Overview of XAI: LIME and SHAP. Should I trust you?

By Pongsakorn Tanupatrasakul & Nat Tangchitnob

“Why should I trust you?” explaining the prediction of any Classifier

“Why Should I Trust You?” Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu

ABSTRACT

Despite widespread adoption, machine learning models remain mostly black boxes. Understanding the reasons behind predictions is, however, quite important in assessing *trust*, which is fundamental if one plans to take action based on a prediction, or when choosing whether to deploy a new model. Such understanding also provides insights into the model, which can be used to transform an untrustworthy model or prediction into a trustworthy one.

In this work, we propose LIME, a novel explanation technique that explains the predictions of *any* classifier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction. We also propose a

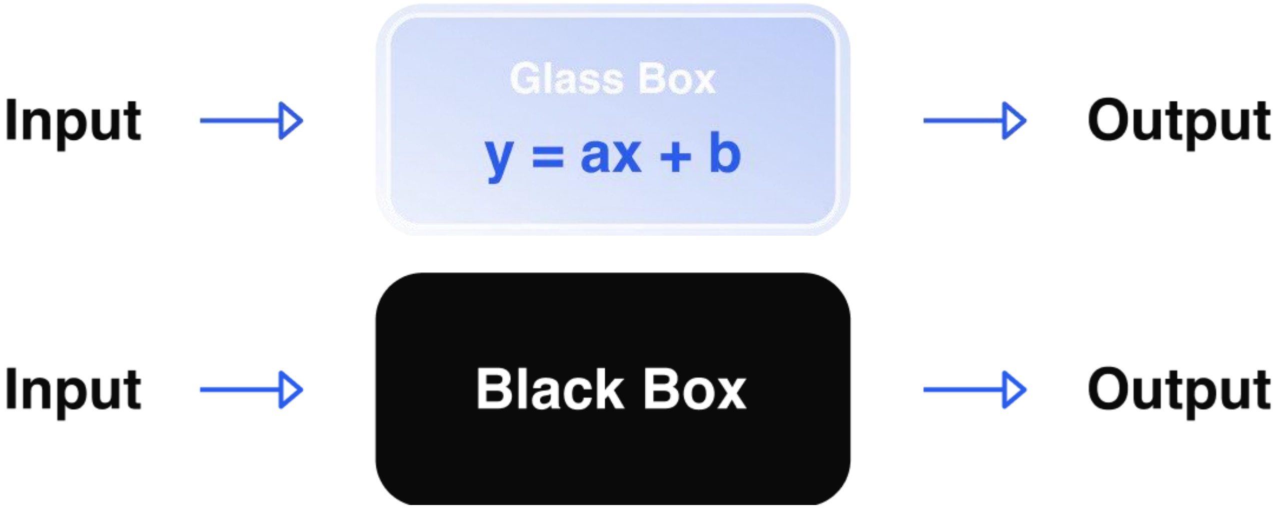
how much the human understands a model's behaviour, as opposed to seeing it as a black box.

Determining trust in individual predictions is an important problem when the model is used for decision making. When using machine learning for medical diagnosis [6] or terrorism detection, for example, predictions cannot be acted upon on blind faith, as the consequences may be catastrophic.

Apart from trusting individual predictions, there is also a need to evaluate the model as a whole before deploying it “in the wild”. To make this decision, users need to be confident that the model will perform well on real-world data, according to the metrics of interest. Currently, models are evaluated using accuracy metrics on an available validation dataset.

Explainable AI?

- Machine learning & Deep learning are widespread
- Incomprehensible to explain coming to the term “Blackbox”
- What is our model learning? Which feature is important for ?
- How does the model work?(Explainable AI)



Deep learning don’t need feature engineering, its do it own feature extraction
Million of parameters!

Explainable AI vs Interpretable AI

Explainable AI aims to explain complex model such as Blackbox on their decision-making processes in a way that is understandable to humans.

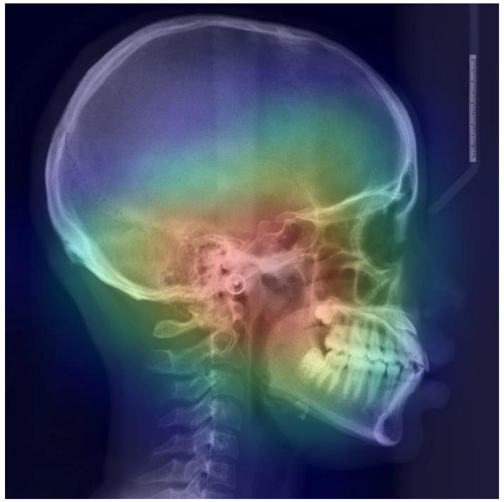
Interpretable AI refers to the ability to inspect the internal workings of a model and determine how it arrives at its output translating to human understandable .

<u>Interpretable AI visualization</u>: OSA classification project

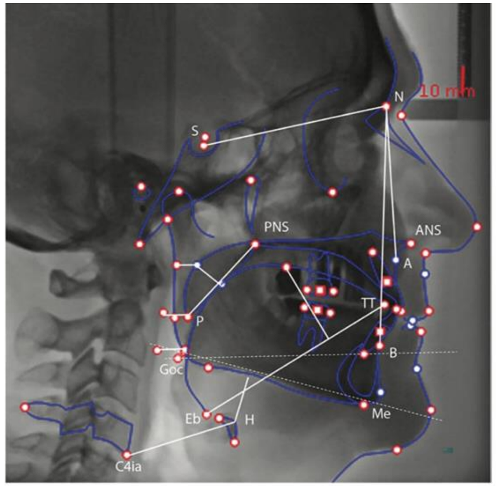
Saliency map

Human/ explainable AI

Raw Cephalometric



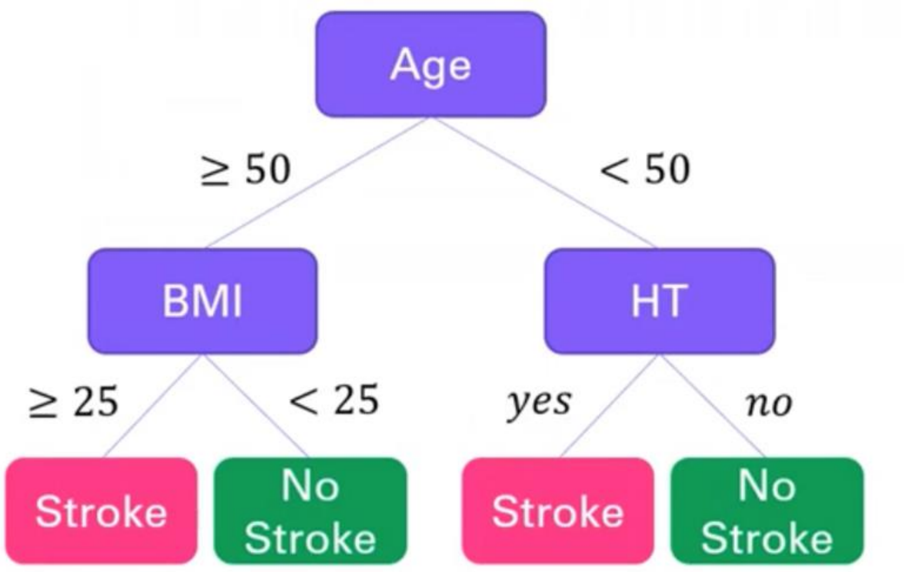
Model interpretable visualization



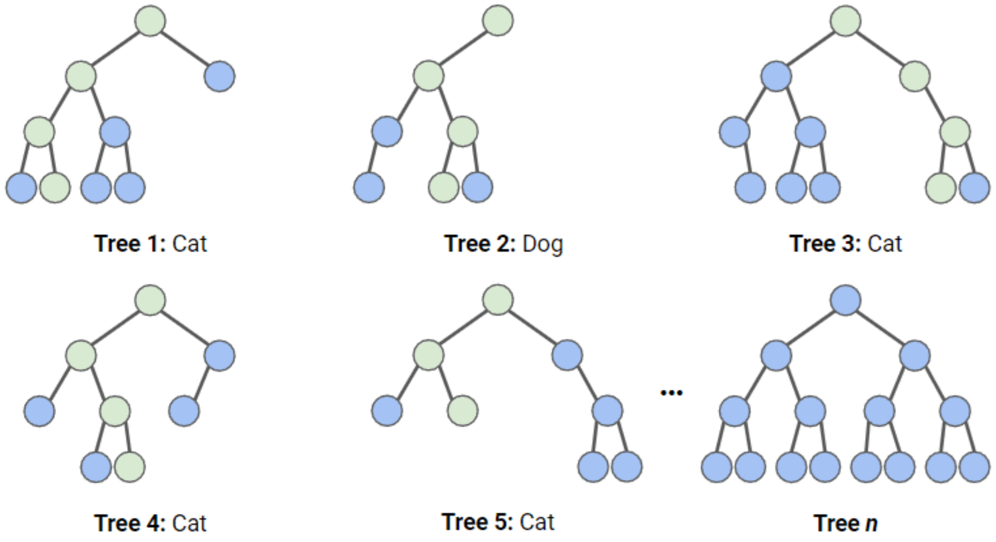
How human/XAI interpret cephalometric

More example of
Interpretable AI visualization

Simple Decision tree for stroke



Random forest

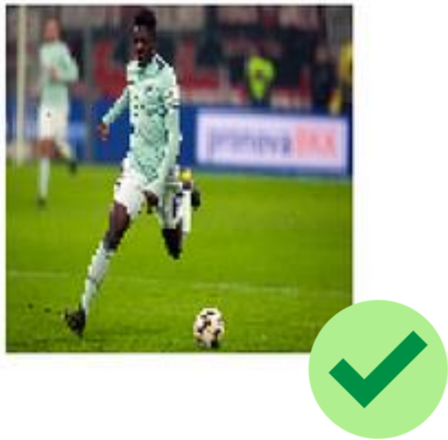


Accuracy and trust?

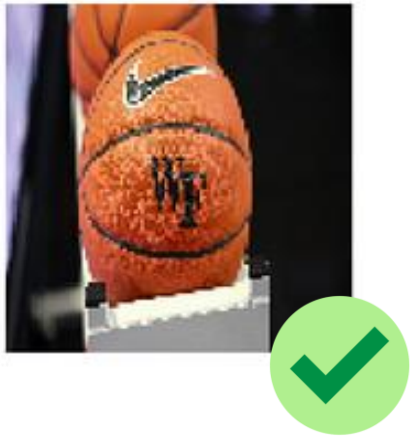
Basketball or Football classifier

It's a basketball!

It's a football!

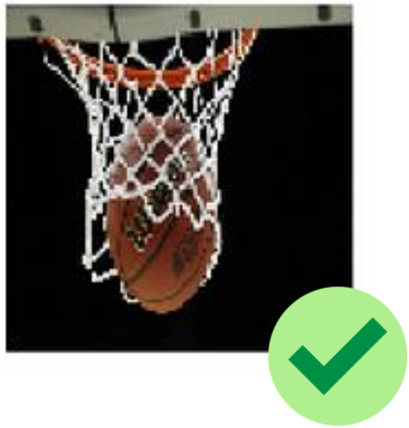


It's a basketball!

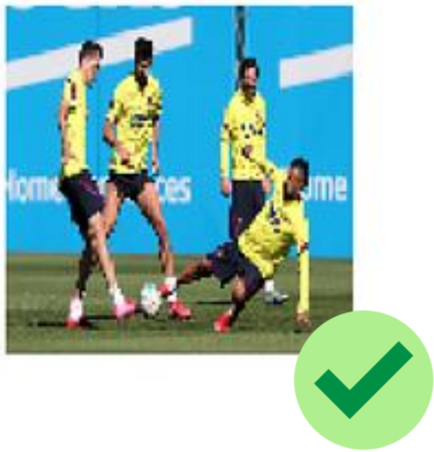


It's a football!

It's a basketball!



It's a football!



Explained Trust?

Still Trust?

It's a basketball!

It's a football!

It's a basketball!

It's a football!

It's a basketball!

It's a football!

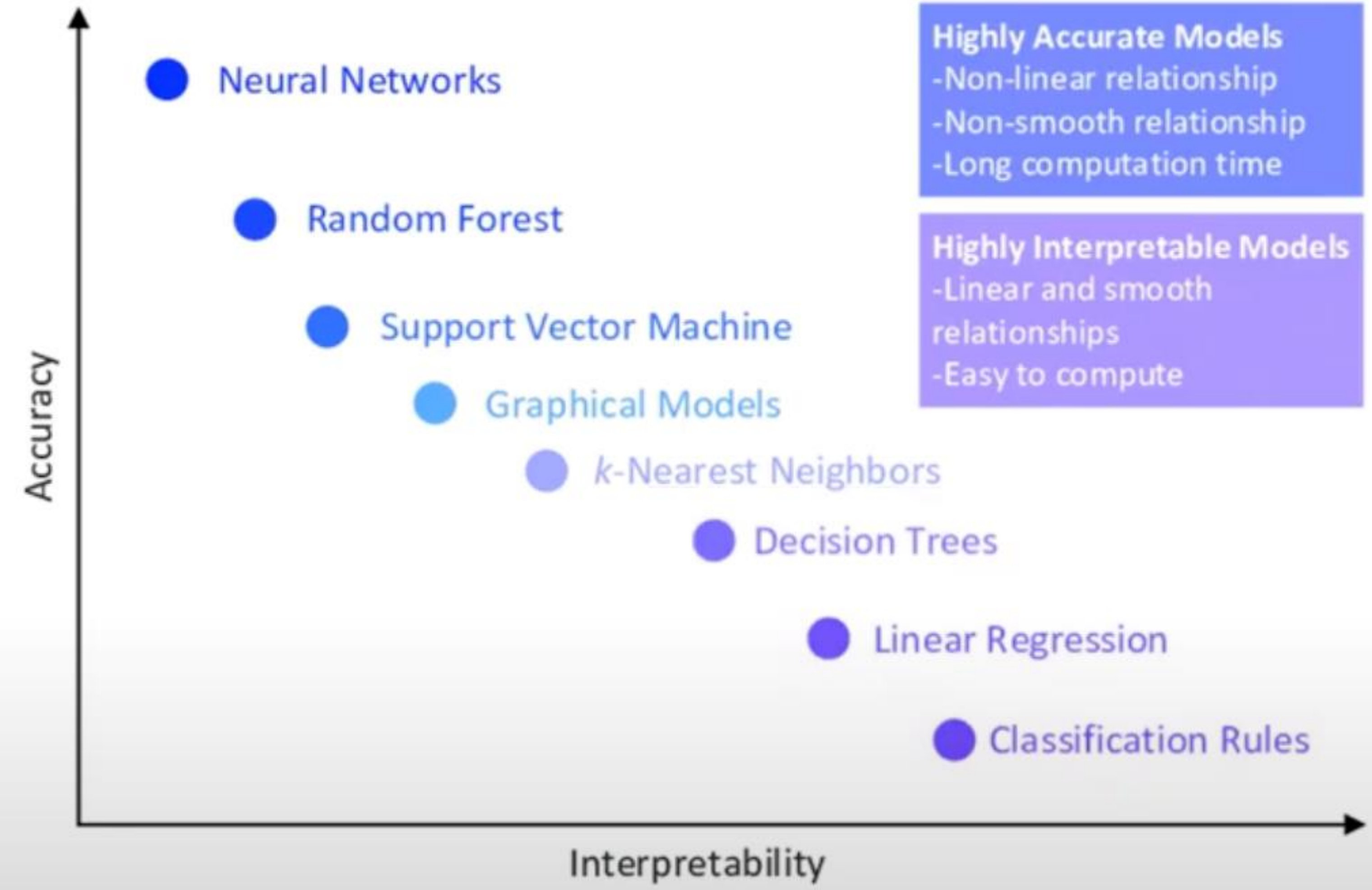
Model understanding Benefit for Stakeholders

Engineers/ Data Scientist	Consumer/ Doctor	Regulator
Increase Understanding	Increase Trust	Increase Trust
Improve Performance	Transparency & Bias	Transparency & Bias
Invent Better algorithm	Understand the Impact	Compliance
Produce Models	Report & Analyses	Report

Benefit of understanding model behavior

Explain	Verify	Present	Debug and fix
• Prediction to support decision process	• That Model behavior is acceptable	• The Model to the stakeholder to increase trust	• Unexpected behavior

Complexity – Explainability trade off



Source: Machine Learning for 5G/B5G Mobile and Wireless Communications: Potential, Limitations, and Future Directions

Explainable AI
Cheat sheet ex.pegg.io v.0.2

Start

Explaining a model's decisions?

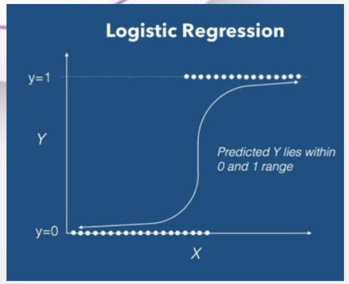
No — Ah, so we're exploring the model and its internals

Neural Representations
- SVCCA
- Activation maximization
- Probes
- Feature visualization
- TCAV

Yes → **Is the model interpretable by design?**

Yes — Good choice! *Especially* for high-stakes decisions

Interpretable Models
- Logistic regression
- KNN
- Linear models



No — So less-interpretable methods perform better for your use-case? We understand!

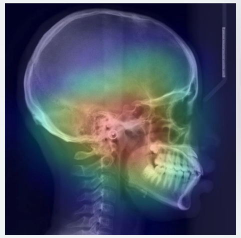
Need a method that works on all models?

Yes → **Model Agnostic**
- SHAP
- Perturbation
- LIME

No → **Model Specific**
- Gradient saliency
- Integrated gradients
- Attention

Also yes — Using special examples from the dataset could uncover insights about the model

Example Based
- Adversarial
- Influence functions
- Counterfactual



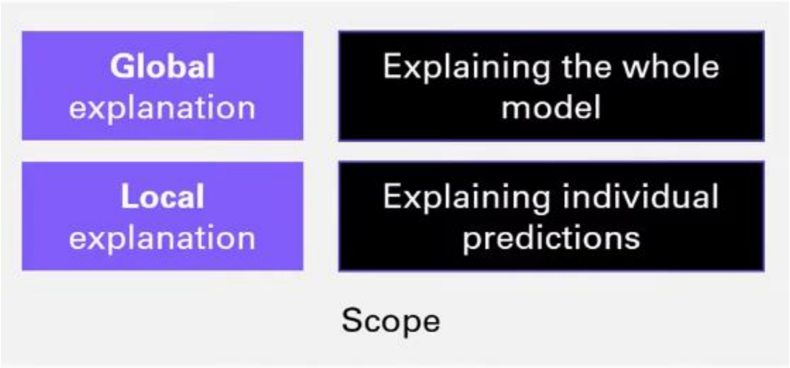
Arpeggio

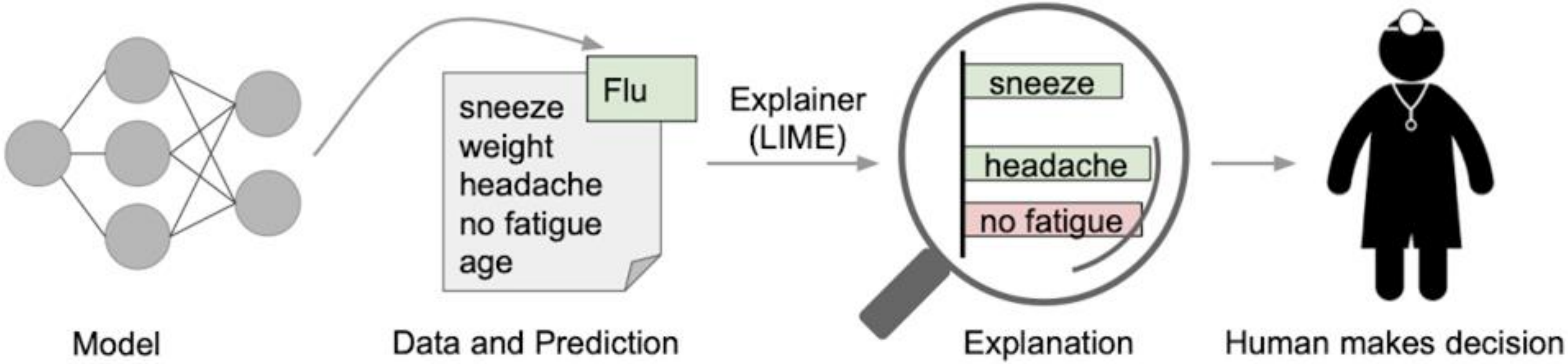
LIME - Local Interpretable Model-agnostic Explanations

LIME focuses on an individual prediction made by the model, by identifying the feature importance for prediction from the input features (like variables or data points).

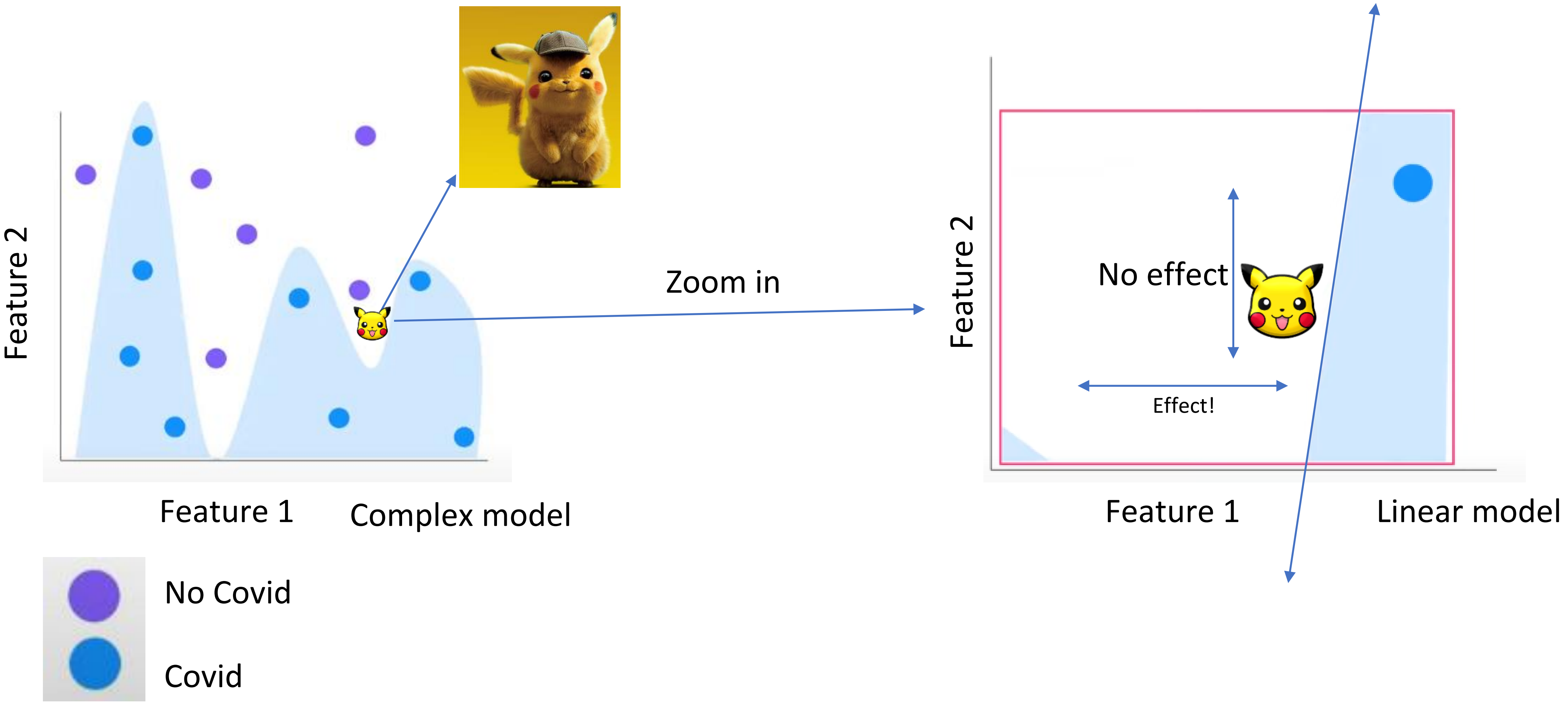
LIME tweak small changes on input feature and observe changes to see how it influence the output of the model. (what influence the model the most?)

LIME supports explanations for tabular, text, and image.





LIME Motivation



LIME "the Math for Geek"

Complex model

Kernel

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \; \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

Age | Gender | BMI

Regularization similar to L1 Lasso

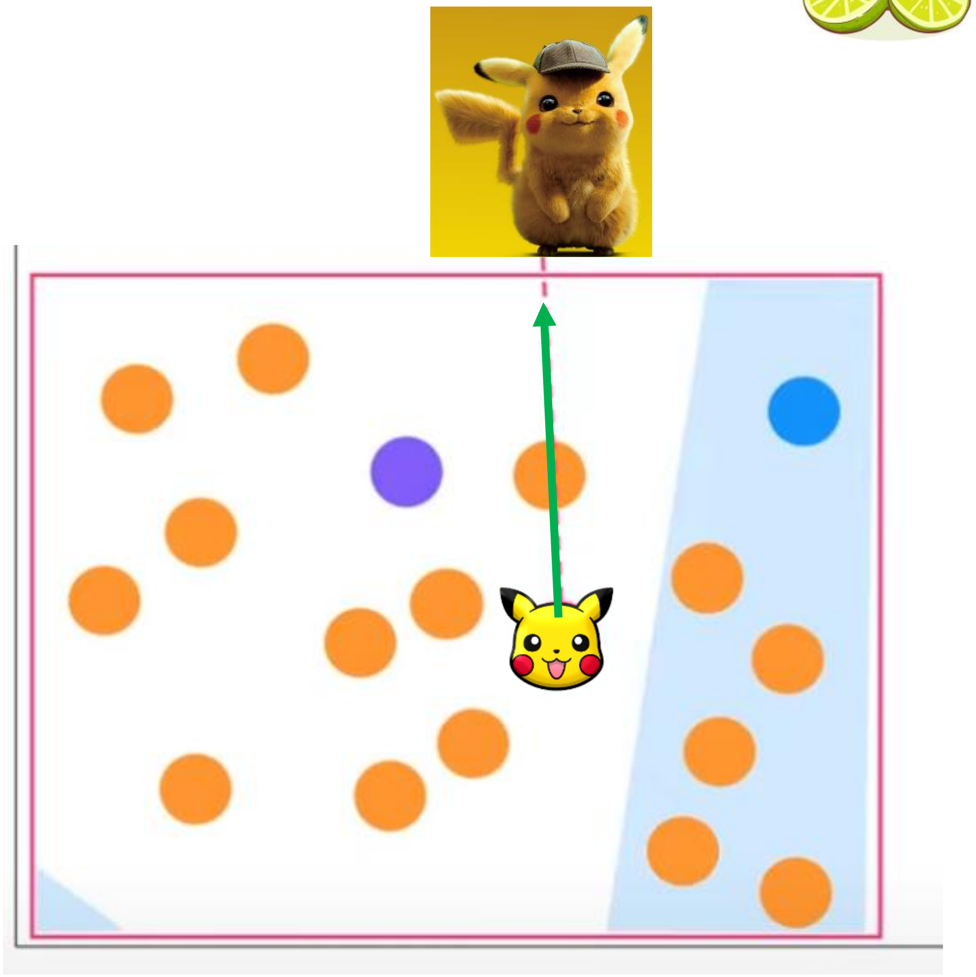
LIME

1. Select instance of interest x that want to explain

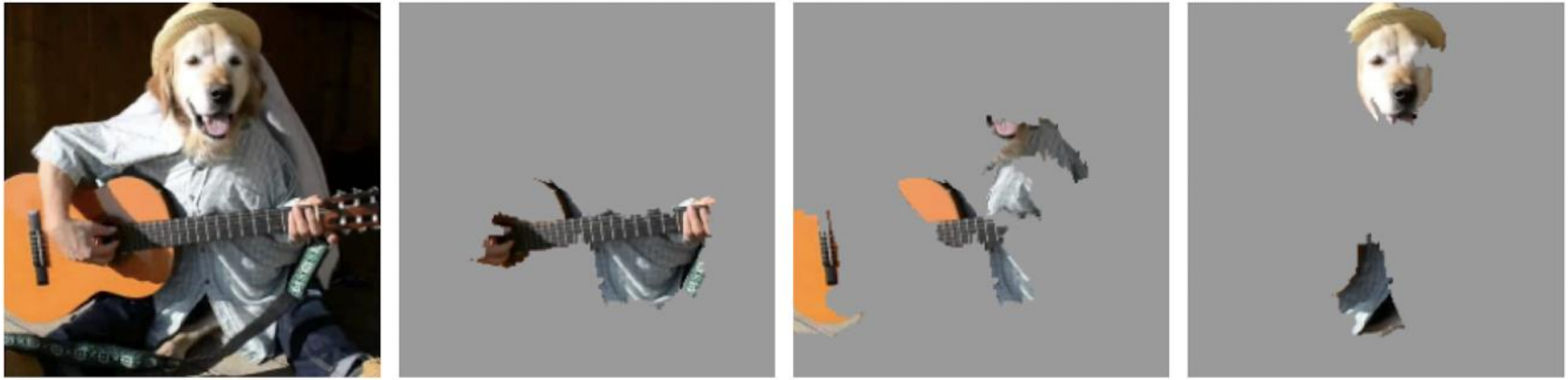
2. Perturb the instance in several way to generate dataset new samples(orange color)

3. Train interpretable model on the newly generated dataset

4. Explain the prediction by interpreting interpretable model



Google interception image



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Do explanations lead to insights?

Wolf/ husky classfier? Or Snow classfier

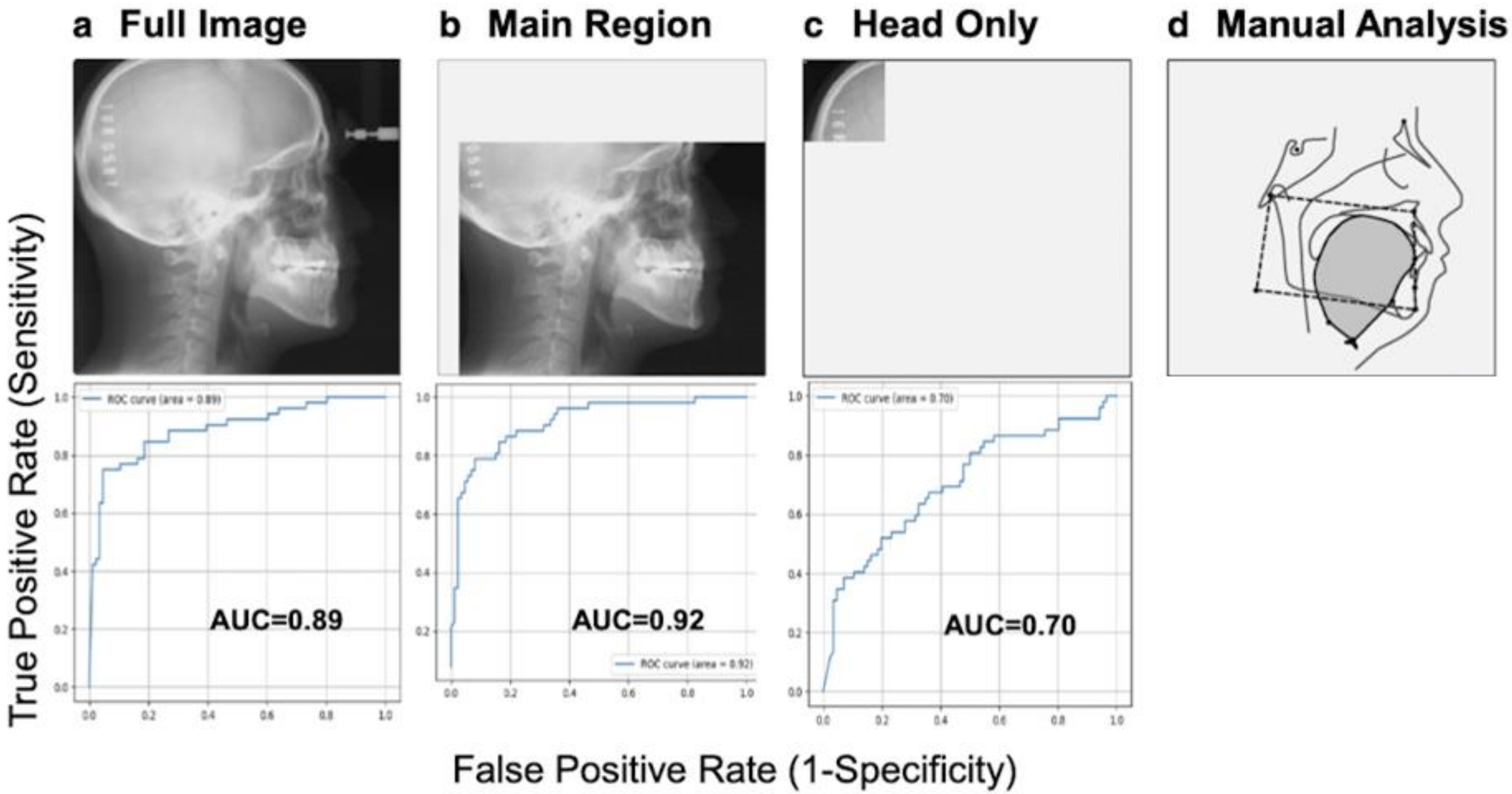
(a) Husky classified as wolf (b) Explanation

Logistic regression (binary classification) on inception neural network condition:

- All train image of wolf are with snow
- All image with huskies no snow

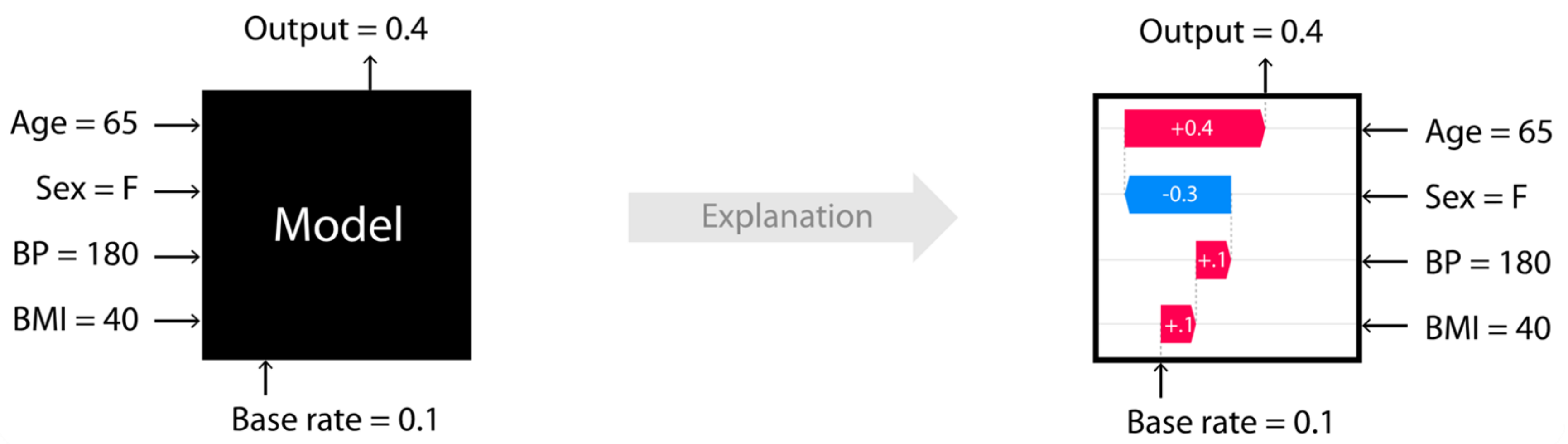
	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

OSA Binary Classifier



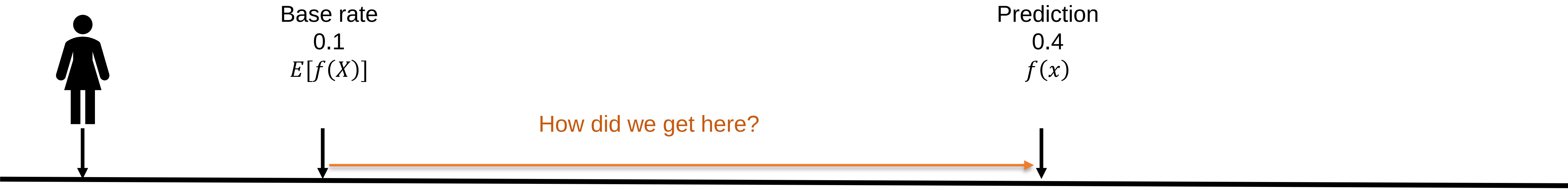
SHAP – SHapley Additive exPlanations

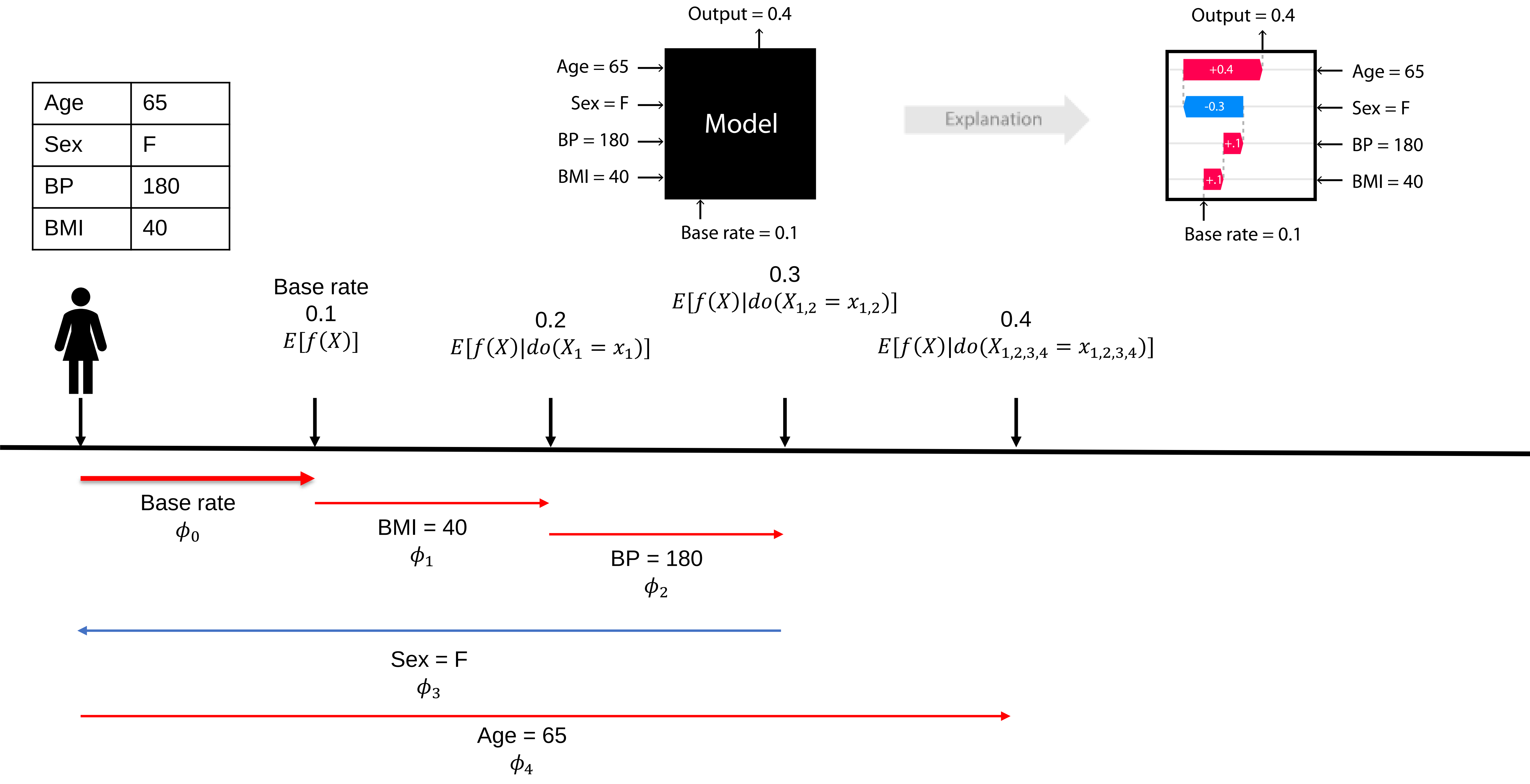
- Explain the prediction of an observation
- By computing the contribution of each feature to the prediction



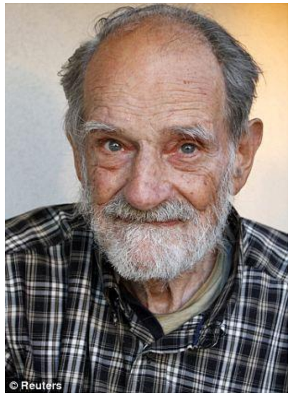
https://github.com/slundberg/shap

Age	65
Sex	F
BP	180
BMI	40





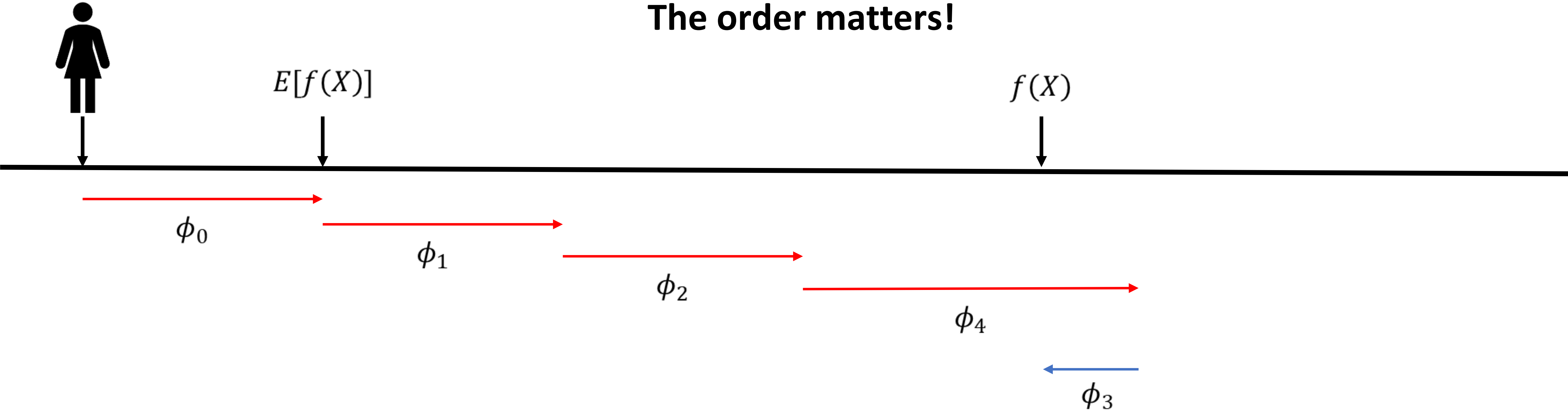
Age 65
Sex F
BP 180
BMI 40
Output = 0.4
Age = 65
Sex = F
BP = 180
BMI = 40
Model
Base rate = 0.1
Explanation
Output = 0.4
+0.4
-0.3
+.1
+.1
Age = 65
Sex = F
BP = 180
BMI = 40
Base rate = 0.1
Base rate
0.1
$E[f(X)]$
0.2
$E[f(X)|do(X_1 = x_1)]$
0.3
$E[f(X)|do(X_{1,2} = x_{1,2})]$
0.4
$E[f(X)|do(X_{1,2,3,4} = x_{1,2,3,4})]$
Base rate
ϕ_0
BMI = 40
ϕ_1
BP = 180
ϕ_2
Sex = F
ϕ_3
Age = 65
ϕ_4



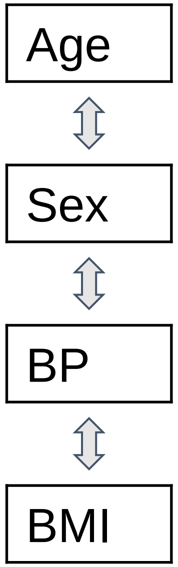
Shapley value, 1951

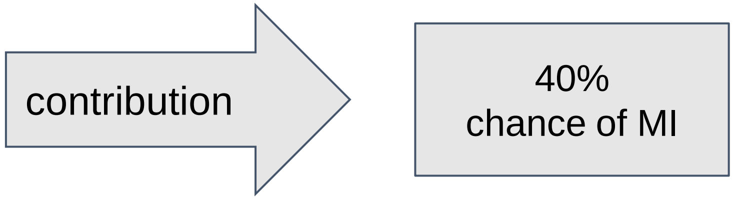
Lloyd Shapley
2012 Nobel Memorial Prize
in Economic Sciences

The order matters!

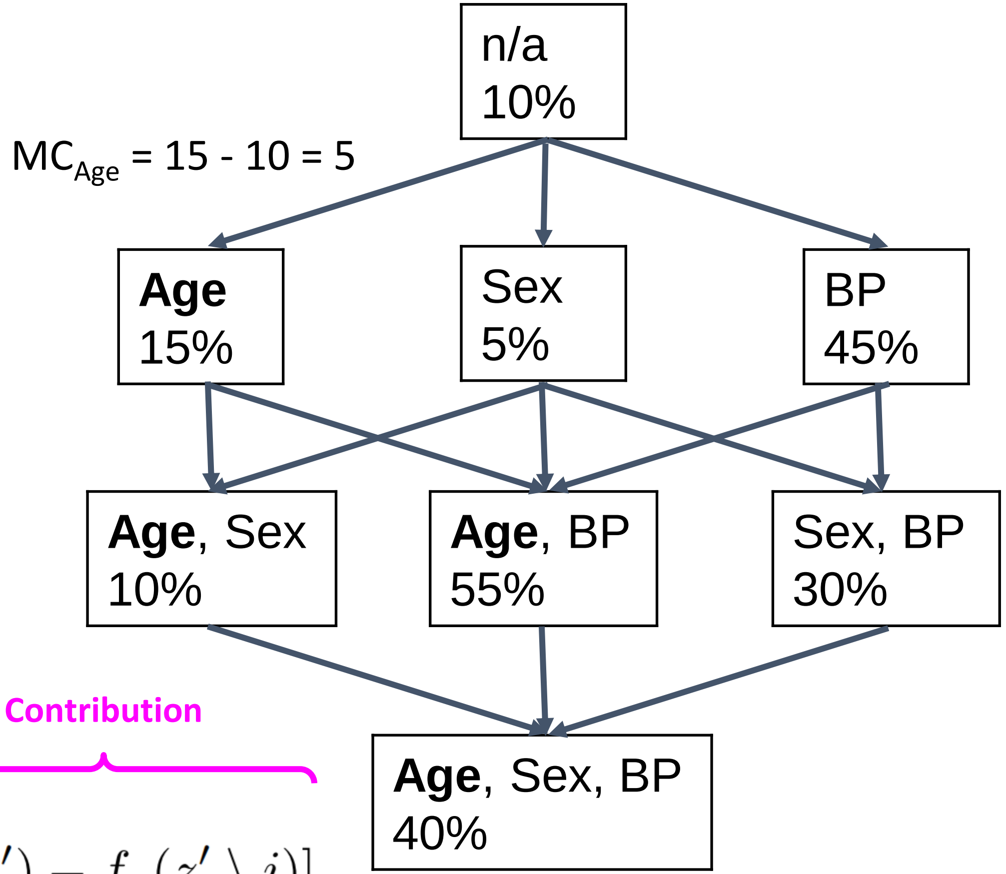


Shapley values





Marginal contributions of Age



$$\phi_i(f,x) = \sum_{z' \subseteq x'} \overbrace{\frac{|z'|!(M - |z'| - 1)!}{M!}}^{\text{Weight}} \overbrace{[f_x(z') - f_x(z' \setminus i)]}^{\text{Contribution}}$$

SHAP_Age = 10

LIME vs SHAP

LIME

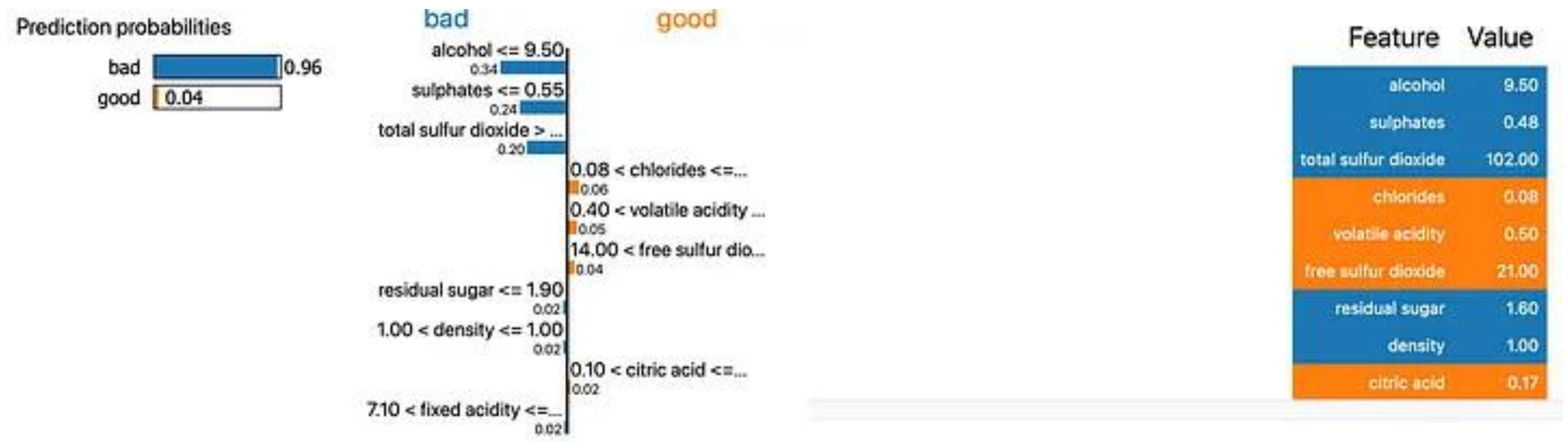
- Works on all kind of models
- Tabular data, text and images
- Only local explanation
- Faster

SHAP

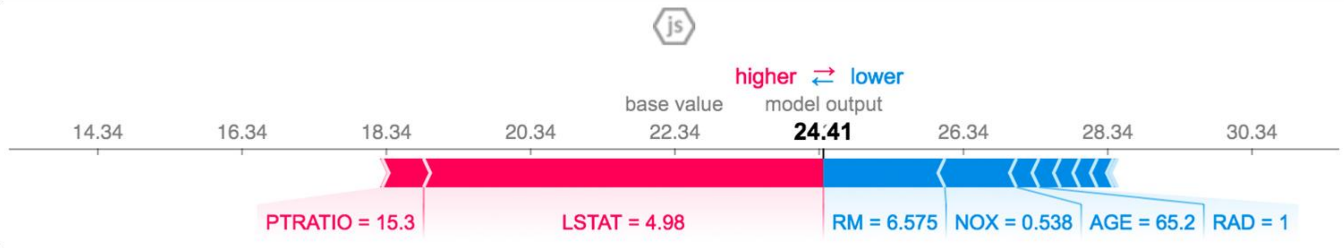
- Works on all kind of models
- Tabular data, text and images
- Both global and local explanation

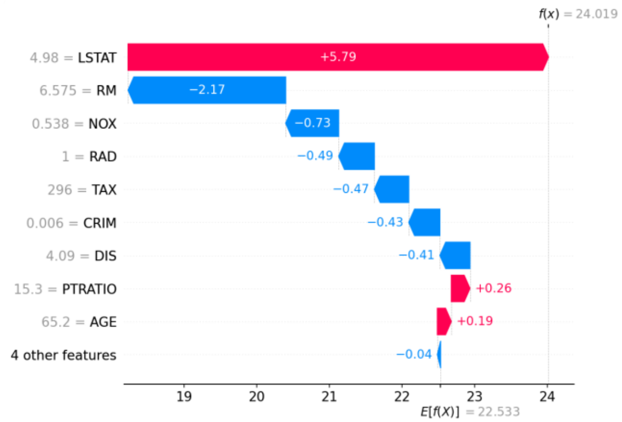
LIME vs SHAP local explanation

LIME



SHAP

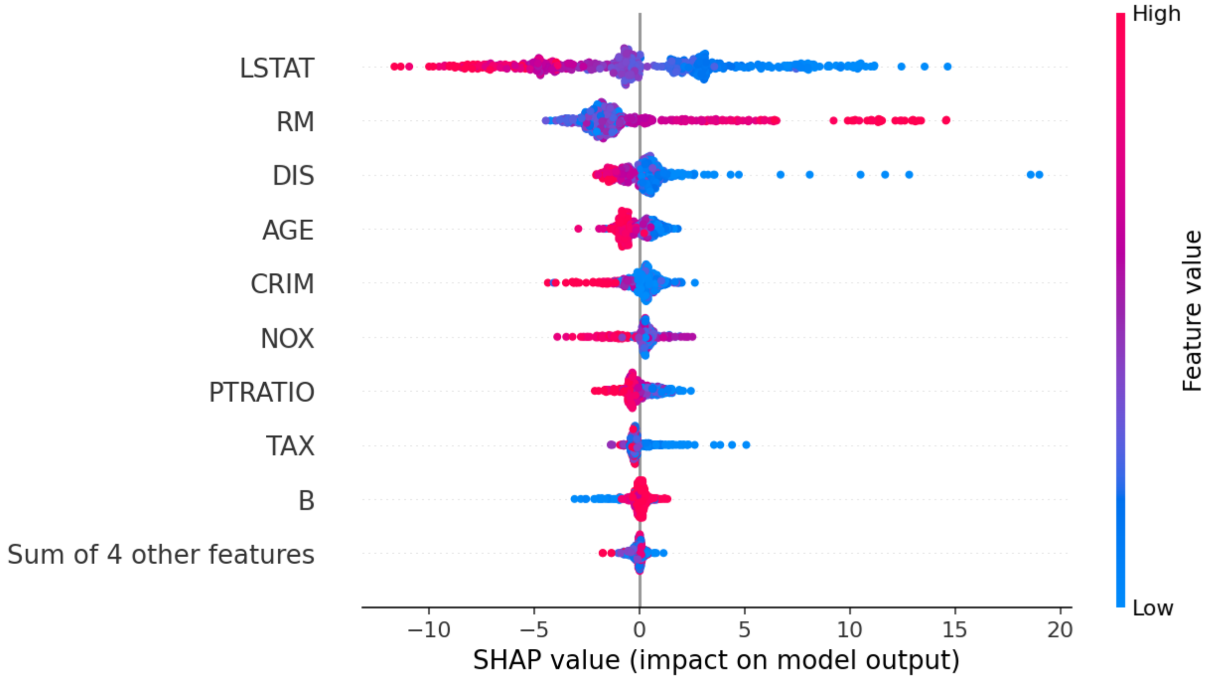




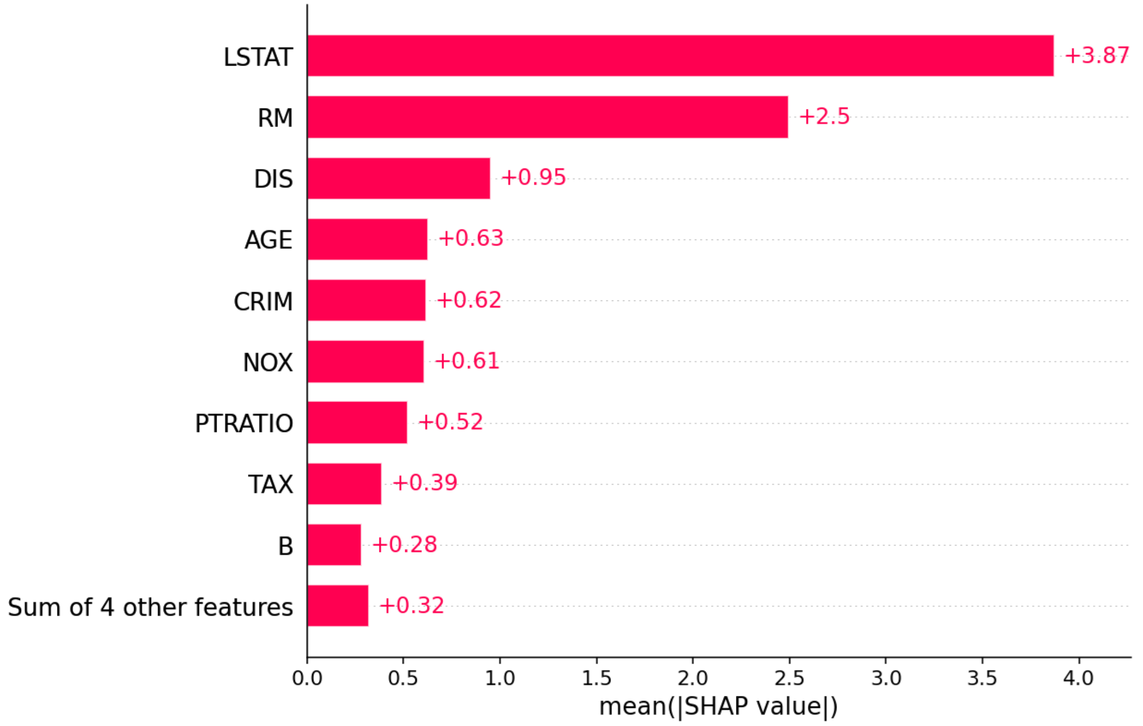
https://github.com/slundberg/shap

SHAP global interpretation

Beeswarm plot



Bar plot



https://github.com/slundberg/shap

Limitation of XAI

1. Correlation not causality
   - Explains the variable correlation of the model's features
2. Model dependency
   - How important a feature is to the model, not reality
3. Consistency in feature importance and signage
   - SHAP values is strongly related to the "objective" of the model
4. Multicollinearity issue
   - If there are variables with high degree of multicollinearity, the SHAP values would be high for one of the variables and zero/very low for the other
5. No performance guarantees
   - The performance of explanations is rarely tested at all, and most tests that are done rely on heuristic measures rather than explicitly scoring the explanation from a human perspective.

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, *3*(11), e745–e750.

Challenges of XAI in healthcare

1. Organizational problems paralyze decision-making, which in turn causes uncertainty, delays, and confusion in the practical implementation of AI
2. Understandable explanations by professionals in the medical field
3. Appropriate user interfaces for effective presentation of explanations
4. Unusual diseases might not be detected or cause false results
5. Insufficient explainability in the healthcare sector
6. Existing ML workflows need to be extended by integrating XAI approaches
7. Awareness of the limitations of explainable AI as it currently exists
8. Explainability in combination with privacy is a key concern

Nazar, M., Alam, M.M., Yafi, E., Su'ud, M.M.: A Systematic Re-view of Human–Computer Interaction and Explainable Artificial Intelligence in Healthcare With Artificial Intelligence Techniques. IEEE Access 9, 153316–153348 (2021).


THANK YOU FOR
YOUR ATTENTION
APPLAUSE BUT
DO NOT ASK
memegenerator.es