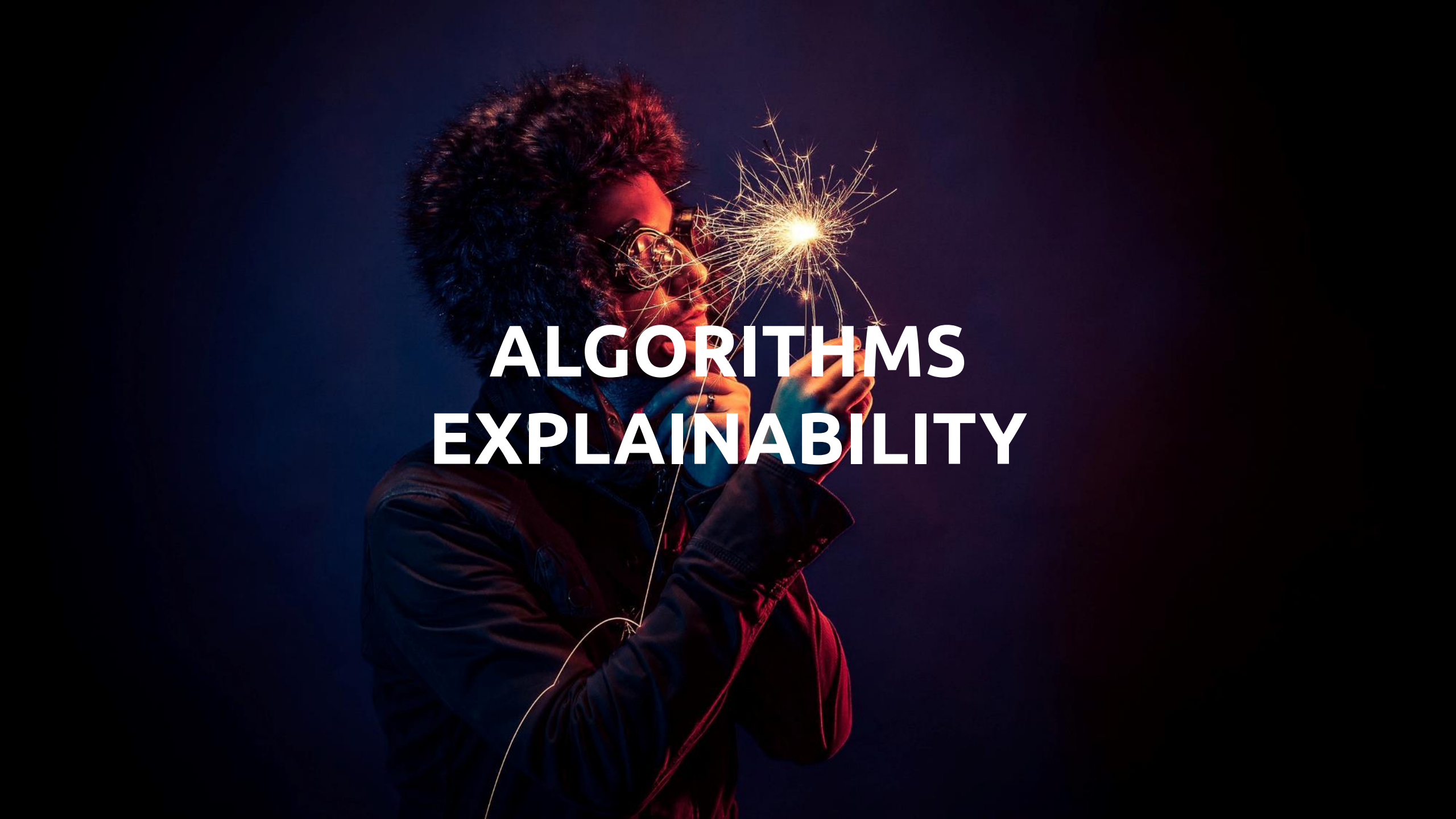


ALGORITHMS EXPLAINABILITY

Survey

01

Have you already heard of the explainability of Machine Learning algorithms ?

02

Do you know why it can be interesting to use explainability methods ?

03

Do you know the difference between local and global explainability ?

04

Do you know some explainability methods or techniques ?

AGENDA

01 CHAPTER 1: AN OVERVIEW OF MACHINE LEARNING ALGORITHMS EXPLAINABILITY

- The beginning of explainability
- Explainability is used in multiple sectors
- Explainability brings value to different stakeholders
- Trade off Performance/Explainability
- Local vs global explainability
- Illustrative Use Cases

02 CHAPTER 2: DEEP DIVE ON EXPLAINABILITY METHODS

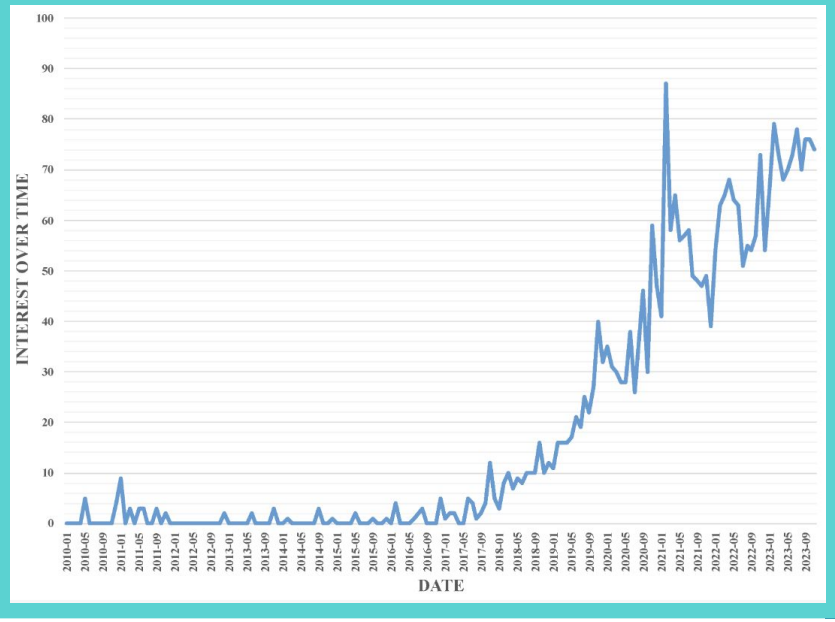
- LIME
  - For tabular
  - For text
  - For images
- SHAP for tabular

Explainability is still in its early stages

- Explainability is the degree to which a **human** can understand the cause of a **decision** (Tim Miller, 2017).
- One model is more easily interpretable than another if its decisions are easier to understand for one human than the decisions of another.

Explainability field has gained substantial importance in the recent years

XAI popularity evolution

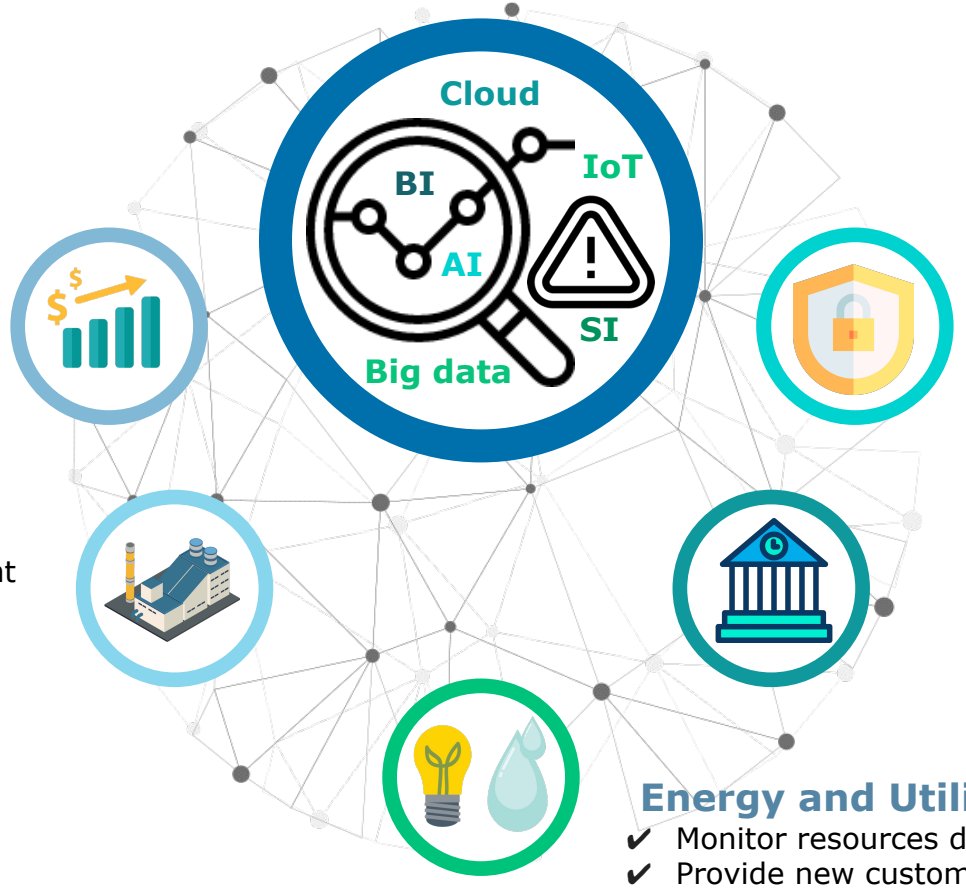


- With the growing importance of understanding Machine Learning models decision making, **scientific efforts** have **exponentially grown** over the last few years
- However, in the upcoming years, the scientific community should **converge towards more advanced and mature explainability methods**
- As such, the research efforts must be maintained and reinforced to provide further relevant and specific explanations for ML models' growing complexity
- The **current explainability state of art** already provides **essential insights** to understand why and how an IA model makes a given prediction
- These explanations must be analyzed **jointly by Business Experts and Data Scientists** to ensure their relevance and quality

Source: Explainable AI in Manufacturing and Industrial Cyber–Physical Systems: A Survey
https://www.mdpi.com/2079-9292/13/17/3497

Explainability is now a key enabler for digital companies of every sector

Making a prediction is not often the end story for a company.
Explainability is the starting point of answering several challenges in the following sectors



Financial Services

- ✔ Decide whether to issue someone a credit card or a loan
- ✔ Detect credit card frauds
- ✔ Provide Anti Money Laundering capabilities

Manufacturing

- ✔ Monitor machine sensors for intelligent maintenance
- ✔ Predict delays in deliveries
- ✔ Quality Inspection with cameras
- ✔ Drive with autonomous vehicles

IT & Cybersecurity

- ✔ Detect unknown attacks on networks
- ✔ Improve IT support by flagging anomalous tickets

Public Sector

- ✔ Help to detect cancerous lesions
- ✔ Prevent criminal recidivism
- ✔ Monitor public infrastructure

Energy and Utilities

- ✔ Monitor resources distribution network (electricity, water, gas)
- ✔ Provide new customer services such as consumption alerts

The importance of explainability varies depending on impacted stakeholders' concerns

DATA SCIENTISTS

Research & enhance models

- Explainability can be used to better understand the model from a "scientific" point of view to improve the product's efficiency & research new functionalities

Use cases illustrations

- Computer vision applications (object detection, face recognition, …)
- Natural Language Processing sentiment analysis, text classification…)

BUSINESS EXPERTS

Improve Business insights

- Explainability may be needed to enrich AI decisions with insights on how the decision was taken, allowing the identification of new business drivers

Use cases illustrations

- Credit allocation
- Virus detection in cybersecurity
- Fraud detection
- Churn detection
- Predict sales volume
- Predictive maintenance

INDIVIDUALS

(Clients, Employees,…)

Protect individual rights

- Whenever an AI model has an impact on people, explainability is required to ensure relevancy of outputs and transparency with final users

Use cases illustrations

- Credit allocation
- Recruitment process decision
- Automated school essays grading

REGULATORS

Answer legal requirements

- Complying with existing laws and regulations such as BASEL III may require to explain, trace and document complete decision-making processes, including when parts of these processes are automatized through artificial intelligence

Use cases illustrations

- Anti Money Laundering & Counter terrorism financing monitoring (transaction release)
- Credit risk modeling

- The required level of explainability should be identified for each project as soon as the design phase following the stakeholders' concerns. It will allow the selection of a fitting model for the project
- As such, it is highly recommended to carefully analyze the tradeoff between explainability and performance as soon as the design phase and implement the most fitting machine learning model

The performance vs. explainability trade-off should be decided according to the application domain and the targeted users

Neural network

Ensemble Methods
(Xgboost, Random Forest)

Characteristics of Highly accurate models

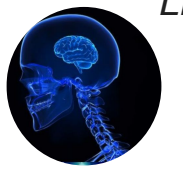
- Non-linear relationships
- Non-smooth relationships
- Long computation Times

SVM, KNN ..

Characteristics of Highly interpretable models

- linear & smooth
- Well-defined relationship
- Easy to compute

Decision Trees



Linear regression

High

Performance

Low

Low

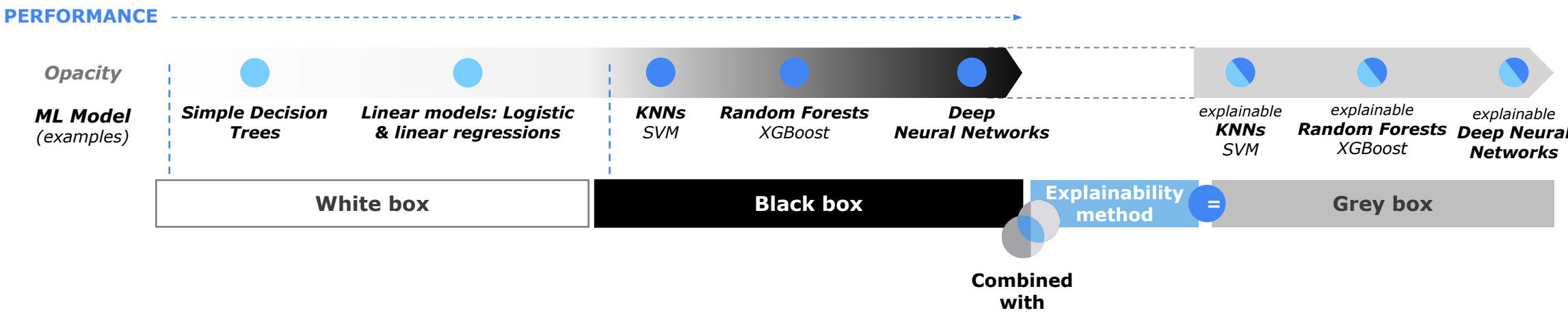
Explainability

High

Explainability is tightly linked to the complexity of the algorithm

A tradeoff between performance & explainability

- To match a growing need for highly performing Machine Learning (ML) models to answer business needs, algorithms are becoming more complex.
- This leads to a **decrease in the ability** of both Data Scientists and other stakeholders to **understand why and how the model led to a specific prediction**



As such, Machine Learning models can be classified in **3 categories**:

- **White Boxes –** A completely transparent model where the process through which inputs have been transformed into outputs is comprehensible
- **Black boxes –** A model that does not have an apparent structure; only inputs and outputs can be observed without visibility on the process
- **Grey boxes –** A "between the two" model where black boxes' decision process is partially explained using specific explainability methods (XAI)

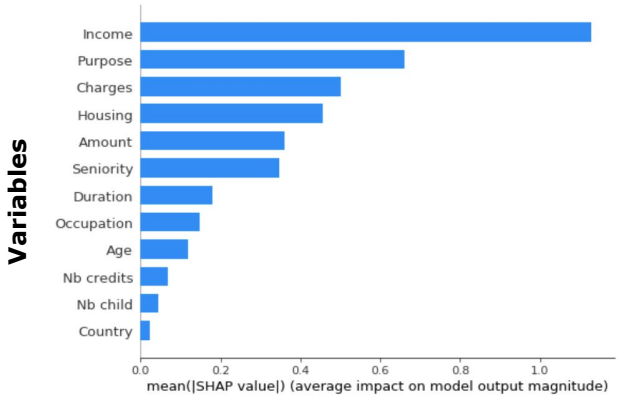
Explainability can be approached in two ways: Globally and Locally

- **Explainable AI** or XAI is an **emerging research field** aiming at developing methods to **interpret complex machine learning models' decision making**
- As of today, there are two approaches towards achieving machine learning explainability:

GLOBAL EXPLAINABILITY

Attempts to understand the high-level concepts and reasoning used by a model by analyzing feature importance globally (analyzing average contribution of each variable based on their contribution to a large number of individual outputs)

Illustration of a global explainability – Display of each variable's average contribution to the overall model

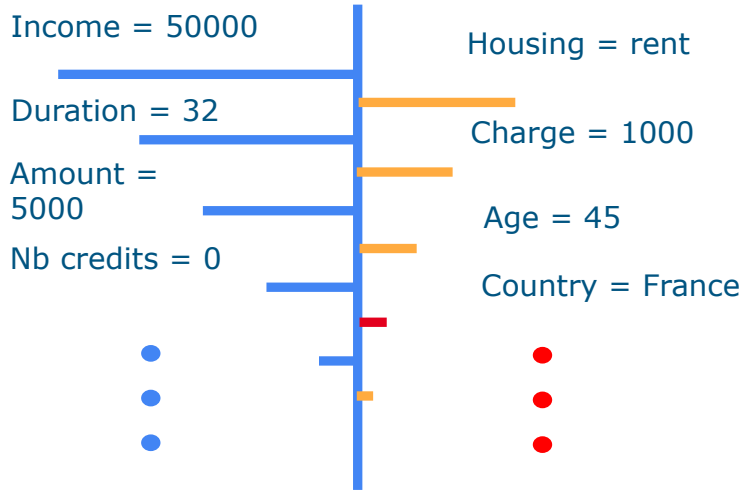


Average impact on model outcome

Based on SHAP method

LOCAL EXPLAINABILITY

Aims to explain the model's behavior for a specific input by analyzing feature importance of determined outputs



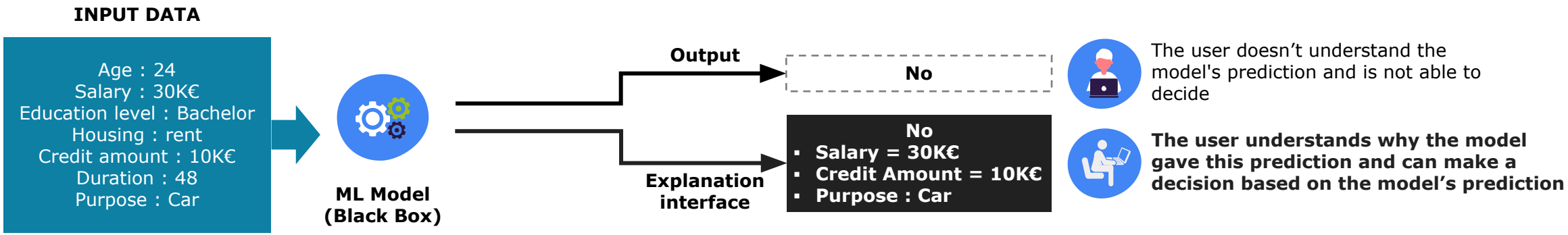
Credit score = 0.99

Illustrative use case of local explainability

Basic example : Credit scoring

- The Machine Learning model analyzes the data and provides a prediction
- Consider a case where, the subject is 24 years old with a 30K€ salary requesting a credit of 10K€ to buy a car
- The ML models gives a negative response but doesn't explain the underlying reasons
- Explainability techniques enable the user to identify the variables that influence the ML model's prediction: e.g. the salary is equal to 30K€, the credit amount equals 1/3 of salary and the purpose of the loan (car)

By using the explainability model, the user is able to understand the predictions of the ML model and can decide whether to give the customer the credit or not



Industrial Process optimisation powered by explainability

Context

- A great European aircraft manufacturer has a high rate of quality issues during its pylons manufacturing process, and the plant management team has no vision on the parameters impacting the quality of their products.

Objective

- Identify the Key Process Parameters (parameters impacting the products quality) and understand the best of the best and worst of the worst products process profiles.

Data

- Process Data (Temperatures, rotation speeds,…)
- Supplier Data (Tools supplier,…)
- Quality Data (Product measurements in tolerance,…)

Approach

Data gathering & preparation

Target Features

PN	is_OOT	Process Data			Supplier Data	
		D1	D2	D3	D4	D5
1	True	0,3	True	145	Sup_1	10
2	False	0,9	True	176	Sup_2	8
…	…	…	…	…	…	…

Model building & training

Model Training to predict our target thanks to our features

Model Explainability

Global Explainability for KPP Identification

Local Explainability for Best of Best Identification

Illustration on Explainability

Typical project

Context

- A great European aircraft manufacturer must perform data enrichment from in-service airline data. Today, this task is done manually and requires high level of expertise and is time consuming.

Objective

- Automatize the tedious task of classifying documents & provide classification suggestions to humans using machine learning.

Data

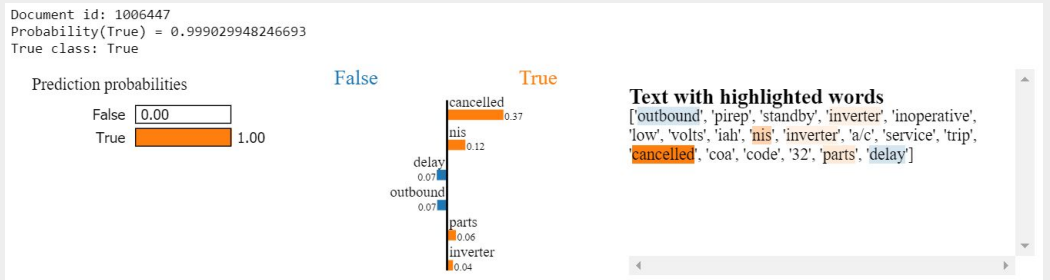
- Flight delays of more than 15 minutes

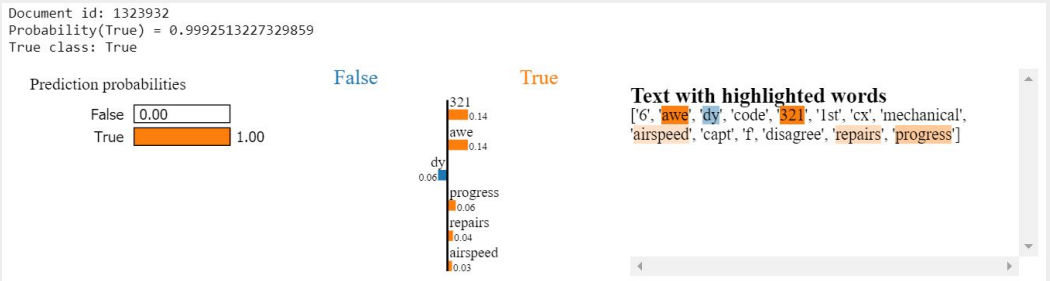
1.7M docs

hybrid data

15 targets

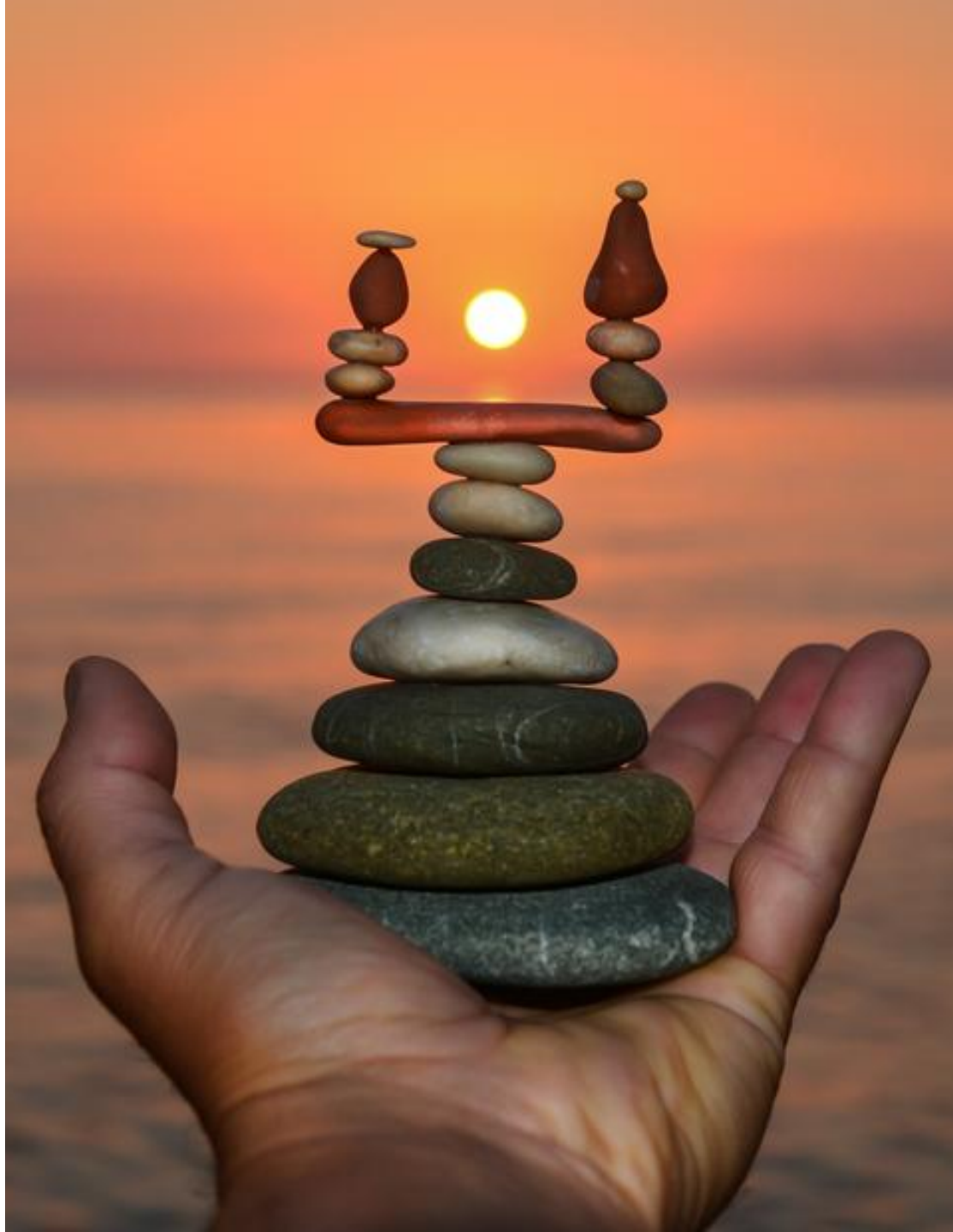
Model explainability





Key Takeaways

ML algorithms explainability is widely used (all sectors for multiple stakeholders)

Explainability methods move the lines of the trade-off between Performance & Explainability by making complex models more interpretable

ML models explanations can be global (understanding of a model behaviour as a whole) or local (understanding of a model decision in a specific case)