The world is becoming perpetually connected. IoT, the Internet of Things where every device is becoming connected to the internet, is a growing concern. We understand cyber security is become more important by the day. Therefore, exploring current technologies and how they are growing is essential for growing in cyber defense. Improving cyber security is important because threats are always evolving. My aim is to understand how big data is being used in cyber defense. After understanding how these emerging technologies are shaping cyber security, I also want to understand how cryptography (the science of manipulating and protecting data), an integral field in security is evolving.

Cisco highlights malware, phishing, man-in-the-middle attack, denial-of-service attack, SQL (Structured Query Language) injection, and zero-day exploit as the most common attacks. Malware, phishing, man-in-the-middle attack, SQL injection, and zero-day exploit can potential access data which attacker should not have access to. Malicious software, malware, breaches a network and can covertly transmit data. Phishing can steal sensitive information through fraudulent communications. Information is compromised by a man-in-the-middle attack because attackers insert themselves between a two-party transaction. SQL injection is an attack where attackers insert malicious code to a server using SQL and can access information, they should not have access to. During zero-day exploit attackers emphasize on an announced exploit that has not been patched and so if the exploit allows for access to data, data is sufficiently released.

There are multiple purposes for these attacks and every attack we highlighted can do serious damage other than granting attackers illegitimate access to data. In 2018 there were thousands of data breaches and the top 19 data breaches affected more than 1 million users each with the top breach affecting 1 billion people.

Big Data is collecting larger and more complex data sets from data sources and analyzing said data. Recent advances allow for improved data collecting, storage, and processing. Analytics of Big Data allows for threat detection and detection of malicious or suspicious threats.

Sapienza and a group of researchers are researching a method to use big data to anticipate cyber-attacks. They named their method Discover. They note that cyber attackers typically identify vulnerabilities, acquire expertise to use the vulnerabilities, choose targets, recruit participants, and plan/execute the attack using online forums. Other signals for potential cyber-attacks such as professionals discussing vulnerabilities, threats, and defense measures also reside on online forums. To mine the data they would analyze, Sapienza and her team compiled a list of cyber-security experts and collected data from their social media. This data is collected every hour and stored to be later recovered. They then formed a list of cyber security experts with blogs. From these blogs they extracted related data with a focus on date published, URL, and contents of the blog. The last place they collected from was dark web forums. Sapienza and researchers collected a list of 263 sites that are forums or marketplaces for hacking. Discover is set up to collect data from the dark web forums relating to the information form the cyber-security experts’ social media/blogs and possibly emerging threats. They collected from these forums relating to malicious hacking three times per week. Then they designed 4 dictionaries to filter out words they know will not be beneficial for analysis. The first words filtered out are common English words that would not relate to cyber threats. Then they removed English words that do not mean anything like “to, on, a, for”, etc. Technical terms are filtered out because they do not represent emerging threats. Then general cyber threat terms are filtered out because they do not stand for an individual threat themselves. Since some of the expert’s tweet in Italian, common Italian words were filtered out. From the remaining words they test for recurring words that occur with a term from the threat dictionary, and this is how they generate errors to possible attacks. According to their results they reached an average threat prediction of 84%. The Discover framework has proven to effectively generate warning from a month before attack date like with ransomware Wannacry to a day early like the data breach with cloudpets.

In this case Big Data is used for analyzing data from multiple sources for threat detection which is an improvement over previous examples where only a single source of data was used. (transition to prediction nkorea theats)

North Korea has been systematically improving their cyber warfare capabilities since the 1990s. Most of their cyber attacks are well-coordinated attack that require 3 to 7 months of preparation. These attacks are carried out in neighboring countries and target financial, press, or government institutions. South Korea’s deals with North Korean Cyber attacks by minimizing the damage caused by these attacks. Since these attacks can be sector specific and sometimes not completely traceable to north Korea, than analysis of the attacks is limited and restricts learning about the attacks. A countermeasure system where sharing the information and subject of an attack is initially proposed for sharing malicious code, origin of attack, and hacking info. Lee explains developed countries in surveillance have set up cross established surveillance systems. Currently big data analysis in cyber security is large-scale log analysis, abnormal transactions/actions, and detection of malicious code. Lee proposes real-time information sharing based on a standardized format. With vast amounts of information collected in real time, lee proposes the use of analysis methodologies such as machine learning and cluster analysis to detect changes in cyber-attacks. The infrastructures would prevent more North Korea cyber-attacks because resources can be concentrated on predicted points.

While this is a more theoretical and proposed system. It is a use of big data to prevent cyber attacks. It is very similar in structure to the Discover framework, where information on emerging threats is collected and analyzed to predict attacks. It has a larger-scale affect because it would potentially limit the cyber-warfare abilities of a country with malicious intent.

Attacks on nuclear power plants are a serious concern because they can lead to real threats or terror. Well-known attacks on a nuclear plant are mitigated by “signature-based detection methods or vaccines”. These signature-based methods have patterns they match and block to similar patterns in networks or URLs containing. Signature-based methods struggle identifying emerging threats, therefore analyzing logs improves the security measures. The volume of raw logs has increased and become segmented and coupled with increasing network flow data volume and number of full-packet logs/atypical logs allows for the categorizing of security threats. SIEM is a next gen log analyzing method since it applies self-established rules to store, analyze, and delete logs. With current SIEM methods a detection scenario is needed. Therefore in developing the security scenario, hacking codes are statically and dynamically analyzed in security events to create signatures by observing status of internal information leaks. A set of base rules for analyzing are complemented by a set of threat rules that create a detection. Signatures are iteratively applied to security equipment in the scenario with broadening rules until the threat is detected. Events that caused misdetections are excluded and events that lead to a positive threats lead to the creation of scenarios based on level of threat. Using the system assets and system status hypotheses and theories are used for security scenarios where a security breach will be detected.

During the development phase of the detection scenario, the status of internal information leaks is analyzed followed by the determination of a speciﬁc feature in the hacking codes by performing static and dynamic analyses of security events to create a signature based on the Snort or Yara rule. The signature is applied to the security equipment to check a misdetection or a correct detection. If the created scenario is unable to make a correct detection, additional rule(s) will be added to repeat the detection process. The detection can be performed by establishing a basis for analyzing the logs in the information system ﬁrst, and then applying the threat analysis rules in addition to additional rules generated through the veriﬁcation scenario. When categorizing the events, items that caused misdetections should be excluded. A scenario is then designed following the category deﬁned in the threat scenario, log extracted based on the user deﬁnition, and speciﬁc keyword search. After completing the analyses of the status of information system assets as well as operation status, several hypotheses and theories are established to create a security scenario wherein security breach will be detected based on the rules initially given. Figure 2 shows the rule generation and detection ﬂow in the security scenario. 400 million logs are created daily, with 65% during the working day. In the first process the data from the raw logs are analyzed to determine the data from them is being collected in real time or “in the form of their deployment”. The second process analyzes the equipment tasks and field/code values. This categorizes characteristics of equipment, code/field values. In the third process the data is filtered and extracted by categorizing based on the threat scenario. Ultimately, the refined data is used to analyze the threat scenarios and correlations which will allow for threat detection.

Cyber-physical systems are systems that tightly integrate physical and computing processes by monitoring/controlling data through underlying networks. The data these networks used is sensed is used to compute and determine actions. These networks are complex and prove to be cumbersome with scalability, therefore software defined networks have been more warranted. Software defined networks are nice because they provide flexibility, customizability, and lower processing expenses. Although, these networks are more susceptible to security threats/attacks. Two common attacks are Link Discovery Attack where a link is falsified which gives an attacker control over the traffic and an ARP attack where the attacker pretends to be a specific gateway an creates an invalid address mapping. Both of these attacks are the entry to other attacks. A lot of techniques that deal with these attacks are complex rendering it practically useless or require a subjective threshold which can be unreliable. In this example a predictive model is built to predict the presence of attacks by leveraging 4 machine learning algorithms(Regression, BayesNet, Decision Tree, and Decision Table). The use of big data analytics overcomes the obstacle of monitoring multiple critical measures of communication. The author uses Mininext data to emulate the virtual SDN networks. In the Link Discovery Attack a recipient node sends a message to a controller of a packed using a sender’s ID and address. In the ARP spoofing attack, the attacker finds a compromised node and sends it a message pretending to be a legitimate node leading the target node to leak its information. In the experimental setup there was a small network of 13 nodes, a medium network of 21 nodes. In each attack the attack was simulated near the source and near the destination. On a small network all the machine learning algorithms performed over 99%. The mid sized network had lower results with the highest being 79.96 and ranging down to 71.23 accuracy. Performance of predictive models are size sensitive and structure of networks. The taining time of these algorithms is near linear. The dynamic nature of SDN are dealt with through the predictive models built with machine learning algorithms which prove to predict the presence of attacks.