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

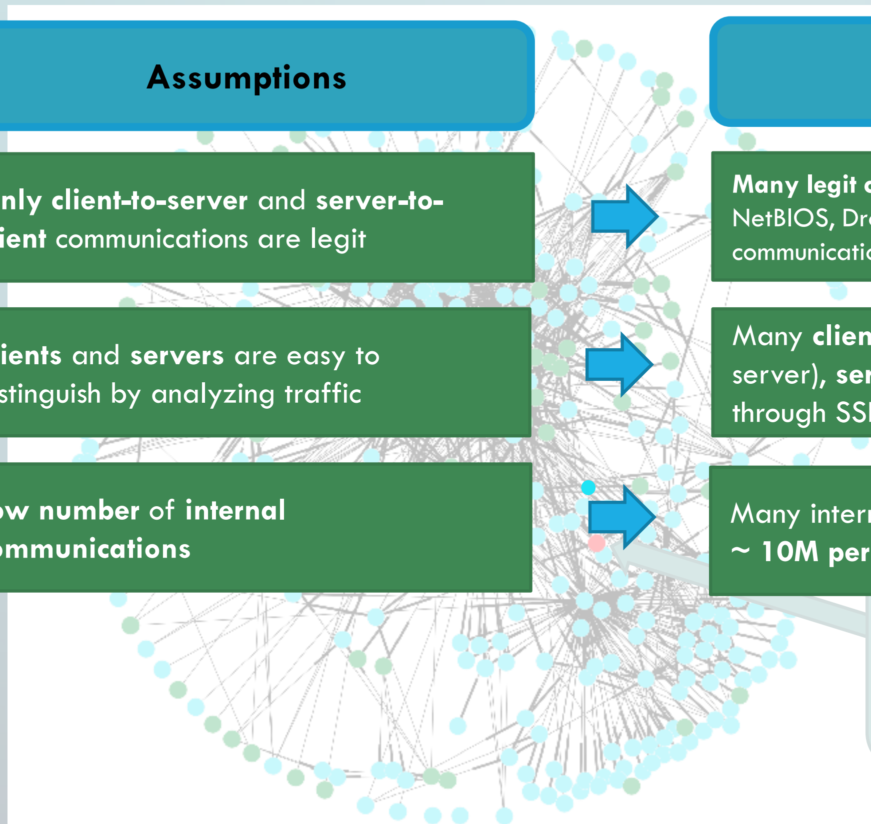
- At the same time, the volume of generated data **is exploding**

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

CONTEXT

Example

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



Assumptions		Reality
Only client-to-server and **server-to-client** communications are legit	→	**Many legit client-to-client** communications (Windows NetBIOS, Dropbox, Skype), and also **server-to-server** communications (e.g., to DNS and storage servers)
Clients and **servers** are easy to distinguish by analyzing traffic	→	Many **clients expose legitimate services** (e.g., SSH server), **servers are often used as clients** (e.g., through SSH or as proxies)
Low number of **internal communications**	→	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)

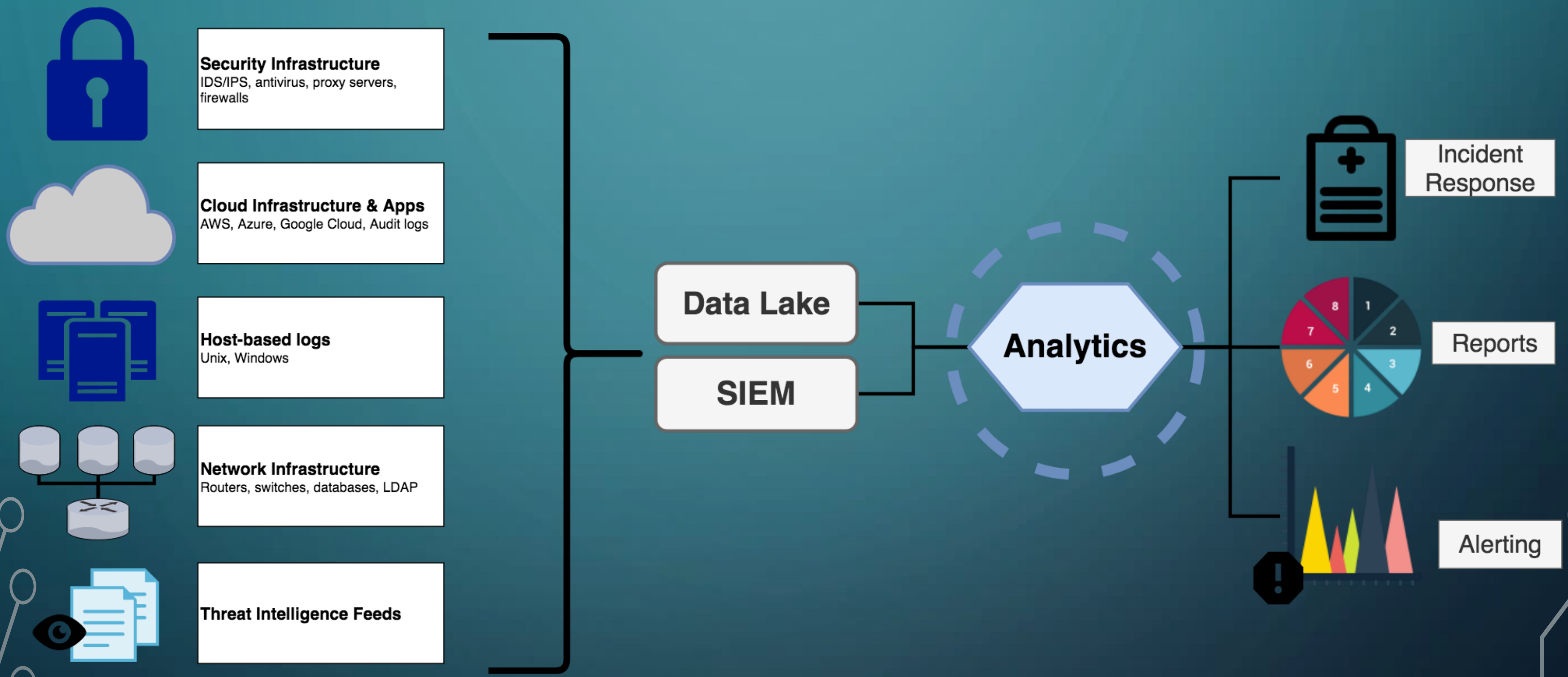
- Reporting
- Information sources: various log formats
- Advanced correlation
- Signature-based alerting
- Increasing devices: >1000
- Events per second: >10000
- Storage: Terabytes
- Focus on threat detection and response, breach response slow, dependent on security analyst skills

2014+ Security Analytics

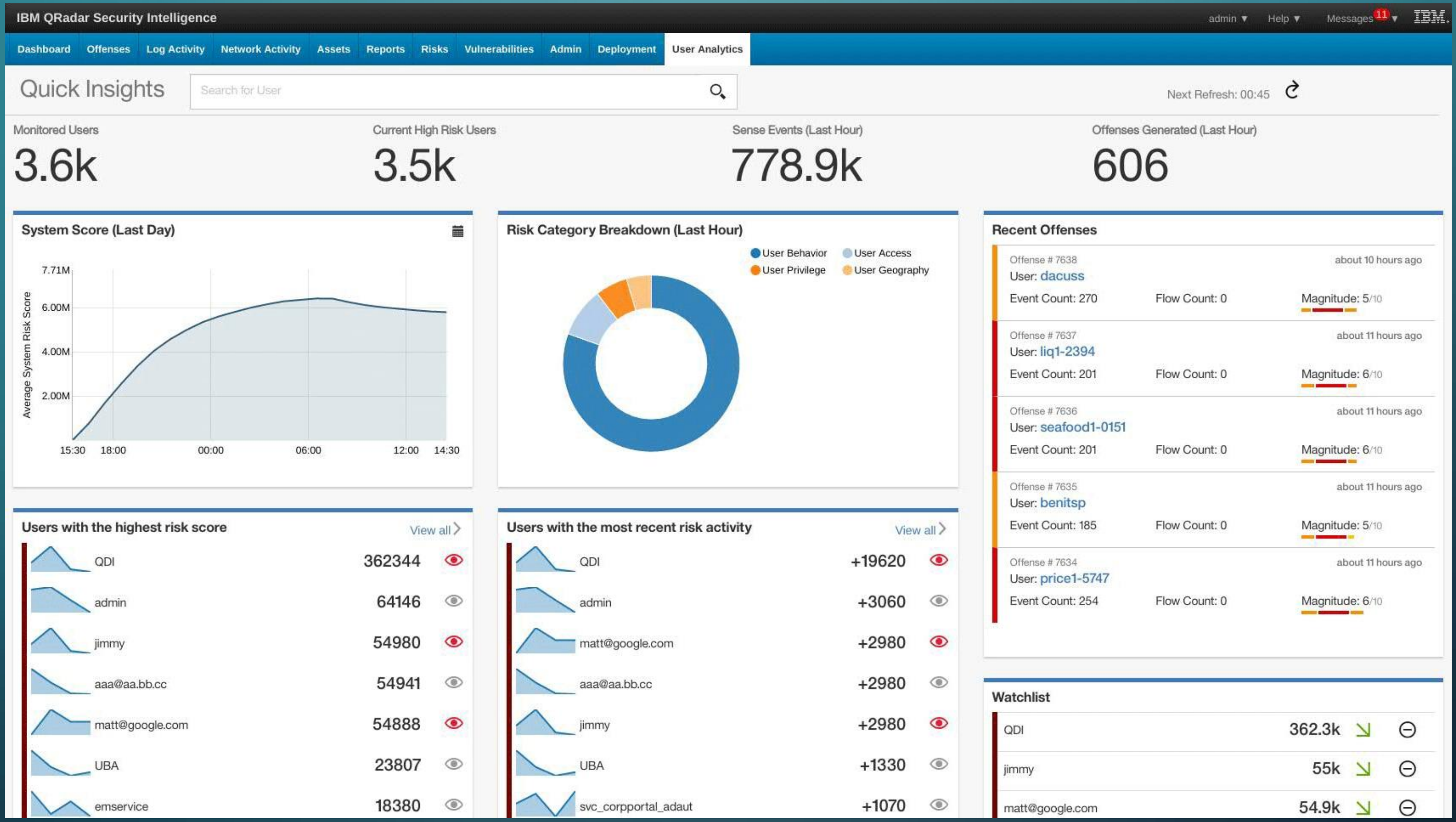
- Feeds from applications, databases, endpoints
- Threat detection
- Advanced analytics with additional security context
- **User** and **network** behavior
- Heterogeneous data: **Netflow**, threat intelligence feeds, multiple log sources
- Huge number of devices: >5000
- Events per second: >100000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

Sophistication, volume, velocity, scalability, complexity

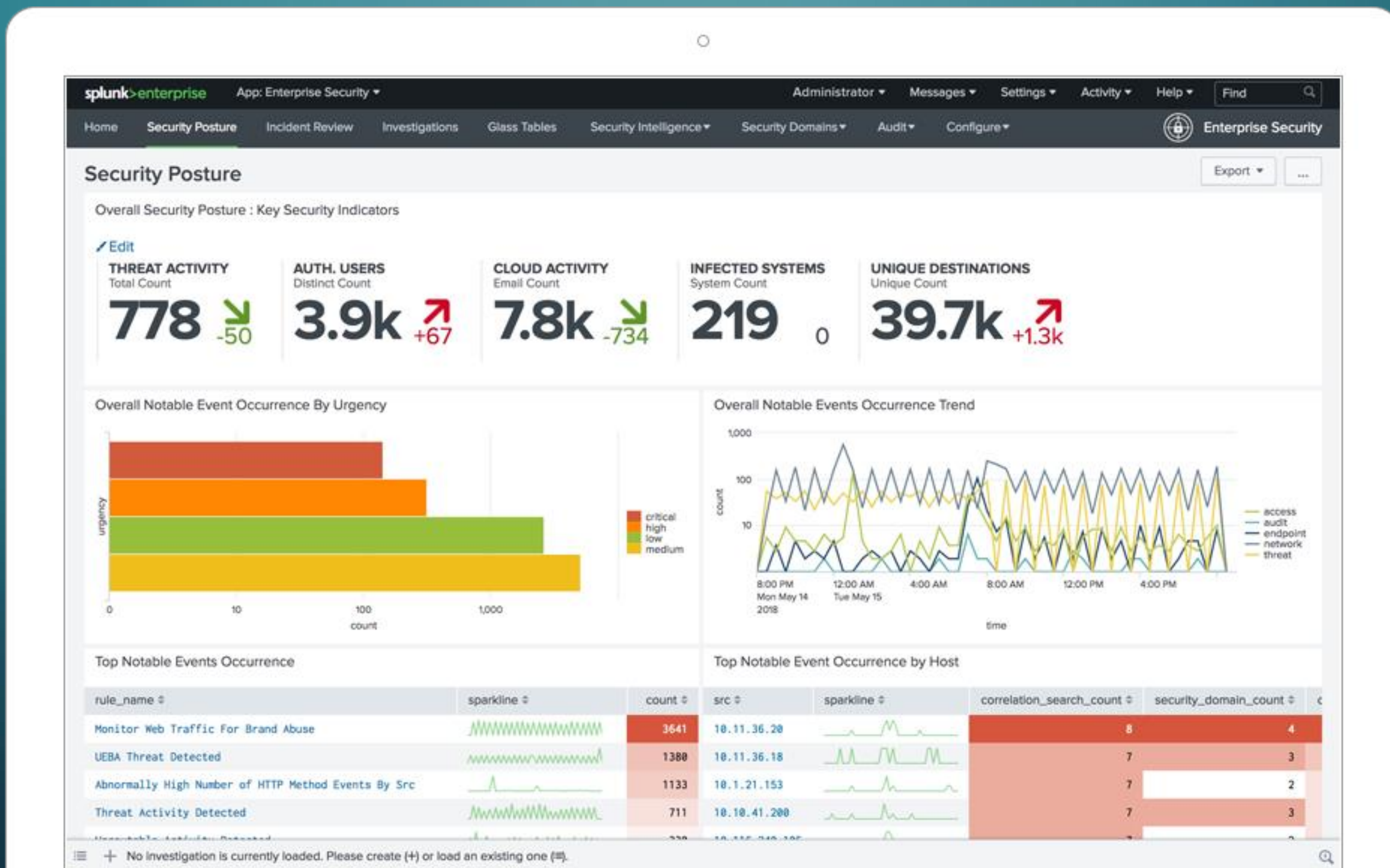
STATE-OF-THE-ART SECURITY ANALYTICS



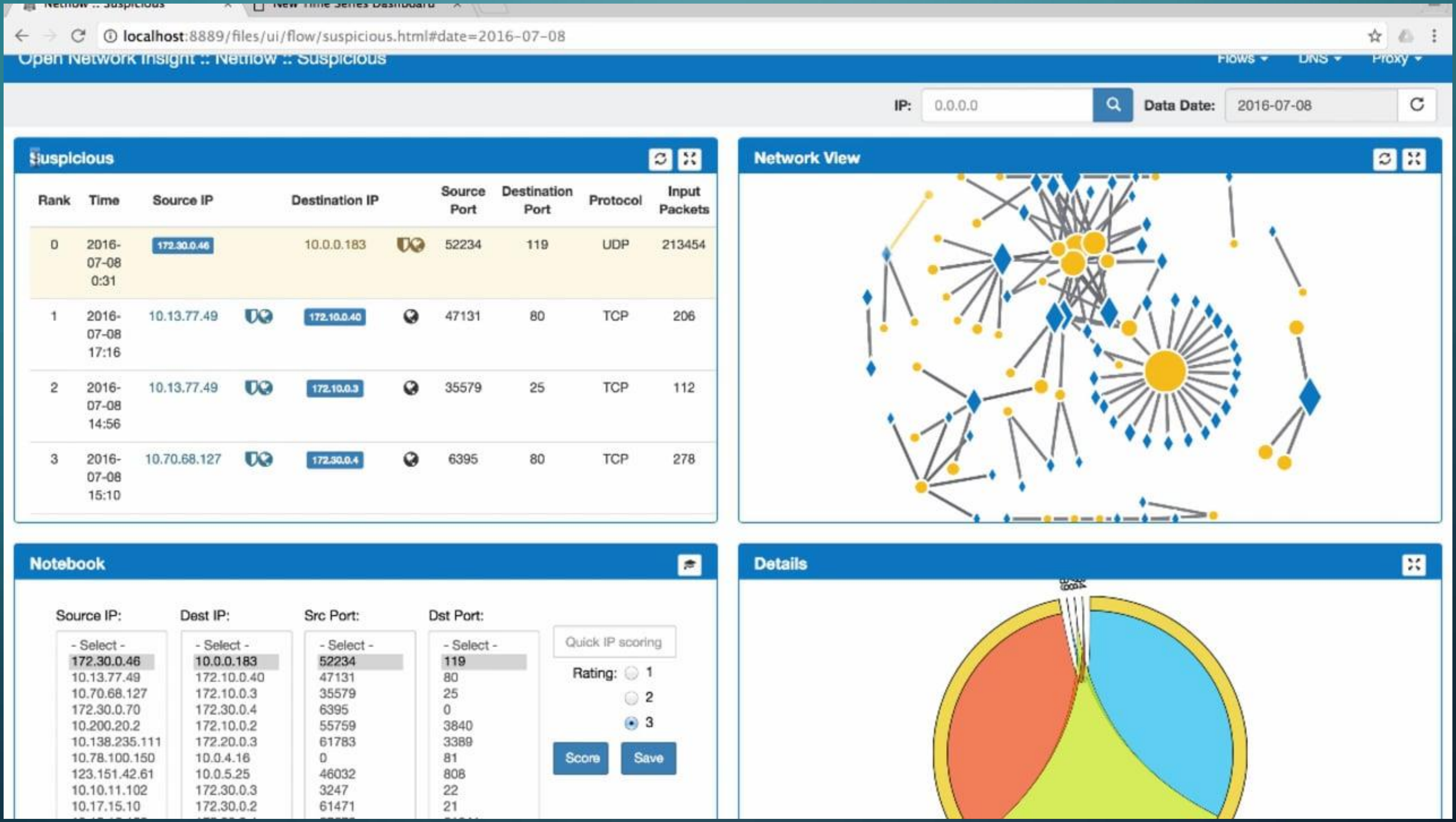
EXAMPLES: QRADAR



EXAMPLES: SPLUNK



EXAMPLES: APACHE SPOT

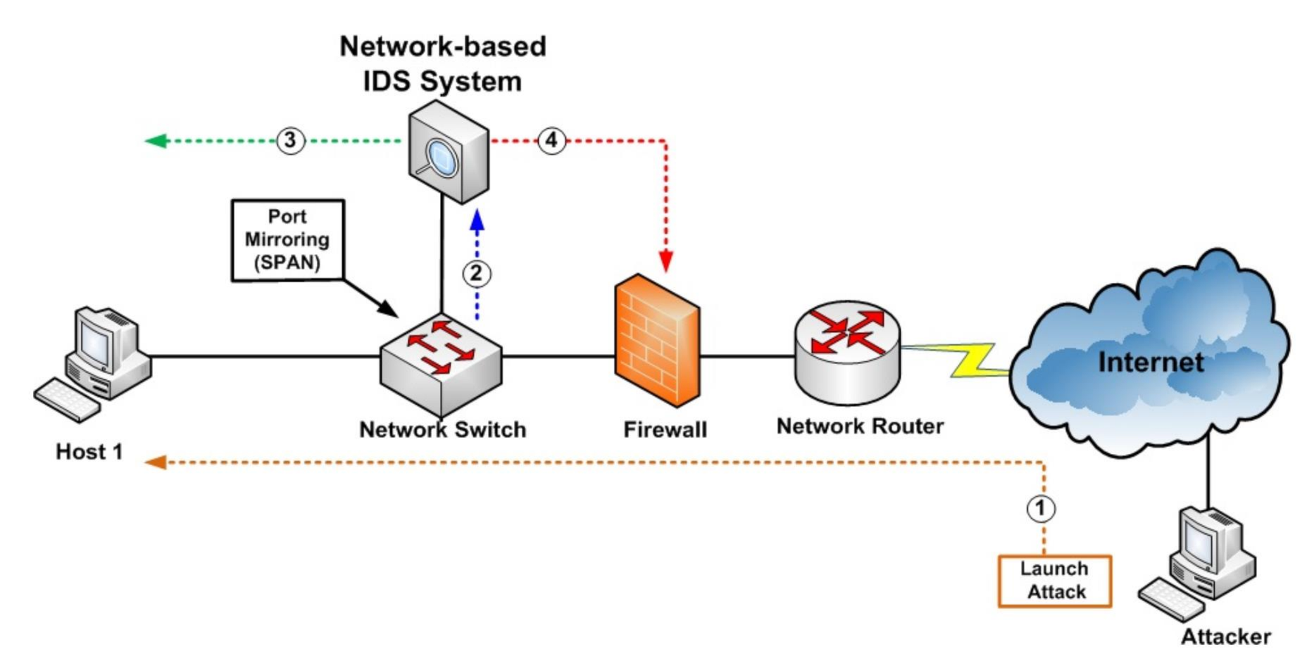


BRIEF RECAP

Intrusion Detection System (IDS)

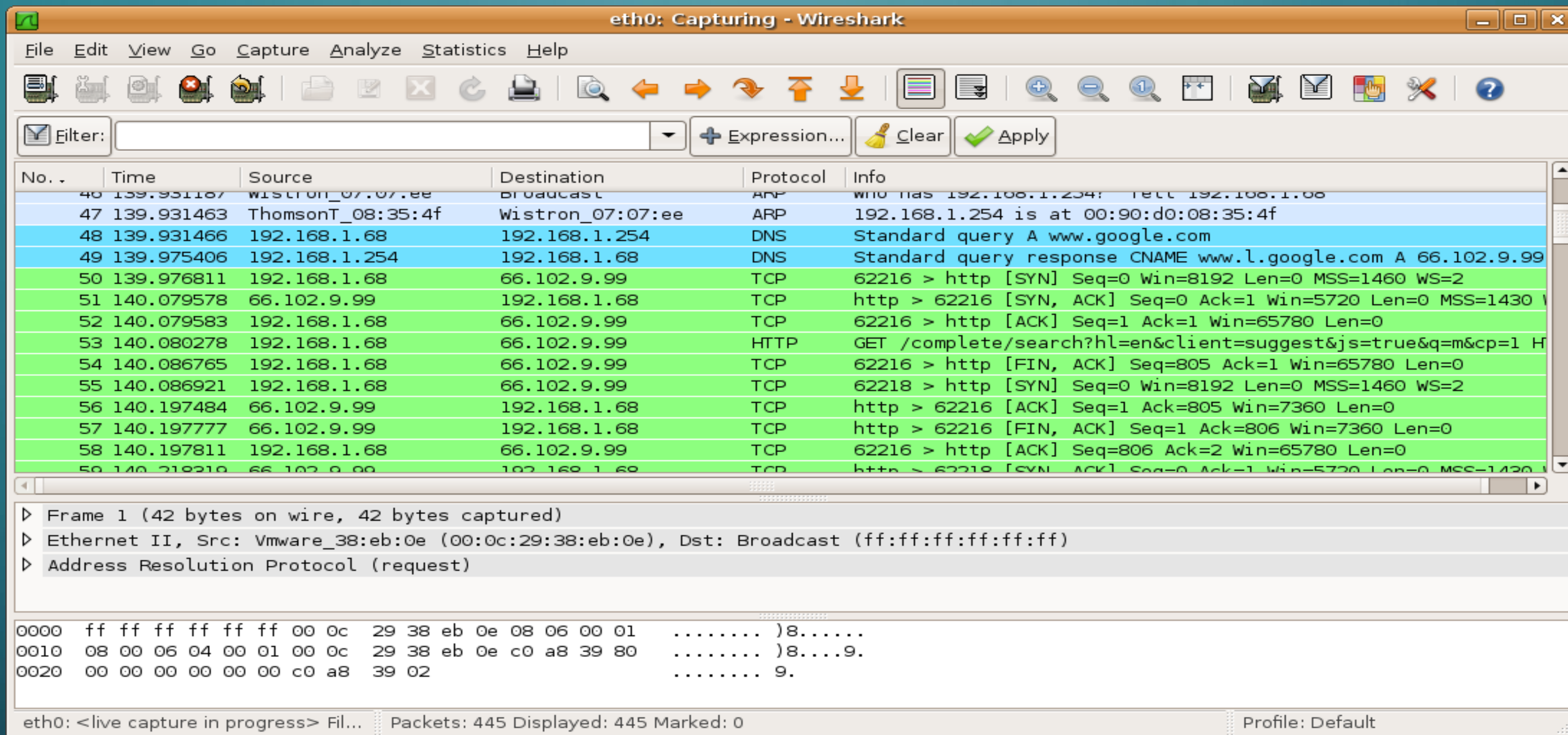
Host-based
Intrusion Detection System
(HIDS)

Network-based
Intrusion Detection System
(NIDS)

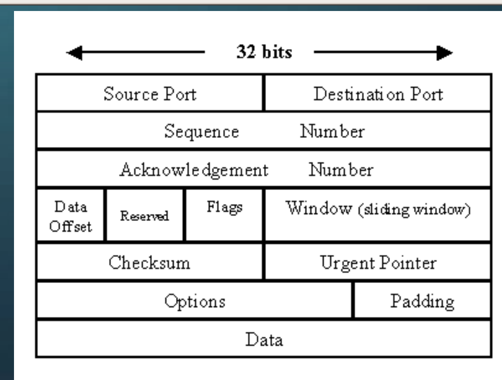


BRIEF RECAP

Network Traffic – Full Packet Capture (PCAP)



Example: TCP Packet

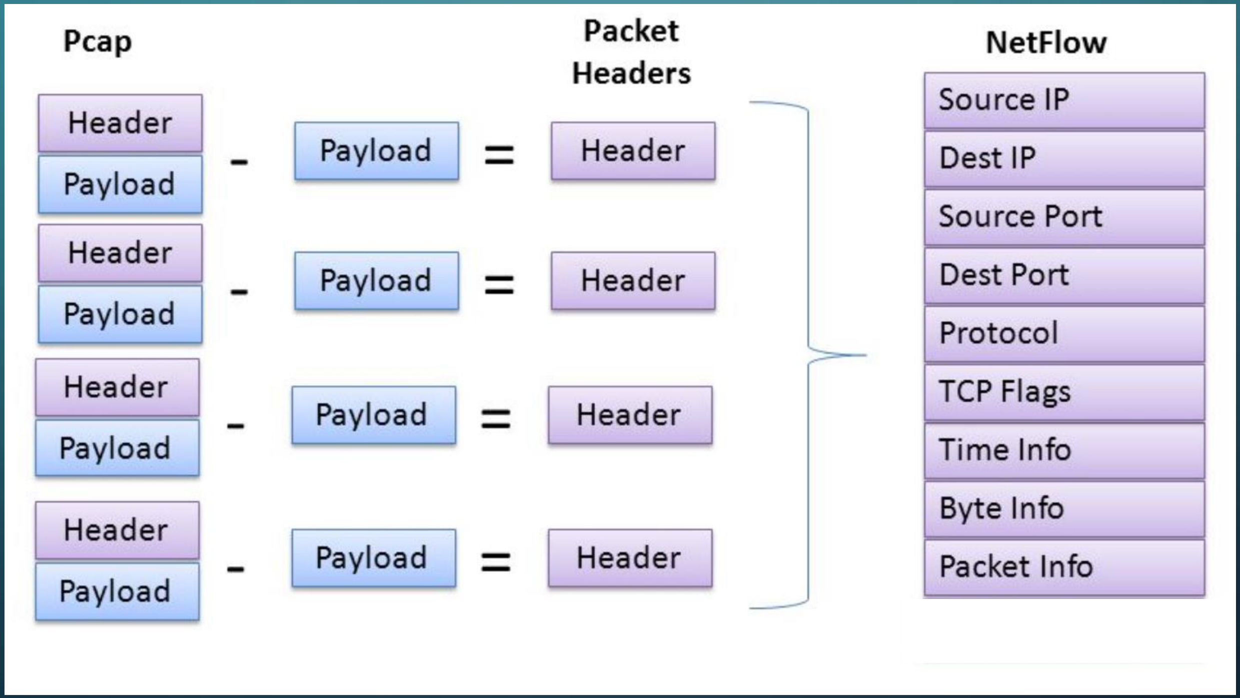


BRIEF RECAP

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

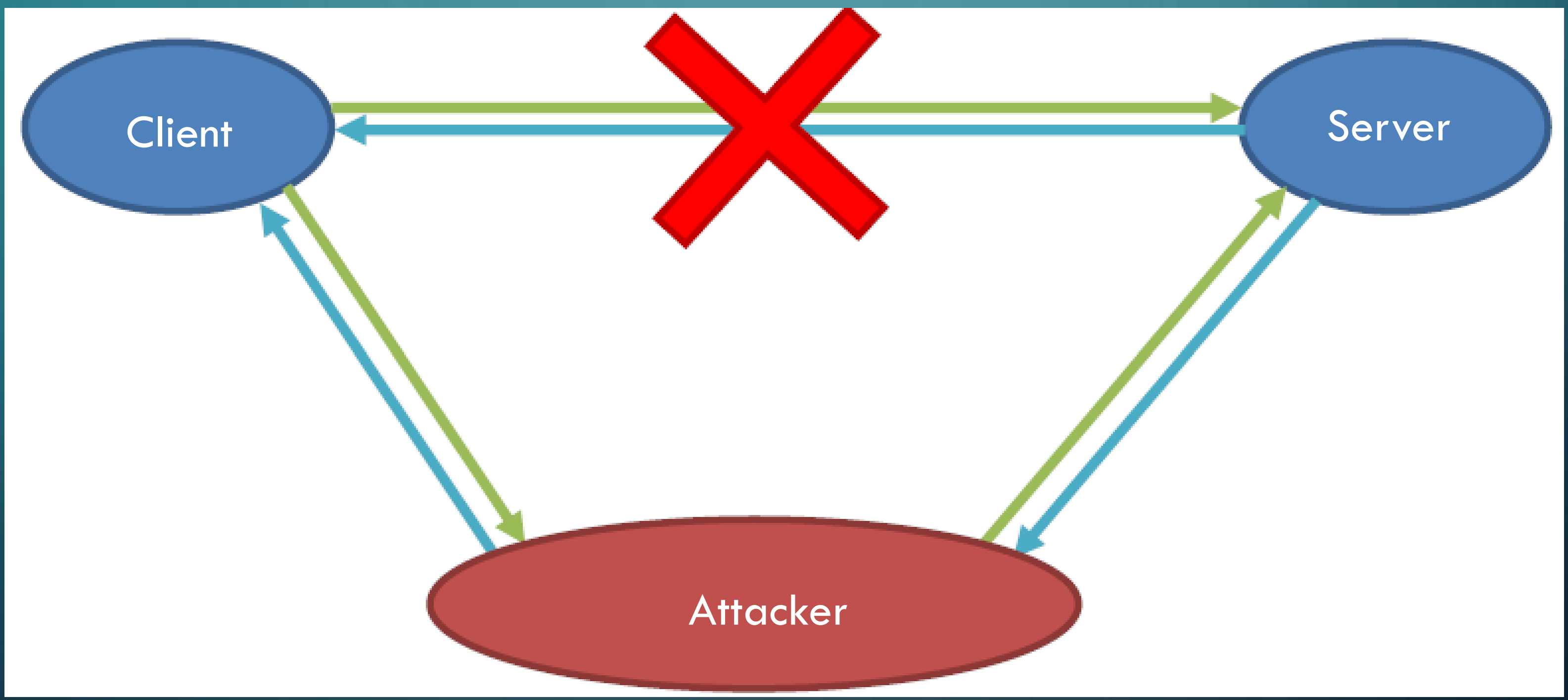
Analysis $\neq$ Analytics

Part 2

Use-cases

MAN-IN-THE-MIDDLE

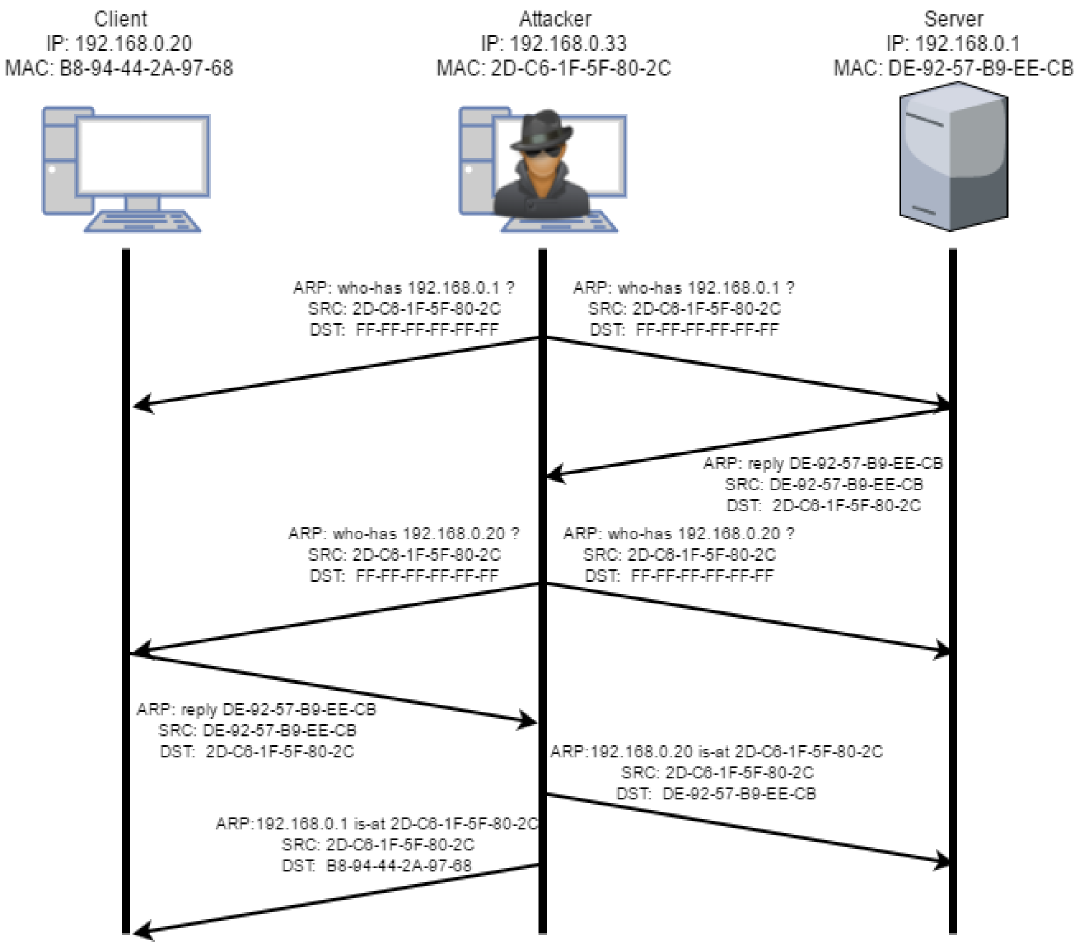
through *ARP Spoofing*



MAN-IN-THE-MIDDLE

through *ARP Spoofing*

Step-by-step

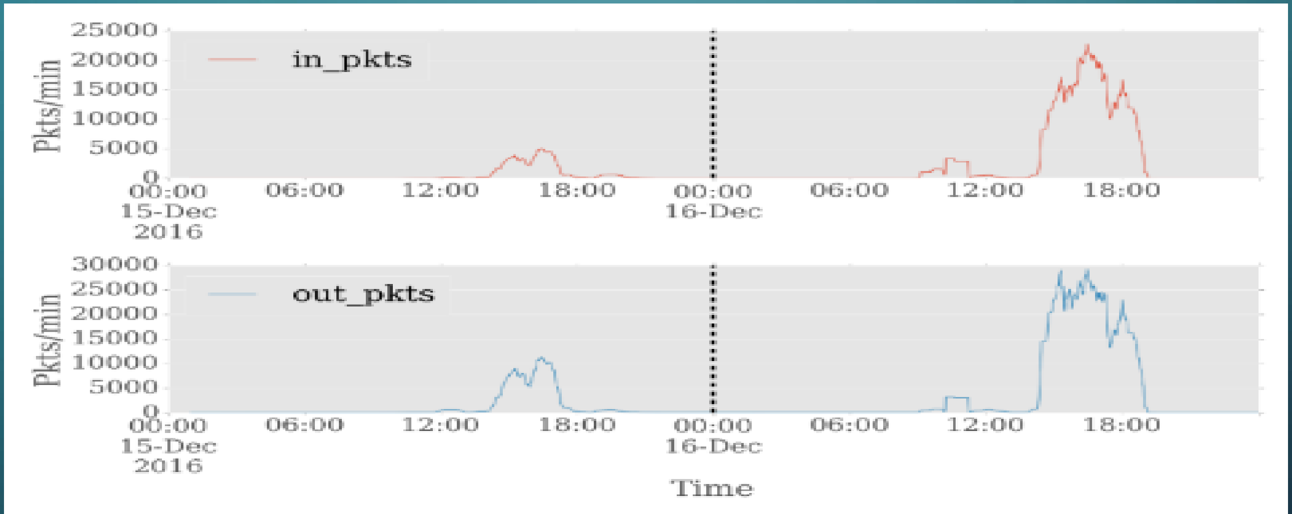


MAN-IN-THE-MIDDLE

through *ARP Spoofing*

Intuition: all packets are doubled!

Check Packets!



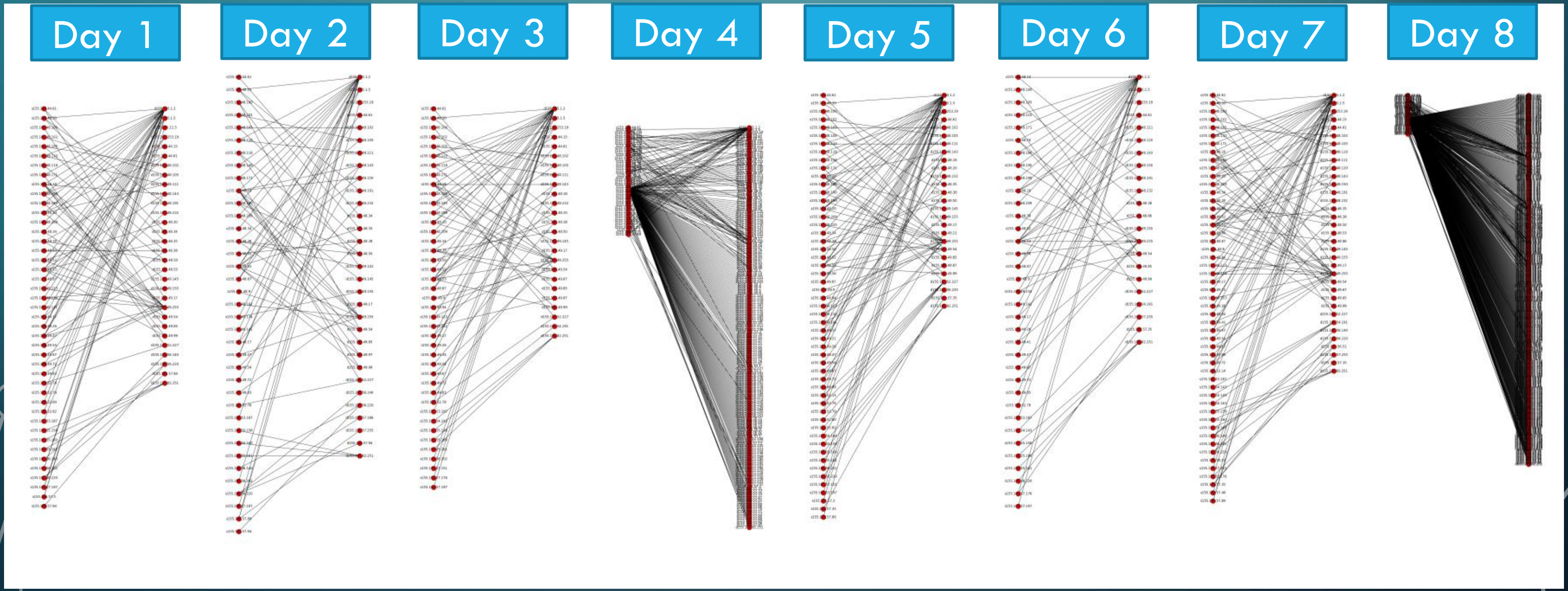
HOWEVER!

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

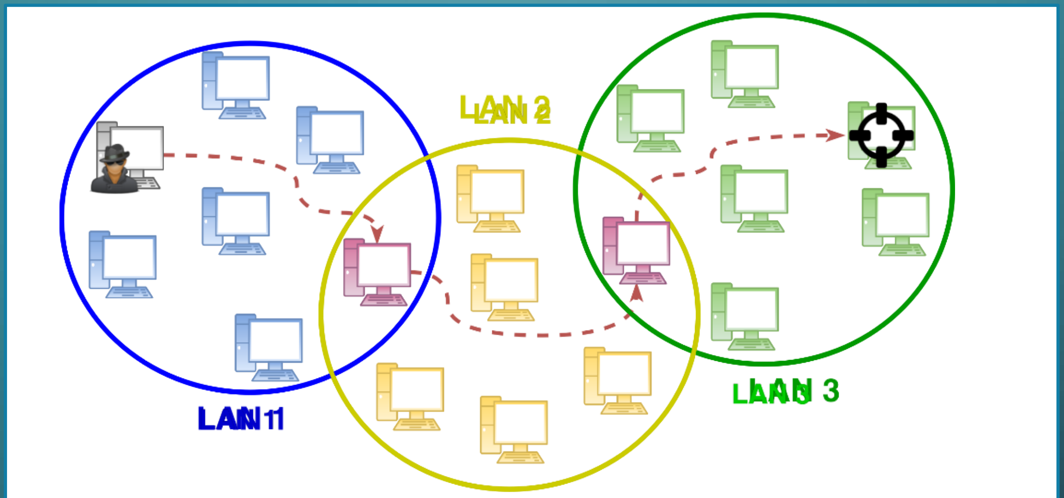
through ***horizontal port-scanning***

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

LATERAL MOVEMENT

through *Pivoting*

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



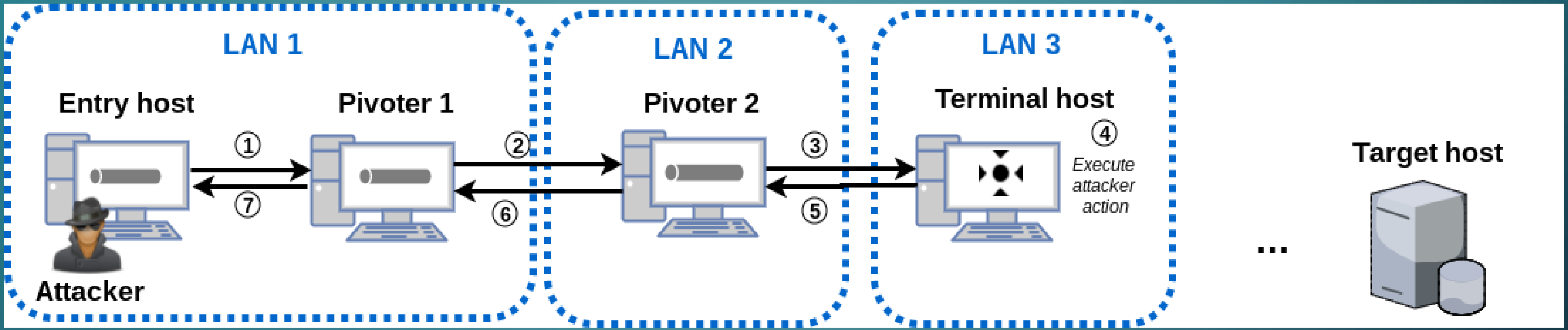
Pivoting: any action in which a *command propagation tunnel* is created among <u>three</u> or more hosts

NB: Pivoting activities are not necessarily malicious.

LATERAL MOVEMENT

through *Pivoting*

Pivoting example



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

Flow-sequence

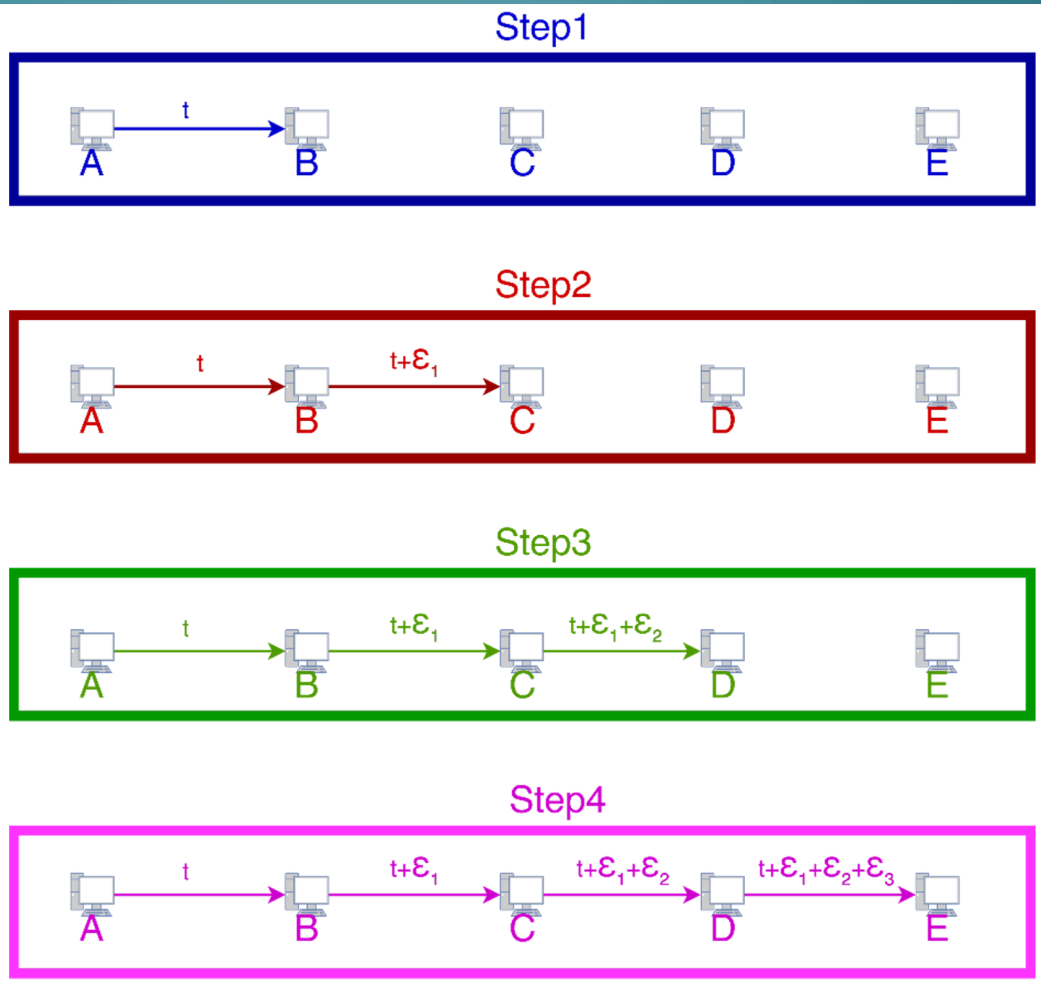
Ordered set of flows where consecutive flows are:

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

LATERAL MOVEMENT

through *Pivoting*

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



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 algorithms are effective…

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

MACHINE LEARNING & CYBERSECURITY

FORTINET

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

JUNIPER NETWORKS

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

Machine Learning: New Frontiers in Advanced Threat Detection

Symantec

SOPHOS

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

MACHINE LEARNING HELPS US FIND NEW ATTACKS

F-Secure

KASPERSKY lab

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

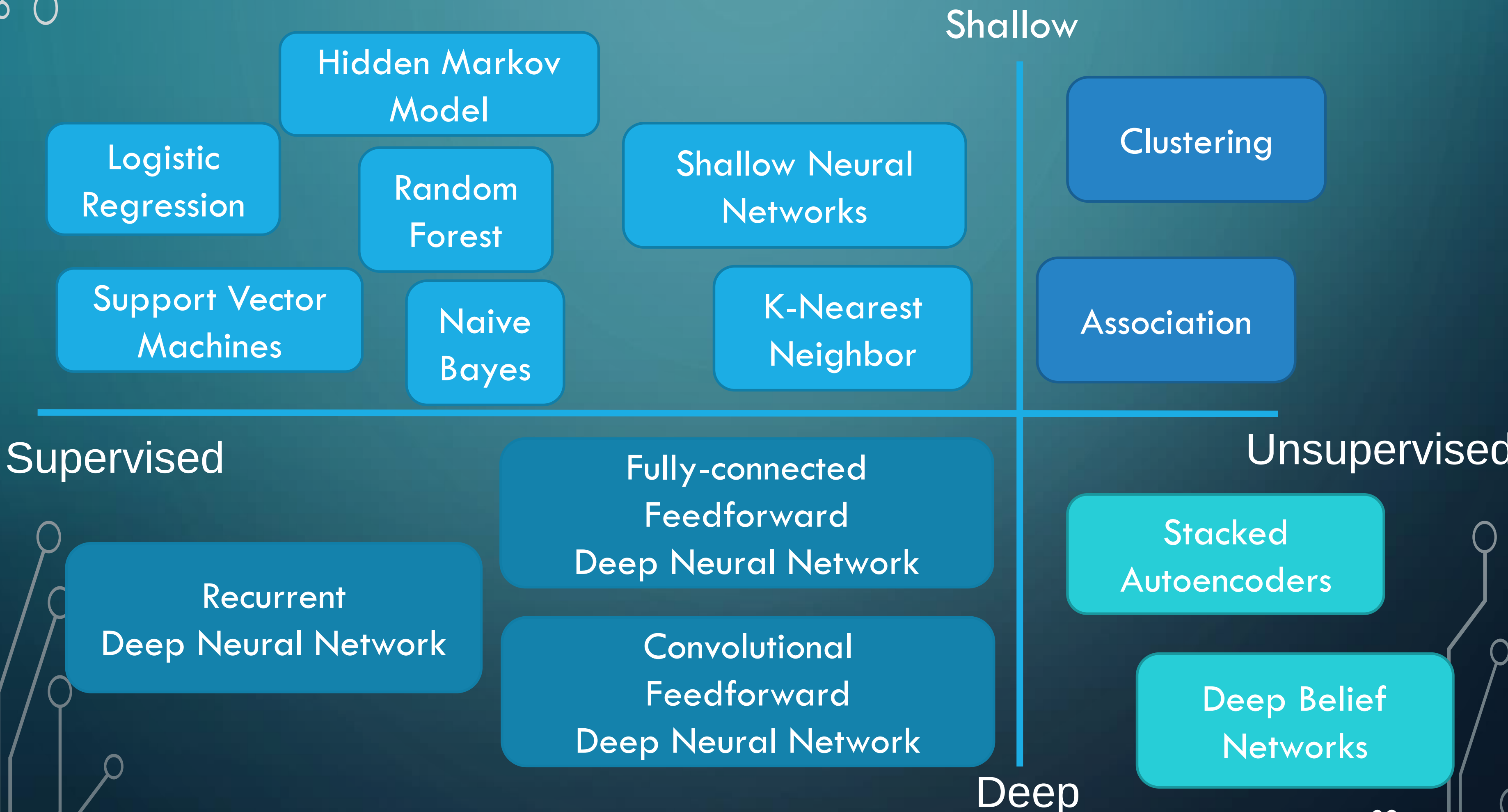
FireEye

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

RAPID7

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?

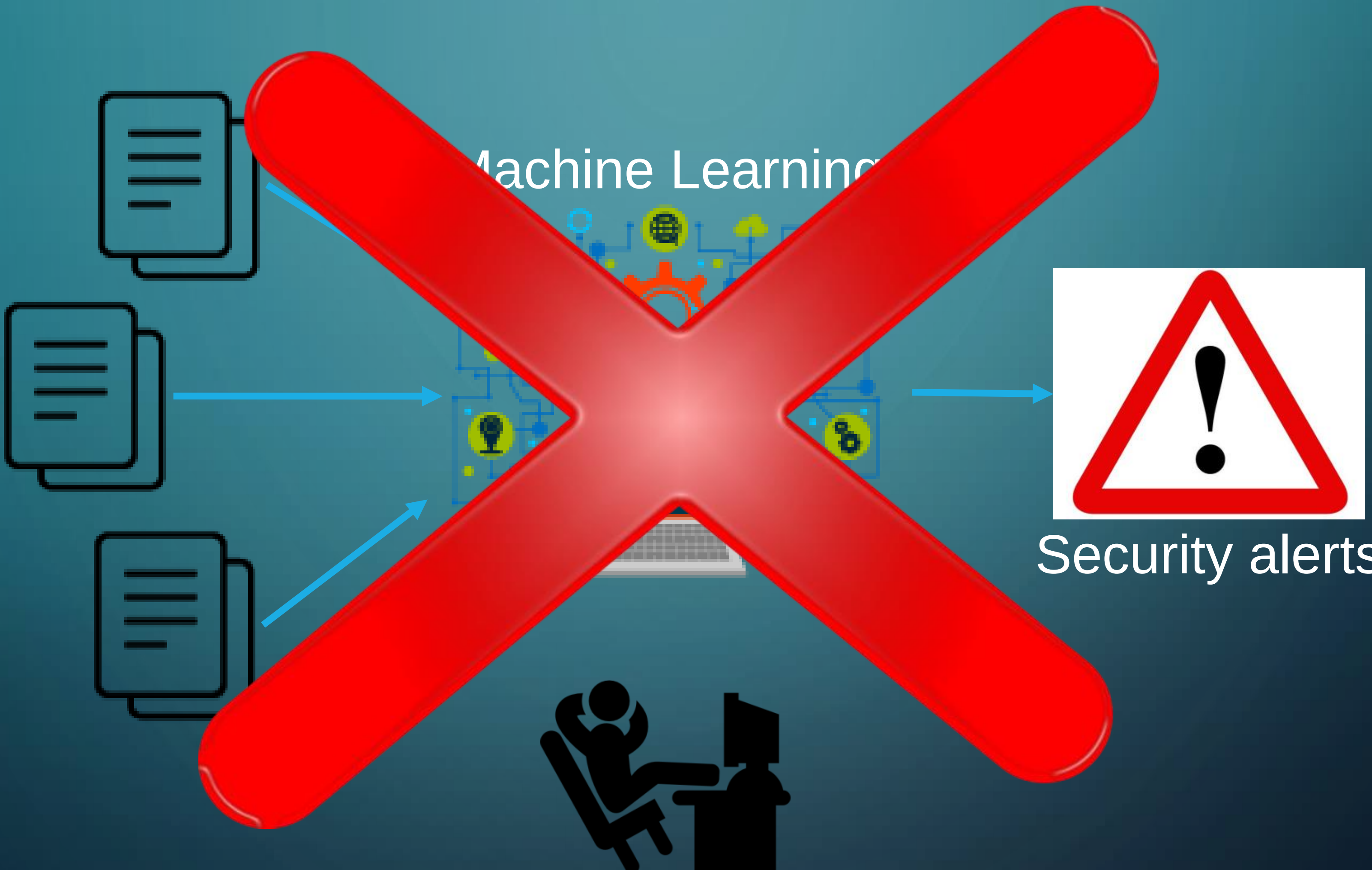
Evolution over time (concept drift)
- How frequently should the model be re-trained?

Explainability
- Results are not explainable (yet)

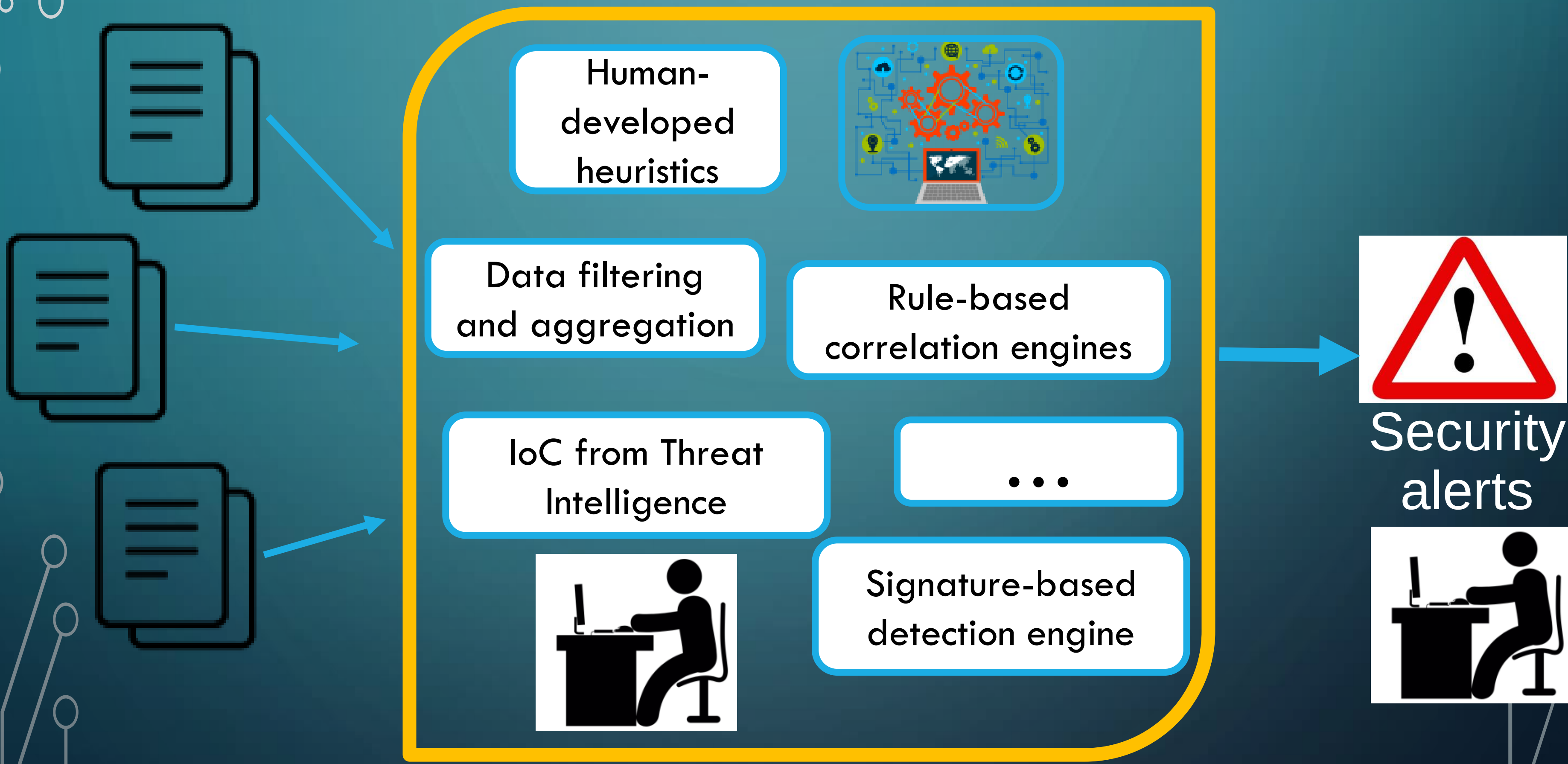
False positives and false negatives
- 1% false positive rate in large organization = **thousands** of daily false alarms

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

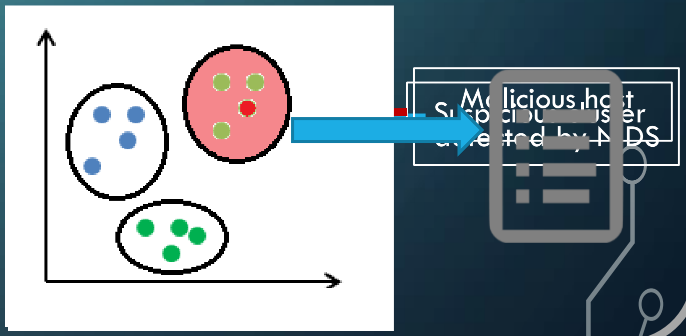
Novel malware variants are likely to evade detection…
…but some features of malware behavior persist and are shared even by novel variants

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

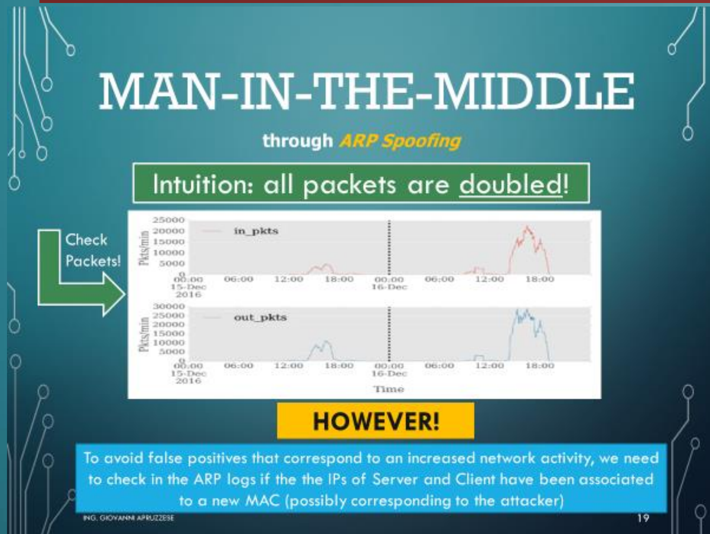
We showed several use-cases of CyberDetection:

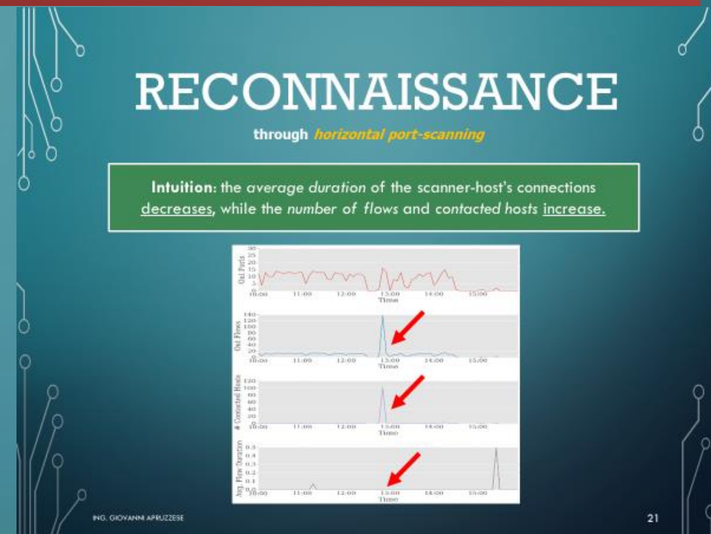
- Man in the Middle
- Reconnaissance
- Lateral Movement
- Periodic Communications

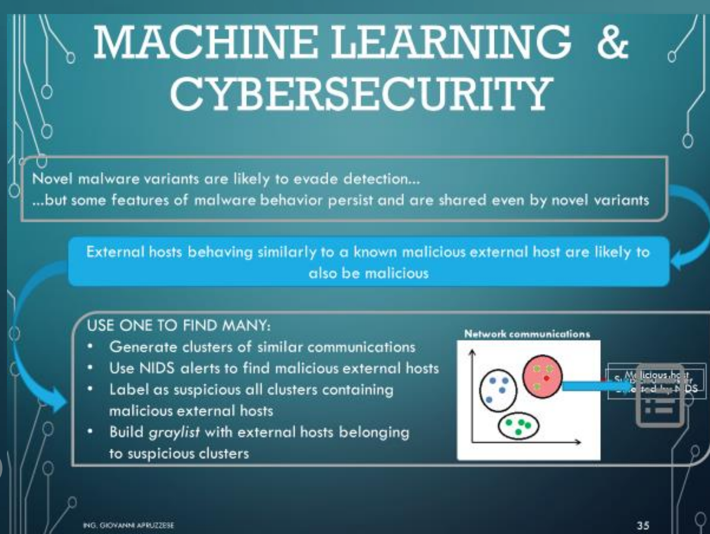
If you were an *attacker*, what would you do against these detection schemes?

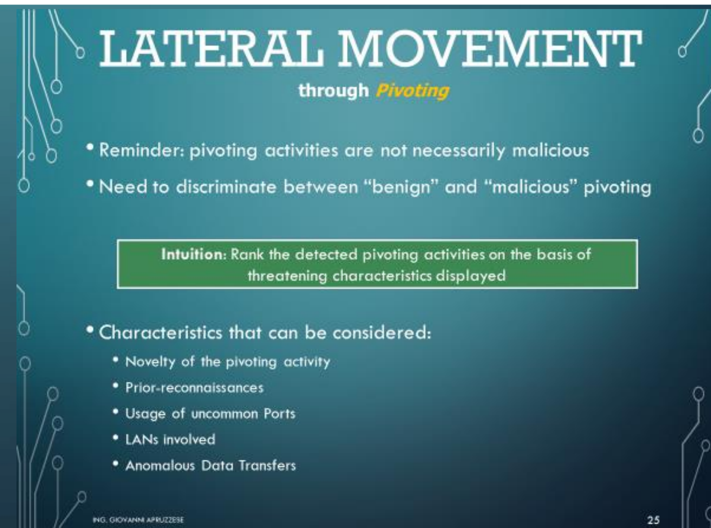
QUESTION

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









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