



# The Challenges of Leveraging Threat Intelligence to Stop Data Breaches

Amani Ibrahim\*, Dhananjay Thiruvady, Jean-Guy Schneider and Mohamed Abdelrazek

School of Information Technology, Deakin University, Geelong, VIC, Australia

Despite the significant increase in cybersecurity solutions investment, organizations are still plagued by security breaches, especially data breaches. As more organizations experience crippling security breaches, the wave of compromised data is growing significantly. The financial consequences of a data breach are set on the rise, but the cost goes beyond potential fines. Data breaches could have a catastrophic impact not only in loss of company's reputation and stock price, but also in economic terms. Threat Intelligence has been recently introduced to enable greater visibility of cyber threats, in order to better protect organizations' digital assets and prevent data breaches. Threat intelligence is the practice of integrating and analyzing disjointed cyber data to extract evidence-based insights regarding an organization's unique threat landscape. This helps explain who the adversary is, how and why they are comprising the organization's digital assets, what consequences could happen following the attack, what assets actually could be compromised, and how to detect or respond to the threat. Every organization is different and threat intelligence frameworks are custom-tailored to the business process itself and the organization's risks, as there is no "one-size-fits-all" in cyber. In this paper, we review the problem of data breaches and discuss the challenges of implementing threat intelligence that scales in today's complex threat landscape and digital infrastructure. This is followed by an illustration of how the future of effective threat intelligence is closely linked to efficiently applying Artificial Intelligence and Machine Learning approaches, and we conclude by outlining future research directions in this area.

## OPEN ACCESS

### Edited by:

ASM Kayes,  
La Trobe University, Australia

### Reviewed by:

Tauhidul Alam,  
Louisiana State University in  
Shreveport, United States  
Abdur Rahman Bin Shahid,  
Concord University, United States

### \*Correspondence:

Amani Ibrahim  
amani.ibrahim@deakin.edu.au

### Specialty section:

This article was submitted to  
Computer Security,  
a section of the journal  
Frontiers in Computer Science

Received: 14 May 2020

Accepted: 17 July 2020

Published: 28 August 2020

### Citation:

Ibrahim A, Thiruvady D, Schneider J-G  
and Abdelrazek M (2020) The  
Challenges of Leveraging Threat  
Intelligence to Stop Data Breaches.  
Front. Comput. Sci. 2:36.  
doi: 10.3389/fcomp.2020.00036

**Keywords:** data breaches, threat intelligence, data intelligence, machine learning, cybersecurity, artificial intelligence

## 1. INTRODUCTION

Data breaches are one of the top cybersecurity problems affecting the digital economy (Confente et al., 2019; Tao et al., 2019). Recent studies show that data breaches continue to grow year after year, and 2019 broke all previous records with millions of dollars of losses (Ponemon Institute, 2019). For example, in May 2019, Canva suffered a data breach that exposed email addresses, usernames, names and salted and hashed passwords of 137 million users. Equifax, one of the largest credit bureaus experienced a data breach as a result of vulnerability in their website, exposing confidential information of about 147.9 million consumers. The fitness app MyFitnessPal was hit where 617 million customer accounts were leaked and offered for sale on the Darkweb. Unfortunately, data breaches comes with multiple direct and indirect cost factors that are significant to the survival and competitiveness of businesses, including financial, reputational, operational, regulatory aspects. Due to the exponential increase in data breaches, new compliance and regulation laws such as the

GDPR and NDBS regulations (Adams and Bennett, 2018) have been introduced, in an attempt to mitigate their impacts (Voigt and Von dem Bussche, 2017; Hopkins et al., 2019).

According to many data breach detection gap analysis studies (Cheng et al., 2017; Verizon, 2017), most data breach incidents could take several months or even more than a year to be discovered. The latest report from Verizon's security research (Verizon, 2017), that analyzed hundreds of data breach incidents reported by tens of organizations, shows that more than a quarter of incidents had gone unnoticed for many months and one in 10 had gone unnoticed for over a year. For instance, Sony's latest data breach incident had gone unnoticed for almost a year, with an estimated breach loss of 1 Billion \$. A data breach not only affects the competitive edge of an enterprise in the marketplace, but also reduces trust from a viewpoint of the consumers. Rosati et al. (2019) conducted a study to analyse the effect of data breach announcements on market activity and stock price. They found that data breach announcements have a critical effect on both bid-ask spread and trading volume. For example, the recent Equifax data breach announcement had a negative impact on their market activity and stock price. Avoiding reputation deterioration and regulatory sanctions deliver non-quantifiable businesses benefits but have a great added-value to business operations and success, respectively.

Whilst *data loss* and *data leakage* are both aspects of data breaches, the handling of data loss and data leakage are addressed very differently. Data loss is usually a result of an insider hacker or enterprise user that could be unintentional and is usually relatively straightforward to handle using Data loss prevention (DLP) solutions. DLP ensures that internal users do not send sensitive information outside the corporate network. DLP uses business rules to classify and protect business data so that unauthorized users cannot accidentally or maliciously share data. Data leakage is almost always intentional, where enterprise data is placed at significant risk for malicious purposes, and is conducted by external hackers. DLP solutions focus mainly on internal hackers and enterprise users to detect data breaches (Taal et al., 2017). However, the greater source of data breaches is from external hackers (George and Emmanuel, 2018; Dongre et al., 2019). Moreover, incident response plans to a data breach usually take place after data is leaked outside corporate walls. The security team of an enterprise then analyses system logs as the primary way of conducting forensics, and properly managed logs can be used as evidence in a court of law for prosecution purposes. This approach does not limit the effect of a data breach nor stop a data breach.

Threat intelligence has been recently introduced as an enabler to predict future potential security threats even before they reach targeted organizations, by applying basic building blocks of data intelligence and data-driven architectures. Data intelligence are influencing many new technologies with the support of AI to achieve predictive powers. Threat intelligence is changing the current reactive defense approach to a proactive approach that can defend against threats that emerge outside the business threat landscape before they even take place. Threat intelligence is about prioritizing, reducing false alarms that overwhelm security operations and discovering potential threats the organization is

most vulnerable to. Threat intelligence allows security teams to know if the existing defense controls can actually handle those threats or not. At this stage of research, the kind of research questions that are asked are: (i) How to classify and separate threat data feeds (adversarial machine learning) from genuine cyber threat intelligence? and (ii) What kind of intelligence can help organizations predict threats, and how to develop and implement these intelligent solutions?

Threat Intelligence has been recently introduced to enable greater visibility of cyber threats that creates a significant difference to the organization's ability to maneuvering threat countermeasure mechanisms into place, prior to and during the attack. The aim is to enable predicting future potential security threats even before they reach targeted organizations, by applying basic building blocks of machine learning with the support of AI to achieve predictive powers. Moreover, we see a change in the current reactive defense approach where, rather, the approaches are proactive to ensure defending against threats that emerge outside the business threat landscape before they even take place. Threat intelligence is about prioritizing, reducing false alarms that overwhelm security operations and discovering potential threats the organization is most vulnerable to.

In this paper, we discuss the problem of data breaches and the challenges of implementing threat intelligence to stop advanced security threats such as data breaches. The paper is structured as follows: first, we discuss the data breaches problem from both defensive and offensive perspectives, and how threat intelligence can become key enabler to stop data breaches. Then, we discuss the challenges of enabling threat intelligence, and finally, we point out future research directions.

## 2. BACKGROUND

Organizations handle a vast amount of sensitive personal financial and business data, some of which are governed by laws and regulations in local and international jurisdictions. Organizations must view protection of sensitive data as a top priority, given the potentially severe consequences of data breaches. Organizations need to secure their digital infrastructure by adopting appropriate risk management plans that enable businesses to comply with federal laws; reduce financial losses that result of confidential data leakage and ensuring a secure digital environment to business customers and partners to support competitiveness in the marketplace. Existing DLP solutions focus on deploying data discovery agents within an enterprise's digital infrastructure to track and monitor corporate data by monitoring internal users. Unfortunately, this does not actually stop a data breach or even limit its impact.

To be able to defend against data breaches, the first step is to understand how data breaches takes place. A data breach, or what we technically call it data exfiltration, is the process of transmitting confidential data outside the enterprise network boundaries to the internet (Giani et al., 2006; D'Orazio et al., 2016). It is commonly achieved after hackers establish a foothold in an organization's internal network by using sophisticated techniques to remain hidden for long periods of time while

actively hunting for valuable data. Hackers can use multiple pathways to steal data, but the one that is often unknowingly left open is using botnets and Advanced Persistent Threats (APTs) (Chen et al., 2014a). APTs are aggressive types of attacks that enable hackers remain anonymous and hidden thereby allowing them to gain access to enterprise systems, compromise infrastructure and steal data (Marchetti et al., 2016). APT attacks are highly targeted attacks with clear goals and targets are typically governments or enterprises possessing substantial intellectual property value or digital assets that bring competitive advantage or strategic benefits.

The actors behind APTs are typically a group of skilled hackers, working in a coordinated way. They may work in a government/military cyber unit, or be hired as cyber mercenaries by governments and enterprises. This provides them with the ability to work for a long period, and have access to zero-day vulnerabilities and attack tools. When they are state-sponsored, they may even operate with the support of military or state intelligence (Chen et al., 2014b). APT attacks are stealthy, possessing the ability to stay undetected, concealing themselves within enterprise network traffic, and interacting just enough to achieve the defined objectives. Hackers find their target data using various data collection and monitoring tools. Once found, the hackers then need to extract as much data from the enterprise network and slowly exfiltrate the data to avoid detection. To transmit data, hackers typically use backdoors or exploit a vulnerability in the operating system to establish a shell between the compromised host and the hackers servers using a predetermined protocol to facilitate exfiltration such as Domain Name System (DNS) or Hypertext transfer protocol secure (HTTPS). For example, DNS can be misused by hackers to facilitate command and control with a compromised host, move malicious code into a network and exfiltrate data. In this approach, hackers send and receive data via DNS by effectively converting it into a covert transport protocol (Nadler et al., 2019). HTTPS can also be used to exfiltrate data to minimize the risk of detection, as its flexible structure provides a lot of benefits to hackers. It can facilitate command and control with a compromised host and enable undetected large data transfers.

The final step is exfiltrating stolen data to remote servers in encrypted traffic through anonymous networks (e.g., Tor and I2P networks). Anonymous networks consist of a network of relay servers that run by volunteers all over the world. When a hacker connects to the Tor network using a Tor client, a path is created from the user to the destination server to which the hacker needs to connect. This path consists of three relay servers and all the communications through the Tor network are relayed through this pre-built path. All the data going through the Tor network is completely encrypted such that nobody who intercepts the communication has a clue as to who the sender is (Winter et al., 2014). This makes it challenging to identify the source of the attack. With APTs, data leakage might not actually occur until several months after a target system has been compromised. The time hackers take to exfiltrate data depends on many factors such as attack strategy, data size, link speed and installed detection defenses at the target network. For example, the Carbanak APT is an ultra-massive money-stealing campaign

with total losses summing up to 1B to date; the campaign has been active since December 2013, with peak infections and compromised banking systems recorded in June 2014. Hackers may also use other attacks such as distributed denial of service (DDoS) as a means of distraction from the real thrust of data exfiltration. For example, hackers used DDOS against Carphone Warehouse websites to distract its IT team from a coordinated data breach of their customer database that resulted in the theft of 2.4 million customer details.

Detecting APTs is very challenging as it requires a deep analysis of system events at all the Open Systems Interconnection (OSI) model layers that are spaced out over a large period (i.e., months) in a distributed environment that originates from different networks, systems and applications. In other words, detecting APTs and preventing data breaches requires greater visibility into all layers of the digital infrastructure, digital asset activities and the threat landscape. Since APT actors use various stealthy and evasive techniques, there is no known pattern that traditional security solutions could and due to the massive amount of data the need to be analyzed, traditional anomaly detection techniques are no longer an effective solution to detection APTs. Moreover, with the complexity of today's businesses and their digital infrastructure, enterprises now need to understand why and how a breach has happened with logical reasoning to enable effective security operations and risk management plans. Also, when it comes to decision making (i.e., the actions that need to be taken); currently depend on notifying security teams to take investigate the incident and take an action, which limits system capabilities. Automating security operations is no longer a luxury in enterprises, it is a must. Not only because of the increasingly complexity of managing different complex security systems manually, it is also inefficient as it adds human errors factor into the equation. Human intelligence is no longer able to provide a reasonable reasoning and quick and thoroughly make decisions in such complex environments. In a Security Operation Center (SOC), it is difficult to get security teams to streamline attack response and mitigation actions. With the complexity of today's business digital infrastructure and security threats, security teams need a new way to become more agile and autonomous.

Intelligence-driven solutions, such as *threat intelligence* and *data analytics*, have been recently introduced to mitigate the risks of such APT threats and data breaches. Threat Intelligence gives organizations a better visibility of their cyber threats, especially data breaches, to better protect their digital assets (Culnan and Williams, 2009; Roberds and Schreft, 2009; Chou, 2013). Threat intelligence is the practice of analyzing, integrating disjointed cyber and business operational data to extract evidence-based insights regarding an organization's unique threat landscape. This helps to explain who the adversary is, how and why they are comprising the organization's digital assets, what consequences could follow an attack, what assets are possibly compromised, and how to detect and respond to a threat. Proper threat intelligence implementation enables organizations to predict and prevent data breaches and APTs threats targeting to move data outside an organization's secure perimeters at the initial stages before data exfiltration can take place. The key idea behind

threat intelligence can be easily explained with the following use case; consider that a security team of an organization analyses all data breaches that happened in the past along with all the vulnerabilities and exploits that led to the breaches. The outcome of the analysis may be informative to them to gain an understanding what happened in the past but does not help in stopping possible future threats, because it does not provide the power of predicting potential future threats that could lead to a breach. Threat intelligence should provide accurate insights into the implications of the threat landscape, allowing organizations to reduce cybersecurity risks (Mavroeidis and Bromander, 2017).

A key enabler to implement threat intelligence is data intelligence, or in other words, *cyber data management*. Threat intelligence will then turn the cyber data into useful insights to drive effective decisions to defend against potential future cybersecurity threats. Data intelligence provides organizations with a better understanding of their business and threat data to automate and speed risk management and incident response, allowing developing effective and measurable security controls. Applying basic building blocks of data intelligence enables predicting future potential security threats even before they could reach targeted organizations. Data intelligence provides context-rich transparency, visibility and interpretation of cyber data and decisions, to improve productivity, efficiency and effectiveness across organization security posture (Sillaber et al., 2016).

Data and threat intelligence can be achieved with the support of Artificial Intelligence (AI) and Machine Learning (ML) to achieve predictive powers by automating and enhancing the process of regression, analysis, classification and prediction. Roughly speaking, artificial intelligence is the science of finding solutions to complex problems like humans do, where a decision mechanism that is similar to a real human decision mechanism is modeled with some algorithms. Machine learning at its most basic is the practice of utilizing different algorithms to parse data, learn from it, and then make a decision or prediction something about the world based on the outcome of the training and the learning stages. ML is capable of learning from experience, not only to achieve the AI goals (e.g., imitating human behavior), but also to reduce efforts and time spent to take decisions with high accuracy. ML systems learn constantly, make decisions based on data rather than programmed algorithms, and thus change their behaviors accordingly. Cybersecurity is a promising area for AI/ML and in the following section we discuss the hype around the ability of AI-powered security solutions that claim to “do it all.”

### 3. MACHINE LEARNING FOR CYBERSECURITY

Cybersecurity is a critical area in which AI/ML is becoming more significant. Implementing the building blocks of practical AI and ML together with security solutions, facilitates automation and orchestration to build autonomic security solutions that can keep up with the scale, speed, complexity and adaptability of today's cybersecurity threats. Over the past decade ML techniques have been used heavily to enable systematic learning and building

of enterprise systems' normal profiles to detect anomalies and zero-day threats. The core focus has been on detecting security threats in real-time in different contexts such as networks, operating systems, traffic, etc. rather than achieving predictive powers. Hence, with all the hype surrounding AI/ML for cybersecurity, one potential question is how AI/ML techniques can be utilized in cybersecurity to achieve predictive powers to solve different cybersecurity problems such as data breaches problem. In the real-world, not all machine learning techniques are implemented equally or designed as a one-size-fits-all solution. The effectiveness of an ML model is usually determined by its accuracy in predicating the future or making accurate decisions. Implementing ML in cybersecurity to achieve threat and data intelligence has long-standing challenges that require methodological and theoretical handling. These challenges are discussed in the below sections.

#### 3.1. The Complexity of the Threat Landscape

The increasing volume and the quick evolution of the threat landscape makes organizations less immune to the evolving capabilities of modern cyberthreats that consist of a multitude of complex attacks. While organizations are adopting AI/ML to better protect their digital infrastructure, hackers are also adopting them to better identify and more quickly exploit vulnerabilities. This increases the potential for serious impact affecting organizations as it becomes more challenging to predict possible future data leakage strategies. With the significant potential of AI in the threat landscape, hackers are weaponizing it to automate and scale up their hacking activities to avoid detection. Hackers are leveraging new automated hacking and scripting techniques to drastically increase the speed and scale of their attacks by adopting AI to automatically map networks, assess vulnerabilities, define attack vectors and compromise systems.

AI is being used to automate attacks on a larger scale, and not relying anymore on the human element to execute attacks, thus giving birth to new types of security threats. Hackers are undergoing their own digital transformation and leveraging agile development to quicken the pace of malware development to outpace threat analysis techniques and outmaneuver modern security solutions. Countering security threats is a constant game of cat and mouse. The eventual adoption of AI will accelerate this process further. Threat intelligence should deal with the evolving nature of threats and to be adapted continuously for better detection and performance results.

#### 3.2. The Complexity of Cyber Data

To stay protected, enterprises who want to run an agile business need log analysis to navigate the complex world of cyber threats in search of actionable mitigation. Enterprises generate an immense volume of cyber data, which presents security teams with both an opportunity and a challenge. Businesses progressively work toward collecting, aggregate and correlating logs for cybersecurity analytics. All applications, services, operating systems, and networking appliances produce logs full of both useful and useless information. But without

an agile-based log management, much of this data can not be utilized. Logging and log analysis requires substantial amounts of operational time to properly develop analysis rules, criteria and alerts to enable protection (Roberts, 2018; Stevens and Wirth, 2018). The exponential rate of growth of data over the last few years has led to the coinage of novel terminology, “big data.” Every minute, millions of cyber data from different sources in multiple formats are being collected; generating a massive amount of log data and alerts that creates a deafening noise level. With such massive amount of cyber data collected frequently, many organizations are unable to prioritize the most meaningful data, accurately discern patterns and pinpoint trends. It is vital that organizations address these shortcomings to systematically analyse, classify and make sense of the data, in order to discover data-driven competitive features.

Log analytics could be ineffective for threat analysis if it is poorly mined. Poor data management and analytics can lead to too many false alarms being raised, and thus, a higher risk of successful breaches. Cyber data management must deal with a massive flow of extremely granular and diversified data produced in real-time. Clean cyber data is the key to effectively implementing ML approaches. Data leads to metadata that is used by the ML models, and unclean data can lead to inaccurate metadata and, subsequently, wrong results and predictions.

Automation is necessary to get rid of all irrelevant data, while extracting useful insights coming from all kinds of unstructured data sources. Without a precise systematic analysis to classify and make sense of this data, organizations are at risk of not being able to discover data-driven competitive features. Cyber data usually comes in raw formats from different sources, including tracing, logging and resource monitoring events. Hence, log data on its own is insufficient for a holistic analysis and predicting threatening behavior. Other data types such as tracing, performance, operational and business data are important to externalize the state of the system by combining different aspects of the data from an end-to-end execution path with structured and related execution traces.

Another important challenge is training data from multiple distributed sources consumes large amounts of computational resources for a thorough analysis. One common approach to training a global data collected from multiple sources is to collect all training data in a central storage location (Chen et al., 2005; Hazelwood et al., 2018). While such an approach has benefits, it consumes a significant amount of network resources because of the large data transmission and continuous generation nature of training data. For the purposes of dealing with distributed data, the concept of distributed data training has been introduced, where edge computing has been utilized to locally train data and then exchange the model parameters with the other edge servers (Wang and Joshi, 2018; Park et al., 2019). This presents a new challenge in implementing intelligence-driven security architectures.

### 3.3. Feature Engineering

Machine learning algorithms require a significant amount of data to build classifiers that can effectively identify malicious behavior. Determining the right data sources, data sets, and exactly how

much data can be considered to be enough to train a ML model is a challenge, and is associated with the feature engineering problem. This problem is often one of the challenging stages in the development of ML models for cybersecurity. Features are simply the information that characterize a given data sample, and feature engineering is the process of pre-processing existing data to build new and more interesting features. The quality of features selected is more important than the number of features fed into the system. For example, if training data is paired to the wrong set of features, the resulting model produced can be highly unreliable. Hence, it is of vital importance that the right features are used to train an ML algorithm.

Nonetheless, feature engineering is usually guided by domain knowledge, and with the inherent complexity of cyber data, this approach is typically ineffective. This makes it almost impossible to implement an efficient ML model because of the aforementioned complexities, the continuous threat landscape and the evolution of digital transformation. This is why the focus in recent times has been shifted toward deep learning, as it does not require the process of feature engineering.

Deep learning is a type of supervised learning that uses many layers of inter-connected mathematical processes Schmidhuber (2015). Thus, it can be considered as a highly non-linear decision-making engine. The deep learning approach for cybersecurity enables creating a classifier that could identify malicious activities. This can be achieved in multiple ways, but the basic approach is usually the adaptations of traditional artificial neural network (ANN), such as Convolutional Neural Network (CNN) or the Bidirectional Long Short-Term Memory Network (BiLSTM).

ANNs, the underlying structure of deep learning, mimic the human brain. When you provide a neural network with a training set, it runs the set through layers of artificial neurons, which then adjust their inner parameters to classify future data with similar properties. Neurons are the atomic unit of a biological neural network. Each neuron is composed of dendrites, nucleus, and axons. They receive signals through dendrites which are carried by axons. The computations are performed in the nucleus and the entire network is made up of a chain of neurons. In similar fashion, the ANN consists of atomic units, called neurons, that accumulate and sum up inputs from the other neurons and then call an activation learning function. A collection of neurons are the able to classify whether or not a set of inputs belong to a specific class (Schmidhuber, 2015; Albawi et al., 2017). In the context of cybersecurity, ANNs enable taking a decision by looking at past behavior and applying reasoning to understand the behavior by closely considering current and predictive data. A common implementation of ANN architecture is a Feed Forward Neural Network (FFNN), where neurons are arranged linearly in the forward direction inside the network. The first layer consists of input neurons and the input neurons connect to neurons in the hidden layer. In turn, the neurons of the hidden layer are connected to the neurons of the output layer.

Deep learning-based classifiers often outperform traditional classifiers, especially when large datasets are being used. For example, if one needs to train a classifier to analyse data to detect data exfiltration activities, then the output layer is the

only one that needs retraining, and all the other layers can be kept the same. Whereas other ML classifiers will need to be retrained for the entire dataset, which significantly affects their performance and resources required increase drastically. In deep learning, the aim is to use existing data to learn a hierarchy of representations useful for a certain task with no feature engineering involved. The model learns the best representation of the data by itself, enabling scalability and accuracy. Given these advantageous characteristics, deep learning can currently be considered as the most suitable ML technique to address the complexity and dynamic nature of cyber data. Previous research in the field of cybersecurity has indeed shown that this is the case, where for example, deep learning can produce results with over 99% accuracy in detecting unknown threats with 0.0001% false alarms, compared to traditional ML with 50–70% accuracy for unknown threats and 1–2% false positives (Hains et al., 2018; Jiang et al., 2020).

The fact that domain experts are not involved enables continuous training and this can address the problem of the ever-changing threat landscape. A feature-less approach also provides great flexibility without requiring that domain experts and data scientists continually tweak the system. Deep learning is capable of (i) training directly on raw data, with no need for feature extraction, (ii) scales well in dealing with the complexity of cyber data training samples, (iii) continuously improves as training datasets get large, and (iv) picks up on complex patterns, insights and correlations in raw data. These unique features of effective training and learning allows both axiomatic and predictive capabilities, thereby effectively predicting security flaws that might lead to future sudden or gradual data leakage. This makes deep learning unique among all other ML approaches for mitigating data breaches (Guo et al., 2018; Choi et al., 2019).

### 3.4. Transparency and Visibility

Human intelligence is unable to provide useful and effective reasoning, and hence, well thought-out decisions in such complex environments. Businesses need a greater transparency into how ML operates in practice and the entire workflow of training and learning. ML transparency enables taking the right actions that might be needed to fine-tune the ML model to achieve more accurate results and also help businesses to understand their threat landscape to automate incident response and risk management. This understanding is increasingly important as findings obtained from ML models can be admitted as official evidence. The key points of explainability or interpretability of ML models can be expected to play a key role in cybersecurity. Thus, transparency in decision-making is a very important factor to assess ML model efficiency and accuracy in cybersecurity. ML algorithms should be able to tell what data was used to reach a particular conclusion and this is a key problem with many ML approaches.

ML models have a reputation of being “black boxes,” where usually the conclusions found are hard to understand or explain. One of the key issues with ML, and especially deep learning, is that they do not provide explanation or reasoning for the decisions they make (Lipton, 2018). Deep learning models are no different, in that they tend to be black boxes. It is even more

complex with deep learning models due to the complexity of the neural networks and mathematical function employed. The explainability problem of ML in threat intelligence is crucial to learn more about potential attack vectors that could be exploited to lead to data breaches. Hence, an explainable multi-layered neural network is needed in cybersecurity. Misleading ML explanations and lack of transparency of ML predictive models could have serious consequences in the risk management plans of organizations.

Roughly speaking, explainable ML models can be of two types. *First*, local interpretation that explains a single prediction made by a model, or explains a group of predictions. Typically, this is the most common interpretation approach. One of the simple methods to implement a local scope interpretation is using dependency graphs to represent the dependency of a target prediction, however, this is ineffective when considering higher-order interactions. Another local approach is Permutation Importance, which assessed the impact of a feature on the performance of an ML algorithm. This is simply achieved by removing the target features from the learning dataset to assess the impact of the feature on the prediction sample. This approach suffers from the inherent problem that it may give varying results based on which features are being removed in each iteration as it is greatly influenced by correlated features. *Second*, global interpretation to explain the entire model’s behavior and this is challenging even for algorithms with the capacity to achieving interpretability, such as linear models (Hohman et al., 2020). For example, explaining the behavior of a linear model with 100 parameters requires a network of 100 dimensions. This makes it compulsory to utilize feature engineering to achieve interpretable model and this is not an effective approach in cybersecurity as discussed above.

Technically, we can also categorize ML explaining models into two types. *First*, Ante-hoc models that give explanations starting from the beginning of the model with an indication of how certain an ANN is about its predictions (Lipton, 2018). Bayesian Rule List (BRL) (Letham et al., 2015) is an example of an ante-hoc model that yields a posterior distribution over possible decision lists which consist of a series of if-then-statements. “if” statements define a partition of a set of features and the “then” statements correspond to the predicted outcome. This is Similar to DeepRED (Zilke et al., 2016), which applies an if-then-rule for each neuron, layer by layer. The CIE model (Hainmueller and Hazlett, 2013) is used to compare between different predictions and explains why a decision it taken compared to another. *Second*, *post-hoc* models add explainability to a model from its outcome—that is, what part of the input data is responsible for the final decision. Yosinski et al. (2015) introduced a new method to interpret neural networks in a global way by considering the number of neurons activations in each layer.

The existing explainability work is applied to images, texts and tabular data, but not to cyber data yet. Until now, ML explainability in cybersecurity has seen little to almost no research. In cybersecurity, ML-based models should be inherently interpretable, so they are able to provide their own explanations. Previously, we used to think that ML explainability decreases for better prediction accuracy, as the prediction

accuracy increase according to complexity of the ANN. Complex ML models are not necessarily more accurate; in other words, a black box is not necessary for accurate predictive performance. In cybersecurity, a key difference between different ML algorithms is their ability to interpret results and process the data better at the next iteration.

### 3.5. Adversarial Machine Learning

Machine learning will never be a silver bullet in cybersecurity industry to stop hackers, in comparison to fields such as image recognition or natural language processing (two areas where machine learning is prospering). Hackers will always try to find weaknesses in ML models to bypass the implemented security mechanisms, especially since many more hackers are now able to utilize ML to carry out their nefarious endeavors. This has led to the field of adversarial machine learning, which has been a topic of substantial interest in the last few years (Huang et al., 2011; Kurakin et al., 2016; Biggio and Roli, 2018). The problem arises in systems which employ ML models. ML models are typically trained and tested on input data that are assumed to come from the same original distributions. However, if malicious inputs are fed as input to the ML models, the systems could be compromised and security issues such as data breaches can become a lot more straightforward to achieve.

A machine learning model is expected to perform well on unseen testing data after being deployed in the real-world. However, this is often violated in an adversarial real-world, where hackers tries to morph its input adversarial data to increase the mis-classification rate in the ML model. Adversarial attacks begin with a reconnaissance step to observe the behavior of the model and then the gathered information are leveraged to morph the adversarial data. This way hackers can feed their malicious payloads without being detected to execute data exfiltration attacks. Adversarial attacks in systems underpinned by ML can pose a serious threat and feeding adversarial data into deep learning is arguably more complex. Deep learning algorithms are only as good as their data and hackers could feed an ANN with carefully tailored training data that can compromise the model behavior. Because of the opaque nature of neural networks, finding and fixing the adversarial examples of a deep learning algorithm is extremely complex. Ren et al. (2020) investigated deep learning in this context and have confirmed its susceptibility with increasingly complex threats.

There are several open questions, and substantial research must be conducted at the intersection of ML and cybersecurity to deal with the adversaries of ML models. Using multi-faceted ML models could potentially overcome this challenge by detecting difference between results of multiple ML models used to detect a threat scenario. However this approach is not effective with reverse engineering adversarial where the trained model is probed to reverse engineer it to get a better understanding of the prediction model in the underlying system. Adversarial reverse engineering enables hackers to subvert the model itself, where new training data can be created without being detected. Obfuscating the model results is a possible solution that require further research to ascertain its applicability.

Moreover, deep learning has a number of vulnerabilities that could affect the accuracy of the results as discussed by Xiao et al. (2018). The authors analyzed the top 10 deep learning algorithms and has proved that hackers could exploit these algorithms to launch DoS attacks that crash or hang the system, or control-flow hijacking attacks that lead to either system compromise or recognition evasions. It is also notable that the explainability problem could be a double edged sword for deep learning. Full transparency into how ML models operate may expose them to adversarial attacks to alter how they make inferences from live cyber data or to poison them by injecting poison data into the training workflows. Even exposing partial information about how ML algorithms work make them more vulnerable to adversarial attacks.

## 4. RELATED WORK

Machine Learning has proven to be useful in detecting security threats, by analyzing security and log data to identify potential threats. Over the past decade ML techniques have been widely used to enable systematic learning and building of enterprise systems' normal profiles to detect anomalies and zero-day threats (Conti et al., 2018). ML includes a large variety of models in continuous evolution, presenting weak boundaries and cross relationships, and has already been successfully applied within various contexts in cybersecurity (Dua and Du, 2011; Ford and Siraj, 2014; Singh and Silakari, 2015; Buczak and Guven, 2016; Fraley and Cannady, 2017; Ghanem et al., 2017; Yadav et al., 2017; Apruzzese et al., 2018). The book by Dua and Du (2011) provides a comprehensive guide to how ML and data mining are incorporated in cybersecurity tools, and in particular, it provides examples of anomaly detection, misuse detection, profiling detection, etc. This study also provides a thorough analysis of where ML approaches can achieve maximum impact and a discussion of their limitations. The concluding chapters discuss emerging challenges and how ML and data mining can be used to effectively deal with them.

Buczak and Guven (2016) survey ML and data mining approaches in intrusion detection whilst Fraley and Cannady (2017) discussed the future possibilities of incorporating ML into the cybersecurity landscape. In particular, problems such as malware detection, data breaches, profiling, etc. can significantly enhance the threats to organizations. Deep learning in cybersecurity has also been investigated (e.g., by Apruzzese et al., 2018). This study looks at whether current state of the art approaches in ML are effective for identifying malware, spam and intrusions and also allude to the current limitations of these approaches. Studies have also considered support vector machines for dealing with cybersecurity issues (Singh and Silakari, 2015; Ghanem et al., 2017; Yadav et al., 2017). Singh and Silakari (2015) explore support vector machines for cyber attack detection, and in similar fashion, Yadav et al. (2017) focus on the problem of classifying cyberattacks. The study by Ghanem et al. (2017) develop an intrusion detection system which is enhanced by support vector machines. Among other cybersecurity issues,

the study by Ford and Siraj (2014) investigates ML approaches for detecting phishing, intrusions, spam detection, etc.

As we have previously mentioned, a few studies have focused on supervised and unsupervised learning and in particular, methods such as support vector machines, artificial neural networks and deep learning for dealing with a range of cybersecurity related issues. Given the availability of large amounts of data, these approaches and their enhancements provide great potential for future work. Additionally, recent trends in ML have shown that reinforcement learning can be very effective (Sutton and Barto, 2018). In this direction, a recent study demonstrates the usefulness of a deep reinforcement learning approach for cybersecurity (Nguyen and Reddi, 2019). A number of problems such as the detection of intrusions, breaches, etc. can be effectively dealt with this approach given that they are constantly evolving. Another promising ML technique is modeling with Bayesian networks (BNs), which developed in the ML community since the late 1980s (Neapolitan, 2003; Korb et al., 2010). They are causal probabilistic models and there are several studies in a number of domains that demonstrate the applicability (Straub, 2005; Bonafede and Giudici, 2007; Fenton and Neil, 2012; Sýkora et al., 2018). The book by Fenton and Neil (2012) provides a comprehensive overview of how BNs can be applied to risk modeling in different domains, such as systems reliability, law, finance, etc. The recent study by Sýkora et al. (2018) shows how BNs can be used for risk assessment in energy. Straub (2005) use BNs to study the risks associated with natural hazards and the study by Bonafede and Giudici (2007) investigates enterprise risk via BNs. As these studies show, they are particularly suited to modeling risk and can be very effective for the probable of modeling threats associated with data breaches.

Preliminary studies have already demonstrated the usefulness of BNs in cybersecurity (Ramakrishnan, 2016; Wang et al., 2020). The study by Wang et al. (2020) shows that BNs can more accurately classify cybersecurity risk, especially compared to previously known Monte Carlo models. Furthermore, BNs also prove to be more flexible. Ramakrishnan (2016) shows how BNs can be used to model and visual the causal models underlying cyber risks. Unlike other ML approaches, BNs are not black-boxes. Their main advantages are the ease of explaining their findings and the ability to perform a systematic sensitivity analysis. In the context of threats in cybersecurity, another key advantage of BNs, is that they can be used to build a causal model of the factors that contribute to threats. This can be achieved through expert elicitation (Kuusisto et al., 2015) (i.e., through knowledge derived from experienced professionals in the field) or built from data sources or a combination of both. In particular, BNs can be applied to problems in security to predict threats and potential data breaches and also to diagnose how these threats came about.

## 5. DISCUSSION AND FUTURE RESEARCH DIRECTIONS

Threat Intelligence provides accurate insights into the implications of the threat landscape, allowing organizations to build reliable defense strategies to reduce cyber risks. Every

organization is different and threat intelligence frameworks are custom-tailored to the business process itself and the organization's risks, as there is no "one-size-fits-all" in cyber.

We do not have yet the required threat intelligence that can be generalized across different architectures or businesses without complex development and customization. Hence, most of the existing tools still focus on using ML to detect/predict specific security threat scenarios with well-defined inputs (i.e., data streams and logs) and well-defined outputs (security indicators) that can then be integrated into other security metrics and tools, or manually investigated by security analysts in the threat intelligence teams. Utilizing AI/ML to support implementing intelligence-driven security solutions has a number of requirements, as follows:

- A good ML solution should be aligned to both business objectives and security standards, to enable a better understanding of the data patterns that affect important business events and the ability to use this information to provide relevant insights. This helps businesses to understand where and how the solution will be used in the entire cybersecurity process.
- ML models are hypothesis-driven, in other words, given a dataset; can we build a model to find certain scenarios? Thus, it is very important to have a clear articulation of the threat landscape to be addressed using ML. It is also important to understand and design ML models to focus on specific scenarios. This makes it easier to solve the explainability problem that greatly impacts. This makes it easier to drill down into the business threat landscape details and fine-tune the ML models for new threat scenarios.
- ML models will be as good as the quality and quantity of the training datasets. A poor quality training data for your machine learning model will not give accurate results. It is important that datasets are not biased and covers enough variations of the threat landscape to be addressed by the ML models. The success of ML models is highly reliant on the quality of the used training dataset. A training dataset that accounts for all variations of the variables in real world results in developing more accurate models. While a huge datasets is important, the right kind of data is more important as the system learns from this data. Having a sophisticated model is not going to help if poor data is used to train these systems. Training a system on a poor dataset will eventually end up learning wrong lessons and generating wrong results. Data cleansing, or data wrangling, is necessary because if the dataset has flaws or if it contains inaccurate data, it may not be processed by machine learning systems optimally. Furthermore, relying on the training dataset requires assuring that the training dataset is not poisoned by malicious inputs which might lead to a malfunctioning ML model.
- Machine learning continuous learning plan to address environment changes and concept drifting that usually lead to performance degradation of the ML model. There is a need to implement a continuously learning platform for the ML model so we can keep deploying new ML models on regular basis to cope with new threat scenarios in the dataset. It is important to adopt the human-in-the-loop model that allows experts to

assess ML models accuracy and provide feedback to the model to improve response. This feedback is very crucial to know when the ML model starts to degrade or report false alarms, and to use the feedback to enrich the training data to retrain the ML model on more recent datasets.

- To effectively apply deep learning in cybersecurity, Root Cause Analysis (RCA) should be implemented. RCA allows a better understating of the digital infrastructure and threat landscape inter-dependencies. Root cause analysis is a method of problem-solving used to identify problem antecedent and underlying causes to maintaining reliable operations. While the symptom and immediate cause might be easy and quick to solve, failing to detect and treat the root cause will likely lead to the problem recurring. In today's interrelated, complex threat landscape, root cause analysis requires different types of data from a number of monitoring tools including business, threat and risk data. Root cause analysis in the context of threat intelligence should be fully automated to identify dependencies between system events to help you find the root cause of problems and fix them faster.
- Implementing threat intelligence require continuous monitoring and improvement as the environment changes and as new threats emerge over-time. One of the biggest challenges in cybersecurity is monitoring staggering volume of extremely granular and diversified data produced in real-time, while making sense of it to turn raw data into intelligence. An organization's security architecture and security program require continuous monitoring to ensure operations are within an acceptable level of risk, despite any changes that occur. Every organization is different and therefore threat intelligence should be custom-tailored to the business process itself and the organization's risks, as there is no "one-size-fits-all" in cybersecurity.
- Despite the fact that machine learning proposes promising solutions for many cybersecurity problems. However, machine learning itself introduces a new set of vulnerabilities, when used in real-world, which makes it susceptible to adversarial activity. Real-world training and testing data is dynamic and change over time due to uncontrollable operational factors. In an adversarial real-world environment, the problem of concept drift is exacerbated and a static ML model might fail in a dynamic environment. The inability to account for a dynamic and adversarial nature creates a new class of ML risks. Data scientists need to be aware of ML limitations in real-world environments and the unique requirements of the cybersecurity industry.

In the cybersecurity industry, developing a reference model to implement intelligence-driven architectures that can utilize ML to support security analytics and threat intelligence has become an urgent need. Google technical debt of AI systems indicates that in real-world ML applications, the ML model itself is the smallest component of the architecture (Sculley et al., 2015) and the biggest challenge is in the data preparation stages. An intelligence-driven architecture that can utilize ML should have three key components:

- *DataOps*: includes any data related activities such as collection, aggregation, correlation, validation, cleansing and wrangling.

*DataOps* is process-oriented methodology to improve the quality and reduce the cycle time of data analytics and make it possible to meet the data analytics needs in data-driven architectures (Atwal, 2020). *DataOps* governs the end-to-end life cycle of data, including: *first*, data pipeline orchestration to build a directed data workflow that contains all data access, integration, modeling and visualization steps in the data analytic production process. *Second*, automated testing to monitor the production quality of all artifacts in the data analytics process. Automated testing should operate at every step of the data pipeline to eliminate data errors and adversaries that might corrupt analytics. *Third*, deployment automation to allow continuous code and configuration moving from development environments into production, including tracking, updating, synchronization, integration and maintenance of the code, files and other artifacts that drive the data-analytics pipeline. *Fourth*, data model deployment to make reproducible development environments and reusable analytics components, standardize widely used functionalities and facilitate data migration across different environments.

- *AIOps*: the application of AI for IT operations, by combining AI and ML to provide full visibility into a system state and performance. *AIOps* aims to *first*, capture large data sets of any type, from different sources and in varying velocity and volume, while maintaining data fidelity for comprehensive analysis; and *second*, to apply automated analysis to predict the threat landscape by leveraging ML, including model selection, model training and tuning, respectively.
- *DevOps*: is about how the *DevOps* engineers put these *AIOps* and *DataOps* together to embrace the scale and speed needed for data-driven architectures and deliver the cutting-edge data-driven solutions. *DevOps* improves systems agility and flexibility of *AIOps* platforms by automating the path from development to production, predicting the effect of deployment on production and automatically responding to changes in how the production environment is performing.

## 6. SUMMARY

Utilizing artificial intelligence and machine learning to apply threat and data intelligence strengthens an enterprise's security by empowering stakeholders with evidential information on what and how cyber threats are relevant to their business. Adopting AI/ML to predict and stop data breaches requires a holistic, organization-wide threat intelligence strategy that is fully-integrated in the organizational security management framework. This makes it possible to find the needle in the haystack before it pricks you.

The key for businesses in dealing with threats is to find elusive patterns and information that will yield the relevant intelligence. Hence threat intelligence, nowadays, is an essential part of businesses and enterprises and requires combining, processing, aggregating, and analyzing historical threat data. Via threat intelligence, it is possible to identify vulnerabilities, threat actors, existing and potential attack vectors, and models thereby improving the cyber security of businesses.

Intelligence-driven architectures enable greater visibility of cyber threats and hence minimize the threat landscape. While implementing intelligence-driven architectures have long-standing challenges that require methodological and theoretical handling, they indicate a clear trend in future cyber defense technologies. Implementing building blocks of practical AI and ML together within security solutions, facilitates automation and orchestration to build autonomous security solutions that can keep

up with the scale, speed, complexity and adaptability of today's cybersecurity threats.

## AUTHOR CONTRIBUTIONS

AI: cybersecurity work. DT and J-GS: ML work. MA: ML and cybersecurity work. All authors contributed to the article and approved the submitted version.

## REFERENCES

- Adams, M. A., and Bennett, S. (2018). Corporate governance in the digital economy: the critical importance of information governance. *Govern. Direct.* 70, 631–639.
- Albawi, S., Mohammed, T. A., and Al-Zawi, S. (2017). "Understanding of a convolutional neural network," in *2017 International Conference on Engineering and Technology (ICET)* (Antalya), 1–6.
- Apruzzese, G., Colajanni, M., Ferretti, L., Guido, A., and Marchetti, M. (2018). "On the effectiveness of machine and deep learning for cyber security," in *Proceedings of 10th International Conference on Cyber Conflict (CyCon '18)* (Tallinn: IEEE), 371–390.
- Atwal, H. (2020). *Practical DataOps: Delivering Agile Data Science at Scale*. Berkeley, CA: Apress. doi: 10.1007/978-1-4842-5104-1\_7
- Biggio, B., and Roli, F. (2018). Wild patterns: ten years after the rise of adversarial machine learning. *Pattern Recogn.* 84, 317–331
- Bonafede, C. E., and Giudici, P. (2007). Bayesian networks for enterprise risk assessment. *Phys. A* 382, 22–28. doi: 10.1016/j.physa.2007.02.065
- Buczak, A. L., and Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Commun. Surveys Tutorials* 18, 1153–1176.
- Chen, P., Desmet, L., and Huygens, C. (2014a). "A study on advanced persistent threats," in *IFIP International Conference on Communications and Multimedia Security* Berlin: (Springer), 63–72. doi: 10.1007/978-3-662-44885-4\_5
- Chen, P., Desmet, L., and Huygens, C. (2014b). "A study on advanced persistent threats," in *Communications and Multimedia Security*, B. De Decker and A. Zúquete (Berlin; Heidelberg: Springer), 63–72.
- Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M., et al. (2005). MXNet: A Flexible and Efficient Machine Learning Library for Heterogeneous Distributed Systems. *arXiv*. arXiv:1512.01274.
- Cheng, L., Liu, F., and Yao, D. (2017). Enterprise data breach: causes, challenges, prevention, and future directions: Enterprise data breach. *Wiley Interdisc. Rev.* 4:e1211. doi: 10.1002/widm.1211
- Choi, Y.-H., Liu, P., Shang, Z., Wang, H., Wang, Z., Zhang, L., et al. (2019). *Using Deep Learning to Solve Computer Security Challenges: A Survey*. *arXiv*. arXiv:1912.05721.
- Chou, T.-S. (2013). Security threats on cloud computing vulnerabilities. *Int. J. Comp. Sci. Inf. Technol.* 5, 79–88.
- Confente, I., Siciliano, G. G., Gaudenzi, B., and Eickhoff, M. (2019). Effects of data breaches from user-generated content: a corporate reputation analysis. *Euro. Manag. J.* 37, 492–504. doi: 10.1016/j.emj.2019.01.007
- Conti, M., Dargahi, T., and Dehghantanha, A. (2018). "Cyber threat intelligence: challenges and opportunities," in *Cyber Threat Intelligence* (Cham: Springer International Publishing), 1–6.
- Culnan, M. J. and Williams, C. C. (2009). How ethics can enhance organizational privacy: lessons from the choicepoint and TJX data breaches. *Mis. Q.* 33, 673–687.
- Dongre, S., Mishra, S., Romanowski, C., and Buddhadev, M. (2019). "Quantifying the costs of data breaches," in *Critical Infrastructure Protection XIII*, eds J. Staggs, and S. Sheno (Cham: Springer International Publishing), 3–16.
- D'Orazio, C. J., Choo, K.-K. R., and Yang, L. T. (2016). Data exfiltration from internet of things devices: iOS devices as case studies. *IEEE Internet Things J.* 4, 524–535.
- Dua, S., and Du, X. (2011). *Data Mining and Machine Learning in Cybersecurity*, 1st edn. Auerbach Publications.
- Fenton, N., and Neil, M. (2012). *Risk Assessment and Decision Analysis with Bayesian Networks*, 1st edn. CRC Press, Inc.
- Ford, V., and Siraj, A. (2014). "Applications of machine learning in cyber security," in *Proceedings of the 27th International Conference on Computer Applications in Industry and Engineering* (Kota Kinabalu: IEEE Xplore).
- Fraleigh, J. B., and Cannady, J. (2017). "The promise of machine learning in cybersecurity," in *SoutheastCon 2017*. Charlotte, NC: IEEE. 1–6. doi: 10.1109/SECON.2017.7925283
- George, J. and Emmanuel, A. (2018). Cyber hygiene in health care data breaches. *Int. J. Privacy Health Inf. Manag.* 6, 37–48. doi: 10.4018/IJPHIM.2018010103
- Ghanem, K., Aparicio-Navarro, F. J., Kyriakopoulos, K. G., Lambbotharan, S., and Chambers, J. A. (2017). "Support vector machine for network intrusion and cyber-attack detection," in *2017 Sensor Signal Processing for Defence Conference (SSPD)*, 1–5.
- Giani, A., Berk, V. H., and Cybenko, G. V. (2006). "Data exfiltration and covert channels," in *Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense V*, Vol. 6201, ed E. M. Carapezza (Orlando, FL: International Society for Optics and Photonics), 5–15. doi: 10.1117/12.670123
- Guo, W. Mu, D., Xu, J., Su, P., Wang, G., and Xing, X. (2018). "LEMNA: explaining deep learning based security applications," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security* (New York, NY: Association for Computing Machinery), 364–379. doi: 10.1145/3243734.3243792
- Hainmueller, J., and Hazlett, C. (2013). Kernel regularized least squares: reducing misspecification bias with a flexible and interpretable machine learning approach. *Polit. Anal.* 22, 143–168. doi: 10.1093/pan/mpt019
- Hains, G., Jakobsson, A., and Khmelevsky, Y. (2018). "Towards formal methods and software engineering for deep learning: security, safety and productivity for dl systems development," in *2018 Annual IEEE International Systems Conference (SysCon)* (Vancouver, BC), 1–5.
- Hazelwood, K., Bird, S., Brooks, D., Chintala, S., Diril, U., Dzhulgakov, D., et al. (2018). "Applied machine learning at facebook: a datacenter infrastructure perspective," in *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)* (Vienna), 620–629.
- Hohman, F., Park, H., Robinson, C., and Polo Chau, D. H. (2020). Summit: scaling deep learning interpretability by visualizing activation and attribution summarizations. *IEEE Trans. Vis. Comp. Graphics* 26, 1096–1106.
- Hopkins, D., Mooney, L., and Mooney, L. (2019). Caring about the notifiable data breach: the human impact on victims. *Govern. Direct.* 71:433.
- Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B. I. P., and Tygar, J. D. (2011). "Adversarial machine learning," in *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence* (New York, NY: Association for Computing Machinery), 43–58. doi: 10.1145/2046684.2046692
- Jiang, F., Fu, Y., Gupta, B. B., Liang, Y., Rho, S., Lou, F., et al. (2020). Deep learning based multi-channel intelligent attack detection for data security. *IEEE Trans. Sustain. Comp.* 5, 204–212.
- Korb, K. B., Nicholson, A. E., and Nicholson, A. E. (2010). *Bayesian Artificial Intelligence*, 2nd edn. Chapman & Hall/CRC Press, 29–54. doi: 10.1201/b10391-4
- Kurakin, A., Goodfellow, I., and Bengio, S. (2016). Adversarial machine learning at scale. *arXiv [Preprint]*. arXiv:1611.01236.

- Kuusisto, F., Dutra, I., Elezaby, M., Mendonça, E. A., Shavlik, J., and Burnside, E. S. (2015). "Leveraging expert knowledge to improve machine-learned decision support systems," in *AMIA Summits on Translational Science Proceedings* Vol. 2015 (American Medical Informatics Association, (2015)).
- Letham, B., Rudin, C., McCormick, T., Madigan, D. (2015). Interpretable classifiers using rules and Bayesian analysis: building a better stroke prediction model. *Annals Appl. Stat.* 9, 1350–1371. doi: 10.1214/15-AOAS848
- Lipton, Z. C. (2018). The myths of model interpretability. *Queue* 16, 31–57. doi: 10.1145/3236386.3241340
- Marchetti, M., Guido, A., Pierazzi, F., and Colajanni, M. (2016). "Countering advanced persistent threats through security intelligence and big data analytics," in *8th International Conference on Cyber Conflict (CyCon)* (Tallinn), 243–261. doi: 10.1109/CYCON.2016.7529438
- Mavroeidis, V., and Bromander, S. (2017). "Cyber threat intelligence model: an evaluation of taxonomies, sharing standards, and ontologies within cyber threat intelligence," in *2017 European Intelligence and Security Informatics Conference (EISIC)* (Athens: IEEE), 91–98.
- Nadler, A., Aminov, A., and Shabtai, A. (2019). Detection of malicious and low throughput data exfiltration over the DNS protocol. *Comp. Sec.* 80, 36–53.
- Neapolitan, R. E. (2003). *Learning Bayesian Networks*. Prentice-Hall, Inc.
- Nguyen, T. T., and Janapa Reddi, V. (2019). Deep reinforcement learning for cyber security. *arXiv*. arXiv:1906.05799.
- Park, J., Samarakoon, S., Bennis, M., and Debbah, M. (2019). Wireless network intelligence at the edge. *Proc. IEEE* 107, 2204–2239. doi: 10.1109/JPROC.2019.2941458
- Ponemon Institute (2019). Ponemon Institute's 2019 Cost of a Data Breach Study: Global Overview. IBM. Available online at: <https://www.ibm.com/security/data-breach>
- Ramakrishnan, V. (2016). Cyberrisk assessment using bayesian networks. *ISACA J.* 5.
- Ren, K., Zheng, T., Qin, Z., and Liu, X. (2020). Adversarial attacks and defenses in deep learning. *Engineering* 6, 346–360. doi: 10.1016/j.eng.2019.12.012
- Roberds, W. and Schreft, S. L. (2009). Data breaches and identity theft. *J. Monet. Econ.* 56, 918–929.
- Roberts, S. (2018). Learning lessons from data breaches. *Netw. Sec.* 2018, 8–11. doi: 10.1016/S1353-4858(18)30111-9
- Rosati, P., Deeney, P., Cummins, M., Van der Werff, L., and Lynn, T. (2019). Social Media and Stock Price Reaction to Data Breach Announcements: Evidence from US Listed Companies. *Res. Int. Bus. Finance* 47, 458–469. doi: 10.1016/j.ribaf.2018.09.007
- Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Netw.* 61, 85–117. doi: 10.1016/j.neunet.2014.09.003
- Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., et al. (2015). "Hidden technical debt in machine learning systems," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2* (Cambridge, MA; Montreal, QC: MIT Press), 2503–2511.
- Sillaber, C., Sauerwein, C., Mussmann, A., and Breu, R. (2016). "Data quality challenges and future research directions in threat intelligence sharing practice," in *Proceedings of the 2016 ACM on Workshop on Information Sharing and Collaborative Security* (New York, NY: Association for Computing Machinery), 65–70. doi: 10.1145/2994539.2994546
- Singh, S. and Silakari, S. (2015). Cyber attack detection system based on improved support vector machine. *Int. J. Sec. Appl.* 9, 371–386.
- Stevens, C. and Wirth, T. (2018). *Contingency Planning for Data Breaches*. New York, NY: Routledge. 247–262. doi: 10.4324/9780203728703-19
- Straub, D. (2005). "Natural hazards risk assessment using bayesian networks," in *9th International Conference on Structural Safety and Reliability (ICOSSAR 05)* (Rome), 19–23.
- Sutton, R. S., and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. Cambridge: MIT press.
- Sýkora, M., Marková, J., and Diamantidis, D. (2018). Bayesian network application for the risk assessment of existing energy production units. *Reliabil. Eng. Syst. Saf.* 169, 312–320. doi: 10.1016/j.ress.2017.09.006
- Taal, A., Le, J., León, A., Sherer, J. A., and Jenson, K. S. (2017). Technological and information governance approaches to data loss and leakage mitigation. *Comp. Sci. Inf. Technol.* 5, 1–7.
- Tao, H., Alam Bhuiyan, M. Z., Rahman, M. A., Wang, G., Wang, T., Ahmed, M. M., et al. (2019). Economic perspective analysis of protecting big data security and privacy. *Future Gen. Comp. Syst.* 98, 660–671. doi: 10.1016/j.future.2019.03.042
- Verizon (2017). *Verizon's 2017 Data Breach Investigations Report*. Verizon. Available online at: <http://www.verizonenterprise.com/verizon-insights-lab/dbir/2017/>
- Voigt, P. and Von dem Bussche, A. (2017). *The EU General Data Protection Regulation (GDPR)*. A Practical Guide, 1st edn. Cham: Springer International Publishing.
- Wang, J., and Joshi, G. (2018). *Adaptive Communication Strategies to Achieve the Best Error-Runtime Trade-off in Local-Update SGD*. *arXiv*. arXiv:1810.08313.
- Wang, J., Neil, M., and Fenton, N. (2020). A bayesian network approach for cybersecurity risk assessment implementing and extending the FAIR model. *Comp. Security* 89:101659. doi: 10.1016/j.cose.2019.101659
- Winter, P., Köwer, R., Mulazzani, M., Huber, M., Schrittwieser, S., Lindskog, S., et al. (2014). "Spoiled onions: exposing malicious tor exit relays," in *Privacy Enhancing Technologies*, ed E. De Cristofaro, and S. J. Murdoch (Cham: Springer International Publishing), 304–331.
- Xiao, Q., Li, K., Zhang, D., and Xu, W. (2018). "Security risks in deep learning implementations," in *Proceedings of 2018 IEEE Security and Privacy Workshops (SPW)* (San Francisco, CA), 123–128. doi: 10.1109/SPW.2018.00027
- Yadav, K., Pai, T., and Rane, R. (2017). Classification of cyber attacks using support vector machine. *Imperial J. Interdiscipl. Res.* 3, 94–97.
- Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., and Lipson, H. (2015). Understanding neural networks through deep visualization. *arXiv*. arXiv:1506.06579.
- Zilke, J. R., Loza Mencía, E., and Janssen, F. (2016). "DeepRED-rule extraction from deep neural networks," in *Discovery Science*, eds T. Calders, M. Ceci, and D. Malerba (Cham: Springer International Publishing), 457–473.

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Ibrahim, Thiruvady, Schneider and Abdelrazek. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.