ENIGMA

Classifiers Under Attack

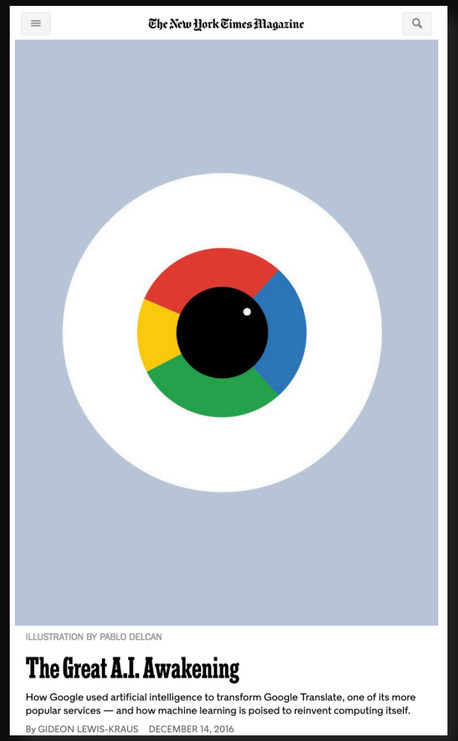
1 February 2017

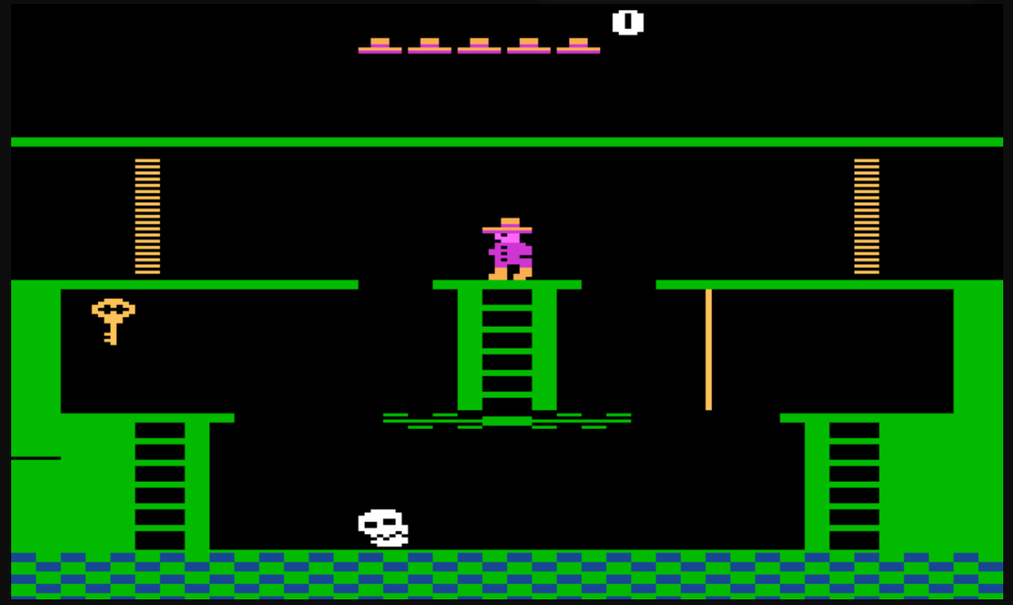
David Evans
work with **Weilin Xu** and **Yanjun Qi**
University of Virginia

evans@virginia.edu
evadeML.org

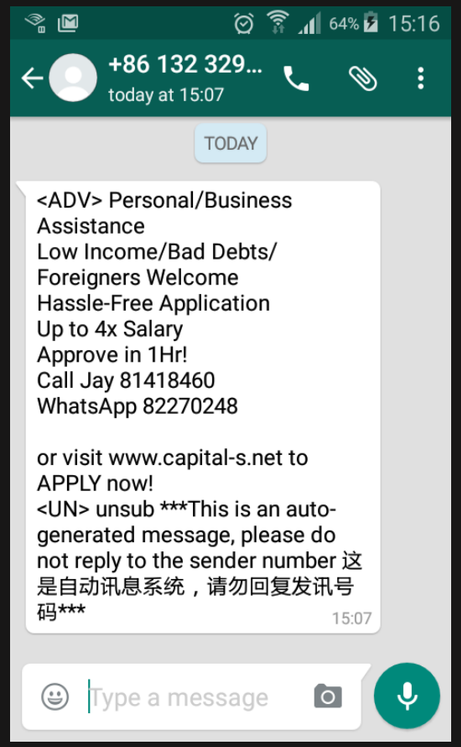
Machine Learning Does Amazing Things

FeatureSmith





… and can solve all Security Problems!



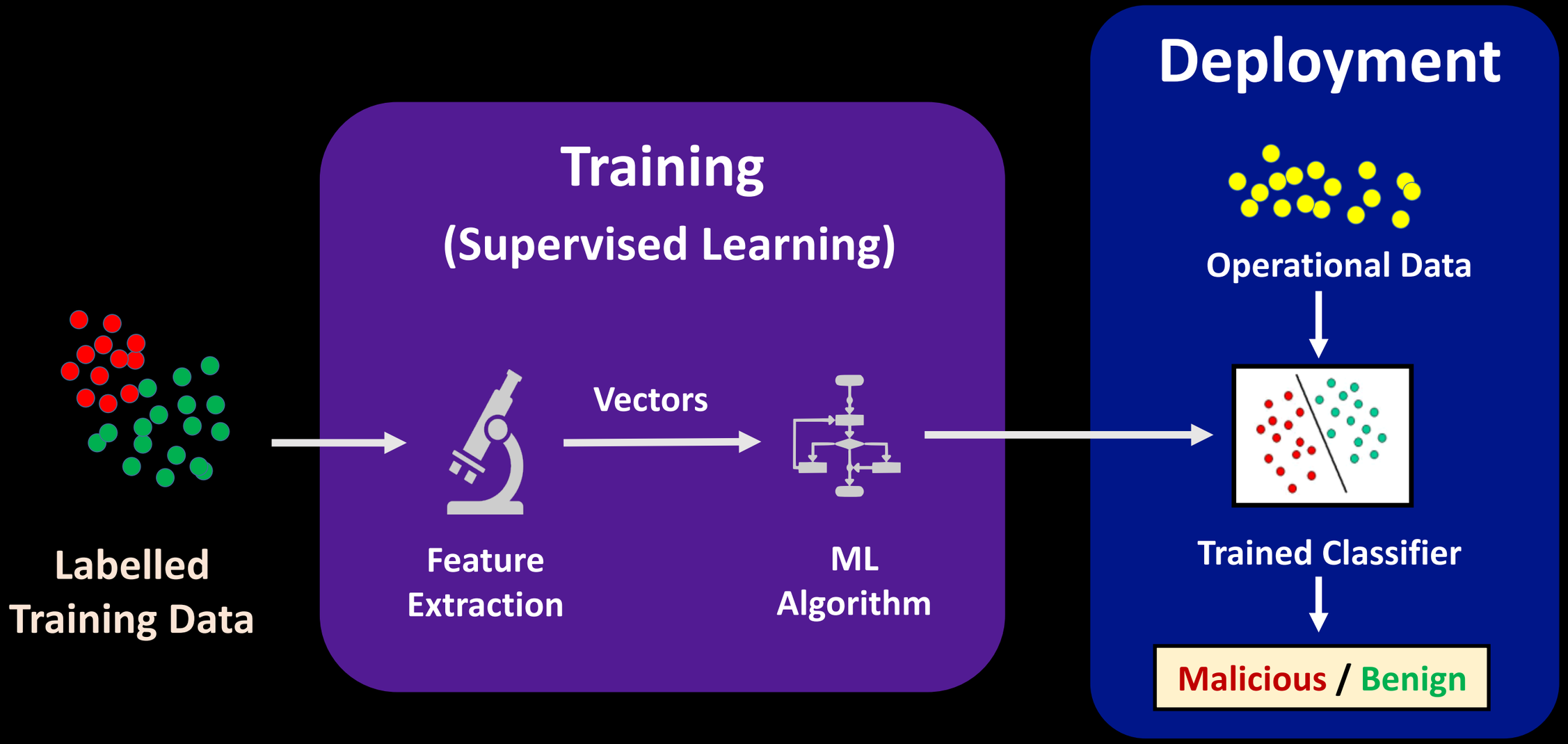
Spam

IDS

Malware

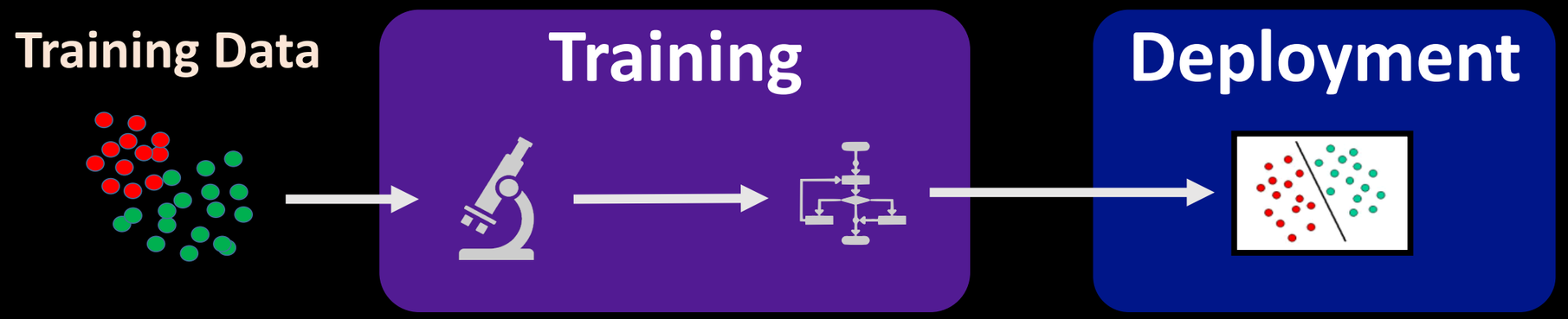
Fake Accounts

…



Assumption: Training Data is *Representative*

Adversaries Don’t Cooperate

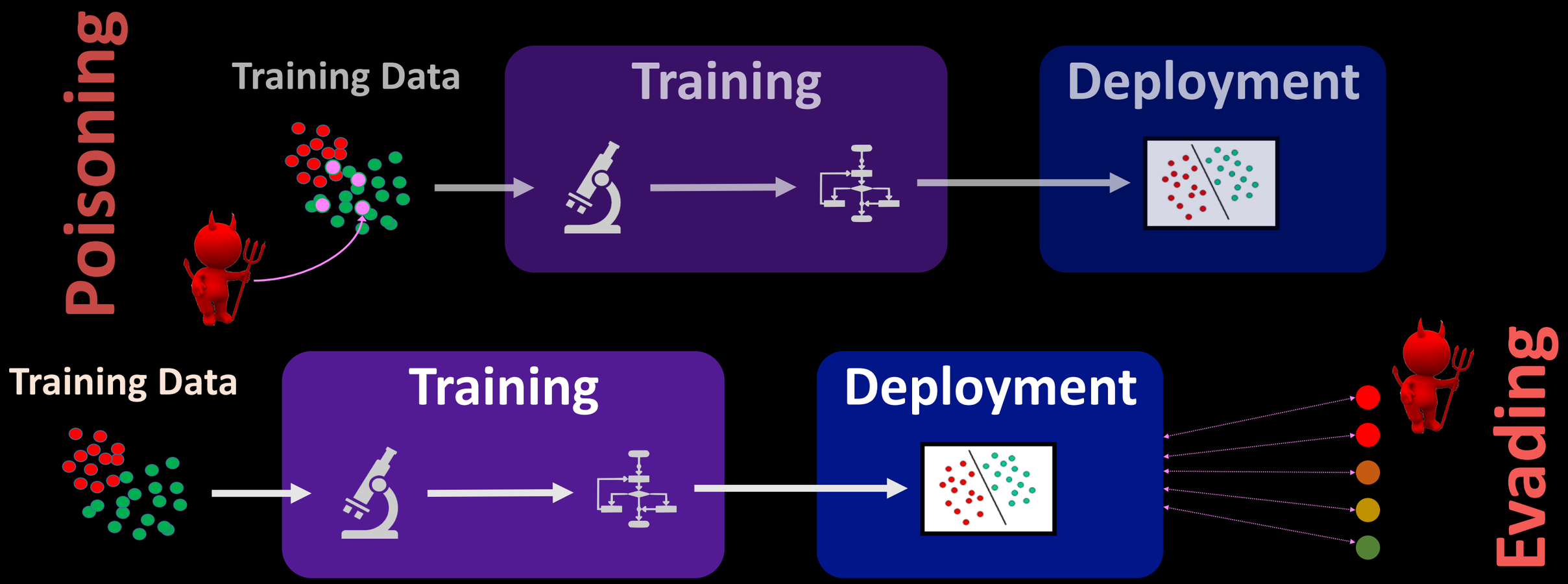


Assumption: Training Data is *Representative*

Adversaries Don't Cooperate

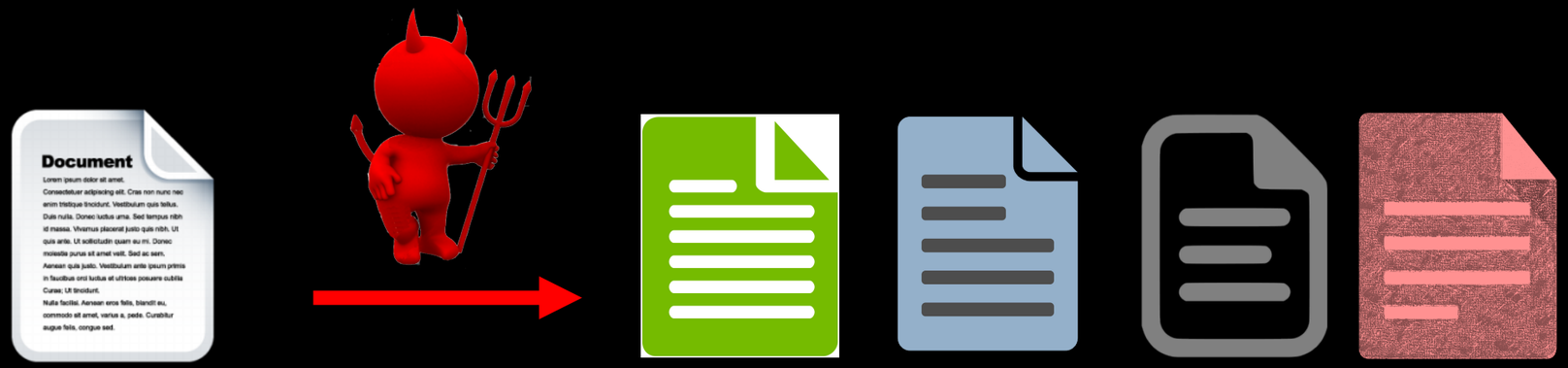
~~**Assumption:** Training Data is *Representative*~~

Adversaries Don't Cooperate



~~**Assumption:** Training Data is *Representative*~~

Focus: **Evasion Attacks**



Goal: Automatically **simulate adaptive adversary** against generic classifier

Purpose: **Understand** classifier robustness

Build better classifiers (or give up)

Case study:
Evading **PDF Malware** Classifiers

THE DUKES

7 years of Russian cyberespionage

TLP: WHITE

This whitepaper explores the tools - such as MiniDuke, CosmicDuke, OnionDuke, CozyDuke, etc- of the Dukes, a well-resourced, highly dedicated and organized cyberespionage group that we believe has been working for the Russian Federation since at least 2008 to collect intelligence in support of foreign and security policy decision-making.

F-SECURE LABS THREAT INTELLIGENCE

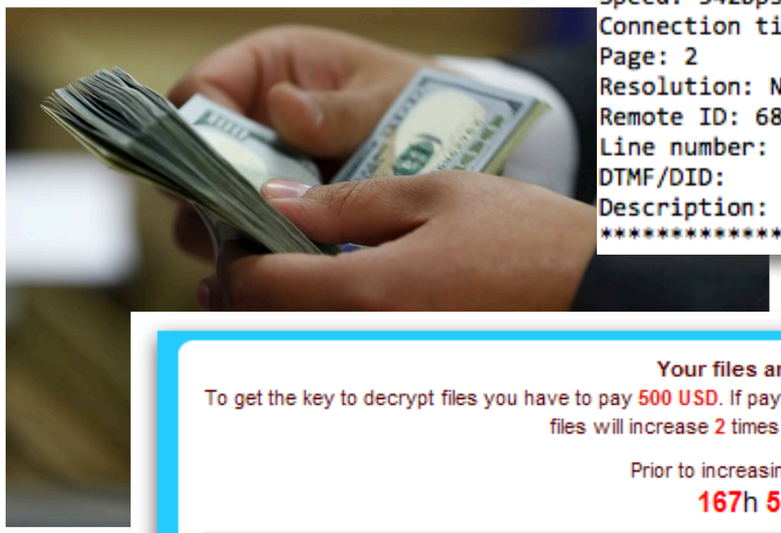
Whitepaper

SECTIONS The New York Times

DealB%k WITH FOUNDER ANDREW ROSS SORKIN

Hackers' $81 Million Sneak Attack on World Banking

By MICHAEL CORKERY APRIL 30, 2016



Anxieties have arise
an international ban

```
From: Incoming Fax [mailto:no-reply@efax-delivery.com]
Sent: Tuesday, October 21, 2014 3:14 PM
To: [redacted]
Subject: Incoming Fax Report

************************************
INCOMING FAX REPORT
************************************

Date/Time: Tuesday, 21.10.2014
Speed: 342bps
Connection time: 01:08
Page: 2
Resolution: Normal
Remote ID: 681-748-172435
Line number: 8
DTMF/DID:
Description: Internal only
************************************
```

Your files are encrypted.

To get the key to decrypt files you have to pay 500 USD. If payment is not made before 20/07/15 - 19:41 the cost of decrypting files will increase 2 times and will be 1000 USD/EUR

Prior to increasing the amount left:

167h 56m 11s

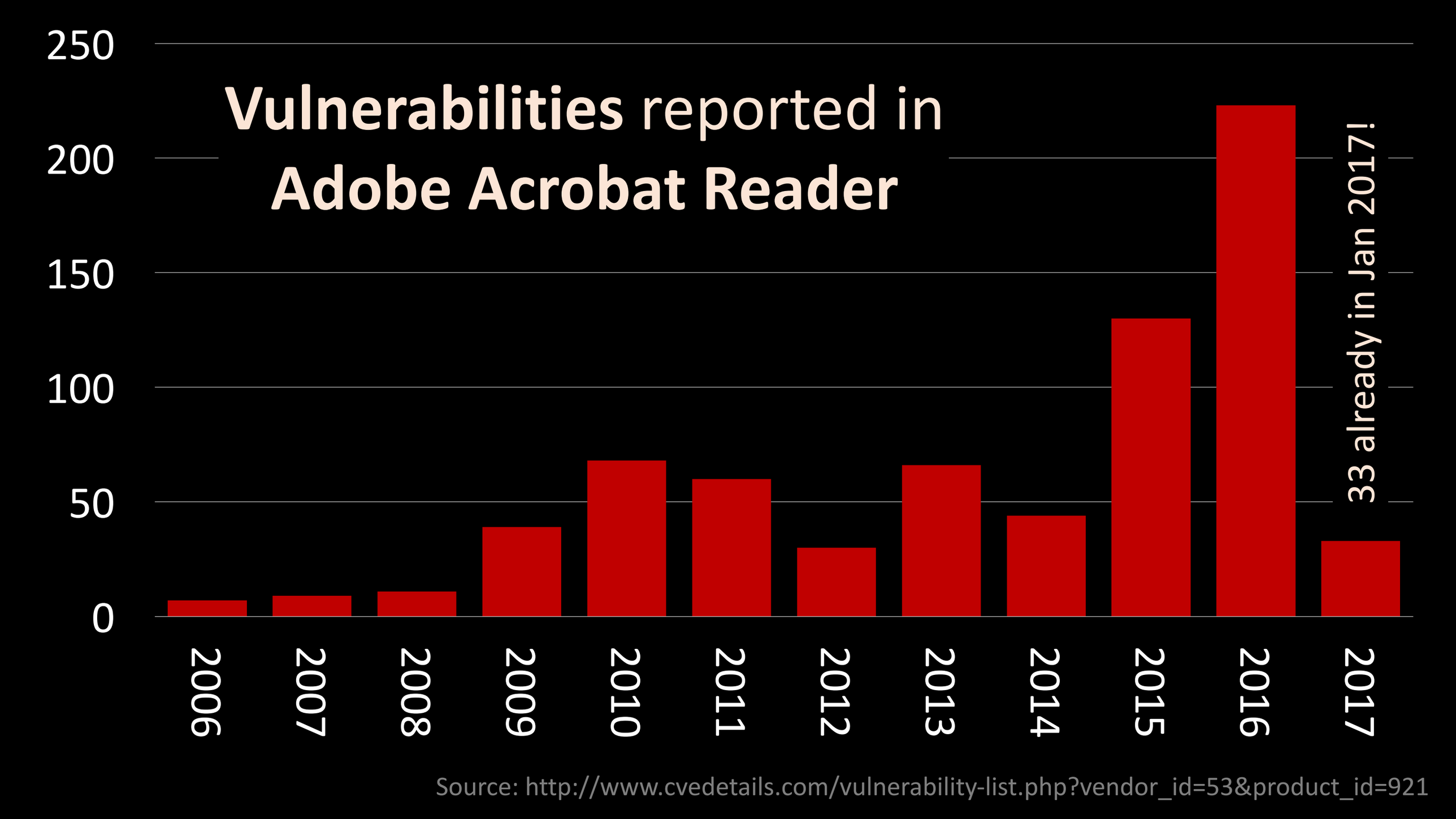
Your system: Windows XP (x32) First connect IP: [redacted] Total encrypted 330 files.

Refresh | Payment | FAQ | Decrypt 1 file for FREE | Support

We give you the opportunity to decipher 1 file free of charge! You can make sure that the service really works and after payment for the CryptoWall program you can actually decrypt the files.

Your file is successfully decoded. You can download it

Download decrypted file


Vulnerabilities reported in
Adobe Acrobat Reader
250
200
150
100
50
0
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
33 already in Jan 2017!
Source: http://www.cvedetails.com/vulnerability-list.php?vendor_id=53&product_id=921

PDF Malware Classifiers

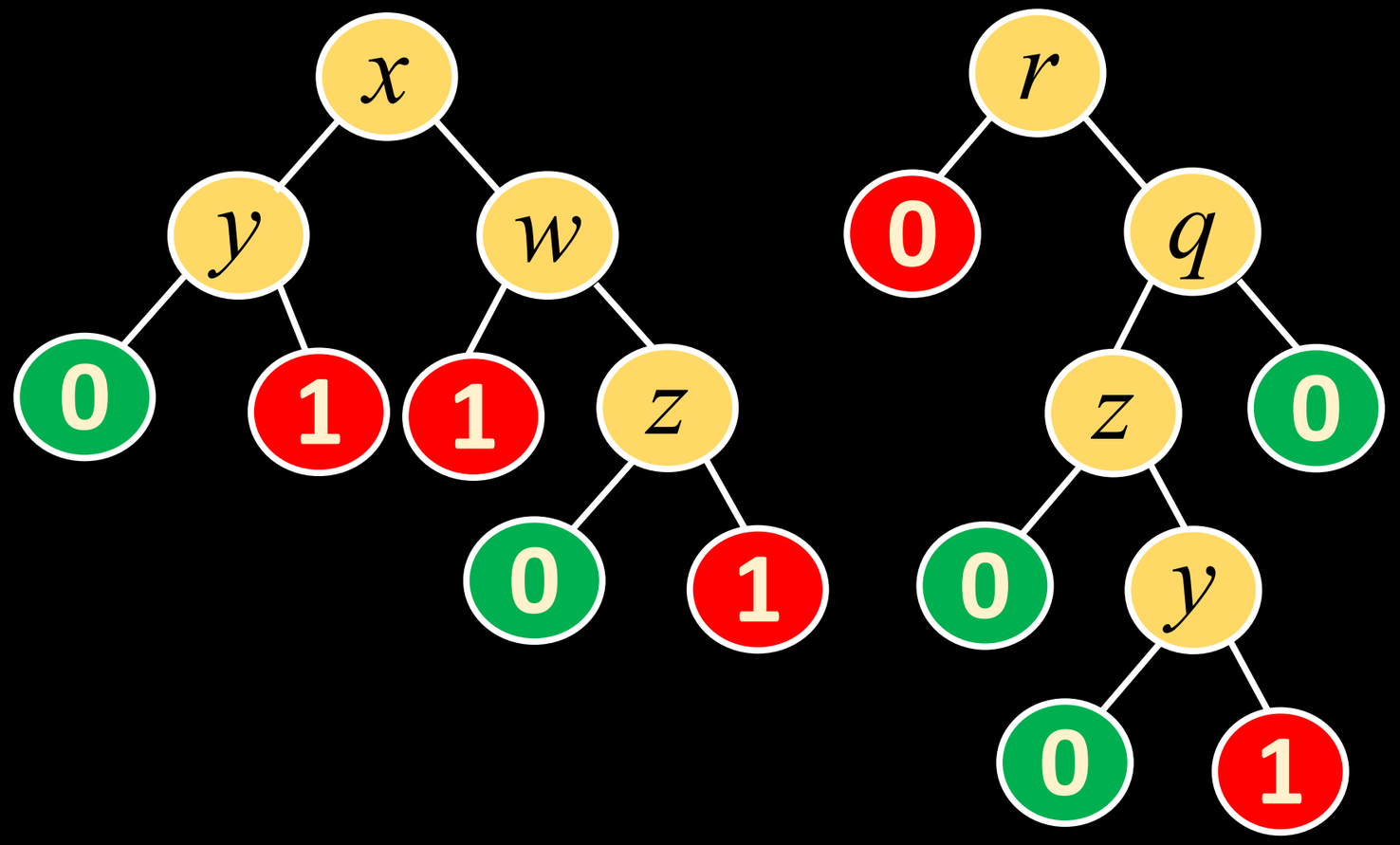
PDFrate	**Hidost[13]**	**Hidost[16]**
[ACSA 2012]	[NDSS 2013]	[JIS 2016]
Random Forest	Support Vector Machine	**Random Forest**

Classifier Accuracy

0.9976	0.9996	0.9996

* Mimicus [Oakland 2014], an open source reimplementation of PDFrate.

Random Forest



Generate many **random decision trees**

Train independently

Select best trees

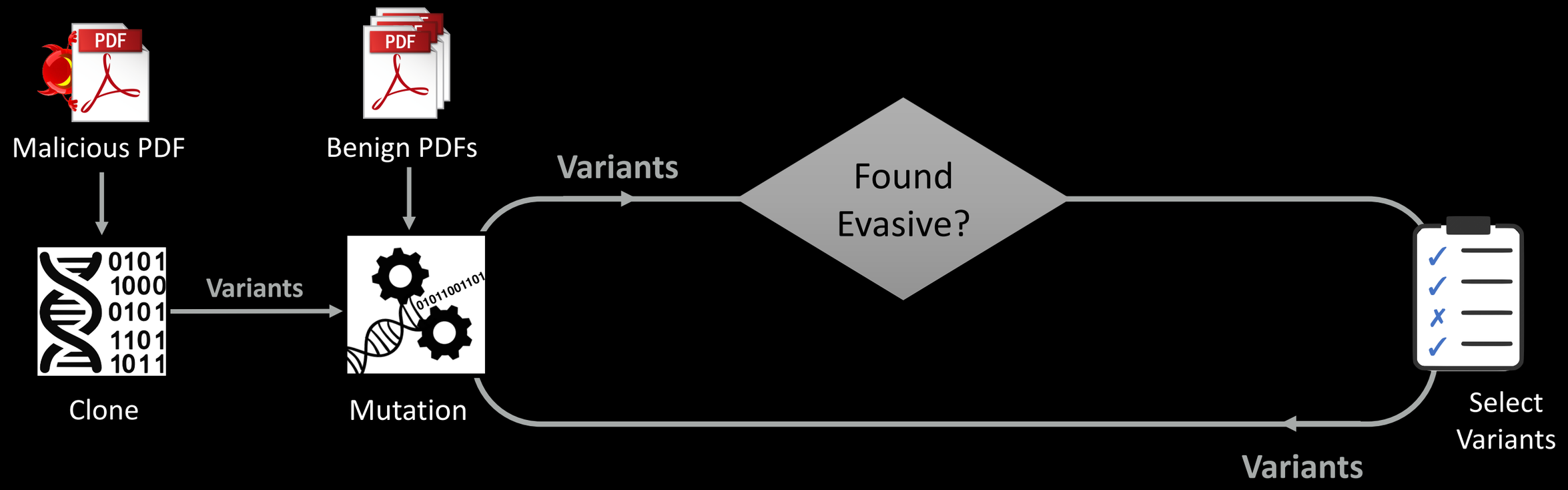
Vote on result

PDF Malware Classifiers

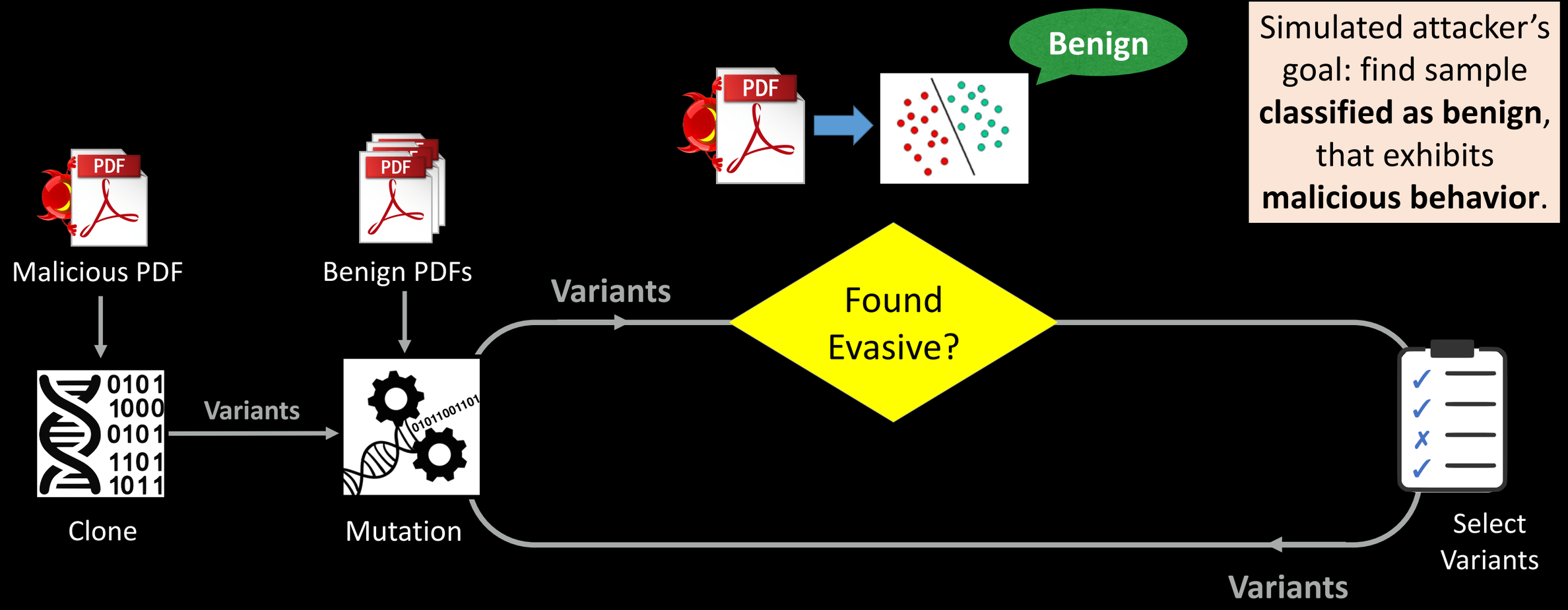
	PDFrate [ACSA 2012]	**Hidost[13]** [NDSS 2013]	**Hidost[16]** [JIS 2016]
	Random Forest	Support Vector Machine	Random Forest
Features	Object counts, lengths, positions, …	**Object structural paths** *Very robust against "strongest conceivable mimicry attack".*	
	Manual Features	**Automated Features**	

Automatically Evading Classifiers

Automated Classifier Evasion Using **Genetic Programming**

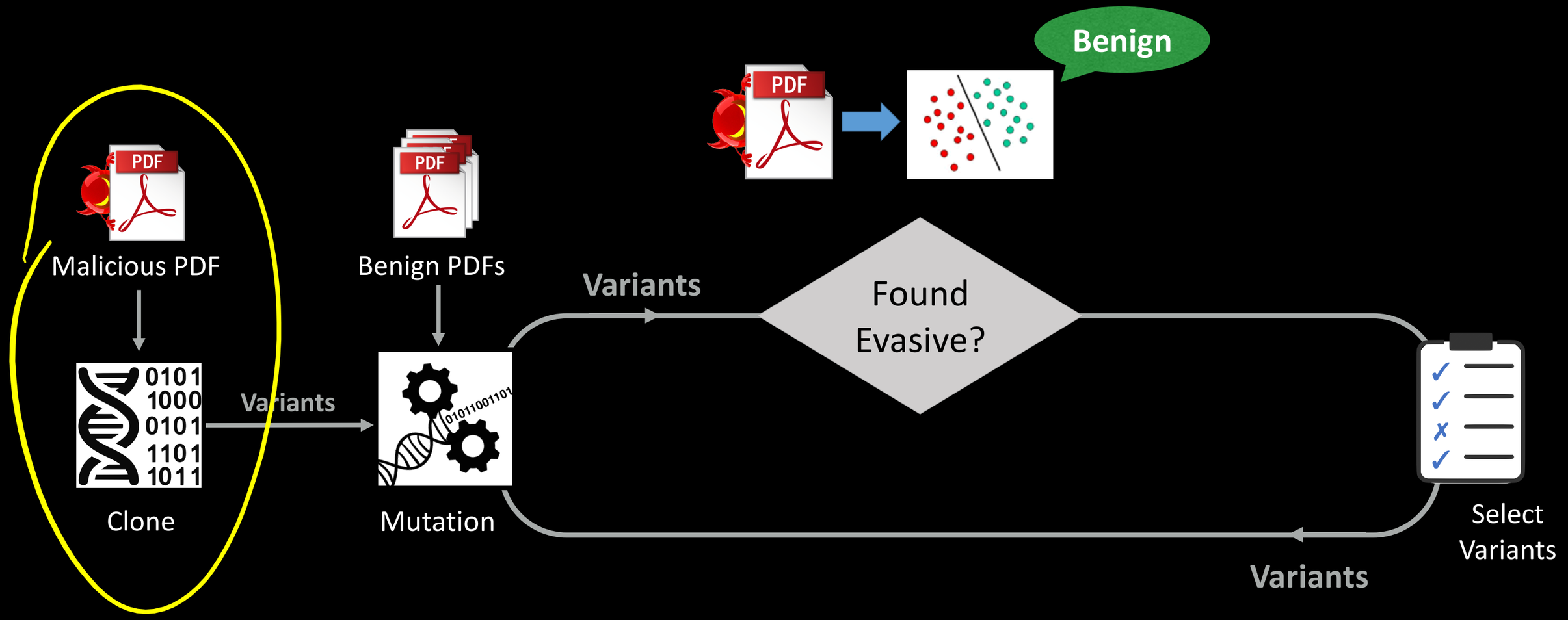


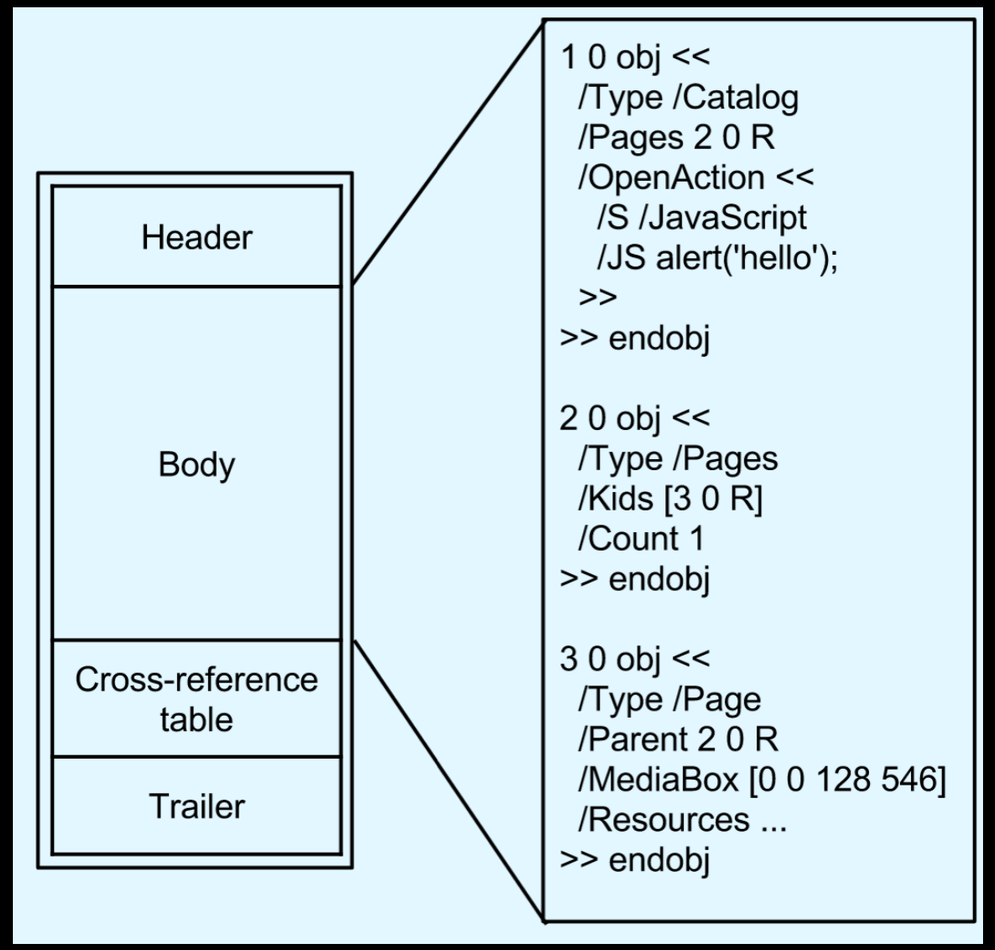
Goal: **Find Evasive Variant**

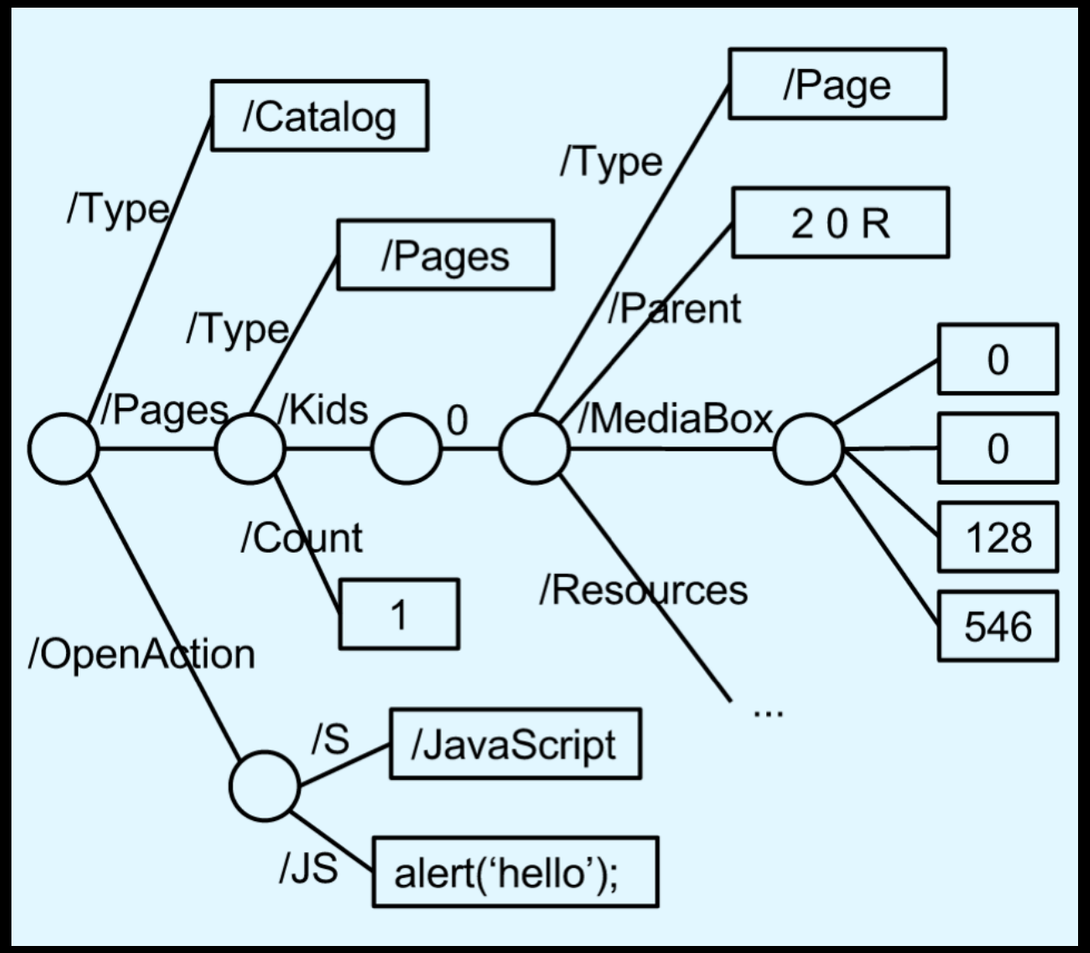


Simulated attacker's goal: find sample **classified as benign**, that exhibits **malicious behavior**.

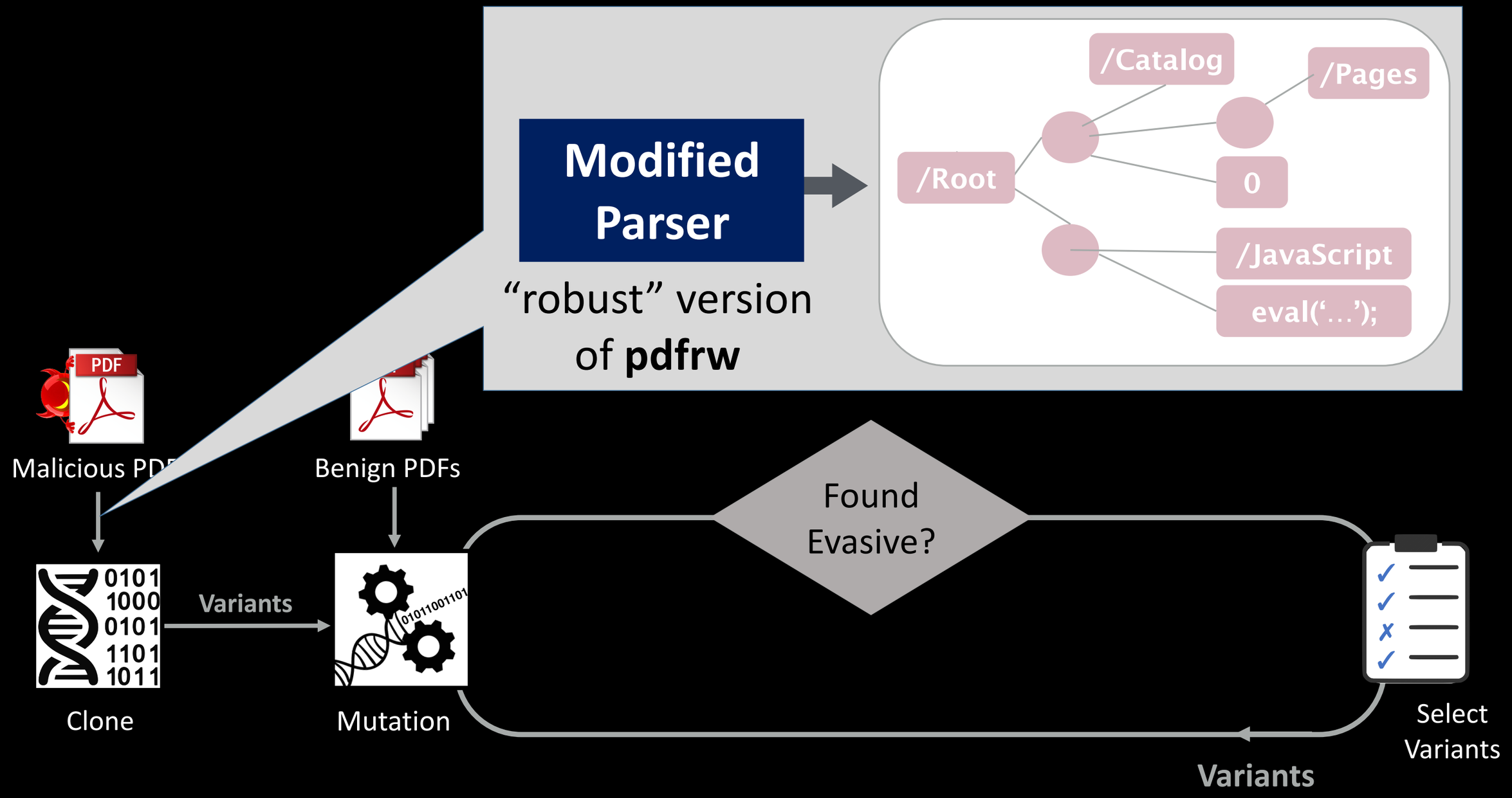
Start with **Malicious Seed**



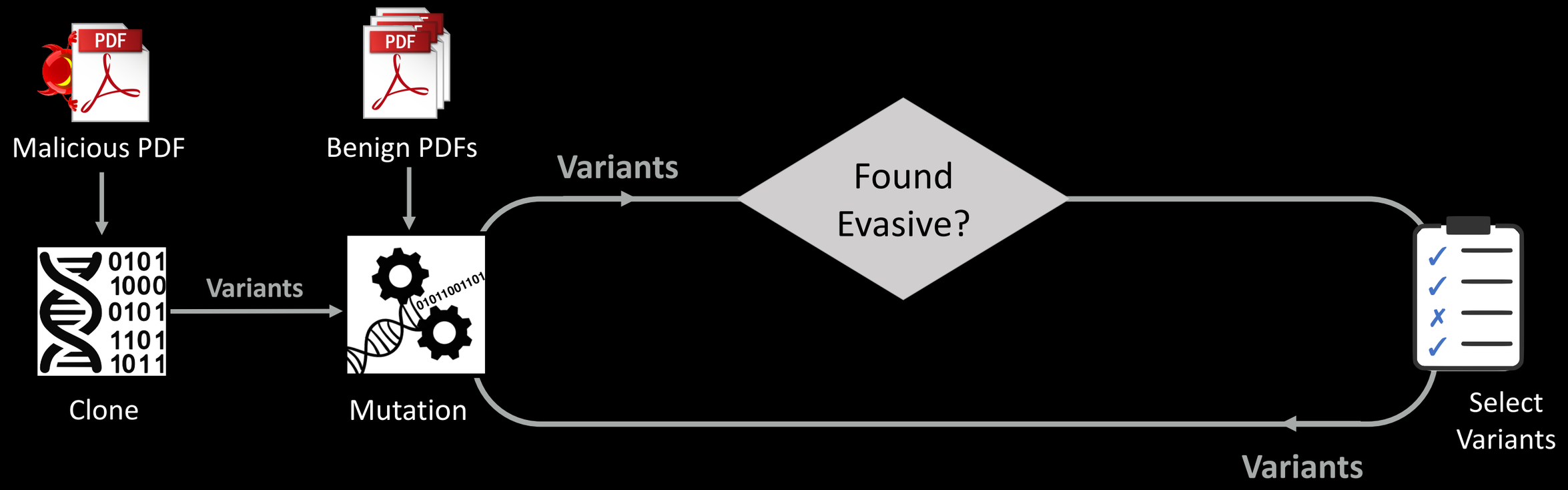




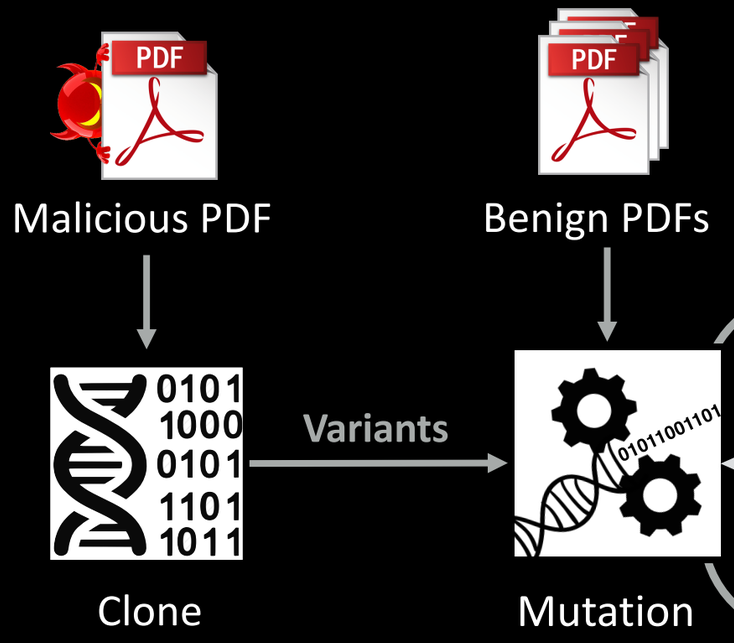
PDF Structure

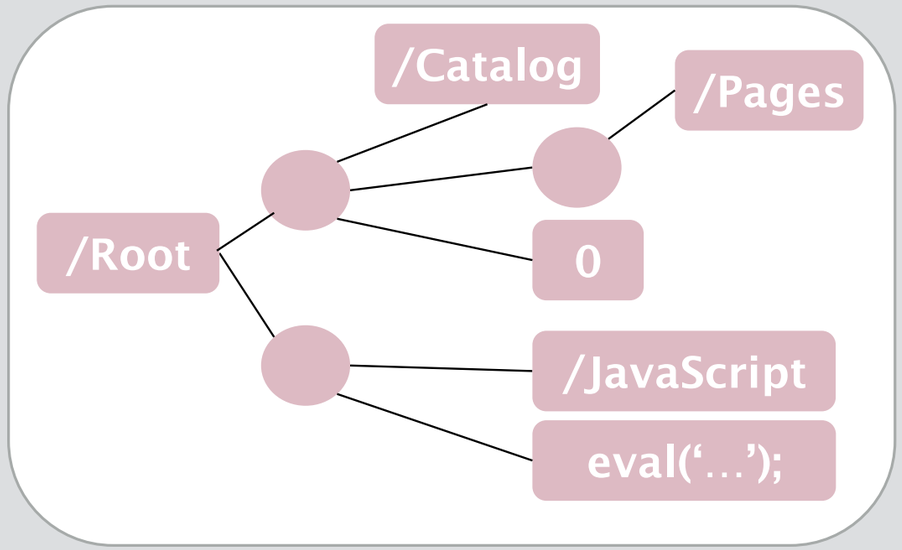

Modified Parser
"robust" version of pdfrw
/Catalog
/Pages
/Root
0
/JavaScript
eval('...');
PDF
Malicious PDF
Benign PDFs
Found Evasive?
0101
1000
0101
1101
1011
Variants
Clone
Mutation
Select Variants
Variants

Generating **Variants**



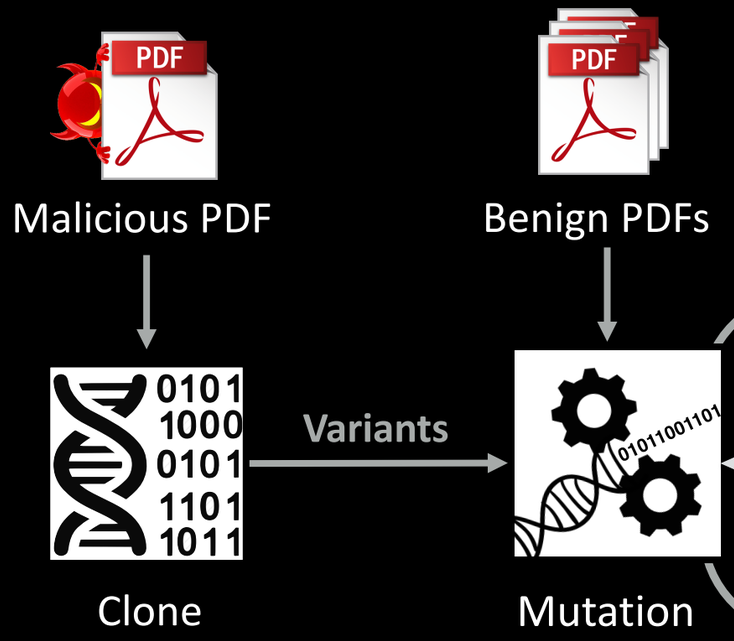
Generating **Variants**

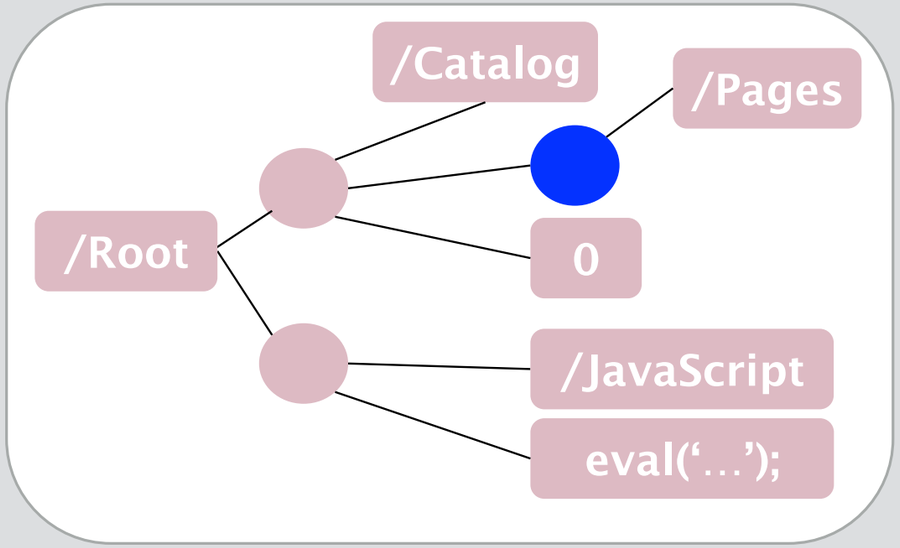




Select **random** node

Generating **Variants**





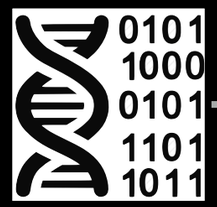
Select **random** node

Random transform: **delete**, insert, replace

Generating **Variants**

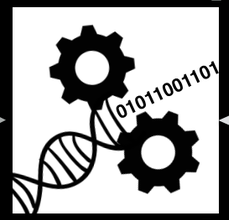
Malicious PDF

Benign PDFs

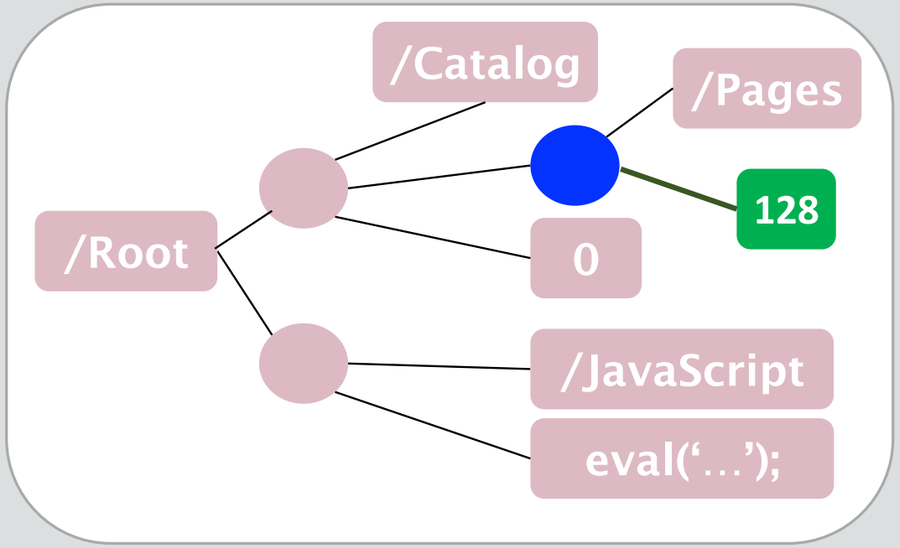


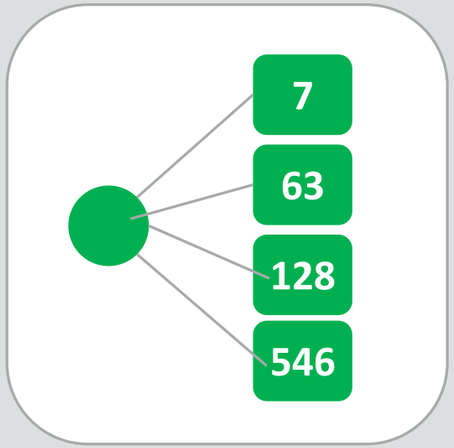
Clone

Variants



Mutation



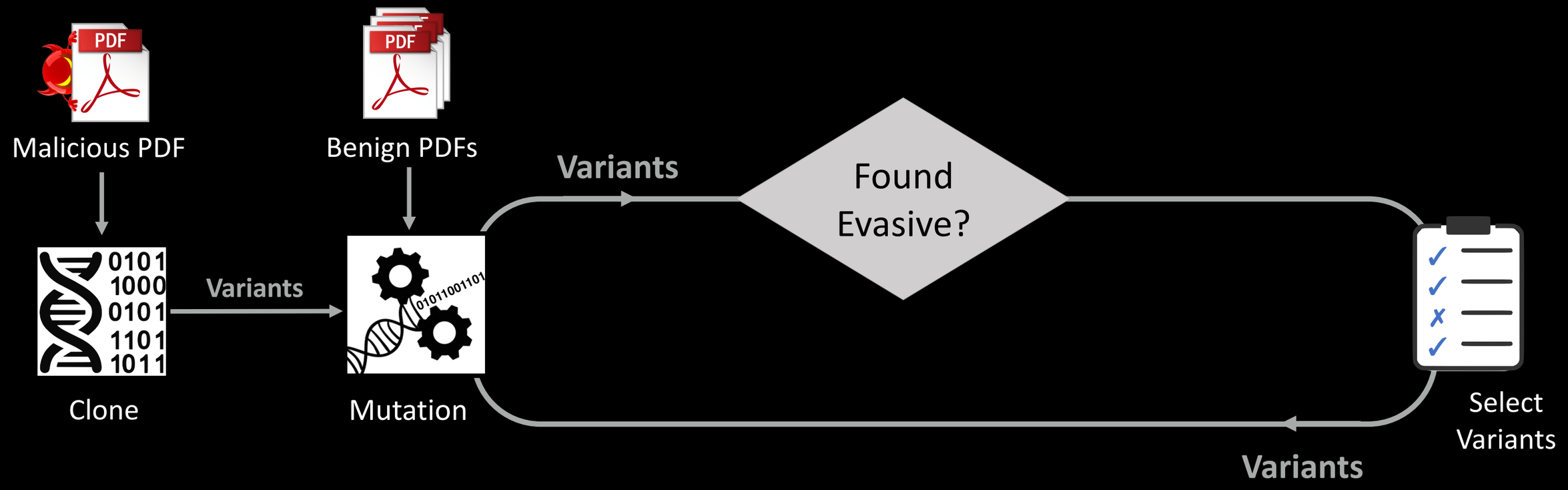


Nodes from **Benign PDFs**

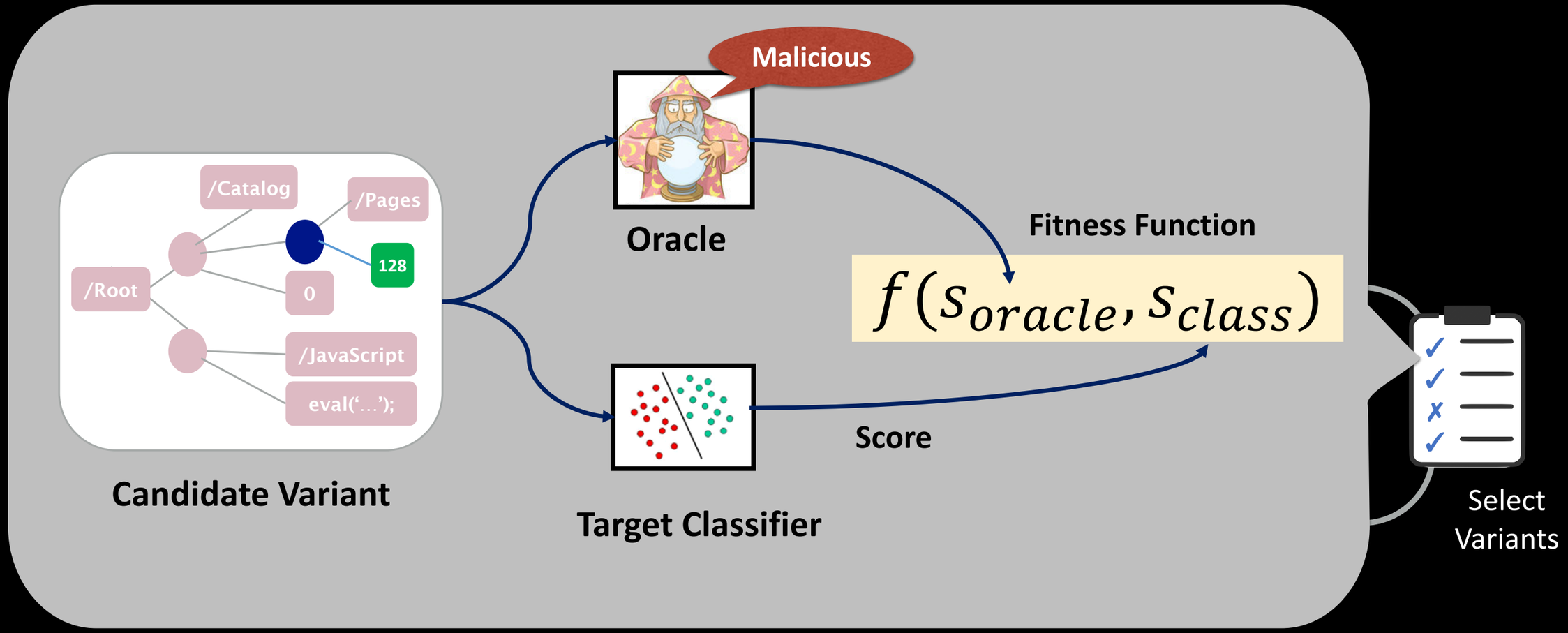
Select **random** node

Random transform: **delete**, **insert**, replace

Selecting Promising Variants



Selecting Promising Variants



Oracle

Execute candidate in vulnerable Adobe Reader in virtual environment

Simulated network: INetSim

Cuckoo

https://github.com/cuckoosandbox

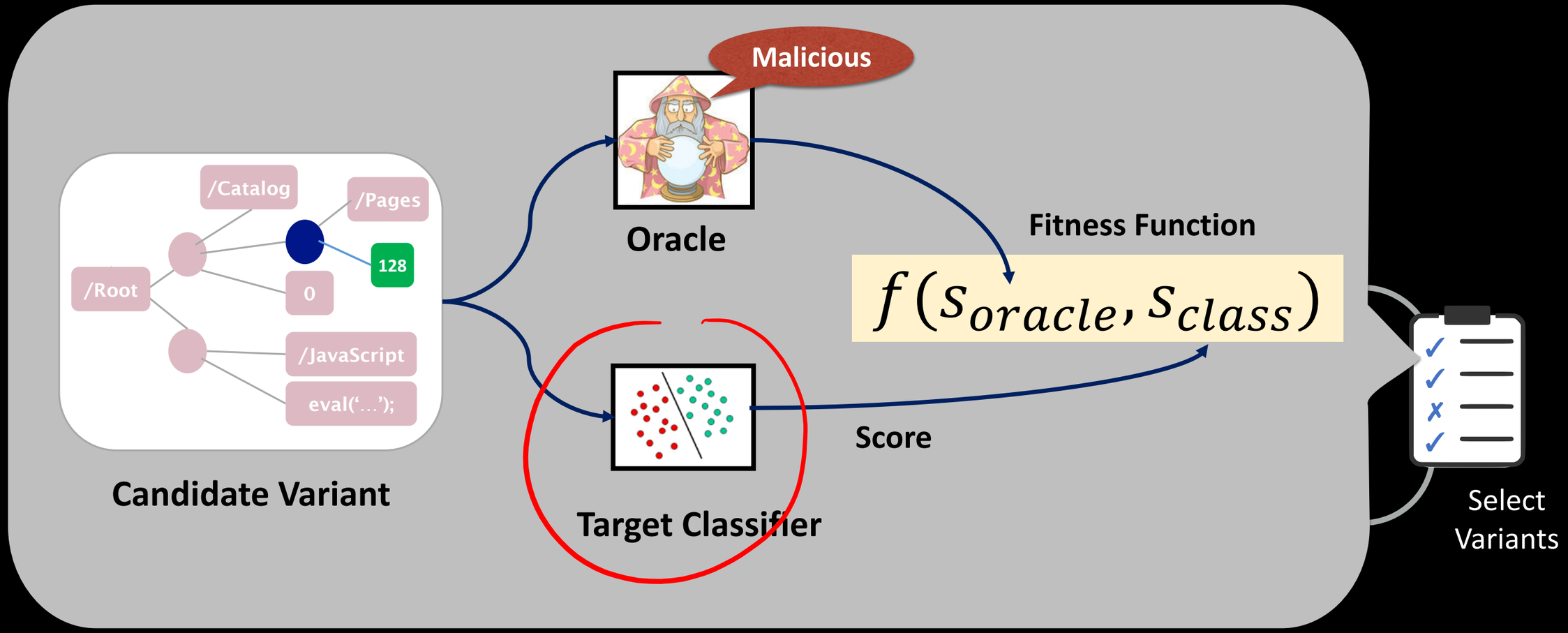
Behavioral signature:

malicious if signature matches

HTTP_URL + HOST
extracted from API traces

Advantage: we know the target malware behavior

Selecting Promising Variants



Fitness Function

Assumes lost malicious behavior will not be recovered

$$f(v) = \begin{cases} .5 - classifier\_score(v) & \text{if } oracle(v) = \text{"malicious"} \\ -\infty & \text{otherwise} \end{cases}$$

$classifier\_score \geq 0.5$: labeled **malicious**

Experimental Results

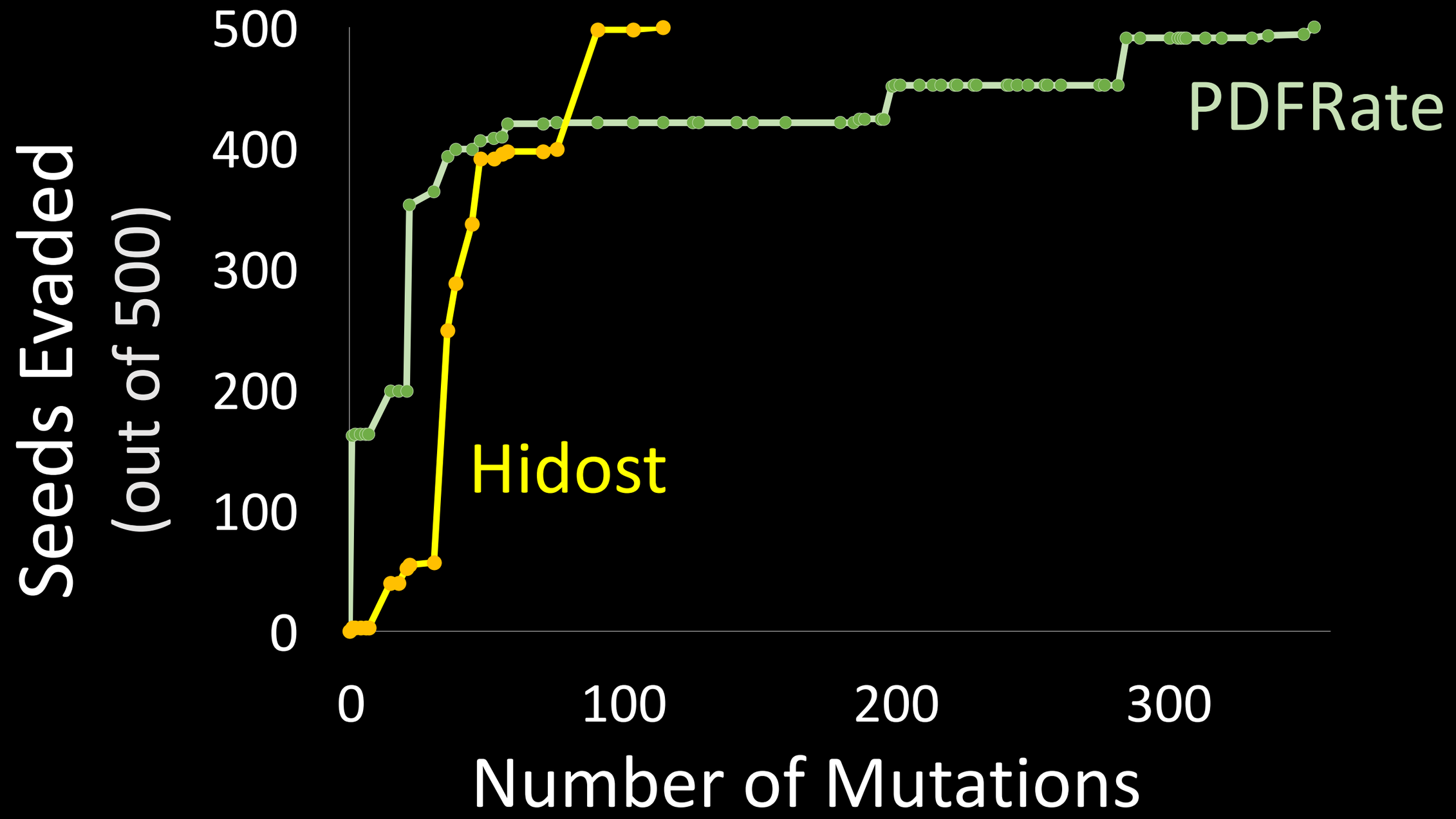
Classifier Performance

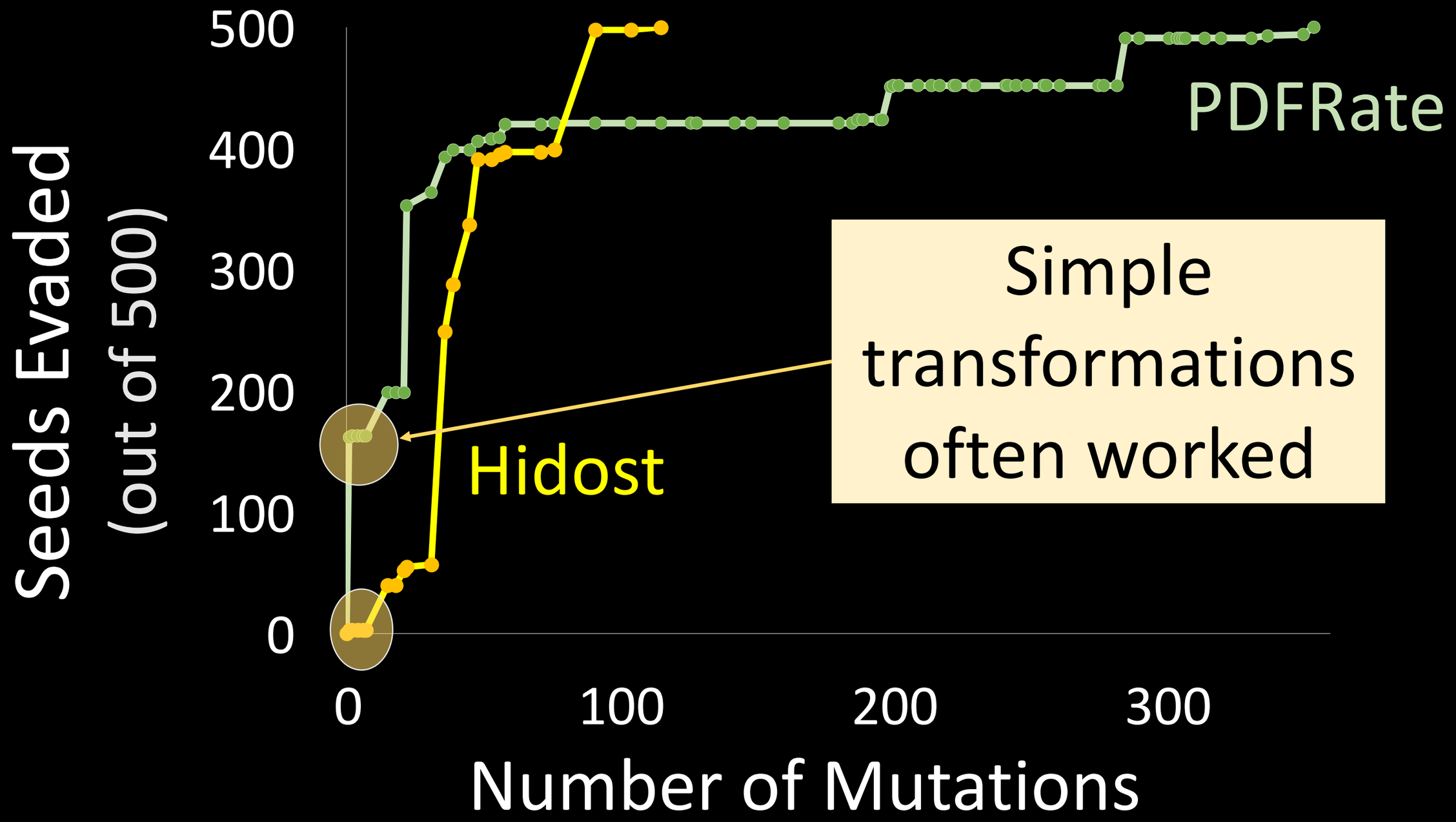
	PDFrate	Hidost
Accuracy	0.9976	0.9996
False Negative Rate	0.0000	0.0056

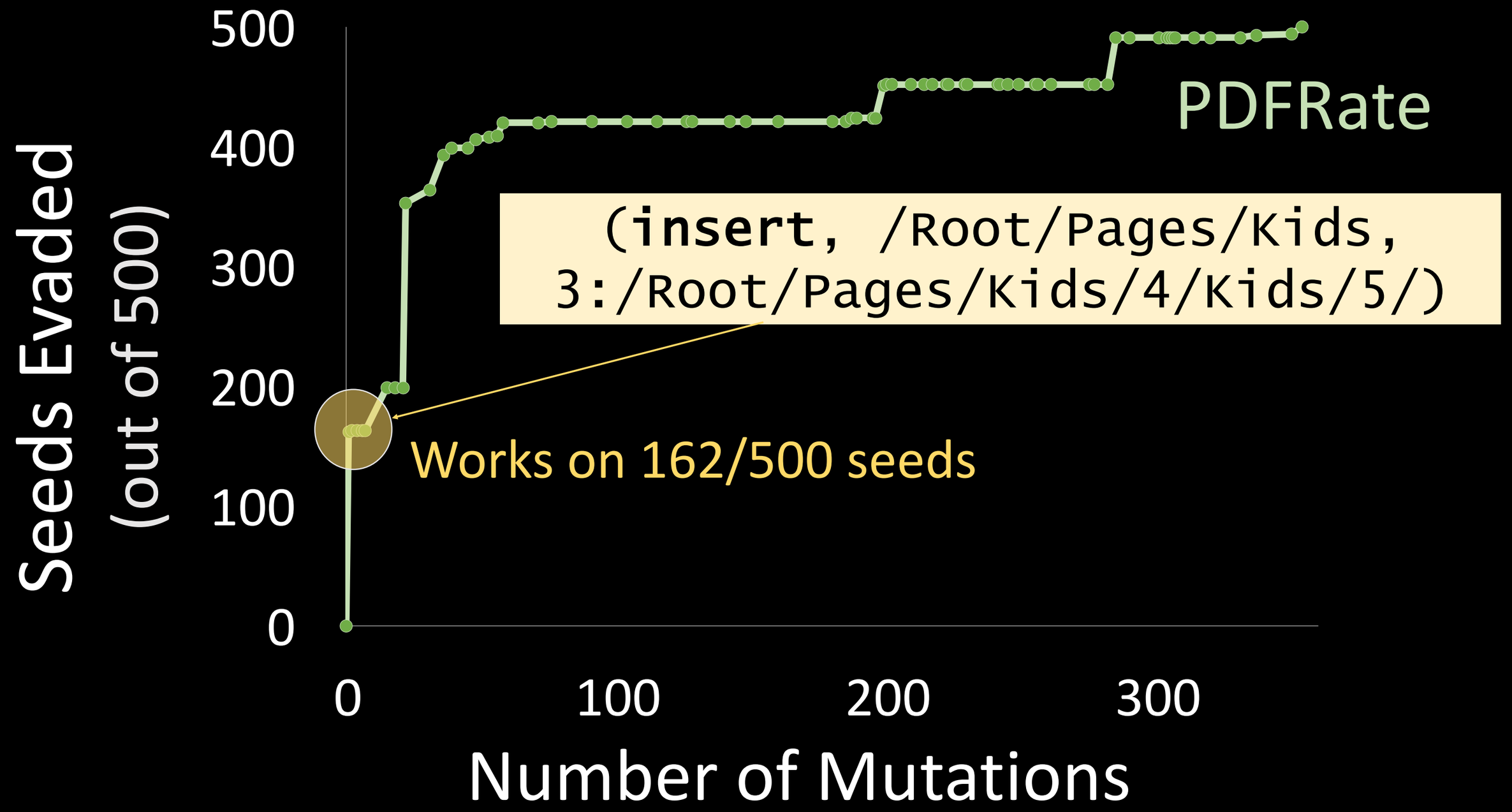
Results on non-adversarial samples

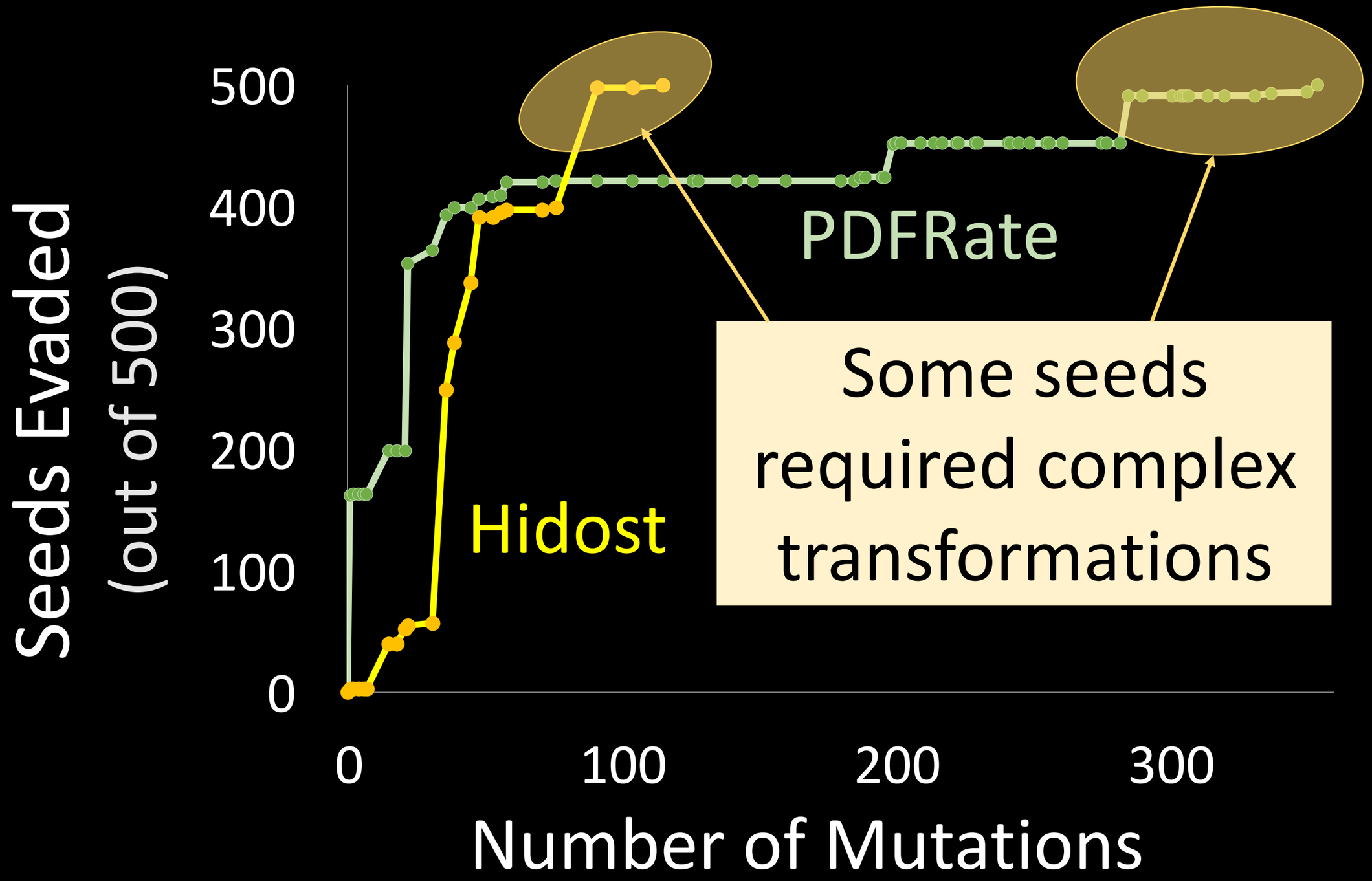
Classifier Performance

	PDFrate	Hidost
Accuracy	0.9976	0.9996
False Negative Rate	0.0000	0.0056
False Negative Rate against Adversary	1.0000	1.0000


Seeds Evaded
(out of 500)
500
400
300
200
100
0
0
100
200
300
Number of Mutations
PDFRate
Hidost


Seeds Evaded
(out of 500)
500
400
300
200
100
0
0
100
200
300
Number of Mutations
PDFRate
Hidost
Simple transformations often worked


Seeds Evaded
(out of 500)
500
400
300
200
100
0
0
100
200
300
Number of Mutations
PDFRate
(insert, /Root/Pages/Kids,
3:/Root/Pages/Kids/4/Kids/5/)
Works on 162/500 seeds


Seeds Evaded
(out of 500)
500
400
300
200
100
0
0
100
200
300
Number of Mutations
PDFRate
Hidost
Some seeds required complex transformations

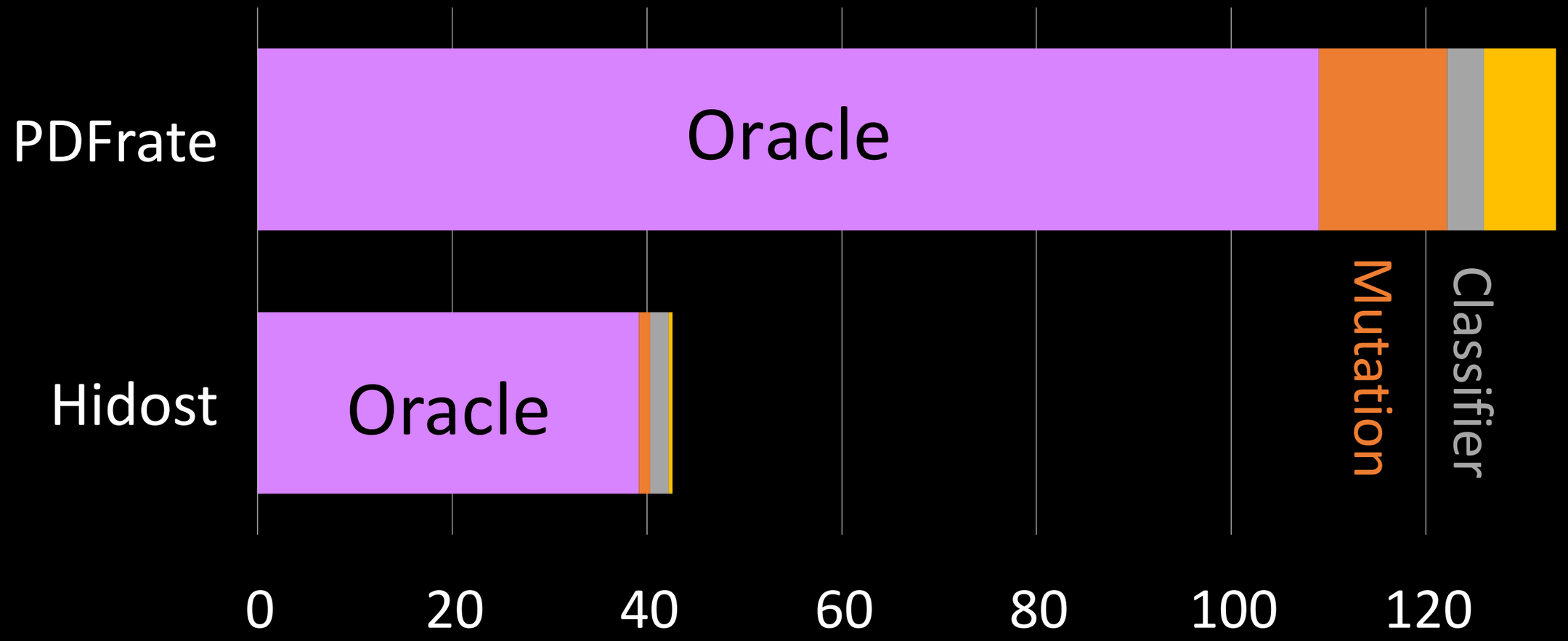
Insert: Threads, ViewerPreferences/Direction, Metadata, Metadata/Length, Metadata/Subtype, Metadata/Type, OpenAction/Contents, OpenAction/Contents/Filter, OpenAction/Contents/Length, Pages/MediaBox

Delete: AcroForm, Names/JavaScript/Names/S, AcroForm/DR/Encoding/PDFDocEncoding, AcroForm/DR/Encoding/PDFDocEncoding/Differences, AcroForm/DR/Encoding/PDFDocEncoding/Type, Pages/Rotate, AcroForm/Fields, AcroForm/DA, Outlines/Type, Outlines, Outlines/Count, Pages/Resources/ProcSet, Pages/Resources

85-step mutation trace evading Hidost

Effective for 198/500 seeds

Execution Cost

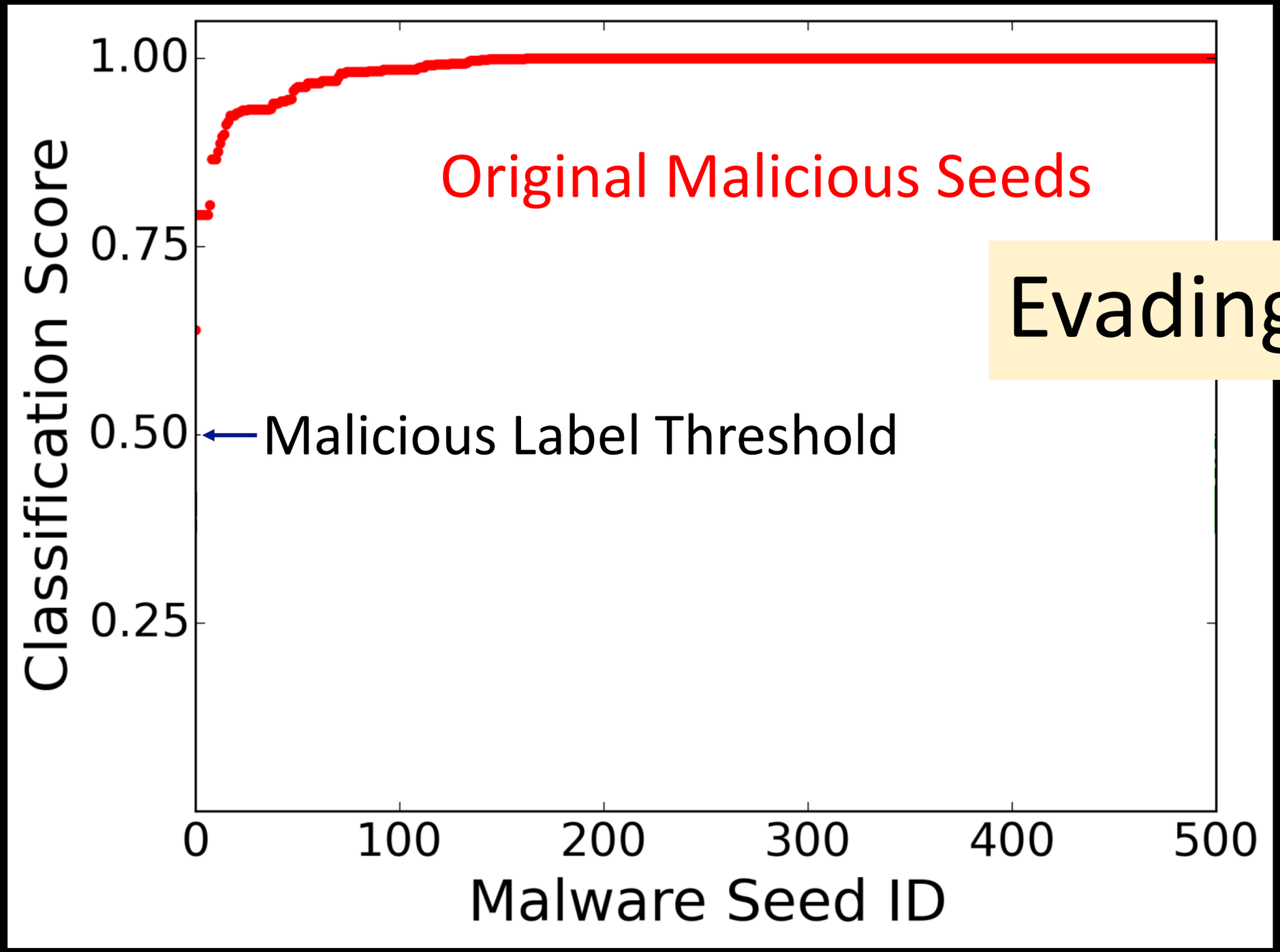


Hours to find all 500 variants on one desktop PC

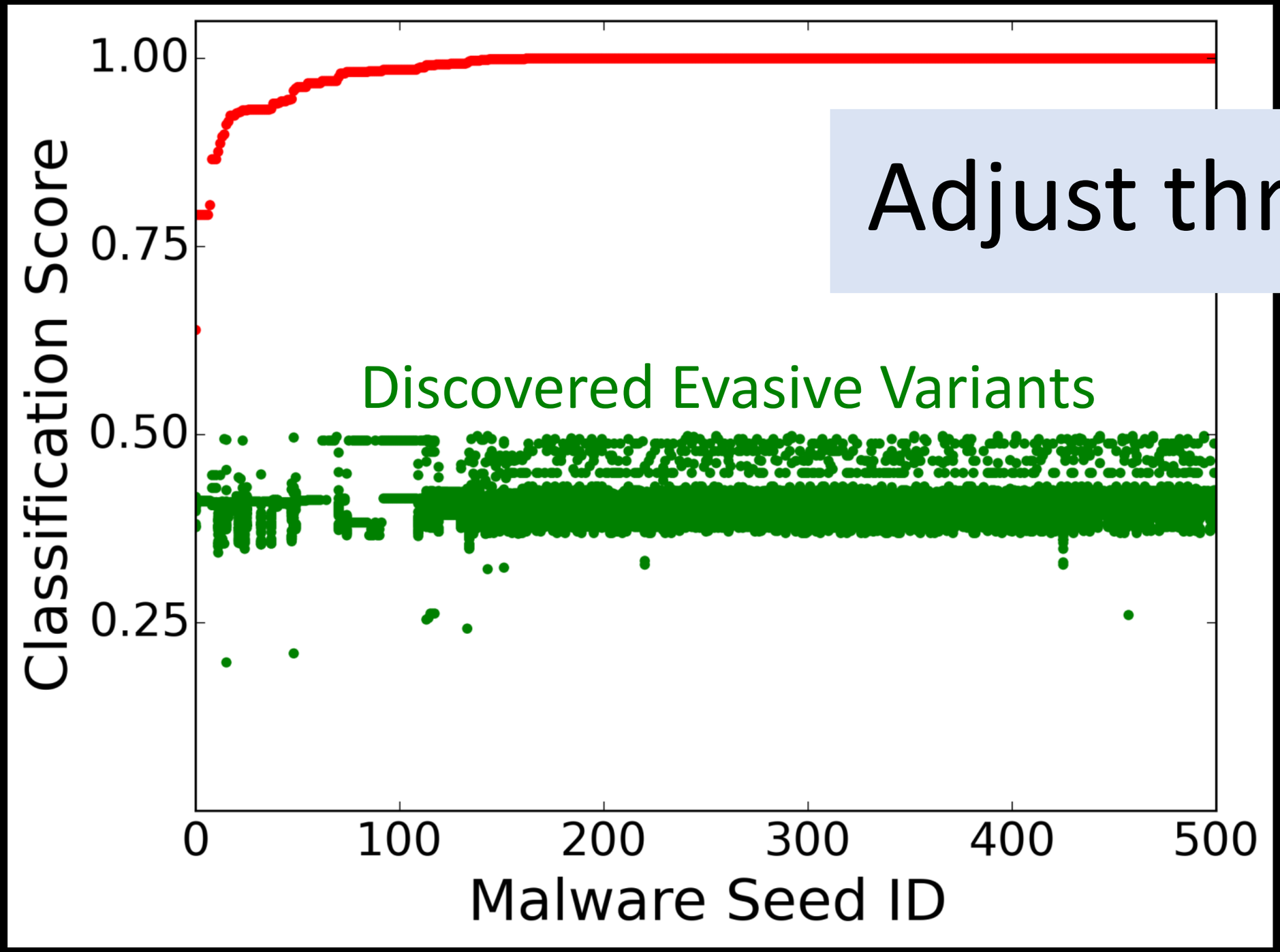
Possible Defenses

Possible Defense:
Adjust Threshold

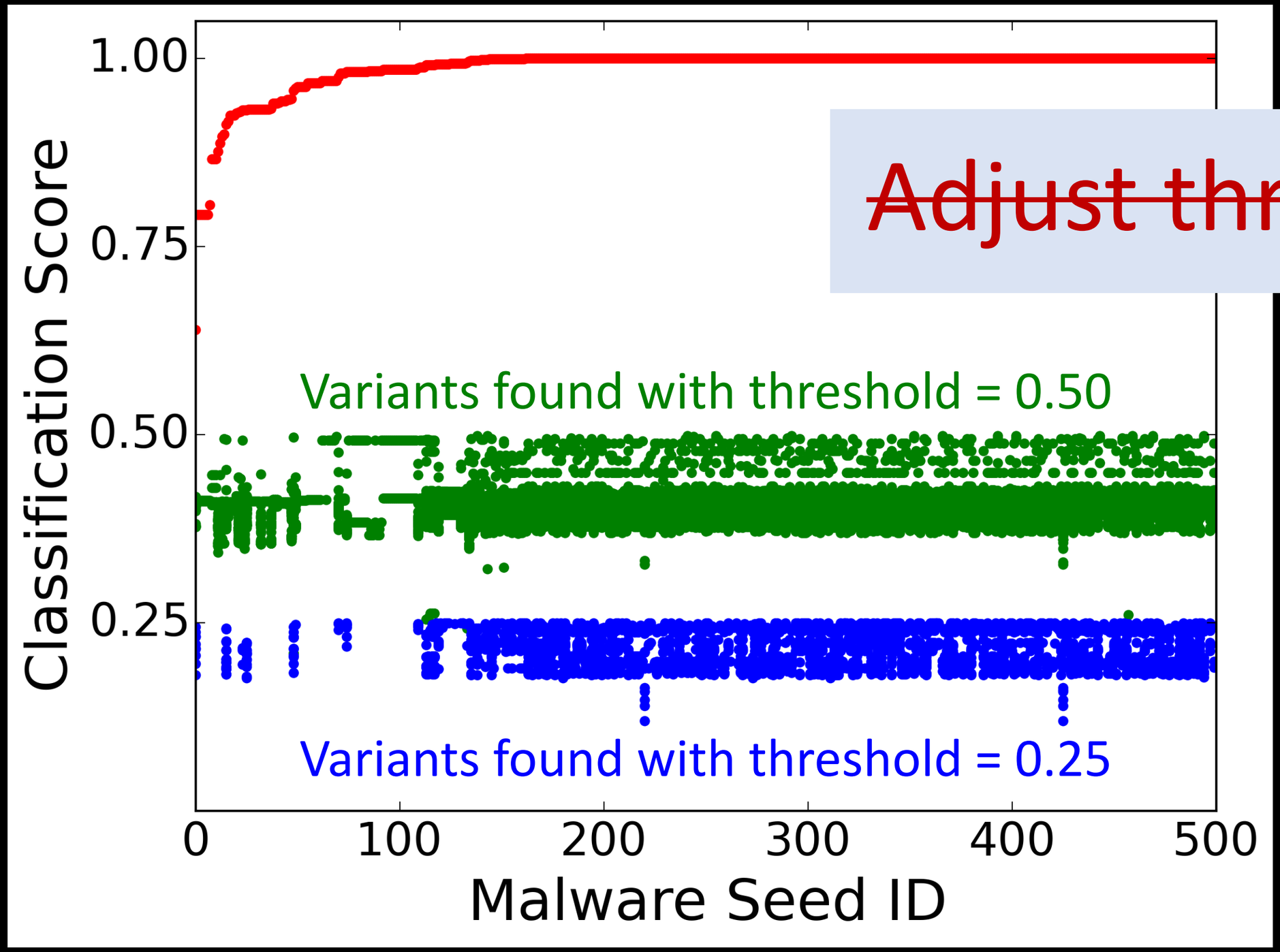
Charles Smutz, Angelos Stavrou. *When a Tree Falls: Using Diversity in Ensemble Classifiers to Identify Evasion in Malware Detectors*. NDSS 2016.



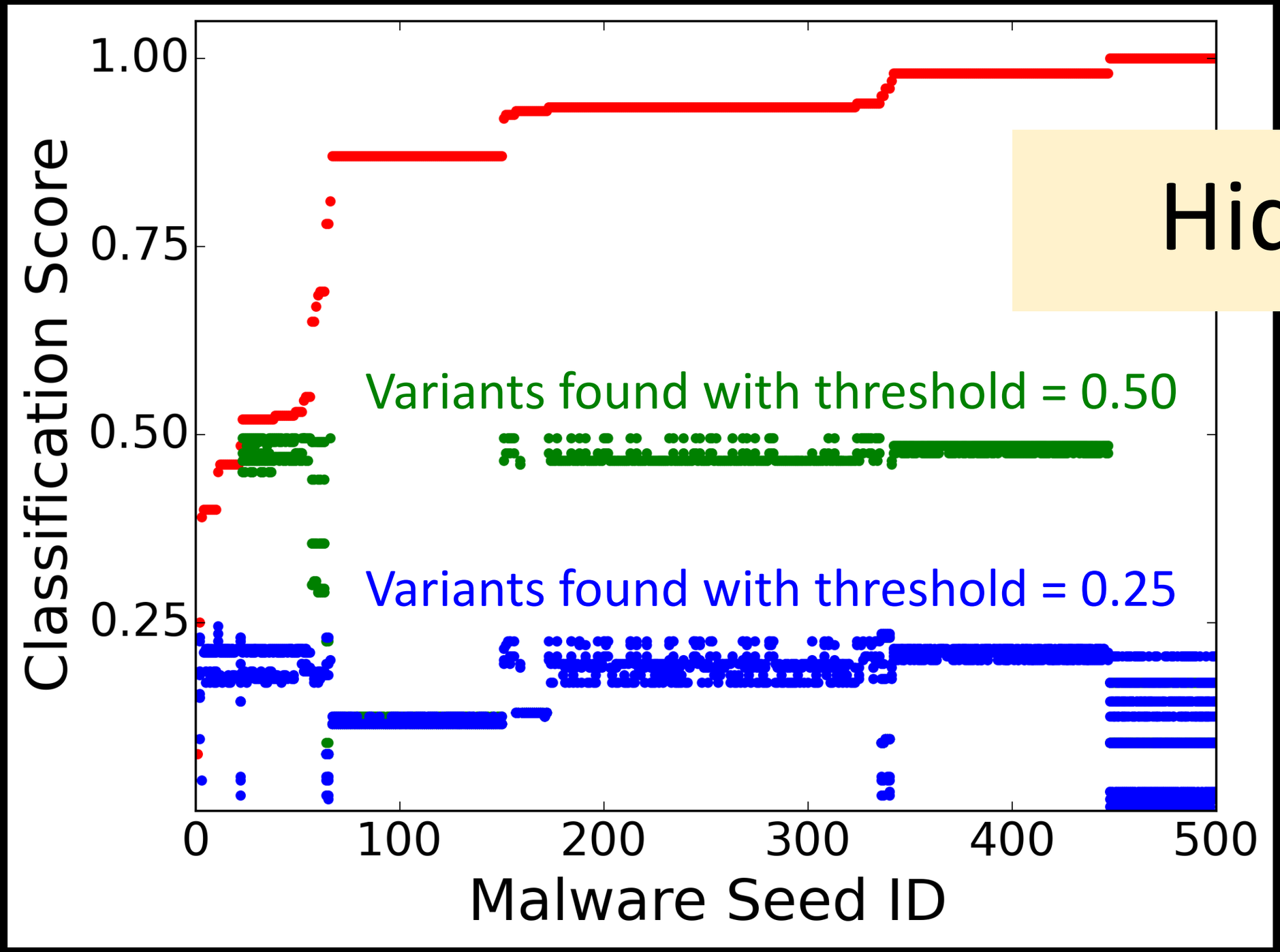
Evading PDFrate



Adjust threshold?



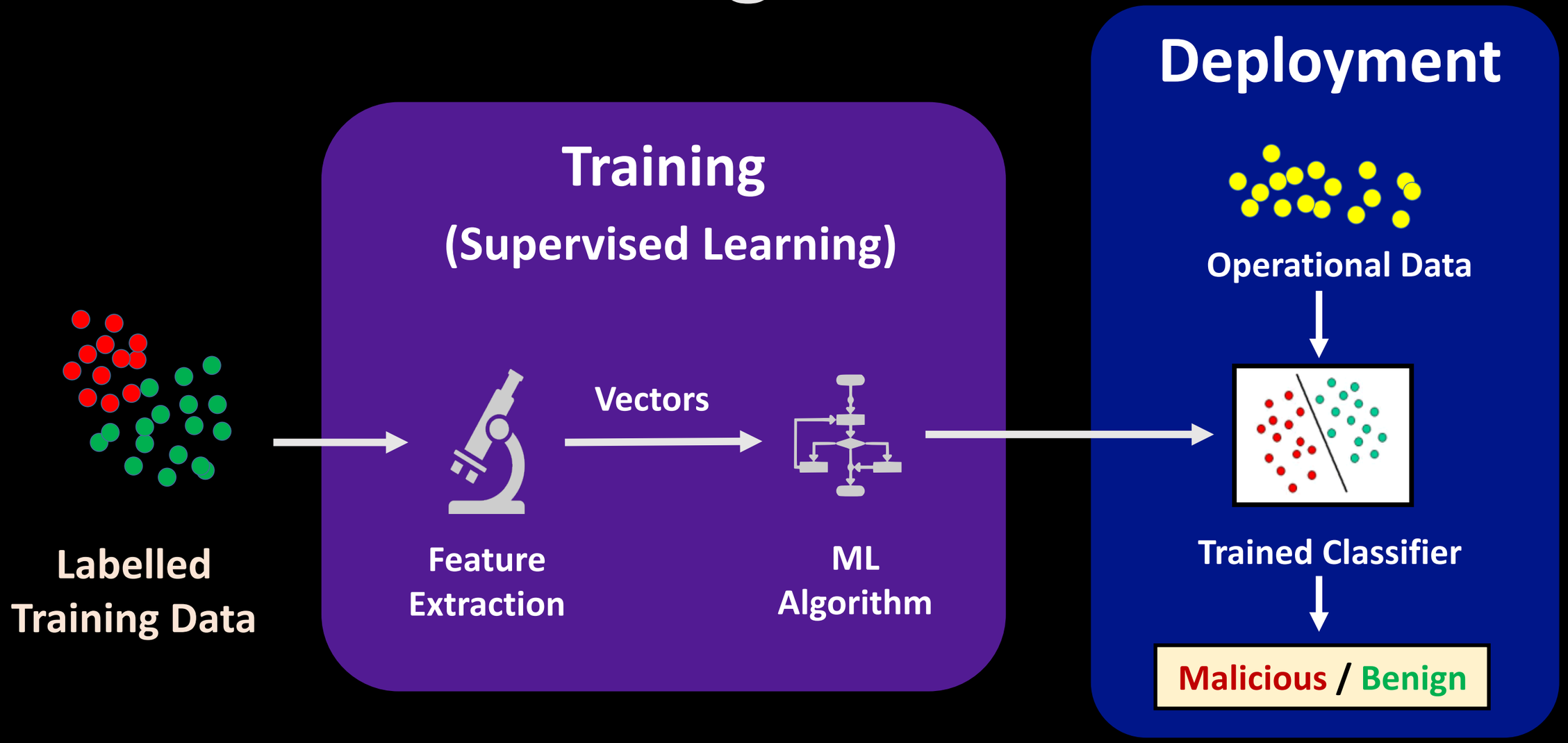
~~Adjust threshold?~~

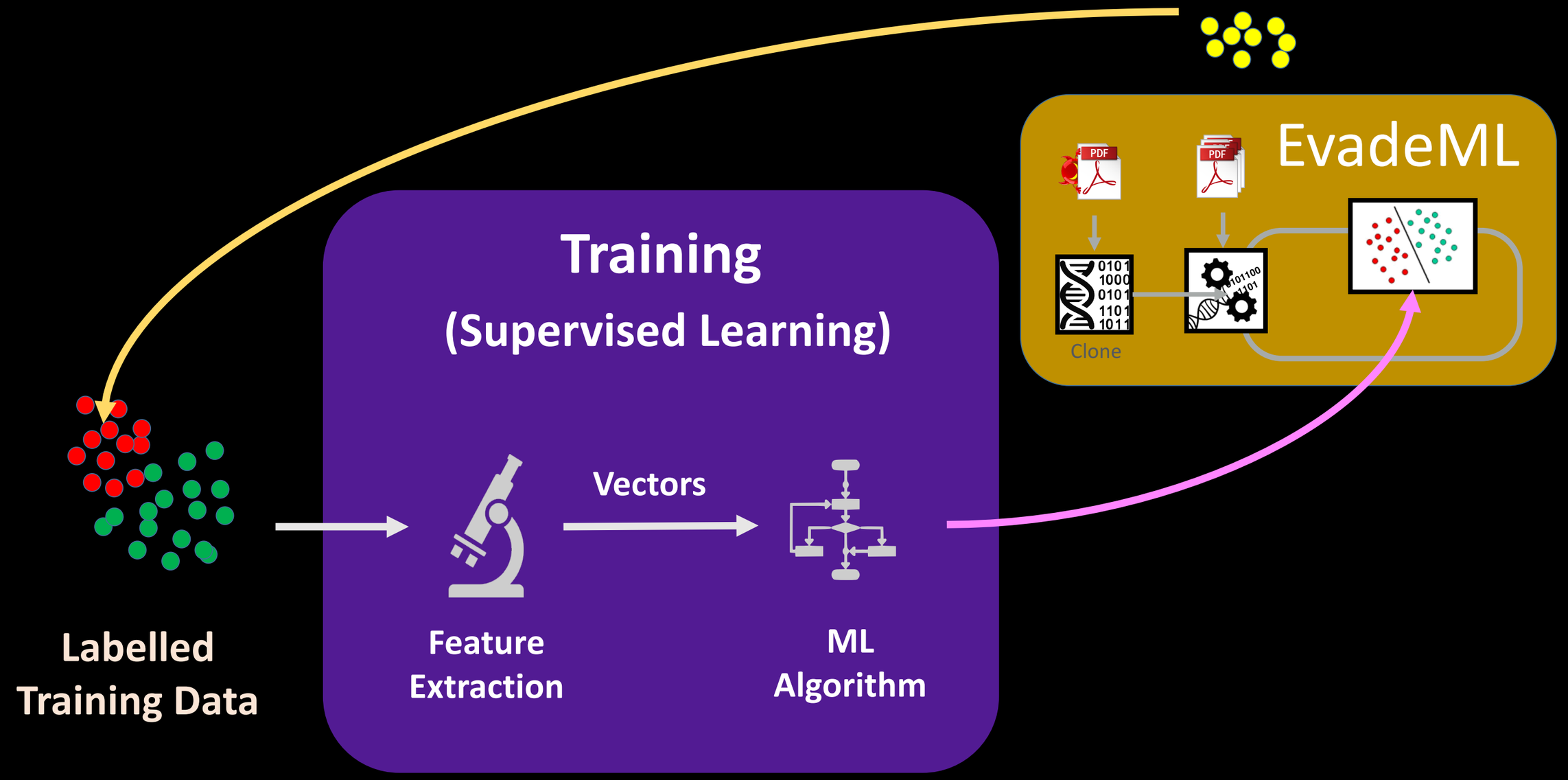

Hidost16
Variants found with threshold = 0.50
Variants found with threshold = 0.25
Classification Score
1.00
0.75
0.50
0.25
0
100
200
300
400
500
Malware Seed ID

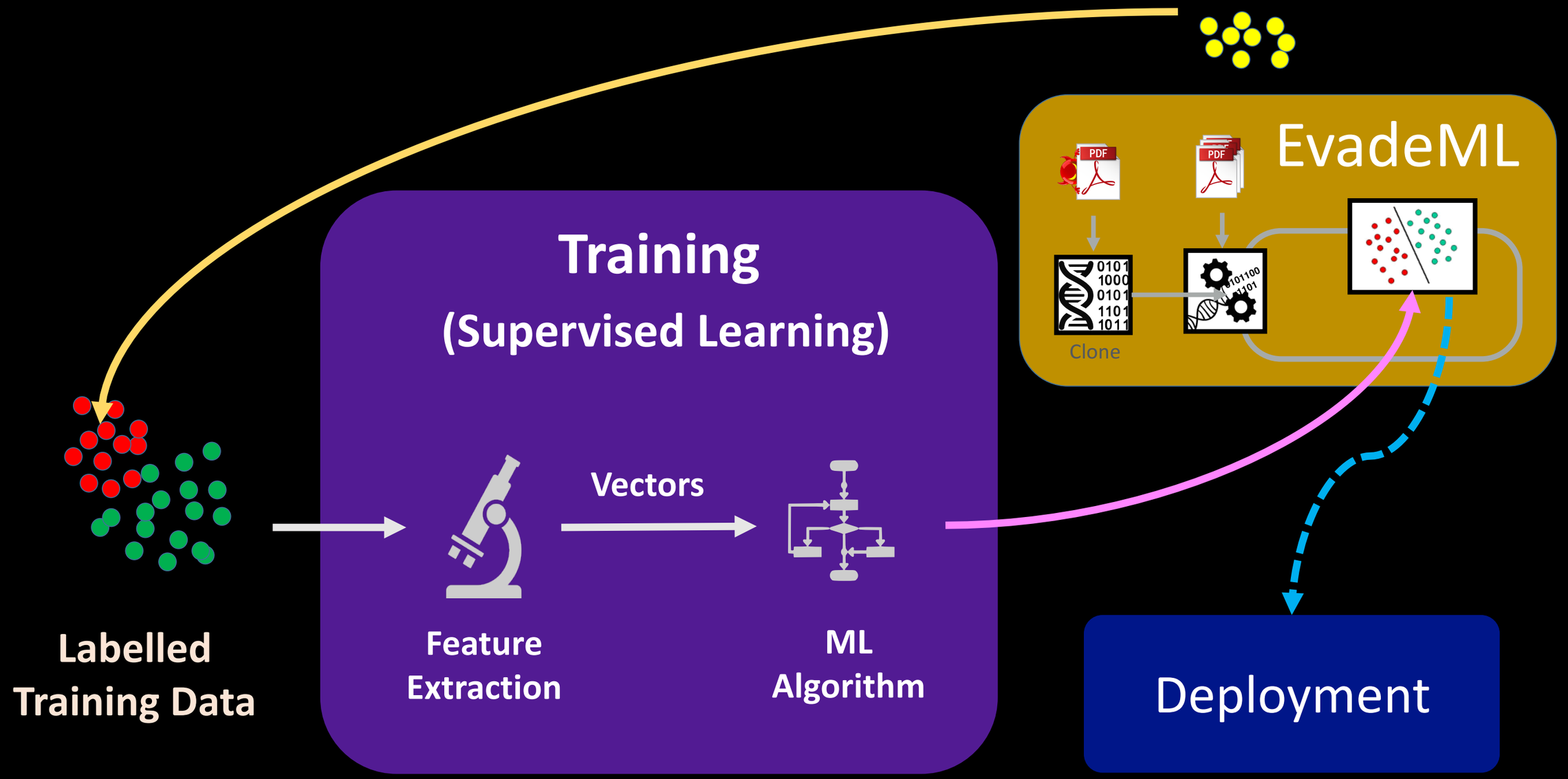
Possible Defense:

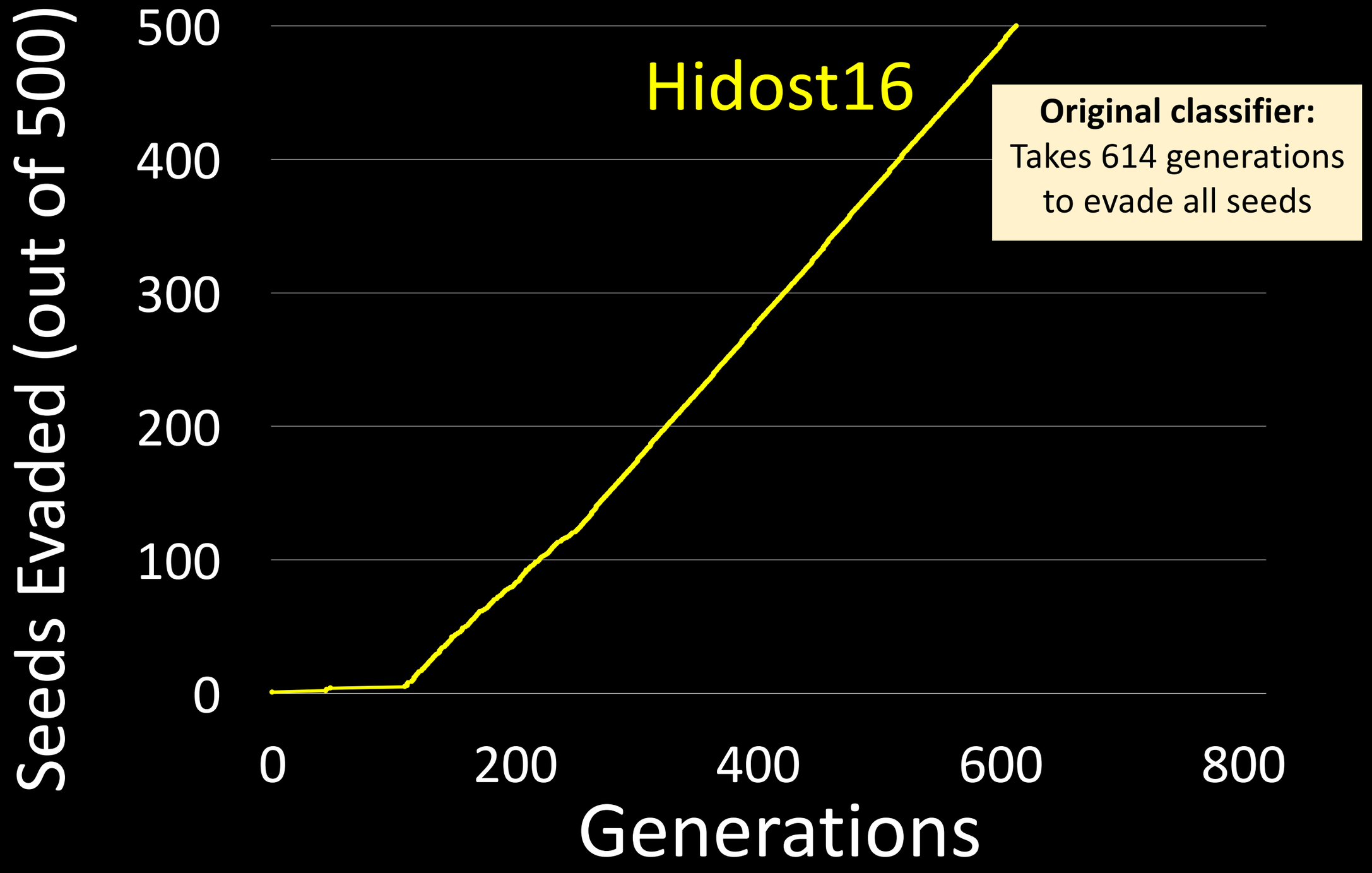
Retrain Classifier

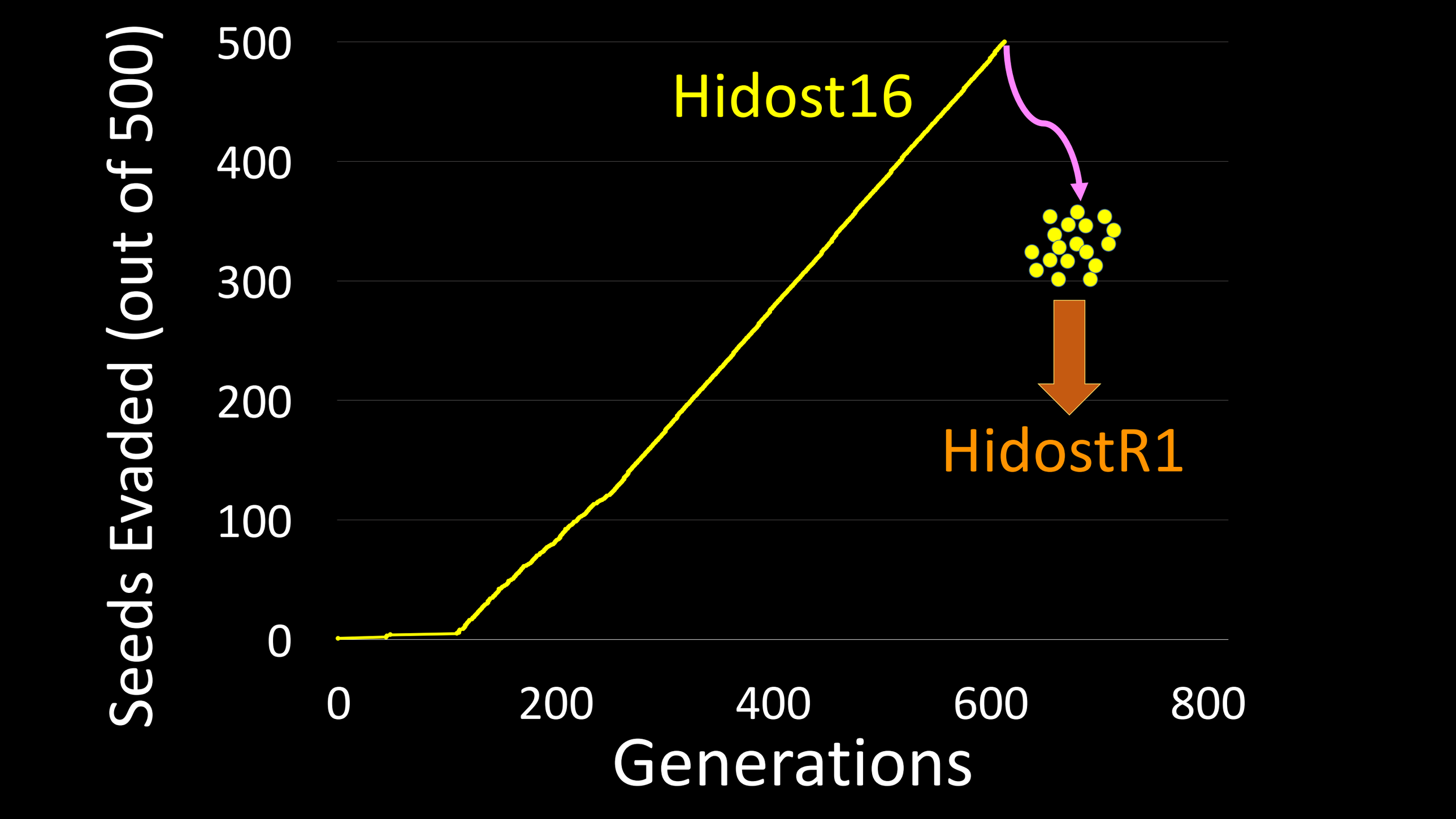
Retraining Classifier

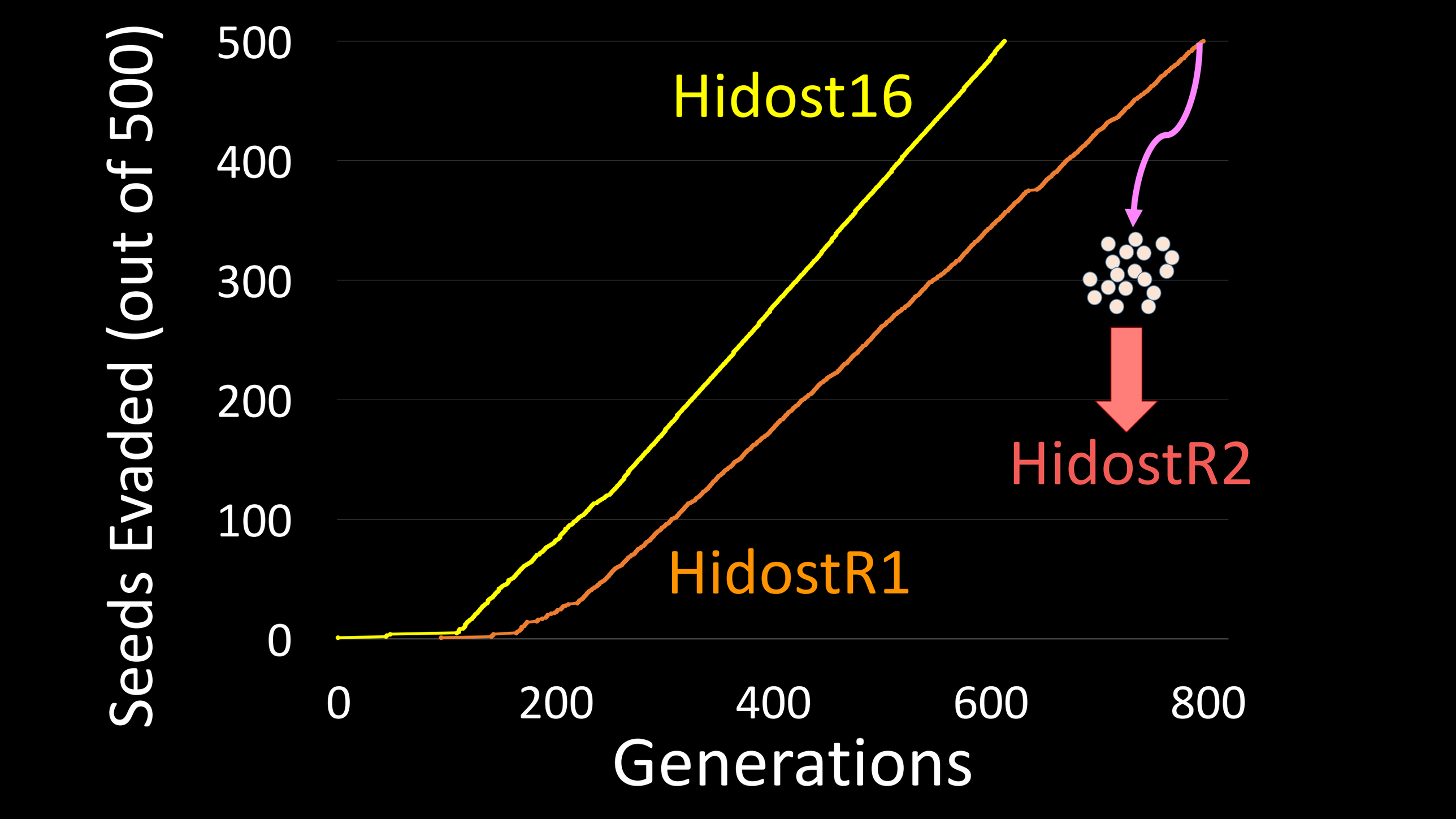


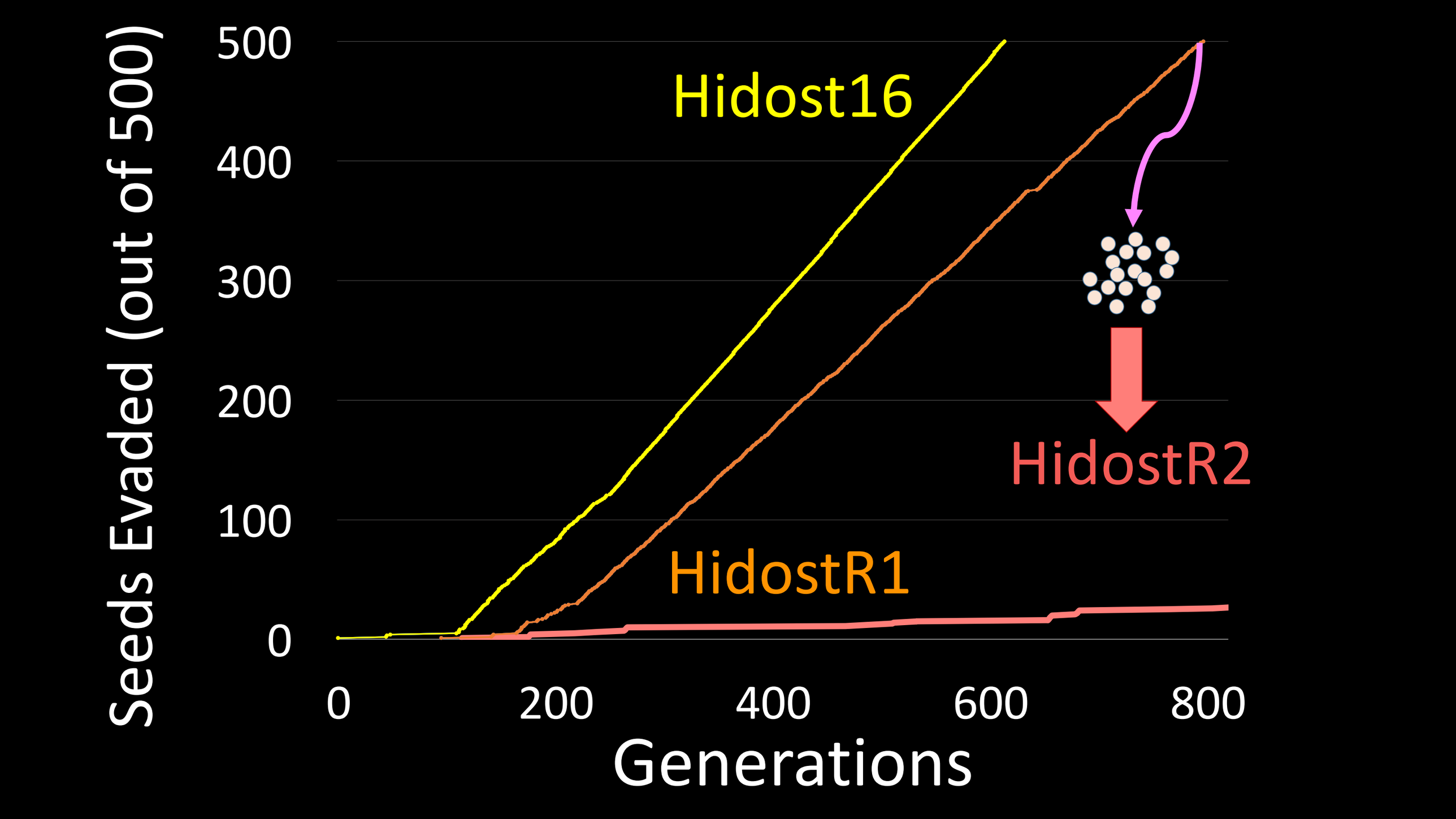

EvadeML
PDF
PDF
Clone
Training
(Supervised Learning)
Vectors
Feature
Extraction
ML
Algorithm
Labelled
Training Data

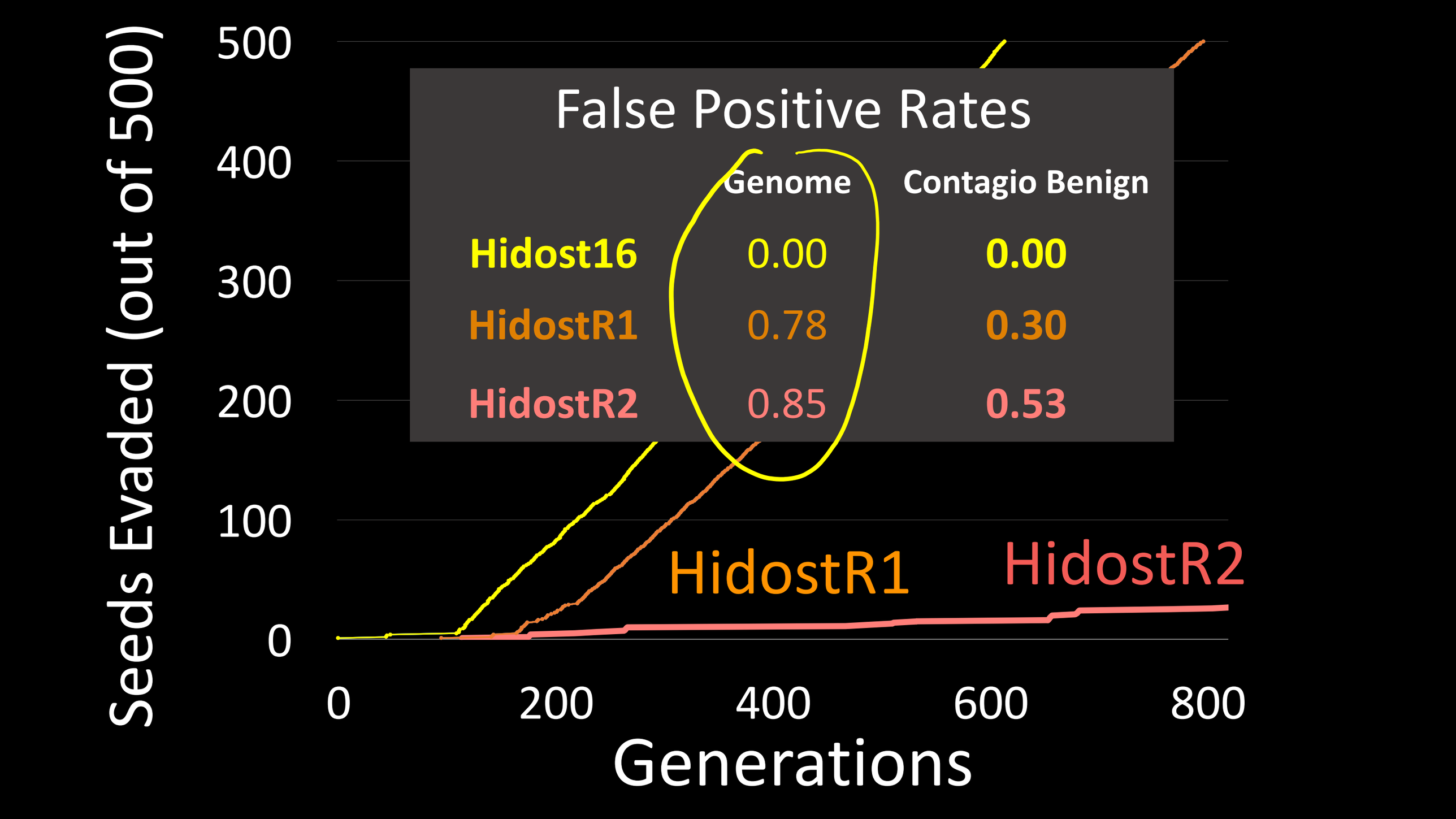

EvadeML
PDF
PDF
Clone
Training
(Supervised Learning)
Vectors
Feature
Extraction
ML
Algorithm
Labelled
Training Data
Deployment


Hidost16
Original classifier:
Takes 614 generations
to evade all seeds
Seeds Evaded (out of 500)
500
400
300
200
100
0
0
200
400
600
800
Generations


Seeds Evaded (out of 500)
500
400
300
200
100
0
0
200
400
600
800
Generations
Hidost16
HidostR1

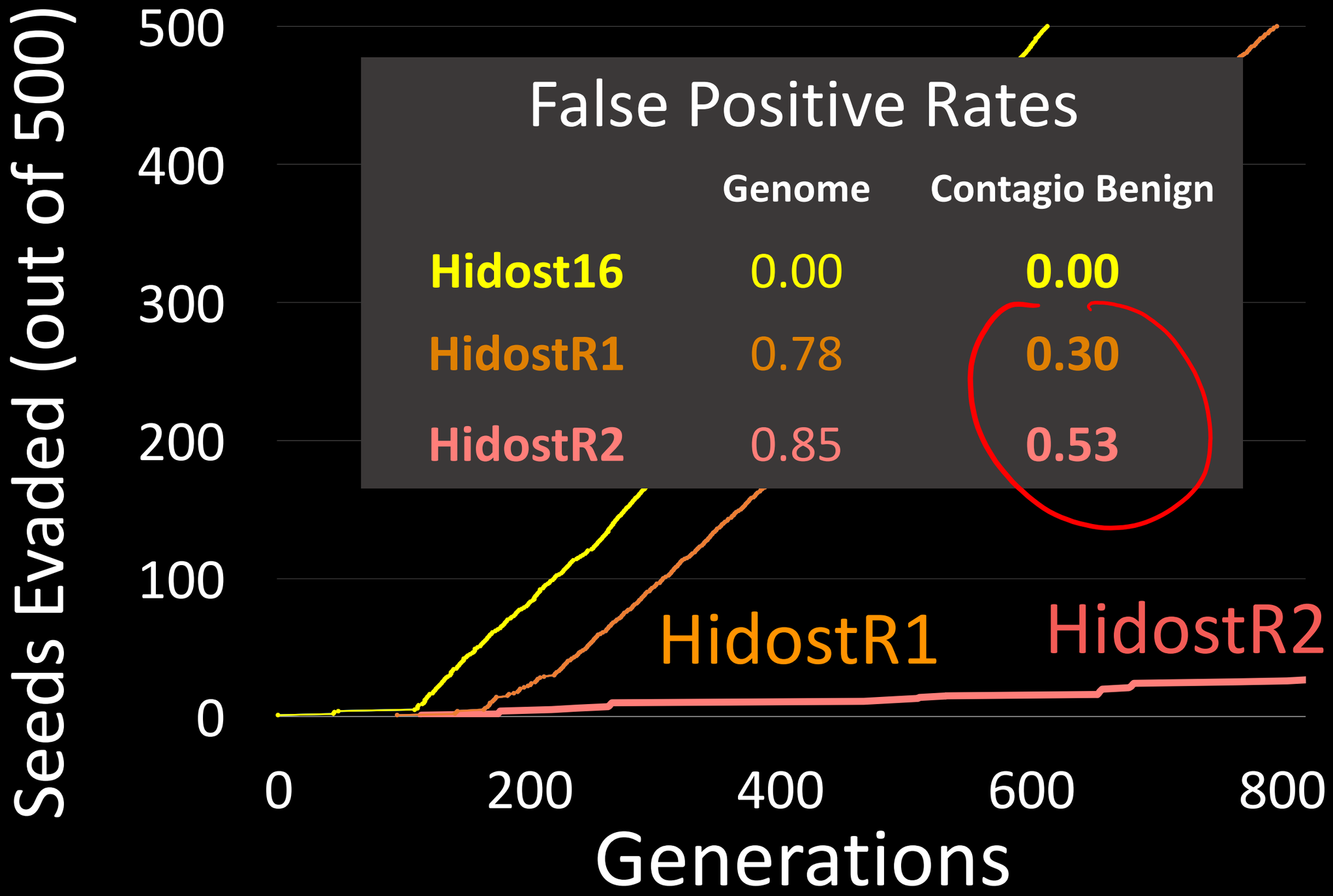

Seeds Evaded (out of 500)
500
400
300
200
100
0
0
200
400
600
800
Generations
Hidost16
HidostR1
HidostR2


Seeds Evaded (out of 500)
500
400
300
200
100
0
0
200
400
600
800
Generations
Hidost16
HidostR2
HidostR1



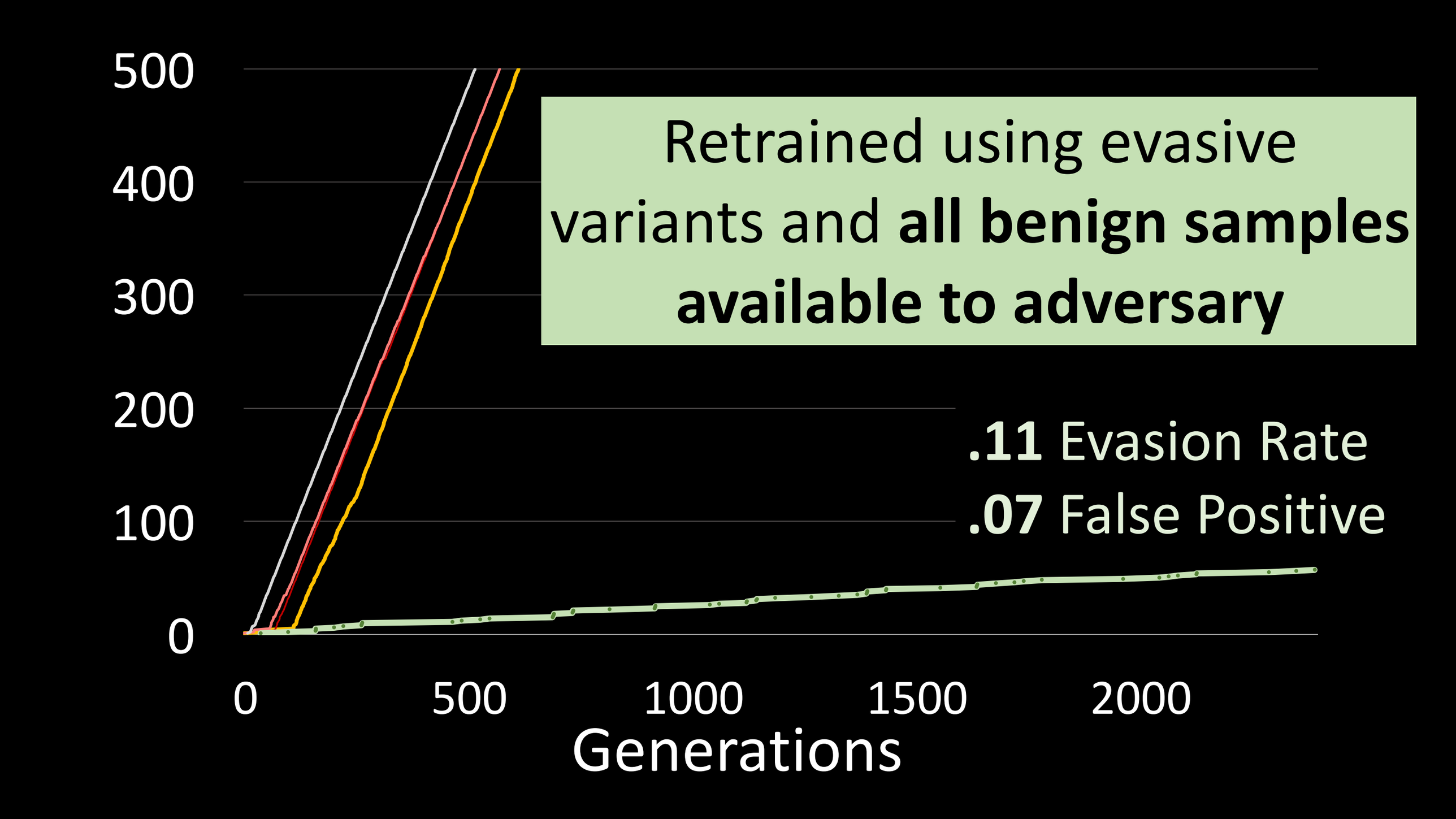
False Positive Rates

	Genome	Contagio Benign
Hidost16	0.00	**0.00**
HidostR1	0.78	**0.30**
HidostR2	0.85	**0.53**



False Positive Rates

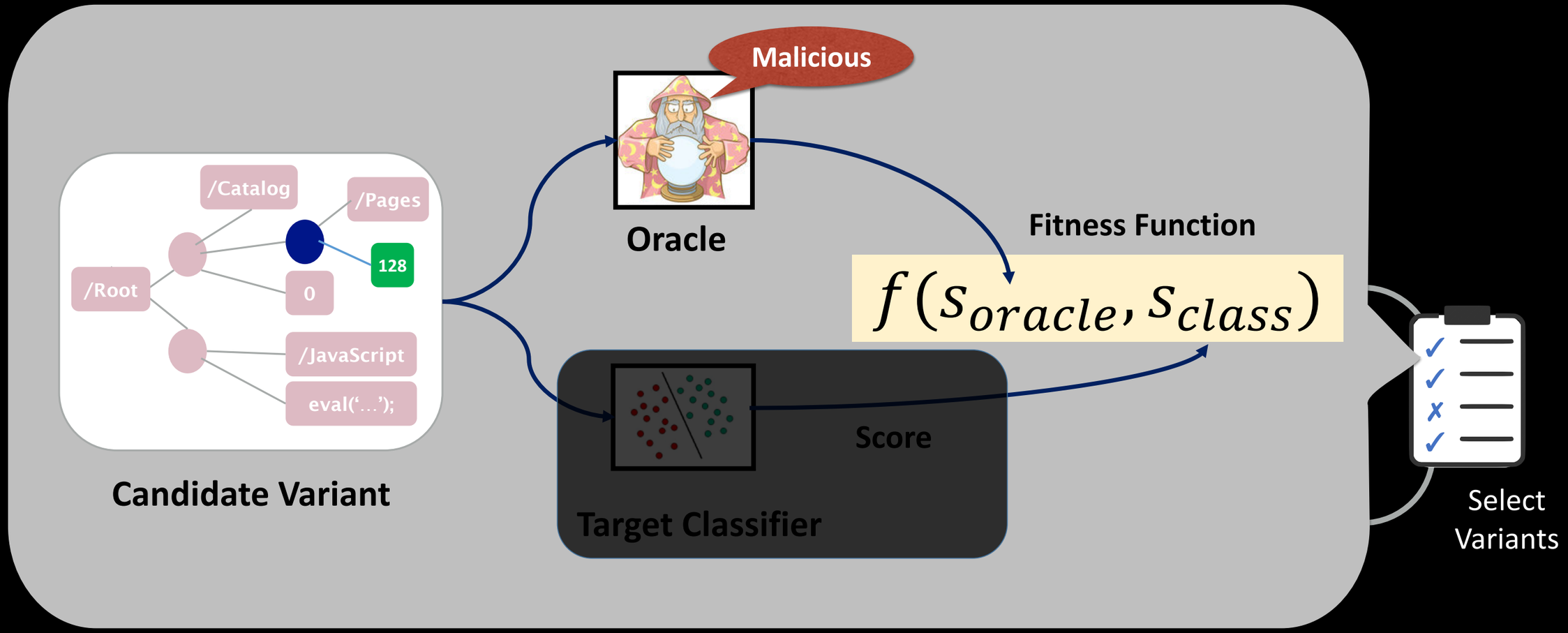
	Genome	Contagio Benign
Hidost16	0.00	**0.00**
HidostR1	0.78	**0.30**
HidostR2	0.85	**0.53**


500
400
300
200
100
0
0
500
1000
1500
2000
Generations
Retrained using evasive variants and **all benign samples available to adversary**
.11 Evasion Rate
.07 False Positive

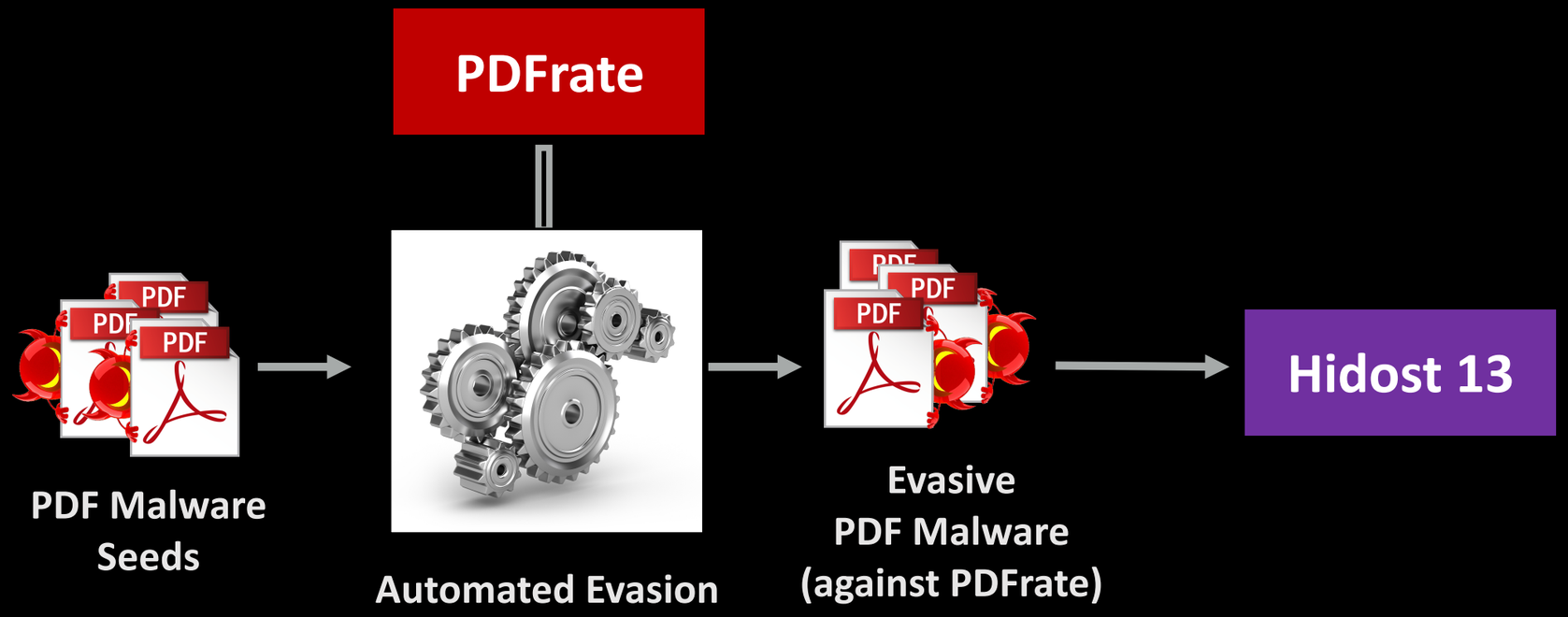
Possible Defense:
Hide Classifier

Hide Classifier

"Security Through Obscurity"



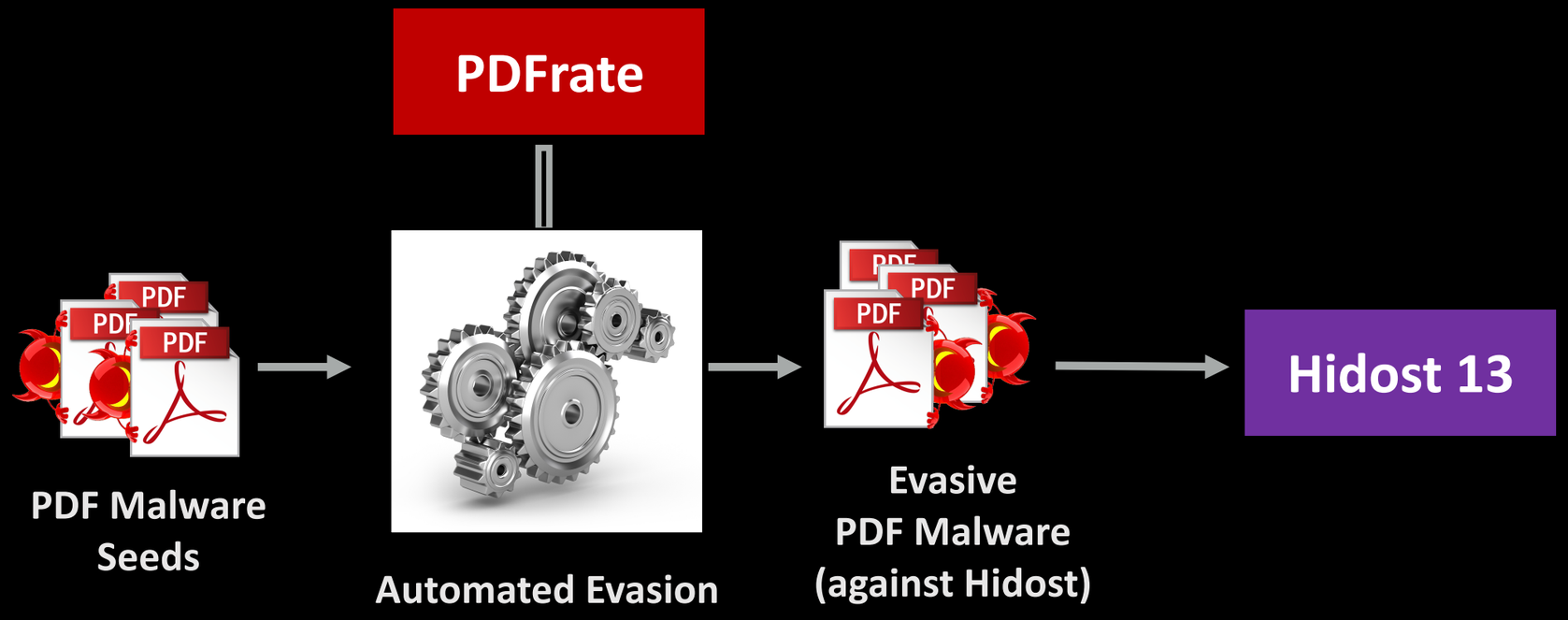
Cross-Evasion Effects



2/500 Evasive
(**0.4%** Success)

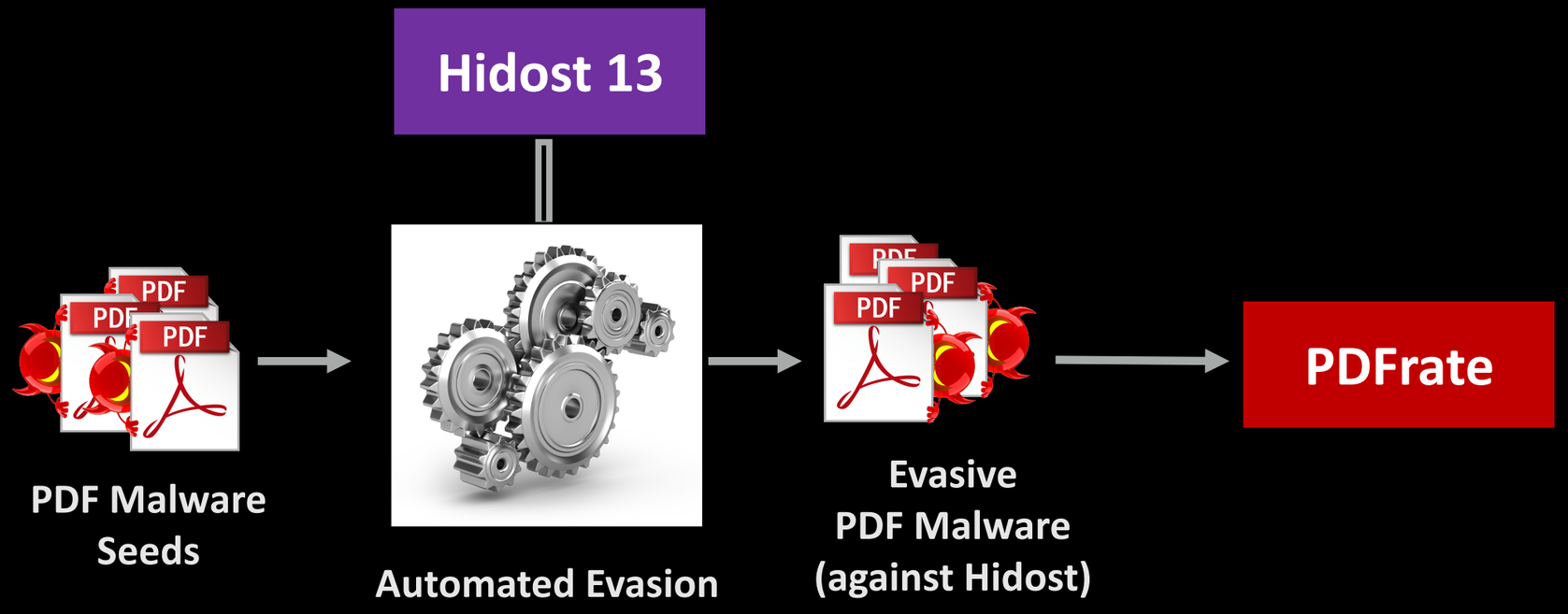
Potentially Good News?

Cross-Evasion Effects



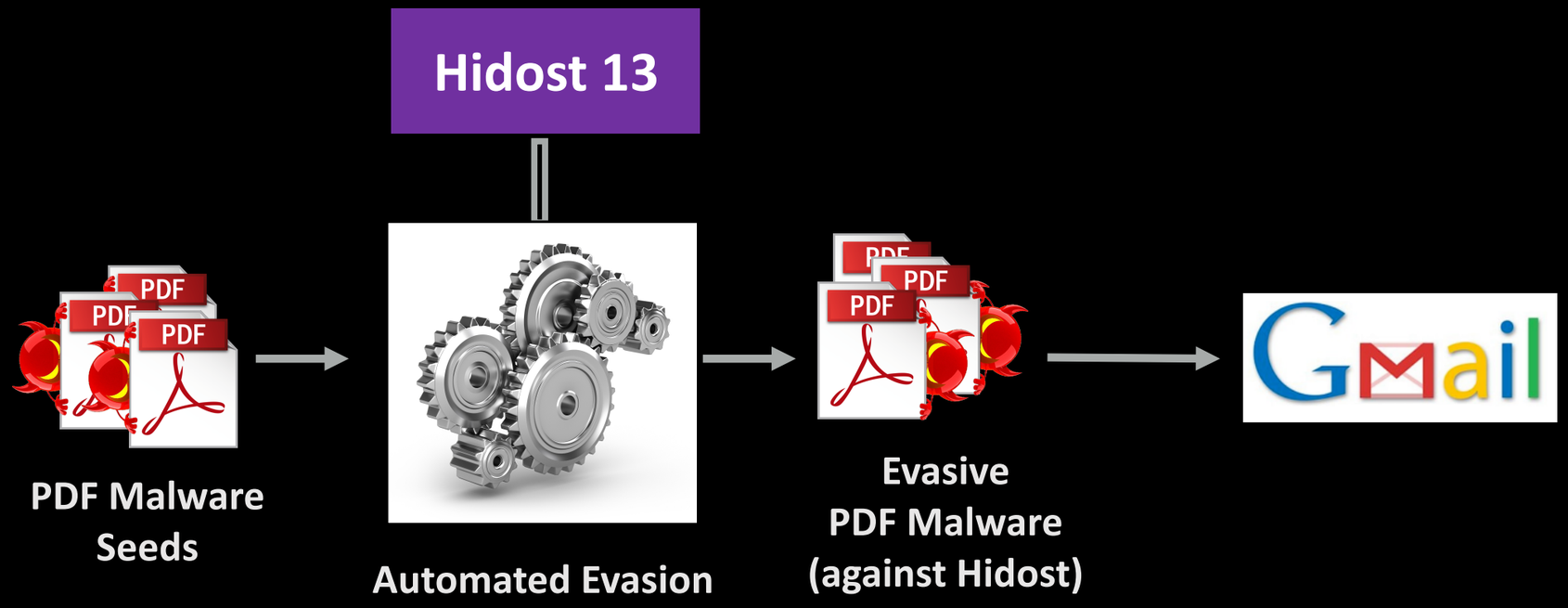
387/500 Evasive
(**77.4%** Success)

Cross-Evasion Effects



387/500 Evasive
(**77.4%** Success)

Cross-Evasion Effects



6/500 Evasive
(**0.6%** Success)

Evading Gmail's Classifier

```
for javascript in pdf.all_js:
    javascript.append_code("[redacted];“)

if pdf.get_size() < [redacted]:
    pdf.add_padding([redacted])
```

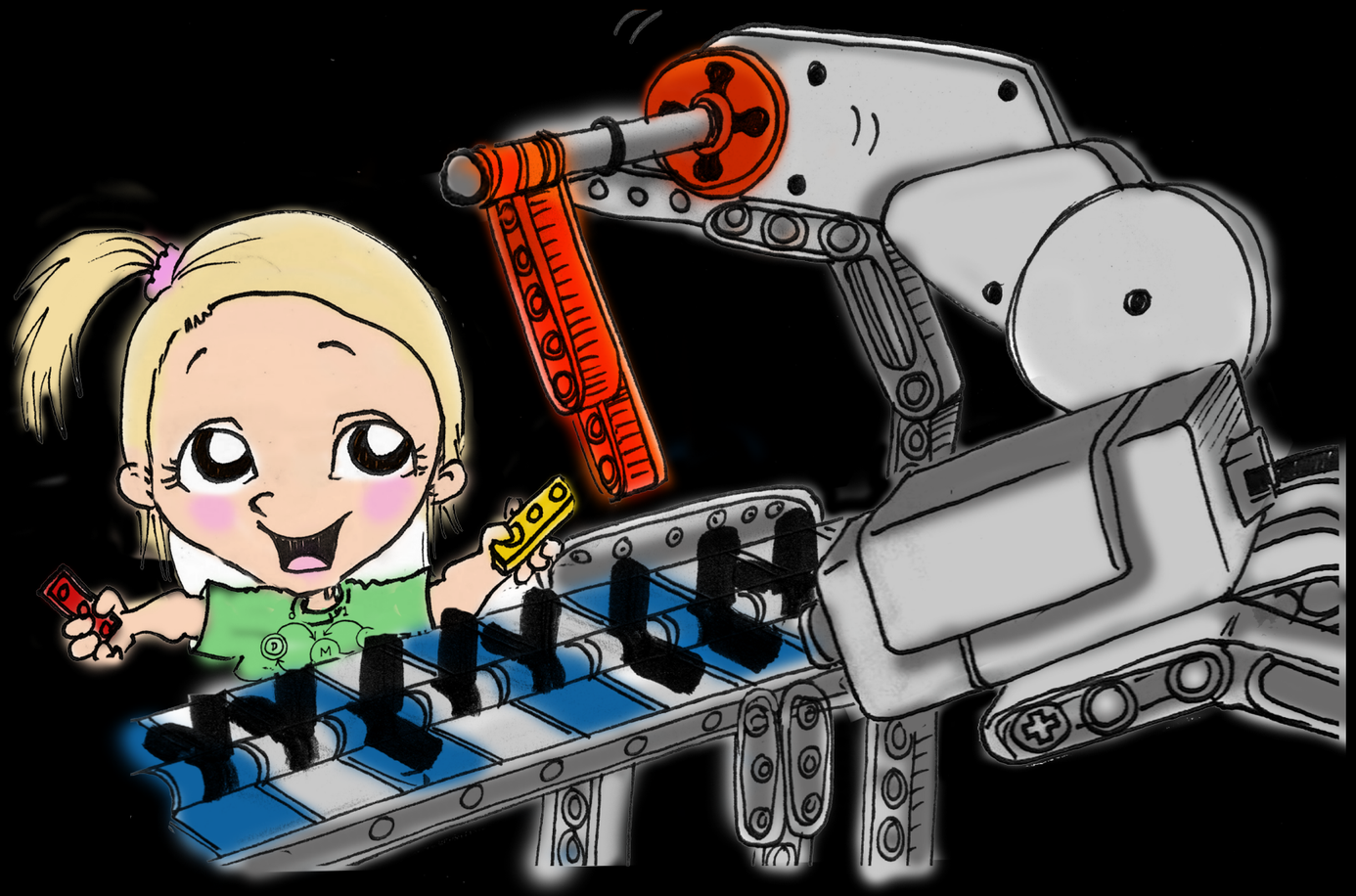
Evasion rate on Gmail: 179/380 (**47.1%**)

Conclusion

Conclusion

Domain Knowledge is Not Dead

Trust Demands Understanding



David Evans

University of Virginia

evans@virginia.edu

EvadeML.org

source code, papers

Credits: Weilin Xu, Yanjun Qi