RSAConference2019
San Francisco | March 4 – 8 | Moscone Center

BETTER.

SESSION ID: SPO1-T07

Accelerate and Simplify Incident Response with Security Automation

Nick Bilogorskiy

Cybersecurity Strategist
Juniper Networks
@belogor

Agenda

- Advanced Threats TTPs
- Modern SOC Problems
- Machine Learning Demystified
- Automation of Incident Response
- Questions

Trends: Passwords are the New Exploits

- 32% of hackers say accessing privileged accounts is the fastest way to hack.
- 81% of breaches leveraged stolen or weak passwords
- Brute forcing a website with a set of stolen passwords is called credential stuffing

Trends: Attacks on 2-factor authentication

SIM swapping is tricking a mobile provider into moving the victim's phone number to another SIM card that is controlled by the attacker.

Trends: Software Supply Chain attacks

- Supply Chain Attacks Surged 200% in 2017
- 42% of companies had a data breach caused by a cyber attacks against third parties
- Two thirds (66 percent) grant privileged account access to third-party partners, contractors or vendors.

statcounter

python™ Package Index

me doc
МІЙ ЕЛЕКТРОННИЙ ДОКУМЕНТ

Trends: Attack automation and packaging

NETWORKComputing | Join us live at Interop ITX

Authors | Slideshows | Video | Tech Library | University

NETWORKING | STORAGE | WIRELESS | DATA CENTER | NET SECURITY | DATA

MOBILE

04/25/2012
12:20 PM

Mathew Schwartz
News
Connect Directly

Anonymous Hackers' New Best Friend: Automation

Anonymous hacktivists and crime syndicates favor free, automated tools to easily and quickly exploit website vulnerabilities. How can enterprises fight back?

SECURITYWEEK NETWORK: Information Security News |

SECURITYWEEK
INTERNET AND ENTERPRISE SECURITY NEWS, INSIGHTS & ANALYSIS

Subscribe (Free) | CISO Forum

Malware & Threats | Cybercrime | Mobile & Wireless | Risk & Compliance | Security Architecture | Secu

Vulnerabilities | Email Security | Virus & Malware | IoT Security | Endpoint Security

Home › Vulnerabilities

AutoSploit: Automated Hacking Tool Set to Wreak Havoc or a Tempest in a Teapot?

By Kevin Townsend on February 01, 2018

Share | G+ | Tweet | Recommend 21 | RSS

AutoSploit Automatically Finds Vulnerable Targets via Shodan and Uses Metasploit Exploits to Compromise Hosts

BBC | News | Sport | More | Search

NEWS

Home | Video | World | US & Canada | UK | Business | Tech | More

Technology

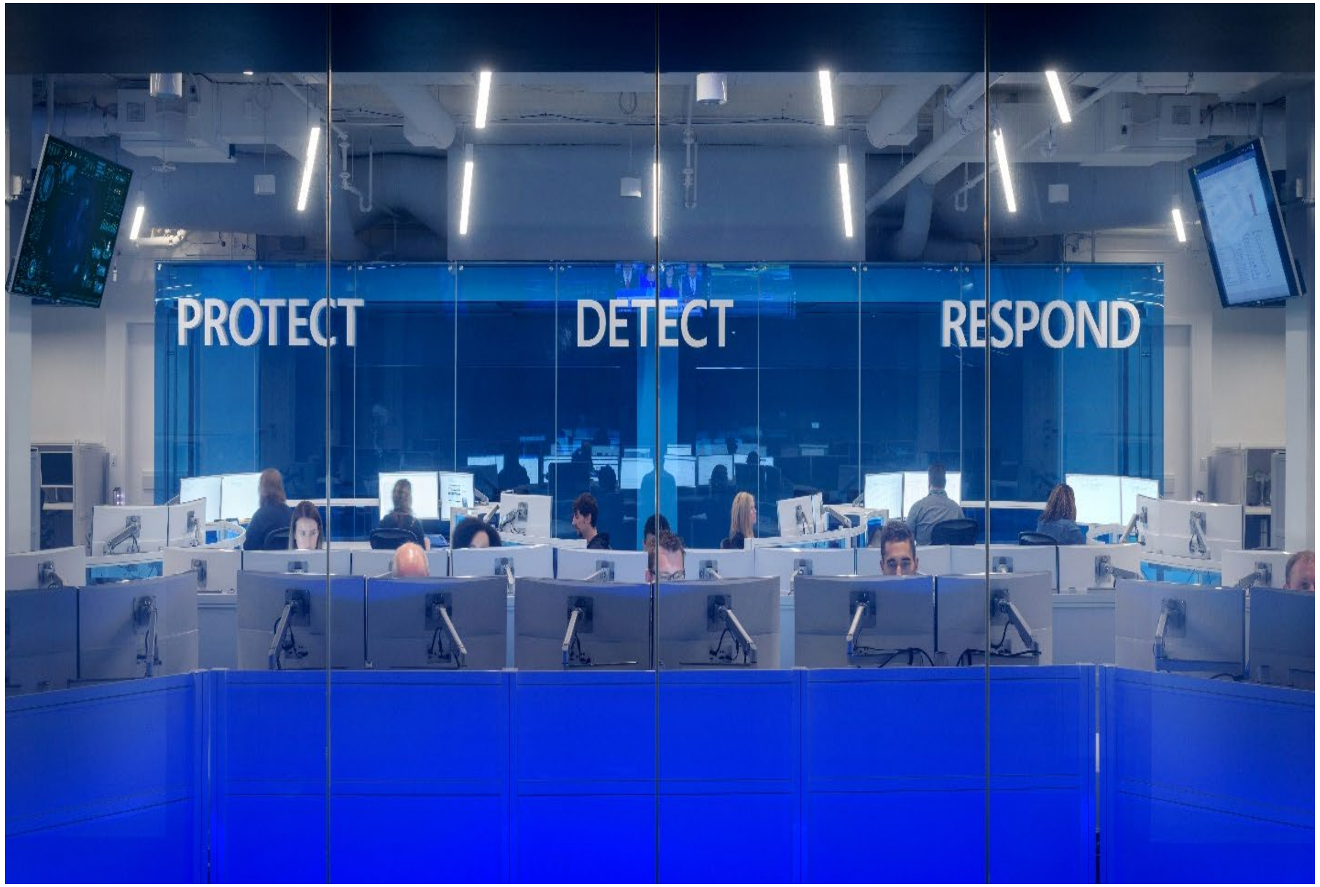
'Lazy hackers' turn to automated attack tools

17 April 2018

Share

Modern SOC problems

- Alerts Overload
- Staffing Challenges
- Complexity
- Threats Evolving Faster Than Defenses

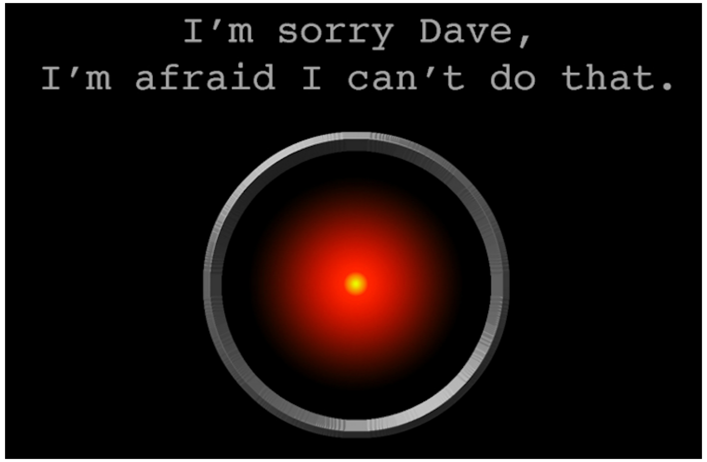


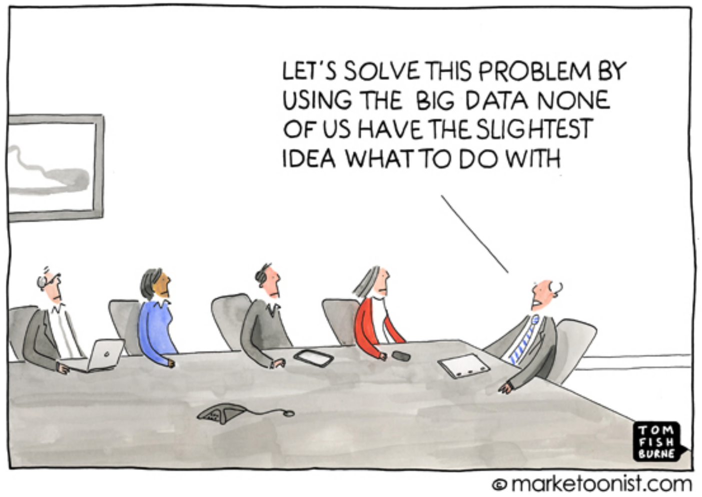
“Assume Breach”

RSAConference2019

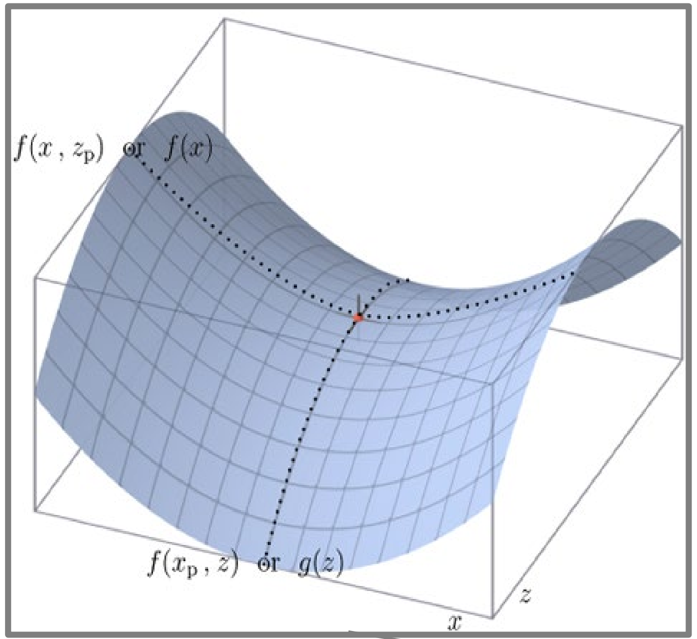
Machine Learning Demystified

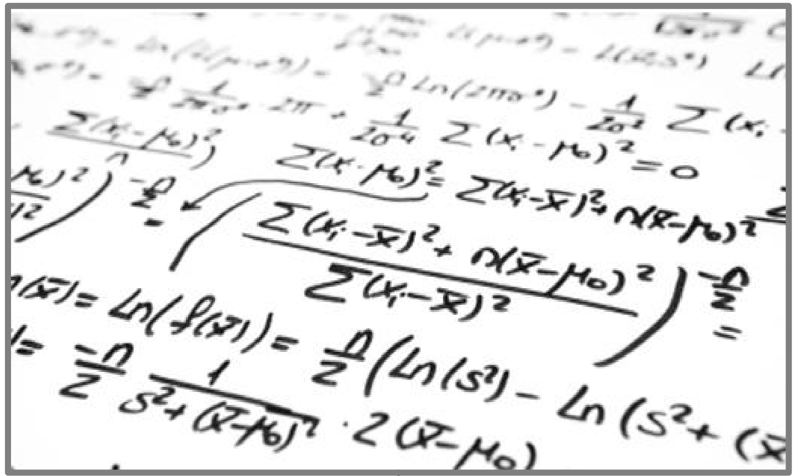
Hype vs. Reality: The Hype

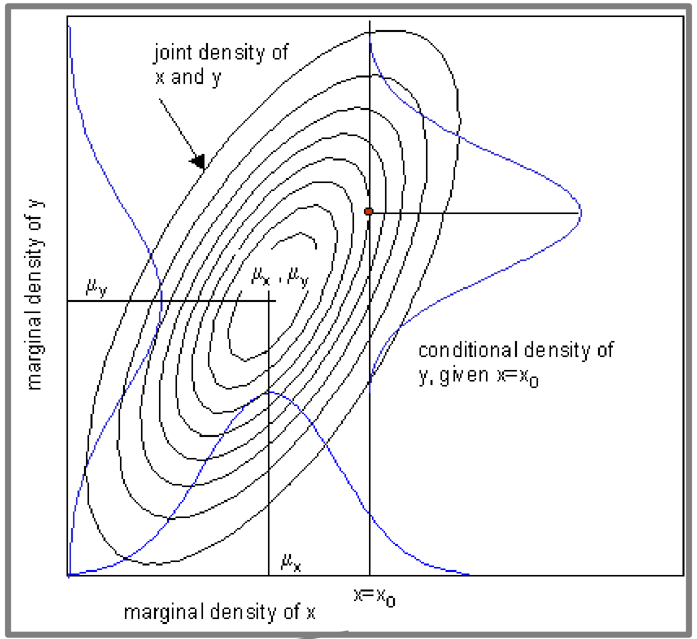




Hype vs. Reality: Reality







Data, numerical software, high performance computing

Prediction, classification, pattern discovery

Security Applications

- Given information about a file or event, answer:
  - Is a file or event malicious? (Yes, No)
  - If malicious, what type of malware is it? (Trojan, Worm, Adware, etc.)
  - How can I *quantify* the risk of the attack? (High, Medium, Low)

Traditional Approaches TO THREAT DETECTION

1	Static	Packer, file type, file size, code obfuscation Detection by checksum match, static property signatures Fast but lacking coverage of newest samples (see WannaCry, for ex.)
2	Reputation	Crowdsourcing multiple detection engines (VirusTotal) Combine detections based on file hash Good coverage but detection lags due to nature of crowdsourcing. Feedback effects (vendors alter detection based on VT data)
3	Behavioral	Log behavior from sandboxing (file creation, CnC activity, etc) Manually create “behavioral signatures” Naïve Bayesian score based on signatures Can detect unknown samples but takes time (1-10 minutes)

Benefits of ML Applied to Behavioral Detection

- Can detect malware using indirect indicators
  - IOC – indicator of compromise, i.e. an action only taken by malware
  - Indirect IOC, action that is not necessarily malicious
    - i.e. looking in a window vs breaking a window
- Indirect indicators are difficult to disguise
  - Relative frequency of certain actions
  - Combinations of actions
- Indirect indicators may provide more generalized detection
  - Able to detect different families that share “tradecraft”

Benefits of ML Applied to Behavioral Detection

- Can easily customize detection focus
  - Using malware training set with particular composition
    - for example, with or without adware
- Can adapt to deployment environment
  - Using benign samples from a given organization

In ML Data is King

All machine learning models need to be "trained" on data.

File/event samples

(Training Data)

→

Feature extraction

- Static
- Behavioral
- Reputation
- etc.

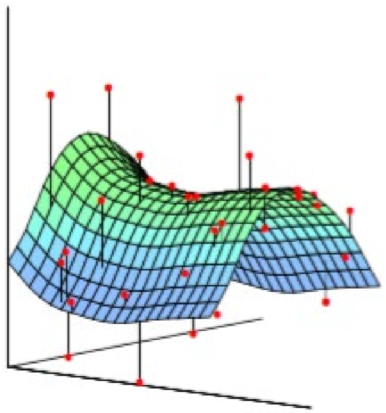
→

Analyze, clean and normalize data

$$
\begin{array}{cc}
 & n \text{ columns} \\
A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{i1} & a_{i2} & a_{i3} & \cdots & a_{in} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix} & m \text{ rows}
\end{array}
$$

→

Train Model



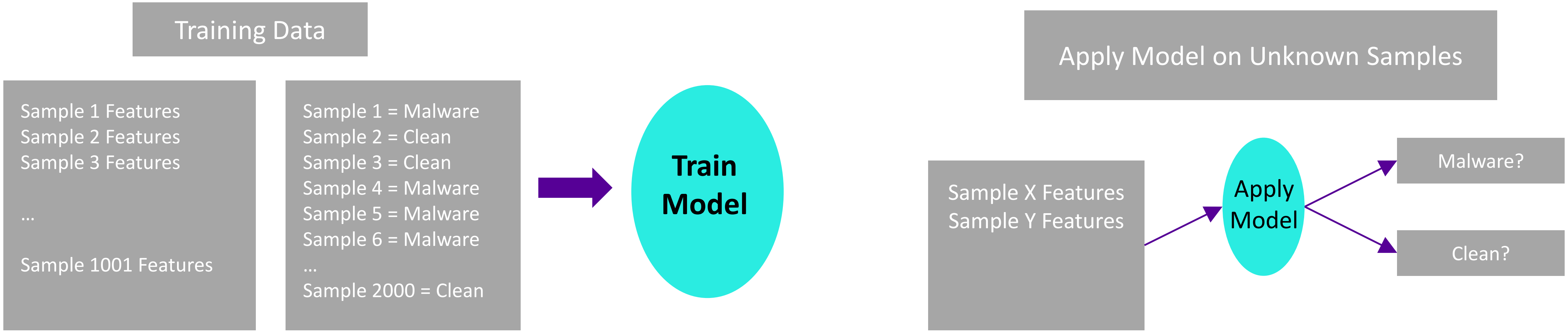
The training data is <u>*the most important factor*</u> in the success of the model.

The Machine Learning Toolkit

- Supervised Learning
- Unsupervised Learning
- Semi-supervised Learning
  - Combination of supervised + unsupervised

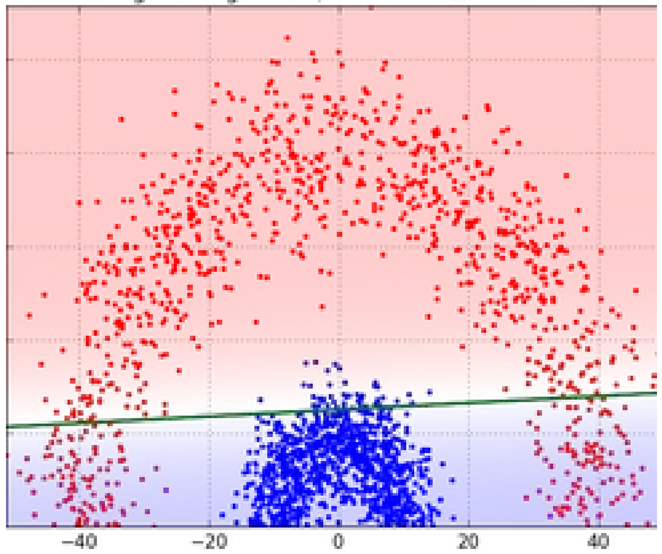
Supervised Learning: Binary Classification

- The outcome of each training sample is already known
- Training Techniques (i.e. Model Types):
  - Linear/Logistic Regression
  - Support Vector Machines (SVM)
  - Classification Trees, Random Forests, Boosted Trees (XGBoost)
  - Neural Networks ("Deep Learning": CNN, RNN)

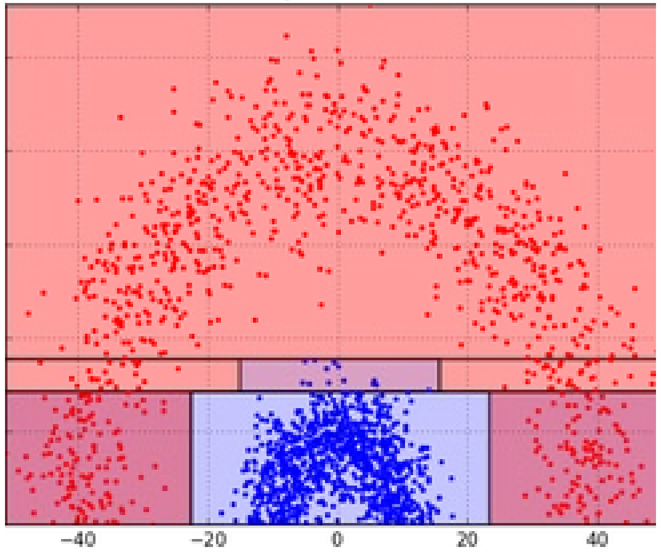


Linear/Logistic Regression vs Decision Trees

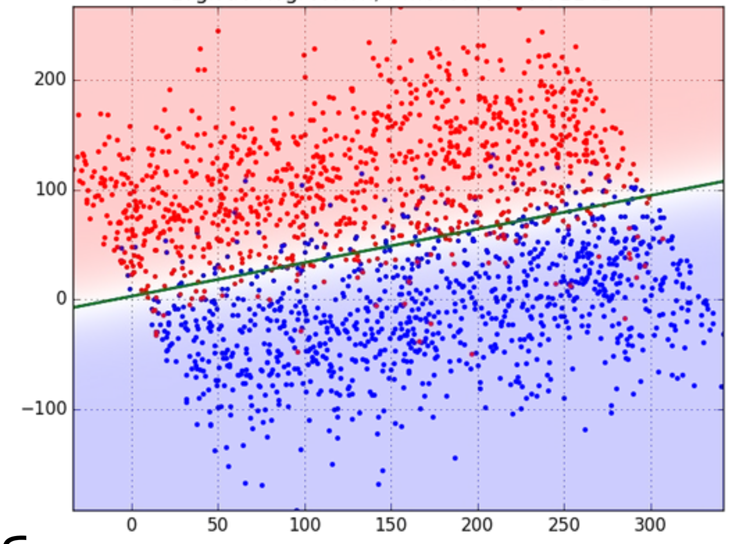
Logistic Regression, f-measure = 0.854290



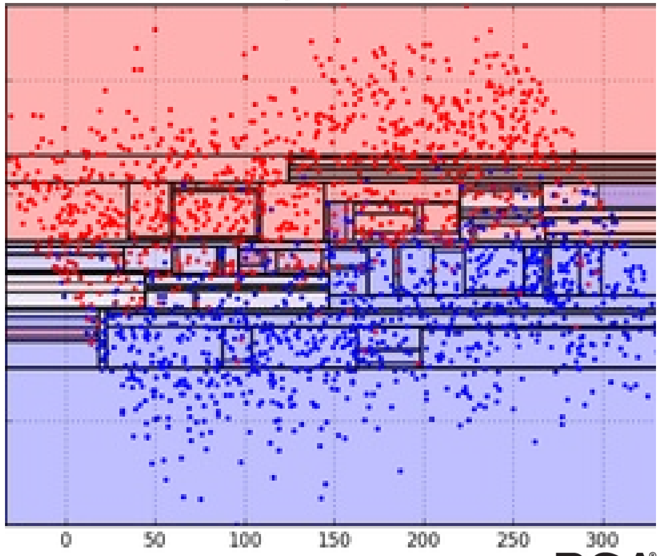
Decision Tree, f-measure = 1.000000



Logistic Regression, f-measure = 0.922420

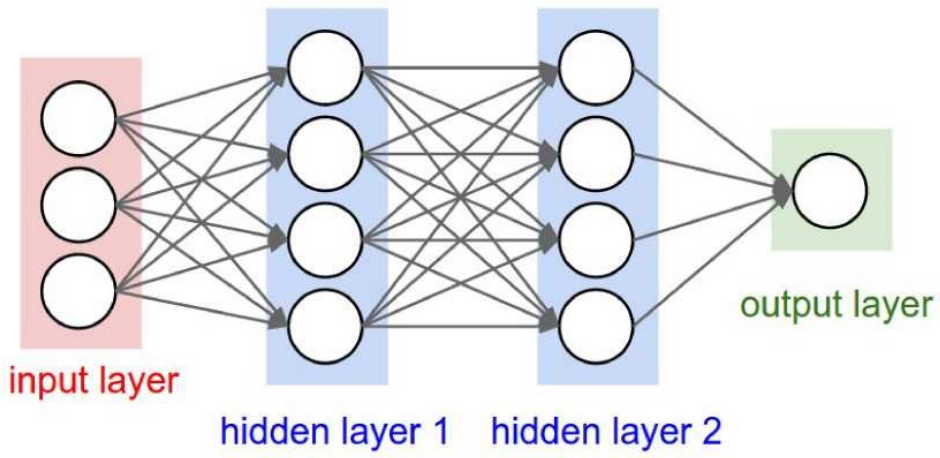


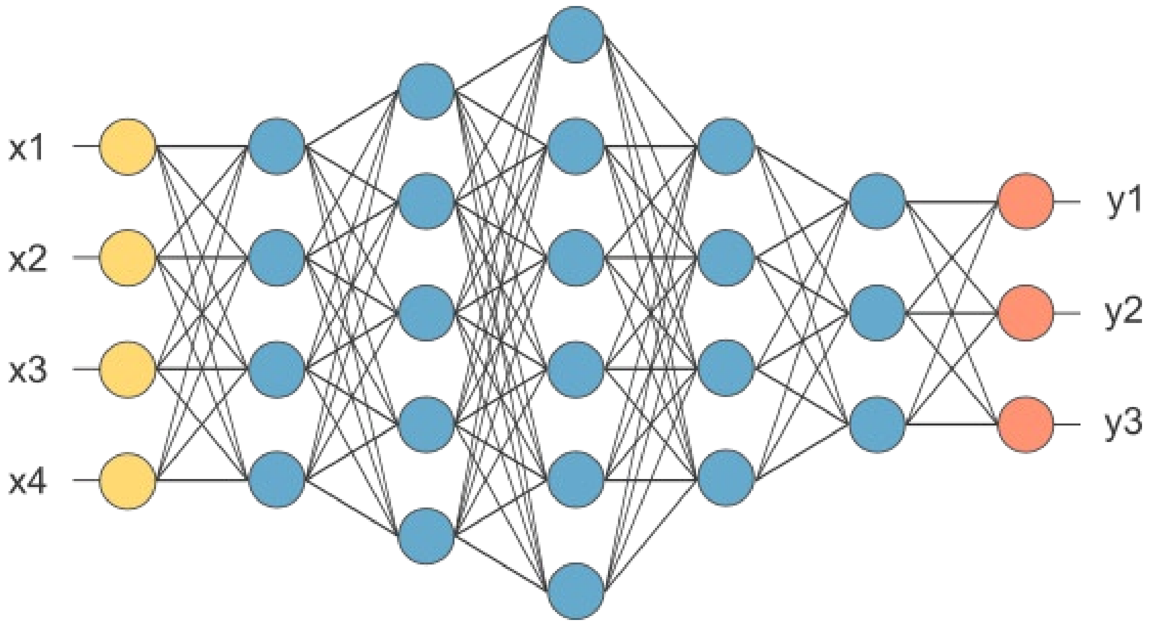
Decision Tree, f-measure = 0.889780



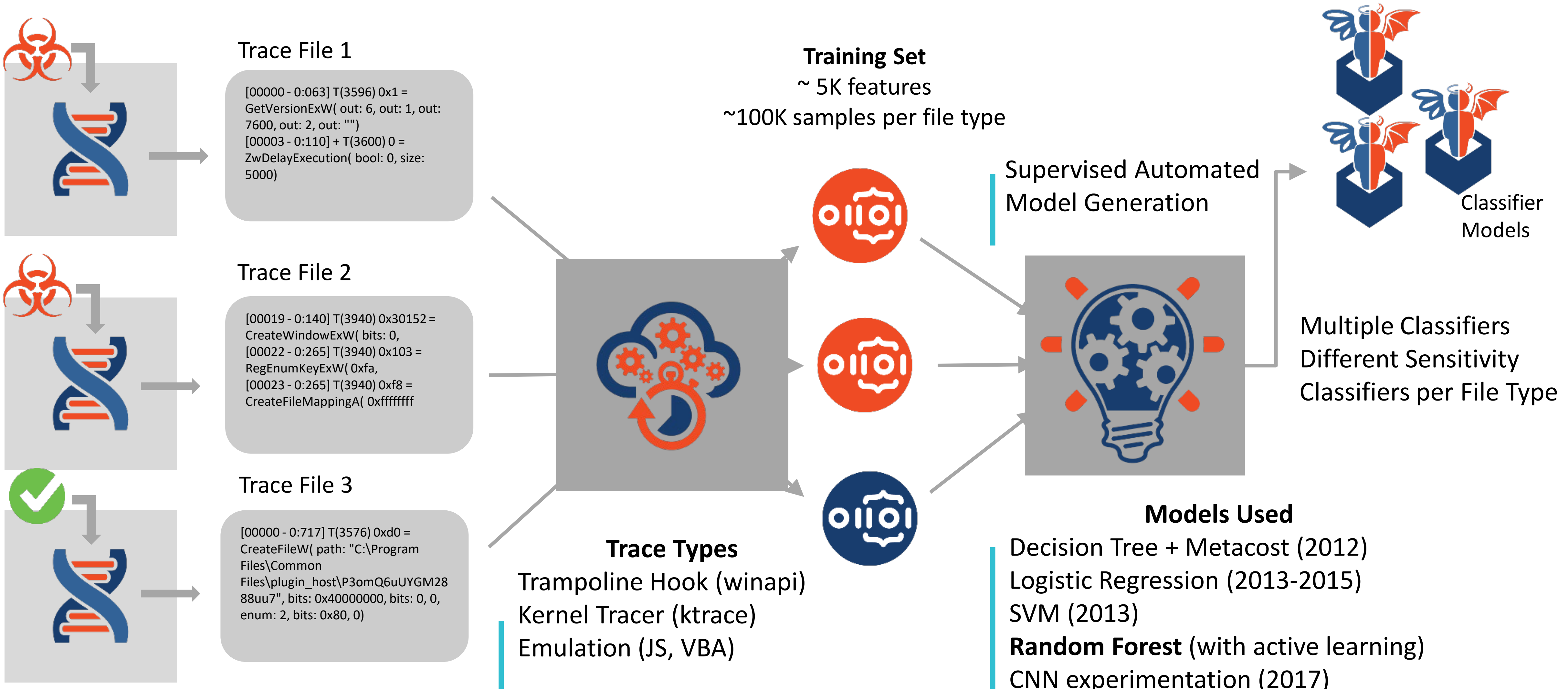
Tangent: What Is Deep Learning?

- Deep learning does not mean "deep understanding"
  - Deep learning uses a Neural Network as the ML model
  - "Deep" refers the number of hidden layers in the network





Machine Learning Model Generation



Trace Types
Trampoline Hook (winapi)
Kernel Tracer (ktrace)
Emulation (JS, VBA)

Models Used
Decision Tree + Metacost (2012)
Logistic Regression (2013-2015)
SVM (2013)
Random Forest (with active learning)
CNN experimentation (2017)

Unsupervised Learning

- The ”outcome” of each training sample is unknown
- Example: Finding families of malware
- Techniques:
  - Clustering algorithms
  - Self-organizing maps
  - Principal Components Analysis (PCA)
  - Archetypal Analysis

Training Data

Sample 1 Features
Sample 2 Features
Sample 3 Features

…

Sample 1001 Features

Train Model

Discover similar “groupings” of samples

Group 1
Sample 17
Sample 1
Sample 264

…

Group 2
Sample 107
Sample 6
Sample 2

…

Group 3
Sample 936
Sample 851
Sample 1001

…

The Machine Learning Toolkit - Clustering

Clustering is a popular ML tool in malware analysis.

(Feature = “Dimension”)

But things break down in higher dimensions!

Separating the Signal from the Noise

A "Needle in the haystack" situation: 1 out of every 100,000 samples

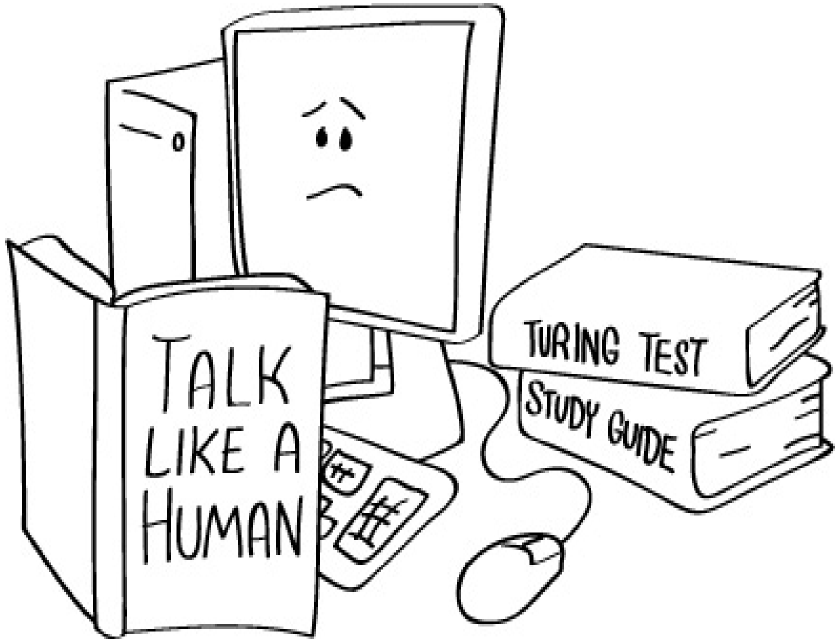
- Build a *classifier* which can detect 95% of threats with a 1% FP rate
  - 1 FP for every 100 objects, or 1000 FPs

- Note: an FP is 1000x more likely than a detection!
  - Leads to a very high **False Discovery Rate** (99%)
  - FP rate closer to 0.001% gives an FDR of 50%
    - Which for Security/Incident Response is maybe still too high

ML may be able to detect the signal, but perhaps not without too much noise.

Takeaways

- The “Gold Standard” for successfully using machine learning
  - Know your data – “extend a hypothesis”
  - Know the benefits and pitfalls of your algorithms
  - Be ready to iterate, rinse, and repeat



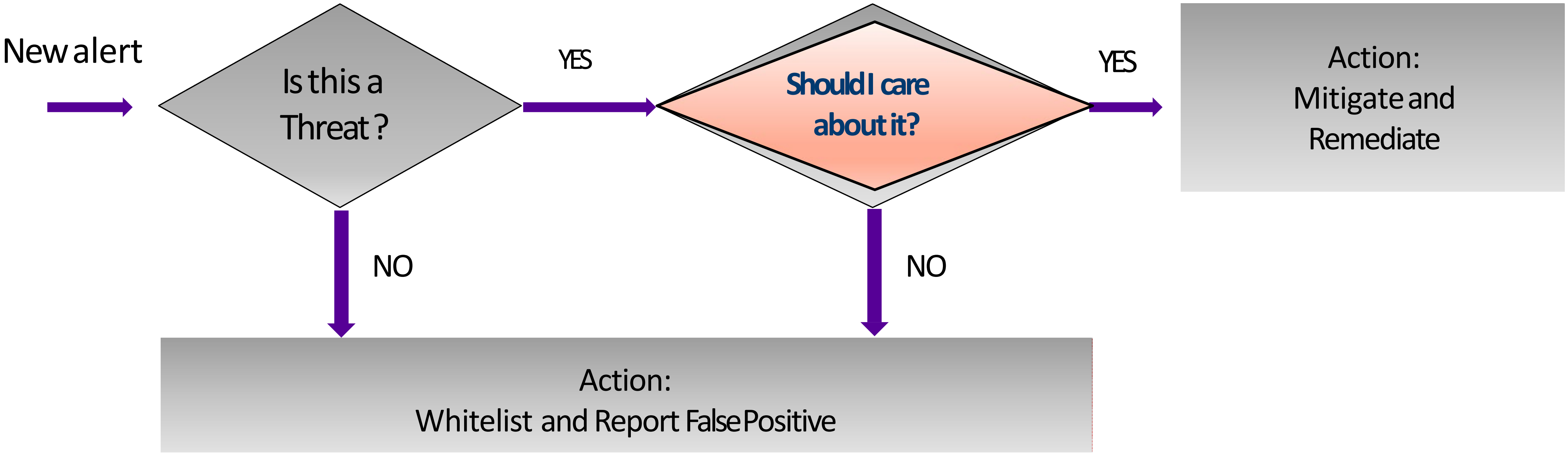
RSAConference2019

Automation of Incident Response

How AUTOMATION Can Help

- Collects, correlates and understands data from multiple sources to identify advanced threats.
- It continuously learns threat behaviors and automatically works with security tools to contain threats.
- Increases detection accuracy and provides security pros with better data with which to make decisions.

Typical Incident Response Process



You Should Care if Incident Risk is High

Goal: Better prioritization of effort

Intersect incident targets with asset values

E.g. Guest network activity vs. data center network anomaly

Factor in scope and progression context

How close to “Action on Objectives”

Has attack been disabled by other controls?

Prioritization of Effort

- Source, target, payload, etc.
- Threat vector – web, email, document, lateral spread
- Behavior – Trojan, reconnaissance, C&C, exfiltration
- Prioritized consolidated threat profiles for IR team
- Extract end-user information from active directory
- Allows incidents to be identified by username rather than IP address or DNS machine name

Attack Evidence, Scope and Progression

Collect malicious objects: files, PCAPS, network telemetry

- Needed to verify incident
- Needed to determine effective and appropriate mitigation

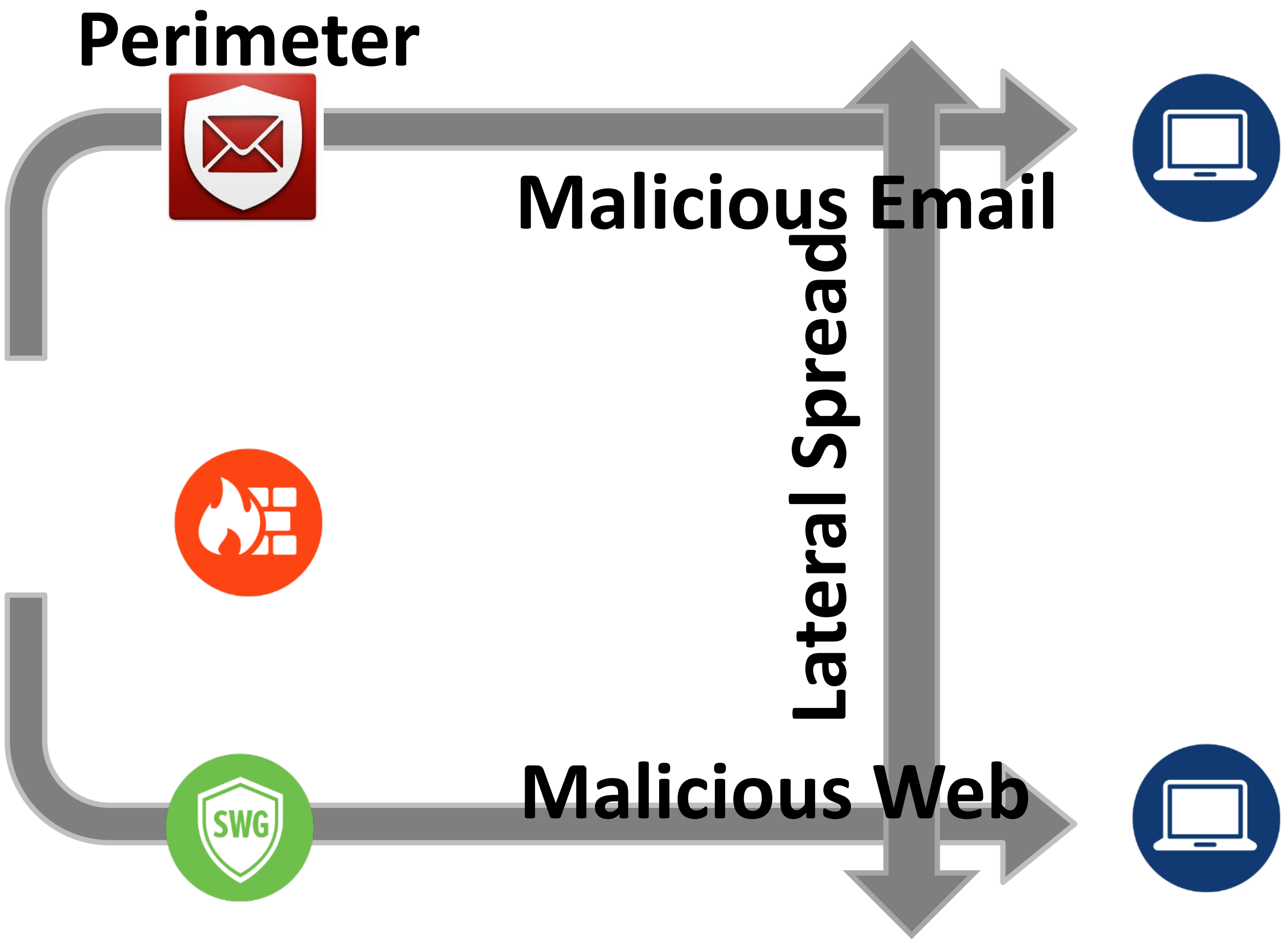
Attack Scope

- Which devices/users are affected?
- How long has attack been active?
  - Requires time series data normalized by resource extending back weeks, months, (years?)

Use Cases

Team	Use case	Question
Threat Intel hunters	Moving from big data to the endpoint to find infections	“Who got infected?”
Digital Forensics Incident Response (DFIR) hunters	Moving from infected endpoint backwards to big data to find root cause	“How they got hit?”

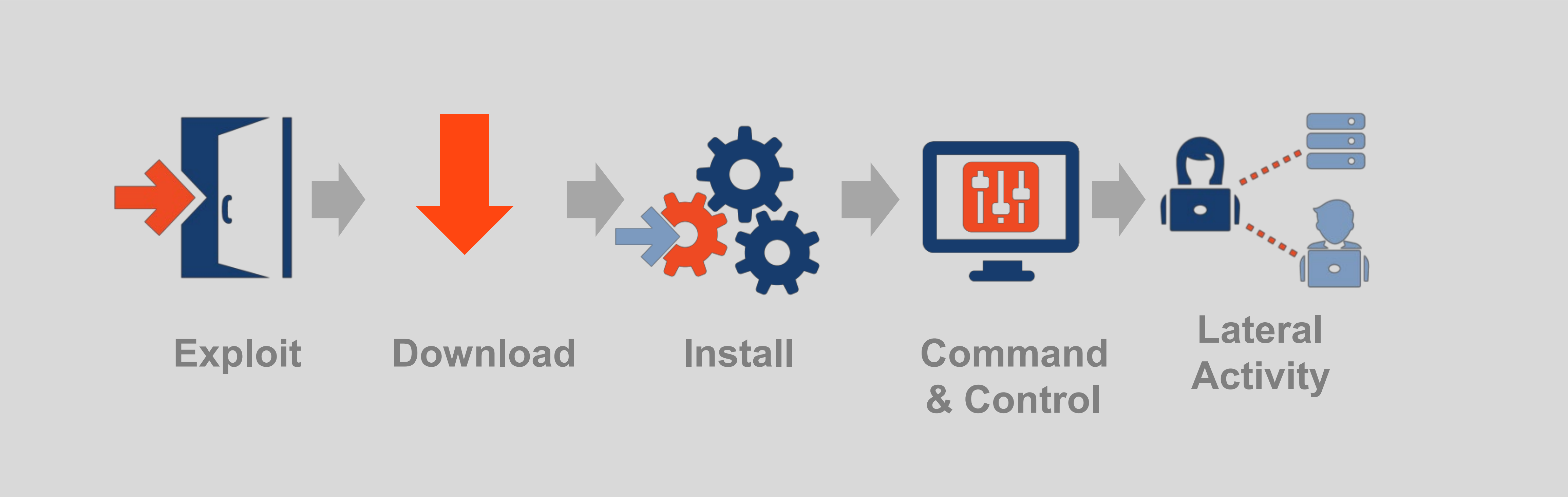
Primary Attack Vectors



Events from all primary attack vectors: Web, Email and Lateral spread

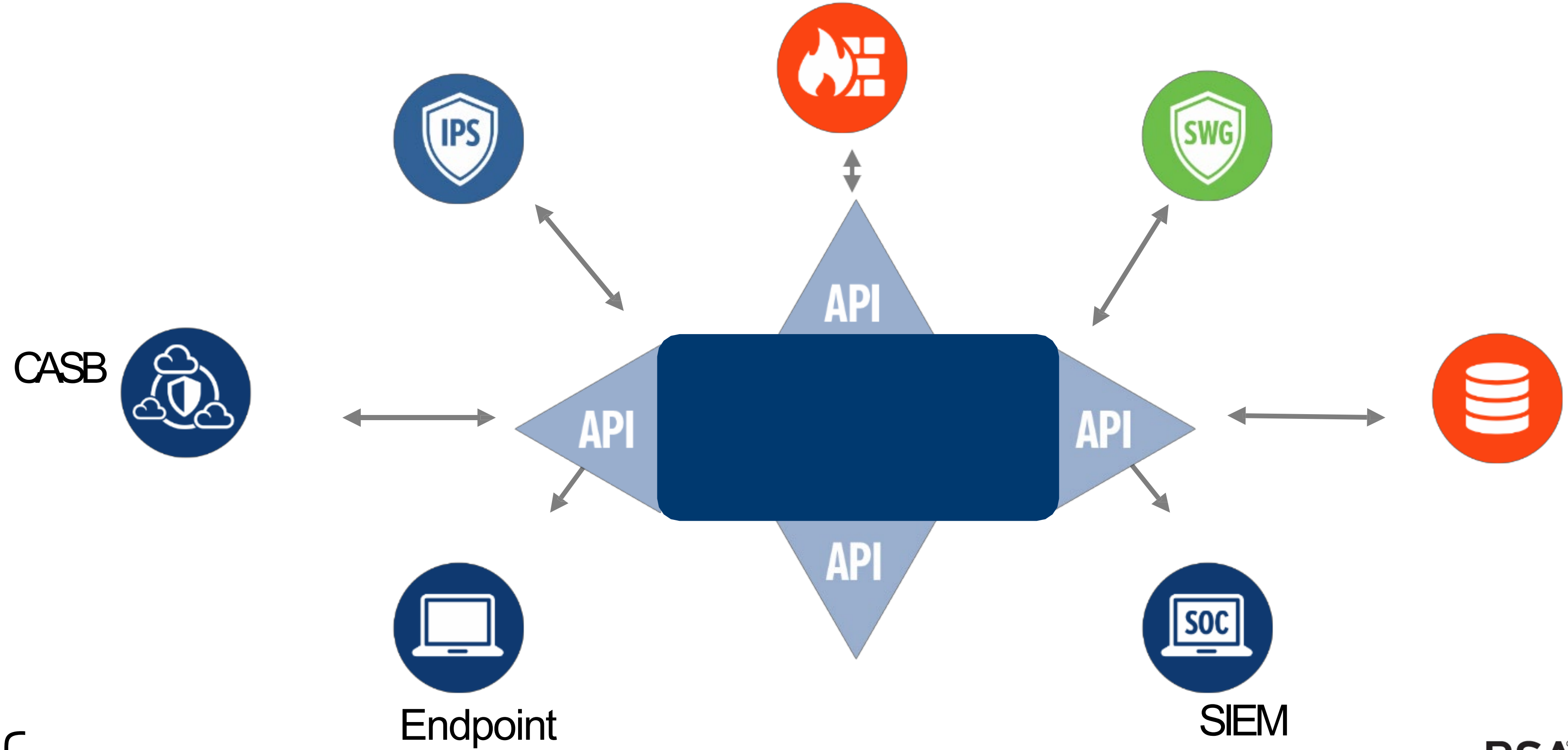
Killchain

Events span all parts of the killchain

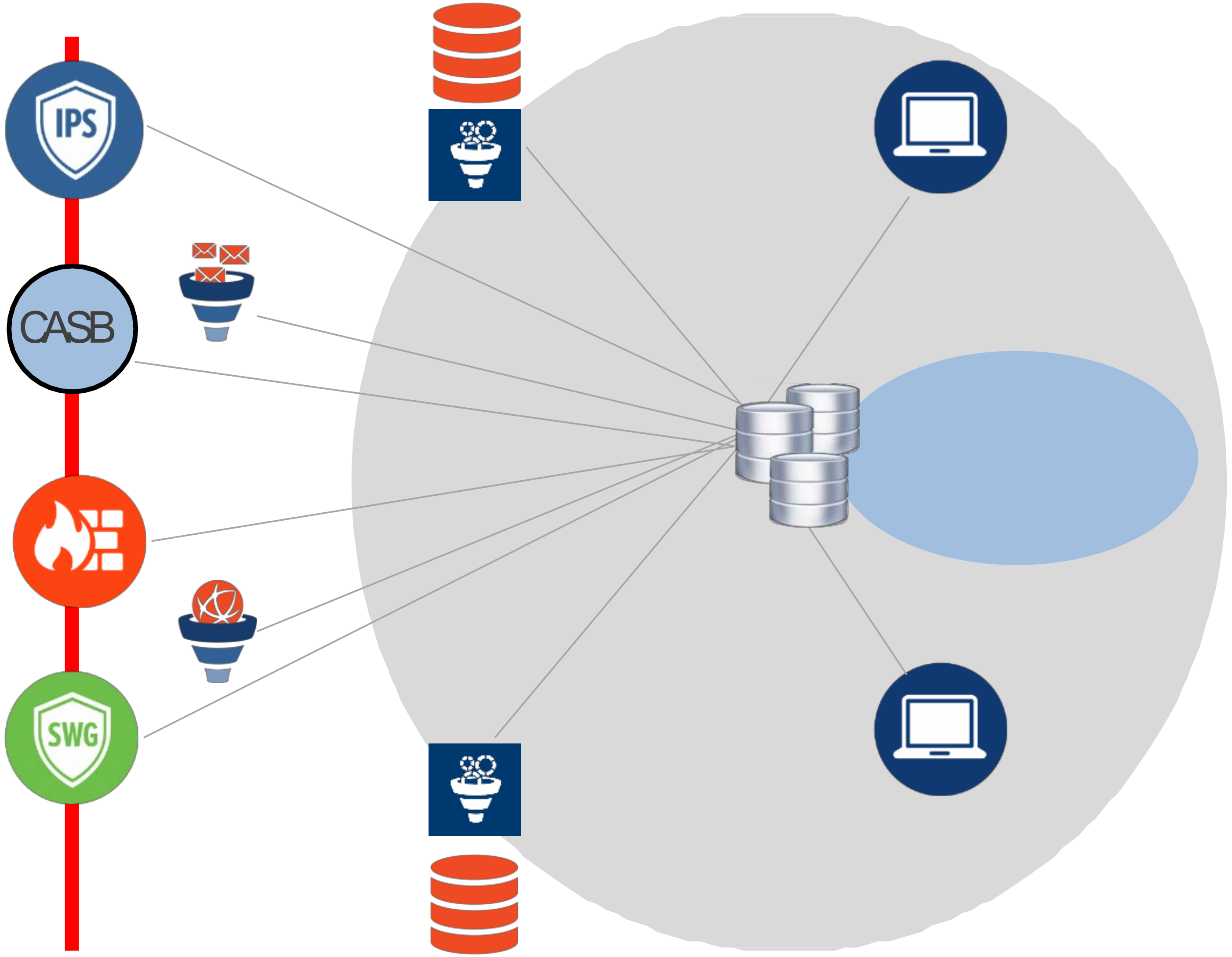


Open APIs

Automation solutions should rely on Open APIs to enable information exchange with other vendors



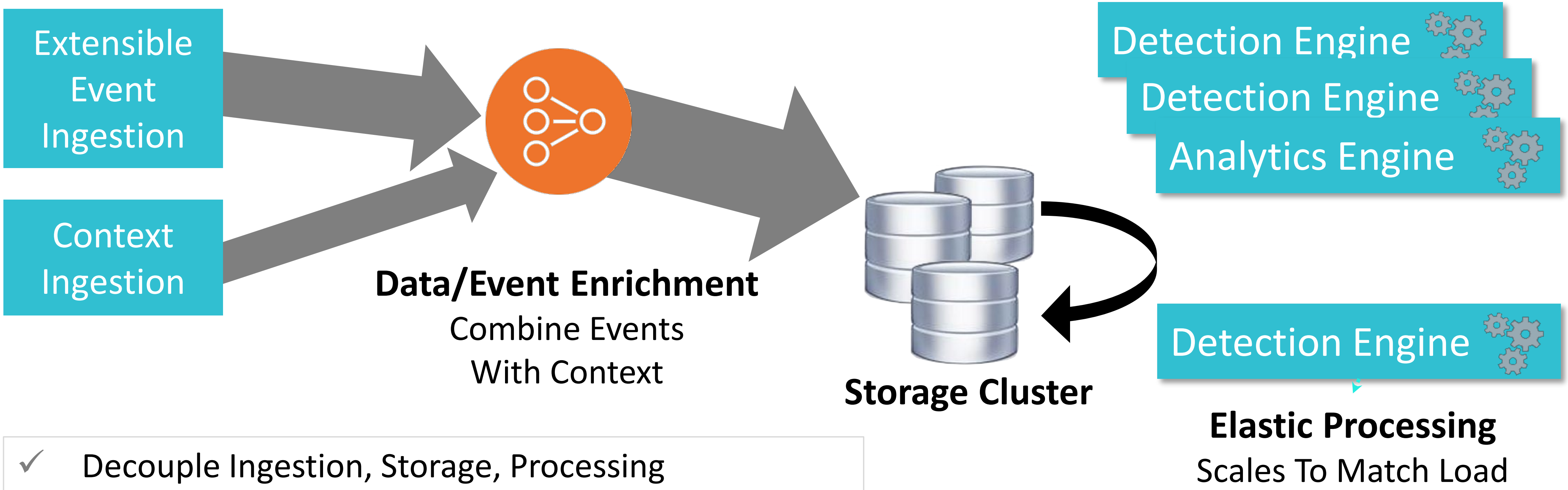
Incident Response Tasks



- Collect data from web, email, etc.
- Analyze/detect advanced threat
- Identify infected host/user
- Ingest meta data from all sources
- Correlate all related host events
- Consolidate events on timeline
- Present as one security incident

- Reduces noise from SIEM alerts
- Eliminates manual correlation
- Provides insight into threat
- Simplifies incident response

Architecture



- ✓ Decouple Ingestion, Storage, Processing
- ✓ Collect raw data for detection, not just logs
- ✓ Add Endpoint Identity to all data
- ✓ Extend to arbitrary time horizon
- ✓ Elastic Detection processing

Automation of Common IR Tasks

Malware Investigation Tasks	Manual Effort Time
Identify Host and User	10 min
Collect AV and EDTR data for given host	25 min
Collect network data (NGFW, SWG)	25 min
Analyze & Correlate	35 min
Determine progression and scope	15 min
Contain the threat	10 min
TOTAL TIME	2 hours

Automation in Action

Investigation Task	Using Automation	Manual Process
Chasing False Positives	38 hours	390 hours
Post-breach Mitigation	37 hours	195 hours
Investigating Breach Indicators	55 hours	177 hours
Total time taken	130 hours	722 hours

Automation gives ~80% Time Savings over Manual Processes

Reducing Cybersecurity Costs & Risks Through Automation Technologies, November 2017

Remember

- Attackers are embracing automation, so should we.
- When done right, machine learning maximizes threat detection.
- Need a human expert to interpret machine learning results.
- Accelerate incident response by <u>automating</u> common incident response tasks.

RSA Conference 2019

Questions

Email:mikolab@juniper.net

Twitter:@belogor