IP Reputation Analysis of Public Databases and Machine Learning Techniques

Jared Lee Lewis

Geanina F. Tambaliuc

Husnu S. Narman

Wook-Sung Yoo

Weisberg Division of Computer Science

Marshall University

narman@marshall.edu

https://hsnarman.github.io/

February 2020

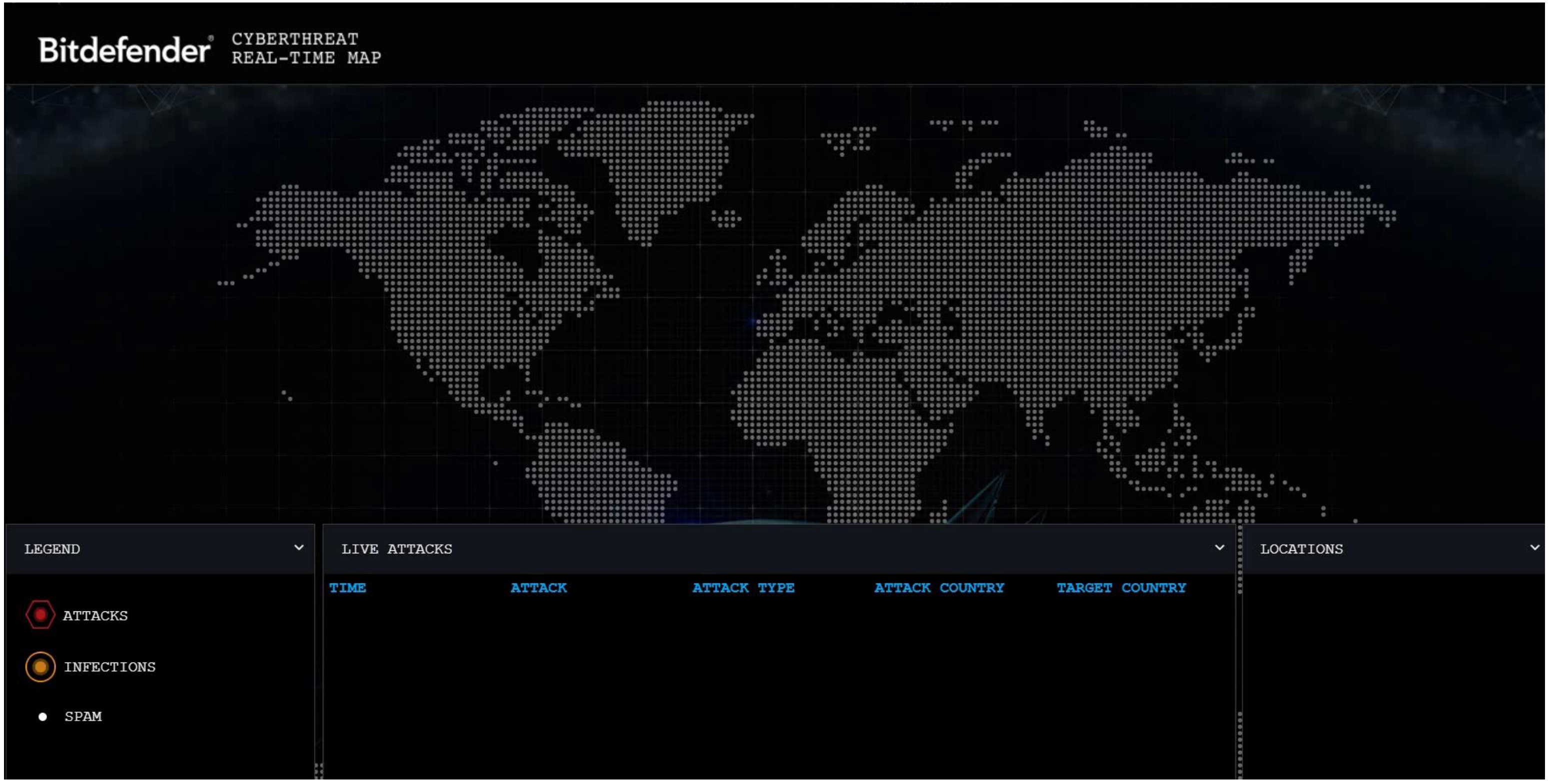
Outline

- Introduction
- Blacklists
- Machine Learning Techniques
- System Model
- Results
- Conclusion

Introduction

- The common usage of Internet adds many challenges in terms of protecting user data.
- Unfortunately, applications cannot protect the user privacy and become a threat to user data security because of new malware.
- 4 new malware samples discovered / sec
- More than 200 million new malware samples / year

Introduction



TIME	ATTACK	ATTACK TYPE	ATTACK COUNTRY	TARGET COUNTRY

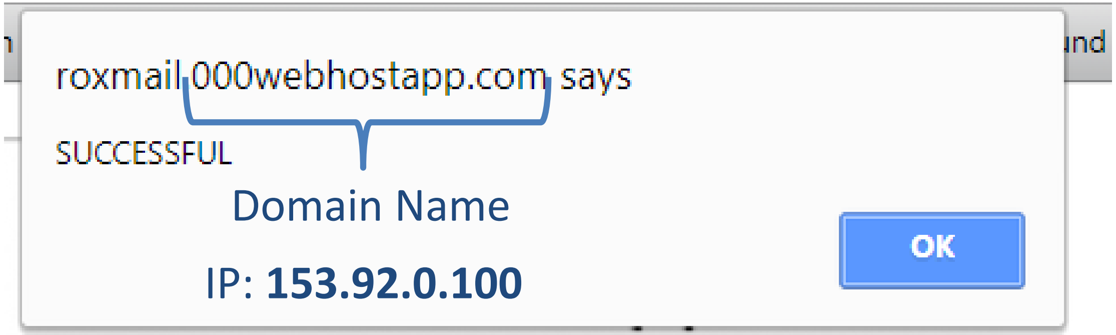
Microsoft Exchange

To prevent the users from spam and phishing email, Microsoft Exchange uses 8 filtering criteria:

- Connection Filtering
- Sender Filtering
- Recipient Filtering
- Sender ID
- Content Filtering
- Sender Reputation
- Attachment Filtering
- Junk Email Filtering

The Importance of DNS

The Domain Name System (DNS) plays an important role in filtering and protection techniques because DNS protocol is used by both cyber-attacks and authorized services.



roxmail.000webhostapp.com says

SUCCESSFUL

Domain Name

IP: **153.92.0.100**

Objective

The objective of this research is to analyze the public databases and machine learning techniques to detect malicious IP addresses and domains and introduce Automated IP Reputation Analyzer Tool (AIRPA), which uses both approaches to check the reputations of IPs and domains.

Public Blacklist Databases

- Seven main databases:
  - VirusTotal
  - URLVoid
  - MyIP.MS
  - Censys
  - AbuseIPDB
  - Apility.io
  - Shodan

  and 102 sub-databases.

Limitations of Public Blacklist Databases

Unfortunately, the public blacklists have some limitations (Free versions):

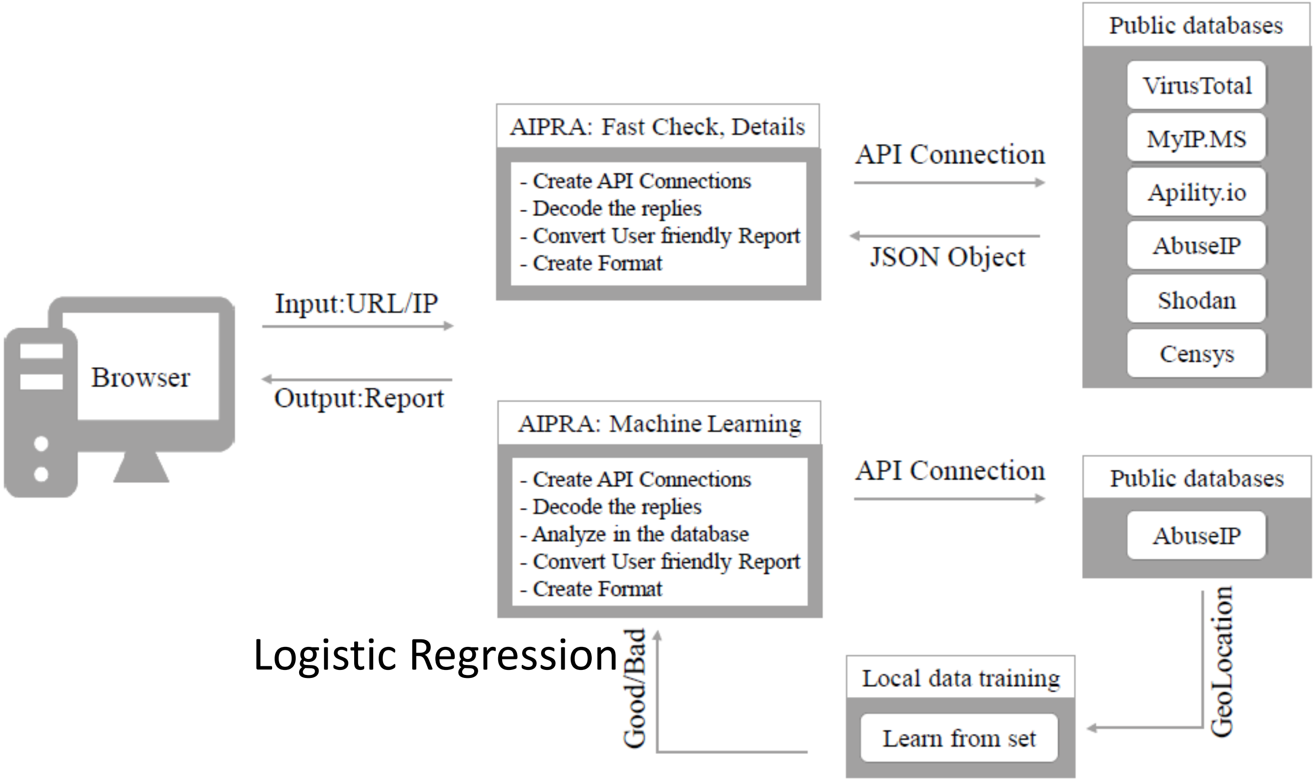
- VirusTotal: 4 requests / minute
- AbuseIPDB: 1,000 reports and checks per day and 60 requests per minute
- Shodan: 1 request/ second
- MyIP.MS: 150 requests/month
- Apility.io: 250 requests/day and 50 requests/minute
- Censys: 250 requests/month
- May not regularly update
- Wrong information

Machine Learning Models

With 80,000 good and 80,000 bad domains

- Logistic Regression
- Bayes
- Random Forest
- Logistic Regression with geolocation
- Bayes with geolocation
- Random Forest with geolocation

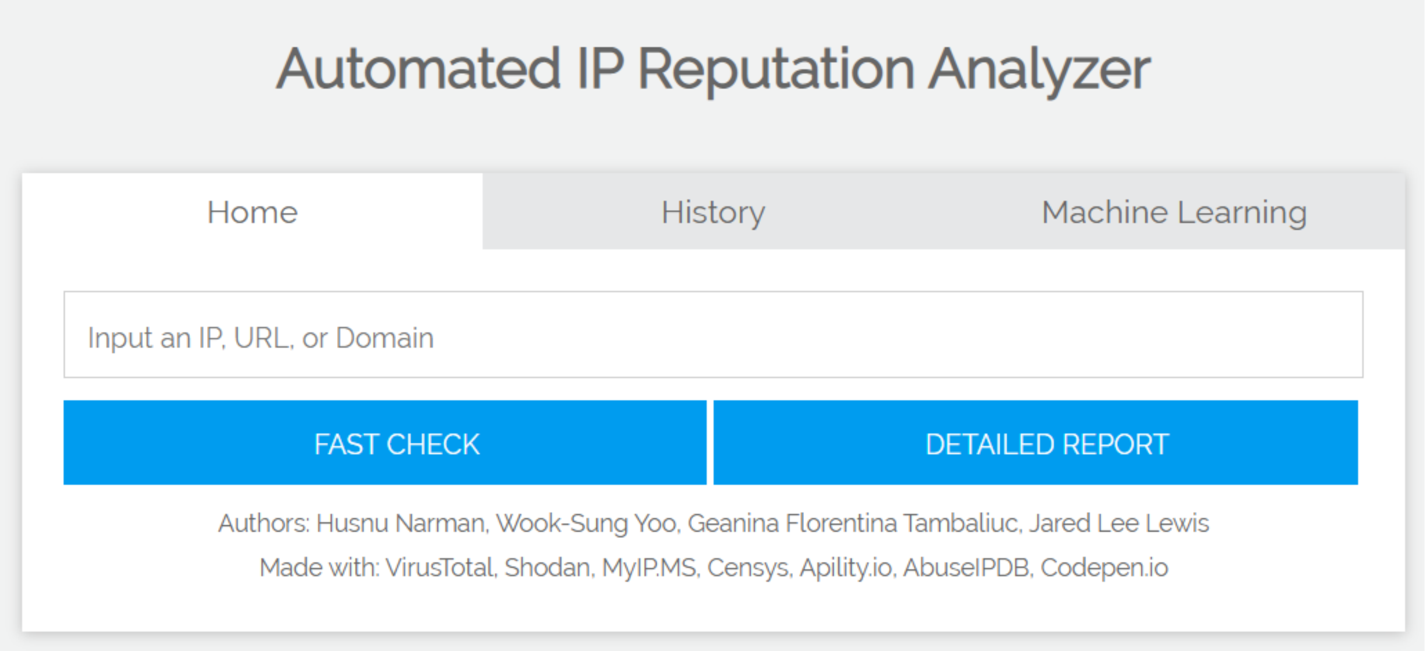
System Model and App: http://ipreputation.herokuapp.com/



Logistic Regression

Husnu S. Narman

App: http://ipreputation.herokuapp.com/

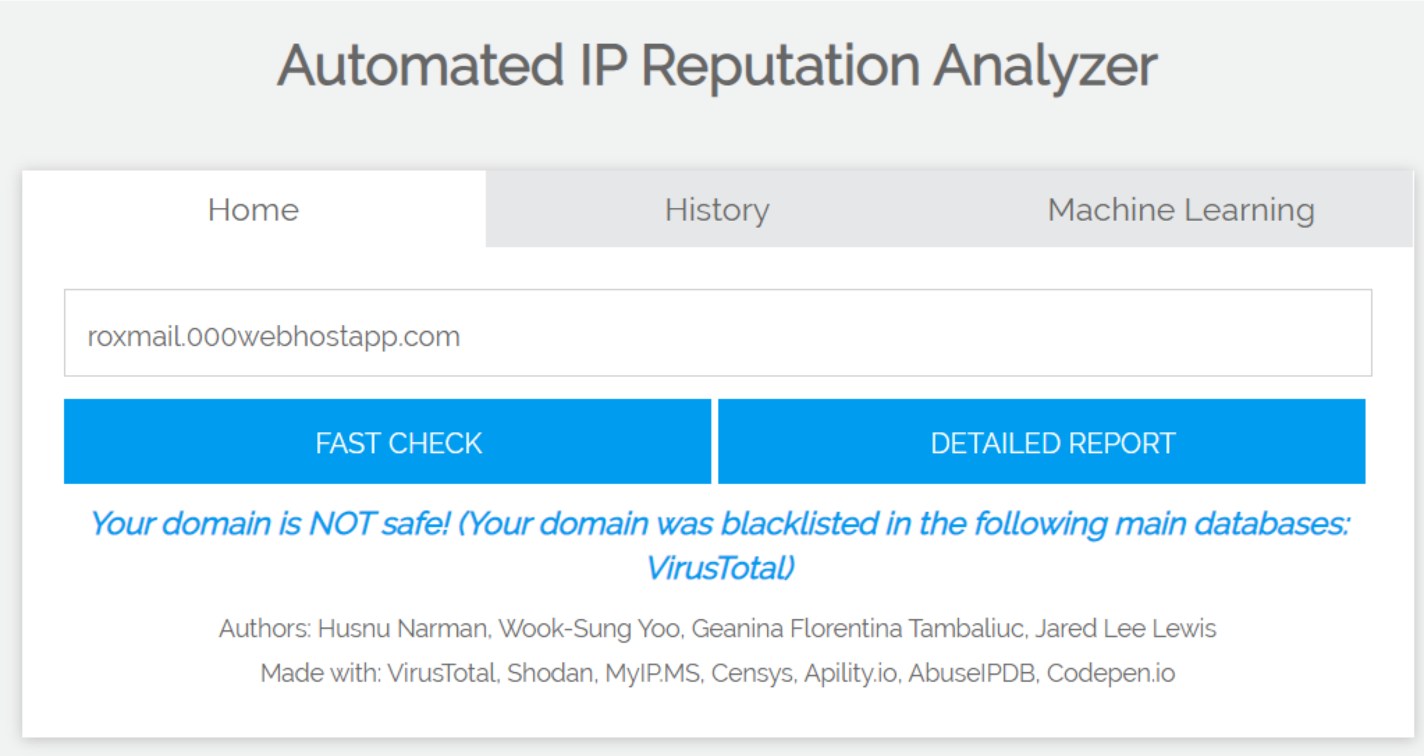


Automated IP Reputation Analyzer

Authors: Husnu Narman, Wook-Sung Yoo, Geanina Florentina Tambaliuc, Jared Lee Lewis

Made with: VirusTotal, Shodan, MyIP.MS, Censys, Apility.io, AbuseIPDB, Codepen.io

App Fast Check: http://ipreputation.herokuapp.com/



Automated IP Reputation Analyzer

roxmail.000webhostapp.com

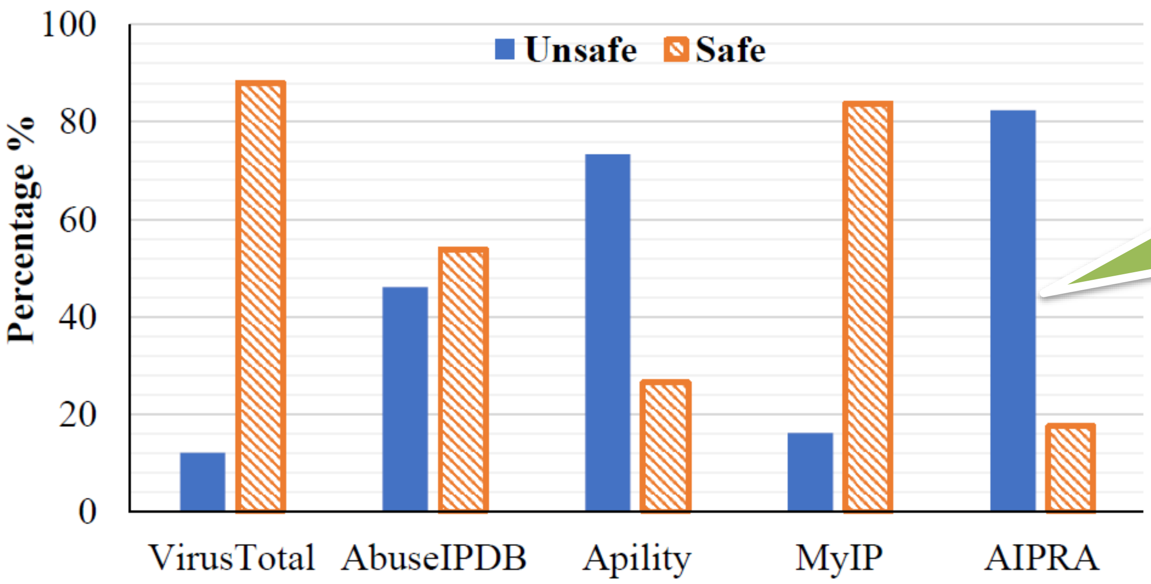
Your domain is NOT safe! (Your domain was blacklisted in the following main databases: VirusTotal)

Authors: Husnu Narman, Wook-Sung Yoo, Geanina Florentina Tambaliuc, Jared Lee Lewis

Made with: VirusTotal, Shodan, MyIP.MS, Censys, Apility.io, AbuseIPDB, Codepen.io

Results

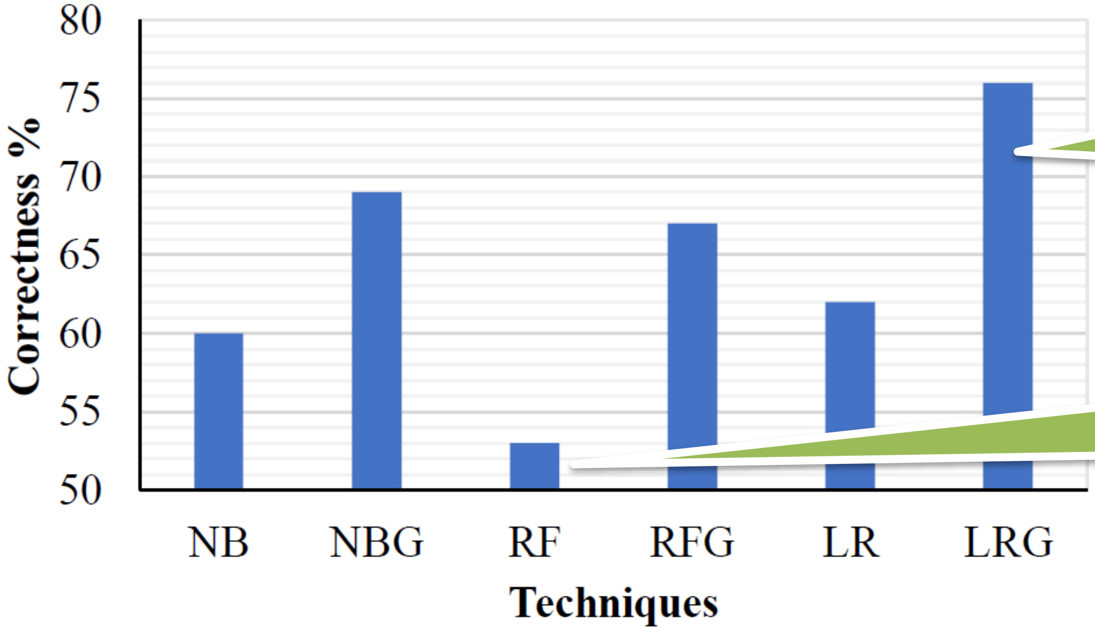
Result for testing unsafe 1586 IPs in public databases and AIRPA



AIRPA has the highest correctness rate with cross check

Results

Result for testing distinct learning techniques with/without geolocation



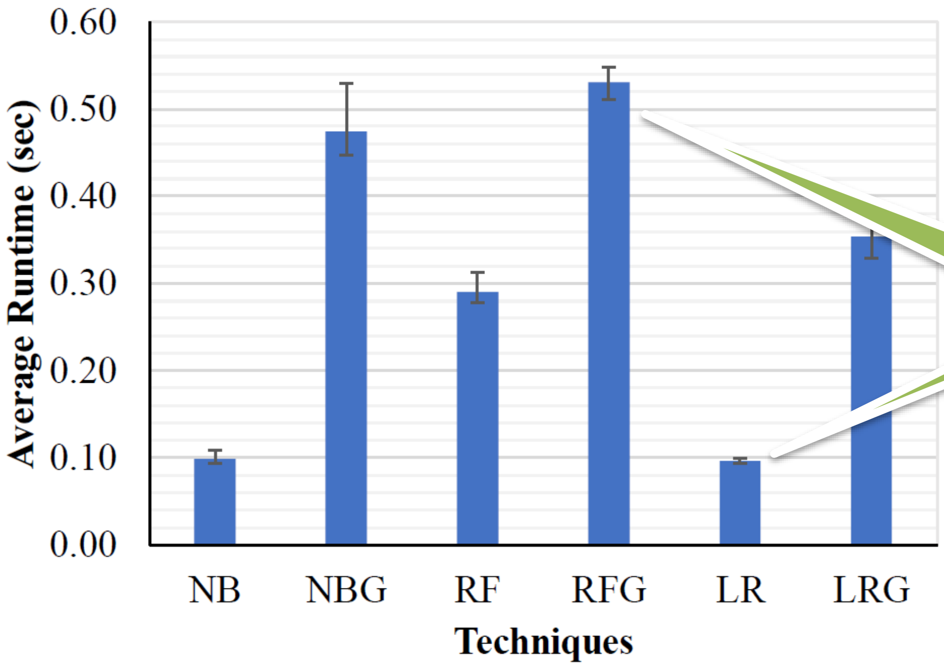
Logistic Regression with geolocation has the highest correctness.

Random Forest without geolocation has the lowest correctness.

Husnu S. Narman

Results

Result for Runtime of distinct learning techniques with / without geolocation.



Logistic Regression has the lowest running time.

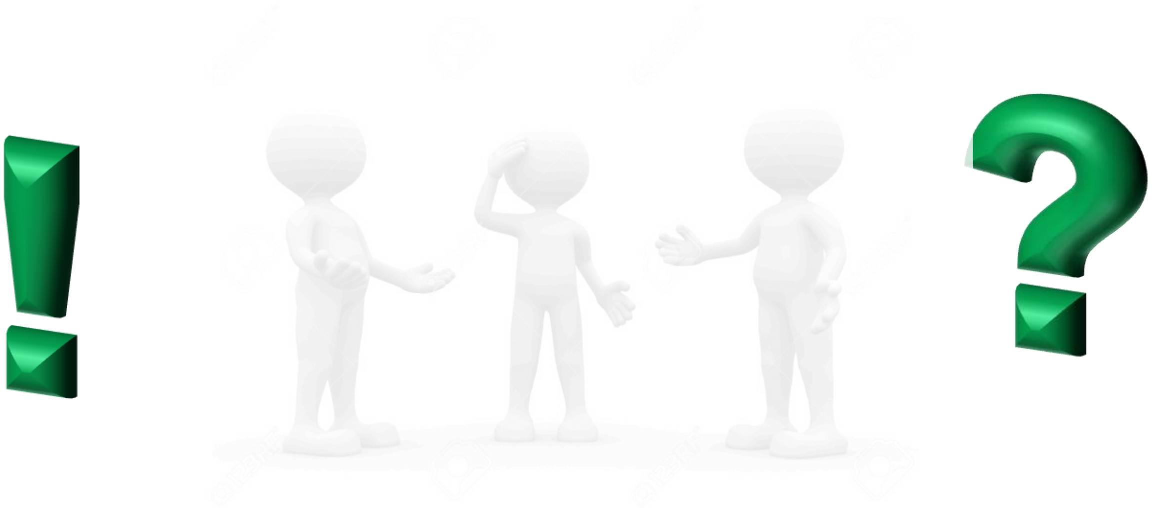
Random Forest with geolocation has the highest running time.

Conclusion

Cross-checking system is better in terms of detection the malicious IPs in public databases but also decrease false positives.

Considering additional parameters with machine learning techniques to find IPs' reputations can affect the obtained results in a better way but increase runtime

Ability in public databases and Logical Regression in machine learning techniques have higher detection rates.



Thank You

narman@marshall.edu

https://hsnarman.github.io/

Husnu S. Narman