opentext™ | Cybersecurity

Position Paper

What Makes ArcSight Intelligence Different Part 2

As one of many vendors in the cybersecurity space competing for budget based on buzzwords like AI, machine learning, threat hunting, accelerated threat detection, advanced analytics, big data, deep learning, neural networks, and more, we understand your skepticism when you see and hear those exact words from us.

This series of ArcSight Intelligence position papers is broken down into two parts that demonstrate the key differentiators that customers tell us are unique to us after evaluating dozens of security analytics or user and entity behavioral analytics (UEBA) products.

Table of Contents

Our AI Is Different. Our Approach Is Different

This paper explores ArcSight Intelligence's approach to threat detection using unsupervised Machine Learning in more detail.

ArcSight Intelligence's platform is built from the ground-up to execute unsupervised machine learning algorithms at enormous scale.

ArcSight Intelligence Threat Detection Platform

ArcSight Intelligence's platform is built from the ground-up to execute unsupervised machine learning algorithms at enormous scale. These algorithms extract the available entities (individual users, machines, IP Addresses, web servers, printers, etc.) from log files and observe events that relate to these entities to determine what is normal or expected behavior. As new information comes through the analytics process, it is evaluated against previously observed behavior and dynamically measured statistical peer groups to assess potential risk.

This results in a highly scalable platform that ingests data from security information and event management (SIEM) systems, endpoint software, and other security tools, and processes this data through highly efficient storage and lightning-fast transactional capabilities. To avoid storage issues, ArcSight Intelligence collects metadata and doesn't re-create logs, but it does allow security teams to determine the time frame of log metadata for historic forensic investigation.

Machine Learning for Faster, More Accurate Threat Detection

The following sections provide an overview of the unique combination of mathematical algorithms deployed at scale by ArcSight Intelligence. ArcSight Intelligence's threat detection capability is powered by four stages: Acquire Data, Create Baselines, Detect Anomalies, and Threat Leads.

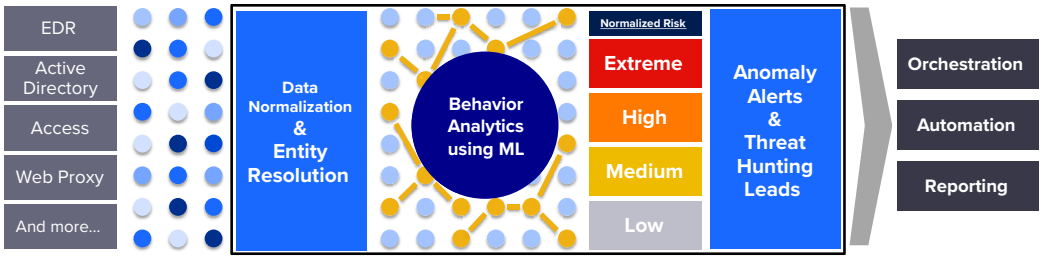

Figure 1. Behavior Analytics using Machine Learning

Acquire Data to Maximize Risk Visibility

The initial stage, Acquire Data, involves the transportation and transformation of data to make it available for analytics. The objective of this stage is to create risk visibility by looking at a wide range of behaviors in the available data. The broader and more diverse the types of behaviors exposed in given data sources, the more holistic a portrait of the enterprise can be generated, enabling more accurate risk assessment. Machine learning will remain valuable against even a single set of data, as this often generates meaningful insight that is impossible to produce from any existing rules- and thresholds-based security tools.

Machine learning will remain valuable against even a single set of data, as this often generates meaningful insight that is impossible to produce from any existing rules- and thresholds-based security tools.

Depending on the threat detection use case, a different set of data sources can be utilized. The following table outlines the different options for data sets that can be used for analytics models for detection of six types of threats: account misuse, compromised account, data staging/theft, infected host, internal recon, and lateral movement.

Log Type	Account Misuse	Compromised Account	Data Staging/ Theft	Infected Host	Internal Recon	Lateral Movement
Authentication	X	X			X	X
Endpoint	X	X	X	X	X	X
File Share	X	X			X	X
IP Repository	X	X	X		X	X
Remote Access	X	X				
Resource Access	X	X			X	X
Web Proxy	X		X	X		

Figure 2. Threat types and the logs used to detect them

The Right Data Is Critical for Accuracy

Data acquisition is difficult because data comes in many different shapes and sizes. Log files can come from different systems and generate data in different ways, so making different datasets compatible with each other is critical to accurate analysis. Imagine if the timestamp of 1:00 p.m. was misinterpreted as 1:00 a.m.? The expected user behavior associated with these two timestamps can be drastically different from each other (i.e., an employee may be expected to work at 1:00 p.m. but not at 1:00 a.m.).

Overcoming these significant yet time-consuming obstacles is one of the major problems of any machine learning and big data analytics project. Typically, data scientists or data engineers spend up to 80% of their time cleaning and preparing data for analysis, as clean data is critical for accurate analysis. Garbage in, garbage out.

For effective analytics and to avoid adding noise to the results, it is important that only data that surfaces behaviors indicative of threats be ingested. ArcSight Intelligence can build a detailed profile with only one data source and identify unusual risky behaviors. For example, Endpoint Detection and Response (EDR) systems provide very rich telemetry to which many of ArcSight Intelligence's behavioral models apply. However, with additional data sources comes the potential for applying more behavioral models. This can lead to a more granular understanding of all entities' normal behavior. Just like a partial fingerprint, while useful, is less revealing than a full fingerprint, ArcSight Intelligence is smart with one data source, but can become even smarter with the right additional data.

Just like a partial fingerprint, while useful, is less revealing than a full fingerprint, ArcSight Intelligence is smart with one data source, but can become even smarter with the right additional data.

ArcSight Intelligence's architecture is able to ingest, stream, parse, and normalize data from multiple sources into a consistent format from which the analytics models can be run. Critically, it can resolve different forms of entity identifiers (e.g., SID, user principle name, user ID) into a single concept of the entity. Otherwise, separate baselines of normal behavior would be maintained for each of an entity's identifiers, potentially leading to false positives or worse, missed opportunities to detect threats.

These capabilities are only reproducible through data analytics experience honed through extensive work with security data. This experience is built into the ArcSight Intelligence architecture, which has been uniquely developed to ingest, stream, parse, and normalize very large quantities of data efficiently and quickly.

Even events that cannot be profiled from behavioral perspectives can be ingested through a data enrichment framework. For example, alerts from third-party malware detection engines can be used to elevate the risk of entities noted as having been exposed to the malware. These alerts should not be profiled for "normal" behavior—they represent suspicious (and one would hope, unusual) occurrences. Nevertheless, they can be considered by ArcSight Intelligence's risk scoring engine as it updates entity risk.

Experience to Recognize the Right Data

This approach delivers the data that ArcSight Intelligence's machine learning models need by performing any required transformations. Time-series data from log files and devices are ingested into ArcSight Intelligence's analytics schema. This "mapping," which defines the relationship between the data source output columns and input variables for each of the supported data types, allows both common and novel data sources to be mapped to the schemas of well-defined data types (e.g., network data types, such as VPN, and web proxy; authentication data sources, such as IAM and Active Directory; repository data sources, such as file share and source code; and endpoint data sources). As the data is ingested or streamed, the input columns are mapped to model features, which are extracted from the data sources.

Taking the universe of possible data sources and transforming them into a finite number of data types provides a balance between two less effective extremes. In one extreme, attempting to support a universe of all possible data settings using a generic anomaly detection method would lead to a tremendous volume of false positives and noise. Such a gross simplification would ignore important semantic differences between data sources—what "bad" looks like in Active Directory is almost certainly different than what "bad" looks like for web proxies. In the other extreme, only supporting very specific data source types without any mapping flexibility would mean the important ability to add multiple data sources would be limited and inflexible. Other authentication systems would not be able to be supported by the Active Directory models, even though, intuitively, there should be authentication models that support both. By creating an intermediate layer that deals with mapped, semantically equivalent data types, you allow for effective data science models concomitant with data source flexibility.

Create Millions of Individual Baselines by Learning

The second stage in the ArcSight Intelligence architecture is the measurement and creation of baselines. As mentioned earlier, ArcSight Intelligence creates unique behavioral baselines for every entity and its relationship to every other entity. This involves measurement of "unique normal" for every user, machine, domain, IP address, share, website, file, project, server, printer, cloud app and resource, as well as measurement of "unique normal" of the interactions between the mix of these entities.

ArcSight Intelligence creates unique behavioral baselines for every entity and its relationship to every other entity. This involves measurement of "unique normal" for every user, machine, domain, IP address, share, website, file, project, server, printer, cloud app and resource, as well as measurement of "unique normal" of the interactions between the mix of these entities.

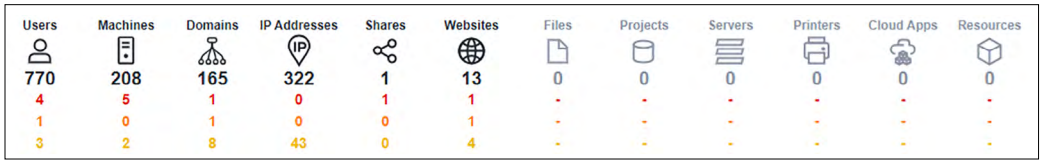

Figure 3. Entity risk breakdown report

Understand "Unique Normal" with Unsupervised Machine Learning

ArcSight Intelligence creates baselines through unsupervised machine learning to mathematically discover patterns. Machine learning involves two stages: training (where the models become smarter and smarter based on new data) and scoring (where the models are used to detect threats and compute risk). Although there are many other AI and machine learning techniques used in cybersecurity today, only unsupervised machine learning can discover relevant patterns without unlabeled datasets. This is different from supervised machine learning, such as deep learning, which requires large volumes of labeled data and is better suited for use cases such as malware detection, where labeled datasets are common.

The purpose of all this is to learn from the data what is normal behavior for every entity in the population (which may include thousands of users, millions of files, hundreds of servers, and thousands of machines), as well as how each entity interacts with other entities. We refer to the result of these observations as “unique normal.”

ArcSight Intelligence achieves this through the power of math—hundreds of algorithms across the various data types automatically learn the normal behavioral patterns for every entity. This learning of baselines results in a statistically driven, comprehensive “fingerprint” for every entity that is automatically learned “online” or in situ within the customer environment. Online learning can be contrasted with “offline” learning, where machine learning is done manually in a data science lab by data scientists. Furthermore, all ArcSight Intelligence’s algorithms take an unsupervised machine learning approach. The algorithms learn the probability density functions associated with each entity’s normal behaviors, the vast majority of which are straightforward univariate or bivariate models.

All ArcSight Intelligence’s algorithms take an unsupervised machine learning approach. The algorithms learn the probability density functions associated with each entity’s normal behaviors, the vast majority of which are straightforward univariate or bivariate models.

A System That Is Principled and Responsible

Having many straightforward models running in parallel has several advantages. First, the algorithms chosen by ArcSight Intelligence to learn unique normal are all well understood and proven statistical learning algorithms—maximum likelihood estimation and kernel density estimation—that have been studied for hundreds of years. This means that they are battle tested, principled, and theoretically sound.

Second, these algorithms can provide human-readable explanations of the specific behavioral anomalies that lead to the detection of a threat. It is very difficult, if not impossible, to generate human-readable explanations from deep learning neural networks, for example. However, such explanations are critical in an industry that requires auditability, traceability, and responsible, transparent use of data.

A Self-Learning System That Automatically Adapts to Changes

The use of hundreds of algorithms tracking different behaviors means that, in practice, a threat can only escape detection if it behaves normally in these hundreds of different ways—a task that becomes increasingly difficult as more data and models are actively deployed within the system. Since these behaviors are learned “online” rather than “offline” in a lab, the learning process not only happens more quickly, but differences between customers or entities within a customer’s population are automatically accounted for by the online learning of “unique normal” for each entity.

This results in a system that is not only precise because it self-learns to customize to the specifics and nuances of each customer environment, but also a system that can react to, adapt to, and keep up with changes in the environment. This self-learning aspect minimizes or even eliminates the manual effort required in traditional rules- and threshold-based systems, where changes in the environment require continuous, manual tuning. By relying on unsupervised approach, there is no need for large and clean sets of labeled data—something which rarely exists in cybersecurity—for effective results. Finally, the use of simple univariate and bivariate models results in low-dimensional learners that converge rapidly and generalize well from customer to customer.

This self-learning aspect minimizes or even eliminates the manual effort required in traditional rules- and threshold-based systems, where changes in the environment require continuous, manual tuning. By relying on unsupervised approach, there is no need for large and clean sets of labeled data—something which rarely exists in cybersecurity—for effective results.

Detect Anomalies with Analytic Models

The third stage of the ArcSight Intelligence architecture is anomaly detection. This is where ArcSight Intelligence applies hundreds of analytic models to measure aberrations in entity behavior and statistically determine which combinations of differences from “unique normal” become prioritized threat leads. The analytic models include many different types of statistical methodologies and machine learning algorithms, such as likelihood estimation, probability density estimation, expectation-maximization, clustering methods like k-means, Gaussian mixture models and power iteration clustering, and dimensionality reduction methods like PCA.

The output of the analytical models is a relative risk score that feeds the threat leaderboard, upon which security practitioners can focus their time for more productive threat detection, threat hunting, and threat investigation.

The models detect and quantify differences between normal, expected behavior, and observed, current behavior to predict the risk of a threat. Analytical models enable the ArcSight Intelligence threat detection platform to detect threats without relying on a team of cybersecurity data scientists to write code and create models, greatly accelerating the time to value of a cybersecurity analytics platform. Often, cybersecurity projects involving large amounts of data require the combination of a security team with deep expertise and a data science team with coding and statistical modeling skills. This is resource-intensive. Additionally, the custom coding required to build the models also takes significant effort, and it typically takes weeks or months before these models are ready to be applied to threat detection. ArcSight Intelligence’s readily available models save months of effort.

Anomalies Measured by Probability Density Functions

By building the probability density functions (PDFs) that describe normal, a distance function between observed behavior (from the data) and expected behavior (from the PDF) affords us a statistical way to detect and quantify the probability of an anomalous and potentially risky behavior. The specific input features, PDF, and distance metric, which together we can term the anomaly model, vary based on the data type and target use case. Model weights allow for different behaviors to be more sensitive to anomalies, while entity weights (entity importance) allow for different entities to be more sensitive to distances.

A Probabilistic Approach Subsumes Binary Nature of Rules and Thresholds

The probabilistic approach means we do not need to ignore behaviors that fall below an arbitrary threshold (and therefore miss low-slow attacks and negatively impact recall) or assume that all behaviors above an arbitrary threshold are always bad (and therefore increase noise and negatively impact precision). In other words, we avoid the binary nature of threshold-based alerts that are prone to false positives. By having the individual models emit continuous numbers that are “fuzzy” in nature, we allow for a more probabilistic, statistical approach. Finally, the model and entity weights allows a customer to provide and integrate business context to the risk scores. For example, a company may have a keen interest in share-related anomalies and therefore increase the share model weights, or it may want to track temp employees within two weeks of the end of their contract more closely by increasing their associated entity weights.

...model and entity weights allows a customer to provide and integrate business context to the risk scores. For example, a company may have a keen interest in share-related anomalies and therefore increase the share model weights, or it may want to track temp employees within two weeks of the end of their contract more closely by increasing their associated entity weights.

Generate Prioritized Threat Leads with Statistical Evidence

The fourth stage of ArcSight Intelligence’s architecture gathers all the statistical evidence across all the models and data sets to compute a single measured risk score for each entity. This is where all the individual risk scores are weighted against each other, to create a prioritized set of threat leads for security practitioners to work from. This stage stores the analytical results and risk scores and makes them available for visualization and instantaneous exploration via ArcSight Intelligence’s simple and intuitive UI.

Apply Statistics to Create Entity Risk Scores

The probabilities emitted from all models are integrated together for each entity across common periods. Intuitively, we are trying to capture the notion of increasing and mounting evidence that something truly risky is going on. The integral is then sent to a logistic or Pareto squashing function that effectively bounds the range of the final value to [0,1]. This final number, which we term the entity risk score, surfaces in the UI as a number between 0 and 100. The shape parameters of the squashing function can be adjusted to create a target distribution across all entity risk scores.

Avoid False Positives

By integrating multiple anomaly models together, we greatly decrease the incidence of false positives, as a large entity risk value becomes only possible with a large number of multiple anomalies (intuitively, a large amount of evidence from multiple sources) or a small number of truly egregious and high-risk anomalies. The guarantee of an output number with a range of [0,1] (or [0,100] within the UI) means that the output risk scores are stable and not unbounded, making orchestration and the development of playbooks more practical. It is difficult to define a playbook if there is no upper limit on the risk score, or if that upper limit changes whenever a new data source or additional model is added.

Finally, the ability to target a specific distribution of entity risk scores means the distribution of low to high entity risk scores can satisfy business, statistical, and aesthetic goals (e.g., a low tolerance for false negatives, a target precision/recall, a "pleasing" distribution of red-orange-yellow-gray, etc.).

By integrating multiple anomaly models together, we greatly decrease the incidence of false positives, as a large entity risk value becomes only possible with a large number of multiple anomalies (intuitively, a large amount of evidence from multiple sources) or a small number of truly egregious and high-risk anomalies.

Next Steps

The following table summarizes the key components of ArcSight Intelligence's approach for user and entity behavioral analytics. As in many situations, it is the execution of the concepts and technologies outlined in this paper that bring true value to an enterprise. CISOs, security architects, and other individuals looking to increase risk visibility and accelerate accuracy and speed of threat detection are encouraged to visit **arcsight.com/intelligence** for further information.

	Stage 1	Stage 2	Stage 3	Stage 4
Purpose	Maximize Risk Visibility	Measure "Unique Normal"	Detect Anomalies	Threat Leads
Action	Acquire Data	Create Baselines	Assess Risk with Mathematical Probabilities	Normalize Entity Risk Scores
Data Science Processes	Data ingestion, normalization, transformation and enrichment	Feature extraction machine learning model online training and validation	Machine learning scoring, weighing expected utility theory-based behavioral risk scoring	Weighing, integration of entity risk score across common entities and time periods, and statistical normalization

Figure 4. Intelligence Key Components chart

Connect with Us

www.opentext.com

opentext™ | Cybersecurity

OpenText Cybersecurity provides comprehensive security solutions for companies and partners of all sizes. From prevention, detection and response to recovery, investigation and compliance, our unified end-to-end platform helps customers build cyber resilience via a holistic security portfolio. Powered by actionable insights from our real-time and contextual threat intelligence, OpenText Cybersecurity customers benefit from high efficacy products, a compliant experience and simplified security to help manage business risk.