WEBROOT®
Smarter Cybersecurity™

# What Makes Real Threat Intelligence

Defining terms for a rapidly evolving discipline

**Written by**

Hal Lonas
Chief Technology Officer
Webroot

## Table of Contents

## Introduction

First popularized in terminology several years ago, threat intelligence means many things to many people. This is due, in part, to the wide variety of producers, consumers, formats, intended uses, and quality of this type of data. This wide variety of definitions is exacerbated by the spectrum of qualitative and quantitative data called threat intelligence.

This paper will arm you with a set of criteria to gauge threat intelligence by its source, intended audience, and use cases to help narrow down the field to a few basic types. It also explores the quality of threat intelligence by examining positive and negative aspects of these types and how they are derived.

We'll also discuss how to gauge the real value of threat intelligence, and how, when properly developed and applied, it can bring enterprise-grade security to businesses and individuals with limited budgets and personnel.

## Defining Threat Intelligence

The concept of "threat intelligence" has different meanings to different audiences. It can refer to:

» Human-readable information to guide or inform threat researchers

» Machine-readable raw data flowing into a system from device logs or telemetry

» Unvetted and/or crowd-sourced lists

» The results of analysis of one or more of the above forms to produce high-quality information exhibiting broad coverage and high accuracy

## Confusing Terminology

You've likely heard adjectives like "actionable" and "real time" as vendors attempt to differentiate themselves, but such terms only contribute to more confusion. For example, while "actionable threat intelligence" may sound strong, it also creates a number of questions; who or what takes which action and by what means? Does the threat intelligence inform a threat researcher to take defensive action against a threat by isolating endpoints or closing a network loophole? Or does the threat intelligence trigger an action through some form of automated orchestration at a policy enforcement point, or a point in time when a potentially malicious activity can be stopped in the network or at execution? As you can see, terms like "actionable" don't lend much meaning without deeper explanation.

"Real time" is another overused term. Some companies have used "real time" to mean that the collection of data is done in real time, but the collected data sits on a server awaiting processing. One should define the sequence of events that occurs, which data flows are important enough to warrant operating in real time, and which are not. For example, some threat intelligence systems can detect endpoint browsing to a given URL that has attributes of a financial or social media website, and classify it as phishing or not—all in real time and without the user noticing any delay in browsing behavior.

## Cybersecurity Hinges on Threat Intelligence

At its core, cybersecurity is an information problem. If you knew that a URL was malicious, you wouldn't click the link. If your firewall recognized that an incoming IP was from a spammer, it wouldn't accept the connection. If your mobile device was informed that a free app in the store was bad, it wouldn't download or run it. Getting the right information to the right place at the right time is crucial, whether the defense point is the endpoint, a network security choke point, or left with a human decision maker to interpret.

The most sophisticated form of threat intelligence contains enough information to make an informed policy decision; it's not just raw data. That's what we mean when we say "real threat intelligence".

## Use Cases for Real Threat Intelligence

1. The inbound eCommerce security layer of a well-known, low-cost airline was being attacked by bad actors probing the airline's systems and defenses for weaknesses. The airline layered in real threat intelligence at the network perimeter, which enabled them to automatically and precisely block the malicious forays at the outermost defense level. This real threat intelligence was integrated in a way that added virtually no latency to the inbound requests. The added defense removed a small, but highly malicious, fraction of inbound network traffic and has saved the airline thousands of hours in lost IT time, not to mention savings in terms of cost and customer trust had a breach occurred.

2. A popular WiFi access point vendor was seeing rapid sales growth to small businesses and other segments. However, customers were complaining that WiFi users were visiting low-reputation websites and getting infected with malware. The vendor implemented a real threat intelligence layer on the devices that adapted to the limited device resources, down to the smallest amount of memory and storage, and provided comprehensive protection against malicious URLs. WiFi users were immediately protected from these attacks on the network.

3. A small business hired a managed service provider (MSP) to manage their IT infrastructure. The MSP had adopted a modern technology stack, including next generation cloud-connected endpoint security and Domain Name System (DNS) protection from the same vendor. Both products used real threat intelligence to protect users from rogue applications, malware, malicious IPs, and URLs. The MSP could also run security awareness training through their cybersecurity provider to ensure regulatory compliance and to educate end users on good security practices. The small business reports a 55% reduction in downtime due to infected machines, and the MSP's support cases have dropped significantly since the new products were introduced.

## Importance of Visibility and High Fidelity Source Data

Raw threat data can come from many sources. These might include sensors, crawlers, honeypots, virtual machines, crowdsourcing, and endpoints. Some of these sources, like honeypots, are set up for the express purpose of attracting attacks. Others, like sandboxes, appear to malware to be an exploitable system and are typically run on a virtual machine (VM) to scale systems for demand. Further sources, like crawlers, attempt to visit every page on the internet to classify URLs. Crowdsourcing solicits input from humans to create threat intelligence. While each of these is widely used and has some advantages, each one also has serious drawbacks.

## Common Data Sources and Their Drawbacks

**Crawlers** must simulate a browser environment via the user agent and geography of the crawling entity, and they can't fully simulate cookies, URL arguments, and other artifacts of real humans browsing web pages. Additionally, many web-borne attacks will not be not triggered by a crawler that does not have the right combination of these factors, rendering the crawler blind to possible malware URLs.

**Honeypots** must appear to be an infect-able machine, but not too predictably, or else the attack will not display its true nature. Honeypots simulate open resources on the internet, including mail servers, proxies, database and web servers, and unprotected or unpatched web services. Attackers constantly run internet probes from rapidly changing IPs, hoping to find and exploit under-defended internet resources. Many modern attacks can now identify a honeypot simulation and therefore will not attack. This limits the usefulness of the honeypot's ability to catch modern attacks and their sources.

**Sandboxing** is another technology security vendors have used over the years. Sandboxes attempt to emulate a real user's endpoint environment. Some security products run a possible malware executable in the sandbox to check its behavior before allowing it to run on real users' systems. However, simulating the wide variety of possible endpoint environments is a daunting task, and many malware variants rely on environmental factors when they run. Some of the environmental variables are:

» Operating system and patch level (Windows® 7, 8, 10, Apple® macOS®, Android™, iOS®, etc.)

» Browser, version, and user agent (Apple® Safari, Microsoft® Edge, Google® Chrome, Windows® Internet Explorer, etc.)

» Installed applications (Microsoft® Office, Adobe® Acrobat, Windows® Media Player, Apple® iTunes, etc.)

Sandboxes are also typically set up in a virtual environment for throughput, scale, and cost considerations. Many variants of malware now detect virtual environments and therefore will not run, thus evading detection by the very sandboxes set up to catch them. Furthermore, sandboxing introduces latency into the system while users wait for the sandbox verdict. Knowing this, malware authors can code their attacks to delay malicious behavior long enough to force the sandboxes to give up, time-out, and return an all-clear.

**Crowdsourcing** relies on humans to report real or perceived threats they have experienced or observed. Studies have shown that most humans are not good at detecting modern malware and phishing attacks, and that they either under-report or report false positives. Therefore, crowdsourced data is typically very noisy and full of inaccurate data.

## High Fidelity, Real Source Data

The best visibility into threat data will comes from a live or "real" system which includes:

» **Real people.**

Real people take realistic actions online, such as clicking links in emails, inadvertently installing rogue or unwanted applications, visiting malicious websites, and getting infected with malware. It's a challenge to protect people while simultaneously using observations of their behaviors to create new threat intelligence, but the practice provides great visibility and fidelity. Additionally, these users' privacy must be protected. Systems should only collect threat telemetry, not sensitive user information.

» **Real machines.**

There are millions of environmental variations on real machines, such as Windows, Apple, Android, and internet of things (IoT) devices. When you consider CPUs, graphics, networks, attached hardware, browsers, installed applications, and data flows, it's an astonishing variety. All of this has an impact on security events. Typically, when a large group of endpoints is exposed to malware, some percentage of them will see the malware activate and carry out its mission, while others will not. As with sandboxing, a lot of modern malware can tell when it's not in a real environment.

» **Real time.**

The only constant in the threat landscape is change. Change occurs at many different levels. Some changes are relatively slow moving, such as the evolution of malware to ransomware to cryptoware. Other changes occur relatively quickly, like the minutes-long lifespan of a phishing URL or the duration of time that an attacking IP might be used to probe for weaknesses in a given target's defenses. Regardless, time is a critical element of threat data, and blacklists distributed via traditional means get stale so fast that they may become irrelevant within seconds of being published. Modern threat intelligence systems have to be implemented as services to keep up with such rapid variability. These services might be implemented with polling, or publish/subscribe models to propagate recent changes. Time to live (TTL) of any data must also be considered, since there can be a tradeoff between the efficiency of communications and the accuracy of the intelligence. In other words, if the TTL is too short, the system can be very demanding on network bandwidth. But if the TTL is too long, the system may not be able to keep up with the rapid changes mentioned above.

## Other Aspects of High Fidelity Source Data

Threat visibility across various entities is also important. It must take consumers and businesses of all sizes and geographic diversity into account. Many malware variants and phishing URLs exhibit different behavior depending on the geographic location of the IP of the accessing endpoint.

Additionally, it's critical to get information across a variety of endpoints and networks. For instance, DNS usage information can reveal very specific threats, such as botnets, and it can do so better than other detection methods.

## Threat Context Can Make Important Connections

Above all, threat intelligence must be contextual. Each IP address, URL, file, domain, application, or other internet object must be considered not only individually, but also collectively. Its relationships to other internet objects are important in determining both its current state and its potential for future malicious activity. For example, imagine that an executable file is categorized as malicious. Knowing the URL source of that malicious executable, combined with the right kind of contextual threat intelligence, could prevent the executable from ever landing on an endpoint by preventing a user from visiting the bad URL in the first place. Knowing what IPs it uses to coordinate or where it sends stolen intellectual property can help take down other parts of the malicious network. As a further step, combining probable botnet observations from a DNS service with threat telemetry from endpoints can quickly identify and protect a network from an evolving threat that is actively propagating on a business local area network (LAN).

Contextual analysis also helps triangulate and improve confidence in the reputation of a given URL, file, IP, or application. In other words, if we suspect a URL is malicious and we see a suspicious or unknown file from that same source, we can use the accumulated evidence to enhance our evaluation of the file and protect users more proactively. Additionally, Whois and domain registration information, or even a lack thereof, can signal certain threat relationships.

## Aggregating and Sharing Data vs. Mining and Creating Real Threat Intelligence

Many vendors who provide threat intelligence do not generate it themselves; they simply aggregate from other data sources. Their data may consist of open source lists and feeds, for instance. These sources and aggregators are typically of very low quality because such simplistic solutions cannot effectively analyze the constituent inputs for accuracy. They also cannot resolve conflicts within the data.

For example, VirusTotal.com lists the opinions of a variety of vendor engines on malicious files and URLs. While this data undoubtedly has value, it is very difficult for the average consumer to make such information useful. Do you vote and use a threshold of opinions to guide your actions, or do you have a favorite source among the many listed, or some combination of the above?

Users who attempt to derive their own threat intelligence using low quality, incomplete, or conflicting sources are subject to a wealth of issues with the accuracy and consistency of their data. An overabundance of false positives and false negatives are typical with these do-it-yourself solutions.

## Incorporating Multiple Sources Via Reputational Analysis

A better approach is to incorporate a variety of source data opinions into threat intelligence, and then use comprehensive analysis to develop a reputation per source. This reputation can be derived over time by comparing the accuracy of a source's opinions with factual observations. As a source's accuracy rises or falls, its inputs can be properly weighted to lend to the overall accuracy of threat intelligence. This type of analysis isn't simple; it requires sophisticated statistical math to improve the results. It also requires access to a factual observation.

Keep in mind that data derived from these sources is far from complete. Depending on others for threat intelligence feeds is reminiscent of the days when security providers would share virus signatures. While providers did share, they often imposed delays on data feeds to retain their intellectual property and maintain an advantage over their competition.

## Discovering New Threats

For a complete approach, the supplier must have access to raw information that can only come from real products and real machines in real time. In short, companies with real products are best positioned to become the true producers of accurate threat intelligence.

Other sources—even crawlers, honeypots, etc. that can introduce issues as described previously—can add value to a complete threat intelligence solution.

» Sophisticated crawlers can be seeded with URLs where endpoints have encountered malware files. This can lead to the discovery and cataloguing of more malware files on other pages and subdomains related to the seed URL used for crawling.

» Through careful configuration, modern honeypots can be almost indistinguishable from a real exploitable resource. However, honeypots are passive, which means they have to wait for someone to attack them.

» Proactive IP scanning sends out probes to the IP space that can take the form of web requests, DNS requests, requests to proxy/forward web requests or forward emails, etc. Responses can then be analyzed to determine if they are coming from a legitimate service or an imposter or bad actor.

Each of these methods produces large amounts of raw data, which must be collected and processed on an ongoing and timely basis to be useful as threat intelligence.

Once we have access to threat telemetry from real people, real products, feeds, and lists from a variety of sources (complete with conflicts, missing information, and inaccuracies), the next step is to turn this avalanche of raw information into useful threat intelligence.

## From Raw Data to Real Threat Intelligence

Converting raw data to real threat intelligence is a daunting prospect. There aren't enough threat researchers to watch indicators of compromise (IOCs) on screens in a network operations center (NOC) or to sift through all the data manually.

There is no single machine learning algorithm or approach, no single source of data, no single infrastructure, no perfect database, no single delivery method, and no single policy enforcement technique that works across all inputs, threats, use cases, and outputs.

## Fine-Grained Reputation Scores vs. Binary Threat Determinations

Threat intelligence should give a more complete picture of internet objects than simply good or bad. It should explain the relative risk of a given threat, and allow you to research why the object is currently classified as good or bad, and offer insight into the nature of the object and how it has behaved or changed status over time.

An analogy from our everyday lives is our credit score—financial institutions don't use a binary "yes" or "no" when they make a lending decision. Instead, they look at an individual's credit score and consider the amount of the loan, interest rate, and degree of risk that the lending institution is prepared to underwrite. Our credit score rises or falls based on our financial capability and history, and informs the lending institution of the risk they take on by lending us a sum of money.

In Figure 1, a domain which has had a good reputation score for some time suddenly gets attacked by malware. Perhaps the web server was not patched or malware masquerading as an advertisement was hosted. We rapidly detect the new security posture of the URL and update its reputation which drops to a low value, and endpoint security providers block or warn users trying to access it.

At some point the website is cleaned up and it becomes safe to visit. However, its good reputation is not instantly regained. The curve climbing to Good reputation may be fast—on the order of minutes—for a highly reputable site such as the fictitious Celebrity.com, or it may be slow—on the order of days—for an unmaintained or lax-in-security site such as (the also fictitious) Sketchysite.com. Slow recovery may indicate a high likelihood that the URL was not completely patched or that further exploits will be successful.

Once the reputation climbs above a certain Trust Threshold value, most endpoints will again allow access to the site. Exposing the reputation score to the policy enforcement point allows various entities to set their own risk threshold to be more or less conservative depending on many factors, ultimately guided by their specific tolerance for risk.

While fine-grained reputation is important, it's also necessary that updates propagate through the system rapidly.

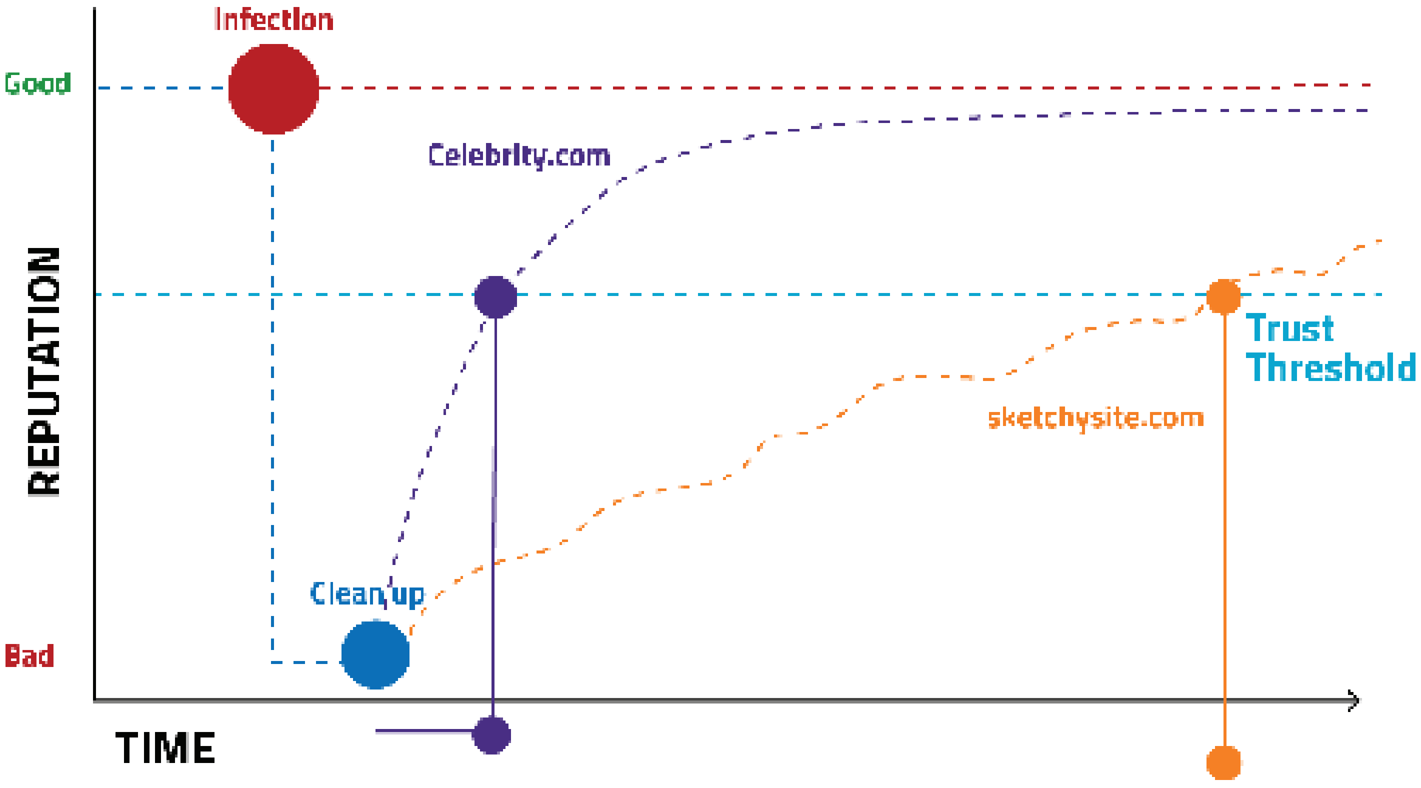

**Figure 1: Reputation over time. Note: this requires extensive data and computational power.**

## Speed of the Collection, Processing, and Delivery Pipeline

With the velocity and variability of today's threats, speed of processing and delivery are essential. Modern implementations of real threat intelligence can leverage modern infrastructure as a service (IaaS) systems, such as Amazon® Web Services (AWS), Microsoft® Azure, and Google® Cloud. Doing so can shorten the interval between discovery, analysis, and universal protection for threat intelligence subscribers to just a few minutes. These IaaS offerings provide huge scale, greater than 5 9's availability and uptime, geographically distributed services, high speed, low latency, redundancy, scalability, load-balancing, reporting, management, financial predictability, and cost control.

Real threat intelligence can leverage IaaS and supply policy enforcement points with relevant reputation information. Some implementations can even collect information in real time at an endpoint as it renders a browser page, simultaneously producing a machine learning feature vector from the data, and send it to a cloud-based service for a live determination on whether the page is safe or not. This can all take place in the sub-second time it takes for a fake page to render on a browser and when a user is asked to enter a username/password combination, i.e. less than half a second. Since most websites take several seconds to load, this small imposition is not noticeable, and makes a world of difference for the user's overall protection.

## Other Required Components of a Real Threat Intelligence System

In addition to speed and scale, there are many other required components of a threat intelligence system and implementation, which are too detailed for this discussion, but essential to its overall success. These include, but are not limited to:

» Accuracy of the machine learning classification system to minimize false positives and false negatives.

» Feedback from customers to the threat intelligence system on false positives and false negatives and incorporation of that information into retraining.

» Implementation of whitelists and blacklists to allow local administrators to override classifications and policy.

» Efficient integration of real threat intelligence in the policy enforcement endpoint or network to minimize latency and impact on end users.

## Real Threat Intelligence is an Enabler

If you're a small business, you have probably outsourced IT to an MSP. For most SMBs and MSPs, employing even one threat researcher is neither in budget nor a real possibility. Larger businesses may employ one or more IT professionals, but maintaining a network operations center (NOC) or security operations center (SOC) is beyond their means.

The solution to this problem is to bring enterprise-grade security defense capability into your business. The answer is to use products that incorporate real threat intelligence that has leveraged automation in combination with machine learning trained by world-class threat researchers.

## Automation is Key

The key to producing real threat intelligence is to rely primarily on automation, but leverage human assets wherever possible and in the proper way. In the current competitive IT environment, no one can hire enough technical talent, and that's true across a broad IT spectrum including machine learning, statisticians, business intelligence, software engineers, threat researchers, and customer support.

Instead of hiring armies of threat researchers to triage every new threat we see, we can hire relatively few threat researchers, a few statisticians and a few machine learning experts and leverage their expertise and knowledge into automation that can do the repetitive and non-creative part of the job.

## The Role of Humans

Machine learning does not replace human analysts' and researchers' roles in threat intelligence. We in no way suggest that machines by themselves can take the place of human ability to inquire, create, and discover.

To be successful, machine learning models must be guided by human intelligence and training. Machines can take on the repetitive, laborious, and monotonous aspects of a threat researcher's job, such as classifying the vast majority of new internet objects that appear every day. Still, a tiny fraction of those objects will be so new and different from previous examples as to be unclassifiable by the automated machines. In these cases, a threat researcher will have to use their human capacity to inquire, create, and draw nuanced conclusions from data to parse the nature, purpose, and inner workings of the new threat. Once that work is done, the machine learning model can be retrained with this new information, and the work of the threat researcher is then leveraged a thousand or million-fold as that updated model can now classify an entirely new genre of previously unseen or zero-day threats.

Machine learning can also incorporate the expertise of many talented threat researchers into a single model and classifier. In the same way that you want to see the best doctor when you get fall ill, you want to have the best threat researcher on staff when you experience a cyberattack. Machine learning lets you leverage an entire team of world-class threat experts, even at 2 a.m. on a Saturday. The machines never get sick, tired, or take a day off.

## Conclusion

Although there is confusion around the data, products, and the practices called threat intelligence, and there are a variety of pros and cons with the various forms it takes, there are a number of concrete qualities to look for when selecting a threat intelligence solution.

**Real threat intelligence must:**

1. Leverage a variety of thoroughly vetted, validated sources
2. Be based on realistic user actions and behaviors, not just simulations
3. Consider numerous environmental factors and variations
4. Give reputation scores, rather than binary good/bad determinations
5. Leverage human subject matter experts to guide machine learning for automated classification
6. Use cloud-based IaaS for scale, speed, and reliability in data collection, processing, and delivery
7. Be timely, complete, accurate, and adaptive
8. Provide context for deeper research and insight into determinations

Real threat intelligence means that the widest spectrum of threats can be blocked automatically, with the highest accuracy, and without a human in the loop. Real threat intelligence can augment humans in the field, but is generally meant to be completely useful without human intervention to enable protection from threats.

Real threat intelligence is synthesized from high fidelity data sources which are processed primarily by automation and a discipline of artificial intelligence known as machine learning, the training of which is guided by experts in threat research. We have shown that rapid delivery of newly discovered threats to policy enforcement points is key to its usefulness, as the shelf life is short. Real threat intelligence stops known as well as previously unknown zero-day threats the moment they are first encountered on live systems, and brings enterprise-grade security to businesses and individual users who do not have the extensive budgets and means of an enterprise-size company.

**About Webroot**

Webroot was the first to harness the cloud and artificial intelligence to protect businesses and individuals against cyber threats. We provide the number one security solution for managed service providers and small businesses, who rely on Webroot for endpoint protection, network protection, and security awareness training. Webroot BrightCloud® Threat Intelligence Services are used by market leading companies like Cisco, F5 Networks, Citrix, Aruba, Palo Alto Networks, A10 Networks, and more. Leveraging the power of machine learning to protect millions of businesses and individuals, Webroot secures the connected world. Headquartered in Colorado, Webroot operates globally across North America, Europe, and Asia. Discover Smarter Cybersecurity® solutions at webroot.com.

**World Headquarters**
385 Interlocken Crescent
Suite 800
Broomfield, Colorado 80021 USA
+1 800 772 9383

**Webroot EMEA**
6th floor, Block A
1 George's Quay Plaza
George's Quay, Dublin 2, Ireland
+44 (0) 870 1417 070

**Webroot APAC**
Suite 1402, Level 14, Tower A
821 Pacific Highway
Chatswood, NSW 2067, Australia
+61 (0) 2 8071 1900