Security Big Data Analytics: Past, Present and Future

Security big data analytics (or cyber security analytics) is a rising force that is helping security analysts and tool vendors do much more with log and event data. In the past SIEMs (https://www.exabeam.com/siem-guide/) were limited to manually defining correlation rules, which were brittle, hard to maintain, and resulted in many false positives.

New machine learning techniques can help security systems identify patterns and threats with no prior definitions, rules or attack signatures, and with much higher accuracy. However, to be effective, machine learning needs very big data. The challenge is storing so much more data than ever before, analyzing it in a timely manner, and extracting new insights.

In this Chapter You Will Learn:

How big data analytics helps combat cyber threats - both traditional and advanced analytics techniques.

Key concepts in big data and security - including data science, machine learning, deep learning and User Entity Behavioral Analytics (UEBA).

Three algorithms for detecting anomalies - Random Forest, Dimension Reduction and Isolation Forest.

How SIEMs leverage big data analytics - to provide new security capabilities.

How Can Security Big Data Analytics Combat Cyber Threats?

Traditionally, security technologies used two primary analytical techniques to detect security incidents:

- **Correlation rules—**manually defined rules specifying a sequence of events that indicates an anomaly, which could represent a security threat, vulnerability or active security incident.
- **Network vulnerabilities and risk assessment—** scanning networks for known attack patterns and known vulnerabilities, such as open ports and insecure protocols.

The common denominator of these older techniques is that they are good at detecting known bad behavior. However they suffer from two key drawbacks:

- **False positives—**Because they are based on rigid

predefined rules and signatures, there is a high level of false positives, leading to alert fatigue.

- **Unexpected events—**what happens if a new type of attack is attempted that no one had created a rule for? What happens if an unknown type of malware infects your systems? Traditional systems based on correlation rules find it difficult to detect unknown threats.

Next Gen SIEM

Advanced Threat Analytics Powered by Machine Learning

Addressing unknown risks—including insider threats, which are trickly to detect because they are users legitimately logged into corporate systems—requires advanced analytics. Advanced threat analytics technology can:

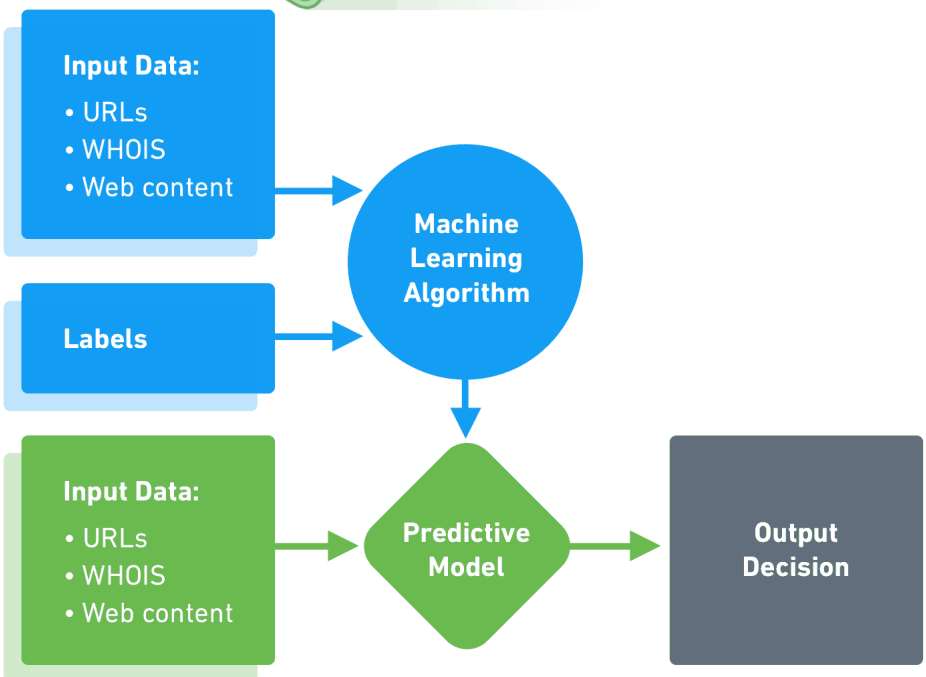
- **Identify anomalies in personnel or device behavior—**creating a model of "normal behavior" for a person, a device or group of devices on the network, and intelligently identifying anomalies, even ones that were not predefined as rules.

- **Detect anomalies in the network—** creating a model of network traffic and intelligently identifying anomalies in traffic. Is something happening that is different than usual for this period or time of day?

- **Perform machine learning based malware detection—**intelligently analyzing binaries transmitted by email or downloaded, even if not flagged by antivirus, to understand if it is a benign program or more likely to be a malicious program.

- **Perform machine learning based intrusion detection**—identifying patterns in network traffic or access controls that are similar to historic intrusions or attacks.

In order to achieve these types of analysis, new analytics methods are needed, as well as access to bigger data than ever before.



Supervised learning for phishing domain detection

Data Science, Machine Learning and Cyber Security

What is Data Science?

Data science is a new discipline that leverages scientific and mathematical analysis of data sets, as well as human understanding and exploration, to derive business insights from big data.

IN THE CONTEXT OF SECURITY:

Data science is helping security analysts and security tools make better use of security data, to discover hidden patterns and better understand system behavior.

What is Machine Learning in Cyber Security?

Machine learning is part of the general field of Artificial Intelligence (AI). It uses statistical techniques to allow machines to learn without being explicitly programmed.

IN THE CONTEXT OF SECURITY:

Machine learning goes beyond correlation rules, to examine unknown patterns and use algorithms for prediction, classification and insight generation.

Important note

Artificial Intelligence (AI) is claimed to be a part of many security analytics solutions. Don't take vendor claims for granted—check what exactly is included in the term "AI". How are vendors building their models? Which algorithms are used? Look under the hood to understand what exactly is being offered.

Supervised vs. Unsupervised Learning

SUPERVISED MACHINE LEARNING

In supervised learning, the machine learns from a data set that contains inputs and known outputs. A function or model is built that makes it possible to predict what the output variables will be for new, unknown outputs.

IN THE CONTEXT OF SECURITY:

Security tools learn to analyze new behavior and determine if it is “similar to” previous known good or known bad behavior.

UNSUPERVISED MACHINE LEARNING

In unsupervised learning, the system learns from a dataset that contains only input variables. There is no correct answer, instead the algorithm is encouraged to discover new patterns in the data.

IN THE CONTEXT OF SECURITY:

Security tools use unsupervised learning to detect and act on abnormal behavior (without classifying it or understanding if it is good or bad).

What is Deep Learning in Cyber Security?

Deep learning techniques simulate the human brain by creating networks of digital “neurons” and using them to process small pieces of data, to assemble a bigger picture. Deep learning is most commonly applied to unstructured data, and can automatically learn the significant features of data artifacts. Most modern applications of deep learning utilize supervised learning.

IN THE CONTEXT OF SECURITY:

Deep learning is primarily used in packet stream and malware binary analysis, to discover features of traffic patterns and software programs and identify malicious activity.

What is Data Mining in Cyber Security?

Data mining is the use of analytics techniques, primarily deep learning, to uncover hidden insights in large volumes of data. For example, data mining can uncover hidden relations between entities, discover frequent sequences of events to assist prediction, and discover classification models which help group entities into useful categories.

IN THE CONTEXT OF SECURITY:

Data mining techniques is used by security tools to perform tasks like anomaly detection in very large data sets, classification of incidents or network events, and prediction of future attacks based on historic data.

What is User Entity Behavioral Analytics (UEBA)?

UEBA solutions are based on a concept called baselining. They build profiles that model standard behavior for users, hosts and devices (called entities) in an IT environment. Using primarily

machine learning techniques, they identify activity that is abnormal or risky compared to the established baselines, and detect security incidents.

The primary advantage of UEBA over traditional security solutions is that it can detect unknown or elusive threats, such as zero day attacks and insider threats. In addition, UEBA reduces the number of false positives because it adapts and learns actual system behavior, rather than relying on predetermined rules which may not be relevant in the current context.

Algorithms for Detecting Outliers and Anomalies

Random Forest

Random Forest is a powerful supervised learning algorithm that addresses the shortcomings of classic decision tree algorithms. A decision tree attempts to fit behavior to a hierarchical tree of known parameters.

For example, in the tree below customer satisfaction is distributed according to two variables, product color and customer age. A decision tree algorithm will inaccurately predict that a different color or slightly different age is a good predictor of satisfaction. This is called *overfitting*—the model uses insufficient or inaccurate data to make predictions on new data.

Random Forest automatically breaks up decision trees into a large number of sub-trees or *stumps*. Each sub-tree emphasizes different information about the population under analysis. It then obtains the result of each sub-tree, and takes a majority vote of all the sub-trees to obtain the final result (a technique called *bagging*).

By combining all the sub-trees together, Random Forest can cancel out the errors of each individual tree and dramatically improve model fitting.

In a security context: Random Forest can help analyze sequential event paths and improve predictions about new events, even when the underlying data is insufficient or improperly structured.

Dimension Reduction

Dimension Reduction is the process of converting a data set with a high number of dimensions (or parameters describing the data) to a data set with less dimensions, without losing important information.

For example, if the data includes one dimension for the length of objects in centimeters and another dimension for inches, one of these dimensions is redundant and does not really add any information, as can be seen by their high correlation. Removing one of these dimensions will make the data easier to explain.

Generally speaking, a Dimension Reduction algorithm can determine which dimensions do not add relevant information and reduce a data set with n dimensions to k, where k<n.

Besides correlation analysis, other ways to remove redundant dimensions include analysis of missing values; variables with low variance across the set; using decision trees to automatically pick the least important variables, and augmenting those trees with Random Forest; factor analysis; Backward Feature Elimination (BFE); and Principal Component Analysis (PCA).

IN A SECURITY CONTEXT

: Security data typically consists of logs with a large number of data points about events in IT systems. Dimensional Reduction can be used to remove the dimensions that are not necessary for answering the question at hand, helping security tools identify anomalies more accurately.

Image Source: Inside Big Data

Isolation Forest

Isolation Forest is a relatively new technique for detecting anomalies or outliers. It isolates data points by randomly selecting a feature of the data, then randomly selecting a value between the maximum and minimum values of that feature. The process is repeated until the feature is found to be substantially different from the rest of the data set.

The system repeats this process for a large number of features, and builds a random decision tree for each feature. An anomaly score is then computed for each feature, based on the following assumptions:

- **Features which are really anomalies** will take only a small number of isolation steps to be far off from the rest of the data set.
- **Features which are not anomalies** will take numerous isolation steps to become far off from the data set.

A threshold is defined, and features which require relatively long decision trees to become fully isolated are determined to be "normal", with the rest determined to be "abnormal".

IN THE CONTEXT OF SECURITY:

Isolation Forest is a technique that can be used by UEBA and other next-gen security tools to identify data points that are anomalous compared to the surrounding data.

SIEM and Big Data Analytics

Security Information and Event Management (SIEM) systems are a core component of large security organizations. They capture, organize and analyze log data and alerts from security tools across the organization. Traditionally, SIEM correlation rules were used to automatically identify and alert on security incidents.

Because SIEMs provide context on users, devices and events in virtually all IT systems across the organization, they offer ripe ground for advanced analytics techniques. Today's SIEMs either integrate with advanced analytics platforms like UEBA, or provide these capabilities as an integral part of their product.

Next-generation SIEMs can leverage machine learning, deep learning and UEBA to go beyond correlation rules and provide:

- **Complex threat identification—**modern attacks are often comprised of several types of events, each of which might appear innocuous on its own. Advanced data analytics can look at data for multiple events over a historic timeline, and capture suspicious activity.
- **Entity behavior analysis—**SIEMs can learn the normal baseline behavior of critical assets like servers, medical equipment or industrial machinery, and automatically discover anomalies that suggest a threat.
- **Lateral movement detection—**attackers who penetrate an organization typically move through a network, accessing different machines and switching credentials, to escalate their access to sensitive data. SIEMs can analyze data from across the network and multiple system resources, and use machine learning to detect lateral movement.
- **Inside threats**—SIEMs can identify that a person or system

resource is behaving abnormally. They can "connect the dots" between a misbehaving user account and other data points, to discover a malicious insider, or compromise of an insider account.

- **Detection of new types of attacks—**by leveraging advanced analytics, SIEMs can capture and alert on zero day attacks, or malware which does not match a known binary pattern.

Exabeam is an example of a next-generation SIEM that comes with advanced analytics capabilities built in (https://www.exabeam.com/product/exabeam-advanced-analytics/)—including complex threat identification, automatic event timelines, dynamic peer grouping of similar users or entities, lateral movement detection and automatic detection of asset ownership.

Prev
SIEM Use Cases

Next
Incident Response and Automation (https://www.exabeam.com/siem-guide/incident-response-and-automation/)

CH01

What is SIEM (https://www.exabeam.com/siem-guide/what-is-siem/)

Components, best practices, and next-gen capabilities

CH02

SIEM Architecture (https://www.exabeam.com/siem-guide/siem-architecture/)

How SIEMs are built, how they generate insights, and how they are changing

READ MORE

CH03

Events and Logs (https://www.exabeam.com/siem-guide/events-and-logs/)

SIEM under the hood - the anatomy of security events and system logs

READ MORE

CH04

UEBA (https://www.exabeam.com/siem-guide/ueba/)

User and Entity Behavioral Analytics detects threats other tools can't see

READ MORE

CH05

SIEM Use Cases (https://www.exabeam.com/siem-guide/siem-use-cases/)

Beyond alerting and compliance - SIEMs for insider threats, threat hunting and IoT

READ MORE

CH06

exabeam (/)

Why Exabeam (https://www.exabeam.com/why-exabeam/)

Get a Demo (/contact/get-a-demo)

SIEM Analytics (https://www.exabeam.com/siem-guide/siem-analytics/)

From correlation rules and attack signatures to automated detection via machine learning

READ MORE

Products (https://www.exabeam.com/product/)

Solutions (https://www.exabeam.com/product/solutions/)

CH07

English

Get a Demo (/contact/get-a-demo)

Incident Response and Automation (https://www.exabeam.com/siem-guide/incident-response-and-automation/)

Security Automation and Orchestration (SOAR) - the future of incident response

READ MORE

Customers (https://www.exabeam.com/customers/)

Resources (/library)

Blog (https://www.exabeam.com/information-security-blog/)

CH08

The SOC, SecOps and SIEM (https://www.exabeam.com/siem-guide/the-soc-secops-and-siem/)

A comprehensive guide to the modern SOC - SecOps and next-gen tech

READ MORE

CH09

Evaluating and Selecting SIEM Tools - A Buyer's Guide (https://www.exabeam.com/siem-guide/siem-buyers-guide/)

Evaluation criteria, build vs. buy, cost considerations and compliance

READ MORE

CH10

SIEM Essentials Quiz (https://www.exabeam.com/siem-guide/siem-quiz/)

SIEM Essentials Quiz

READ MORE

Outsmart The Odds.

1.844.EXABEAM (TEL:18443922326)
info@exabeam.com (mailto:info@exabeam.com)

1051 E. Hillsdale Blvd. 4th Floor
Foster City, CA 94404

© 2021 Exabeam Terms and Conditions (/terms-and-conditions) Do Not Sell My Personal Information (Privacy Policy) (/privacy-policy) Ethical Trading Policy (/ethical-trading-policy) Sitemap (/site-map)