Cybersecurity Data Science (CSDS)

Best Practices in an Emerging Profession

Scott Allen Mongeau
INFORMS CAP®

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

s.mongeau@edp1.nyenrode.nl
scott@sark7.com
scott.mongeau@sas.com

@SARK7 #CSDS2020

INTRODUCTION

- ~30 years
  - IT / data analysis and data management
  - Statistics, analytics, simulation, data science...
- Cybersecurity Data Science
  - *SAS Institute & Deloitte (~7 yrs)*
- Technical & management consulting
  - Bio-pharma, telecom, finance, public sector
  - Military, defense, intelligence, security, policing
- Guest lecturer / PhD candidate
  - *Nyenrode University, Netherlands*

Cybersecurity Data Science (CSDS): Best Practices in an Emerging Profession

Scott Mongeau, EDP PhD candidate

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

NYENRODE BUSINESS UNIVERSITY

Cybersecurity Data Science (CSDS): Best Practices in an Emerging Profession

Scott Mongeau, EDP PhD candidate

I. Research Overview
II. Literature
III. Interviews
IV. Designs
V. Conclusions

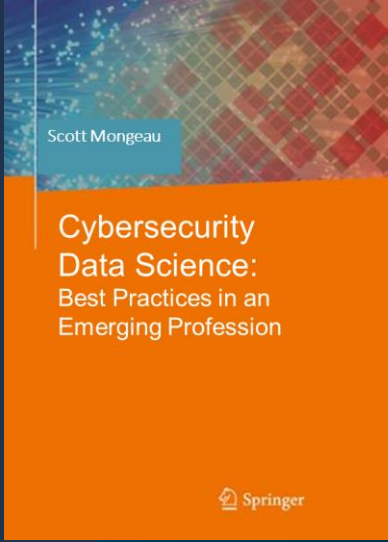
NYENRODE BUSINESS UNIVERSITY

LEADERSHIP,
ENTREPRENEURSHIP,
STEWARDSHIP

I. Research Overview

PhD academic research / book

- ~July 2020 release



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

I. <u>Literature</u>: What is CSDS? Status as a profession?

II. <u>Interviews</u>: 50 CSDS practitioners

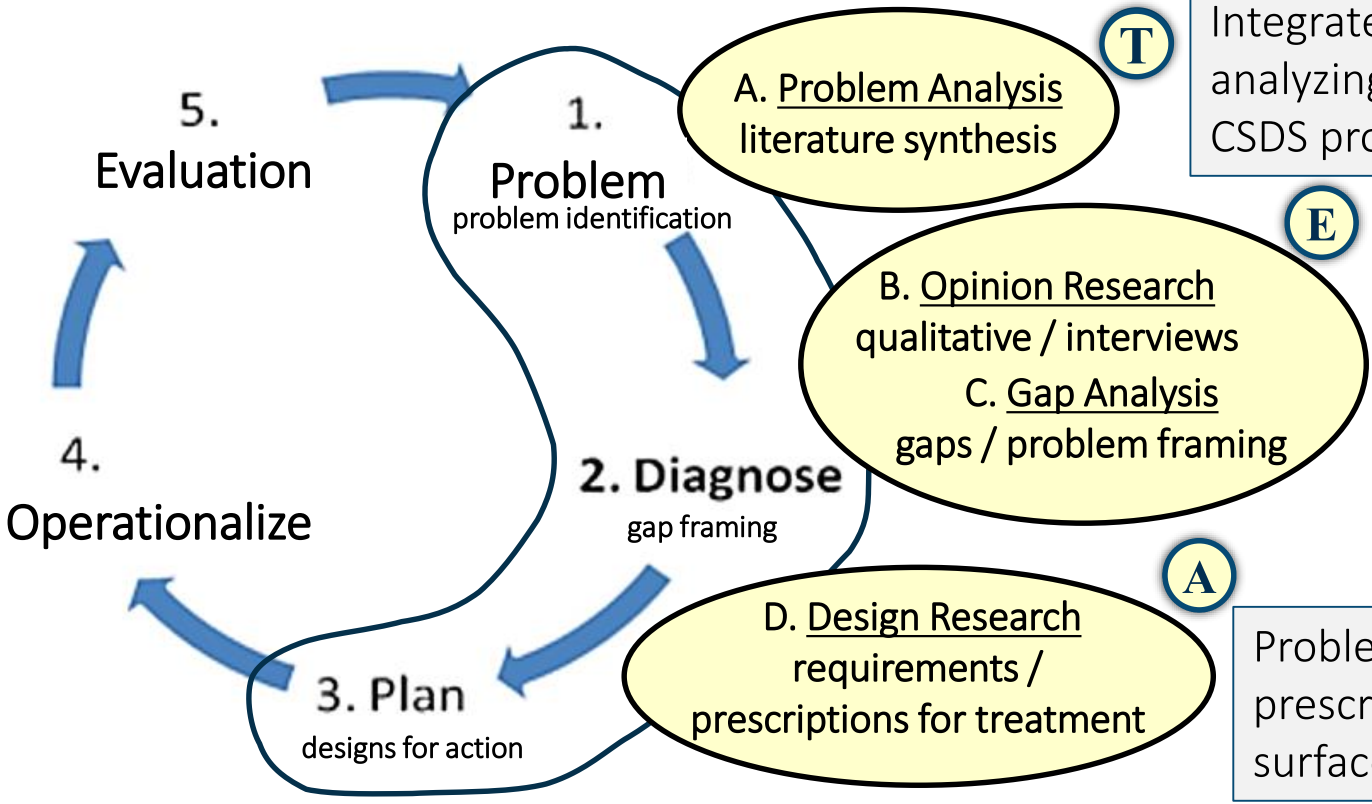
III. <u>Designs</u>: Approaches to address challenges

PhD academic research / book

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

- What is data science with respect to cybersecurity?
  - Professionalization maturity / best practices gap diagnosis
- Triangulated mixed methods
  - Qualitative and quantitative (inductive focus)
  - Literature review, interview coding, text analytics
- Gap analysis leading to design prescriptions

Practitioner Diagnostic & Design Research



Integrated literature review analyzing challenges related to CSDS professionalization

Interviews with 50 CSDS professionals leading to diagnosis of perceived gaps impeding CSDS practice

Problem solving design prescriptions to address surfaced gaps

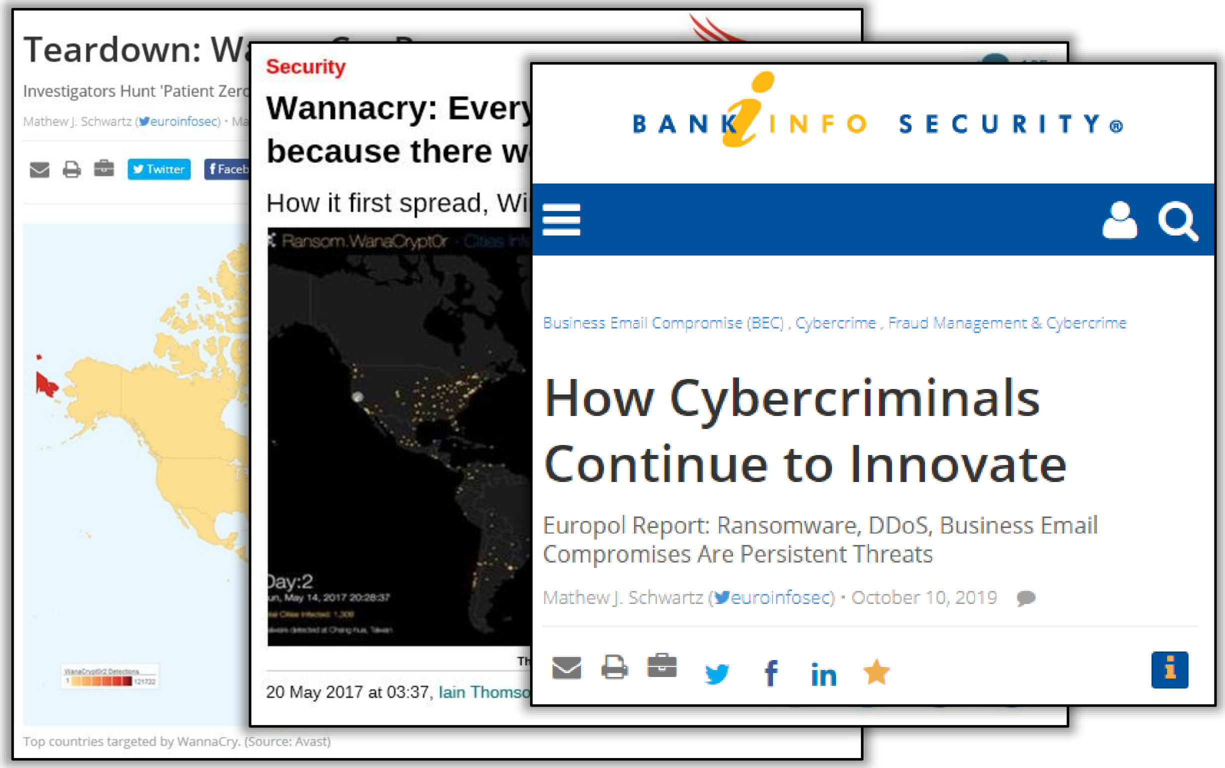
Management of Information Systems (MIS)

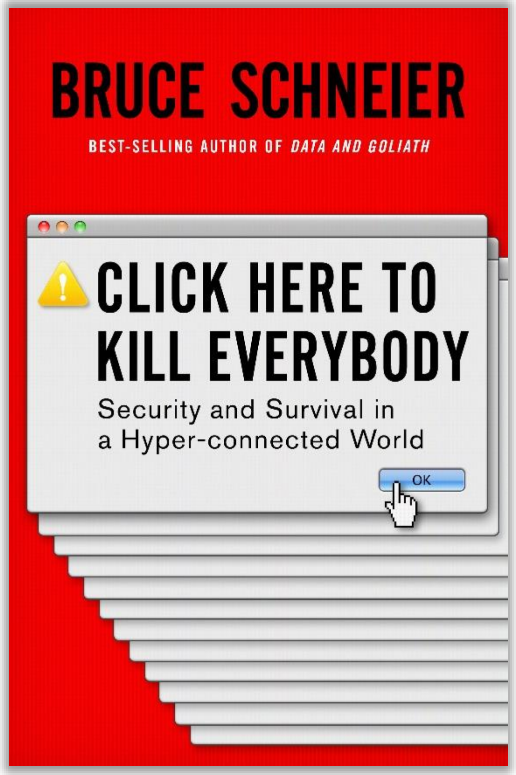
Haag & Cummings, 2012
Hsu, 2013
Laudon & Laudon, 2017
Pearlson, Saunders, & Galletta, 2016
Sousa & Oz, 2014

II. CSDS Literature

FUD Fear, Uncertainty, Doubt

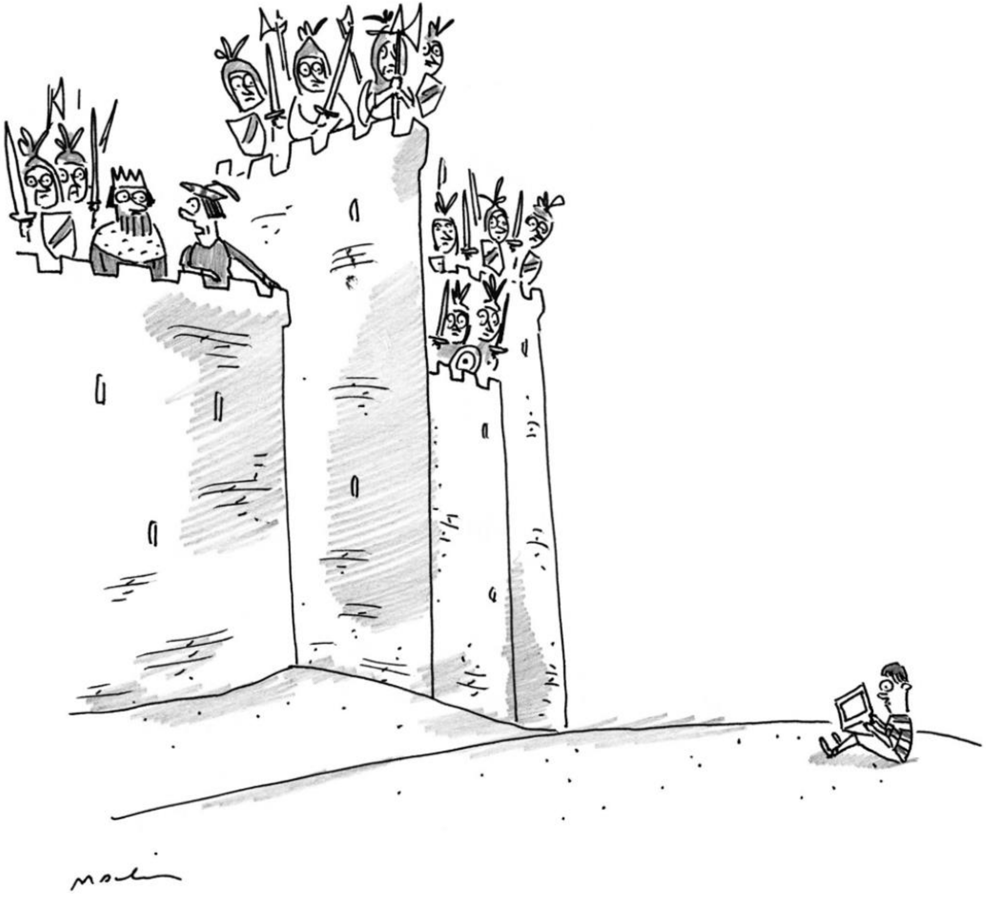
Expansion of exposure and targets >!< Increasing sophistication, frequency, and speed of attacks



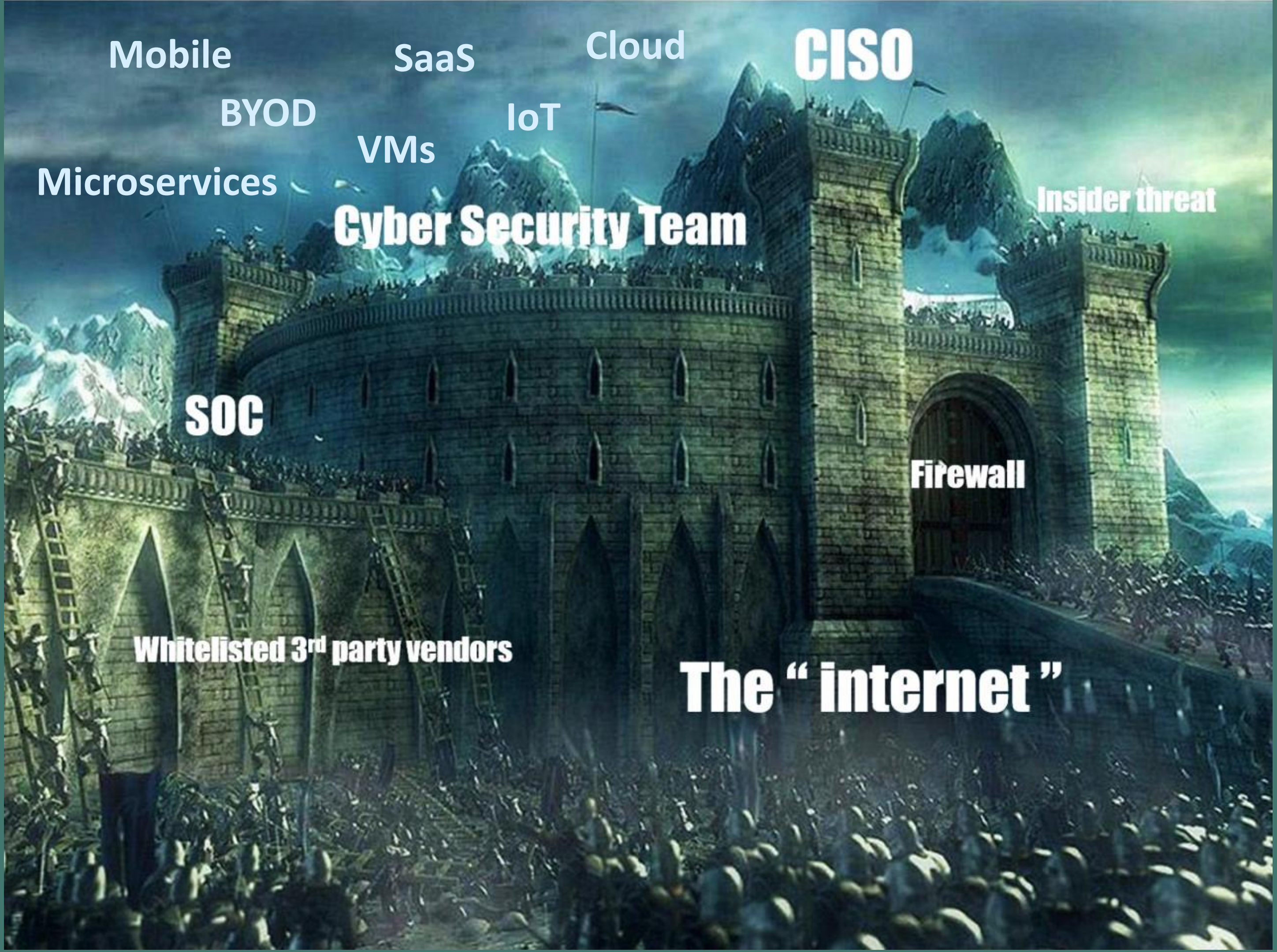


Castle and Moat

How quaint!



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


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

Cybersecurity Challenges

DATA DISCONNECTED & FRAGMENTED

LACK OF CONTEXT

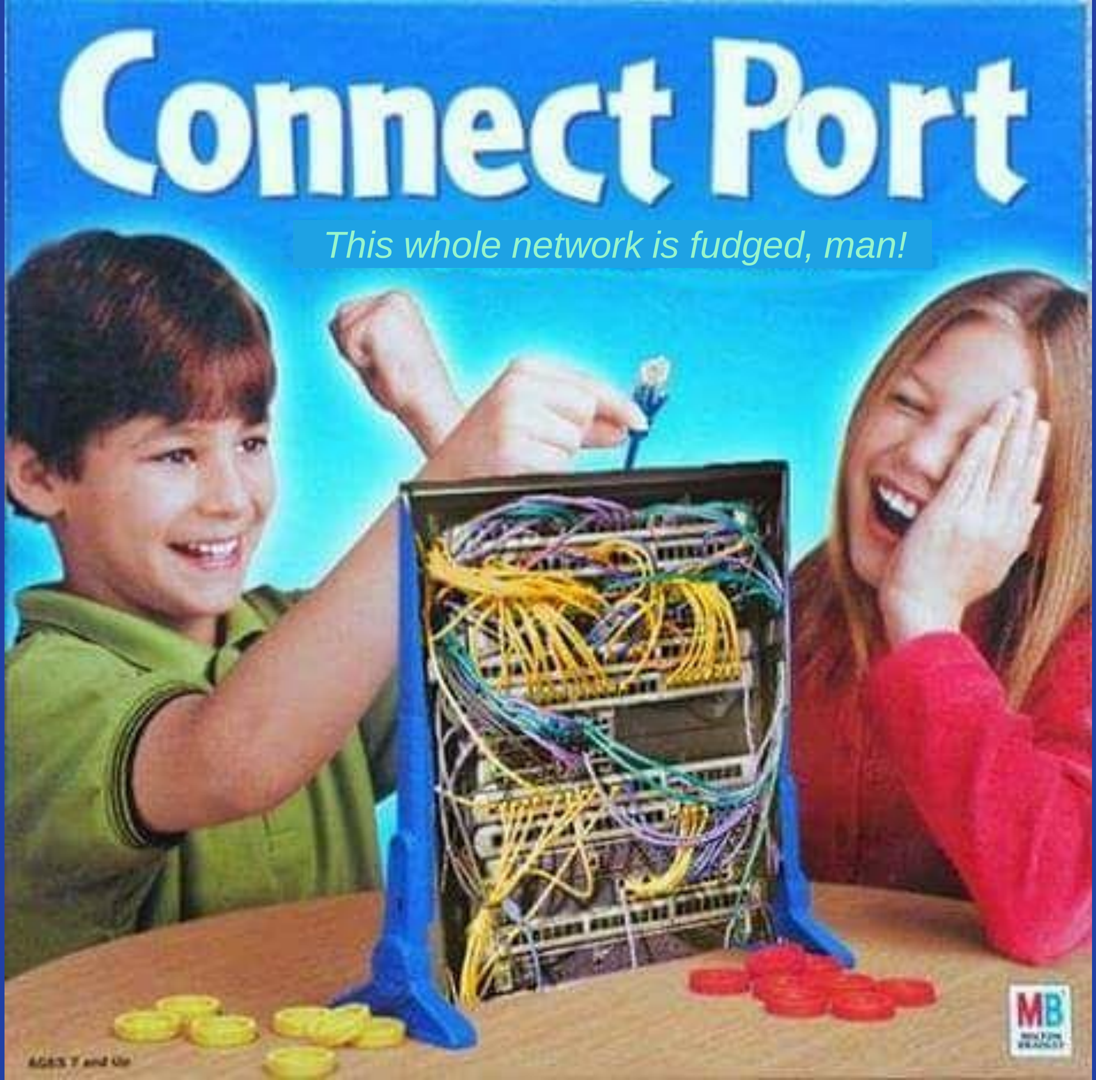
LIMITED STAFF

DATA VOLUME & SPEED

MULTIPLE SYSTEMS & ALERTS

Data Science

New hope amidst complexity and confusion…



CSDS

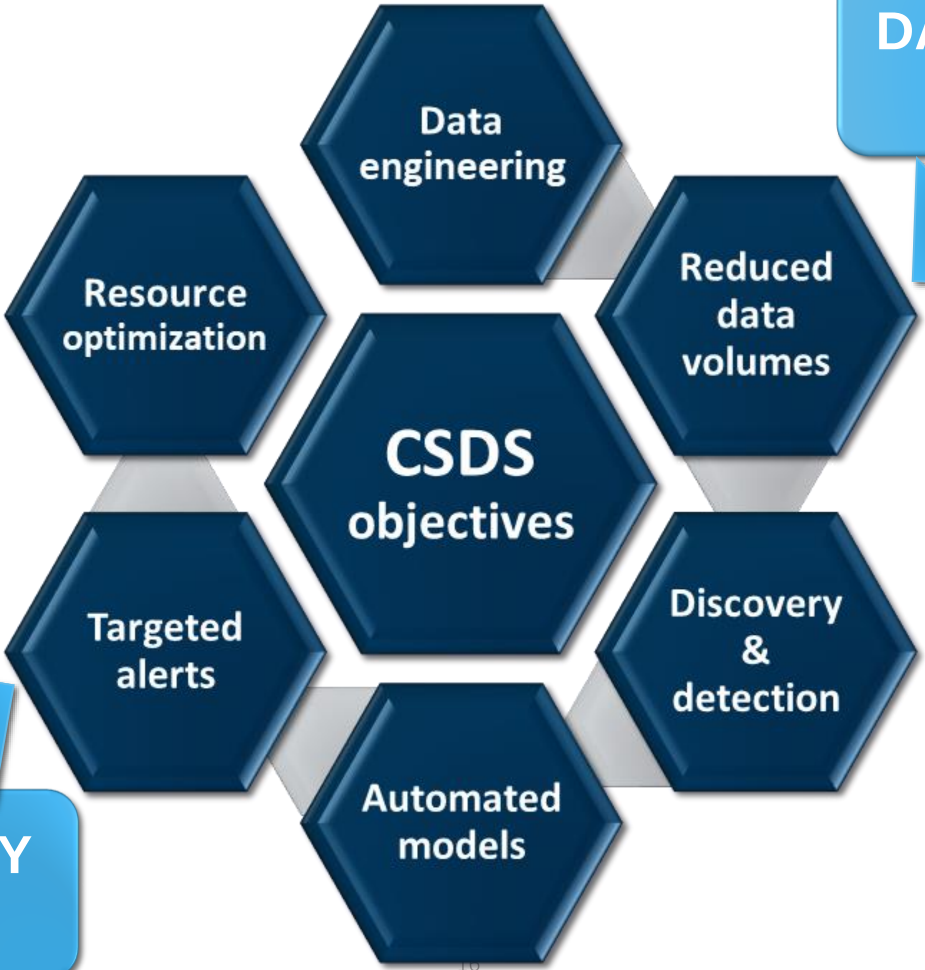
Cyber

Security

Data

Science

DATA SCIENCE METHODS

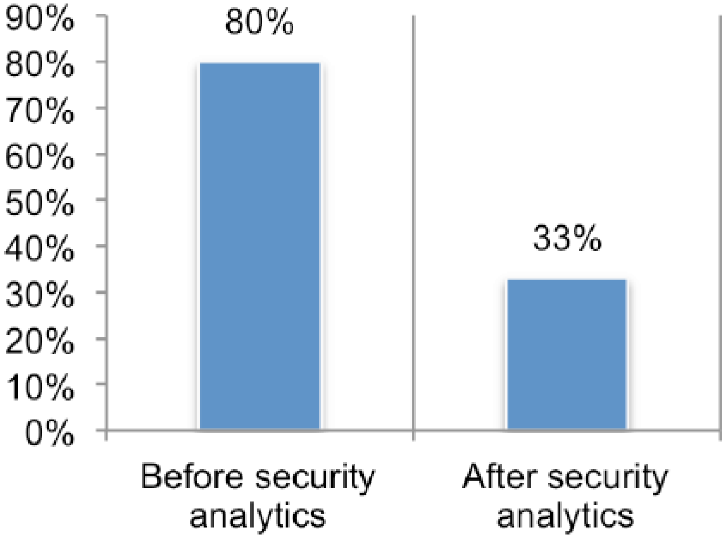


CYBERSECURITY GOALS

CSDS: Existing Professionals + Demonstrated Efficacy

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*

EXAMPLE CSDS PRACTICAL APPLICATIONS

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

Derived Professionalization Assessment Model

Professional maturity

1	Systematic body of theory
2	Authority and judgement recognized by client
3	Community sanctions authority
4	Ethical code of stewardship
5	Professional culture supported by associations

Greenwood, E. (1957). Attributes of a Profession. *Social Work, 2*, 11.

Van der Krogt, T. (2015). Professionals and Their Work.

Professional emergence

1	Active, focused interest from diverse participants
2	Active professionals with associated job titles & roles
3	Emerging and informal training
4	Informal professional groups
5	Professional and industry literature
6	Research literature
7	Formalized training
8	Formal professional groups
9	Professional certifications
10	Standards bodies
11	Independent academic research disciplinary focus

Beer, J. T., & Lewis, W. D. (1963). Aspects of the Professionalization of Science. *The MIT Press, 92*(4), 20.

Freidson, E. (2001). *Professionalism: The Third Logic*. Cambridge, MA, U.S.: Polity Press.

CYBERSECURITY PROFESSION

STRATEGIC

TACTICAL

Operational Security Design

- Trend and historical monitoring
- Security architecture and models
- Identity and access management (IAM)
- Software & application development
- Encryption architecture and infrastructure
- Systems integration planning
- Cloud administration
- IoT and ICS security architecture
- Asset & physical security design and documentation

Management and Stewardship

- Assurance, audit, and compliance
- Governance, policy, procedures, & standards
- Security leadership and management practices
- Risk management
- Law, investigation, and ethics
- Training and development
- User education

SECURITY ASSURANCE

Security Monitoring

- Continuous monitoring
- Communication and network security administration
- Cryptographic mechanisms and administration
- Endpoints and device security monitoring
- Asset, physical, and environmental security
- Cloud security monitoring
- Internet of Things (IoT) monitoring
- Industrial control systems monitoring

Readiness and Response

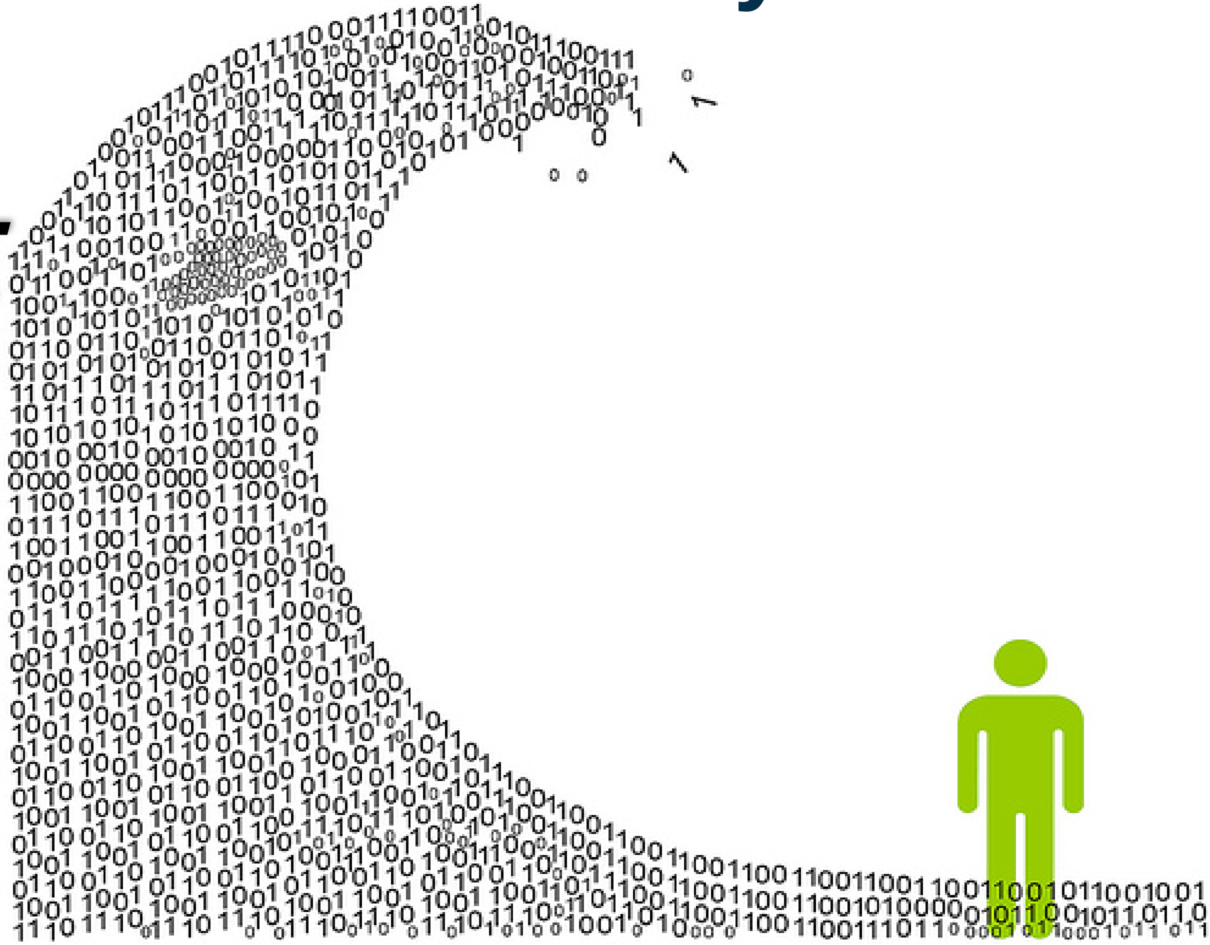
- Business continuity and disaster recovery
- Incident response
- Computer / digital forensics
- Vulnerability assessment
- Threat assessment, intelligence, and research
- Adversarial techniques
- Ethical hacking
- Penetration testing
- Red teaming

TECHNICAL ⟷ ORGANIZATIONAL

Data Science is Everywhere!

DATA SCIENCE PROFESSION

Market Hype



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?"

Phantom Patterns: Correlation ≠ Causation

The Ghost of Christmas Overfitting comes to visit

Are you or a friend addicted to predictive machine learning?

Key warning signs:

- Throwing 800 variables into a model and running with a good ROC score
- Need to retrain your model every three weeks?
- "Explanation!? We don't need no stinkin' explanation!"

If so, call 1-800-DIAGNOSTICS now!

CSDS Body of Literature (book length works)

1	Machine Learning and Data Mining for Computer Security: Methods and Applications	* Maloof ed., 2006
2	Intrusion Detection: A Machine Learning Approach	Yu & Tsai, 2011
3	Data Mining and Machine Learning in Cybersecurity	Dua & Du, 2011
4	Network Anomaly Detection: A Machine Learning Perspective	Bhattacharyya & Kalita, 2013
5	Applied Network Security Monitoring	Sanders & Smith, 2013
6	Network Security Through Data Analysis	Collins, 2014
7	Data Analysis for Network Cyber-Security	* Adams & Heard eds., 2014
8	Data-Driven Security	Jacobs & Rudis, 2014
9	Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques	Baesens, Van Vlasselaer, & Verbeke, 2015
10	Essential Cybersecurity Science	Dykstra, 2016
11	Dynamic Networks and Cyber-Security	Adams & Heard, 2016 *
12	Cybersecurity and Applied Mathematics	Metcalf & Casey, 2016

13	How to Measure Anything in Cybersecurity Risk	Hubbard & Seiersen, 2016
14	Data Analytics and Decision Support for Cybersecurity	* Carrascosa, Kalutarage, & Huang eds., 2017
15	Research Methods for Cybersecurity	Edgar & Manz, 2017
16	Introduction to Machine Learning with Applications in Information Security	Stamp, 2017
17	Information Fusion for Cyber-Security Analytics	* Alsmadi, Karabatis, & AlEroud eds., 2017
18	Machine Learning & Security	Chio & Freeman, 2018
19	Data Science for Cybersecurity	Heard, Adams, Rubin-Delanchy, & Turcotte eds., 2018
20	AI in Cybersecurity	* Sikos ed., 2018
21	Malware Data Science: Attack Detection and Attribution	Saxe & Sanders, 2018
22	Machine Learning for Computer and Cyber Security	* Gupta & Sheng eds., 2019
23	Cybersecurity Analytics	Verma & Marchette, 2019

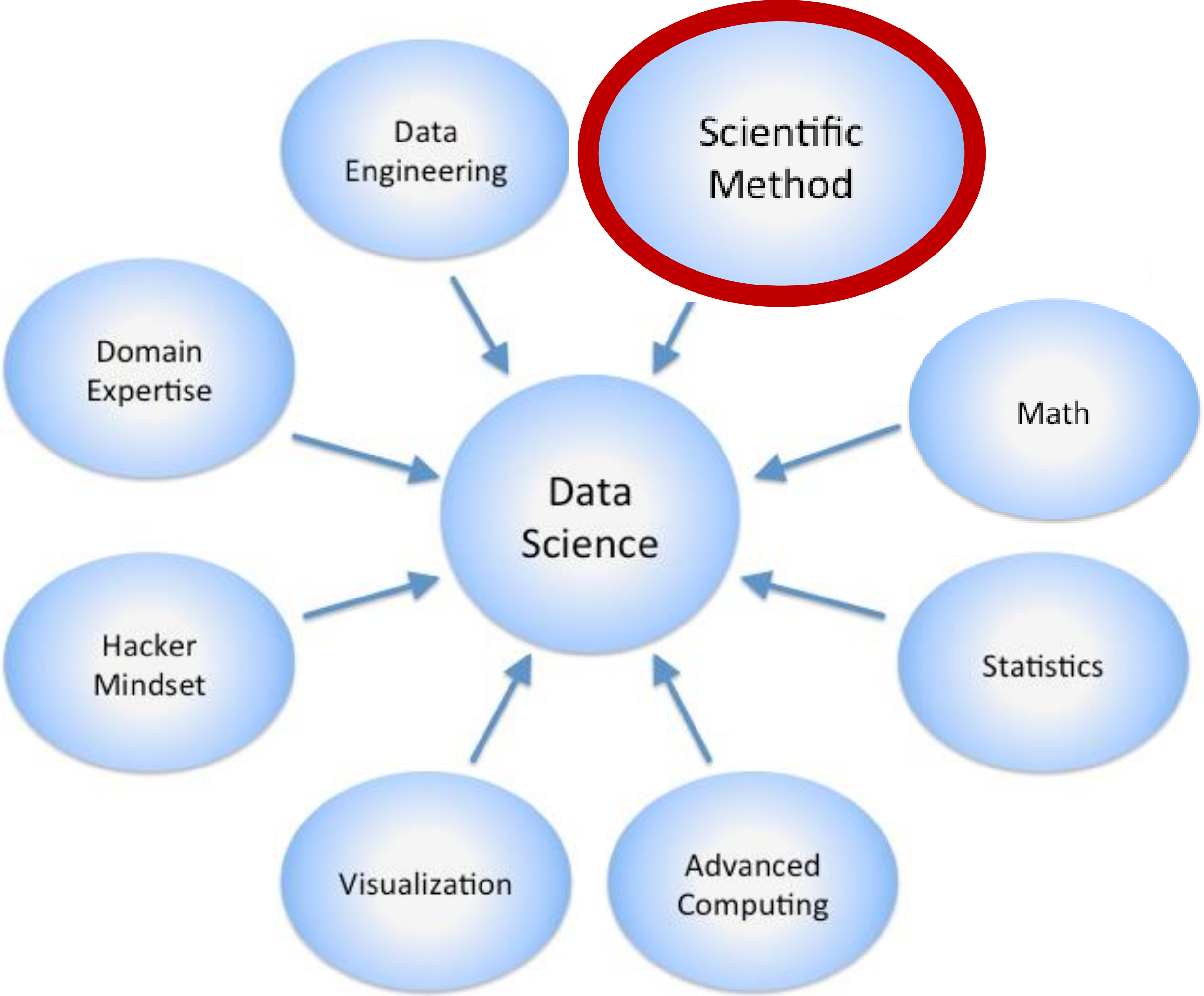
Email me if there is a CSDS book you feel should be added! scott@sark7.com

#	Title	Authors	Focused Use Cases	Risk Quantification	Decision Support	Data Management	Data Collection	Scientific Methods	Feature Engineering	Statistical Methods	Anomaly Detection	Machine Learning	Model Management	Visualization	Adversarial Methods	Organizational Management
1	Intrusion Detection: A Machine Learning Approach	Yu & Tsai, 2011	✓							✓	✓	✓			✓	
2	Data Mining and Machine Learning in Cybersecurity	Dua & Du, 2011	✓		✓	✓			✓	✓	✓	✓	✓	✓		
3	Network Anomaly Detection: A Machine Learning Perspective	Bhattacharyya & Kalita, 2013	✓		✓		✓		✓	✓	✓	✓	✓	✓	✓	
4	Applied Network Security Monitoring	Sanders & Smith, 2013	✓	✓	✓	✓	✓		✓	✓	✓			✓		✓
5	Network Security Through Data Analysis	Collins, 2014	✓		✓	✓	✓									
6	Data Analysis for Network Cyber-Security	Adams & Heard, 2014 *	✓		✓		✓									
7	Data-Driven Security	Jacobs & Rudis, 2014	✓	✓	✓	✓	✓									✓
8	Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques	Baesens, Van Vlasselaer, & Verbeke, 2015	✓	✓	✓	✓	✓									✓
9	Essential Cybersecurity Science	Dykstra, 2016	✓	✓	✓	✓	✓									✓
10	Dynamic Networks and Cyber-Security	Adams & Heard, 2016 *	✓	✓			✓									
11	Cybersecurity and Applied Mathematics	Metcalf & Casey, 2016			✓											
12	How to Measure Anything in Cybersecurity Risk	Hubbard & Seiersen, 2016		✓	✓			✓		✓			✓	✓		✓
13	Data Analytics and Decision Support for Cybersecurity	Carrascosa, Kalutarage, & Huang, 2017 *	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	
14	Introduction to Machine Learning with Applications in Information Security	Stamp, 2017	✓						✓	✓	✓	✓	✓	✓	✓	
15	Information Fusion for Cyber-Security Analytics	Alsmadi, Karabatis, & AlEroud, 2017 *	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	
16	Machine Learning & Security	Chio & Freeman, 2018	✓		✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	
17	Data Science for Cybersecurity	Heard, Adams, Rubin-Delanchy, & Turcotte,	✓	✓	✓		✓		✓	✓	✓	✓	✓		✓	
18	AI in Cybersecurity	Sikos, 2018 *	✓		✓	✓			✓			✓	✓		✓	
19	Malware Data Science: Attack Detection and Attribution	Saxe & Sanders, 2018	✓				✓		✓	✓	✓	✓	✓	✓	✓	
20	Machine Learning for Computer and Cyber Security	Gupta & Sheng, 2019 *	✓	✓	✓				✓	✓	✓	✓	✓		✓	
			90%	50%	80%	50%	65%	25%	90%	100%	90%	75%	80%	80%	80%	25%

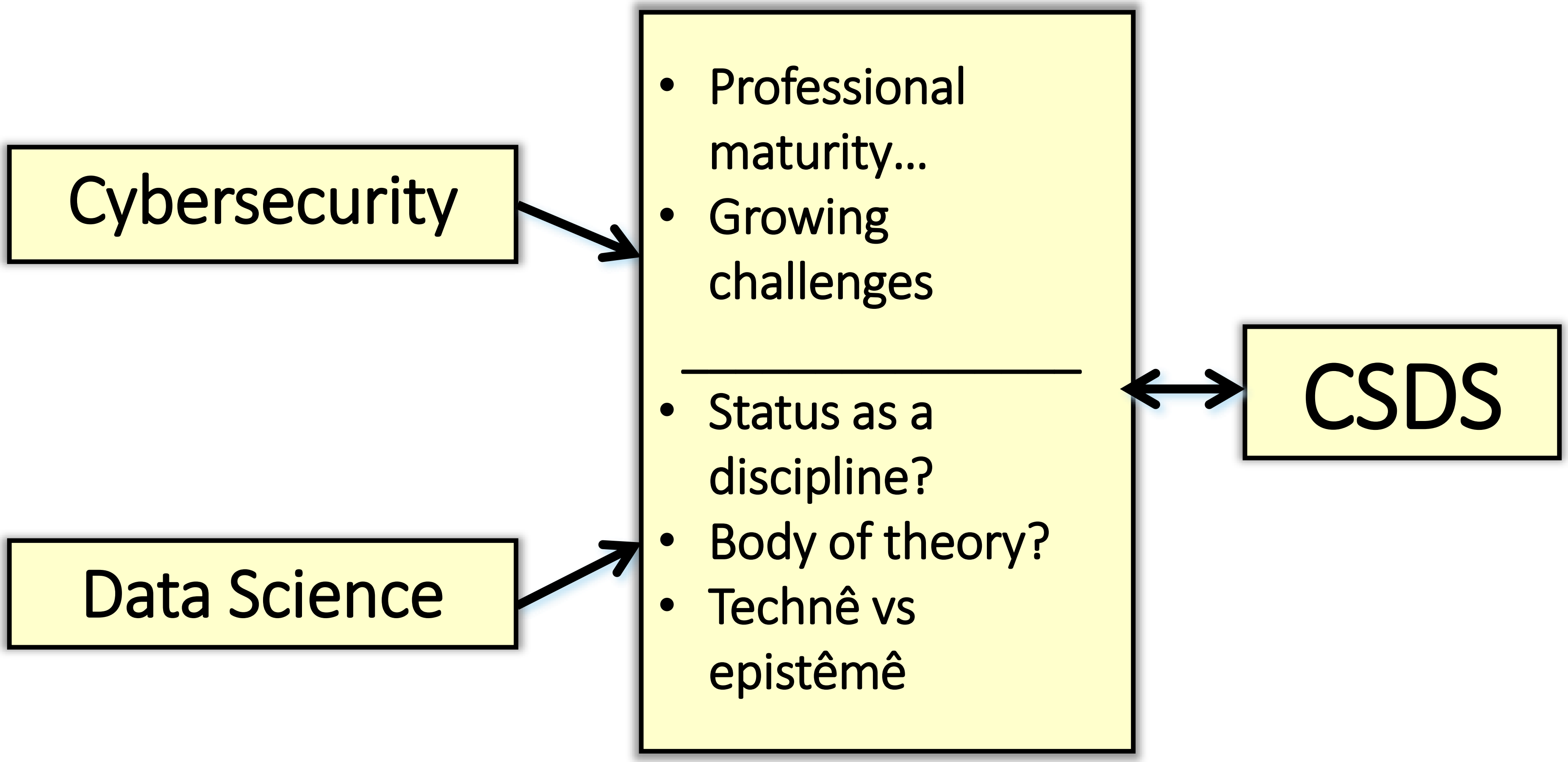
Table 2.11: CSDS topic coverage across central literature

Relatively less coverage <= 50%

- Risk quantification: 50%
- Data management: 50%
- Scientific methods: 25%
- Organizational management: 25%



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
• Professional maturity…
• Growing challenges
• Status as a discipline?
• Body of theory?
• Technê vs epistêmê
CSDS

‘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?

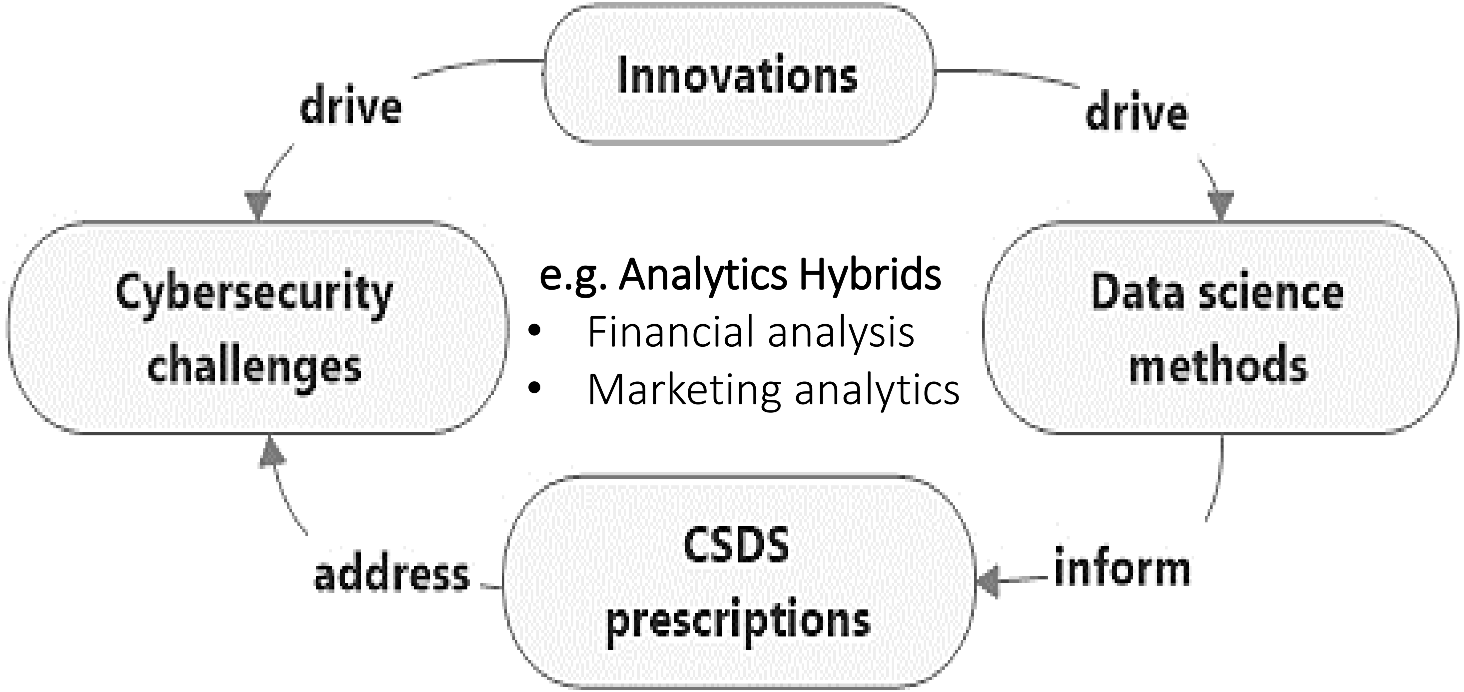
CSDS ≈ Medieval Medicine?

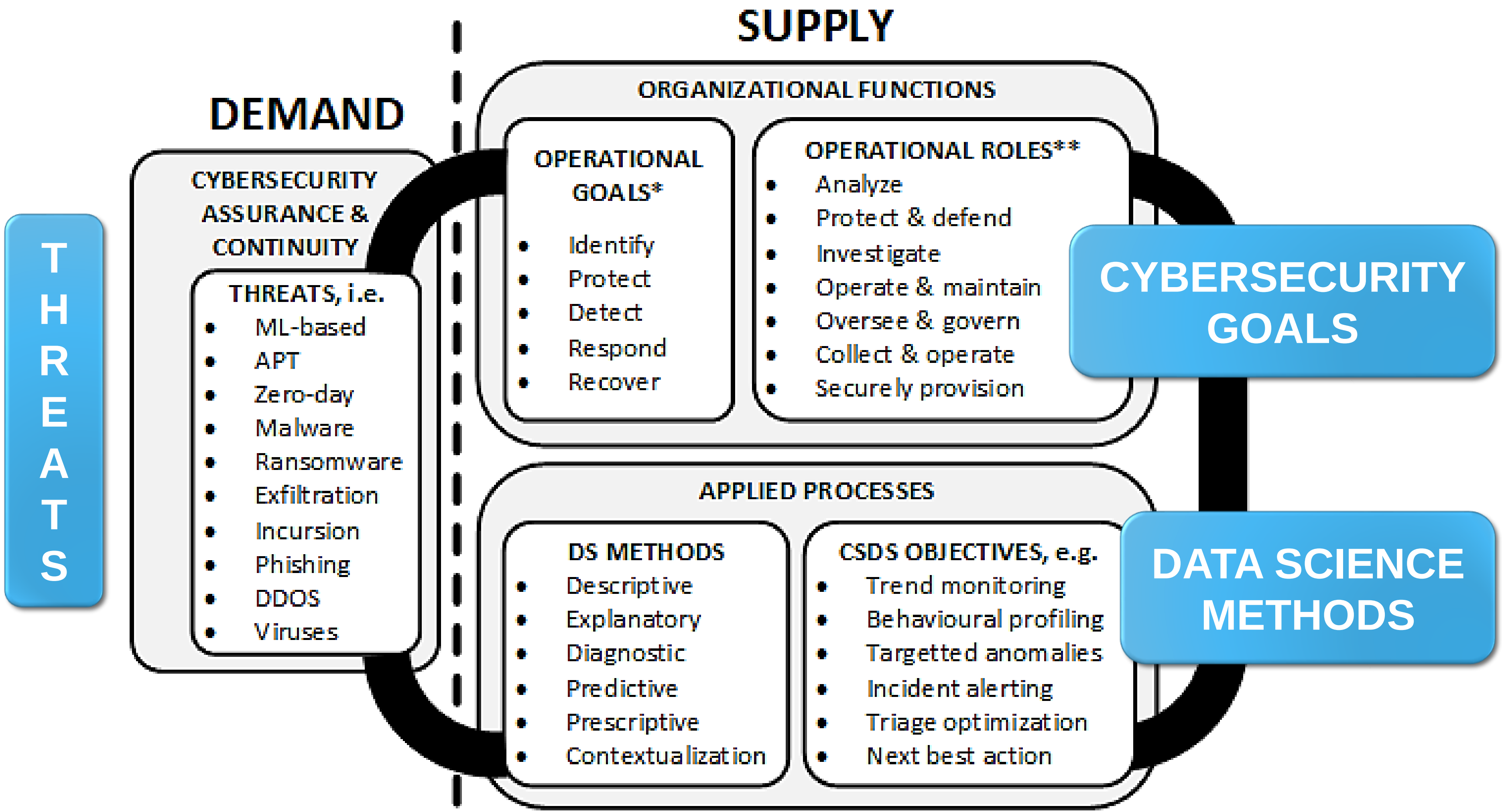
Medieval Medicine	CSDS
Understandings of basic anatomy	Good knowledge of networking, devices & architectures
Surgical treatments are extremely painful and dangerous	Interventions frequently involve leaches, saws, knives, and hammers
Poor understanding of functional biotic processes and interaction of organs	Security field lacking in strong scientific foundations & general theory
Just about anyone can be a physician	Just about anyone can be a (cybersecurity) data scientist

https://isseicreekphilosophy.wordpress.com/2012/04/28/sympathetic-magic-the-weapon-salve-and-the-powder-of-sympathy-in-the-17th-century-europe/

CYBERSECURITY DATA SCIENCE PROFESSION

CSDS =
Cybersecurity Gaps ⇔ DS Methods




SUPPLY
DEMAND
ORGANIZATIONAL FUNCTIONS
OPERATIONAL GOALS*
Identify
Protect
Detect
Respond
Recover
OPERATIONAL ROLES**
Analyze
Protect & defend
Investigate
Operate & maintain
Oversee & govern
Collect & operate
Securely provision
CYBERSECURITY GOALS
CYBERSECURITY ASSURANCE & CONTINUITY
THREATS, i.e.
ML-based
APT
Zero-day
Malware
Ransomware
Exfiltration
Incursion
Phishing
DDOS
Viruses
THREATS
APPLIED PROCESSES
DS METHODS
Descriptive
Explanatory
Diagnostic
Predictive
Prescriptive
Contextualization
CSDS OBJECTIVES, e.g.
Trend monitoring
Behavioural profiling
Targetted anomalies
Incident alerting
Triage optimization
Next best action
DATA SCIENCE METHODS

III. CSDS Interviews

Participants - Sample

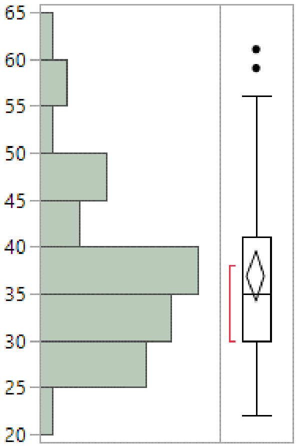
50 participants + 150 years collective CSDS experience (3 yr mean)

- Linked-In search
  - ‘cybersecurity’ + (‘data scientist’ or ‘analytics’)
- ~350 professionals globally
  - Direct outreach
  - Follow-on referrals
- Gating to exclude ‘ceremonial CSDS’
  - i.e. sales, recruiting, marketing, technology strategists
- Aspects of methodological integrity addressed in write-up
  - i.e. selection bias, representativeness of sample, etc.

Demographic Profile (n=50)

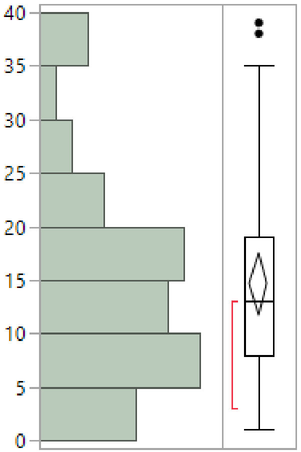
LinkedIn => 350 candidates => 50 participants

Age*



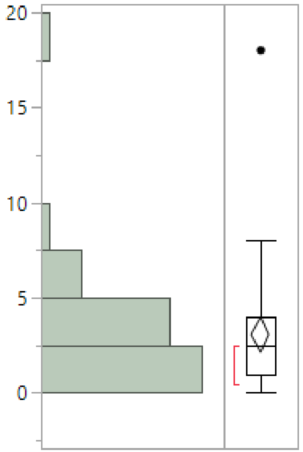
Mean	36.8
StdDev	9.1

Yrs Employed*



Mean	14.2
StdDev	9.5

Yrs CSDS*

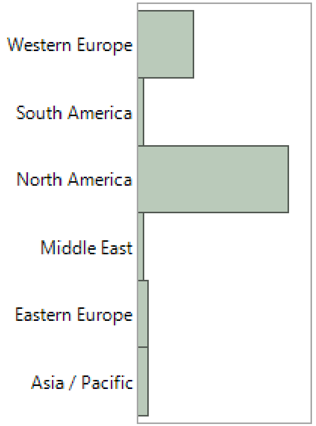


Mean	2.9
StdDev	1.9

** Estimates inferred from LinkedIn profile data*

Demographic Profile (n=50)

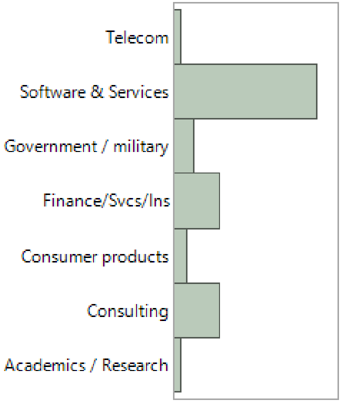
Current Region



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

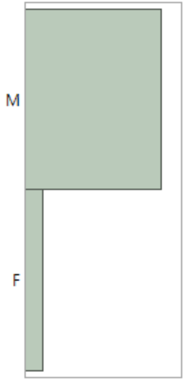
22% (n=11) relocated from native region
18% (n=9) relocated to US specifically
10% (n=5) relocated specifically from Asia/Pacific to US

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 Practitioner Interview Research

Qualitative: 30 minute open response interviews

- ENTRY: How did you become involved in domain?
- What TRENDS are emerging?
- What are perceived central CHALLENGES?
- What are key BEST PRACTICES?
- METHODS: Borrowing from adjacent domains?
- THREATS: Trends on the adversarial side?

Methodology: Interview Topic Labeling (CODING)

Inductive Extrapolation and Deductive Refinement

+scientist,science,+activity,+data scientist,cyber
+instance,+positive,false,+false positive,+obtain
+behavior,+anomaly,detection,+attack,false
right,+risk,+day,+case,+aspect
machine,machine learning,learning,+industry,ml
quality,+process,+process,collection,data quality
cyber security,+tool,+little,+hard,malicious
+tool,+integrate,job,+user,knowledge

Topic extraction

Agglomerative => multi-doc

- Text analytics processing
  - Engine: SAS Contextual Analysis
  - Natural Language Processing (NLP)
  - Latent Semantic Indexing (LSI)
  - Singular Value Decomposition (SVD)

training +industry 'machine learning' +apply pretty 'data science' +market analysis ml +area machine +algorithm +domain +defense 'as well'
+behavior false +anomaly +positive 'as well' +event +'false positive' detection +point well important +solution +automate learning +label
+instance +'false positive' +allow +depend +extract +obtain +amount +'different thing' +add +deal +positive +collect +mention false information
+integrate 'cyber security' +trend +approach cyber better +business +field +depend +large +know +good +machine +hard +scientist
cybersecurity definitely +address +increase +automate +complexity +defense +industry +mention +threat +attacker +issue right +device +tool
'big data' privacy +implement +process +decision +technique +big quality +algorithm +bring +solve difficult +method +year +apply
+buy +day money +long +aspect +source +network especially +case right +area +start +bring cybersecurity +big

Concept clustering

Divisive => unique doc

Content analytics extrapolated themes

Domain literature: sensitizing concepts

Practitioner review

Key topics (codes)

'Coding' of processed interview transcripts

Challenges: 11 Codes

ORGANIZATION

Ownership?

Marketing hype

Regulatory uncertainty

Few resources

PROCESS

False alerts volume

Decision uncertainty

Scientific process?

DATA & TECHNOLOGY

Data preparation / quality

Own infrastructure & shadow IT?

Normal vs. anomalous?

Lack of labeled incidents

Challenge codes
CH1: Data preparation (access, volume, integration, quality, transformation, selection)
CH2: Unrealistic expectations proliferated by marketing hype
CH3: Contextual nature of normal versus anomalous behavioral phenomenon
CH4: Lack of labeled incidents to focus detection
CH5: Own infrastructure, shadow IT, and proliferation of exposure
CH 6: Uncertainty leads to ineffective reactive stance
CH 7: Traditional rules-based methods result in too many alerts
CH 8: Program ownership, decision making, and processes
CH 9: Resourcing, developing, & hosting in house
CH 10: Expanding breadth and complexity of cyber domain
CH 11: Policy, privacy, regulatory, and fines

CSDS 'CHALLENGES': 11

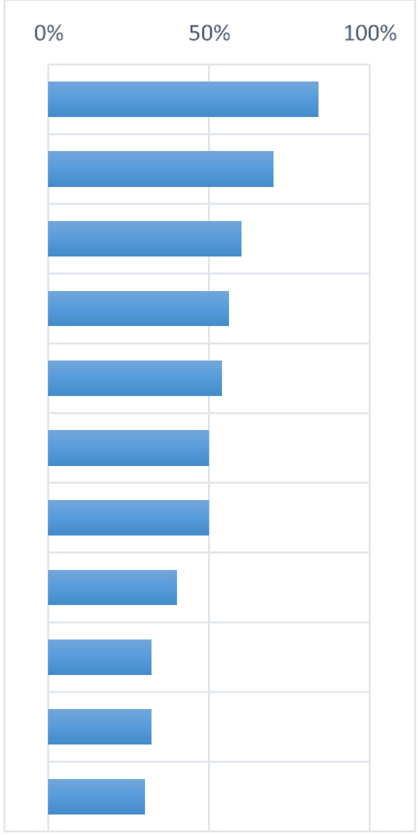
DATA PREPARATION! 84%

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%



Best Practices: 26 Codes

OR

- Manageme
- Training &

Best practice codes*			
BP1: Structured data preparation, discovery, engineering process	Proc	**BP14:** Cloud and container-based tools and data storage	Tech
BP2: Building process focused cross-functional team	Org	**BP15:** Distinct exploration and detection architectures	Tech
BP3: Cross-training team in data science, cyber, engineering	Org	**BP16:** Participate in data sharing consortiums and initiatives	Tech
BP4: Scientific method as a process	Proc	**BP17:** Deriving probabilistic and risk models	Org
BP5: Instill core cyber domain knowledge	Org	**BP18:** Upper management buy in and support	Org
BP6: Vulnerability, anomaly & decision automation to operational capacity	Tech	**BP19:** Human-in-the-loop reinforcement	Proc
BP7: Data normalization, frameworks & ontologies	Tech	**BP20:** Survey academic methods and techniques	Org
BP8: Model validation and transparency	Proc	**BP21:** Cyber risk as general enterprise risk & reward	Org
BP9: Data-driven paradigm shift away from rules & signatures	Org	**BP22:** Segment risk programmatically and outsource components	Org
BP10: Track and label incidents and exploits	Proc	**BP23:** Adding machine learning to SIEM	Tech
BP11: Cyclical unsupervised and supervised machine learning	Proc	**BP24:** Preventative threat intelligence	Org
BP12: Address AI hype and unrealistic expectations directly	Org	**BP25:** Hosting and pushing detection to endpoints	Tech
BP13: Understand own infrastructure & environment	Org	**BP26:** Honeypots to track and observe adversaries	Tech

Y

actices~

lization

- Architecture-driven solutions

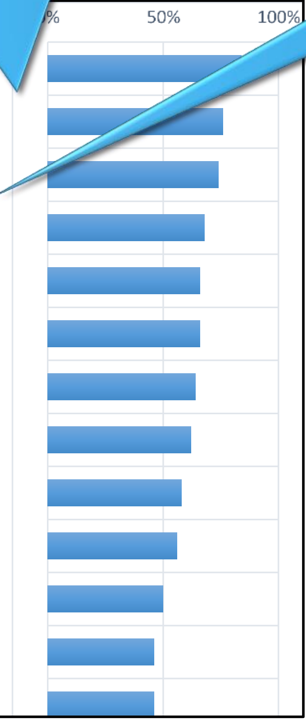
DATA PREPARATION! 84%

Cross-domain collaboration 76%

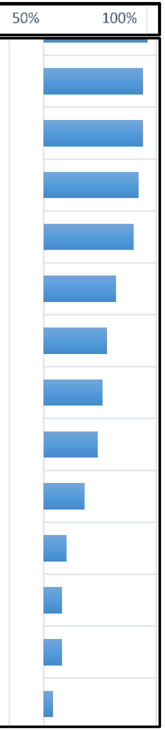
Scientific rigor 68%

CSDS 'BEST PRACTICES': 26

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%



Factor Analysis: 6 Challenge and 6 Best Practice Themes

Exploratory factor analysis (extraction of latent factors across responses)

CH F1
Expansive complexity

CH F2
Tracking & context

CH F3
Data management

CH F4
Expectations versus limitations

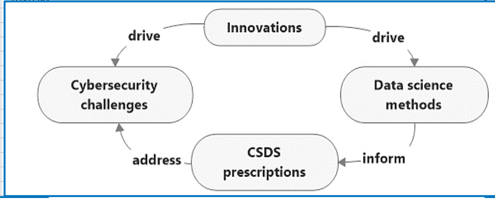
CH F5
Unclear ownership

CH F6
Data policies

Challenge Factor Rotated Factor Scores (per respondent)

#	FACTOR1	FACTOR2	FACTOR3	FACTOR4	FACTOR5	FACTOR6
1	-1.10951	-1.28479	0.35935	0.706186	-0.61424	-0.65698
2	-0.65954	0.826596	0.346772	0.658827	1.416078	0.069793
3	-1.14351	0.858178	-2.30938	0.582112	-0.54697	0.54427
4	0.27474	0.984315	0.385166	-1.42545	1.240915	1.300297
5	0.185896	1.062432	0.562463	0.78656	1.188467	1.298155
6	-0.98246	-1.32727	0.273863	0.616202	1.437153	-0.4533
7	-1.19556	-1.36513	0.230082	0.598131	1.492765	-0.64768
8	-1.08428	0.629378	0.129993	-1.55677	-0.47443	-0.98028
9	[illegible]	-1.19096	0.30925	-1.4047	-0.66512	0.249976
10	-0.19805	[illegible]	0.411884	-1.34841	-0.72229	0.914774
11	0.771806	-1.22723	[illegible]	-1.47342	-1.0147	0.004531
12	-0.93501	0.76347	0.213409	-1.47372	-0.52214	-0.26985
13	1.374426	-1.3837	0.525116	0.505144	0.860759	-0.83992
14	0.740622	0.650381	0.426068	0.499477	[illegible]	-0.92164
15	-0.95034	0.965299	0.460065	0.816931	-0.60279	[illegible]
16	0.889892	0.784473	0.509483	0.582528	1.037561	-0.21121
17	-0.03689	0.80463	0.409688	0.624868	1.264377	-0.07003
18	[illegible]	-0.81167	0.787572	0.963632	-1.06151	1.808688
19	0.971184	[illegible]	-2.10882	0.490955	0.94173	1.072211
20	-1.17033	0.549047	0.000725	-1.66482	1.632577	-0.97099
21	1.328284	-1.2434	[illegible]	0.427606	0.860188	0.634095
22	0.092641	0.917448	0.480335	[illegible]	1.223261	[illegible]
23	-0.13444	-1.00191	0.463944	-1.29747	[illegible]	1.240221
24	0.402174	-1.10421	[illegible]	[illegible]	1.145655	[illegible]
25	-0.37696	[illegible]	0.298401	-1.47246	-0.72714	-0.24082
26	[illegible]	-1.28479	0.35935	0.706186	-0.61424	-0.65698
27	0.827517	0.759206	0.419901	-1.53194	-0.95627	-0.35647
28	1.460472	-1.30337	0.654385	0.6132	-1.24625	-0.84922
29	-1.16343	0.927441	0.416284	0.798861	-0.54718	[illegible]
30	-0.16308	0.875596	0.35974	-1.42462	[illegible]	0.300757
31	0.558327	0.780959	[illegible]	[illegible]	1.187568	0.204428
32	0.024778	[illegible]	0.632878	0.856401	-0.91698	0.818443
33	[illegible]	0.827283	-2.26368	0.545722	-0.6712	0.360713
34	-0.15817	0.490192	-2.37321	0.31032	-0.7599	-1.45588
35	1.399657	0.530476	0.29576	-1.75781	1.000571	-1.16323
36	0.175996	-1.05357	0.545164	0.758866	1.129298	0.965849
37	0.624724	-1.31926	0.547687	0.629087	-1.03894	[illegible]
38	-0.64063	-1.1469	-2.36111	-1.54722	-0.67359	[illegible]
39	0.978066	0.587325	-2.27081	0.279948	1.031358	[illegible]
40	-0.88673	-1.02699	0.512125	0.867878	-0.69015	[illegible]
41	-0.7452	-1.29979	0.315417	0.626722	1.376429	[illegible]
42	1.333037	0.785159	0.641959	0.627234	-1.15099	[illegible]
43	1.246992	0.704828	0.51269	0.519178	0.956019	[illegible]
44	-1.02385	0.841588	0.390705	0.738291	-0.57459	[illegible]
45	1.333037	0.785159	0.641959	0.627234	-1.15099	[illegible]
46	-0.95034	0.965299	0.460065	0.816931	-0.60279	[illegible]
47	-1.02385	0.841588	0.390705	0.738291	-0.57459	[illegible]
48	1.277203	0.705515	-2.09776	0.406074	-1.13126	[illegible]
49	1.333037	0.785159	0.641959	0.627234	-1.15099	[illegible]
50	-1.02385	0.841588	0.390705	0.738291	-0.57459	[illegible]

- Least squares
- Correlation



Best Practice Rotated Factor Scores (per respondent)

#	FACTOR1	FACTOR2	FACTOR3	FACTOR4	FACTOR5	FACTOR6
1	-1.10951	-1.28479	0.35935	0.706186	-0.61424	-0.65698
2	-0.65954	0.826596	0.346772	0.658827	1.416078	0.069793
3	-1.14351	0.858178	-2.30938	0.582112	-0.54697	0.54427
4	0.27474	0.984315	0.385166	-1.42545	1.240915	1.300297
5	0.185896	1.062432	0.562463	0.78656	1.188467	1.298155
6	-0.98246	-1.32727	0.273863	0.616202	1.437153	-0.4533
7	-1.19556	-1.36513	0.230082	0.598131	1.492765	-0.64768
8	-1.08428	0.629378	0.129993	-1.55677	-0.47443	-0.98028
9	-0.65831	-1.19096	0.30925	-1.4047	-0.66512	0.249976
10	-0.19805	0.990378	0.411884	-1.34841	-0.72229	0.914774
11	0.771806	-1.22723	0.460708	-1.47342	-1.0147	[illegible]
12	-0.93501	0.76347	0.213409	[illegible]	-0.52214	-0.26985
13	1.374426	-1.3837	[illegible]	0.505144	0.860759	-0.83992
14	[illegible]	0.650381	0.426068	0.499477	1.085276	-0.92164
15	-0.95034	0.965299	0.460065	0.816931	-0.60279	0.385758
16	0.889892	0.784473	0.509483	0.582528	1.037561	[illegible]
17	-0.03689	0.80463	0.409688	0.624868	[illegible]	-0.07003
18	0.548646	-0.81167	[illegible]	[illegible]	-1.06151	1.808688
19	0.971184	[illegible]	-2.10882	0.490955	0.94173	1.072211
20	[illegible]	0.549047	0.000725	-1.66482	1.632577	-0.97099
21	1.328284	-1.2434	-2.10561	0.427606	0.860188	0.634095
22	0.092641	0.917448	0.480335	0.694619	1.223261	0.527547
23	-0.13444	-1.00191	0.463944	-1.29747	-0.80965	1.240221
24	0.402174	-1.10421	0.397592	-1.43948	1.145655	1.109696
25	-0.37696	-1.28099	0.298401	-1.47246	-0.72714	-0.24082
26	-1.10951	-1.28479	0.35935	0.706186	-0.61424	-0.65698
27	0.827517	0.759206	0.419901	-1.53194	-0.95627	-0.35647
28	[illegible]	-1.30337	0.654385	0.6132	-1.24625	-0.84922
29	-1.16343	[illegible]	0.416284	0.798861	-0.54718	0.191376
30	-0.16308	0.875596	[illegible]	-1.42462	-0.72156	0.300757
31	0.558327	0.780959	0.319014	[illegible]	1.187568	0.204428
32	0.024778	-1.00072	0.632878	0.856401	[illegible]	0.818443
33	-0.62933	0.827283	-2.26368	0.545722	-0.6712	0.360713
34	-0.15817	0.490192	-2.37321	0.31032	-0.7599	[illegible]
35	1.399657	0.530476	0.29576	-1.75781	1.000571	-1.16323
36	[illegible]	-1.05357	0.545164	0.758866	1.129298	0.965849
[illegible]	[illegible]	[illegible]	0.547687	0.629087	-1.03894	-0.90378
[illegible]	[illegible]	-1.1469	[illegible]	-1.54722	-0.67359	1.207946
[illegible]	[illegible]	0.587325	-2.27081	[illegible]	1.031358	-0.54349
[illegible]	[illegible]	-1.02699	0.512125	0.867878	[illegible]	0.711205
[illegible]	-0.7452	-1.29979	0.315417	0.626722	1.376429	[illegible]
[illegible]	[illegible]	0.785159	0.641959	0.627234	-1.15099	-0.65862
[illegible]	[illegible]	0.704828	0.51269	0.519178	0.956019	-0.64932
[illegible]	[illegible]	0.841588	0.390705	0.738291	-0.57459	-0.272
[illegible]	[illegible]	0.785159	0.641959	0.627234	-1.15099	-0.65862
[illegible]	[illegible]	0.965299	0.460065	0.816931	-0.60279	0.385758
[illegible]	[illegible]	0.841588	0.390705	0.738291	-0.57459	-0.272
[illegible]	[illegible]	0.705515	-2.09776	0.406074	-1.13126	-0.3584
[illegible]	[illegible]	0.785159	0.641959	0.627234	-1.15099	-0.65862
[illegible]	[illegible]	0.841588	0.390705	0.738291	-0.57459	-0.272

BP F1
Scientific process

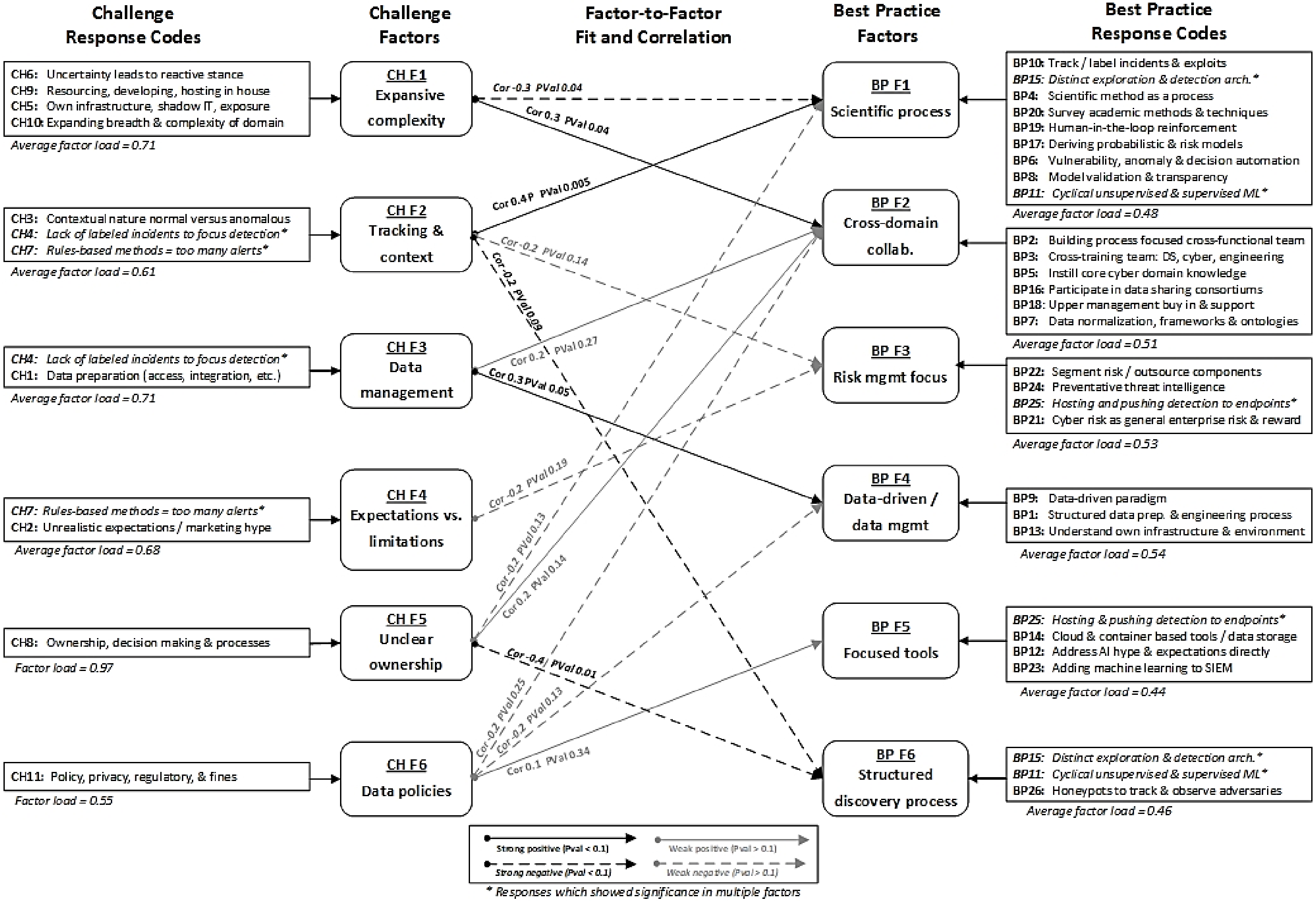
BP F2
Cross-domain collaboration

BP F3
Risk management focus

BP F4
Data-driven / data management

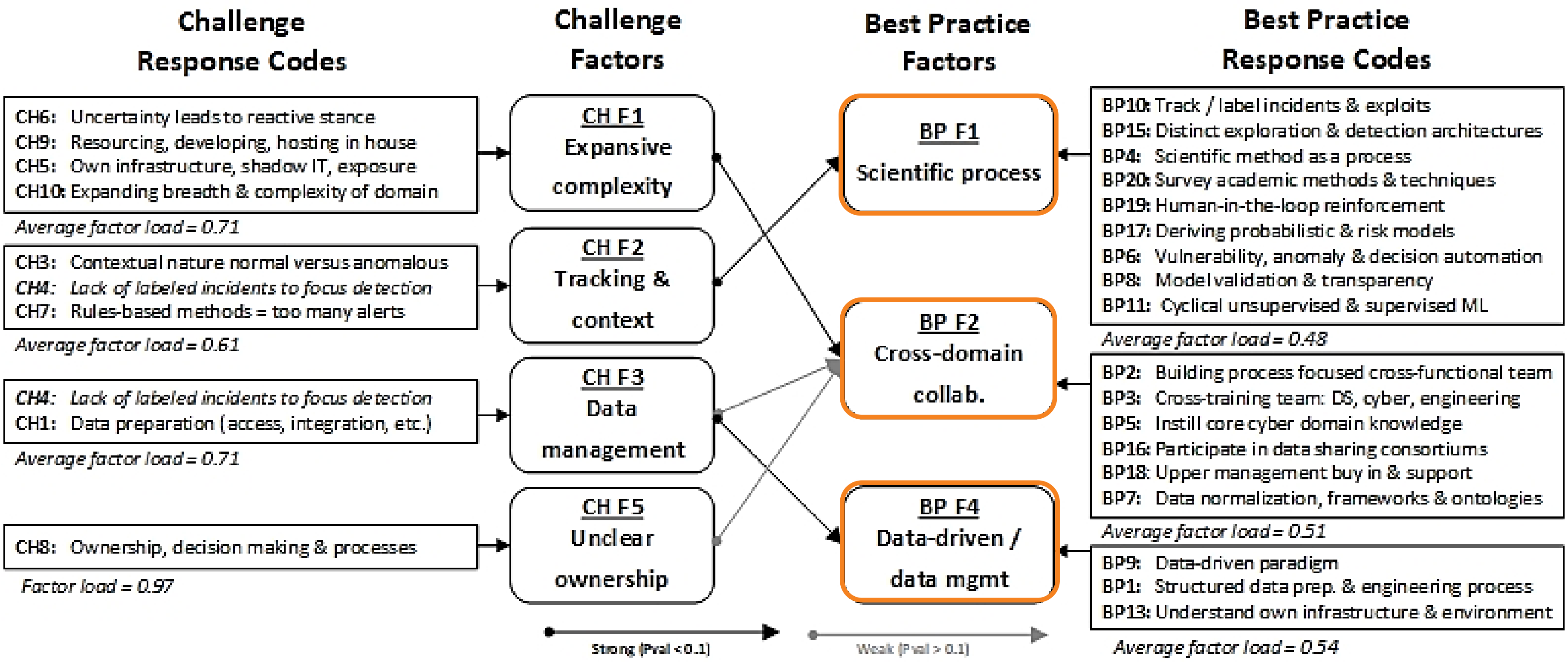
BP F5
Focused tools

BP F6
Structured discovery process



Challenge Response Codes
Challenge Factors
Factor-to-Factor Fit and Correlation
Best Practice Factors
Best Practice Response Codes
CH6: Uncertainty leads to reactive stance
CH9: Resourcing, developing, hosting in house
CH5: Own infrastructure, shadow IT, exposure
CH10: Expanding breadth & complexity of domain
Average factor load = 0.71
CH3: Contextual nature normal versus anomalous
CH4: Lack of labeled incidents to focus detection*
CH7: Rules-based methods = too many alerts*
Average factor load = 0.61
CH4: Lack of labeled incidents to focus detection*
CH1: Data preparation (access, integration, etc.)
Average factor load = 0.71
CH7: Rules-based methods = too many alerts*
CH2: Unrealistic expectations / marketing hype
Average factor load = 0.68
CH8: Ownership, decision making & processes
Factor load = 0.97
CH11: Policy, privacy, regulatory, & fines
Factor load = 0.55
CH F1 Expansive complexity
CH F2 Tracking & context
CH F3 Data management
CH F4 Expectations vs. limitations
CH F5 Unclear ownership
CH F6 Data policies
Cor -0.3 PVal 0.04
Cor 0.3 PVal 0.04
Cor 0.4 P PVal 0.005
Cor -0.2 PVal 0.14
Cor -0.2 PVal 0.09
Cor 0.2 PVal 0.27
Cor 0.3 PVal 0.05
Cor -0.2 PVal 0.19
Cor -0.2 PVal 0.13
Cor 0.2 PVal 0.14
Cor -0.4 PVal 0.01
Cor -0.2 PVal 0.25
Cor -0.2 PVal 0.13
Cor 0.1 PVal 0.34
BP F1 Scientific process
BP F2 Cross-domain collab.
BP F3 Risk mgmt focus
BP F4 Data-driven / data mgmt
BP F5 Focused tools
BP F6 Structured discovery process
BP10: Track / label incidents & exploits
BP15: Distinct exploration & detection arch.*
BP4: Scientific method as a process
BP20: Survey academic methods & techniques
BP19: Human-in-the-loop reinforcement
BP17: Deriving probabilistic & risk models
BP6: Vulnerability, anomaly & decision automation
BP8: Model validation & transparency
BP11: Cyclical unsupervised & supervised ML*
Average factor load = 0.48
BP2: Building process focused cross-functional team
BP3: Cross-training team: DS, cyber, engineering
BP5: Instill core cyber domain knowledge
BP16: Participate in data sharing consortiums
BP18: Upper management buy in & support
BP7: Data normalization, frameworks & ontologies
Average factor load = 0.51
BP22: Segment risk / outsource components
BP24: Preventative threat intelligence
BP25: Hosting and pushing detection to endpoints*
BP21: Cyber risk as general enterprise risk & reward
Average factor load = 0.53
BP9: Data-driven paradigm
BP1: Structured data prep. & engineering process
BP13: Understand own infrastructure & environment
Average factor load = 0.54
BP25: Hosting & pushing detection to endpoints*
BP14: Cloud & container based tools / data storage
BP12: Address AI hype & expectations directly
BP23: Adding machine learning to SIEM
Average factor load = 0.44
BP15: Distinct exploration & detection arch.*
BP11: Cyclical unsupervised & supervised ML*
BP26: Honeypots to track & observe adversaries
Average factor load = 0.46
Strong positive (Pval < 0.1)
Weak positive (Pval > 0.1)
Strong negative (Pval < 0.1)
Weak negative (Pval > 0.1)
* Responses which showed significance in multiple factors

Interpretation: Best Practice as Perceived 'Gap' (Required Objective)

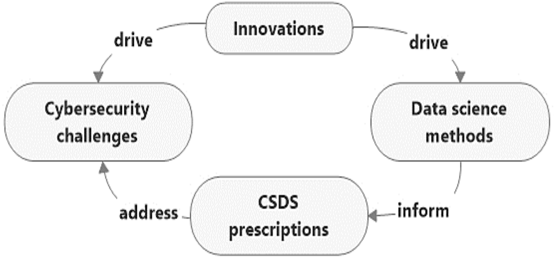


Challenge to Best Practice Factor Correlation

CODED RESPONSES: Perceived Challenges
CH1: Data preparation (access, volume, integration, quality, transformation, selection)
CH2: Unrealistic expectations proliferated by marketing hype
CH3: Contextual nature of normal anomalous behavioral phenomena
CH4: Lack of labeled incidents to fo
CH5: Own infrastructure, shadow proliferation of exposure
CH 6: Uncertainty leads to ineffect stance
CH 7: Traditional rules-based meth too many alerts
CH 8: Program ownership, decision processes
CH 9: Resourcing, developing, & hosting in house
CH 10: Expanding breadth and complexity of cyber domain
CH 11: Policy, privacy, regulatory, and fines

Challenge factors: diagnosed gaps	Best practice factors: prescribed treatments
CH F1: Expansive complexity	**BP F2:** Cross-domain collaboration
CH F2: Tracking and context	**BP F1:** Scientific process
CH F3: Data management	**BP F4:** Data-driven / data management **BP F2:** Cross-domain collaboration
CH F5: Unclear ownership	**BP F2:** Cross-domain collaboration

Best practice codes*			
BP1: Structured data preparation, discovery, engineering process	Proc	**BP14:** Cloud and container-based tools and data storage	Tech
BP2: Building process focused cross-functional team	Org	**BP15:** Distinct exploration and detection architectures	Tech
	Org	**BP16:** Participate in data sharing consortiums and initiatives	Tech
	Proc	**BP17:** Deriving probabilistic and risk models	Org
	Org	**BP18:** Upper management buy in and support	Org
	Tech	**BP19:** Human-in-the-loop reinforcement	Proc
	Tech	**BP20:** Survey academic methods and techniques	Org
	Proc	**BP21:** Cyber risk as general enterprise risk & reward	Org
	Org	**BP22:** Segment risk programmatically and outsource components	Org
BP10: Track and label incidents and exploits	Proc	**BP23:** Adding machine learning to SIEM	Tech
BP11: Cyclical unsupervised and supervised machine learning	Proc	**BP24:** Preventative threat intelligence	Org
BP12: Address AI hype and unrealistic expectations directly	Org	**BP25:** Hosting and pushing detection to endpoints	Tech
BP13: Understand own infrastructure & environment	Org	**BP26:** Honeypots to track and observe adversaries	Tech



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

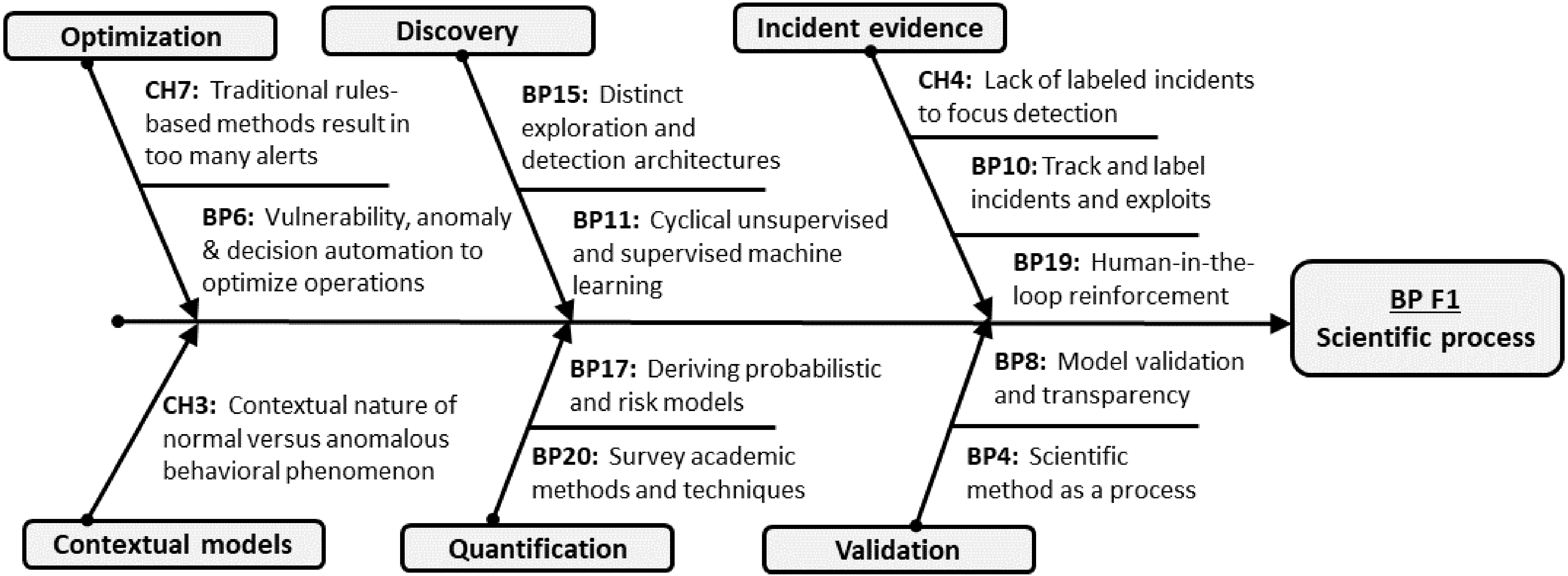
BP F3
Risk management focus

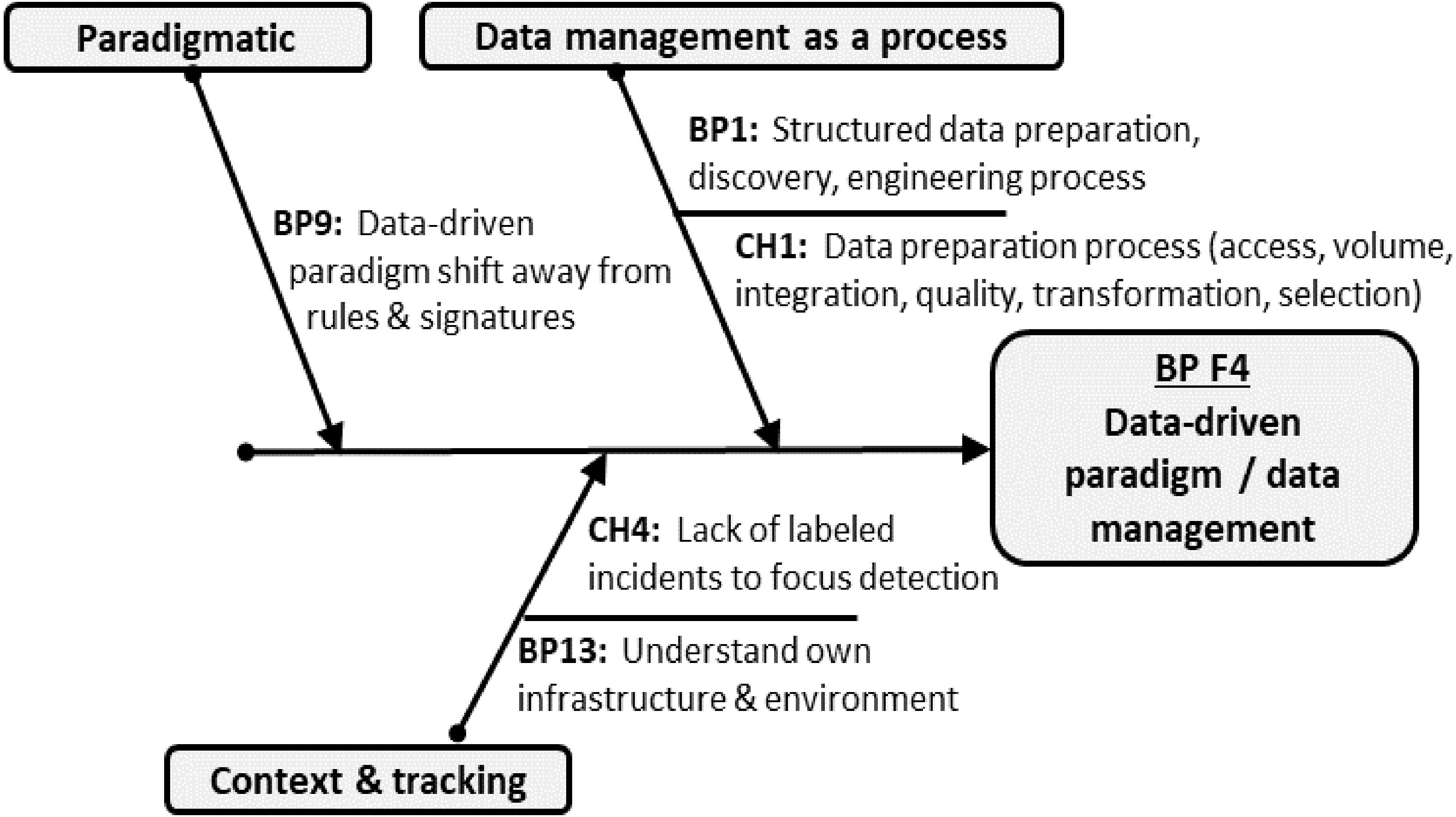
BP F4
Data-driven / data management

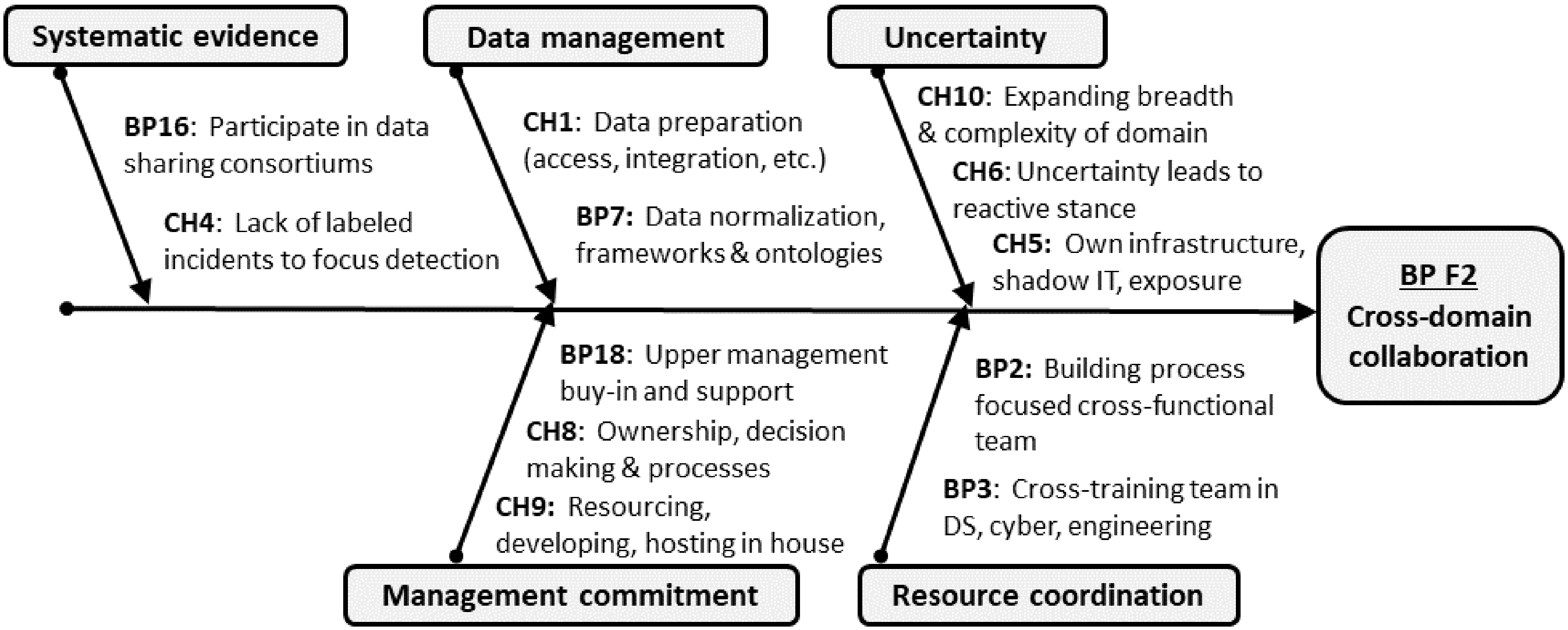
BP F5
Focused tools

BP F6
Structured discovery process

Root Cause Analysis: Fishbone / Ishikawa Diagram



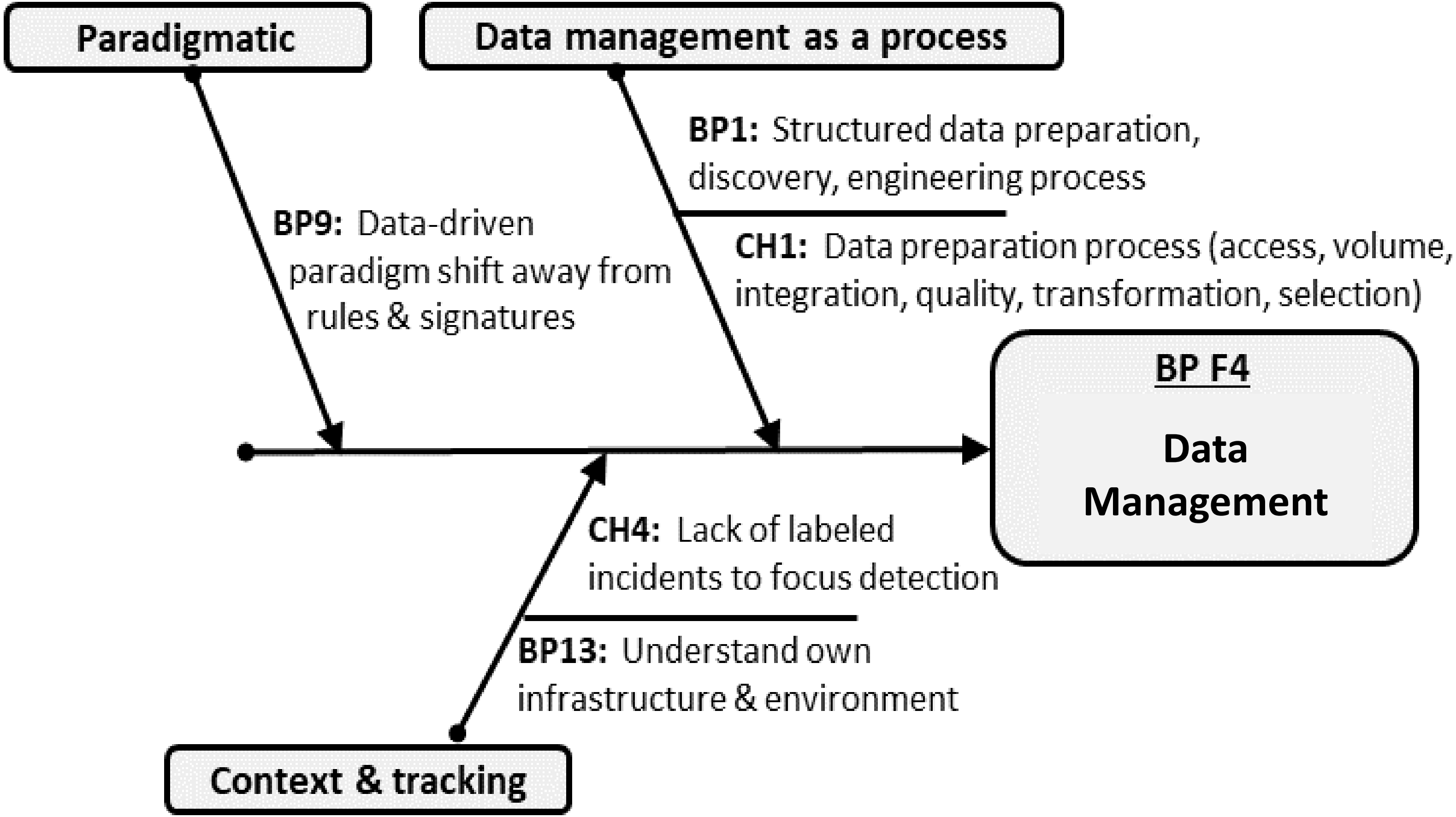

Paradigmatic
Data management as a process
BP1: Structured data preparation, discovery, engineering process
CH1: Data preparation process (access, volume, integration, quality, transformation, selection)
BP9: Data-driven paradigm shift away from rules & signatures
BP F4
Data-driven paradigm / data management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



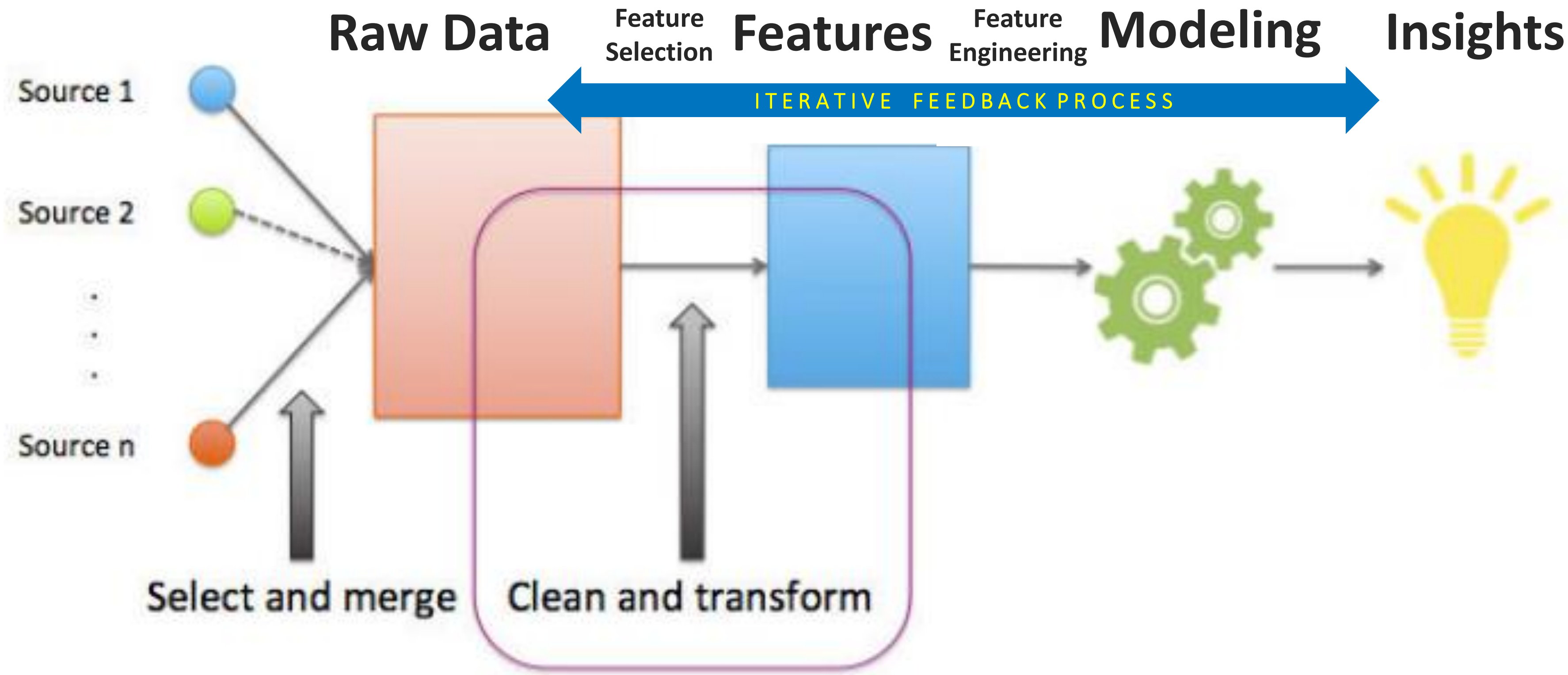
Systematic evidence
Data management
Uncertainty
BP16: Participate in data sharing consortiums
CH4: Lack of labeled incidents to focus detection
CH1: Data preparation (access, integration, etc.)
BP7: Data normalization, frameworks & ontologies
CH10: Expanding breadth & complexity of domain
CH6: Uncertainty 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

IV. CSDS Designs


ORGANIZATION
TECHNOLOGY
PROCESS
Data Management

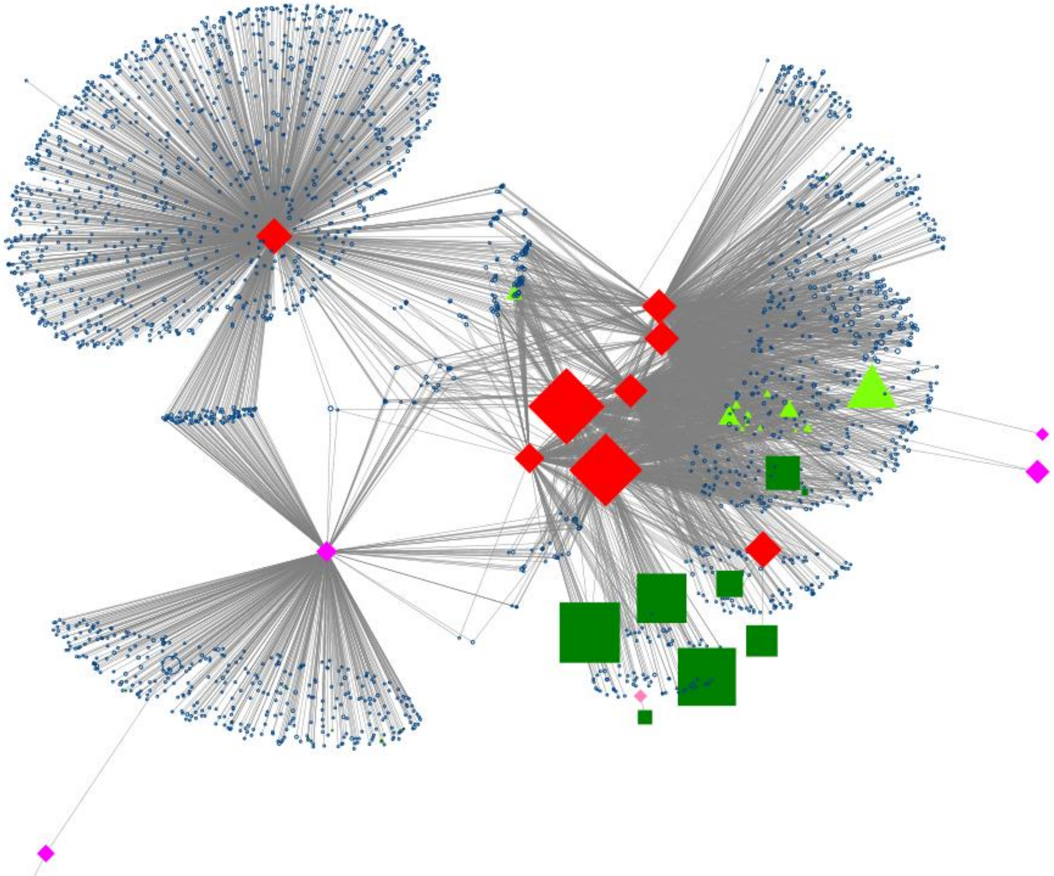

Paradigmatic
Data management as a process
BP1: Structured data preparation, discovery, engineering process
CH1: Data preparation process (access, volume, integration, quality, transformation, selection)
BP9: Data-driven paradigm shift away from rules & signatures
BP F4
Data Management
CH4: Lack of labeled 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

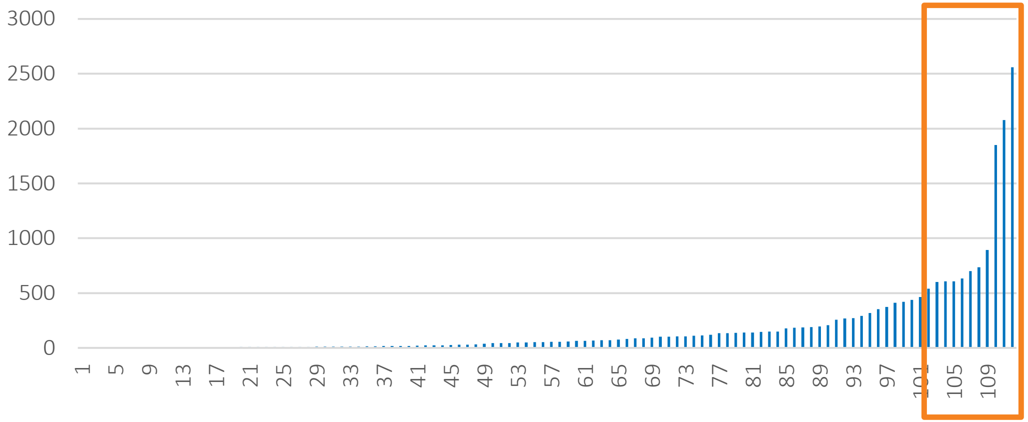


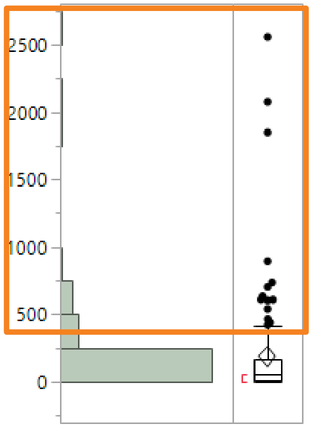
Host	Server
System User	System Interface
Human Users	

Exploratory Data Analysis (EDA): Example – Probabilistic Analysis

Exception Events

Exception messages per user (ranked)





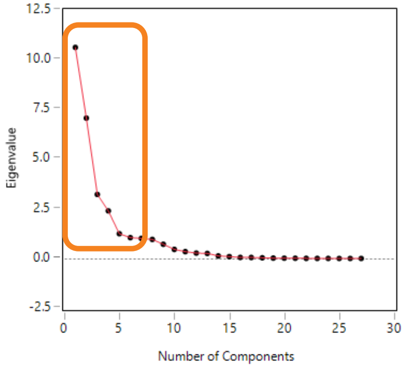
Quantiles

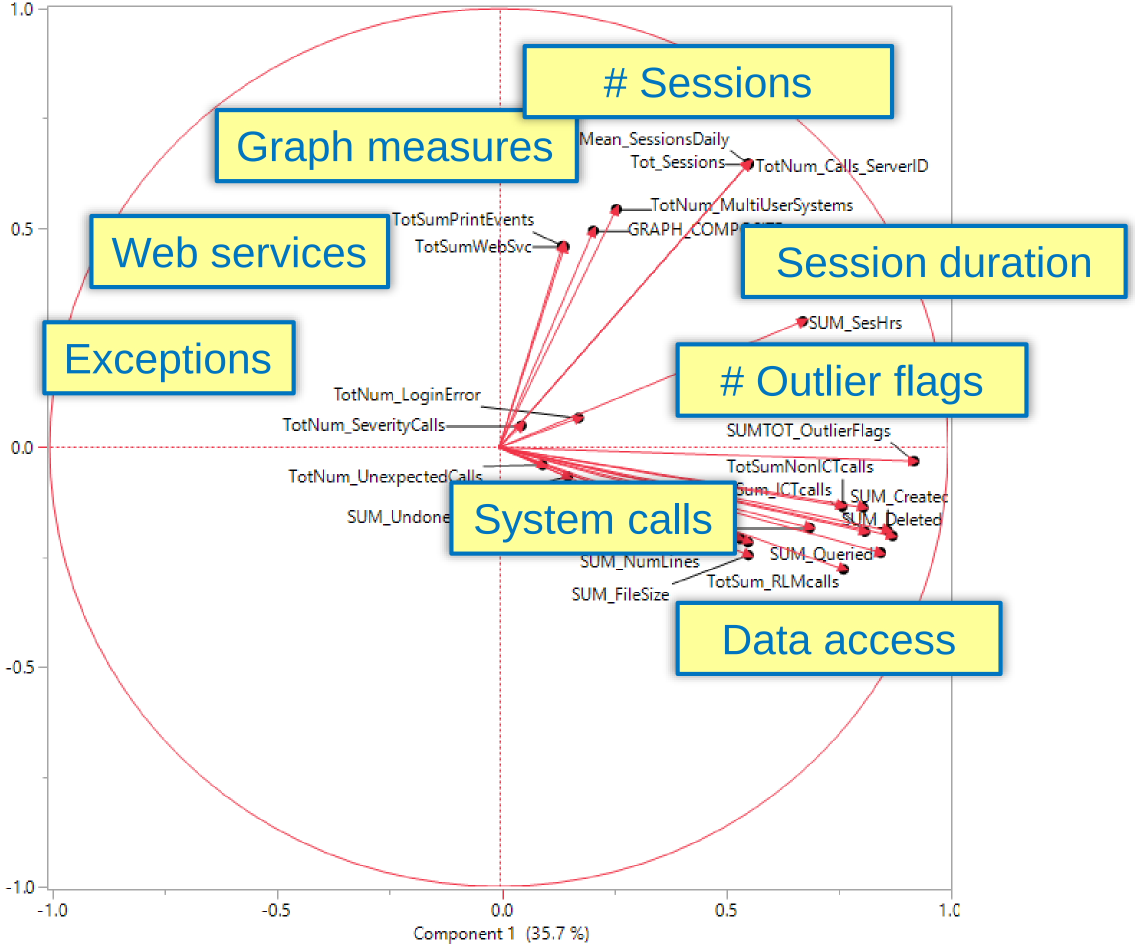
100.0%	maximum	2559
99.5%		2559
97.5%		1889.725
90.0%		517.5
75.0%	quartile	172.75
50.0%	median	55.5
25.0%	quartile	9.75
10.0%		3.3
2.5%		1.825
0.5%		1
0.0%	minimum	1

Summary Statistics

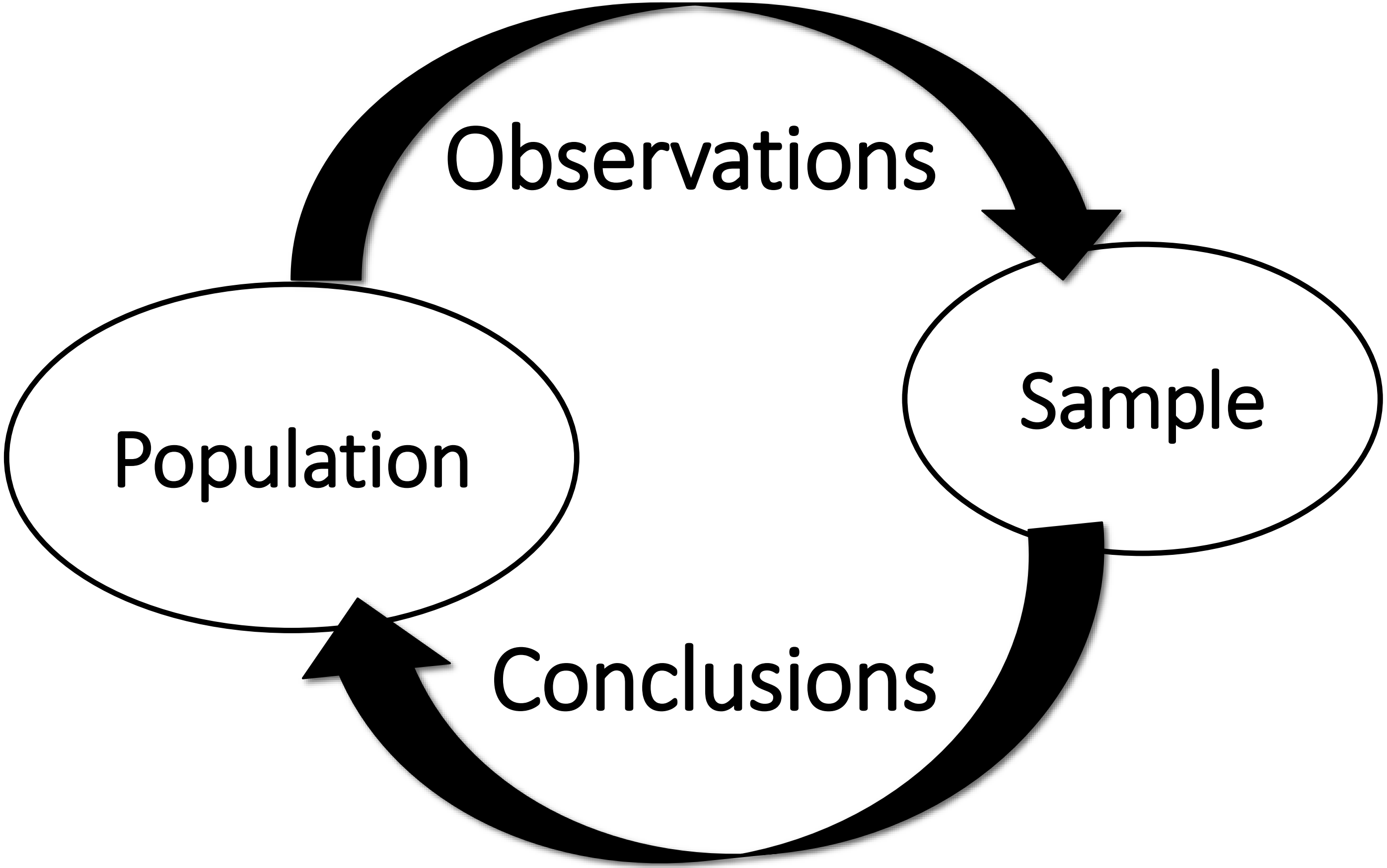
Mean	184.01786
Std Dev	380.96684
Std Err Mean	35.997982
Upper 95% Mean	255.35026
Lower 95% Mean	112.68545
N	112

Feature Reduction: Example - Principal Component Analysis (PCA)

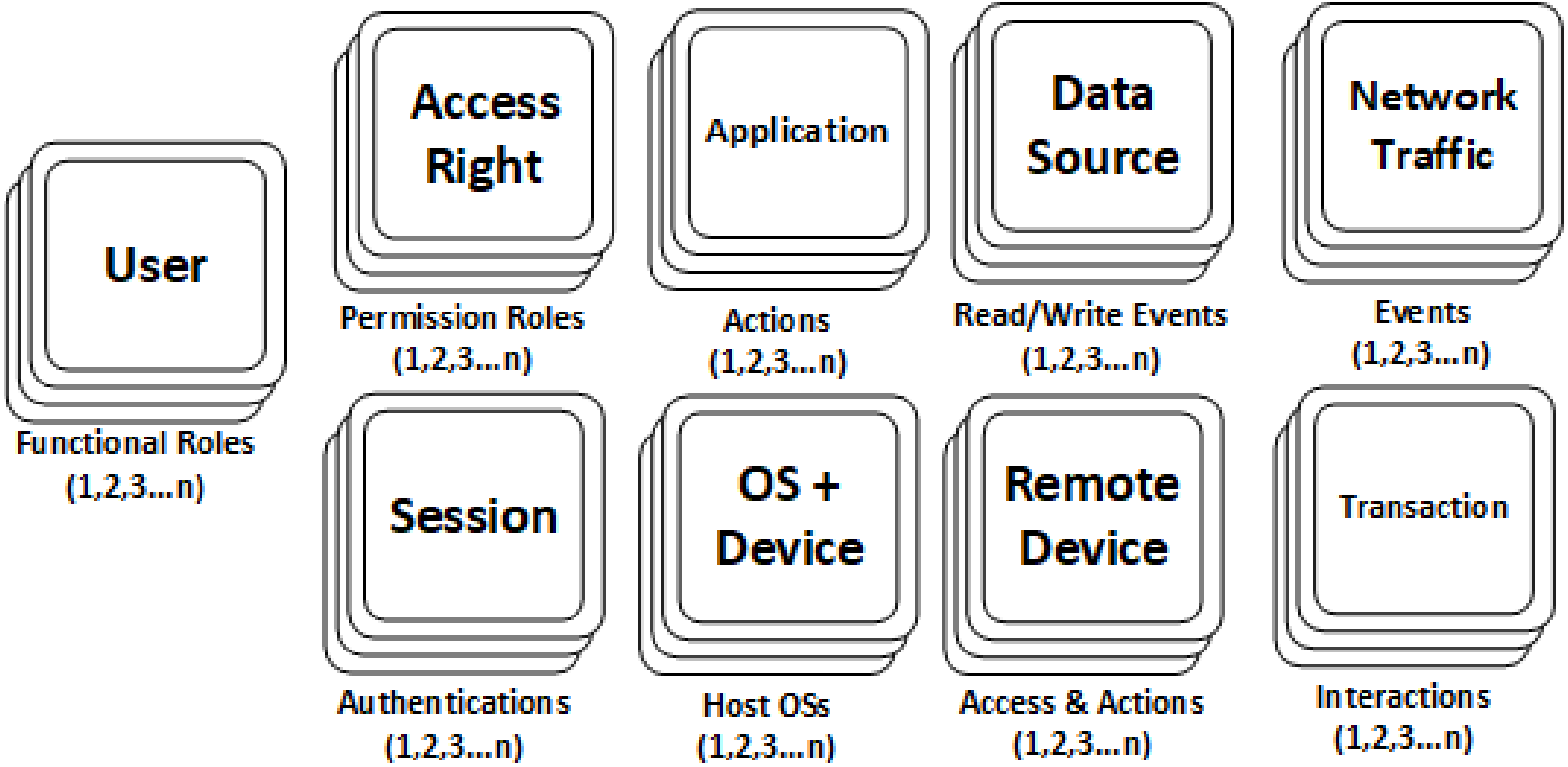




Inferential Statistics

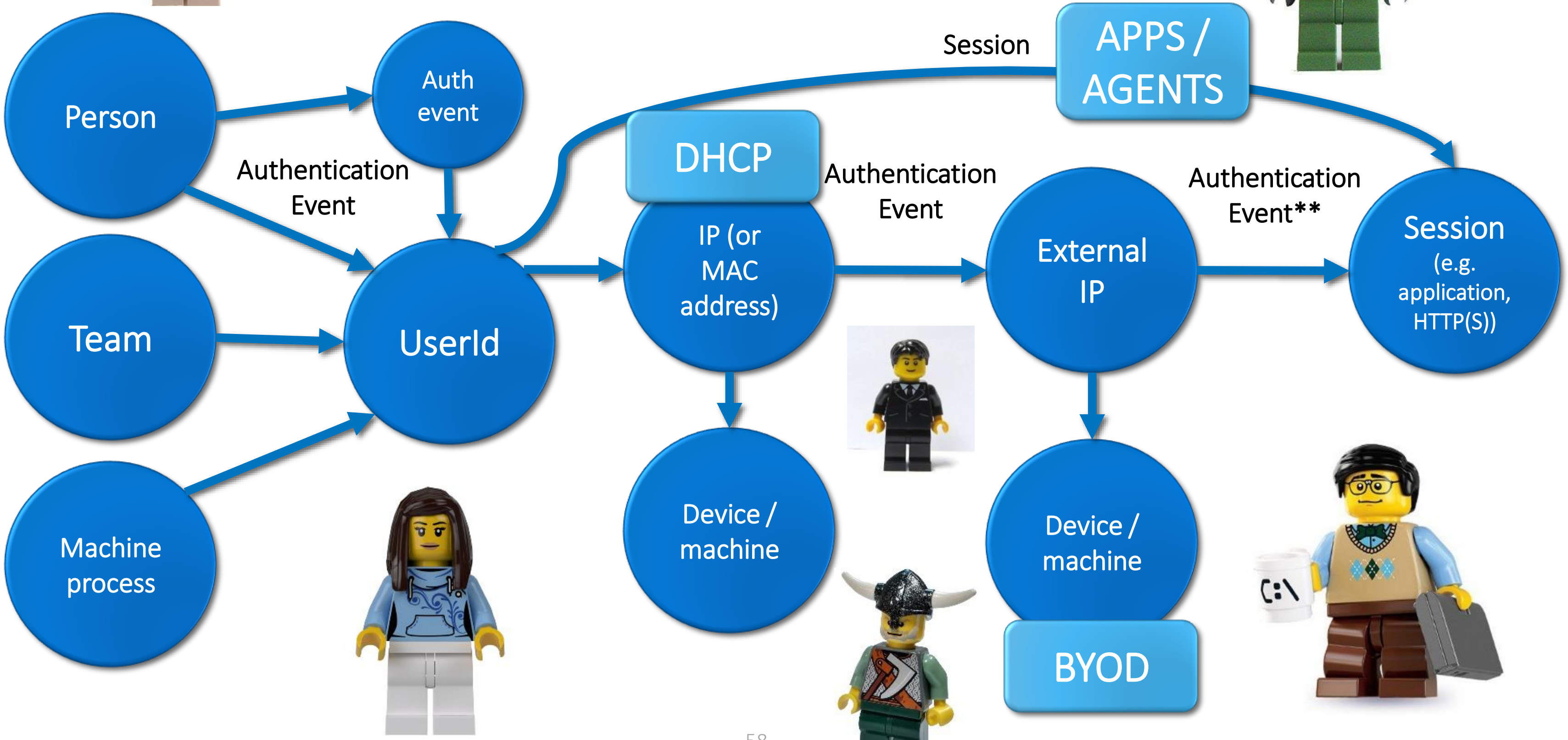


Entity Resolution

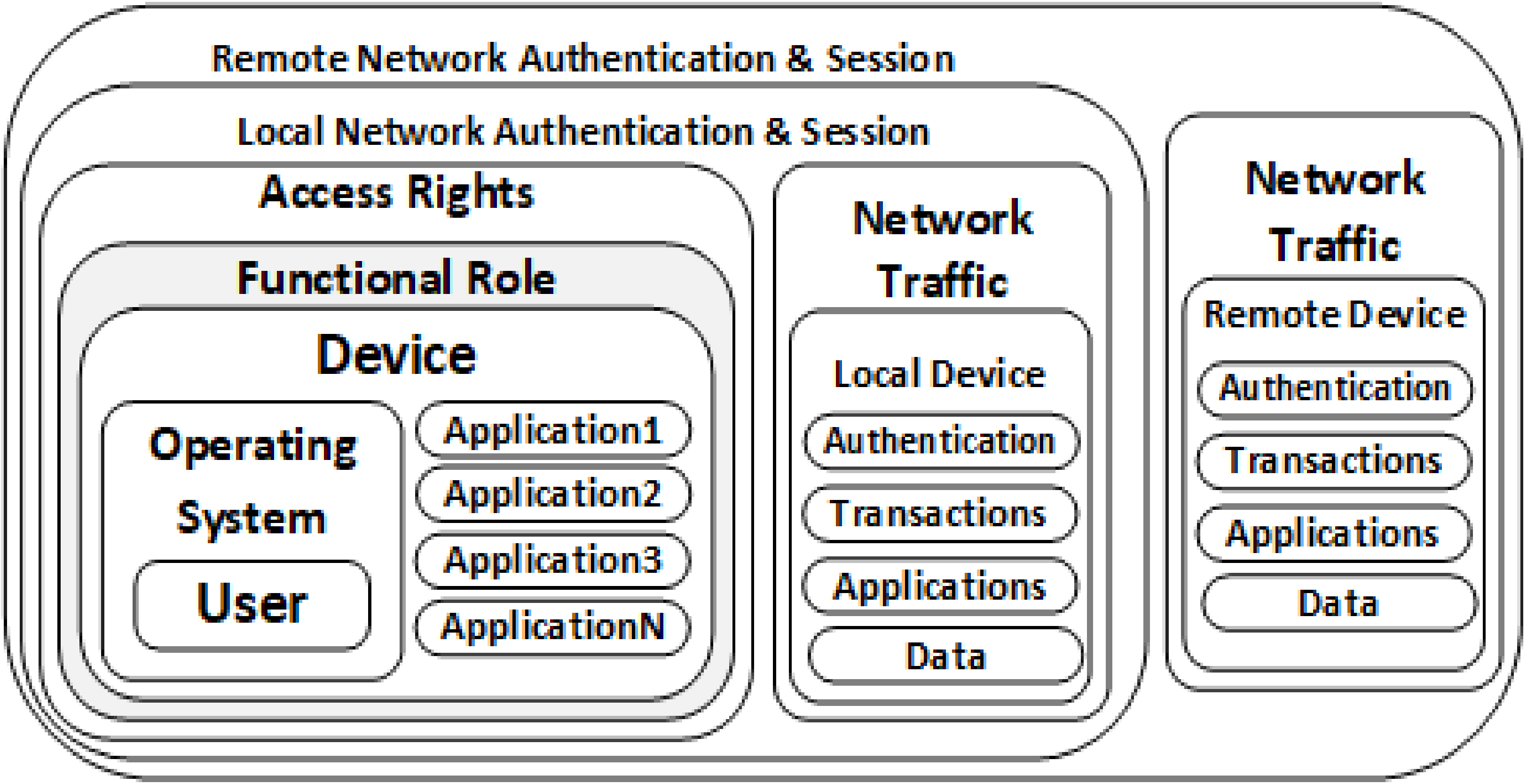


What is a User, anyway?

What is an IP address, anyway?



Entity Relational Specification



‘Cyborg’ behavioral profile

CSDS Data Processing

EDA + Feature Engineering (example)

1. **Data Acquisition**
   Unstructured or Semi-structured
   - Streaming or batch
   Sources 1,2,3,N

2. **Aggregation**
   Semi-structured
   - Physical co-location
   - Primary structuring

3. **Integration**
   Structured
   - Filtering
   - Combinations of primary events

4. **Normalization**
   Tabular & relational
   - Primary entity resolution by event

5. **Binding**
   Blocks/cubes
   - Time epoch event aggregation
   - By event & entity

6. **Cubing Aggregation**
   Multivariate cross-sectional & longitudinal
   - Time epoch event aggregation
   - By entity & measure over time

7. **Cubing Summarization**
   Multivariate cross-sectional & longitudinal
   - Summarized / refined measures
   - By entity & summary measure over time

8. **Featurization**
   - Transforms
   - Binning
   - Parameterize etc.

9. **Enhancement**
   - Calculated measures (e.g. mean, SD, rate of change, etc.

10. **Derivation**
    - Graphs measures, multivariate parameters, etc.

11. **Discernment**
    - Mislabeled?
    - Misbehaving?
    - Outlier?
    - Exception?
    - Segmentation

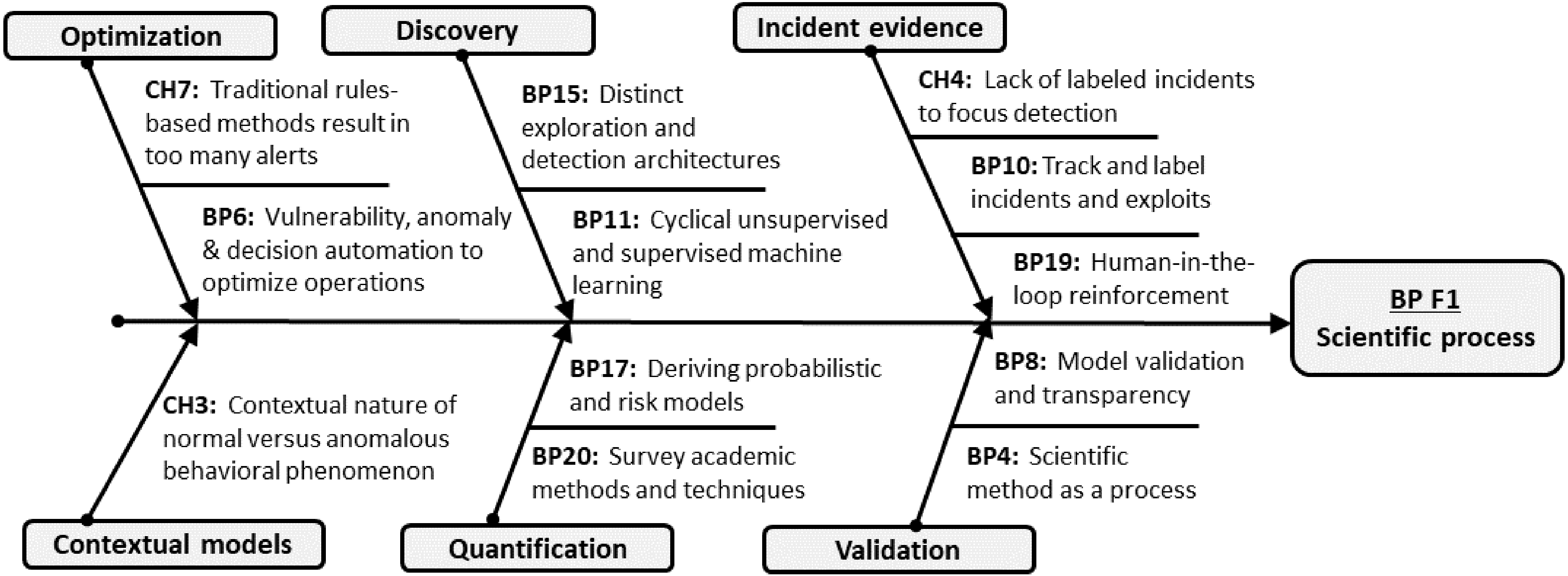
12. **Segmentation**
    - Multivariate segmentation
    - e.g. peer group extrapolation

13. **Validation**
    - Back- and hold-out testing
    - Variable significance
    - Predictive integrity assessment

14. **Heuristics**
    - Labeling for contextual operational handling & support


ORGANIZATION
TECHNOLOGY
PROCESS
Scientific Process

Root Cause Analysis: Fishbone / Ishikawa Diagram

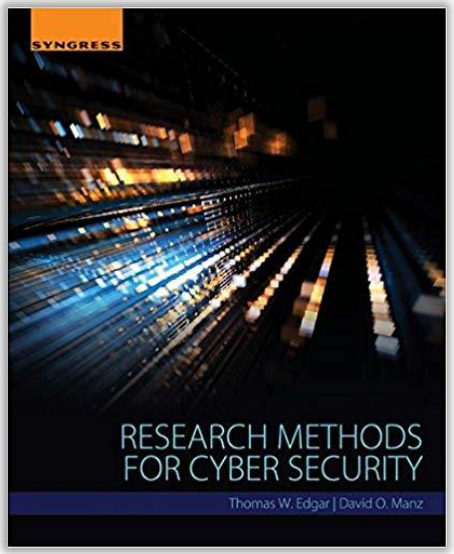


** Resulting from factor analysis and factor-to-factor fitting*

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

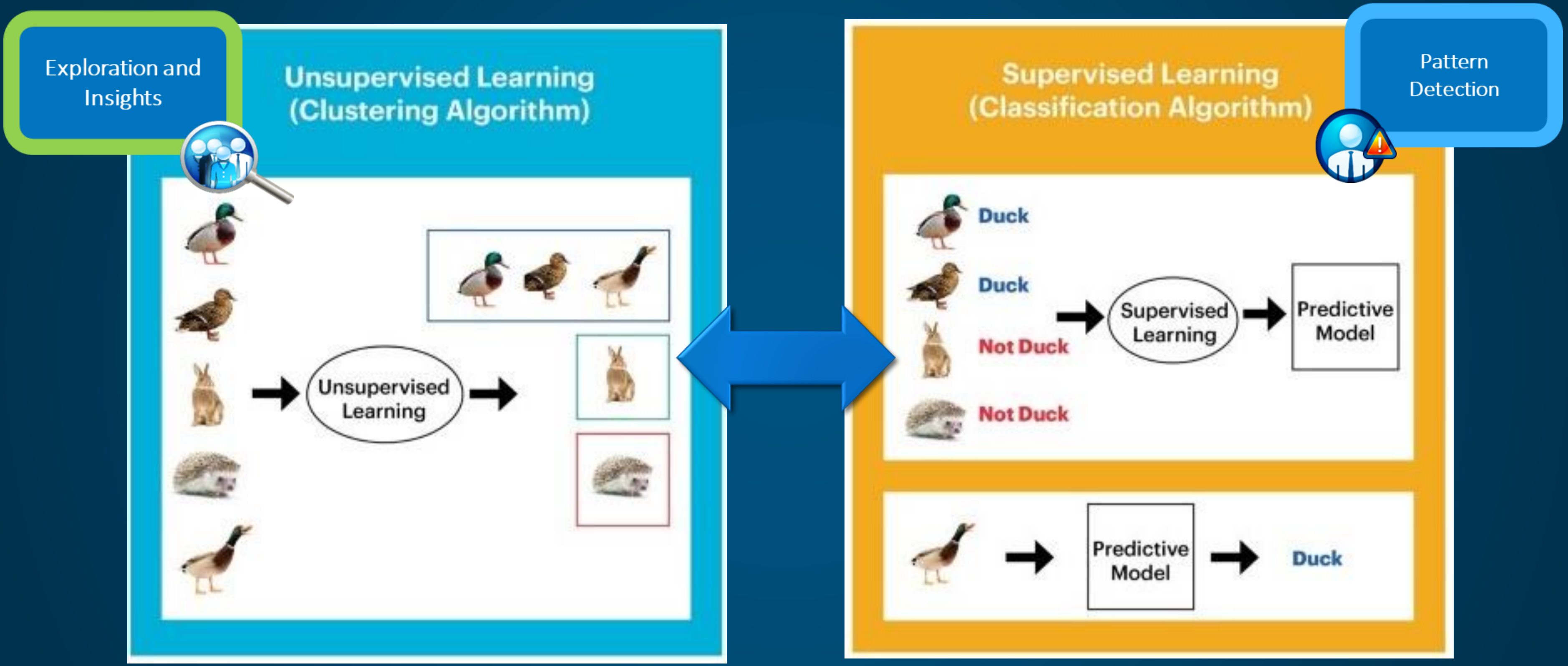
Labels: What constitutes 'evidence'?

	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)

Discovery ⇔ Detection



SEGMENTATION

CATEGORIZATION

<u>Technology</u>: Architect Exploratory & Detection Platforms*

Functional Architectural Segmentation

Exploratory ‘big data’ repository

Operationally focused detection

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

Canonical ontology / schemas

<u>Analytical models</u>
- Descriptive
- Unsupervised

<u>Analytical models</u>
- Semi-supervised
- Human-in-the-loop
- Reinforcement

<u>Analytical models</u>
- Statistical
- Supervised

** Runs counter to the industry vendor stance of store ‘all-the-data-all-the-time’*

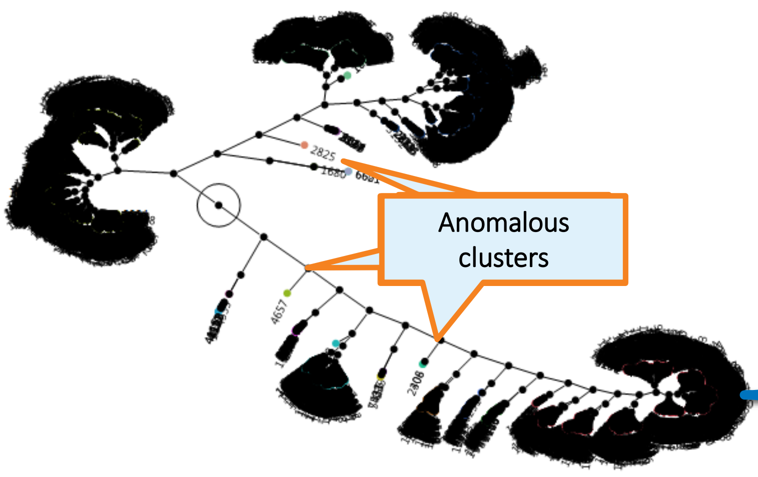

Auntie Anne's
WAY OUT WEST
NEWSSTAND

CSDS as a Process: Discovery and Detection

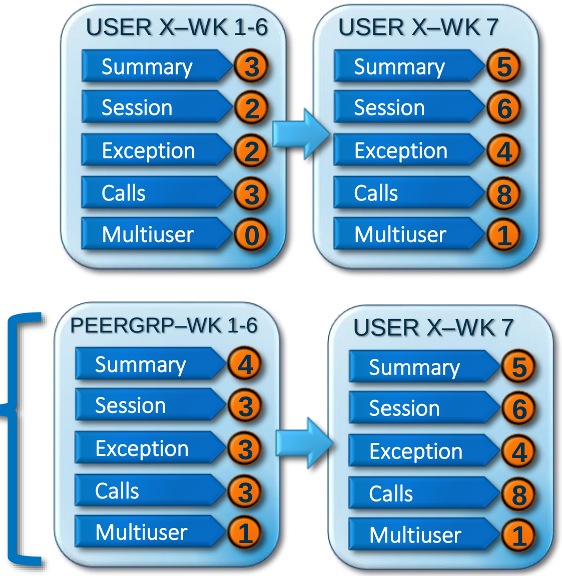

HUMAN
ORGAN
FOR TRANSPLANT

Unsupervised Discovery

Disassociating ‘Normal’ from ‘Abnormal’



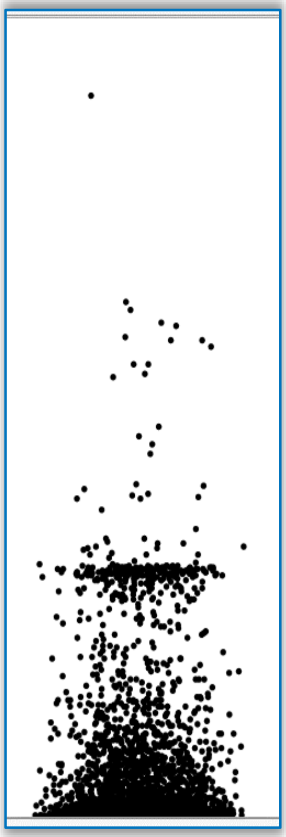
UNSUPERVISED MACHINE LEARNING

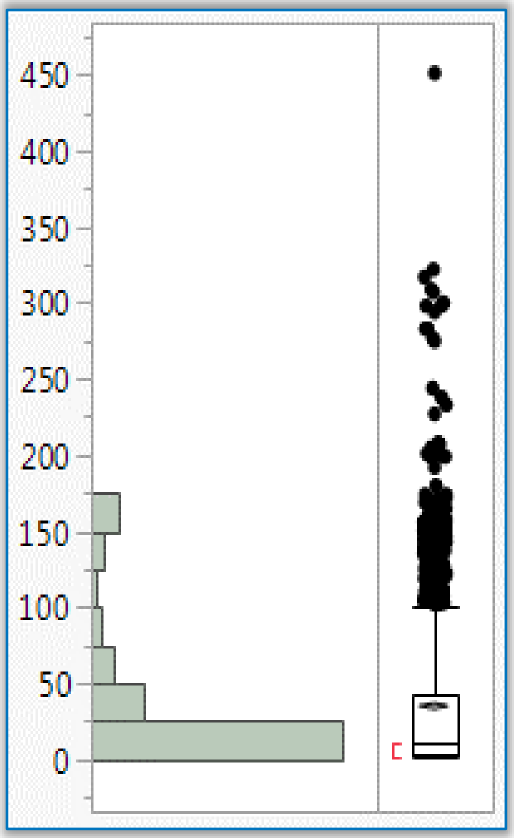


DEVIATION FROM OWN PATTERNS
(OWN & PEER GROUP)

CSDS Theory Development

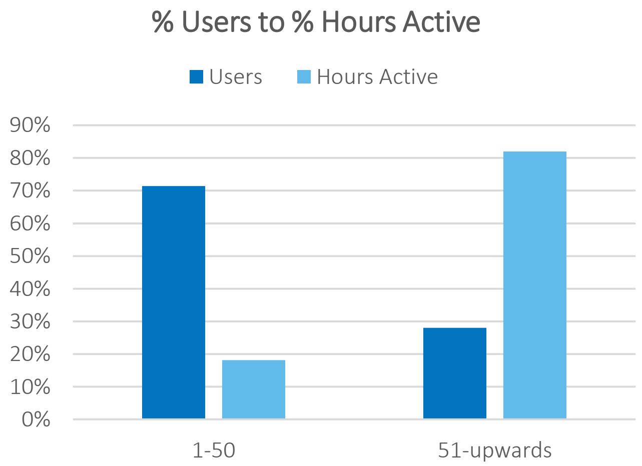
Example: Cyborg Network Behavioral Principals





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



‘The Normals’*

22 weeks of behavioral clustering

SIX MAJORS PEER GROUP CLUSTERS

1: Infrequent users (~50%)

2: Sporadic use / low activity (~20%)

3: Active / specialized (~15%)

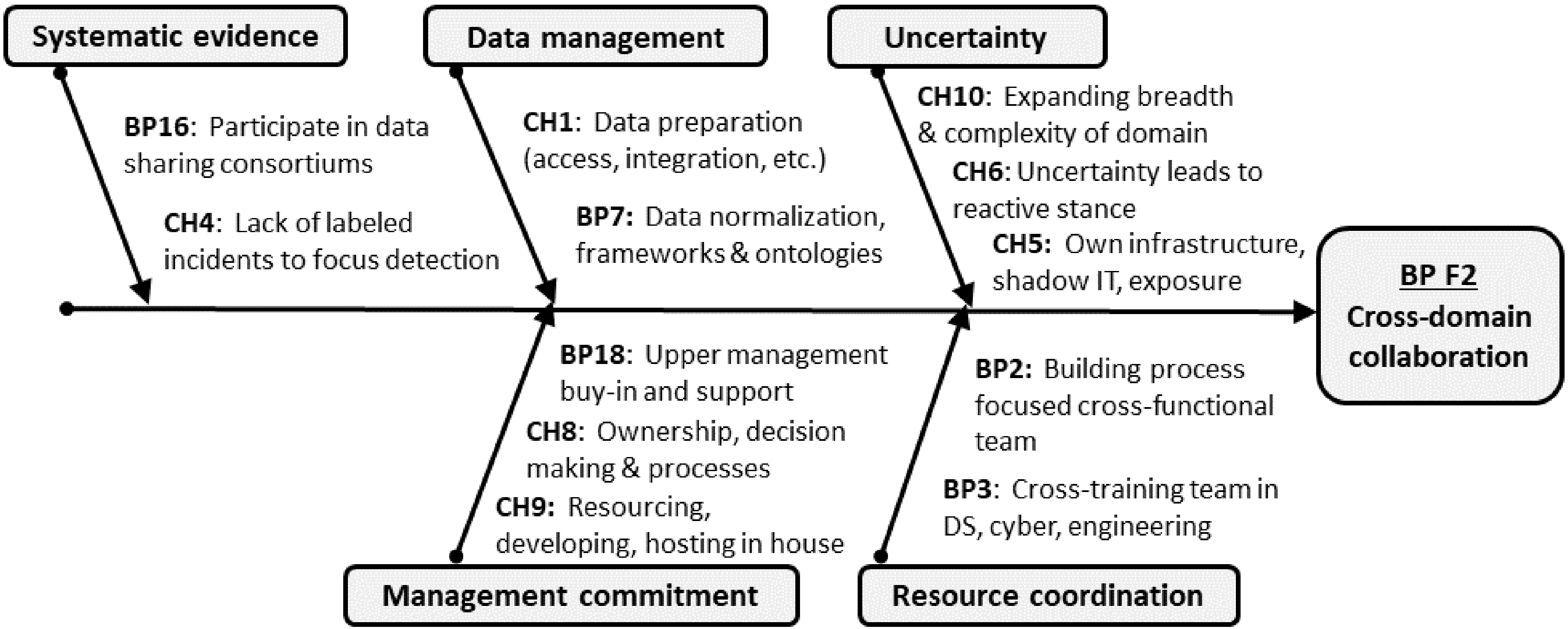
4: Active generalists (~6%)

5: Very active / specialized (~6%)

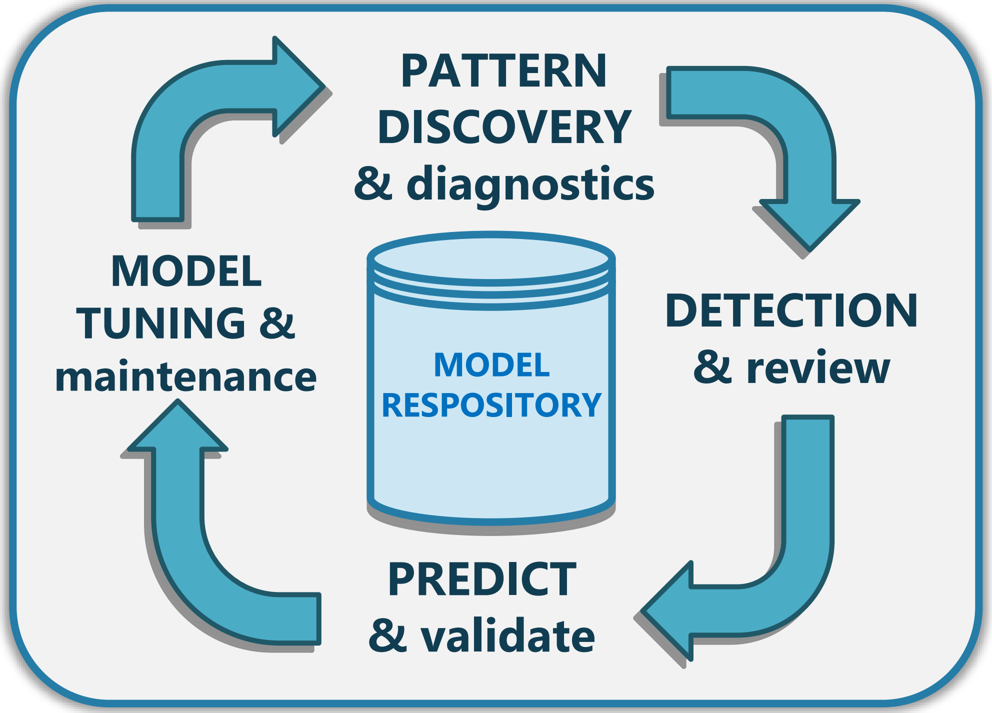
6: Sporadic high-low active (~3%)

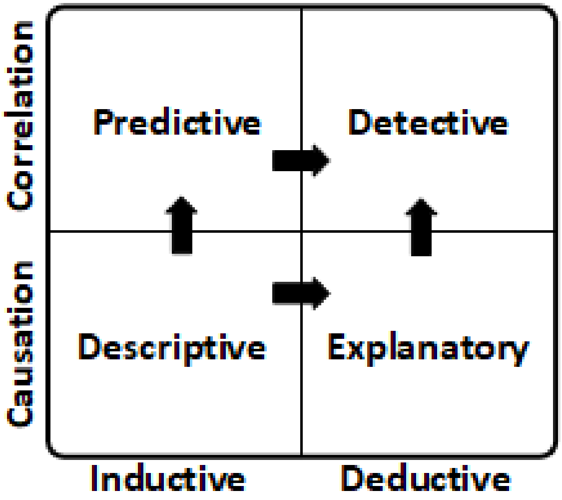
** After 2% ‘unusuals’ removed*

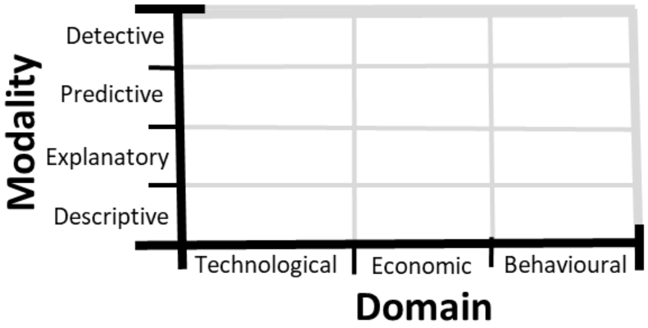

Cross-Domain Collaboration
ORGANIZATION
TECHNOLOGY
PROCESS

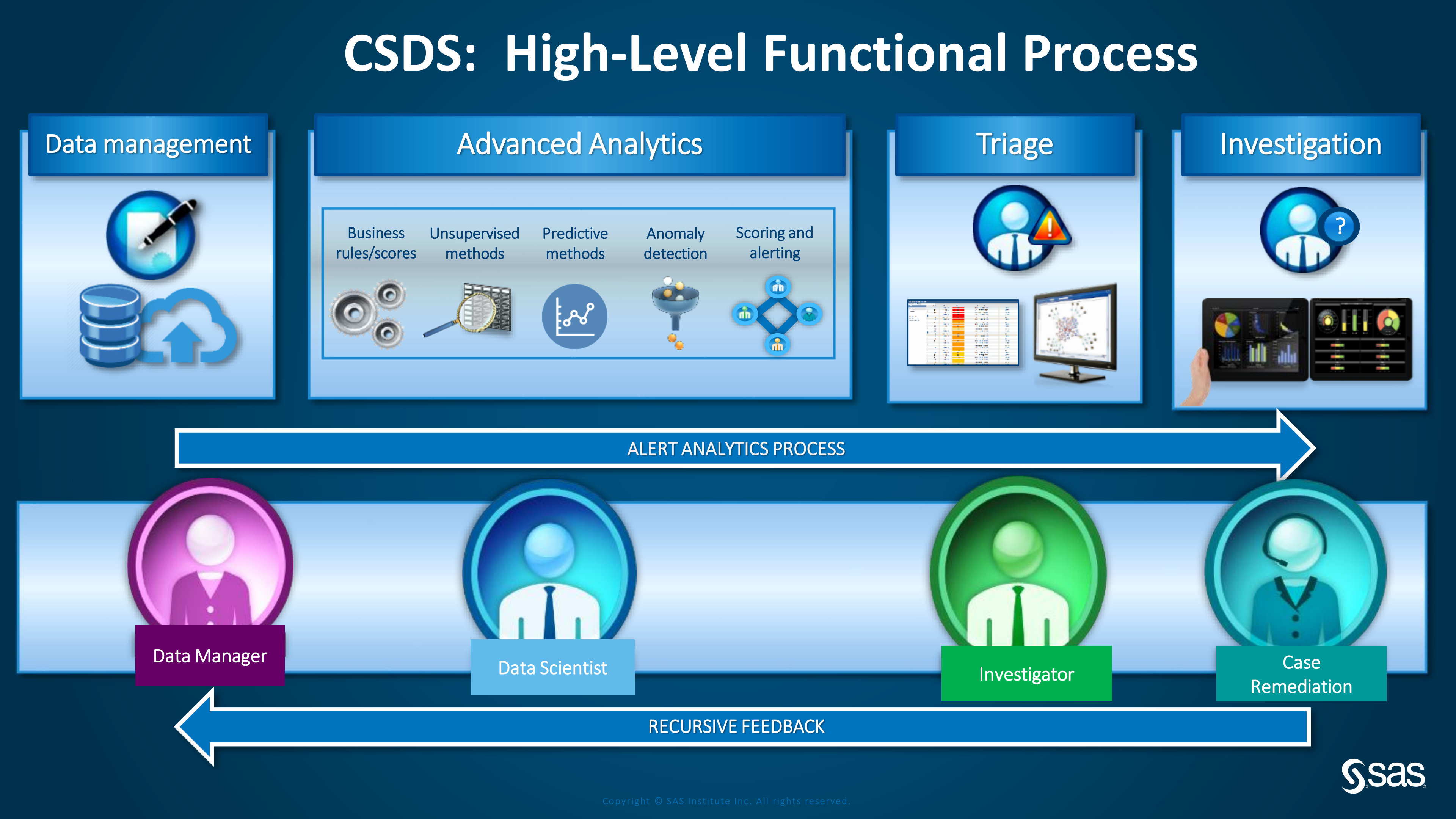

Systematic evidence
Data management
Uncertainty
BP16: Participate in data sharing consortiums
CH4: Lack of labeled incidents to focus detection
CH1: Data preparation (access, integration, etc.)
BP7: Data normalization, frameworks & ontologies
CH10: Expanding breadth & complexity of domain
CH6: Uncertainty 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

Staged Discovery Process





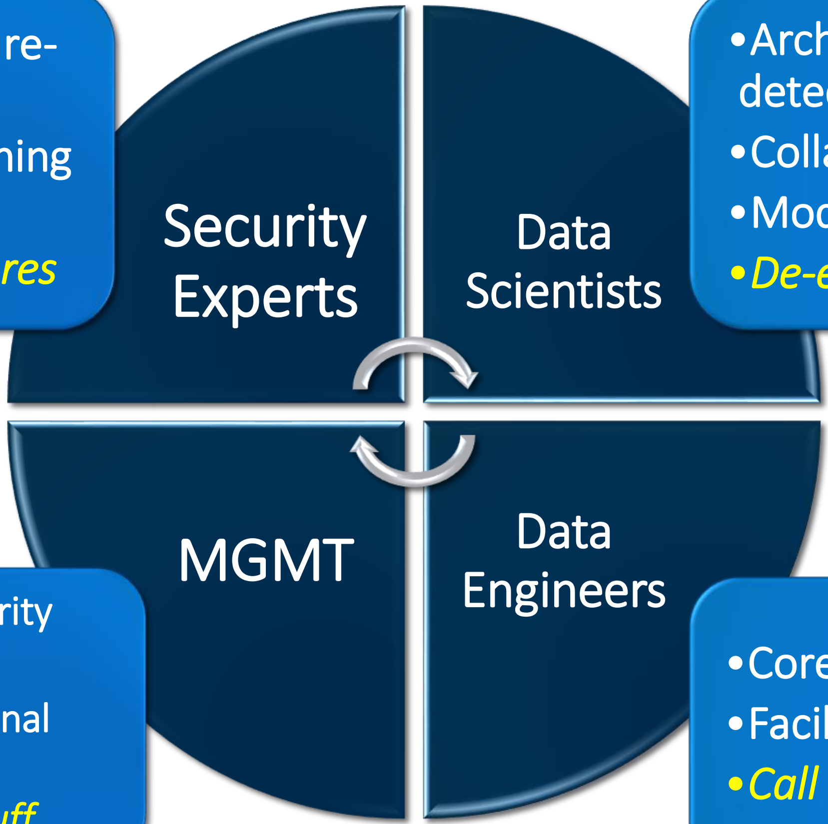



CSDS: High-Level Functional Process
Data management
Advanced Analytics
Business rules/scores
Unsupervised methods
Predictive methods
Anomaly detection
Scoring and alerting
Triage
Investigation
ALERT ANALYTICS PROCESS
Data Manager
Data Scientist
Investigator
Case Remediation
RECURSIVE FEEDBACK
Copyright © SAS Institute Inc. All rights reserved.

<u>Organization</u>: Interdisciplinary Collaboration

- Collaborate in process re-engineering
- Collaborate in establishing model context
- *Admit limits of signatures*

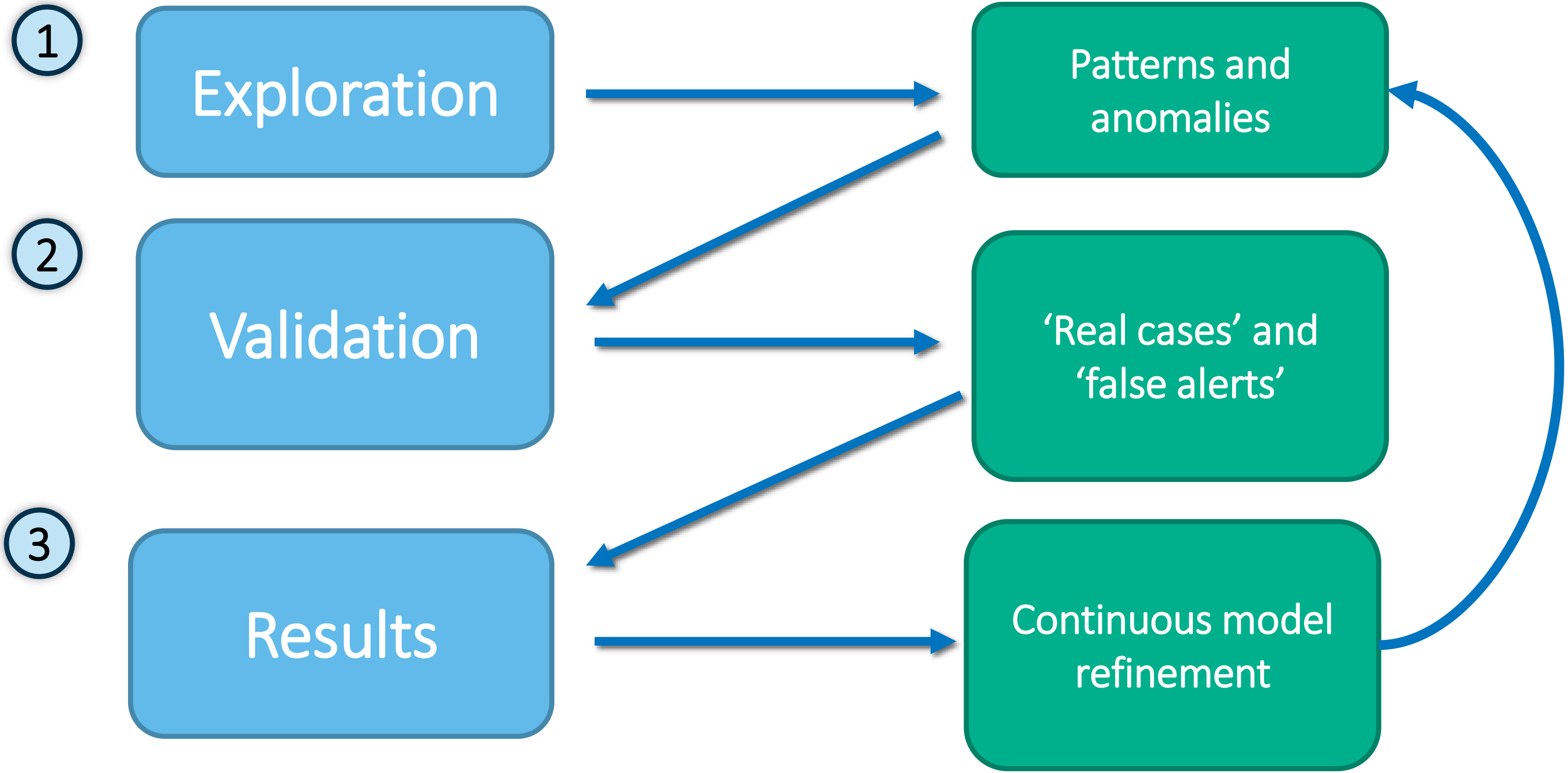
- Architect exploration and detection processes
- Collaborative model building
- Model transparency
- *De-escalate "AI hype cycle"*



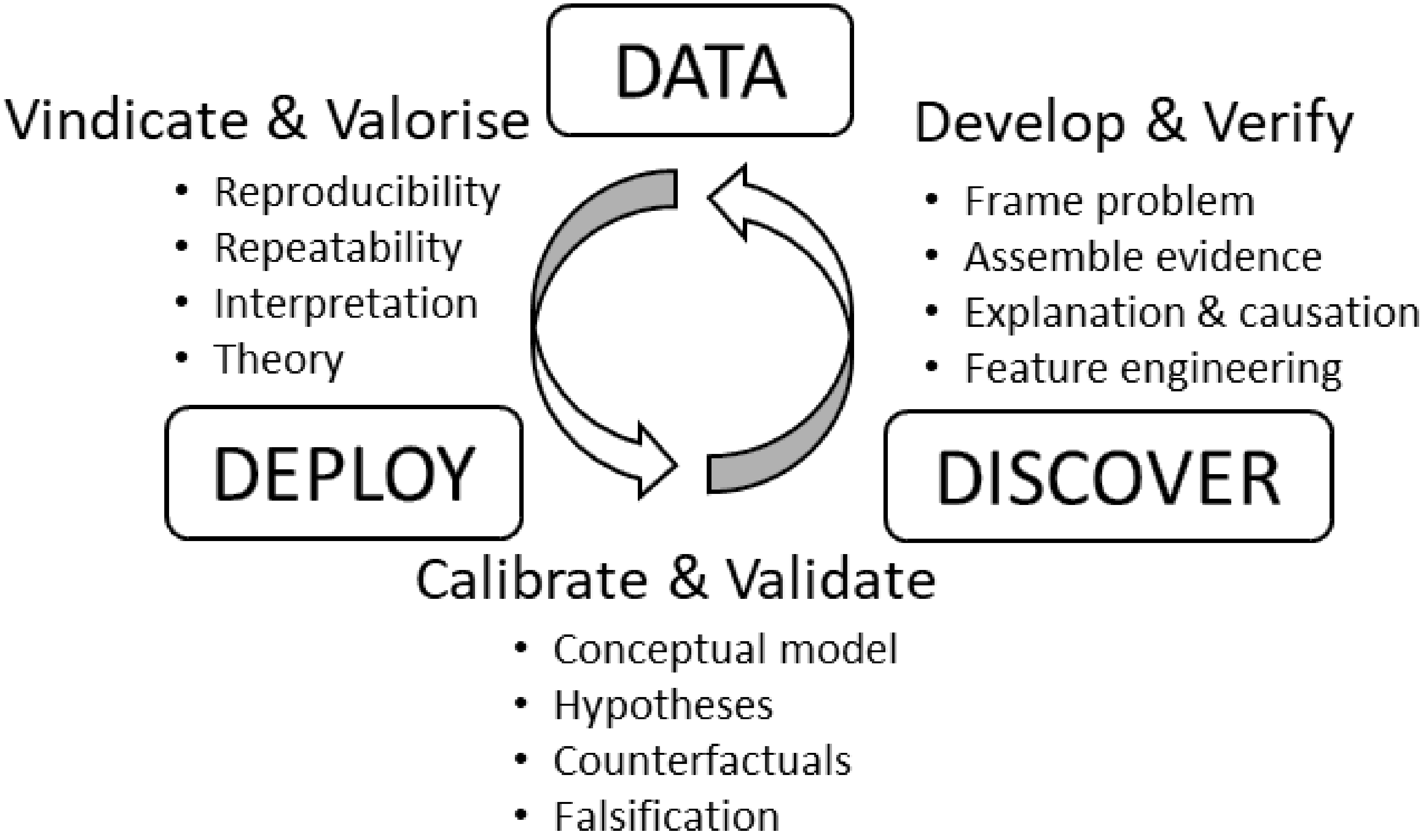
- Decision & ownership clarity
- Training & team building
- Orchestrate cross-functional collaboration (incentives)
- *Call "AI = automation" bluff*

- Core data 'pipeline' processing
- Facilitate processes / quality
- *Call "data lake = strategy" bluff*

Continuous Detection Improvement Process



CSDS Model Development Process



V. Conclusions

Cybersecurity

Data

Science

Not so much…

but, ASPIRATIONAL!

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

Foundation: CSDS Maturity Framework

Anomalies

- Big data overload
- Flags, rules, and alerts

Chasing phantom patterns

Discovery

Understanding

- Feature engineering
- Diagnostics
- *Unsupervised ML*

Prediction

Learning

- Human-in-the-loop reviews
- *Combined supervised and unsupervised machine learning*

Optimization

Optimal

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

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

Cybersecurity Data Science (CSDS)

Best Practices in an Emerging Profession

Scott Allen Mongeau
INFORMS CAP®

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

s.mongeau@edp1.nyenrode.nl
scott@sark7.com
scott.mongeau@sas.com

@SARK7 #CSDS2020

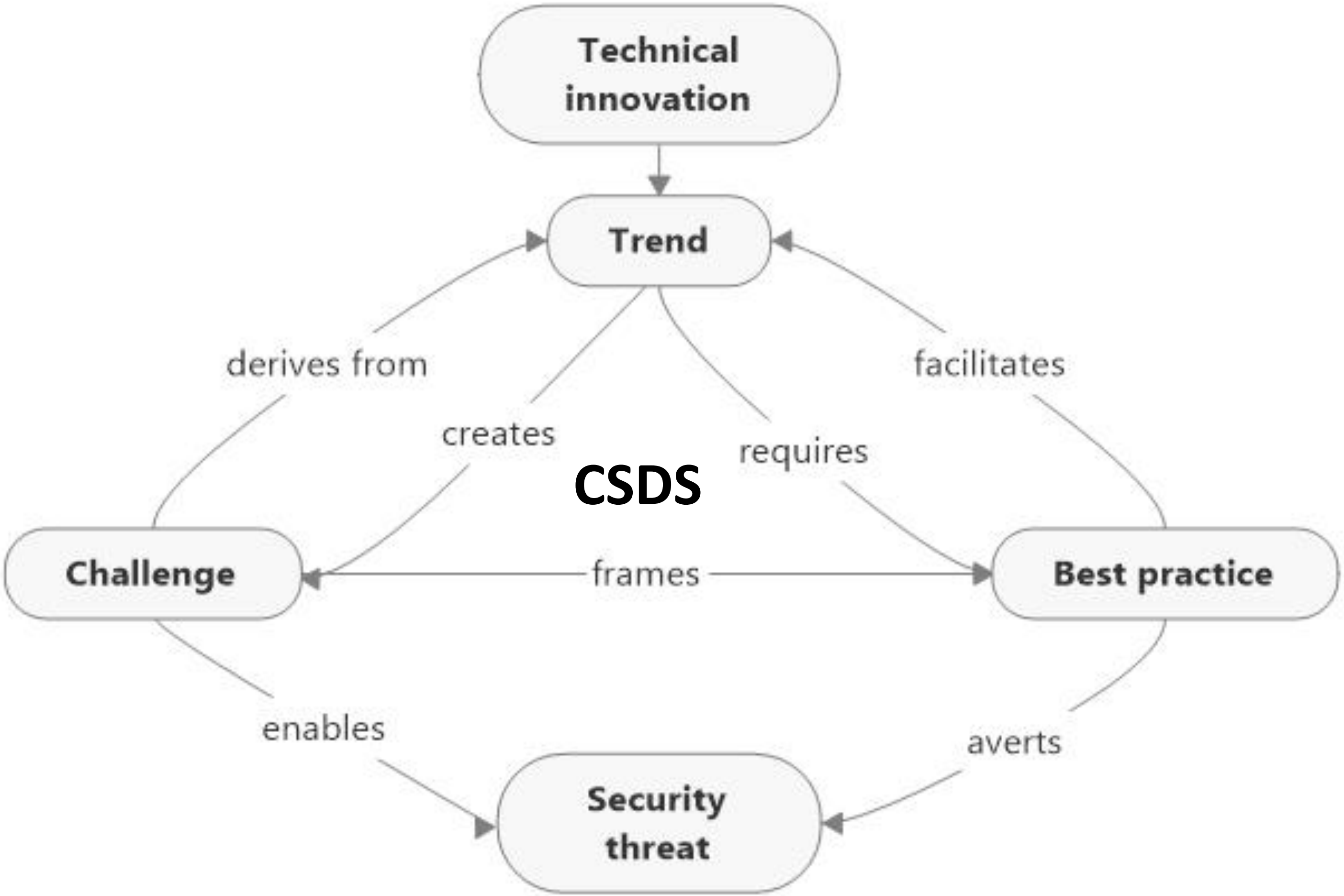
APPENDIX

Section	Phase	Method	Result
I 1. CSDS as an emerging profession T	Diagnostic background analysis	Integrated literature review	CSDS professional maturity gaps Sensitizing concepts
II 2. CSDS practitioner interviews E	Diagnostic opinion research Diagnostic gap analysis	Qualitative interview research Quantitative analysis of themes	Key challenge & best practice themes Diagnosis of CSDS gaps
III 3. CSDS methodological prescriptions A	Design requirements	Structural requirement analysis	CSDS methodological design prescriptions

Research Overview

RESEARCH OBJECTIVE: <u>Diagnose</u> and <u>prescribe</u> treatment <u>designs</u> to address <u>gaps</u> *impeding the development of CSDS professional practice*

- DIAGNOSTIC RESEARCH: Undertaken to analyse, diagnose, and prescribe design treatments to address gaps resident in CSDS practice
- BUSINESS GOAL: Facilitate professional advancement of the CSDS domain by addressing 'body of theory' gaps
- <u>ACADEMIC CONTRIBUTION</u>
  - <u>Diagnosis</u> for a novel topic
    definition and awareness of a problem
    => *addresses research lacuna*
  - <u>Design prescriptions</u> to address empirically identified gaps
    conceptual and theoretical suggestions to address practical shortcomings => *addresses management theory need*


Technical innovation
Trend
derives from
creates
CSDS
requires
facilitates
Challenge
frames
Best practice
enables
averts
Security threat

CSDS High-Level Overview

- Represents a partial paradigm shift from traditional cybersecurity
  - **Cybersecurity** = rule-and-signature-based and focuses on boundary protection
  - **CSDS** = situational awareness and assumes persistent and prolific threats
- CSDS is data focused
  - Applies quantitative, algorithmic, and probabilistic methods
  - Attempts to quantify risk
  - Focuses on producing focused and efficacious alerts
  - Promotes inferential methods to categorize behavioral patterns
  - Ultimately seeks to optimize cybersecurity operations
- Emerges from two parent domains…
  - Which themselves are undergoing rapid transformation
  - As such, ‘body of theory’ surrounding CSDS is evolving

CSDS Definition

- The practice of data science…
- to assure the continuity of digital devices, systems, services, software, and agents…
- in pursuit of the stewardship of systemic cybersphere stability,…
- spanning technical, operational, organizational, economic, social, and political contexts

CSDS Curriculum Design I

- 1.0 Introduction to the CSDS field 1.1. Cybersecurity basics and challenges
  - 1.2. Data science basics and challenges
  - 1.3. CSDS as a focused hybrid domain
  - 1.4. Differentiating analytics goals and methods
  - 1.5. Framing the cybersecurity analytics lifecycle
  - 1.6. Introducing cybersecurity analytics maturity

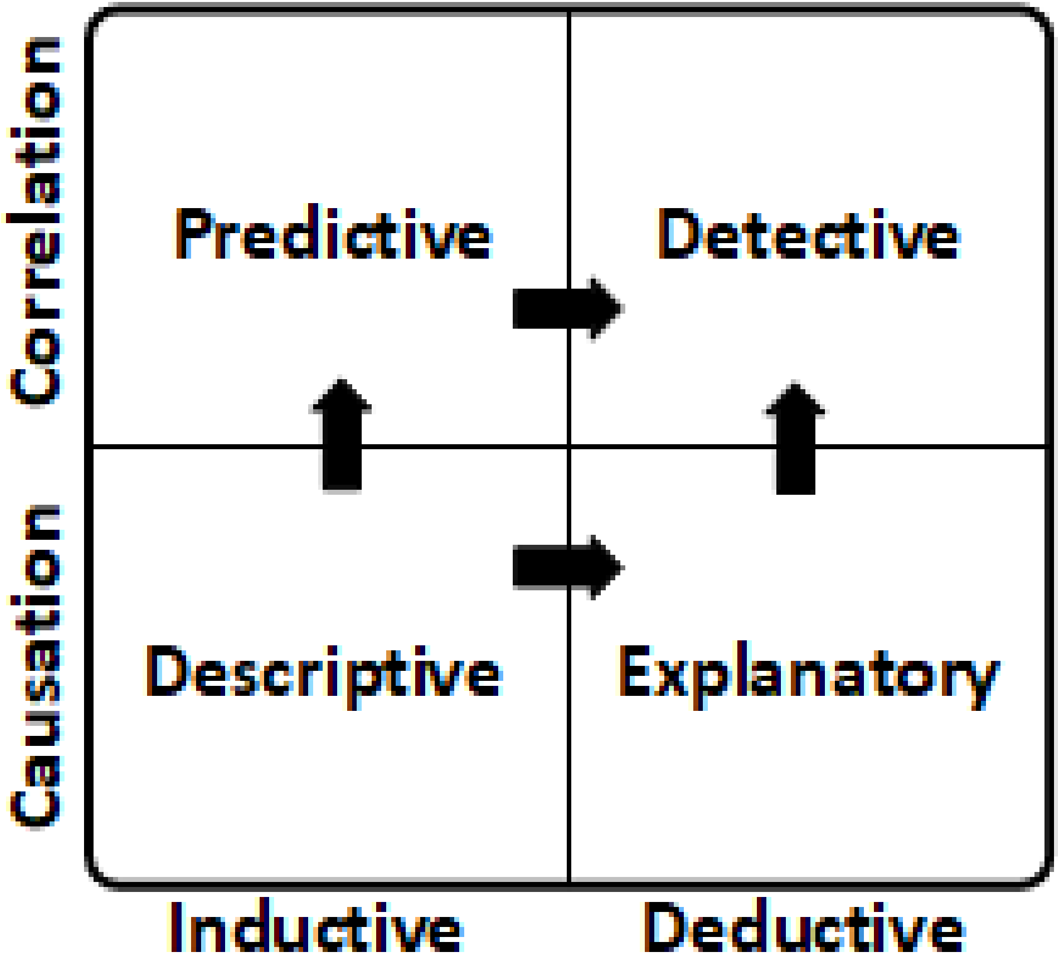
- 2.0 Cybersecurity data: challenges, sources, features, methods
  - 2.1. Sources of cybersecurity data, research datasets, types of evidence
  - 2.2. Examples: log files and network traffic
  - 2.3. Data preparation, quality, and processing
  - 2.4. Statistical exploration and analysis (EDA)
  - 2.5. Feature engineering and selection
  - 2.6. Feature extraction and advanced methods
  - 2.7. Positioning and handling real-time and streaming data

CSDS Curriculum Design II

- 3.0 Exploration and discovery: pattern extraction, segmentation, baselining, and anomalies
  - 3.1. Building contextual knowledge
  - 3.2. Segmentation and categorization
  - 3.3. Multivariate analysis
  - 3.4. Parameterization and probability
  - 3.5. Outliers and differentiating normal from abnormal
  - 3.6. Anomaly types, anomaly gain, and detection
  - 3.7. Unsupervised machine learning
  - 3.8. Establishing a foundation for prediction
- 4.0 Prediction and detection: models, incidents, and validation
  - 4.1. Distinguishing explanation versus prediction
  - 4.2. Framing detective analytics: combining explanation and prediction
  - 4.3. Econometric approaches
  - 4.4. Predictive machine learning (supervised machine learning)
  - 4.5. Deep learning
  - 4.6. Reinforcement learning
  - 4.7. Model diagnostics and management
  - 4.8. Bootstrapping detection: semi-supervised machine learning

CSDS Curriculum Design III

- 5.0 Operationalization: CSDS as-a-process
  - 5.1. Analytics process management: integrating discovery and detection
  - 5.2. Human-in-the-loop: integrating investigations and investigative feedback
  - 5.3. Robo-automation, online machine learning, and self-improving processes
  - 5.4. Technical and functional architectures
  - 5.5. Systems integration and orchestration
  - 5.6. Cybersecurity analytics maturity recap
  - 5.7. Cybersecurity risk and optimization
  - 5.8. Guidance on implementing CSDS programs


Correlation
Causation
Predictive
Detective
Descriptive
Explanatory
Inductive
Deductive