Recorded Future

WHITE PAPER

4 Ways Machine Learning Is Powering Smarter Threat Intelligence

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

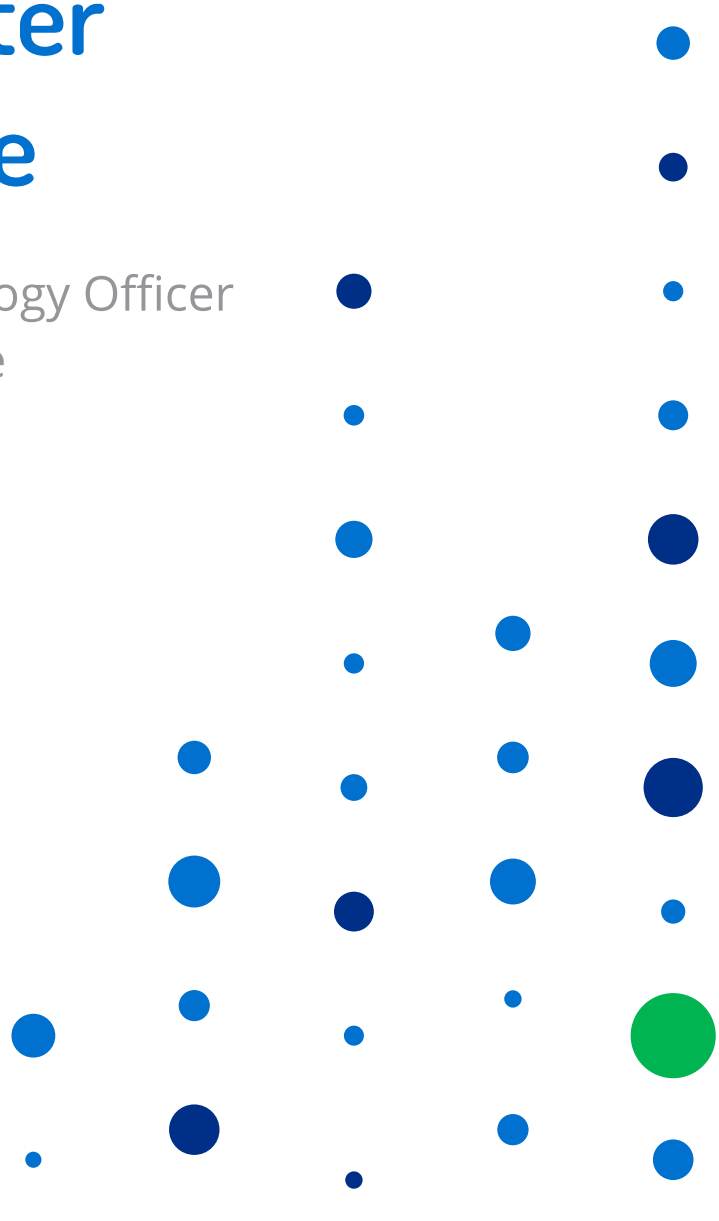
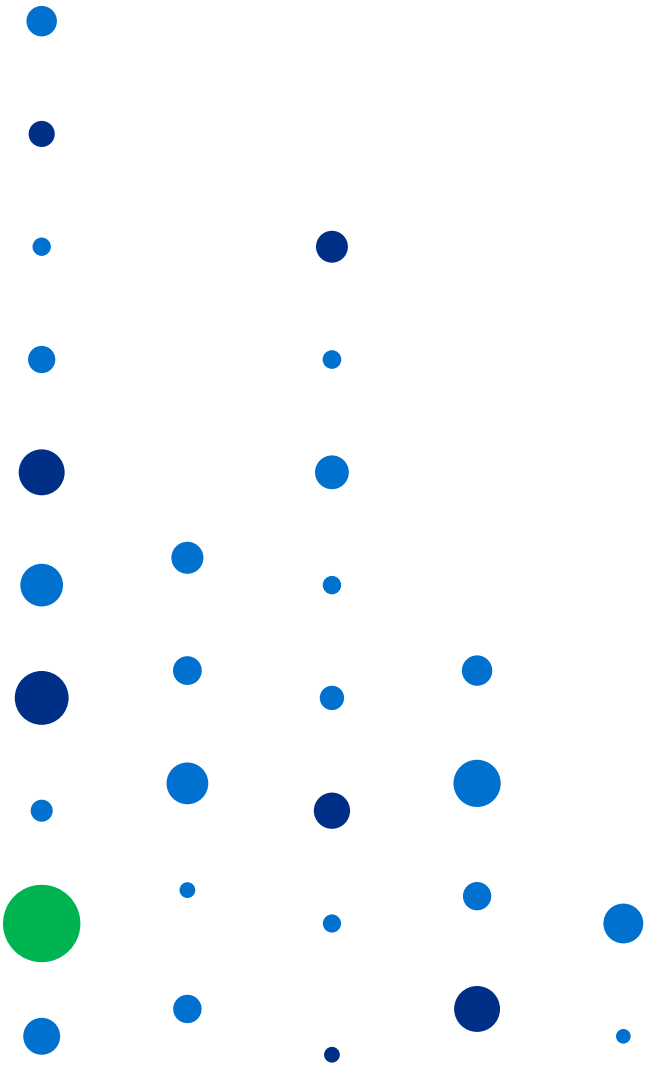


Table of Contents



Recorded Future

Introduction

Artificial intelligence (AI), and in particular machine learning, has seen huge strides in recent years and is now set to really start impacting all aspects of society and business. 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 machines 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 cybersecurity is no exception.

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 though that definition is fairly vague, it actually does effectively capture the difficulty in grasping what AI really means.

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 most humans are capable of solving.

AI is now being applied in a variety of problem domains, such as natural language processing, robotic planning and navigation, computer vision, etc., and relies on a number of underlying technologies such as rule-based systems, logic, neural networks, and statistical methods like machine learning. As in other areas, there is a lot of fashion in what techniques are preferred, as exemplified by the recent hype around deep learning.[1] Our experience is that a mix of different techniques is needed to tackle a complex problem domain.

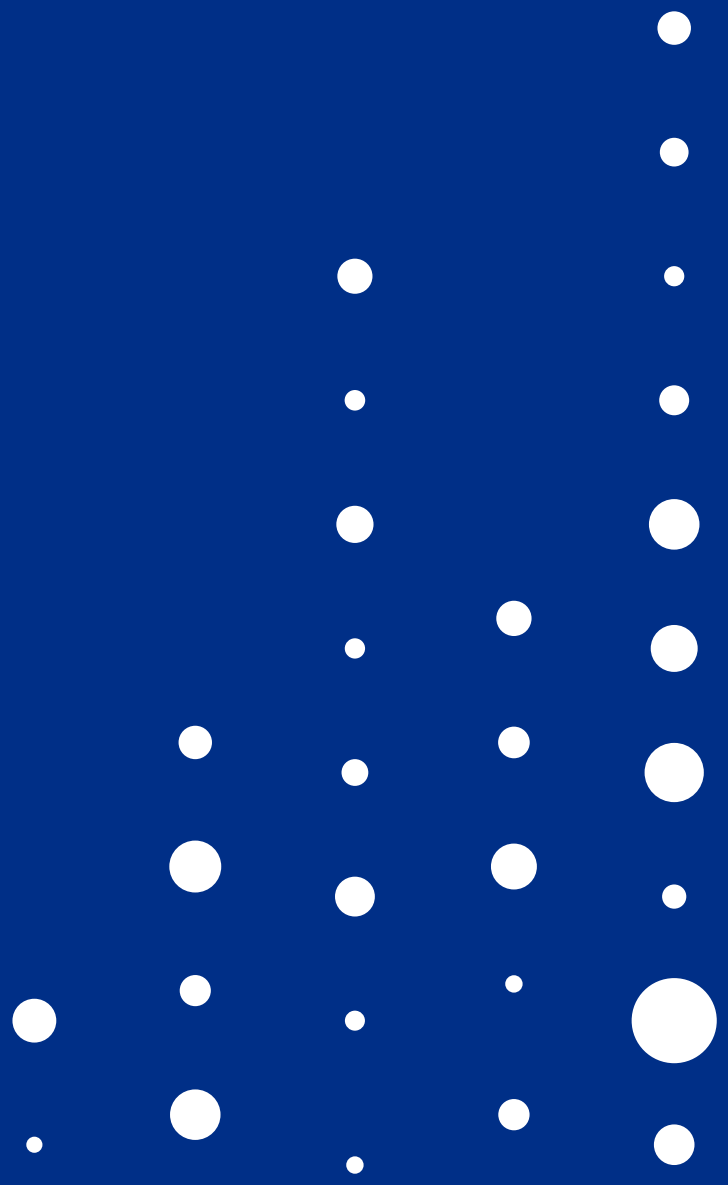
In the end, like for all other computer systems, implementing these techniques in the real world boils down to two things: data structures and algorithms. Whether what you build using those components is "AI" or not depends on if a human observer believes the system behaves "intelligently" wherever it's applied.

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.

[1] http://karpathy.github.io/2015/05/21/rnn-effectiveness/

“In recent years, humanity’s ability to make accurate predictions has improved in many fields thanks to a combination of augmented sensor capabilities and new prediction algorithms. As an example, today’s weather forecasts benefit both from improved sensing by weather satellites and from new algorithms run on powerful parallel computers. In a similar way, applying novel AI techniques to threat intelligence provides new sensing capabilities which work at scale and can be applied to new domains like predicting future cyber threats.”

Staffan Truvé, CTO and Co-Founder
Recorded Future



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 at 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 come a long way in giving us the ability to use these new computing resources on the massive data sets now available.

Why AI and Machine Learning for Threat Intelligence?

The field of threat intelligence presents unique challenges and opportunities for AI systems. The intelligence process deals with gathering, analyzing, and presenting a variety of statistical and narrative data. This means that to make available information useful and actionable, the machinery must be able to:

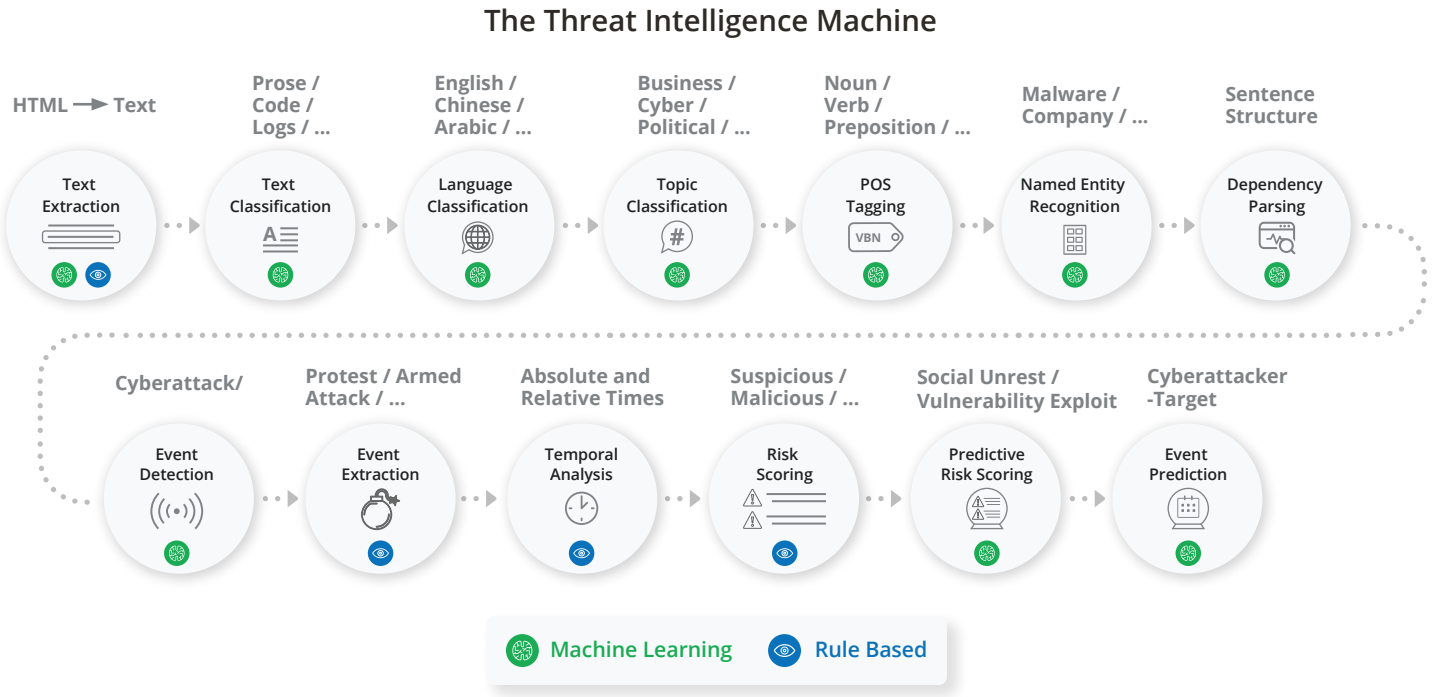
- Deal with the complexity of the available data.
- Be able to automate high volumes of the available data.

Recorded Future is using a combination of rule-based, statistical, and machine-learning techniques to meet these challenges. Our goal is both to automate and scale up tedious and almost trivial human tasks in threat intelligence analysis, as well as apply more complex analytics to challenges, like predicting future cyber threats.

This white paper will demonstrate a number of ways in which Recorded Future uses AI, and machine learning in particular, to analyze data, structure this information at scale, and present threat intelligence for sharing with humans or security systems:

1. **Natural Language Processing:** Transform unstructured text in multiple languages into a structured representation.[2]
2. **The Threat Graph:** Represent structured knowledge of the world, showing connections between things, people, places, and time.
3. **Malware Entity Recognition:** Identify that new words in the right context are names of malware.
4. **Predictive Analytics:** Forecast events and entity properties by building predictive models from historic data.

[2] Currently, we do deep linguistic analysis in English, French, Spanish, German, Russian, Arabic, Farsi, and Chinese, and are continuously adding new languages.



The processing pipeline of Recorded Future's Threat Intelligence Machine uses machine learning and rule-based algorithms to transform unstructured information from open, technical, and dark web sources into actionable threat intelligence.

1. Natural Language Processing

Natural language processing transforms unstructured, natural language text into a structured, language-independent representation. In our system, this means identifying entities and events, and time associated with those events to make available information more easily understandable to a human. Once these entities and events are put into context, the machine can then structure this information and reveal connections to more technical threat indicators like IP addresses, hashes, and domains.

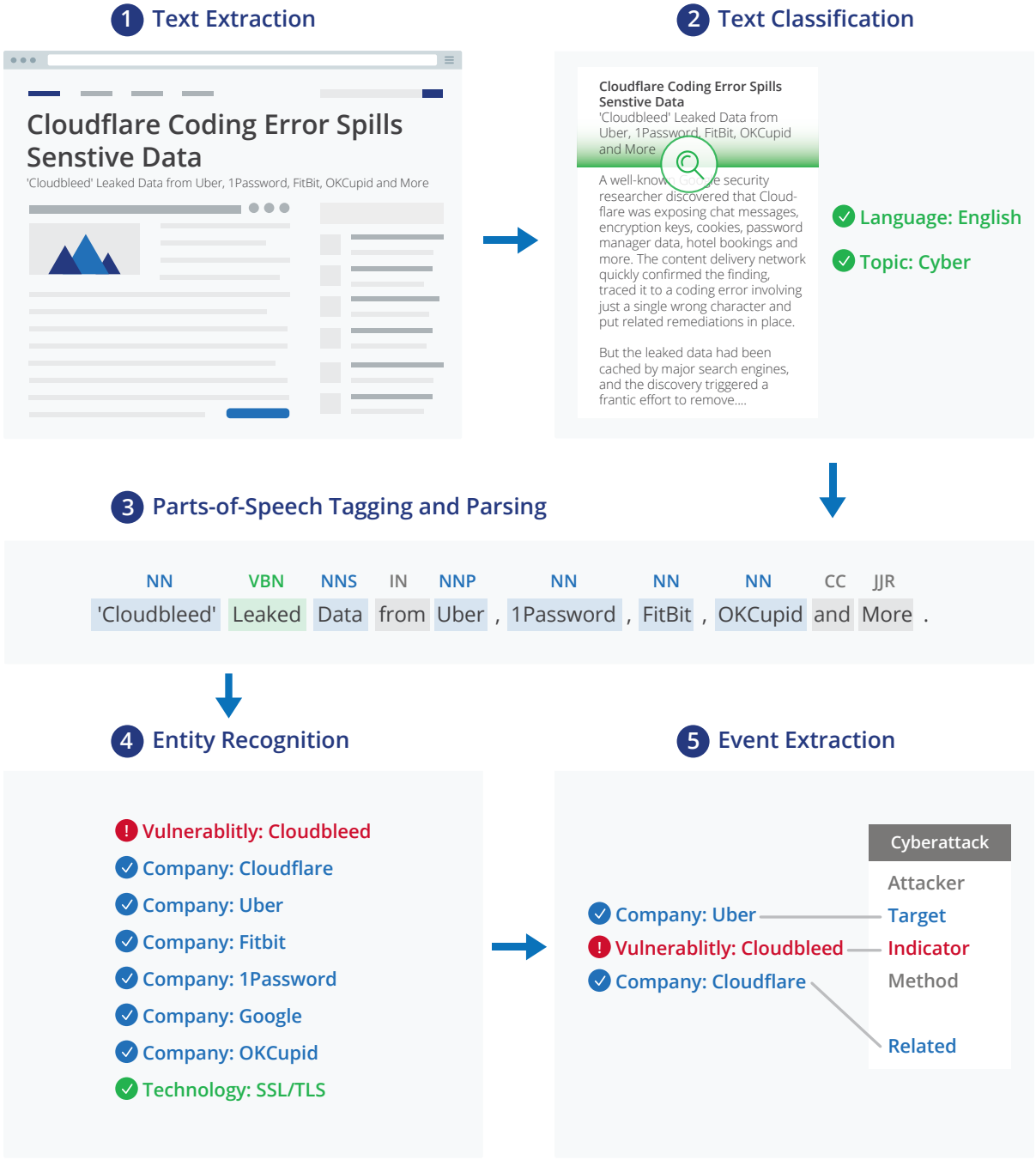
The graphic below illustrates the phases of natural language processing inside Recorded Future. We've developed a machine-learning module that initially determines which text is relevant and what should be ignored, stripping away advertising or links to other unrelated content.[3]

Next, the machine identifies the nature of the text; it has learned how narrative text is constructed differently to programming code or data logs, so it is recognizing nouns, verbs, adjectives, etc. After this, we use supervised machine learning to extract entities from the text to reveal if this sentence is, for example, about a particular company, industry, or technology.

As part of our natural language processing, the machine will also automatically disambiguate between different entities with the same name.

[3] https://en.wikipedia.org/wiki/Gibbs_sampling

For example, where the word “Zeus” appears, the machine can see the context of that word to classify it either as “Zeus” the Greek god, or “Zeus” the malware.



In the Real World

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

Obviously, humans can read and understand text, but machines can do this at massive scale to process the huge amounts of available threat data.

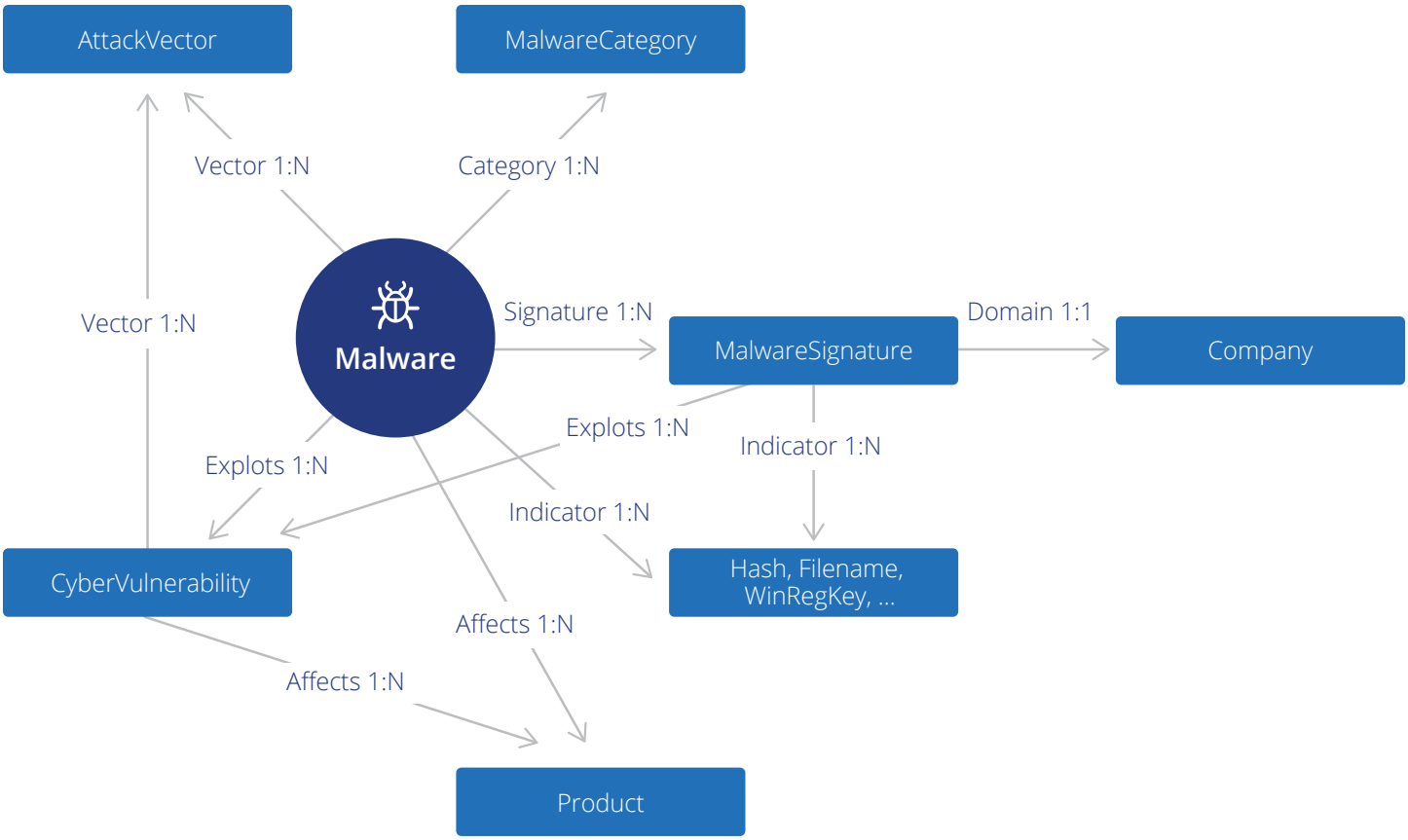
The power of applying machines to the job of collecting and structuring text massively scales human capacity and speeds up the human process identifying contextualized threat intelligence.

2. The Threat Graph

At the heart of Recorded Future is a structured representation of the world that we refer to as the Threat Graph. The graph identifies and structures more than a hundred **entity** types like "Malware," "Vulnerability," and "Threat Actor." The graph contains information about the relationships between these entities, such as hierarchies. For example, "Zeus" is a malware and a member of the categories "Botnet" and "Banking Trojan."

This example shows the Threat Graph for a particular malware, using available intelligence to create relationships between related vulnerabilities, the malware category, specific threat actors, and any indicators of compromise like malicious hashes, IP addresses, or domains.

This same methodology is applied to the target entities like businesses and organizations — providing context to the relationships between brands, domains, key personnel, identification numbers, and more.



Our Threat Intelligence Machine can also identify **events** that affect **entities** in the Threat Graph. In Recorded Future we represent real-world events in a language-independent, structured way. These range from people and corporate, to geopolitical, environmental, and cyber-related events. Event detectors help us to classify an event even if the wording used to describe it are different. 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" and "Travel" event where "John" is the traveler, and "Paris" is the destination.

Each event type has a set of named attributes. For example, a "Cyberattack" event relates an "Attacker" to a "Target," and could include additional information about the attack method used, as well as any 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 to describe the attack.

In the Real World

Machine learning involves connecting the dots in the data to add context to vulnerabilities, attack methods, and targets. This ongoing processing and analysis of these entities, ontologies, and events removes a significant burden from security analysts and allows them to get an understanding of cyber events much more quickly.

The Threat Graph also provides a powerful way of searching over categories. For example, imagine searching for "All cyberattacks against finance in 2015." The machine knows that a "cyberattack" will mean events that contain references to cyber events that include an attacker and a target. "finance" captures all company entities classified as "Finance," which includes brand names, domains, email addresses, and even Bank Identifier Numbers (BINs). By being able to search for these entities and events instead of just using keywords, a human using the system can focus on abstract concepts to uncover the intelligence they need, and not the many ways and different languages in which a source might talk about them.

3. Malware Entity Recognition

Recognizing new malware names presents a fairly unique linguistic challenge in cybersecurity. Newly discovered malware is generally named by a security vendor, a researcher, or sometimes the threat actor themselves. In many cases, these names are invented by putting existing or invented words (sometimes not English) together to create new nouns, or using unusual combinations of words. For example:

Panda Banker	**Onion Duke**	**Sub7**	**WannaCry**
Turla	**FALLCHILL**	**C99Shell**	

Our Threat Intelligence Machine applies supervised machine learning[4] to detect new malware entities. The model is trained on a large number of examples of text mentioning known malware names, and learns the contexts and language constructs used when talking about malware. If the model determines that the probability is high enough it will decide that there *is* a reference to a new malware. You can see the outcome of the process here:


```
"name" : "Googlian",
"status" : "ACCEPTED",
"sentence" : "The malware, dubbed Googlian, has been designed to steal Google
credentials from affected devices.",
"created" : ISODate("2016-11-30T18:06:10.739Z"),
"entity_context_score" : 0.897466719579364,
"entity_score" : 0.9999919409381894
```


Our machine learning identifies the context in which a new word appears to determine if that word refers to malware.

The machine is continuously retrained, learning new uses of words and context that increase its accuracy.

In the Real World

Trying to stay on top of new threats as they are reported in the news is a significant challenge for humans. There are clear advantages in applying machines to this particular problem — the technology scales to discover these new names as well as references to them, and makes the intelligence

[4] *http://blog.echen.me/2012/01/03/introduction-to-conditional-random-fields/*

available in real time. The machine can also address another problem: the fact that different security vendors and researchers will choose to use different names for the same malware, consolidating all the names into a single malware entity.

Malware aliases consolidated into a single view of available intelligence.

4. 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 the likelihood of product vulnerabilities being exploited, and to assess the risk that an IP address will behave maliciously in the future, and even before it has ever been referenced anywhere.

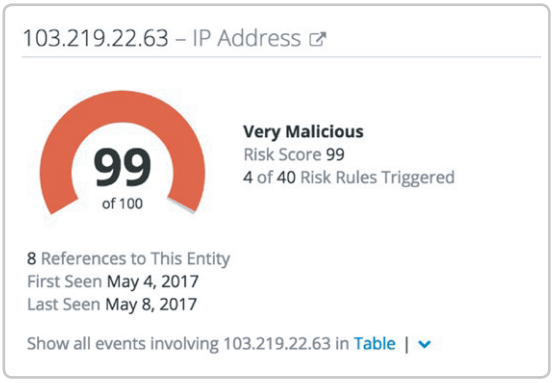
The challenge in all of these cases is to identify relevant context on which to base the predictions, and most of all, to get access to enough ground-truth[5] training data to be able to generate models that can be used to make predictions with the required accuracy.

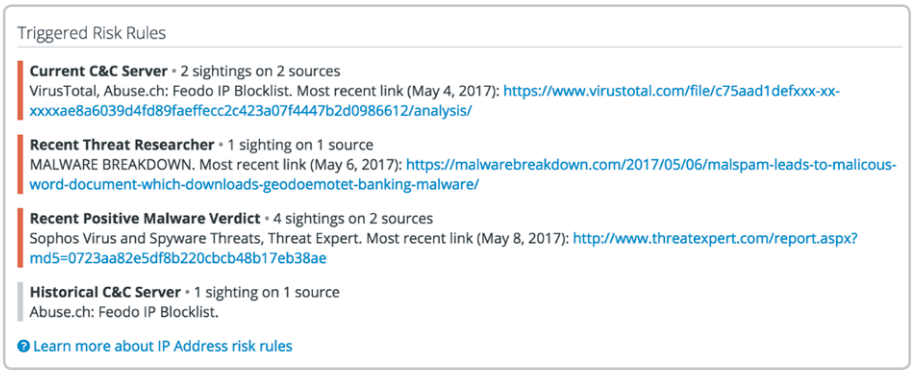
Prediction generation is an example of a task that is hard, or even impossible, for a human analyst to carry out due to the complexity and large volume of data needed. Algorithms and machines scale much better to problems of this kind.

[5] *https://datascience.stackexchange.com/questions/17839/what-is-ground-truth*

In the Real World

Let's consider a specific example. 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 (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 cybersecurity — 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, so they can't be used until something has happened, either locally or somewhere else.

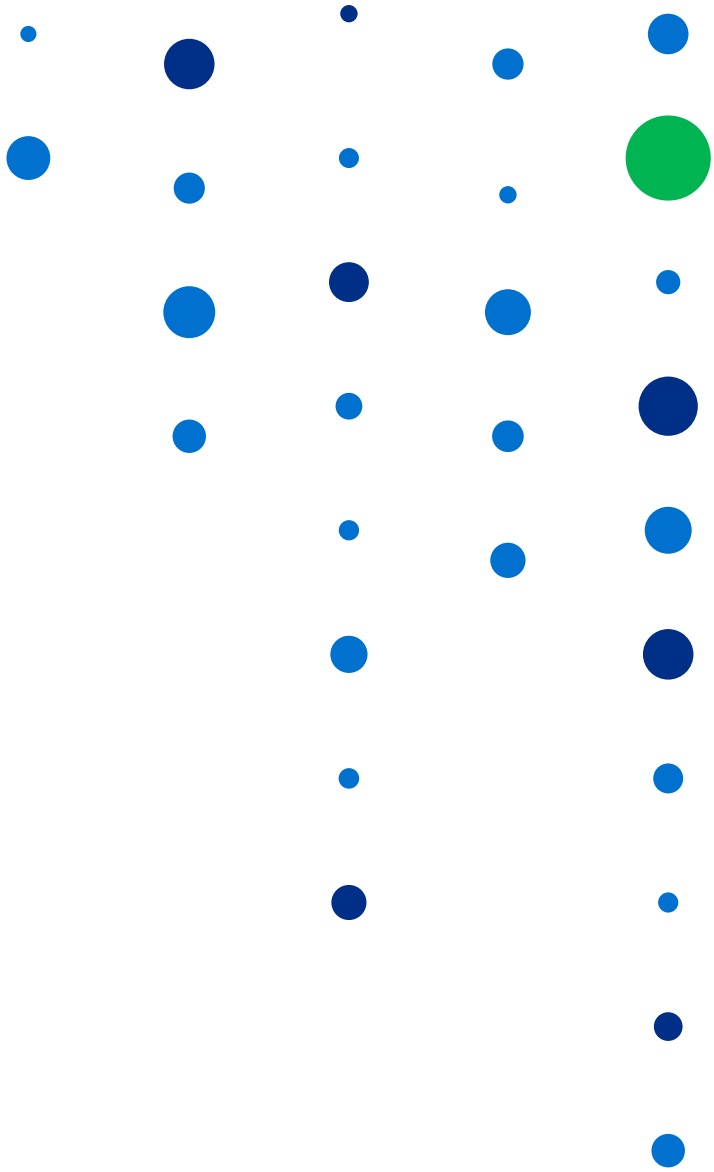
To further assist threat analysts and SOC operators, we've developed predictive risk scores. These scores are produced using a machine-learning model, which is trained on historic information from both threat lists and open source information. The algorithm can assign a predictive risk score to an as of yet unseen IP address. The predictive risk scores are based on historic and current risk scores for neighboring 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 of entities like domain names to identify likely new variants being used in typosquatting for phishing attacks.

Conclusion

Applying machine learning delivers two significant gains in the domain of threat intelligence. First, the processing and structuring of such huge volumes of data, including analysis of the complex relationships within it, is a problem almost impossible to address with manpower alone. Augmenting the machine with a reasonably capable human, means you're more effectively armed than ever to reveal and respond to emerging threats. The second is automation — taking all these tasks, which we as humans can perform without a problem, and using the technology to scale up to a much larger volume we could ever handle. We estimate that it would take around 10,000 humans to do the analysis done automatically by Recorded Future on a daily basis.

Though people can't process the volume of data machines can, machines struggle to deal with nuances in language, behavior, and motivation. While machines continue to develop advanced skills to overcome these limitations, it remains most effective to pair machine-led intelligence with human skills to maximize the benefits of both to proactively defend against threats.



Recorded Future

www.recordedfuture.com

@RecordedFuture

About Recorded Future

Recorded Future arms security teams with the only complete threat intelligence solution powered by patented machine learning to lower risk. Our technology automatically collects and analyzes information from an unrivaled breadth of sources and provides invaluable context in real time and packaged for human analysis or integration with security technologies.