Machine Learning and Next-Generation Intrusion Prevention System (NGIPS)

Building a smarter NGIPS

» How Trend Micro is using machine learning to tackle today's complex security threats

EXECUTIVE SUMMARY

Modern security threats are short-lived and constantly evolving, at times limiting the effectiveness of traditional signature- and hash-based detection mechanisms. And with regular expressions unable to detect and block certain highly evolved modern threats, newer and smarter security techniques are required.

This white paper explains how Trend Micro has incorporated machine learning into its next-generation intrusion prevention system (NGIPS), taking an in-depth look at the three main steps in its approach — and how machine learning has significantly improved the effectiveness of its TippingPoint NGIPS in detecting and blocking certain classes of complex and unknown threats.

CONTENTS

WHEN REGEXES ARE NO LONGER ENOUGH

Among the many challenges faced by IT security managers, one of the most difficult is the short-lived, ever-changing nature of today's security threats. The only constant is that they are constantly evolving, using a mix of encryption, obfuscation and zero-day exploits to stay one step ahead of traditional signature-based detection mechanisms. By the time signatures are developed and deployed for a specific threat instance, they're already outdated or can detect only a subset of the malicious content.

While powerful string- and pattern-matching tools like regular expressions (regexes) continue to be critical for blocking vulnerabilities, they aren't flexible or dynamic enough on their own to continuously detect and block a wide range of advanced and highly volatile threats.

Exploits kits are a perfect example of a malware-delivery system capable of evading regex detection, especially when coupled with obfuscation techniques. First, exploit kit operators avoid reputation checks by compromising what are normally considered to be safe websites. They inject encoded objects into the HTML that are then decoded by JavaScript or VBScript evaluation to redirect users to a separate landing page. Once on this page, the user's vulnerabilities (related to Flash and other plug-ins, for instance) are fingerprinted and the exploit kit's control server delivers a specific shellcode that then downloads, decrypts and deploys malware to the user's system.

The security challenge here lies in the fact that the attack code in the exploit kit's landing page changes dynamically every time a user clicks through to it, making it extremely difficult to write a regex to handle it.

While it's true that the top exploit kits such as Angler, Nuclear and Neutrino have all gone underground following the arrests of 50 people in Russia in 2016 — with Trend Micro data showing that exploit kit attacks that year were only a third of what they were in 2015 — they are a threat that can easily return. Other exploit kits still remain active, such as RIG and Sundown. And the developers of the top exploit kits have proven adaptable in the past and are expected to evolve their arsenal to adopt more effective obfuscation techniques.

Still, exploit kits were just the leading edge of attacks that use obfuscated HTML and script. That trend will continue whether exploit kits resurface or not, as already seen though obfuscated JavaScript that conceals Trojan droppers, malicious Flash and PDF files, polymorphic malware and the latest fileless malware, which downloads nothing on the victim's system and instead runs entirely through a web browser's script interpreter.

The bottom line? When used in isolation, traditional signature- and hash-based detection won't work against these modern threats — meaning newer, smarter security techniques will be needed to address them.

Using machine learning to address modern security threats

Statistical data modeling powered by machine learning will be essential to closing the security gaps exploited by obfuscation and encryption — and in doing so, improve the security effectiveness of technologies like next-generation intrusion prevention systems (NGIPS).

By analyzing the attributes and characteristics of a dataset that contains both 'known good' and 'known bad' examples, machine learning algorithms can compute a mathematical model that an NGIPS can then use to make a decision — in-line and in real-time — about whether the network traffic passing through it is benign or malicious. By constantly evaluating the model against newly collected samples, the algorithms can 'train' detection engines to adapt to the latest threats and be more efficient in blocking new and unknown malicious items.

At the highest level, machine learning is focused on identifying and blocking generalized techniques or patterns of obfuscation rather than one particular threat instance or vulnerability. So rather than writing a regex to block a specific string of code, the highly refined and trained machine learning algorithms can recognize unusual patterns that are typically associated with malicious code — the volume and frequency of use of certain types of characters, for example, or how often the code transitions between letters and digits — and then, based on how many times those identifying patterns are seen, separate the harmful traffic from the benign. This means filters can be developed for an NGIPS that can cast a wide net to detect and block thousands of variations of the same threat, even those that have yet to be created.

Trend Micro is the first security vendor to incorporate machine learning into its NGIPS to detect and eliminate some of today's more complex and sophisticated threats in-line, at wire speed. Machine learning is applied through an iterative, three-staged approach: gathering the data, selecting the appropriate features from that data, and then using those features to build and validate statistical models for identifying threats

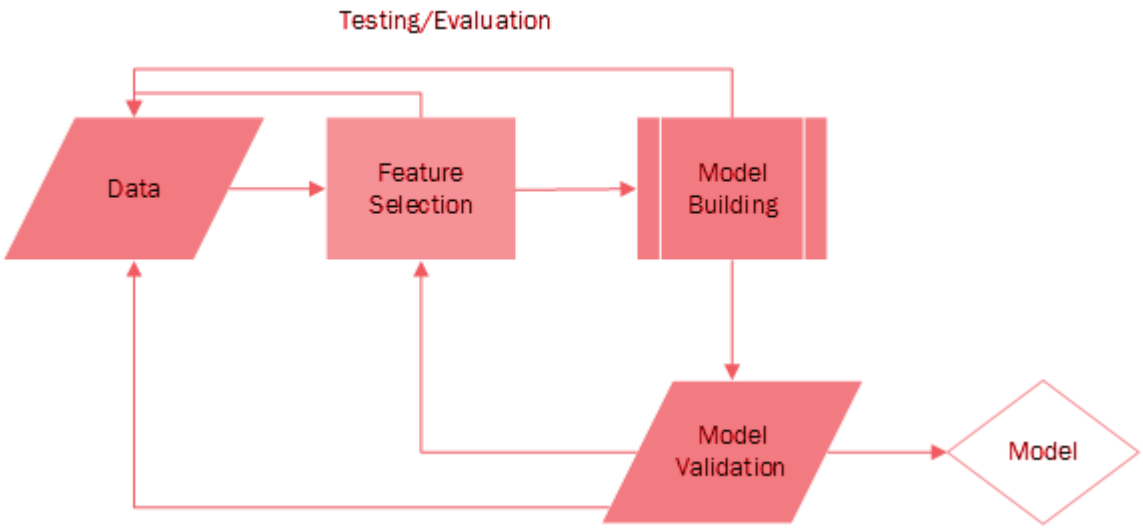

Figure 1. How Trend Micro applies machine learning to its NGIPS

STEP ONE: DATA GATHERING

At its core, machine learning is simply data science — and you can't do data science without, well, data.

When applied to the context of an NGIPS, machine learning is about performing classification tasks on data: describing and separating it into different classes based on certain characteristics (in this case, whether something is or is not malicious). To make a precise and accurate classification, you first need to gather and analyze as much data as possible about the items you're trying to classify.

For example, say you wanted to make a program to identify whether something is or isn't a tree. If the only data you've collected says a tree has blue-green needles, your program will incorrectly assume that only conifers can be considered trees. In other words, if your data is limited, the conclusions that can be reached through that data will be limited as well.

It's also important that the collected data be as "clean" as possible. If information on things other than trees find its way into your positive classification of a tree, your determinations will not be 100 percent accurate. After the data has been gathered, techniques such as anomaly detection and clustering can be used to detect outliers and better understand the latent groups within the data, ensuring only the most suitable data is used for classification.

STEP TWO: FEATURE SELECTION

Once the data has been collected and cleaned, the next step is to isolate and quantify the features that will form the basis of the statistical model used to distinguish between benign and malicious data.

Going back to the example of the tree-identifying program, some of the specific features built into that model would include the height, color, shape and area occupied by the object to be classified. By comparing and contrasting the right parameters, it should be easy for the model to tell the difference between a tree and any another object.

When measuring and categorizing a webpage as either malicious or benign, Trend Micro's machine learning algorithms count the words (i.e., alphabetic strings), non-linguistic bigrams (i.e., character pairs that don't appear in Indo-European languages, such as ZX or QF) and linguistic bigrams (i.e., common character pairs such as BU or DE) appearing on the page. A person (and by extension, a real landing page) will use a greater number of words and linguistic bigrams to express meaning; the randomly generated obfuscated code of an exploit kit landing page will not.

Trend Micro's algorithms also measure the frequency at which different character classes appear on the webpage: digits, hex characters, punctuation, uppercase and lowercase characters, whitespace and non-printable characters. Also assessed are the transitions between these various character classes: how often an uppercase letter is followed by a lowercase letter, for example, or when and where two whitespaces in a row are used.

A closer look at exploit kit obfuscation

To illustrate the features assessed by Trend Micro, consider the following example from a landing page generated by the Neutrino exploit kit, which contains several different forms of obfuscation. The top "div" element uses word-like obfuscation, featuring a lot of punctuation as well as a high frequency of lowercase-to-lowercase and lowercase-to-whitespace transitions. On their own, those traits aren't very indicative of malicious obfuscation; however, the human brain can quickly tell this content is obfuscated because it simply doesn't resemble anything we encounter in any spoken language.

This sample also features numeric obfuscation, but as many webpage elements present numbers in this way, there's nothing particular about that code that flags it as being malicious.

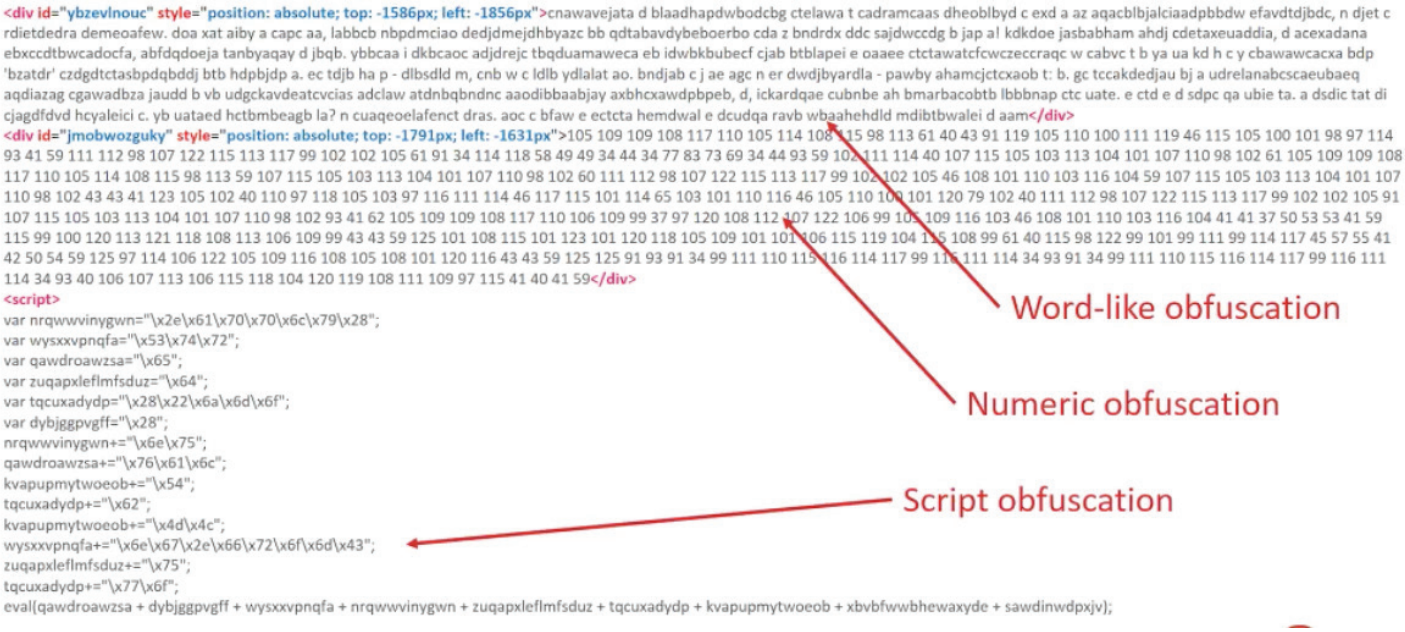

Figure 1. How Trend Micro applies machine learning to its NGIPS

Similar types of obfuscation can be seen in the Angler exploit kit. For example, one variation of Angler starts with a set of seemingly random words and phrases taken from classic novels (to make it look like the page was written by humans and bypass detection). Further down the page, however, is a large block of hex-like characters (i.e., numbers in combination with only the letters "A" through "F") that are clearly not meant for human interpretation. These hex-like characters also feature more uppercase characters and whitespace than is typically seen in real webpage copy, along with a higher proportion of unusual uppercase-to-digit and digit-to-uppercase transitions — more flags of potentially malicious code.

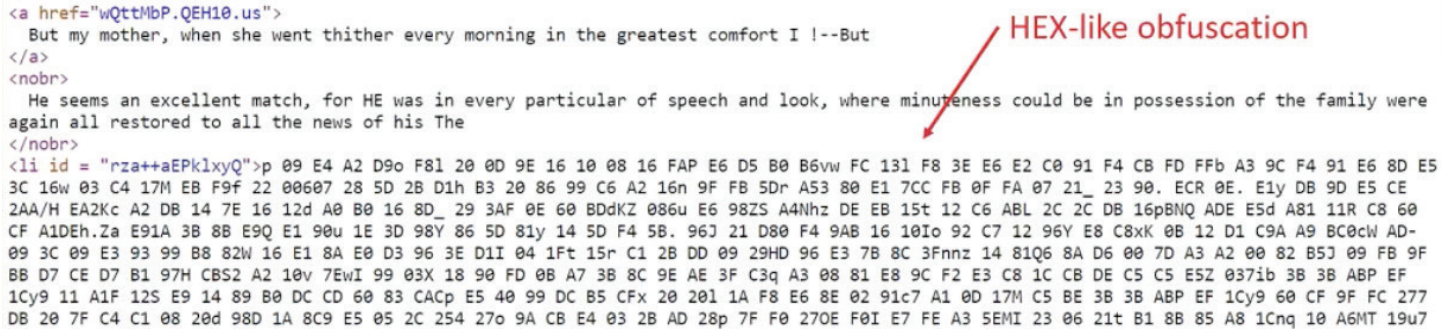

Figure 3. Hex-like obfuscation used by the Angler exploit kit

Again, it must be emphasized that there's nothing inherently good or bad about an uppercase letter or digit. When they appear in certain ratios, however, that can indicate a potential exploit kit landing page or other obfuscated content. As shown in figure 4, which compares character and transition counts from Neutrino and two variations of the Angler exploit kit with those of Google and Tumblr, it's easy to see that these 'top sites' have much higher counts of lowercase characters, punctuation characters and linguistic bigrams.

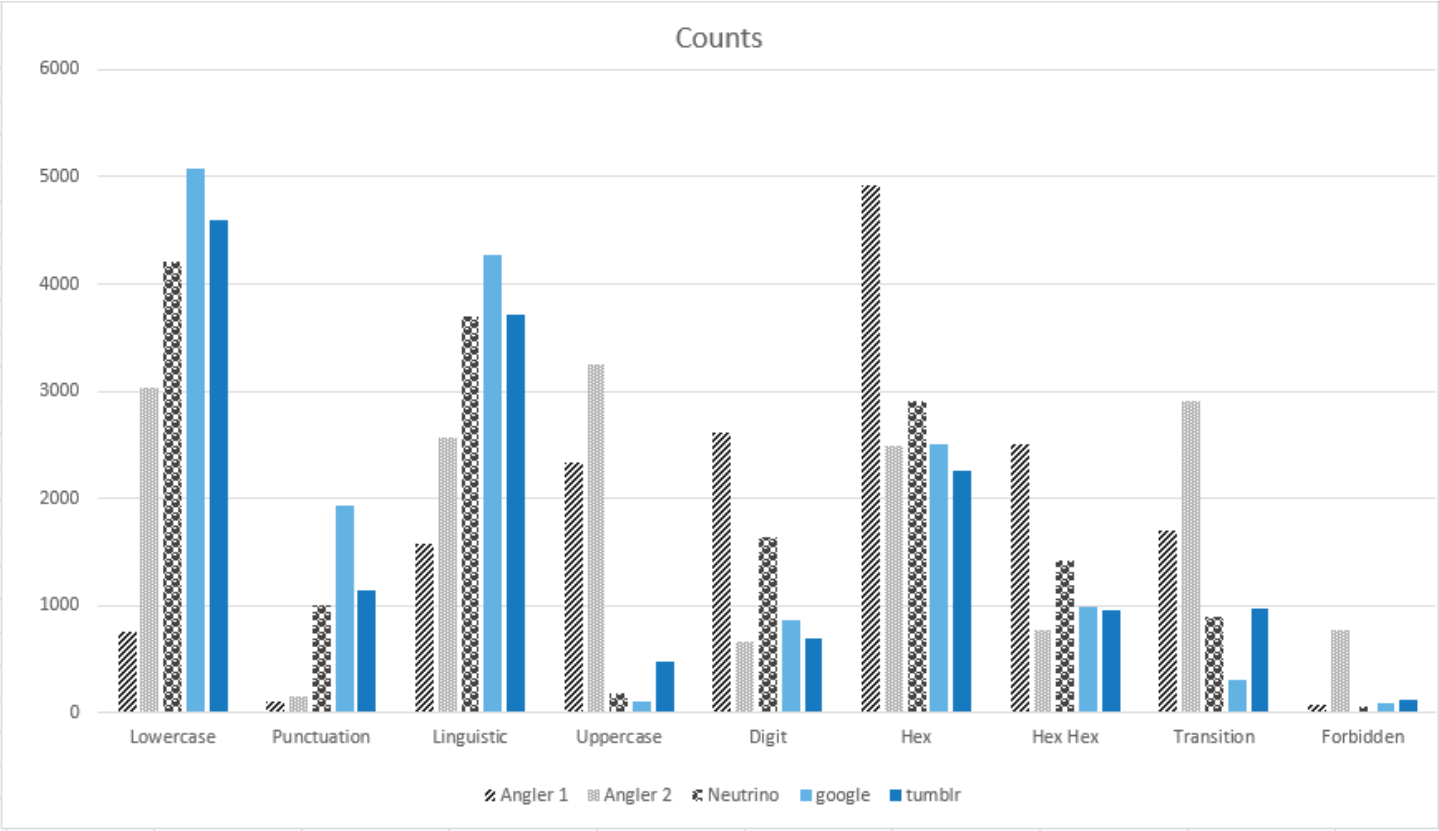

Figure 4. Character counts from Angler and Neutrino samples versus top sites

Surprisingly, these are the only features required to build a statistical model that can accurately classify whether a webpage is benign or a malicious exploit kit landing page. And as a general rule, it is better to incorporate as few features as possible into the statistical model.

For example, while Trend Micro measures counts of lowercase characters, it does not count instances of every letter in the alphabet. Doing so would lead to a model that 'overfits' the collected data, meaning the model would be very good at predicting things that fit within its parameters but very fragile should it encounter new pieces of data. As an NGIPS has to deal with new and unknown threats all the time, it is better to build a model that relies on a smaller set of features so that it can generalize to a greater degree, both within and outside its dataset.

STEP THREE: MODEL BUILDING

With the features selected, the next step in the process involves building a generalizable model for classifying benign and malicious data.

First, the selected features are assigned either a positive or negative weighting. Features such as lowercase-to-uppercase transitions, uppercase-to-whitespace transitions, and the use of digits and non-linguistic bigrams are weighted positively, meaning they're more likely to indicate malicious content. Features common to top sites (such as linguistic bigrams and the use of punctuation) are assigned a negative weight. The counts for each feature are multiplied by their respective weighting coefficient and then added together, producing a final score. If the score is positive, the content will be classified as malicious.

That sounds simple enough — until you consider that the statistical models used by Trend Micro incorporate thousands of different data points. Attempting to visualize that data for further analysis would result in a thousand-dimensional graph, which is very difficult for the human brain to comprehend. That's why Trend Micro uses a process known as principal component analysis, which removes redundant information in the features to take the collected data from n-dimensional to three-dimensional space, making it possible to generate an easy-to-understand 3D visualization.

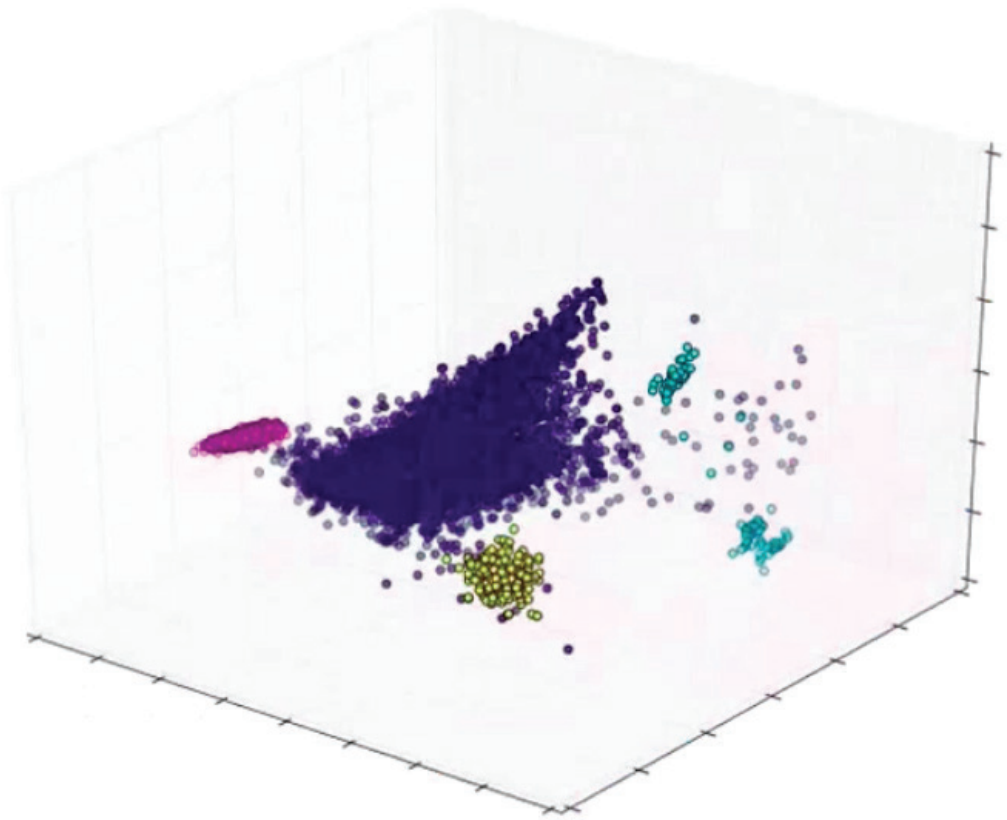

Figure 5. 3D visualization showing clear separation between malicious and top site data

In figure 5 above, the large dark-blue cluster represents data from top sites. The yellow and light blue represent the two variations of the Angler exploit kit, while the magenta represents Neutrino. With just one glance, it's easy to see that, for the most part, the exploit kit data is set off by itself, away from the top-site data. As a result, it is possible to draw lines and planes separating the various classes of benign and malicious content.

When clear separation of the data is evident, the 3D visualization can then be converted into a simple linear model to be used by the NGIPS with little to no impact on benign traffic.

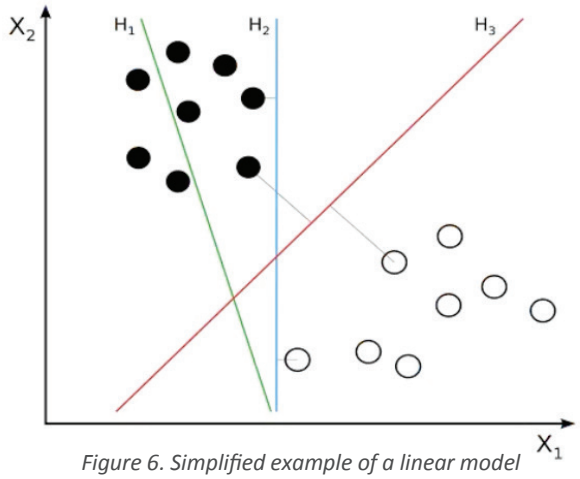

Figure 6. Simplified example of a linear model

In the example shown in figure 6, everything on the right side of the red line are the features that likely indicate benign content (i.e., those with a negative weight, such as linguistic bigrams). On the left of the red line are the features common to malicious files.

As the linear model reflects essentially a weighted sum of character counts and their coefficients, classifications can be performed quickly by the NGIPS with just a single pass over the incoming traffic — limiting induced latency to lessen the overall impact on network performance.

A more flexible approach to detection

An additional benefit of the linear model is the flexibility it offers, as different models can be created easily for general or specific threats.

As shown in figure 7, creating a model to separate all benign data from all malicious data is relatively straightforward: you just draw a line between the two. That said, it's not possible to get 100 percent separation, as there is some overlap between the blue and green dots in places. No matter where you place the line, there will likely be a green dot on the left side of it, which would result in a false positive from the NGIPS. However, if we're trying to develop a more general model to catch as much malicious content as possible, a small number of false positives might be considered acceptable — and you might even shift the green line further to the right to be absolutely sure that no malicious traffic can slip through the NGIPS.

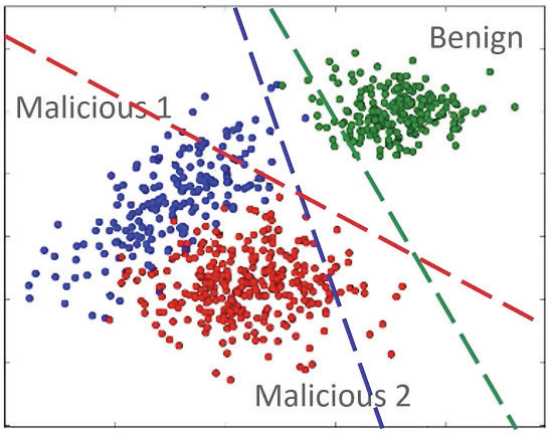

Figure 7. Creating different models for different purposes

The linear model can then be divided even further to create specific models that separate the blue and red obfuscation techniques, for example, or to isolate the red malicious data from the green benign data.

This is exactly what Trend Micro has done with the Digital Vaccine® (DV) filters deployed in its TippingPoint NGIPS, creating a general model to hunt for unknown and undiscovered threats (while allowing for a higher number of false positives) as well as models designed for specific exploit kits (with much lower false-positive rates). For example, DV filter 23799 addresses almost all variants of Angler as well as several other exploit kits that leverage obfuscated HTML and JavaScript, detecting and blocking this malicious content with a false-positive rate of less than one percent on top 100,000 websites.

DV filters have also been created to detect domain generation algorithms (DGAs), which are used in many malware families to randomly generate domain names from which to contact their command-and-control servers. These filters include classifiers developed through machine learning techniques, using a combination of syntactical rules and logistic regression to detect families of DGAs with over 95 percent accuracy.

Making the model better

The "learning" aspect of "machine learning" comes into play when the model is exposed to massive amounts of real-world data — and then adjusted as necessary to improve the precision of its classifiers. Trend Micro is constantly improving its models through test deployments with real enterprise traffic, checking to see if the model is detecting what it is supposed to or if it's leading to too many false positives. If so, it might be necessary to go back the first two steps in the process to optimize and improve things further — for example, by gathering more data, choosing different features, or finding new ways to group the data via additional rounds of clustering and anomaly detection.

TREND MICRO: AN INDUSTRY-LEADING INNOVATOR

As security threats become increasingly sophisticated and complex — and too dynamic for traditional signature-based detection to handle — statistical models powered by machine learning will become increasingly important to keeping enterprises protected.

Trend Micro's NGIPS security filters are frequently tested by the NSS Labs Cyber Advanced Warning System. Since the first machine learning filters were deployed in April 2016, Trend Micro's block rate for live exploits has jumped from 90 percent to 99.5 percent — proof of the effectiveness and detection power offered by the machine learning approach.

Just as important, however, will be enterprises' adoption of a smart, optimized and connected approach to network security: one that utilizes a variety of cross-generational threat protection techniques, including the TippingPoint NGIPS, with seamless integration and sharing of real-time threat intelligence between security products for improved detection accuracy. This is the core philosophy behind Trend Micro XGen security, which protects against the full range of known, unknown and undisclosed threats by leveraging multiple detection techniques, all working together and building on each other's strengths to catch as many malicious elements as possible.

By continuously adapting to the evolving threat and IT landscape, Trend Micro is the first major security vendor to:

- Detect and block attacks in-line and in real-time by embedding machine learning techniques in its NGIPS
- Offer an NGIPS that delivers up to 100 Gbps inspection throughput with low latency for data centers and high-performance enterprise networks

A recognized leader in security
Trend Micro is a:

- Recommended vendor for next-generation intrusion prevention systems (NSS Labs, 2016)
- Recommended vendor for breach detection systems for three consecutive years (NSS Labs, 2016)
- Leader in the 2017 Gartner Magic Quadrant for Intrusion Detection and Prevention Systems
- Leader in 2016 Forrester Wave for Advanced Malware Analysis

Drawing on three decades of experience and extensive history of innovation, Trend Micro's TippingPoint NGIPS solutions, powered by XGen security, are helping enterprises efficiently and effectively protect themselves against a wide range of complex, unknown and undiscovered threats such as exploit kits and obfuscated HTML and JavaScript — including those that can now slip by the formerly powerful regexes they used to depend on.

TREND MICRO™
Securing Your Journey to the Cloud

Trend Micro Incorporated is a pioneer in secure content and threat management. Founded in 1988, Trend Micro provides individuals and organizations of all sizes with award-winning security software, hardware and services. With headquarters in Tokyo and operations in more than 30 countries, Trend Micro solutions are sold through corporate and value-added resellers and service providers worldwide. For additional information and evaluation copies of Trend Micro products and services, visit our Web site at www.trendmicro.com.

TREND MICRO INC.
U.S. toll free: +1 800.228.5651
phone: +1 408.257.1500
fax: +1 408.257.2003