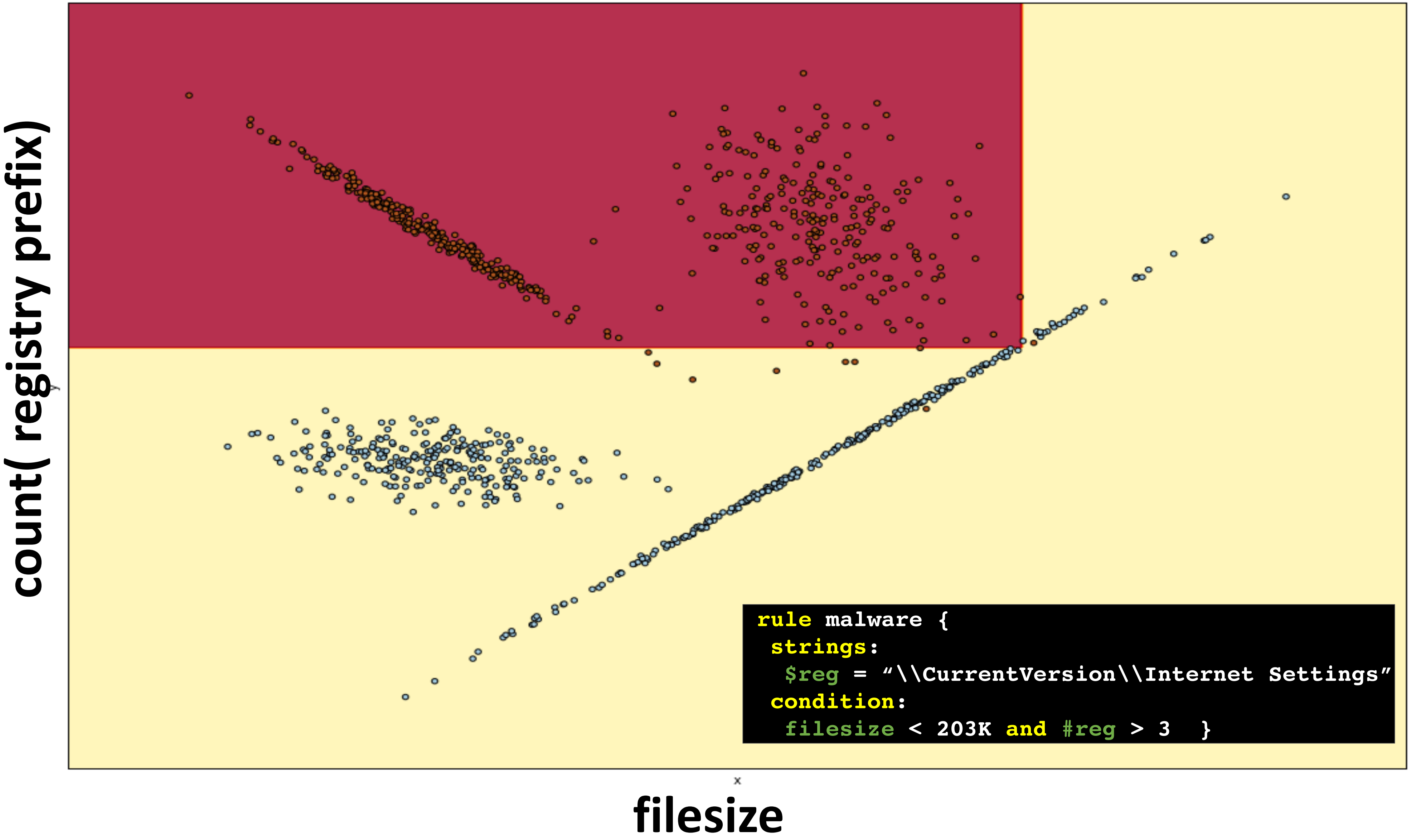
Bot vs. Bot:
Evading Machine Learning Malware Detection

Hyrum Anderson

hyrum@endgame.com @drhyrum /in/hyrumanderson

black hat USA 2017

Why Machine Learning?

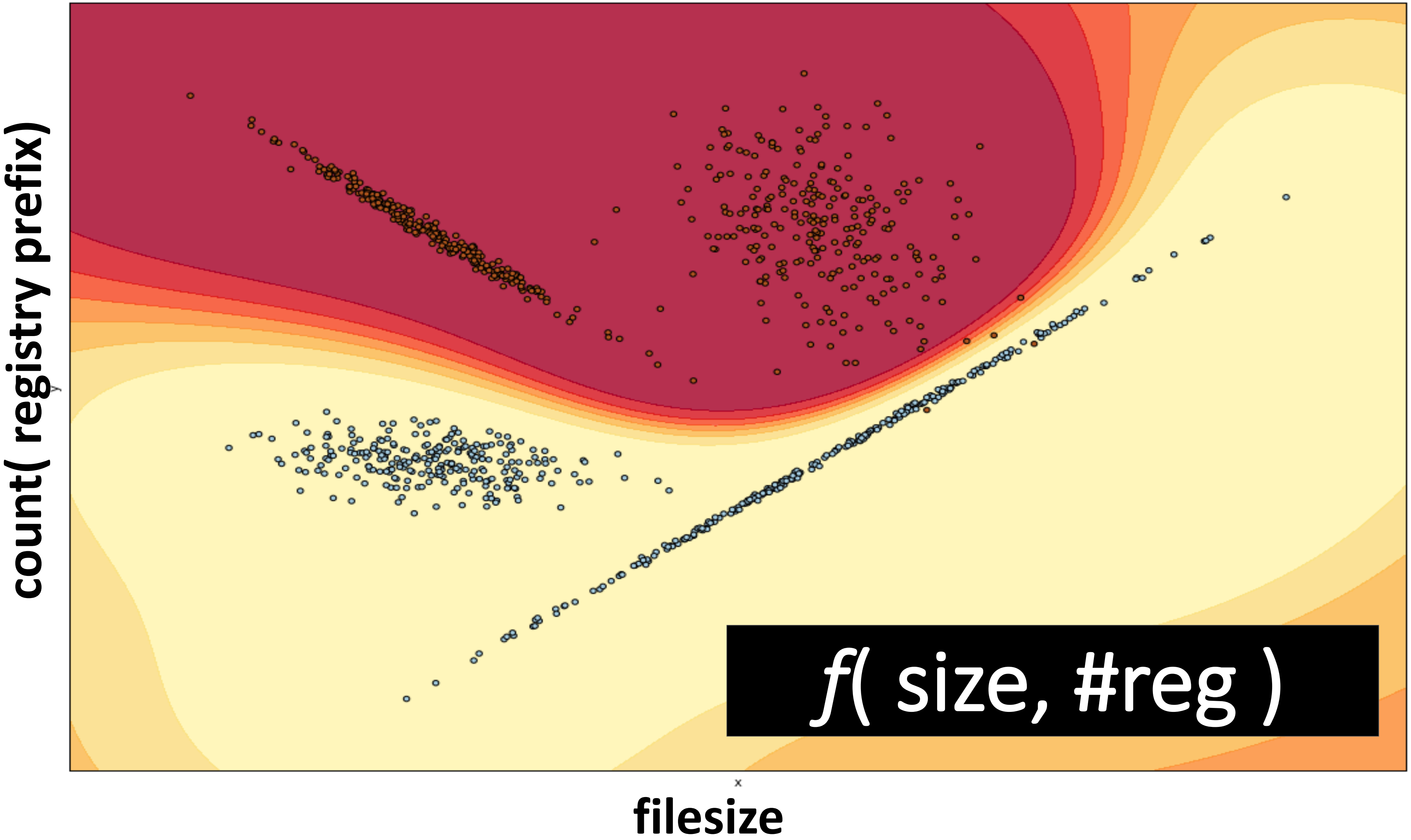


Why Machine Learning?

Automated

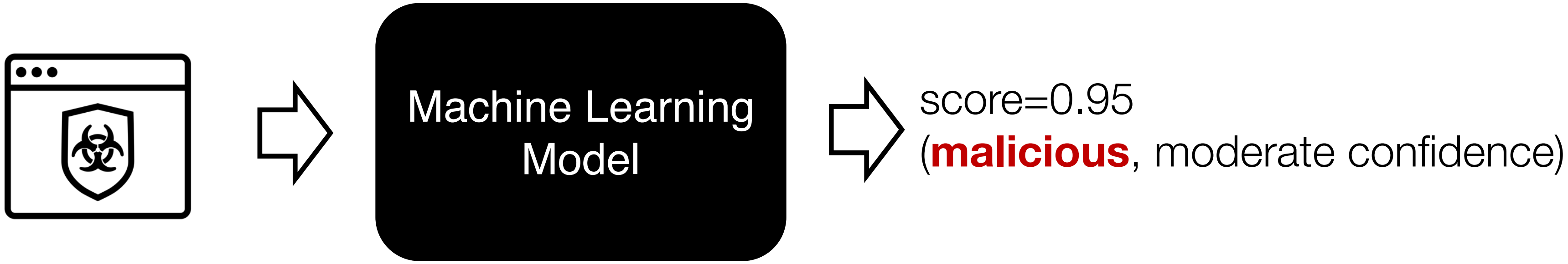
Sophisticated relationships

Generalizes

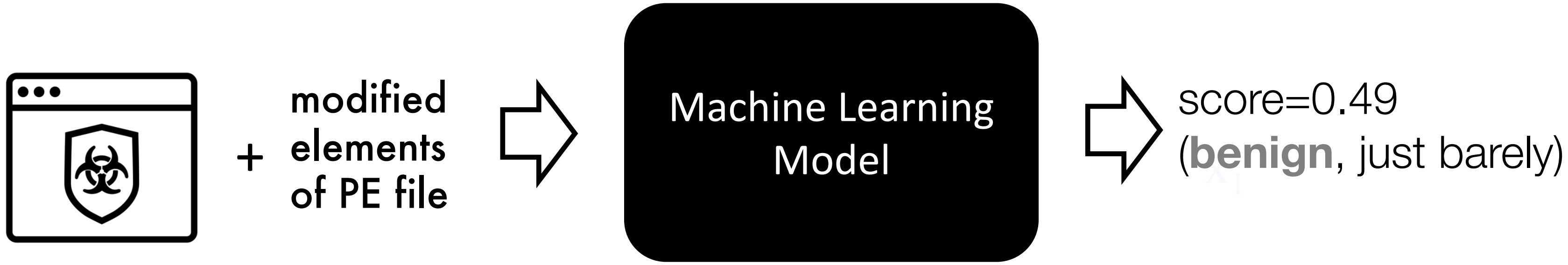


Goal: Can You Break Machine Learning?

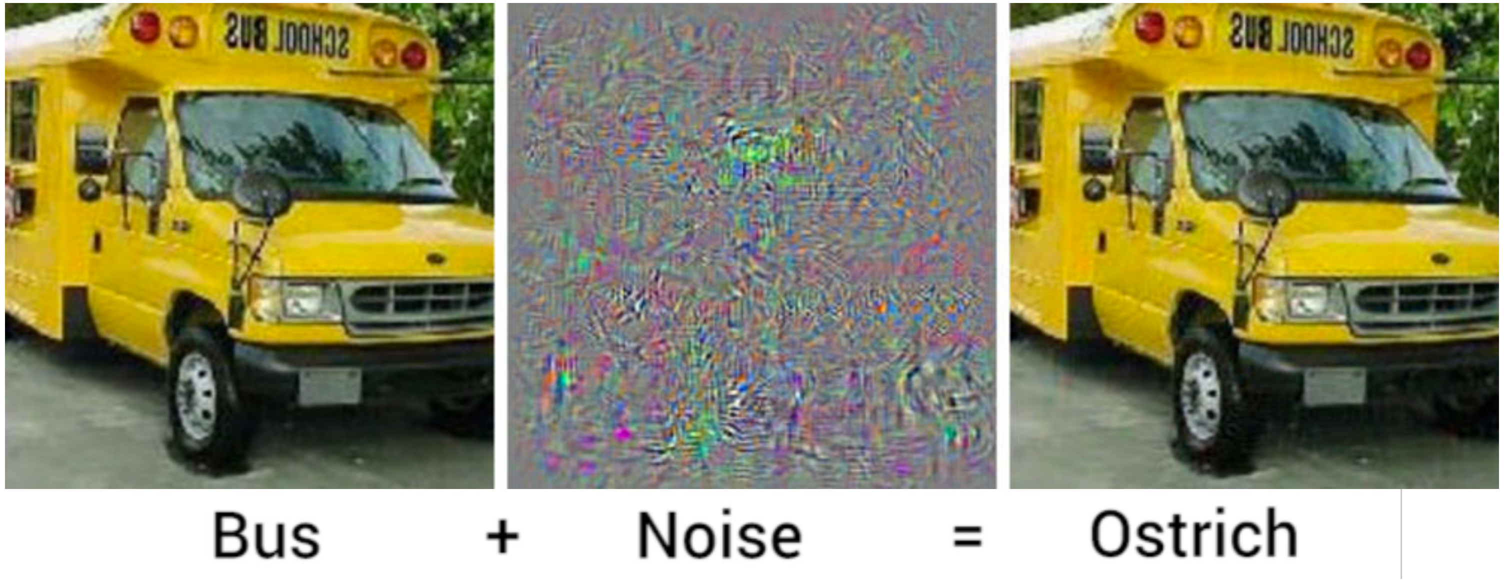
- Static machine learning model trained on millions of samples



- Simple structural changes that don’t change behavior
  - upx_unpack
  - ‘.text’ -> ‘.foo’ (remains valid entry point)
  - create ‘.text’ and populate with ‘.text from calc.exe’



Yes! And it can be automated!



- Machine learning models have blind spots / hallucinate (modeling error)
- Depending on model and level of access, they can be straightforward to exploit
- Adversarial examples can generalize **across** models / model types (Goodfellow 2015)
  - blind spots in MY model may also be blind spots in YOUR model

image credit: http://www.popsci.com/byzantine-science-deceiving-artificial-intelligence

Taxonomy of ML Attacks in infosec

adversary's knowledge about your model

An adversary...

...has your model

- known architecture + weights
- direct attack on model
- "easy" for deep learning
  - gradient perturbation
    [for Android malware]
    (Papernot et al. 2016)
  - dueling models / GAN
    [for DGA detection]
    (Anderson et al. 2016)

...can get a score

- black box...
- ...can arbitrarily probe and retrieve a **score / confidence**

EvadeML [for PDF malware]
(Xu, Qi, Evans, 2016)

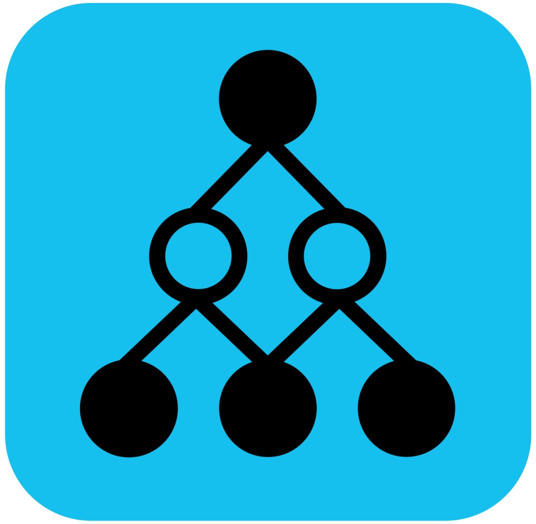
...can get good/bad

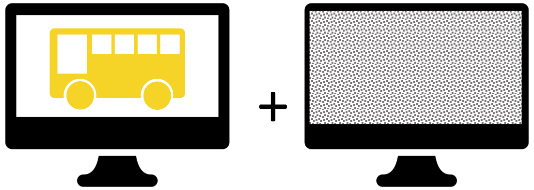
- black box...
- ...can arbitrarily probe for **malicious / benign**

difficulty for adversary to bypass

Related Work: full access to model

Bus (99%), Ostrich (1%)

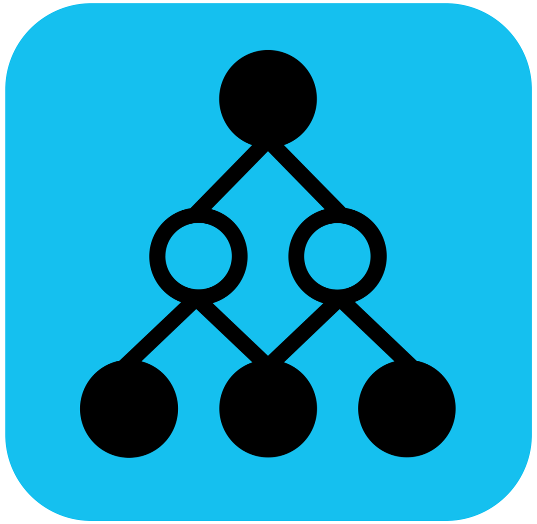




+ = image

BUT...

Malware (90%), Benign (10%)



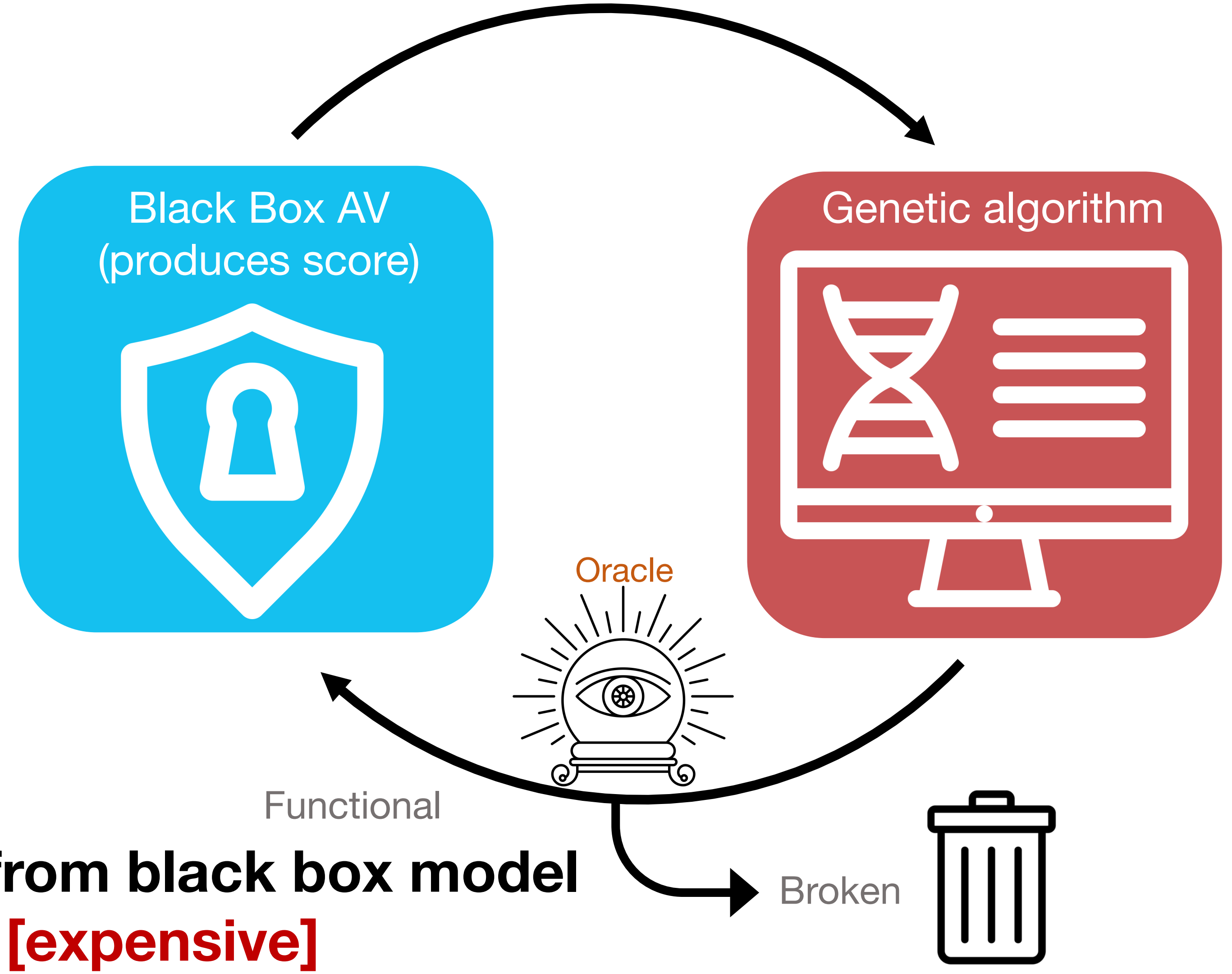
+ modified bytes or features !=

break PE format destroy function

Attacker requires full knowledge of model

Generated sample may not be valid PE file

Related Work: attack score-reporter



Requires score from black box model
Oracle/sandbox [expensive]

EvadeML [for PDF malware]
(Xu, Qi, Evans, 2016)

Summary of Previous Works

Gradient-based attacks: perturbation or GAN

- Attacker requires **full knowledge** of model structure and weights
- Generated sample may not be valid PE file

Genetic Algorithms

- Attacker requires **score** from black box model
- Requires oracle/sandbox [expensive] to ensure that functionality is preserved

Goal: Design a machine learning agent that

- bypass **black-box** machine learning using
- **format-** and **function-preserving** mutations

Reinforcement Learning!

Atari Breakout

Nolan Bushnell, Steve Wozniak, Steve Bristow

Inspired by Atari Pong

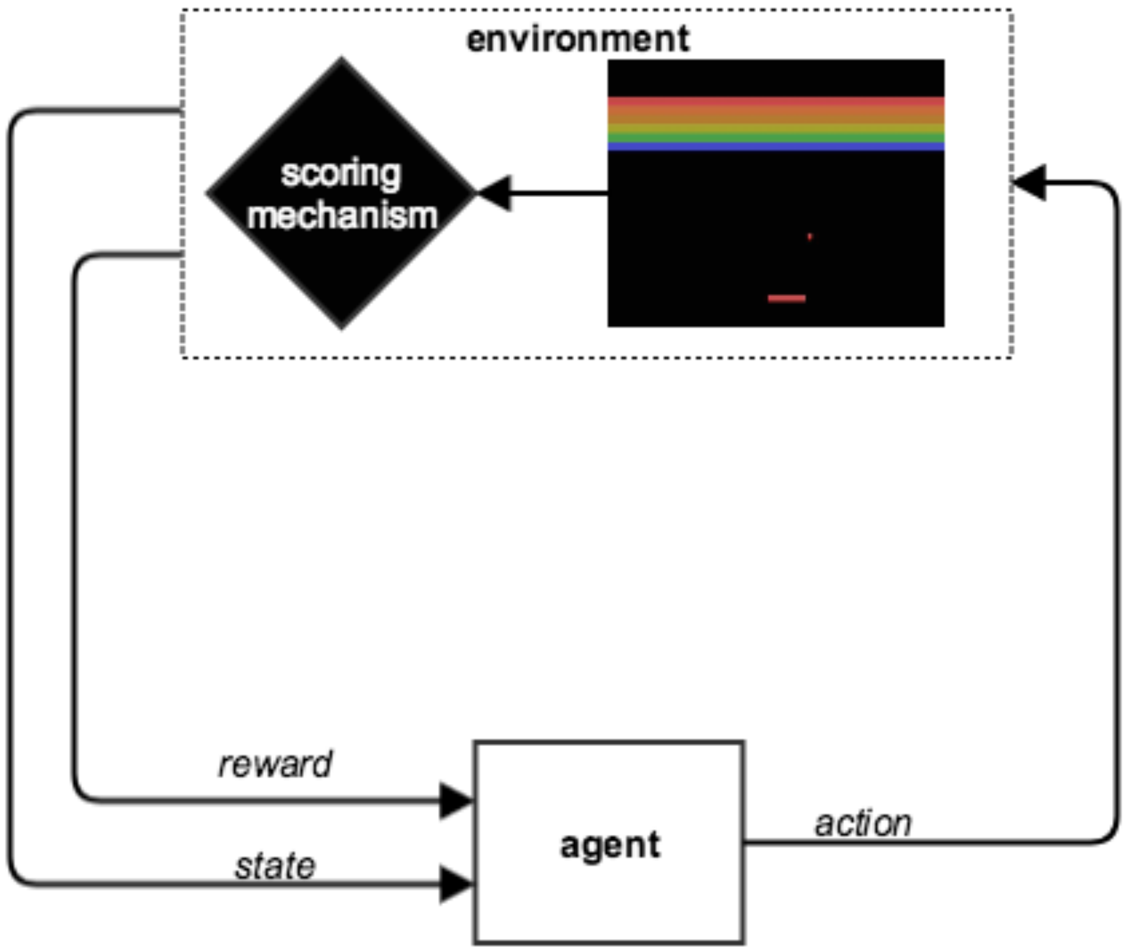
"A lot of features of the Apple II went in because I had designed Breakout for Atari"

(The Woz)

Game

- Bouncing ball + rows of bricks
- Manipulate paddle (left, right)
- Reward for eliminating each brick

Atari Breakout: an AI



- **Environment**
  - Bouncing ball + rows of bricks
  - Manipulate paddle (*left, right, nothing*)
  - Reward for eliminating each brick

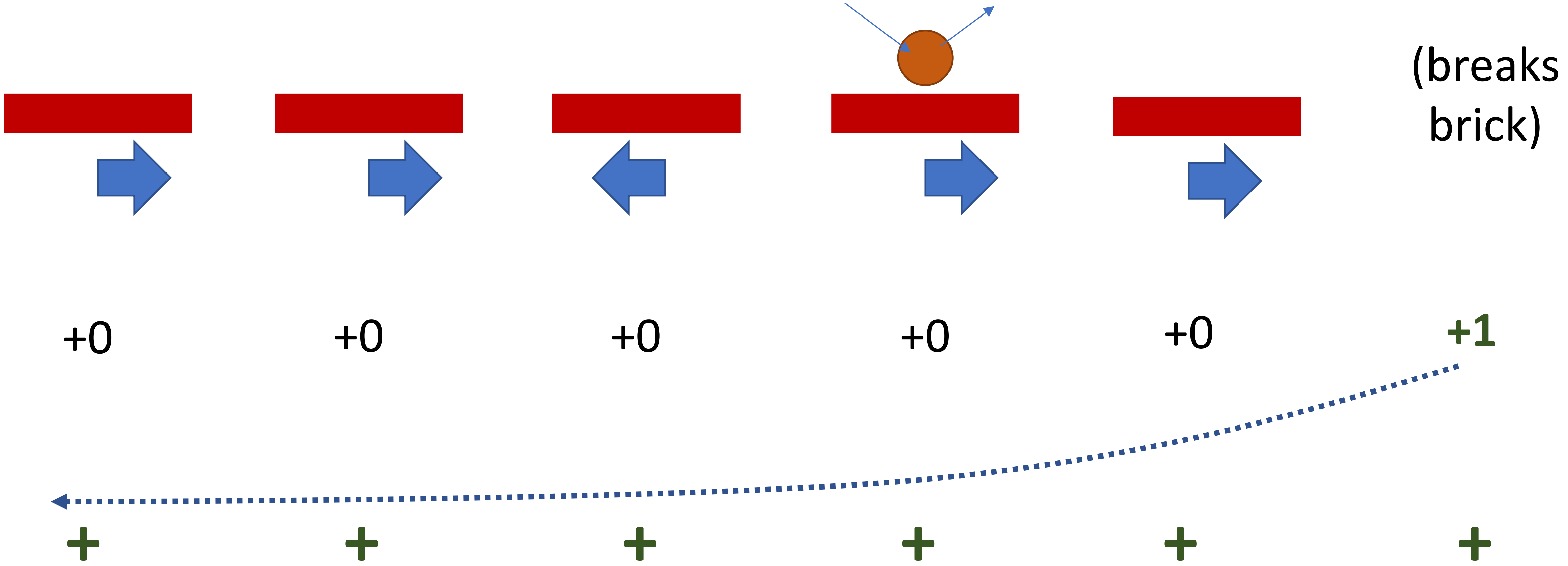
- **Agent**
  - Input: **environment state** (*pixels*)
  - Output: **action** (*left, right*) via **policy**
  - Feedback: delayed **reward** (*score*)

- Agent learns through 1000s of games:

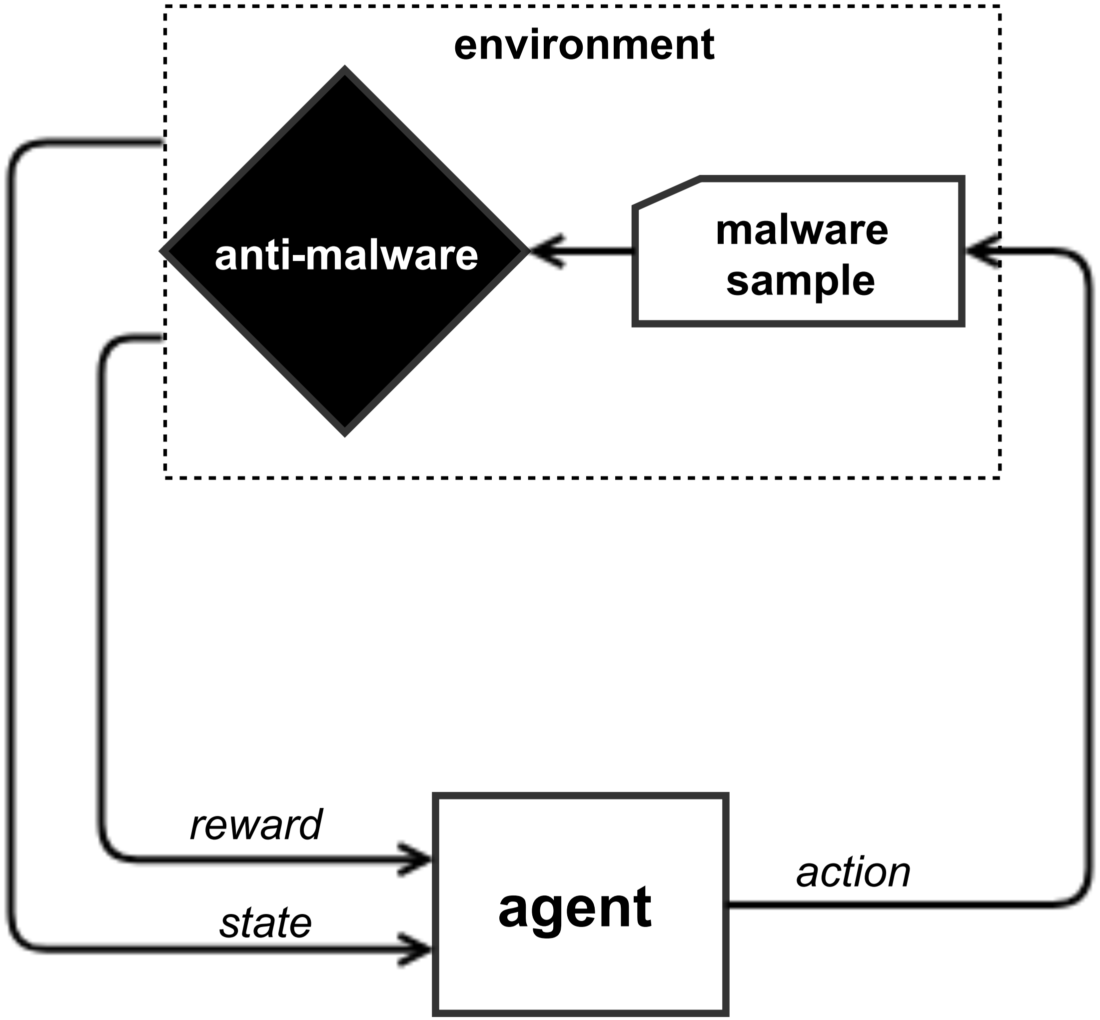
  what action is most useful given a screenshot of the Atari gameplay?

https://gym.openai.com/envs/Breakout-v0

Learning: rewards and credit assignment



Anti-malware evasion: an AI

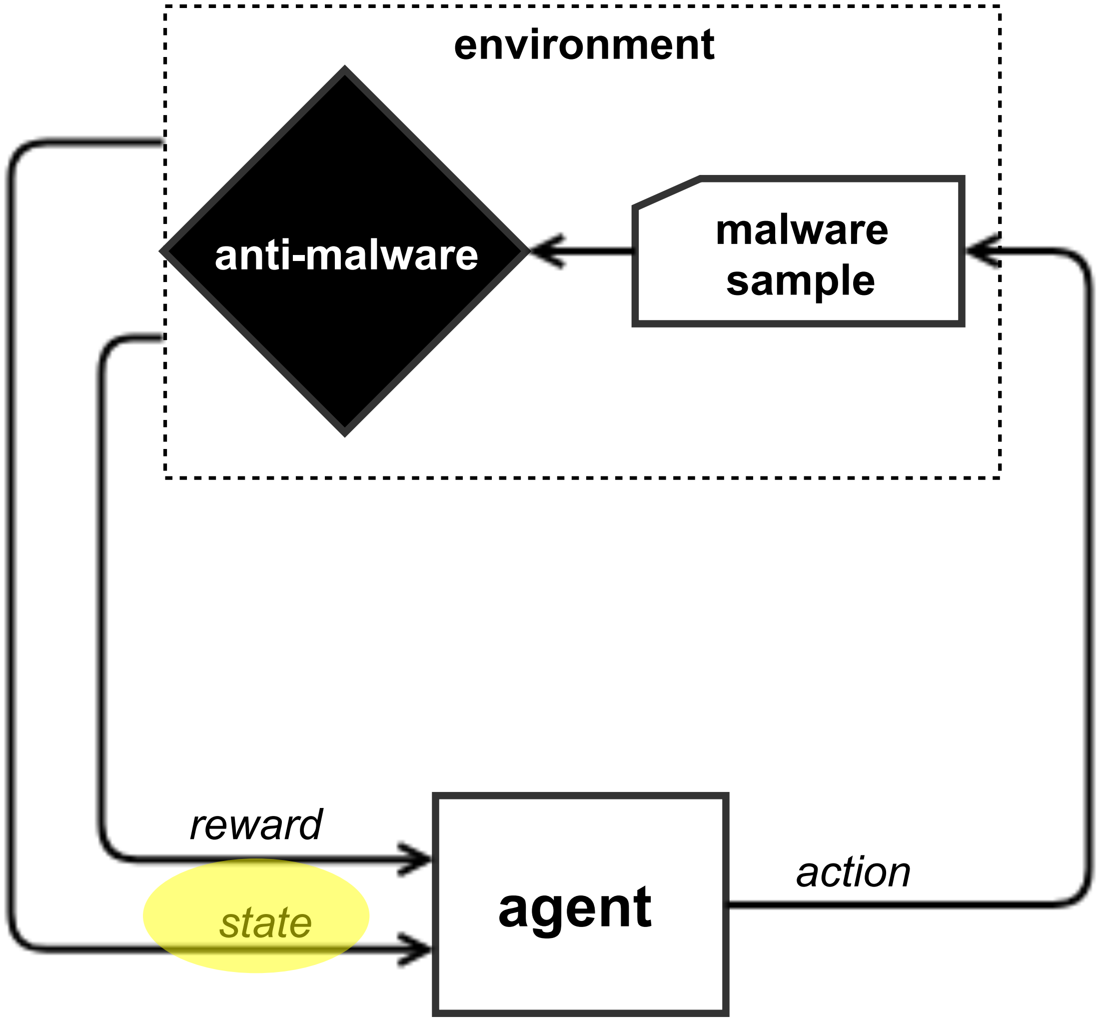


- **Environment**
  - A malware sample (*Windows PE*)
  - Buffet of malware mutations
    - *preserve format & functionality*
  - Reward from static malware classifier

- **Agent**
  - Input: **environment state** (*malware bytes*)
  - Output: **action** (*stochastic*)
  - Feedback: **reward** (*AV reports benign*)

https://github.com/endgameinc/gym-malware

The Agent's State Observation

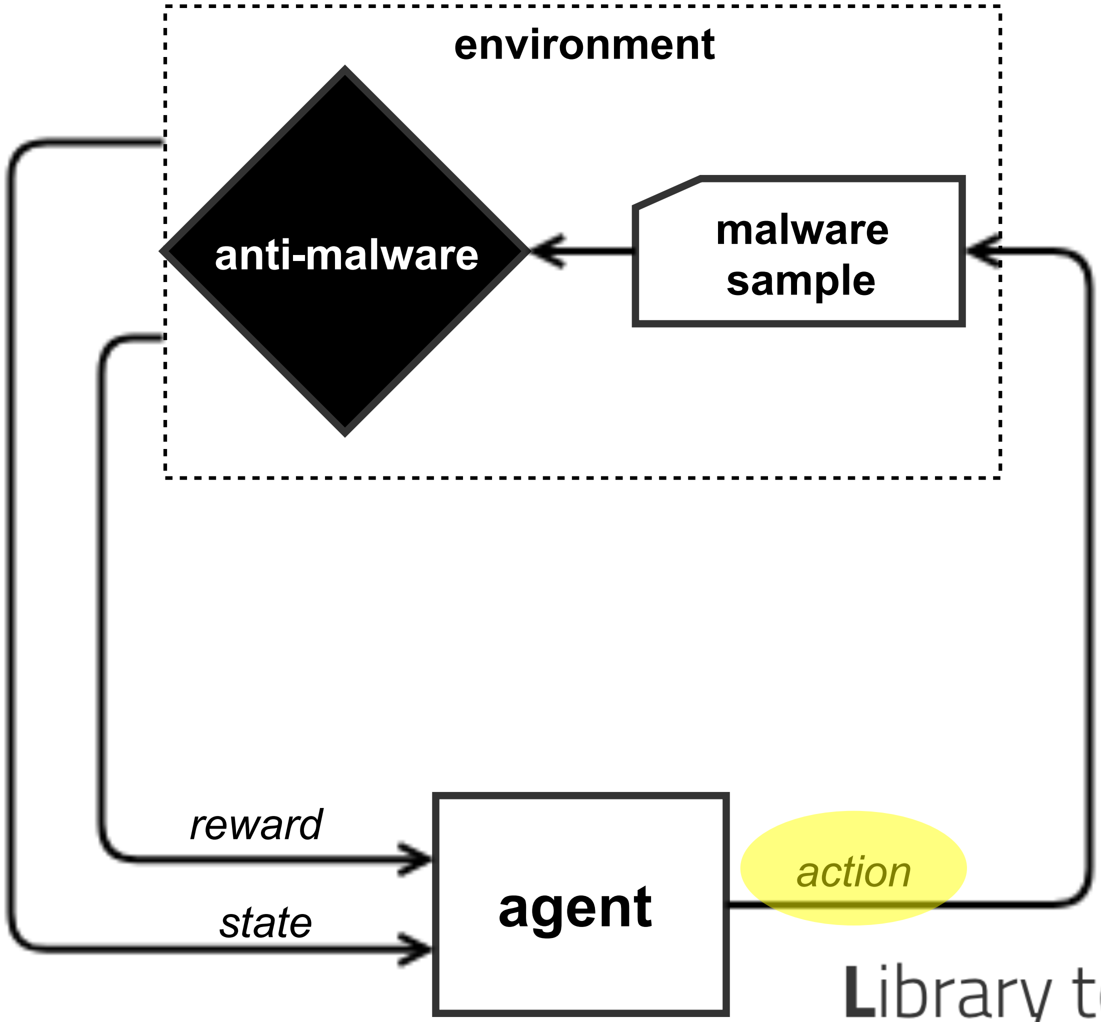


Features

- Static Windows PE file features compressed to 2350 dimensions
  - General file information (size)
  - Header info
  - Section characteristics
  - Imported/exported functions
  - Strings
  - File byte and entropy histograms

- Feed a neural network to choose the best action for the given "state"

The Agent's Manipulation Arsenal

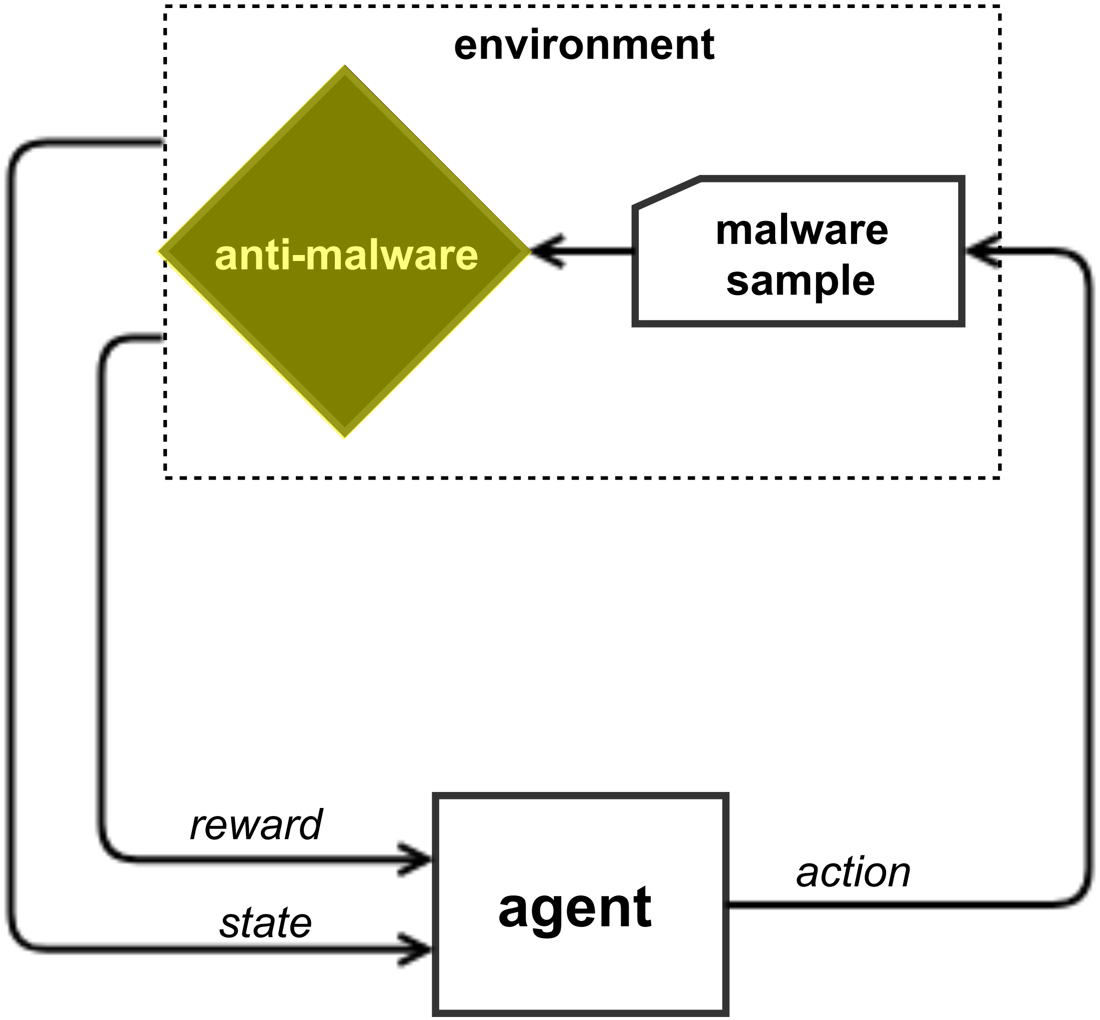


Functionality-preserving mutations:

- **Create**
  - New Entry Point (w/ trampoline)
  - New Sections
- **Add**
  - Random Imports
  - Random bytes to PE overlay
  - Bytes to end of section
- **Modify**
  - Random sections to common name
  - (break) signature
  - Debug info
  - UPX pack / unpack
  - Header checksum
  - Signature

Quarkslab
Library to **I**nstrument **E**xecutable **F**ormats

The Machine Learning Model



Static PE malware classifier

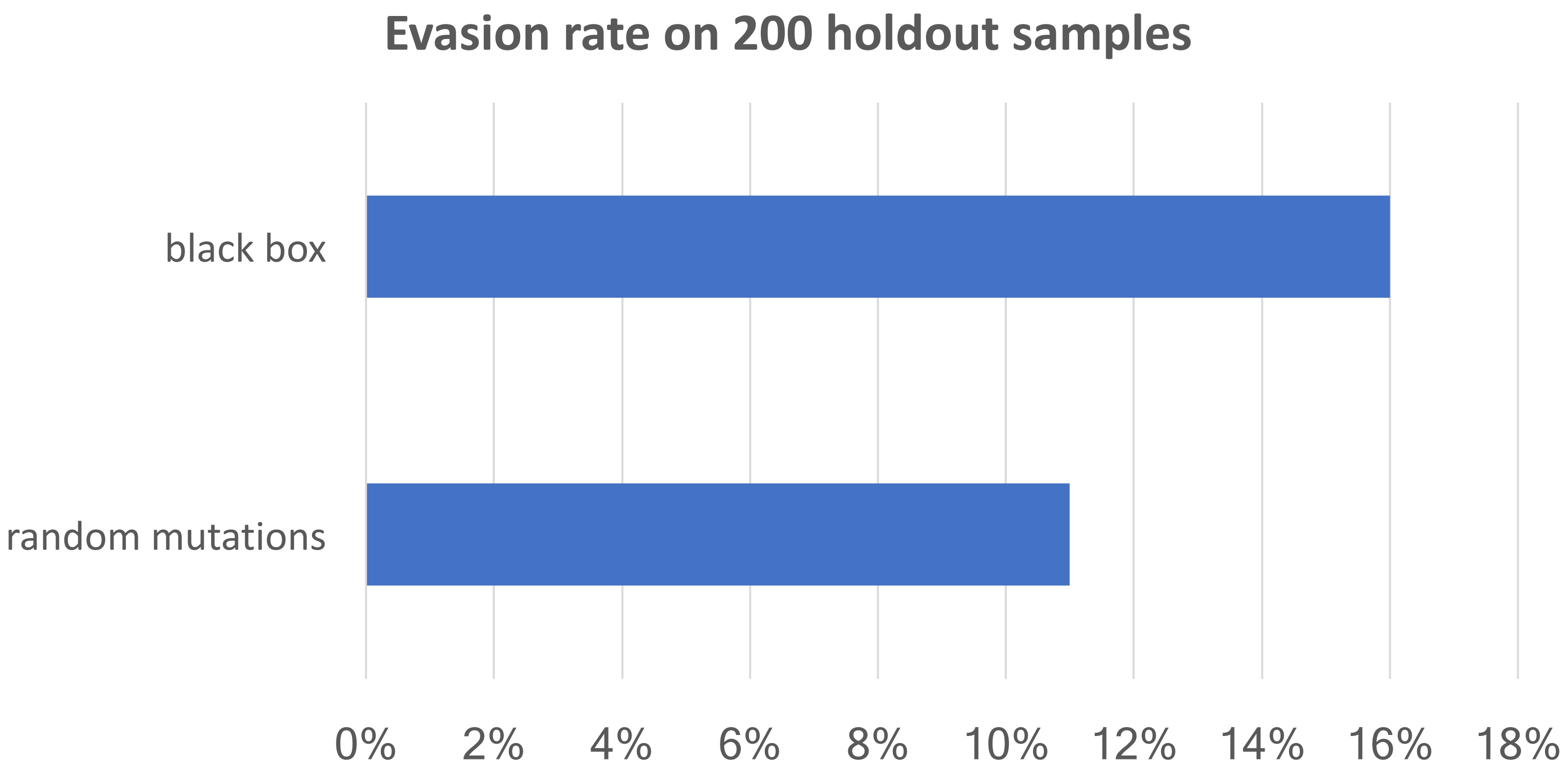
- gradient boosted decision tree (for which one *can't directly do gradient-based attack)*
- need not be known to the attacker

Machine learning malware model for demo purposes only. Resemblance to Endgame or other vendor models is incidental.

Ready, Fight!

Evasion Results

- Agent training: 15 hours for 100K trials (~10K games x 10 turns ea.)
- Using malware samples from VirusShare



Cross-evasion: detection rate on VirusTotal (average)

- from 35/62 (original)
- to 25/62 (evade)

Model Hardening Strategies

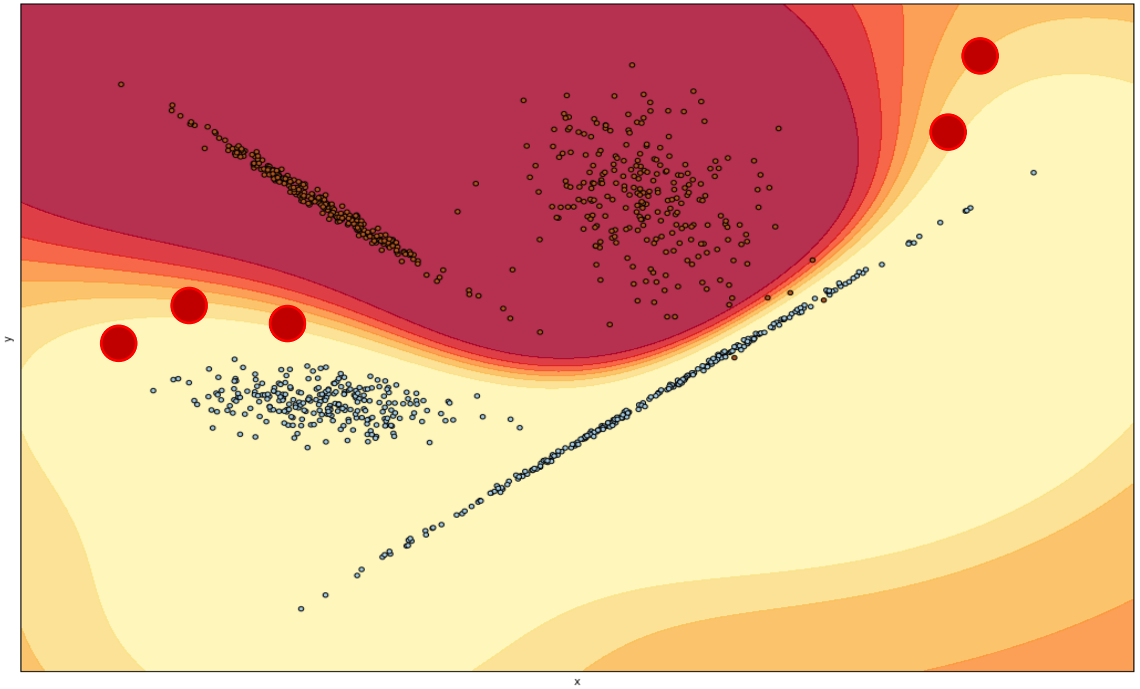
Feedback to the human

category	evasion %	dominant action sequence
ransomware	12%	upx_unpack -> overlay_append -> section_rename
virut	5%	upx_unpack -> section_add -> imports_append

Adversarial training

(train with new evasive variants)

Ransomware evasion drops from 12% to 8%



Big Picture

- Attack your own model to discover and fix blind spots!
- Limit isolated exposure of your model
- Limit exposing a score (easier attacks possible)

https://github.com/endgameinc/gym-malware

Thank you!

Hyrum Anderson
Technical Director of Data Science

hyrum@endgame.com @drhyrum /in/hyrumanderson

Co-contributors:

Anant Kharkar, University of Virginia
Bobby Filar, Endgame
Phil Roth, Endgame

ENDGAME.