IBM Research Security

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

AI

AI-powered attacks
to strengthen and automate attacks

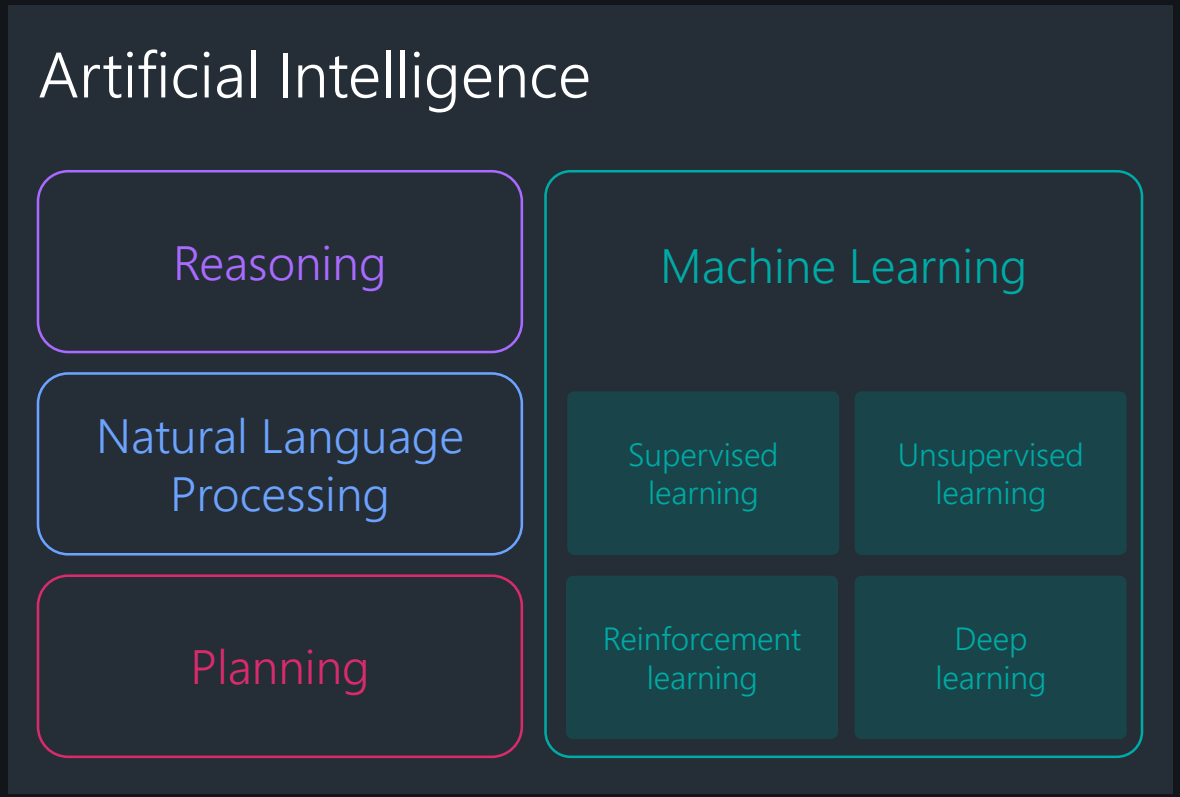
AI-powered defenses
to proactively protect and disable AI-powered attacks

Leveraging adversarial AI
to defeat protections and steal or influence AI

Countering adversarial AI
to fortify AI to withstand adversarial environments

AI

Artificial Intelligence



Posture Dimension

Identify

Protect

Adopted from NIST Cybersecurity Framework (NIST CF)

Posture Dimension

Identify

Protect

Threat Dimension

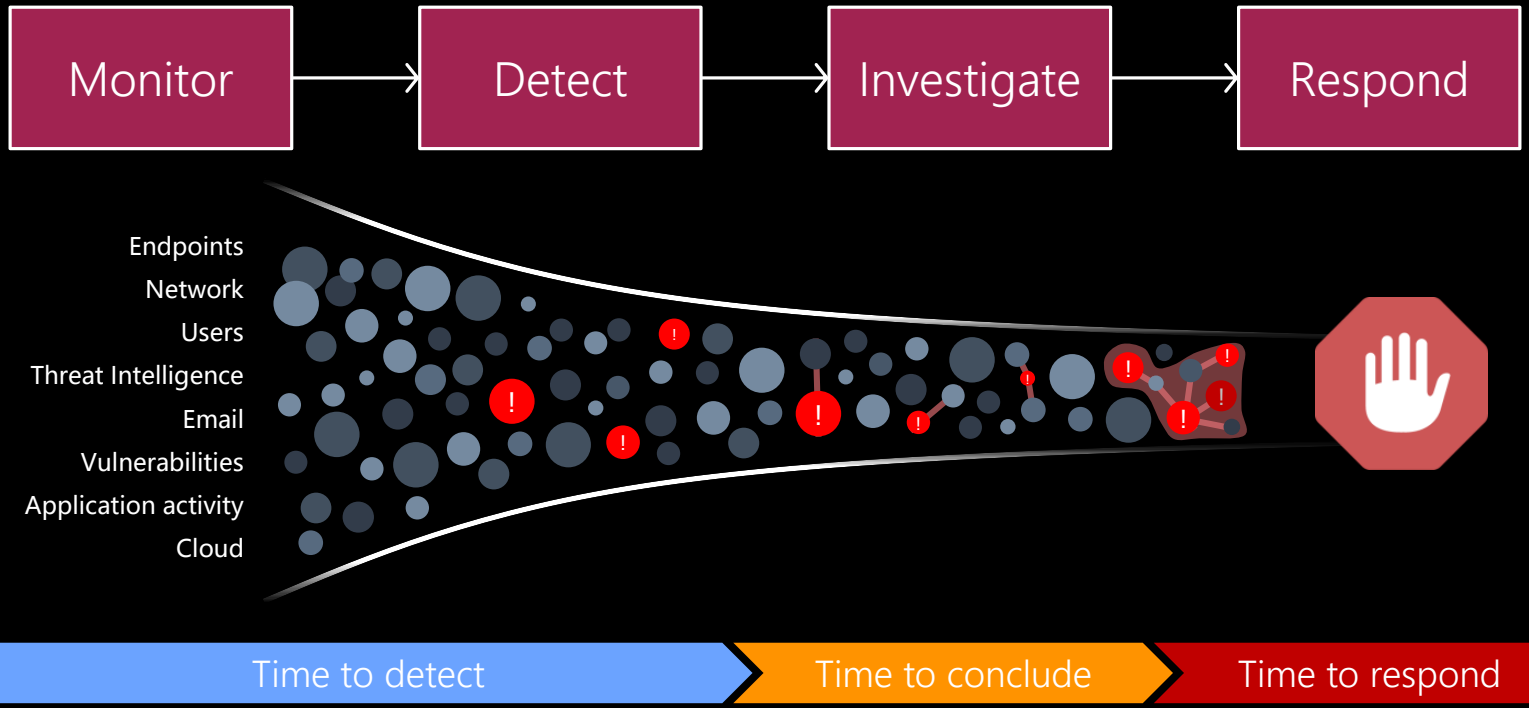
Detection

Response

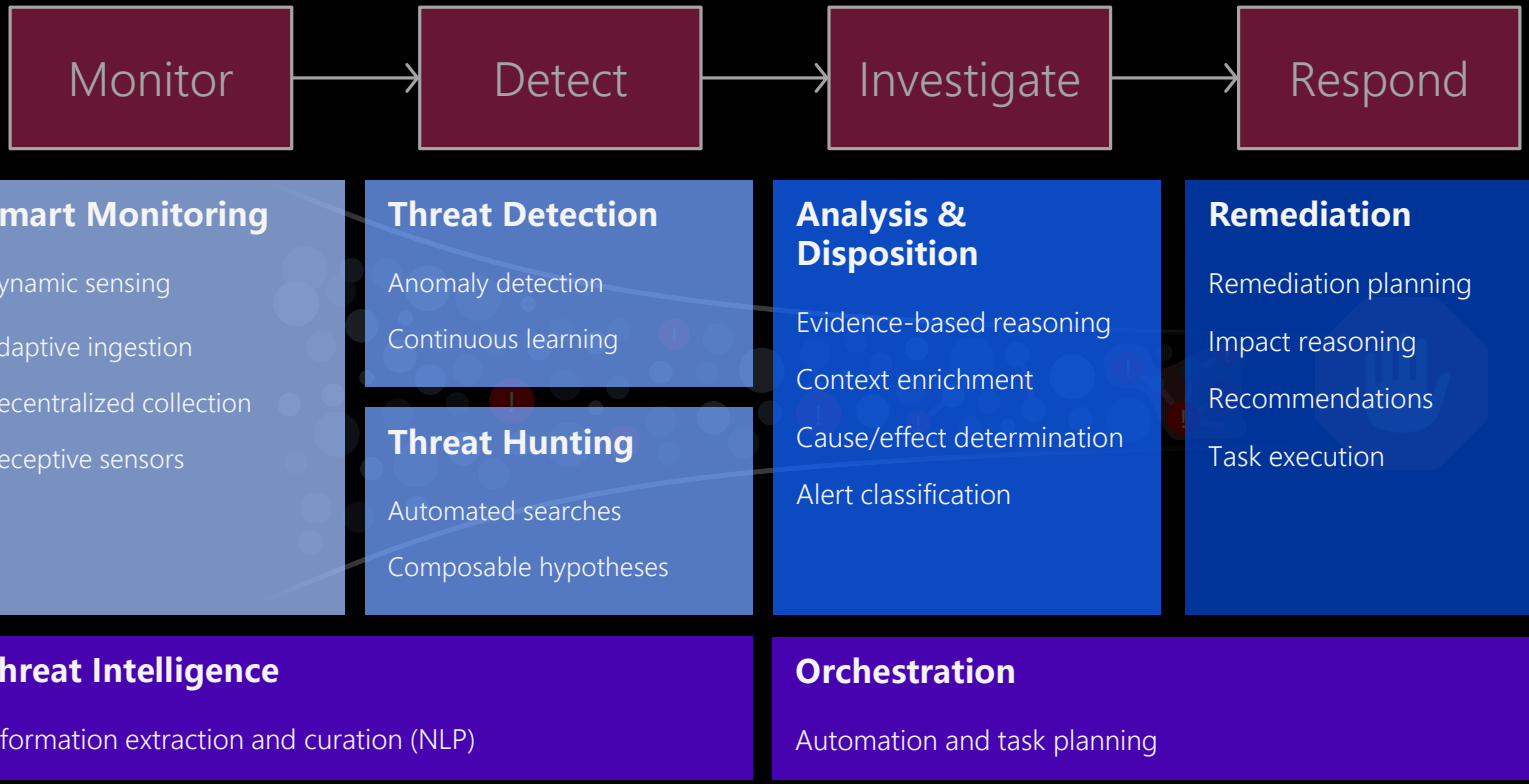
Recovery

Adopted from NIST Cybersecurity Framework (NIST CF)

Threat Management Lifecycle



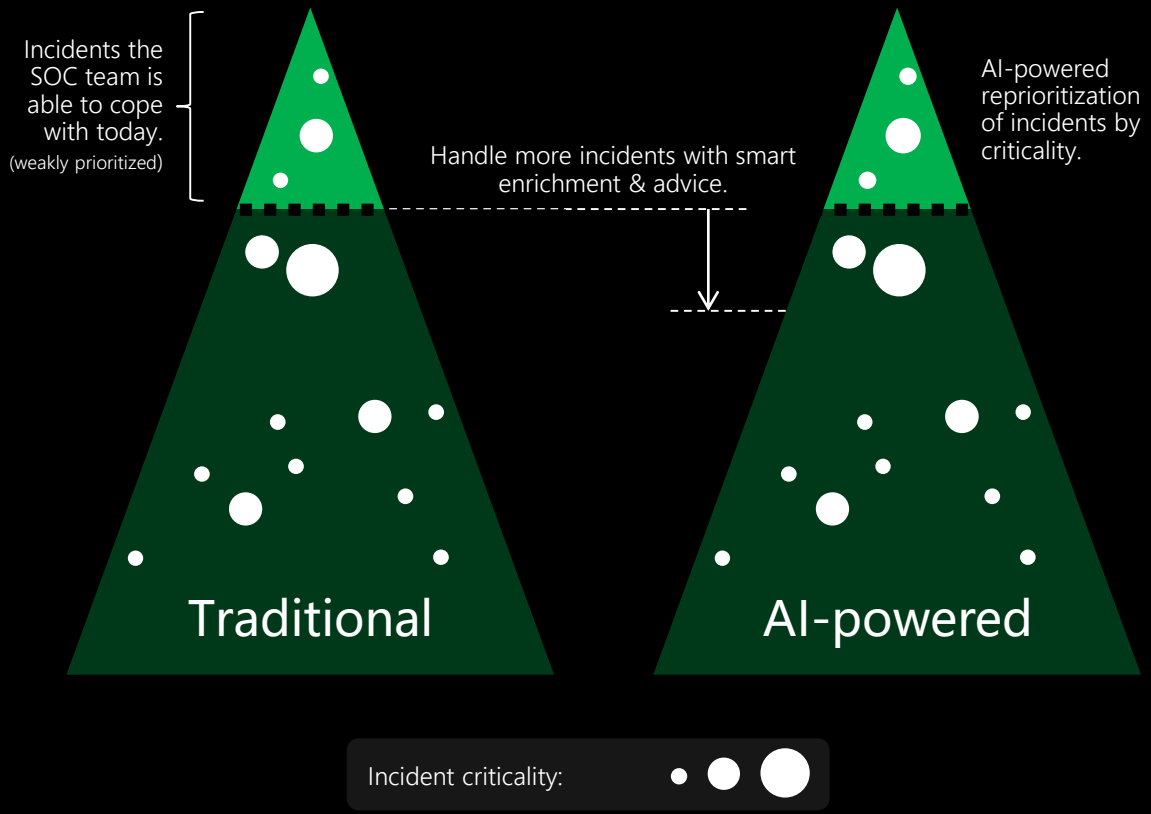
Use of AI in the Threat Management Lifecycle



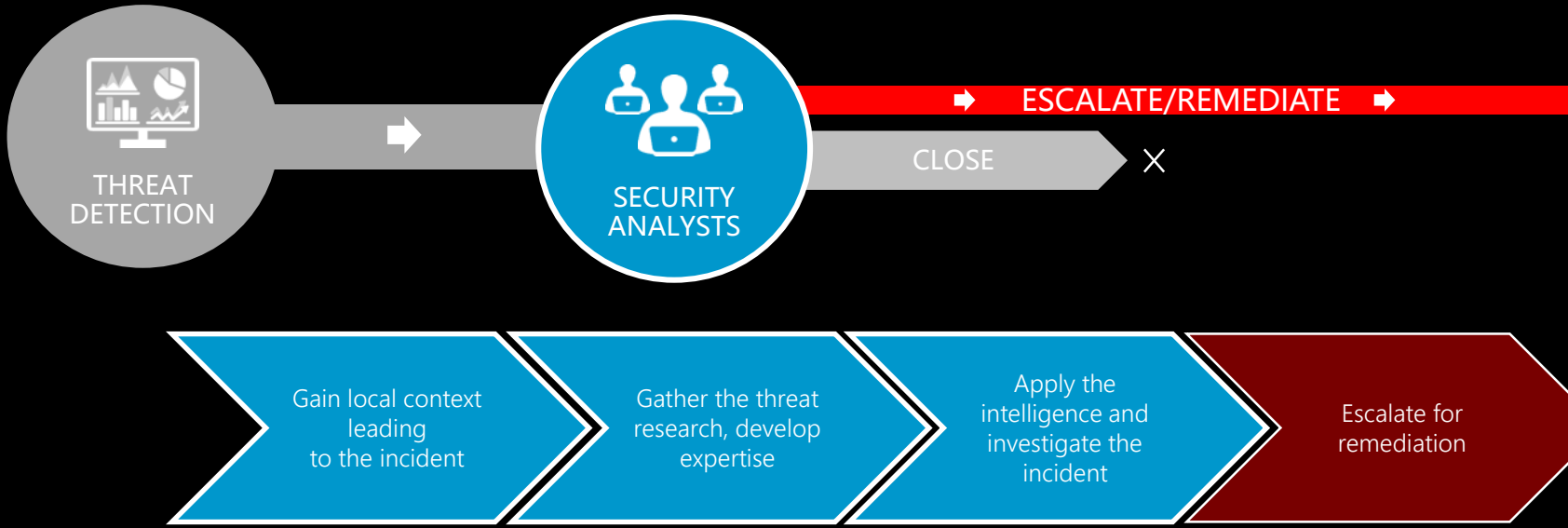
Pain points of Security Operations teams

- Flood of alerts
- Shortage of qualified staff
- Too many tools, too many technologies
- Increasing threat landscape

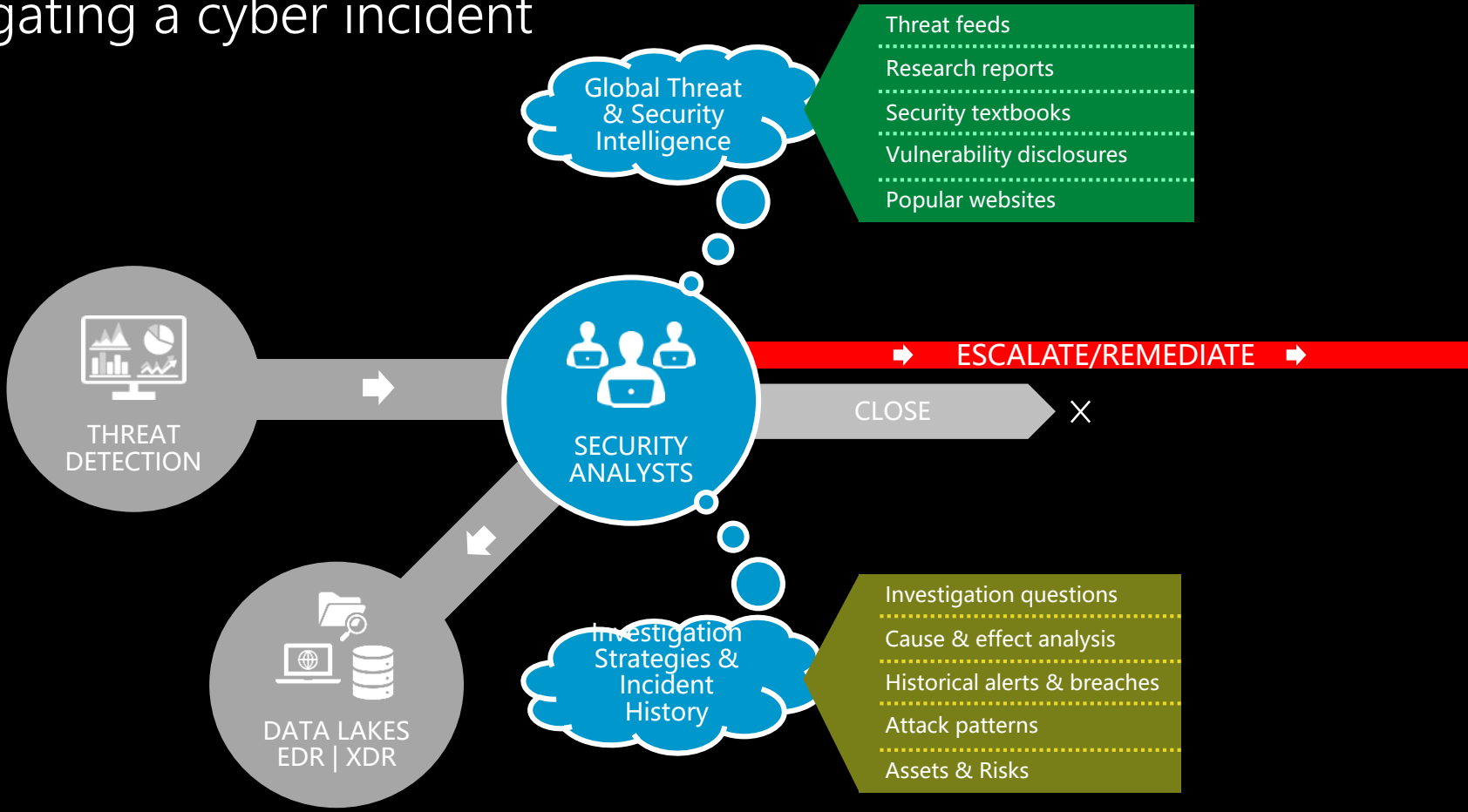
Addressing the Investigation Bottleneck



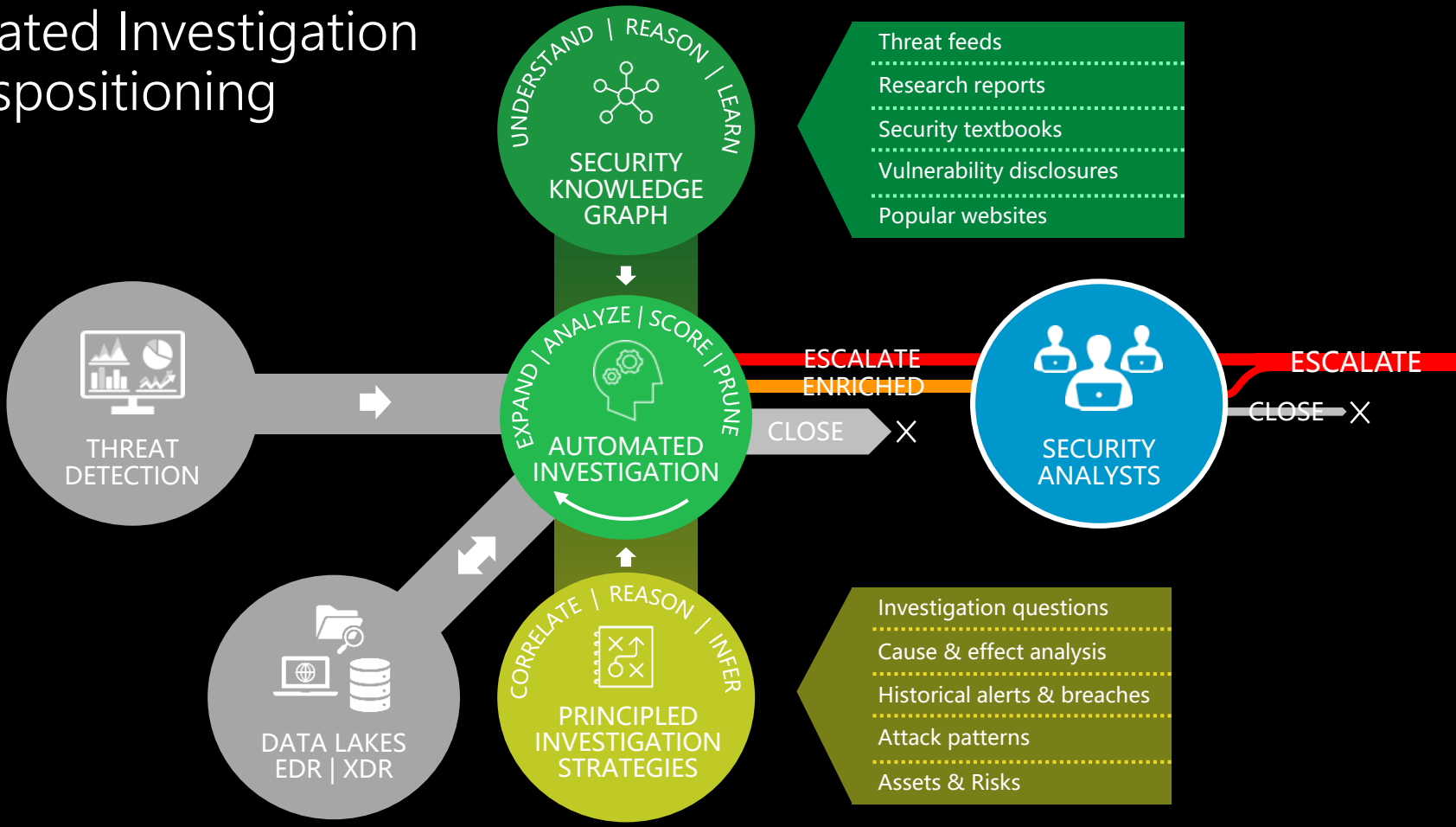
Investigating a cyber incident



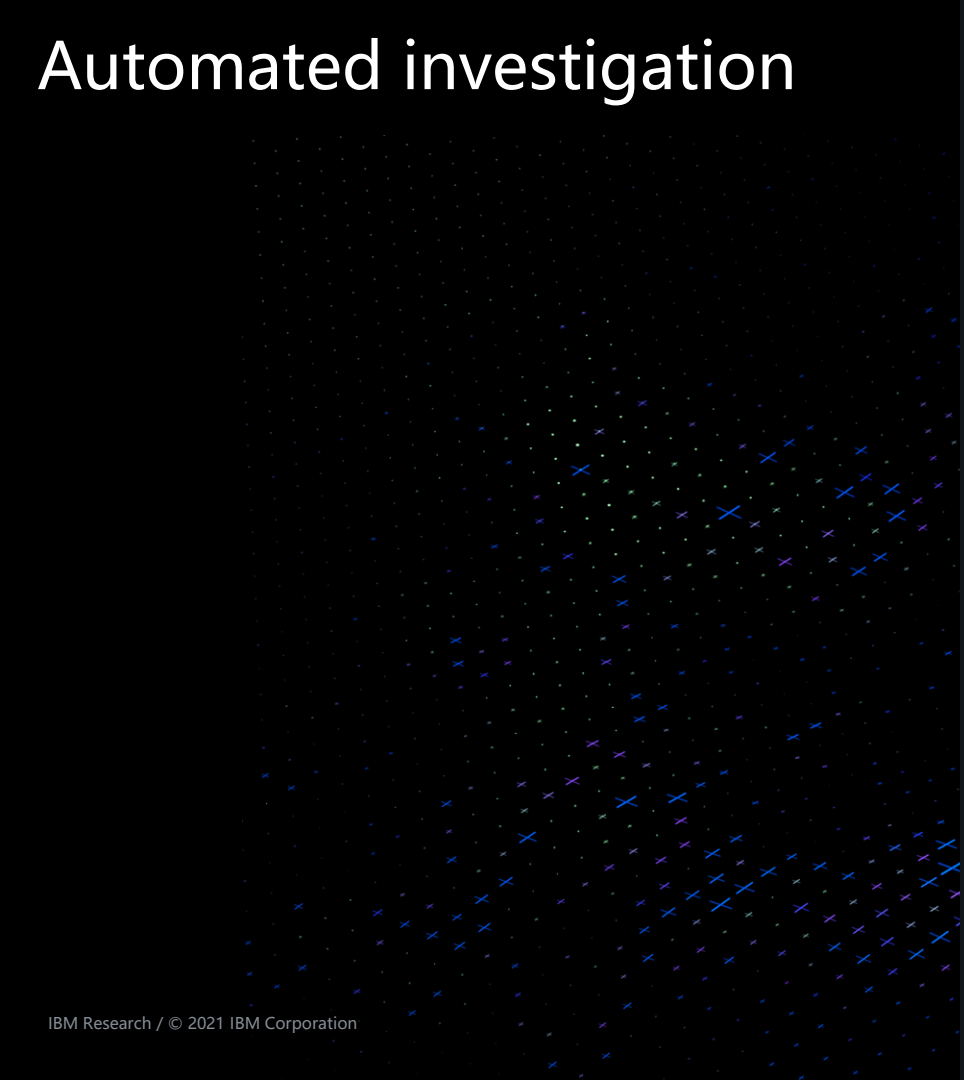
Investigating a cyber incident

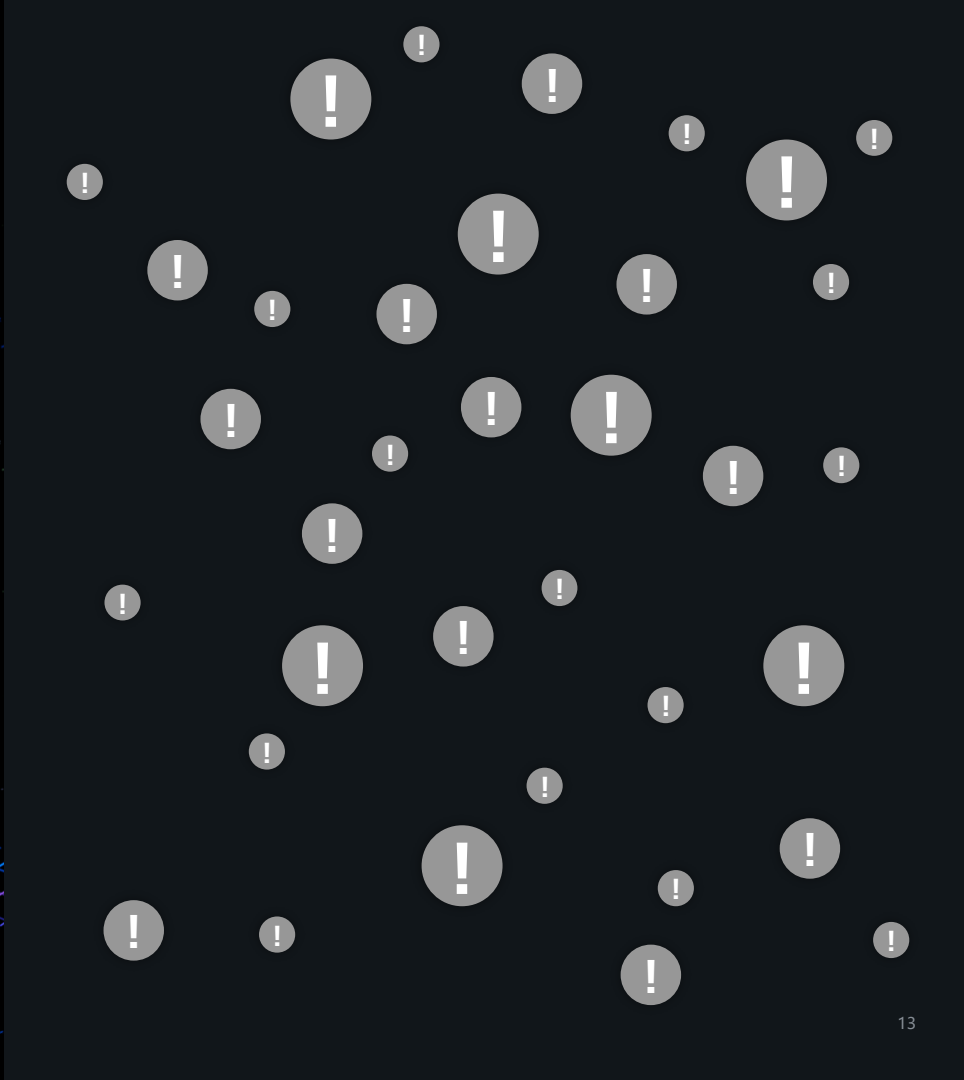


Automated Investigation and Dispositioning

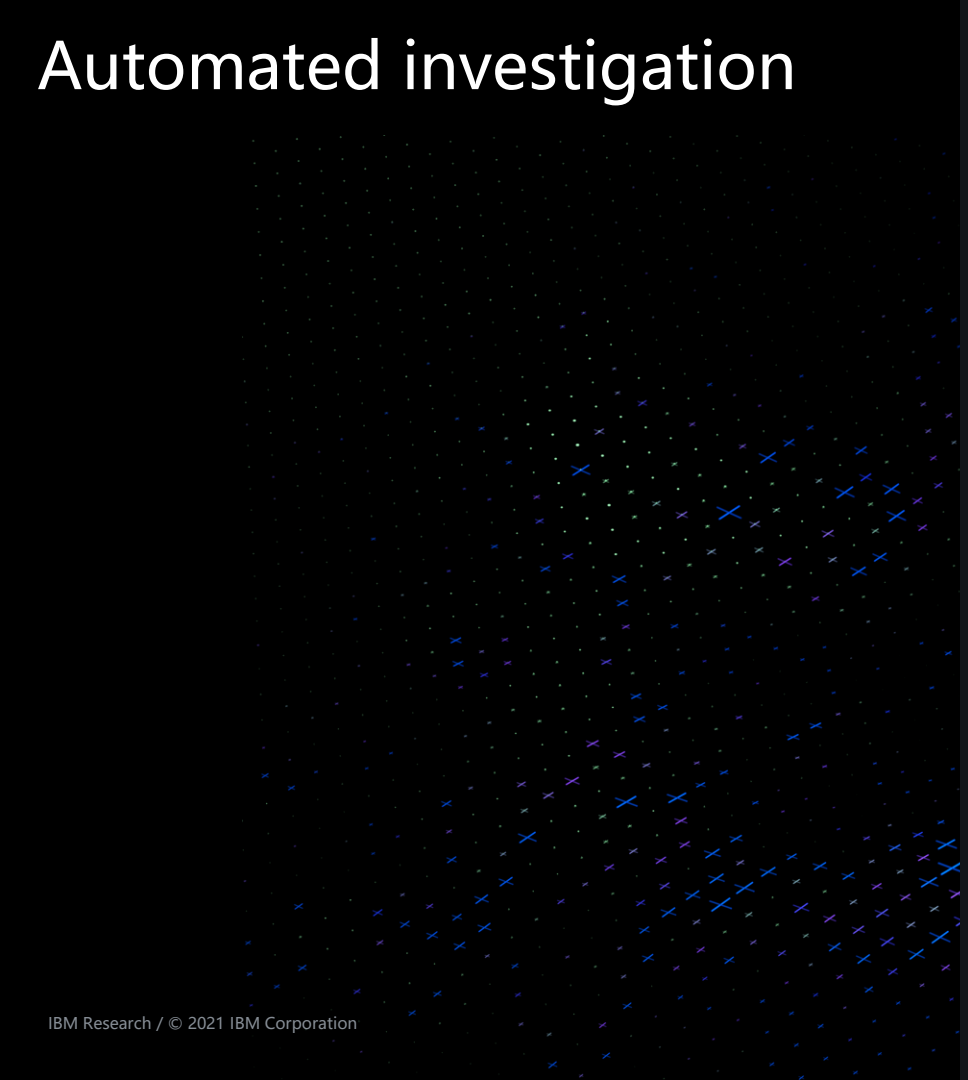


Automated investigation



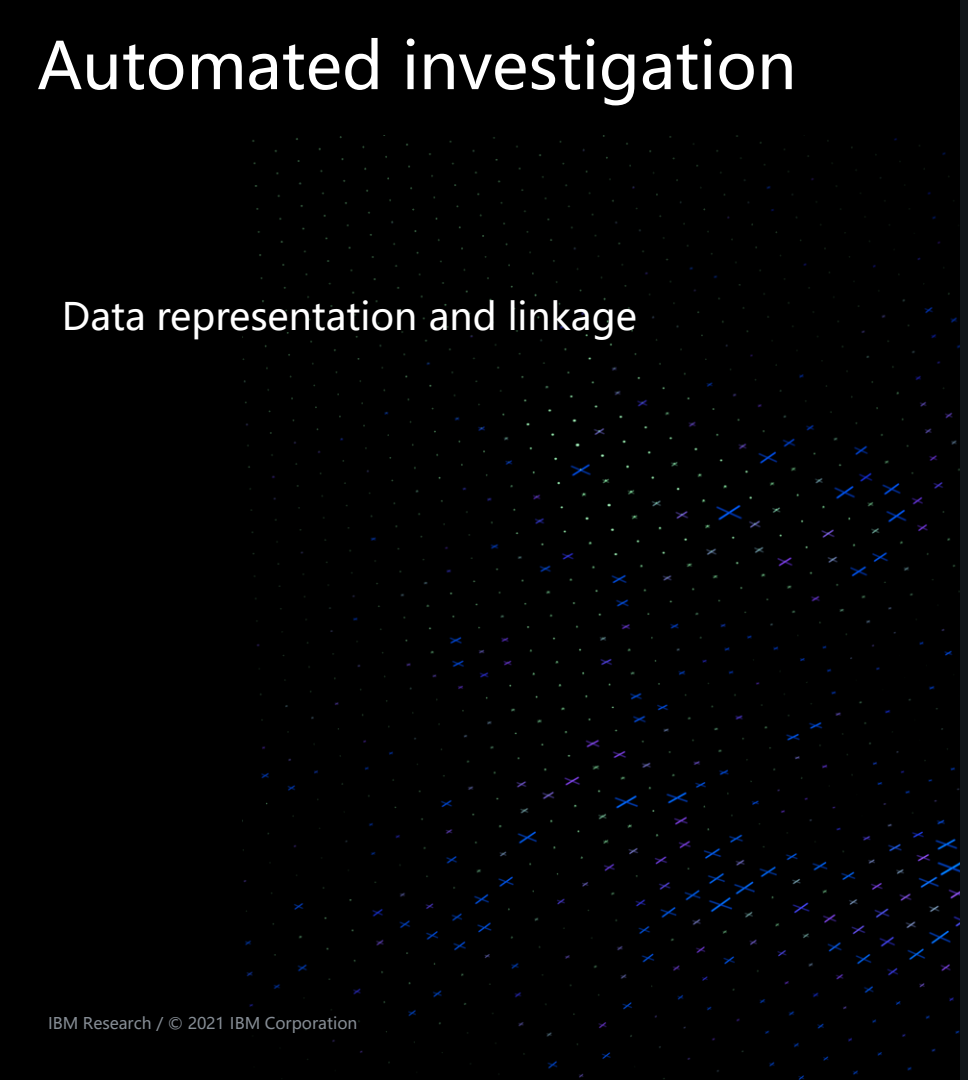


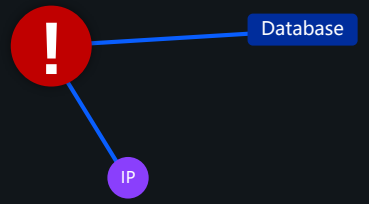
Automated investigation



Automated investigation

Data representation and linkage

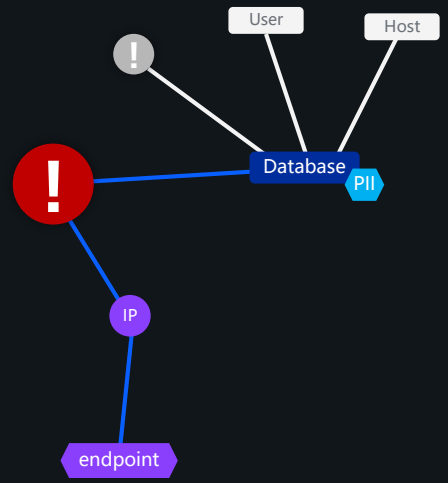




Automated investigation

Data representation and linkage

Environment context

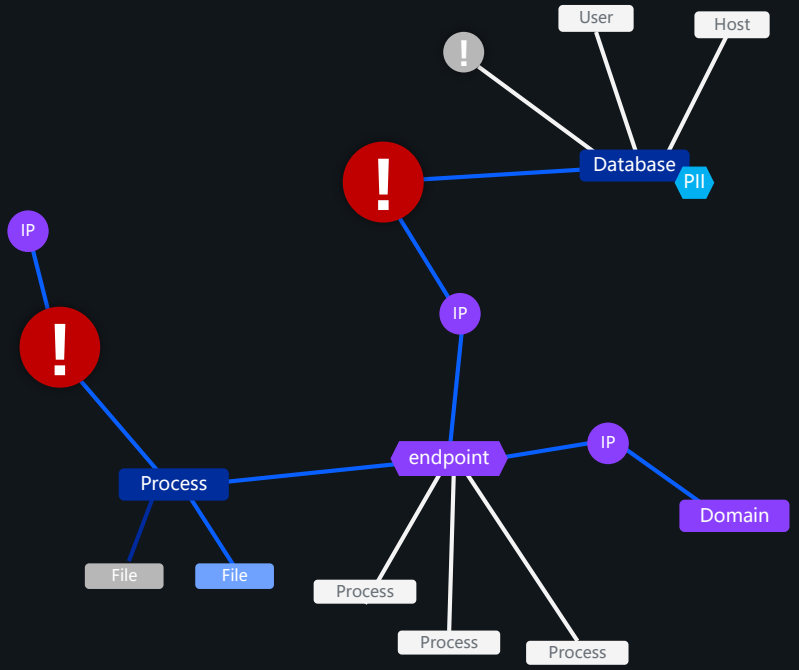


Automated investigation

Data representation and linkage

Environment context

Forensic reasoning



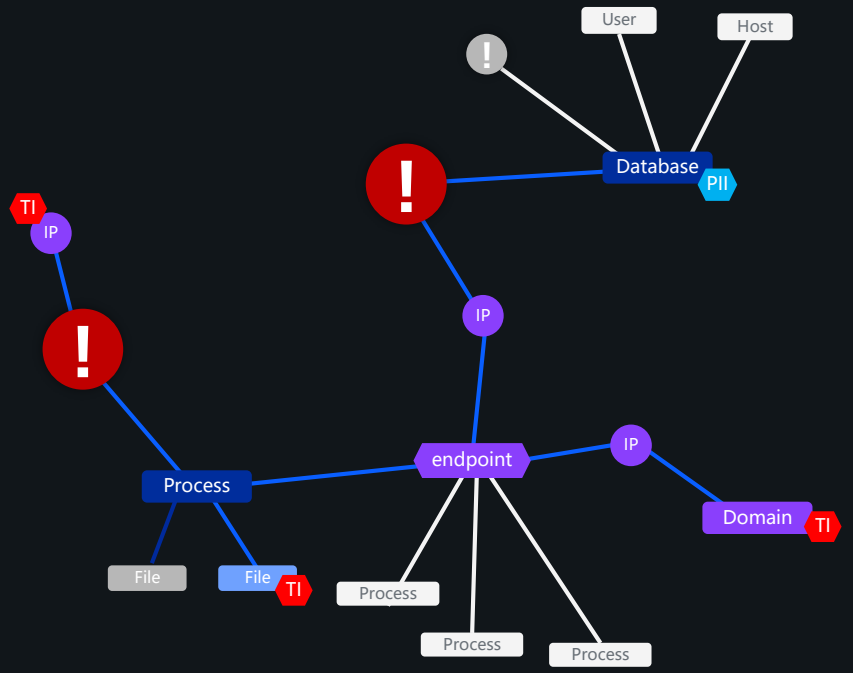
Automated investigation

Data representation and linkage

Environment context

Forensic reasoning

Threat intelligence



Automated investigation

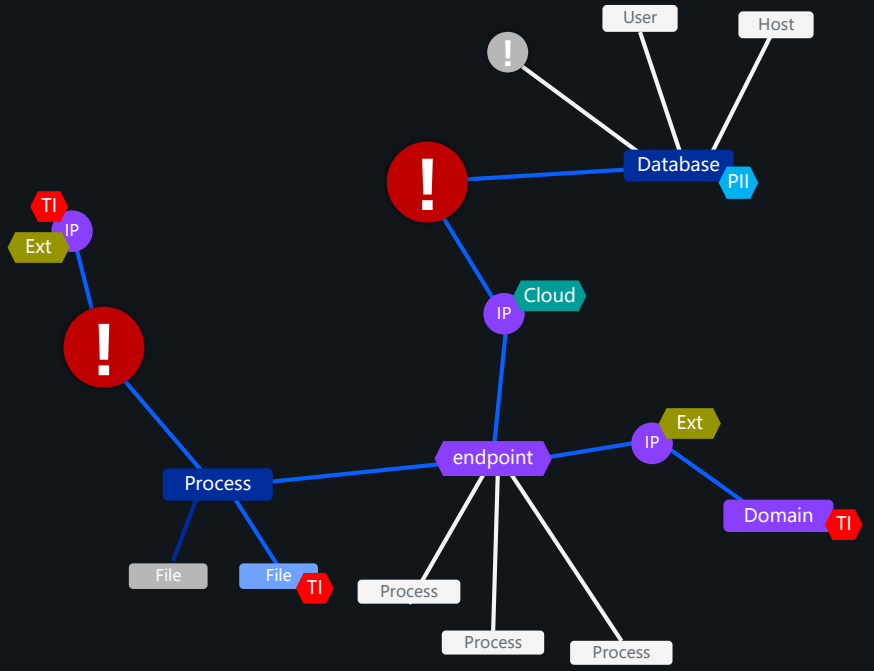
Data representation and linkage

Environment context

Forensic reasoning

Threat intelligence

Domain knowledge



Automated investigation

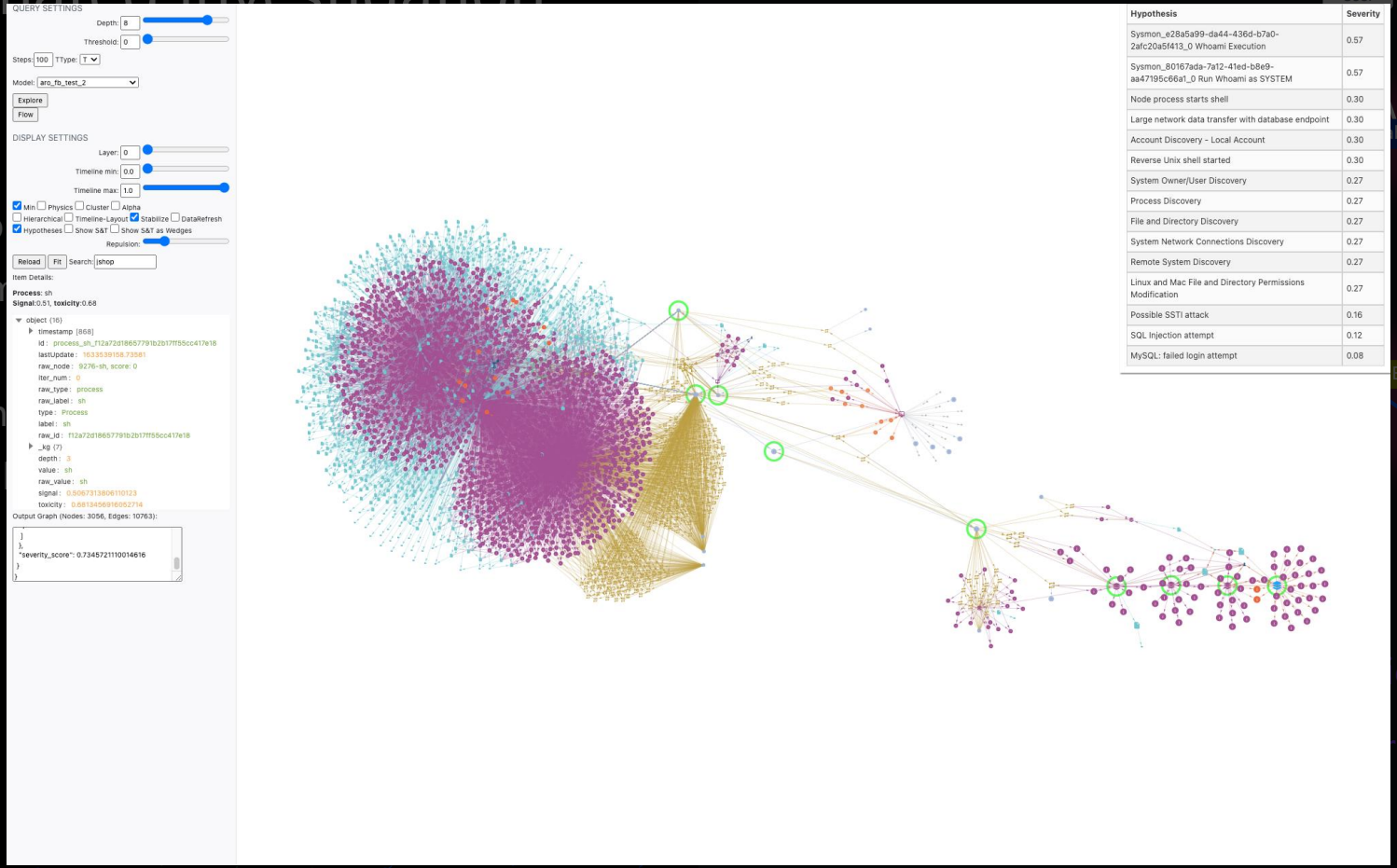
Data rep

Environm

Forensic

Threat in

Domain



Hypothesis	Severity
Sysmon_e28a5a99-da44-436d-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

User

Host

base

PII

Ext

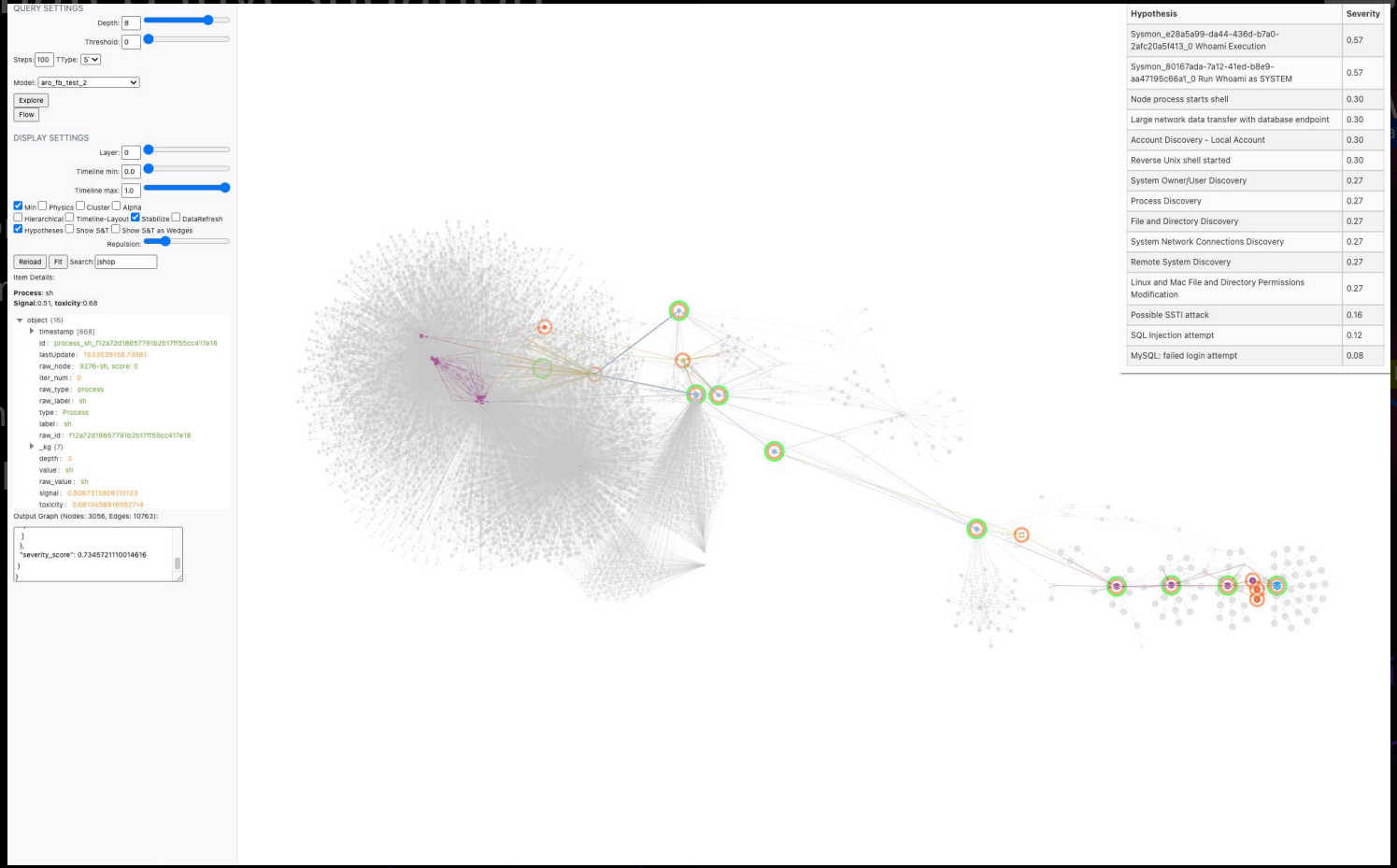
Domain

TI

c reasoning

ration

Automated investigation



Hypothesis	Severity
Sysmon_e28a5a89-da44-436d-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

Automated investigation

Data representation and linkage

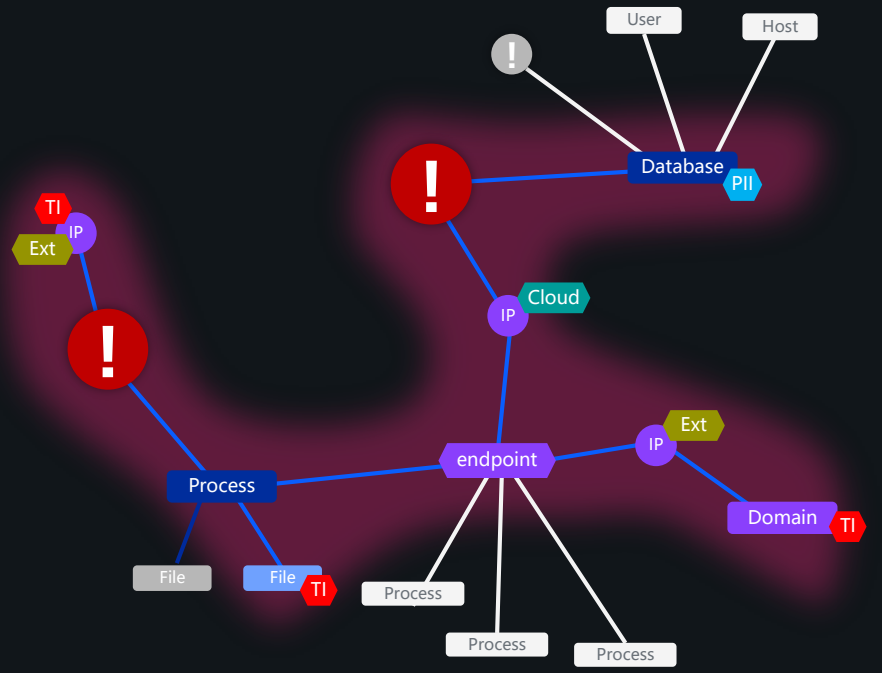
Environment context

Forensic reasoning

Threat intelligence

Domain knowledge

Data reduction



Query Optimization

Semantic reasoning

Graph Theory

Hypothesis generation

STIX

MITRE | ATT&CK®

Automated investigation

Data representation and linkage

Environment context

Forensic reasoning

Threat intelligence

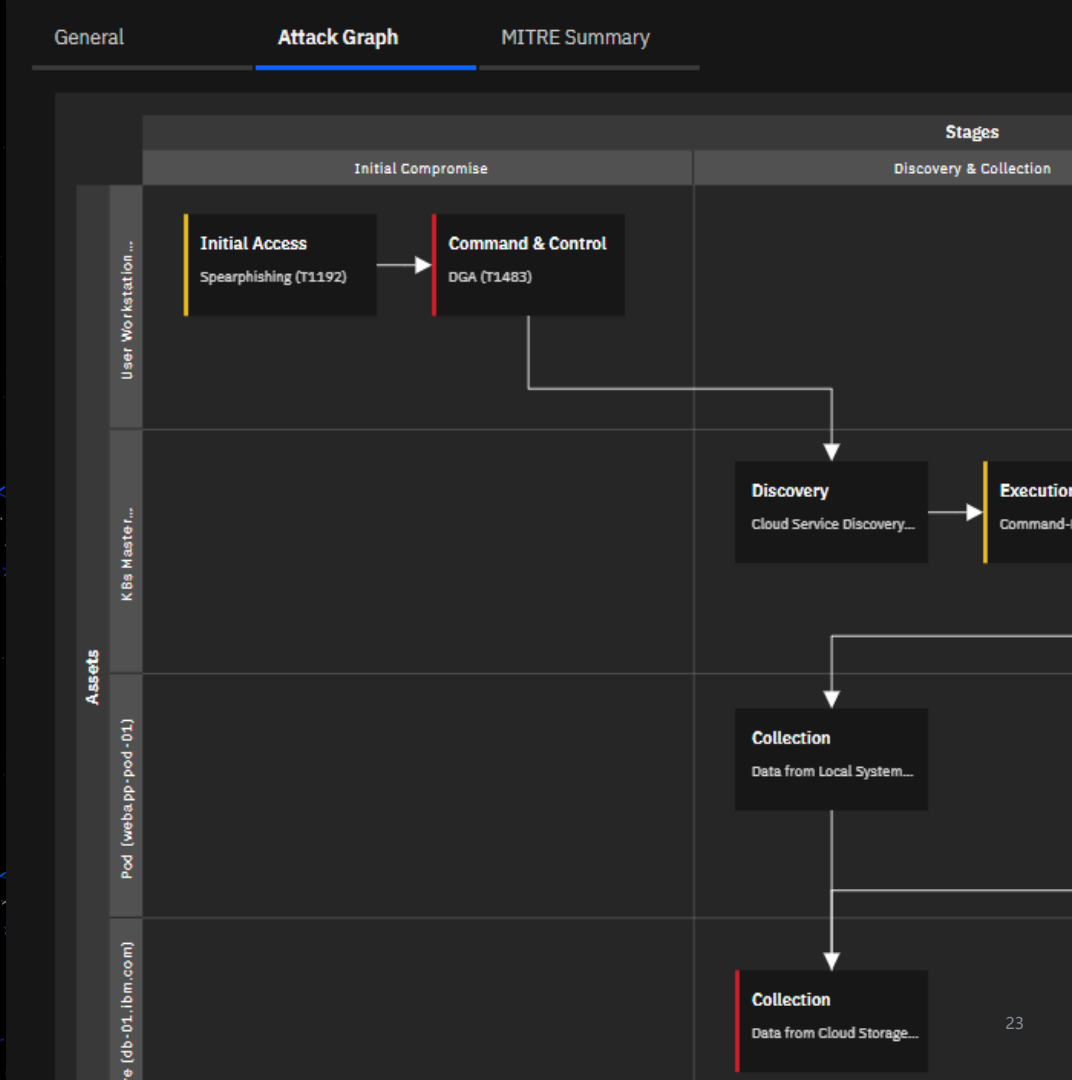
Domain knowledge

Data reduction

Attack kill chain reasoning

Root cause analysis

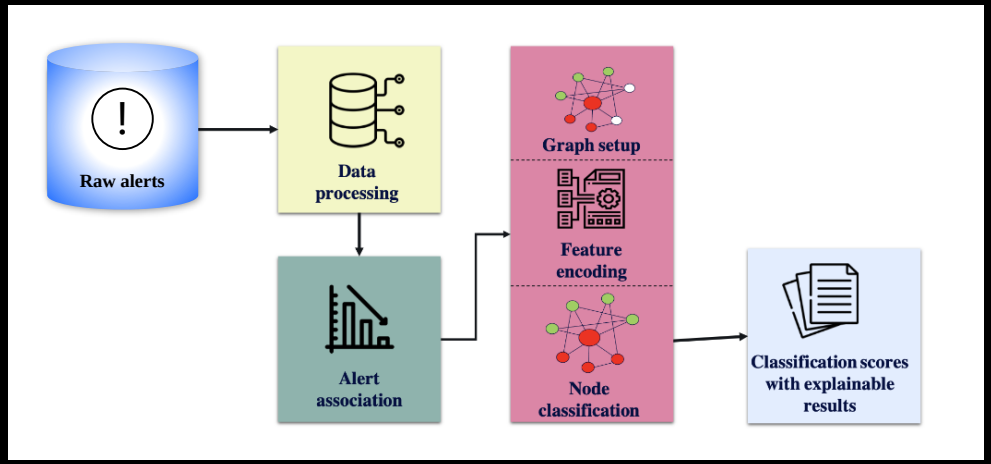
Extent & effect determination

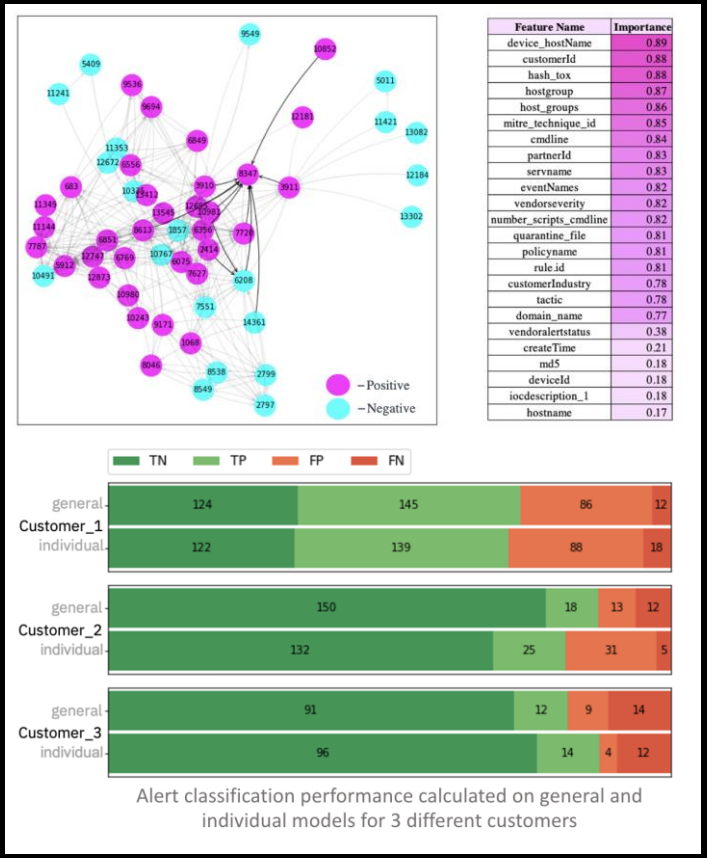


Alert Similarity Analysis and Classification

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

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





Feature Name	Importance
device_hostName	0.89
customerId	0.88
hash_tox	0.88
hostgroup	0.87
host_groups	0.86
mitre_technique_id	0.85
cmdline	0.84
partnerId	0.83
servname	0.83
eventNames	0.82
vendorseverity	0.82
number_scripts_cmdline	0.82
quarantine_file	0.81
policyname	0.81
rule.id	0.81
customerIndustry	0.78
tactic	0.78
domain_name	0.77
vendoralertstatus	0.38
createTime	0.21
md5	0.18
deviceId	0.18
iocdescription_1	0.18
hostname	0.17

AI and Security: Considerations and Reflections

Data

Quality
Quantity
Granularity

Applicability

Precision
Speed
Maintenance

Trustworthiness

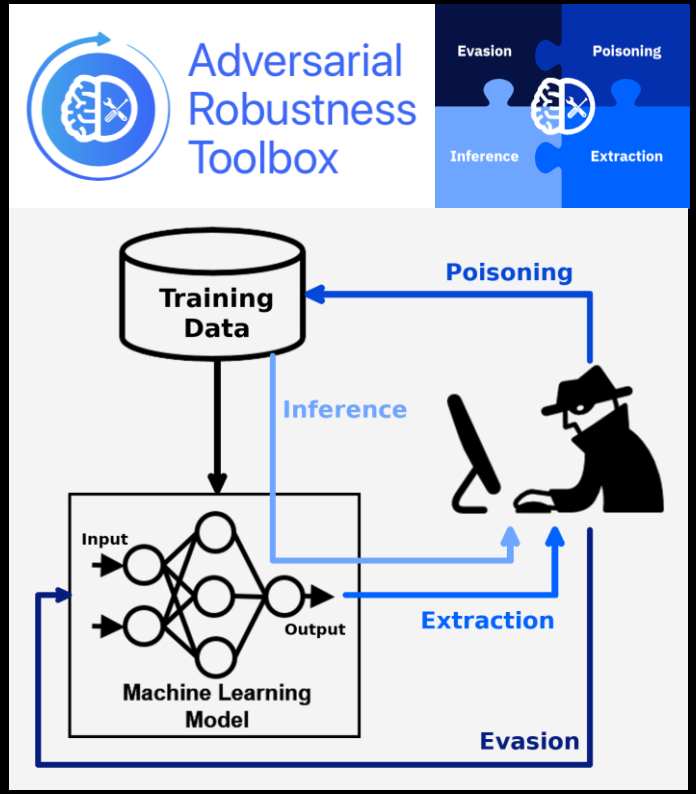
Robustness
Explainability
Privacy

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.



https://github.com/Trusted-AI/adversarial-robustness-toolbox

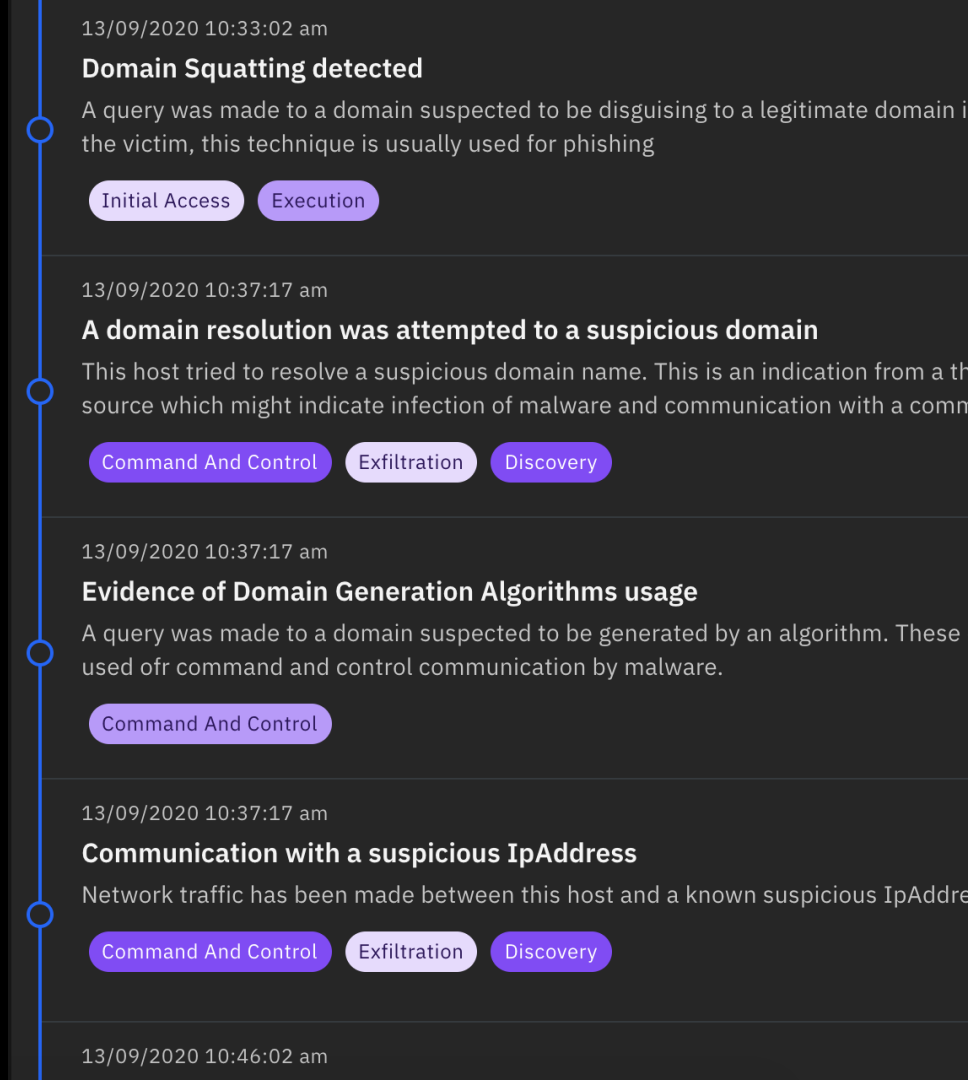
Summary and Next Steps

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 …



IBM **Research** Security

THANK YOU

Marc Ph. Stoecklin

mtc@zurich.ibm.com

IBM