AIRBUS CYBERSECURITY

European Cyber Week / C&ESAR Conference

Artificial Intelligence and cyber security

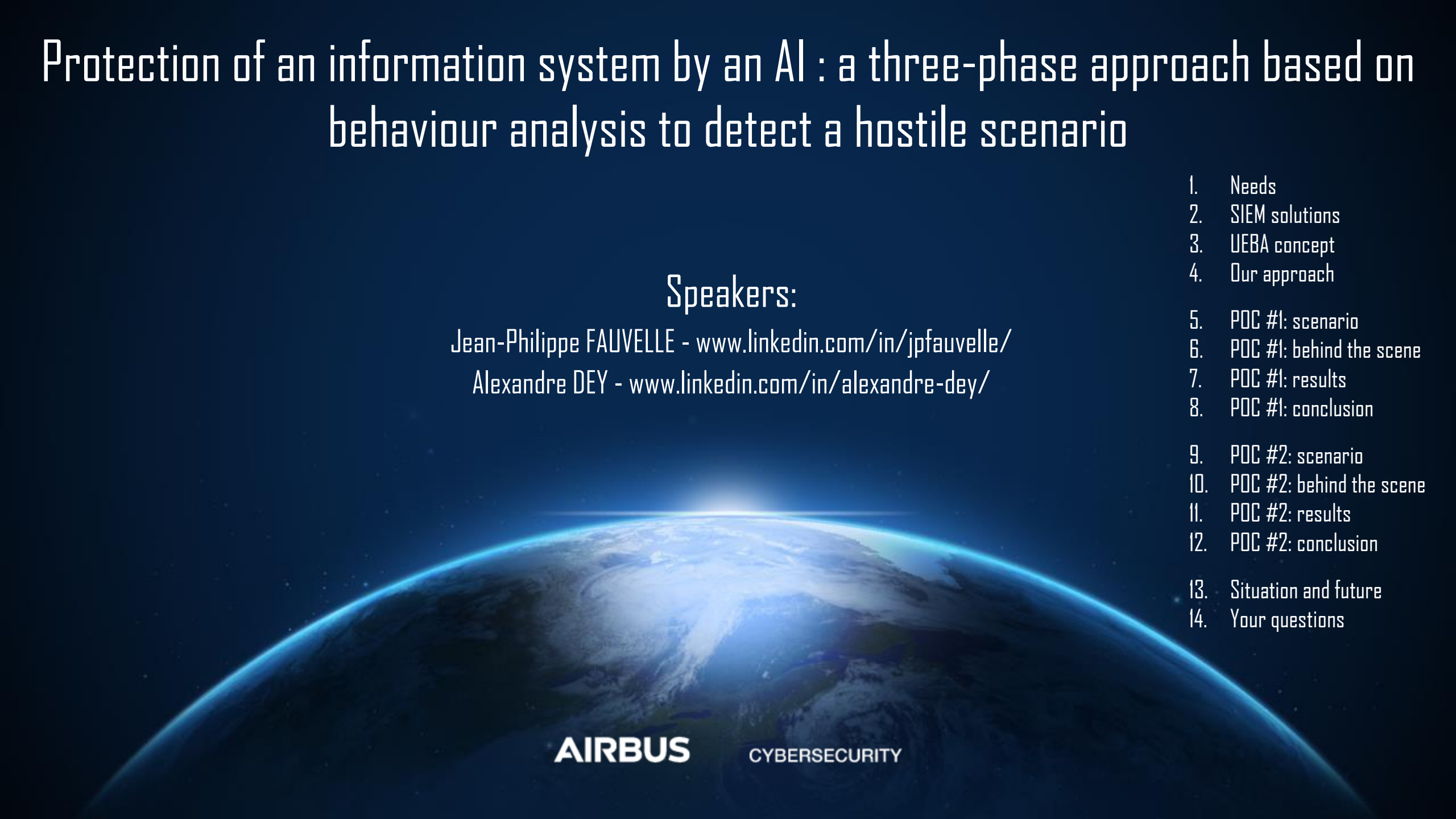
19-22 November 2018 / Rennes / France

Protection of an information system by an AI : a three-phase approach based on behaviour analysis to detect a hostile scenario

Speakers:

Jean-Philippe FAUVELLE - www.linkedin.com/in/jpfauvelle/

Alexandre DEY - www.linkedin.com/in/alexandre-dey/

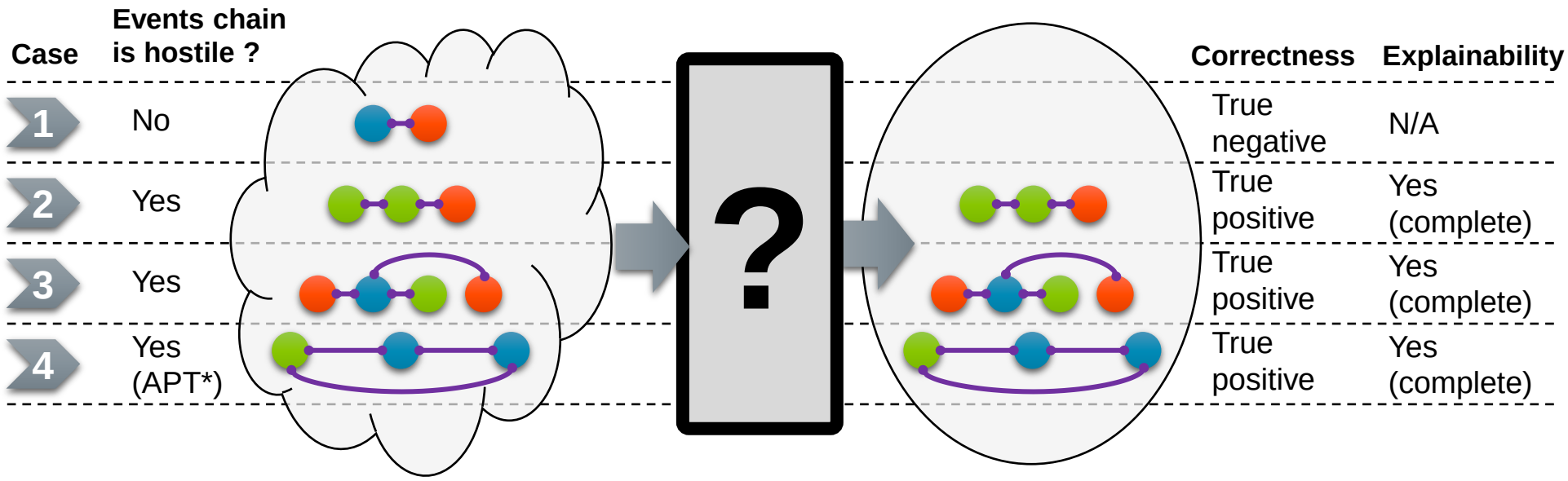


Real world

- ❑ Growing and evolving threats.
- ❑ Hostile actions over wide time periods, including APT*.
- ❑ Cyber and non-cyber events.
- ❑ Weak signals, noises, pollution.
- ❑ Increasing volume of data.

Main needs

- ❑ Detect hostile actions over wide time periods, including APT*.
- ❑ Produce **explainable** alerts.
- ❑ Automatically adapt to changing threats and behaviors.
- ❑ Reduce false positives/negatives.
- ❑ Horizontal scaling.



Events chains spread over a wide time period

Expected detection and quality

Average signal Weak signal Strong signal

(*) APT – Advanced Persistent Threat

A B

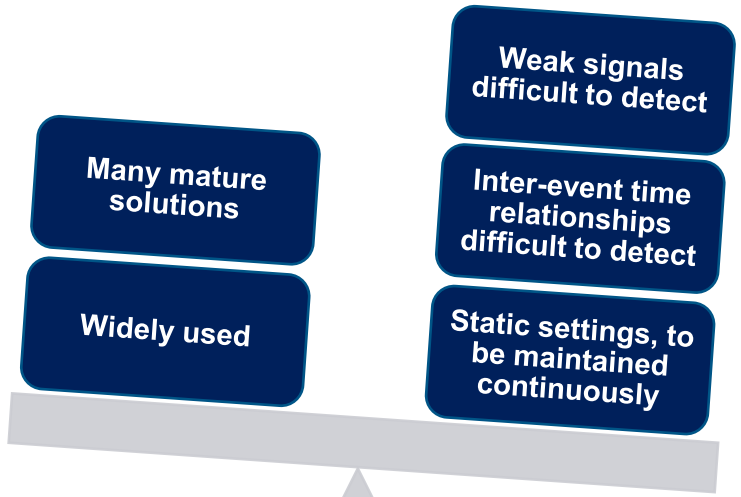
1. Needs

A. Real world: 4 cases to illustrate detection completeness and quality.

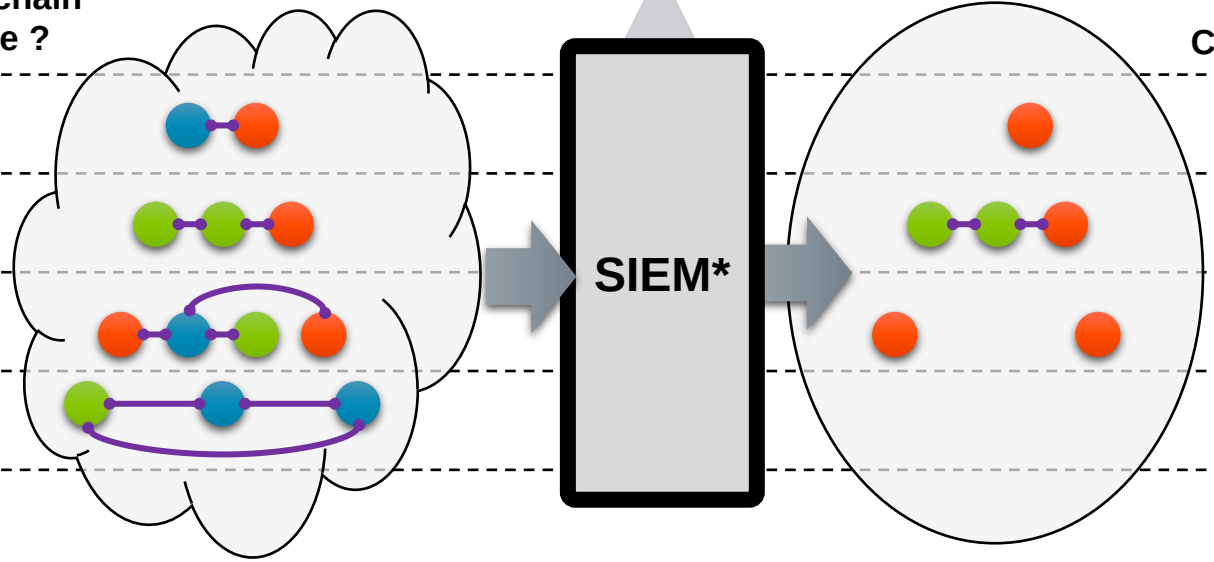
B. Needs.

AIRBUS

Pros & Cons Of current SIEM* solutions



Case	Events chain is hostile ?	Correctness	Explainability
1	No	False positive	N/A
2	Yes	True positive	Yes
3	Yes	True posit. Two alerts	Yes (partial)
4	Yes (APT)	False negative	N/A



Events chains spread over a wide time period

Usual detection/quality of SIEM* solutions

Average signal Weak signal Strong signal

(*) SIEM – Security Information and Event Management

2. SIEM* solutions

A. SIEM pros and cons.
B. Four cases to show limits.

QUICK FACTS CONCERNING UEBA*

- ❑ Learning of behaviours.
- ❑ Method agnostic to Good/Evil: detects behaviour changes (incongruities).
- ❑ Two training methods:
  - Once for all training (eg: embarked).
  - Continuous training: assimilation and forgetting of behaviours, permanent adaptation, non-supervised.
- ➧ **UEBA with continuous training meets our needs.**

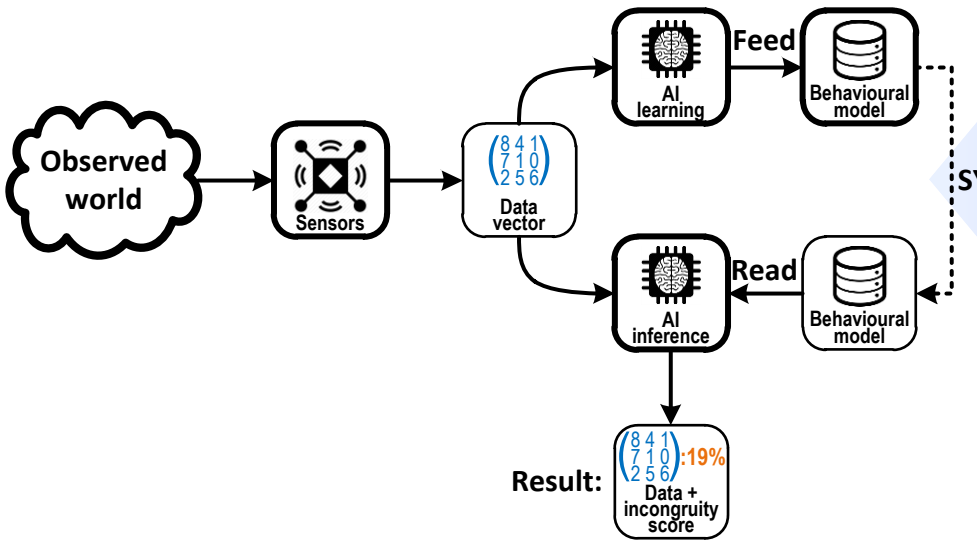
But

MAIN BIASES OF AVAILABLE SOLUTIONS

- ❑ Training performance.
- ❑ Many false positives (or negatives).
- ❑ **Slightly explainable result** (black box).
- ❑ Over-simplification of problems to solve.
- ❑ Almost systematic presence of a simple time window alerts counter.
- ❑ Little consideration of events temporality.
- ❑ Low management of behavioural model, boiled frog paradox (see below).

More

UEBA PRINCIPLE AND BOILED FROG PARADOX



- ❑ Assimilate new behaviours:
  ▶ Need for quick synchronism.
- ❑ Avoid boiled frog paradox:
  ▶ Need for slow synchronism.

➧ **Conflicting needs: synchronism is an unsatisfactory compromise.**

(*) UEBA – User and Entity Behaviour Analytics

A B C

3. UEBA concept

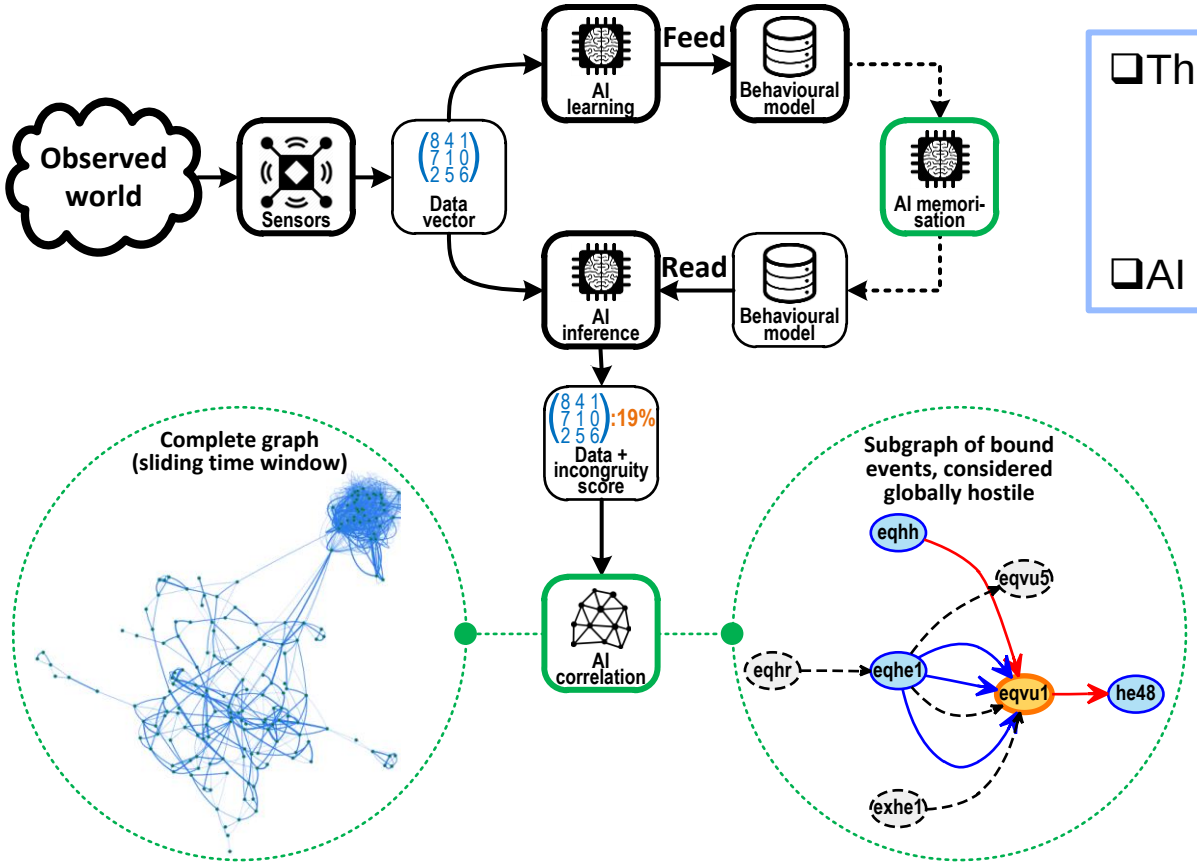
A. Facts concerning UEBA.
B. Biases of current solutions.
C. Principle overview and boiled frog paradox.

AIRBUS

APPROACH

- ❑ POC #1 (finished): simulated activity on an information system (with synthetic data).
- ❑ POC #2 (almost finished): real activity on a workstation (with real data).
- ❑ Keep in mind **biases**.
- ❑ Focus on **explainability** of results.
- ❑ Continue the work with a PhD Thesis (2019).

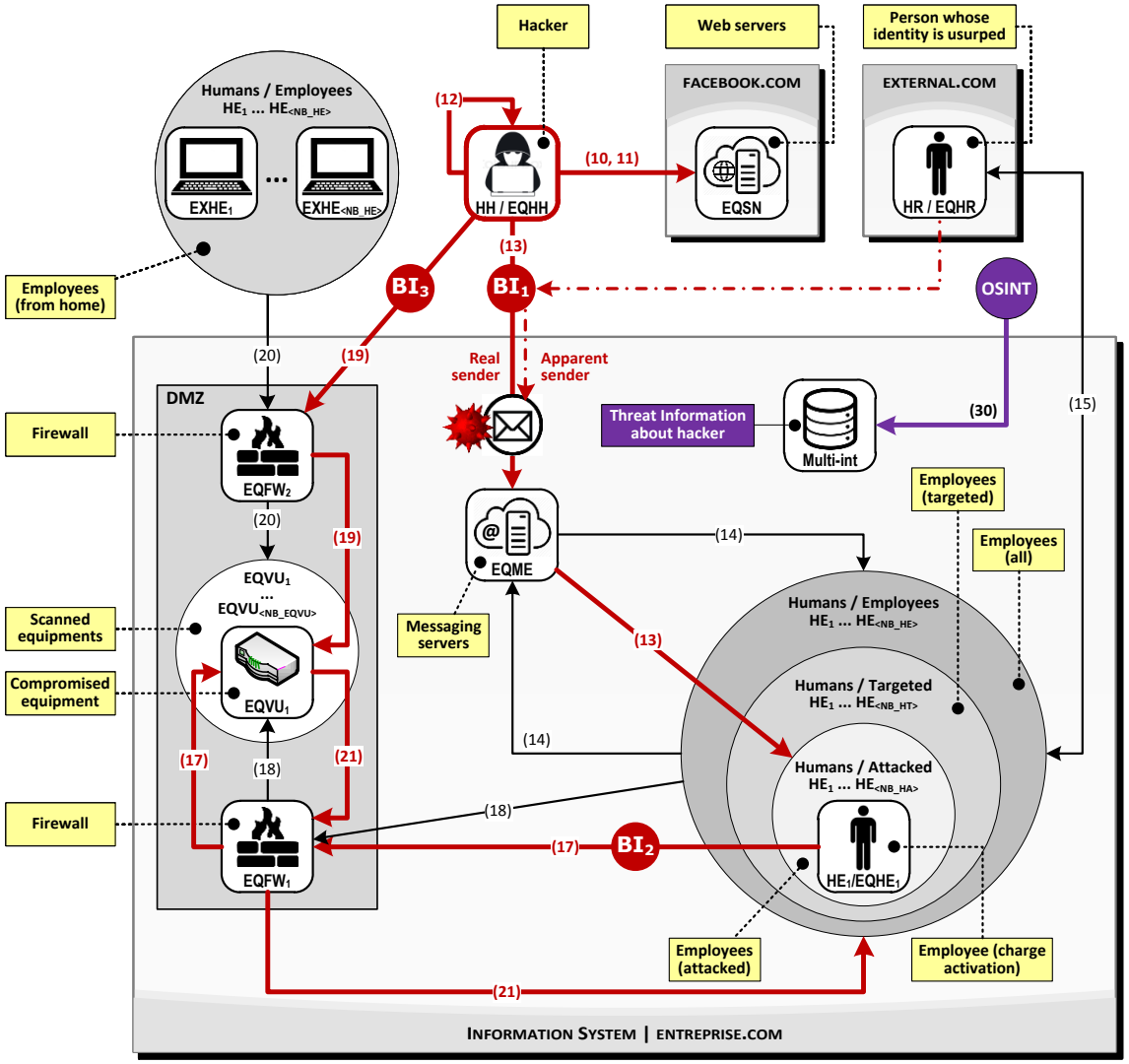
PRINCIPLE



- ❑Three phases AI :
  - Learning (coutinuous).
  - Inference.
  - Correlation.
- ❑AI for memorisation (to be done).

4. Our approach

Compromising documents on a company's information system, by screening / targeting, identity theft, malicious attachment, and exploitation of a vulnerability.



HE: Human-Employee
EXHE: From external / Human-Employee
EQHE: Equipment of / Human-Employee
HH: Human-Hacker
EQHH: Equipment of / Human-Hacker
HR: Human-Referent
EQHR: Equipment of / Human-Referent

EQFW: Equipment / Firewall
EQME: Equipment / Messaging
EQSN: Equipment / Social Network
EQVU: Equipment / Vulnerable

<NB_EQVU>: Number of equipments vulnerable
<NB_HA>: Number of humans attacked
<NB_HE>: Number of humans employed
<NB_HT>: Number of humans targeted

BI Behavioural incongruity

(xx): Rule reference

Usual behaviours (extract)

14 15	Normal sending of internal and external emails.
18	Normal solicitations of equipments / ports.
20	Normal activity between the external and the equipment compromised.

Hostile scenario

10 11	The hacker performs a screening and targeting.
12	The hacker prepares an attack kit.
13 BI_1	The hacker sends an email with malicious attachment to 2 targeted employees by usurping a third-party identity.
16	Targeted employee opens the attachment and activates the charge.
17 BI_2	The charge scans ports on vulnerable equipment and compromises one.
19 BI_3	The hacker connects to the compromised equipment and takes control of it.
21	The hacker exploits the vulnerability to collect sensitive documents.
30	An OSINT* source reports hacker.

(*) OSINT – Open Source Intelligence

A B C

5. POC #1: scenario

A. Scenario theatre.
B. Usual behavior.
C. Hostile behavior.

AIRBUS

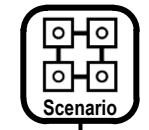
6. POC #1: behind the scene

A. Scenario details.
B. Metrics generation.
C. More about AI.
D. Correlation and graphs.

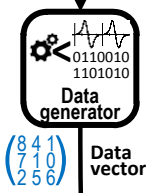
1 month
Normal activity

2 days
Normal + hostile activities

1 month
Normal activity



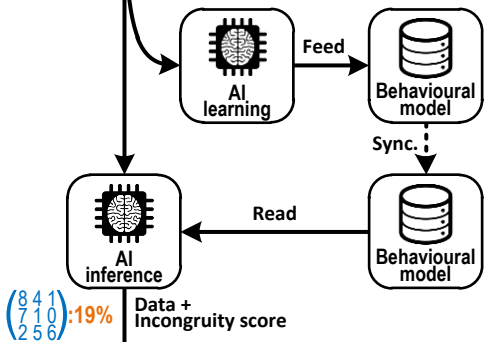
- A company, 100 employees working on site and from their home.
- Theatre: an IS (internal/external PC, messaging, network flows, firewalls, routers).
- internal, external, mixed flows.
- A social network used for screening / targeting.



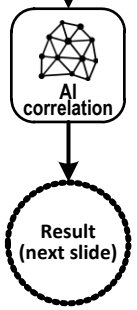
- Our own massive, coherent data generator.
- 500K metrics generated.
- Data enrichment (eg: aggregations / counts on sliding time windows).

Metrics

- Flow (source, destination).
- Email (sender, recipient, attachment).
- Protocols, ports, timestamp.
- OSINT source.



- Input metrics: converted to numbers.
- Algorithm: **isolation forest**, **unsupervised**.
- Output scores: **neither normalised nor filtered**, so that the correlation phase (see below) receives all the information including weak signals.
- **Real time** performance : ~5K metrics / s. on a single PC.

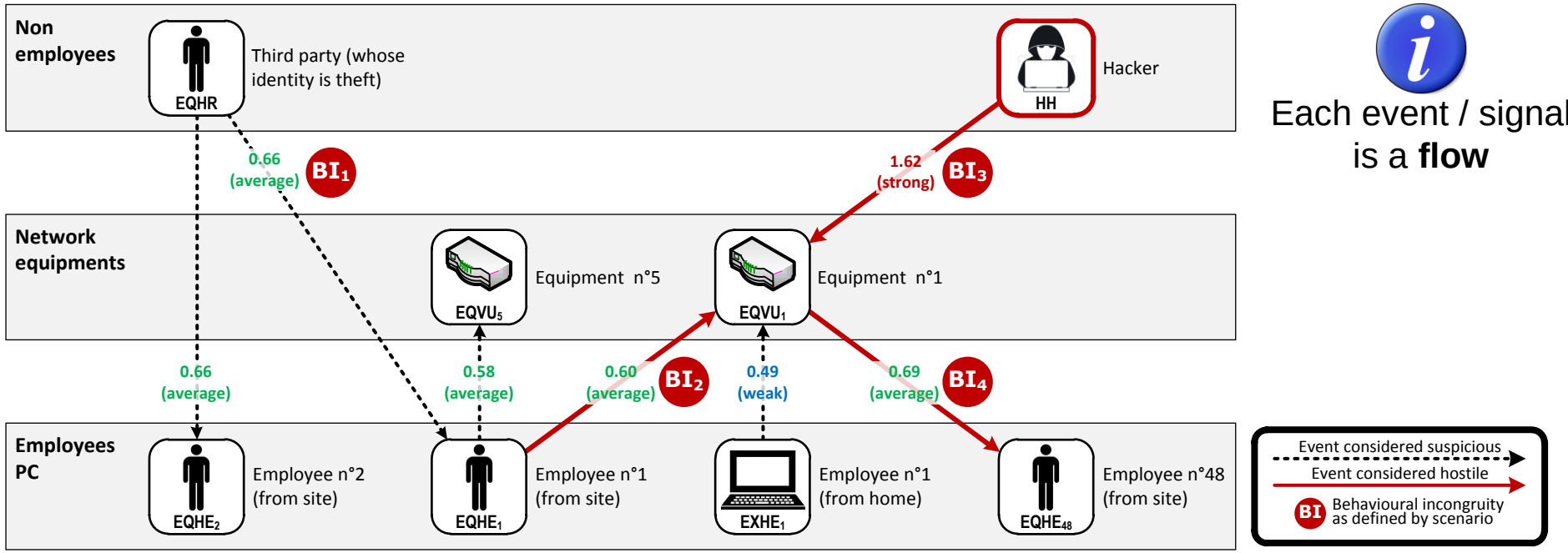


- Discovery of **major interest graphs**, with an algorithm working on 3 spaces:
  1. Metrics concentration (quasi-twins).
  2. Search for related events.
  3. Search for major interest graphs made of strong / weak / normal signals via a relevance function.

Relevance function

Based on temporal feedback, hysteretic effect, forgetfulness, incongruity score, signal type, topological properties, time scales, probabilities.

AIRBUS



Each event / signal is a **flow**

MAIN RESULTS: DETECTION OF HOSTILE BEHAVIOURS HAVING DIRECT IMPACT

	SCENARIO	DETECTED
BI_1	The hacker sends malicious attachment to 2 targeted employees by usurping a third-party identity.	Event is considered only **suspicious** but nevertheless contributes to the globally hostile events chain.
BI_2	The charge scans ports on vulnerable equipment and compromises one.	Event is considered **incongruous** (average score) within hostile events chain.
BI_3	The hacker takes control of compromised equipment.	Event is considered **incongruous** (strong score) within hostile events chain.

❑ Few false positives (during calibration).

UNEXPECTED: DETECTION OF HOSTILE BEHAVIOURS HAVING INDIRECT IMPACT

❑ Detection of suspicious flow: sending of the same malicious attachment to the employee's PC n° 2.

❑ Detection of a fourth behavioural incongruity BI_4 : the hacker downloads sensitive documents located on PC n° 48.

➡ **Detection is <u>complete</u> with <u>good explainability</u>.**

A B

7. POC #1: results

A. Achieved expected results.
B. Unexpected results.

AIRBUS

Main biases of available solutions	Our results for POC #1	Focus (for POC #2)
Training performance	❑ Learning: partially scalable. ❑ Inference + correlation: horizontal scaling.	
Many false positives	❑ Few false positives, only during first month (calibration). ❑ No false negatives.	
Over-simplification of problems	❑ Training on the entire dataset. ❑ Multivariate events of different types.	
Slightly explainable result	❑ Detection is complete with good explainability.	
Frequent presence of a simple time window alerts counter	❑ We don't use counters but graphs on sliding and variable time windows over wide temporal ranges.	
Little consideration of events temporality	❑ Our algorithm uses events temporality, it adapts to any time scale, from microseconds to years.	
Low management of behavioural model, boiled frog paradox	❑ To be done, we will use AI for synchronisation of the behavioural model.	
Other limitations	❑ Synthetic data. ❑ Simplistic scenario. ❑ Too little data.	

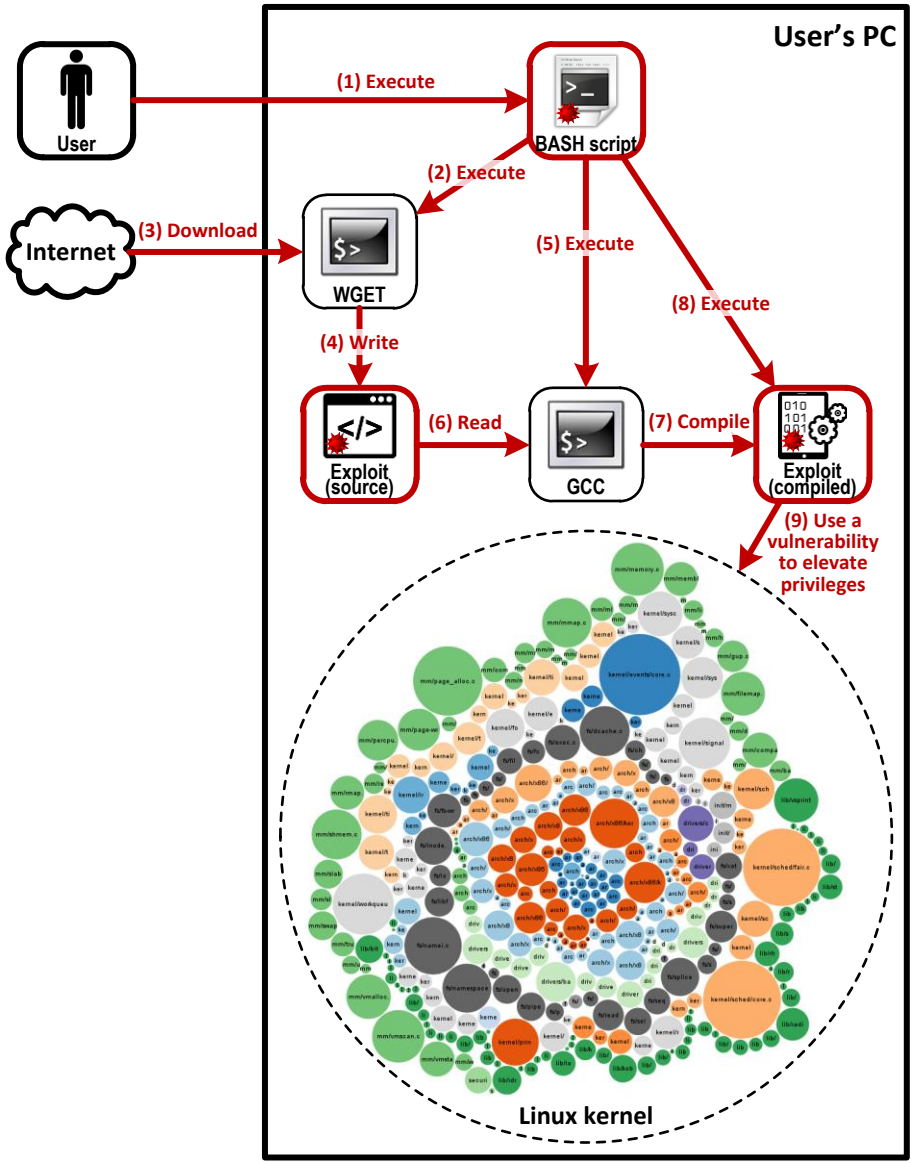
A

8. POC #1: conclusion

A. Biases versus progress.

AIRBUS

On his Linux PC, a user unwisely executes a malicious script which downloads an exploit from the Web in order to use a kernel vulnerability to elevate its privileges.



Usual behaviours
The user performs office tasks (eg: word processing, messaging, Internet browsing).
The user executes commands and scripts.

Hostile scenario	
1	The user executes a malicious script, via a BASH* command.
2, 3, 4	The malicious script downloads source code of an exploit from the web, via a WGET* command.
5, 6, 7	The malicious script compiles the exploit, via a GCC* command.
8, 9	The malicious script executes the compiled exploit, which tries to elevate its privileges using a vulnerability of the operating system kernel.

(*)
- BASH : standard command for executing scripts.
- WGET : standard command for downloading files from the Web.
- GCC : standard command for compiling programming languages.

A

9. POC #2: scenario

A. Scenario overview.

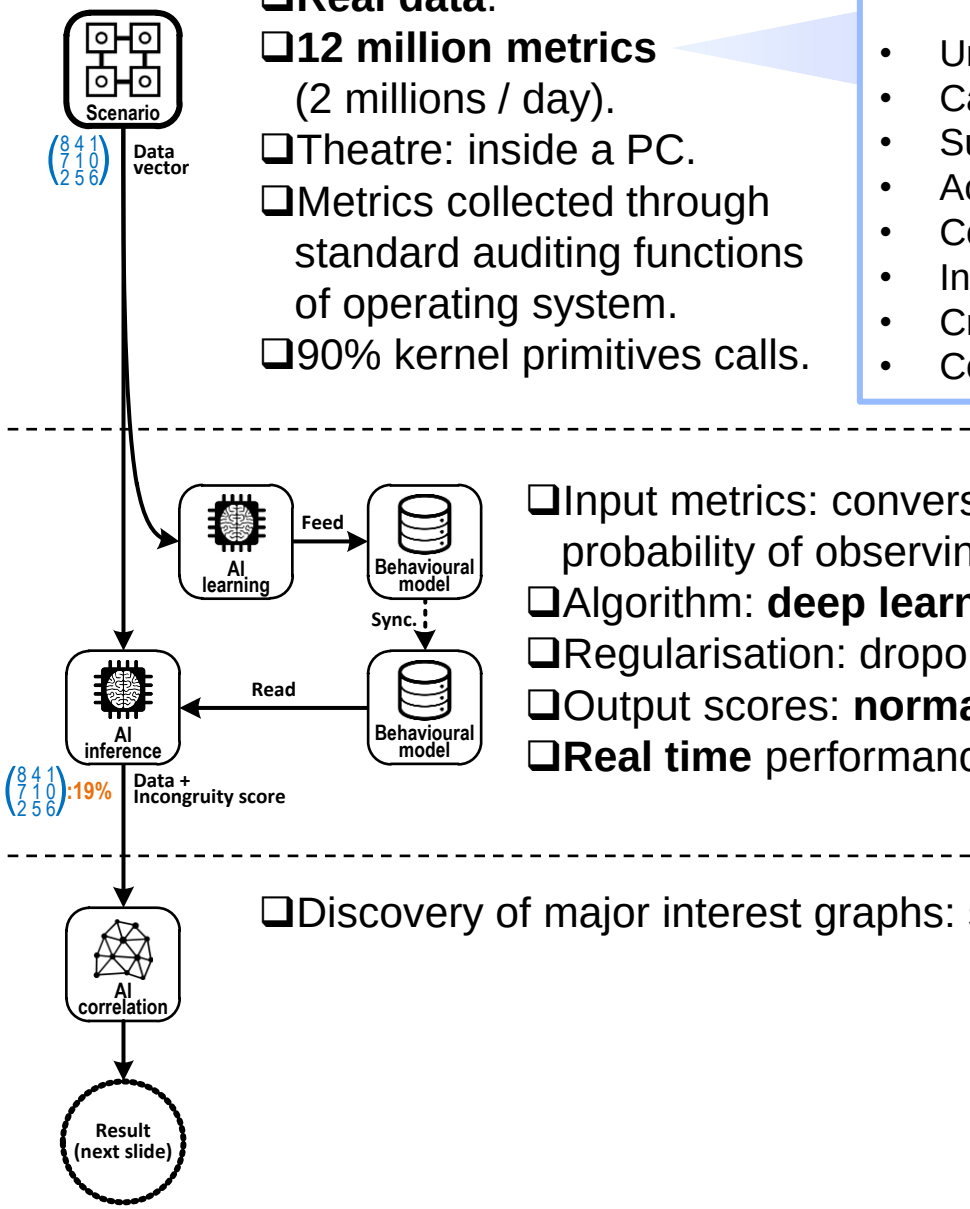
AIRBUS

10. POC #2: behind the scene

A. Metrics.
B. More about AI.
C. Correlation and graphs.

1 week
Normal activity

1 minute
Normal + hostile activity



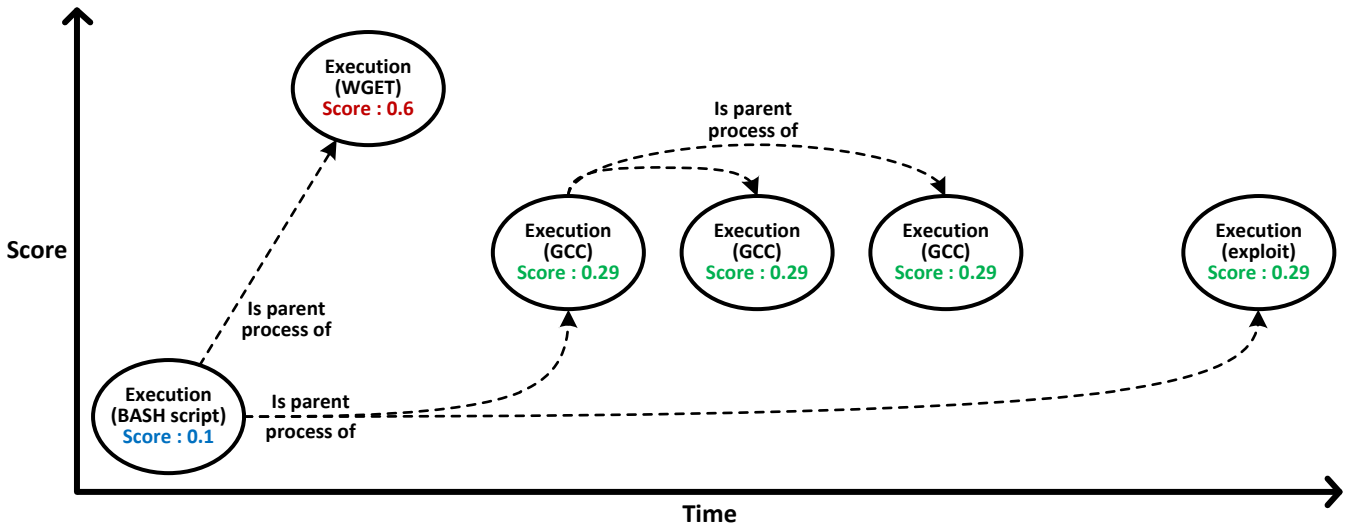
- ❑**Real data**.
- ❑**12 million metrics** (2 millions / day).
- ❑Theatre: inside a PC.
- ❑Metrics collected through standard auditing functions of operating system.
- ❑90% kernel primitives calls.

Metrics

- Unauthorised actions.
- Calls to functions/commands for modifying kernel/modules.
- Suspicious actions (eg: nmap, wget, tcpdump).
- Access to monitored files (eg: config., binaries, temp. files).
- Commands executed.
- Invocations of potentially dangerous kernel primitives.
- Credentials (eg: user, group).
- Context (eg: path, timestamp, parent process).

- ❑Input metrics: conversion of categorical variables to numerical using probability of observing couples of values after observing others.
- ❑Algorithm: **deep learning autoencoder**, **unsupervised**.
- ❑Regularisation: dropout, noise addition, early stopping.
- ❑Output scores: **normalised, not filtered**.
- ❑**Real time** performance : ~2K metrics / s. on a single PC with GPU.

- ❑Discovery of major interest graphs: same as for POC #1.



Each event / signal is an **action**

MAIN RESULTS

- **Detection is <u>complete</u> with <u>good explainability</u>** :
  - Execution of the BASH script (score 0.1).
  - Execution of the WGET command (score 0.6).
  - Three executions of the GCC command (score 0.29).
  - Execution of the exploit (score 0.29).
- The BASH process has a low incongruity score, but it still contributes to the major interest graph because it connects other actions.
- Some false positives resulting from rare actions, which could be avoided by optimising training.
- No false negatives.

Main biases of available solutions	Our results for POC #2
Training performance	❑ Learning + inference + correlation : horizontal scaling (cloud friendly).
Many false positives	❑ Few false positives, but could be avoided. ❑ No false negatives.
Over-simplification of problems	❑ Training on the entire dataset, directly from raw logs. ❑ Multivariate events of different types.
Slightly explainable result	❑ Detection is complete with good explainability.
Frequent presence of a simple time window alerts counter	❑ We don’t use counters but graphs on sliding and variable time windows over wide temporal ranges.
Little consideration of events temporality	❑ Our algorithm uses events temporality, it adapts to any time scale, from microseconds to years.
Low management of behavioural model, boiled frog paradox	❑ To be done, we will use AI for synchronisation of the behavioural model.
Other limitations	❑ Simplistic scenario.

A

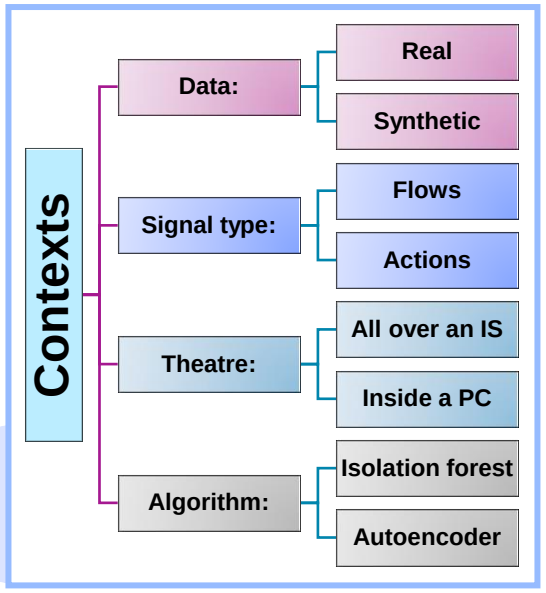
12. POC #2: conclusion

A. Biases versus progress.

AIRBUS

SITUATION

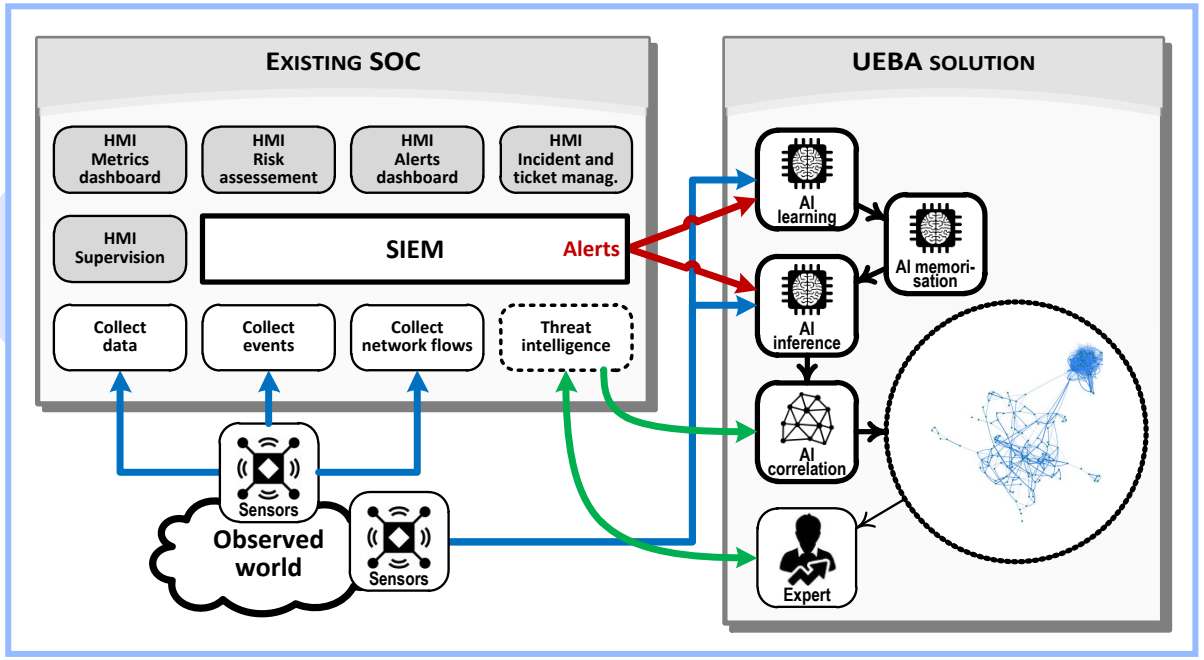
- ❑ Effective association of UEBA with correlation process.
- ❑ Good explainability of alerts.
- ❑ Few but avoidable false positives.
- ❑ Temporality taken into account from microseconds to years.
- ❑ Real time 3 phases algorithm + horizontal scaling.
- ❑ Integration issues partially addressed (ELK).
- ➧ **Encouraging results.**
- ➧ **Results confirmed in various contexts.**



FUTURE

- ❑ More realistic scenarios.
- ❑ Adversarial AI*.
- ❑ Memorisation AI*.
- ❑ Interoperation with SIEM.

(*) PhD thesis 2019 : « *Continuous Model Learning for Anomaly Detection In the Presence of Highly Adaptative Cyberattacks* ».



A B

13. Situation and future

A. Progress and limits.

B. Remaining work.

AIRBUS

Questions (and answers !)