Carnegie Mellon University
Software Engineering Institute

Cybersecurity Data Science (CSDS)

Best Practices in an Emerging Profession

Scott Allen Mongeau

Cybersecurity Data Scientist – SAS Institute
PhD candidate - Nyenrode Business University (Netherlands)

s.mongeau@edp1.nyenrode.nl
scott.mongeau@sas.com

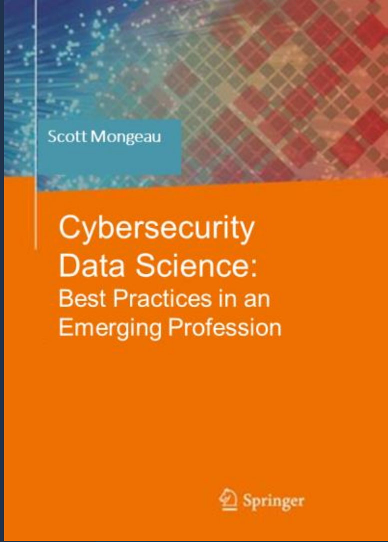
@SARK7 #CSDS2020 #FloCon2020

FloCon 2020

JANUARY 6-9, 2020 | SAVANNAH, GA

PhD academic research / book

- ~June 2020 release



Research on cybersecurity data science (CSDS) as an emerging profession

I. Literature: What is CSDS and is it a profession?

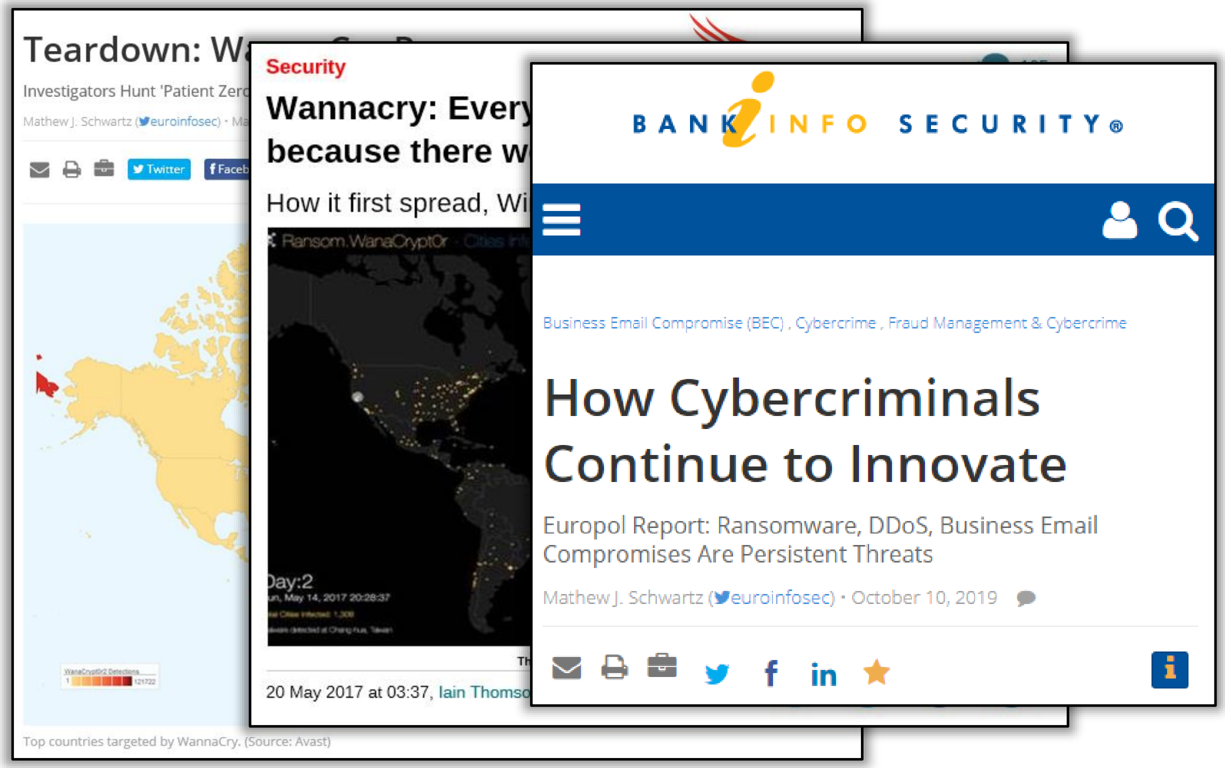
II. Interviews: 50 CSDS practitioners

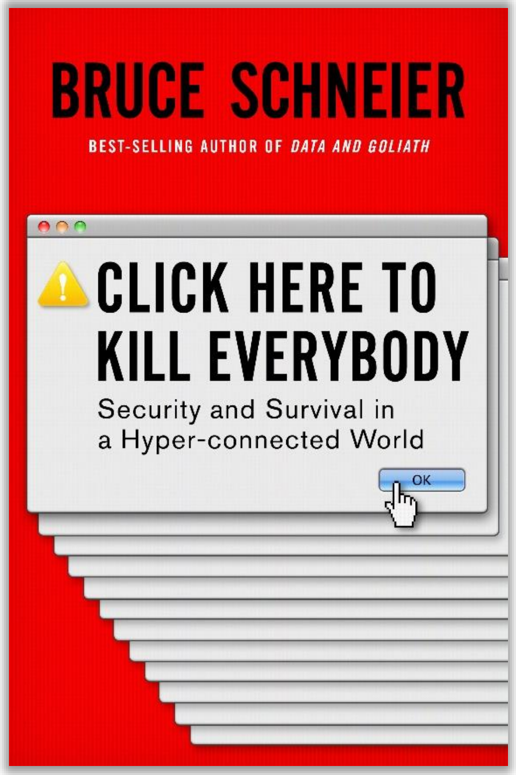
III. Designs: Approaches to address challenges

I. CSDS Literature

FUD Fear, Uncertainty, Doubt

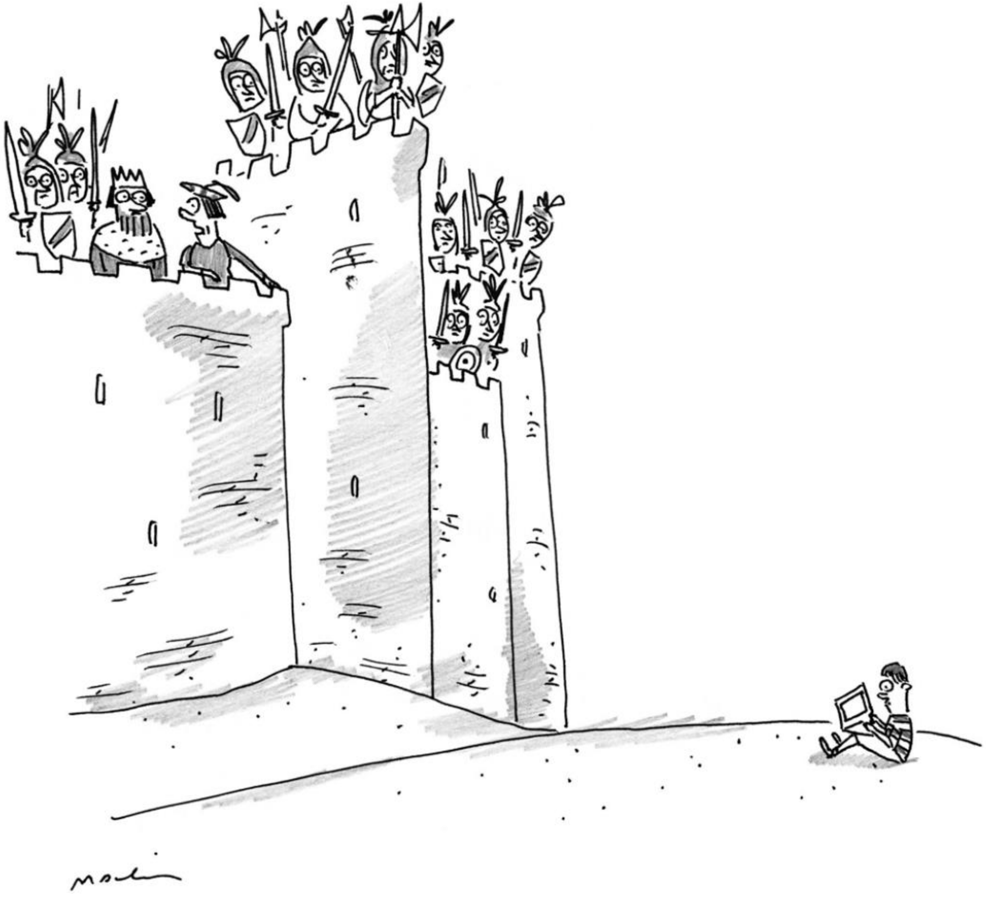
Expansion of exposure and targets >!< Increasing sophistication, frequency, and speed of attacks



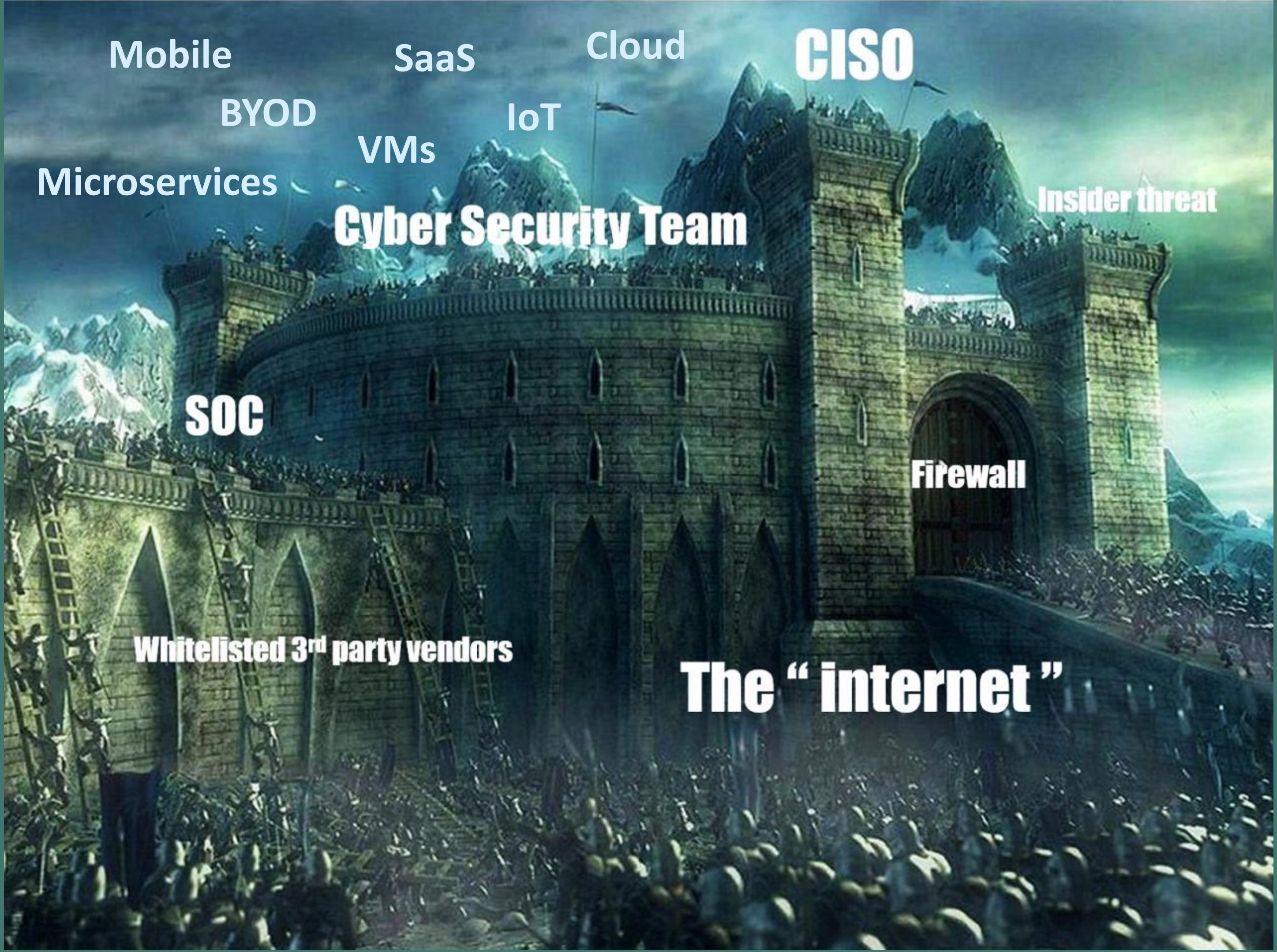


Castle and Moat

How quaint!



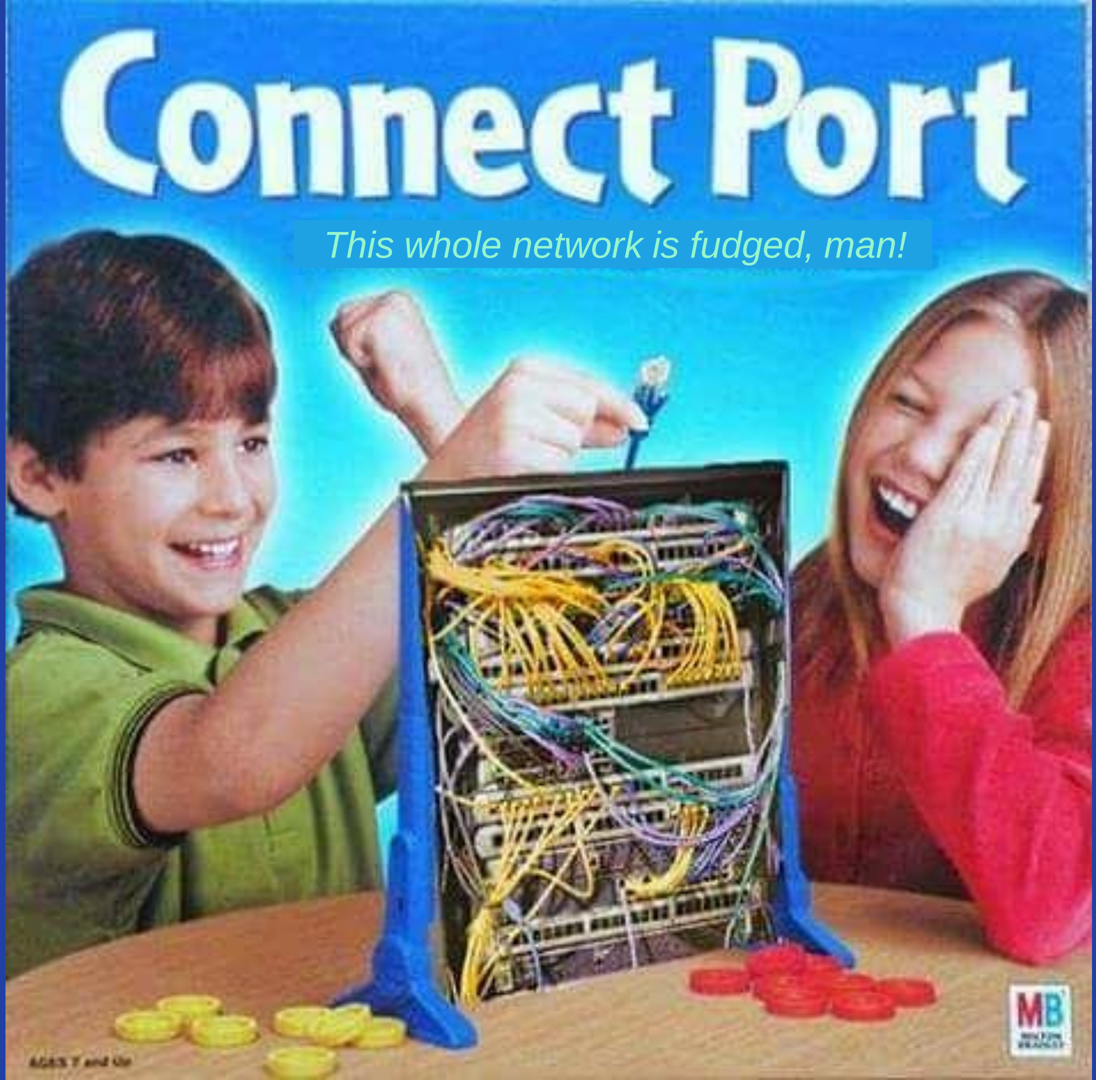
"Bad news, Your Majesty—it's a cyberattack."


Mobile
SaaS
Cloud
CISO
BYOD
IoT
VMs
Microservices
Insider threat
Cyber Security Team
SOC
Firewall
Whitelisted 3rd party vendors
The " internet "

Cybersecurity Challenges

Data Science

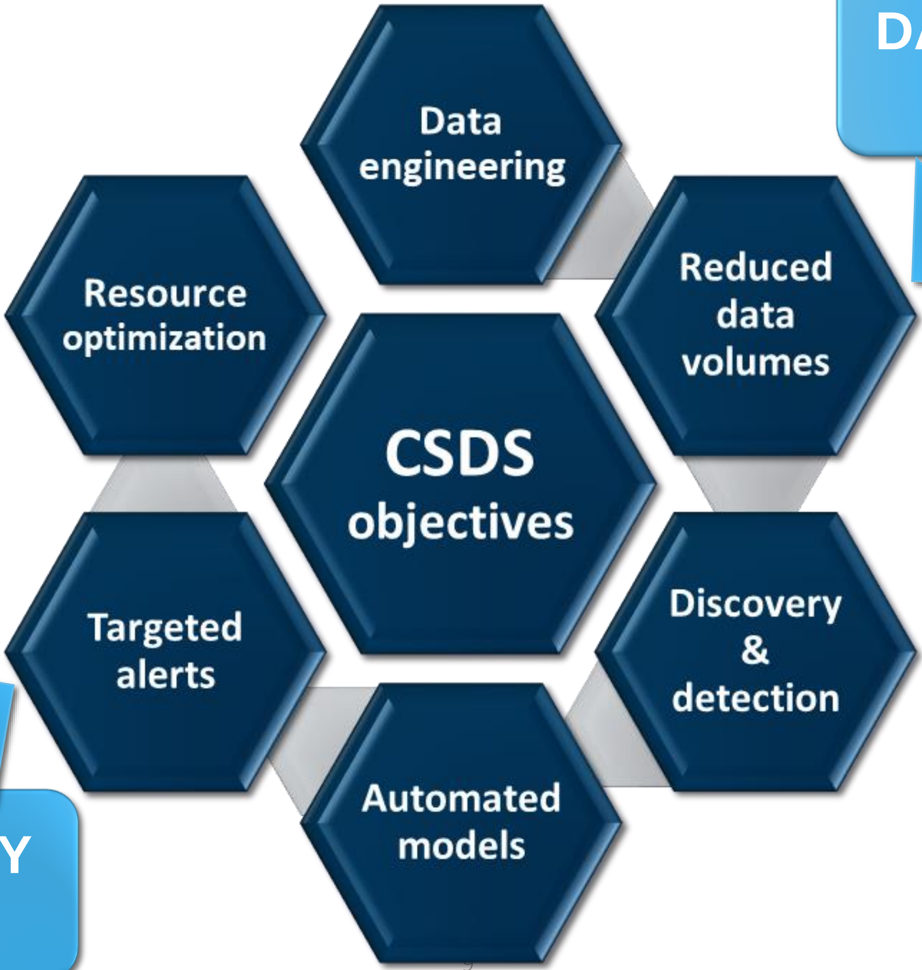
New hope amidst complexity and confusion…



CSDS

Cyber Security Data Science

DATA SCIENCE METHODS



CYBERSECURITY GOALS

CSDS: Existing Professionals + Demonstrated Efficacy

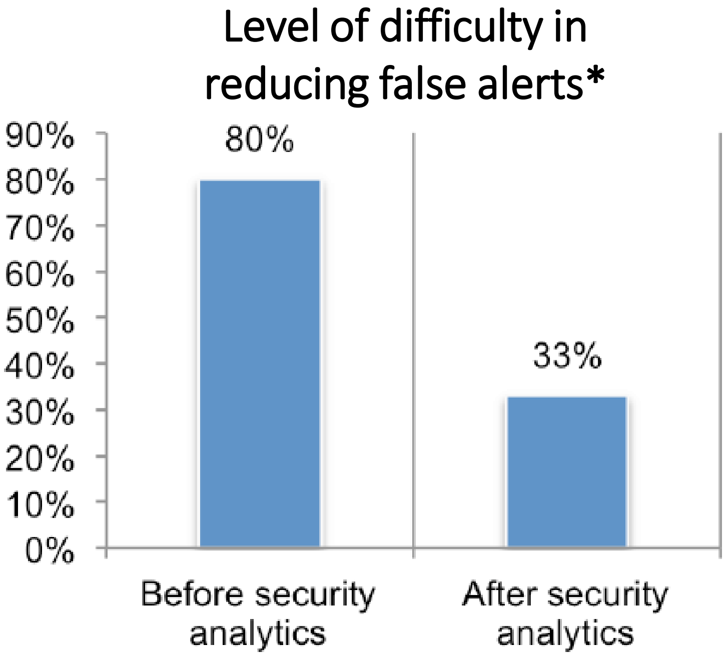

Ponemon INSTITUTE

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*

EXAMPLE CSDS PRACTICAL APPLICATIONS

- Spam filtering
- Phishing email detection
- Malware & virus detection
- Network monitoring
- Endpoint protection

‘Professional Maturity’ Comparison

#	CRITERIA	CYBER	DS	CSDS
1	Broad interest	●	●	●
2	People employed	●	◕	◕
3	Informal training	●	●	◑
4	Informal groups	●	●	◑
5	Professional literature	●	●	◕
6	Research literature	◕	◕	[illegible]
7	Formal training	●	[illegible]	◔
8	Formal prof. groups	●	◑	○
9	Professional certificates	◕	◔	○
10	Standards bodies	●	◔	○
11	Academic discipline	◕	◔	○

CYBER =
Growing challenges +
rapid paradigm shift

DATA SCIENCE =
Poorly defined standards
“whatever you want it to be!”

CSDS =
At risk problem child?

The Blessing and Curse of Data Science

PROS

- Commercial interest
- Range of methods
- Freedom to experiment
- Delivers efficiencies
- Big data engineering
- Insightful questions
- Power of machine learning

CONS

- Hype & noise
- Befuddling array of approaches
- Lack of standards
- Myth of automation
- Big data ipso facto is not solution
- Wait, what is the question?
- "Throwing the statistical baby out with grampa's bathwater?"

II. CSDS Interviews

CSDS Practitioner Interviews

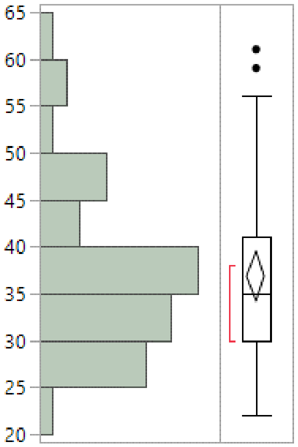
30 minutes per interviewee

- ENTRY: How did you become involved in domain?
- What are perceived central CHALLENGES?
- What are key BEST PRACTICES?

Demographic Profile (n=50)

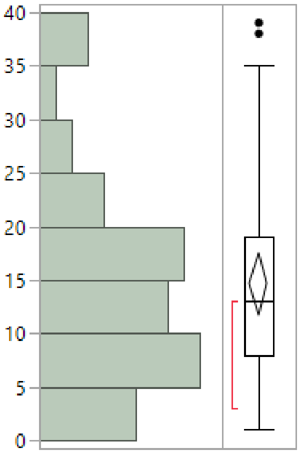
LinkedIn => 350 candidates => 50 participants

<u>Age</u>*



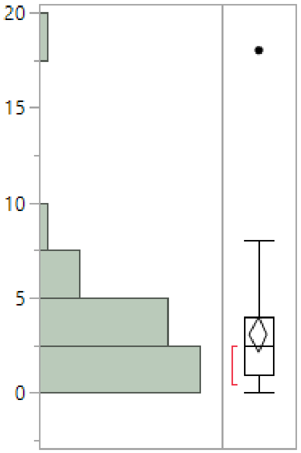
Mean	36.8
StdDev	9.1

<u># Yrs Employed</u>*



Mean	14.2
StdDev	9.5

<u># Yrs CSDS</u>*

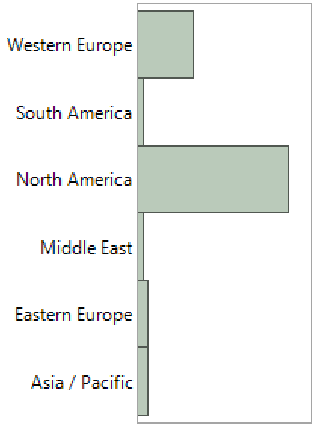


Mean	2.9
StdDev	1.9

** Estimates inferred from LinkedIn profile data*

Demographic Profile (n=50)

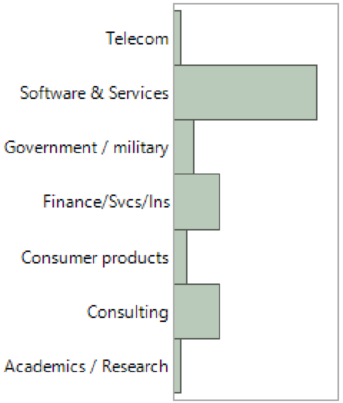
Current Region



Current Region[1]	n	%
North America	35	70%
Western Europe	10	20%
Eastern Europe	2	4%
Middle East	2	4%
South America	1	2%

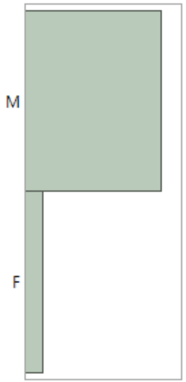
22% (n=11) relocated from native region
18% (n=9) relocated to US specifically
10% (n=5) relocated specifically from Asia/Pacific to US

Current Industry



Industry	n	%
Software and services	28	56%
Consulting	7	14%
Finance/financial services/insurance	7	14%
Government / military	3	6%
Consumer products	2	4%
Academics / research	2	4%
Telecom	1	2%

Gender



Gender	n	%
Male	43	86%
Female	7	14%

CSDS ‘CHALLENGES’: 11

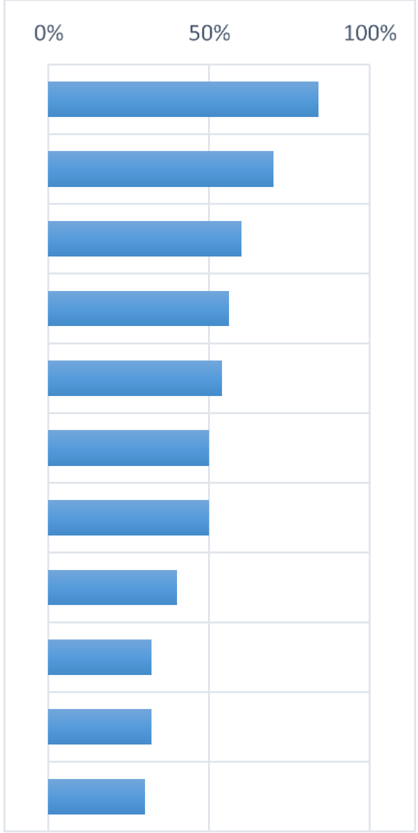
DATA PREPARATION! 84%

Marketing hype 70%

Establishing context 60%

Labeled incidents (evidence) 56%

CODED RESPONSES: Perceived Challenges	N	%
CH1: Data preparation (access, volume, integration, quality, transformation, selection)	42	84%
CH2: Unrealistic expectations proliferated by marketing hype	35	70%
CH3: Contextual nature of normal versus anomalous behavioral phenomenon	30	60%
CH4: Lack of labeled incidents to focus detection	28	56%
CH5: Own infrastructure, shadow IT, and proliferation of exposure	27	54%
CH 6: Uncertainty leads to ineffective reactive stance	25	50%
CH 7: Traditional rules-based methods result in too many alerts	25	50%
CH 8: Program ownership, decision making, and processes	20	40%
CH 9: Resourcing, developing, & hosting in house	16	32%
CH 10: Expanding breadth and complexity of cyber domain	16	32%
CH 11: Policy, privacy, regulatory, and fines	15	30%



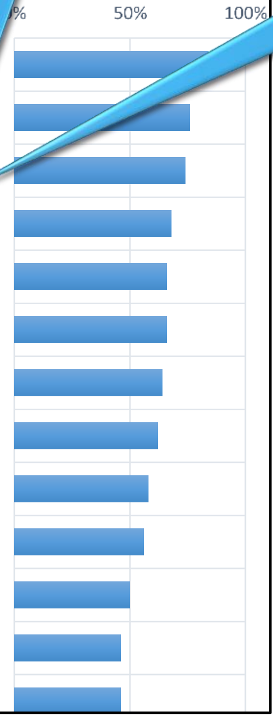
CSDS 'BEST PRACTICES': 26

DATA PREPARATION! 84%

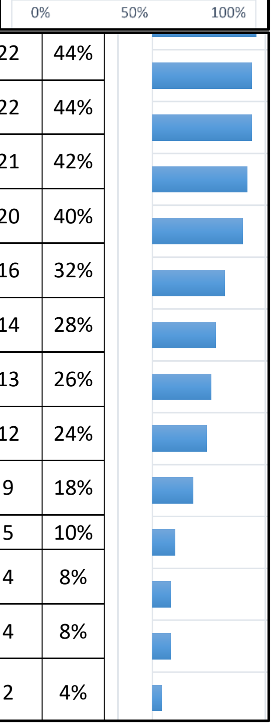
Cross-domain collaboration 76%

Scientific rigor 68%

RESPONSES: Advocated best practices	Family	N	%
BP1: Structured data preparation, discovery, engineering process	Proc	42	84%
BP2: Building process focused cross-functional team	Org	38	76%
BP3: Cross-training team in data science, cyber, engineering	Org	37	74%
BP4: Scientific method as a process	Proc	34	68%
BP5: Instill core cyber domain knowledge	Org	33	66%
BP6: Vulnerability, anomaly & decision automation to operational capacity	Tech	33	66%
BP7: Data normalization, frameworks & ontologies	Tech	32	64%
BP8: Model validation and transparency	Proc	31	62%
BP9: Data-driven paradigm shift away from rules & signatures	Org	29	58%
BP10: Track and label incidents and exploits	Proc	28	56%
BP11: Cyclical unsupervised and supervised machine learning	Proc	25	50%
BP12: Address AI hype and unrealistic expectations directly	Org	23	46%
BP13: Understand own infrastructure & environment	Org	23	46%



RESPONSES: Advocated best practices	Family	N	%
BP14: Cloud and container-based tools and data storage	Tech	22	44%
BP15: Distinct exploration and detection architectures	Tech	22	44%
BP16: Participate in data sharing consortiums and initiatives	Tech	21	42%
BP17: Deriving probabilistic and risk models	Org	20	40%
BP18: Upper management buy in and support	Org	16	32%
BP19: Human-in-the-loop reinforcement	Proc	14	28%
BP20: Survey academic methods and techniques	Org	13	26%
BP21: Cyber risk as general enterprise risk & reward	Org	12	24%
BP22: Segment risk programmatically and outsource components	Org	9	18%
BP23: Adding machine learning to SIEM	Tech	5	10%
BP24: Preventative threat intelligence	Org	4	8%
BP25: Hosting and pushing detection to endpoints	Tech	4	8%
BP26: Honeypots to track and observe adversaries	Tech	2	4%



KEY CSDS GAPS: Factor-to-Factor Fitting

CH F1
Expansive complexity

CH F2
Tracking & context

CH F3
Data management

CH F4
Expectations versus limitations

CH F5
Unclear ownership

CH F6
Data policies

BP F1
Scientific process

BP F2
Cross-domain collaboration

BP F3
Risk management focus

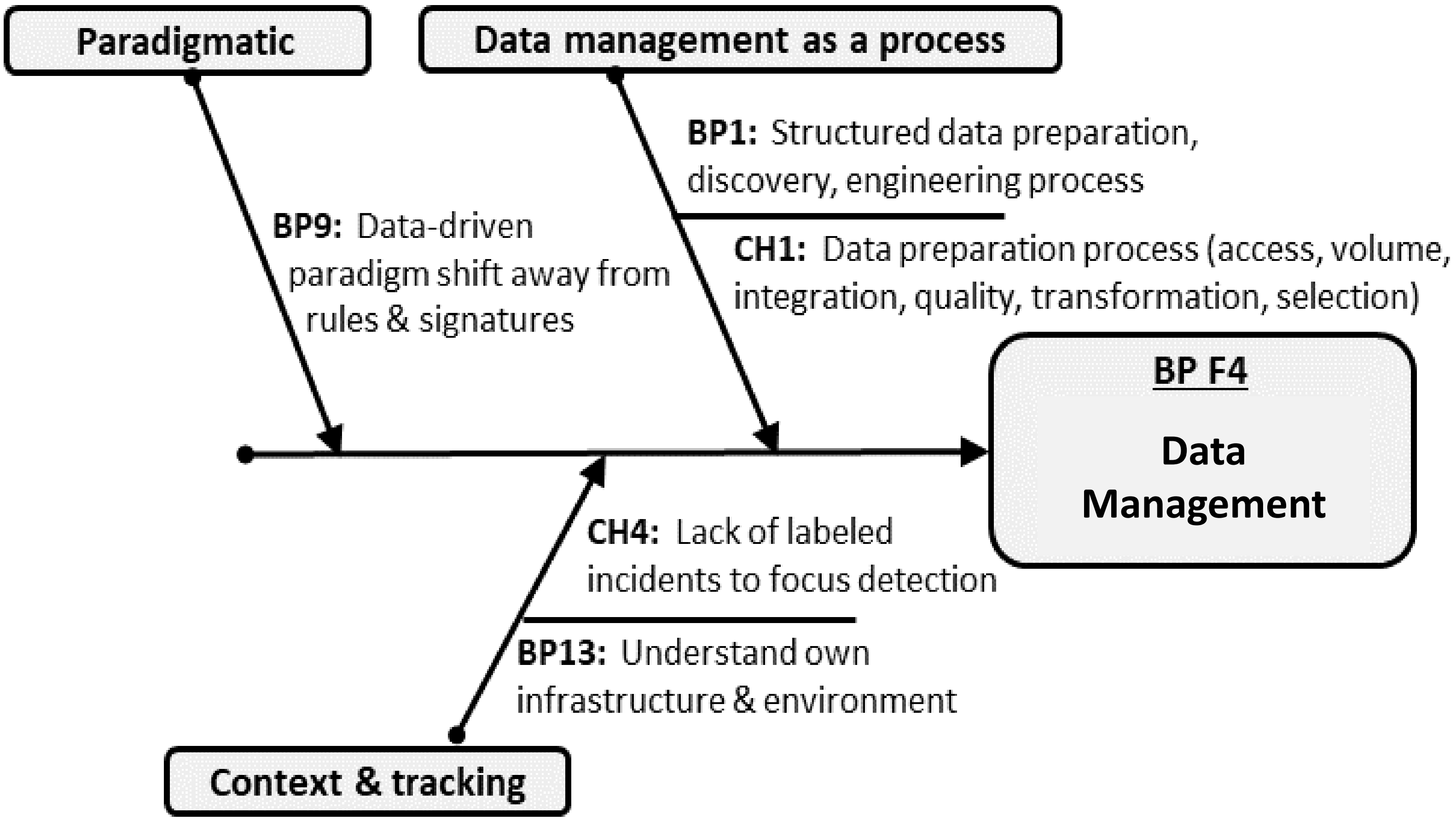
BP F4
Data-driven / data management

BP F5
Focused tools

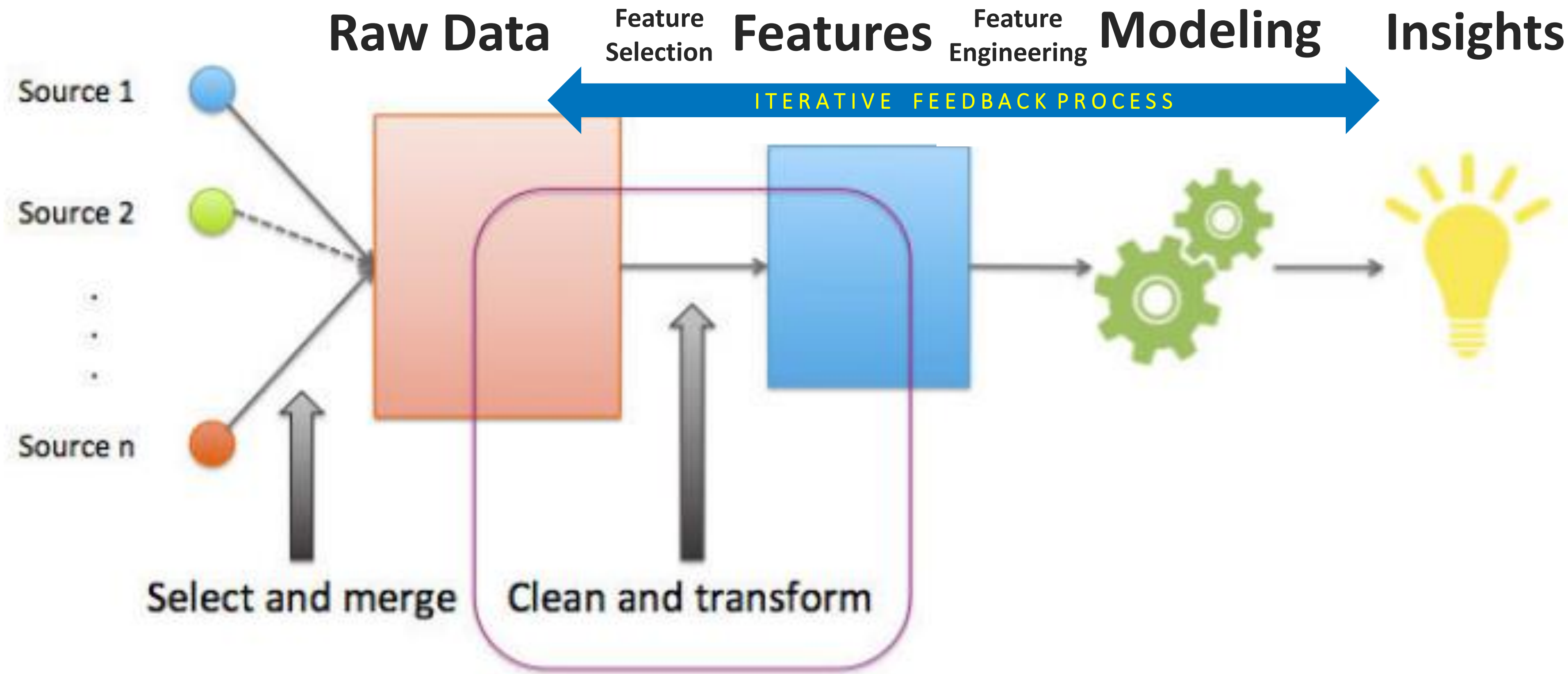
BP F6
Structured discovery process

III. CSDS Designs


ORGANIZATION
TECHNOLOGY
PROCESS
Data Management

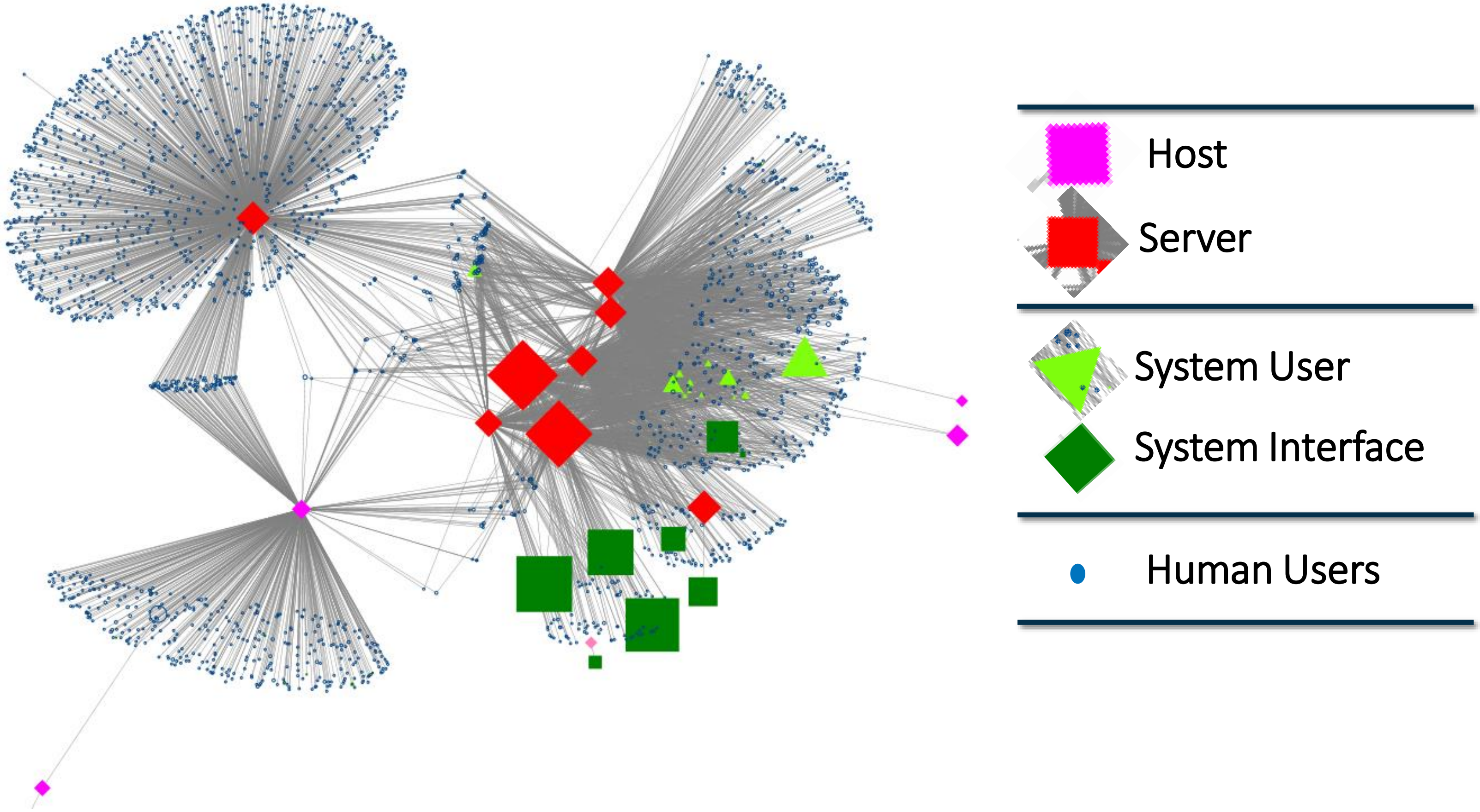

Paradigmatic
Data management as a process
BP1: Structured data preparation, discovery, engineering process
CH1: Data preparation process (access, volume, integration, quality, transformation, selection)
BP9: Data-driven paradigm shift away from rules & signatures
BP F4
Data Management
CH4: Lack of labeled incidents to focus detection
BP13: Understand own infrastructure & environment
Context & tracking

Data Management: EDA Process + Feature Engineering

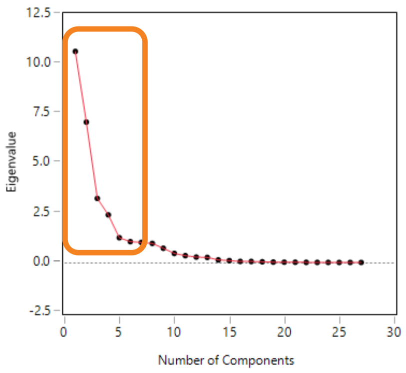


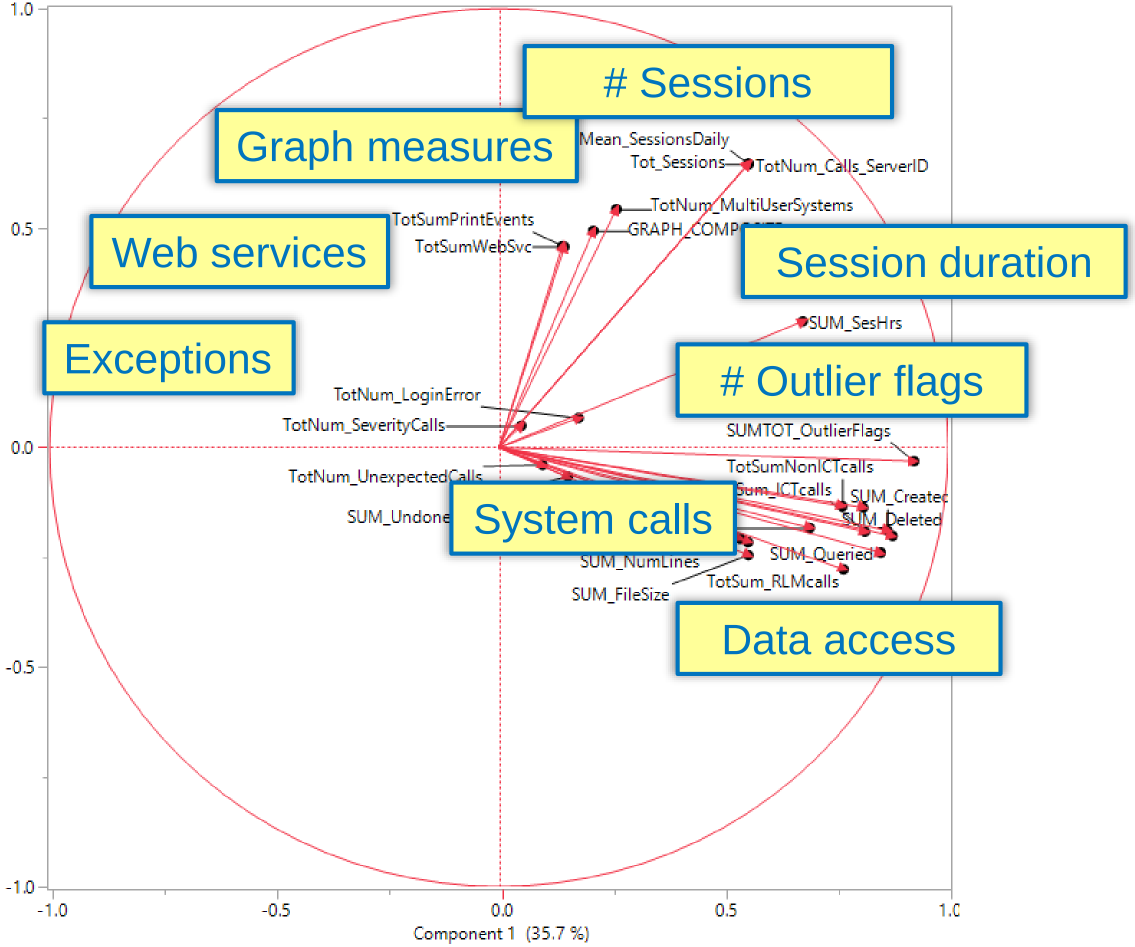
SOURCE: Alice Zheng, Amanda Casari. 2016. [Feature Engineering for Machine Learning Models](). O'Reilly Media.

Featurization: Example - Graph Analytics



Feature Reduction: Example - Principal Component Analysis (PCA)

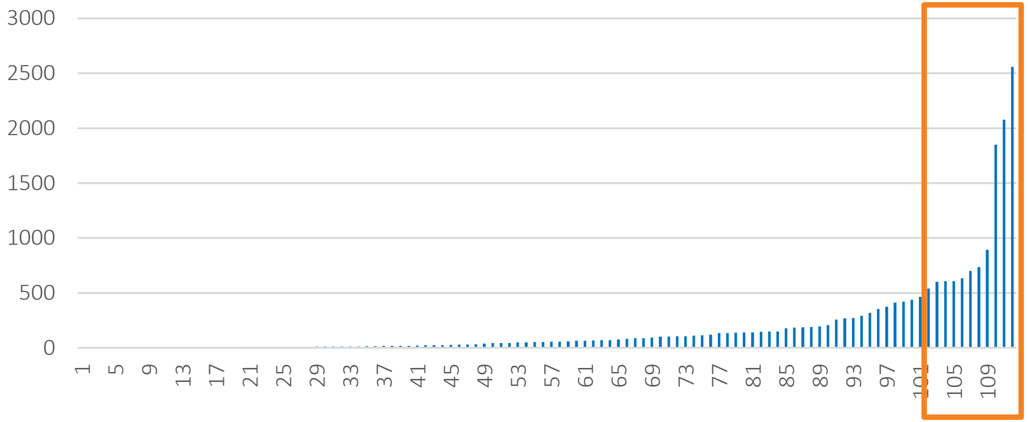


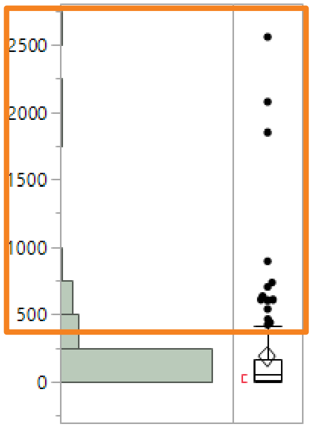


Exploratory Data Analysis (EDA): Example – Probabilistic Analysis

Exception Events

Exception messages per user (ranked)





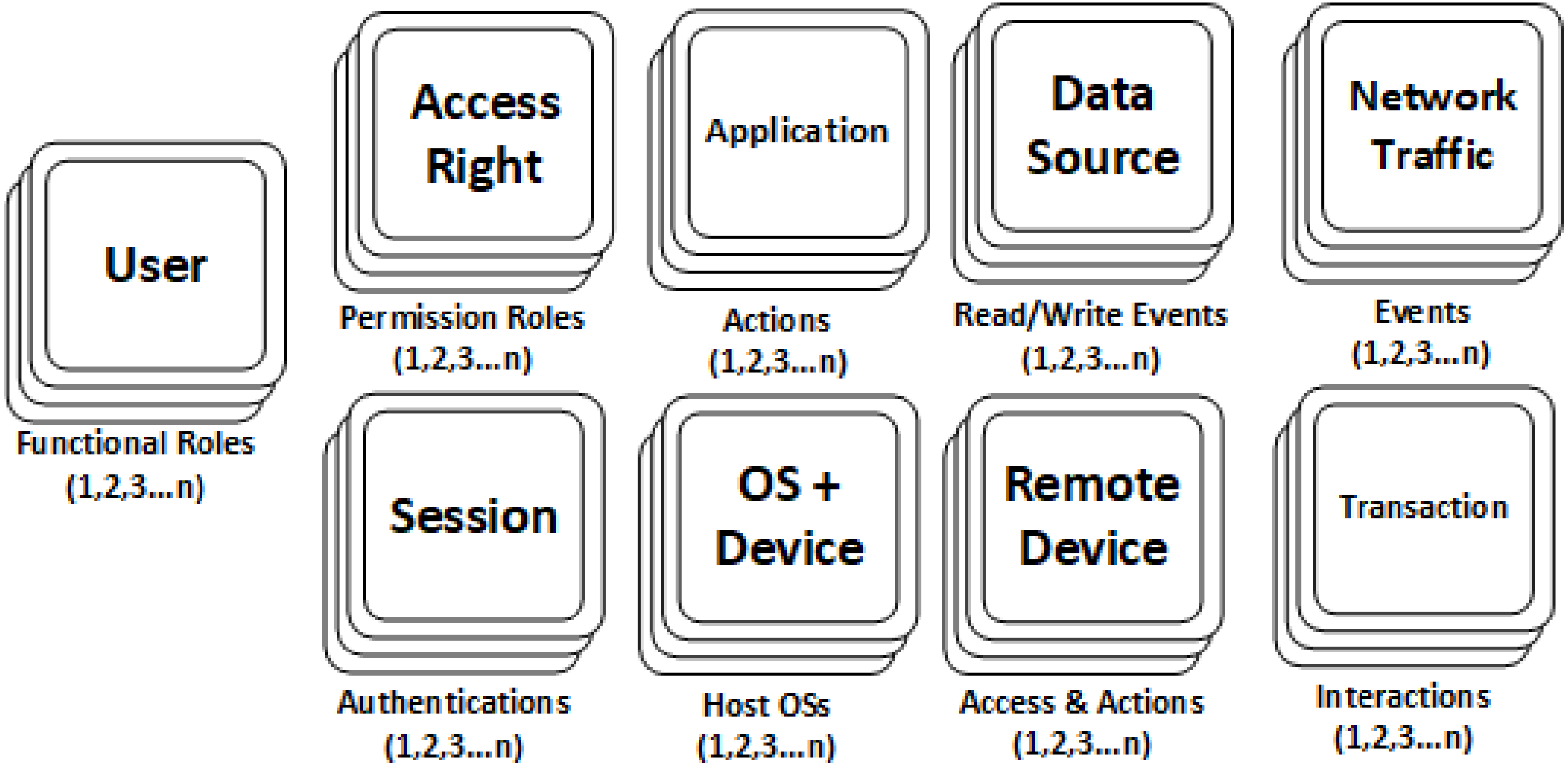
Quantiles

100.0%	maximum	2559
99.5%		2559
97.5%		1889.725
90.0%		517.5
75.0%	quartile	172.75
50.0%	median	55.5
25.0%	quartile	9.75
10.0%		3.3
2.5%		1.825
0.5%		1
0.0%	minimum	1

Summary Statistics

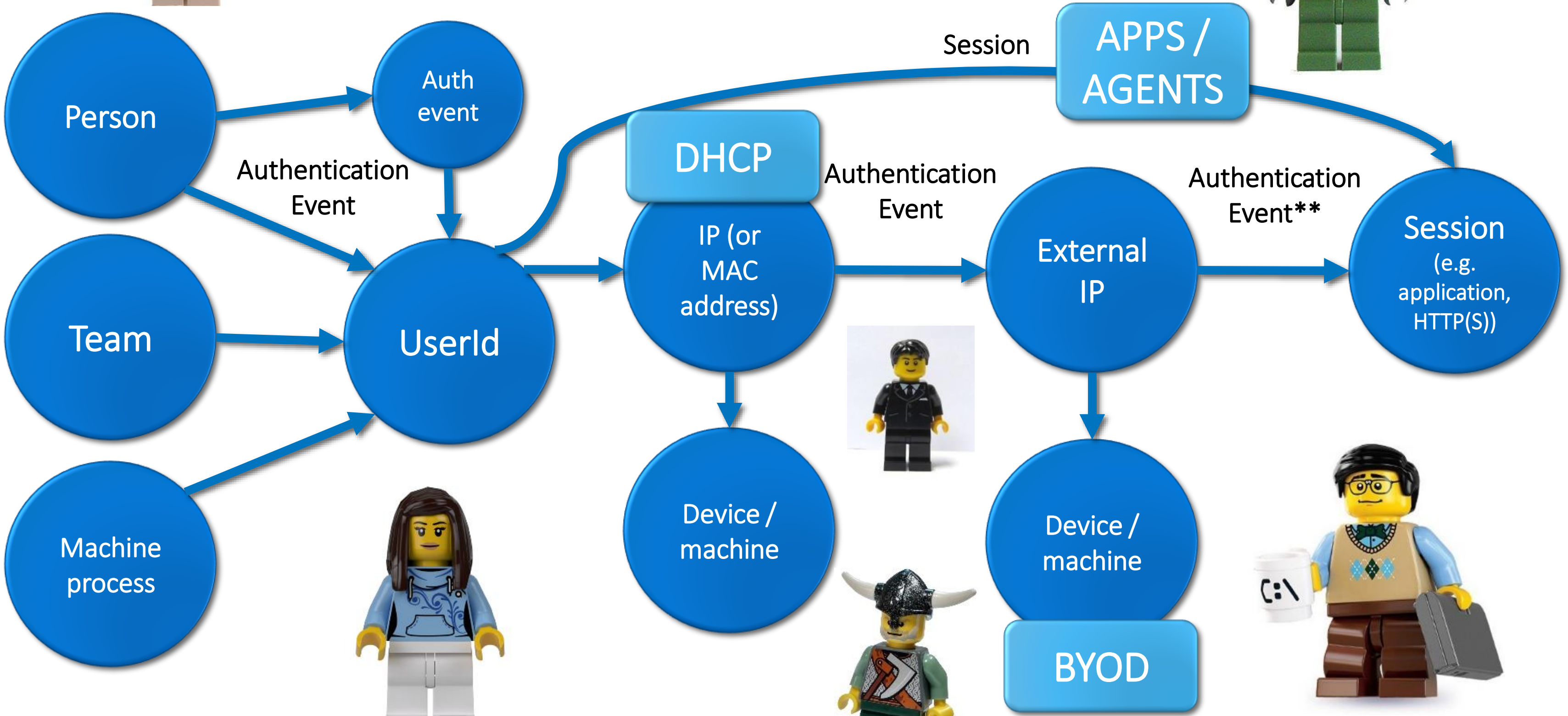
Mean	184.01786
Std Dev	380.96684
Std Err Mean	35.997982
Upper 95% Mean	255.35026
Lower 95% Mean	112.68545
N	112

Entity Resolution

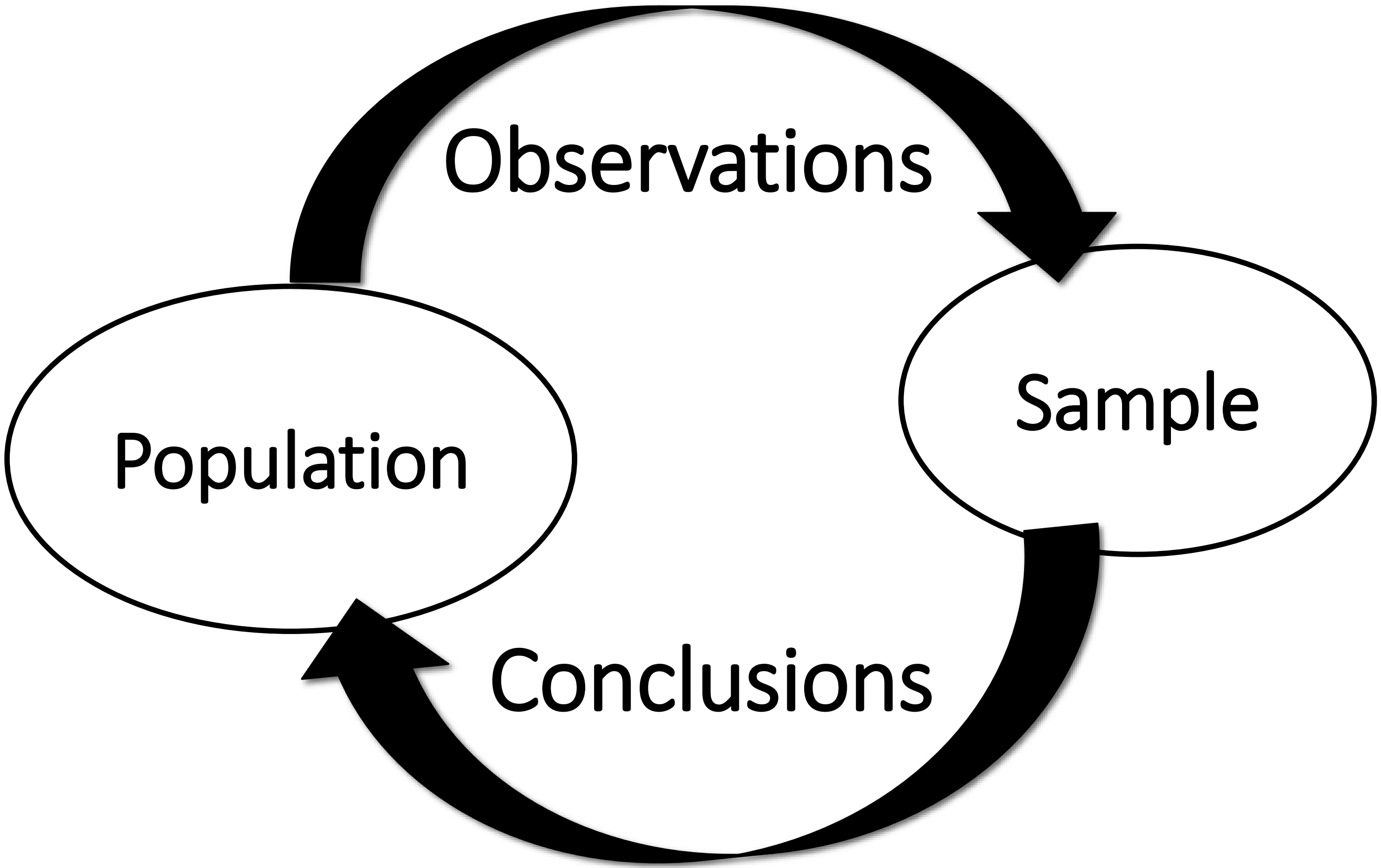


What is a User, anyway?

What is an IP address, anyway?

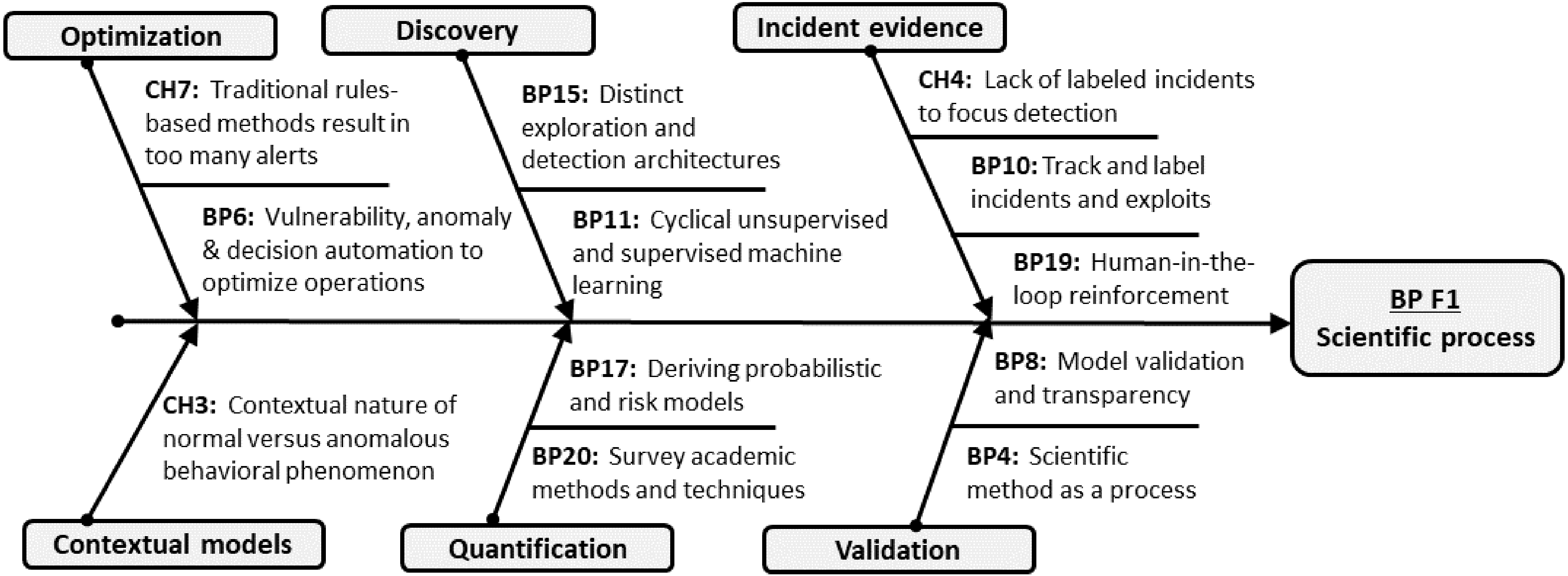


Inferential Statistics




ORGANIZATION
TECHNOLOGY
PROCESS
Scientific Process

Root Cause Analysis: Fishbone / Ishikawa Diagram

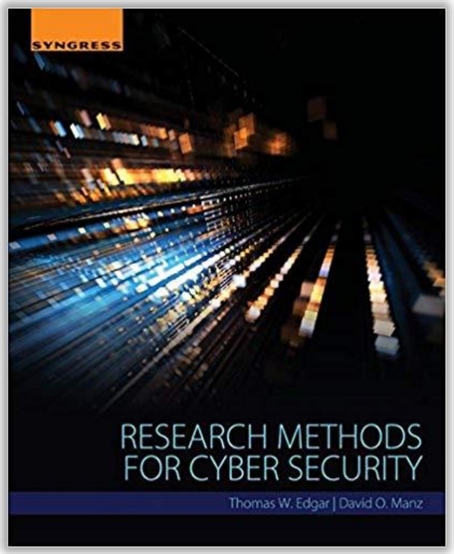


** Resulting from factor analysis and factor-to-factor fitting*

CSDS: What type of science is it?

Controlled experiments
versus
Pattern extrapolation

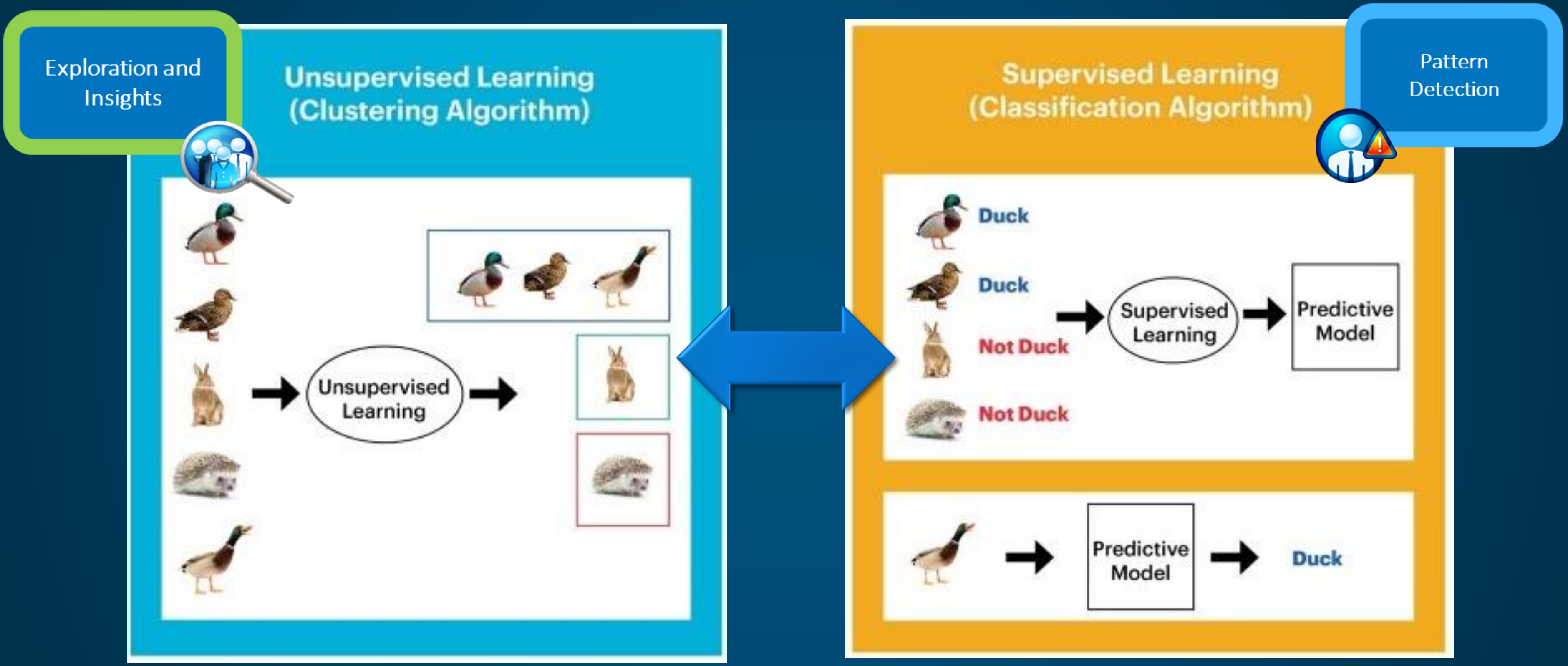
Research Methods for Cybersecurity



Manz, D. and Edgar, T. (2017)
Research Methods for Cyber Security

- *Experimental*
  - i.e. hypothetical-deductive and quasi-experimental
- *Applied*
  - i.e. applied experiments and observational studies
- *Mathematical*
  - i.e. theoretical and simulation-based
- *Observational*
  - i.e. exploratory, descriptive, machine learning-based

Discovery ⇔ Detection



SEGMENTATION

CATEGORIZATION

Labels: What constitutes 'evidence'?

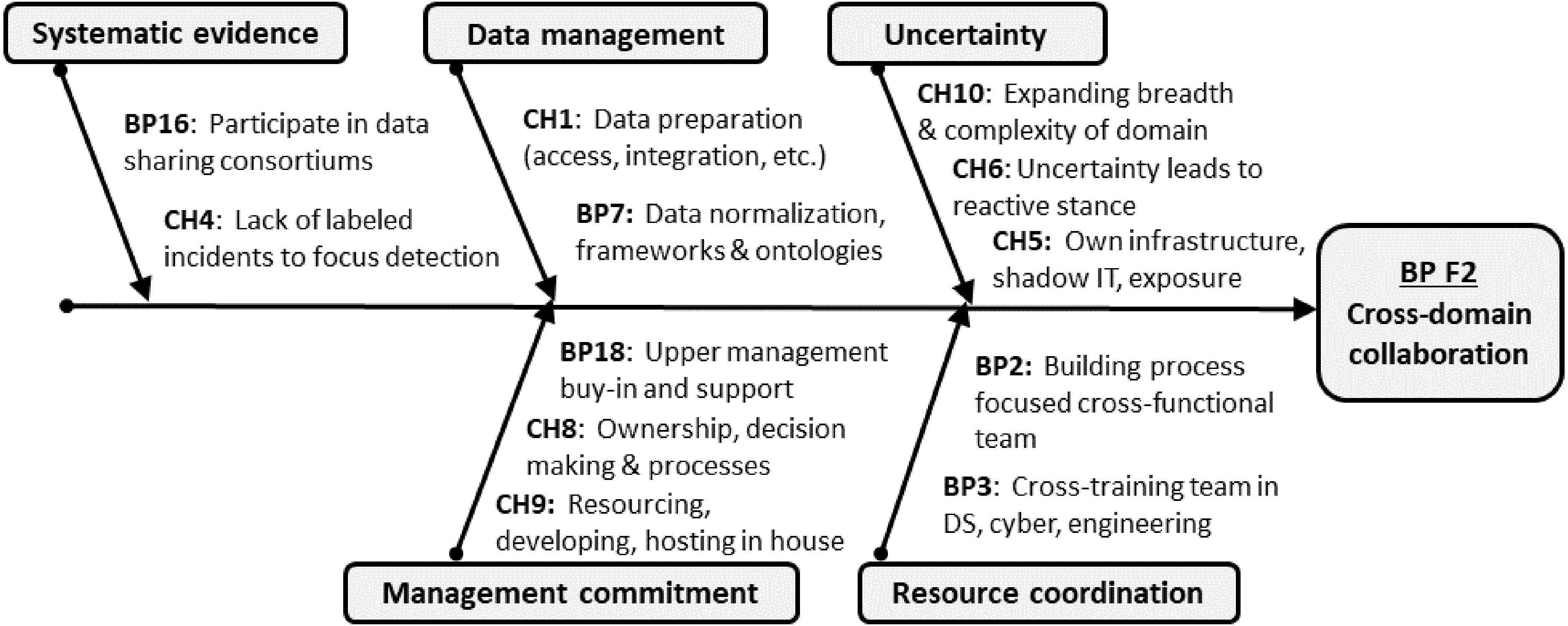
	Inductive	Deductive
Collected	- Field evidence - Probing & testing - 3rd party sourced	- Rules & signatures - Research & threat intelligence
Synthesized	- Red Teaming - Simulations - Laboratory	- Expert opinion - Thought experiments

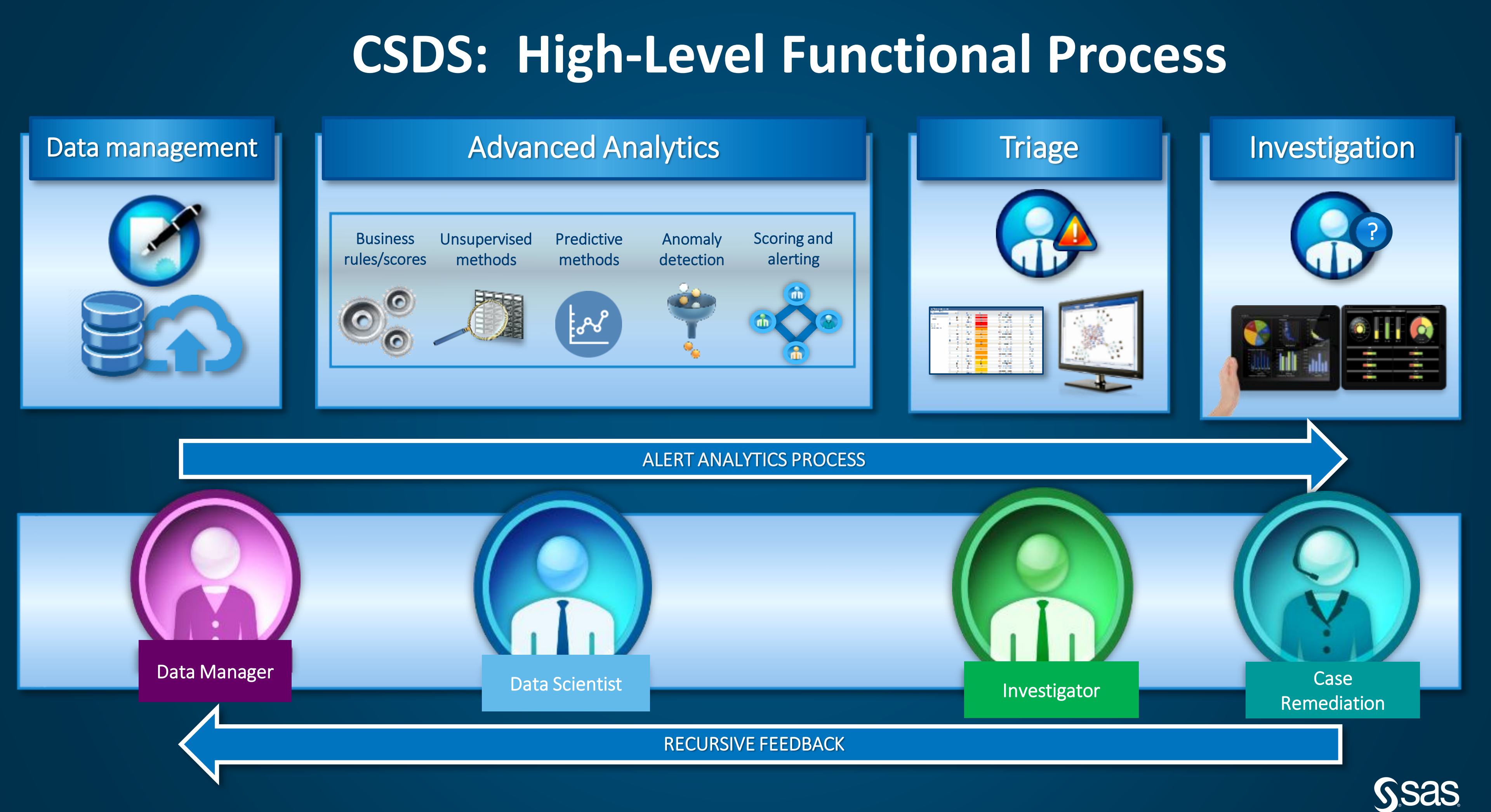
EXAMPLES OF SECURITY EVIDENCE

1. Field evidence (e.g. observed incidents)
2. Sourcing own data from field testing (e.g. local experiments)
3. Honeypots
4. IDSs (Intrusion Detection Systems)
5. Simulation findings
6. Laboratory testing (e.g. malware in a staged environment)
7. Stepwise discovery (iterative interventions)
8. Pen testing (attempts to penetrate the network)
9. Red teaming (staged attacks to achieve particular goals)
10. Incidents (records associated with confirmed incidents)
11. Reinforcement learning (self-improving ML to achieve a goal)
12. Research examples (datasets recording attacks from research)
13. Expert review (opinion and guidance from experts)
14. Intelligence feed (indications from a 3rd party service)
15. Thought experiments (e.g. boundary conditions, counterfactuals)

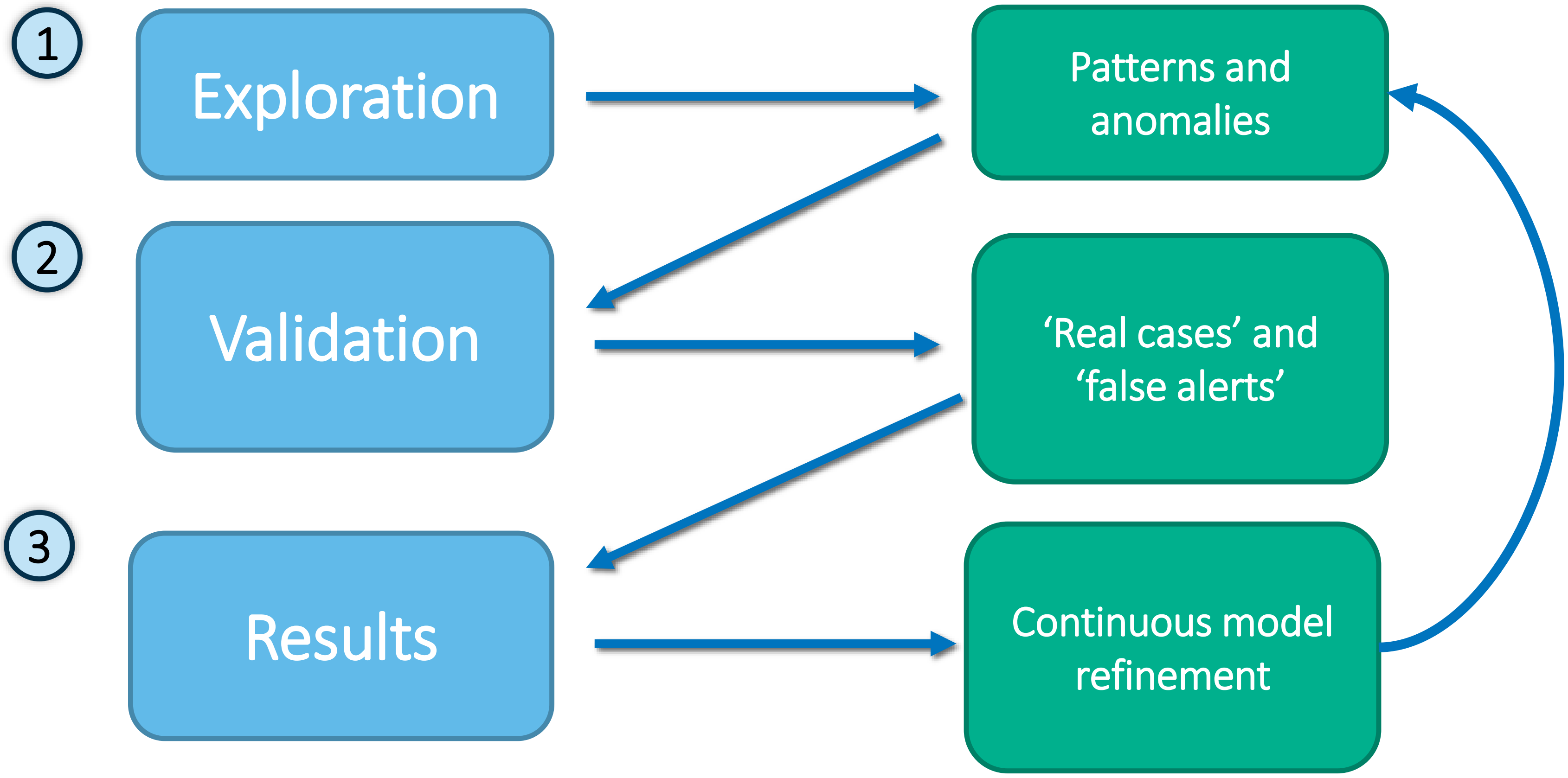
CSDS as a Process: Discovery and Detection


Cross-Domain Collaboration
ORGANIZATION
TECHNOLOGY
PROCESS

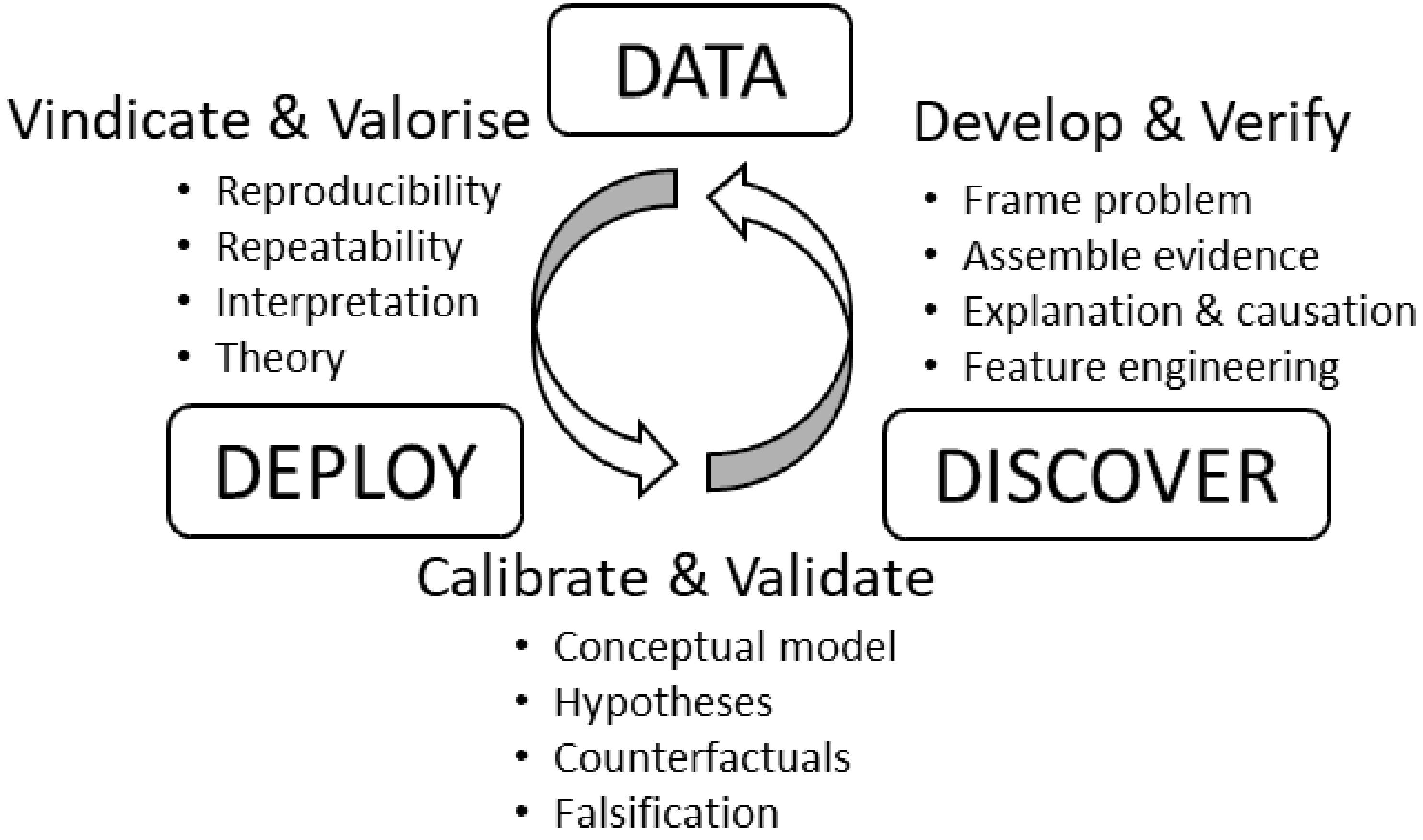

Systematic evidence
Data management
Uncertainty
BP16: Participate in data sharing consortiums
CH4: Lack of labeled incidents to focus detection
CH1: Data preparation (access, integration, etc.)
BP7: Data normalization, frameworks & ontologies
CH10: Expanding breadth & complexity of domain
CH6: Uncertainty leads to reactive stance
CH5: Own infrastructure, shadow IT, exposure
BP F2
Cross-domain collaboration
BP18: Upper management buy-in and support
CH8: Ownership, decision making & processes
CH9: Resourcing, developing, hosting in house
BP2: Building process focused cross-functional team
BP3: Cross-training team in DS, cyber, engineering
Management commitment
Resource coordination


CSDS: High-Level Functional Process
Data management
Advanced Analytics
Business rules/scores
Unsupervised methods
Predictive methods
Anomaly detection
Scoring and alerting
Triage
Investigation
ALERT ANALYTICS PROCESS
Data Manager
Data Scientist
Investigator
Case Remediation
RECURSIVE FEEDBACK
sas

Continuous Detection Improvement Process



CSDS Model Development Process



Conclusions

Cybersecurity ✓

Data ✓

Science ?

Not so much...

but, ASPIRATIONAL!


EST COFFEE
Auntie Anne's
WAY OUT WEST
NEWSSTAND


HUMAN
ORGAN
FOR TRANSPLANT

CSDS: A Work in Progress

- Process of Professionalization
  - Named professionals
  - Set of methods and techniques
  - Standards, best practices

Training programs

Certifications

Academic degree programs

Focused research journals

Formal sub-specialization

Specialist Researcher Primary Care

Surgeon Diagnostician Emergency Care

APPENDIX

References

- Aggarwal, C. (2013). "Outlier Analysis." Springer. http://www.springer.com/la/book/9781461463955
- 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://www.sas.com/en_us/whitepapers/hbr-defending-your-networks-108030.html
- Longitude Research. (2014). "Cyberrisk in banking." Available at https://www.sas.com/content/dam/SAS/bp_de/doc/studie/ff-st-longitude-research-cyberrisk-in-banking-2316865.pdf
- 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
- SAS Institute. (2019). "Data Management for Artificial Intelligence." Available at www.sas.com/en_us/whitepapers/data-management-artificial-intelligence-109860.html
- 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
- 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