

Big Data Security Analytics: Opportunities and Issues

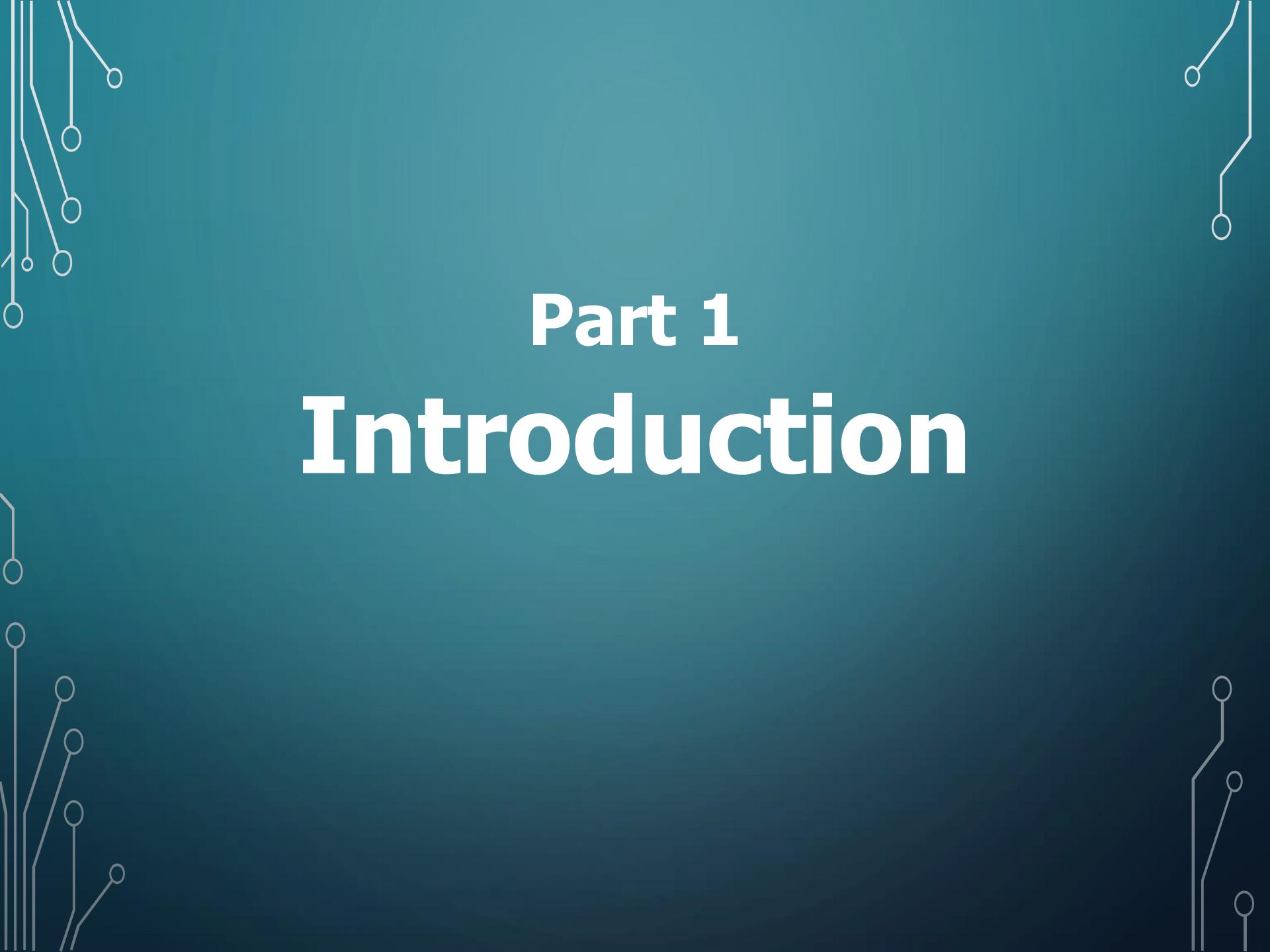
December 12th, 2019

Ing. Giovanni Apruzzese

PhD Candidate in Information and Communication Technologies

✉ giovanni.apruzzese@unimore.it

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



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

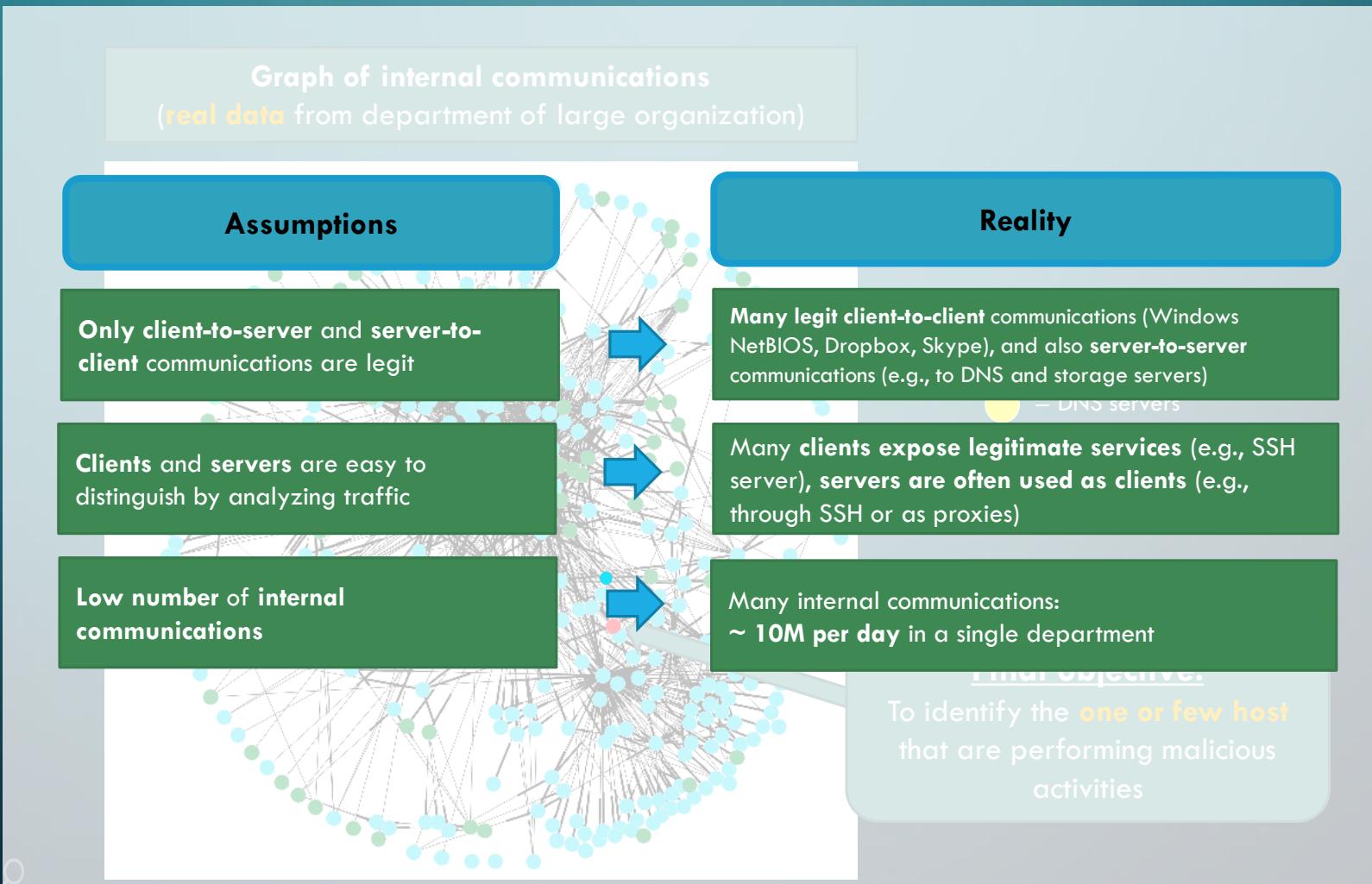
CONTEXT

- 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



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)

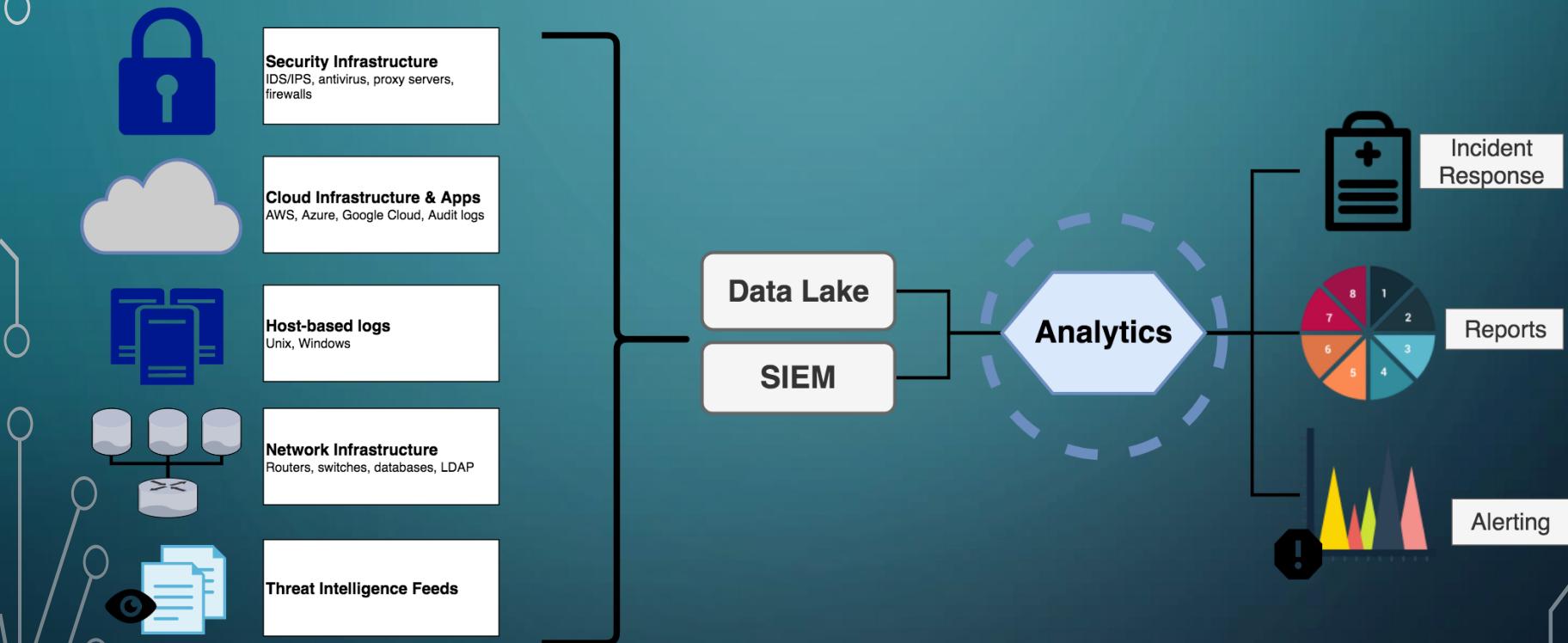
- Reporting
- Information sources: various log formats
- Advanced correlation
- Signature-based alerting
- Increasing devices: >1000
- Events per second: >10000
- Storage: Terabytes
- Focus on threat detection and response, breach response slow, dependent on security analyst skills

2014+ Security Analytics

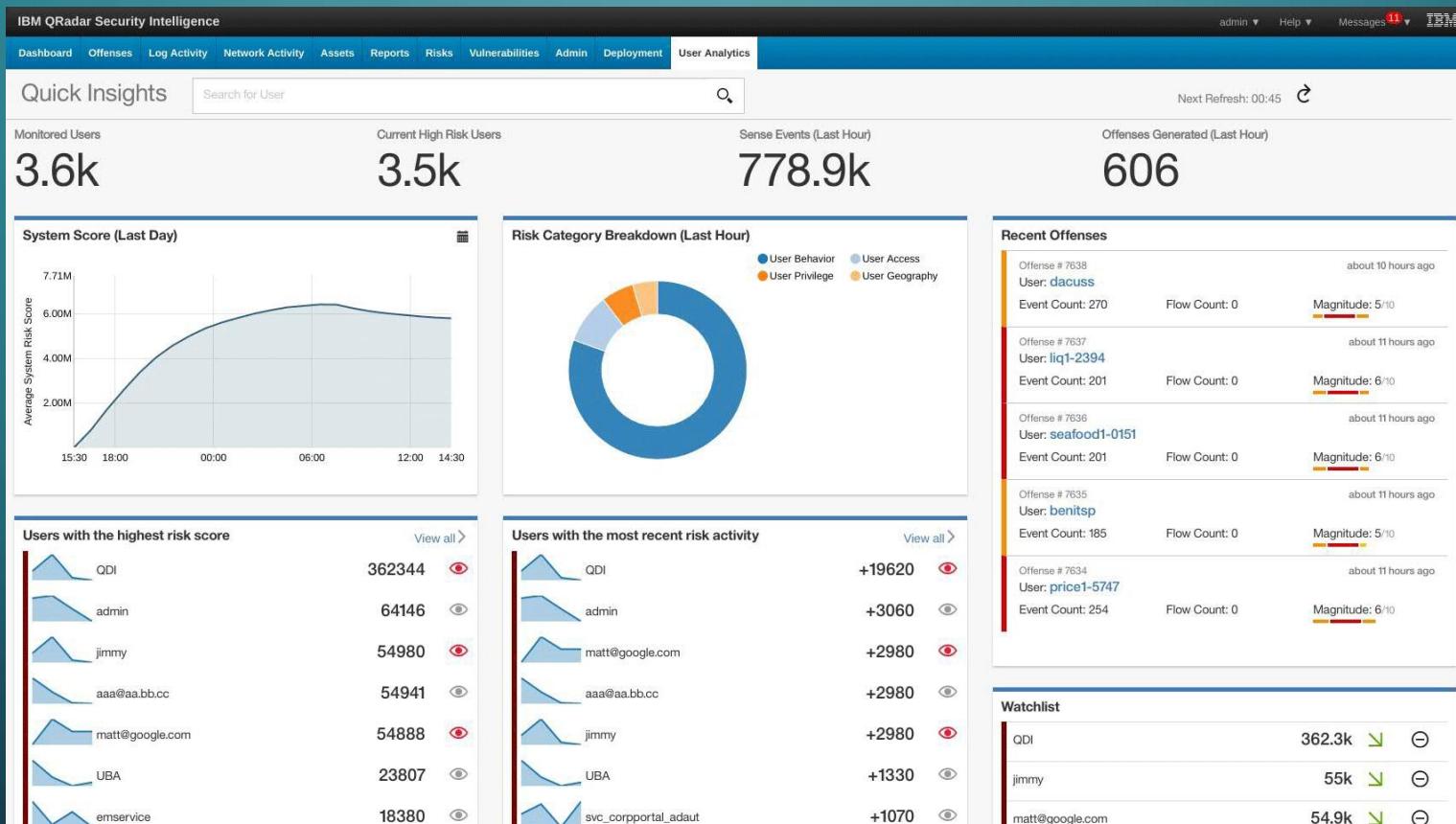
- Feeds from applications, databases, endpoints
- Threat detection
- Advanced analytics with additional security context
- User and **network** behavior
- Heterogeneous data: **Netflow**, threat intelligence feeds, multiple log sources
- Huge number of devices: >5000
- Events per second: >100000
- Storage: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

Sophistication, volume, velocity, scalability, complexity

STATE-OF-THE-ART SECURITY ANALYTICS



EXAMPLES: QRADAR



EXAMPLES: SPLUNK

The screenshot shows the Splunk Enterprise Security application interface. The top navigation bar includes links for Home, Security Posture, Incident Review, Investigations, Glass Tables, Security Intelligence, Security Domains, Audit, Configure, Administrator, Messages, Settings, Activity, Help, Find, Export, and three dots for more options.

The main dashboard section is titled "Security Posture" and displays "Overall Security Posture : Key Security Indicators". It features five cards:

- THREAT ACTIVITY**: Total Count 778 (down 50), with a green downward arrow icon.
- AUTH. USERS**: Distinct Count 3.9k (up +67), with a red upward arrow icon.
- CLOUD ACTIVITY**: Email Count 7.8k (down -734), with a green downward arrow icon.
- INFECTED SYSTEMS**: System Count 219 (0), with a black zero icon.
- UNIQUE DESTINATIONS**: Unique Count 39.7k (up +1.3k), with a red upward arrow icon.

Below these are two charts:

- Overall Notable Event Occurrence By Urgency**: A horizontal bar chart showing event counts for critical, high, low, and medium urgency levels. The critical and high bars are orange, while low and medium are green. The y-axis is labeled "urgency" and the x-axis is "count" from 0 to 1,000.
- Overall Notable Events Occurrence Trend**: A line chart showing the count of events over time from 8:00 PM on Monday May 14, 2018, to 4:00 PM on Tuesday May 15, 2018. The chart tracks four categories: access (blue), audit (green), endpoint (dark blue), and network (yellow). The y-axis is "Count" from 0 to 1,000, and the x-axis is "time".

At the bottom, there are two tables:

- Top Notable Events Occurrence**: A table listing notable events with their rule names, sparkline visualizations, counts, source IP addresses, correlation search counts, and security domain counts. Examples include "Monitor Web Traffic For Brand Abuse" (count 3641) and "UEBA Threat Detected" (count 1380).
- Top Notable Event Occurrence by Host**: A table listing notable events by host IP address, showing counts and security domain counts. Examples include "10.11.36.20" (count 8) and "10.11.36.18" (count 7).

At the very bottom, a note states: "No investigation is currently loaded. Please create (+) or load an existing one (≡)".

EXAMPLES: APACHE SPOT

The screenshot displays the Apache Spot web interface with four main panels:

- Suspicious:** A table listing suspicious network flows. The columns are Rank, Time, Source IP, Destination IP, Source Port, Destination Port, Protocol, and Input Packets. The data shows:

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

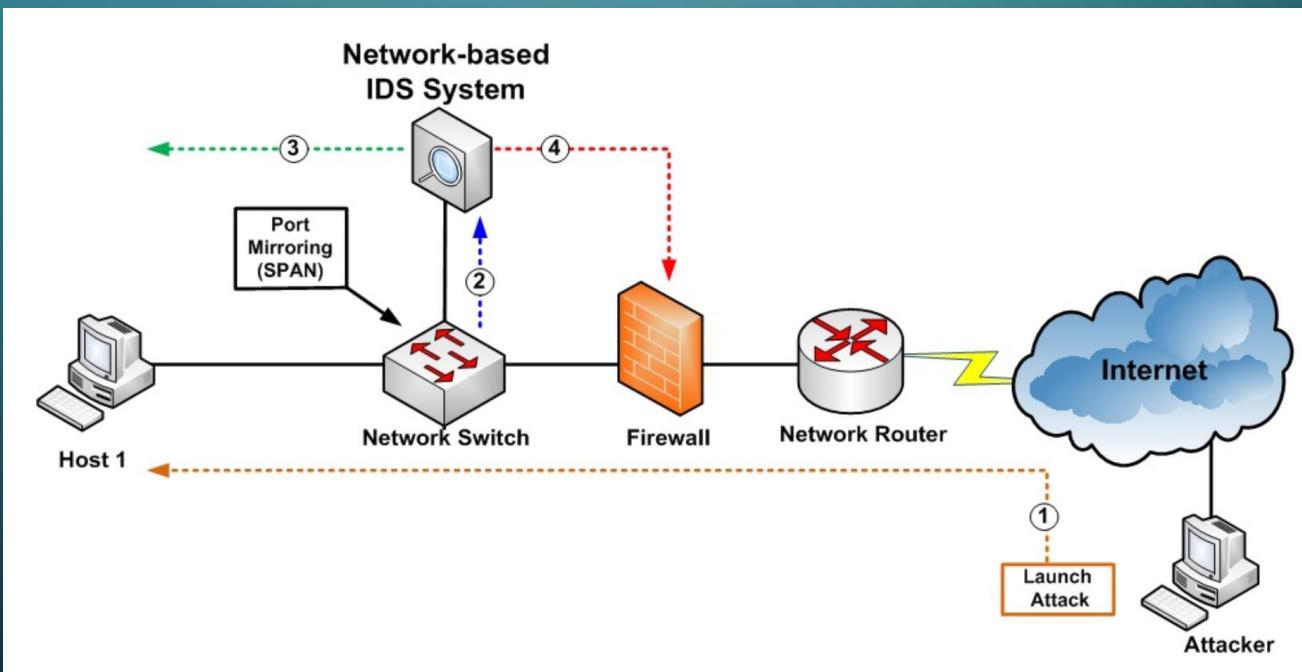
- Network View:** A network graph visualization showing connections between nodes. Nodes are represented by circles of varying sizes and colors (blue, yellow, orange), and connections are shown as lines.
- Notebook:** A form for selecting network parameters. It includes dropdown menus for Source IP (172.30.0.46), Dest IP (10.0.0.183), Src Port (52234), Dst Port (119), and a "Quick IP scoring" section with a rating scale from 1 to 3. Buttons for "Score" and "Save" are also present.
- Details:** A pie chart showing the distribution of scores for the selected IP. The chart has three segments: a large orange segment (Rating 1), a medium blue segment (Rating 2), and a small yellow segment (Rating 3).

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)

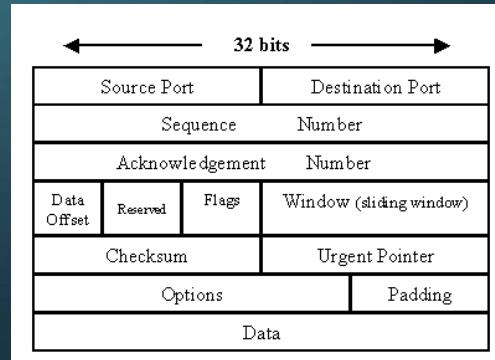
The screenshot shows the Wireshark interface capturing traffic on the eth0 interface. The packet list pane displays several TCP and DNS packets. The details pane shows the structure of a selected TCP packet, which is highlighted in green. The bytes pane shows the raw hex and ASCII data of the selected packet. The status bar at the bottom indicates "eth0: <live capture in progress>" and "Packets: 445 Displayed: 445 Marked: 0".

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.

Profile: Default

Example: TCP Packet

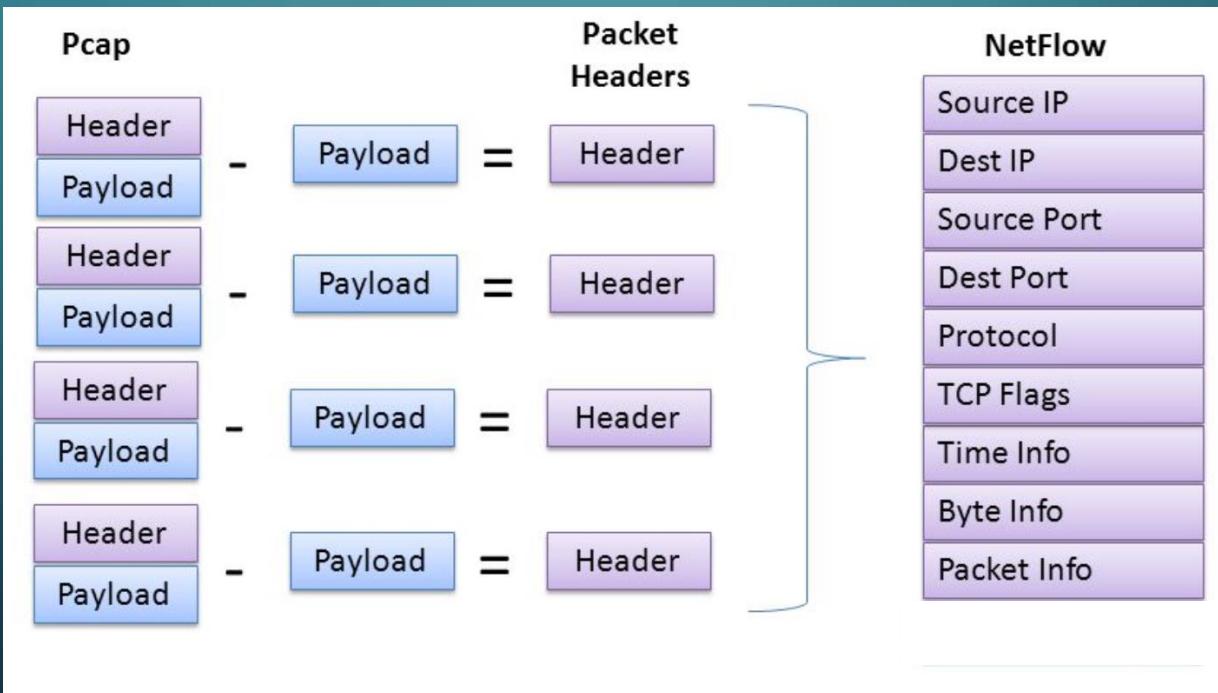


BRIEF RECAP

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)

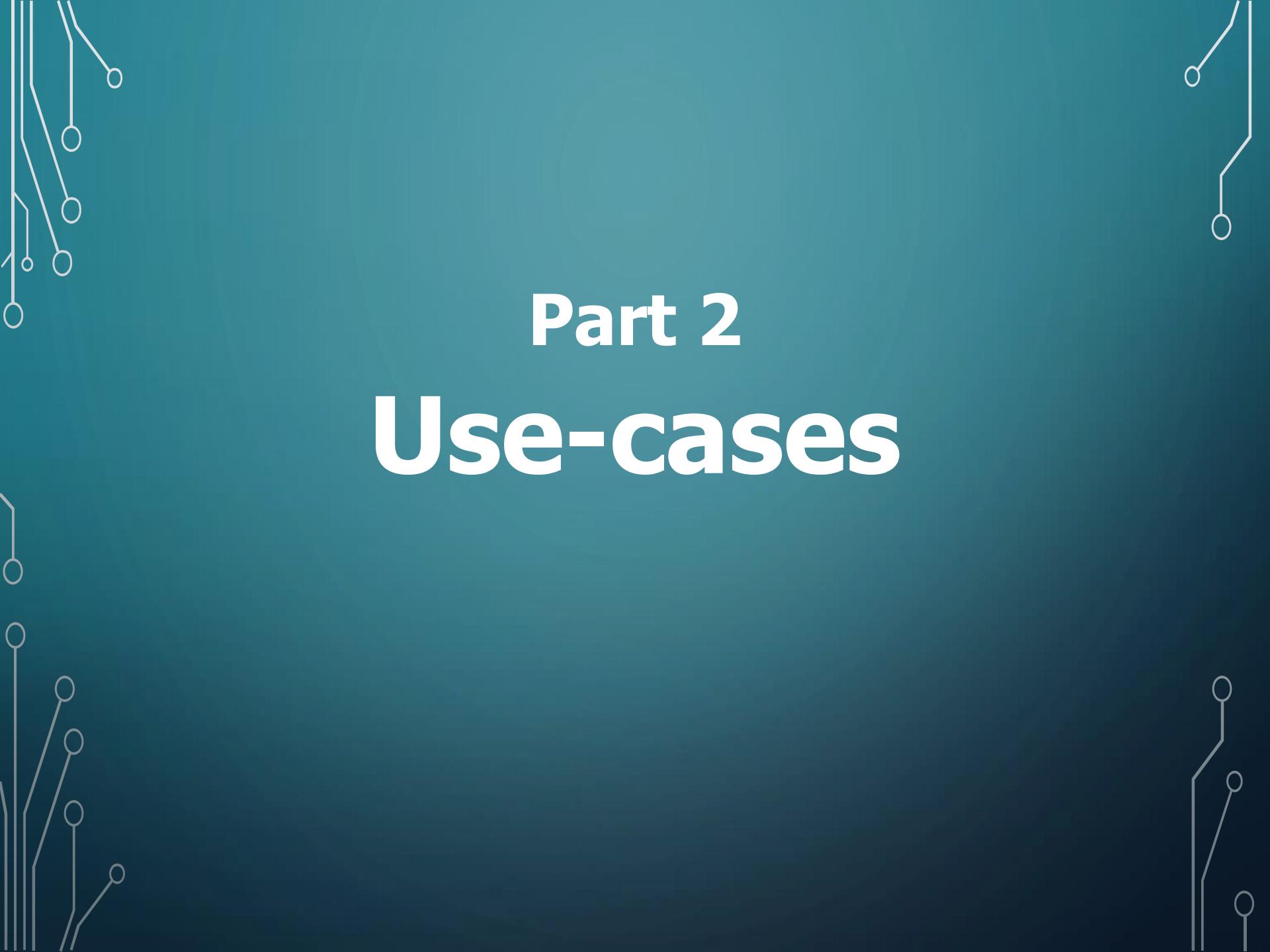


REMINDER

Analysis



Analytics

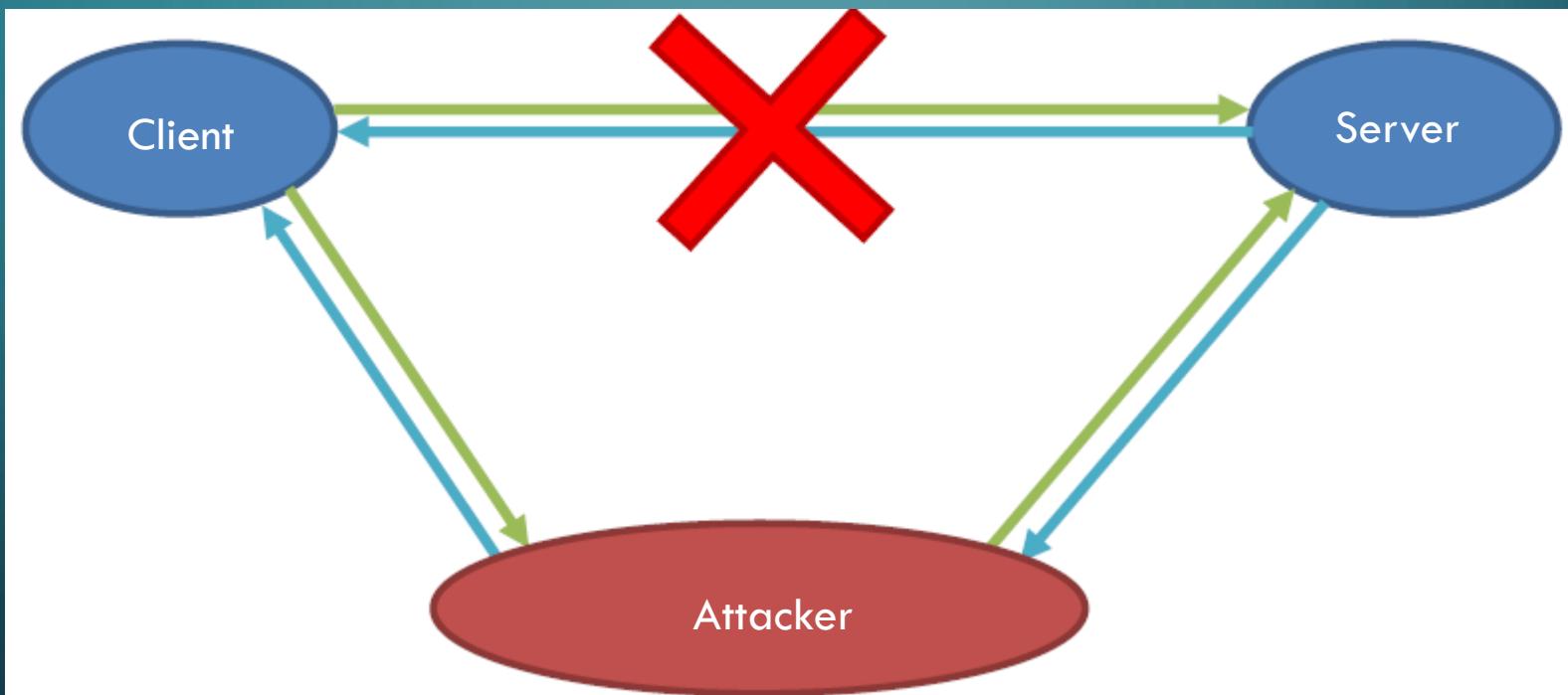


Part 2

Use-cases

MAN-IN-THE-MIDDLE

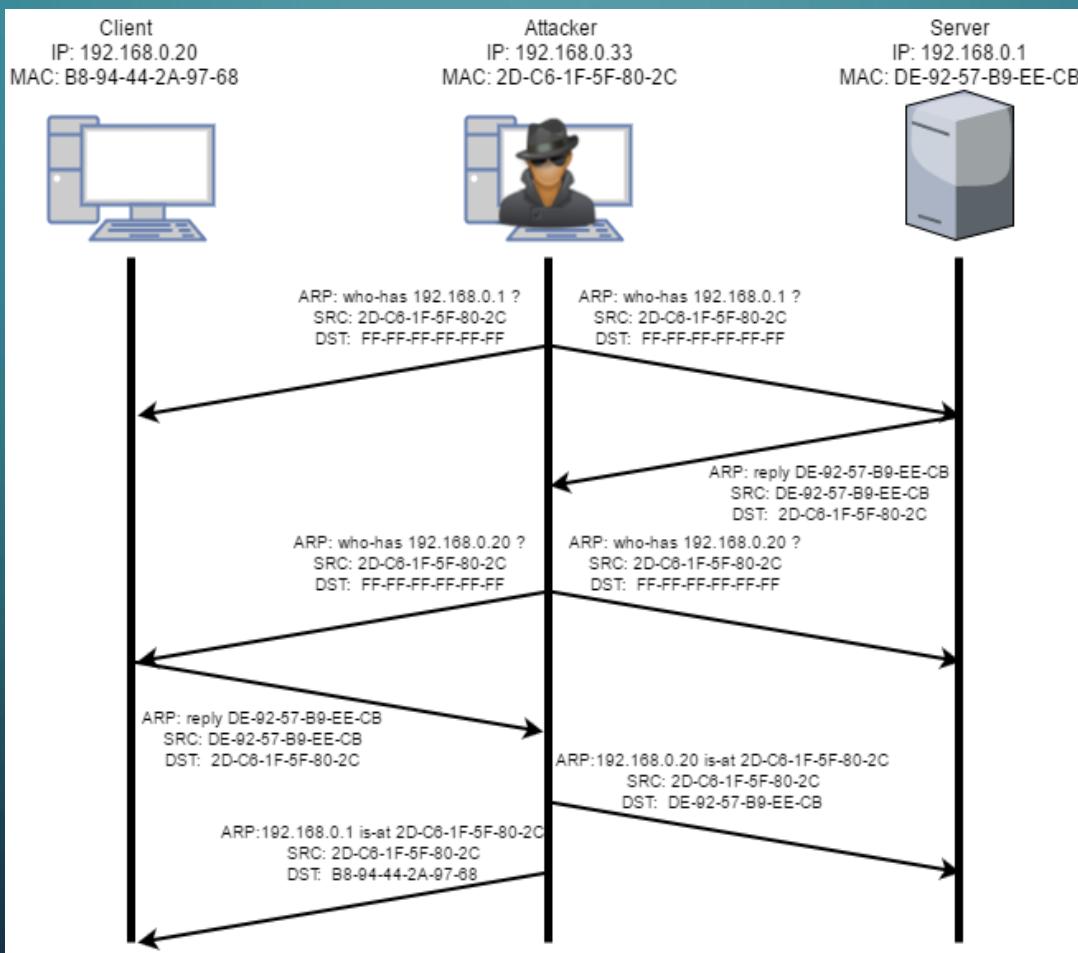
through *ARP Spoofing*



MAN-IN-THE-MIDDLE

through *ARP Spoofing*

Step-by-step



MAN-IN-THE-MIDDLE

through *ARP Spoofing*

Intuition: all packets are doubled!

Check
Packets!



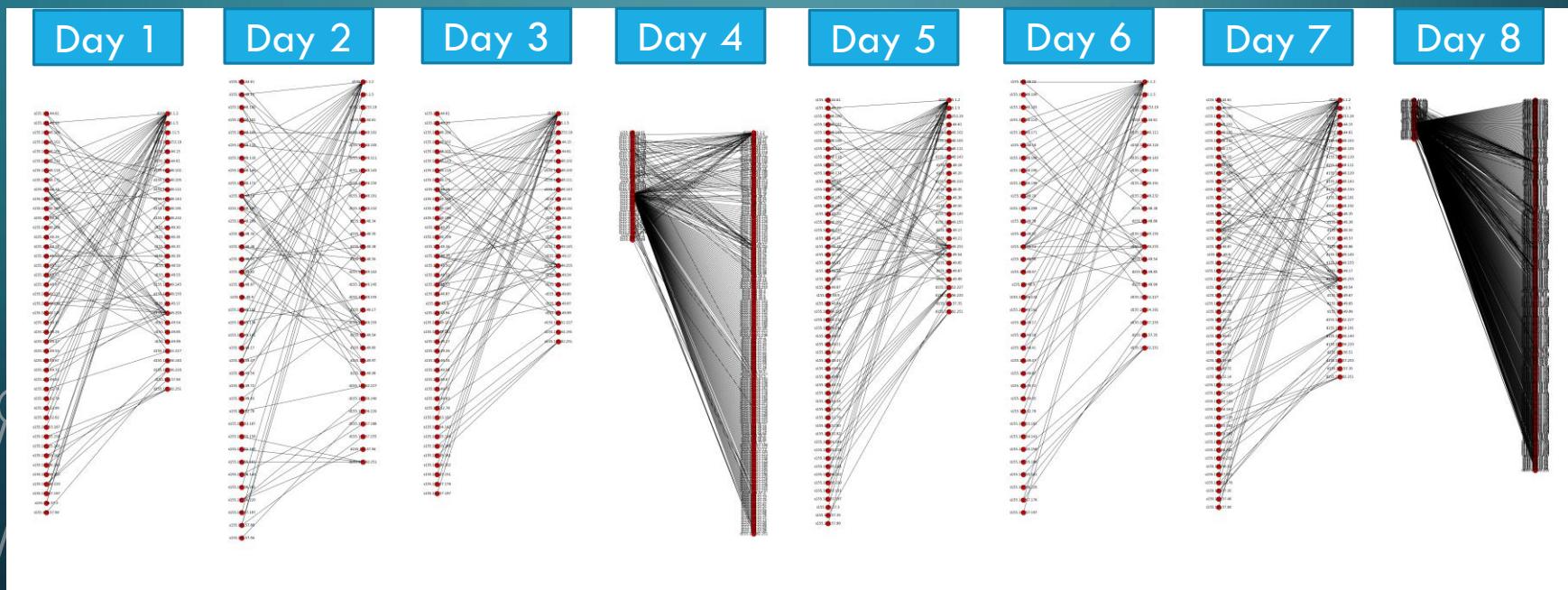
HOWEVER!

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*

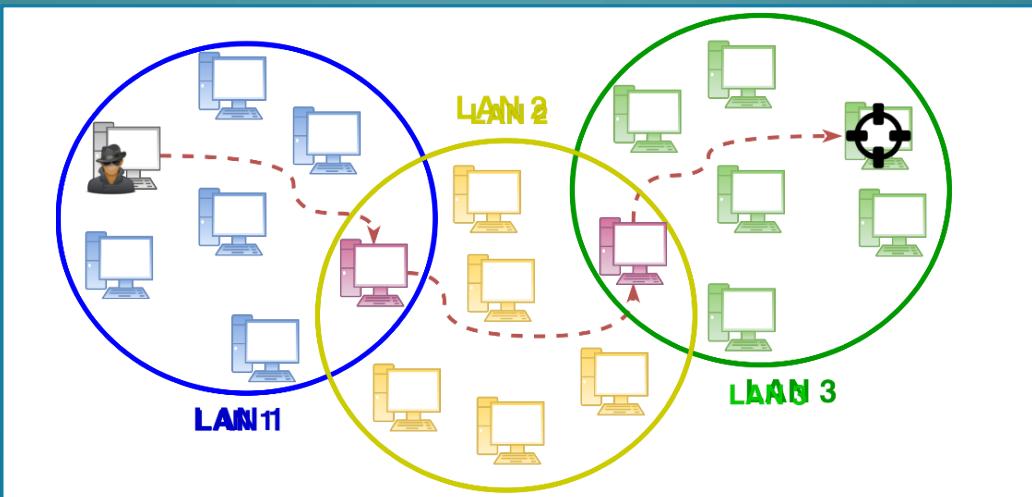
Intuition: the average duration of the scanner-host's connections decreases, while the *number of flows* and *contacted hosts* increase.



LATERAL MOVEMENT

through *Pivoting*

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



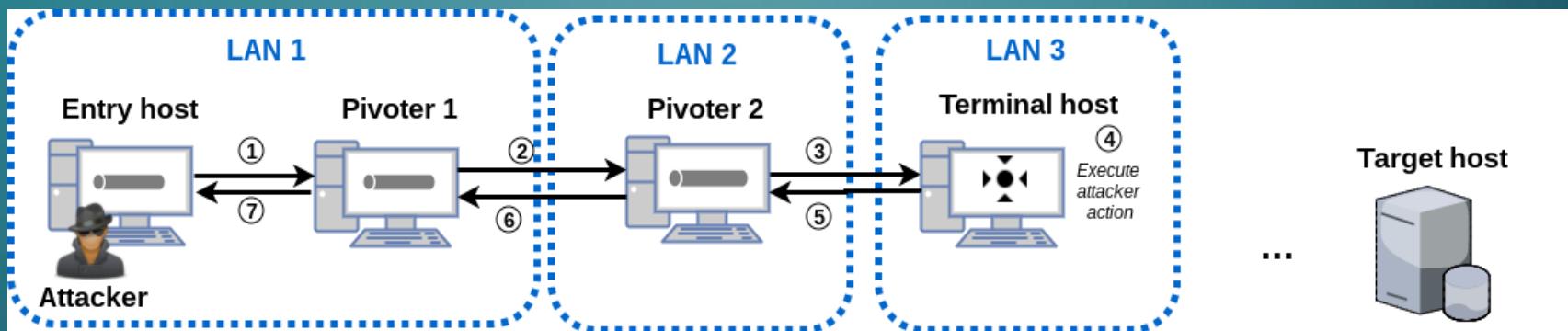
Pivoting: any action in which a *command propagation tunnel* is created
among three or more hosts

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*

Step1

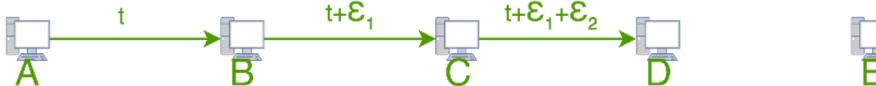


Step2

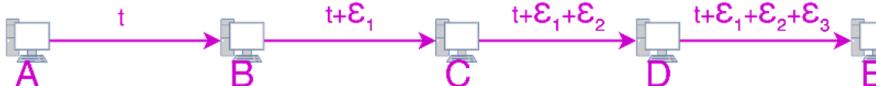
$$\varepsilon_i \leq \varepsilon_{max}, \forall i$$



Step3



Step4



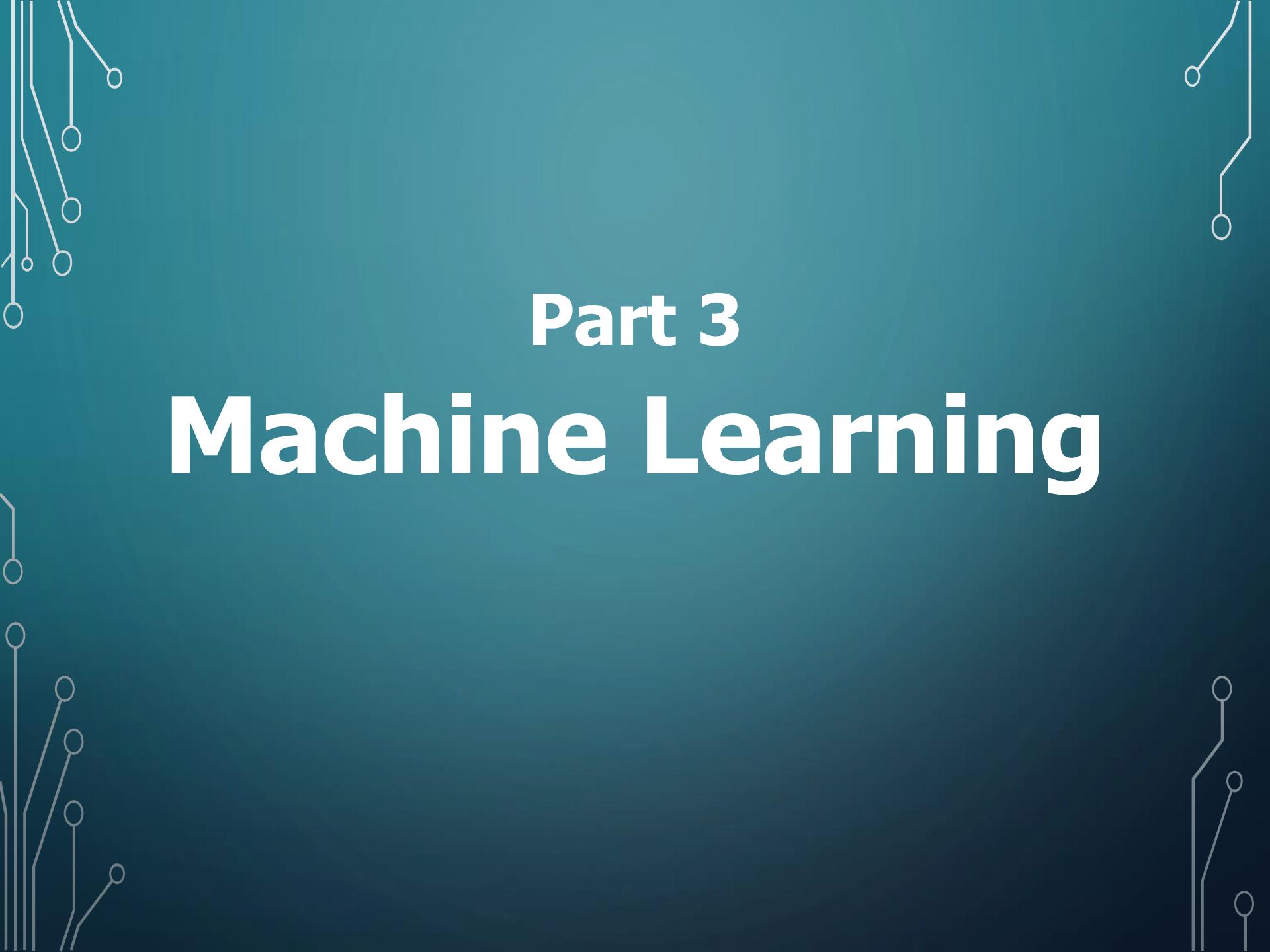
LATERAL MOVEMENT

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



Part 3

Machine Learning

MACHINE LEARNING

The popularity of machine learning is skyrocketing.



MACHINE LEARNING & CYBERSECURITY



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

Machine Learning: New Frontiers in Advanced Threat Detection



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

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

MACHINE LEARNING HELPS US FIND NEW ATTACKS



Machine learning in Kaspersky Endpoint Security 10 for Windows



TREND
MICRO™

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



CYBERARK®

MACHINE LEARNING PREVENTS PRIVILEGE ATTACKS AT THE ENDPOINT

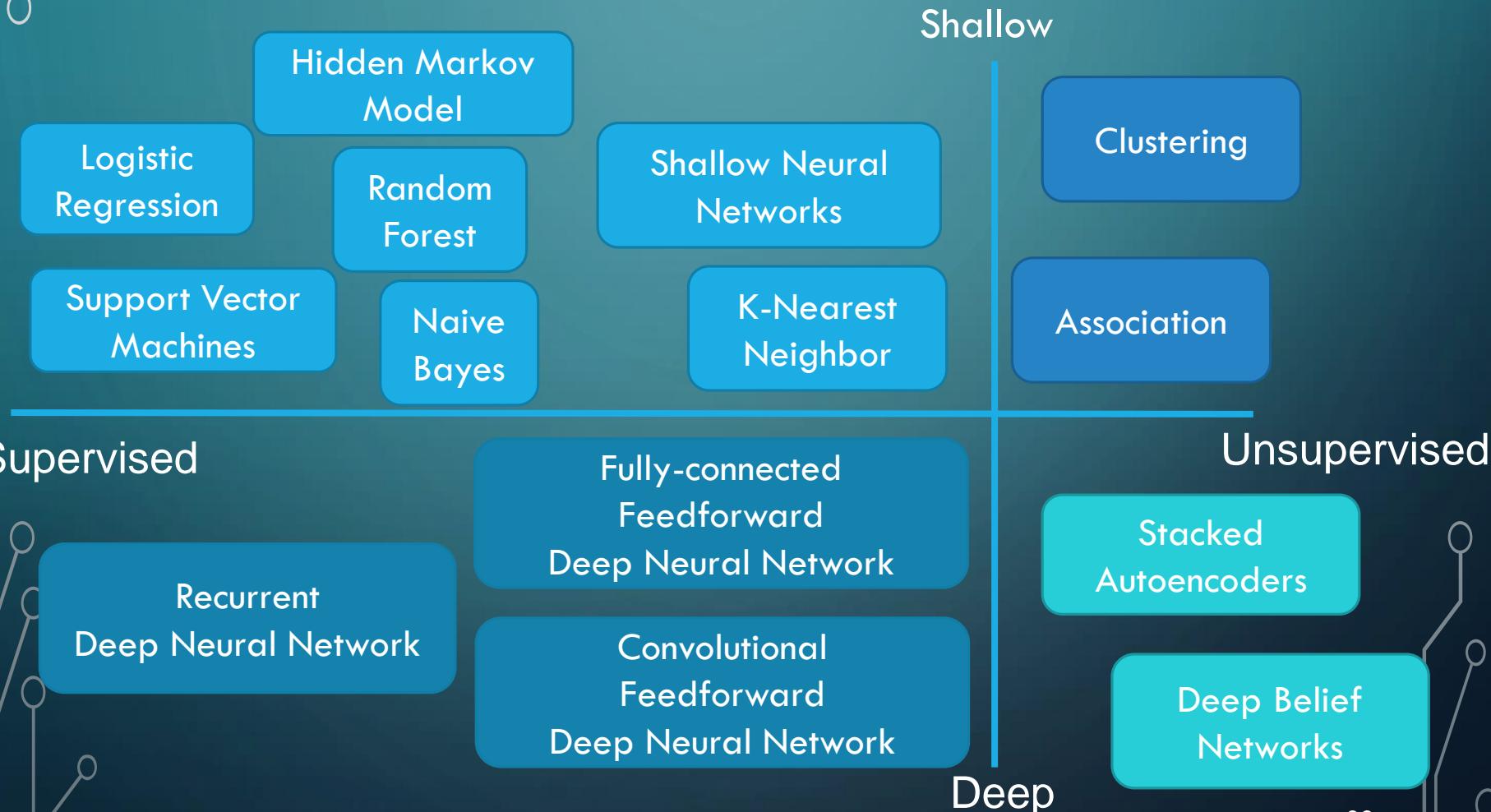


Rapid7 Attacker Behavior Analytics Brings Together Machine Learning and Human Security Expertise



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)

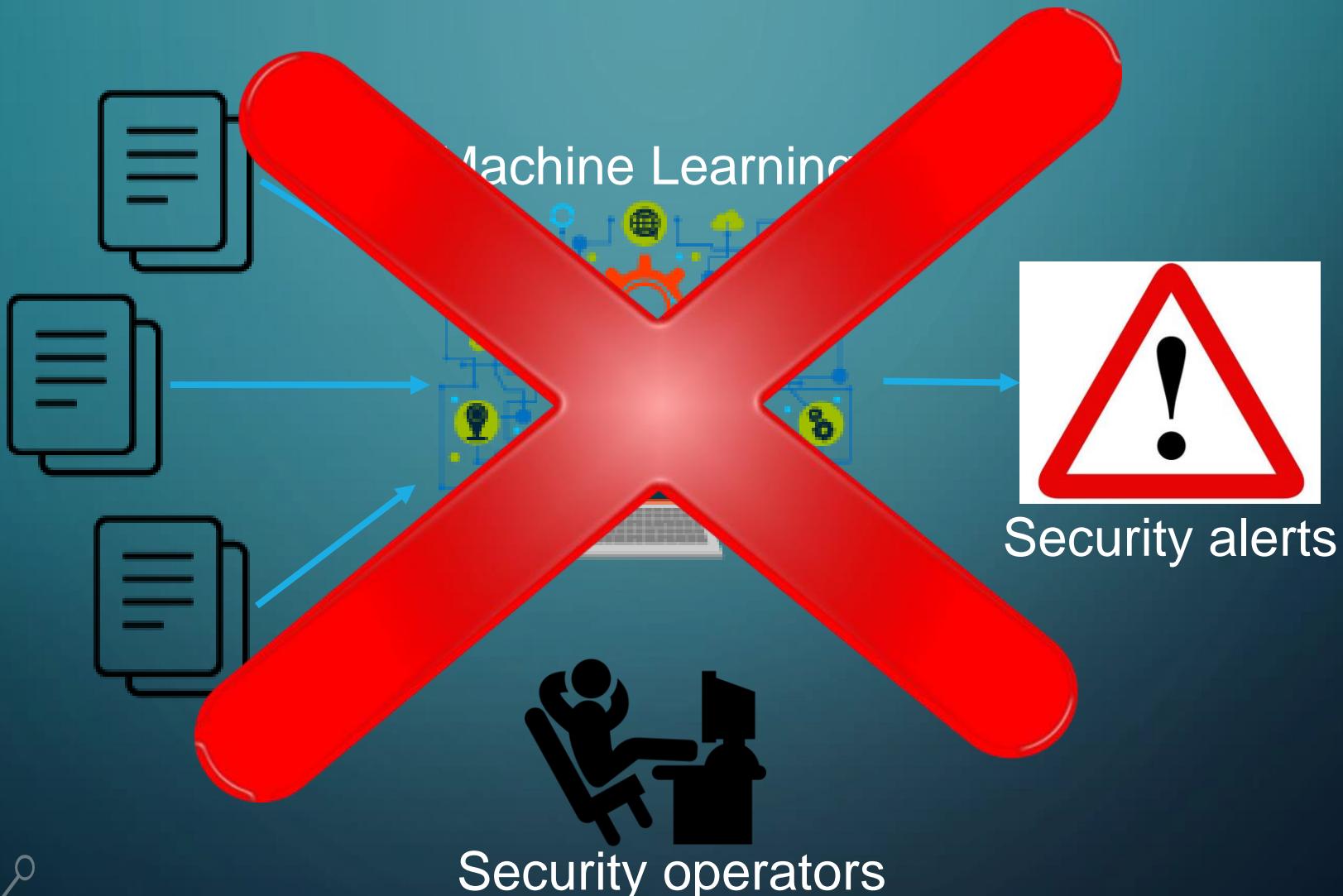
False positives and false negatives

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

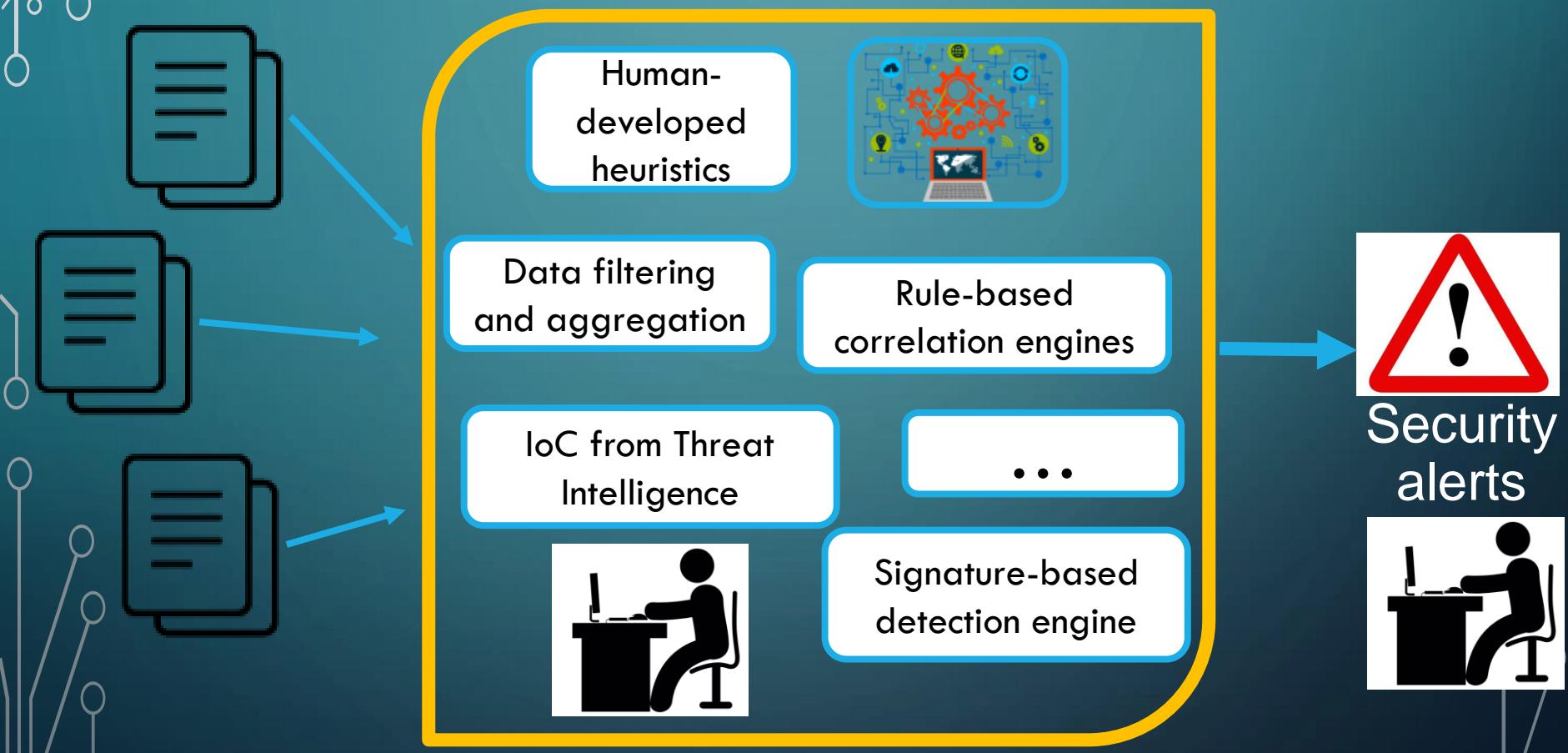
Adversarial attacks

- 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

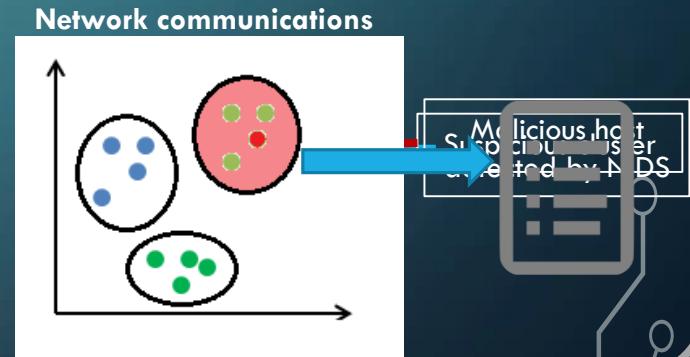
MACHINE LEARNING & CYBERSECURITY

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



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?

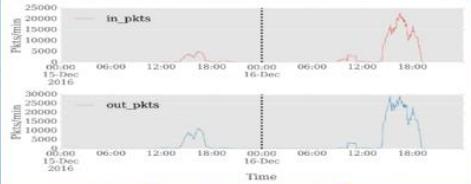
QUESTION

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 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.



21

MACHINE LEARNING & CYBERSECURITY

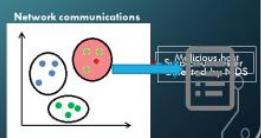
Novel malware variants are likely to evade detection...

...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>