How to Build Realistic Machine Learning Systems for Security?

Sadia Afroz
ICSI and Avast

Rajarshi Gupta
Avast

Machine Learning is necessary for detecting malware at scale

BREAKING NEWS

U.N. weathers storm of Emotet-TrickBot Malware

THE EXTORTION ECONOMY

Like Voldemort, Ransomware Is Too Scary to Be Named

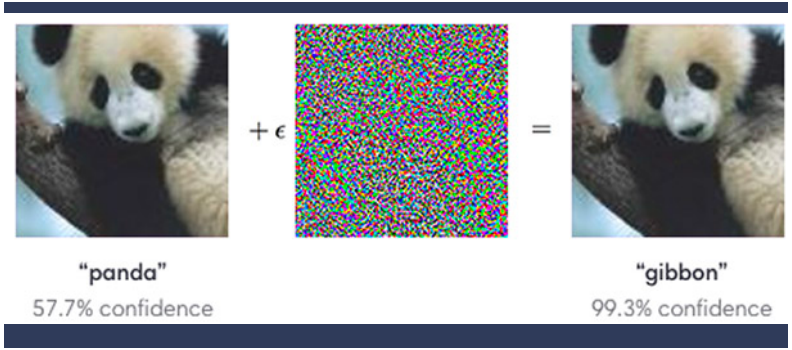
Wary of alarming investors, companies victimized by ransomware attacks often tell the SEC that "malware" or a "security incident" disrupted their operations.

A Cyberattack 'the World Isn't Ready For'

Mobile Banking Malware Up 50% in First Half of 2019

Every single Yahoo account was hacked - 3 billion in all

…but Machine Learning is unreliable, inexplicable and easily fooled

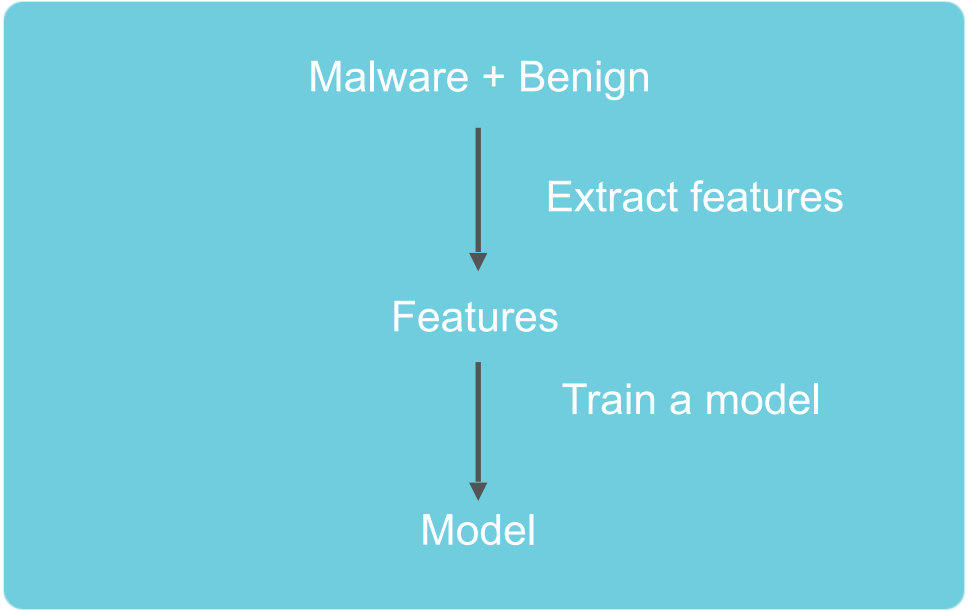


Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014).
Explaining and harnessing adversarial examples.
arXiv preprint arXiv:1412.6572.

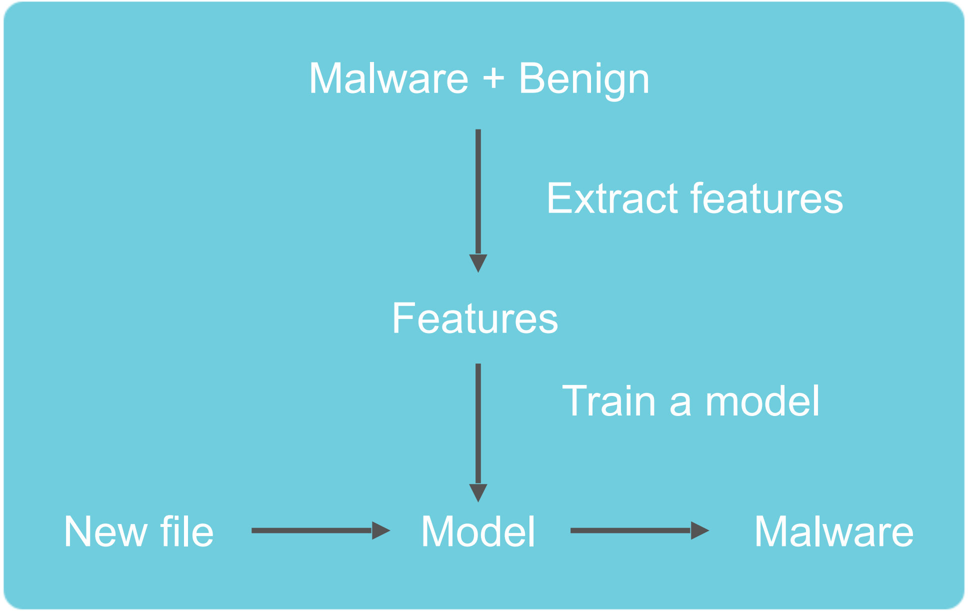
Evtimov, Ivan, et al. (2017).
”Robust physical-world attacks on deep learning models."
arXiv preprint arXiv:1707.08945.

Is machine learning useful for security?

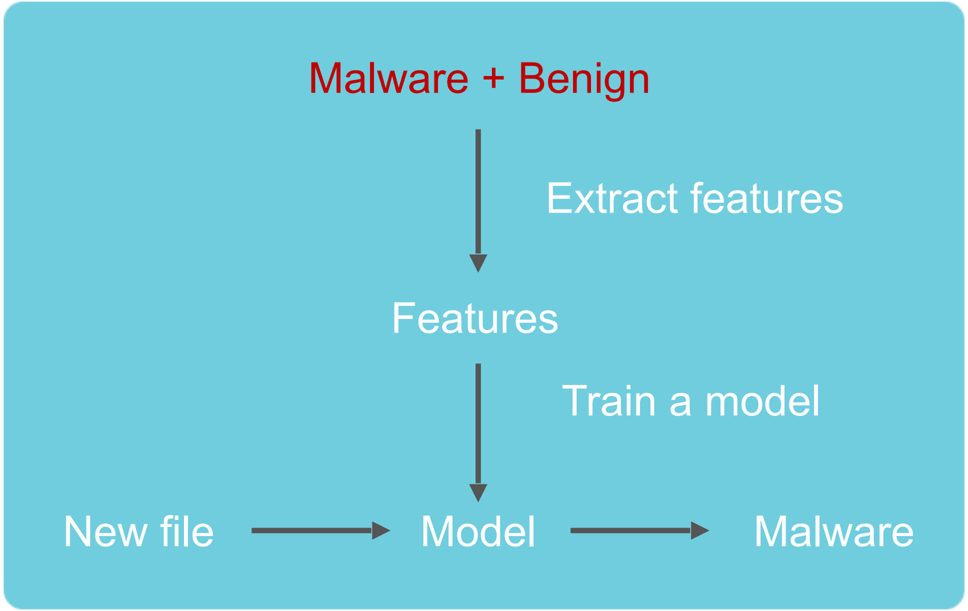
Let’s build a malware detector using machine learning



Let’s build a malware detector using machine learning



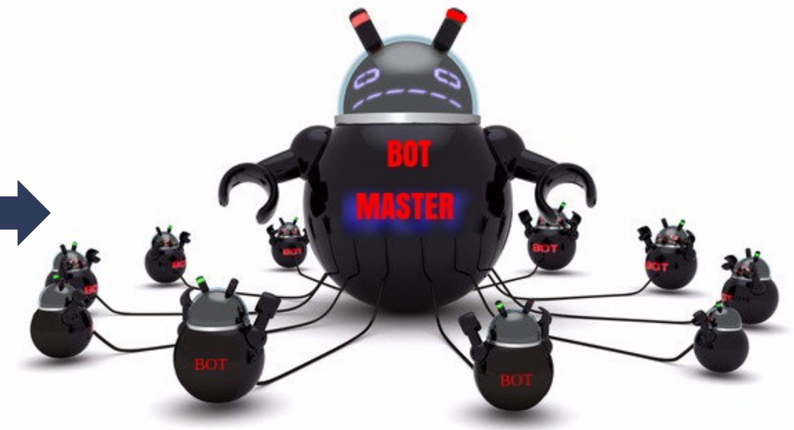
Let's build a malware detector using machine learning



Quality of the data ==> Quality of the model

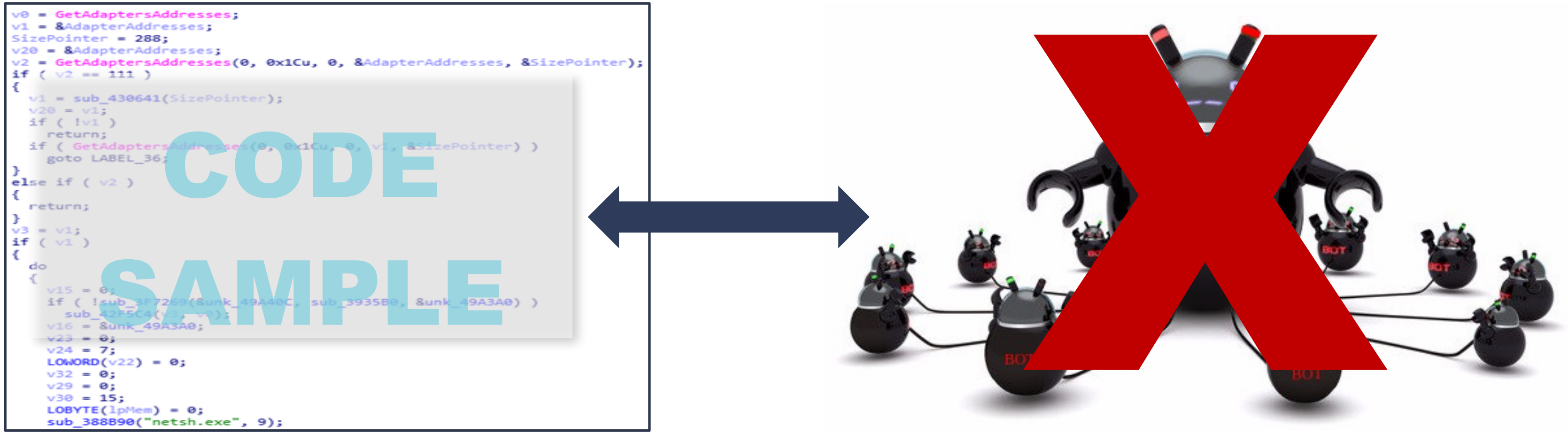

v0 = GetAdaptersAddresses;
v1 = &AdapterAddresses;
SizePointer = 288;
v20 = &AdapterAddresses;
v2 = GetAdaptersAddresses(0, 0x1Cu, 0, &AdapterAddresses, &SizePointer);
if (v2 == 111)
{
v1 = sub_430641(SizePointer);
v20 = v1;
if (!v1)
return;
goto LABEL_36;
}
else if (v2)
{
return;
}
v3 = v1;
if (v1)
{
do
{
v16 = &unk_49A3A0;
v23 = 0;
v24 = 7;
LOWORD(v22) = 0;
v32 = 0;
v29 = 0;
v30 = 15;
LOBYTE(lpMem) = 0;
sub_388B90("netsh.exe", 9);
CODE
SAMPLE


BOT
MASTER
BOT
BOT



Is this malware?


v0 = GetAdaptersAddresses;
v1 = &AdapterAddresses;
SizePointer = 288;
v20 = &AdapterAddresses;
v2 = GetAdaptersAddresses(0, 0x1Cu, 0, &AdapterAddresses, &SizePointer);
if (v2 == 111)
{
v1 = sub_430641(SizePointer);
v20 = v1;
if (!v1)
return;
if (GetAdaptersAddresses(0, 0x1Cu, 0, v1, &SizePointer))
goto LABEL_36;
}
else if (v2)
{
return;
}
v3 = v1;
if (v1)
{
do
{
v15 = 0;
if (!sub_3F7269(&unk_49A40C, sub_3935B0, &unk_49A3A0))
v16 = &unk_49A3A0;
v23 = 0;
v24 = 7;
LOWORD(v22) = 0;
v32 = 0;
v29 = 0;
v30 = 15;
LOBYTE(lpMem) = 0;
sub_388B90("netsh.exe", 9);
CODE SAMPLE

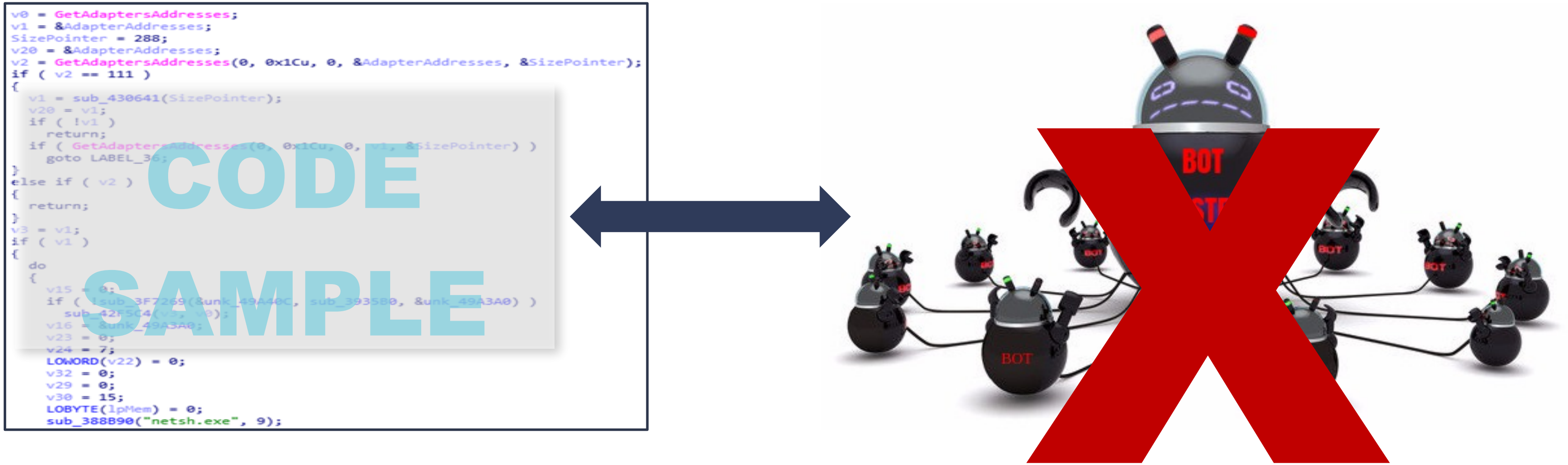


Is this malware?

Is this malware?

The answer depends on **WHO** you ask and **WHEN** you ask

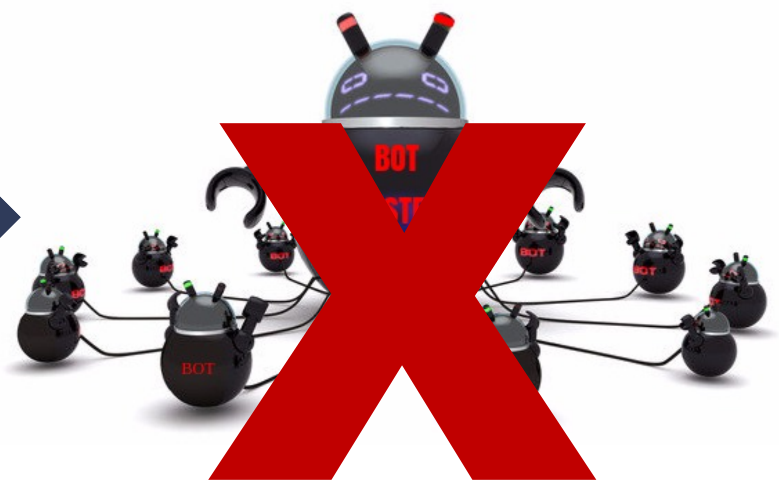
According to VirusTotal…



https://www.virustotal.com/gui/file/3120b563781b5ead9fdebc906818836329f362bf8e3ea7ee3dbfd4ceb0ebd8dd/detection

According to VirusTotal…

Sep 2019

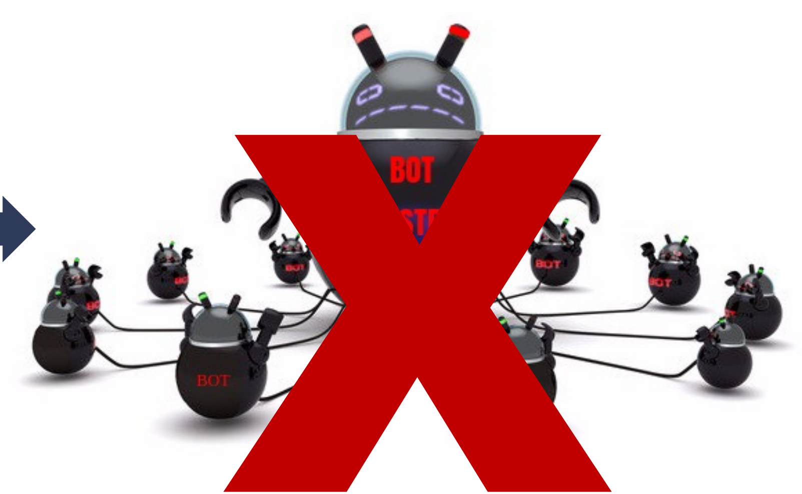


https://www.virustotal.com/gui/file/3120b563781b5ead9fdebc906818836329f362bf8e3ea7ee3dbfd4ceb0ebd8dd/detection

According to VirusTotal…

Sep 2019

~42% AVs
considered it malware



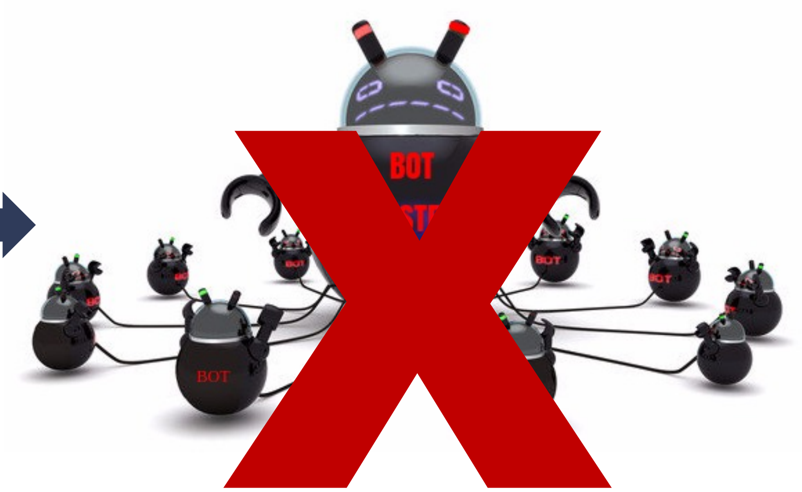
https://www.virustotal.com/gui/file/3120b563781b5ead9fdebc906818836329f362bf8e3ea7ee3dbfd4ceb0ebd8dd/detection

According to VirusTotal…

Sep 2019

~42% AVs
considered it malware

Jan 2020



https://www.virustotal.com/gui/file/3120b563781b5ead9fdebc906818836329f362bf8e3ea7ee3dbfd4ceb0ebd8dd/detection

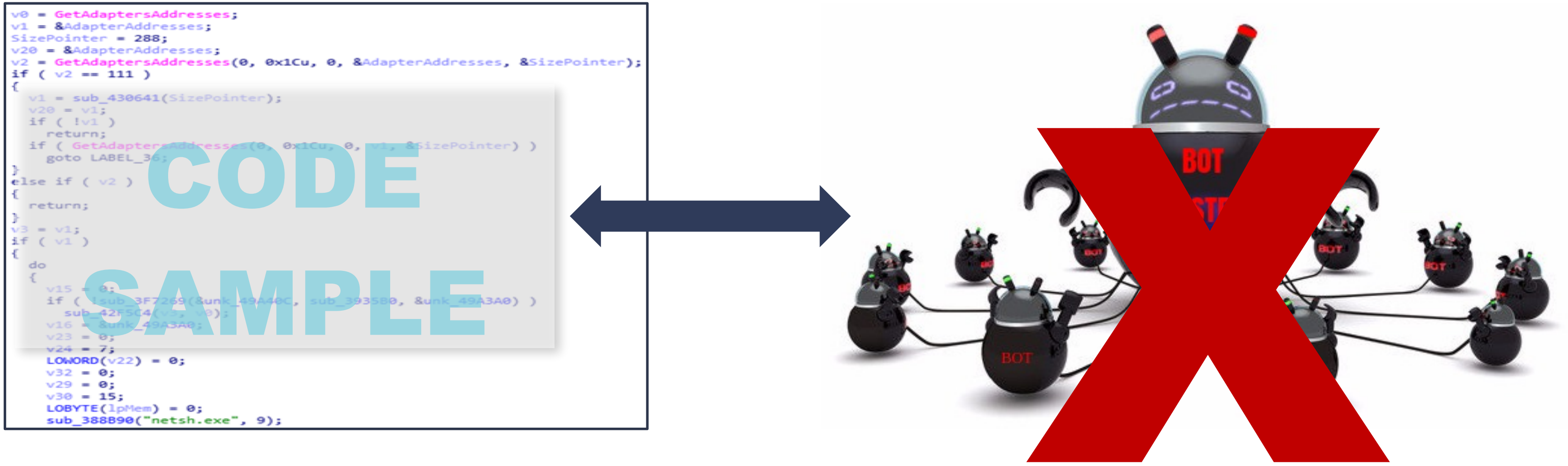
According to VirusTotal…

Sep 2019

~42% AVs
considered it malware

Jan 2020

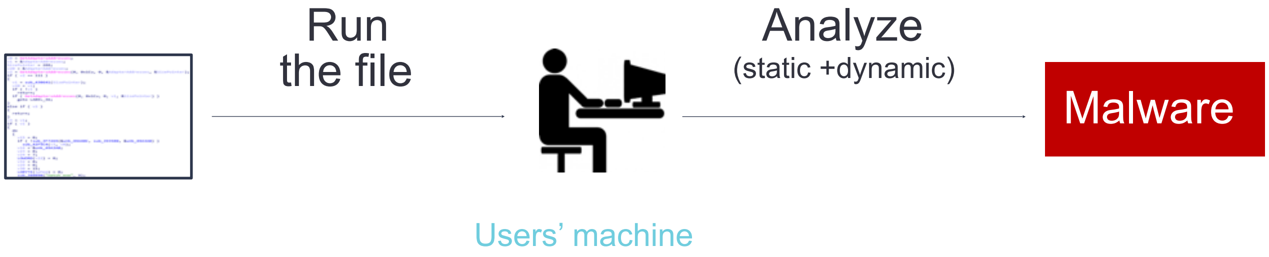
~72% AVs
considered it malware



https://www.virustotal.com/gui/file/3120b563781b5ead9fdebc906818836329f362bf8e3ea7ee3dbfd4ceb0ebd8dd/detection

How can we protect users from malware when we don’t know what malware is?

What is malware?



What is malware?

Run
the file

Virtual machine

Analyze
(static +dynamic)

Malware

What is malware?

Run the file

cuckoo

Analyze (static +dynamic)

Malware

Sandbox

What is malware?

Run
the file

Analyze
(static +dynamic)

Malware

Sandbox

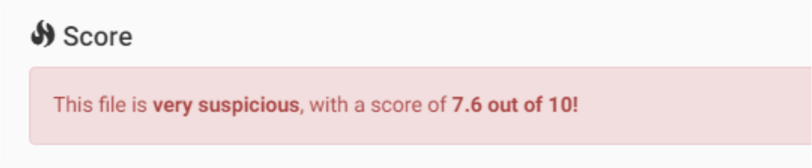
What is malware?

Run
the file

Analyze
(static +dynamic)

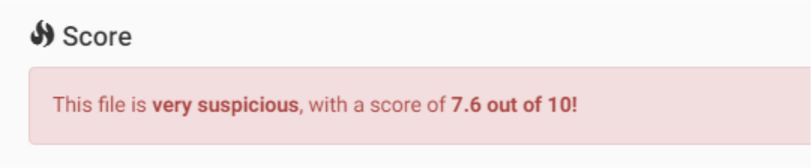
Malware

Sandbox



Malware is highly suspicious files

What is malware?



Malware is highly suspicious files

Too time consuming!

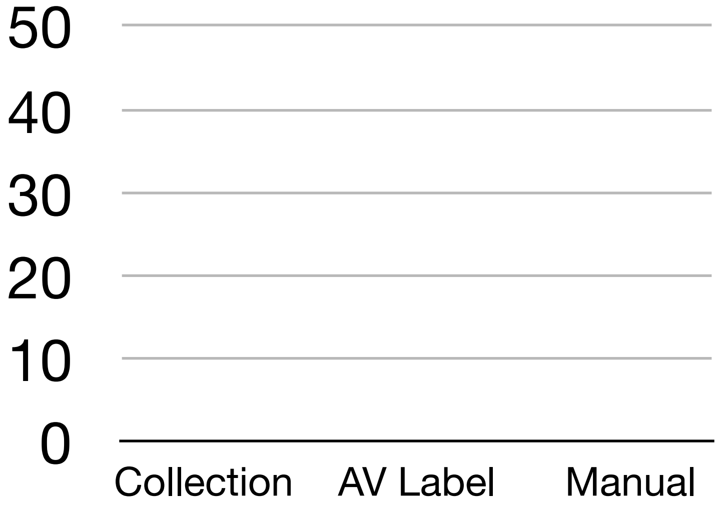
What is malware?

Solution: Get labels from other sources

We studied 40 papers from 2001-2019 to check where they get their ground truth from

What is malware?

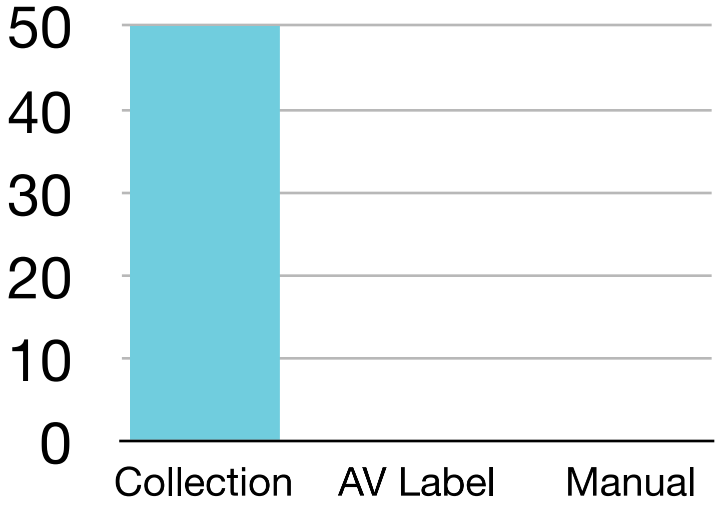
Solution: Get labels from other sources



We studied 40 papers from 2001-2019 to check where they get their ground truth from

What is malware?

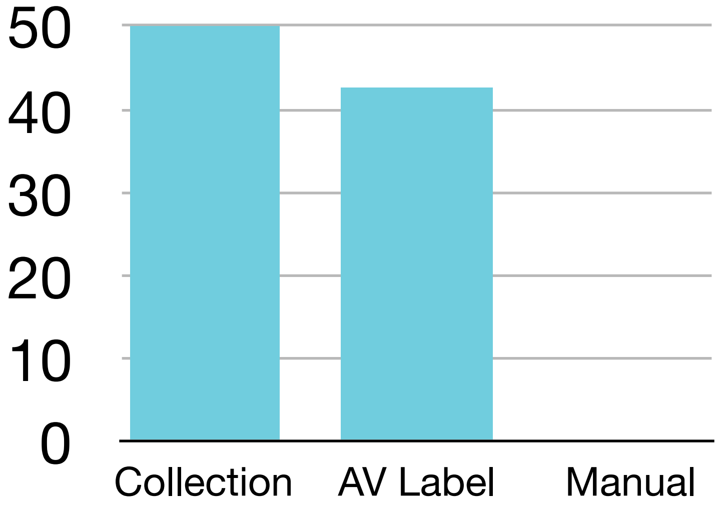
Solution: Get labels from other sources



We studied 40 papers from 2001-2019 to check where they get their ground truth from

What is malware?

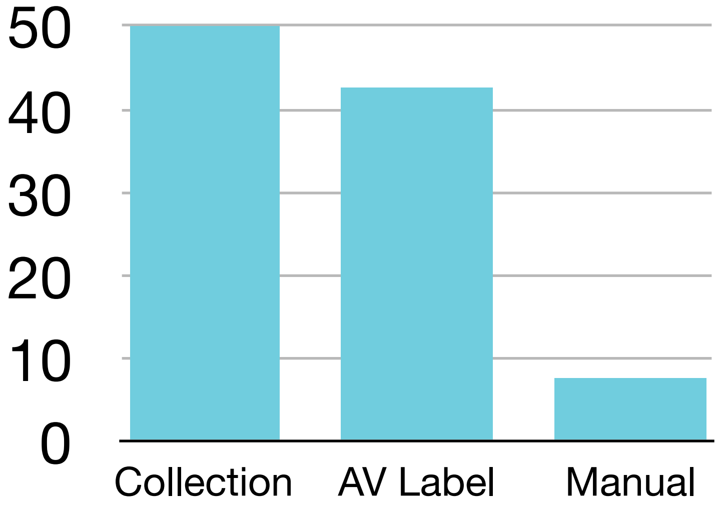
Solution: Get labels from other sources



We studied 40 papers from 2001-2019 to check where they get their ground truth from

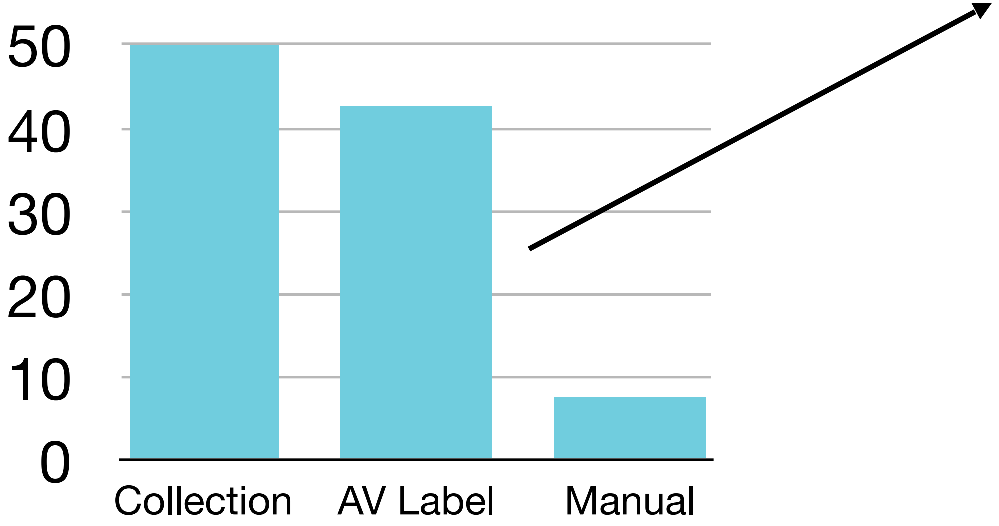
What is malware?

Solution: Get labels from other sources



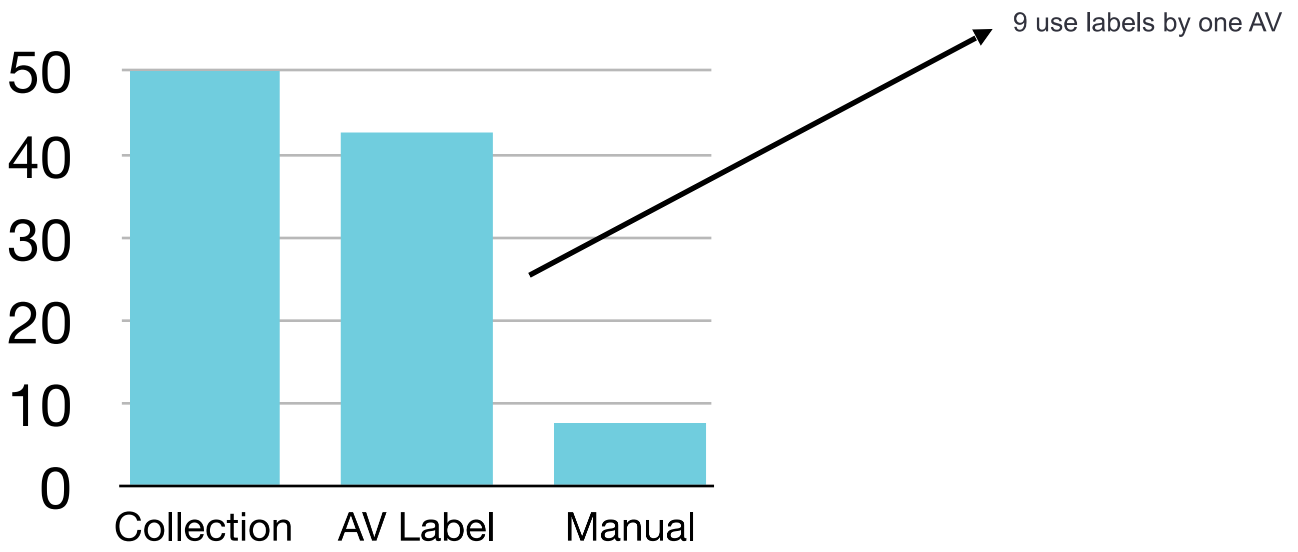
We studied 40 papers from 2001-2019 to check where they get their ground truth from

What is malware?



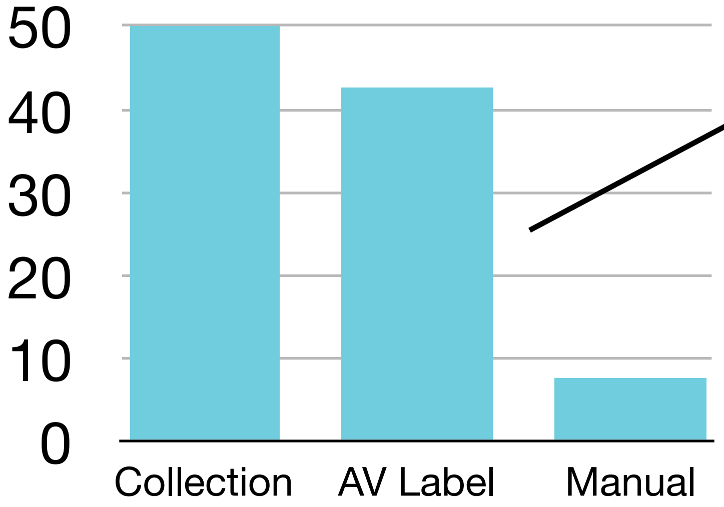
We studied 40 papers from 2001-2019 to check where they get their ground truth from

What is malware?



We studied 40 papers from 2001-2019 to check where they get their ground truth from

What is malware?

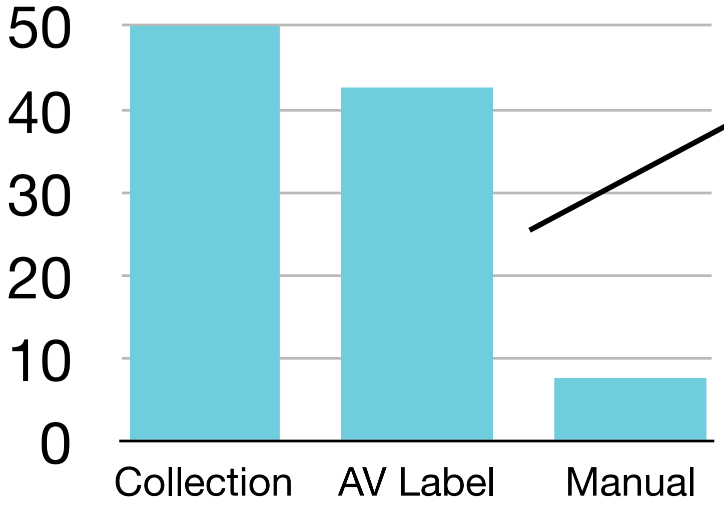


9 use labels by one AV

2 papers:
Malware >=4, Benign == 0

We studied 40 papers from 2001-2019 to check where they get their ground truth from

What is malware?



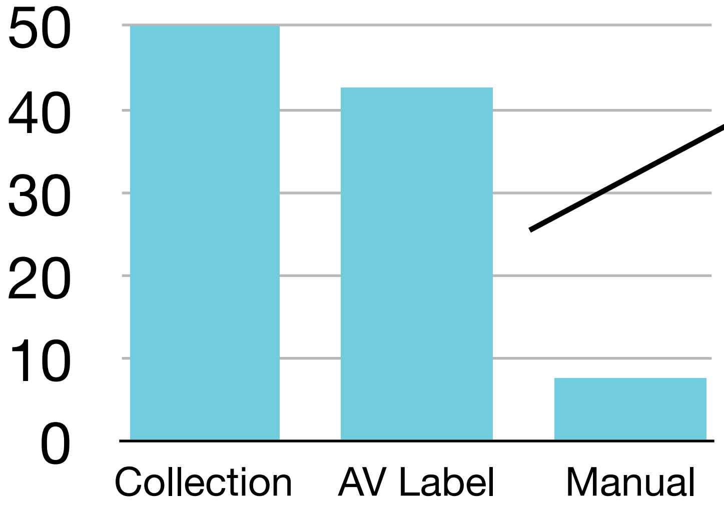
9 use labels by one AV

2 papers:
Malware >=4, Benign == 0

2 papers:
Malware >=5, Benign <=1

We studied 40 papers from 2001-2019 to check where they get their ground truth from

What is malware?



9 use labels by one AV

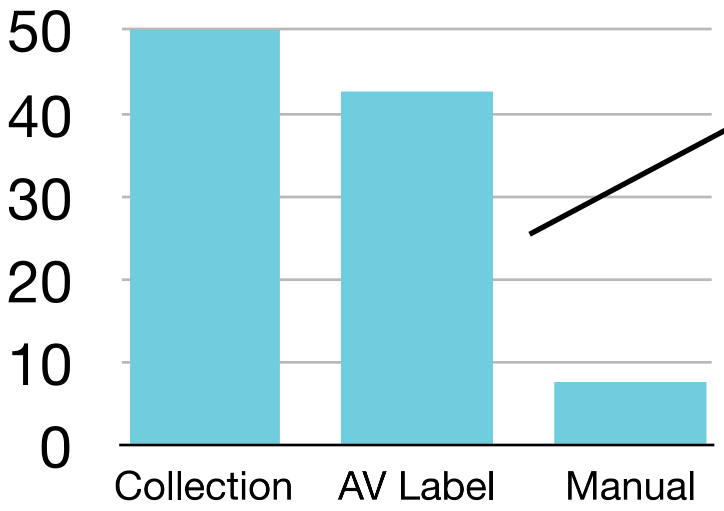
2 papers:
Malware >=4, Benign == 0

2 papers:
Malware >=5, Benign <=1

1 paper:
Malware >=10, Benign == 0

We studied 40 papers from 2001-2019 to check where they get their ground truth from

What is malware?



We studied 40 papers from 2001-2019 to check where they get their ground truth from

9 use labels by one AV

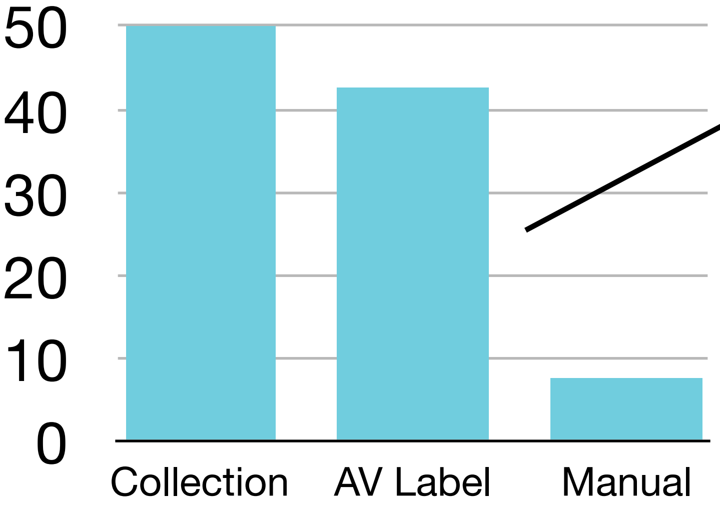
2 papers:
Malware >=4, Benign == 0

2 papers:
Malware >=5, Benign <=1

1 paper:
Malware >=10, Benign == 0

1 paper:
Malware == ALL, Benign == 0

What is malware?



We studied 40 papers from 2001-2019 to check where they get their ground truth from

9 use labels by one AV

2 papers:
Malware >=4, Benign == 0

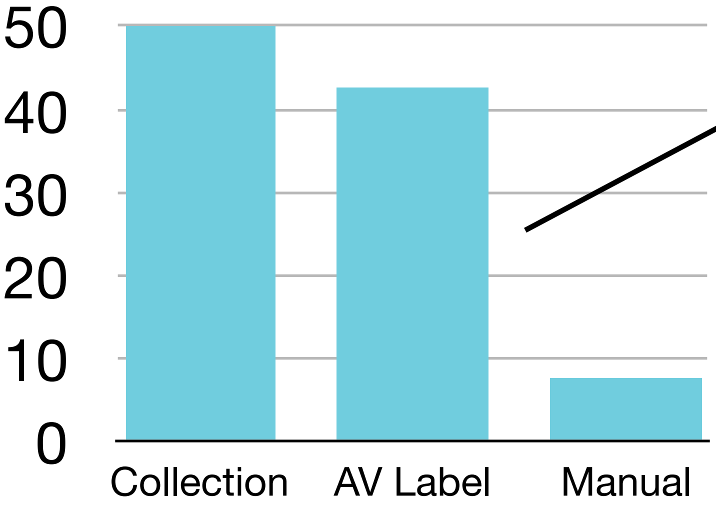
2 papers:
Malware >=5, Benign <=1

1 paper:
Malware >=10, Benign == 0

1 paper:
Malware == ALL, Benign == 0

1 paper:
Malware == Majority, Benign == 0

What is malware?



We studied 40 papers from 2001-2019 to check where they get their ground truth from

9 use labels by one AV

2 papers:
Malware >=4, Benign == 0

2 papers:
Malware >=5, Benign <=1

1 paper:
Malware >=10, Benign == 0

1 paper:
Malware == ALL, Benign == 0

1 paper:
Malware == Majority, Benign == 0

1 paper:
Malware == Weighted Majority, Benign == 0

How to compare different approaches?

What is malware?

AISec 2015

Better Malware Ground Truth:
Techniques for Weighting Anti-Virus Vendor Labels

Alex Kantchelian
UC Berkeley

Michael Carl Tschantz
International Computer Science Institute

Sadia Afroz
UC Berkeley

Brad Miller
UC Berkeley

Vaishaal Shankar
UC Berkeley

Rekha Bachwani
Netflix*

Anthony D. Joseph
UC Berkeley

J. D. Tygar
UC Berkeley

What is malware?

AISec 2015

Better Malware Ground Truth: Techniques for Weighting Anti-Virus Vendor Labels

Alex Kantchelian
UC Berkeley

Michael Carl Tschantz
International Computer Science Institute

Sadia Afroz
UC Berkeley

Brad Miller
UC Berkeley

Vaishaal Shankar
UC Berkeley

Rekha Bachwani
Netflix*

Anthony D. Joseph
UC Berkeley

J. D. Tygar
UC Berkeley

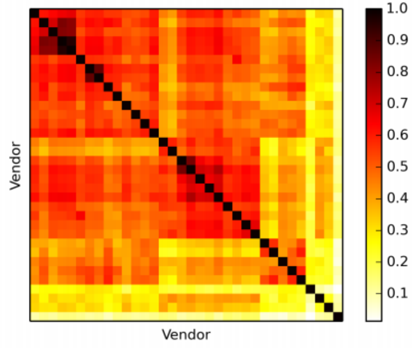


Figure 3: Label correlations between vendors.

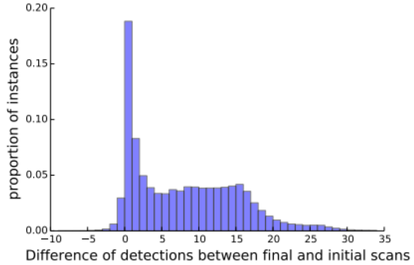
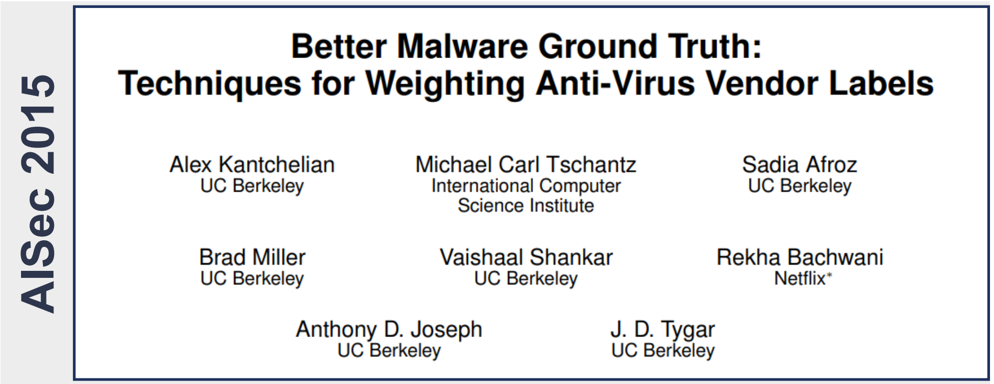


Figure 4: Difference in the number of positive detections between the first and last scans for each instance.

What is malware?

- Number of very large and professional companies share their labels on VirusTotal



AISec 2015

Better Malware Ground Truth: Techniques for Weighting Anti-Virus Vendor Labels

Alex Kantchelian, UC Berkeley
Michael Carl Tschantz, International Computer Science Institute
Sadia Afroz, UC Berkeley
Brad Miller, UC Berkeley
Vaishaal Shankar, UC Berkeley
Rekha Bachwani, Netflix*
Anthony D. Joseph, UC Berkeley
J. D. Tygar, UC Berkeley

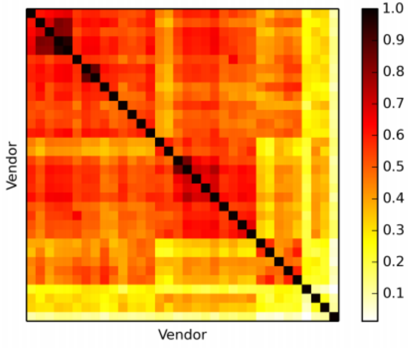


Figure 3: Label correlations between vendors.

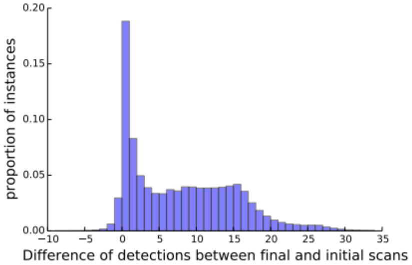


Figure 4: Difference in the number of positive detections between the first and last scans for each instance.

What is malware?

- Number of very large and professional companies share their labels on VirusTotal
- Great correlation in general, especially for top companies
  - 96% agreement after 3 days
  - 99% agreement after 3 weeks

AISec 2015

**Better Malware Ground Truth:
Techniques for Weighting Anti-Virus Vendor Labels**

Alex Kantchelian
UC Berkeley

Michael Carl Tschantz
International Computer Science Institute

Sadia Afroz
UC Berkeley

Brad Miller
UC Berkeley

Vaishaal Shankar
UC Berkeley

Rekha Bachwani
Netflix*

Anthony D. Joseph
UC Berkeley

J. D. Tygar
UC Berkeley

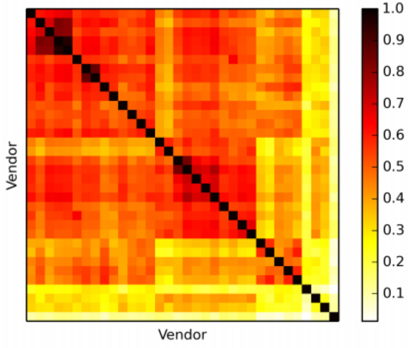


Figure 3: Label correlations between vendors.

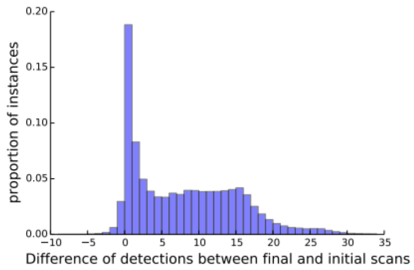
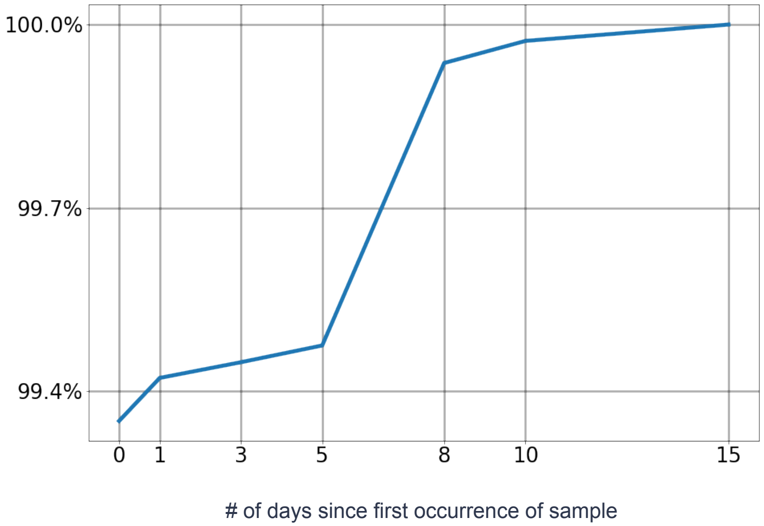


Figure 4: Difference in the number of positive detections between the first and last scans for each instance.

Professional Heuristics for Ground Truth



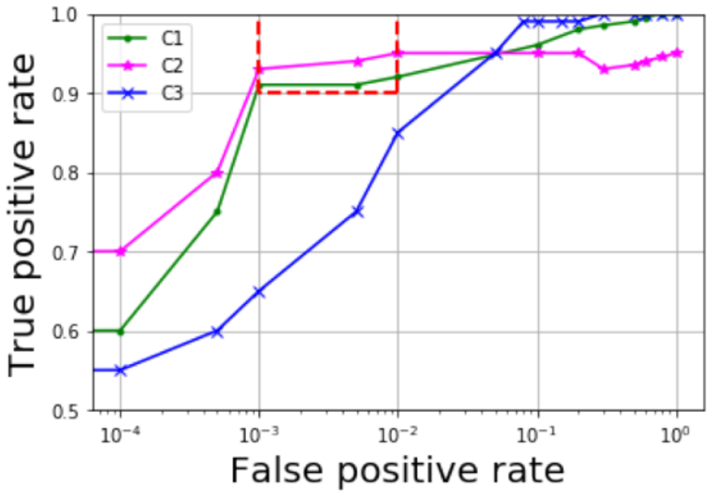
Avast Results

(100k samples in Sep 2019)

Our (professional) rule of thumb of malware ground truth:
One week delayed results on VT from Top Few (<10) companies is good enough

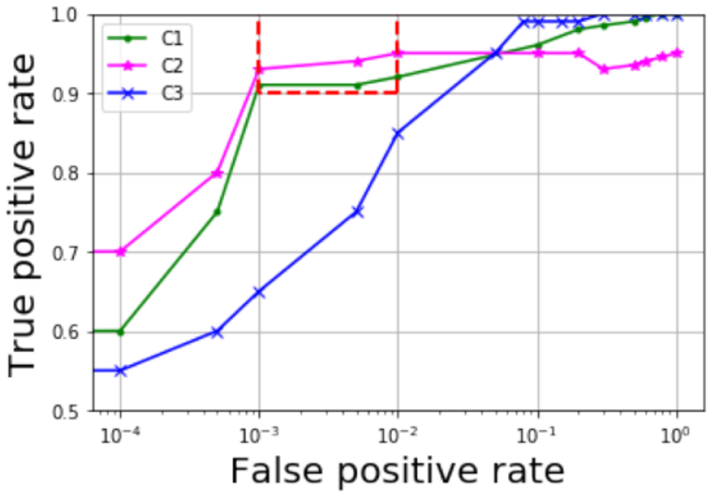
Does the overall performance of the classifiers matter?

Does the overall performance of the classifiers matter?



Which of the classifiers are best?

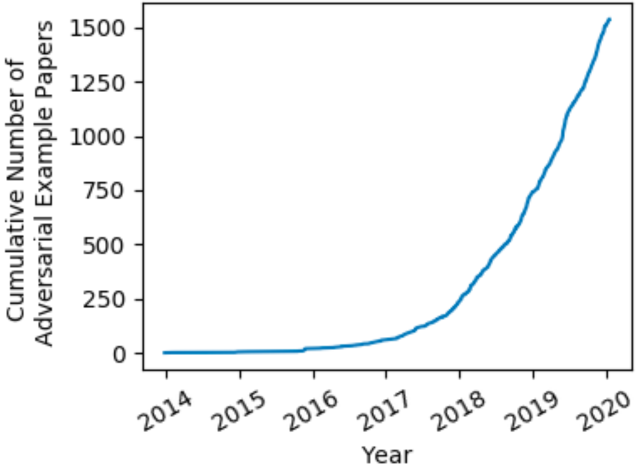
Does the overall performance of the classifiers matter?



Which of the classifiers are best?

Depends upon where you look!

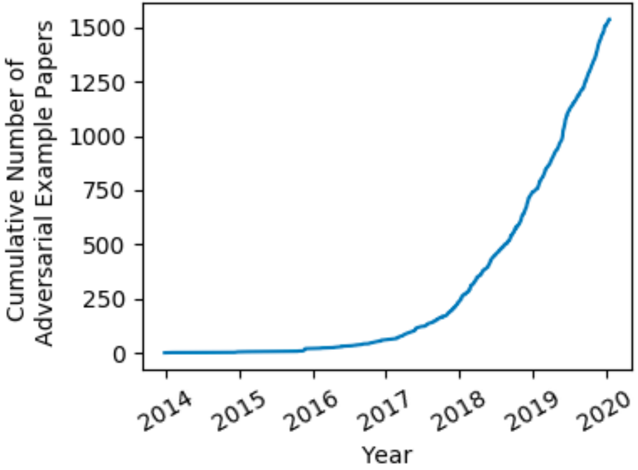
Adversarial attacks



Graph credit: Nicholas Carlini, Google Brain;

More than 1500 papers on adversarial ML

Adversarial attacks



Graph credit: Nicholas Carlini, Google Brain;

More than 1500 papers on adversarial ML

Only 36 (2.4%) papers focus on evading malware detectors

Can adversarial malware evade malware detectors?

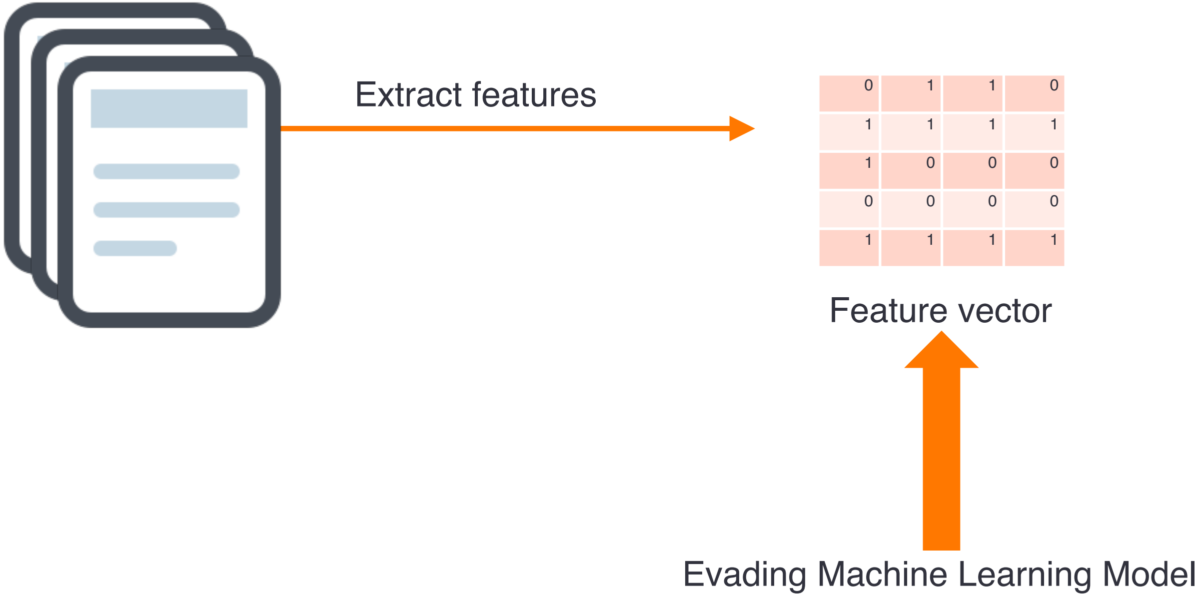
~~Can adversarial malware evade malware detectors?~~

~~Can adversarial malware evade malware detectors?~~

Are adversarial attacks harmful for users?

Adversarial attacks: feature space vs problem space

Adversarial attacks: feature space vs problem space



Adversarial attacks: feature space vs problem space

Extract features

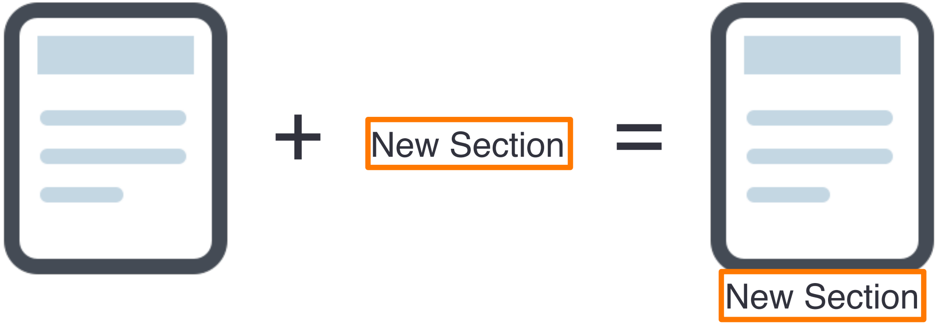
0	1	1	0
1	1	1	1
1	0	0	0
0	0	0	0
1	1	1	1

Feature vector

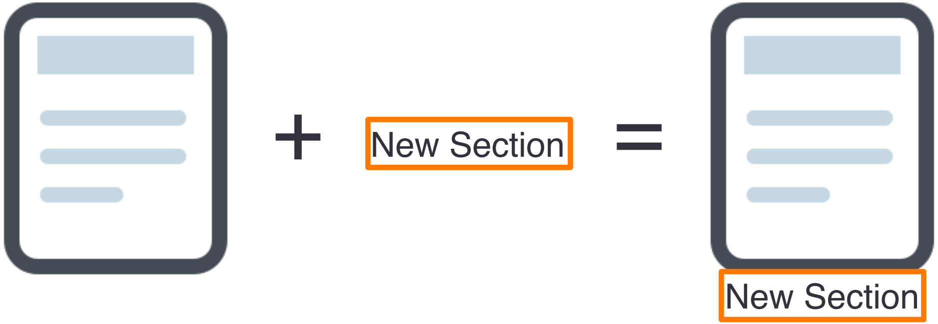
Checking Harm to Users

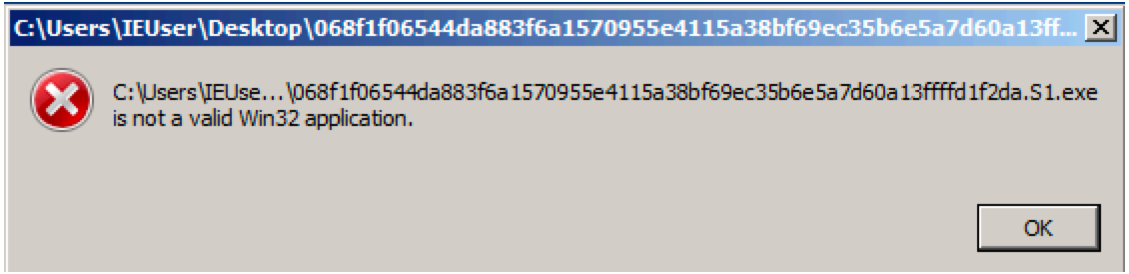
~~Evading Machine Learning Model~~

Adversarial attacks: feature space vs problem space



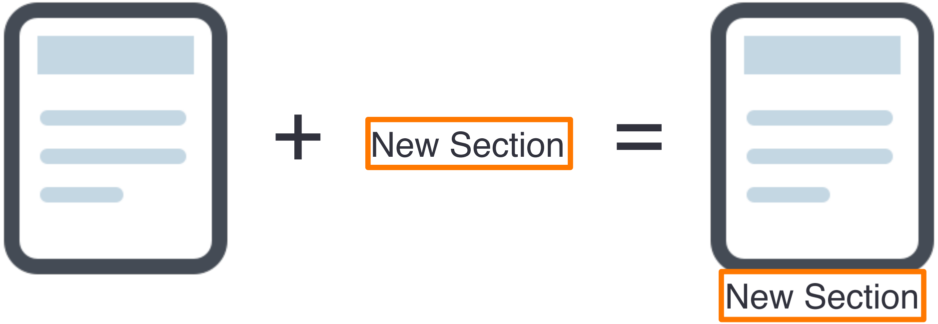
Adversarial attacks: feature space vs problem space

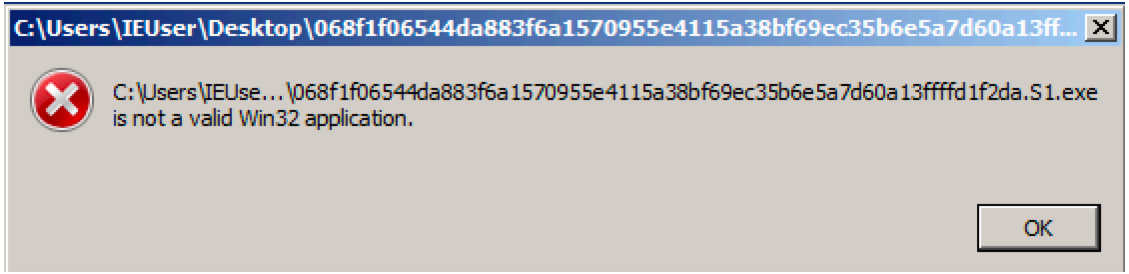




Adversarial attacks: feature space vs problem space

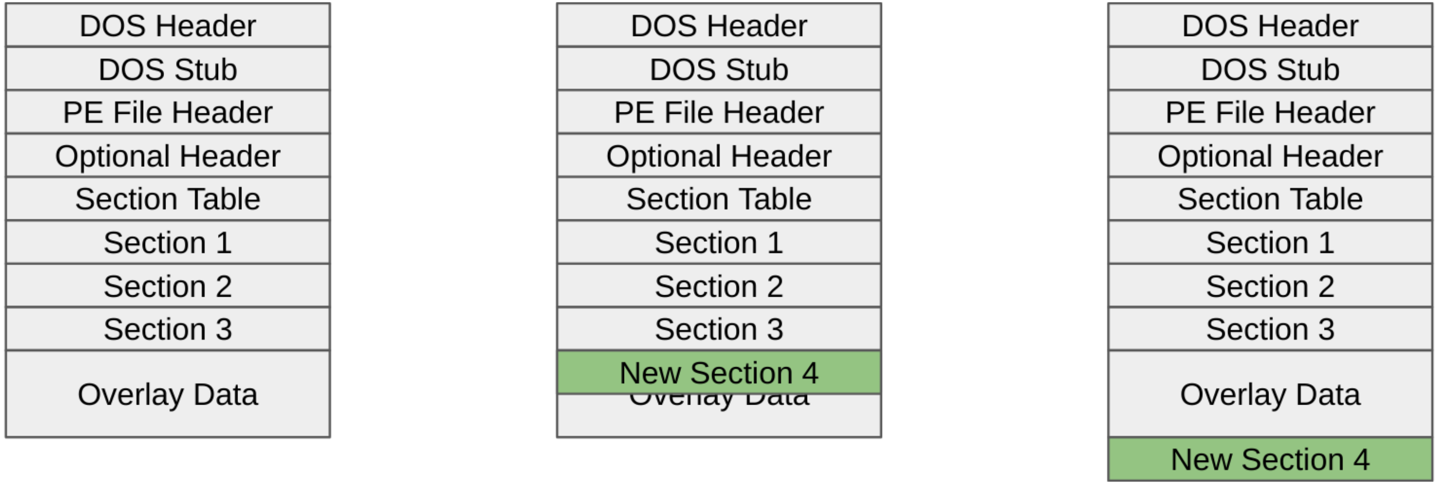
The new section can override an existing section



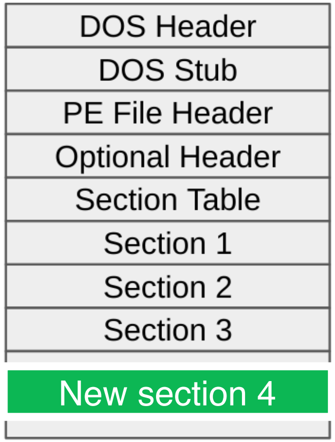


Adversarial attacks: feature space vs problem space

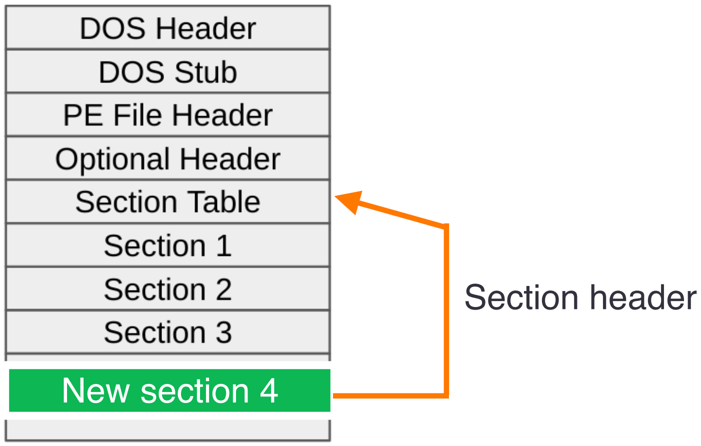
When adding a new section at the end of the last section,
if the sample has overlay data, the new section will overwrite the overlay data.



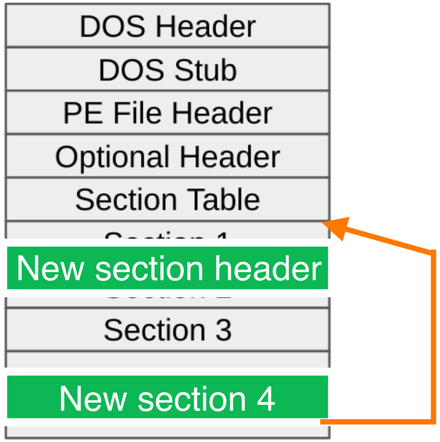
Adversarial attacks: feature space vs problem space



Adversarial attacks: feature space vs problem space



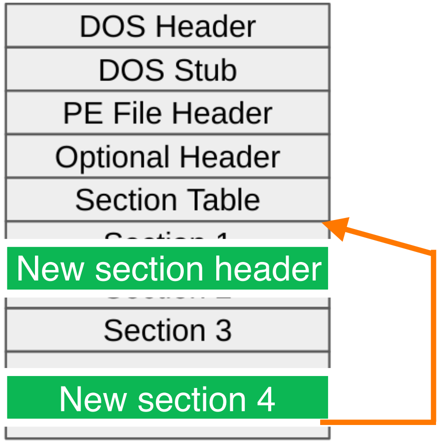
Adversarial attacks: feature space vs problem space



Override existing sections

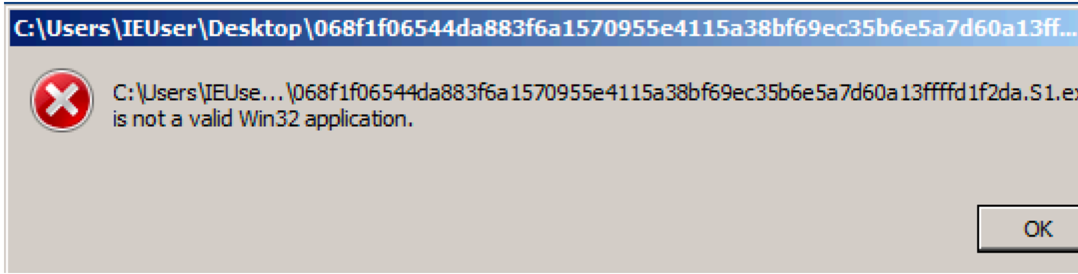
Section header

Adversarial attacks: feature space vs problem space



Override existing sections

Section header



Are adversarial attacks harmful to users?

Are adversarial attacks harmful to users?

papers changed
the malware files

Are adversarial attacks harmful to users?

9/36

papers changed the malware files

Are adversarial attacks harmful to users?

9/36

papers changed the malware files

papers tried to execute the adversarial samples

Are adversarial attacks harmful to users?

9/36

papers changed the malware files

4/36

papers tried to execute the adversarial samples

Are adversarial attacks harmful to users?

9/36

papers changed
the malware files

4/36

papers tried
to execute the
adversarial
samples

papers check if the
modified malware
is harmful to users

Are adversarial attacks harmful to users?

9/36

papers changed the malware files

4/36

papers tried to execute the adversarial samples

0/36

papers check if the modified malware is harmful to users

Are adversarial attacks harmful to users?

From: <security@google.com>
Date: Tuesday, February 23, 2016
Subject: [5-0014000010203] other in https://mail.google.com

Hey!

Thanks for your feedback. I think generally because of the ways anti-viruses work there's not really much we can do in this case, but thanks for letting us know!

Eduardo
Google Security Team

[1] Xu et al., NDSS Talk: Automatically Evading Classifiers (including Gmail's).

Is evading one classifier enough?

* Hashes and hand written rules

Is evading one classifier enough?

Sample

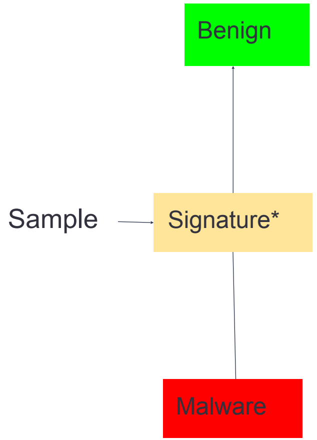
* **Hashes and hand written rules**

Is evading one classifier enough?

Sample → Signature*

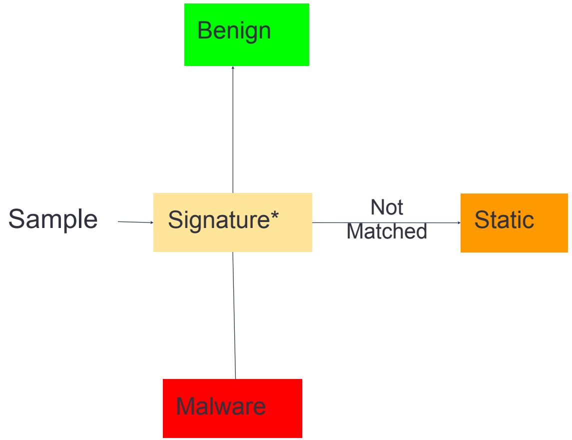
* **Hashes and hand written rules**

Is evading one classifier enough?



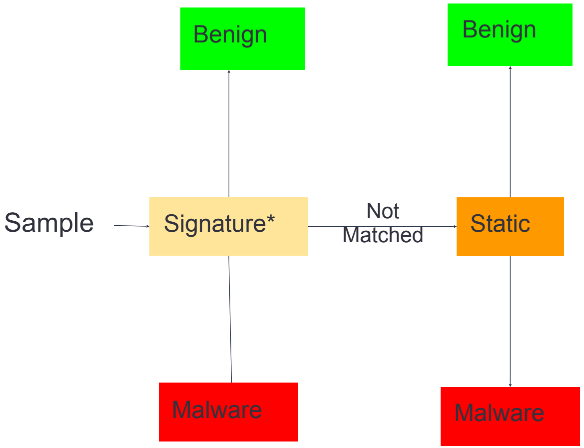
* **Hashes and hand written rules**

Is evading one classifier enough?



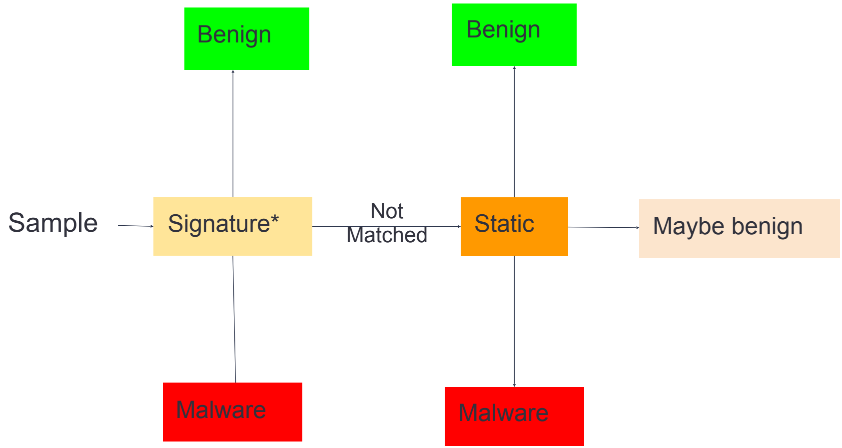
* **Hashes and hand written rules**

Is evading one classifier enough?



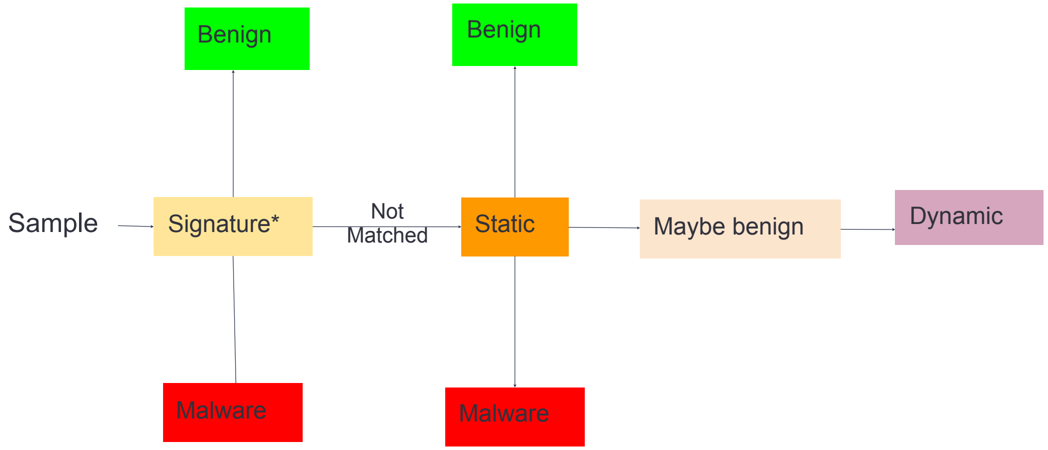
* **Hashes and hand written rules**

Is evading one classifier enough?



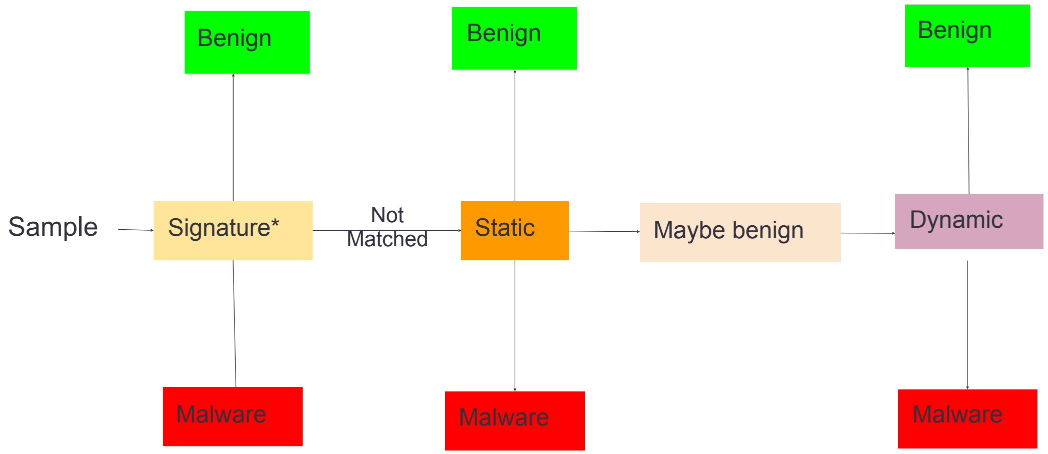
* **Hashes and hand written rules**

Is evading one classifier enough?



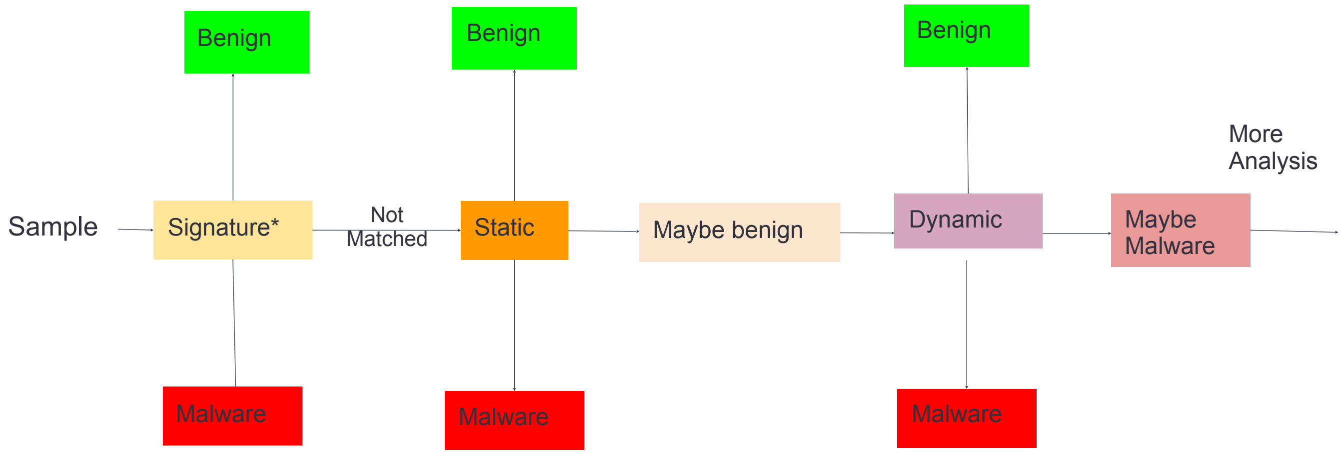
*** Hashes and hand written rules**

Is evading one classifier enough?



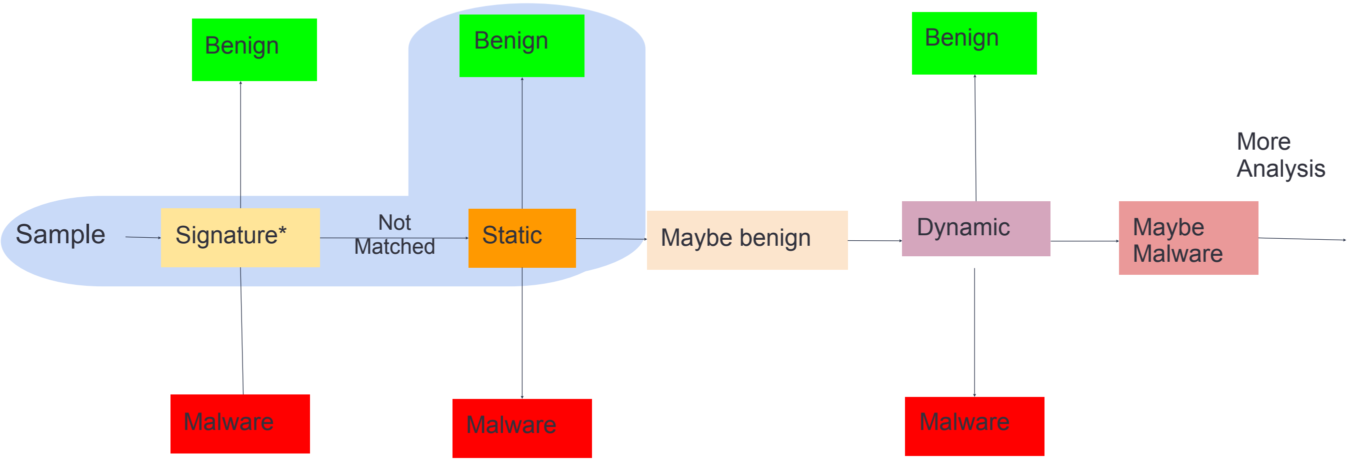
* **Hashes and hand written rules**

Is evading one classifier enough?



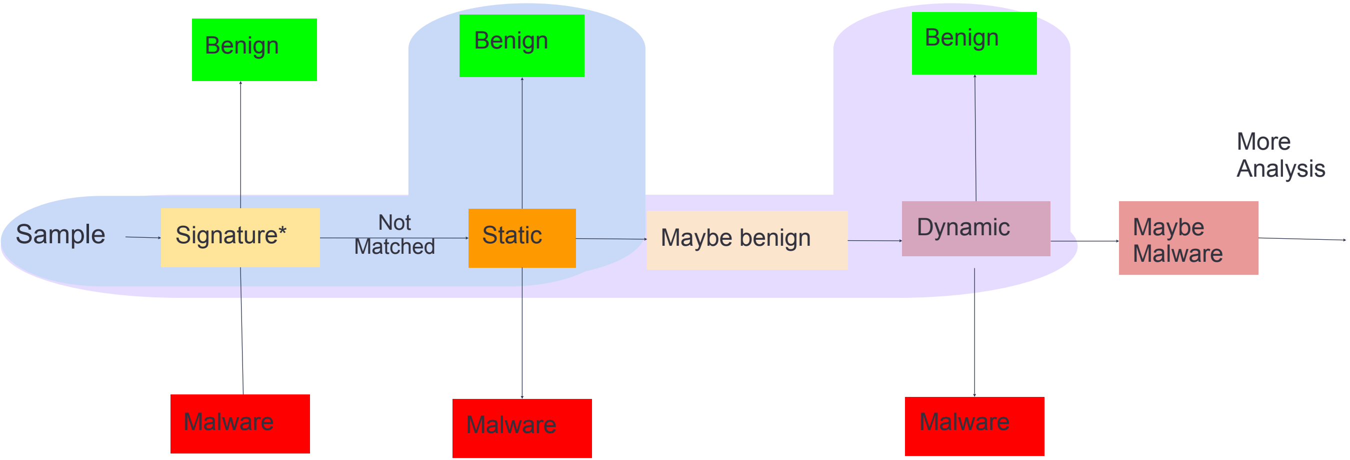
* **Hashes and hand written rules**

Is evading one classifier enough?



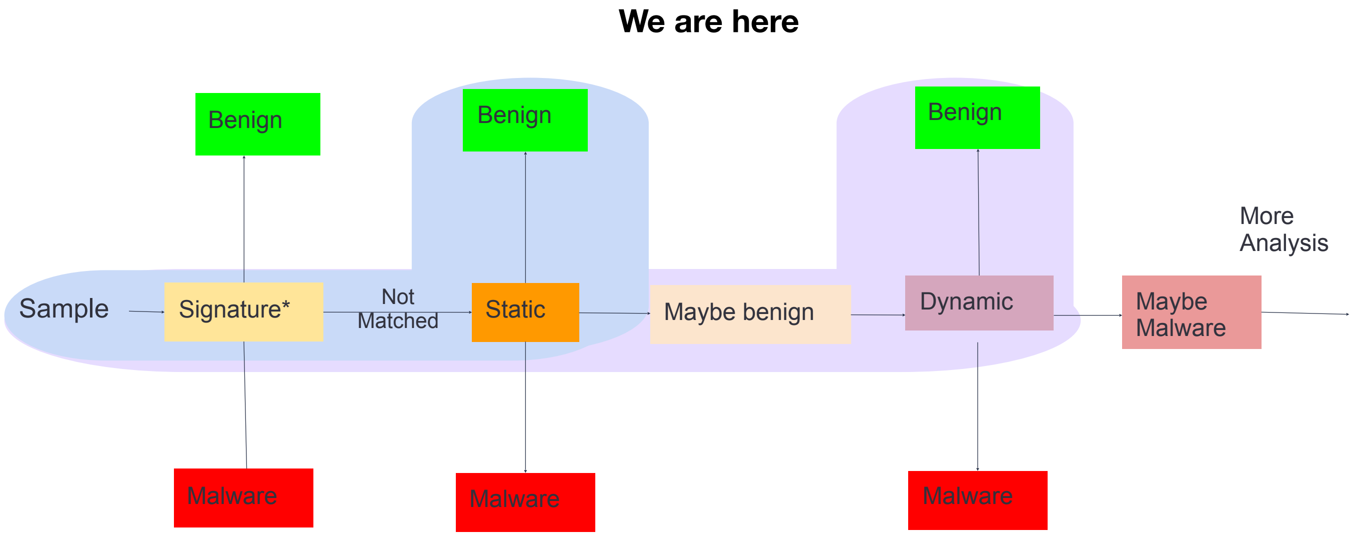
* **Hashes and hand written rules**

Is evading one classifier enough?



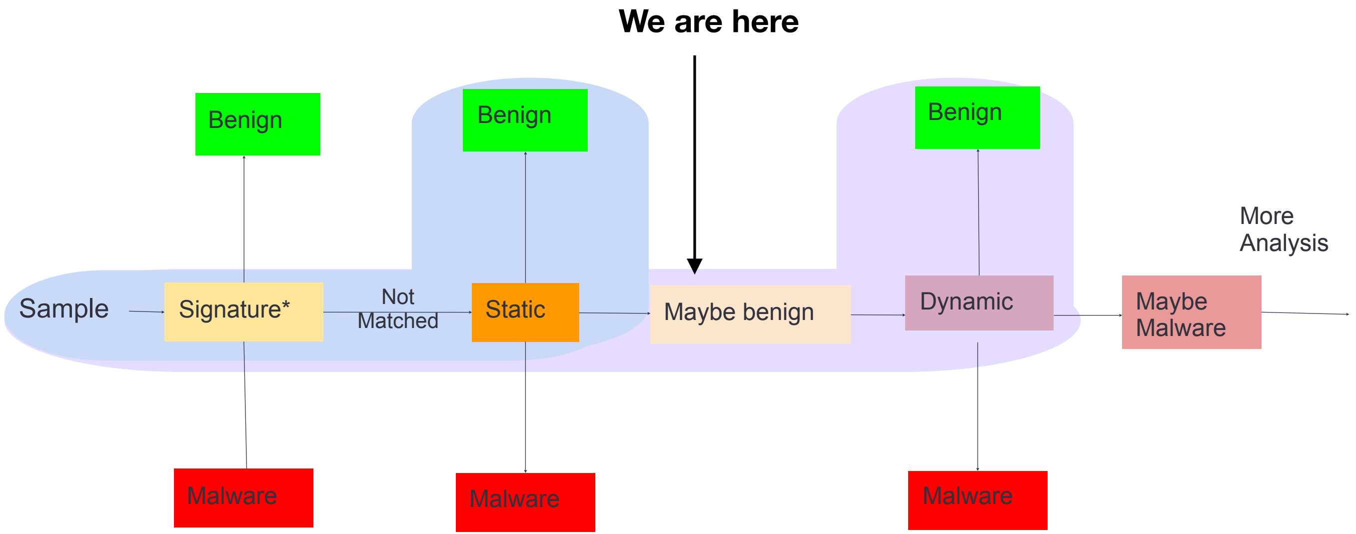
* **Hashes and hand written rules**

Is evading one classifier enough?



*** Hashes and hand written rules**

Is evading one classifier enough?



* **Hashes and hand written rules**

Who is the adversary?

White box

Adversary has full access

Black box

Adversary has no access

Who is the adversary?

White box

Adversary has full access

Black box

Adversary has no access

Who is the adversary?

White box

Adversary has full access

Grey box

Black box

Adversary has no access

Who is the adversary?

White box	Grey box	Black box
Adversary has full access	Adversary has full access to the features	Adversary has no access

Who is the adversary?

White box

Adversary has full access

Grey box

Adversary has full access to the features

Adversary can do unlimited queries

Black box

Adversary has no access

Who is the adversary?

White box

Adversary has full access

Grey box

Adversary has full access to the features

Adversary can do unlimited queries

Adversary has access to the training data

Black box

Adversary has no access

Who is the adversary?

White box

Adversary has full access

Grey box

Adversary has full access to the features

Adversary can do unlimited queries

Adversary has access to the training data

Adversary can build substitute classifiers

Black box

Adversary has no access

How to Build Realistic Machine Learning Systems for Security?

Consistent ground truth

Proper evaluation

Measurable adversary

Questions?

Sadia Afroz

sadia.afroz@avast.com

Research contributors

Rajarshi Gupta
VP, Head of AI
Avast

Deepali Garg
Senior Data Scientist
Avast

Fabrizio Bondi
AI Manager
Avast

Heng Yin
Associate Professor
UC Riverside

Wei Song
PhD Student
UC Riverside

Xuezixiang Li
PhD Student
UC Riverside

avast