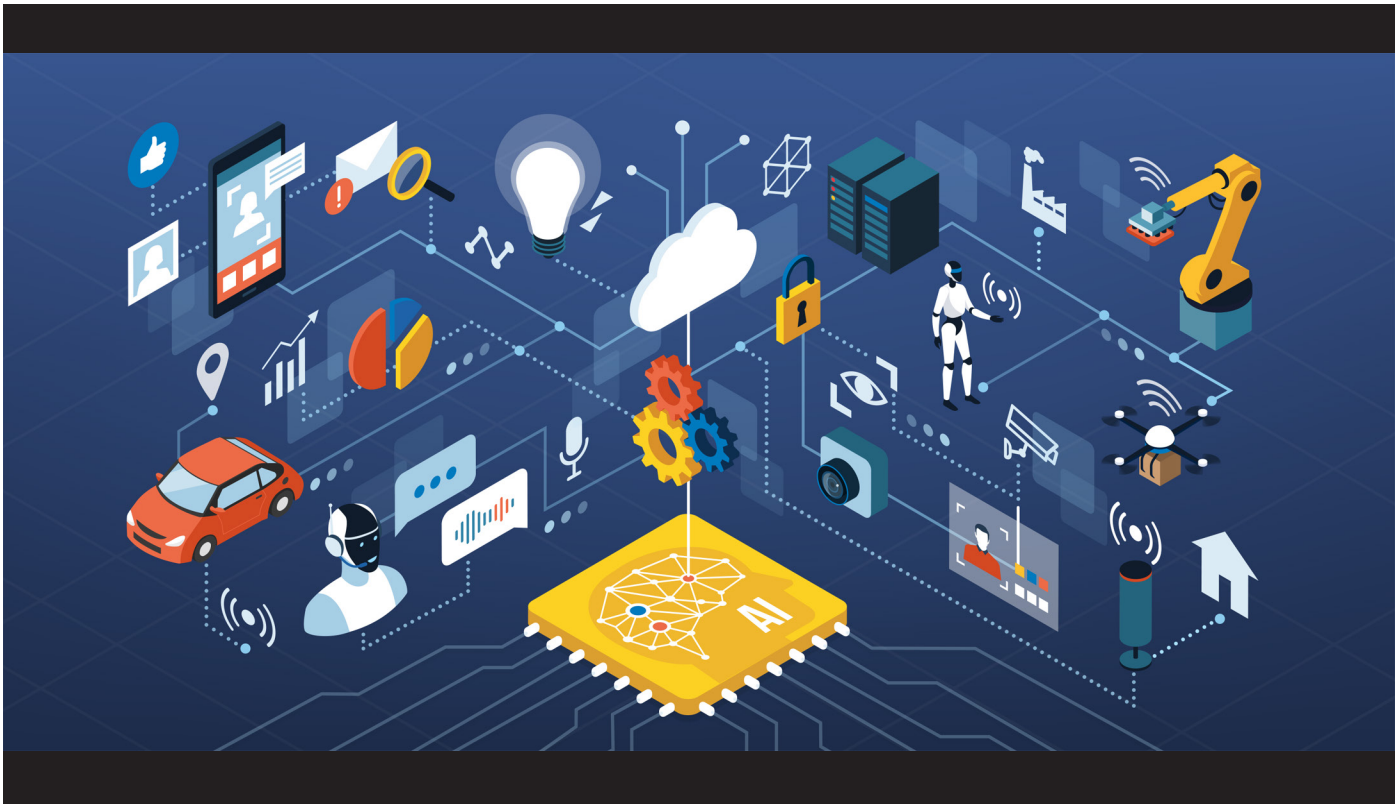
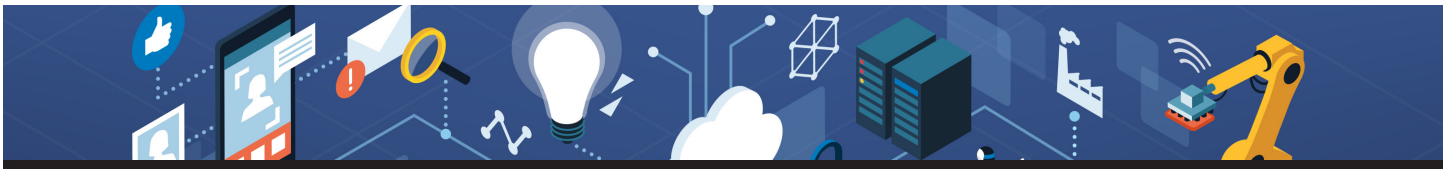
EPRI | ELECTRIC POWER RESEARCH INSTITUTE

DATA FOUNDATIONS FOR OPERATIONS TECHNOLOGY CYBER SECURITY ANALYTICS, ARTIFICIAL INTELLIGENCE, AND OTHER DATA-INTENSIVE APPLICATIONS



November 2019



Executive Summary

Analytics, artificial intelligence (AI), and other data-intensive applications have significant value for utility operations technology (OT) cyber security use cases. AI techniques will form the basis for automated threat mitigation, an emerging technology that has potential to help utilities more quickly detect advanced persistent threats and insider threats. But the successful use of these capabilities to achieve the most impactful outcomes requires proper data preparation.

Utilities are advised to first ensure that:

- *Data management and governance plans are updated to accommodate OT cyber security data needs.*
- *Computing and communications network infrastructures can support OT cyber security use cases that rely on predictive and prescriptive analytics.*

Table of Contents

This white paper was prepared by EPRI.

Introduction: The Promise of Cyber Security Data

The value of data is defined by how its variety, velocity, volume, and veracity are leveraged to deliver actionable information and support human or computer-automated actions. As traditional utility grids transform into advanced grids, the volume, velocity, and variety of data dramatically increase. The vast difference in volume between traditional meter and smart meter data is the most well known example, but any advanced grid deployment may offer new and different types of data to help monitor and manage utility operations. Data growth means more opportunities and greater challenges for utilities.

An increasing number of utilities are using analytics capabilities in their integrated security operations centers (ISOCs). Recent Enterprise Strategy Group research into security analytics across all business sectors reported that 77% of all organizations were gathering, managing, and analyzing "significantly more" or "somewhat more" security data than they did two years ago. The security data sources include their internal networks, applications, and endpoints, plus external threat intelligence and all as-a-service activity. Having real-time situational awareness of these data streams is essential to help utilities identify vulnerabilities, audit operations, and respond to threats. It is particularly critical given the proliferation and integration of sensors, smart inverters, and other devices that help utilities remotely monitor and manage their generation plants, distribution, and transmission grids. Anything connected to a utility communications network is a potential attack vector for cyber incursions.

However, all that collected data also creates opportunities for utility cyber security teams.

The data generated by these devices holds potential for improved situational intelligence and enhanced monitoring of utility devices, applications, and networks to detect and respond to physical and cyber threats. But the increasing volume and velocity of this data overwhelm human abilities to assimilate and act on it. The promise of all this data hinges on artificial intelligence (AI)[1] and AI-based applications to process enormous volumes of bits and bytes in real time, to detect patterns, to identify anomalies, and thus to react and alert appropriate staff to threats. These capabilities are the foundation for automated threat mitigation (ATM) technology, which will analyze large volumes of security data from a variety of sources at high velocity to automatically identify cyber threats, respond to attacks, and potentially resolve vulnerabilities.

ATM technology is not yet commercially available, but that gives utilities sufficient time to prepare their data for it. This proactive data preparation may also help utilities identify potential use cases, improve project scopes, and accelerate returns on investment for advanced cyber security analytics and AI applications.

1 Machine learning, neural networks, and other computer learning techniques are types of AI. The use of the term AI in this document includes all such techniques.

Data Fundamentals

Utilities can prepare for advanced operations technology (OT) cyber security that relies on data by understanding and acting on two principles:

- The immutable data condition is still "garbage in, garbage out." Bad data leads to bad information and decisions.[2]
- Good data is data that has been validated, standardized, and normalized. In other words, raw data must undergo some preparation to be usable.

These principles apply to data analytics projects. According to the SAS Institute, 40% of all analytics projects fail because of insufficient or inadequate data preparation. The institute also notes that about 80% of the time spent on any data analytics project is dedicated to data modeling and management. These are crucially important statistics to keep top of mind when considering incorporation of data-intensive applications such as analytics, AI, and machine learning in your OT cyber security portfolio of tools and practices.

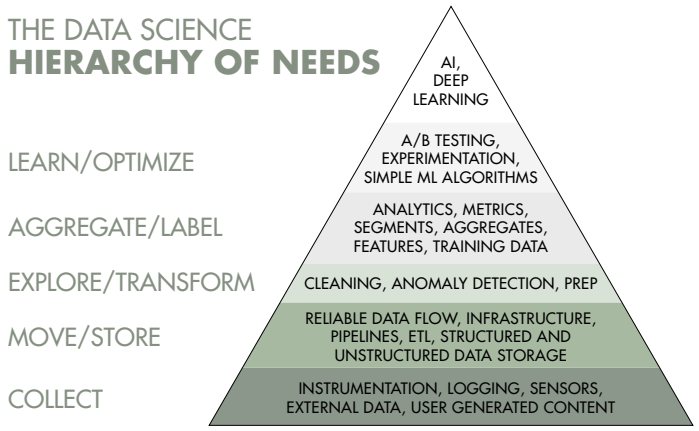


Figure 1 – From Monica Rogati: Data Science Hierarchy of Needs

Figure 1 elegantly illustrates how more sophisticated data processes at the top of the pyramid rely on a solid foundation of good data. It also highlights the increasing complexity of data activities, culminating in applications like AI and, specifically, complex learning models such as deep learning. No organization can achieve optimal results using analytics or AI applications with suboptimal or bad data.

2 Recent news disclosed that an AI-based resume-screening application had a built-in bias based on inaccurate and incomplete data that excluded qualified female candidates from consideration (bad decisions). https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Analytics and AI Basics

Analytics is a process of taking data, developing information and insights, and subsequently using the insights to make informed decisions. There are three types of analytics, described here with cyber security use cases:

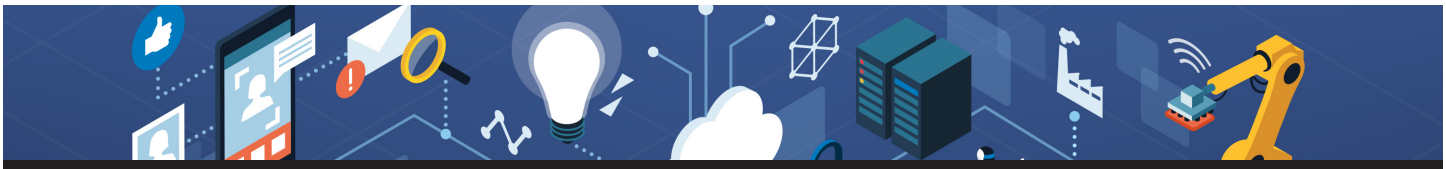
- Descriptive: Provide reports and statistics on attacks and breaches
- Predictive: Deliver information on vulnerabilities, risks, and attacks
- Prescriptive: Prevent and automatically respond to attacks

Utilities currently use analytics to identify changes in asset performance or behavior of human resources for cyber security monitoring. For instance, utility ISOCs collect data from a variety of sources and standardize, normalize, and analyze that data to assess threats to OT and information technology systems. The Electric Power Research Institute (EPRI) conducts extensive research in ISOCs. The research reports listed at the end of this paper provide detailed information about planning and implementing ISOCs, including data management considerations.

Predictive analytics are less common in OT cyber security, but the growing volume and variety of data, coupled with more powerful computing capabilities, offer interesting possibilities for utilities. ATM technologies will leverage predictive and prescriptive analytics with AI to inform machine-based detection and response to cyber security events.

AI is a science focused on creating software models that can train a computing platform to perform learning, accomplish problem-solving, and facilitate pattern recognition, all on a far larger data scale than humans can manage. Machine learning is an AI application that enables a computer to automatically learn and improve from experience without additional programming.

AI and all its machine-learning capabilities are based on algorithms. Algorithms are the coded instructions that computers apply to data to achieve desired outcomes. An omelet recipe—a set of instructions—is an example of an algorithm. Ingredients for the recipe are examples of different types of data. Eggs are one form of data. Salt and pepper are other data sources. Data preparation is much like food prep. Eggs need to be "normalized" into a liquid blend. Vegetables are standardized or sliced based on recipe instructions. Execution of an algorithm produces a result, just as following a recipe does. And in both cases, the quality of the ingredients determines the quality of the result.



Data-intensive applications like analytics and AI have significant promise for utility OT cyber security use, but only if the proper time and resources are invested in ensuring reliable data flow, readily accessible storage, and data formats for use. Data collection, transmission, and storage are determined by an organization's data governance policies and data management practices.

Data governance is strategic. It is a series of policy decisions at an organizational level that determines:

- What data must be collected
- Data formats
- How data is classified
- How data is transmitted and stored
- Where and how long data is stored

Data management is tactical—it covers the life cycle of data from creation to deletion, and it helps ensure data quality and availability. It identifies the tools and practices that will manage data to conform with data governance policies. Data governance and data management responsibilities are established within utilities. What's new and different are the increase in OT cyber security data and uses of that data, as well as new AI applications.

As noted before, this new data may help utilities detect and respond to threats. The analysis of this data must occur in or near real time to support any ATM solutions, and that raises a number of data intensity questions that must be addressed in data governance and management. Development of answers and policies should include discussions with domain experts—the utility resources who work with OT cyber security data. Many of the questions focus on the velocity and veracity of data:

- Velocity
  - What new OT data must be immediately available to utility security operations centers?
  - How and where should this data be stored? (Fast access means higher storage costs and siting near decision-making points.)
- Veracity
  - What preparation is required (normalization, standardization, etc.) for the data to be readily utilized by OT security tools and applications?
  - Do we have accurate time stamps on this data?

Data veracity or accuracy is often assumed to focus on the correct information, such as the current address of a business or the proper designation of a phone number as a mobile or a landline. There are many options to eliminate data duplication and cleanse data to build confidence in its use in analytics and more advanced data science applications. **But for OT environments, the *time stamp* on data determines accuracy and usability**.

Several research activities are investigating equipment vulnerabilities that impact time stamps based on global positioning system, Network Time Protocol, or Precision Time Protocol options. EPRI is conducting research into timing vulnerabilities to help utilities and equipment vendors ensure data veracity for mission-critical systems, such as protective relaying, wide area protection systems, and MPLS networks, that rely on absolute precision in data measurements.

The growing adoption of OT cyber security analytics and AI solutions will have impacts on utility operations infrastructure—specifically, data networks and computing capabilities. As noted above, new sources can potentially provide real-time data that helps utilities improve their security posture, but that means networks must have the capacity and capability to transport data to AI-enhanced tools.

Conclusion

Utilities are facing new varieties, volumes, and velocities of OT cyber security data. This data will be leveraged in analytics as well as AI and other data-intensive applications to supplement and enhance threat detection and response in utility ISOCs. These capabilities are the precursors to the ATM-based solutions that will be needed to handle the exponential growth of data from an ever-increasing number of devices connected to utility networks and grids. Utilities will need rock-solid data (veracity) in formats that support real-time, frictionless accessibility (velocity, volume, variety). These attributes define the value of data and are impacted by the principles funda-

mental to delivering good data. Utilities should address OT cyber security data considerations within their data governance and data management plans and ensure their computing and data communications infrastructure are prepared to handle the data volumes and speeds needed to support real-time analysis and decision making. These plans should include input from OT cyber security resources to ensure that the appropriate domain expertise is applied to these decisions. These actions will help utilities prepare for advanced analytics and AI applications like ATM.

Additional EPRI Information Resources

Guidelines for Planning an Integrated Security Operations Center. 3002000374

Guidelines for Integrating Control Center Systems into an Integrated Security Operations Center. 3002003739

Guidelines for Integrating Substation and Field Domain Events into an Integrated Security Operations Center. 3002005946

The Integrated Security Operations Center (ISOC) Guidebook. 3002013903

An Introduction to AI, Its Use Cases, and Requirements for the Electric Power Industry. 3002017143

Timing Security Assessment and Solutions. 3002008952

Timing Security Assessment and Solutions: Phase II. 3002016546

EPRI RESOURCES

Christine Hertzog, *Technical Leader, Principal*
650.314.8111, chertzog@epri.com

Cyber Security

The Electric Power Research Institute, Inc. (EPRI, www.epri.com) conducts research and development relating to the generation, delivery and use of electricity for the benefit of the public. An independent, nonprofit organization, EPRI brings together its scientists and engineers as well as experts from academia and industry to help address challenges in electricity, including reliability, efficiency, affordability, health, safety and the environment. EPRI also provides technology, policy and economic analyses to drive long-range research and development planning, and supports research in emerging technologies. EPRI members represent 90% of the electricity generated and delivered in the United States with international participation extending to 40 countries. EPRI's principal offices and laboratories are located in Palo Alto, Calif.; Charlotte, N.C.; Knoxville, Tenn.; Dallas, Texas; Lenox, Mass.; and Washington, D.C.

Together . . . Shaping the Future of Electricity

3002017455 November 2019

Electric Power Research Institute
3420 Hillview Avenue, Palo Alto, California 94304-1338 • PO Box 10412, Palo Alto, California 94303-0813 USA
800.313.3774 • 650.855.2121 • askepri@epri.com • www.epri.com

© 2019 Electric Power Research Institute (EPRI), Inc. All rights reserved. Electric Power Research Institute, EPRI, and TOGETHER . . . SHAPING THE FUTURE OF ELECTRICITY are registered service marks of the Electric Power Research Institute, Inc.