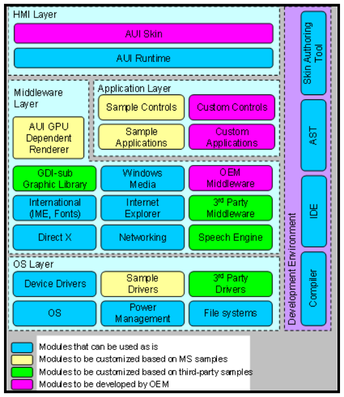
Learning from Big Malwares

Linhai Song, Heqing Huang, Wu Zhou,
Wenfei Wu, and Yiying Zhang

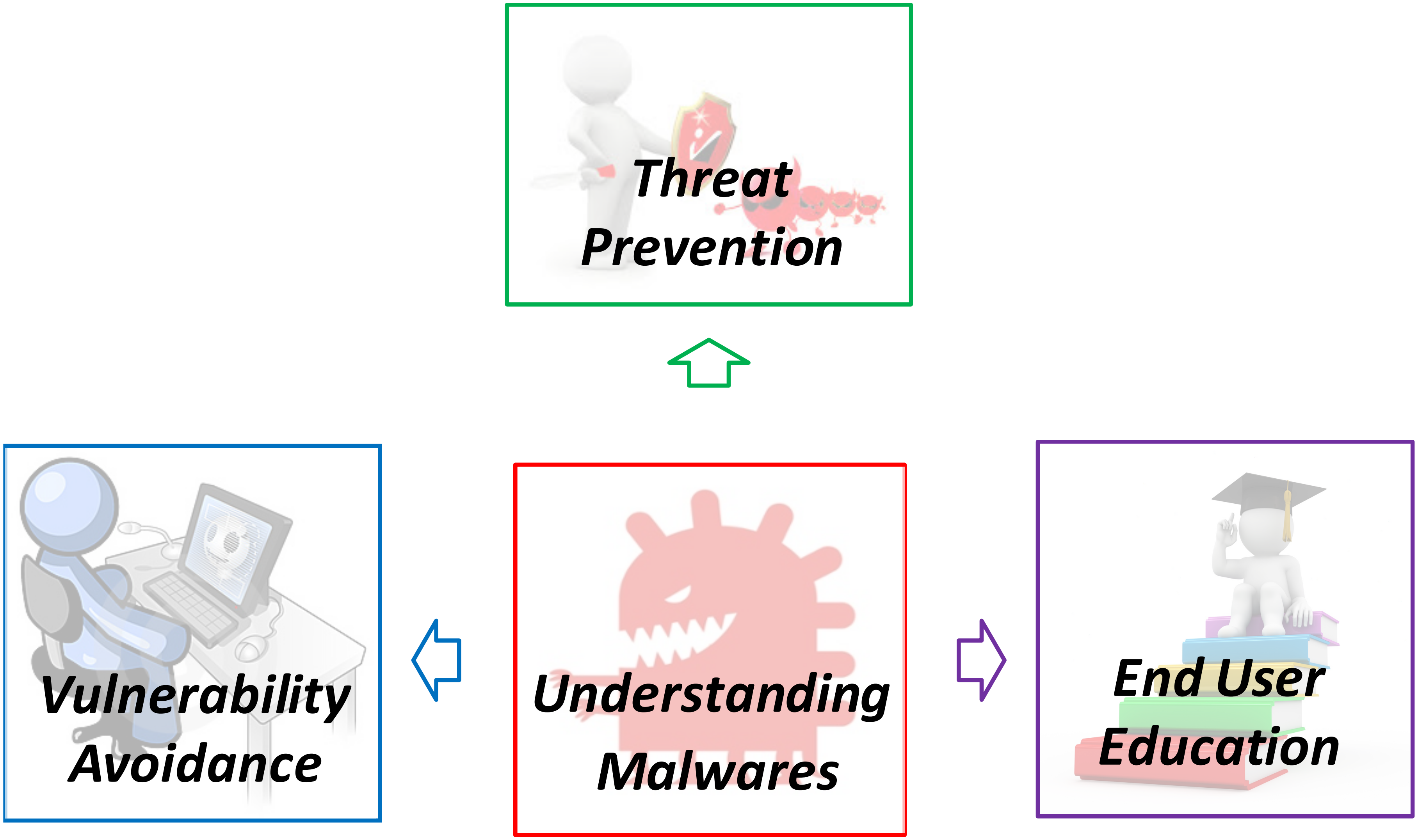
WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Combating Malwares is Critical

- Definition of malwares
  - A variety of hostile or intrusive software
- Malwares are common and severe
  - 140 million new malwares appeared in 2015
  - 2 millions attempts to steal money via online bank
- Fighting malwares is increasing important



How to Fight Malwares?



Threat Prevention

Vulnerability Avoidance

Understanding Malwares

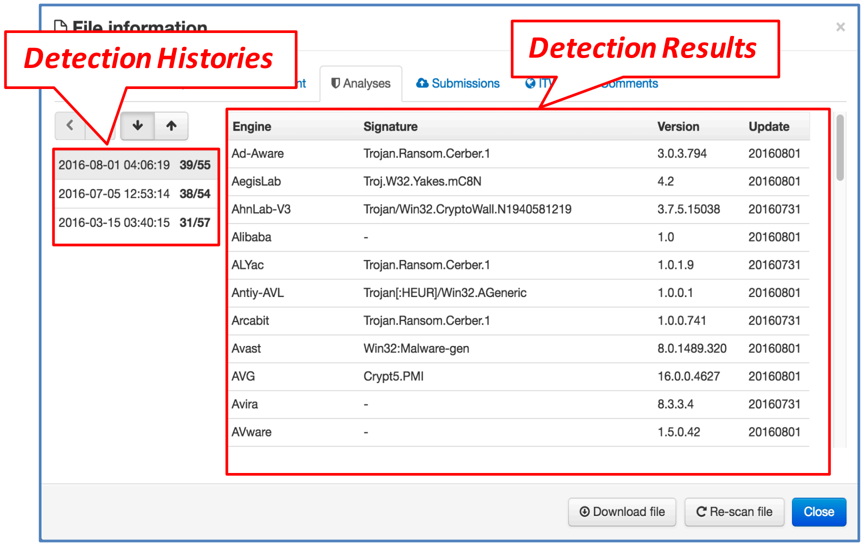
End User Education

Why Studying Big Malwares?

- Previous works on studying malwares
  - Provide invaluable insights
  - Only on a limited amount of malwares
- Studying big malwares
  - “Big”: in large scale and with high diversity
  - Exposes new insights

VirusTotal (VT)

- An online service to analyze suspicious files
  - Containing a huge amount of real-world files
    - 43 million suspicious files submitted last Nov.
  - Applying a host of latest anti-virus engines
  - Providing rich metadata

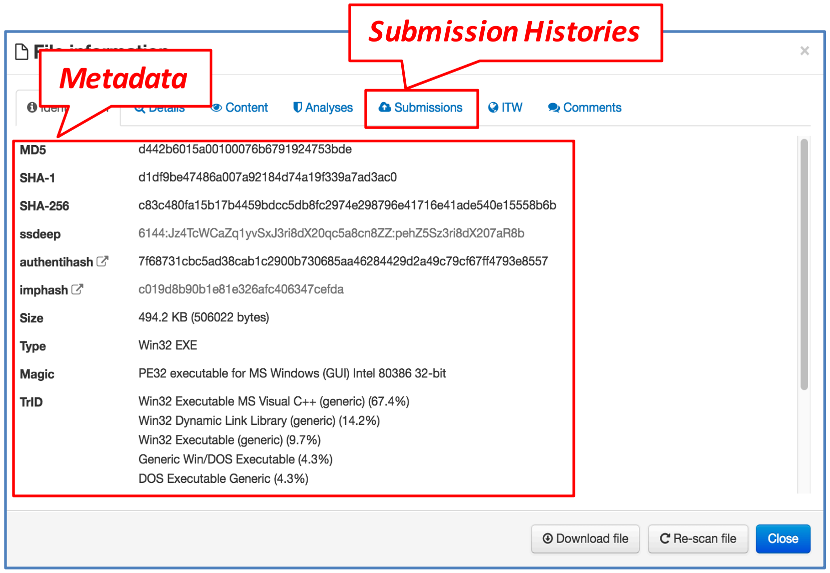


Detection Histories

Detection Results

2016-08-01 04:06:19	**39/55**
2016-07-05 12:53:14	**38/54**
2016-03-15 03:40:15	**31/57**

Engine	Signature	Version	Update
Ad-Aware	Trojan.Ransom.Cerber.1	3.0.3.794	20160801
AegisLab	Troj.W32.Yakes.mC8N	4.2	20160801
AhnLab-V3	Trojan/Win32.CryptoWall.N1940581219	3.7.5.15038	20160731
Alibaba	-	1.0	20160801
ALYac	Trojan.Ransom.Cerber.1	1.0.1.9	20160731
Antiy-AVL	Trojan[:HEUR]/Win32.AGeneric	1.0.0.1	20160801
Arcabit	Trojan.Ransom.Cerber.1	1.0.0.741	20160731
Avast	Win32:Malware-gen	8.0.1489.320	20160801
AVG	Crypt5.PMI	16.0.0.4627	20160801
Avira	-	8.3.3.4	20160731
AVware	-	1.5.0.42	20160801



Submission Histories

Metadata

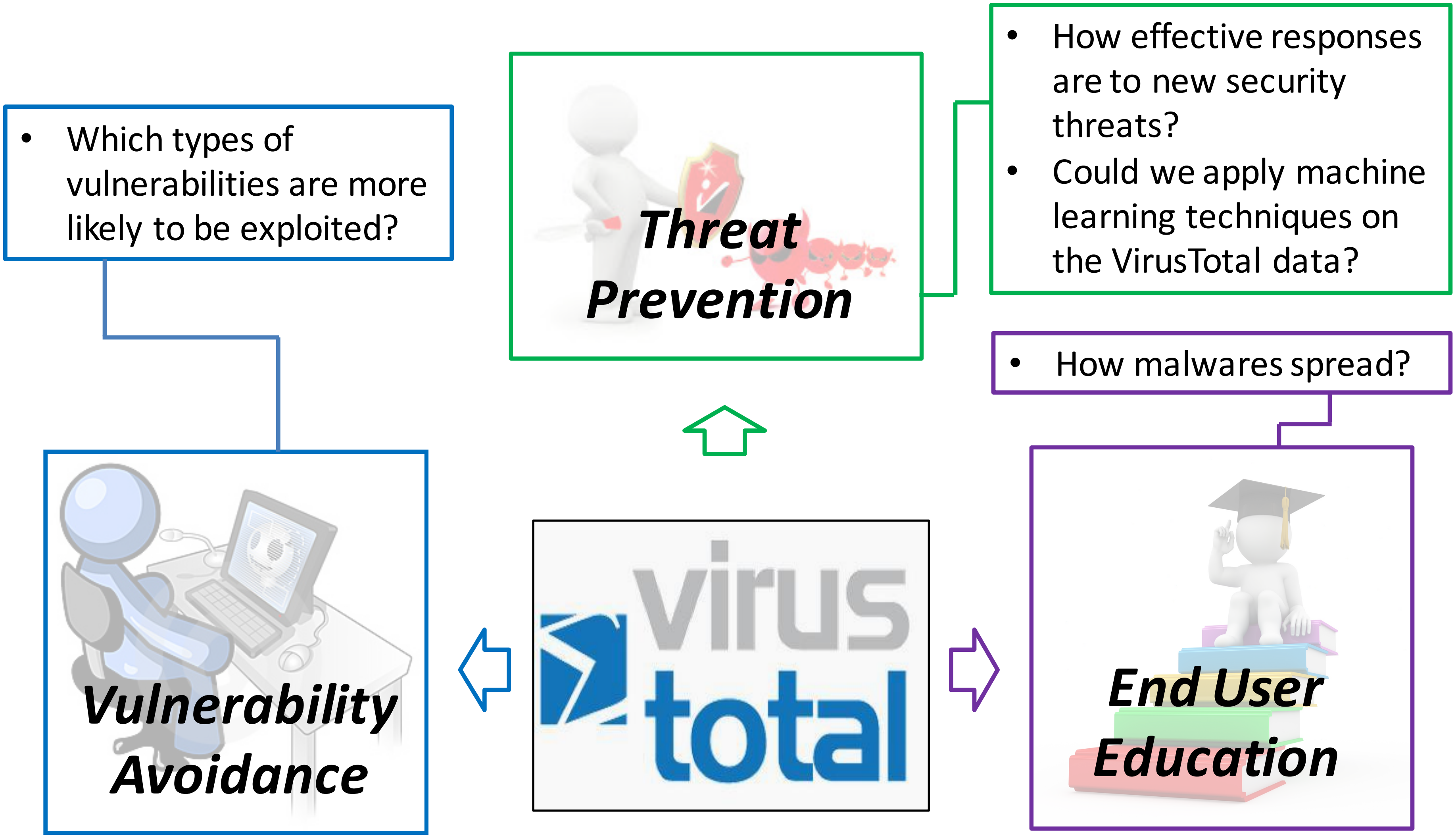
MD5	d442b6015a00100076b6791924753bde
SHA-1	d1df9be47486a007a92184d74a19f339a7ad3ac0
SHA-256	c83c480fa15b17b4459bdcc5db8fc2974e298796e41716e41ade540e15558b6b
ssdeep	6144:Jz4TcWCaZq1yvSxJ3ri8dX20qc5a8cn8ZZ:pehZ5Sz3ri8dX207aR8b
authentihash	7f68731cbc5ad38cab1c2900b730685aa46284429d2a49c79cf67ff4793e8557
imphash	c019d8b90b1e81e326afc406347cefda
Size	494.2 KB (506022 bytes)
Type	Win32 EXE
Magic	PE32 executable for MS Windows (GUI) Intel 80386 32-bit
TrID	Win32 Executable MS Visual C++ (generic) (67.4%) Win32 Dynamic Link Library (generic) (14.2%) Win32 Executable (generic) (9.7%) Generic Win/DOS Executable (4.3%) DOS Executable Generic (4.3%)

Existing Usage of VirusTotal

- Anti-virus vendors in industry
  - Identify FPs and FNs in their products
  - Fail to consider correlations

- Researchers in academia
  - Identifying users using VT as a test platform
  - Very few other works

Research Opportunities



- Which types of vulnerabilities are more likely to be exploited?

Threat Prevention

- How effective responses are to new security threats?
- Could we apply machine learning techniques on the VirusTotal data?

- How malwares spread?

Vulnerability Avoidance

virus total

End User Education

Contributions

- An early-stage empirical study on VT data
  - Temporal analysis
    - Submission frequency and family generation rate
    - Burstiness of malwares
  - Distribution study
    - Skewness of malware families
    - Identifying hot malware families
- Identifying key research opportunities from VT

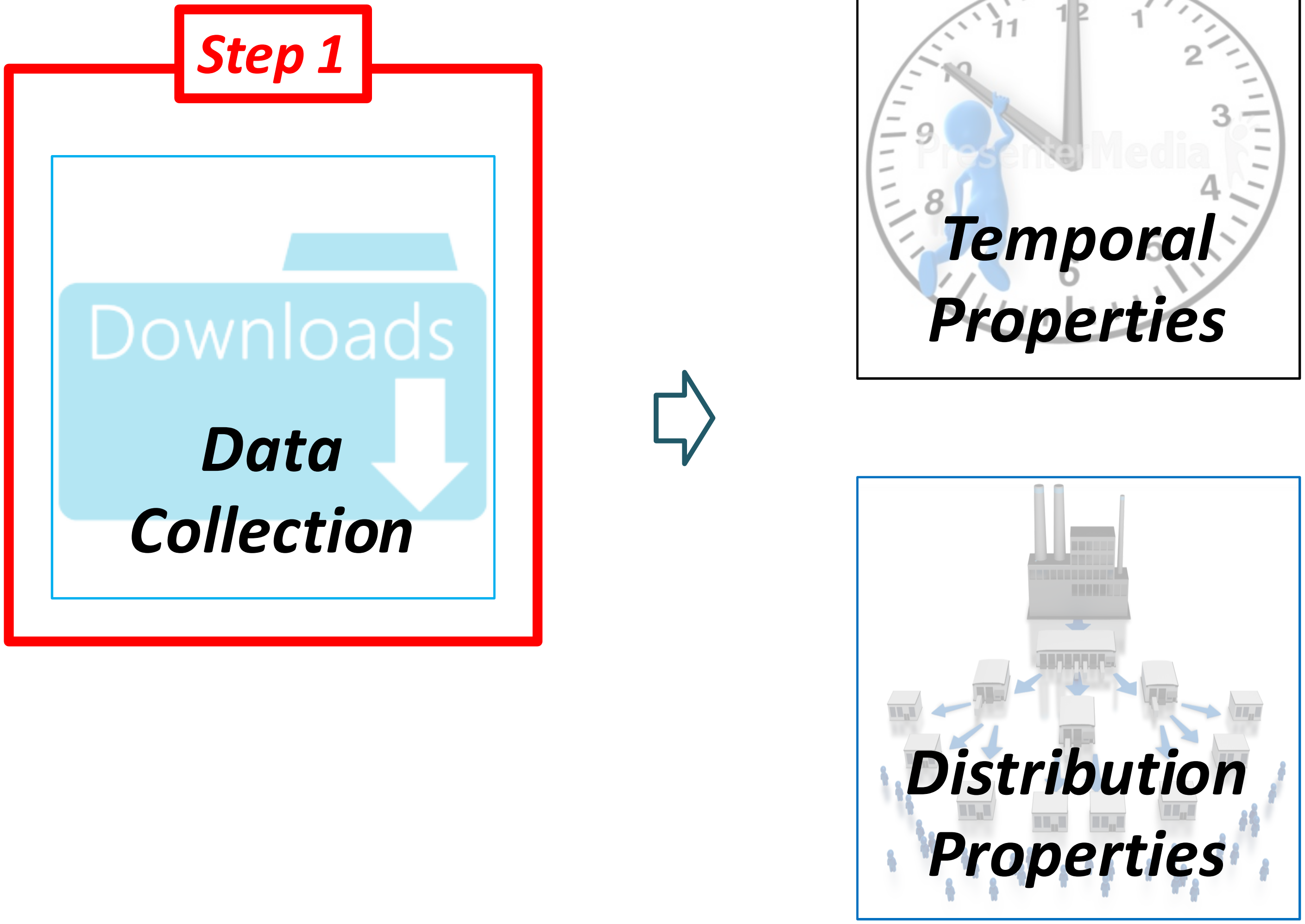
Outline

- Introduction
- Empirical Study on VirusTotal Data
- Research Opportunities
- Conclusion

Outline

- Introduction
- Empirical Study on VirusTotal Data
- Research Opportunities
- Conclusion

Empirical Study



Step 1

Data Collection

Temporal Properties

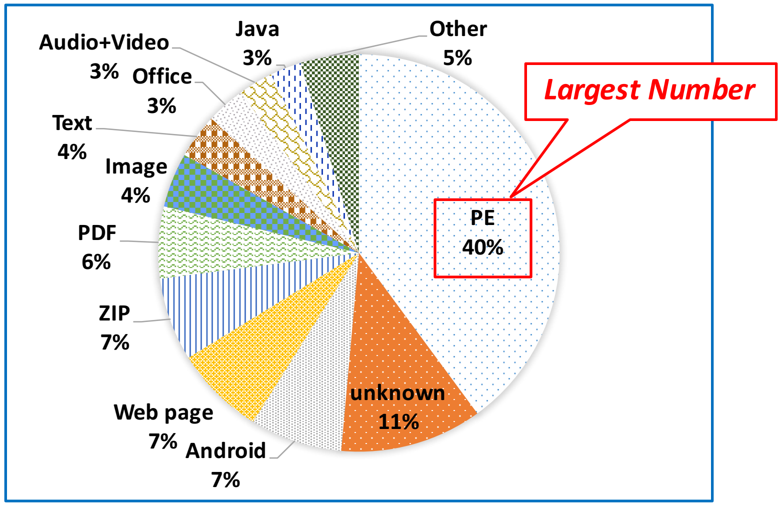
Distribution Properties

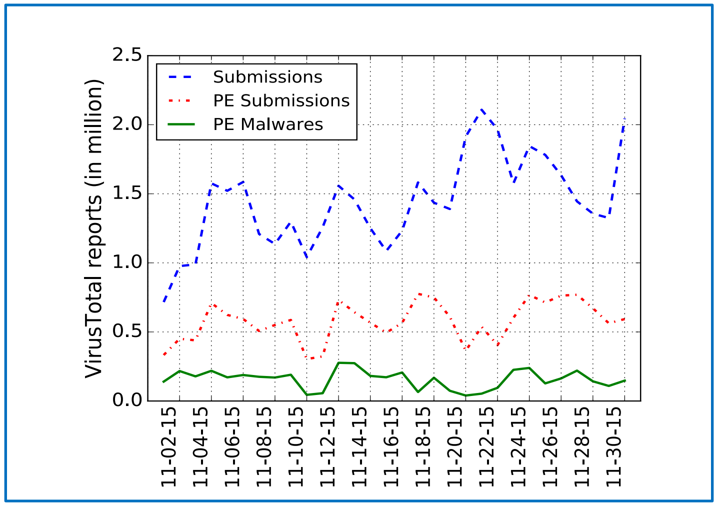
Data Collection

- What to collect for each submission?
  - Metadata
    - File information: size, type
    - Submission information: timestamp, ID, country
    - Different hashes: ssdeep, sha256, md5
  - Analysis results
    - Roughly 50 engines used for each file
- All 43 million submissions in 2015/11

Preprocessing

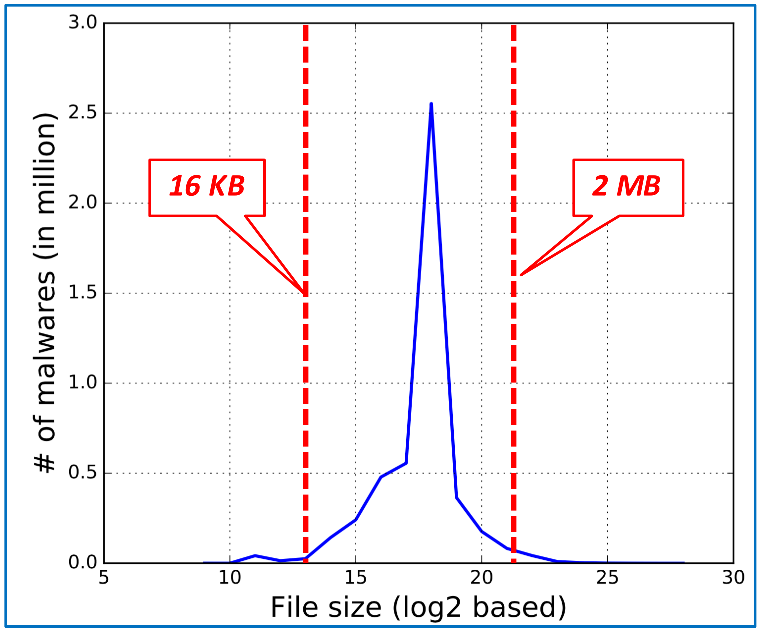
- Focusing on PE files
- Merging redundant submission reports
- Leveraging Microsoft engine
  - Identifying malwares from benign files
  - Deciding malwares’ families



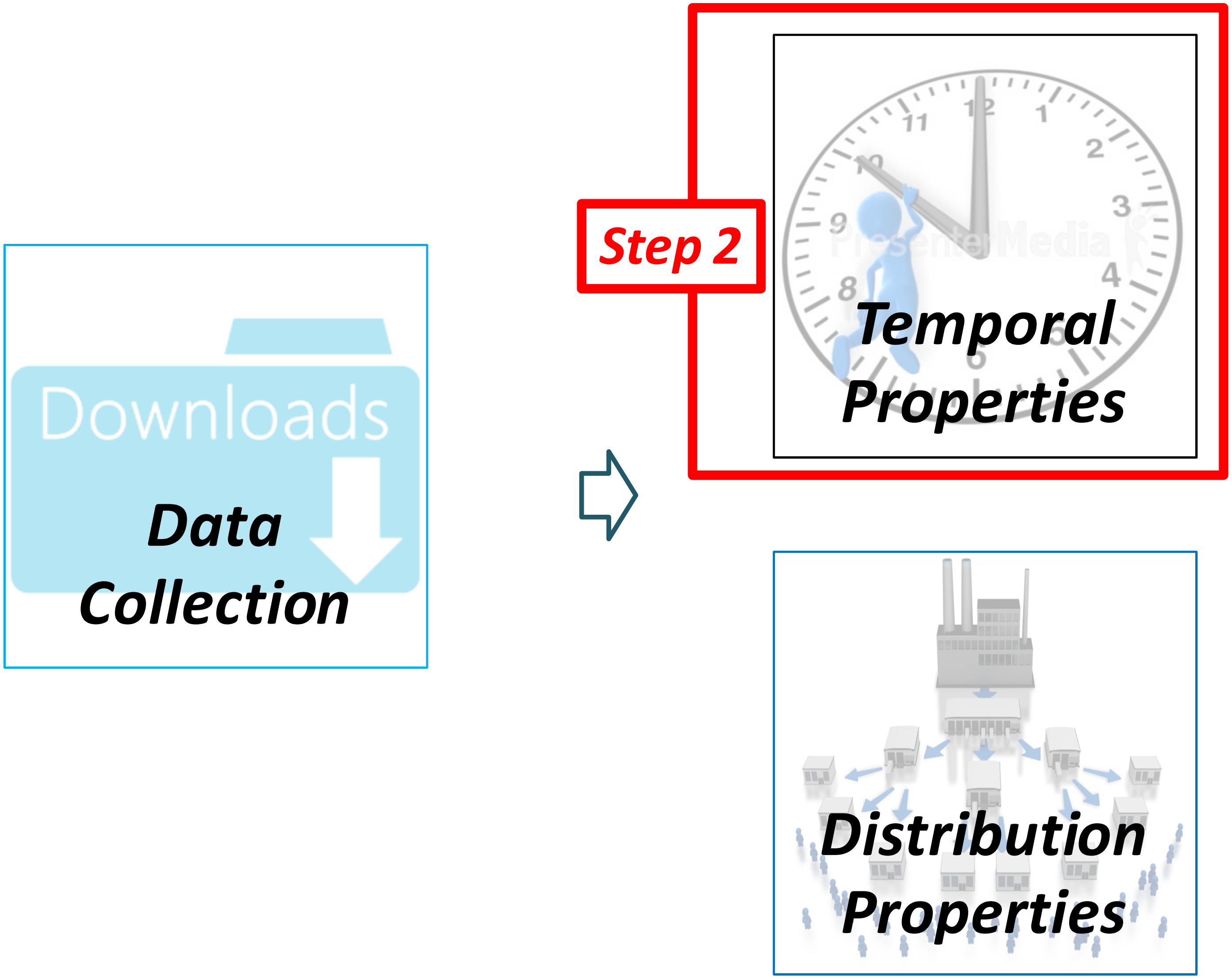


Basic Properties

- Most malwares submitted once in 2015/11
  - Average submission number is 1.17
- Most malwares > 16 KB && < 2MB
- Most malwares are 32-bit



Empirical Study



Data Collection

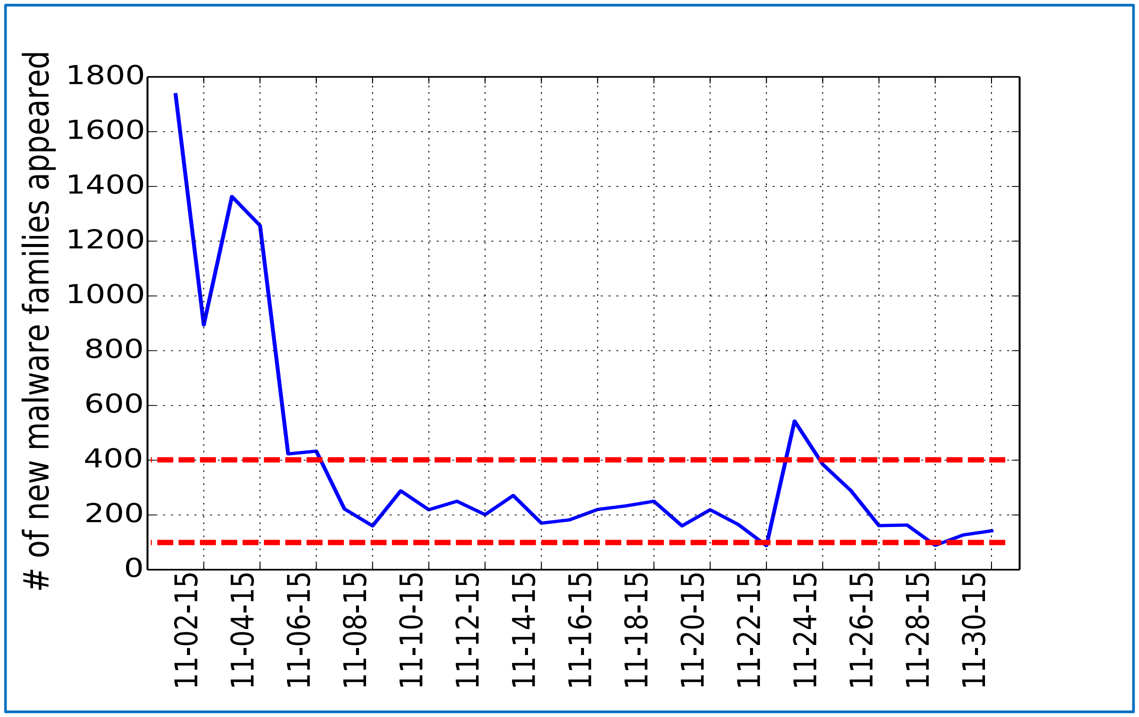
Step 2

Temporal Properties

Distribution Properties

Malware Family Generation Rate

Observation 1: 100-400 new malware families appear each day.

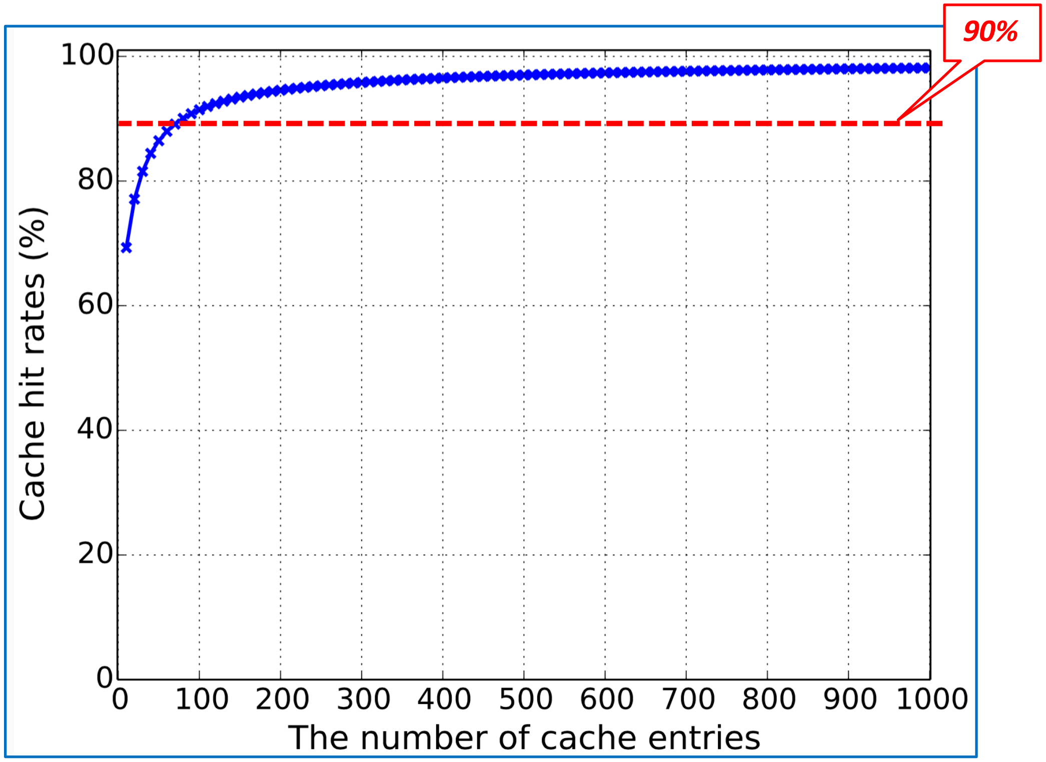


Temporal Locality (I)

- Definition
  - How bursty malwares in the same family appear
- Cache mechanism
  - Cache design
    - Address: malware family
    - Time: submission timestamp
    - Cache hit: new submission's family in the cache
  - Cache setting
    - Setting block size to be 1, no prefetching, LRU

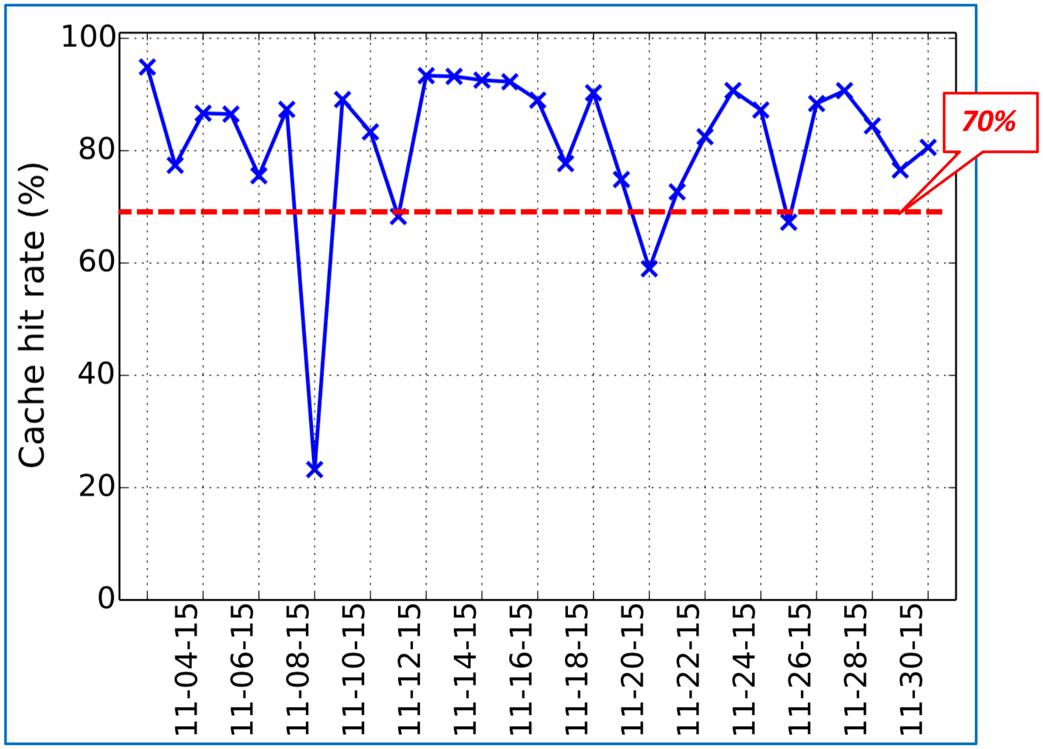
Temporal Locality (II)

Observation 2: The occurrence of malwares in each family has strong temporal locality.

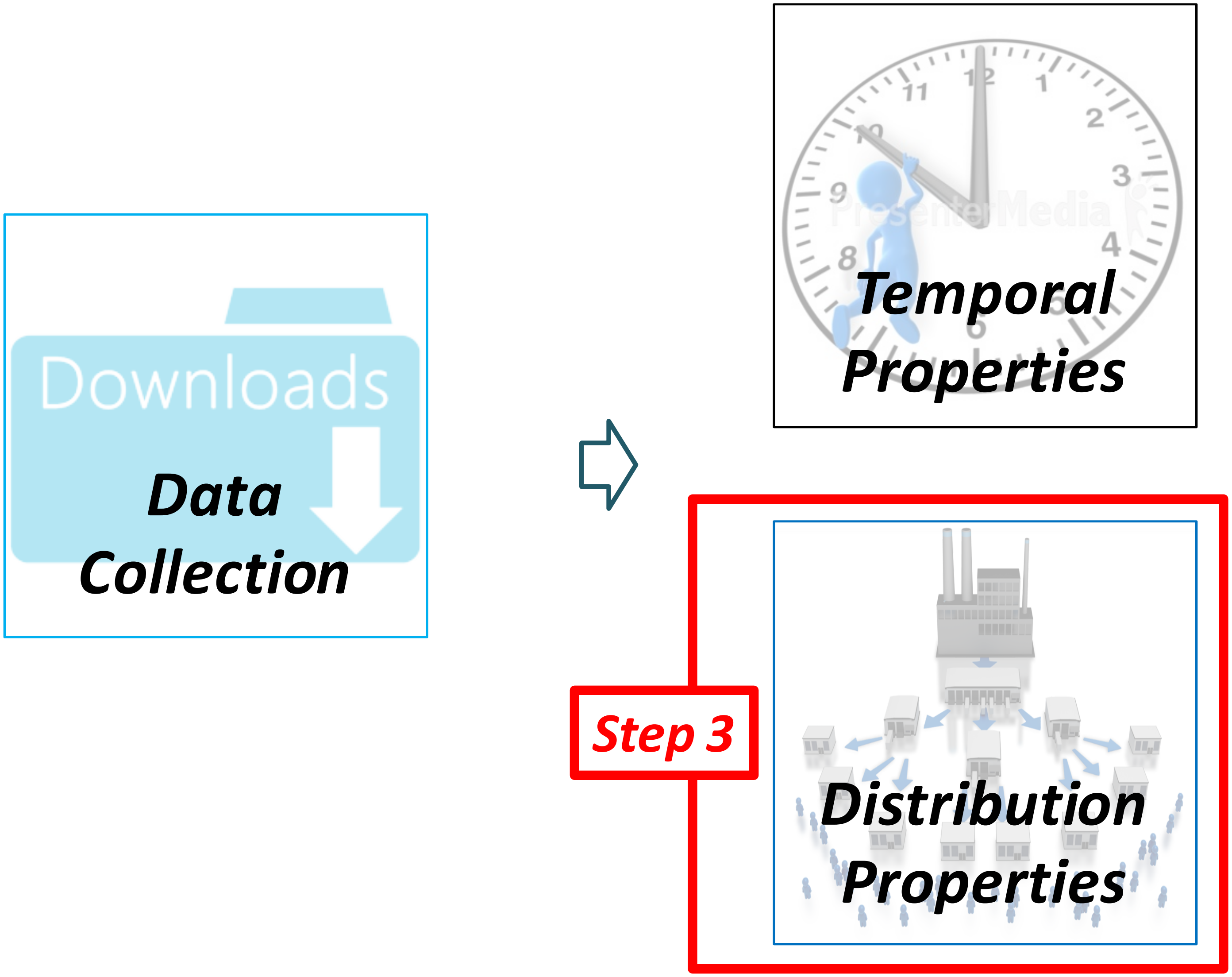


Temporal Locality (III)

- Online malware occurrence prediction
  - Updating cache content once a day
  - Fixing cache size to be 200



Empirical Study



Data Collection

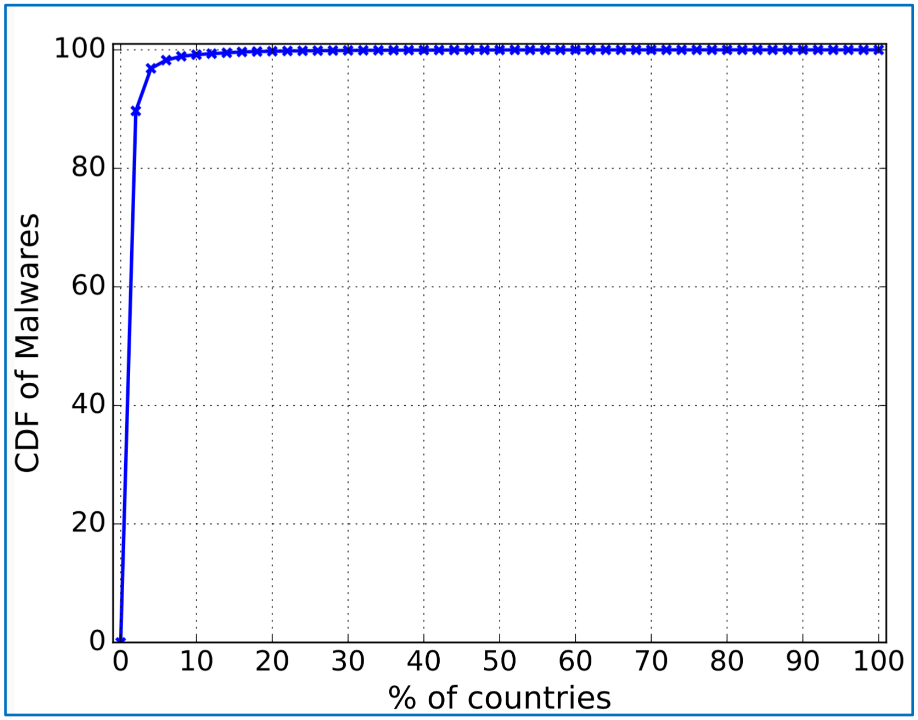
Temporal Properties

Step 3

Distribution Properties

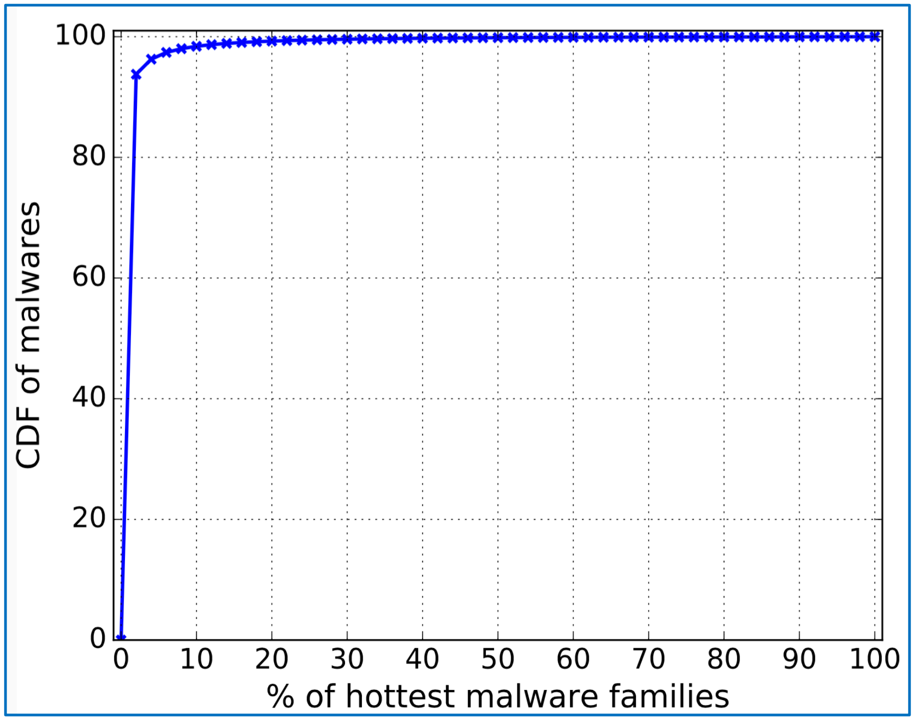
Submission Country Distribution

- Submitted from 164 countries
- Top 5 countries include
  - Canada, USA, China, France, and Germany



Malware Family Distribution

Observation 3: Distributions of malwares are highly skewed in countries and malware families.



Outline

- Introduction
- Empirical Study on VirusTotal Data
- Research Opportunities
- Conclusion

Correlation Mining

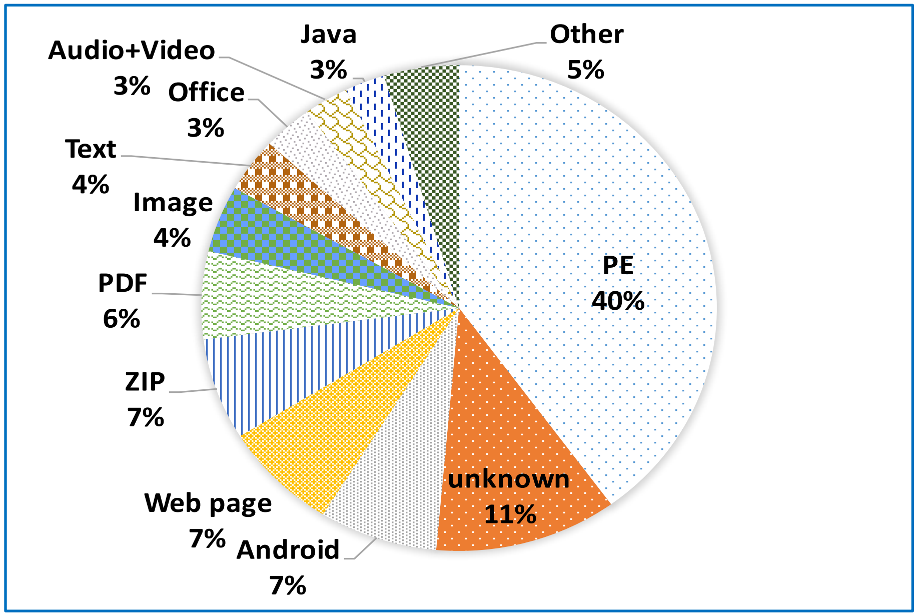
- Information on VirusTotal
  - Metadata fields
  - Static features from executable
  - Dynamic behaviors
- Correlation mining
  - Which features/behaviors are more suspicious?
  - Which features/behaviors are ignored?

Evaluating Vendors’ Reports

- 50+ different engines used for each submission
  - Detailed detection results
  - How detection results change
- Questions to answer?
  - Are there influences between different vendors?
  - How to combine results from different vendors?

Studying Other File Types

- We only study PE files
- Question to answer?
  - How other malicious files distribute?
  - How other malicious files behave?



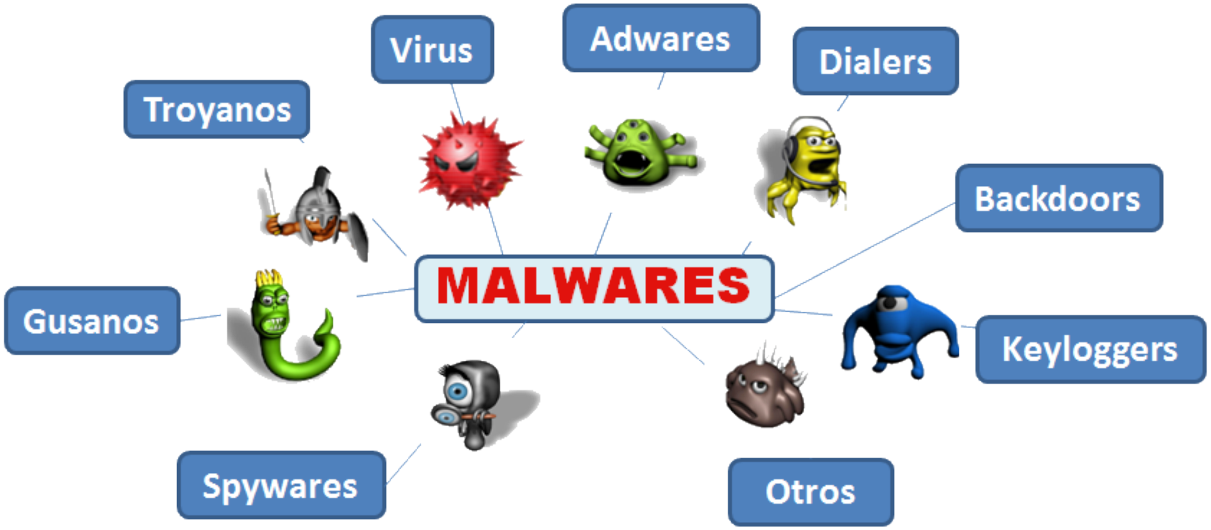
Machine Learning

- A huge set of labeled malwares on VirusTotal
- How about applying machine learning?
  - Training models using VT data
  - Using trained models to detect/classify malwares
- Questions to answer?
  - Which features on VT are useful?
  - Whether extracting features not on VT scalable?

Conclusion

- An early-stage empirical study on VT data
  - Temporal properties
  - Distribution properties
- Research Opportunities
  - Leveraging more information
  - Mining correlations
  - Applying machine learning

Thanks a lot!


Virus
Adwares
Dialers
Troyanos
Backdoors
MALWARES
Gusanos
Keyloggers
Spywares
Otros