

# AI and Security

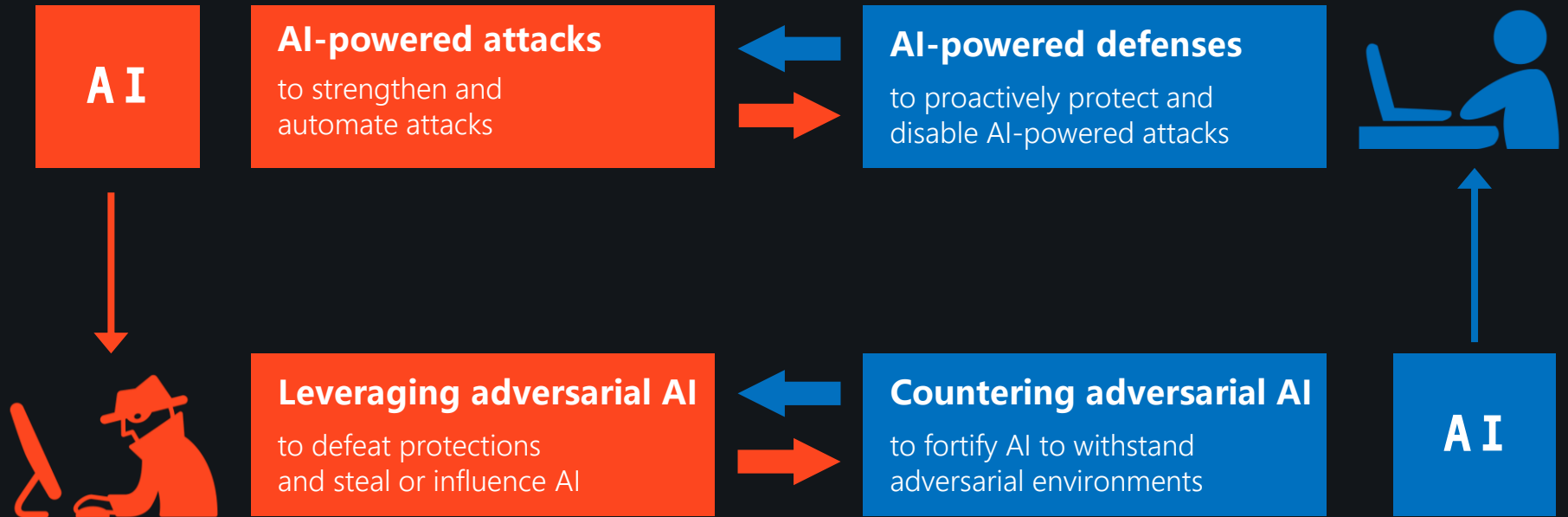
Promises and realities in automated threat management and hunting

**Dr. Marc Ph. Stoecklin**

Principal Research Scientist

Department Head, Security Research

# AI and Security: A Love and Hate Story



# Artificial Intelligence

Reasoning

Natural Language  
Processing

Planning

Machine Learning

Supervised  
learning

Unsupervised  
learning

Reinforcement  
learning

Deep  
learning

# Posture Dimension

Identify  
Protect



# Posture Dimension

Identify  
Protect

# Threat Dimension

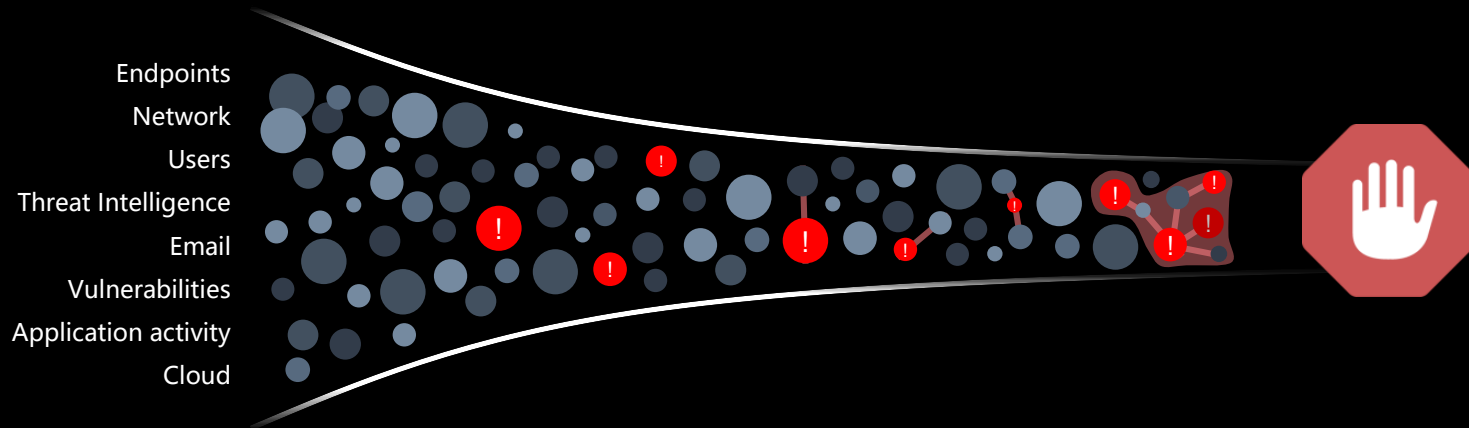
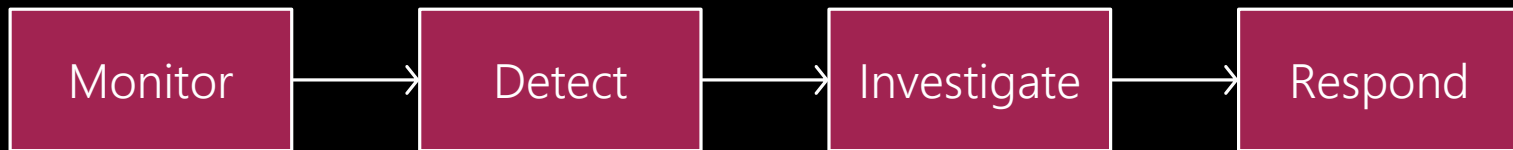
Detection  
Response



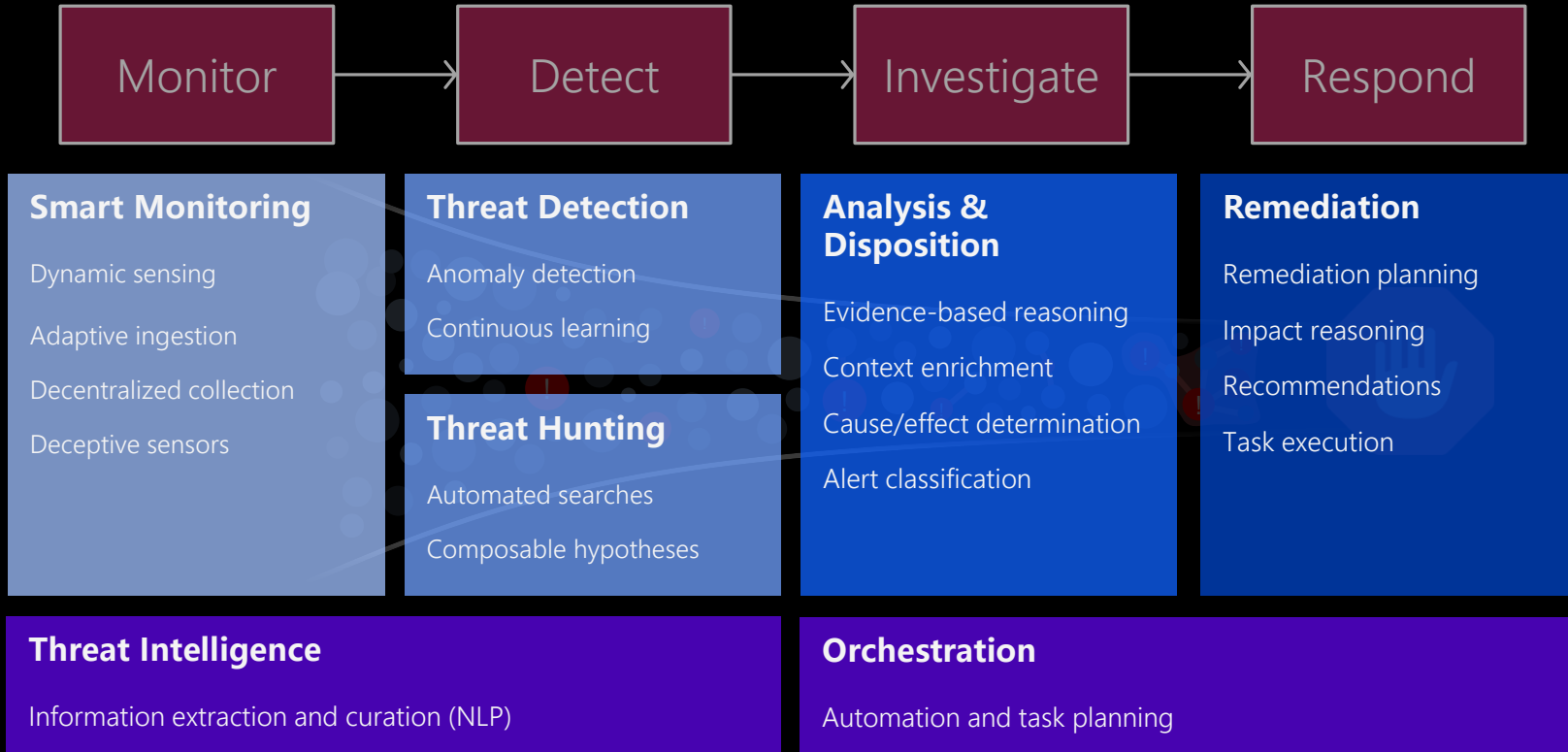
Recovery



# Threat Management Lifecycle



# Use of AI in the Threat Management Lifecycle



# Pain points of Security Operations teams



Flood of alerts



Shortage of qualified staff



Too many tools,  
too many technologies

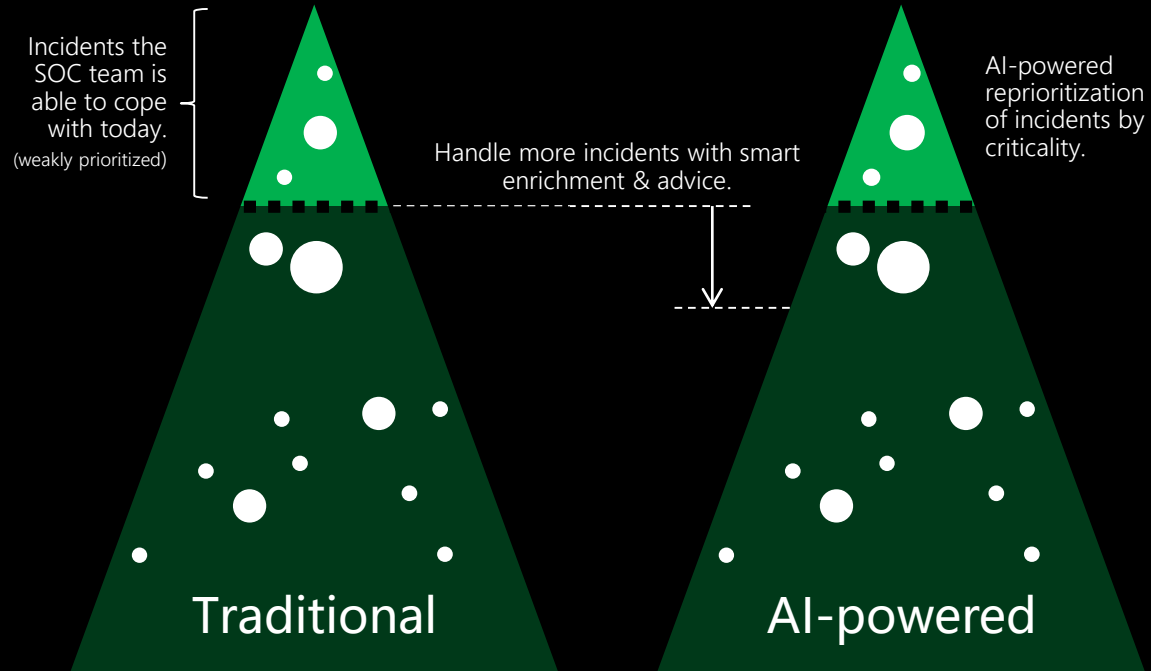


Increasing threat landscape





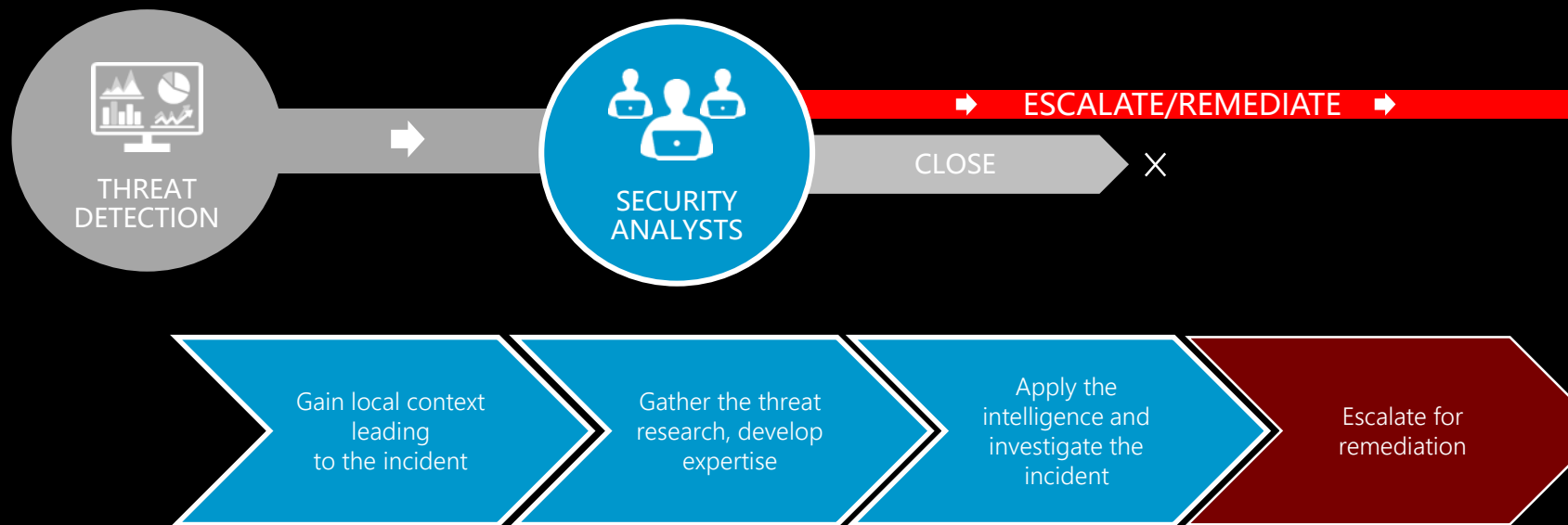
# Addressing the Investigation Bottleneck



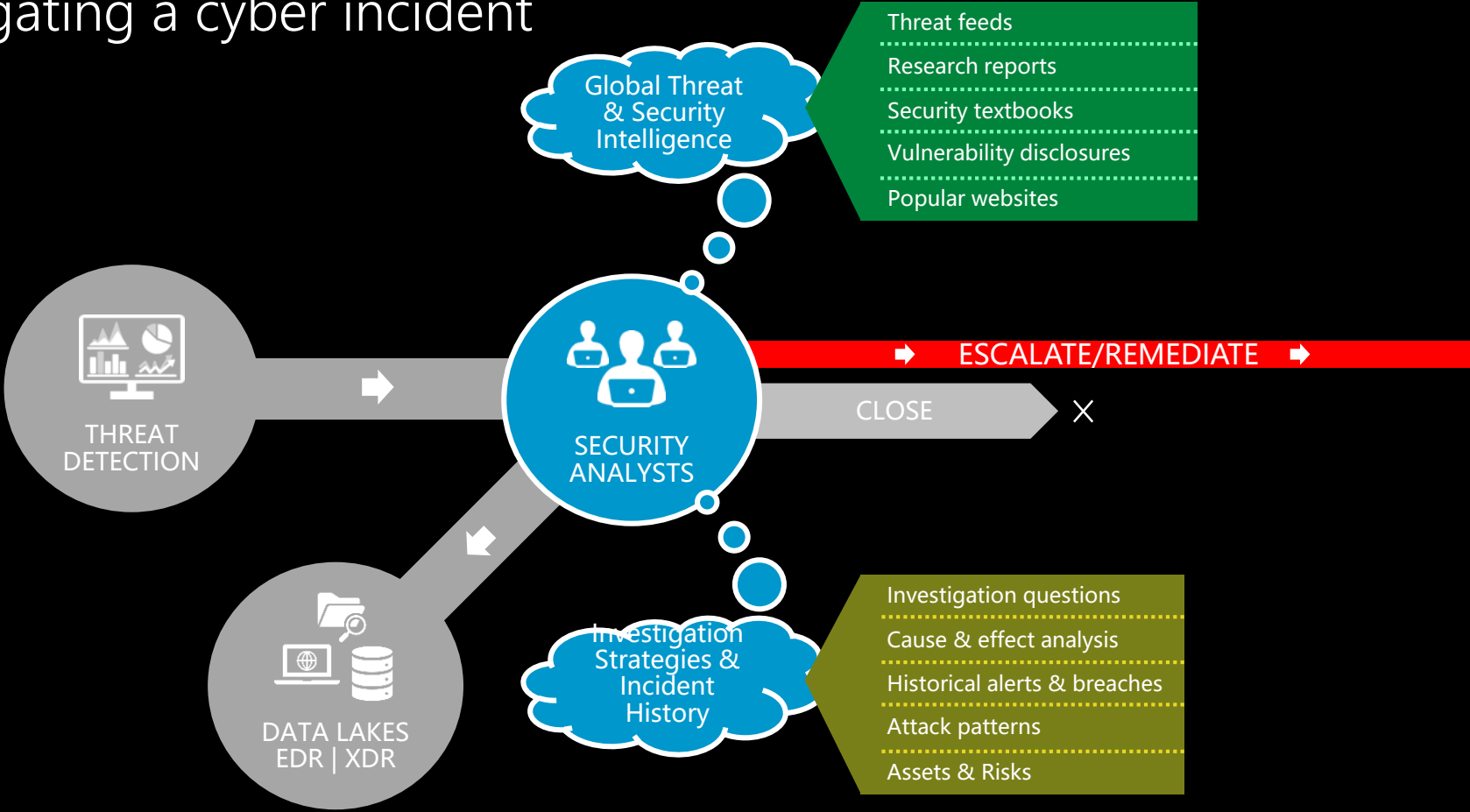
Incident criticality:



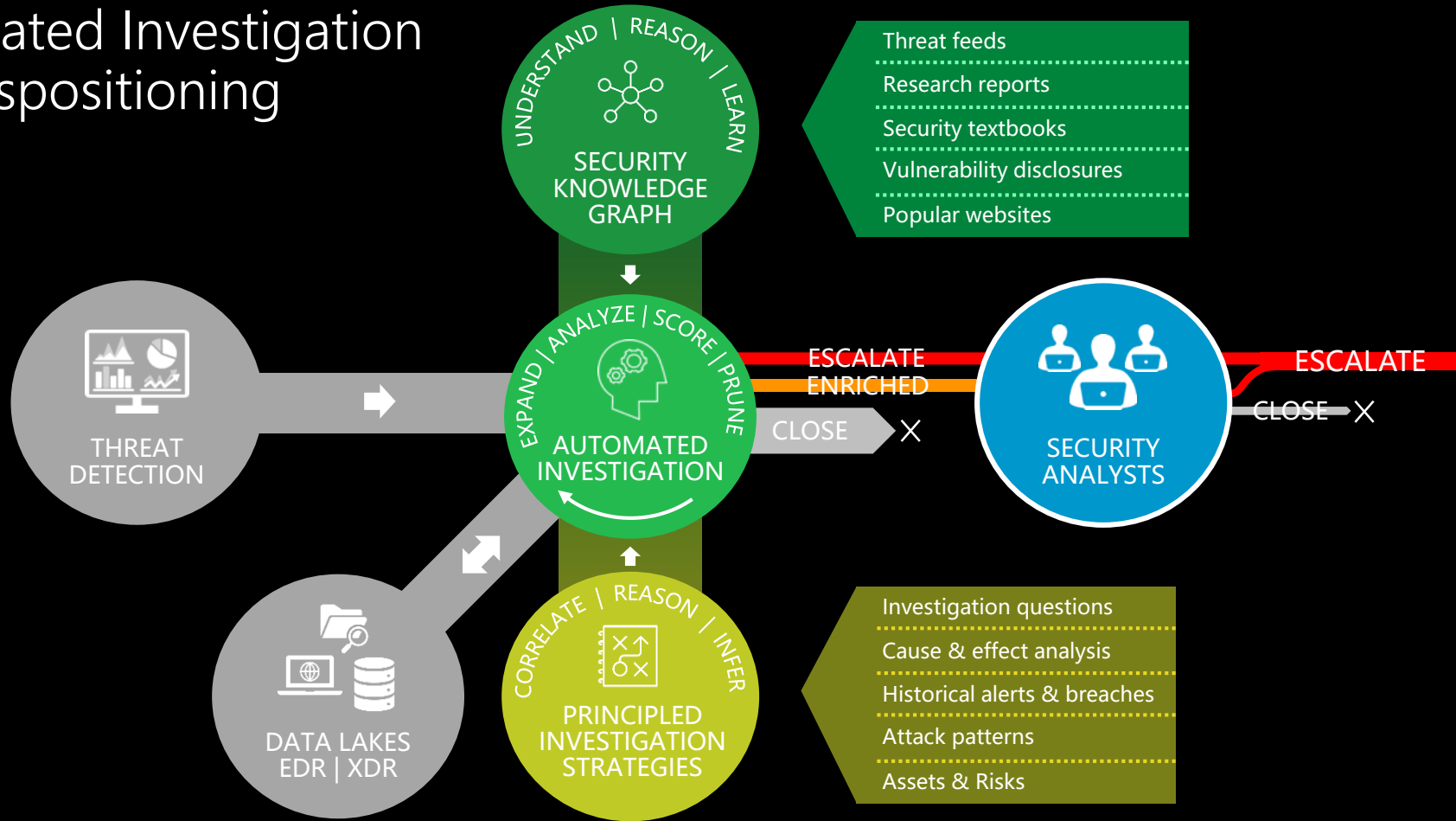
# Investigating a cyber incident



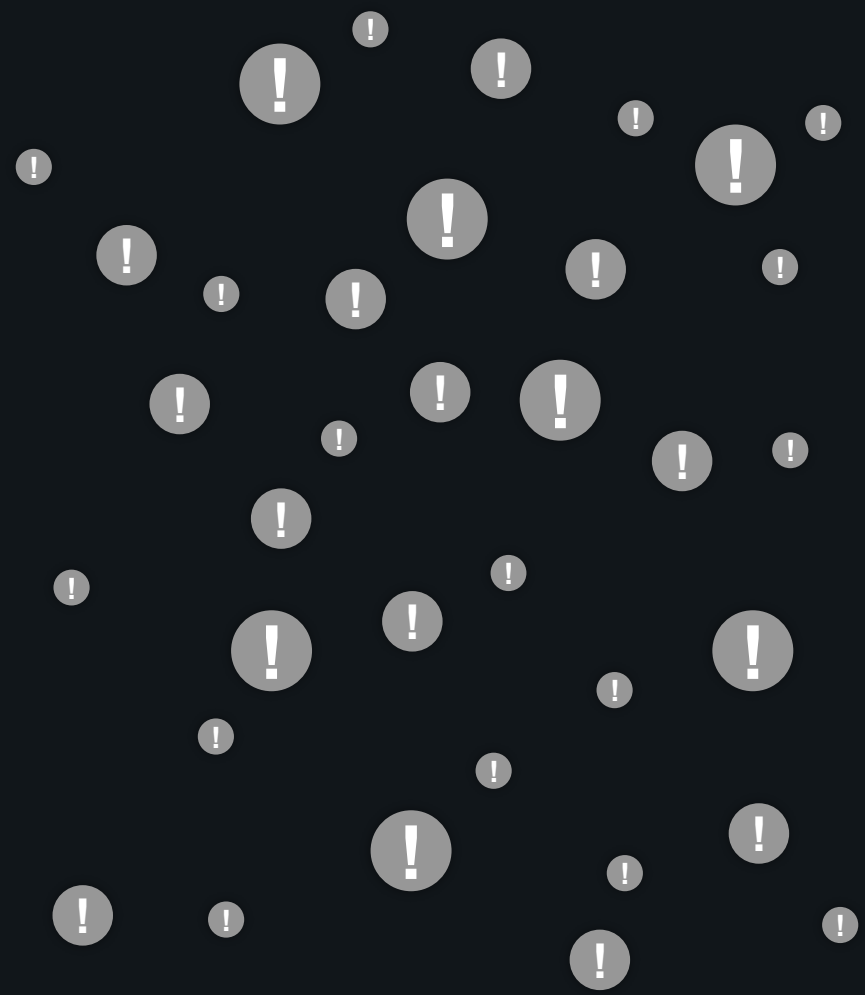
# Investigating a cyber incident



# Automated Investigation and Dispositioning



# Automated investigation

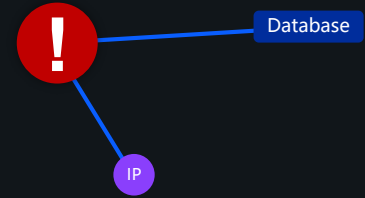


# Automated investigation



# Automated investigation

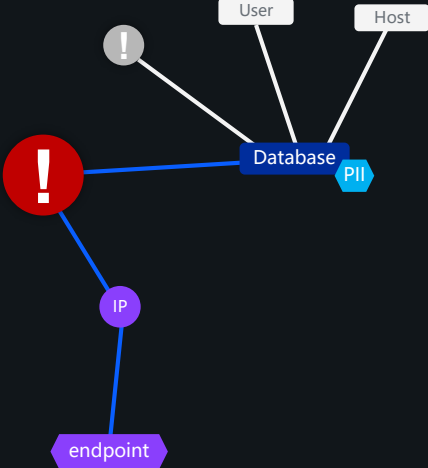
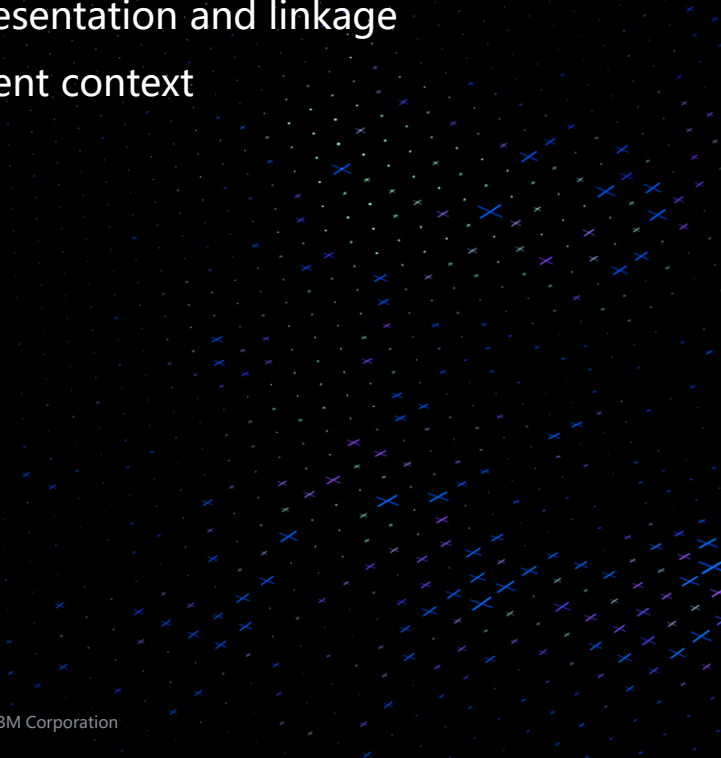
Data representation and linkage



# Automated investigation

Data representation and linkage

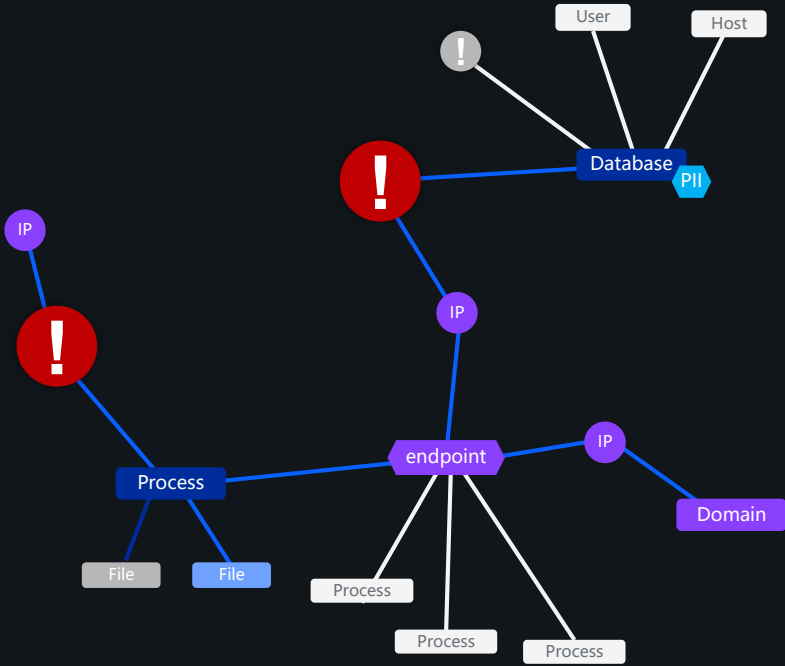
Environment context





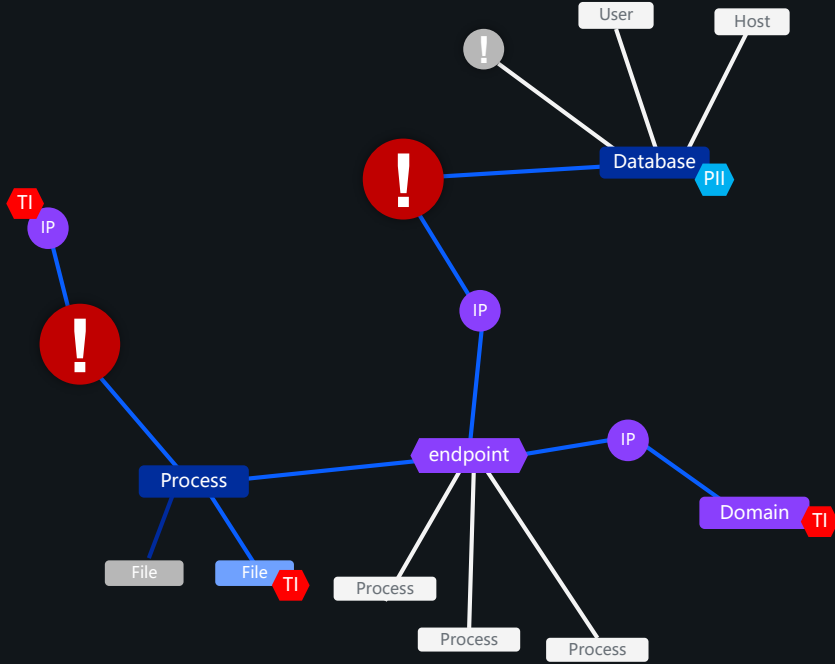
# Automated investigation

Data representation and linkage  
Environment context  
Forensic reasoning



# Automated investigation

- Data representation and linkage
- Environment context
- Forensic reasoning
- Threat intelligence





# Automated investigation

Data rep  
Environ  
Forensic  
Threat in  
Domain

### QUERY SETTINGS

Depth: 8  
Threshold: 0  
Steps: 100 | TType: T  
Model: aro\_fb\_test\_2  
Explore  
Flow

### DISPLAY SETTINGS

Layer: 0  
Timeline min: 0.0  
Timeline max: 1.0  
 Min  Physics  Cluster  Alpha  
 Hierarchical  Timeline-Layout  Stabilize  DataRefresh  
 Hypotheses  Show S&T  Show S&T as Wedges  
Repulsion: [slider]  
Reload | Fit | Search | Stop

Item Details:  
Process: sh  
Signal: 0.51, toxicity: 0.68

object (16)  
▶ timestamp [868]  
id: process\_sh\_112a72018657791b2b17755cc417a18  
lastUpdate: 1633539106.73261  
raw\_node: 9276-sh, score: 0  
Ref\_Num: 0  
raw\_type: process  
raw\_label: sh  
type: Process  
label: sh  
raw\_id: 112a72018657791b2b17755cc417a18  
▶ \_kg (7)  
depth: 3  
value: sh  
raw\_value: sh  
signal: 0.50871380611023  
toxicity: 0.6818456916052714

Output Graph (Nodes: 3056, Edges: 10763):

```
{
  "severity_score": 0.734572110014616
}
```

Hypothesis	Severity
Sysmon_e28a5a99-da44-436d-b7a0-2af620a5f413_0 Whoami Execution	0.57
Sysmon_80167ada-7a12-41ed-b8e9-aa47195c66a1_0 Run Whoami as SYSTEM	0.57
Node process starts shell	0.30
Large network data transfer with database endpoint	0.30
Account Discovery - Local Account	0.30
Reverse Unix shell started	0.30
System Owner/User Discovery	0.27
Process Discovery	0.27
File and Directory Discovery	0.27
System Network Connections Discovery	0.27
Remote System Discovery	0.27
Linux and Mac File and Directory Permissions Modification	0.27
Possible SSTI attack	0.16
SQL Injection attempt	0.12
MySQL: failed login attempt	0.08

User

Host

base

PII

Ext

Domain

TI

# Automated investigation

**QUERY SETTINGS**

Depth:

Threshold:

Steps: 100 TType: S

Model: **arc\_fb\_test\_2**

Explore  
Flow

**DISPLAY SETTINGS**

Layer:

Timeline min:

Timeline max:

Min  Physics  Cluster  Alpha

Hierarchical  Timeline-Layout  Stabilize  DataRefresh

Hypotheses  Show S&T  Show S&T as Wedges

Reputation:

Reload | Fit | Search: /shop

Item Details:

Process: sh  
Signal: 0.0, toxicity: 0.68

object (16)

- timestamp (868)
- id: sprocsh\_sh\_112a72d18657791b2b37ff55cc417a18
- lastUpdate: 163303916873561
- raw\_node: 9276-sh, score: 0
- iter\_num: 0
- raw\_type: process
- raw\_label: sh
- type: Process
- label: sh
- raw\_id: 112a72d18657791b2b37ff55cc417a18
- \_kg (7)
- depth: 3
- value: sh
- raw\_value: sh
- signal: 0.00731806110133
- toxicity: 0.6813456916052714

Output Graph (Nodes: 3056, Edges: 10763):

```

{
  "severity_score": 0.7345721110014616
}

```

**Hypothesis**

Hypothesis	Severity
Sysmon_a28a5a99-d944-436d-b7a0-2afc20a5f413_0 Whoami Execution	0.57
Sysmon_80167ada-7a12-41ed-b8e9-aa47195c66a1_0 Run Whoami as SYSTEM	0.57
Node process starts shell	0.30
Large network data transfer with database endpoint	0.30
Account Discovery - Local Account	0.30
Reverse Unix shell started	0.30
System Owner/User Discovery	0.27
Process Discovery	0.27
File and Directory Discovery	0.27
System Network Connections Discovery	0.27
Remote System Discovery	0.27
Linux and Mac File and Directory Permissions Modification	0.27
Possible SSTI attack	0.16
SQL injection attempt	0.12
MySQL: failed login attempt	0.08

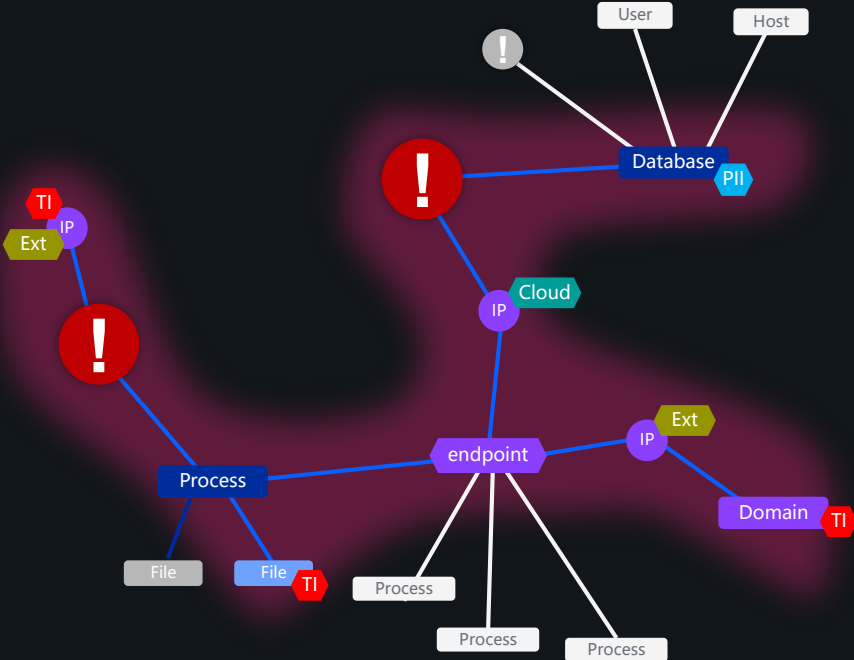
Data rep  
Environm  
Forensic  
Threat in  
Domain



c reasoning  
ation

# Automated investigation

- Data representation and linkage
- Environment context
- Forensic reasoning
- Threat intelligence
- Domain knowledge
- Data reduction



Query Optimization    Semantic reasoning

Graph Theory

Hypothesis generation



# Automated investigation

Data representation and linkage

Environment context

Forensic reasoning

Threat intelligence

Domain knowledge

Data reduction

Attack kill chain reasoning

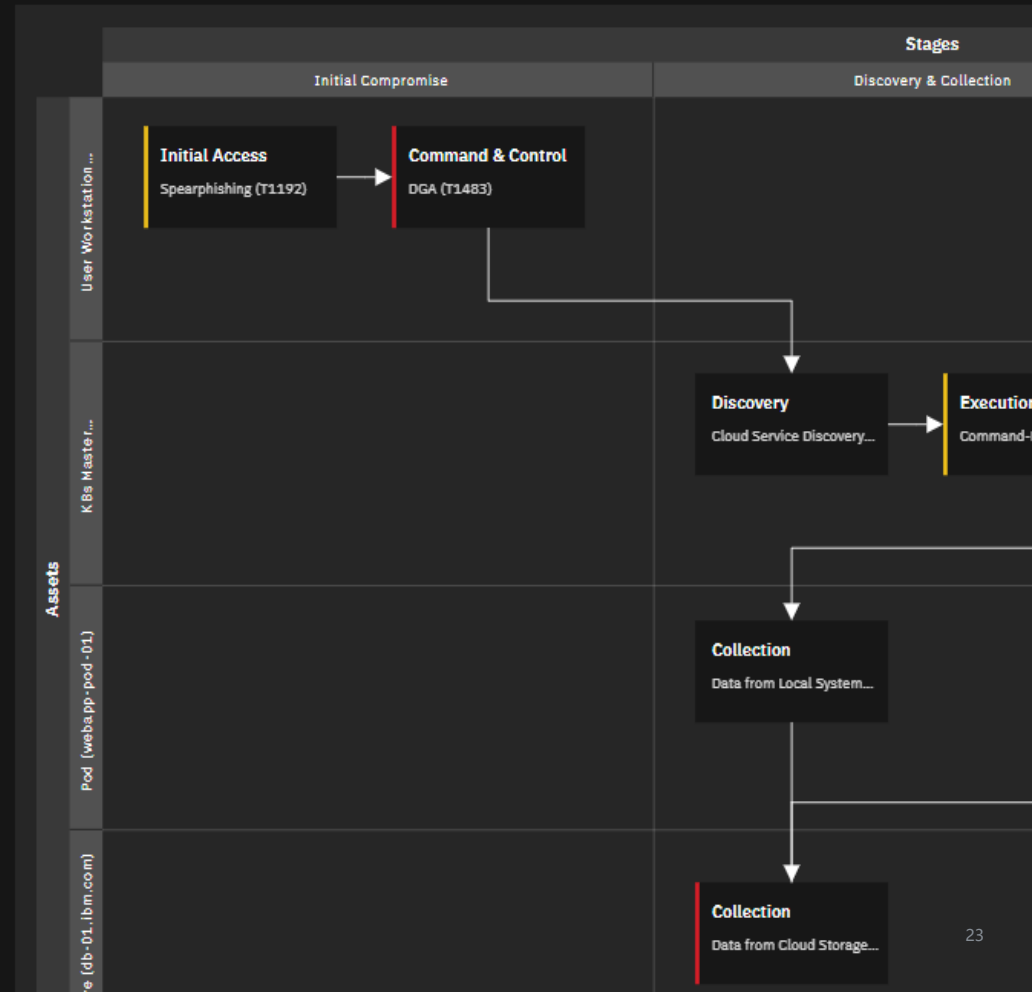
Root cause analysis

Extent & effect determination

General

Attack Graph

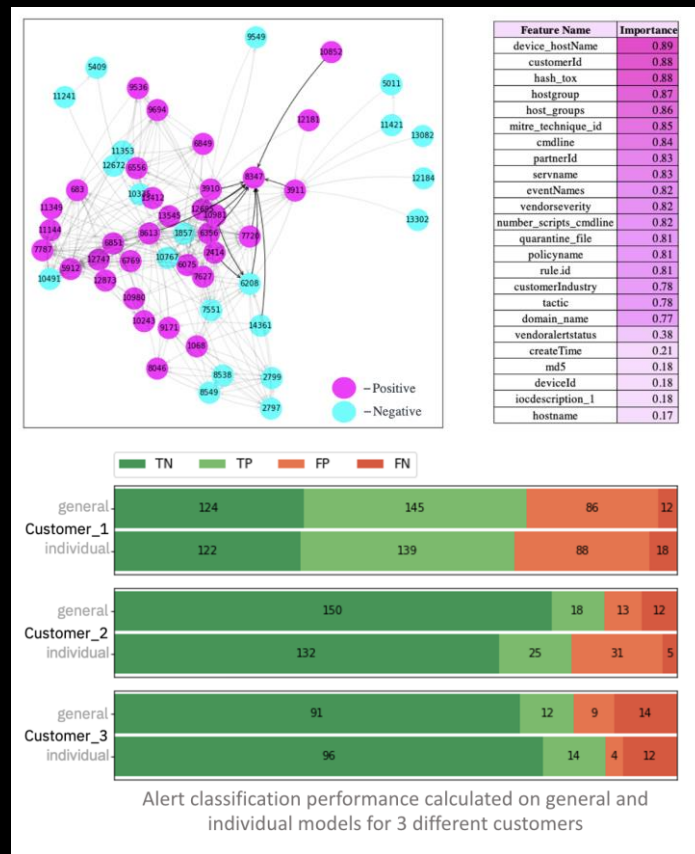
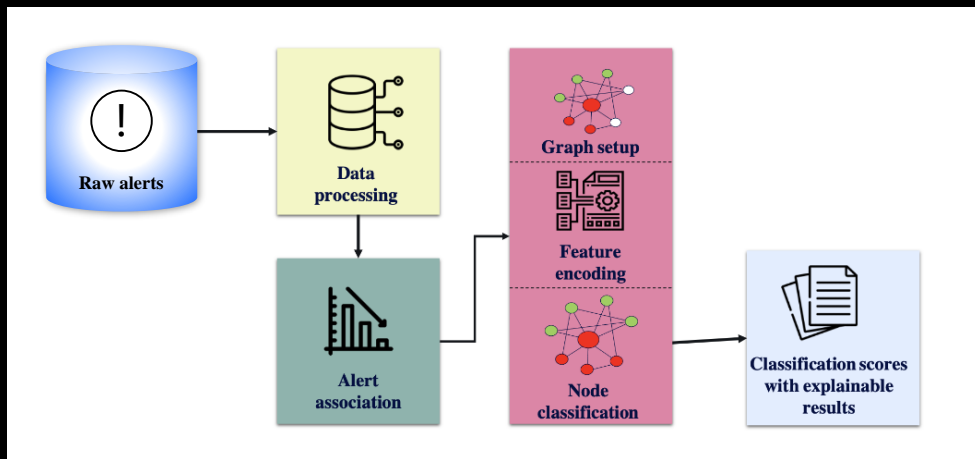
MITRE Summary



# Alert Similarity Analysis and Classification

Train graph neural network models to classify security alerts like security analysts:

- **Close ticket** (as a false alarm)
- **Escalate ticket** (likely true positive)



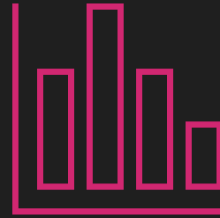


# AI and Security: Considerations and Reflections



Data

Quality  
Quantity  
Granularity



Applicability

Precision  
Speed  
Maintenance



Trustworthiness

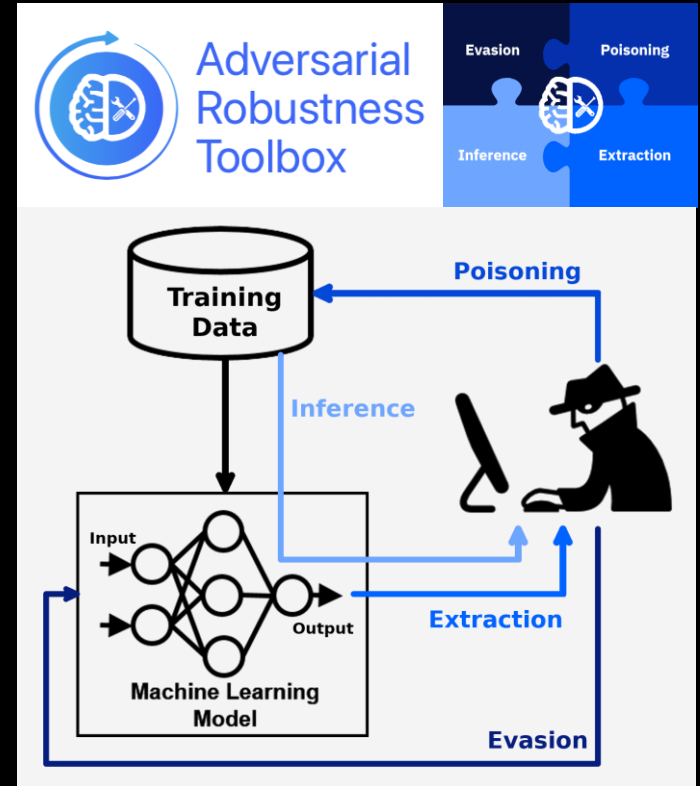
Robustness  
Explainability  
Privacy

# Adversarial Robustness Toolkit (ART)

Tools to defend and evaluate machine learning models and applications against the adversarial threats of Evasion, Poisoning, Extraction, and Inference.

Support for

- all popular machine learning frameworks: TensorFlow, Keras, PyTorch, MXNet, scikit-learn, XGBoost, LightGBM, CatBoost, GPy, etc.
- all data types: images, tables, audio, video, etc.)
- all machine learning tasks: classification, object detection, speech recognition, generation, certification, etc.



# Summary and Next Steps

AI technologies help security teams identifying and countering attacks

Automation and recommendations to expedite for investigation

Data is critical, models need protection

Many more AI applications in security to follow

...

13/09/2020 10:33:02 am

## Domain Squatting detected

A query was made to a domain suspected to be disguising to a legitimate domain in the victim, this technique is usually used for phishing

Initial Access

Execution

13/09/2020 10:37:17 am

## A domain resolution was attempted to a suspicious domain

This host tried to resolve a suspicious domain name. This is an indication from a third source which might indicate infection of malware and communication with a command and control server.

Command And Control

Exfiltration

Discovery

13/09/2020 10:37:17 am

## Evidence of Domain Generation Algorithms usage

A query was made to a domain suspected to be generated by an algorithm. These domains are used for command and control communication by malware.

Command And Control

13/09/2020 10:37:17 am

## Communication with a suspicious IpAddress

Network traffic has been made between this host and a known suspicious IpAddress.

Command And Control

Exfiltration

Discovery

13/09/2020 10:46:02 am

IBM **Research** Security

# THANK YOU

Marc Ph. Stoecklin

[mtc@zurich.ibm.com](mailto:mtc@zurich.ibm.com)

