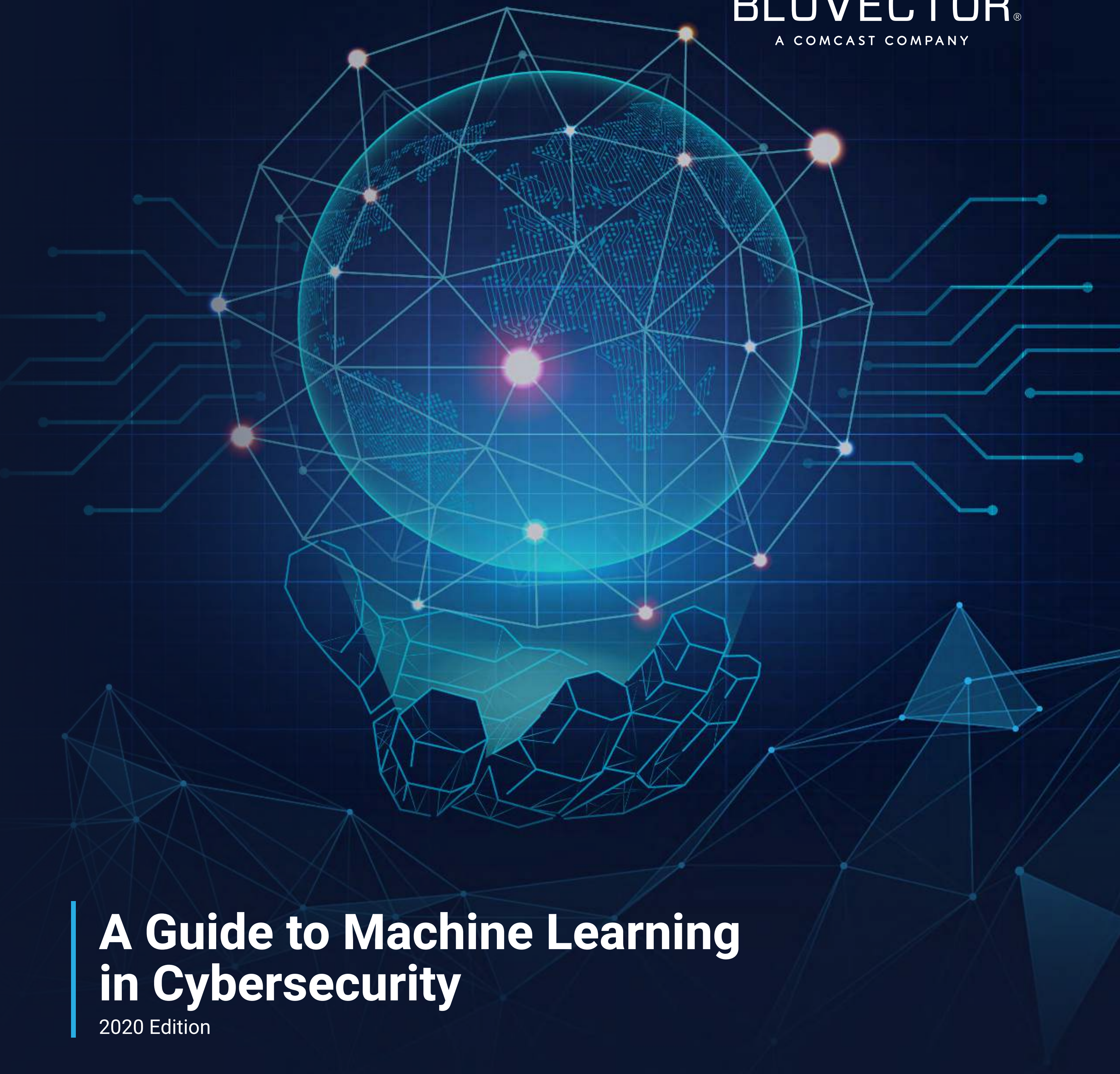
BLUVECTOR®
A COMCAST COMPANY



A Guide to Machine Learning in Cybersecurity

2020 Edition

Cyber threat actors keep reinventing their approaches to successfully evade cybersecurity defenses and wreak more havoc. Machine learning is the latest tool for security teams to get ahead of obfuscated, new and known threats. With every company claiming machine learning capabilities, cybersecurity professionals are confused on how to differentiate among the claims.

READ THIS GUIDE *to understand the different applications of machine learning in cybersecurity and how to determine what will work best for your security defense needs.*

Executive Summary: The Advent of Machine Learning

As threat actors improve their efforts with sophisticated technologies and tactics, enterprises are faced with the reality that their security stack lacks the capabilities to effectively detect, prevent, or respond to all attacks.

Most security solutions (e.g., anti-virus and firewalls) currently rely on methods such as signatures to judge a file or a source's malignancy. Yet, signature-based solutions cannot predict potential threats that haven't already been cataloged as such (i.e., zero-day threats.) And, while zero-day threats have long been seen as a menace, it's the use of old, known threats combined with new obfuscation and evasion techniques that are quickly becoming common and increasingly able to bypass many of the tools designed to detect them. This is a growing trend and the result of simple to access and easy to use tools that can turn even less experienced attackers into successful breachers.

Moreover, most solutions lack the ability to scan the millions of files that enter and exit networks on a daily basis. Compounded by the exponential growth in the number of endpoints from distributed networks, hybrid IT, mobile, IoT and other devices, security teams are often unable to detect malware that could cause serious harm to their networks, leaving them to rely on network traffic analysis to raise alerts long after a successful breach has been executed.

Most current solutions aren't detecting these threats. A new approach is required to keep pace with these new threats, and that is where machine learning can play an important role. Machine learning that uses data and analytics to predict threats and catch obfuscated and zero-day exploits is a fundamental transformation of cyber defenses. But confusion persists on what exactly machine learning is, which has led to it being a check box vendors use to lure customers. The fact is that all machine learning implementations are not the same and much depends on data collection, preparation and method of model development and refinement (i.e.,training).

The intent of this guide is to provide clarity into how machine learning works for combating cyber attacks and explain why not all machine learning implementations are equally effective.

The Problem: Defenses Rooted in Reaction

While threats have changed over the course of the last 30 years, the methodologies and goals used to produce cyber defenses have remained mostly reactive in nature.

For instance, anti-virus (AV) software was a reaction to the Morris Worm of 1988 and subsequent malware throughout the 1990s. AV engines work by using pre-built signatures of known viruses and malware after at least one victim has been reported and the threat identified. Through analysis, a file's signature would be added to a growing list of threats that would allow others to protect themselves. Once created, the AV software company would then be tasked with pushing these signatures out to a client's endpoints in the form of updates.

Similarly, firewalls were built to filter traffic before it could enter the network. Like AV software, firewalls are rules-based, checking for exact matches of hash values, IP addresses, ports and protocols for known threats. But as an exact detection approach, potential victims in a firewall attack would have to meet all the same conditions as those in the rule. So any variance in or obfuscation of the threat would reduce the likelihood of that threat being detected.

The primary drawback of both AV software and firewalls has been their inability to predict and protect against threats that have yet to be seen in the wild (aka "zero-day threats"). During Q3 2019, 50% of all malware detections that avoided signature-based detection were identified as zero-day attacks[1]. Relying on these old-school security technologies to stop malware and other attacks is akin to using a bow and arrow to stop a tank.

[1] "As malware and network attacks increase in 2019, zero day malware accounts for 50% of detections" (Help Net Security December 13, 2019)

The next cybersecurity innovation was the sandbox, virtual machines that performed automated detonation and forensic analysis to identify malicious behavior. Yet, sandboxes cannot typically examine every file. Instead, sandbox solutions must rely on a filter to perform random samplings of traffic, along with signature matches, to determine what to analyze. This technology was effective at finding new threats by observing their behaviors, only when those files happen to be chosen by random sampling. The obvious drawback for sandboxes was their inability to analyze anything beyond files that are identified and sent for analysis. Moreover, threat actors quickly developed new sandbox detection and evasion techniques that would simply hide malicious intent when the malware detected that it was being executed within a sandbox.

The Machine Learning Advantage

Put simply, signature-based cybersecurity approaches cannot recognize well-disguised, altered or new cyber threats that hadn't been seen before. Moreover, the continuing rise in pace and volume of data breaches and hacks underscores the need for a means to stop so-called "unknown unknowns," including such exploits as ransomware, phishing and destructive malware, before they can wreak havoc on an organization.

Machine learning systems trained on a substantial collection of threat samples can recognize new and unknown types of threats because they do not rely on a list of file feature(s) but rather on complex and often subtle combinations of features found in malware, that the system discovered through its training data.

Machine learning systems can help supply answers for these current intrusion detection needs and challenges that dated, signature-based methods cannot:

1. Unique Malware

Most malware detection solutions lack the ability to detect malware in a threat landscape where 70 to 90 percent of all malware is unique to an organization.[2] Detectors trained on adequate malware samples can detect new, never seen before malware entering a network.

2. Network Volume

Sandboxes that safely execute suspicious files are only capable of analyzing a fraction of network traffic and can be evaded or fooled by clever threat actors. Detectors using machine learning algorithms can rapidly examine each file and are nearly impossible to fool.

3. Not Device Specific

An increase in the use of distributed networks, mobile devices, PoS (Point of Sale) devices and IoT (Internet of Things) installations provides threat actors with a long list of ways to attack systems. Machine learning solutions can be specifically trained on a variety of threat vectors.

4. Dwell Time

Organizations have shifted from detecting exploits to hunting for threats already inside their network. The dwell time of the average threat from compromise to discovery was 56 days in 2019[3] — a 22 day drop from 2018 but still way too long for effective remediation. Machine learning reduces intrusion and in more advanced products, can provide a vulnerability score for every file that does enter the system to assist in threat hunting.

Understanding Machine Learning

Machine learning is a tool for finding patterns in data and turning those patterns into rules or knowledge. The mathematical process or algorithm for learning patterns in a dataset is called training. Once a machine learning system is trained, it can be used to recognize patterns, assign classifications or make predictions on new data. There are different types of algorithms that can be employed but a machine learning system's effectiveness is highly dependent on the quality and quantity of the training data.

[2] "Surrounded: Unique malware up 13.7%; AV threats up 523%, backdoors 134%; banking trojans 61%." SC Media UK, Dec. 13, 2019.

[3] Mandiant. "M-Trends 2020: A View From the Front Lines." (Mandiant/FireEye, 2020), 7.

There are three important aspects to consider for each machine learning component:

1. Training Data

Training data is the collected data samples to be used by the machine to learn from. Training data must be of the right type, in sufficient quantity and contain examples of variations that might occur in the wild. In the case of identifying malware, the dataset needs to contain samples of all known malware as well as ample examples of benign files so the machine can learn what benign files look like.

2. Feature Space

The feature space requires specific data characteristics to be included in the learning algorithm. Collecting the data is just the first step. For most types of data, each sample will be pre-processed so that the important features are converted into a numerical representation that can be used in a machine learning model. The critical features must be identified and encoded. In the example of malware, this could include file names, size and bit pattern. The selection of features can impact both accuracy and speed of the analysis.

3. Learning Algorithm

This component will perform the “learning” in building a model. The objective of the model is to make a prediction on unseen data based on the training data that the algorithm has previously seen. The choice of model depends on the type of prediction that is to be made as well as the nature of the data collected. Some of the most common algorithms include:

- K-means
- Deep Learning
- Decision Trees
- Random forest
- Regression
- kNN

Machine learning algorithms primarily fall into two[4] categories: “unsupervised” and “supervised.” The choice of category is dictated by the knowledge of the training data and its features. To understand how a machine learning algorithm is applied, it is important to understand how these categories differ in their approach and application.

Supervised vs Unsupervised Machine Learning

Algorithms in a machine learning model largely determine how the features or characteristics of the training data are used to make predictions of newly observed samples. The two types, supervised and unsupervised machine learning, are both used in cybersecurity but for different purposes. What follows is a discussion about how the learning process works for each category, primary use case scenarios and respective strengths and limitations of each.

Unsupervised Machine Learning

Unsupervised learning seeks to make predictions by grouping data according to similarities the algorithm has found amongst the training instances. Unsupervised machine learning’s job is to devise a method for sorting these training instances. How the training instances are sorted depends on:

- The features used to describe the instances to the algorithm;
- The parameters of the sorting process, such as how many groups are desired, minimum group size or density or whether the sorting should be hierarchical.

Figure 1 shows how a learning algorithm might sort some training instances of colored shapes. Three different sample models are shown illustrating how multiple reasonable groupings can be formed. One major drawback of unsupervised learning is there is no way to predict or generate descriptions of groups the algorithm forms. An unsupervised model may find some similarity between instances placed into the same

[4] In addition to supervised and unsupervised learning, there are two other categories of machine learning that are much less common. First,semi-supervised learning is a hybrid approach in which relatively few training instances are labeled but most are not. Second, reinforcement learning is a variant of supervised learning concerned with teaching software agents how to interact with complex environments by providing rewards and penalties over time (it’s akin to learning by trial-and-error).

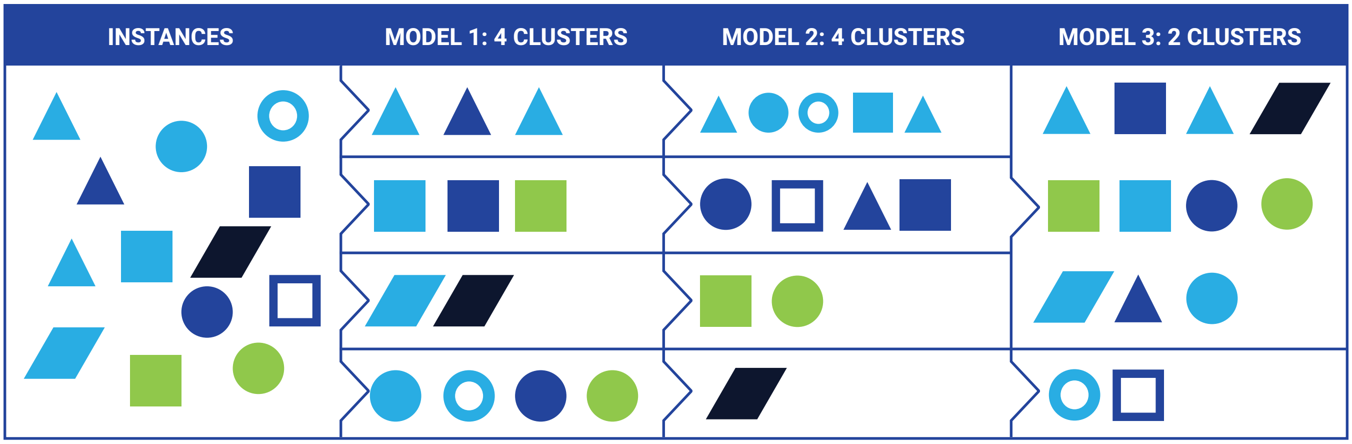


FIGURE 1. Examples of three different sorting models that could be produced by an unsupervised machine learning algorithm.

group or cluster, but the users may not readily understand why or what makes them similar.

Unsupervised Learning for Anomaly Detection

The most common application of unsupervised machine learning in cybersecurity is anomaly detection. Several cybersecurity approaches apply unsupervised machine learning to a wide variety of user behavior and network traffic analyses. Most of these approaches use a form of anomaly detection, making reasonable conclusions about instances by grouping them as normal or abnormal This grouping, such as determining whether something is normal or abnormal, can be dependent on the context in which the determination has been made.
Nearly all anomaly detection cybersecurity solutions focus on the behavior detected. This relies on the assumption that perimeter defenses cannot succeed at detecting anything new, and thus, anomaly detection will alert on an attack post-breach as it begins performing malicious activity.

As we know from looking at the cyber kill chain (See Figure 2), reconnaissance and ultimately exfiltration of sensitive data or confidential information may occur days, weeks or even months after initial malware delivery, lying dormant until that point. Moreover, normal is not immutable. New network processes, flows and behaviors happen all the time which can lead to false alerts or false positives that put strain on a security team.

Supervised Machine Learning Explained

Supervised machine learning algorithms work by building a classification model from “labeled” training data. Labeled here means that someone has assigned a category of interest to each training instance. The label is dependent on the classifications. The labels, the training instances and the desired task are inextricably linked.

Labeling may seem trivial, but in general, it can be a difficult, expensive and time-consuming process to attain enough training instances of each label to produce highly accurate machine learning models. Subject matter experts are used to manually look at every training instance to determine its label. This can be a labor intensive activity that not all cybersecurity companies claiming to do machine learning have the resources or experience to perform diligently on their training data.

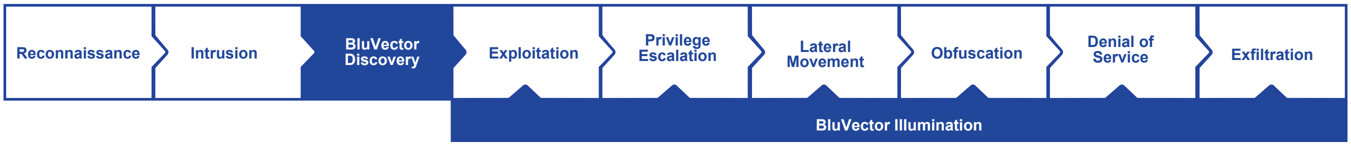


FIGURE 2. An example of events in a kill chain to illustrate where BluVector helps.

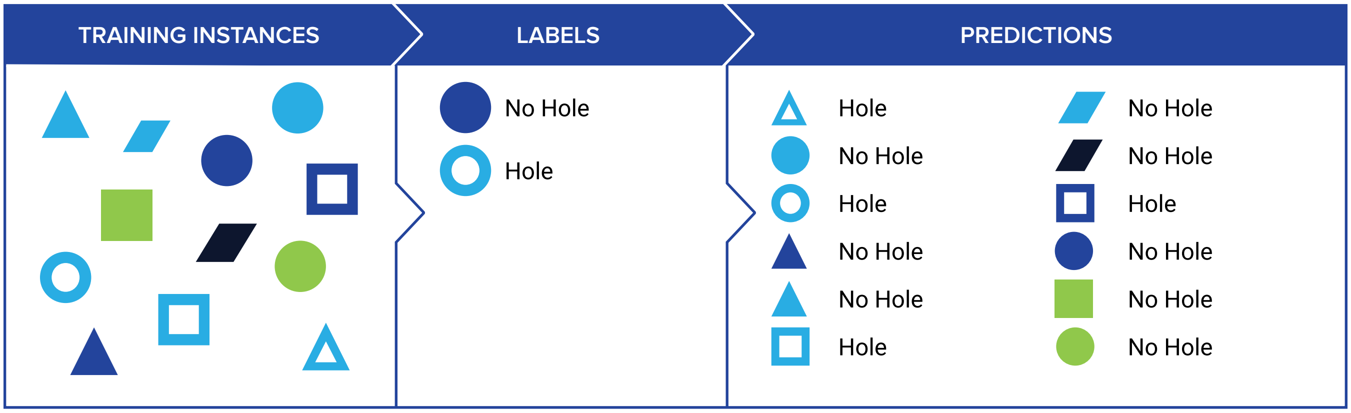


FIGURE 3. Example of how a supervised learning algorithm could sort training instances.

One way to think about supervised learning is that during the learning process, a “teacher” will tell the algorithm when it is predicting labels correctly and when it is making mistakes. That teacher is the subject matter expert who labeled all the training instances used by the machine learning algorithm. The machine uses the labeled examples to learn to make accurate predictions. In Figure 3 above, the teacher trains the machine to predict whether a sample has a “hole” and labels each sample with its prediction.

Supervised Machine Learning for Pre-Breach Detection

Supervised machine learning allows models to be based on samples that have been carefully analyzed, curated and determined to be benign or malign.

Supervised machine learning is looking for characteristics or features similar to how a signature will look to match a byte sequence in a file to determine if it’s something malicious and previously seen. This is where the similarities end. The differences from here are significant and are what makes supervised machine learning incredibly accurate in its application.

The supervised machine learning algorithm, in its analysis of hundreds of thousands of training instances is capable of identifying combinations of characteristics or features — even very small and rare — across all layers of a file. For example, a Windows executable may have combinations of characteristics across the metadata, code and raw bytes layers. Taking it a step further, unlike signatures, a machine learning model can also consider the absence of features as part of its prediction making.

Conclusion

The advancement and proliferation of cyber threats illustrates the difficult landscape organizations must now operate in and the threats they will likely face in the near future. Cybersecurity systems that utilize machine learning have a distinct advantage but not all machine learning products are equally effective. Data collection, feature selection and algorithm expertise all play a significant role in the performance of a machine learning system.

BluVector’s long-term experience and patented approach in machine learning offers organizations award winning detection and performance and gives their security teams an upper hand in the fight against cyberattacks — both now and well into the future.

Advantages of BluVector's Machine Learning Approach for Cybersecurity

Challenge	BluVector's Solution
Zero-Day Malware Detection	BluVector examines hundreds of machine-learned file features to identify malicious hidden content or not previously seen malware.
Comprehensive Training Data	BluVector uses a proprietary malware dataset which began from its work on a cyber-intelligence project with the U.S. government nearly a decade ago. This extensive, industry-leading dataset is continuously updated and validated with the latest threat intelligence data giving the customer the latest and most effective detection capability right out of the box.
Enterprise and Industry-Specific Malware Classification	BluVector provides the capability for a customer's security team to add their own enterprise-specific files to the training data to re-train and customize each detector if desired.
Network Volume	BluVector's Machine Learning Engine is scalable and highly optimized for hardware performance so that every file coming into a network can be classified in nearly real-time without delay or impact on business operations.
Expanding Attack Surface	BluVector's patented approach to intrusion detection separately trains for each file type to account for variations in attack surfaces and file specific vulnerabilities.
Reducing Dwell Time	BluVector provides a threat score for every file as that file enters the system and can assist immediate threat hunting activities and analysis.

About BluVector, A Comcast Company

As a leader in advanced threat detection, BluVector is empowering security teams to get answers about real threats, allowing businesses and governments to operate with greater confidence that data and systems are protected.

BluVector MLE

BluVector MLE is a patented supervised Machine Learning Engine that was developed within the defense and intelligence community to accurately detect zero-day and polymorphic malware in real time. Unlike unsupervised machine learning, which is leveraged by most security vendors today, BluVector MLE algorithms were pre-trained to immediately identify malicious content embedded within common file formats like Office documents, archives, executables, .pdf, and system updates. The result: 99.1%+ detection accuracy upon installation.

BluVector SCE

BluVector SCE is the security market's first analytic specifically designed to detect fileless malware as it traverses the network. By emulating how the malware will behave when it is executed, the Speculative Code Execution engine determines, at line speed, what an input can do if executed and to what extent these behaviors might initiate a security breach. By covering all potential execution chains and focusing on malicious capacity rather than malicious behavior, the analytic technology vastly reduces the number of execution environments and the quantity of analytic results that must be investigated.

BLUVECTOR®

A COMCAST COMPANY

www.bluvector.io