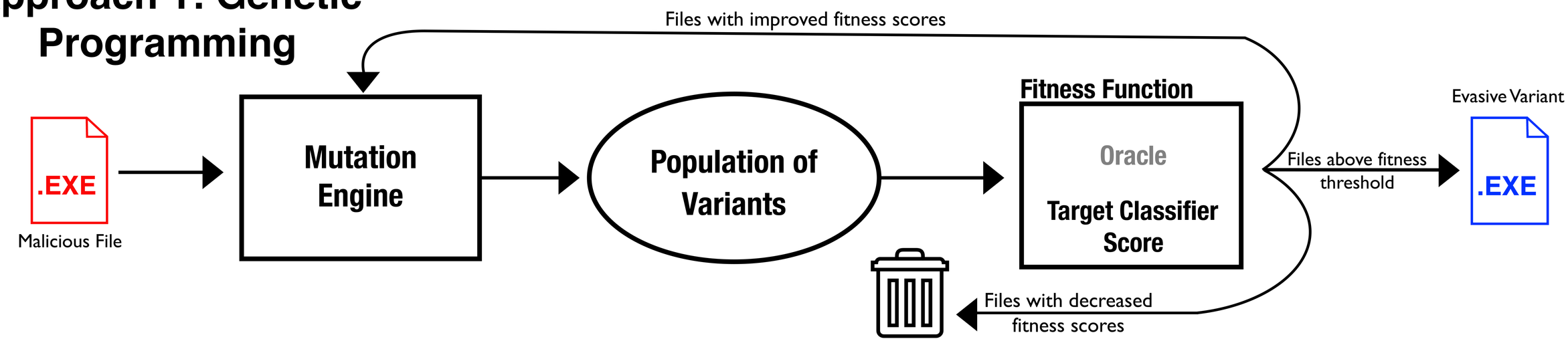
Approaches to Evading Windows PE Malware Classifiers

Anant Kharkar, Helen Simecek, Weilin Xu, David Evans, *University of Virginia*
Hyrum S. Anderson, *Endgame*

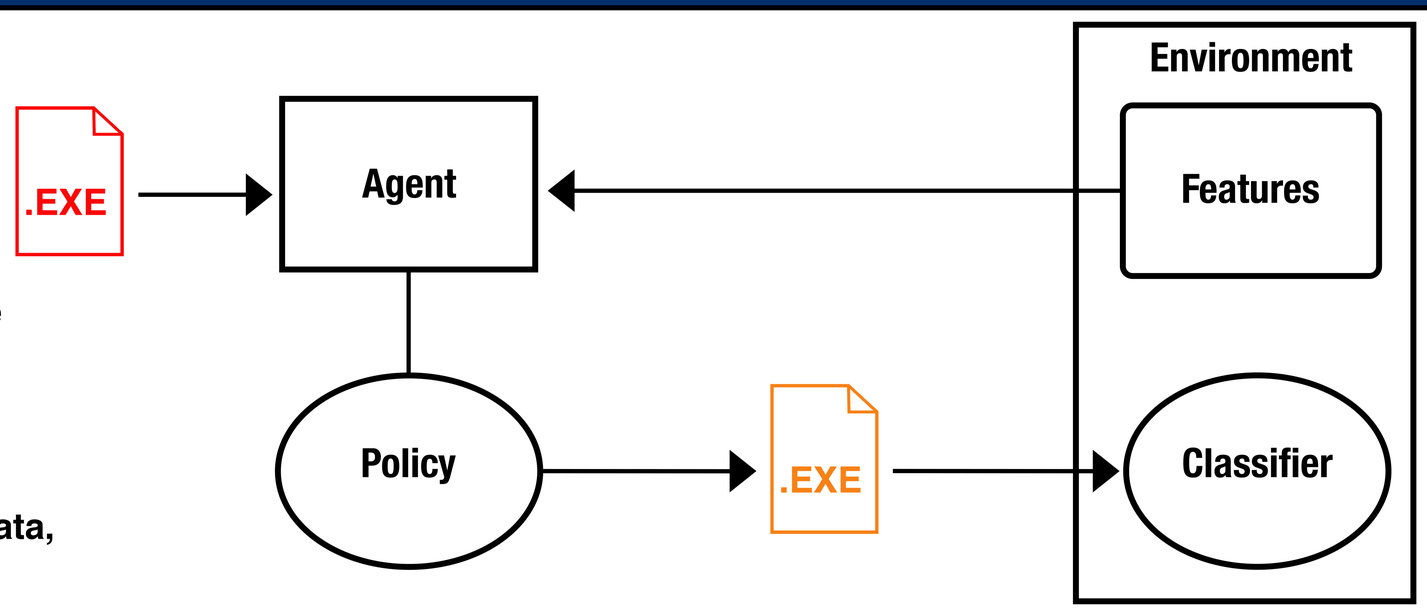
Research Objective: **While machine learning is proving itself to be a valuable tool for malware classification, its security guarantees remain an open question. We are exploring techniques to manipulate malware to appear harmless to machine learning classifiers, yet retain its malicious behavior (an evasion attack). We apply two heuristic search strategies to Windows PE malware.**

Approach 1: Genetic Programming



Approach 2: Reinforcement Learning

- Agent acting in its environment
  - Agent – malware mutator
  - Environment – set of features describing malware
- Actor-critic model w/experience replay (ACER)
- Agent learns mutation policy
  - Optimal mutations given prior sequence of states
- Environment features
  - Byte entropy, PE header metadata, section metadata, Import/Export table metadata



Preliminary Experiments

Target Classifier

Endgame's MalwareScore™ *

Samples

10,000 Windows PE malware samples from VirusShare

File Manipulations

We use a set of 10 file manipulations developed to preserve the malware sample's malicious behavior. If we can verify that all manipulations are behavior preserving, then we can guarantee that all variants remain malicious without the need for an oracle. Our current implementation of these mutations does not provide strict guarantees of malice preservation.

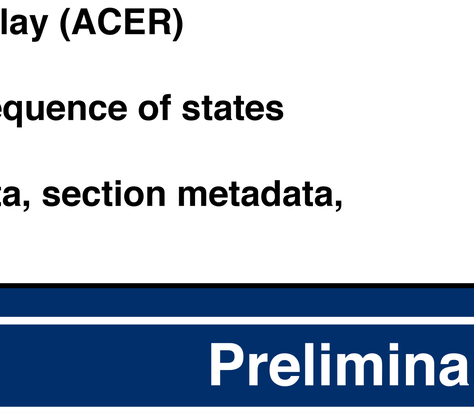
Set of Manipulations

create_new_entry	section_rename
section_add	section_append
upx_pack	imports_append
upx_unpack	overlay_append
remove_signature	
break_optional_header_checksum	

GP Current Status

Developing a genetic programming framework using the current implementations of the file manipulations to mutate malware samples.

An experiment on 130 random samples found:
76.9% achieved a 'benign' classifier score
23.1% did not reach this threshold in 20 generations



Classifier outputs a real number between 0.0 (most malicious) to 1.0 (most benign)

Next Steps

- Verify that all mutations retain the file's malicious behavior
- Improve success rates:
  - Fitness function with more features than classifier score
  - A more rigorous variant selection process
  - Adjust mutations and variants per generation
  - Increase probability of more effective mutations occurring

RL Current Status

- RL advantageous in the most difficult attack scenario
  - Target classifier returns only boolean (benign/malicious)
- Continuous score provides modest increase in performance
- RL outperformed by GP when more information is available

Next Steps

- Verify malicious behavior of variants
- Improve the performance of the RL agent
  - Improved features / feature engineering
  - Improved model architectures
  - Expanded mutation set

*Core machine learning model that drives a recent May 2017 version of Endgame's MalwareScore™