CYLANCE

# MATH VS. MALWARE

A Cylance Whitepaper

# Table of Contents

# Part I
# Introduction

*"Mathematics is a more powerful instrument of knowledge than any other that has been bequeathed to us by human agency."*

*– Rene Descartes*

The problem—although few want to admit it—is that enterprise security personnel are defending a castle riddled with holes, filled with secret passageways, and protected by ineffective barriers. These weak points are a consequence of poor quality security software, inferior hardware, and—in some cases—backdoors planted by malicious insiders. The end result is a begrudging acceptance that the attackers are winning the war.

Attacks are motivated by a variety of reasons, originate from various locales, and continue to evolve in complexity as technology progresses. As part of this evolution, modern threats commonly employ evasion techniques designed to bypass existing security measures. Simply detecting these advanced threats after the fact is hard enough, let alone protecting an entire organization against them beforehand.

What if there was a better way?
What if the castle could be defended?
What if the threat could be stopped long before the damage was done?

# Part II
# The Human Factor

In order to keep up with modern attackers, security technologies need to evolve alongside them—without relying on human intervention. That's where math and machine learning have the advantage. If we can objectively classify "good" files from "bad" based on mathematical risk factors, then we can teach a machine to make the appropriate decisions on these files in real time.

In the pages that follow, we posit that a math and machine learning approach to computer security will fundamentally change the way we understand, categorize, and control execution of every file. We'll also discuss how Cylance products leverage this approach, and demonstrate just how different they are from every other security offering on the market.

For years industries such as healthcare, insurance, and high-frequency trading have applied the principals of machine learning to analyze enormous quantities of business data and drive autonomous decision-making. At the core of each implementation is a massively scalable data-processing 'brain' capable of applying highly tuned mathematical models to enormous amounts of data in near real-time.

# Part III
# Machine Learning & Security

## Applying Machine Learning to File Classification

Over the past few decades, billions of files have been created—both malicious and non-malicious. In the file creation evolution, patterns have emerged that dictate how specific types of files are constructed. There is variability in these patterns as well as anomalies, but as a whole, the computer science process is reasonably consistent.

The patterns become even more consistent as one looks at different development shops such as Microsoft®, Adobe®, and other large software vendors. These patterns increase in consistency as one looks at development processes used by specific developers and attackers alike. The challenge lies in; identifying patterns, understanding how they are manifest across millions of attributes, across millions of files, and what consistent patterns tell us about the nature of these files.

Because of the magnitude of the data involved, the tendency towards bias, and the number of computations required, humans are incapable of leveraging this data to make a determination as to whether the file is malicious or not. Unfortunately most security companies still rely on humans to make these determinations. They hire a large number of people to look through millions of files to determine which are "good" and which are "bad".

Humans have neither the brainpower nor the physical endurance to keep up with the overwhelming volume and sophistication of modern threats. Advances have been made in behavioral and vulnerability analysis, as well as identifying indicators of compromise, but these "advances" all suffer from the same fatal flaw. They are all based on the human perspective and analysis of a problem—and humans err towards over-simplification.

Machines, however, do not suffer from this same bias.

### Machine Learning Defined

> Machine learning, a branch of artificial intelligence, concerns the construction and study of systems that can learn from data ... The core of machine learning deals with representation and generalization. Representation of data instances are part of all machine learning systems. Generalization is the property that the system will perform well on unseen data instances; the conditions under which this can be guaranteed are a key object of study in the subfield of computational learning theory.
>
> — Wikipedia, 02/13/2014

## How It Works

Machine learning and data mining go hand-in-hand. Machine learning focuses on prediction, based on properties learned from earlier data. This is how Cylance differentiates malicious files from safe or legitimate ones. Data mining focuses on the discovery of previously unknown properties of data, so those properties can be used in future machine learning decisions.

Machine learning leverages a four phase process: collection, extraction, learning , and classification.

## Collection

Much like a DNA analysis or an actuarial review, file analysis starts with the collection of a massive amount of data—in this case files of specific types (executables, .pdf, .doc. Java, flash, etc.). Hundreds of millions of files are collected via 'feeds' from industry sources, proprietary organizational repositories, and live inputs from active computers with Cylance agents on them[1].

The goal with collection is to ensure that:

- One has a statistically significant sample size.
- One has sample files that cover the broadest possible range of file types and file authors (or author groups such as Microsoft, Adobe, etc.)
- One has not biased the collection by over-collecting specific file types

Once these files are collected, they are reviewed and placed into 3 buckets; known and verified valid, known and verified malicious, and unknown. It's imperative to ensure that these buckets are accurate—including malicious files in the valid bucket or valid files in the malicious bucket would create incorrect bias.

## Extraction

The next phase in the machine learning process is attribute extraction. This process is substantively different from the process of behavior identification or malware analysis currently conducted by threat researchers.

Rather than looking for things which people believe are suggestive of something that is malicious, Cylance leverages the compute capacity of machines and data mining techniques to identify the broadest possible set of characteristics of a file. These characteristics can be as basic as the PE file size or the compiler used, and as complex as a review of the first logic leap in the binary. We extract the uniquely atomic characteristics of the file depending on its type (.exe, .dll, .com, .pdf, .java, .doc, .xls, .ppt, etc.).

1 Files are only uploaded from active computers if the customer chooses to enable this option.

## Extraction (Cont.)

By identifying the broadest possible set of attributes, Cylance removes the bias introduced by the manual classification of files. Use of hundreds of thousands of attributes, also substantially increases the cost for an attacker to create a piece of malware that is not characterized by Cylance.

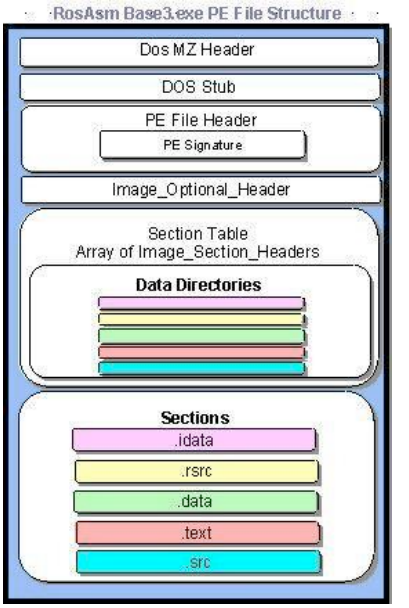


Figure 1: Structure of a PE File

The result of this attribute identification and extraction process is the creation of a file genome very similar to that used by biologists to create a human genome. This genome is then used as the basis for which mathematical models can be created to determine expected characteristics of files, much like human DNA analysis is leveraged to determine characteristics and behaviors of cells.

## Learning & Training

Once the attributes are collected, the output is normalized and converted to numerical values that can be used in statistical models. It's here where the vectorization and machine learning are applied to eliminate the human impurities and speed analytical processing. Leveraging the millions of attributes of files identified in extraction, Cylance mathematicians then develop statistical models that accurately predict whether a file is valid or malicious.

Dozens of models are created with key measurements monitored to ensure the predictive accuracy of the final models used by Cylance products. Ineffective models are scrapped, and effective models are run through multiple levels of testing. The first level starts with a few million known files, and later stages involve the entire file corpus (tens of millions of files). The final models are then extracted from the test corpus and loaded into Cylance's production environment for use in file classification.

It's important to remember that for each and every file, thousands of attributes are analyzed to differentiate between legitimate files and malware. This is how the Cylance engine identifies malware—whether packed or not, known or unknown—and achieves an unprecedented level of accuracy. It divides a single file into an astronomical number of characteristics, and analyzes each one against hundreds of millions of other files to reach a decision about the normalcy of each characteristic.

## Classification

Once the statistical models are built, the Cylance engine can be used to classify files which are unknown (e.g., files that have never been seen before or analyzed by another white list or black list). This analysis takes only milliseconds and is extremely precise, because of the breadth of the file characteristics analyzed.

## Classification (Cont.)

Because the analysis is done using statistical models, the classification is also not completed in a black box. Cylance provides the user with a "confidence score" as part of the classification process. This score provides the user with incremental insight that they can use to weigh decisions around what action to take on the specific file—block, quarantine, monitor or analyze further.

There is an important distinction between the machine learning approach and a traditional threat research approach. With the mathematical approach, Cylance builds models that specifically determine if a file is valid or malicious. It will also return a response of "suspicious" if our confidence about its malicious intent is less than 20% and there are no other indications of malicious intent. In so doing, the enterprise gains a holistic perspective on the files running in their environment. It also eliminates the current industry bias in which threat researchers only determine if a file is malicious and white list vendors only determine if a file is good.

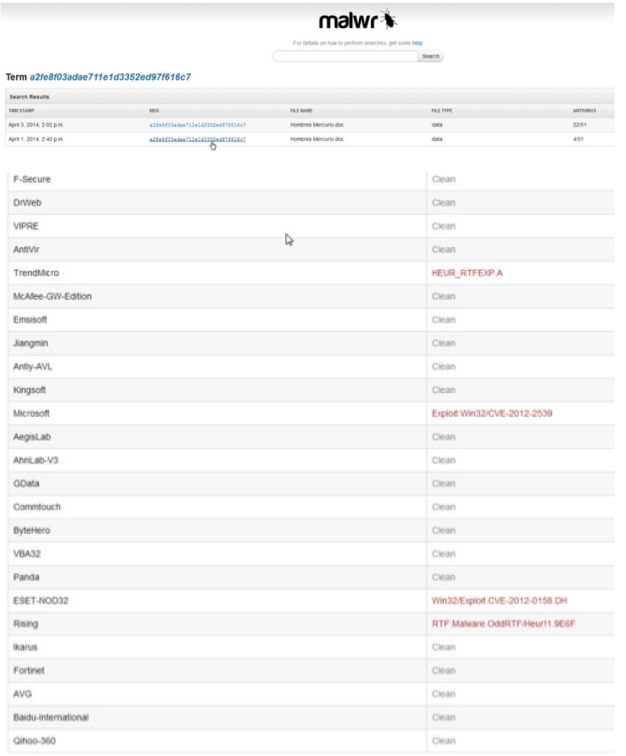
Other than the obvious benefits of detecting a larger amount of threats, there are more subtle benefits to this approach; every file that is analyzed is evaluated using the classification algorithms. While this may seem straight-forward, traditional antivirus (AV) vendors only evaluate a specific file against a finite list of signatures designed to detect malware based on human analysis. Even if they use some automated techniques, they are limited to creating signatures based on specific parts of files that were previously identified as known malware by a human. Not only is there little-to-no proactivity possible with these techniques, they simply classify objects that do not match any particular signature as good. The Cylance engine, on the other hand, analyzes every sample, and provides a definitive classification of all files whether they bad, suspicious or good. This provides the security team with a clear understanding of exactly what is running in their environment.

# Part IV
# Cylance Vs. The Real World

As an example, Cylance prevented 0-days like the Microsoft Word RTF vulnerability (CVE-2014-1761) before it was even observed in the wild—without any foreknowledge.

*(See next page for details.)*

News about a new Microsoft Word vulnerability—one that opens the door for remote code execution—finally made the rounds on 04/01/2014 to malwr.com and, at the time, was detected by only 4 out of 51 AV engines.


malwr

Term a2fe8f03adae711e1d3352ed97f616c7

| | |
|---|---|
| F-Secure | Clean |
| DrWeb | Clean |
| VIPRE | Clean |
| AntiVir | Clean |
| TrendMicro | HEUR_RTFEXP.A |
| McAfee-GW-Edition | Clean |
| Emsisoft | Clean |
| Jiangmin | Clean |
| Antiy-AVL | Clean |
| Kingsoft | Clean |
| Microsoft | Exploit:Win32/CVE-2012-2539 |
| AegisLab | Clean |
| AhnLab-V3 | Clean |
| GData | Clean |
| Commtouch | Clean |
| ByteHero | Clean |
| VBA32 | Clean |
| Panda | Clean |
| ESET-NOD32 | Win32/Exploit.CVE-2012-0158.DH |
| Rising | RTF.Malware.OddRTF/Heur!1.9E6F |
| Ikarus | Clean |
| Fortinet | Clean |
| AVG | Clean |
| Baidu-International | Clean |
| Qihoo-360 | Clean |

## Microsoft Security Advisory 2953095

This topic has not yet been rated – Rate this topic

### Vulnerability in Microsoft Word Could Allow Remote Code Execution

Published: March 24, 2014 | Updated: April 8, 2014

**Version:** 2.0

On this page
General Information
Acknowledgments
Other Information

### General Information

#### Executive Summary

Microsoft has completed the investigation into a public report of this vulnerability. We have issued MS14-017 to address this issue. For more information about this issue, including download links for an available security update, please review MS14-017. The vulnerability addressed is the Word RTF Memory Corruption Vulnerability – CVE-2014-1761.

### Acknowledgments

Microsoft thanks the following for working with us to help protect customers:

- Drew Hintz, Shane Huntley, and Matty Pellegrino of the Google Security Team for reporting the Word RTF Memory Corruption Vulnerability (CVE-2014-1761)

### Other Information

#### Microsoft Active Protections Program (MAPP)

To improve security protections for customers, Microsoft provides vulnerability information to major security software providers in advance of each monthly security update release. Security software providers can then use this vulnerability information to provide updated protections to customers via their security software or devices, such as antivirus, network-based intrusion detection systems, or host-based intrusion prevention systems. To determine whether active protections are available from security software providers, please visit the active protections websites provided by program partners, listed in Microsoft Active Protections Program (MAPP) Partners.

#### Feedback

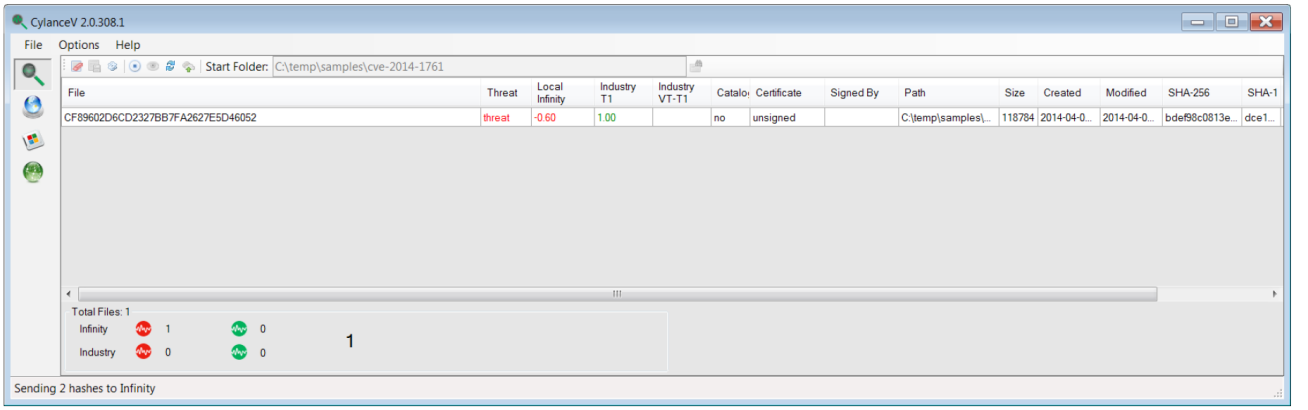
- You can provide feedback by completing the Microsoft Help and Support form, Customer Service Contact Us.

#### Support

- Customers in the United States and Canada can receive technical support from Security Support. For more information, see Microsoft Help and Support.
- International customers can receive support from their local Microsoft subsidiaries. For more information, see International Support.
- Microsoft TechNet Security provides additional information about security in Microsoft products.

The Cylance engine, however, detected the same malware *(a2fe8f03adae711e1d3352ed97f616c7)* instantaneously—without the need for any updates. CylancePROTECT (not pictured) prevented this exploit from executing.

*CylanceV product pictured below.*



# Part V
# Future-Proof Security

By applying math models to the endpoint, the Cylance engine easily surpasses all traditional methods of malware detection and prevention. Our mission is to stop the execution of bad files before they can cause any damage. With this approach, the endpoint remains secure and unviolated even if the file is resident on disk.

## CylancePROTECT

CylancePROTECT is our flagship enterprise product that harnesses the power of the Cylance engine to prevent the execution of advanced threats in real-time on each endpoint in the organization.

### Key Features:

- Protection and detection of previously undetectable advanced threats.
- Cloud enabled, but not cloud dependent for sensitive environments.
- No daily .DAT updates which eliminates the need for an "always on" connection.
- Extremely low performance impact. Runtime execution dramatically reduces overhead.
- Easy to deploy and manage with a purpose-built web interface.

PROTECT provides real-time detection and prevention of malware. It operates by analyzing potential file executions for malware in both the Operating System (O/S) and memory layers and prevents the delivery of malicious payloads. The memory protection is designed to be extremely low-touch as to not incur a heavy performance overhead. Instead, memory protection strengthens the basic O/S protection features like DEP, ASLR, and EMET by providing an additional layer to detect and deny certain behaviors which are very commonly used by exploits.

These two core functions are supported by a variety of ancillary features necessary for enterprise functionality including:

- Whitelist and blacklist support for administrative granularity
- Detect-only mode (audit mode)
- Self-protection (prevention against user tampering)
- Complete control, update and configurability from the management console

## CylanceV

CylanceV is an Incident Response and Forensics tool that enables enterprises to seamlessly integrate the power of the Cylance engine into the SOC, CSIRT, Helpdesk and everywhere in-between. CylanceV is a lightweight, easy to integrate utility for endpoint file analysis that scales to meet any deployment size. CylanceV is available in two forms. The first is a REST API that provides direct access to the Cylance engine. The API can be leveraged programmatically (Python samples are available from Cylance). This option requires the enterprise to pass files up to Cylance's secure cloud for analysis.

## CylanceV (cont.)

The second form is a local utility that includes the attribute extraction engine and the statistical models that can be run on any Windows or Linux machine within the boundaries of the enterprise (eliminating the need to pass files to the Cylance's secure cloud).

Built for the purpose of rapid deployment and enterprise wide assessments, CylanceV puts the power of machine learning in a format built for malware hunters.

## Cylance Professional Services

The expert Cylance Professional Services team empowers organizations in several ways. Our team can; unearth vulnerabilities in existing infrastructure, identify previously undetectable threats, assist in attack recovery, implement "best practice" polices and deploy products to detect—and prevent—attacks before they can impact the business.

### Key services offered:

- Compromise Assessments
- Penetration Testing
- Forensic Investigation
- Specialty and Custom Services (CIKR, Embedded, ICS)
- Secure Software Development

The Cylance Professional Services team leverages the power of the Cylance engine for each and every service, allowing for deep insight and analysis which cannot be offered by any other vendor.

## Summary

Cylance truly believes that mathematical modeling and machine learning are the keys to a secure future. Each product and service we offer is tightly integrated into the Cylance engine, providing unparalleled accuracy and insight into the modern threat landscape. Best of all, by continuously learning and training based on new data, the Cylance engine is truly "future-proof" and will not lose efficiency over time—even as the attackers morph their strategies.

### About Cylance

Cylance, Inc. is a global provider of cybersecurity products and services that is changing the way companies, governments, and end-users proactively solve the world's most difficult security problems. Through our holistic security methodology, Cylance couples the understanding of a hacker's mentality with algorithmic intelligence and best practices to be truly predictive and preventive against advanced threats.



+1 (877) 973-3336

sales@cylance.com

www.cylance.com

46 Discovery, #200
Irvine, CA 92618
USA