The Future of Machine Learning in Cybersecurity

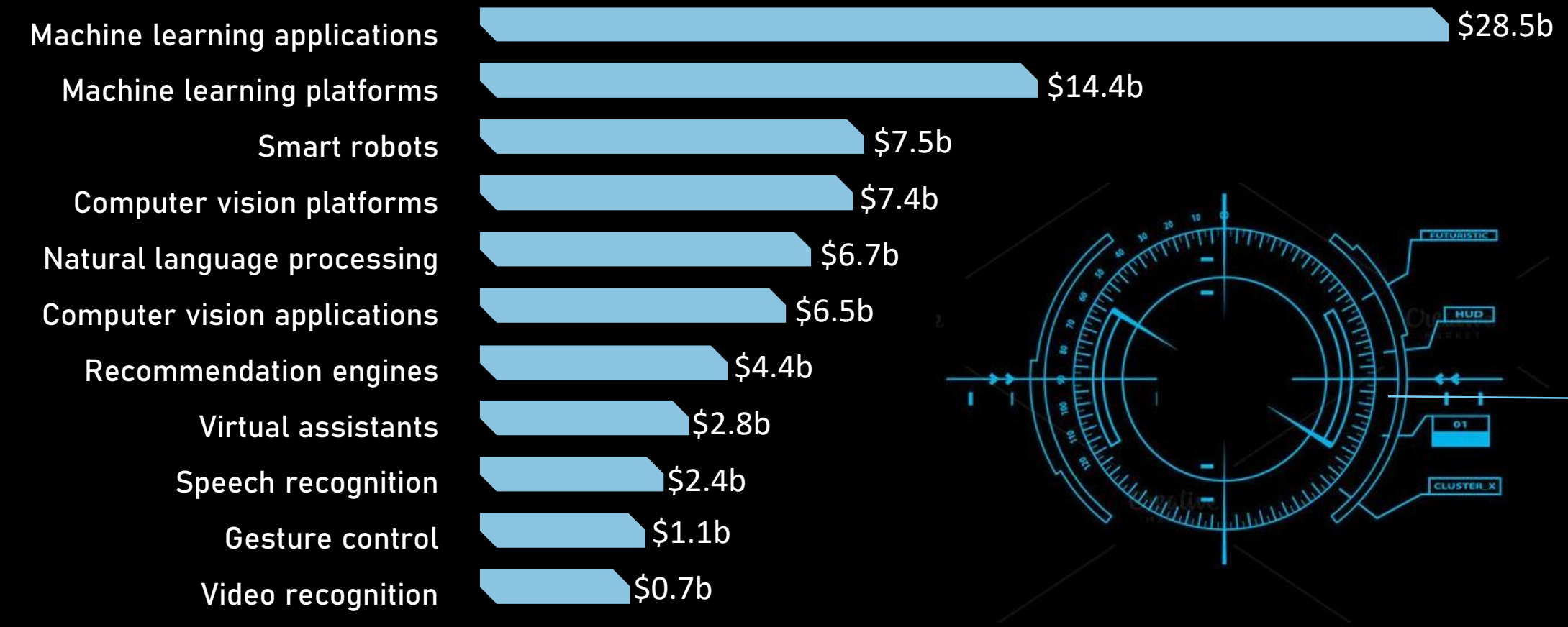
Machine learning and artificial intelligence gradually become integral part of multiple IT solutions across a variety of industries. From transportation and logistics, through banking and finance services to e-commerce and business applications – machine learning (ML) opens the door to adoption of innovative solutions that change industry operational models we have been accustomed to for decades.

As digital transformation is disrupting entire industries, the problem with securing and protecting organizations' digital assets is becoming a business-critical mission. Today's cybersecurity tools apply advanced technologies such as heuristics, deep packet inspection and behavioral analysis and implement a proactive approach toward cyber threats but the malware landscape is evolving so rapidly that we need even more advanced technology to fight against the ever growing number of increasingly sophisticated viruses, Trojans, spear-phishing attacks and ransomware.

Machine Learning Funding Accounts for Bulk Part of AI Investments

Machine Learning Tops AI Dollars

AI funding worldwide cumulative through March 2019 (in billion U.S. dollars), by category



As ML enters the domain of cybersecurity, both IT security experts and organizations face new challenges due to the very complex nature of machine learning and AI algorithms powering a new generation of cybersecurity defense tools. If you are to adopt cybersecurity tools having machine learning and AI capabilities, you need to have deep understanding what ML and AI really are and exactly how you can use these technologies to proactively protect your digital assets and corporate networks.

What is Machine Learning, Really

English logician and crypto analyst Alan Turing gives birth to the modern concept of computing by theorizing how a machine might decipher and execute a set of instructions in 1936. By 1950s, scientists define AI defined as a machine that is able to perform a task that would typically require human intelligence. About the same time the term ‘machine learning’ as a crucial component of any ‘intelligent machine’, or a machine that possesses artificial intelligence.

The history of modern ML algorithms starts in 1952, when researcher Arthur Samuel of IBM managed to develop a computer program that plays checkers and improves its gaming skills over time. In fact, Arthur Samuel coins the term “Machine Learning” in 1952.

During the late 1970s and early 1980s, AI and ML separate into two different research fields with AI researchers focusing on logical and knowledge-based approaches while ML researchers turned to neural networks, algorithms and methods widely used in probability theory and statistics. AI and ML still share common approaches and methods but ML is now mostly used to train AI programs.

The shortest definition of machine learning defines ML as a method and technology to drive a computing device to take actions without being explicitly programmed. This is not exactly artificial intelligence, especially if we bear in mind that AI has higher levels on which we may not be even able to understand the motives behind one or another AI action, but it is very close to an intelligent machine taking actions based on its past experience and predictive analytics.
As of now, businesses can leverage the power of ML in a number of fields such as:

- Sales data analysis by streamlining the data
- Real-time personalization based on past experiences
- Fraud detection through detecting pattern changes
- Product recommendations using data for past purchases
- Decision-making though adoption of ML in business management systems
- Dynamic pricing based on supply and demand
- Natural language processing in various software apps
- Image and face recognition

As you can see, at a high level machine learning is a method for teaching a computer system how to make accurate predictions on the basis of data it gets from one or more external sources. The distinguishing characteristics of ML lie with the fact that the software developer does not instruct the program how to tell the difference between one human face and another in the process of facial recognition, for instance.
Instead, ML is teaching a machine how to predict an answer. Thus, an ML machine is able to respond to situations that it is not familiar with and which it has never encountered before. Scientists achieve this by feeding the ML system with multiple examples from a dataset and then applying a set of rules and algorithms to these datasets.
And you can achieve this in more than one way.

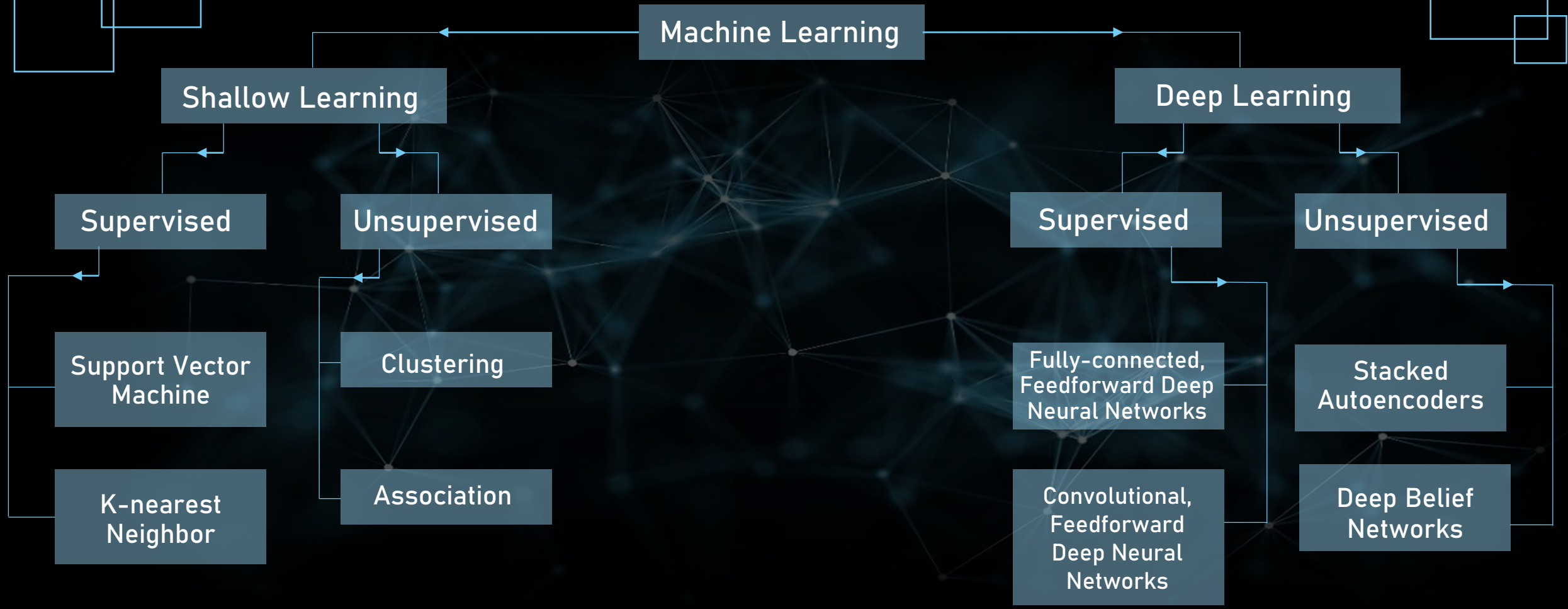
Main Types of Machine Learning

Modern computer science recognizes three main categories of ML, which are:

- Supervised learning,
- Unsupervised learning and
- Reinforcement learning.

Some cybersecurity experts claim that reinforcement learning is a sub-category of unsupervised learning and such a statement is acceptable to a certain degree. Thus, the taxonomy of machine learning methods will look like as on the chart shown below.

Taxonomy of Machine Learning Techniques



1

What Is Supervised Learning

Simply put, supervised learning is teaching a machine by example. The ML system gets large amounts of labelled data with instructions what the input data represents and what answers it should tell. This supervised learning method requires vast amounts of input data, sometimes millions of data records for an ML system to start recognizing shapes, words or image objects and provide correct answers.

Supervised learning is mostly used in ML systems dealing with natural language processing, image recognition, video recognition and the like.

2

What Is Unsupervised Learning

Unsupervised learning takes advantage of algorithms for the ML system to learn to spot patterns in the data provided and then categorize these data.

When you get a newsfeed consisting of news on a specific topic or news category on your connected devices, this is made by a ML system that is powered by unsupervised learning algorithms. Actually, the ML algorithm does not look for a particular type of data but is looking for similarities or anomalies and then groups the news into one category.

In fact, the machine does not know what the input data represents and does not know the expected answers but finds specific patterns and decides on the expected output.

3

What Is Semi-supervised Learning

This ML training method combines the use of a small amount of labelled data and a large amount of unlabeled data. Researches use the labelled data to train a machine-learning model in part, and the resulting partly trained model is used to label the unlabeled input data (pseudo-labelling). The next stage is to train the ML system using the mix of the labelled and pseudo-labelled data.

A new generation of Generative Adversarial Networks (GANs), which are machine-learning systems using labelled data to generate brand new data, have capabilities to train new ML models, which eventually will replace the current methods to train ML systems by training them by the means of generated data.

What Is Reinforcement Learning

Reinforcement learning is very similar to unsupervised learning, but in this case, the ML machine gets feedback on the final result. The best example for reinforcement learning is how you learn to play a game you have never played before. You do not know the rules of the game or how to control it but you learn by experimenting with different buttons, relationships between a pressed button and what happens on your screen as well by seeing your in-game score resulting from certain actions or inactions.

Over time, after playing many game cycles, the ML system will learn what actions are effective and bring higher score and what actions are ineffective. This way, the machine builds a number of winning strategies with every win reinforcing the validity of certain actions or inactions to get the final victory.
This game-based example actually applies to any complex process where a specific outcome, a win, is expected.

4

What Is Deep Learning

In short, deep learning is a sub-category of machine learning where you train your ML model by feeding it huge amounts of data, which the algorithm is able to learn unsupervised and then apply to new data sets.

Algorithms in deep learning originate from human neural networks and that is why deep learning is often referred to as deep neural learning or deep neural network.

The major distinguishing characteristics of deep learning are that in reinforcement learning the ML system learns dynamically with a trial and error method to maximize the outcome, while deep reinforcement learning is learning from existing data and applies it to a new data set.

5

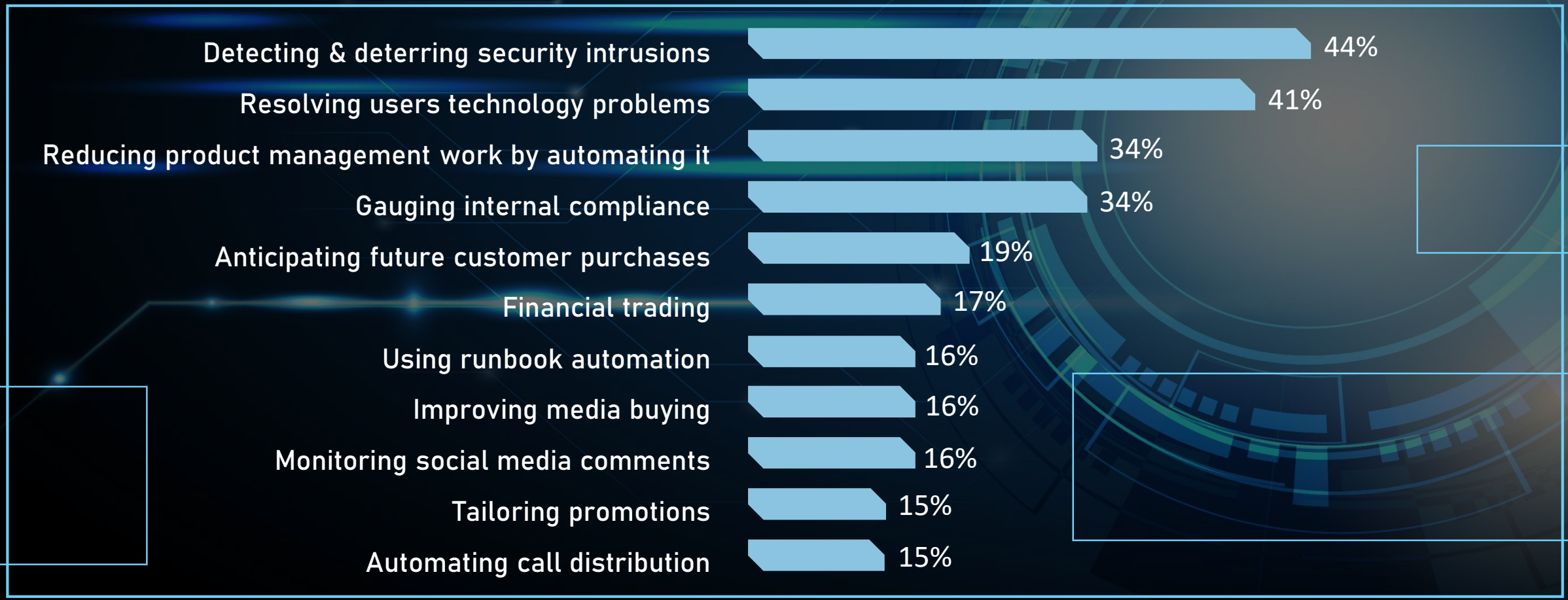
Application of Machine Learning in Cybersecurity

Machine learning powers cybersecurity systems with algorithms that analyze patterns and use what they have learned to detect similar cyber-attacks. ML is able to proactively respond to changing behavior of sophisticated malware and take actions in real time. Furthermore, with ML systems you can automate multiple routine tasks and spend more time on researching possible and emerging threats while putting more time and efforts in building a solid cybersecurity strategy.

AI and ML Evolve into Top Cybersecurity Tools

Detecting Security Intrusions Is Top AI Application in 2018

Application areas of artificial (AI) in organizations worlswide in 2018



The problem with ML in cybersecurity, as well as in any other field, is that you need reliable and accurate data and you need throve of it. For your ML cybersecurity algorithms to be effective, you need to have vast amounts of data originating from everywhere to ensure you have data representing as many potential outcomes from as many potential scenarios as possible.

Your data should be relevant and contextual be it from an endpoint, on the network or in the cloud. You also need to clean the data to be able to understand what happens when a certain event occurs and define outcomes based on the structured clean data. Once you have the right data in sufficient amounts, you are ready to start applying ML across your cybersecurity systems where you can perform a number of tasks that utilize different approaches and methods.

Major Machine Learning Tasks in Cybersecurity

These are mostly tasks that do not involve human actions and work entirely on the basis of ML algorithms that use rich data in context.

- Regression (or prediction) - a task of predicting the next value based on the previous values.
- Classification - a task of separating things into different categories.
- Clustering - similar to classification but the classes are unknown, grouping things by their similarity.
- Association rule learning (or recommendation) - a task of recommending something based on the previous experience.
- Dimensionality reduction - or generalization, a task of searching common and most important features in multiple examples.
- Generative models - a task of creating something based on the previous knowledge of the distribution.

You can use a single approach for some of the tasks or you can apply multiple approaches to solve other cybersecurity jobs.

Those approaches are based on the main types of machine learning we have explored above. Hence, you can choose between supervised, semi-supervised and unsupervised learning. ML researchers also use ensemble learning where different simple models, which they combine to solve a task using the supervised learning approach.

The most current trends in ML bet on developing models through reinforcement learning and active learning. Active learning derives from the reinforcement learning method but here a researcher is manually correcting errors and behavior of a ML model in a changing environment.

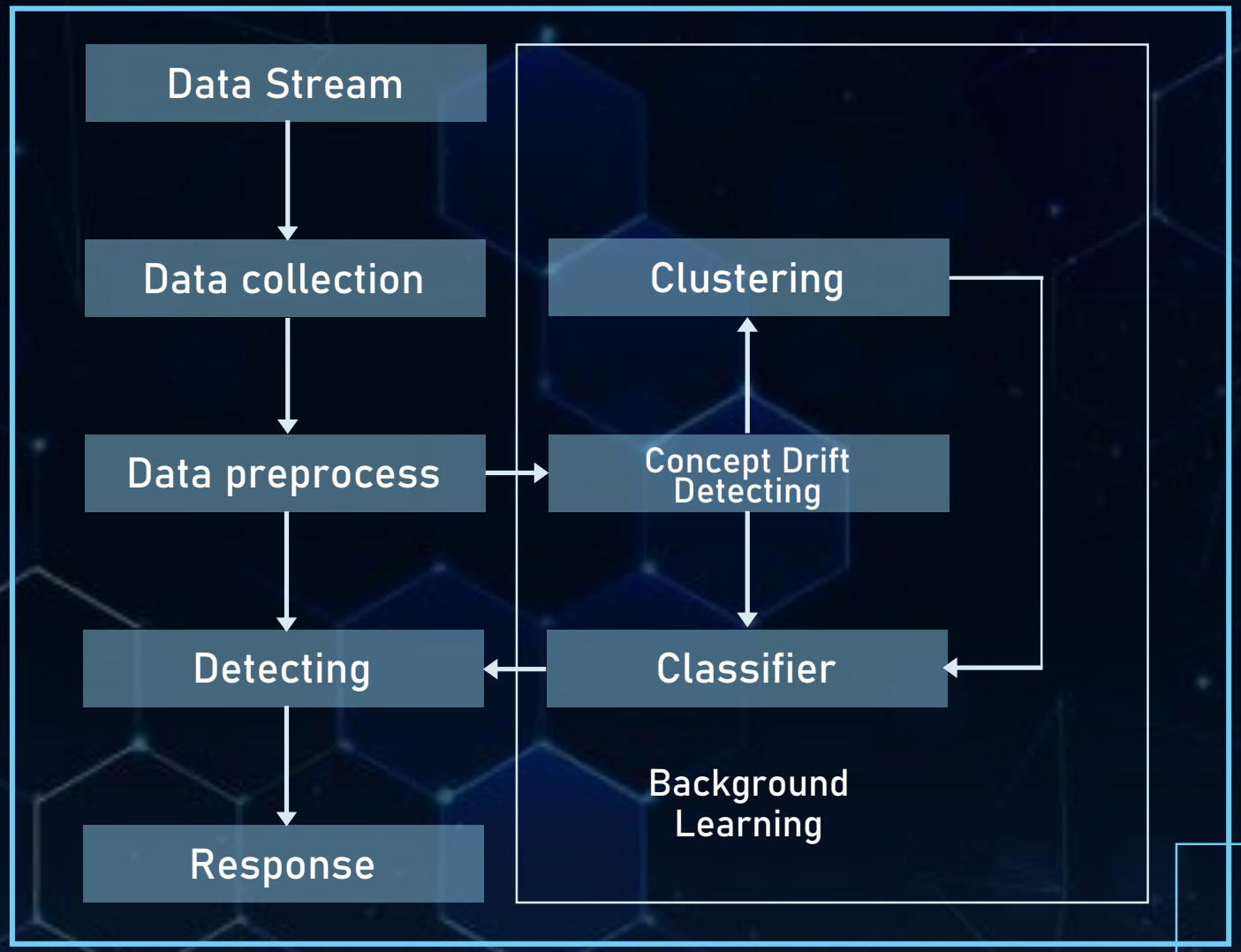
Machine Learning Tasks in Cybersecurity

Prediction, prevention, detection and response are the core of Gartner's PPDR model, which defines five categories of security tasks:

- Prediction
- Prevention
- Detection
- Response
- Monitoring.

In the context of ML cybersecurity, we can illustrate this framework for improved anomaly detection by the following diagram where the ML system deals mostly with detection, response and monitoring but the ML algorithm can be taught also to predict cyber threats.

Basic Model of Cybersecurity ML System

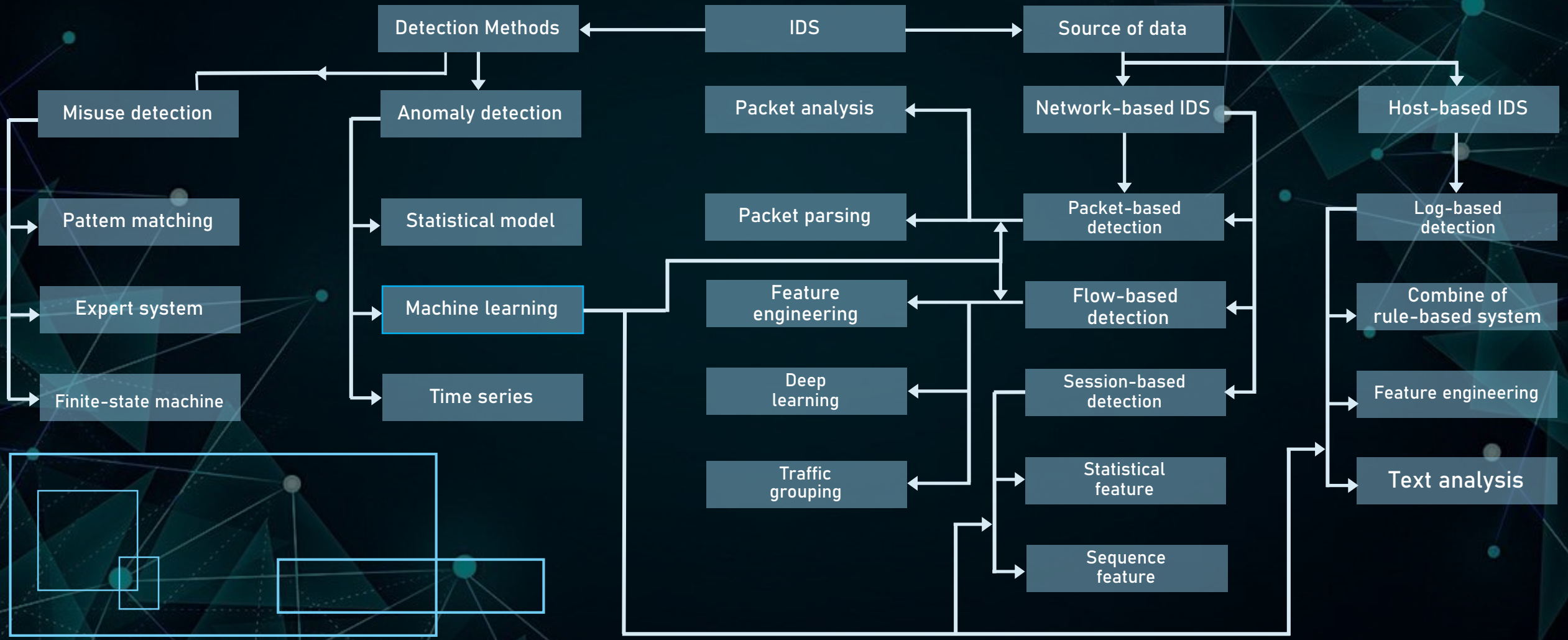


It this sample, intrusion detection works by detecting abnormal behaviors to protect network security. The model is using data mining technology to improve the performance of anomaly detection while some algorithms are improved with abnormal system behavior in mind. Over time, the background learning unit will be able to detect and respond to a growing number of threats and will be able to develop predictive capabilities to detect unknown threats. This is a very simplified model of how a cybersecurity ML model works but still explains the basic steps and the overall concept.

A full-featured Intrusion Detection System (IDS) utilizing machine learning models is capable of performing all the five categories of security tasks by implementing advanced technologies such as statistical models, pattern matching, deep learning, payload analysis, rule-based analysis, text analysis and feature engineering among others.

Taxonomy System of ML-based Intrusion Detection System

In this advanced model for building your IDS defenses, the ML-powered system is able to detect and respond to misuse and anomalies, which in turn protects you against both internal and external cyber threats. You should know that attempts for system misuse could occur both internally and externally when you open your connected apps or online services to third-party users or customers.



Real-life Machine Learning Applications in Cybersecurity

The use of specific methods listed above can solve any cybersecurity tasks that experts allocate to ML systems.

Regression

Regression uses both machine learning and deep learning to get results. When the ML model has learned sufficient amount of data, it can apply it to analyze new data.

For instance, when you have enough knowledge of the amount of suspicious financial transaction, location, frequency of fraud attempts, etc., your ML model is able to predict the probability of attempted frauds. Furthermore, the ML system can detect abnormal actions that indicate that a fraud might be attempted.

Machine leaning and deep learning methods applied for regression include: Liner regression, Polynomial regression, Ridge regression, Decision trees, Support Vector Regression, Random forest as well as Artificial Neural Network, Recurrent Neural Network, Neural Turing Machines and Differentiable Neural Computer.

Clustering

Clustering is widely used in IDS software and works in a similar fashion to classification. This unsupervised learning method differs from classification because you have no information about the classes of the data. It is also unknown if you can classify the training data.

Because of this, clustering is broadly used in forensic analysis. Where you need to classify all activities, which are obscure, to find anomalies. This method is also suitable for detecting suspicious files.

Security teams use the following machine learning methods for clustering: K-nearest neighbors, K-means, Mixturemodel (LDA) DBSCn, Bayesian, Gaussian Mixture Model, Agglomerative, Mean-shift as well as deep learning methods like Self-organized Maps or Kohonen Networks.

Classification

Classification uses a supervised learning approach in which classes and labels for specific data samples are known.

An example for classification ML models are spam filters, which detect new spam messages on the basis of previous experience with unwanted online communication. When you have a large data set containing unwanted messages, you can relatively easily and reliably determine what is spam and what is not.

Machine learning methods in use for classification include: Logistic Regression, K-Nearest Neighbors, Support Vector Machine (SVM), Kernel SVM, NaiveBayes, Decision Tree Classification and Random Forest Classification. You can also use deep learning methods such as Artificial Neural Network or Convolutional Neural Networks.

Association Rule Learning

Association rule learning is a method enabling a ML model to respond adequately when an incident occurs. Incidents usually differ in nature and organizations should have various types of responses.
This way, the ML system is learning what type of response is appropriate for a particular incident.
Algorithms in use for association rule learning include Apriori, Euclat and FP-Growth as well as deep learning methods such as Deep Restricted Boltzmann Machine, Deep Belief Network and Stacked Autoencoder.

Dimensionality Reduction

Dimensionality reduction is required when you need to secure complex systems processing unlabeled data having many features. This method helps you to replace clustering, in which you can deal with limited number of features by eliminating unnecessary features.
On the other hand, dimensionality reduction as well as clustering, is a major task in a more complex ML model. Dimensionality reduction is one of the major methods in use for face recognition tasks in cybersecurity solutions.
Machine learning methods in use for dimensionality reduction include Principal Component Analysis, Singular-value decomposition, T-distributed Stochastic Neighbor Embedding, Linear Discriminant Analysis, Latent Semantic Analysis, Factor Analysis, Independent Component Analysis and Non-negative Matrix Factorization.

Generative Models

Generative models do not rely on existing data but instead simulate the actual data on the basis of the previous decisions.

Generative ML models are in wide use for generating a list of input parameters for testing an application for injection vulnerabilities or scanning a web application for vulnerabilities.

A specific ML module is checking files for unauthorized access while searching for mutated filenames. For instance, the ML system can find a system file named login.php, and then try mutations such as login_1.php or login_backup.php.

Generative models are developed using methods like Markov Chains and Genetic algorithms. Deep learning algorithms in use for generative models include Variational Autoencoders, Generative adversarial networks and Boltzmann Machines.

What the Future Holds for Machine Learning and AI in Cybersecurity

As you can see, machine learning has various and practicable applications in the field of cybersecurity. With ML algorithms, you can reduce human involvement to a healthy minimum and focus on core activities such as analyzing possible unknown threats while paying more attention to developing your core business activities. Deployment of ML algorithms reduces both your cybersecurity workload while eliminating human error. Not that those ML algorithms cannot miss a threat but it is because the algorithms are not perfect and not because they have omitted to spot a malicious action or a piece of code.

On the other hand, ML is adaptive, so they can evolve to detect and respond to a growing category of cyber threats and, with a little help from AI algorithms, to spot previously unknown malicious software and suspicious actions, coming both form outside and occurring inside your corporate networks. Furthermore, ML algorithms are not fixed and you can modify an algorithm to the specific needs and requirements of your organization, which is not quite true for traditional cybersecurity tools.

Moreover, we are witnessing a massive growth of Internet-of-Things (IoT) networks and deployment of smart and connected devices is booming ranging from smart TVs through connected light bulbs to home security systems and smart home and office appliances. You cannot actually secure all these billions of connected devices without some automation and keep them secure without deployment of specific ML algorithms. We are talking about billions and billions of devices distributed across millions of homes and office around the world with many of them connected to a global network where cyber threats are roaming freely. Securing IoT devices' firmware and software is one thing but continuously monitoring those networks for existing and unknown threats definitely requires working cybersecurity algorithms.

Nonetheless, ML models are not a silver bullet, mostly because hackers have the same ML tools at their disposal and utilize them skillfully. There are reports how bad actors have successfully deceived ML programs with misleading data inputs where the algorithm accepts a malicious code or file as a legitimate one. You can use a machine-learning algorithm both ways: you can teach it to detect malware or you can teach it how to defeat advanced cyber defenses. We are now coming to a point where ML and AI algorithms are fighting each other on both sides of the fence but this is an inevitable development as with most other advanced technologies, which are in use by good and bad guys alike.

The future of AI and ML algorithms in cybersecurity is only partially related to detection and response to existing cyber threats. Existing enterprise-grade antivirus suites and IDS systems are doing a good job in detecting known malware and proven types of cyber-attacks.

The future of ML systems, coupled with AI algorithms, is in pro-actively detecting unknown or mutating threats and taking the necessary actions to remediate, isolate and clean each and every suspicious file or code trying to enter your perimeter. As cyber threats evolve, an ML algorithm should also be able to detect unknown attack types that do not necessarily attempt to plant malicious code onto your business systems.

There are attack methods that work passively on your systems and which traditional cybersecurity tools are unable to detect. The future of cybersecurity ML models is to actively seek and respond to such abnormal code and user behaviors that are outside the reach of the classic cyber defenses we have. Equipped with AI and ML algorithms, cybersecurity teams will gradually progress from reactive cybersecurity practices toward pro-active and preventive IT security solutions that require minimal human involvement beyond the creation and continuous development of cybersecurity machine-learning models and systems.