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

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

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

@SARK7 #CSDS2020 #FloCon2020
Research on cybersecurity data science (CSDS) as an emerging profession

I. Literature: What is CSDS and is it a profession?
II. Interviews: 50 CSDS practitioners
III. Designs: Approaches to address challenges
I. CSDS Literature
FUD  Fear, Uncertainty, Doubt

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

How quaint!

“Bad news, Your Majesty—it’s a cyberattack.”
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

CSDS objectives

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

DATA SCIENCE METHODS

CYBERSECURITY GOALS
CSDS: Existing Professionals + Demonstrated Efficacy

**EXAMPLE CSDS PRACTICAL APPLICATIONS**
- Spam filtering
- Phishing email detection
- Malware & virus detection
- Network monitoring
- Endpoint protection

*Survey of 621 global IT security practitioners*
### ‘Professional Maturity’ Comparison

<table>
<thead>
<tr>
<th>#</th>
<th>CRITERIA</th>
<th>CYBER</th>
<th>DS</th>
<th>CSDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Broad interest</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>2</td>
<td>People employed</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>3</td>
<td>Informal training</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>4</td>
<td>Informal groups</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>5</td>
<td>Professional literature</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>6</td>
<td>Research literature</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>7</td>
<td>Formal training</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>8</td>
<td>Formal prof. groups</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>9</td>
<td>Professional certificates</td>
<td>●</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>10</td>
<td>Standards bodies</td>
<td>●</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>11</td>
<td>Academic discipline</td>
<td>●</td>
<td>○</td>
<td>○</td>
</tr>
</tbody>
</table>

**CYBER** = Growing challenges + rapid paradigm shift

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

**CSDS** = At risk problem child?
# The Blessing and Curse of Data Science

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

**Cons**
- Hype & noise
- Befuddling array of approaches
- Lack of standards
- Myth of automation
- Big data ipso facto is not solution
- "Wait, what is the question?"
- “Throwing the statistical baby out with grampa’s bathwater?”
II. CSDS Interviews
CSDS Practitioner Interviews
30 minutes per interviewee

• ENTRY: How did you become involved in domain?
• What are perceived central CHALLENGES?
• What are key BEST PRACTICES?
Demographic Profile (n=50)
LinkedIn => 350 candidates => 50 participants

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

<table>
<thead>
<tr>
<th>Region</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>North America</td>
<td>35</td>
<td>70%</td>
</tr>
<tr>
<td>Western Europe</td>
<td>10</td>
<td>20%</td>
</tr>
<tr>
<td>Eastern Europe</td>
<td>2</td>
<td>4%</td>
</tr>
<tr>
<td>Middle East</td>
<td>2</td>
<td>4%</td>
</tr>
<tr>
<td>South America</td>
<td>1</td>
<td>2%</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Industry</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Software and services</td>
<td>28</td>
<td>56%</td>
</tr>
<tr>
<td>Consulting</td>
<td>7</td>
<td>14%</td>
</tr>
<tr>
<td>Finance/financial services/insurance</td>
<td>7</td>
<td>14%</td>
</tr>
<tr>
<td>Government / military</td>
<td>3</td>
<td>6%</td>
</tr>
<tr>
<td>Consumer products</td>
<td>2</td>
<td>4%</td>
</tr>
<tr>
<td>Academics / research</td>
<td>2</td>
<td>4%</td>
</tr>
<tr>
<td>Telecom</td>
<td>1</td>
<td>2%</td>
</tr>
</tbody>
</table>

Gender

<table>
<thead>
<tr>
<th>Gender</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>43</td>
<td>86%</td>
</tr>
<tr>
<td>Female</td>
<td>7</td>
<td>14%</td>
</tr>
</tbody>
</table>

Female
CSDS ‘CHALLENGES’: 11

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

- **DATA PREPARATION!** 84%
- **Marketing hype** 70%
- **Establishing context** 60%
- **Labeled incidents (evidence)** 56%
### RESPONSES: Advocated best practices

| 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% |

### Cross-domain collaboration 76%

- **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%)

### Scientific rigor 68%

- **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%)

**CSDS ‘BEST PRACTICES’: 26**

**DATA PREPARATION! 84%**

- **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%)

**Cross-domain collaboration 76%**

- **BP4**: Scientific method as a process (Proc, 34, 68%)

**Scientific rigor 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**

- **BP14**: Cloud and container-based tools and data storage
- **BP15**: Distinct exploration and detection architectures
- **BP16**: Participate in data sharing consortiums and initiatives
- **BP17**: Deriving probabilistic and risk models
- **BP18**: Upper management buy in and support
- **BP19**: Human-in-the-loop reinforcement
- **BP20**: Survey academic methods and techniques
- **BP21**: Cyber risk as general enterprise risk & reward
- **BP22**: Segment risk programmatically and outsource components
- **BP23**: Adding machine learning to SIEM
- **BP24**: Preventative threat intelligence
- **BP25**: Hosting and pushing detection to endpoints
- **BP26**: Honeypots to track and observe adversaries
### I. Data Management

<table>
<thead>
<tr>
<th>Challenge Factor Score (per respondent)</th>
<th>Rotated Factor Score (per respondent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CH F1 Expansive complexity</td>
<td>BP F1 Scientific process</td>
</tr>
<tr>
<td>CH F2 Tracking &amp; context</td>
<td>BP F2 Cross-domain collaboration</td>
</tr>
<tr>
<td>CH F3 Data management</td>
<td>BP F3 Risk management focus</td>
</tr>
<tr>
<td>CH F4 Expectations versus limitations</td>
<td>BP F4 Data-driven / data management</td>
</tr>
<tr>
<td>CH F5 Unclear ownership</td>
<td>BP F5 Focused tools</td>
</tr>
<tr>
<td>CH F6 Data policies</td>
<td>BP F6 Structured discovery process</td>
</tr>
</tbody>
</table>

### II. Scientific Processes

### III. Cross-Domain Collaboration

KEY CSDS GAPS: Factor-to-Factor Fitting

<table>
<thead>
<tr>
<th>Data context</th>
<th>Tracking &amp; complexity</th>
<th>Expansive</th>
<th>Ownership</th>
<th>Expectations</th>
<th>CH F4 Data</th>
<th>CH F3 Limited</th>
<th>CH F2 Tracking</th>
<th>CH F1 Expansive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation</td>
<td>CH F5 Unclear</td>
<td>CH F2</td>
<td>CH F1</td>
<td>CH F3</td>
<td>CH F4</td>
<td>CH F1</td>
<td>CH F6</td>
<td>CH F3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Factor to Factor Fitting</th>
<th>Factor Scores (per respondent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CH F1 Expansive complexity</td>
<td>-0.05401 0.09968 1.09968</td>
</tr>
<tr>
<td>CH F2 Tracking &amp; context</td>
<td>0.03401 2.09968 1.09968</td>
</tr>
<tr>
<td>CH F3 Data management</td>
<td>0.02401 1.09968 1.09968</td>
</tr>
<tr>
<td>CH F4 Expectations versus limits</td>
<td>0.01401 0.09968 1.09968</td>
</tr>
<tr>
<td>CH F5 Unclear ownership</td>
<td>0.00401 0.09968 1.09968</td>
</tr>
<tr>
<td>CH F6 Data policies</td>
<td>0.00001 0.09968 1.09968</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Challenge Factor Score (per respondent)</th>
<th>Rotated Factor Score (per respondent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CH F1 Expansive complexity</td>
<td>BP F1 Scientific process</td>
</tr>
<tr>
<td>CH F2 Tracking &amp; context</td>
<td>BP F2 Cross-domain collaboration</td>
</tr>
<tr>
<td>CH F3 Data management</td>
<td>BP F3 Risk management focus</td>
</tr>
<tr>
<td>CH F4 Expectations versus limitations</td>
<td>BP F4 Data-driven / data management</td>
</tr>
<tr>
<td>CH F5 Unclear ownership</td>
<td>BP F5 Focused tools</td>
</tr>
<tr>
<td>CH F6 Data policies</td>
<td>BP F6 Structured discovery process</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data context</th>
<th>Tracking &amp; complexity</th>
<th>Expansive</th>
<th>Ownership</th>
<th>Expectations</th>
<th>CH F4 Data</th>
<th>CH F3 Limited</th>
<th>CH F2 Tracking</th>
<th>CH F1 Expansive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation</td>
<td>CH F5 Unclear</td>
<td>CH F2</td>
<td>CH F1</td>
<td>CH F3</td>
<td>CH F4</td>
<td>CH F1</td>
<td>CH F6</td>
<td>CH F3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
III. CSDS Designs
Paradigmatic

Data management as a process

BP9: Data-driven paradigm shift away from rules & signatures

BP1: Structured data preparation, discovery, engineering process

CH1: Data preparation process (access, volume, integration, quality, transformation, selection)

CH4: Lack of labeled incidents to focus detection

BP13: Understand own infrastructure & environment

Context & tracking

BP F4

Data Management
Data Management: EDA Process + Feature Engineering

Featurization: Example - Graph Analytics
Feature Reduction: Example - Principal Component Analysis (PCA)
Exception Events

Exception messages per user (ranked)

Quantiles
- 100.0% maximum: 2559
- 99.5% maximum: 2559
- 97.5% quartile: 1889.725
- 90.0% quartile: 517.5
- 75.0% quartile: 172.75
- 50.0% median: 55.5
- 25.0% quartile: 9.75
- 10.0% minimum: 3.3
- 2.5% minimum: 1.825
- 0.5% minimum: 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
What is a User, anyway?

What is an IP address, anyway?

- Person
- Team
- Machine process
- Userld
  - Authentication Event
  - Auth event
  - DHCP
    - IP (or MAC address)
    - Authentication Event
  - External IP
    - Authentication Event**
  - Device / machine
  - Device / machine
  - BYOD
  - Session (e.g. application, HTTP(S))
  - APPS / AGENTS
Inferential Statistics

Observations

Population

Conclusions

Sample
Root Cause Analysis: Fishbone / Ishikawa Diagram

**Optimization**
- **CH7:** Traditional rules-based methods result in too many alerts
- **BP6:** Vulnerability, anomaly & decision automation to optimize operations

**Discovery**
- **BP15:** Distinct exploration and detection architectures
- **BP11:** Cyclical unsupervised and supervised machine learning
- **BP3:** Contextual nature of normal versus anomalous behavioral phenomenon

**Incident evidence**
- **CH4:** Lack of labeled incidents to focus detection
- **BP10:** Track and label incidents and exploits
- **BP19:** Human-in-the-loop reinforcement
- **BP8:** Model validation and transparency
- **BP4:** Scientific method as a process

**Contextual models**
- **BP17:** Deriving probabilistic and risk models
- **BP20:** Survey academic methods and techniques

**Quantification**

**Validation**

* 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

- **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

Research Methods for Cyber Security
Discovery ↔ Detection

SEGMENTATION

CATEGORIZATION

Unsupervised Learning (Clustering Algorithm)

Supervised Learning (Classification Algorithm)

Pattern Detection

Exploration and Insights

Unsupervised Learning

Supervised Learning

Not Duck

Predictive Model

Duck

Predictive Model
**Labels:** What constitutes ‘evidence’?

**Examples of Security Evidence**

<table>
<thead>
<tr>
<th>Inductive</th>
<th>Deductive</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Field evidence</td>
<td>- Rules &amp; signatures</td>
</tr>
<tr>
<td>- Probing &amp; testing</td>
<td>- Research &amp; threat</td>
</tr>
<tr>
<td>- 3rd party sourced</td>
<td>- intelligence</td>
</tr>
<tr>
<td>- Red Teaming</td>
<td>- Expert opinion</td>
</tr>
<tr>
<td>- Simulations</td>
<td>- Thought experiments</td>
</tr>
<tr>
<td>- Laboratory</td>
<td></td>
</tr>
</tbody>
</table>

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

- **Problem Framing**
- **Data Preparation**
- **Data Exploration**
- **Transform & Select**
- **Model Building**
- **Evaluate & Monitor Results**
- **Model Deployment**
- **Targeted Alerts**

*Exploration and Insights*
**Systematic evidence**

- **BP16**: Participate in data sharing consortia
- **CH4**: Lack of labeled incidents to focus detection

**Data management**

- **CH1**: Data preparation (access, integration, etc.)
- **BP7**: Data normalization, frameworks & ontologies
- **BP18**: Upper management buy-in and support
- **CH8**: Ownership, decision making & processes
- **CH9**: Resourcing, developing, hosting in house

**Uncertainty**

- **CH10**: Expanding breadth & complexity of domain
- **CH6**: Uncertainty leads to reactive stance
- **CH5**: Own infrastructure, shadow IT, exposure

**Management commitment**

**Resource coordination**

**BP F2**: Cross-domain collaboration

- **BP2**: Building process focused cross-functional team
- **BP3**: Cross-training team in DS, cyber, engineering
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
Continuous Detection Improvement Process

1. Exploration
2. Validation
3. Results

Patterns and anomalies
‘Real cases’ and ‘false alerts’
Continuous model refinement

Results

Exploration

Validation

Patterns and anomalies
‘Real cases’ and ‘false alerts’
Continuous model refinement
CSDS Model Development Process

Vindicate & Valorise
- Reproducibility
- Repeatability
- Interpretation
- Theory

DEPLOY

DATA

Develop & Verify
- Frame problem
- Assemble evidence
- Explanation & causation
- Feature engineering

DISCOVER

Calibrate & Validate
- Conceptual model
- Hypotheses
- Counterfactuals
- Falsification
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

- Surgeon
- Diagnostician
- Researcher
- Primary Care
- Emergency Care
APPENDIX
References


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