Application of Cybersecurity Data Science in Manufacturing Operations

Civil Aviation Cybersecurity Subcommittee

Stefan Schwindt – WG Chair (GE Aerospace)

Working Group Membership:

David Harvie	Embry-Riddle Aeronautical University (WG Chair)
Joe Reisinger	Astronautics Corporation of America
Kathleen Finke	Astronautics Corporation of America
Sean Crouse	Embry-Riddle Aeronautical University
Stefan Schwindt	GE Aerospace
Gabe Elkin	MIT Lincoln Laboratory
Jonathan Lee	MIT Lincoln Laboratory
Lily Lee	MIT Lincoln Laboratory
Daniel Stabile	MIT Lincoln Laboratory
Patrick Morrissey	Collins Aerospace
Alim Mohammad	Boeing
Taylor Lamb	Boeing
Chace Wilcoxson	Boeing
Tom McGoogan	Boeing (retired)
Greg Wellerman	Collins Aerospace

Contents

Figures

EXECUTIVE SUMMARY

This white paper examines how Cybersecurity Data Science (CSDS) can be applied in Manufacturing Operations (MO) within the aviation industry. It achieves this by utilizing the CSDS Aviation Architecture Framework (AAF) to identify unusual behavior and cyber threats in IT/OT environments that are interconnected. CSDS utilizes data-focused analytics, such as machine learning (ML), to analyze vast amounts of data from network traffic, system logs, and sensors, to answer three main questions:

Is a cyber event imminent?

Is there an active attack?

Was the event related to cyber?

The MO domain, which encompasses Industrial Control Systems (ICS) and assets such as CNC machines, presents unique challenges that differ from those in other domains, as it requires real-time operation and handles vast amounts of data.

The paper provides examples of how to utilize it, essential considerations for implementation, data lifecycle processes, challenges, and recommendations for enhancing cyber resilience without compromising operations.

Key Challenges

Manufacturing environments generate a significant amount of data, with two notable challenges that make it more difficult to identify threats: collecting irrelevant data adds noise and failing to capture essential data. Both cyberattacks and benign failures can cause process abnormalities, which makes it harder to link an anomaly to an attack. The CSDS Data Sphere model presents a way to address challenges that arise in acquiring, filtering, and analyzing data within the ICS cyber domain:

- Data is acquired but ignored or underutilized
- Data is desired but is unavailable

The expansion of the Data Sphere requires the addition of sensor data. However, additional data increases the challenges introduced in preparing the data for ML usage (labeling, synchronization, etc.) that consumes resources. Outside factors, such as environmental noise and non-cyber related events can further complicate the model training.

Use Cases and Lessons Learned

Three progressive use cases have been utilized to demonstrate application of CSDS:

• **Use Case #1**: Simulated a manufacturing (cube production) environment using synthesized data via Flexible Industry Relevant Environment (FIRE). ML models (Prophet, Long Short-Term Memory (LSTMs), Transformers) detected anomalies in quality metrics.

A key lesson of UC 1: Feature-rich, high-resolution data improves detection, while tunable thresholds balance false positives and lag.

• **Use Case #2**: UC2, built on top of UC1, simulated using contextual data and sophisticated attacks. Models put abnormalities into two groups: benign and malignant.

A key lesson of UC 2: Simpler models suffice for efficiency purposes, while high-density data boosts accuracy. However, rare events are hard to detect such that data preparation dominates effort.

• **Use Case #3**: Evaluate if real data from a motor setup with PLC, sensors, and acoustic input provide the same capabilities as a simulated setup. Custom ML models (transformers, LSTMs, Python Outlier Detection (PyOD)) were analyzed to extract torque profiles.

A key lesson of UC 3: Factory equipment requires modifications for ML-ready data and must prioritize the quality of the data. It is crucial to synchronize sensor data in a timely manner; however, acoustic sensors often struggle to distinguish between external noise right out of the box.

In general, ML is capable of finding threats that affect the CIA trinity (Confidentiality, Integrity, Availability) and safety but requires iterative improvements.

Recommendations

1. Invest in ML-Ready Infrastructure: When designing a system, make data collecting, synchronization, and labeling a top priority. For precision, connect sensors to PLCs.

2. Describe Threat Scenarios: Guidelines for sensor selection and requirements are used to distinguish between noise and abnormalities.

3. Make sensors more configurable: identify signatures in controlled situations and tailor features to reduce noise.

4. Use iterative processes: Use trial and error to check the Desired Data and simulate assaults to train your model.

5. Evaluate existing data sources in the system to determine the viability to correlate system data, infrastructure data, and manufacturing system data to detect or identify potential cyber events.

1 Introduction to CSDS in Manufacturing Operations (MO)

1.1 Intro to CSDS AAF

Cybersecurity Data Science (CSDS) is the application of data science to cybersecurity to generate actionable cyber-analytical insights from large and ever-increasing quantities of generated data. CSDS uses a data-focused approach to identify potential threats. Malware often behaves differently from normal network traffic or software processes. The problem is how to detect anomalies that may indicate a cyber attack. The CSDS Aviation Architecture Framework (AAF) applies CSDS to the complex, interconnected aviation ecosystem that includes manufacturing which has a mix of information technology (IT) and operation technology (OT) in distinct and specialized environments. In this research, the AAF was primarily used in the identification and placement of sensors.

1.2 Challenge of detecting anomaly/cyber event

The aviation ecosystem contains tremendous volumes of raw data in the form of network traffic, systems logs, and application logs generated through merged IT/OT systems. Within that ocean of raw data, the CSDS AAF describes how data can be acquired, pre-analyzed, collected, and further analyzed with the intent to answer the following three key questions:

1. Is there a cyber-event pending?
2. Is there an attack occurring now?
3. Was an incident/event caused by cyber activity?

Critical to answering those questions is acquiring the relevant data which is not trivial. Collecting too much irrelevant data unnecessarily slows down analysis and inhibits the production of meaningful results by introducing too much noise. Similarly, there may be CSDS relevant data that is not collected for analysis. Thus, the system can have both too much irrelevant data and not have the relevant data needed to answer those three questions. Answering these three questions can be difficult. For example, the third question is difficult to answer as attribution requires much more contextual information.

1.3 Need for CSDS AAF in MO

The Manufacturing Operation (MO) environment in the Aircraft Original Equipment Manufacturer (OEM) domain is where aviation equipment and parts are designed and produced. The MO environment utilizes many assets, including OT (e.g., Industrial Control Systems (ICS)), with computational and network capabilities which could make them potential targets for cyberattacks. Traditional IT-centric cybersecurity techniques of broad scope real-time data monitoring and collection are not suitable for the MO environment. The MO environment generates enormous data that becomes unfeasible to collect it all. Likewise, IT-based monitoring techniques (e.g., active scanning) can interfere with OT real-time requirements. The CSDS AAF in this domain seeks to detect anomalies in MO data and determine whether those anomalies are due to cyber activity or not, without interfering with MO requirements. MO Use Cases are necessary to experiment with and refine the CSDS AAF ability to detect anomalies in the manufacturing environment.

2 Data Science Implementation Considerations

Data science/analytics have been applied to many domains to achieve a variety of goals, including optimizing performances, detecting anomalies, and mitigating adverse impacts, all of which are relevant to the manufacturing operations environment. There are three key elements to consider in the implementation effort: the process of implementation, the data characteristics, and the analytics. We will capture the process of implementation and the data characteristics in this white paper. The analytics implementation is specific to the desired task goal and is not discussed here.

2.1 Data Science Implementation Process

The process to implement data science applications generally follows the steps:

1. Define the analytics goal
2. Define relevant data domain and collect data
3. Select and develop algorithms/analytics tools, including computation resources
4. Evaluate analytics on relevant data
5. Iterate algorithm selection and/or data collection until desired outcome is achieved
6. Deploy to application environment

Analytics can take the form of traditional statistics-based techniques or more modern techniques. Artificial intelligence/machine learning is an advanced data analytics tool that has shown promising results in many domains and is likely to be applicable to manufacturing domains. The demands on data for some advanced machine learning applications are often unique to ML and have not traditionally been collected and available. Considerations need to be made for investing in AI-ready data (further described in Section 2.2) collection capabilities to enable future applications of advanced analytics capabilities.

2.2 AI-Ready Data Characteristics

Modern data science and analytics have leaned significantly on application of AI/ML techniques. In order to successfully apply machine learning techniques to MO environment, the MO data needs to have a number of AI-ready characteristics:

1. Large quantity of data. The data sets need to be sufficiently large to reliably represent the nominal variabilities present in the MO environment.
2. Rich and relevant features. The data needs to contain features that are relevant to the analytics task.
3. Data labels. At least some of MO data need to be labeled with categories of interest to the analytics task. While not all machine learning algorithms require labeled data in their training phase, the evaluation phase of analytics development require truth labels.
4. Entity association. The data collection needs to contain elements that enable the association of features and data points to entities of interest. For example, all QA features on a part need to be associated with a specific part or batch of parts, and all measurements about machine performance and behavior need to be associated with a specific machine.
5. Data alignment. Often there are multiple sources of information available in a production system that pertain to a specific process. To enhance the usefulness of the data to capture information about the process, it is important to align the data sources, which often take the form of synchronization.

A large quantity of data from the operational MO environment is preferred over synthetic data. Where that is not possible, then synthetic data, either based on engineering knowledge or augmented from a small quantity of real data, can be sufficient. While it is possible to assign data labels post-hoc, after data collection, this potentially requires significant amount of effort. It is far more efficient to design the data synthesis/collection system to generate the data labels in situ during data collection/synthesis. Similarly, entity association is also far more efficiently generated/collected during the data synthesis/collection process.

Section 3 describes the 3 use case implementations under the CSDS MO project. AI-readiness has been integral in the 3 use cases. In the implementation of UC1 and UC2, MIT LL's synthesis engine produced data labels and entity associations during data generation. A significant amount of effort in the implementation of UC3 was devoted to the automation of data collection, data labeling and entity association. These data acquisition investments resulted in datasets that have a high level of AI-readiness and significantly reduced hurdles to applying machine learning techniques.

3 CSDS MO Use Cases

As a demonstration of data science applications to manufacturing operations technology domain, FAA initiated a progression of MO use cases to develop and broaden understanding of implementation considerations and benefits to MO cybersecurity missions. Each use case involves the following steps:

1. Define a use case of interest
2. Develop data acquisition capabilities
3. Develop data science/analytics capabilities
4. Evaluate performance of analytics on data based on the goal of the use case, with targeted exploration of trade space in data and analytics

3.1 Industry Interest in MO Cybersecurity

AIA members and ERAU engaged with MIT Lincoln Laboratory (MIT LL) in the definition of manufacturing process definition and threat scenario prioritization and made significant contributions through their expertise. A generic manufacturing process involving parts manufacturing and parts inspection was chosen as the experimental domain. AIA members prioritized three threat scenario impacts to manufacturers. These threats correspond to the Confidentiality, Integrity, and Availability (CIA) Triad as well as product, human, and environmental safety:

2. **Data exfiltration**. A threat actor may exfiltrate OT relevant data to monitor the manufacturing process and gain access to sensitive data or intellectual property (IP). This type of threat may be observable in potential impact to manufacturing throughput, in addition to direct detection of exfiltration signals. Data exfiltration is a concern for system confidentiality.
3. **Parts quality degradation**. A threat actor may tamper with the controls of manufacturing assets to result in increasing occurrences of defective parts. Tampering may happen at the machine code level so the programmable logic controller (PLC) program may be corrupted, or the machine code to a computer numerical control (CNC) machine may be corrupted. This type of threat is most immediately observable through sensor data collected in the quality inspection process. Parts quality degradation is a concern for system integrity
   Parts quality degradation could impact product, human, and/or environmental safety.
4. **Damage to manufacturing assets**. A threat actor may tamper with the controls of manufacturing assets to cause catastrophic failure or increasing wear and tear to reduce a system's lifespan. This type of threat may be observable in potential impact to manufacturing throughput and parts quality, in addition to direct detection of malicious instructions to manufacturing assets. Damage to manufacturing assets is a concern for system availability.
   Damage to manufacturing assets could impact human and/or environmental safety.

3.2 MO Use Case #1

Initial discussion with AIA members in defining MO use case 1 (UC#1) had identified a challenge in the lack of data appropriate for development and evaluation of analytics/data science applications. Specifically, techniques such as supervised machine learning require annotated data that label the events of interest. Additionally, regardless of techniques used, the need for annotated data is necessary for evaluation of any analytics capabilities. To address this shortcoming, MIT LL developed the capability, Flexible Industry Relevant Environment (FIRE), to synthesize OT environment data for the purpose of training machine learning models to detect anomalies or threat activities in an OT environment. Through discussions with industry participants, the resulting simulation scenario is an idealized manufacturing process that captures the events and data among a maker (which creates the cube and hole), an inspector (which collects quality assurance (QA) data), and a data logger (which stores QA and historical data), in the manufacturing of a simple cube. This simulation enabled experimentation with the capture of and analysis of data.

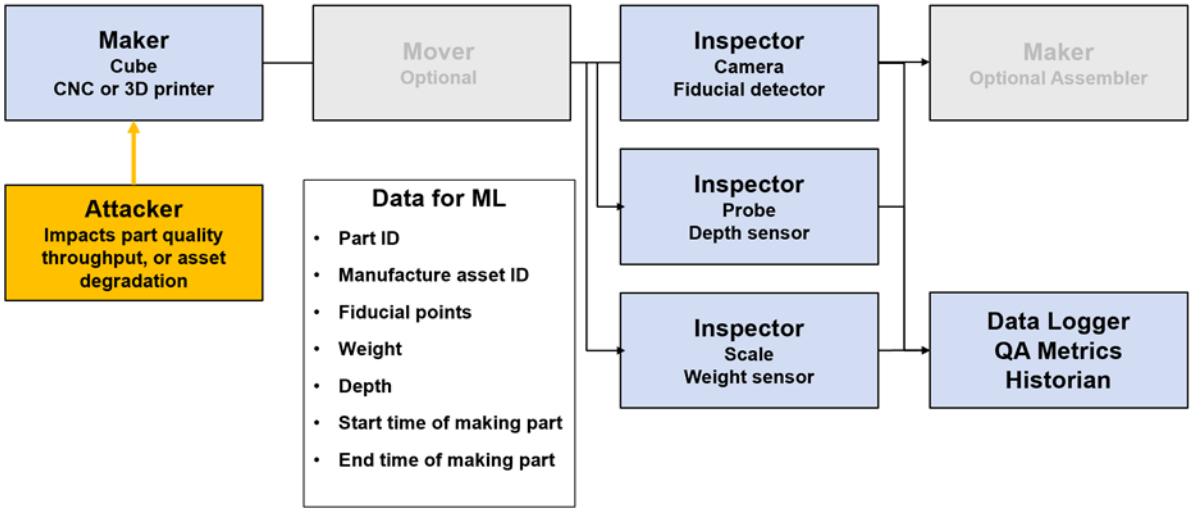

Figure 1: Idealized Manufacturing Process

AIA industry members desire the capabilities to detect anomalous events that may be indicative of errors in a manufacturing process or loss of IP, either due to inherent machine failures or cyberattacks. MIT LL's synthesis process generated normal data, associated with normal operations of a manufacturing system. Additionally, through discussion and agreement with AIA members, several anomaly scenarios were selected that would indicate a sudden, gradual, or sporadic change in parts quality metrics, or in system meta data. The synthesis of anomaly events was statistically generated based on distributions agreed upon by AIA members.

MIT LL implemented a number of analytics capabilities, ranging from classical statistics-based classification techniques to advanced machine learning techniques, such as transformers, and applied them to the synthetic data to evaluate their efficacy. All techniques were trained on normal data and applied to anomalous data for evaluation.

3.2.1 Lessons Learned

Three ML models – built from Prophet, Long Short-Term Memories (LSTMs), and Transformers – effectively detected anomalies in the UC#1 dataset. The detection capabilities of these models improved when utilizing feature-rich data sets. Also, increasing temporal resolution reduces lag in detecting anomalies.

The anomaly detection threshold can be tuned to balance the trade-off between the false positive rate and detection lag and missed detections. Anomaly detection model performance metrics can be used to optimize operations management. However, such optimization of operations is a tradeoff that has the potential to impact detection accuracy.

3.3 MO Use Case #2

MO Use Case #2 (UC#2) was developed to address the limitations identified in UC#1, which primarily concentrated on differentiating between cyber and non-cyber anomalies using synthesized QA data. Contextual data from operation, such as tool usage hours and environmental metrics, were integrated into the AI/ML pipelines. The rise in sophisticated OT-specific and Living Off the Land (LotL) style cyberattacks undermines traditional cybersecurity's perimeter defense. Also, it is impractical to perform broad-scope monitoring in IT/OT environments with limited resources.

The UC#2 Manufacturing Process Model (MPM) consists of a cube created by a maker who uses a computer numerical control (CNC) machine or 3D printer (see Figure 2). The maker device consists of a mill and drill system that produces a cube with holes, followed by autonomously bolting a plate to the cube. MIT-LL incorporated contextual data to develop data synthesis, FIRE V4, and three QA scenarios. The three QA scenarios were: 1) normal MO operations, 2) benign system failures, and 3) malicious attacks.

MIT LL experimented with multiple algorithms to assess their accuracy in detecting anomalous data. Once data is classified as anomalous, that data is passed on to a diagnostic classifier model for further analysis.

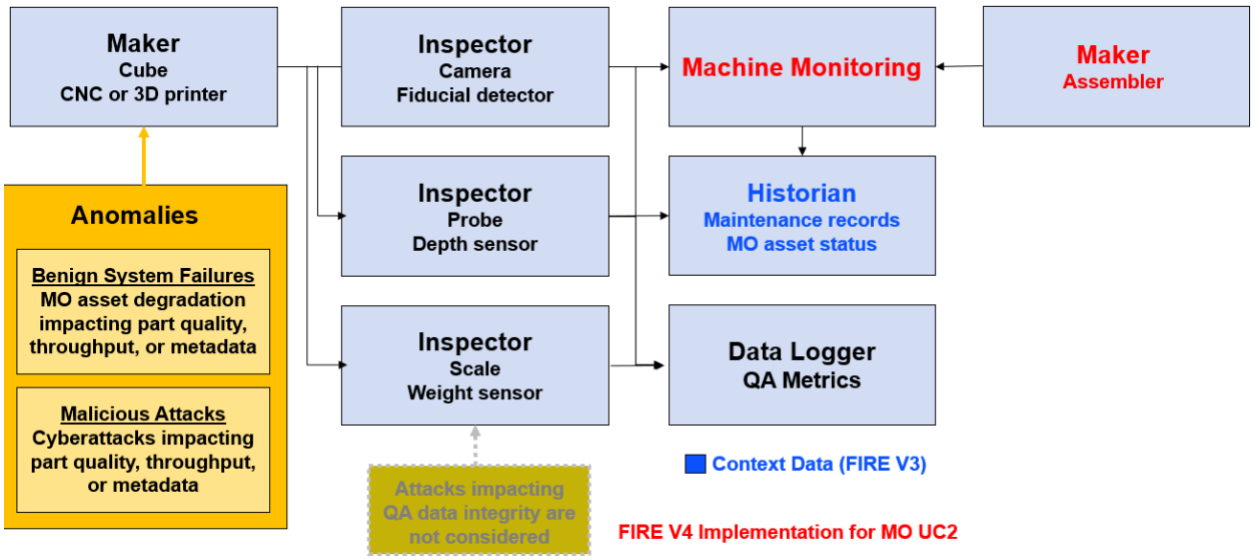

Figure 2: Manufacturing Process Model for MO UC#2

3.3.1 Lessons Learned

ML techniques effectively detected a wide range of anomalies in MO including benign system failures and malicious attacks. Simpler AI/ML models performed comparably to more complex models, but they have the benefit of requiring less computational resources. The ML model accuracy to detect anomalies improves when high-density, feature-rich data is utilized. It is still very difficult to detect rare, low-frequency events as there are insufficient examples to train the model with. A significant portion of resources, up to 80%, (CSDS Research Del #MO.2.4, 2025) of time is consumed in data preparation, including gathering, cleaning, and labeling datasets.

3.4 MO Use Case #3

Whereas prior iterations relied on synthetically generated data, MO UC#3 introduced the collection and use of real data. Partnering with Astronautics Corporation of America (Astronautics), an environment was developed by Astronautics to collect sensor data of a motor mimicking the profile of a torque wrench. Data from a PLC, the motor, a connected alternator, and an external audio sensor were collected. A significant amount of development time went into ensuring that the data was ready for ML which included synchronizing data streams and labeling with ground truth. The setup was fully automated and ran continuously for thousands of iterations to collect the necessary amount of data to train ML models. Malicious anomalies were introduced by instructing the torque-controlled motor to follow handcrafted torque profiles that deviated from normal.

Model development is currently underway, and analysis of model performance and the utility of different data features is to be done. The use of custom models built upon transformers and LSTMs will continue along with the introduction of algorithms from the python outlier detection library (PyOD).

3.4.1 Lessons Learned

Original factory equipment was not designed to deliver AI-ready data and required modifications to properly format and label for ML. With standard off-the-shelf anomaly detection algorithms readily available and ready to use, most of the development time was dedicated to building the infrastructure to collect, format, and label data with numerous iterations needed to identify and address issues in sensor outputs, alignments, and quality. Effective ML prioritizes feature quality over quantity through principled feature selection based on domain knowledge, rather than indiscriminate data collection.

Cyber-induced anomalies in a manufacturing process are challenging to distinguish from environmental noise. AI/ML models require labeled data to differentiate normal operations from cyber-induced anomalies. In the FIRE-M experiments, audio feature data collected from the audio sensor demonstrated the potential for detecting simulated torque manipulations. However, a lack of sound characterization of environmental noise versus anomalies hindered model training, revealing a critical barrier to AI/ML feasibility.

Data time synchronization is critical for AI/ML data correlation. Effective anomaly detection with AI/ML processing hinges on precise synchronization between the PLC and sensor data. As designed, the audio sensor could not be integrated into the PLC. An independent time synchronization function was required on the FIRE-M system to synchronize external sensor data and sensor data collected via the PLC. Although functional, the synchronization method could not achieve a fine enough resolution to distinguish environmental noise from the FIRE-M-induced system noise effectively. This misalignment risks obscuring subtle cyber-attack signatures.

4 Applying Cybersecurity Data Science (CSDS) to MO

4.1 Availability of Data

The following is a highly simplified summary of the typical steps from taking the design of a part, sending the design to a work cell, and resulting in a physical part:

- Design files are created and typically reside on a server in the enterprise IT or OT domains. The work cell pulls the design file from the designated server.
- Machine instructions are generated within the work cell based on design files or maybe resident on a server located in the Enterprise IT or OT domain.
- Machines, while creating the product as defined by the machine instructions, generate a varied data set related to machine function, which may or may not be collected.
- Work cells may also include technicians using tools to manually assemble parts from components.
- Quality assurance (QA) testing may be done in a specialized QA cell to determine whether to accept or reject the part based on design specifications.

Data is generated and used throughout product creation. It is preferred to determine the data requirements for the work cell prior to the work cell being built. This allows the OEM to direct the data to be collected, and the manufacturing entity would then need to design the work cell to create and store the required data. However, not all of this data is necessary or even useful to detect anomalies and potentially cyber events.

The CSDS Data Sphere, shown in Figure 3, can be applied to help identify what types of data are relevant to security objectives and the availability of that data. Most of the data acquired is ignored as it is neither feasible nor desired to collect all acquired data for analysis. Only the data that is collected is available for analysis. It is important to note that there is desired or required data that exists outside the Data Sphere (i.e., that data is never acquired). A more detailed description of the Data Sphere is in Appendix A: Data Sphere Detailed Description

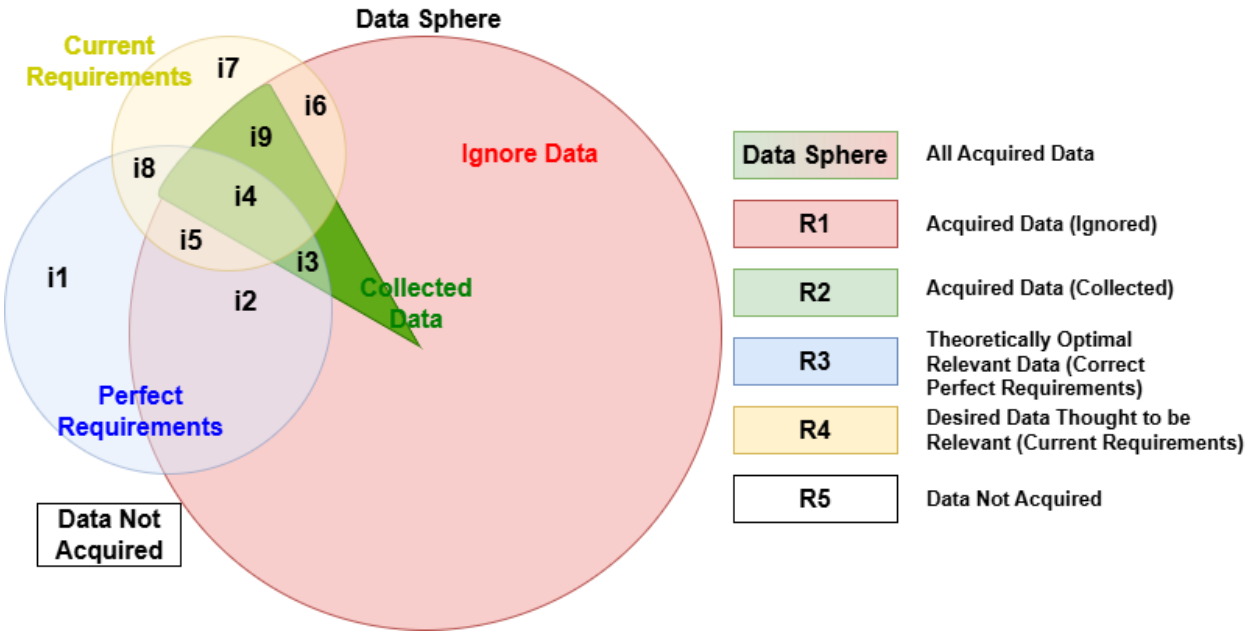

Figure 3. CSDS Data Sphere

4.2 AAF Process (Data Life Cycle/Engineering Process)

The CSDS Data Life Cycle consists of six (6) phases:

1. Acquire
2. Pre-analyze
3. Collect
4. Curate
5. Advanced Analytics
6. Information Sharing

These Data Life Cycle phases are shown at the top of Figure 4. Data Acquisition Sensors (DAS) monitor and capture data from hardware components and software processes during the Acquire Phase. All acquired data are illustrated by the R1 region (pink area) and the R2 region (green area) in Figure 3 constitute the Data Sphere. Most of that acquired data is not relevant to collect. During the Pre-analyze Phase, the Ignore Data illustrated by R1 is filtered out and the Available Data, illustrated by R2, remains. The Available Data is stored in non-volatile memory devices. The perfect set of data requirements is illustrated by R3. Required data identified by CSDS requirements, also known as Desired Data, is illustrated by R4. It is important to note that part of R4 (see Figure 3) is outside the Data Sphere. The portion outside of the Data Sphere represents data that is not acquired by any DAS. The intersection of the Desired Data and Available Data known as the Available Desired Data (see Figure 4) is what is available to the Curate Phase to extract cyber-relevant data prior to Advanced Analytics Phase. It is also important to note that the perfect set of requirements (R3) and current set of requirements (R4) do not directly align.

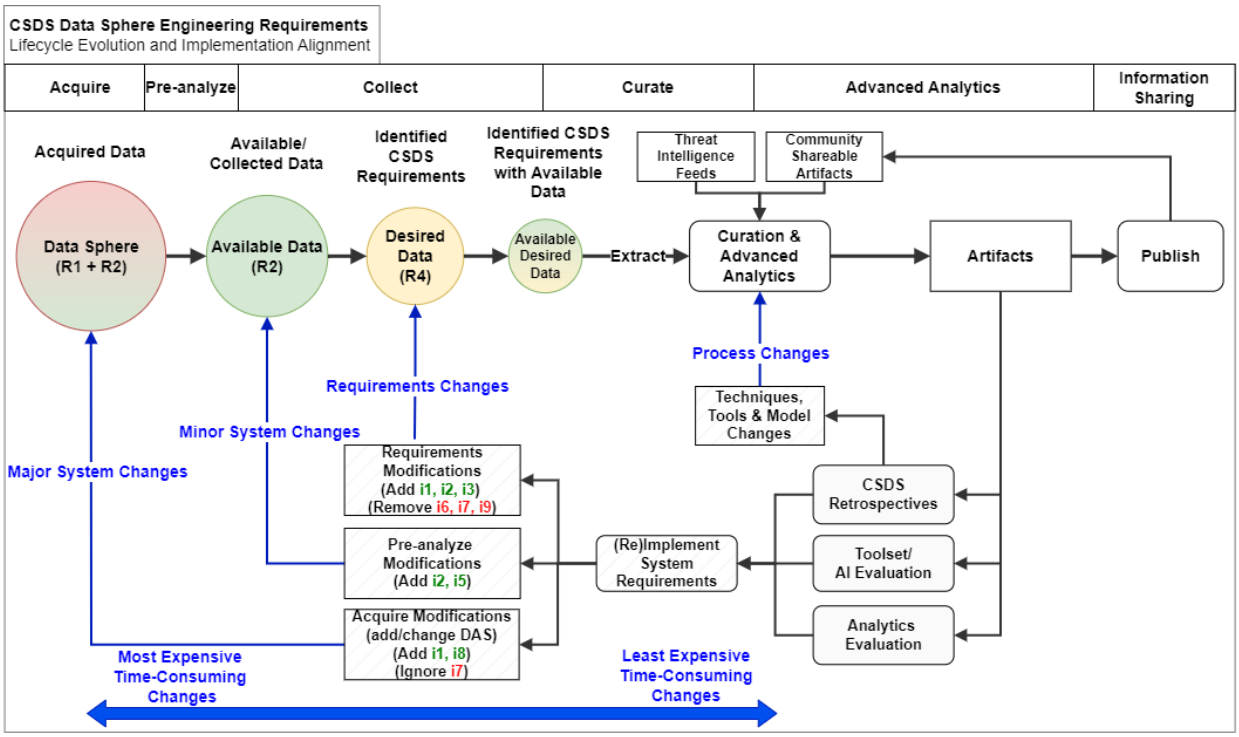

Figure 4. CSDS Data Sphere Engineering Requirements

4.3 Formatting of Data

The acquisition of the data is only part of the Data Life Cycle. Advanced analytics algorithms often require data to be curated and formatted in specific ways. Data from various sources must be synchronized and aligned, sensor drift needs to be corrected, and missing values filled in. For many ML algorithms, data needs to be numeric which necessitates numerical encodings of non-numeric fields. If using supervised learning, training data must be labelled, and regardless of the learning approach, any evaluation of a model's performance also requires labels indicating normal vs anomalous.

5 Challenges to Expanding the Data Sphere

5.1 Identifying Need for a New Data

Determining what should be Desired Data is difficult. Part of the Desired Data (i.e., data that is cyber relevant) may exist outside of the Data Sphere. In order to acquire this data outside of the Data Sphere, a DAS must be added to the system. As shown in Figure 4, adding a DAS is both expensive and time-consuming because it represents major system changes. However, acquiring the new data is just the beginning. It must be processed through the CSDS AAF Data Life Cycle in terms of filtering out the extraneous data, formatting and tagging the data for analytics, and then analyzing the data for actionable cyber analytical artifacts.

5.2 Practical Challenges with Expanding the Data Sphere with the addition of the acoustic sensor

During UC#3, it was hypothesized that an acoustic sensor could potentially detect anomalies in the operation of the motor. This hypothesis placed acoustic data within the region of Desired Data. However, the acoustic

data was outside the Data Sphere as there was no DAS equipped to acquire that type of data. A new DAS was introduced to the system.

The introduction of the acoustic sensor in the experimentation was not a panacea to detecting anomalies in the motor operation. Section 3.4.1 provides lessons learns from this experimentation. While the acoustic signals were useful for detecting anomalies, the torque and motor velocity data performed better. The introduction of the acoustic sensor illustrates an important point regarding the difficulty of ascertaining correct Desired Data. Determining the correct Desired Data requires trial and error.

6 Recommendations

The following are recommendations regarding applying CSDS to MO data.

1. Design production system with data acquisition in mind. Traditionally the design of production system had been focused on meeting the production goals. There is a significant amount of information from the production systems that can be used to evaluate the production quality and anomalies.
2. Design data acquisition systems to be AI-ready. Section 2 described the concept of AI-ready MO data. Capturing MO data with high levels of AI-readiness will reduce the burden of analytics development and deployment.
3. Design data acquisition systems with highly flexible sensor integration and configuration capabilities. A production system may be initially built with a certain collection of sensors, but hanging environment and conditions may precipitate the need for a different or expanded collection of sensors or configurations. A flexible sensor integration capability will reduce the burden of future updates.
4. Design data pipelines with interfaces for analytics insertions. MO environments often have high requirements on timely analytics. Realtime data from the production systems may be beneficial for timely analysis, as opposed to the data being archived and then processed at a later time.
5. Capture human and machine evaluations from the production process. Machine evaluation could be quality assurance results on parts or calibration on machines, and human evaluation could be human assessment on production quality or machine status. CSDS MO use cases leveraged a significant amount of labeled data to evaluate analytics performances, through synthetic data generation or a dedicated data collection system. This level of data collection may be challenging to accomplish in a production system. Sources such as machine or human evaluations can function as data labels, which are important for training and evaluating analytics. Collecting them from production systems will likely lower the cost of generating data and increase the relevance to the systems, and as a result, increase the applicability of analytics.
6. Develop data simulation capabilities. In situations where collection of real MO data presents a challenge to production systems, simulated data sets are viable supplemental sources to train machine learning analytics. Specific manufacturing processes and cyber-attack scenarios can be simulated more comprehensively with robust detection requirements defined upfront. Expert engineering knowledge is needed to produce high quality simulation data.

7 Abbreviations

Astronautics	Astronautics Corporation of America
AIA	Aerospace Industries Association
ARINC	Aeronautical Radio, Incorporated
CIA	Confidentiality, Integrity, and Availability
CNC	Computer Numerical Control
COTS	Commercial-Off-The-Shelf
CSDS	Cybersecurity Data Science
CVE	Common Vulnerabilities and Exposures

CVSS	Common Vulnerability Scoring System
ERAU	Embry-Riddle Aeronautical University
FIRE-M	Flexible Industry Relevant Environment - Manufacturing
IP	Intellectual Property
IUEI	Intentional Unauthorized Electronic Interaction
LotL	Living Off the Land
LRU	Line Replaceable Unit
LSTM	Long Short-Term Memory
MIT LL	MIT Lincoln Laboratory
ML	Machine Learning
MO	Manufacturing Operation
MPM	Manufacturing Process Model
NVD	National Vulnerability Database
OEM	Original Equipment Manufacturer
PLC	Programmable Logic Controller
QA	Quality Assurance
SBOM	Software Bill of Materials
UC	Use Case

8 List of References

Reference	**Title**
AIA SW Bill of Materials Report	Civil Aviation Cybersecurity Recommendations on the Use of Software Bill of Materials in Aviation
RTCA DO-326B (*equivalent to EUROCAE ED-202B*)	Airworthiness Security Process Specification
RTCA DO-356A (*equivalent to EUROCAE ED-203A*)	Airworthiness Security Methods and Considerations
RTCA DO-392 (*equivalent to EUROCAE ED-206*)	Information Security Event Management
NIST SP 800-82 Revision 3	Guide to Operational Technology (OT) Security
CSDS AAF	Cyber Security Data Science (CSDS) Aviation Architecture Framework (AAF), Parts 1-4
CSDS Research Del #MO.2.4	Findings and Recommendations Report – Final Draft – Use Case #2 (March 17, 2025)

Appendix A: Data Sphere Detailed Description

Data relevancy is a critical concept in CSDS because not all data is useful or helpful. Also, there may be gaps between the data currently being collected and what should be collected. From a CSDS perspective, data is relevant if it can help answer one of the three primary questions:

1. Is there a cyber-event pending?
2. Is there an attack occurring now?
3. Was an incident/event caused by cyber activity?

The Data Sphere (see Figure 5) is defined as the set of all data that is acquired. Only a small portion of the originally acquired data is collected into the Data-Store and is available for analysis. If the portion of the Data Sphere that serves as an input to the CSDS process is incorrect or insufficient, then the following issues may happen:

- Analytical toolsets are slowed down processing unnecessary data.
- Analytical toolset results are skewed leading to biased or faulty conclusions.
- Unnecessary noise and data prevent the production of meaningful results.
- Unnecessary or excessive configuration of network monitoring devices and IIS may stress the networks and system.

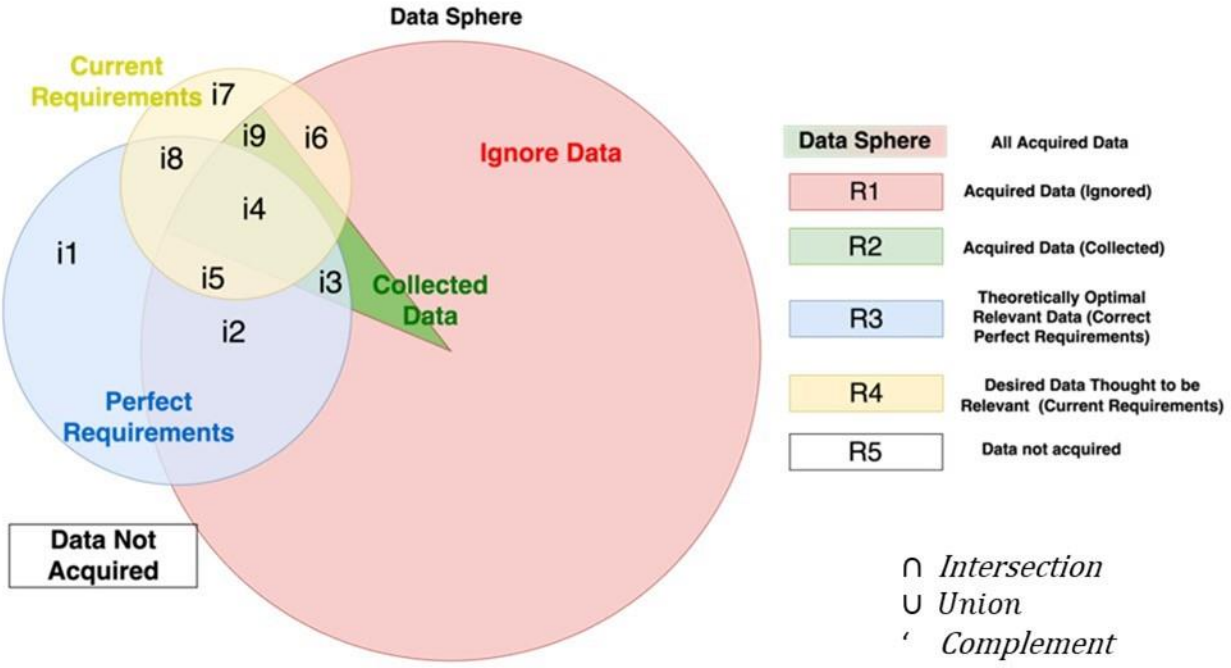

Figure 5. Data Sphere, Regions, and Intersections

Determining what is relevant data for CSDS is difficult. Figure 5 shows the Data Sphere with five regions and nine intersections which are described below:

1. The Data Sphere [R1 ∪ R2] represents all acquired data within a given Stakeholder's Environment of Operation from IIS. For data to be acquired from these systems, a Data Acquisition Sensor must be integrated into the IIS. As more Data Acquisition Sensors are integrated into a Domain Stakeholder's Environment of Operation, the Data Sphere will expand accordingly.

2. R1 represents all data that has been Acquired by Data Acquisition Sensors but ignored due to certain Pre-Analysis logic. R1 will grow or shrink in size as re-configuration to pre-analyzers occur.
3. R2 represents all data that has been Acquired and Collected due to certain pre-analyzer logic that deems it “potentially relevant data”. This data is said to be “Available” for CSDS and will require the CAC to correctly extract the available data from the Data Store based on the 4 Extraction Modes discussed. R2 will grow or shrink in size as re-configuration to pre-analyzers occurs.
4. R3 is a theoretical region representing the “Optimal data” for a specific CSDS Use Case. R3 indicates the correct CSDS data requirements for a particular Use Case (i.e., what data must be collected). It is important to understand that the Optimal data for CSDS is not always obvious, known at the start of a CSDS effort, or may not be possible to discover. Via adjustments to the yellow region, it typically takes multiple iterations of trial and error by Data Scientists and Human Analysts to determine the Optimal Data.
5. R4 represents data that the CAC desires to extract and curate for a specific CSDS Use Case (e.g., Malware Detection, Intrusion Detection, LMD, Spam Filtering, DDoS Detection, etc.). This represents the data that is part of the current data requirement for doing CSDS, but these requirements have not yet been validated as being part of the theoretically optimal R3. Note that just because a CAC considers the data requirements to be optimal and therefore desires this data, it does not mean the data is actually optimal (i.e., the requirements have not yet been proven to be valid).
6. i1 [R3 ∩ "Data Sphere"'] represents an intersecting region in which the Optimal data has not yet been acquired. To get this data, changes to the IIS configuration are required to acquire, pre-analyze, and collect this. This can often be an expensive and time-consuming process to accomplish.
7. i2 [R1 ∩ R3] represents an intersecting region in which the Optimal data is being Acquired but has been ignored due to faulty/incorrect pre-analyze logic. To address this problem, the Pre-Analyzer in the Data Acquisition Sensors must be reconfigured to account for this new data.
8. i3 [R2 ∩ R3] represents an intersecting region in which the Optimal data is being collected, but the CAC has not yet recognized/identified the data as relevant to the specific CSDS Use Case. Fixing this requires a process-driven scientific approach by Data Scientists to assist them in recognizing that available data is missing from the analysis. For near-term CSDS Use Cases, this represents the Optimal Data Set that is most useful to current efforts.
9. i4 [R2 ∩ R3 ∩ R4] represents an intersecting region in which the theoretically optimal data has been identified for a specific CSDS Use Case and is actively being collected/extracted by the CAC for conducting advanced analytics. This is the best outcome.
10. i5 [R3 ∩ R4 ∩ "Data Sphere" ∩R2'] represents an intersecting region in which the CORRECT data has been identified to be relevant for a specific CSDS Use Case but is currently not being collected due to faulty/incorrect pre-analyze logic.
11. i6 [R3' ∩ R4 ∩ "Data Sphere" ∩R2'] represents an intersecting region in which data that has been recognized/identified as relevant for a specific CSDS Use Case is not the CORRECT data to be used. To fix this, requires a process-driven scientific approach by Data Scientists to assist them in recognizing that some data being analyzed is not needed and should be removed.
12. i7 [R4 ∩ "Data Sphere"' ∩ R3'] represents an intersecting region in which data that has been recognized/identified as relevant for a specific CSDS Use case is not being acquired. i7 is detrimental to CSDS efforts and the business as expensive and time-consuming changes to systems will be made to acquire NEW data that is not optimal for the specific CSDS Use Case. (i.e., wrong requirements).
13. i8 [R4 ∩ "Data Sphere"' ∩ R3] represents data that is not being acquired and has correctly been identified as a CSDS requirement.
14. i9 [R2 ∩ R3' ∩ R4] represents data that is being collected and has incorrectly been identified as a CSDS requirement.

The full documentation for the CSDS AAF is available for download at https://erau.edu/research/industry-collaboration/center-for-aerospace-resilient-systems/csds