The Game of Threat Hunting

By

Akash Sarode

Author: Akash Sarode, akky_sanj@yahoo.com
CEH, CHFI, ISO27001 LA, ACE, Cyber security researcher, Author of whitepapers – Machine learning-learning cybersecurity, Instructor of the course: "Machine Learning: The Future"

Abstract

Threat Hunting is proactively looking out for threats in your organization. Manual hunting is totally dependent on hunter skills and expertise. The entire world is moving towards automation so why can't we add this flavor of automation in our threat hunting process. We all know threat hunting can never be fully automated but let's try making it semi-automated with cyber analytics approach. Machine learning along with threat hunting on a manual level will not only identify known threats but also zero days and APT whose behavior is unknown to the entire world.

Table of Contents

What is Threat hunting?

In traditional security monitoring approach, most of the blue teamers look out for threats based on the alerts being triggered out by SIEM or other security devices. In addition to alert–driven approach, why can't we add a continuous process for finding stuff from the data without any alerts driving us for incident. That's the process of Threat hunting, proactively looking out for threats in network. Those threats which are not identified by your existing security solutions or attacks which bypassed your solutions can be hunted down using this process.

Instead of depending on tools for giving us alert, we can manually hunt down for threats in our organization based on real world threats discovered on a daily basis. But do you think it's possible manually considering the scarce number of skilled threat hunters. In order to cope up with this issue, we need some automation in our hunting process and cyber security analytic can be useful.

Threat Hunting is not a Technology but Approach

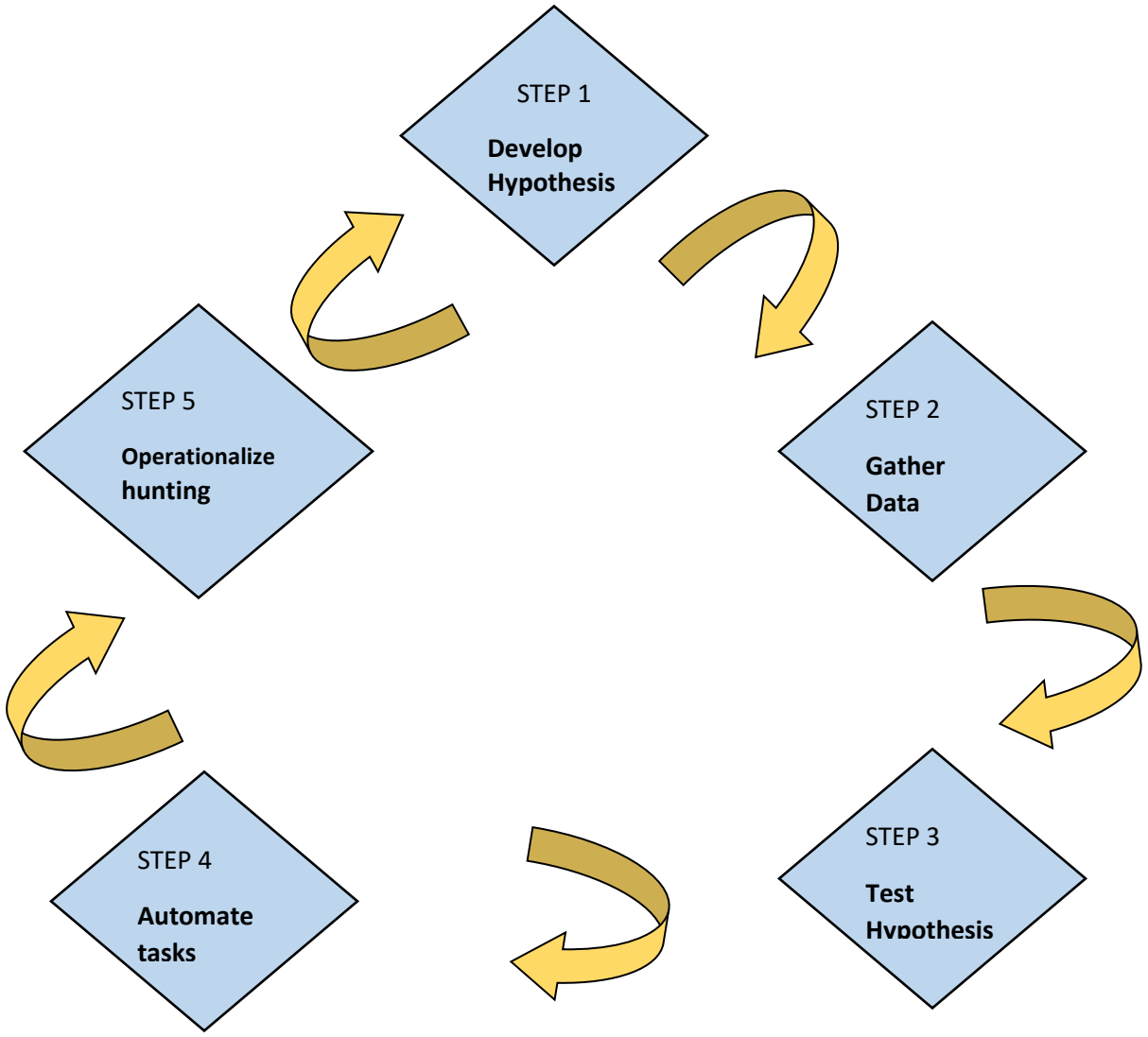
As a Security analyst, Threat Hunting is applying our knowledge in an effective way to look out for any anomalies in the environment. A threat hunter uses critical-thinking skills and creativity to look at patterns of normal behavior and be able to identify behavior anomalies.

Different ways of hunting

- **Manual** – Analyst need to continuously looking for anything that could be evidence/indicator of intrusion.
    - Important for the threat hunter to keep current on the latest security research.

- **<u>Automated/Machine-assisted</u>** – Analyst uses software that leverages "Machine Learning" and "UEBA" to inform analyst about potential risks.
    - It helps in providing Predictive and Prescriptive analytics.

Threat Hunting Process



1) **Develop Hypothesis** – Hypothesis means what you want to look for, like looking out for powershell commands making connection to internet etc.

2) **Gather data** – Based on the hypothesis, look out for data you need to collect for hunting.

3) **Test hypothesis and gather hunting** – Once data is collected, look out for threats based on behavior, search queries.

4) **Automate certain tasks** – Threat hunting can never be fully automated but semi-automated.

5) **Operationalize Threat Hunting** – Now, instead of ad-hoc hunting, operationalize your hunting program so that we can perform continuous threat hunting.

Gathering Hypothesis

Hypothesis is obtained from multiple sources:-

- Reading articles
- security news
- New APT public report
- Twitter
- Frameworks
- Blogs
- Sigma rules

Threat hunting is carried out on every kind of data sources like endpoint, network, perimeter, etc. Its just applying our knowledge in effective way to find anomalies. Critical thinking skills are required. Threat Intelligence which consists of Indicators of Compromise (IOC's) also plays important role in performing hunting.

MITRE ATT&CK for hunting

Most of the threat hunting platforms uses "Attack MITRE" adversary model.MITRE ATT&CK™ is a globally-accessible knowledge base of adversary tactics and techniques based on real-world observations.

MITRE ATT&CK™ is a globally-accessible knowledge base of adversary tactics and techniques based on real-world observations. The ATT&CK knowledge base is used as a foundation for the development of specific threat models and methodologies in the private sector, in government, and in the cybersecurity product and service community.

ATT&CK™

The knowledge base consists of many Tactics, corresponding techniques to those tactics and its underlying procedure to follow. Majorly tactics consists of following content:-

- **Initial access**
- **Execution**
- **Persistence**
- **Privilege escalation**
- **Credential access**
- **Defense Evasion**
- **Discovery**
- **Lateral movement**
- **Collection**
- **Command and control**

· **Exfiltration**

· **Impact**

Attack MITRE has also came up with a project name “CAR” Cyber Analytics Repository.

The Mitre team has listed down all those adversary behaviors and attack vectors carries out by an adversary on a victim machine. It provides you with description as well as some references regarding the threats, based on historical outbreak. It uses TTP’s Tactics, Techniques and Procedures and maps it to Cyber Kill chain. Most of the threat hunting methodology uses Mitre framework to carry out hunting process.

Analytics & Machine learning

Cyber Security Analytics is the process of applying analytics to billions to logs which are collected from various IT infrastructure devices to identify any kind of unusual behavior or threat. Analytics is nothing but generating some kind of pattern from data. Analytics can be Stack counting, Machine learning and much more. In Cyber analytics, we apply some kind of formula/analytic to large amount of data and generate output which plays an important role in detecting threats. Output can be in the form of graphs, pie charts, etc. Anomalies, outliers can be detected using analytical approach. In addition to manual process, we can also use the output which we get from machine learning models used to identify any kind of outliers in our environment so that we can further hunt the threat down.

The data which is required for cyber analytics can be collected in various forms such as:-

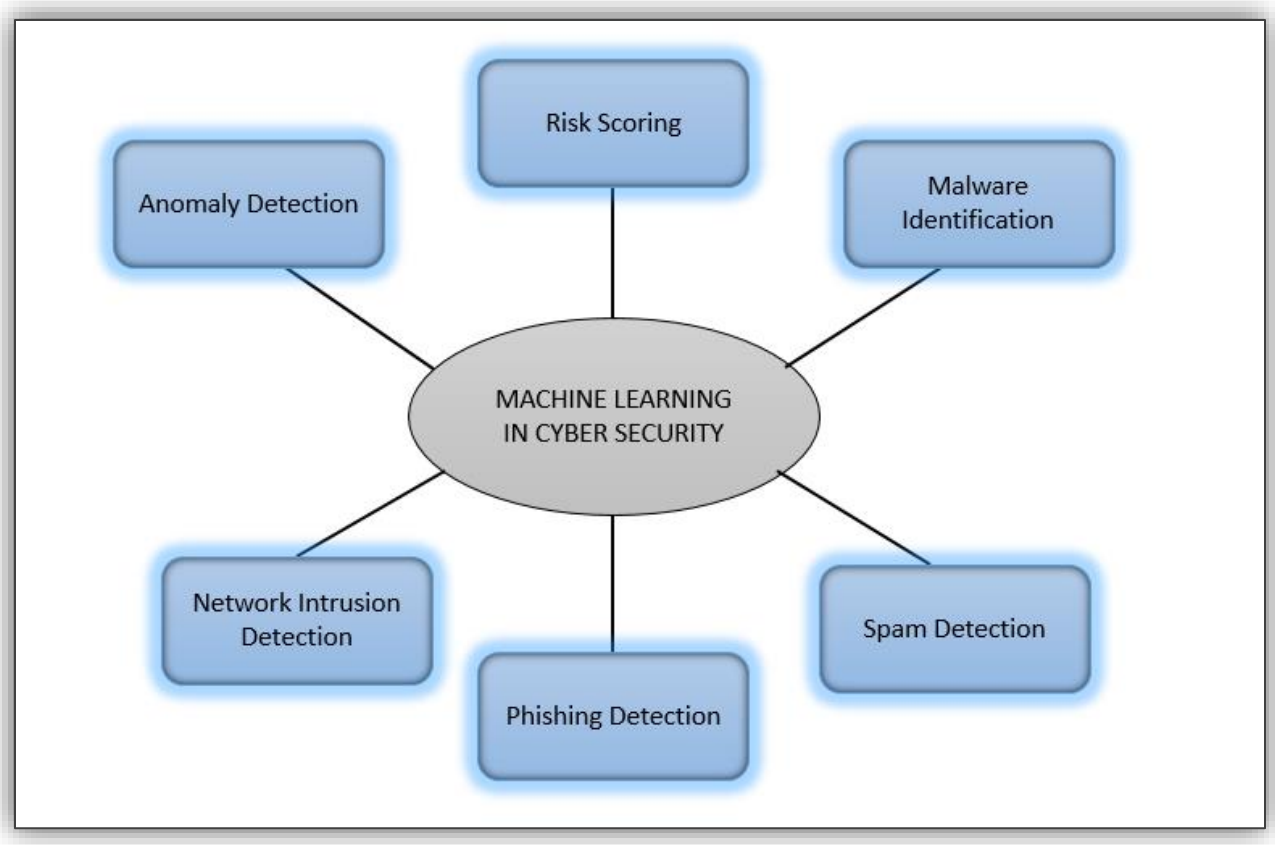
- Endpoint logs
- User and entity data
- Network Traffic

- Cloud resources
- Identity & access management
- Threat Intelligence feeds

Machine Learning plays an important role in analytical solution. Machine learning consists of various algorithms which can be applied to data and output can be generated based on the type of algorithm used. Mostly classification and clustering algorithms are used in Cyber security. It involves training set and test set wherein we initially train our data based on the training set and once, the algorithm has learned our data, we used new set of data or test set to generate output based on the algorithm.

Machine Learning & Cyber Security

Machine Security is of utmost importance for any organization and we need to blend in both these technologies together i.e. Machine Learning and Cyber Security to achieve this. There are multiple ways in which machine learning can be used in cyber security.



Anomaly detection, malware identification, Spam detection, Phishing detection, network intrusion detection, and many more are various application of Machine Learning in Cyber Security. So our problem is to understand whether Machine Learning is really worth learning data and if yes, how is it learning and how can it be applied to cyber security. To answer all these questions, we will practically demonstrate an application of ML in cyber threat Hunting.

Entire manual process of hunting may take time but this automation to some extent may ease down a work of threat hunter.

Threat Hunting can never be fully automated, but it can be semi-automated!!

Let's understand how we can use this semi-automated process.

Practical Threat hunting with automation

Let the Games of Threat Hunting begin….In this paper, I will be focusing on the web server logs which we collect from web servers similar to what we do in our SIEM. Various tools are available in market which performs the process of analytics using logs but we are going to perform each and every step manually, so that you understand what exactly happens behind the tool.

STEP 1 –

We need to have some web server traffic which can be distinguished as Normal Traffic and Anomalous traffic. This would be raw logs as we see in our SIEM or logs folder on server. Once, we have those logs we need to prepare a data set out of those logs which we can ingest to our machine learning algorithm. This data set can be in any format, For the purpose of simplicity, let's use csv format.

Creating data set out of traffic is very trivial. We will be using basic python functions for the same. Please find the below mentioned code snapshot:-

```
import pandas as pd
import re
import matplotlib.pyplot as plt


dframe = pd.DataFrame()
requesttype = []
responsecode = []
bytesl=[]
referer=[]
url=[]
useragent = []
nature = []


def anomalous(filename):
  f = open(filename,'r')
    doc = f.read()
    line = doc.split("\n")

    for l in line:
        if l[:4].__contains__("GET") or l[:4].__contains__("POST") or l[:4].__contains__("PUT") :
            url.append(l.split(" ")[1])
            nature.append("Malicious")
        if l[:10].__contains__("User-Agent"):
            useragent.append(l.split(": ")[1])
```

```
def normal(filename):
    f = open(filename,'r')
    doc = f.read()
    rl = re.findall(r"(GET|POST|OPTIONS|TRACE|PUT|DELETE)\s(.*?)\s.*?(\d{3})\s(\d+)\s\"(.*?)\"\s\"(.*?)\"", doc)

    for x in rl:
        requesttype.append(x[0])
        url.append(x[1])
        responsecode.append(x[2])
        bytesl.append(x[3])
        referer.append(x[4])
        useragent.append(x[5])
        nature.append("Non-Malicious")


anomalous("D:\\anomalousTrafficTest.txt")
normal("D:\\normal_test.txt")

dframe['URL'] = url
dframe["USERAGENT"] = useragent
dframe["REQUESTTYPE"] = requesttype
dframe["RESPONSECODE"] = responsecode
dframe["BYTES"] = bytesl
dframe["REFERER"] = referer
dframe.to_csv("D:\\NewFile.csv")
```

STEP 2-

Once, we get this data set csv file containing labeled data as Malicious and Non-Malicious, we need to read this data. Multiple fields are available in our data set such as UserAgent, URL, bytes, Referer, request type, but we will be using Query part of URL field to perform some analytics. The idea is based on the query parameter which are already labeled, our machine learning model will learn from data set and then once it has learned, it will be able to predict out which log is malicious or normal based on its learning.

Have a look at data set which has been formed:

A	B	C	D	E	F	G	H	I
Request M	url query	Response	Bytes	Referer	length	user agent	label	
GET	c/ caridad s/n	200	4263	-	14	Mozilla/5.0 (Windows NT 6.0; rv:3	norm	
GET	campillo, el	200	4494	http://almhuette-raith.at/ad	12	Mozilla/5.0 (Windows NT 6.1; WO	norm	
GET	40184	200	4263	-	5	Mozilla/5.0 (Windows NT 6.1; WO	norm	
GET	1.44E+15	200	4494	http://almhuette-raith.at/ad	16	Mozilla/5.0 (compatible; Googleb	norm	
GET	nue37	200	4263	-	5	Mozilla/5.0 (Linux; Android 4.4.2;	norm	
GET	nuda drudes	200	4494	http://almhuette-raith.at/ad	11	Mozilla/5.0 (Windows NT 6.1; WO	norm	
GET	tufts3@joll.rs	200	4263	-	14	Mozilla/5.0 (Windows NT 5.1; rv:6	norm	
GET	22997112x	200	4494	http://almhuette-raith.at/ad	9	T-Online Browser (Windows NT 6.	norm	
GET	c/ del ferrocarril, 152,	200	4263	-	24	Mozilla/5.0 (Windows NT 6.1; WO	norm	
GET	arenas de san juan	200	4494	http://almhuette-raith.at/ad	18	Mozilla/5.0 (X11; Linux i686) Apple	norm	
GET	19245	200	4263	-	5	Mozilla/4.0 (compatible; MSIE 6.0;	norm	
GET	2.07E+15	200	4494	http://almhuette-raith.at/ad	16	Mozilla/5.0 (compatible; bingbot/	norm	
GET	fennell	200	4263	-	7	Mozilla/5.0 (Windows NT 6.1; WO	norm	
GET	d50allecido	200	4494	http://almhuette-raith.at/ad	11	Mozilla/5.0 (compatible; AhrefsBo	norm	

GET	1'+(select 'ndpx' where 4061=4061 c	200	8600	http://stanmore.menczykow	135		anom
GET	-1591 where 3291=3291 union all se	200	6562	http://stanmore.menczykow	84		anom
GET	1")) or updatexml(1808,concat(0x2e	200	560659	-	114		anom
GET	1) as svkf where 6503=6503 or char(	200	646404	-	164		anom
GET	1%')) or 8514=benchmark(5000000,r	200	4852	-	49		anom
GET	-9240' union all select 6538,6538,65	200	4576	-	39		anom
GET	1))) or 2633=dbms_pipe.receive_m	200	174	http://stanmore.menczykow	95		anom
GET	1'||(select 'jpso' from dual where 9	200	11572	http://stanmore.menczykow	203		anom
GET	-6360') as yfrk where 3683=3683 uni	200	151	http://stanmore.menczykow	94		anom
GET	-1922%') union all select 2335,2335,	200	152	http://stanmore.menczykow	77		anom

So, we need to tokenize each word in url query field and use an NLP (Natural Language processing) model to convert text to number based on weights as machine learning model is based on numbers. We wil be using Word2Vec model to perform these steps.

```
df = pd.read_csv("D:/NewFile.csv")
df.head()

df.shape
tokenized = []
from nltk.tokenize import word_tokenize

for line in df["url query"]:
    tokenized.append(word_tokenize(line))


import gensim
from gensim.test.utils import common_texts
model = gensim.models.Word2Vec(tokenized,size=150, window=10, min_count=1, sg=1, workers=10)
w2v = dict(zip(model.wv.index2word, model.wv.syn0))
print(common_texts)

x = model.wv.index2word

A = []

for lst in tokenized:
    a = [0]*34862
    for token in lst:
        a[x.index(token)] = 1
    A.append(a)

import numpy as np
A = np.array(A)
At = np.transpose(A)
B = np.array(model.wv.syn0)
C = A.dot(B)
```

STEP 3-

After converting the text to numbers based on their importance, we have data set which contains data in the form of numerals. Now, we will be using Support Vector machine (SVM) classifier model as our ML model to perform learning.

Our new file after second step is renamed as payload_final.csv

```
dataset = pd.read_csv('D:/payload_final.csv')
y = dataset.iloc[:,7].values
from sklearn.preprocessing import LabelEncoder, OneHotEncoder

onehotencoder = OneHotEncoder(categorical_features = [0])

labelencoder_y = LabelEncoder()
y = labelencoder_y.fit_transform(y)

print(y)
```

We will be splitting our data into training set and test set

```
from sklearn.model_selection import train_test_split
C_train, C_test, y_train, y_test = train_test_split(C, y, test_size = 0.25, random_state=0)

from sklearn.svm import SVC
classifier = SVC(kernel = 'linear', random_state = 0)
classifier.fit(C_train, y_train)

y_pred = classifier.predict(C_test)
```

STEP 4 –

Once, the model has undergone learning, cross check the false positive ratio using confusion matrix

```
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_pred)
```

STEP 5 –

Start predicting the traffic based on our machine learning model as malicious or normal. This will provide us with outlier which many of the modern day tools perform.

```
y_prediction = classifier.predict()
```

These steps are the basic building block of any threat hunting / cyber security analytics tool which helps us to add some automation to our manual threat hunting process. I will soon be publishing the code in my github account as well - https://github.com/akky2892

Conclusion

There always seems to be a race involved between an attacker and defender/hunter to stay ahead in their respective responsibility and this game requires some assistance to stay ahead. One such assistance to threat hunter is what we have demonstrated. We need to be Gamer and not a Noob in this exercise.

This assistance is of utmost importance when there is recent type of attack which doesn't have any kind of signature or behavior description anywhere on the internet. Even after performing some manual searches, we are not able to identify the attack vectors and in such cases, machine learning plays an important part in our detection process.

That's how the Game of Threat hunting can be won !!!

References

- www.kaggle.com – *Place to do data science projects.*
- Machine Learning and security: Protecting systems with Data and Algorithms – *A book by Clarence Chio and David Freeman.*
- https://threathunting.net
- https://cyberwardog.blogspot.com/
- https://attack.mitre.org/
- https://medium.com/@cyb3rops
- https://vincentyiu.com

Ways to connect

https://twitter.com/akky2892 

https://www.youtube.com/channel/UCzJnWB-dI8DrLGM-LZ-4FWw 

https://in.linkedin.com/in/akash-sarode-b0193091