THREAT INTELLIGENCE WHITE PAPER

Machine Learning in Cyber Security: Age of the Centaurs

Staffan Truvé, PhD, Chief Technology Officer and Co-Founder, Recorded Future

Recorded Future

brookcourt
Collaborative Solutions

Artificial intelligence (AI), and in particular machine learning, has taken huge strides and is now set to really start impacting all aspects of industry and society. This development has been fueled by decades of exponential improvement in raw computing power, combined with progress in algorithms and, perhaps most importantly, a huge increase in the volume of data for training and testing that is readily available on the internet. The combination of these three factors is now giving us everything from voice-controlled digital assistants to autonomous cars. It it safe to say that "this changes everything," and cyber security is no exception.

AI and Machine Learning: What's the Difference?

As these concepts move into mainstream consciousness it's inevitable that there will be confusion in defining these technologies and what they do. Big data and analytics expert Bernard Marr endeavours to broadly define both AI and machine learning:

"Artificial Intelligence is the broader concept of machines being able to carry out tasks in a way that we would consider 'smart.' and, machine Learning is a current application of AI based around the idea that we should really just be able to give machines access to data and let them learn for themselves."[1]

Systems based on AI, sometimes referred to as cognitive systems, are helping us automate many tasks which until recently were seen as requiring human intelligence. However, AI allows us to not only automate and scale up tasks that so far have required humans, but also lets us tackle problems which are more complex than what most humans are capable of solving. Even so, many tasks are best performed by a machine and a human working together.

Centaurs in greek mythology are creatures with the upper body of a human and the lower body of a horse. In computer chess, a "centaur" is a human and computer playing together as a team, to take advantage of their complementary strengths:[2] the speed and storage capacity of the machine and the creativity and strategic eye of the human. For two decades following the landmark defeat of Garry Kasparov, the reigning world champion in chess, by IBM's Deep Blue computer in 1996,[3] the combination of human and computer has created the strongest chess players of them all, far surpassing top human players and beating even the best chess computers.

Similarly, we believe *threat analyst centaurs* — man and machine working together — will be teams capable of tackling the most difficult cyber adversaries.

Some claim that computers have become fast enough that the efficiency afforded by human collaboration on a centaur team may no longer make a difference in chess, but the space of possible actions for threat actors is far more expansive than a chess game and there is little tolerance for unexpected blind spots. Also, unlike in chess,[4] the counterparts in cyber security do not follow any rules.

At Recorded Future we look to the centaur model to create the best possible threat analysts, combining the speed and depth of AI with the strategic vision of a human expert.

What Is Artificial Intelligence, Really?

Webster's Dictionary defines artificial intelligence as "an area of computer science that deals with giving machines the ability to seem like they have human intelligence," and even if that definition is fairly vague it actually does capture the difficulty in defining AI quite nicely.

[1] https://www.forbes.com/sites/bernardmarr/2016/12/06/what-is-the-difference-between-artificial-intelligence-and-machine-learning
[2] https://en.wikipedia.org/wiki/Advanced_Chess
[3] https://en.wikipedia.org/wiki/Deep_Blue_versus_Kasparov,_1996,_Game_1
[4] http://www.businessinsider.com/computers-beating-humans-at-advanced-chess-2013-11

AI is now being applied in a variety of problem domains, such as natural language processing, robotic navigation, computer vision, etc., and relies on a number of underlying technologies such as rule-based systems, logic, neural networks, and statistical methods such as machine learning. Like in other areas, there is a lot of fashion in what techniques are preferred, as exemplified by the recent hype around deep learning.[5] Our experience is that a mix of different techniques is needed to tackle a complex problem domain.

Deep Learning

Deep learning is the study of artificial neural networks and related machine learning algorithms that contain more than one hidden layer. These deep nets:

- Use a cascade of many layers of nonlinear processing units for feature extraction and transformation. Each successive layer uses the output from the previous layer as input. The algorithms may be supervised or unsupervised and applications include pattern analysis (unsupervised) and classification (supervised).
- Are based on the (unsupervised) learning of multiple levels of features or representations of the data. Higher level features are derived from lower level features to form a hierarchical representation.
- Are part of the broader machine learning field of learning representations of data.
- Learn multiple levels of representations that correspond to different levels of abstraction; the levels form a hierarchy of concepts.

Deep learning is part of a broader family of machine learning methods based on learning representations of data. One of the promises of deep learning is replacing handcrafted features with efficient algorithms for unsupervised or semi-supervised feature learning and hierarchical feature extraction.

In the end, like for all other computer systems, it all boils down to two things: data structures and algorithms. Whether what you build using those components is "AI" or not depends on if a human observer believes the system behaves "intelligently" in its target application domain.

Unlike older AI systems that primarily performed one task using one AI technology (such as Deep Blue playing chess or MYCIN and Dendral[6] giving expert advice in a very narrow medical domain), today's systems must use AI technology in many different places, to automate or facilitate the tasks at hand. A successful use of AI technology today will to a large extent be invisible, like in Google Photos where images are categorized, augmented, and merged into stories using technologies the end user does not and need not be aware of. Another example is Siri, which does speech recognition, but also uses other underlying systems, like Wolfram|Alpha, to answer questions in different domains.

It is also worth emphasizing that building an AI-based product is in almost all cases a systems engineering challenge, requiring not only a few clever algorithms but also a massive investment in supporting technologies like scalable computing infrastructure, monitoring systems, quality control, and data curation. These more mundane aspects may not be immediately visible to an end user, but they are essential for a working solution.

[5] http://karpathy.github.io/2015/05/21/rnn-effectiveness/
[6] http://artificialintelligence-notes.blogspot.se/2010/07/expert-systems-dendralmycin.html

Why Now?

AI has become such a focal point of attention for both researchers and entrepreneurs during the last few years due to several factors contributing to a "perfect storm":

› Never before has so much information been available in digital form, ready for use. All of humanity is, on a daily basis, providing more information about the world for machines to analyze. Not only that — through crowdsourcing and online communities we are also able to give feedback on the quality of the machines' work on an unprecedented scale.

› Computing power and storage capacity continue to grow exponentially, and the cost for accessing these resources in the cloud are decreasing. Incredible resources are now available not only to the world's largest corporations but to garage startups as well.

› Research in algorithms has taken huge strides in giving us the ability to use these new computing resources on the massive data sets now available.

Singularity

The technological singularity (also, simply, the singularity) is the hypothesis that the invention of artificial superintelligence will abruptly trigger runaway technological growth, resulting in unfathomable changes to human civilization. According to this hypothesis, an upgradable intelligent agent (such as a computer running software-based artificial general intelligence) would enter a 'runaway reaction' of self-improvement cycles, with each new and more intelligent generation appearing more and more rapidly, causing an intelligence explosion and resulting in a powerful super intelligence that would, qualitatively, far surpass all human intelligence. John von Neumann first uses the term "singularity" (c. 1950s), in the context of technological progress causing accelerating change: "The accelerating progress of technology and changes in the mode of human life, give the appearance of approaching some essential singularity in the history of the race beyond which human affairs, as we know them, can not continue".

Even though we have not yet reached the technological singularity,[7] we are living in a time when the limit of what machines can accomplish is being pushed every day. With these rapid advancements come perhaps understandable fears surrounding the continuing viability of the human race. As this acceleration continues we see a world where the centaur model is amplified and applied to numerous domains which ultimately results in continuing benefits to humanity as a whole.

AI at Recorded Future

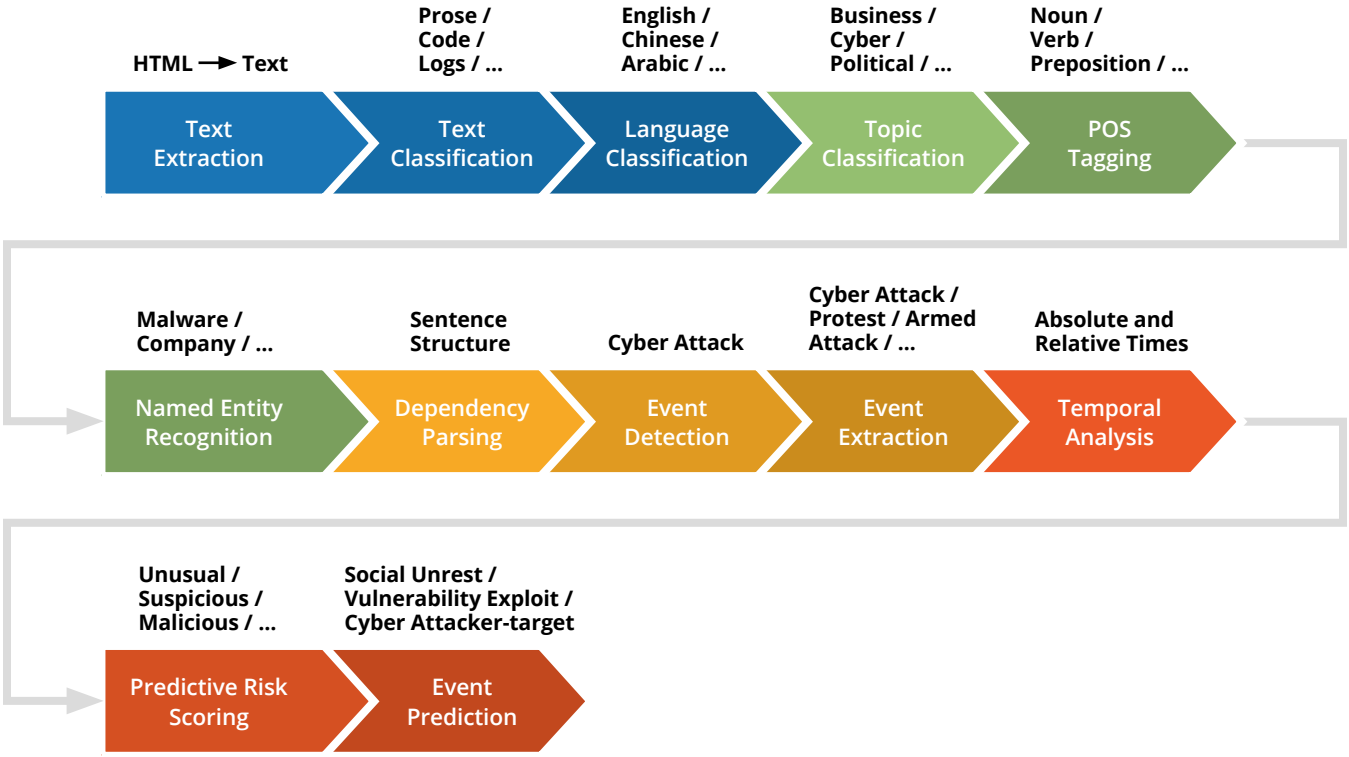
Recorded Future is driven to provide the most relevant threat intelligence from the greatest breadth of available sources. This mission requires us to use the latest technologies for acquiring, aggregating, analyzing, and presenting threat intelligence. We use AI for all of these tasks, both to automate routine work and to solve extremely complex problems like prediction of future threats.

[7] https://en.wikipedia.org/wiki/The_Singularity_Is_Near

We use AI techniques to:

1. Represent structured knowledge of the world, using ontologies[8] and entities plus events — as described below.
2. Transform unstructured text in multiple languages[9] into a language-independent, structured representation, using natural language processing.
3. Classify events and entities, primarily to help decide if they are important enough to require a human analyst to perform a deeper investigation.
4. Forecast events and entity properties by building predictive models from historic data.

We use a combination of rule-based, statistical, and machine-learning techniques to deliver these capabilities, as described below. Our goal is both to automate and scale up tedious and almost trivial human tasks, and to enable more complex analytics (e.g., for predictions).



The Recorded Future processing pipeline uses machine learning and rule-based algorithms to transform unstructured information from the web into actionable threat intelligence.

Knowledge Representation

At the heart of Recorded Future is a structured representation of the world, separated into two parts: ontologies and events.

Our ontologies are used to represent entities such as persons, organizations, places, technologies, and products. Specifically for the cyber security domain, we detect domain-specific entities such as malware, malware categories, vulnerabilities, and technical indicators (hashes, file names, IP addresses, domain names, etc.). The ontologies also contain information about

[8] Ontologies are formal representations of a set of concepts/entities within a domain and the relationships between those concepts/entities.
[9] Currently we do deep linguistic analysis in English, French, Spanish, German, Russian, Arabic, Farsi, and Chinese — with more to follow.

relationships between these entities, such as hierarchies ("X is part of Y" — e.g., the entity "Paris" is a City in a Country called "France", and "Zeus" is a Malware and a member of Categories "Botnet" and "Banking Trojan"). These ontologies provide, among other things, a powerful way of searching over categories, like "find all 'Heads of State' who traveled to 'Africa' in 2015," where "Heads of State" will mean all Person entities with that attribute, and "Africa" captures all geographic entities (Country, City, etc.) on that Continent.

Recorded Future events are used to represent real-world events in a language-independent, structured way. They range from person- and corporate-related to geopolitical, environmental, and cyber-related events. Event detectors abstract away from the exact wording used to describe an event. For example, "John flew to Paris," "John visited Paris," "John took a trip to Paris," "Джон прилетел в Париж," and "John a visité Paris" are all different ways of expressing the same event: a Person Travel event where "John" is the traveler and "Paris" is the destination. This is the core idea behind events: By being able to search for an event instead of just using keywords, the analyst can focus on abstract concepts and not the many ways and different languages in which an author might talk about them.

Each event type has a set of (sometimes optional) named attributes. For example, a Cyber Attack event relates an attacker to a target, and includes additional information about the attack method used, and related hacktivist operation hashtags. At least one attacker or one target must be specified, the rest is optional. Multiple mentions of an event are grouped together to simplify analysis, even if the original text is in different languages or uses different words for the attack.

Natural Language Processing

Natural language processing (NLP) transforms an unstructured, natural language text into a structured, language-independent representation. In our system, this means identifying entities, events, and time associated with those events. There are several steps to this using different AI techniques:

- To extract the relevant text from an HTML web page, we have developed a machine-learning-based module using Gibbs Sampling[10] to extract the actual content (e.g., to decide what text should be ignored, such as advertising).
- We use supervised machine-learning algorithms to classify texts, such as determining in which language a text is written or if a text is prose, a data log, or programming code.
- Supervised machine learning (e.g., using Conditional Random Fields[11]) is also used to do named entity extraction from text.
- We are also using machine-learning based classifiers to automatically disambiguate between different entities that have the same name (e.g., "Zeus" the Greek god vs. "Zeus" the malware) based on the context in which they are mentioned.
- We use a data-driven dependency parser, MaltParser, to analyze the structure of sentences.[12] This is an implementation of inductive dependency parsing, where the syntactic analysis of a sentence amounts to the derivation of a dependency structure, and where inductive machine learning is used to guide the parser.
- We have developed two distinct proprietary rule-based systems for extracting events and temporal information.

The combination of statistical/machine-learning and rule-based components has allowed us to build a system that maximizes precision and recall. In the future we foresee more components being based on machine learning, but the significant costs associated with producing annotated data needed for training such components has so far motivated the hybrid approach.

The use of NLP has allowed us to build a system capable of analyzing millions of documents per day, in eight different languages (English, French, Spanish, Russian, Farsi, Arabic, German, and Chinese), and to transform that data into a

[10] https://en.wikipedia.org/wiki/Gibbs_sampling
[11] https://en.wikipedia.org/wiki/Conditional_random_field
[12] http://www.maltparser.org/intro.html

representation that gives analysts insight, independent of language skills. This use of AI thus addresses two of the major challenges an analyst faces: the need to find information written in several languages, and the capacity to read and organize the massive amounts of security-related information being published every day.

Event and Entity Classification

The third area where AI techniques are used is for classification of entities and events. We classify the importance of events to help analysts prioritize what they should focus their attention. We also classify the maliciousness of technical entities (e.g. IP addresses and internet domain names) to support analysts and to enable the automatic configuration of network equipment and network management systems.

Event classification is done using statistical methods to detect anomalies (e.g., an unusual number of event references related to a certain cyber attacker or target).

Entity classification is used to decide the maliciousness of an entity — an IP address, for example — and to assign a risk score that can be used by an analyst or an automated system to decide how to act. Risk scores are assigned using two different systems, one rule based and one machine-learning based. The rule-based system is based on human intuition about which contexts, sources, co-occurrences, threat list mentions, etc. are useful in deciding if an entity is associated with some kind of risk. The machine-learning-based classifier, on the other hand, has been trained on a large data set, using trusted threat list sources as ground truth for what constitutes a malicious entity.

One important aspect of both event and entity classifiers is that they must provide not only a judgement ("this event is critical" or "this IP address is malicious"), but also a human-readable motivation for that judgement ("this IP address is considered malicious because it has been called out as a malware command and control center by three independent sources").

Rule-based systems can fairly easily generate motivations like these, but it is equally important in machine-learning systems to be able to generate motivations, at least in terms of which significant features of an entity contributed the most to a judgement stating that it has a certain property. We are working to derive similar explanations from our machine-learning-based components.

Automated classification of entities and events allows our system to support the analyst, who can spend significantly less time on deciding what topics to focus on, and instead use that time for improved analysis of prioritized threats.

Predictive Analytics

Cyber defenders today are almost always one step behind, trying to patch systems and configure protection mechanisms against known attacks and existing breaches. With predictive information, defenders might instead start being proactive, and protect their systems against future threats. We believe that, in specific domains, predictive threat intelligence is derivable from historic and current data.

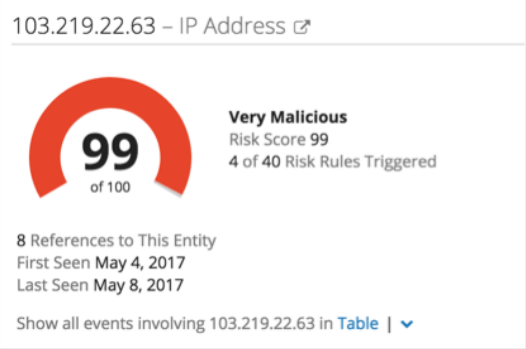
We use machine learning to generate predictive models that can be used to forecast events or classify entities. We have, for example, created models to predict future risk of social unrest, the likelihood of product vulnerabilities being exploited, and to assess the risk that an IP address will behave maliciously in the future, even though no such activity has yet been observed.

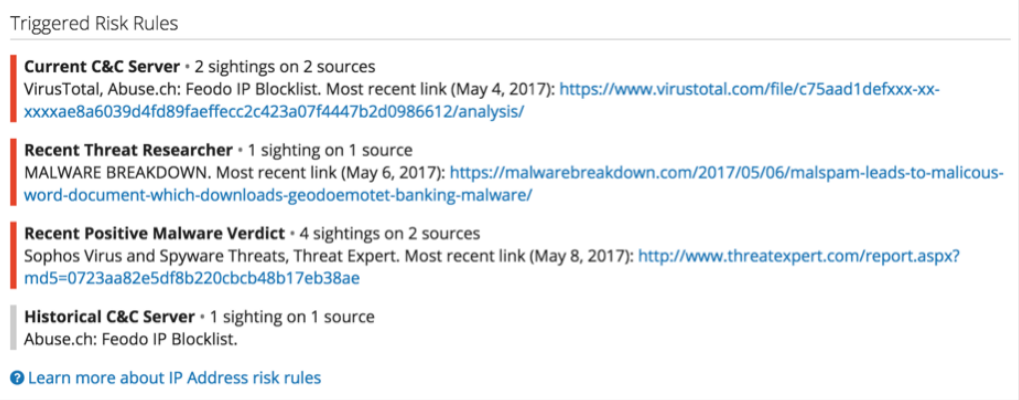
The challenge in all these cases is to identify relevant features on which to base the predictions, and most of all to get access to enough ground truth training data to be able to generate models that can be used to make predictions with the required accuracy.

Prediction generation is an example of a task that is hard or even impossible for a human analyst to carry out, due to the complexity and large volume of data needed. Algorithms and machines scale much better to some problems of this kind. Below, we illustrate predictive analytics with a concrete example.

Predictive IP Risk Scoring

As previously mentioned, Recorded Future assigns risk scores to entities such as IP addresses and vulnerabilities. These scores are set using a rule based system, and are derived from historic observations around an entity (e.g., which sources it is being mentioned in, its presence on threat lists, occurrence together with known threat actors and malware, etc.)





Risk scores have a practical use in cyber security — a security operations center (SOC) operator can quickly assess and possibly block an IP address on a network, for example — but the big limitation is of course that they are based on historic data, and thus cannot be used until something has happened, either locally or somewhere else.

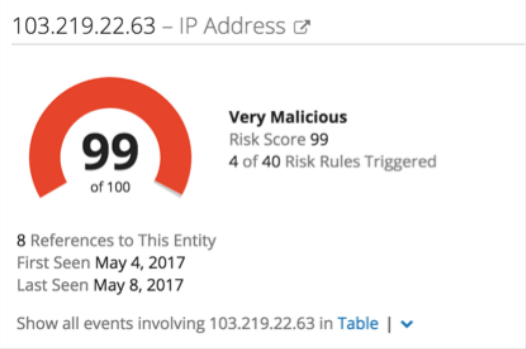
To further assist threat analysts and SOC operators we've developed predictive risk scores. These scores are produced using a machine-learning model, which is trained on historic information from both threat lists and open source information, and can assign a predictive risk score to a hitherto unseen IP address. The predictive risk scores are based on historic and current risk scores for neighbouring or related addresses, and the ways in which these are being mentioned in open source discussions.

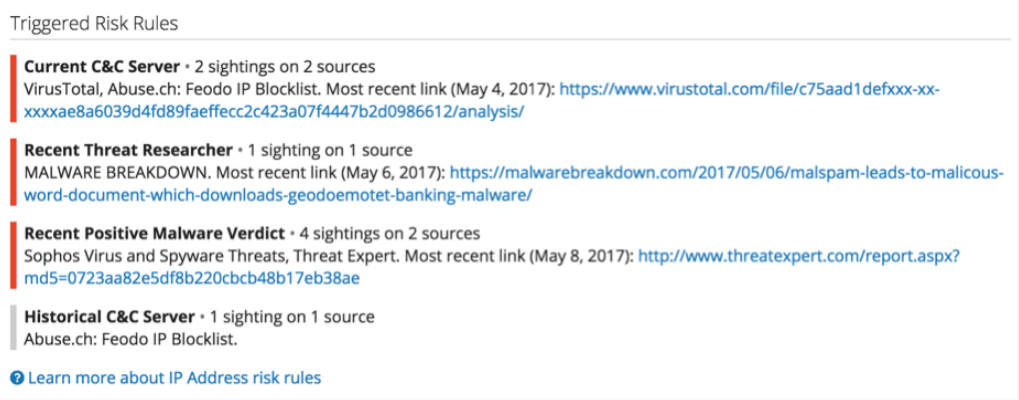
Predictive IP risk scoring has proven to be very valuable, and we are now applying similar methods for predictive scoring (e.g., internet domain names) and to identify likely new targets of cyber attacks.

Prediction generation is an example of a task that is hard or even impossible for a human analyst to carry out, due to the complexity and large volume of data needed. Algorithms and machines scale much better to some problems of this kind. Below, we illustrate predictive analytics with a concrete example.

Predictive IP Risk Scoring

As previously mentioned, Recorded Future assigns risk scores to entities such as IP addresses and vulnerabilities. These scores are set using a rule based system, and are derived from historic observations around an entity (e.g., which sources it is being mentioned in, its presence on threat lists, occurrence together with known threat actors and malware, etc.)





Risk scores have a practical use in cyber security — a security operations center (SOC) operator can quickly assess and possibly block an IP address on a network, for example — but the big limitation is of course that they are based on historic data, and thus cannot be used until something has happened, either locally or somewhere else.

To further assist threat analysts and SOC operators we've developed predictive risk scores. These scores are produced using a machine-learning model, which is trained on historic information from both threat lists and open source information, and can assign a predictive risk score to a hitherto unseen IP address. The predictive risk scores are based on historic and current risk scores for neighbouring or related addresses, and the ways in which these are being mentioned in open source discussions.

Predictive IP risk scoring has proven to be very valuable, and we are now applying similar methods for predictive scoring (e.g., internet domain names) and to identify likely new targets of cyber attacks.

What Does the Future Hold?

As can be seen from the description above, Recorded Future uses AI techniques in multiple parts of its system to create functionality that mimics or even extends human intelligence. This enables analysts to work together with the machines, creating extremely capable centaurs.

We have come some way in using AI techniques to improve the capabilities of human analysts, but there is much more to be done in terms of automation. Our next step is to include analysis of remediation/action advice mentioned together with a vulnerability or malware, for example, allowing our system to recommend an action to mitigate the threat. This will require our system to both understand a customer's operating environment and the discussions and advice about how to handle threats and fix vulnerabilities, including possible dependencies between both different fixes and product versions.

We foresee a future where the collaboration between man and machine will continue to grow, and where new tasks will be automated, freeing up time for the human analysts to focus on what they are best at.

About Recorded Future

Recorded Future delivers threat intelligence powered by machine learning, arming you to significantly lower risk. We enable you to connect the dots to rapidly reveal unknown threats before they impact your business, and empower you to respond to security alerts 10 times faster. Our patented technology automatically collects and analyzes intelligence from technical, open, and dark web sources to deliver radically more context than ever before, updates in real time so intelligence stays relevant, and packages information ready for human analysis or instant integration with your existing security systems.

@RecordedFuture | www.recordedfuture.com

About Brookcourt Solutions Ltd

Brookcourt Solutions Ltd is an award-winning organisation; committed to providing independently sourced, leading edge protective technology solutions. Working together to ensure your organisation stays safe, confident and empowered in today's challenging Cyber Security and networking environment.

We're not just a reseller, we approach every client strategically – understanding the specific business landscape, infrastructure and associated threats.

We are an adept team – we won't throw bodies at a problem – we provide the best people for the job. Our analytical and independent approach to your specific circumstances means we deliver carefully throughout, purpose built solutions to ensure your business is protected and your network is consistent, intelligent and efficient.

@TweetBrookcourt | www.brookcourtsolutions.com