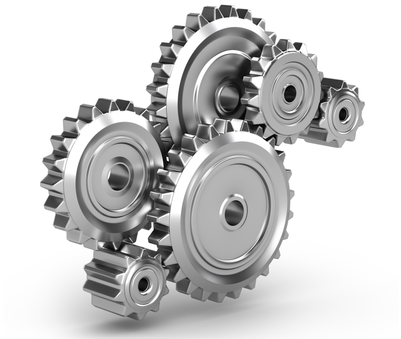
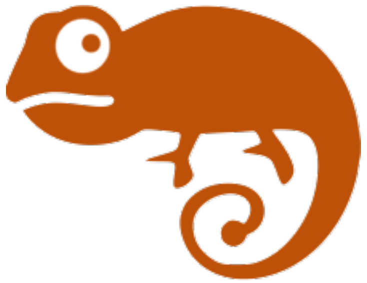
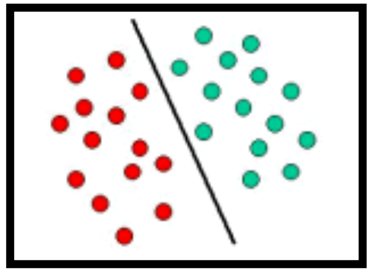
Automatically Evading Classifiers

A Case Study on PDF Malware Classifiers







Weilin Xu

David Evans

Yanjun Qi

University of Virginia

Machine Learning is Solving Our Problems

…

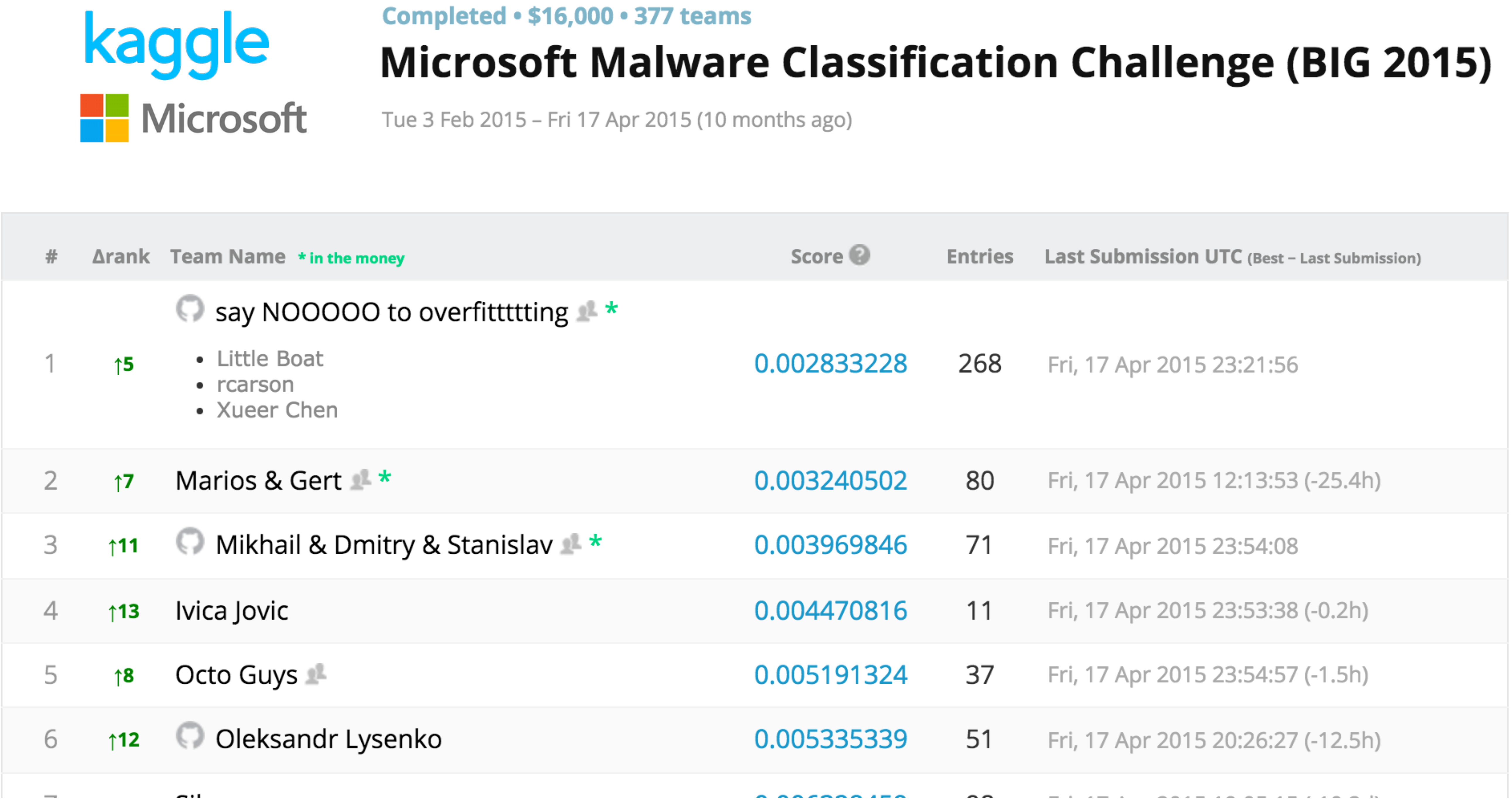
Spam

IDS

Fake Accounts

Malware

…

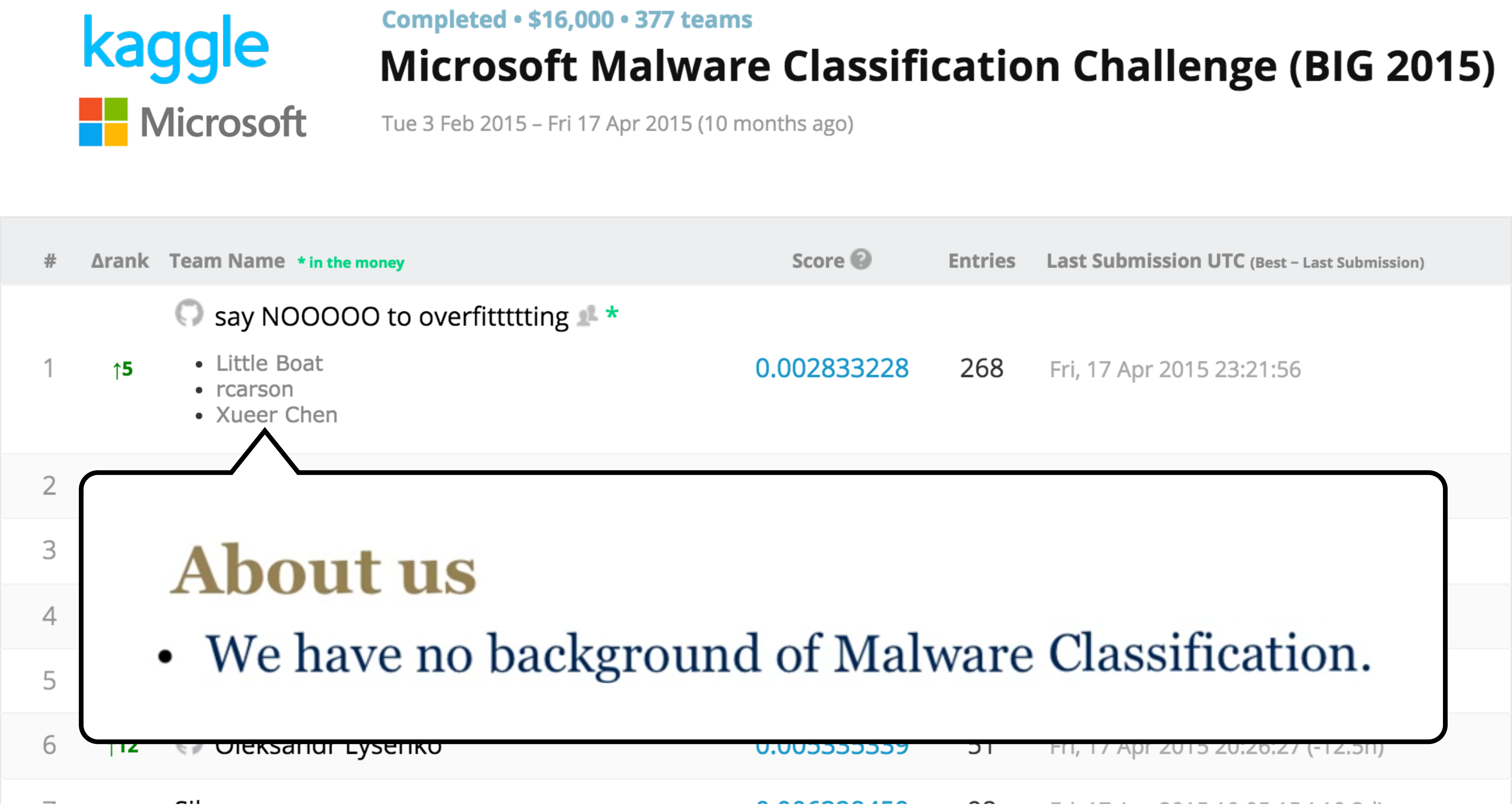


Completed • $16,000 • 377 teams

Microsoft Malware Classification Challenge (BIG 2015)

Tue 3 Feb 2015 – Fri 17 Apr 2015 (10 months ago)

#	Δrank	Team Name * in the money	Score	Entries	Last Submission UTC (Best – Last Submission)
1	↑5	say NOOOOO to overfitttting * • Little Boat • rcarson • Xueer Chen	0.002833228	268	Fri, 17 Apr 2015 23:21:56
2	↑7	Marios & Gert *	0.003240502	80	Fri, 17 Apr 2015 12:13:53 (-25.4h)
3	↑11	Mikhail & Dmitry & Stanislav *	0.003969846	71	Fri, 17 Apr 2015 23:54:08
4	↑13	Ivica Jovic	0.004470816	11	Fri, 17 Apr 2015 23:53:38 (-0.2h)
5	↑8	Octo Guys	0.005191324	37	Fri, 17 Apr 2015 23:54:57 (-1.5h)
6	↑12	Oleksandr Lysenko	0.005335339	51	Fri, 17 Apr 2015 20:26:27 (-12.5h)



Completed • $16,000 • 377 teams

Microsoft Malware Classification Challenge (BIG 2015)

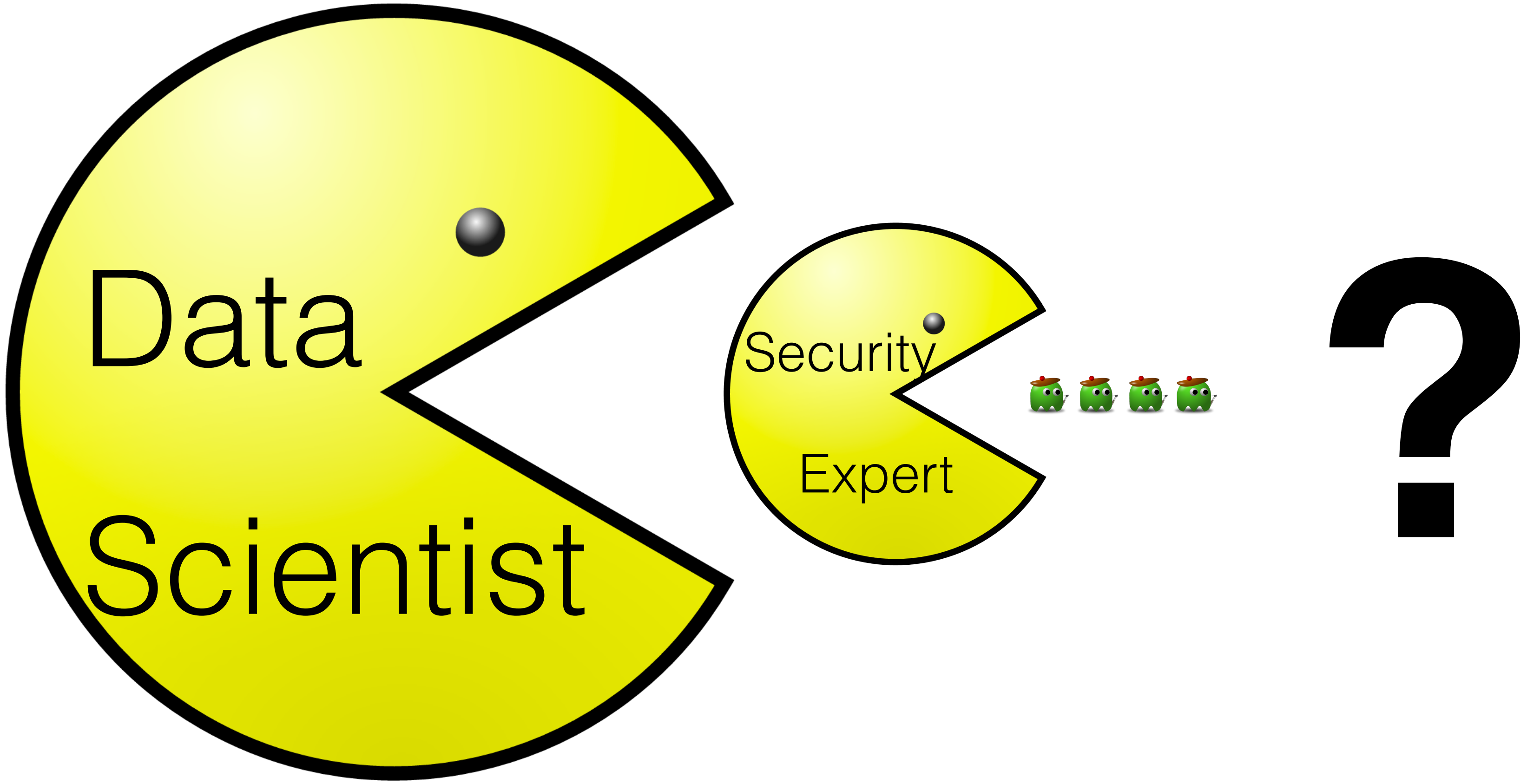
Tue 3 Feb 2015 – Fri 17 Apr 2015 (10 months ago)

#	Δrank	Team Name * in the money	Score	Entries	Last Submission UTC (Best – Last Submission)
1	↑5	say NOOOOO to overfittttting * • Little Boat • rcarson • Xueer Chen	0.002833228	268	Fri, 17 Apr 2015 23:21:56
2					
3					
4					
5					
6	↑12	Oleksandr Lysenko	0.005333339	51	Fri, 17 Apr 2015 20:26:27 (-12.5h)

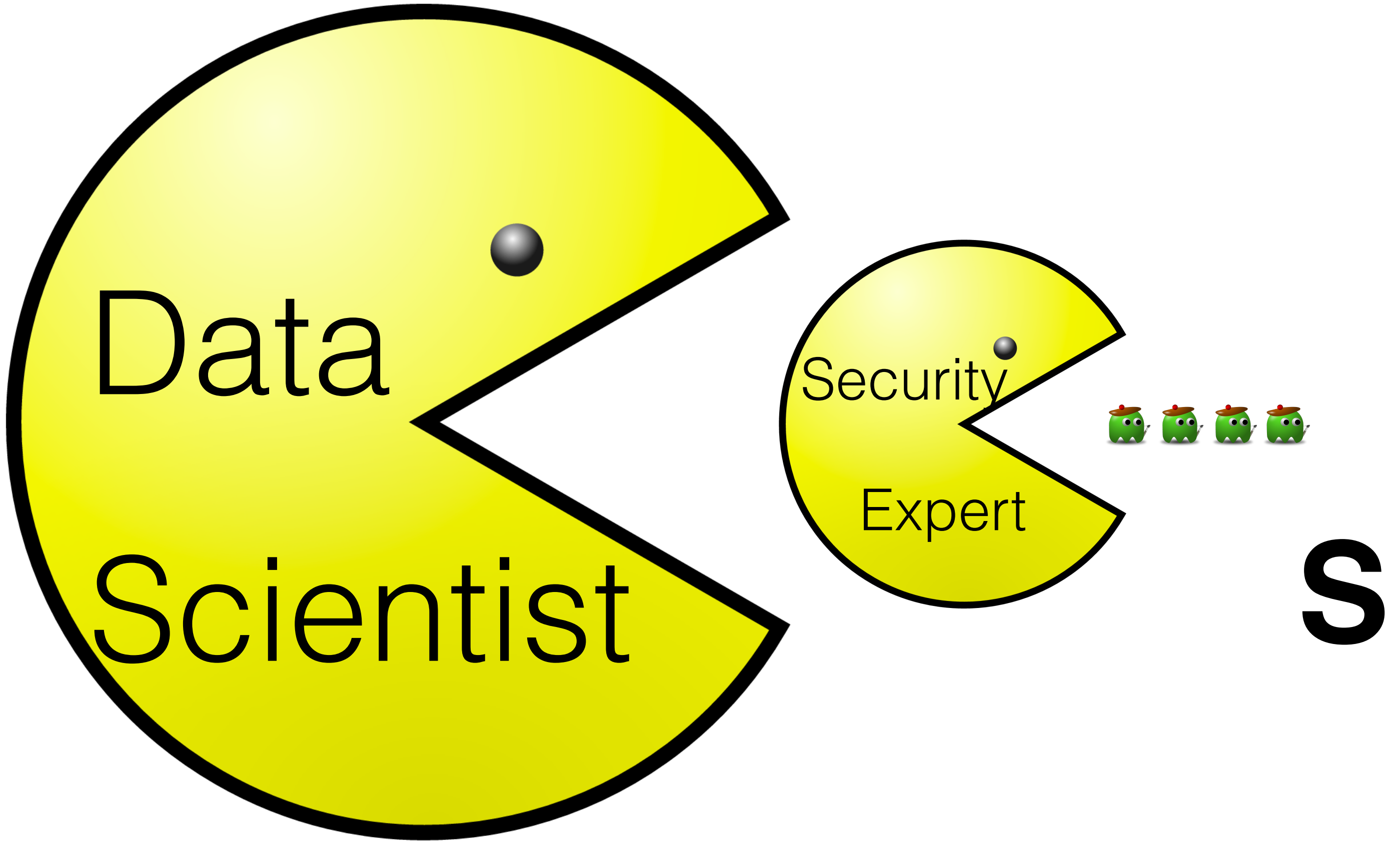
About us

- We have no background of Malware Classification.

Machine Learning is Eating the World



Machine Learning is Eating the World



No!

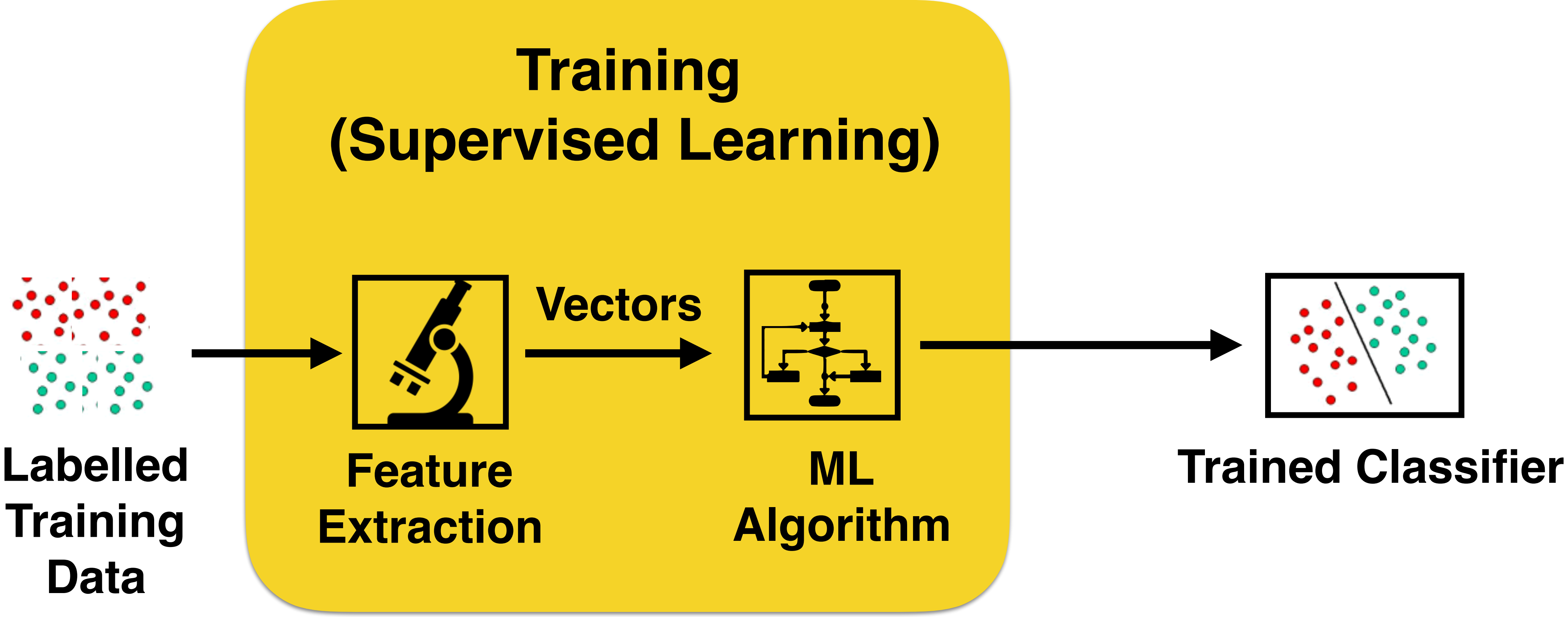
Security is different.

Security Tasks are Different: Adversary Adapts

Goal: Understand classifiers under attack.

Results: Vulnerable to automated evasion.

Building Machine Learning Classifiers



Training
(Supervised Learning)

Labelled Training Data

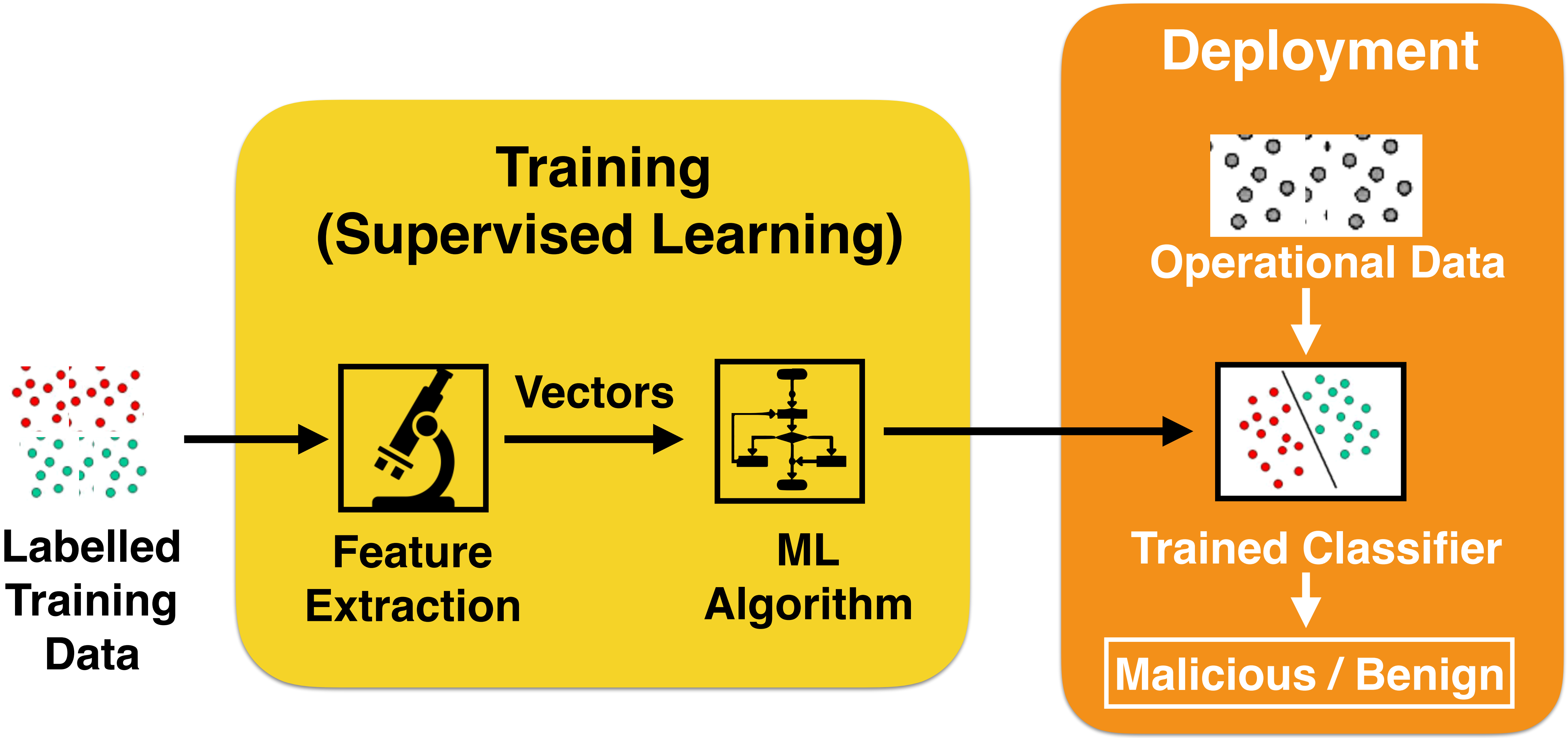
Feature Extraction

Vectors

ML Algorithm

Trained Classifier

Assumption: Training Data is Representative



Labelled Training Data

Training (Supervised Learning)

Feature Extraction

Vectors

ML Algorithm

Deployment

Operational Data

Trained Classifier

Malicious / Benign

Results: Evaded PDF Malware Classifiers

	PDFrate* [ACSAC'12]	Hidost [NDSS'13]
Accuracy	0.9976	0.9996
False Negative Rate	0.0000	0.0056
False Negative Rate with Adversary	**1.0000**	**1.0000**

* Mimicus [Oakland '14], an open source reimplementation of PDFrate.

Results: Evaded F sifiers

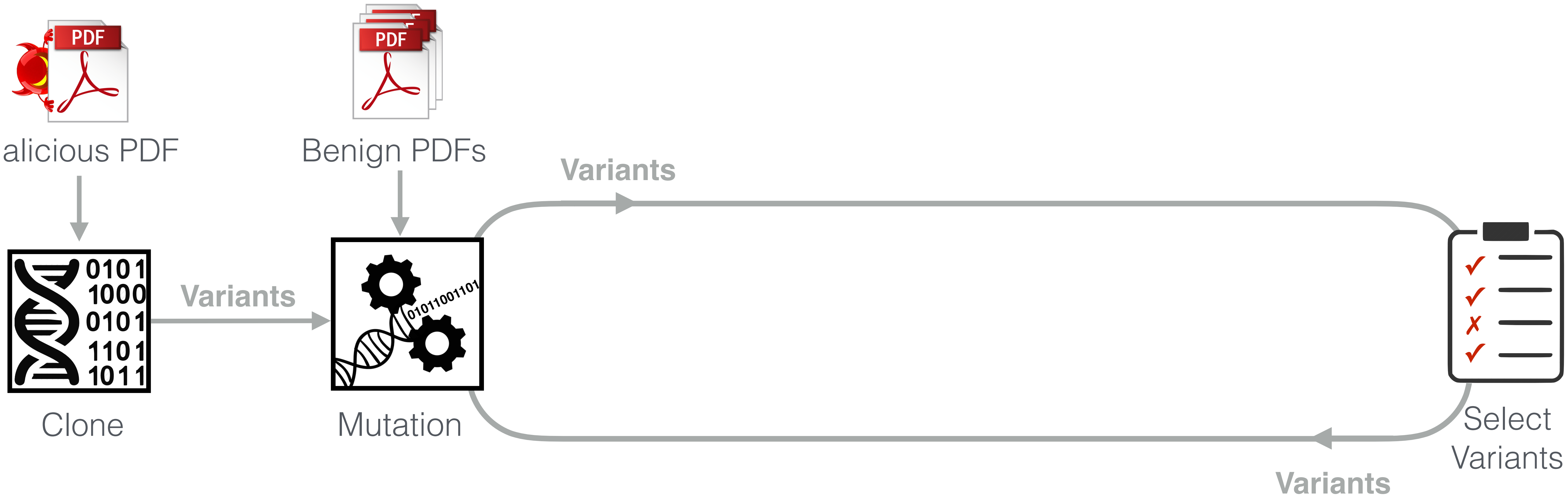
Very robust against “strongest conceivable mimicry attack”.

	PDFrate* [ACSAC’12]	Hidost [NDSS’13]
Accuracy	0.9976	0.9996
False Negative Rate	0.0000	0.0056
False Negative Rate with Adversary	**1.0000**	**1.0000**

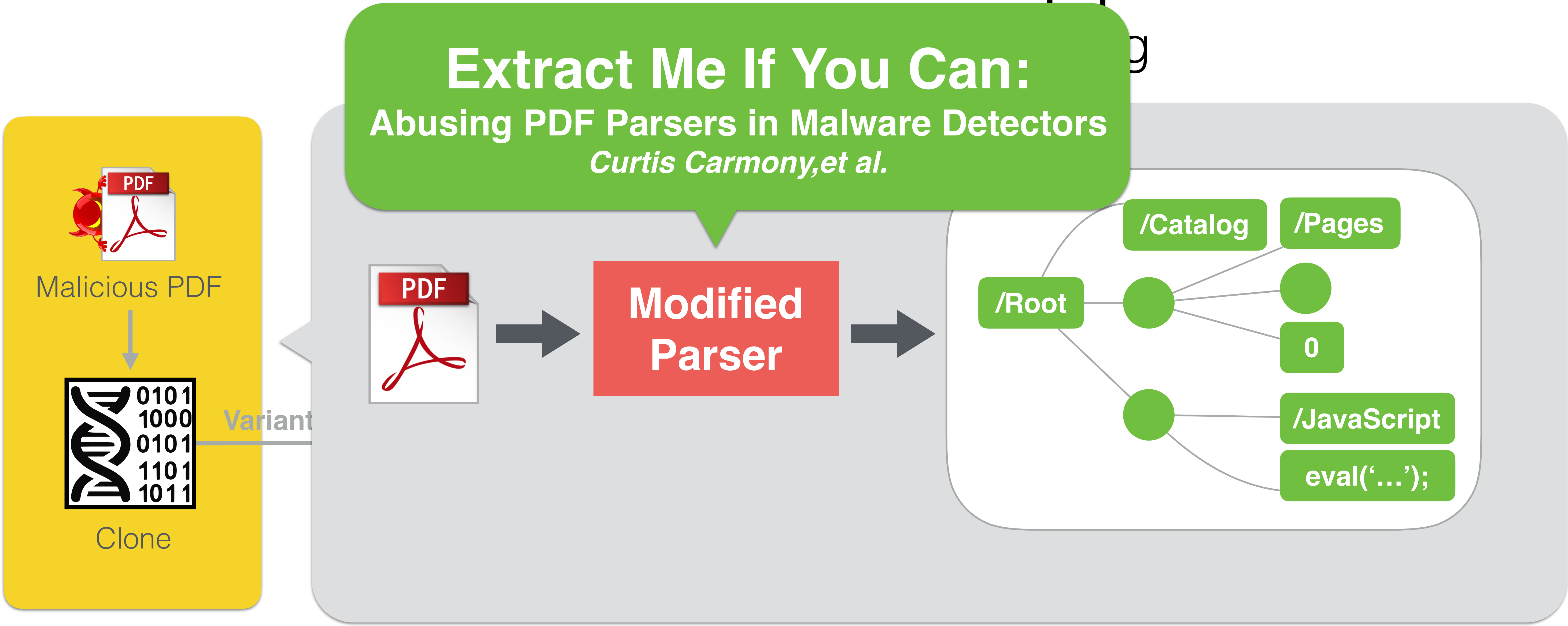
* Mimicus [Oakland ’14], an open source reimplementation of PDFrate.

Automated Evasion Approach

Based on Genetic Programming



Automated Evasion Approach



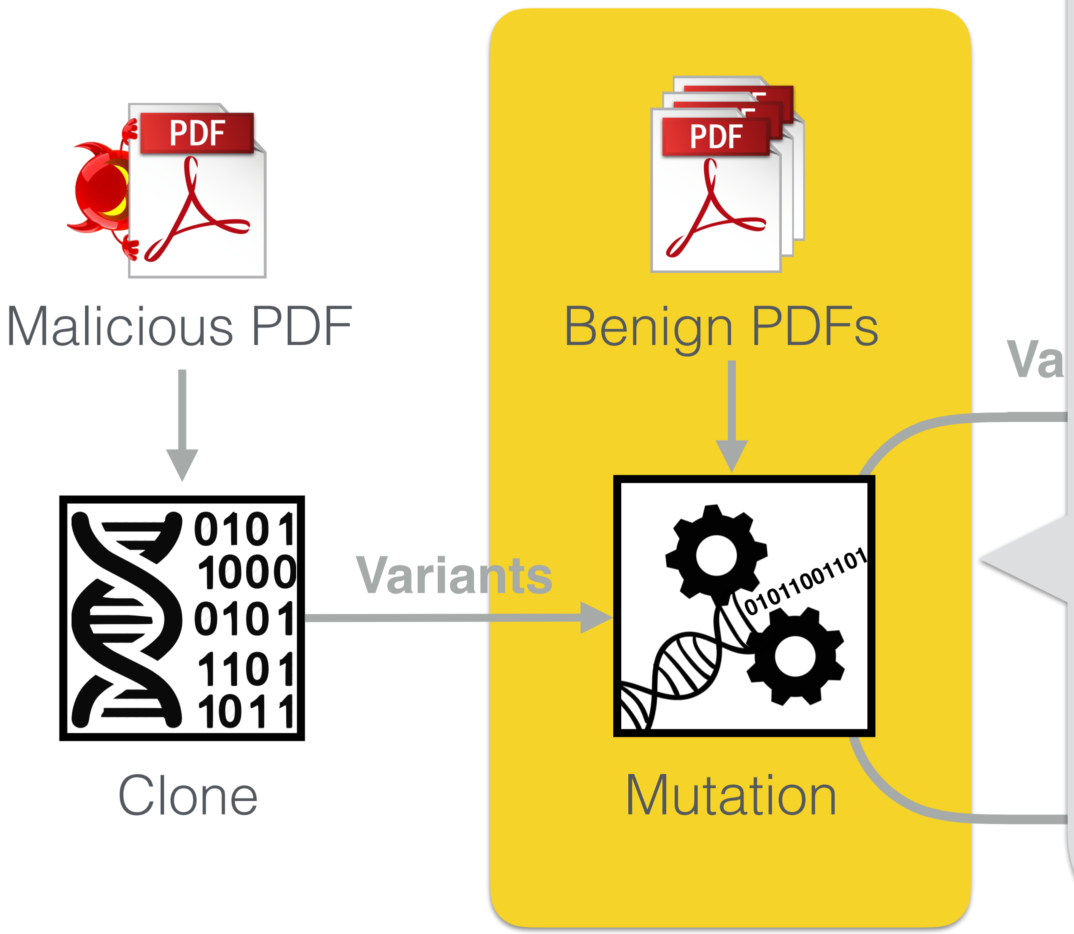
Extract Me If You Can:

Abusing PDF Parsers in Malware Detectors

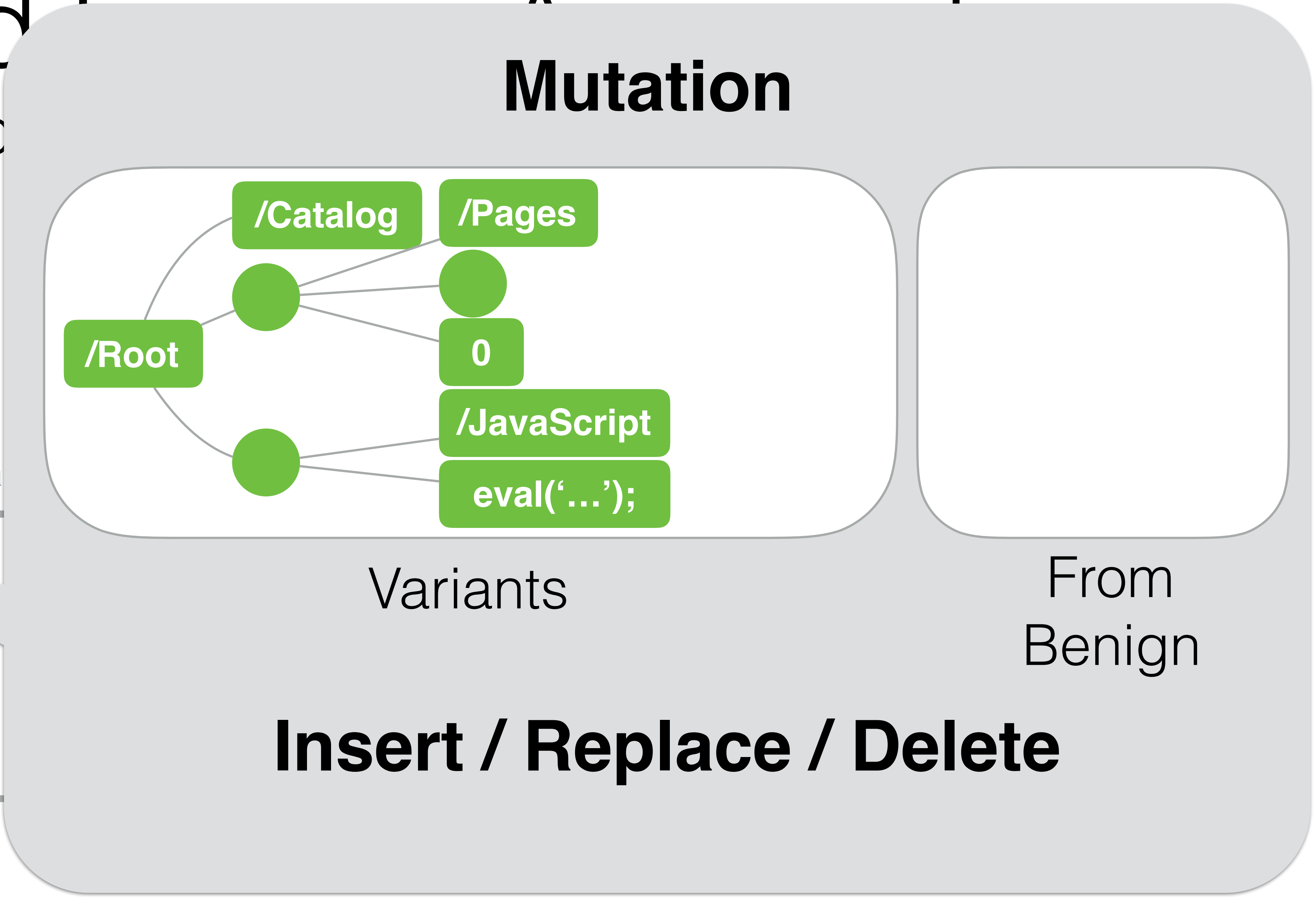
Curtis Carmony,et al.

Automated

Based c



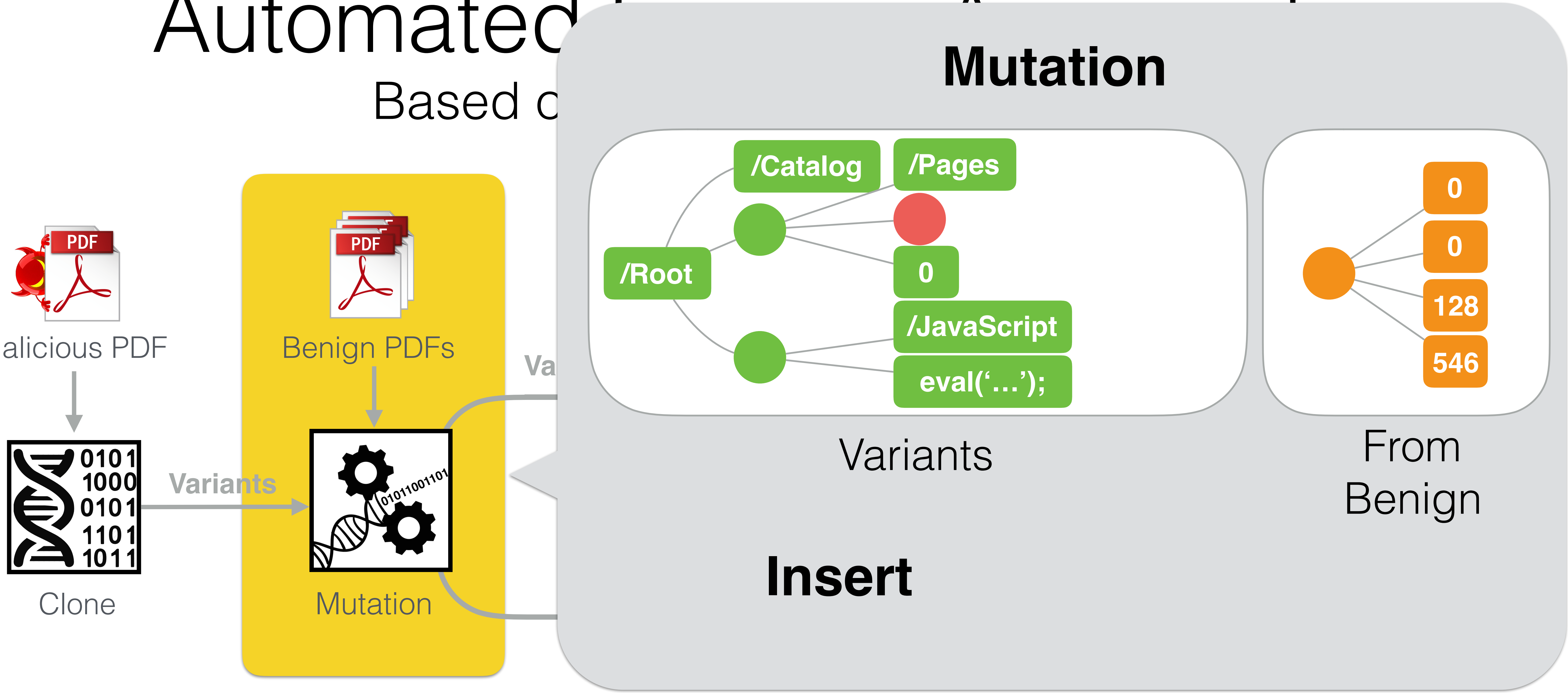
Mutation



Insert / Replace / Delete

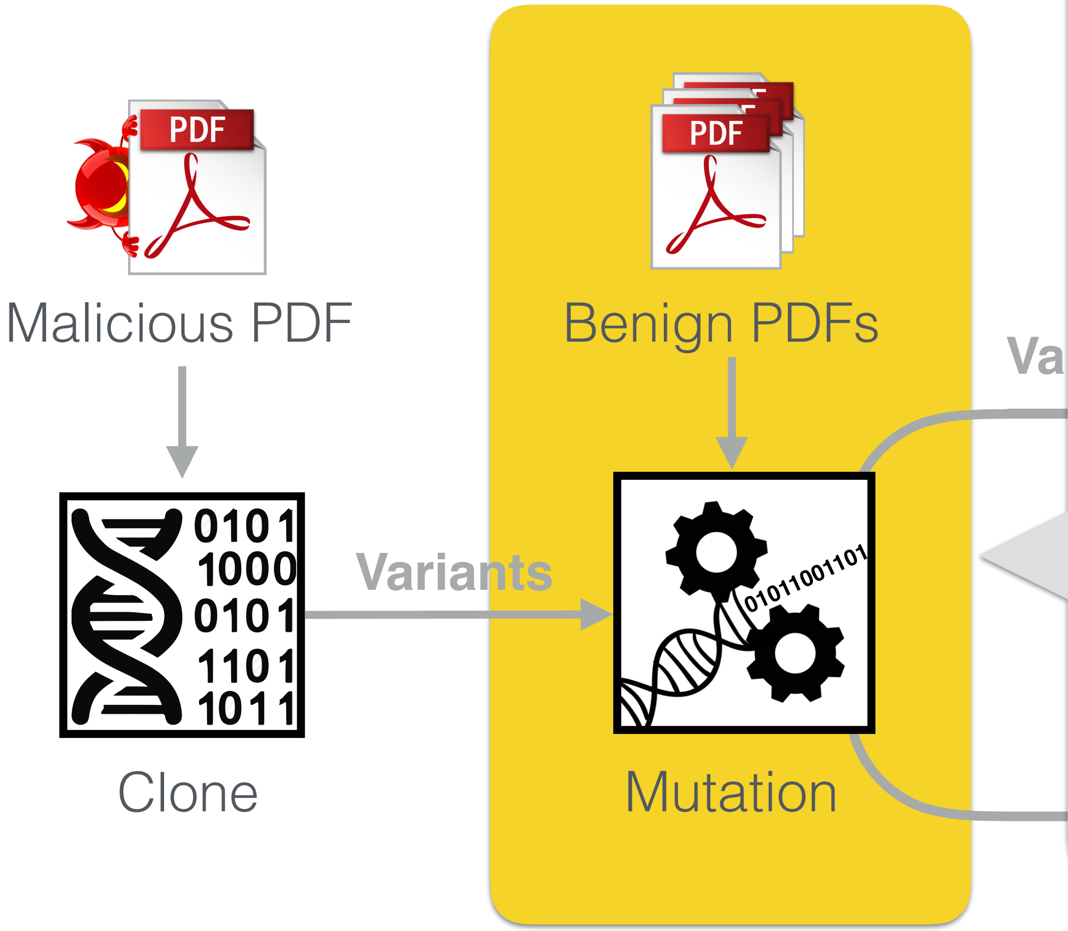
Automated E

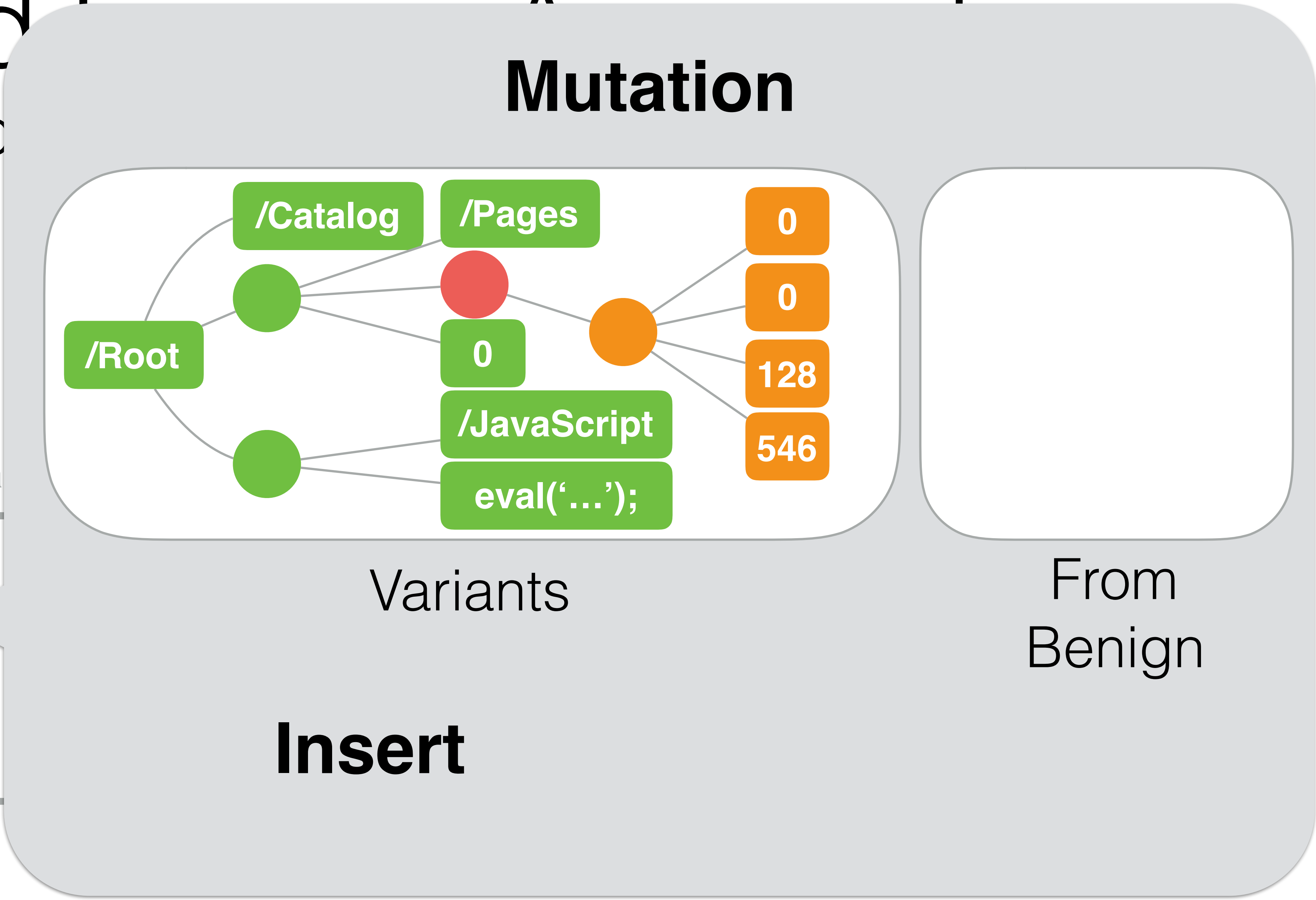
Based c



Automated

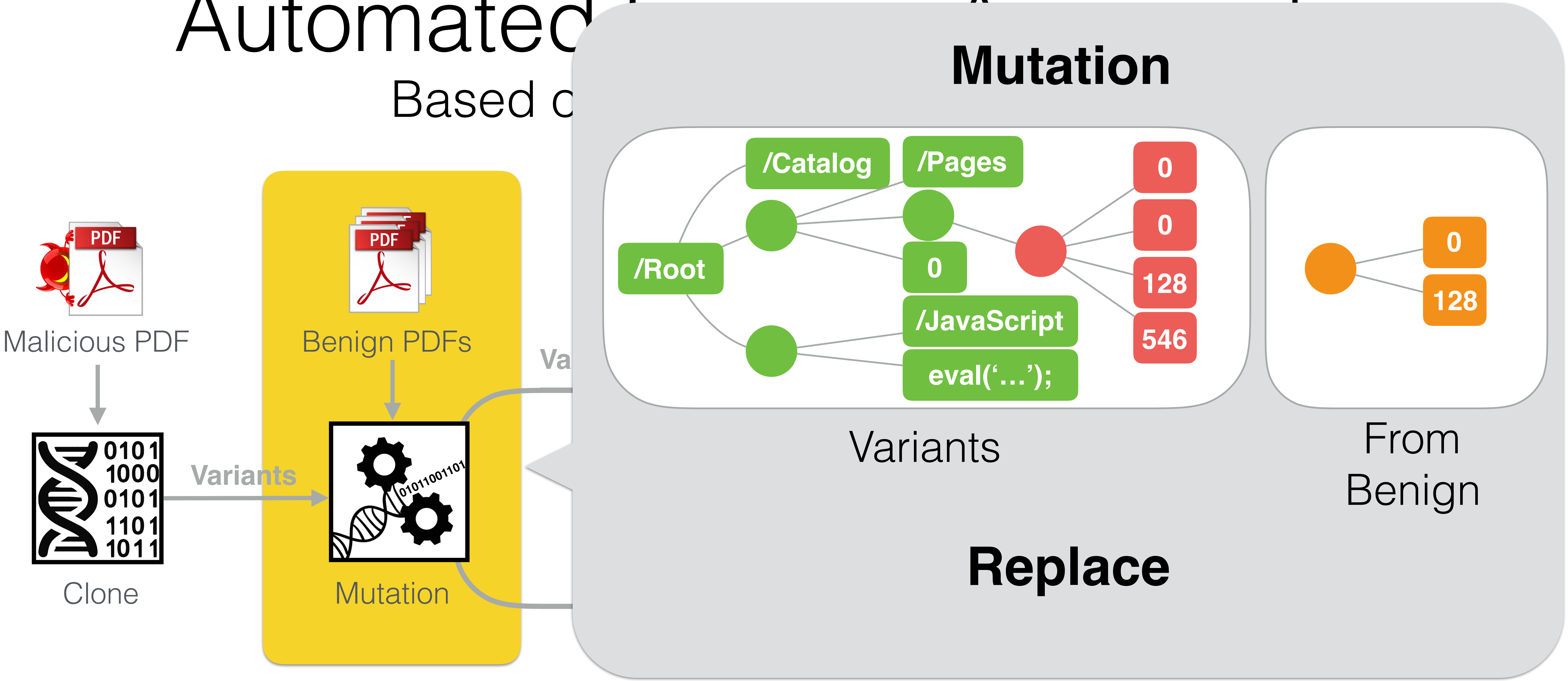
Based o





Automated

Based o



Malicious PDF

Clone

Variants

Benign PDFs

Mutation

Va

Mutation

/Root

/Catalog

/Pages

0

/JavaScript

eval('...');

0

0

128

546

Variants

0

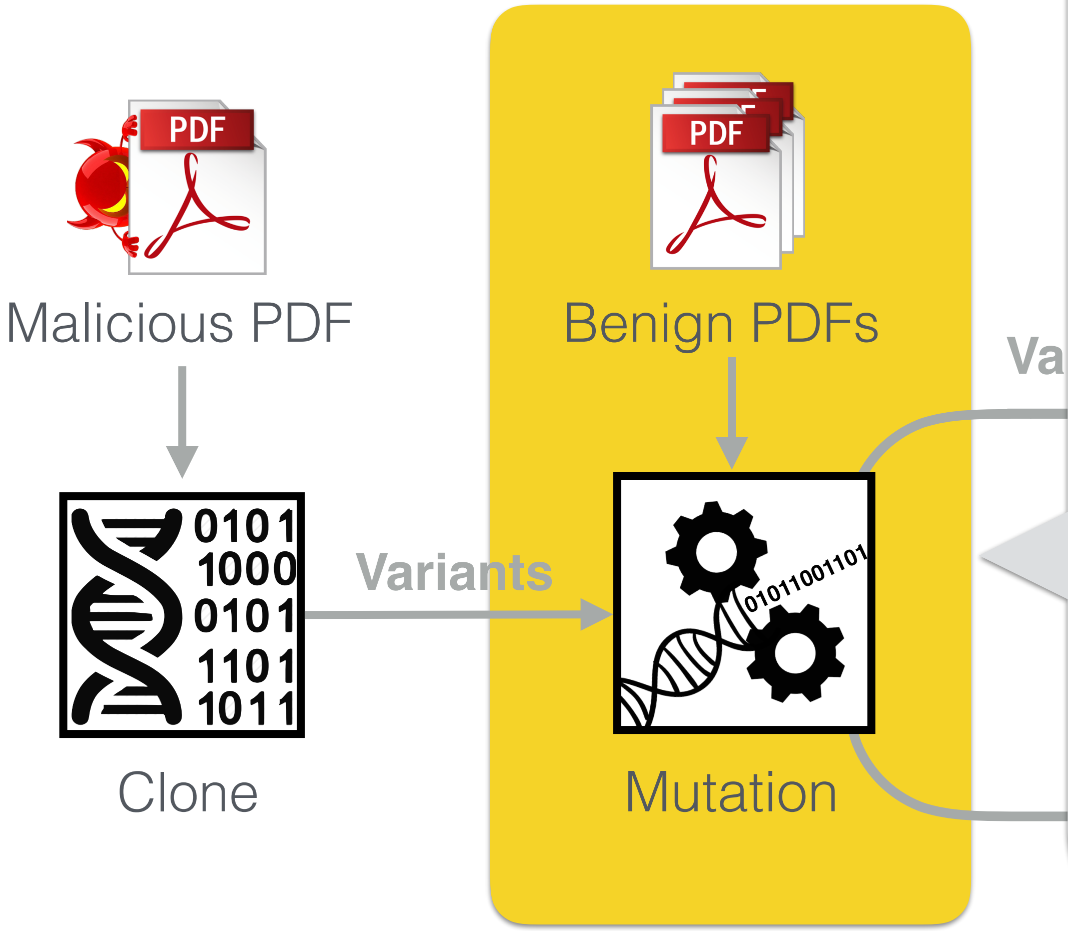
128

From Benign

Replace

Automated

Based o



Malicious PDF

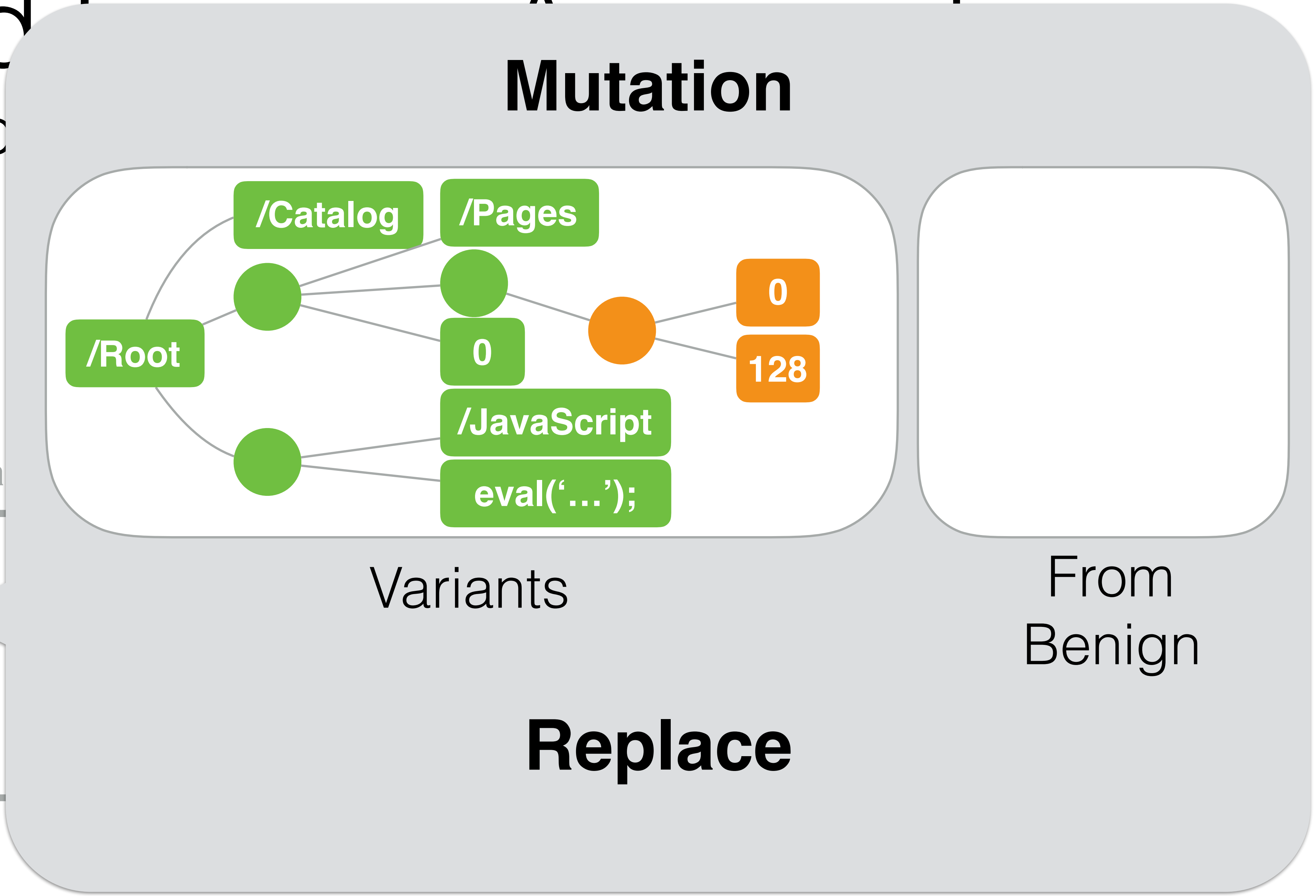
Benign PDFs

Clone

Variants

Mutation

Va



Mutation

/Catalog

/Pages

/Root

0

0

128

/JavaScript

eval('...');

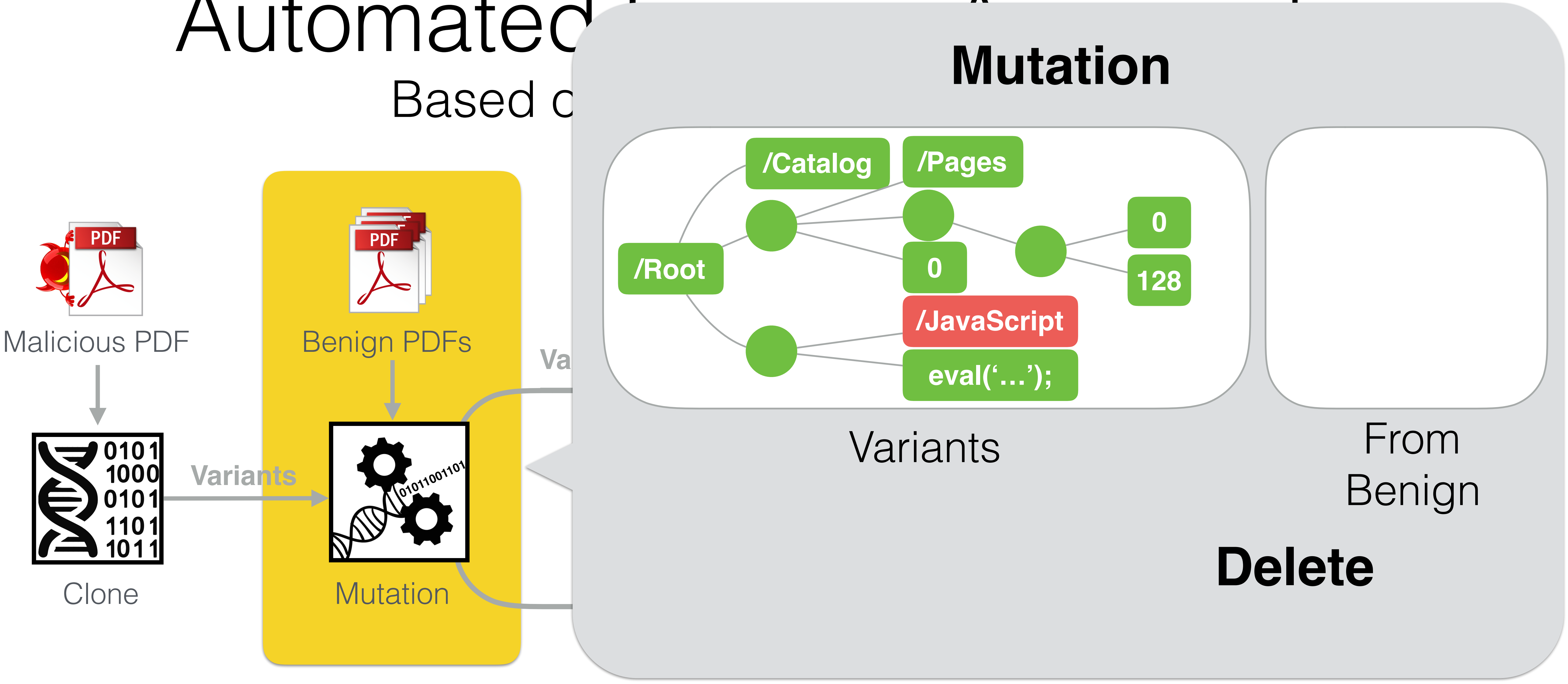
Variants

From Benign

Replace

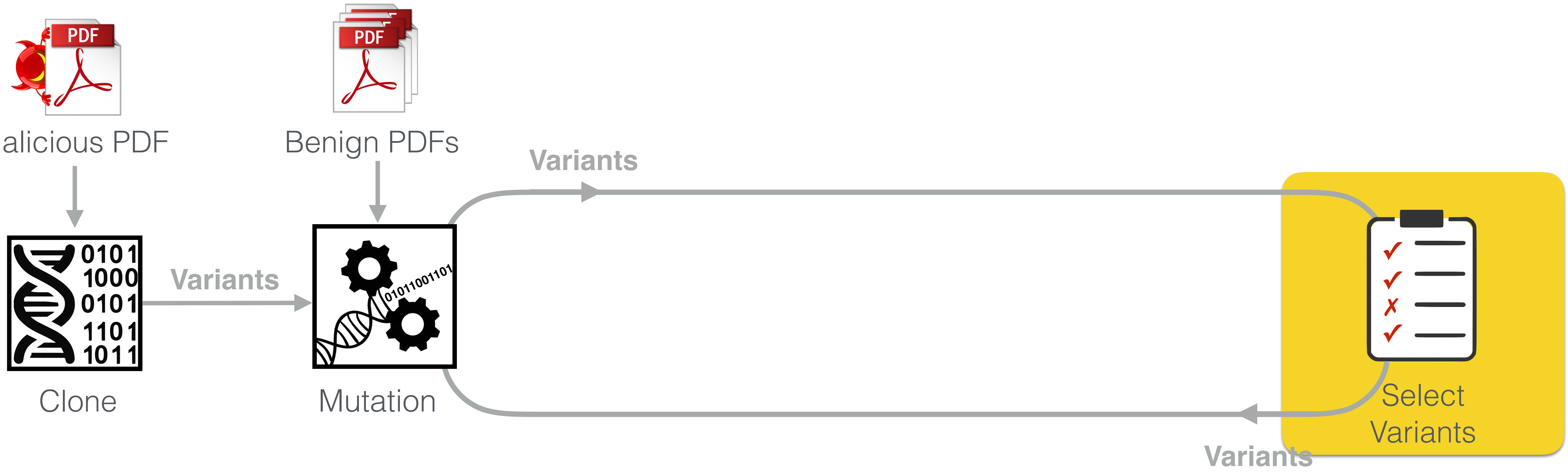
Automated

Based o



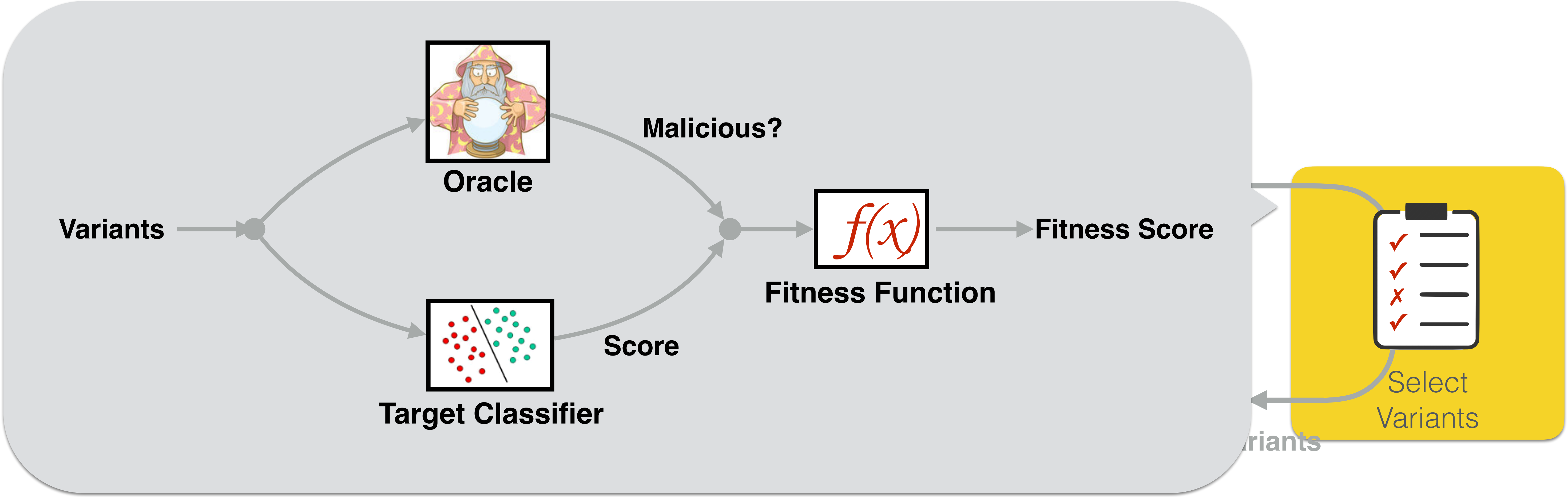
Automated Evasion Approach

Based on Genetic Programming



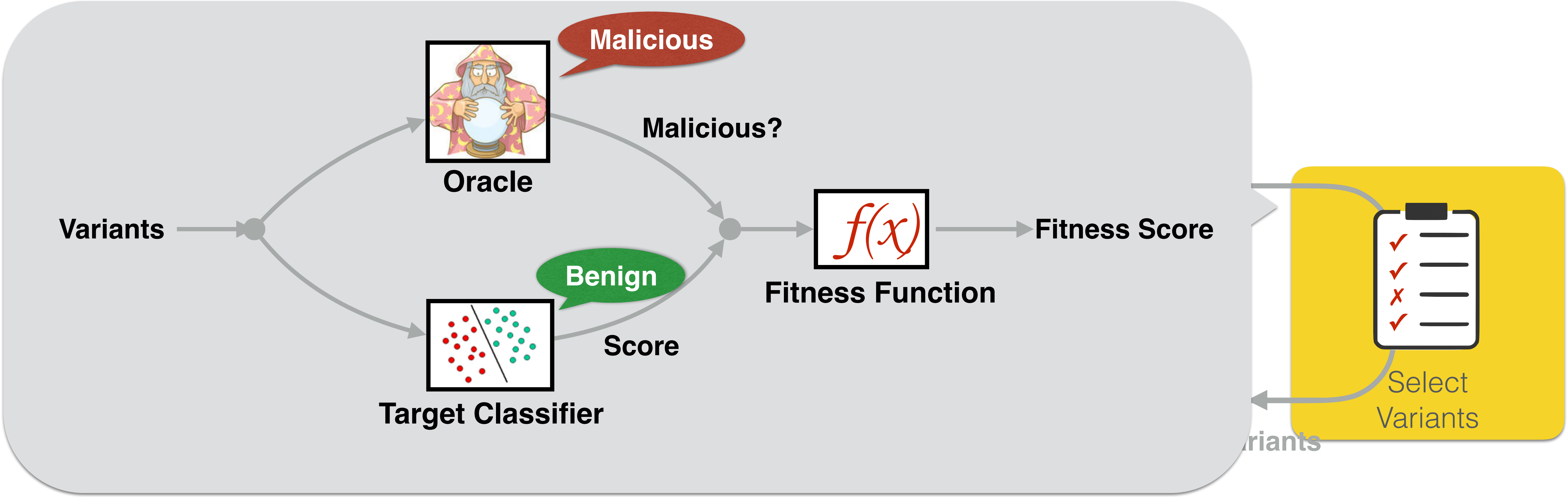
Automated Evasion Approach

Based on Genetic Programming



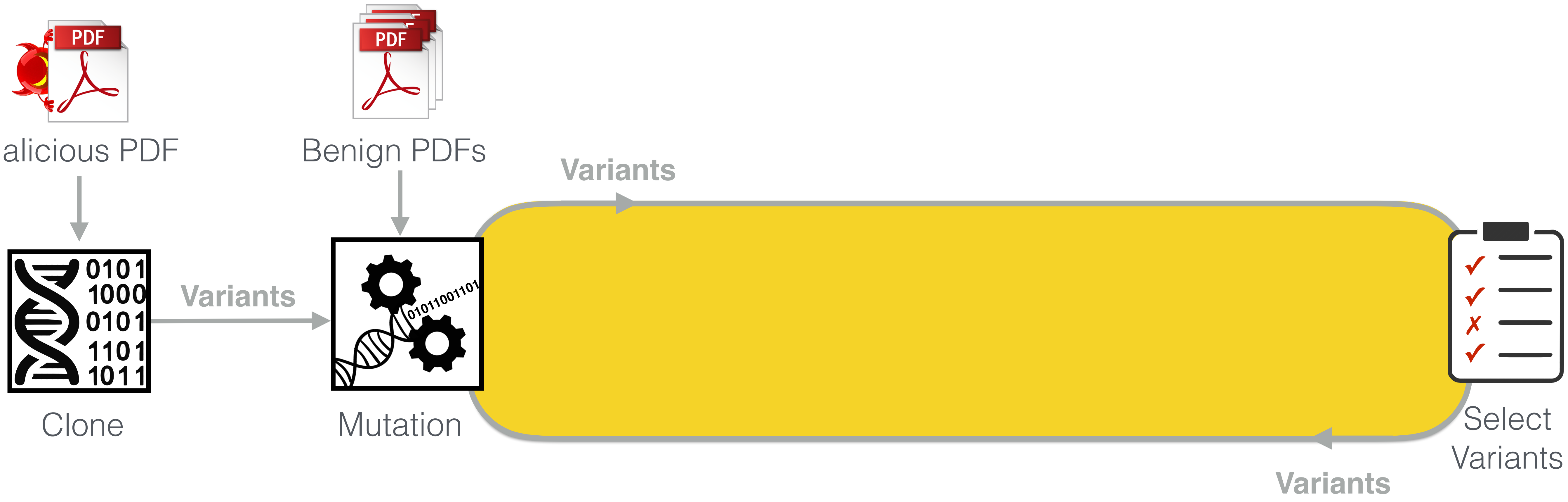
Automated Evasion Approach

Based on Genetic Programming

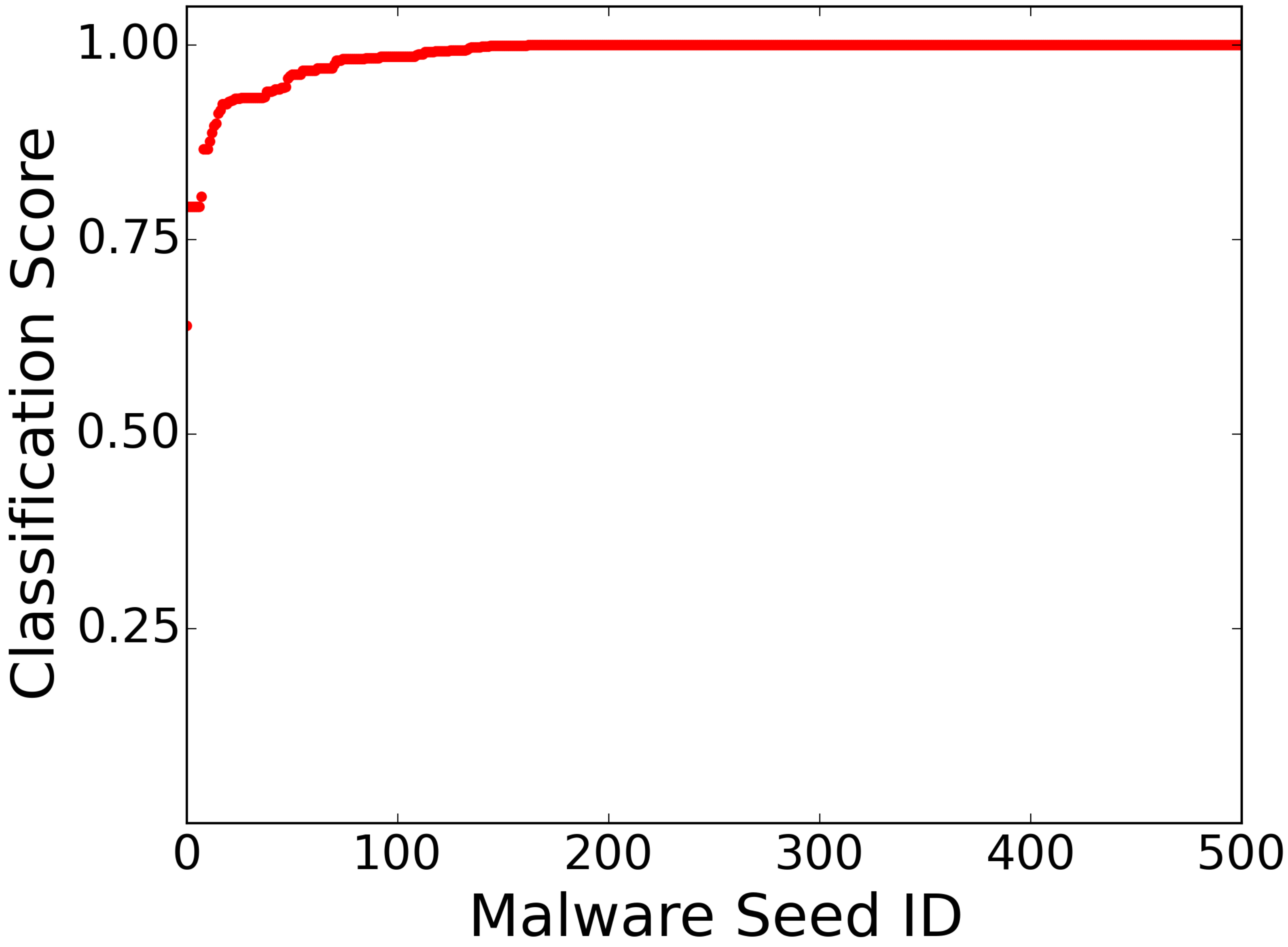


Automated Evasion Approach

Based on Genetic Programming

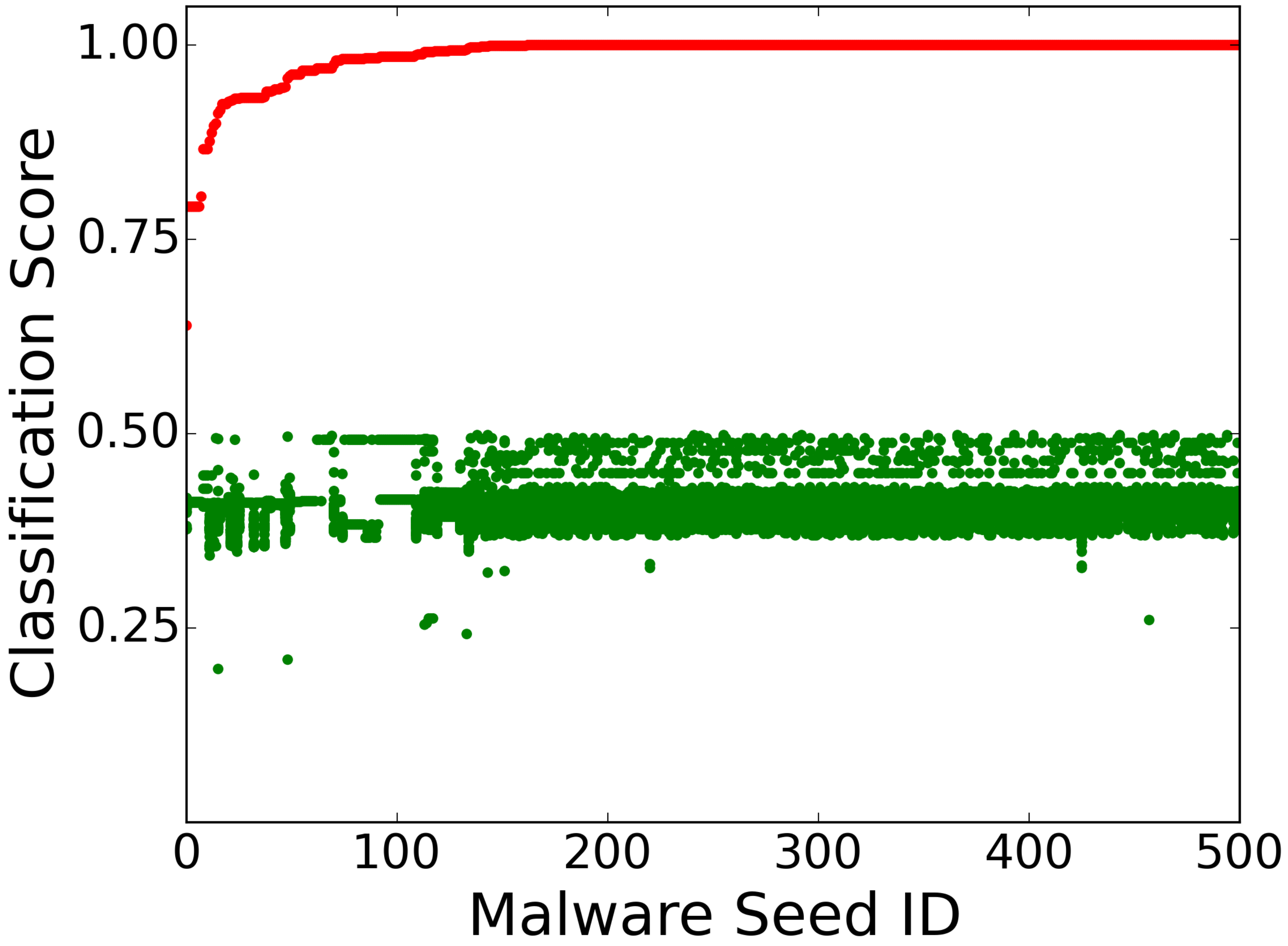


Results: Evaded PDFrate 100%



Original Malware Seeds

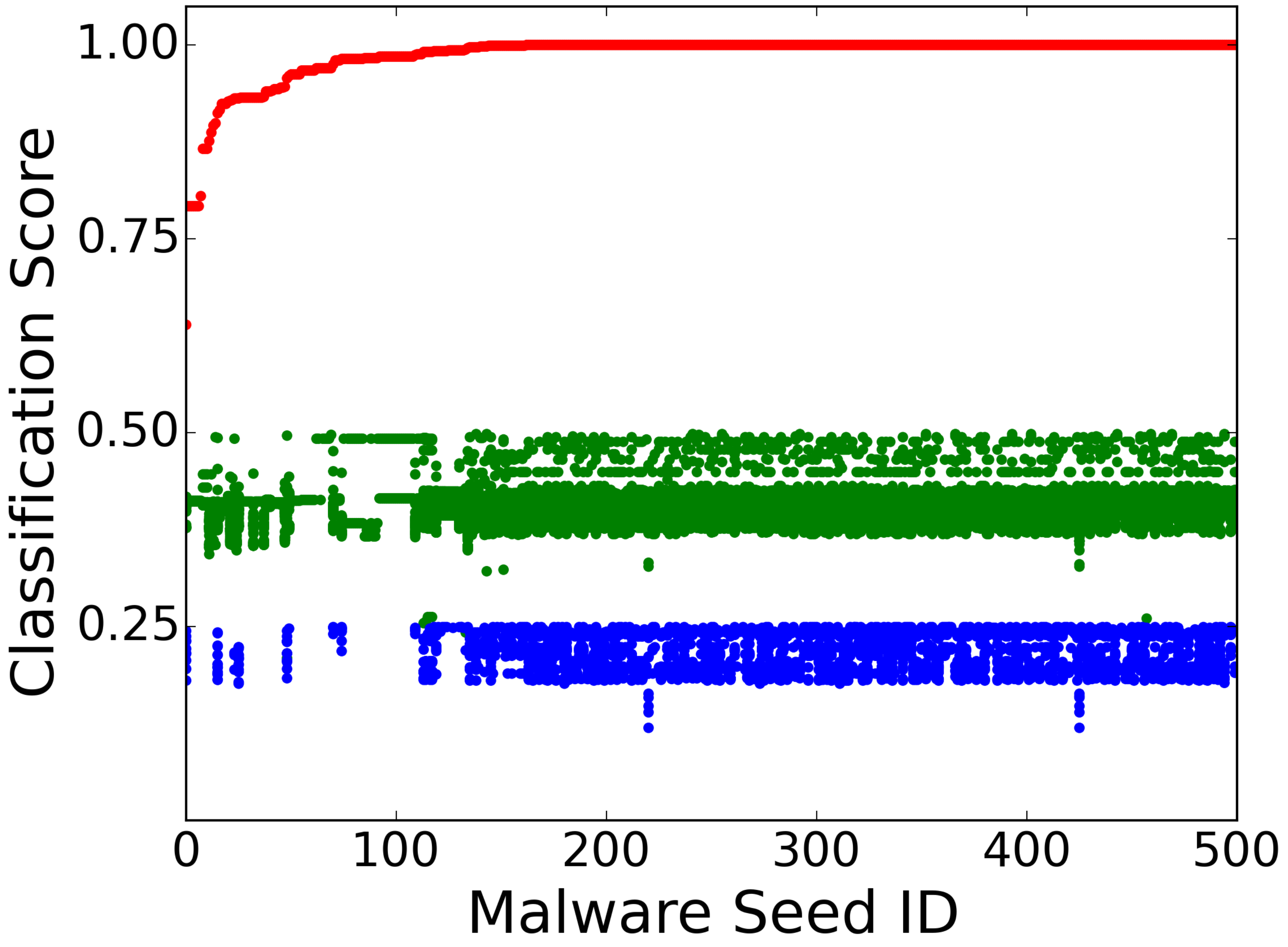
Results: Evaded PDFrate 100%



Original Malware Seeds

Evasive Variants

Evaded PDFrate with Adjusted Threshold

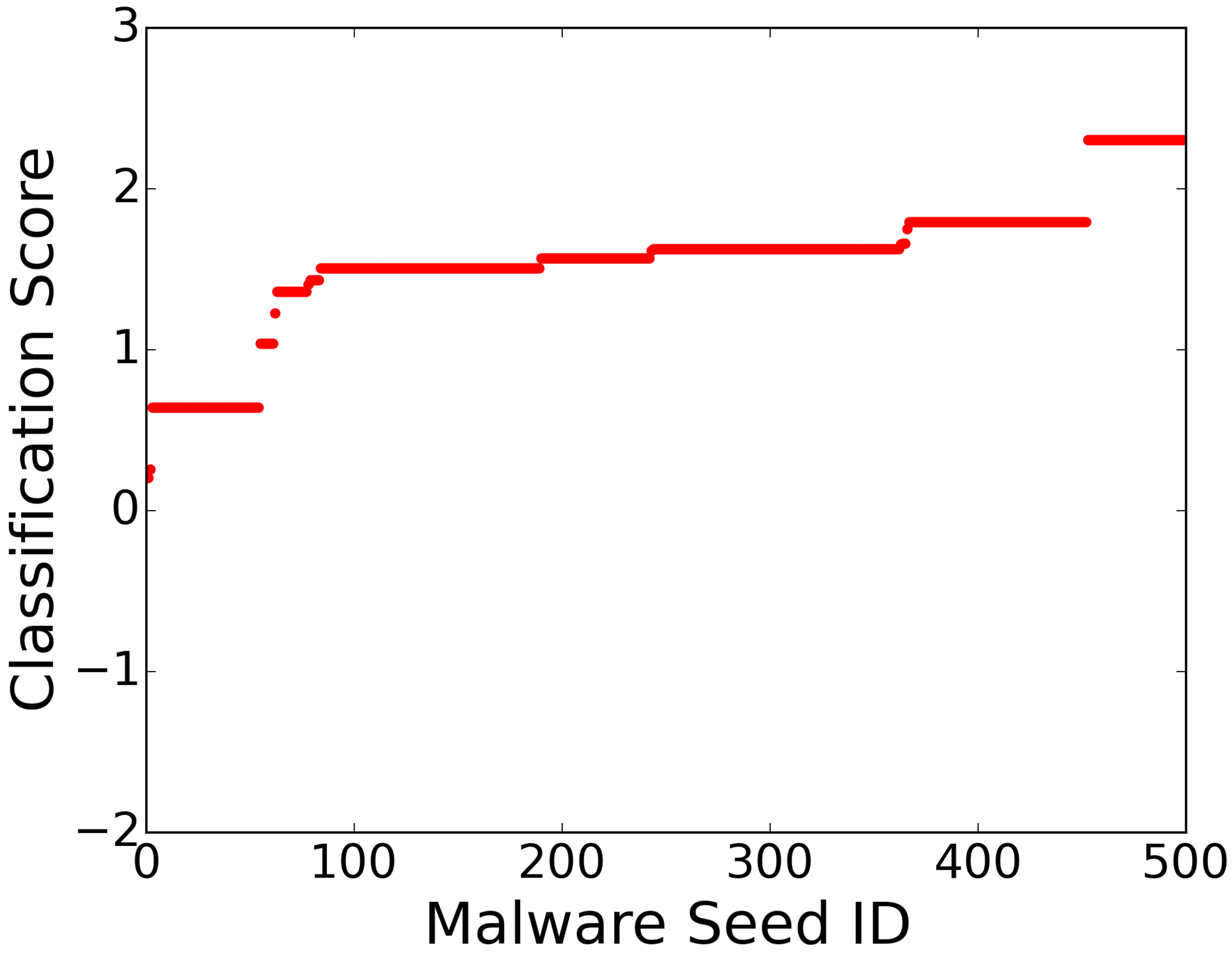


Original Malware Seeds

Evasive Variants

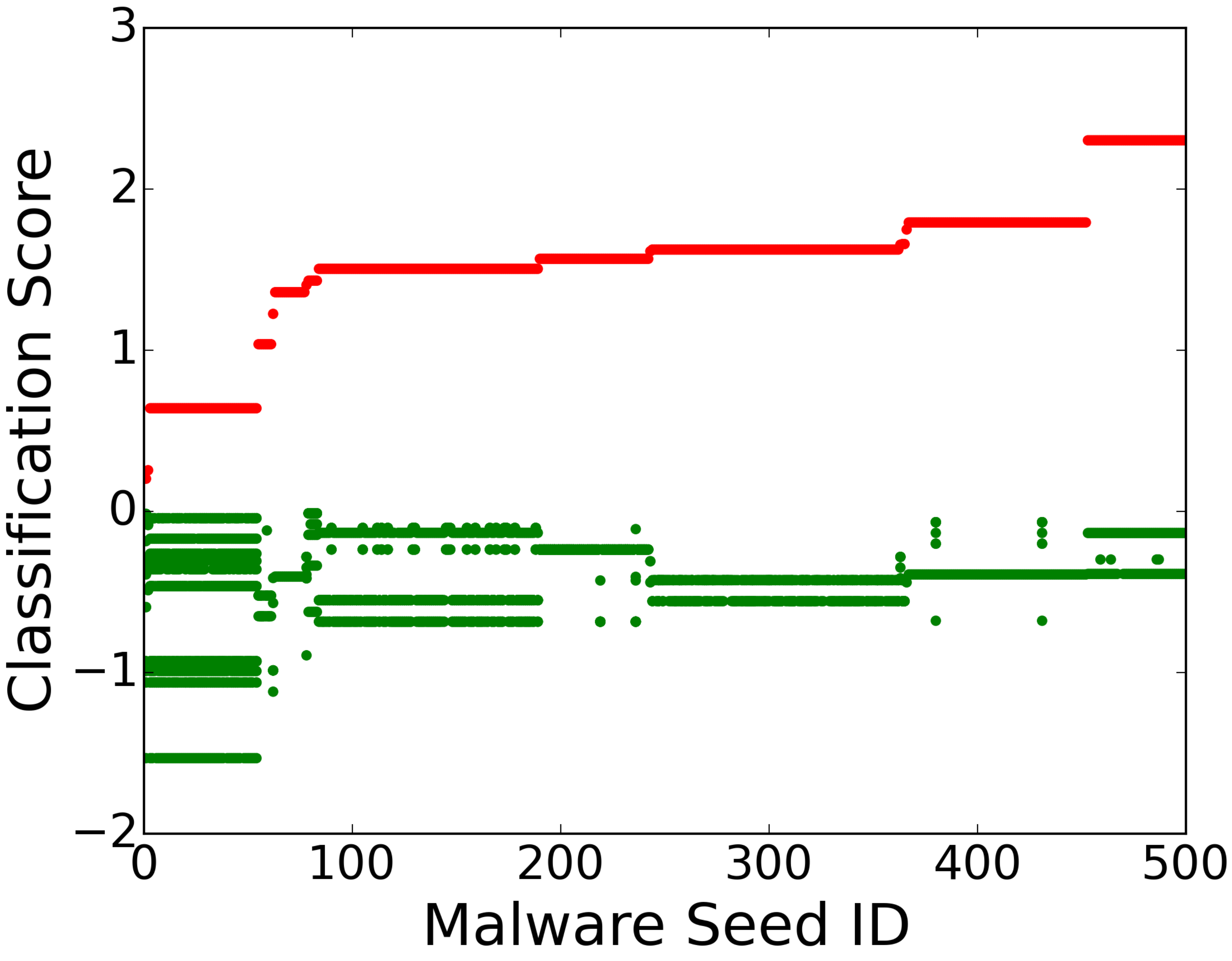
Evasive Variants with lower threshold

Results: Evaded Hidost 100%



Original Malware Seeds

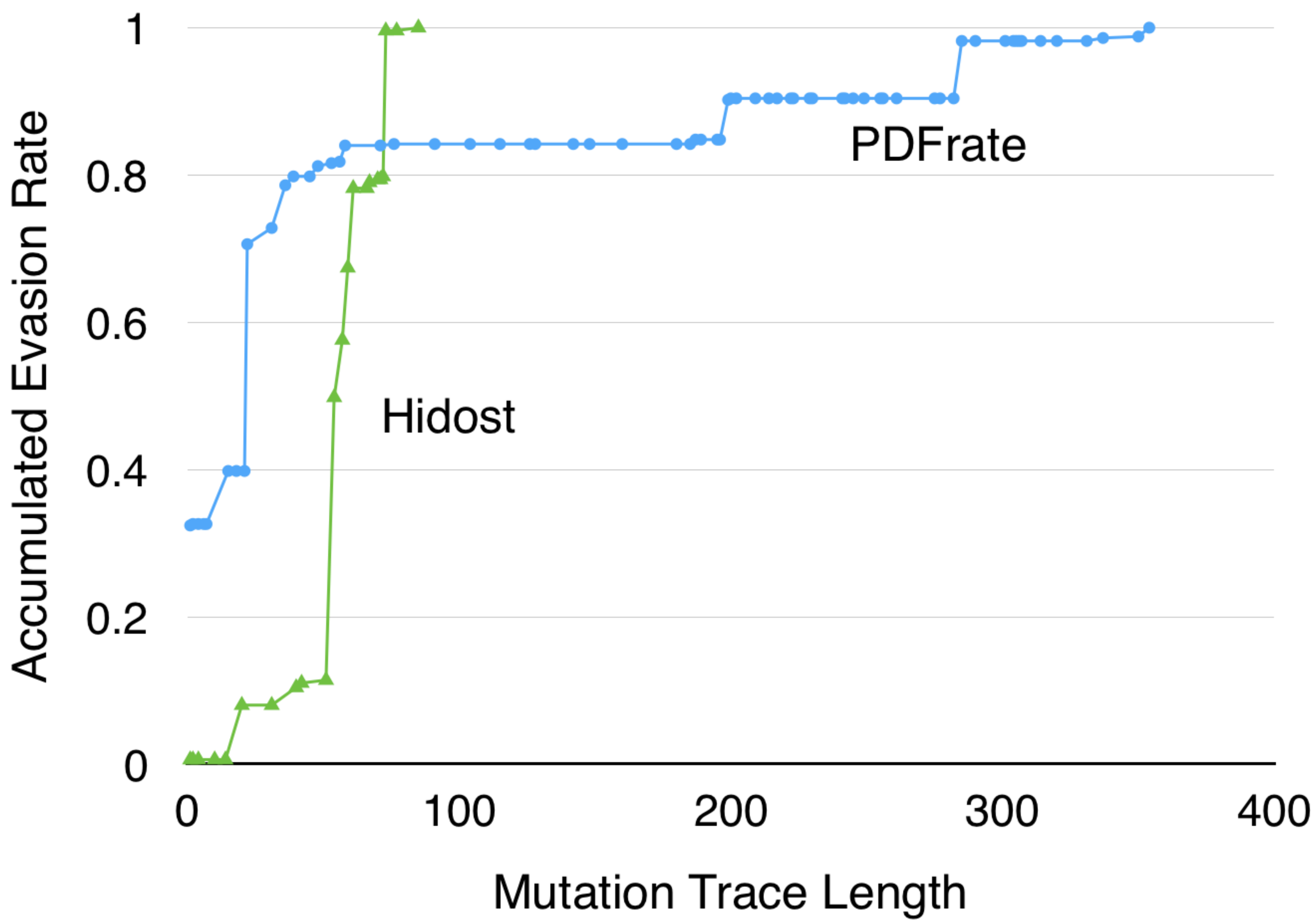
Results: Evaded Hidost 100%



Original Malware Seeds

Evasive Variants

Results: Accumulated Evasion Rate



Difficulty varies by seed

Simple mutations often work

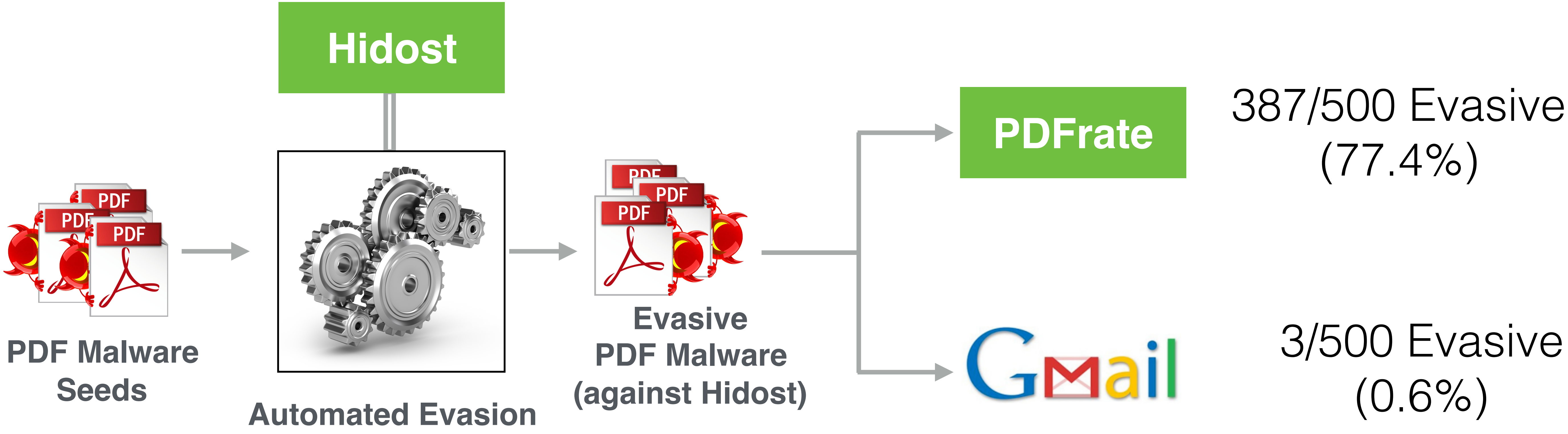
Complex mutations sometimes needed.

Difficulty varied by targets:

PDFrate: 6 days to evade all

Hidost: 2 days to evade all

Cross-Evasion Effects



Hidost

PDF Malware Seeds

Automated Evasion

Evasive PDF Malware (against Hidost)

PDFrate

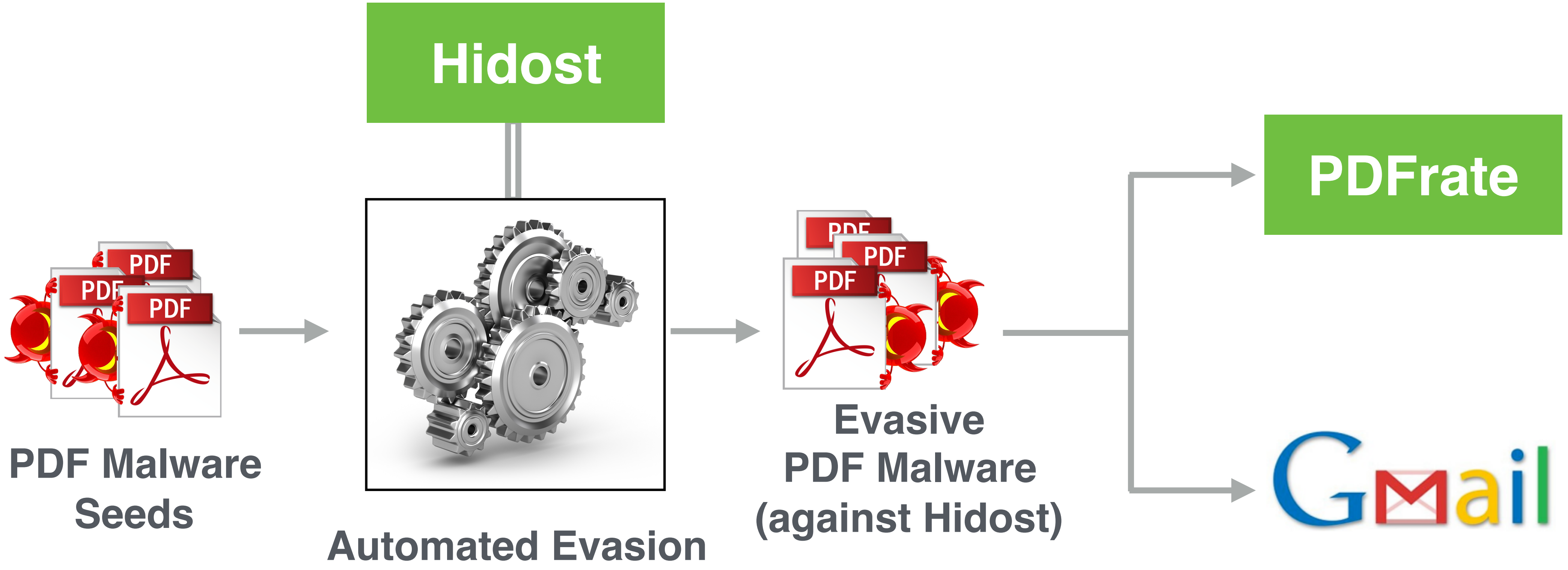
387/500 Evasive (77.4%)

Gmail

3/500 Evasive (0.6%)

Gmail’s classifier is secure?

Cross-Evasion Effects



Hidost

PDF Malware Seeds

Automated Evasion

Evasive PDF Malware (against Hidost)

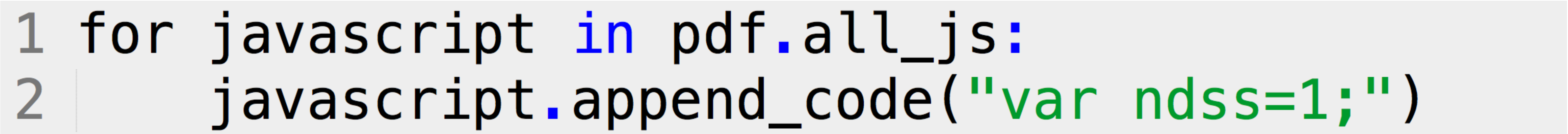
PDFrate

387/500 Evasive (77.4%)

3/500 Evasive (0.6%)

Gmail's classifier is [redacted] different.

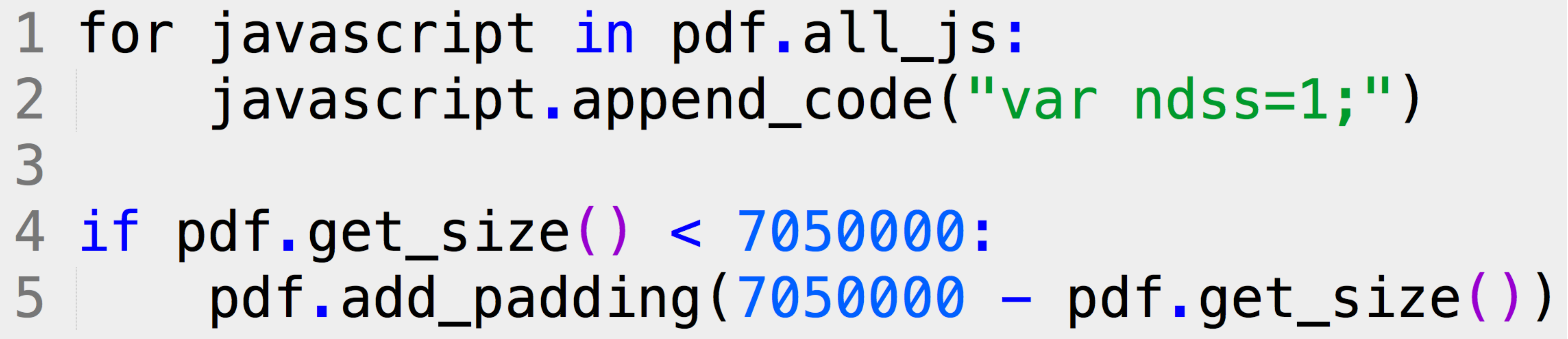
Evading Gmail's Classifier



```
for javascript in pdf.all_js:
    javascript.append_code("var ndss=1;")
```

Evasion rate on Gmail : 135/380 (35.5%)

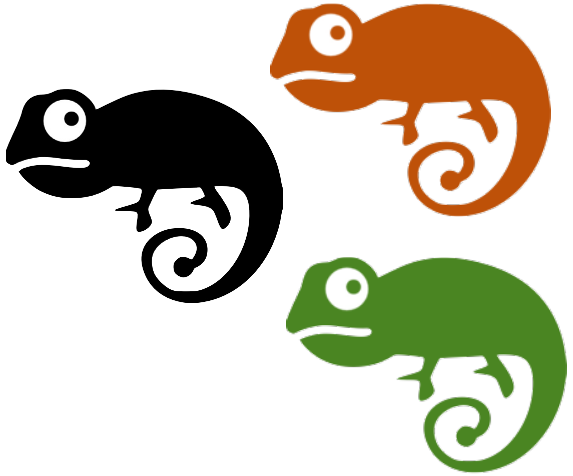
Evading Gmail’s Classifier



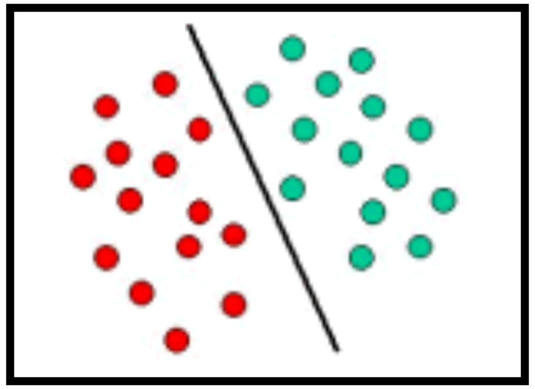
```
for javascript in pdf.all_js:
    javascript.append_code("var ndss=1;")

if pdf.get_size() < 7050000:
    pdf.add_padding(7050000 - pdf.get_size())
```

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

Conclusion



Vs.



Who will win this arm race?

Source Code: http://EvadeML.org