Analysis of Malware Behavior: Type Classification using Machine Learning

Steven S. Hansen
Radu S. Pirscoveanu
Thor M. T. Larsen
Matija Stevanovic
Jens M. Pedersen
Alexandre Czech

Introduction

- Malware is a threat to the modern society
- Approximately 390.000 new malware emerge each day according to AV-TEST
  - Many of them are variants originating from the same code

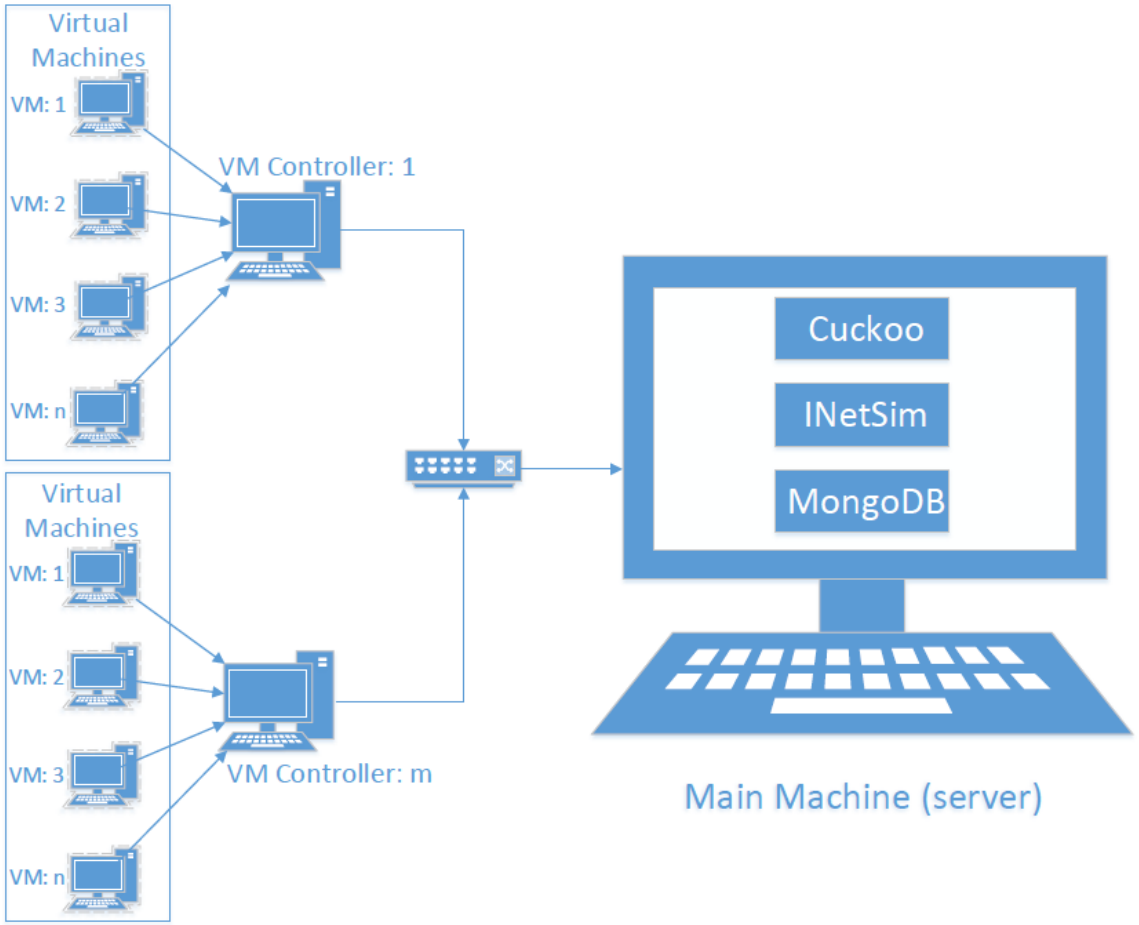
How to classify a large amount of malware using machine learning ?

Introduction

- Dynamic analysis
  - Executing malware in a secure environment
  - Collects behavioral data from the samples
- Pre-filtering application
  - Filter known malware from novel malware
  - Proof of concept

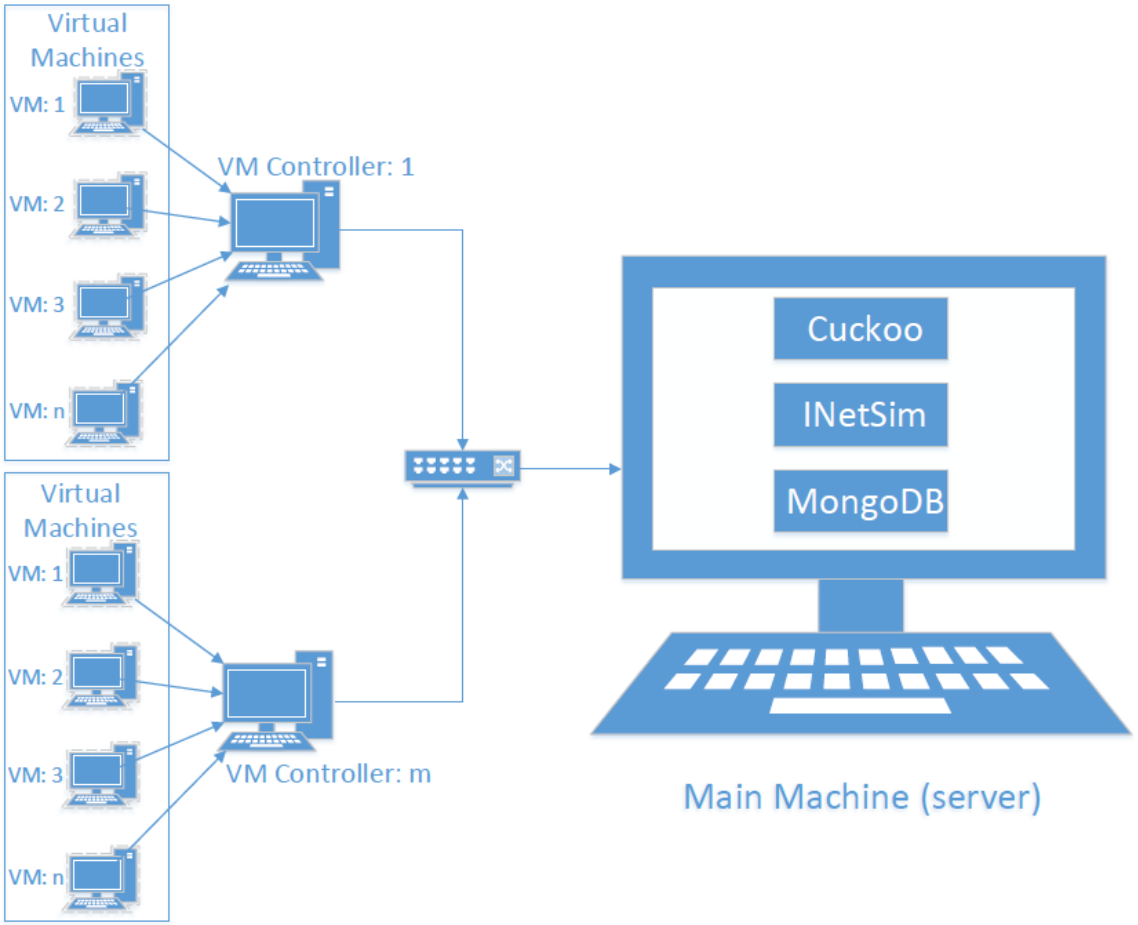
Analysis Setup

- Scalable and distributed



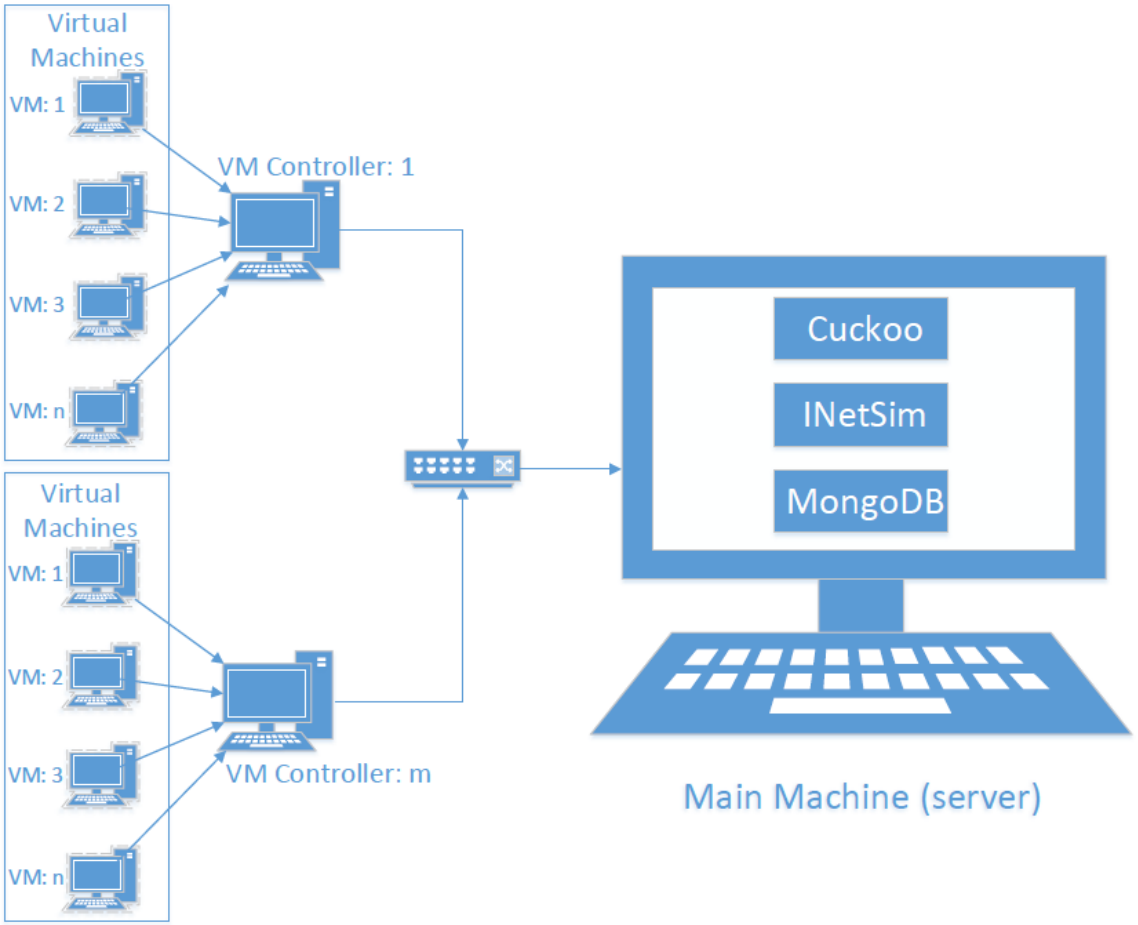
Analysis Setup

- Emulate Internet services using INetSim



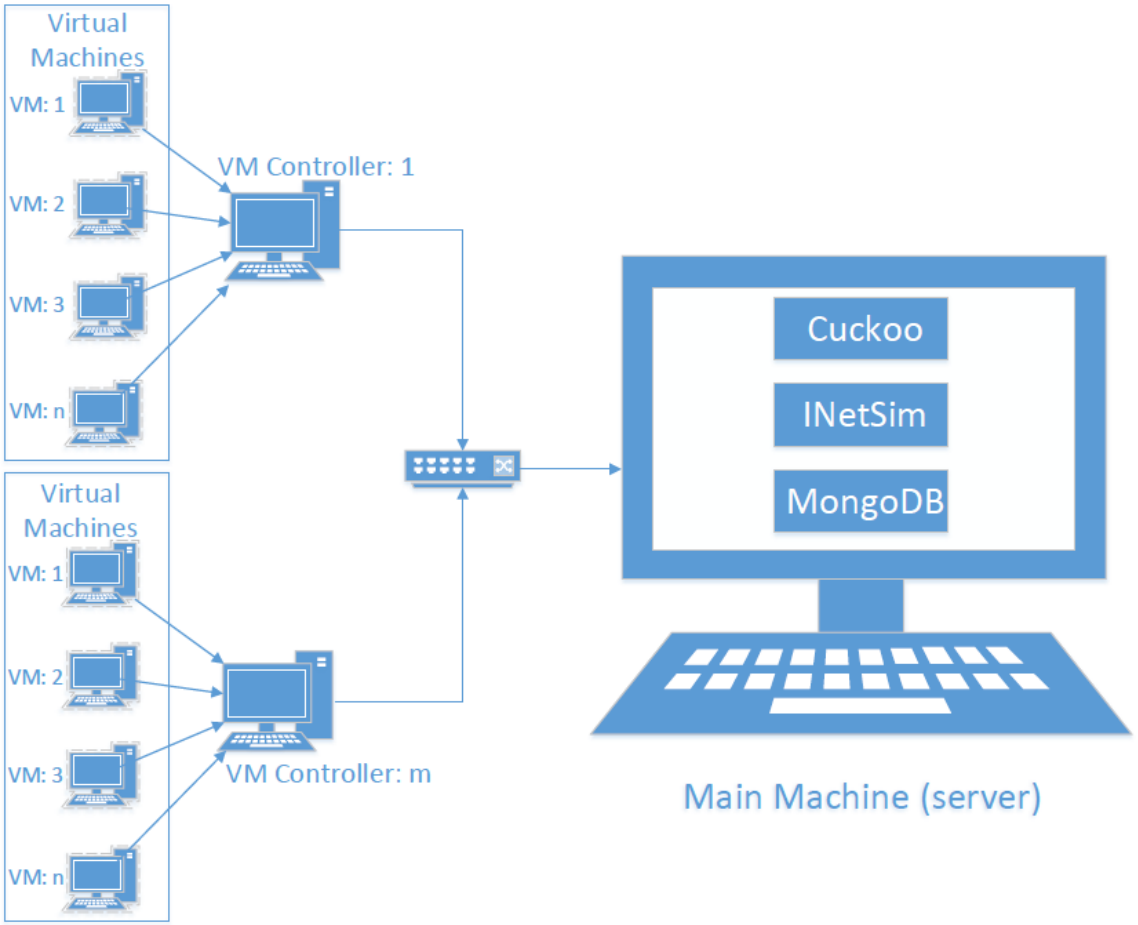
Analysis Setup

- VMs are personalized



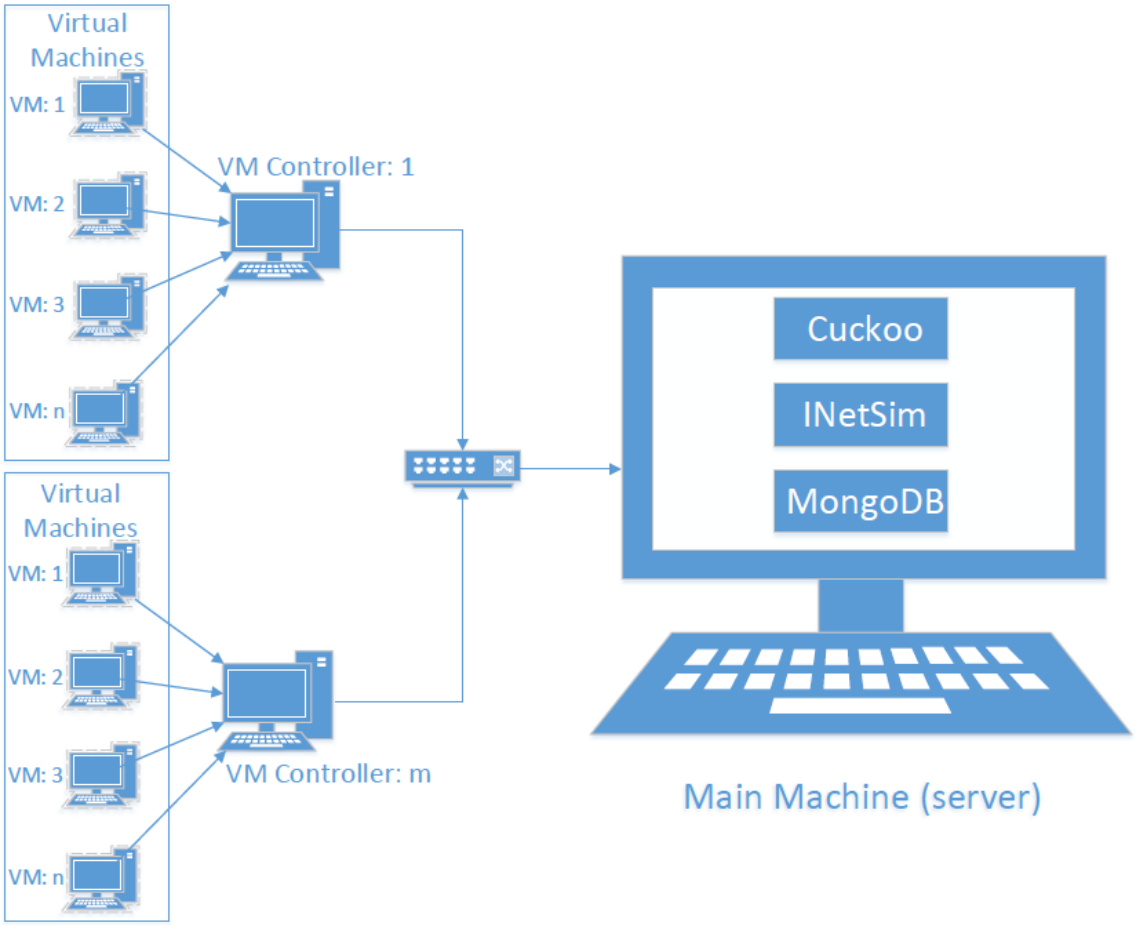
Analysis Setup

- 80.000 samples are analyzed



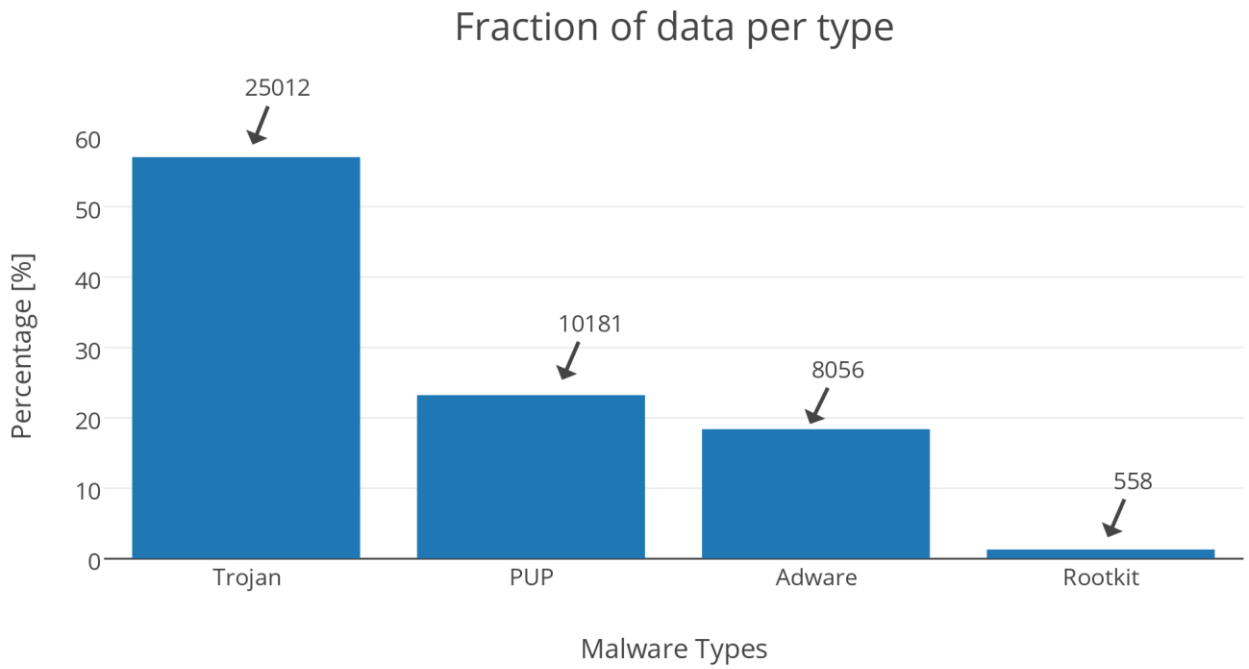
Analysis Setup

- Each sample is analyzed for 200 sec



Malware samples

- Supervised machine learning
- Avast is used to extract labels
  - Approximately 42.000 samples are labeled



Features

- Main parameter
  - API calls
- Secondary parameters
  - Mutexes
  - Registry Keys
  - Files
  - DNS requests

Feature Representation

- Sequence
  - Distinct API calls

	SEQ_1	SEQ_2	$\cdots$	SEQ_{40}	BIN_1	BIN_2	$\cdots$	BIN_{24}	CTR_1	CTR_2	CTR_3	CTR_4
S_1	API_{14}	API_{12}	$\cdots$	API_{112}	677	43	$\cdots$	83	4	23	502	12
S_2	API_{76}	API_{47}	$\cdots$	API_{146}	4	8785	$\cdots$	51	63	3	20	322
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
S_m	API_{97}	API_3	$\cdots$	API_{32}	414	27	$\cdots$	7397	0	9	89	3

Feature Representation

- Frequency Bins
  - Frequency of all APIs for each bin are summed

	SEQ_1	SEQ_2	$\cdots$	SEQ_{40}	BIN_1	BIN_2	$\cdots$	BIN_{24}	CTR_1	CTR_2	CTR_3	CTR_4
S_1	API_{14}	API_{12}	$\cdots$	API_{112}	677	43	$\cdots$	83	4	23	502	12
S_2	API_{76}	API_{47}	$\cdots$	API_{146}	4	8785	$\cdots$	51	63	3	20	322
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
S_m	API_{97}	API_3	$\cdots$	API_{32}	414	27	$\cdots$	7397	0	9	89	3

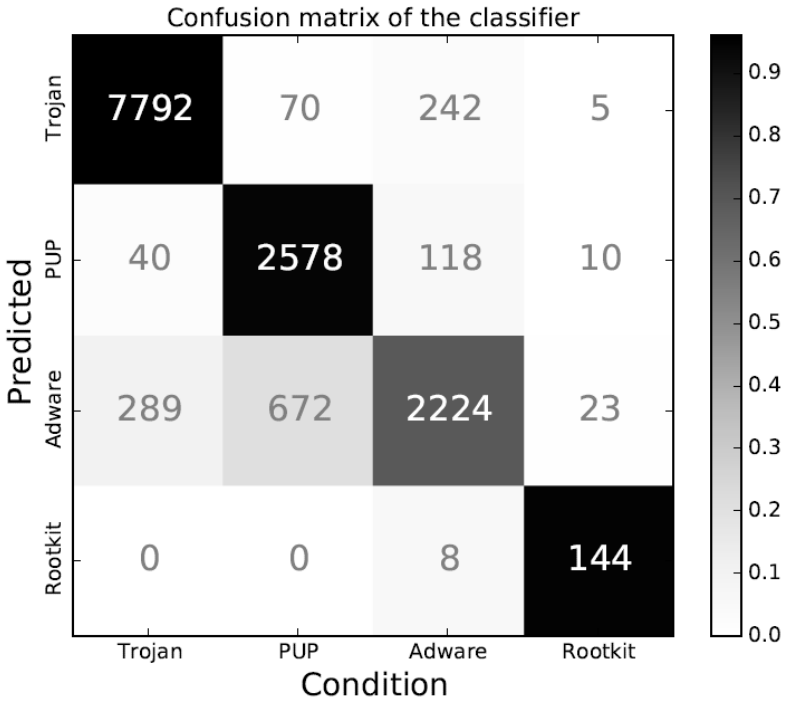
Feature Representation

- Counters
  - Count number of actions performed for each secondary parameter

$$
\begin{array}{c|cccc|cccc|cccc}
 & SEQ_1 & SEQ_2 & \cdots & SEQ_{40} & BIN_1 & BIN_2 & \cdots & BIN_{24} & CTR_1 & CTR_2 & CTR_3 & CTR_4 \\
S_1 & API_{14} & API_{12} & \cdots & API_{112} & 677 & 43 & \cdots & 83 & 4 & 23 & 502 & 12 \\
S_2 & API_{76} & API_{47} & \cdots & API_{146} & 4 & 8785 & \cdots & 51 & 63 & 3 & 20 & 322 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\
S_m & API_{97} & API_3 & \cdots & API_{32} & 414 & 27 & \cdots & 7397 & 0 & 9 & 89 & 3
\end{array}
$$

Results

- Random Forests
  - 160 trees



Class	**F-measure**	**AUC**
Trojan	0.960	0.989
PUP	0.850	0.978
Adware	0.767	0.955
Rootkit	0.862	0.970
Weighted Avg.	0.898	0.980

Conclusion

- Cuckoo Sandbox
- Feature representation
- Random Forests
- Pre-filtering application

Future Work

- Uniform dataset
- Ambiguous type description
- Project is continued

- Q&A
- Email: steven9220@gmail.com