WHITE PAPER

McAfee™
Together is power.

Advanced Analytics and Machine Learning: A Prescriptive and Proactive Approach to Security

Table of Contents

Advanced Analytics and Machine Learning: A Prescriptive and Proactive Approach to Security

Everyone involved in information security—from solutions vendors to analysts to CISOs—can agree that analytics is an essential tool in the continuing effort to disrupt adversaries, who are using increasingly advanced and complex attack methods. With the advent of Big Data, the Internet of Things (IoT), and 24/7 connectivity, data scientists are now at the forefront of security product development, as vendors like McAfee expand advanced analytics and machine learning capabilities from its research lab into its products. Overwhelmed by the immense volume of security intelligence and alerts, human analysts need machine learning to augment and accelerate their efforts. Machine learning expands the scope of security analytics from diagnostic and descriptive to prescriptive and proactive, which leads to faster and more accurate detection and improved capabilities to act on threat intelligence today and in the future.

> "At McAfee Labs nine years ago, we saw less than 200 new threats per day. Today, we see almost 400,000. Our response has to be of equal scale if we are to beat our enemy."
>
> —McAfee® Labs

Evolution of Analytics

Let's begin by taking a look at the evolution of analytics, which spans numerous areas—from data mining and data monitoring to forecasting and machine learning. As defined by McAfee Chief Data Scientist Celeste Fralick, analytics is "the scientific process of transforming data into insight for making better decisions."[1] In the security world, this definition can be expanded to mean the collection and interpretation of security event data from multiple sources and in different formats for the purpose of identifying threat characteristics and improving protection, detection, and correction.

The science of analytics has undergone a transformation in a relatively short period of time:

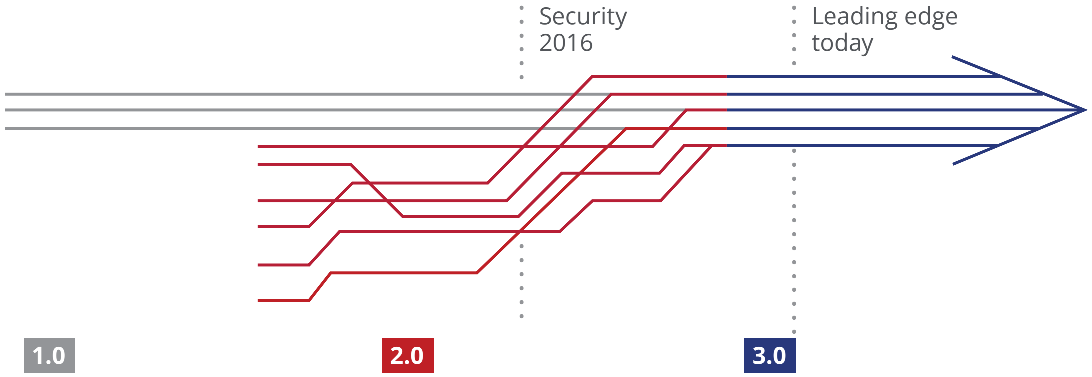
- **Analytics 1.0:** In the early stages, data statisticians spent their time dissecting internally sourced structured data sets, most often in reaction to a specific problem. This type of analytics was descriptive and diagnostic, answering the questions "What happened?" and "Why did it happen?" Most vendors

Connect With Us

are extremely competent in this area, applying the knowledge they gather to rule sets and decision trees. In fact, it's imperative for vendors to continually react, respond, and learn through this type of analysis, which, along with a layered approach, is vital for truly effective security coverage.

- **Analytics 2.0:** In the era of Big Data, connectivity, and microprocessors, the quantity of security data, which is being culled from both internal and external sources, has been growing in volume and complexity. Vendors have become adept at the very essential task of churning through mountains of information and making sense of it, though the emphasis has been on descriptive and diagnostic analytics.
- **Analytics 3.0:** Security vendors are now beginning to move in the direction of predictive and prescriptive analytics, which enables accelerated and proactive discovery and insights. Machine learning applied to Big Data utilizing deep learning methodologies (which may include cognitive computing) is the foundational technology. Predictive solutions for current threats—such as ransomware, advanced malware, and botnets—are already being rolled out.

The evolution of analytics



Analytics 1.0

- Internally sourced, structured data sets
- Descriptive and diagnostic analytics
- Reactive, but useful

Analytics 2.0

- Big data: Large, complex, unstructured
- Data from internal and external sources

Analytics 3.0

- Uses machine learning with big data, deep learning, and cognitive computing
- Fast, proactive discovery and insight

Figure 1. The evolution of analytics. (Used with the permission of Dr. Tom Davenport.)

Algorithms: At the Foundation of Machine Learning

Machine learning is one of the driving technologies differentiating Analytics 3.0. Machine learning leverages automation to learn and adapt over time as new data comes in and utilizes predictive and prescriptive algorithms. These types of algorithms need to be trained how to learn, and this process involves several steps:

- **Framing the problem:** Developing a useful predictive and prescriptive algorithm requires an understanding of how the analytics will solve a given problem; knowing the variables, inputs, and outputs; and grasping how the solution will make your business healthy and more secure. At this stage, proper scoping is critical in order to ensure that the right problem is being addressed and that the analytics yield conclusive and accurate results that can help drive a viable solution. Key questions to ask before you begin the training associated with machine learning include the following:
  - Are analytics being developed as a way to detect suspicious activity?
  - How is suspicious activity defined when each environment is different?
  - What does it mean for activity to be suspicious?
  - What are the acceptable false positive and false negative rates?
  - What is the risk of not alerting?
  - What level of explanation and information needs to be included in the output, along with the answer? (In other words, how proactive does the solution need to be?)
- **Cleansing and processing data:** This is a time-consuming process, but an important one. It entails removing corrupt, incomplete, irrelevant, and inaccurate data. By getting this step right, organizations will be better equipped to identify spurious data, improper readings, and data trending.
- **Statistical analysis of the data:** For this step, data scientists perform a variety of often complicated mathematical operations that will help them determine whether they need to apply data normalization to reduce redundancy and improve data integrity and/or data transformation to translate the format to a destination system.

  Once this is accomplished, the data scientist will manipulate the data to fit a model. For example, let's assume that, in the wild, 99% of the data is non-malicious, but the focus on collection has been on finding the malicious samples. In this scenario, the training needs to be adjusted to compensate so that it doesn't statistically align with real-world traffic in a given sample ration but rather specifically emphasizes identification of malware.

The Terminology of Analytics

- **Descriptive:** The first stage of analytics reviews and summarizes historical data, attempts to determine causes of events and behaviors, and prepares data for further analysis.
- **Diagnostic:** This stage asks the question: "Why did it happen?" Techniques used include data discovery, data mining, and correlations.
- **Predictive:** This mode of analytics uses descriptive and diagnostic analytics as a basis for answering the question "What will happen?"
- **Prescriptive:** Building on predictive analytics, this area of analytics states that "This is what is recommended because certain events will likely happen."

- **Choosing a model:** Choosing a model is the biggest challenge faced by data scientists. The data will direct the choice of a model. The data scientist needs to ensure that the model aligns with customer needs and that the data fits the model precisely and accurately and in a way that can be repeated. Above all, the data scientist needs to verify that the model answers the question that is being asked. In addition, the data scientist needs to able to show that the model will adapt as changes to the data stream occur and to verify that the model can actually learn rather than just memorize. Often, the data scientist decides to test three to five models to find the best match.
- **Verifying the model:** In this step, the key question is whether the model performs as expected. There are many ways to determine this. One of the more common methods is call "cross-validation," where a specified percentage of the data is channeled into a training set and the rest is held back for testing. For example, 80% of data could be channeled into a training set, which helps discover predictive relationships or correlations, and 20% of the data could be channeled into a validation set, which is used to determine whether these relationships hold true.
- **Operationalizing the model:** Unlike other disciplines that use analytics (such as web trend analysis), security has some unique requirements. Models need to be out in the wild and need to be continuously measured to ensure that they are performing effectively in a highly dynamic environment. The data scientist is responsible for developing methods to monitor the machine learning model, retrain it, and report on its effectiveness.

Machine Learning Methods and Types

As we've mentioned earlier, machine learning is an automated way of learning about the attributes of security events over a period of time, as new information comes in from different sources. There are many types of machine learning applications that are commonly associated with security systems, but let's focus on the top four.

- **Neural networks:** In the same way that the human brain hard codes a certain skill and becomes expert at it (think of swimming or riding a bicycle), a neural network uses transformational and validation algorithms for continual crosschecking of data, becoming extremely accurate and adept at its task.[2] It can rapidly analyze millions of samples and classify them as false positives, false negatives, true positives, and true negatives. This form of machine learning is an effective tool for identifying rapidly morphing or evasive malware that is undetected by traditional antivirus solutions.[3]
- **Deep learning:** This methodology, which is associated with artificial intelligence, employs neural networks and multiple complex algorithms to reach conclusions by looking at what happened in the past, applying

reasoning (“if this, then that”), and by paying attention to current and predictive data. This form of machine learning, which is associated with self-driving cars, image recognition, and voice recognition, is increasingly being used in cybersecurity solutions. Deep learning is effective because the more it sees, the more it knows.

- **Cognitive computing:** Also a form of deep learning, cognitive computing consists of a complex array of algorithms and mimics the human brain, behavior, and thought processes. According to Lynne Parker, director of the division of Information and Intelligent Systems for the National Science Foundation, cognitive computing, “refers to computing that is focused on reasoning and understanding at a higher level, often in a manner that is analogous to human cognition—or at least inspired by human cognition.” Instead of processing pure data or sensor streams, it works with symbolic and conceptual information and is used for complex decision-making. For example, it can be used to mitigate threats based on analyzing patterns of activity and determining which activities look normal and expected and which appear to be anomalies. It then can raise alerts and take appropriate action to contain threats.[4]
- **Ensemble learning:** In this application, the data is run through several different models, which can improve the accuracy of the results.

McAfee Advances Security Defenses with Advanced Analytics and Machine Learning

McAfee is leading the market by moving advanced analytics and machine learning out of the engineering department and into its product offerings—from research labs to endpoint defenses to advanced solutions, such as sandboxing technologies. These advanced capabilities complement human analysis by our own researchers and enterprise security analysts and dramatically enhance threat detection and remediation. Most importantly, advanced analytics and machine learning are valuable tools for making sense and making use of vast quantities of threat data and enabling faster response to complex threats with fewer resources.

McAfee Labs

McAfee Labs has seen a dramatic escalation in the number of malware samples in a relatively short period of time: “At McAfee Labs nine years ago, we saw less than 200 new threats per day. Today, we see almost 400,000. Our response has to be of equal scale if we are to beat our enemy.”[5]

The enormous volume of malware samples flowing from multiple sources, including McAfee customers, keeps McAfee Labs innovating advanced analytics and classification engines capable of handling vast amounts of threat data at high processing speeds.

Components of this innovation include:

- **A cloud infrastructure connected to millions of global sensors.** Data is gathered from millions of endpoints, gateways, and mobile devices and a broad sweep of IT environments, geographies, and threat actors. McAfee Labs is able to respond to millions of requests around the world via nine data centers, whose data is refreshed every five minutes.
- **Automation to collect and transform data sourced from sensors, third-party sharing communities, historical repositories, and customers.** Data comes from many sources, such as customer submissions, internal submissions by researchers, third-party intelligence feeds (including content from the Cyber Threat Alliance), and products used out in the field. This is further enhanced by 25 years' worth of accumulated data to provide enterprises with knowledge for deep analytics and trend mapping.

  This broad set of data types is stored in the McAfee Labs proprietary classification system and covers one petabyte of data—which is equivalent to 13.3 years of HDTV content (approximately 58,292 movies). The framework and infrastructure gather this data and then pipe it through the various analytics engines, prioritizing and adapting the flow of the data based on the metadata that is acquired.
- **A broad variety of machine learning, advanced analytics, and human interpretations from a variety of Big Data sets.** Suspicious files, both clean and malicious, are consumed and processed at a capacity of one million files per day. These capabilities, which operate without human intervention, collect and transform threat insights, such as file types, indicators of compromise (IoCs), reputation lists, and exploits, into knowledge within minutes, rather than hours.

  At every stage, analytics comes into play, directing the infrastructure on what to do next. For example, analytics helps identify what the submission is (file type, for example), prioritize the file, determine how it should be processed by which modules and in what order based on its attributes, allow the engine to combine modules, and then enable the modules to work with one another. Modules within the analytics engine have a variety of capabilities, including gathering static attributes to identify files, dynamic and run-time analysis, writing to databases, creating reports, and much more.

Real Protect: Dedicated Endpoint Detection of Zero-Day Malware Through Machine Learning

Real Protect is one example of how McAfee is expanding advanced analytics and machine learning out of the realm of research and engineering and into its product offerings. Real Protect leverages multiple machine learning methods, along with other capabilities and security layers, to help enterprises detect, protect, and correct more accurately and swiftly. Real Protect monitors suspicious activity and zero-day malware on endpoints by using machine learning to compare all aspects of a file—both static and dynamic—to swiftly separate clean files from malware.

Static analysis

Real Protect Static trains an algorithm to make the distinction between malware and clean files. It starts by taking millions of malicious and clean file samples that have been stored in the McAfee Labs malware database and other databases and submits them to Real Protect. Real Protect gathers a whole host of static features of the code—file type, compiler used, import hash, entry point, strings, source code language, application programming interfaces (APIs), and many more. It then uses the statistical comparisons created by the machine learning algorithms to develop machine learning models of various file families—both malware and clean.

The next step is to perform statistical analysis and correlation to create a model for later comparisons. When unknown malware samples are submitted, Real Protect Static queries the model library or does a match lookup in the cloud. If there's a match, and the file does indeed look like malware, the endpoint security solution is notified, and then it proceeds with an appropriate security action, such as blocking the file, informing other threat intelligence sources of its findings, or some other action.

Real Protect Static can be used either offline or with an instant cloud lookup, depending on how the solution is designed. The benefit is that Real Protect Static can immediately recognize the difference between clean and dirty files on threat families it is trained for, without needing a traditional antivirus signature. This makes it ideal for detecting zero-day malware, including various kinds of obfuscated (polymorphic, packed, encrypted) threats.

Dynamic analysis

Real Protect Dynamic focuses on behaviors of suspicious files on endpoints. When a suspicious file is detected, Real Protect does an assessment, allowing the file to execute so it can track behaviors like behavioral sequence, process tree, file system changes, registry events, and network communication events. A trace report is generated and sent to the Real Protect Cloud for analysis. As in static analysis, machine learning looks at the report and compares the behavioral attributes to clean files and other malware. Results are then relayed back to the endpoint, and if the file is deemed malicious, appropriate remediation is performed.

The strategy of using both static and dynamic machine learning offers a more complete and effective solution. Static analysis alone would miss key elements of malware and may allow it to run indefinitely on the endpoint. By detecting unusual behaviors—such as piggybacking on legitimate applications or exploitation of vulnerabilities in applications—dynamic analysis provides automated, proactive analysis and can block processes before they do further harm.

McAfee Advanced Threat Defense: Using Machine Learning to Identify Hidden Malware

In another application, McAfee is implementing a sophisticated, advanced machine-learning model based on deep neural network classification in McAfee Advanced Threat Defense.

While many traditional defenses, including sandboxes, utilize rules and automation based on behaviors and static features to detect malware, use of machine learning dramatically scales this technology's ability to identify malicious markers.

After McAfee Advanced Threat Defense's standard sandboxing techniques, machine learning algorithms are applied to detect behavioral similarities and patterns of behavior to identify malicious files. This added logic is continually trained against one of the largest malware data sets in the industry with over 25 years of data and 2 billion files to reference.

Available in McAfee Advanced Threat Defense physical and virtual appliances, the addition of machine learning to sandboxing capabilities improves detection precision. Machine learning is the next step towards providing an autonomous detection solution that will be least susceptible to evasive techniques used by complex and advanced malware.

Conclusion: Better Results Faster with Machine Learning

Machine learning offers the depth, creative problem-solving capabilities, and automation to help security organizations gain significant ground against attackers. It's a powerful tool for processing massive amounts of data for the purpose of malware classification and analysis, especially for unknown threats. Through supervised learning, human researchers can continually develop new training models that expand the understanding and competency of machine learning systems.

The rich heritage of McAfee Labs combined with its continued commitment to invest in the latest advancements in data science materially enrich the insights and models we use to train our cloud analytics technology. Machine learning will result in notable acceleration of detection and remediation, along with a reduced burden on security analysts. Cloud-based machine learning is advantageous because it enables continual updates and cloud sourcing, with no maintenance or effort on the part of end users or enterprises that benefit from it. The more the solution learns from millions of endpoints out in the field, the better it becomes at detecting and adapting to new threats.

Significantly, the power of machine learning lies in its ability to detect never-before-seen threats that human analysts would be unlikely to discover, as we lack the capacity for visualizing and processing large-scale data sets in the same way. By expanding machine learning to a cloud-based system, McAfee Labs and McAfee solutions can provide these benefits to everyone.

1. *McAfee Labs Threats Report, September 2016*
2. http://ubiquity.acm.org/article.cfm?id=958078
3. https://www.technologyreview.com/s/542971/antivirus-that-mimics-the-brain-could-catch-more-malware/
4. http://www.computerworld.com/article/3068185/personal-technology/traditional-security-is-dead-why-cognitive-based-security-will-matter.html
5. Chris Young, address at RSA Conference, 2016

About McAfee

McAfee is the device-to-cloud cybersecurity company. Inspired by the power of working together, McAfee creates business and consumer solutions that make our world a safer place. By building solutions that work with other companies' products, McAfee helps businesses orchestrate cyber environments that are truly integrated, where protection, detection, and correction of threats happen simultaneously and collaboratively. By protecting consumers across all their devices, McAfee secures their digital lifestyle at home and away. By working with other security players, McAfee is leading the effort to unite against cybercriminals for the benefit of all.

www.mcafee.com

McAfee™
Together is power.

2821 Mission College Boulevard
Santa Clara, CA 95054
888 847 8766
www.mcafee.com

McAfee and the McAfee logo are trademarks or registered trademarks of McAfee, LLC or its subsidiaries in the US and other countries.
Other marks and brands may be claimed as the property of others. Copyright © 2018 McAfee, LLC. 3902_0418
APRIL 2018