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
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”

COMPUTER HOLY WARS

HOLD IT RIGHT THERE, BUDDY.

THAT SCRUFFY BEARD... THOSE SUSPENDERS... THAT SMUG EXPRESSION...

YOU'RE ONE OF THOSE CONDESCENDING UNIX COMPUTER USERS!

HERE'S A NICKEL, KID. GET YOURSELF A BETTER COMPUTER.
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?’
2018: “THE DATA REVOLUTIONARIES”

ENERGY AND PERSISTENCE CONQUER ALL THINGS.
2018: 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?’

FloCon 2019 – New Orleans
• Deafening market / vendor buzz
• But, caveats abound: ‘Many are drowning in data lakes’
2019: Drowning in Data Lakes
2019: 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

CYBERSECURITY GOALS

DATA SCIENCE METHODS

CSDS objectives
- Data engineering
- Reduced data volumes
- Discovery & detection
- Automated models
- Targeted alerts
- Resource optimization

DATA SCIENCE METHODS
"Data scientists and practitioners can talk past each other."

- Rapid emergence
- Early stages of professionalization
- Affected by maturity of ‘data science’ more generally
Data Science in 30 Seconds...

- **Understanding Patterns**
- **Identifying Factors & Causes**
- **Data Engineering**
- **Business Intelligence**
- **Data Visualization**
- **Forecasting & Probabilities**
- **Optimizing Systems**
- **Network Context & Meaning**
- **Semantic Data**
- **Prescriptive**

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

Current Industry
<table>
<thead>
<tr>
<th>Industry</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>North America</td>
<td>27</td>
<td>63%</td>
</tr>
<tr>
<td>Western Europe</td>
<td>10</td>
<td>23%</td>
</tr>
<tr>
<td>Asia / Pacific</td>
<td>2</td>
<td>5%</td>
</tr>
<tr>
<td>Eastern Europe</td>
<td>2</td>
<td>5%</td>
</tr>
<tr>
<td>Middle East</td>
<td>1</td>
<td>2%</td>
</tr>
<tr>
<td>South America</td>
<td>1</td>
<td>2%</td>
</tr>
<tr>
<td>Total</td>
<td>43</td>
<td>100%</td>
</tr>
</tbody>
</table>

Gender
<table>
<thead>
<tr>
<th>Gender</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>38</td>
<td>88%</td>
</tr>
<tr>
<td>Female</td>
<td>5</td>
<td>12%</td>
</tr>
<tr>
<td>Total</td>
<td>43</td>
<td>100%</td>
</tr>
</tbody>
</table>

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

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

- **Domain literature review**
- **Practitioner review**
- **Key topics (codes)**
- ‘Coding’ of processed interview transcripts

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

**Domain literature review**

**Practitioner review**

**Key topics (codes)**

‘Coding’ of processed interview transcripts

---

**Topic extraction**

Agglomerative => multi-doc

**Concept clustering**

Divisive => unique doc

---

**Engine**

SAS Contextual Analysis

**Natural Language Processing (NLP)**

- Latent Semantic Indexing (LSI)
- Singular Value Decomposition (SVD)

---

**Domain literature review**

**Practitioner review**

**Key topics (codes)**

‘Coding’ of processed interview transcripts

---

**Concept clustering**

Divisive => unique doc
Threats & Adjacent Domains
CSDS Professional Perspectives
THREATS: 13 Adversarialial 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
- False alerts volume
- Decision uncertainty
- Scientific process?

**TECHNOLOGY**
- Data preparation / quality
- Normal vs. anomalous?
- Own infrastructure & shadow IT?
- 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

<table>
<thead>
<tr>
<th>Challenge Themes</th>
<th>Best Practice Themes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Leadership has 'lost the plot' • Management-driven change</td>
<td>• Training &amp; program governance</td>
</tr>
<tr>
<td>2. Can't do it all! • Organizational process engineering</td>
<td>• Focused scientific processes</td>
</tr>
<tr>
<td>3. Between a rock and a hard place… (limits of rules vs. hype)</td>
<td>• Architecture-driven solutions</td>
</tr>
<tr>
<td>4. Scientific contextualists • Training &amp; program governance</td>
<td>• Data engineering practices</td>
</tr>
<tr>
<td>5. Data cleansing: 'the ugly stepchild' • Management-driven change</td>
<td>• Management-driven change</td>
</tr>
<tr>
<td></td>
<td>• Training &amp; program governance</td>
</tr>
<tr>
<td></td>
<td>• Structured risk quantification</td>
</tr>
<tr>
<td></td>
<td>• Focused scientific processes</td>
</tr>
<tr>
<td></td>
<td>• Data engineering practices</td>
</tr>
<tr>
<td></td>
<td>• Ontologies &amp; normalization</td>
</tr>
</tbody>
</table>
Organization: Interdisciplinary Collaboration

Data Engineering

Advanced Analytics
- Diagnostics & patterns
- Establishing baselines
- Predictive modelling
- Anomaly detection
- Behavioral insights

Triage / Validate

Remediate

CYBER RISK ANALYTICS PROCESS

Data Engineer
Data Scientist
Cyber Investigator
Infosec Response

Organization: Interdisciplinary Collaboration

Cyber Investigator
Infosec Response
Organization: Interdisciplinary Collaboration

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

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

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

- Core data ‘pipeline’ processing
- Facilitate processes / quality
  - Call “data lake = strategy” bluff
People: Anomaly Detection - Simply Complex

Identifying targeted anomalies amongst an ocean of noise...

SOURCE
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

Operationally focused detection

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

- 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

[Image of healthcare professionals: Specialist, Researcher, Primary Care, Surgeon, Diagnosticians, Emergency Care]
Thank You!

Interested to participate?

scott.mongeau@sas.com
REFERENCES
REFERENCES


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

<table>
<thead>
<tr>
<th>Challenge Themes (Factors)</th>
<th>Best Practice Themes (Factors)</th>
</tr>
</thead>
</table>
| 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… | • Architecture-driven solutions  
| (limits of rules vs. hype) | • 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

Anomaly Detection
- Big data overload
- Flags, rules, and alerts

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

Cybersecurity Analytics Maturity Model

Chasing phantom patterns
Cyber Defense Economics: Optimizing Accessibility Versus Exposure

Invest to point of optimality

Profits from digital participation

Profits

Costs

under-invested
global optimal
premium investing
break even

(Q) Quantity of cyber threat assurance

(P)

SOURCE
Partnering for Cyber Resilience: Towards the Quantification of Cyber Threats
WEF report in collaboration with Deloitte:
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**