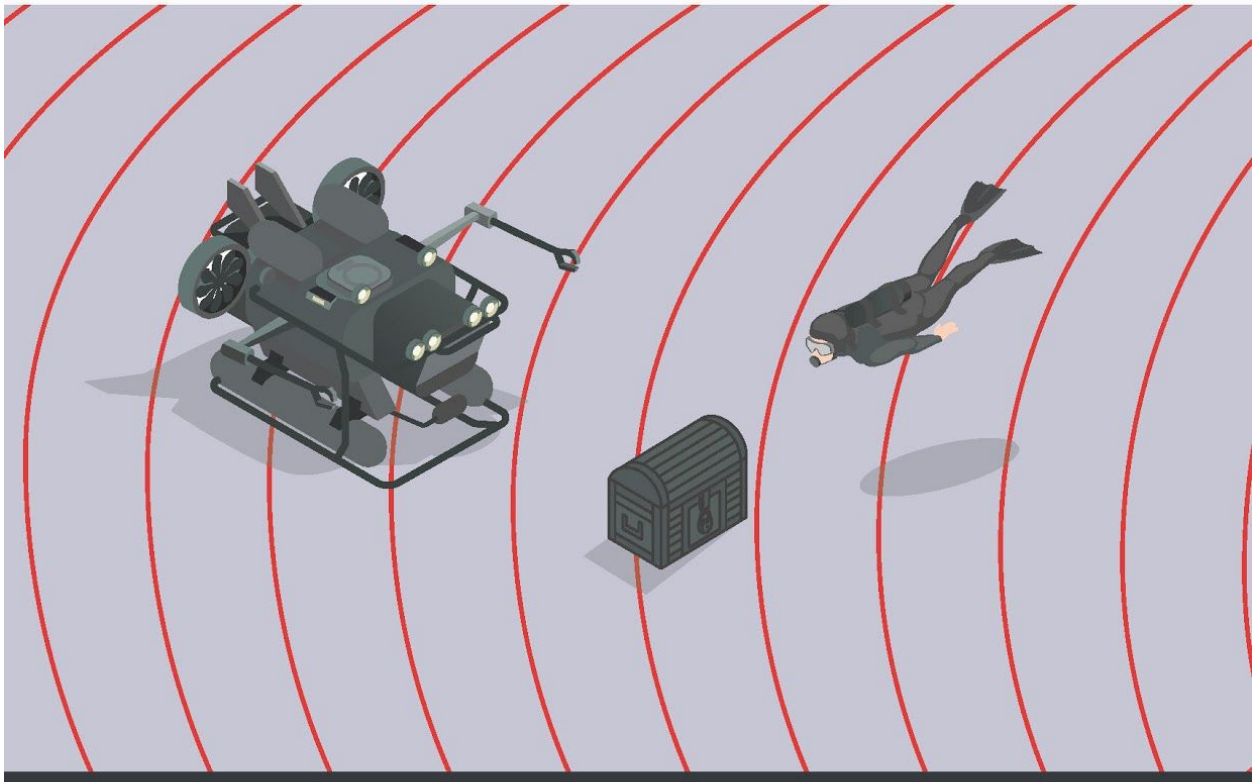
INQUEST



InQuest Machine Learning (ML)

Augmenting Human Analysts to Tackle the Ever-increasing Talent Gap

InQuest Machine Learning (ML)

Augmenting human analysts to tackle the ever-increasing talent gap.

Table of Contents **2**

InQuest Machine Learning (ML)

Augmenting human analysts to tackle the ever-increasing talent gap.

Executive Summary

With data collection methods becoming increasingly sophisticated, information related to the malware and its spread is more expansive and accessible than ever before. Along with this explosion of new data, questions have arisen about how to best harness it, with machine learning being one of many big data techniques for tapping into this vital new resource. The proper application of ML-based modeling to the malware data space has helped to significantly advance the field of cybersecurity, even as it gives attackers the tools necessary to circumvent these new developments.

Machine Learning (ML) results are certainly impressive, yet still suffer from a failure to stay relevant against an ever-changing rogue's gallery of malicious threats. A model trained on data from three months ago might suddenly find itself faced with an entirely new kind of malware it simply wasn't prepared for. The problem is further amplified by adversaries deliberately working to weaken dataset integrity, through the inclusion of "red herring" samples that bias models towards irrelevant criteria.

InQuest's philosophy when tasked with this problem is a sort of "buddy cop" approach, with humans forming teams with ML-generated models to cover each other's backs. This is augmented by InQuest's unique DFI capability, allowing raw data to be efficiently processed into as much useful information as possible for its organic and artificial learners to grow from.

InQuest Company and Solution Overview

The InQuest platform provides high-throughput Deep File Inspection (DFI) for threat detection, data-leakage detection, and threat hunting. We ingest, dissect, and catalog files for real-time and retrospective analysis, leveraging the power of hindsight to apply today's threat intelligence to yesterday's data. Built by SOC analysts for SOC analysts, we empower defenders to save their organizations most precious and limited commodity, human cognition, by democratizing advanced malware analysis skills to reduce analyst fatigue and frustration and increase return on investment with regards to personnel.

We aim to automate and scale the expert knowledge of a typical SOC analyst. Available on-premise, cloud-based, or as a service, the InQuest platform leverages a variety of sources in our automated decision-making engine. This includes bi-directional orchestration with multi-scanning and sandbox platforms, unique threat intelligence sources, and a seasoned signature development team augmented by machine learning.

Founded in 2013, the InQuest leadership and engineering teams are composed of passionate security researchers hailing from both the public and private sectors. Our mission is to deliver our decades of lessons-learned to protect users and organizations everywhere. We strive to maintain our hands-on familiarity with a wide range of prevention, analysis, and monitoring solutions, continuously exploring, examining, and validating security vendors. We continue our community involvement through contributions by way of talks, publications, open-source software, and threat research collaboration. Two unmatched advantages fuel our team's differentiators:

1. We've worked with thousands of real-world exploits from vulnerability discovery and exploitation specialists all around the world.
2. We've vetted nearly every major security vendor under the Sun, having hands-on experience with the best of breed Commercial Off The Shelf (COTS) and Open Source Software (OSS) solutions across the spectrum.

Regardless of how the InQuest solution is deployed, our goals are to:

1. Reduce analyst frustration and fatigue by acting as a force multiplier to support the needs and scale of businesses ranging from small offices to global enterprises.
2. Expose deeply embedded malicious logic through novel methods, automating typically human-intensive tasks to democratize expert-level skill sets to a wider audience.
3. Utilize any and all analyst and threat intelligence resources available in customer environments to automate the identification and validation of threats and data theft.

Machine Learning: A 50,000 Foot View

While machine learning is often associated with artificial intelligence, and while the two are related, this comparison is incorrect. Rather, machine learning (ML) is just one of many components that make up AI, a group that also includes simple computer logic and the optimization of search strategies. In general, an ML algorithm is a strategy for forming mathematical models from preexisting data, models that are able to generalize unseen data despite not having seen it before. This can include algorithms as complicated as deep neural networks built to recognize faces, to a single linear equation that can distinguish spam mail from legitimate content. Thus, ML not only applies to solve AI-related problems but can be useful in a statistical capacity as well.

Most ML algorithms fall into one of two camps, supervised and unsupervised, though some methods blur the line between the two. What follows is a brief overview of these two categories, though more can be found on our Ex Machina blog[1]. A supervised algorithm builds its models out of data that's already been divided up into two or more labeled categories, with the end goal being a model that is able to accurately classify future data by those same labels. Such learning

[1] https://inquest.net/blog/2018/11/14/Ex-Machina-Man-Plus-Machine

methods include random forests, logistic regression, and gradient boost. Classifying samples using these methods is known as “pattern recognition,” one of the two main goals of ML.

Unsupervised methods, conversely, are fed unlabeled data, which they themselves divide up into various categories of their own creation. This can serve a lot of different purposes, from limited search functions that can return similar predigested data points for any given input, to simply elucidating groupings that humans weren’t previously aware of. It can even be used as another form of classification; if most of the members of a group are found to be malicious, for example, the remaining samples considered benign might qualify for in-depth human examination as potential false negatives. Finding interesting samples in this way, where they’re detected for *not* conforming to the model’s expectations, is known as “anomaly detection,” the second goal of ML. Methods of this type include Kmeans[2], DBSCAN[3], and OPTICS[4].

Machine Learning Within Information Security

While often only talked about when it’s used to defeat professionals in games like chess or Jeopardy, machine learning has a variety of widely-used practical applications across a variety of disciplines, from business to medicine. As advancements in data collection and storage continue to be made, its use will continue to grow. Cybersecurity has embraced ML techniques no less readily than any of these other fields, putting it to use in network anomaly detection, semantic analysis, image recognition, and, of course, malware classification. Whether it’s for prediction, prevention, detection, response or monitoring, ML has been tried and applied to solving problems. However, it is far from a universal solution, and, like all tools, works best as part of a larger ensemble.

For all of its advantages, machine learning does have one major weakness: models are inherently prone to bias. No model will ever be a perfect representation of the factors that do or do not determine malware, and if an attacker manages to learn the logic behind one, it can become trivial to modify a malicious program to conform to the standards by which a benign candidate is qualified. This was exemplified last year, when reverse engineers found a way to consistently fool Cylance, a popular antivirus product built on several machine learning algorithms[5].

In order to avoid such adversarial manipulations, it is important to not only regularly update your models with new data, but to recognize that ML is not a replacement for human analysis. Similar to a bloodhound, it’s able to pick up on clues humans cannot, but should not be left to solve an entire case by itself.

---

[2] https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1
[3] https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering/
[4] https://www.geeksforgeeks.org/ml-optics-clustering-explanation/
[5] ttps://blog.morphisec.com/machine-learning-cant-protect-you-from-fileless-attacks

Applications Within InQuest

When it comes to our own ML applications, it must be understood that one can't have good models without good data. Our good data comes from our proprietary DFI technology. DFI is a core tenet of our solution. DFI is a static-analysis engine that peers deep beyond layer 7 of the OSI model, essentially automating the work of your typical SOC analyst or security researcher. Regardless of the novelty of nesting employed by an attacker, DFI will rapidly dissect common carriers to expose embedded logic (macros, scripts, applets), semantic context (e.g. cells of the spreadsheet, words in a presentation), and metadata (e.g. author, edit time, page count). Images discovered to be embedded are processed through a machine vision layer (OCR, perception hashing), adding to the semantic context extracted from the original file. Common evasive characteristics and encoding mechanisms are automatically discovered and deciphered. The DFI process typically results in three times (300%) the amount of analyzable content. For example, 6MB of data may be derived from a 2MB file, resulting in 8MB of total inspectable content.

Once our data is gathered, it can be passed on to our classifiers. Our current machine learning endeavors have resulted in the creation and implementation of two different ensembles, supervised and unsupervised, for classifying malware based on pattern recognition and anomaly detection, respectively.

Our supervised classifiers come in three flavors:

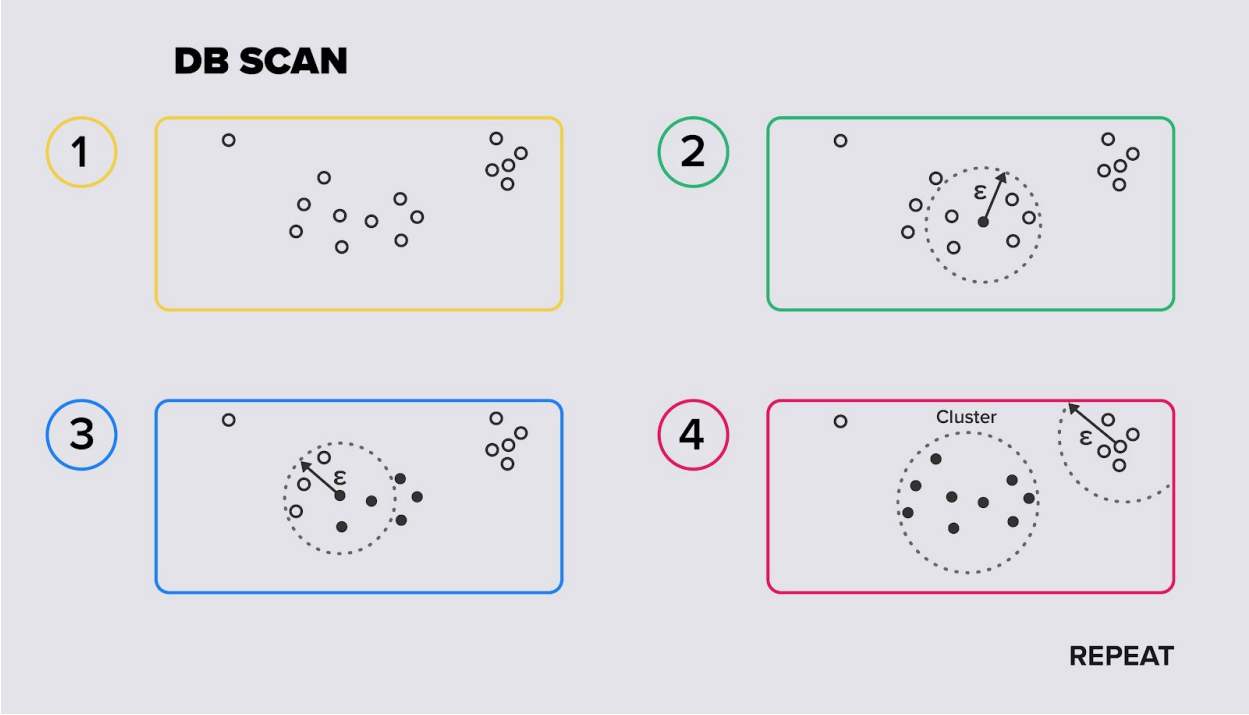
1. Logistic regression: the simplest machine learning model of all. It works to define a single linear equation which, when applied to any given input data, outputs a number that can be related to the probability of that data belonging to one class or another.
2. Random forests: an in-depth explanation can be found on our machine learning blog[6]. In short, it's a large collection of yes/no splits (known as decision trees) resulting in a final guess. Input data is run through all trees simultaneously, with the guess that reoccurs the most being used as the final prediction.
3. Gradient boosting: similar to random forests, only rather than data being run through all trees simultaneously, it goes through one tree at a time, with the results from one tree being slightly modified based on the results of the tree before.

Our five unsupervised clusters provide another method of classification. While traditionally local sensitivity hashing, K-means, and other such clustering algorithms have been used as either search functions or for defining certain "families" of samples (and are also used as such by InQuest,) we've also found them useful for binary classification. Similar to the logic behind nearest-neighbors classification, which assumes that it is likely for all members of a cluster to

[6] https://inquest.net/blog/2018/12/28/Ex-Machina-Random-Forests

share a label, we're able to use clustering to contribute to our labeling while simultaneously organizing our data for further use elsewhere. Our five unsupervised classifiers are as follows:

1. TLSH: a minhash-related clustering technique[7]. Converts data to hashes and compares the hashes to quickly find similar samples.
2. SSDeep: another hash-related method. Note that both this and TLSH aren't ML algorithms in and of themselves, but can be used to aid in ML efforts.
3. K-means: a process that starts with the data being divided up randomly into a specified number of groups, and then gradually adjusts the boundaries between these divisions until the differences in the feature data of all samples within a group is at the minimum possible for the dataset.
4. DBSCAN: measures a specified distance out from a single data point. If a specified number of other data points lie within that distance, they are all grouped into a single cluster. Then this process is repeated with other points in that cluster, with more and more points being added as long as they meet this criteria. Once no more points are found that reach the set number of points within the set distance, a random point from outside the cluster is chosen and the process repeats, until all points have either been assigned a cluster or been found to be isolated, in which case they are marked as noise.
5. OPTICS: similar to DBSCAN, but optimized for dealing with larger datasets.



[7] https://inquest.net/blog/2019/02/28/Ex-Machina-Family-Matters

Conclusion

In this paper we provided readers with high-level background information regarding machine learning and its current uses in information security. We've demonstrated our own personal approach to the problem, and touched on our unique advantages, such as DFI. As our data collection continues to improve, so too will our machine learning applications.

Readers interested in interacting with a lightweight version of DFI to better grasp the plausible benefits availed from additional dimensions of data (DFI) and time (RetroHunting), are encouraged to tinker with an open data portal maintained by InQuest Labs at:

https://labs.inquest.net

It should be noted that this portal hosts known malware for download to aid defenders in their research and development efforts.

For further details or to get in touch, visit us at www.inquest.net.