WEBROOT
Smarter Cybersecurity™

Automating Threat Defense

Using Machine Learning to Prevent Modern Cyberattacks

Table of Contents

What does “Machine Learning” Mean? 3

How Does Machine Learning Work? 3

The Sorry State of Information Security 3

Machine Learning in Action 3

- Breadth and Depth: Better Visibility 3
- More Accurate Threat Detection 4
- Keeping Pace with the Volume of Threats 4
- Consistent and Dependable Results 4
- Self-Evolving Techniques 4
- Predicting Future Threat Vectors 5

The Benefit of Experience 5

Summary 5

What does "Machine Learning" Mean?

As technology continues to advance, managing and analyzing massive data sets just can't be accomplished by humans alone. It requires huge amounts of memory and storage, as well as the high-speed processing power of the cloud. Machine learning is a process that applies advanced mathematical algorithms and powerful computing capabilities to quickly and efficiently analyze those data sets and identify patterns. In the case of IT security, correctly determining patterns helps create accurate predictions and detect behaviors that may be associated with malware or other attacks. Doing so in real time, or as near to real time as possible, can help prevent breaches from occurring successfully in the first place.

How Does Machine Learning Work?

The algorithms determine how to interpret the data, and process it to produce predictable outputs. Machine learning then helps to decipher the data to identify the patterns, make sense of them, and enable security tools and personnel to take actions.

Tuning a machine learning model is complex and challenging work, and there's no substitute for time and experience when developing an accurate machine learning model. The key is continually refining the algorithms. Initially, data is fed into the model for training while human intelligence and analysis help to fine-tune it. This process, referred to as "supervised machine learning", takes highly skilled machine learning experts, data scientists, and statisticians, all working to train and test the algorithms over time. While the refinement process can be very involved, it increases both speed and accuracy, and enables machine learning systems to produce truly meaningful and actionable results.

The Sorry State of Information Security

At its core, information security is a numbers game. Organizations have to monitor tremendous volumes of incoming and outgoing information in real time, not to mention the packets that traverse their networks. Within that data, they then have to search for the minute combinations and patterns that may indicate malicious activity that could jeopardize their IT operations, steal sensitive customer and employee data, or critical internal intellectual property, result in significant financial loss, and put their organization's name in the headlines.

The threat landscape continues to grow and evolve in speed, sophistication, and persistence of attacks. Traditional security infrastructure is utterly overwhelmed by the flood of new attack vectors and the rapid expansion of new variants of polymorphic malware and crypto ransomware, while security researchers can't hope to keep up when armed only with manual identification, classification, and pattern matching techniques. InfoSec teams now find themselves inundated with alerts, and must research an ever increasing number of security incidents, without being able to prioritize their response. Unfortunately, this locks organizations into a reactive state as they attempt to catch up, rather than being able to proactively strategize and prepare to keep their data and customers secure.

Machine Learning in Action

It is more important than ever for organizations to be proactive in preventing security issues and attacks. Although human analysis and classification alone aren't feasible solutions anymore, companies can get ahead of the threats by augmenting existing defenses and security appliances with real-time dynamic intelligence that leverages advanced machine learning. Machines have a tremendous ability to crunch through massive amounts of data to find the worst and most current threats to businesses and individuals, and are incredibly accurate and consistent once the model is trained. Machine learning can automate security intelligence to enable accurate categorization and proactive threat detection in near real time. By taking advantage of the near infinite expandability of cloud computing platforms such as Amazon Web Services, machine learning can also scale with the increasing volume and complexity modern attacks.

Today, the most advanced machine learning platforms incorporate human feedback loops or active feedback, and active learning. Through active learning, they can become self-improving, and, essentially self-evolving. The active learning, massive scale, and accurate classifications can also be employed to drive predictive analytics, combining or contextualizing information on different threat types across disparate systems and domains to accurately predict where new threats will originate. Blocking or monitoring these potential threat sources enables security administrators to be more proactive than ever when anticipating attackers' next moves.

Breadth and Depth: Better Visibility

To create a system of learning, you have to start with a data set that's large enough for the algorithms to operate on to see patterns. This pattern recognition is what's used to make predictions and perform statistical analysis on the data set as a whole. Supplying a substantial data set for the model is crucial to enable the machine to learn, adapt, and produce the desired output.

A machine learning system also needs access to a wide variety of data for analysis. Broad and varied data sources are key to effective machine learning analysis. For example, this may start with standard sources, such as sophisticated internet crawlers that catalog all URLs, IP addresses, files, and mobile applications. Thanks to the seemingly infinite scalability of cloud infrastructures, crawlers are now able to catalog the entire IPv4 space in a matter of minutes. Large passive internet sensor networks, or honeypots, that attract malicious connections such as exploitable spam relays, are another effective way to gather data.

Active scanning, e.g., pinging on different levels within the IPv4 space to get IPs to respond, adds incremental data. It is also possible to subscribe to a variety of third party lists from financial networks, partner networks, the Defense Information Systems Agency (DISA), the FBI, and various malware lists. These lists are often riddled with false positives, but can still be quite valuable if the data is vetted and scrubbed prior to use. It is also necessary to analyze actual web traffic to determine the types of sites that may be targeted by malware based on usage.

All that said, when it comes to finding the richest and most highly differentiated source of input for machine learning-based security, nothing beats real-world endpoint and web sensor data. Organizations that incorporate real-world data from millions of endpoint sensors are better positioned to identify never-before-seen and zero-day threats the moment they emerge, anywhere in the world.

More Accurate Threat Detection

When applied to information security, sophisticated machine learning provides fast and accurate threat detection, including zero-day and previously unknown threats. Advanced heuristics and rules allow machine learning models to help determine in near real time if a file, URL, IP, or application is a threat, and then communicate that information broadly.

Although machine learning takes on the repetitive and tedious tasks InfoSec teams don't have the time or resources to process, it doesn't eliminate the need for people. Human threat researchers must be closely involved in the automated classification process in order to improve the model. It's a symbiotic relationship, as human efforts are highly leveraged when training the machine to be more accurate, while the machines improve upon human classification effectiveness by dramatically increasing speed and scalability.

The keys to speed and accuracy in a machine learning system are:

» **Vector dimensionality**, or maximizing the number of factors or characteristics used to classify each object

» **Processing speed**, i.e. boosting the number of classifications per second

» **Prioritizing** to enable the machine to choose the most important features for classifying objects.

Being able to rely on the machine is critical, but it requires data to be accurately classified before assigning different automated classifiers to different threat types. Classifier reputation and accuracy are measured and the results are integrated to strengthen the model.

Keeping Pace with the Volume of Threats

Cybercriminals are constantly developing new methods and approaches. They launch targeted attacks against organizations, specific data and information, as well as individuals, while representing organized crime, nation-state, and hacktivist groups. In order to successfully discover and block today's polymorphic malware, ransomware, phishing attacks, and other advanced and targeted threats, billions of events must be analyzed daily. Machine learning analytics need to be trained to look for modern attack methods, such as malware's polymorphic behaviors.

Leading threat intelligence platforms are cloud-hosted and can process tens of thousands of requests per second per CPU. They classify and re-classify billions of IP addresses across hundreds of millions of domains, and categorize billions of IP address and file behavior records, tens of millions of mobile apps, and have millions of connected sensors.

Continuous updates in a machine learning environment are possible through highly differentiated cloud-based architectures. These are similar to the big data architectures used by Amazon, eBay, Facebook, and Twitter. Real-time globally distributed Cassandra database clusters offer high performance at a massive scale with tremendous reliability, while scale for high volume transaction processing and throughput can be provided by Hadoop and other big data application frameworks. These technologies, combined with hundreds of classification models to cover different threat types and content languages, are the only way to automatically classify the millions of brand-new zero day file executables that can appear on a daily basis.

Consistent and Dependable Results

Despite an ever changing threat landscape, machine learning enables very high detection rates over time. To be successful, security intelligence must be continually published and updated; threat information on many types of threats—including files, applications, URLs, IP addresses, and phishing sites—is too dynamic to rely on email exchanges, black lists, or the like. There's just too much volume.

The goal of any machine learning system is to minimize the total amount of time from when the threat is detected until systems and users are protected. To provide timely, real-time protection for both systems and users, machine learning database architectures are designed for extremely fast updates. At the most, it only takes a few seconds from the time a change is detected until all systems and users are protected. As an example, with continuous updates, machine learning can identify a negative change to a web reputation score in fractions of a second and then alert on the potential new phishing source. Machine learning helps to make detection, communication, and protection a timely, consistent, and repeatable process.

Self-Evolving Techniques

» **First Generation: Bayesian algorithms**
Very simplistic analytics that were often inaccurate

» **Second Generation: Support Vector Machines (SVM)**
Basic supervised classification that used a given a set of training examples

» **Third Generation: Maximum Entropy Discrimination (MED)**
A major leap forward, MED began as an extension of SVM that was then fine-tuned to become quite effective for web analytics and detecting malware and phishing attacks

» **Fourth Generation: Active Learning (AL)**
An evolution of MED that added a human feedback loop from experienced threat researchers into the machine learning equation

» **Fifth Generation: Active Feedback (AF)**
Another evolution of MED that included managed human feedback loops, such as crowd sourcing techniques, lower signal/noise ratio for detection, and exceptionally high-performing algorithms

An active learning system uses incoming streams of data to select what are potentially the most informative data points. Using a machine learning model with active learning and retraining enables the system to reach model capacity much faster than if it relied solely on passive learning. Few cybersecurity firms use fifth generation machine learning with active feedback today because of the enormous technical challenges involved in implementing the technique successfully. This includes the need for both skilled personnel and the massive computing infrastructure to run it.

Predicting Future Threat Vectors

Using historical data, machine learning can actually predict a future outcome. In the case of cybersecurity, machine learning supports contextual analysis techniques that drive the development of actionable threat intelligence, which can predict the behaviors of unknown objects to find previously unidentified threats. Machine learning-based threat intelligence can assess every internet object based on its history and relationship to other known malicious objects to assign a risk score. That score indicates the likelihood that the object will launch an attack in the future.

Example 1: A specific URL may have had a reputation for serving malware for the last five years. If a new file that has never been seen before is associated with that URL, then it has a higher likelihood of being malicious than an unknown file associated only with benign sites. Using this guilt-by-association model, the file would be automatically flagged as a potential threat due to its relationship with the known malicious site. This model can be used to links objects, or nodes, such as files, URLs, mobile apps, and IP addresses.

Example 2: Let's say a user were to launch an app on a mobile device and it immediately began to access their contact list it. This already means it could be a bad app. Now let's say that the app contacted an IP address that was not previously classified as malicious, but now appears ready to accept the contact list. That IP address would be flagged immediately as being potentially malicious due to its relationship with the bad mobile app.

The benefit of contextualization provided by machine learning-based threat intelligence is understanding how known bad and known good objects communicate with other things online. It's not just a few connections here and there; it's the ability to analyze billions of associations across the same and different object types, combined with history on how millions of objects have behaved over time, which results in the predictive nature of threat intelligence driven by advanced machine learning.

The Benefit of Experience

There are significant differences in levels of experience with machine learning. Some firms have deep experience in machine learning and automated classification, and have actively tracked cyber threats for over a decade, while other firms are relative newcomers. Time in market allows for a stronger coverage model and minimizes the number of uncategorized objects. It allows accurate classification of websites and enables collection and analysis of historical records and the ability to track changes to objects such as URLs and IPs over time.

Commitment to machine learning can be measured not only by time in market, but through staffing dedicated teams of data scientists and threat researchers who specialize in security and who can access big data sources of security information from millions of endpoints, crawlers, active and passive sensors, and third-party sources. Machine learning analysis tends to be in the cloud, so security offerings and back-end services that were cloud-architected from inception have a distinct advantage. Machine learning systems should be live and in production, updated continuously, 24x7x365.

A security firm's level of experience with machine learning is critical, not only in terms of tracking cyber threats for a long period of time and having tremendous visibility over the threat landscape, but having the experience to build a robust threat database and then analyze that data. The associated learning and historical viewpoint of experienced machine learning-based security providers cannot be underestimated and is virtually impossible to replicate.

Summary

Security technologies enabled by machine learning have broad application. As the network perimeter expands to include employees working at home, on the road, on their company-owned devices, their own mobile devices, and their emerging IoT devices, this new network edge creates new points of entry for cybercriminals to breach systems and plant malware.

Machine learning technology is now emerging as a critical component across all of the various domains of cybersecurity. Traditional security appliances, such as next-generation firewalls, SIEMs, UTM, IDS, IPS, and more, can all be augmented by machine learning driven security intelligence. Machine learning can also power next-generation endpoint protection, mobile protection, threat intelligence, web security and network anomaly detection offerings. It enables threat activity across multiple security domains to be contextually associated in real time. Only systems using fifth generation machine learning algorithms can manage the numerous and disparate data flows from sources such as endpoint security, next-generation firewalls, network behavioral analytics, and external IP, URL, domain threat intelligence, and so forth.

The capability to understand and visualize threat vectors across multiple domains simultaneously using advanced machine learning is critical to maintaining an organization's security posture. When examining a breach, for example, organizations need to connect all cross-domain threat activity that occurred from the beginning of an attack through the kill chain to the actual crime itself, whether it involves data exfiltration or disabling critical infrastructure through a DDoS attack. By using advanced machine learning to automate contextual associations across all domains, organizations shift from a reactive mode to a predictive and preventive mode, and become proactive in their defense.

About Webroot

Webroot delivers next-generation network and endpoint security and threat intelligence services to protect businesses and individuals around the globe. Our smarter approach harnesses the power of cloud-based collective threat intelligence derived from millions of real-world devices to stop threats in real time and help secure the connected world. Our award-winning SecureAnywhere® endpoint solutions, BrightCloud® Threat Intelligence Services, and FlowScape® solution protect millions of devices across businesses, home users, and the Internet of Things. Trusted and integrated by market-leading companies, including Cisco, Citrix, F5 Networks, Aruba, Palo Alto Networks, A10 Networks, and more, Webroot is headquartered in Colorado and operates globally across North America, Europe, and Asia. Discover Smarter Cybersecurity™ solutions at webroot.com.

World Headquarters
385 Interlocken Crescent
Suite 800
Broomfield, Colorado 80021 USA
+1 800 772 9383

Webroot EMEA
6th floor, Block A
1 George's Quay Plaza
George's Quay, Dublin 2, Ireland
+44 (0) 870 1417 070

Webroot APAC
Suite 1402, Level 14, Tower A
821 Pacific Highway
Chatswood, NSW 2067, Australia
+61 (0) 2 8071 1900

WP _ 011817 _ US