Carnegie Mellon University
Software Engineering Institute

Cybersecurity Data Science

Best Practices from the Field

Scott Allen Mongeau
scott.mongeau@sas.com

Cybersecurity Data Scientist – SAS Institute
Lecturer / PhD candidate – Nyenrode Business University

@SARK7 #CSDS2020 #FloCon19

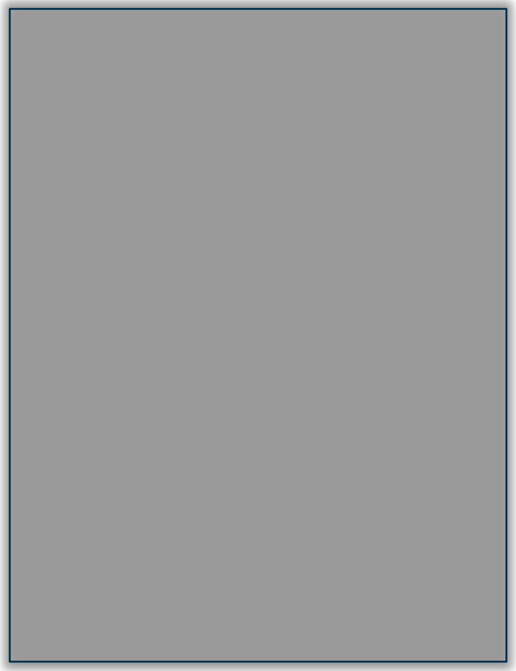
FloCon 2019
JANUARY 7-10, 2019 | NEW ORLEANS, LA

INTRODUCTION

- Cybersecurity Data Science practitioner – *SAS Institute*
- Lecturer / PhD candidate – *Nyenrode Business University*

- Qualitative research
  - 43 global cybersecurity data scientists
  - Key challenges and best practices
  - Organizational & methodological guidance
  - Book early 2020 #CSDS2020

  ‘Cybersecurity Data Science: Prescribed Best Practices’



Research Motivation: Genesis in Six Memes

Three Year Genesis of This Talk

FloCon 2017 – San Diego

- Interest in data analytics percolates
- But… cautious: *'I'll know it when I see it'*

2017: “THE CAUTIOUS TRADITIONALISTS”

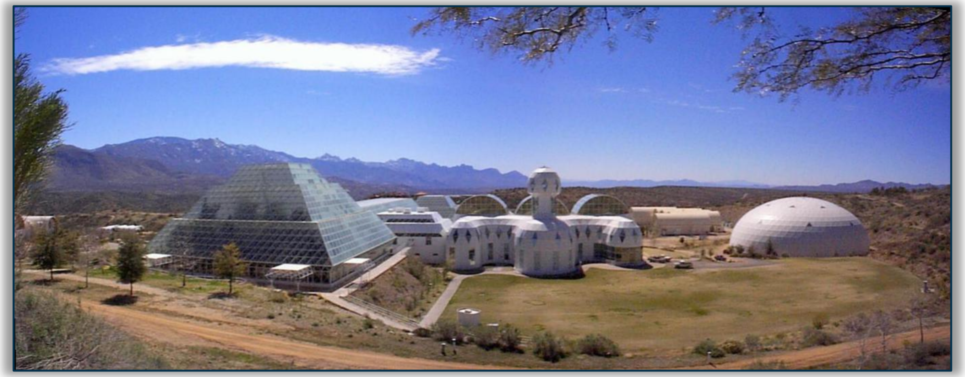
Three Year Genesis of This Talk

FloCon 2017 – San Diego

- Interest in data analytics percolates
- But… cautious: *'I'll know it when I see it'*

FloCon 2018 – Tucson

- Spike in analytics and machine learning cases
- But… questions emerge: *'How do we get from here to there?'*



Labeled for non-commercial reuse Wikipedia Commons

2018: "THE DATA REVOUTIONARIES"

<u>2018</u>: SAY ‘DATA SCIENCE’…

ONE… MORE… TIME!

Three Year Genesis of This Talk

FloCon 2017 – San Diego

- Interest in analytics percolates
- But…: *'I'll know it when I see it'*

FloCon 2018 – Tucson

- Spike in analytics and ML cases
- But…: *'How do we get there?'*

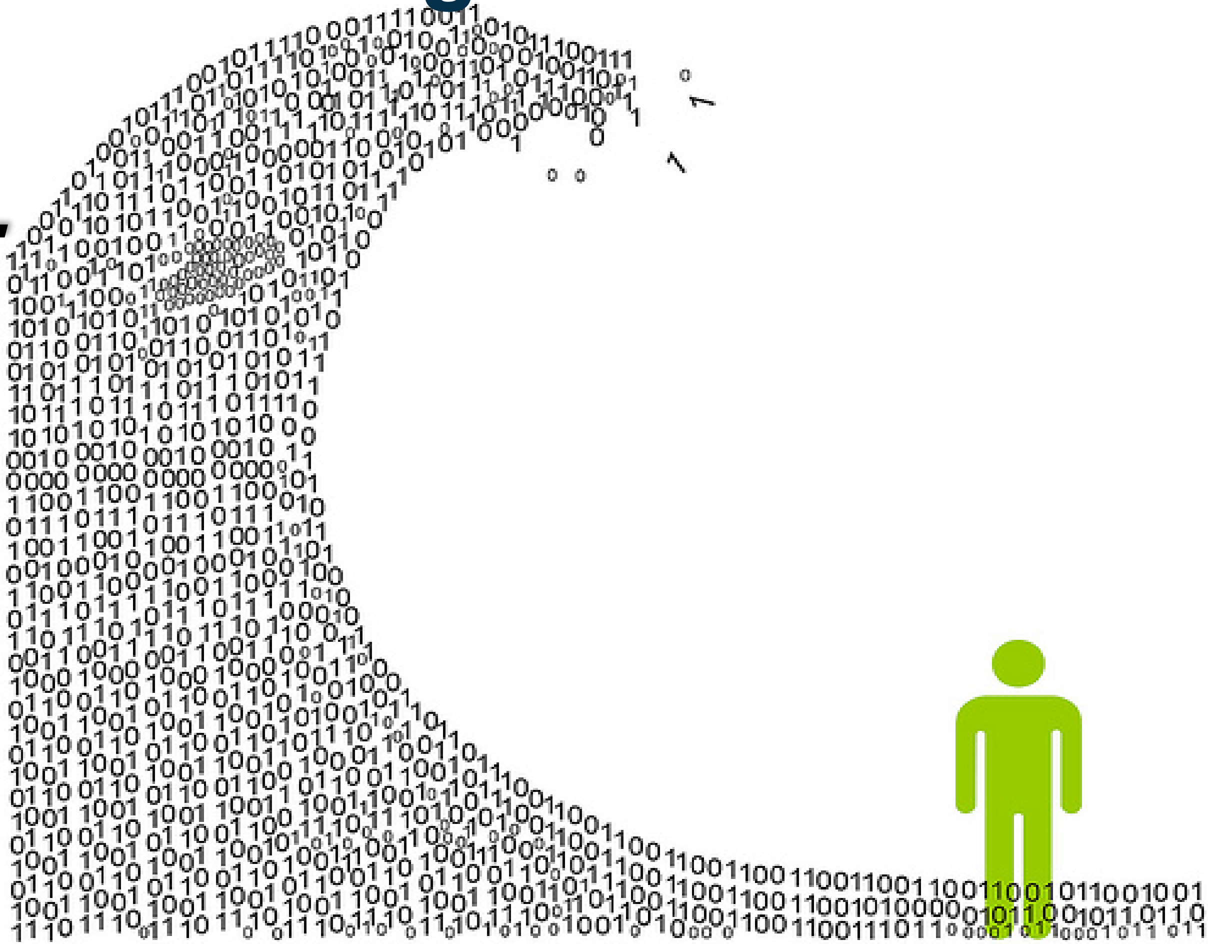
Labeled for non-commercial reuse Wikipedia Commons

FloCon 2019 – New Orleans

- Deafening market / vendor buzz
- But, caveats abound: '*Many are drowning in data lakes*'

<u>2019</u>: Drowning in Data Lakes

Vendor Hype

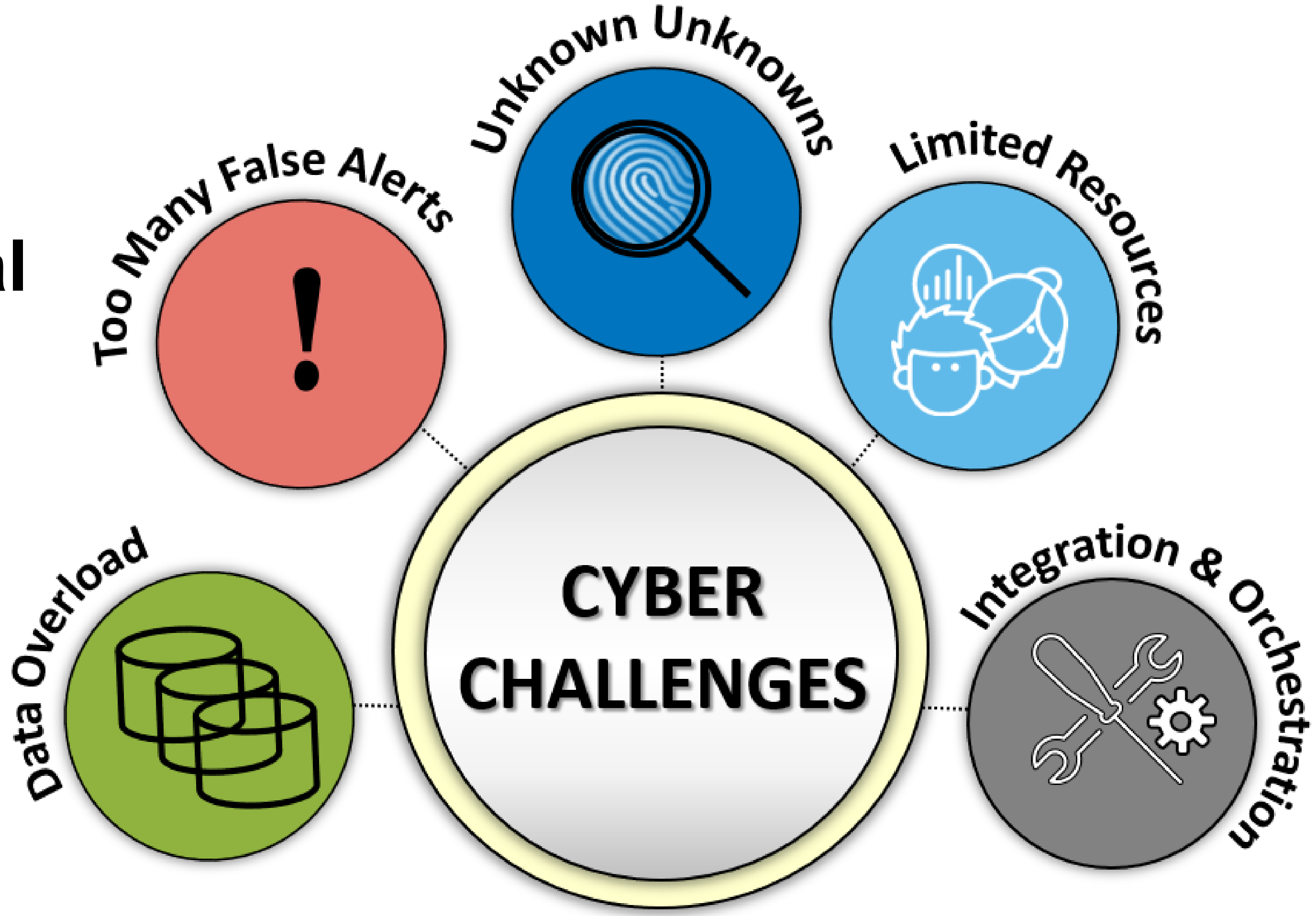


<u>2019</u>: ONE DOES NOT SIMPLY…

“PUSH A DEEP LEARNING MODEL TO PRODUCTION”

2019

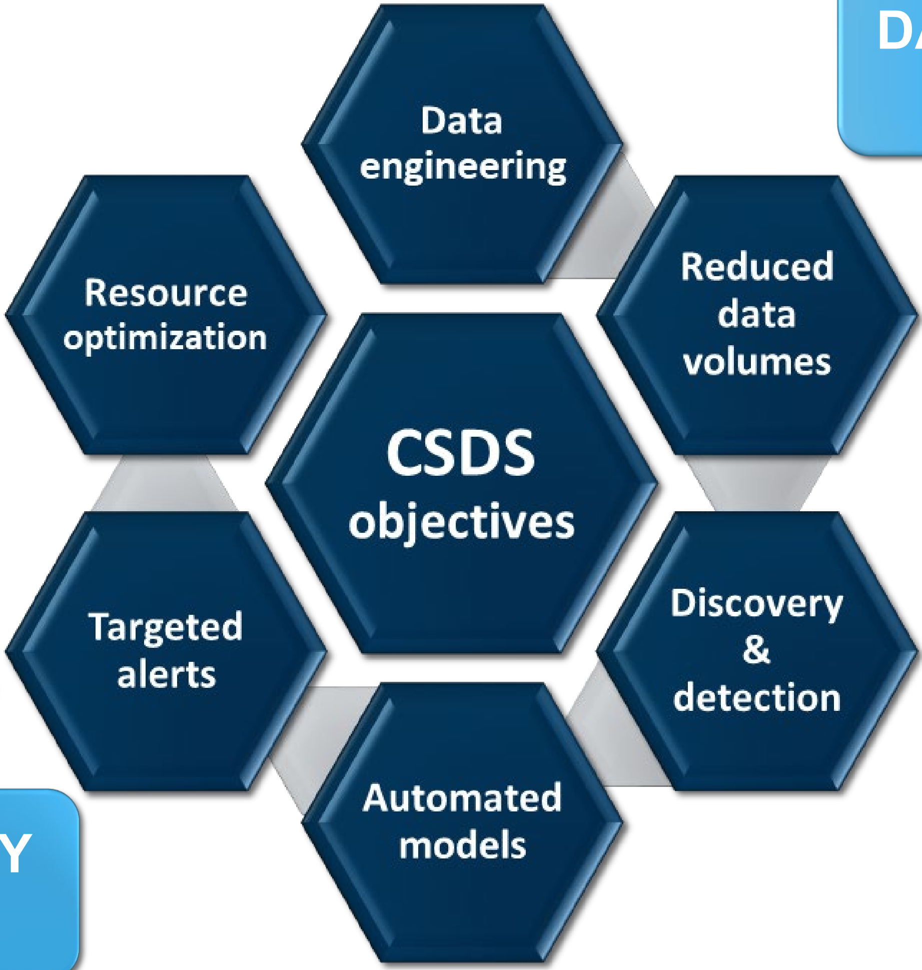
But… substantial issues grow



2019: Reactive militarization

2019
CSDS
Cyber Security Data Science

DATA SCIENCE METHODS



CYBERSECURITY GOALS

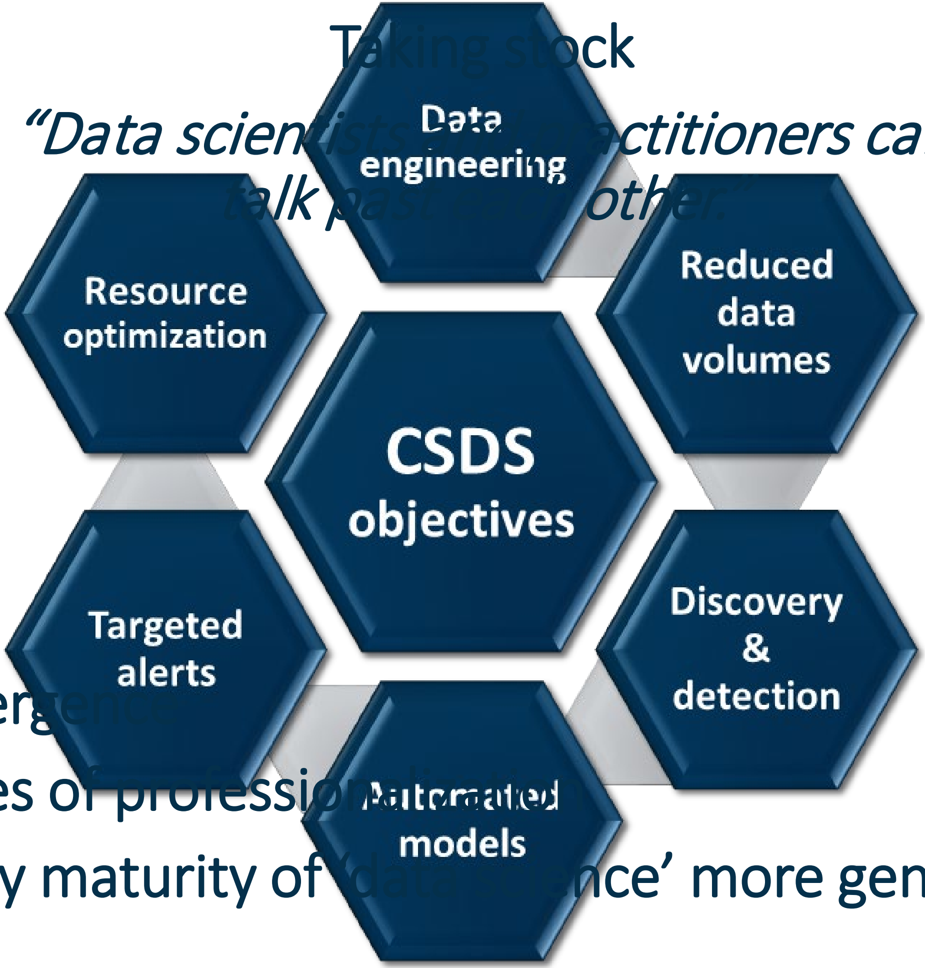
2019

CSDS

Cyber Security Data Science

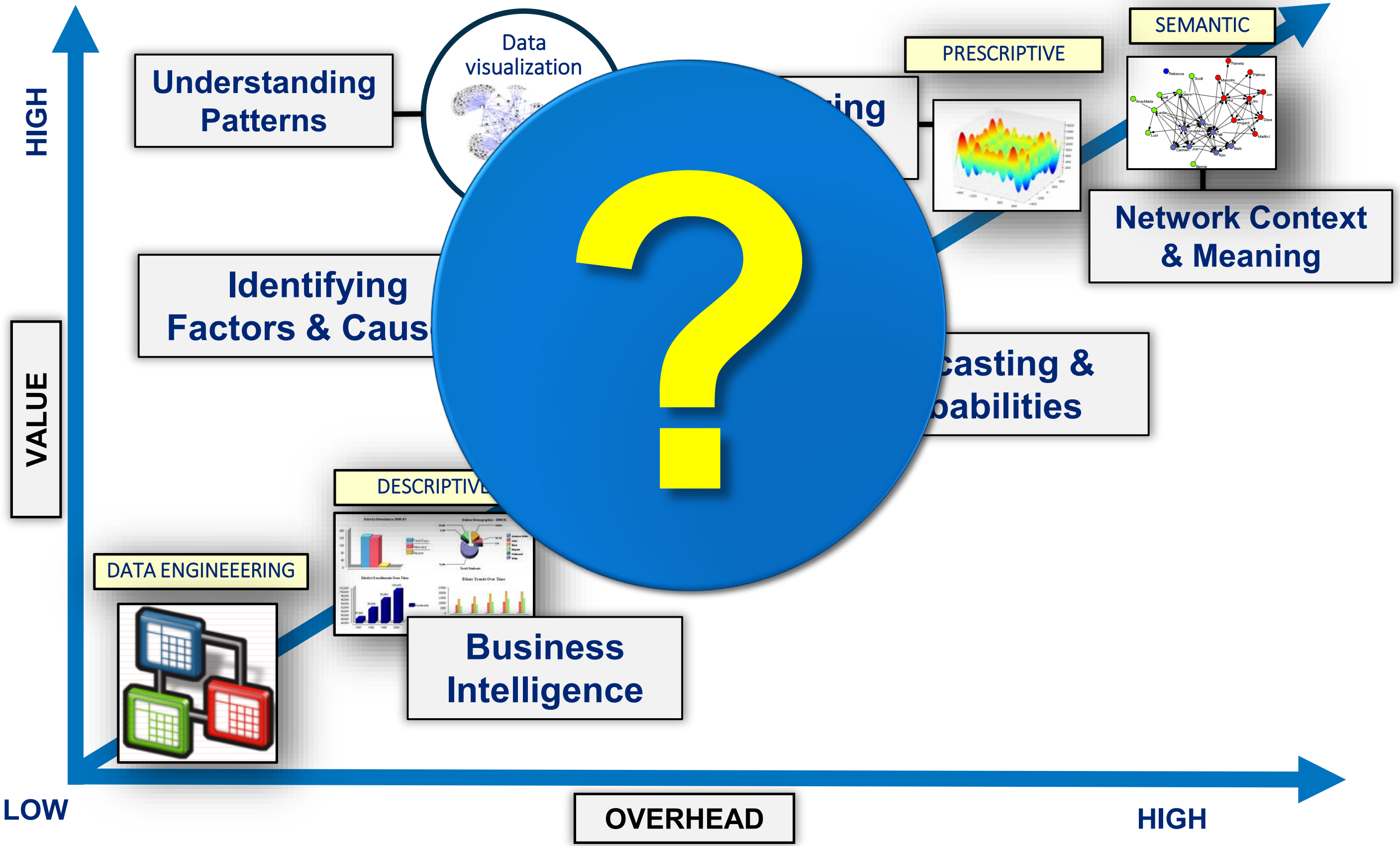
Taking stock

"Data scien… …ctitioners can talk … other."



- Rapid emer…
- Early stages of professio…
- Affected by maturity of '…ce' more generally

Data Science in 30 Seconds…



See YouTube lectures: https://bit.ly/SS9rCT

CSDS Interview Research

What Type of Data Science is CSDS?

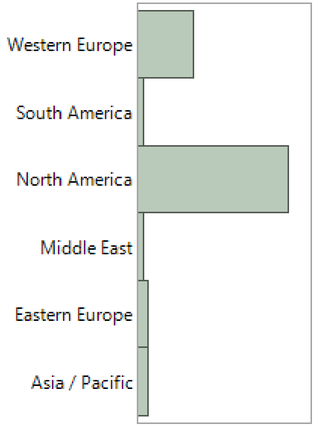
Participants - Sample

43 participants + 130 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

Demographic Profile (n=43)

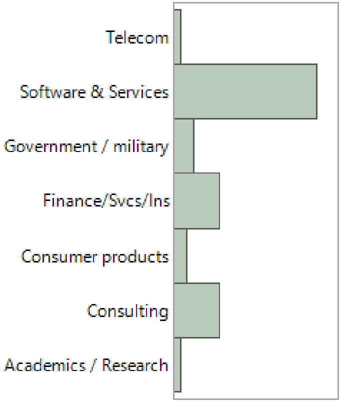
Current Region



	n	%
North America	27	63%
Western Europe	10	23%
Asia / Pacific	2	5%
Eastern Europe	2	5%
Middle East	1	2%
South America	1	2%
Total	**43**	**100%**

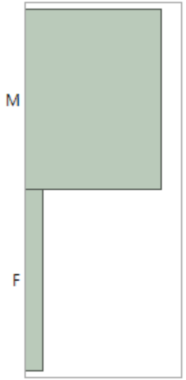
25% (n=11) relocated from native region
19% (n=8) relocated to US specifically
12% (n=5) relocated from Asia to US

Current Industry



	n	%
Software & Services	22	51%
Consulting	7	16%
Finance/Svcs/Ins	7	16%
Government / military	3	7%
Consumer products	2	5%
Academics / Research	1	2%
Telecom	1	2%
	43	**100%**

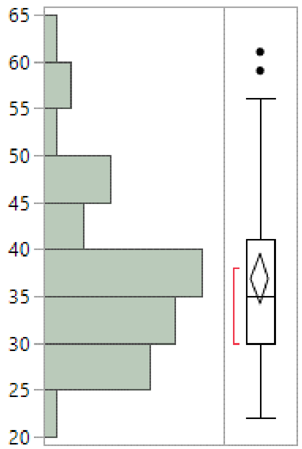
Gender



	n	%
Male	38	88%
Female	5	12%
	43	**100%**

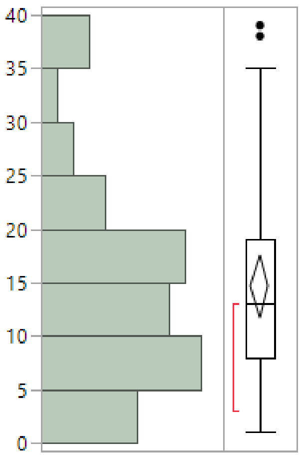
Demographic Profile (n=43)

Age*



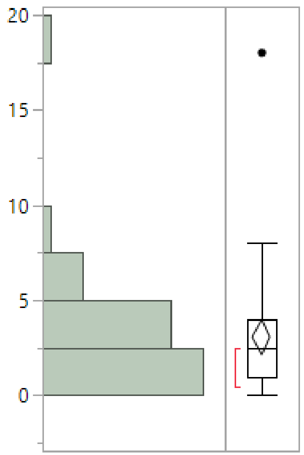
Mean	37
StdDev	9

Yrs Employed*



Mean	15
StdDev	10

Yrs CSDS*



Mean	3
StdDev	3

** Estimates inferred from LinkedIn profile data*

Interview Questions and Analysis

43 Cybersecurity Data Scientists (Dis-)Agree…

CSDS Practitioner Interview Research

Qualitative: Open Response 30 Minute 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 review

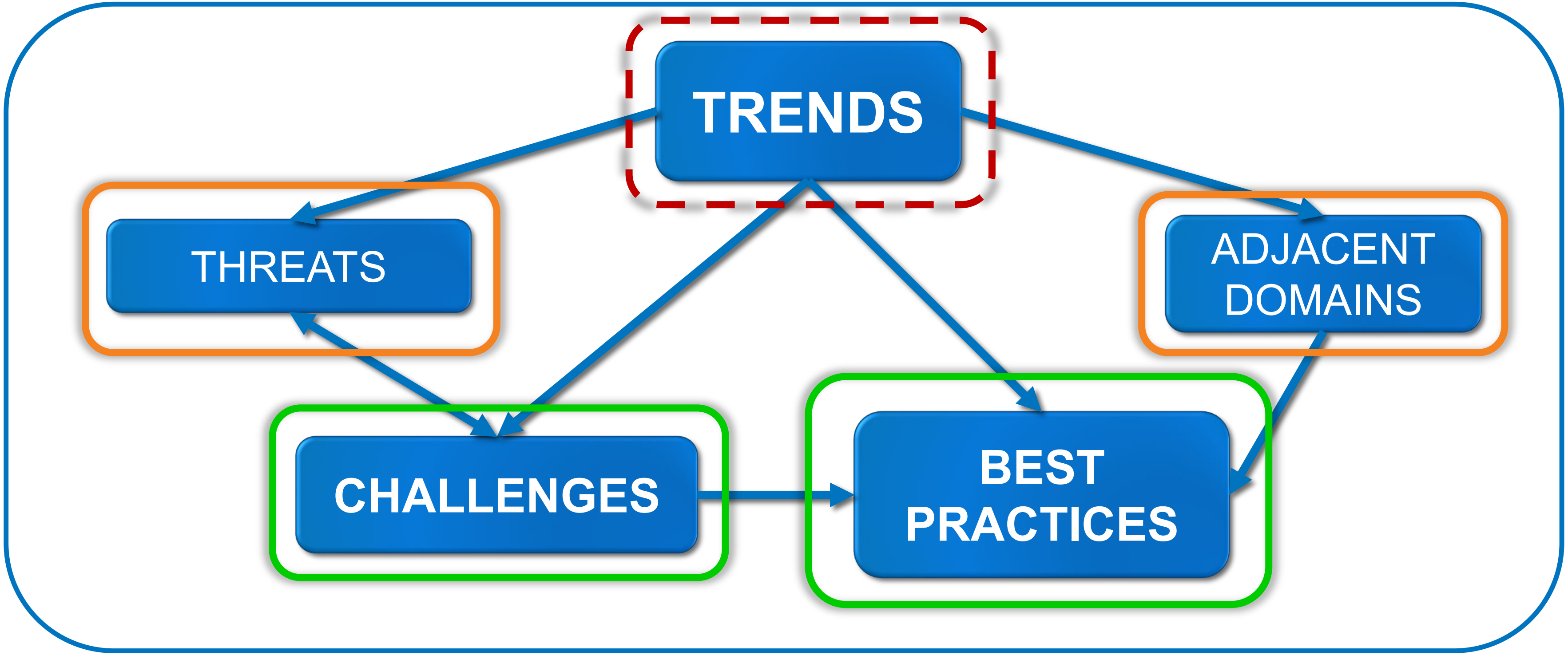
Practitioner review

Key topics (codes)

'Coding' of processed interview transcripts

CSDS Objectives - Conceptual Model for Responses

Framing and Relationships Amongst Topics



Threats & Adjacent Domains

CSDS Professional Perspectives

THREATS: 13 Adversarial Trends

- Internal threats
- Inherent vulnerabilities
- Reverse engineering detection
- Automated attacks increasing
- Exploiting new tech vectors
- Social engineering
- Ransomware-as-a-service
- Crypto-jacking
- Continual adaptation
- State actors => machine learning
- Time-to-detection / dwell time
- Industry-specific attacks
- Adversarial ML

White hat tools (i.e. PEN testing) often quickly end up being repurposed for black hat purposes…

Adversarial objectives evolve to optimize economic risk-reward

Much disagreement, from indignant disbelief to notion of manifest destiny

i.e. Reverse engineering and confusing / tricking ML models (seeding false data)… Although a 'hot topic' in academic research, few indications of incidents.

METHODS: 8 Influential Adjacent Domains

- Social & behavioral sciences
- Fraud / forensics / criminology
- Medical, epidemiological, ecological
- Enterprise risk management
- Network graph analytics
- NLP & semantic engineering
- Forecasting / time-series analysis
- Computer vision / deep learning

QUOTE: "It is almost a crime how little we learn from the fraud domain being as they have been at it for almost a century."

QUOTE: "As networks and devices become increasingly complex and intertwined, they begin to resemble organic systems and act in biological ways."

QUOTE: "Whereas cybersecurity seeks to safeguard, it isn't going to get very far without quantifying risks and impacts."

QUOTE: "Still a work in progress, and one does need to step over the hype, but there are some early indications that deep learning can be quite efficacious if one is handling immense amounts of labeled data."

CHALLENGES

Perceived CSDS Gaps

Challenges: 12 Topics

ORGANIZATION

- Confusion
- Marketing hype
- Regulatory uncertainty
- Few resources

PROCESS

- Inherent costs
- Decision uncertainty
- False alerts volume
- Scientific process?

TECHNOLOGY

- Data preparation / quality
- Own infrastructure & shadow IT?
- Normal vs. anomalous?
- Lack of labeled incidents

Challenges: 12 Topics => 5 Themes*

* Utilizing exploratory factor analysis (extraction of latent factors)

1. **Leadership has 'lost the plot'**
   - Uncertainty: nature of threats, what is being protected, how to react
2. **Can't do it all!**
   - Expansive domain: not cost effective to cover everything in house
3. **Between a rock and a hard place…**
   - Rules-based approaches failing, but alternate approaches overhyped
4. **Scientific contextualists**
   - Need to improve *representation of environment* & *tracking of events*
5. **Data cleansing: 'the ugly stepchild'**
   - Critical underinvestment in data engineering to stage analytics

Best Practices

Perceived CSDS Treatments

Best Practices: 26 Topics => 8 Themes*

* Utilizing exploratory factor analysis (extraction of latent factors)

ORGANIZATION

- Management-driven change
- Training & program governance

PROCESS

- Organizational process engineering
- Structured risk quantification
- Focused scientific processes

TECHNOLOGY

- Data engineering practices~
- Ontologies & normalization
- Architecture-driven solutions

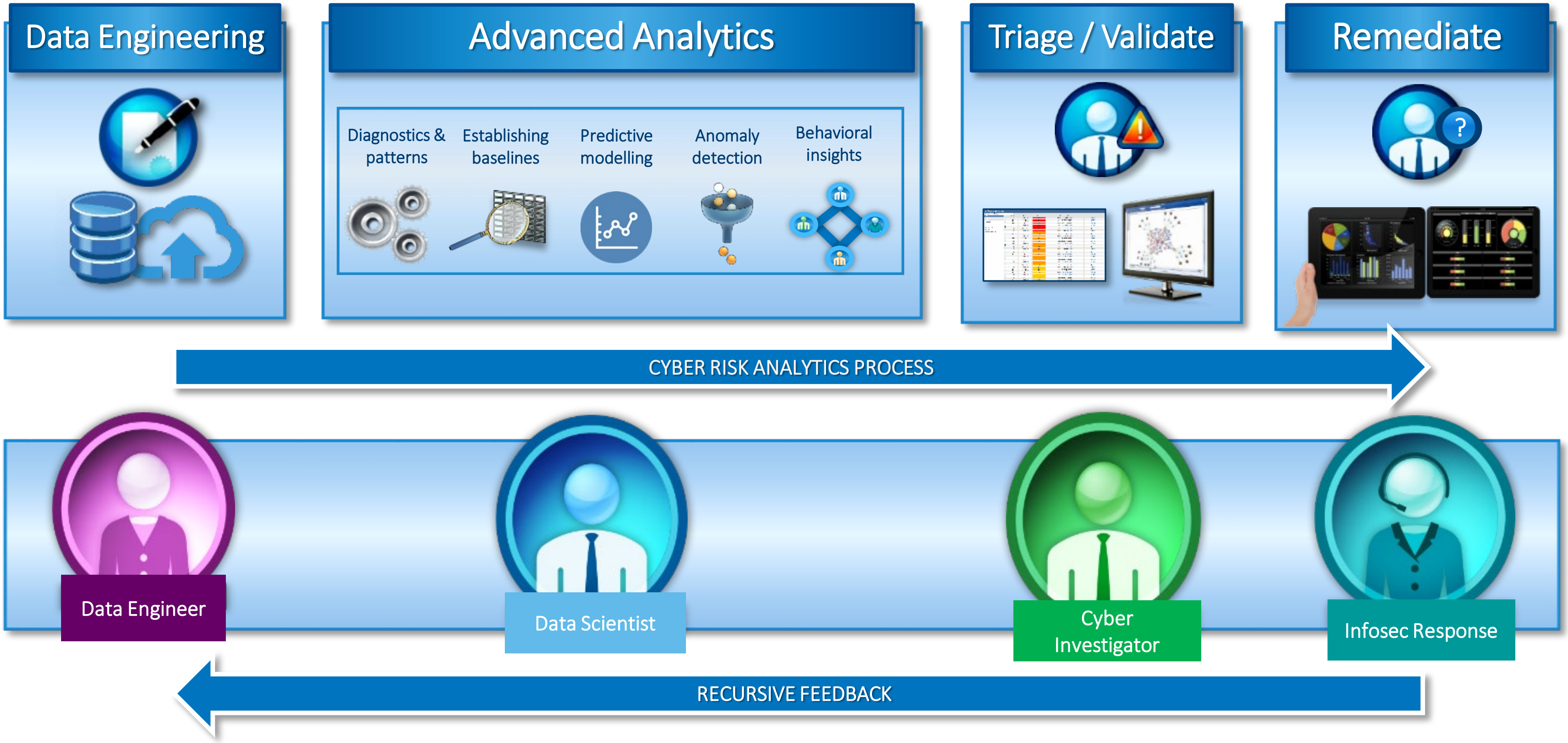
Key Guidance

CSDS Gap Prescriptions

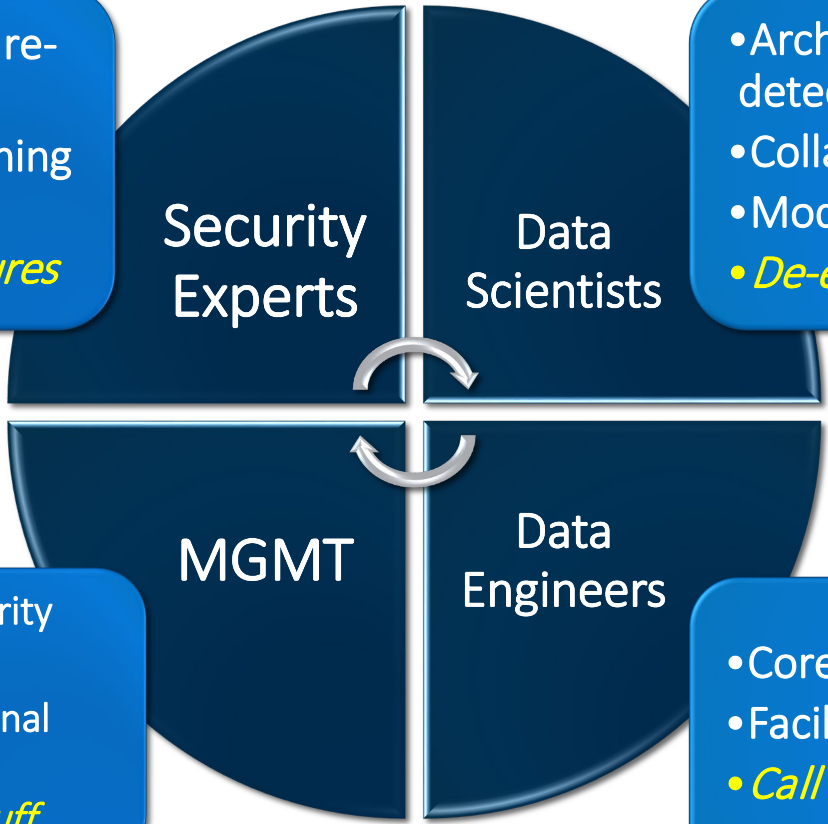
Key Prescribed Treatments: Correlation Between Factors

Challenge Themes	Best Practice Themes

Organization: Interdisciplinary Collaboration



Organization: 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*

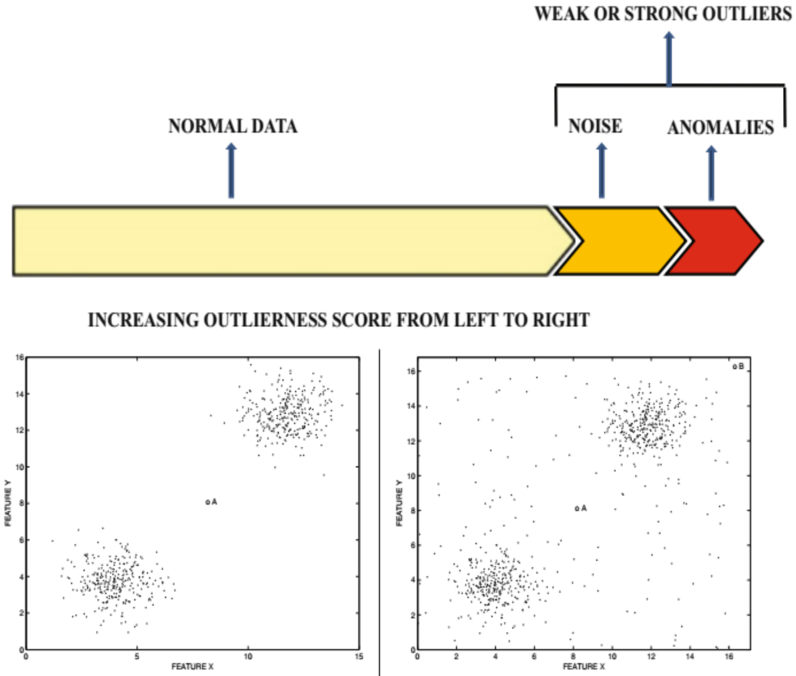
- Core data 'pipeline' processing
- Facilitate processes / quality
- *Call "data lake = strategy" bluff*

People - Process - Technology

Management of Information System

People: Anomaly Detection - Simply Complex

Identifying targeted anomalies amongst an ocean of noise…

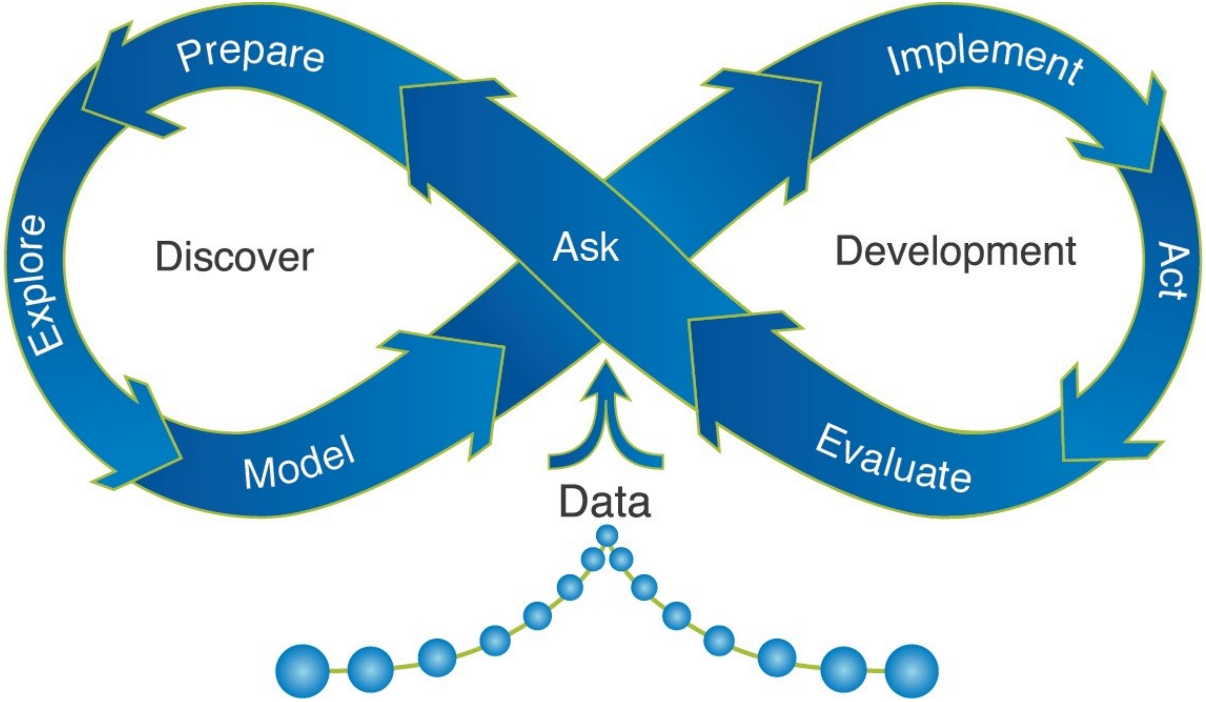


(a) No noise (b) With noise

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

Process: Analytics Life Cycle

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



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

Technology: Architect Exploratory & Detection Platforms*

Functional Architectural Segmentation

Exploratory 'big data' repository

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

Analytical models
- Descriptive
- Unsupervised

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

Operationally focused detection

Canonical ontology / schemas

Analytical models
- Statistical
- Supervised

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

Summary

Cybersecurity Data Science (CSDS)

- Process of Professionalization: a work in progress
  - Named professionals
  - Set of methods and techniques

Standards, best practices

Training programs

Certifications

Academic degree programs

Focused research journals

Formal sub-specialization

Scott Mongeau
Cybersecurity
Data Scientist

+31 68 370 3097
(Netherlands GMT+1)

Thank You!

Interested to participate?

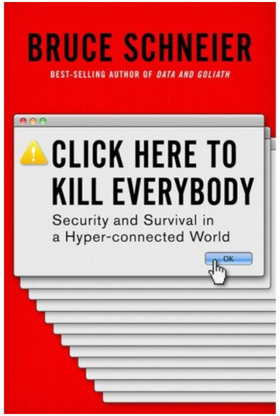
scott.mongeau@sas.com

REFERENCES

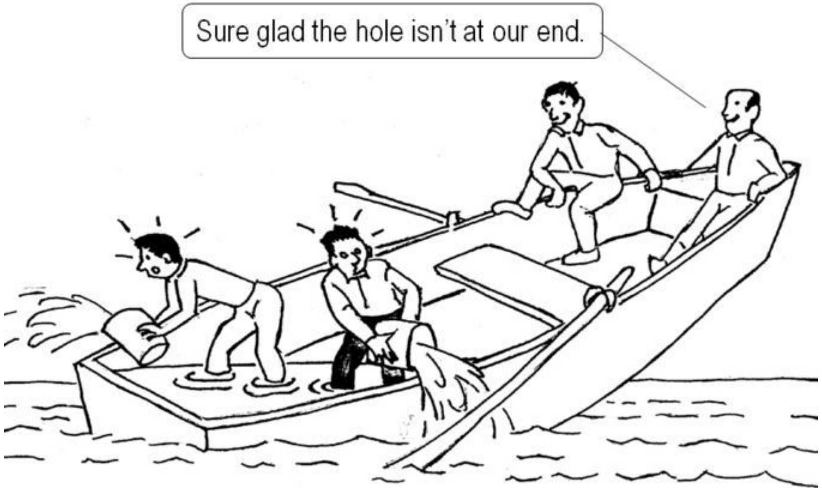
REFERENCES

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

APPENDIX



Organization: Building Disciplinary Bridges

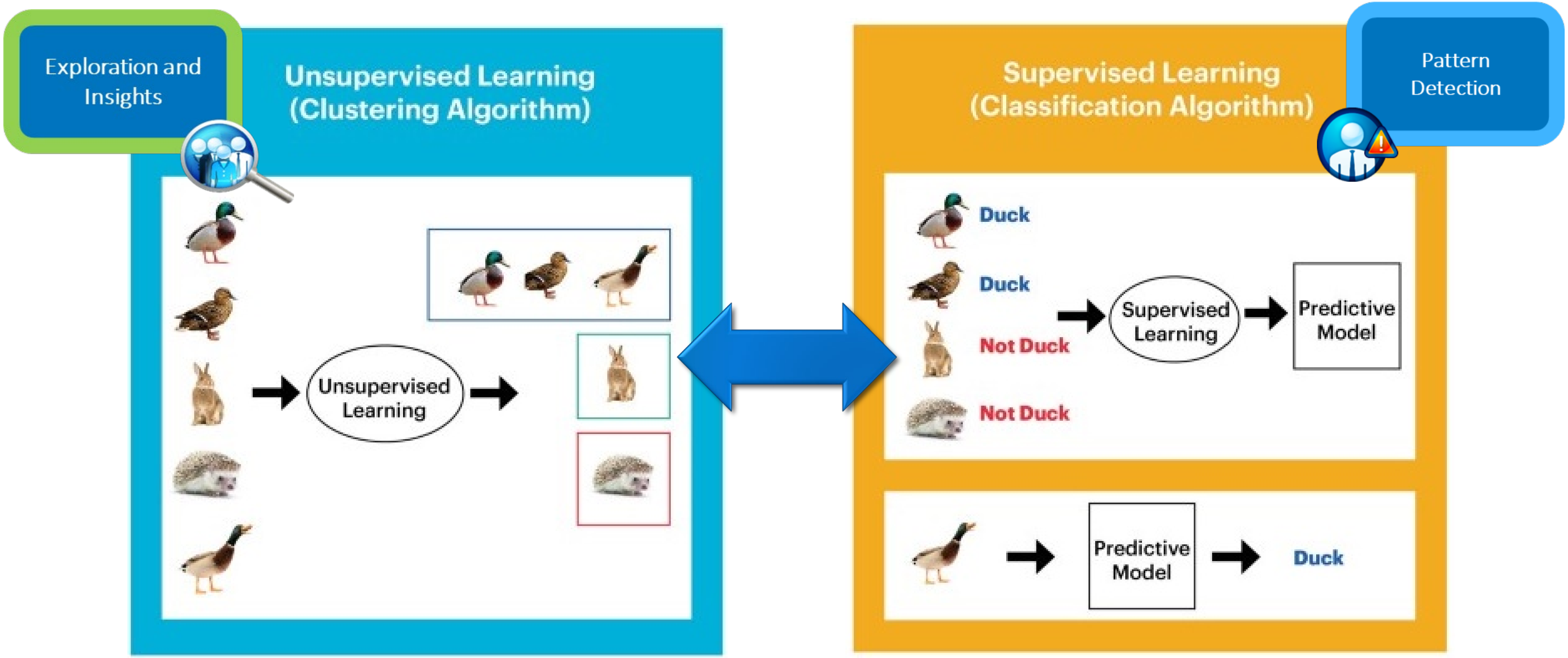


- Growing pressure/urgency
  - Cyber = general enterprise risk
- Structured processes
  - Meshing discovery, model building/ validation, alerting/remediation
- Data engineering as a process
  - Discovery / exploration
  - Detection / remediation

Key Prescribed Treatments: Correlation Between Factors

Challenge Themes (Factors)	Best Practice Themes (Factors)
1. Leadership has 'lost the plot'	• Management-driven change • Training & program governance
2. Can't do it all!	• Organizational process engineering • Focused scientific processes
3. Between a rock and a hard place… (limits of rules vs. hype)	• Architecture-driven solutions • Semantic frameworks
4. Scientific contextualists	• Training & program governance • Data engineering practices
5. Data cleansing: 'the ugly stepchild'	• Management-driven change • Training & program governance • Structured risk quantification • Focused scientific processes • Data engineering practices • Semantic frameworks

Process: Machine Learning Segmentation versus Classification



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

Cybersecurity Analytics Maturity Model

Anomaly Detection

- Big data overload
- Flags, rules, and alerts

Chasing phantom patterns

Data-aware Investigations

Understanding

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

Predictive Detection

Learning

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

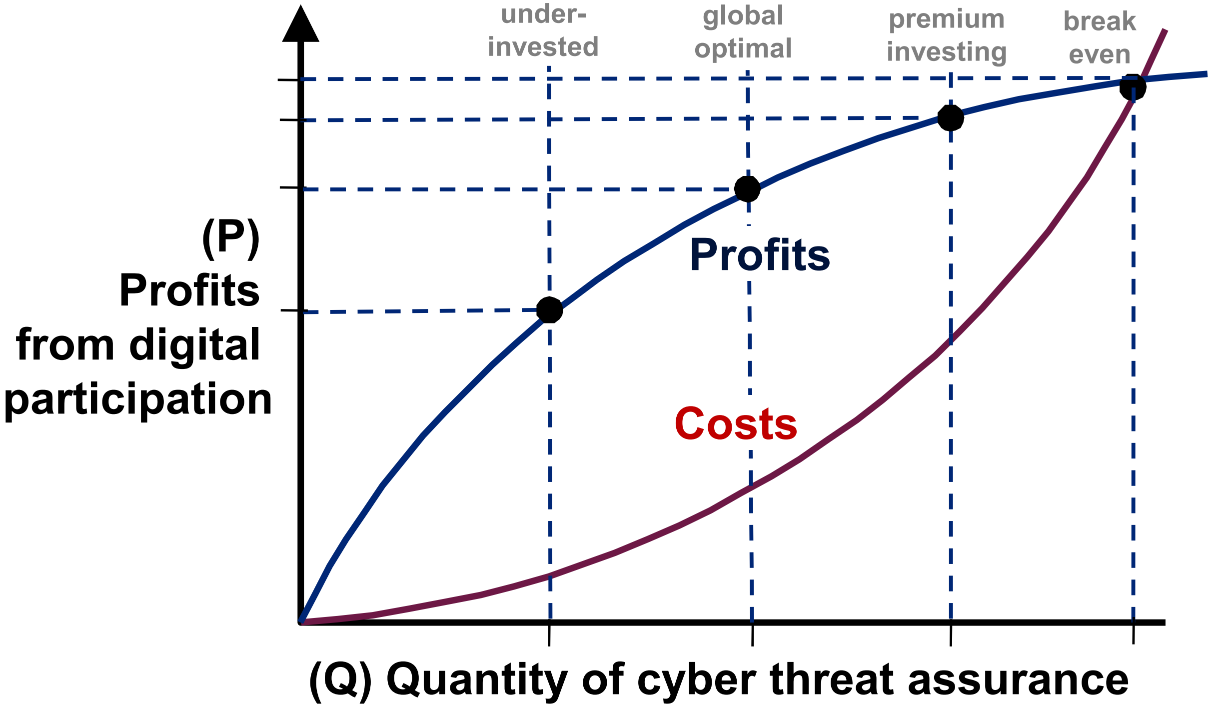
Risk Awareness / Resource Optimization

Risk Optimal

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

Cyber Defense Economics: Optimizing Accessibility Versus Exposure

Invest to point of optimality



SOURCE
Partnering for Cyber Resilience: Towards the Quantification of Cyber Threats
WEF report in collaboration with Deloitte:
http://www3.weforum.org/docs/WEFUSA_QuantificationofCyberThreats_Report2015.pdf

The ‘Meta Picture’ for Technologists and Methodologists

- **Cybersecurity**: hybrid techno-economic-behavioral context = many latent variables
- Research methodology
  - Multivariate inferential statistics
  - Social science: grounded theory (inductive)
  - Cross-applicability to ‘core’ cybersecurity?
  - e.g. Increase in complex multi-domain models?
- Extrapolating & validating patterns
  - *Content analysis / text analytics*
  - *Cluster Analysis*
  - *Principal Component Analysis (PCA)*
  - *Discriminant Analysis*
  - *Factor Analysis* => latent factors*
  - *Correspondence Analysis*
  - *Structural equation modeling (SEM)*
- Extrapolating latent behavioral indicators
  - i.e. User IT ‘*technical sophistication*’
  - ‘*Organizational importance*’ of a device
  - ‘*Adversarial determination*’
- Validating theoretical models