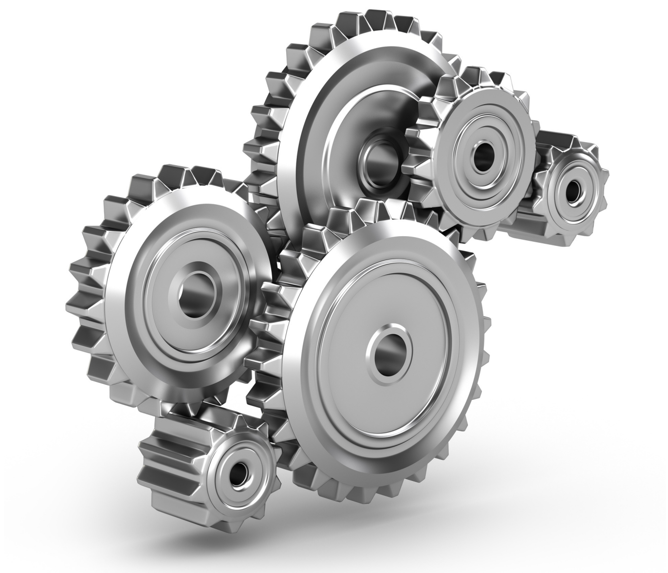
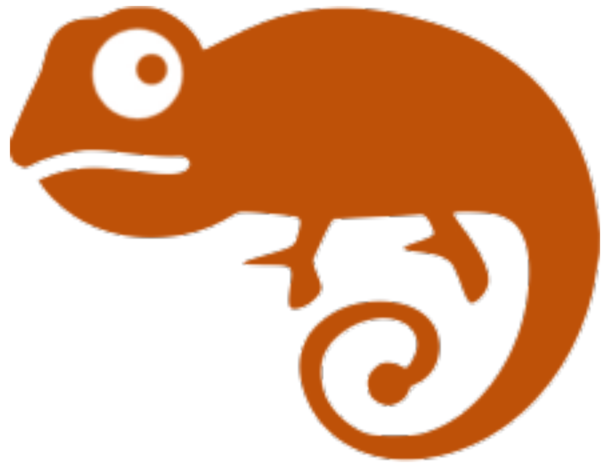
Automatically Evading Classifiers

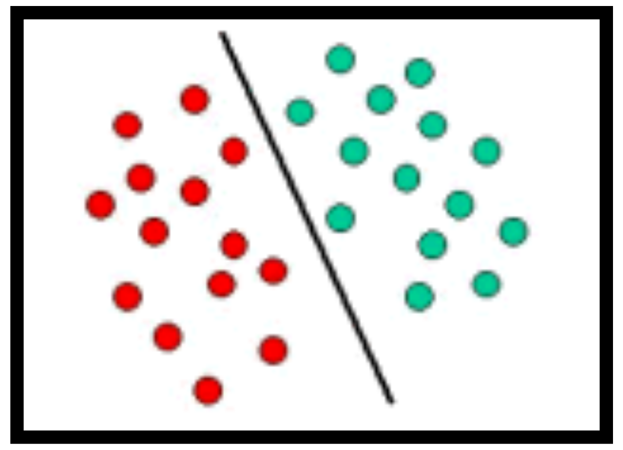
A Case Study on PDF Malware Classifiers



Weilin Xu



David Evans



Yanjun Qi

University of Virginia

Machine Learning is Solving Our Problems

Spam

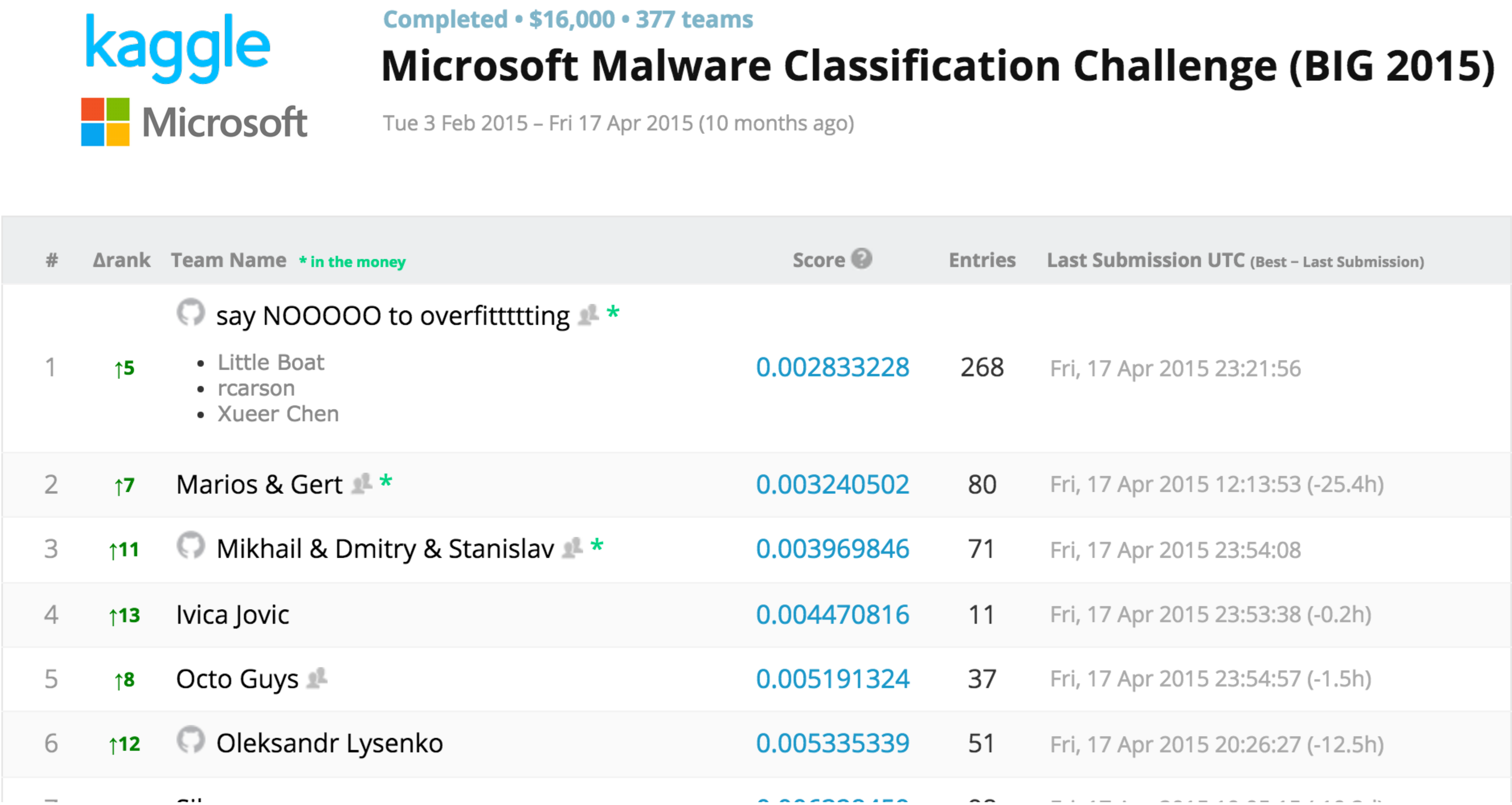
IDS

Fake Accounts

Malware

. . .

. . .

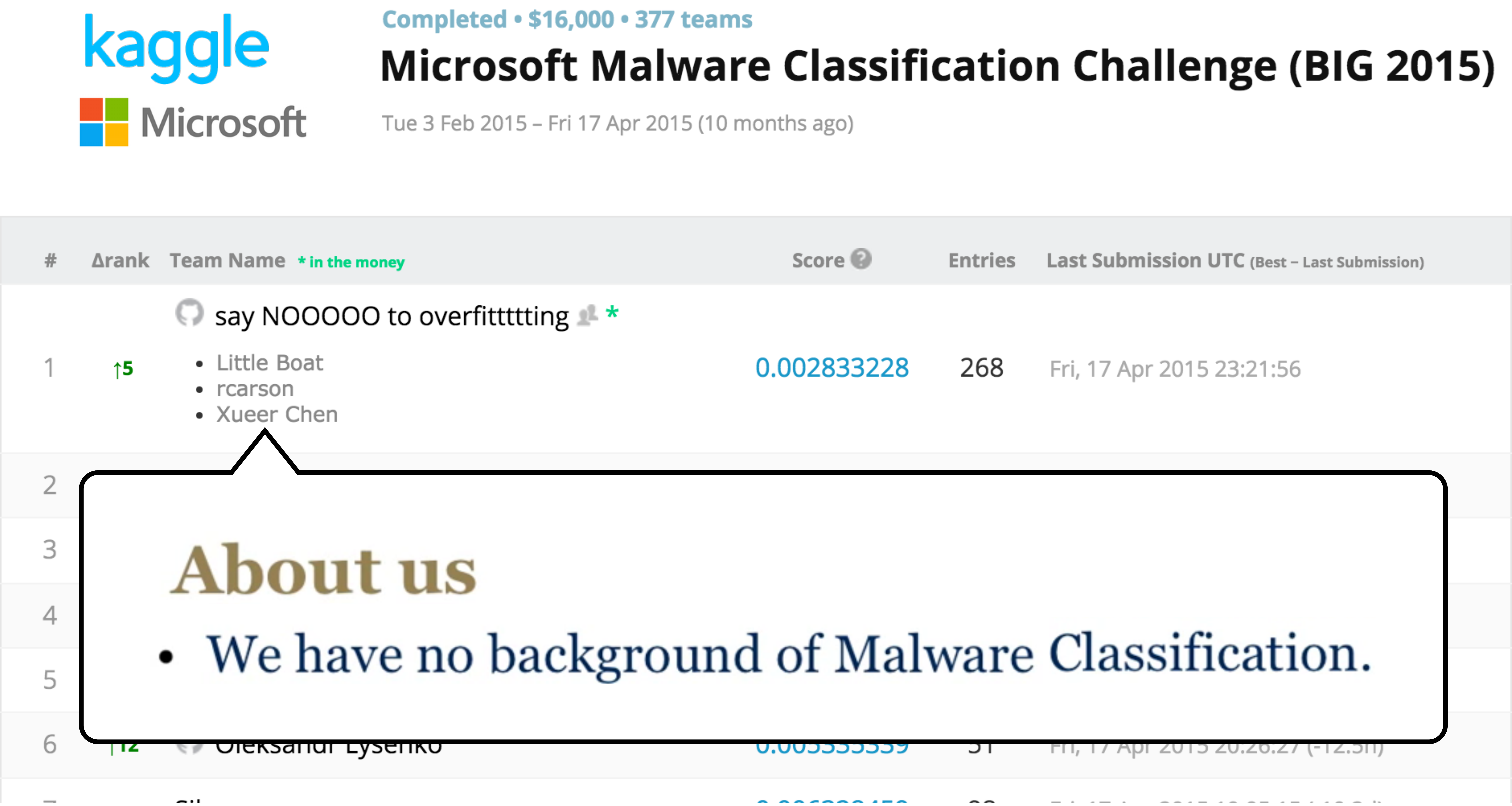


Completed • $16,000 • 377 teams

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 overfittttting * • 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)



kaggle

Microsoft

Completed • $16,000 • 377 teams

Microsoft Malware Classification Challenge (BIG 2015)

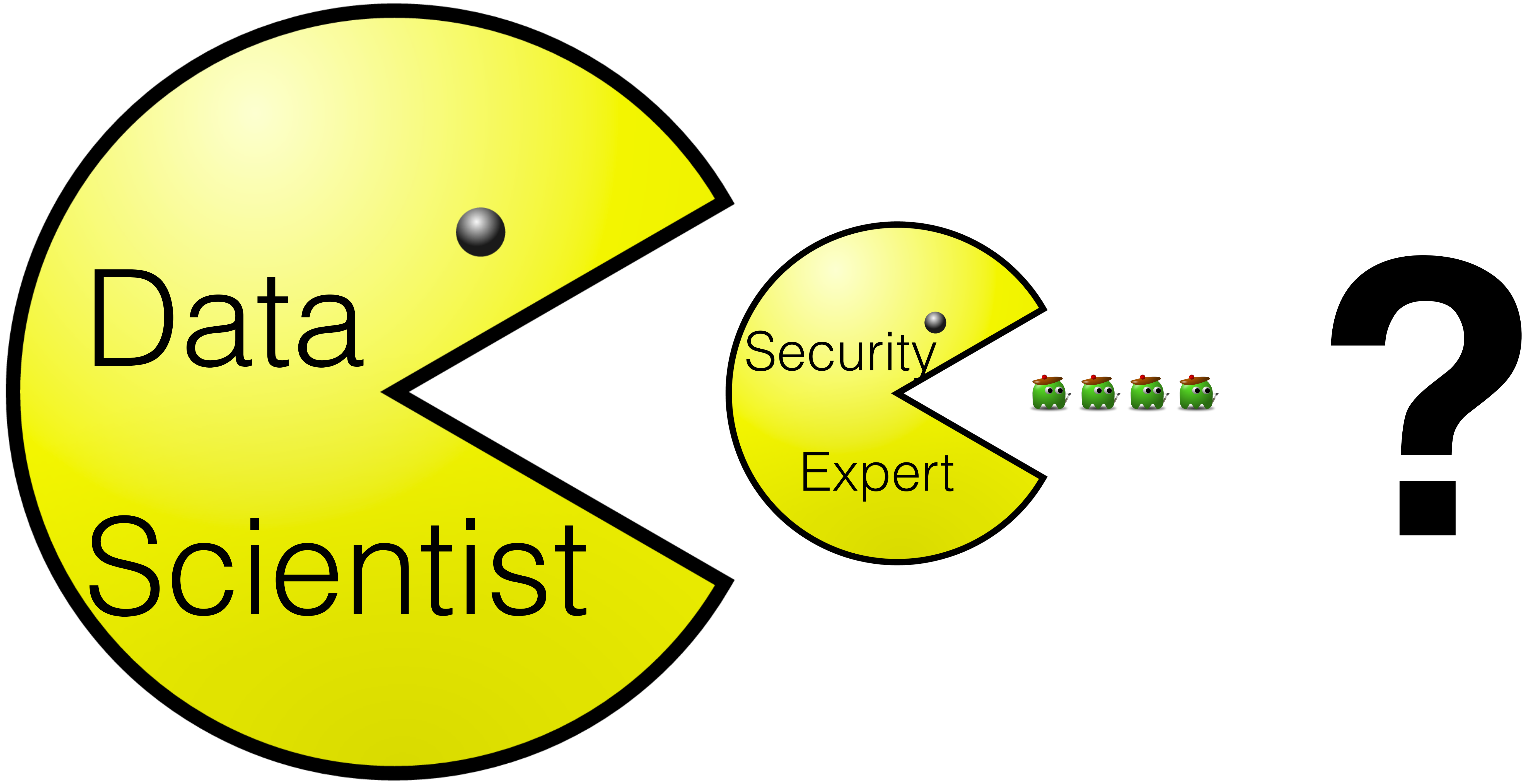
Tue 3 Feb 2015 – Fri 17 Apr 2015 (10 months ago)

#	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)
7		[illegible]	[illegible]	[illegible]	[illegible]

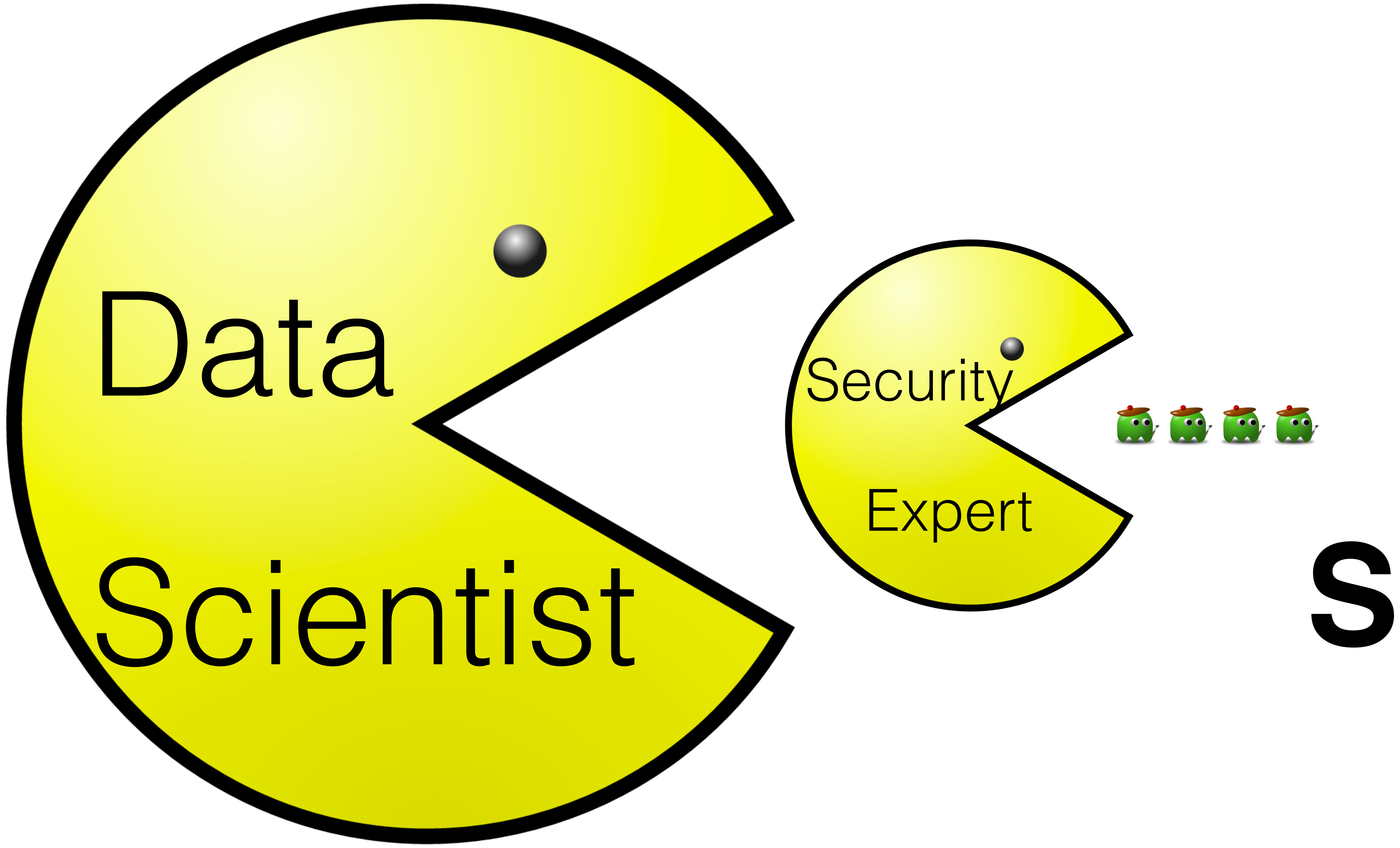
About us

- We have no background of Malware Classification.

Machine Learning is Eating the World



Machine Learning is Eating the World



No!
Security is different.

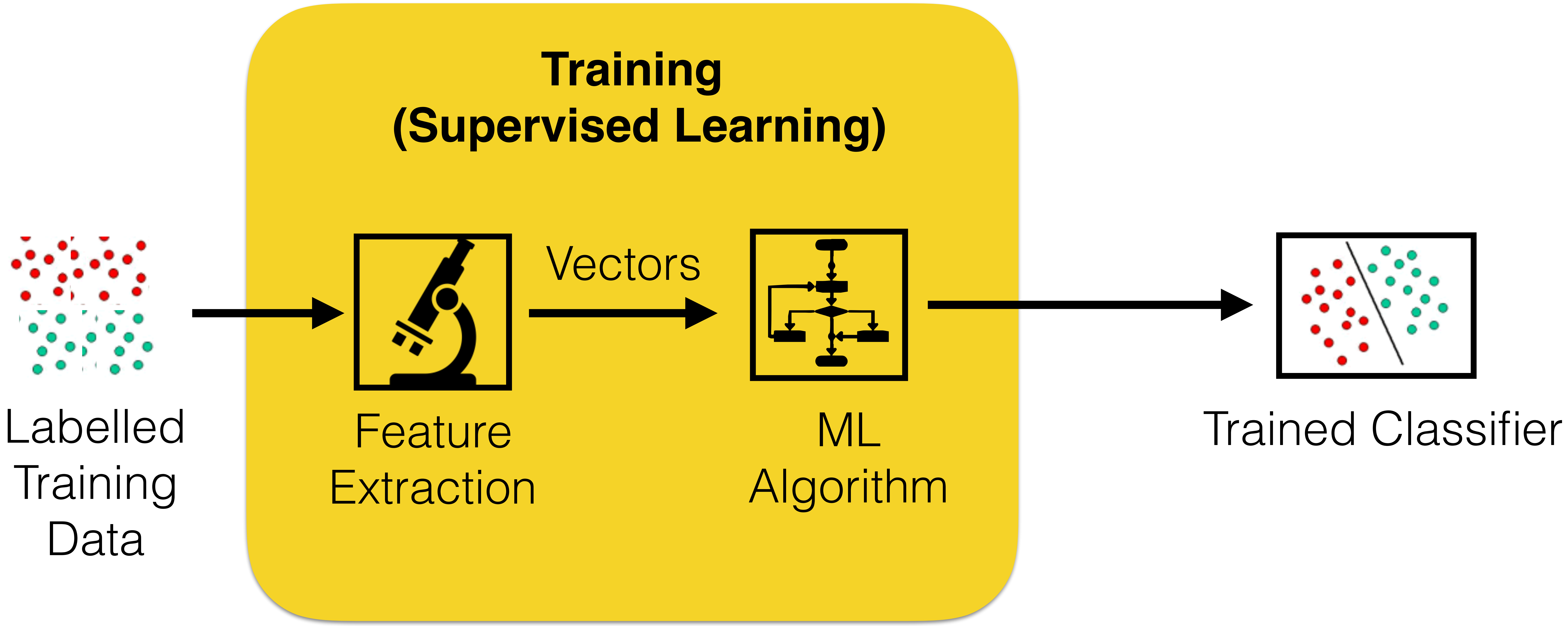
Security Tasks are Different: Adversary Adapts

Document
Lorem ipsum dolor sit amet.
Consectetuer adipiscing elit. Cras non nunc nec
enim tristique tincidunt. Vestibulum quis tellus.
Duis nulla. Donec luctus urna. Sed tempus nibh
id massa. Vivamus placerat justo quis nibh. Ut
quis ante. Ut sollicitudin quam eu mi. Donec
molestie purus sit amet velit. Sed ac sem.
Aenean quis justo. Vestibulum ante ipsum primis
in faucibus orci luctus et ultrices posuere cubilia
Curae; Ut tincidunt.
Nulla facilisi. Aenean eros felis, blandit eu,
commodo sit amet, varius a, pede. Curabitur
augue felis, congue sed.

Goal: Understand classifiers under attack.

Results: Vulnerable to automated evasion.

Building Machine Learning Classifiers



Training
(Supervised Learning)

Labelled Training Data

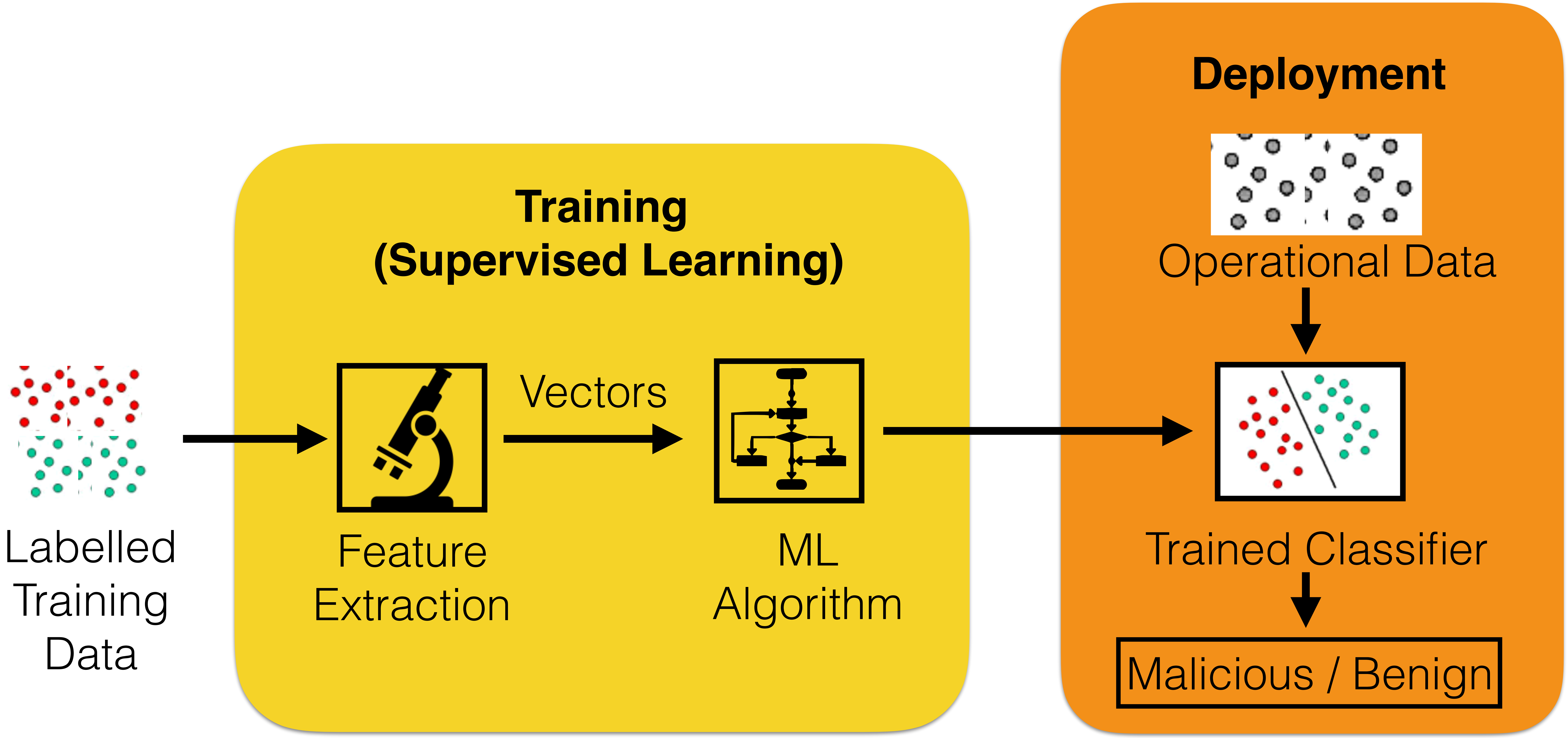
Feature Extraction

Vectors

ML Algorithm

Trained Classifier

Assumption: Training Data is Representative



Training

(Supervised Learning)

Deployment

Labelled Training Data

Feature Extraction

Vectors

ML Algorithm

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

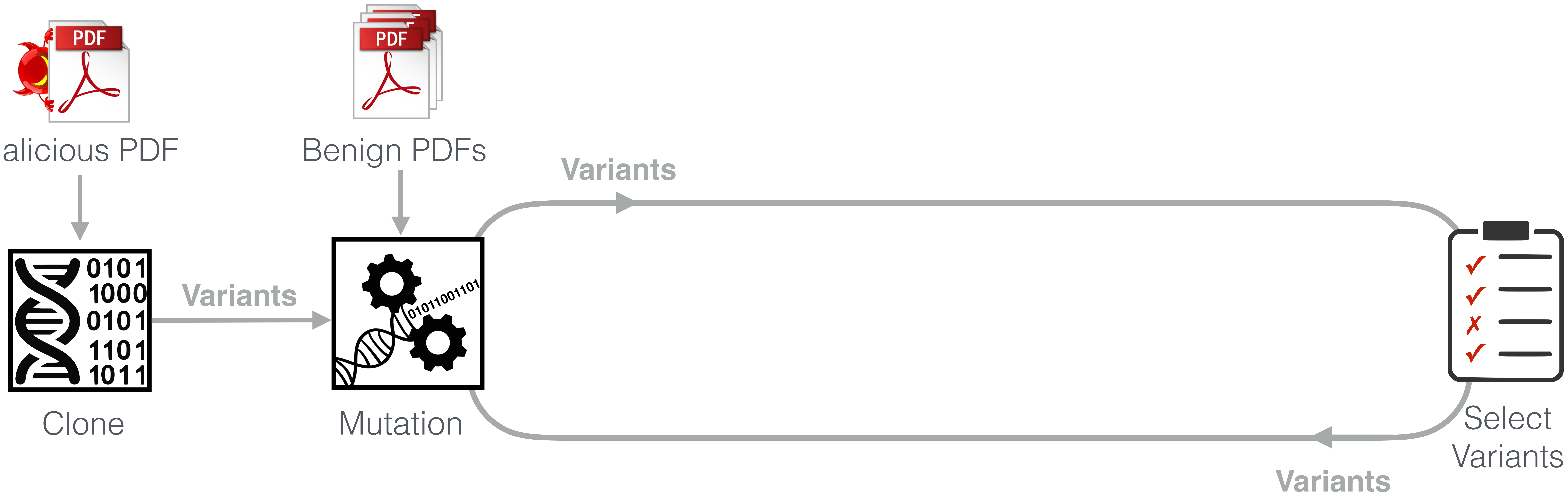
Very robust against "strongest conceivable mimicry attack".

	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.

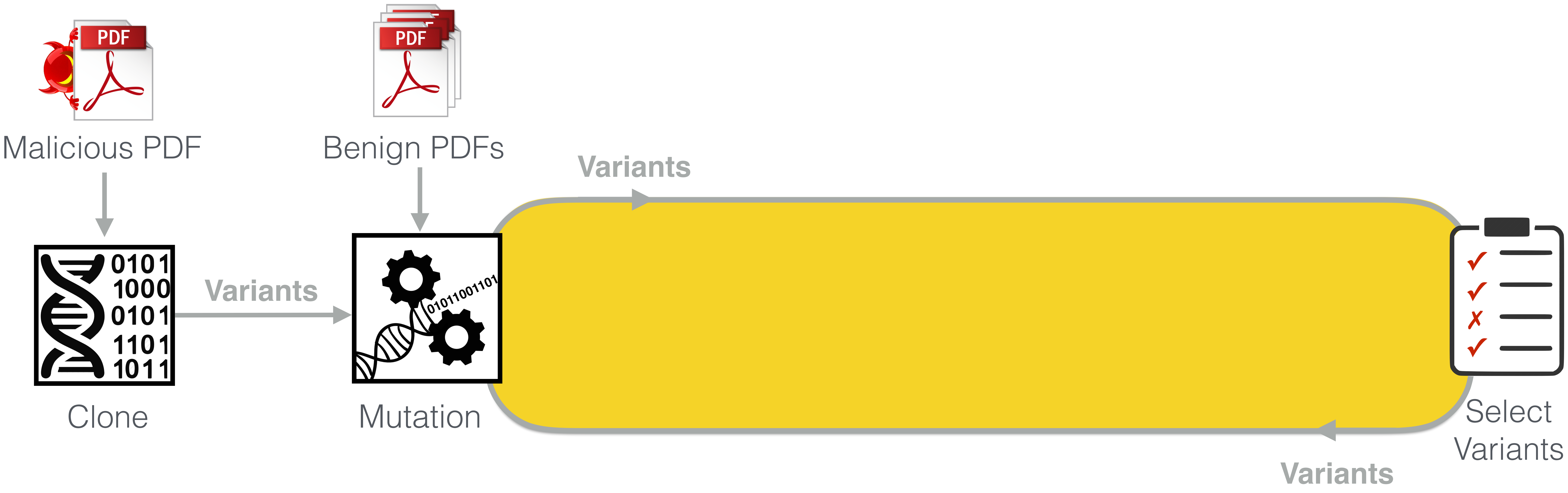
Automated Evasion Approach

Based on Genetic Programming



Automated Evasion Approach

Based on Genetic Programming



Malicious PDF

Benign PDFs

Clone

Variants

Mutation

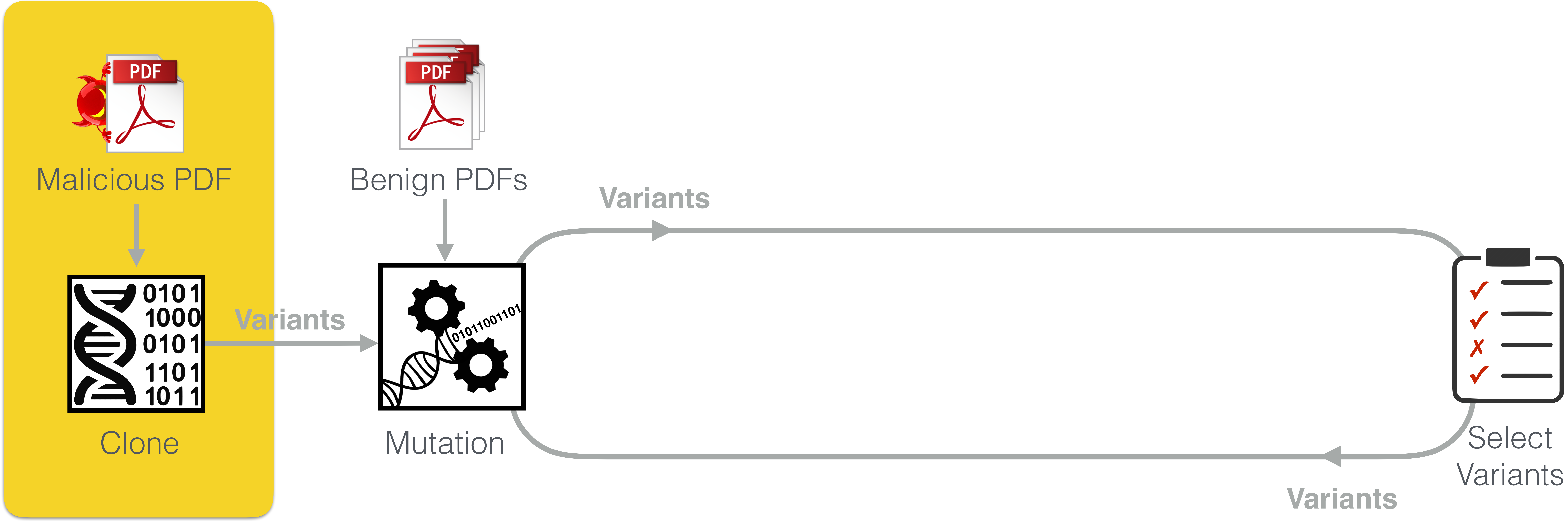
Variants

Select Variants

Variants

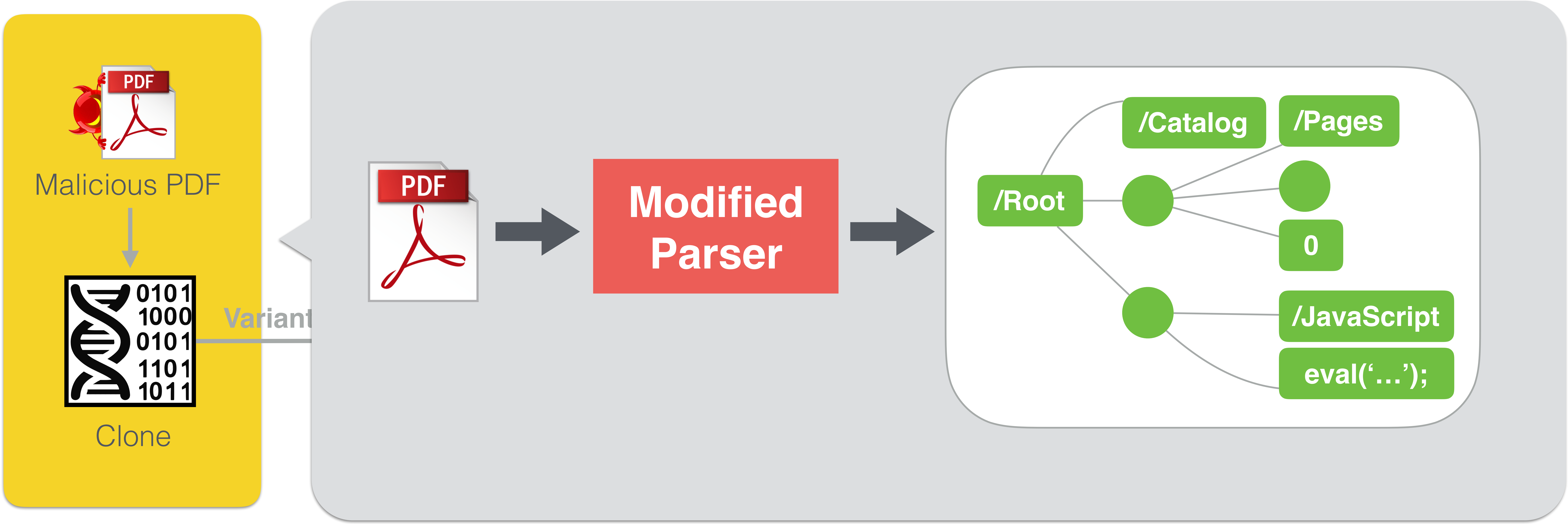
Automated Evasion Approach

Based on Genetic Programming



Automated Evasion Approach

Based on Genetic Programming

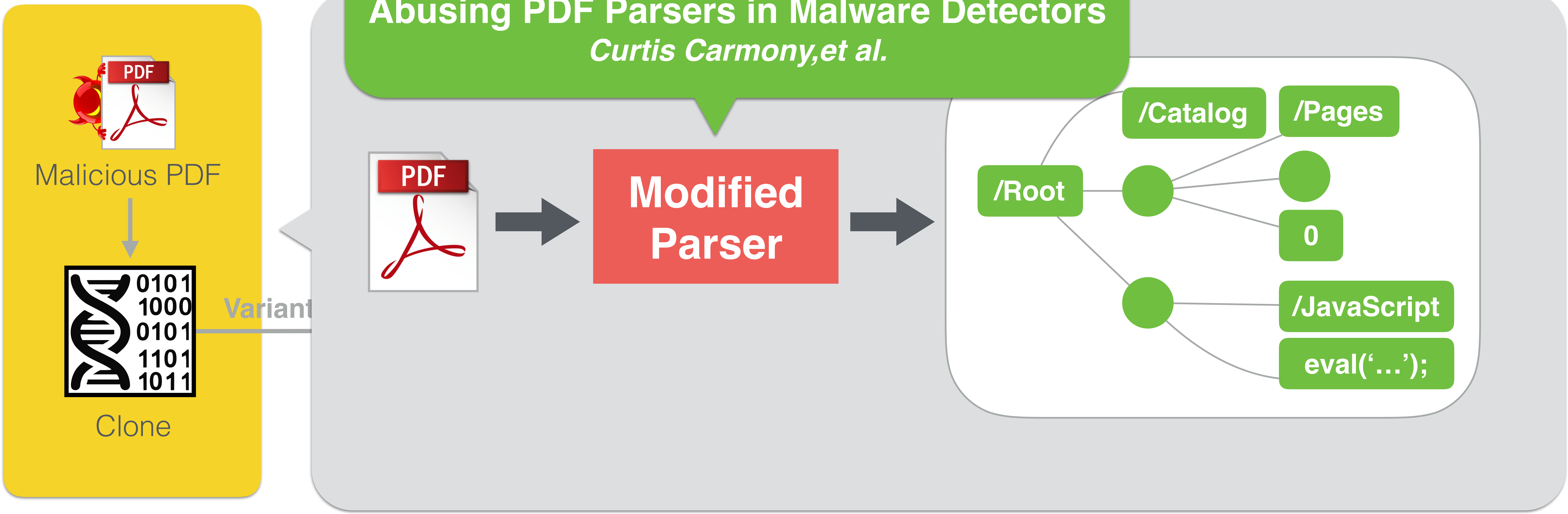


Automated Evasion Approach

Extract Me If You Can:

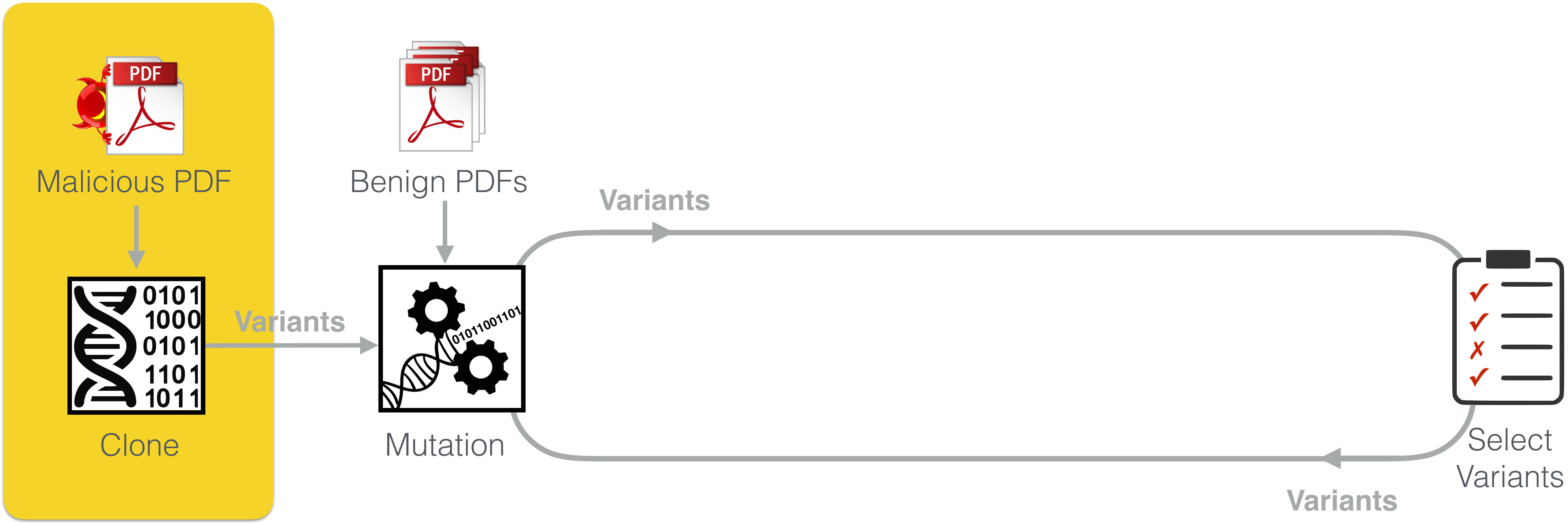
Abusing PDF Parsers in Malware Detectors

Curtis Carmony,et al.



Automated Evasion Approach

Based on Genetic Programming



Malicious PDF

Benign PDFs

Variants

Clone

Variants

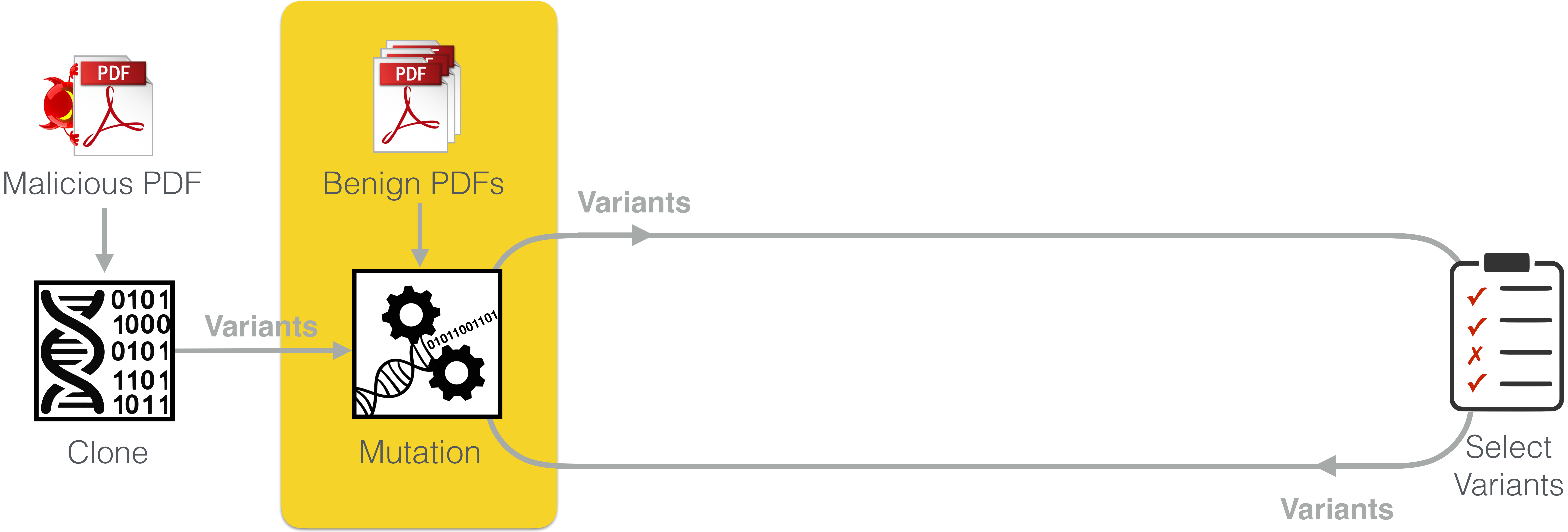
Mutation

Variants

Select Variants

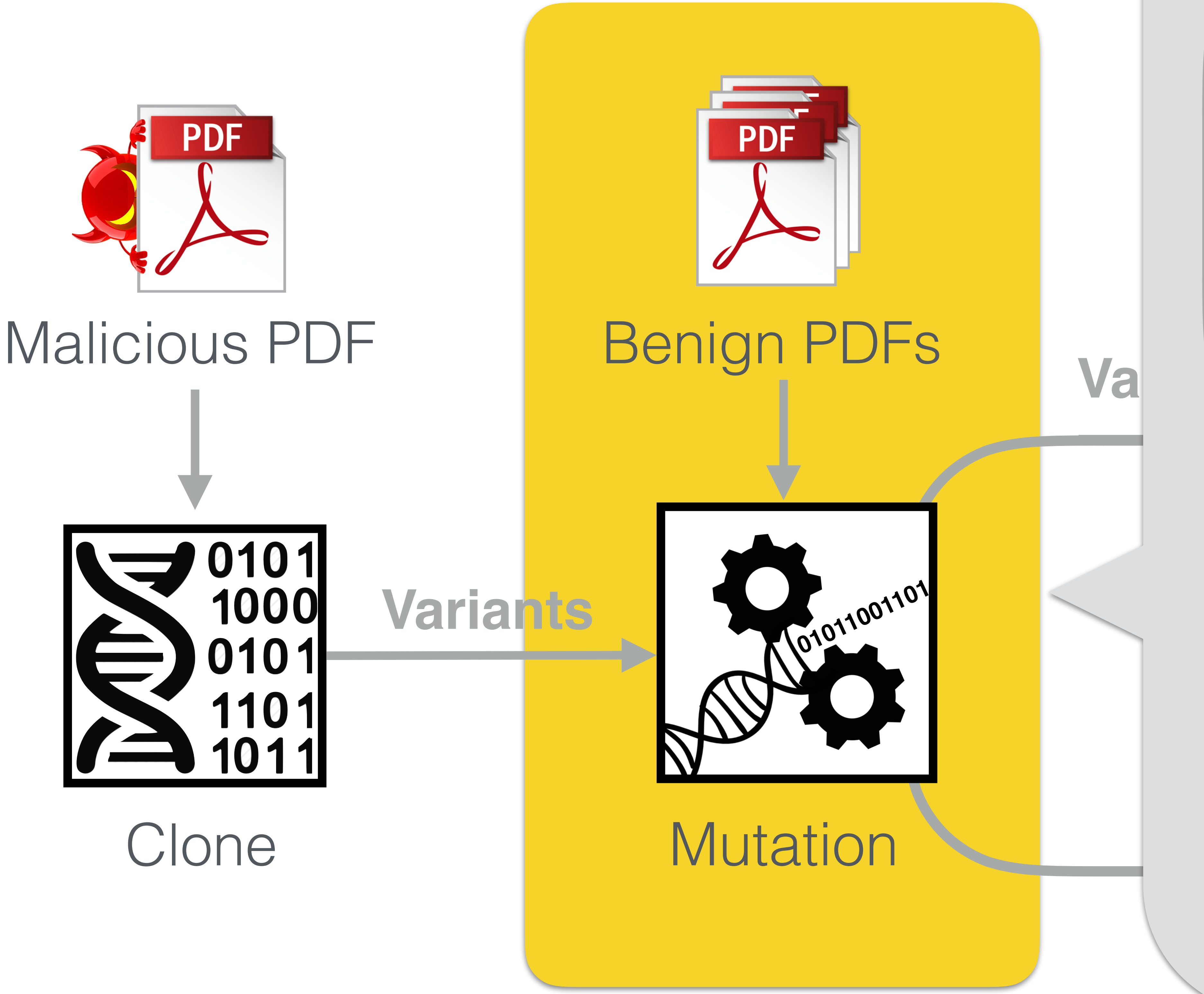
Automated Evasion Approach

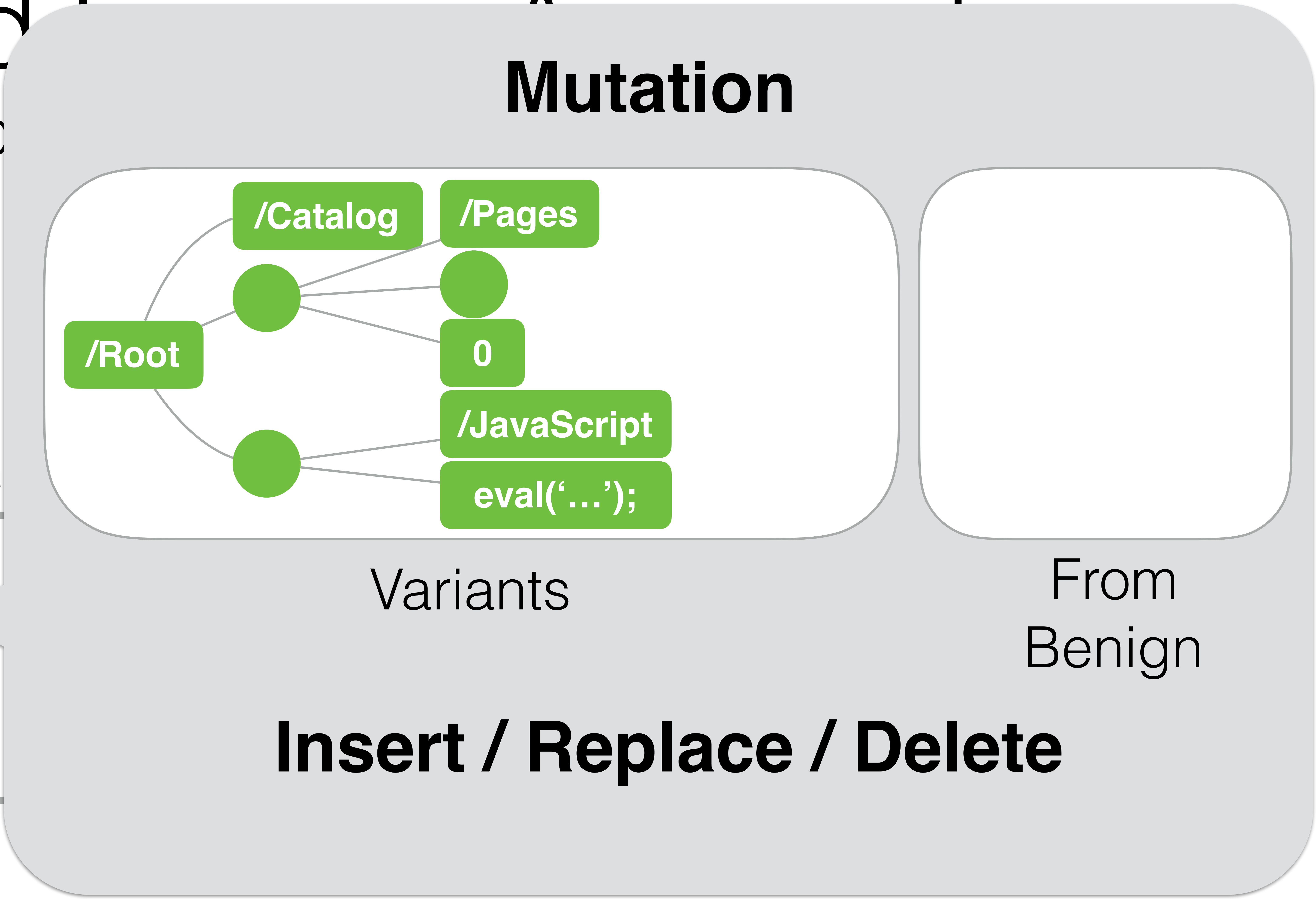
Based on Genetic Programming



Automated

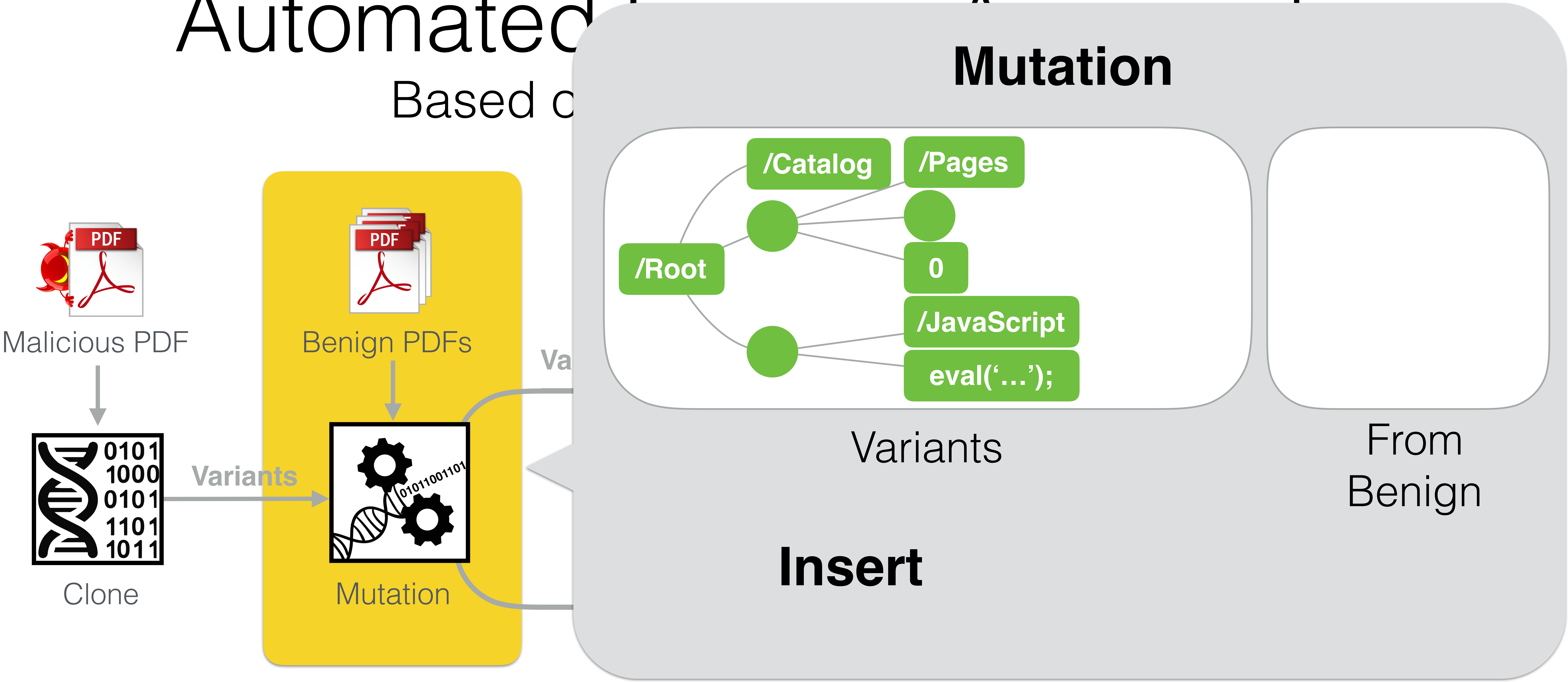
Based c





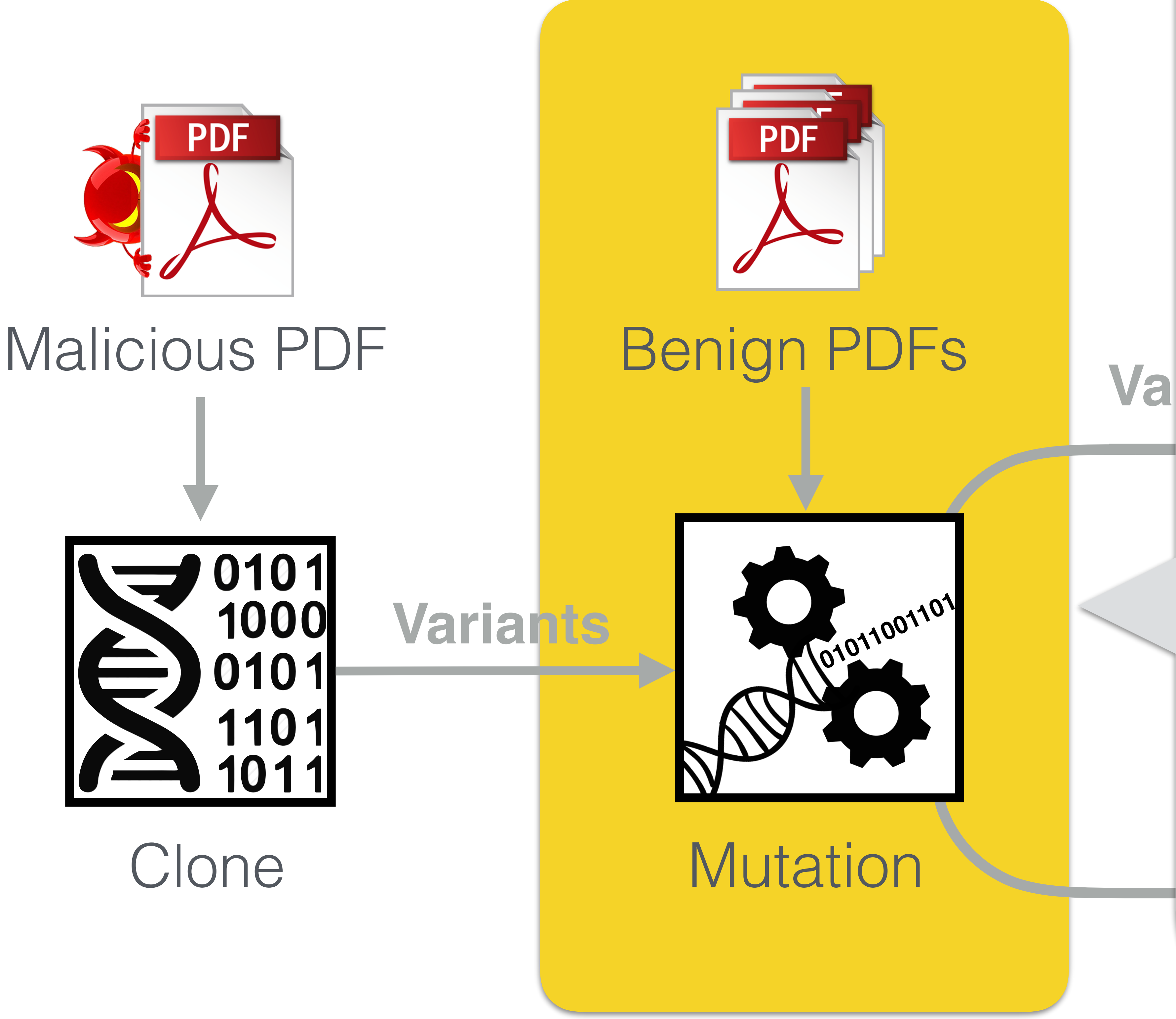
Automated

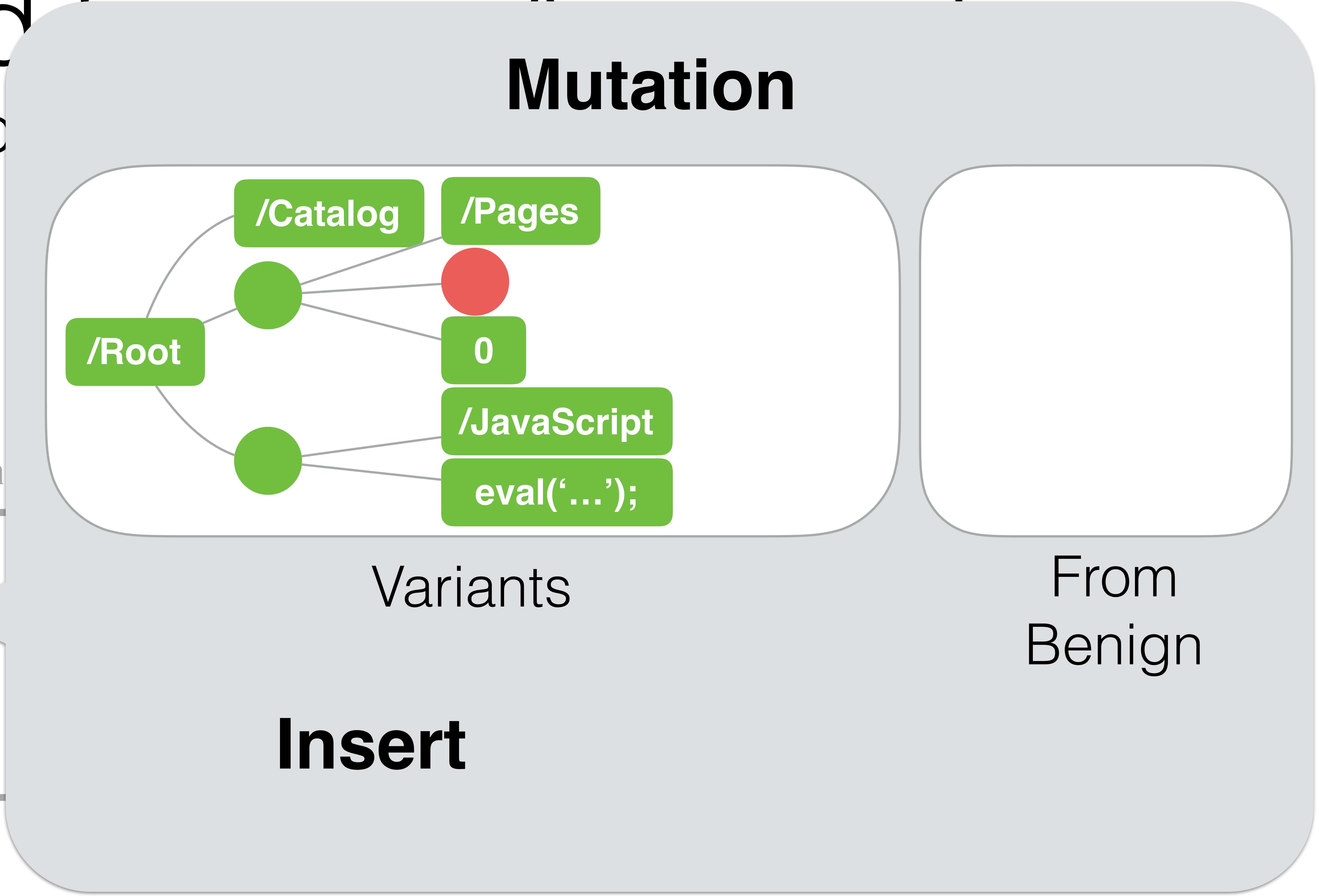
Based c



Automated

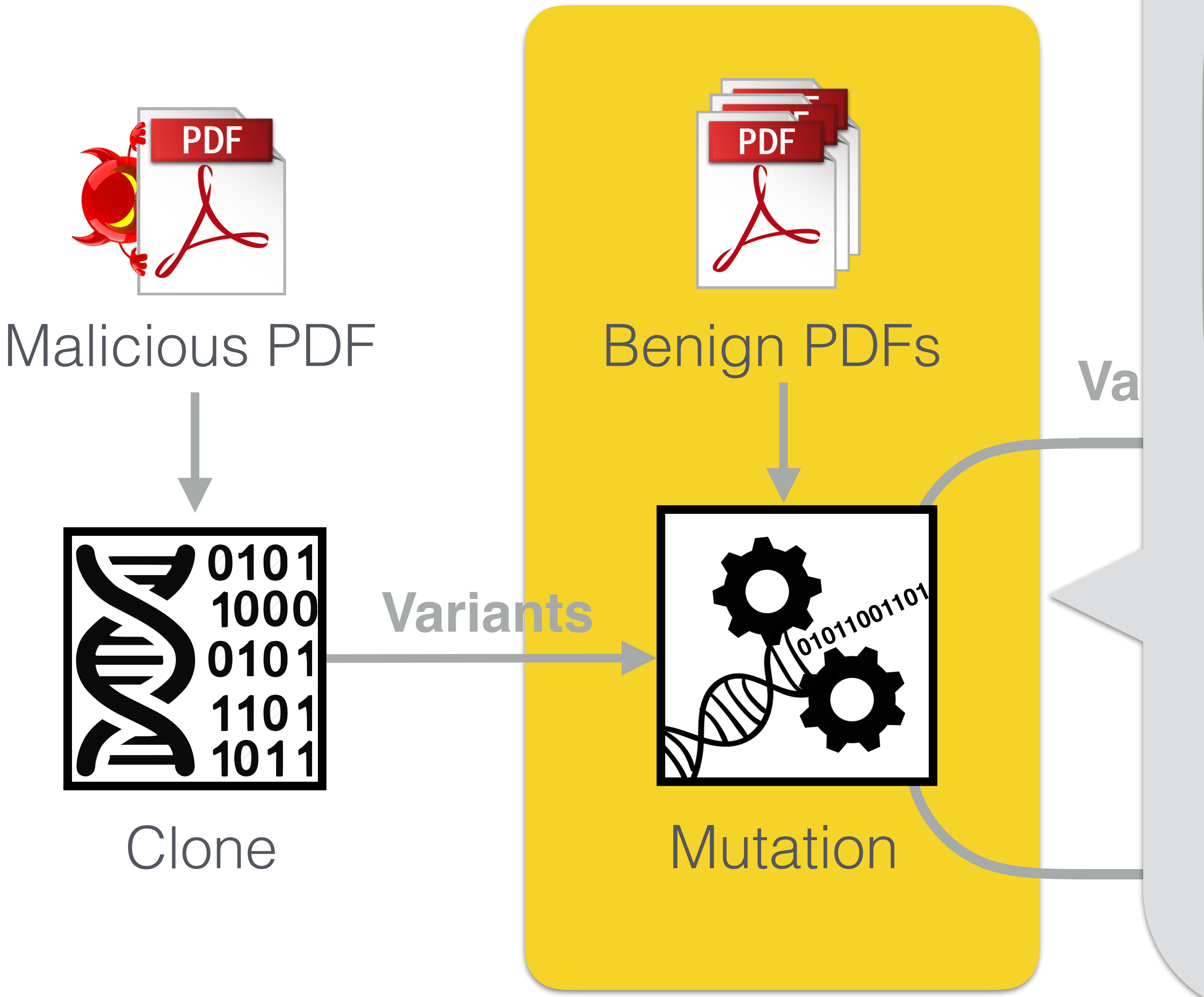
Based c



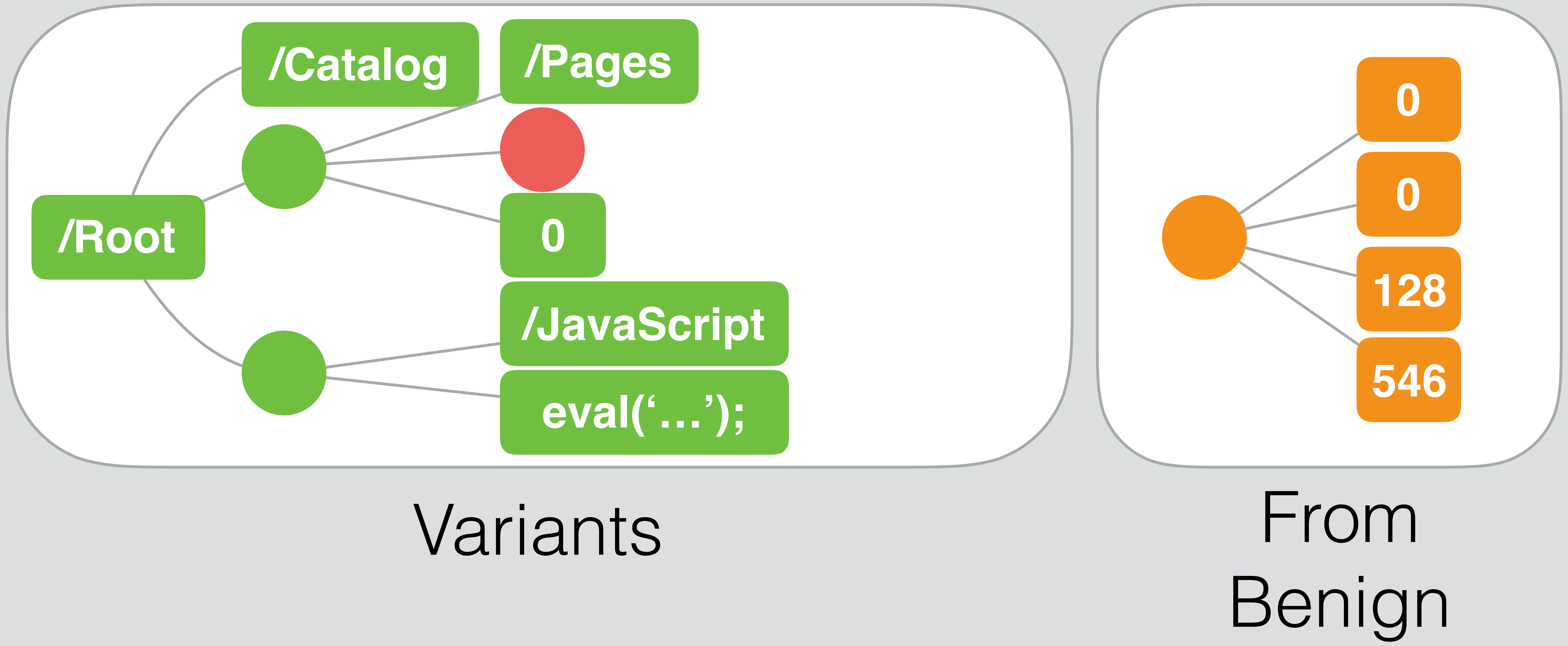


Automated

Based o



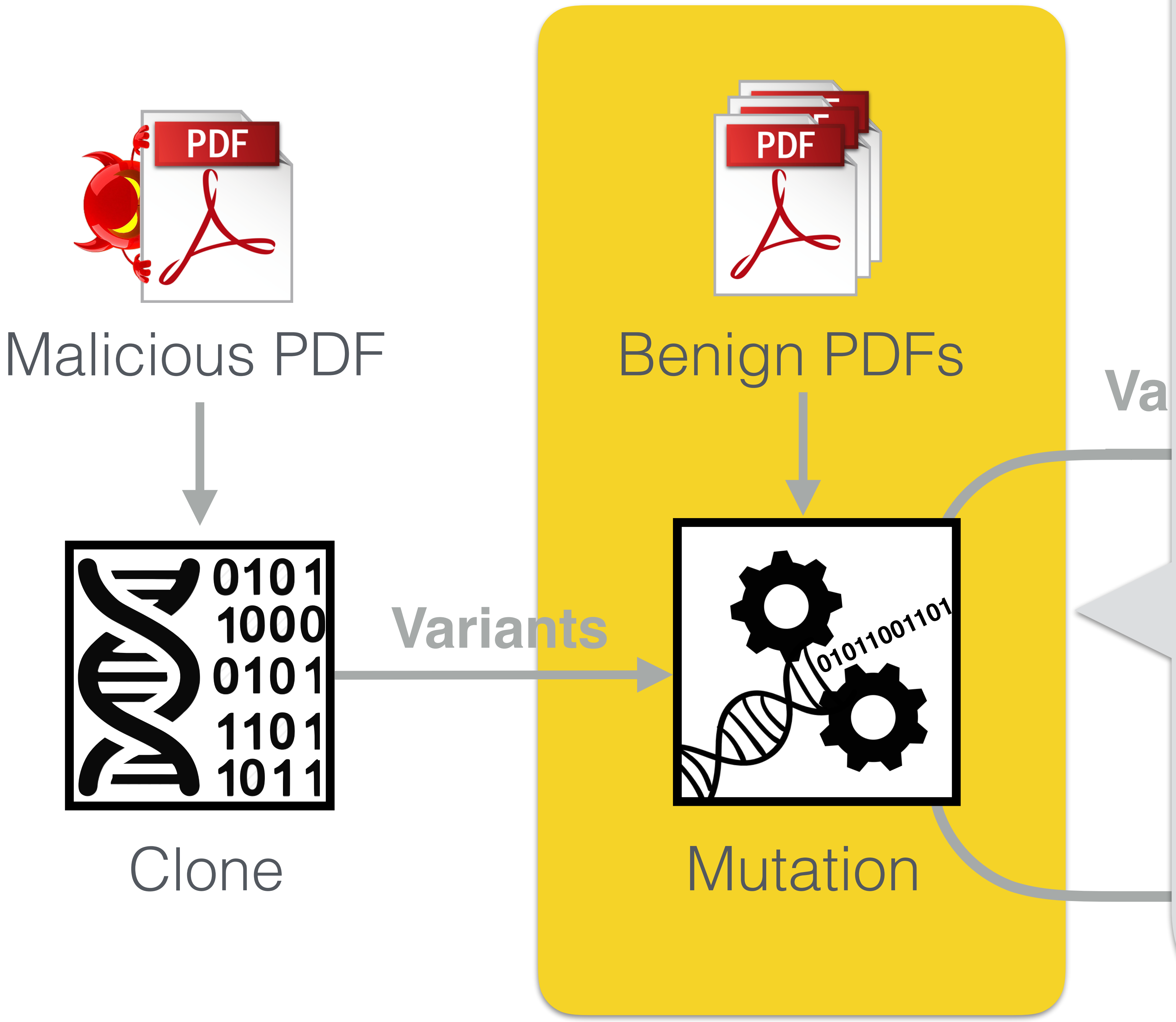
Mutation



Insert

Automated

Based c



Malicious PDF

Clone

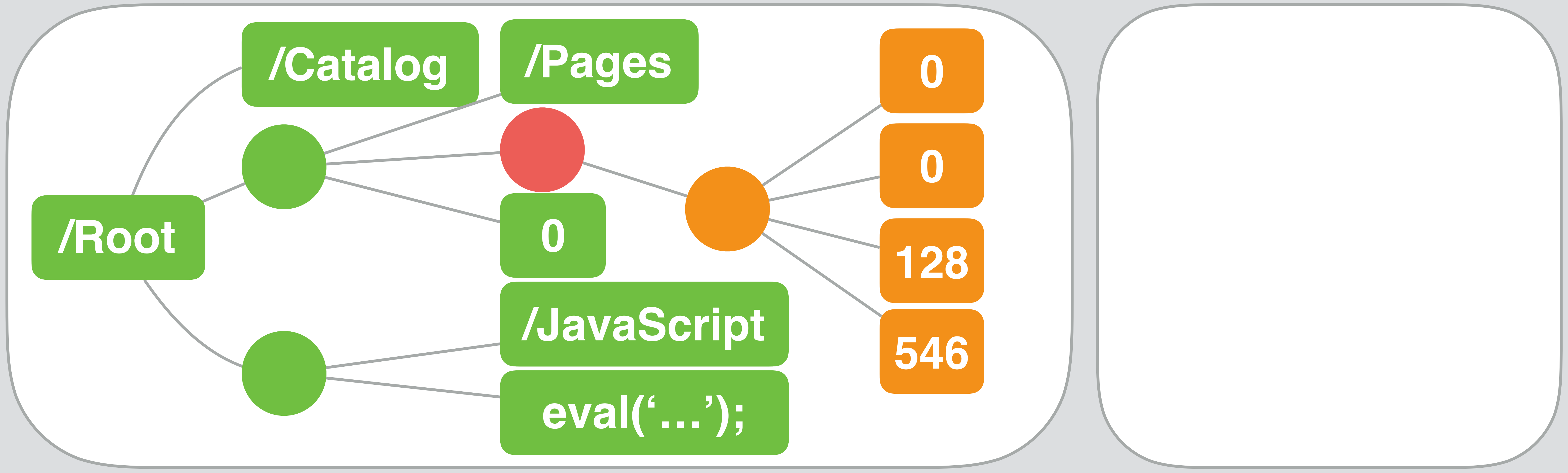
Variants

Benign PDFs

Mutation

Va

Mutation



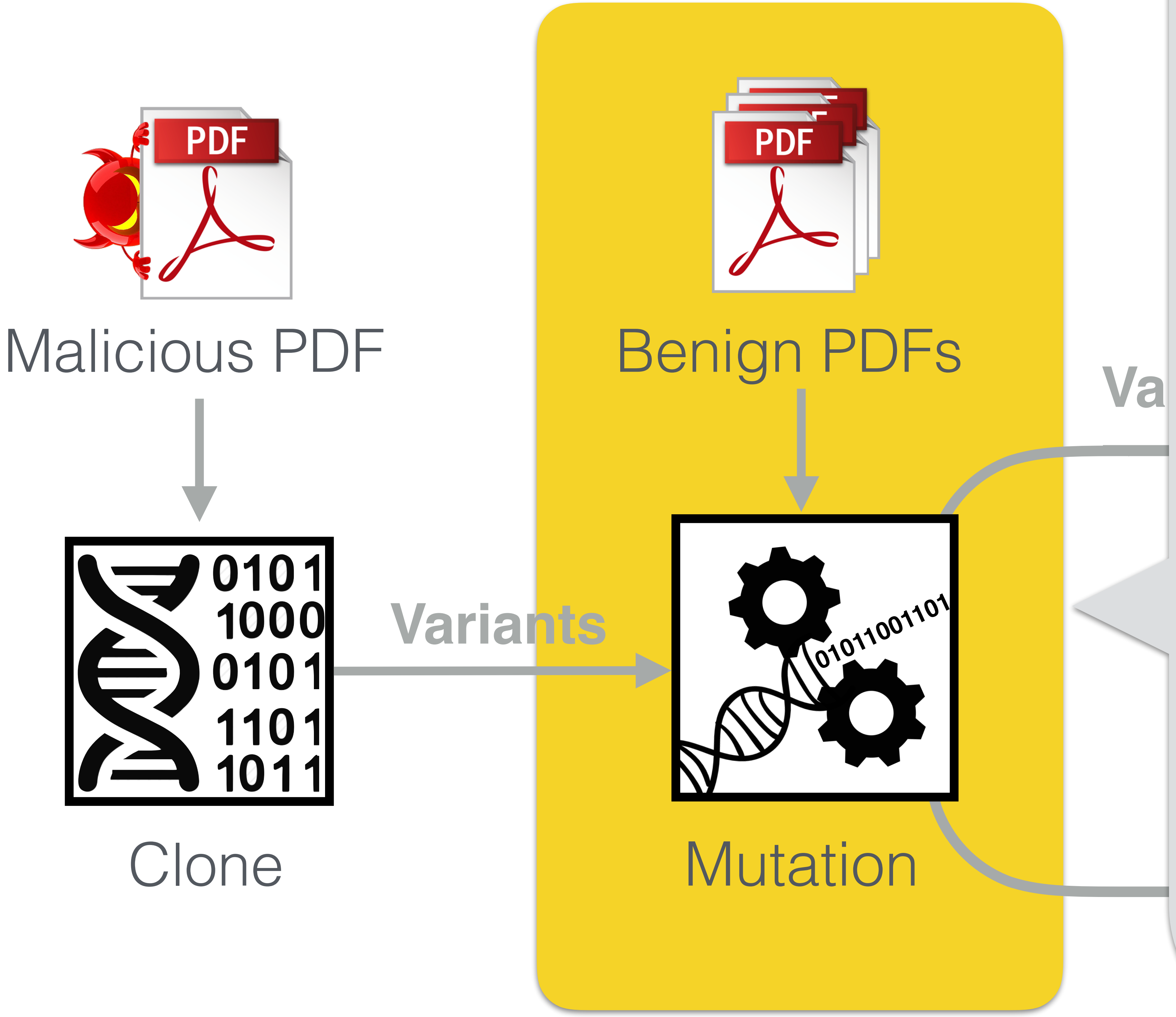
Variants

From Benign

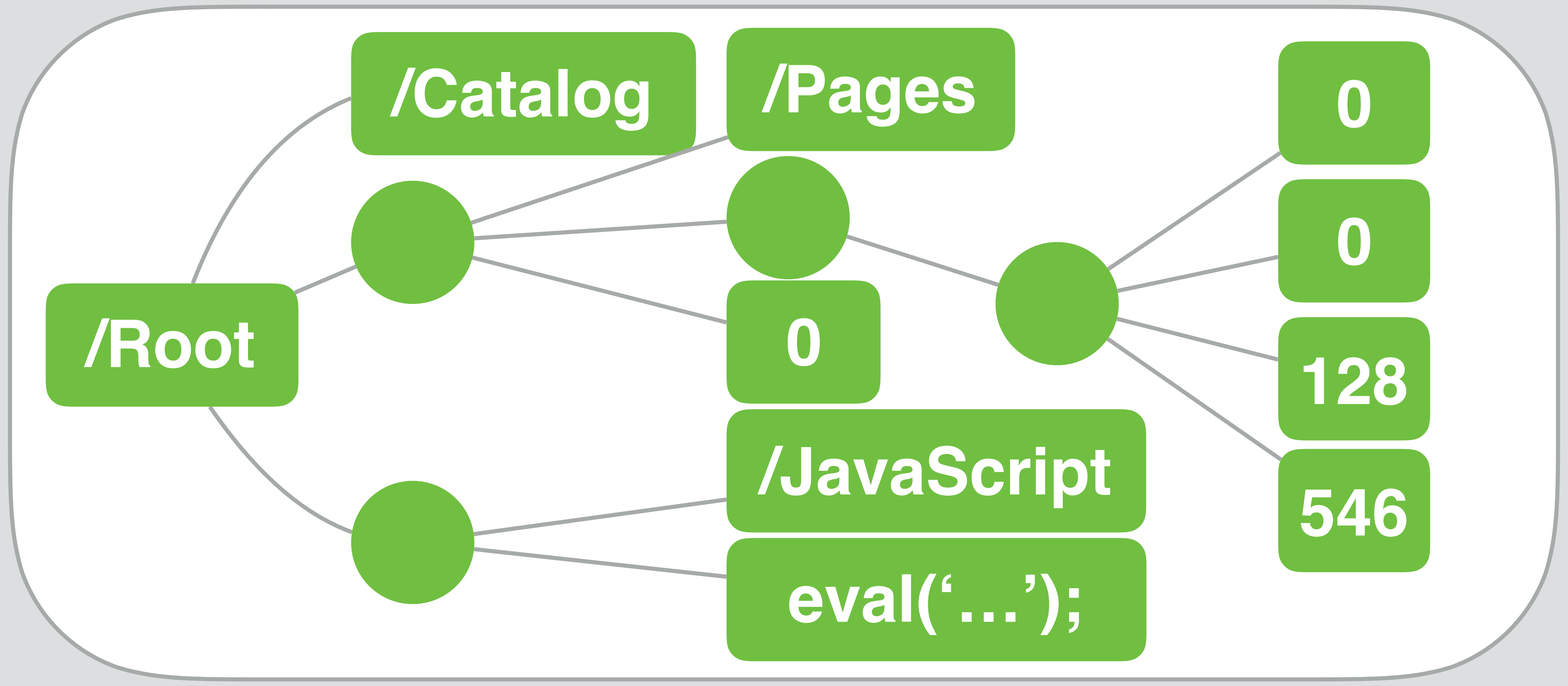
Insert

Automated

Based c



Mutation



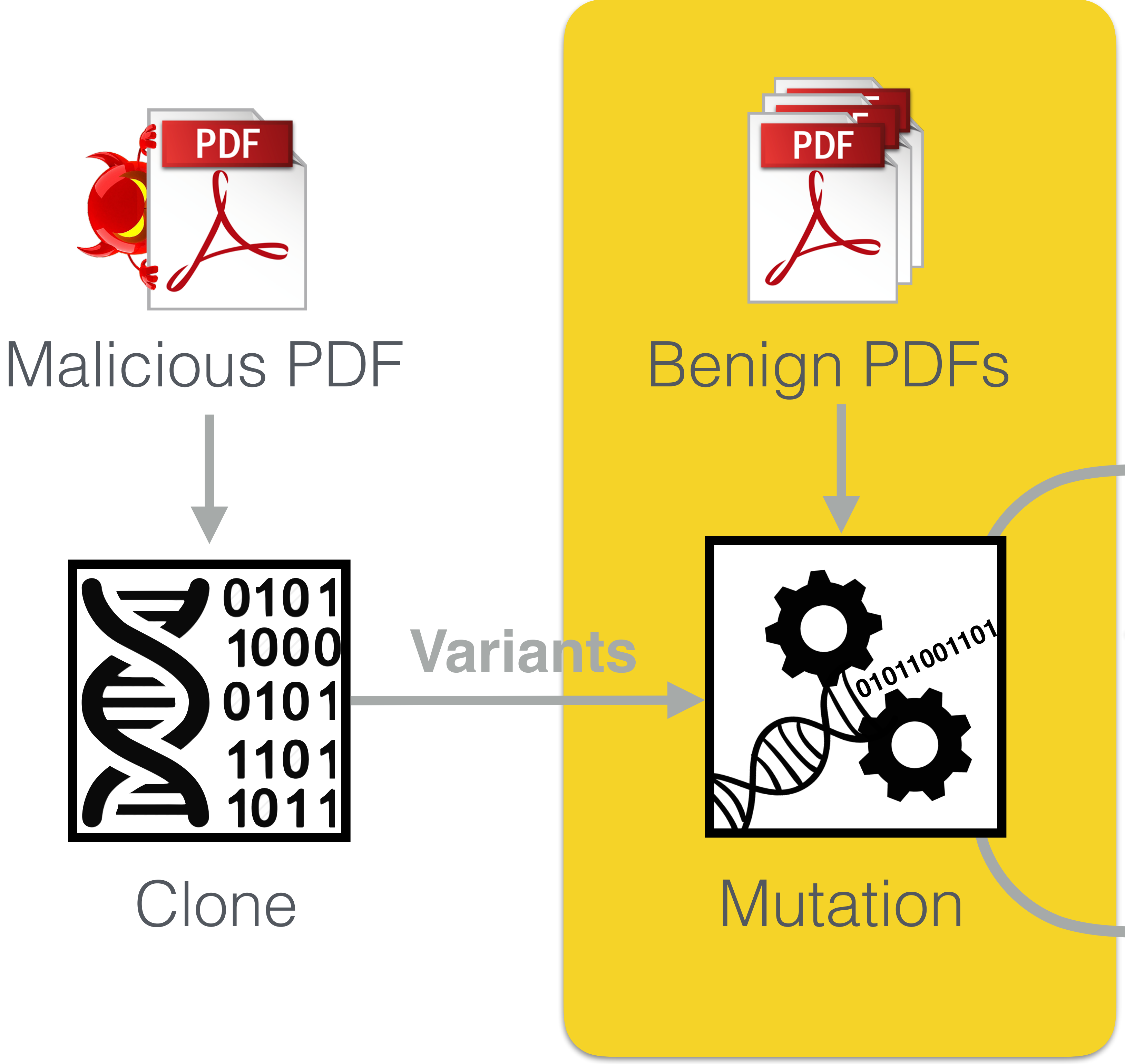
Variants

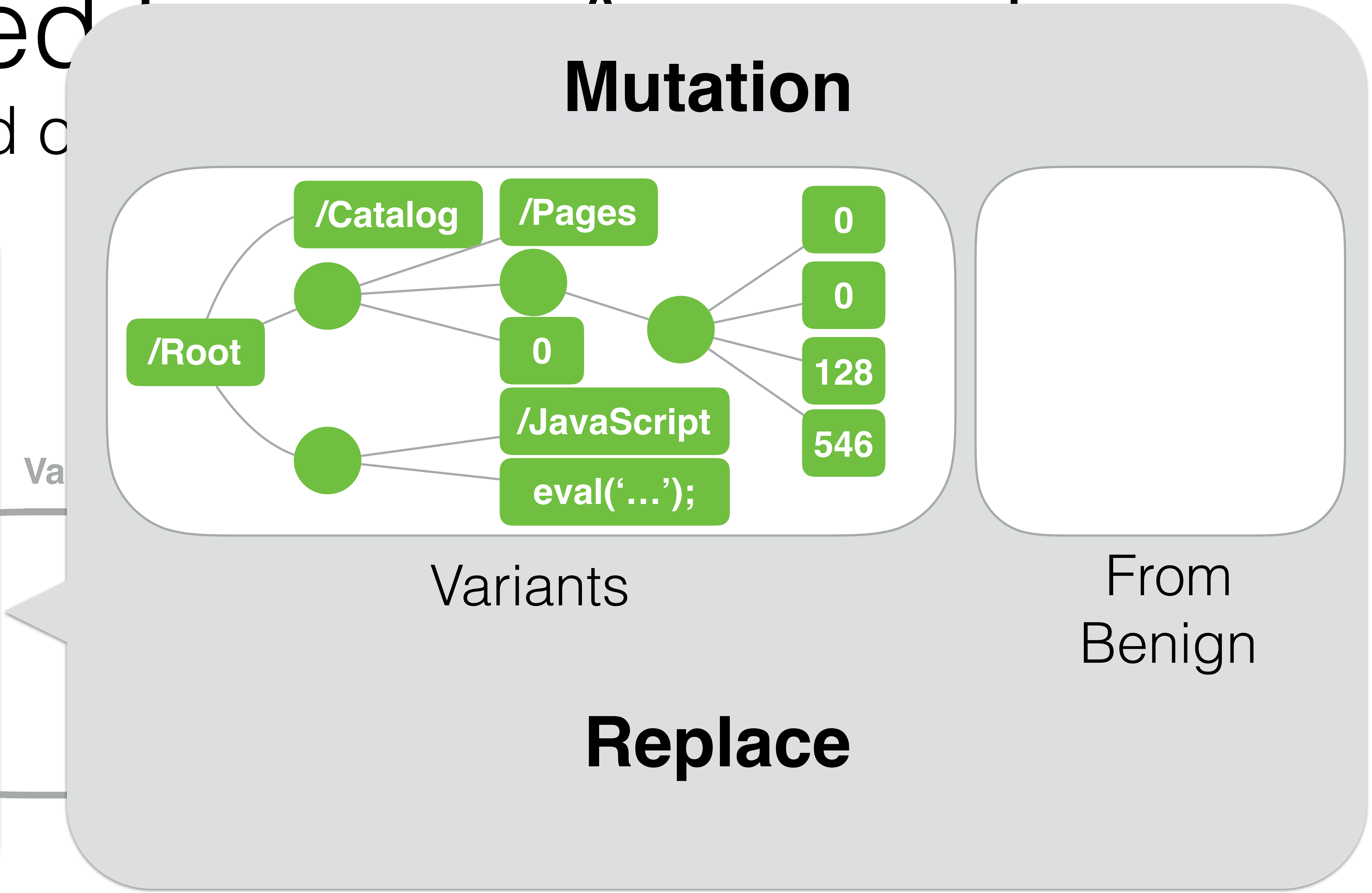
From
Benign

Insert

Automated

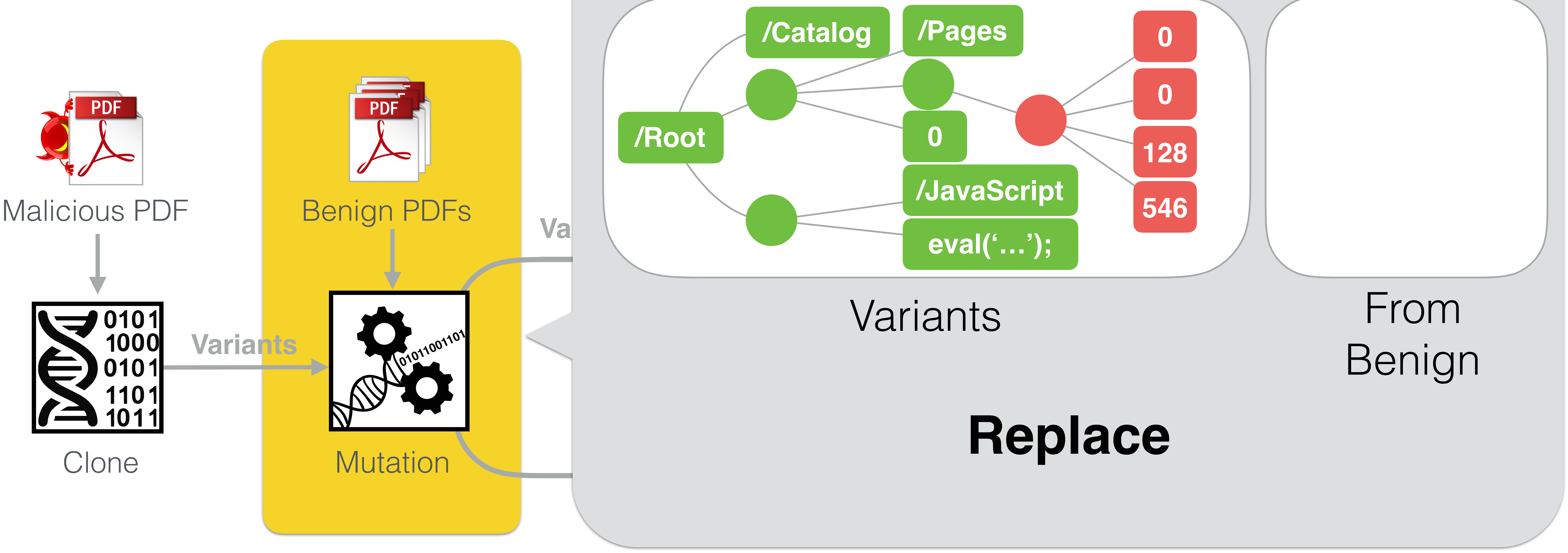
Based c





Automated

Based o



Malicious PDF

Clone

Variants

Benign PDFs

Mutation

Va

Mutation

/Root

/Catalog

/Pages

0

/JavaScript

eval('...');

0

0

128

546

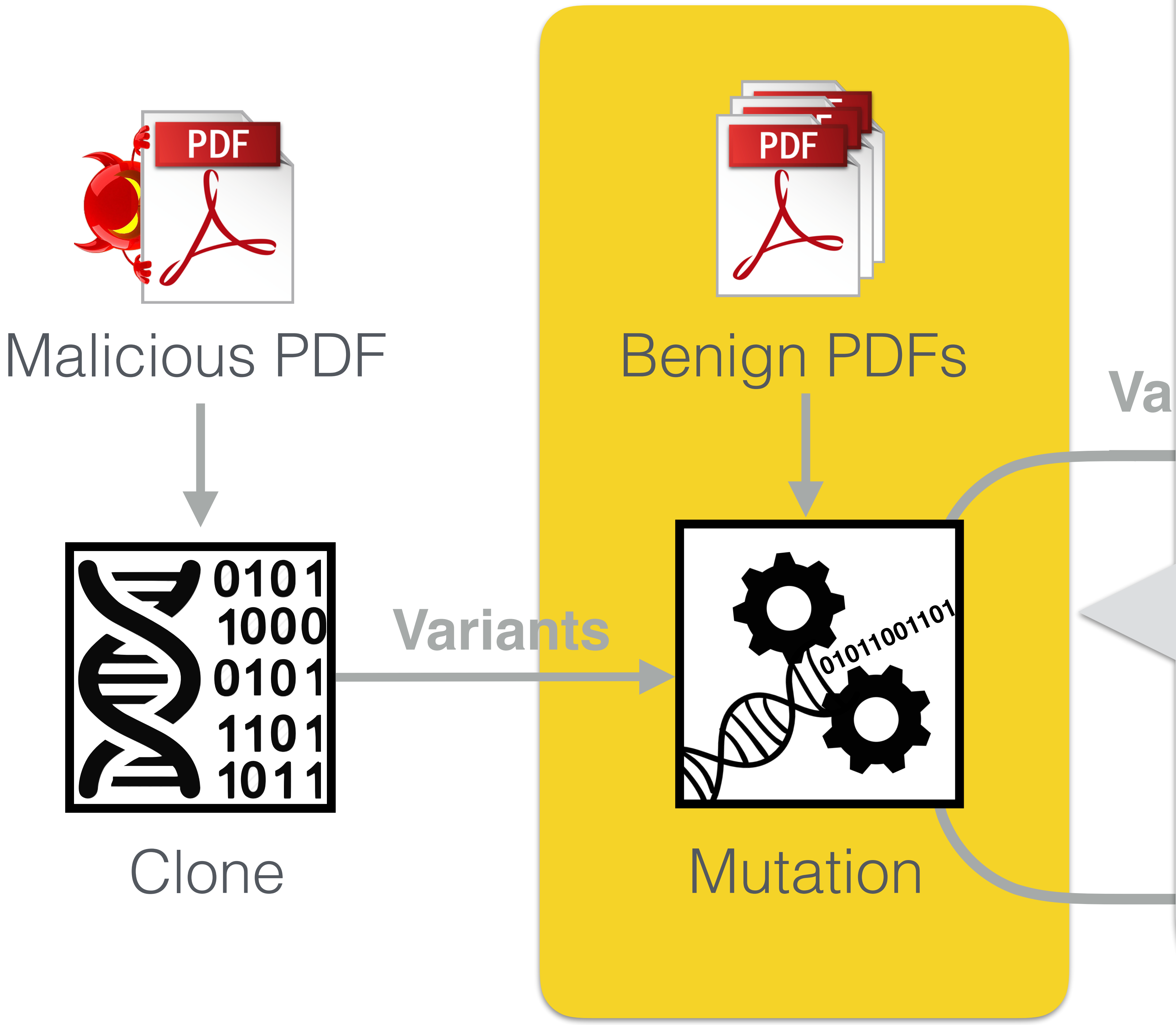
Variants

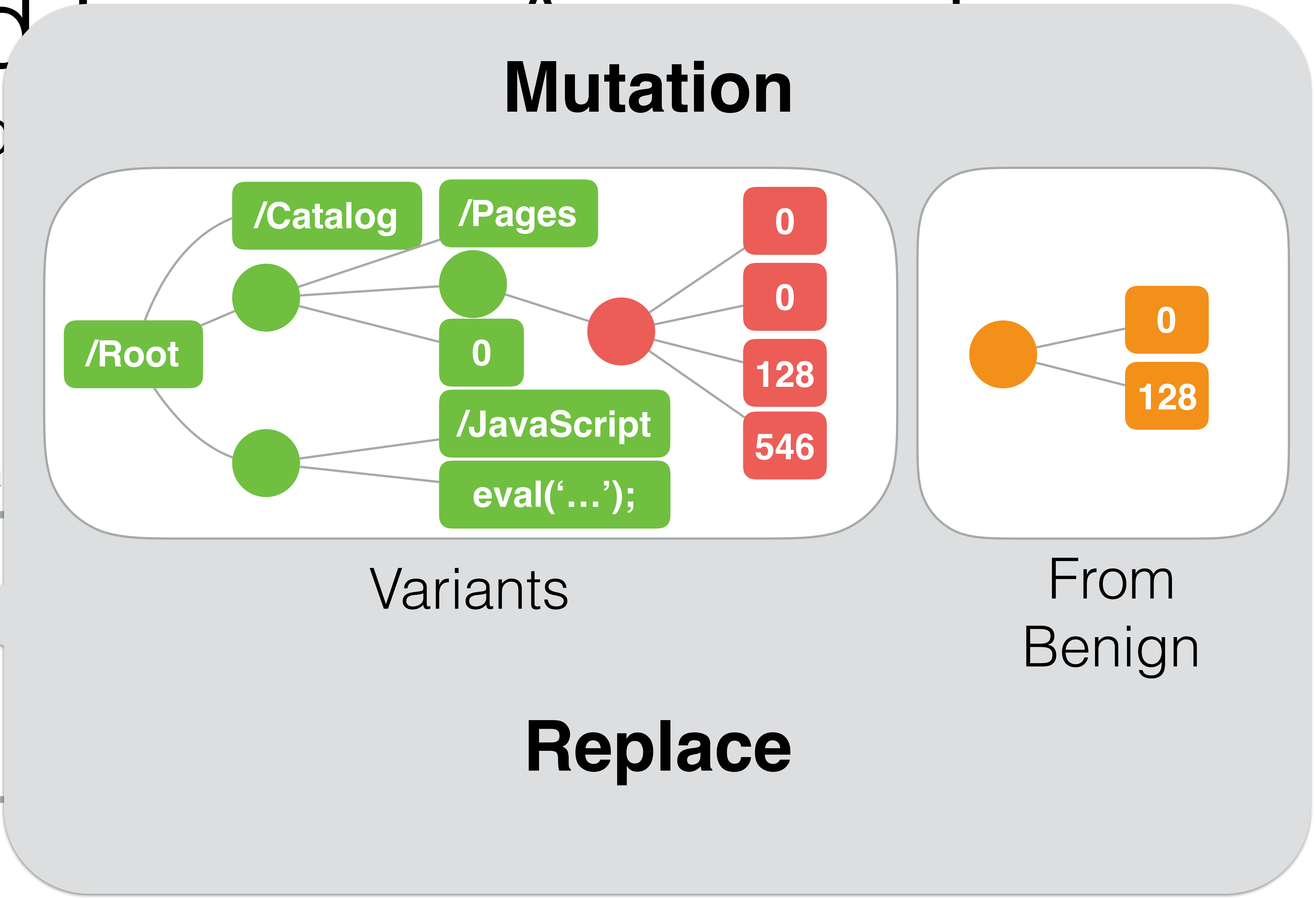
From Benign

Replace

Automated

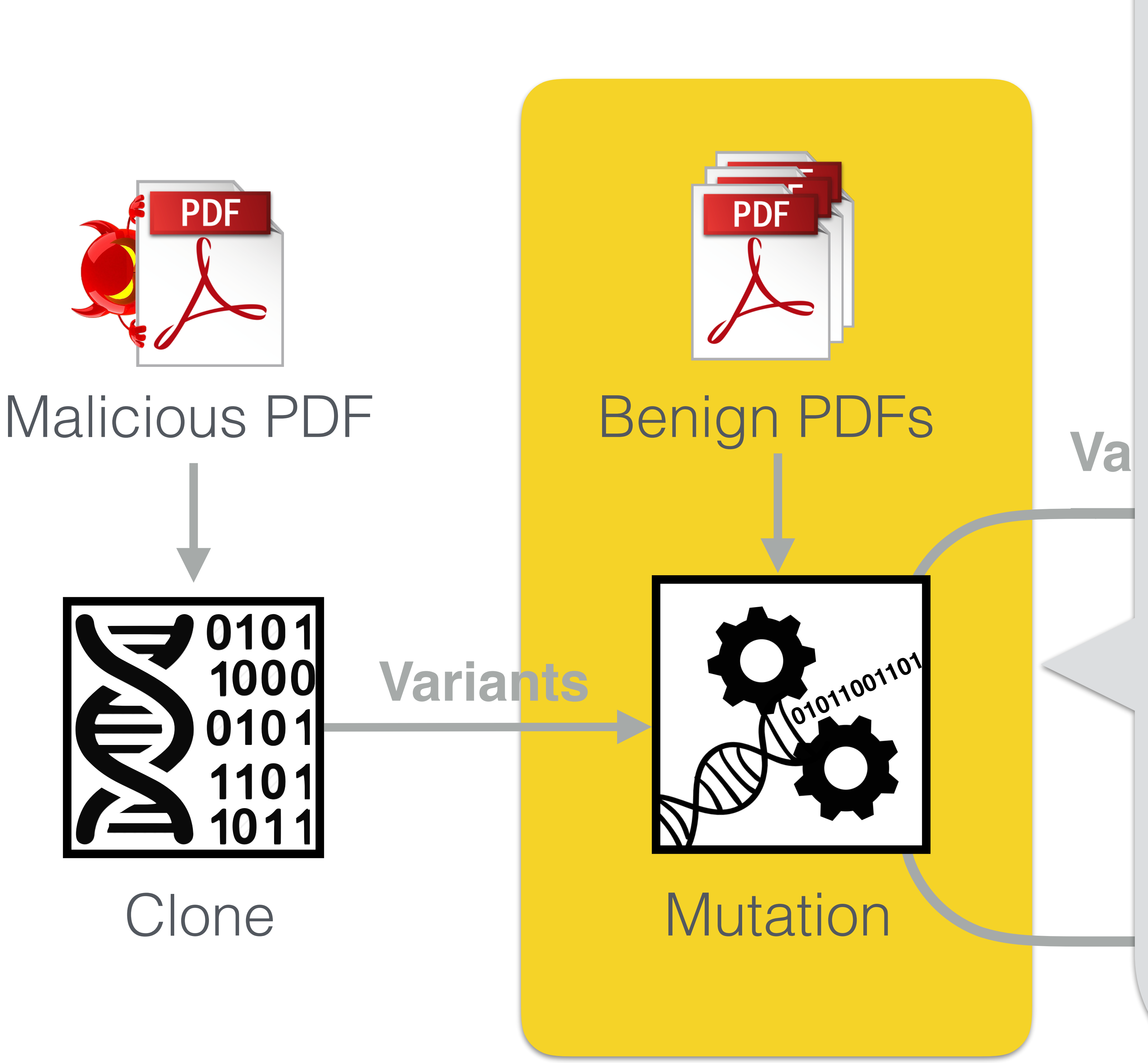
Based o



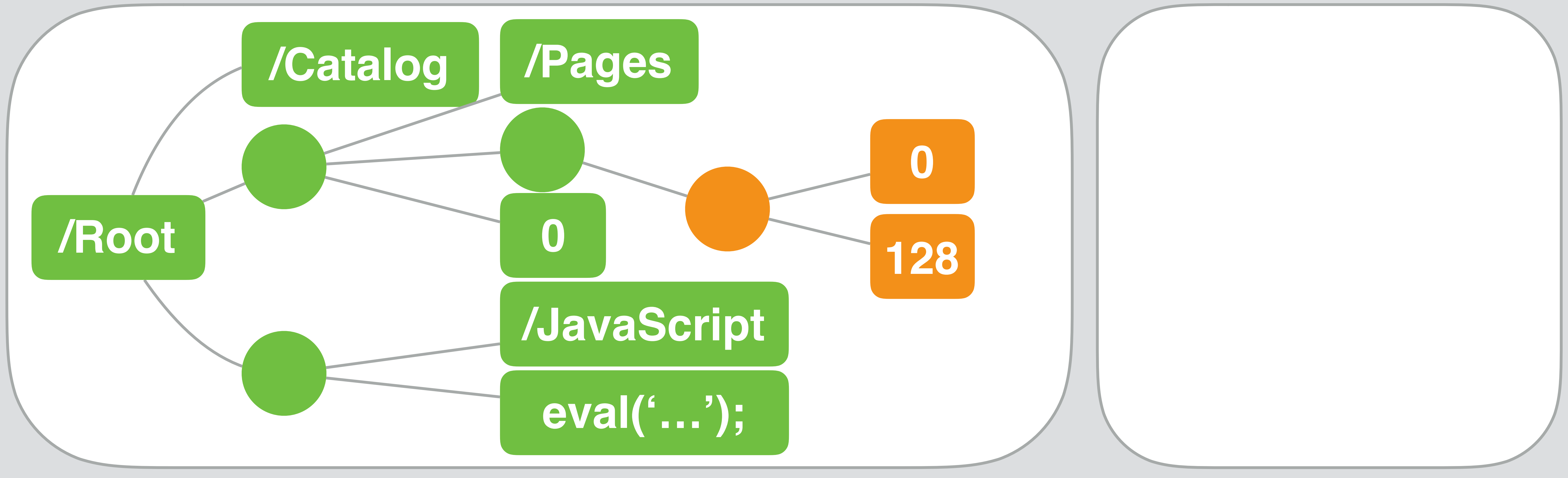


Automated E

Based o



Mutation



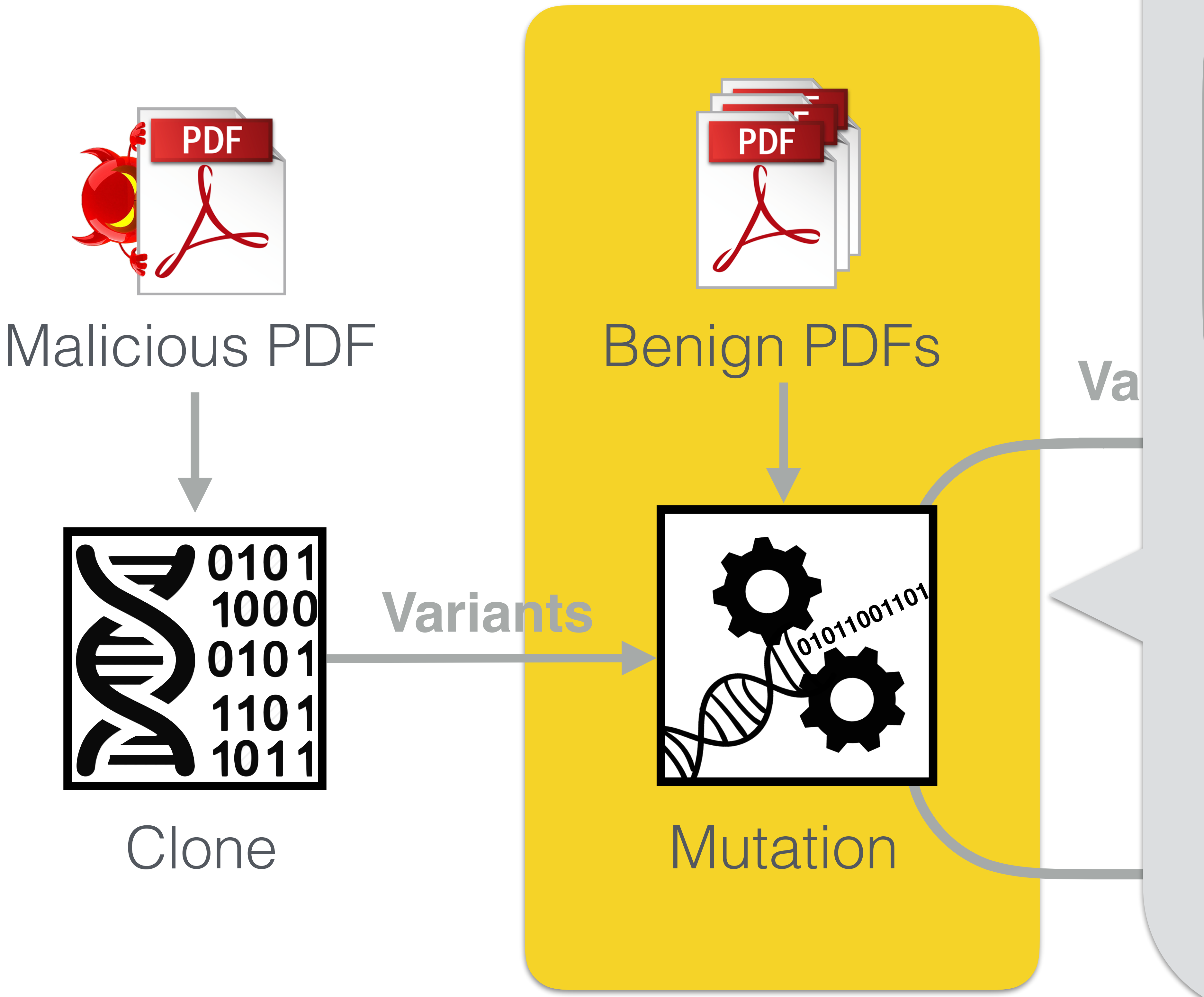
Variants

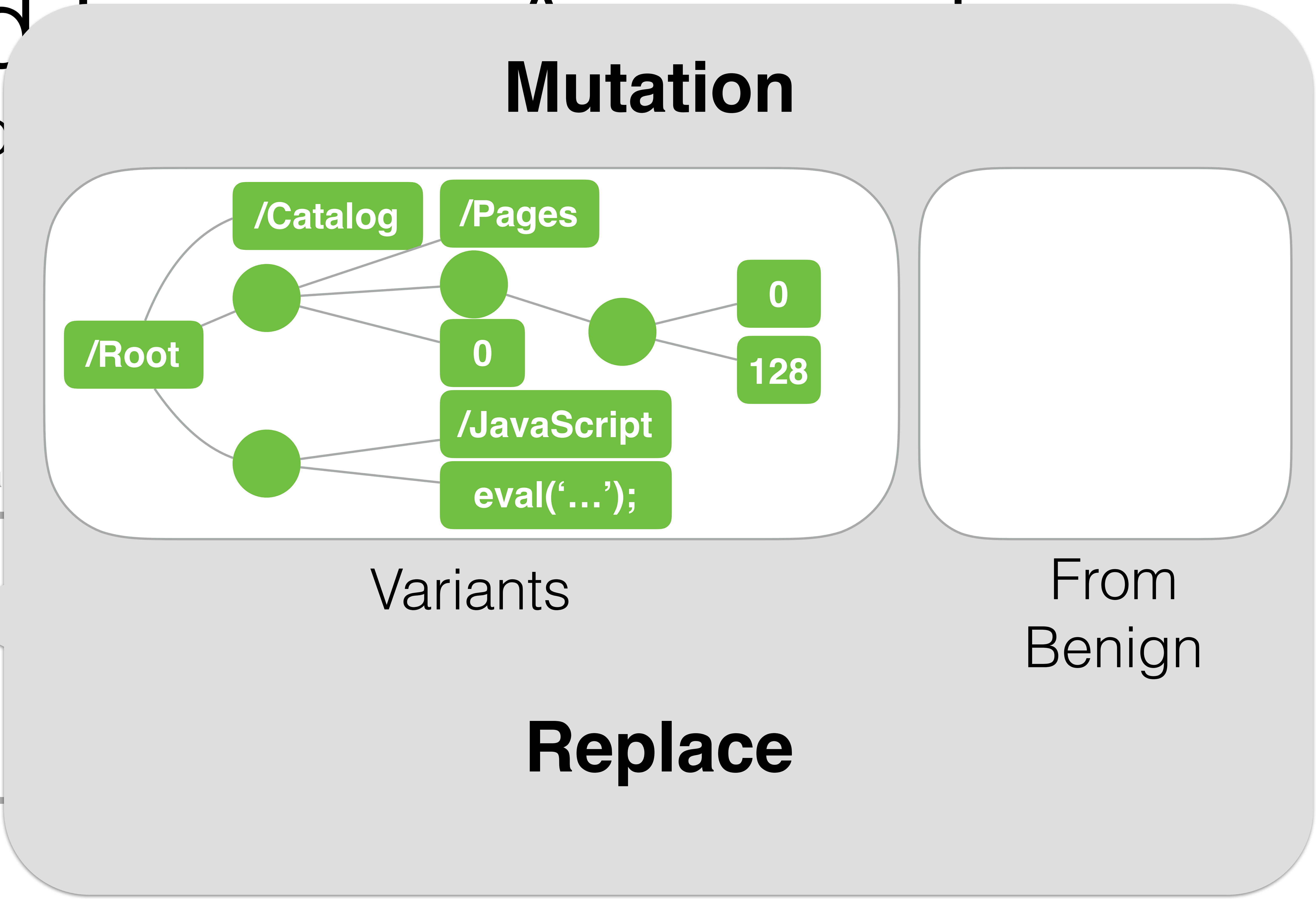
From Benign

Replace

Automated

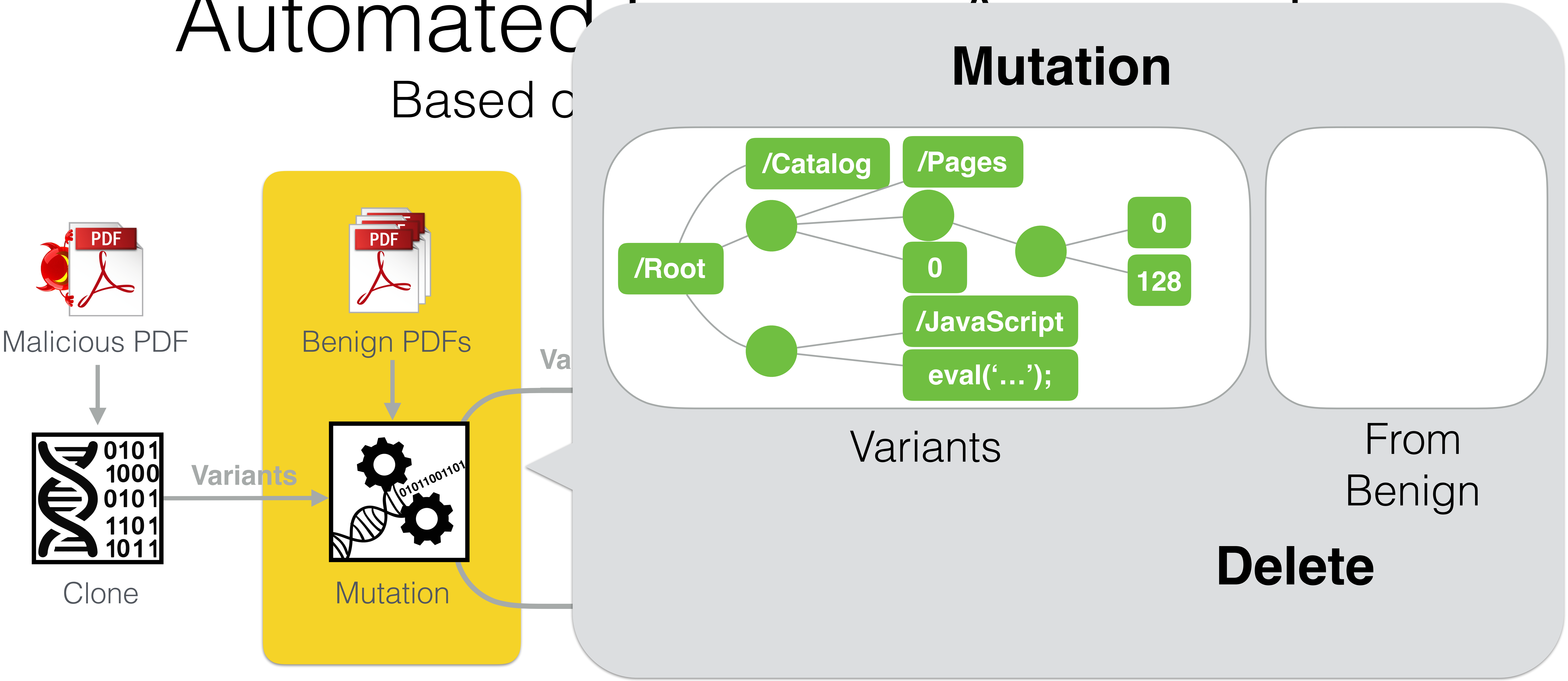
Based c





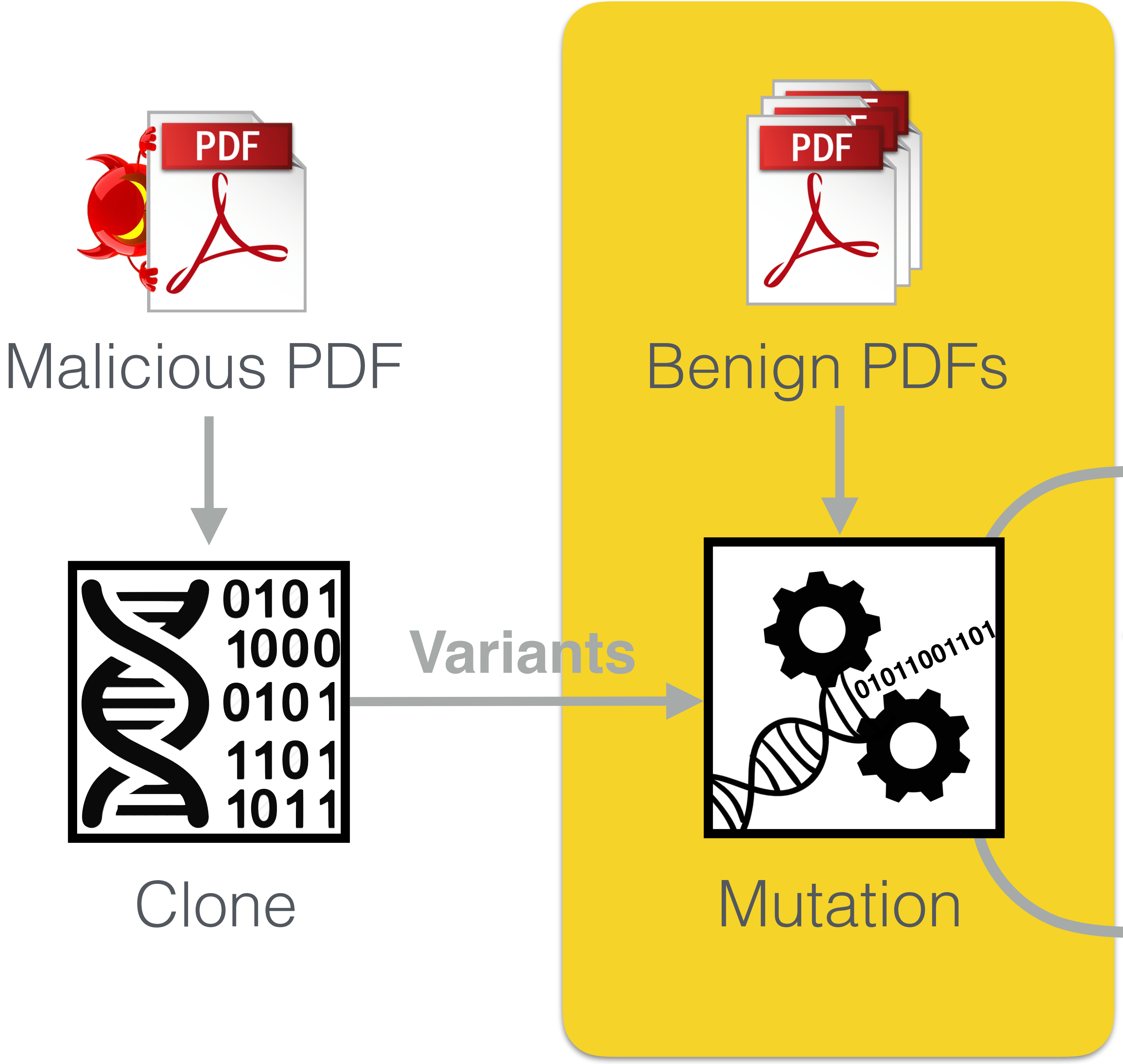
Automated

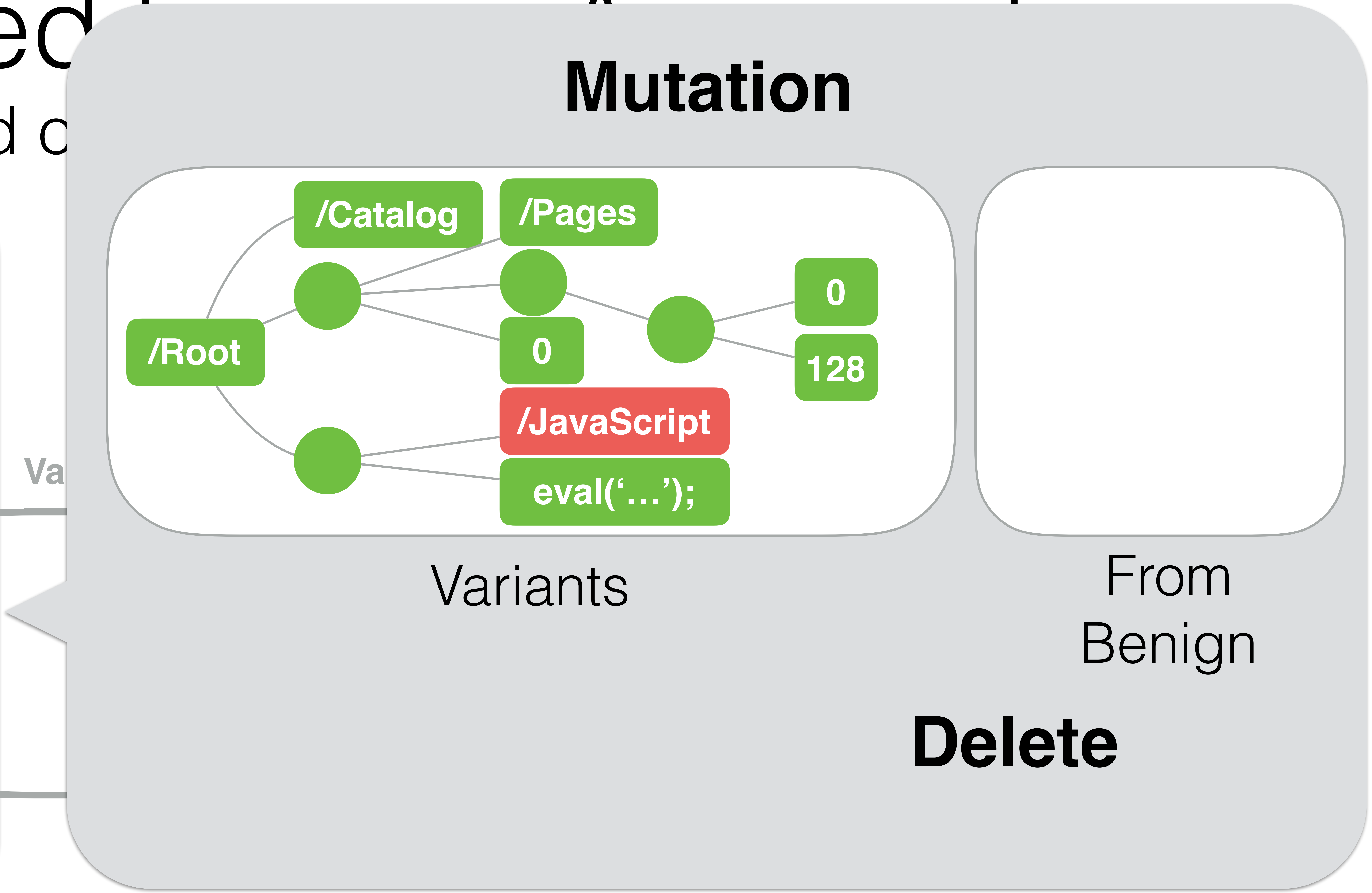
Based o



Automated

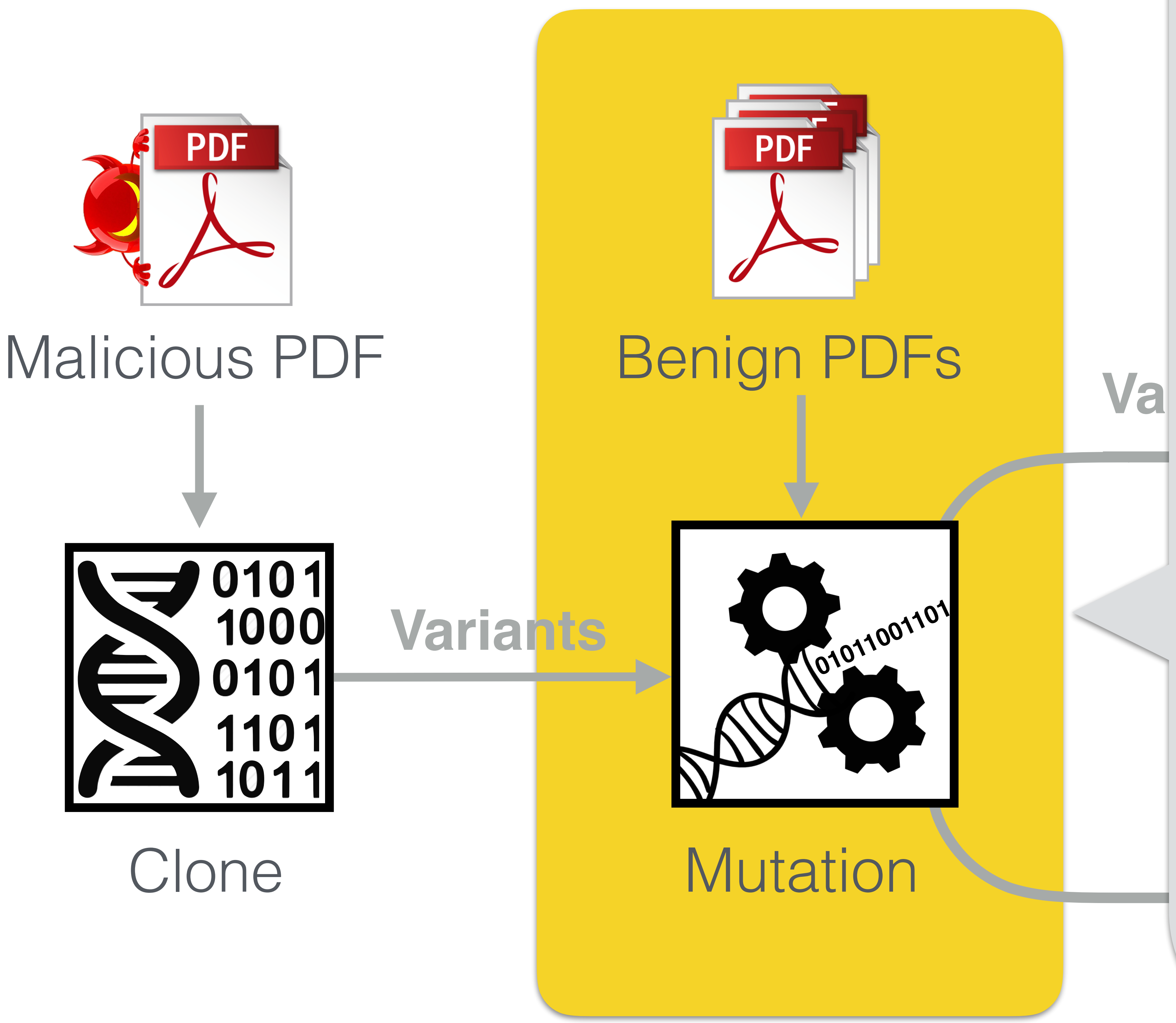
Based c



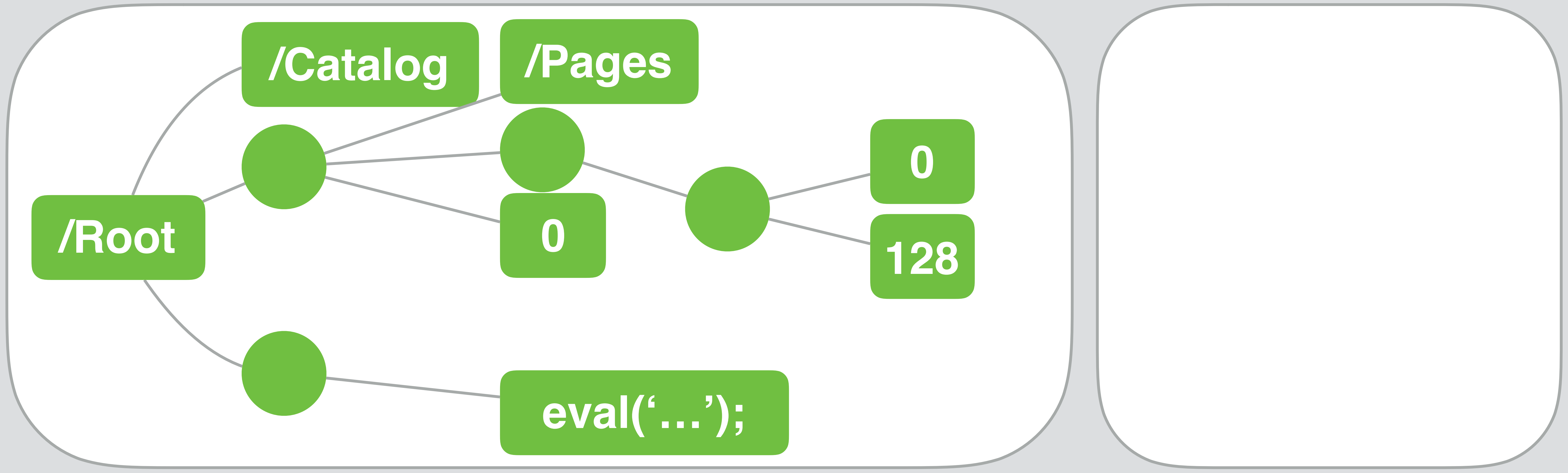


Automated

Based o



Mutation



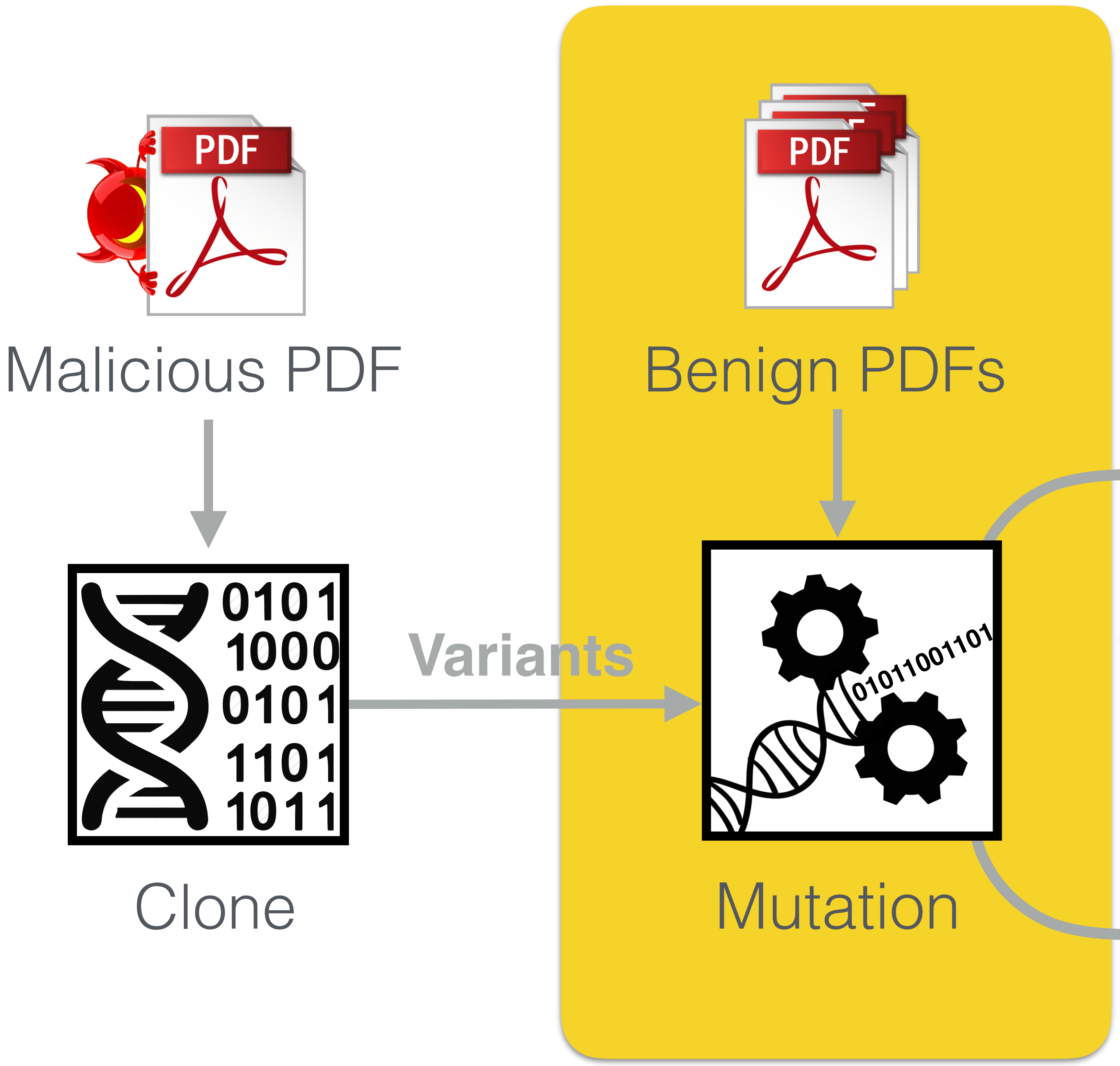
Variants

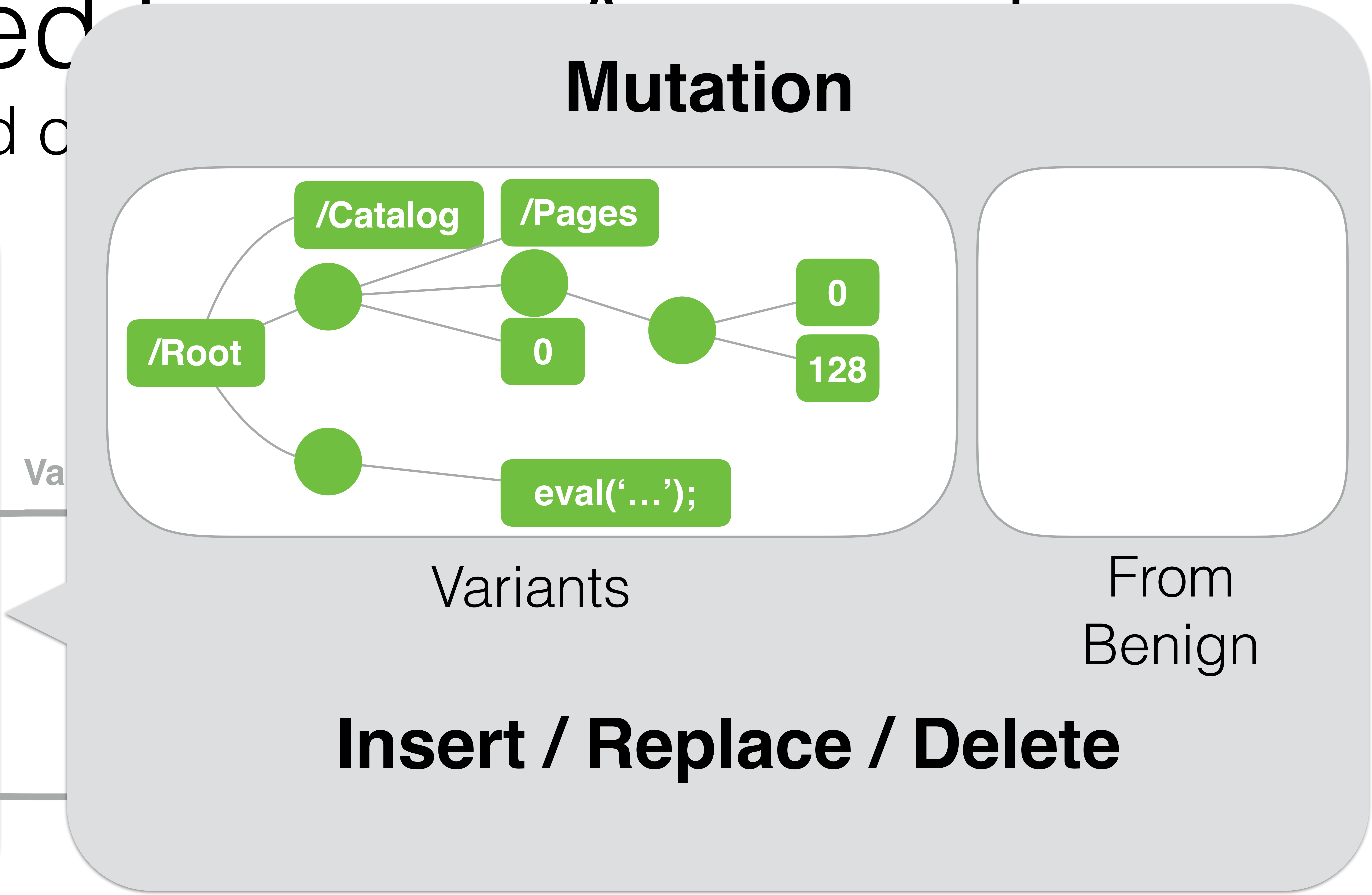
From Benign

Delete

Automated

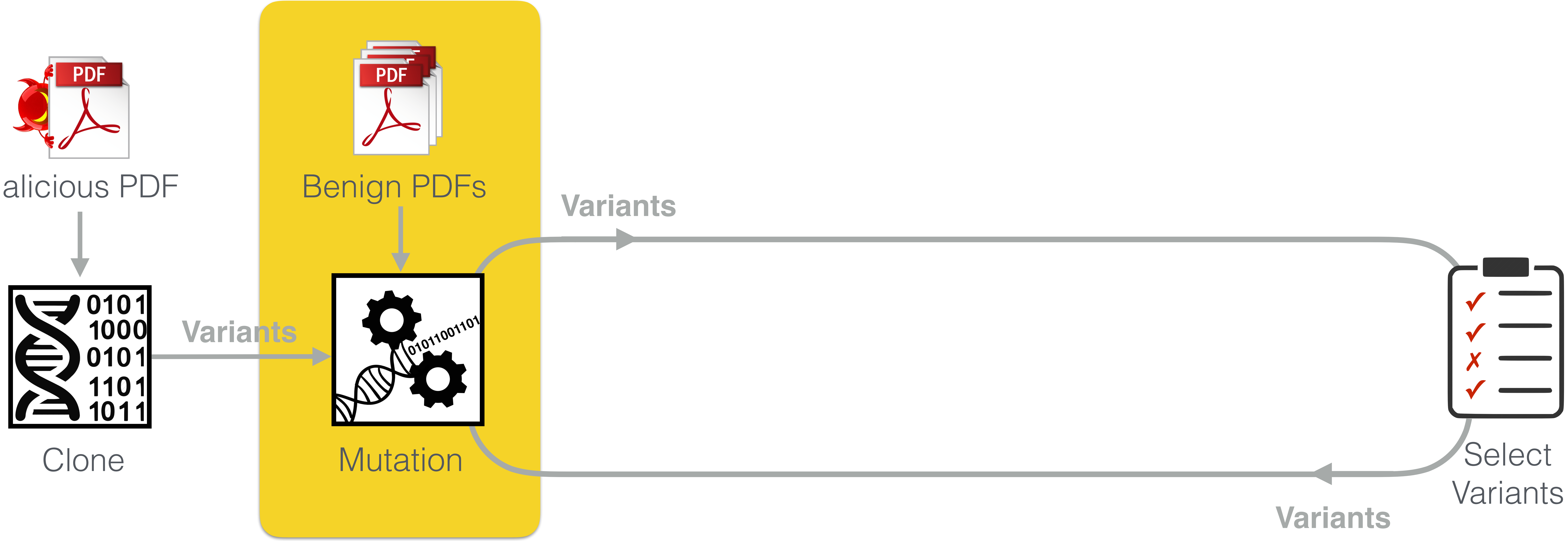
Based o





Automated Evasion Approach

Based on Genetic Programming



Malicious PDF

Benign PDFs

Clone

Variants

Mutation

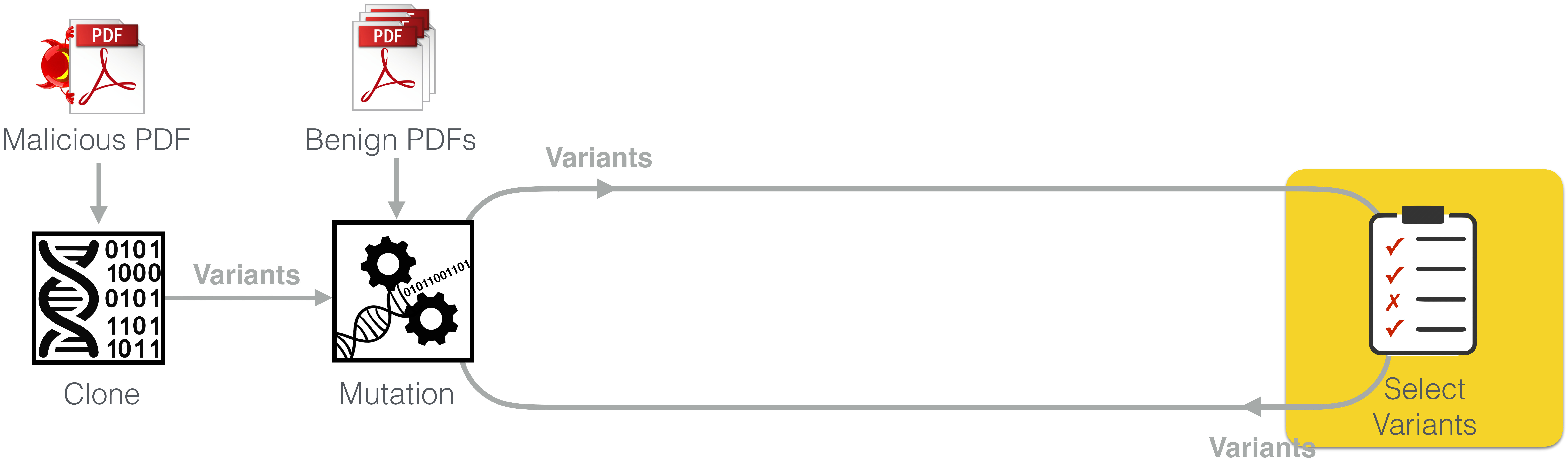
Variants

Select Variants

Variants

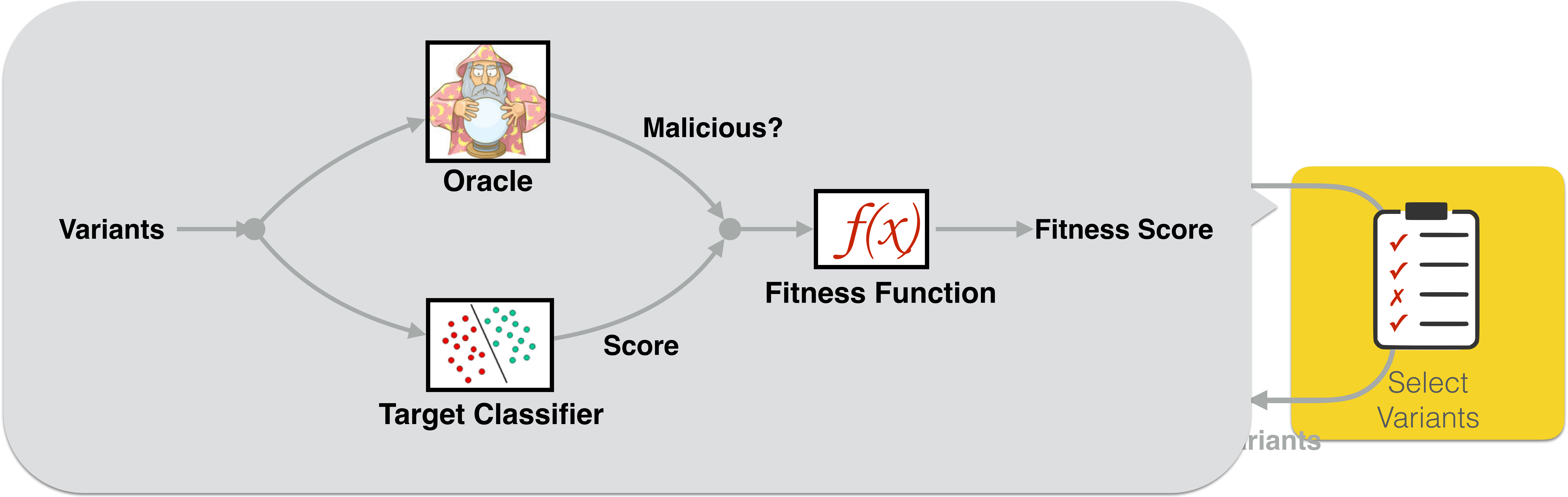
Automated Evasion Approach

Based on Genetic Programming



Automated Evasion Approach

Based on Genetic Programming



Variants

Oracle

Malicious?

Target Classifier

Score

Fitness Function

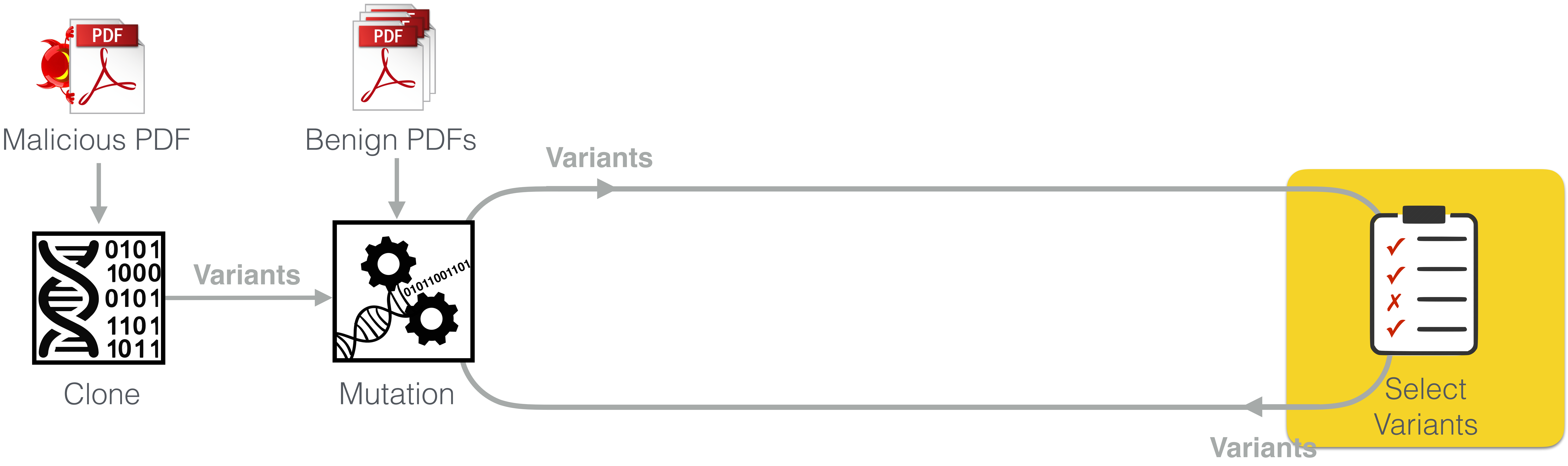
Fitness Score

Select Variants

riants

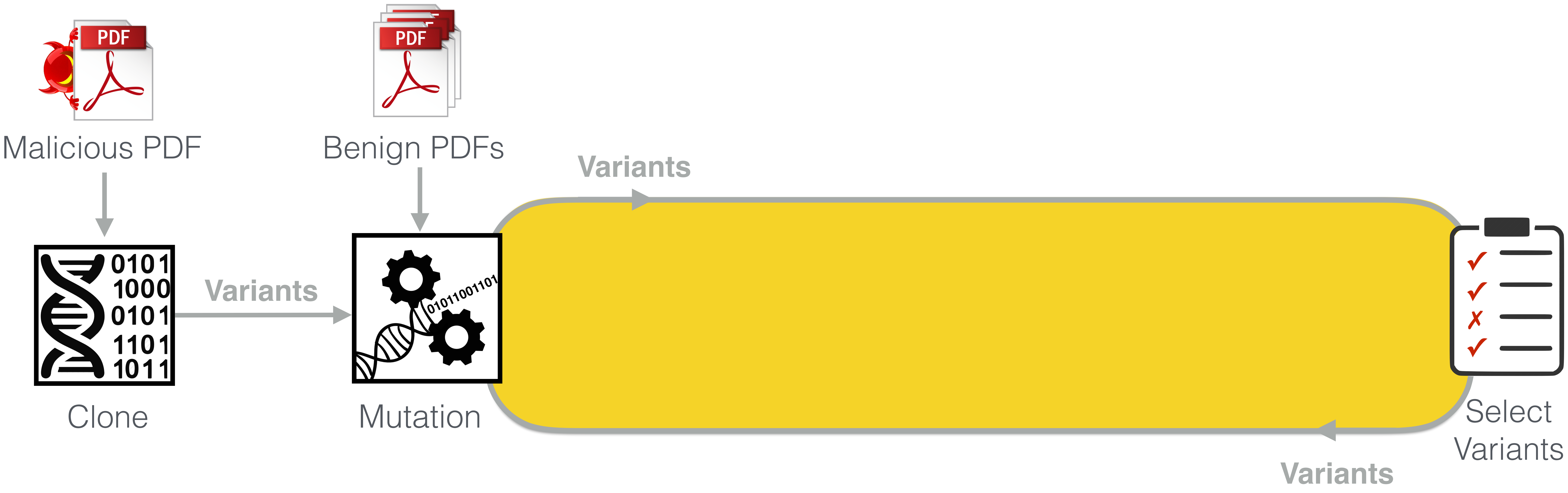
Automated Evasion Approach

Based on Genetic Programming

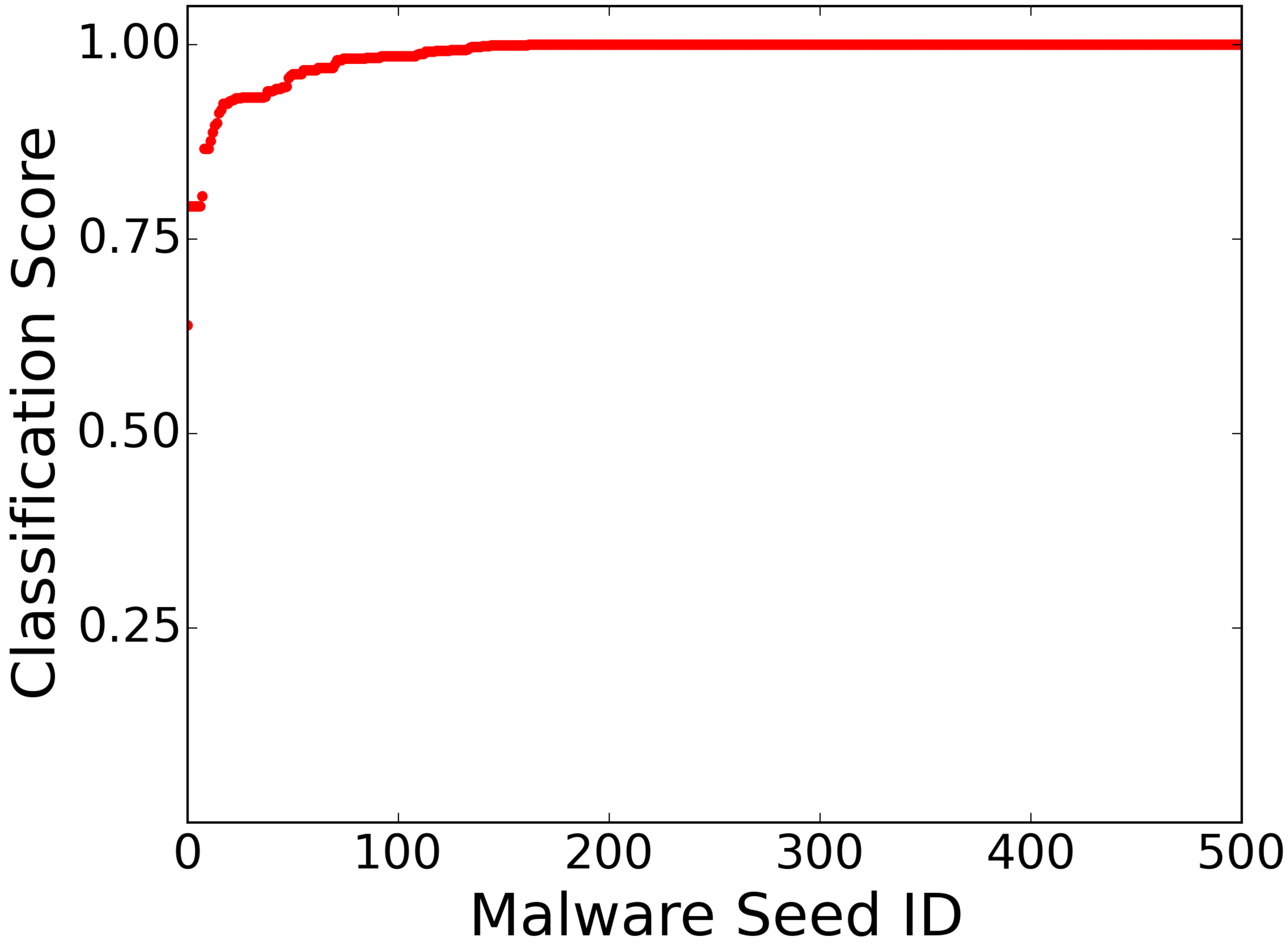


Automated Evasion Approach

Based on Genetic Programming

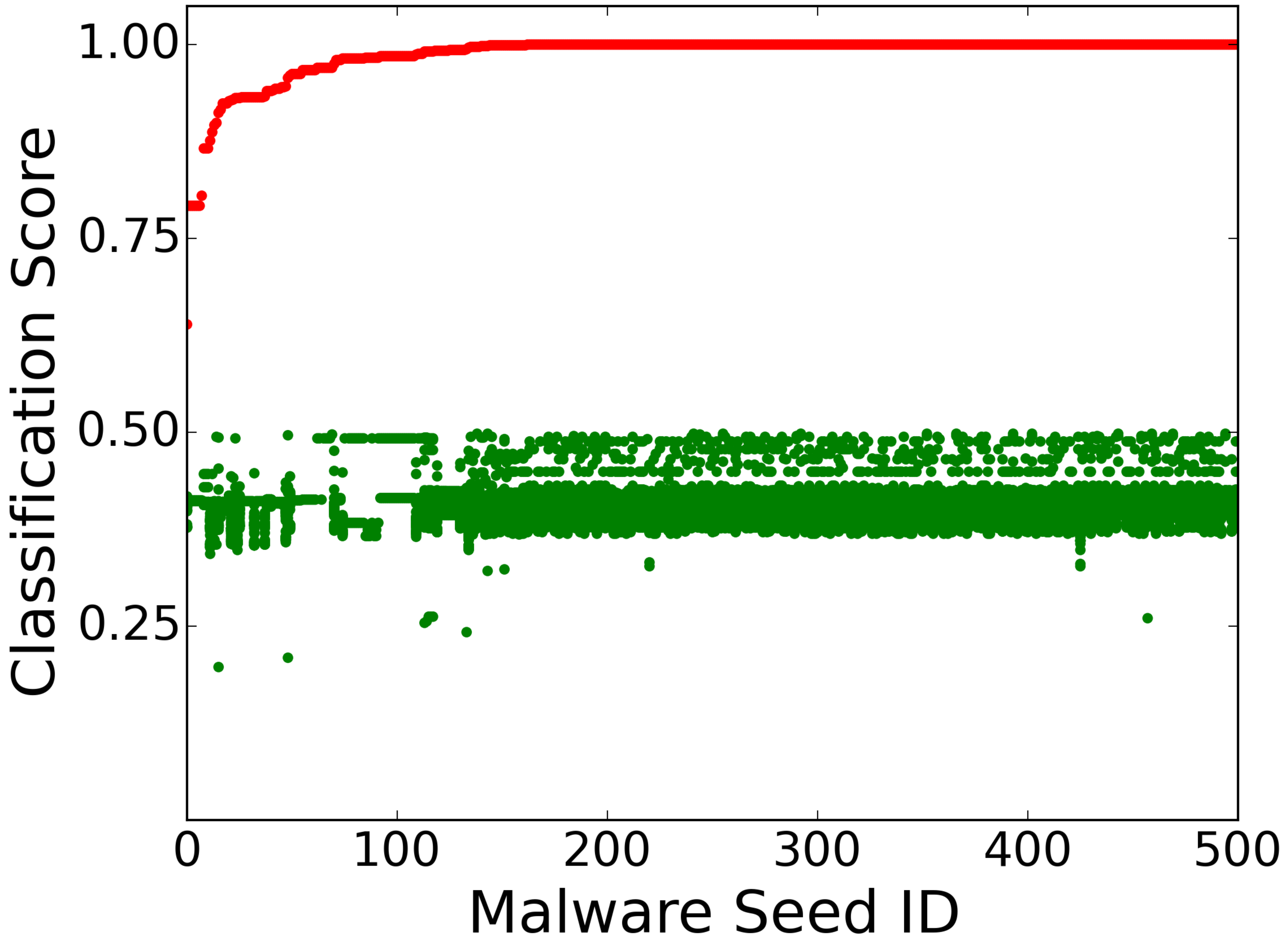


Results: Evaded PDFrate 100%



Original Malware Seeds

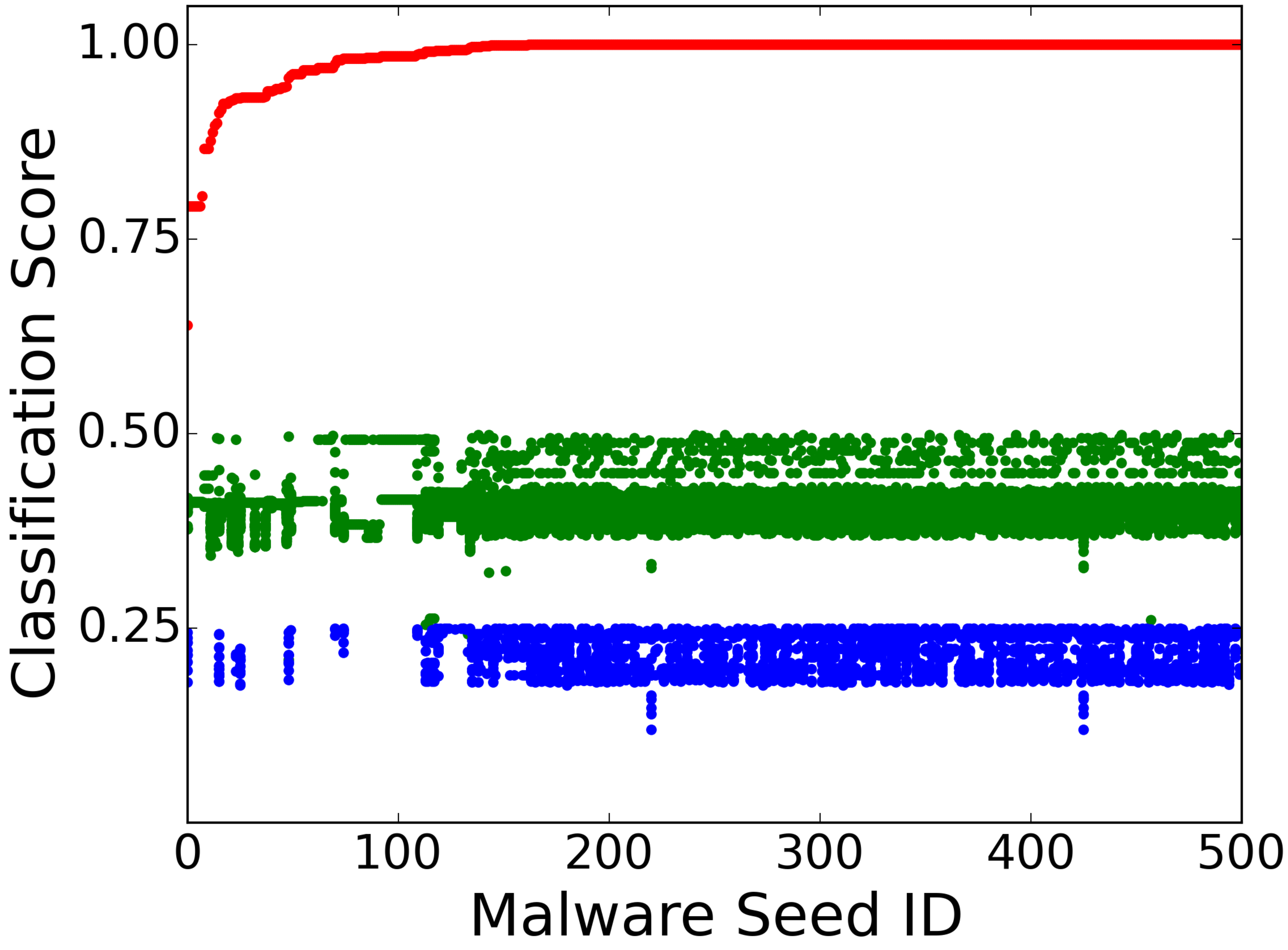
Evaded PDFrate with Adjusted Threshold



Original Malware Seeds

Evasive Variants

Evaded PDFrate with Adjusted Threshold

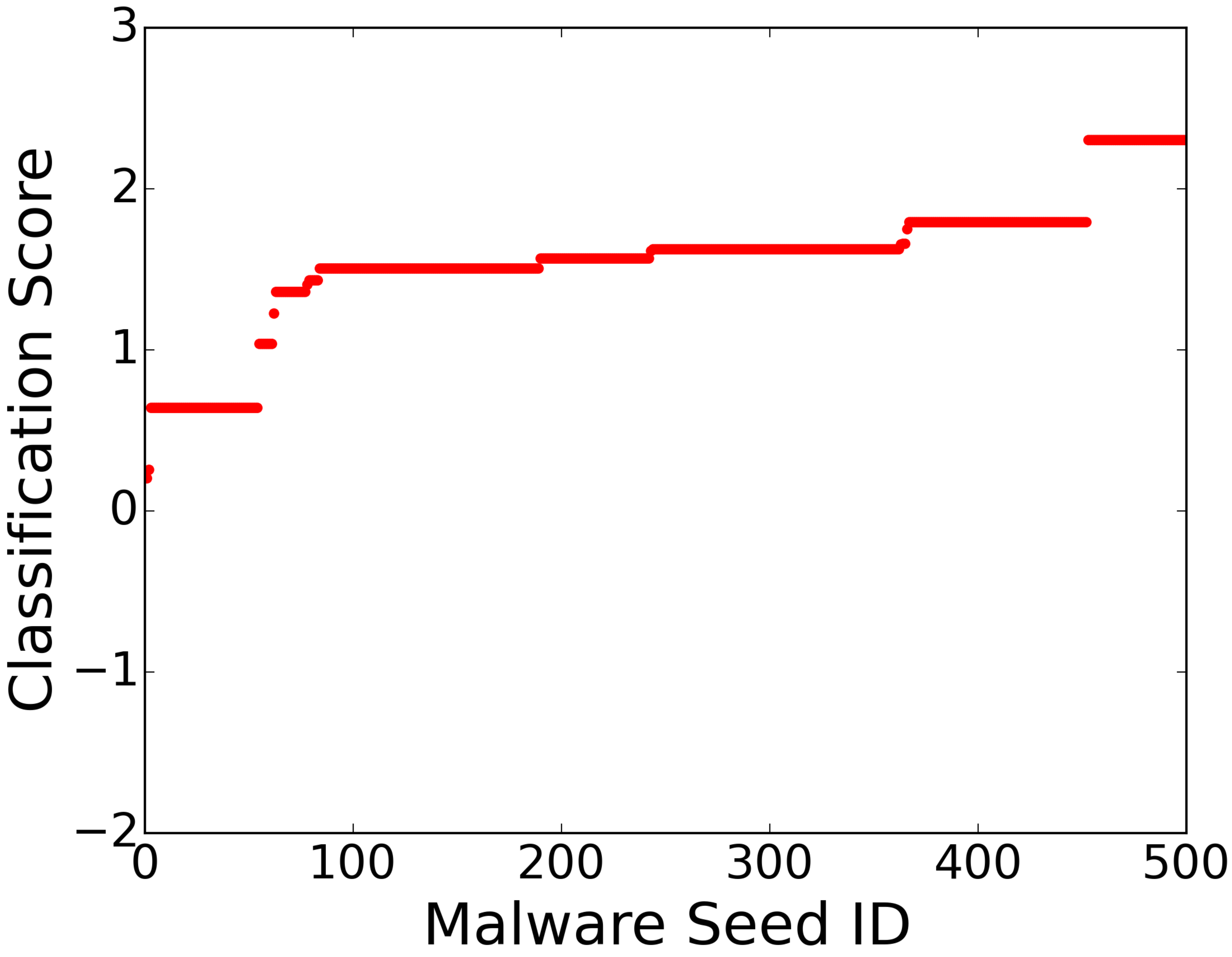


Original Malware Seeds

Evasive Variants

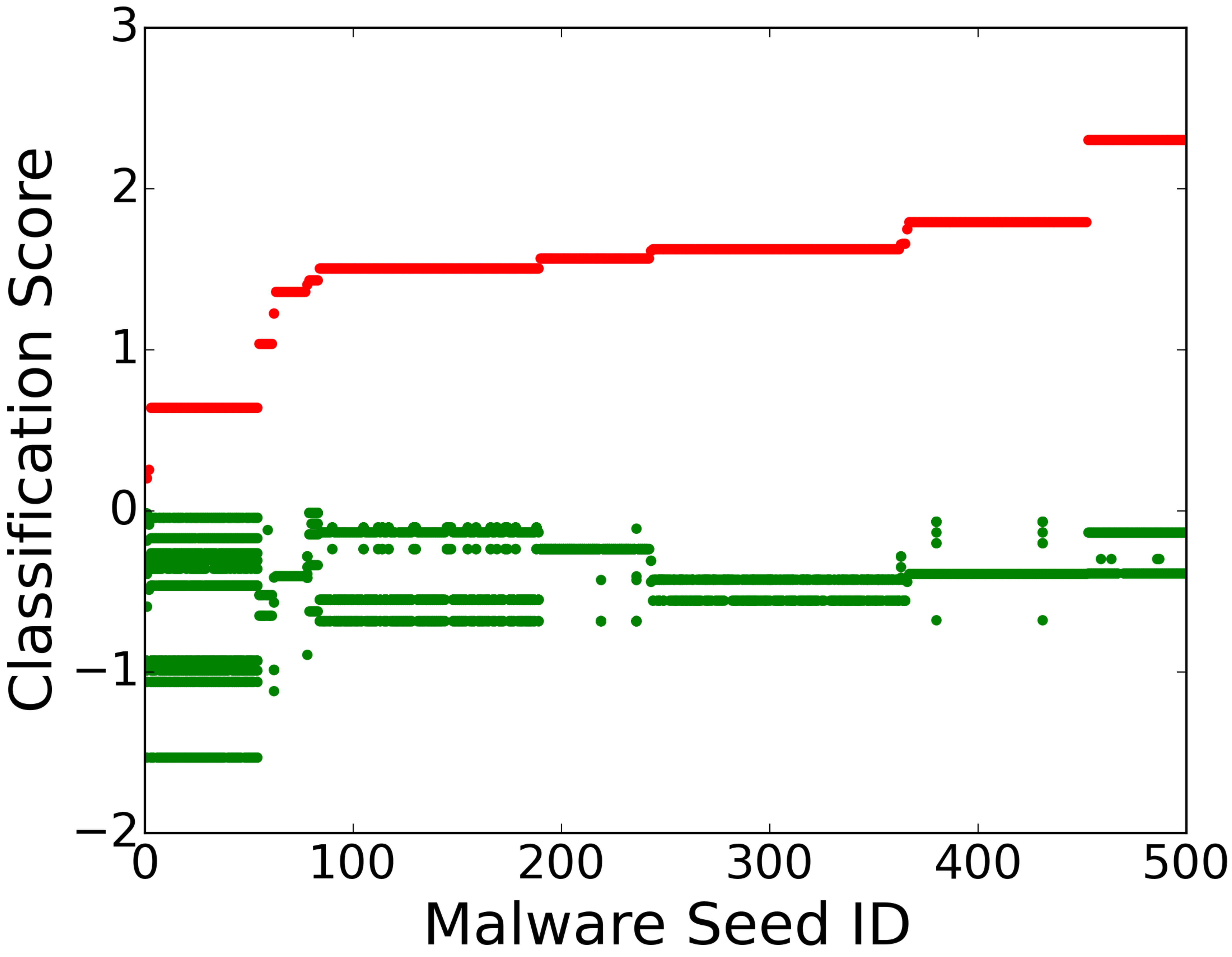
Evasive Variants with lower threshold

Results: Evaded Hidost 100%



Original Malware Seeds

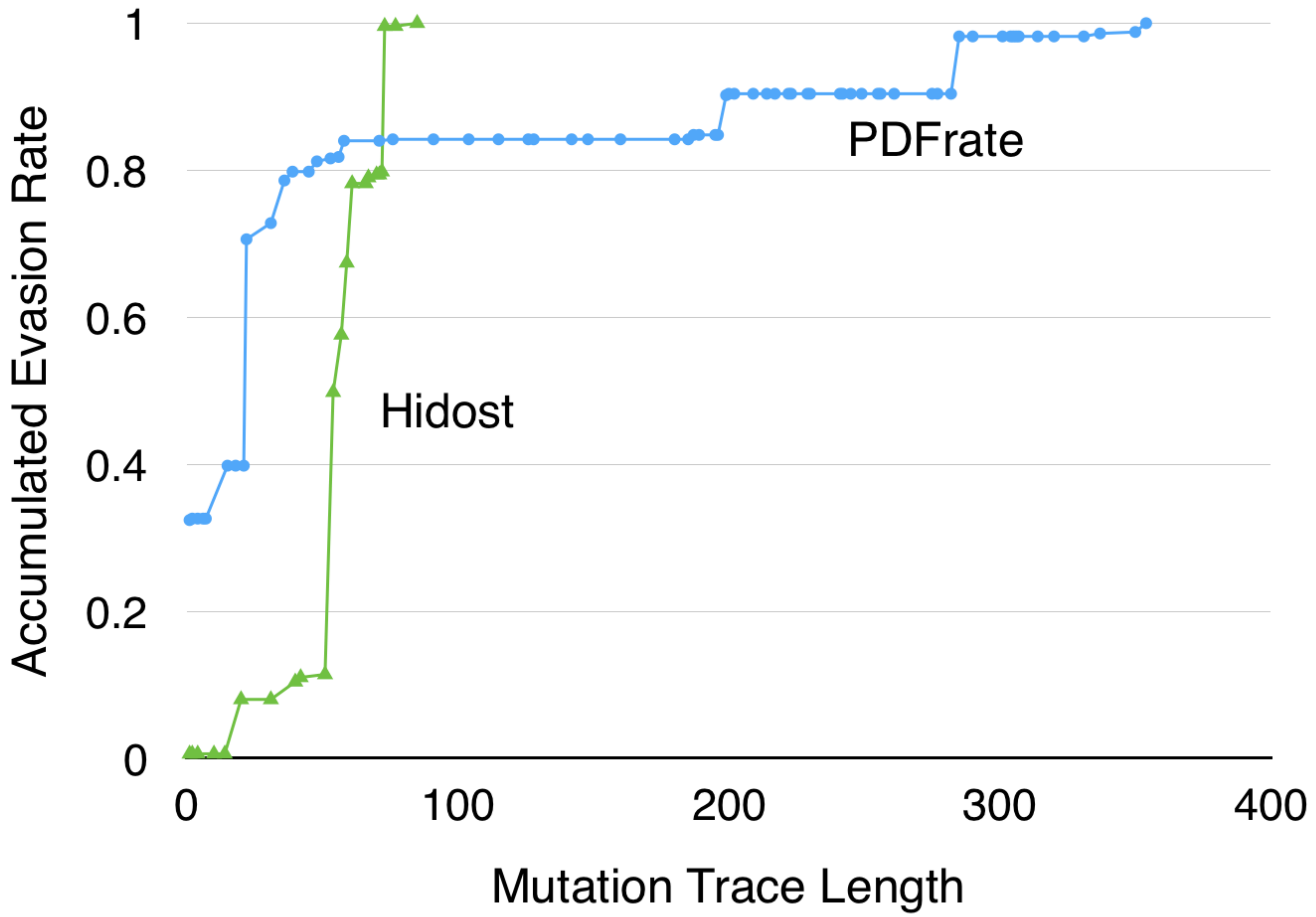
Results: Evaded Hidost 100%



Original Malware Seeds

Evasive Variants

Results: Accumulated Evasion Rate



Difficulties varied on targets.

Evaded PDFrate in 6 days.

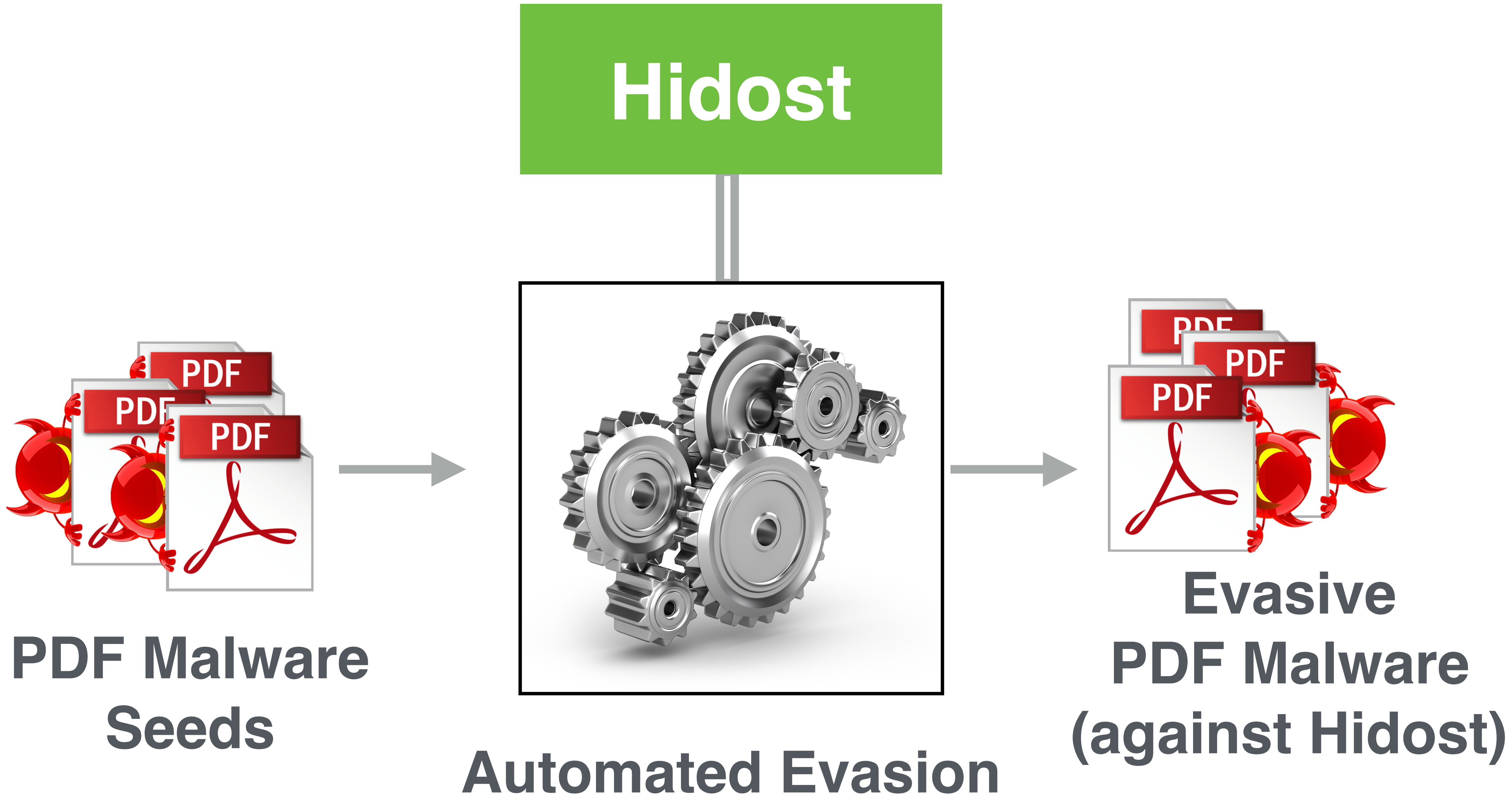
Evaded Hidost in 2 days.

Difficulties varied on seeds.

Simple mutations worked.

Complex mutations required.

Cross-Evasion Effects



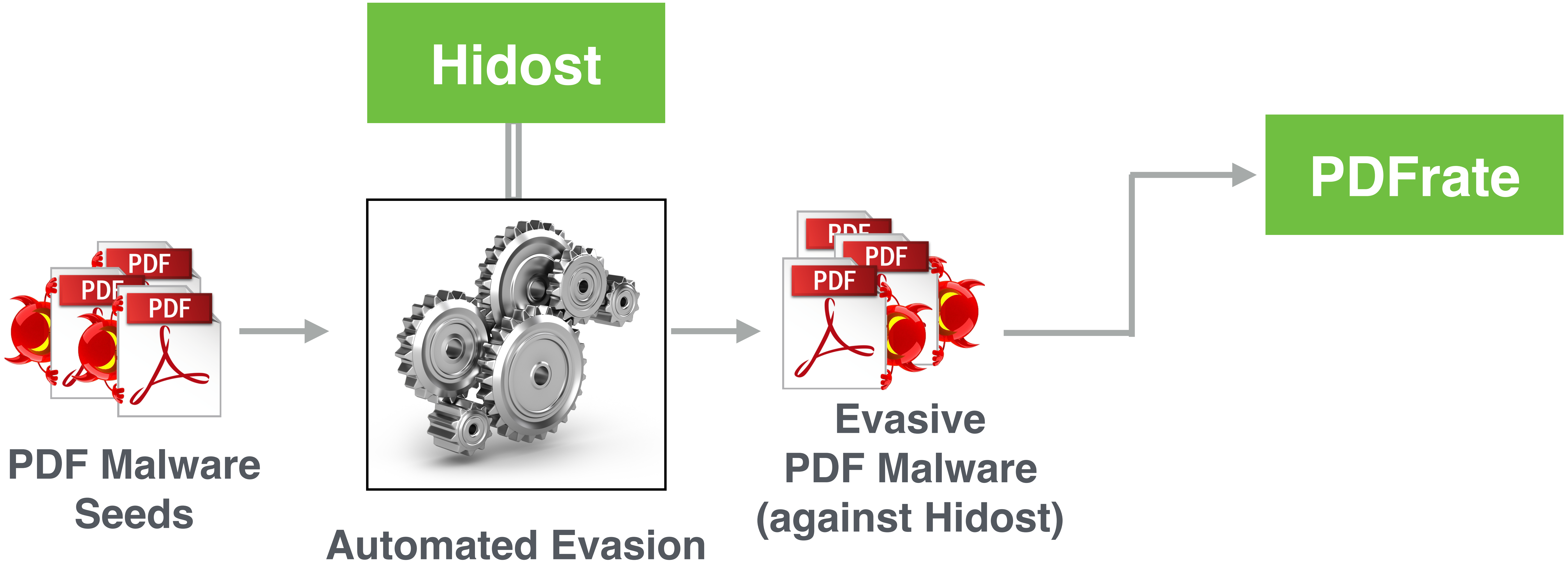
Hidost

PDF Malware Seeds

Automated Evasion

Evasive PDF Malware (against Hidost)

Cross-Evasion Effects



Hidost

PDF Malware Seeds

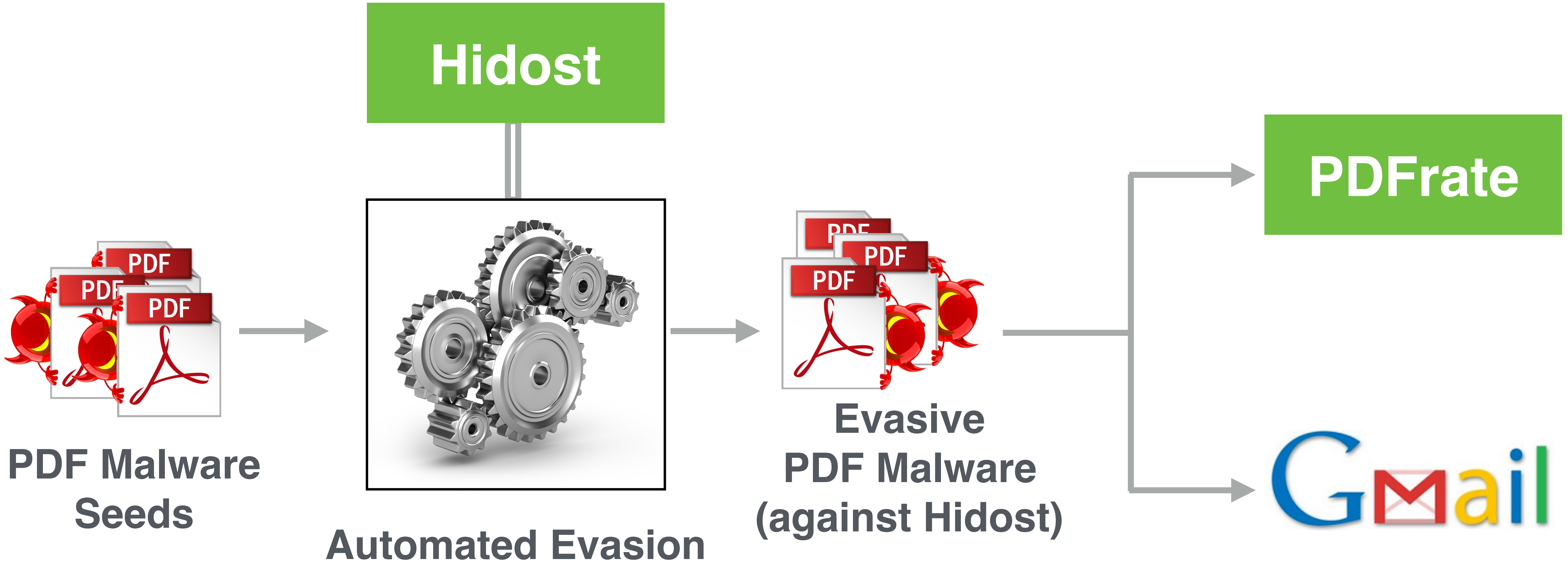
Automated Evasion

Evasive PDF Malware (against Hidost)

PDFrate

387/500 Evasive
(77.4%)

Cross-Evasion Effects



Hidost

PDF Malware Seeds

Automated Evasion

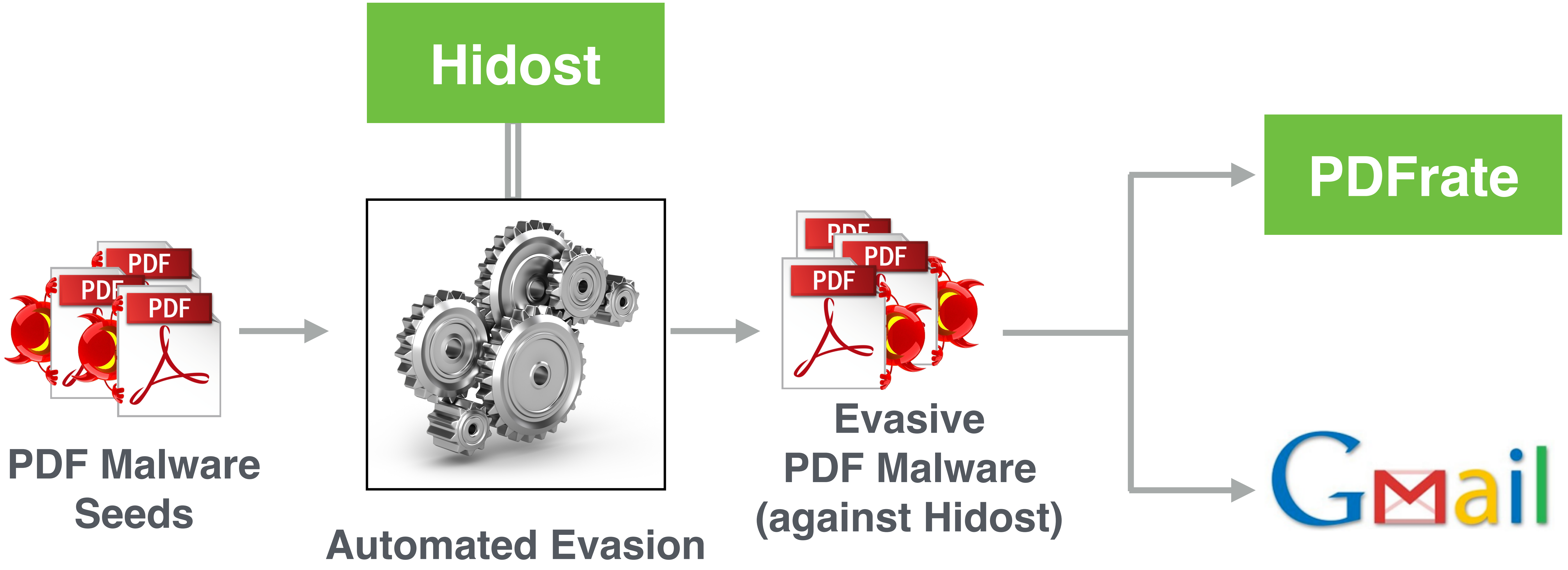
Evasive PDF Malware (against Hidost)

PDFrate

Gmail

387/500 Evasive (77.4%)

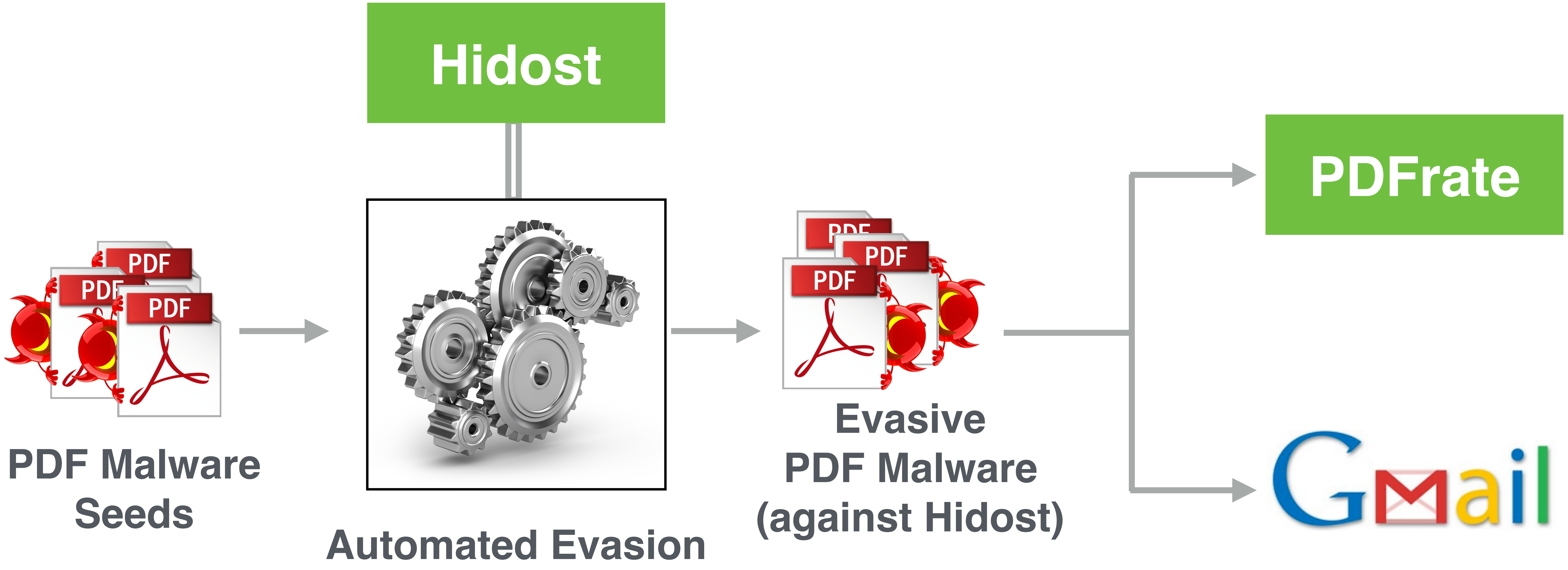
Cross-Evasion Effects



387/500 Evasive
(77.4%)

3/500 Evasive
(0.6%)

Cross-Evasion Effects



Hidost

PDF Malware Seeds

Automated Evasion

Evasive PDF Malware (against Hidost)

PDFrate

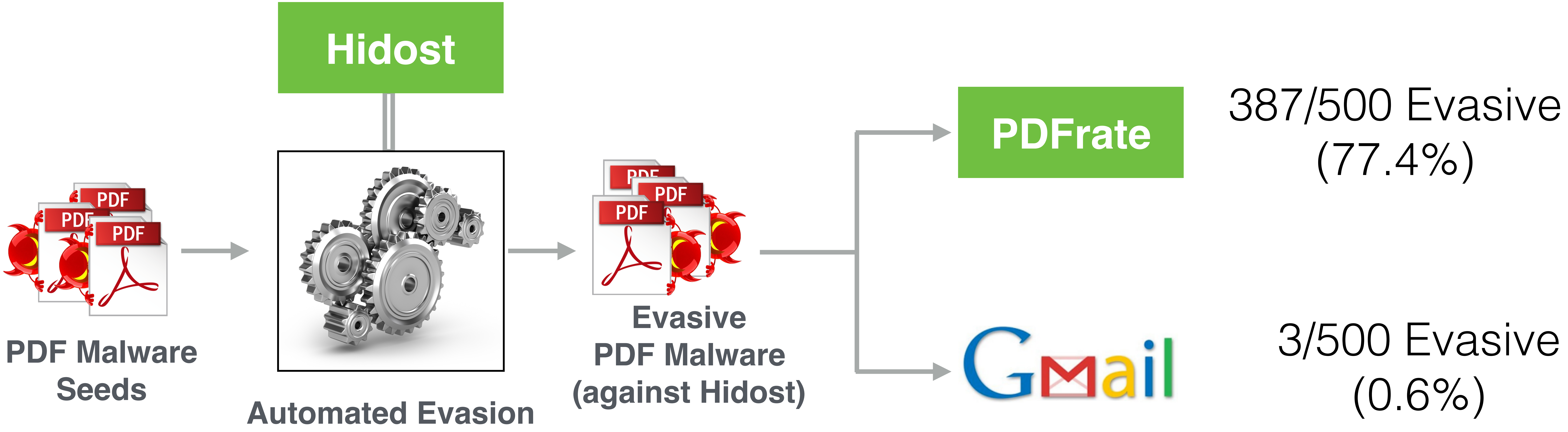
387/500 Evasive (77.4%)

Gmail

3/500 Evasive (0.6%)

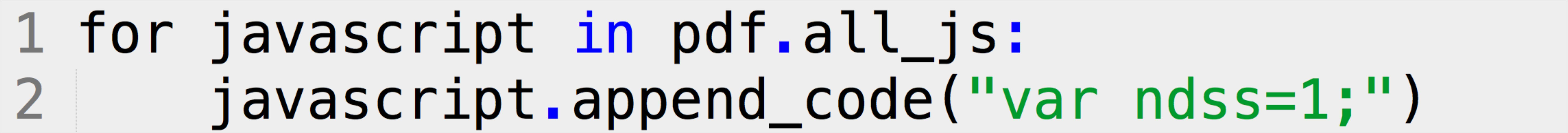
Gmail's classifier is secure?

Cross-Evasion Effects



Gmail's classifier is [redacted] different.

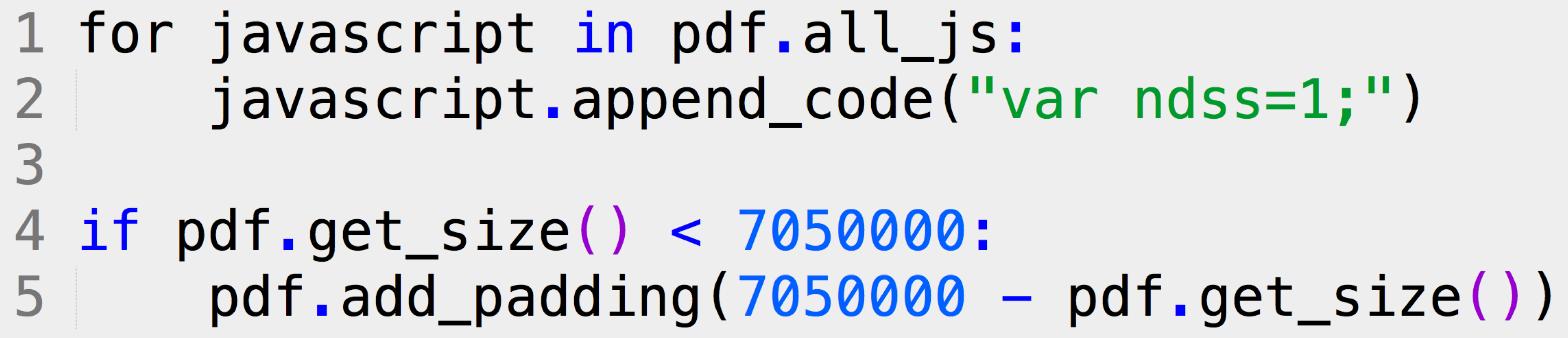
Evading Gmail’s Classifier



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

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

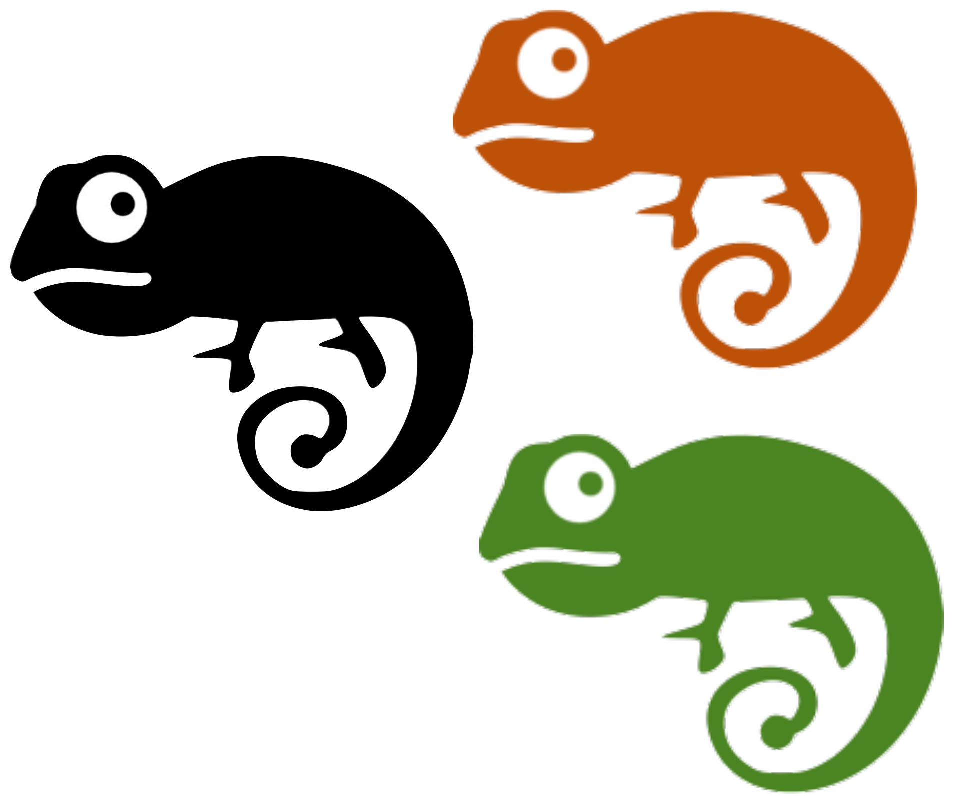
Evading Gmail's Classifier



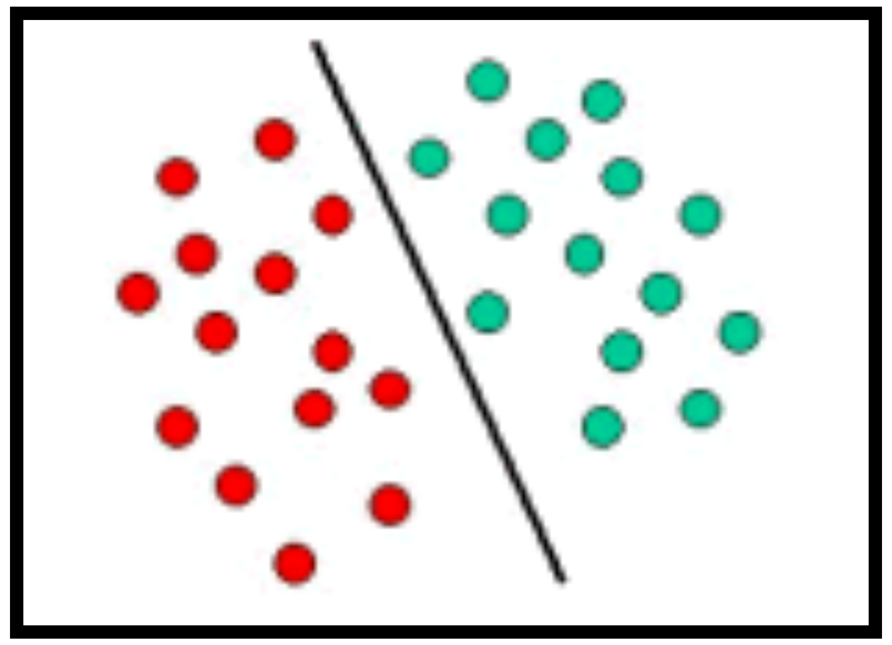
```
for javascript in pdf.all_js:
    javascript.append_code("var ndss=1;")

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

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

Conclusion



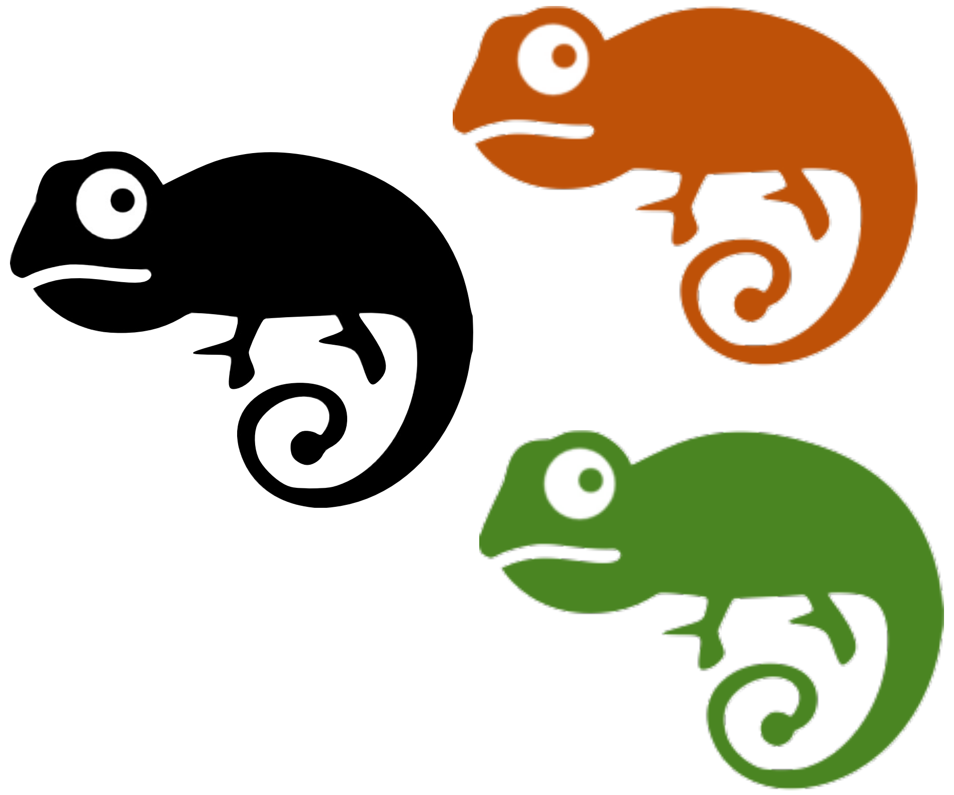
Vs.



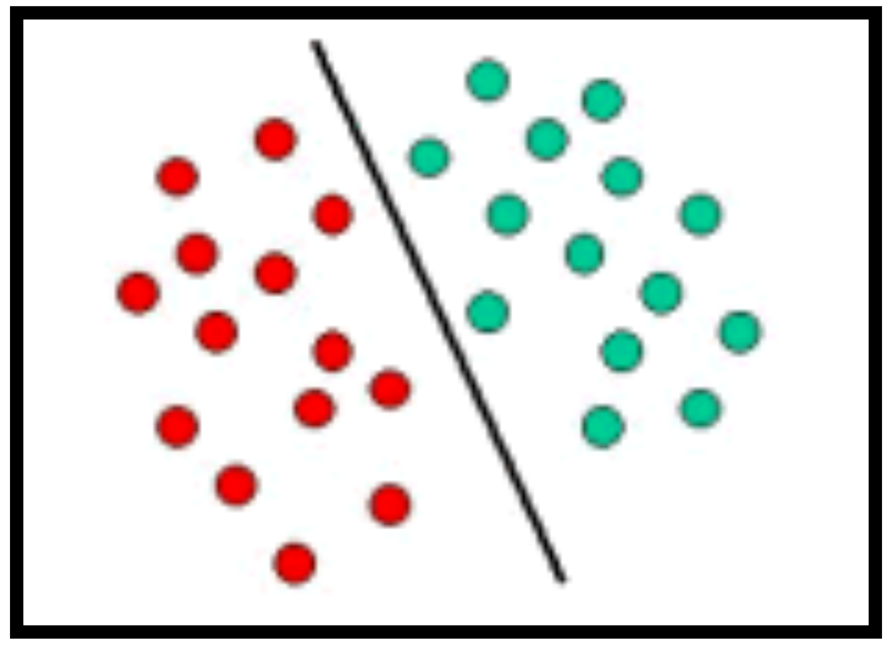
Who will win this arm race?

Source Code: http://www.EvadeML.org

Conclusion



Vs.



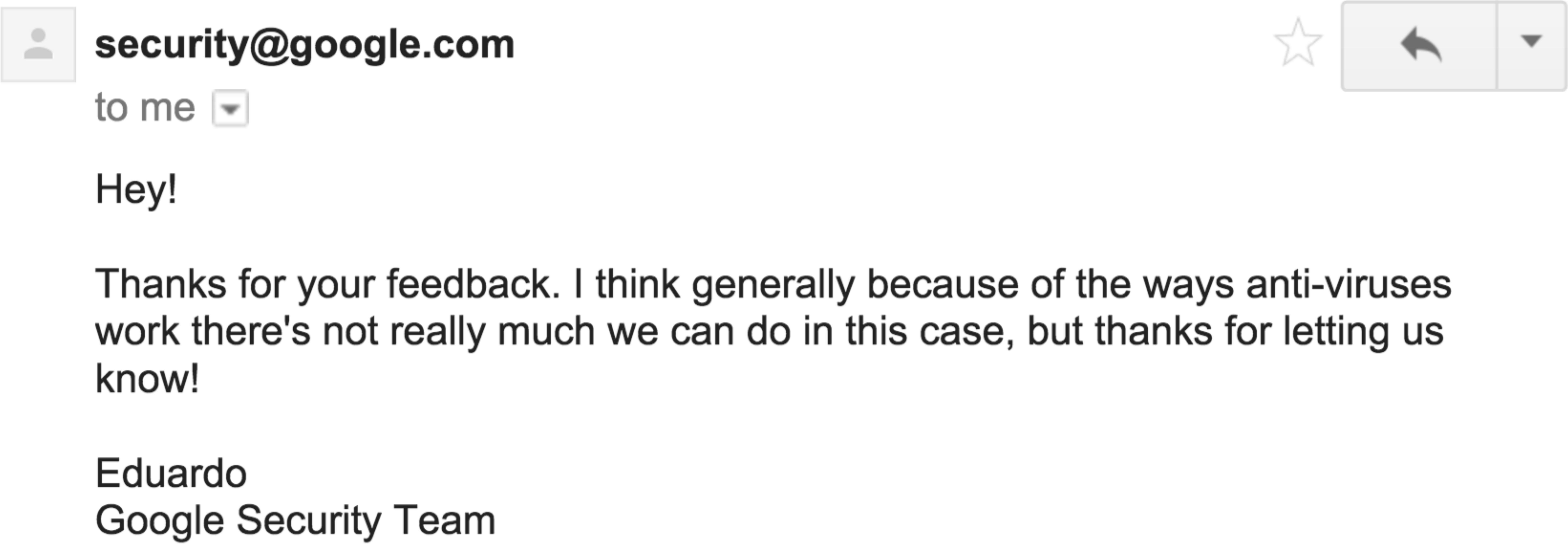
Who will win this arm race?

Source Code: http://www.EvadeML.org

Ad: Weilin is seeking summer internship opportunities.

Backups

They Don’t Care



security@google.com

to me

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