



UNIVERSIDADE  
CATÓLICA  
PORTUGUESA

BRAGA

# Machine Learning

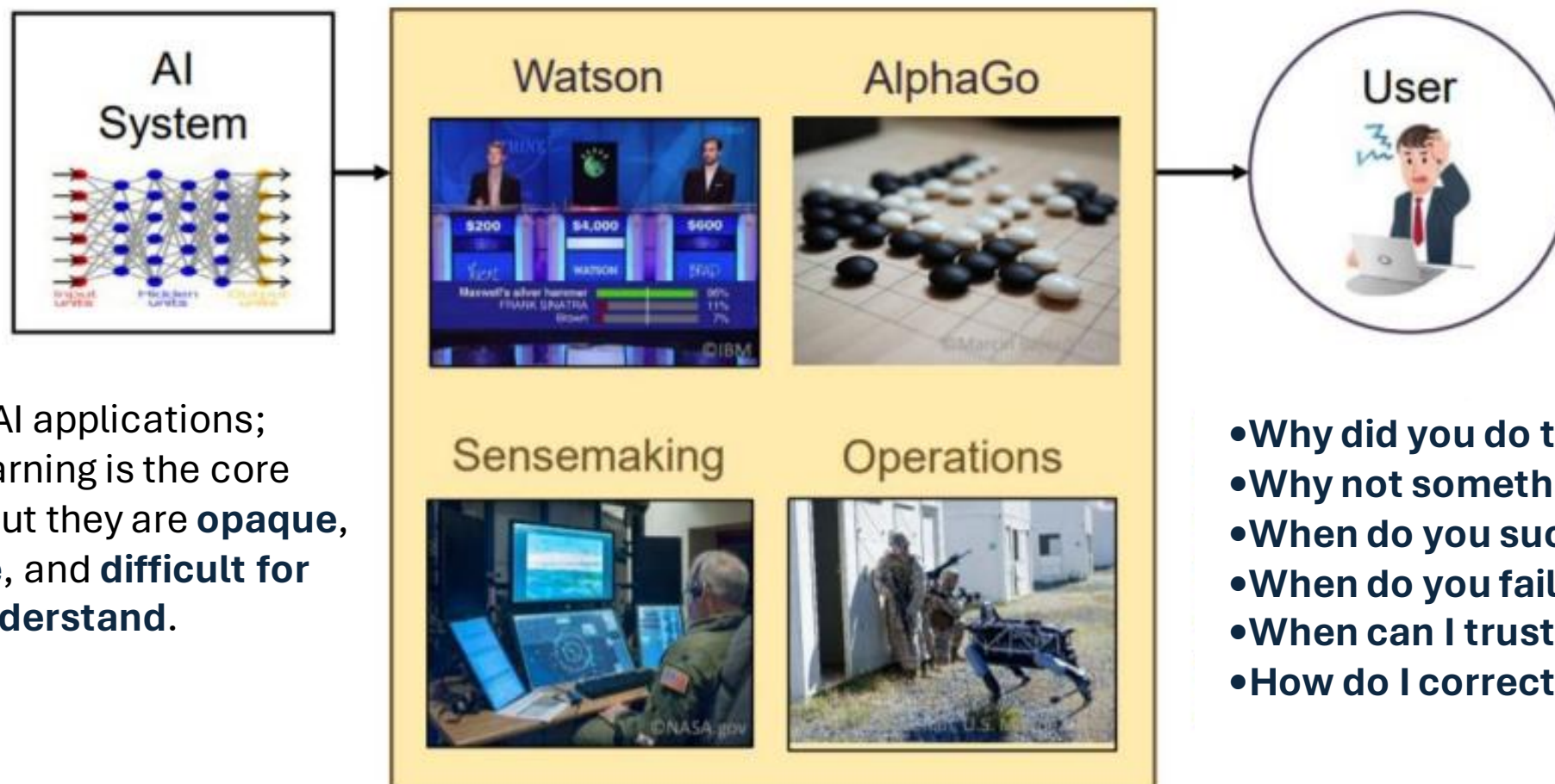
Session 23 - T

## Explainable AI (XAI)

Ciência de Dados Aplicada

2023/2024

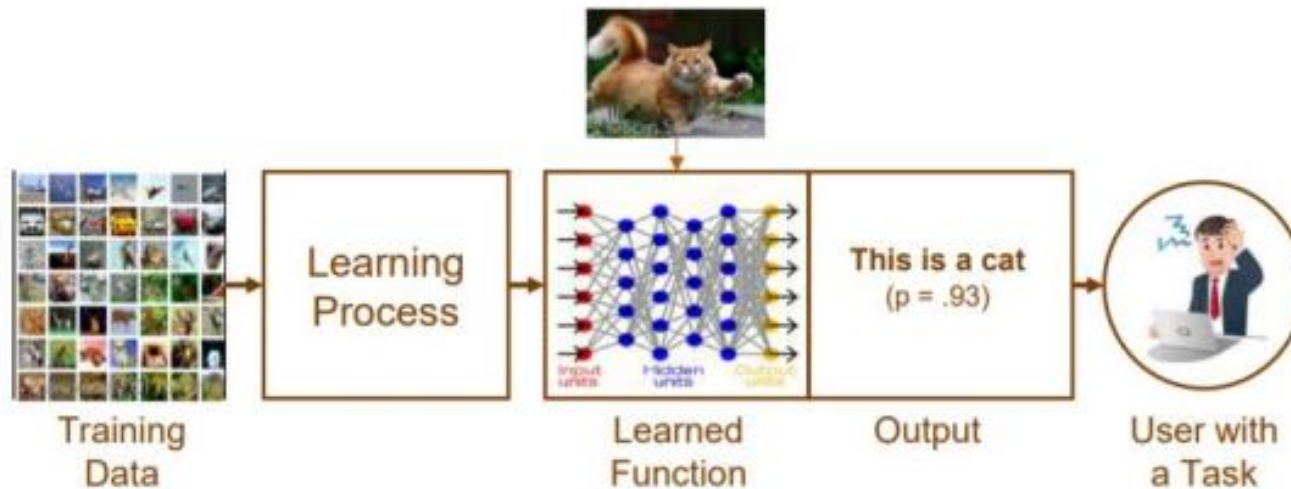
# Explainable AI (XAI) Motivation



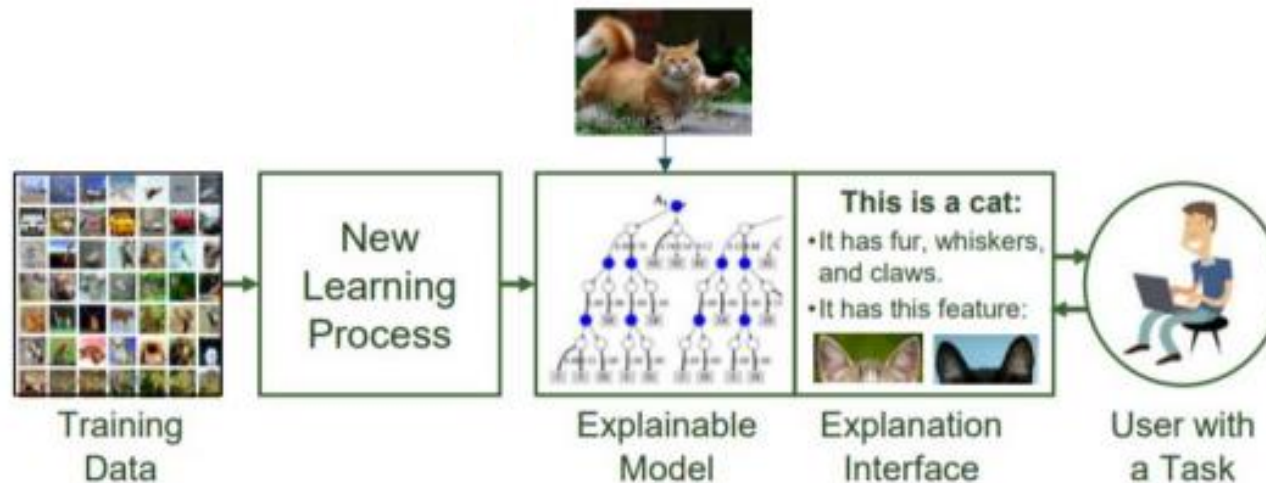
- New era of AI applications;
- Machine learning is the core technology, but they are **opaque**, **non-intuitive**, and **difficult** for people to understand.

- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

# XAI Objective



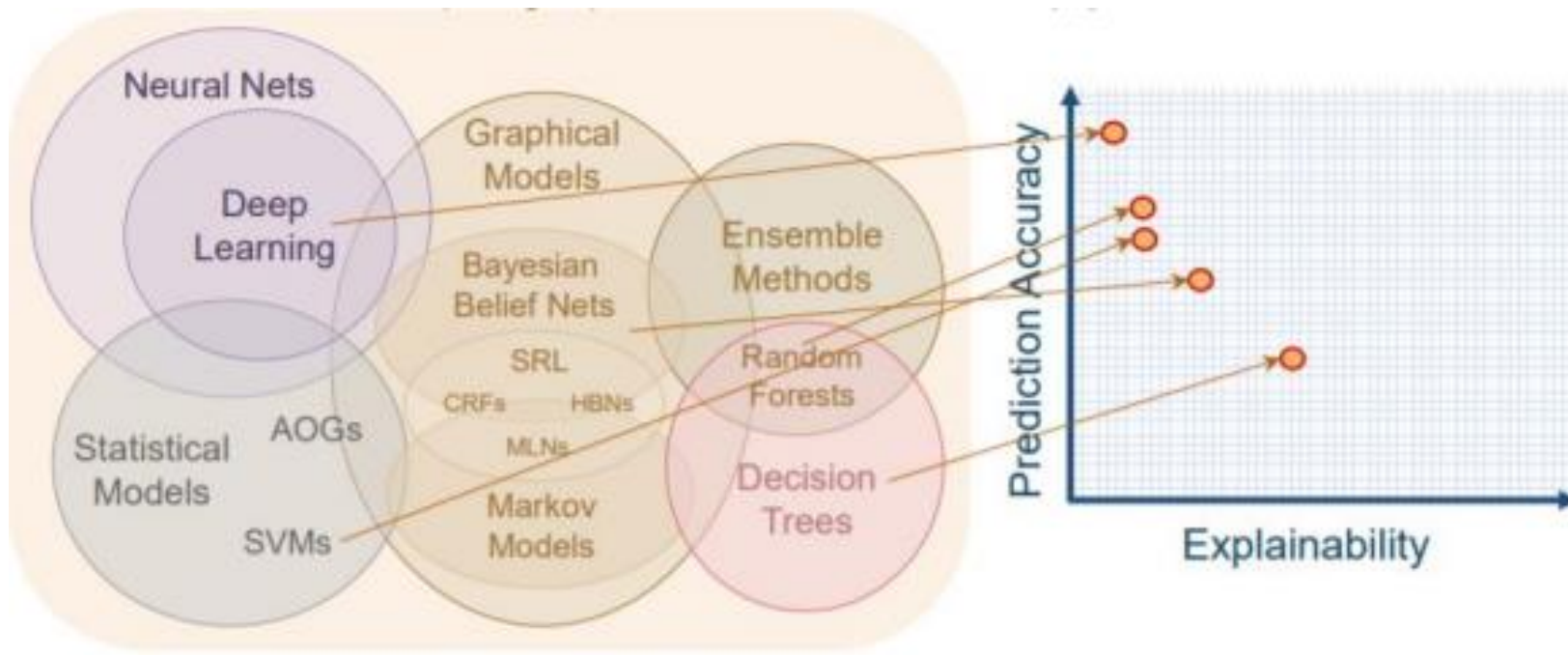
- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?



- I understand why.
- I understand why not.
- I know when you will succeed.
- I know when you will fail.
- I know when to trust you.
- I know why you made a mistake.

# Performance Vs Explainability

- **Challenge:** Develop machine learning techniques that produce more **explainable models** while maintaining a **high level of performance**.



# What is a good explanation?

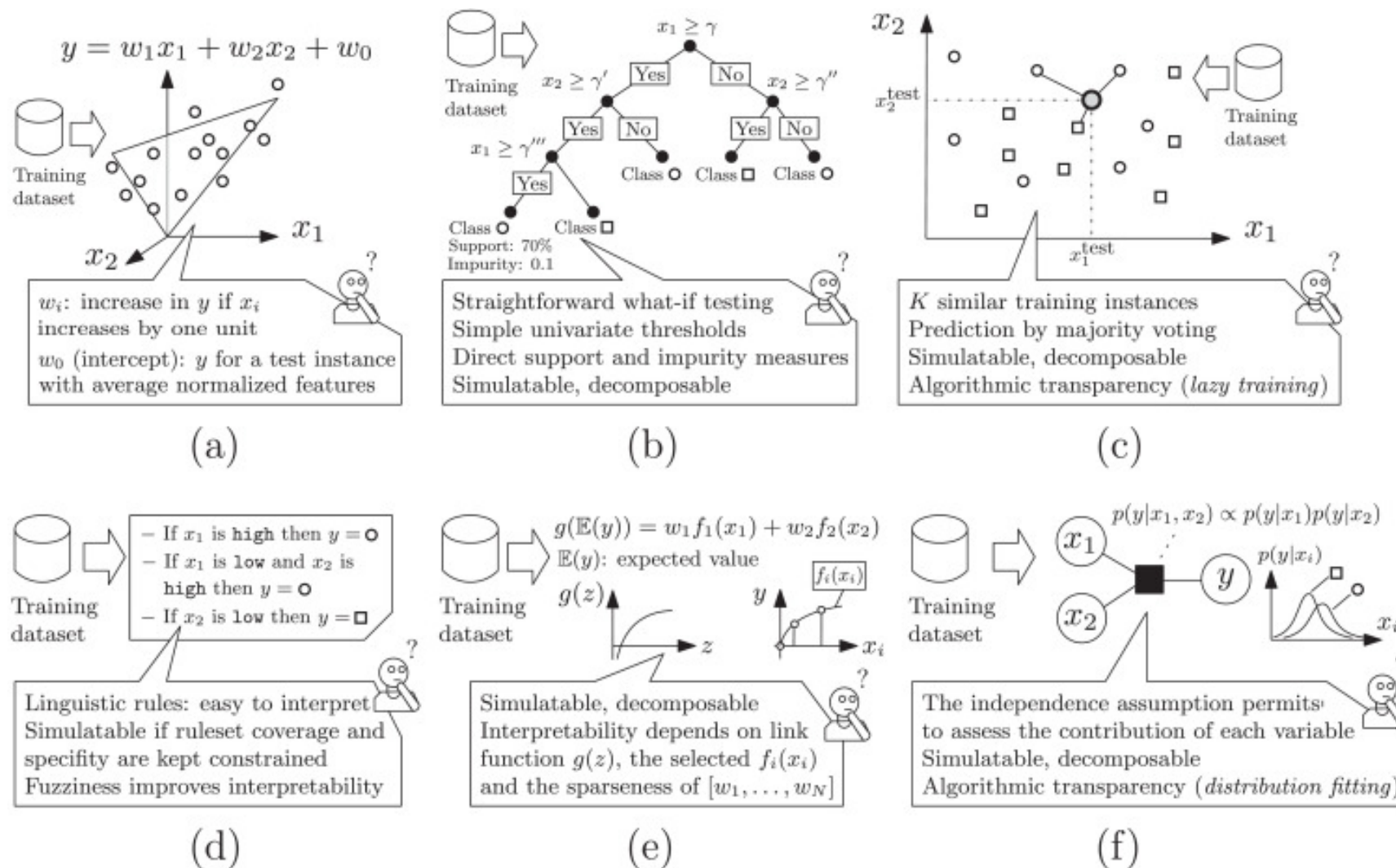
- Explanation not only answers "**why this**", but also "**why this instead of that**"!
- **Q:** “Why did Jane get the promotion (while Bob didn’t)?”
- **A1:** “Jane completed her project successfully.”
  - But John also completed his project successfully!
    - That doesn’t explain why she got the promotion!
- **A2:** “Jane completed her project successfully and consistently demonstrated leadership skills.”
  - Bob struggled with leadership, so this explains why Jane got the promotion and John did not.

# What is a good explanation?

- Explanation must be based on **relevant information!**
- **Q:** “Why did Jane get the promotion (while Bob didn’t)?”
- **A1:** “Jane completed her project successfully and wore glasses.”
  - But John also completed his project successfully, and wearing glasses shouldn't affect the promotion decision.
    - That doesn't explain why she got the promotion!
- But how do we decide that wearing glasses is not relevant, even if it might be **statistically significant**?



# XAI in Various ML Models



(a) Linear regression; (b) Decision trees; (c) K-Nearest Neighbors;  
(d) Rule-based Learners; (e) Generalized Additive Models; (f) Bayesian Models.

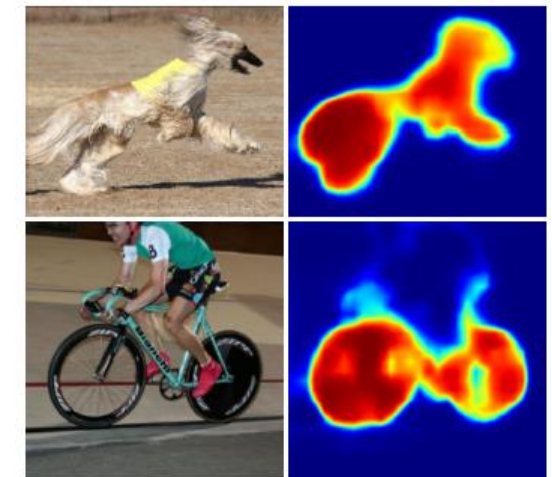
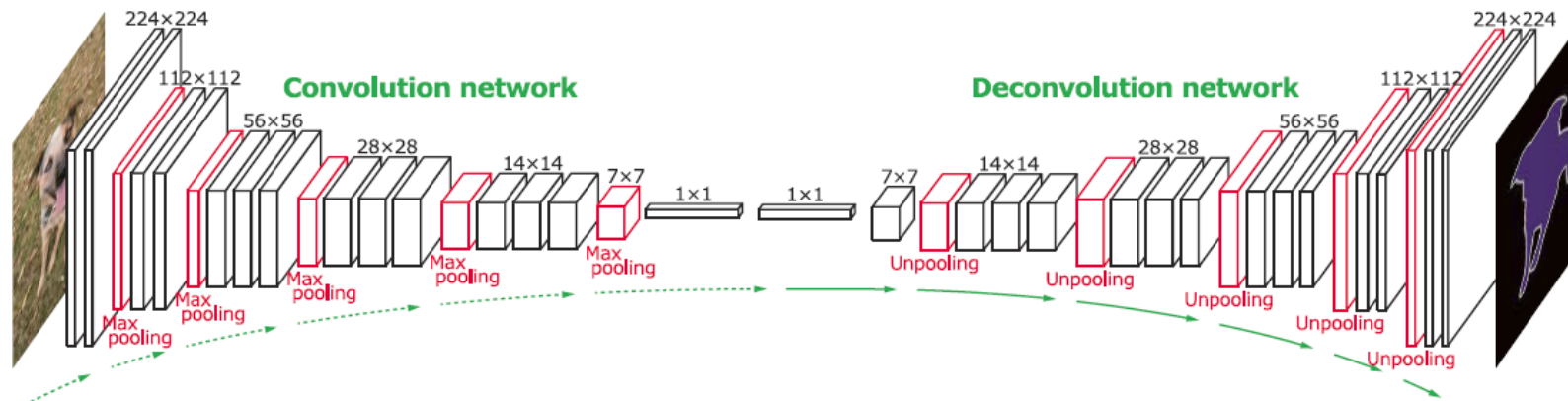
# XAI Approaches

- **Post-hoc:**
  - Applied to already developed models in order to understand how one produces predictions for given input;
  - Produce a separate algorithm which reads the end-to-end process.
- **In built:**
  - Build the decision-making algorithm so that traces have within them the basis for explanation.



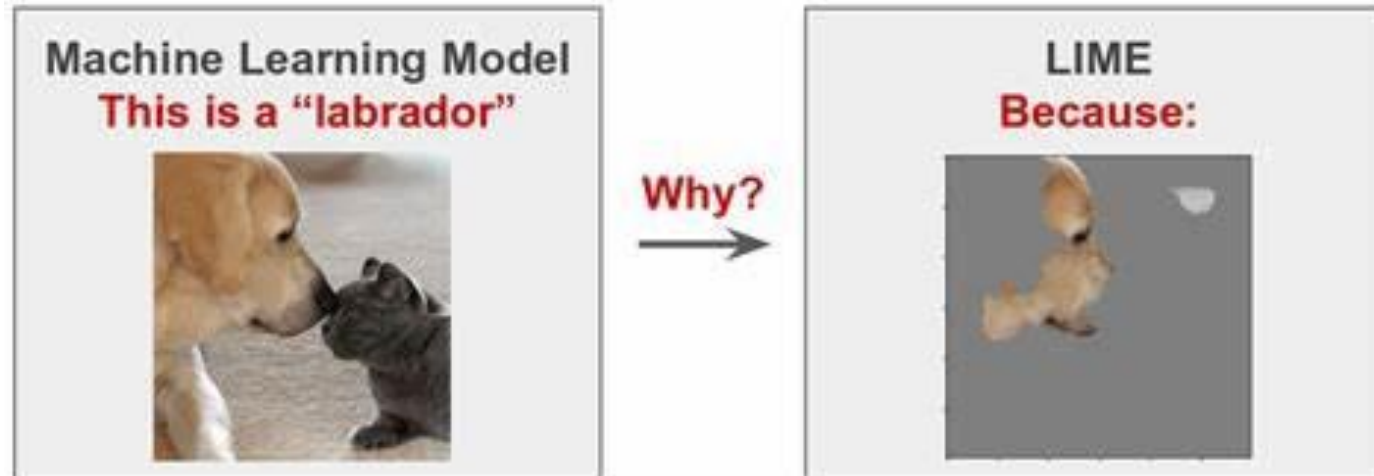
# Post-hoc Techniques: Model-Specific

- Post-hoc approach can be categorized into two approaches: **model-specific** and **model-agnostic**;
- One popular technique used in model-specific approaches is to **map back the output/prediction of a given input**, through the learned model, see which parts of the input were discriminative for the output.



# Post-hoc Techniques: Model-Agnostic

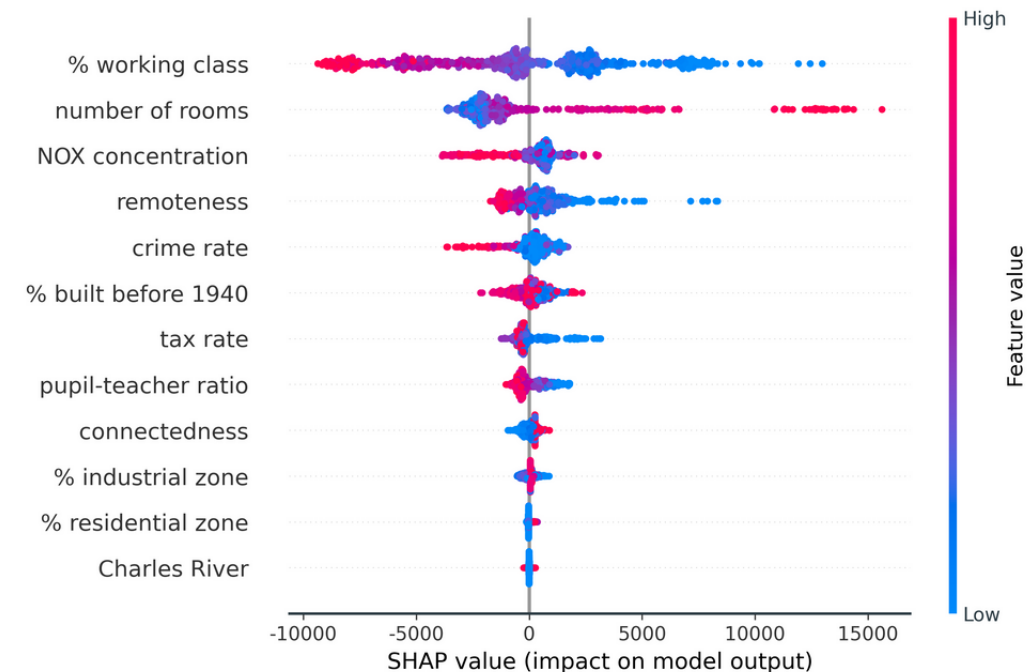
- Techniques used in **model-agnostic** approaches (i.e. treat the original model as a **black box**) are categorized into two groups:
  - **Explanation by simplification** approaches aim to extract underlying rules or na approximate **interpretable model** from the original model.
  - “Local Interpretable Model-Agnostic Explanations” (**LIME**) system:



# Post-hoc Techniques: Model-Agnostic

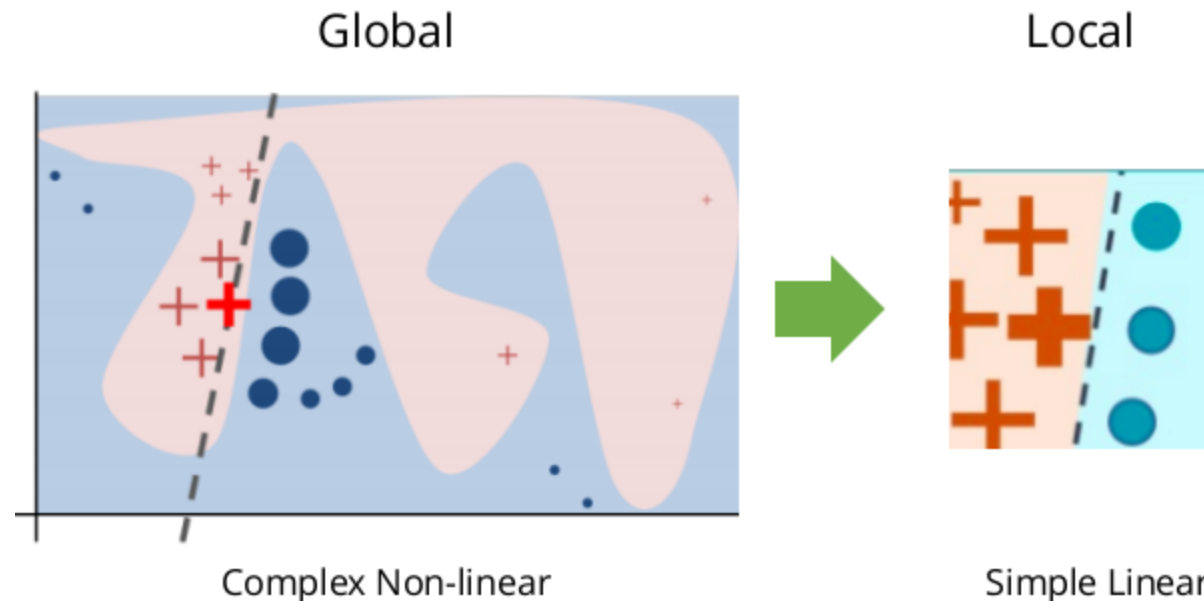
- Techniques used in **model-agnostic** approaches (i.e. treat the original model as a **black box**) are categorized into two groups:
  - **Feature relevance explanation** approach aims to describe the functioning of an opaque model by **measuring the influence and relevance of each feature** on prediction output.

- “Shapley additive explanations” (**SHAP**) system:



# Local Interpretable Model-Agnostic Explanations (LIME)

- LIME method was originally proposed by Ribeiro, Singh, and Guestrin (2016);
- The key idea behind it is to approximate a global model (which is a black-box) by **local models** which are simpler and transparent.

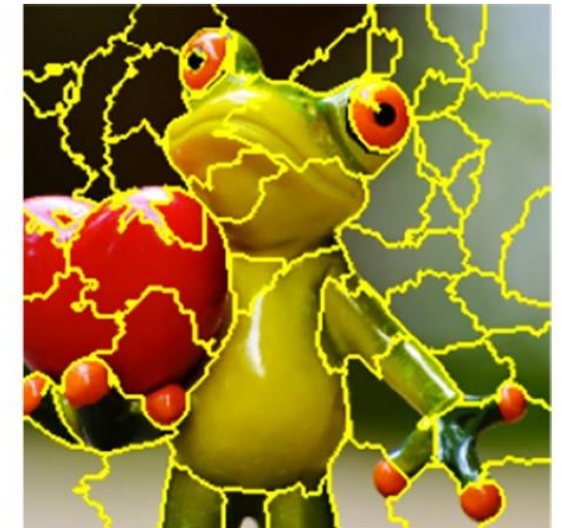


# LIME Method

- In order to be model-agnostic, LIME can't peak into the model. What LIME does to learn the behavior of the underlying model is to first **perturb the input** (e.g., removing words or hiding parts of the image);
- For images, an original image is divided into interpretable components (contiguous superpixels).



Original Image



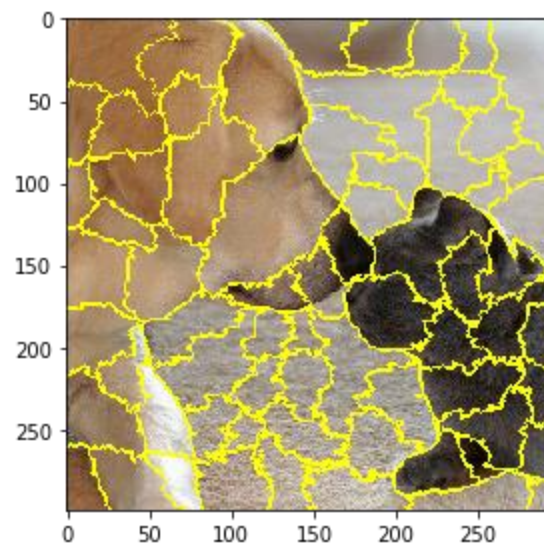
Interpretable  
Components



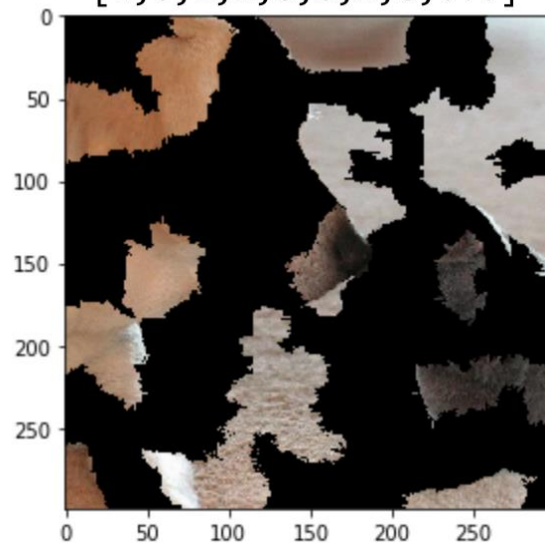
# LIME Method



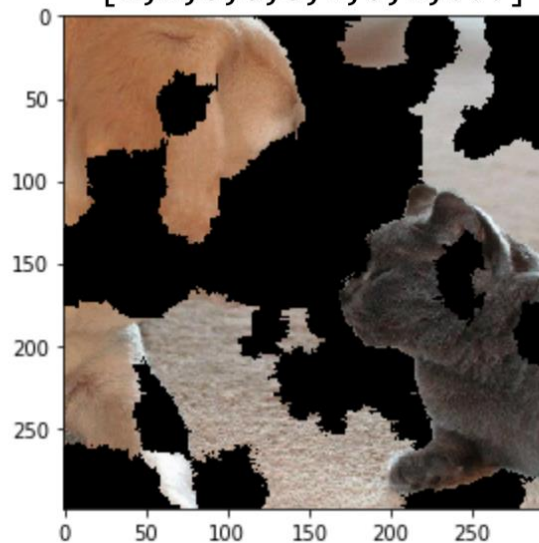
Original image  
segmented into  
150 superpixels.



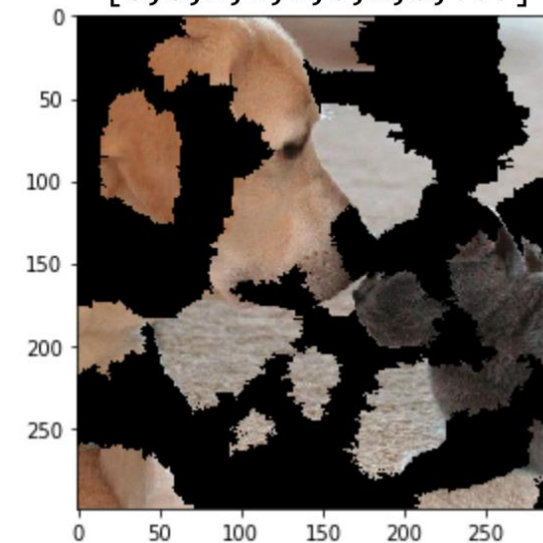
perturbation1=  
[1,0,1,1,0,0,1,0,...]



perturbation2=  
[1,1,0,0,0,1,0,1,...]



perturbation3=  
[0,0,1,1,1,0,1,0,...]











- **Perturbation for text data:**

- For example, if we are trying to explain the prediction of a text classifier for the sentence:
  - “I hate this movie”, we will perturb the sentence and get predictions on sentences such as
    - “I hate movie”,
    - “I this movie”,
    - “I movie”,
    - “I hate”, etc.

# LIME Method

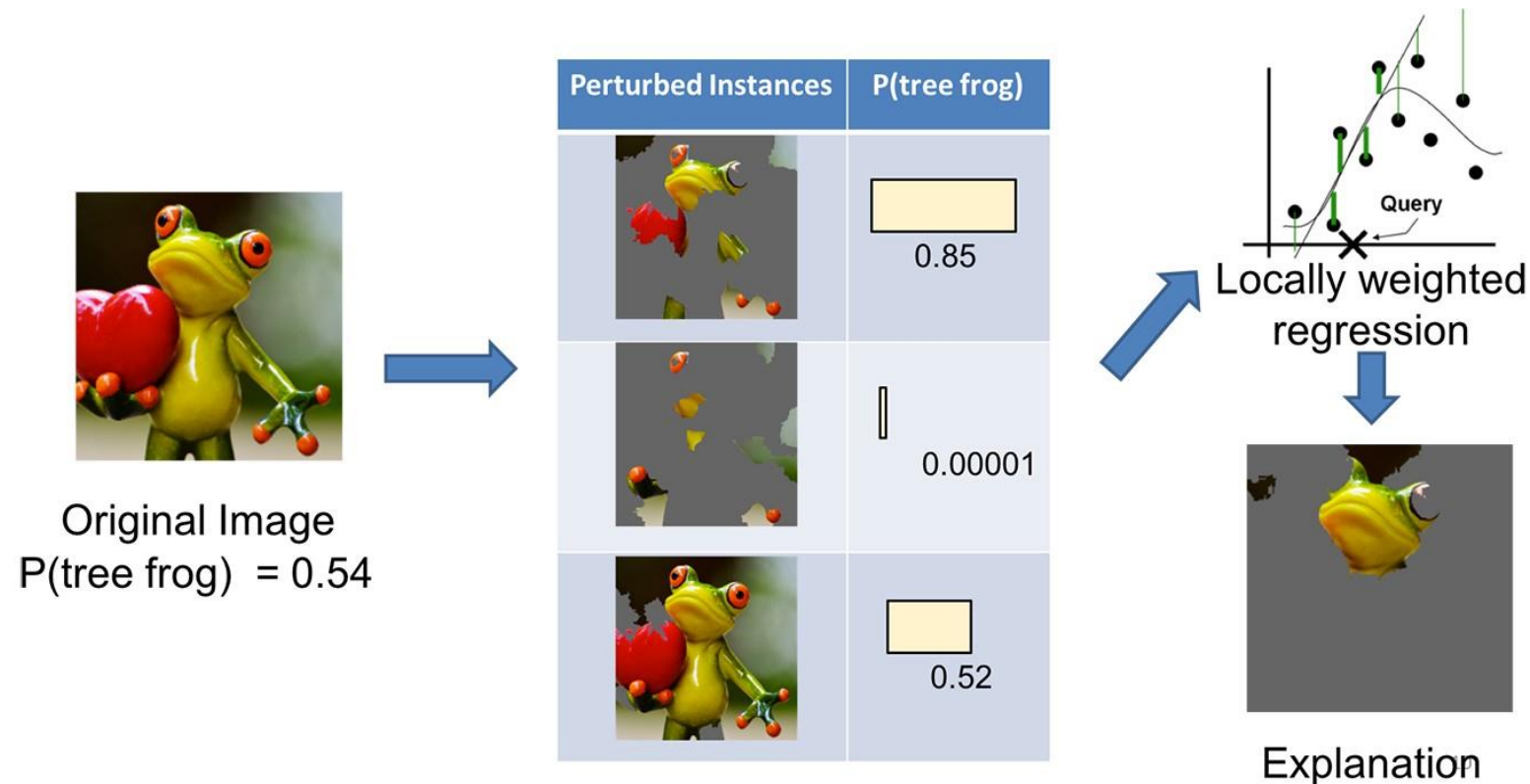
- Then LIME run the perturbed data in the model and see how the predictions change.



Perturbed Instances	$P(\text{tree frog})$
	 0.85
	 0.00001
	 0.52

# Resources

- Then LIME weights these perturbed data points by their proximity to the original example and learns an interpretable model on those and the associated predictions.

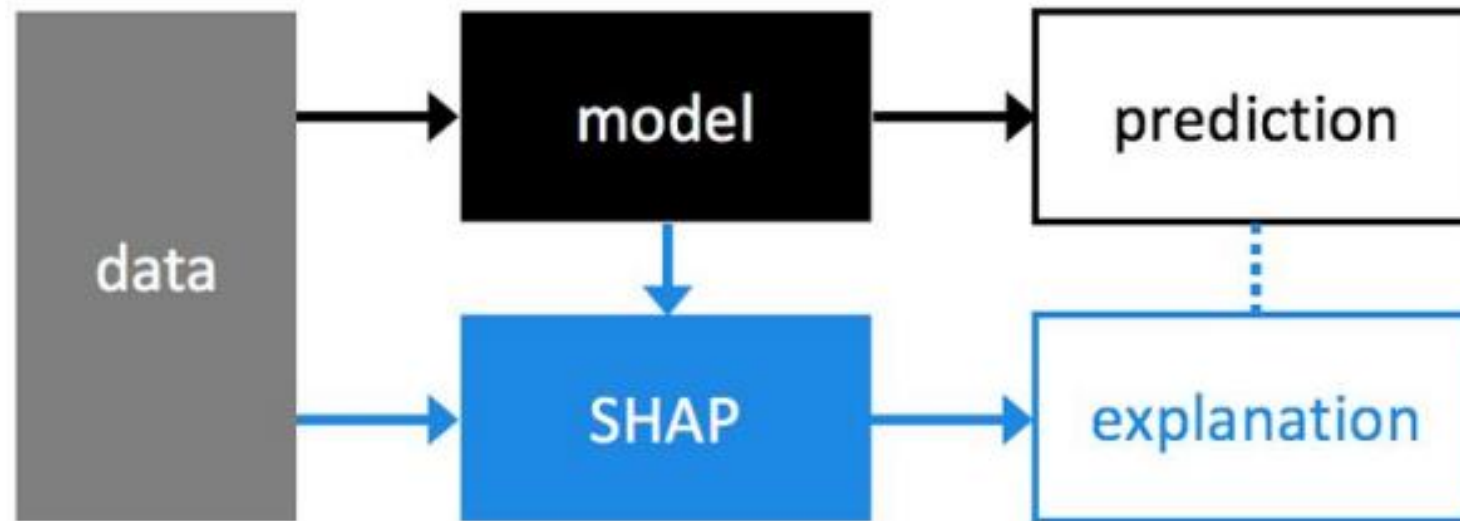


# LIME Algorithm

1. Sample the locality around the selected single data point uniformly and at random and generate a dataset of **perturbed data points** with its corresponding prediction from the model we want to be explained.
2. Use the specified feature selection methodology to select the number of **features** that is required for explanation.
3. Calculate the **sample weights** using a kernel function and a distance function. (this captures how close or how far the sampled points are from the original point).
4. Fit an interpretable model (**locally weighted linear regression**) on the perturbed dataset using the sample weights to weigh the objective function (e.g. **squared error**).
5. Provide local explanations using the newly trained interpretable model.

# SHapley Additive exPlanations (SHAP)

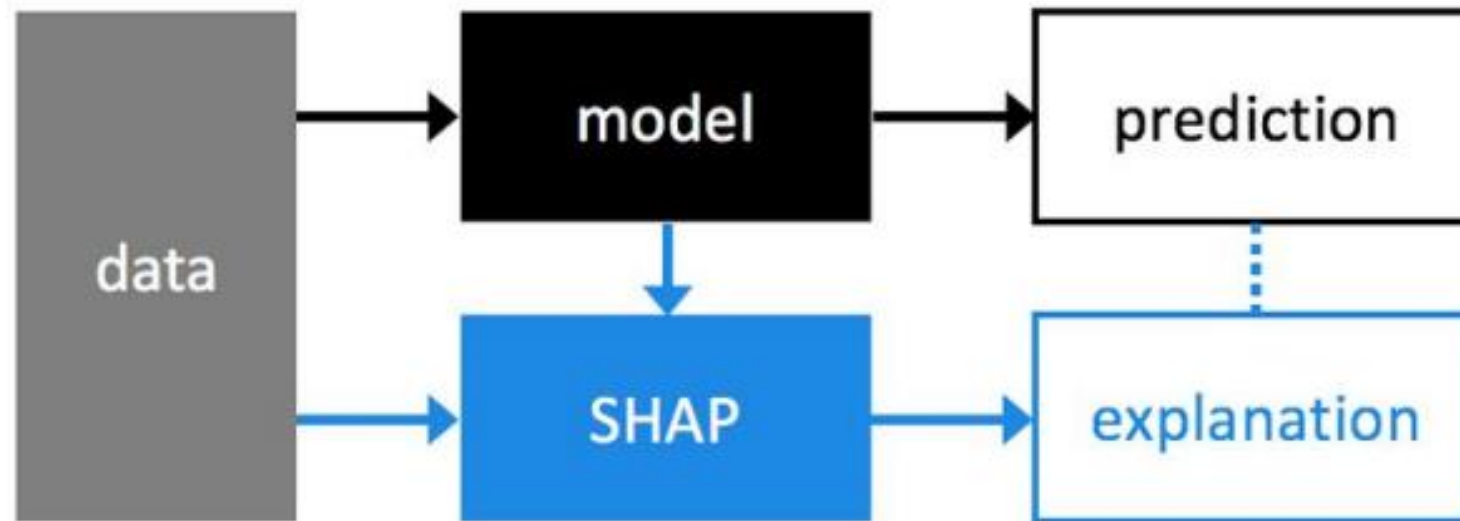
- Additive **feature attribution** method to explain the output of any ML model.



- It assigns each feature an importance value for a particular prediction.

# SHapley Additive exPlanations (SHAP)

- Additive **feature attribution** method to explain the output of any ML model.

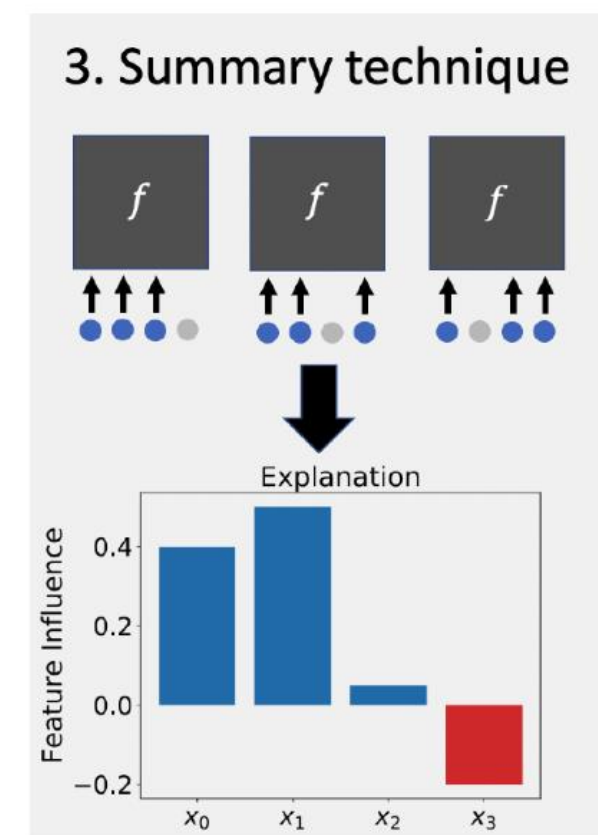
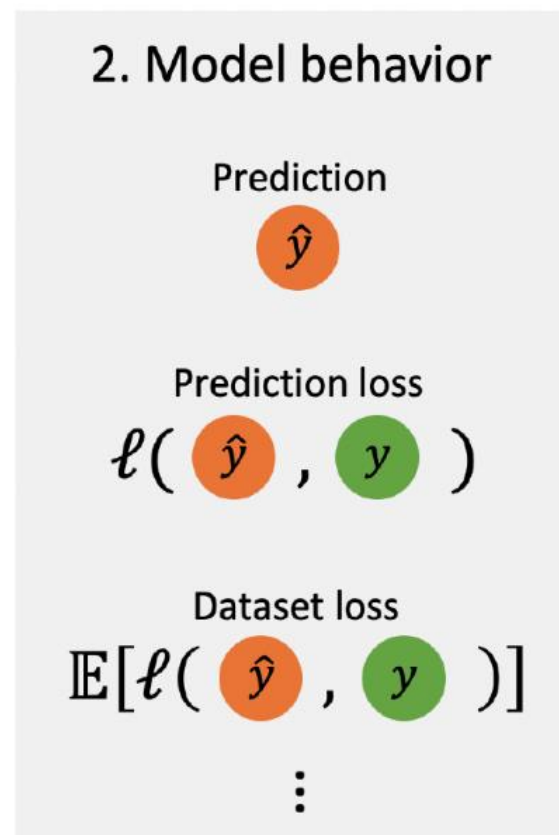
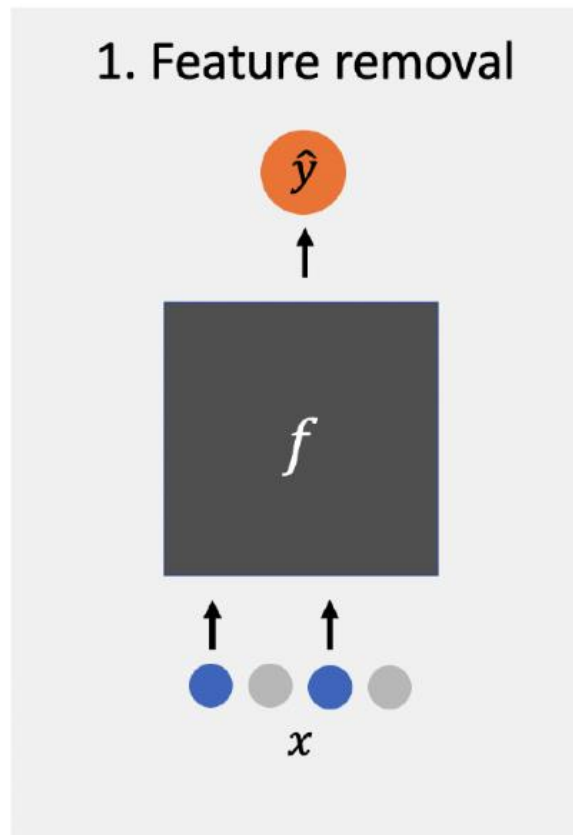


- It assigns each feature an importance value for a particular prediction.

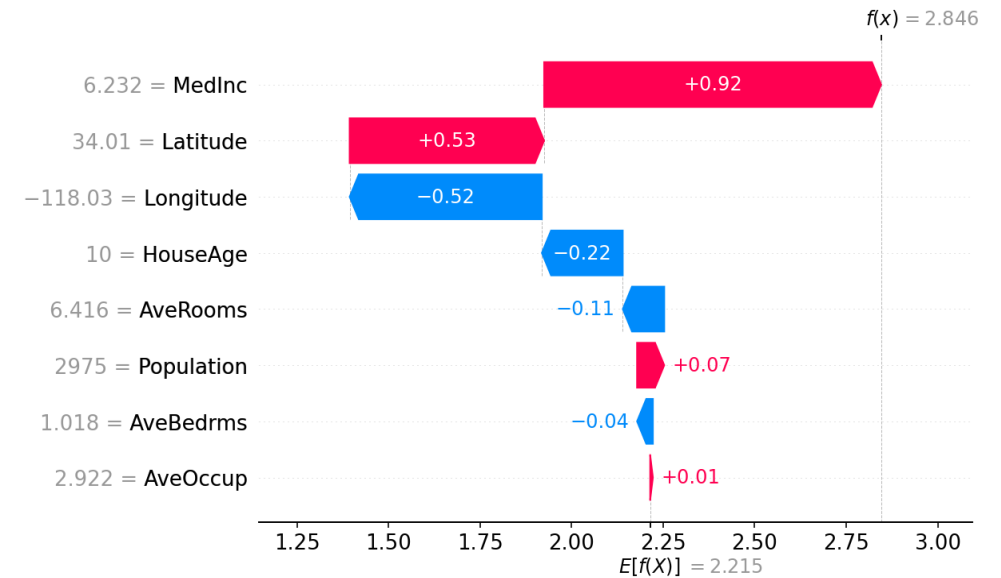
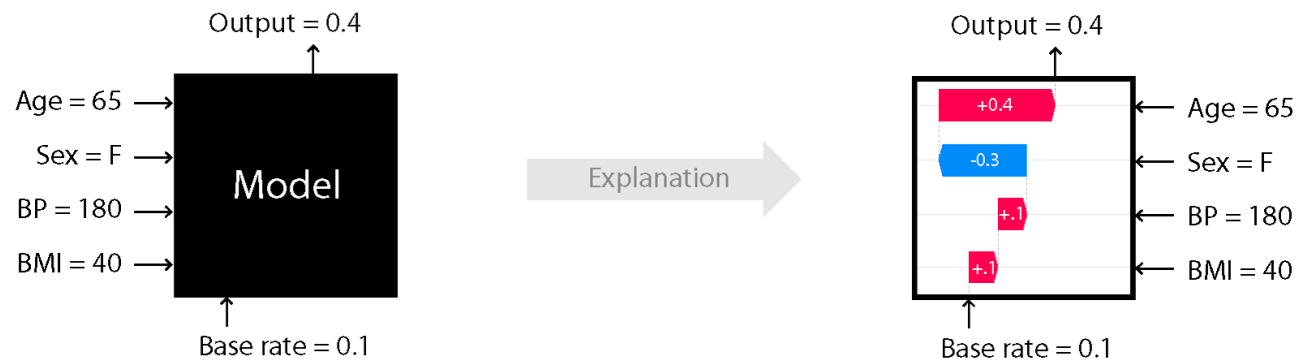


# SHapley Additive exPlanations (SHAP)

- The removal-based explanations framework



# SHAP



# Resources

- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why Should I Trust You?": Explaining the Predictions of Any Classifier (Version 3). arXiv. <https://doi.org/10.48550/ARXIV.1602.04938>
- <https://github.com/marcotcr/lime/tree/master>
- Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.1705.07874>
- <https://github.com/shap/shap>