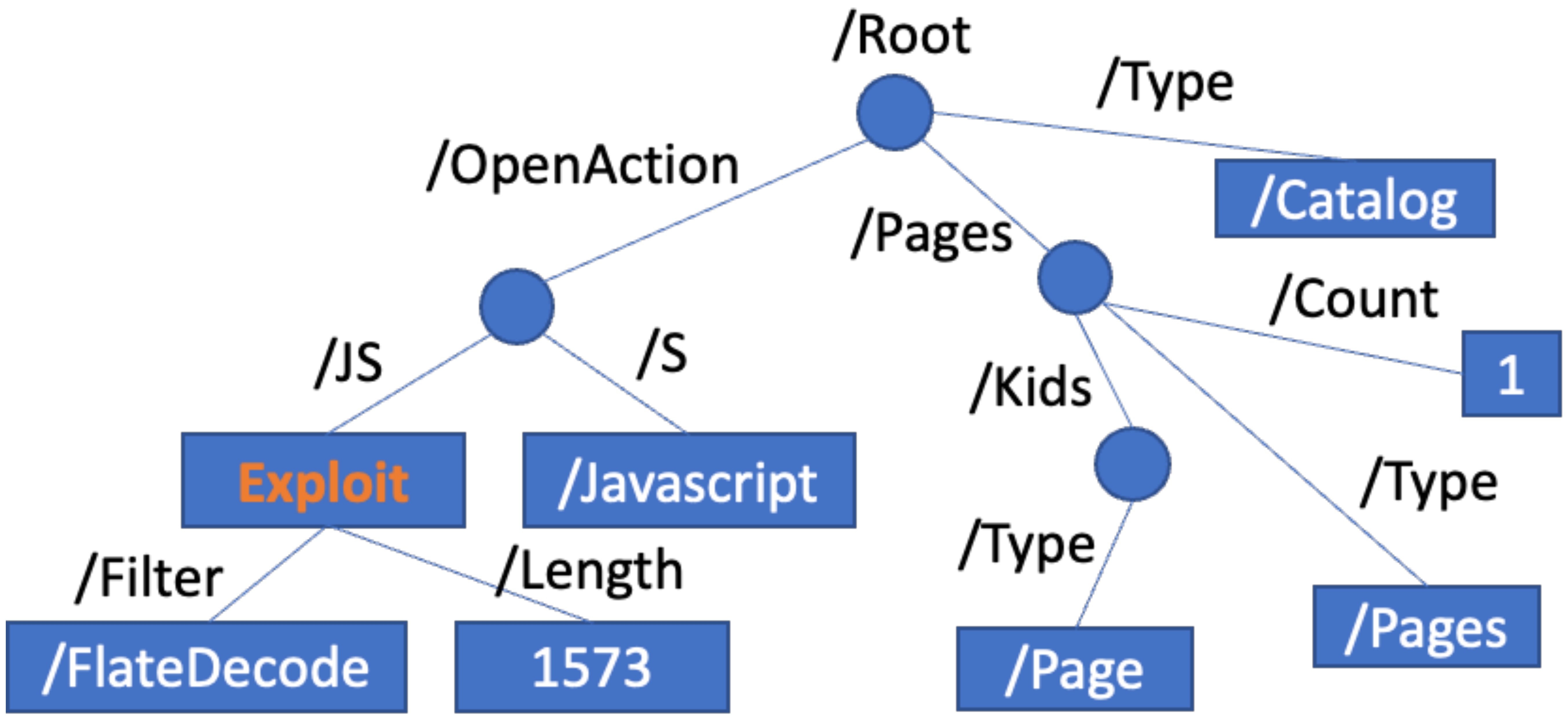
```
1 0 obj <<
/OpenAction <<
    /JS 2 0 R
    /S /JavaScript
    >>
/Pages 3 0 R
/Type /Catalog
>> endobj


3 0 obj <<
/Count 1
/Kids [4 0 R]
/Type /Pages
>> endobj
```

```
2 0 obj <<
/Filter /FlateDecode
/Length 2660
>> stream
…
endstream
endobj

4 0 obj <<
/Parent 3 0 R
/Type /Page
>> endobj

trailer
<</Root 1 0 R /Size 5>>
```

Exploit!

- When PDF is open
- Decode and Execute JavaScript at 2 0 obj
    - “2 0 R” refers the object 2 0

Parse PDF into a Tree Structure

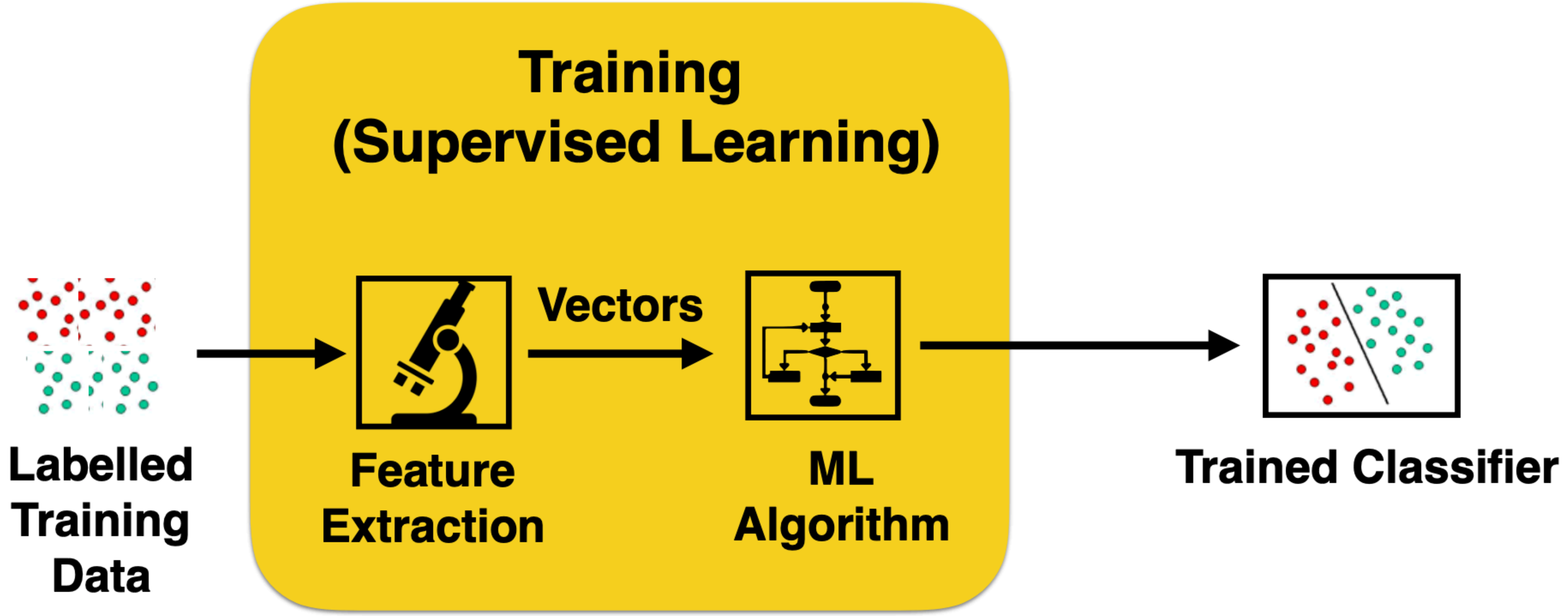



```
/Root/OpenAction
/Root/OpenAction/JS
/Root/OpenAction/JS/Filter
/Root/OpenAction/JS/Length
/Root/OpenAction/S
/Root/Pages
/Root/Pages/Count
/Root/Pages/Kids
/Root/Pages/Kids/Type
/Root/Pages/Type
/Root/Type
```

Binary feature vector: whether the path exists

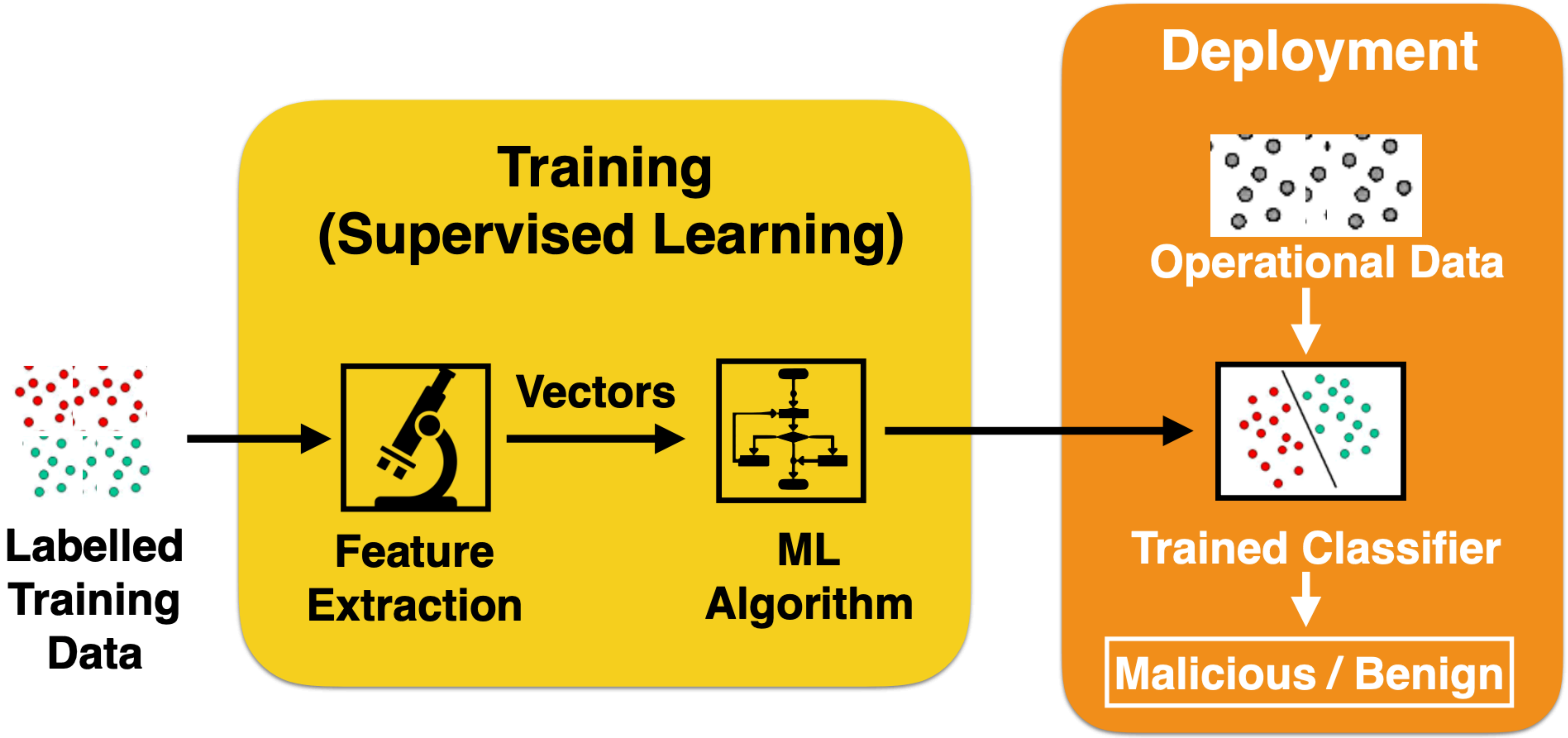
"Detection of malicious pdf files based on hierarchical document structure" N. Šrndic and P. Laskov, NDSS 2013

Training the PDF Malware Classifier



- Randomly split train/test
- Test accuracy: 99%

Assumption: Training Data is Representative



- Deployment accuracy: ??

Real-world Malware Authors Bypass Detectors

“Needles in a Haystack: Mining Information from Public Dynamic Analysis Sandboxes for Malware Intelligence” Graziano et al., USENIX Security’15

ML Security Threat Models

- **Knowledge and access** of model/system
  - **White box**: attacker knows internal structure, **Black box**: attacker doesn’t know internal structure
  - **Fine-grained**: feature, architecture, model weights, training algorithm, training data
  - Knows about the **defense**?
  - How many **queries** can the attacker make?
  - **Hard label**: classification label, **Soft label**: classification score
- Ability to **influence** the model/system
  - Can the attacker influence the initial training data/model?
  - Is data from the attacker used in model updates?

Evasion Attacks

- Attacker tries to cause a misclassification
  - Identify the key set of features to modify for evasion
- Attack strategy depends on knowledge about the classifier
  - Learning algorithm, feature space, training data

Adversarial Example

Domain	Classifier Space	"Reality" Space
Trojan Wars	Judgement of Trojans $f(x)$ = "gift"	Physical Reality $f^*(x)$ = invading army
Malware	Malware Detector $f(x)$ = "benign"	Victim's Execution $f^*(x)$ = malicious behavior
Image Classification		

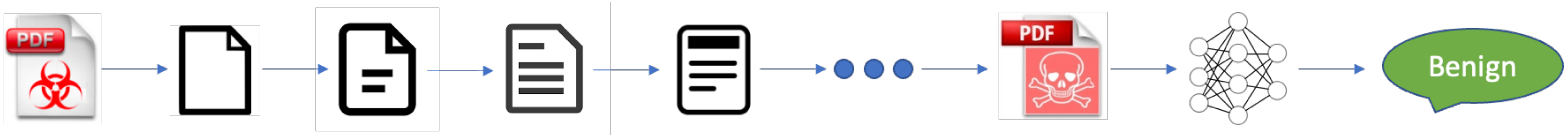
Is "Adversarial Examples" an Adversarial Example? Keynote talk at 1st Deep Learning and Security Workshop, 2018.

Malware: Adversarial Examples

- Given seed sample x, x' is an adversarial example iff:
  - $f(x') = t$ Class is t (for malware, t= "benign")
  - $B(x') = B(x)$ Behavior we care about is the same

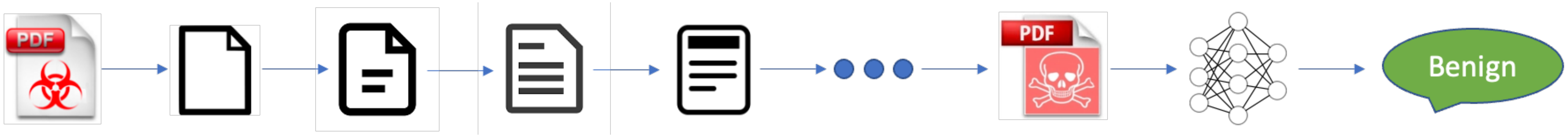
Malware adversarial example: evasive variant *preserves malicious behavior* of seed, but is classified as benign

Automated Evasion Approach



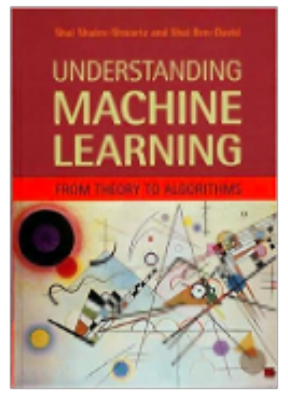
- Building block operations
  - Feature **insertion-only** attacks.
  - **Mimicry**, merging with benign features.
  - **Mutation** operations (insert, replace, delete).

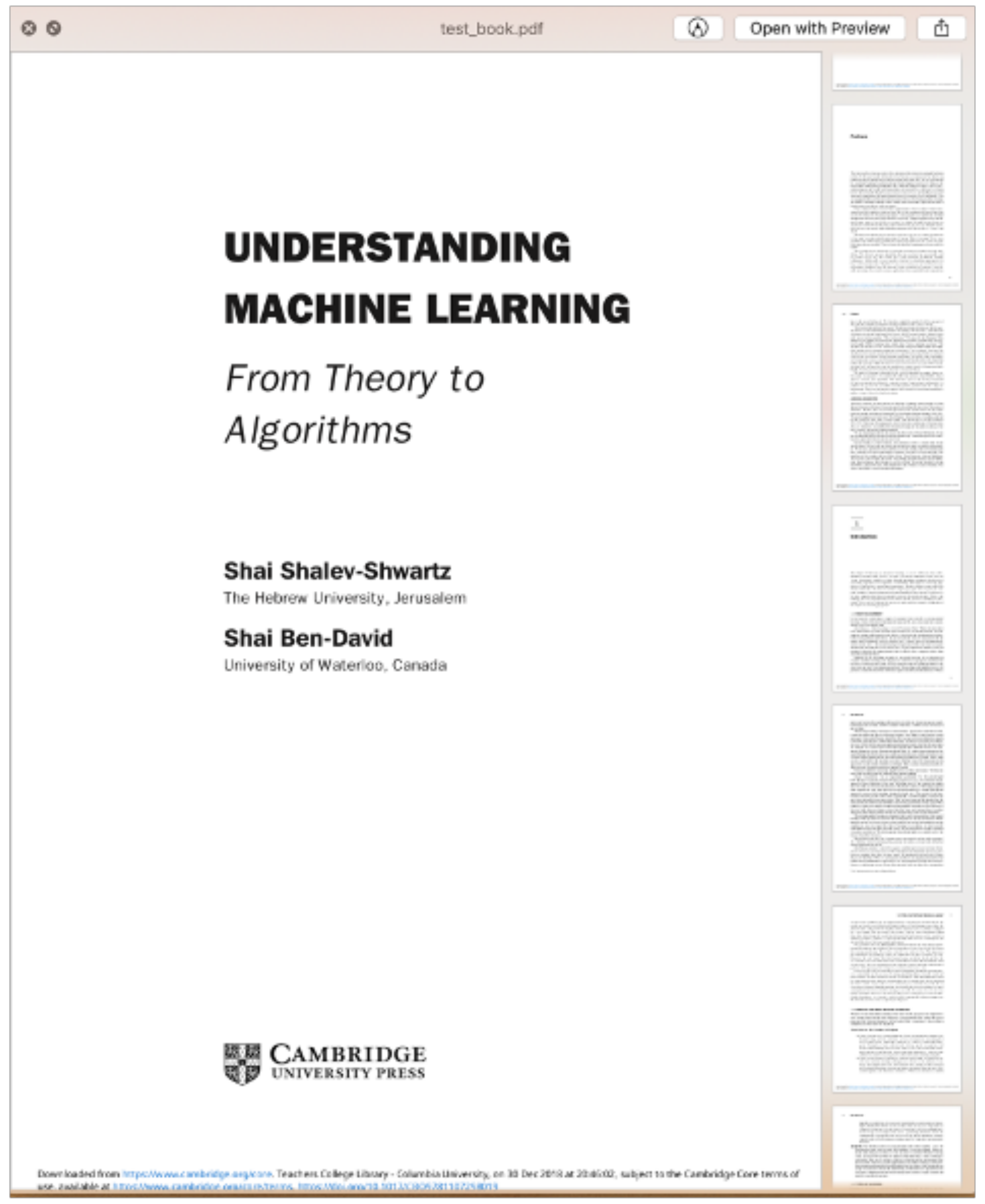
Automated Evasion Approach



- Building block operations
  - Feature **insertion-only** attacks.
  - **Mimicry**, merging with benign features.
  - **Mutation** operations (insert, replace, delete).
- Optimization: slowly change the input according to the prediction
  - Greedy
  - Genetic Evolution

Mimicry Attack, insertion only: Evading Gmail’s PDF Malware Classifier

Inserted /Root/Pages from  to  → → Benign



caa5b1d1d0f94a3f480b68dae40f473c193cbe7... (99K)
Virus detected! Help ×

test_book.pdf (8,174K) ×

The PDF is still malicious

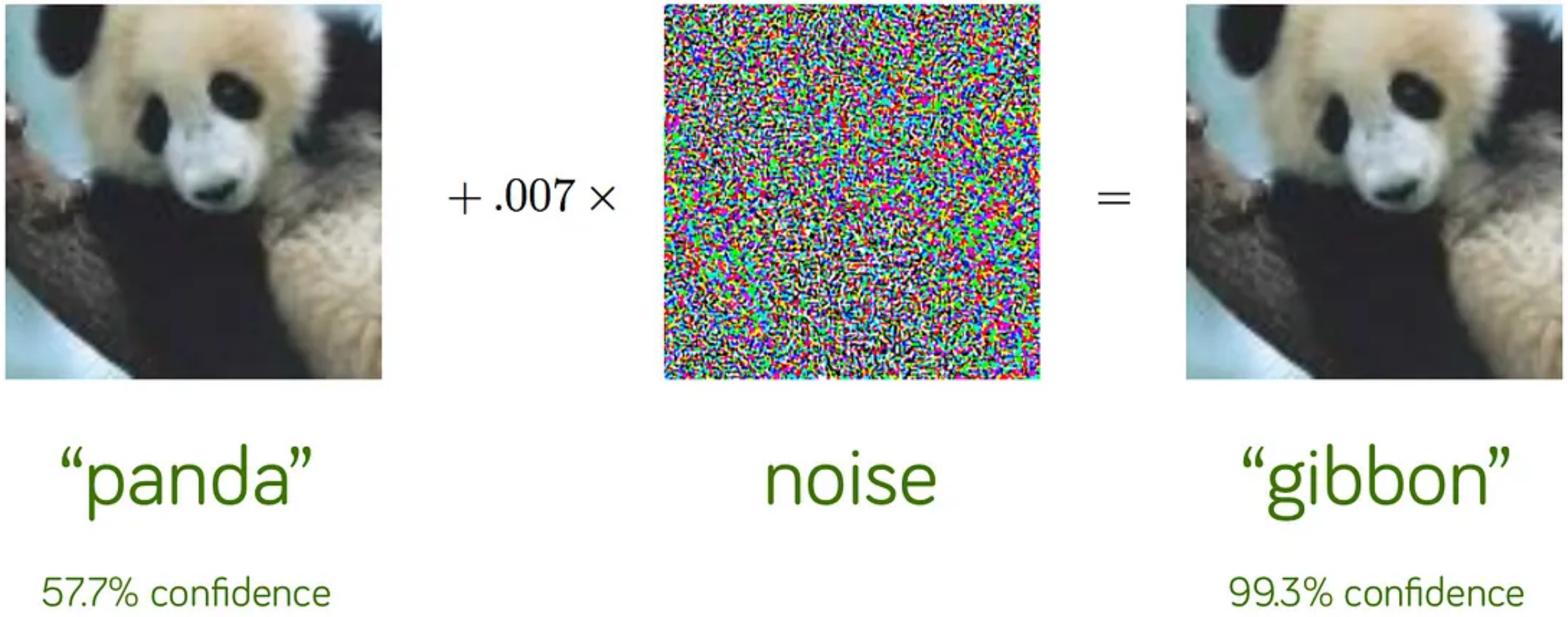
Attack worked in 2018

Adversarial Example

Domain	Classifier Space	"Reality" Space
Trojan Wars	Judgement of Trojans $f(x)$ = "gift"	Physical Reality $f^*(x)$ = invading army
Malware	Malware Detector $f(x)$ = "benign"	Victim's Execution $f^*(x)$ = malicious behavior
Image Classification	DNN Classifier $f(x) = t$	Human Perception $f^*(x) = c$

Is "Adversarial Examples" an Adversarial Example? Keynote talk at 1st Deep Learning and Security Workshop, 2018.

Image Classification: Adversarial Example



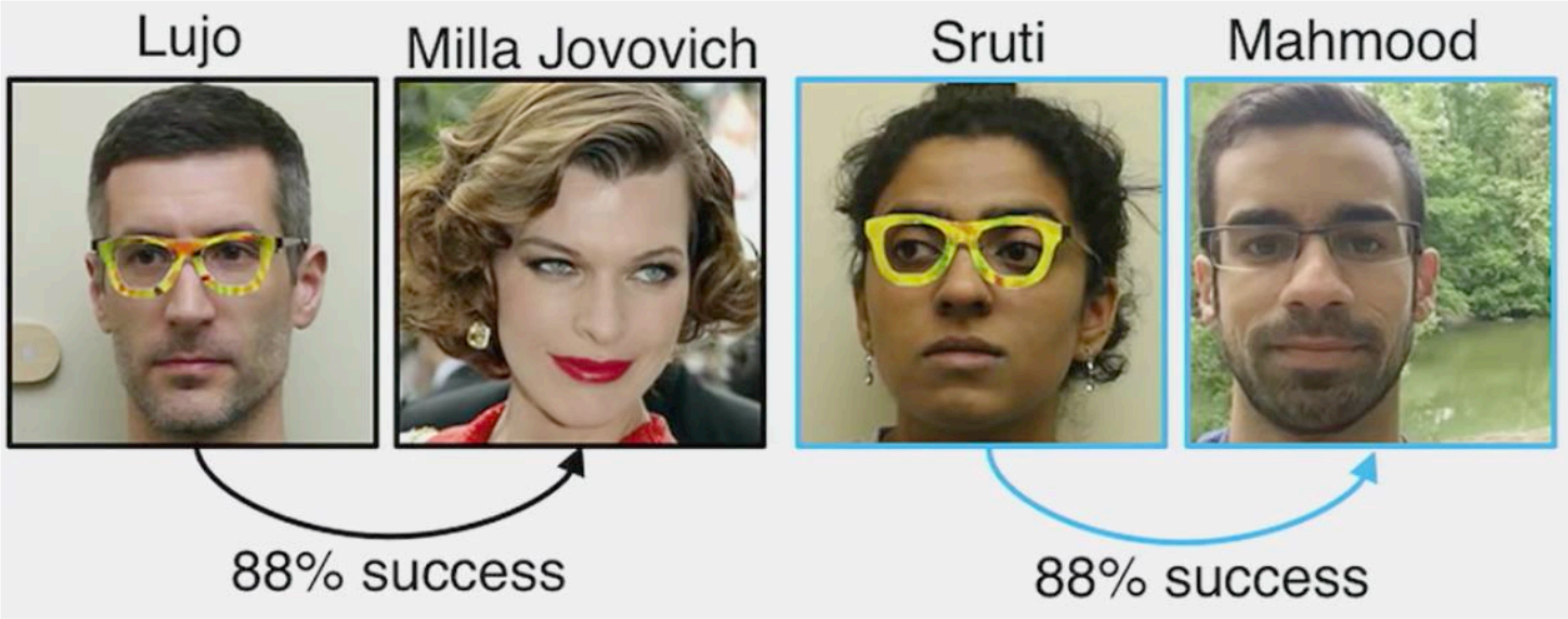
“Explaining and Harnessing Adversarial Examples”, Goodfellow et al, ICLR 2015.

Image Classification: Adversarial Example

- Given seed sample x, x' is an adversarial example iff:
  - $f(x') = t$ Class t is a wrong class, chosen t or arbitrary t
  - $B(x') = B(x)$ **Small imperceptible noise**

Adversarial example: looks the same to human, but classified differently by a neural network model

Evasion Attacks in the Physical World



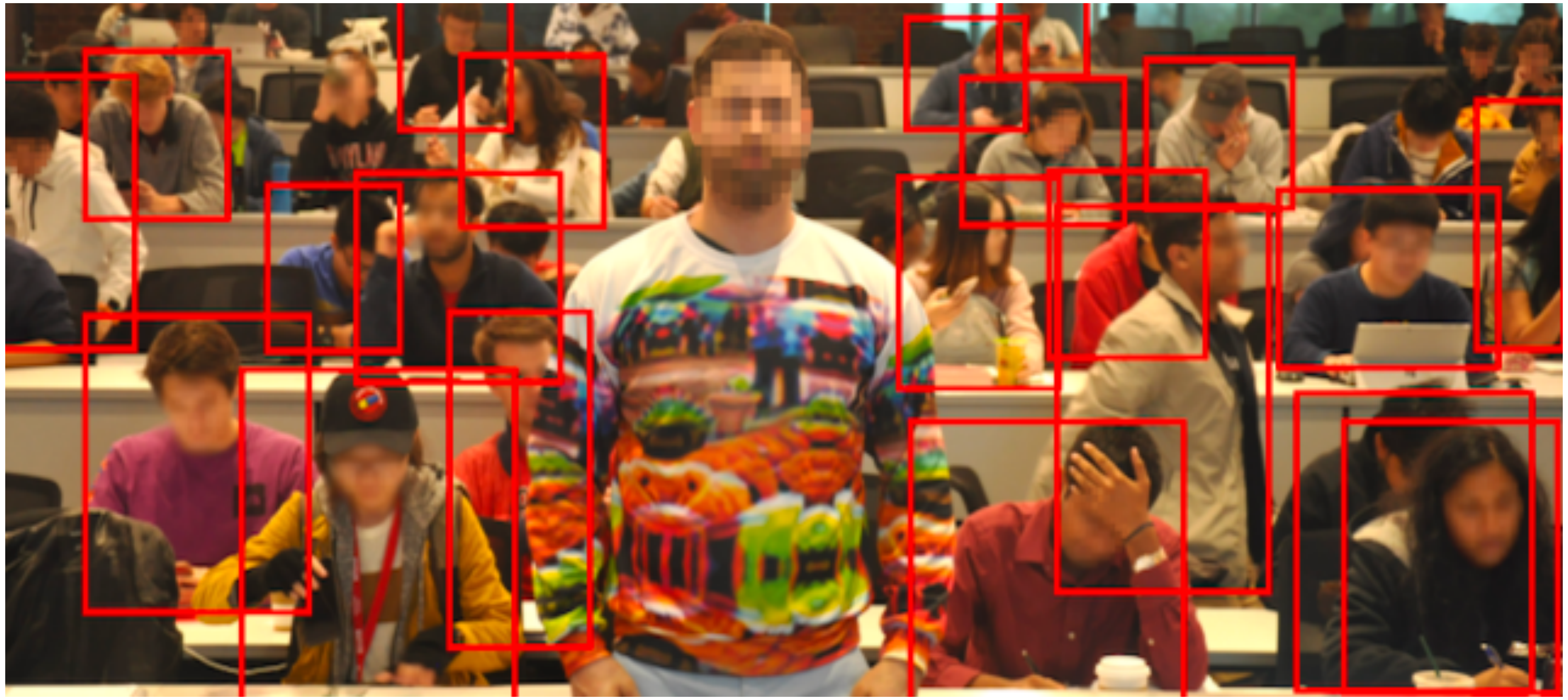
Sharif, Bhagavatula, Bauer, Reiter, Accessorize to a Crime: Real and Stealthy Attacks on State-Of-The-Art Face Recognition, CCS 2016

Evasion Attacks in the Physical World

Misclassified as Speed Limit 45 Sign

Eykholt et al., Robust Physical-World Attacks on Deep Learning Models, CVPR 2018

Evasion Attacks in the Physical World



Object detection: person disappears

"Making an Invisibility Cloak: Real World Adversarial Attacks on Object Detectors", Zuxuan et al, ECCV 2020

Neural Network Model Evasion Attack Idea

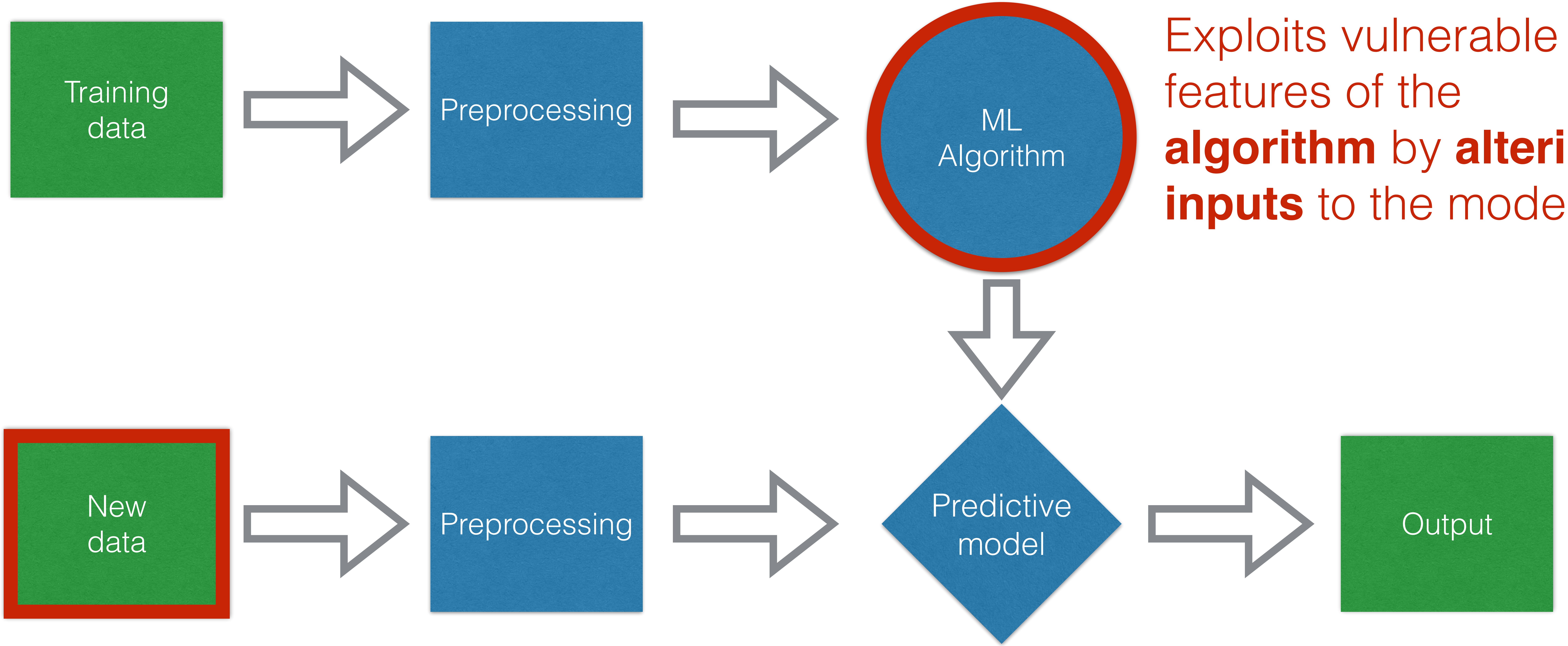
- $f_\theta(x) = \hat{y}, \quad$ i.e., $\; f(x, \theta) = \hat{y}$
  - Model f, parameters θ, input x, label y, predicted $\hat{y}$
  - The parameters θ and input x are **symmetric** to the Neural Network model
- Training: optimize θ, so we have small errors between $\hat{y}$ and y

Attack needs to change x that predicts differently

Neural Network Model Evasion Attack Idea

- $f_\theta(x) = \hat{y}$, i.e., $f(x, \theta) = \hat{y}$
  - Model f, parameters θ, input x, label y, predicted $\hat{y}$
  - The parameters θ and input x are **symmetric** to the Neural Network model
- Training: optimize θ, so we have small errors between $\hat{y}$ and y
- Evasion attack: optimize x, so we have small errors between $\hat{y}$ and a target class
  - Subject to small perturbation constraints

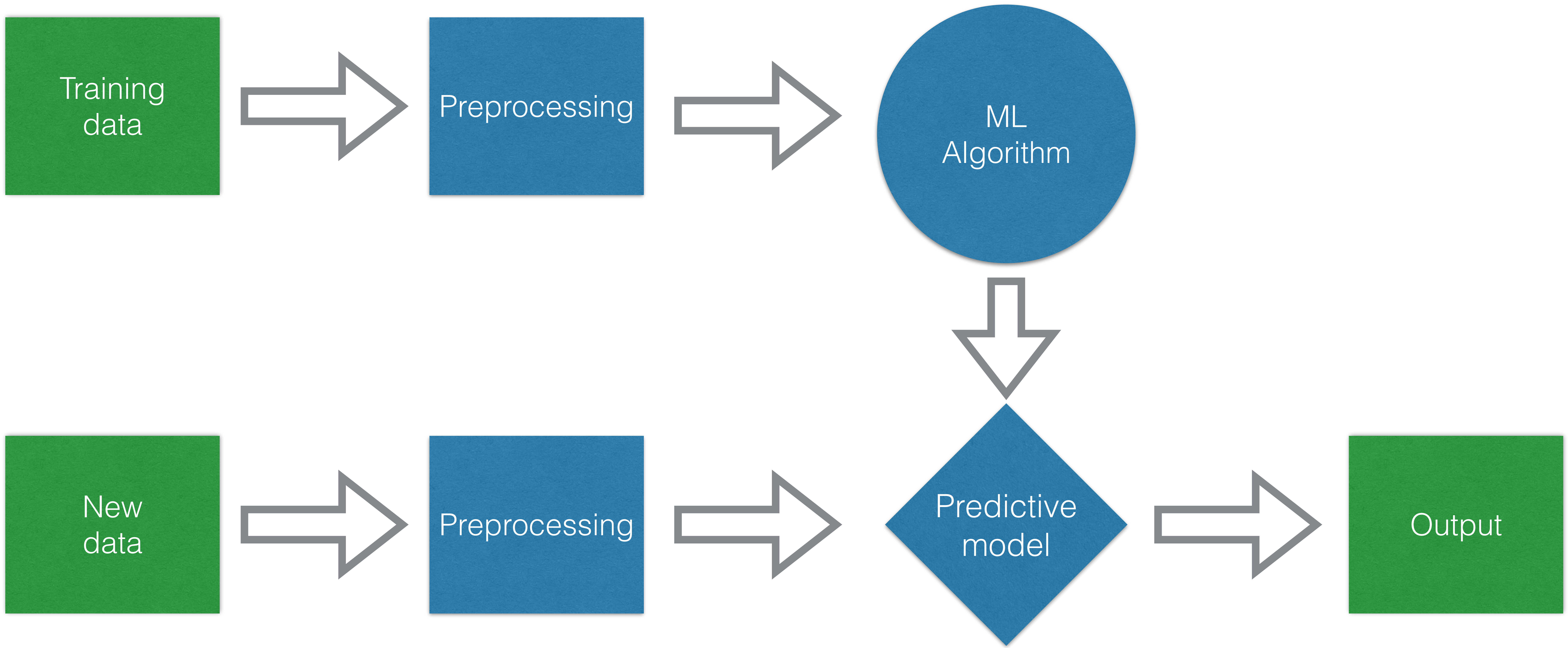
Evasion attacks



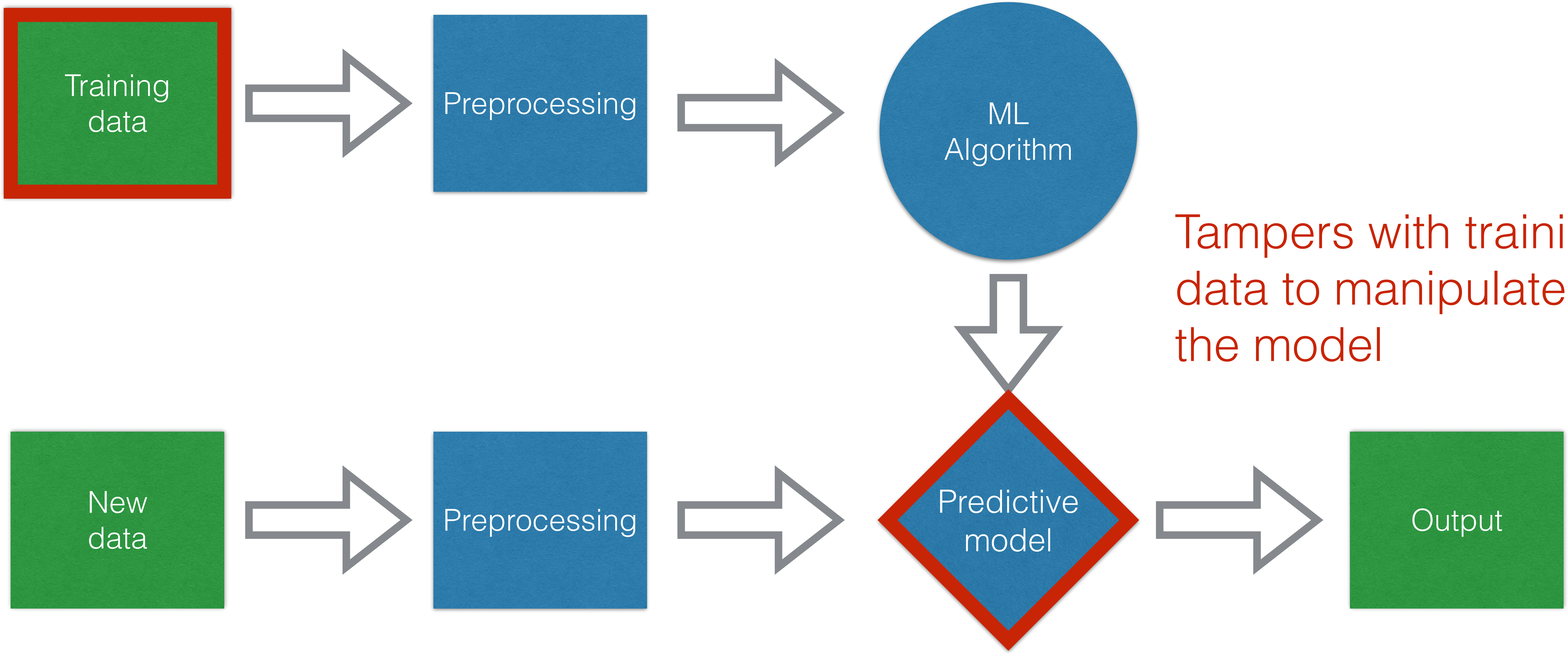
Exploits vulnerable features of the **algorithm** by **altering inputs** to the model

Threat model for attacks in ML

What else can the adversary attack?

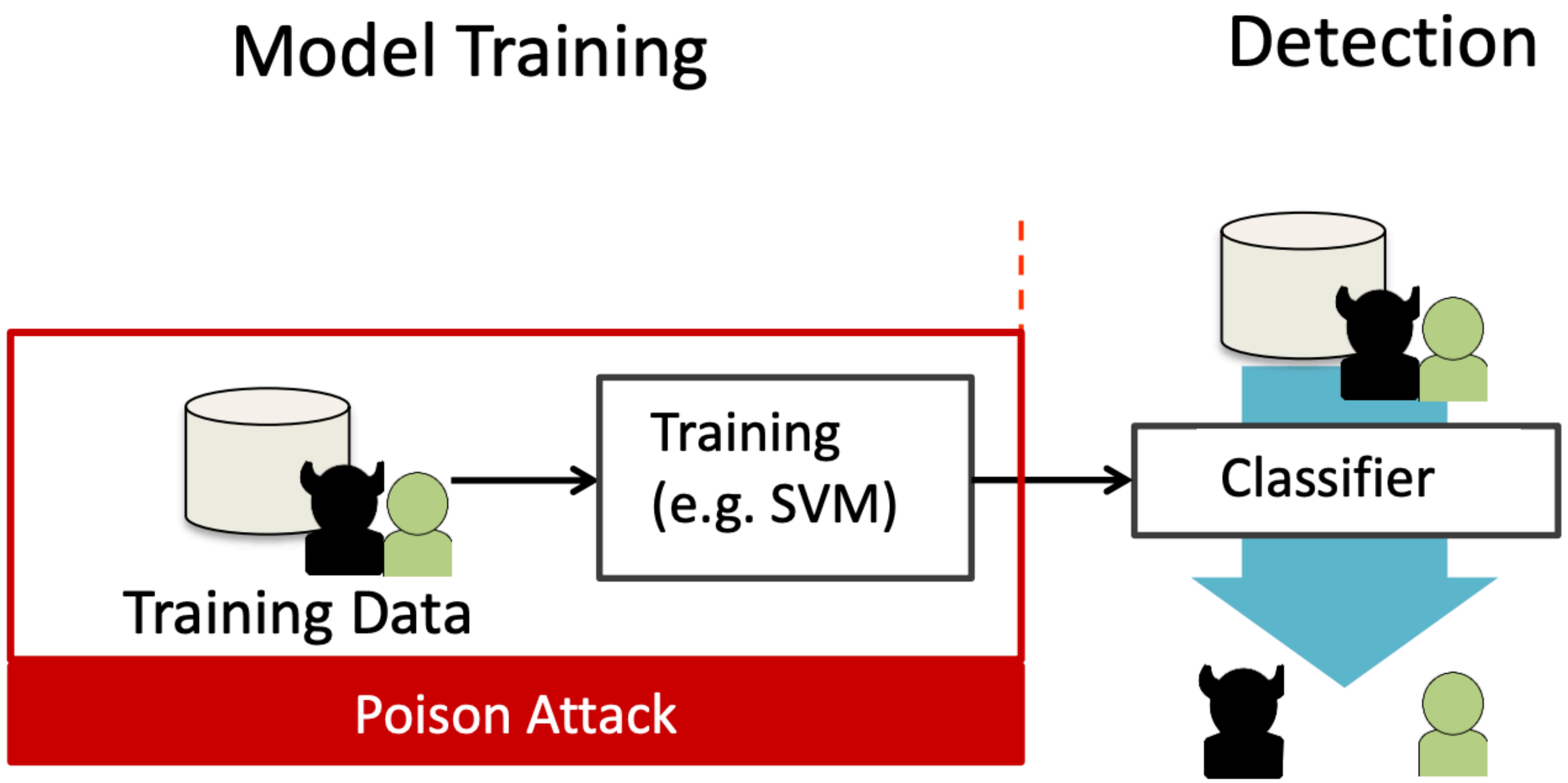


Poisoning attacks



Tampers with training data to manipulate the model

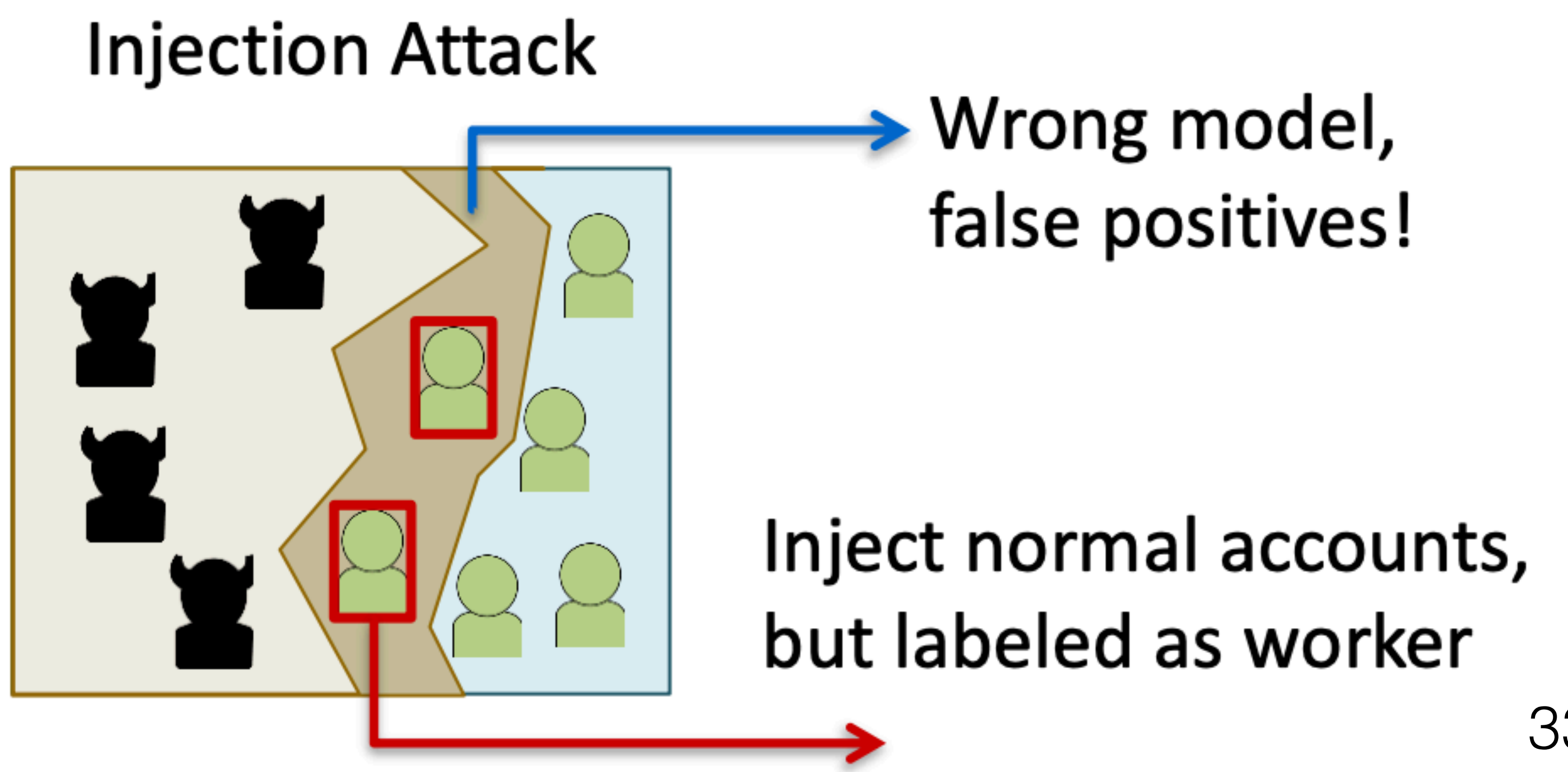
Poisoning attack

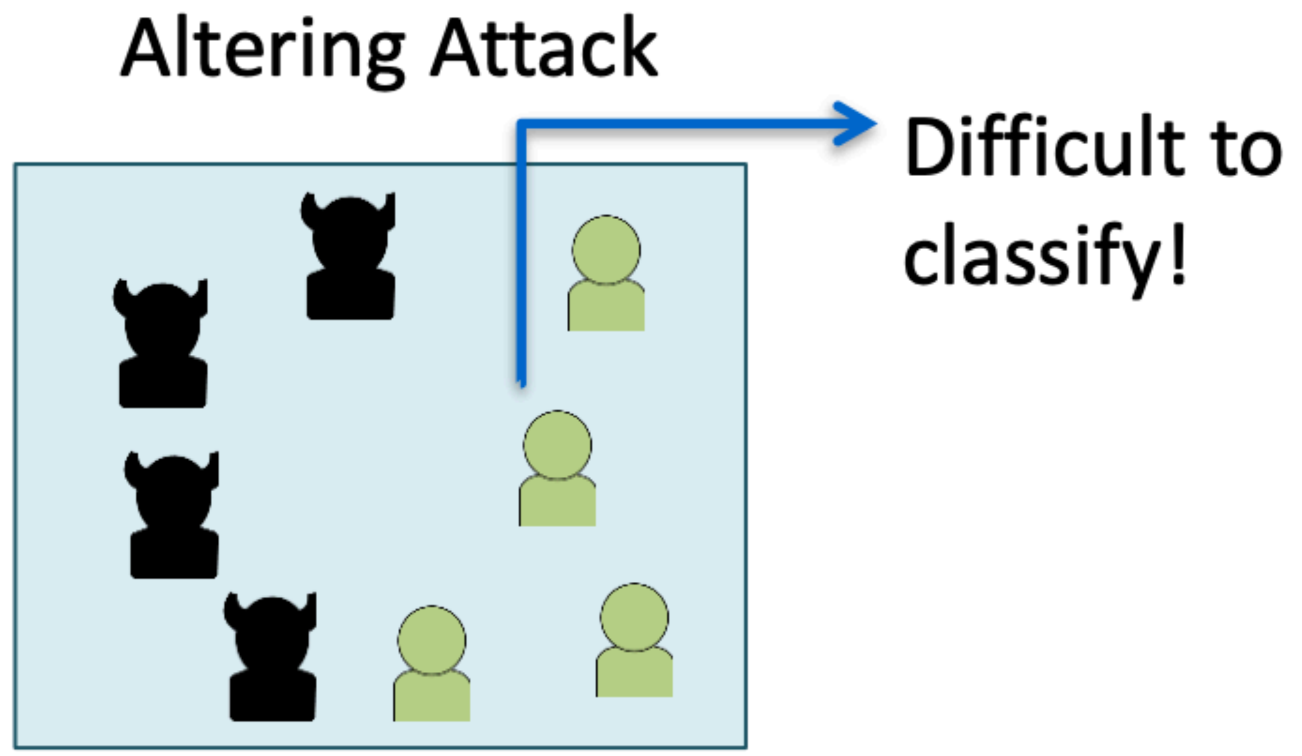


Poisoning attacks

- Tamper with training data to manipulate model

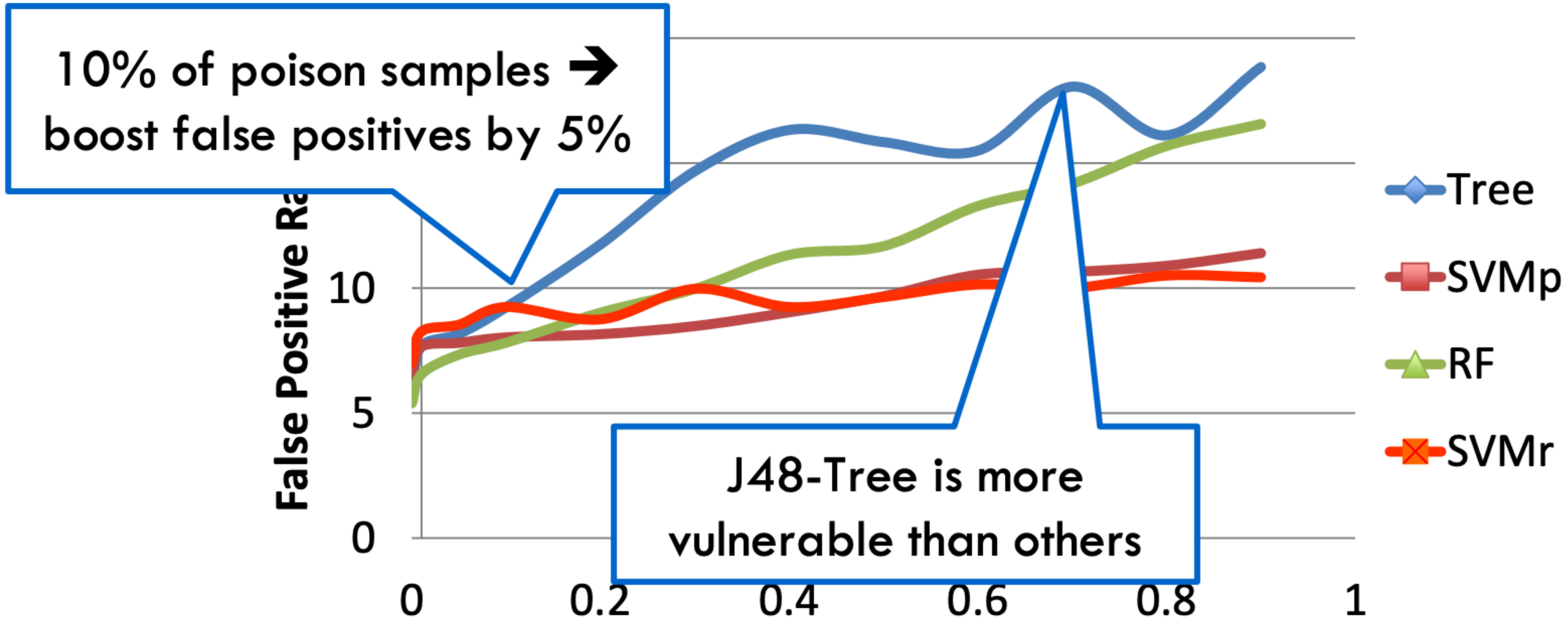
- Two practical poisoning methods:
  - **Inject** mislabeled samples to training data
    - ➔ wrong classifier
  - **Alter** worker behaviors uniformly by enforcing system policies
    - ➔ *harder to train accurate classifiers*





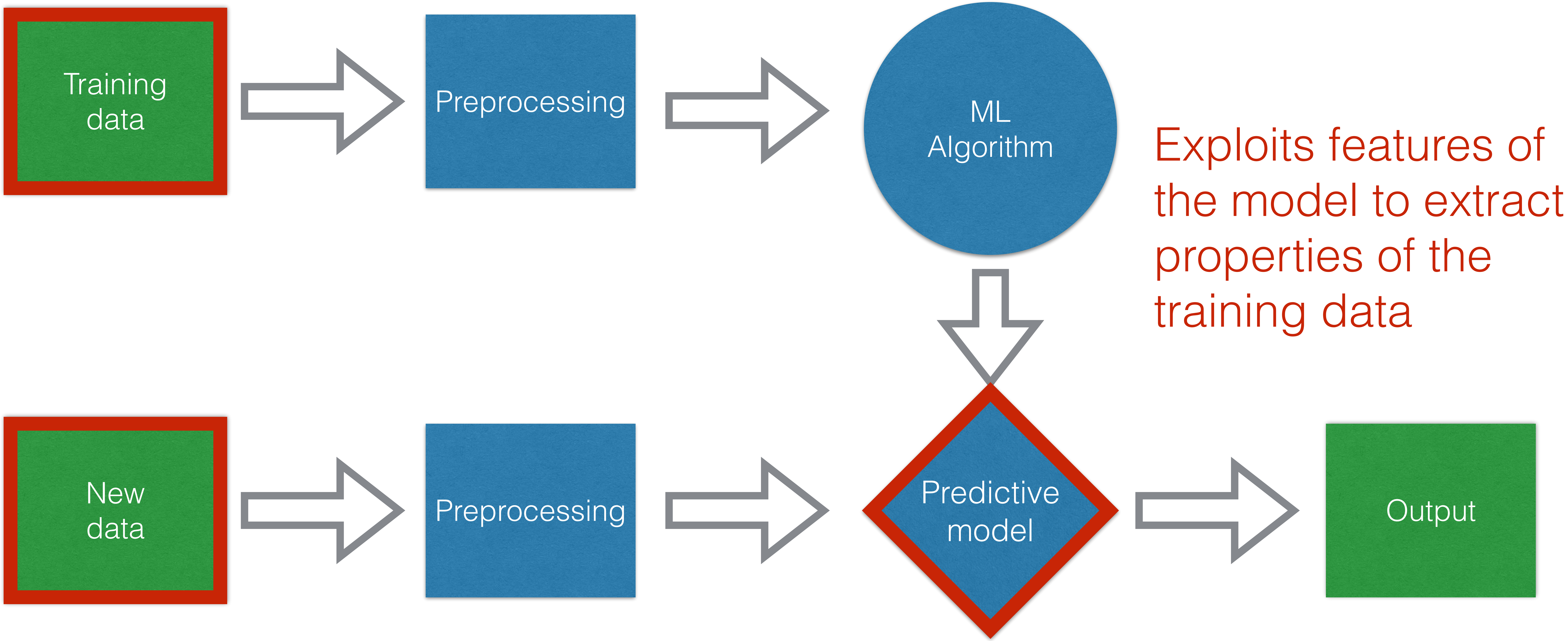
Injecting Poison Samples

- Injecting benign accounts as “workers” into training data
  - Aim to trigger false positives during detection



Poisoning attack is highly effective
More accurate classifiers often more vulnerable

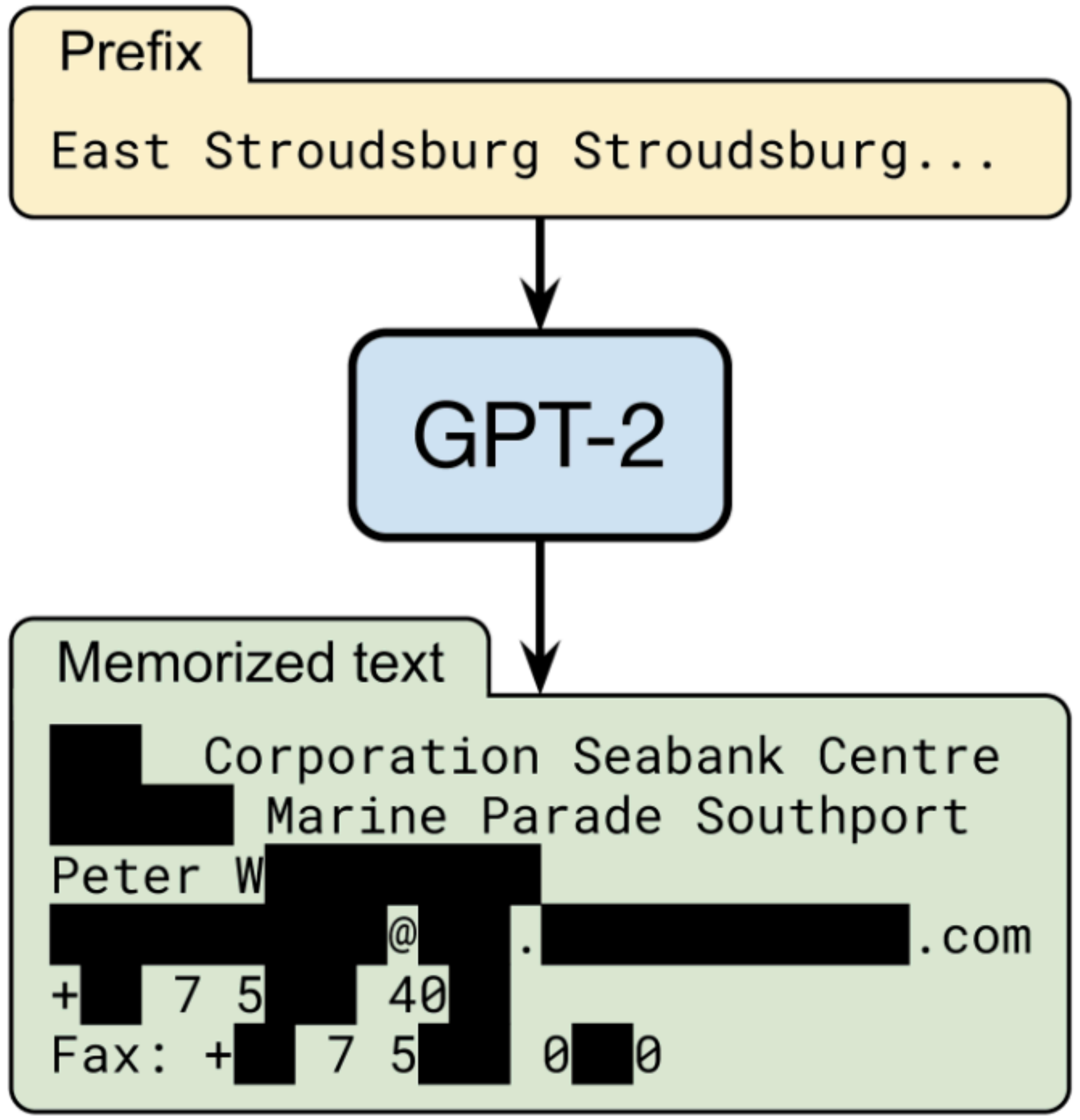
Model inversion attack



Model Inversion Attack

- Extract private and sensitive inputs by leveraging outputs and ML model

Figure 1: An image recovered using a new model inversion attack (left) and a training set image of the victim (right). The attacker is given only the person's name and access to a facial recognition system that returns a class confidence score.



https://bair.berkeley.edu/blog/2020/12/20/lmmem/

Model Extraction Attack

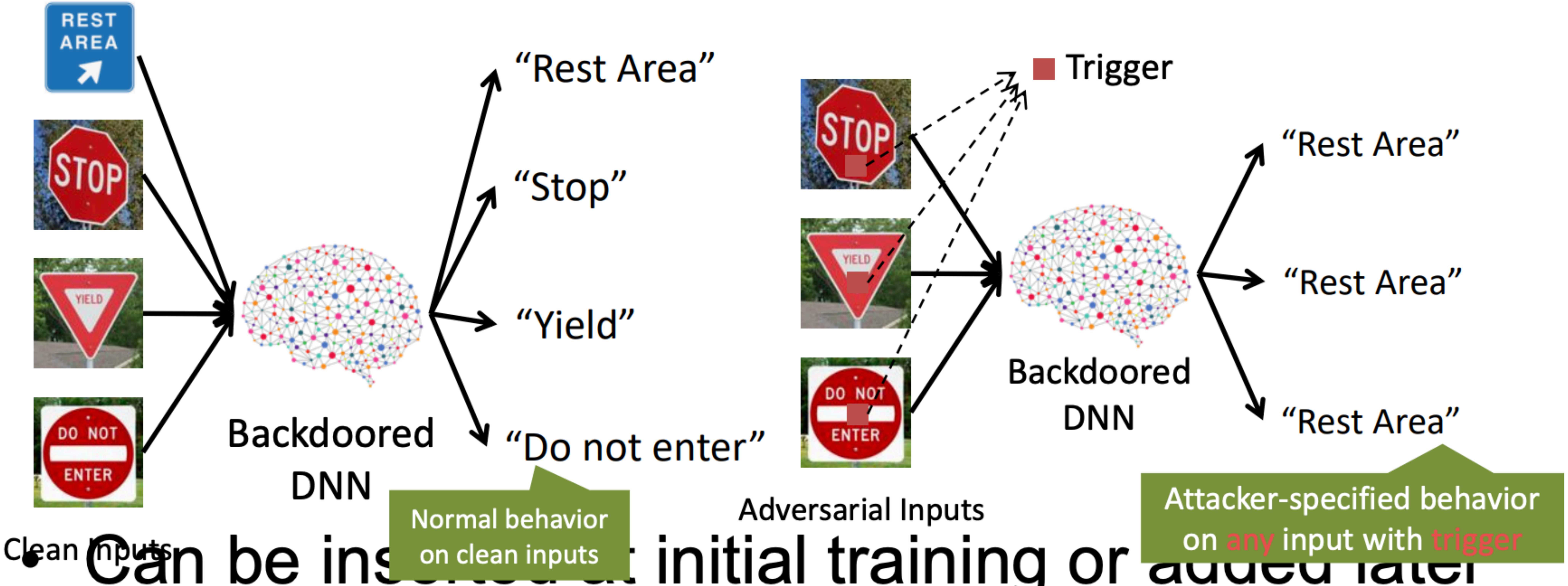
- **Extract model parameters** by querying model

Model	OHE	Binning	Queries	Time (s)	Price ($)
Circles	-	Yes	278	28	0.03
Digits	-	No	650	70	0.07
Iris	-	Yes	644	68	0.07
Adult	Yes	Yes	1,485	149	0.15

Table 7: **Results of model extraction attacks on Amazon.** OHE stands for one-hot-encoding. The reported query count is the number used to find quantile bins (at a granularity of 10^{-3}), plus those queries used for equation-solving. Amazon charges $0.0001 per prediction [1].

Backdoors

- Hidden behavior trained into a DNN



- Can be inserted at initial training or added later

Deepfakes

Deepfakes

The New York Times

Your Loved Ones, and Eerie Tom Cruise Videos, Reanimate Unease With Deepfakes

A tool that allows old photographs to be animated, and viral videos of a Tom Cruise impersonation, shined new light on digital impersonations.

39

A looping video of the Rev. Dr. Martin Luther King Jr. was created using a photograph and a tool on the MyHeritage genealogy site.

By Daniel Victor

March 10, 2021 Updated 1:07 p.m. ET

Deepfakes

- Content generation
- Video alterations
- Video/audio mimicry using LSTMs
  - e.g. Lyrebird.ai

Takeaway

If you use AI, there are new components in the system, so they allow more attacks…