Explainable Machine Learning

Robert Haase

Robert Haase
@haesleinhuepf
AI4Medicine
Sept 24th 2025

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.

TECHNISCHE UNIVERSITÄT DRESDEN

UNIVERSITÄT LEIPZIG

Explainable Artificial Intelligence (XAI)

- "Es gibt derzeit noch keine allgemein akzeptierte Definition von XAI."
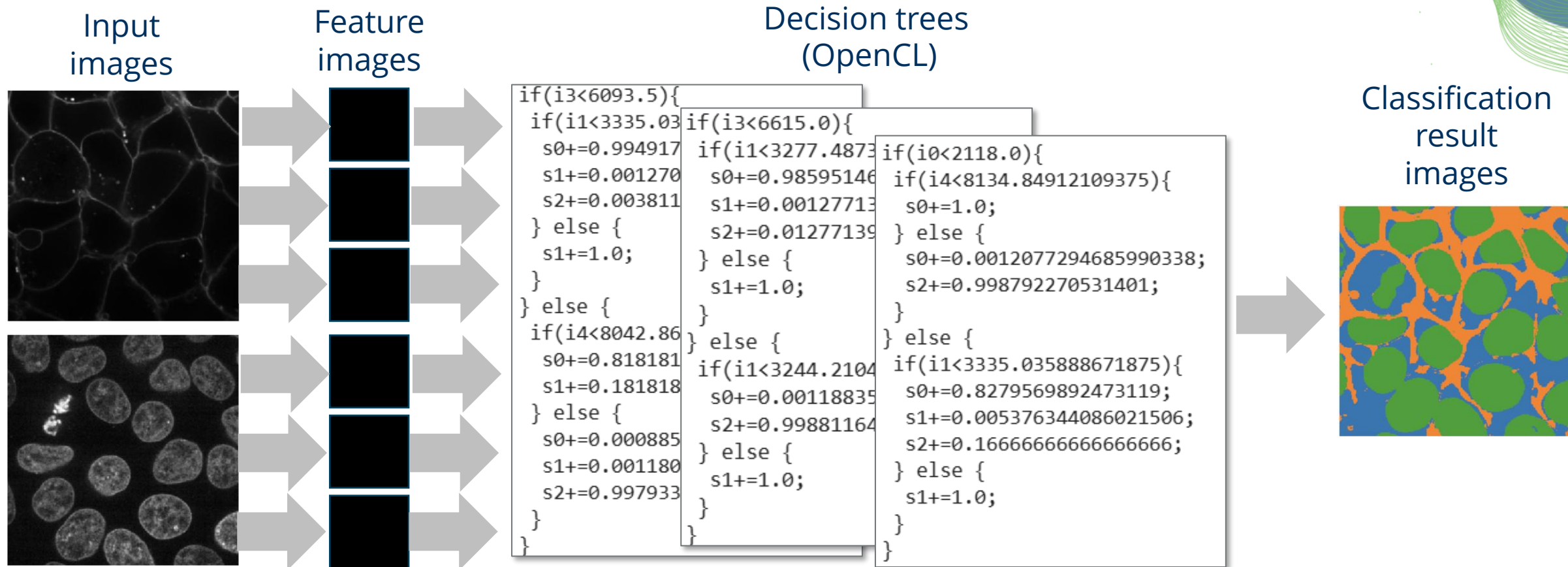
  Wikipedia [1]

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.

Robert Haase
@haesleinhuepf
AI4Medicine
Sept 24th 2025

[1] https://de.wikipedia.org/wiki/Explainable_Artificial_Intelligence Slide 2

TECHNISCHE UNIVERSITÄT DRESDEN

UNIVERSITÄT LEIPZIG

ScaDS.AI DRESDEN LEIPZIG

Explanation of Random Forest Classifiers

... by reading code   ... is quite useless

Explainability

A logically consistent line of argumentation that depicts a situation or an algorithm with complete transparency.
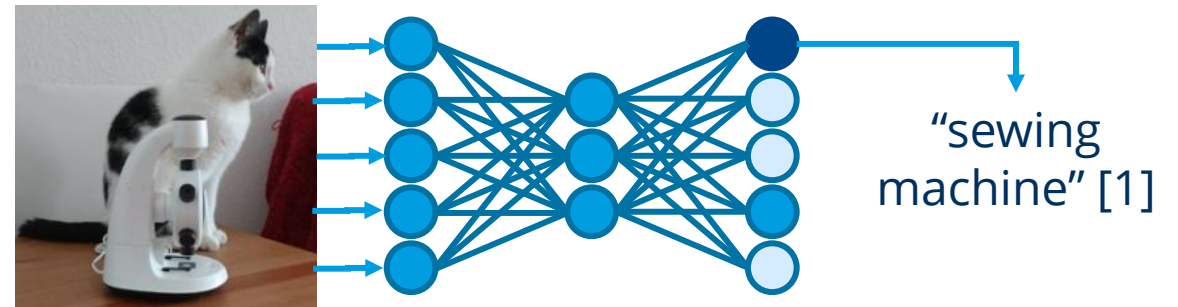
Intrinsically explainable AI-algorithms
- Example: Linear Regression

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

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

Robert Haase
@haesleinhuepf
AI4Medicine
Sept 24th 2025

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
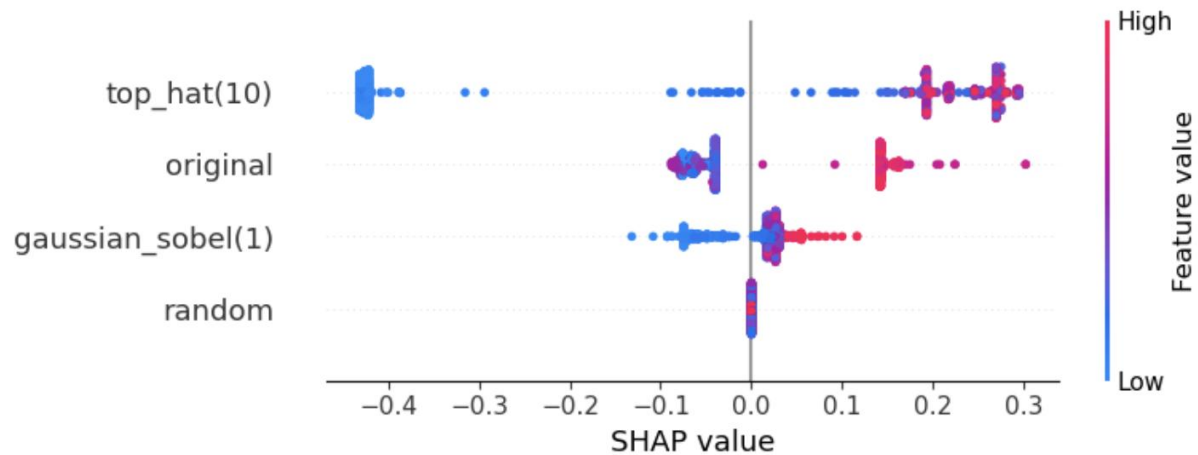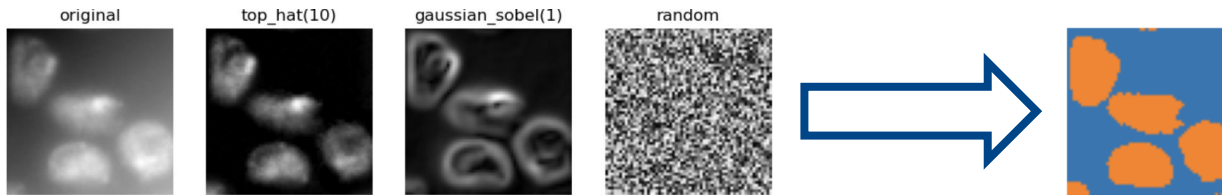- Example: Deep Neural Networks (DNN)

"sewing machine" [1]

Not easily explainable and predictable

Robert Haase
@haesleinhuepf
AI4Medicine
Sept 24th 2025

https://github.com/haesleinhuepf/git-bob-playground/issues/241

TECHNISCHE UNIVERSITÄT DRESDEN

UNIVERSITÄT LEIPZIG

ScaDS.AI DRESDEN LEIPZIG

Interpretability

Visualization of intermediate results and their influence on results

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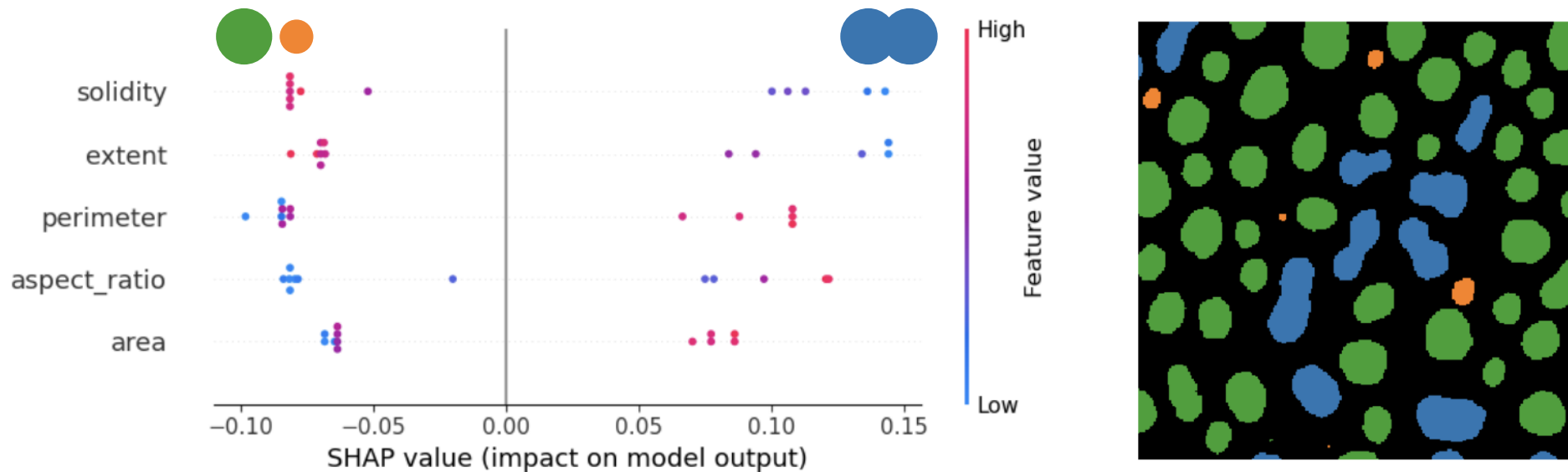Example: Shapley's Additive exPlanations (SHAP)

Robert Haase
@haesleinhuepf
AI4Medicine
Sept 24th 2025

https://haesleinhuepf.github.io/xai/30_shap/pixel_classifier.html   Slide 6

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)

"beach wagon"

Robert Haase
@haesleinhuepf
AI4Medicine
Sept 24th 2025

ScaDS.AI DRESDEN LEIPZIG

TECHNISCHE UNIVERSITÄT DRESDEN

UNIVERSITÄT LEIPZIG
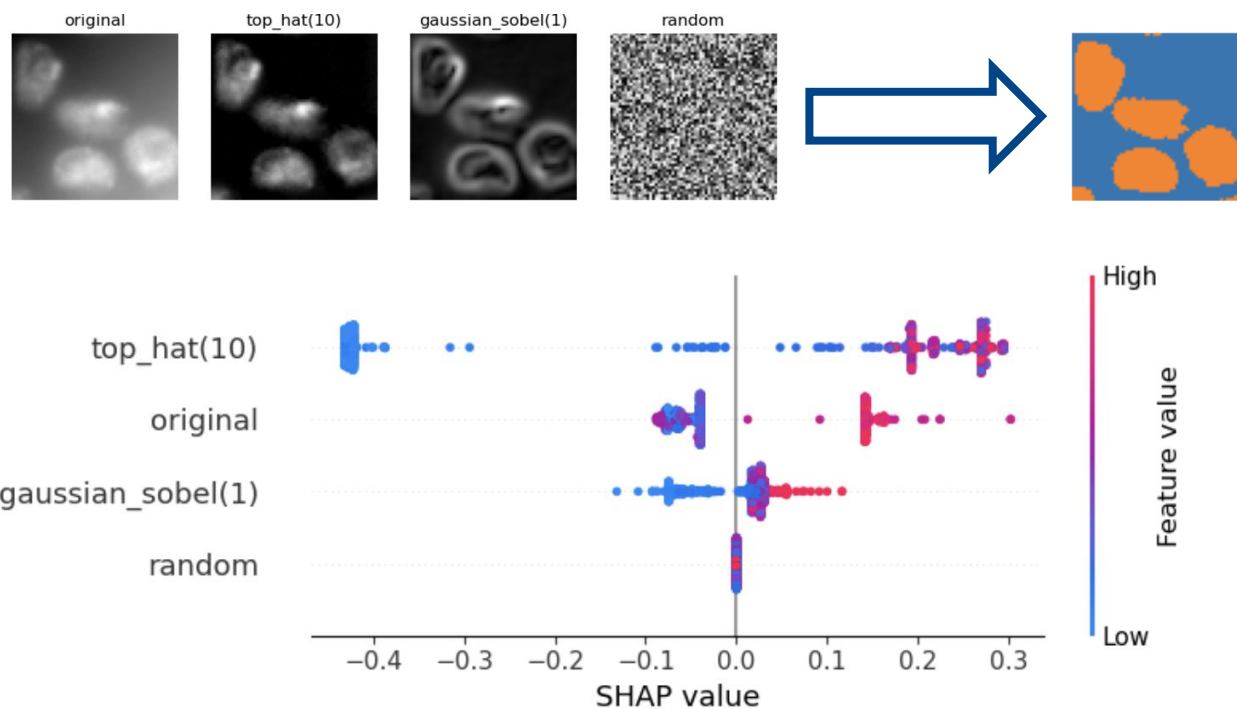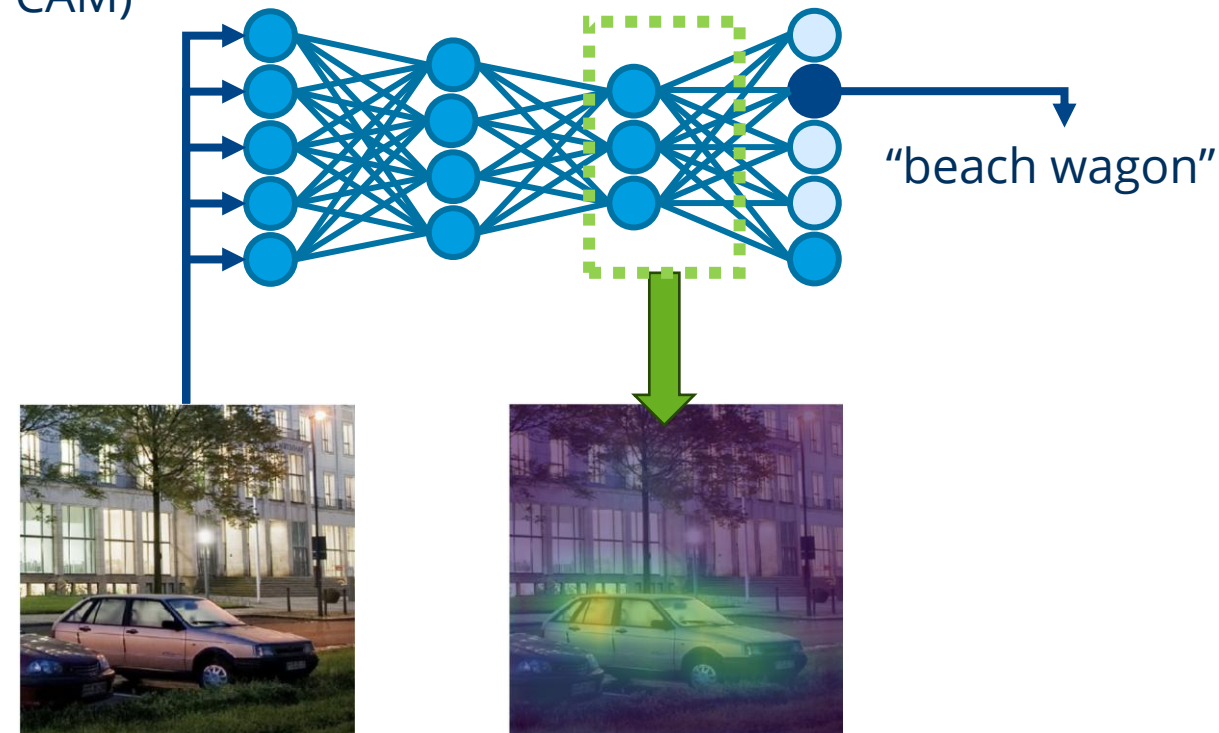
Explainable AI

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?"

Robert Haase
@haesleinhuepf
AI4Medicine
Sept 24th 2025

https://haesleinhuepf.github.io/xai/30_shap/object_classification.html

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

original   top_hat(10)

gaussian_sobel(1)   random
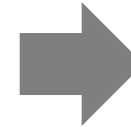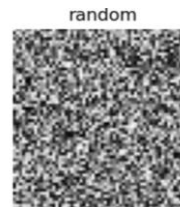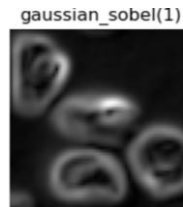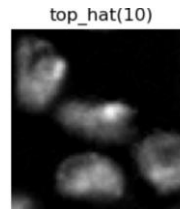
- Area
- Perimeter
- Aspect ratio
- …

- Round
- Elongated

Signal classification   Pixel classification   Object classification

TECHNISCHE UNIVERSITÄT DRESDEN

UNIVERSITÄT LEIPZIG

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

Robert Haase
@haesleinhuepf
AI4Medicine
Sept 24th 2025

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

Robert Haase
@haesleinhuepf
AI4Medicine
Sept 24th 2025
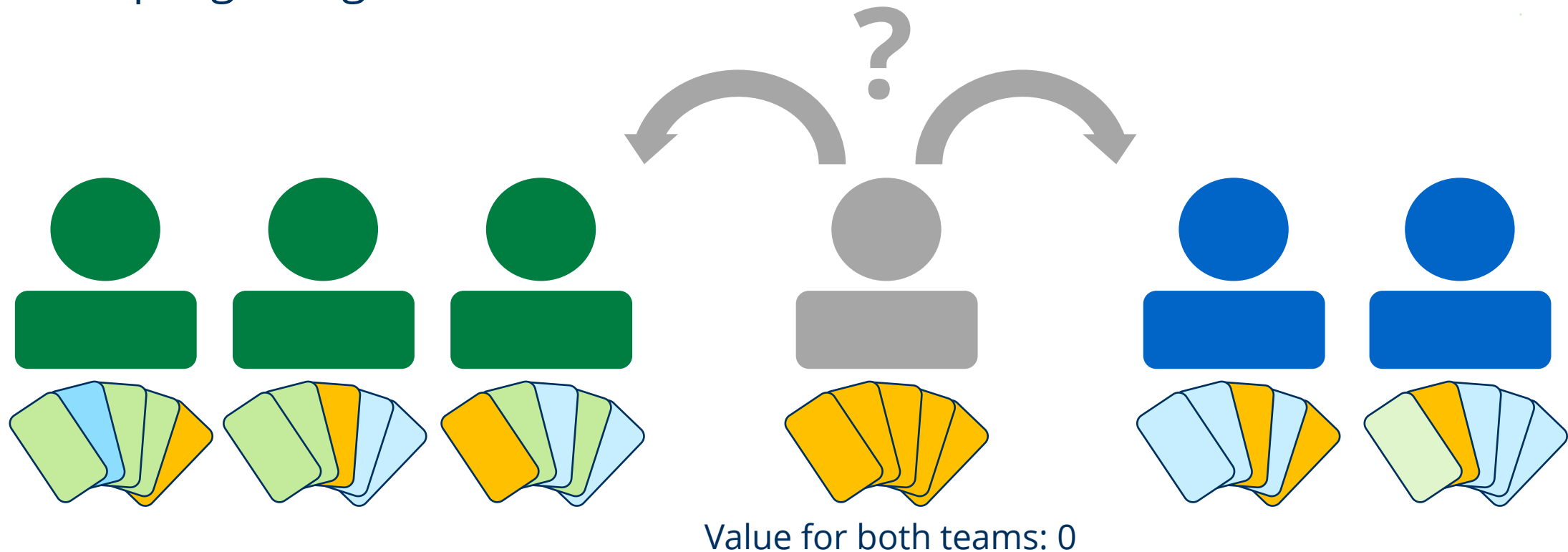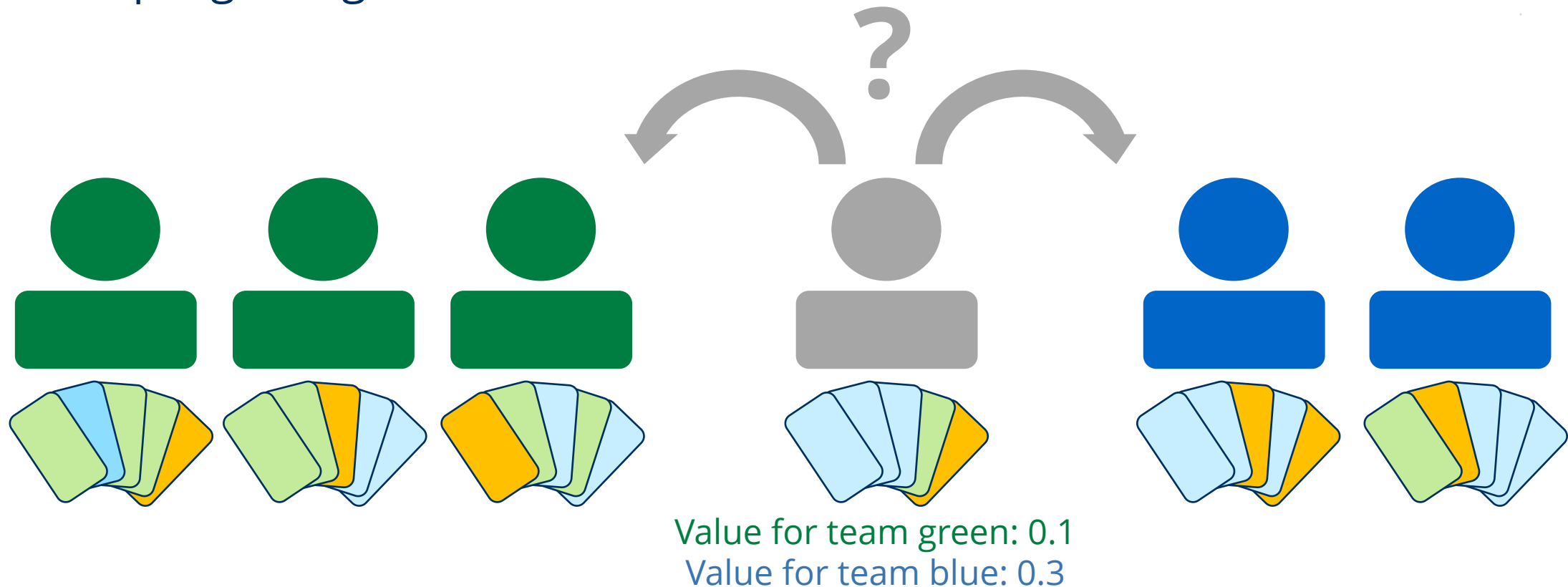
SHAP

SHapley's Additive exPlanations

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

	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

Robert Haase
@haesleinhuepf
AI4Medicine
Sept 24th 2025

SHAP

Allows interpreting [pixel] classificatio

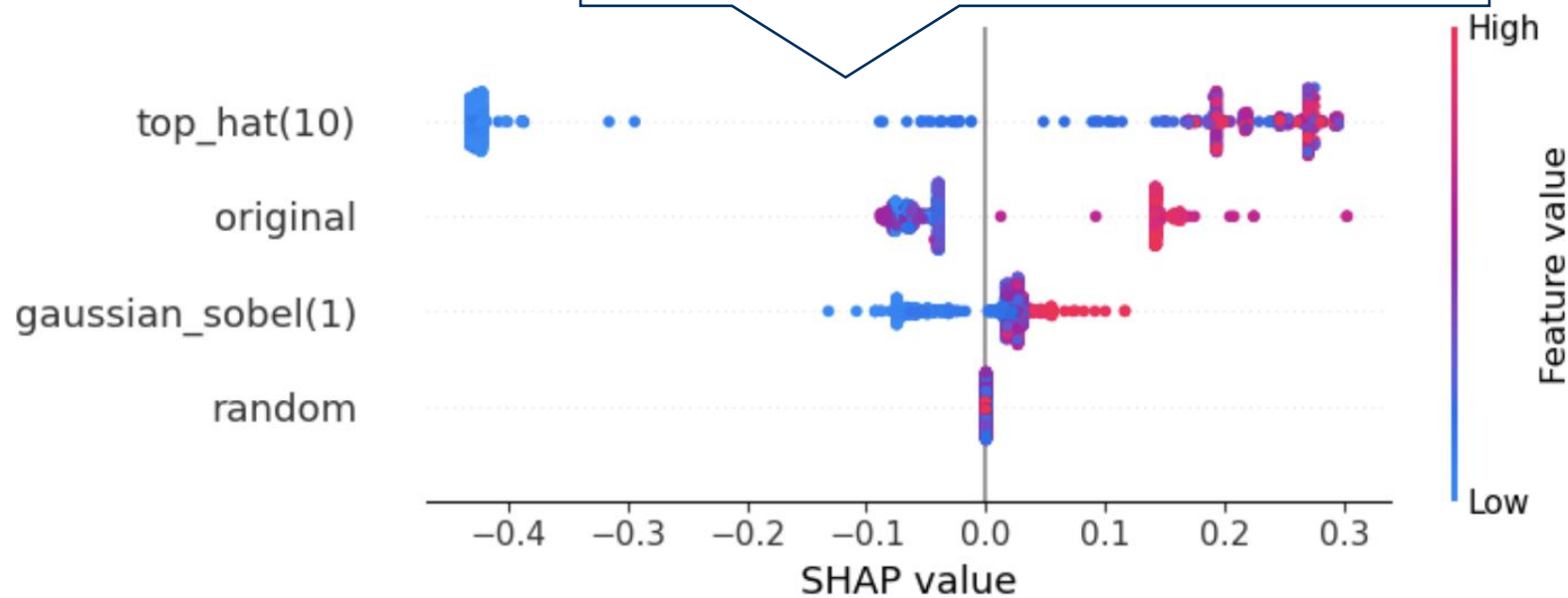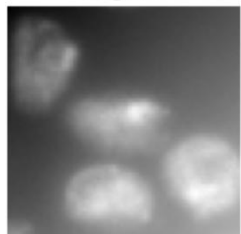"If intensity in the top-hat image is high, the classifier tends to select the positive class (orange)."
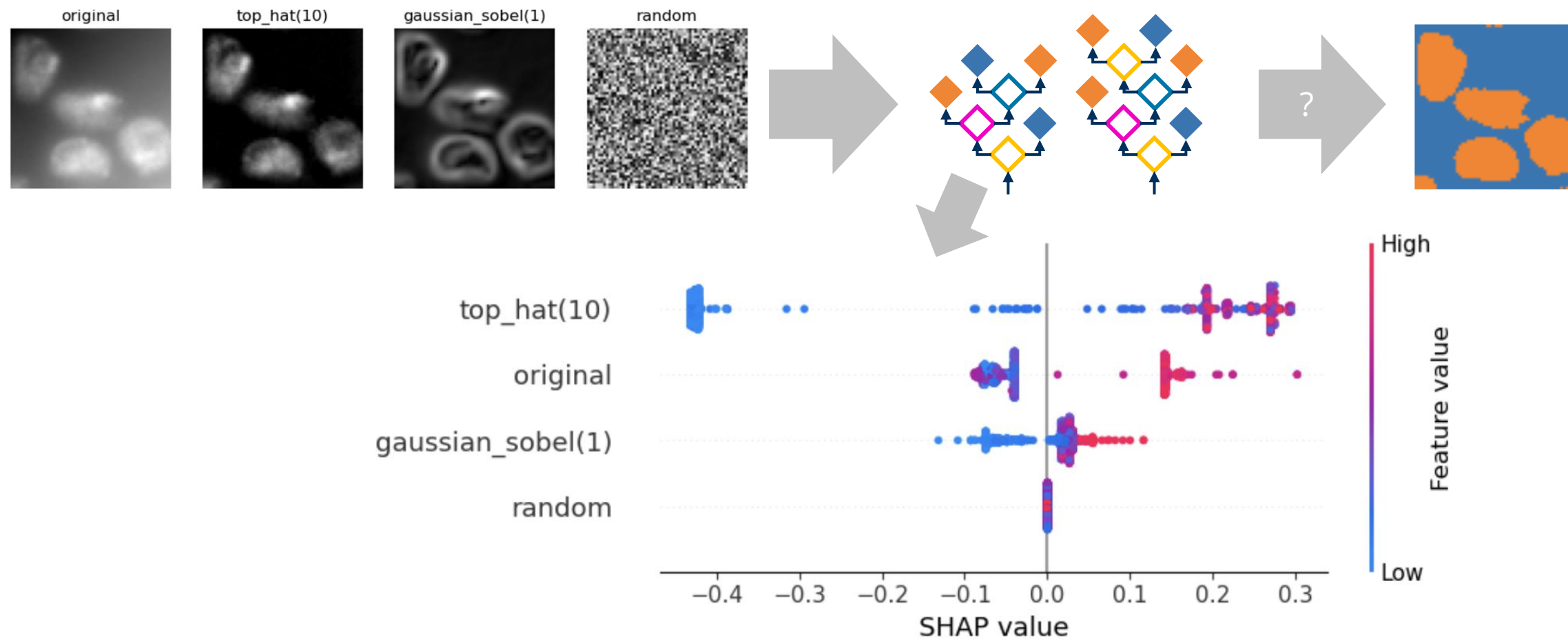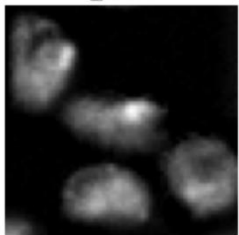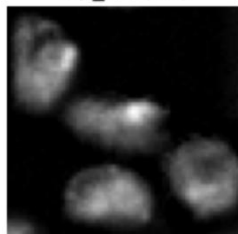
SHAP

Allows interpreting [pixel] classification

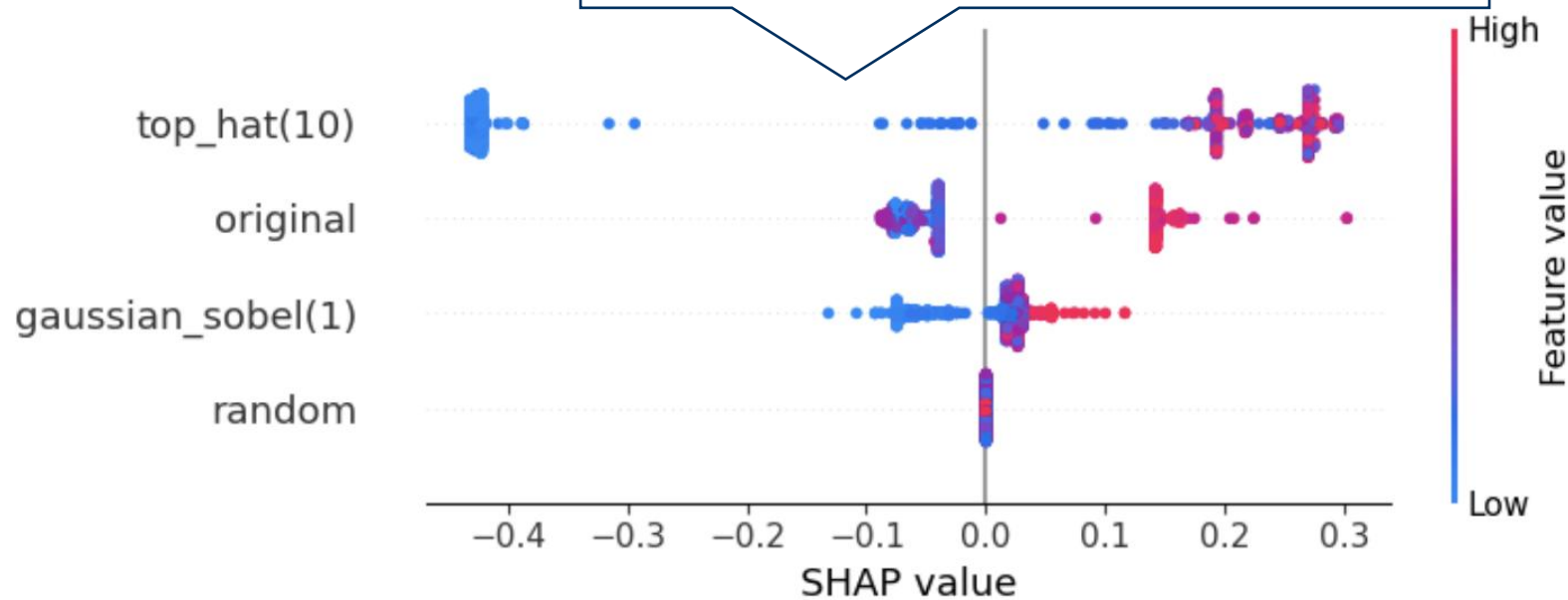"If intensity in the top-hat image is low, the classifier needs to take other features into account."

Robert Haase
@haesleinhuepf
AI4Medicine
Sept 24th 2025

SHAP
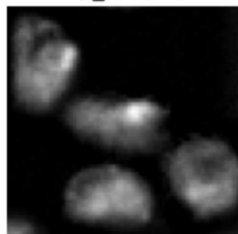
Allows interpreting [pixel] classification
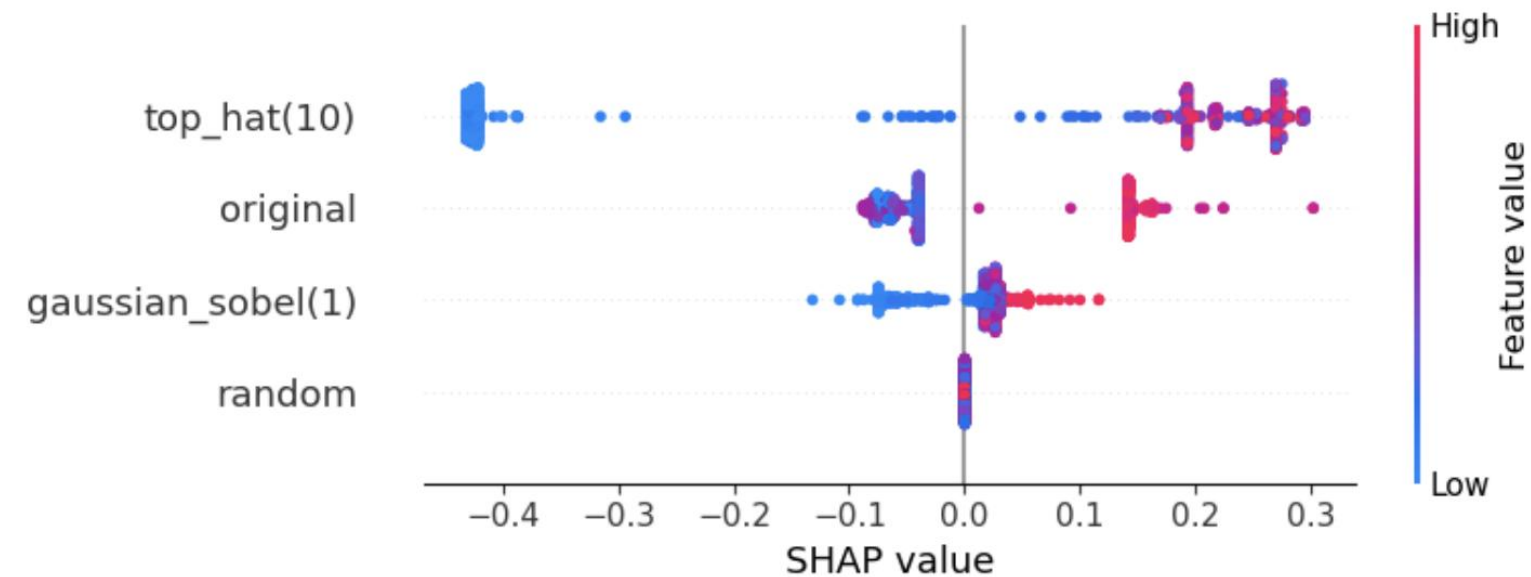

"The random feature has no value for classification."
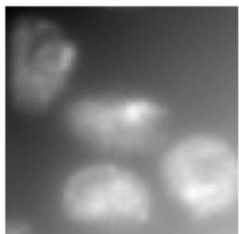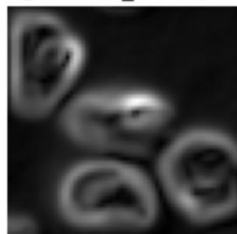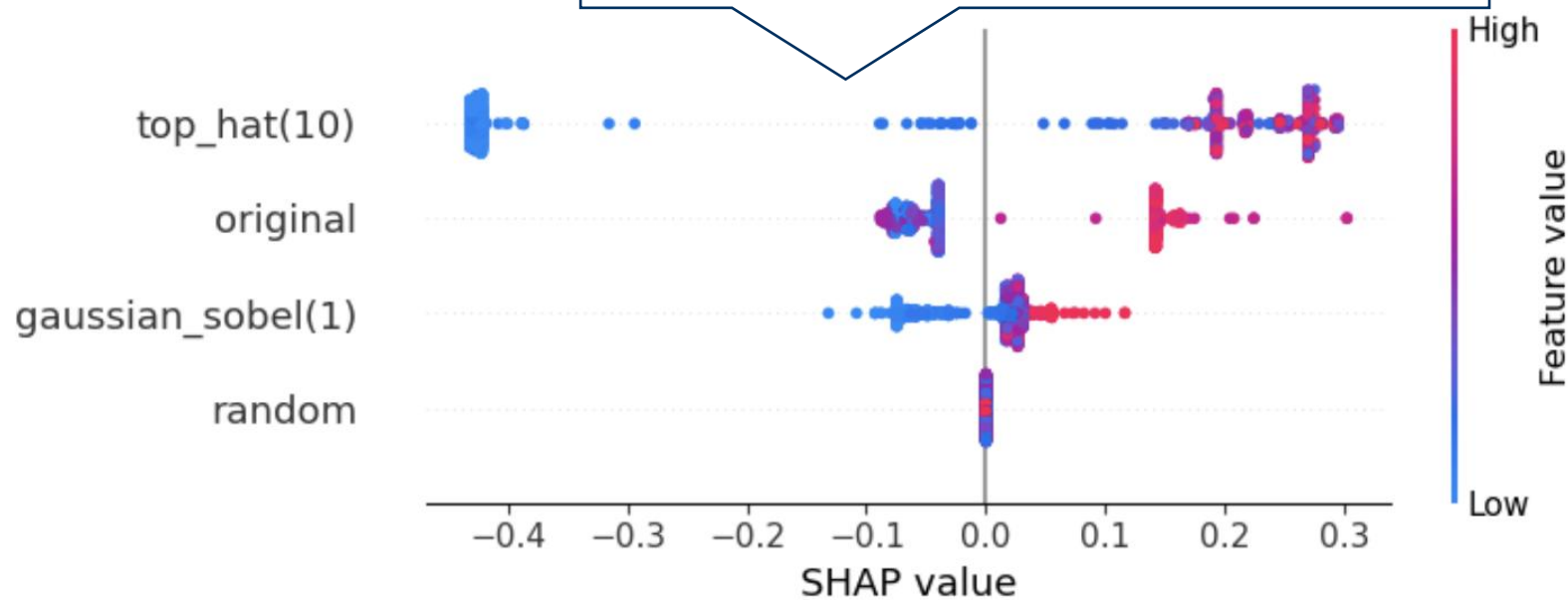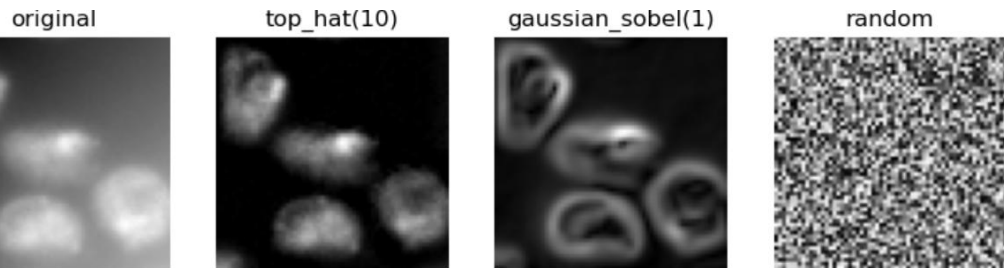
Robert Haase
@haesleinhuepf
AI4Medicine
Sept 24th 2025

Pitfall: Correlation

Correlated features may harm interpretability

Robert Haase
@haesleinhuepf
AI4Medicine
Sept 24th 2025

Pitfall: Correlation

Correlated features may harm interpretability
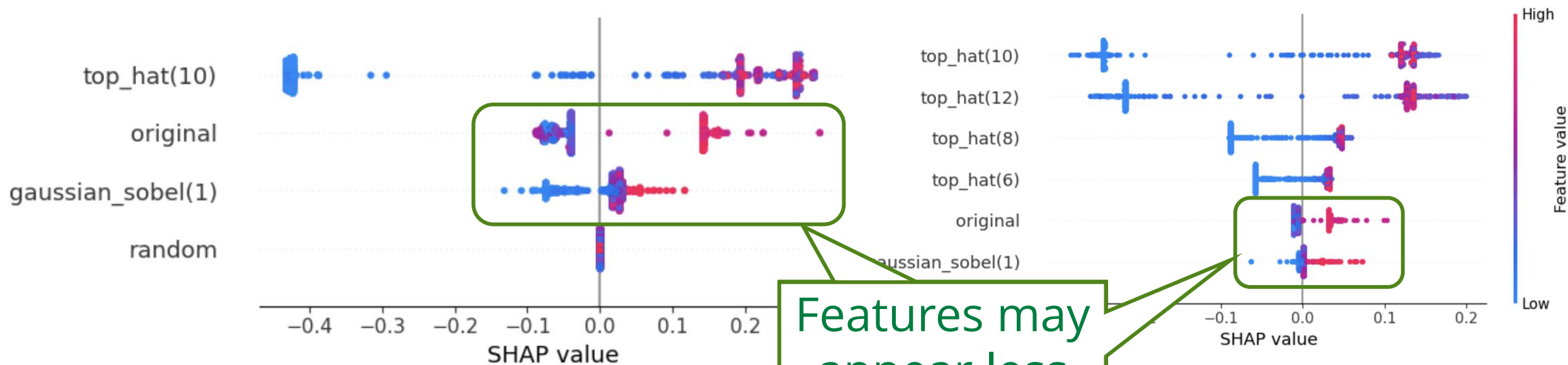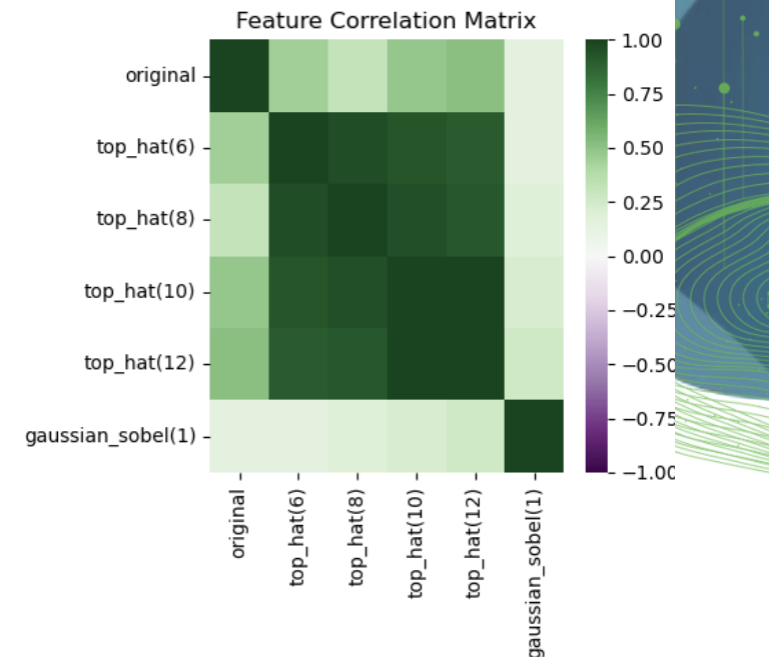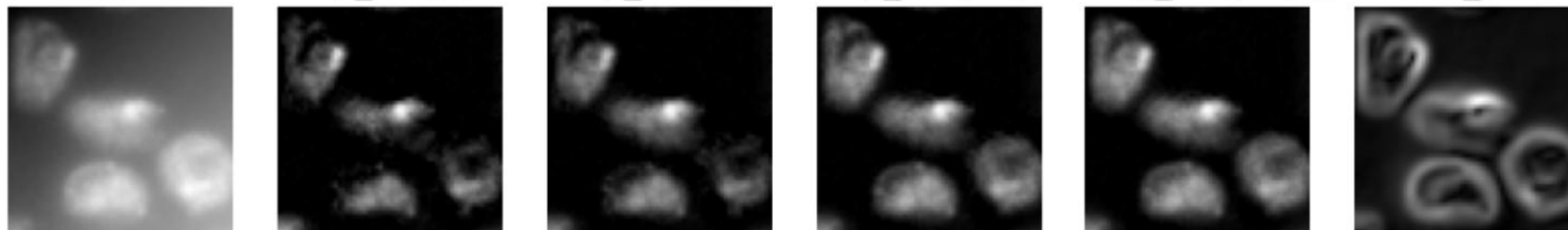
Features may appear less valuable.

Robert Haase
@haesleinhuepf
AI4Medicine
Sept 24th 2025

Interpretability

Visualization of intermediate results and their influence on results

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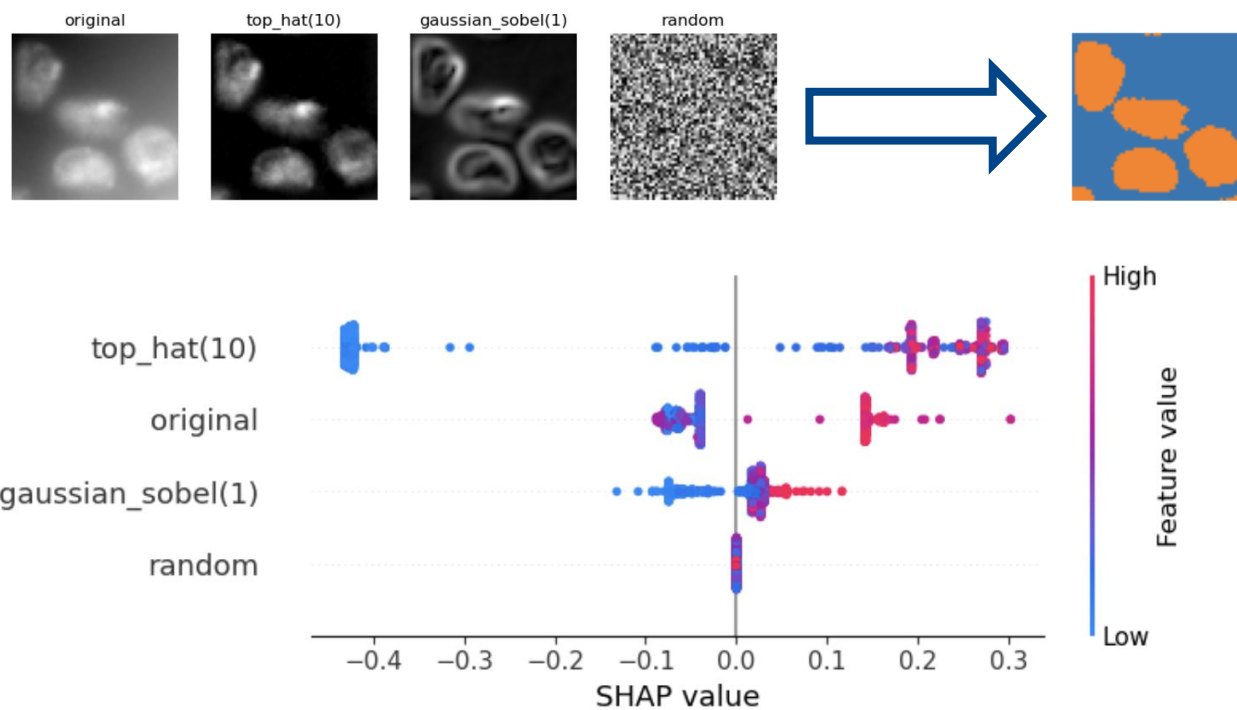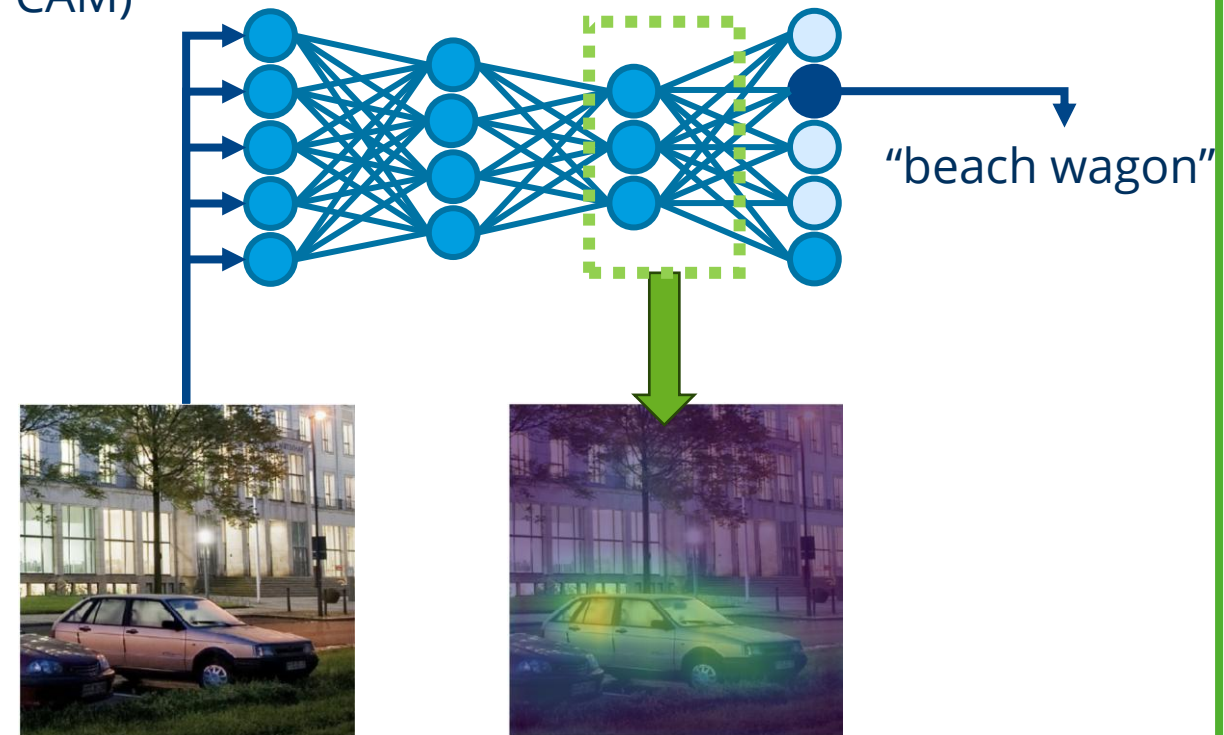Example: Shapley's Additive exPlanations (SHAP)

Model-specific methods
Example: Gradient Class Activation Maps (Grad-CAM)

"beach wagon"

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

Read more...

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

https://www.amazon.de/dp/3030686396
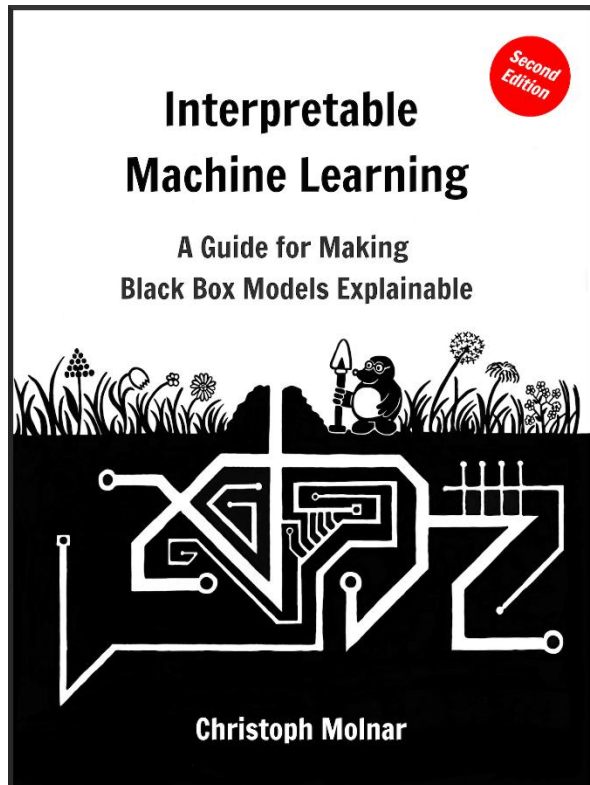
https://www.youtube.com/watch?v=dw63QH_b3Jo
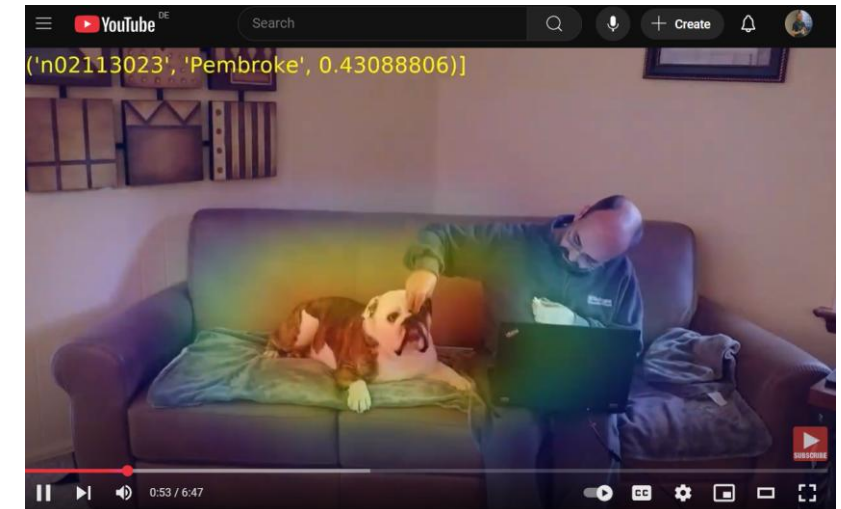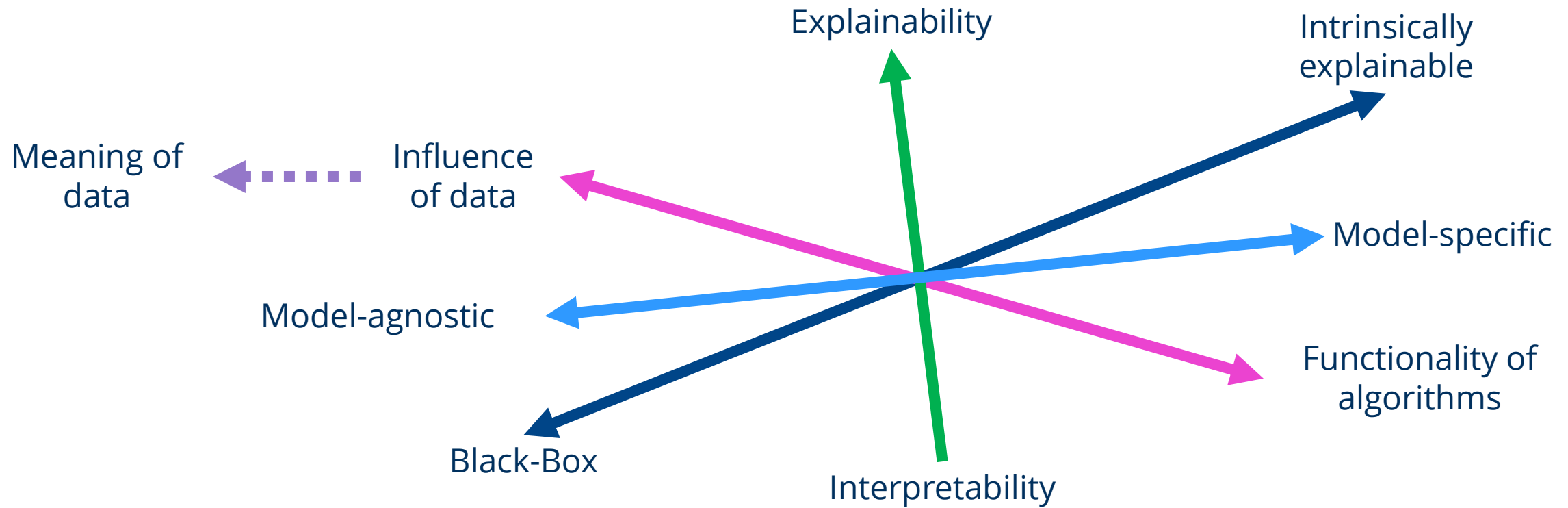
Summary: Explainable AI

Methods of XAI can be classified on different scales

Robert Haase
@haesleinhuepf
AI4Medicine
Sept 24th 2025

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.

Robert Haase
@haesleinhuepf
AI4Medicine
Sept 24th 2025

32

TECHNISCHE
UNIVERSITÄT
DRESDEN

UNIVERSITÄT
LEIPZIG

SHAP Analysis in Python

Use the opportunity and explain SHAP plots like this one!

Robert Haase
@haesleinhuepf
AI4Medicine
Sept 24th 2025