ScaDS.AI
DRESDEN LEIPZIG
CENTER FOR SCALABLE DATA ANALYTICS
AND ARTIFICIAL INTELLIGENCE

Explainable Machine Learning

Robert Haase

Bundesministerium
für Bildung
und Forschung
SACHSEN
Diese Maßnahme wird gefördert durch die Bundesregierung aufgrund eines Beschlusses des Deutschen Bundestages.
Diese Maßnahme wird mitfinanziert durch Steuermittel auf der Grundlage des von den Abgeordneten des Sächsischen Landtags beschlossenen Haushaltes.

Explainable Artificial Intelligence (XAI)

- “Es gibt derzeit noch keine allgemein akzeptierte Definition von XAI.”

Wikipedia [1]

Relevant Aspects:

- Explainability vs. Interpretability of AI-algorithms
- We seek to enable humans to
  - predict results of AI Systems,
  - trust AI-Systems and
  - using AI-Systems effectively.

[1] https://de.wikipedia.org/wiki/Explainable_Artificial_Intelligence

Explanation of Random Forest Classifiers

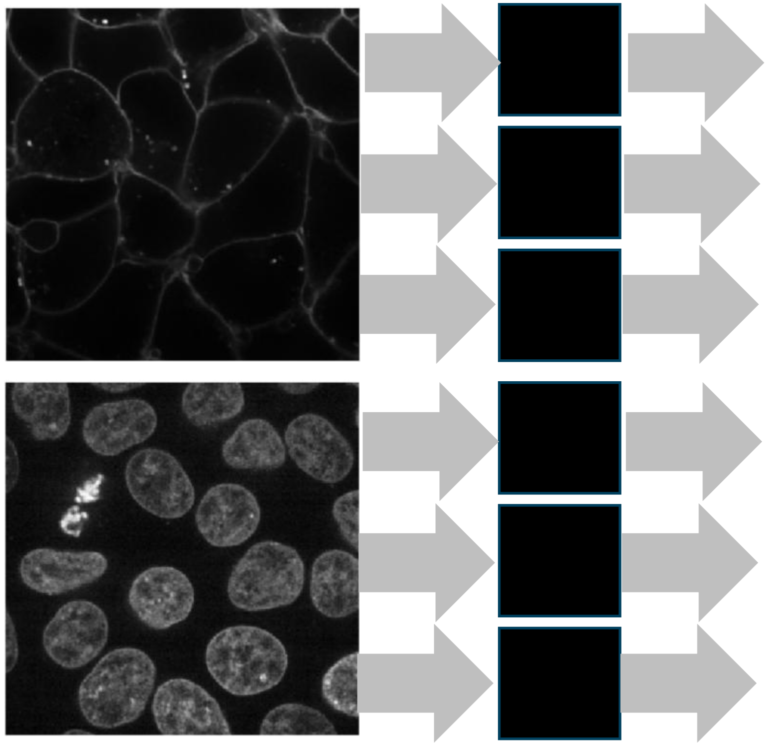
… by reading code

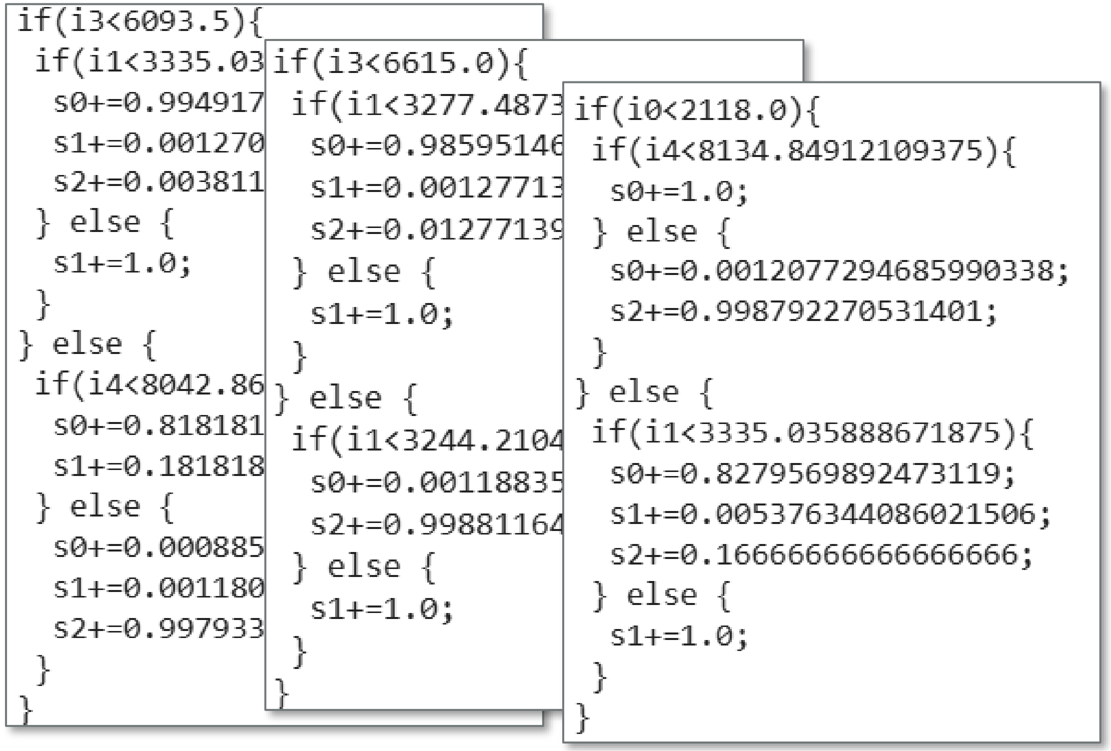
… is quite useless

Input images

Feature images

Decision trees (OpenCL)



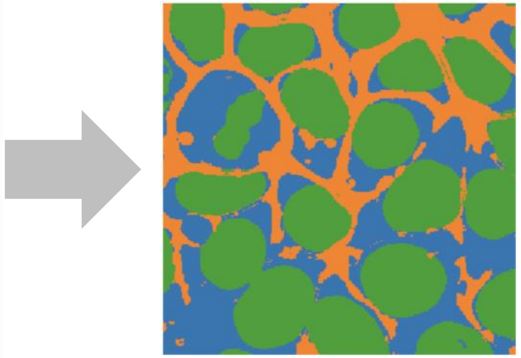


```
if(i3<6093.5){
 if(i1<3335.03
  s0+=0.994917
  s1+=0.001270
  s2+=0.003811
 } else {
  s1+=1.0;
 }
} else {
 if(i4<8042.86
  s0+=0.818181
  s1+=0.181818
 } else {
  s0+=0.000885
  s1+=0.001180
  s2+=0.997933
 }
}
```

```
if(i3<6615.0){
 if(i1<3277.4873
  s0+=0.98595146
  s1+=0.00127713
  s2+=0.01277139
 } else {
  s1+=1.0;
 }
} else {
 if(i1<3244.2104
  s0+=0.00118835
  s2+=0.99881164
 } else {
  s1+=1.0;
 }
}
```

```
if(i0<2118.0){
 if(i4<8134.84912109375){
  s0+=1.0;
 } else {
  s0+=0.0012077294685990338;
  s2+=0.998792270531401;
 }
} else {
 if(i1<3335.035888671875){
  s0+=0.8279569892473119;
  s1+=0.005376344086021506;
  s2+=0.16666666666666666;
 } else {
  s1+=1.0;
 }
}
```

Classification result images



Robert Haase
@haesleinhuepf
AI4Medicine
Sept 24th 2025

https://github.com/haesleinhuepf/apoc/blob/main/demo/mutlichannel_images.ipynb

Explainability

A logically consistent line of argumentation that depicts a situation or an algorithm with complete transparency.

Intrinsically explainable AI-algorithms

- Example: Linear Regression

$$f(x_1, x_2) = w_1 x_1 + w_2 x_2$$

If w_1 is much bigger than w_2, the result depends much more on x_1 compared to x_2.

Model explainable

Results predictable

Explainability

A logically consistent line of argumentation that depicts a situation or an algorithm with complete transparency.

Intrinsically explainable AI-algorithms

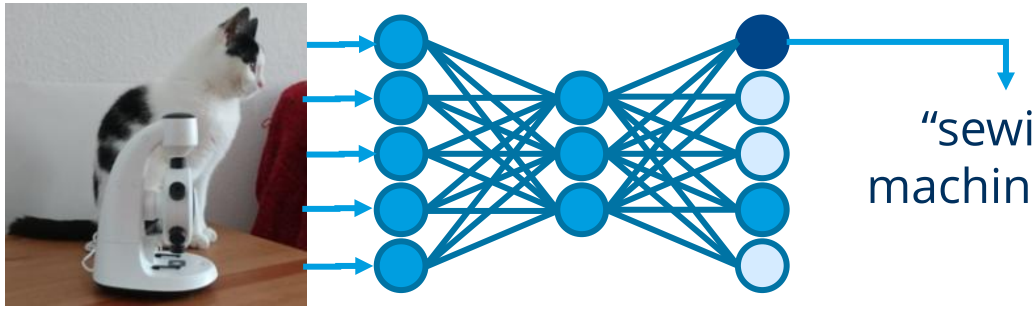
- Example: Linear Regression

$$f(x_1, x_2) = w_1 x_1 + w_2 x_2$$

If w_1 is much bigger than w_2, the result depends much more on x_1 compared to x_2.

Black-Box AI-algorithms

- Example: Deep Neural Networks (DNN)



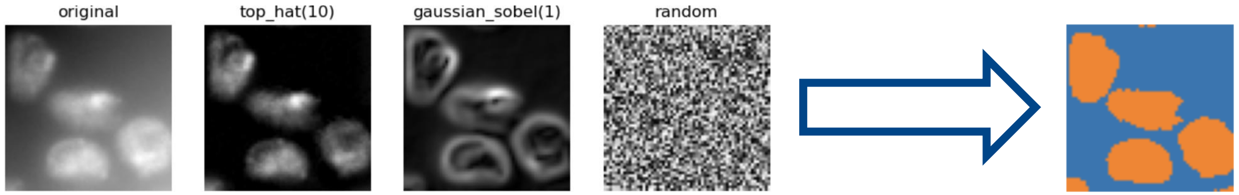
"sewing machine" [1]

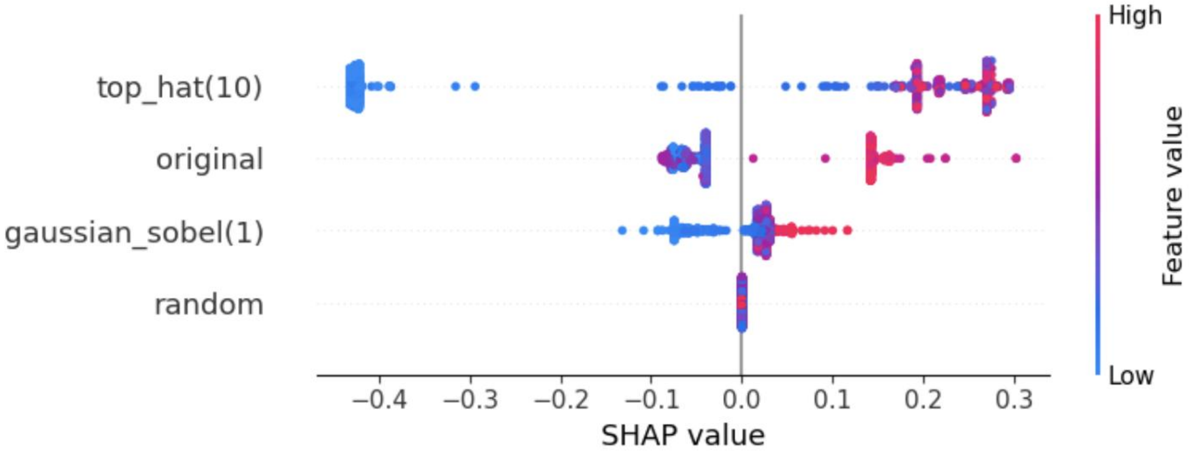
Not easily explainable and predictable

Interpretability

Visualization of intermediate results and their influence on results

Model-agnostic methods
Example: Shapley's Additive exPlanations (SHAP)



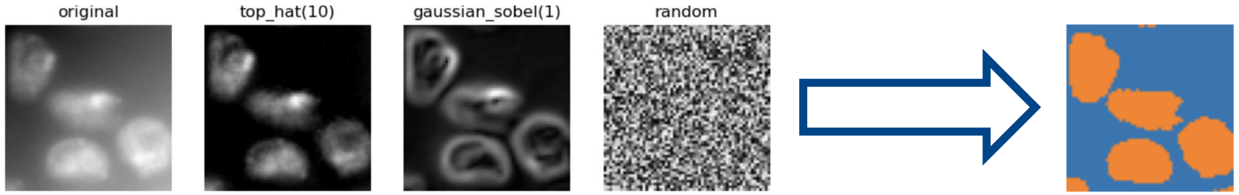


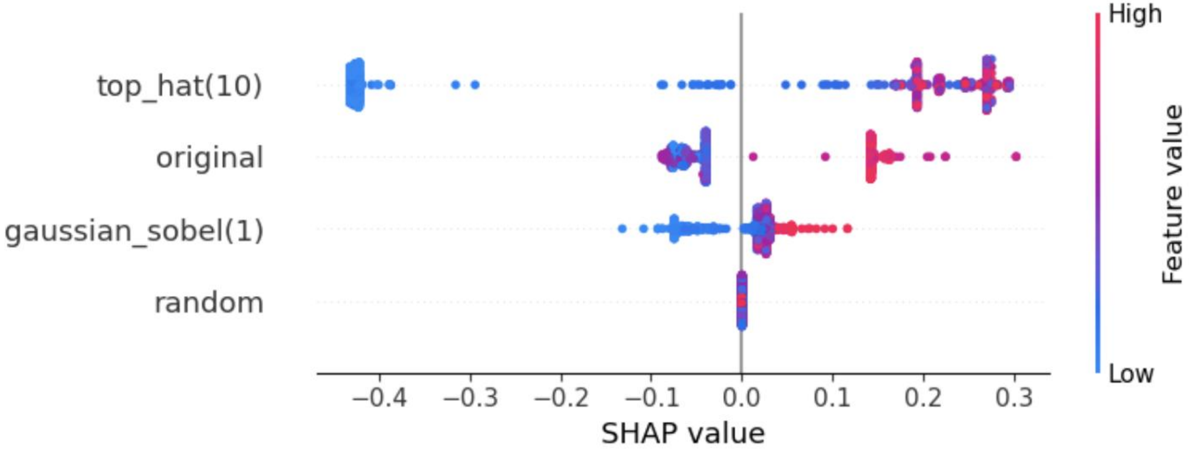
Interpretability

Visualization of intermediate results and their influence on results

Model-agnostic methods

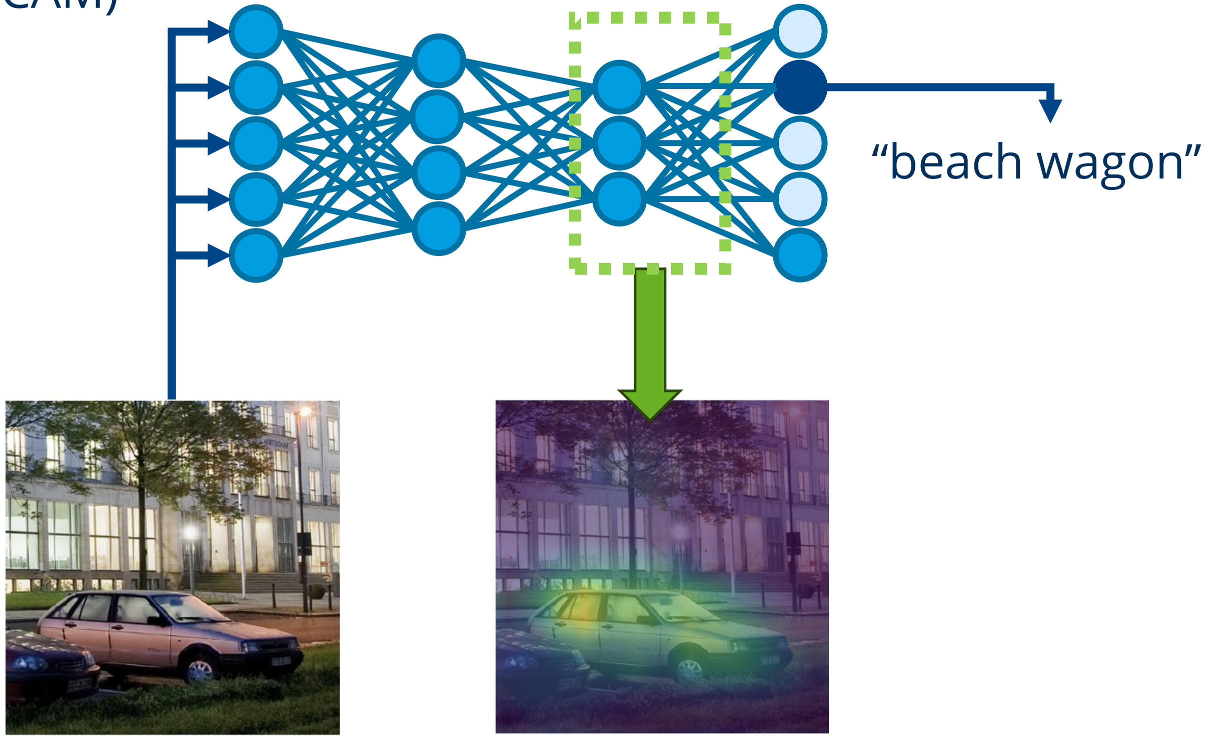
Example: Shapley's Additive exPlanations (SHAP)





Model-specific methods

Example: Gradient Class Activation Maps (Grad-CAM)



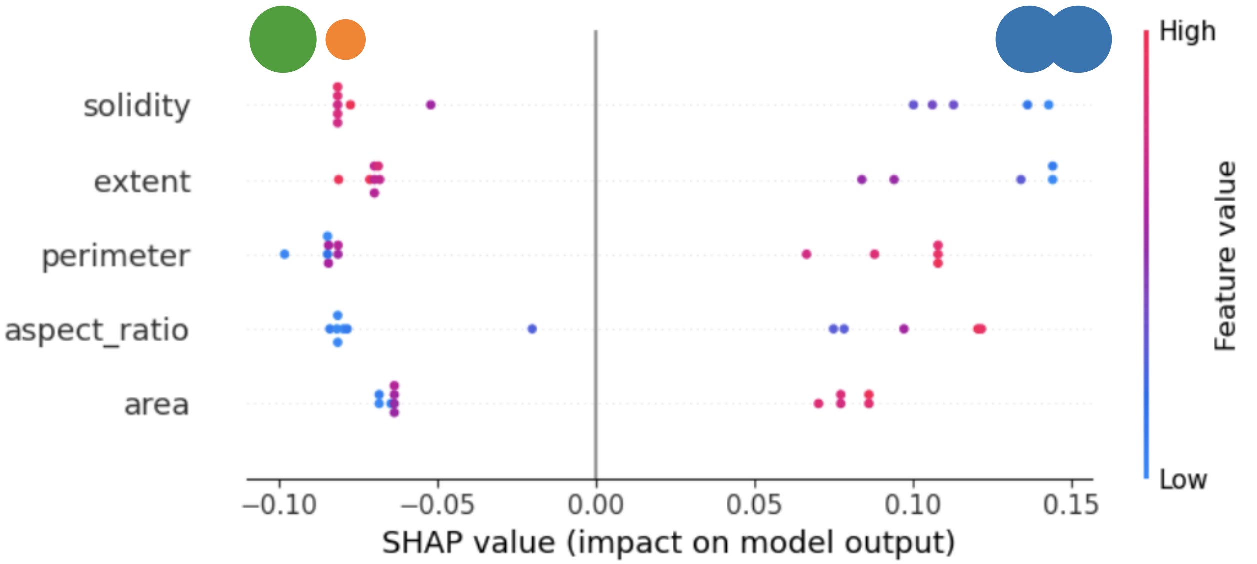
Robert Haase
@haesleinhuepf
AI4Medicine
Sept 24th 2025

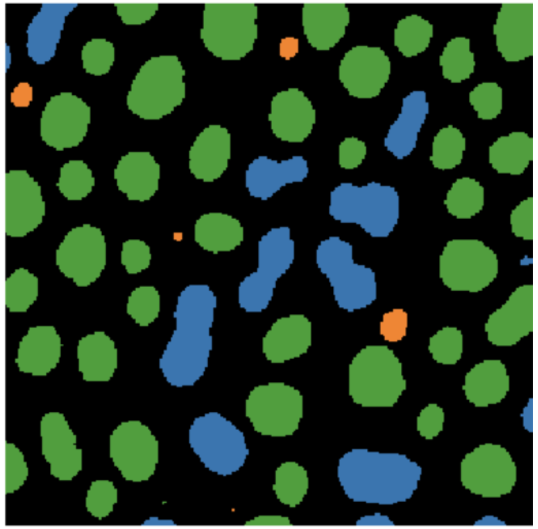
https://haesleinhuepf.github.io/xai/30_shap/pixel_classifier.html
https://haesleinhuepf.github.io/xai/60_grad-cam/classification_resnet.html
Image source: Cropped from HTW Dresden (Fotograf: Peter Sebb) licensed CC BY-SA 3.0
https://commons.wikimedia.org/w/index.php?curid=15652763

Explainable AI

Depending on the target group [for the explanation], the influence of data is more important than how AI algorithms work.

- Many computer scientists want to explain and understand AI methods.
- Biologists use AI as a method to explain biological processes.
- Example: "What parameters distinguish round objects from elongated ones?"





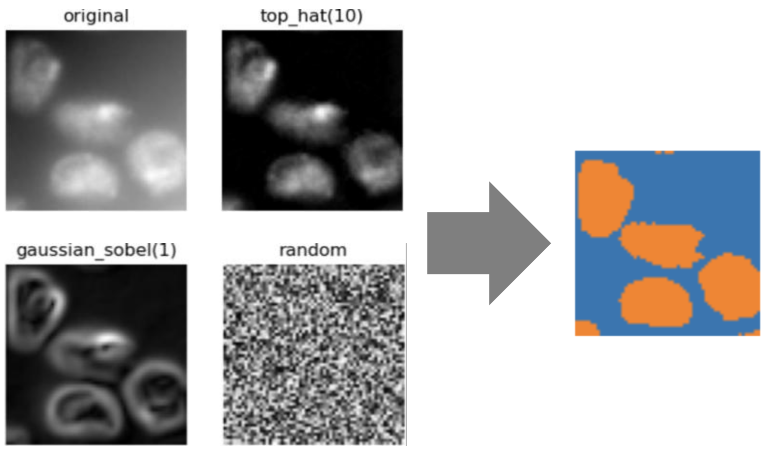
Recap: Feature selection

- Which measurement / parameter / feature is related to the effect I'm investigating?
- Example goals:

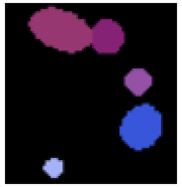
- Amplitude
- Energy
- Duration
- …

- Noise
- Tourists jumping on a sensor
- Earthquake approaching

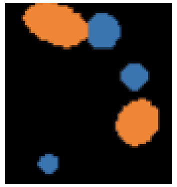
Signal classification



Pixel classification



- Area
- Perimeter
- Aspect ratio
- …



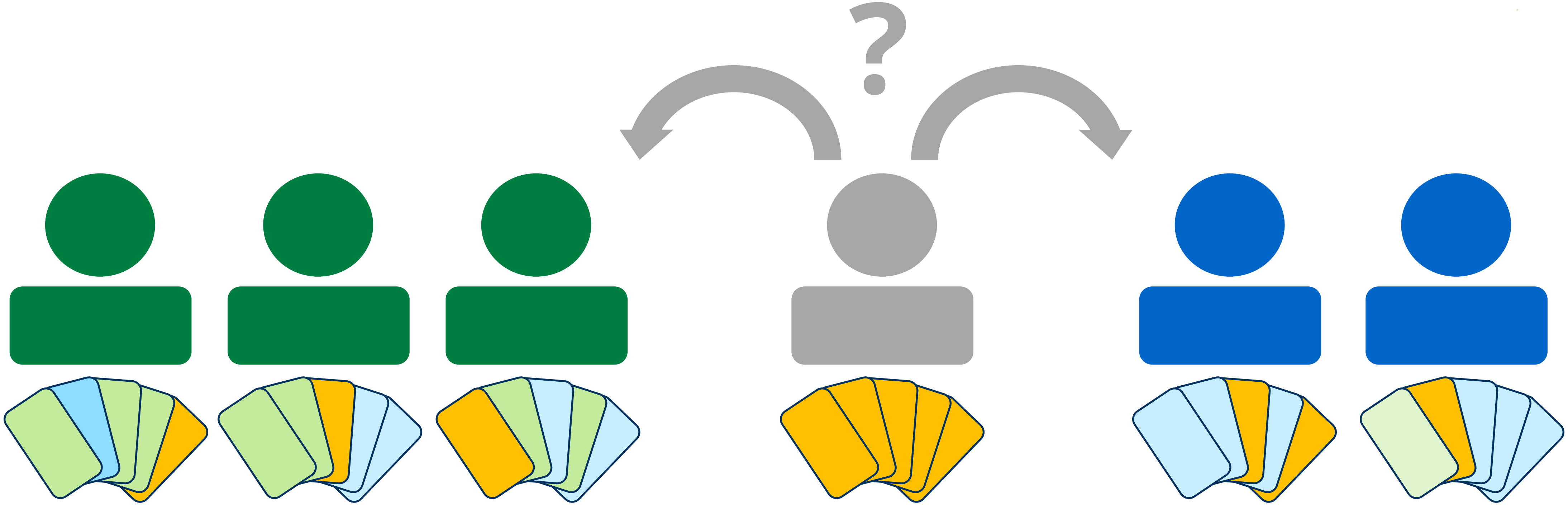
- Round
- Elongated

Object classification

Collaborative game theory

If players collaborate, how is the impact on a team if another player joins?

Example game goal: maximize cards of the same colour.

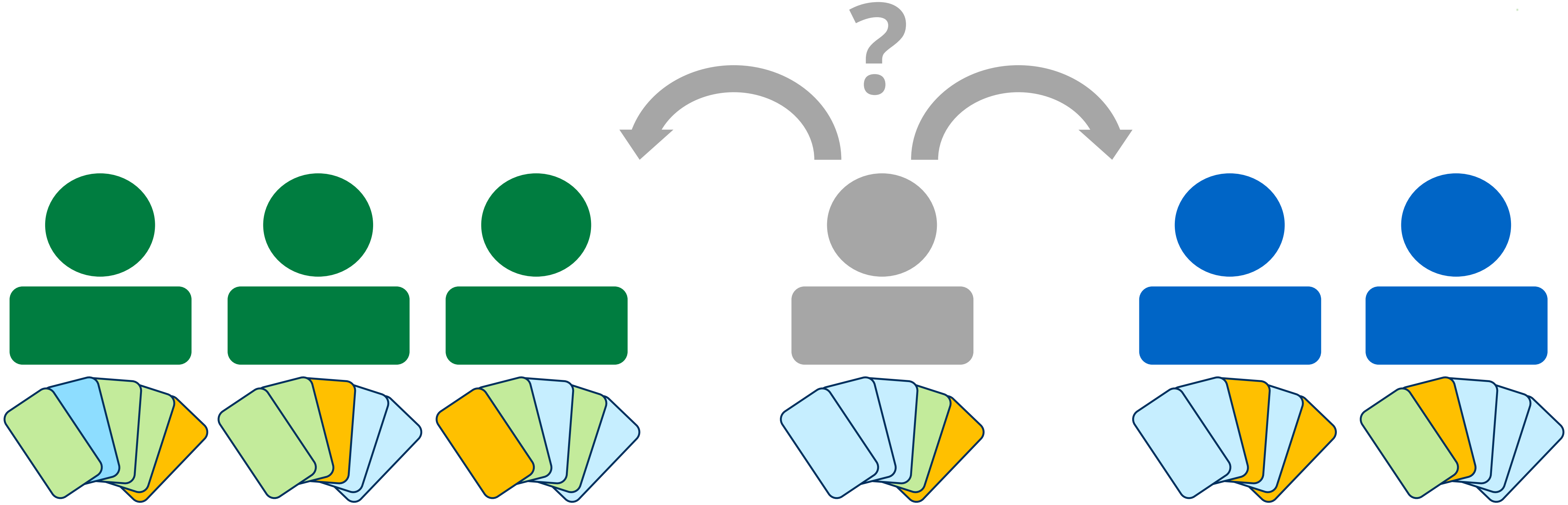


Value for both teams: 0

Collaborative game theory

If players collaborate, how is the impact on a team if another player joins?

Example game goal: maximize cards of the same colour.



Value for team green: 0.1

Value for team blue: 0.3

SHAP

Analogously, this can be done with data points instead of features.

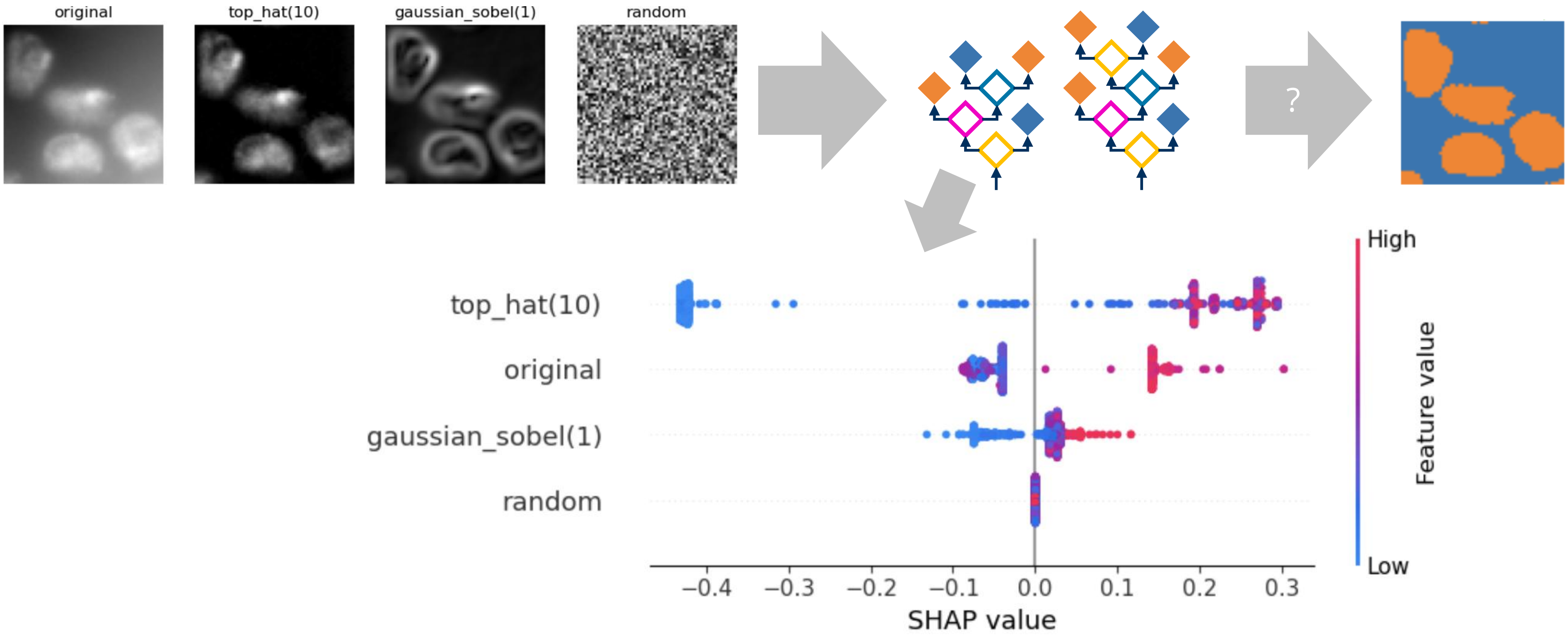
SHapley’s Additive exPlanations

$$\phi_i = \sum_{S\subseteq F\setminus\{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}\left[f_x(S\cup\{i\}) - f_x(S)\right]$$

	SHAP value of feature i	Sum over all Subsets of Features not including i	Weight related to number of used features in relation all players	Quality of classifier using feature i	Quality of classifier *not* using feature i
Game theory	SHAP value of player i	Sum over all Subsets of Players not including i	Weight related to number of players in a coalition in relation to undecided players and all players	Chance to win game of coalition without player i	Chance to win game of coalition *including* player i

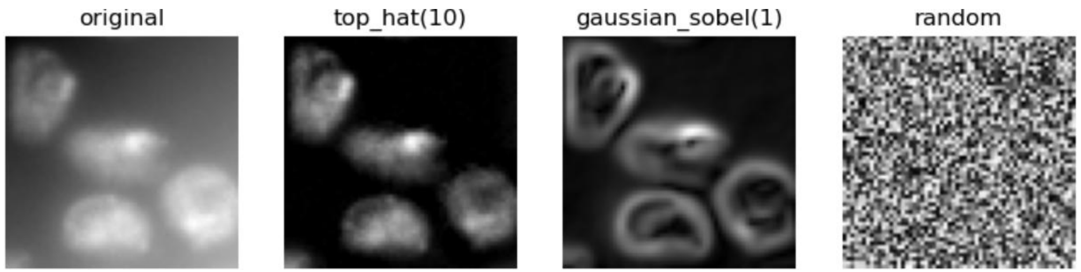
SHAP

Allows interpreting [pixel] classification results

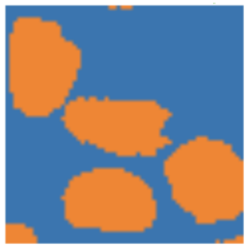


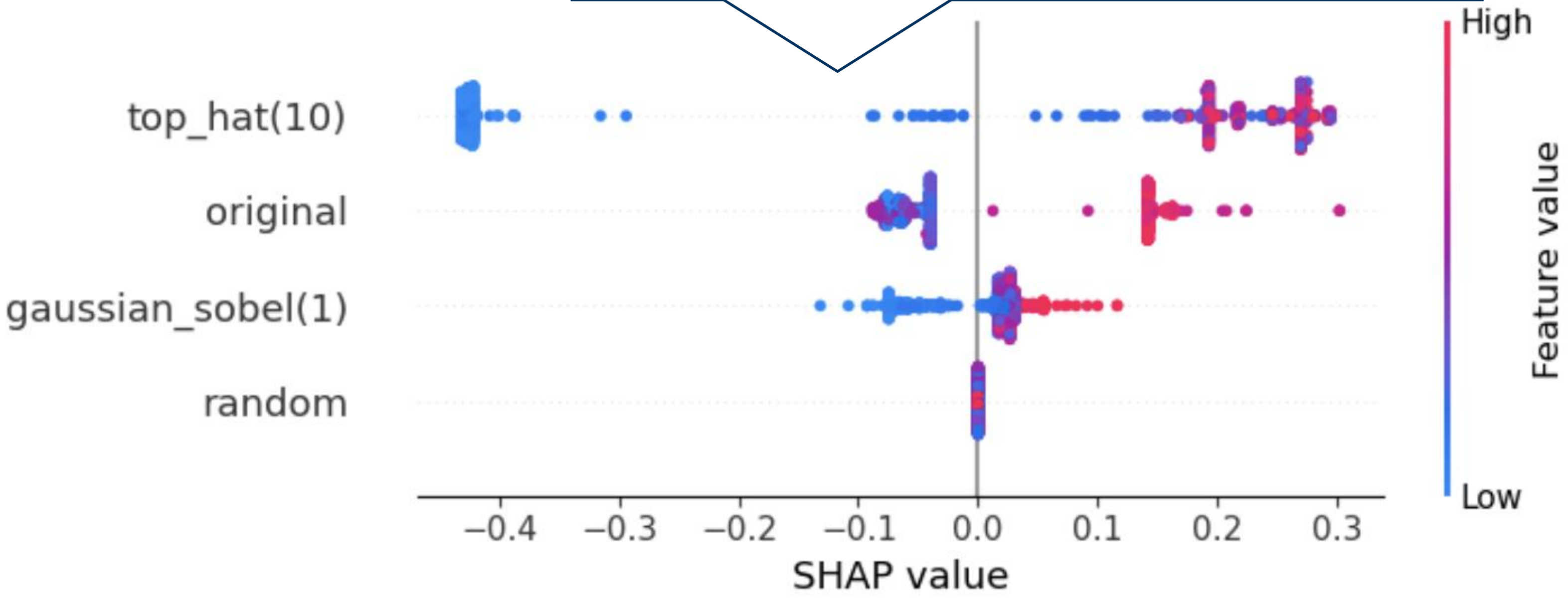
SHAP

Allows interpreting [pixel] classificatio



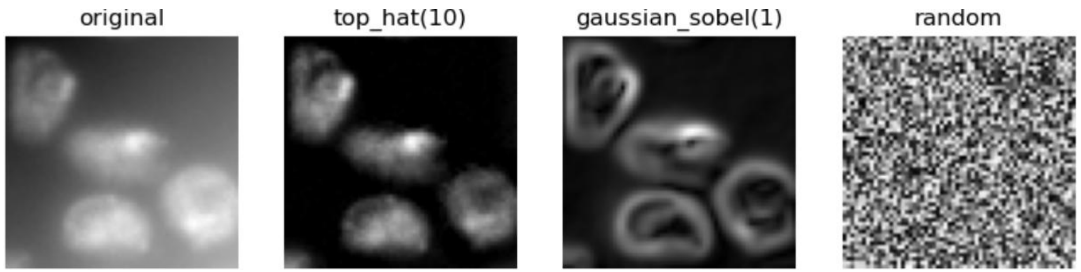
“If intensity in the top-hat image is high, the classifier tends to select the positive class (orange).”



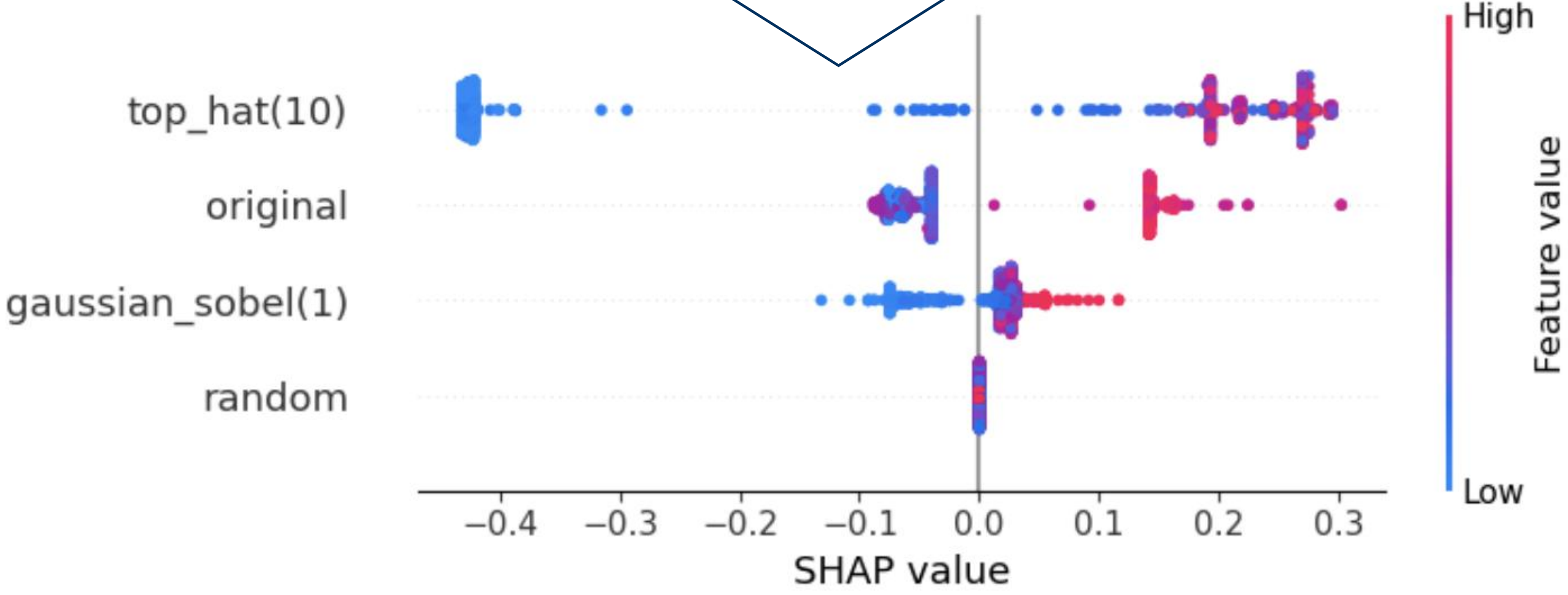


SHAP

Allows interpreting [pixel] classificatio

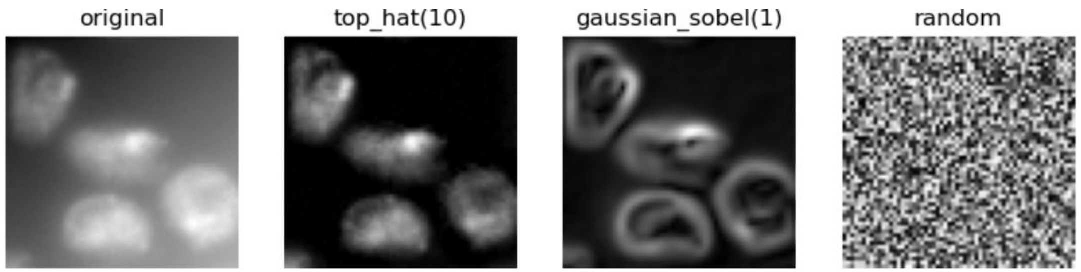


“If intensity in the top-hat image is low, the classifier needs to take other features into account.”

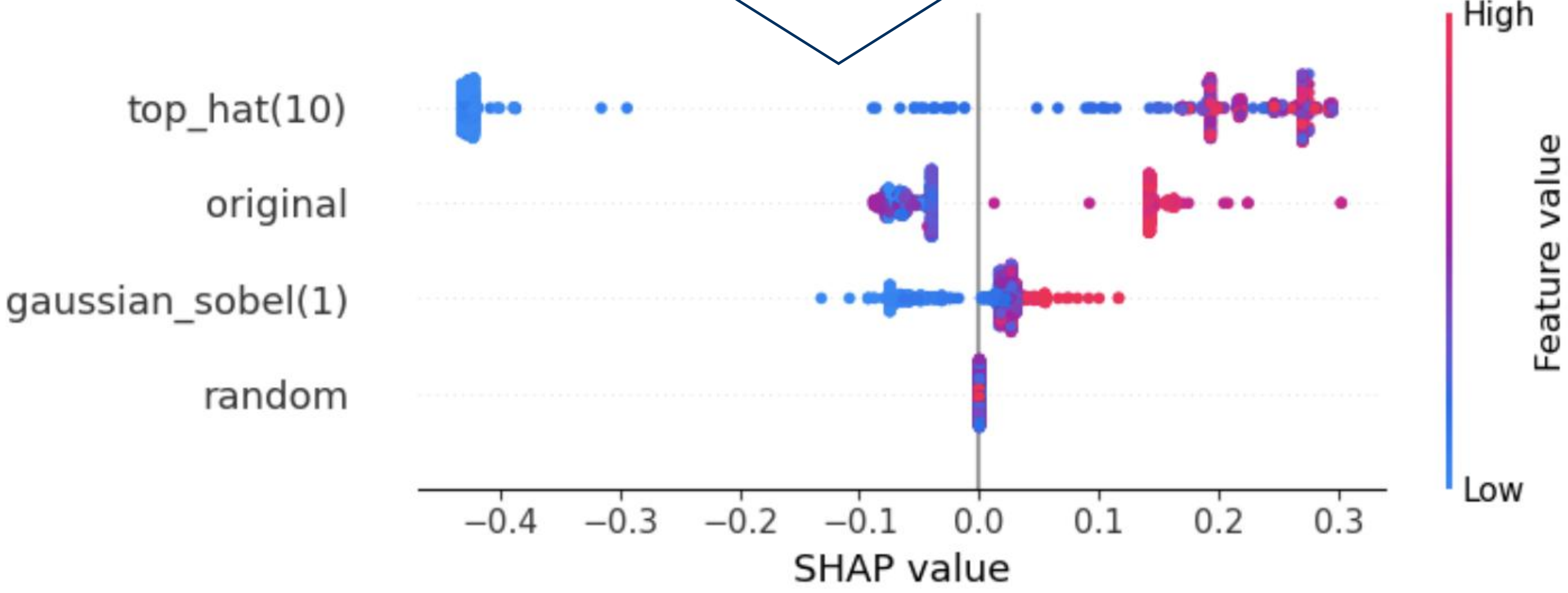


SHAP

Allows interpreting [pixel] classificatio

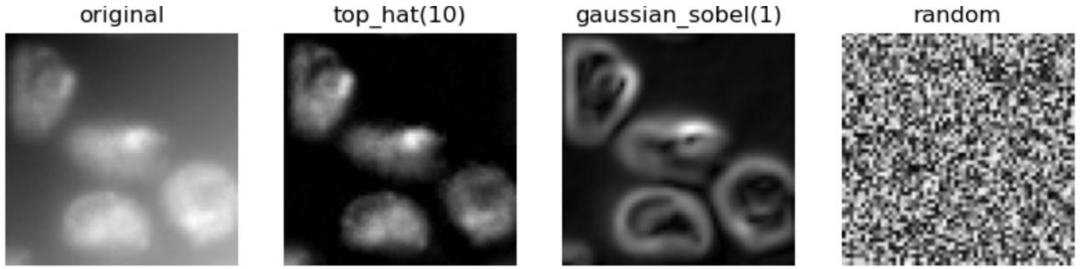


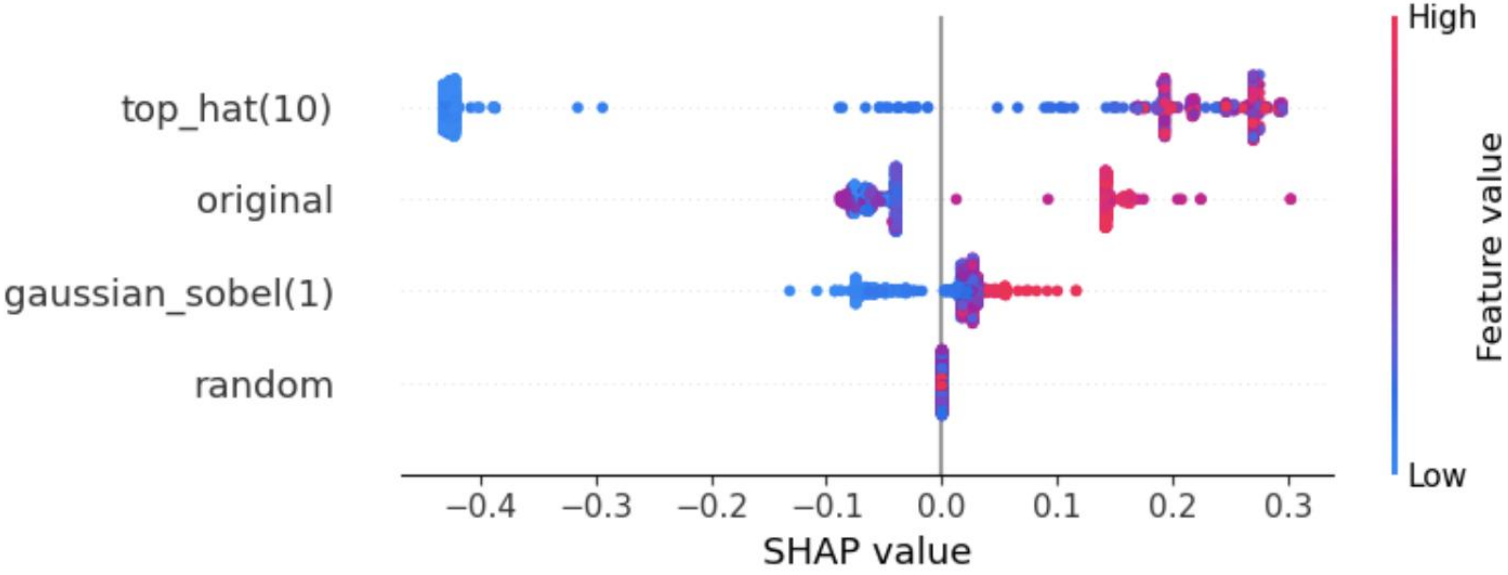
“The random feature has no value for classification.”



Pitfall: Correlation

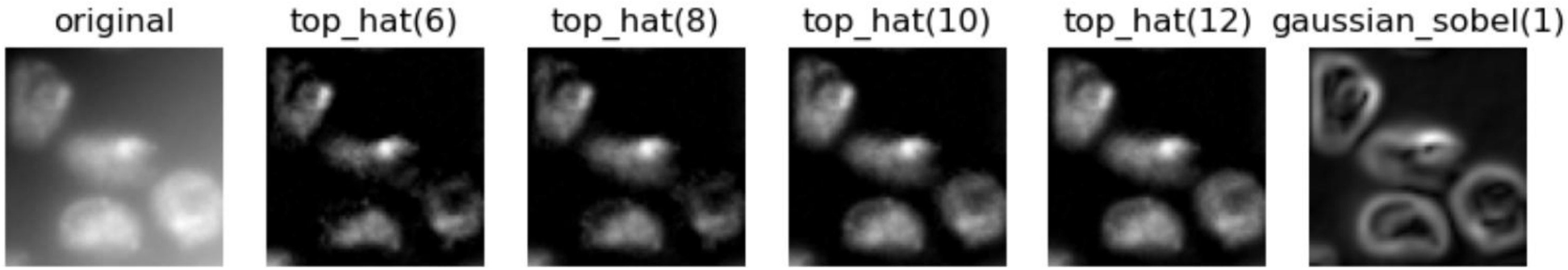
Correlated features may harm interpretability

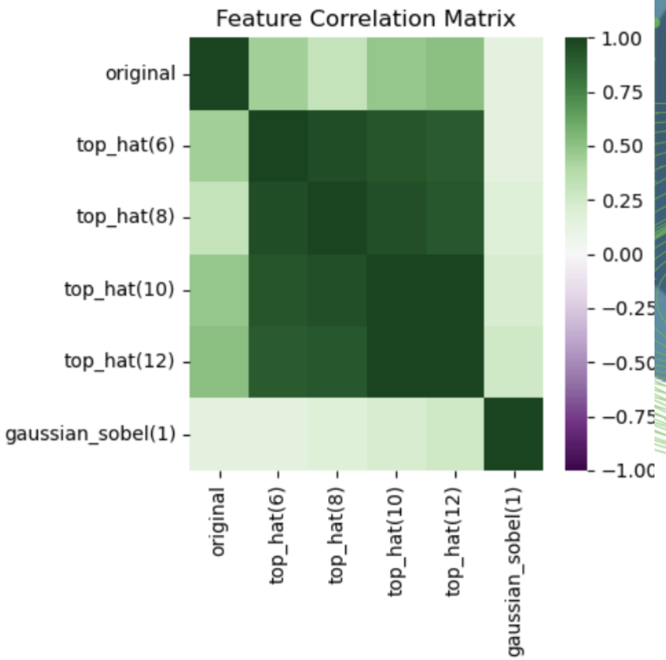


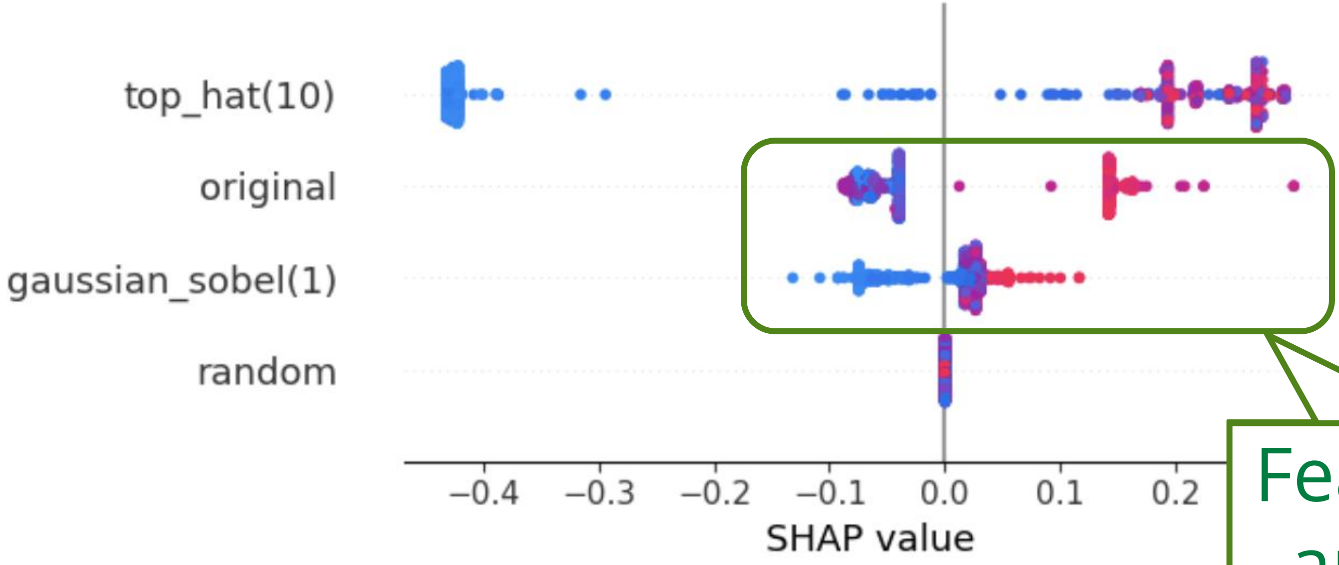


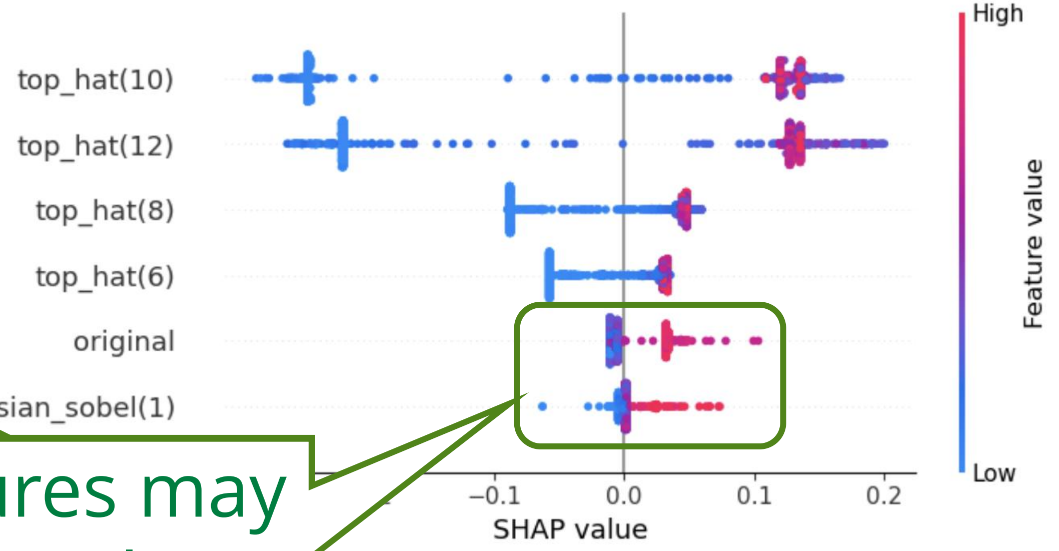
Pitfall: Correlation

Correlated features may harm interpretability









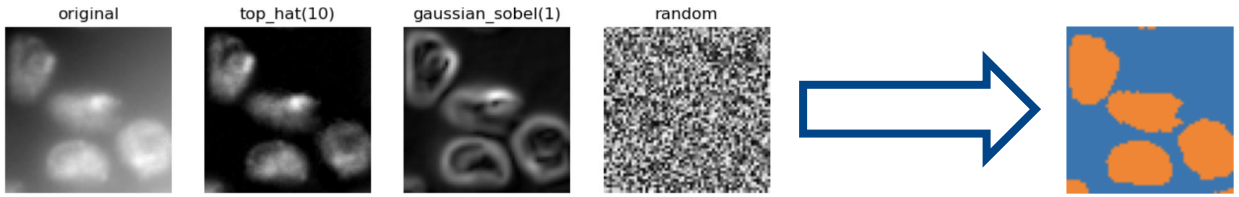
Features may appear less valuable.

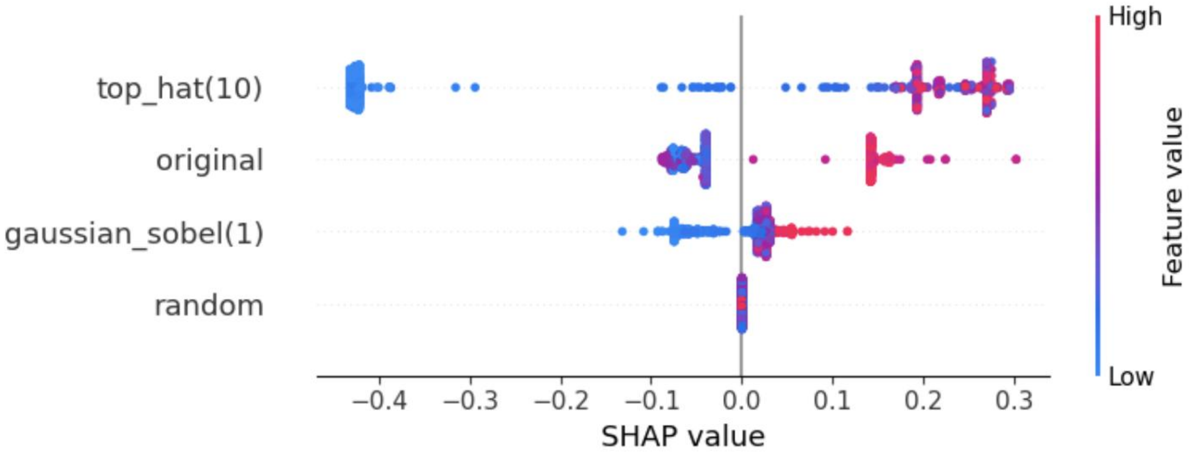
Interpretability

Visualization of intermediate results and their influence on results

Model-agnostic methods

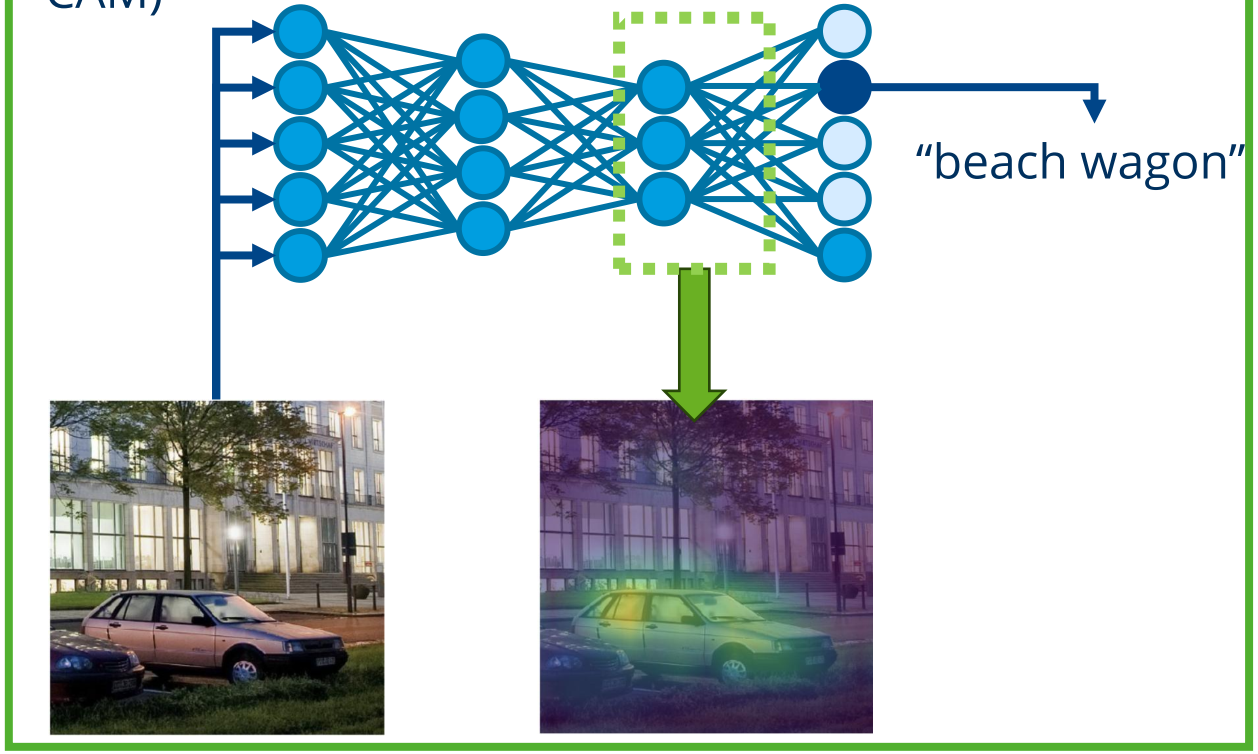
Example: Shapley's Additive exPlanations (SHAP)





Model-specific methods

Example: Gradient Class Activation Maps (Grad-CAM)

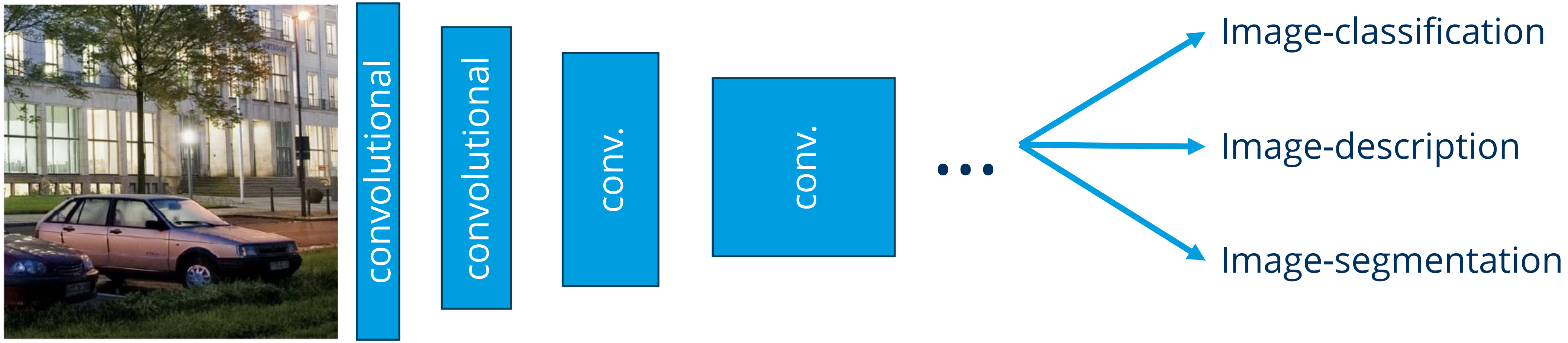


Robert Haase
@haesleinhuepf
AI4Medicine
Sept 24th 2025

https://haesleinhuepf.github.io/xai/30_shap/pixel_classifier.html
https://haesleinhuepf.github.io/xai/60_grad-cam/classification_resnet.html
Image source: Cropped from HTW Dresden (Fotograf: Peter Sebb) licensed CC BY-SA 3.0
https://commons.wikimedia.org/w/index.php?curid=15652763

Gradient Class-Activation Maps (Grad-CAM)

- Works only with NN algorithms that first process input data with convolutional layers. (model-specific)
- Independent of right half of the NN (model-agnostic)
- Visualizes intermediate results to make decision-making in the AI system interpretable



convolutional

convolutional

conv.

conv.

…

Image-classification

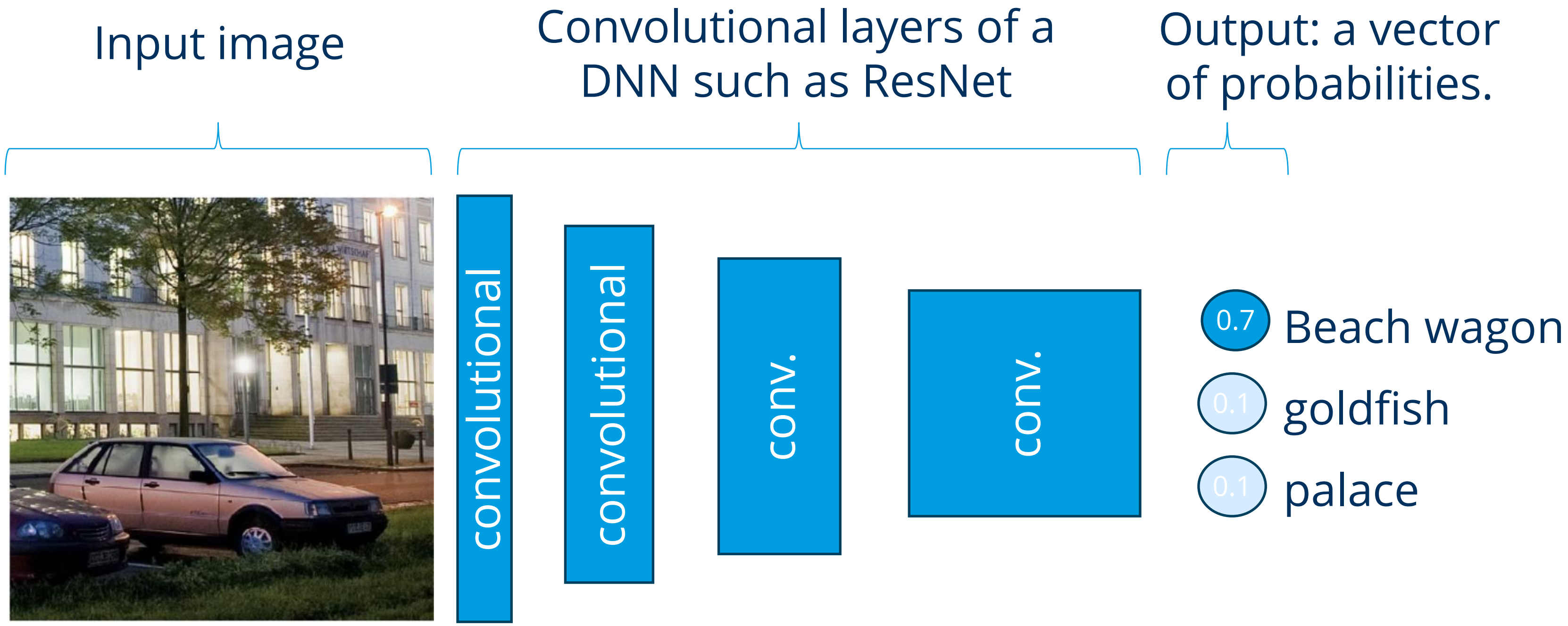
Image-description

Image-segmentation

Image source: Cropped from HTW Dresden (Photographer: Peter Sebb) licensed CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=15652763

Gradient Class-Activation Maps (Grad-CAM)

Is applied to existing network ; no modification of the architecture necessary (post-hoc method).



Input image

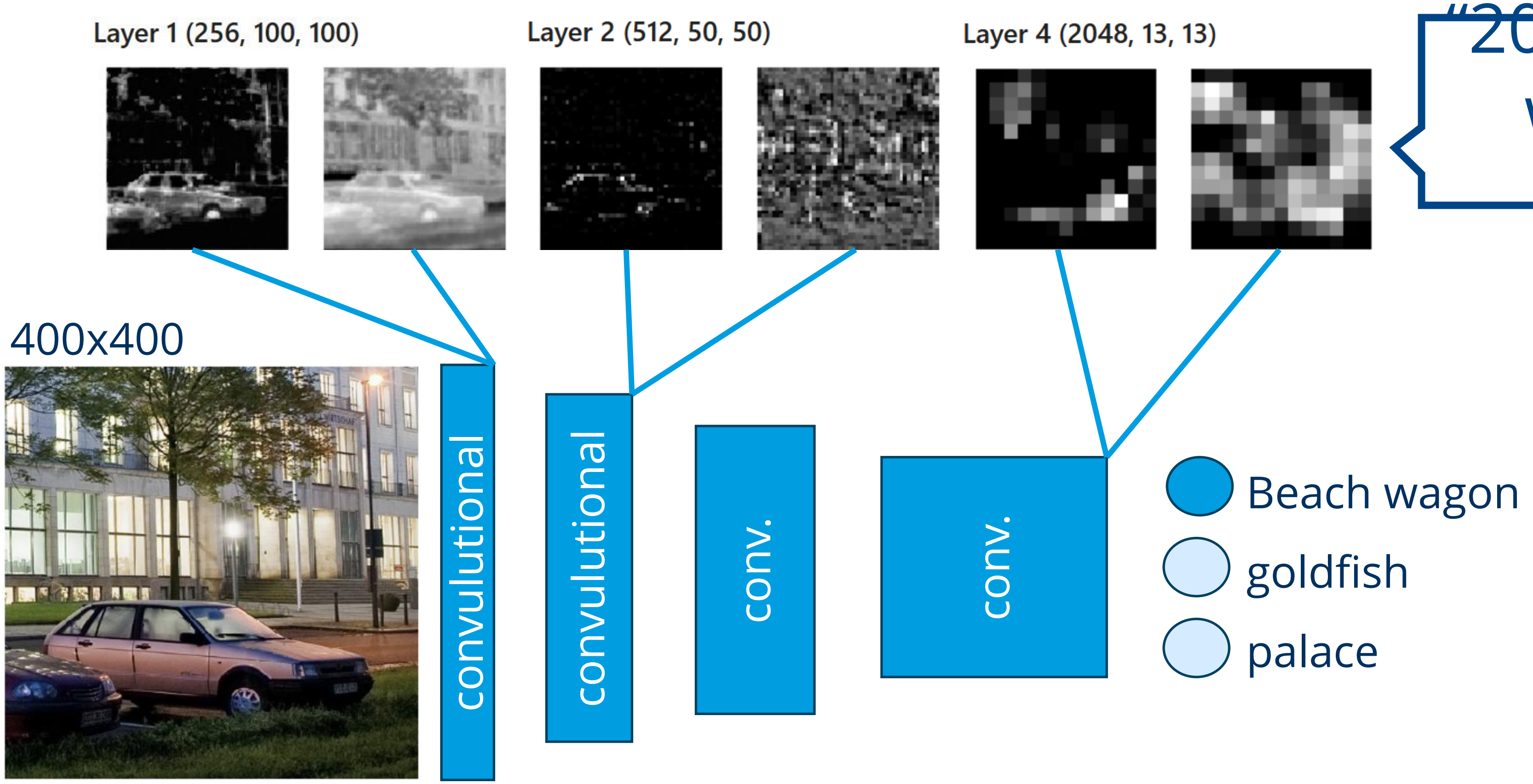
Convolutional layers of a DNN such as ResNet

Output: a vector of probabilities.

Image source: Cropped from HTW Dresden (Fotograf: Peter Sebb) licensed CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=15652763

Gradient Class-Activation Maps (Grad-CAM)

Applied to existing network; no adaptation of the architecture necessary (post-hoc method).



Layer 1 (256, 100, 100)

Layer 2 (512, 50, 50)

Layer 4 (2048, 13, 13)

400x400

convulutional

convulutional

conv.

conv.

● Beach wagon

○ goldfish

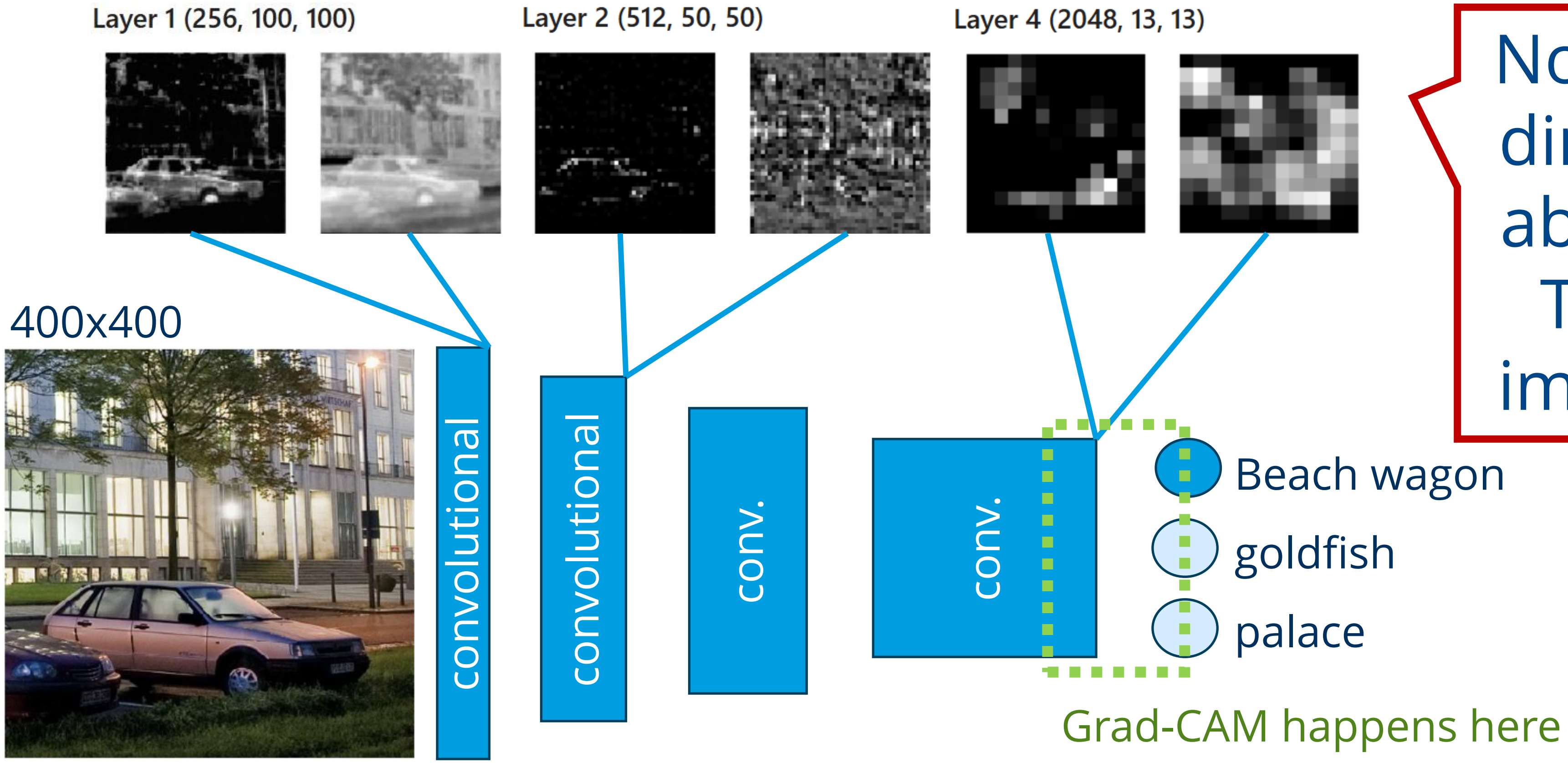
○ palace

"2028 feature images with each 13x13 pixels"

Image source: Cropped from HTW Dresden (Fotograf: Peter Sebb) licensed CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=15652763

Gradient Class-Activation Maps (Grad-CAM)

Applied to existing network; no adaptation of the architecture necessary (post-hoc method).

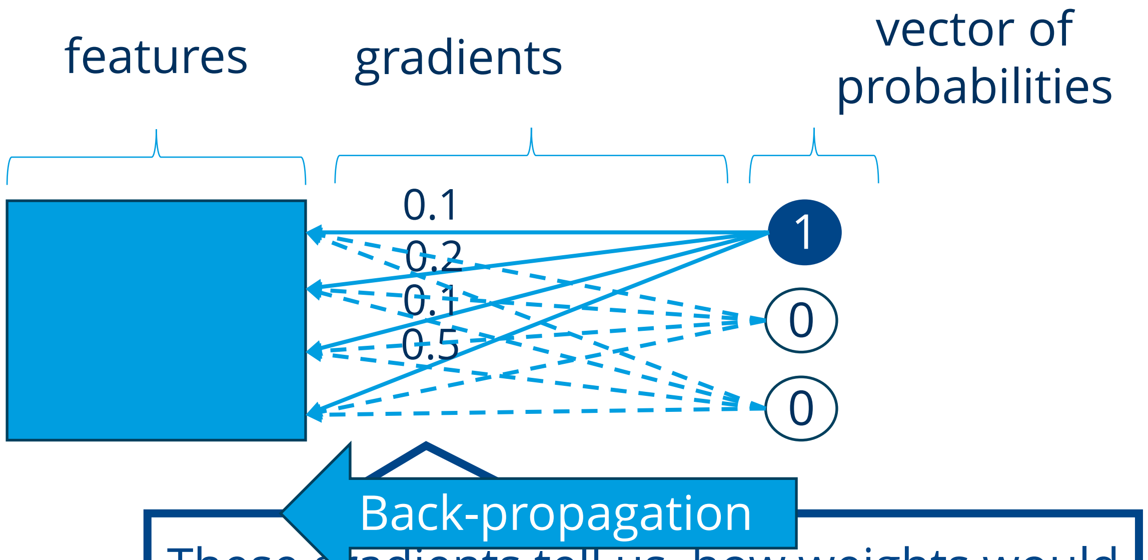


None of these images directly says anything about image content. There is no feature image "Beach wagon"

Image source: Cropped from HTW Dresden (Fotograf: Peter Sebb) licensed CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=15652763

Gradient Class-Activation Maps (Grad-CAM)

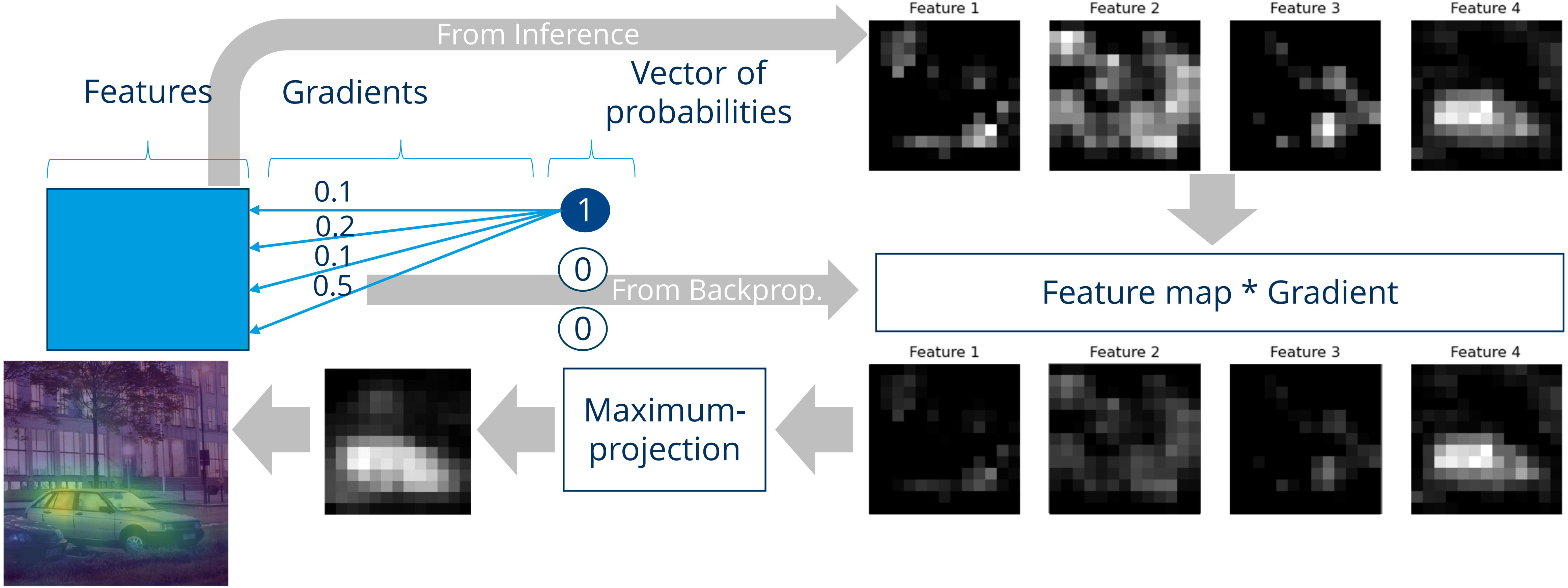
Back-propagation of a perfect classification (1,0,0) gives us gradients (weight changes) to improve the classification.



These gradients tell us, how weights would need *to be adjusted* to produce the desired classification

Gradient Class-Activation Maps (Grad-CAM)

Back-propagation of a perfect classification (1,0,0) gives us gradients (weight changes) to improve the classification.



From Inference

Features

Gradients

Vector of probabilities

0.1

0.2

0.1

0.5

1

0

0

From Backprop.

Feature 1

Feature 2

Feature 3

Feature 4

Feature map * Gradient

Feature 1

Feature 2

Feature 3

Feature 4

Maximum-projection

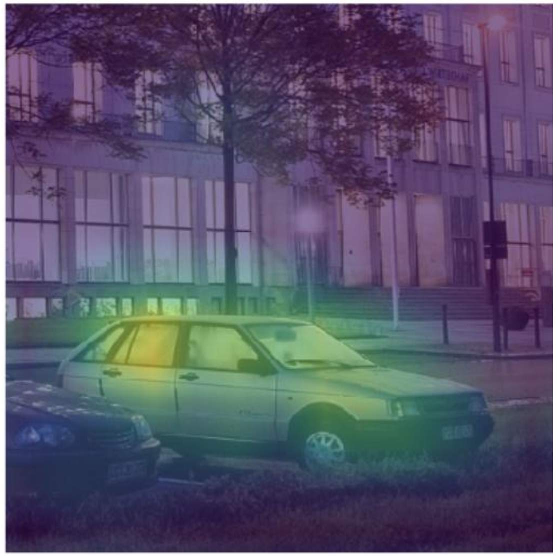
Image source: Cropped from HTW Dresden (Photographer: Peter Sebb) licensed CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=15652763

Gradient Class-Activation Maps (Grad-CAM)

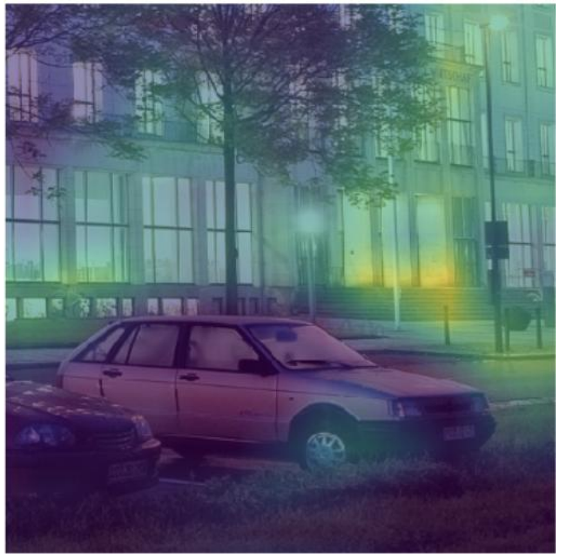
Back-propagation of a perfect classification (1,0,0) gives us gradients (weight changes) to improve the classification.

This also works with other possible classifications, e.g. (0,1,0).

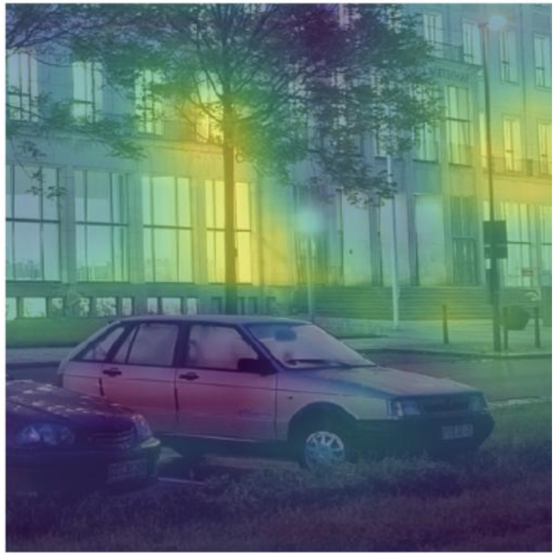
“beach waggon”



“palace”



“flagpole”



“great white shark”

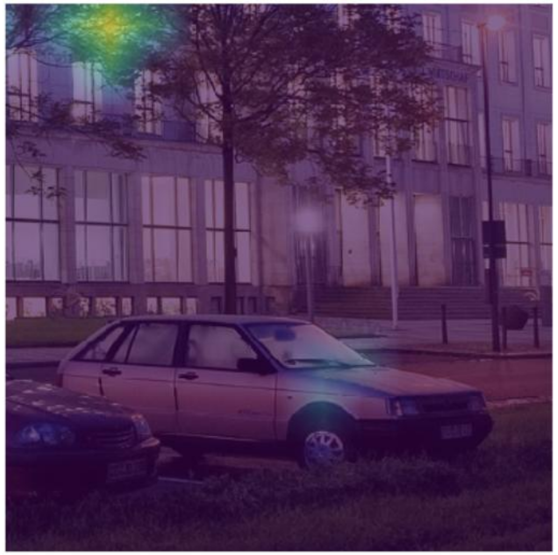
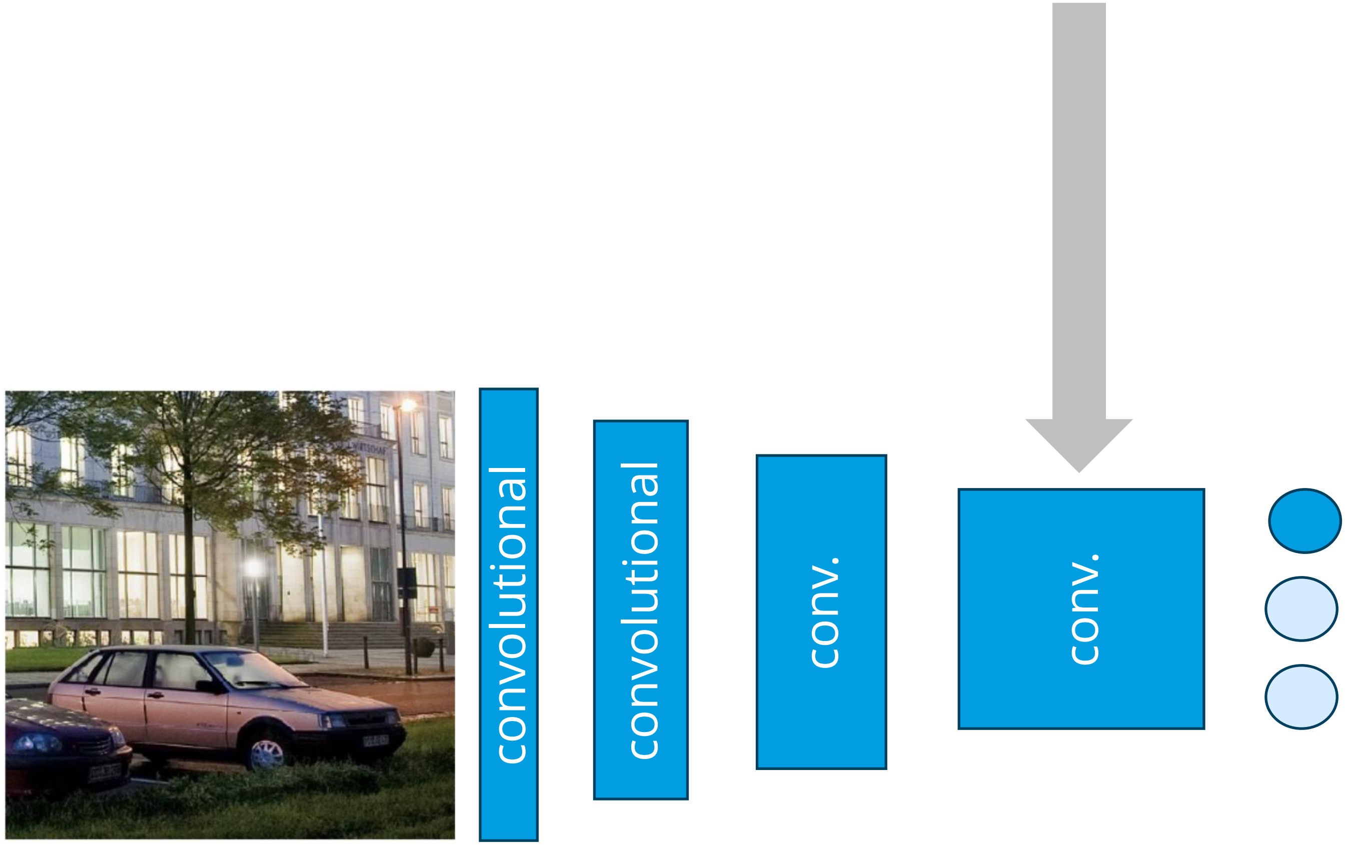


Image source: Cropped from HTW Dresden (Fotograf: Peter Sebb) licensed CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=15652763

Quiz

Assuming, this layer has 2048x13x13 outputs. What does the 2048 stand for?



Number of features

Width of the feature maps

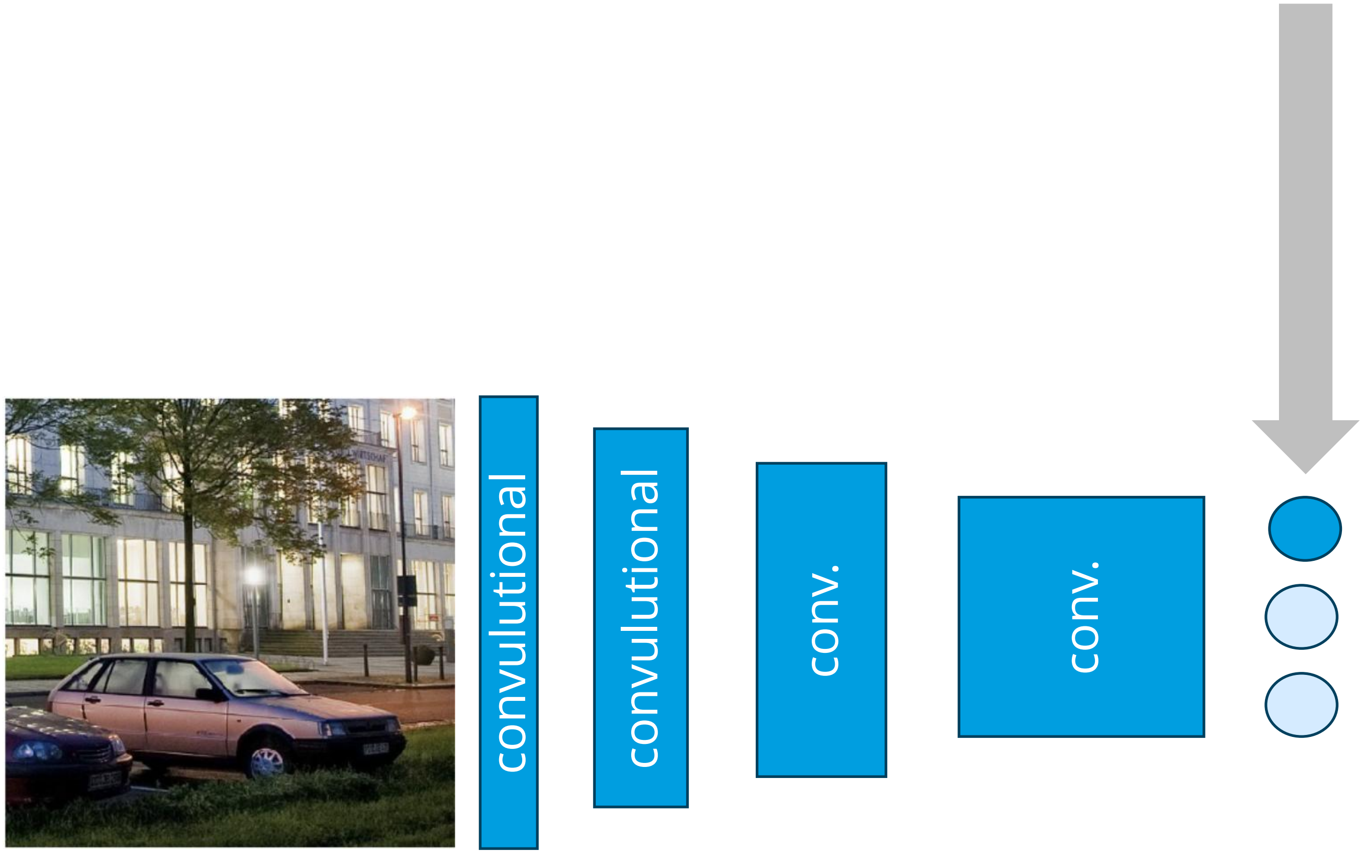
Number of classes

Number of layers

Image source: Cropped from HTW Dresden (Photographer: Peter Sebb) licensed CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=15652763

Quiz

Assume this vector has 1000 elements. What does the 1000 stand for?



convulutional

convulutional

conv.

conv.

Number of features

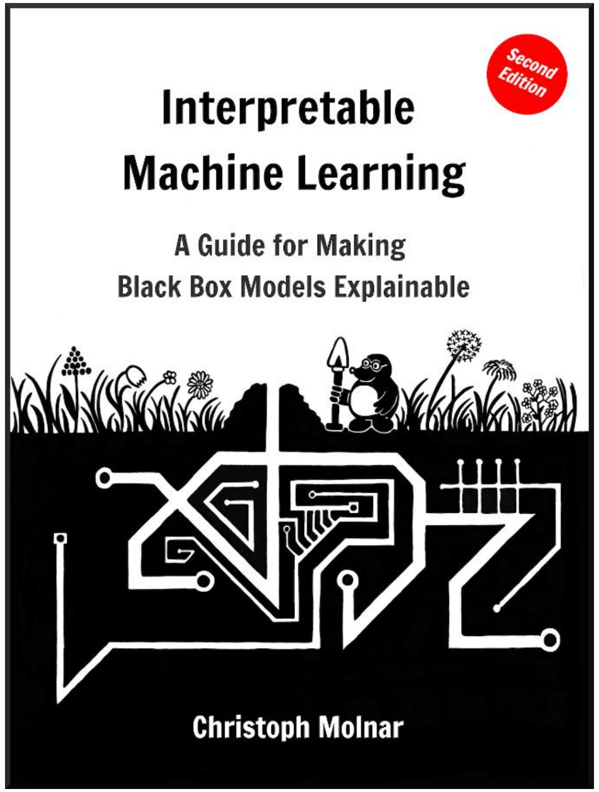
Width of the feature maps

Number of classes

Number of layers

Image source: Cropped from HTW Dresden (Fotograf: Peter Sebb) licensed CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=15652763

Read more...



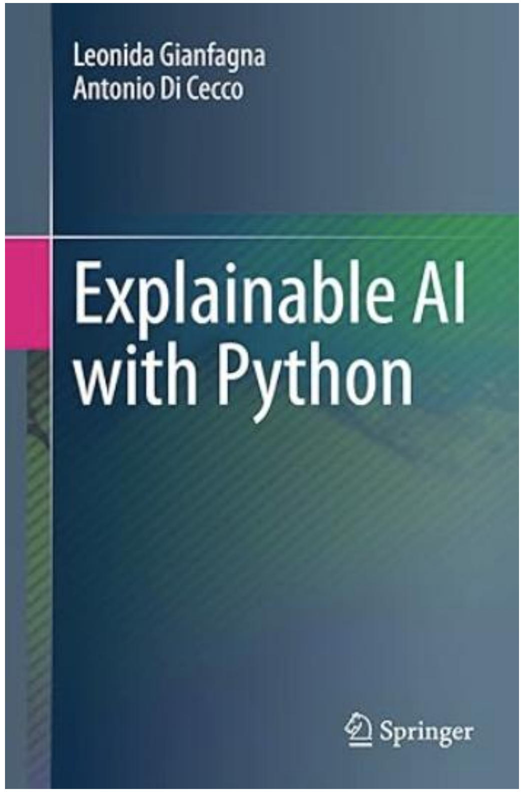
Second Edition

Interpretable Machine Learning

A Guide for Making Black Box Models Explainable

Christoph Molnar

https://christophm.github.io/interpretable-ml-book/

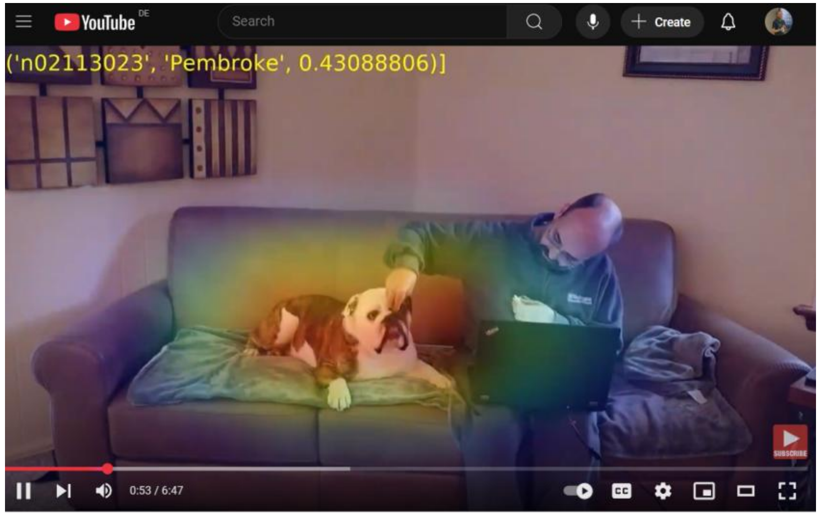


Leonida Gianfagna
Antonio Di Cecco

Explainable AI with Python

Springer

https://www.amazon.de/dp/3030686396

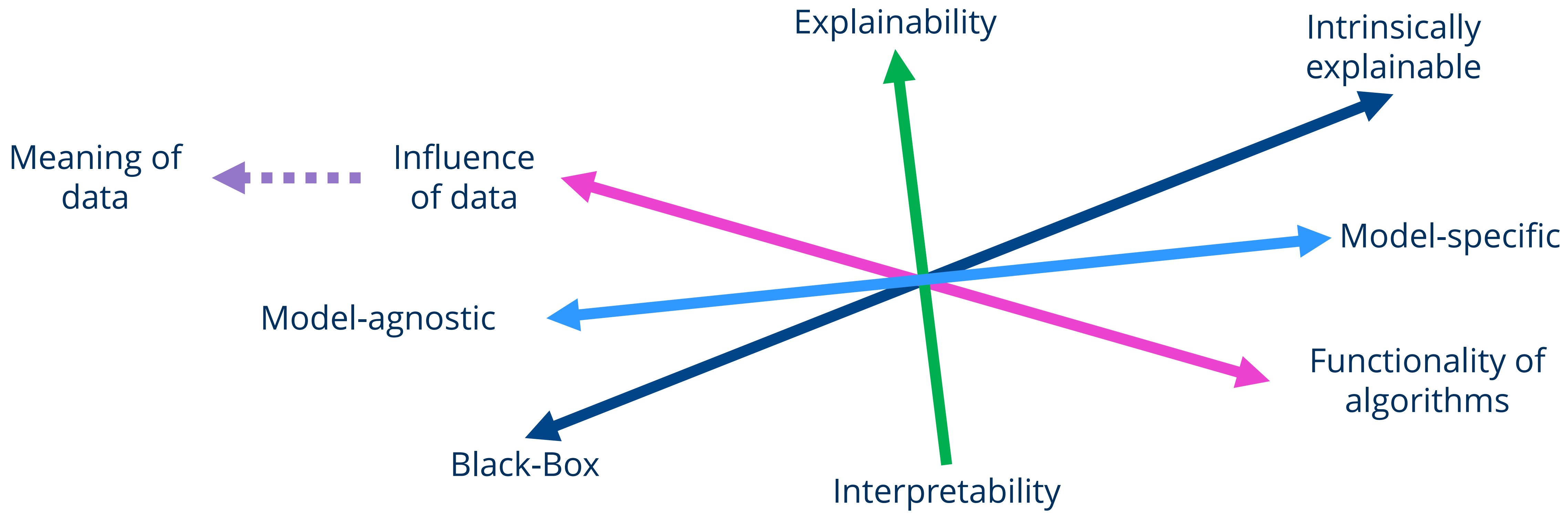


('n02113023', 'Pembroke', 0.43088806)]

https://www.youtube.com/watch?v=dw63QH_b3Jo

Summary: Explainable AI

Methods of XAI can be classified on different scales



ScaDS.AI DRESDEN LEIPZIG

CENTER FOR SCALABLE DATA ANALYTICS AND
ARTIFICIAL INTELLIGENCE

Exercises

Robert Haase

Funded by

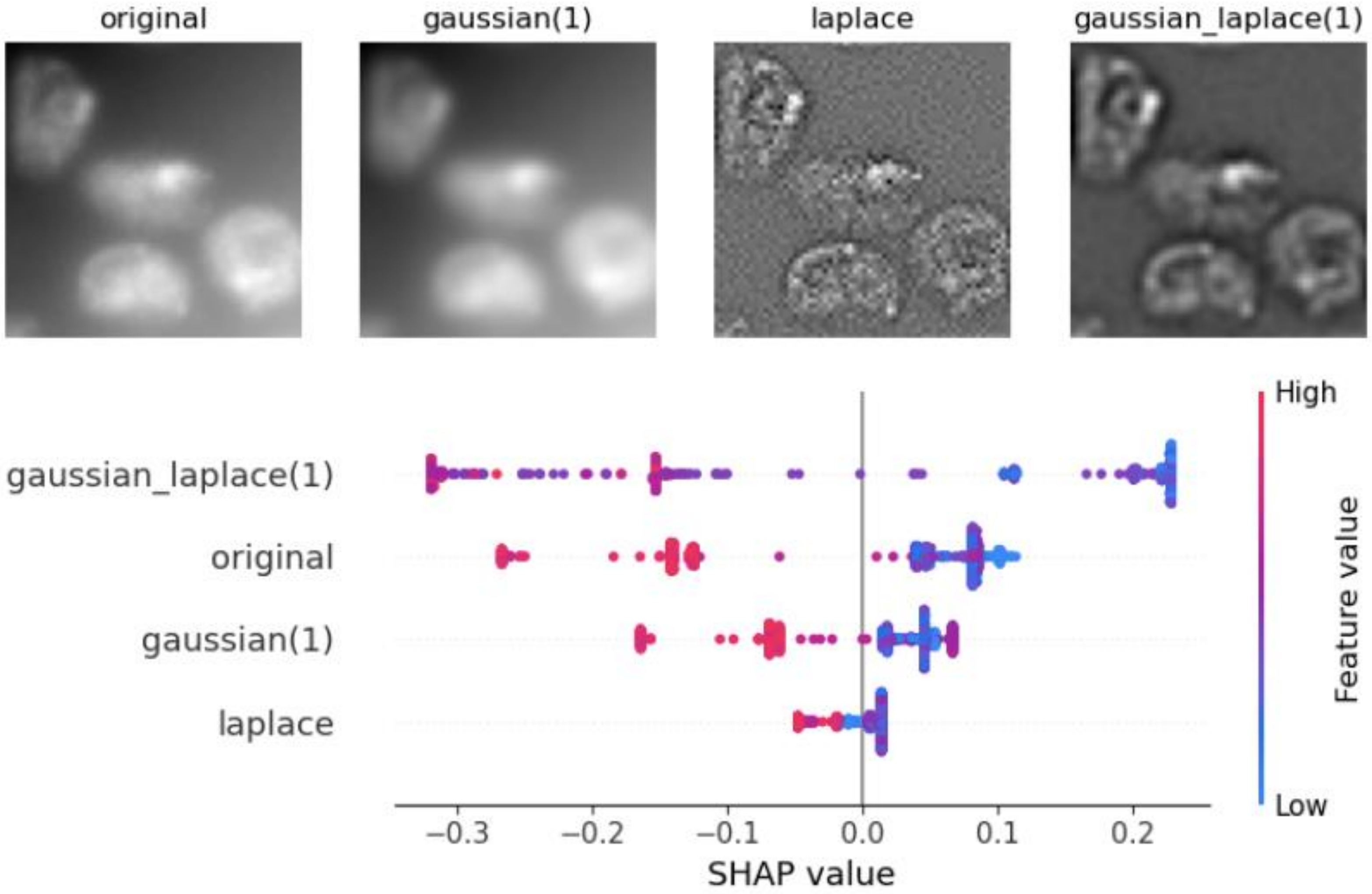
Bundesministerium
für Bildung
und Forschung

SACHSEN

Diese Maßnahme wird gefördert durch die Bundesregierung aufgrund eines Beschlusses des Deutschen Bundestages. Diese Maßnahme wird mitfinanziert durch Steuermittel auf der Grundlage des von den Abgeordneten des Sächsischen Landtags beschlossenen Haushaltes.

SHAP Analysis in Python

Use the opportunity and explain SHAP plots like this one!


original
gaussian(1)
laplace
gaussian_laplace(1)
gaussian_laplace(1)
original
gaussian(1)
laplace
SHAP value
High
Low
Feature value