Chapter 3

Machine Learning in Application Security

Nilaykumar Kiran Sangani and Haroot Zarger

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.68796

Abstract

Security threat landscape has transformed drastically over a period of time. Right from viruses, trojans and Denial of Service (DoS) to the newborn malicious family of ransomware, phishing, distributed DoS, and so on, there is no stoppage. The phenomenal transformation has led the attackers to have a new strategy born in their attack vector methodology making it more targeted—a direct aim towards the weakest link in the security chain aka humans. When we talk about humans, the first thing that comes to an attacker's mind is applications. Traditional signature-based techniques are inadequate for rising attacks and threats that are evolving in the application layer. They serve as good defences for protecting the organisations from perimeter and endpoint-driven attacks, but what needs to be focused and analysed is right at the application layer where such defences fail. Protecting web applications has its unique challenges in identifying malicious user behavioural patterns being converted into a compromise. Thus, there is a need to look at a dynamic and signature-independent model of identifying such malicious usage patterns within applications. In this chapter, the authors have explained on the technical aspects of integrating machine learning within applications in detecting malicious user behavioural pattern.

Keywords: machine learning, cybersecurity, signature-driven solutions, application security, pattern-driven analytical solutions

1. Introduction

Cybersecurity, a niche domain is likely to be compared in parallel to a cat and mouse game where sometime the offensive team (attacker/hacker) has an advantage and sometime the defensive team (cyber sec personnel). This never settling game has changed drastically over a period of time having born to various attack vectors targeting humans or what is largely known as the weakest link in the security chain.

Over the years, Information Technology (IT) has witnessed a massive paradigm shift. Initially, it was about mainframes, client-server model, closed group of systems, and the attacks were very limited and focused towards these only. Down the line of time, the former has been transformed completely into the web-based layer, clouds, virtualisation, and so on, thus adding greater complexity in the whole development-deployment architecture—of applications and infrastructure—thus making the attack surface further difficult for the hackers. What has remained constant is the human factor and the same is being exploited in large to circumvent the protection mechanisms which are in place.

Traditional signature-based solutions are functioning great in preventing against known attacks, but the paradigm shift of the technologies is making the signature-based systems inadequate against the newborn attacks and malicious exploits. Thus, the need of the hour is to implement which is a dynamic and signature—less thus evolved machine learning (ML).

Machine learning is not a new domain or technology. It has been in use in other areas since the 1950s. The missing link is the intersection of cybersecurity and machine learning. One of the best examples of early use of machine learning in security is the case of spam detection.

In this chapter, we cover how cybersecurity has evolved over a period of time and how attacks have become more tactical and sophisticated. We also talk about what is machine learning and its associated components. In this part, we cover how combination of machine learning and security adds value to an organisation. Later on, we focus on the application layer and web applications in specifics. And, finally, we talk about focusing on merging machine learning and applications to provide a pattern-based analytics of security within applications.

The second section covers in detail how cybersecurity has evolved over a period of time and how attacks have become more tactical and sophisticated. Section 3 focuses on the application layer and web applications in specifics. It will also cover how web applications have grown over time and the threats associated with them. Section 4 talks about what is machine learning and its associated components. This section, in addition, will also cover how combination of machine learning and security adds value to an organisation.

Section 5 targets on the merger of machine learning and application to provide a pattern-based analytics layer of security within applications.

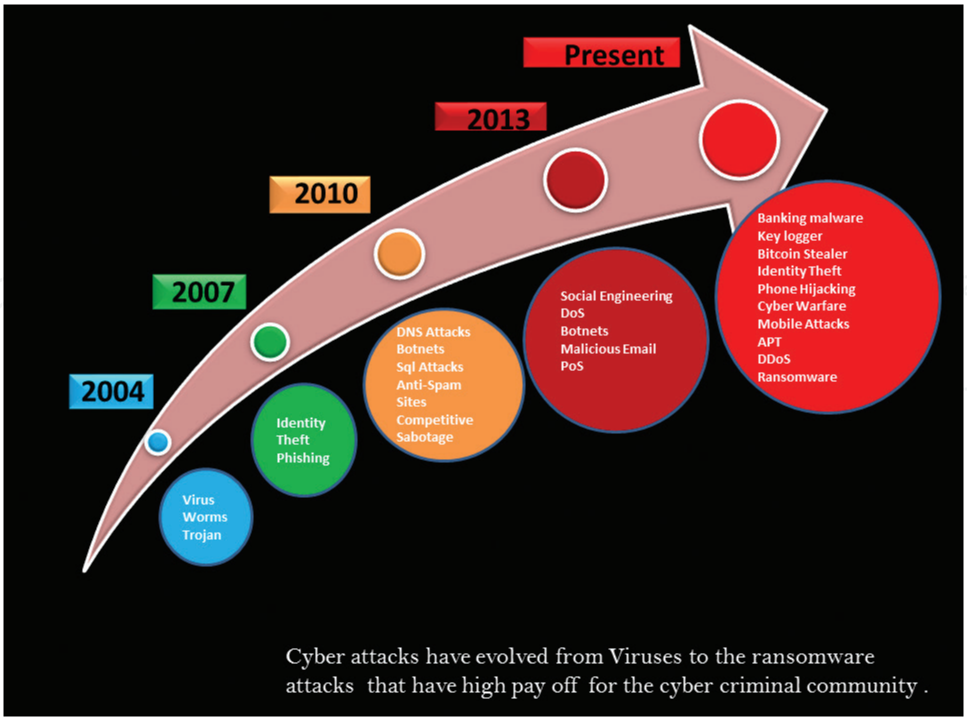
2. Evolution of cybersecurity

By definition, Cybersecurity can be defined as 'the body of technologies, processes and practices designed to protect networks, computers, programs and data from attack, damage or unauthorised access'. One of the most challenging elements of cybersecurity is the quickly and constantly evolving nature of security risks. Adam Vincent pronounces the problem [1]:

'The threat is advancing quicker than we can keep up with it. The threat changes faster than our idea of the risk. It's no longer possible to write a large white paper about the risk to a particular system. You would be rewriting the white paper constantly.'

Initially cybersecurity used to be relatively simple. The enterprise network comprised of mainframes, client-server model, closed group of systems and the attacks were very limited with viruses, worms and Trojan horses being the major cyber threats. The focus was more towards malwares such as virus, worms and trojans with purpose of causing damage to the systems. It started with virus which needed to be executed in order to cause a malfunction or damage to the system. As this was something where manual intervention was required for propagation, a new type of malware came into existence, that is, 'Worm' similar to virus but with self-replicating feature, that is, they do not require a human intervention or a program to execute. These cyber threats randomly targeted computers directly connected to the Internet but posed little threat. Within the enterprise networks with firewalls on the perimeter and antivirus protection on the inside, the enterprise appeared to be protected and relatively safe. Occasionally an incident would occur and security teams would fight it.

The initial attacking methodology was attacking the infrastructure. This involved the traditional approach of compromising the systems by getting inside the network though loopholes such as open ports, unknown services, and exploiting system-related vulnerabilities in the infrastructure. At this time, the offensive teams started to recognise and closed these gaps as much as possible, reducing the attack surface. Over a period of time, as the infrastructure changed, the former has been transformed completely into the web applications, web-based layer, clouds, virtualisation, mobility, and so on, thus adding greater complexity in the whole development, deployment architecture of applications and infrastructure and changing the attack surface further as shown in **Figure 1**. Attackers started getting inside the enterprise networks, and once they were inside they operated in stealthy mode. By attaining access, they controlled the



Cyber attacks have evolved from Viruses to the ransomware attacks that have high pay off for the cyber criminal community .

Figure 1. Cybersecurity attacks evolution over time.

infected machines and managed them through command and control systems (C&C servers). Vulnerable systems were exploited within the enterprise for lateral movement among computers on the network, capturing user credentials and other critical information of more and more users within the organisation. The final nail in the coffin was privilege escalation, art to gaining elevated access to the machine, get control of the systems administrator accounts in charge of everything. Once these attackers got administrative control of the enterprise, they were able to do anything they wanted. It was like 'Keys to the Kingdom'.

Similarly, to the cat and mouse game or as we have seen in Tom and Jerry, to overcome each other as they used to change the tactics, the very same applies when we talk about attackers versus defenders in cyber space. Attackers take the advantage of zero-day exploits, vulnerabilities, and so on to compromise the systems, whereas defenders use secure mechanism such as hardening, patching, segmentation and other security controls to reduce the surface attack. This way the enterprise is locked down up to a certain extent, thus reducing the attack surface. For the attackers with less threat surface to attack due to the lock down, the only possibility seen by them towards a breach lies in web application.

2.1. Why web applications are vulnerable?

Before we begin, let us have a basic understanding of web applications. A web application or web app is a software application in which the client (or user interface) runs in a browser. Common web applications include webmail, online retail sales, online auctions, wikis, instant messaging services and many other functions [2]. For organisations, whether they are a private entity or government, to conduct business online, it has to provide services to the outside world. Over a period or so, the web has been embraced by millions of businesses as an inexpensive medium to communicate and exchange information with customers [3]. Therefore, they are vital to businesses for expanding their online presence, thus fashioning long-term and beneficial relationships with customers. There is no doubt in saying that web applications have become such a universal phenomenon over a period of time. Web applications are convoluted and multifarious in nature, and due to this behavior, they are widely mysterious and completely misinterpreted [3].

Regardless of the advantages, web applications do raise a number of security concerns. Severe weaknesses or vulnerabilities allow hackers to gain direct and public access to databases in order to extract sensitive data. Many of these databases contain critical information (personal, official, financial details, etc.) making them a frequent target of hackers. Although defacing corporate websites are still commonplace, nowadays, hackers prefer gaining access to the sensitive data residing on the database server because of the immense pay-offs in selling the data.

The greater complexity, including the web application code and underlying business logic, and their potential as a vector to sensitive data in storage, or in process, makes web application servers an obvious target for attackers [12].

In **Figure 2**, it is easy to see how a hacker can quickly access the data on the database through creativity and negligence or human error, leading to vulnerabilities in the web applications.

As mentioned, websites use databases to store and fetch the required information to the users. If a web application is vulnerable, that is, it can be exploited by the attackers, then the database

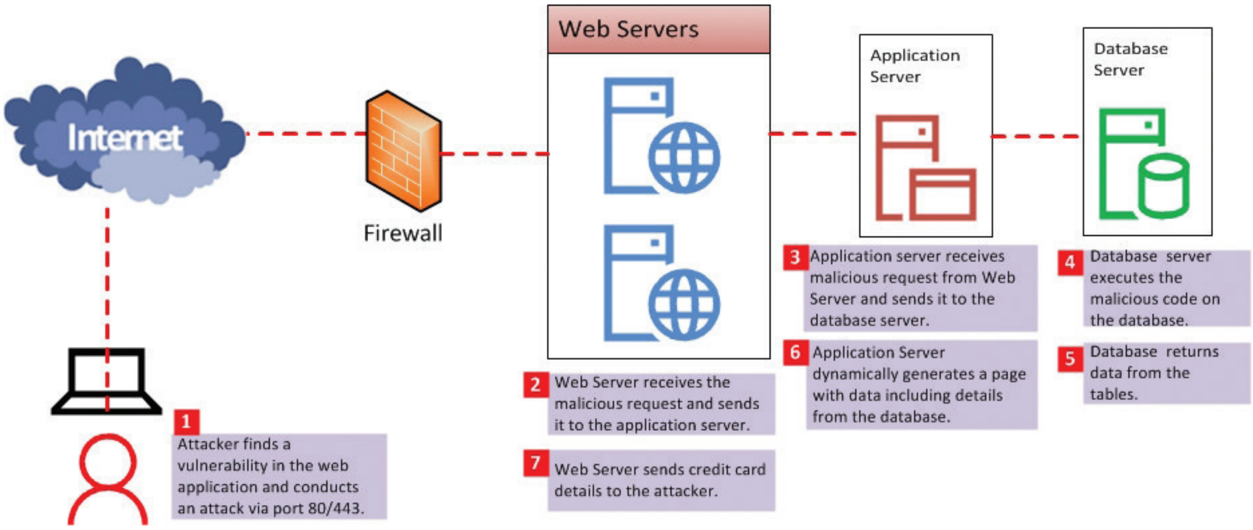

Internet
Firewall
Web Servers
Application Server
Database Server
3 Application server receives malicious request from Web Server and sends it to the database server.
4 Database server executes the malicious code on the database.
6 Application Server dynamically generates a page with data including details from the database.
5 Database returns data from the tables.
2 Web Server receives the malicious request and sends it to the application server.
7 Web Server sends credit card details to the attacker.
1 Attacker finds a vulnerability in the web application and conducts an attack via port 80/443.

Figure 2. How an attacker exploits a web application.

associated with the web application is at serious risk, as it contains all the critical information that the attackers are looking for. Recent research shows that 75% of cyberattacks are done at web application level [3].

Web application vulnerabilities have drastically increased over the past few years, as companies demand faster web application releases to fulfil the end-user requirements. Vulnerabilities associated with web applications are risky for organisations as they endanger not only brand and reputational damage but also loss of data, legal action and financial penalties associated with these incidents. The outcomes continue to confirm the majority understanding that the web application vector is a foremost and less protected path for attackers [11].

Web application scene is altering continuously over time. Evolution of web layer has enhanced rich experiences and functionalities directly within/from the browser. As a result of this flexibility and scalability that web applications provide, web applications and web services are rapidly replacing the legacy applications, and, as a result, broadening the surface attack which increases attacker's chances of exploitation, primarily since traditional network layer security controls such as firewalls and signature-based intrusion prevention and detection systems (IPS/IDS) have little or no role to play in detecting and preventing an attack occurring via the web application.

2.2. Cybersecurity attacks

In the past few years, the trend has played out in more and more breaches hitting the headlines. Some of the cyberattacks that shock the IT world include the following:

- RSA SecurID breach: Year 2011 [4]

  In 2011, RSA's enterprise was breached and the security keys for many of its customers were believed to have been stolen. This breach prompted RSA to replace millions of its SecurID tokens to restore security for its customers.

- Columbian Independence Day Attack: Year 2013

  In 2013, a large-scale cyberattack held on 20 July—Columbian Independence Day—against 30 Colombian government websites. As the most successful single-day cyberattack against a government, most websites were either defaced or shut down completely for the entire day of the attack. Attacks included both web and network vectors including web application and network Distributed Denial of Service (DDoS) attacks. [5]

- eBay Data Breach: Year 2014 [6]

  eBay went down in a blaze of embarrassment as it suffered this year's biggest hack so far. In May 2014, eBay revealed that hackers had managed to steal personal records of 233 million users.

- Sony Picture Entertainment: Year 2014 [7]

  On 25 November 2014, something new happened in the history of data theft activity. A group calling itself GOP or The Guardians of Peace hacked into Sony Pictures, causing severe damage to the network for days and leaked confidential data. The data included personal information about employees and their families, e-mails and copies of then-unreleased Sony films and other information.

- Dyn Cyber Attack: Year 2016 [9]

  The largest cyberattack in recorded history happened on 21 October 2016, causing temporary shutdown of websites such as Twitter, Netflix, Airbnb, Reddit and SoundCloud. The threefold hack caused mass Internet outage for large parts of the USA and Europe.

These incidents are a few of the numerous cybersecurity breaches and attacks that have occurred over the past few years [8]. The trend indicates that the attacks are more towards personal identities, financial accounts and healthcare information and getting such information on millions or tens of millions of people. Looking at the trend here, these types of cyberattacks are moving down market over time. In simple terms, the techniques that nation states were using few years back are being used by cyber criminals currently [10]. In the real-world scenario, we have to expect that these types of less known attacks will become more public in the near future as exploits and techniques will surge and become available to larger communities. These types of threats may be affecting a small group of organisations at a given time, but progressively they will become more common. Organisations have to be regularly evolving their defences [10].

2.3. Web application threat trend

As per Verizon's [12] recently released Data Breach Investigation Report (DBIR) for 2016 which is constructed on real-world investigations and security incidents:

1. When we compared this year's data to last year's data, the total number of attacks this year was significantly higher than last year (see below).
2. Conventional web attacks rose by 200 and 150%, respectively, continuing the trend from last year, with larger numbers and larger volumes of scanning campaigns across the Internet.

3. The volume and persistency of attacks indicate industrialisation of and automation behind organised efforts.
4. Ninety-five per cent of confirmed web app breaches were financially motivated [12].
5. Web application attacks are the #1 source of data breaches [12].
6. Data breaches caused by web application attacks are rapidly rising. The percentage of data breaches that leveraged web application attacks has increased rapidly in the last. This indicates that the web applications in many organisations are not just exposed but are also extremely susceptible compared to other points of attack [12].

Figure 3 illustrates the occurrence rates of different attack methods that resulted in data loss. The grey bars indicate the corresponding figure for the past year, that is, 2015. It clearly shows that web application attacks accounted for the highest proportion of attacks that resulted in breaches.

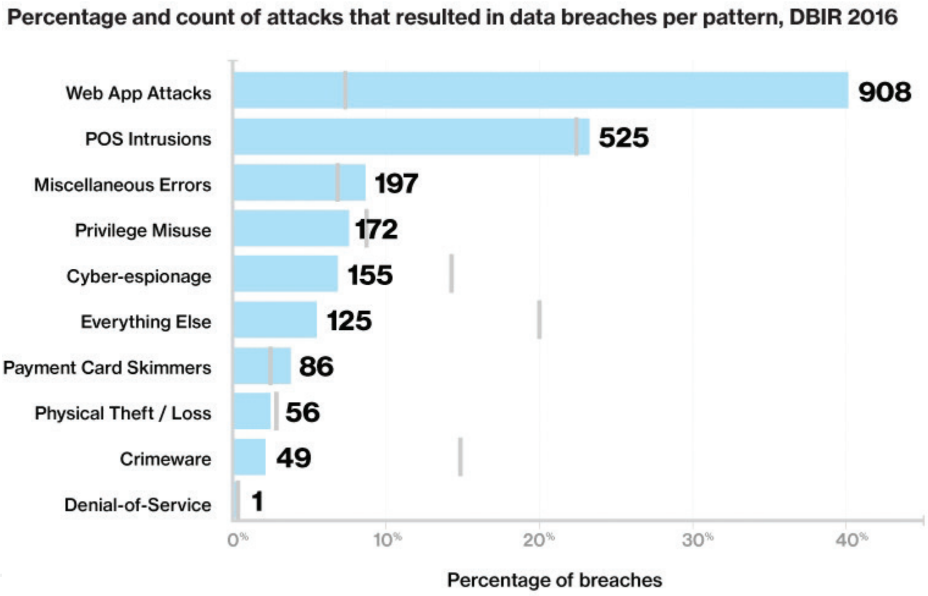


Figure 3. DBIR statistics report.

3. Web application threats

3.1. Web application security: a new boundary break

In the recent era, each and every business has web applications to showcase its online presence, conduct business online, and so on. These applications are hosted on multiple online servers, databases, infrastructures, and so on. And, thus, inherits security risks from the underlying technologies and its associated components. An interesting fact, in 2012 alone, there has been reported around more than 800 hacking events and around 70% of them where via issues in web applications, thus, making the web the new boundary for security, which is not as easy to pull a kill switch like the network [28].

These days we see application development is more focused towards the web, creating applications for every business and personal needs. Looking at an increase in web applications, hackers are more focused to alter their threat attack model targeting these applications instead of a complete protected infrastructure, networks, and so on [29]. Web applications are susceptible to attack from the time they go online. As more inventive attack strategies and structures seem on the Internet, end users and the organisations that provide web services need to shield their systems from being compromised. According to Gartner, around 75% of all external attacks occur at the application level [18]. Web 2.0 helps enterprises in conducting their business; however, an understanding needs to be adhered to that it also introduces a surfeit of damaging risks [18]. The beauty about web application is that in the past the applications were created with scripts as there were no frameworks to support a web developer. But these days, rise in various web development languages, such as Java, NET, WordPress, PHP, Ajax and JQuery, allows a developer to create web application with delivering wide range of functionality in less than no time. Thus come the security issues with the underlying frameworks.

3.2. Security risks in a web application

Application security risks are universal and can pose an unswerving threat to business availability. Business world works using web-based applications and web-based software. Because of the propagation of web-based apps, vulnerabilities within the web applications are the new attack path for malicious actors/hackers. An attack of a web-based application possibly will produce information that should not be available, browser spying, identify theft, theft of service or content, damage to corporate image or the application itself and the feared Denial of Service (DoS). The nature of http, hackers find it very easy and lucrative to modify the parameters and execute functionality that was not envisioned to be performed as a function of the application [30, 32, 33].

Businesses and organisations are anticipating large amount of capital in expenditure to safeguard and secure their complete networks (internal/online) and servers. And yet, when it comes to web application security, there is a huge ignorance towards its protection, or, at the very least, considered as an undervalued aspect within the threat model architecture. This notion of thought is considered to be ill-fated, as it has been seen that most security attacks occur online via applications. As per Gartner Group, '75% of cyber-attacks and internet security violations are generated through internet applications'. It is just that organisations are unable to comprehend the security loopholes which exists in web applications [31].

Open Web Application Security Project (OWASP) Top 10 increases cognizance of the challenges organisations face safeguarding web application security in a swiftly fluctuating application security environment. Let us focus on the OWASP Top 10 from 2013 as described below [34].

Injection: Injecting aka inserting code to trick an application in triggering unplanned activities which serve as a deviation from the business functional logic. One of the most preferred injection hacks, which is being actioned by the hackers, is a SQL Injection (SQLi) attack. In SQLi type of attack, malicious actor (hacker) injects a SQL declaration within the application to perform malicious actions like deleting the database, retrieving sensitive database records, and so on.

Broken Authentication and Session Management: Hackers can take over user identities, unauthenticated pages and hide behind a genuine user ID to gain easy admittance to your data and programs.

Cross-Site Scripting (XSS): XSS inserts malicious scripts within the web applications. This malicious script lies within the client side code (browser) and targeted for a different user of the application.

Insecure Direct Object References: Most websites store user records based on a key value that the user controls. When a user inputs their key, the system regains the equivalent information and presents it to the user. An Insecure Direct Object Reference occurs when an authorisation system fails to avert one user from gaining access to another user's information.

Security Misconfiguration: Security misconfiguration is a reference to application security systems that are half-finished or ailing managed. Security misconfiguration can occur at any level and in any part of an application and, thus, is both highly common and effortlessly noticeable.

Sensitive Data Exposure: Inadvertent data leak is a grave problem to everyone using a web application that contains user data.

Missing Functional Level Access: Wrongly configured user access control system can allow users the capacity to achieve functions above their level.

Cross-Site Request Forgery (CSRF): The attack functions on a web application in which the end users’ client (browser) has performed an undesired action (user has no knowledge until the task has been performed) in which the very same user is authenticated.

Using Components with Known Vulnerabilities: Open source development practices drive innovation and reduce development costs. However, the 2016 Future of Open Source Survey found that momentous encounters remain in security and management practices. It is critical that organisations gain perceptibility into and control of open source software in their web applications.

Unvalidated Redirects and Forwards: When a web application accepts unverified input that affects URL redirects, malicious actors/hackers can redirect users to malicious websites. In addition, hackers can alter automatic forwarding routines to advance access to sensitive information.

3.3. Associated motive in a web application hack

Users’ accessing web application(s) are indirectly accessing the critical resources such as the web server and database server (if applicable). Software developers intend to spend vast amount of their project allocated time in developing the functionality and ensuring a timely release thus binding less or no time to security requirements. The reason for this can be due to lack of understanding/implementing security measures/controls in a web application [19]. For whatever reason, applications are often peppered with vulnerabilities that are used by attackers to advance access to either the web server or the database server. Some of the aspects what an attacker seeks for [19]—defacement, redirect the user to a malicious website, inject malicious code, steal user's information, steal bank account details, access unauthorised and restricted content, and so on.

4. Machine learning (ML)

4.1. What is ML?

An ardent subset of artificial intelligence dedicated to the formal study of learning systems. Machine learning is a methodology of performing data analysis which automates an analytical model [13]. In other words, machine learning is all about learning to do a task better in the future based upon its previous learned patterns in the past [14]. ML being a subsection of artificial intelligence provides systems/computers with the power to learn without being explicitly programmed [14]. One of the reasons why ML is picking up traction in the IT world is because as and when patterns are developed with new data, ML algorithms has the ability to independently adapt and learn from the data and information. With ML, computers are not being programmed but are altering and refining algorithms by themselves [13, 14].

Looking at other definitions, ML discovers the study and construction of algorithms that can learn and make predictions of data. In other words, it focuses on prediction-making through the use of computers [13–15].

With the rise in new-generation-technologies being witnessed in the twenty-first century, ML today is something that cannot be compared to what it used to be in the historical past. The past has been witnessed in the rise of various ML algorithms and the complexity of the calculations being carried out; however, it is just during the recent times, the recent ML algorithms have been tuned in such fashion that the whole complex mathematical calculations, analysing big data at a much greater and faster—a very recent development [16, 17]. Some of the underneath examples (but not limited to) have adapted ML within their service space:

- Google's Self-Driving Cars: ML algorithms are used to create models in classifying various types of objects in different situations [18].
- Netflix: ML is used in improving the member experience [19].
- Twitter: ML is being applied to enhance its video strength [20].

4.2. Rise in ML

An immense amount of popularity is being gained over Data Mining and Bayesian analysis due to a fast-pace adaptation of ML in solving business problems. Computational power, availability and various type of data, cheap and powerful data storage is ever growing which is some of the few attractive factors towards adapting ML. What this mean is that it is quickly possible to fire up an automated predictive model which can analyse larger and complex datasets and deliver accurate final outcomes [25]. This results in an additional value towards predictions which leads in creating smarter real-time decisions without human intervention [25]. Within the software vertical, artificial intelligence is being a popular technology to integrate within a service as the mandate for analytics is motivated more by growth in type of both structured and unstructured data [21].

In the 1930s and 1940s, the pioneers of computing, such as Alan Turing, began framing and playing with the most basic aspects of ML such as a neural network which has made today's ML probable [27].

As per [25], humans create couple of models every week; with ML, thousands of models are created within a week. Upsurge in computing power is one of the prime reason from a transformational shift from theoretical to practical implementation. High number of researchers and industry expertise are contributing towards the advancements in this space as it is constantly being used in solving some real issues across industries including (but not limited to) healthcare, automotive, financial service, cloud, oil and gas, governments, and so on. Data (be it small or large) residing within these types of industries contain a large number of patterns and insights. ML creates the ability to discover various patterns and trends within these giving rises to substantial results.

The rise of cloud computing, massive data storage, devices connecting with each other (Internet of Things [IoT]), and mobile devices play a huge role in the adaption of ML.

4.3. ML methodologies

Supervised Learning: Algorithms are trained on labelled data, essentially leading to its meanings where an input having a looked-for output. In other words, a supervised learning algorithm with an input variable denoted as P and an output variable denoted as Q and algorithms are used to create and learn a mapping function (f) via the input to the output.

$$Q = f(P)$$

The goal of a supervised learning algorithm is to achieve an estimate mapping function so that for every new input (P), a new predicted output (Q) is created. In other words, the learning algorithm receives a set of inputs with their corresponding outputs, and the algorithm learns by equating its concrete output with correct outputs in order to find errors and have the learning model modified accordingly. Supervised learning algorithms make use of patterns to predict the values of the label on unlabelled data. This is achieved by classification, regression, prediction, and so on [25, 22].

Supervised learning is used to predict probable future events within applications having vast amount of historical data [25]. An example is detecting likely fraud patterns in credit card transactions.

Unsupervised Learning: Unsupervised learning is where only an input data (P) is available with no equivalent output variables. The aim of unsupervised learning is to model the construction of the data in order to learn more about the data. Algorithms are required to discover a structure, an inference and meaning within the data in order to arrive to a conclusion. These algorithms do not have any type of historical data in order to predict the output unlike supervised algorithms [25]. Unsupervised learning does not have any explicit outputs and nor exists a dependency environment factor within the input variables; it brings to accept preceding predispositions as to what aspects of the structure of the input should be seized in the output [23]. In an aspect, unsupervised learning locates patterns in the data which succours in arriving to a constructive meaningful decision.

4.4. Adaption of ML in industry

ML has been widely adopted across various sectors within the industry to solve real-life business statements. Data is available within the whole global space and to derive a deep understanding from it, ML is the methodology to be consumed for such derivations. We live in the golden era of innovative technologies and ML is one of them [24]. ML has created an ability to solve problem declarations horizontally and vertically across aviation, oil and gas, finance, sales, legal, customer service, contracts, security, and so on, due to its greatest capability of learning and improving. ML algorithms has been a great stimulus in creating applications and frameworks to analyse data which brings in a great predictive accuracy and value to enterprise's data, leading to a sundry company-wide strategy ensuing faster and stimulating more profit [25].

One such example is of the revenue teams across the industry, they are converting the practical aspect of ML in augmenting promotions, compensations and rebates driving the looked-for behaviour across various selling streams. **Figure 4** draws a mind of ML applications within some of the industries [25].

ML has been a chosen integration within the industries for its skill to constantly learn and improve. As we have seen, ML algorithms are very iterative in nature, having the flexibility to make it learn towards a vision of achieving an optimised and a useful outcome [25]. ML's data-driven acumen is infusing every corner of every industry and it's starting

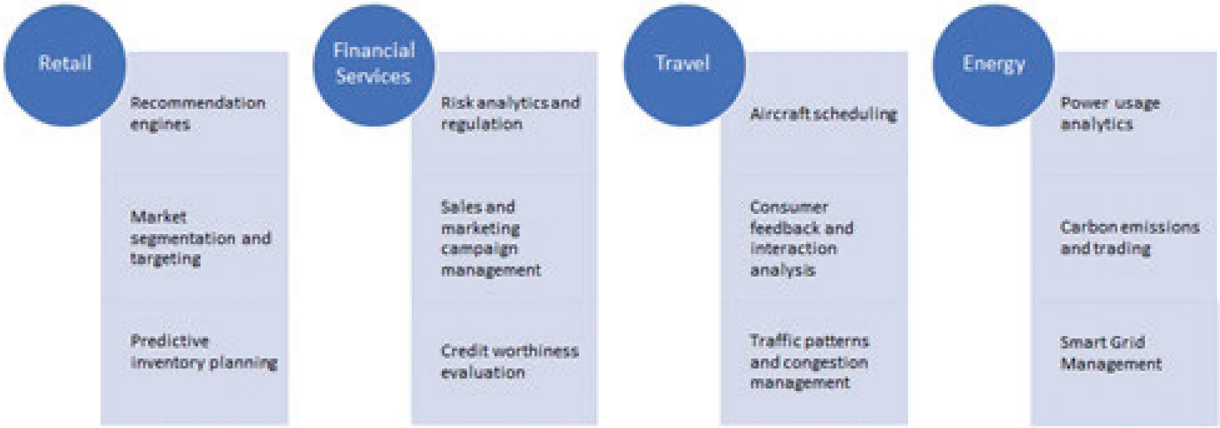

Retail
Recommendation engines
Market segmentation and targeting
Predictive inventory planning
Financial Services
Risk analytics and regulation
Sales and marketing campaign management
Credit worthiness evaluation
Travel
Aircraft scheduling
Consumer feedback and interaction analysis
Traffic patterns and congestion management
Energy
Power usage analytics
Carbon emissions and trading
Smart Grid Management

Figure 4. ML applications in industries.

to disrupt the way business is done worldwide. Leveraging ML has enabled processes to be re-calibrated inevitably and improved for reduced cycle times, created a higher quality of delivered goods and allowed for new products to be established and tested. The ability to influence data for more accurate decision-making in place of instinctive feel [26]. According to a representative from Gartner, as quoted, 'Ten years ago, we struggled to find 10 machine learning – based business applications. Now we struggle to find 10 that don't use it' [26].

ML's rise in the industry makes data a growing vital part of how a business makes decisions. Because of this, data scientists will take up a complete central focused role in organisational strategies as data is becoming a core agent of change within a business. It is forecasted that with a wealth of data in business given the occurrence of sensors and IoT implementation, the wide ability to influence data will be critical to building competitive advantage [26].

ML should not be understood as a technology component, and, by the rise within the current era, it confidently is not a short-term trend. With its impact within the industries and various business sectors by bringing out toppling business models and with its rising maturity in terms of sophisticated algorithms being advanced, it will continue to be the solitary driver in shifting the complete viewpoint in decision making and having a truly workable conclusion [26].

4.5. Machine learning usage in cybersecurity

Machine learning (ML) is not something new that security domain has to adapt or utilise. It has been used and is being used in various areas of cybersecurity. Different machine learning methods have been successfully deployed to address wide-ranging problems in computer security. Following sections highlight some applications of machine learning in cybersecurity such as spam detection, network intrusion detection systems and malware detection [39].

4.5.1. Spam detection

Traditional approach of detecting spam is usage of rules also known as knowledge engineering [39]. In this method, mails are categorised as spam- or genuine-based set of rules that are created manually either by the user. For example, a set of rules can be:

- If the subject line of an email contains words 'lottery', its spam.
- Any email from a certain address or from a pattern of addresses is spam.

However, this approach is not completely effective, as a manual rule doesn't scale because of active spammers evading any manual rules.

Using machine learning approach, there is no need specifying rules explicitly; instead, a decent amount of data pre-classified as spam and not spam is being used. Once a machine learning model with good generalisation capabilities is learned, it can handle previously unseen spam emails and take decisions accordingly [40].

4.5.2. Network intrusion detection

Network intrusion detection (NID) systems are used to identify malicious network activity leading to confidentiality, integrity or availability violation of the systems in a network. Many intrusion detection systems are specifically based on machine learning techniques due to their adaptability to new and unknown attacks [39].

4.5.3. Malware detection

Over the last few years, traditional anti-malware companies have stiff competition from new generation of endpoint security vendors that major on machine learning as a method of threat detection [41]. Using machine learning, machines are taught how to detect threats, and, with this knowledge, the machine can detect new threats that have never been seen before. This is a huge advantage over signature-based detection which relies on recognising malware that it has already seen.

4.5.4. Machine learning and security information and event management (SIEM) solution

Security information and event management (SIEM) solutions have started leveraging machine learning into its latest versions, to make it quicker and easier to maximise the value machine data can deliver to organisations [42]. Certain vendors are enabling companies to use predictive analytics to help improve IT, security and business operations.

5. Uniting machine learning and application security

In the last few sections, we have seen that the web application attacks are constantly evolving, and building protection mechanism on the fly has been a complex task. So, with all of the recent threats and attack trends on web application, one may ask what exactly is machine learning and how is it applied in these situations.

Inferring from a much wider scope and having it elucidated, machine learning imparts the understanding as a line of drills where the algorithm would 'train' a machine in cracking a problem. In order to understand the above statement, we need to comprehend it via an example, let us imagine a task to determine if the animal in the photo is a lion or an elephant. Prior to coming out with this conclusion, it is imperative to train the machine by providing 'n' photos of elephants and 'm' photos of lion. As the machine trains, a picture can be supplied and the output will be predicted if the supplied picture is a lion or an elephant.

The effectiveness of a machine learning model is determined in the accuracy of its predictions; in other words, a predictive analytical model needs to be derived. In order to explain this, let us now provide the model with around 10 pictures of elephants and the output imparts eight being elephants and two depicted as lions. In this case, we derive the model to be 80% precise. Looking at this being on the brink of accuracy, there is a way to improve the model. And the improvement will be by providing more data; in other words, deliver knowledges to improve its proficiencies meaning to provide a large number of photos to train the machine as increase in data volume rises large developments aiming at an acceptable accuracy of the model. An implausible frequency of growth of web applications over the years produces large sum of logs which leads to a methodology in improving the precision over a period of time.

Let's explain the above perspective in web application scenario. Any three-tier web application consists of web traffic logs, application logs (normally terms as business layer) and the database logs (normally termed as data access layer). When we look at the logs, let's say we look at one category, that is, login attempts on the application. The output of the login can be either a successful attempt or a failure attempt. Compared to our example of elephants and lions, to train the failure or successful attempts we provided it with 100 logs of successful attempts and 100 logs of failed attempts. Once the model or the machine is trained, we can provide a log and it can tell me if it is a failed attempt or a successful attempt.

Now for predictive analysis, if we provide the model with 10 web logs of successful login attempts, out of that it says that seven are successful attempt and three are failed attempts, we can say that the model is 70% accurate. One way to improve a machine learning system is to provide more data, essentially provide broader experiences to improve its capabilities and with the application logs this is not a challenge. Any application which is accessed my thousands of users can generate huge number of logs on daily basis, thus increasing the accuracy of the machine learning model or algorithm.

5.1. ML detecting application security breaches

Researchers are constantly working on implementing ML techniques in detecting various application security level hacks. But the authors have proposed an extraction algorithm, which is based upon various ML algorithms. The authors [36] adapted various ML algorithms, such as SVM, NB and J48, to develop the vulnerability prediction model. They have emphasised on vulnerability prediction prior releasing an application. In an environment where time and resource are very minimal, web application security personnel require an upper-level support in identifying vulnerable code. A complete practical methodology in bringing out predicted vulnerable code will surely assist them in prioritising the secure code vulnerabilities.

Inferring from this thought process, authors [37] have worked towards bringing out a substantial pattern that illustrates both input validation and sanitisation code which are expected to be the predicted vectors of web application vulnerabilities. They have applied both supervised and semi-supervised learning when building vulnerability predictors based on hybrid code attributes. Security researchers are utilizing ML towards web application vulnerability detection. SQL Injection being one of the most preferred attack vector of hackers, authors [38] have displayed their work by coming with a classifier for detection of SQL Injection attacks. The classifier implements Naïve Bayes ML algorithm in conjunction with application security principle of Role-Based Access Control implementation for detection of such attacks.

5.2. Anomaly detection and predictive analysis

Anomaly detection is the documentation of items, events or observations which do not conform to an expected pattern or other items within a dataset. Anomalies are also termed as outliers. These outliers will detect an issue which is not normal compared to its learned model. Industries are adapting anomaly detection techniques in identifying medical problems, financial frauds, and so on.

Anomaly detection is not limited just to security but it is being utilised in various other domains such as financial fraud uncovering, fault detection systems for structural defects, event detection in sensor networks used in petroleum industries and many other. It is used in preprocessing the data, to eliminate any abnormal data from the dataset. By eliminating the abnormal data in supervised learning results in a statistically significant increase in accurateness.

Looking at the vast amount of cyberattacks increasing on web applications, the authors of Azane were inspired by the complete study of anomalies and patterns which led them to present a research towards the implementation of an ML engine comprising of an anomaly detection and predictive analysis framework at an application level to detect certain user behaviour in order to predict if it is a normal usage or an attack. The authors have explained a prototype model that will describe the Azane which is a machine learning framework [35] for web applications. Azane as a proof-of-concept algorithm designed by the authors has played a major role at the applications log level to detect anomalies at the application workflow level and also serve as a prediction base for any future events. The workflow in **Figure 5** comprises multiple stages: Application Logs → Pre-processing → Training Data → ML Algorithm → Test Phase → Predictive Model Output.

Let us understand each phase in general:

1. Logs

This is the first and the foremost phase upon which the whole model depends. We need to understand that in order for the model to work, logs are necessary. The authors have taken the dual aspects of logging into consideration and have applied their algorithms in order to derive a meaningful context. This phase is more about collection of logs and verifying whether the log contains the parameters that are required for analytical purpose.

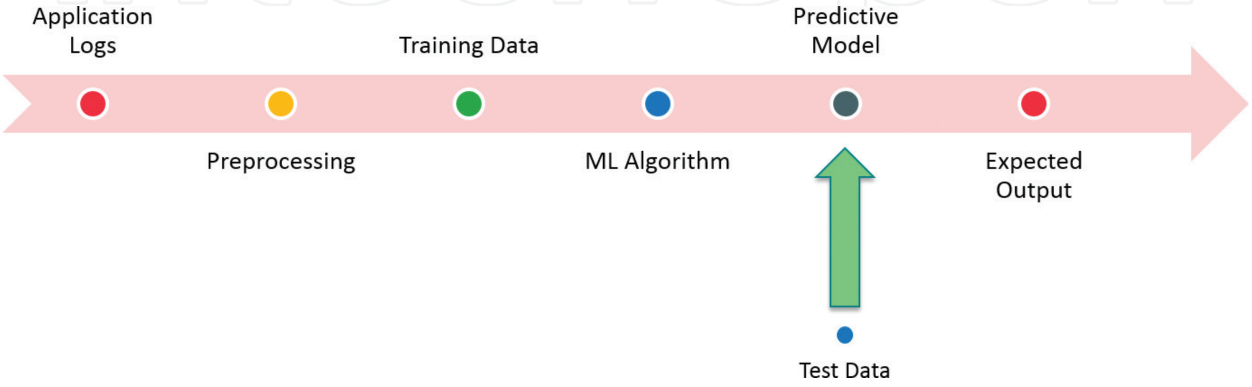


Figure 5. Anomaly detection workflow.

2. Pre-processing

This phase is more of transforming a given dataset into a format in which the ML algorithm can deduce and learn from it. This phase emphasises on making your data compatible with the machine learning algorithms. A challenge which can occur is that various algorithms make different assumptions about the data which may require to conduct individual analysis to see which algorithm is more suitable to the business needs. Further, when you follow all of the rules and prepare your data, sometimes algorithms can deliver better results without the pre-processing.

Pre-processing in general includes the following steps:

a. Load the data

b. Split the dataset into the input and output variables for machine learning.

c. Apply a pre-processing transform to the input variables.

d. Summarise the data to show the change.

3. Training data

Training data phase is more of a data that is an outcome of pre-processing and will be used to train the algorithm. This data plays a vital role to feed in the ML algorithm as it has the right amount of input and output data. Training data is more about making the machine learning algorithms aware about the data attributes and their values.

4. Machine learning algorithm

This phase is used to identify the algorithm which suits your dataset and final outcome.

5. Test phase (Learning: Predictive Model)

In simpler terms, training data is bought in to create a learning set which will service as a predictive model against the selected algorithm to validate the prediction or the accuracy of the model. A training set is learnt and this particular set of learnt data is used to discover potentially predictive relationships. This whole analysis is based on the training data which forms the baseline of the predictive analysis model.

6. Predictive model output

Now the final step is to test the predictive model with the accuracy with the new data known as the test data. A test set is a set of data used to assess the strength and utility of a predictive relationship.

Azane was developed to unite machine learning and application security in order to protect web applications from sophisticated type of attacks by predicting the application's user pattern.

ML algorithms yield a near-to-accurate result when huge amount of data is fed and trained in order to aid in spotting malicious patterns. Data should be consistent for the ML algorithm to work to its fullest. Combining ML output with other infrastructure devices, such as IPS and firewall, will strengthen the correlation and assist in drilling down to its correctness and validation of a web application attack. Finally, once the patterns are identified and analysed and blocked, this can be integrated to a SIEM solution for a complete centralised management metrics reporting.

In this chapter, we have seen how machine learning can be integrated within application security in order to prevent attacks on web-driven applications.

6. Conclusion

Our daily life, economic growth and a country's security is highly dependable on a safe and secure cyber universe. Hackers are always on a lookout in breaching the cyber universe in identifying vulnerable loopholes to steal information, data, money, having services disrupted, and so on. The inventiveness of hackers has led to the advance of new attack vectors and new ways of exploiting bugs in a web application. Web application breaches have evolved the cyber war between application owners and hackers. As companies cope with a more urbane threat landscape, they will have no choice but to innovate, automate and predict hacking identifications, attempts and breaches in their web applications.

Talking about prediction, machine learning as a technology has erupted vastly in the whole cyber implementation space. These decision-making algorithms are known to solve several problems as seen in the above illustrations. Following a simple principle of prediction, machine learning has shown itself as the problem solver for any given type of problem occurring within the complete technology space. Looking at the in-depth capability of machine learning, the cybersecurity industry started its adaptation. The collection and storage of large amount of data points is rapidly rising in cybersecurity where machine learning plays a huge role in analyzing different use case patterns. Another facet where machine learning is being utilized is in identifying and defending against vulnerabilities in the complete cyber eco-system and web application being a part of it.

Integrating machine learning in web applications are proving to serve as an identification and prevention against web hacking breaches by analysing the usage patterns of the web application. As seen within the above sections, machine learning has been a success in identifying various attacks, and research works have been carried out. Future of web application security lies in the hands of machine learning as we are stepping in the space of large data residing in web applications, logs being written every millisecond and attacks being witnessed at large.

Conflict of interest

All work presented in this chapter is our own research/views and not those of our present/past organizations and institutions. It does not represent the thoughts, intentions, plans or strategies of our present/past organizations and institutions.

Author details

Nilaykumar Kiran Sangani[1]* and Haroot Zarger[2]

*Address all correspondence to: sanganinilay@hotmail.com

1 BITS Pilani-Dubai Campus, Dubai, United Arab Emirates

2 Abu Dhabi Company for Onshore Petroleum Operations Ltd., Abu Dhabi, United Arab Emirates

References

[1] Rouse M. WhatIs.com. What is cyber security? Definition from WhatIs.com [Internet]. [Updated: November 2016]. Available from: http://whatis.techtarget.com/definition/cybersecurity [Accessed: December 2016]

[2] Magicwebsolutions. The benefits of web-based applications [Internet]. Available from: http://www.magicwebsolutions.co.uk/blog/the-benefits-of-web-based-applications.htm [Accessed: December 2016]

[3] Acunetix. Web Applications: What are They? What of Them? [Internet]. Available from: http://www.acunetix.com/websitesecurity/web-applications/ [Accessed: December 2016]

[4] Markoff J. The New York Times. SecurID Company Suffers a Breach of Data Security [Internet]. 2011. Available from: http://www.nytimes.com/2011/03/18/technology/18secure.html [Accessed: November 2016]

[5] ITBusinessEdge. The Most Significant Cyber Attacks of 2013 [Internet]. Available from: http://www.itbusinessedge.com/slideshows/the-most-significant-cyber-attacks-of-2013-02.html [Accessed: December 2016]

[6] McGregor J. The Top 5 Most Brutal Cyber Attacks Of 2014 So Far [Internet]. 2014. Available from: http://www.forbes.com/sites/jaymcgregor/2014/07/28/the-top-5-most-brutal-cyber-attacks-of-2014-so-far/#212d8c5321a6 [Accessed: December 2016]

[7] Risk Based Security. A Breakdown and Analysis of the December, 2014 Sony Hack [Internet]. 2014. Available from: https://www.riskbasedsecurity.com/2014/12/a-breakdown-and-analysis-of-the-december-2014-sony-hack/ [Accessed: December 2016]

[8] Business Insider. The 9 worst cyberattacks of 2015 [Internet]. 2015. Available from: http://www.businessinsider.com/cyberattacks-2015-12/#hackers-breached-the-systems-of-the-health-insurer-anthem-inc-exposing-nearly-80-million-personal-records-1 [Accessed: December 2016]

[9] Woolf N. The Guardian. DDoS attack that disrupted internet was largest of its kind in history, experts say [Internet]. 2016. Available from: https://www.theguardian.com/technology/2016/oct/26/ddos-attack-dyn-mirai-botnet [Accessed: January 2017]

[10] Donaldson S, Siegel S, Williams CK, Aslam A. Enterprise Cyber Security How to Build a Successful Cyberdefense Program Against Advanced Threats. 1st ed. Apress; p. 536

[11] Acunetix. Acunetix Web Application Vulnerability Report 2016 [Internet]. Available from: http://www.acunetix.com/acunetix-web-application-vulnerability-report-2016/ [Accessed: December 2016]

[12] Verizon. Verizon DBIR 2016: Web Application Attacks are the #1 Source of Data Breaches [Internet]. 2016. Available from: https://www.verizondigitalmedia.com/blog/2016/06/verizon-dbir-2016-web-application-attacks-are-the-1-source-of-data-breaches/ [Accessed: December 2016]

[13] SAS. Machine Learning—What it is & why it matters [Internet]. Available from: http://www.sas.com/en_sg/insights/analytics/machine-learning.html

[14] Princeton. COS 511: Theoretical Machine Learning [Internet]. 2008. Available from: http://www.cs.princeton.edu/courses/archive/spr08/cos511/scribe_notes/0204.pdf

[15] WhatIs.com. Machine Learning [Internet]. Available from: http://whatis.techtarget.com/definition/machine-learning

[16] Forbes. A Short History of Machine Learning [Internet]. Available from: http://www.forbes.com/sites/bernardmarr/2016/02/19/a-short-history-of-machine-learning-every-manager-should-read/#57f757ac323f

[17] Wikipedia. Machine Learning [Internet]. Available from: https://en.wikipedia.org/wiki/Machine_learning

[18] Madrigal AC. The Trick That Makes Google's Self-Driving Cars Work [Internet]. 2014. Available from: http://www.citylab.com/tech/2014/05/the-trick-that-makes-googles-self-driving-cars-work/371060/

[19] Basilico J, Raimond Y. http://techblog.netflix.com/search/label/machine%20learning [Internet]. 2016. Available from: http://techblog.netflix.com/search/label/machine%20learning

[20] Trefis Team. Here's Why Twitter Is Increasing Its Focus On Machine Learning [Internet]. 2016. Available from: http://www.forbes.com/sites/greatspeculations/2016/06/22/heres-why-twitter-is-increasing-its-focus-on-machine-learning/#2a378e915aff

[21] Bloomberg Intelligence. Rise of artificial intelligence and machine learning [Internet]. 2016. Available from: https://www.bloomberg.com/professional/blog/rise-of-artificial-intelligence-and-machine-learning/

[22] Brownlee J. Supervised and Unsupervised Machine Learning Algorithms [Internet]. 2016. Available from: http://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/

[23] Dayan P. Unsupervised Learning [Internet]. Available from: http://www.gatsby.ucl.ac.uk/~dayan/papers/dun99b.pdf

[24] De Vos T. Cool Machine Learning Examples In Real Life [Internet]. Available from: http://itenterprise.co.uk/cool-machine-learning-examples-real-life/

[25] Columbus L. Machine Learning Is Redifining the Enterprise in 2016 [Internet]. 2016. Available from: https://whatsthebigdata.com/2016/07/22/machine-learning-applications-by-industry/

[26] Ramasubramanian G. Machine Learning Is Revolutionizing Every Industry [Internet]. 2016. Available from: http://observer.com/2016/11/machine-learning-is-revolutionizing-every-industry/

[27] Pyle D, San Jose C. An Executive's Guide to Machine Learning [Internet]. 2015. Available from: http://www.mckinsey.com/industries/high-tech/our-insights/an-executives-guide-to-machine-learning

[28] Hoff J. A Strategic Approach to Web Application Security [Internet]. Available from: https://www.whitehatsec.com/wp-content/uploads/2016/01/A_Strategic_Approach_to_Web_Application_Security_White_Paper.pdf

[29] Ionescu P. The 10 Most Common Application Attacks in Action [Internet]. 2015. Available from: https://securityintelligence.com/the-10-most-common-application-attacks-in-action/

[30] Trend Micro. How's your business on the web? [Internet]. Available from: http://www.trendmicro.com.sg/cloud-content/us/pdfs/business/tlp_web_application_vulnerabilities.pdf

[31] AppliCure. Available from: http://www.applicure.com/solutions/web-application-security

[32] Networking Exchange. The Top 10 Web Application Security Risks [Internet]. Available from: https://networkingexchangeblog.att.com/enterprise-business/the-top-10-web-application-security-risks/

[33] Commonplaces. 6 Threats To Web Application Security & How To Avoid It [Internet]. Available from: http://www.commonplaces.com/blog/6-threats-to-web-application-security-<-how-to-avoid-it/

[34] Blier N. OWASP Top 10: Application Security Risks [Internet]. 2016. Available from: http://blog.blackducksoftware.com/owasp-top-10-application-security/

[35] RSA. Available from: https://www.rsaconference.com/events/ad15/downloads-and-media

[36] Gupta KM, Govil CM, Singh G. Predicting Cross-Site Scripting (XSS) Security Vulnerabilities in Web Applications. In: 2015 12th International Joint Conference on Computer Science and Software Engineering (JCSSE); 22-24 July 2015; IEEE; 2015. DOI: 10.1109/JCSSE.2015.7219789

[37] Shar KL, Briand CL, Tan KBH. Web application vulnerability prediction using hybrid program analysis and machine learning. IEEE Transactions on Dependable and Secure Computing. 2015;**12**(6):688-707. DOI: 10.1109/TDSC.2014.2373377

[38] Joshi A, Geeta V. SQL Injection Detection using Machine Learning. In: 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT); 10-11 July 2014; IEEE; 2014. DOI: 10.1109/ICCICCT.2014.6993127

[39] V. Ford and A. Siraj, "Applications of Machine Learning in Cyber Security", in Proceedings of the 27th International Conference on Computer Applications in Industry and Engineering, October 13-15, 2014.

[40] Tretyakov K. Machine Learning Techniques in Spam Filtering. In: Data Mining Problem-oriented Seminar, MTAT.03.177; 2004

[41] SecurityWeek Network. Symantec Adds Machine Learning to Endpoint Security Lineup [Internet]. 2016. Available from: http://www.securityweek.com/symantec-adds-machine-learning-endpoint-security-lineup

[42] Splunk Inc. Splunk Empowers IT, Security and Business Teams with Better Data Decisions from Machine Learning [Internet]. Available from: https://www.splunk.com/en_us/newsroom/press-releases/2016/splunk-empowers-it-security-and-business-teams-with-better-data-decisions-from-machine-learning.html [Accessed: DD-Month-YYYY]