Interpretable Machine Learning

Innopolis University
Applied Machine Learning Learning
BEKKOUCH Imad Eddine Ibrahim

Learning Outcomes

Interpretability & Interpretable models

The importance, applications, scope of interpretability, and simple models.

Local Surrogate (LIME)

Using a simple model to explain a more complicated one.

GradCam

Class level explanations, Won’t the Grad-CAM Heatmap Be Too Small?

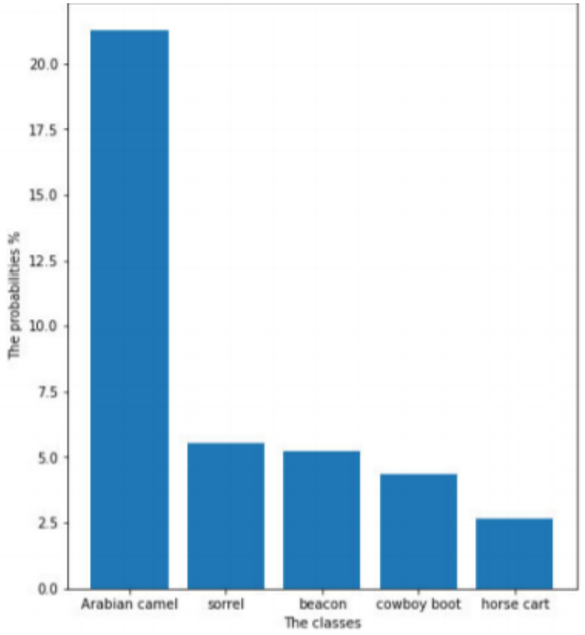
Interpretability

Interpretability is the degree to which a human can understand the cause of a decision.

- **Interpretability**
  What is it and why is it important
- **Scope of interpretability**
  types of interpretations
- **Interpretable models**
  Decision tress, logistic regression

Interpretability

Interpretability is the degree to which a human can consistently predict the model's result



Importance of Intepretability

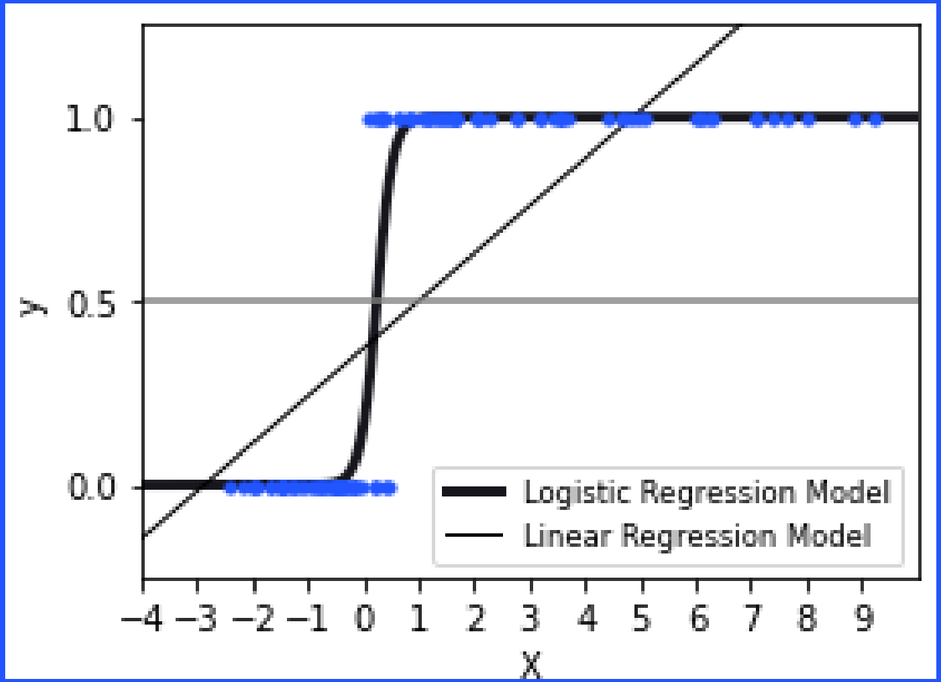
why do not we just trust the model and ignore why it made a certain decision?

- Satisfying the human curiosity
- Improving recommendations
- safety measures
- Detecting biases
- debugged and audited

Scope of Intepretability

What are the different levels of interpreting a model?

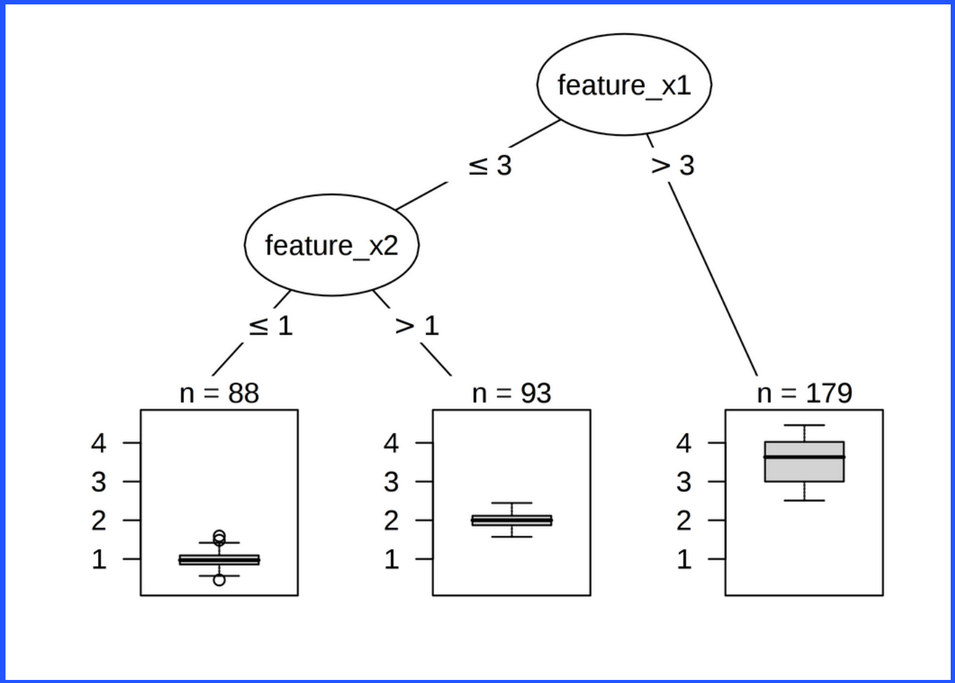
- Algorithm Transparency
- Global, Holistic Model Interpretability
- Global Model Interpretability on a Modular Level
- Local Interpretability for a Single Prediction
- Local Interpretability for a Group of Predictions



Logistic Regression

Interpretation:

1. Numerical Features
2. Binary Features
3. Categorical Features
4. Bias



Decision Trees

How to explain a prediction on a certain sample?
How to change the class of a prediction?
How to measure the importance of features?

Model Interetability

1. Interpretability

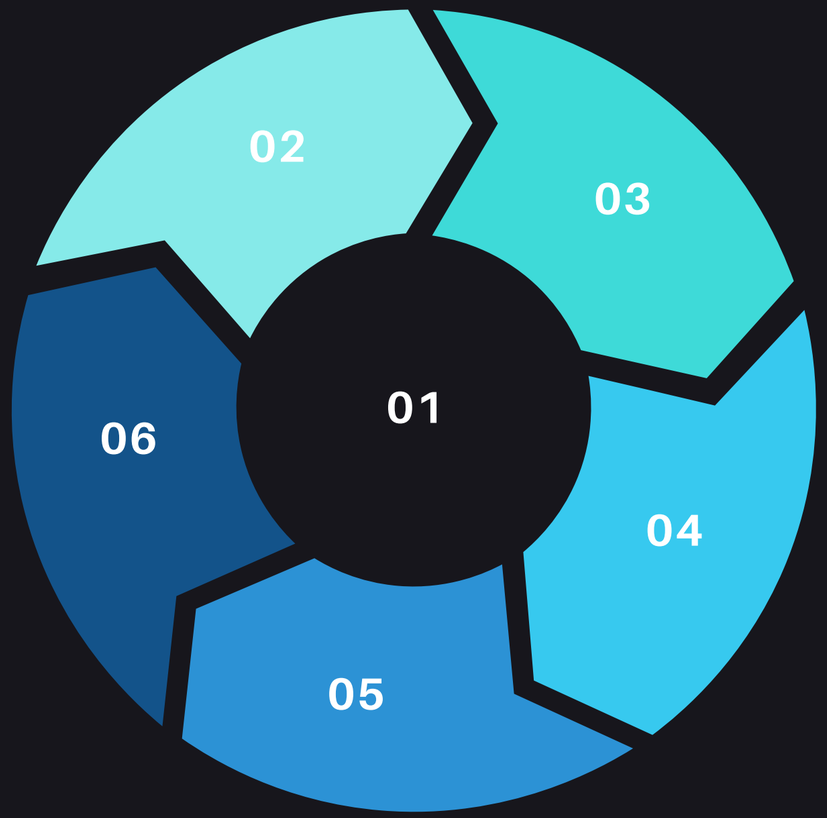
Can a human understand why?

2. Importance of interpreatibly

Medical imaging, loans

3. Scope of interpretability

Understanding the model and understand a decision made by the model



4. Interpretable models

logistic regression

5.Lime

Use a simple model to explain a more complicated one

6. Grad cam

class-level explanations on images

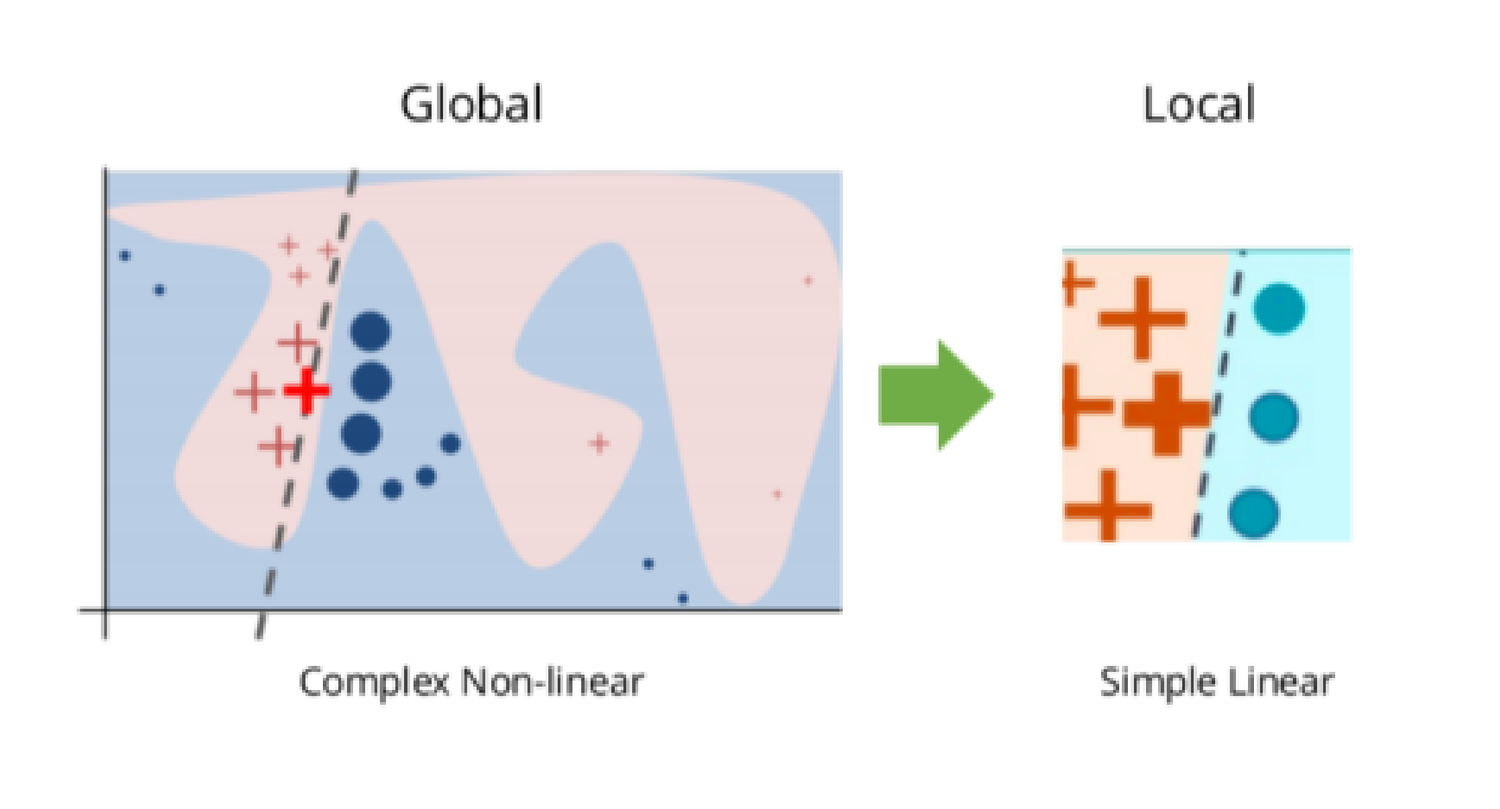
Local Surrogate (LIME)

Local interpretable model-agnostic explanations (LIME)

- **Tabular Data**
  Steps, example
- **Text**
  Steps, example
- **Images**
  Steps, example

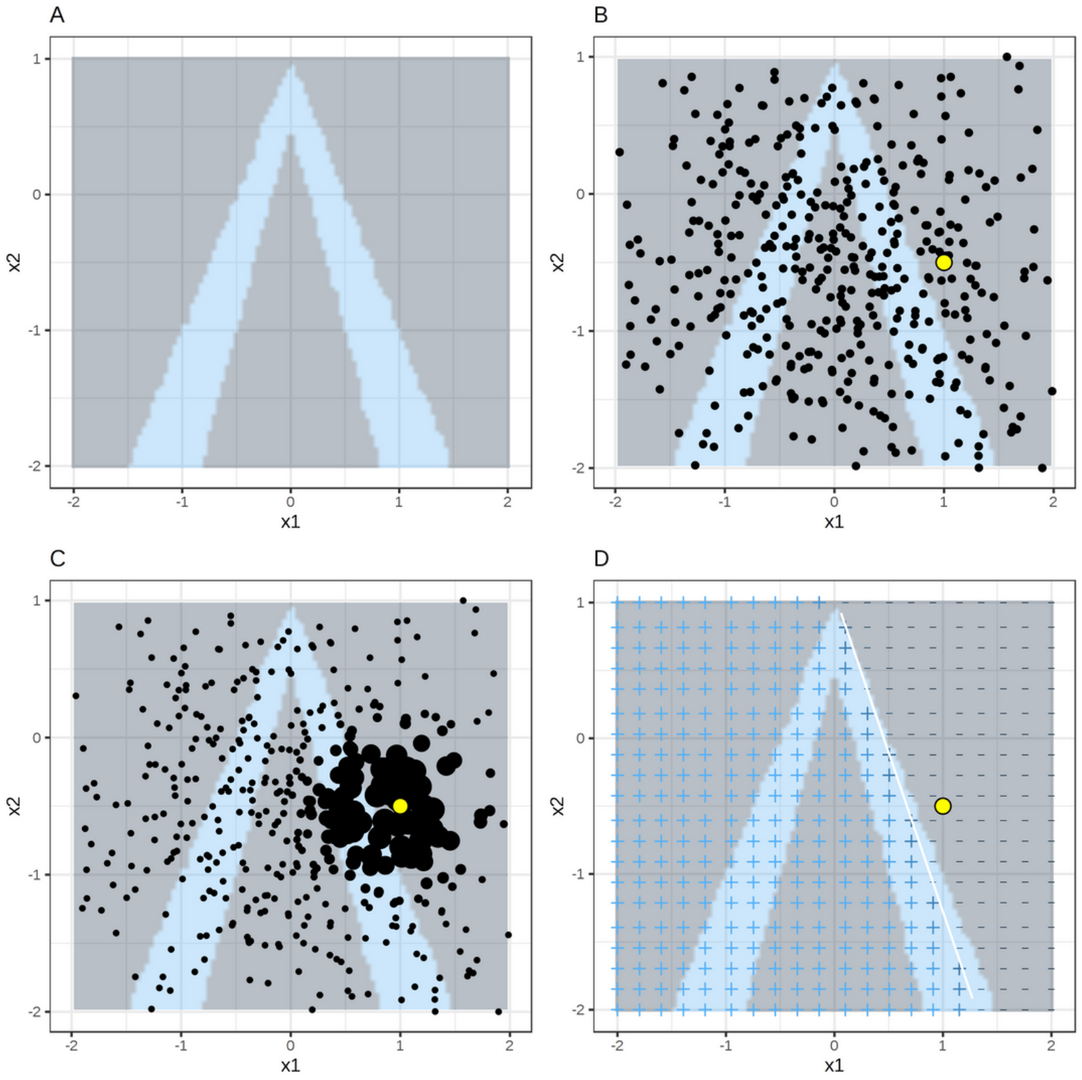
LIME

Approximate the decision locally using a simple model.



Lime for tabular data

- Select your instance of interest.
- Perturb your dataset and get the black box predictions for these new points.
- Weight the new samples according to their proximity to the instance of interest.
- Train a weighted, interpretable model on the dataset with the variations.
- Explain the prediction by interpreting the local model.



Lime for textual data

For Christmas Song visit my channel! ;) - Spam detection

	For	Christmas	Song	visit	my	channel!	;)	prob	weight
2	1	0	1	1	0	0	1	0.17	0.57
3	0	1	1	1	1	0	1	0.17	0.71
4	1	0	0	1	1	1	1	0.99	0.71
5	1	0	1	1	1	1	1	0.99	0.86
6	0	1	1	1	0	0	1	0.17	0.57

Lime for textual data

For Christmas Song visit my channel! ;) - Spam detection

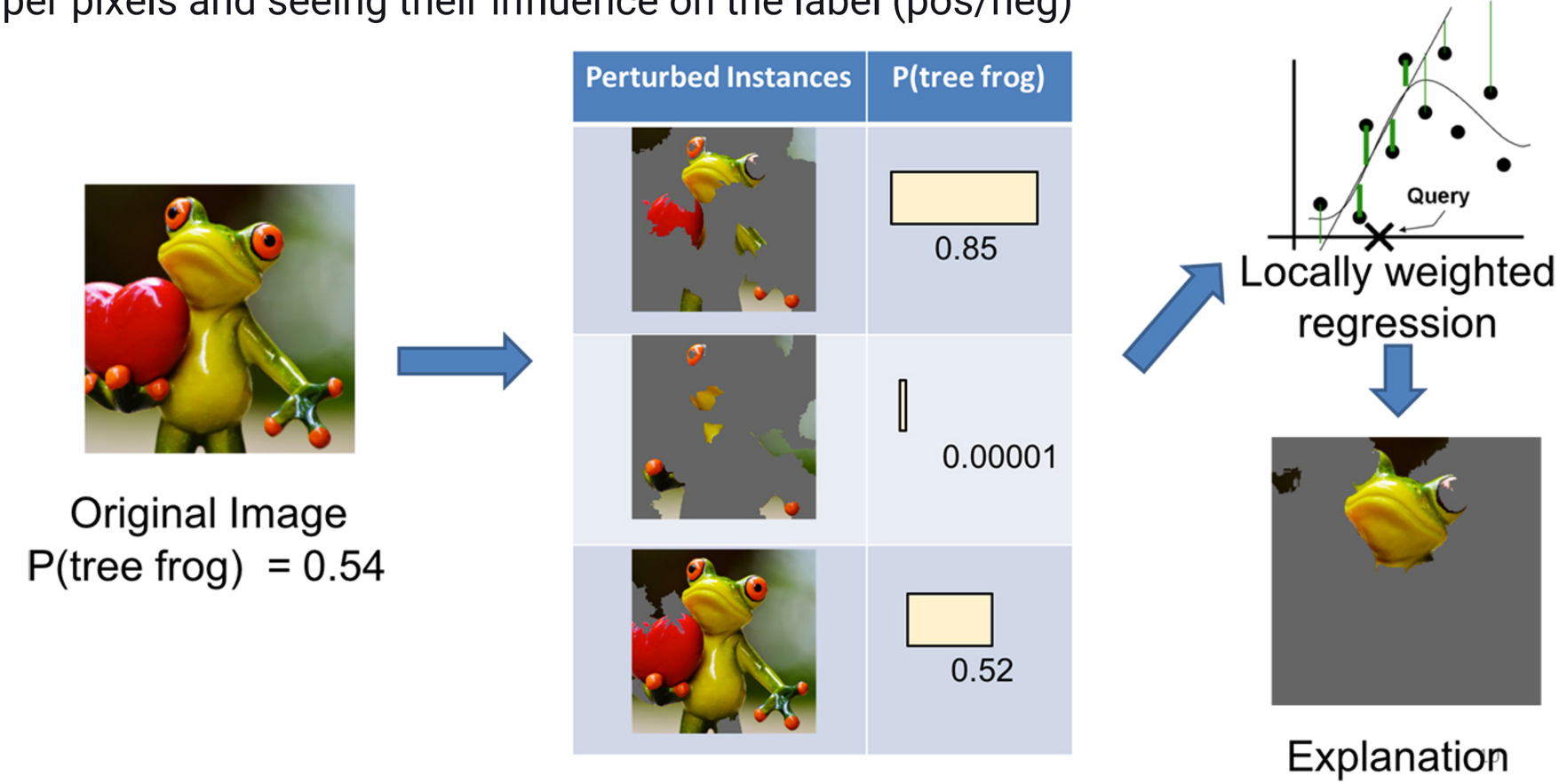
case	label_prob	feature	feature_weight
1	0.1701170	is	0.000000
1	0.1701170	good	0.000000
1	0.1701170	a	0.000000
2	0.9939024	channel!	6.180747
2	0.9939024	Christmas	0.000000
2	0.9939024	Song	0.000000

Super pixels

Superpixels are interconnected pixels with similar colors

Lime for image data

Switching super pixels and seeing their influence on the label (pos/neg)



Lime for image data

Switching super pixels and seeing their influence on the label (pos/neg)

Model Interetability

1. Interpretability

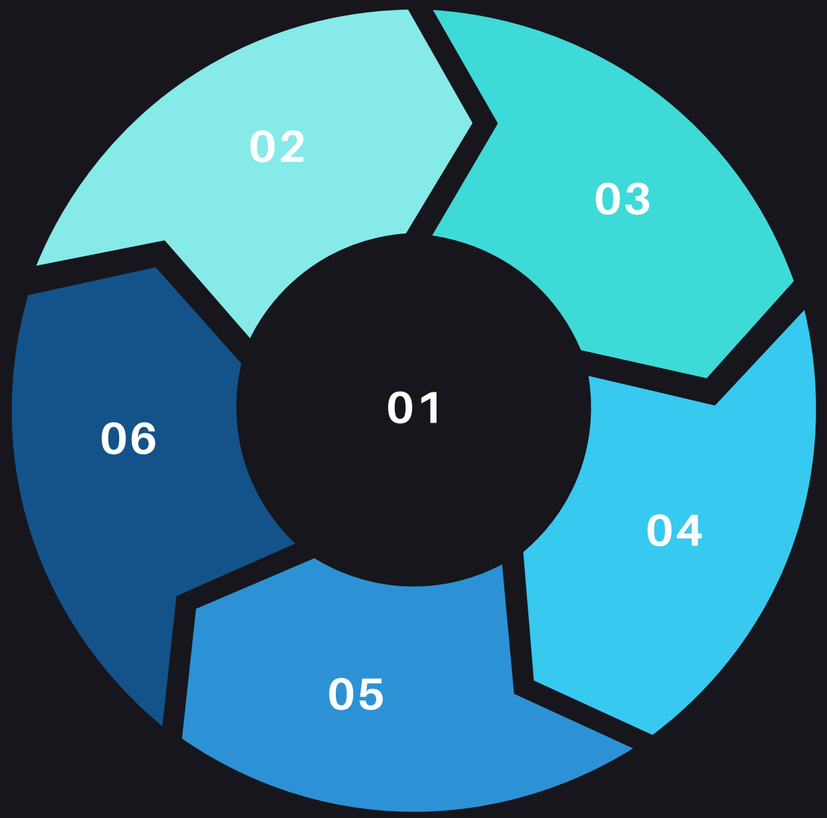
Can a human understand why?

2. Importance of interpreatibly

Medical imaging, loans

3. Scope of interpretability

Understanding the model and understand a decision made by the model



4. Interpretable models

logistic regression

5.Lime

Use a simple model to explain a more complicated one

6. Grad cam

class-level explanations on images

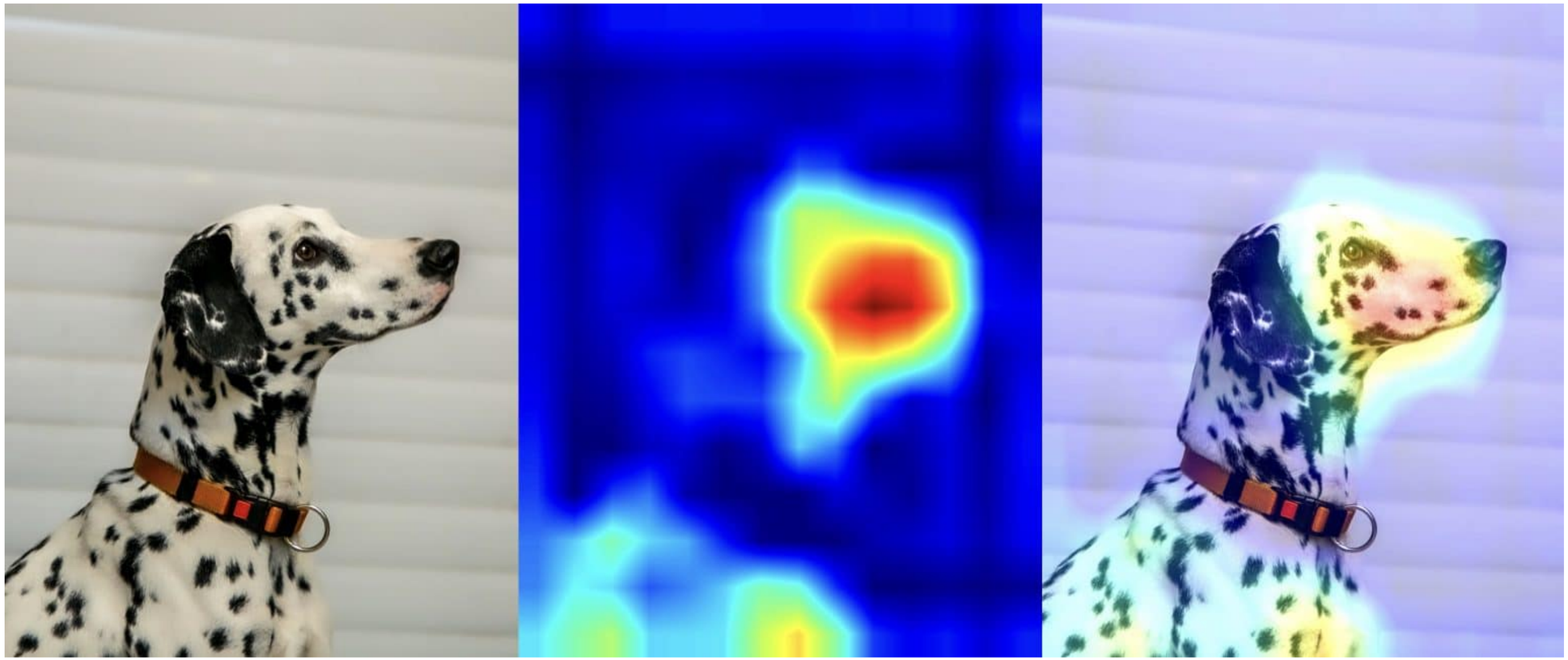
Grad Cam

Grad-CAM is a popular technique for creating a class-specific heatmap based on a particular input image, a trained CNN, and a chosen class of interest.

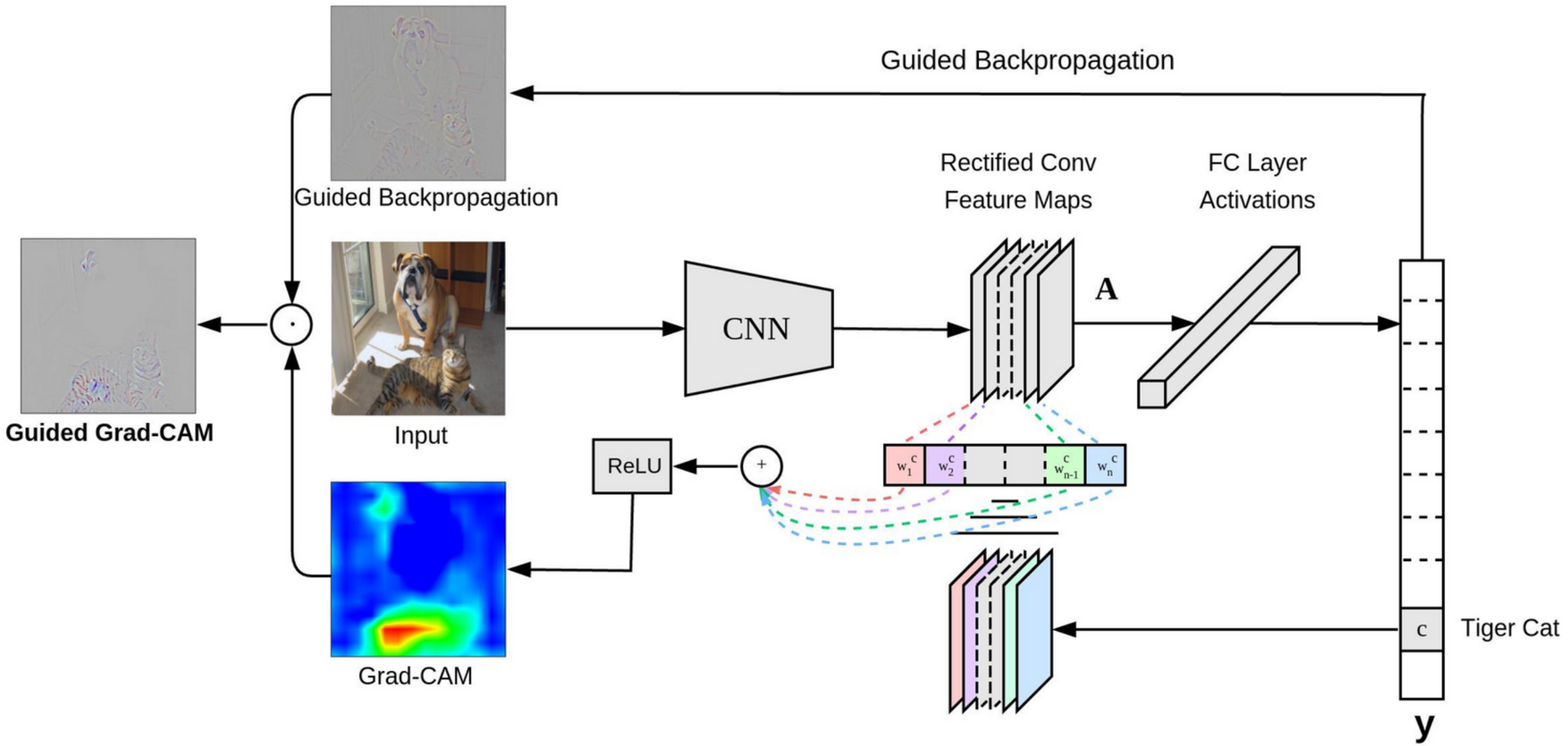
- **Examples**
  Seeing how it works
- **Intuition**
  Steps and method

Grad Cam

Class-specific heat maps



Grad Cam


Guided Backpropagation
Guided Backpropagation
Rectified Conv
Feature Maps
FC Layer
Activations
A
CNN
Guided Grad-CAM
Input
ReLU
Grad-CAM
c
Tiger Cat
y

Grad Cam steps

- Compute Gradient

Step 1: Compute the gradient of y_c with respect to the feature map activations A^k of a convolutional layer, *i.e.* $\frac{\partial y^c}{\partial A^k}$

- Calculate Alphas by Averaging Gradients

Step 2: Global average pool the gradients over the width dimension (indexed by i) and the height dimension (indexed by j) to obtain neuron importance weights α_k^c:

- Calculate Final Grad-CAM Heatmap

Step 3: Perform a weighted combination of the feature map activations A^k where the weights are the α_k^c just calculated:

Model Interetability

1. Interpretability

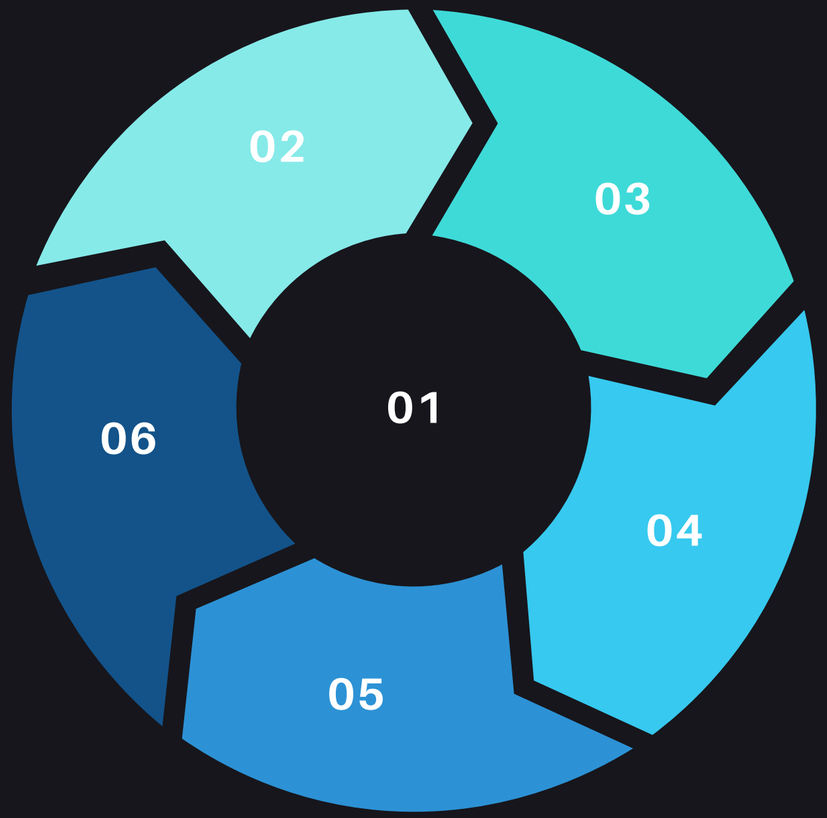
Can a human understand why?

2. Importance of interpreatibly

Medical imaging, loans

3. Scope of interpretability

Understanding the model and understand a decision made by the model



4. Interpretable models

logistic regression

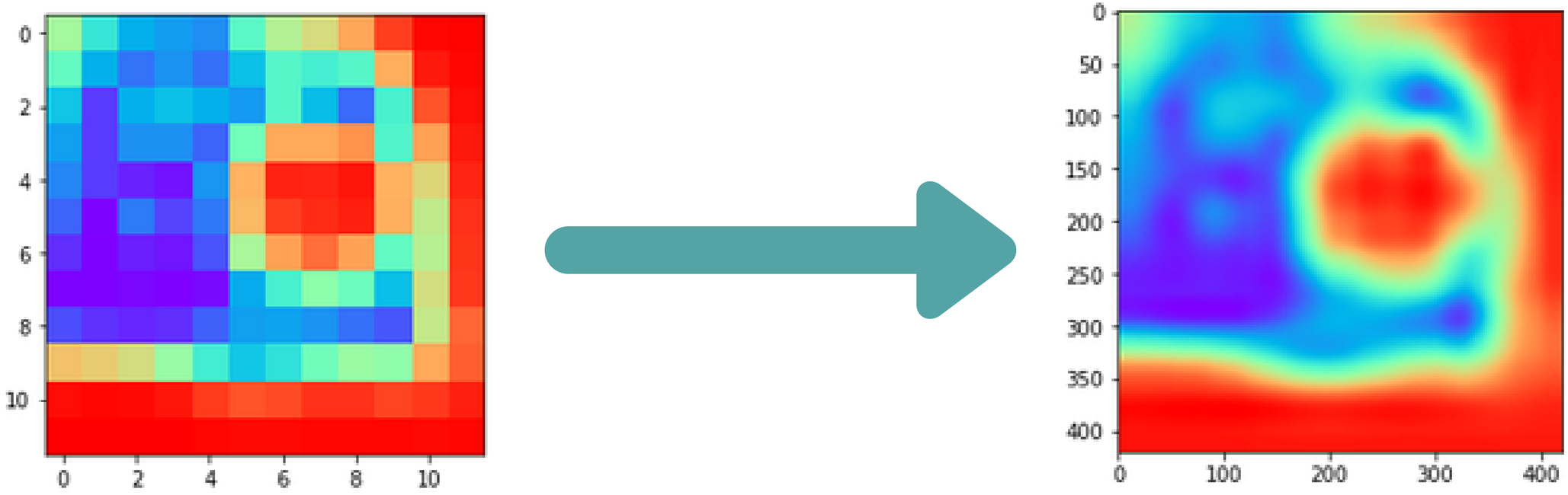
5.Lime

Use a simple model to explain a more complicated one

6. Grad cam

class-level explanations on images

Upsampled heatmaps



Ressources

- Grad-CAM: Visual Explanations from Deep Networks
- Investigate Network Predictions Using Class Activation Mapping
- Grad-CAM Reveals the Why Behind Deep Learning Decisions
- Gradient-weighted Class Activation Mapping - Grad-CAM-
- Interpretable Machine Learning