MACHINE LEARNING AND ENDPOINT SECURITY – SEPARATING HYPE FROM VALUE

Learn how effective machine learning can help prevent malware and protect your endpoints

Antivirus software has long been the key defense in protecting the endpoint. Today, AV, designed to detect known threats based on signatures, is fighting a losing battle as new threats evolve and signatures become rapidly outdated. Even additional techniques, such as whitelisting, sandboxing and behavioral detection, are ineffective in the face of the tidal wave of new malicious files appearing daily. The most promising weapon in the endpoint security arsenal is machine learning, with its ability to quickly learn, make instant decisions and enable rapid response to prevent threats rather than dealing with them during execution or after the fact. One of the main advantages of machine learning is that it can capture minor deviations in an executable in a way that signature-based approaches cannot. As with any new approach, vendors have been quick to jump on the bandwagon and claim the benefits of machine learning for their products. With all the buzz in the market, it's important to understand the role machine learning can and should play, and how to separate hype from reality in effectively preventing malware.

The Vulnerable Endpoint

The endpoint remains the weakest link in enterprise security. Whether it's an unsuspecting employee clicking on a link in a phishing email, users relying on cloud-based collaboration tools, or simply leaving a system unattended, bad actors have endless opportunities to introduce malware into the corporate network. Today, it's easy and inexpensive to buy or rent malware and exploit kits that can evade traditional detection tools. The result is more than 1 million new malware files appearing each day.[1] Coupled with the fact that 80 percent of organizations receiving 500 or more severe/critical alerts per day report they currently investigate less than 1 percent of them,[2] it's easy to understand how security teams are hard-pressed to keep up.

High-profile breaches continue to be a headline staple in the press. Data from one million Verizon® customers was stolen in March of 2016; the credit card processing system used by Oracle®, MICROS was breached in August; 2.2 million patients' information from Century Oncology was compromised in October; and the Democratic National Committee was subject to a much-publicized breach and subsequent leak of information via WikiLeaks. These are only the most prominent examples.

The problem is compounded by the fact that malware comes in a variety of forms, including spyware, keyloggers, backdoors, command and control, exfiltration, and many others. Phishing attacks are a prevalent point of entry, adding to the complexity, alongside the fact that many known vulnerabilities – some several years old – continue to be exploited. Traditional tools have proven ineffective at facing this growing threat, and the consequences are potentially serious.

What We Need to Win the Battle Against Malware

Security teams need to shift their focus to preventing attacks, rather than detecting them during execution or after the damage has been done. With ruthless adversaries who frequently change their attack methods and approaches, we will continue to play catch-up unless we employ automation to identify threats. Automated approaches need to fulfill three basic requirements:

- **Accuracy** – security analysts can't waste time dealing with false positives, and the organization can't afford the exposure from false negatives.
- **Speed** –every minute counts when the network is under attack.
- **Efficiency –** quickly identifying threats without impeding end-user productivity.

Taken in the aggregate, these requirements point in the direction of a solution based on machine learning. With speed, accuracy and a light endpoint footprint, a solution that meets these requirements using machine learning would be able to detect threats and respond at speed.

What Is Machine Learning, and How Does It Work?

We've been hearing about artificial intelligence for years. More recently, the specific buzz has been around machine learning. Artificial intelligence is the science and engineering of making intelligent machines.[3] Machine learning is a type of AI that provides computers with the ability to learn without being explicitly programmed.[4] Machine learning is focused on sifting through data to look for patterns and then adjusting program actions accordingly.

We're familiar with many types of machine learning in use today, such as spam filtering programs and recommendation systems used by online retailers to encourage you to buy additional products based on your preferences. (For a detailed description of how machine learning works, see this blog post on Google® AlphaGo™.[5]) While machine learning systems can employ a number of different algorithmic approaches, they all have one thing in common: they need to be trained. Their predictions can only be made based on properties learned from earlier data. They need an abundance of high-quality data and intelligence.

To understand how machine learning works, we can use the example of spam filtering. Data scientists built a database that encapsulated their understanding of typical spam characteristics and trained the machine to parse data and make determinations based on that initial knowledge. Over time, all of us add to the body of knowledge each time we mark an email message as spam. As a result, the system gets better and better at identifying what is really spam and what is legitimate email. The more extensive the base information, and the more often it is updated and enhanced with rich data, the better the machine learning results.

How Does Machine Learning Improve the Effectiveness of Endpoint Protection?

Incorporating machine learning into endpoint security efforts would allow decisions to be made quickly based on what has been learned in the past, while continuing to learn and improve as new malware variants are seen. Machine learning overcomes the weakness of signature-based approaches that can't be updated quickly enough to deal with evolving threats. Signatures are easy to avoid, as attackers can write software that takes malware and makes multiple small obfuscations, each of which requires its own signature. This leads to an explosion of the number of signatures, as well as the fact that the signatures often don't match new malware.

A machine learning solution on the endpoint would be able to identify real threats more quickly and accurately than other approaches. It could rapidly determine whether a file is benign or malware and perform immediate triage to stop suspected malware at the first point of entry. The best solution would not stop there: Once it blocked a threat, it would continue to learn all it could about the threat, and, if the suspect file were not really malware (a false positive), it could act quickly to maintain end-user productivity. If it were a real threat, the machine learning model could then be enhanced with more details to help it make better, faster determinations in the future.

How to Evaluate a Machine Learning-Based Solution for Endpoint Protection

Many vendors, seeing the growing interest in machine learning, have rushed to adopt the term in their marketing literature. As with most promising technologies, we see overstated capabilities and hollow promises. We need to look carefully at the three principal characteristics – accuracy, speed and efficiency – to determine if the solution truly incorporates machine learning.

Accuracy: Since the solution will both intercept potential malware and provide information for future decision-making, it must have a low rate of false positives. This is a result of analyzing a tremendous amount of raw data – security analysts who establish baseline rules classify robust data sets. Ongoing training of the model similarly relies on continuous access to large amounts of new data. But the rate of false positives can be extremely high if the data set is not robust.

It is important to note that the machine must have access to both benign and malicious data in order to accurately distinguish between the two. Training a model based solely on bad data increases the chance of high false-positive rates.

Ask the vendor:

- Upon how much data does the machine learning solution base its decisions? Is it enough?
- From where does the data come? Is there a wide variety of sources, or are they dependent on third-party threat aggregator sites?
- How often is the data collected?
- How often are new models trained and propagated to the customer?
- How is the system trained? Is it trained through a constant supply of rich data sets, so properties discovered can be used in future machine learning decisions?
- How does the vendor handle false positives?
- How does the vendor handle false negatives that the vendor later discovers (after the customer has run the malware)?

Speed: The solution must quickly determine, with a high degree of certainty, that a threat has been detected.

Ask the vendor:

- How quickly can the solution make a determination that leads to action?
- How quickly can it obtain enough relevant new data to influence the decisions it makes?

Efficiency: The solution must be accurate and fast, and it must support a productive end-user environment. Many solutions on the market today are slow, and they have a big impact on end-user productivity.

Ask the vendor:

- Where and how quickly does the analysis take place?
- What is the impact on the end-user system?
- What type of analysis is done on incoming files? On endpoints only, on cloud only, or a combination?
- Does it rely on post-event analysis (detecting rather than preventing)?

How Palo Alto Networks Traps Provides Advanced Endpoint Protection

Using these three characteristics as a benchmark, it is easy to see that Palo Alto Networks® Traps™ advanced endpoint protection employs advanced machine learning.

Accuracy

Accuracy depends on data: abundant, rich and timely. Traps is uniquely capable of building and enhancing an extremely rich source of files that are used to train its predictive machine learning classifier. The company has access to a vast number of files, thanks to its distributed sensor system that pulls data from more than 15,500 enterprise, government and service provider customers. Its large firewall deployment and the massive adoption of its threat intelligence cloud, WildFire™ threat analysis service (with almost three times as many customers as the nearest competitor), as well as the many files uploaded from Traps endpoints, more than satisfy the data quantity requirement.

In addition, because WildFire actually runs every file, Palo Alto Networks obtains fantastic ground truth on a large scale, while other organizations without comparable capabilities are dependent on third-party threat aggregator sites. And because accuracy also depends on the quality of the data scientists who determine the labeling and classification of data, Palo Alto Networks again has a big advantage with some of the world's experts. We understand how to extract thousands of unique features from each file and how best to classify them, producing results unattainable with static or dynamic analysis alone.

Speed

Speed is the only way to defeat new threats and protect the endpoint. Traps is extremely fast while operating on the endpoint. This first line of defense provides a high-speed determination, but it doesn't stop there. With Traps, any unknown, suspicious file is uploaded to the cloud for additional, more thorough dynamic and static analysis. This multi-method prevention approach helps improve accuracy and end-user safety.

Efficiency

Traps is designed specifically to identify evolving threats before they can do damage, without slowing down end-user productivity. The Traps multi-method prevention approach does away with the need for constant signature updates. It avoids the bandwidth issues of traditional antivirus, and does not slow down the endpoint with manual or scheduled scans or memory issues. In fact, the Traps agent represents less than 1 percent CPU usage. This means that, while it is effectively preventing threats from turning into breaches, users remain fully productive.

Endpoint Security Solutions That Deliver

The probability of falling victim to an attack is growing far faster than traditional tools or human analysis can handle, and automation is clearly the answer. Machine learning is proving to be a key component in an effective endpoint security solution. But not all machine learning solutions are equal. When evaluating an endpoint security solution based on machine learning, get answers to key questions to ensure the solution you choose will be truly effective. Once you fully understand the three most important characteristics – accuracy, speed and efficiency – it will be clear that Palo Alto Networks Traps can deliver advanced endpoint protection based on superior machine learning technology.

1. Virginia Harrison and Jose Pagliery. Cyber Attack Hacks Security. CNN Tech, April 2016. Retrieved from http://money.cnn.com/2015/04/14/technology/security/cyber-attack-hacks-security/
2. David Monahan. Achieving High-Fidelity Security. An Enterprise Management Associates End-User Research Report. April 2016. Retrieved from https://www.savvius.com/files/marketing/white_papers/EMA_Savvius_High_Fidelity_Security_2016.pdf
3. John McCarthy. What is Artificial Intelligence? Computer Science Department, Stanford University. 2007. Retrieved from http://www-formal.stanford.edu/jmc/whatisai/node1.html
4. AWS analytics tools help make sense of big data. TechTarget. Retrieved from http://whatis.techtarget.com/definition/machine-learning
5. Demis Hassabic. AlphaGo using machine learning to master the ancient game of Go. January 2015. Retrieved from https://blog.google/topics/machine-learning/alphago-machine-learning-game-go/

4401 Great America Parkway
Santa Clara, CA 95054

Main: +1.408.753.4000
Sales: +1.866.320.4788
Support: +1.866.898.9087

www.paloaltonetworks.com