WHITEPAPER 04/2022

# MACHINE LEARNING
## in Advanced Threat Detection

# Content

## Executive Summary

**Artificial Intelligence (AI)** is now everywhere. It is pretty clear that applications of AI add huge value in many business areas and use-cases. But there is also a lot of noise, hype, and misconceptions around this concept.

While AI is a broad term which refers to machines solving a problem in a smart way, Machine Learning (ML) is its most significant subset, especially for business world applications. The aim of machine learning is to train the machine to find a mathematical function – or model – that links given inputs to outputs. Once the machine is trained, then it's expected to turn new data into reliable predictions. When it comes to recognizing new patterns and finding new correlations within data that is generated in high volumes and high speed, ML has proven to be a strong tool for almost all industries.

Cyber security is no exception. In fact, machine learning can make a significant impact on cyber security by improving detection and automating response, resulting in **faster and more efficient** security operations.

The cyber threat landscape is expanding with enormous pace. And the World Economic Forum (WEF) claims that attackers are becoming more impactful in terms of speed, scale, precision, and stealth. This means, as WEF claims, that we need AI-supported defenses to keep-up with this fast-evolving threat landscape. But what is the value that comes with AI, and how can we get the best out of it?

To understand how to reach the true potential of Machine Learning, we should first know the stages to develop a machine learning model. In this paper, you will find brief explanations about how Machine Learning flow looks like, how it works, and what is needed to excel in each of these stages.

And here's a hint: to get the best of Machine Learning, it's essential to build and evaluate the models with **high-quality and trustworthy input:** **accurate, relevant, noise-free, and in-depth data.**

It is also important to both support and validate the output of ML with non-AI technologies, because ML will not be sufficient to provide ultimate security as a stand-alone solution. It brings its true value when it's supported and balanced by a broader set of cutting-edge technologies.

As pioneers of automated malware detection and analysis, we will approach AI from a realistic point-of-view, to explore the true value and capabilities of Machine Learning, beyond the myth.

# 1 Why the support of machine learning is needed in cyber security

We are on the edge of the 4th platform of transformation. First, there was only computing. Then came networks. The third platform combined the web, social media and cloud computing, and the fourth will be about “ambient computing” with practically infinite network connections interactively linking users, apps, intelligence and sensors. This means that everything about our professional and personal lives is becoming subject to increased cyber risks.

Market data reveals that digital threats already keep growing exponentially **in volume, frequency, and impact.** One reason for this is the growing scale of cyberspace that expands the threat surface (IoT, e-commerce, remote work, BYOD, etc.)[1]

In addition, cyber crimes are becoming more professional and coordinated. The dark web is full of online shops and individuals selling products and services: viruses, zero-day attacks, exploits, and infamous software to penetrate online banking. The “as a service” cyber crime model, the cloudification of everything, the rise of crypto currencies and even DevOps approaches used by attackers enable them to accumulate the budget and data to invest in R&D. This in turn, empowers threat actors to create more optimized and impactful attacks with higher volume, variety, and velocity.

**“We continue to see the cybercrime supply chain consolidate and mature,”**[1]

| | |
|---|---|
| **Attacker for hire** | $250 per job |
| **Ransomware kits** | $66 up front or 30% of the profit |
| **Spear phishing for hire** | $100 to $1,000 per account takeover |
| **Denial of service** | $311.88 per month |
| **Compromised mobile devices** | $82 to $2.78 |
| **Stolen usernames/ passwords** | $97 per 1,000 (avg) |

SOC (Security Operations Center) teams, which already have a heavy workload, are struggling to deal with this exponential growth. Given attackers’ professionalization in terms of speed, scale, precision, and stealth, it’s increasingly a challenge to keep up with this **fast-evolving landscape** by simply relying on traditional rule-based and signature-based systems. As the complexity of the predictive problems increase, we need too many rules to cover all the cases. Even then, there will be exceptions that will need to be handled. Needless to say, rules are difficult to design, maintain and adopt.

This is where Machine Learning can change the game. Against such a growing criminal ‘ecosystem’, AI provides robust algorithms that will help defend organizations against the cyber plague. More specifically, AI will help by improving detection, reducing false positives, sorting large volumes of information, monitoring logs, and shaping better software with fewer flaws.

Attackers keep getting better in terms of

SPEED
PRECISION
SCALE &
STEALTH

[1] Microsoft Digital Defense Report, October 21

This is why World Economic Forum (WEF) lists AI as second most important trend to shape cyberspace, stating that “It is critical that the cyber security community quickly prepares to combat fast-emerging AI-enabled attackers, by continuing to evolve technologies and operational capabilities that can match their pace, dynamism and sharpened predictive capabilities. While **non-AI risk controls will form an important baseline,** this likely means using faster and more dynamic AI-enabled defenses.” The WEF also listed dynamic analysis among the promising use cases of AI in cyber defense.

It is significant that WEF highlighted the importance of non-AI risk controls forming the baseline upon which AI and Machine Learning add extra value by improving the outcome of existing technologies. This implies that AI is needed, but it is – by itself – not a silver bullet that has an answer to everything. We need a combination of cutting-edge technologies and Machine Learning; supporting and feeding each other. We will revisit this topic in detail later in this paper.

## 2 Key definitions

Before we develop in detail of Artificial Intelligence and Machine Learning, it's worth providing a brief reminder of the relevant key terms.

Many of these terms have deeper roots than people generally imagine and go back in history to long before the emergence of computer science. Early works about probability start with general probability theories around the dice (Newton, 17th century), Linear Regression (introduced by Gauss in the19th century), and the Logistic Regression concept that Berkson coined in 1944. Early ideas about machine learning started in 1945, with the introduction of Neural Networks by McCulloch, followed by Regression Trees, Classification Trees and the concepts about Random Forest and much more.

This deeper history of theoretical and mostly mathematical studies became more applicable – and thus, more popular – as we have **higher quality and volume of data** to work on, and more powerful computers to process the data.

**Artificial Intelligence** is a concept that lets computers solve a problem in a smart way. It is a type of intelligence exhibited by non-human entities (machines) which tends to overlap on how humans think and understand but goes beyond human cognition. Its primary and most common objective is to find an optimal solution to a problem.

AI has many subsets from simple rule-based classification and automation to rule-based expert systems, statistical inspired algorithms, and Machine Learning.

**46bn $**

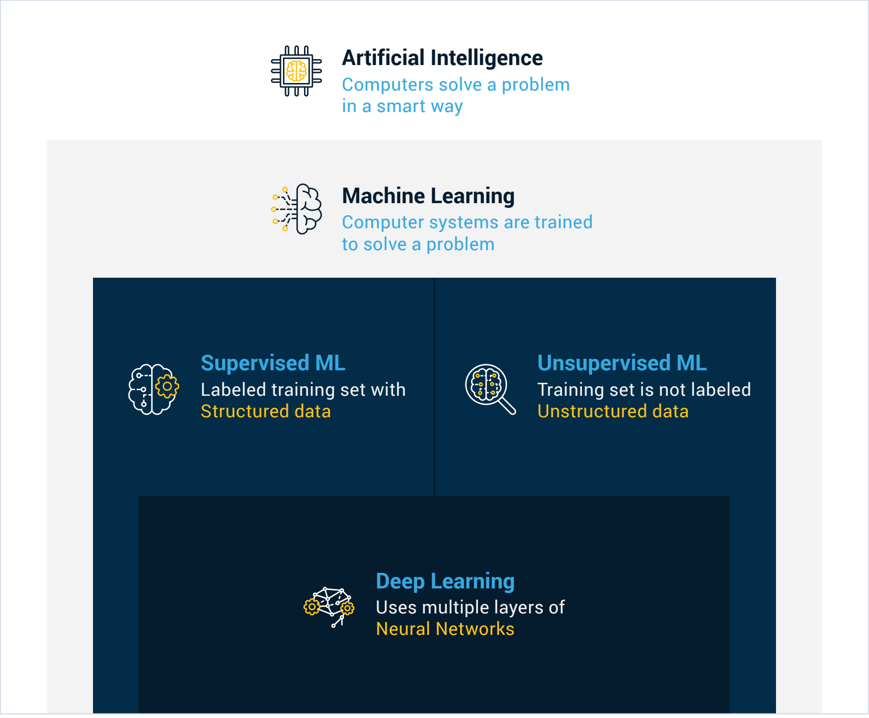
Estimated global value of **AI in cyber security** by the year 2027. Meticulous Research

**600%**

The increase in **phishing** schemes in May 2020, just two months after Covid-19 lockdowns began. United Nations

**>500,000**

video conferencing users had their personal data stolen and sold on the dark web between February and May 2020. Deloitte




Key Terms in Artificial Intelligence

**Machine Learning** is a subset of AI. This time the computer systems are trained to find an optimized solution to a problem, without being explicitly programmed to something exactly.

When we say "programming", it means that the computer/machine is given a function. The developers define the function, and all that is left for the computer to do is perform the computation. To give an example, it is similar to programming the computer to apply the function: "f(x)= 2x". From this point on, when you give a new input as "x = 5", the computer will perform the "2x" function and will simply give the output as "10".

In Machine Learning, however, the machine is not given the exact function. Instead, **it is trained to find and understand a mathematical function** that connects two sides of the equation. The process is a bit similar to the famous game show "Countdown", (or with the original name: "Des Chiffres et des Lettres") where contestants were supposed to find closest to a certain result by using given numbers.

This time we provide the machine with the input and output, and we expect it to find the correlation. Following the same example, we tell the machine that when "x = 5", then the output is "10". Now it has endless options to connect these two numbers, and it will be our job to **assess and validate the correctness** of the functions it finds.

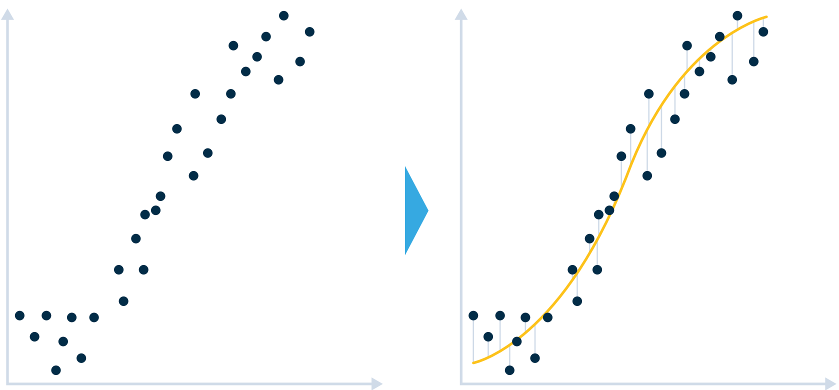

The Machine Learning Process turns input from past and current data into an output as a prediction

We assess the correctness of these functions by validating the algorithm with new data. When the new input is "x = 10", and the machine gets the results as "20", then it means that it possibly has found the right mathematical function as "2x". Otherwise, we have to tell it to "try again". With iterations, the machine is trained (taught) to develop an algorithm that makes **correct predictions with new input data.**

Obviously, the real computation is much more complicated than this simple arithmetic problem of finding "2x". Typically, there will be a huge number of input variables – the machine needs to calculate the weights and thresholds – and there will be different types of desired outputs.

We, as humans, have two critical roles in this training process: to provide the machine with **high quality input**, and **assess and validate** its predictions accurately. For this, we need to be confident and precise about the actual correct output that the model is supposed to predict.

In summary, the prediction is a function of the input data, and the machine learns to **define the "function" that connects the input and output,** through training and iterations.



The aim of Machine Learning is to find the optimized and generalizable mathematical function to connect the dots.[2]

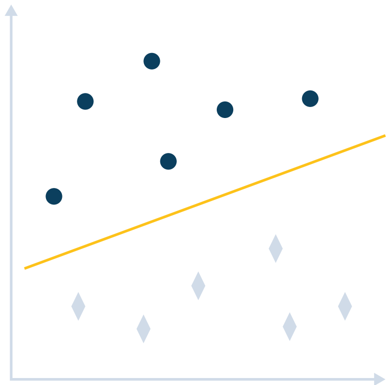
[2] An Introduction to Statistical Learning, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

The renowned mathematician Alan Turing envisioned a "child-machine" that would experience and learn in the same way as humans: through heredity, mutation, experiments, and choices following experiments.

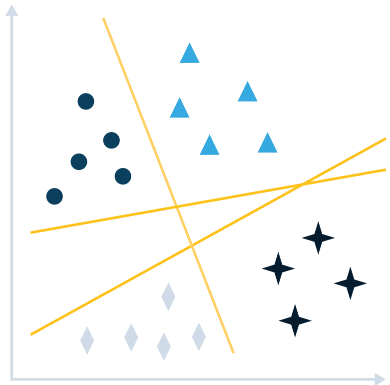
Similar to human learning, there are several ways machines can learn, but the two broad categories are: from lessons learned in a structured curriculum (supervised learning) or from random and unpredicted data (unsupervised learning). There are also semi-supervised and reinforcement learning options, which are not detailed within the scope of this paper.

**Supervised Machine Learning** is widely used, due to its practical value and ability to capture the domain expertise and experience of human experts. **Structured and labeled data** is used to train and test the algorithm. Through an iterative process, the model learns about the relationship between the input and output, improving its predictive accuracy until the trainer is satisfied that the model performs well with new data.

Supervised machine learning includes linear and logistic regression, decision trees, trees of trees, random forests and many more algorithms.



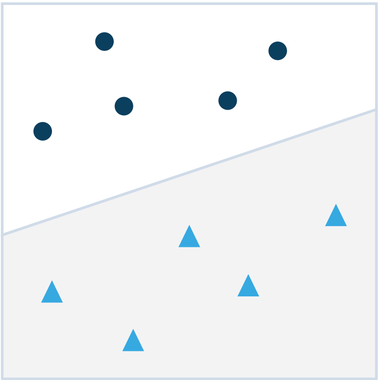
Binary classification generally addresses a "Yes/No" question



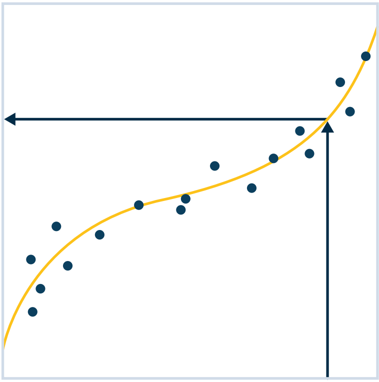
Multi-class classification separates into multiple groups

For discrete outputs, the result is in the form of a **class or label,** either binary or multi-class, depending on the use case addressed. Examples include determining whether an email is malicious or not, predicting whether a customer will buy again, etc.

If the desired output is continuous (and not discrete), then **regression models** can be applied. Examples are sales, price or weather forecasts.



Classification of discrete output



Regression for continuous output

With **Unsupervised Machine Learning** the learning model is trained and tested using large datasets that are **unstructured and unlabeled.** For example, the system reads texts and images and then creates clusters of objects according to similarity. Here, the model is trained based on some measure of similarity or distance and is expected to explore whether there are any groups, for example, with similar behavioral patterns.

**Common uses for unsupervised learning include:**

- Identifying hidden patterns, trends, and relationships within large datasets
- Clustering similar objects into groups for purposes like customer segmentation or recommender systems
- Pre-processing data before labeling it for use in supervised learning
- And many more.

Cluster analysis or clustering is the task of grouping a set of objects in a way that objects in the same group are more similar to each other than to those in other groups.

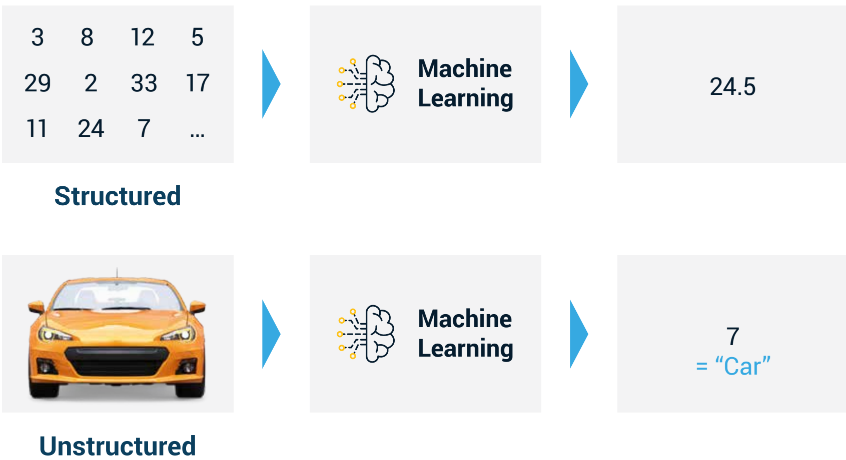
Clusters might be determined by attributes such as age, payment history, purchase history, etc. So, it will be important to find the most relevant type of clusters for the desired result. The below image shows how the same data set can be clustered differently, according to different attributes (for example according to shape or color in below example).

Raw Data

Clustered according to proximity

Clustered according to shape

One of the main challenges of working with unstructured data is that you need to turn the input into mathematical vectors. This is easier to do with structured data: turning "Yes or No" into "1 or 0". It's trickier to translate an image, sound or text into a mathematical representation so the machine can make its computations.




Processing structured vs unstructured data into a mathematical format

**Deep Learning** is a subset of machine learning, usually dealing with extremely large amount of unstructured data, about which we have almost no knowledge. Leveraging a **multi-layer neural network,** often running in GPU-powered machines, the system is trained to handle combinations and activations of learned parameters, ranging in number from hundreds of thousands to hundreds of millions.

Non-linearity is considered, allowing the network to learn and generalize. The system can also construct and modify the neural network architecture, extracting relevant features without the tedious process of feature engineering.

In this way, it's quite similar to the fictional supercomputer in *The Hitchhiker's Guide to the Galaxy:* you give the data and ask a question, and don't have a clue about even what features the model is taking into consideration for its decision.

One of the most conspicuous examples of Deep Learning are virtual assistants, like Siri and Alexa, autonomous vehicles, and NLP (Natural Language Processing).

## 3 AI use cases in cyber security

In general, Artificial Intelligence can be used in four different ways:

- For **descriptive analytics,** which examine "what happened?"
- For **explanatory analytics** that reveal "what factors contributed to this outcome?"
- For **predictive analytics** that tell us "what is likely to happen in the future?
- For **prescriptive analytics** which – when used for optimization purposes – focus on the effects that different actions have on end results.

Let's get into more detail about the use cases of AI and Machine Learning, specifically in the field of cyber security.

According to Gartner, the main requirements for AI in security are **improving detection and decreasing FP's.** Supporting SOC tasks and automating investigation and response come next as further evolution points. So, the first steps of introducing AI within an organization are almost always focused on improving existing solutions.[5]

Another Gartner study ranks 19 potential AI use cases in cyber security – starting with the most promising and valuable – according to an index of business value and feasibility. This study can be seen as a strong indicator that the threat detection category potentially offers the highest value to the market, compared to "policy" and "response" categories. Of the 19 use cases listed on Gartner's "AI use-case prism", the top 5 are all linked to threat detection.[6]

**AI methods and techniques** are being integrated into products in all security market segments, potentially making this technology, in aggregate, the **largest impact on attack detection** development for the next five to eight years. Gartner

[5] Emerging Technologies: Tech Innovators in AI in Attack Detection — Product and Business, Gartner Report, November 21
[6] Gartner AI Use-case Prism for Cybersecurity, 2021

**Top 5 Use-Cases of AI in Cyber security:**

- Transaction fraud detection,
- File-based malware detection,
- Process behavior analysis,
- Abnormal system behavior detection,
- Web domain and reputation assessment

Two of these top five are already among the core strengths of VMRay: web and domain reputation assessment, and file-based malware detection. VMRay's Machine Learning Lab has developed and deployed a machine learning model to **enhance threat detection** capabilities that are part of VMRay's comprehensive stack of cutting-edge technologies.

Gartner also profiles several vendors who are using machine learning in diverse and innovative ways to address a range of cyber security challenges and vulnerabilities:

| Methods and techniques | Impact | Machine learning model |
|---|---|---|
| Classify OS process behavior in real time to catch suspicious processes | Block phishing attacks from diverse sources. Reduce FPs and noise | ML and deep learning |
| Identify repackaged threats according to binary code similarities | Detect malware that are remodeled and modified to evade next-gen antivirus tools (NGAV). | ML, binary similarity analysis |
| Offload part of SOC analysts' activities by validating alerts | Streamline investigations and mitigate skills shortages by automated investigation and response. | Deep learning, deep neural networks |
| Capture and encode the expertise of senior analysts in scenario pools | Improve SOAR automation by recommending remediation steps based on prior expert knowledge | Supervised and unsupervised ML methods, AI-assisted data labeling |
| Use pattern analysis across multiple API traffic vectors to detect and stop API attacks | Address known gaps in API security | AI/ML for pattern analysis of API vectors |

Emerging Technologies: Tech Innovators in AI in Attack Detection — Product and Business, Gartner Report, November 21

## 4 Our aproach: combining the best of two worlds

Artificial intelligence is not a silver bullet and shouldn't be seen as a stand-alone solution; it won't be sufficient to provide ultimate protection by itself. As suggested in the above-mentioned WEF report, non-AI technologies should form the baseline, and to get the best out of AI, you need a **carefully arranged combination of cutting-edge technologies and machine learning.**

When you have a broader set of technologies, that includes Machine Learning, you can address different threats with whatever technology that fits. Only then, instead of having Machine Learning as one major tool, you will have it as a part of a broad set of technologies that enable you to **detect the undetectable.**

When it comes to creating the Machine Learning model, you need the most reliable technologies to support this process. Because what makes a Machine Learning model stronger is how you train and develop the model: namely, the qualities of the data and the expertise of the team.

Both the input that you feed into the model and the accuracy of how you assess and validate its predictions have vital importance. Thus, Machine Learning can only add value **when it's based on an already advanced technology platform** with outstanding detection capabilities. Otherwise, neither the input would be reliable, nor – as a consequence – is the prediction.

VMRay is developing its machine learning model on top of our Advanced Threat Detection Platform. The platform itself was built on our groundbreaking sandbox developed by our founders, the pioneers of advanced threat detection. Throughout the years, we have gone far beyond this strong basis, adding new building blocks on this foundation by creating **20+ cutting-edge technologies.** And this number grows with each new release.

These technologies make VMRay Platform resistant against even the most evasive malware and phishing threats, as it detonates the samples in a safe environment. Our platform observes all phases of malicious behavior from outside this environment. Since the threat is never aware of being observed, it displays it real, genuine behavior while the platform logs every necessary detail.

This is why VMRay is trusted by the largest companies and most critical public institutions around the world. VMRay enables its customers to detect the undetectable threats: **unknown and zero-day threats** as well as **advanced and targeted attacks,** such as those including password-protected attachments or multi-step phishing URLs that activate malicious behavior only after a number of steps.

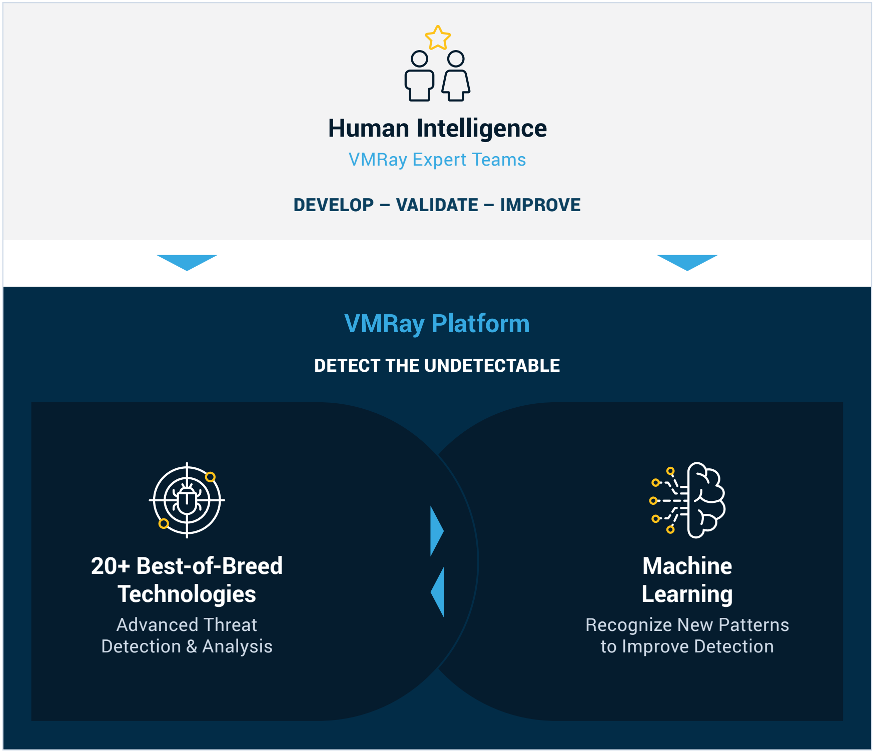
Only when you have a comprehensive technology platform you can ensure that the Machine Learning model is continuously fed with the best input derived and filtered by best-of-breed technologies. And **our approach is to bring the best of two worlds together,** creating a virtuous cycle between our existing technologies and our machine learning module.

Our unique monitoring technologies feed the Machine Learning model with the **highest quality input:** accurate, uncontaminated, noise-free and in-depth.

The Machine Learning module responds by recognizing additional patterns and bringing additional detection capabilities.

It's important to validate ML predictions. Our existing technologies provide the accuracy and precision needed to assess and validate the predictions.

The prediction of ML module is not delivered as a stand-alone verdict. It becomes a part of our unique VTI rule set and in a way, both **validated and balanced** by the "decisions" of other technologies.




In any case, it's hardly possible to use Artificial Intelligence without "Human Intelligence": the experts in cyber security and data science. On top of the technologies and Machine Learning, the **experience and domain expertise of the team are needed.** This is especially necessary while defining the "features", the variables that ML is expected to consider, as well as evaluating and fine-tuning the data models and algorithms.

Standing on vast experience and know-how about unknown and undetectable threats, the VMRay experts and ML lab is continuously developing, improving, and validating the outcome of our Machine Learning module.

In addition to that, the predictions of the Machine Learning module are **not presented as a stand-alone verdict.** It's a part of a bigger set of rules that define the final verdict. This way, the ML outcome is also balanced – and in a way, validated – by the other technologies within the platform.

So, our approach is to use Machine Learning along with our best-of-breed technologies to support and enhance detection capabilities to perfection, by **combining the best of two worlds.**

In the next sections, we will dive deeper into the necessary qualities of the input and the challenges in evaluating the model, and how this approach creates real value to our customers.

## Explore all 20+ unique technologies similar to these:

**Intelligent Monitoring:**
Allows VMRay to stay invisible to evasive malware as it runs solely in the hypervisor layer and without affecting the analysis environment.

**Smart Memory Dumping:**
Advanced triggers to accurately dump and store relevant memory buffers of analyzed malware in real time that enables timely detection.

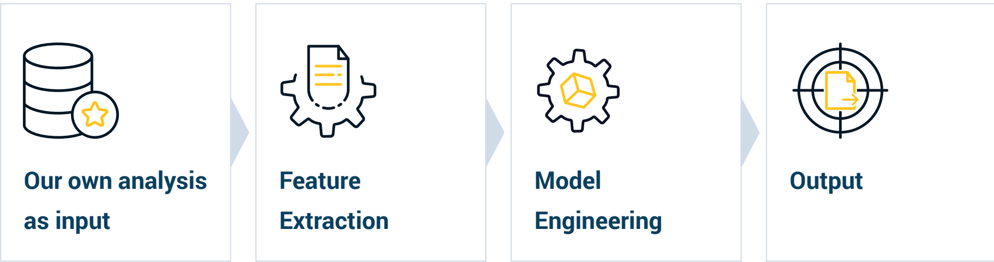
**Adaptive Browser Simulation:**
Detects and clicks on buttons that activatephishing attacks to automatically trigger the payload delivery.

Learn more on VMRay.com

## 5 How It Works

Below, we briefly describe how Machine Learning models are created:

VMRay's Machine Learning flow:




### Input data

The input data is probably the most critical component of a machine learning model. Everything from the model creation process to the actual predictions of the model is dependent on the data that you use as input. We will discuss the necessary data requirements in "What it Takes to Create a Reliable Model" section.

We use our own analysis as input to feed and train the model. The data that we accumulated from our past analyses is used for training the model.

When it comes to analyzing a new sample, the model works as a module of the VMRay analysis platform – following the static and dynamic analysis engines – and gets the actual data about the real behavior of that sample. By limiting the input to only using output from our previous analyses, we ensure that the model is trained with and makes its prediction based on the **highest quality input.**

### Feature extraction

Feature extraction is the phase where the features are selected and translated into a mathematical form that the model can understand. This task gets more complicated as you switch from using structured data to unstructured data. So, this is the stage where the machines are taught – or they decide – how to **determine which input variables** they need to take into consideration.

Features indicate which input variables the model should consider.

When you build a model, you have a number of parameters or features that can be used to predict a desired outcome. Often times however, a significant portion of these features are redundant. Feature selection is the process of identifying the most important features and eliminating the irrelevant or redundant ones.

VMRay uses a supervised machine learning model, so we define the features **by the expertise and experience of our experts.** This way, we capture the know-how of our experts in the model as a reusable asset. The most relevant features are picked among numerous potential indicators such as URL string entropy, white space percentage and many more.

### Model engineering

A model is a paramaterized function through which we can map inputs to outputs. Once the best possible way to match the input and output is created, the model can then be used to compute the output, based on a set of new input variables.

The models are created through a careful and meticulous engineering and experimentation process of **selecting, validating and evaluating the models.** Both the type of input data and the type of outcome are critical in choosing the most suitable algorithm. We trained our models by creating a Machine Learning workflow, which will enable us to explain the significant parts of the workflow that contributed to the prediction, such as sample set collection, feature engineering, feature weights and inference, etc.

In most cases, the success of model engineering is a matter of **avoiding overfitting** to the training data set and finding the optimized balance between the accuracy and false positives of the outcome.

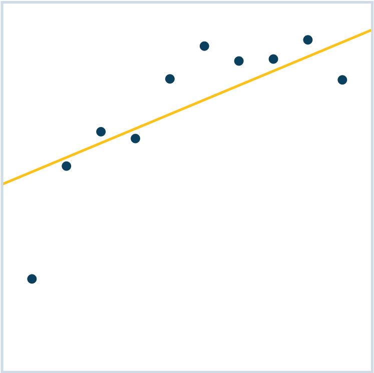
When a machine learning model is given a set of input variables and the outcomes, it tends to find the function that represents an exact match. This leads to a tendency to overfit. However, as a general principle, we do not need an exact match. Instead, **we need a generalizable model:** one that will be almost equally reliable when it sees a new dataset.

By experimenting with the number of hyperparameters, we can control the complexity of the model. In the regression example in below figure, you can see why an optimal level of model complexity is needed.

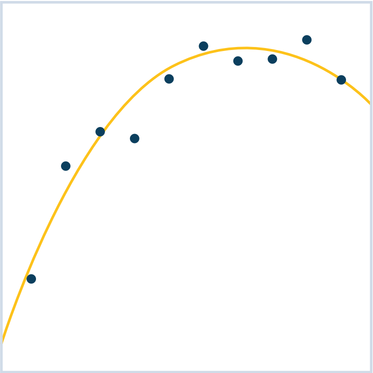
When the model finds the exact match of input and output in the training dataset, as in the overfitting example, it would be hard to apply this model to external data. On the other hand, if the model is too simple, then it will not be able to perform well even with the training data.

**Reliable** data and analysis technologies are required for model

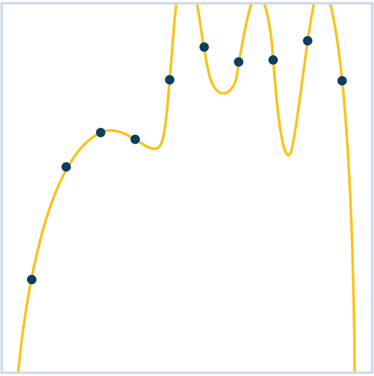
- Development
- Selection
- Validation
- Evaluation
- Improvement



Underfitting



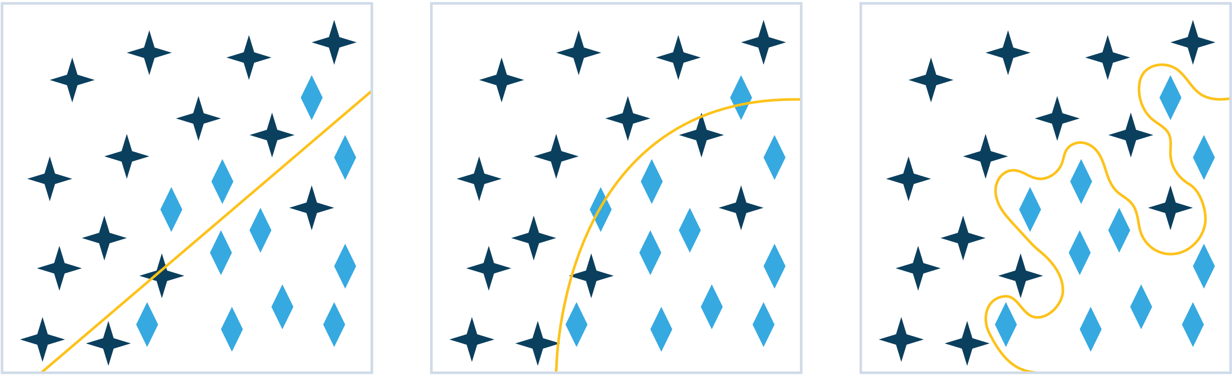
Optimal



Overfitting

Hyperparameters are parameters that control the complexity of a model. Specific hyperparamenters turn a “model family” into a concrete model. In this example, hyperparameter is the “degree” of the polynomial function.

The figure below shows another example, this time on classification. The graph on the left has a model that cannot reliably divide the data distribution. The one on the right, on the other hand, finds a function that performs perfectly in the training dataset. But it sticks to the training data so firmly that it cannot be generalized. The optimal model in the middle defines an acceptable balance between the accuracy and the false positives.



Underfitting　　Optimal　　Overfitting

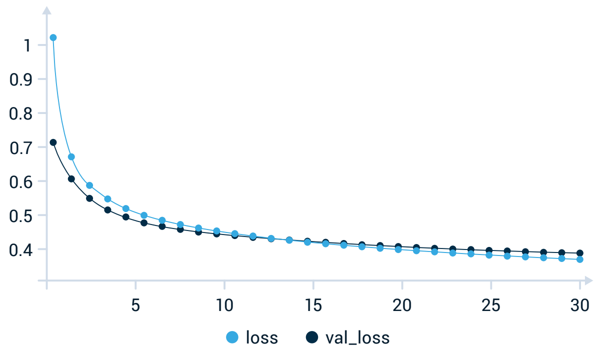
**In short, we need a model that does not "memorize" the data, we need it to "learn" the underlying function.**

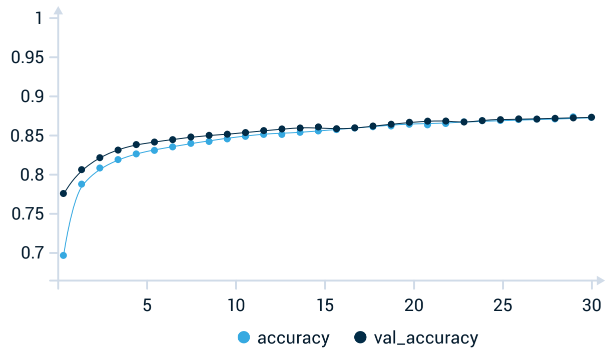
To achieve this, we need a process called "Model Evaluation", which basically means that we're testing and improving the model so that it performs well not only within the training set, but also with new data.

This stage is where **the quality of the data, and the expertise** pays off. Because you need to be confident in evaluating the accuracy of the prediction. For this, you need to have the reliable data and the expertise in the first place. Here's how it works:

For model evaluation, the data set is divided into two subgroups. The first one – generally the smaller group – is the training set, while the rest is called "validation set". The model is trained to find the function that ties the input and output within the training set, and then this function is tested with the validation set, which is outside the data upon which the model is trained.

There are different methods to evaluate the performance on the validation set and finding the optimal balance, such as comparing loss (bad prediction) with validation loss, and accuracy (correct prediction) with validation accuracy. These are expected to get closer as the model "learns" through iterations. The closer the performance within the training set and validation set are, the more generalizable the model.



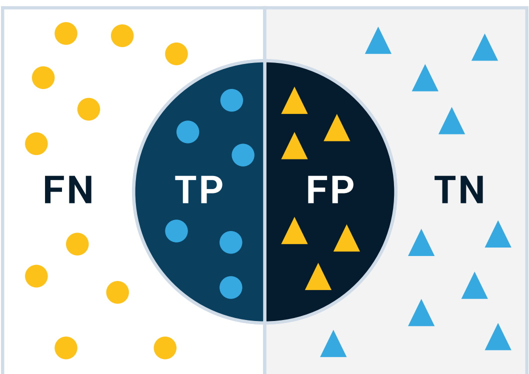

Source: An Introduction to Statistical Learning, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

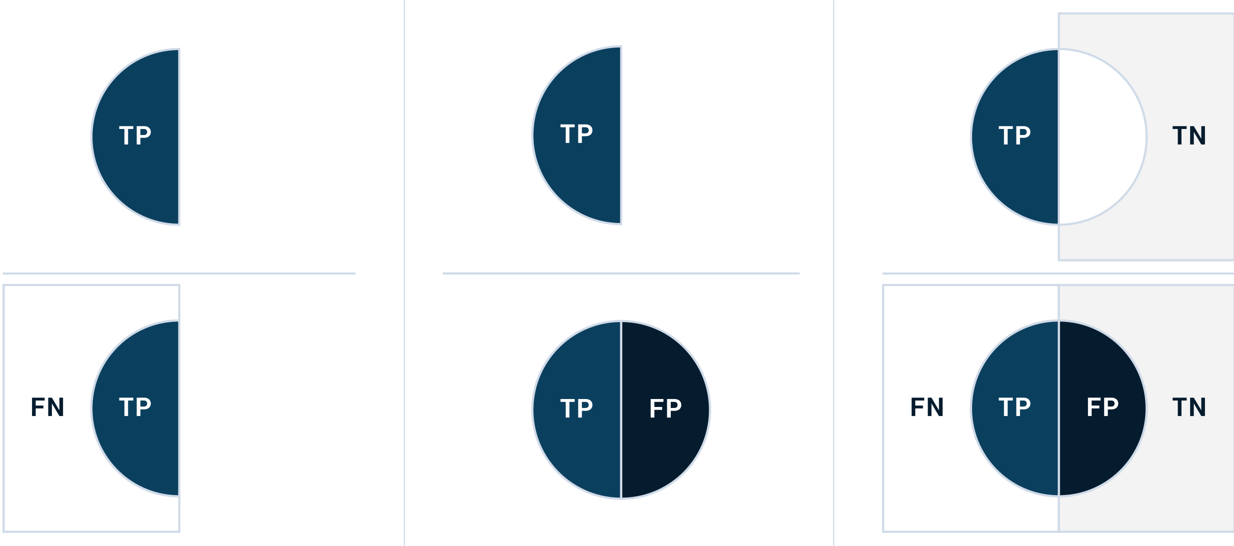
Another very important tool for model evaluation is the Confusion Matrix, where we can have both a general overview and conduct deeper analysis about the **balance of accuracy and FP.**

| | Predicted Class | |
|---|---|---|
| **Actual Class** | **Yes** | **No** |
| **Yes** | True Positives | False Negatives |
| **No** | False Positives | True Negatives |

**Confusion Matrix**

For example, when evaluating a model that predicts whether a sample is malware or not, we can use this tool to calculate:






**Recall:** How much of the actual malware is caught, calculated by dividing predicted malware to all malware: TP/(FN+TP)

**Precision:** How much of the "positive" predictions are correct, calculated by dividing "correctly predicted malware" to all predicted malware: TP/(TP+FP)

**Accuracy:** How much of the predictions are correct, basically, true predictions divided by all predictions: (TP+TN) / (TP+TN+FP+FN)

Model engineering is a tedious process of developing, selecting, validating, evaluating and improving models. For this process, we need the **highest quality and trustworthy material** that we can rely on. This is why VMRay limits the source material by only using its own analysis results.

VMRay Platform analyses samples with its cutting-edge advanced detection technologies, observes and reports the "actual" behavior of malicious samples, and brings **accurate and noise-free outputs.** This is why VMRay is trusted by the largest private and public organizations as the ultimate source of truth. VMRay's platform already offers the accuracy and clarity needed for validating alerts and false positives, and these capabilities makes the real difference when evaluating the performance of a machine learning model.

## Output

Within the VMRay platform, the machine learning module is an additional layer that is fed by the outcome of static and dynamic analysis. The output contributes to our unique VTI (VMRay Threat Identifiers) scoring methodology, as a separate VTI rule. Although we do not deliver the ML prediction as a standalone verdict, users can still see the ML prediction on the detailed VTI list.

## 6 Creating the best in class

So, what's in it for customers? What is the value that the additional capabilities of VMRay's Machine Learning module bring?

The two main outcomes of our Machine Learning module will be improving detection capabilities and reducing false positives.

**Improving threat detection:**
With a model that continuously evolves, the customers will have ultimate protection against emerging threats, and feel more confident that their ML-enabled security solution is always up to date, compared to having only the rule-based, heuristics-based systems.
This, combined with the supplementary approach of VMRay, means that organizations can leverage the capabilities of our machine learning model without any need to make a drastic change in their security stack.

**Instead of replacing** their current security solutions, they will be able to **augment and complete** those solutions by covering the blind spots, thus **boosting the value** gained from existing security investments.

Cover blind spots

Complete your security

Increase ROI

**Perfecting the Balance between Accuracy and false positives (FPs):**
False positives are always a big concern for EDR/XDR and SOC automation; especially for the efficiency and efficacy of security teams. As VMRay sees the true face of the threats, the output of our analysis is accurate and noise-free. We provide concise, to-the-point reports, delivering the customers all they need, and only what they need.

Accurate, in-depth, yet noise-free input is a must-have to enable effective automation. With these features, along with **seamless integrations** with all popular Email Security, EDR/XDR, SIEM, SOAR and Threat Intelligence platforms, VMRay brings the capability to validate false positives and empower automation with accurate, detailed yet noise-free input. The Machine Learning module is bringing these qualities a step closer to perfection.

With **noise-free and concise output that enable full automation,** VMRay empowers organizations to improve team efficiency, addressing the major challenges in the industry: the skills shortage and the need to empower junior security analysts to make better decisions. This way, our customers can focus on more strategic tasks to move forward in their digital transformation journey with ultimate peace of mind.

Empower automation

Efficient security teams

Focus on strategic tasks

## 7 What it takes to create a reliable model

When we speak about what is the most important requirement for Artificial Intelligence and Machine Learning, **it always comes down to the input data.** It's the data that makes the difference because data is the raw material of ML.

But not all data is created equal.

We need certain qualities to rely on the input data to train and run a model. And those qualities are not easily found. An HFS Study shows that 75% of executives do not have high level of trust in their data. According to another study from Gartner, 40% of enterprise data is either inaccurate, incomplete, or unavailable. So, the first thing needed to create the Machine Learning model is to find the most trustworthy input.

And this is where VMRay excels. We provide the model with highest quality input in three aspects:

### Veracity

"Access to **good data** is one of the major challenges of AI/ML development" says Gartner, highlighting the importance of reliable data.

Our core technologies and the new innovations we keep creating and introducing with every release enable us to **see the true face** of the enemy. While we analyze a file or URL, it displays its genuine behavior, because it is not aware of being observed. We can bring the most accurate data to the table, which is very much needed for building a reliable ML model.

In short, the input that we use to feed, train, and validate the model is **accurate and noise-free data.**

### Scope

Covid19 showed us once again that the data from the past loses relevance in a short time, because the world is changing in an exponential pace. As a result of this ongoing disruptive transformation, a new concept is emerging: "Small and wide data", where "small" refers to the increasing importance of relevant, to-the-point data, instead of big volumes of it.

We specialize in **what matters the most:** the types of threats that others miss. The unknown, zero-day threats and the sophisticated and targeted attacks such as attacks that use evasive techniques or URL's that activate malicious behavior 3-4 steps into the process, etc.

Our data and expertise are right to-the-point, when it comes to detecting what matters the most. Created with this **relevant expertise and data,** our models know where to look for and how to create an algorithm for accurately predicting undetectable threats.

**75%**

of executives do not have high level of trust in their data.
HFS Research

**40%**

of enterprise data is either inaccurate, incomplete, or unavailable. Gartner

**70%**

of organizations will shift their focus from big to small and wide data by 2025. Gartner

## Range

To create the most reliable machine learning model that helps detecting new threats, you need data that includes a wide variety of threat types, targets, vectors. And behavior patterns.

We have a diverse client portfolio in terms of verticals, regions, and company sizes.

**We're working with top companies:**

- **14 out of Fortune** 100
- **4 of top 5** Global tech giants
- **3 of "Big 4"** Accounting companies
- **17 of the World's** Most Valuable Brands

In addition to the private companies, we're working with more than 50 critical Government customers from 17 countries. This adds a huge range to our expertise and know-how.

And when we analyze a URL or file, we provide in-depth vision to the actual, genuine behavior of the threat. We log **every necessary detail in every step** of its execution. This adds enormous breadth to our data.

The bottom line: our strength is in the **data and expertise** that we use to build the machine learning model – both the past data that we use to train and validate the model, and the actual analysis data to create the output.

**VERACITY**
Reliable

**SCOPE**
Relevant

**RANGE**
Wide

## 8 Conclusion

VMRay is dedicated to creating the best Machine Learning module to detect undetectable threats and adding this module to support its comprehensive platform of advanced threat detection technologies.

Two pillars are crucial to develop a reliable Machine Learning model:

### THE RAW MATERIAL

To create a premium product, you always need the best raw material.

We dig deep to find the gold which is uncontaminated and freed from the mud and earth: the data to be used for training, evaluating, and validating a model.

As we see the real face and genuine behavior of a threat, our data is **accurate, relevant, noise-free, and in-depth.** This means that it has all the qualities that could be expected from the input data to build a model.

Accurate

Noise-free

Relevant

In-depth

### THE CRAFTSMANSHIP

How you process this raw material is equally important.

Our processing is also reliable because we define the features and create the models based on our expertise. And from the very beginning, we specialize on the **threats that others miss.**

We are the experts of the unknown. We can "see" in the deepest parts of the ocean where there's no light, and we have the know-how about the uncovered deep-sea creatures. Our teams have the expertise, experience and excellence needed to create the best Machine Learning model for advanced threat detection.

Expertise

Experience

Excellence

**When both the raw material and the process are reliable, so is the result.**

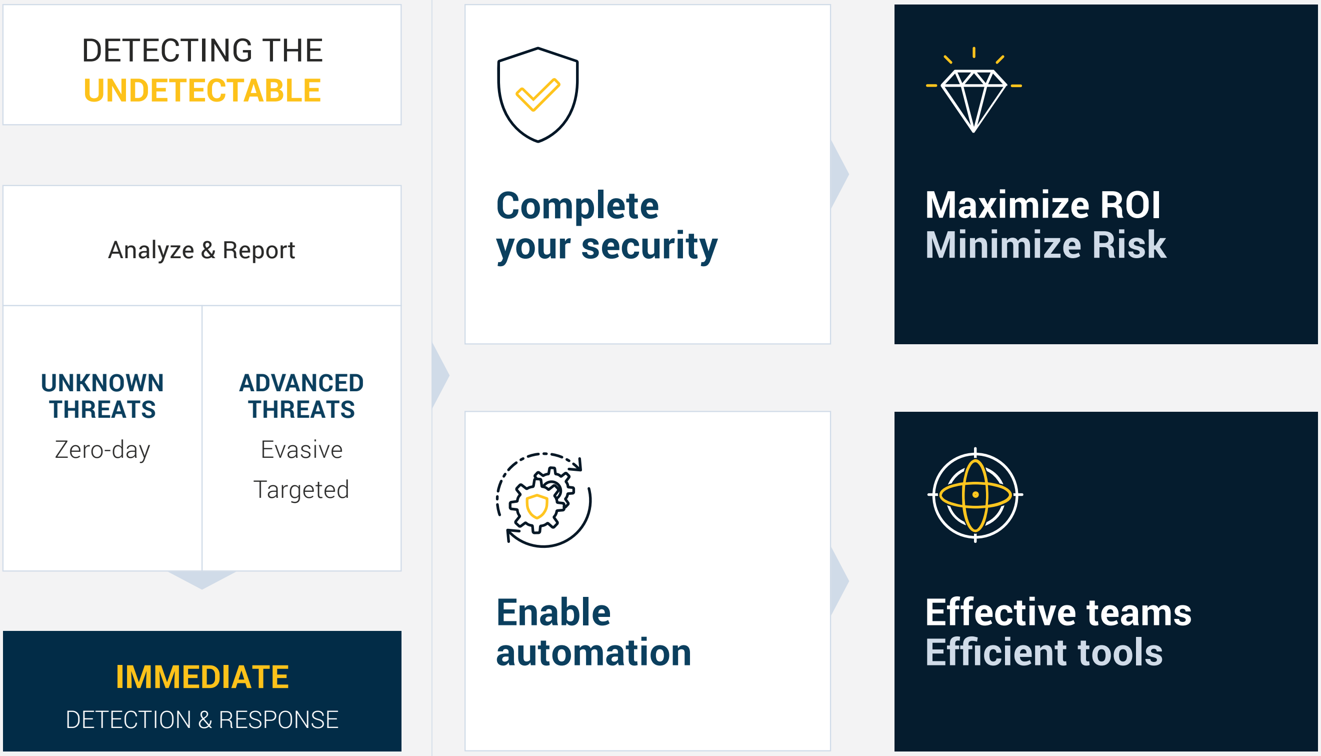
## About VMRay

At VMRay, our purpose is to liberate the world from undetectable digital threats.

Led by reputable cyber security pioneers, we develop best-of-breed technologies to detect unknown threats that others miss. Thus, we empower organizations to augment and automate security operations by providing the world's best threat detection and analysis platform.

We help organizations build and grow their products, services, operations, and relationships on secure ground that allows them to focus on what matters with ultimate peace of mind. This, for us, is the foundation stone of digital transformation.

Read more about our solutions at vmray.com





**Contact Us**
**Email:** sales@vmray.com
**Phone:** +1 888 958-5801

**VMRay GmbH**
Universitätsstraße 142
44799 Bochum • Germany

**VMRay Inc.**
22 Boston Wharf Road, 7th Floor
Boston, MA 02210 • USA