An attempt to understand our models' predictions

Parul Pandey, Machine Learning Engineer

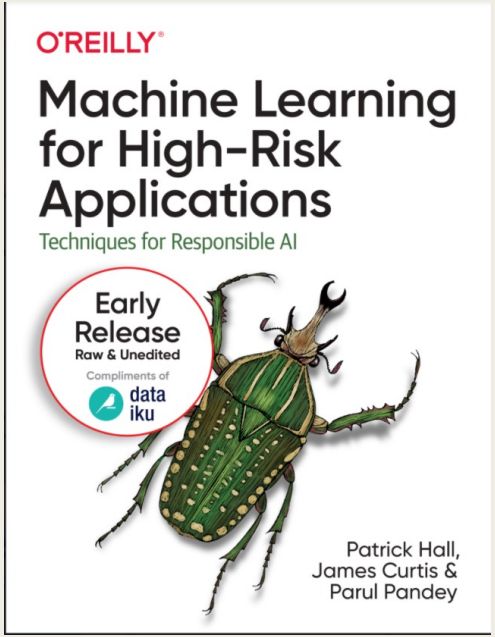
Who am I?

- ML Engineer @ Weights & Biases
- Electrical Engineer
- Kaggle Grandmaster(Notebooks)
- Author

kaggle

Who am I?

- ML Engineer @ Weights & Biases
- Electrical Engineer
- Kaggle Grandmaster(Notebooks)
- Author



https://learning.oreilly.com/library/view/machine-learning-for/9781098102425/

Motivation

Complex systems tend to drift toward unsafe conditions unless constant vigilance is maintained.

— Closing the AI Accountability Gap, Google Research

Motivation

Machine Bias

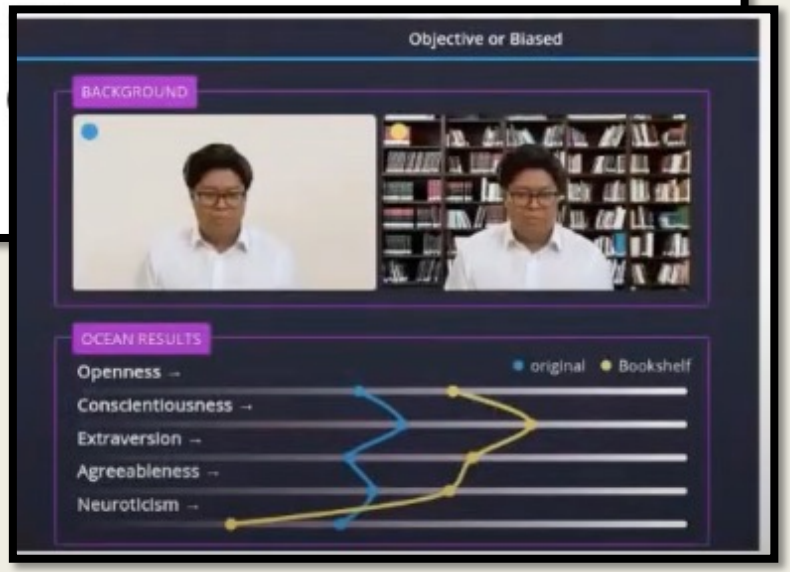
There's software used across the country to predict future criminals.
And it's biased against blacks.

by Ju

Amit Datta*, Michael Carl Tschantz, and Anupam

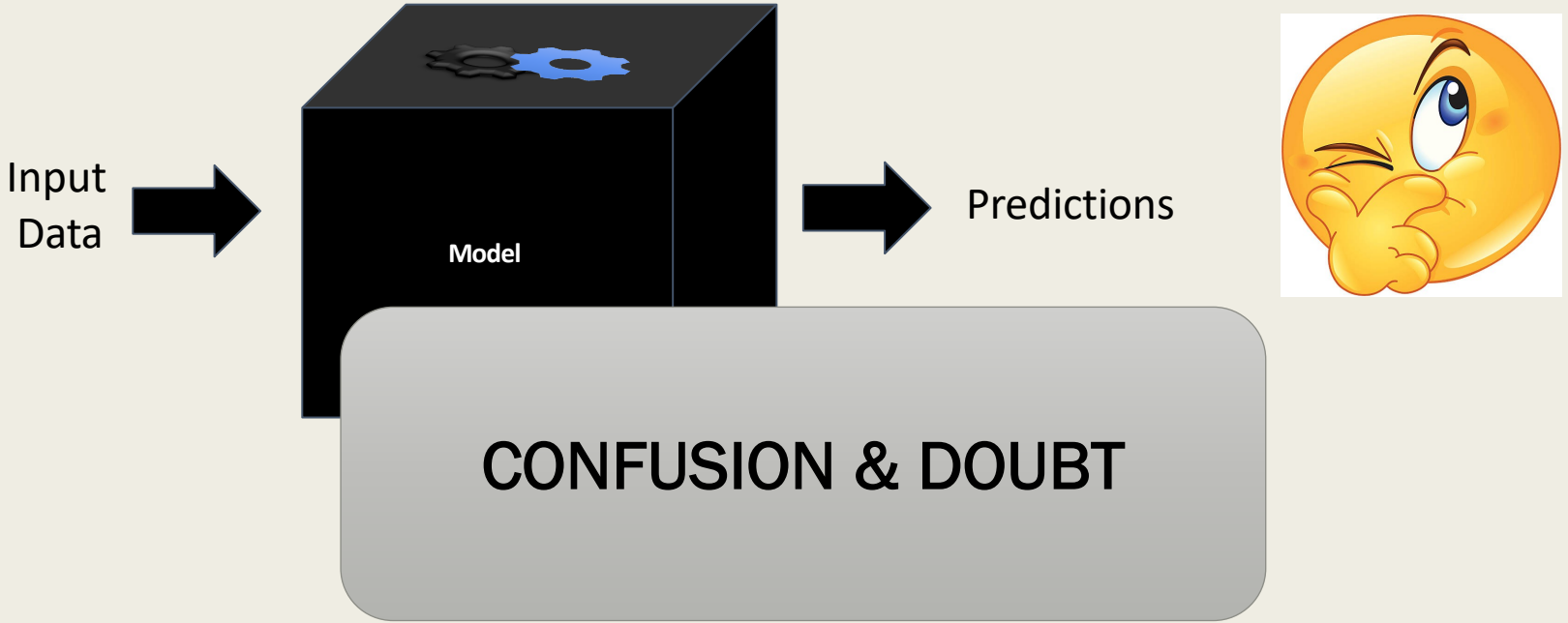
Automated Experiments

A Tale of Opacity, Choice, and Discrimination



- Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
- https://www.andrew.cmu.edu/user/danupam/dtd-pets15.pdf
- Washington Post

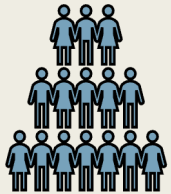
Complex & Opaque Models



Why Explainability

- To inculcate Trust in models
- To debug predictions
- To detect bias
- Ensure suitability of models for deployment
- Global AI Regulation

Data Scientists and ML Engineers



End Users

Regulators

Explanations

The degree to which a humans can understand and trust a ML model's predictions.

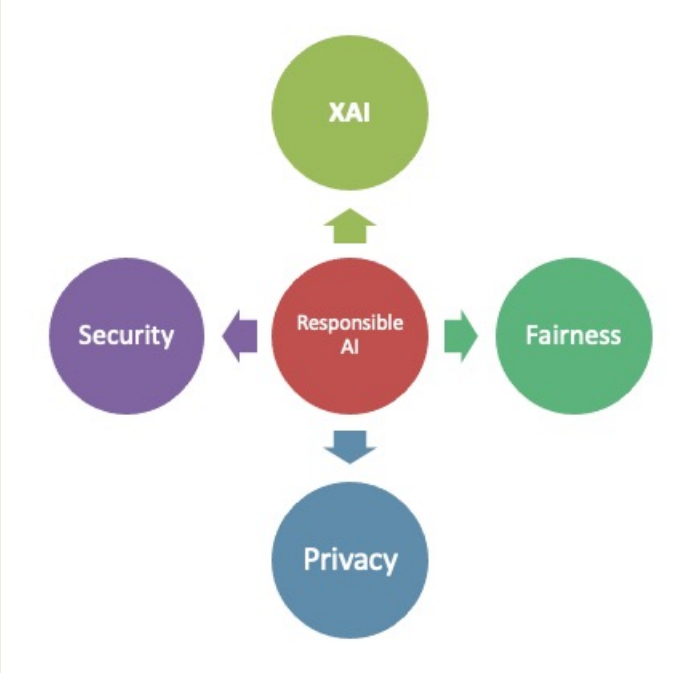
- Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences"
- Kim, Been, Rajiv Khanna, and Oluwasanmi O. Koyejo. "Examples are not enough, learn to criticize! Criticism for interpretability." Advances in Neural Information Processing Systems (2016).

Interpretable vs Explainable AI (XAI)

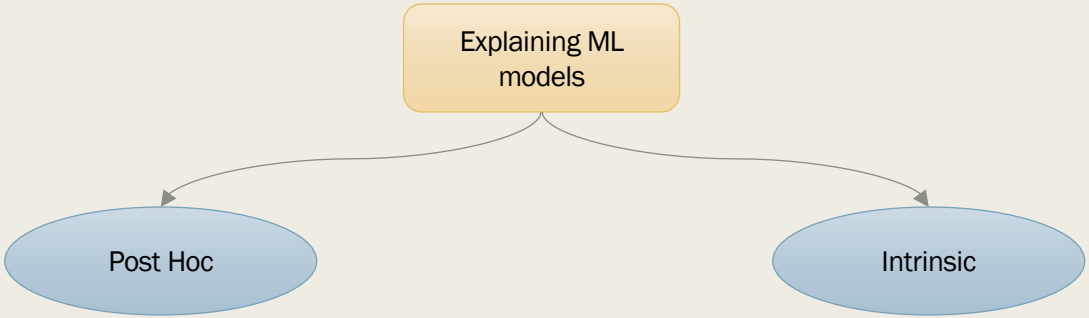
- Describes the internals of a system in a way which is understandable to humans
- Summarize the reasons for their behaviour.

Explaining Explanations: An Overview of Interpretability of Machine Learning

Responsible AI



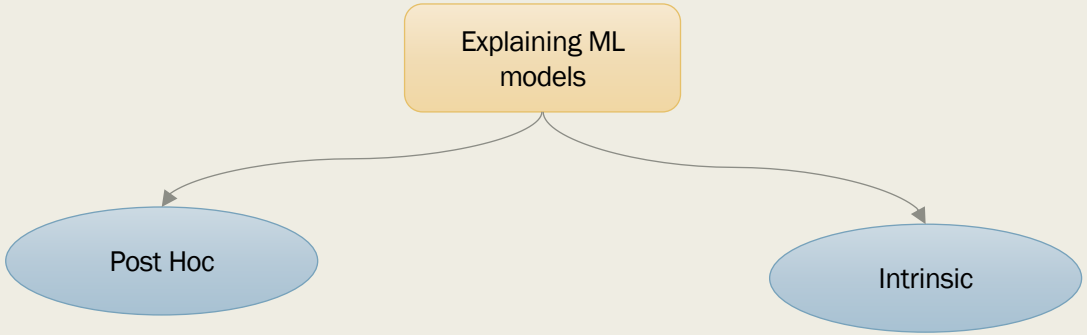
Taxonomy



After Model has been trained

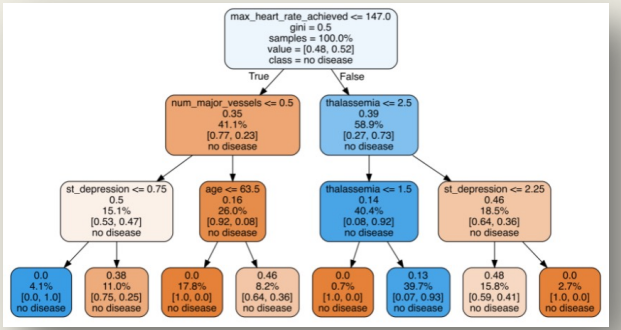
Create Inherently Interpretable models e.g.
Linear regression,
Logistic regression,
Decision trees etc

Taxonomy



After Model has been trained

Create Inherently Interpretable models e.g.
Linear regression,
Logistic regression,
Decision trees etc



Does a person have a heart disease ?

Taxonomy

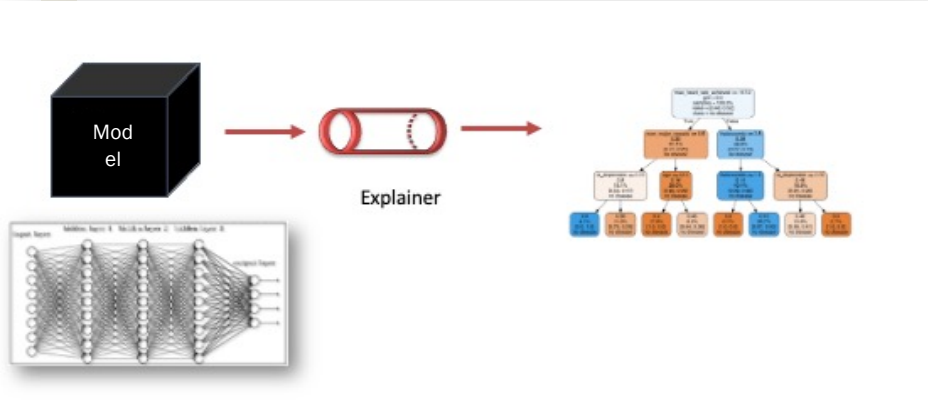
Explaining ML models

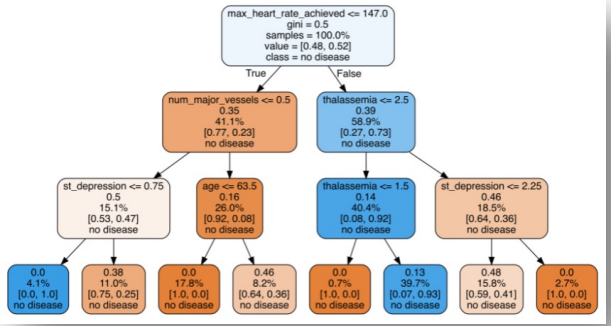
After Model has been trained

Post Hoc

Intrinsic

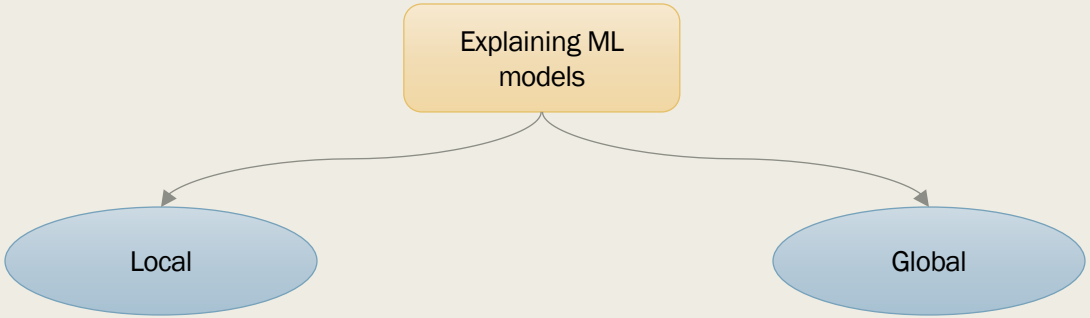
Create Inherently Interpretable models





Does a person have a heart disease ?

Taxonomy



explaining an individual prediction or a small part of the model's prediction space

explaining the entire model behaviour

Interpretable Models

- Penalized Regression
- Additive Models – GAMs
- Decision Trees
- Constrained XGBoost Models

Post Hoc Techniques

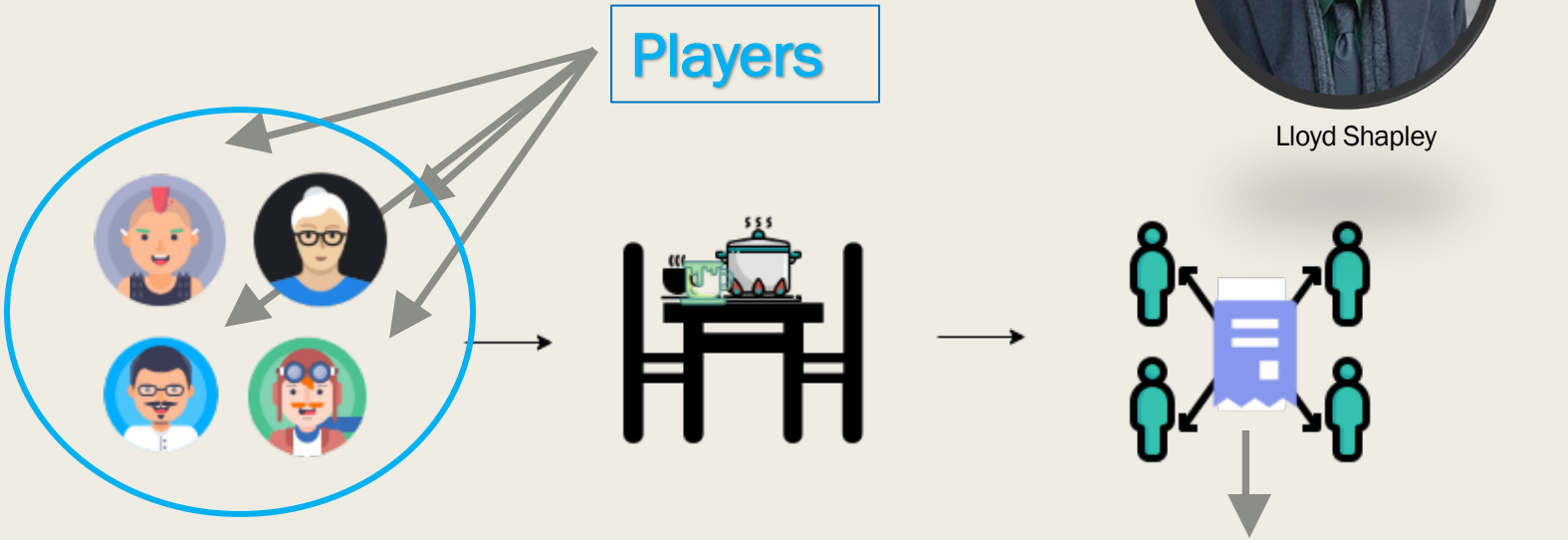
- Partial Dependence Plots
- SHAP
- LIME
- Counterfactuals
- Saliency Maps
- Occlusion

SHAP values

- SHAP stands for SHapley Additive exPlanations.
- It leverages the concept of **Cooperative Game theory** to break down a prediction to measure the impact of each feature.

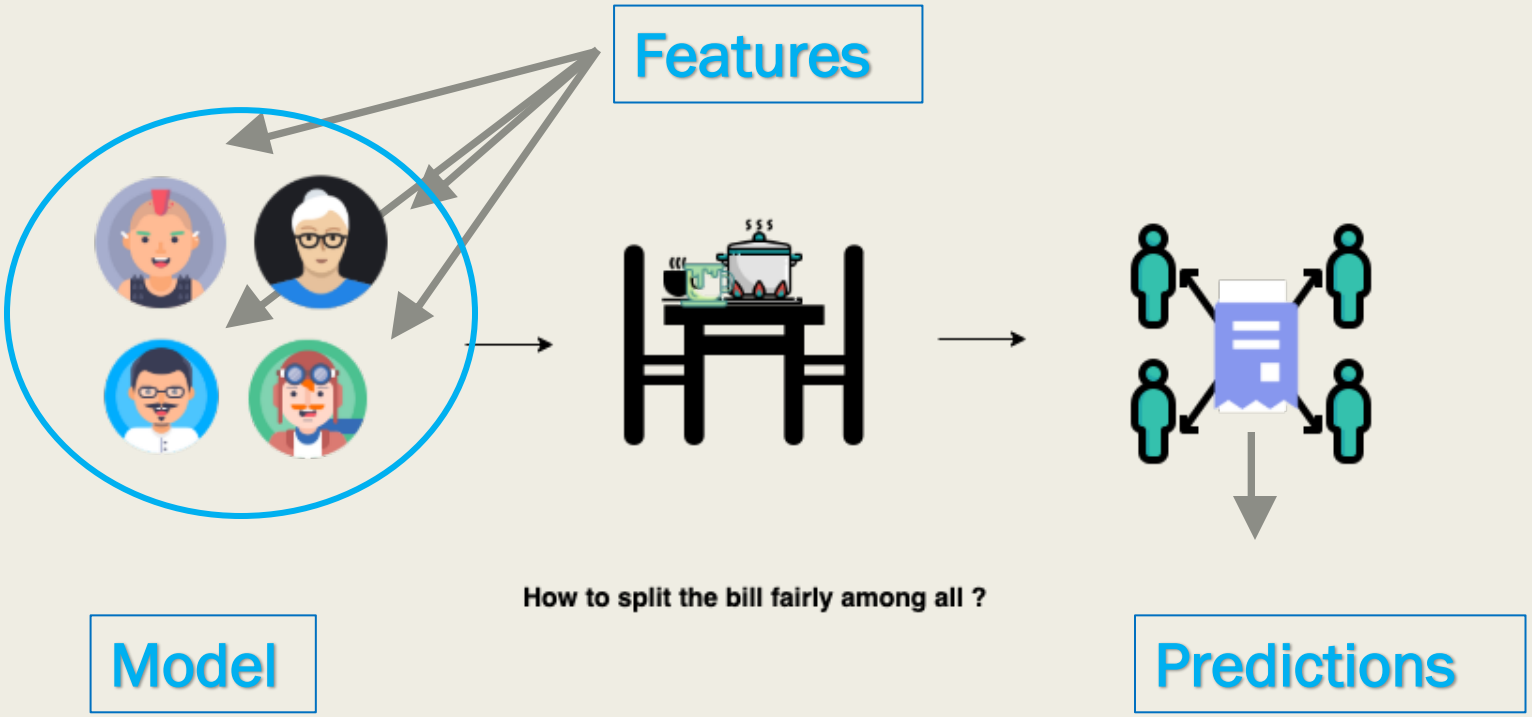
Cooperative Game Theory

Lloyd Shapley

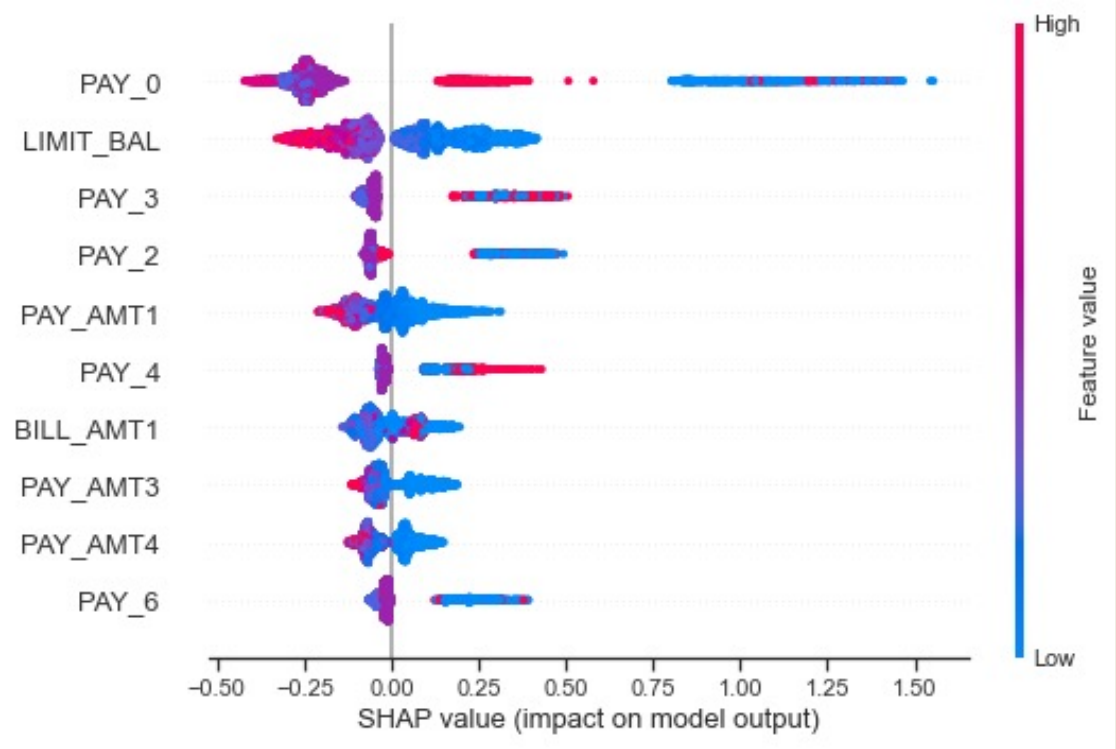


Lloyd Shapley, Nobel Prize Winner 2012 in economics. Bengt Nyman

SHAP values : Machine Learning context



Who will default on a Credit Card Payment?

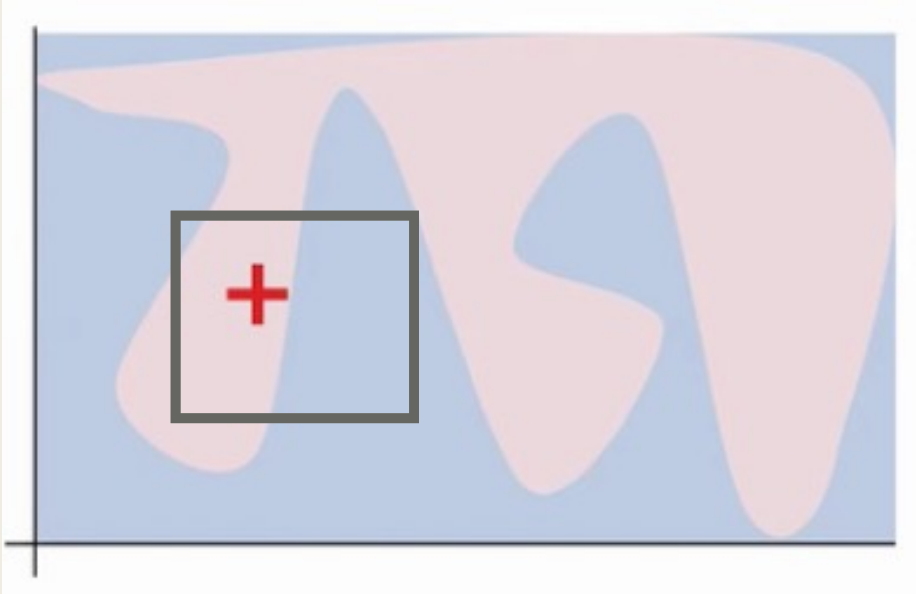


Data: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients#

Lime

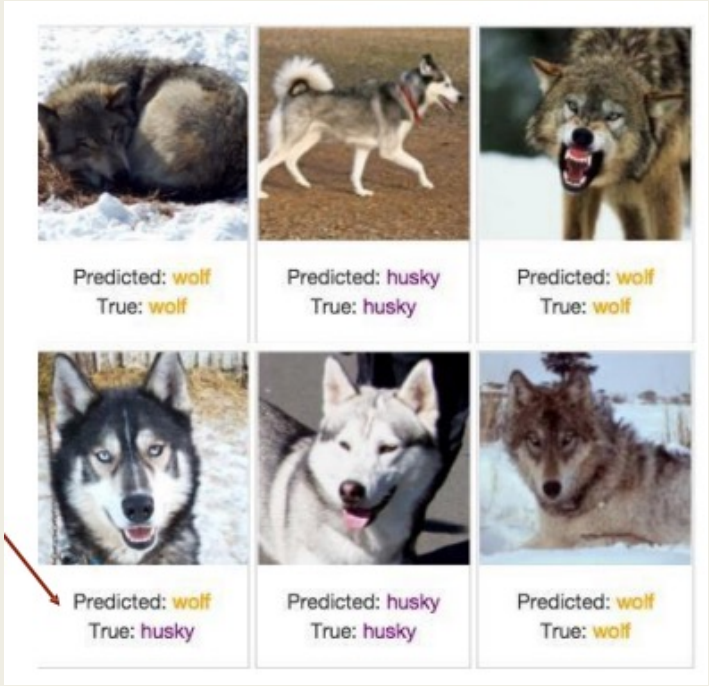
- LIME stands for Locally Interpretable Model-Agnostic Explanations.
- A technique to explain the predictions of any machine learning classifier – Model Agnostic.
- Approximates a given model by an interpretable one (such as a linear model)

Intuition



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

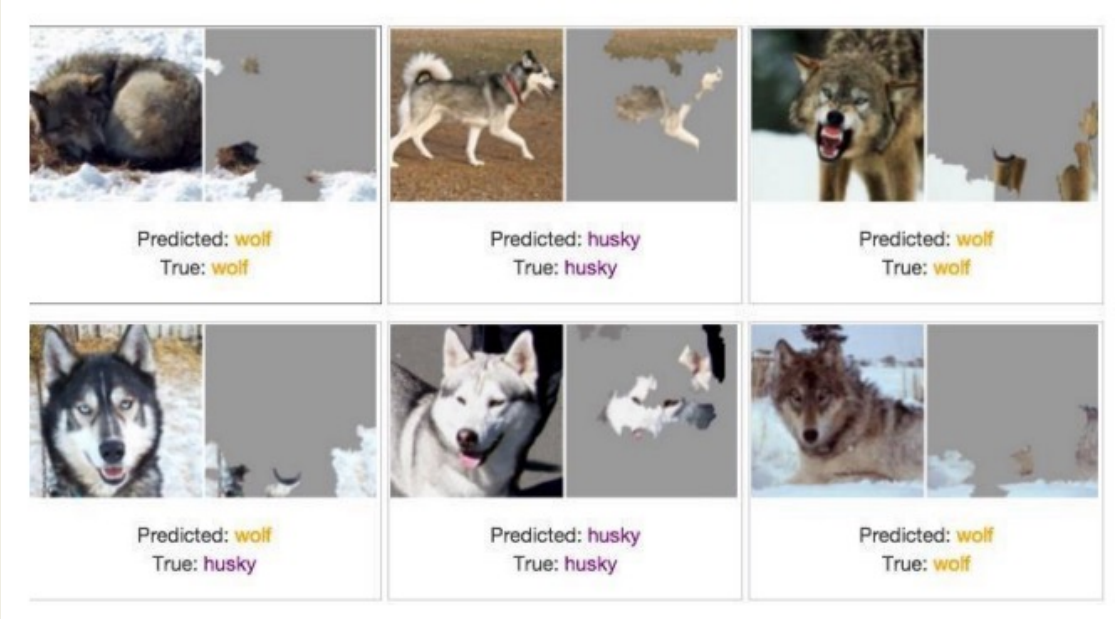
Lime for Image data



Husky or Wolf ?

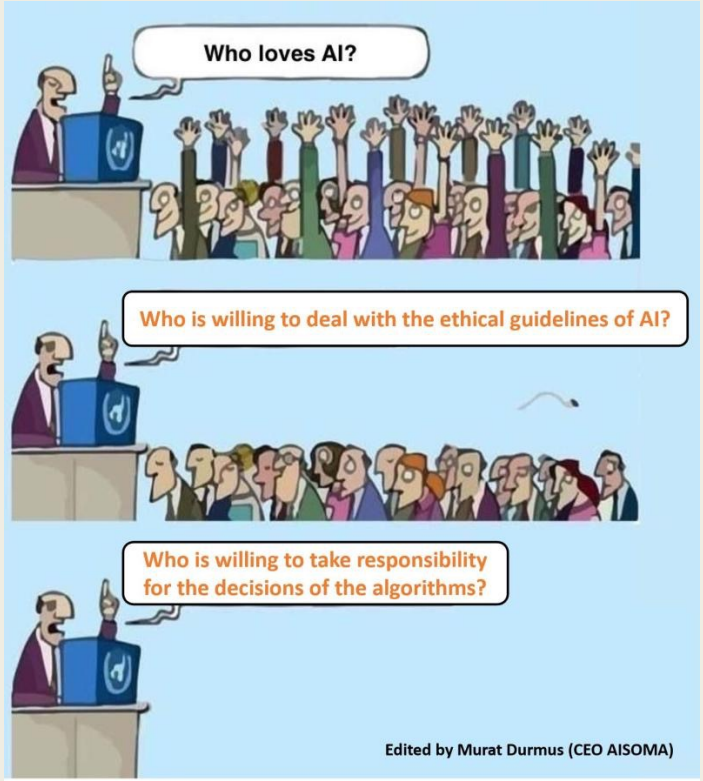
Explaining Black-Box Machine Learning Predictions - Sameer Singh, UCI

Lime for Image data



A great SNOW detector

Explaining Black-Box Machine Learning Predictions - Sameer Singh, UCI


Who loves AI?
Who is willing to deal with the ethical guidelines of AI?
Who is willing to take responsibility for the decisions of the algorithms?
Edited by Murat Durmus (CEO AISOMA)

Thankyou