

Big Data Security Analytics: Opportunities and Issues

December 12th, 2019

Ing. Giovanni Apruzzese

PhD Candidate in Information and Communication Technologies

✉ giovanni.apruzzese@unimore.it

🌐 <https://weblab.ing.unimo.it/people/apruzzese>

The background is a dark teal gradient. In the four corners, there are white line-art illustrations of circuit traces and nodes, resembling a printed circuit board (PCB) layout. These lines are thin and connect to small white circles representing nodes or components.

Part 1

Introduction

CONTEXT



- **Cyber threats are on the rise...**



More than **4 billion** records compromised in 2016
→ a 566% increase from 2015

- **...they become more advanced...**



Some examples of recent **cyber attacks**:

- BlackEnergy (2015)
- MEDJACK (2016)
- Archimedes (2017)
- Wannacry (2017)
- Meltdown & Spectre (2018)

- **...and the penalties are steep**

\$3.6 Million avg cost of a data breach

CONTEXT

- On average, it takes **191 days** to identify a threat, and **66 days** to triage it
- At the same time, the volume of generated data **is exploding**

A medium-sized enterprise can easily produce **TBs** of daily network traffic data

CONTEXT

Example

Graph of internal communications
(**real data** from department of large organization)

Assumptions

Only **client-to-server** and **server-to-client** communications are legit

Clients and **servers** are easy to distinguish by analyzing traffic

Low number of internal communications

Reality

Many legit **client-to-client** communications (Windows NetBIOS, Dropbox, Skype), and also **server-to-server** communications (e.g., to DNS and storage servers)

Many **clients expose legitimate services** (e.g., SSH server), **servers are often used as clients** (e.g., through SSH or as proxies)

Many internal communications:
~ **10M per day** in a single department

Final objective:
To identify the **one or few host** that are performing malicious activities

SOLUTION

- **(Big Data) Security Analytics**

Definition: process of using data collection, aggregation, and analysis tools for security monitoring and threat detection

EVOLUTION OF SECURITY ANALYTICS

1995-2000 (SEM)

- Focus on network security
- Event filtering and basic correlation
- Single layer inspection
- Log management and retention
- Events per second: <5000
- Storage: Gigabytes
- Manual breach response
- Limited scalability

2005-2014 (SIM)

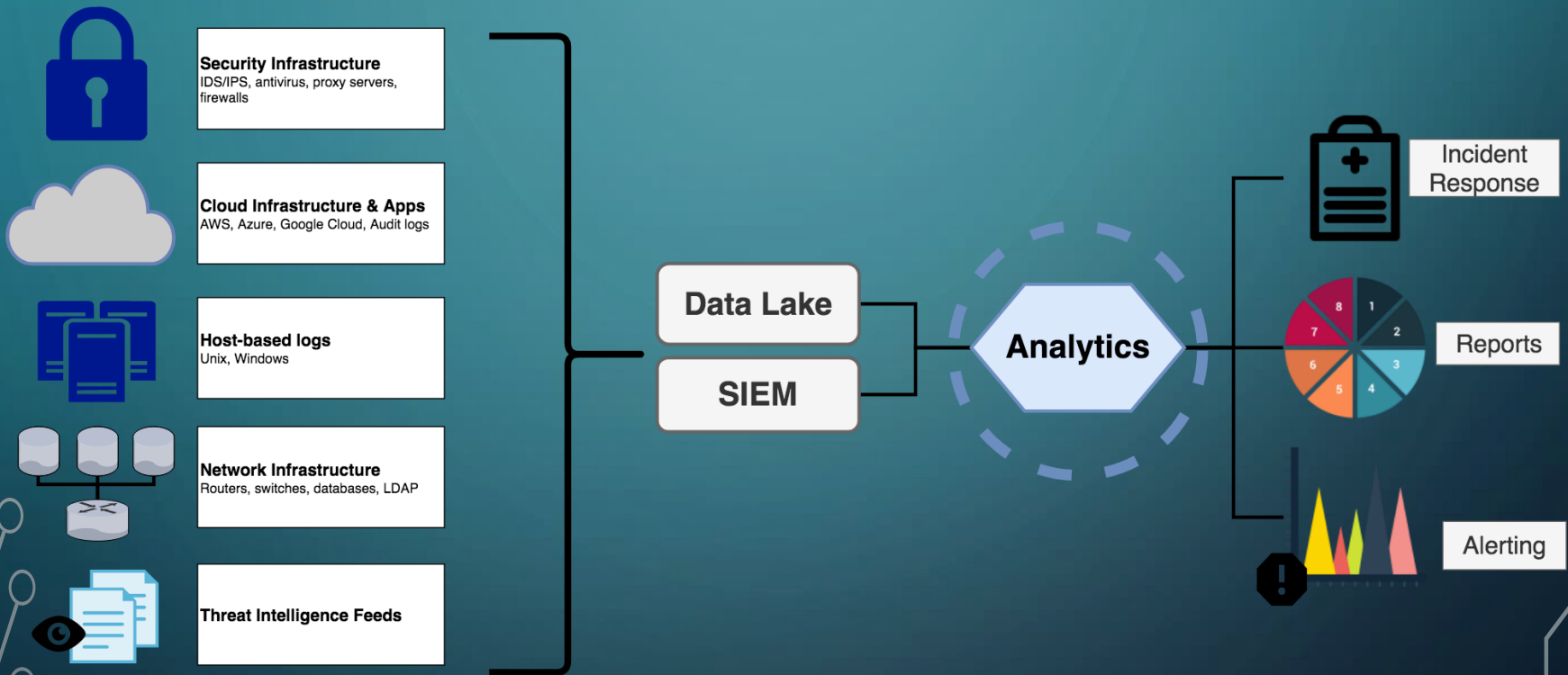
- Reporting
- Information sources: various log formats
- Advanced correlation
- Signature-based alerting
- Increasing devices: >1000
- Events per second: >10000
- Storage: Terabytes
- Focus on threat detection and response, breach response slow, dependent on security analyst skills

2014+ Security Analytics

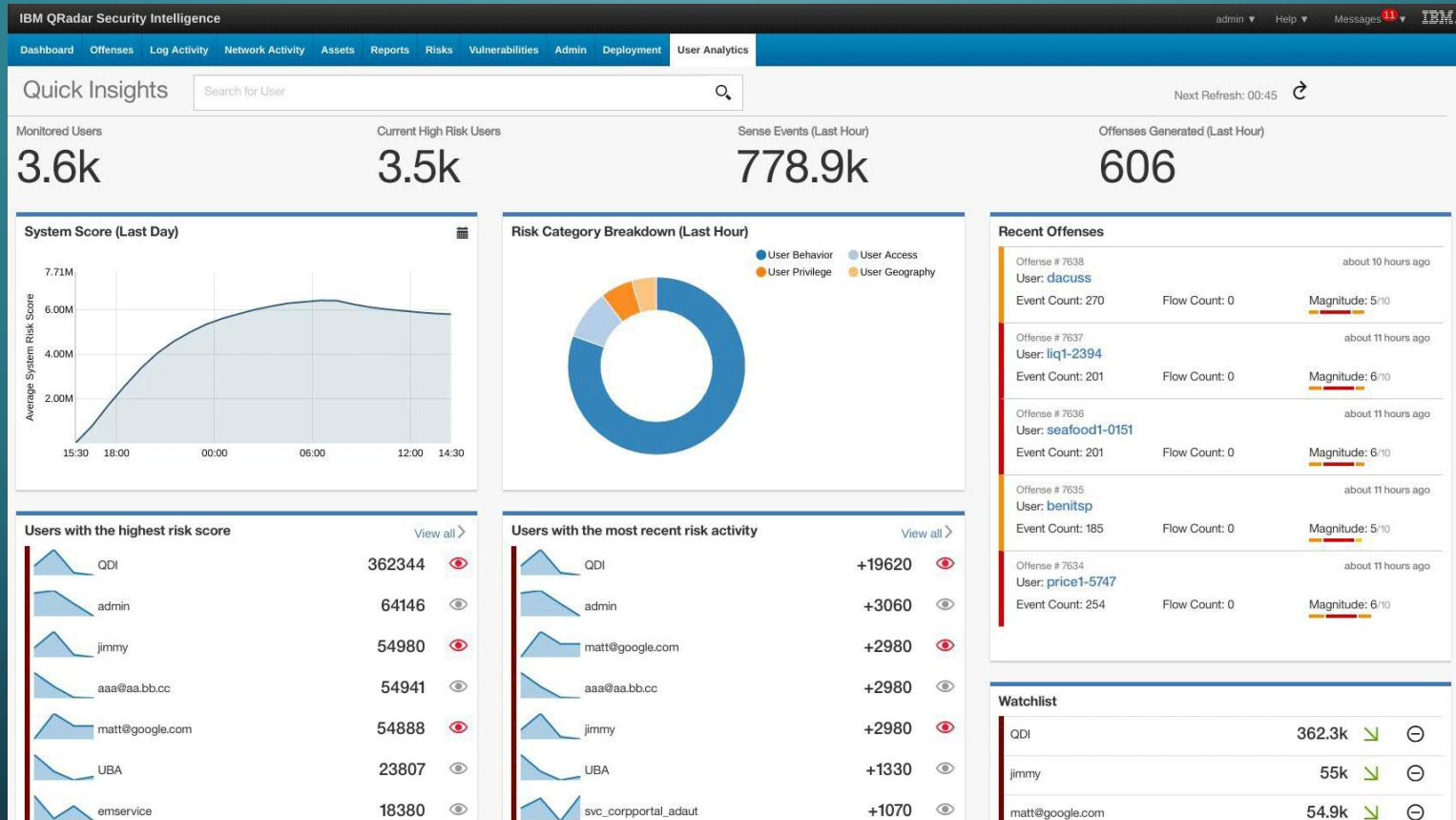
- Feeds from applications, databases, endpoints
- Threat detection
- Advanced analytics with additional security context
- **User and network** behavior
- Heterogeneous data: **Netflow**, threat intelligence feeds, multiple log sources
- Huge number of devices: >5000
- Events per second: >100000
- Storage: Petabytes
- Near real-time breach response

Sophistication, volume, velocity, scalability, complexity

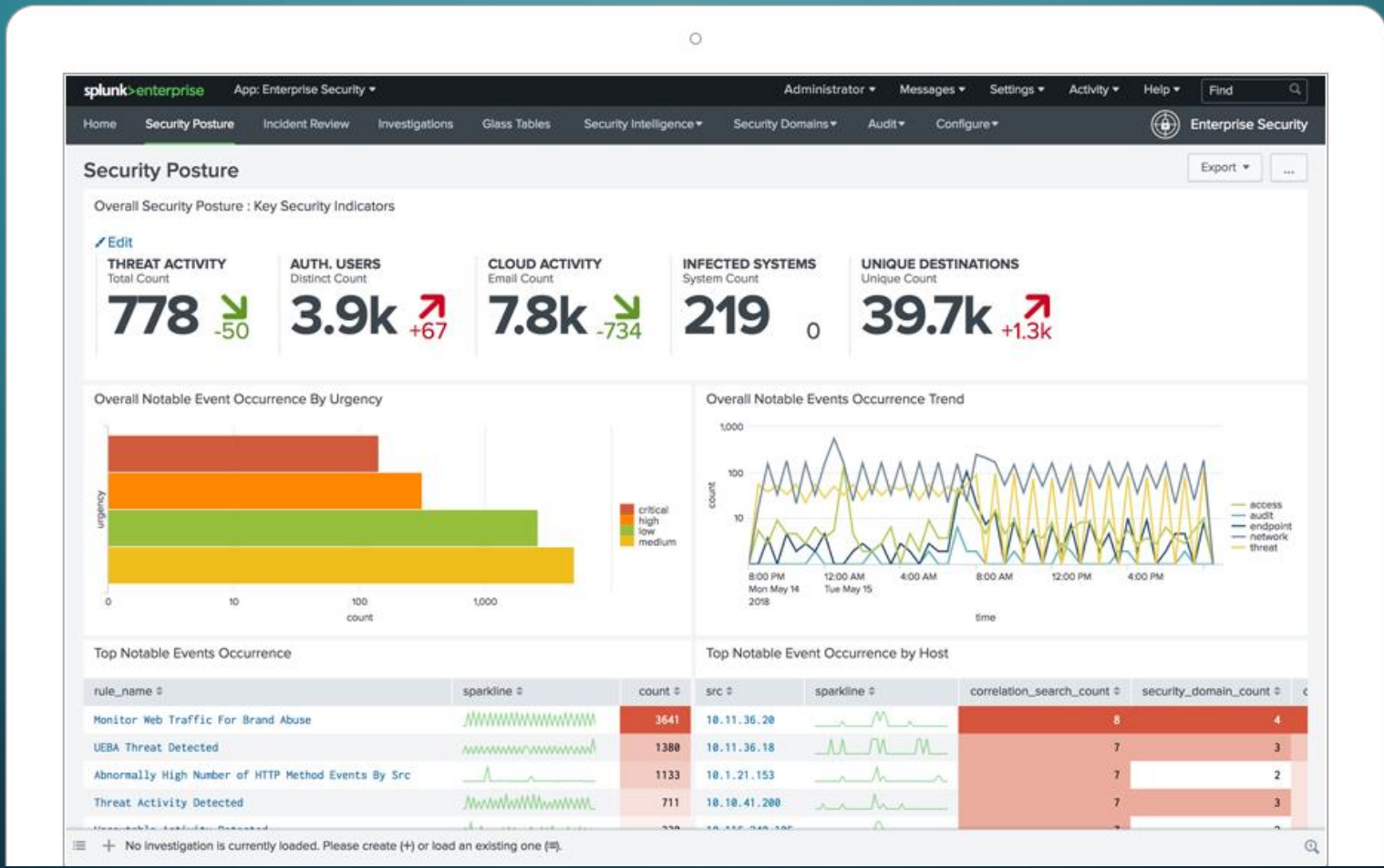
STATE-OF-THE-ART SECURITY ANALYTICS



EXAMPLES: QRADAR



EXAMPLES: SPLUNK



EXAMPLES: APACHE SPOT

localhost:8889/files/ui/flow/suspicious.html#date=2016-07-08

Open Network Insight :: Retflow :: Suspicious

IP: 0.0.0.0 Data Date: 2016-07-08

Suspicious

Rank	Time	Source IP	Destination IP	Source Port	Destination Port	Protocol	Input Packets
0	2016-07-08 0:31	172.30.0.46	10.0.0.183	52234	119	UDP	213454
1	2016-07-08 17:16	10.13.77.49	172.10.0.40	47131	80	TCP	206
2	2016-07-08 14:56	10.13.77.49	172.10.0.3	35579	25	TCP	112
3	2016-07-08 15:10	10.70.68.127	172.30.0.4	6395	80	TCP	278

Network View

A network graph visualization showing a complex web of connections between nodes. The nodes are represented by blue diamonds and yellow circles, connected by lines. The graph is dense and multi-directional, with a central hub-and-spoke structure.

Notebook

Source IP:	Dest IP:	Src Port:	Dst Port:
- Select -	- Select -	- Select -	- Select -
172.30.0.46	10.0.0.183	52234	119
10.13.77.49	172.10.0.40	47131	80
10.70.68.127	172.10.0.3	35579	25
172.30.0.70	172.30.0.4	6395	0
10.200.20.2	172.10.0.2	55759	3840
10.138.235.111	172.20.0.3	61783	3389
10.78.100.150	10.0.4.16	0	81
123.151.42.61	10.0.5.25	46032	808
10.10.11.102	172.30.0.3	3247	22
10.17.15.10	172.30.0.2	61471	21

Quick IP scoring

Rating: 1 2 3

Score Save

Details

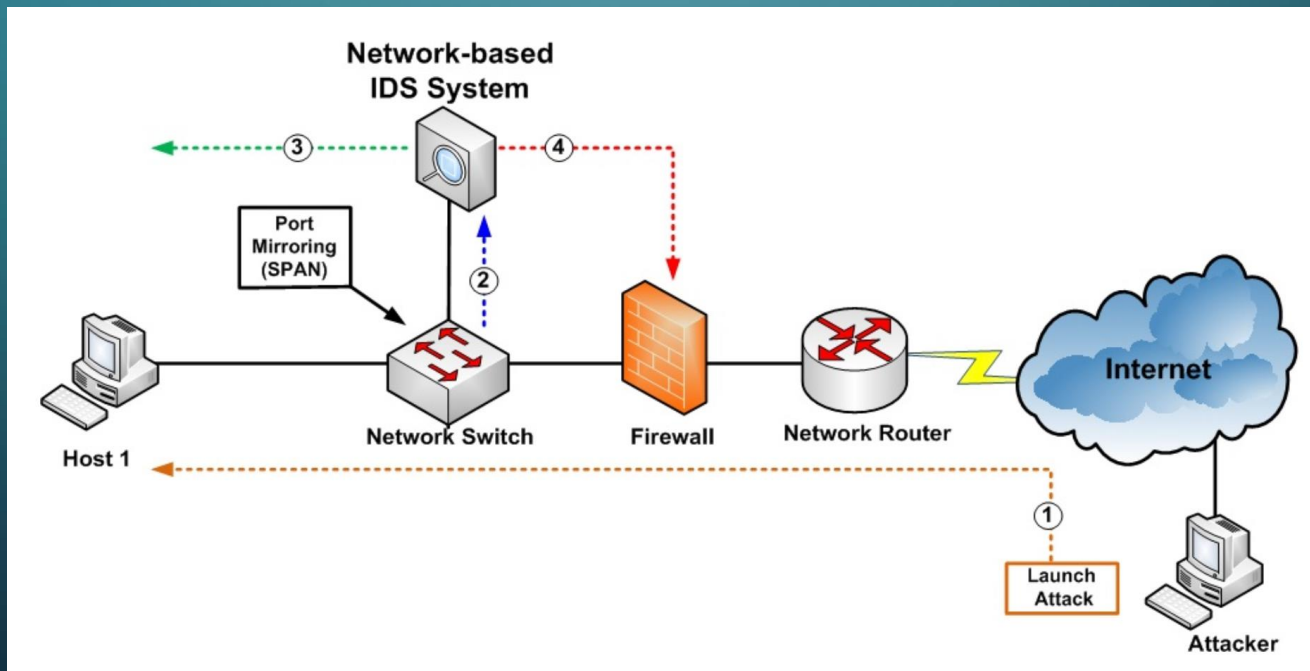
A pie chart visualization showing the distribution of data. The chart is divided into four segments: a large red segment, a large blue segment, a smaller green segment, and a very thin white segment. The segments are separated by yellow borders.

BRIEF RECAP

Intrusion Detection System (IDS)

Host-based
Intrusion Detection System
(HIDS)

Network-based
Intrusion Detection System
(NIDS)



BRIEF RECAP

Network Traffic – Full Packet Capture (PCAP)

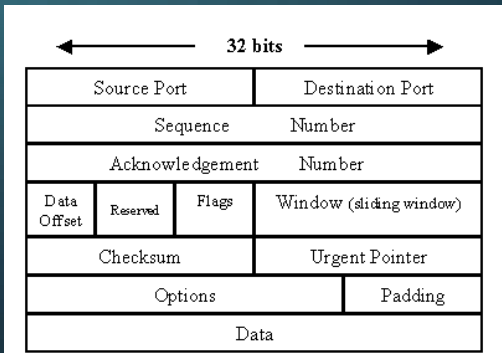
The screenshot shows the Wireshark interface with a packet capture on the eth0 interface. The packet list pane shows several packets, including ARP requests and DNS queries. The packet details pane is expanded to show the structure of a frame, including Ethernet II, ARP, and IP headers. The packet bytes pane shows the raw hex and ASCII data of the captured frame.

No.	Time	Source	Destination	Protocol	Info
46	139.931187	wistron_07:07:ee	Broadcast	ARP	who has 192.168.1.254? tell 192.168.1.68
47	139.931463	ThomsonT_08:35:4f	Wistron_07:07:ee	ARP	192.168.1.254 is at 00:90:d0:08:35:4f
48	139.931466	192.168.1.68	192.168.1.254	DNS	Standard query A www.google.com
49	139.975406	192.168.1.254	192.168.1.68	DNS	Standard query response CNAME ww.l.google.com A 66.102.9.99
50	139.976811	192.168.1.68	66.102.9.99	TCP	62216 > http [SYN] Seq=0 Win=8192 Len=0 MSS=1460 WS=2
51	140.079578	66.102.9.99	192.168.1.68	TCP	http > 62216 [SYN, ACK] Seq=0 Ack=1 Win=5720 Len=0 MSS=1430
52	140.079583	192.168.1.68	66.102.9.99	TCP	62216 > http [ACK] Seq=1 Ack=1 Win=65780 Len=0
53	140.080278	192.168.1.68	66.102.9.99	HTTP	GET /complete/search?hl=en&client=suggest&js=true&q=m&cp=1 H
54	140.086765	192.168.1.68	66.102.9.99	TCP	62216 > http [FIN, ACK] Seq=805 Ack=1 Win=65780 Len=0
55	140.086921	192.168.1.68	66.102.9.99	TCP	62218 > http [SYN] Seq=0 Win=8192 Len=0 MSS=1460 WS=2
56	140.197484	66.102.9.99	192.168.1.68	TCP	http > 62216 [ACK] Seq=1 Ack=805 Win=7360 Len=0
57	140.197777	66.102.9.99	192.168.1.68	TCP	http > 62216 [FIN, ACK] Seq=1 Ack=806 Win=7360 Len=0
58	140.197811	192.168.1.68	66.102.9.99	TCP	62216 > http [ACK] Seq=806 Ack=2 Win=65780 Len=0
59	140.219210	66.102.9.99	192.168.1.68	TCP	http > 62218 [SYN, ACK] Seq=0 Ack=1 Win=5720 Len=0 MSS=1430

Frame 1 (42 bytes on wire, 42 bytes captured)
Ethernet II, Src: Vmware_38:eb:0e (00:0c:29:38:eb:0e), Dst: Broadcast (ff:ff:ff:ff:ff:ff)
Address Resolution Protocol (request)

```
0000  ff ff ff ff ff ff 00 0c 29 38 eb 0e 08 06 00 01  .... )8.....
0010  08 00 06 04 00 01 00 0c 29 38 eb 0e c0 a8 39 80  .... )8.....9.
0020  00 00 00 00 00 00 c0 a8 39 02  .... 9.
```

Example: TCP Packet

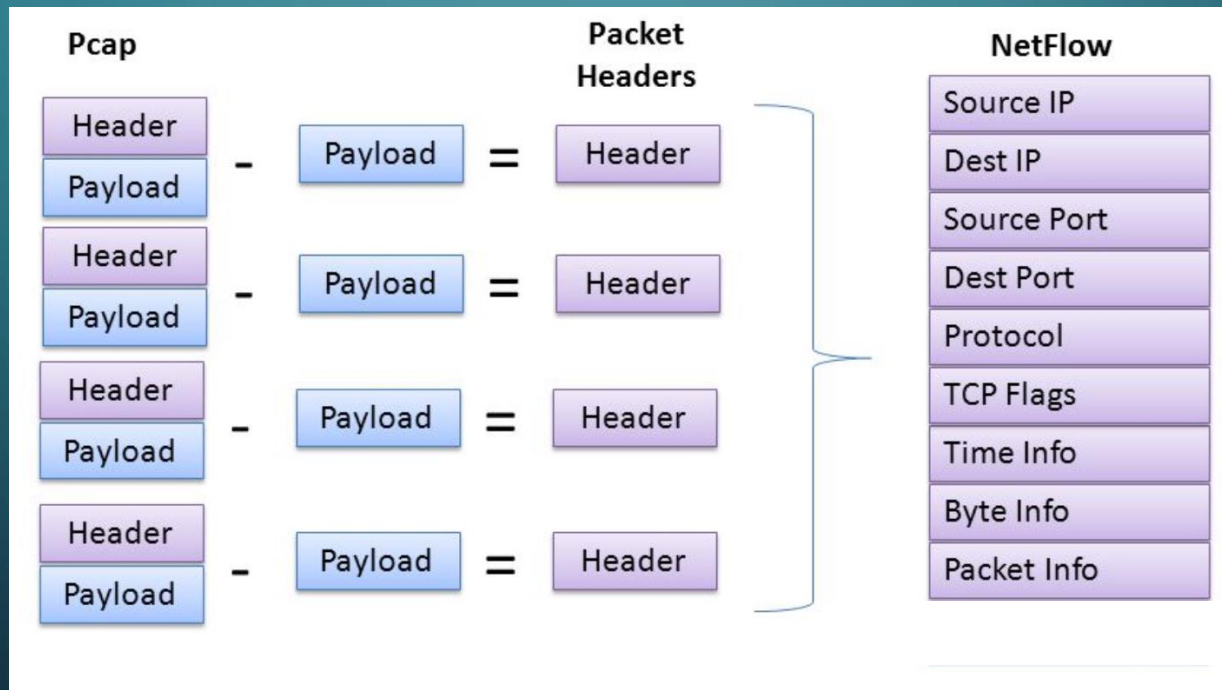


BRIEF RECAP

Network Traffic – Network Flow (NetFlow)

Network flow: **sequence** of packets that share:

- Source IP address
- Destination IP address
- IP protocol
- Source port
- Destination port
- IP Type of Service (ToS)

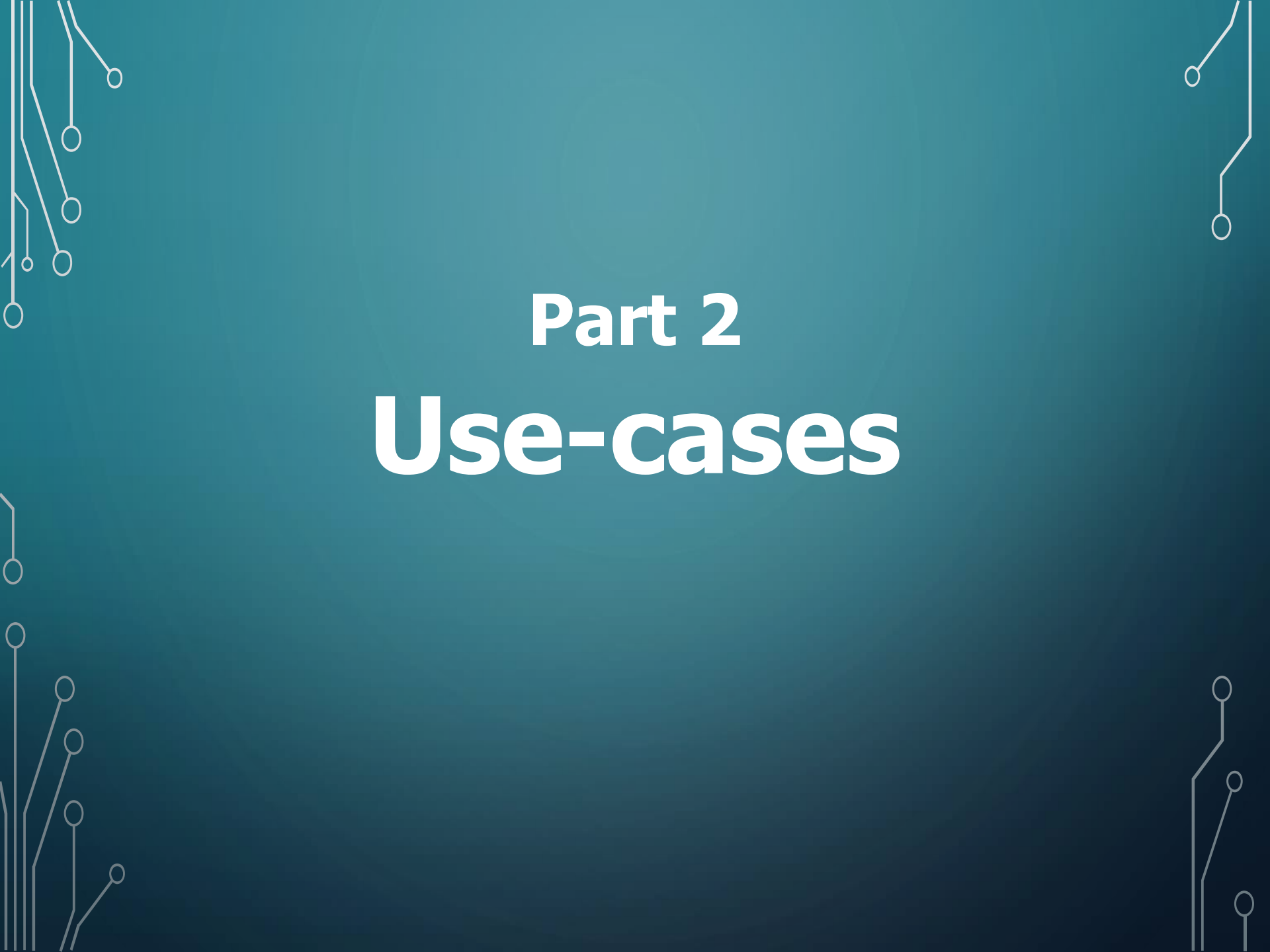


REMINDER

Analysis



Analytics

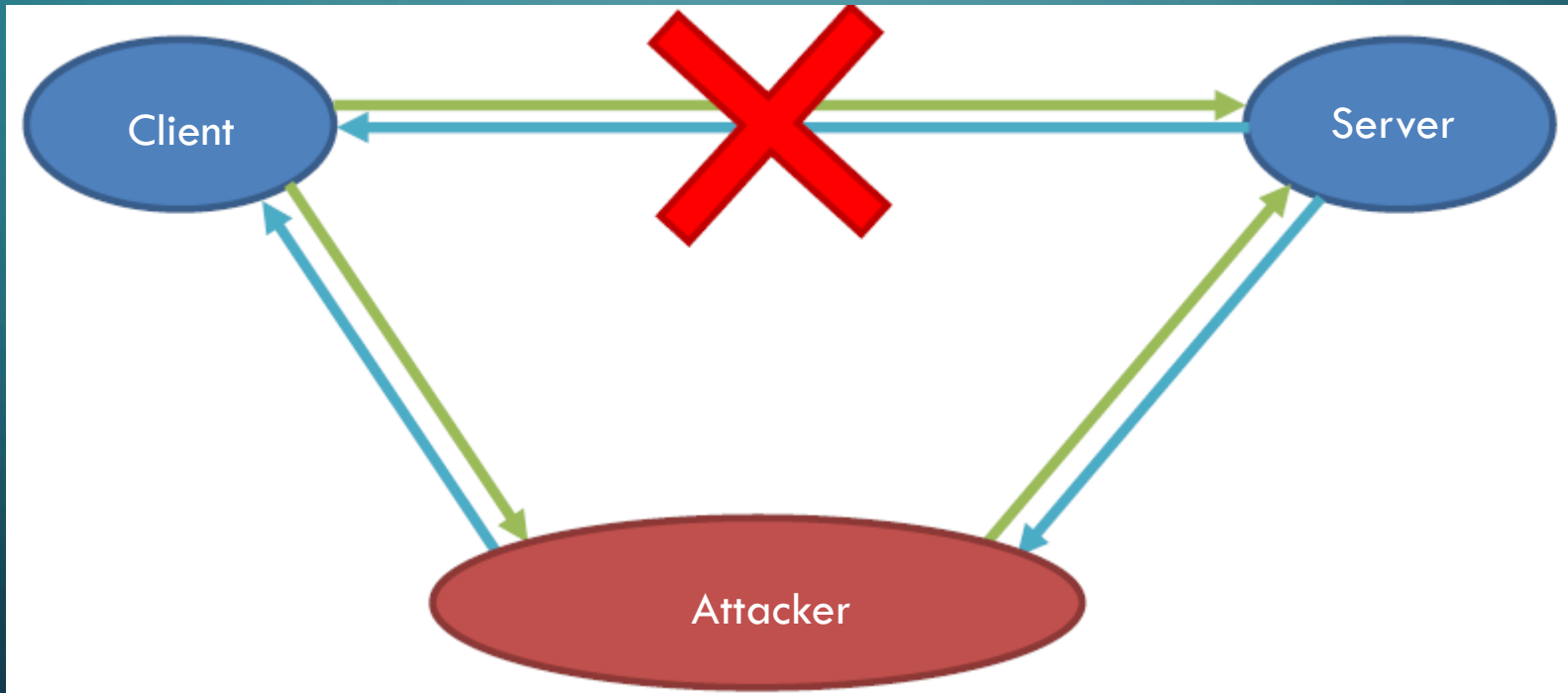
The background is a dark teal gradient. In the four corners, there are white line-art illustrations of circuit traces and nodes, resembling a printed circuit board layout. These elements are positioned in the top-left, top-right, bottom-left, and bottom-right corners, framing the central text.

Part 2

Use-cases

MAN-IN-THE-MIDDLE

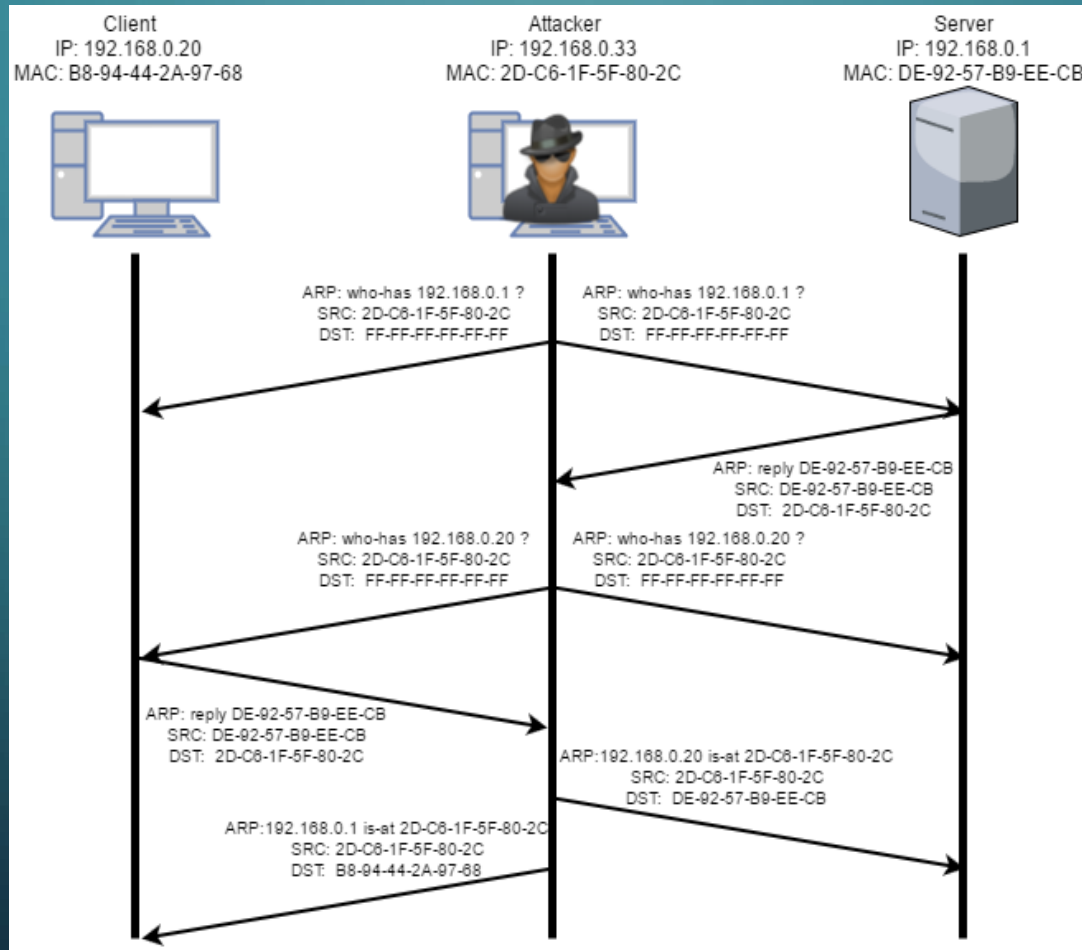
through *ARP Spoofing*



MAN-IN-THE-MIDDLE

through *ARP Spoofing*

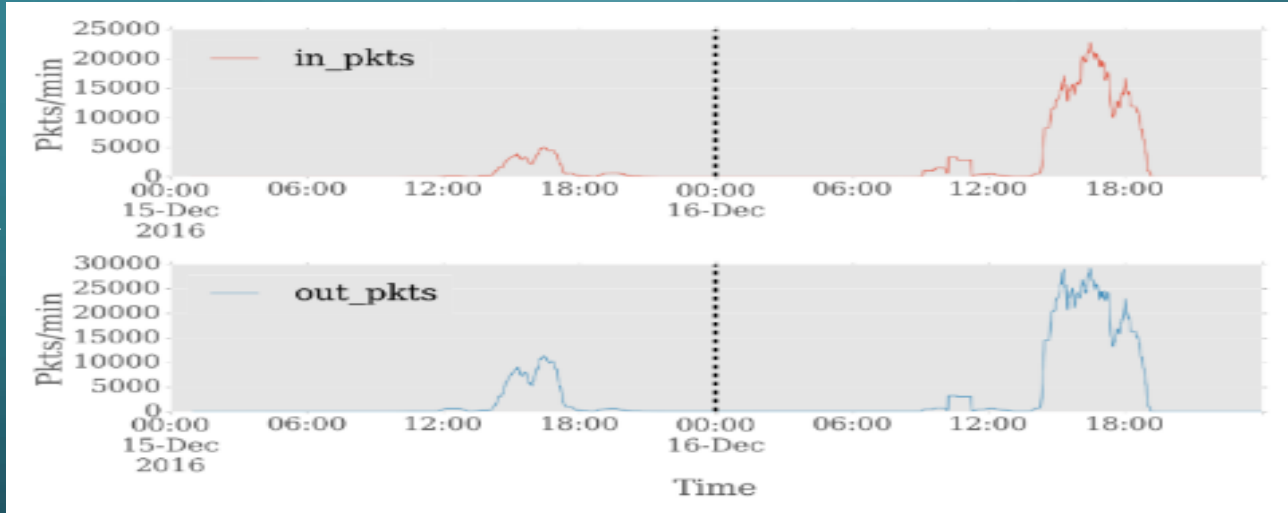
Step-by-step



MAN-IN-THE-MIDDLE

through *ARP Spoofing*

Intuition: all packets are doubled!



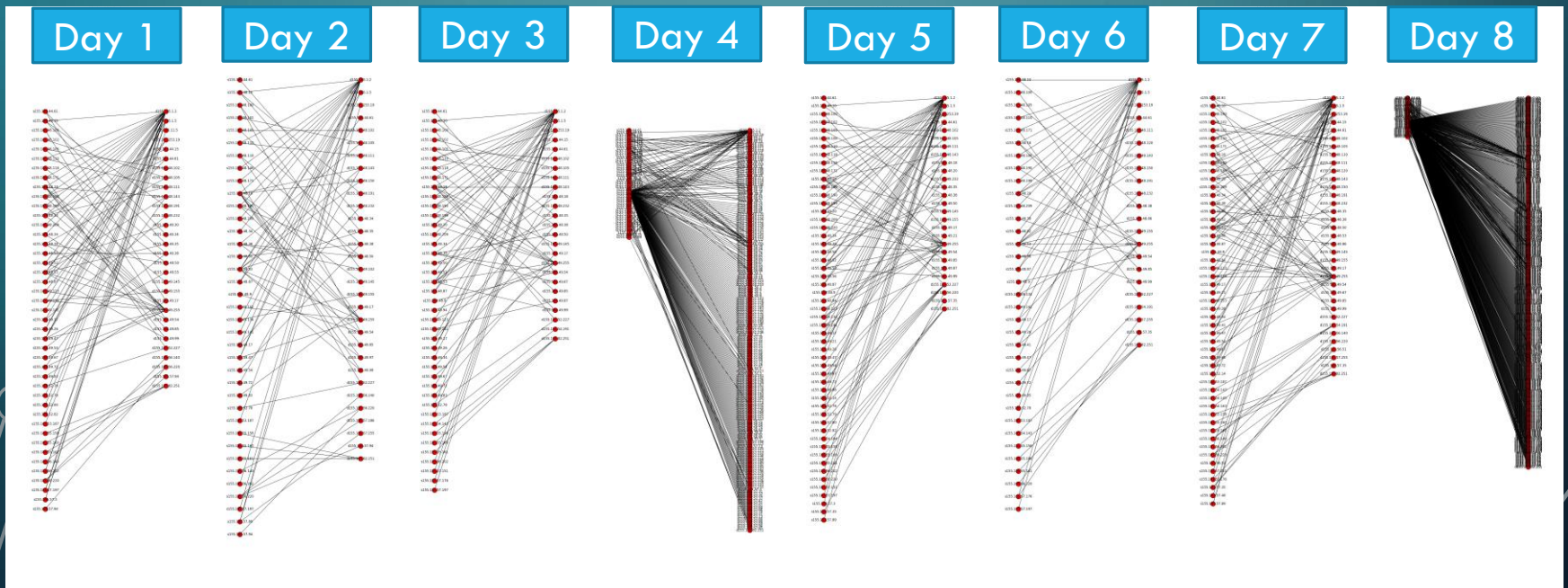
HOWEVER!

To avoid false positives that correspond to an increased network activity, we need to check in the ARP logs if the the IPs of Server and Client have been associated to a new MAC (possibly corresponding to the attacker)

RECONNAISSANCE

through *horizontal port-scanning*

```
$nmap -p80 192.168.0.0/24
```



RECONNAISSANCE

through *horizontal port-scanning*

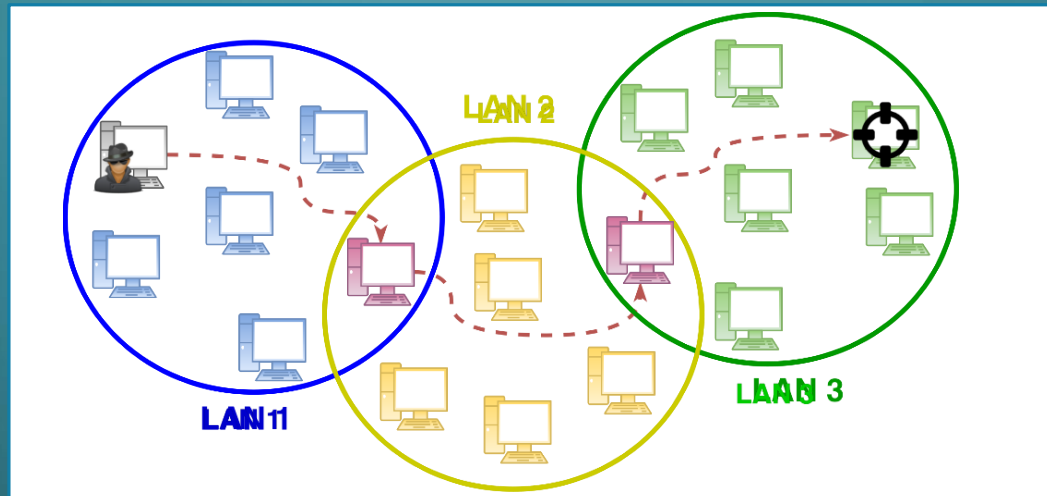
Intuition: the *average duration* of the scanner-host's connections decreases, while the *number of flows* and *contacted hosts* increase.



LATERAL MOVEMENT

through *Pivoting*

Attackers want to control hosts with
higher privileges or more valuable data.



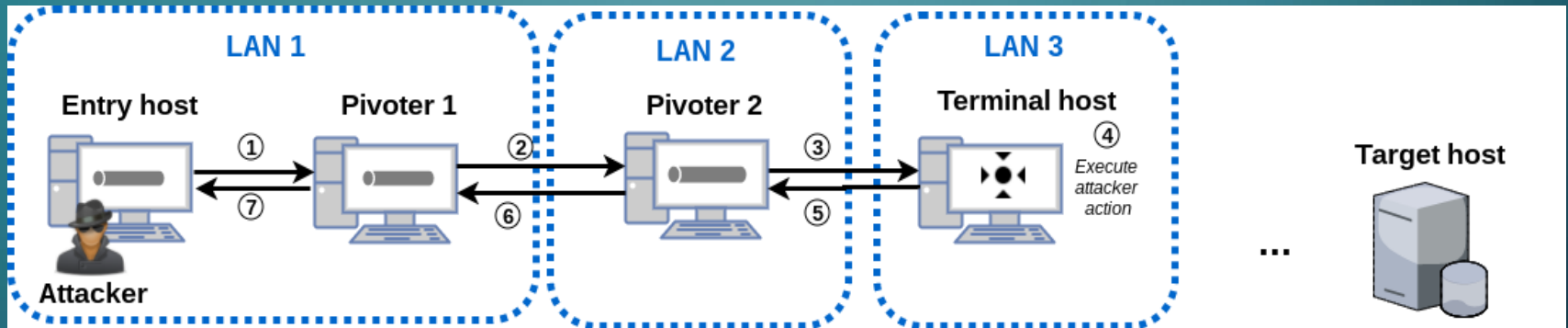
Pivoting: any action in which a *command propagation tunnel* is created among three or more hosts

NB: Pivoting activities are not necessarily malicious.

LATERAL MOVEMENT

through *Pivoting*

Pivoting example



Intuition: pivoting activities can be modelled through *Flow-sequences*

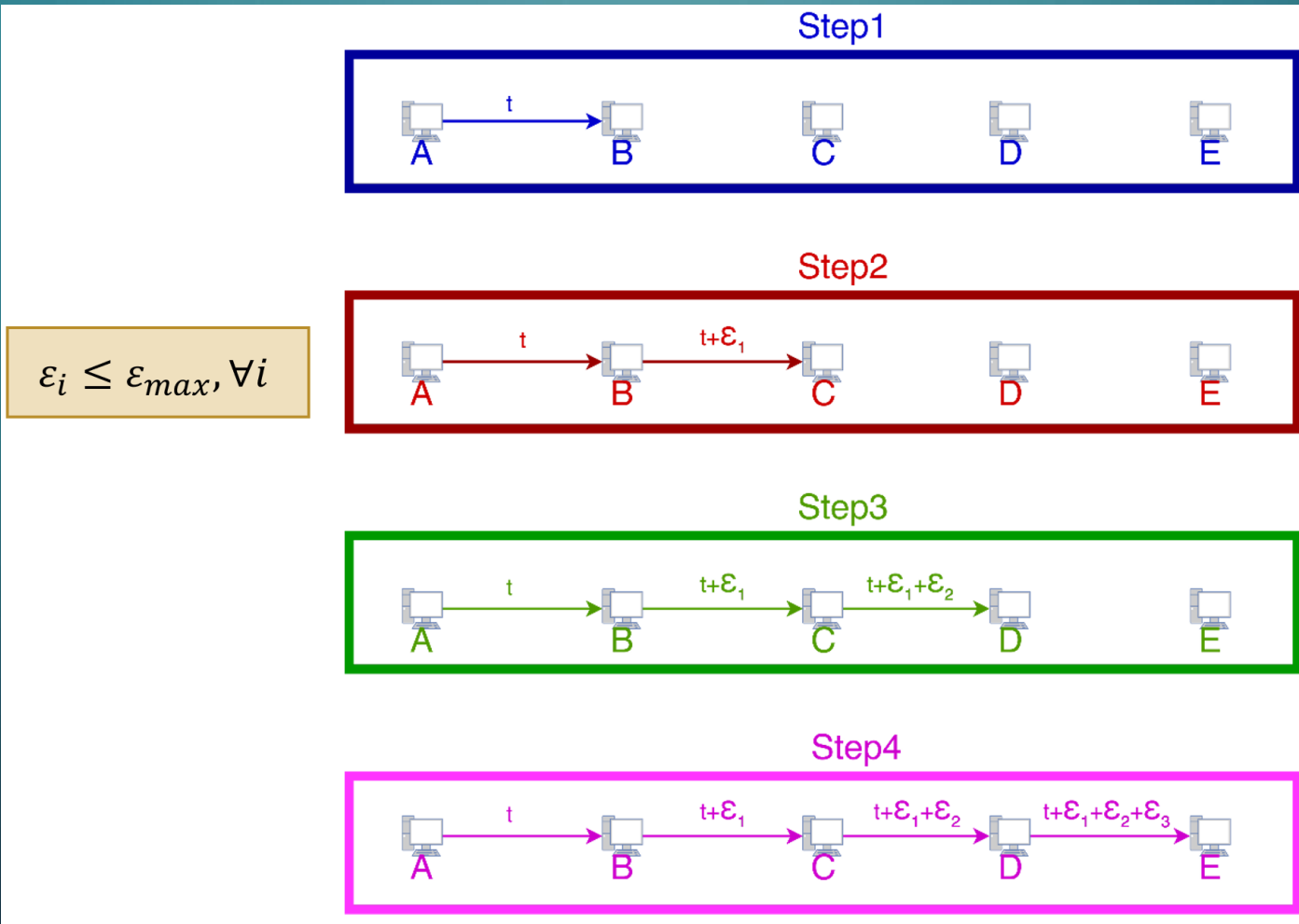
Flow-sequence

Ordered set of flows where consecutive flows are:

- Chronologically ordered
- Separated by at most ϵ_{max} time units
- Adjacent
- Not cyclical

LATERAL MOVEMENT

through *Pivoting*



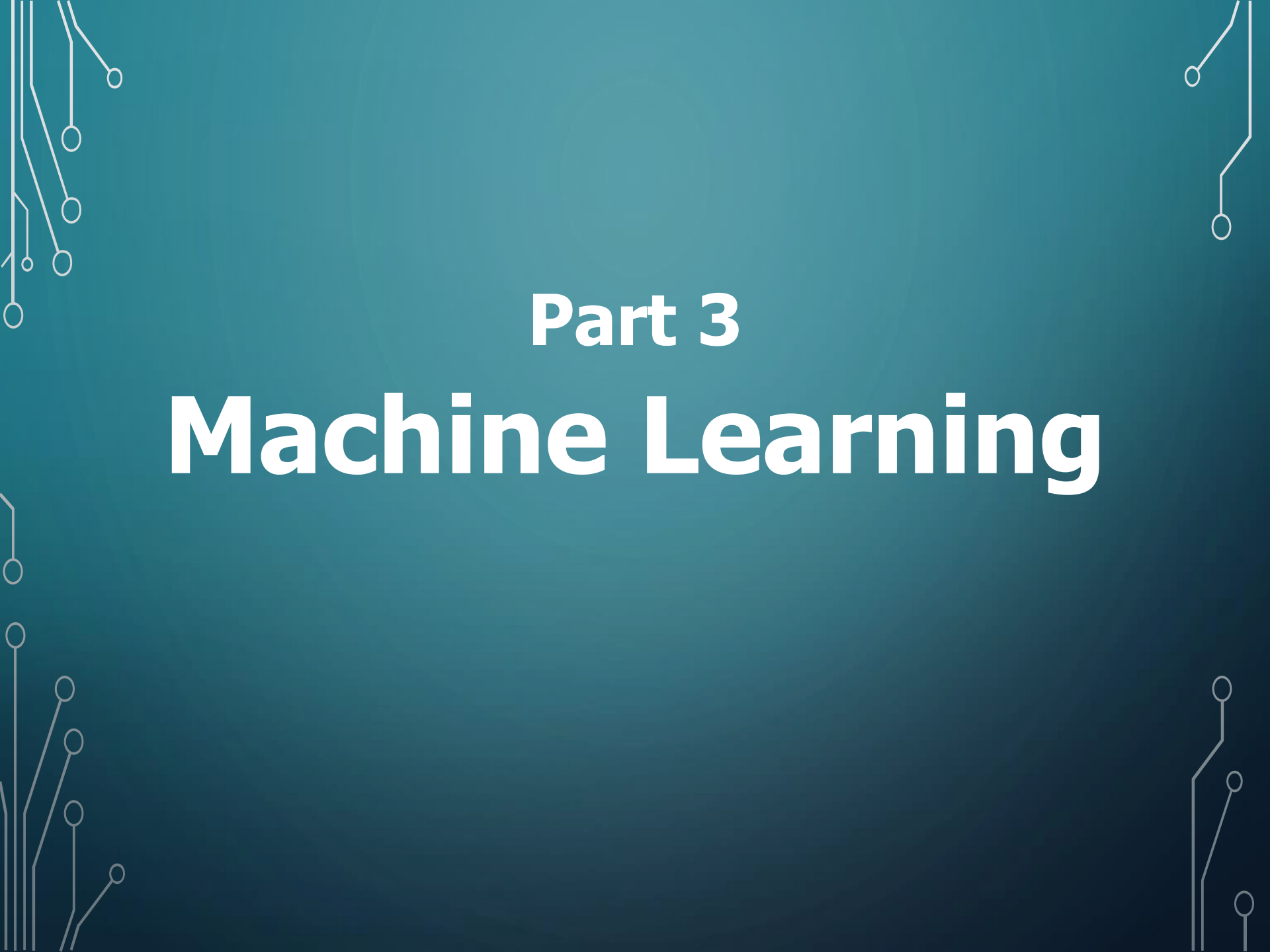
LATERAL MOVEMENT

through *Pivoting*

- Reminder: pivoting activities are not necessarily malicious
- Need to discriminate between “benign” and “malicious” pivoting

Intuition: Rank the detected pivoting activities on the basis of threatening characteristics displayed

- Characteristics that can be considered:
 - Novelty of the pivoting activity
 - Prior-reconnaissances
 - Usage of uncommon Ports
 - LANs involved
 - Anomalous Data Transfers

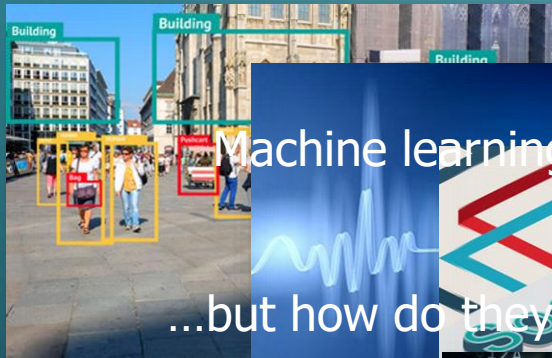
The background is a dark teal gradient. In the four corners, there are white, stylized circuit board traces. These traces consist of straight lines that turn at right angles and terminate in small circles, resembling electronic components or connection points.

Part 3

Machine Learning

MACHINE LEARNING

The popularity of machine learning is skyrocketing.



Machine learning algorithms are effective...

...but how do they behave for **cyber security**?



MACHINE LEARNING & CYBERSECURITY



FortiGuard Artificial Intelligence (AI) Delivers Proactive Threat Detection at Machine Speed and Scale



Machine learning moves to the front lines of defense against an expanding threat surface.

Machine Learning: New Frontiers in Advanced Threat Detection



Sophos Adds Advanced Machine Learning to Its Next-Generation Endpoint Protection Portfolio



MACHINE LEARNING HELPS US FIND NEW ATTACKS



Machine learning in Kaspersky Endpoint Security 10 for Windows



The truth is Trend Micro has been using machine learning since 2005.



MACHINE LEARNING PREVENTS PRIVILEGE ATTACKS AT THE ENDPOINT

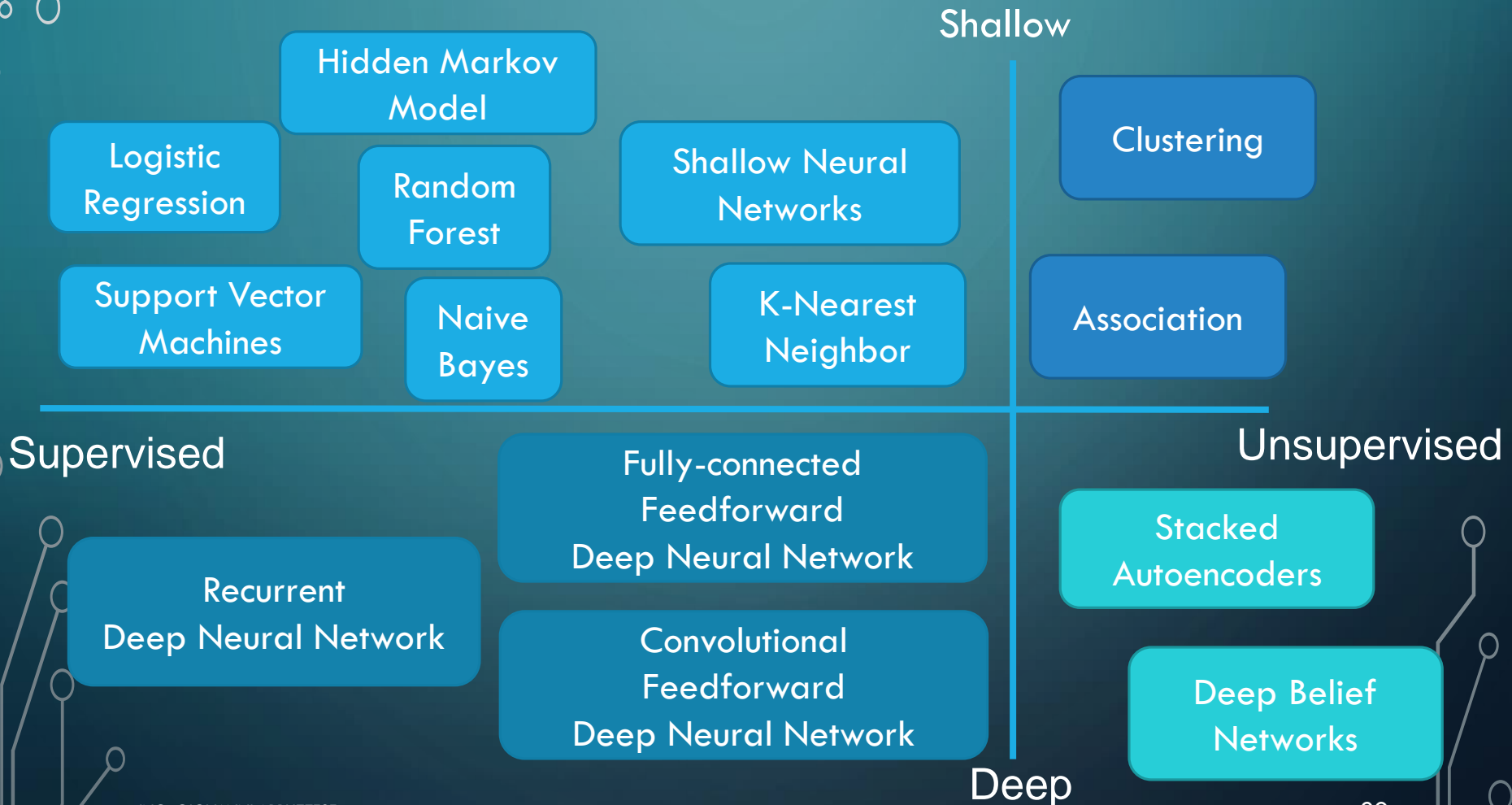


Rapid7 Attacker Behavior Analytics Brings Together Machine Learning and Human Security Expertise



MACHINE LEARNING & CYBERSECURITY

Lots and lots of algorithms...



MACHINE LEARNING & CYBERSECURITY

Several criticalities

Model training

- Where and how to find high quality and labeled training dataset?

Model deployment

- Is a pre-trained model applicable to my environment?

Model evaluation and selection

- How to compare different ML approaches?

Evolution over time (concept drift)

- How frequently should the model be re-trained?

Explainability

- Results are not explainable (yet)

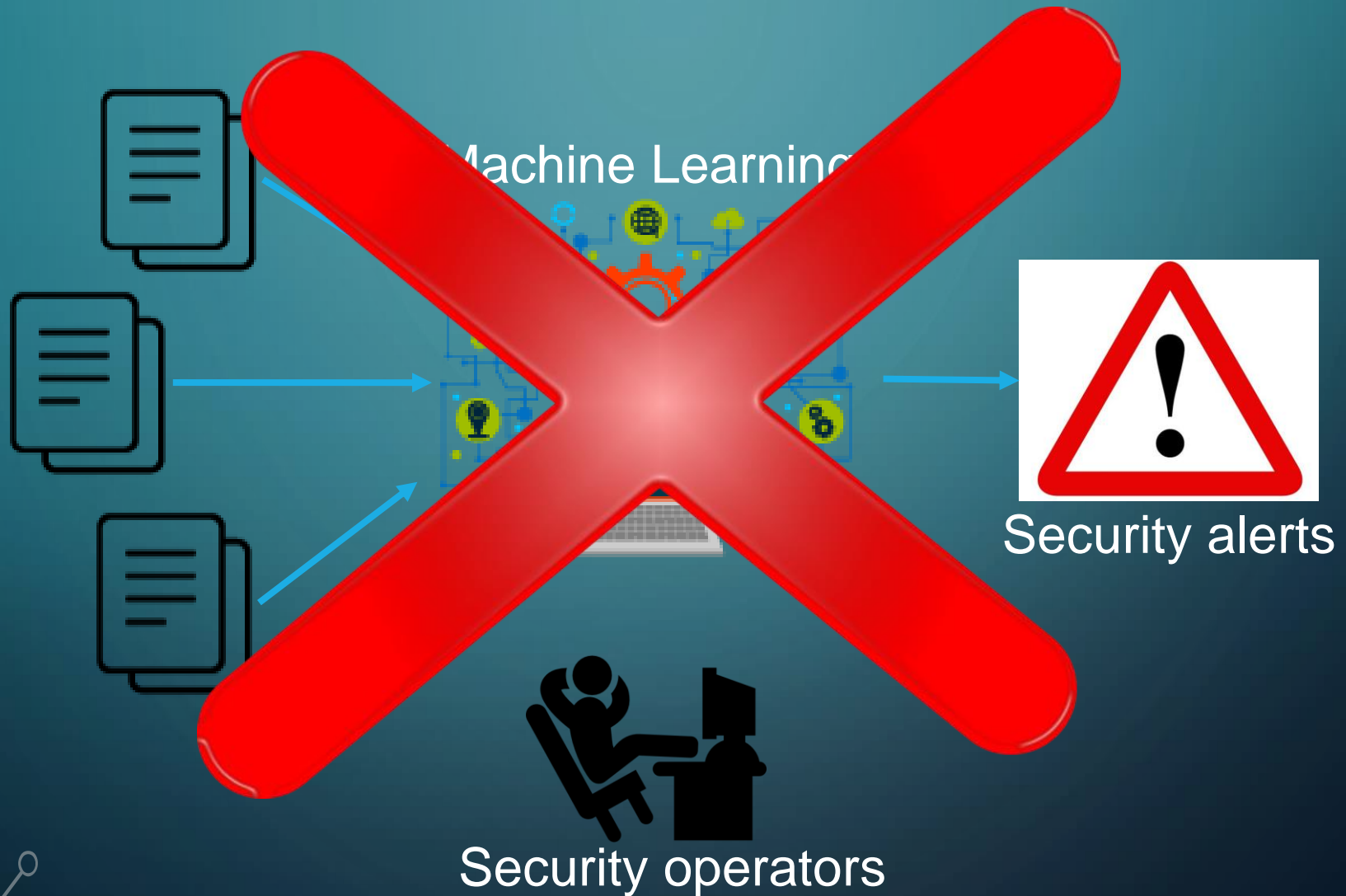
False positives and false negatives

- 1% false positive rate in large organization = **thousands** of daily false alarms

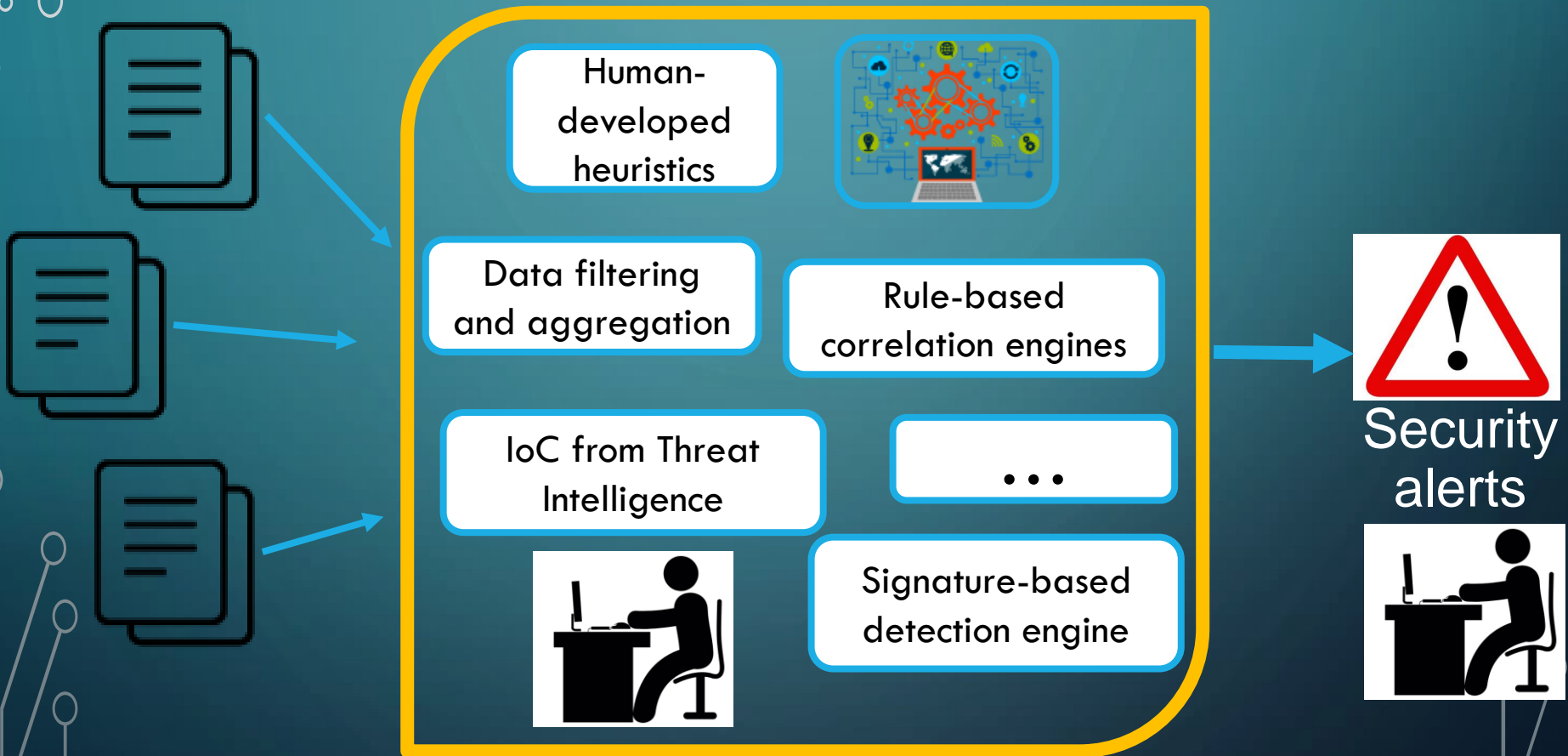
Adversarial attacks

- More on this later...

MACHINE LEARNING & CYBERSECURITY



MACHINE LEARNING & CYBERSECURITY



MACHINE LEARNING & CYBERSECURITY

Use-case:

Identifying malicious hosts involved in periodic communications

The defense of large information systems is still based on Network Intrusion Detection Systems (**NIDS**)

NIDS are currently affected by **two major issues**:

1. **Incapability of detecting all attacks**
2. **Excessive amount of info generated**

Necessity to **support the security analyst** with:

- **Automatic and timely security analyses**
- **Concise information**
- **Knowledge of ongoing novel attack variants**

MACHINE LEARNING & CYBERSECURITY

Our focus

External hosts performing *beaconing* activities

Intuition: Periodic activities tend to be more malicious

Goal

Graylist of external hosts with high likelihood of maliciousness

MACHINE LEARNING & CYBERSECURITY

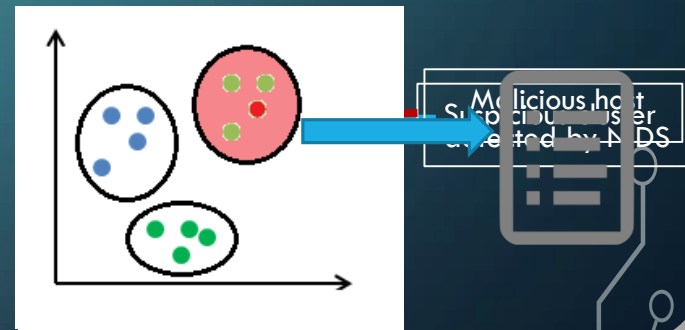
Novel malware variants are likely to evade detection...
...but some features of malware behavior persist and are shared even by novel variants

External hosts behaving similarly to a known malicious external host are likely to also be malicious

USE ONE TO FIND MANY:

- Generate clusters of similar communications
- Use NIDS alerts to find malicious external hosts
- Label as suspicious all clusters containing malicious external hosts
- Build *graylist* with external hosts belonging to suspicious clusters

Network communications



MACHINE LEARNING & CYBERSECURITY

Results for 7 days of traffic inspection in a large organization

Day	External hosts	External hosts with periodic behavior	External hosts in graylist	Malicious hosts in graylist	Malicious hosts detected by NIDS
1	296 943	3139	127	19 (14,96%)	3 (2,36%)
2*	105 884	2284	90	17 (18,89%)	3 (3,33%)
3*	89 283	2123	70	6 (8,57%)	3 (4,29%)
4	298 241	3194	31	3 (9,68%)	3 (9,68%)
5	314 313	3288	120	17 (14,17%)	4 (3,33%)
6	249 768	3044	119	7 (5,58%)	3 (2,52%)
7	258 439	3034	115	15 (13,04%)	4 (3,48%)

Much more manageable!

QUESTION

We showed several use-cases of CyberDetection:

- Man in the Middle
- Reconnaissance
- Lateral Movement
- Periodic Communications

If you were an *attacker*, what would you do against these detection schemes?

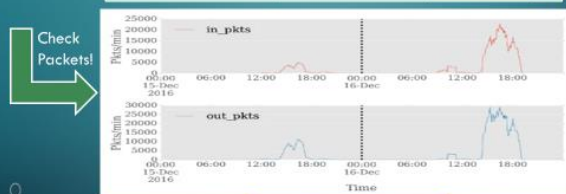
QUESTION

If you were an *attacker*, what would you do against these detection schemes?

MAN-IN-THE-MIDDLE

through *ARP Spoofing*

Intuition: all packets are doubled!



HOWEVER!

To avoid false positives that correspond to an increased network activity, we need to check in the ARP logs if the IPs of Server and Client have been associated to a new MAC (possibly corresponding to the attacker)

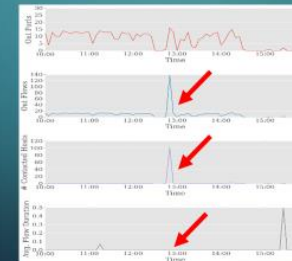
ING. GIOVANNI APRUZZESE

19

RECONNAISSANCE

through *horizontal port-scanning*

Intuition: the average duration of the scanner-host's connections decreases, while the number of flows and contacted hosts increase.



ING. GIOVANNI APRUZZESE

21

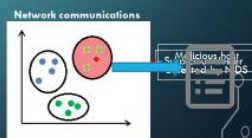
MACHINE LEARNING & CYBERSECURITY

Novel malware variants are likely to evade detection...
...but some features of malware behavior persist and are shared even by novel variants

External hosts behaving similarly to a known malicious external host are likely to also be malicious

USE ONE TO FIND MANY:

- Generate clusters of similar communications
- Use NIDS alerts to find malicious external hosts
- Label as suspicious all clusters containing malicious external hosts
- Build *graylist* with external hosts belonging to suspicious clusters



ING. GIOVANNI APRUZZESE

35

LATERAL MOVEMENT

through *Pivoting*

- Reminder: pivoting activities are not necessarily malicious
- Need to discriminate between "benign" and "malicious" pivoting

Intuition: Rank the detected pivoting activities on the basis of threatening characteristics displayed

- Characteristics that can be considered:
 - Novelty of the pivoting activity
 - Prior-reconnaissances
 - Usage of uncommon Ports
 - LANs involved
 - Anomalous Data Transfers

ING. GIOVANNI APRUZZESE

25

38

Big Data Security Analytics: Opportunities and Issues

December 12th, 2019

Ing. Giovanni Apruzzese

PhD Candidate in Information and Communication Technologies

✉ giovanni.apruzzese@unimore.it

🌐 <https://weblab.ing.unimo.it/people/apruzzese>