Shrinking and Exploring Adversarial Search Spaces

ARO Workshop on Adversarial Learning
Stanford, 14 Sept 2017

evadeML.org

David Evans
University of Virginia

Weilin Xu

Yanjun Qi

Machine Learning is Eating Computer Science

2017 Google PhD Fellows

Algorithms, Optimizations and Markets
Chiu Wai Sam Wong, *University of California, Berkeley*
Eric Balkanski, *Harvard University*
Haifeng Xu, *University of Southern California*

Human-Computer Interaction
Motahhare Eslami, *University of Illinois, Urbana-Champaign*
Sarah D'Angelo, *Northwestern University*
Sarah Mcroberts, *University of Minnesota - Twin Cities*
Sarah Webber, *The University of Melbourne*

Machine Learning
Aude Genevay, *Fondation Sciences Mathématiques de Paris*
Dustin Tran, *Columbia University*
Jamie Hayes, *University College London*
Jin-Hwa Kim, *Seoul National University*
Ling Luo, *The University of Sydney*
Martin Arjovsky, *New York University*
Sayak Ray Chowdhury, *Indian Institute of Science*
Song Zuo, *Tsinghua University*
Taco Cohen, *University of Amsterdam*
Yuhuai Wu, *University of Toronto*
Yunhe Wang, *Peking University*
Yunye Gong, *Cornell University*

Machine Perception, Speech Technology and Computer Vision
Avijit Dasgupta, *International Institute of Information Technology - Hyderabad*
Franziska Müller, *Saarland University - Saarbrücken GSCS and Max Planck Institute for Informatics*
George Trigeorgis, *Imperial College London*
Iro Armeni, *Stanford University*
Saining Xie, *University of California, San Diego*
Yu-Chuan Su, *University of Texas, Austin*

Mobile Computing
Sangeun Oh, *Korea Advanced Institute of Science and Technology*
Shuo Yang, *Shanghai Jiao Tong University*

Natural Language Processing
Bidisha Samanta, *Indian Institute of Technology Kharagpur*
Ekaterina Vylomova, *The University of Melbourne*
Jianpeng Cheng, *The University of Edinburgh*
Kevin Clark, *Stanford University*
Meng Zhang, *Tsinghua University*
Preksha Nama, *Indian Institute of Technology Madras*
Tim Rocktaschel, *University College London*

Privacy and Security
Romain Gay, *ENS - École Normale Supérieure*
Xi He, *Duke University*
Yupeng Zhang, *University of Maryland, College Park*

Programming Languages, Algorithms and Software Engineering
Christoffer Quist Adamsen, *Aarhus University*
Muhammad Ali Gulzar, *University of California, Los Angeles*
Oded Padon, *Tel-Aviv University*

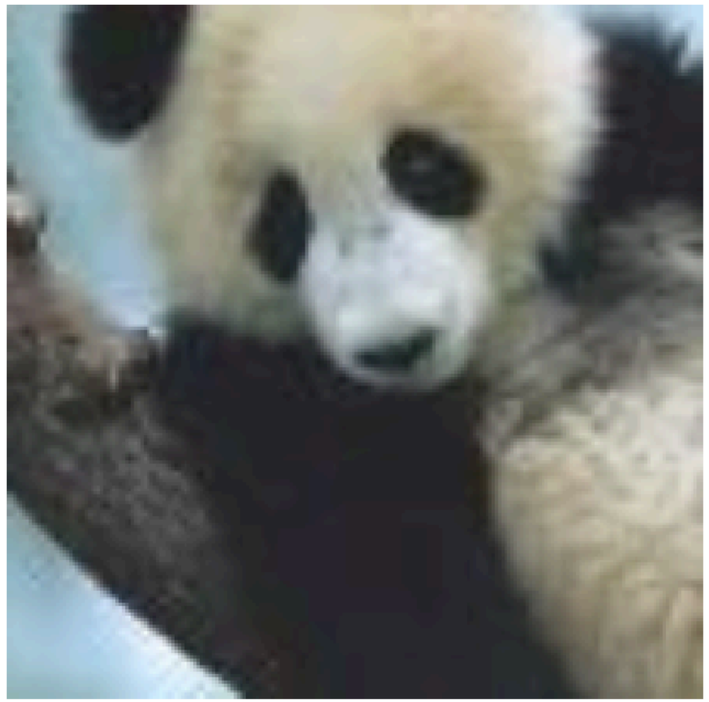
Structured Data and Database Management
Amir Shaikhha, *EPFL CS*
Jingbo Shang, *University of Illinois, Urbana-Champaign*

Systems and Networking
Ahmed M. Said Mohamed Tawfik Issa, *Georgia Institute of Technology*
Khanh Nguyen, *University of California, Irvine*
Radhika Mittal, *University of California, Berkeley*
Ryan Beckett, *Princeton University*
Samaneh Movassaghi, *Australian National University*

Security State-of-the-Art

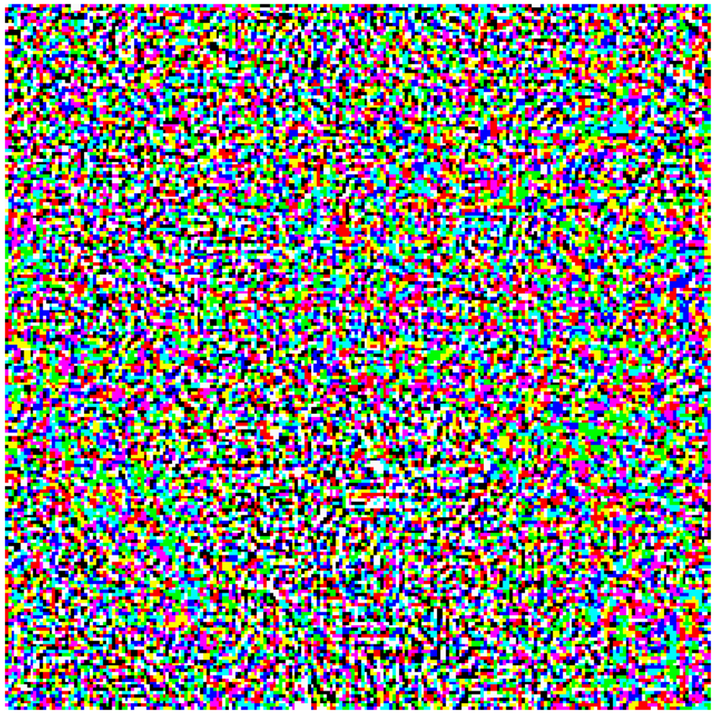
	Random guessing attack success probability	Threat models	Proofs
Cryptography	2^{-128}	information theoretic, resource bounded	required
System Security	2^{-32}	capabilities, motivations, rationality	common
Adversarial Machine Learning	2^{-11} *; 2^{-6}	white-box, black-box	rare!

Adversarial Examples



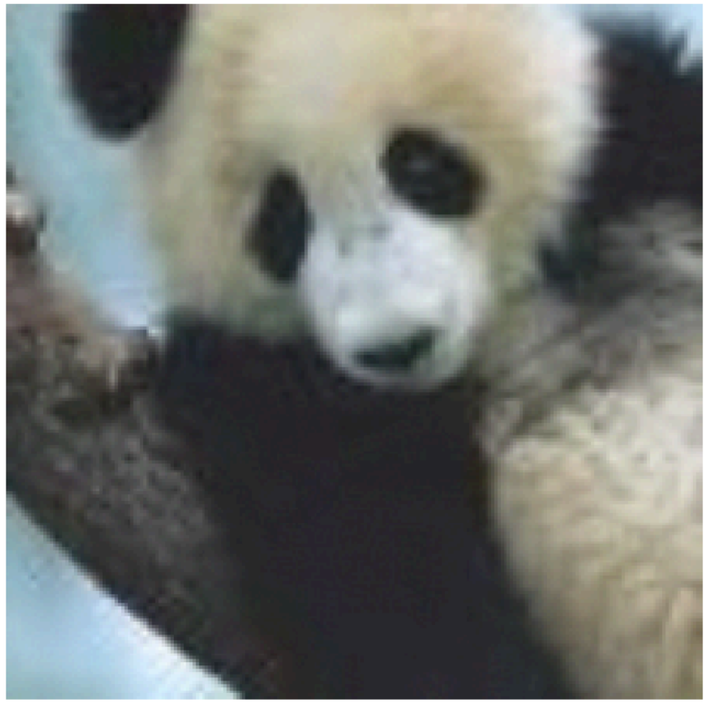
“panda”

+



$0.007 \times [noise]$

=



“gibbon”

Example from: Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy. *Explaining and Harnessing Adversarial Examples*. 2014.

Adversarial Examples Game

Given seed sample, x, find x' where:

$f(x') \neq f(x)$ Class is different (untargeted)

$f(x') = t$ Class is t (targeted)

$\Delta(x, x') \leq \delta$ Difference below threshold

$\Delta(x, x')$ is defined in some (simple!) metric space:

L_0 "norm (# different), L_1 norm, L_2 norm ("Euclidean"), L_∞ norm:

Detecting Adversarial Examples

Input

Model

$Prediction_0$

Squeezer$_1$

Model

$Prediction_1$

Squeezer$_2$

Model

$Prediction_2$

Squeezer$_k$

Model'

$Prediction_k$

$f(pred_0, pred_1, \ldots, pred_k)$

Adversarial

Yes

No

Legitimate

“Feature Squeezing”

$\boldsymbol{x}$ [0.054, 0.4894, 0.9258, 0.0116, 0.2898, 0.5222, 0.5074, …]

Squeeze: $f_i = \text{round}(f_i \times 4)/4$

[0.0, 0.5, 1.0, 0.0, 0.25, 0.5, 0.5, …]

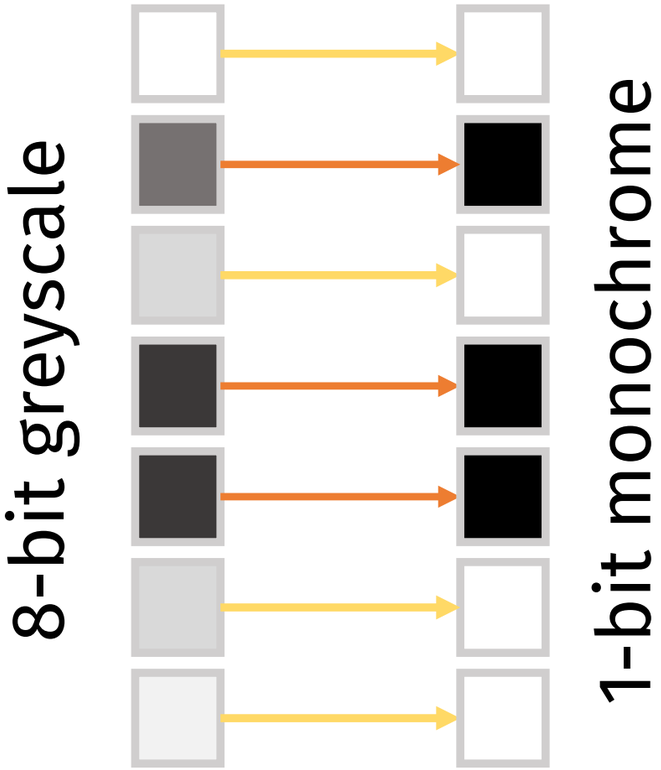
$$\text{squeeze}(\boldsymbol{x}') \approx \text{squeeze}(\boldsymbol{x}) \implies f(\text{squeeze}(\boldsymbol{x}')) \approx f(\text{squeeze}(\boldsymbol{x}))$$

[0.0, 0.5, 1.0, 0.0, 0.25, 0.5, 0.5, …]

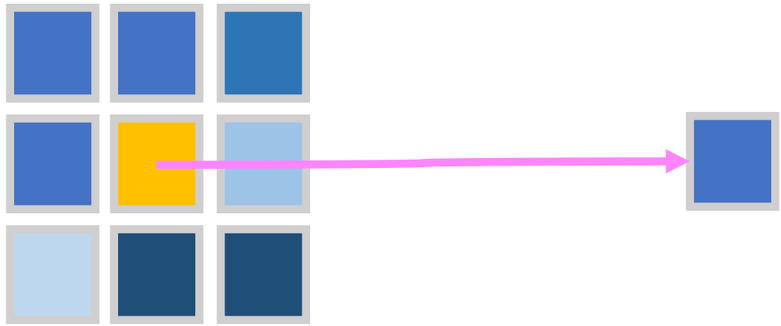
Squeeze: $f_i = \text{round}(f_i \times 4)/4$

$\boldsymbol{x}'$ [0.0491, 0.4903, 0.9292, 0.009, 0.2942, 0.5243, 0.5078, …]

Example Squeezers



Reduce Color Depth



3x3 smoothing:
Replace with median of pixels and its neighbors

Median Smoothing

Simple Instantiation

Input

Model (7-layer CNN)

$Prediction_0$

Bit Depth-1

Model

$Prediction_1$

Median 2×2

Model

$Prediction_2$

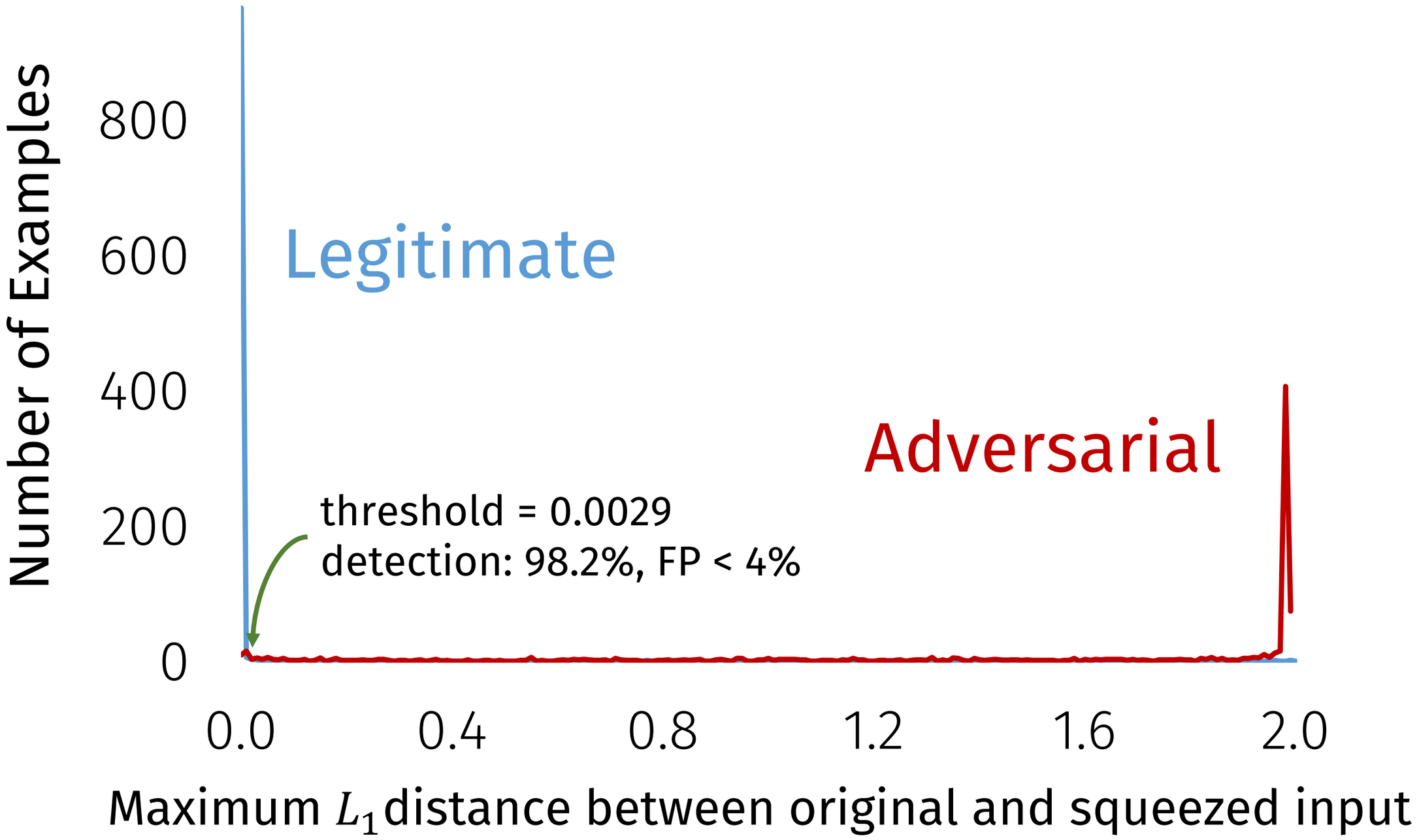
$\max\left(L_1(p_0, p_1), L_1(p_0, p2)\right) > t$

Adversarial

Yes

No

Legitimate


Number of Examples
800
600
400
200
0
Legitimate
Adversarial
threshold = 0.0029
detection: 98.2%, FP < 4%
0.0
0.4
0.8
1.2
1.6
2.0
Maximum L_1 distance between original and squeezed input

ImageNet Configuration

Input

Model (MobileNet)

$Prediction_0$

Bit Depth-5

Model

$Prediction_1$

Median 2×2

Model

$Prediction_2$

Non-local Mean

Model

$Prediction_3$

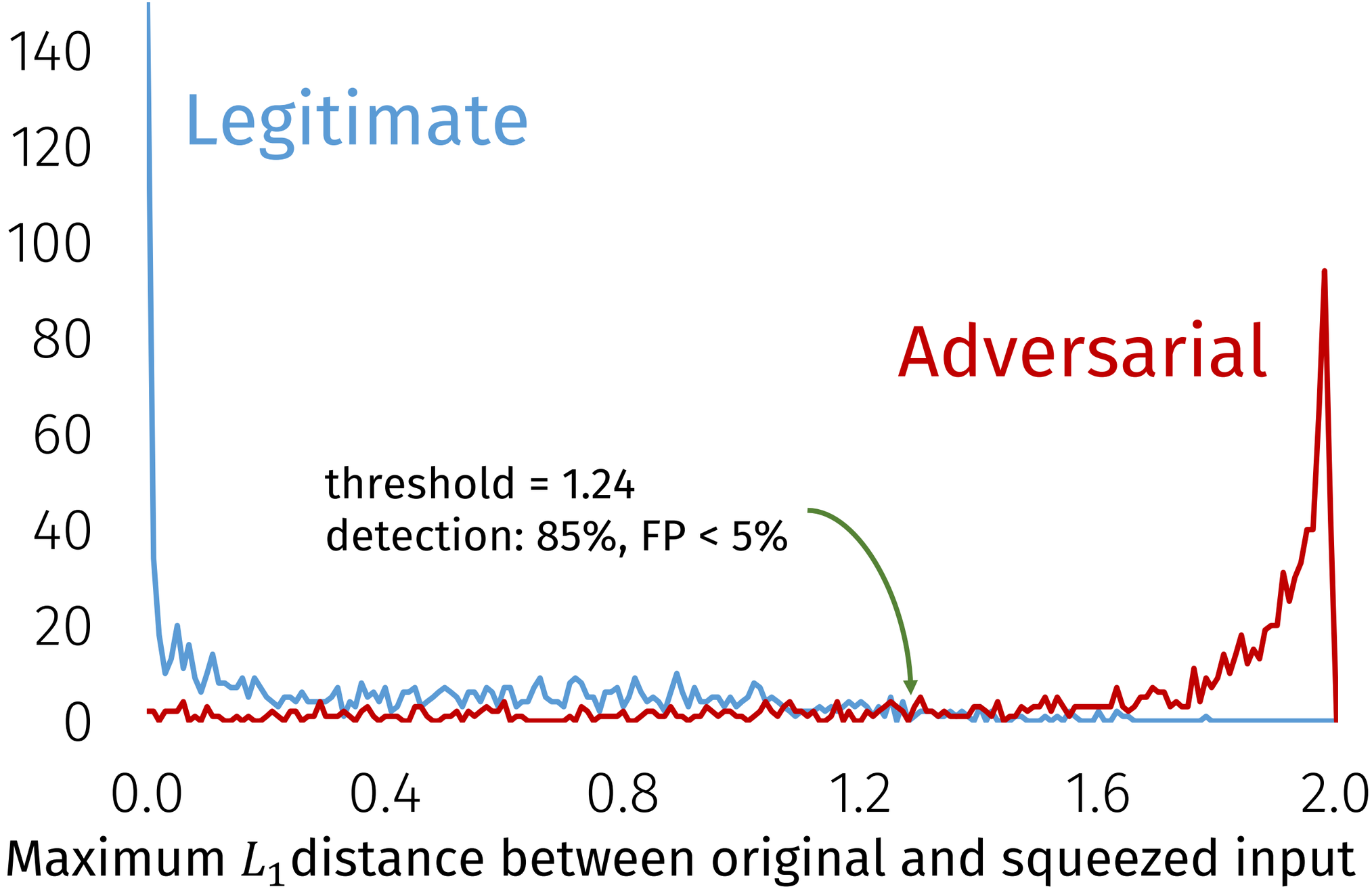
$\max(L_1(p_0, \{p_1, p_2, p_3\}) > t$

Adversarial

Yes

No

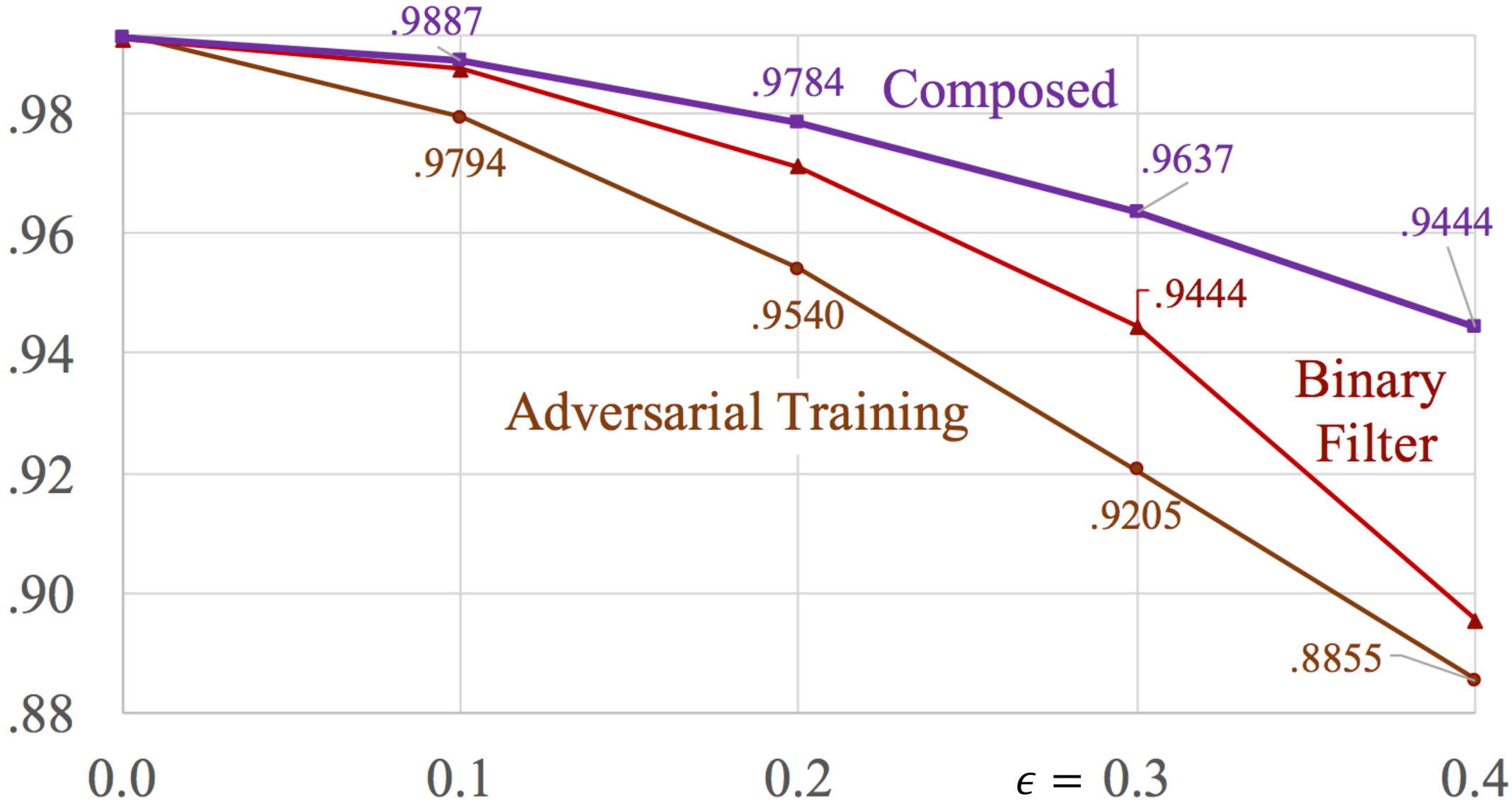
Legitimate


Legitimate
Adversarial
threshold = 1.24
detection: 85%, FP < 5%
140
120
100
80
60
40
20
0
0.0
0.4
0.8
1.2
1.6
2.0
Maximum L_1 distance between original and squeezed input

Detection Performance Summary

Dataset	Most Effective Squeezers Found	Threshold	Detection Rate		False Positive Rate	ROC-AUC	
			Successful AEs	Failed AEs		Including FAEs	Excluding FAEs
MNIST	Bit Depth (1-bit), Median (2x2)	0.0029	98.2%	20.0%	3.98%	94.5%	99.6%
CIFAR-10	Bit Depth (5-bit), Median (2x2), Non-local Mean (13-3-2)	1.1402	85.0%	9.1%	4.93%	95.7%	95.9%
ImageNet	Bit Depth (5-bit), Median (2x2), Non-local Mean (11-3-4)	1.2476	85.2%	25.0%	4.70%	94.0%	94.5%

Detection Performance
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
MNIST
CIFAR-10
ImageNet
FGSM
BIM
CW_∞ (Next)
CW_∞ (LL)
DeepFool
CW_2 (Next)
CW_2 (LL)
CW_0 (Next)
CW_0 (LL)
JSMA (Next)
JSMA (LL)



Composes with model-based defenses

Arms Race?

Adversarial Example Defenses: Ensembles of Weak Defenses are not Strong

Warren He
UC Berkeley

James Wei
UC Berkeley

Xinyun Chen
UC Berkeley

Nicholas Carlini
UC Berkeley

Dawn Song
UC Berkeley

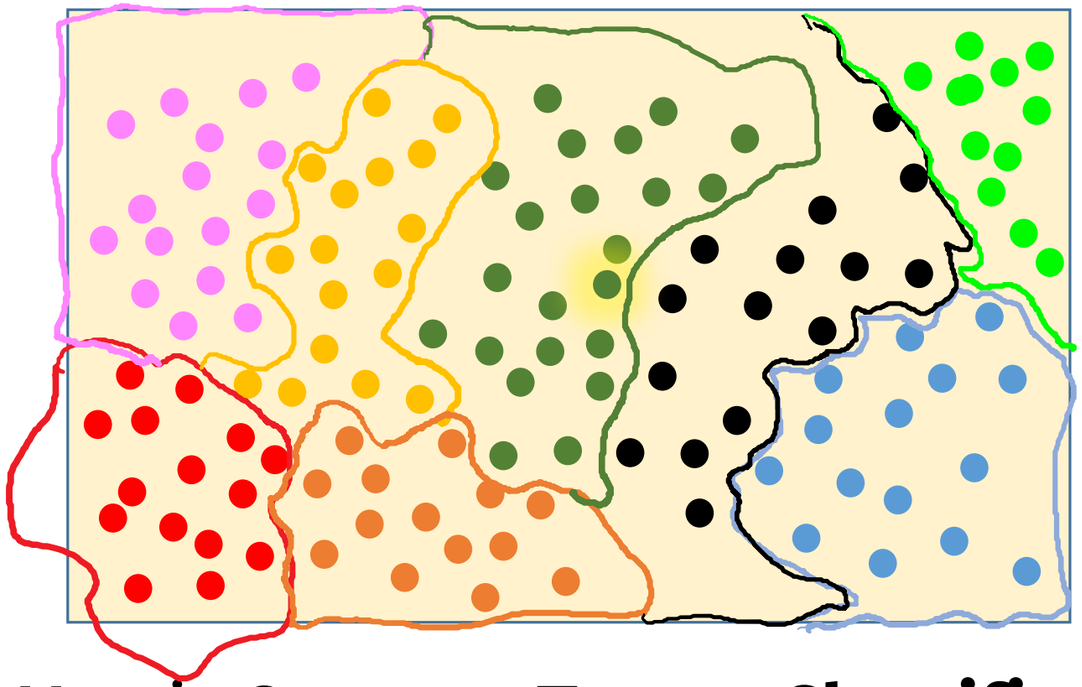
WOOT (August 2017)

Incorporate L_1 squeezed distance into loss function

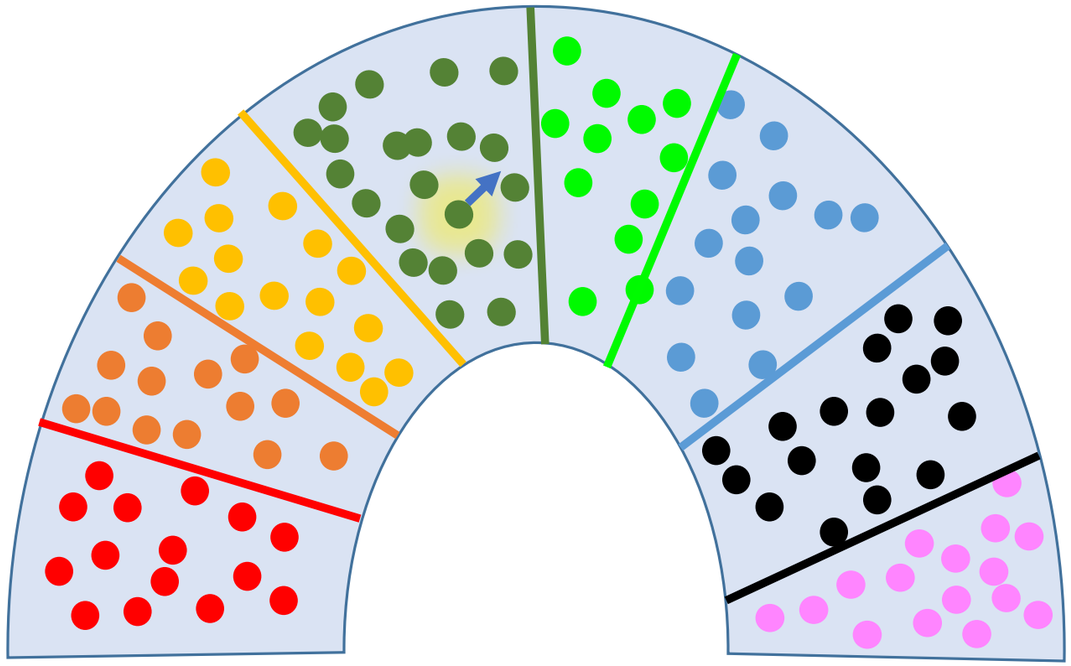
Untargeted	Targeted (Next)	Targeted (Least Likely)
64%	41%	21%

(Adversary success rate on MNIST)

Raising the Bar or *Changing the Game?*



Metric Space 1: **Target Classifier**



Metric Space 2: **“Oracle”**

Before: find a small perturbation that changes class for classifier, but imperceptible to oracle.

Now: change class for both **original** and **squeezed** classifier, but imperceptible to oracle.

“Feature Squeezing” Conjecture

For any *distance-limited* adversarial method, there exists *some* feature squeezer that accurately detects its adversarial examples.

Intuition: if the perturbation is *small* (in some simple metric space), there is some squeezer that coalesces original and adversarial example into same sample.

Defender's Entropy Advantage

random seed

Input

Model

$Prediction_0$

Squeezer$_1$

Model

$Prediction_1$

Squeezer$_2$

Model

$Prediction_2$

Squeezer$_k$

Model'

$Prediction_k$

$f(pred_0, pred_1, \ldots, pred_k)$

Yes

Adversarial

No

Legitimate

More Complex Squeezers + Entropy

CCS 2017

MagNet: a Two-Pronged Defense against Adversarial Examples

Dongyu Meng
ShanghaiTech University
mengdy@shanghaitech.edu.cn

Hao Chen
University of California, Davis
chen@ucdavis.edu

Pick a random autoencoder

Changing the Game

Option 1:

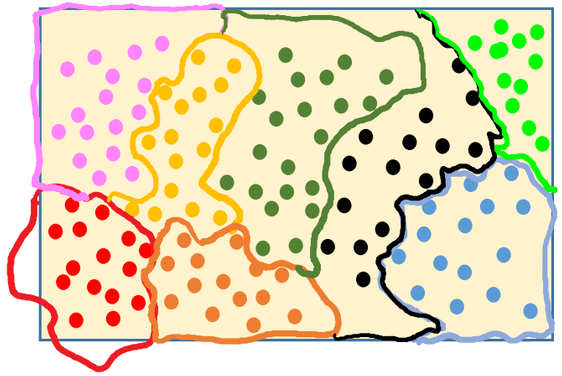
Find distance-limited adversarial methods for which it is intractable to find effective feature squeezers.

Option 2:

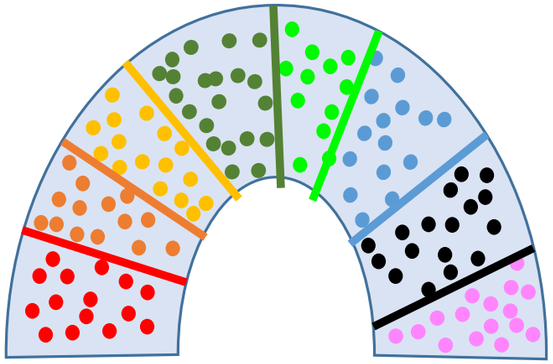
Redefine adversarial examples so *distance is not limited* in a simple metric space…

focus of rest of the talk

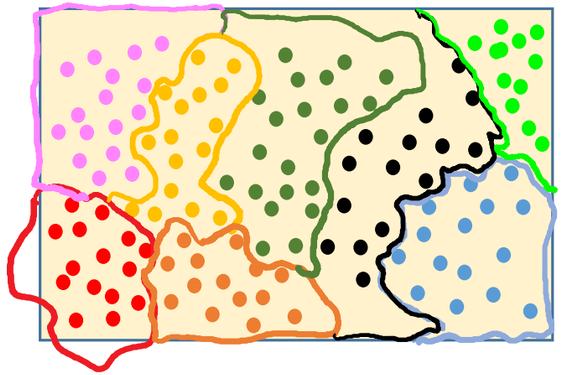
Do Humans Matter?



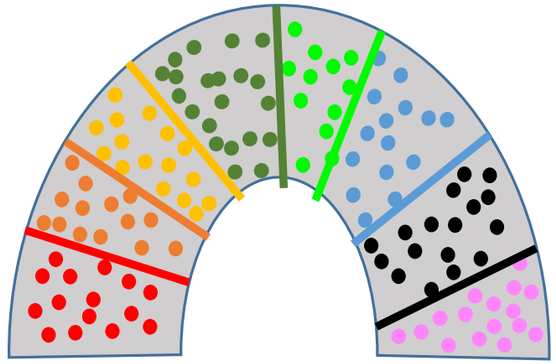
Metric Space 1:
Machine



Metric Space 2:
Human

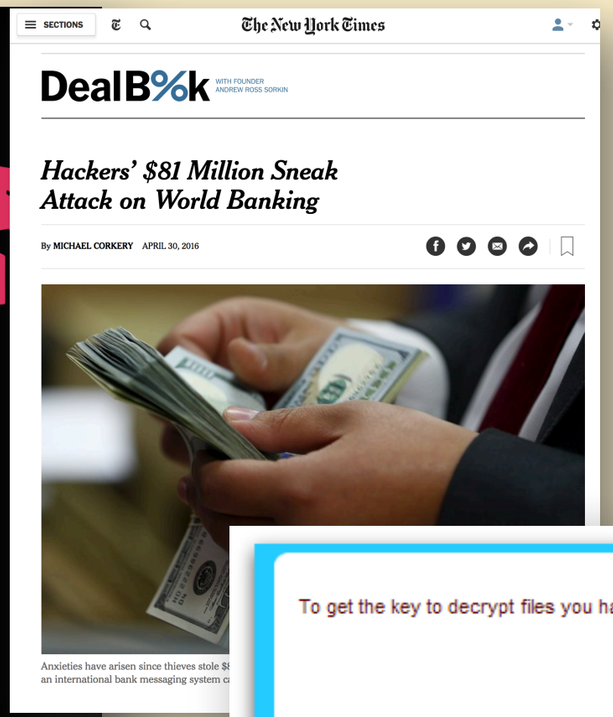


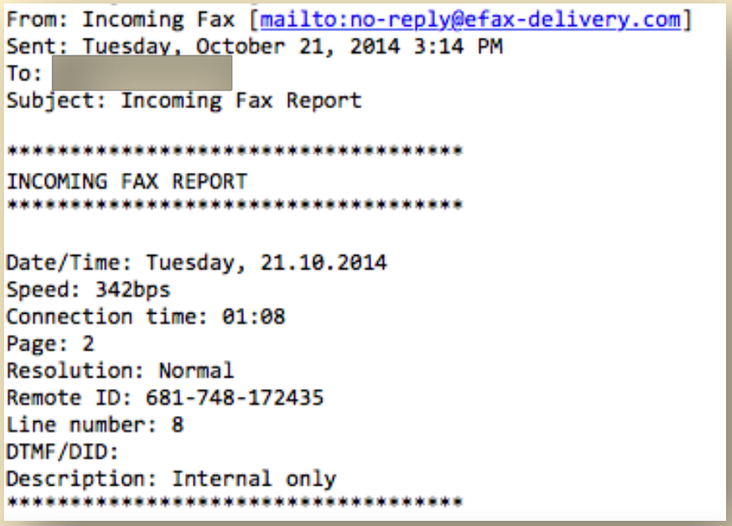
Metric Space 1:
Machine 1



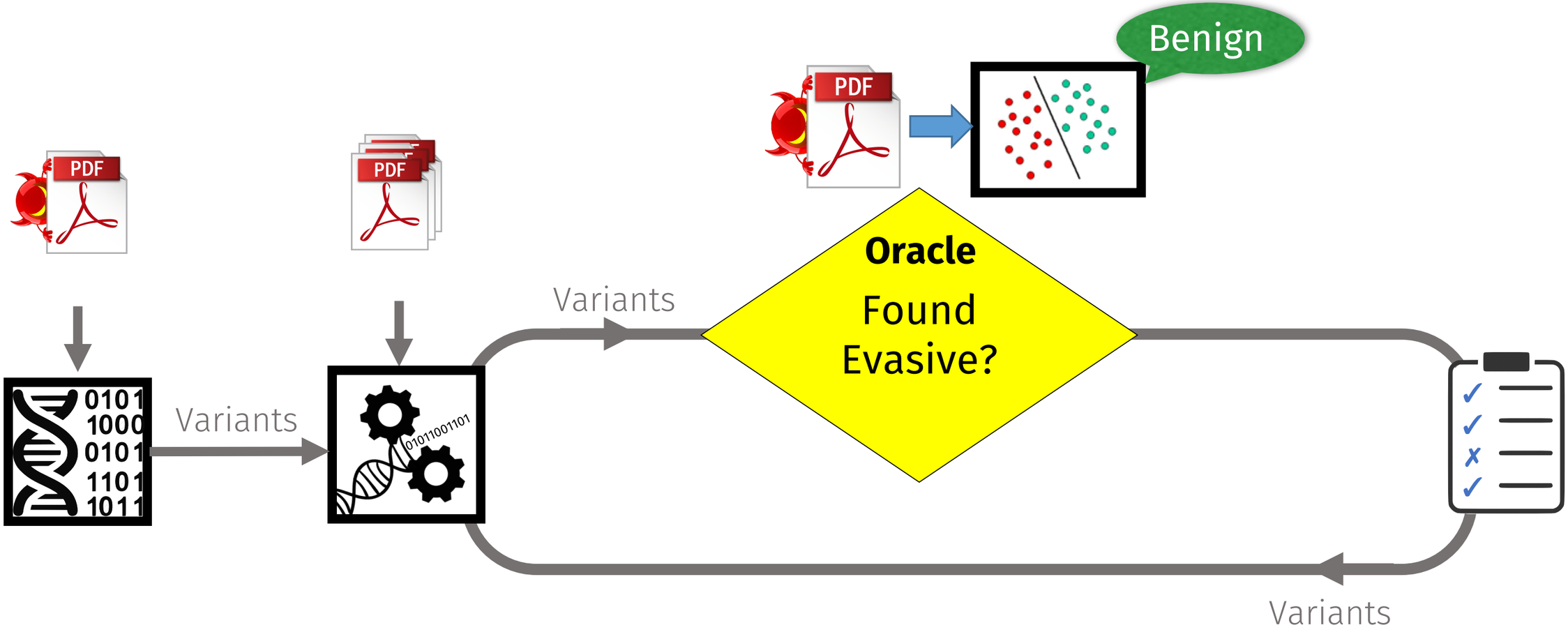
Metric Space 2:
Machine 2

Malware Classifiers

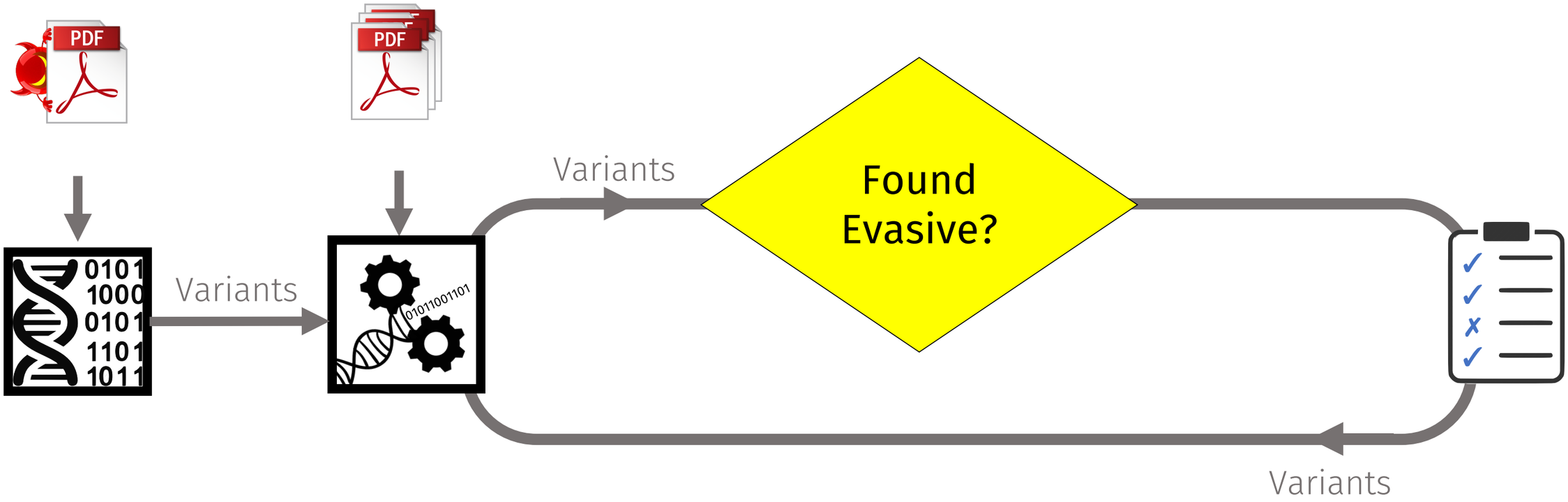




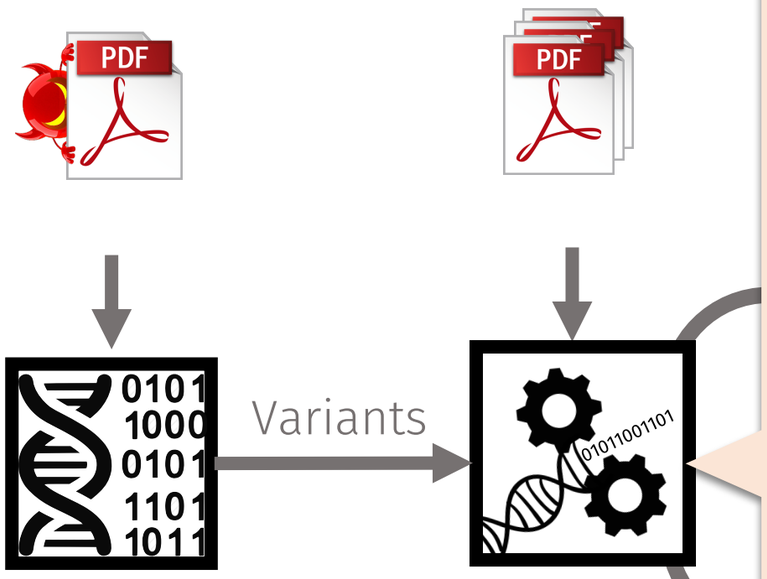
Automated Classifier Evasion Using **Genetic Programming**

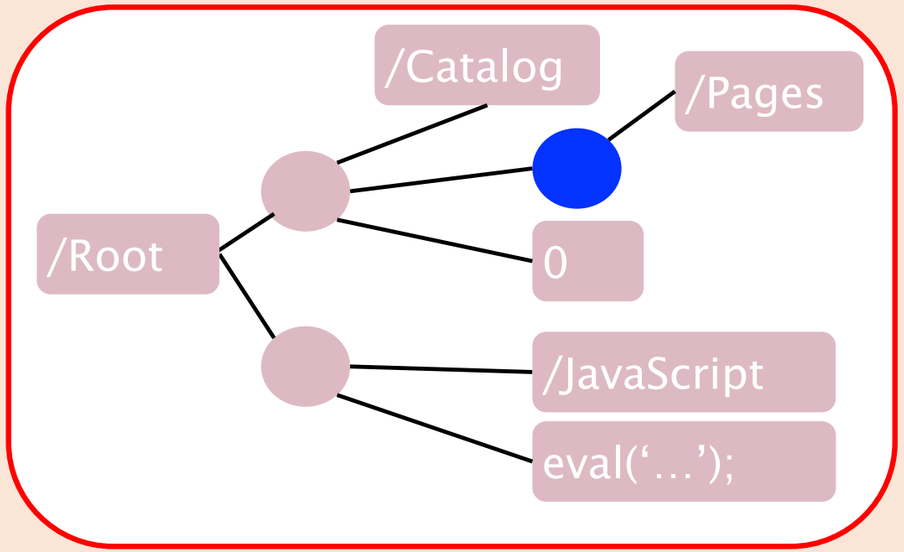


Generating Variants



Generating Variants

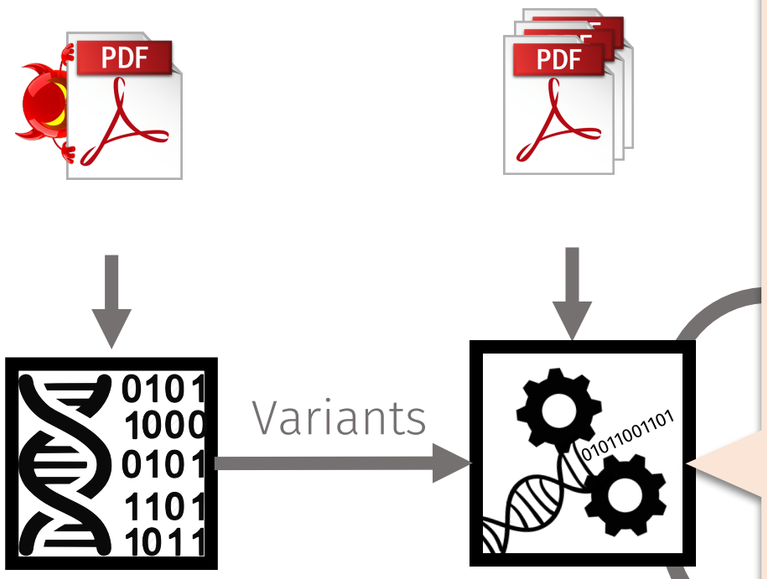


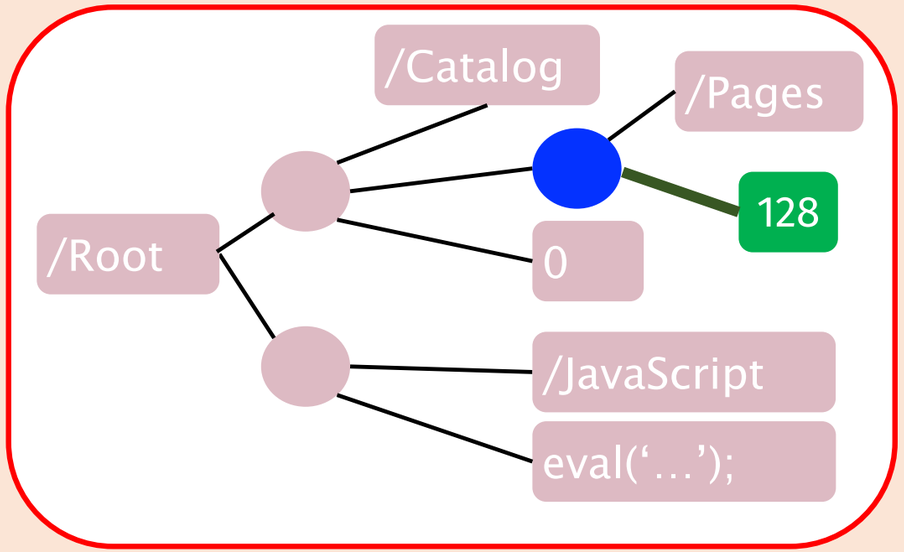


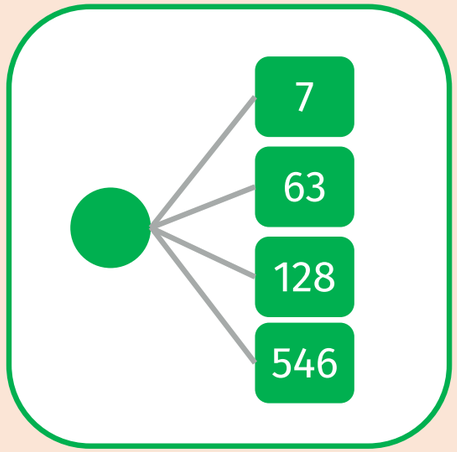
Select **random** node

Randomly transform: **delete**, insert, replace

Generating Variants





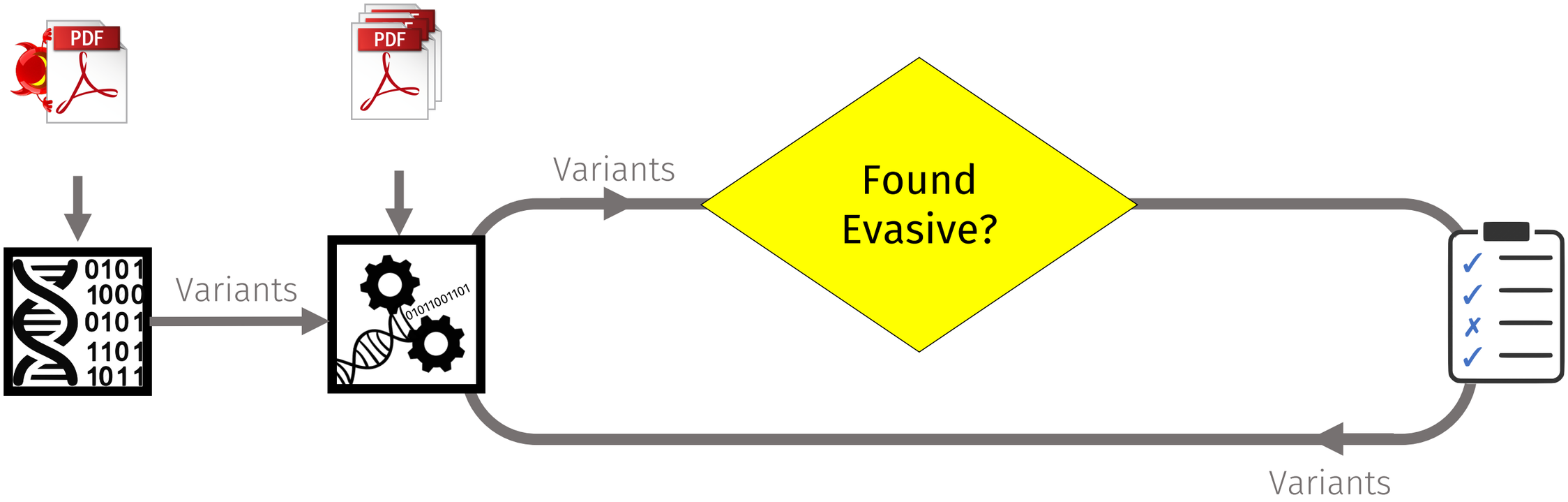


Nodes from Benign PDFs

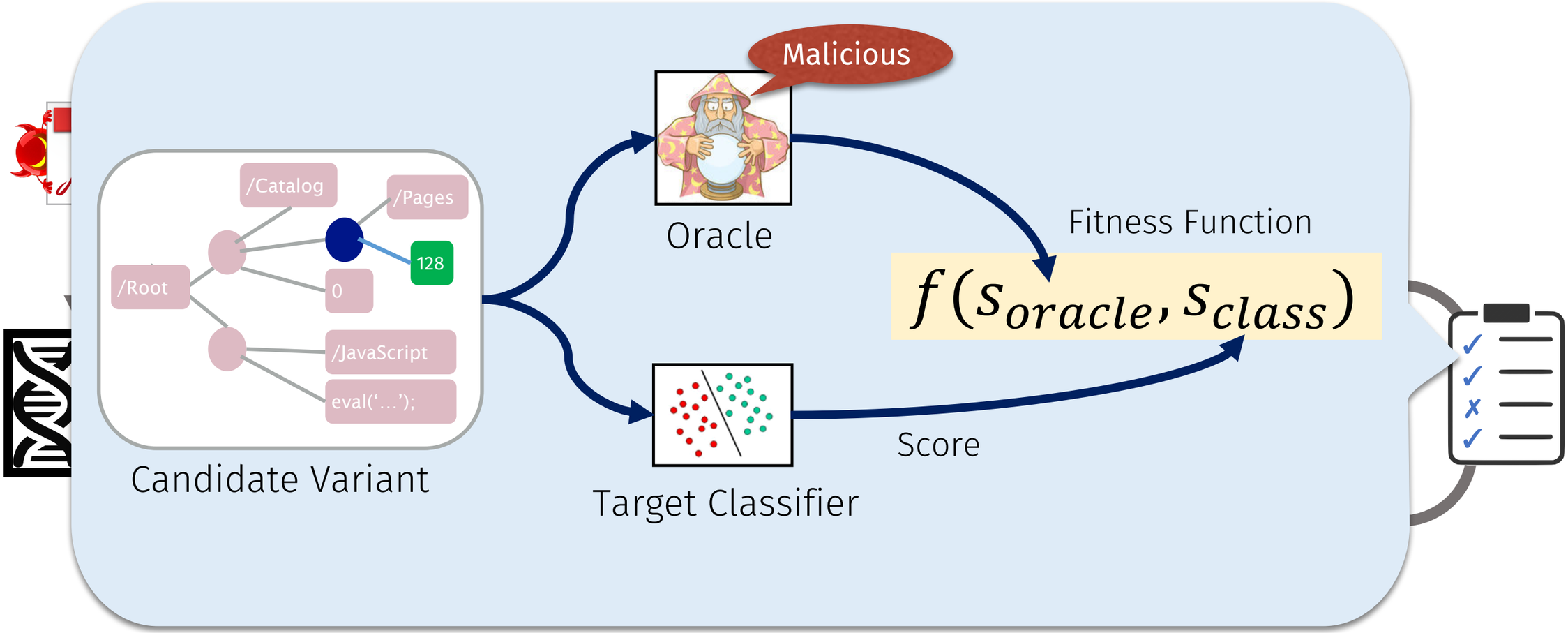
Select **random** node

Randomly transform: delete, **insert**, replace

Selecting Promising Variants



Selecting Promising Variants



Oracle

Execute candidate in vulnerable Adobe Reader in virtual environment

Cuckoo

https://github.com/cuckoosandbox

Simulated network: INetSim

Behavioral signature:

malicious if signature matches

HTTP_URL + HOST
extracted from API traces

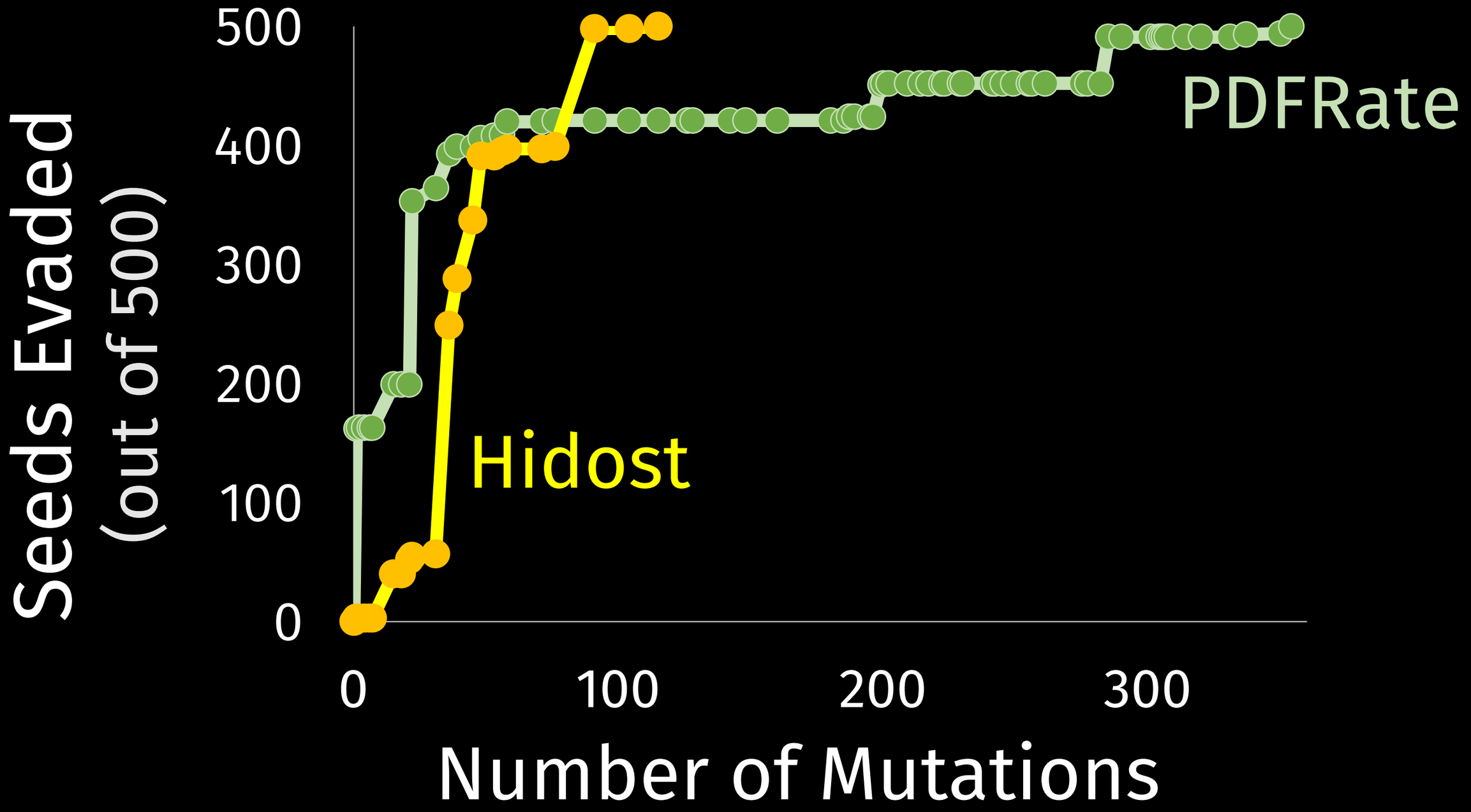
Advantage: we know the target malware behavior

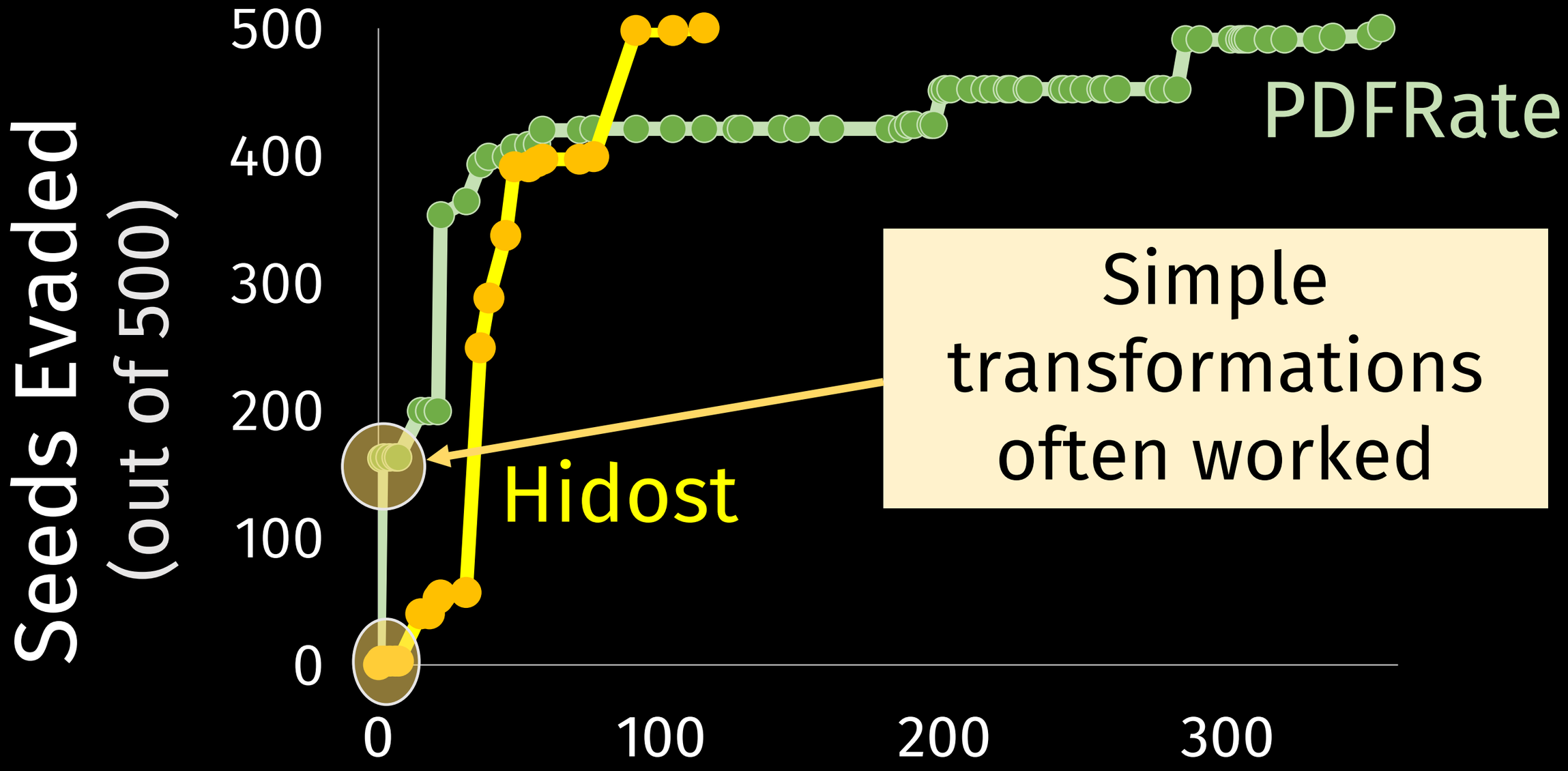
Fitness Function

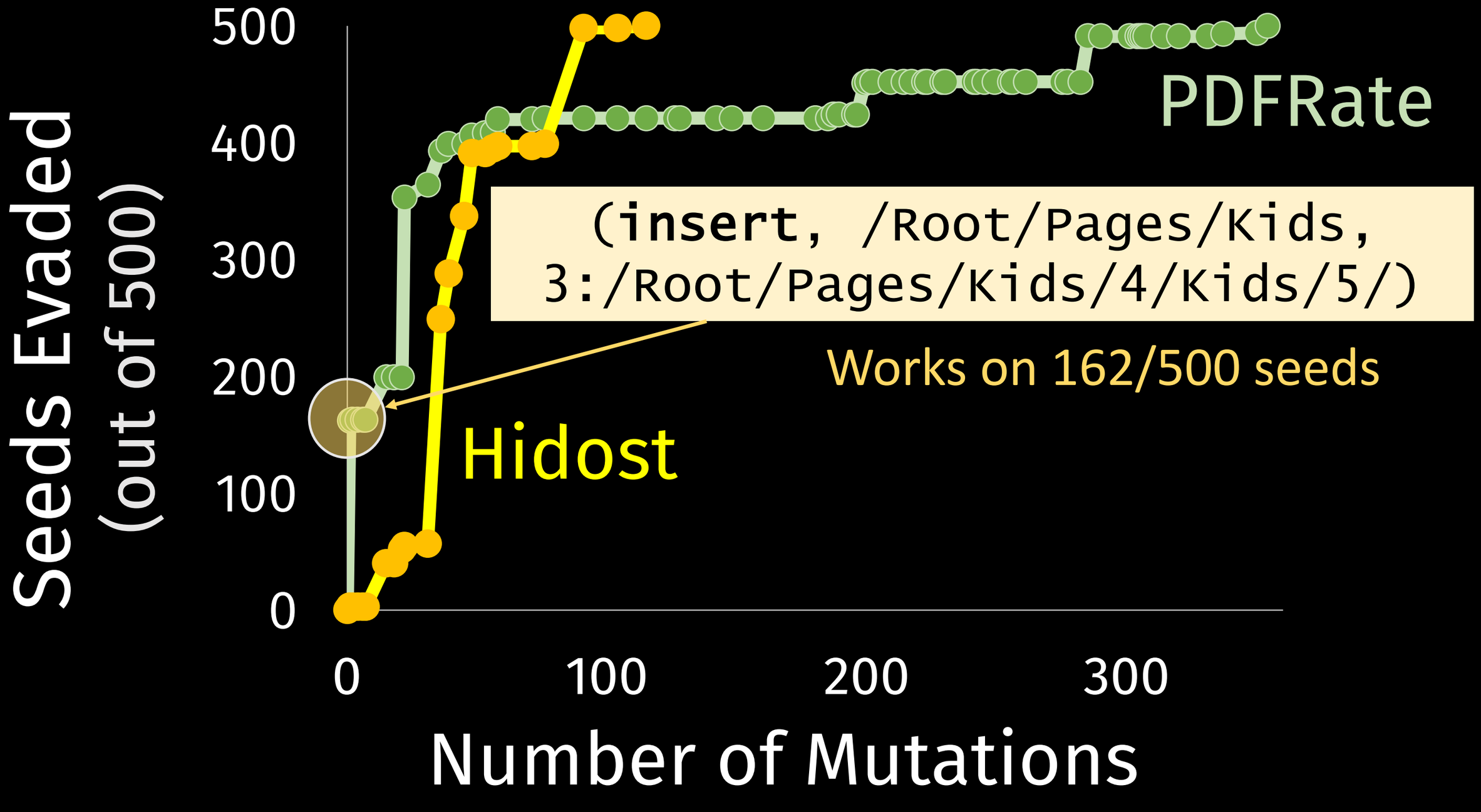
Assumes lost malicious behavior will not be recovered

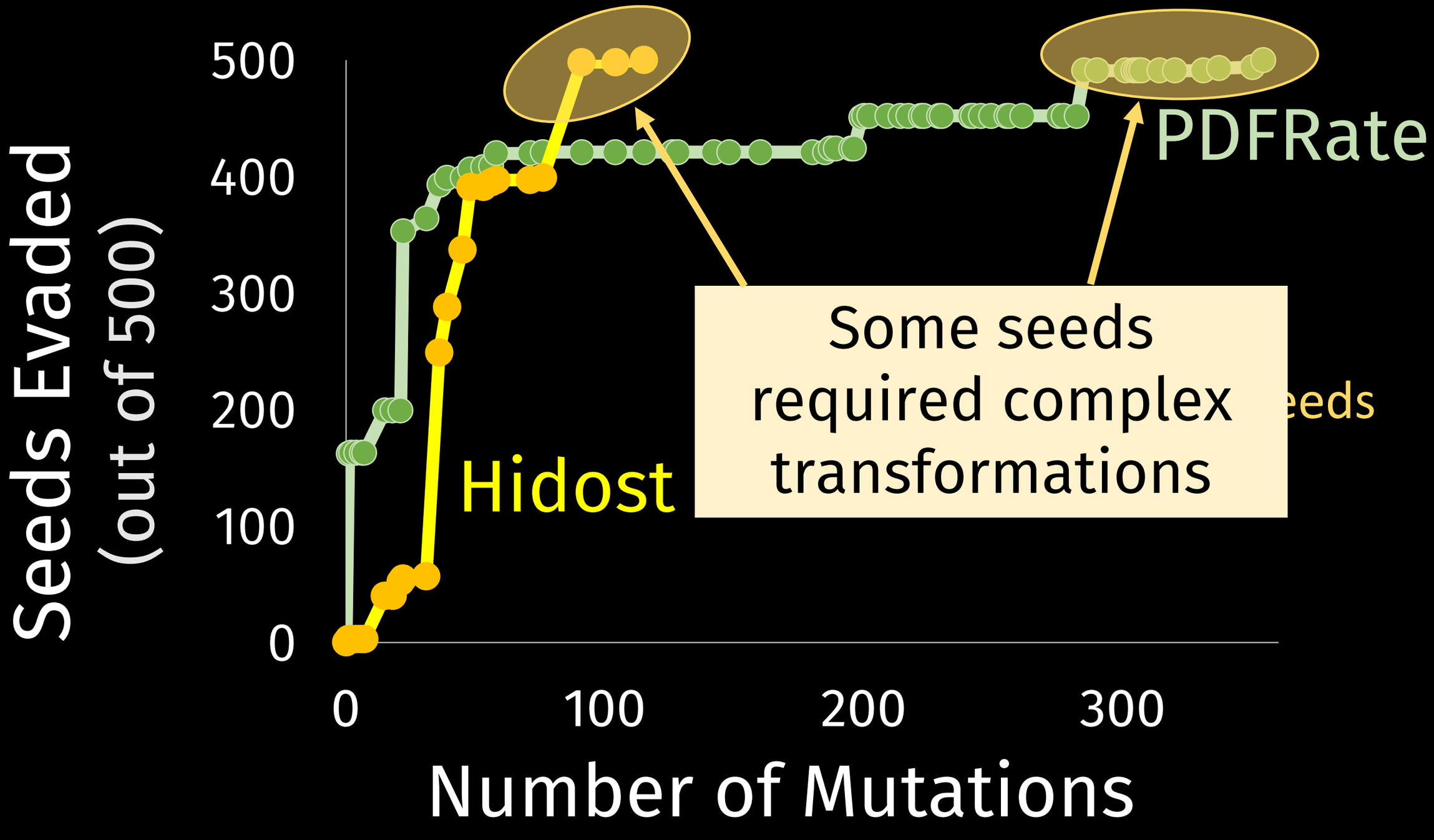
$$f(v) = \begin{cases} .5 - \mathit{classifier\_score}(v) & \text{if } \mathit{oracle}(v) = \text{"malicious"} \\ -\infty & \text{otherwise} \end{cases}$$

$\mathit{classifier\_score} \geq 0.5$: labeled malicious


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

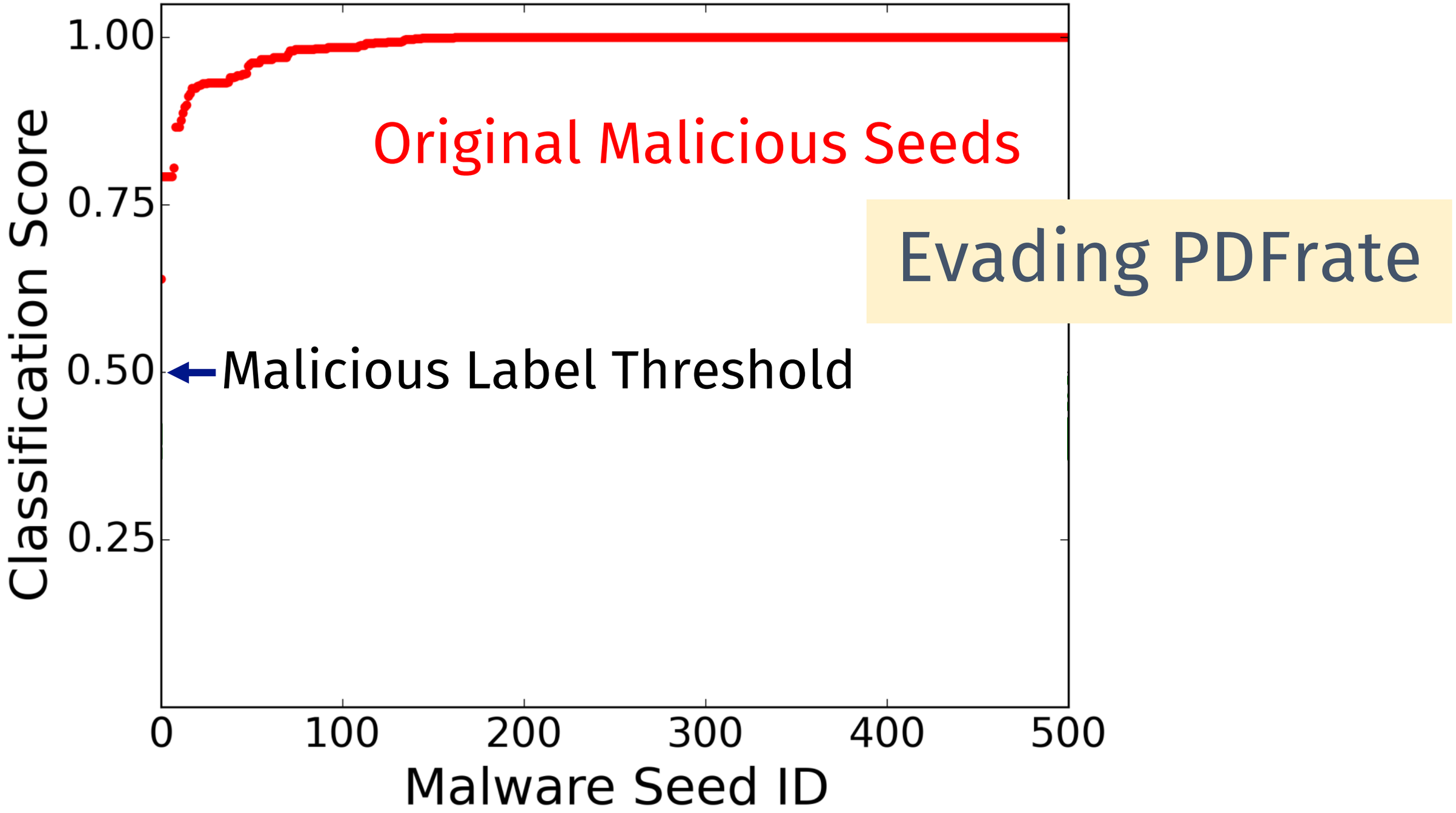

Seeds Evaded
(out of 500)
500
400
300
200
100
0
0
100
200
300
Number of Mutations
PDFRate
(insert, /Root/Pages/Kids,
3:/Root/Pages/Kids/4/Kids/5/)
Works on 162/500 seeds
Hidost


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

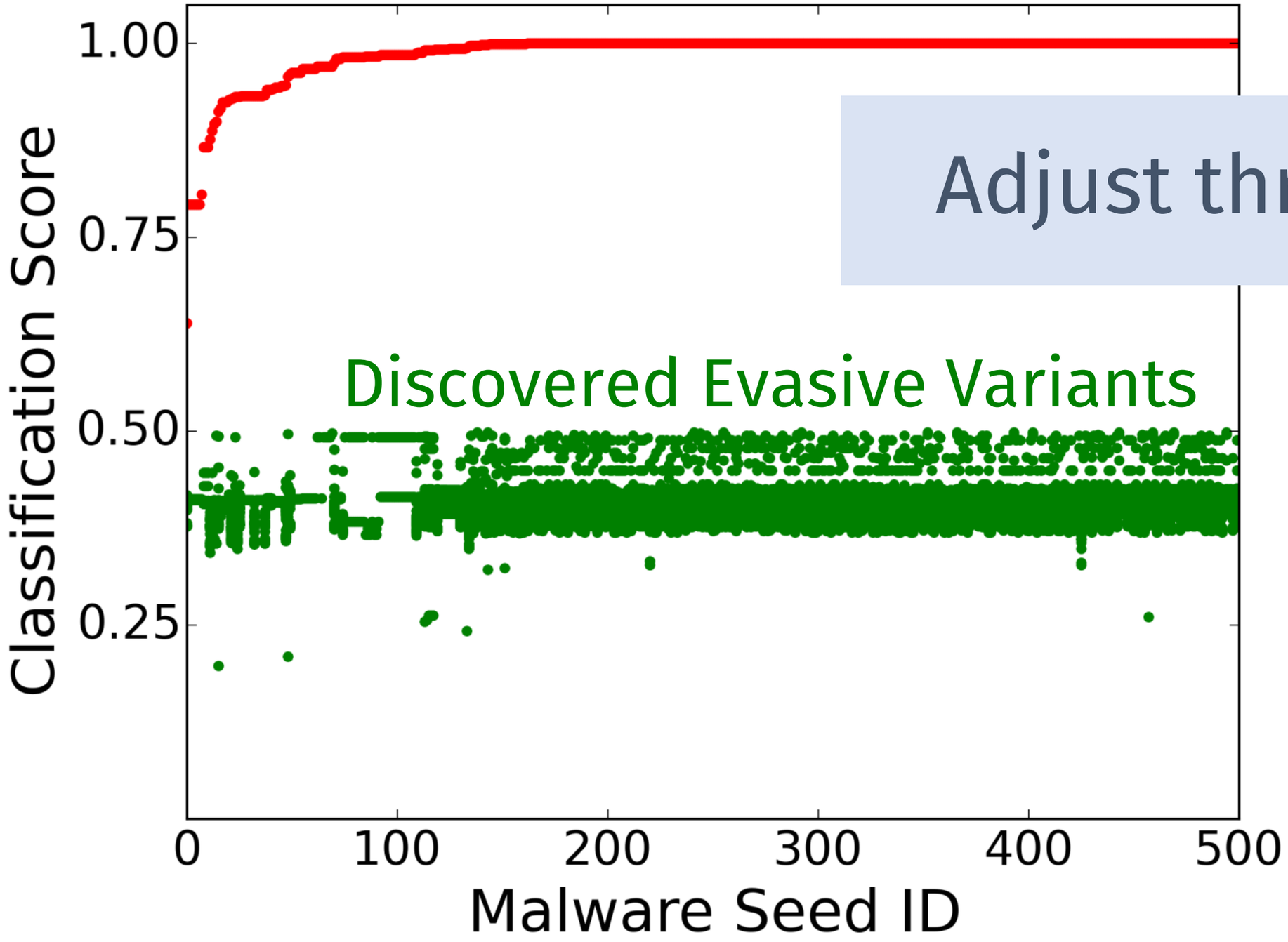
Possible Defenses

Possible Defense:
Adjust Threshold

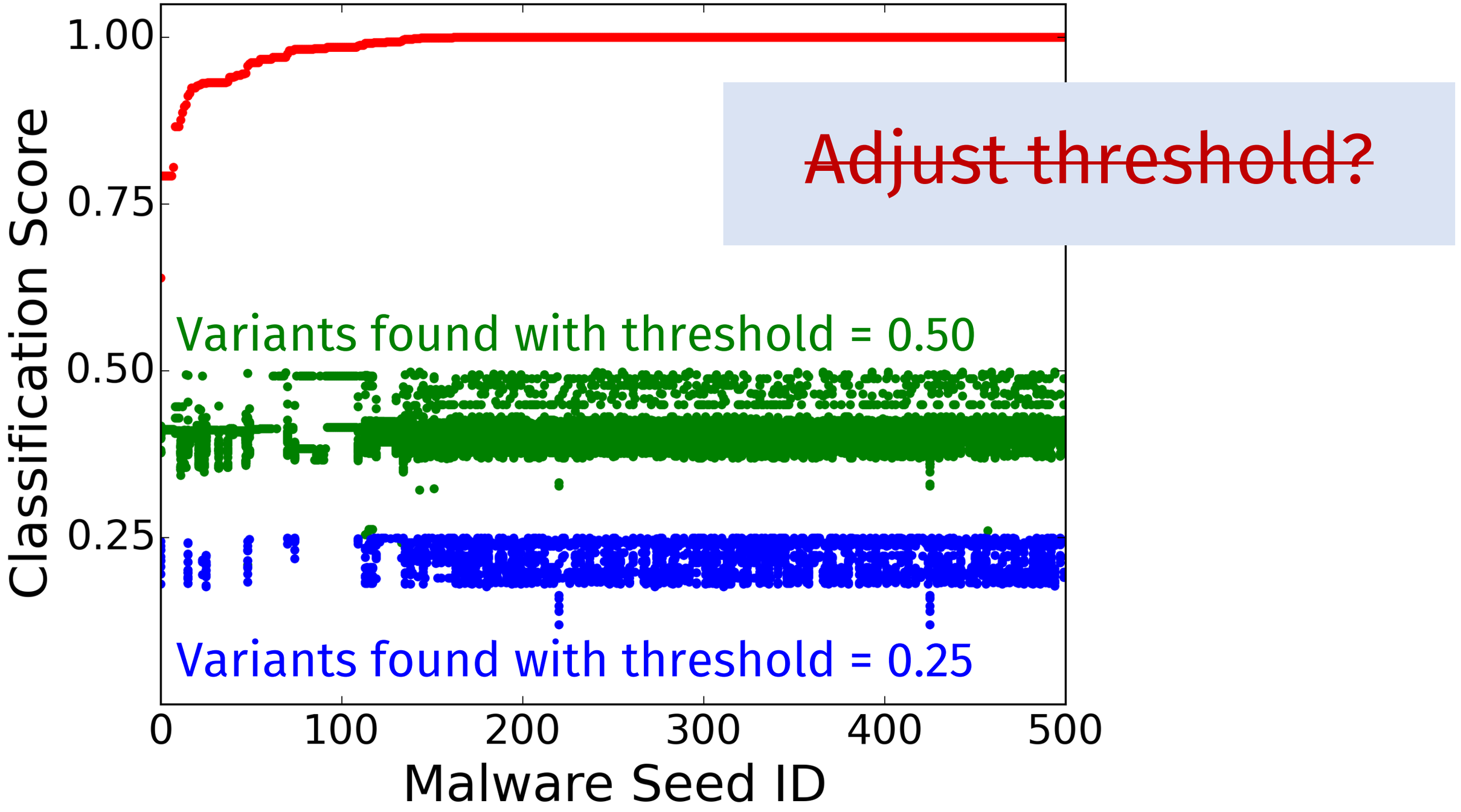
Charles Smutz, Angelos Stavrou. When a Tree Falls: Using Diversity in Ensemble Classifiers to Identify Evasion in Malware Detectors. NDSS 2016.



Evading PDFrate



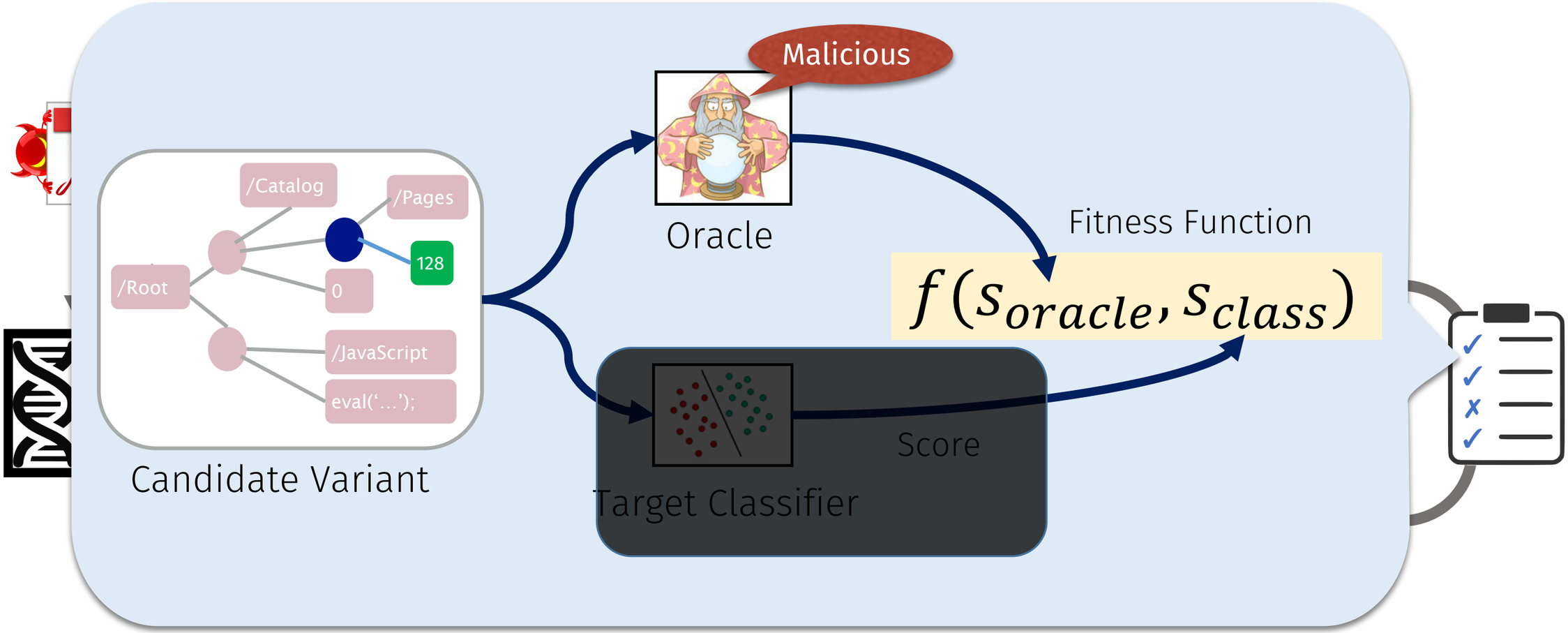
Adjust threshold?



~~Adjust threshold?~~

Possible Defense:
Hide Classifier

Hide the Classifier Score?



Binary Classifier Output is Enough

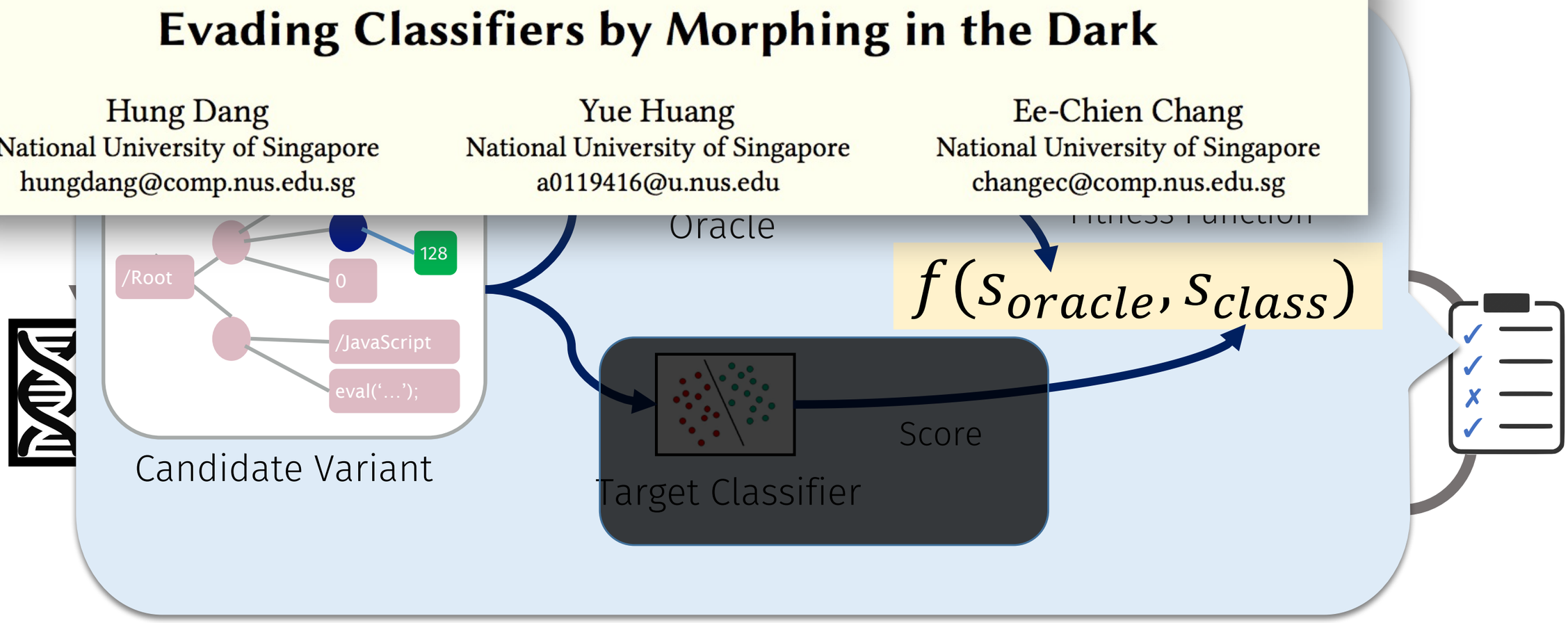
ACM CCS 2017

Evading Classifiers by Morphing in the Dark

Hung Dang
National University of Singapore
hungdang@comp.nus.edu.sg

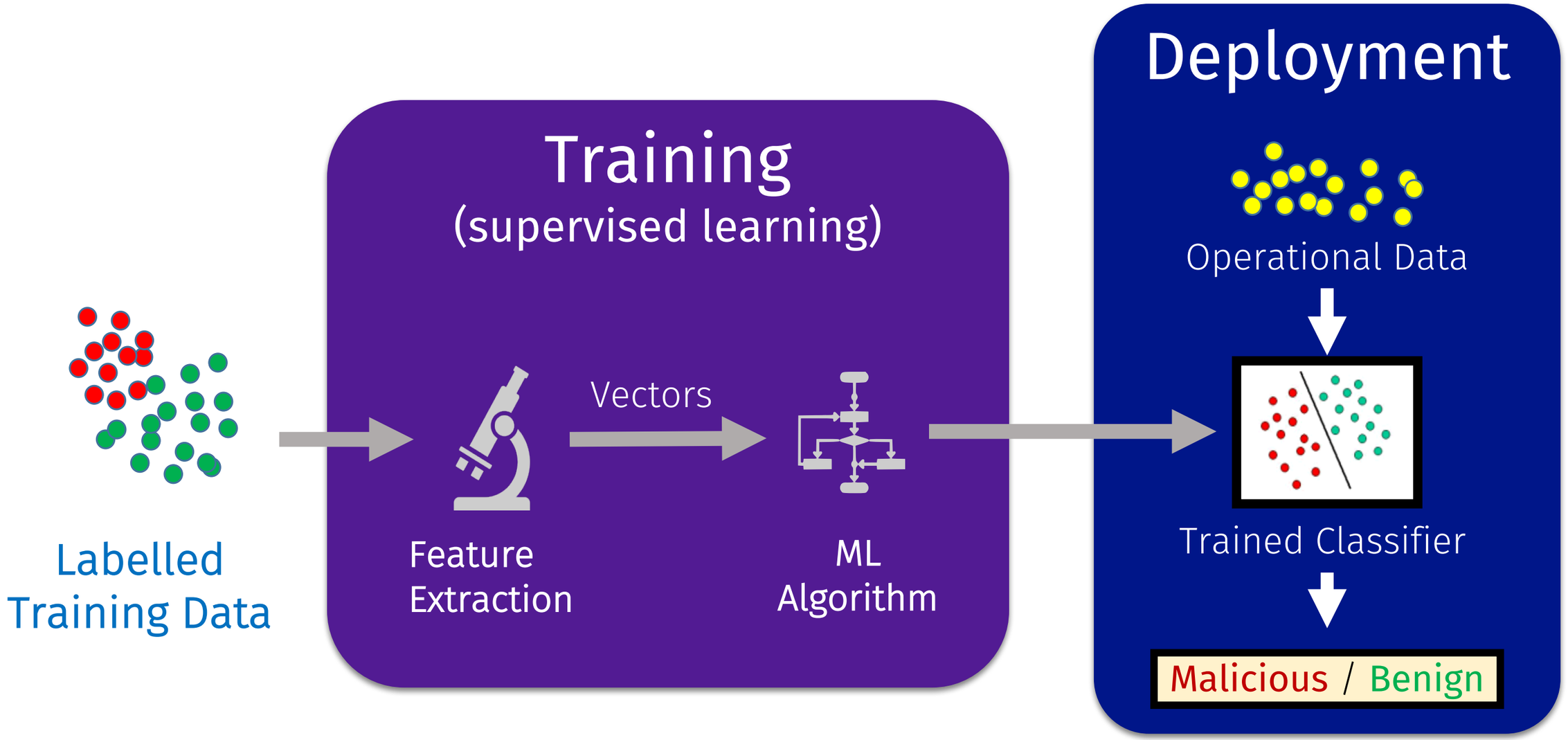
Yue Huang
National University of Singapore
a0119416@u.nus.edu

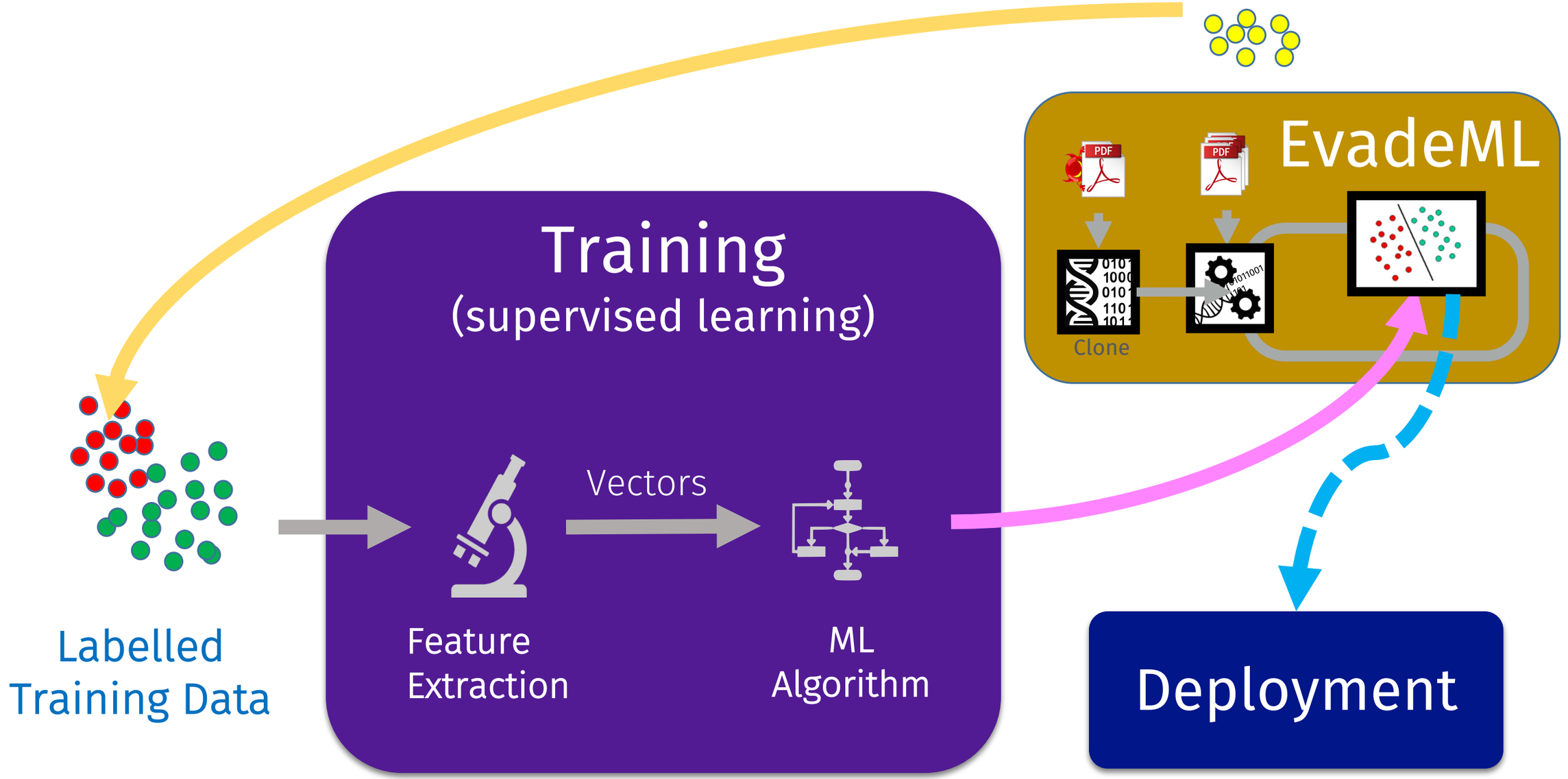
Ee-Chien Chang
National University of Singapore
changec@comp.nus.edu.sg

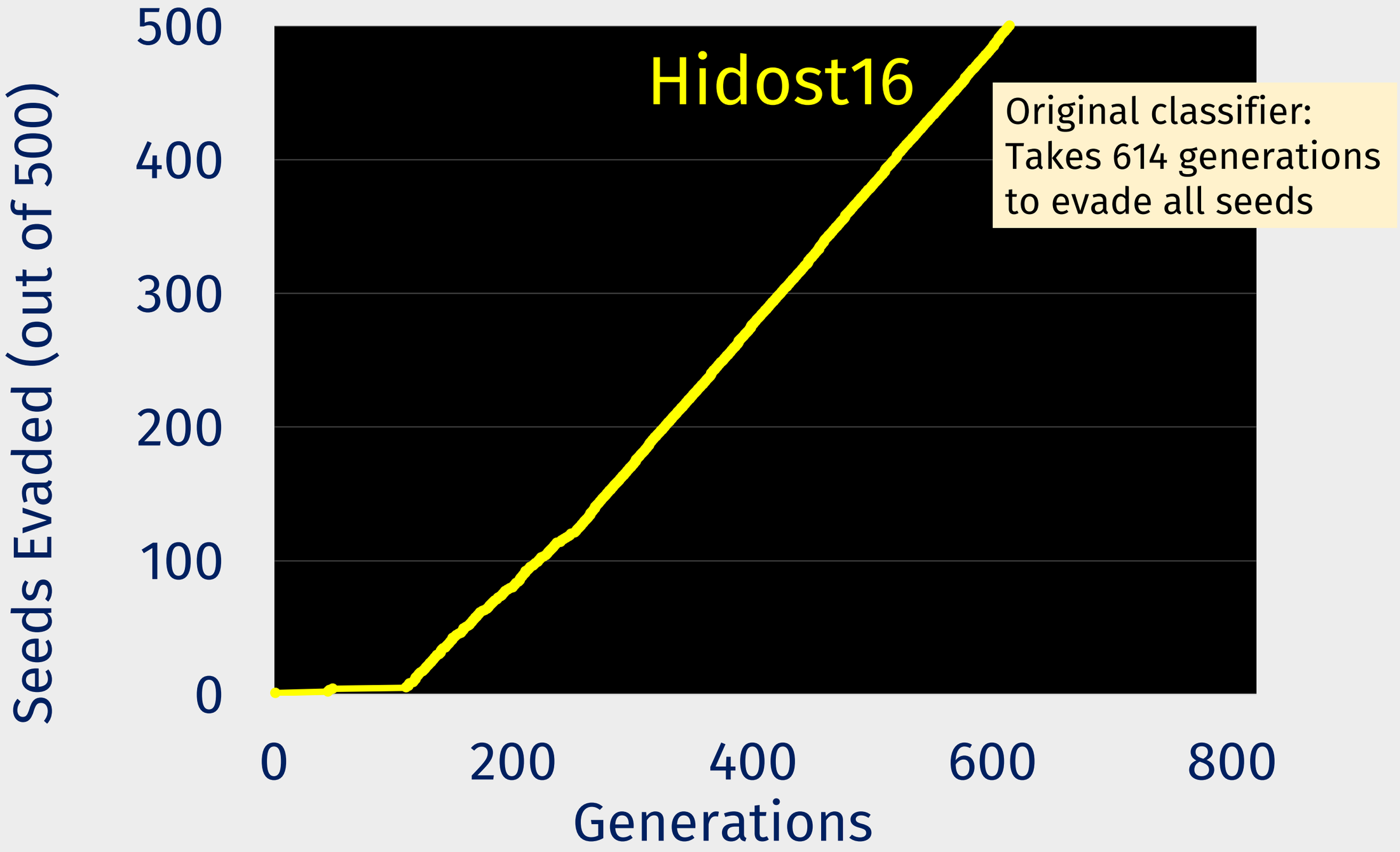


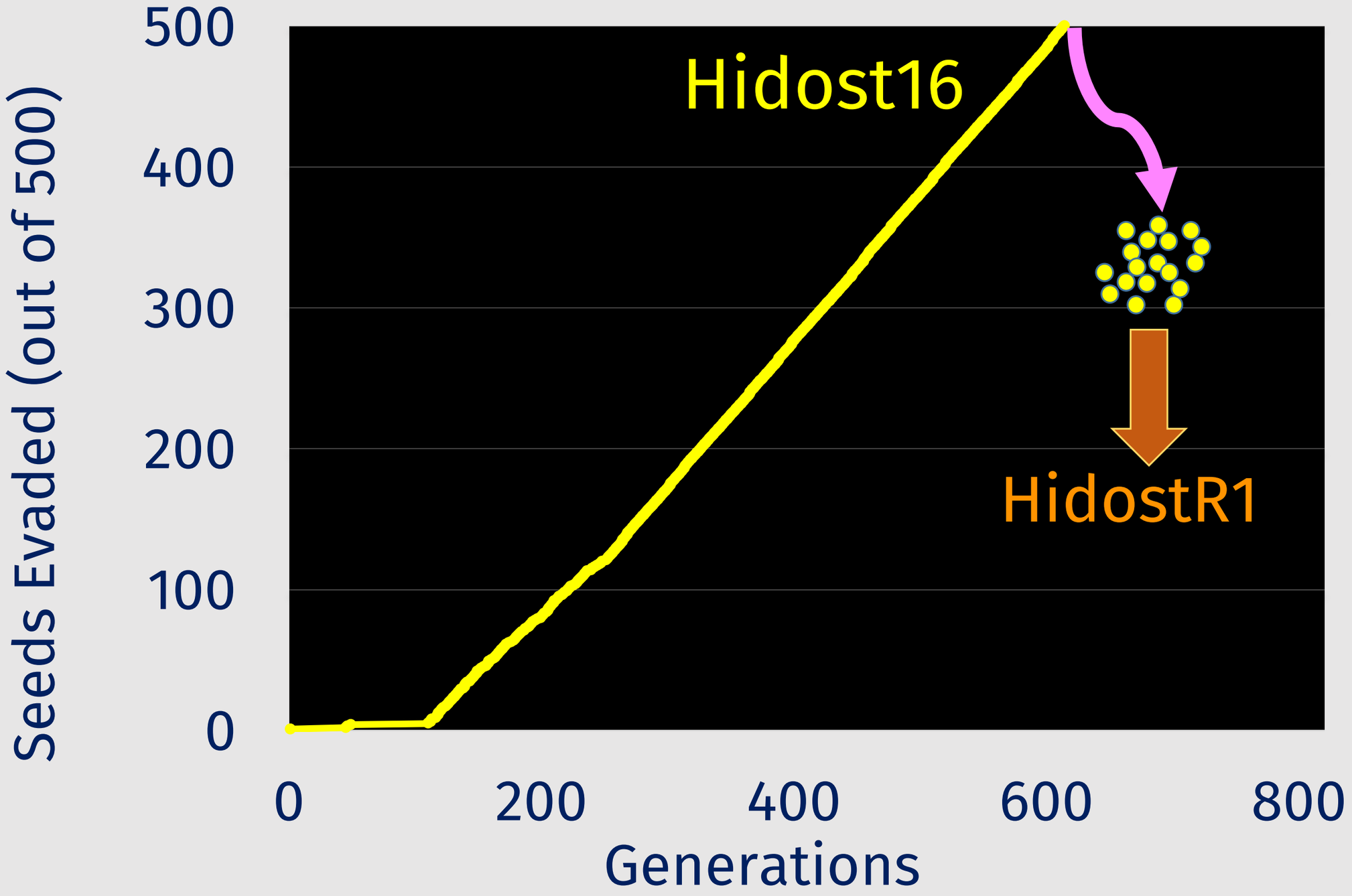
Possible Defense:
Retrain Classifier

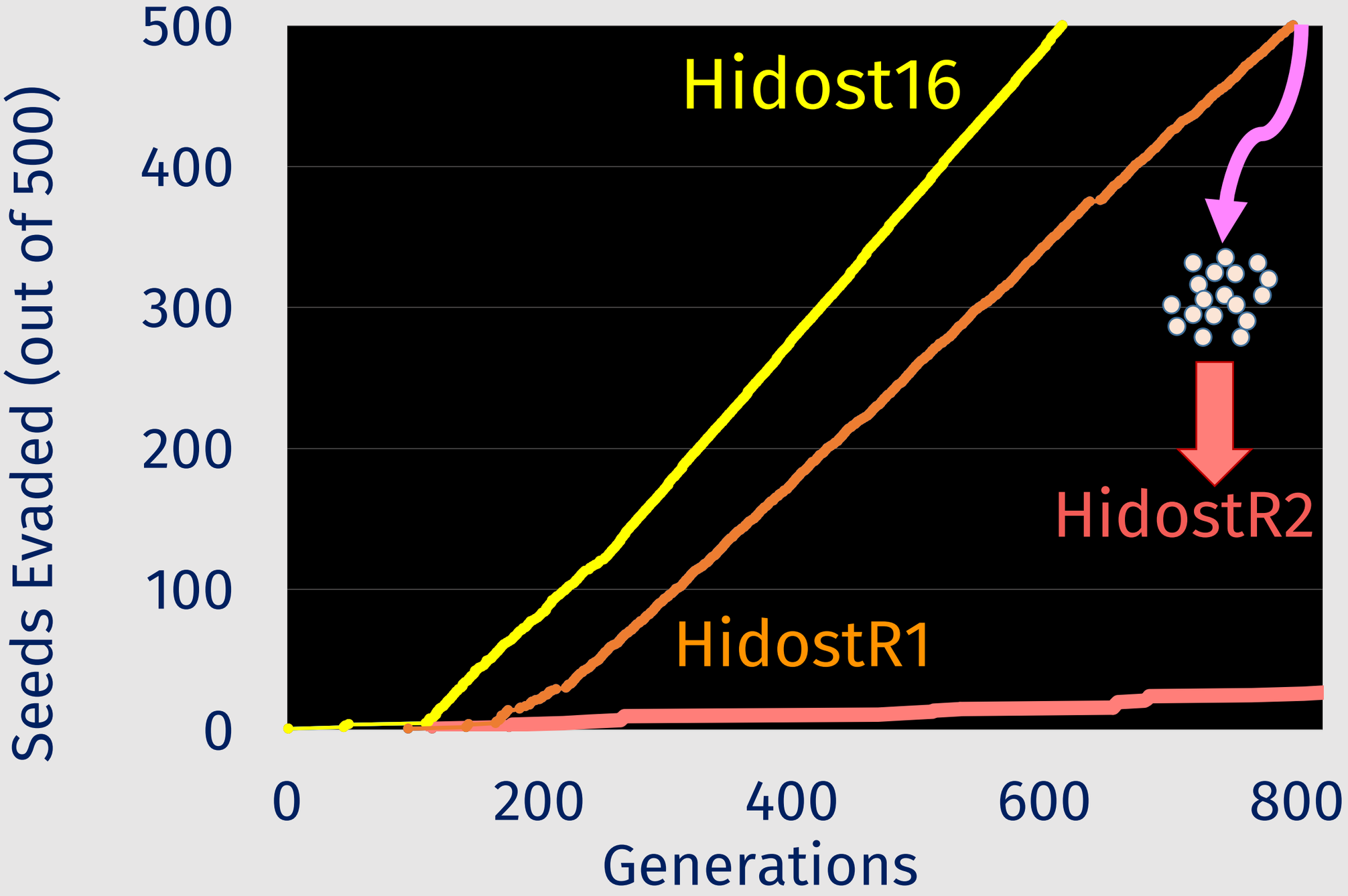
Retrain Classifier

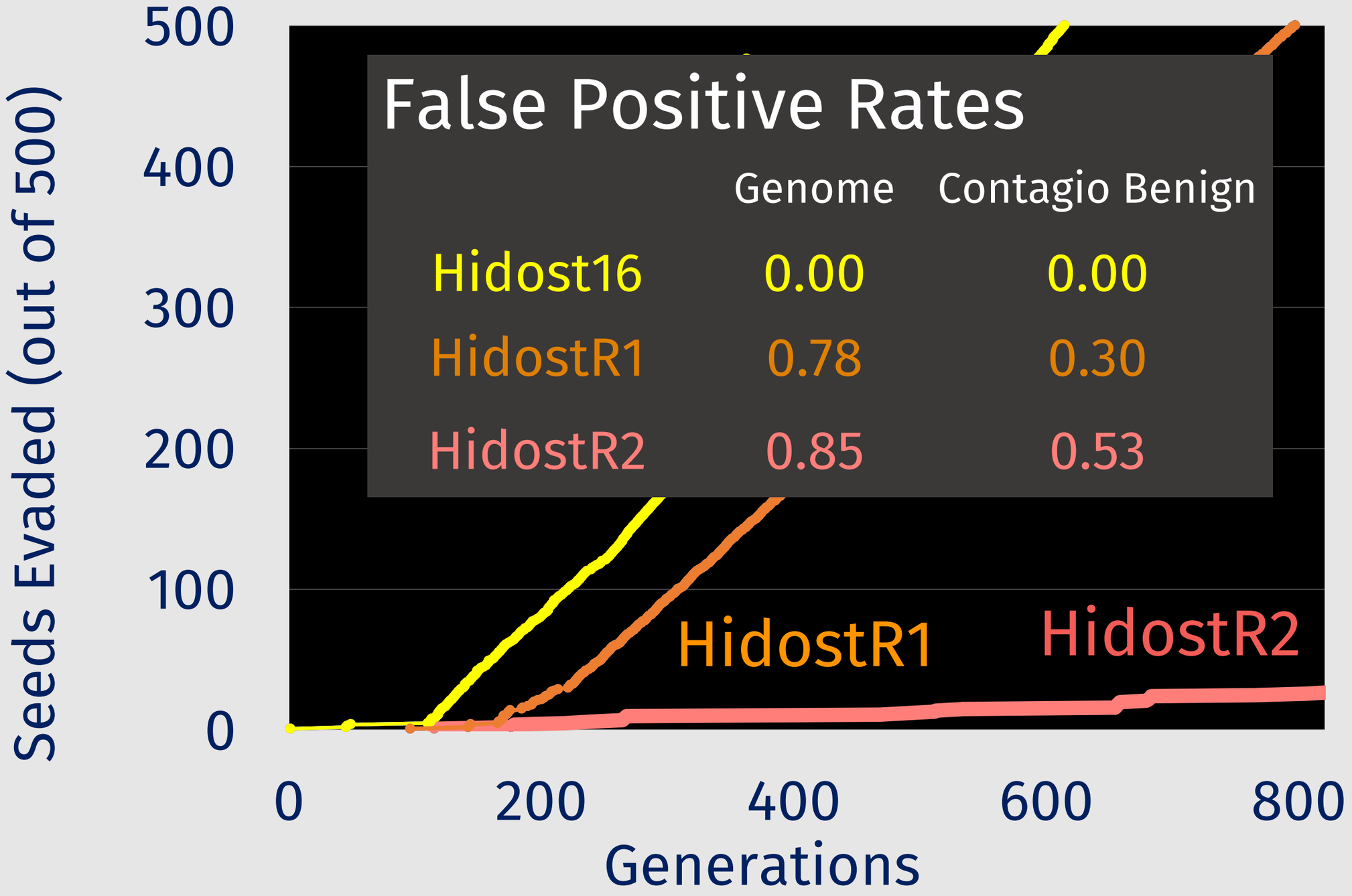



EvadeML
PDF
PDF
Clone
Training
(supervised learning)
Vectors
Feature
Extraction
ML
Algorithm
Labelled
Training Data
Deployment


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


Hidost16
HidostR1
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
Hidost16
HidostR2
HidostR1
0
200
400
600
800
Generations


False Positive Rates
Genome
Contagio Benign
Hidost16
0.00
0.00
HidostR1
0.78
0.30
HidostR2
0.85
0.53
HidostR1
HidostR2
Seeds Evaded (out of 500)
500
400
300
200
100
0
0
200
400
600
800
Generations

Feature Conservation in Adversarial Classifier Evasion: A Case Study

Liang Tong
Vanderbilt University
liang.tong@vanderbilt.edu

Bo Li
University of California, Berkeley
crystalboli@berkeley.edu

Chen Hajaj
Vanderbilt University
chen.hajaj@vanderbilt.edu

Yevgeniy Vorobeychik
Vanderbilt University
yevgeniy.vorobeychik@vanderbilt.edu

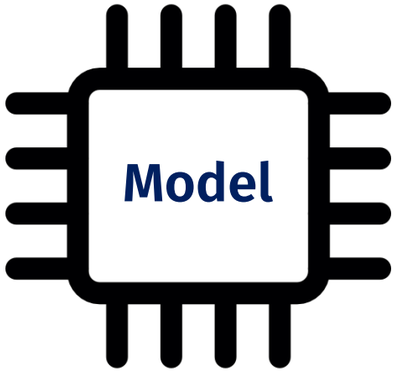
/Names
/Names /JavaScript
/Names /JavaScript /Names
/Names /JavaScript /JS
/OpenAction
/OpenAction /JS
/OpenAction /S
/Pages

Only 8/6987 robust features (Hidost)
Robust classifier
High false positives

evademl.org/zoo

EvadeML-Zoo: an AML Toolbox

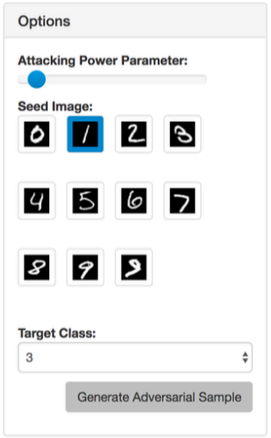
MNIST
CIFAR-10
ImageNet



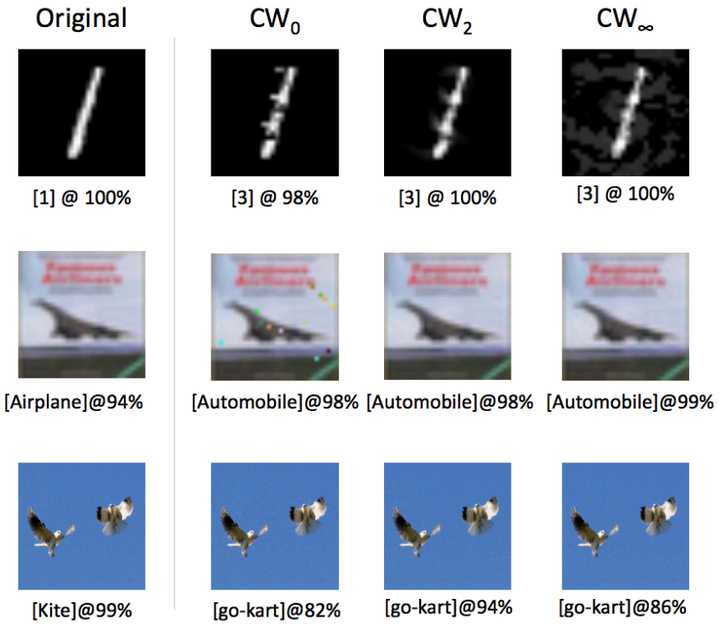
CNN
DenseNet
MobileNets

FGSM, BIM,
JSMA, DeepFool,
CW_2, CW_∞, CW_0

Feature Squeezing



Visualization



Weilin Xu, Andrew Norton,
Noah Kim, Yanjun Qi

Open Questions

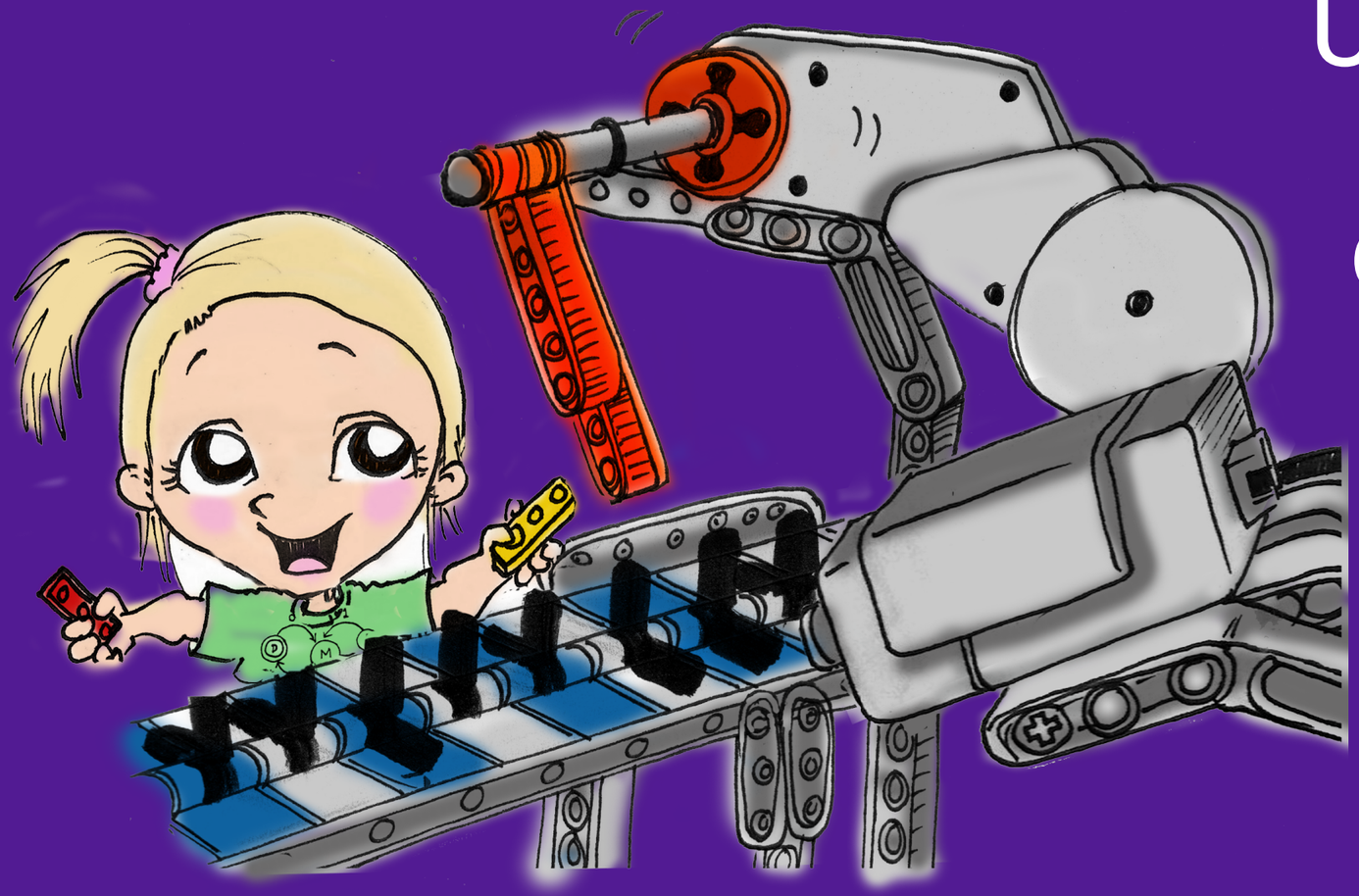
Can we close the gap between experimental techniques (that work on complex models) and formal methods (that work on small models)?

Reducing adversarial search space

Will classifiers ever be good enough to apply “crypto” standards to adversarial examples?

Is PDF Malware the MNIST of malware classification?

EvadeML.org



David Evans

University of Virginia

evans@virginia.edu

EvadeML.org

source code, papers