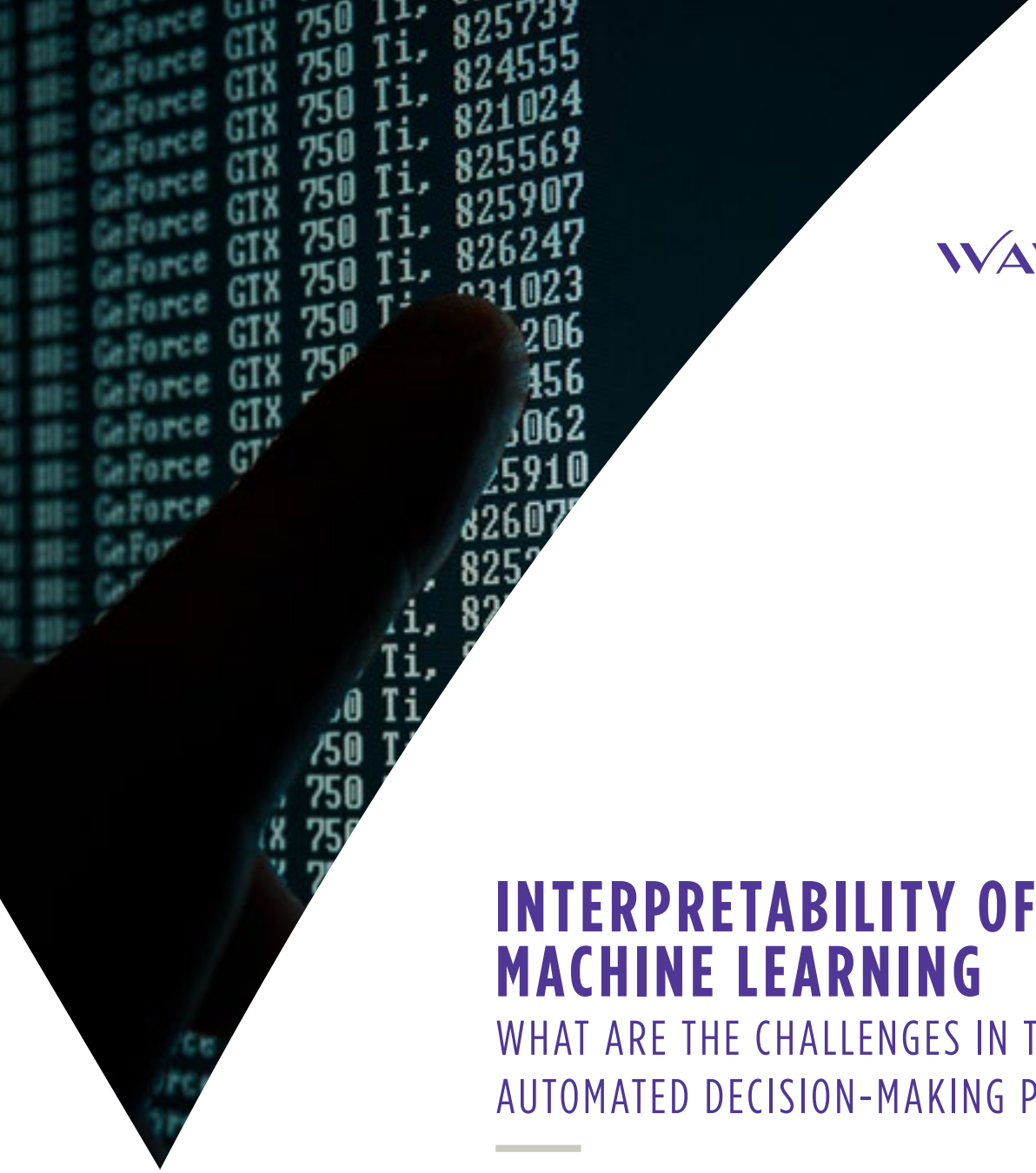


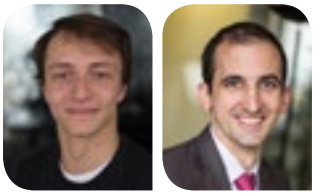
The Positive Way

WAVESTONE

INTERPRETABILITY OF MACHINE LEARNING

WHAT ARE THE CHALLENGES IN THE ERA OF AUTOMATED DECISION-MAKING PROCESSES?

AUTHORS



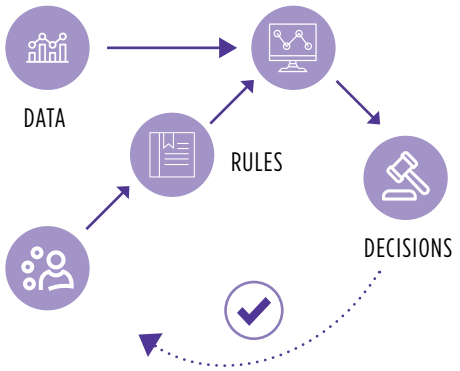
ALEXANDRE VÉRINE
alexandre.verine@wavestone.com

STÉPHAN MIR
stephan.mir@wavestone.com

The evolution of Artificial Intelligence since the 70's has intrinsically reinvented Decision-Making processes. The rise of Machine Learning has made possible to learn directly form data instead of human knowledge with a strong emphasis on accuracy. The lack of interpretability and the introduction of possible biases has led to ethical and legal issues. The EU General Data Protection Regulation took actions and a growing concern has risen on the Interpretability of Machine Learning algorithms.

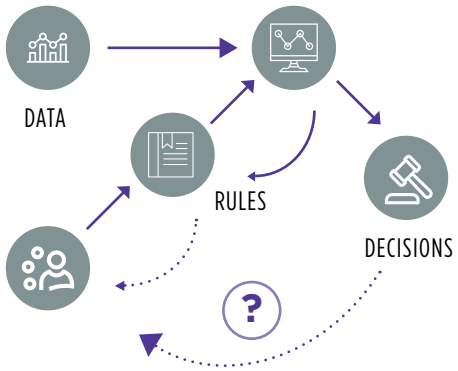
On the other hand, Interpretability can also help companies leverage the new AI tools to gain deeper insights on their decision-making processes.

Deterministic Automated Decision Making



HUMANS DETERMINES THE RULES THEREFORE CAN EXPLAIN THE DECISION

Decision Making based on Machine Learning



THE INTERPRETABILITY TOOLS ENABLES TO EXPLAIN THE RULES OR THE DECISION IN HUMAN-UNDERSTANDABLE TERMS

AI : FOLLOWING THE BLACK BOX, INTERPRETABILITY?

The main difference between the AI of the 70's and current AI is that it is not purely deterministic. In other words, decision rules and mathematical equations were set by humans and implemented to be automatically processed. Today, with Machine Learning algorithms, these rules can be directly extracted from the data. Consequently, the automated decision lost its inherent interpretability since the rules may easily be hidden by the complexity of algorithms or the multiplicity of inputs.

Interpretability – The ability to explain or to present in understandable terms to a human –has been set to be a minimal requirement for some automated processes. In 2018, the European Union General Data Protection Regulation (GDPR) regulated any significant or legally related decision to be explainable. The subject can require human intervention to challenge the decision. There are other regulations in specific domains such as the Code of Federal Regulations of the US which states that every credit action has a well-established right for explanation.

Science wise, the CNRS has been discussing the limits and consequences of such regulations and developing solutions. Whereas the DARPA (US Army – Defense Advanced Research Projects Agency) currently has annual funding of 400M$/year for the Next Generation AI project. It includes the XAI (eXplainable AI) project which aspires to increase the interpretability of Machine Learning.

INTERPRETABILITY OR EXPLAINABILITY?

The most general way to define Interpretability may be the ability to explain or present in understandable terms to human. However, another definition can be given by differentiating the interpretation and explanation. This enables to establish the difference between two scales of observation. The interpretation is "how does an algorithm make a decision?" and an explanation is "why has a singular decision been made by the algorithm?".

In other terms, the interpretation is the global evaluation of a decision-making process. It aims to represent the relative importance of every feature. On the other hand, the explanation provides a local insight on which features were determinants in a specific decision.

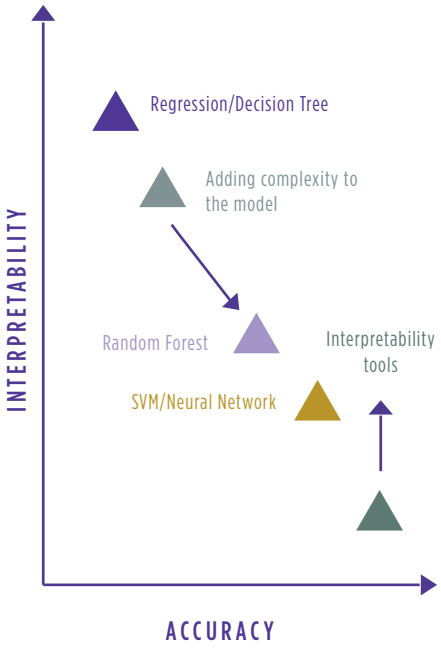
GDPR Recital 71

"The data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and **which produces legal effects concerning him or her or similarly significantly affects him or her**, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention. [...] In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and **the right to obtain human intervention**, to express his or her point of view, **to obtain an explanation of the decision** reached after such assessment and to challenge the decision."

SHOULD INTERPRETABILITY BE GIVEN PRIORITY OVER PRECISION?

In the area of Machine Learning methods and algorithms, the levels of Interpretability may vary greatly. Some methods are human friendly since they are highly interpretable. Others are too complex to apprehend and thus require ad-hoc methods to obtain an interpretation.

General Tendency of the Accuracy vs. Interpretability Trade off



With the use of Big Data, the number of features and the high dimensionality complexifies the method comprehension. For instance, a Decision Tree is a sequence of decisions in order to split the data. Those decisions are easy to understand if the sequence is not too long. However, the Random Forest is an ensemble of Decision Trees and visualizing every sequence is not suitable to human intelligence. Deep Learning or Neural Network is a massive number of links by addition and multiplication with non-linearities. It is hard to keep track of the relevant computations.

Thus, why interpretable models are barely used? The complexity introduced in Machine Learning Models has consequently increased the performances in most domains. Therefore, a trade-off depending on the application has then occurred: Accuracy vs. Interpretability.

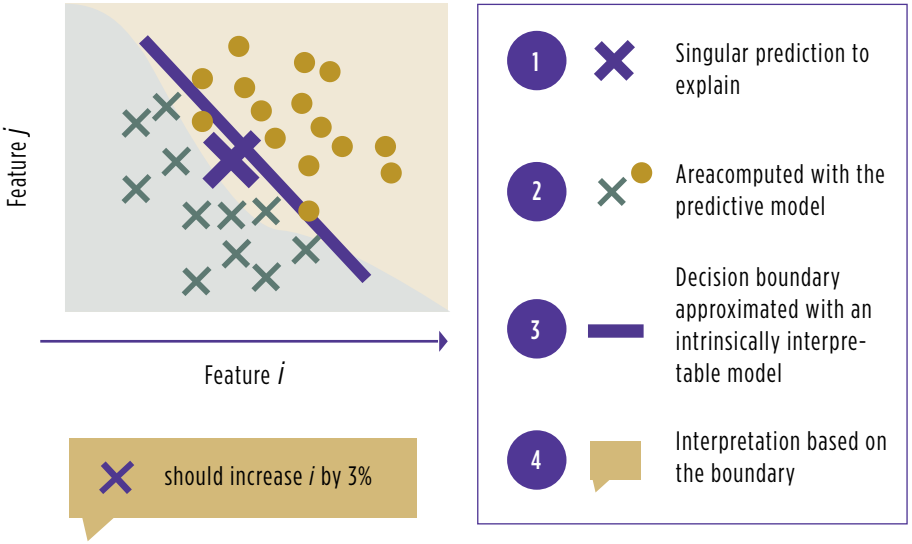
IS IT NECESSARY TO CHOOSE BETWEEN STABILITY OF RESULTS, CALCULATION TIME AND SPECIALIZATION OF THE ALGORITHM?

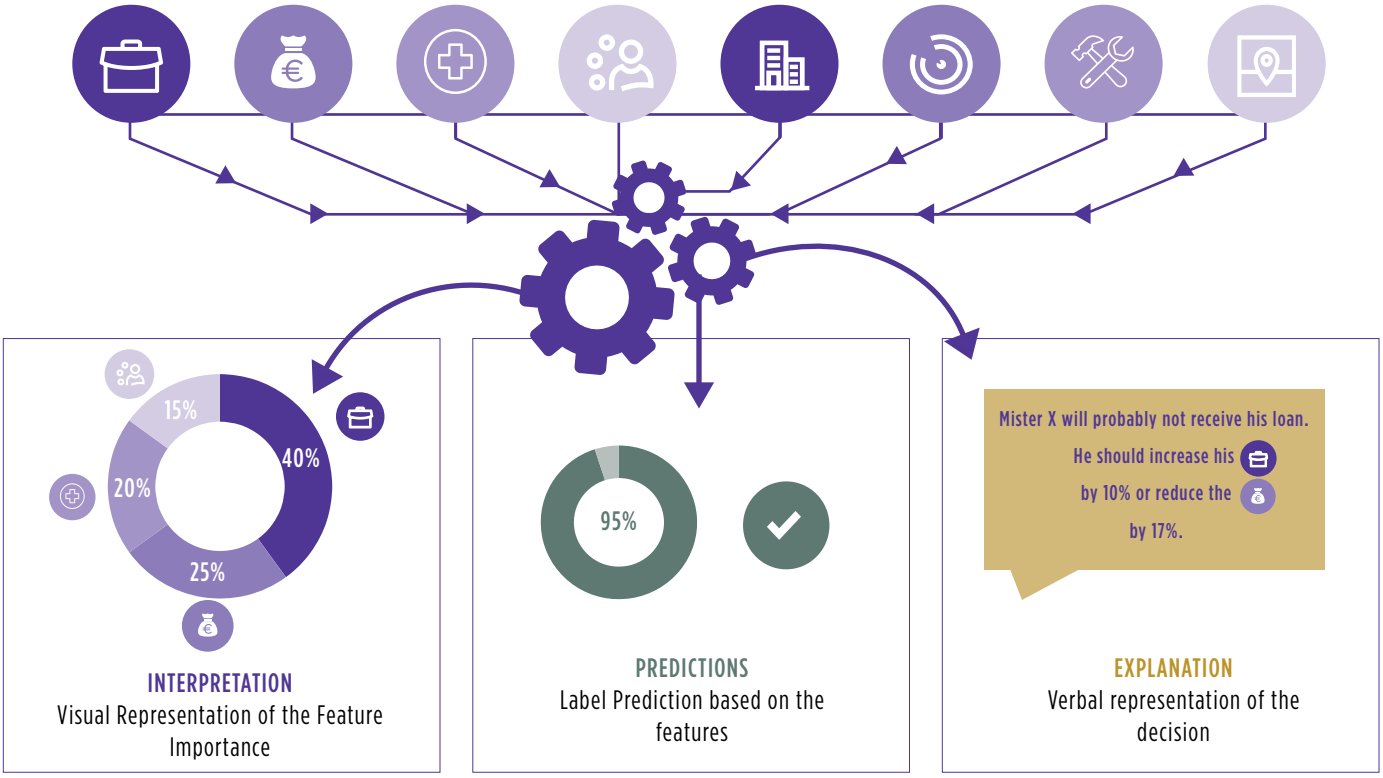
Even for high accuracy algorithms, there are methods to obtain either the interpretation or the explanation. However, there is not one single method which can be applied safely with a stable result to every machine learning model. Every method has its pros and cons.

A new trade-off occurs between the stability of results, the computation time and how specialized the algorithm is.

"Tree-Specific Feature Importance", for instance, is a method deriving from Information Theory. It evaluates how well every feature splits the data to make the decision and thus produce a global interpretation. It is fast to compute but suffers from instability and only works with Decision Trees and therefore Random Forest. The "SHAP method" on the other hand, is longer to compute and is very stable. Based on Game Theory, it evaluates how every group of features affect the results and then gives an interpretation and an explanation for every model. Another example, "LIME" can approximate a single prediction with a simple highly interpretable geometric model. It is quick to compute. However, the area of a single prediction is ill-defined and thus, the result may greatly vary.

LIME Method





AS SIMPLE AS "INTERPRETABILITY"?

The first step is to determine the actual need for Interpretability. For instance, Quantitative Analysts have a legal obligation to use highly interpretable models are thus stuck to a specific range of highly interpretable algorithms. On the contrary, Postal sort has no legal nor significant effect on the subject. Thus, complex Deep Neural Network complete the decision task. In any case, the need for interpretability is variable and to be assessed. For instance, at Wavestone, half of the recent projects required Interpretability according to the GDPR.

Then the interpretation method should be chosen depending on the desired outcome – Interpretability or Explainability – and the current implementation of the Machine Learning algorithm.

Once the method is chosen, it is important to apply it carefully mainly due to the lack of fidelity of some methods. Finally, the interpretation and the explanation can be represented in many ways. Ones may prefer to have a table to ease the automated processing, another may prefer a sentence to be sent directly to a customer and another may want a visual representation.

However, both interpretation and explanation are often based on a certain dataset, to a specific area or a part of the data space. Misinterpretation can be easy. Some interpretation methods miss the correlations between features or only offer one counterfactual explanation where multiple ones could have been given. Despite these limitations, the tools are sufficiently powerful to produce a GDPR compliant Interpretability.

INTERPRETABILITY: THE 2020 CHALLENGE FOR CIOS

Nevertheless, even without the GDPR regulation, the interpretation of a complex learning algorithms helps to optimize the model overall. In a loan attribution problem, if the important features are determined, it is possible to directly track customers who fits important features, thus optimizing client acquisition and marketing costs.

The analysis of the requirement for interpretability and the optimization of problem through the Interpretability methods are soon to be a basic step of any Machine Learning use-case.