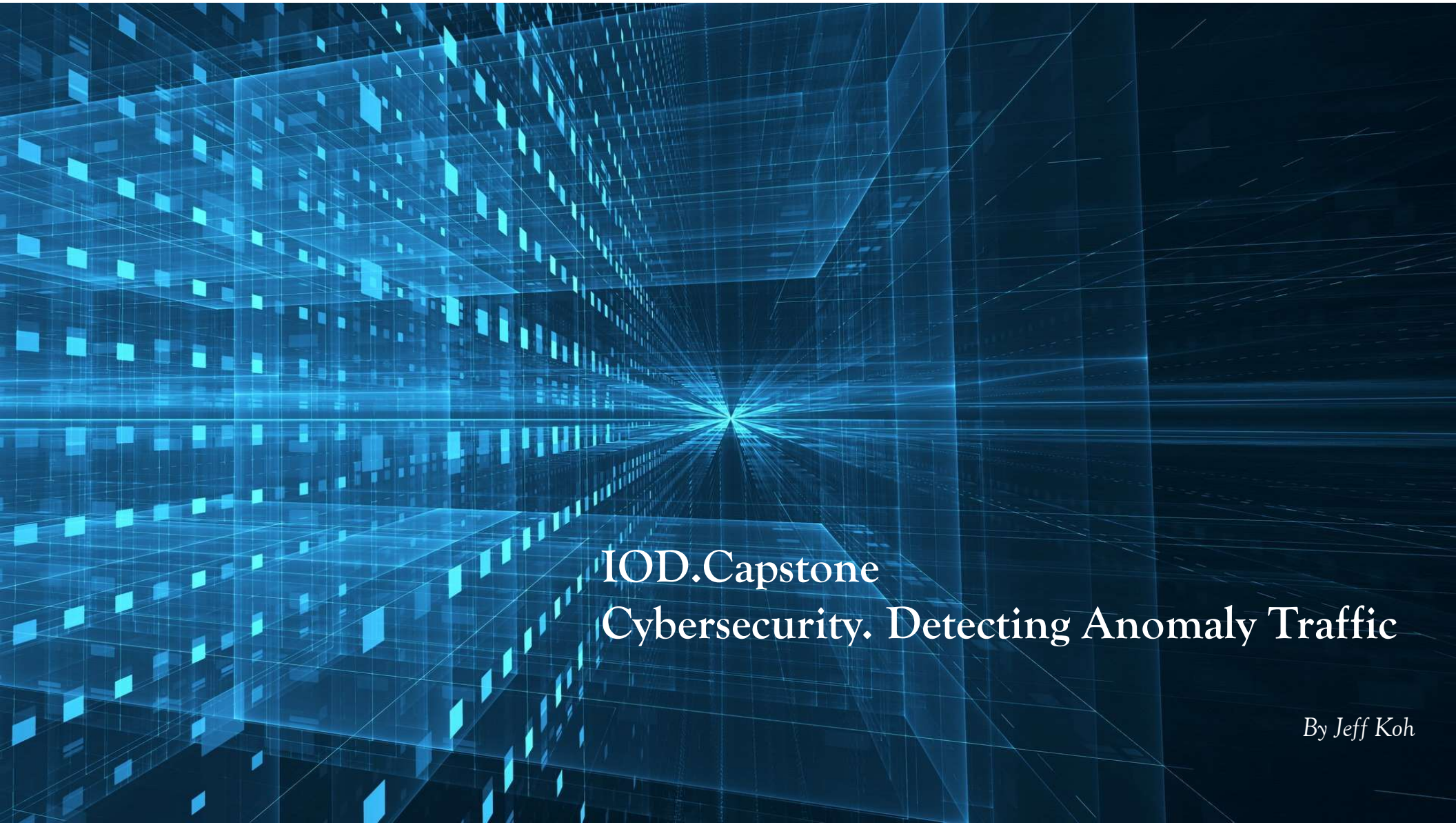


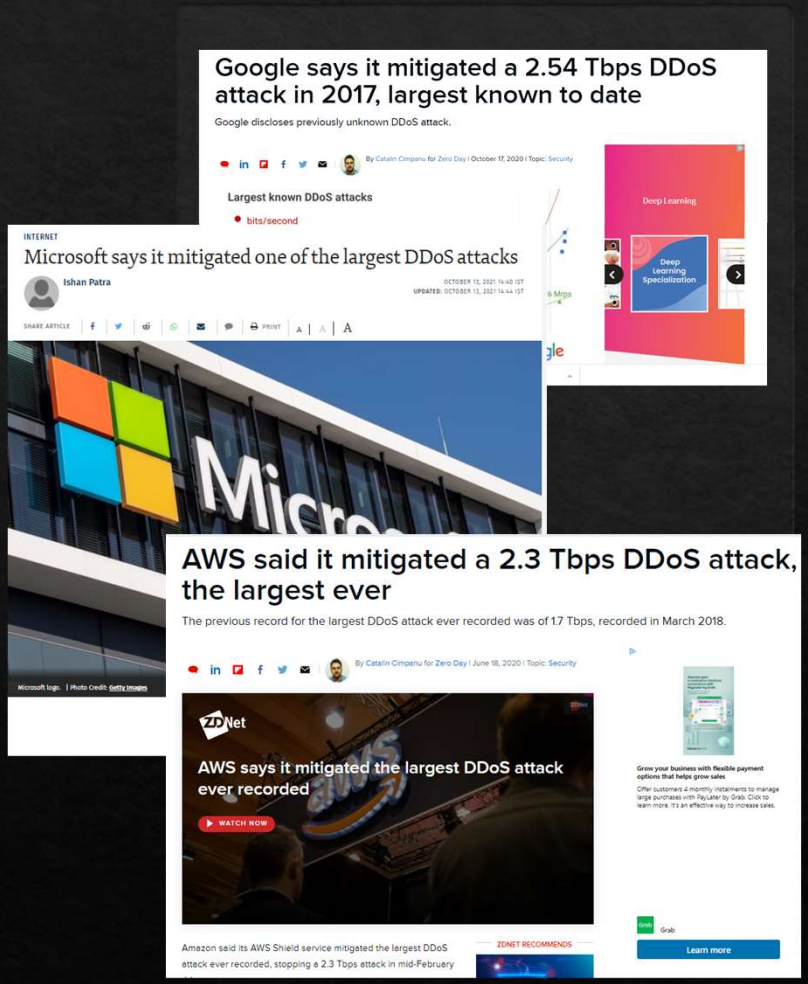
IOD.Capstone
Cybersecurity. Detecting Anomaly Traffic

By Jeff Koh

1. Introduction

- Cyber attacks are now so common, recent reports show that hackers attack a computer in the US every 39 seconds*! Once an attack happens millions of people could be harmed.
- State-run organizations can be shut down, services can't be provided to citizens. Businesses from small to medium to large corporation will be affected, millions of dollars can be lost, and it does not stop there, can be dreadful.
- Any government, organization, companies and individual needs to be aware of cybersecurity

* https://www.securitymagazine.com/articles/87787-hackers-attack-every-39-seconds

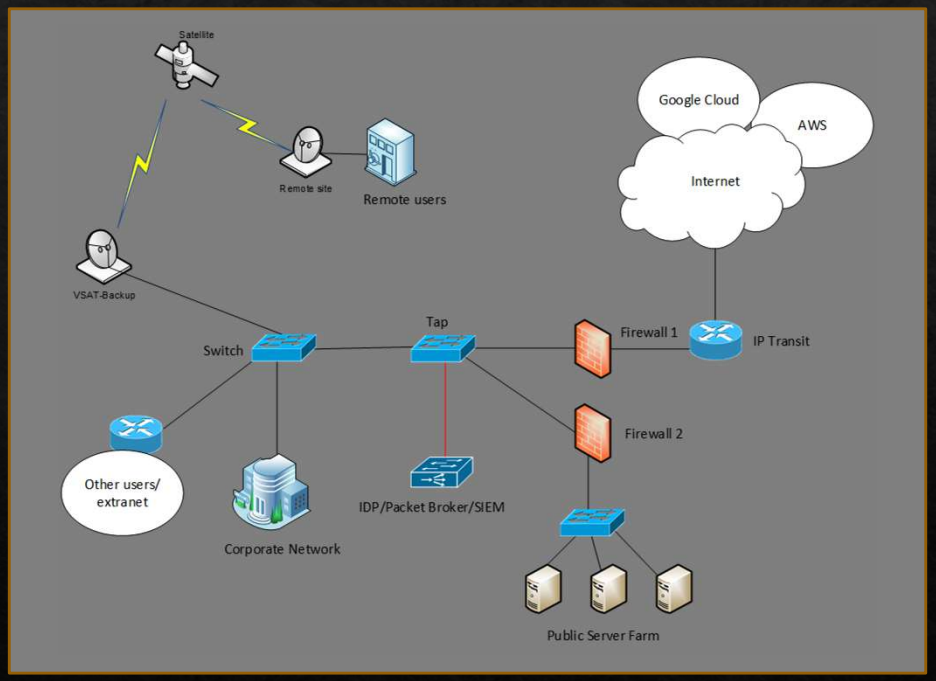


2. Business Requirement

- Company – Pluto, an Internet Service Provider was a victim of a recently DDoS attack. The attack causes part of their core network unable to provide Internet access. Their customers who rely on Internet selling services eg e-commerce, webhosting, DNS were disconnected, which amounting to heavy financial losses.
- Even though the attack has been mitigated and security ringfenced around the border, but the immensity of potential attack is always presence.
- They have engaged us to provide Cybersecurity consultancy. With AI and ML gaining popular in cybersecurity. They would like to see how these technology can help them in providing the additional layer of security.

3. Business Insight

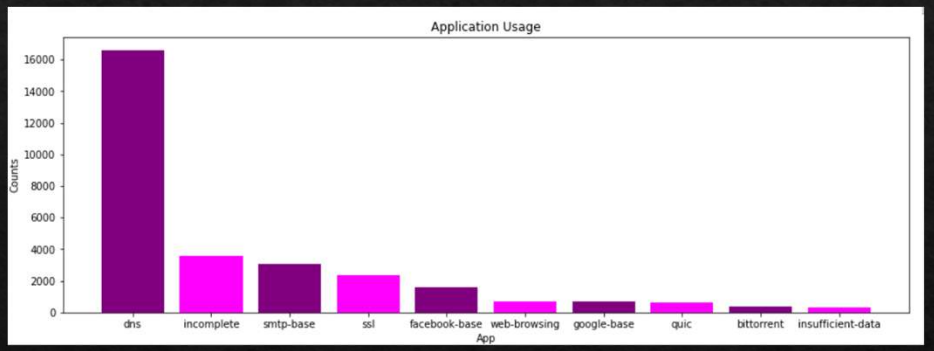
- The company provide multi services in their public server farm in different countries across this region.
- Each POP established connectivity with major cloud provider – Google, AWS via the local IP transit. The public Server farm provides caching, accounting services to their remote users.
- In addition, each POP is connected via dark fiber, VSAT back to their users and corporate network.
- We collated data from the offline Packet broker, SIEM log and filer out the unnecessary traffic. This data will be feed to the machine learning.



4. Implementation

- The data consist of 31,645 rows and 26 columns. Among these, the traffic capture records the source IP address, source port, destination IP address, destination port, bytes size which breaks into sent and received.
- First – we look at the number of application passing through the network. Plot out the top 10 application. From there we were able to see what frequent application, website our users accessing (Fig 1.)
- DNS topple the top, this is no surprise given that DNS provide name resolution to IP for access to any application, website. The remaining list shown application been commonly used within the network.
- We can drill into details of each individual application however, this is not what we want at this stage. The objective is to look at anomaly traffic - **identifying rare events or observations which can raise suspicions by being statistically different from the rest of the observations.**
- Lets dive deep into "incomplete" feature, which is the second highest in the list.

Fig 1



5. Discovery

- Based on the top 10 source IP address we observed there the distribution of packet sizes different significantly. Over 80% of the source address does not receive Bytes Received (Fig 2).
- Over 80% of Bytes Sent are below 200 bytes. A device that conducting attack, such as a TCP SYN flood is trying to open as many connection request as possible. The objective is to exhaust the victim's server/ PC resources. Thus an attacker will want to keep the size of the packet as small as possible.
- Now, lets drill down on the first source address – 10.255.164.60, let see what we can find.
- We can see that this source (Fig 3) is sending to different destination address using port 25 with a typical bytes of 66 and 264 but with no bytes received! Port 25 is a smtp, it is used to send out email to other email server.
- This source is either sending out mass email marketing or something fishy (Fig 4).

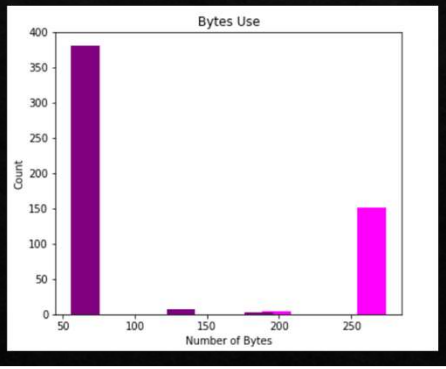
Fig 2

Source address	Application	Bytes Sent	Bytes Received	
10.255.164.60	incomplete	66	0	373
182.54.147.14	incomplete	66	0	206
10.255.166.95	incomplete	186	66	158
10.255.164.60	incomplete	264	0	152
10.255.173.23	incomplete	66	0	142
182.54.147.45	incomplete	246	126	124
10.255.174.96	incomplete	66	0	106
10.255.173.23	incomplete	264	0	80
182.54.147.45	incomplete	186	126	72
203.21.141.10	incomplete	66	0	63

Fig 3

Source address	Destination address	Destination Port	IP Protocol	Bytes Sent	Bytes Received
10.255.164.60	8.31.233.70	25	tcp	264	0
10.255.164.60	17.178.102.78	25	tcp	264	0
10.255.164.60	199.59.148.144	25	tcp	66	0
10.255.164.60	52.24.249.124	25	tcp	66	0
10.255.164.60	185.79.118.151	25	tcp	264	0
10.255.164.60	34.209.30.116	25	tcp	66	0
10.255.164.60	8.31.233.27	25	tcp	66	0
10.255.164.60	52.72.197.234	25	tcp	264	0
10.255.164.60	210.128.48.74	25	tcp	66	0
10.255.164.60	34.217.152.37	25	tcp	66	0

Fig 4



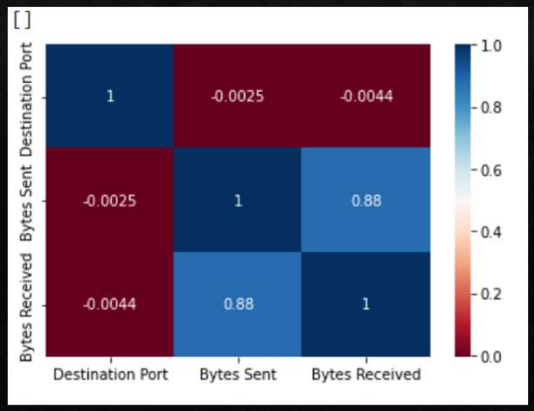
6. Machine Learning

- We will create a label tag to identify this source IP.
- Grouping 4 features into one feature – Application, Destination Port, Bytes Sent, Bytes Received. And creating a target with label as the feature (Fig 5).
- It will be interesting to see if there is any correlation ship between the 4 features, as we can see here in fig 6, there is none, which telling us these are independent. There is no connection related to each other.

Fig 5

	Application	Destination Port	Bytes Sent	Bytes Received
31640	smtp-base	25	306	401
31641	incomplete	25	264	0
31642	incomplete	25	66	0
31643	incomplete	25	66	0
31644	smtp-base	25	579	587

Fig 6



6. Machine Learning

- Before we decide to use the machine learning algorithm. We created 3 cluster to identify the grouping base on destination port and bytes sent (Fig 7).
- We tested 4 machines learning algorithms to distinguish normal traffic and suspicious traffic:
  - K-nearest neigbhors (KNN).
  - Logistic Regression (LR).
  - MLP Classifier (MLP).
  - Decision Tree (DT).

Fig 7

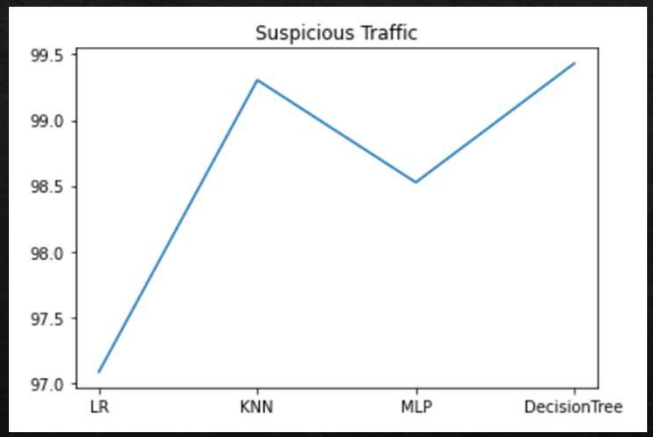
	Destination Port	Bytes Sent
0	53	106
1	53	102
2	53	95
3	53	95
4	53	91
...	...	...
31640	25	306
31641	25	264
31642	25	66
31643	25	66
31644	25	579

31645 rows × 2 columns

6. Machine Learning

- We trained these 4 classifier on a training set of 70% of the combine of normal and suspicious traffic. We calculated the accuracy of using the remaining traffic as test set (30%). The accuracy range from 97% to 99%.
- The decision tree classifier performed well, achieving an accuracy of 99.4% suggesting that the data can be segmented in a higher feature.

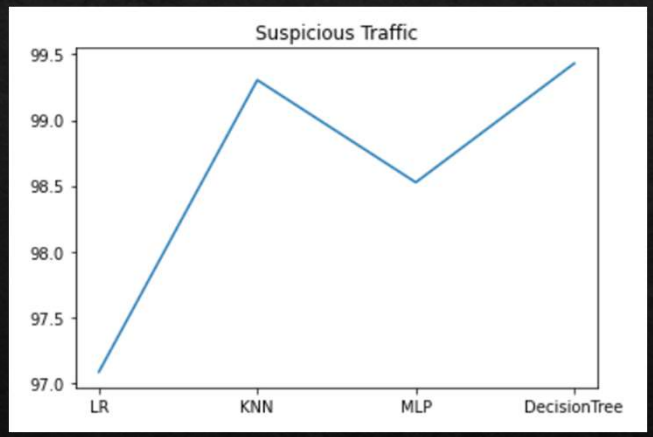
Fig 8



7. Future work

- We have seen how the 4 machine learning algorithms works and the accuracy of it. This based on the selected features available in the data collected.
- We could look at protocol level eg how much TCP or UDP are going through the network.
- Timescale, what time and how long does the suspicious traffic start and end.
- All these will provide additional understanding of the traffic, provide re-active with pro-active data response to the end users.

Fig 8



Thank you.