Unknown Unknowns

"There are known knowns: there are things we know we know. We also know there are known unknowns: that is to say we know there are some things we do not know. But there are also unknown unknowns — the ones we don't know we don't know. And if one looks throughout the history of our country and other free countries, it is the latter category that tends to be the difficult ones." *Donald Rumsfield, former US Secretary of Defense.*

New Waters: Navigating the Threat Landscape

To borrow the framework created by the former US Secretary of Defense, Donald Rumsfield, cyber-attacks can be broadly categorized in three ways:

1. **'Known knowns':** This is our low-hanging fruit, like spam and viruses, which has been seen before. It can be found through open-source intelligence (OSINT).
2. **'Known unknowns':** This is malware which has been updated to alter its underlying code or signature. Although it is new, such attacks are often anticipated and can be traced back to known threats.
3. **'Unknown unknowns':** These are the threats which are not expected. The techniques are novel, no signatures exist, and to the majority of security tools they are invisible.

This report shines a light on the fundamental limitations of traditional defenses in catching never-before-seen attacks. It offers a more nuanced solution which abandons rules and signatures in favor of a self-learning approach to security.

Rules are Made to be Broken

A rules-based approach to defense involves taking broad brushstrokes to determine what activity should be allowed and what should not. This often comes in the form of a series of "if / then" statements, for example: "if a file over 1 gigabyte is sent to country X, then raise an alert."

While at first sight this seems like a reasonable model, in practice, it's unworkable in the modern enterprise.

Such an approach not only fails to account for novel or disguised attacks but can also lead to a barrage of false positives. This overwhelms security teams, and it doesn't take long before the alerts are habitually ignored.

Moreover, the rules must be continually updated and edited as the company changes, new employees join, new divisions are organized, new products are launched, and it's an impossible task to keep up. In the case of a seismic change – a sudden shift to remote working, for example – the playbook must be rewritten from scratch.

Zero-day exploits have been responsible for some of the most high-profile attacks in the past year, including SolarWinds and Hafnium.

Email-borne threats leverage new domains in their thousands, which can be bought for pennies and bypass 'bad domain' deny lists.

New ransomware strains, updated malware, and morphed malicious files do not have any signatures associated with them.

Limitations of Signatures

Most existing security tools rely on the premise that any given cyber-attack has been seen before. Its signature is therefore known, and can be recognized when seen again in future attacks. This is the basis of antivirus, firewalls, and other perimeter security.

Over time, as the list of known threats became progressively larger and less manageable, this approach became antiquated.

Recognizing that the vast number of existing viruses and malware was no longer knowable, defenders sought novel solutions to detect new variations of known threats (sometimes referred to as 'known unknowns').

The idea behind 'known unknown' threats is that while their precise nature is not understood in detail, security solutions 'know' they are out there and are familiar with their general profile based on similar threats these solutions have encountered. Security defenders have therefore worked hard to create more sophisticated antivirus solutions and 'next-generation' firewalls that identify these cyber-attacks from common patterns found in their code.

'Unknown Unknowns' and the Death of Signatures

However, there is one more category of threat that has proven both to be the most evasive and the most damaging of all. This kind of attack can be referred to as an 'unknown unknown': not only are the indicators of compromise absent from any static deny lists but traditional security teams and solutions don't even know to look for them.

With a traditional approach, organizations either go extremely narrow in their criteria, in the form of specific signatures, or incredibly broad, in the form of blanket policies. Both of these approaches rely on the defender's ability to define the threat in advance.

Attackers generally know about the cyber security tools they are trying to evade and take measures to get around them. Cyber-criminals constantly update their attack infrastructure (email domains, IP addresses, strains of malware) and innovate (fundamentally changing techniques, targeting supply chains, employing armies of botnets) because they know most tools are blind to new attacks.

Detecting the Undefinable: A Self-Learning Approach

Breaches like SolarWinds have proven that attackers will continue to innovate and get inside your systems – whether through the supply chain, a careless employee through a phishing email, or a vulnerability in your cloud infrastructure.

No one can predict the next vector of attack, so defenders' mindset must shift to asking, 'What do we do once the threat gets inside?' A security system must be able to identify attacks once they get in.

Self-Learning AI does not rely on pre-defined signatures and rules to find novel attacks. Instead it learns the digital DNA of an organization to identify threats.

Traditional security tools are always one step behind the attacker, reacting to a new threat and quickly updating lists once the first attack has run its course. Instead, Self-Learning AI proactively understands 'normal' for every user and device within an organization, and all the connections between them. This enables it to detect subtle anomalies in real time, stopping truly novel threats as they emerge – not just those previously seen in the wild.

Self-Learning AI: Real-World Threat Finds

Signatureless Ransomware Stopped by Darktrace Antigena

At an electronics manufacturer, Darktrace Antigena, Darktrace's Autonomous Response capability, stopped a never-before-seen ransomware attack in its earliest stages.

An infected device was observed making an unusually large number of connections, writing multiple SMB files, and transferring data internally to a server it did not usually communicate with. Hundreds of Dropbox-related files were then accessed on SMB shares, with several of these files becoming encrypted, appended with a [HELP_DECRYPT] extension.

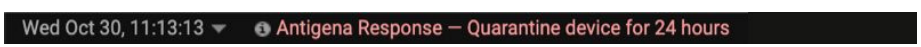

Figure 1: Darktrace Antigena responds 1 second after ransomware was detected

Darktrace Antigena kicked in a second later. Powered by Self-Learning AI, it understood the organization's normal 'pattern of life' and recognized this behavior as a systematic ransomware attack, stepping in within 2 seconds to stop the encryption. By the time it took action, only four of these files had been successfully encrypted.

This strain of ransomware was not associated with any publicly known indicators of compromise. Nevertheless, Darktrace was able to detect this attack based purely on its comprehensive understanding of 'normal' across the organization.

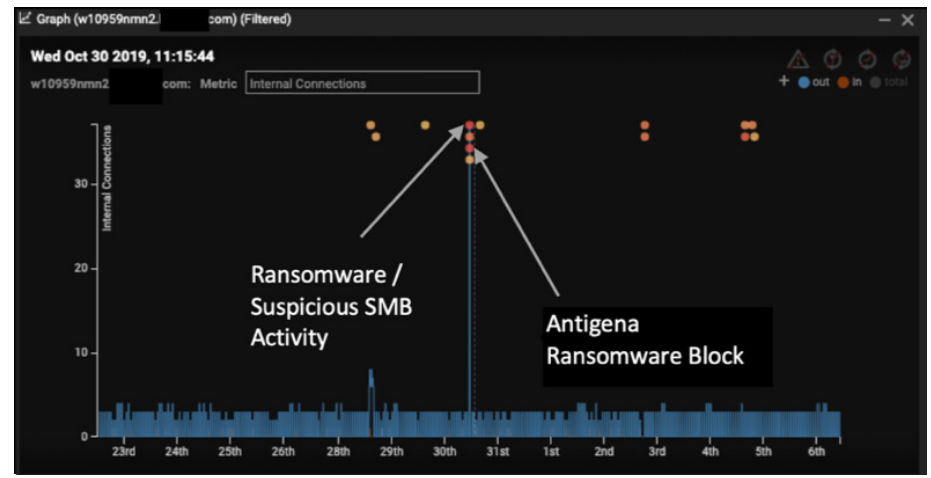

Figure 2: Four model breaches observed on October 30th and a dotted line representing Darktrace Antigena's actions

"With its self-learning technology that requires no training data sets or manual configuration, Darktrace is capable of uncovering 'unknown unknowns' and alerting us to abnormal behavior in real time."

CISO, SNCF/ Avancial

Hafnium Zero-Day Exploit Neutralized by Self-Learning AI

In early December 2020, Darktrace autonomously detected and investigated a sophisticated cyber-attack that had targeted a customer's Exchange server. On March 2nd, 2021, Microsoft disclosed an ongoing campaign by the Hafnium threat actor group leveraging Exchange server zero-day vulnerabilities.

Based on similarities in techniques, tools, and procedures (TTPs) observed, Darktrace has assessed with high confidence that the attack in December was the work of the Hafnium group.

The intrusion was detected at a critical national infrastructure organization in South Asia. One hypothesis is that the Hafnium group was testing out and refining its TTPs, potentially including the Exchange server exploit, before running a broad-scale campaign against Western organizations in early 2021.

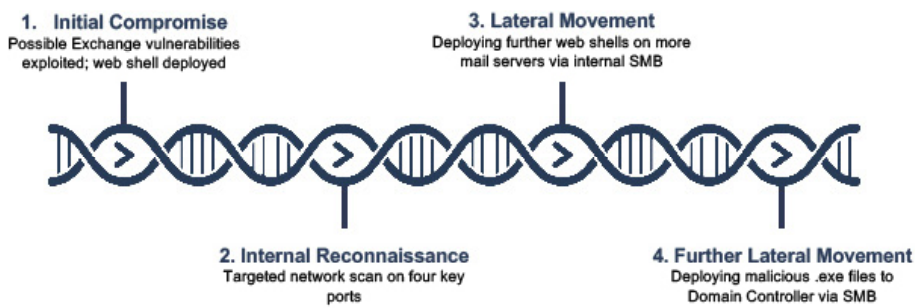

Figure 3: Timeline of the attack from early December 2020

As soon as the attackers gained access via a web shell, they used the Exchange server to scan all IPs in a single subnet on ports 80, 135, 445, and 8080.

This particular Exchange server had never made such a large number of new failed internal connections to that specific subnet on those key ports. As a result, Darktrace instantly detected the anomalous behavior, indicative of a network scan.

A single click in the Darktrace user interface revealed further details about the written files. The full file path for the newly deployed China Chopper web shells was:

ProgramFiles\Microsoft\ExchangeServer\V15\FrontEnd\HttpProxy\owa\auth\Current\themes\errorFS.aspx

The file path and file name of the actual .aspx web shell bear very close resemblance to the Hafnium campaign details published by Microsoft and others in March 2021.

The organization had several thousand devices defended by Darktrace. Nevertheless, over the period of one week, the Hafnium intrusion was in the top five incidents highlighted by Darktrace's autonomous investigation tool, Cyber AI Analyst.

Even a small or resource-stretched security team, with only a few minutes available per week to review the highest-severity incidents, could have seen and inspected this threat.

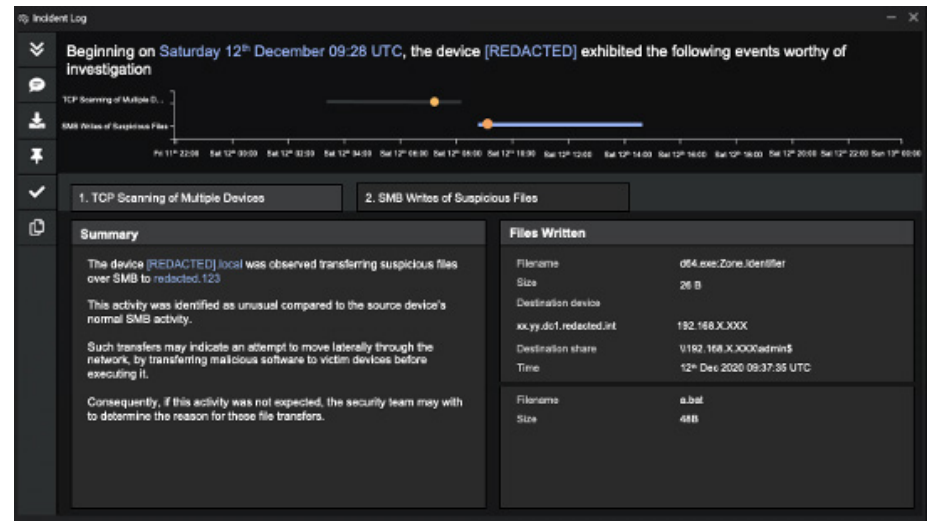

Figure 4: A Cyber AI Analyst report showing unusual SMB activity

COVID-19 Fearware Phishing Emails Held Back

Last year, Darktrace for Email observed an email threat trend where attackers claimed to be from the Center for Disease Control and Prevention, purporting to have emergency information about COVID-19. Exploiting a sense of collective fear, uncertainty, and doubt is a common tactic for cyber-criminals, but such campaigns are difficult to defend against with a rules-based approach as they contain unique terms and content.

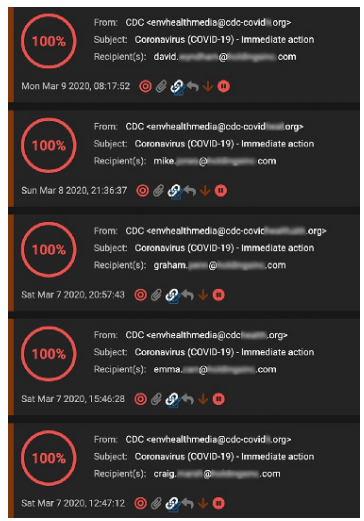

Figure 5: While other defenses failed to block these emails, Darktrace for Email immediately marked them as 100% unusual and held them back from delivery

Taking a sample COVID-19 email seen in a customer's environment, Darktrace for Email saw a mix of domains used in what appears to be an attempt to avoid pattern detection. It would be improbable to have the domains used on a list of 'known bad' domains anywhere at the time of the first email, as it was received a mere two hours after the domain was registered.

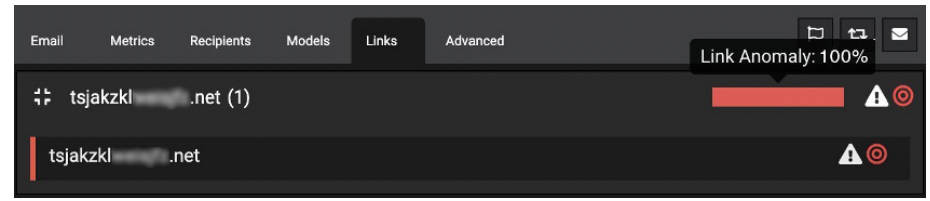

Figure 6: The link was determined to be 100% rare for the enterprise

Darktrace determined that the domain in the 'From' address was rare by correlating contextual information across the customer's entire digital environment, including network data. The emails' KCE, KCD, and RCE scores indicate that it was the first time the sender had been seen in any email: there had been no correspondence with the sender in any way, and the email address had never been seen in the body of any email.

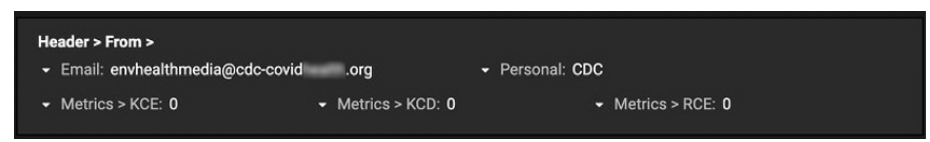

Figure 7: KCE, KCD, and RCE scores indicate no sender history with the organization.

Powered by Self-Learning AI, Darktrace for Email correlated the above findings to discern that these emails were anomalous to the business and immediately removed them from the recipients' inboxes. It did this for the very first email and every malicious email thereafter.

"For us, deploying Darktrace wasn't an option; it was a necessity in staying ahead of today's advanced and unpredictable threats."

Director of Innovation and Technology, City of Auburn