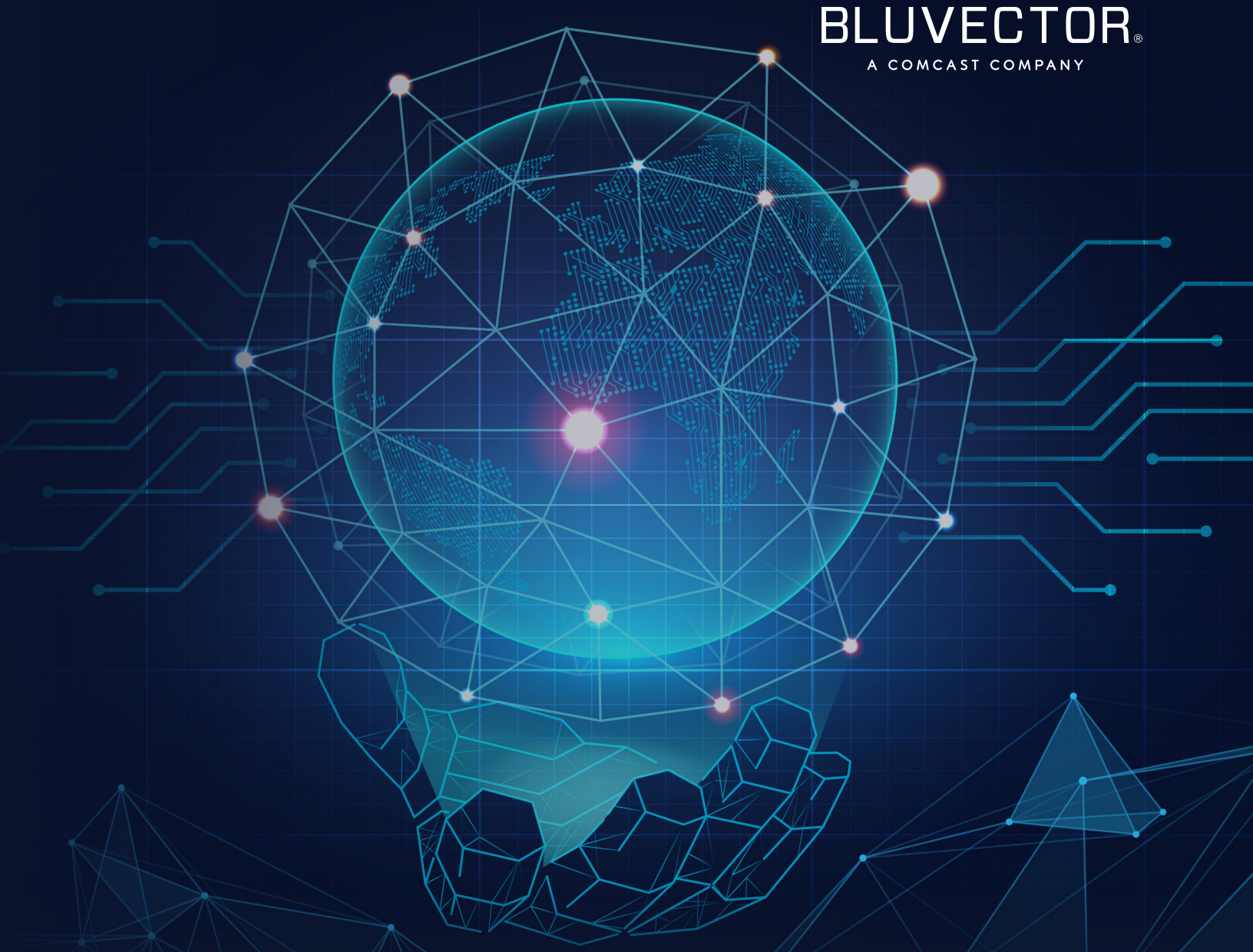
BLUVECTOR®
A COMCAST COMPANY



A Guide to Machine Learning in Cybersecurity

Cyber threat actors have reinvented their approach time and time again to evade defense and wreak more havoc. Machine learning is the latest tool for security teams to get ahead of obfuscated, new and unknown threats. But everyone says "machine learning," as though it's all the same, leaving cybersecurity professionals confused on how to differentiate among the tools.

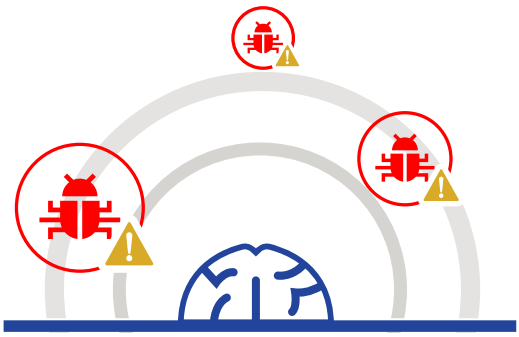
READ THIS GUIDE *to understand the different applications of machine learning in cybersecurity and how to approach deciding what will work best for your security defenses.*

Executive Summary: The Advent of Machine Learning

As threat actors ratchet up ever more sophisticated and complex technologies and tactics, enterprises are grappling with the reality that their current security stack lacks the capabilities to effectively stop attacks. Tools to obfuscate malware are widespread, easy to get and easy to use by anyone. Many organizations face the same ordinary threats, but the evasion techniques available to bypass defenses make it increasingly difficult to stop them.

The majority of security technologies rely on methods such as signatures to judge a file's malignancy and therefore cannot predict potential threats that haven't already been cataloged as such. Moreover, most technologies lack the ability to scan the millions of files that enter and exit networks on a daily basis. This is compounded by the exploding number of endpoints from distributed networks, hybrid IT, mobile devices, IoT and other devices.

As a result, machine learning has become the latest buzzword among cybersecurity experts. Its use of data and analytics to predict threats and catch zero-day exploits is a fundamental transformation of cyber defenses. But confusion reigns on what exactly machine learning is, which has led to it being a check box vendors use to lure customers. The intent of this guide is to provide clarity into what the different types of machine learning actually are and help refocus the discussion on how they work and their different value in combating cyberattacks.



Machine Learning's use of data and analytics to predict threats and catch zero-day exploits is a fundamental transformation of cyber defenses.

What Is Machine Learning?

Machine learning is a technology-based process where computer programs make accurate predictions through exposure and analysis of data. Machine learning works much the way a human mind works in the way it learns to make decisions or predictions.

Mathematician and computer scientist Alan Turing argued that machine learning should simulate the process in which a child's mind learns. Turing broke down the process into three components:

1. The initial state of the mind at birth;
2. Its formal education;
3. Additional experiences that mind has experienced that do not fall under formal education.[1] In understanding the application of machine-learning algorithms to cybersecurity, it's not necessary to understand the algorithm itself—many solutions may even apply several within a category.

[1] Turing, Alan M. Computing Machinery and Intelligence. (Mind, 1950)

Similar components may be considered in how a machine learns and have significant impact on the performance and accuracy of the resulting model:

1. **Feature Space:** The characteristics to be included in the learning algorithm. For example, the feature space to distinguish dogs and cats may consider the volume of the nose, length of the muzzle, pupil shape and how the tail moves.
2. **Training Data:** The samples from which the machine will learn. In the case of dogs and cats, the data scientist building the machine-learning model may want it to distinguish between the two. The training data would include the defined characteristics or features in the feature space. For instance, the data of a sample dog may include a 9.6cm nose, 8cm muzzle, round pupils and a tail that moves only at the base.
3. **Learning Algorithm:** The component that performs the "learning" in building a model. The objective of the model is to make a prediction on newly provided data based on the training data the algorithm has observed. In building a model, the machine may choose from a large number of algorithms, including:

- K-means
- Apriori algorithm
- Regression
- Decision Trees
- Random forest
- kNN

Instead, In understanding the application of machine-learning algorithms to cybersecurity, it's not necessary to understand the algorithm itself — many solutions may even apply several within a category. Instead, it's important to look at the category or type of the algorithm used as there are two[2] primarily used in cybersecurity: "unsupervised" and "supervised." Selecting which to apply depends on the available training data and the desired outcome of the resulting machine learning model, as is often the case in the application within cybersecurity. This paper will review how these categories differ in their approach and application to educate readers about making a more informed decision in choosing the type of machine learning that is right for your cyber defense.

The Problem: Defenses Rooted in Reaction

Since the dawn of the public internet about 30 years ago, cyber defenses that were built to detect malicious threats have been reactive in nature. Anti-virus (AV) software was a reaction to the Morris Worm of 1988 and subsequent malware throughout the 1990s. AV engines work by building signatures of known viruses and malware and pushing them down in the form of updates to endpoint machines via patches, a process which cannot take place without first knowing the malware exists.

Similarly, firewalls were built to filter traffic before it could enter the network as packets. Like AV software, firewalls are rules-based, checking for exact matches of hash values, IP addresses, ports and protocols for known threats. The primary drawback of both AV and firewalls has been their inability to predict and protect against threats that have yet to be seen in the wild. As of December 2016, around **323,000 new malware files were being identified daily** (up from 70,000 files per day in 2011)[3]; relying on these old-school security technologies to stop malware and other attacks is akin to using a slingshot to stop a tank.

[2] 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).

[3] Seals, Tara. "323K Malware Files Are Detected Daily." (Infosecurity, December 7, 2016)

The next cybersecurity innovation was the sandbox, virtual machines (VMs) that performed automated detonation and forensic analysis to identify malicious behavior. Sandboxes rely on a filter to perform random samplings of traffic, along with signature matches, to determine what to analyze in the sandbox. This technology was effective at finding some new threats by observing their behavior, particularly with the addition of random sampling.

Windows	57 6 9 6e 64 6f 7 7 73 2 0
WiNdOwS	57 69 4e 64 4f 77 53 2 0
wINDOWS	77 4 9 4e 44 4f 5 7 53 20

FIGURE 1: The byte sequence is easily altered without meaningful changes to the content.[4]

The obvious drawback to sandboxes was their inability to analyze anything beyond files sent for analysis through the use of signatures, pattern matching or added random samples. Moreover, threat actors quickly developed sandbox evasion techniques.

Environmental checks for signs of sandbox, including:

- Common analysis tools
- Mouse clicks and movement
- Content on the clipboard
- CPUs with multiple cores
- Disk size
- Arbitrary sleep statements
- Exploits targeting specific versions of Adobe Flash Player or PDF viewers
- Requiring user interaction before malware launches
- Requiring target-specific data, such as a cookie, certification or second file to execute malware payload

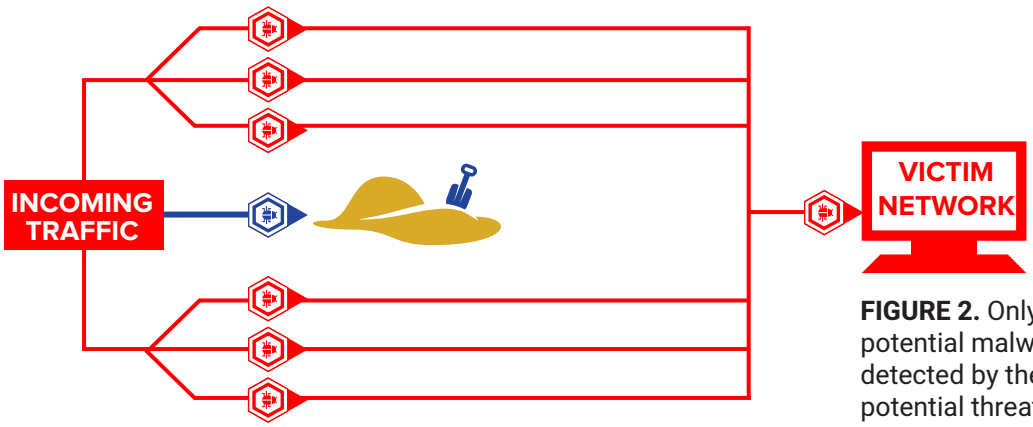


FIGURE 2. Only a small subset of potential malware is inspected or detected by the sandbox. Many potential threats continue into the victim network untouched / undetected.

[4] Signatures today are more complex than simply relying on a single byte sequence; however, they are still easy to bypass by deviating from the signature pattern. Shown is an overly simplistic example of how changes in character case results in very different bytes yet the word "windows" is left intact — just as complex code can be altered yet yield the same, malicious behavior.

The Impetus Behind Machine Learning

Put simply, past cybersecurity technologies cannot anticipate new cyber threats. Along with the limitations already discussed, here is a list of drawbacks faced when relying solely on these dated methods:

- Inability to detect new malware in a threat landscape where 70 to 90 percent of all malware is unique to an organization;[5]
- Sandboxes performing more complex analysis (beyond basic signatures and matching) are only capable of analyzing a fraction of network traffic and are commonly evaded;
- Organizations have shifted from trying to detect exploits to trying to find threats already inside the network using such techniques as threat hunting. The dwell time of the average threat from compromise to discovery was 99 days in 2016[6] — an almost 50-day drop from the previous year but still way too long for effective remediation;
- An expanding attack surface area resulting from an increase in distributed networks, mobile devices, PoS (Point of Sale) devices and IoT (Internet of Things) installations, providing threat actors with a cornucopia of ways to attack systems.

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 your organization.

Understanding the Two Types of Machine Learning

As discussed in the introduction, 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, both unsupervised and supervised machine learning, are used both 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 reasonable conclusions the algorithm has given to the training instances. Unsupervised machine learning's job is to devise a method for sorting these training instances. There is no "right" conclusion but rather reasonable ones. 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.

[5] Verizon. "2015 Data Breach Investigations Report." (Verizon Enterprise Solutions, 2015), 10
[6] Mandiant. "M-Trends 2017: A View From the Front Lines." (Mandiant/FireEye, 2017), 47.

Figure 3 below shows how a learning algorithm may sort some training instances of colored shapes. Three different models are shown illustrating how multiple reasonable groupings can be formed from even this trivial example. One major drawback of unsupervised learning is there is no way to predict or generate descriptions of groups the algorithm forms. So, a user may know there is some similarity between instances placed into the same group or cluster, but they may not understand why or what makes them similar.

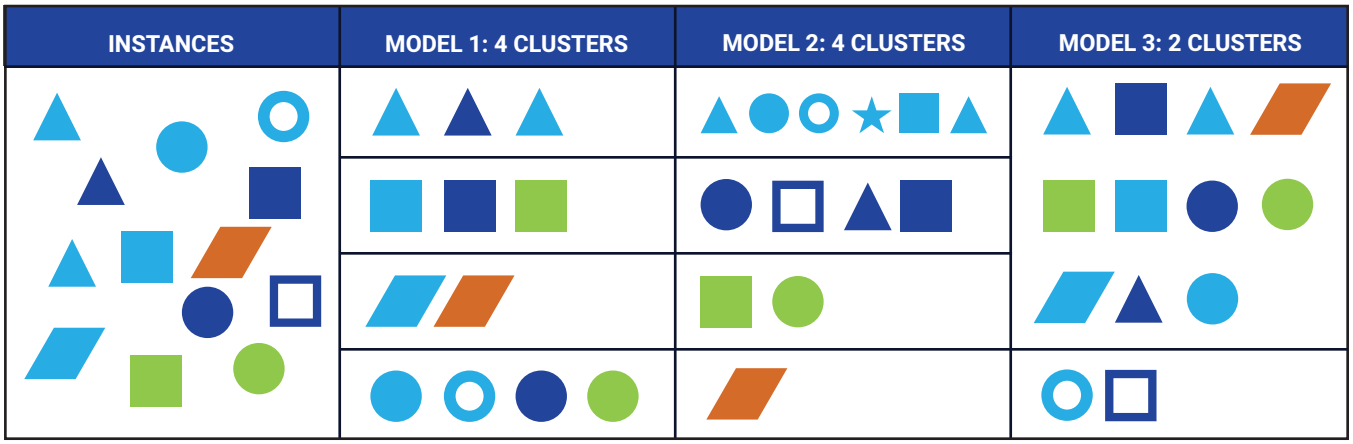


FIGURE 3. Example of how an unsupervised learning algorithm could sort training instances.

Anomaly Detection, Stopping Bad From Getting Worse

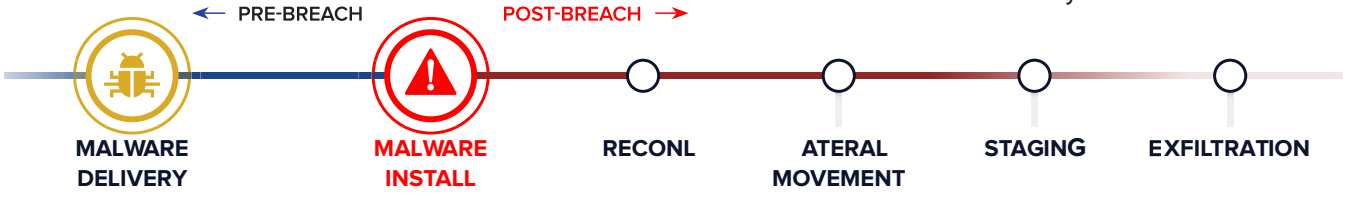
The most common application of unsupervised machine learning is anomaly detection. Several cybersecurity products apply unsupervised machine learning to a wide variety of user behavior and network traffic log analyses. These products are almost all performing a form of anomaly detection, making reasonable 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 environment in which the determination has been made. For example, a cat at a dog show may be labeled as abnormal, while a cat at an animal shelter may be considered normal.

Anomaly detection typically takes about two weeks to observe traffic on a network to identify it as "normal." Over the course of this time, it will make reasonable conclusions about the data, and anything new being observed will be placed into the clustering created by the algorithm.

Nearly all anomaly detection cybersecurity products 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 adversary post-breach as it begins performing malicious activity. As we know from looking at the cyber kill chain (See Figure 4), 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.

FIGURE 4. Cyber kill chain of events



Supervised Machine Learning

Supervised machine learning algorithms work by building a model based on “labeled” training data. Labeled here means that someone has assigned a category of interest to each training instance. The best label is dependent on what task the machine learning program is seeking to accomplish. For example, if one wants a program to distinguish between cats and dogs, labels of “cat” and “dog” are sufficient. If one wishes to distinguish among breeds of cat or dog, labels such as Pug, Dalmatian, American Shorthair, Siamese, Collie and German Shepard are required (See Figure 5).

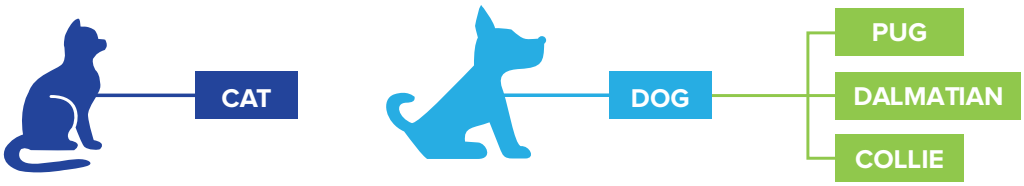


FIGURE 5. The difference in specificities of labels.

If you instead want to distinguish between young and old animals, you would choose labels related to age, such as newborn, adolescent, adult and senior. The labels, the training instances and the desired task are inextricably linked.

Labeling may seem trivial for something as familiar as dogs and cats, 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. The number of training instances required can be in the tens or hundreds of thousands.

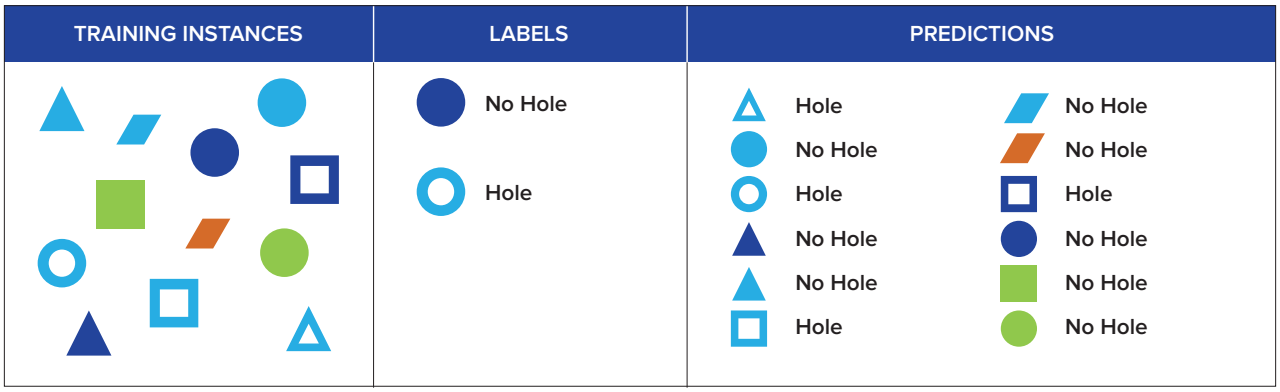


FIGURE 6. Example of how an unsupervised learning algorithm could sort training instances.

One way to think about supervised learning is that during the learning process, a “teacher” is available that will tell the algorithm when it is predicting labels correctly and when it is making mistakes. That teacher is the subject matter expert (or experts) who labeled all the training instances used by the machine learning algorithm. The machine uses the teacher to make more accurate predictions. In Figure 6 above, the teacher trains the machine to predict whether a sample has a “hole” and label each sample with its prediction.

Supervised Machine Learning, 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. Basing the learning on the ground truth of the content samples themselves allows for the production of models relevant across a wide variety of industries and customers, reducing noise in the training process and providing detection functionality right out of the box (no onsite training needed).

The way supervised machine learning models make their predictions has more similarities to traditional signature-based approaches than it does the anomaly detection capabilities of unsupervised machine learning applications. Like a signature, supervised machine learning is looking for a characteristic or feature similar to how a signature will look to match a byte sequence in a file to determine it's something malicious 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.

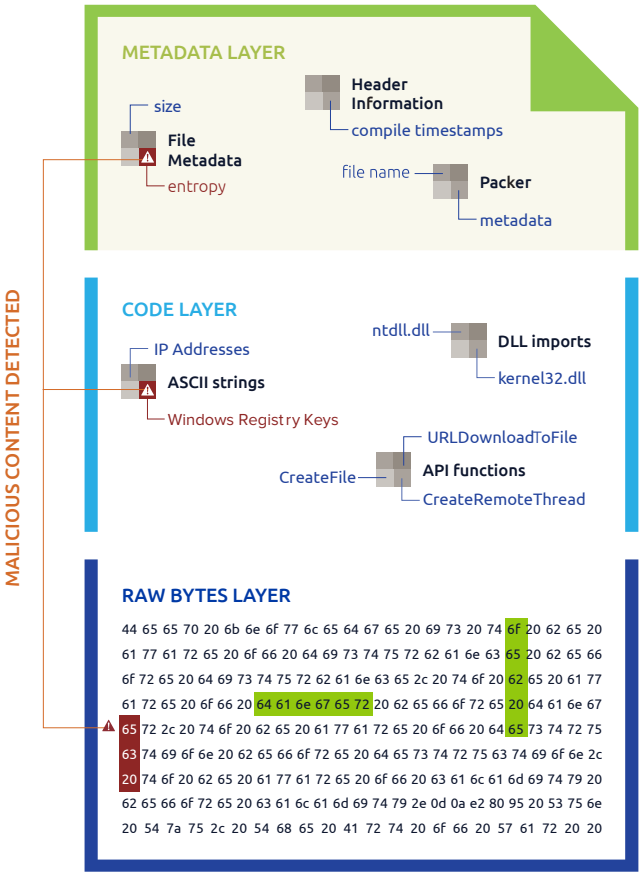


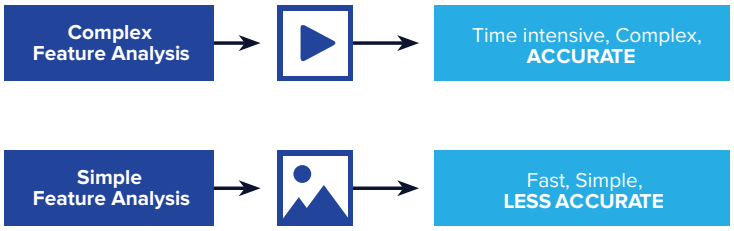
FIGURE 7. An abstract illustration of the type of features at each layer of a Windows executable file type. Supervised machine learning is able to consider the combination of features both present and absent across all layers. Other file types are different, but the concept remains the same.

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 across the "metadata," "code" and "raw bytes" layers as shown in Figure 7. Taking it a step further, unlike signatures, a machine learning model can consider the absence of features as part of its prediction making. Looking back at the "dog" vs. "cat" model, a Chihuahua (which may not have been among the training instances) may be nearly confused with a cat until we consider its tail because a cat moves its tail at the tip. The Chihuahua's absences of the ability to move its tail from the tip would result in it being classified as a dog. A Chihuahua only moves its tail from the base, something which a cat can do as well (See sidebar on Feature Selection vs. Speed).

Feature Selection vs. Speed

In the example of determining if a Chihuahua would be classified as a dog or cat, we look at the movement of its tail and observe due to its limited movement only at the base of the tail and lack of movement from the tip, it must be a dog.

This introduces a challenge when attempting to achieve analysis at speed. To include the movement of an animal in our feature set, we would need a way to interpret the movement in addition to having motion picture recordings among our dataset.



This may be more time intensive and complicated for the analysis vs. simply analyzing the pictures. The training instances required also become more complicated.

There is a similar impact when analyzing files for malicious content — the feature selection has an impact on both accuracy and speed of the analysis. This results in differences among one supervised machine learning model to the next and is best assessed with independent lab tests or onsite trials of models.

Conclusion

The application of machine learning in cybersecurity differs greatly both in application and objective. While unsupervised machine learning focuses on post-breach activity later in the cyber kill chain, supervised machine learning with the right balance of feature selection is capable of analyzing content with impressive accuracy at speeds suitable for analysis of all content at the network edge to provide the best network protection.

The advancement and proliferation of cyber threats illustrates the threats organizations now face and foreshadows the threats they will likely face in the near future. To combat destructive threats, it's imperative that organizations arm themselves with proactive information that gives their security teams an upper hand in the fight against cyberattacks — both now and well into the future.

About BluVector, A Comcast Company

As a leader in network security, 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 571.565.2100