# Cybersecurity Data Science as a Process

## Practitioner Insights and Best Practices

DATA
DETECTION
DISCOVERY

John Hopkins IACD cyber conference
Baltimore
May 2 – 3, 2019

https://www.iacdautomate.org/may-2019-integrated-cyber

## Patrick Alcorn

Data Scientist Cybersecurity

## John Stultz

Cyber Analytic Platform Architect

sas

40+ Years of BUSINESS ANALYTICS

14,000 SAS employees worldwide

93 of the top 100 companies on the FORTUNE® GLOBAL 500 LIST

#1 World's LARGEST privately held software company

80,000+ Customer sites in 148 countries

23% Annual reinvestment in R&D

# Cybersecurity Context

# Evolving Threats

Internal Threats

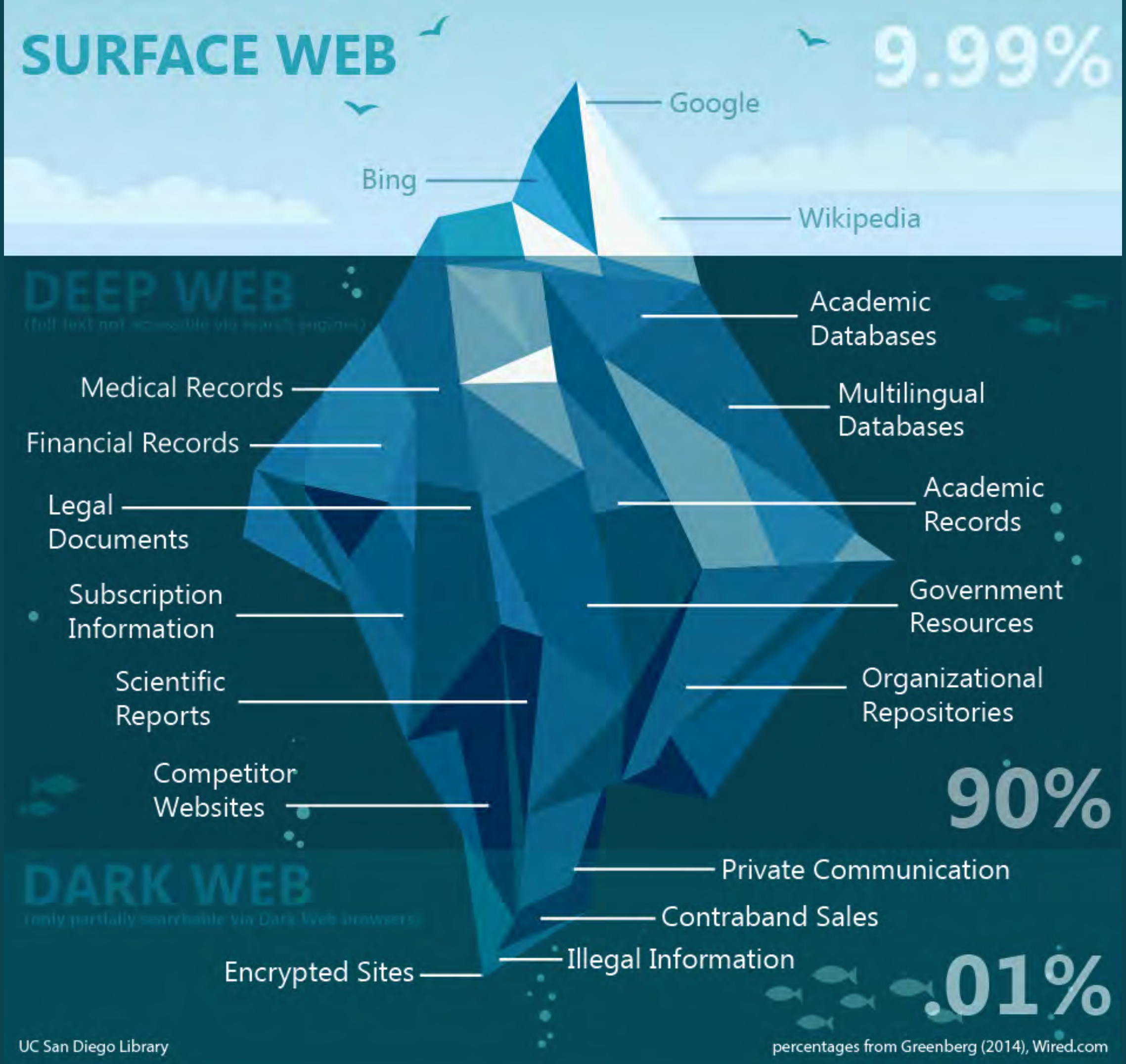
Automated Attacks

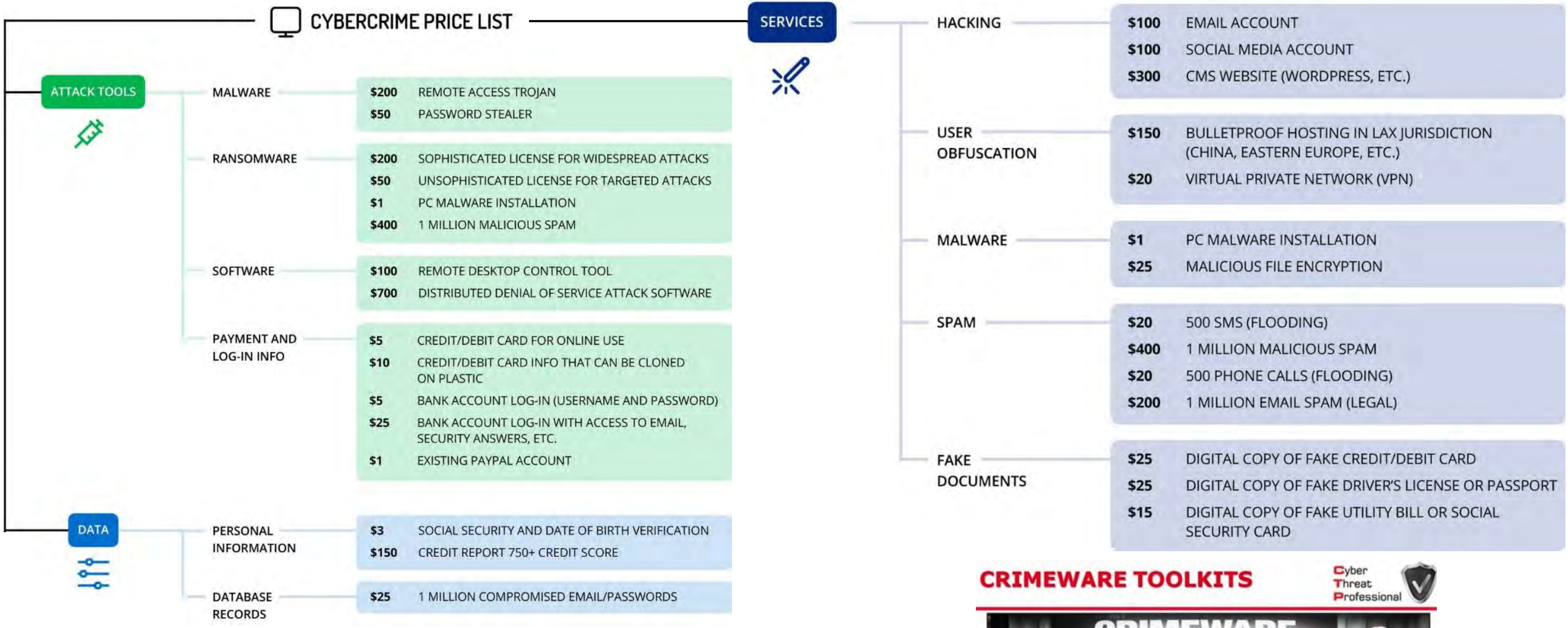
Social Engineering

Fraud-Cyber Hybrids

Ransomware & Cryptojacking

State Actors


SURFACE WEB
9.99%
Google
Bing
Wikipedia
DEEP WEB
(full text not accessible via search engines)
Academic Databases
Medical Records
Multilingual Databases
Financial Records
Academic Records
Legal Documents
Government Resources
Subscription Information
Scientific Reports
Organizational Repositories
Competitor Websites
90%
DARK WEB
(only partially searchable via Dark Web browsers)
Private Communication
Contraband Sales
Illegal Information
Encrypted Sites
.01%
UC San Diego Library
percentages from Greenberg (2014), Wired.com



## CYBERCRIME PRICE LIST

### ATTACK TOOLS

| | | |
|---|---|---|
| MALWARE | $200 | REMOTE ACCESS TROJAN |
| | $50 | PASSWORD STEALER |
| RANSOMWARE | $200 | SOPHISTICATED LICENSE FOR WIDESPREAD ATTACKS |
| | $50 | UNSOPHISTICATED LICENSE FOR TARGETED ATTACKS |
| | $1 | PC MALWARE INSTALLATION |
| | $400 | 1 MILLION MALICIOUS SPAM |
| SOFTWARE | $100 | REMOTE DESKTOP CONTROL TOOL |
| | $700 | DISTRIBUTED DENIAL OF SERVICE ATTACK SOFTWARE |
| PAYMENT AND LOG-IN INFO | $5 | CREDIT/DEBIT CARD FOR ONLINE USE |
| | $10 | CREDIT/DEBIT CARD INFO THAT CAN BE CLONED ON PLASTIC |
| | $5 | BANK ACCOUNT LOG-IN (USERNAME AND PASSWORD) |
| | $25 | BANK ACCOUNT LOG-IN WITH ACCESS TO EMAIL, SECURITY ANSWERS, ETC. |
| | $1 | EXISTING PAYPAL ACCOUNT |

### DATA

| | | |
|---|---|---|
| PERSONAL INFORMATION | $3 | SOCIAL SECURITY AND DATE OF BIRTH VERIFICATION |
| | $150 | CREDIT REPORT 750+ CREDIT SCORE |
| DATABASE RECORDS | $25 | 1 MILLION COMPROMISED EMAIL/PASSWORDS |

### SERVICES

| | | |
|---|---|---|
| HACKING | $100 | EMAIL ACCOUNT |
| | $100 | SOCIAL MEDIA ACCOUNT |
| | $300 | CMS WEBSITE (WORDPRESS, ETC.) |
| USER OBFUSCATION | $150 | BULLETPROOF HOSTING IN LAX JURISDICTION (CHINA, EASTERN EUROPE, ETC.) |
| | $20 | VIRTUAL PRIVATE NETWORK (VPN) |
| MALWARE | $1 | PC MALWARE INSTALLATION |
| | $25 | MALICIOUS FILE ENCRYPTION |
| SPAM | $20 | 500 SMS (FLOODING) |
| | $400 | 1 MILLION MALICIOUS SPAM |
| | $20 | 500 PHONE CALLS (FLOODING) |
| | $200 | 1 MILLION EMAIL SPAM (LEGAL) |
| FAKE DOCUMENTS | $25 | DIGITAL COPY OF FAKE CREDIT/DEBIT CARD |
| | $25 | DIGITAL COPY OF FAKE DRIVER'S LICENSE OR PASSPORT |
| | $15 | DIGITAL COPY OF FAKE UTILITY BILL OR SOCIAL SECURITY CARD |

Source: Recorded Future via Fortune Magazine 'A Hacker's Tool Kit'
http://fortune.com/2017/10/25/cybercrime-spyware-marketplace/

# Security Operations Center (SOC)

# Emerging SOC Operational Drivers

**Big & fast streaming data** needs to be stitched into ‘smart data’

Limitations of traditional signature and rules-based approaches, **requiring probabilistic and risk-focused models**

**Integrated situational awareness** of network, device, and user behavior while **reducing false alerts**

Need to build and validate efficacious **machine learning models**

**Automation of manual investigation** and remediation processes

# Cybersecurity Data Science (CSDS)

# Level of difficulty in reducing false alerts*

Ponemon INSTITUTE

**When Seconds Count: How Security Analytics Improves Cybersecurity Defenses**

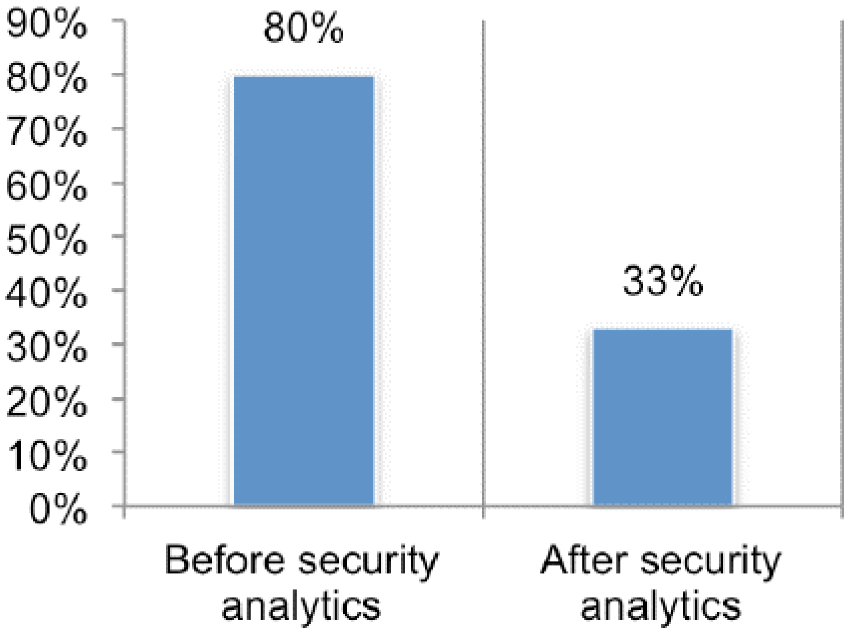
**Sponsored by SAS Institute**

Independently conducted by Ponemon Institute LLC

Publication Date: January 2017

Ponemon Institute© Research Report

https://www.sas.com/en_us/whitepapers/ponemon-how-security-analytics-improves-cybersecurity-defenses-108679.html



** Survey of 621 global IT security practitioners*

# CSDS: Cybersecurity Data Science

Replacing rules with machine learning to reduce false alerts

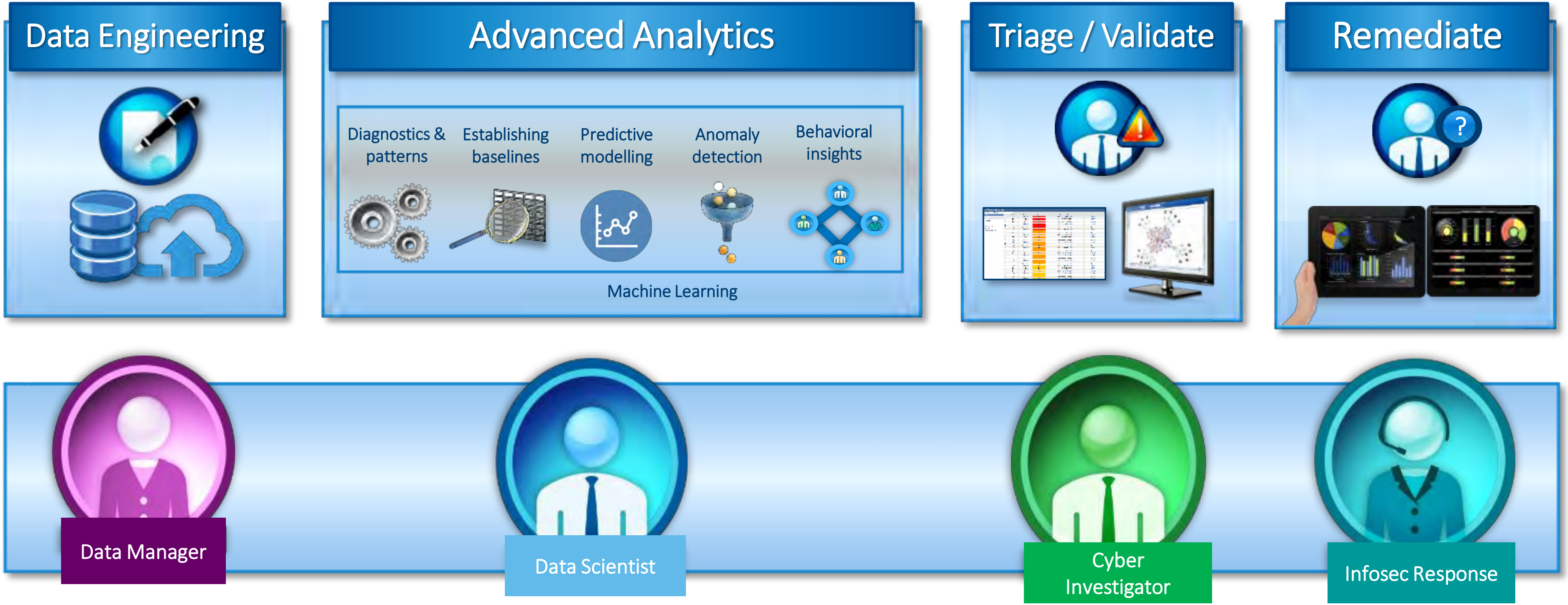
Moving to real time detection and decisioning

Automation of manual processes and routine decisions

Data engineering to structured and integrate distributed big data into ‘smart data’

Investigation tools that visualize complexity to improve investigator efficiency and decision making

# Cybersecurity Analytics as-a-Process



Data Engineering

Advanced Analytics

Diagnostics & patterns

Establishing baselines

Predictive modelling

Anomaly detection

Behavioral insights

Machine Learning

Triage / Validate

Remediate

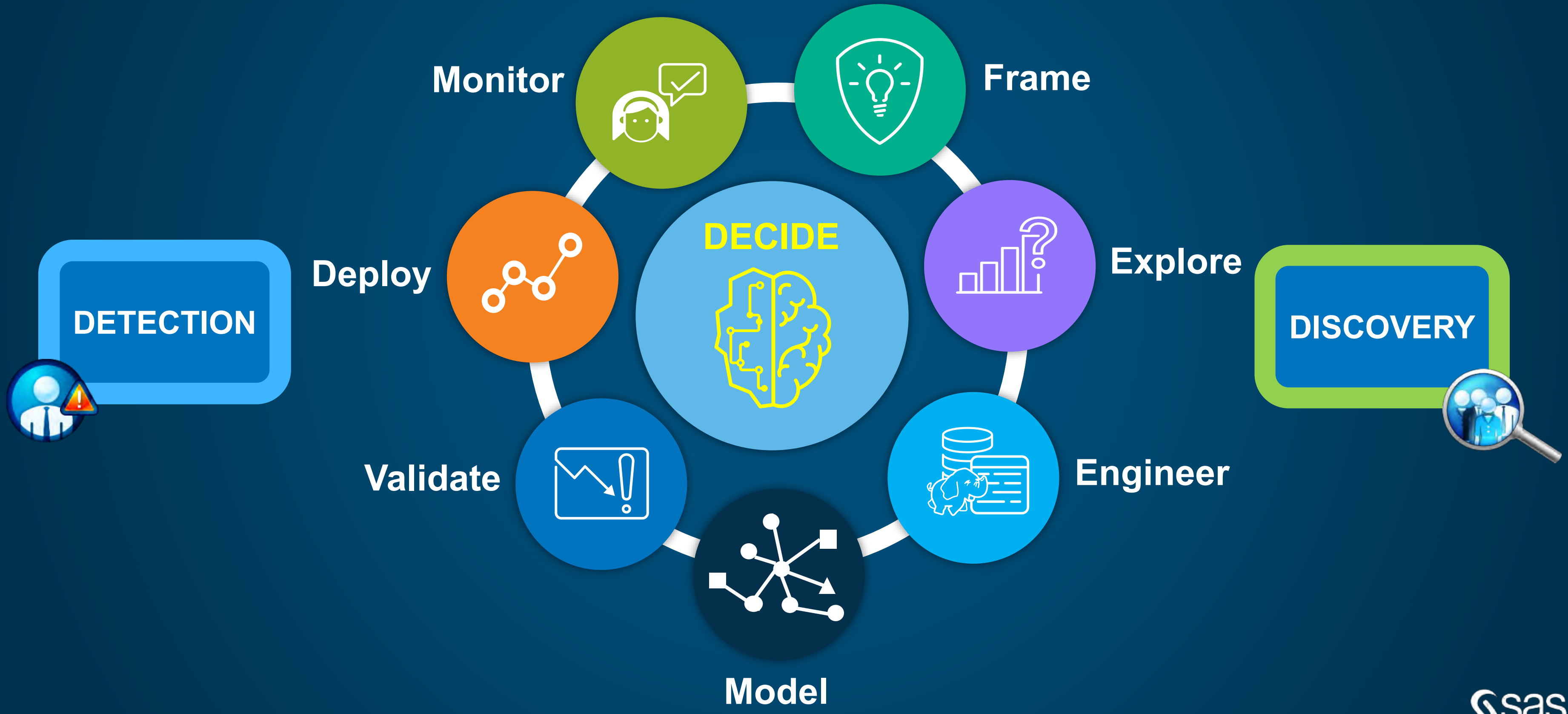
Data Manager

Data Scientist

Cyber Investigator

Infosec Response

# Cybersecurity Data Science (CSDS) Lifecycle



DISCOVERY

Frame

Explore

Engineer

Model

Validate

Deploy

Monitor

DECIDE

DETECTION

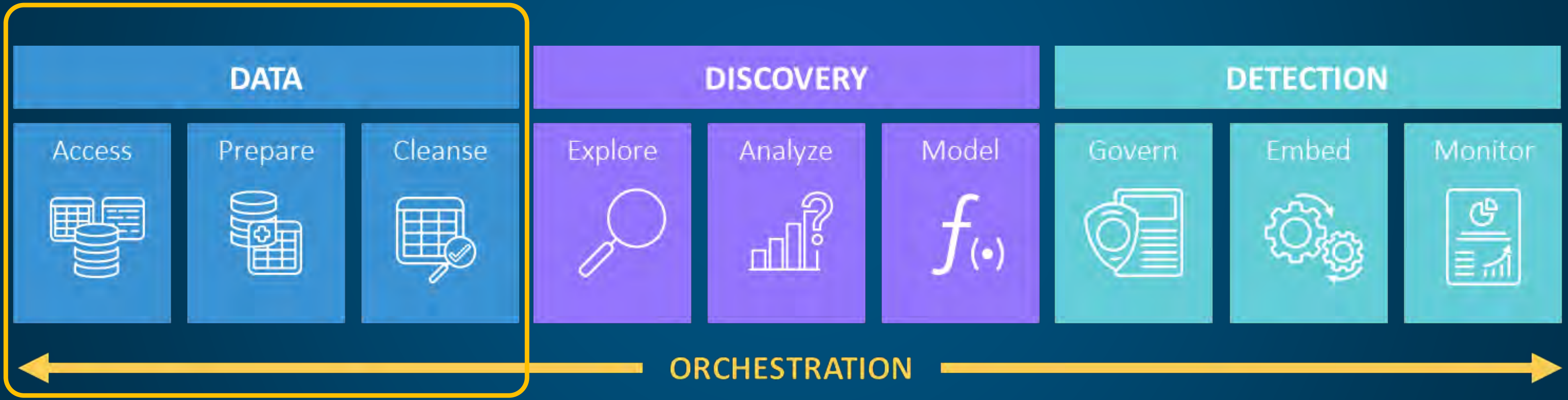
DATA
DETECTION
DISCOVERY

# CSDS Process

# CSDS Data

# CSDS Process

## Unified Orchestration



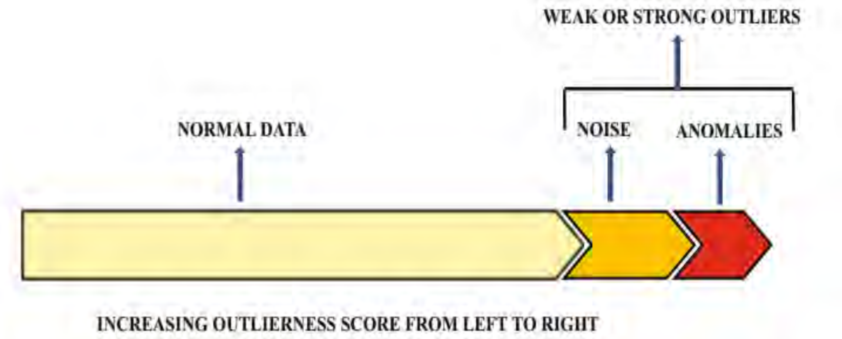
# The devil is in the data

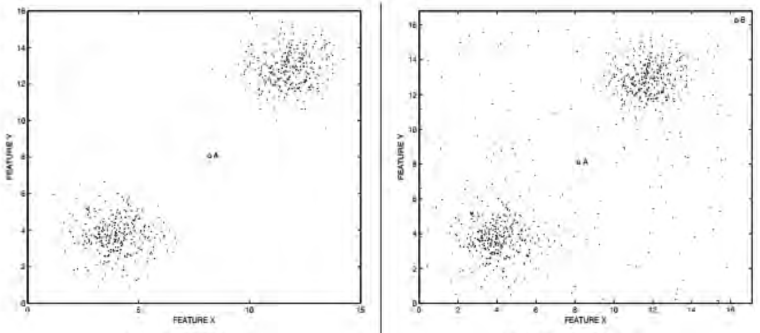
010001011010001000101100000001101010111100010000110001
100000110100101000101100100111011100101001000001100100
00001101101100110011011100110010101111😈000100001000011
010101101000100010010001000100101011011101101010011101
1110000100110011001001111000110101111100001001100100001
1110110100101😈0111000100101000110100110010011001010010
010111110100100011101000100011010100111011000101100111
110001001011001000100110100011010101101000010000100001
010001011010001000010000000😈01101010101100001010010010
101010110100010001001111000110101011010100100001000001
11100100😈11001000110000100011010101111001010000111001
0111000110100011100001011111100001110011😈001110101001
101001011000001000110000100011010101111000111100100010
0101011010111010001010001000100101010111100001000000011
011101000001100100100011001001010110111011010100110101

# Hidden threats want to remain hidden (in the data)

# Anomaly Detection:  Simply Complex

## Identifying focused anomalies amongst an ocean of noise…





(a) No noise

(b) With noise

*SOURCE*  Aggarwal, Charu C. (2017). “Outlier Analysis: Second Edition”. Springer International Publishing AG.

# Challenges preventing successful use of cybersecurity analytics*

Ponemon INSTITUTE

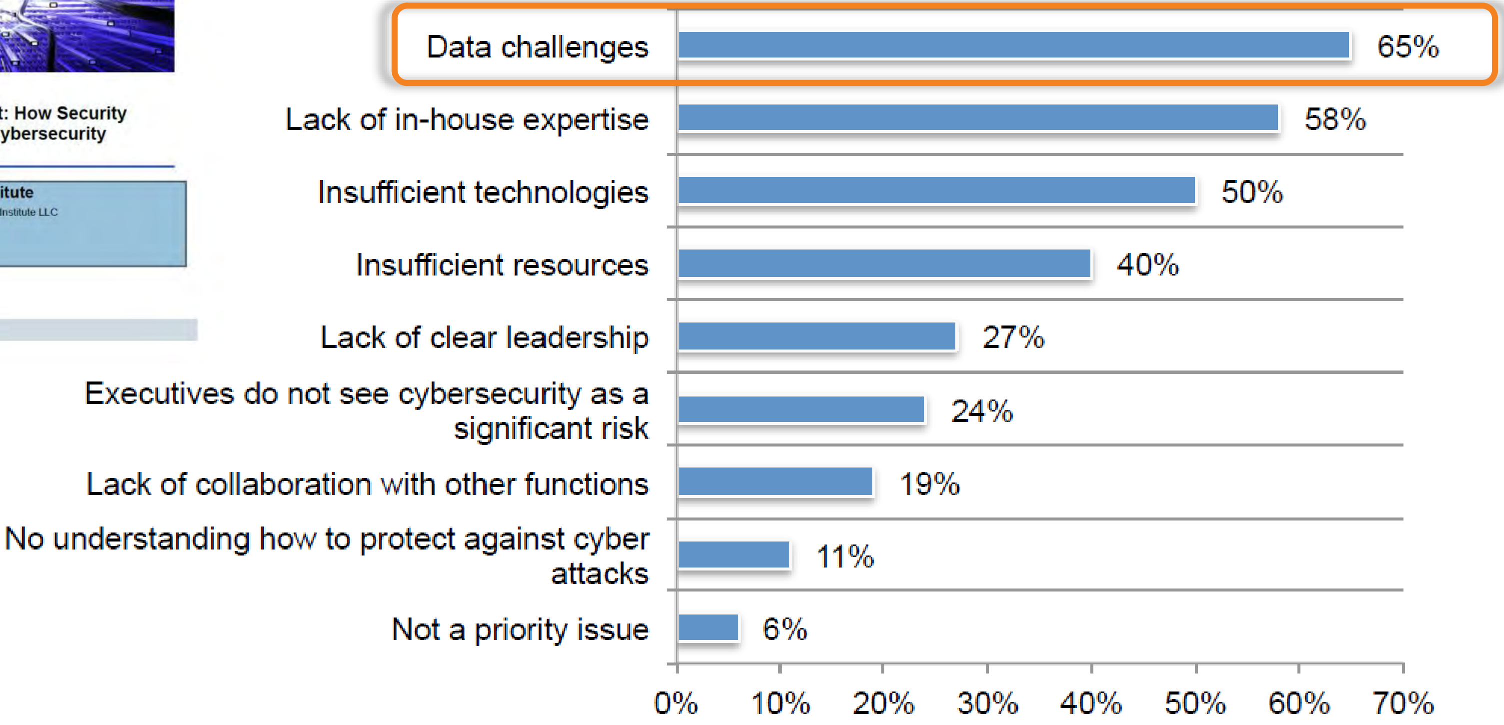
When Seconds Count: How Security Analytics Improves Cybersecurity Defenses

Sponsored by SAS Institute

Independently conducted by Ponemon Institute LLC

Publication Date: January 2017

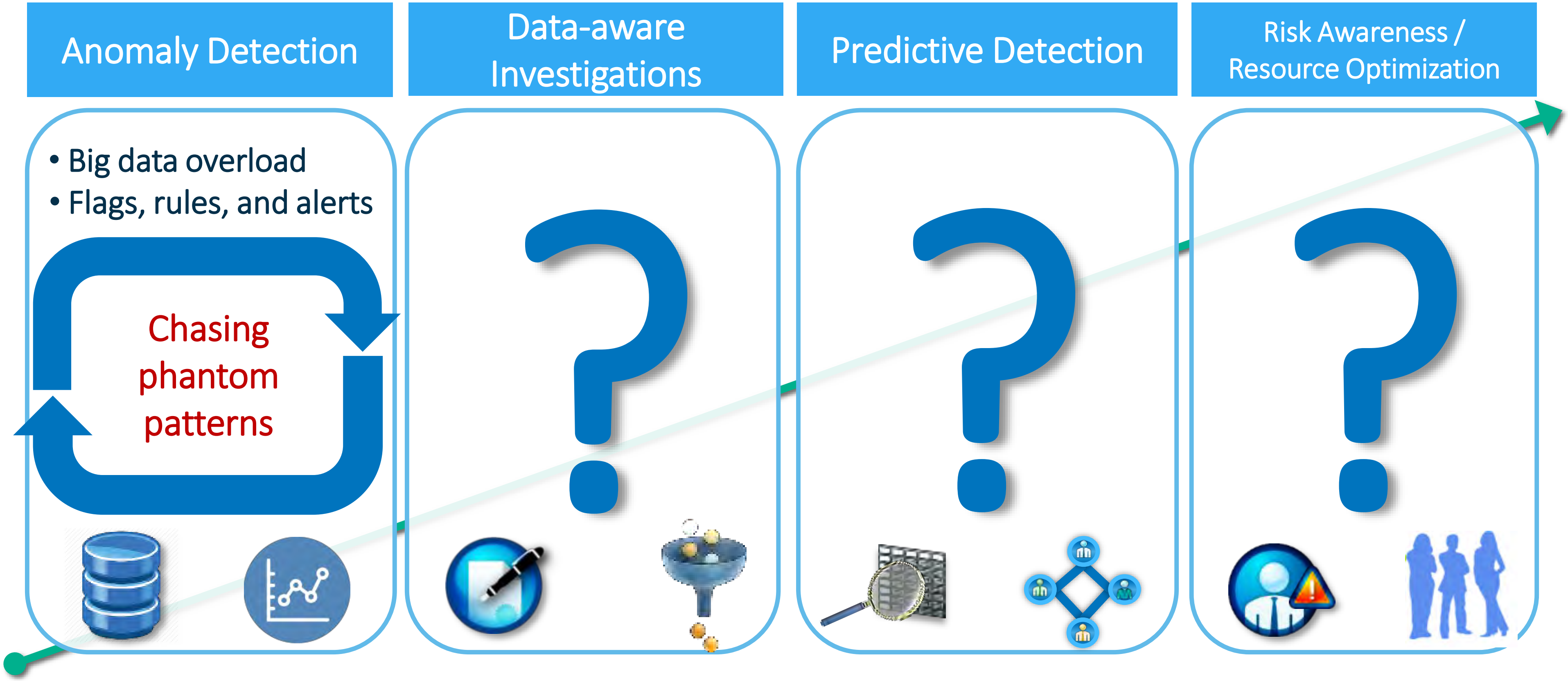
Ponemon Institute© Research Report



https://www.sas.com/en_us/whitepapers/ponemon-how-security-analytics-improves-cybersecurity-defenses-108679.html

** Survey of 621 global IT security practitioners*

# Cybersecurity Analytics Maturity Curve


Anomaly Detection
Data-aware Investigations
Predictive Detection
Risk Awareness / Resource Optimization
• Big data overload
• Flags, rules, and alerts
Chasing phantom patterns
?
?
?

SC Security Brief

MAGAZINE

Nearly two-third of respondents to a new SC Magazine survey said they were in the process of implementing security analytics.

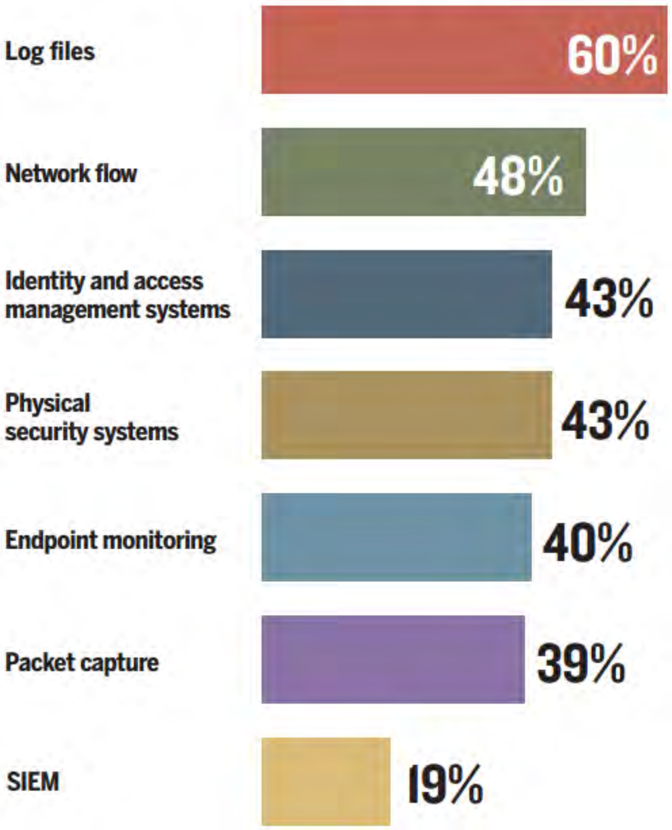
Analyze this!

Sponsored by sas

*SOURCE*

Security Brief Magazine. (2016). “Analyze This! Who’s Implementing Security Analytics Now?” Available at https://www.sas.com/en_th/whitepapers/analyze-this-108217.html

## What data sources are available within your organization, should a security analytics program happen?

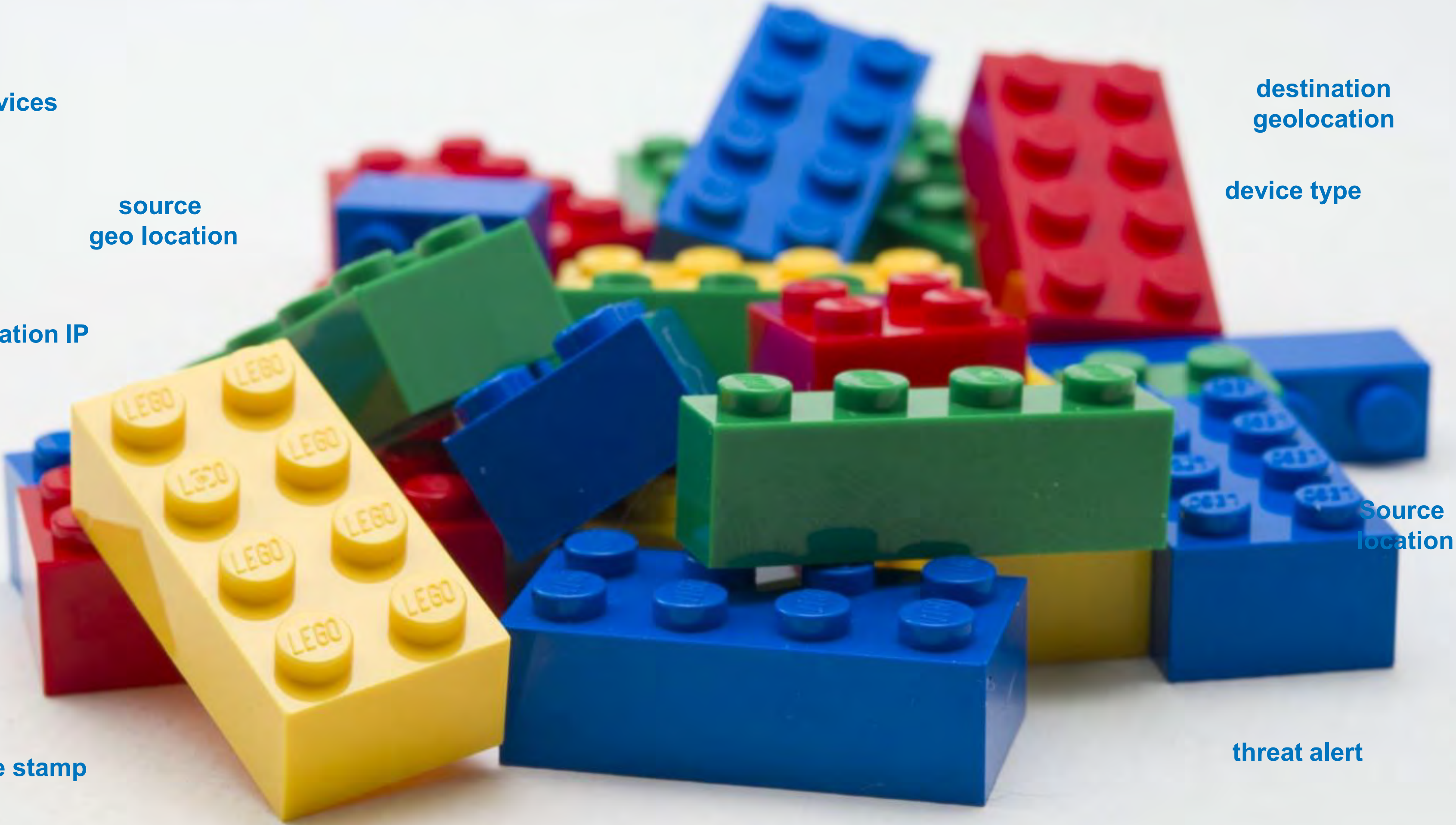


IP address

time stamp

userid

destination port



devices

destination geolocation

device type

source geo location

destination IP

Source location

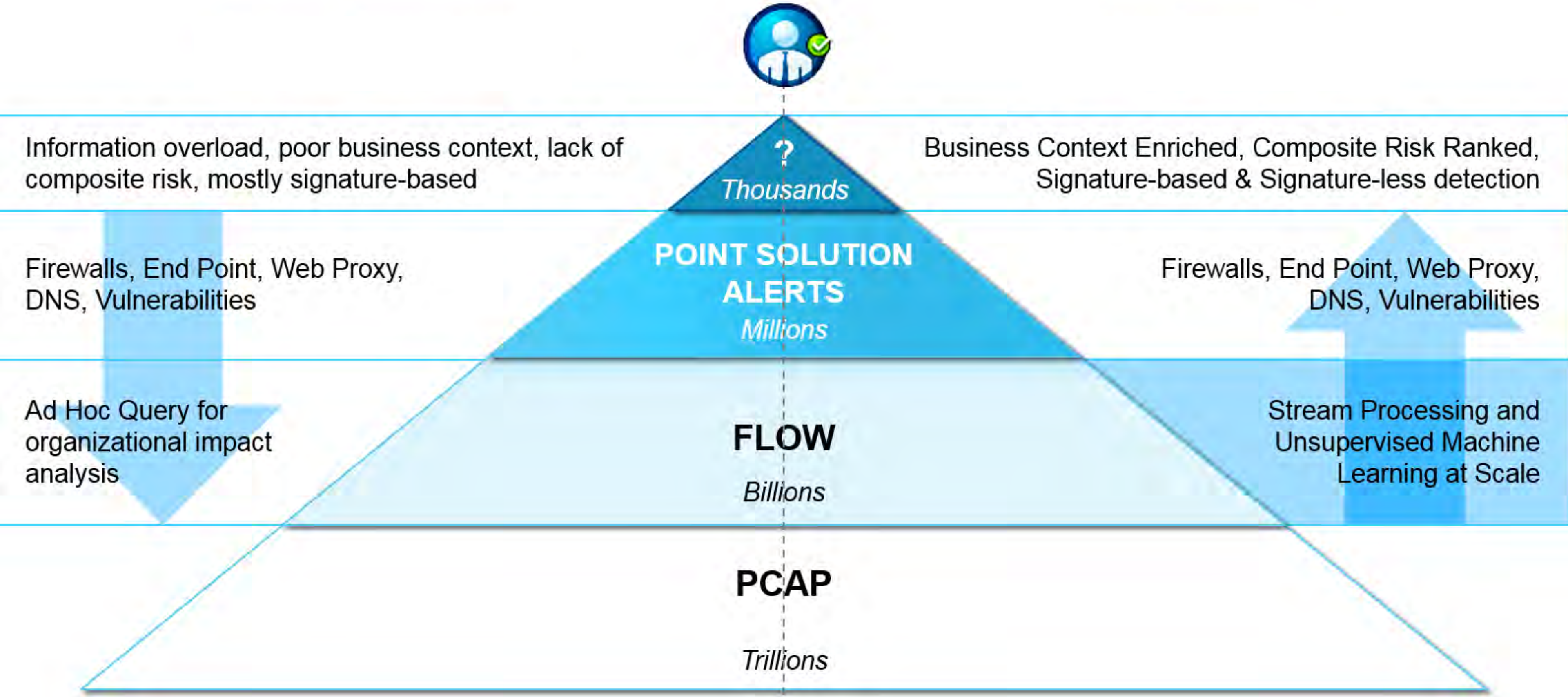
date stamp

threat alert

destination port

IP address

# Security Data Management Challenge: Speed and Volumes



Information overload, poor business context, lack of composite risk, mostly signature-based

Firewalls, End Point, Web Proxy, DNS, Vulnerabilities

Ad Hoc Query for organizational impact analysis

?

*Thousands*

POINT SOLUTION ALERTS

*Millions*

FLOW

*Billions*

PCAP

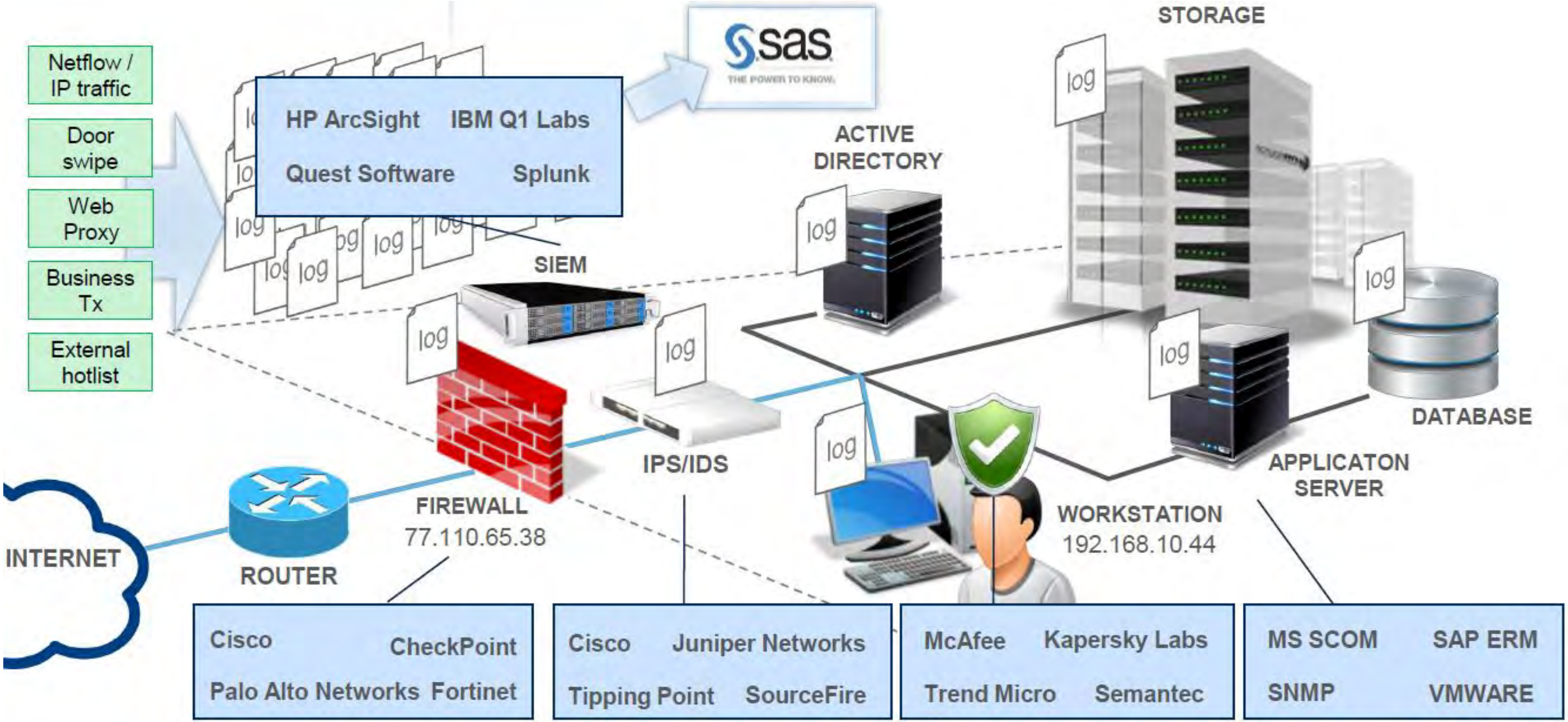
*Trillions*

Business Context Enriched, Composite Risk Ranked, Signature-based & Signature-less detection

Firewalls, End Point, Web Proxy, DNS, Vulnerabilities

Stream Processing and Unsupervised Machine Learning at Scale

# Many data sources... increasing data volumes



Netflow / IP traffic

Door swipe

Web Proxy

Business Tx

External hotlist

HP ArcSight IBM Q1 Labs

Quest Software Splunk

SIEM

ACTIVE DIRECTORY

STORAGE

DATABASE

APPLICATON SERVER

WORKSTATION 192.168.10.44

IPS/IDS

FIREWALL 77.110.65.38

ROUTER

INTERNET

Cisco CheckPoint

Palo Alto Networks Fortinet

Cisco Juniper Networks

Tipping Point SourceFire

McAfee Kapersky Labs

Trend Micro Semantec

MS SCOM SAP ERM

SNMP VMWARE

# High false alerts... slow investigation processes

LACK OF CONTEXT

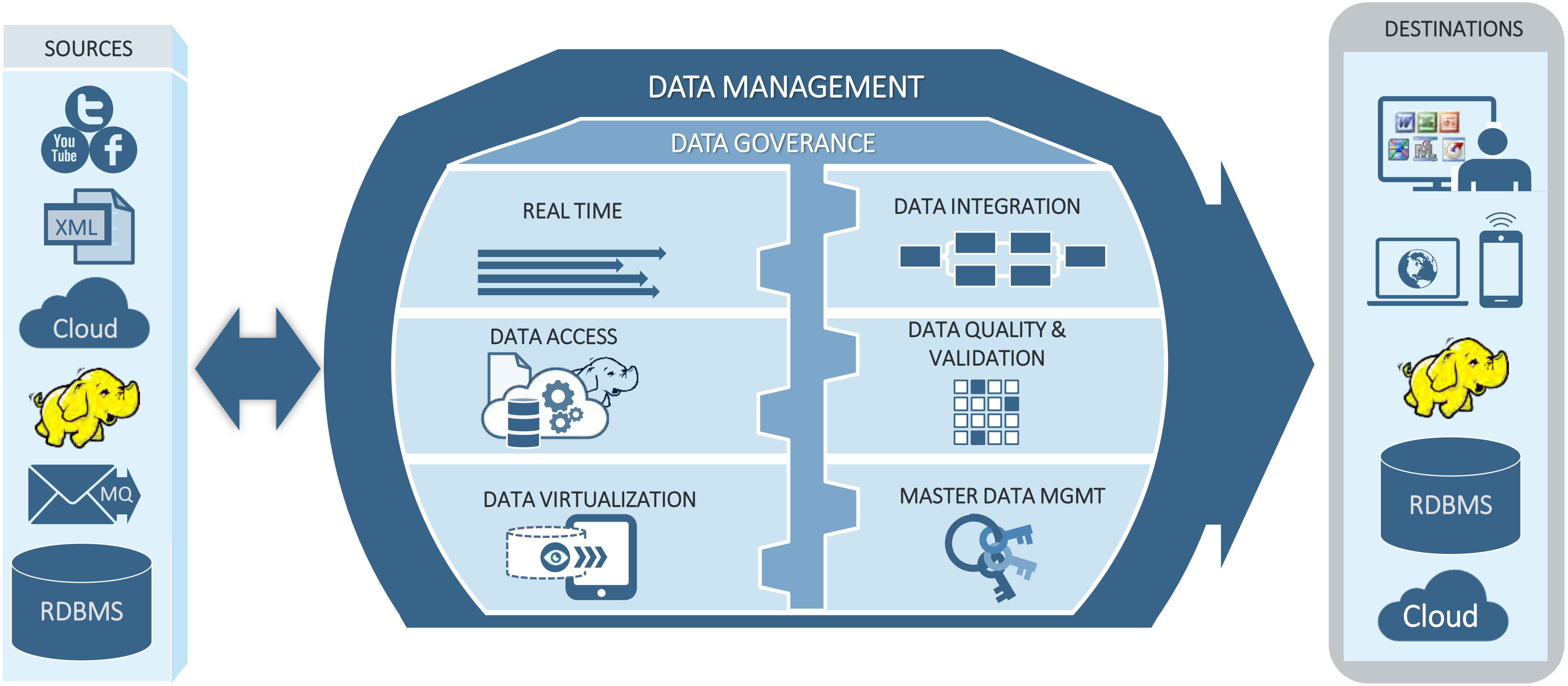
DISCONNECTED & FRAGMENTED

UNVALIDATED ALERTS

VOLUME & SPEED

MULTIPLE SYSTEMS

# Data Engineering: Fusion, Quality and Delivery



"Organizing data is a critical first step in figuring out what data means"

Larry Alton, Information Management Feb 14th, 2019

# Cybersecurity Frameworks & Ontologies

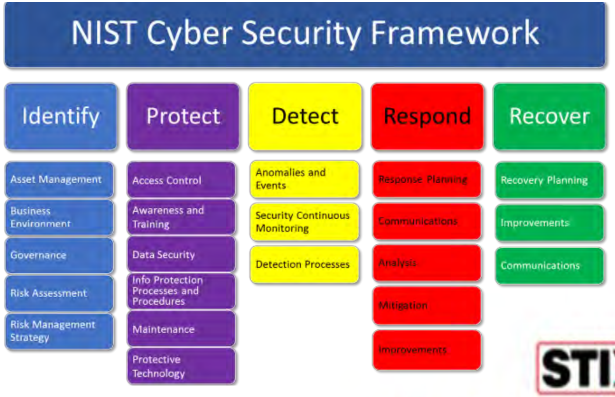
## FRAMEWORKS

- MITRE Cyber Observable eXpression
- NIST Cybersecurity Framework
- Intrusion Kill Chain (Lockheed Martin)

## ONTOLOGIES

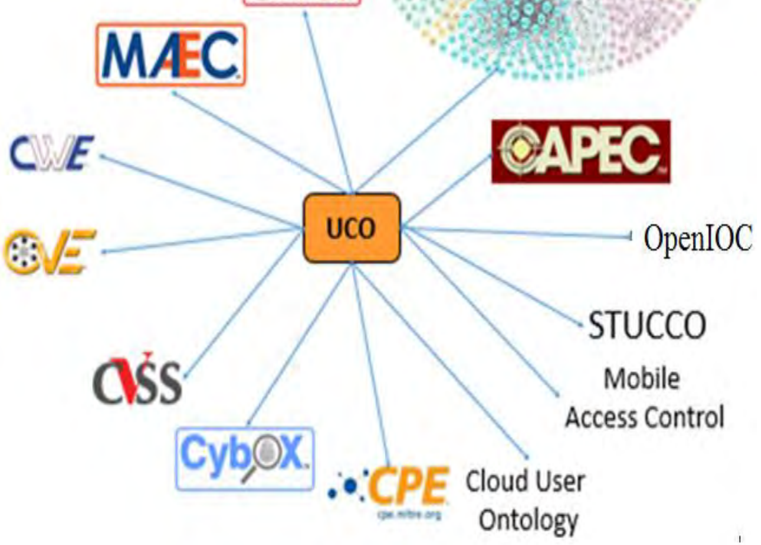
- **DFAX** Digital Forensic Analysis eXpression
- **CVE** Cyber Intelligence Ontology
- **ICAS** Information Security (example)
- **UCO / UCO (OWL)**

  Unified Cybersecurity Ontology

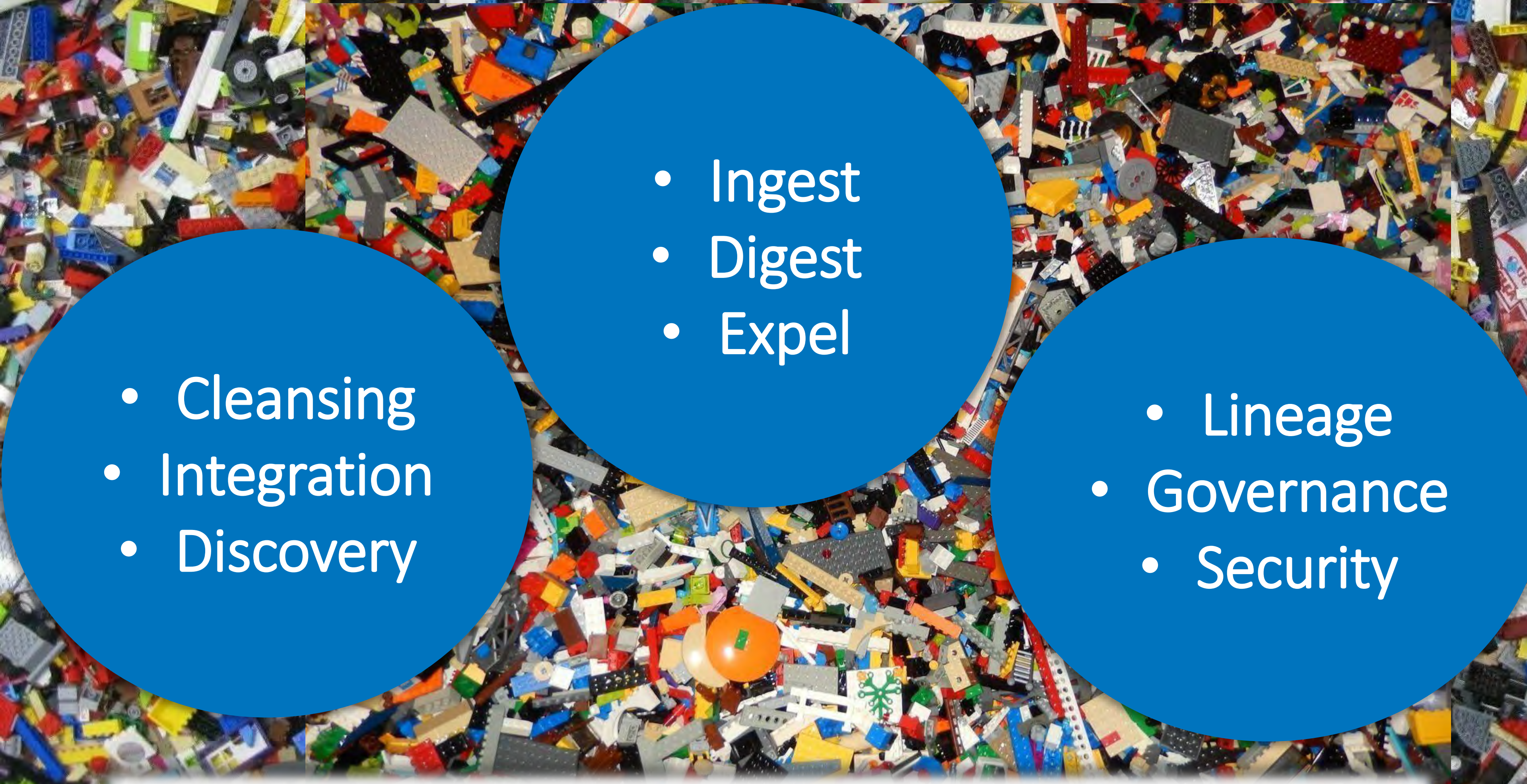


NIST Cyber Security Framework

| Identify | Protect | Detect | Respond | Recover |
|---|---|---|---|---|
| Asset Management | Access Control | Anomalies and Events | Response Planning | Recovery Planning |
| Business Environment | Awareness and Training | Security Continuous Monitoring | Communications | Improvements |
| Governance | Data Security | Detection Processes | Analysis | Communications |
| Risk Assessment | Info Protection Processes and Procedures | | Mitigation | |
| Risk Management Strategy | Maintenance | | Improvements | |
| | Protective Technology | | | |



POLICE



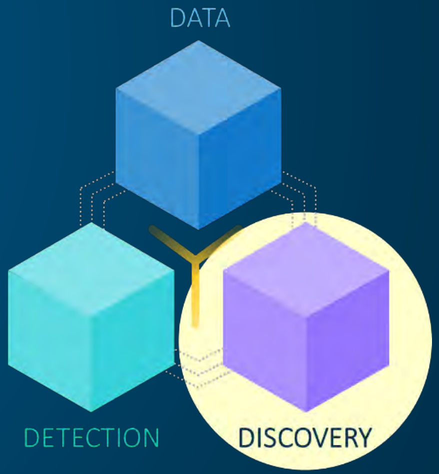
- Ingest
- Digest
- Expel

- Cleansing
- Integration
- Discovery

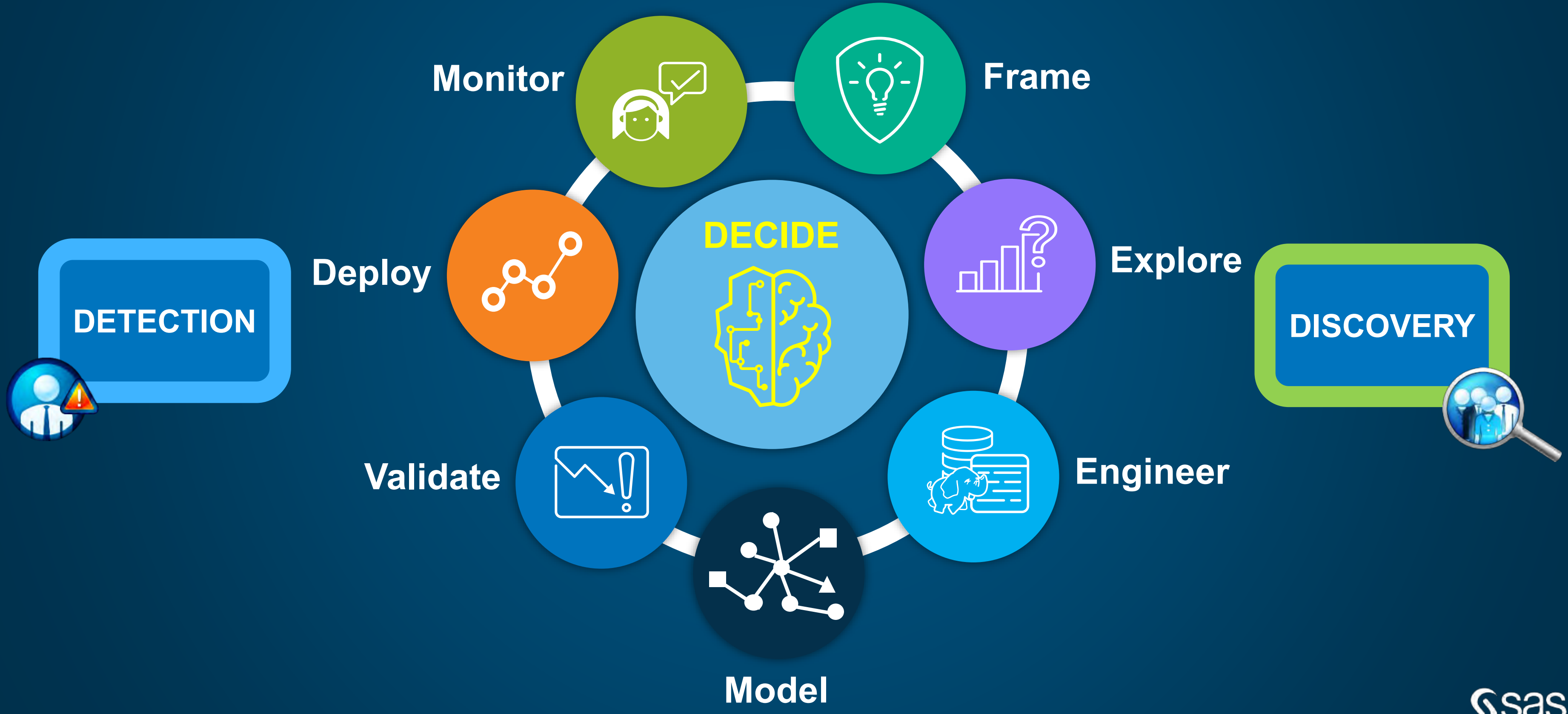
- Lineage
- Governance
- Security

[Whitepaper: A Comprehensive Approach to Big Data Governance, Data Management and Analytics]()

# CSDS Discovery



# Cybersecurity Data Science (CSDS) Lifecycle



DECIDE

Frame

Explore

Engineer

Model

Validate

Deploy

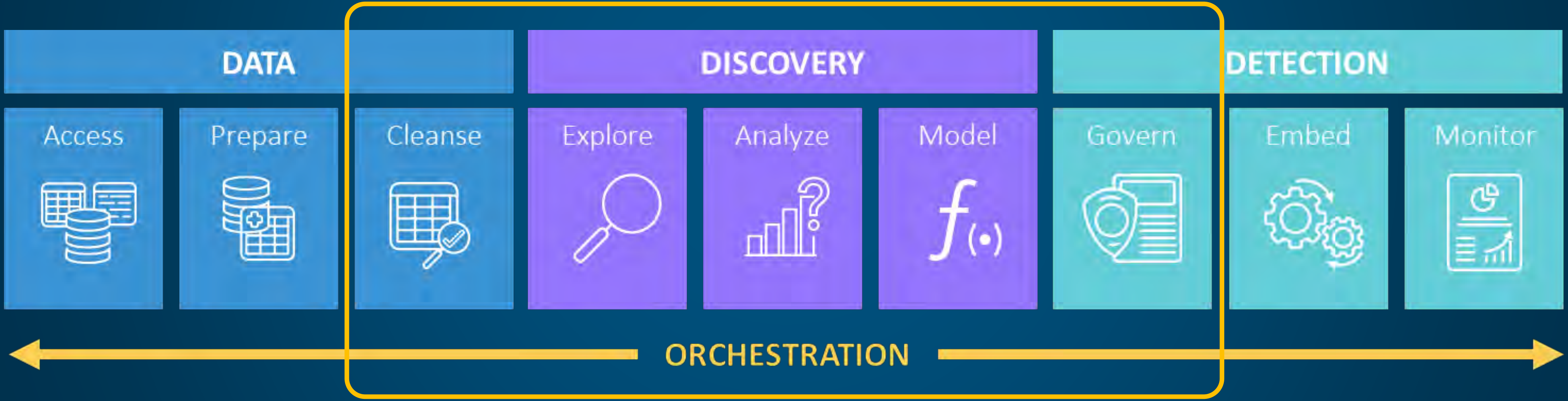
Monitor

DETECTION

DISCOVERY

# CSDS Process

## Unified Orchestration



DATA

- Access
- Prepare
- Cleanse

DISCOVERY

- Explore
- Analyze
- Model

DETECTION

- Govern
- Embed
- Monitor

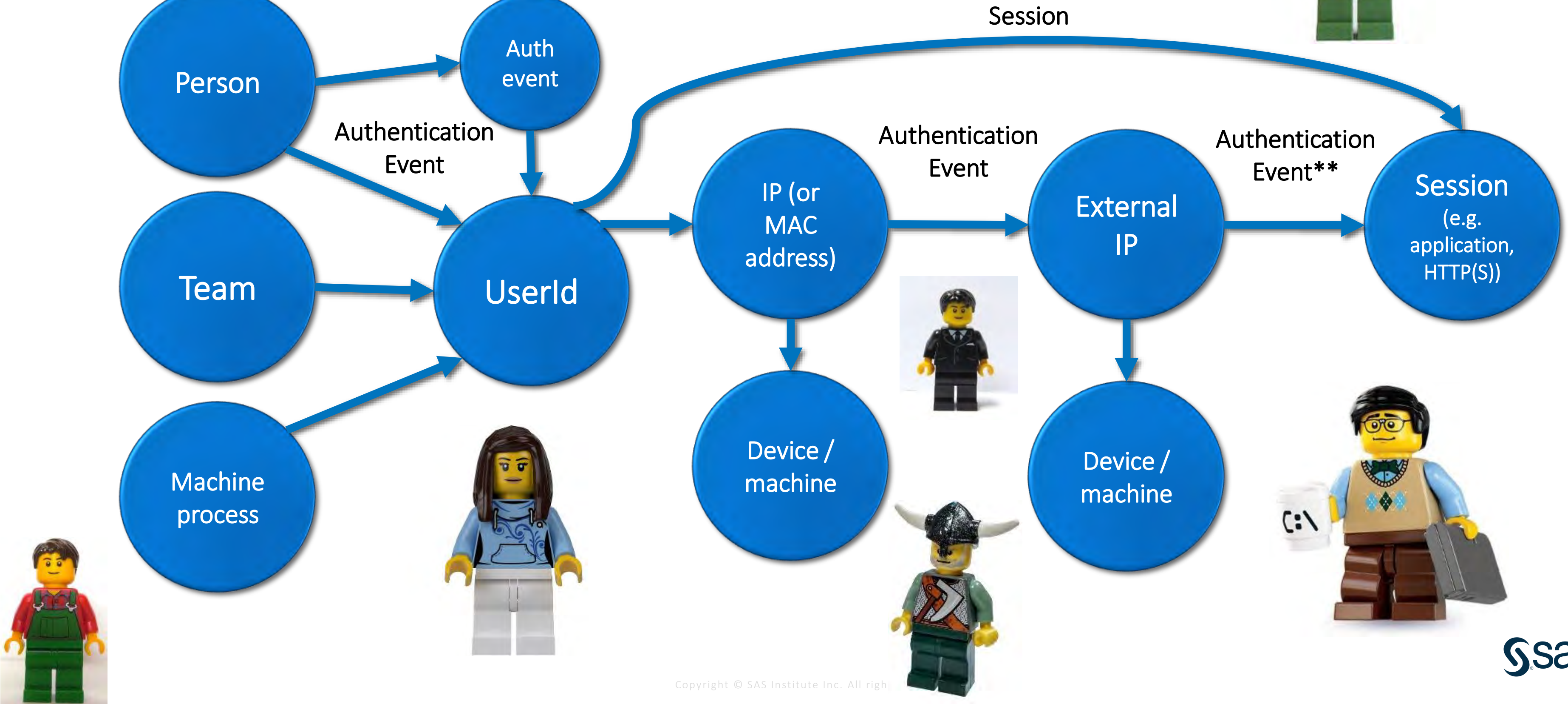
ORCHESTRATION

# Cybersecurity Events

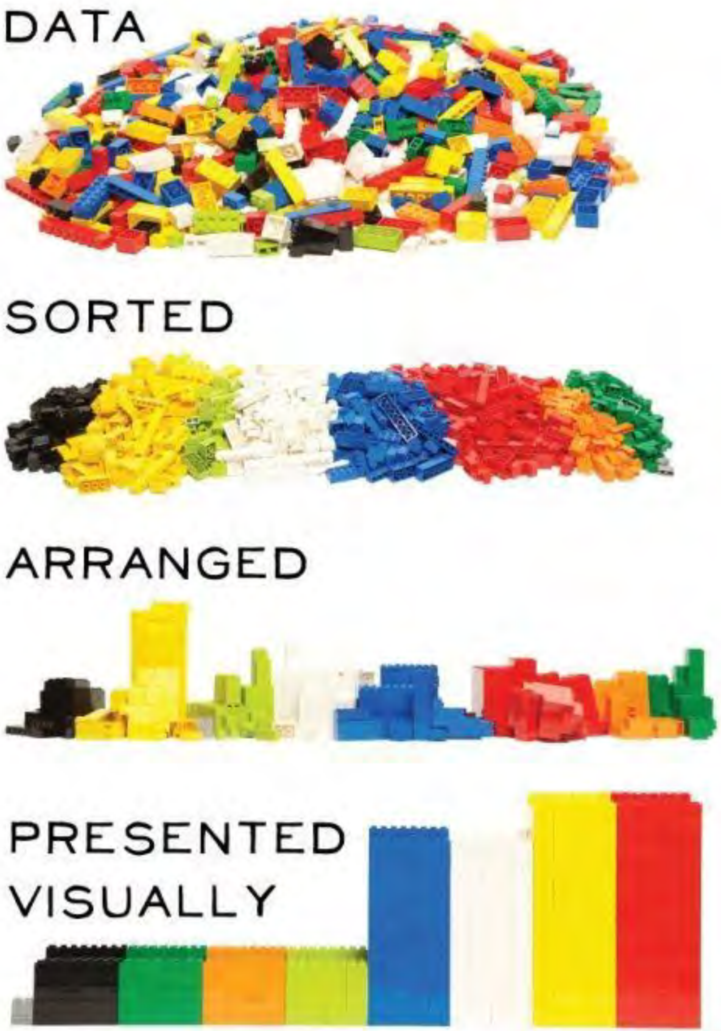
## Irregular and Complex Events

time's arrow…

# What is a User, Anyway?



# behavioral profile



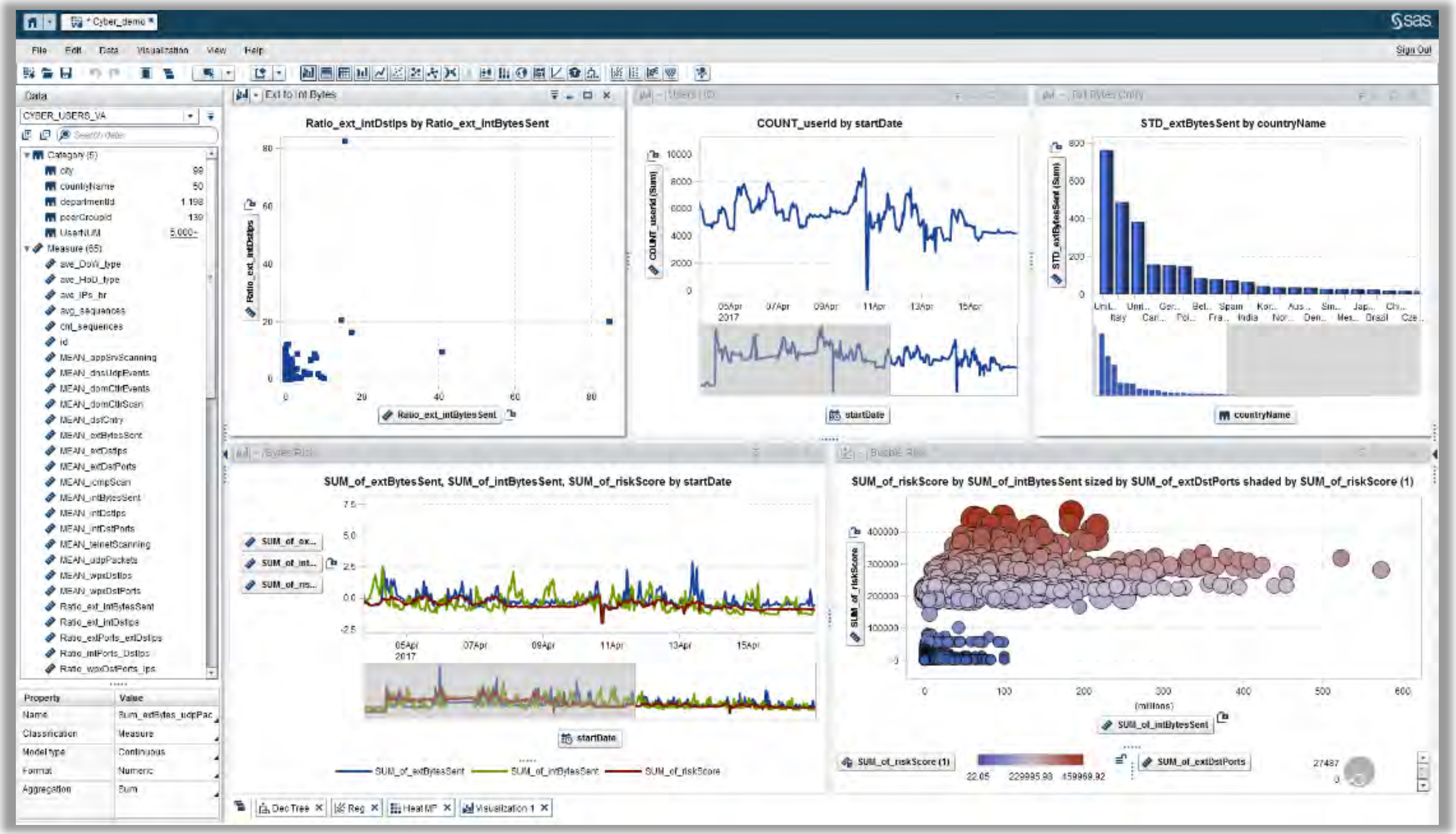
DATA

SORTED

ARRANGED

PRESENTED VISUALLY

# Self-Service Visual Analytics

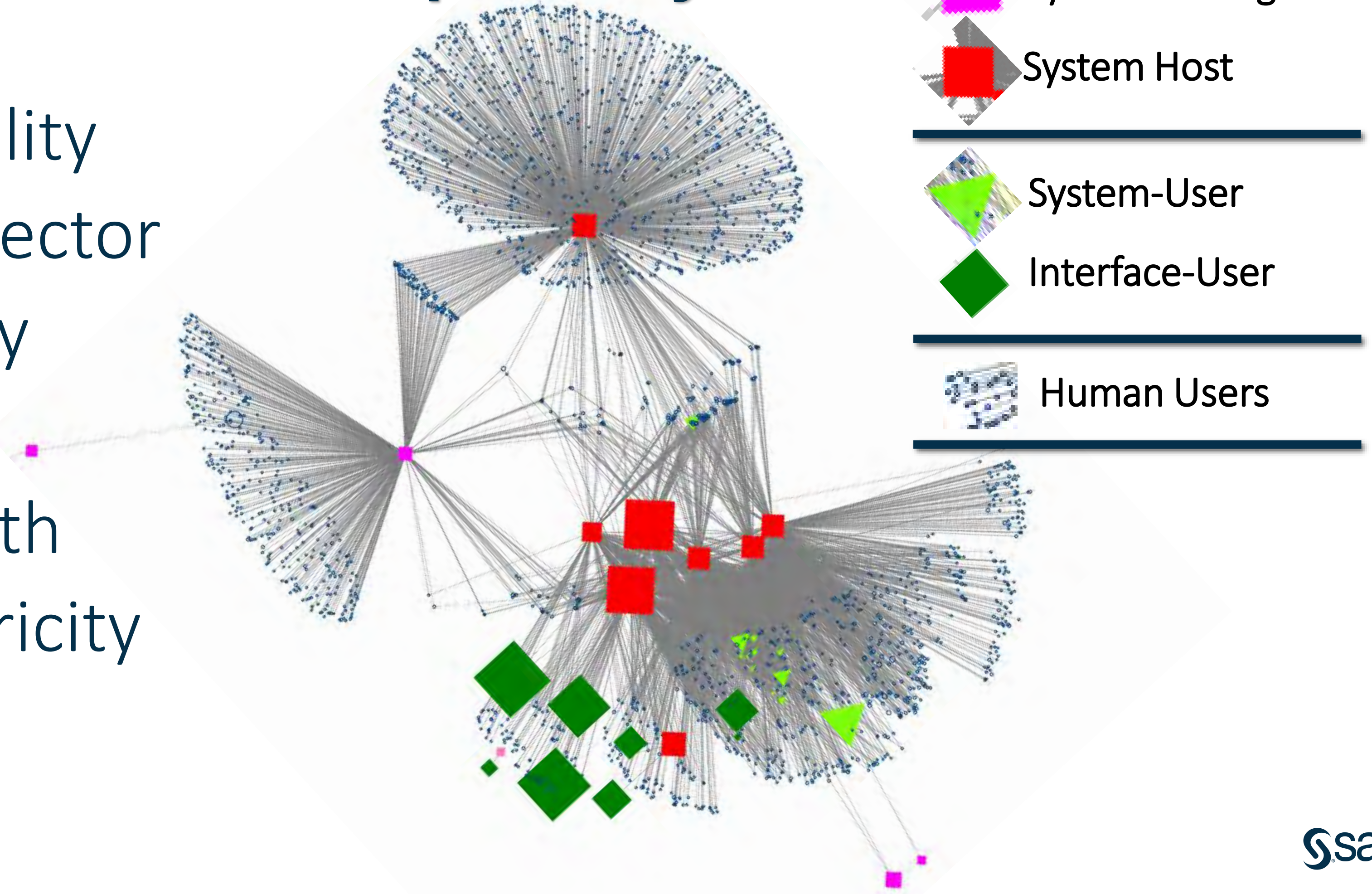


| | | |
|---|---|---|
| Category (5) | | |
| | city | 99 |
| | countryName | 50 |
| | departmentId | 1 198 |
| | peerGroupId | 139 |
| | UserNUM | 5,000+ |
| Measure (65) | | |
| | ave_DoW_type | |
| | ave_HoD_type | |
| | ave_IPs_hr | |
| | avg_sequences | |
| | cnt_sequences | |
| | id | |
| | MEAN_appSrvScanning | |
| | MEAN_dnsUdpEvents | |
| | MEAN_domCtlrEvents | |
| | MEAN_domCtlrScan | |
| | MEAN_dstCntry | |
| | MEAN_extBytesSent | |
| | MEAN_extDstIps | |
| | MEAN_extDstPorts | |
| | MEAN_icmpScan | |
| | MEAN_intBytesSent | |
| | MEAN_intDstIps | |
| | MEAN_intDstPorts | |
| | MEAN_telnetScanning | |
| | MEAN_udpPackets | |
| | MEAN_wpxDstIps | |
| | MEAN_wpxDstPorts | |
| | Ratio_ext_intBytesSent | |
| | Ratio_ext_intDstIps | |
| | Ratio_extPorts_extDstIps | |
| | Ratio_intPorts_DstIps | |
| | Ratio_wpxDstPorts_Ips | |

| Property | Value |
|---|---|
| Name | Sum_extBytes_udpPac |
| Classification | Measure |
| Model type | Continuous |
| Format | Numeric |
| Aggregation | Sum |

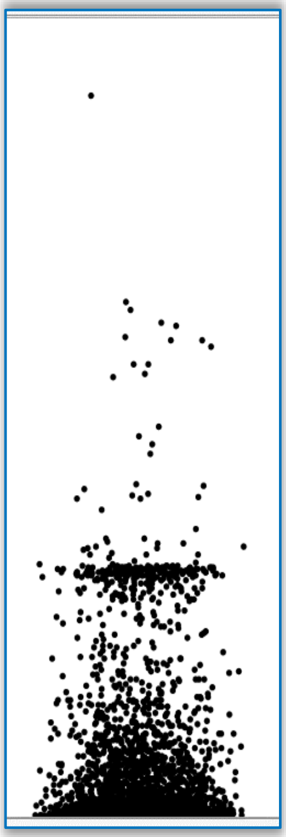
# Network Graph Analytics

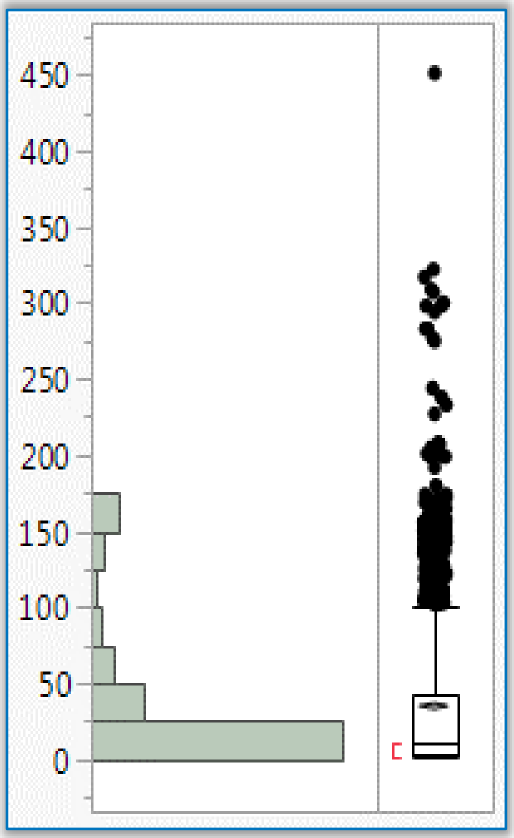
- Centrality
- Eigenvector
- Density
- Reach
- Strength
- Recopricity



# Feature Selection / Extraction

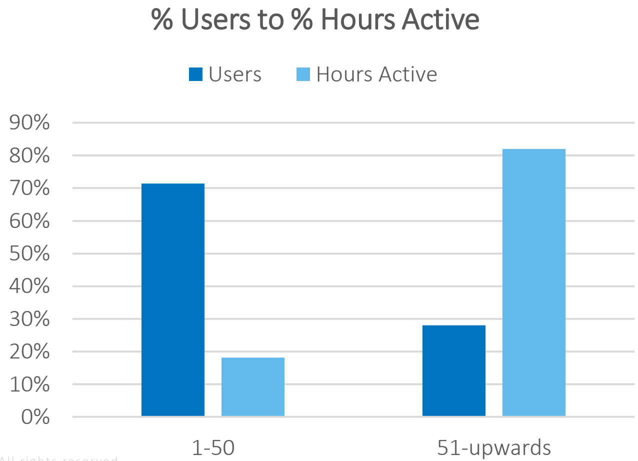
## Understanding Network Behavioral Patterns





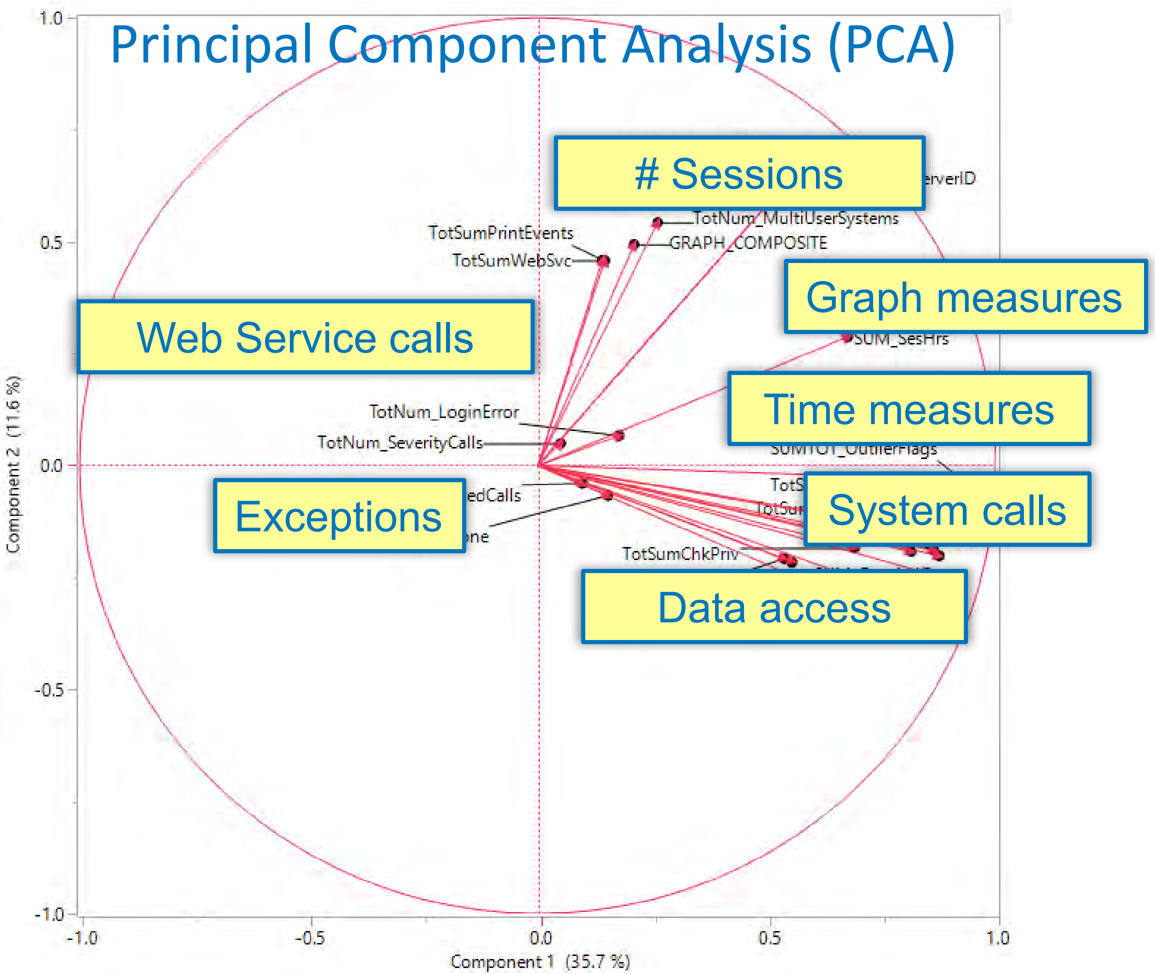
### Pareto Principle

- **80/20%** pattern in network-usage
- *Outliers*: multiple devices 24 hours online
- High correlation: hrs online and breadth of activities
- Pattern observed across multiple networks



# Dimensionality Reduction

## Principal Component Analysis (PCA)



# Sessions

Graph measures

Web Service calls

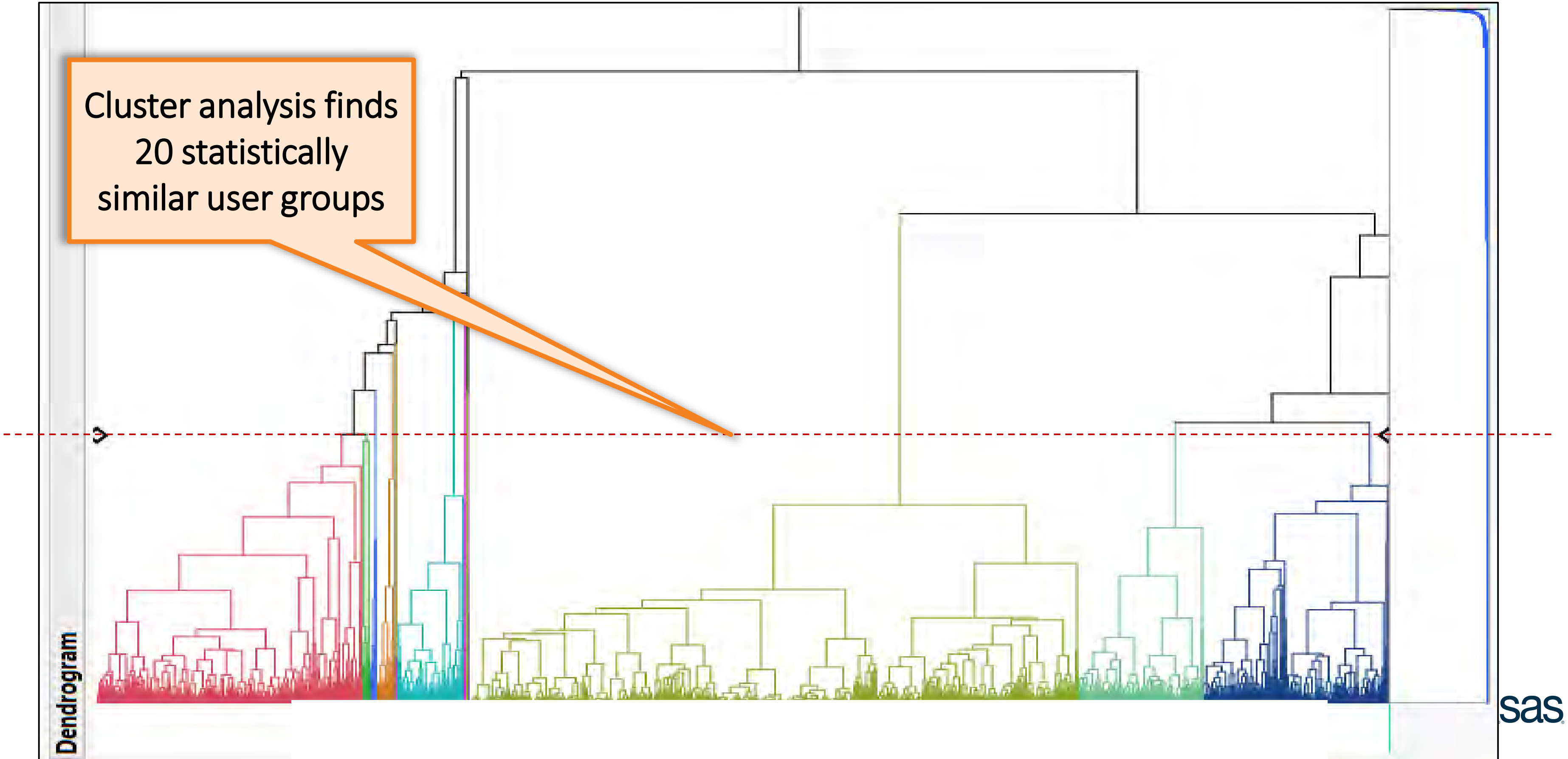
Time measures

Exceptions

System calls

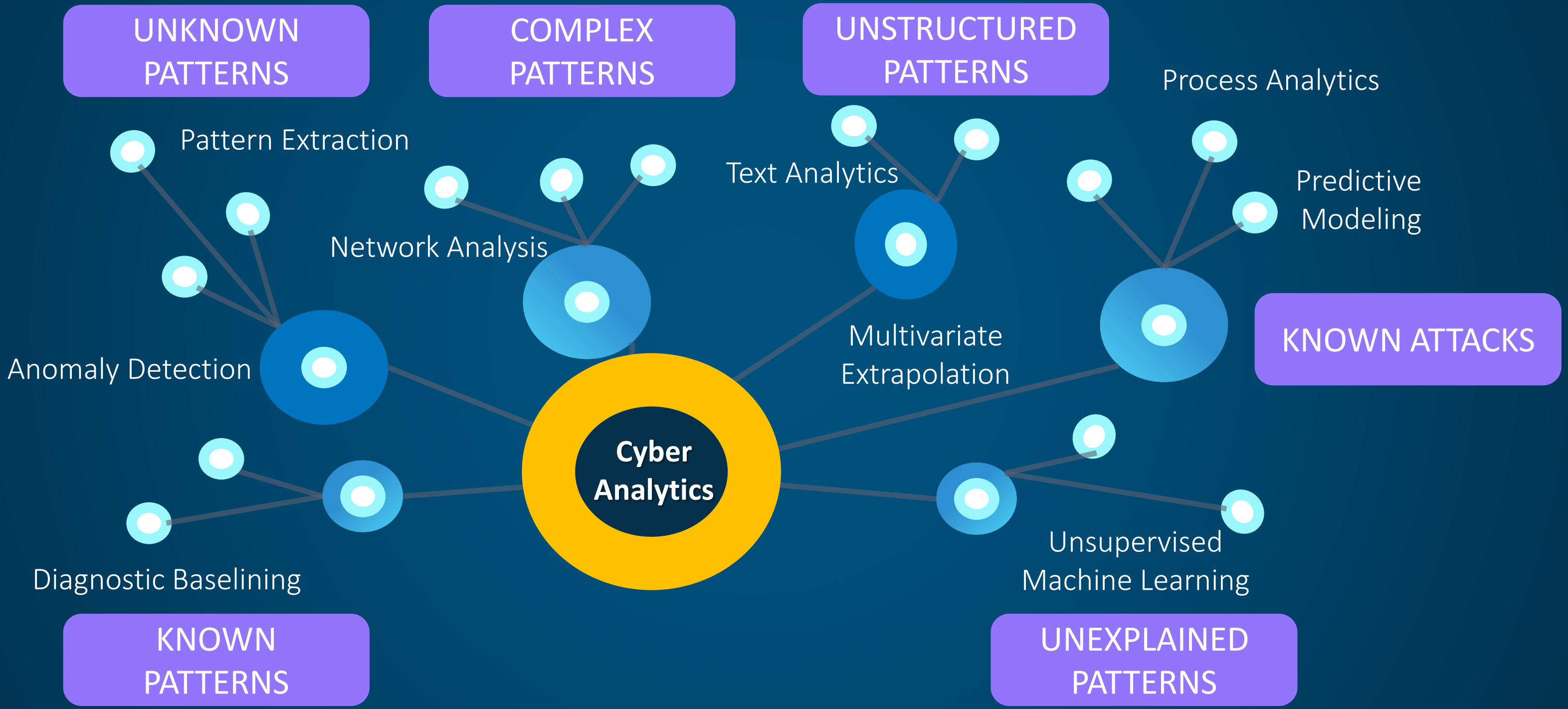
Data access

# Pattern Extrapolation Machine Learning (Unsupervised)



Cluster analysis finds 20 statistically similar user groups

# CSDS: Diverse Analytics Toolkit



UNKNOWN PATTERNS

COMPLEX PATTERNS

UNSTRUCTURED PATTERNS

Process Analytics

Pattern Extraction

Text Analytics

Network Analysis

Predictive Modeling

Anomaly Detection

Multivariate Extrapolation

KNOWN ATTACKS

Cyber Analytics

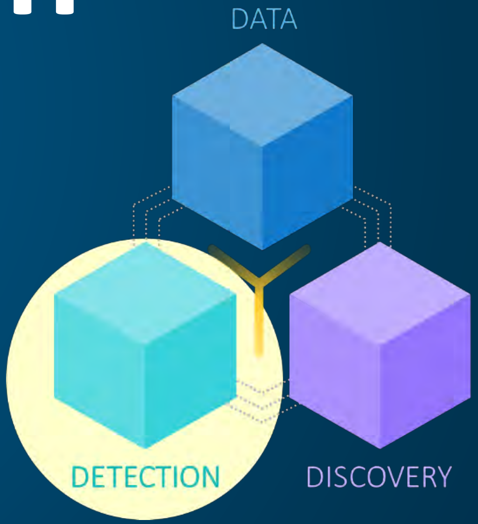
Diagnostic Baselining

Unsupervised Machine Learning

KNOWN PATTERNS

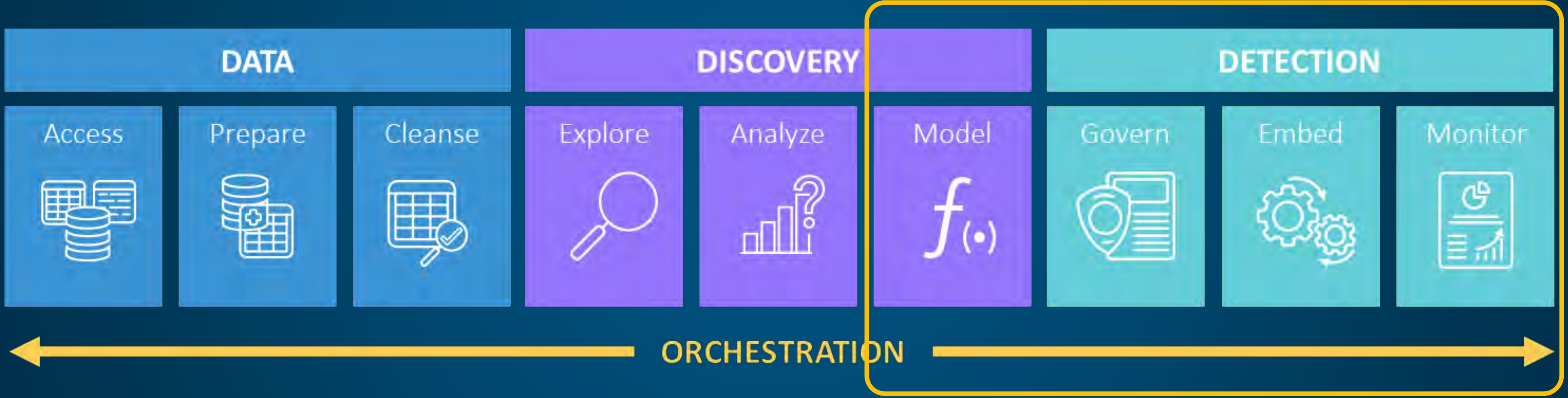
UNEXPLAINED PATTERNS

# CSDS Detection



# CSDS Process

## Unified Orchestration



DATA

Access

Prepare

Cleanse

DISCOVERY

Explore

Analyze

Model

DETECTION

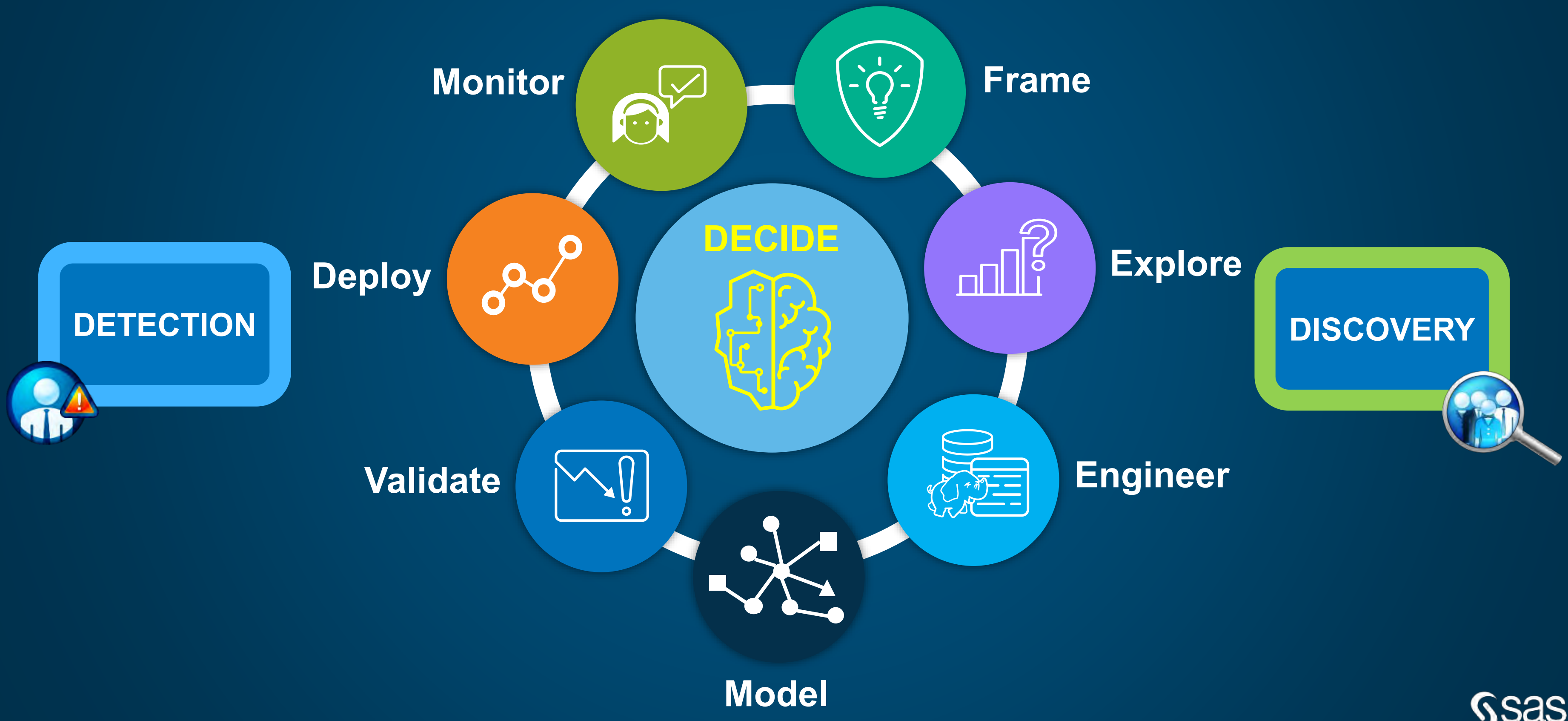
Govern

Embed

Monitor

ORCHESTRATION

# Cybersecurity Data Science (CSDS) Lifecycle



DETECTION

Deploy

Monitor

DECIDE

Frame

Explore

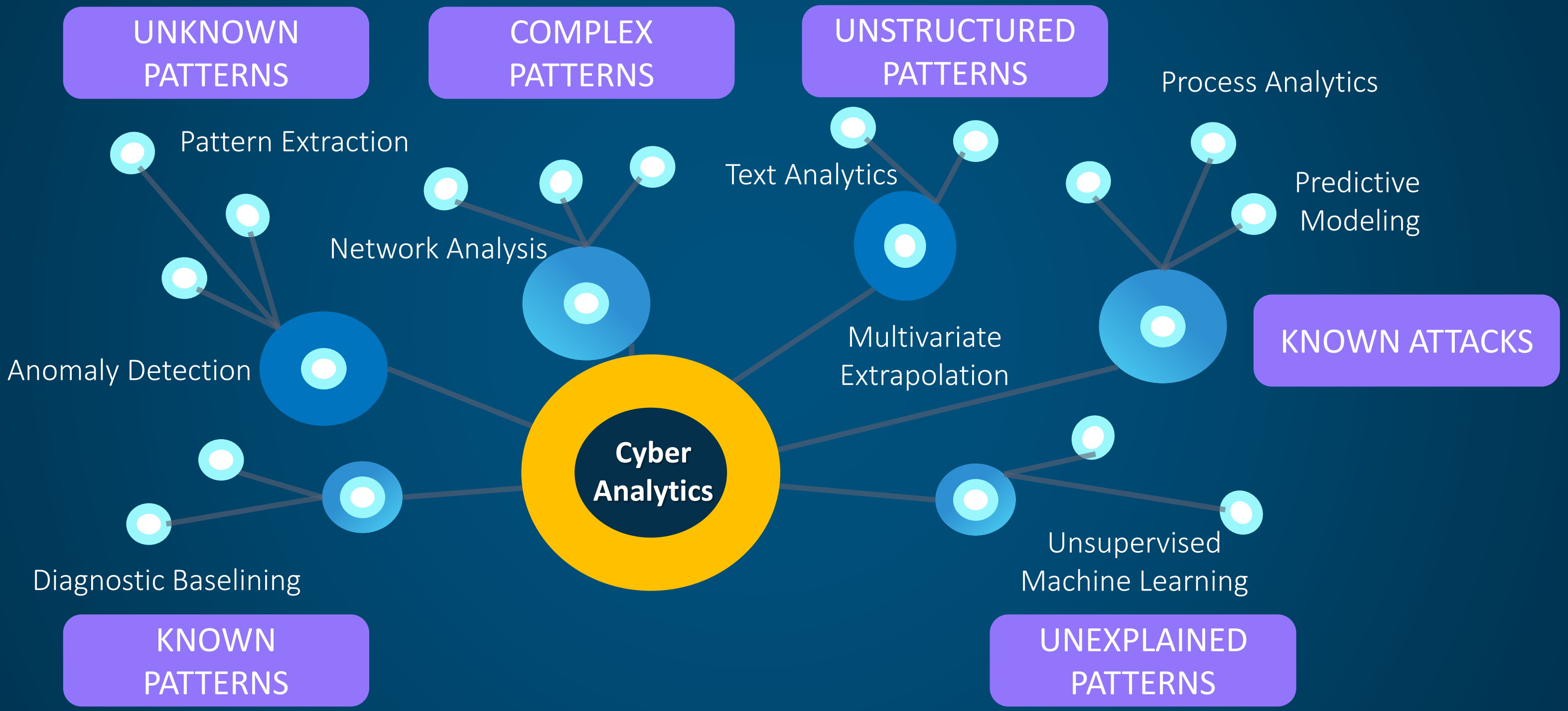
DISCOVERY

Engineer

Model

Validate

# CSDS: Diverse Analytics Toolkit



UNKNOWN PATTERNS

COMPLEX PATTERNS

UNSTRUCTURED PATTERNS

Process Analytics

Pattern Extraction

Text Analytics

Predictive Modeling

Network Analysis

Anomaly Detection

Multivariate Extrapolation

KNOWN ATTACKS

Cyber Analytics

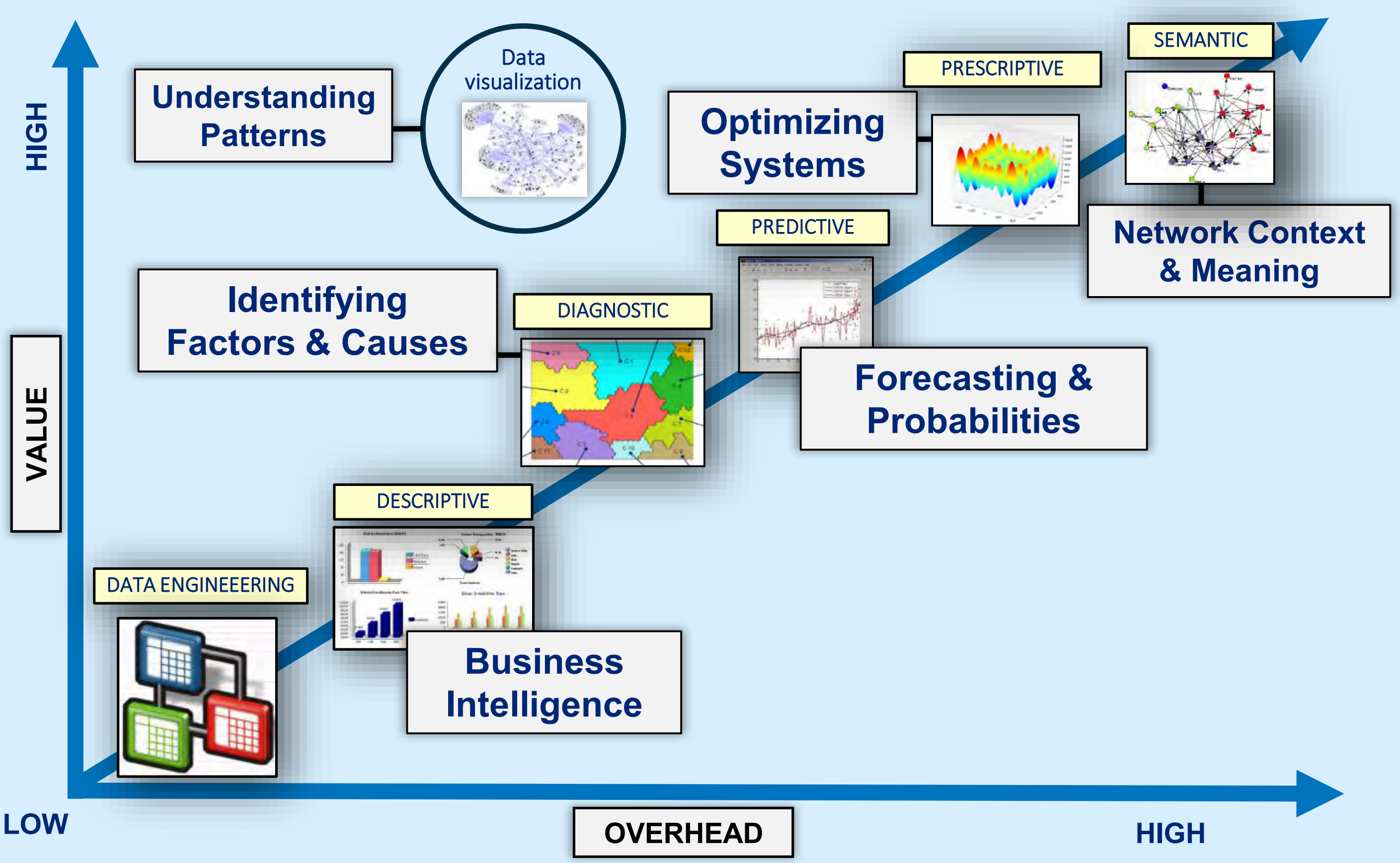
Unsupervised Machine Learning

Diagnostic Baselining

KNOWN PATTERNS

UNEXPLAINED PATTERNS

# Data Science: Continuum of Analytics Methods



VALUE

HIGH

LOW

OVERHEAD

HIGH

DATA ENGINEERING

DESCRIPTIVE

Business Intelligence

DIAGNOSTIC

Identifying Factors & Causes

PREDICTIVE

Forecasting & Probabilities

PRESCRIPTIVE

Optimizing Systems

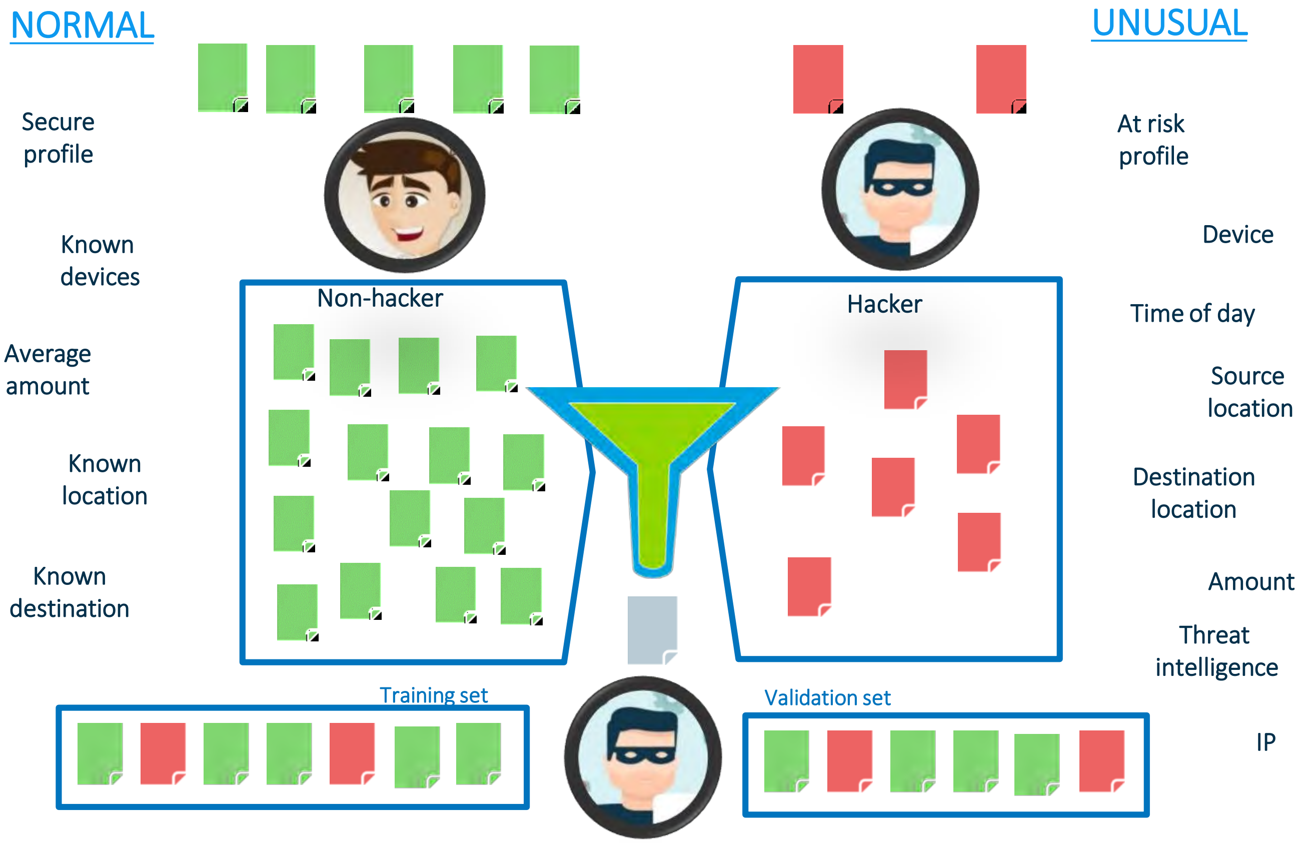
SEMANTIC

Network Context & Meaning

Understanding Patterns

Data visualization

# Predictive Machine Learning (Supervised)



## SUPERVISED MACHINE LEARNING

Decision trees

Gradient boosting

Random forests

Naïve Bayes

SVM

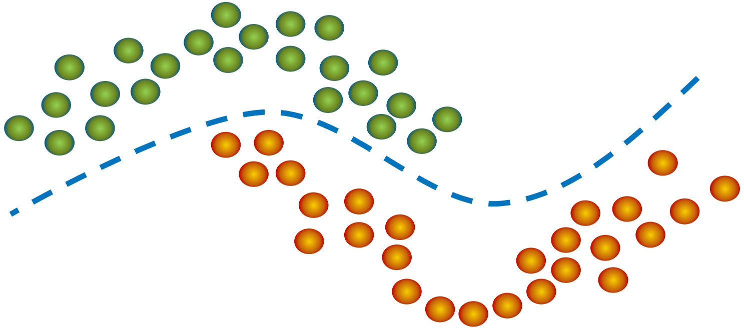
Gaussian processes

# Supervised Learning

- Trained on labeled examples. We have a target we are predicting.
- Map inputs to desired output.
- Suitable for classification and prediction.

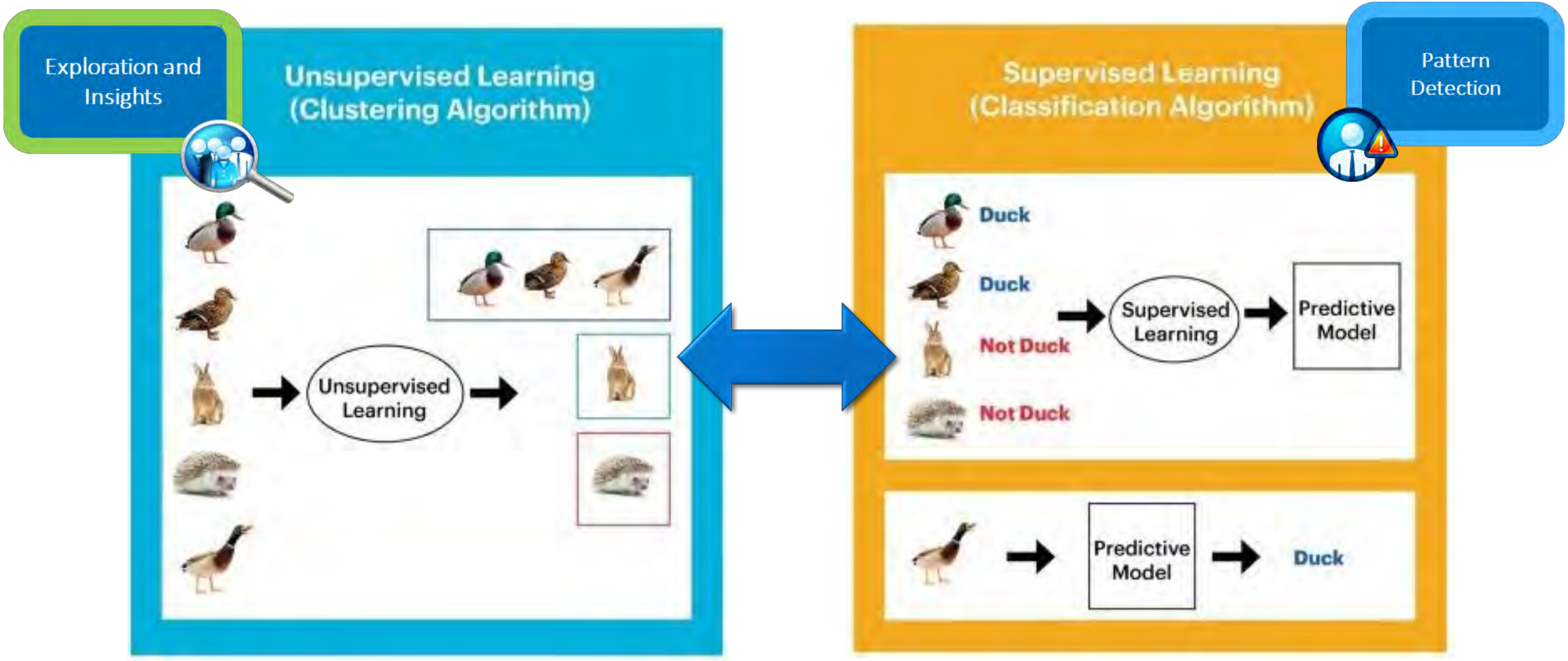
### Considerations

- Obtaining labeled data for rare events can be a challenge
- Suspicion is not a cyber incident!
- Data is skewed 99-1



# Machine Learning Model = Active Data Vehicle

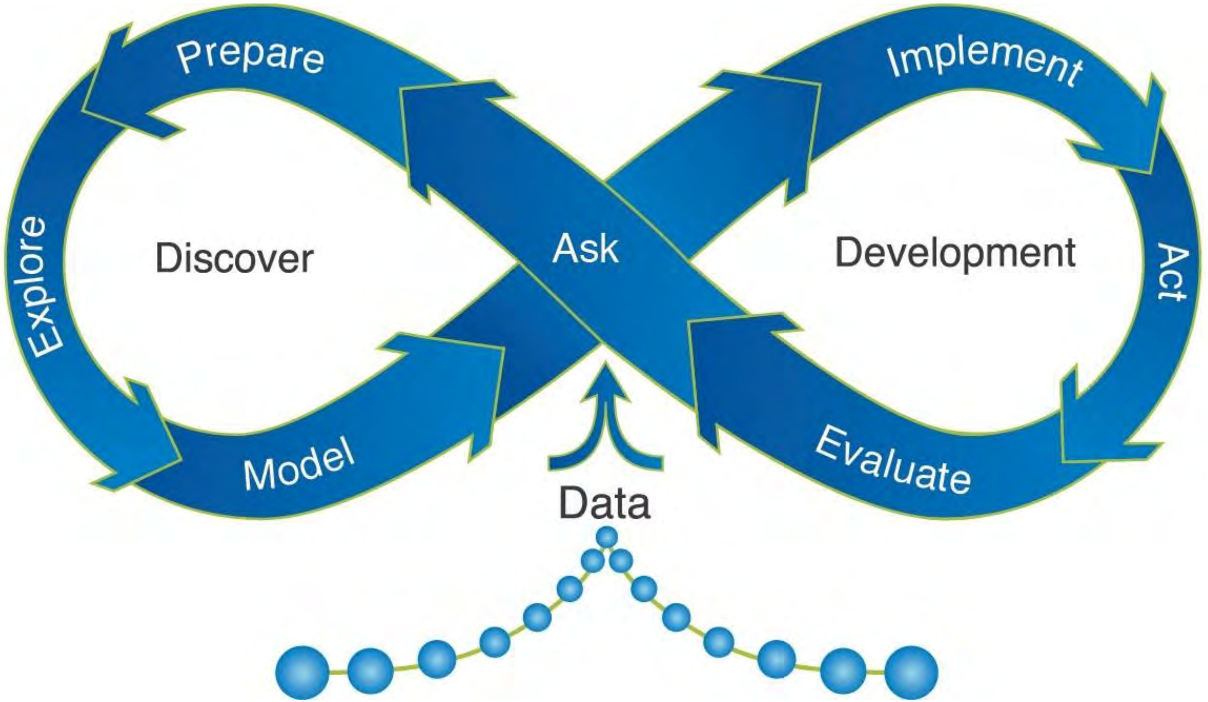
# Machine Learning Segmentation and Classification



https://medium.com/datadriveninvestor/differences-between-ai-and-machine-learning-and-why-it-matters-1255b182fc6

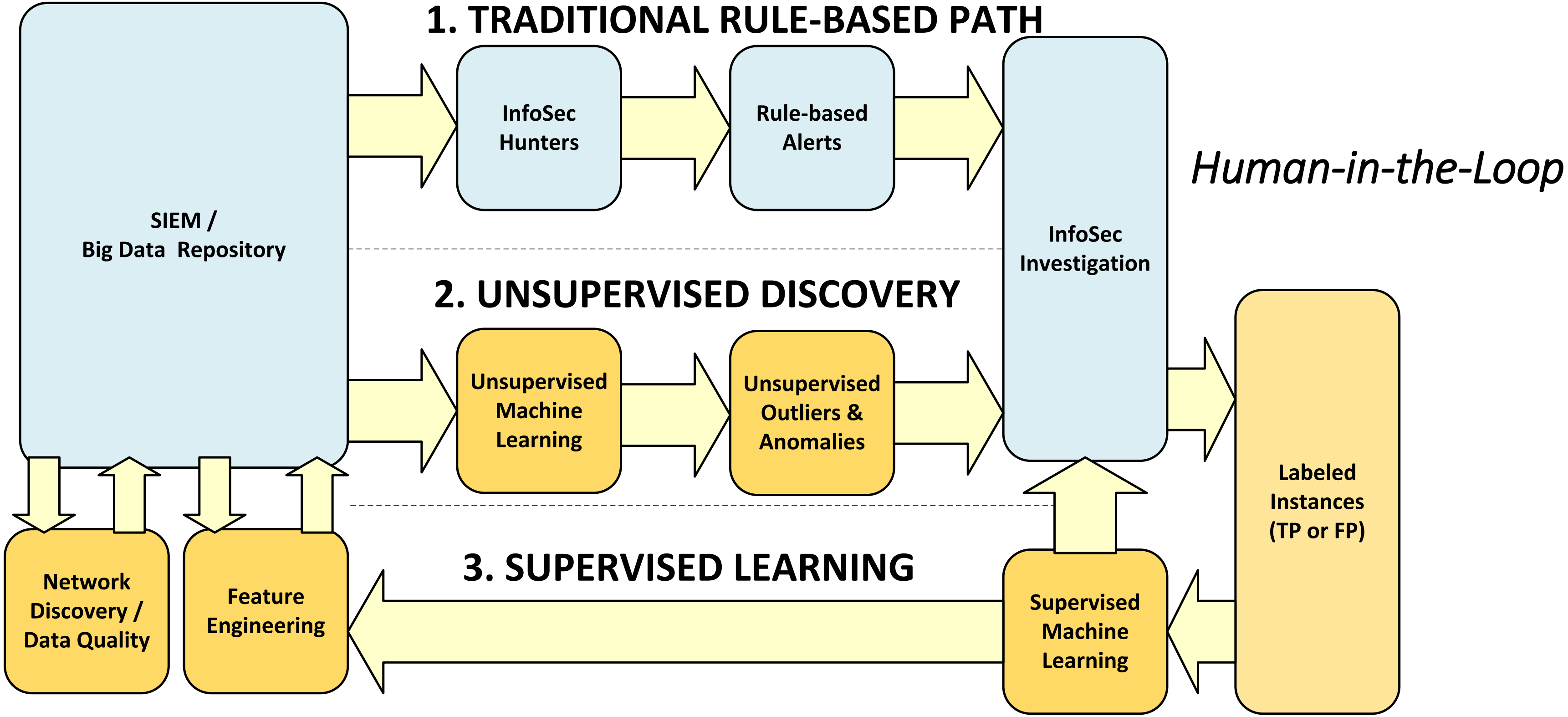
# Model Building Process ⇔ Analytics Life Cycle

Raw Data | Feature Selection | Features | Feature Engineering | Modeling | Insights



SAS: ‘Managing the Analytics Life Cycle for Decisions at Scale’

# Bootstrapping Machine Learning Facilitated Cyber Detection



1. TRADITIONAL RULE-BASED PATH

2. UNSUPERVISED DISCOVERY

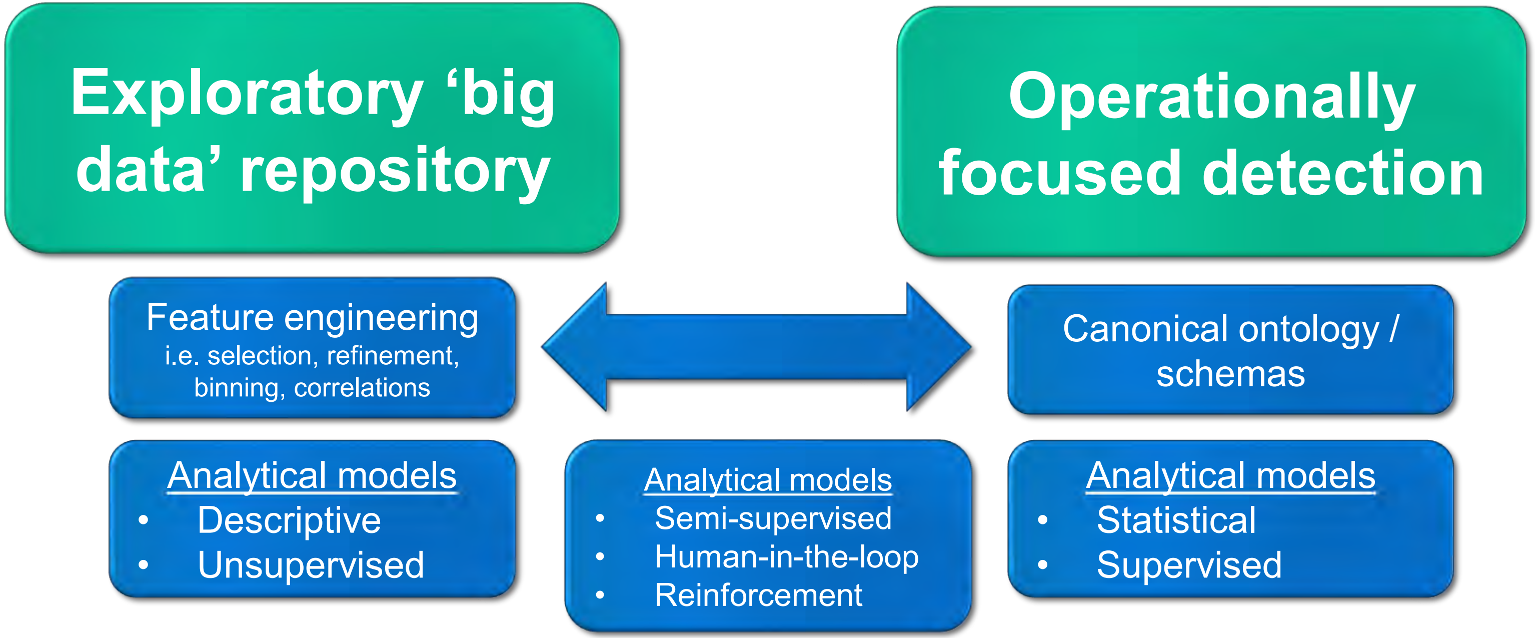
3. SUPERVISED LEARNING

*Human-in-the-Loop*

SIEM / Big Data Repository
InfoSec Hunters
Rule-based Alerts
InfoSec Investigation
Unsupervised Machine Learning
Unsupervised Outliers & Anomalies
Labeled Instances (TP or FP)
Network Discovery / Data Quality
Feature Engineering
Supervised Machine Learning

# Architecture: Exploratory & Detection Platforms*

## Functional Architectural Segmentation



Exploratory 'big data' repository

Operationally focused detection

Feature engineering
i.e. selection, refinement, binning, correlations

Canonical ontology / schemas

Analytical models
- Descriptive
- Unsupervised

Analytical models
- Semi-supervised
- Human-in-the-loop
- Reinforcement

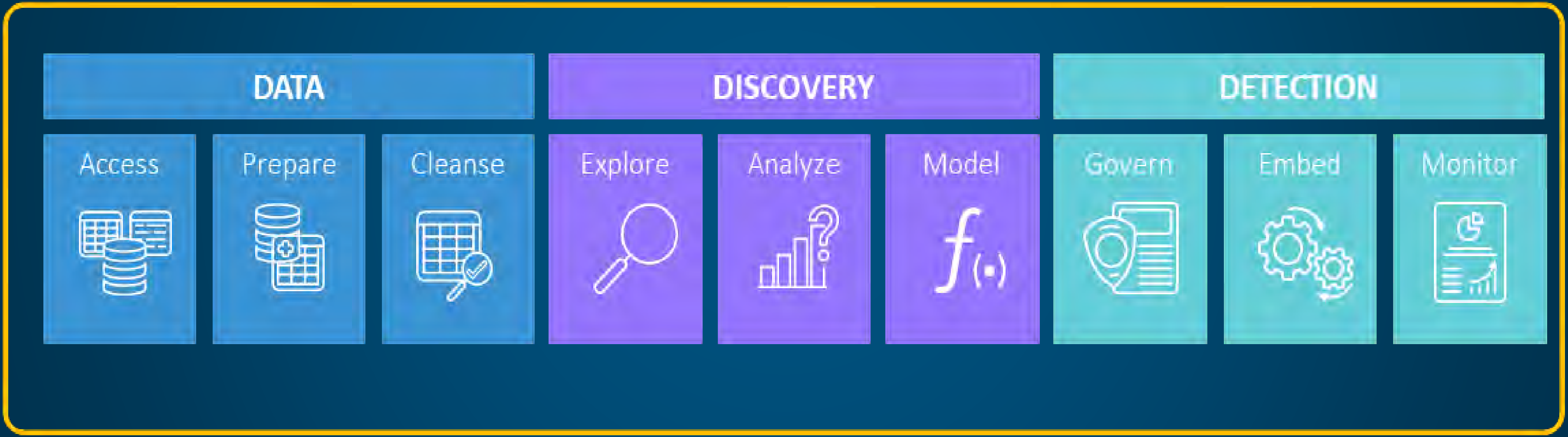
Analytical models
- Statistical
- Supervised

** Runs counter to the vendor stance of store 'all-the-data-all-the-time'*

# Summary

# Accelerate the Analytics Lifecycle

## A Unified Process Approach



DATA

- Access
- Prepare
- Cleanse

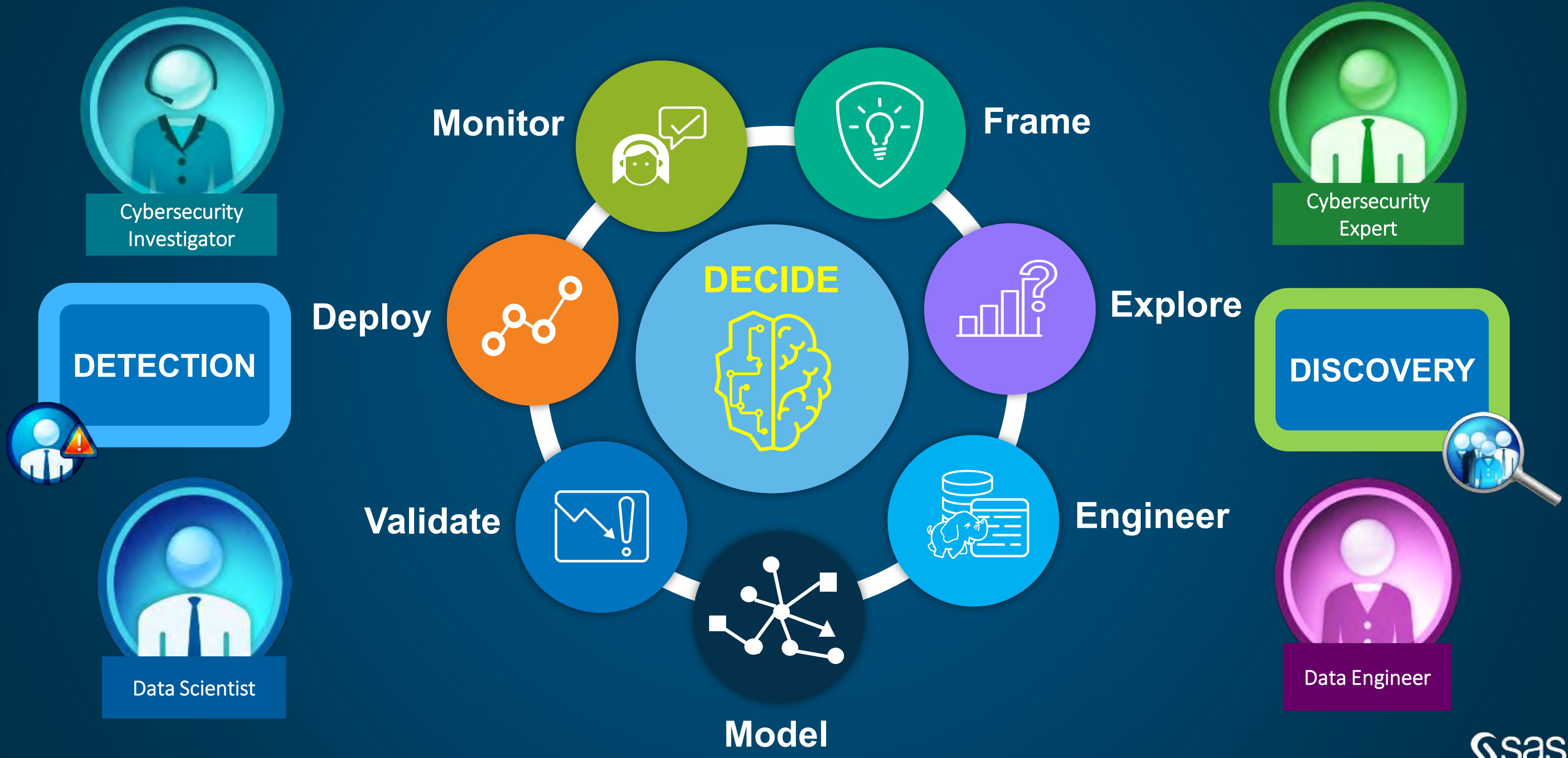
DISCOVERY

- Explore
- Analyze
- Model

DETECTION

- Govern
- Embed
- Monitor

# Cybersecurity Data Science (CSDS) Lifecycle



Cybersecurity Investigator

DETECTION

Data Scientist

Monitor

Frame

DECIDE

Deploy

Explore

Validate

Model

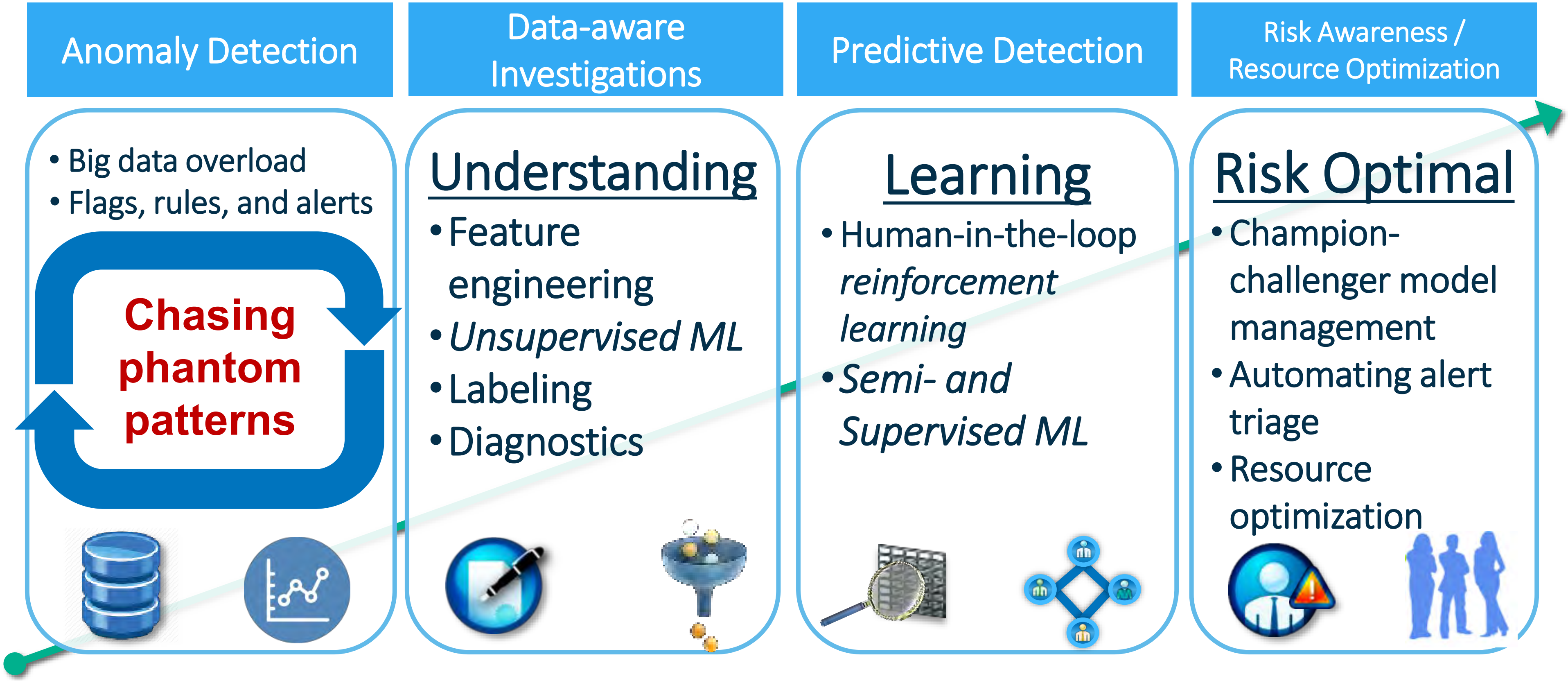
Engineer

Cybersecurity Expert

DISCOVERY

Data Engineer

# Cybersecurity Analytics Maturity Model



## Anomaly Detection

- Big data overload
- Flags, rules, and alerts

**Chasing phantom patterns**

## Data-aware Investigations

### Understanding

- Feature engineering
- *Unsupervised ML*
- Labeling
- Diagnostics

## Predictive Detection

### Learning

- Human-in-the-loop *reinforcement learning*
- *Semi- and Supervised ML*

## Risk Awareness / Resource Optimization

### Risk Optimal

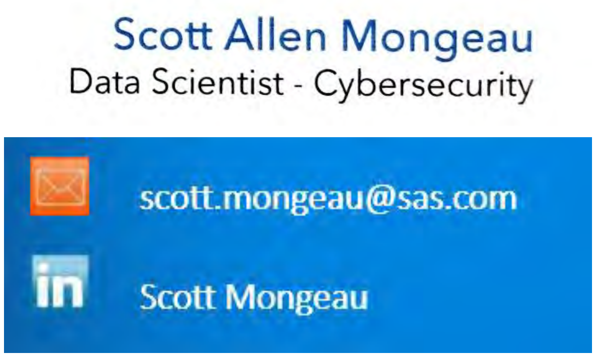
- Champion-challenger model management
- Automating alert triage
- Resource optimization

# Want to Know More?

SAS whitepaper ‘*Data Management for Artificial Intelligence*’

SAS Cybersecurity Solution (SCS)

www.sas.com/en_us/software/cybersecurity.html



Scott Allen Mongeau
Data Scientist - Cybersecurity

✉ scott.mongeau@sas.com

in Scott Mongeau

www.sas.com/en_us/whitepapers/data-management-artificial-intelligence-109860.html

# REFERENCES

# REFERENCES

- Aggarwal, C. (2013). “Outlier Analysis.” Springer. http://www.springer.com/la/book/9781461463955
- Harris, H., Murphy, S., and Vaisman, M. (2013). “Analyzing the Analyzers.” O’Reilly Media. Available at https://www.oreilly.com/data/free/analyzing-the-analyzers.csp
- Kirchhoff, C., Upton, D., and Winnefeld, Jr., Admiral J. A. (2015 October 7). “Defending Your Networks: Lessons from the Pentagon.” Harvard Business Review. Available at https://hbr.org/webinar/2015/10/defending-your-networks-lessons-from-the-pentagon
- Mongeau, S. (2018). “Cybersecurity Data Science (CSDS).” SCTR7.com. https://sctr7.com/2018/12/03/cybersecurity-data-science-csds-how-not-to-drown-in-your-cyber-data-lake/
- Mongeau, S. (2017). “Cybersecurity Big Data Overload?” SCTR7.com. https://sctr7.com/2017/10/22/cybersecurity-big-data-overload/
- Ponemon Institute. (2017). “When Seconds Count: How Security Analytics Improves Cybersecurity Defenses.” Available at https://www.sas.com/en_us/whitepapers/ponemon-how-security-analytics-improves-cybersecurity-defenses-108679.html
- SANS Institute. (2015). “2015 Analytics and Intelligence Survey.” Available at https://www.sas.com/en_us/whitepapers/sans-analytics-intelligence-survey-108031.html
- SANS Institute. (2016). “Using Analytics to Predict Future Attacks and Breaches.” Available at https://www.sas.com/en_us/whitepapers/sans-using-analytics-to-predict-future-attacks-breaches-108130.html
- SAS Institute. (2016). “Managing the Analytical Life Cycle for Decisions at Scale.” Available at https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/manage-analytical-life-cycle-continuous-innovation-106179.pdf
- SAS Institute. (2017). “SAS Cybersecurity: Counter cyberattacks with your information advantage.” Available at https://www.sas.com/en_us/software/fraud-security-intelligence/cybersecurity-solutions.html
- UBM. (2016). “Dark Reading: Close the Detection Deficit with Security Analytics.” Available at https://www.sas.com/en_us/whitepapers/close-detection-deficit-with-security-analytics-108280.html