DARKTRACE

Darktrace AI: Combining Unsupervised and Supervised Machine Learning

Technical White Paper

A New Age of Cyber-Threat

Executive Summary

Darktrace is a research and development-led company that first applied unsupervised machine learning to the challenge of detecting threatening digital activity within corporate networks, with the aim of combatting novel cyber attacks that bypass rule-based security tools. This white paper examines the multiple layers of machine learning that make up Darktrace's Cyber AI, and how they are architected together to create an autonomous, system that self-updates, responding to, but not requiring, human input. This system is today used by over 4,000 organizations globally.

$$\tilde{p}(x_t|y_t) = \sum_{i=1}^{N} w^{(i)} \times \delta\left(x_t^{(i)}\right)$$

A Changing Threat Landscape

The last decade has seen an unmistakable escalation in cyber conflict, as criminals, nation states and lone opportunists take advantage of digitization and connectivity to compromise networks and gain advantage – whether financial, reputational, or strategic.

Our adversaries in cyber space are constantly innovating too, launching new attack tools and technologies to get through traditional cyber security defenses, such as firewalls and signature-based gateways, and into the file shares or accounts that are most valuable to them.

Attackers have also exploited to their advantage the digital complexity of the average organization of today, which shares data across multiple locations, devices and technology services– from cloud services and SaaS tools, to untrusted home networks and non-official IoT devices. Meanwhile, malicious insiders remain a constant threat.

The new age of cyber security is defined by a constantly evolving threat landscape: no sooner have you identified the adversary, than it has shape-shifted into something unrecognizable.

The emergence of these new threats created a technical change: how can we use machine learning to detect what we don't know to anticipate? In other words, without relying on data sets on previous attacks, how can we build a system that learns what a threat is, by comparing its behavior to everything else going in that environment?

This white paper explains Darktrace's approach to machine learning and shines a light on the unique interplay between unsupervised machine learning, supervised machine learning, and deep learning behind the world's leading cyber AI technology.

Traditional Approaches to Cyber Security

According to the traditional security paradigm, firewalls, endpoint security methods, and other tools such as SIEMs and sandboxes are deployed to enforce specific policies and provide protection against known threats. While these tools have a part to play in an organization's overall defense posture, they are ill-equipped to tackle the new age of rapidly evolving cyber-threats. Many have become defunct as enterprise infrastructures diversify.

Fundamental Limitations

Perimeter controls are dependent on signatures, rules and heuristics – if they miss an attack at the point of entry, they have failed and cannot take further action.

Endpoint security also depends on signatures, and only detects attacks that have been previously identified – ineffective against unseen threats.

Log tools and SIEM databases require manual effort to ensure data is consistently collected across the entire organization and matched against the security team's predictions of threats. It relies on the security team imagining everything that might possibly go wrong, without overwhelming analysts with alarms.

So-called 'behavioral analytics' rely on the rules-based paradigm of configuring how certain job titles or devices 'should' behave and then looking for deviations in that behavior. This approach fails to scale to the complexity and size of modern businesses.

Ultimately, legacy systems have been outpaced by modern business complexity and attacker innovation, suffering from these fundamental constraints:

- They need to know about all previous attacks.
- They need to perfectly understand your business and business-specific rules.
- They need a flawless way of sharing high quality information about new attacks.
- They need to guess what all future attacks and software vulnerabilities look like.
- They need to be able to turn all the above insights into rules or signatures that work.

Most significantly, legacy tools require victims before they can provide solutions. The age of unpredictable, fast-moving attacks has rendered this approach woefully deficient.

Applying AI Across Diverse Digital Environments

Darktrace's Cyber AI technology is powered by multiple machine learning approaches, which operate in combination to power the world's first AI platform for cyber defense, working in any digital environment. This allows Darktrace to protect the entire digital estate of the 4000 organizations that it protects, including corporate networks, cloud computing services and SaaS, IoT, Industrial Control Systems, and email systems.

Plugged into the heart of the organization's infrastructure and services, the AI ingests and analyzes the data and its interactions within the environment, and forms an understanding of the normal behavior of that environment, right down to the granular details of specific users and devices, using unsupervised machine learning. A self-learning approach, the system continually revises its understanding about 'what is normal' based on evolving evidence.

This evolving understanding of normal means that the AI can identify, with a high degree of precision, events or behaviors that are anomalous, and unlikely to be benign. The ability to identify highly subtle activity that represents the first footprints of an attacker, without any prior knowledge or intelligence, lies at the heart of the AI's efficacy in keeping pace with today's threat actors. The AI detects what humans cannot, amid the immense noise of legitimate, day-to-day digital interactions.

An Examination of Supervised Machine Learning

Supervised learning works by using previously-classified data, from which the machine learns the classification system. For scenarios where behaviors are well understood and classifications are easy to determine, the output of such systems can be highly accurate.

For example, state-of-the-art image classification systems are outperforming humans in some cases. Indeed, what makes supervised machine learning so powerful is its ability to learn to deal with the errors and noise of the real world, through a statistical approach.

Thus, supervised machine learning systems are best equipped to give you an explicit answer based on prior knowledge. For example, we can feed a system with lots of examples of known ransomware and it will learn the common indicators of that malware and be able to detect similar attacks in the future. However, overfitting is a common problem in supervised machine learning, where model parameters are too finely tuned to the training data.

Instead of learning the essence of a category, the machine learns a particular example – for example, a machine may learn to recognize a German Shepherd, but fail to understand 'dogs' as a category, when distinguishing between 'dogs' and 'cats,' despite recognizing the features that make that German Shepherd pertain to the group.

Supervised Machine Learning and Cyber Security

In the information security context, supervised machine learning is used to train a database of previously seen behaviors, where each behavior is known to be either malicious or benign and is labeled as such.

New activities are then analyzed to see whether they more closely match those in the malicious class, or those in the benign class. Any that are evaluated as being sufficiently likely to be malicious are again flagged as threats.

Systems that rely entirely on supervised machine learning have fundamental constraints:

- Malicious behaviors that deviate sufficiently from those seen before will fail to be classified as such, hence will pass undetected.
- A large amount of human input is needed to label the training data.
- Any mislabeled data or human bias introduced can seriously compromise the ability of the system to correctly classify new activities.

Machine learning has presented a significant opportunity to the cyber security industry. New machine learning methods can vastly improve the accuracy of threat detection thanks to the greater amount of computational analysis they can handle. They are also heralding in a new era of autonomous response, as machine learning systems are sufficiently intelligent to understand how and when to fight back against in-progress threats.

An Examination of Unsupervised Machine Learning

Unsupervised machine learning is critical because, unlike supervised approaches, it does not require labeled training data. Instead it is able to identify key patterns and trends in the data, without the need for human input. Unsupervised learning can therefore take computer processing beyond what programmers already know or can imagine, and discover previously unknown relationships.

Darktrace uses unique unsupervised machine learning algorithms to analyze enterprise data at scale, and make billions of probability-based calculations based on the evidence that it sees.

Instead of relying on knowledge of past threats, it independently classifies data and detects compelling patterns. From this, it forms an understanding of ‘normal’ behaviors across the infrastructure, pertaining to devices, users, or cloud containers and sensors, and detects deviations from this evolving ‘pattern of life’ that may point to a developing threat.

Darktrace Machine Learning: Combining Unsupervised with Supervised

Darktrace technology is powered, at its core, by unsupervised machine learning algorithms that uncover rare and previously-unseen threats that other approaches fail to detect. Over the years, our R&D team in Cambridge, England, has continually developed and deepened the capability of its set of proprietrary technologies, benefiting from the knowledge and experience from the thousands of deployments of Cyber AI across the world.

Today, the Darktrace AI architecture is made up of deep learning techniques that supplement the unsupervised algorithms with expert knowledge from the field.

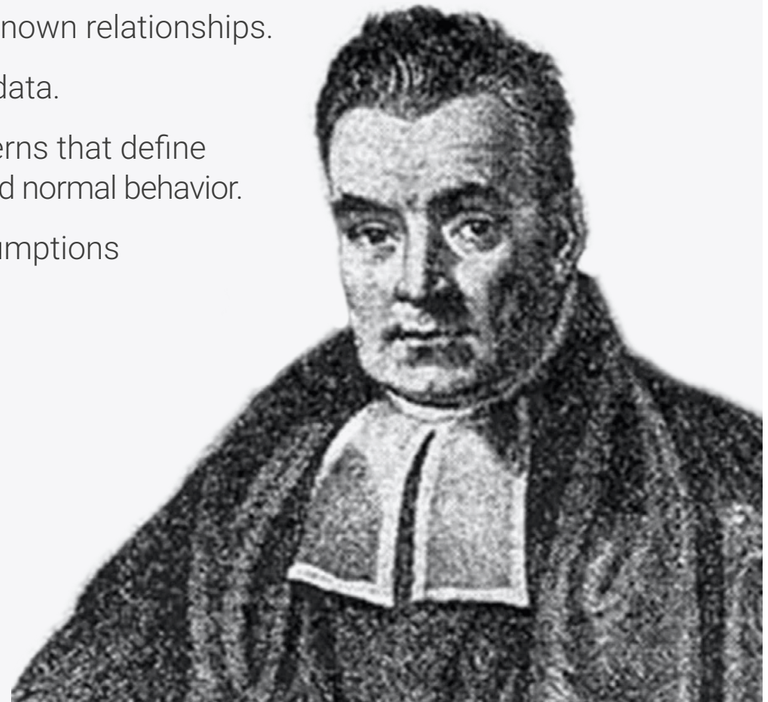
$$\tilde{p}(x_t|y_t) = \sum_{i=1}^{N} w^{(i)} \times \delta(x_t^{(i)})$$

Reverend Thomas Bayes

The mathematics at the forefront of Darktrace's machine learning approach are anchored in the seminal work of British mathematician Thomas Bayes (1702–1761). His theory of conditional probability provides a mathematical bridge between objective, developed methods and the subjective world that we populate. An advanced approach to Bayesian theory, developed by mathematicians from the University of Cambridge, provides a filter to ascertain the true meaning of vague and profuse data.

Darktrace's use of Bayesian probability as part of its unsupervised machine learning approach uniquely enables Darktrace's technology to:

- Discover previously unknown relationships.
- Independently classify data.
- Detect compelling patterns that define what might be considered normal behavior.
- Work without prior assumptions when needed.



Core Principles of Darktrace's Machine Learning:

- Learns 'on the job' – it does not depend upon knowledge of previous attacks.
- Thrives on complexity and diversity of modern businesses.
- Constantly revises assumptions about behavior, using probabilistic mathematics.
- Always up to date, and not reliant on human input.

The impact of Darktrace's unsupervised machine learning on cyber security is transformative. Its Cyber AI technology has quickly proven itself capable of seeing hitherto undiscovered cyber events, from a variety of threat sources, which would otherwise have gone unnoticed.

These include:

- Insider threat – malicious or accidental
- Zero-day attacks – previously unseen, novel exploits
- Latent vulnerabilities
- Machine-speed attacks – ransomware and other automated attacks that propagate and/or mutate very quickly
- Cloud and SaaS-based attacks
- Silent and stealthy attacks
- Advanced spear-phishing

Ranking Threat

Darktrace's AI accounts for ambiguities by distinguishing between the subtly differing levels of evidence that characterize network data. Instead of generating the simple binary outputs 'malicious' or 'benign', Darktrace's mathematical algorithms produce outputs marked with differing degrees of potential threat. This enables users of the system to rank alerts in a rigorous manner, and prioritize those which most urgently require action.

Meanwhile, it avoids the problem of numerous false positives associated with a rule-based approach.

At its core, Darktrace mathematically characterizes what constitutes 'normal' behavior, based on the analysis of a large number of different measures of a device's network behavior, including:

- Server access
- Data volumes
- Timings of events
- Credential use
- Connection type, volume, and directionality
- Directionality of uploads/downloads
- File type
- Admin activity
- Resource and information requests

Clustering Methods

In order to model what should be considered as normal for a device or cloud container, its behavior is analyzed in the context of other similar entities on the network. Darktrace uses unsupervised machine learning to algorithmically identify significant groupings, a task which is impossible to do manually.

To create a holistic image of the relationships within the network, Darktrace employs a number of different clustering methods, including matrix-based clustering, density-based clustering, and hierarchical clustering techniques. The resulting clusters are then used to inform the modeling of the normative behaviors.

$$\tilde{p}(x_t|y_t) = \sum_{i=1}^{N} w^{(i)} \times \delta\left(x_t^{(i)}\right)$$

Modeling Dynamic Environments

A major challenge in modeling the behaviors of a dynamically evolving infrastructure is the huge number of potential predictor variables. For the observation of packet traffic and host activity within an enterprise LAN or WAN, where both input and output can contain many inter related features (protocols, source and destination machines, log changes, and rule triggers), learning a sparse and consistent structured predictive function is crucial.

In this context, Darktrace employs a, cutting-edge large- scale computational approach to understand sparse structure in models of network connectivity based on applying L1- regularization techniques (the lasso method). This allows Darktrace's AI to discover true associations between different elements of a network which can be cast as efficiently solvable convex optimization problems and yield parsimonious models.

> **Within one week of installing Darktrace, the Enterprise Immune System notified us to threats and vulnerabilities we had been totally unaware of.**
>
> - Gabe Cortina, Chief Technology Officer, Bunim/Murray Productions

Recursive Bayesian Estimation

To combine these multiple analyses of digital activity, Darktrace leverages the power of Recursive Bayesian Estimation (RBE). Using RBE, Darktrace's mathematical models are able to constantly adapt to new information as it becomes available to the system.

Continually recalculating threat levels in the light of new data, the Darktrace Immune System can discern significant patterns in data flows indicative of attacks, where conventional signature-based methods see only chaos.

> **As we shifted to the new mode of operation with people being remote, Darktrace very quickly gave us the ability to have the same functionality that we had when everybody was working on campus.**
>
> **- Irving Bruckstein, CIO, Salve Regina University**

Enhancing Detection

Deep learning is a subset of machine learning that uses the cascading interactions of layered mathematical processes – known as neural nets – to give intelligent systems a higher degree of insight. Multi-layered neural nets can improve the detection and remediation of certain threats, for example, in the identification of DNS anomalies, which are less effectively tracked by other machine learning methods. Darktrace's deep learning system assigns a score to all DNS data from a device or digital entity, with the purpose of identifying suspicious activity even faster.

Darktrace clusters devices into peer groups, based on its own understanding of how those devices behave, and uses supervised learning to uncover sequences of breaches, unusual patterns, or to detect aberrant activity at a higher, more holistic level. For example, the well-known WannaCry ransomware was detected by Darktrace as it breaches a number of different 'pattern of life' models. Combining this approach to detection with supervised machine learning, Darktrace can replicate the process of a human interpreting various sets of breaches for a device, network or data environment over time and so present correlated alerts instead of a multitude.

Supervised Machine learning is also used by Darktrace to understand more about the environment, without a human having to label it. By observing millions of different smartphones, for example, Darktrace gets faster and faster at identifying a new device as a 'smartphone', and even what type of smartphone it is.

Using a combination of supervised and unsupervised techniques to complement its core unsupervised machine learning algorithms, Darktrace builds up unique, contextual knowledge about activity and integrates the insights of our global deployments to improve threat detection.

Autonomous Response

Because Darktrace's artificial intelligence is capable of understanding the 'pattern of life' across the entire digital infrastructure at a granular level, detect specific deviations from normal, benign activity, in addition, it can come to autonomous decisions about how to appropriately and proportionately respond to an in-progress attack.

AI response can be triggered by a significant deviation from the derived 'normal' for the device and its peer group, by the detection of specific malicious indicators or unwanted activities, or by a combination of small but meaningful indicators and subtle deviations from expected behavior. This autonomous response technology, known as Antigena, is supported by this unsupervised 'pattern of life' detection, as well as a range of cutting-edge supervised and unsupervised classifiers that measure associations between users, activity patterns and user intents.

Antigena can generate proportionate and targeted surgical responses to deliver these precise actions without disrupting daily business operations. Actions are targeted to the source of the threat, and escalated only when necessary. For instance, Antigena may strip active parts from an email attachment, sever an unusual FTP connection, or block access to Office 365 from an anomalous IP range.

Thanks to the combination of core unsupervised AI and machine learning, this solution can also learn from itself, as well as learn passively from the data that it observes. For example, when Antigena generates an autonomous response action, a feedback reinforcement loop is triggered.

Available across cloud, email, IoT, and on-premise networks, Darktrace Antigena is a crucial part of the data-agnostic Cyber AI Platform that works across an organization's entire digital business. Used in this way, Cyber AI technology does not replace the human's function, but rather serves to enhance it. Darktrace Antigena acts faster than a human, buying the security team precious time to catch up.

> **We no longer live in an era where cyber-attacks are limited to the desktop or server. Darktrace's machine learning fights the battle before it has begun.**
>
> **- Michael Sherwood, CIO, City of Las Vegas**

Automating Threat Investigation Processes

Finally, Darktrace also uses various machine learning techniques to automate repetitive and time-consuming tasks carried out during investigation workflows. By analyzing how expert cyber analysts interact with the AI's output, for example how they triage threat alerts and how they use third-party sources, Darktrace is able to replicate those expert behaviors and automate certain analyst functions. This allows for increasingly efficient and simplified investigations for analysts of all maturity levels. It also gives security teams the crucial time they need to focus on higher-value strategic work, such as managing risk and focusing on broader improvements to the business.

Darktrace leverages supervised learning in a capability that mimics the way a human carries out the threat investigation process, in the form of a capability known as the Cyber AI Analyst. In the initial stages of the investigation process, the technology will make broad hypotheses about what is happening, and then will query and analyze this information as a human would – using custom algorithms and other machine learning techniques.

Once the investigation has been launched, the results of these can be classified using supervised machine learning to determine incidents of interest, at a speed and scale only possible with AI.

The Cyber AI Analyst can often detect details that a human might miss, or might not have time to identify, and can decide whether or not an initial hypothesis holds in a matter of minutes. More crucially, the technology can classify and store the results of these investigations, allowing for only a small number of high priority incidents to be presented at any one time.

It communicates its findings and recommendations immediately in the user interface, and additionally as detailed PDF reports, which present only a few high-priority incidents at any one time in natural language. These are then enriched with context and security insights that can be reviewed and understood by executives and end-users alike.

Crucially, the Cyber AI Analyst is able to adapt to new and unprecedented situations on the fly, enabling users to spend less time trawling through alerts, and more time prioritizing the strategic work that matters.

> **"It's mind-bending that Darktrace has been able to do this. Having the Cyber AI Analyst stitching together multiple security alerts at once helps us get to the high-value work quickly."**
>
> **- Phillip Miller, CISO, Brooks Brothers**

Conclusion

Our generation is witnessing the machine learning revolution. We are seeing shifts in working practices brought about by the replacement of muscle with machine, the automation of repetitive tasks, and now the replacement of low value, thoughtful tasks with machines capable of handling big data and making vast calculations.

As networks have grown in scope and complexity, the opportunities for attackers to exploit the gaps have increased. Walls are no longer enough to protect the corporate networks spilling into home environments, and rules-based tools cannot keep up with all possible attack vectors, and cannot respond fast enough if a machine-speed attack hits. A constantly evolving cyber-attack landscape requires a step up in our detection capability, using machine learning to understand the environment, filter the noise and take action where threats are identified.

Darktrace's technology has become a vital tool for security teams attempting to understand the scale of their network, observe levels of activity, and detect areas of potential weakness. Machine learning technology is the fundamental ally in the defense of systems from the hackers and insider threats of today, and in formulating response to unknown methods of cyber-attack. It is a momentous change in cyber security.

> **Darktrace is one of the few in the threat analysis space doing it right.**
>
> **- Alissa Knight, Senior Analyst, Aite group**