Cybersecurity Data Science (CSDS)

Practitioner Methods and Best Practices

PART 1: FRAME

Scott Allen Mongeau
Cybersecurity Data Scientist
SAS Institute
scott.mongeau@sas.com

Cybersecurity Data Science (CSDS)

TOPIC
1. FRAME
2. DATA
3. DISCOVER
4. DETECT
5. DEPLOY

Introductions, Expectations, Agenda

Scott Allen Mongeau
Cybersecurity Data Scientist
MA GD MA MBA PhD (ABD)

+31 (0)68 370 3097

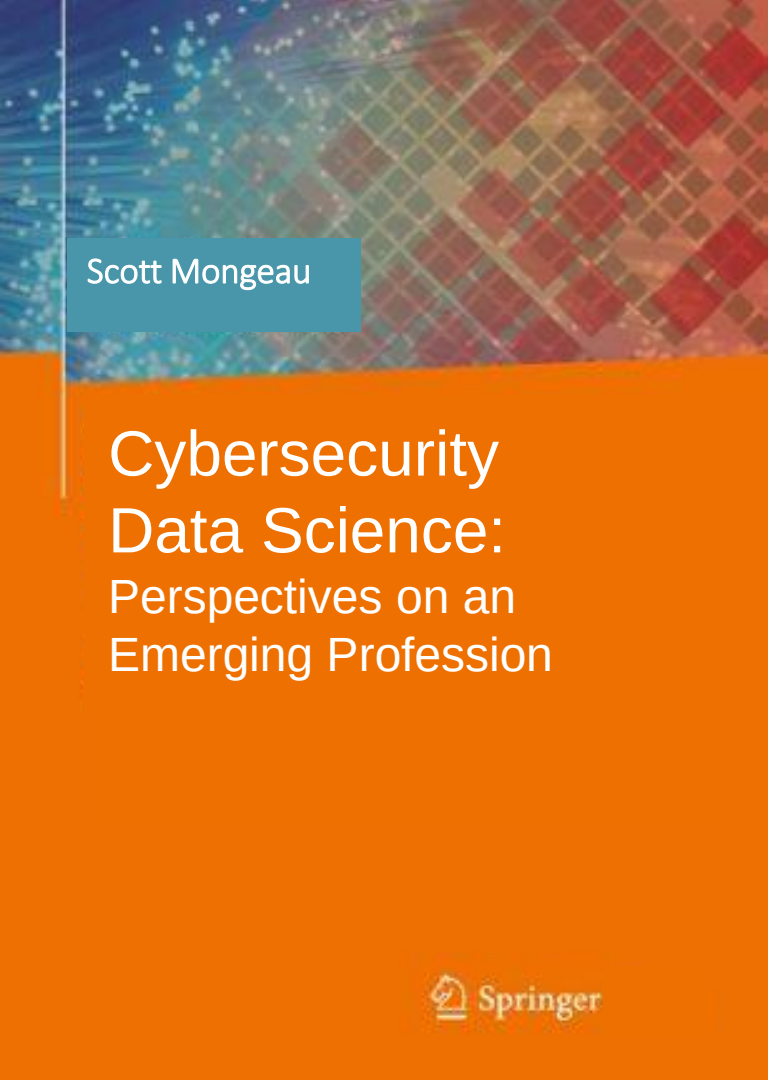
scott@sark7.com

Scott Allen Mongeau

SARK7

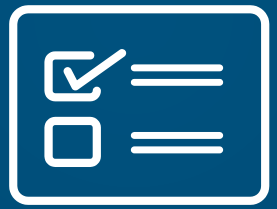
Cybersecurity Data Science (CSDS)

- **SAS (2015 – present)**
  - F&SI Cyber – Global (EU/Netherlands based)
  - Fraud & Financial Crime Analytics (London)
- **SARK7 (2008 – present)**
  - Data analytics consulting (Leiden)
- **Deloitte (2013 – 15)**
  - Fraud, Financial Crime, Cyber (Amsterdam)
- **Genentech Inc. (2000 – 2008)**
  - Data analytics (San Francisco, CA)



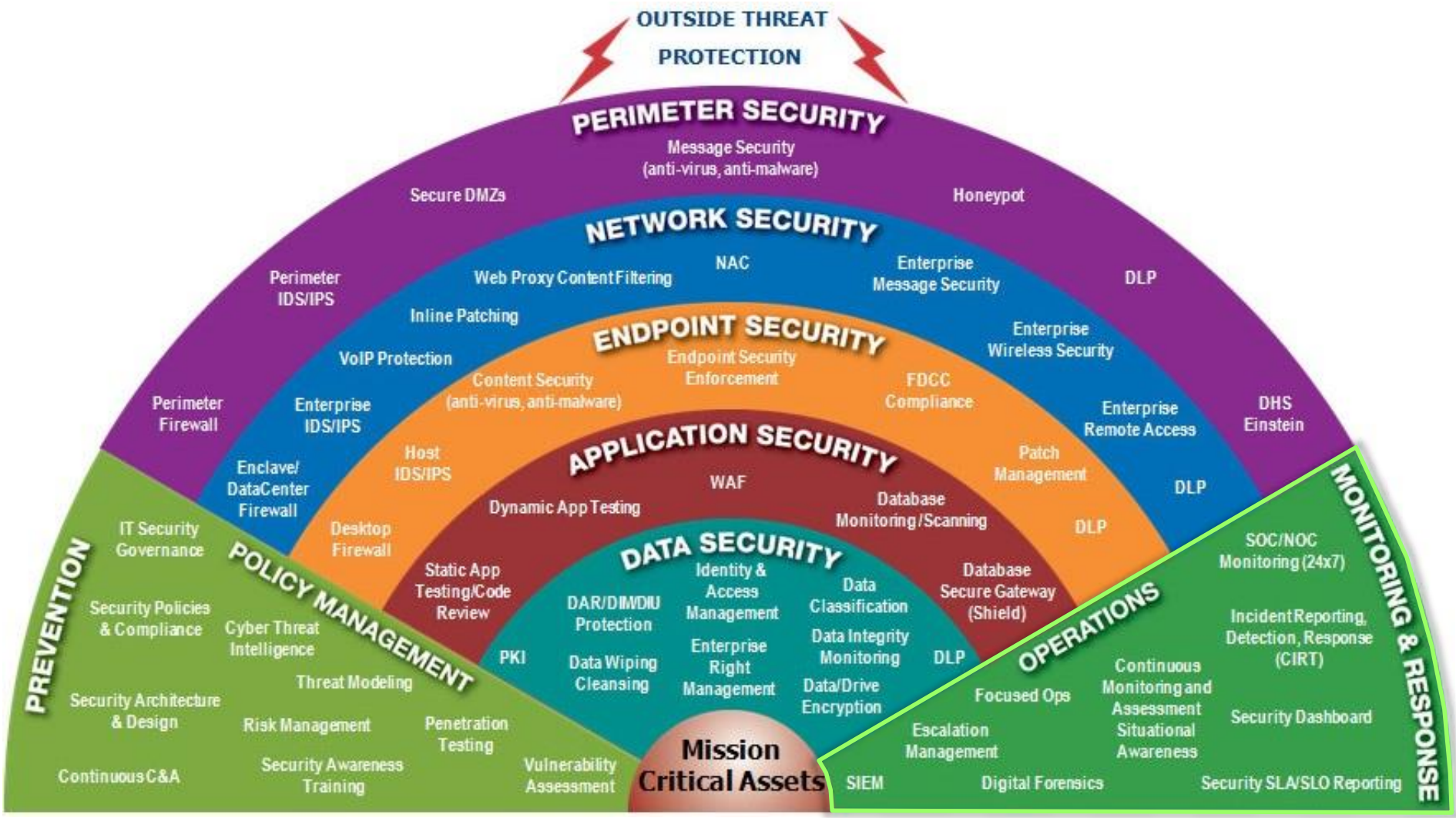
- Defining key CSDS challenges
- Outline best practices
- Guidance to managers and practitioners in implementation of CSDS programs

Key Learning Objectives



Cybersecurity Services and Solutions: Overview of Major Areas


1
Identify
AM: Asset Management
BE: Business Environment
GV: Governance
RA: Risk Assesment
RM: Risk Management Strategy
2
Protect
AC: Access Control
AT: Awareness Training
DS: Data Security
IP: Information Protection
Processes and Procedures
PT: Pretective Technology
3
Detect
AE: Anomolies and Events
CM: Security Continuous Monitoring
DP: Detection Processes
4
Respond
RP: Response Planing
CO: Communications
AN: Analysis
MI: Mitigation
IM: Improvements
5
Recover
RP: Recovery Planning
IM: Improvements
CO: Communications
NIST
Cybersecurity
Framework


OUTSIDE THREAT
PROTECTION
PERIMETER SECURITY
Message Security
(anti-virus, anti-malware)
Secure DMZs
Honeypot
NETWORK SECURITY
Perimeter
IDS/IPS
Web Proxy Content Filtering
NAC
Enterprise
Message Security
DLP
Inline Patching
ENDPOINT SECURITY
Enterprise
Wireless Security
VoIP Protection
Endpoint Security
Enforcement
Content Security
(anti-virus, anti-malware)
FDCC
Compliance
Perimeter
Firewall
Enterprise
IDS/IPS
Enterprise
Remote Access
DHS
Einstein
Host
IDS/IPS
APPLICATION SECURITY
Patch
Management
Enclave/
DataCenter
Firewall
WAF
DLP
Dynamic App Testing
Database
Monitoring/Scanning
Desktop
Firewall
DLP
IT Security
Governance
DATA SECURITY
SOC/NOC
Monitoring (24x7)
Static App
Testing/Code
Review
Identity &
Access
Management
Data
Classification
Database
Secure Gateway
(Shield)
PREVENTION
POLICY MANAGEMENT
OPERATIONS
MONITORING & RESPONSE
Security Policies
& Compliance
DAR/DIM/DIU
Protection
Incident Reporting,
Detection, Response
(CIRT)
Cyber Threat
Intelligence
Data Integrity
Monitoring
PKI
Data Wiping
Cleansing
Enterprise
Right
Management
DLP
Continuous
Monitoring and
Assessment
Situational
Awareness
Threat Modeling
Focused Ops
Security Architecture
& Design
Data/Drive
Encryption
Security Dashboard
Risk Management
Penetration
Testing
Escalation
Management
Mission
Critical Assets
Security Awareness
Training
Vulnerability
Assessment
Continuous C&A
SIEM
Digital Forensics
Security SLA/SLO Reporting

Cybersecurity Data Science as a Process

Data Engineering

Advanced Analytics

Diagnostics & patterns

Establishing baselines

Predictive modelling

Anomaly detection

Behavioral insights

Triage / Validate

Remediate

Data Manager

Data Scientist

Cyber Investigator

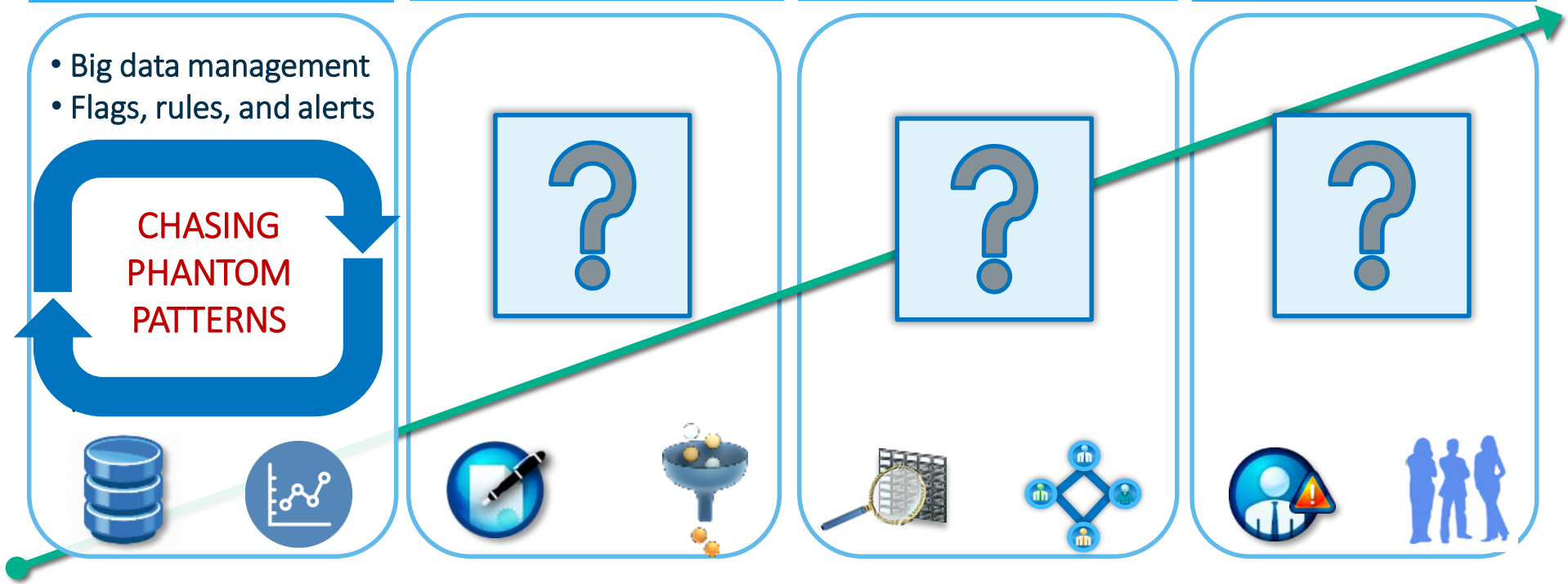
Infosec Response

Cybersecurity Analytics Maturity Curve

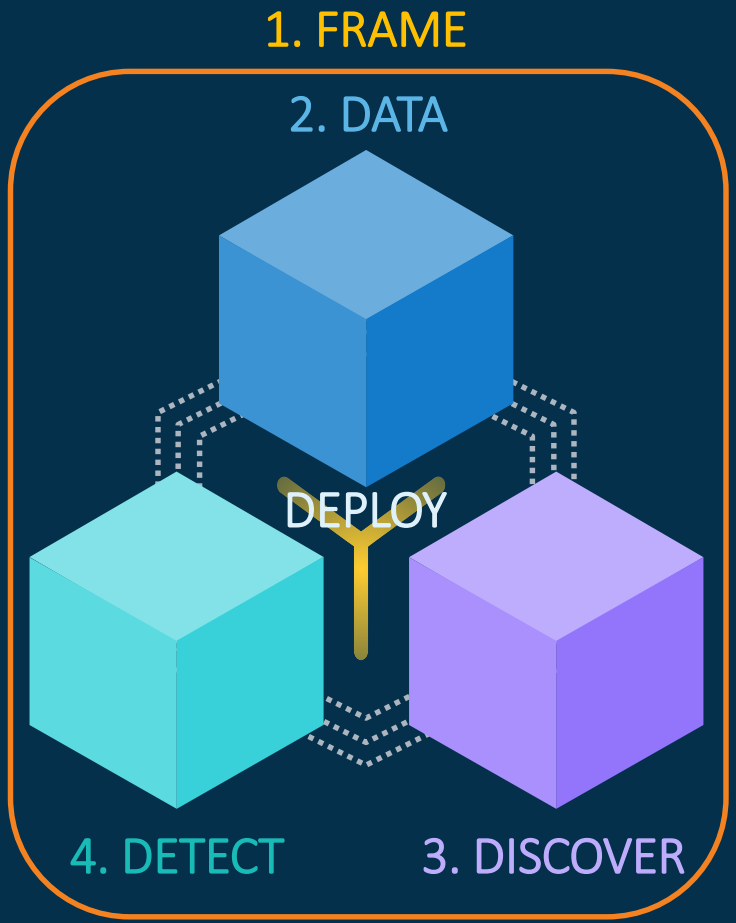
Anomaly Detection	Data-aware Investigations	Predictive Detection	Risk Optimization

- Big data management
- Flags, rules, and alerts

CHASING PHANTOM PATTERNS



CSDS Process



Cybersecurity Data Science (CSDS) Lifecycle

InfoSec Response

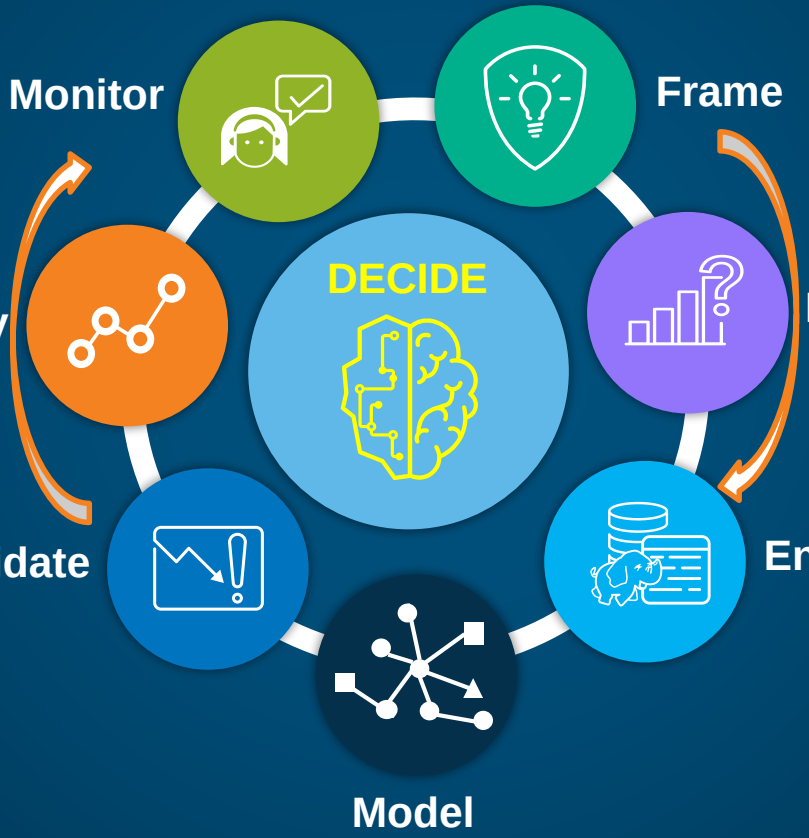
Data Manager

DETECTION

DISCOVERY

Data Scientist

Cyber Investigator



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

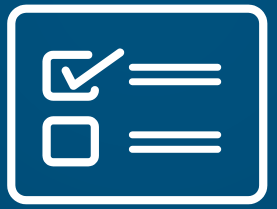
1. FRAME

Cybersecurity, Data Science, Logs

Cybersecurity Data Science (CSDS)

TOPIC
1. FRAME
2. DATA
3. DISCOVER
4. DETECT
5. DEPLOY

Learning Objectives



Objectives of Context Setting

Establishing a foundation

- Data challenges (opportunities) in cybersecurity
- Data science foundations
- Demonstration / exercises
  - Insights into the ‘data deluge’ (network analytics)
  - Log file analytics – from unstructured to structured

CSDS Process

Unified Orchestration

Cybersecurity Context

Evolving Threats

Internal Threats

Automated Attacks

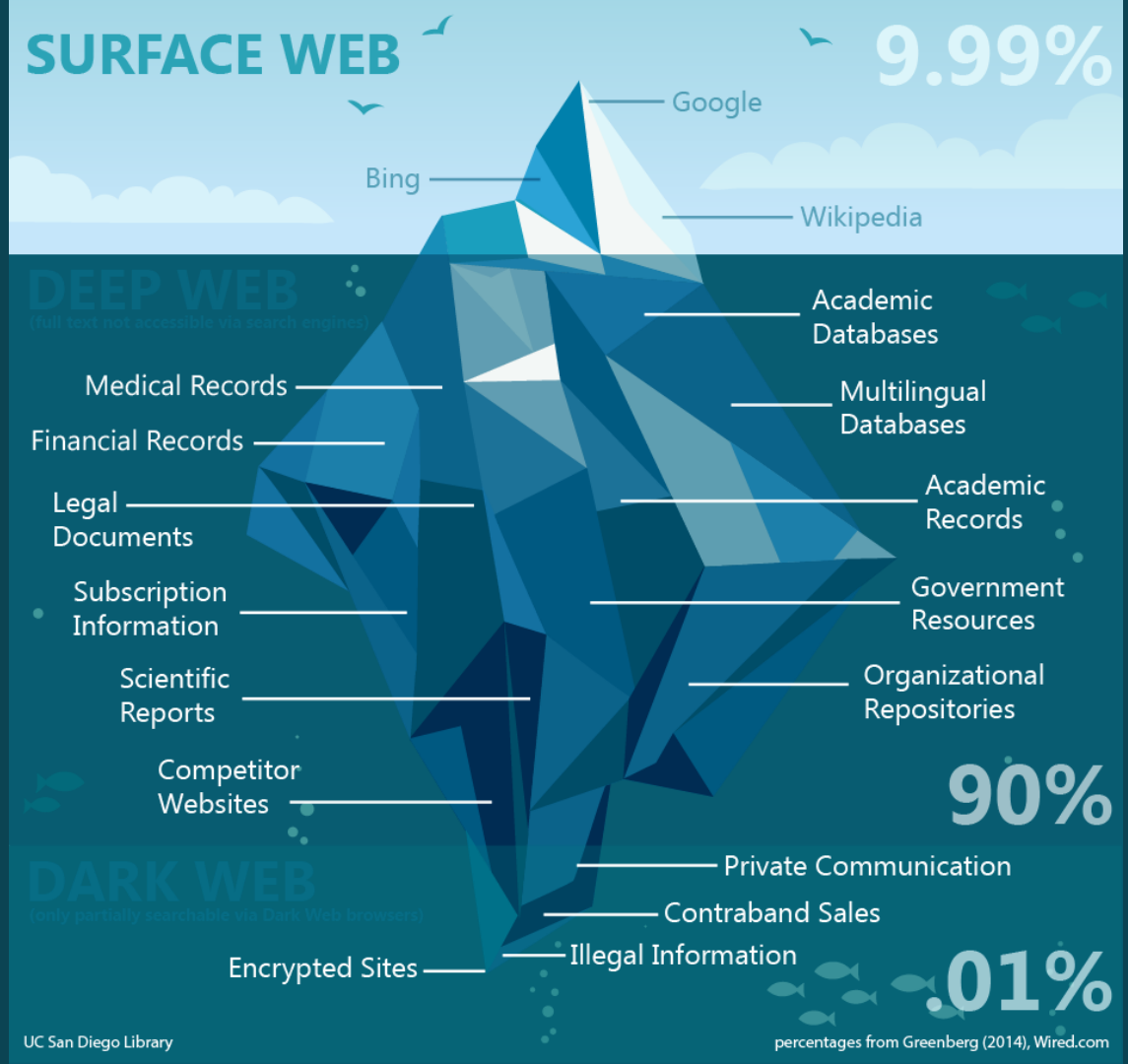
Social Engineering

Fraud-Cyber Hybrids

Ransomware & Cryptojacking

State Actors

Hidden threats want to remain hidden (in the data)


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
Subscription Information
Government Resources
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/

CRIMEWARE TOOLKITS

Cyber Threat Professional

FUD Fear, Uncertainty, Doubt

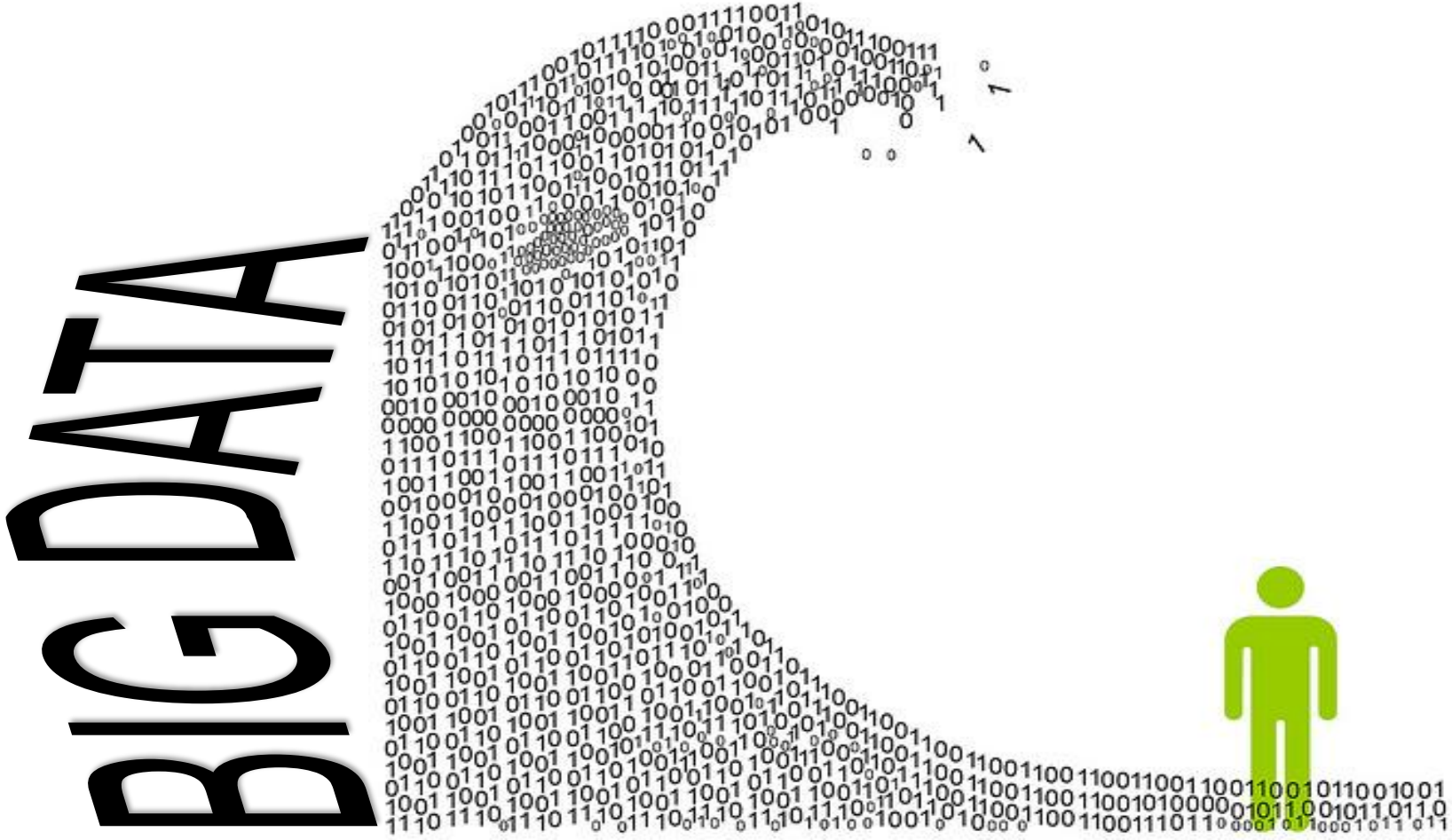
Increasing FREQUENCY, sophistication, and speed of attacks

LIST OF COGNITIVE BIASES

https://en.wikipedia.org/wiki/List_of_cognitive_biases

Ambiguity effect	Distinction bias	Information bias	Reactive devaluation
Anchoring or focalism	Dread aversion	Insensitivity to sample size	Recency illusion
Anthropocentric thinking	Dunning–Kruger effect	Interoceptive bias	Regressive bias
Anthropomorphism or personification	Duration neglect	Irrational escalation	Restraint bias
Attentional bias	Empathy gap	Law of the instrument	Rhyme as reason effect
Automation bias	Endowment effect	Less-is-better effect	Risk compensation / Peltzman effect
Availability heuristic	Exaggerated expectation	Look-elsewhere effect	Selection bias
Availability cascade	Experimenter's or expectation bias	Loss aversion	Selective perception
Backfire effect	Focusing effect	Mere exposure effect	Semmelweis reflex
Bandwagon effect	Forer effect or Barnum effect	Money illusion	Social comparison bias
Base rate fallacy or Base rate neglect	Form function attribution bias	Moral credential effect	Social desirability bias
Belief bias	Framing effect	Negativity bias or Negativity effect	Status quo bias
Ben Franklin effect	Frequency illusion or Baader–Meinhof effect	Neglect of probability	Stereotyping
Berkson's paradox	Functional fixedness	Normalcy bias	Subadditivity effect
Bias blind spot	Gambler's fallacy	Not invented here	Subjective validation
Choice-supportive bias	Groupthink	Observer-expectancy effect	Surrogation
Clustering illusion	Hard–easy effect	Omission bias	Survivorship bias
Confirmation bias	Hindsight bias	Optimism bias	Time-saving bias
Congruence bias	Hostile attribution bias	Ostrich effect	Third-person effect
Conjunction fallacy	Hot-hand fallacy	Outcome bias	Parkinson's law of triviality
Conservatism (belief revision)	Hyperbolic discounting	Overconfidence effect	Unit bias
Continued influence effect	Identifiable victim effect	Pareidolia	Weber–Fechner law
Contrast effect	IKEA effect	Pessimism bias	Well travelled road effect
Courtesy bias	Illicit transference	Planning fallacy	Zero-risk bias
Curse of knowledge	Illusion of control	Post-purchase rationalization	
Declinism	Illusion of validity	Present bias	
Decoy effect	Illusory correlation	Pro-innovation bias	
Default effect	Illusory truth effect	Projection bias	
Denomination effect	Impact bias	Pseudocertainty effect	
Disposition effect	Implicit association	Reactance	

‘Drowning’ in Data Lakes



Network Traffic

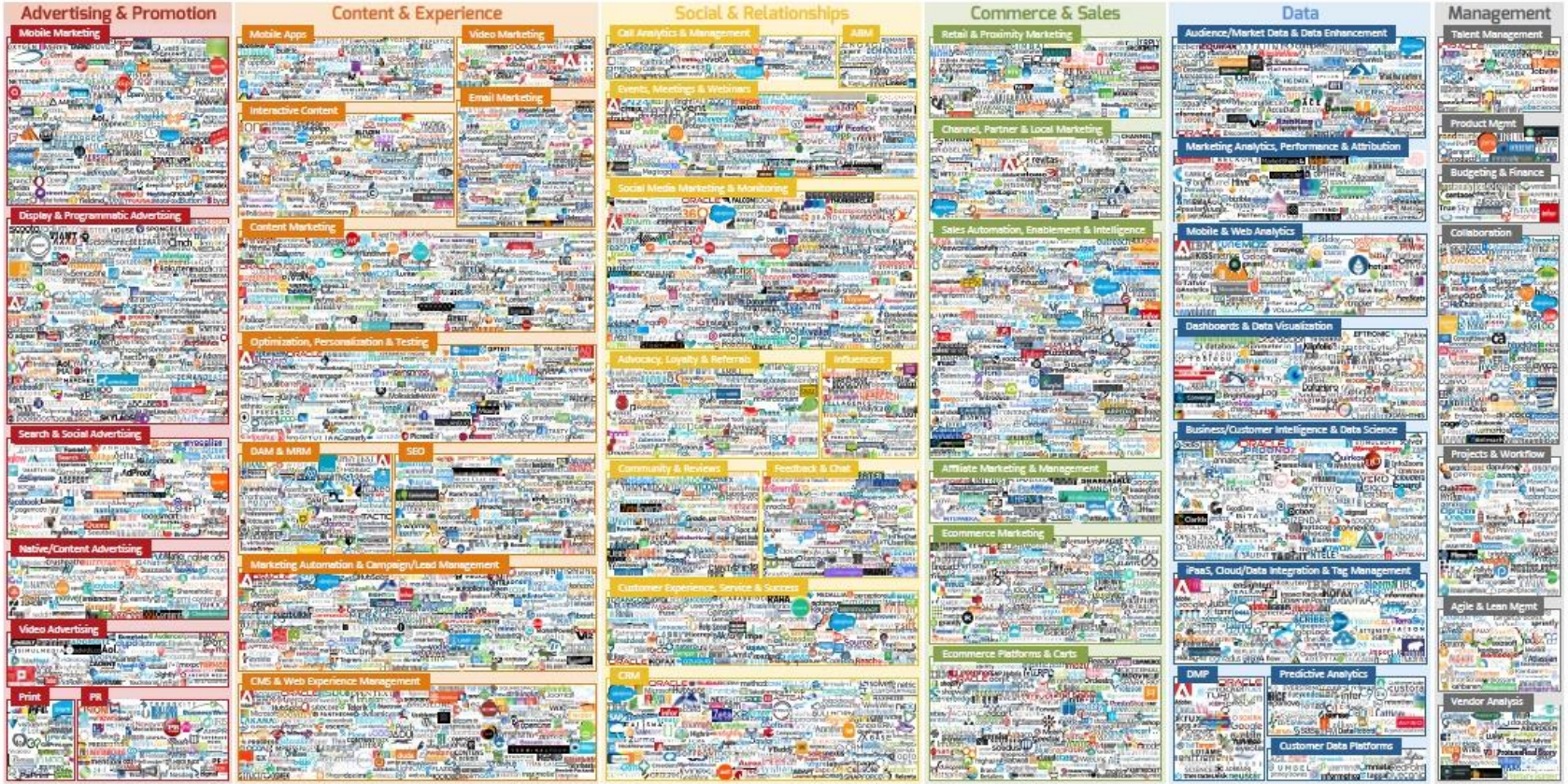
Tracking network traffic on a single device

Exponential Growth in Network Traffic

Using cookies and other tracking techniques, many "normal" web sites aggressively access and store our:

- Device details
- IP address
- Present geolocation
- Changing physical locations
- Browser history
- Online purchases
- Profile information
- Buying behavior
- Personal finances
- Religious beliefs
- Political affiliations
- Health concerns and problems
- Camera / photos / microphone

Marketing Analytics Ecosystem (>4000 companies)



Sources: CabinetM (http://cabinetm.com), Capterra, G2 Crowd, Google, Growthverse, LUMA Partners, Siffery, TrustRadius, VBProfiles — see http://chiefmartec.com/2016/03/marketing-technology-supergraphic-2016/ for details.

Created by Scott Brinker (@chiefmartec).

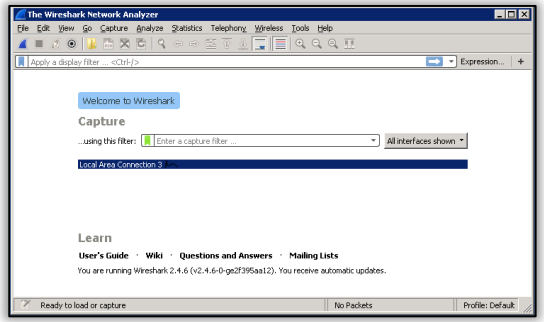
March 2016

chiefmartec.com Marketing Technology Landscape

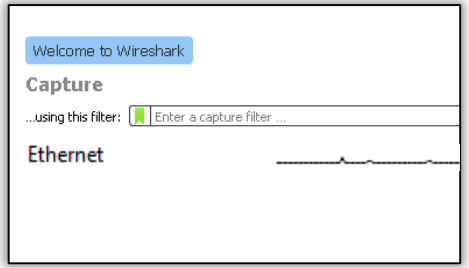
Demonstration / Exercise

Network Traffic from Web Browsing

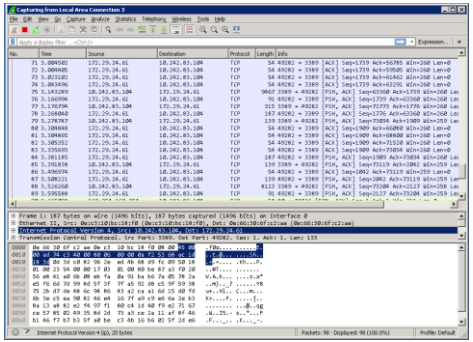
1. Open ‘Wireshark’



2. Click ‘*Ethernet*’



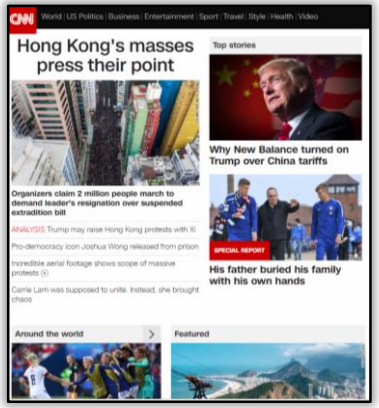
3. Observe results



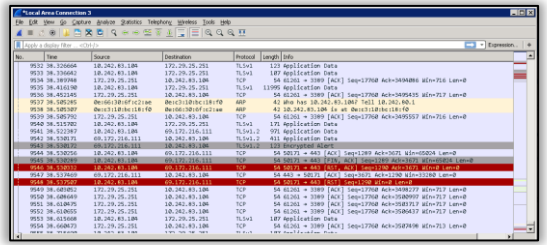
Demonstration / Exercise

Network Traffic from Simple Web Browsing

4. Open ‘Internet Explorer’ (next to Wireshark)
5. Go to a popular news website



6. Monitor Wireshark – what do you see?
7. Sort on destination
7. View ‘red’ & ‘**black**’ events



See codes View => Coloring Rules

Note ‘Destination’ IP

8. Go to http://ip-lookup.net

Where and what are these IPs?

Wireshark Coloring Rules (default)

Wireshark · Coloring Rules Default

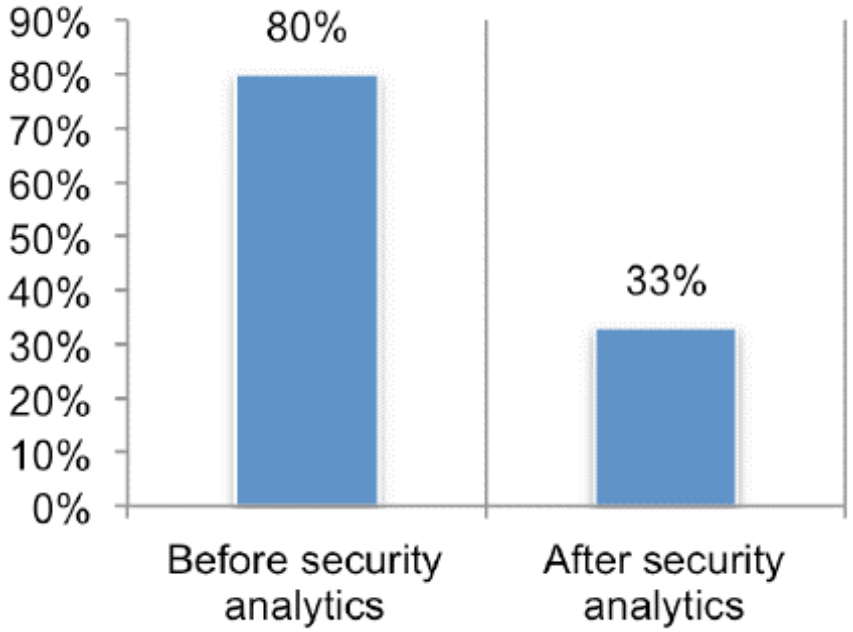
Name	Filter
Bad TCP	tcp.analysis.flags && !tcp.analysis.window_update
HSRP State Change	hsrp.state != 8 && hsrp.state != 16
Spanning Tree Topology Change	stp.type == 0x80
OSPF State Change	ospf.msg != 1
ICMP errors	icmp.type eq 3 || icmp.type eq 4 || icmp.type eq 5 || icmp.type eq 11 || icmpv6.type eq 1 || icmpv6.type eq 2 || icmpv6.type eq 3 || icmpv6.type eq
ARP	arp
ICMP	icmp || icmpv6
TCP RST	tcp.flags.reset eq 1
SCTP ABORT	sctp.chunk_type eq ABORT
TTL low or unexpected	(! ip.dst == 224.0.0.0/4 && ip.ttl < 5 && !pim && !ospf) || (ip.dst == 224.0.0.0/24 && ip.dst != 224.0.0.251 && ip.ttl != 1 && !(vrrp || carp))
Checksum Errors	eth.fcs.status=="Bad" || ip.checksum.status=="Bad" || tcp.checksum.status=="Bad" || udp.checksum.status=="Bad" || sctp.checksum.status==
SMB	smb || nbss || nbns || netbios
HTTP	http || tcp.port == 80 || http2
DCERPC	dcerpc
Routing	hsrp || eigrp || ospf || bgp || cdp || vrrp || carp || gvrp || igmp || ismp
TCP SYN/FIN	tcp.flags & 0x02 || tcp.flags.fin == 1
TCP	tcp
UDP	udp
Broadcast	eth[0] & 1
System Event	systemd_journal || sysdig


C/B
A la Sainte
CORSET
TURKISH TROPHIES
10 for 10
Why Pay More?
WILLIAMS
TALCUM POWD
FOLLIES OF THE DAY
A GENTLEMAN
MISSISSIPPI

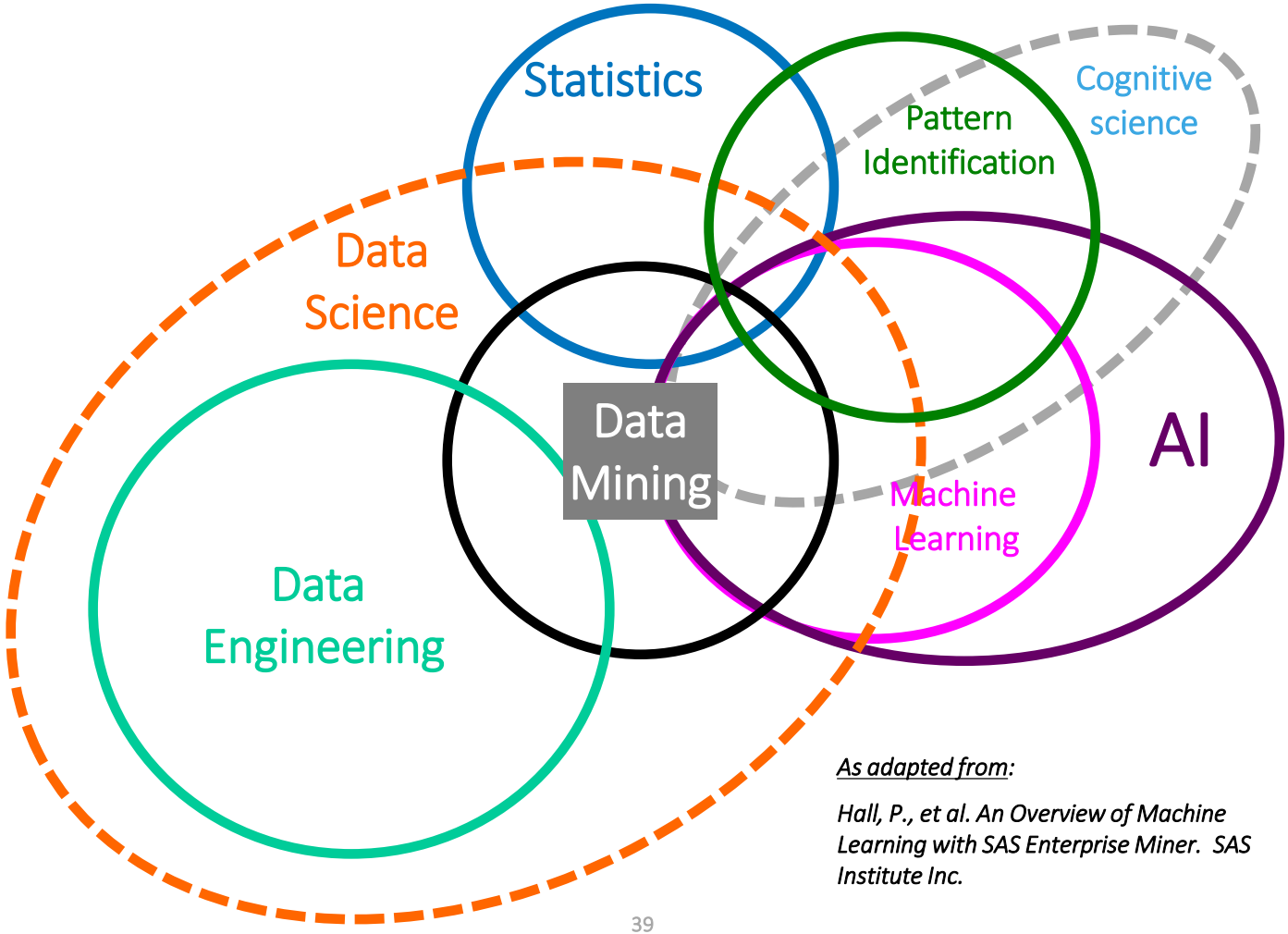
Data Science Foundations

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

Level of difficulty in reducing false alerts*



** Survey of 621 global IT security practitioners*

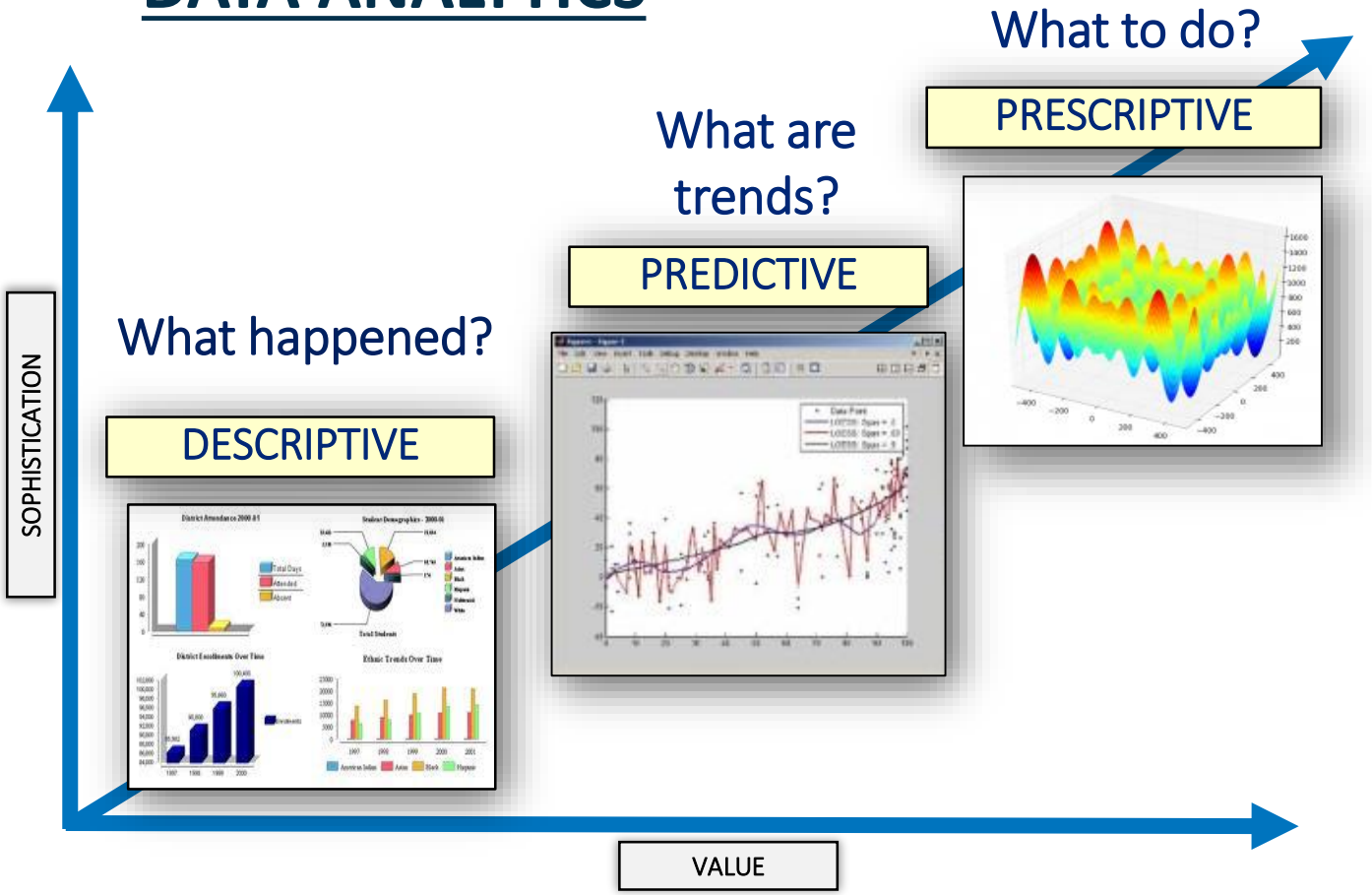


As adapted from:

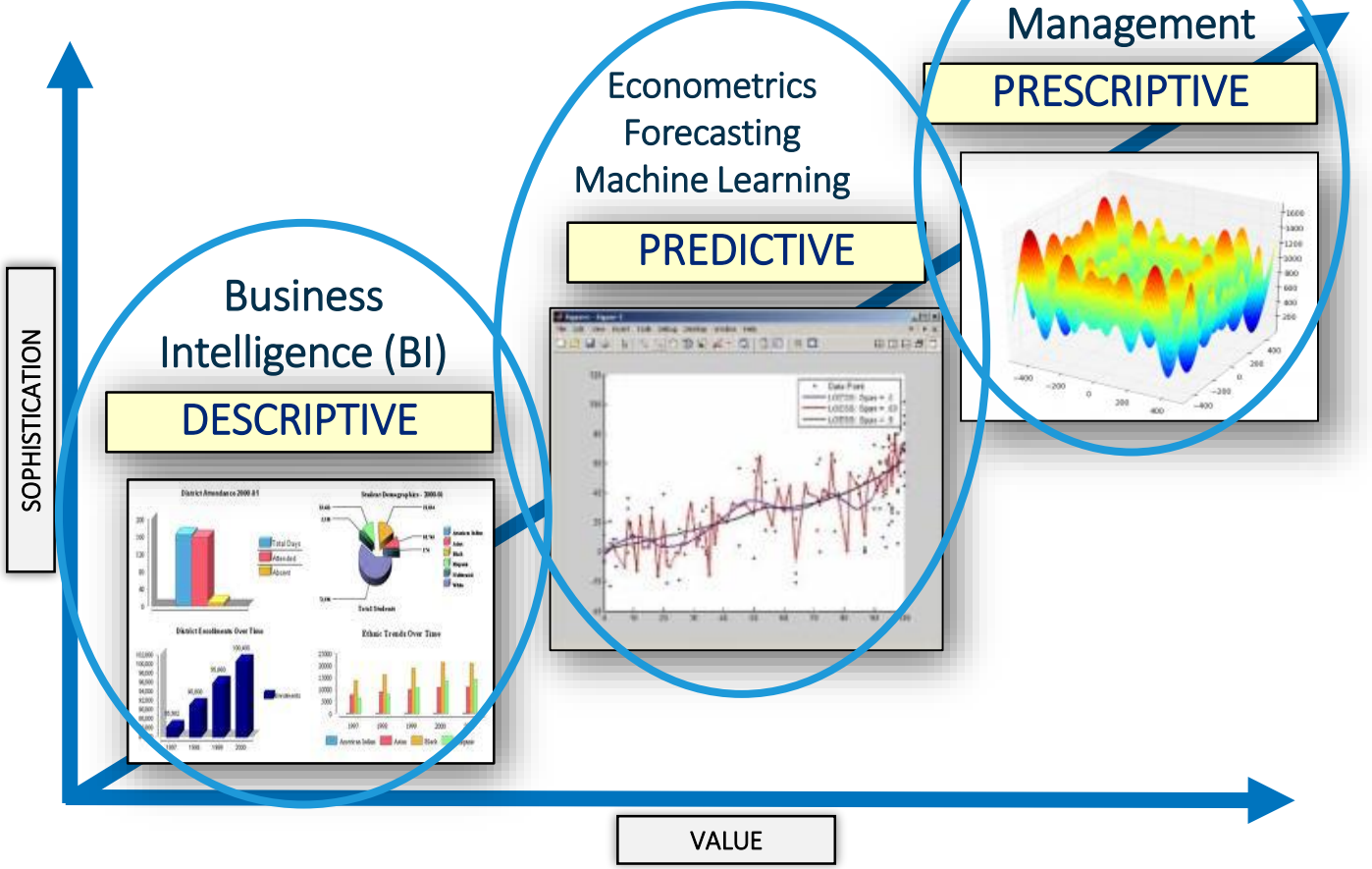
Hall, P., et al. An Overview of Machine Learning with SAS Enterprise Miner. SAS Institute Inc.

39

DATA ANALYTICS



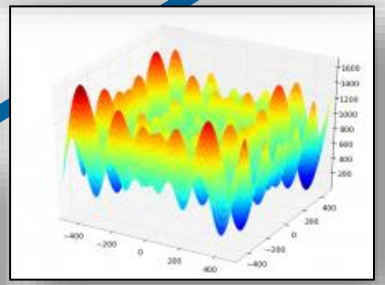
DATA ANALYTICS



DATA ANALYTICS

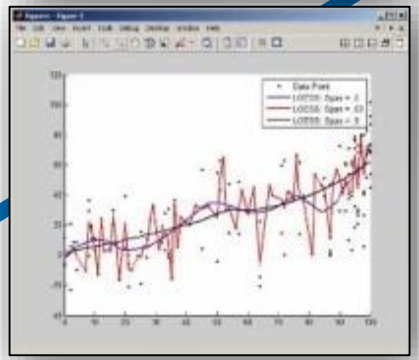
How can we optimize remediation from focused alerts?

PRESCRIPTIVE



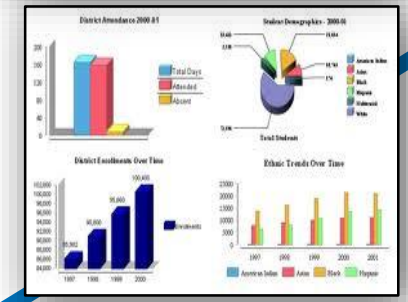
What is the 'normal' behaviour of particular users or devices?

PREDICTIVE



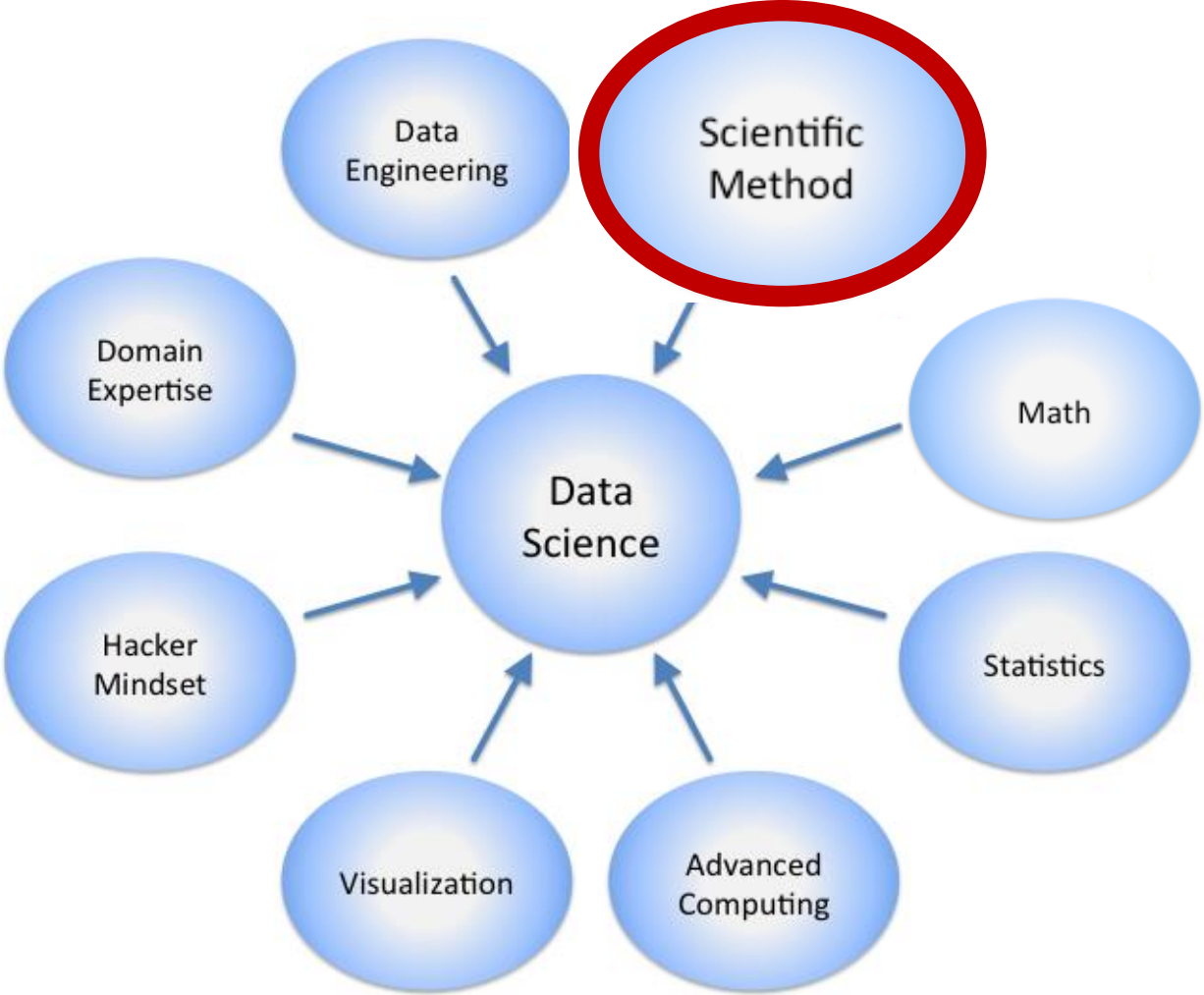
What are trends and patterns in network and device usage?

DESCRIPTIVE

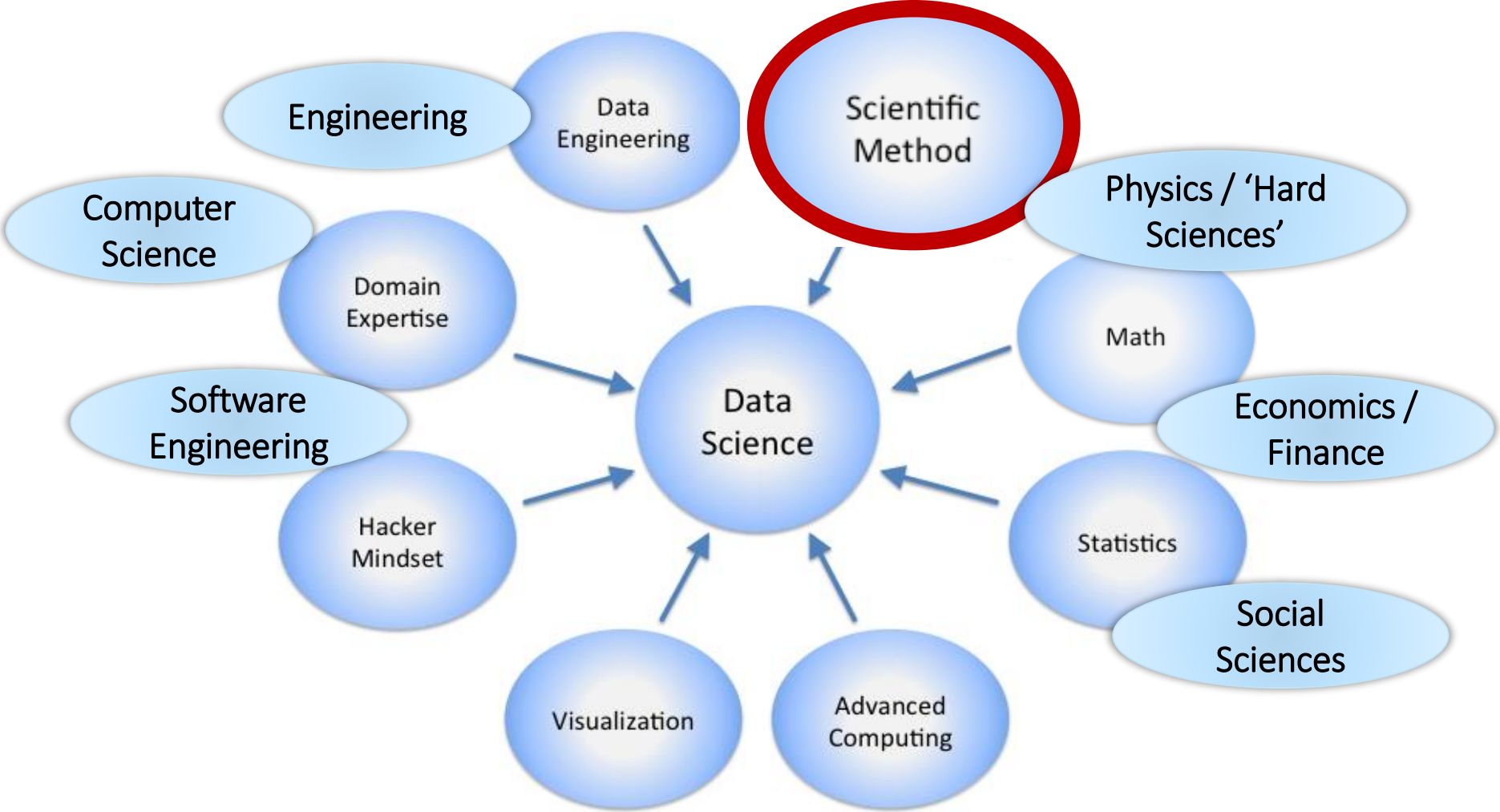


SOPHISTICATION

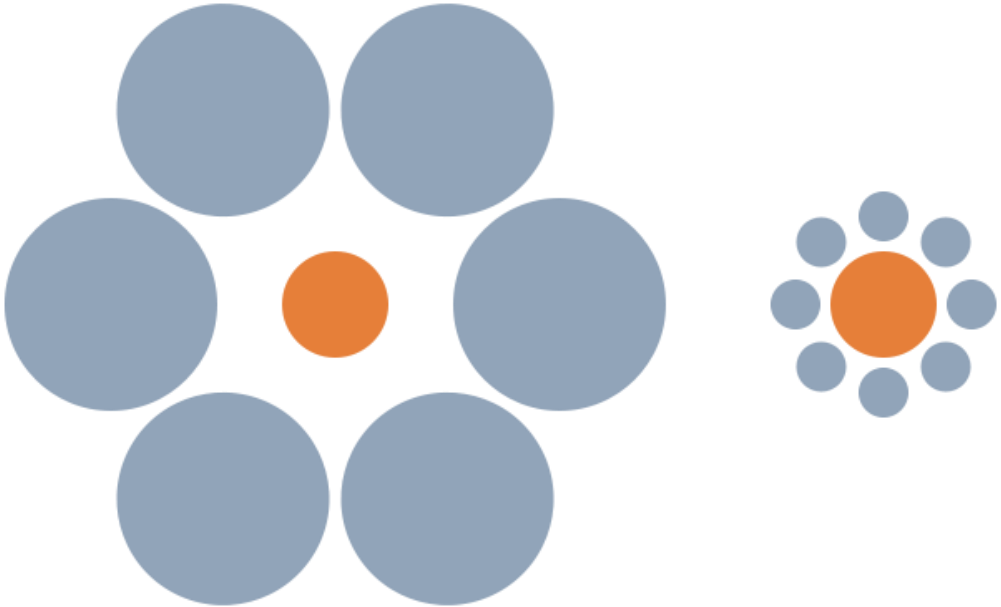
VALUE



Calvin.Andrus (2012) Depicts a mash-up of disciplines from which Data Science is derived http://en.wikipedia.org/wiki/File:DataScienceDisciplines.png

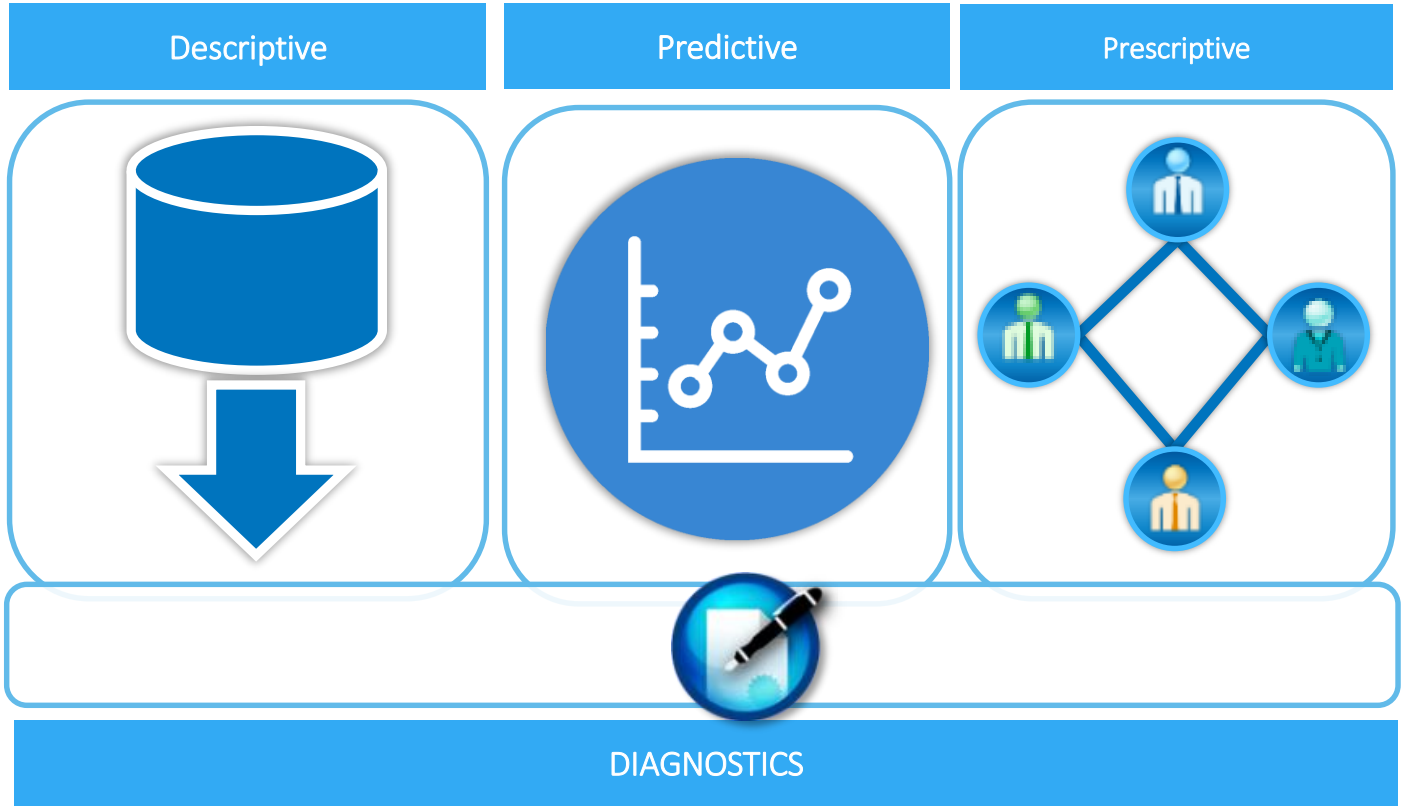


Calvin.Andrus (2012) Depicts a mash-up of disciplines from which Data Science is derived http://en.wikipedia.org/wiki/File:DataScienceDisciplines.png



Scientific test…

VALIDATE WITH DIAGNOSTICS



i.e. statistical tests, causal and explanatory models, experiments, validation, model performance, etc.

Data Science Methods

How can we support experts with visualizations?

What are the underlying human factors?

SOPHISTICATION

VALUE

Data visualization

Understanding Patterns

Optimizing Systems

SEMANTIC

PRESCRIPTIVE

Understanding Context & Meaning

PREDICTIVE

Validating Factors & Causes

DIAGNOSTICS

Forecasting & Probabilities

How do we substantiate our assumptions and hypothesizes?

DESCRIPTIVE

DATA QUALITY

Is data quality reliable enough? What are limitations?

Business Intelligence

CSDS: Cybersecurity Data Science

Replacing rules with machine learning to reduce false alerts

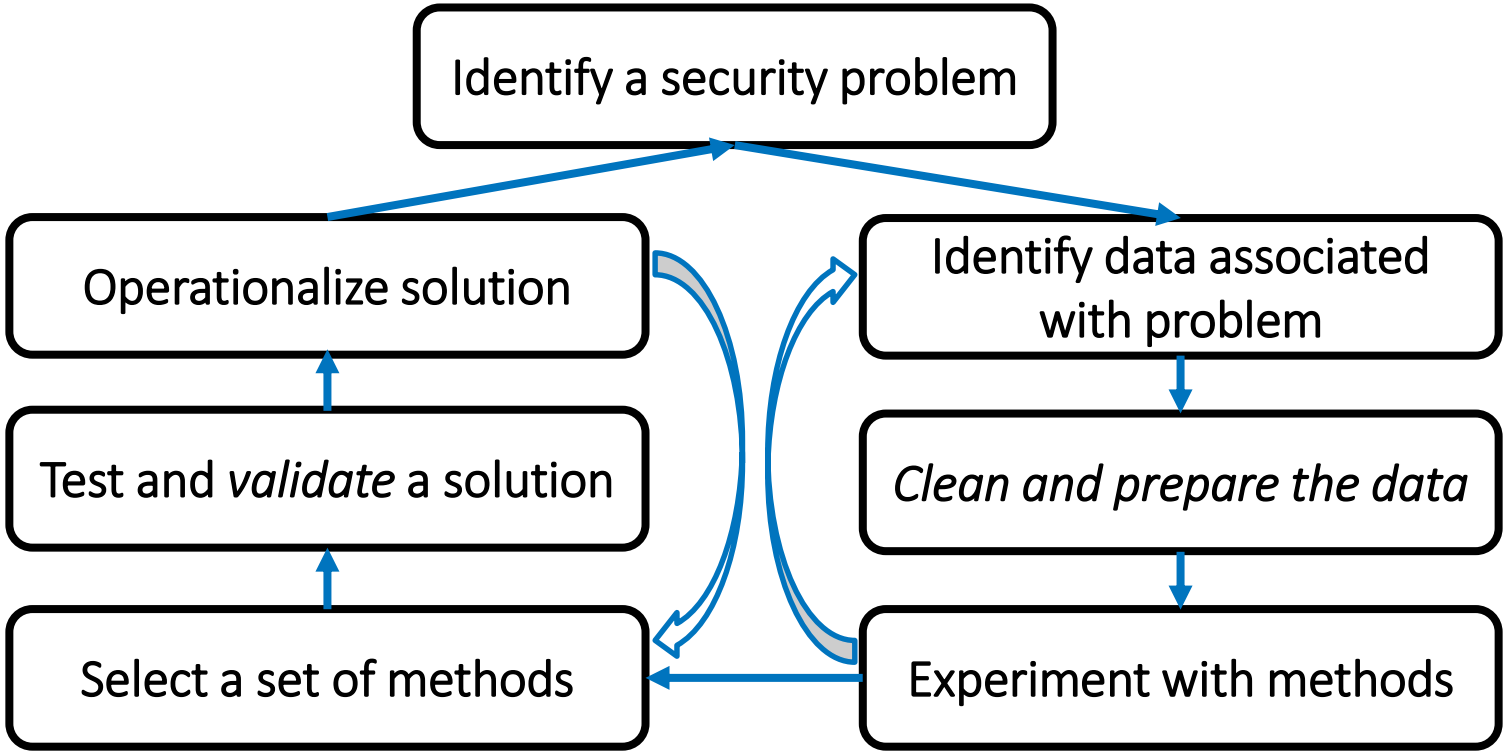
Moving to real time detection and decisioning

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 Data Science: Typical Sequence



Data Foundations: Log File Analytics

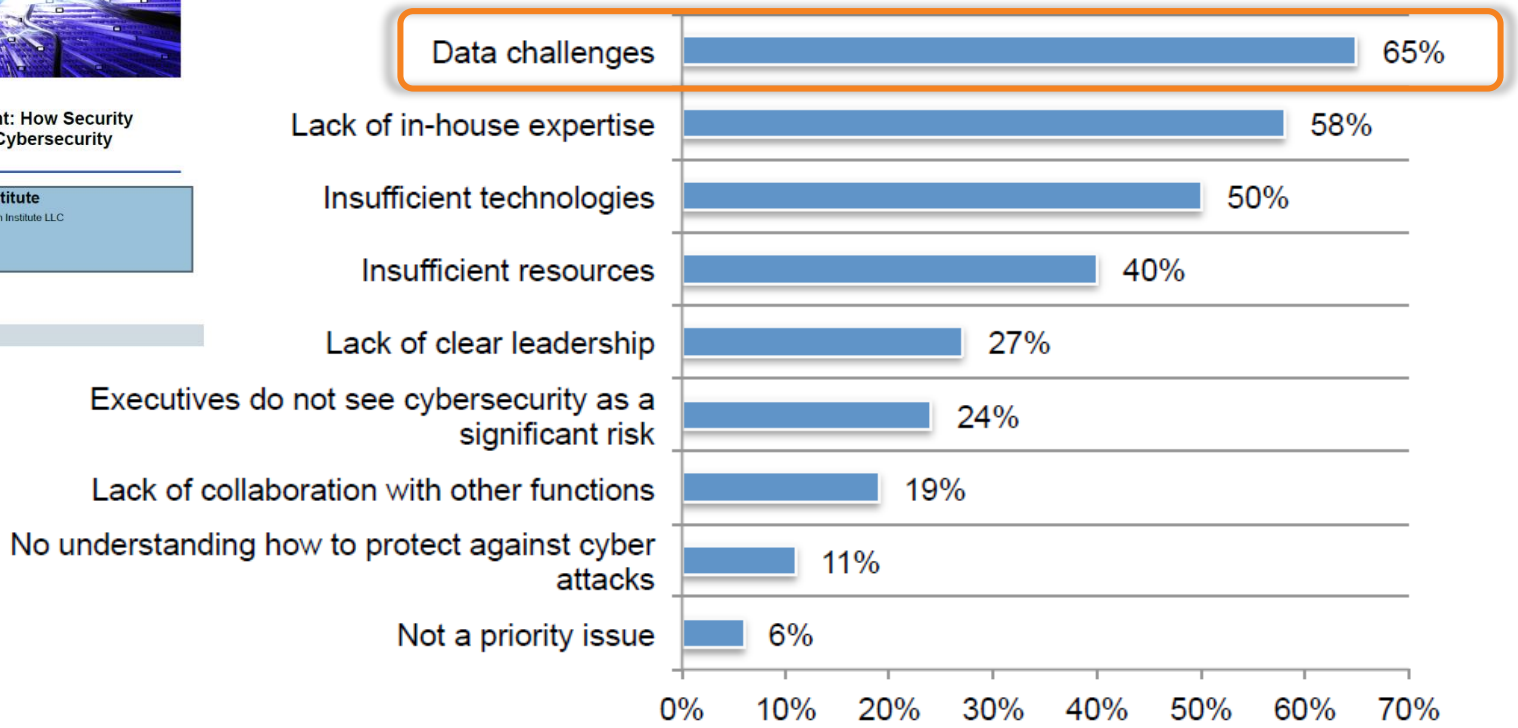
The devil is in the data

Challenges preventing successful use of cybersecurity analytics*

When Seconds Count: How Security Analytics Improves Cybersecurity Defenses

Sponsored by SAS Institute
Independently conducted by Ponemon Institute LLC
Publication Date: January 2017

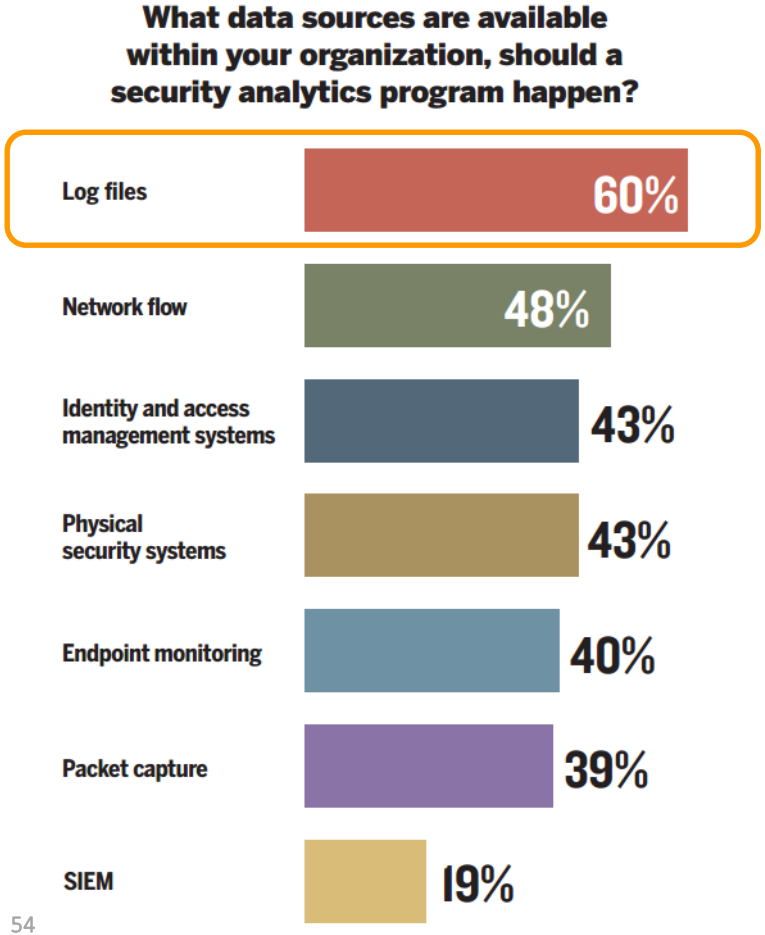
Ponemon Institute® Research Report



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

** Survey of 621 global IT security practitioners*

Log files are the most common source of cybersecurity data…



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

Security Log Data Sources

Operating systems
- Windows & registry events
- Unix logs

Data access
- File system
- ODBC
- RMDB (i.e. Oracle, SQL Server)

Authentication and Authorization Reports
- Active Directory Services
- Kerberos
- Proxy logs

Network logs
- Firewall logs
- Router
- Traffic flows / packet captures

- Web server
  - IIS W3C Logs, Apache
  - HTTP Error Logs, URL Scans
- Resource access and performance
- Physical security access
- Malware and endpoint activity
- Failure and critical error reports
- Specialized
  - Cloud applications (i.e. SaaS)
  - *Specialized systems / applications (i.e. ERP, thin client, specialized or home-grown)*

SOURCE: Marty, Raffael. Applied Security Visualization. Pearson Education.

Use Cases: Security Log File Analytics

- **User-entity behavioral analytics (UEBA)***
- Performance insights
- Optimization of resources
- Asset tracking
- Refining focused alerts
- Refining risk indicator metrics
- Enriching SIEM or other repositories

- **Data / asset protection**
- Preventive insights
- Identify / attribute attacks
- Incident response
- Incident root-cause analysis
- Compliance / risk reporting
- CISO dashboard

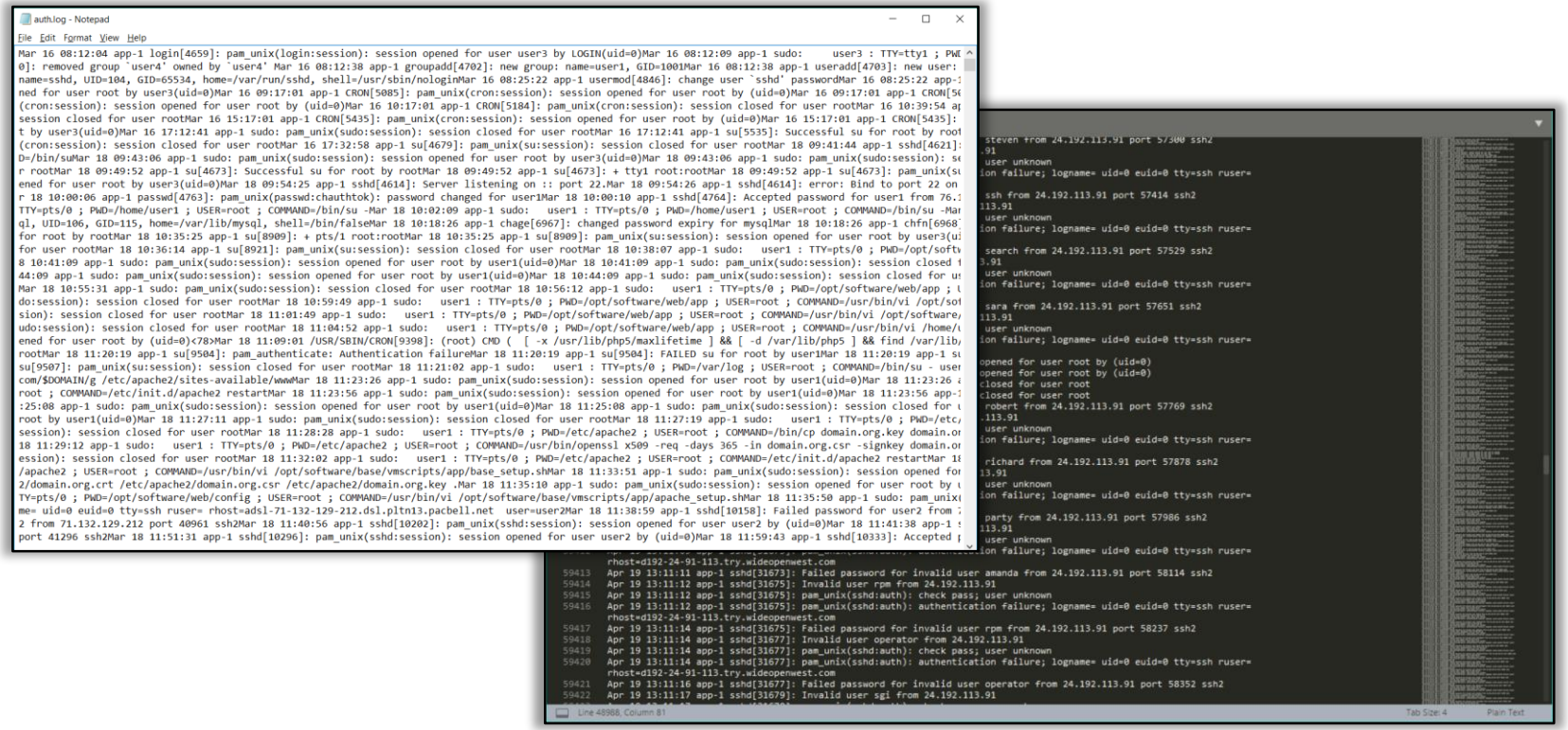
* *Global Behavioral Analytics Market $3.5 billion projected market value by 2024*

Diving into Cybersecurity Log Data

This exercise illustrates an example of examining cybersecurity log data with a variety of tools.

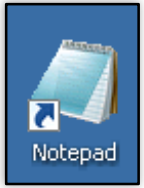
Diving Right In!

EXERCISE: Assessing Raw Log Data



Demonstration / Exercise

EXERCISE: Assessing Raw Log Data

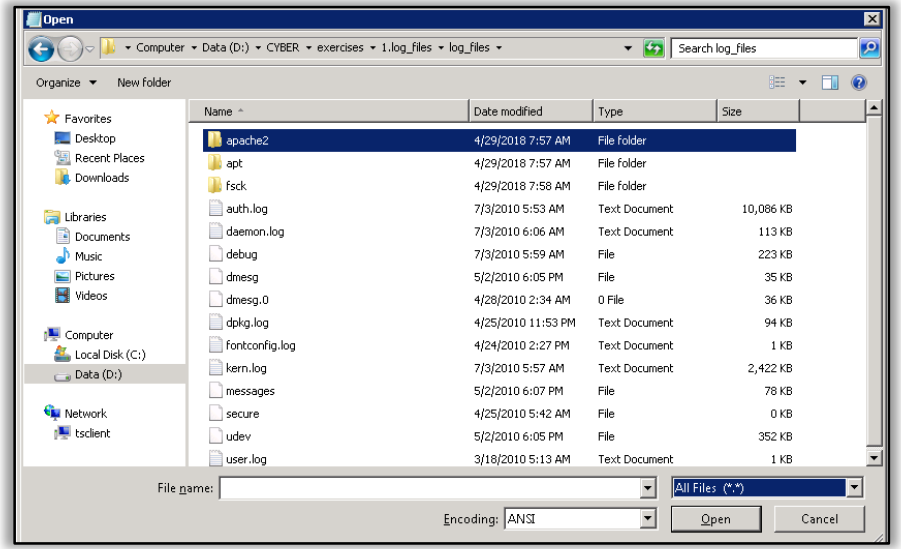


1. Open ‘Notepad’

2. File => Open

3. Select ‘All Files (*.*)’ (lower right)

3. D: => @CYBER => 1.FRAME => B.Log_Files => log_files => ***auth.log***



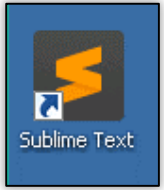
auth.log - Notepad

File Edit Format View Help

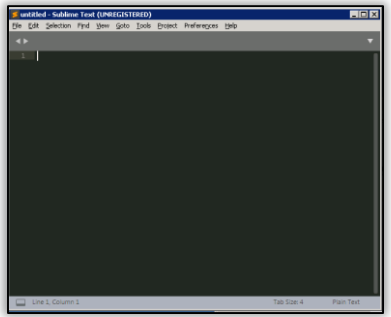
```
Mar 16 08:12:04 app-1 login[4659]: pam_unix(login:session): session opened for user user3 by LOGIN(uid=0)Mar 16 08:12:09 app-1 sudo:     user3 : TTY=tty1 ;
0]: removed group `user4' owned by `user4' Mar 16 08:12:38 app-1 groupadd[4702]: new group: name=user1, GID=1001Mar 16 08:12:38 app-1 useradd[4703]: new use
name=sshd, UID=104, GID=65534, home=/var/run/sshd, shell=/usr/sbin/nologinMar 16 08:25:22 app-1 usermod[4846]: change user `sshd' passwordMar 16 08:25:22 ap
ned for user root by user3(uid=0)Mar 16 09:17:01 app-1 CRON[5085]: pam_unix(cron:session): session opened for user root by (uid=0)Mar 16 09:17:01 app-1 CRO
(cron:session): session opened for user root by (uid=0)Mar 16 10:17:01 app-1 CRON[5184]: pam_unix(cron:session): session closed for user rootMar 16 10:39:5
session closed for user rootMar 16 15:17:01 app-1 CRON[5435]: pam_unix(cron:session): session opened for user root by (uid=0)Mar 16 15:17:01 app-1 CRON[543
t by user3(uid=0)Mar 16 17:12:41 app-1 sudo: pam_unix(sudo:session): session closed for user rootMar 16 17:12:41 app-1 su[5535]: Successful su for root by
(cron:session): session closed for user rootMar 16 17:32:58 app-1 su[4679]: pam_unix(su:session): session closed for user rootMar 18 09:41:44 app-1 sshd[46
D=/bin/suMar 18 09:43:06 app-1 sudo: pam_unix(sudo:session): session opened for user root by user3(uid=0)Mar 18 09:43:06 app-1 sudo: pam_unix(sudo:session)
r rootMar 18 09:49:52 app-1 su[4673]: Successful su for root by rootMar 18 09:49:52 app-1 su[4673]: + tty1 root:rootMar 18 09:49:52 app-1 su[4673]: pam_uni
ened for user root by user3(uid=0)Mar 18 09:54:25 app-1 sshd[4614]: Server listening on :: port 22.Mar 18 09:54:26 app-1 sshd[4614]: error: Bind to port 22
r 18 10:00:06 app-1 passwd[4763]: pam_unix(passwd:chauthtok): password changed for user1Mar 18 10:00:10 app-1 sshd[4764]: Accepted password for user1 from
TTY=pts/0 ; PWD=/home/user1 ; USER=root ; COMMAND=/bin/su -Mar 18 10:02:09 app-1 sudo:   user1 : TTY=pts/0 ; PWD=/home/user1 ; USER=root ; COMMAND=/bin/su
ql, UID=106, GID=115, home=/var/lib/mysql, shell=/bin/falseMar 18 10:18:26 app-1 chage[6967]: changed password expiry for mysqlMar 18 10:18:26 app-1 chfn[6
for root by rootMar 18 10:35:25 app-1 su[8909]: + pts/1 root:rootMar 18 10:35:25 app-1 su[8909]: pam_unix(su:session): session opened for user root by user
for user rootMar 18 10:36:14 app-1 su[8921]: pam_unix(su:session): session closed for user rootMar 18 10:38:07 app-1 sudo:   user1 : TTY=pts/0 ; PWD=/opt/s
8 10:41:09 app-1 sudo: pam_unix(sudo:session): session opened for user root by user1(uid=0)Mar 18 10:41:09 app-1 sudo: pam_unix(sudo:session): session close
44:09 app-1 sudo: pam_unix(sudo:session): session opened for user root by user1(uid=0)Mar 18 10:44:09 app-1 sudo: pam_unix(sudo:session): session closed fo
Mar 18 10:55:31 app-1 sudo: pam_unix(sudo:session): session closed for user rootMar 18 10:56:12 app-1 sudo:   user1 : TTY=pts/0 ; PWD=/opt/software/web/app
do:session): session closed for user rootMar 18 10:59:49 app-1 sudo:   user1 : TTY=pts/0 ; PWD=/opt/software/web/app ; USER=root ; COMMAND=/usr/bin/vi /opt/
sion): session closed for user rootMar 18 11:01:49 app-1 sudo:   user1 : TTY=pts/0 ; PWD=/opt/software/web/app ; USER=root ; COMMAND=/usr/bin/vi /opt/softwa
udo:session): session closed for user rootMar 18 11:04:52 app-1 sudo:   user1 : TTY=pts/0 ; PWD=/opt/software/web/app ; USER=root ; COMMAND=/usr/bin/vi /hom
ened for user root by (uid=0)<78>Mar 18 11:09:01 /USR/SBIN/CRON[9398]: (root) CMD (  [ -x /usr/lib/php5/maxlifetime ] && [ -d /var/lib/php5 ] && find /var/
rootMar 18 11:20:19 app-1 su[9504]: pam_authenticate: Authentication failureMar 18 11:20:19 app-1 su[9504]: FAILED su for root by user1Mar 18 11:20:19 app-
su[9507]: pam_unix(su:session): session closed for user rootMar 18 11:21:02 app-1 sudo:   user1 : TTY=pts/0 ; PWD=/var/log ; USER=root ; COMMAND=/bin/su -
com/$DOMAIN/g /etc/apache2/sites-available/wwwMar 18 11:23:26 app-1 sudo: pam_unix(sudo:session): session opened for user root by user1(uid=0)Mar 18 11:23:
root ; COMMAND=/etc/init.d/apache2 restartMar 18 11:23:56 app-1 sudo: pam_unix(sudo:session): session opened for user root by user1(uid=0)Mar 18 11:23:56 a
:25:08 app-1 sudo: pam_unix(sudo:session): session opened for user root by user1(uid=0)Mar 18 11:25:08 app-1 sudo: pam_unix(sudo:session): session closed f
root by user1(uid=0)Mar 18 11:27:11 app-1 sudo: pam_unix(sudo:session): session closed for user rootMar 18 11:27:19 app-1 sudo:   user1 : TTY=pts/0 ; PWD=/
session): session closed for user rootMar 18 11:28:28 app-1 sudo:   user1 : TTY=pts/0 ; PWD=/etc/apache2 ; USER=root ; COMMAND=/bin/cp domain.org.key domai
18 11:29:12 app-1 sudo:   user1 : TTY=pts/0 ; PWD=/etc/apache2 ; USER=root ; COMMAND=/usr/bin/openssl x509 -req -days 365 -in domain.org.csr -signkey domai
ession): session closed for user rootMar 18 11:32:02 app-1 sudo:   user1 : TTY=pts/0 ; PWD=/etc/apache2 ; USER=root ; COMMAND=/etc/init.d/apache2 restartMa
/apache2 ; USER=root ; COMMAND=/usr/bin/vi /opt/software/base/vmscripts/app/base_setup.shMar 18 11:33:51 app-1 sudo: pam_unix(sudo:session): session opened
2/domain.org.crt /etc/apache2/domain.org.csr /etc/apache2/domain.org.key .Mar 18 11:35:10 app-1 sudo: pam_unix(sudo:session): session opened for user root
TY=pts/0 ; PWD=/opt/software/web/config ; USER=root ; COMMAND=/usr/bin/vi /opt/software/base/vmscripts/app/apache_setup.shMar 18 11:35:50 app-1 sudo: pam_u
```

Demonstration / Exercise

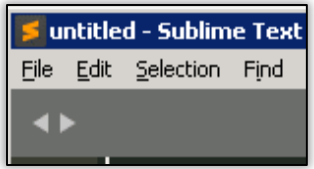
EXERCISE: Assessing Semi-Structured Log Data



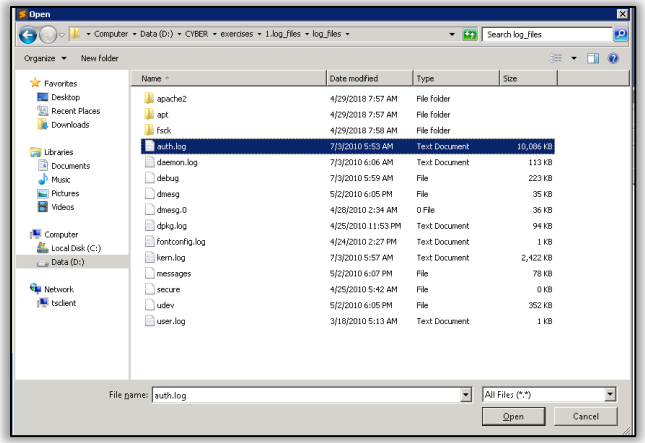
1. Open ‘Sublime Text’

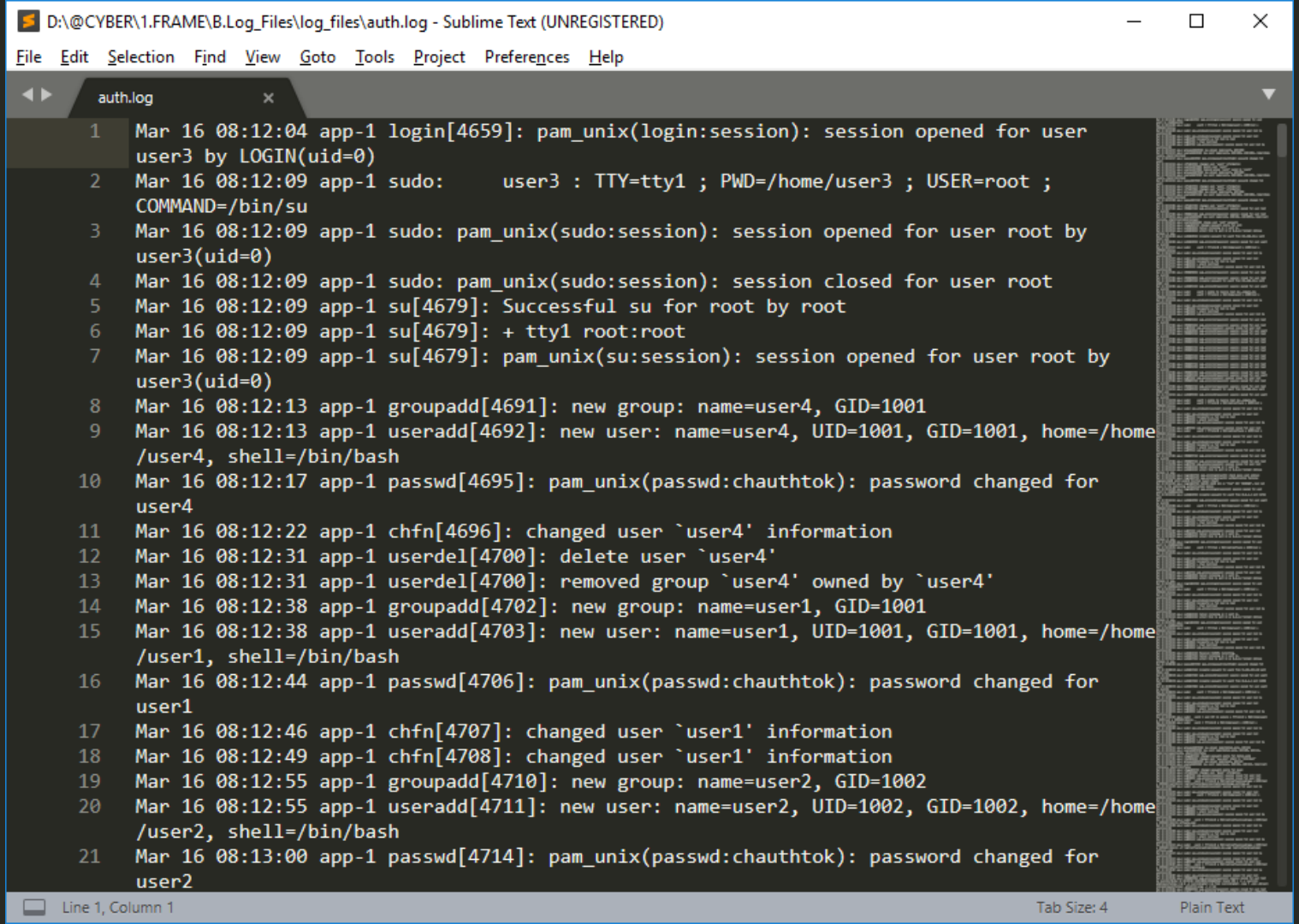


2. File => Open File



3. D: => @CYBER => 1.FRAME => B.Log_Files => log_files => ***auth.log***




D:\@CYBER\1.FRAME\B.Log_Files\log_files\auth.log - Sublime Text (UNREGISTERED)
File Edit Selection Find View Goto Tools Project Preferences Help
auth.log
1 Mar 16 08:12:04 app-1 login[4659]: pam_unix(login:session): session opened for user user3 by LOGIN(uid=0)
2 Mar 16 08:12:09 app-1 sudo: user3 : TTY=tty1 ; PWD=/home/user3 ; USER=root ; COMMAND=/bin/su
3 Mar 16 08:12:09 app-1 sudo: pam_unix(sudo:session): session opened for user root by user3(uid=0)
4 Mar 16 08:12:09 app-1 sudo: pam_unix(sudo:session): session closed for user root
5 Mar 16 08:12:09 app-1 su[4679]: Successful su for root by root
6 Mar 16 08:12:09 app-1 su[4679]: + tty1 root:root
7 Mar 16 08:12:09 app-1 su[4679]: pam_unix(su:session): session opened for user root by user3(uid=0)
8 Mar 16 08:12:13 app-1 groupadd[4691]: new group: name=user4, GID=1001
9 Mar 16 08:12:13 app-1 useradd[4692]: new user: name=user4, UID=1001, GID=1001, home=/home/user4, shell=/bin/bash
10 Mar 16 08:12:17 app-1 passwd[4695]: pam_unix(passwd:chauthtok): password changed for user4
11 Mar 16 08:12:22 app-1 chfn[4696]: changed user `user4' information
12 Mar 16 08:12:31 app-1 userdel[4700]: delete user `user4'
13 Mar 16 08:12:31 app-1 userdel[4700]: removed group `user4' owned by `user4'
14 Mar 16 08:12:38 app-1 groupadd[4702]: new group: name=user1, GID=1001
15 Mar 16 08:12:38 app-1 useradd[4703]: new user: name=user1, UID=1001, GID=1001, home=/home/user1, shell=/bin/bash
16 Mar 16 08:12:44 app-1 passwd[4706]: pam_unix(passwd:chauthtok): password changed for user1
17 Mar 16 08:12:46 app-1 chfn[4707]: changed user `user1' information
18 Mar 16 08:12:49 app-1 chfn[4708]: changed user `user1' information
19 Mar 16 08:12:55 app-1 groupadd[4710]: new group: name=user2, GID=1002
20 Mar 16 08:12:55 app-1 useradd[4711]: new user: name=user2, UID=1002, GID=1002, home=/home/user2, shell=/bin/bash
21 Mar 16 08:13:00 app-1 passwd[4714]: pam_unix(passwd:chauthtok): password changed for user2
Line 1, Column 1
Tab Size: 4
Plain Text

Processing Raw Cybersecurity Data – Example

EXERCISE: Assessing Semi-Structured Log Data

1. What are we looking at – what is this?
   - What system? http://honeynet.org/challenges/2010_5_log_mysteries
   - What types of logs / format? https://help.ubuntu.com/community/LinuxLogFiles
   - How 'verbose'? Inherent structure (or lack thereof)?
   - How might we build context (structure more granularly)?
     - Schema availability? Data dictionary? Guidance from experts?
     - For example, https://help.ubuntu.com/community/LinuxLogFiles
     - Codes and indicators - descriptive versus opaque?

Cyber Forensics: Log File Analysis – HoneyNet Challenge

Virtual Server Log File Analysis (Virtual Ubuntu Linux Server)

1. **Was the system compromised** and when? How do you know that for sure?
2. If the system was compromised, **what was the method used**?
3. Can you locate how many attackers failed? If some succeeded, how many were they?
4. How many stopped attacking after the first success?
5. What happened after the brute force attack?
6. Locate the authentication logs. Was a brute force attack performed? if yes, how many?
7. What is the timeline of significant events? How certain are you of the timing?
8. Anything else that looks suspicious in the logs? Any misconfigurations? Other issues?
9. Was an automatic tool used to perform the attack? if yes, which one?
10. What can you say about the attacker's goals and methods?

http://honeynet.org/challenges/2010_5_log_mysteries

Processing Raw Cybersecurity Data – Example

EXERCISE: Assessing Semi-Structured Log Data

2. How can we reduce and structure?
   - Parsing and extracting – many options
     - Commercial tools (e.g. Splunk, Excel) and free tools (e.g. MS Log Parser)
     - Scripting/programmatic (e.g. Perl, R, Python, SAS DS2, UNIX GREP, PowerShell)
3. Where do we store and transform subsequently?
   - Flat files
   - Database
   - SIEM or other specialized repository
   - 'Data Lake' (note: also can dump the raw logfile there too)
   - Analytics platform

Demonstration / Exercise

EXERCISE: Assessing Semi-Structured Log Data

1. Open ‘Log Parser 2.2’
2. Examples of structured queries to try:

More fun with LogParser!
https://mlichtenberg.wordpress.com/2011/02/03/log-parser-rocks-more-than-50-examples/

ALL FAILED AUTHENTICATION EVENTS:

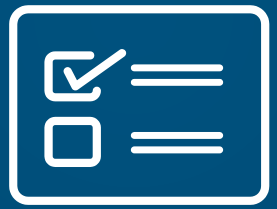
```
logparser -i:TEXTLINE -o:csv "SELECT * INTO
'D:\@CYBER\1.FRAME\B.Log_Files\results\log_extract_failed.txt' FROM
'D:\@CYBER\1.FRAME\B.Log_Files\log_files\auth.log' WHERE Text LIKE '%Failed password%'"
```

ALL FAILED ROOT ACCESS AUTHENTICATION EVENTS – STRUCTURED:

```
logparser -i:TEXTLINE -o:csv "SELECT EXTRACT_TOKEN (Text, 8, ' ') AS UserName,
EXTRACT_TOKEN (Text, 10, ' ') AS IPSource, EXTRACT_TOKEN (Text, 12, ' ') AS Port,
EXTRACT_TOKEN (Text, 10, ' ') AS Protocol INTO
'D:\@CYBER\1.FRAME\B.Log_Files\results\log_extract_failed_root.txt' FROM
'D:\@CYBER\1.FRAME\B.Log_Files\log_files\auth.log' WHERE Text LIKE '%Failed password for root%'"
```

RESULTS *D:\@CYBER\1.FRAME\B.Log_Files\results*

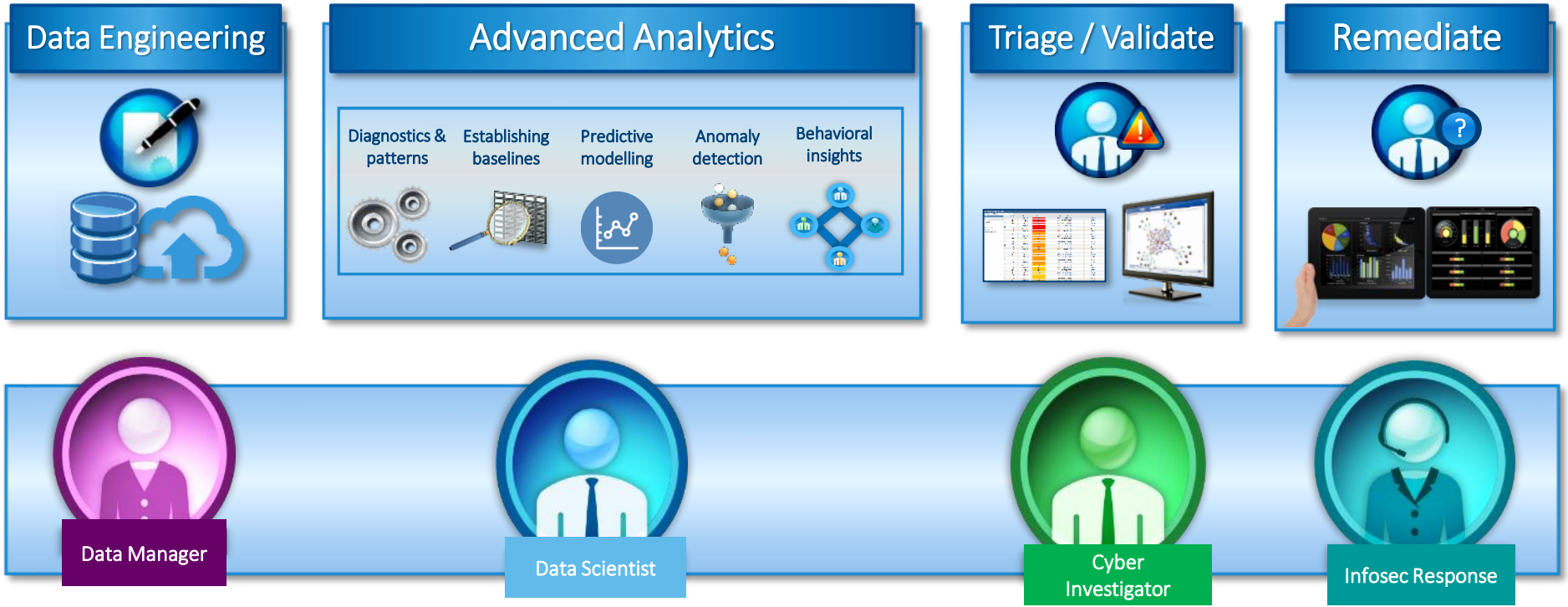
Exercise Review



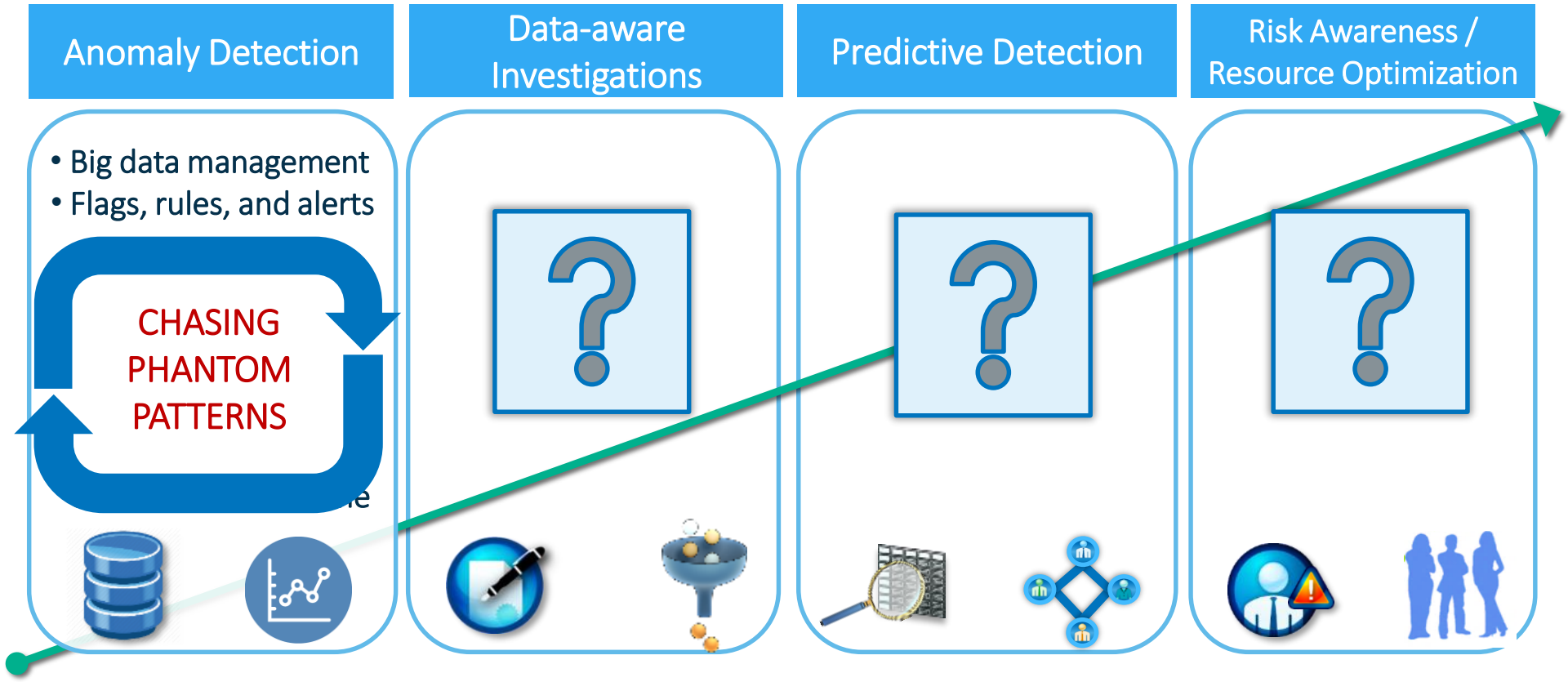
Conclusion

Section Review

Cybersecurity Data Science as a Process



Cybersecurity Analytics Maturity



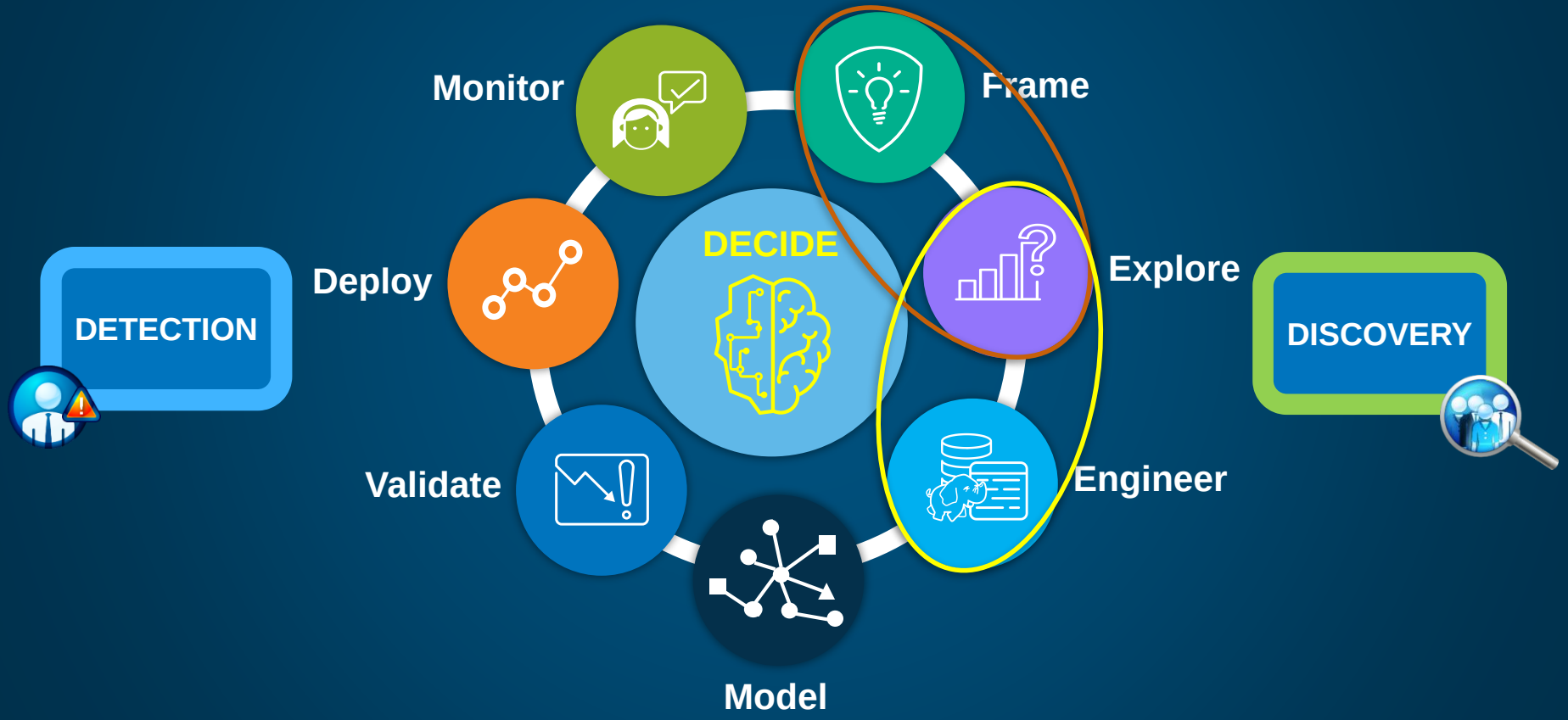
Cybersecurity Analytics Maturity

Anomaly Detection	Data-aware Investigations	Predictive Detection	Risk Awareness / Resource Optimization
• Big data management • Flags, rules, and alerts --- • Structured data • Counts of key measures • Foundation for comparisons across entities and over time	?	?	?






Cybersecurity Data Science (CSDS) Lifecycle



Cybersecurity Data Science (CSDS)

TOPIC
1. FRAME
2. DATA
3. DISCOVER
4. DETECT
5. DEPLOY