Science of Cybersecurity

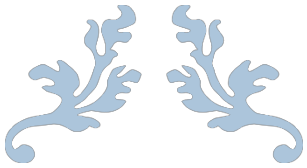
Developing Scientific Foundations for the Operational Cybersecurity Ecosystem

Solving the grand challenge --- How to develop an organized, cohesive body of knowledge that informs the field of cybersecurity

CSCSS
CENTRE FOR STRATEGIC
CYBERSPACE + SECURITY SCIENCE

by

Shawn Riley
Executive Vice President
Strategic Cyber Science



SCIENCE OF SECURITY

Developing Scientific Foundations for the Operational Cybersecurity Ecosystem

AUGUST 11, 2015

SHAWN P. RILEY
EVP, Strategic Cyberspace Science
Centre for Strategic Cyberspace + Security Science

Table of Contents

Introduction

According to PwC's 18th Annual CEO Survey [1], 90% of US CEOs say cybersecurity is strategically important, 87% are concerned about cyber threats, and 45% are extremely concerned about them. Unfortunately, knowing about the problem and fixing it are two different things entirely. Especially when a 2015 ISACA survey [2] showed 86% of respondents say there is a global shortage of skilled cybersecurity professionals with the required knowledge.

Finding people with the right knowledge is hard. In fact, 92% of ISACA's survey respondents whose organizations will be hiring cybersecurity professionals in 2015 say it will be difficult to find skilled candidates. There are several compounding issues at play when looking at this skills shortage. There are dozens of different cybersecurity job roles across security engineering, security operations, and security intelligence with more roles being identified as the field continues to evolve.

Organizations spend a lot of time and effort trying to find the much sought after skilled employees with the right set of knowledge. The problem is employees don't tend to stick around as long as they once did. The average worker today stays at each of his or her jobs for 4.4 years, according to data from the Bureau of Labor Statistics, but the expected tenure of the workforce's youngest employees is about half that. 91% of Millennials expect to stay in a job for less than three years, according to the Future Workplace "Multiple Generations @ Work" survey of 1,189 employees and 150 managers [3].

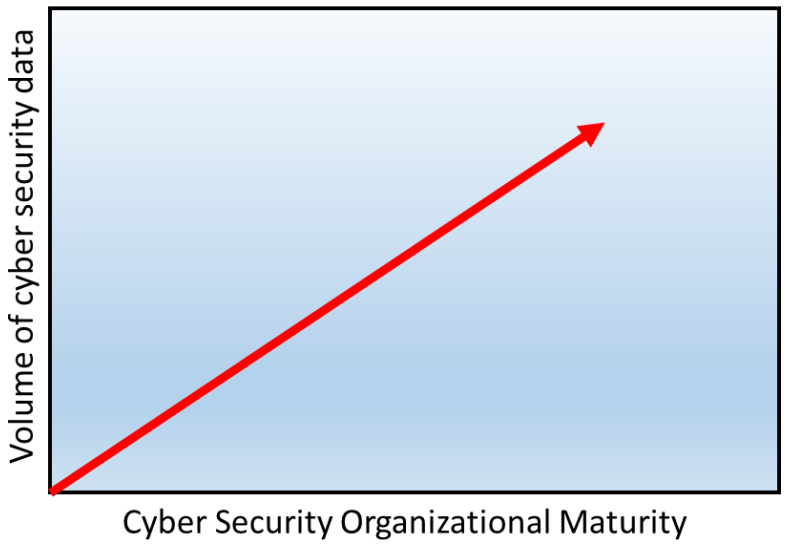
As it often the case when talent is in limited supply, organizations also worry about their much in-demand talent being poached, which is especially true for cybersecurity talent. In January 2015, MasterCard hit Nike with a $5M cyber talent poaching suit [4]. The suit noted that companies are desperate for information security talent amid highly publicized data breaches at Target Corp. and Home Depot Inc. While the area is fast-growing, skilled workers are limited and in demand, according to MasterCard.

Having the right skills and knowledge is critical to an organization's cybersecurity program. This brings us to the next issue compounding the skills shortage. As we look across different industries and organizations there are varying levels of maturity in the execution of each organization's cybersecurity program. Some organizations are just starting their cybersecurity journey and beginning activities such as Cyber Hygiene while

others organizations have expanded beyond hygiene and are focusing more on tactical cyber threat information and intelligence.

Organizations will have different skills and knowledge requirements as they progress up the maturity model. The early years of an effective cybersecurity program is focused on cyber hygiene activities which are aimed at understanding the organization's cybersecurity health to counter the lower 80% of cyber threats. During this stage an organization produces detailed knowledge on the security of their enterprise. This knowledge can be used to measure the security of the enterprise, determine risk, threat susceptibility, and countermeasure based risk remediation activities. However, organizations have to have people with the knowledge and skills to understand the data and information as well as how to measure and score it to produce standardized repeatable results an organization can use.

Once that initial task of cyber hygiene has become institutionalized, organizations expand their focus to the internal and external threats to the organization. This is done in hopes of expanding their current level of knowledge and understanding of the events causing incidents and breaches and how to best prevent them. Essentially, the data analysis requirements will continue to increase as the amount of threat data and requirements for intelligence from open-sources grows.



Many organizations today struggle with making the leap from analyzing raw security data and identifying patterns in security information to being able to expand or produce new knowledge and enable predictability. Security teams today are overwhelmed with both the different types and volume of data

they have to go through to find the knowledge and get the answers they need as a results of ad-hoc and often non-repeatable approaches. Addressing these challenges requires organizations to begin to think about how they approach these problems using a different, more rigorous and repeatable manner. Organizations should seek to lay a scientific foundation to their cybersecurity program that can help them not with the "big data" problem but with turning big data into knowledge they can use.

Developing a strong, rigorous scientific foundation to the real-world cyber ecosystem is a grand challenge. To help us better understand the grand challenge we will look at 4 documents that were published in 2011 that provide key pieces of background information. The first document, "Trustworthy Cyberspace: Strategic Plan for the Federal Cybersecurity Research and Development Program", will define the problem or puzzle to be solved. The second document, "Science of Security Joint Statement of Understanding", will define the 7 core themes that need to be developed within the cyber ecosystem. The third document, "Enabling Distributed Security in Cyberspace – Building a Healthy and Resilient Cyber Ecosystem with Automated Collective Action", will define the cyber ecosystem and 3 interdependent building blocks of a secure cyber ecosystem. The fourth document, "Semantic eScience: Encoding Meaning in Next-Generation Digitally Enhanced Science" by Peter Fox and James Hendler of Rensselaer Polytechnic Institute, will define the core technology solution to this grand challenge.

We will then build on this background information in the next section where we'll look at how Semantic eScience and semantic technology work in the operational cyber ecosystem. We'll go over the technology stack and components that make up a Semantic eScience solution customized for the Science of Security discipline. We'll address how the Semantic eScience of Security solution supports the 7 core themes within the operational cyber ecosystem.

We'll then address the key interaction points of the configured and operational Semantic eScience of Security solution and how it enables the state-of-the-art Intelligence Community methodologies called Object-Based Production (OBP) and Activity-Based Intelligence (ABI) [5]. We will explain how OBP is used to develop an organized, cohesive body of knowledge that informs the field of cybersecurity and ABI enables us to discover what are called the "unknown unknowns".

Background

In 2011, the White House released the document "Trustworthy Cyberspace: Strategic Plan for the Federal Cybersecurity Research and Development Program [6]", a strategic plan for cybersecurity research and development. 1 of the 4 main thrusts of the program is "Developing Scientific Foundations – minimizing future cybersecurity problems by developing the science of security" on which they write:

> *In anticipation of the challenges in securing the cyber systems of the future, we must develop an organized, cohesive foundation to the body of knowledge that informs the field of cybersecurity. Currently, we spend considerable intellectual energy on a patchwork of targeted, tactical activities, some of which lead to significant breakthroughs while others result in a seemingly endless chase to remedy individual vulnerabilities with solutions of limited scope. A more fruitful way to ground research efforts, and to nurture and sustain progress, is to develop a science of security.*
>
> *Developing a strong, rigorous scientific foundation to cybersecurity helps the field in the following ways: Organizes disparate areas of knowledge – Provides structure and organization to a broad-based body of knowledge in the form of testable models and predictions – Enables discovery of universal laws – Produces laws that express an understanding of basic, universal dynamics against which to test problems and formulate explanations – Applies the rigor of the scientific method – Approaches problems using a systematic methodology and discipline to formulate hypotheses, design and execute repeatable experiments, and collect and analyze data.*

From the above we can identify the core puzzle to solve: How to develop the science of security to enable an organized, cohesive foundation to the body of knowledge that informs the field of cybersecurity. For our efforts, we want to develop the science of security in the operational cyber ecosystem so the body of knowledge is based on the actual cybersecurity implemented in operations.

What is the Science of Security?

In the 2011 "Science of Security Joint Statement of Understanding [7]" federal agencies of the Canadian, United Kingdom, and United States governments identified a set of 7 core themes that together form the foundational basis for the science of security discipline. The themes are strongly inter-related, and mutually inform and benefit each other. They are:

- Common Language
- Core Principles
- Attack Analysis
- Measurable Security
- Risk
- Agility
- Human Factors

In the statement, agencies further clarify that in the context of security, science can be thought of as knowledge that results in the correct predictions or reliable outcomes. The "Science of Security" resides in a particularly complex area, being at the intersection of behavioral sciences, formal sciences, and natural sciences.

From this information we understand that the cybersecurity body of knowledge will need to support the 7 inter-related core themes of the science of security discipline within the cyber ecosystem.

We'll discuss the core themes in more detail in the section of the paper on how the 7 core themes are supported within the operational cyber ecosystem.

What is the cyber ecosystem?

In 2011, the U.S. Department of Homeland Security (DHS) released "Enabling Distributed Security in Cyberspace – Building a Healthy and Resilient Cyber Ecosystem with Automated Collective Action [8]". In it DHS states:

> *This discussion paper explores the idea of a healthy, resilient – and fundamentally more secure – cyber ecosystem of the future, in which cyber participants, including cyber devices, are able to work together in near-real time to anticipate and prevent cyber attacks, limit the spread of attacks across participating devices, minimize the consequences of attacks, and recover to a trusted state. In this future cyber ecosystem, security capabilities are built into cyber devices in a way that allows preventive and defensive courses of action to be coordinated within and among communities of devices. Power is distributed among participants, and near-real time coordination is enabled by combining the innate and interoperable capabilities of individual devices with trusted information exchanges and shared, configurable policies.*

In the 2011 paper, DHS identifies three interdependent building blocks of a healthy cyber ecosystem as:

Building Block 1 – Automation

> *Automated Courses of Action (ACOAs) are strategies that incorporate decisions made and actions taken in response to cyber situations. Automation frees humans to do what they do well – think, ask questions, and make judgments about complex situations. Automation allows the speed of response to approach the speed of attack, rather than relying on human responses to attacks that are occurring at machine speed. With the ability to execute at machine speed, defenders could get inside the turning circles or decision cycles of attackers. Further, automation could make it easier to adopt and adapt new or proven security solutions.*

Building Block 2 – Interoperability

> *Interoperability allows cyber communities to be defined by policies rather than by technical constraints and permits cyber participants to collaborate seamlessly and dynamically in automated community defense. Interoperability enables common operational pictures and shared situational awareness to emerge and disseminate rapidly. The creation of new kinds of intelligence (such as fused sensor inputs), coupled with rapid learning at both the machine and the human levels, could fundamentally change the ecosystem.*

Building Block 3 – Authentication

> *Authentication should enable trusted online decisions. Nearly every decision in an online environment involves resources and actors at a distance. When needed for a decision, authentication provides appropriate assurance that the participants are authentic or genuine, and it should do so in a way that enhances individual privacy. In a healthy ecosystem, authentication could extend beyond persons to include cyber devices (e.g., computers; software, or information).*

To better understand the cybersecurity data and information in the cyber ecosystem we need to focus on the interoperability building block. DHS calls out three types of interoperability that are fundamental to integrating the many disparate participants into a comprehensive cyber defense system that

can create new intelligence and make and implement decisions at machine speed. They are:

- **Semantic Interoperability**. The ability of each sending party to communicate data and have receiving parties understand the message in the sense intended by the sending party
- **Technical Interoperability.** The ability for different technologies to communicate and exchange data based upon well-defined and widely adopted interface standards.
- **Policy Interoperability.** Common business processes related to the transmission, receipt, and acceptance of data among participants.

DHS further clarifies how interoperability is being enabled in the operational cyber ecosystem:

> *Within cybersecurity, all three types of interoperability are being enabled through an approach that has been refined over the past decade by many in industry, academia, and government.* ***It is an information-oriented approach****, generally referred to as [cyber] security content automation and comprises the following elements.*
>
> ***Enumerations.*** *These are lists or catalogs of the fundamental entities of cybersecurity, for example, cyber devices and software items (CPE); device and software configurations (CCE); publicly known weaknesses in architecture, design, or code (CWE); publicly known flaws or vulnerabilities (CVE); or publicly known attack patterns (CAPEC). Enumerations enable semantic interoperability.*
>
> ***Languages and Formats.*** *These incorporate enumerations and support the creation of machine-readable security state assertions, assessment results, audit logs, messages, and reports. Examples include patterns associated with assets, configurations, vulnerabilities, and software patches (XCCDF & OVAL); security announcements (CAIF), events (CEE), malware (MAEC); risk associated with vulnerability (CVSS), sensor collection and correlation (ARF), and US-CERT security bulletins and incident reports (NIEM). Languages and formats enable technical interoperability.*
>
> ***Knowledge Repositories.*** *These contain a broad collection of best practices, benchmarks, profiles, standards, templates, checklists, tools, guidelines, rules, and principles, among others. In many respects, knowledge repositories serve as the cybersecurity community "memory" and enable policy interoperability. Examples include Information Assurance Checklists housed on the National*

> *Checklist Program website (http://checklists.nist.gov/), Department of Defense Security Technical Implementation Guides (STIGs), and vendor guides."*

The Enumerations, Languages, Formats, and Knowledge Repositories are what is better known as the Cybersecurity Measurement and Management Architecture [9] which is led by The MITRE Corporation as a Federally Funded Research and Development Center. MITRE's website the states [10]:

> MITRE, in collaboration with government, industry, and academic stakeholders, is improving the measurability of security through **registries** of baseline security data, providing standardized **languages** as means for accurately communicating the information, defining proper **usage**, and helping establish community approaches for standardized **processes**.
>
> The other activities and initiatives listed here have similar concepts or compatible approaches to MITRE's. Together all of these efforts — be they mature or continuing to build momentum — are helping to make security more measurable by defining the concepts that need to be measured, providing for high fidelity communications about the measurements, and providing for sharing of the measurements and the definitions of what to measure.
>
> Measurable security pertains at a minimum to the following areas:
>
> - Software Assurance
> - Application Security
> - Asset Management
> - Supply Chain Risk Management
> - Cyber Intelligence Threat Analysis
> - Cyber Threat Information Sharing
> - Vulnerability Management
> - Patch Management
> - Configuration Management
> - Malware Protection
> - Intrusion Detection
> - System Assessment
> - Incident Coordination
> - Enterprise Reporting
> - Remediation
>
> The basic premise of the "Making Security Measurable" effort is that for any enterprise to operate, measure, and manage the security of

> their cyber assets they are going to have to employ automation. For an enterprise of any reasonable size that automation will have to come from multiple sources. To make the finding, sharing, and reporting issues consistent and composable across different tools and partners there has to be a set of standardized definitions of the things that are being examined, reported, and managed by those different tools and described by different information sources. That standardization is what comprises the core of the "Making Security Measurable" efforts.

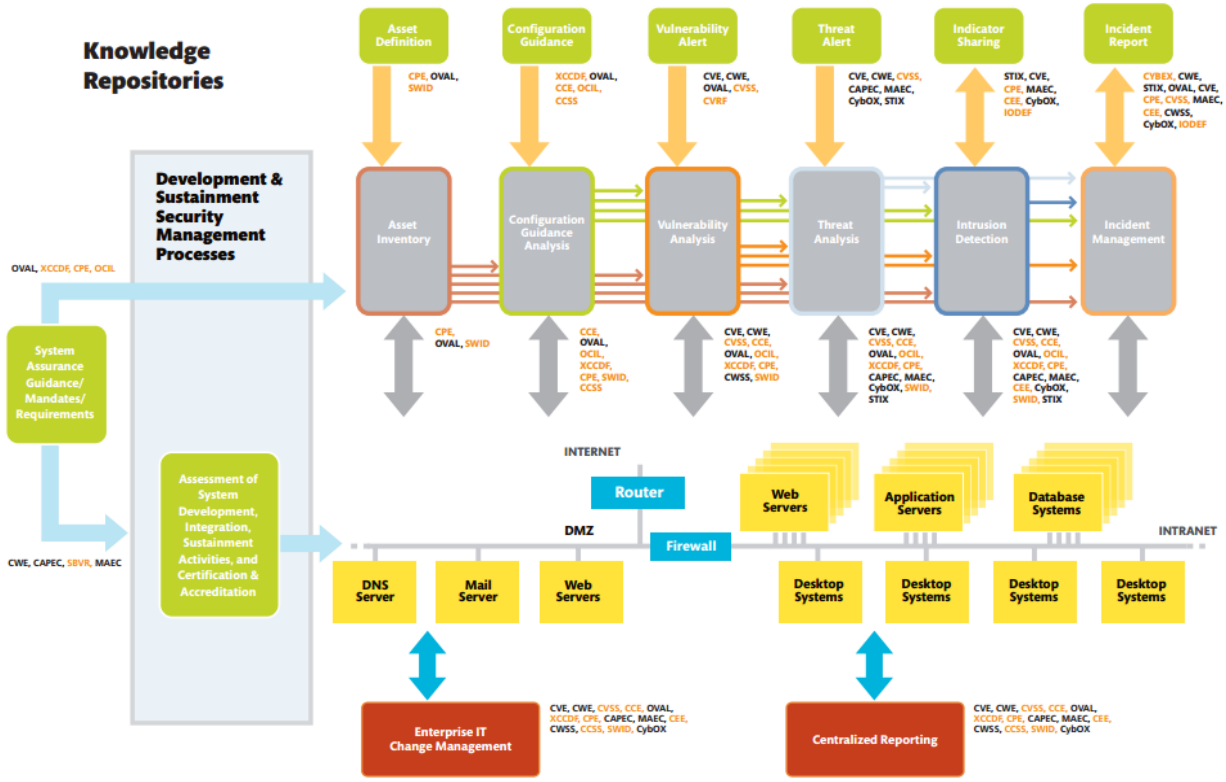
We can observe that security vendors are supporting the development and implementation of the Cybersecurity Measurement and Management Architecture by looking at compatible cybersecurity products and services for both older languages like the Common Vulnerability Enumeration [11] (CVE) here and newer languages such as the Structured Threat Information eXpression [12] (STIX) found here.

From the information provided by DHS we can identify that for more than a decade the cybersecurity community has taken an "information-oriented" or informatics approach and developed enumerations, languages, formats, and repositories for the cybersecurity data and information. We can further observe security technologies have integrated or are working towards integration of the making security measurable standards in order to collaborate in the operational cyber ecosystem. For the purposes of this paper, we want to build on and take advantage of the Cybersecurity Measurement and Management Architecture in the operational cyber ecosystem.

While an information-oriented approach was needed to support the interoperability and automation building blocks, it has resulted in the creation of silos of cybersecurity information. We have silos for vulnerabilities, attack patterns, weaknesses, malware characterization, threat information, etc.

The ever increasing threat of cyber-attack has spurred the need to share more and more cybersecurity information about the attacks that are happening as part of shared situational awareness. This has driven the formation of new Information Sharing and Analysis Centers [13] (ISAC) for vertical industries as well new government threat sharing initiatives such as the creation of the recent concept of Information Sharing and Analysis Organizations [14] (ISAO). While these new entities aggregate and share threat intelligence, they still form yet another silo of security information since the information isn't shared uniformly across all industries and organizations.

Now that we better understand the operational cyber ecosystem and the significant investment made by the international cybersecurity community to develop and build the Cybersecurity Measurement and Management Architecture [9] at the information layer, how do we develop the science of security so that we can transform the cybersecurity information into an organized, cohesive body of cybersecurity knowledge?



To help us identify a solution, we can observe that the majority of the cybersecurity languages and formats created over the past decade by the cybersecurity community have leveraged the World Wide Web Consortium (W3C) Extensible Markup Language (XML) standard [15]. We can also observe that the Science of Security is a data intensive science, a characteristic of eScience. From these observations we were able to identify the Web Science area of Semantic eScience as being the best candidate for developing the Science of Security within the operational cyber ecosystem that would result in an organized, cohesive body of knowledge that informs the field of cybersecurity.

What is Semantic eScience?

In 2011, the Microsoft Research book, "The Fourth Paradigm: Data-Intensive Scientific Discovery [16]" contained an excellent essay by Peter Fox and

James Hendler of Rensselaer Polytechnic Institute titled "Semantic eScience: Encoding Meaning in Next-Generation Digitally Enhanced Science [17]". While the essay does not address the topic of cybersecurity, the following 2 paragraphs gives a good insight to Semantic eScience for the purposes of this paper.

> An important insight into dealing with heterogeneous data is that if you know what the data "means," it will be easier to use. As the volume, complexity, and heterogeneity of data resources grow, scientists increasingly need new capabilities that rely on new "semantic" approaches (e.g., in the form of ontologies—machine encodings of terms, concepts, and relations among them). Semantic technologies are gaining momentum in eScience areas such as solar-terrestrial physics, ecology, ocean and marine sciences, healthcare, and life sciences, to name but a few. The developers of eScience infrastructures are increasingly in need of semantic-based methodologies, tools, and middleware. They can in turn facilitate scientific knowledge modeling, logic-based hypothesis checking, semantic data integration, application composition, and integrated knowledge discovery and data analysis for different scientific domains and systems noted above, for use by scientists, students, and, increasingly, non-experts.
>
> The influence of the artificial intelligence community and the increasing amount of data available on the Web (which has led many scientists to use the Web as their primary "computer") have led semantic Web researchers to focus both on formal aspects of semantic representation languages and on general-purpose semantic application development. Languages are being standardized, and communities are in turn using those languages to build and use ontologies—specifications of concepts and terms and the relations between them (in the formal, machine-readable sense). All of the capabilities currently needed by eScience—including data integration, fusion, and mining; workflow development, orchestration, and execution; capture of provenance, lineage, and data quality; validation, verification, and trust of data authenticity; and fitness for purpose—need semantic representation and mediation if eScience is to become fully data-intensive.

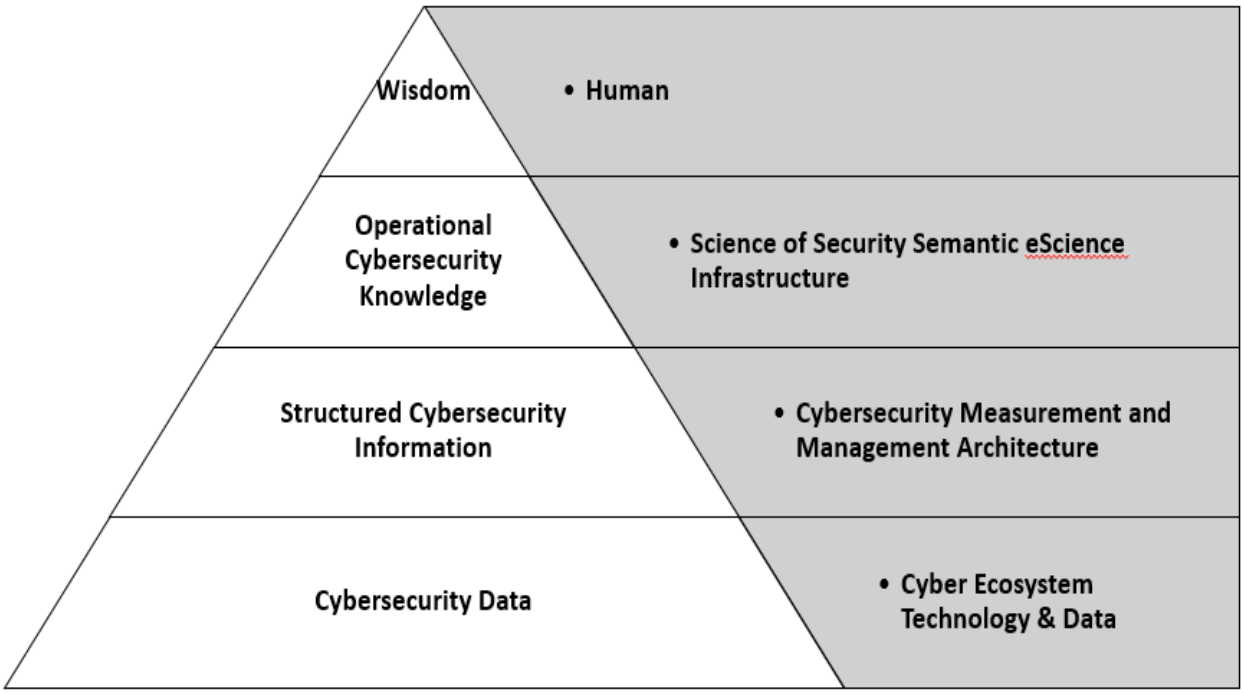
From these paragraphs we can identify that Semantic eScience is the development of the infrastructure needed to facilitate scientific knowledge modeling, logic-based hypothesis checking, semantic data integration,

application composition, and integrated knowledge discovery and data analysis for the scientific domain it was developed to support. Semantic technology being a key component of that infrastructure.

To help pull the Semantic eScience and the operational cyber ecosystem together, we will leverage the Data Information Knowledge Wisdom (DIKW) pyramid. The bottom of the pyramid is the Data Level and it represents the basic cyber ecosystem. People, Processes, Technology and all the cybersecurity data. Every organization has this layer. This is an ecosystem without the automation and interoperability provided by the enumerations, languages, formats, and knowledge repositories that make up the Cybersecurity Measurement and Management Architecture.

The information layer of the cyber ecosystem is the Cybersecurity Measurement and Management Architecture. The cybersecurity data for the cyber ecosystem is now captured as structured cybersecurity information as part of the "information-oriented" approach described in the DHS paper.

The languages that make up the Cybersecurity Measurement and Management Architecture are written using the W3C XML standardized language. The use of XML allows the cybersecurity data to be machine readable and the format of the data validated. Many organizations are slowly maturing in their practice and adoption of the Cybersecurity Measurement and Management Architecture.



The Semantic eScience of Security infrastructure builds a strong, rigorous scientific foundation on top of the information layer to create the knowledge layer. The semantic technology provides an abstraction layer above existing IT technologies that enables bridging and interconnection of data, context, and processes that provides far more intelligence, capable, relevant, and responsive interaction than with information technologies alone. This enables the structured cybersecurity information from the Cybersecurity Measurement and Management Architecture to be federated into an organized, cohesive body of knowledge that is based on the cybersecurity of the real-world operational cyber ecosystem.

Developing an Organized, Cohesive Body of Knowledge

To develop an organized, cohesive body of knowledge from the Cybersecurity Measurement and Management Architecture, we need to be able to integrate the structured information described by the individual common languages (STIX [12], CybOX [18], MAEC [19], CVE [11], CAPEC [20], etc.) into a unified conceptual data model for cybersecurity knowledge. To do this we will turn to some languages that came out of the field of Artificial Intelligence that are specifically designed to express semantic models: W3C Web Ontology Language (OWL) [21] and the W3C Resource Description Framework (RDF) [22].

Semantic technology provides a means to capture the actual semantics of the data with the data itself. In addition, it also enables the ability to capture a meta-description of the different kind of objects, their attributes, associations, and activity into a conceptual model which can then be populated with instances of actual data. Described using the industry-standard OWL/RDF syntax, it's possible to capture the conceptual model, referred to as an "ontology", and represent the data itself in a single, consistent manner that is independent of how it is physically stored.

Ontologies are a formal way to describe taxonomies and classification networks, essentially defining the structure of knowledge for various domains: the nouns representing classes of objects and the verbs representing relationships between the objects. In addition, an ontology forms a vocabulary with well-defined semantics that can be used or combined with other ontologies allowing for the creation of new concepts.

Ontologies are meant to represent information coming from all sorts of heterogeneous data sources. This makes ontologies ideal for dealing with all

the different structured, semi-structured, and unstructured data that comes in various formats and languages from across cyberspace.

The conceptual data model expressed in OWL/RDF is able to express concepts similar to other classical conceptual modeling approaches such as Entity-Relationship or UML class diagrams: an OWL Class is similar to an Entity in an ER diagram or a Class in a UML diagram. But unlike in ERD and UML, attributes referred to as "data properties" and associations referred to as "object properties" are used by reference, not scoped to the definition of a Class or Entity, thus allowing for a single, consistent semantic to be defined for a class, attribute, or association.

When data is mapped against an OWL/RDF ontology, instances of the data are expressed based upon the idea of making statements about resources in the form of **subject–predicate–object** expressions. These expressions are known as *triples* in RDF terminology.

The 'Subject' denotes the object, and the predicate denotes a single semantic trait or aspect of the object that can be a literal value or expressed as a relationship between the subject and another object that is the target of the relationship.

For example, one way to represent the notion "The sky has the color blue" in RDF is as the triple: a **subject** denoting "the Sky", a **predicate** denoting "hasColor", and an **object** denoting "blue". The semantics of the "hasColor" predicate indicate that the object should contain a color. Therefore OWL/RDF uses the object as the subject that would be used in the classical notation of an entity–attribute–value model within object-oriented design; object (sky), attribute (color) and value (blue). The object (Sky) can also have another attribute (contains) that can point to another object (Cloud). The object (Cloud) might have an attribute (produces) another object (Rain). This forms a series of relationships between two or more objects that is the basis for on which directed graphs can be built.

Core to each statement is the concept of a 'predicate' that represent a single semantic concept. This allows the single semantic concept to be learned once and leveraged where needed, thus avoiding the problem of trying to infer meaning based on ambiguous terms. Because predicates can refer to

literal values or other objects, statements about an object that refer to another object can be used to form one or more directed graphs that illustrate the various relationships amongst objects.

As stated before 'objects' are nouns and as such objects represent the people, places, events, and things in cyberspace. The Cybersecurity Measurement and Management Architecture is full of different objects that are relevant to cybersecurity such as: Threat Actors, Campaigns, TTPs, Incidents, Vulnerabilities, Weaknesses, Malware, Indicators, Observables, and many others.

A collection of RDF statements intrinsically represents a directed multi-graph. As such, an OWL/RDF-based data model is more naturally suited to certain kinds of knowledge representation than the relational model because it can fuse data from multiple relationship tables about the same object.

This simplistic approach allows any number of statements to be made about an object and through the use of references to other objects there is virtually nothing that can't be described using this technique. Additionally, it abstracts the user from the physical storage mechanism; there is no need to know about views and tables. And unlike other techniques where the semantics of an attribute or association must be implied or derived from other places (e.g., inferred from the header of a table), OWL/RDF allows the meaning to be directly connected to the data. Therefore the predicate which conveys semantic meaning is always with the data itself.

Because of the critical role provenance plays in Semantic eScience of Security, it is necessary to be able to express provenance information that can be used to form assessments about the quality, reliability or trustworthiness of data. The W3C Provenance (PROV) [23] standard defines an ontology for representing provenance information about entities, activities, and people involved in producing a piece of data or object. But because provenance information could refer to a single attribute or association, such as why there is a relationship between a threat actor and a specific campaign, it is necessary to be able relate the provenance information to a single RDF statement.

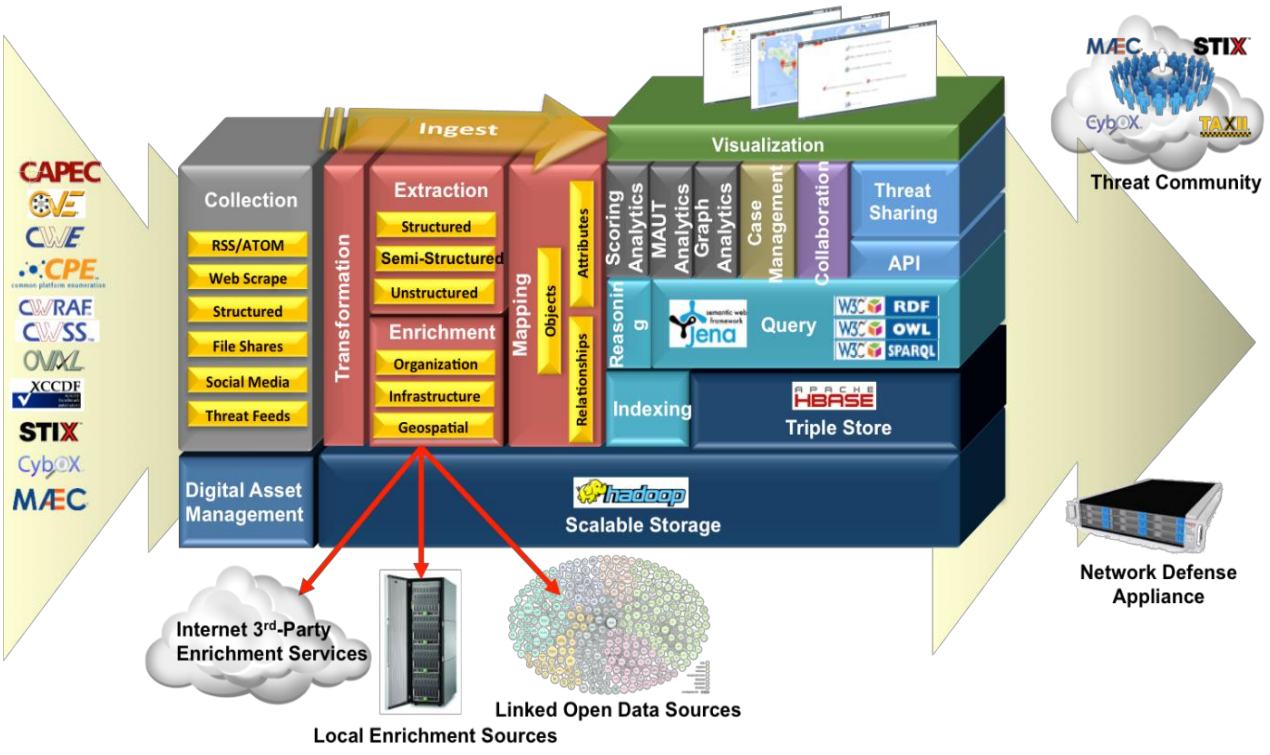
RDF supports the ability to make statements about statements through the concept of reification. Expressed as a collection of RDF statements, reification data can be related to a specific RDF statement allowing information about the statement to be expressed.

The OWL and RDF standards provide ways to unambiguously identify objects and to describe the activity, attributes, and associations of those objects in a

language that enables Semantic eScience of Security solutions to read, write, and comprehend the meaning of the data. This then enables the ability to reason about and across the data using logic based reasoning rules to infer new knowledge based on what it already knows.

Blueprint of Semantic eScience of Security Technology Stack

The actual composition of the technology stack for a Semantic eScience of Security platform can vary based on a number of different factors: area of focus, connectivity, expected number of concurrent users, expected volume and rate of data ingest, just to mention a few. Regardless of these factors, any Semantic eScience of Security platform will have a common set of core capabilities required to achieve its objectives. The illustration below outlines the major subsystems that make up the technology stack with components color-coded that are common to a particular subsystem.



The core of a Semantic eScience of Security technology stack generally consists of subsystems focused on the ingest and fusion of data, the storage of fused data, the ability to make inquiries against the data, analytics run against the data, visualization and navigation of the data, and the

generation and sharing of information in a number of formats and manners. However, one key component not shown in the illustration is the *Conceptual Data Model*.

Conceptual Data Model

The Conceptual Data Model drives a number of decisions in the technology stack. Expressed as a set of one or more ontologies described in RDF/OWL, the conceptual data model plays a prominent role in decisions around collection, fusion, query, storage, and a minor role concerning other major subsystems. The conceptual data model defines the types of data required to answer domain-specific questions. The techniques and ability to evolve the conceptual data model are critical to the overall systems ability to evolve and answer questions yet thought of. Finally, understanding what data is needed drives decisions about where does one go to get that data, what format is it in, and what needs to be done with it.

Collection

Collection is based on automation of an extensible framework of plug-ins, referred to as "collectors" that are responsible for obtaining the raw cyber data and associating critical metadata, such as when and from where it was collected, with the artifact in a data lake possessing the same characteristics as a Digital Asset Management [24] system. The use of an extensible framework supports evolution as the number of formats and structures constantly change. The storage of the artifacts also enables the ability for re-analysis should new analysis techniques or new concepts be added to the conceptual data model. Collectors generally leverage common technology used to collect data from multiple open-sources, such as the Internet: crawling and scraping web pages, monitor threat feeds, ingest of files in a file system, and so forth. The metadata about an artifact serves as provenance used as a basis for confidence in validity of the data from the source, sensitivity associated with others knowing about source, and any other metadata that might be useful.

Ingest and Fusion

Regardless of the method used to acquire the raw cyber data, the Ingest and Fusion capability is responsible for the transformation of the data from its native format into a representation utilized throughout the rest of the platform. It is through the process of ingest and fusion that the raw cyber data is mapped against the conceptual data model. To accomplish this, ingest and fusion leverages translation services, extraction services, enrichment services, and mapping services. The Translation Services perform translations from a native representation to a more common

representation, while the Domain-aware Extraction Services identify and extract relevant entities of interest. The Enrichment Services automate obtaining additional data with which to augment extracted entities, and the Mapping Services then map the extracted and augmented data to a representation based on the conceptual data model.

Translation

Depending upon the native representation of the raw cyber data, a translation engine transforms the raw cyber data into another representation. For example, most malware sandboxes have proprietary formats that can be translated into a common representation format such as MAEC [19]. Translation between human languages is an example of a translation that might occur on unstructured or semi-structured text before proceeding to extraction.

Extraction

Extraction focuses on the accurate identification of entities and the context in which they exist within a single artifact. Context helps increase the accuracy of entity identification. Because cyber data is represented in different forms (structured, unstructured, semi-structured) and formats (XML [15], JSON [25], text), different extraction techniques must be utilized. Structured representations, such as in XML and JSON, are best handled with extractors that are specific to the format since they can leverage the associated metadata to increase the accuracy of identification. Unstructured representations, such as web page and document content, have increased complexities due to the lack of formal context and often require the use of regular expressions or Natural Language Processing (NLP) [26] techniques as a means to identify relevant entities. In addition to entity identification and extraction, the extraction process must keep track of the provenance from which an extracted entity was contained.

Mapping

The mapping services are responsible for properly assigning the appropriate semantic meaning to the entity information identified and fusing that data into the appropriate object with its corresponding attributes and relationships. To achieve this requires each mapping service to have knowledge of the conceptual data model in order to map the entities. With this knowledge, a mapping service is able to map an entity to an appropriate ontology class to form an object, the values associated with the entity as data properties of the object, and any relationships with other entities as object properties.

Additionally, a mapping service must also generate a representation of the provenance metadata and assign an identifier to the object

representing the entity. The object's identifier enables any additional data associated with an entity, identified while processing other artifacts or through enrichment, to "rendezvous" with any previous data for the entity, thus allowing additional relationships or properties to be added. However, determining an appropriate identifier for an entity requires careful consideration. Not all entities have one or more properties than can be used as the basis for a unique identifier.

Enrichment

Enrichment is the automated process, informed by cyber tradecraft and an understanding of what additional data is required, to collect and associate additional information for an entity in order to gain better insights. Augmentation extends an "enrichable" entity with any pertinent information obtained from an "enrichment source". For example, enrichment of IP addresses and domain names can include geo-location and registration information. Because the enrichment process results in the creation of additional potentially "enrichable" cyber data entities, consideration for the detection and escape of potentially infinite loops is critical.

Storage

Given the focus of a Semantic eScience of Security technology stack is on data and the massive amounts of data that needs to be stored, the storage of ingested and generated data is a critical aspect. It must support not only the storage of the results of data fusion in a representation that is optimized for both retrieval and update, but also the source artifacts from which the fusion data was generated, as well as provide mechanisms to support the control of access to sensitive data.

Physical Representation

The format in which fused data is stored needs to take into consideration both retrieval as well as support for continuous evolution since the introduction of unforeseen concepts and relationships are part of normal domain evolution. While there are a number of popular approaches to storing the data, most often lack the ability to associate the semantics directly with the data. Storing the fused data utilizing the N-Triple [27] semantic serialization scheme avoids this limitation since evolution occurs by just adding additional triples. Most other changes, such as changing the type of an object, result in a single update to just one of the triple elements.

Controlling Access to Sensitive Data

Finally, the storage must support mechanisms that aide in the control of access to sensitive data. The exact mechanisms vary based on the

technology utilized, but it is often necessary to store metadata that contains information about the conditions under which a particular entity or a specific attribute is accessible. This includes data markings that can indicate the conditions under which something is shareable.

Inquiry and Retrieval

The ability to inquire upon and retrieve analyzed data is the heart and soul of a Semantic eScience of Security technology stack. Querying capabilities need to support both the ability to ask a direct question such as "show me all malware with a specific country of origin" as well as searches based on keywords such as "show me all articles mentioning 'Poison Ivy'". Additionally, logic based reasoning is required to be applied in order to identify inferred relationships between entities that are not obvious or explicitly stated thus filling in missing pieces of information.

Query

While traditional keyword searches typically generate a list of results that then require a human to perform further processing upon, they often lack the ability to support queries that retrieve inter-connected information that are critical when looking for the preverbal "need in the haystack". As one of the means to address this limitation, the technology stack leverages the lingua franca for querying in the semantic world: SPARQL [28] – a query language that can express queries across diverse data sources, whether the data is stored natively as RDF or viewed as RDF via middleware. SPARQL allows the formulation of complex queries without exposing the need to understand the physical layout of how data is stored. SPARQL also allows for federation of queries across a number of different endpoints, a key architectural approach in scalable solutions that support the concept of moving the compute to the data instead of data to the compute point.

Search

Traditional keyword searching still has its place in the technology stack. The ability to query the contents of an artifact for a keyword remains important. To support a unified query and search capability, the technology stack integrates the query capability with a search engine that has indexed both the content of artifacts and the triples. This combination can accelerate query performance of fused data when the contents of triples are also included in the index.

Reasoning

Even with powerful query and search combined, there is always going to be an issue with finding hidden relations between entities including sub-classification and related entities. Traditionally, it has taken a human to

look at all the data to identify these hidden relations and the accuracy depends almost entirely upon the experience of the individual. The technology stack enables automation of relationship identification using reasoning. A reasoning engine automates the application of logic-based reasoning rules against data, both inline as part of a query or executed out-of-band, to identify relationships that can be persistently stored or generated as temporal relationships.

Analytics

Automated and human analysis of data is only one part of the solution. Using the same data to determine trends, perform risk and vulnerability analysis, and other domain-specific tasks requires the addition of various kinds of analytics. For example, determining the sphere of influence for a set of cyber threat actors requires various forms of graph analytics. Risk and Threat Analysis is based on the operational assets to be utilized requires the use of scoring engines, such as the Portable Format for Analytics (PFA) [29], and the application of Multi-Attribute Utility Theory (MAUT) [30] or Analytic Hierarchy Processing (AHP) [31] to help rank order alternatives. Regardless of the specific purpose, the technology stack must provide an array of analytic capabilities to assist in the creation and proof of hypothesis.

Visualization and Navigation

Humans are naturally visually oriented and thus there remains a challenge of how to best visualize the massive amounts of data contained within a Semantic eScience of Security platform. One of the visual complexities that occurs is how to help users navigate the "sea" of data in a manner that is efficient and simple while allowing them to control how they want to see the data. The technology stack supports the visualization of data in a number of different ways to help users with "living in the data": directed graphs, tables, temporal (in time), and geospatial (geographic location). Inter-linking the various views enables the reflection of any change to effect each of the other views, thus allowing users to pivot quickly between them.

Directed Graph

Visualization of data in the form of directed graphs allows the visual representation of entities and the relationships amongst them and is well suited for discovery and data navigation. With data presented as interconnected nodes and edges, it is easy to visualize how data is inter-related and facilitates the ability to identify difficult to find relations and see data convergence. Each node and edge are labeled by the technology to help users understand the data more easily. Using an interactive progressive disclosure approach to exposing/hiding nodes makes for intuitive navigation while the ability to select one or multiple

nodes enables easy scoping for activities such as similarity analysis and drilling into details.

Tabular

Visualization of data in tabular form works well for lists of data that generally do not require supporting data in order to gain an understanding. Tabular data visualization is well suited for lists of unrelated data, such as malicious infrastructure (IP addresses, domains, URLs) as well as risk and threat assessments, where the order of the data in table is dependent on the values within a given column, such as a score.

Temporal

Visualization of data in a temporal view (timeline) works well for data that has a time element to it. Here, sequences of events and actions plotted against a timeline show occurrences and duration, such as the remediation of an incident or the behavior of malware.

Geospatial

Visualization of data such as infrastructure like IP address, targeted victims, and command & control sites, as points on a map adds an additional dimension of understanding that helps human better understand how the data relates in terms of location. Connections between these various items can visualize aspects that can help the user better grasp involvement, geo-political associations, and similar concepts.

Publishing and Sharing

The ability to generate and publish findings derived through machine and human analysis is the critical final aspect of a Semantic eScience of Security technology stack. These findings can be as simple as watch lists of indicators for perceived malicious network infrastructure such as IP addresses or as complex as characterization of threat actors and the tactics, techniques, and procedures (TTP) they utilize. In addition to the different types of finding reports, the finding report generation must support the generation in a format that is targeted to the particular audience: such as PDF for human consumption or STIX [12] for machine consumption.

Finally, sharing of generated reports and information with others will need to support a variety of mechanisms such as TAXII [32], email, RSS/ATOM feeds, or even as Linked Data [33]. As with query, it is critical to honor any controls and data markings that govern the visibility of sensitive data both in content and in the ability to share.

Generation

The generation of various findings reports must be guided by an explicit description often referred to as a "profile" which provides a meta-description of what is to be generated, its organization, data to be included, and format that is to be used for the target audience: human, machine, or hybrid. Because not all formats are able to represent the all the different kinds of data, the target format defines what kinds of data is included. For example, OpenIOC [34] only supports the representation of indicators whereas the STIX [12] format provides the ability to express a wide variety of different cyber concepts.

Sharing

Sharing of reports requires the ability to target various mechanisms and protocols depending upon the target audience, the trust associated with a specific audience, and the sensitivity of the data contained within the report. Traditional mechanisms, such as email, RSS/ATOM feeds, and blogs, are currently popular for sharing generated cyber reports for human consumption. More recently, a portion of the broader cyber community has adopted TAXII [32] as a common means over which to share threat intelligence amongst members of a trusted community. TAXII provides both a peer-to-peer as well as spoke-n-hub approach that is well suited for machine-to-machine sharing within a trusted community of subscribers.

Linked Data

An alternative to both traditional and TAXII-based sharing mechanisms is the ability to expose the shared information as part of a collection of interrelated datasets on the web referred to as *Linked Data* [33]. Based on W3C standards and as a foundation for the Semantic Web, Linked Data makes the information available in a common format (RDF) that is easily converted to formats more appropriate for a given environment or device, such as JSON-LD [35].

This approach has the benefit that owners/publishers of data maintain ownership and control of sharing data and with whom. In addition, it provides additional scale since query resolution happens where the data resides instead of moving the data to the point of query.

Instead of every "point of sharing" having to keep a complete copy of everything, a Linked Data approach allows the creation of relationships with data contained within the dataset with other datasets. For example, a TTP defined by STIX [12] could reference an attack pattern contained in the CAPEC [20] Knowledge Repository that references a related vulnerability in the NIST national vulnerability database [36] instead of having various identifiers that require human intervention to find more detailed information.

Supporting the 7 Core Themes of the Science of Security

The Semantic eScience of Security technology stack choices were dictated by what was needed to support the 7 core themes of the science of security [7] discipline while enabling us to develop an organized, cohesive body of knowledge. The goal was to prove that the Semantic eScience technology stack could be customized to support the 7 inter-related core themes of the Science of Security discipline within the operational cyber ecosystem. Here is how each of the core themes are supported within the Semantic eScience of Security solution.

Common Language

This theme is about expressing security in a precise and consistent way. By integrating the enumerations, languages and formats, and knowledge repositories from the Cybersecurity Measurement and Management Architecture [9] the solution supports the "common languages" developed through government, industry, and academia collaboration designed to express security in a precise and consistent way. Additionally, the information coming from the various heterogeneous datasets (STIX [12], MAEC [19], CVE [11], etc) is processed into evidence based statements with clear semantics in the industry standard RDF [22] format.

Core Principles

This theme is focused on foundational principles and fundamental definitions of concepts. The solution supports this theme by leveraging the principles and definitions of concepts contained within the Cybersecurity Measurement and Management Architecture. Additionally, the solution can collect actual instances of data that aligns to those principles and concepts allowing for evidence based knowledge to be produced. The Cybersecurity Measurement and Management Architecture also includes guidance on practical application and use cases that supported by the architecture.

Attack Analysis

This theme is focused on analyzing cyber attacks and understand both the threat actor's actions as well as the actions taken by the defenders. This theme also includes sharing the results of attack analysis in the form of cyber threat intelligence. This theme is supported by being able to collect data in all the various formats from across the Cybersecurity Measurement and Management Architecture [9] including STIX [12], CybOX [18], MAEC [19], CAPEC [20], CWE [37], CVE [11], and the STIX extensions such as YARA [38], OpenIOC [34], Snort [39], and OVAL [40].

The solution is able to take the information and produce evidence based statements with clear semantics and full provenance information. Each evidence based statement represents an analytic "pivot" that the solution can automatically assemble and fuse together into a body of knowledge. The solution provides a number of visualization methods for discovering and looking at cyber attack data including directed graph, timeline/temporal analysis, and geographic analysis.

The solution also includes the MITRE developed vocabulary for characterizing effects on the cyber threat actor [41]. The vocabulary allows for stating claims or hypotheses about the effects of cyber mission assurance decisions on threat actor behavior. Cyber mission assurance decisions include choices of cyber defender actions, architectural decisions, and selections and uses of technologies to improve cyber security, resiliency, and defensibility (i.e., the ability to address ongoing threat actor activities). The vocabulary enables claims and hypotheses to be stated clearly, comparably across different assumed or real-world environments, and in a way that suggests evidence that might be sought but is independent of how the claims or hypotheses might be evaluated. The vocabulary can be used with multiple modeling and analysis techniques, including Red Team analysis, game-theoretic modeling, attack tree and attack graph modeling, and analysis based on the cyber-attack lifecycle (also referred to as kill chain analysis or cyber campaign analysis).

Measurable Security

This theme is about techniques to measure security. This theme is supported in the solution by its support for the entire making security measurable collection that make up the Cybersecurity Measurement and Management Architecture [9]. The solution can consume and produce data in the XML [15] formats of the architecture. The solution includes logic based reasoning and inference capabilities as well as mathematical score systems requires to support the range of measureable security activities.

Risk

This theme is focused on making risk assessments more consistent and less subjective. The Semantic eScience of Security solution enables a number of standardized, repeatable ways to score and measure risk at the application, system, and enterprise level. Much of the work in this field has focused on process and methodology, but risk assessment is still based on individual expertise. The focus of this use case is to make risk assessments more consistent and less subjective using the solution's built in mathematical scoring systems. The below scoring systems, risk analysis framework, threat

assessment and risk remediation analysis methodologies that are supported by the Cybersecurity Measurement and Management Architecture have been integrated into the solution allowing a much deeper technical understanding of the risk based on evidence and customized for each individual organization based on their specific operational assets, business mission, and risk tolerance.

Common Weakness Scoring System (CWSS)

CWSS [42] provides a mechanism for scoring weaknesses in a consistent, flexible, open manner while accommodating context for the various business domains. It is a collaborative, community-based effort that is addressing the needs of its stakeholders across government, academia, and industry. CWSS is a part of the CWE project, co-sponsored by the Software and Supply Chain Assurance program in the Office of Cybersecurity and Communications (CS&C) of the US Department of Homeland Security (DHS).

Common Weakness Risk Analysis Framework (CWRAF)

CWRAF [43] provides a framework for scoring software weaknesses in a consistent, flexible, open manner, while accommodating context for the various business domains. It is a collaborative, community-based effort that is addressing the needs of its stakeholders across government, academia, and industry. CWRAF is a part of the Common Weakness Enumeration (CWE) project, co-sponsored by the Software Assurance program in the office of Cybersecurity and Communications of the U.S. Department of Homeland Security (DHS).

Threat Assessment & Remediation Analysis (TARA)

TARA [44] is a methodology to identify and assess cyber threats and select countermeasures effective at mitigating those threats. When applied in conjunction with a Crown Jewels Analysis (CJA) or other means for assessing mission impact, CJA and TARA together provide for the identification, assessment, and security enhancement of mission critical assets, which is the cornerstone of mission assurance.

The focus of TARA to enable an understanding all the attack vectors based on the assets an organization has and understanding how to remediate those attack vectors with countermeasures in a standardized, repeatable manner based on evidence.

Agility

This theme is focused on being more agile to reflect the more dynamic environment that systems now reside in. The solution provides automation of many of the time critical, labor intensive, and high-skilled tasks that must occur in an effective cybersecurity program that results in overall cost reduction, time savings, and better utilization of scarce resources. In short, it can act as a "force multiplier" enabling less-skilled analysts to be more productive and more highly skilled analysts to focus on the identification of unknown threats to the enterprise.

The evolutionary Semantic eScience of Security solution enables non-disruptive and continuous evolution of data ingestion, enrichment, fusion, and analysis capabilities as new cyber tradecraft techniques are developed or adopted allowing the solution to remain at the forefront of the cybersecurity discipline in the operational cyber ecosystem. The solution's design is based on a data-driven architecture approach, where data drives everything in the solution, leverages an extensible set of orchestrated services which can be added to, augmented, replaced, or removed in order to provide the ability to keep pace with the rate of change in the operational cybersecurity ecosystem.

As with the solution architecture, the conceptual data models used by the platform are designed for evolution through the use of extensible ontologies – semantic models of data and how it interconnects – and by representing data as set of N-triples [45] (e.g., subject-predicate-object) statements. The use of N-triple statements allow virtually anything to be described, designed to form a natural directed graph making interlinking of data easy, and is easy to augment and update allowing any new aspects and existing data to be updated without the need for traditional export-transform-import due to changes in storage schema.

The Semantic eScience of Security solution provides the automation and an infrastructure that is designed to be able to adapt to keep pace with a constantly changing environment at a dramatically reduced cost, while continuously providing new capabilities to users. The automation allows organizations that lack in maturity or have insufficiently skilled staff to still take advantage of the knowledge to provide them the insights required to anticipate potential attacks, while allowing them to continue to increase their maturity. The automation of cyber tradecraft helps organizations bridge the gaps of insufficiently skilled resources and increases the effectiveness of their existing staff, whilst keeping costs down.

Human Factors

This theme tackles factors affecting people's security relevant behavior. This includes both defenders as well as threat actors. Defender human factors could range from secure coding to phishing employees as part of security awareness training to various response times and activities such as incident response. Most large organizations have already been focused on understanding and measuring the human factors of those on the defense side but little has been done to understand and measure the human factors of the offense side, the threat actors.

Where the core theme of attack analysis focused primarily on the tactical behaviors and actions of the threat actor over time in cyberspace, the human factors theme is focused primarily on operational measurements of the humans running the operations.

Operational measurements are generally time based and help us to understand how fast this particular Threat Actor's operations cycle is. Here are a few examples to further our understanding.

- Measure the number of days between attacks/sighting/incidents attributed to a specific threat actor. In other words, how long is the time between attack cycles on average? If the APT campaign attacks every 55 days, that knowledge can help the defender plan to ensure they have right resources in play.
- Measure the time between known events and a specific Threat Actor's activity. For example, if a Threat Actor always attacks within 3 days of Microsoft Patch Tuesday, the defender can use this to plan for the attack. Look to see if there are any reoccurring events, holidays, etc. that act as trigger events or quiet periods for specific Threat Actors.
- Measure the time differential between the observed Threat Actor's Exploit Target (0 day, known vulnerability, or configuration issue) and the date it was publically disclosed. In other words if we look at the Threat Actor's exploit, if it was a 0 day, how many days was the Threat Actor observed using it prior to the vulnerability in the platform being disclosed publically. This would give us a negative number. This is also useful for measuring the Threat Actor's level of sophistication and resources. This can help us to rank Threat Actors based on the risk they pose.
- Measure the time between version number changes over time in the Threat Actor's tools and malware to gain insight to the Threat Actor's engineering and development cycle speed and resources.

- Measure how long a Threat Actor will use a Legend / Sock Puppet / Fake Persona accounts created for registering domains or social engineering before abandoning it for another persona.

There are a significant amount of operational measurements that can be taken and captured in solution if the attack data and information is attributed to a Threat Actor.

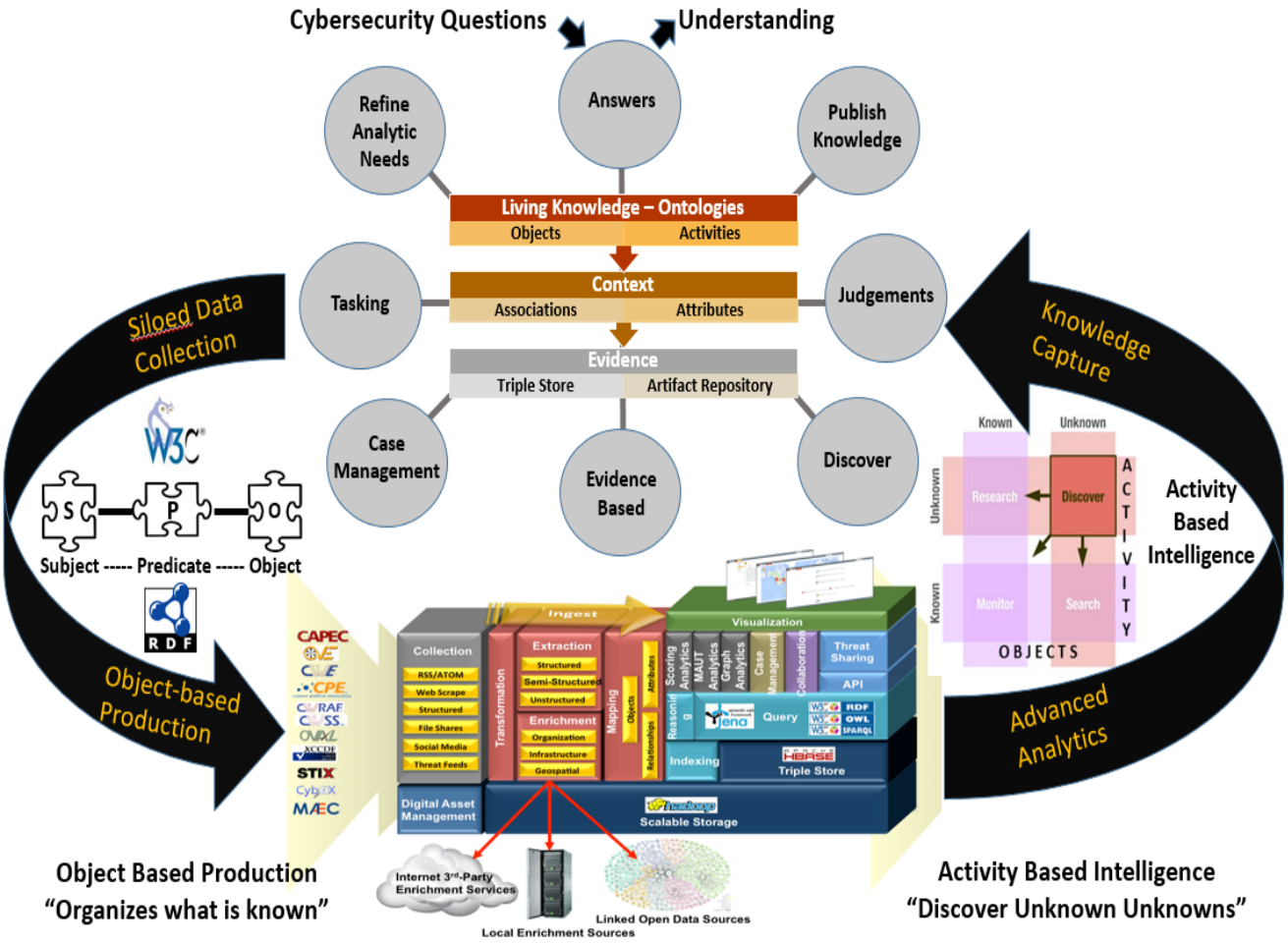
Object-Based Production and Activity-Based Intelligence

Now that we've explained the technology stack for the Semantic eScience of Security solution and how it supports the 7 core themes of the Science of Security discipline, we can look at how humans would use the solution. The goal of this research is to develop an organized, cohesive body of knowledge that informs the field of cybersecurity. Humans will leverage the body of knowledge to answer key questions they have about the cybersecurity of the organization's operational cyber ecosystem. The solution provides a level of automation to help acellorate the time to discovery of the answers they are seeking.

It should be noted that automation isn't about replacing the analysts; it's about helping them be more efficient and effective at their job and moving them "up the stack" to the tasks that require human judgments based on the evidence produced by the automation. As DHS stated in 2011 [8],

> "*Automation frees humans to do what they do well – think, ask questions, and make judgments about complex situations. Automation allows the speed of response to approach the speed of attack, rather than relying on human responses to attacks that are occurring at machine speed. With the ability to execute at machine speed, defenders could get inside the turning circles or decision cycles of attackers.*"

In the below graphic we can see the key human interaction points in relationship to the technology stack. The automation moves the human from "in-the-loop" to a new paradigm of "on-the-loop" meaning the solution isn't waiting for the human to take an action in order for the solution to develop the organized, cohesive body of knowledge. Let's take a closer look at the key interaction points for asking cybersecurity questions and using the solution to discover answers.



Refine Analytic Needs

The first key interaction point is refining the analytic needs required to answer the cybersecurity questions an organization has. Refining the analytic needs to answer a question can be as simple as making sure the right data is available to answer the question or questions being asked so we can find the answers. Consider a few common questions that might be asked that we need to find the answer for. What threat activity are we seeing? What threats should I look for on my network and systems and why? Where has this threat been seen? What does it do? What weaknesses does this threat exploit? Why does it do this? Who is responsible for this threat? What can I do about it? Once the human has refined the analytic needs and identified the data needed to answer the question we can move on to the next interaction area of tasking.

Tasking

Tasking centers around telling the technology what data to collect and analyze based on the refined analytic needs to answer the cybersecurity

questions. Tasking can be set up to support one-time collection or sustained collection as well as controlling how often to perform the collection task. If you are collecting from a source that is updated once a week or once a month then collection doesn't need to hit the source every day and can be configured to collect only when needed. Collection can be inclusive of all data from the source such as collecting all the vulnerability data from the National Vulnerability database or configured to only collect specific instances of data such as only collecting the Whois data for a specific domain. Tasking gives the user the ability to have granular control of exactly what is collected from which sources, at what times, and how often. The data tasked for collection is sent through the Semantic eScience of Security technology stack for automated object-based production of knowledge.

Object-Based Production (OBP) of Knowledge

In terms of Semantic eScience of Security, one might think of ontologies in OWL/RDF as acting as an Object Description Framework (ODF) that enables object-based production of knowledge from each of the common language used in the Cybersecurity Measurement and Management Architecture. These objects are then mapped to the ontology that defines the conceptual semantic object model. The individual semantic object models for each dataset (STIX, MAEC, CVE, etc.) are interconnected by a unifying object model.

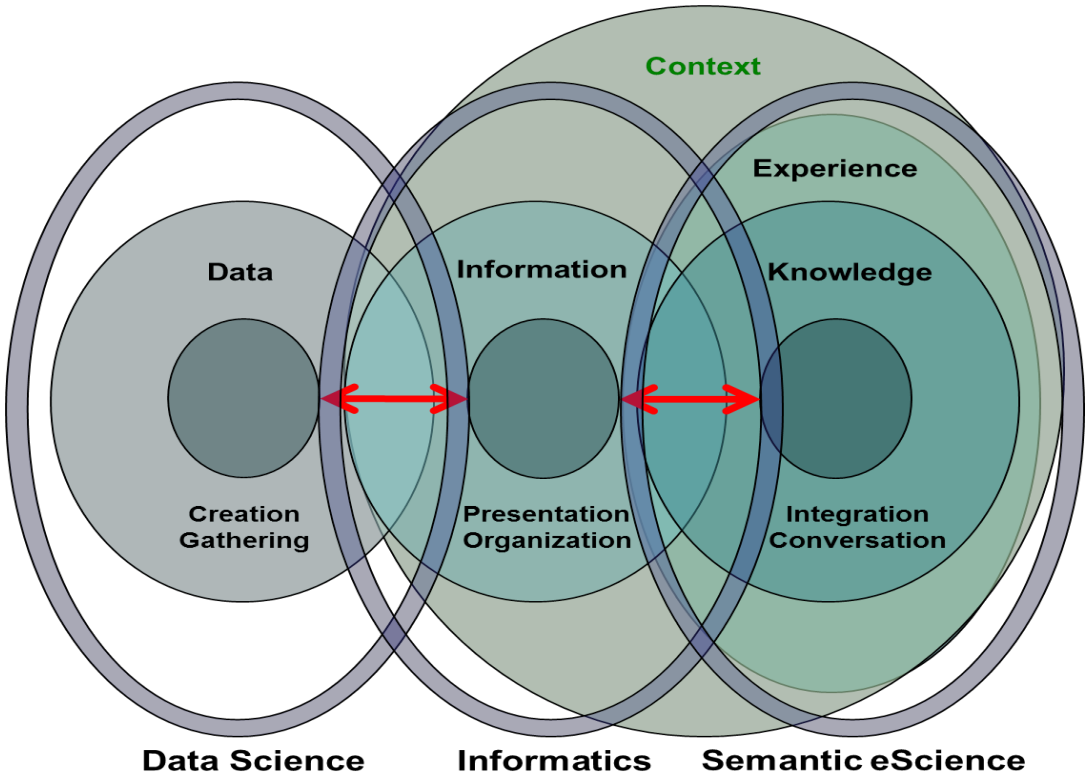
Object-Based Processing (OPB) [5] works by representing these various pieces of data as 'objects' in order to gain greater insights about the nature of the object, the object's attributes, the relationships or associations amongst objects, and observed activity.

Through the modeling of data as objects, attributes, associations, and activities it becomes dramatically easier to understand and categorize objects, often through just an examination of its attributes. It also helps in the identification of behaviors normally attributed to an object. In short, it represents and organizes the data in a manner similar to the way humans think about objects in the natural world. This then enables a focus on the creation and organization of knowledge about what is known so organizations can to do a better job of discovering the unknown through a methodology known as Activity-Based Intelligence (ABI) [46].

In order to get a complete description of an object, it may require multiple statements to be made where each statement describes a specific attribute or association of the object. This collection of statements provides a description of an object and can be easily added to by just adding additional statements as new knowledge about the object is discovered.

Within the Cybersecurity Measurement and Management Architecture [9] at the Information Layer of the cyber ecosystem, cybersecurity information is stored and processed in a linear fashion following a traditional informatics type of approach to organizing information and knowledge. While an informatics approach allows for the capture of context, it has no means to capture the experience of 'working with the data and information' that forms the basis of tradecraft. Experience is also the historical knowledge of what has happened before. As new information and statements enter the body of knowledge they are connected to the knowledge we already gained from experience over time.

Only Semantic eScience of Security provides the foundation required for capturing both cybersecurity context and experience in order to enable the tradecraft to be captured and utilized to automate and formalize that knowledge. This allows the technology to automatically "connect the dots" between the new data and information entering the system and the knowledge already processed. A key enabler to predictability is understanding what happened before.



The Semantic eScience of Security infrastructure focuses on fusing diverse datasets from multiple sources using semantic technology for Object-Based Production (OBP), automated tradecraft based reasoning

rules, and a unifying data model to organize what is "known". There are many diverse types and formats of data, from internal and external sources, that come together to form cybersecurity and threat knowledge. That makes the ability to share that information and reuse it in an automated fashion very difficult. Object-Based Production represents a fundamental shift in how components of information products are stored, increasing productivity and efficiency and allowing analysts to move from searching data and information for knowledge to searching an organized, cohesive body of knowledge for answers.

With the rise of the "Internet of Things", Object-Based Production is key for continued evolution since those "Things" on the Internet are all objects. When you gather the properties of an object, you can identify what sort of thing it is. Aggregating data around objects allows systems to transform data into knowledge. Because of the enormous volume of data available to modern security and intelligence analysts, a degree of automation to connect objects and activity and to build relationships among data is essential to better understanding known threats, discovering unknown threats, and more quickly countering adversaries.

The Semantic eScience of Security related ontologies and Object-Based Production is what enables an organized, cohesive living body of knowledge about the cybersecurity in an organization's operational cyber ecosystem.

Case Management

The next key interaction point is the case management system which provides the tools needed for the human to manage their analytic case workload. Since the solution supports the field of cybersecurity the case management solution needs to support a wide variety of uses across security engineering and operations. An incident response team could manage incident response cases. Security engineers could use the case management solution to track the assets, security configurations, vulnerabilities, weaknesses, attack patterns, and risk scores for each project or program they are supporting. Threat Intelligence analysts could use it for managing threat actors and campaigns to name just a few types of cases that can be supported in the case management system.

The case management solution provides users the ability to manage, track, and be notified about things they care about for their specific use case. Through the case management system, the users can input data via standards based web forms using any of the built-in formats from the Cybersecurity Measurement and Management Architecture [9]. Users can set

up notifications for individual objects so they are notified when a new attribute, association, or activity is observed.

Evidence Based

The Semantic eScience of Security solution develops an organized, cohesive body of knowledge using object-based production to produce individual evidence based statements in RDF [22] format with complete provenance. Each evidenced based statement represents an analytic "pivot" that the technology can assemble together as directed graphs where each object, attribute, association, and activity is labeled with clear semantic meaning.

Discover

Humans can use the solution to discover the evidence that is related to the case they are working on. The solution provides the ability to search for specific instances or types of objects as well searching based attributes, associations, or activities of the objects in the knowledge base. The solution provides a number of different visualizations (directed graph, timeline, geographic, etc.) to help the analyst review and analyze the evidence based knowledge. Since all the knowledge has been organized using object-based production the solution essentially supports activity-based intelligence. Activity-based intelligence [46] allows users to discover the "unknown unknowns".

Activity-Based Intelligence (ABI)

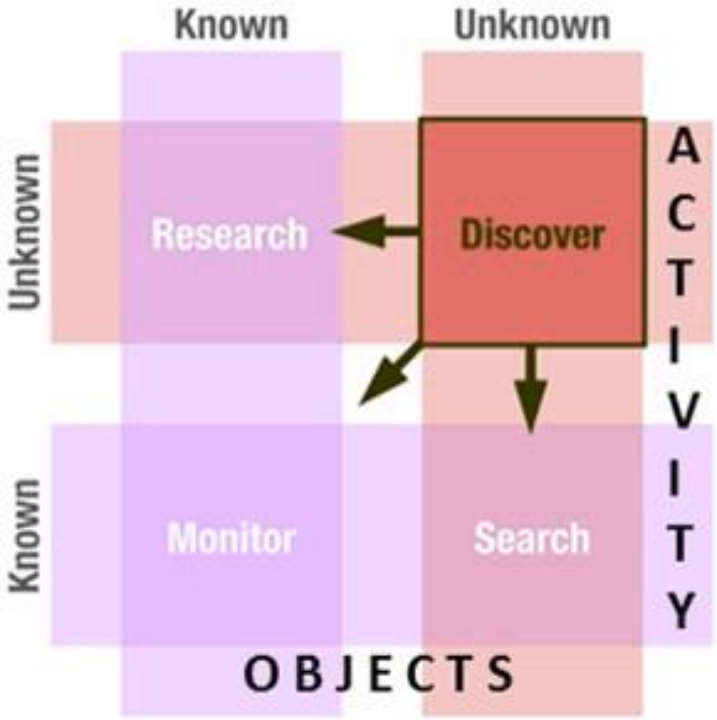
Another key concept in Semantic eScience of Security is enabling activity-based intelligence [46], which uses a multi-source approach to analyze activity and transactional data to develop cyber threat intelligence, drive data collection and resolve what have been called the "unknown unknowns." Activity-based intelligence can be used to develop observable patterns, understand the intent of threat actors, and formulate courses of action.

ABI attempts to meet two challenges with traditional intelligence-gathering. First, there are no clear signatures for and no doctrine governing the activities of non-state actors and insurgents who have emerged as the most important threats to U.S. national security. Second, the volume of big data has become overwhelming and is only set to increase with information sharing efforts.

By researching a known object (person, place, thing, event) we can discover unknown activity, attributes, and associations. By searching for known activity, attributes, and associations, we can discover unknown objects. Object-Based Production [5] organizes what is known and

Activity-Based Intelligence enables the discovery of unknown objects and unknown activities, attributes, and associations across what is known.

ABI is a high-quality methodology for maximizing the value we can derive from "Big Data" that is, making the new discoveries about adversary patterns and networks that give two crucial advantages, unique insights and more decision space to decision makers. ABI also teases out subtle behaviors and relationships between the "known" targets, objects, and networks. To date, ABI has primarily been used in the kinds of government intelligence operations that have defined Iraq and Afghanistan: man hunting and uncovering insurgent networks. Since ABI is more a methodology than a discipline, the intelligence community sees ABI applied to a broad range of problems. We've taking the lead and applied this state-of-the-art intelligence methodology to the field of cybersecurity.



Cybersecurity is filled with all sorts of activity, such as the threat actor moving from stage to stage across the cyber-attack lifecycle or kill chain. Or malware installing on a system, making changes, and connecting to a command and control server. All of this is activity that can be observed and would benefit from ABI analysis.

The goal of Activity-Based Intelligence is to discover and track dynamic activities between objects. Activity-Based Intelligence makes use of data resulting from Object-Based Production as it leverages the conceptual

(semantic) data model that allows associations between objects to be created and represents data as the appropriate objects. Objects are always located in time and space: if you are watching an object as it changes in time and space, that's activity-based intelligence. Activity-Based Intelligence deals more with the transient nature of things. There is some activity going on which we are monitoring and trying to understand what it means and what to do about it.

Object-Based Production supports Activity-Based Intelligence and is also informed by it. There is a feedback loop since the activity itself then becomes an object that can be represented in the knowledge base. Activity-Based Intelligence and Object-Based Production play off one another and are really 2 sides of the same coin.

Judgements

Human judgements about the evidence can be captured in the case management system and become part of the body of knowledge. Users can also make object level annotations to any object in the knowledge base to future provide fine grain judgements or analytic insights as part of the body of knowledge. Judgement, like all other data in the system, would include full provenance information.

Publish Knowledge

Humans can publish knowledge via the standardized formats provided by the web based forms in the case management system to provide answers back to decision makers in order to give them understanding of the situation. Published knowledge can be shared with humans or deployed in machine-readable formats directly to security technology in the user's cyber ecosystem.

Conclusion

Semantic eScience of Security infrastructure is an evolutionary approach that applies a scientific foundation to an organization's cyber ecosystem to support continual evolution of the organization's ability to analyze, assess, mitigate, and monitor the cybersecurity of their cyber ecosystem. A Semantic eScience of Security approach supports the 7 core themes of the Science of Security [7] discipline while enabling Object-Based Production (OBP) [5] and Activity-Based Intelligence (ABI) [46] for cybersecurity. The approach results in the development of an organized, cohesive body of knowledge, based on evidence statements that each represent an analytic pivot with clear semantics and full provenance information.

A Semantic eScience of Security approach has the unique ability to help organizations address critical challenges such as:

- Organizations can't find people skilled in analyzing, assessing, mitigating, and monitoring the threat; When they do find them, Generation X stays about 4 years on the job before leaving, Millennials are only staying about half that if they don't get poached first.
- Organizations are at different levels of maturity in their knowledge, understanding, and application of threat management; maturity is often tied to personnel maturity versus organization maturity (loss of skilled people could set organization maturity backwards).
- As an organization matures, the amount of internal and external data analysts have to go through increases. With the increased focus on information sharing, analysts find themselves swimming in a sea of raw and undifferentiated data that require them to infer the connections and meaning. Engineering, operations, and intelligence teams are often managing cyber threats in isolation without a unified approach that enables continued evolution.
- Organizations struggle with discovering unknown threats and unknown activity (the "unknown unknowns") that require security and intelligence tradecraft to discover.
- Organizations are increasingly looking to implement measurable security that uses standardized, repeatable, quantitative methods for assessing and reducing risk. Much of the work in today's threat and risk assessment fields focuses on process and methodology, but assessment is still based on individual (human) experience and are done using qualitative methods that are descriptive, subjective, or difficult to measure.

In many ways our research "connected the dots" between existing cybersecurity efforts across different government agencies that have been identified by the government as the way forward. From the Science of Security championed by National Security Agency and National Science Foundation, to the Cybersecurity Measurement and Management Architecture championed by the Department of Homeland Security and the Department of Defense, and the revolutionary intelligence methodologies (object-based production / activity-based intelligence) championed by the Intelligence Community and particularly the National Geospatial-Intelligence Agency.

While each of these cybersecurity and intelligence efforts can stand on its own and provide great benefit, bringing these efforts together demonstrates how each agency's investments can see a greater return on investment by working together to develop scientific foundations for the operational

cybersecurity ecosystem. Just like we need infrastructure in areas like Meteorology to understand and predict the weather we need operational cybersecurity science infrastructure to understand and predict events in our cyber ecosystem of the future.

Acknowledgements

I would like to acknowledge the different efforts by government called out in this paper, while this Science of Security research was self-funded it takes advantage of and pulls together years of investment and R&D by various government agencies. I would also like to thank Paul Patrick and Edward Tucker for providing their contributions and input to this paper.

References

[1] "PWC's 18th Annual CEO Survey," 2015. [Online]. Available: http://www.pwc.com/gx/en/ceo-survey/2015/.

[2] "2015 Global Cybersecurity Status Report," 2015. [Online]. Available: http://www.isaca.org/pages/cybersecurity-global-status-report.aspx.

[3] "Job Hopping is the 'New Normal' for Millennials: Three Ways to Prevent a Human Resource Nightmare," 2015. [Online]. Available: http://www.forbes.com/sites/jeannemeister/2012/08/14/job-hopping-is-the-new-normal-for-millennials-three-ways-to-prevent-a-human-resource-nightmare/.

[4] "MasterCard Hits Nike With $5M Cyber Talent Poaching Suit," 2015. [Online]. Available: http://www.law360.com/articles/610735/mastercard-hits-nike-with-5m-cyber-talent-poaching-suit.

[5] D. I. Agency, "Modernizing Defense Intelligence: Object Based Production and Actitity Based Intelligence," 2013. [Online]. Available: https://publicintelligence.net/dia-activity-based-intelligence/.

[6] "TRUSTWORTHY CYBERSPACE: STR ATEGIC PLA N FOR THE FEDERAL CYBERSECURITY RESEARCH AND DEVELOPMENT PROGRAM," 2011. [Online]. Available: https://www.whitehouse.gov/sites/default/files/microsites/ostp/fed_cybersecurity_rd_strategic_plan_2011.pdf.

[7] "Science of Security Joint Statement of Understanding," 2011. [Online]. Available: http://cps-vo.org/node/20575.

[8] "Enabling Distributed Security in Cyberspace – Building a Healthy and Resilient Cyber Ecosystem with Automated Collective Action," 2011. [Online]. Available: http://www.dhs.gov/xlibrary/assets/nppd-cyber-ecosystem-white-paper-03-23-2011.pdf.

[9] "Cyber Security Measurement and Management Architecture," [Online]. Available: http://makingsecuritymeasurable.mitre.org/docs/Cyber_Security_Measurement_and_Management_Poster.pdf.

[10] "Making Security Measurable," [Online]. Available: http://makingsecuritymeasurable.mitre.org/participation/index.html.

[11] "Common Vulnerabilities and Exposures," [Online]. Available: https://cve.mitre.org/index.html.

[12] "Structured Threat Information Expression," [Online]. Available: https://stixproject.github.io/.

[13] "National Council of ISACs," [Online]. Available: http://www.isaccouncil.org/.

[14] "Information Sharing and Analysis Organizations," [Online]. Available: http://www.dhs.gov/isao.

[15] "W3C Extensible Markup Language (XML)," [Online]. Available: http://www.w3.org/XML/.

[16] "The Fourth Paradigm: Data-Intensive Scientific Discovery," 2011. [Online]. Available: http://research.microsoft.com/en-us/collaboration/fourthparadigm/contents.aspx.

[17] "Semantic eScience: Encoding Meaning in Next-Generation Digitally Enhanced Science," 2011. [Online]. Available: http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_part3_fox_hendler.pdf.

[18] "Cyber Observable Expression (CYBOX)," [Online]. Available: https://cyboxproject.github.io/.

[19] "Malware Attribute Enumeration and Characterization," [Online]. Available: http://maec.mitre.org.

[20] "Common Attack Pattern Enumeration and Classification," [Online]. Available: http://capec.mitre.org/.

[21] "Web Ontology Language (OWL)," [Online]. Available: http://www.w3.org/2001/sw/wiki/OWL.

[22] "Resource Description Framework (RDF)," [Online]. Available: http://www.w3.org/RDF/.

[23] "W3C Provenance (PROV) Overview," [Online]. Available: http://www.w3.org/TR/prov-overview/.

[24] "Digital Asset Management," [Online]. Available: https://en.wikipedia.org/wiki/Digital_asset_management.

[25] "JavaScript Object Notation," [Online]. Available: http://json.org/.

[26] "Natural Language Processing," [Online]. Available: https://en.wikipedia.org/wiki/Natural_language_processing.

[27] "RDF 1.1 N-Triples," [Online]. Available: http://www.w3.org/TR/n-triples/.

[28] "SPARQL 1.1 Query Language," [Online]. Available: http://www.w3.org/TR/sparql11-query/.

[29] "Portable Format for Analytics (PFA)," [Online]. Available: http://scoringengine.org/.

[30] "MULTI-ATTRIBUTE UTILITY THEORY (MAUT)," [Online]. Available: http://www.wright.edu/~yan.liu/AdditiveUtility.pdf.

[31] "Analytic Hierarchy Process (AHP)," [Online]. Available: https://en.wikipedia.org/wiki/Analytic_hierarchy_process.

[32] "Trusted Automated eXchange of Indicator Information (TAXII™)," [Online]. Available: https://taxiiproject.github.io/.

[33] "Linked Data - Connect Distributed Data across the Web," [Online]. Available: http://linkeddata.org/.

[34] "OpenIOC - An Open Framework for Sharing Threat Intelligence," [Online]. Available: http://www.openioc.org/.

[35] "JSON for Linking Data," [Online]. Available: http://json-ld.org/.

[36] "National Vulnerability Database," [Online]. Available: https://nvd.nist.gov/.

[37] "Common Weakness Enumeration (CWE)," [Online]. Available: https://cwe.mitre.org/.

[38] "YARA - The pattern matching swiss knife for malware researchers," [Online]. Available: http://plusvic.github.io/yara/.

[39] "SNORT," [Online]. Available: https://www.snort.org/.

[40] "Open Vulnerability and Assessment Language," [Online]. Available: https://oval.mitre.org/.

[41] "Characterizing Effects on the Cyber Adversary," [Online]. Available: http://www.mitre.org/sites/default/files/publications/characterizing-effects-cyber-adversary-13-4173.pdf.

[42] "Common Weakness Scoring System (CWSS)," [Online]. Available: https://cwe.mitre.org/cwss/cwss_v1.0.1.html.

[43] "Common Weakness Risk Analysis Framework (CWRAF™)," [Online]. Available: https://cwe.mitre.org/cwraf/index.html.

[44] "Threat Assessment & Remediation Analysis (TARA)," [Online]. Available: https://www.mitre.org/sites/default/files/pdf/11_4982.pdf.

[45] "W3C N-Triples," [Online]. Available: http://www.w3.org/2001/sw/RDFCore/ntriples/.

[46] "Activity-Based Intelligence Working Group," [Online]. Available: http://usgif.org/community/Committees/_ABI.

About The Cyber Science Directorate

Shawn Riley leads the Cyber Science Directorate. The Cyber Science Directorate mission is to strengthen and support CSCSS, cyberspace security, and cyber resiliency by promoting and coordinating funded cyberspace research and development, innovation across the Centre through creativity, vision, and the delivery of advanced cyber technology solutions for cyberspace. In anticipation of the challenges in securing the cyber systems of the future requires grounded research efforts that nurture and sustain progress to drive and develop realist approach to science of security.

Contact The Author

Shawn Riley
Executive Vice President, Strategic Cyber Science
Shawn.riley@cscss.org
LinkedIn: http://www.linkedin.com/in/shawnriley71 -

About CSCSS

The Centre for Strategic Cyberspace + Security Science / CSCSS is a multilateral, international not-for-profit organization that conducts independent cyber-centric research, development, analysis, and training in the areas of cyberspace, defence intelligence, cyber security, cybercrime, and science while addressing the threats, trends, and opportunities shaping international security policies and national cyberspace cyber security initiatives.

CSCSS, as a strategic leader in cyberspace, works jointly with key partners to address, develop, and define cyber technologies, cyber defence force capabilities, information dominance, and concept operations. We deliver practical recommendations and innovative solutions and strategies to advance a secure cyberspace domain.

CSCSS CENTRE FOR STRATEGIC CYBERSPACE + SECURITY SCIENCE

Contact Us

For more information on the Centre for Strategic Cyberspace + Security Science, its programmes or to find out how we can help you please contact us.

CSCSS / Centre for Strategic Cyberspace + Security
Washington D.C + 571.451.0312
London, United Kingdom +44 2035141784
North America +877.436.6746
Middle East +855.237.8767
Australia +61 2.8003.7553
Email information@cscss.org

CSCSS.org