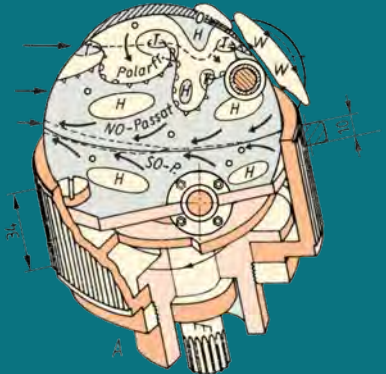
Why & How Machine Learning Models should explain themselves

Machine Learning Interpretability



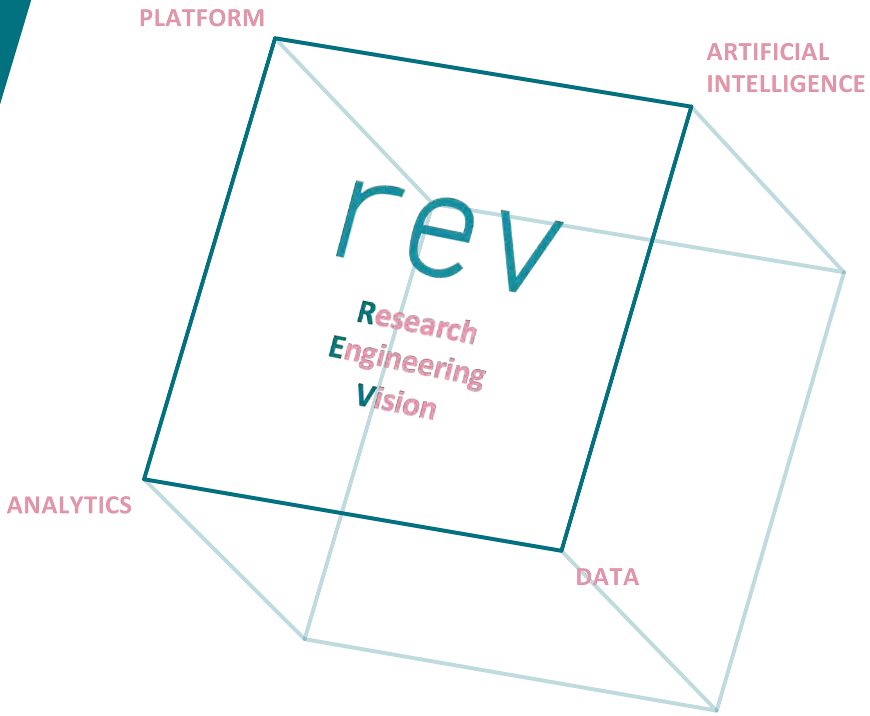
@XavierRenard | @detyniecki

AXA / GO / REV / Research & Development

rev was built to make AXA a tech-led company

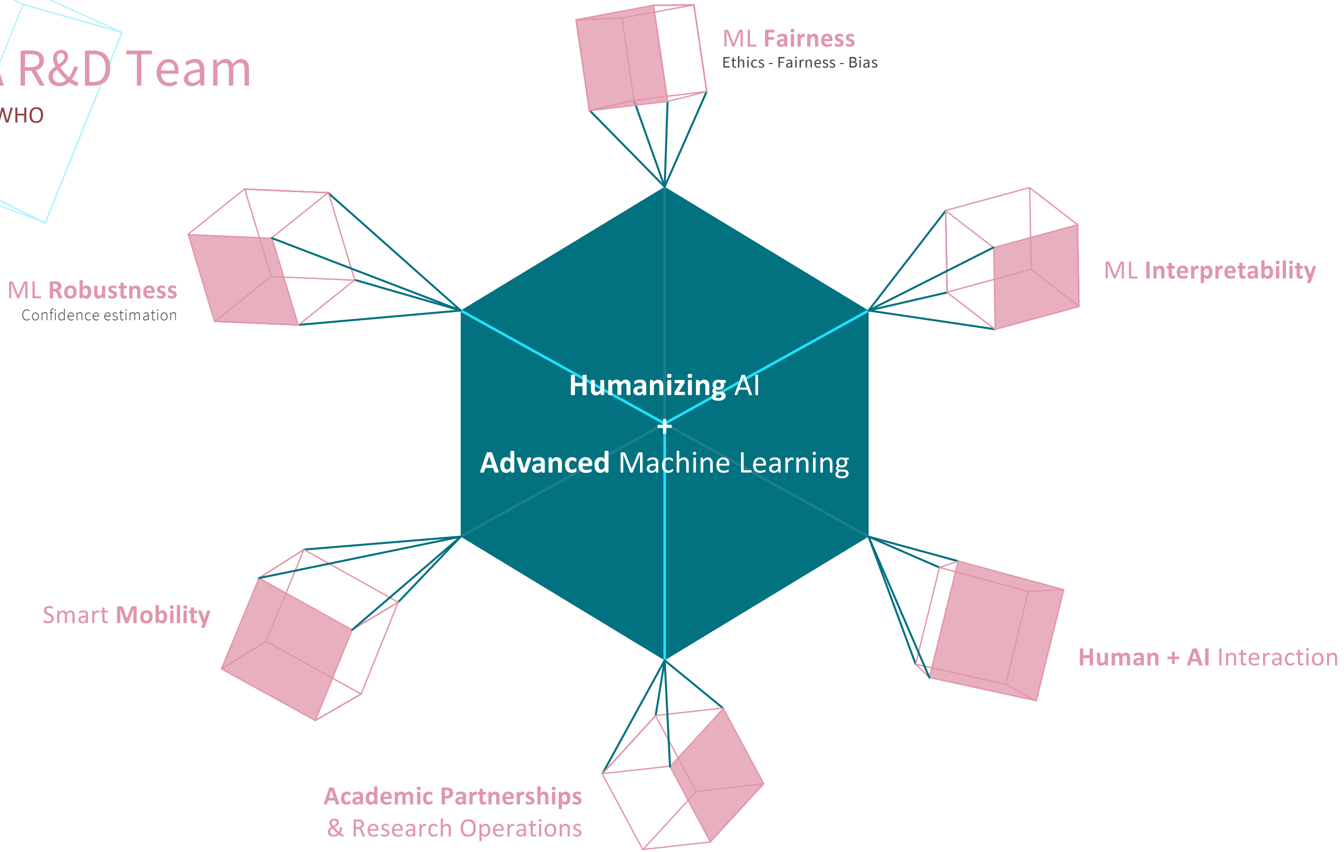
OUR MISSION

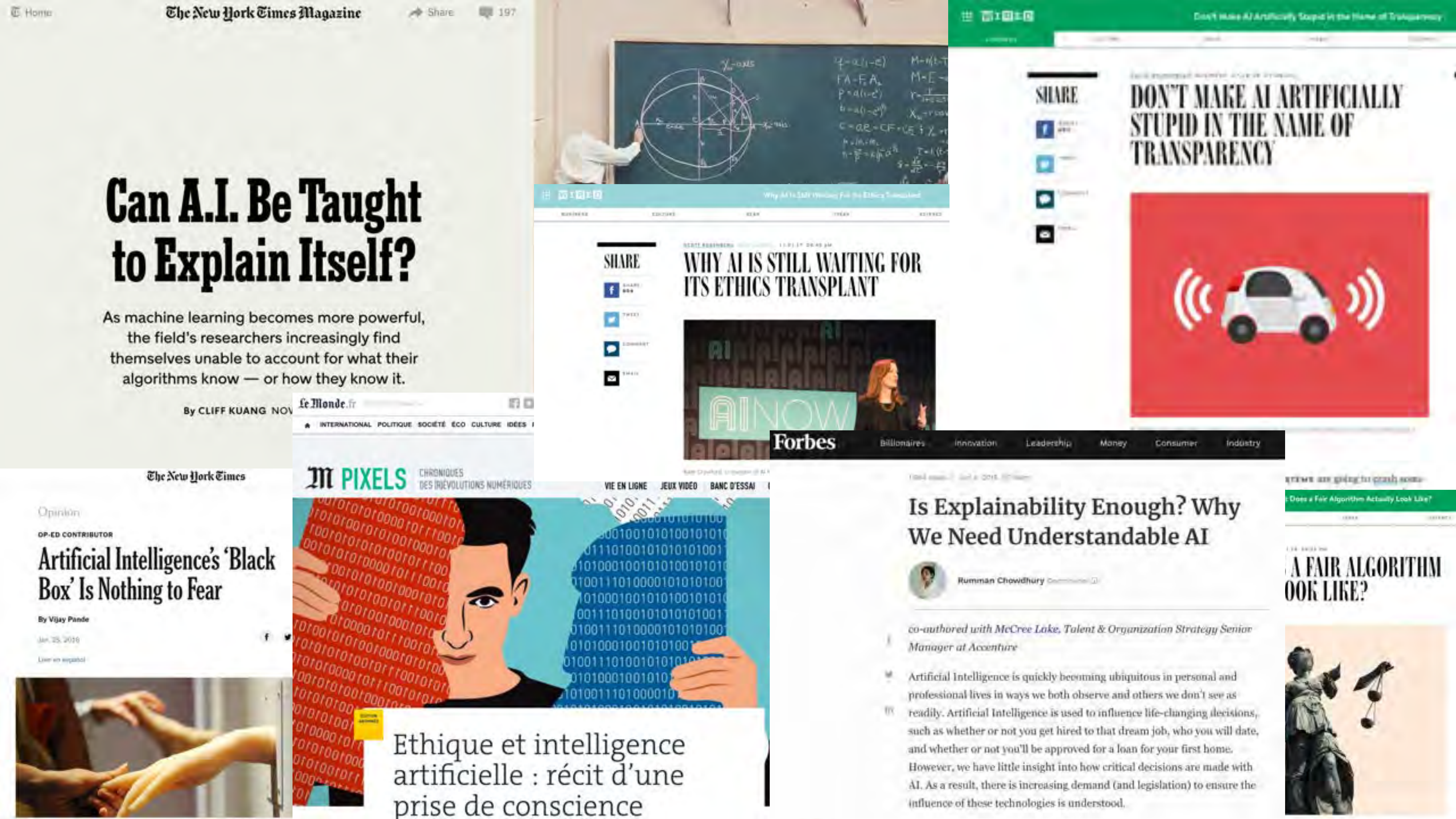
AXA rev (Research, Engineering, & Vision)
explores and scales the **value of data**
and **emerging technologies** with the potential
to **disrupt** the current
insurance business model and
to **shape future opportunities**
in order to be a **better partner**
in our customer’s lives.



AXA R&D Team

WHO’S WHO





The New York Times Magazine

Home Share 197

Can A.I. Be Taught to Explain Itself?

As machine learning becomes more powerful, the field's researchers increasingly find themselves unable to account for what their algorithms know — or how they know it.

By CLIFF KUANG NOV

WIRED

SHARE

WHY AI IS STILL WAITING FOR ITS ETHICS TRANSPLANT

WIRED

SHARE

DON'T MAKE AI ARTIFICIALLY STUPID IN THE NAME OF TRANSPARENCY

The New York Times

Opinion

OP-ED CONTRIBUTOR

Artificial Intelligence's 'Black Box' Is Nothing to Fear

By Vijay Pande

Le Monde.fr

INTERNATIONAL POLITIQUE SOCIÉTÉ ÉCO CULTURE IDÉES

PIXELS CHRONIQUES DES (R)ÉVOLUTIONS NUMÉRIQUES

VIE EN LIGNE JEUX VIDÉO BANC D'ESSAI

Ethique et intelligence artificielle : récit d'une prise de conscience

Forbes

Billionaires Innovation Leadership Money Consumer Industry

Is Explainability Enough? Why We Need Understandable AI

Rumman Chowdhury

co-authored with McCree Lake, Talent & Organization Strategy Senior Manager at Accenture

Artificial Intelligence is quickly becoming ubiquitous in personal and professional lives in ways we both observe and others we don't see as readily. Artificial Intelligence is used to influence life-changing decisions, such as whether or not you get hired to that dream job, who you will date, and whether or not you'll be approved for a loan for your first home. However, we have little insight into how critical decisions are made with AI. As a result, there is increasing demand (and legislation) to ensure the influence of these technologies is understood.

A FAIR ALGORITHM OOK LIKE?

Machine Learning Interpretability Impacts the Business

Improve Model's Quality
- Improve models, features, robustness, fairness, etc.
- Identify data leakage & data drift
- *e.g. Understand origin of wrong predictions*

Reassure Users & Business Owners
- Trust by **explanation**: improve ML acceptance
- Help to take ML prediction-based decision
- *e.g. Assess reasonable behaviour if deployment*

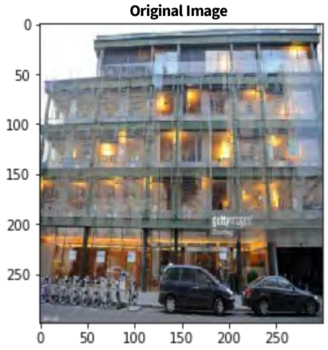
<u>Use-Case in Fraud</u>: Analysts insist to understand why there is an alert

Law & Ethics compliance
- Right to explanation
- Assess model's fairness
- Inform customers

Gain Knowledge on Business' processes
- Insight of revenues or value-generating application
- *e.g. Credit scoring, fraud detection, etc.*

Machine Learning Interpretability

APPLIED TO AXA'S HEADQUARTERS



Classification

With InceptionV3:

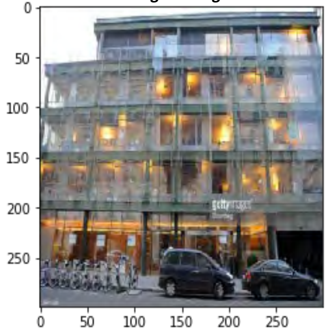
Most probable labels:

- **Building**
- **Minivan**
- **Traffic light**

Machine Learning Interpretability

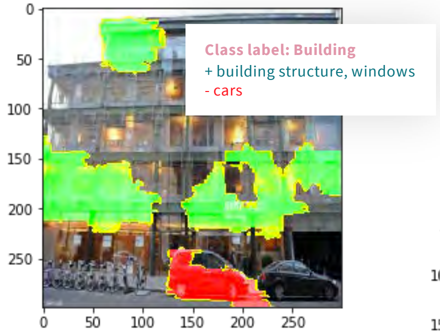
APPLIED TO AXA'S HEADQUARTERS

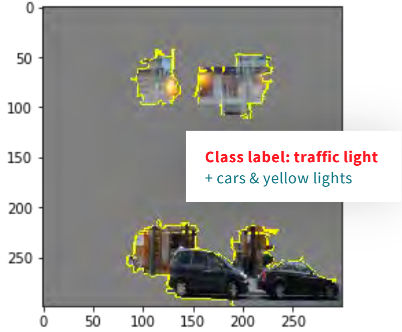
Original Image

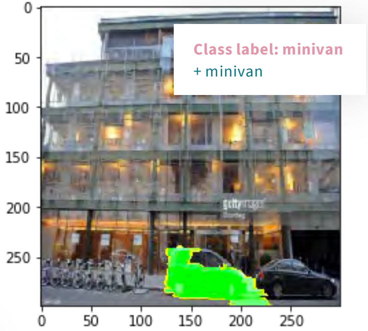


Classification

(With InceptionV3)







Machine Learning Interpretability

EVALUATE MACHINE LEARNING MODELS BEYOND ACCURACY SCORES

What has been learned by the model?

Where is the model {correct ; wrong} ?

Why a particular prediction has been made?

What can be done to change the prediction?

x

Description of the problem to solve
Tabular data, unstructured data, etc.

Machine Learning Model

y

Prediction / Decision

Usually aggregated accuracy score

Is the model robust?

Is the model fair?

How does the model behave in areas with few data?

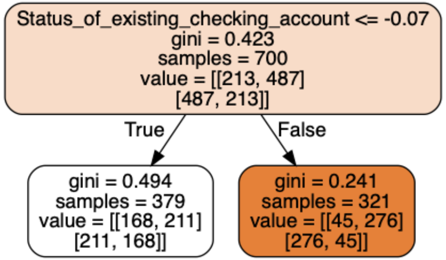
Is the model causal?

How the prediction is affected by small changes in input?

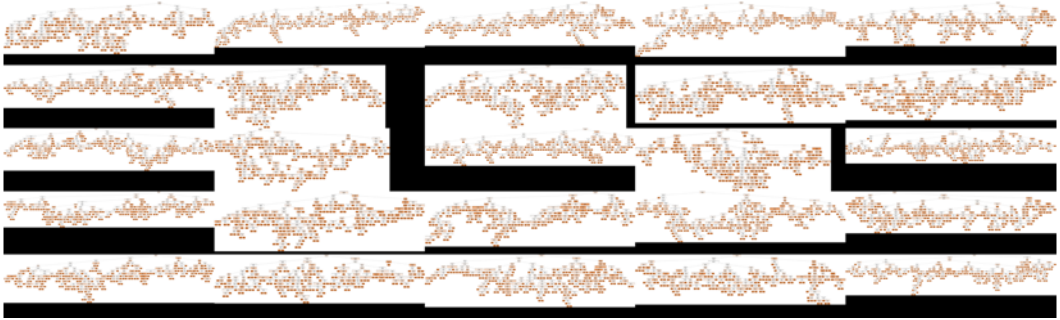
Trade-off Interpretability-Accuracy

Accurate Machine Learning Models are **not Interpretable** (usually)

Simple machine learning model
e.g. Decision Tree



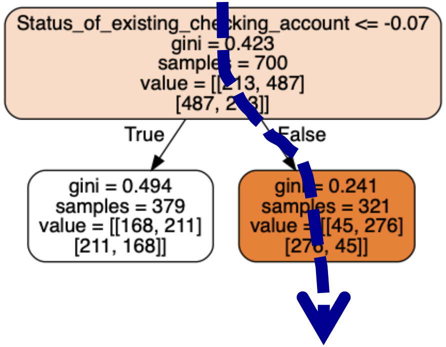
Blackbox machine learning model
e.g. Random Forest, CNN (Inception…)



Trade-off Interpretability-Accuracy

Accurate Machine Learning Models are **not Interpretable** (usually)

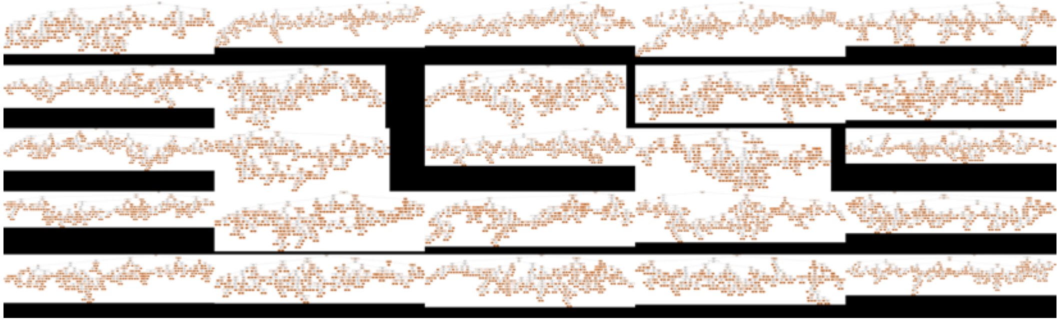
Simple machine learning model
e.g. Decision Tree



Decision: credit or not

One **path** → simple **explanation**

Blackbox machine learning model
e.g. Random Forest, CNN (Inception…)

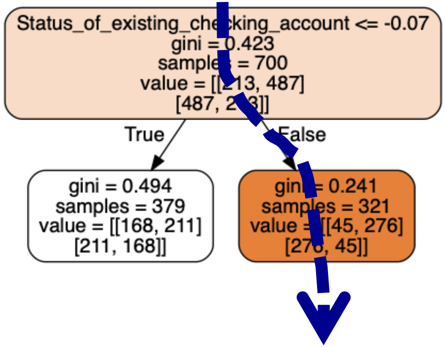


Trade-off Interpretability-Accuracy

Accurate Machine Learning Models are **not Interpretable** (usually)

Simple machine learning model

e.g. Decision Tree

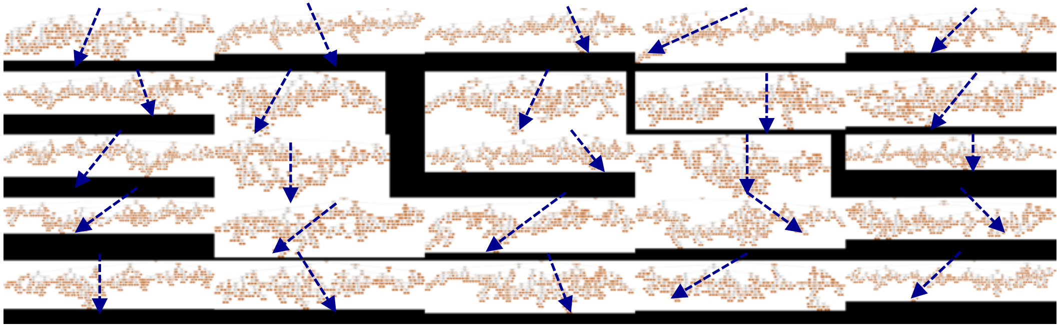


Decision: credit or not

One **path** → simple **explanation**

Blackbox machine learning model

e.g. Random Forest, CNN (Inception…)



One path → One decision **by** base model

Final decision: <u>aggregation</u> of each decision

Explanation: <u>no consensus</u>

Trade-off Interpretability-Accuracy

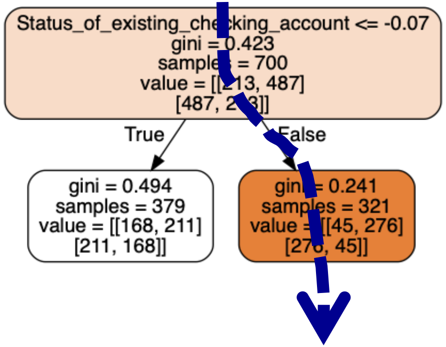
Accurate Machine Learning Models are **not Interpretable** (usually)

Simple machine learning model

e.g. Decision Tree

→ Interpretable

→ Less accurate



Decision: credit or not

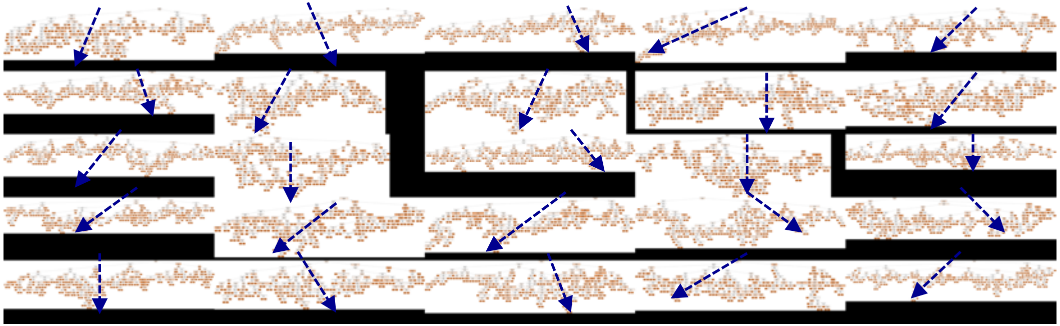
One **path** → simple **explanation**

Blackbox machine learning model

e.g. Random Forest, CNN (Inception…)

→ Uninterpretable

→ More Accurate

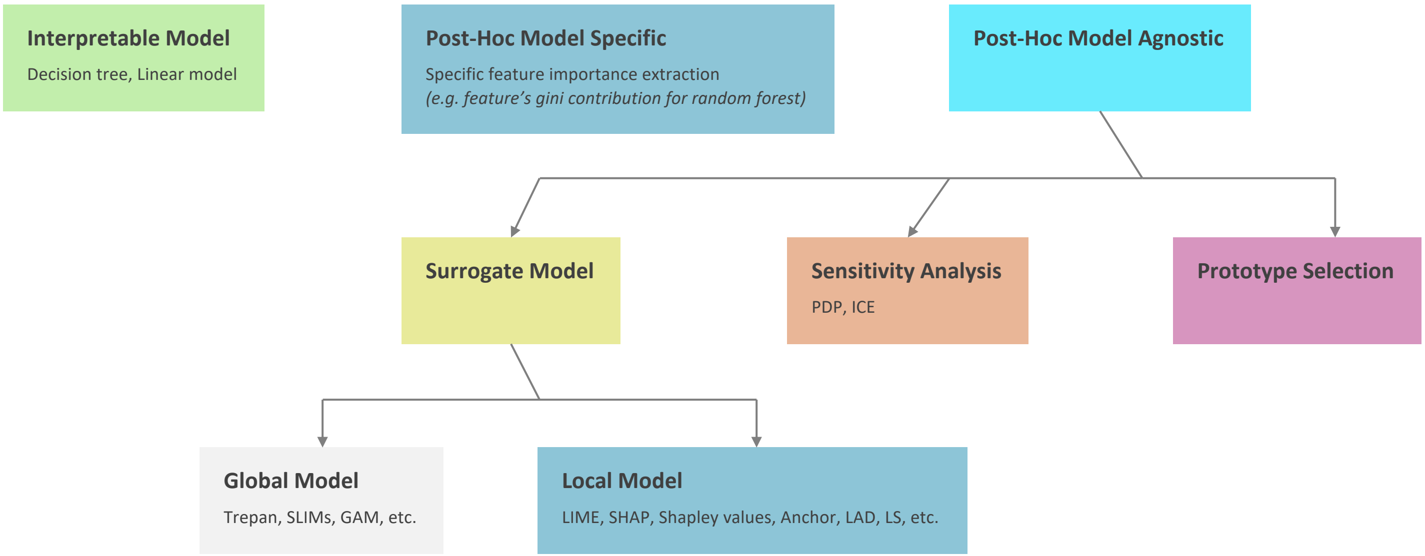


One path → One decision **by** base model

Final decision: aggregation of each decision

Explanation: no consensus

Taxonomy of Interpretability Approaches



Locality Issue *(2018 ICML WHI)*

A widely used approach -LIME- is inaccurate

Defining Locality for Surrogates in Post-hoc Interpretablity

Thibault Laugel [1,*] Xavier Renard [2,*] Marie-Jeanne Lesot [1] Christophe Marsala [1] Marcin Detyniecki [1,2,3]

Abstract

Local surrogate models, to approximate the local decision boundary of a black-box classifier, constitute one approach to generate explanations for the rationale behind an individual prediction made by the back-box. This paper highlights the importance of defining the right *locality*, the neighborhood on which a local surrogate is trained, in order to approximate accurately the *local* black-box decision boundary. Unfortunately, as shown in this paper, this issue is not only a parameter or sampling distribution challenge and has a major impact on the *relevance* and quality of the approximation of the local black-box decision boundary and thus on the *meaning* and accuracy of the generated explanation. To overcome the identified problems, quantified with an adapted measure and procedure, we propose to generate surrogate-based explanations for individual predictions based on a sampling centered on particular place of the decision boundary, relevant for the prediction to be explained, rather than on the prediction itself as it is classically done. We evaluate the novel approach compared to state-of-the-art methods and a straightforward improvement, on four UCI datasets.

1. Introduction

The task of explaining individual predictions made by a black-box classifier aims at providing to a human user the rationale, or at least intuitions, about the factors leading to this prediction and, eventually, showing how the prediction can be altered by changing some of these factors (Doshi-Velez et al., 2017). While it is clear in the current literature that there are still no consensual definitions of "explanations" and "interpretability" for machine learning algorithms and their predictions, we can nevertheless state that there is a consensus to say that providing an explanation for an individual prediction relies on finding the features that actually impact the prediction.

To do so, several types of approaches have been proposed. In this paper we consider *post-hoc* approaches, which are model-agnostic and classically applied to trained machine learning prediction models. For instance, sensitivity analysis (Simonyan et al., 2013; Adler et al., 2018; Koh & Liang, 2017) generates perturbations on the feature values of the instance of the prediction to explain in order to observe the consequences on the output of the black-box to outline the local behaviour of the black-box regarding its local decision boundary. Surrogate models (Craven & Shavlik, 1996; Hara & Hayashi, 2016), on the other hand, train an interpretable model (e.g. linear regression or decision tree with low complexity) to mimic the black-box decisions in order to extract explanations from it. A particular case are the local surrogates (Ribeiro et al., 2016) which, in order to better locally approximate the black-box decision boundary, propose to train an interpretable model locally in the neighborhood of the instance whose prediction by the black-box is to explain.

In this paper, we refine the notion of *locality*, the neighborhood on which a local surrogate is trained and show that it is not trivial to define the right neighborhood. We illustrate this issue using LIME (Ribeiro et al., 2016): we show that choosing an adequate sampling strategy for generating the instances used to fit the surrogate model has a major impact on the quality of the approximation of the local black-box decision boundary and thus on the accuracy of the generated explanation. In particular, the effect of locally important features can be hidden by globally important ones. We show that this issue is not related to a simple parametrization to control the range of the sampled neighborhood. For instance, centering the sampling on the instance of the prediction to explain may not be the best location to approximate the black-box decision boundary. To solve this issue, we propose a novel approach to sample the right neighborhood to fit local surrogate models. The intuition is the following: since a local surrogate aims at approximating the local black-box decision boundary that matters for the prediction to explain, this boundary should be sought first in

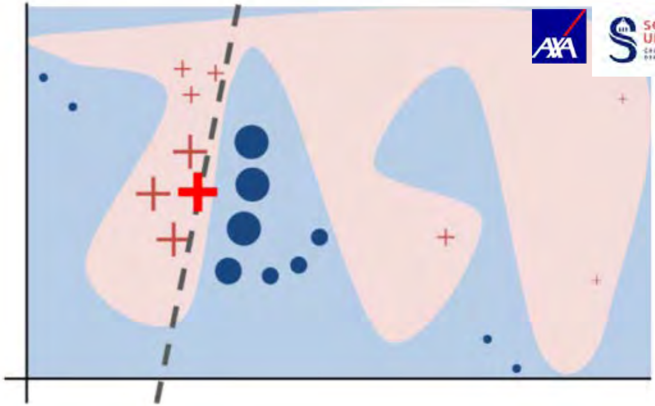
*Equal contribution [1]Sorbonne Université, CNRS, LIP6, Paris, France [2]AXA, Paris, France [3]Polish Academy of Science, IBS PAN, Warsaw, Poland. Correspondence to: Thibault Laugel <thibault.laugel@lip6.fr>, Xavier Renard <xavier.renard@axa.com>.

2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018), Stockholm, Sweden. Copyright by the author(s).

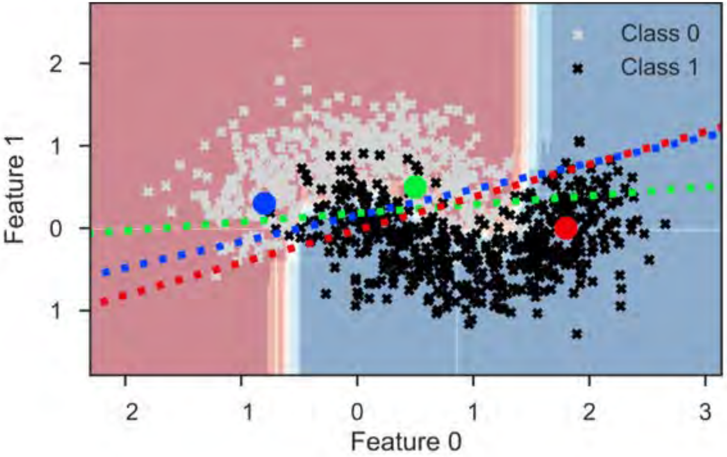
47

LIME

Expectation

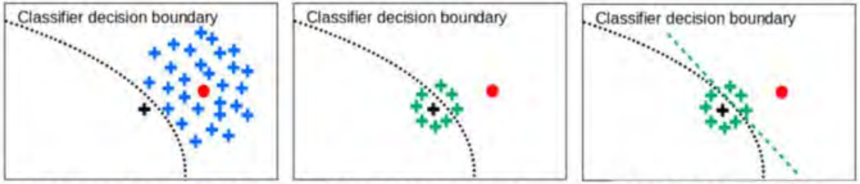


Reality

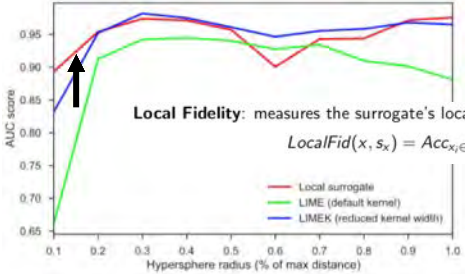


Locality Issue *(2018 ICML WHI)*

Our proposition: find the frontier first



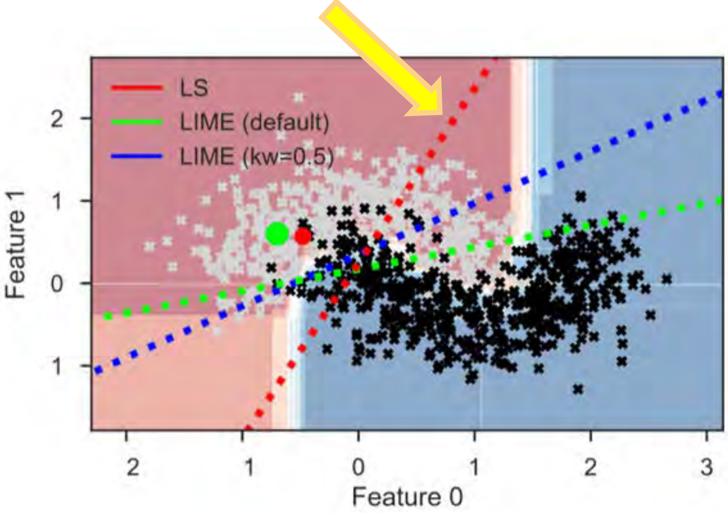
Step 0: Closest border detection **Step 1: Local sampling** **Step 2: Model training**



Local Fidelity: measures the surrogate's local accuracy

$$LocalFid(x, s_x) = Acc_{x_i \in \mathcal{V}_x}(b(x_i), s_x(x_i))$$

Better black-box frontier approximation
→ more accurate explanations



Concept Tree *(2019 ICML WHI)*

Gather Related Variables for More Interpretable (Surrogate) Decision Trees

- Global & Local explanation of a black-box classifier based on a decision tree and **concepts**
  - In the presence of correlated variables
  - Or expert-defined groups of variables

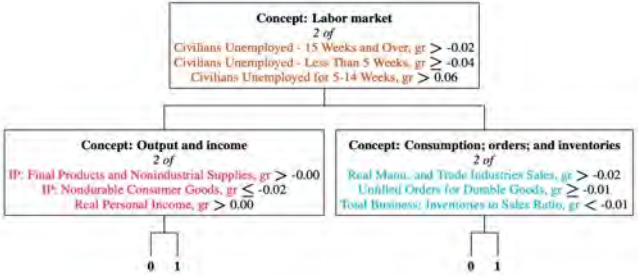
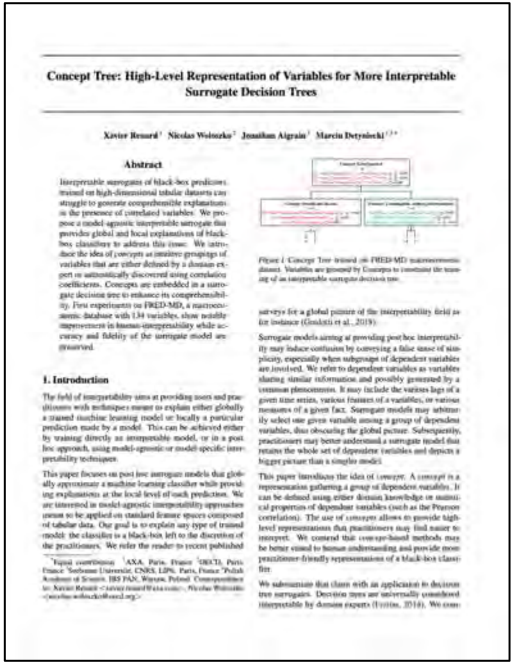


Figure 1. Concept Tree trained on FRED-MD macroeconomic dataset. Variables are grouped by Concepts to constraint the training of an interpretable surrogate decision tree



Concept Tree: High-Level Representation of Variables for More Interpretable Surrogate Decision Trees

Xavier Renard, Nicolas Woloszko, Jonathan Aigrain, Marcin Detyniecki

Abstract

1. Introduction

Unjustified Counterfactual Explanations *(2019 IJCAI)*

The Dangers of Post-hoc Interpretability: Unjustified Counterfactual Explanations

- Instance close to the original observation predicted in a different class
  - They can be a consequence of an **artifact** of the classifier
  - Unjustified by **ground truth** (training data)
  - Lack of robustness of the classifier / ood prediction
- To be **justified** a counterfactual example should be continuously connected to an instance of the training set
- Assessment procedure proposed
- Counterfactual explanation methods **vulnerable** to unjustified counterfactual examples

The Dangers of Post-hoc Interpretability: Unjustified Counterfactual Explanations

Abstract

Post-hoc interpretability approaches have been proven to be powerful tools to generate explanations for the predictions made by a trained black-box model. However, they create the risk of having explanations that are a result of some artifacts learned by the model instead of actual knowledge from the data. This paper focuses on the case of counterfactual explanations and asks whether the generated instances can be *justified*, i.e. continuously connected to some ground-truth data. We evaluate the risk of generating unjustified counterfactual examples by investigating the local neighborhoods of instances whose predictions are to be explained and show that this risk is quite high. Furthermore, we show that most state of the art ap-

shown to be useful solutions [Wachter *et al.*, 2018] that can be easily understood and thus directly facilitate decisions for a user.

However, without any guarantee on existing data, counterfactual examples in the post-hoc paradigm are vulnerable to issues raised by the robustness of the classifier, leading to explanations that are arguably not satisfying in the context of interpretability. More generally, this paper argues that a crucial property that a counterfactual example should satisfy is that it should be connected to the training data of the classifier. Formally, we define this relation using the notion of connectedness and argue that there should be a continuous path between a counterfactual and the training data, which we define in Section 3.

This paper aims at showing that generating such post-hoc counterfactual explanations can be difficult. The contribu-

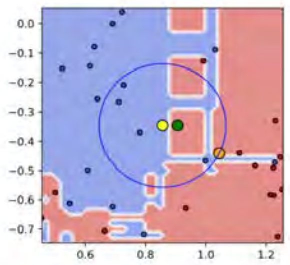


Figure 4: LRA procedure for an instance x with $S(x) = 1$

Imperceptible Adversarial Attacks on Tabular Data (on going)

ML interpretability * Adversarial ML

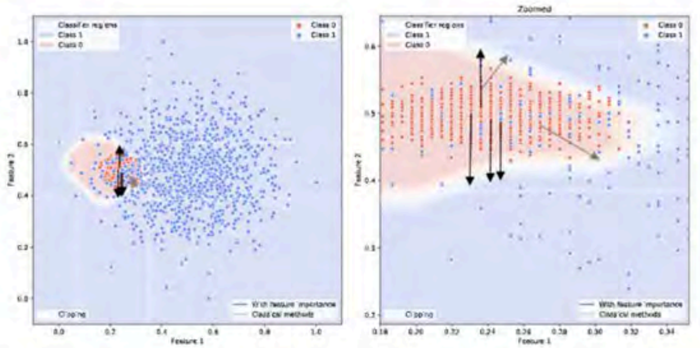
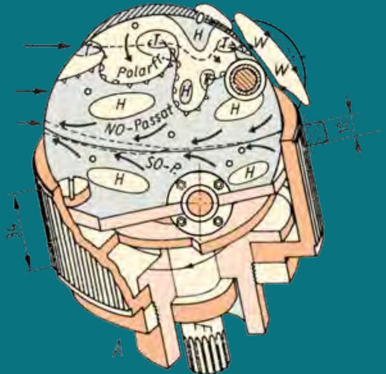


Figure 1: Intuition of adversarial samples generation using feature-importance, in black, compared to traditional methods in gray. We observe that feature-importance based perturbations follow only the Feature 2 directions.

Papers + Ressources
http://axa-rev-research.github.io

Why & How Machine Learning Models should explain themselves

Machine Learning Interpretability



@XavierRenard | @detyniecki
AXA / GO / REV / Research & Development