March 11th 2022, Mauriana Pesaresi Seminar Series

CALIME

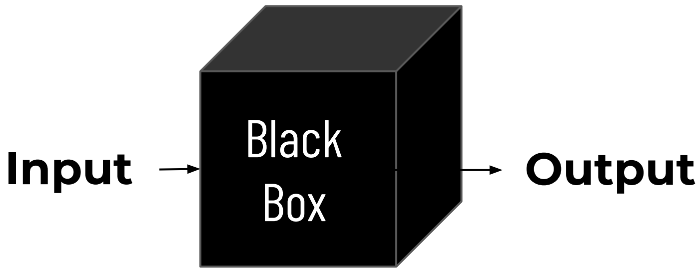
Causality-Aware Local Interpretable Model-Agnostic Explanations

Martina Cinquini
martina.cinquini@phd.unipi.it

Riccardo Guidotti
riccardo.guidotti@phd.unipi.it

What is eXplainable AI (XAI) ?

XAI provides explanations for the decisions of Machine Learning models.



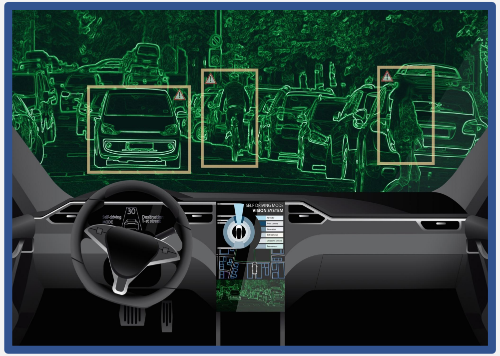
Black box models have an hidden internal structure that humans do not understand e.g. DNNs, SVMs

Source: Google Trends for "Explainable AI"

Why does XAI matter in Machine Learning?

Benefits

1. AI systems are increasingly used in sensitive areas



Self-driving cars

2. ML models can perpetuate existing bias

Racial Bias

3. Automated business decision making requires reliability and trust

Financial Services

Taxonomy

Explainable by Design

Build **interpretable** ML models

Black box Explanation

Derive explanations for **complex** ML models

Local

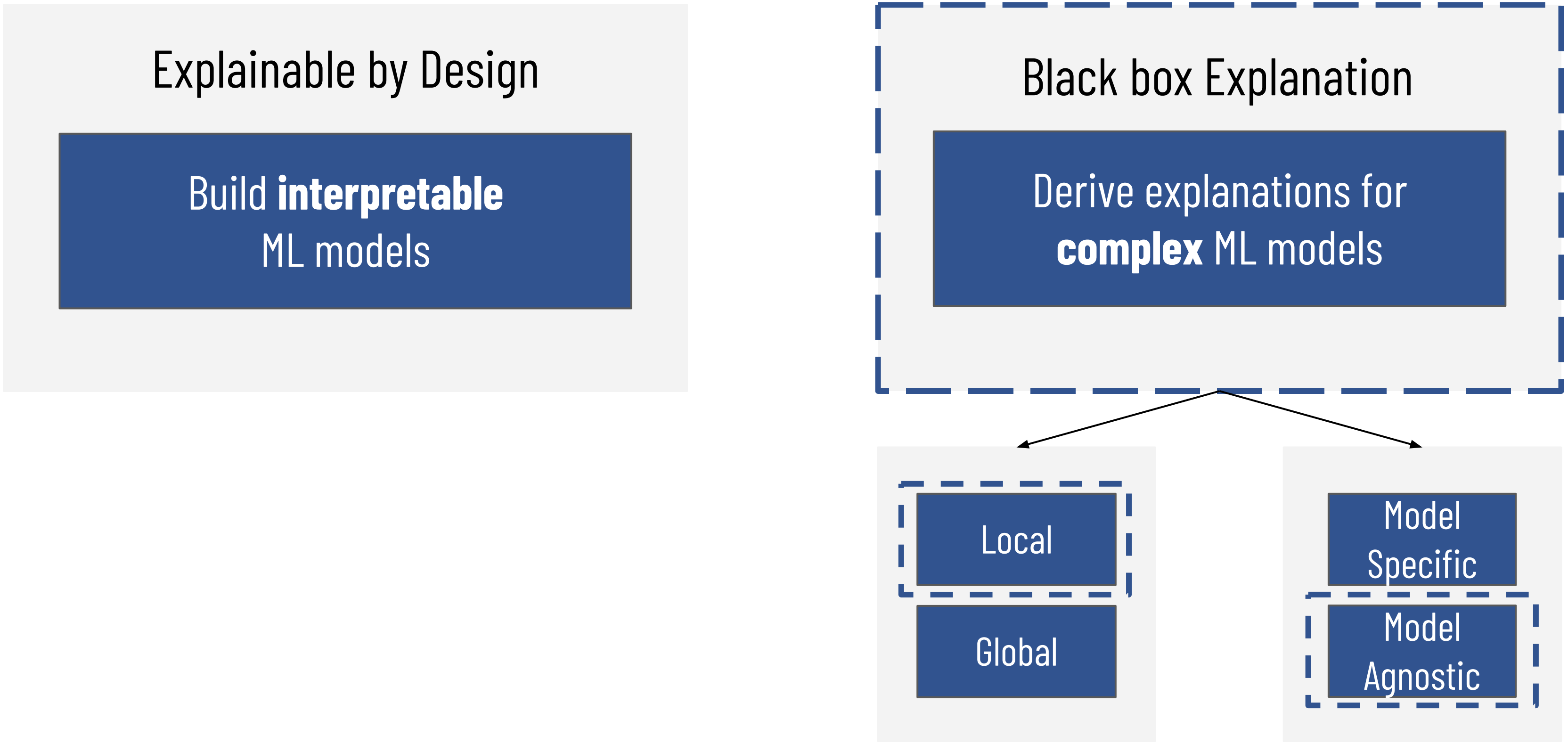
Global

Model Specific

Model Agnostic

[1] A Survey of Methods for Explaining Black Box Models, Guidotti et al., 2018

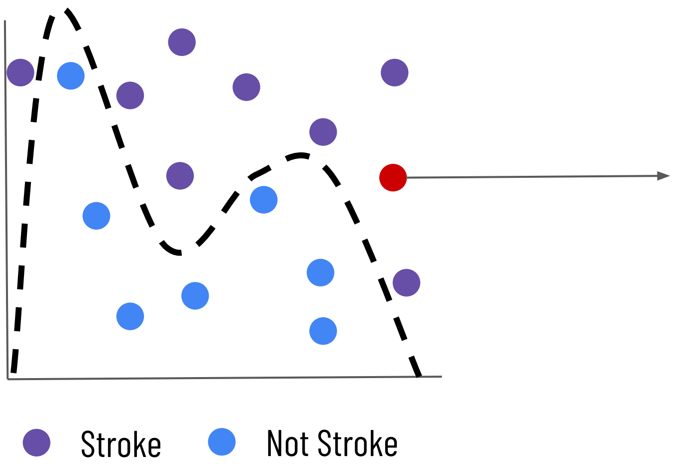
Taxonomy



[1] A Survey of Methods for Explaining Black Box Models, Guidotti et al., 2018

LIME

Local **I**nterpretable **M**odel-**A**gnostic **E**xplanations[2]



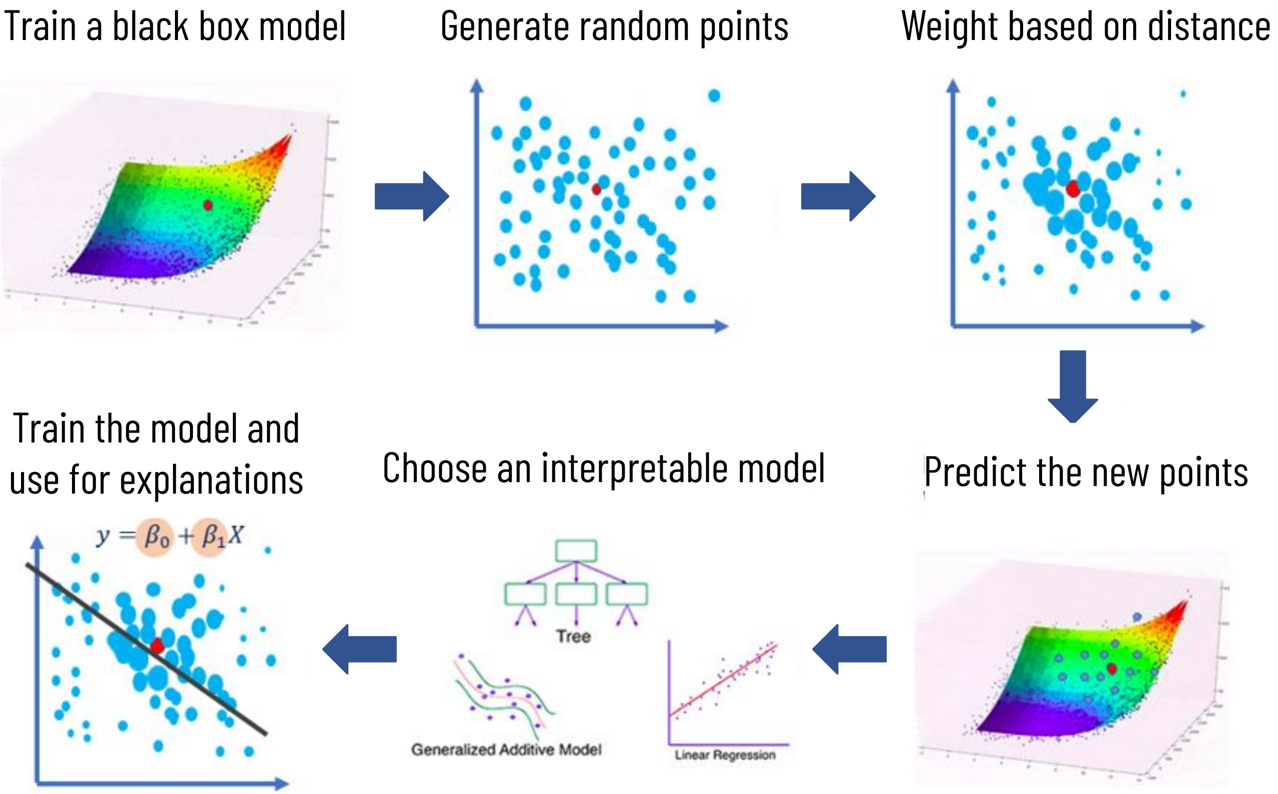
Mark

GOAL

Understand why the ML model made a certain prediction

[2] "Why should I trust you?": Explaining the Predictions of Any Classifier, Ribeiro et al., 2016
Slide example from: https://www.youtube.com/watch?v=d6j6bofhj2M

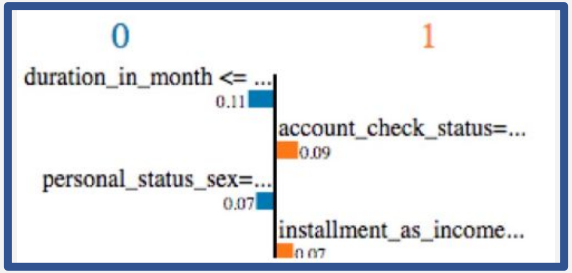
LIME


Train a black box model
Generate random points
Weight based on distance
Predict the new points
Choose an interpretable model
Tree
Generalized Additive Model
Linear Regression
Train the model and use for explanations
$y = \beta_0 + \beta_1 X$

LIME

Explanations

Feature importance



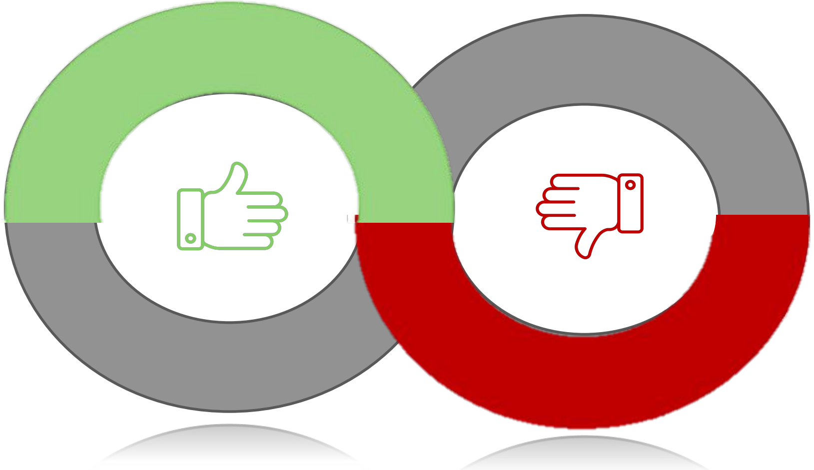
Saliency Maps

LIME

Pros & Cons

It is Model Agnostic

It works on text, images and tabular data



Instability of Explanations

Low Fidelity

It does not consider the causal relationships among input features

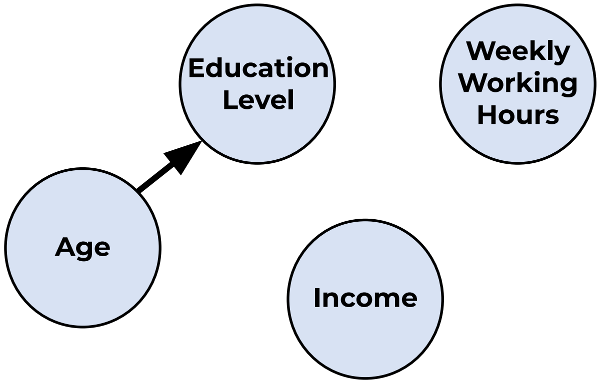
Why do we need causality?

Goal: Can the customer get the loan?

Dataset

Age	Income	Education Level	Weekly working hours
24	800	High School	20
28	1300	Bachelor Degree	35
...	...	...	...

Causal Graph



Why do we need causality?

Goal: Can the customer get the loan?

Age	Income	Education Level	Weekly working hours
24	800	High School	20
28	1300	Bachelor Degree	35
...	...	...	...

Black Box Prediction: No

Lime Explanation: Low education level is mainly responsible for the denied loan

Why do we need causality?

We inspect the neighborhood generated by LIME of the instance to explain

Age	Income	Education Level	Weekly working hours
24	800	High School	20
28	1300	Bachelor Degree	35
...	...	...	...

Generated Neighborhood

24	800	PHD	20

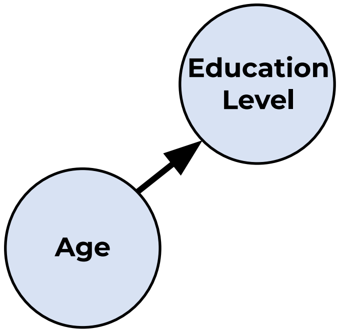
Why do we need causality?

We inspect the neighborhood generated by LIME of the instance to explain

Age	Income	Education Level	Weekly working hours
24	800	**High School**	20
28	1300	Bachelor Degree	35
...	...	...	...

Generated Neighborhood

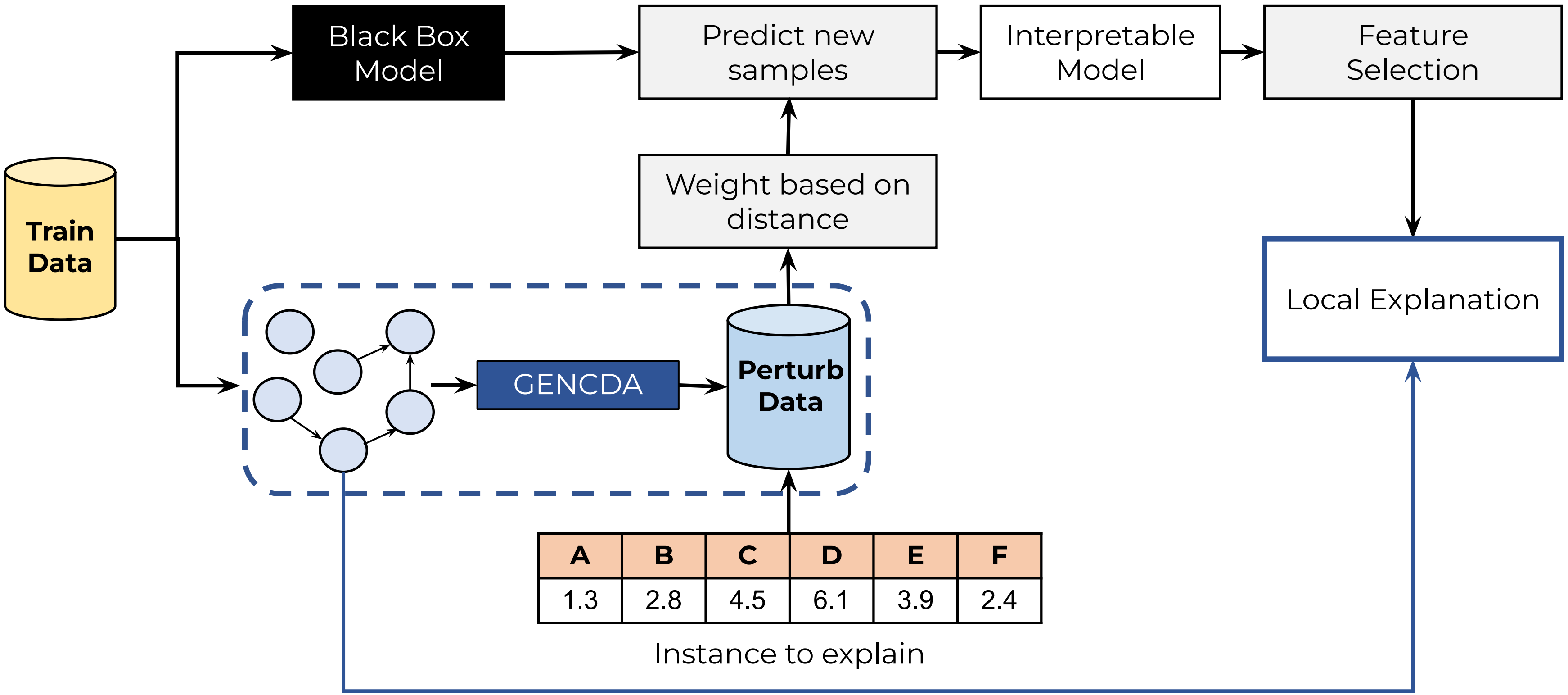
24	800	**PHD**	20



Problem: The generated instance is not plausible.
Generally, a guy who is 24 is too young to have a PhD.

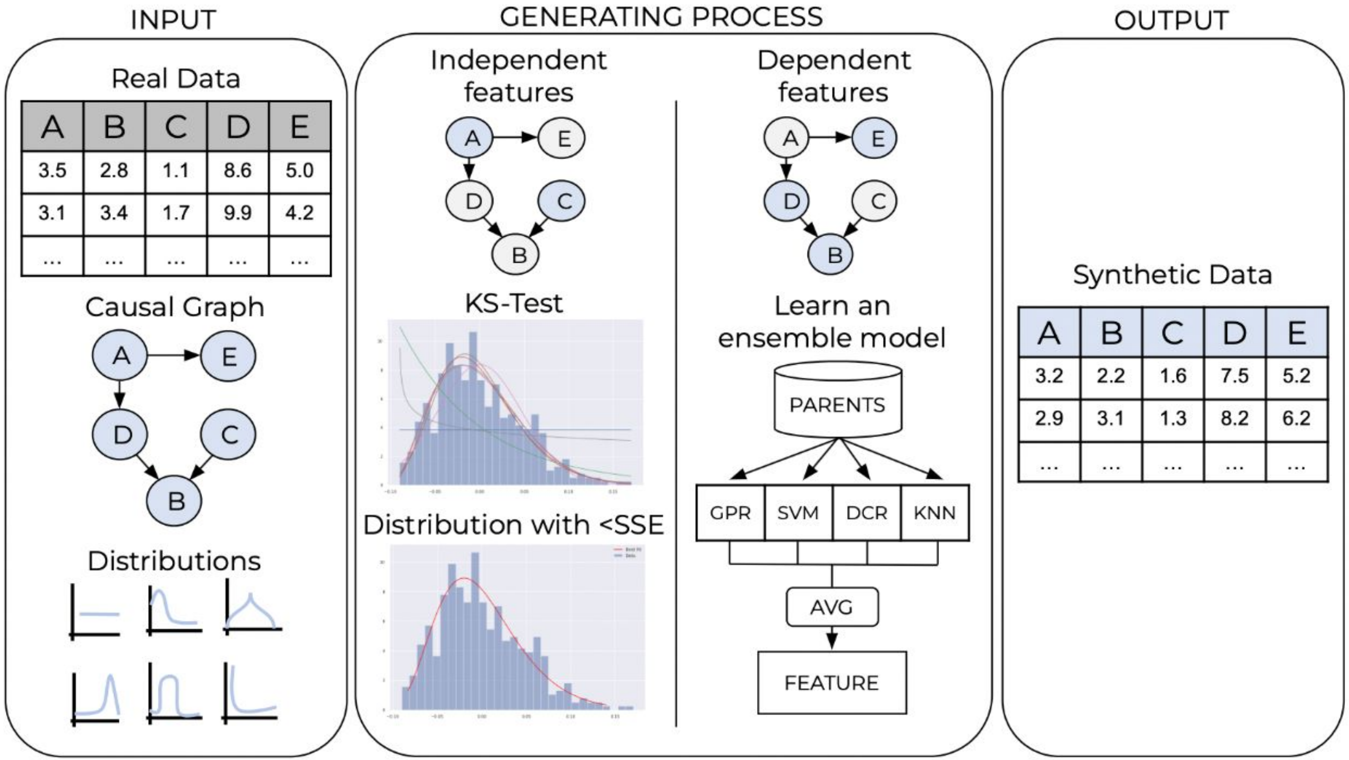
CALIME
Causality-Aware LIME

CALIME workflow



GENCDA

GEnerative **N**onlinear **C**ausal **D**iscovery with **A**priori[3]



[4] Boosting Synthetic Data Generation with Effective Nonlinear Causal Discovery, Cinquini et al., 2021

Example

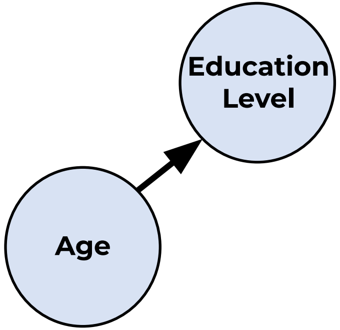
We inspect the neighborhood generated by CALIME of the instance to explain

Age	Income	Education Level	Weekly working hours
24	800	**High School**	20
28	1300	Bachelor Degree	35
...	...	...	...

Generated Neighborhood

34	1500	**PHD**	30

- Education level cannot be changed if age is not changed
- When age is changed also education level must be changed according to the regression model



Experiments

Datasets & DAGs

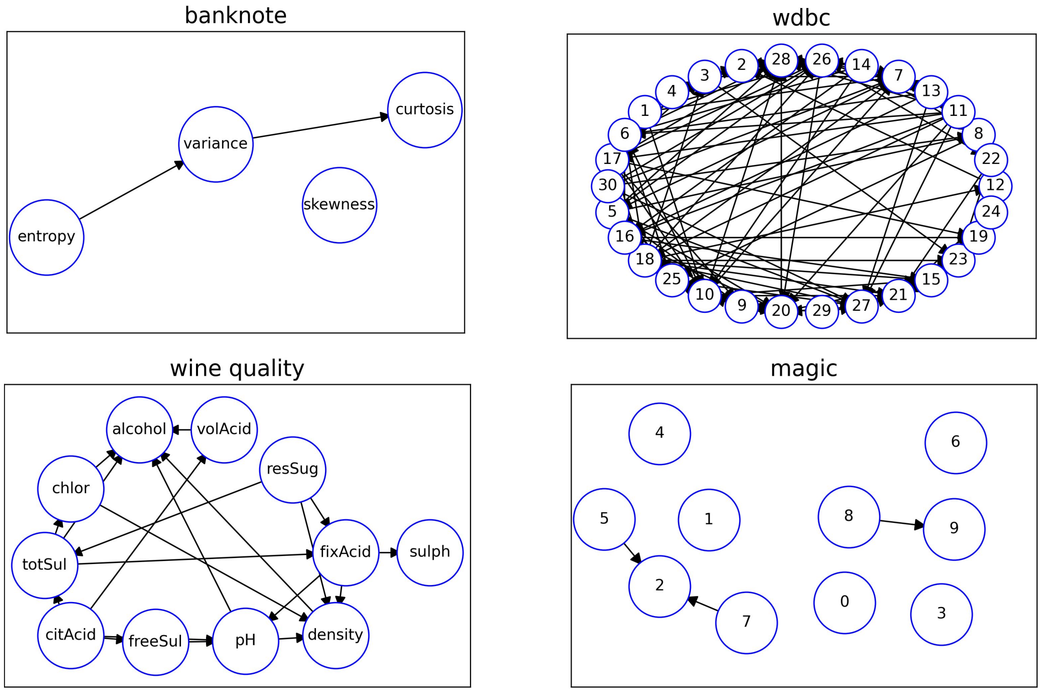
Statistics and classifiers accuracy

	n	m	RF	NN
banknote	1372	4	0.99	1.0
magic	19020	11	0.92	0.85
wdbc	569	30	0.95	0.92
wine-red	1159	11	0.82	0.70

n: # samples

m: # features

DAGs discovered by CALIME



[4] Source: UCI Repository

Evaluation Measures

Fidelity

How well does the explanation approximate the prediction of the black box model?

Plausibility

How convincing the explanations are to humans?

Stability

How similar are the explanations for similar instances?

Fidelity

In our setting, we define fidelity in terms of coefficient of determination R^2

$$R_x^2 = 1 - \frac{\sum_{i=1}^{N}(b(z_i) - r(z_i))^2}{\sum_{i=1}^{N}(b(z_i) - \hat{y})^2} \quad \text{with} \quad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} b(z_i)$$

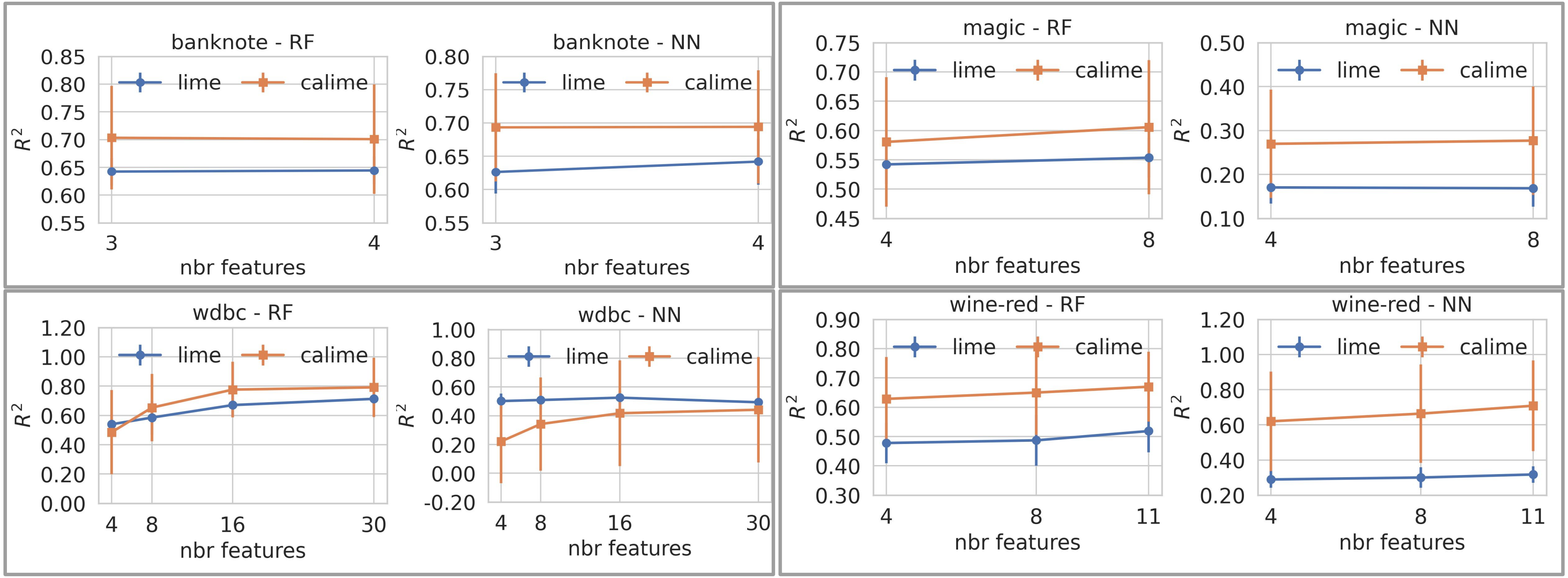
where $z_i \in Z$ is the synthetic neighborhood generated by LIME or CALIME for a certain instance x, and r is the linear regressor with Lasso regularization trained on Z.

R^2 ranges in [−1, 1] :

- 1 indicates that the regression predictions perfectly fit the data
- 0 is obtained by a baseline.

Fidelity

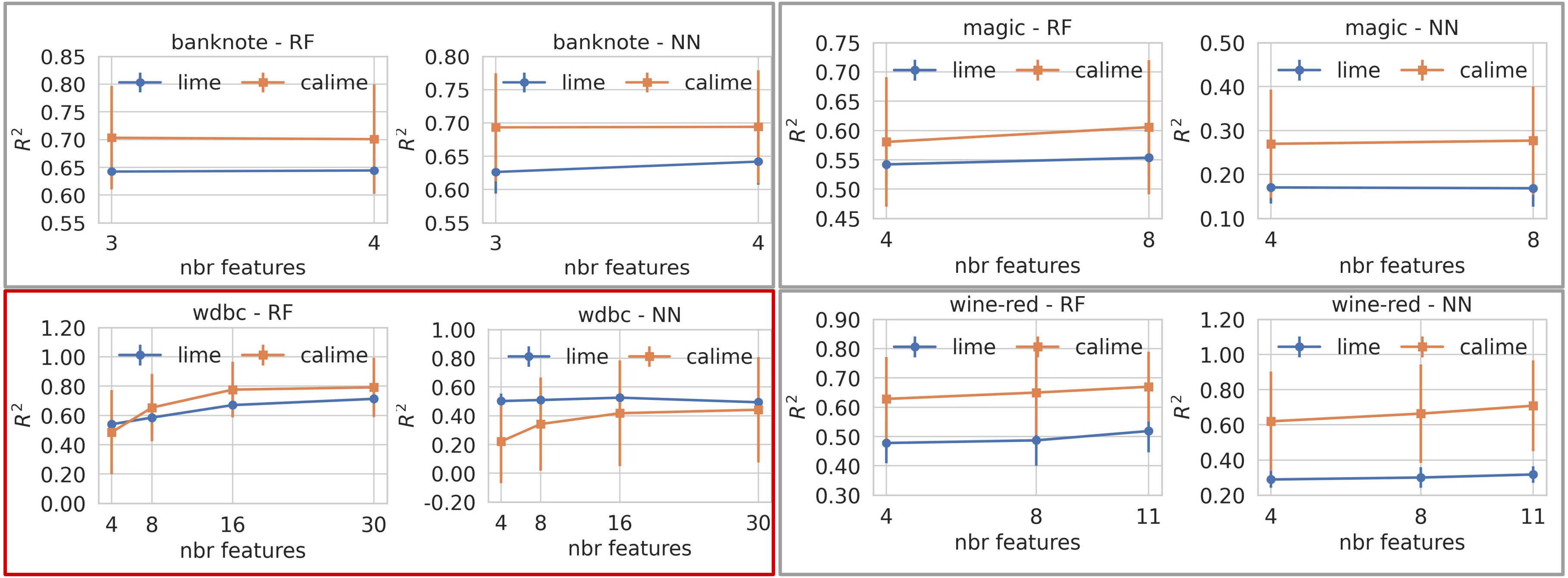
Results



A higher score indicates better fidelity values

Fidelity

Results



A higher score indicates better fidelity values

Plausibility

We evaluate the plausibility of the explanations in terms of the goodness of the synthetic datasets locally generated by LIME and CALIME by using the following metrics based on:

Distance

Average **M**inimum **D**istance

The lower the AMD, the more plausible are the instances in Z.

Outlierness

Average **O**utlier **S**core

-Local Outlier Factor
-Isolation Forest
-Angle-Based Outlier D.

Statistics

Average **S**tatistical **M**etric

-KS Test
-Continuous KL Divergence
-GM Log Likelihood

Detection

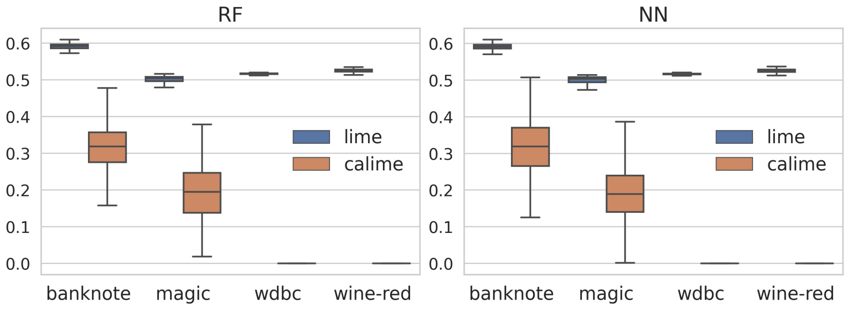
Average **D**etection **M**etric

-Logistic Detector
-SVM

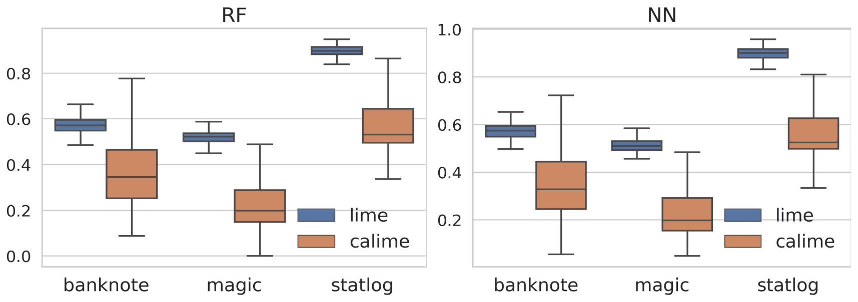
Plausibility

Results

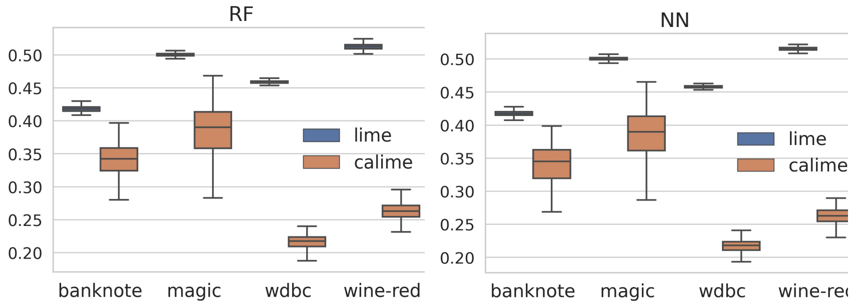
Average Minimum Distance



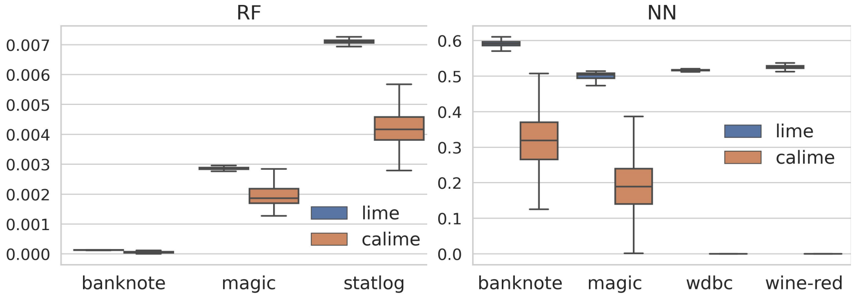
Average Outlier Score



Average Statistical Metric



Average Detection Metrics



Stability

We assess the stability through the local Lipschitz estimation:

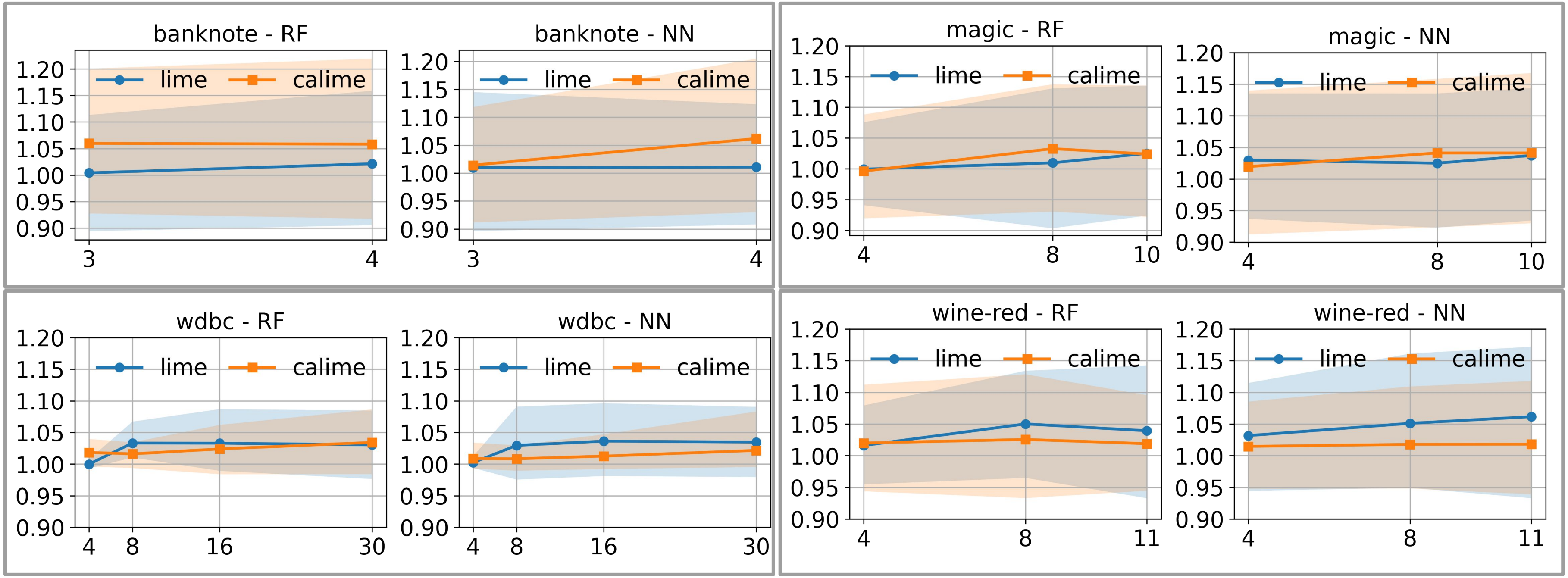
$$LLE_x = \underset{x_i \in \mathcal{N}_x^k}{avg} \frac{\|e_i - e\|_2}{\|x_i - x\|_2}$$

where x is the instance to explain and $N_x^k \subset X$ is the k-Nearest Neighborhood of x with the k neighbors selected from the test set.

The lower the LLE, the higher the stability.

Stability

Results



The lower the LLE, the higher the stability.

Key takeaways

CALIME is the first black-box explanation methods returning features importance as explanations that directly discover and incorporate causal relationships in the explanation extraction process.

Experiments results show that CALIME overcomes the weaknesses of LIME concerning both the fidelity in mimicking the black-box and the stability of the explanations.

CALIME could strengthen user trust in the AI system. It will be especially useful for high-impact domains such as financial services or healthcare (e.g., therapy planning or patient monitoring).

Key takeaways

Disadvantages:

- it suffers from limitations that are typical of black-box explanation methods returning explanations in the form of features importance, e.g. it is parametric w.r.t the number of features;

- it is only suitable for continuous data due to GENCDA

Future Directions:

- Develop causality aware explanation methods suitable for images and time series working in a similar manner of CALIME;

- Employ the knowledge about causal relationships in the explanation extraction process of other model-agnostic explainers like LORE, SHAP or ANCHORS.

Thank you for your attention!