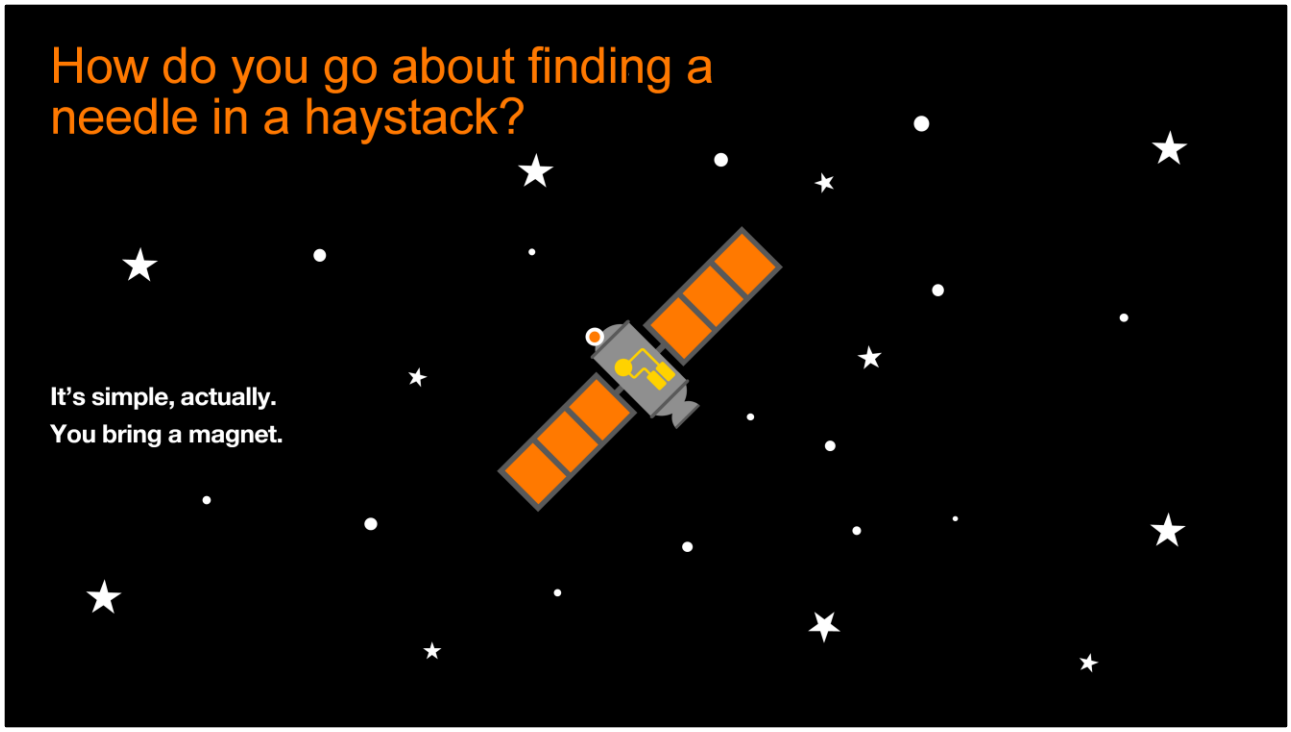
APT Hunting:

Machine Learning powered Threat Detection

Ioan Constantin,
Orange Romania

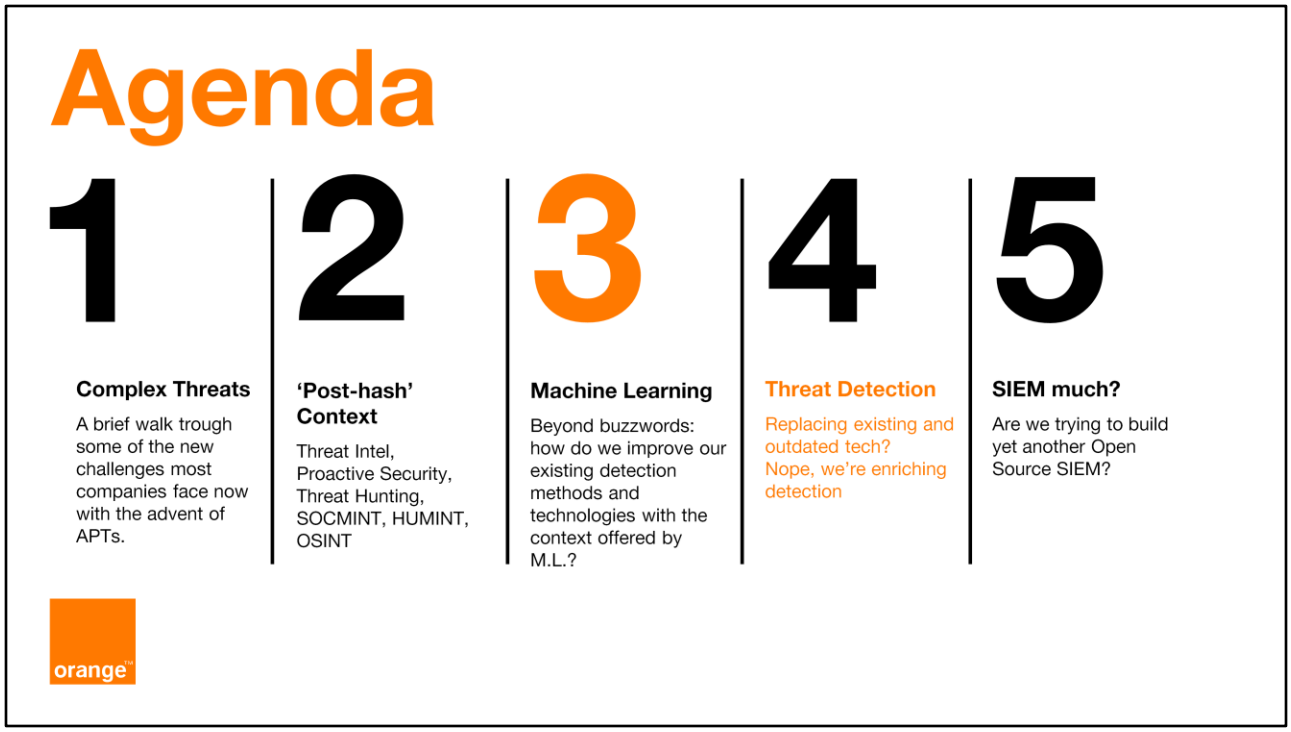
Are robots taking over? If John Connor was right, Skynet is hiding somewhere in a docker container hosting logstash ☺

-We'll skip the buzzwords and see if and how Machine Learning and Log and Event Correlation can prove helpful in detecting APTs



:Funny quote from a funny movie:

-Identifying the challenge:
-We generate a large amount of log data, telemetry, samples and captures;
-Shifting towards cloud-based infrastructure & virtualization means that the sheer volume of data sources is increasing;
-SIEM solutions are great at finding correlation between events, assuring compliance and automating response;
-They do a good job at pointing out the breadcrumbs
-We're looking into improving this response with open-source technology that can point out the entire trail ☺
-We're not looking into building cvasi-SIEMs and replacing our existing ones;
-We're simply moving forward on using advanced machine learning as a means of putting together all the puzzle pieces that can lead to better detecting persistent threats.
-…and we're doing this using Open Source and free software

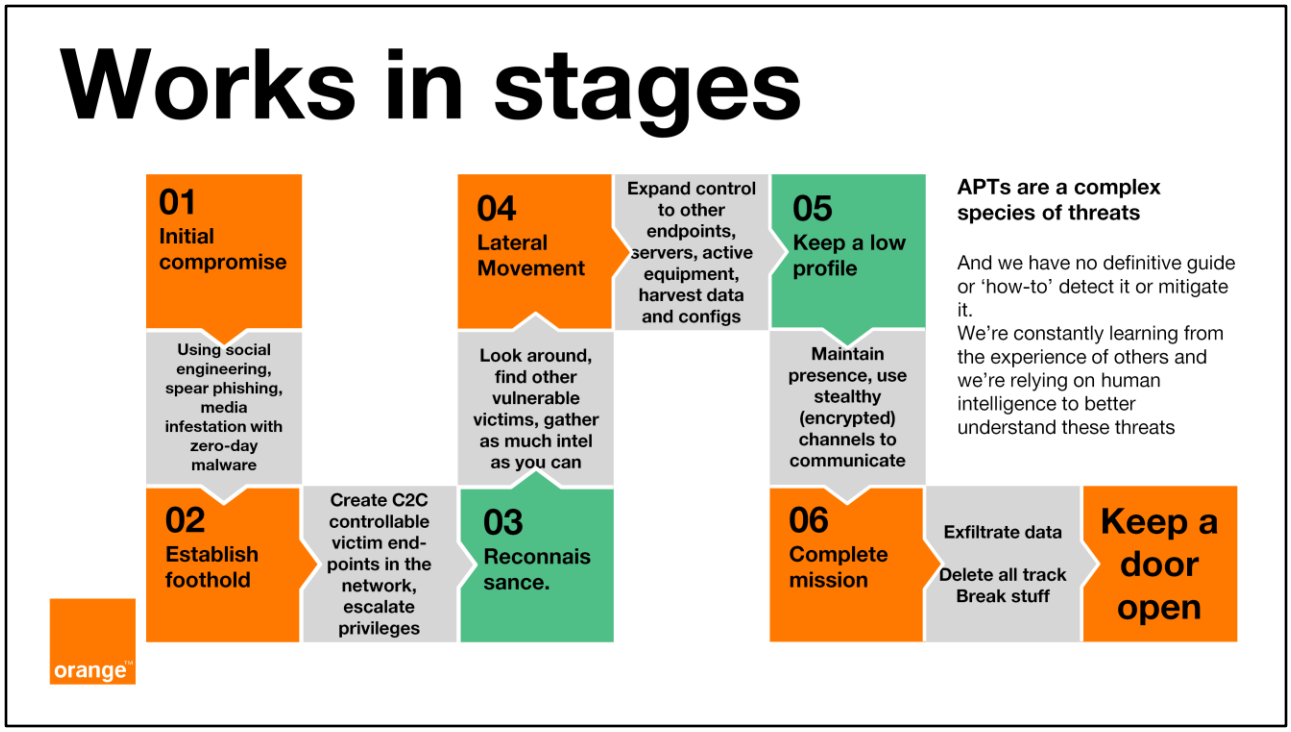


Our agenda for today focuses on 5 points and covers in a high-level presentation our quest for better resilience;
1. We'll discuss the new threats that large companies face, such as APTs; Yes, there's a timeline of well-know APTs some where in here ☺
2. What's beyond signatures, hashes, IoCs and rules? We'll talk Proactive Security
3. Machine Learning is such a overused buzzword, right up there together with Artificial Intelligence and Blockchain. We're not trying to match our smartphone camera's settings to various pets. We're trying to teach some new tricks to a computer.
4. We'll see how this implementation improves the detection of cyber threats beyond the capabilities of 'traditional' signature-based detection tools;
5. We'll discuss shortly the possibility of automation, thus closing the circle, from detection to mitigation and prevention. *Are robots going to steal the hard-working firewalls' jobs? ☺*

APT usually refers to a group, such as a government, with both the capability and the intent to target, persistently and effectively, a specific entity. The term is commonly used to refer to cyber threats, in particular that of Internet-enabled espionage using a variety of intelligence gathering techniques to access sensitive information, but applies equally to other threats such as that of traditional espionage or attacks. Other recognized attack vectors include infected media, supply chain compromise, and social engineering. The purpose of these attacks is to place custom malicious code on one or multiple computers for specific tasks and to remain undetected for the longest possible period. Knowing the attacker artifacts, such as file names, can help a professional make a network-wide search to gather all affected systems. Individuals, such as an individual hacker, are not usually referred to as an APT, as they rarely have the resources to be both advanced and persistent even if they are intent on gaining access to, or attacking, a specific target.

Some defining characteristics of these threats:

- Highly targeted;
- Highly resourceful actors with expert knowledge;
- Specific motivation – political or business (economical) reasons;
- Targets private organizations, government institutions at state-level;
- Actors have access to exploits and backdoors that are 'zero-day' or unbeknown to the general public;
- Highly reliant on social engineering;
- As per boundaries it can be considered as a attack from inside and outside;
- This endeavor usually lasts for months on;
- It is extremely hard to detect with conventional, signature-based solutions;



APTs are a new species of threats and there's no definitive guide or 'how-to"
Never the less, there are specific activities and tools usually in-use to successfully run such an attack. By studying high-profile APTs such as Stuxnet, we identified six main stages for this new type of threat:

1. **The initial compromise usually relies on highly targeted social engineering campaigns, with victims most commonly being able to provide access to within the targeted network(s). The initial vector could be a spear phishing e-mail or a media-infestation (this was the case for the Stuxnet APT). The initial driver could be malware derived from zero-day vulnerabilities with a high degree of targeting. The attacker can – for once- find out what kind of business apps the victim is using, research that software, dependencies and plugins, write (or buy) an exploit for a unkown (zero-day) vulnerability and somehow insert that malware into the target network**
2. **Gaining ground: once the malicious code is copied on an endpoint (laptop, desktop computer, smartphone, tablet) that belongs to a business network the attacker will 'listen' for covert connections, appearing as legitimate traffic to most signature-based detection tools. In some cases, the 'agent' that runs on the endpoint will communicate with the Command and Control Server over port 53, legitimately used for DNS. Most firewalls will 'see' this traffic as being legitimate as they don't posses the required intelligence to discern as to why is a computer creating outbound DNS connection when it DOES NOT receive any DNS Queries? The same applies to traffic 'hidden' over HTTP on ports 80, 8080, 443 etc. If the attacker adds encryption to the mix, most firewalls will simply ignore this traffic, per policy**
3. **Reconnaissance: Once installed on an endpoint, the malicious software can perform several discovery – type functions: it can map the network, it can silently audit the computers on the network for installed software and open ports, etc.**
4. **Lateral movement: based on the findings from the Reconnaissance stage and the**

commands it receives from the C2 server, the malware can 'move laterally' and expand to other systems on the same network or on servers, active equipment such as routers, switches (using something like the payload in MiraiBot or VPNFilter)

5. **Keep a low profile – in all stages, the malware will not alert to it's presence, will use system functionality where available to disguise as legitimate software or as a OS process. In all stages, the malware will usually use encrypted communications to 'talk' to other instances of itself or the C2 server, oblivious to Firewalls, IDS/IPSs etc. This is not a stage in the process, per-se but rather a general requirement encompassing all other stages.**
6. **Complete mission: in this final stage, the malware will actually deliver the results expected by the attacker: exfiltrate information, delete information, disable devices, encrypt disks etc. Usually this is accompanied by a final process of deleting it's own tracks – all data on disk, all relevant logs etc.**

There's a lot of specificity to each APT, as it happens. Some attacks lasts for years on (Stuxnet), others take days to compromise, act and deliver.

Technical Controls used to Protect Against APTs*

*according to a TrendMicro Survey, H1 2017

83%
Antivirus & Antimalware

Mobile Security Gateways, Mobile Anti-Malware, Mobile Device Management
37%

77%
Network Technologies (FWs, Routers, Switches)

66%
Zoning Off (Network Segregation)

64 %
Log Monitoring / Event Correlation / SIEMs / A.I. / M.L.

Key takeaway is that 83% of responders are betting on AV in fighting APTs. Not exclusively by as a first line of protection.
What about BYOD, fileless malware, zero-days & vaults?

Challenges in mitigation

1

Malware Variety

There are hundreds of million of malware variations that can be used in a APT. This makes it challenging to detect.

2

Traditional security is ineffective

As it needs more than signature or rule-based detection. It needs context and some form of automated learning method to expand the context

3

Log analysis and Log correlation has its limits

Because, of course, it also needs context and ‘perspective’.

4

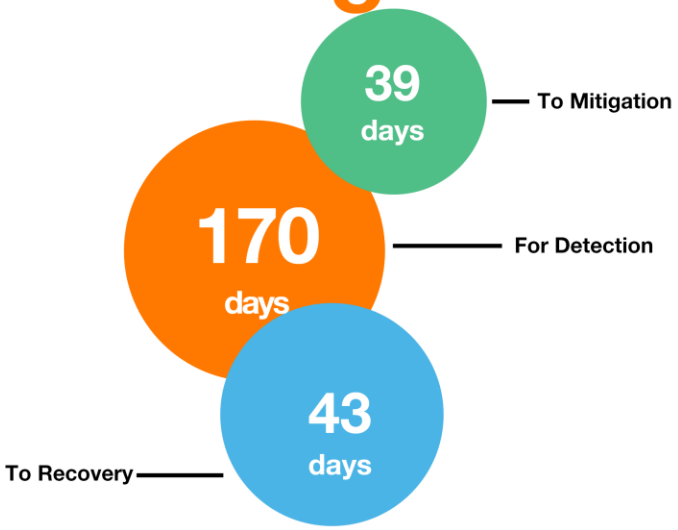
There’s lots of noise

It’s extremely hard to separate noise from legitimate traffic and there’s lots of noise being ‘ingested’ by security equipment

orange™

Challenges in mitigation

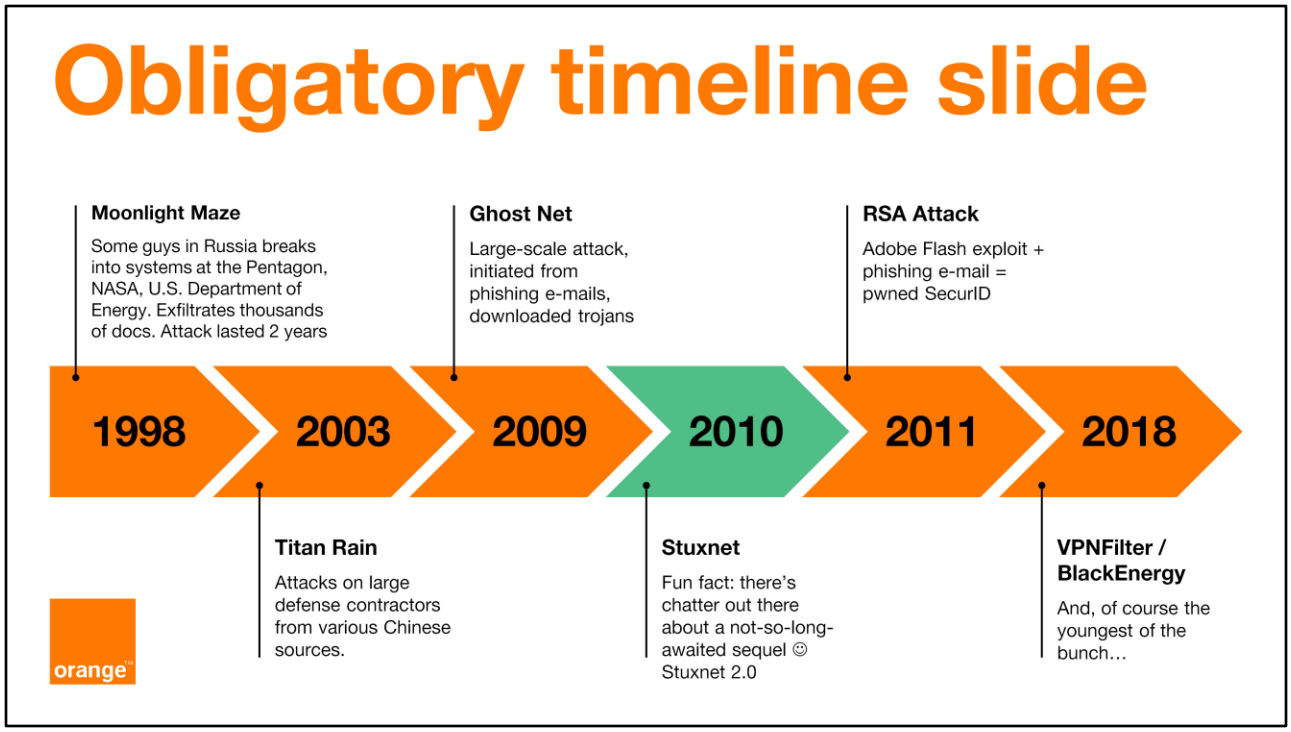
The average company takes 170 days to detect an advanced threat, 39 days to mitigate and 43 days to recover*



orange™ *According to the Ponemon Institute

170 days – . Mars One estimates their mission's travel time from Earth to Mars in the 6 to 8 months region.
That's 180 days to 240 days.

This is anecdotal evidence that it takes less time to get to Mars than to detect and APT in your perimeter.



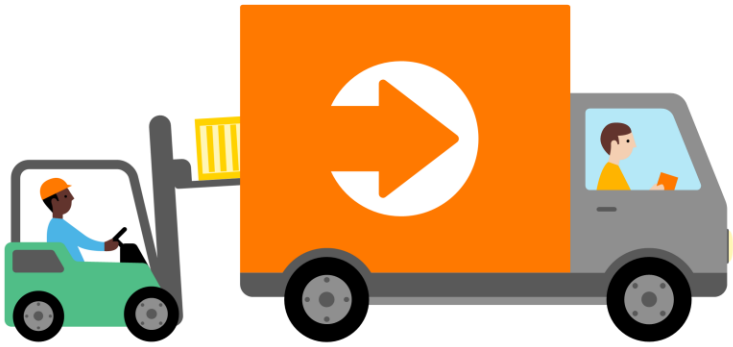
What great diversity, isn't it?
You have everything in one graph, from Russian hackers to Chinese spies, Adobe Flash and nukes. Quite the James Bond on steroids movie, right?

Solutions?

Provide Context.

It's quite the mix: your large datasets and

Machine Learning



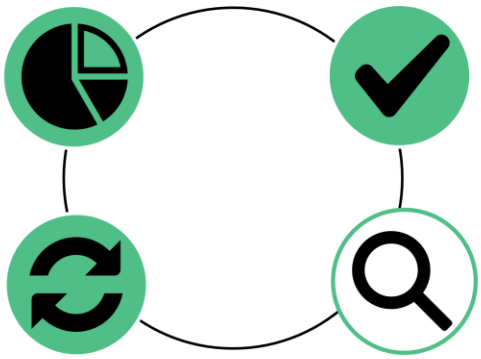
Machine Learning is the science of getting computers to act without being explicitly programmed (Stanford University, Artificial Intelligence Laboratory)

Machine Learning (…) algorithms can figure out how to perform important tasks by generalizing from examples (University of Washington)

Machine Learning

Learn from available data
Using representation, evaluation and optimisation. Use training data

Establish baseline
Define normal behaviour, define normal data.
Don't over use the word 'normal' ☺



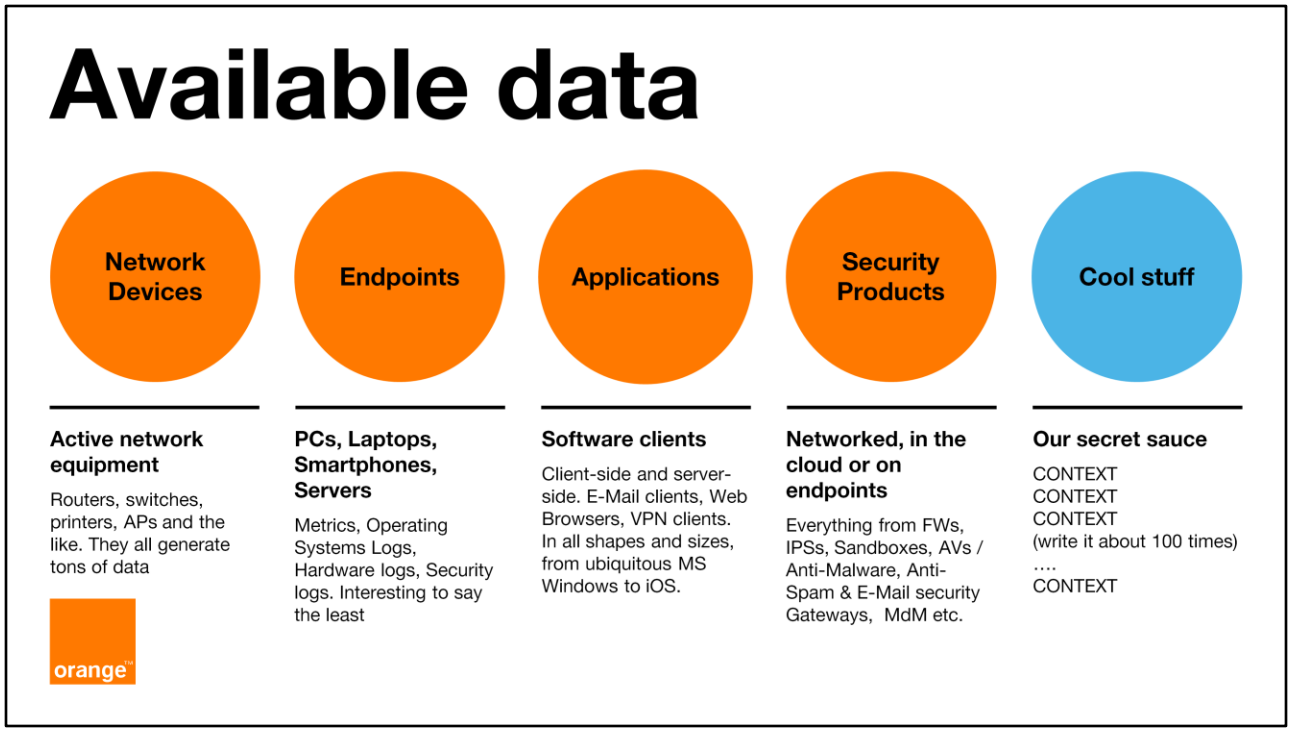
Act upon finding
Classify deviations as anomalies. Notify / Take action

Find Anomalies
Identify deviations from the baseline. Correlate deviations with other input data

orange™

Machine Learning is not necessarily a new topic of discussion among data scientist and CyberSec / InfoSec people. Various ideas and implementations have been tried and in-place for the last couple of decades. Back then, the data sets that machines had to use to learn from was usually low in volume and diversity.
The game changer happened maybe 10 years ago when smartphones, social media and apps become the norm. This drove a increase in data at least SEVERAL orders of magnitude greater that before it as internet providers and service providers adapted to a new content-model where data was GENERATED by its users rather then CONSUMED.
Large volumes of distributed data constitutes larger surfaces of attack that drives upwards the number and complexity of cyber attacks. Numerous and complex cyber attacks become more diverse as technology progresses.
There's direct causality between the complexity, numbers and diversity of cyber threats and the increasing volume of data.
From a security standpoint, each category feeds on the other.

As there's no school of thought reg. Machine Learning, various institutions, researchers and professionals tried to define the concept. We'll focus on the two definitions from the previous slide and extract key points from it. Machine Learning will:
-find patterns in large pools of data
-will establish a baseline, a 'normal' environment where events happen predictably
-will identify deviations from the baseline
-will correlate such deviations to find patterns that can be labeled as anomalies
-(it can) automatize processes and act upon finding.



There's a challenge: Design, Build and Operate a large Telco Infrastructure, deliver powerful various services in a secure, reliable way by using custom built, state-of-the-art technologies, equipment and software. Oh, and try keeping it legacy-compliant for those customers who still love their 2G phone and don't like spending too much time on the internet.

Unfortunately, this is utopic. The reality is that we rely on multiple vendors, various technologies, both commercial and developed in-house. Most of these adhere to some kind of standardization for things like communications protocols. They all like the OSI Layer when it comes to 'being talkative' but unfortunately for security gals and guys, they talk different languages when it comes to security logs and feeds.

Our equipment, software, hardware, firmware and middleware generates a HUGE amount of data relevant to security in various forms. Some like HTTP Restful APIs and JSON. Others like Schemes and XML, there's the ones that will send you CEF and then there's those that only 'speak' encrypted. Then there's our language. We mostly speak Romanian when discussing threats. Some speak English or French. And then there's the Internet. You have Social Intelligence where people write things in forums, on facebook, on twitter, on reddit. Some use sarcasm, others use abbreviations. Some LOL, some are L33t.

Ever browsed virustotal.com? Imagine making business sense from virus samples and md5 hashes. You'll probably end up with a headache. How about reverse engineered exploits? Memory dumps?

You can figure out how easy is to get lost in such an avalanche of information, generated, written or spoken in different languages, using different constructs and schemas.

We're a Telecommunications Operator. We anonymize all the data we gather, into datasets:

- Cellular data: Location data, Radio information, Device data; Data usage; Voice usage;
- Wireless data: SSIDs used, Radio information, Cellular to WiFi Roaming info;
- Security data: Website and applications usage monitoring, malware activity, attack activity

We collaborate with other actors and we rely on proven sources:

- Threat intelligence feeds: Open, free and paid (such as OTX, Virustotal)
- Malware analysis:
- Threat hunting activity: We use Threatmap as a scanning engine for compromised websites (OWASP Vulnerability Scanning, CMS-Specific Scanning, Malware Scanning)

We're white hats & pen-testers:

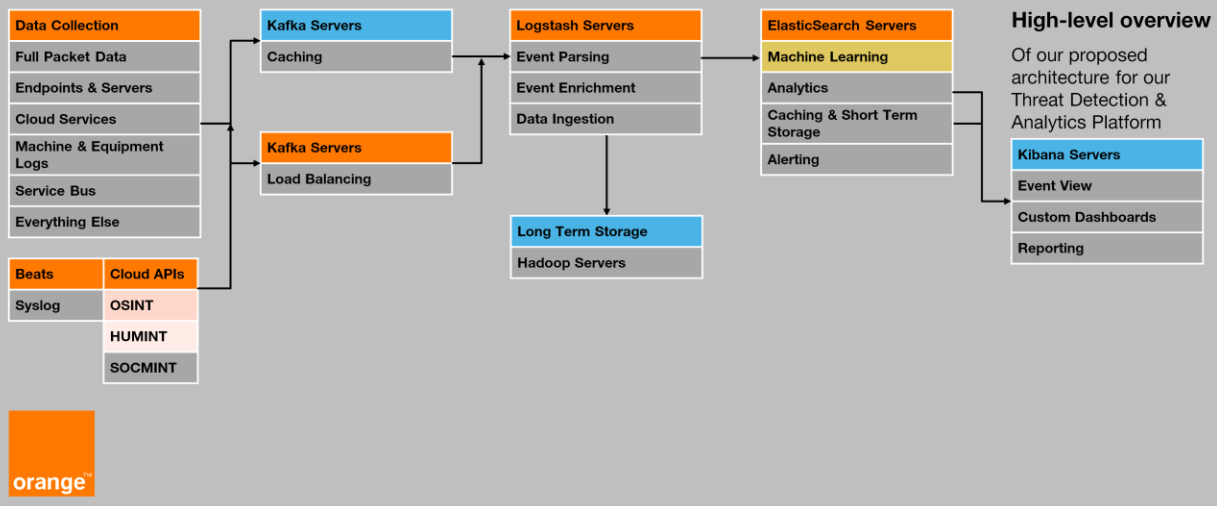
- Vulnerability assessment information: We use up-to-date vulnerability

scanners (mostly open-sourced, free) to identify potential targets for current threats

- CVEs feeds and zero-day info: We input feed data about the latest reported vulnerabilities and affected systems & software and we use social intelligence feeds, with info gathered from discussion forums, reddit, twitter.
- Pentest results: we use info gathered from pentest to better understand the baseline of the current threat model

Architecture

Architecture - II



Modern IT systems are built from larger numbers of individual systems interconnected to satisfy the desired end goal. These systems generate logs and events which can provide valuable insight into the system's current state and internal workings. The more data ingested and analyzed, the better the decision. Currently, given the reduced costs of storage, we target massive collection of useful data from across the entire IT infrastructure and application environments.
Over the years, as the data has grown to Petabyte level, the technologies capable of storing and manipulating these amounts of data has shifted from a traditional file based approach to

unstructured no-sql databases, capable of adapting to any type of data and volumes of ingestion.
Currently Elasticsearch and Hadoop are amongst technologies capable of handling large volumes of data, while at the same time providing for fast searching capabilities and customized visualizations and dashboards.
With the new addition of Machine Learning and Alerting, we can now provide fast analytics and anomaly detection as well as generating alerts concerning threat-related events identified within the client's infrastructure

We are using leading industry no-sql, hdfs and in-memory technologies and consists of:
Data Collection - Open Source and commercial technologies are being installed and customized for collecting data from any source (device, application, sensor, etc.) across the client's environment (cloud and/or on premise) using standard, broadly used, protocols such as syslog, nslog, ftp, smtp, snmp, https or using stand-alone light software agents or API's, if/when available.
Kafka nodes - To ensure high availability for the collection of information within the data ingestion pipeline, we use Apache Kafka

therefore being able to sustain longer periods of downtime in case of disasters.
Logstash/other log shippers nodes - Logstash is the heart of data enrichment, providing abundant capabilities in parsing events from any data source by enriching the data with geo-IP information, customer internal DNS information, threat intelligence and other custom enrichment.
Hadoop cluster: Used for long term storage, and custom post-event enrichment; also stores an unaltered version of your raw data which may be needed for other purposes such as compliance.
Network Load Balancer – Component acts as a one-point collector for the data sources that are trying to send events into the solution; the NLB forwards events to the Kafka servers. NLB is configured with one or more virtual IPs for HA reasons.
Elasticsearch cluster: This is a no-sql database technology, capable of ingesting, storing and searching any amounts of data. Build around a standard cluster design, it provides integrated high availability and can horizontally scale automatically, allowing it to store from tens of GB to hundreds of TB of data. With its new addition of Machine Learning (Prelert), we are

currently capable of detecting anomalies within hundreds of GB of events in seconds, alerting you about a potential infrastructure or security issue.
Visualization nodes: Allows SOC personnel as well as the customer's security team to easily visualize and interact with the data and create custom dashboards.

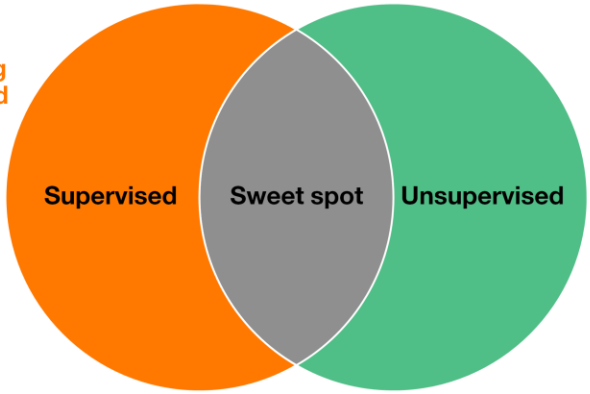
Summary:

We are using a highly scalable containerized deployment of Elastics Search, Logstash, Kafka, Beats and Kibana
We have pipelines that run across the VMs, with one machine's output becoming the next machine's input.
We wrote APIs for custom-built things.
We used the provided APIs for proprietary things.
We ingest syslog where we need it
We capture netflow where needed.
We ingest, store, normalize, anonymize and finally analyze data.
We present it via custom dashboards, in single-pane-of-glass style
We generate reports

Machine Learning Algorithms

Supervised Machine Learning relies on feeding the machines with labeled data, be it 'good' data or 'bad' data.

This helps create a baseline of expected behaviour, in threat detection and, in turn, will yield results in detecting anomalies

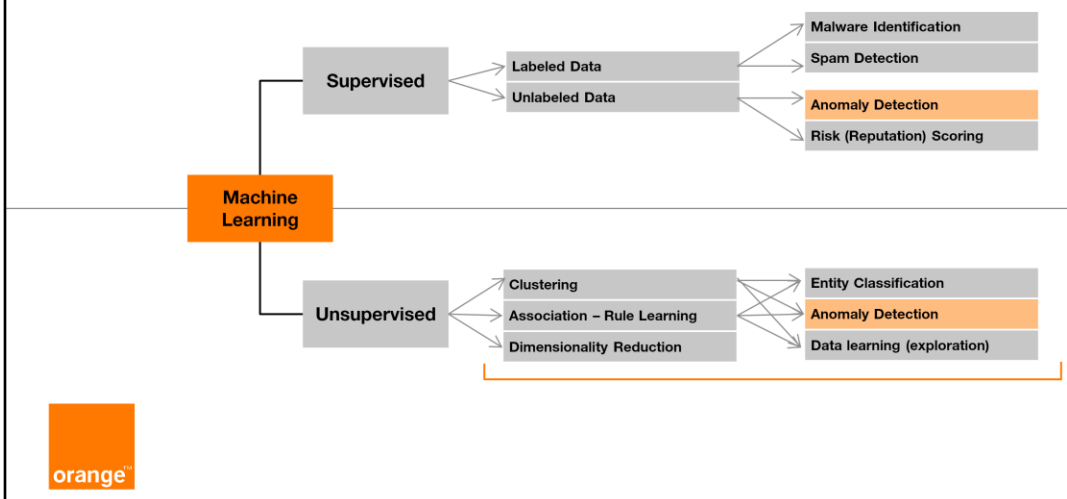


On the Unsupervised side you have to consider several approaches such as dimensionality reduction and association rule learning.

These approaches are useful in making large data sets easier to analyse or understand. They can be used to reduce the complexity (dimensionality) of fields of data to look at or group things together (clustering)

orange™

Machine Learning Algorithms



orange™

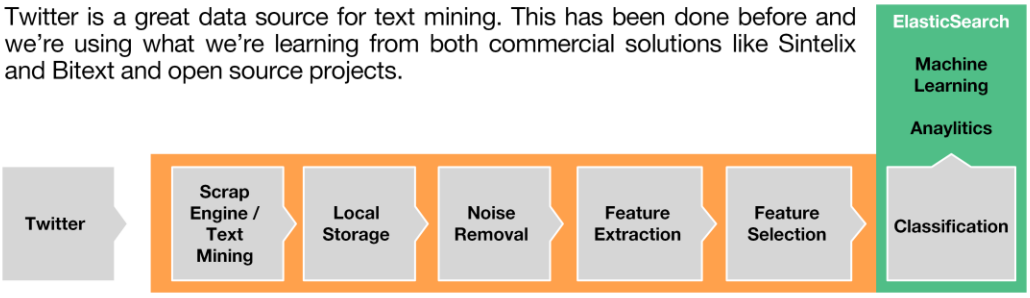
On the supervised side we have the two poster-use-cases: Malware identification and spam detection

On the unsupervised size we're dealing with classification and feature selection for better understanding the data itself.

We're not bound to one method so we can run both types of algorithms on different data sets. We can use the output of our unsupervised algorithms as input for our supervised algorithms, for once.

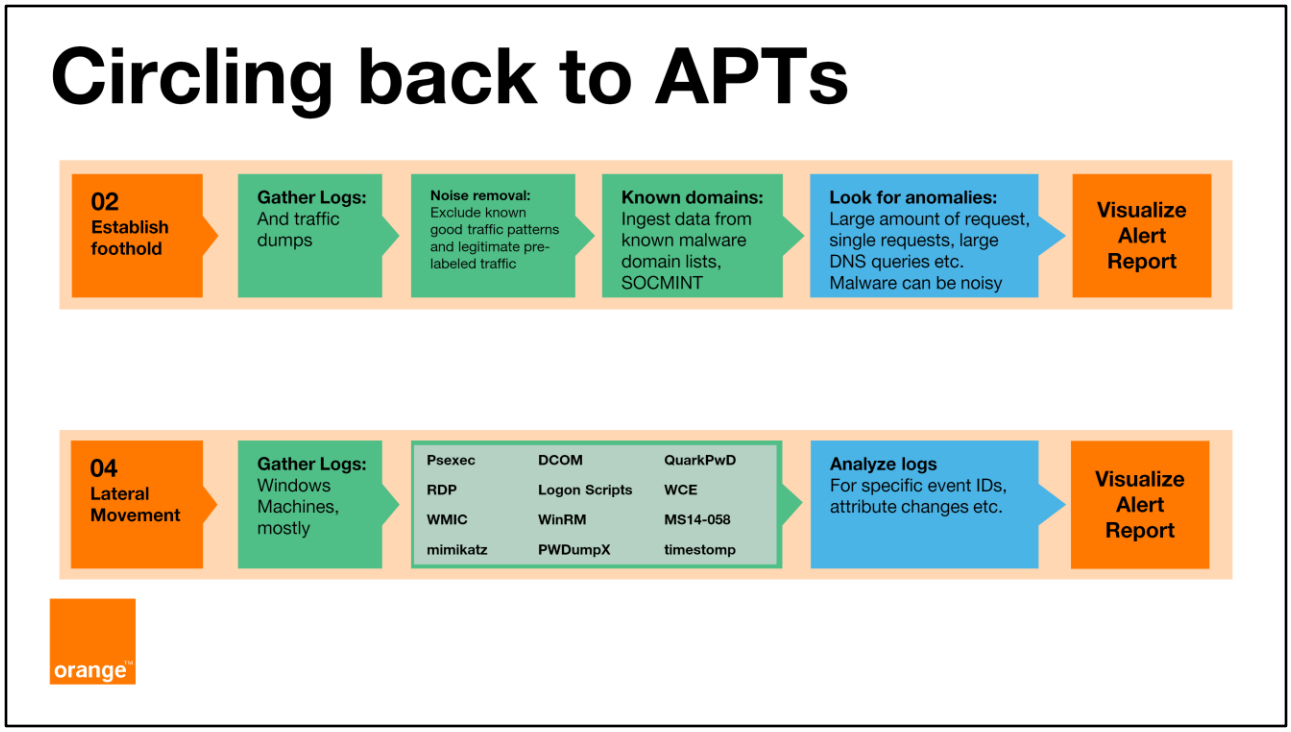
(More than) Text Mining

Twitter is a great data source for text mining. This has been done before and we're using what we're learning from both commercial solutions like Sintelix and Bitext and open source projects.



orange™

We're mainly using Text Mining for classification of alerts, newly reported vulnerabilities and reported attacks.



Okay. What about practical uses?
M.L. algorithms and SOCMINT can prove helpful in detecting APTs and its associated activity like establishing a foothold in the target network(s) and move laterally.
For the first activity, we'll focus on analysing logs and traffic dumps. We'll start by excluding legitimate 'labeled' traffic and we'll continue with feeding our training algorithm with 'good' traffic patterns. We'll use public-open data sets such as those indexed by Kaggle or UCI ML Repo and we'll eventually recycle or own data sets that provide a high-enough level of certainty on their safety.

We then compare outgoing connections to known malware domain lists such as C2 server domains. We ingest these lists from public indexes and from internets chatter (reading tweets)

We're adding an additional loop where we're writing algorithms that correlate seemingly unrelated events to possible threats, such as failed login attempts, keyboard interactive or over VPN-remote, for a given period of time, with repetition or large DNS queries that might hide malicious traffic.
Once our training algorithms have a good grasp on what is 'ok' they can point out what's above or below the baseline – anomalies. Add the second loop in the mix and there you have it, a possible correlation between seemingly unrelated events that in turn will alert the security guys to take action.

How about lateral movement?
There A LOT we can learn from logs, event IDs and Sysmon logs, specially from Windows machines. The tools used by an malicious actor to move laterally in a network can be either Microsoft-made, signed and trusted by the OS on other machines or custom tools made by 3rd parties. What they all have in common is that they generate enough 'breadcrumbs' that – when pieced together – will tell us if their use was legitimate or illegitimate.
We can elaborate a training set for our ML that cycles trough various events gathered from logs and matches those to malicious activities such as privilege escalation by use of *bullets – mimikatz.

Once the training set is in place, the machines can monitor seemingly ‘normal’ activity from the logs gathered and tell if the succession of recognized events had a legitimate outcome or a malicious one.

Expected Output

Dashboards

Configurable dashboard that can be customized to show both live data and analytics for any number of categories such as Malware Detections, C2 Server activity, Phishing Attacks reported trough SOCMINT and our own sensors etc.

Alerting

Customizable alerts per events such as live attacks, threshold breaking, Social Sentiment Analysis and -possibly- any delta over a preset baseline

Reporting

Detailed reporting on all incidents, incident categories, sources of attacks, types of threats etc.

In closing, some nice stats*:

***From a Ponemon survey conducted in H1 2018**

$2.5m

Savings

Companies using M.L. to detect threats save an estimate of 2.5 million US$ in operating costs

60%

Improves Productivity

Companies are positive that deploying M.L.-based security tech improves the productivity of their security personnel.

69%

More speed

The most significant benefit of using M.L. for threat analysis is increased speed. 64% say that the most significant advantage is the acceleration in the containment of infected endpoints, devices, hosts.

60%

Identify Vulnerabilities

Sixty percent of the respondents stated that M.L. identified their application security vulnerabilties

orange™

Thanks ☺