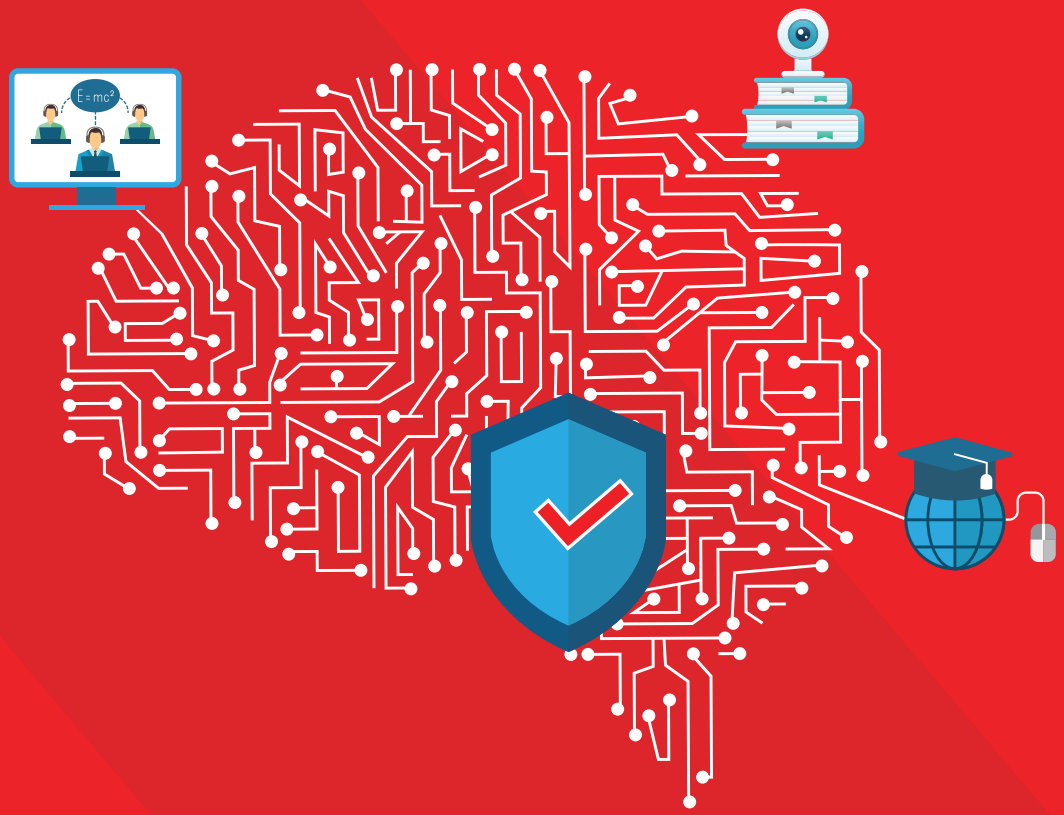


Machine Learning in Digital Security

White Paper

SEQRITE

Table of Contents

This whitepaper was researched and written by:

Abhijit Kulkarni
Associate Director, Quick Heal Security Labs

Harshad Bhujbal
Senior Project Manager, Quick Heal Security Labs

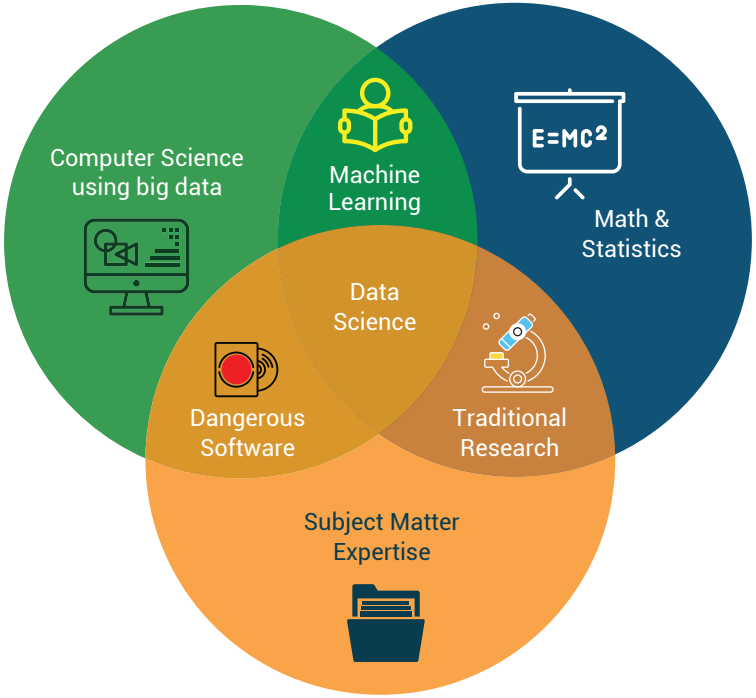
Introduction

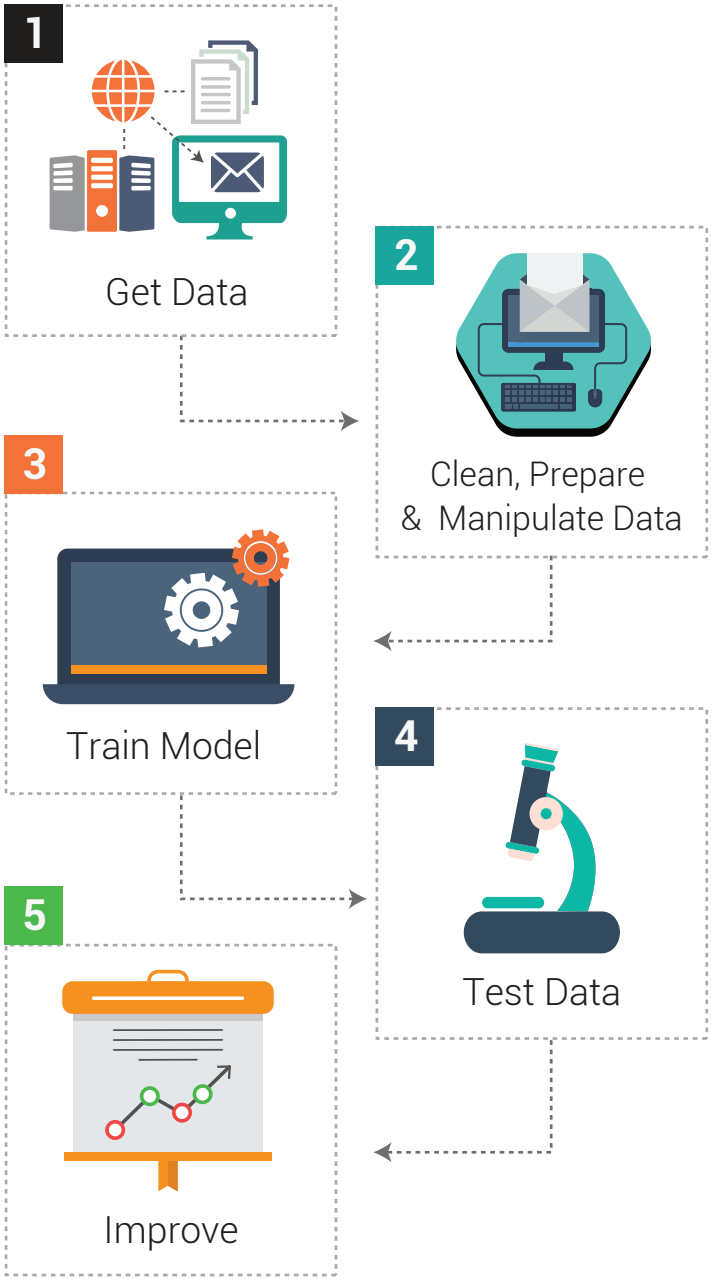
It's a common belief nowadays that data is the new oil. Today's digital world is producing a huge amount of data. But to make sense of this data one needs to mine the data. Data Science and Machine Learning (ML) are the branches which help in this context. And the businesses adopting these 2 technologies have flourished and have edge over those who haven't done it yet.

To briefly describe Data Science, it is a branch of scientific methods to extract insights from data either structured or unstructured and is similar to data mining. Machine Learning is a field of computer science that gives computers the ability to learn from the past/seen data and apply it to the unseen data. Data Mining helps to find new aspects of data which goes as an input to Machine Learning.

Data Science, Machine Learning, and Data Mining technologies are applied in parallel to solve today's complex problems. In this ever-changing world, the cyber threat landscape is getting increasingly complex with multifold increase in threat count. This poses a challenge for the security industry to tackle such complex and huge amount of threats. Manual methods are falling short and hence the industry has already moved to technologies like Big Data, Data Mining and Machine Learning.

In this whitepaper, we will discuss few of the use cases of Machine Learning in the security world.





Introduction to Machine Learning (ML)

Machine learning is a field of computer science that gives computers the ability to learn without being explicitly programmed. Machine learning is the process by which, given data, an algorithm produces a model that when presented with an input (e.g. an image, a binary file, etc.) provides a corresponding output (e.g. there's a mountain in the picture, the file is malicious, etc). In ML, most important step is to prepare data for learning which includes multiple rounds of sanitization of the data. Only then this data is used for training. Machine learning tasks are broadly categorized into 3 types:

- Supervised Learning: Here labelled data is used to train a model which can be later applied to unseen data to label it. The labeling requires domain expertise to label the data to help train the model.
- Unsupervised Learning: Here unlabeled data is used for training which finds patterns or structure in the input data. For example, clustering is an unsupervised method to group similar inputs. Anomaly detection is used to determine whether the input data is deviated from the data on which the unsupervised model was trained. This does not require domain expertise and instead focuses on finding patterns in the data without human intervention.
- Reinforcement Learning: In this category, model uses punishment-reward way of learning. A model attempts to achieve a goal with indirect feedback about the outcome.

Machine Learning usage in Security Industry

ML is being used in almost all the domains and security industry is no exception. In the security industry, it is used at various stages. At Quick Heal, we have huge amounts of threat-specific data which is processed with data mining and Machine Learning algorithm to generate quality detections.

The process of ML starts with extracting features of underlying data and Data Science is extensively used for this purpose. Selecting good features is one of the most important steps in training any ML model. A feature is an individual measurable property or characteristic of the things being observed. For example, samples registering for auto-execution, establishing outbound connection, highly obfuscated code are few of the examples of features. In data, many times there are groups of features that are co-related to each other, i.e. if one feature changes, others also change at some rate. Removing or combining such features saves space and time and improves the performance of machine learning models. This process of reducing the number of unwanted or redundant features is known as dimensionality reduction.

> ***"The global machine learning market is expected to grow from USD 1.41 Billion in 2017 to USD 8.81 Billion by 2022, at a Compound Annual Growth Rate (CAGR) of 44.1%. The main driving factors for the market are proliferation in data generation and technological advancement."***
>
> *- Research and Markets*

At Quick Heal, we use Machine Learning to solve various problems and we have observed significant advantages of it. Here we are listing few of the extensively used cases:

Clustering Samples

We source huge samples per day and machine learning is used to cluster them, map them to existing clusters or generate new ones. Clustering is the task of dividing the samples into a number of groups such that samples in the same groups are more similar to other samples in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters. At regular intervals, these clusters are re-clustered to accommodate newer samples, internally we call it as incremental clustering. Various ML clustering algorithms like Centroid models, Distribution models and Density Models are used for this purpose.

Classifying Samples

A daily job of processing these generated clusters is a mammoth task and Machine Learning is quite extensively used for it. ML algorithms are highly effective to aggregate, analyze large-scale data and to automate the process of classification. Once the clusters are prepared they go to our automated malware classification system wherein we label the data (i.e. clusters) considering the various aspects of examined samples. Many times contextual information is used for enabling insights for classifying sample as malicious. Data mining is extensively used to trace the anomalies among the processed samples to clearly distinguish malicious and benign samples. This process of labeling the data as clean or benign is called as sample classification. These classified samples are then further processed to generate Machine Learning models which can be deployable to endpoints for security.

Deployable Detection Models

The process of Machine Learning model preparation includes various aspects like selecting right set & right ratio of benign & malware samples, dividing the selected set into training & test set and finally selecting right ML algorithms to generate the models.

The generated models are then scrutinized on numerous factors in order to get qualified for endpoint deployment. The judgement process considers several factors like size of the model, the time required to generate model, the time taken by model to scan a sample, quality of model (False positive ratio) etc. The quality of model is a very important factor for the security industry use case as minute false positive ratio may have an adverse effect at the endpoint. These excessively tested, well pruned, highly scrutinized models are considered for endpoint deployment.

Initially, these models operate in passive mode and observe the detection pattern. They are supported by our cloud security platform for mitigating critical errors. The telemetry generated by these passive models is processed in the cloud with our automated systems and based on the results these models are made active.

Conclusion

Machine Learning is one more weapon in our arsenal but to claim it as a silver bullet will be an overstatement. We, at Quick Heal, believe that a combination of Machine Learning, Data Science and human expertise is a winning formula and we are already treading this path. Being a leading security company, we are committed to secure our customers using latest technologies, and these include ML and Data Science.

> ***"Deloitte Global predicts the number of machine learning and implementations will double in 2018 compared to 2017, and double again by 2020."***
>
> *- Deloitte Global Predictions*

Expanding international presence

USA
Quick Heal Technologies America Inc.

JAPAN
Quick Heal Japan KK.

UAE
Quick Heal Technologies (MENA) FZE

KENYA
Quick Heal Technologies Africa Ltd.

Certifications

OPSWAT

AV-TEST APPROVED CORPORATE ENDPOINT PROTECTION WINDOWS TOP PRODUCT

ICSA labs CERTIFIED Excellence in Information Security Testing Award

AV LAB AWARD BEST+++

SEQRITE

Seqrite is a world-class Enterprise Security brand defined by innovation and simplicity. Our solutions are a combination of intelligence, analysis of applications and state-of-the-art technology, and are designed to provide better protection for our customers.

Seqrite is backed by Quick Heal's cutting-edge expertise of producing cybersecurity solutions for over two decades. Our products help secure the networks used by millions of customers in more than 80 countries.

Experience the best-in-class solutions offered by Seqrite and how they can address the security challenges of your enterprise. Boost your cybersecurity,

Request Demo

Quick Heal Technologies Limited
Corporate office: Marvel Edge, Office No. 7010 C & D, 7th Floor,
Viman Nagar, Pune - 411014, India.
Support Number: 1800-212-7377 | info@seqrite.com | www.seqrite.com