#BHASIA @BlackHatEvents

CalMal

Malware clustering using unsupervised Machine Learning

https://github.com/unknownhad/CalMal

By : Himanshu Anand

c/side

Introduction

- whoami
  - Himanshu Anand
  - Security junkie who does some CTF with Water Paddler and BlueWater

(Mal|Soft)ware

The Problem

- challenge of malware proliferation
  - Many Different Variants Present:
  - Packers/Crypters Usage
  - Too Many New Malwares Found Every Day (https://www.av-test.org/en/statistics/malware/)
- Distinguishing between malware and benign software
- Too many new malwares popping everyday

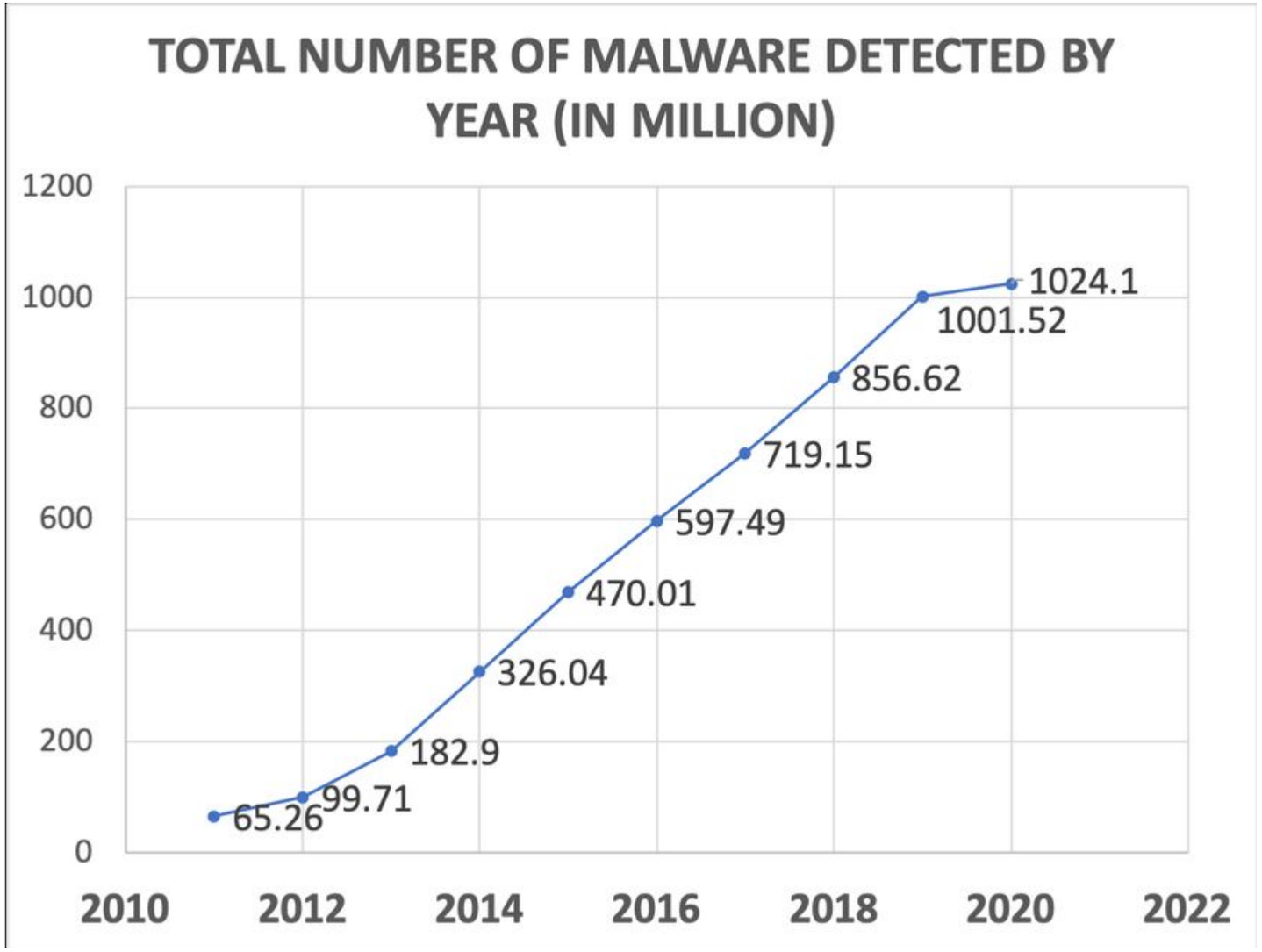


Image from: https://www.researchgate.net/figure/Total-number-of-malware-detected-by-year-in-million-47_fig2_339301928

Blue Team Weapon	Red Team Arsenal
CalMal	Packers Cryptors Compressors Anti Emulation Anti Debugging

How Do we analyze malware effectively?

CalMal Solution

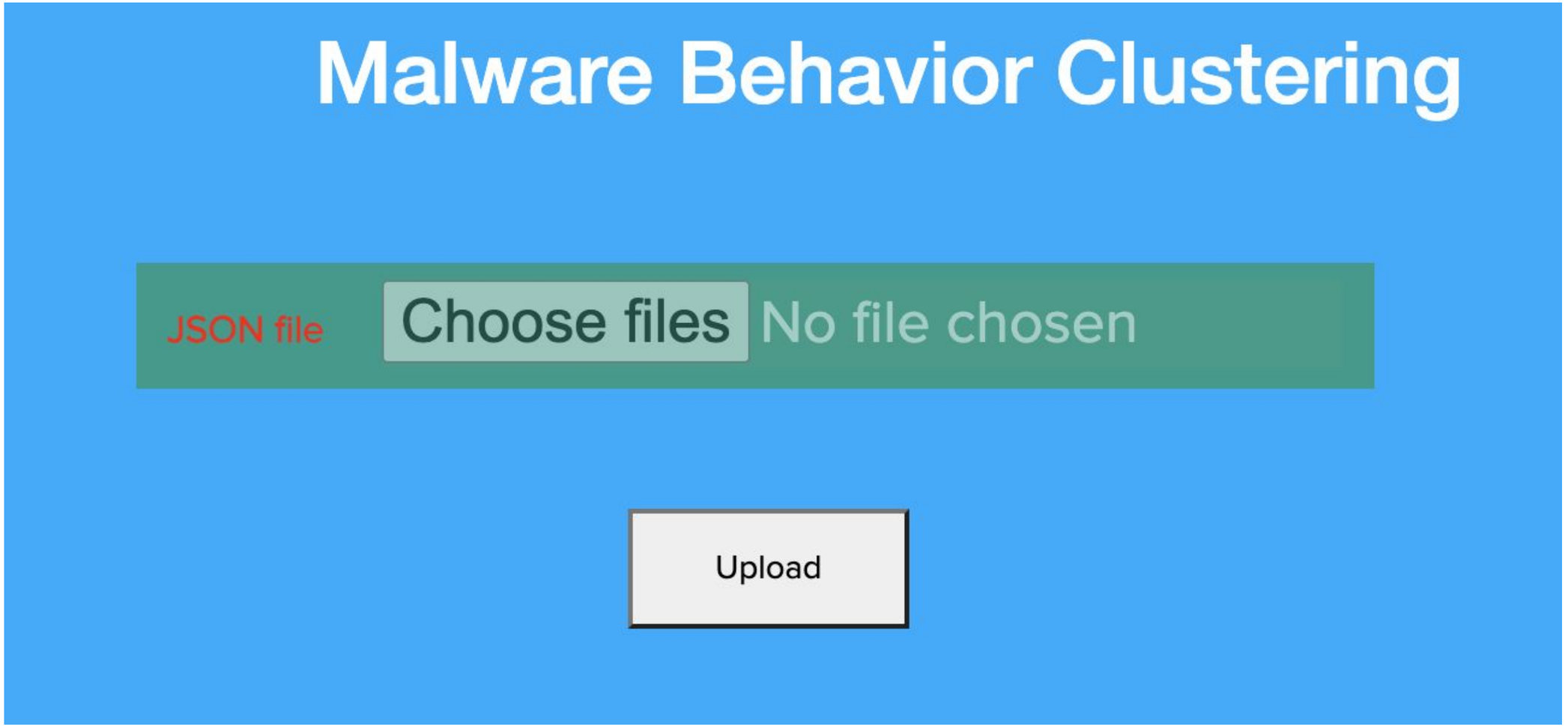
- CalMal’s goal to cluster malware using unsupervised learning, emphasizing its significance in cybersecurity

Why Clustering?

- the choice of unsupervised learning for malware clustering, focusing on its ability to handle the vast number of malware families and variants


HACKERMAN

CalMal Setup



Data Processing Workflow

- Preprocessing JSON Data
- Extracting and Cleaning Unigrams
- Selecting Top Unigrams
- Creating Bit Strings
- Saving the Results

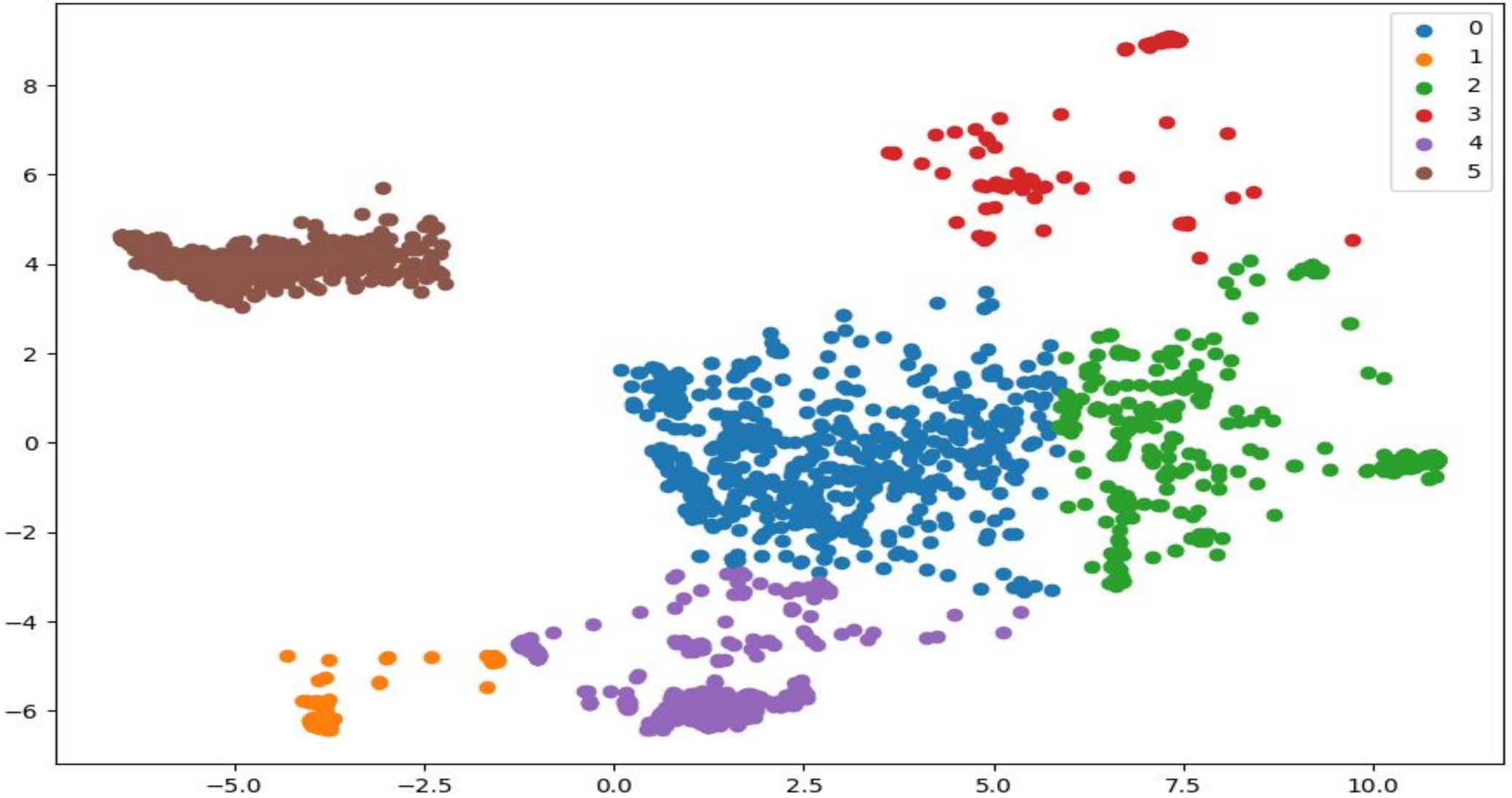
Data Encoder

- Data Preparation
- Data Encoding Neural Network
- Training Preparation
- Checkpointing
- Noise Injection
- Training Loop
- Saving the Model

Model Training and Clustering

- Clustering of malwares
- Dataset Preparation
- Neural Network Architecture
- Training Process
- Evaluation and Checkpointing
- Final Model Save and Evaluation
- Label Export

CalMal Web Service Demonstration


0
1
2
3
4
5
8
6
4
2
0
−2
−4
−6
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

Contact me:

- me@himanshuanand.com
- https://himanshuanand.com
- https://twitter.com/anand_himanshu
- https://bsky.app/profile/noob.bsky.social
- https://infosec.exchange/@N00b
- https://github.com/unknownhad