Article

A Review: Data science And Cybersecurity

Name: Farha Jabin Oyshee
Batch: LISUM13

Description

A number of adjacent strategies and domains are being studied by cybersecurity data scientists in order to report on Cyber Attack Response and rising trends and threats with the fundamental foundation of constant terms. Thus, cybersecurity data science is also a replacement term for some security professionals. It will simply serve as a foundation for some understandings that require considerable time to understand the classical impact of the concern uncertainty and doubt issue, except that it appears that a dramatic and well-publicized incursion occurs every year.

There is no completely different situation than the recent discovery of the star winds attack. In spite of this, there is still a struggle to delineate the main points of how this intrusion has occurred, and it is quite worrying to only mention this because it is another example of the standard security profession failing. It is under a great deal of pressure and a significant part of it is because of the sheer complexity and growth of the networks and the technology that need to be protected along with the limited resources available. Since information science is providing a variety of palliatives and approaches which will

facilitate enhanced security posture because of the large amount of knowledge complexity and alerts that initiate from the infrastructure, it is becoming increasingly difficult to survey and maintain. There is the assurance of security, but it is also true that not all security professionals are enthusiastic about the prospects for data science within the profession. After all, there remains some doubt about what exactly is data science. There is a potential for fair criticism and the emperor's new clothes aspect to data science. As a result, it's difficult for others to comprehend what information science is comprehensive as a result it is probably composed of a variety of topics.

Data science can be considered both an instructional discipline and a set of strategies. Essentially, it is a set of management techniques and technologies as well as a set of marketing terms, some of which are quite vague. The purpose is not to act as the defender of information science, but to demand a pragmatic approach to data science, which could consist of a collection of tools and techniques. It can assist in building choices and in gathering insights from massive data sets, especially regarding complex development systems and complex, sophisticated, complex, and complex systems. As a result, once we define cyber security data science from a realistic viewpoint, we can simply state that the choice of a number of these strategies can assist assistance with security assurance. The protection goals of Cyber Attack Response methods are attainable through data science. There is a multitude of centered disciplines, including monetary analysis, marketing analytics, and research, where data science methods help those domains to make choices and become the most efficient ones among the models, where there is a focus. A pragmatic model of provide and demand may emerge from the analysis, which is that information sciences offer a variety of techniques. The purpose of these techniques is to create cyber security goals that are more efficient, ranging from network monitoring to increase the subject style of security systems. Against threats and incursions, all of these are necessary. There are three main elements involved in cyber security information science: data engineering, data management, and scientific processes. The analytical techniques encourage cross-domain collaboration. This collaboration has incentives and unified goals between data engineers and data scientists. These techniques are being integrated, so the explicit findings from the analysis involved shaping specific pillars.

According to the in-depth analysis, there is a decent example of this in action. As a tangible example, there's a great case study utilizing Google Cloud technology. This is often massive data cloud-based cyber security data science in action. After watching the case study, it is evident that indeed these three pillars of knowledge engineering, data

science, and collaboration are all elements of this. Technology engineering includes a scientific aspect for gain insight, and the human and organizational aspects will argue that this raises the most discussed topic. One of the explanations is that cybersecurity data scientists were interviewed about what they apply to other domains that work. A higher understanding of the actual techniques they used in their work is possible because data science has such a broad toolkit of strategies. The results were attention-grabbing and implied price exploration. As a consequence of the analysis of the actual domains raised, it became apparent that they could actually be split into strategies and disciplines. These active cyber security data scientists used eight specific domains in their work. These domains raised half had to do with particular techniques, so the partner had a lot to do with a wider body of academic disciplines. An outline of everyone is provided by deep learning and it is believed many have detected this might be a machine learning technique. It relies primarily on nested neural networks for analyzing unstructured information, particularly audio and visual. An extensive data set with tagged samples of the thing one wants to observe or determine. In order to observe advanced cyber security incidents, it is necessary to develop a very refined understanding of aspects. Several analyses have already been done on this topic and it looks very promising. Cybersecurity data scientists have observed that despite being often very promising, some early indications are quite sensible. There is still a lot to refine in this domain, even the simplest practices. It is necessary to have giant data sets and also to have tagged incidents or incidents to feed in as not every organization has access to labeled attacks. A reportable strategy was network graph analytics. It has its mathematical roots back to the 1700s and involves the formal study of network dynamics and interconnections between nodes in network patterns, which can be applied to many different domains, including the social sciences. Networks are the substrate of most of the infrastructure being protected, so this is a very natural sympathy. One of the foremost principles and beneficial aspects of the domain is the study of networks. There are formally applied math measures that will be derived from the study of networks. There are two types of network interactions: spatial relation or reach, strength between variables connecting entities on the network, and users or devices. All of these factors will enable us to qualify and quantify the entities. Network dynamics have been widely used in cyber security, so it has been explored.

It has been reported in case studies that natural language processing and semantic engineering were also methods used to analyze unstructured textual data. Natural language processing involves analyzing unstructured textual data to extract topics, themes, and meanings. Although these are somewhat distinct, they overlap when we talk

about natural language processing. Even analyzing log files and security log files can be done using this method, although it takes a lot of computing power. Structured log files can also be analyzed for patterns and meaning. In most cases, however, the human-readable content is the most commonly used format. News, emails, and this type of content can be used as indicators of threats, of course. In addition to thematic content in language, semantic engineering is closely related to security and assurance frameworks. Ontologies have been developed in recent years to encode such frameworks and approaches. This is a formal representation of a domain of knowledge that can be encoded so that computers can read it. For example, automated computer-driven processes are very promising for understanding context and meaning. Cybersecurity analysis and operations triage are automated processes. There is a great deal of knowledge about forecasting and time series analysis, especially in the field of finance. Econometrics uses some well-understood methods for studying patterns. Cybersecurity data can benefit greatly from this information. In particular, the phenomenon that arises over time can be very subtle, particularly when it comes to attackers and threats that are increasingly conducting or staging their incursions over a long period of time. In this instance, the signal is very subtle and must be detected over time in order to be detected. However, it is also beneficial to establish a baseline of what is normal versus abnormal in a network or device or in user behavior.

It’s to mention that it is an excellent foundation for process mining which is understanding patterns and processes so be that an attack or be that the normal actions that are undergone by a device or a user and so that this these techniques can help to feed in and help us to understand uh understand processes and to better improve anomaly detection so having covered these methods will now go into these disciplines that were raised by the cyber security data scientists and these embody more formal academic disciplines or even professional disciplines and the first encompasses fraud forensics and criminology and of course, this is a very mature domain even going back to the 1800s.

One of the interviewees has mentioned that they felt that this wasnot deferred to or referenced enough in the work of cyber security and that more could be done to derive out best practices so in fact there is quite a lot of research in this area so particularly and fraud analytics that could be cross-applied to security analytics so for instance anomaly detection and finding obscure patterns often criminals want to be hidden so there's alot of research and best practices that can be applied from this these domains and cross-applied in security uh so another uh broad discipline that was raised which is uh sort of aggregated because it was mentioned in several respects as bridging medical

epidemiological ecological and bioinformatics disciplines which is the application of more biological and medical analytical techniques and statistical analysis procedures to the security domain so this has some crossover which is natural with topics like network analytics and there's some actual more than metaphoric similarities between looking at things like virus dispersion and infections uh between these two domains that are are increasingly useful so one of the quotes that came up was that as the networks that are being protected are becoming increasingly complex there is more and more resemblance to biological systems in some respects so there is a natural sympathy that can be understood in applying and cross-applying some of the techniques in these domains to security so another one of the disciplines raised was social and behavioral sciences including economics and game theory and uh well this this is quite understandable given that is essentially have quite a lot of social phenomenon that is occurring on networks and particularly the behavior of attackers and also the behavior of users when they're using devices it's helpful to have an understanding or formal understanding of the human element and to understand how principles such as supply and demand and such as the motivations of threat actors influence the dynamics and behavior of of attackers so one of the areas that also was raised was risk management and i think this is of course no surprise and there's quite a lot of research and writing on this topic um however it is worth just to repeat that without some understanding of risk uh the mission of securing networks is quite difficult because we we need to quantify and identify what uh assets need to be protected and to have some method to conduct triage and to value the time spent on protecting certain resources and and less other resources so that's quite understandable so that covers the analysis of of these adjacent areas that are applied in cyber security data science and there's quite a lot of connection once they're separated into methods.

The disciplines have quite a few connections between them, which is interesting to consider that the methods offer utility which is they offer certain benefits and techniques. There are a number of disciplines that provide context for modeling and understanding deeper phenomena, such as pattern recognition, classification and diagnostics. By combining these two perspectives, a perspective on analytical techniques with a perspective on disciplines that frame models, we will be able to classify use cases for cybersecurity data science, in other words, examine the type of goals we wish to accomplish through methods and techniques. Technology, economic resources, or behavioral characteristics of attackers and users are the types of domains we analyze. To make the last observation concerning some of the threats trends, this matrix combines these two perspectives. During the interviews, thirteen particular adversarial trends were

raised and categorized to focus on those that overlapped with data science. One of the most common observations was that machine learning was increasingly fair game for use in adversarial environments. The fact that this is both an attack mechanism and an object of attack is because we are increasingly seeing the proliferation of algorithms and machine learning-driven systems. It governs many different aspects of infrastructure and business. As a result, objects that are vulnerable to attack can be tricked by injecting poison data into them. The other two trends also overlapped somewhat, which is also a cause for concern. For example, pentesting has become increasingly automated through sophisticated software tools and there are indications that this trend is likely to continue. Machine learning is also considered a part of these toolkits, and could also be beneficial to adversarial actors. The last observation was that it has emerged in adversarial toolkits. There is an increase in the hiring of data scientists in cyber states, so all of this combined creates something of a perfect storm that is concerning.

Based on observation, it appears that data science has equal benefits for both defenders and attackers. At a very basic level, machine learning in particular offers great deals of efficiency for attackers in the context of security. Therefore, it can be used both for offense and defense, as it is capable of identifying patterns in large sets of data, identifying anomalies, automating expert decision-making, and automating very routine tasks, such as reconnaissance and scanning. As a result of these techniques, tools, and methods, adversaries have an advantage in that they are able to innovate, explore, and have a great deal of incentive. In this regard, the question arises as to whether the benefit is on the side of the defender and whether it can be made tangible and practical. To improve security posture, users and devices cannot be defined as these basic entities from a theoretical perspective. This might suggest a need for further research on the theory of modern users, for example. It has become apparent that the modern user is somewhat of a cyborg because he or she interacts with and uses various electronic devices. Automated processes apps are running on these devices and sometimes interacting on our behalf autonomously to gather information and transact with cloud-based systems. Since this is a very complex ecosystem, there is a notion that the individual person, who must sleep and use a particular set of devices at certain times, actually has agents acting on his or her behalf. To achieve this objective, it should better understand the patterns and nature of what is human and what is automated as well as how these two can be combined into an aggregate entity. Through the use of the theories, better defensive mechanisms can be put in place based on the results of that testing. An analysis of patterns leads to behavioral models that offer insight into what might be considered normal behavior for a particular user. It is the individuals or groups that, through that understanding, are able to recognize when a user or cyborg user is not behaving as they should. A greater focus on data

science is needed in the security realm, and it can serve as a useful mediator in this regard by providing data and methods for testing and diagnostics.

Virtual machines and microservices-based architectures have also raised the issue of how difficult it is to trace linear and static paths connecting devices and users. The proliferation of the cloud as well as bringing your own devices is another aspect that makes it harder to trace lineage when applying deep learning to cyber security. In the research, the deep learning skeptic came out, but there are several reasons for this, the most common being explainability, although there is a lot of work being done. A technique to support some of the existing products on the market to develop signatures and mechanisms to detect intrusions. Deep learning can be applied to things like detecting spam, detecting malware, which has a very complex data set and hidden patterns. Clustering or other mechanisms that use missing fields will be discussed in the presentation. In some ways, it is an aspect of the security professional's responsibility to obtain the data. It is more efficient to get data from application administrators or systems administrators and not from security professionals since they place a lot of barriers in the way of sharing very valuable metadata. Network administration data that are related to cybersecurity statistics are merely parametric attributes of data. Instead, the core statistical tests are very often frustrating. A manager would read an article on machine learning and encourage them in the solution to why they are wasting their time doing data analysis. Although the two can work together ideally, there is a lot of schism between traditional statisticians and machine learning engineers in the field based on the methodological underpinnings of cybersecurity data core statistical analysis mixed with machine learning. Those cyber security data scientists actually felt these were complementary approaches and that they should indeed be integrated so explanations and predictions can work together. Reinforcement learning may be used in modeling red and blue team scenarios.

More so than previously, that's a hardcore empiricist approach, which is to pursue simulations in an effort to qualify statistical models. Simulations are based on statistical models that include agents. In any case, the parameters and outline for the behaviors in the simulation require an intensive effort in their own right, which is to qualify the phenomenon before simulating it as a realistic observation. As long as that safety threshold is observed, which is that real cases are observed and are used to extrapolate statistical measures that then feed into simulation-based parameters.

Reference:

[1] Cybersecurity Data Science- Best Practices in an Emerging Profession, Authors: Scott Mongeau, Andrzej Hajdasinski, 2021
[2] Cybersecurity data science: an overview from machine learning perspective
Iqbal H. Sarker, A. S. M. Kayes, Shahriar Badsha, Hamed Alqahtani, Paul Watters and Alex Ng3., 2020
[3] What is Cybersecurity Data Science? By Scott Allen Mongeau, 2019
(linkedin.com)