Cybersecurity Data Science Overview

Hands-on analytics practitioner methods and best practices

Scott Mongeau
Cybersecurity Data Scientist

What We Will Address Today

WHY…

- is cybersecurity data science ‘a thing’?
- sharing and socializing best practices from adjacent domains

WHAT…

- do we mean by ‘data science’ in the cyber domain?

HOW…

- is it delivered?
- and given data challenges ***and ever-emerging new incidents***, how do we…
  - make the best of limitations,
  - extract hidden patterns, and
  - make focused detection and remediation a realistic goal.

DATA SCIENCE FOR CYBER

‘How Do I…’ Points to Be Conveyed

TECHNOLOGY

- Implement data processing methods to address cybersecurity data challenges
- Leverage and extend existing cybersecurity alerts and rules
- Address challenges associated with big and fast data

PROCESS

- Implement models for pattern detection using unsupervised machine learning
- Boot-strap extracted patterns to detect targeted anomalies
- Detect focused cybersecurity anomalies with predictive machine learning
- Optimize cybersecurity resource allocation

ORGANIZATION

- Plan and implement an operational cybersecurity data analytics process
- Design and implement big data analytics approaches (high level)
- Integrate analytics into the enterprise cybersecurity process

Context Setting

Introduction to Cybersecurity Data Science

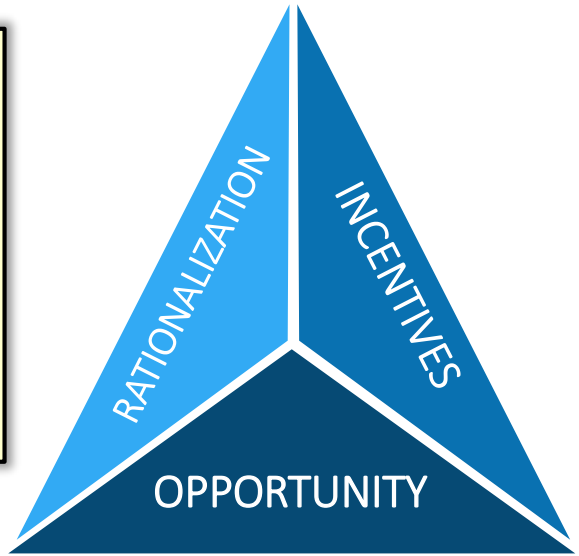
Scott Mongeau
Cybersecurity Data Scientist

Cybersecurity Context

The Fraud Triangle: Not 'IF', but 'When?'

FRACTURED LOGIC

- "I'm only playing around – it is a game"
- "I am not respected"
- "I need / deserve the money"
- "They should not have such poor security – I will teach them a lesson"
- "They had it coming"
- "They are working closely with a country I do not agree with"



SYSTEMIC WEAKNESSES
STRIDE susceptibility

- Spoofing
- Tampering
- Repudiation
- Information disclosure
- Denial of service
- Elevation of privilege

STRATEGIC
- National interests
- Corporate espionage / sabotage

FINANCIAL
- Fraud
- Theft (i.e. credit cards)
- Market manipulation

REPUTATIONAL
- Recognition of expertise / fame in hacker networks
- Making a political statement

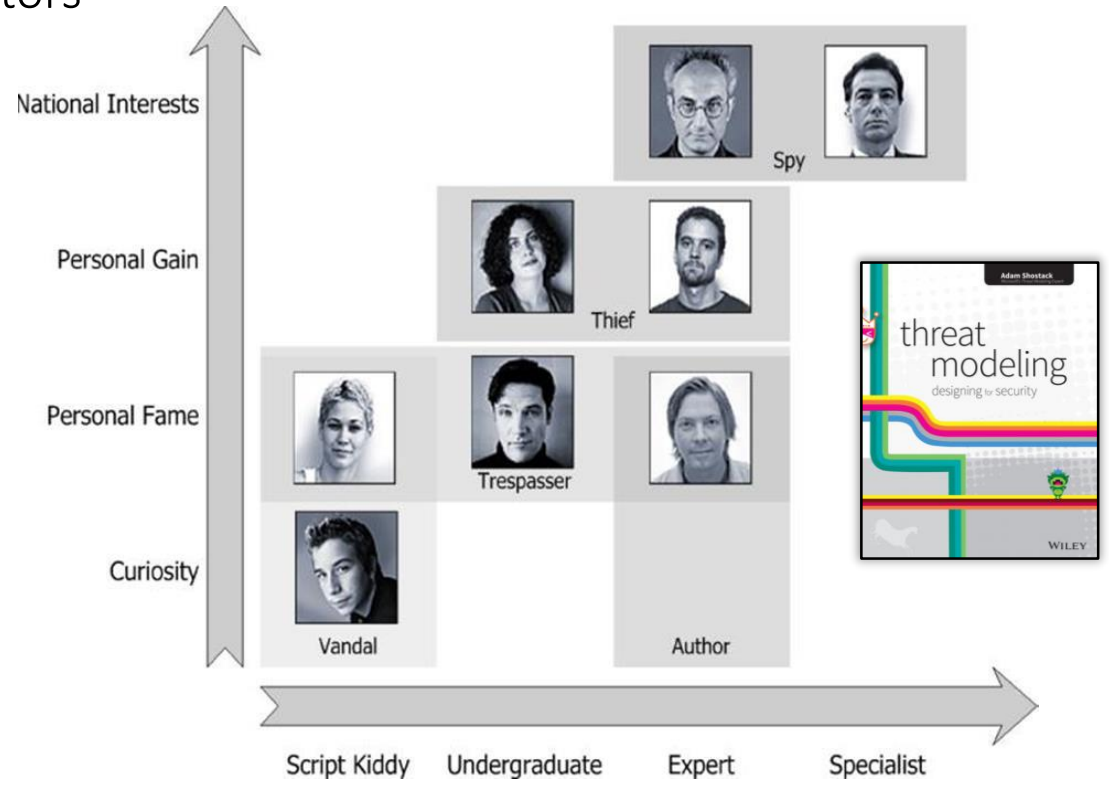
PERSONAL
- Curiousity
- Greed or revenge
- Character flaws (i.e. sociopathic disorder)

Threat Actors

Aims & impacts of threat actors permanent or temporary:

- Destruction
- Damage/Compromise
- Denial of Access
- Theft

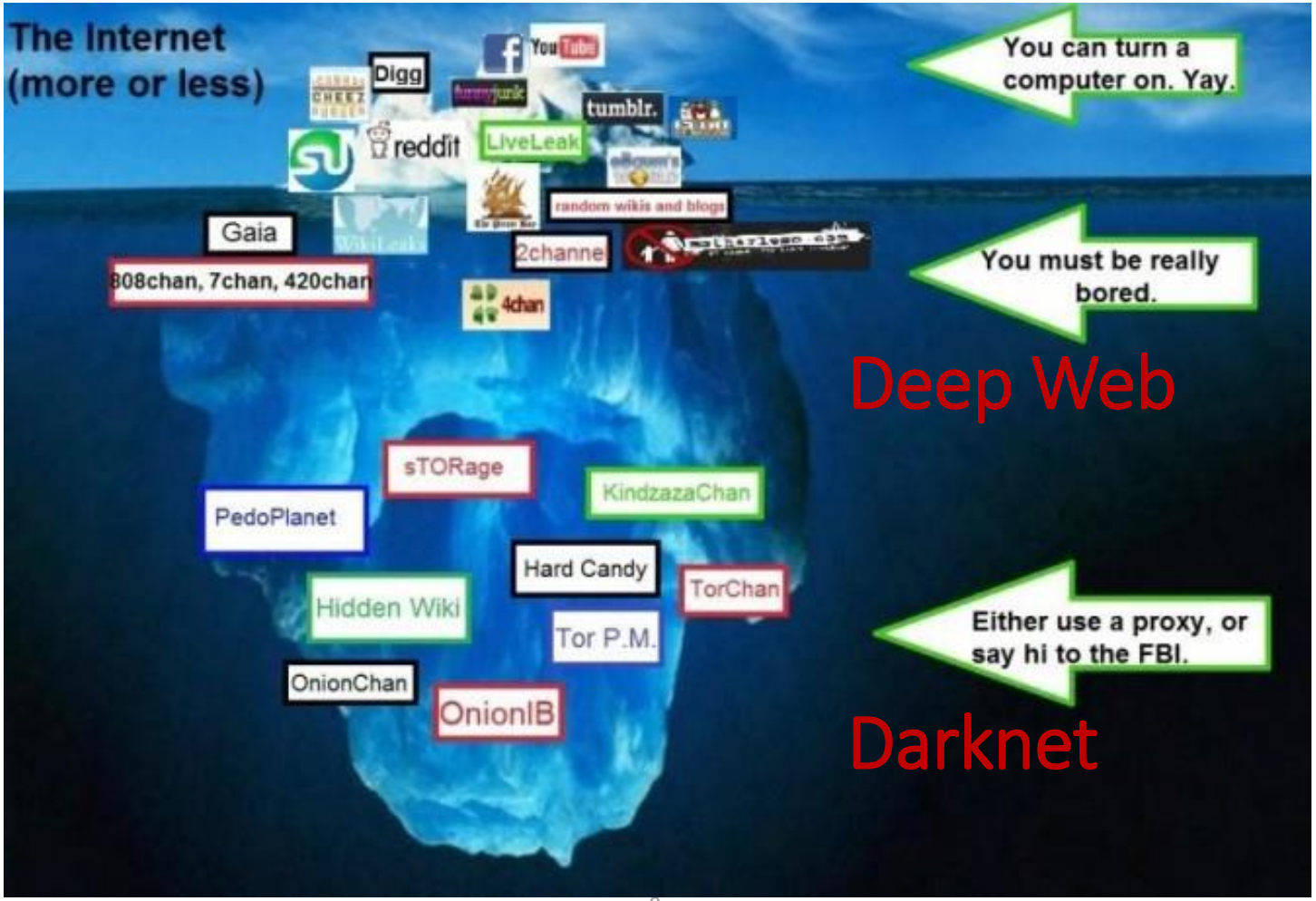
….of data or underpinning physical and virtual infrastructure



Shostack, Adam. 2014. Threat Modeling: Designing for Security. Wiley.

Aucsmith, David. 2003. Threat Personas. Microsoft.

8


The Internet
(more or less)
Digg
YouTube
reddit
LiveLeak
tumblr.
random wikis and blogs
Gaia
808chan, 7chan, 420chan
2channe
4chan
You can turn a computer on. Yay.
You must be really bored.
Deep Web
sTORage
KindzazaChan
PedoPlanet
Hard Candy
TorChan
Hidden Wiki
Tor P.M.
OnionChan
OnionIB
Either use a proxy, or say hi to the FBI.
Darknet

9

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 & 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

SERVICES

HACKING	$100	EMAIL ACCOUNT
	$100	SOCIAL MEDIA ACCOUNT
	$300	CMS WEBSITE (WORDPRESS, ETC.)
USER OBFUSCATION	$150	BULLETPROOF HOSTING IN A 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	500 MALICIOUS EMAIL 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

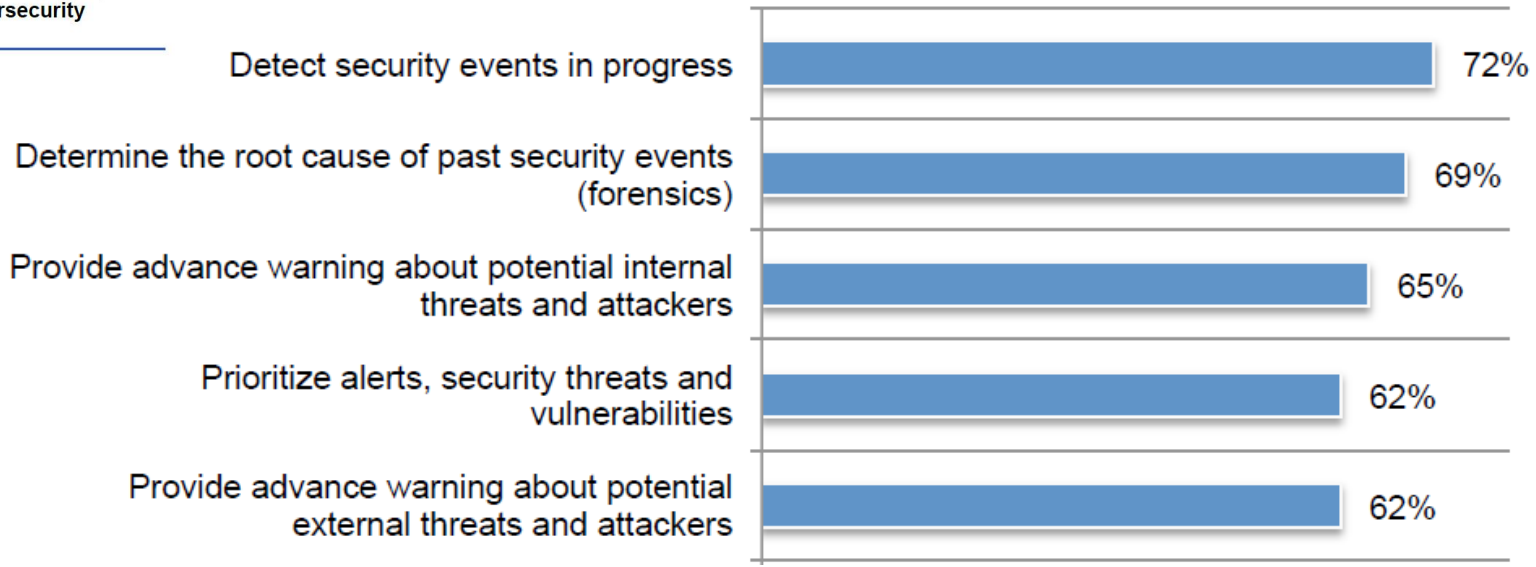
DATA

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

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

When Seconds Count: How Security Analytics Improves Cybersecurity Defenses

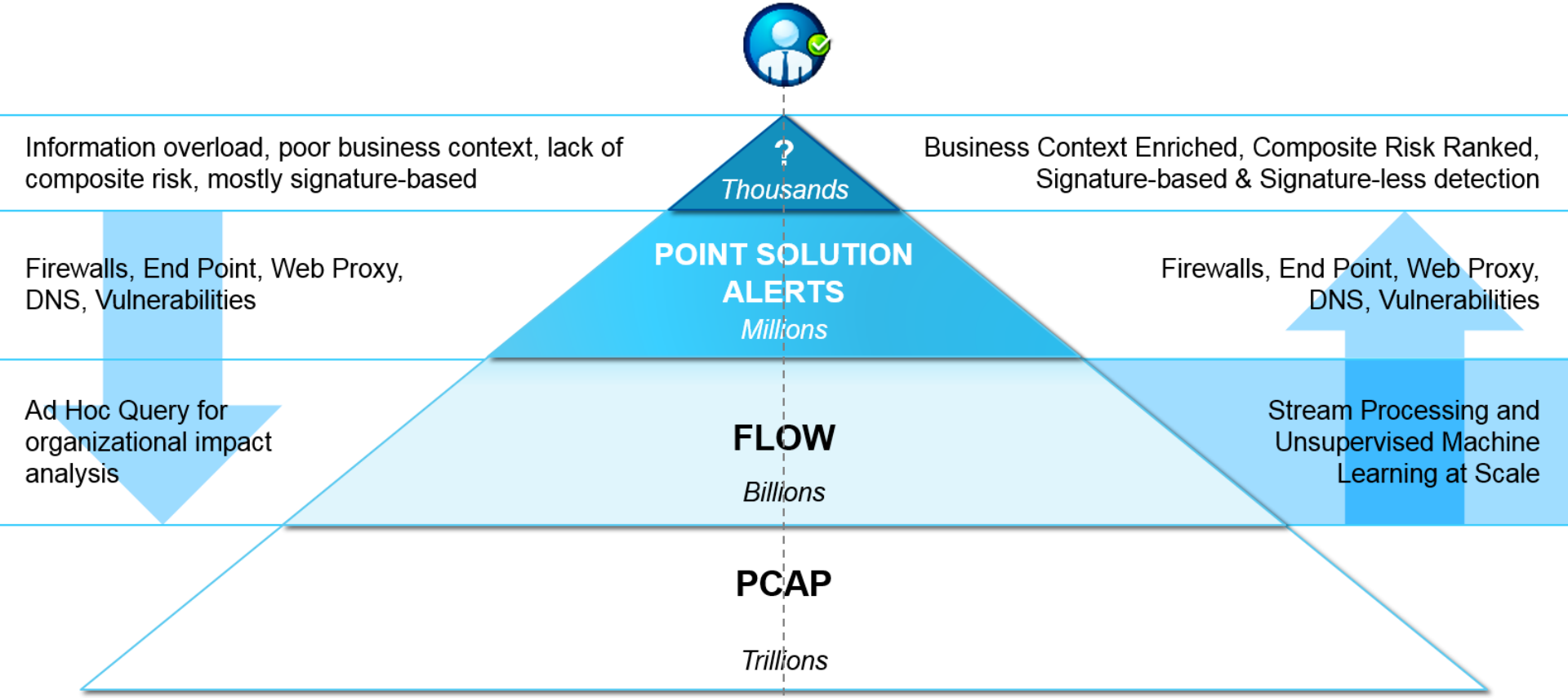
Most important objectives for a cybersecurity analytics solution*



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

** Survey of 621 global IT security practitioners*

Data Volumes and Security Challenge



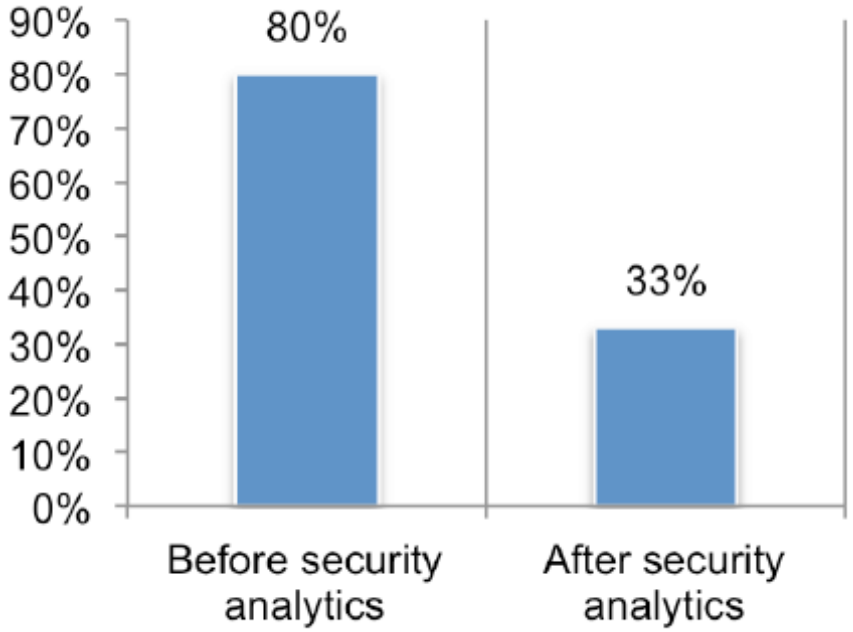
Cybersecurity Context

CHALLENGES → DATA SCIENCE

Cyber detection is challenged by…	Data science addresses…
…high volumes of structured and **unstructured** data	…operation at **big data scale** at **high velocity**
…**disconnected** data sources of **variable quality**	…**assess, extract, transform,** and **aggregate** data
…high **false positive** alerts with rule-based approaches	…unsupervised machine learning identifies **hidden patterns**
…lack of **statistical baselines** to establish **validity**	…effective statistical **diagnostics** for model **validation**
…slow and **manual** investigation processes (needles in the haystack)	…supply hunters with **targeted alerts** based on demonstrable statistical anomalies

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

Level of difficulty in reducing false alerts*



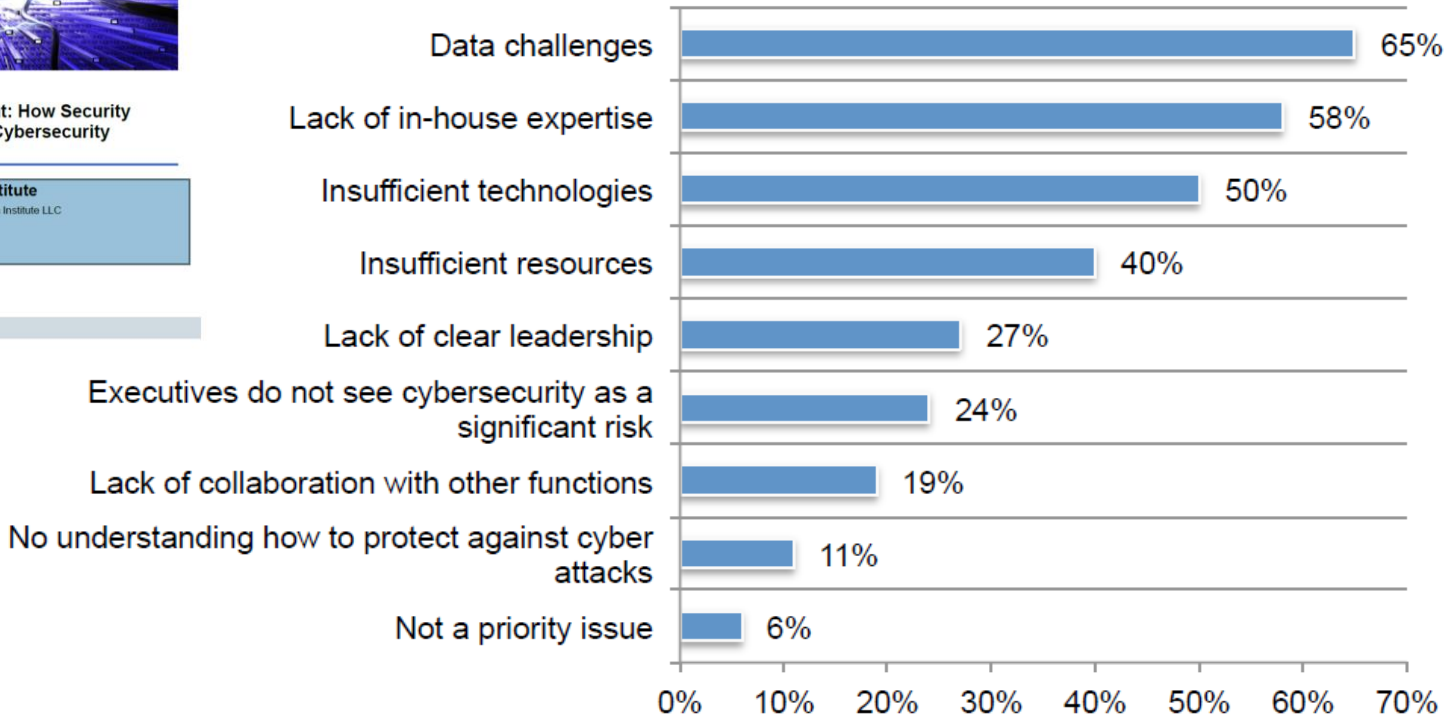
** Survey of 621 global IT security practitioners*

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

Challenges Preventing Successful Use of Cybersecurity Analytics*

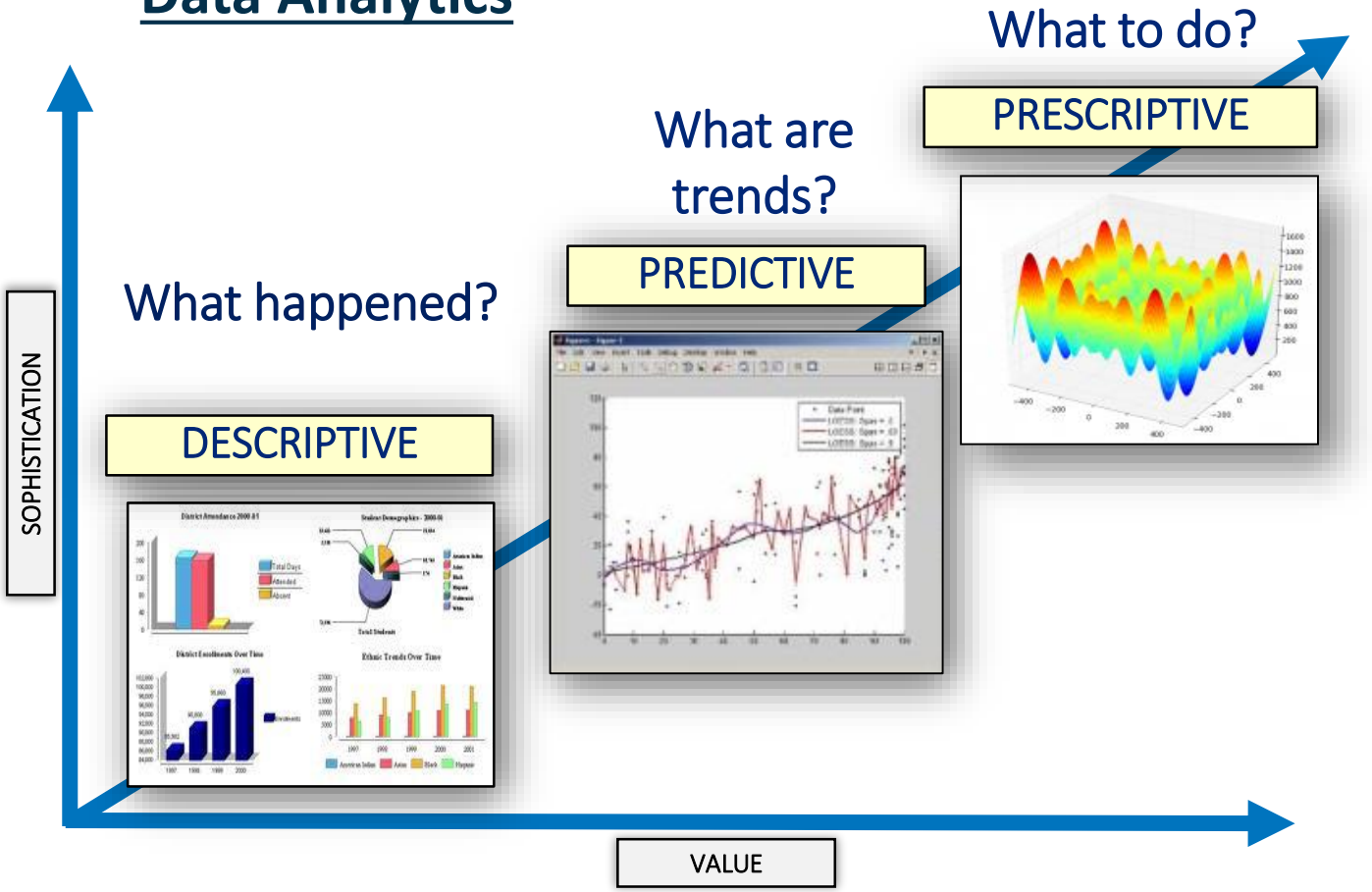


https://www.sas.com/en_us/whitepapers/ponemon-how-security-analytics-improves-cybersecurity-defenses-108679.html

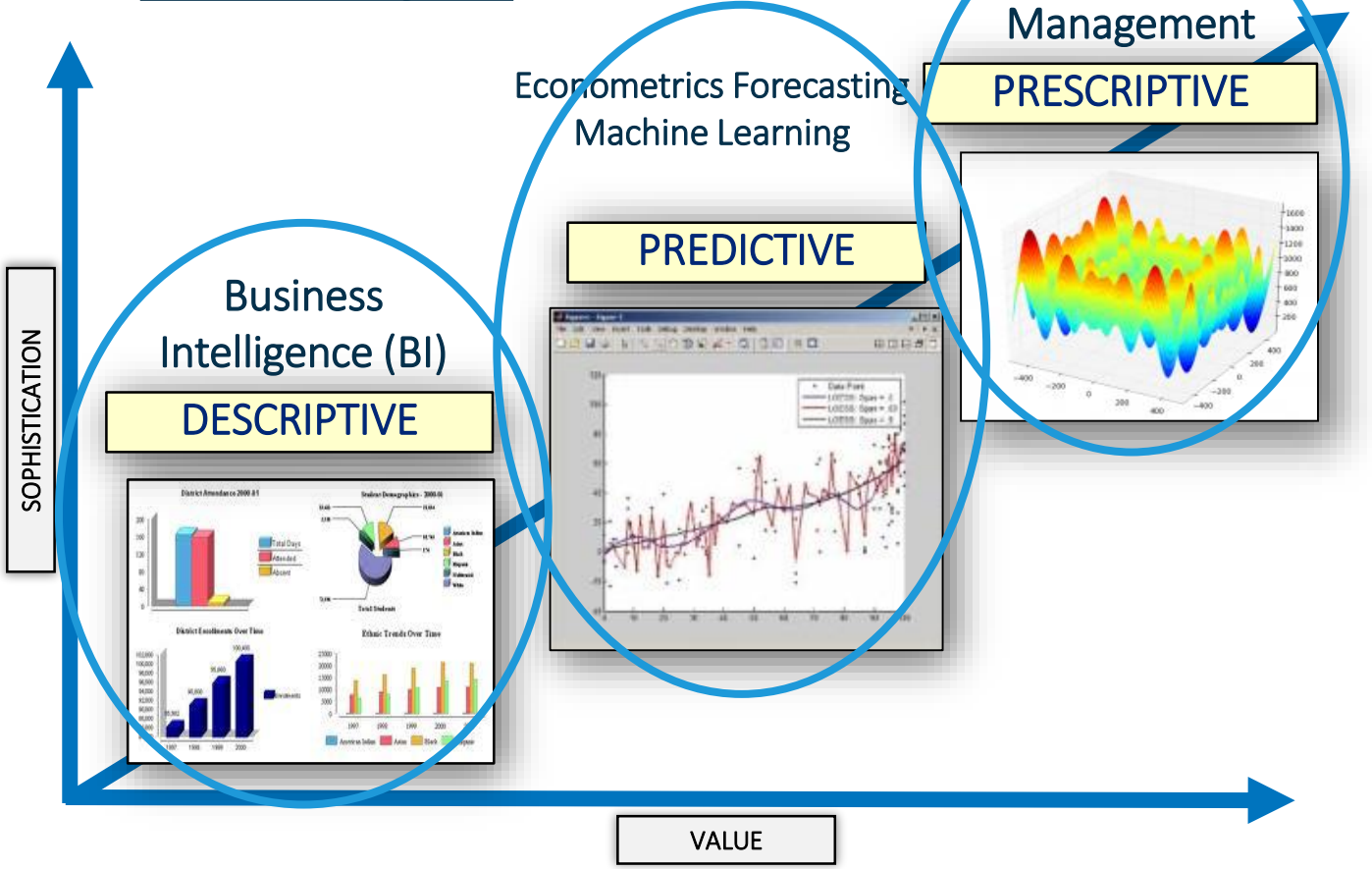
** Survey of 621 global IT security practitioners*

Data Science

Data Analytics



Data Analytics

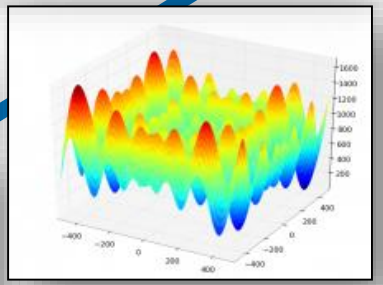


Cybersecurity Analytics

Based on identified incidents, how can we optimize monitoring?

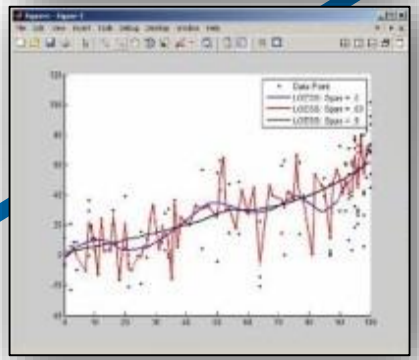
What is the expected behavior of particular users or devices?

PRESCRIPTIVE

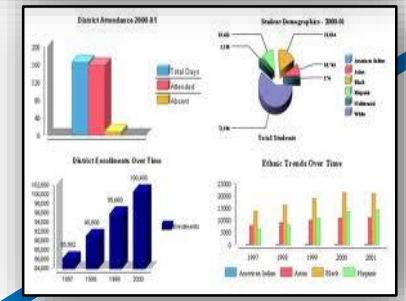


What are trends in network and device usage?

PREDICTIVE



DESCRIPTIVE



SOPHISTICATION

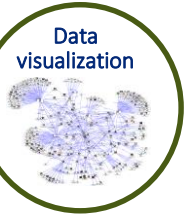
VALUE

Advanced Analytics

How can we support cybersecurity experts with visualizations?

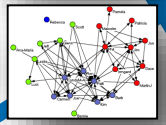
Understanding Patterns

Data visualization



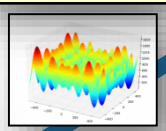
Optimizing Systems

SEMANTIC



What are the underlying human factors associated with the insights gained?

PRESCRIPTIVE



Understanding Social Context & Meaning

PREDICTIVE

Identifying Factors & Causes

DIAGNOSTICS

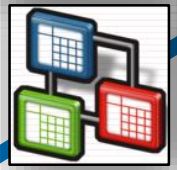
Forecasting & Probabilities

SOPHISTICATION

DESCRIPTIVE

Can we statistically substantiate our assumptions and hypothesizes?

DATA QUALITY

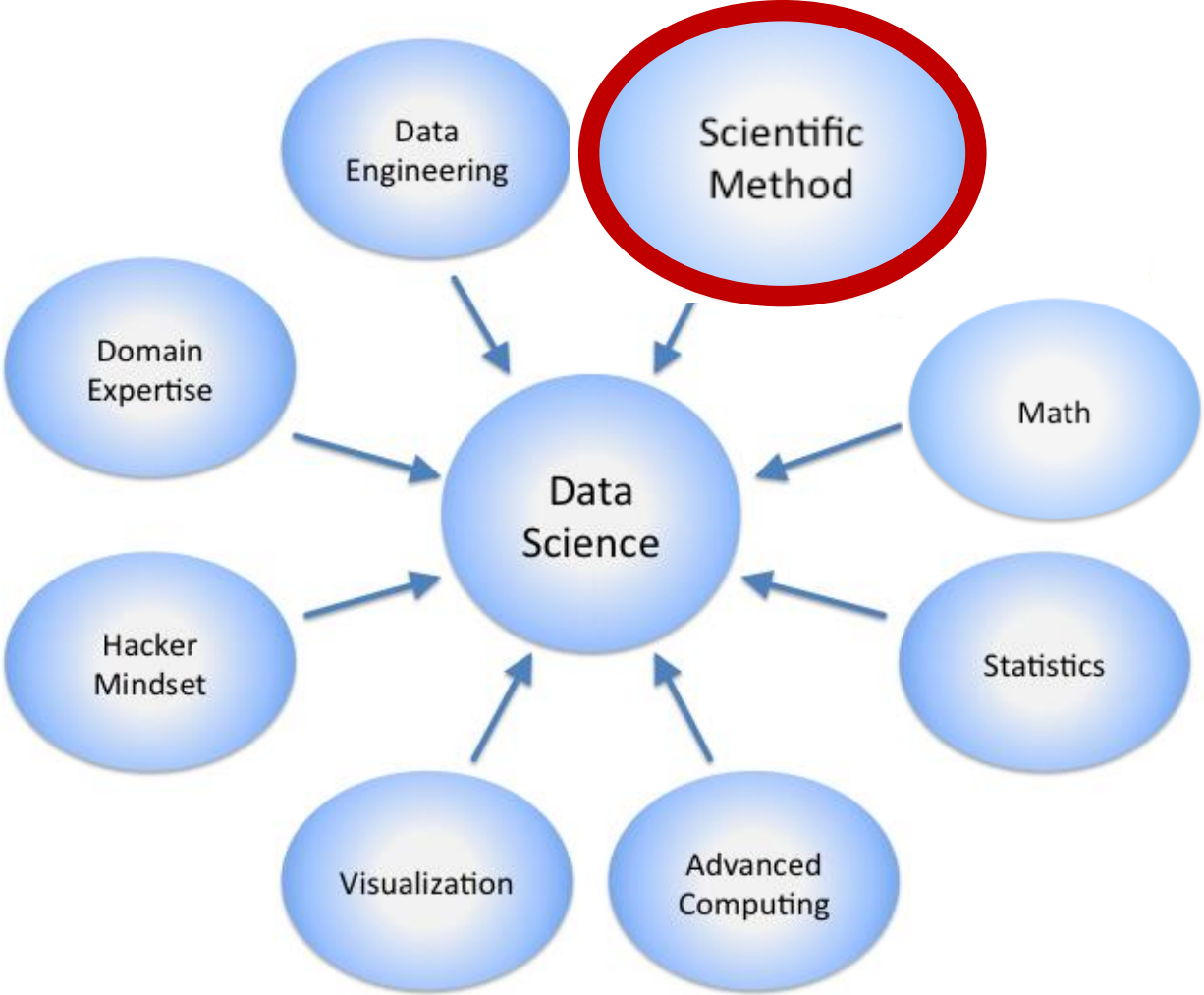


Business Intelligence

Is data quality reliable enough? What are biases/ limitations?

VALUE

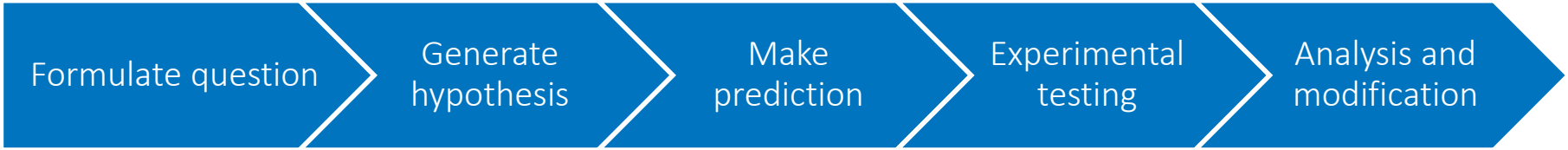
https://www.youtube.com/watch?v=LlSdZLsiAzo



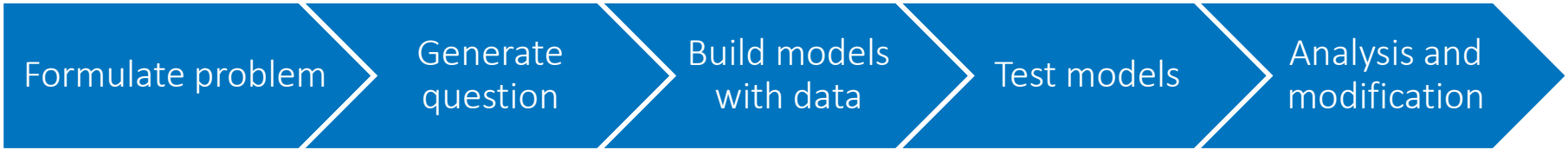
Calvin.Andrus (2012) Depicts a mash-up of disciplines from which Data Science is derived http://en.wikipedia.org/wiki/File:DataScienceDisciplines.png

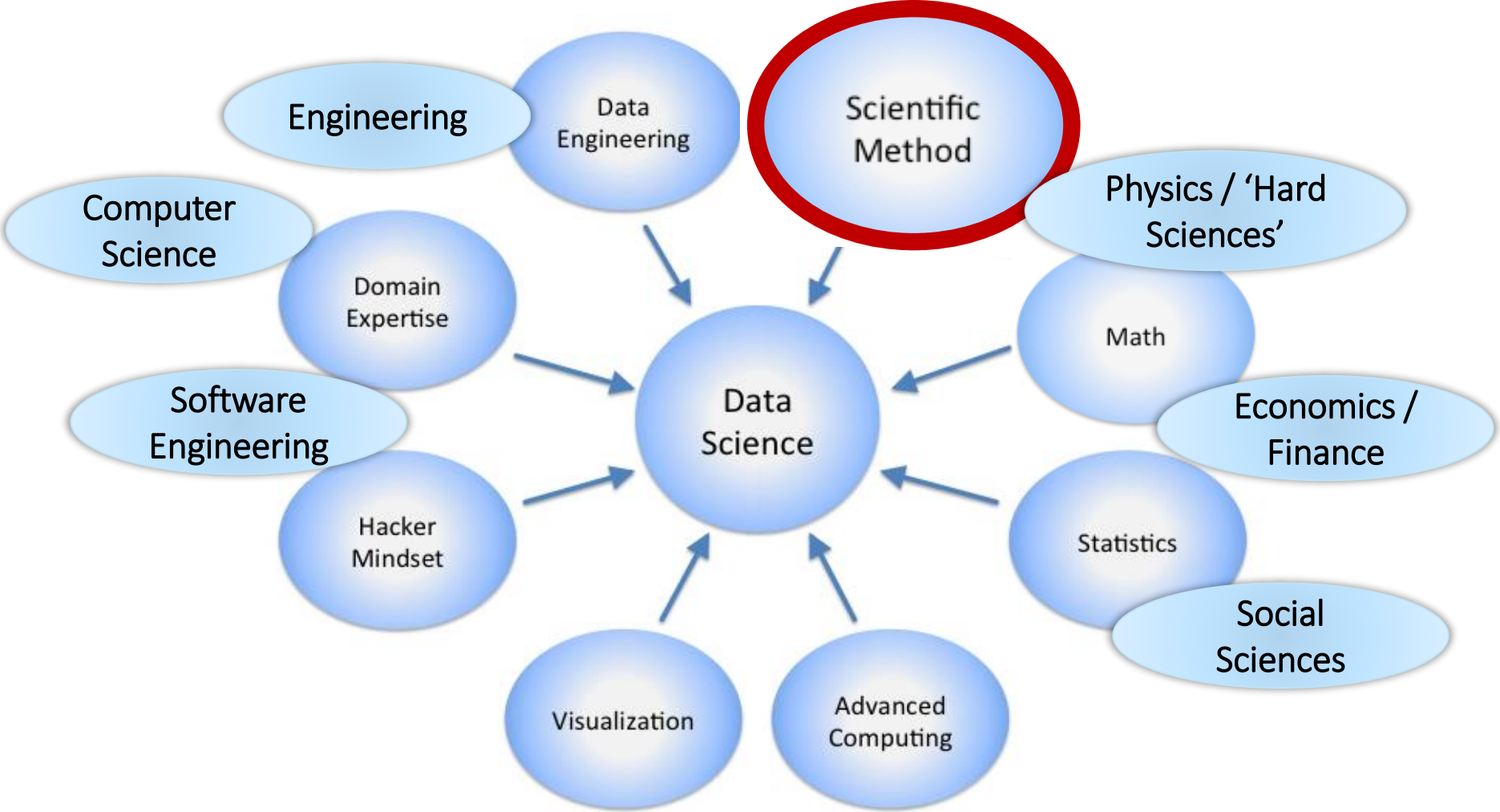
Scientific Method

Scientific Experimentation



Data Science Process



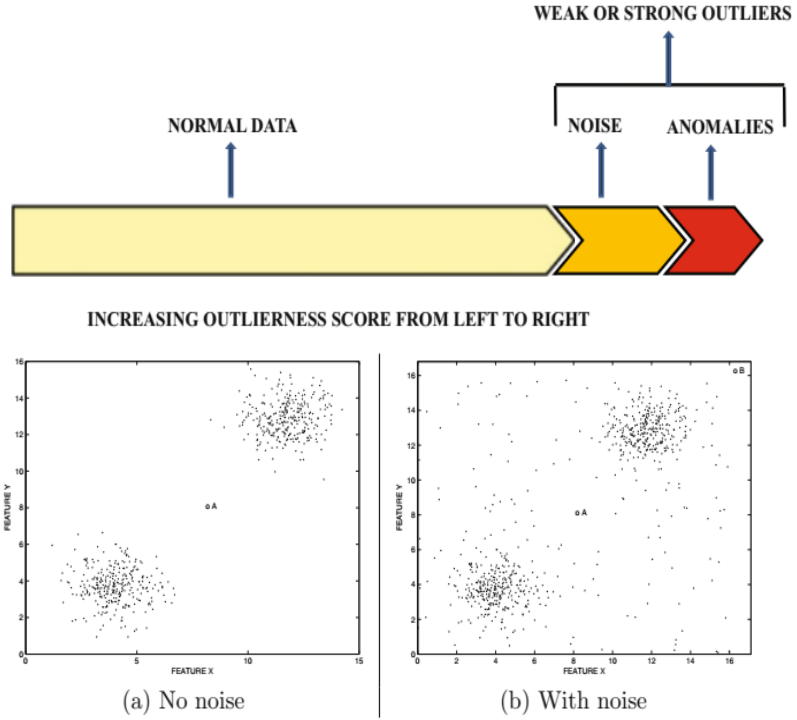


Calvin.Andrus (2012) Depicts a mash-up of disciplines from which Data Science is derived http://en.wikipedia.org/wiki/File:DataScienceDisciplines.png

Cybersecurity Data Science as a Process

Simply Complex

Identifying targeted anomalies amongst an ocean of noise…



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

Simplified Ideal ‘To-Be’ End State

Data Engineering

Advanced Analytics

Business rules/scores

Database searches

Predictive modeling

Anomaly detection

Network Analysis

Targeted Alerts

Streamlined Resolution

CYBER RISK ANALYTICS PROCESS

Data Manager

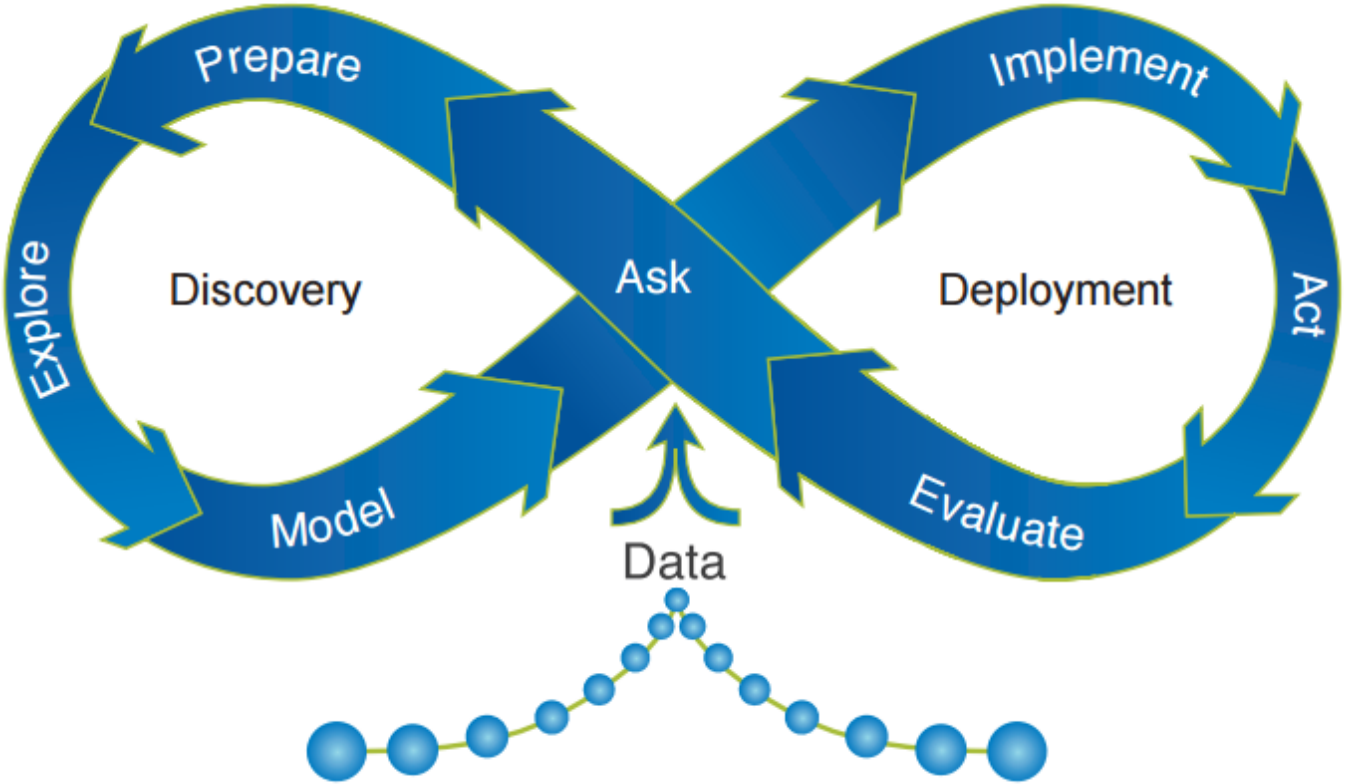
Data Scientist

Cyber Investigator

Case Investigator

RECURSIVE FEEDBACK

Analytics Life Cycle



SOURCE 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

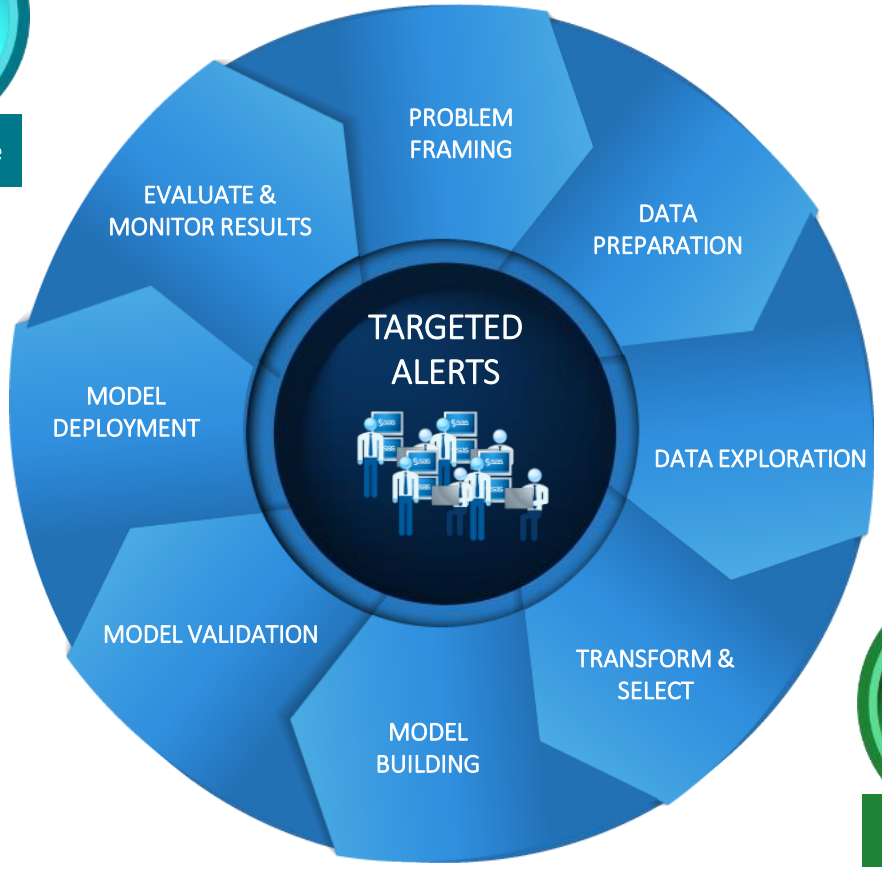
Cyber Analytics Life Cycle

InfoSec Response

Data Manager

Pattern Detection

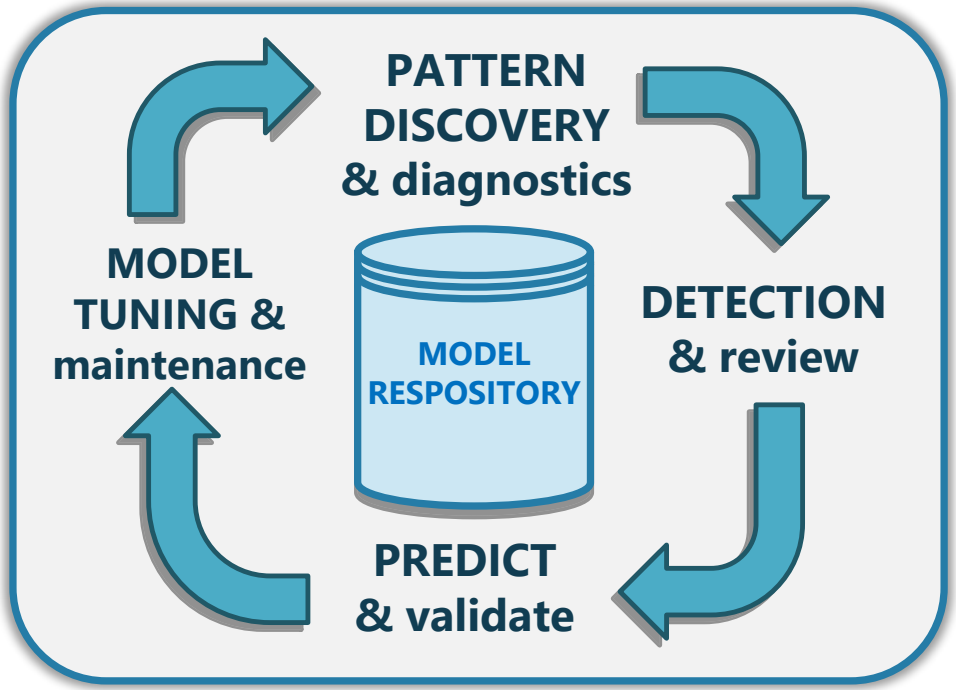
Exploration and Insights



Data Scientist

Cyber Investigator

Data Science for Cybersecurity: High-Level Process



1. Data extraction, aggregation, *quality*, and variable selection
2. Discovery of hidden patterns and groupings (segmentation)
3. Targeted anomaly detection
4. Operationalization of investigative cycle
5. Ongoing risk / operational model refinement and management

Cybersecurity Analytics Maturity Curve (Simplified)

Anomaly Detection

- Big data management
- Flags, rules, and alerts

Chasing phantom patterns

Data-aware Investigations

- Tagging risk indicators
- Tagging incidents and false alerts
- Focused and new measures extracted

Predictive Detection

- Tagged instances drive predictive detection
- Refined alerts fed to investigators
- Follow-on results drive further refinements

Risk Awareness / Resource Optimization

- Focused risk models synthesized, tested, & refined
- Matching of resources (availability/efficacy) to alerts (risk likelihood/impact)