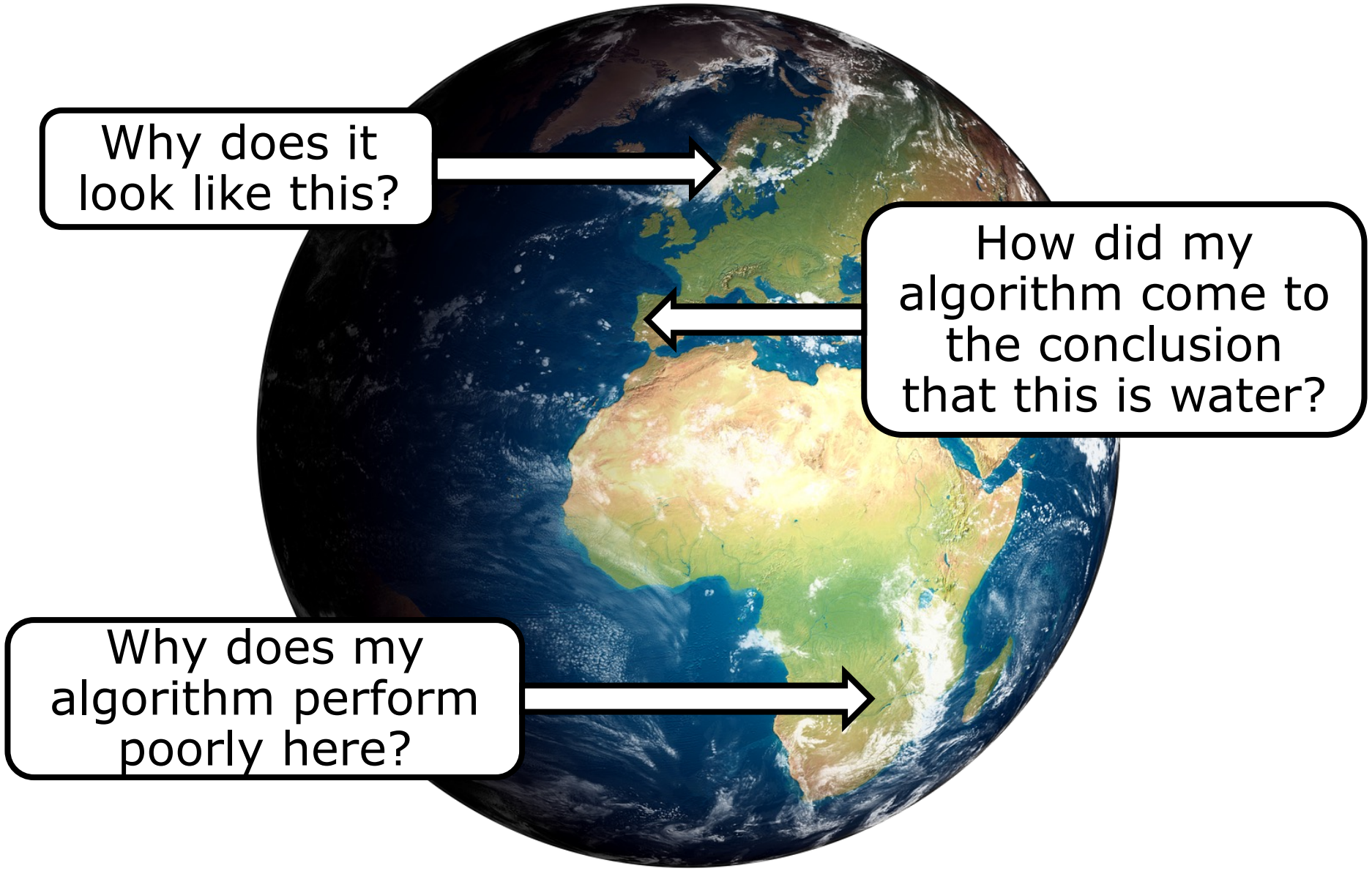
Explainable Machine Learning

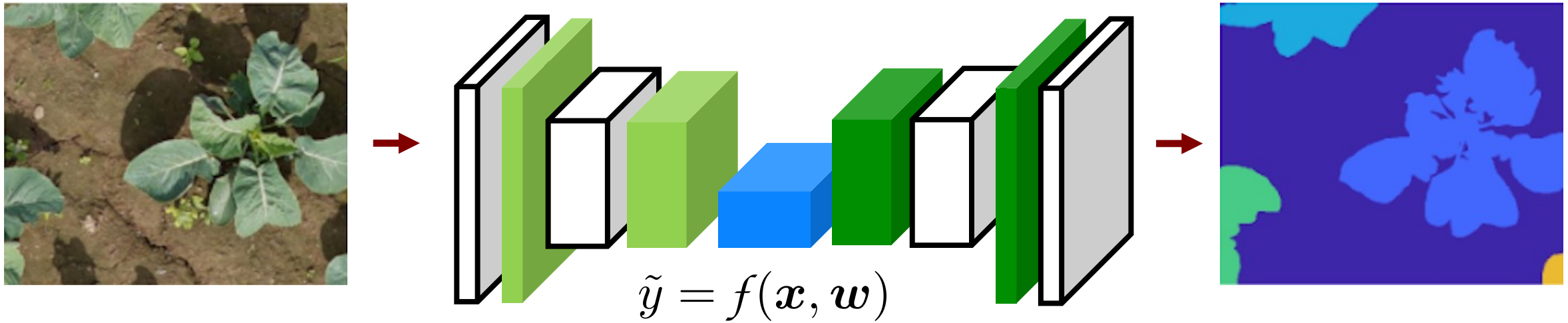
Ribana Roscher

These slides have been created by Ribana Roscher.

UNIVERSITÄT BONN


Why does it look like this?
How did my algorithm come to the conclusion that this is water?
Why does my algorithm perform poorly here?

Challenges and opportunities



Deep neural networks seem to be the prime example of black box models.

Images from: Kierdorf, J., Junker-Frohn, L.V., Delaney, M., Donoso-Olave, M., Burkhart, A., Jeanicke, H., Müller, O. & Roscher, R. (2022). GrowliFlower: A time series dataset for GROWth analysis of cauliflower, submitted.

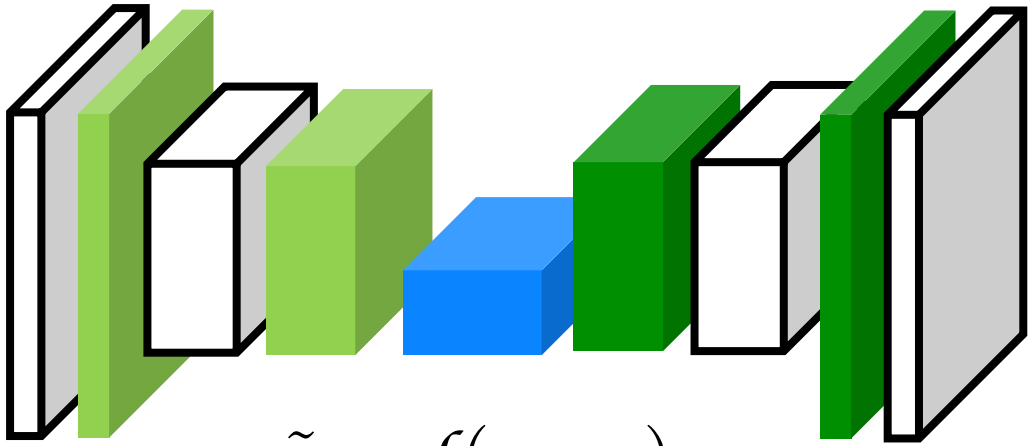
Core elements

- **Transparency**
- **Interpretability**
- **Explainability**

Transparency

Transparency of a machine learning approach concerns its different ingredients such as

- overall model structure
- individual model components
- learning algorithm
- how the specific solution is obtained by the algorithm



$$\tilde{y} = f(\boldsymbol{x}, \boldsymbol{w})$$

Interpretability

Interpretability is about **making sense** of the obtained machine learning model with the aim to present some properties in **understandable terms** to a human such as

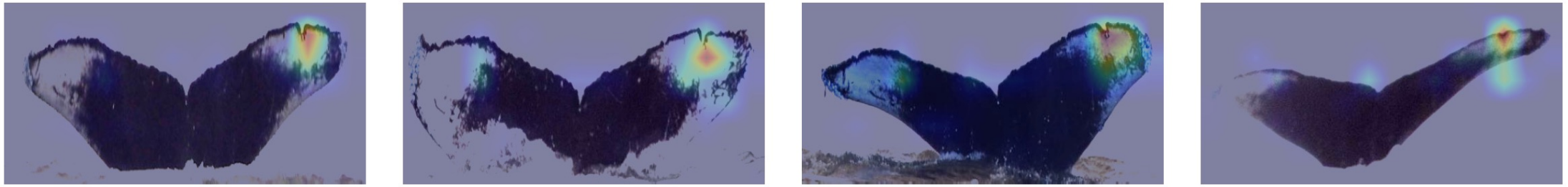
- feature statistics such as feature importance
- data point with special significance such as archetypes or prototypes
- model parameters
- patterns in the model decision process

Heatmaps

Why are these images are identified as the same whale?

Different images of a whale

Heatmaps



Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2019). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, *128*(2), 336-359.

Explainability

- Research into **explainable machine learning** is widely recognized as important
- Joint understanding of the concept of explainability still needs to evolve
- Adadi & Berrada (2018): essentially four reasons to seek explanations
  - to justify decisions
  - to (enhance) control
  - to improve models, and
  - to discover new knowledge

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138-52160.

Interpretability vs. explainability

Interpretability

Present some properties of a machine learning model in understandable terms to a human

Explainability

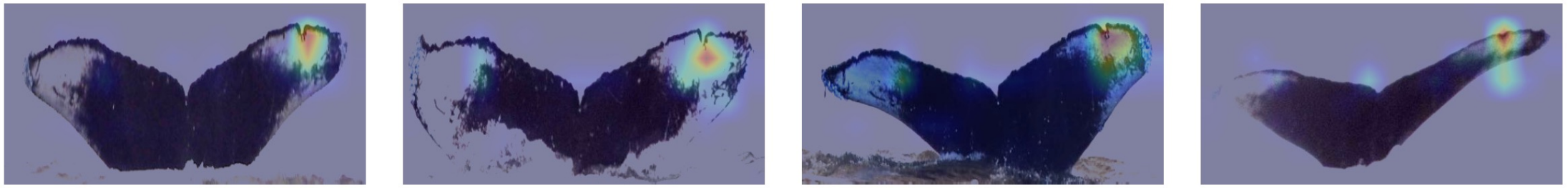
Combine interpretable entities with domain knowledge (and an analysis goal)

Why do we distinguish?

- Explanation depends on the use case

Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. *IEEE Access*, *8*, 42200-42216.

Interpretability vs. explainability



Interpretation

The score for whale ID […] is significantly influenced by the image pattern in the right upper corner of image […].

Explanation

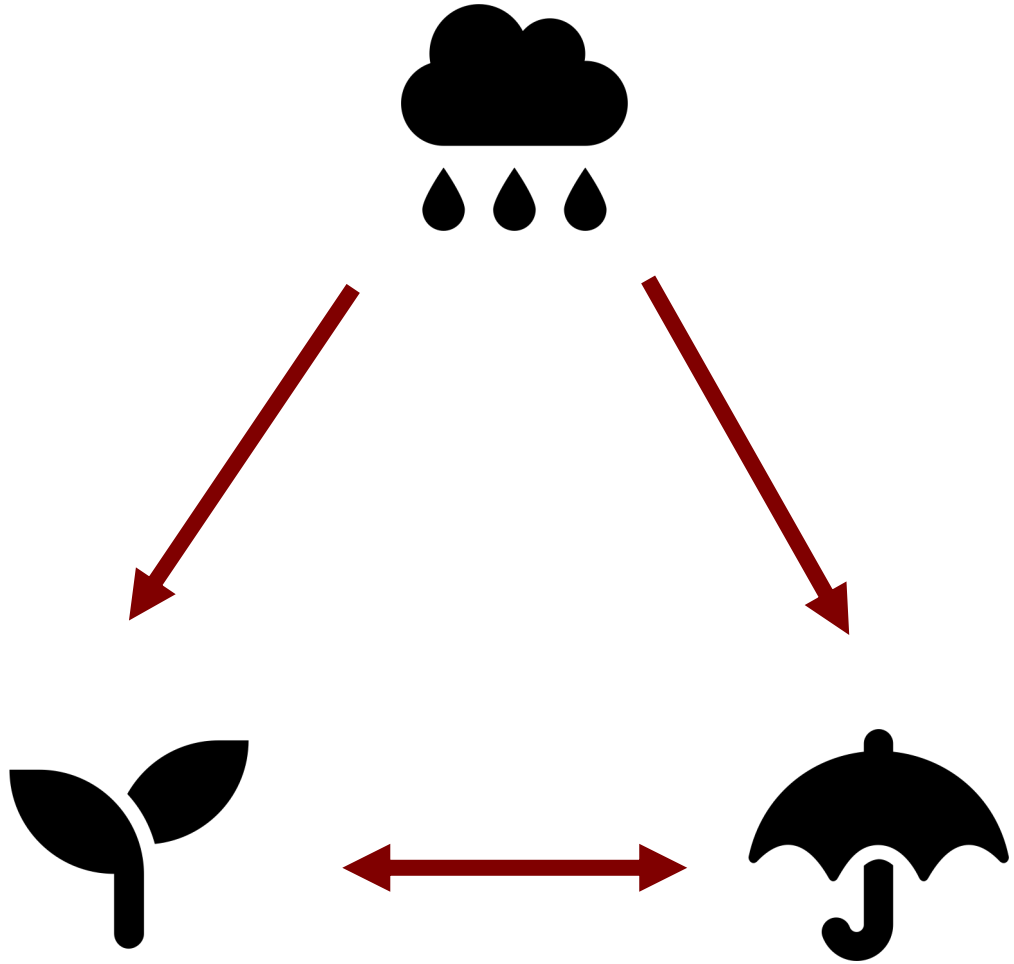
The notch in the fluke of the whale with ID […] is a relevant fluke pattern for identifying this specific whale.

Connection to correlation and causation

Causation means that an output is the result of the occurrence of a specific input (**cause and effect**)

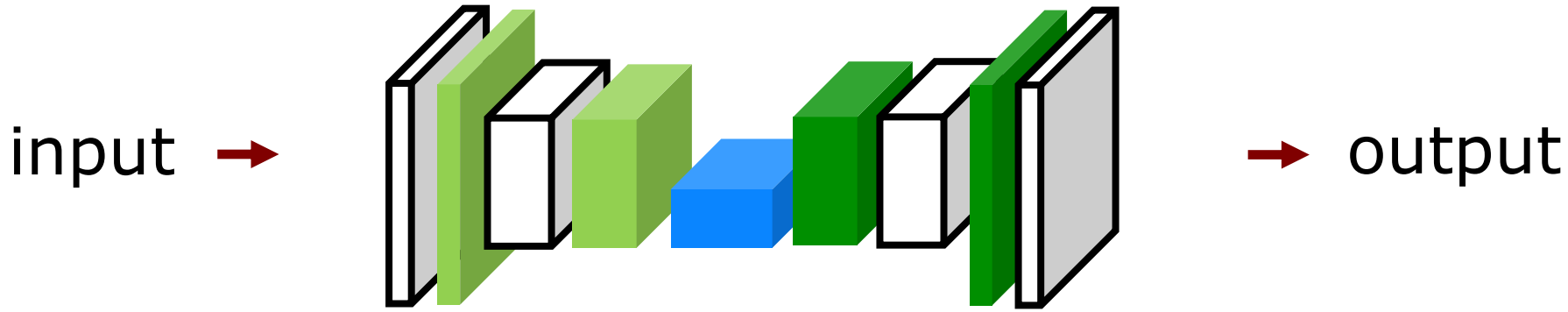
Correlation measures the relationship between input and output

- does not imply causation



Connection to correlation and causation

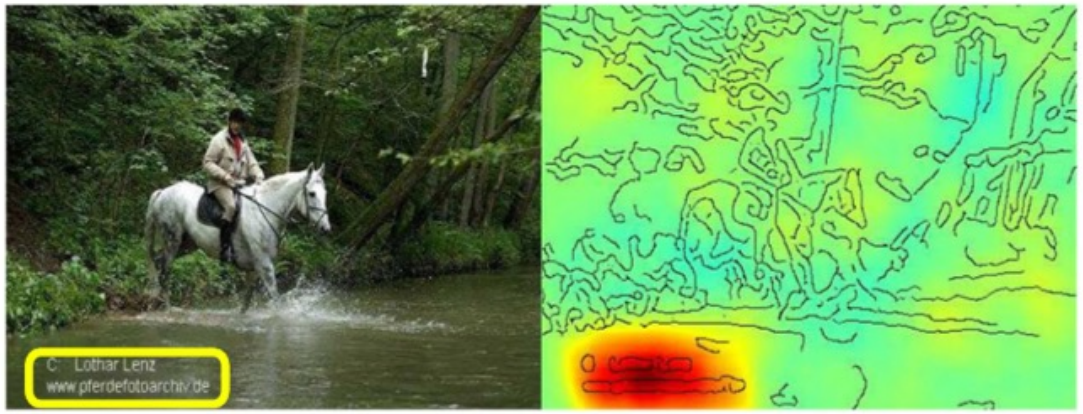
Interpretation tools present properties of a machine learning model and generally build on correlation



Confirmation bias

Underlying tendency to search for explanations which are in line with our existing knowledge

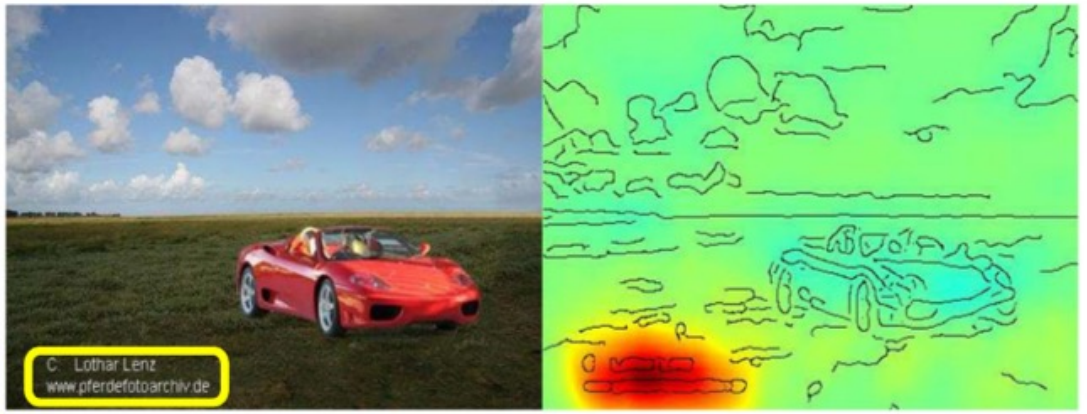
Clever Hans effect



Source tag present

↓

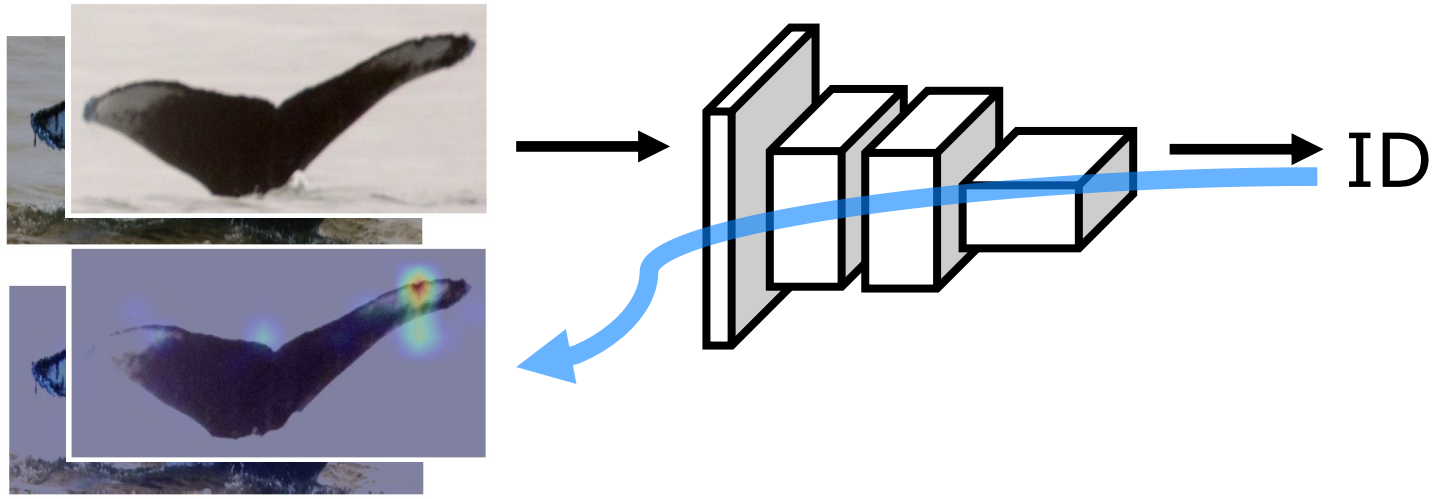
Classified as horse



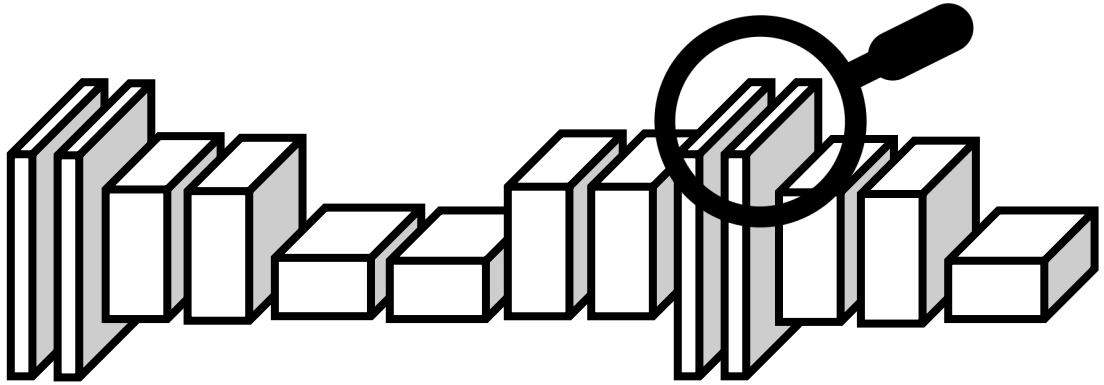
S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking Clever Hans predictors and assessing what machines really learn," Nature Communications, vol. 10, no. 1, p. 1096, 2019.

Approaches categorized by specificity

Explaining output by input (post-hoc, model-agnostic)

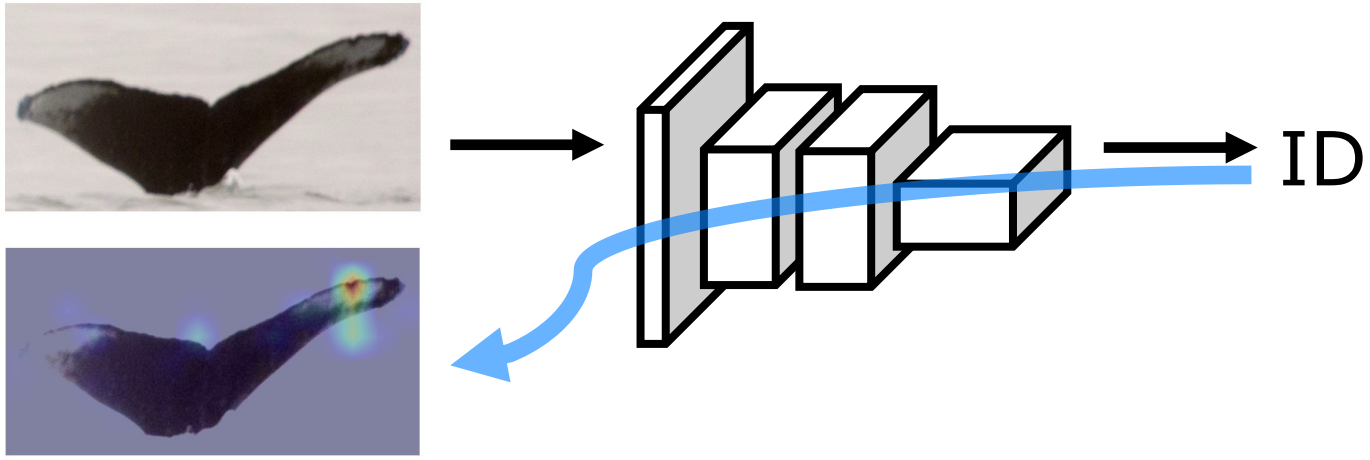


Explaining the whole model or parts (model-specific)

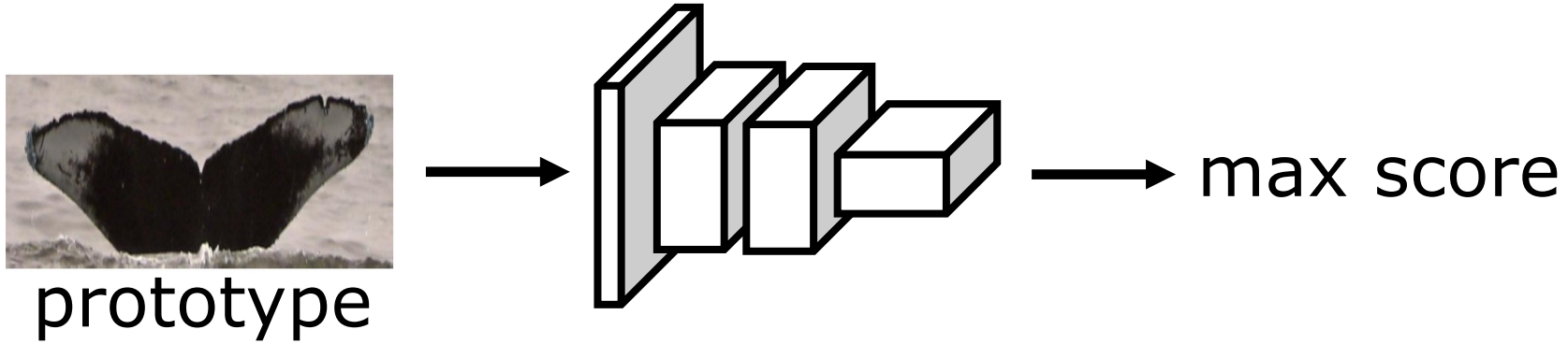


Approaches categorized by locality

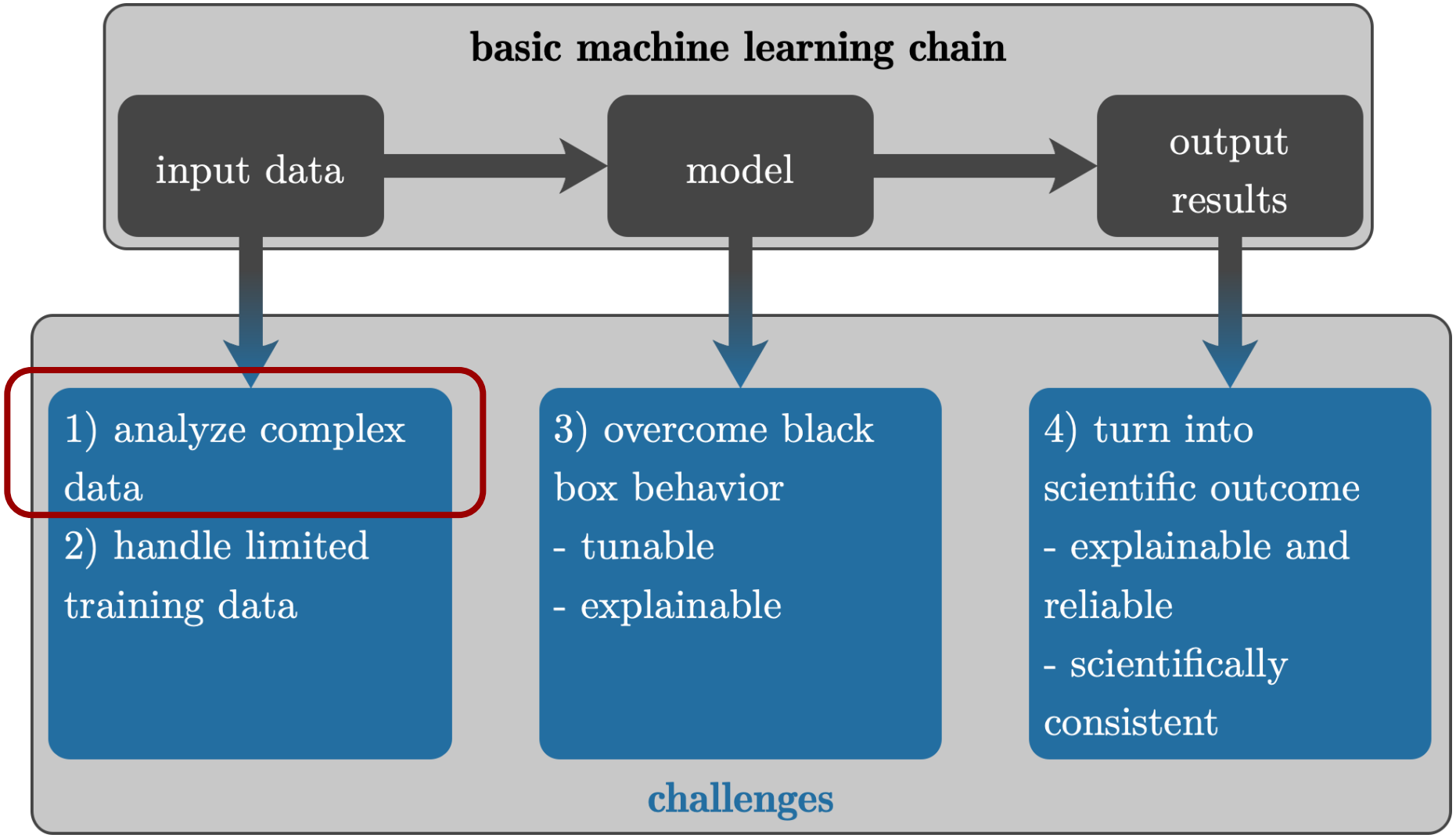
Explain locally (individual output)



Explain globally (entire model)

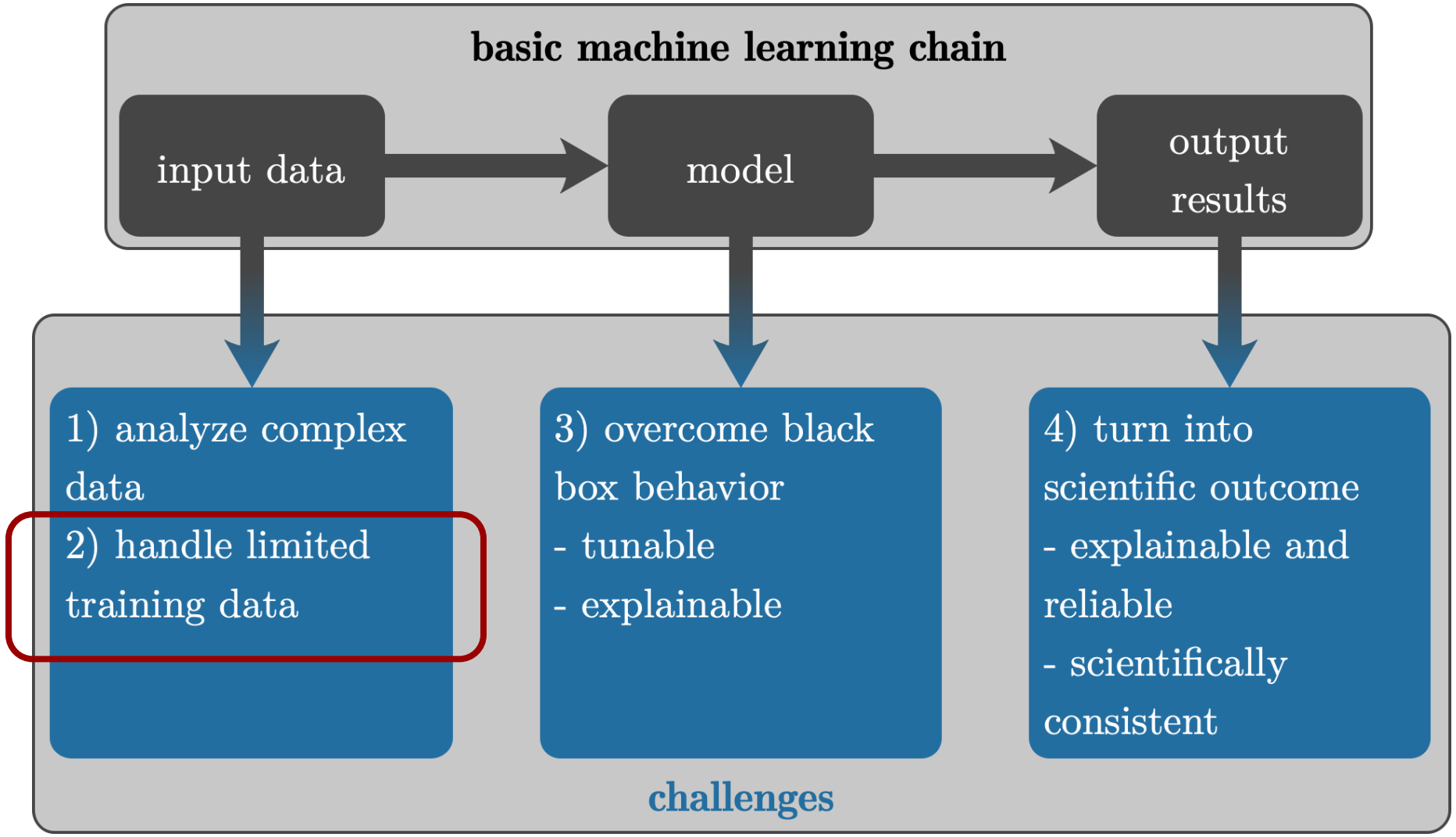


Can explainable ML be useful in RS?



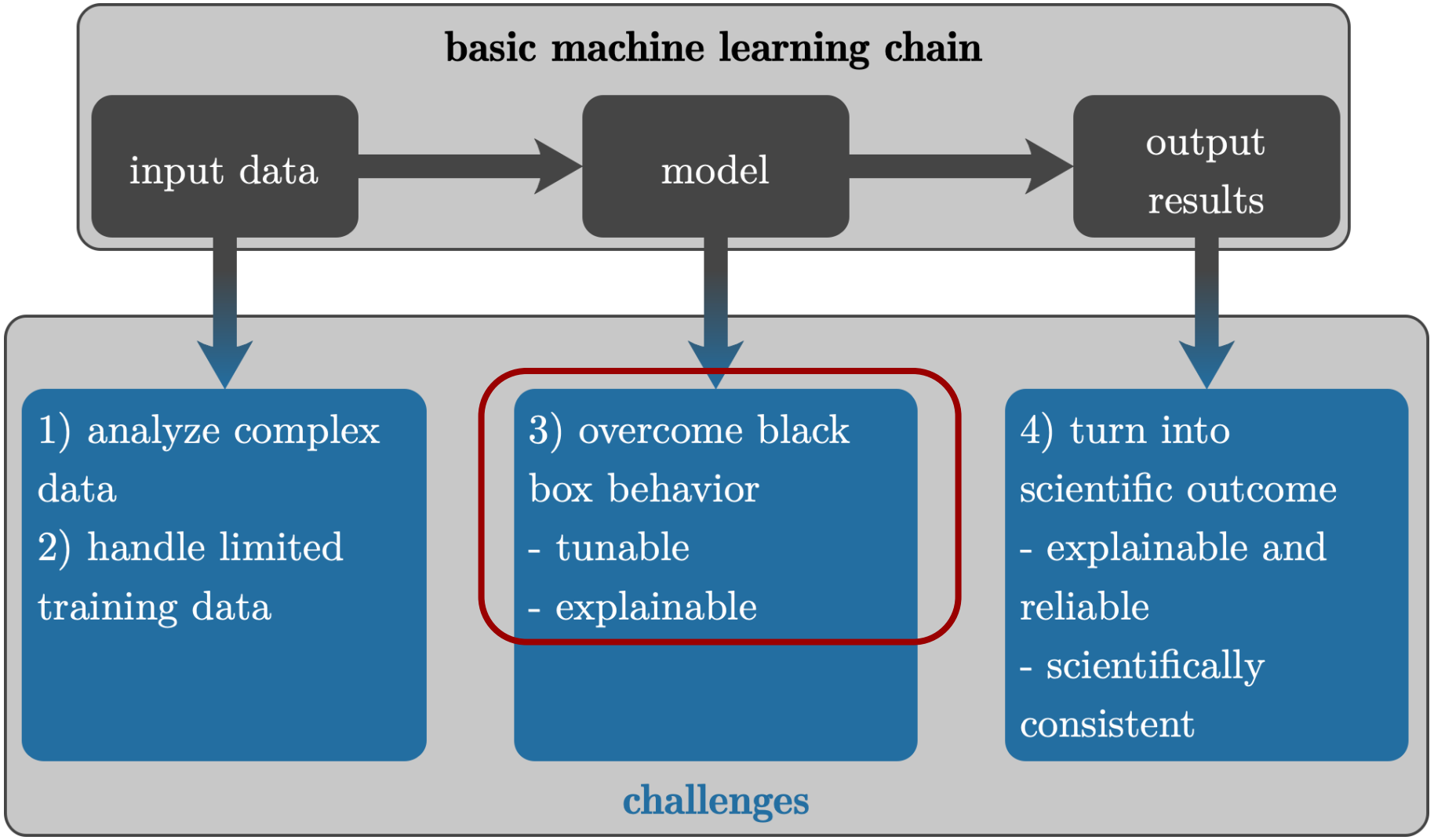
Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explain it to Me-Facing Remote Sensing Challenges in the-and Geosciences with Explainable Machine Learning. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *5*, 817-824.

Can explainable ML be useful in RS?



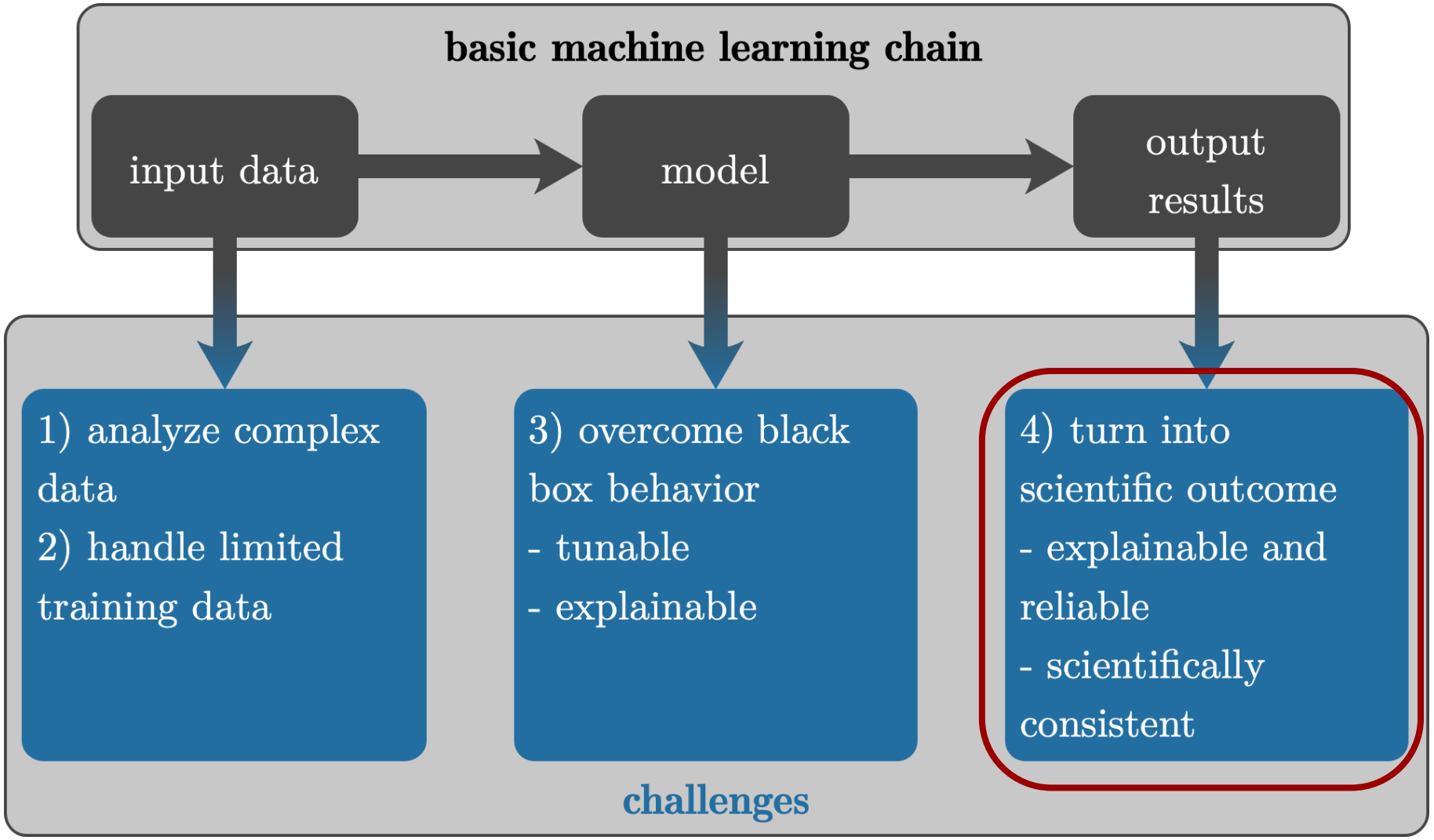
Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explain it to Me-Facing Remote Sensing Challenges in the-and Geosciences with Explainable Machine Learning. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *5*, 817-824.

Can explainable ML be useful in RS?



Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explain it to Me-Facing Remote Sensing Challenges in the-and Geosciences with Explainable Machine Learning. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *5*, 817-824.

Can explainable ML be useful in RS?



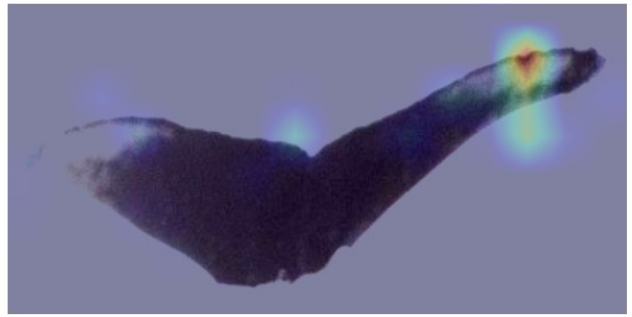
Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explain it to Me-Facing Remote Sensing Challenges in the-and Geosciences with Explainable Machine Learning. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *5*, 817-824.

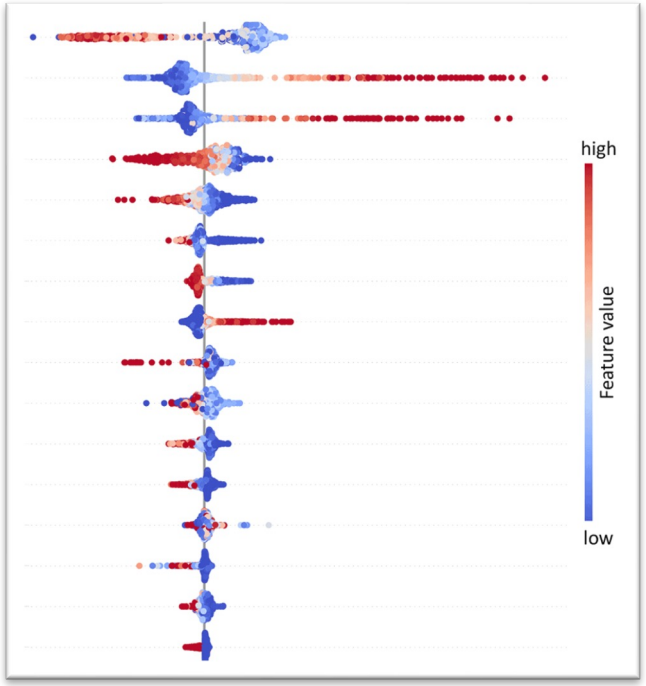
Methods and Applications

Terms connected to interpretability

Reasons to seek explanations

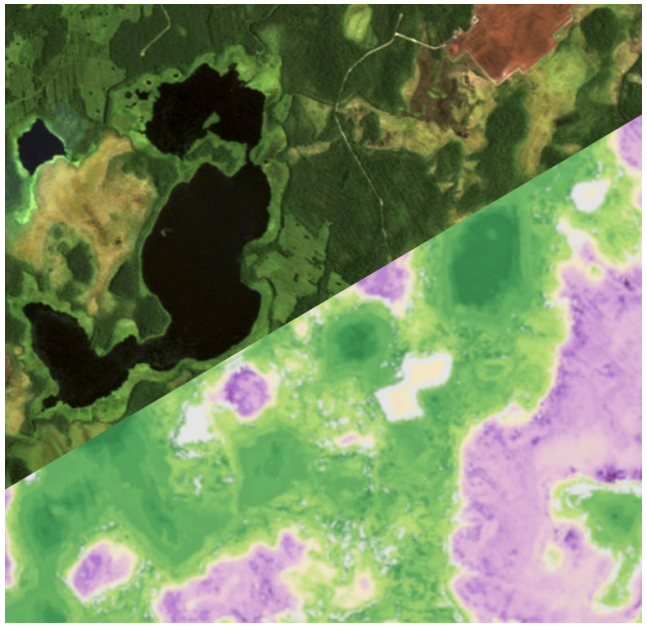
Justify decisions





Example: whale monitoring and ozone value estimation

Discover knowledge



Example: understanding wilderness

Reasons to seek explanations

Justify decisions

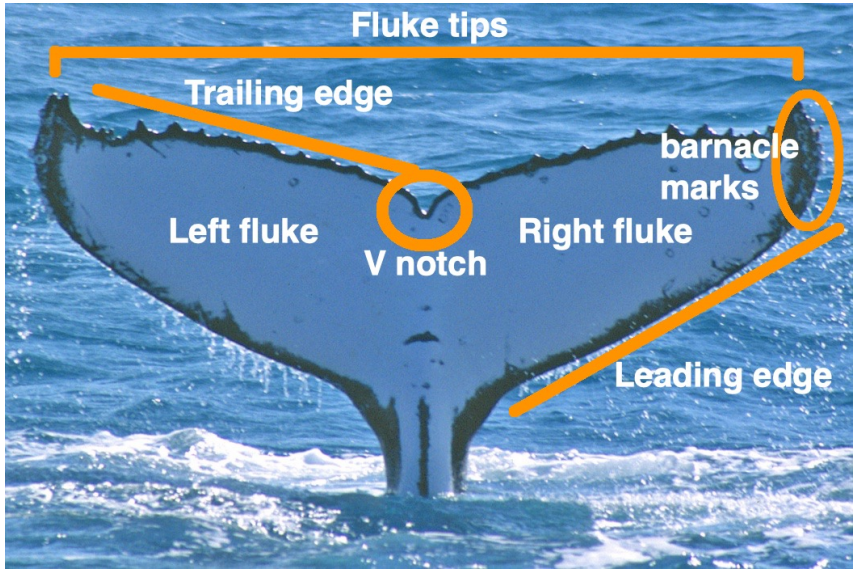
- Occlusion sensitivity maps (OSM)
- Class activation maps (CAM) and Grad-CAM
- Shapley values

Discover knowledge

- Activation space occlusion sensitivity (ASOS)

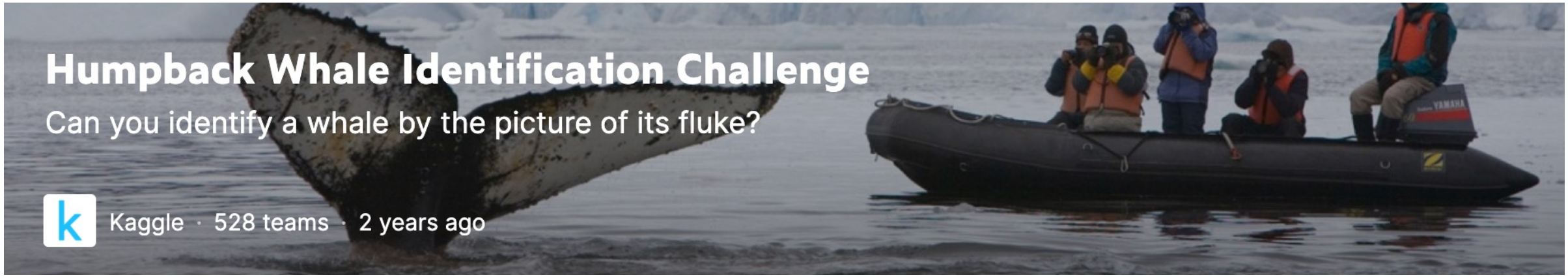
Whale monitoring

- Whale populations are threatened by commercial whaling, ocean warming and competition for food
- Important to monitor whales and spatio-temporal migration



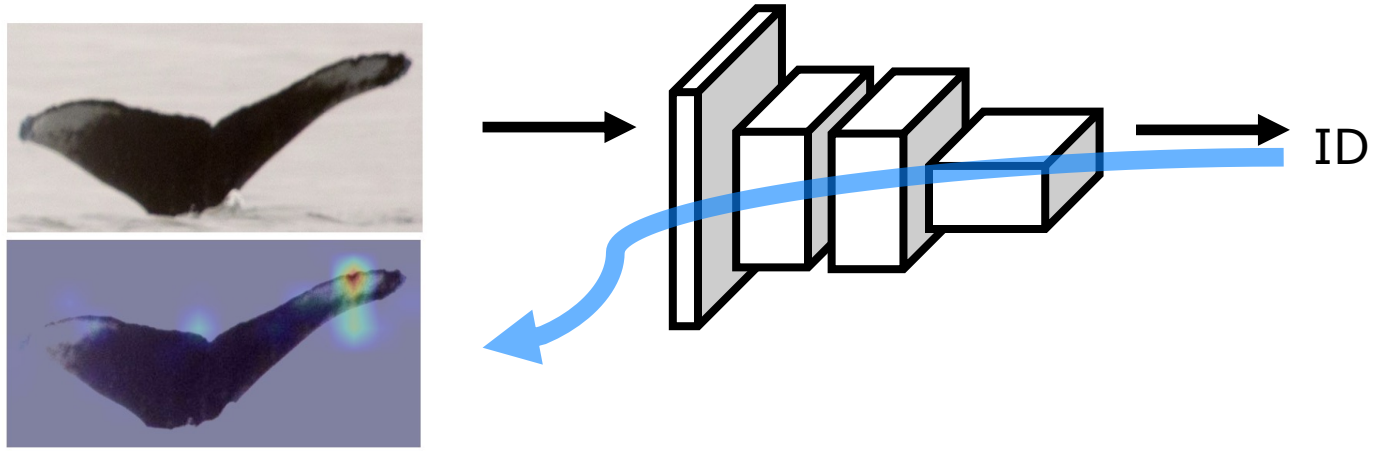
@Happywhale

Neural networks for identification



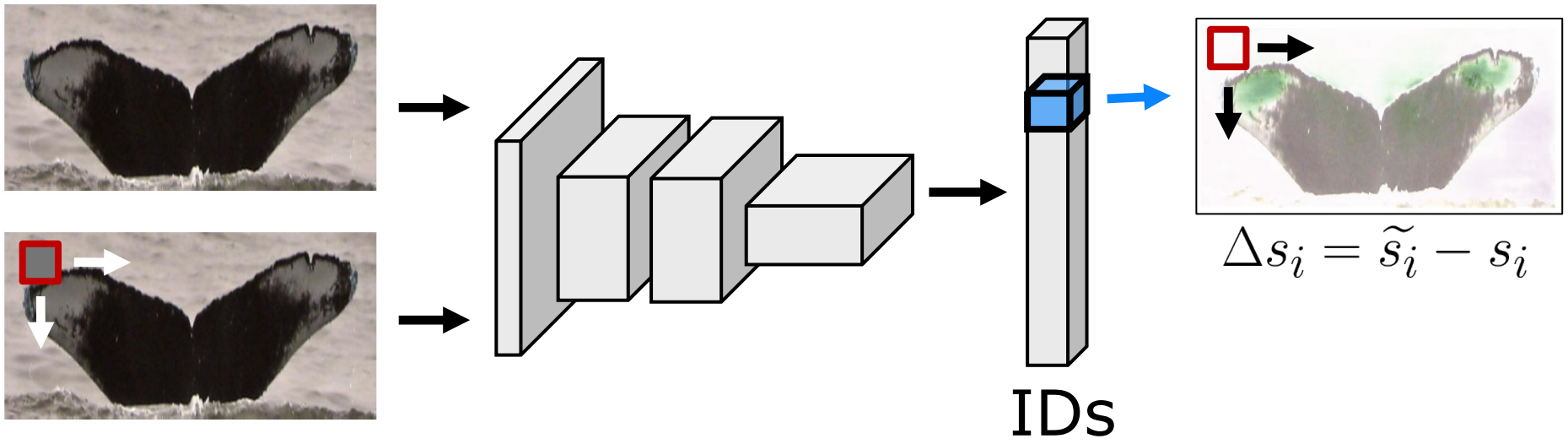
Research questions

- Why does the network identify images as the same whale/different whales?
- Is a whale expert looking at the same area in the image as the network?



Occlusion sensitivity maps

- Evaluate the **sensitivity** of the trained model to **occlusions**
- **Difference** between the original score and the score after applying occlusion



Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818-833)

Occlusion sensitivity maps

Green values

- score decreases → regions are an indicator for the specific whale

Purple values

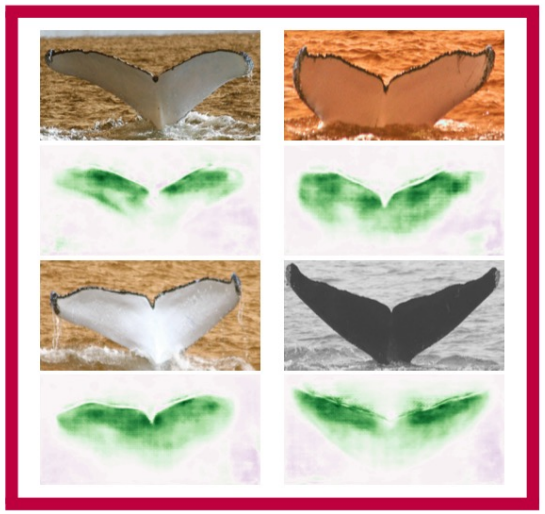
- Score increases → regions are an indicator for other whales/not the specific whale

Advantages and disadvantages

- Simple and versatile
- Model-agnostic
- Definition of the patch is challenging
  - Patch is agnostic to semantic segments which can differ in size
  - Every occlusion is a signal (noise, single value, Gaussian kernel, …)
  - Distribution-shift: patch might be out-of-distribution
- Computationally expensive

Spectral clustering of OSM

Goal: Identify patterns in the decision process of the model



a) Heatmaps of cluster 1

b) Heatmaps of cluster 2

c) Heatmaps of cluster 29

- Clustering highlights and separates different features

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K. R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature communications*, *10*(1), 1-8.

Analysis of OSM with spectral clustering

Step 1

Extract heatmaps from samples of interest

Step 2

Vectorize heatmaps, unify their size, and reduce the dimensionality (e.g. PCA)

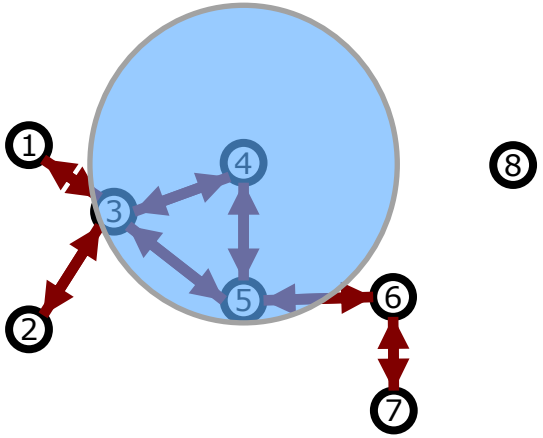
Step 3

Spectral clustering

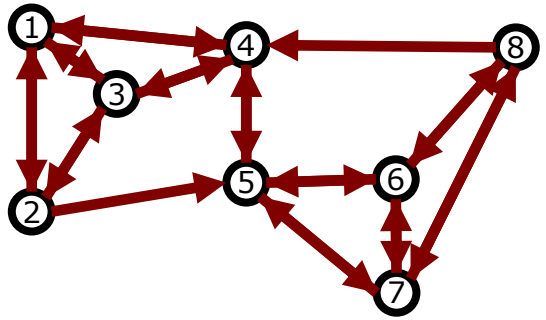
Step 4

Eigengap analysis (differences between two successive eigenvalues)

Spectral clustering

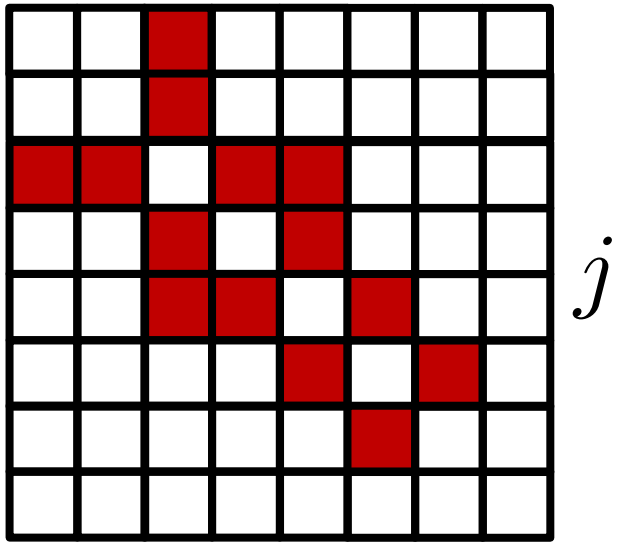


neighbors within radius r



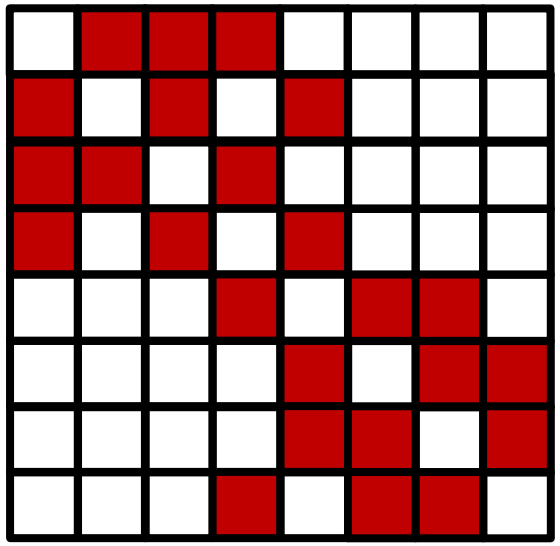
k-nearest neighbors

$$W =$$



j

i



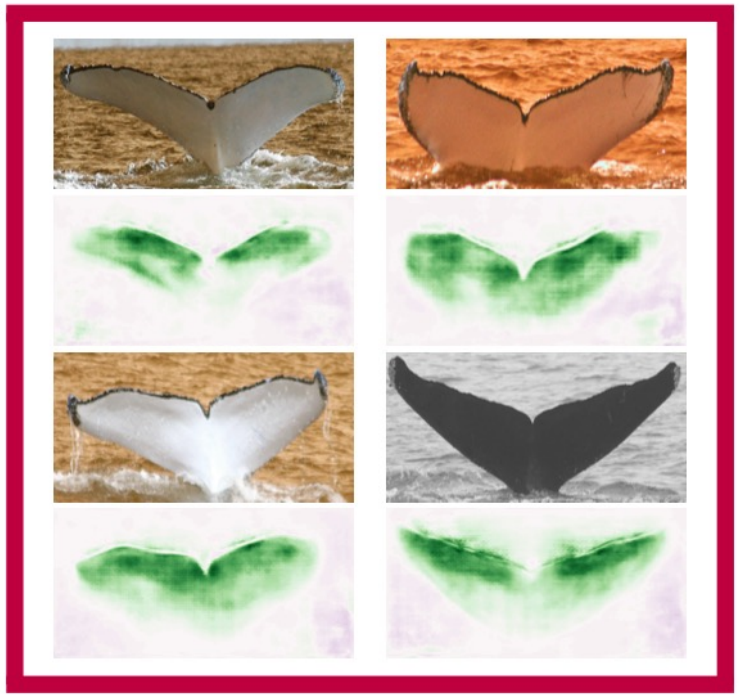
$$D = \mathrm{diag}([d_1, d_2, \ldots, d_N])$$

$$d_i = \sum_j W_{i,j}$$

$$L = D - W$$

eigenvalue decomposition and clustering of eigenvectors

Spectral clustering of OSM

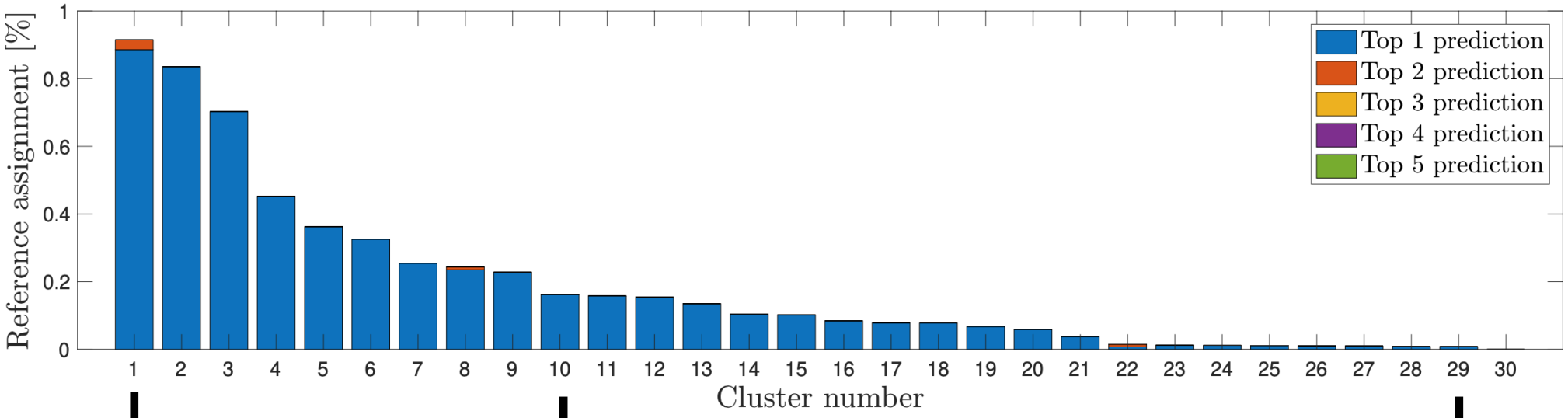


a) Heatmaps of cluster 1

b) Heatmaps of cluster 2

c) Heatmaps of cluster 29

Relation of scores and clusters



90% of heatmaps belong to top 1 predictions

Most heatmaps belong to wrong predictions

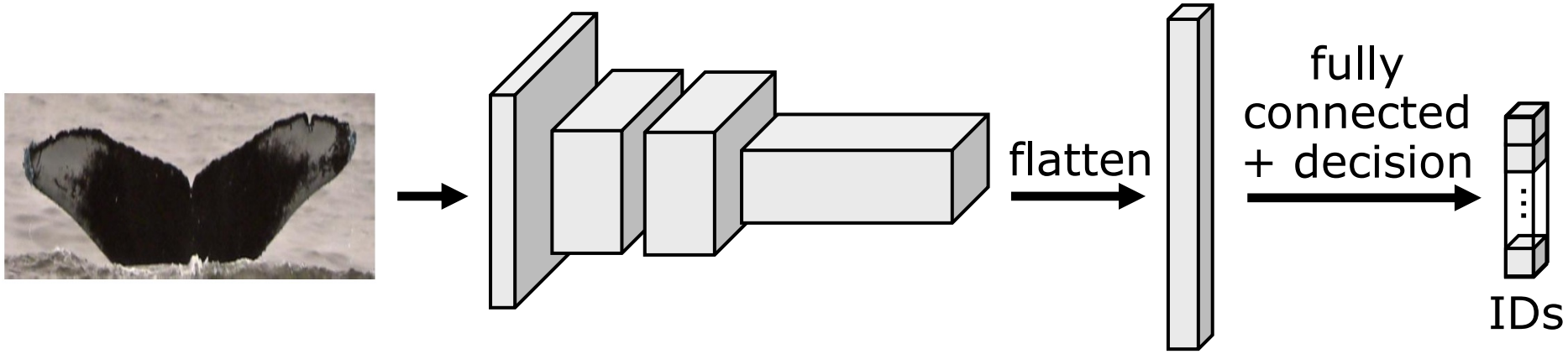
Nearly all heatmaps belong to predictions with small scores for all classes (indicator that the whale is not in the data base)

Insights

- Some clusters can be assigned to high identification rates
  - Specific characteristics are good for identification (e.g., shape of fluke vs. time-variable scars)
- Network looks at the right features (no clever-Hans-effect: right decision for the wrong reasons)
- Real-time evaluation of a photograph (e.g., lighting conditions, viewing angle)

Class activation maps (CAM)

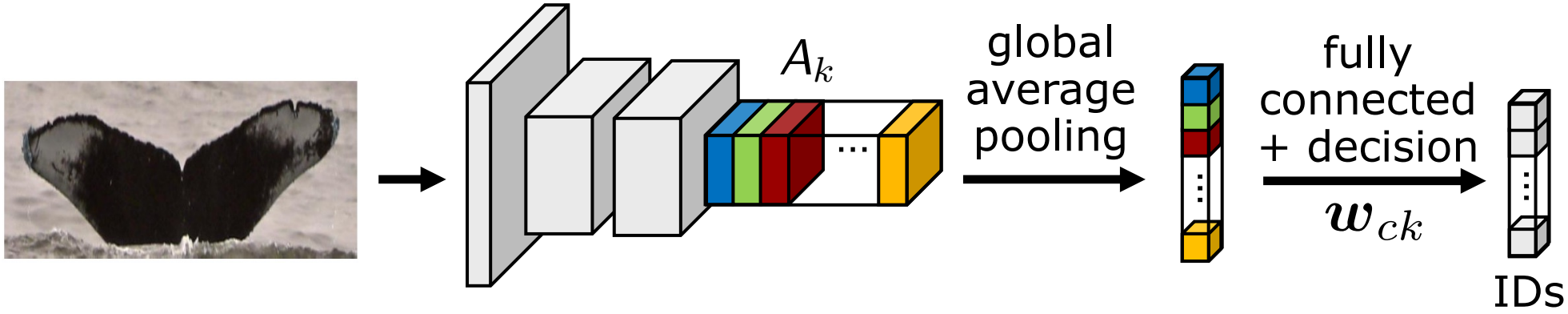
- Specifically designed for CNNs
- **Goal**: indicate the discriminative regions used by the CNNs for classification



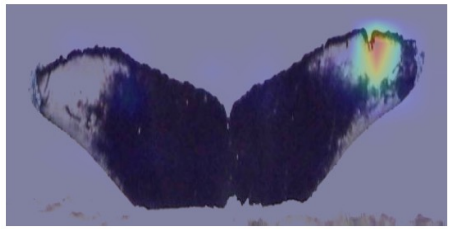
Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2921-2929).

Class activation maps (CAM)

- Specifically designed for CNNs
- **Goal**: indicate the discriminative regions used by the CNNs for classification



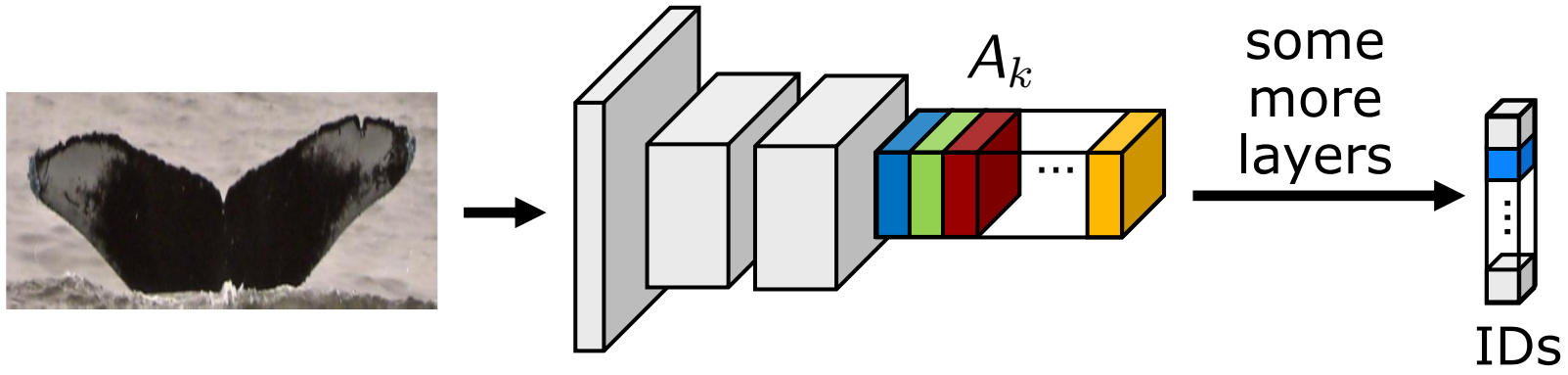
$$\sum_k w_{ck} A_k =$$



Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2921-2929).

Gradient-weighted class activation maps (Grad-CAM)

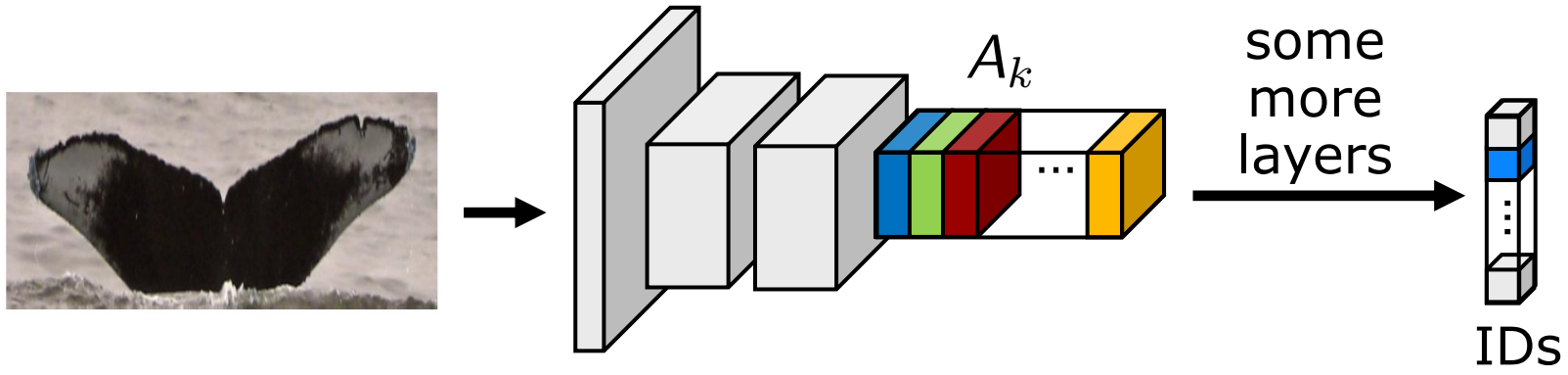
Uses the **gradient** of the learned network to indicate from which part of an image a given convolutional layer takes information



Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2019). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, *128*(2), 336-359.

Gradient-weighted class activation maps (Grad-CAM)

Step 1: Forward propagate the input image through the CNN and obtain the raw score for the class of interest before applying softmax/sigmoid. Set all other scores to zero.

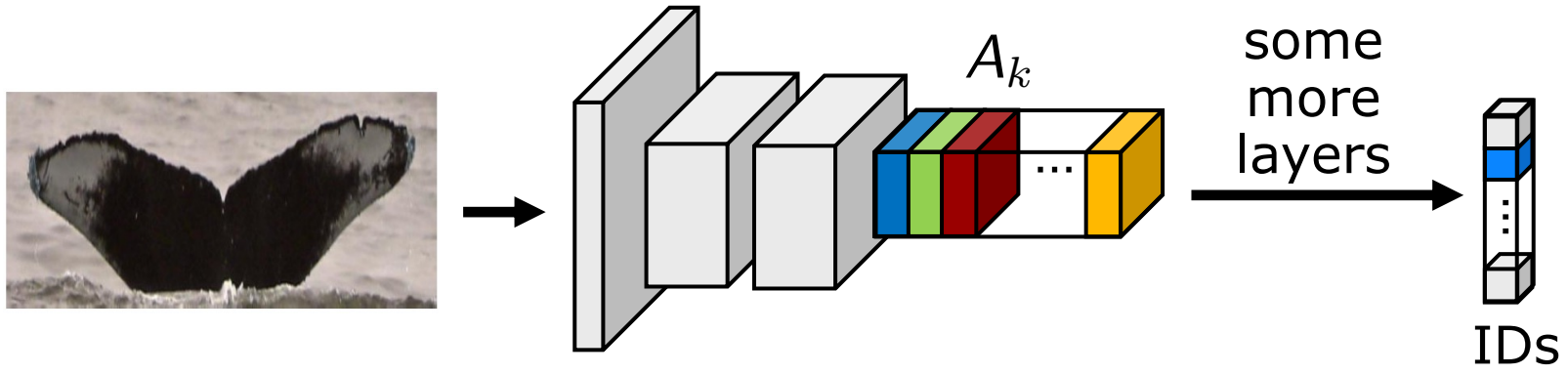


$$\text{ReLU}\left(\sum_k \left(\frac{1}{Z}\sum_i\sum_j \frac{\partial y_c}{\partial A_{ij}}\right) A_k\right)$$

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2019). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, *128*(2), 336-359.

Gradient-weighted class activation maps (Grad-CAM)

Step 2: Backpropagate the gradient of the raw score of a specific whale ID to the convolutional layer you are interested in

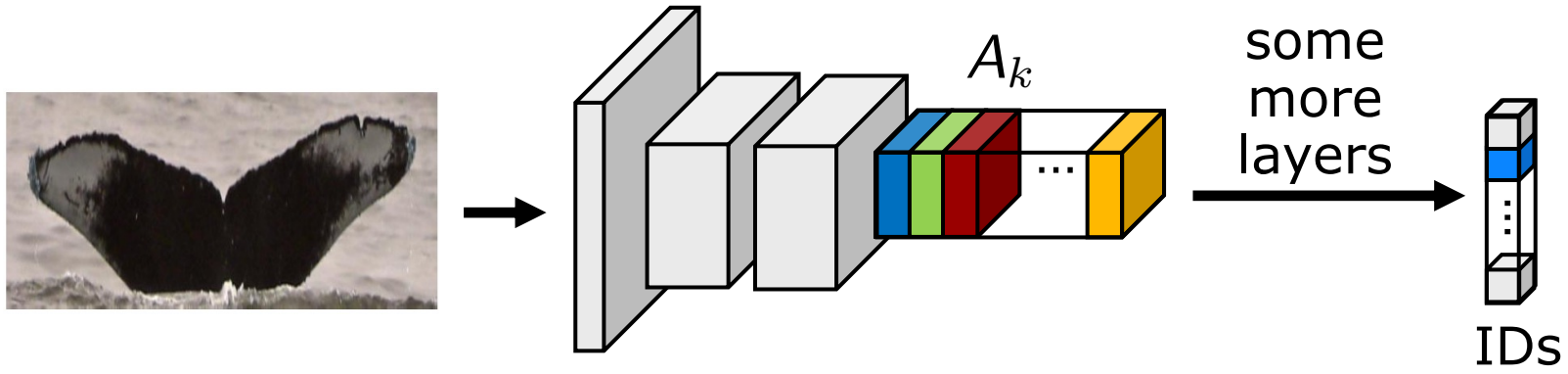


$$\text{ReLU}\left(\sum_k \left(\frac{1}{Z}\sum_i\sum_j \frac{\partial y_c}{\partial A_{ij}}\right) A_k\right)$$

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2019). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, *128*(2), 336-359.

Gradient-weighted class activation maps (Grad-CAM)

Step 3: Perform global average pooling, i.e. gradients are averaged across activation maps ⟶ importance of each activation map for the target class

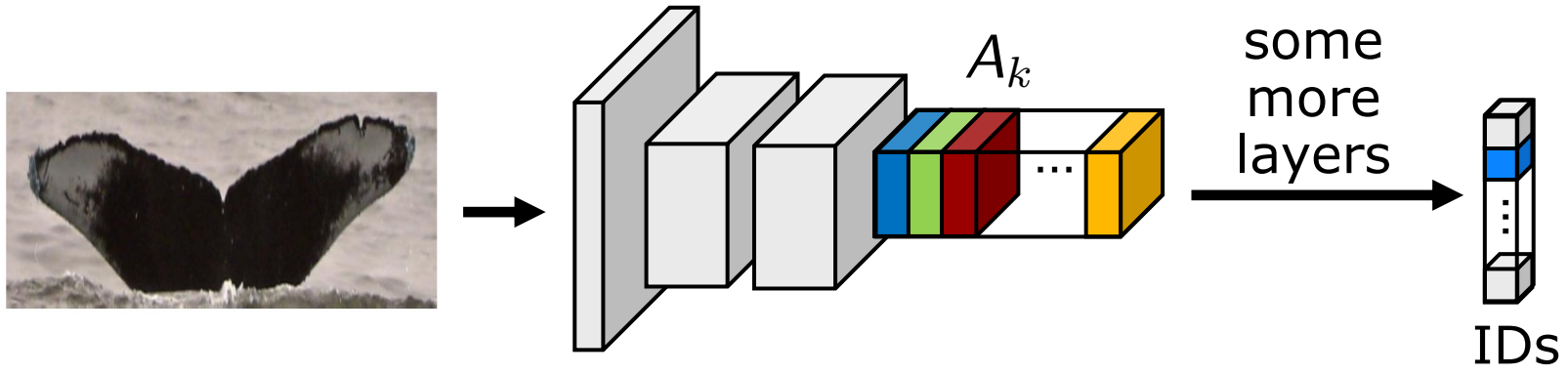


$$\text{ReLU}\left(\sum_k \left(\frac{1}{Z}\sum_i\sum_j \frac{\partial y_c}{\partial A_{ij}}\right) A_k\right)$$

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2019). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, *128*(2), 336-359.

Gradient-weighted class activation maps (Grad-CAM)

Step 4: Average all activation maps, weighted by their importance

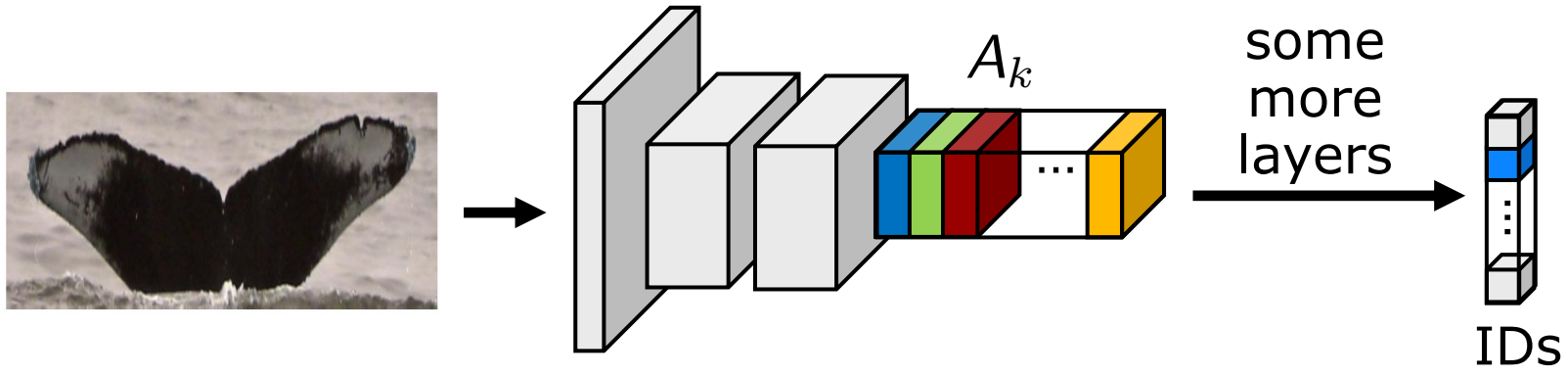


$$\text{ReLU}\left(\sum_k \left(\frac{1}{Z}\sum_i\sum_j \frac{\partial y_c}{\partial A_{ij}}\right) A_k\right)$$

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2019). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, *128*(2), 336-359.

Gradient-weighted class activation maps (Grad-CAM)

Step 6: Apply ReLU to consider only the values with a positive influence



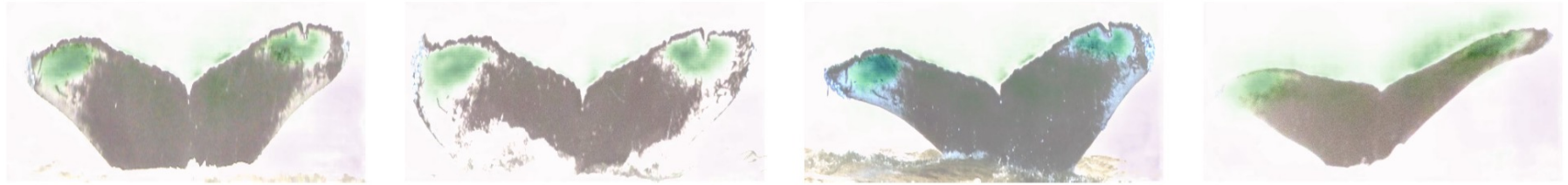
$$\text{ReLU}\left(\sum_k \left(\frac{1}{Z}\sum_i\sum_j \frac{\partial y_c}{\partial A_{ij}}\right) A_k\right)$$

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2019). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, *128*(2), 336-359.

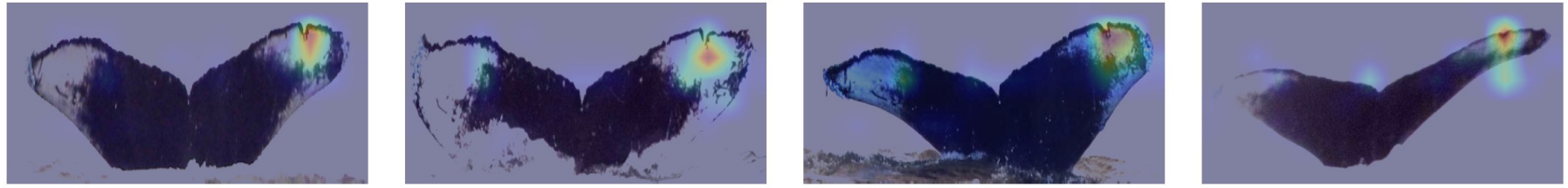
Comparison

Original

Occlusion sensitivity maps



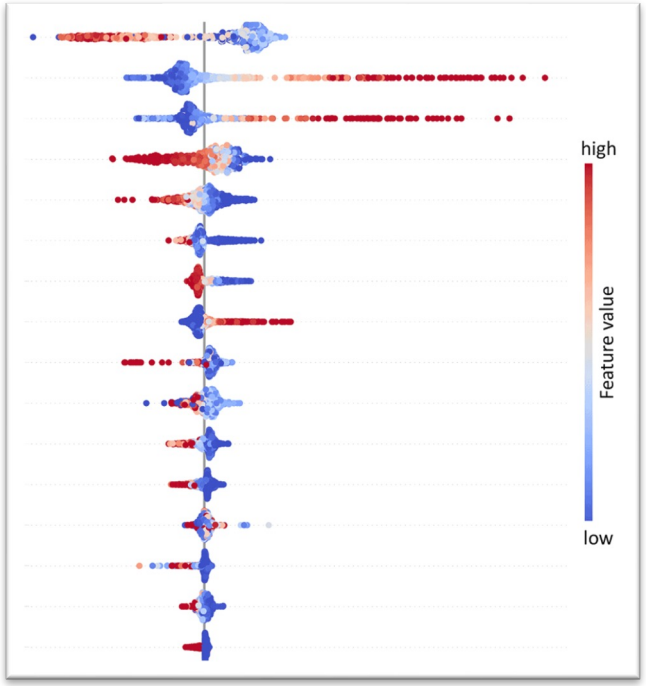
Grad-CAM



Kierdorf, J., Garcke, J., Behley, J., Cheeseman, T., & Roscher, R. (2020). What identifies a whale by its fluke? On the benefit of interpretable machine learning for whale identification. *ISPRS Annals of the Photogrammetry Remote Sensing and Spatial Information Sciences*, *2*, 1005-1012.

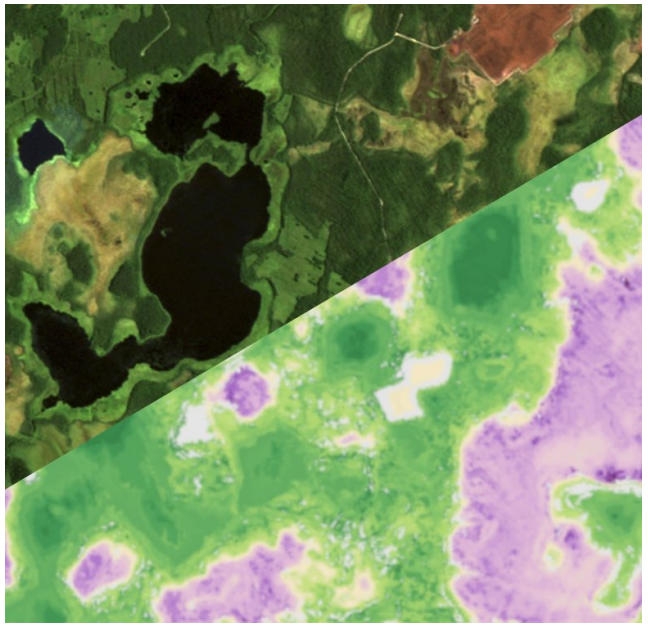
Reasons to seek explanations

Justify decisions



Example: whale monitoring and ozone value estimation

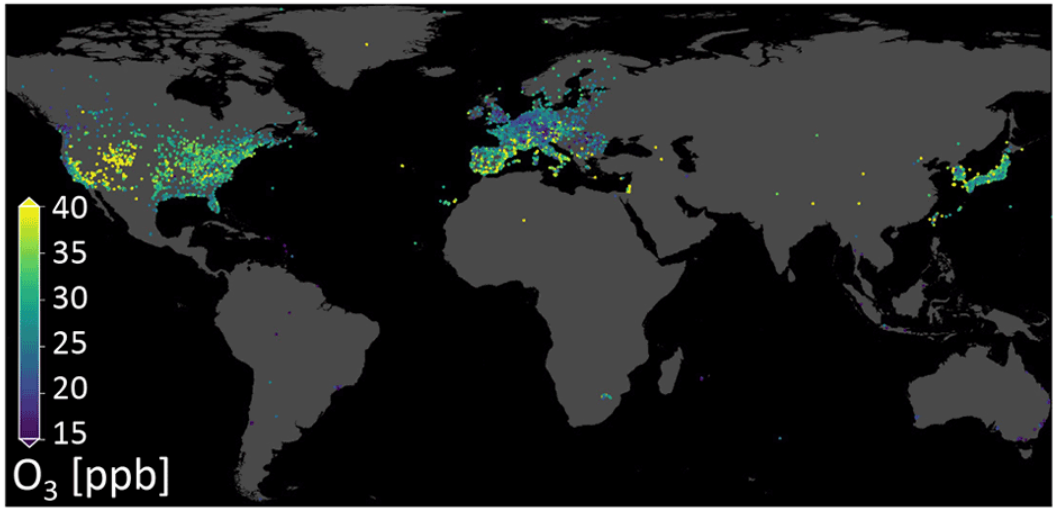
Discover knowledge



Example: understanding wilderness

Mapping of tropospheric ozone

- Toxic greenhouse gas with highly variable spatial distribution
- So far, the relation between geospatial data and ozone metrics is not well understood



Betancourt, C., Stomberg, T. T., Edrich, A. K., Patnala, A., Schultz, M. G., Roscher, R., ... & Stadtler, S. (2022). Global, high-resolution mapping of tropospheric ozone–explainable machine learning and impact of uncertainties. *Geoscientific Model Development*, *15*(11), 4331-4354.

Betancourt, C., Stomberg, T., Roscher, R., Schultz, M. G., & Stadtler, S. (2021). AQ-Bench: a benchmark dataset for machine learning on global air quality metrics. *Earth System Science Data*, *13*(6), 3013-3033.

Shapley values

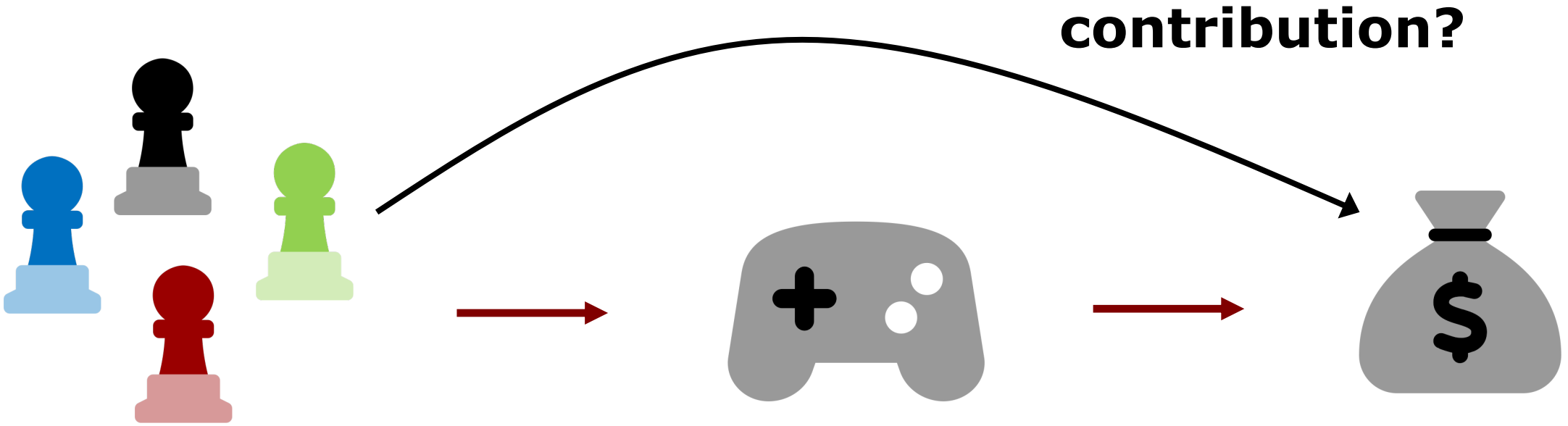
- Local interpretation of the output by means of the input (post-hoc, model-agnostic)
- Theoretical basis from game theory

Question

How much has each feature value contributed to the prediction compared to the average prediction?

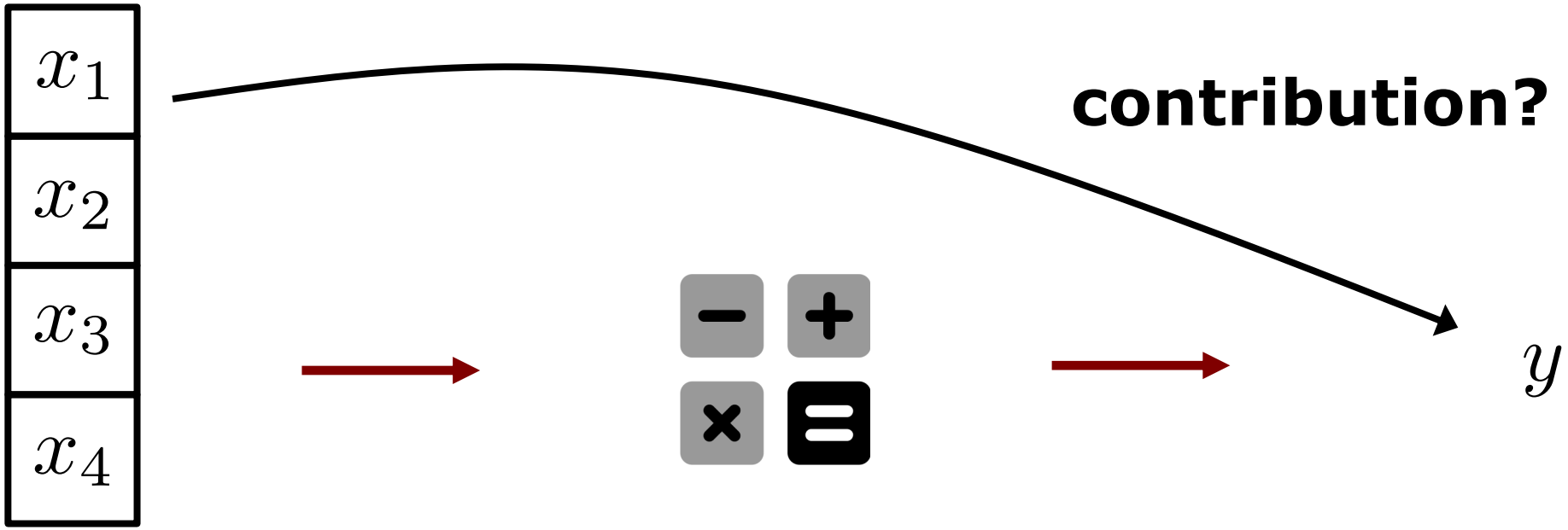
Shapley, Lloyd S. "A value for n-person games." Contributions to the Theory of Games 2.28 (1953): 307-317.

Game theory



- Assignment of a payout to each player based on the contribution

Data science



set of features with given values (players in coalition)

prediction task for a single instance (game)

output (overall payout)

Shapley values

Shapley values

(Weighted) average marginal contribution of a feature value across all possible coalitions

Goal

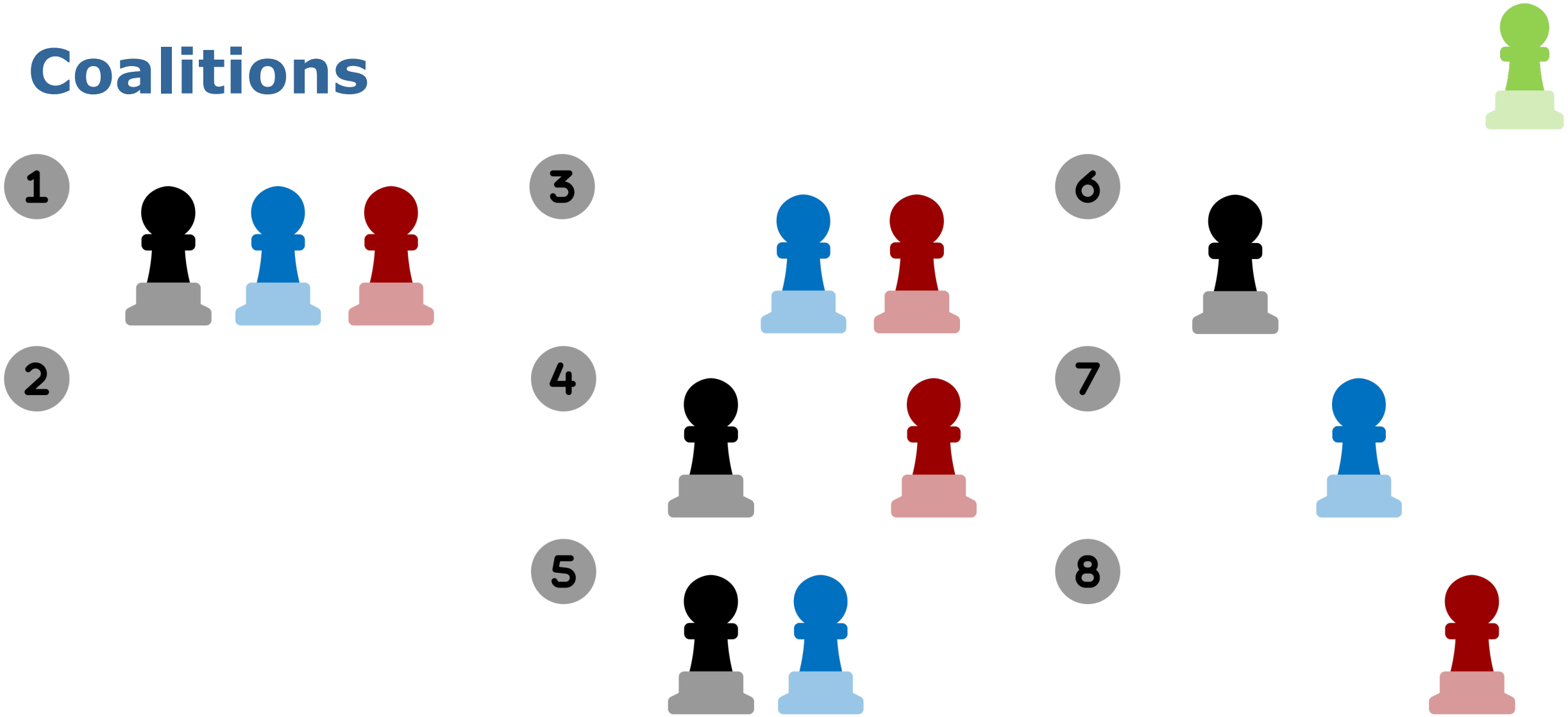
Explain the difference between the actual prediction with specific feature values and the average prediction

Marginal distribution

- Effect of single variables is eliminated (”marginalized out”)
- Marginal distributions add up to 100%

	pizza	burrito	sushi	total
player 1	1	2	7	10
player 2	3	3	4	10
player 3	4	0	6	10
player 4	6	1	3	10
total	14 (35%)	6 (15%)	20 (50%)	40

Coalitions



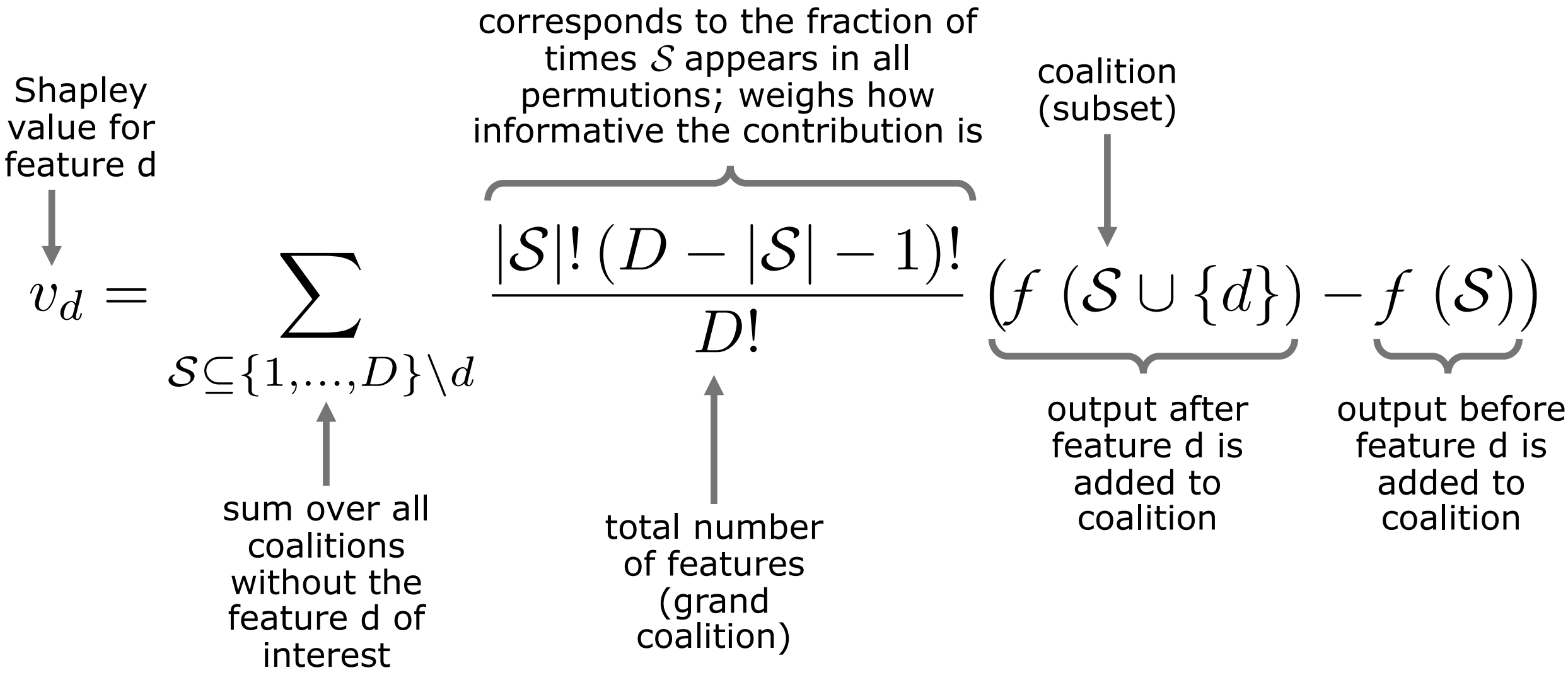
- 8 coalitions with given feature values to which the green player contribute

Challenge

Machine learning model cannot be evaluated with less than D features

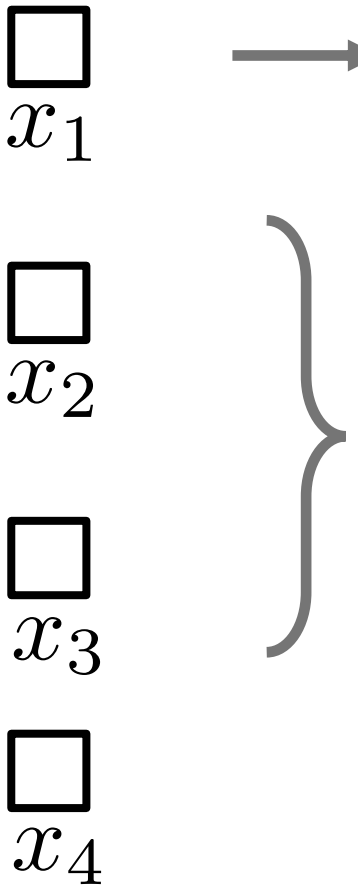
- Training separate models for each number of features is computationally too expensive
- Alternative
  - Feature values in coalition are fixed
  - Feature values outside coalition are randomly sampled from the dataset (treated as unknown)
  - Marginalizing out the effect of other features outside the coalitions

Shapley value



https://edden-gerber.github.io/shapley-part-1/

Contribution of one feature value to one coalition



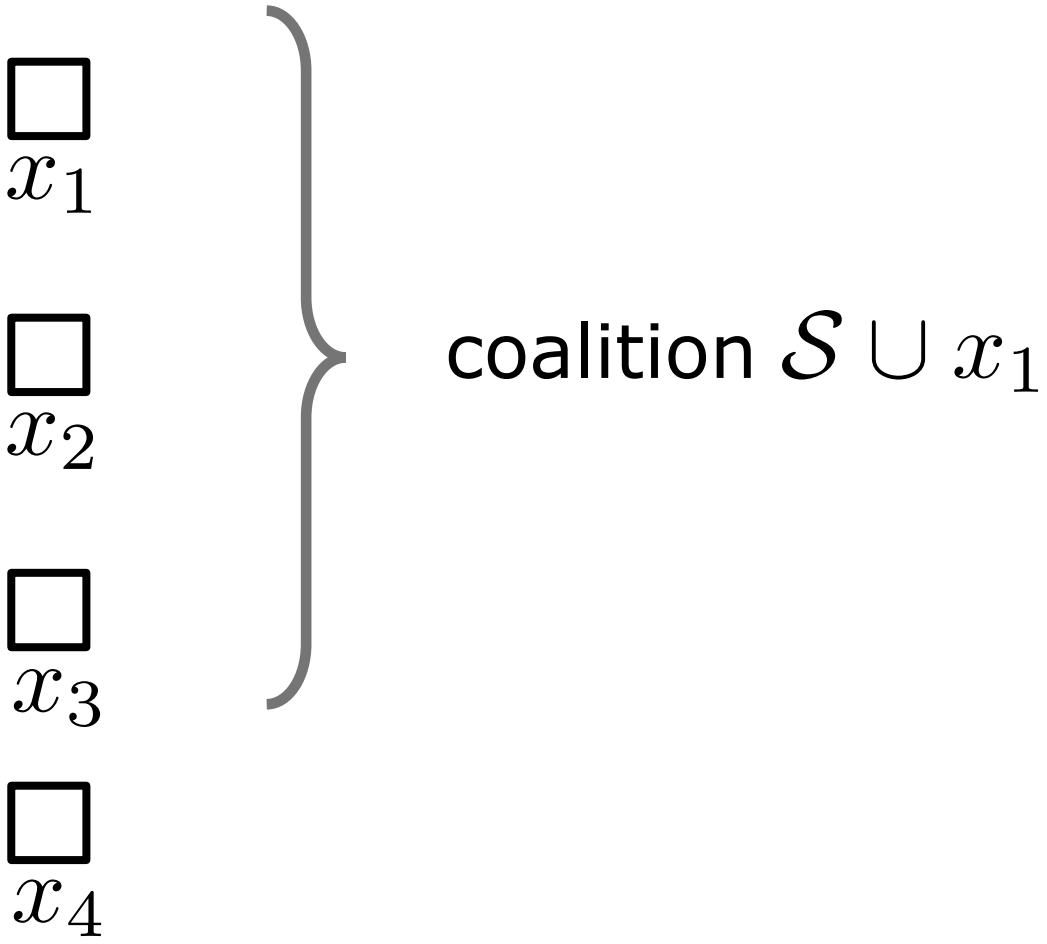
contribution?

coalition $\mathcal{S}$

Step 1

Add x_1 to the coalition

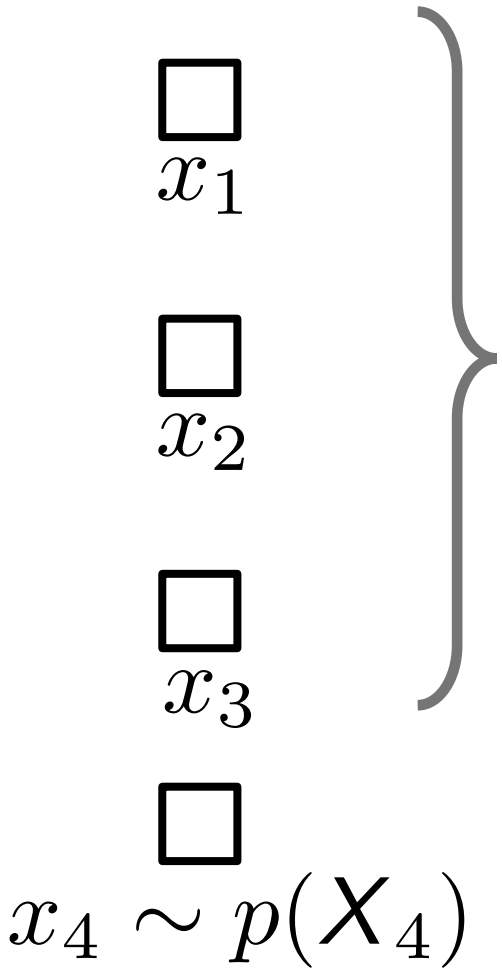
Contribution of one feature value to one coalition



Step 2

Randomly sample a value for x_4 from the dataset and compute the output

Contribution of one feature value to one coalition

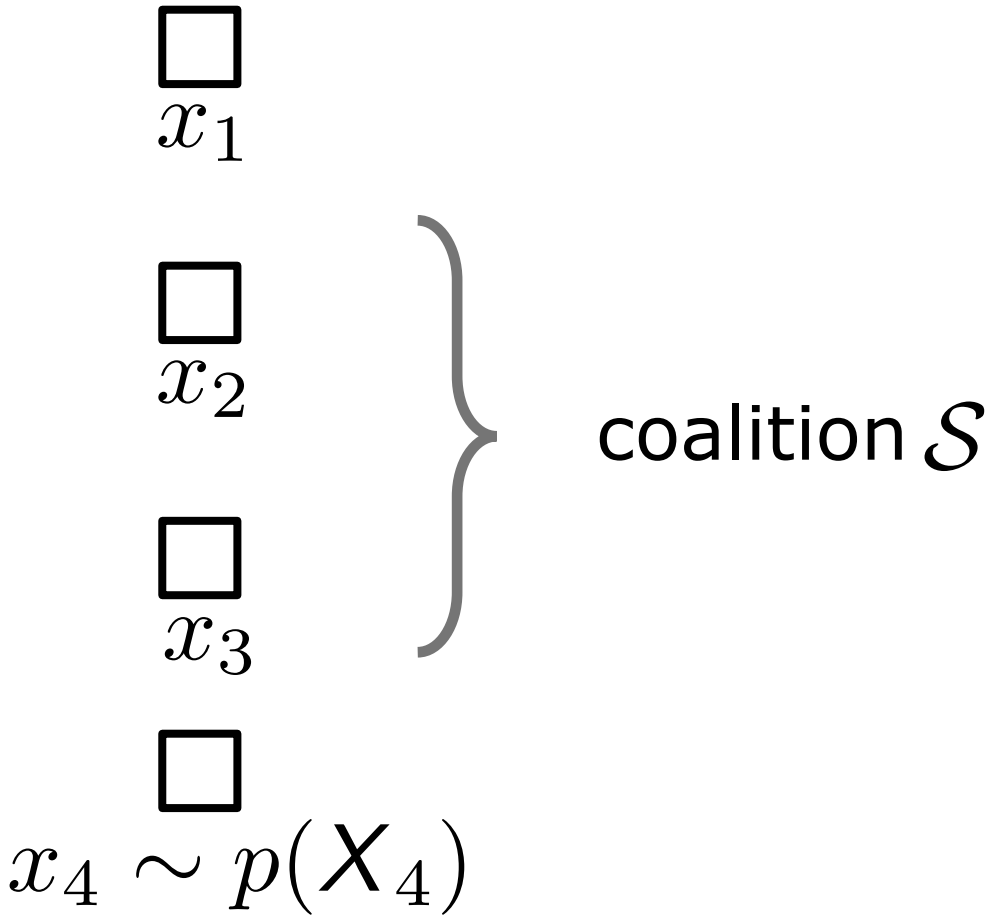


coalition $\mathcal{S} \cup x_1$

Step 3

Remove x_1 from the coalition

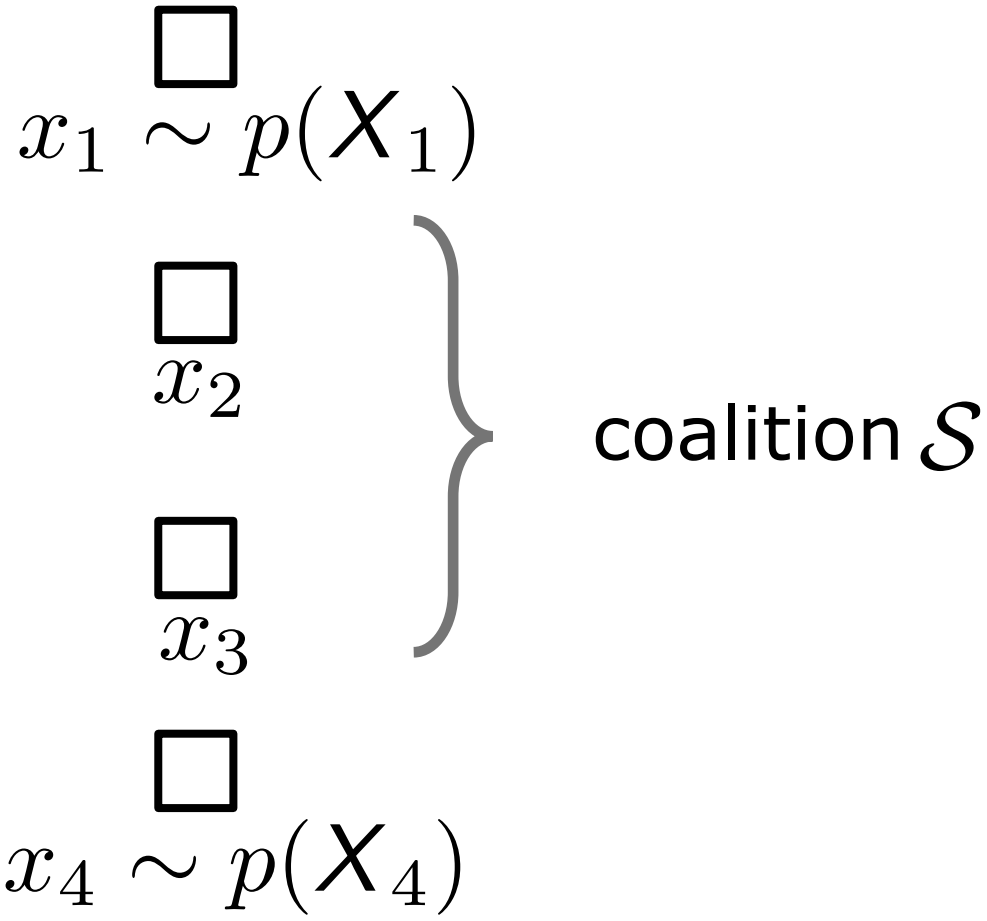
Contribution of one feature value to one coalition



Step 4

Randomly sample a value for x_1 from the dataset and compute the output

Contribution of one feature value to one coalition



Step 5

Compute the contribution of x_1 by comparing the output when the feature is in coalition and when it is not in coalation

- Repeat sampling for feature values outside the coalition

Shapley value

$$v_d = \sum_{\mathcal{S} \subseteq \{1,\ldots,D\} \setminus d} \frac{|\mathcal{S}|!\,(D - |\mathcal{S}| - 1)!}{D!} \left(f\left(\mathcal{S} \cup \{d\}\right) - f\left(\mathcal{S}\right)\right)$$

- (Weighted) average contribution of a feature value to the prediction in different coalitions
- Explains the difference between the actual prediction with specific feature values and the average prediction
- It is not the difference between the actual prediction and the prediction after it was removed from the model training

https://edden-gerber.github.io/shapley-part-1/

Properties

Efficiency

The sum of single Shapley values equals the value of the grand coalition.

Symmetry

The Shapley values of two feature values should be the same if they contribute equally to all coalitions.

Dummy

A feature value that does not change the predicted value should have Shapley value of 0.

Additivity

For predictions that arose from added predictions from individual models, the Shapley values of the individual models should also add up to the Shapley value that arise from the overall model.

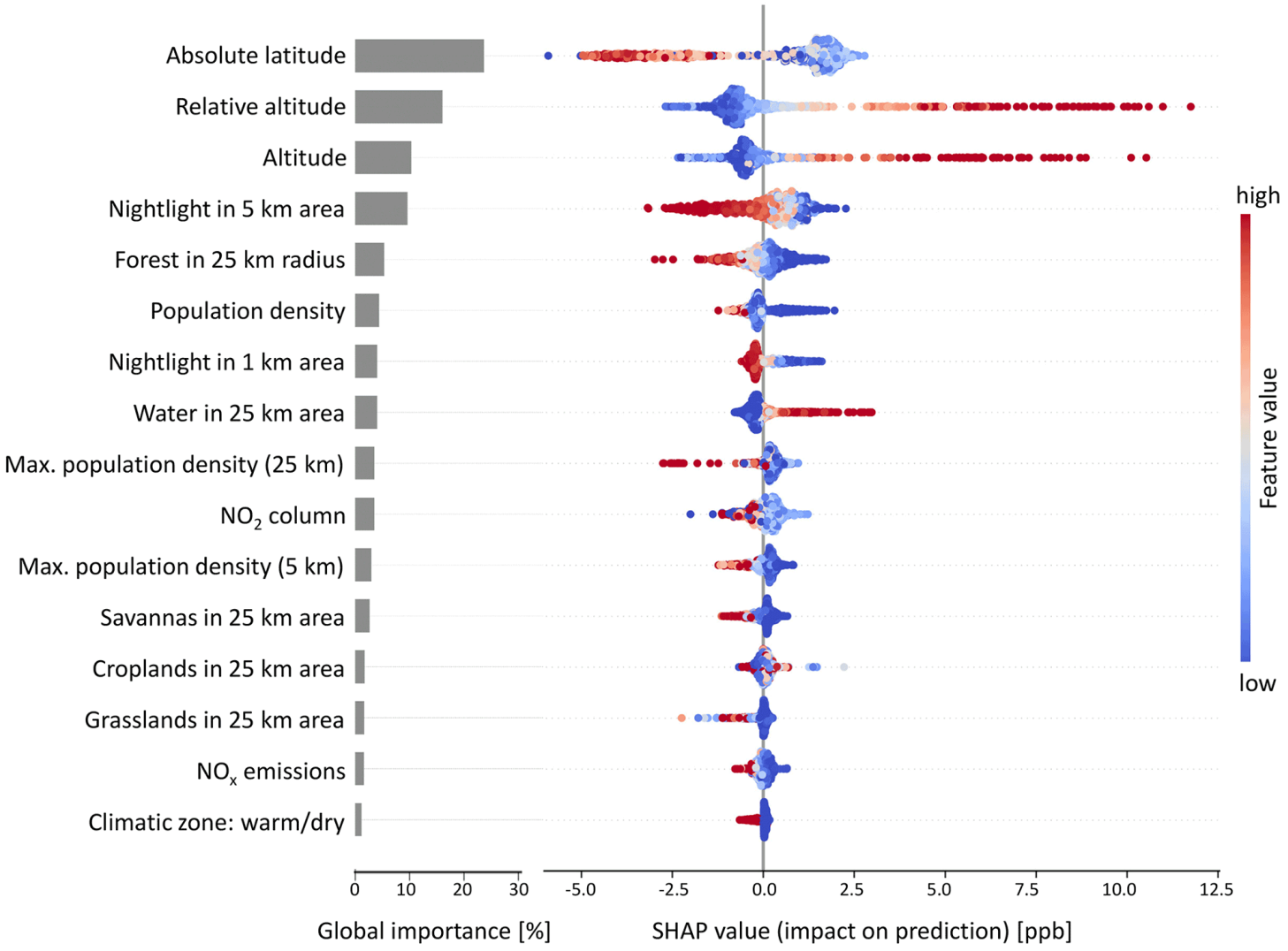
Advantages and disadvantages

- Most of the interpretation methods don’t fulfill all 4 properties
- Shapley values can take into account the whole data set or just a subset (useful for analysis)

- Oftentimes misinterpreted
- Only reliable for datasets with a lot of features
- Access to the dataset is needed
- Computationally expensive
  - Efficient implementation exist, mostly tailored to specific model types/architectures (see SHAP - Shapley additive explanations)

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Ozone value prediction with random forests



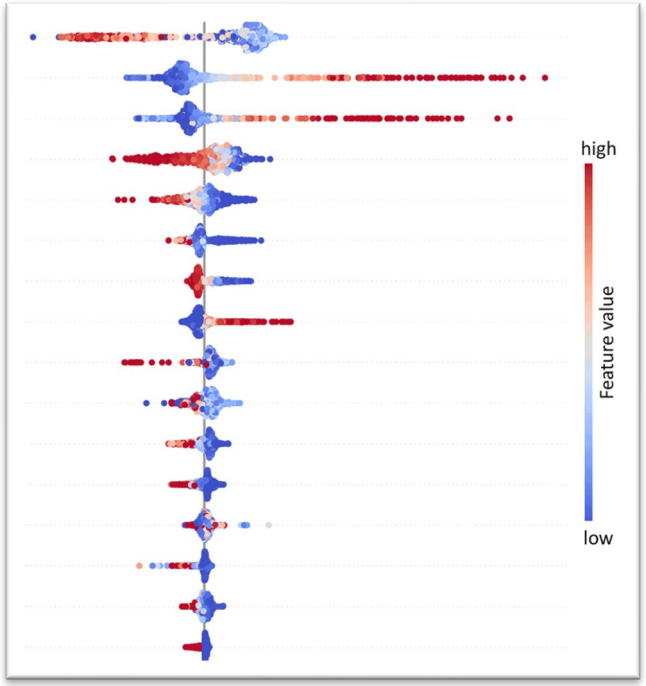
Summary plot

- Global importance: averaged sum of the absolute Shapley values
- Single predictions illustrated in a beeswarm plot

Betancourt, C., Stomberg, T. T., Edrich, A. K., Patnala, A., Schultz, M. G., Roscher, R., ... & Stadtler, S. (2022). Global, high-resolution mapping of tropospheric ozone–explainable machine learning and impact of uncertainties. *Geoscientific Model Development*, *15*(11), 4331-4354.

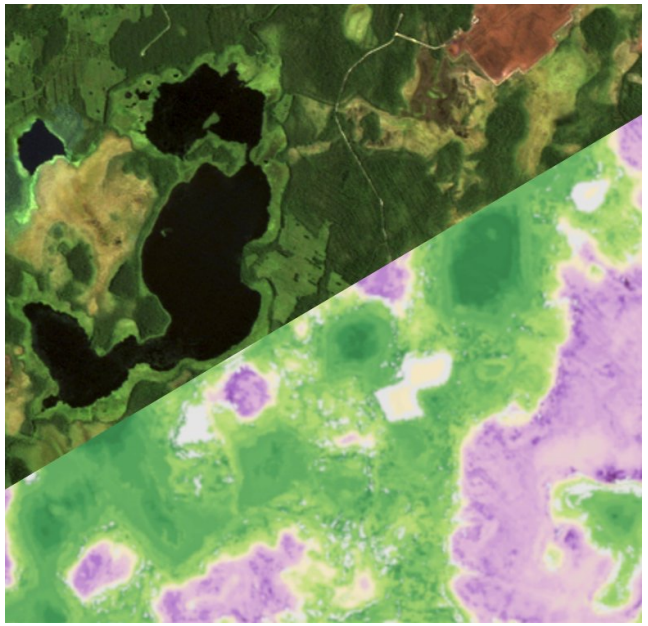
Reasons to seek explanations

Justify decisions



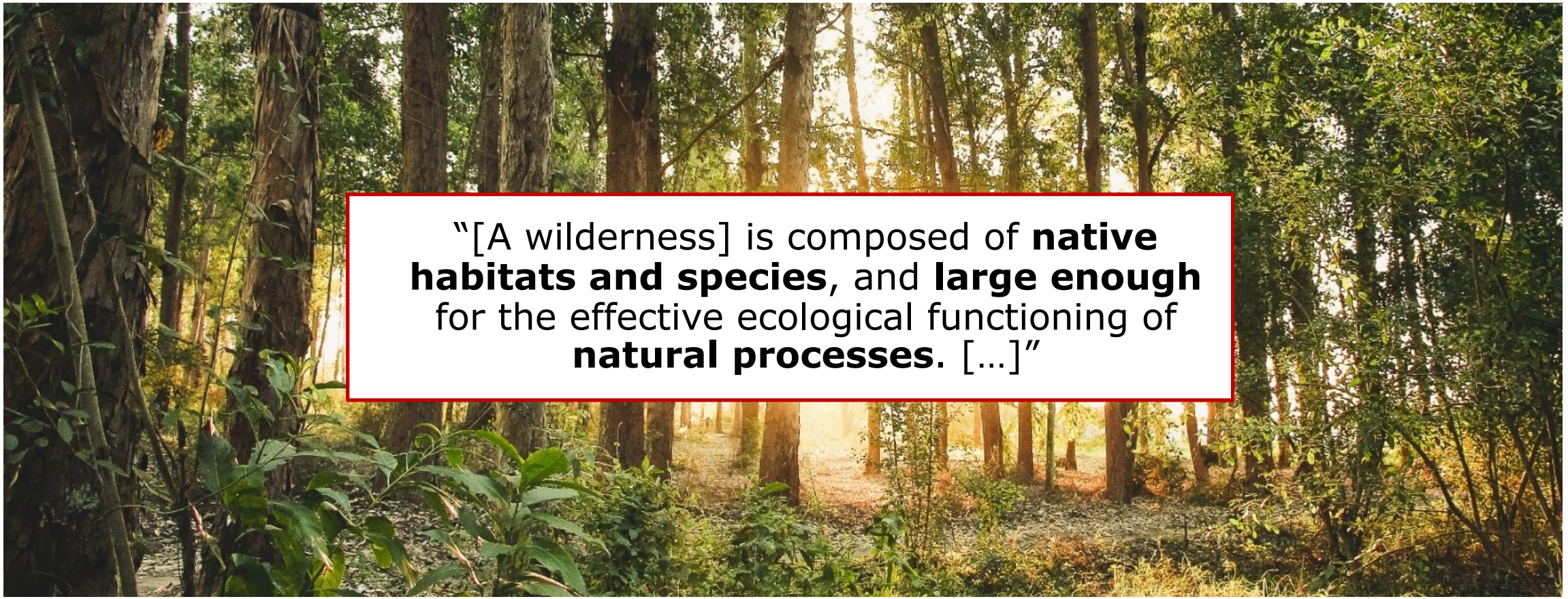
Example: whale monitoring and ozone value estimation

Discover knowledge



Example: understanding wilderness

Discover wilderness characteristics



No existing definition that can be used for machine learning

- **Discover characteristics** of wilderness to deepen our understanding about the land cover class so that it is useful for mapping

Study site

Fennoscandia (Norway, Sweden, Finland)

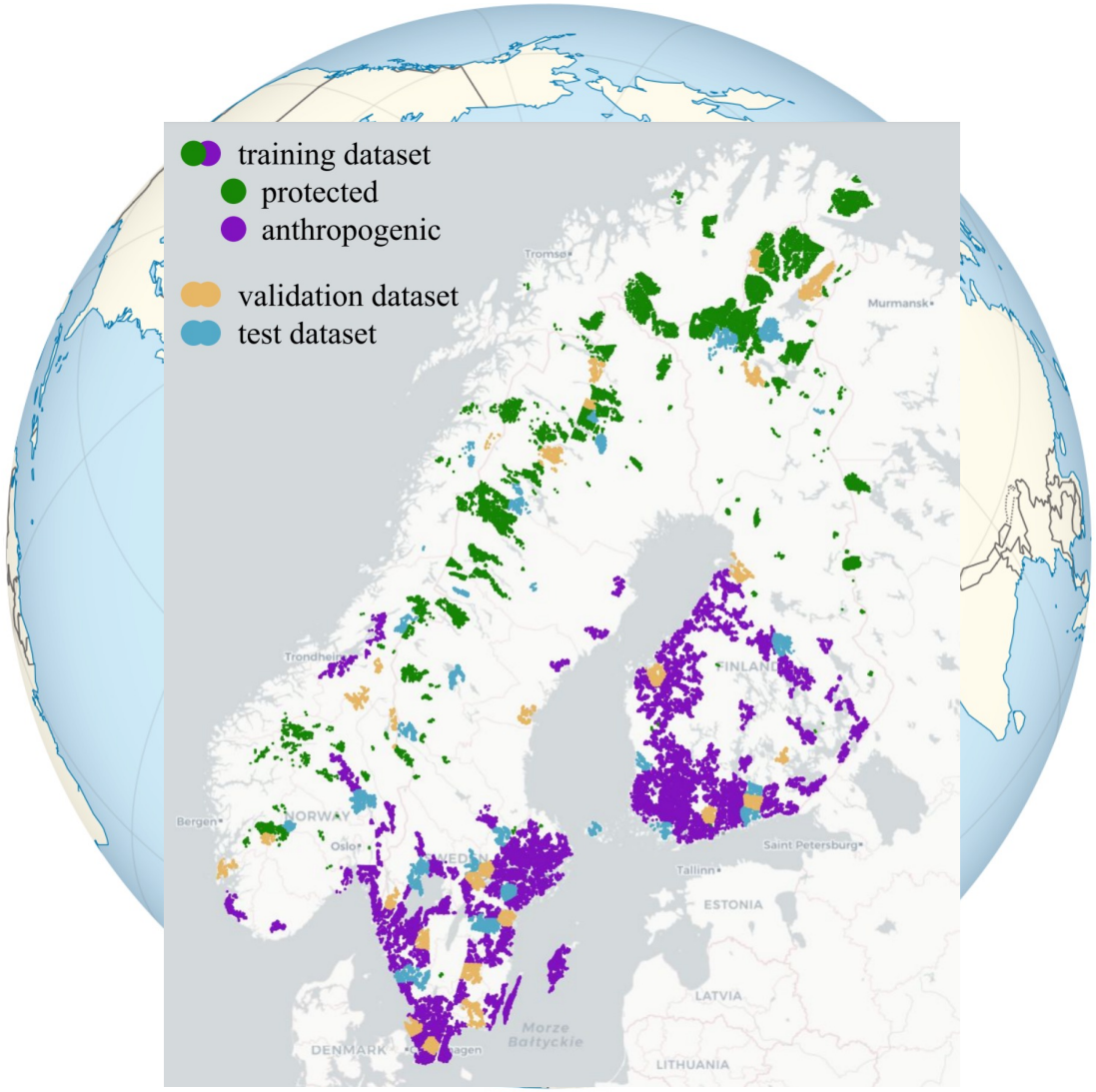
Study site

World Database on Protected Areas (WDPA)

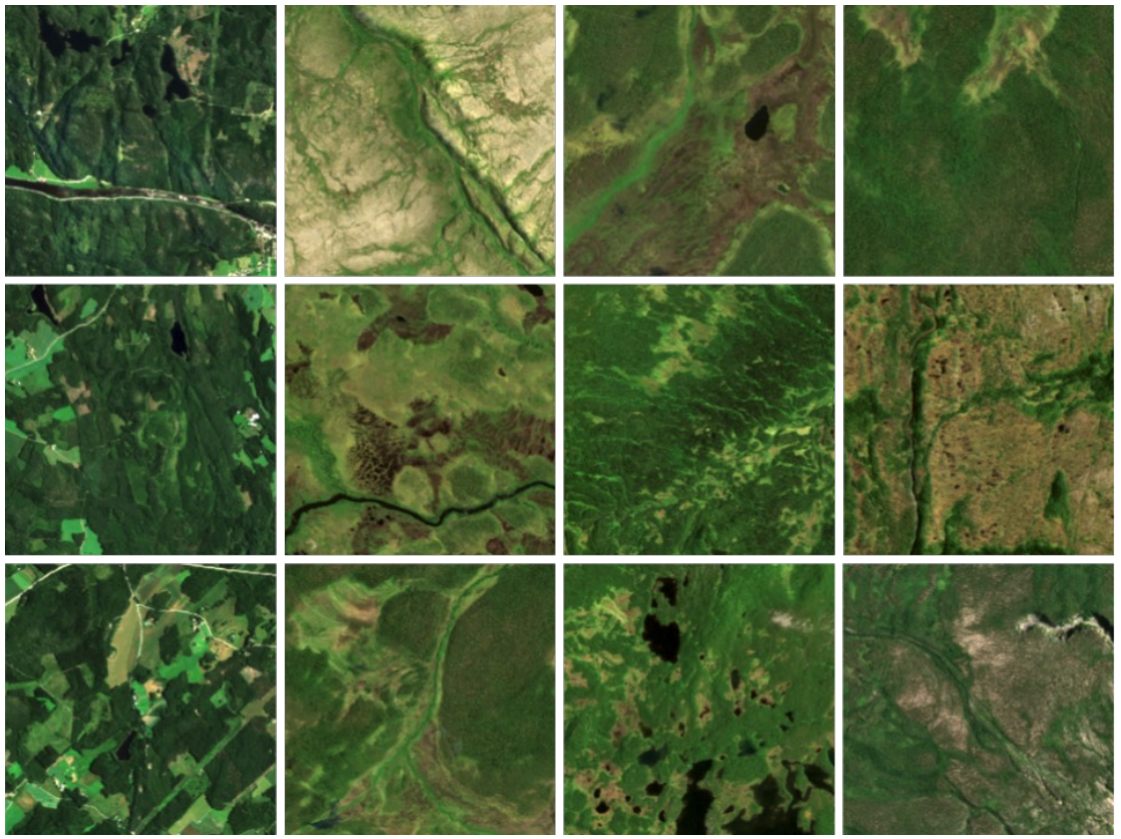
vs.

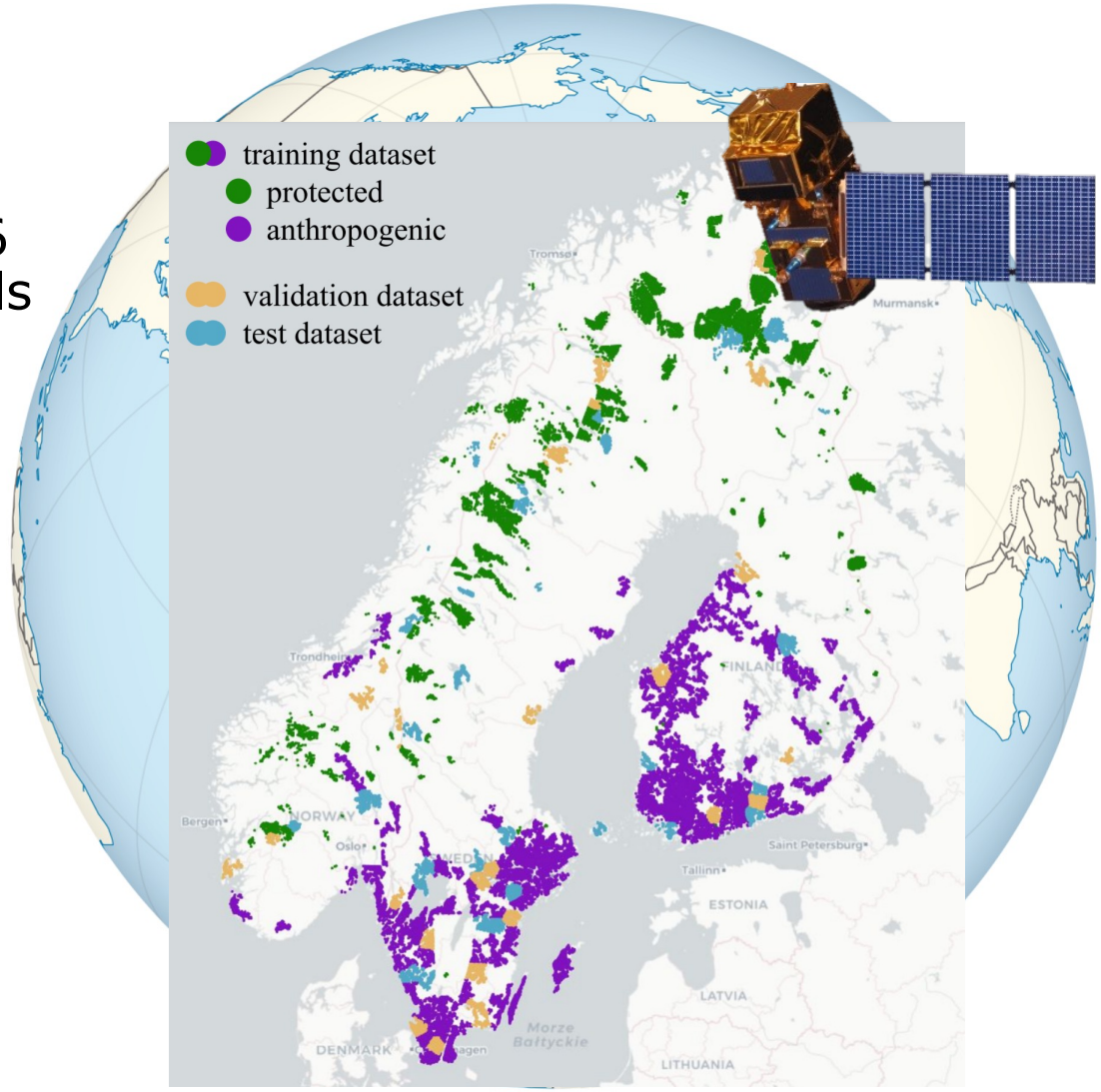
Anthropogenic Areas

(artificial and agricultural surfaces)

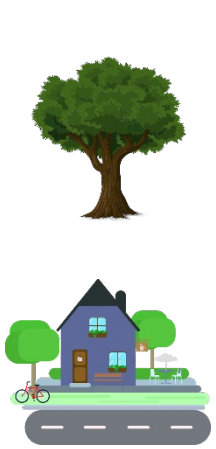


Sentinel-2 data

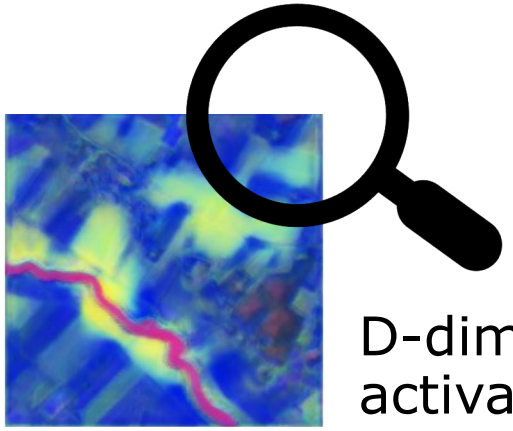




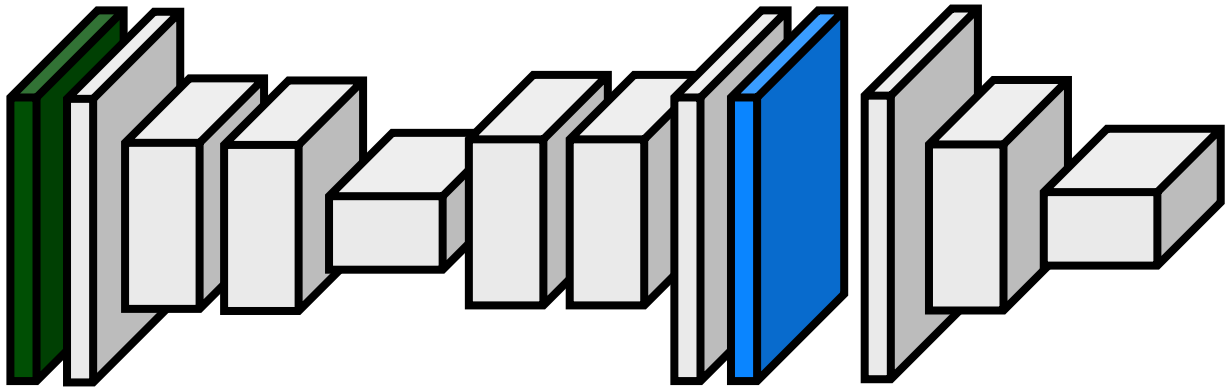
Conceptual framework



multispectral Sentinel-2 images



D-dimensional activation maps



U-net

CNN+FCN

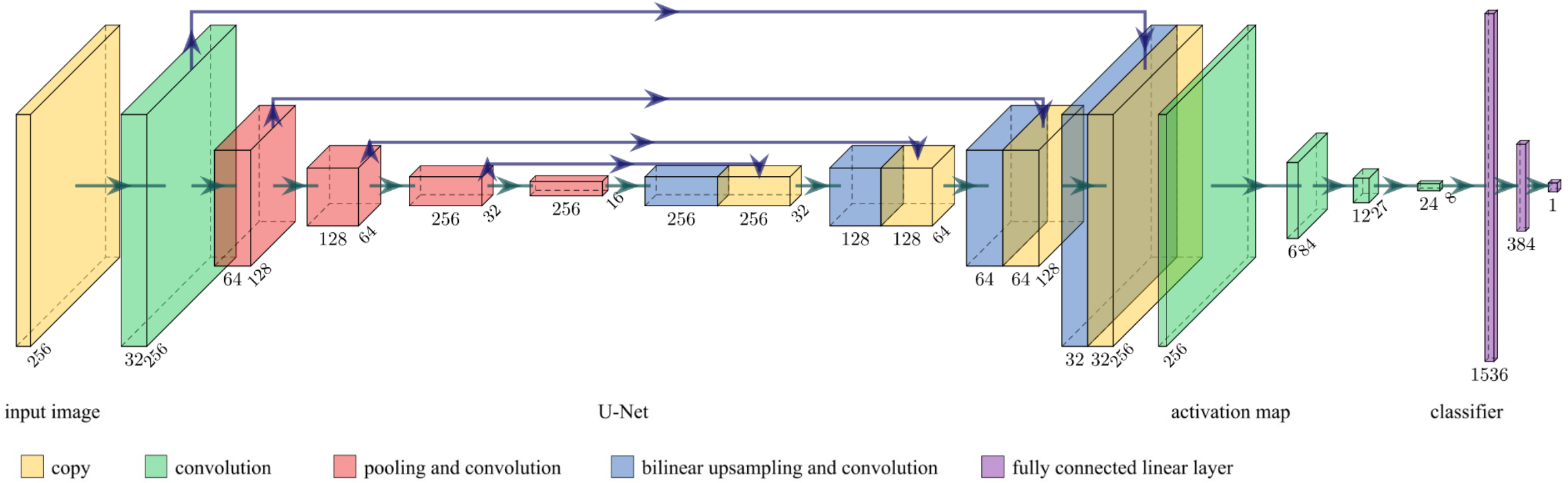
wilderness

non-wilderness

Stomberg, T., Weber, I., Schmitt, M., & Roscher, R. (2021). jUngle-Net: Using explainable machine learning to gain new insights into the appearance of wilderness in satellite imagery. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *3*, 317-324.

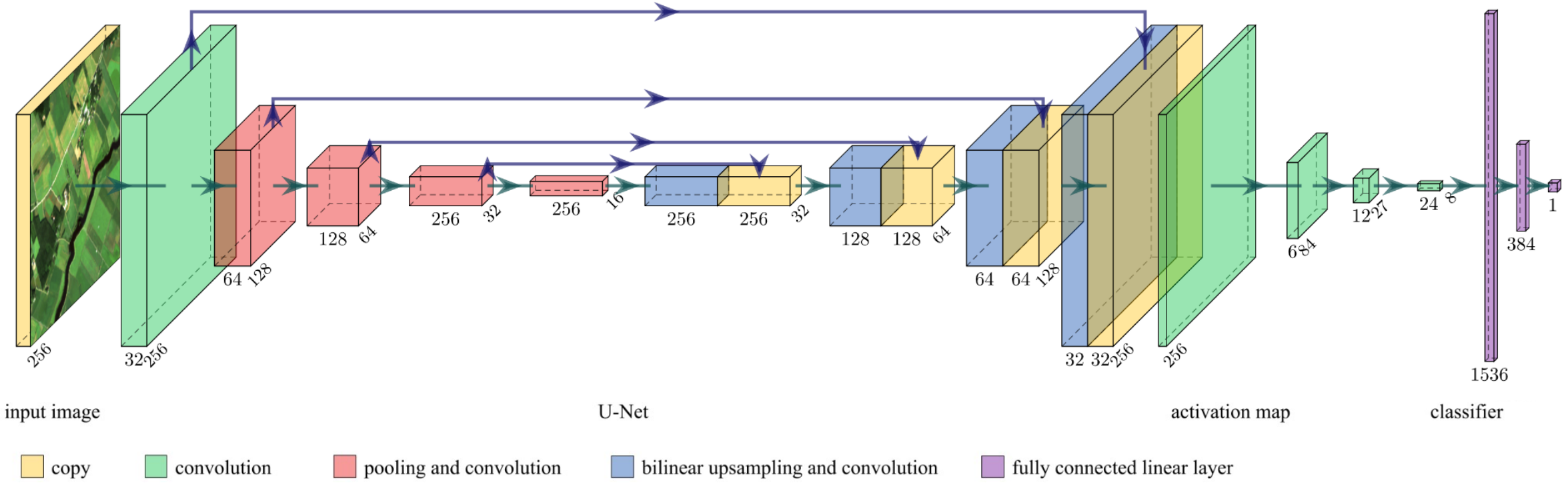
Stomberg, T. T., Stone, T., Leonhardt, J., & Roscher, R. (2022). Exploring Wilderness Using Explainable Machine Learning in Satellite Imagery. *arXiv preprint arXiv:2203.00379*.

Neural network architecture



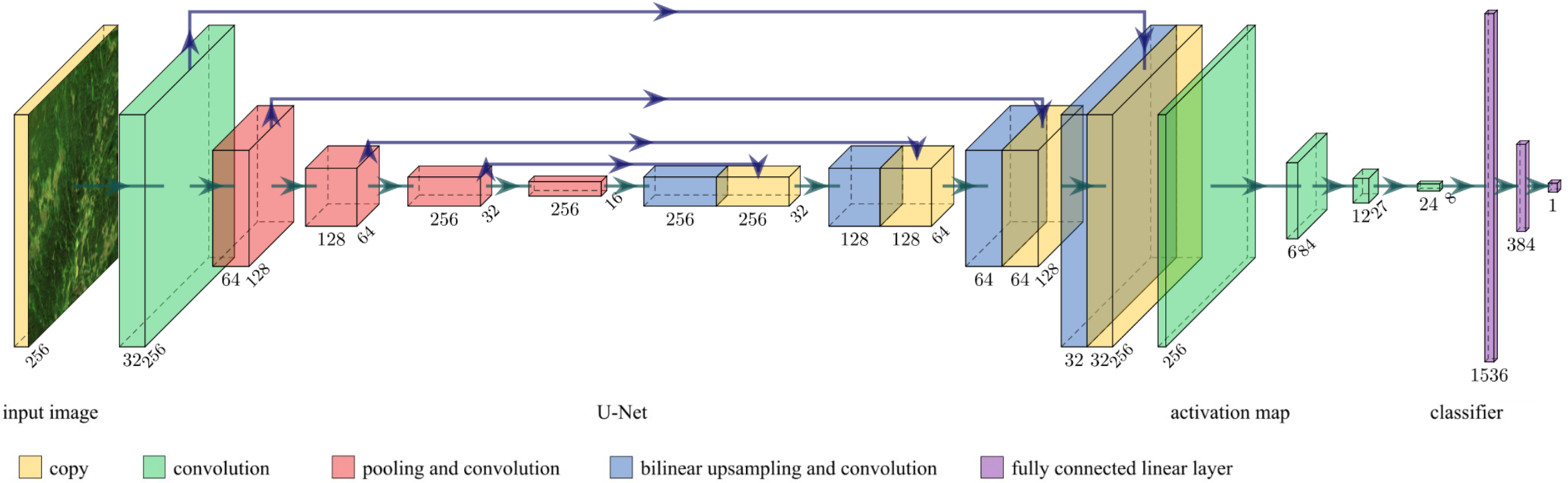
Neural network architecture

anthropogenic



Neural network architecture

wilderness area



Classification results

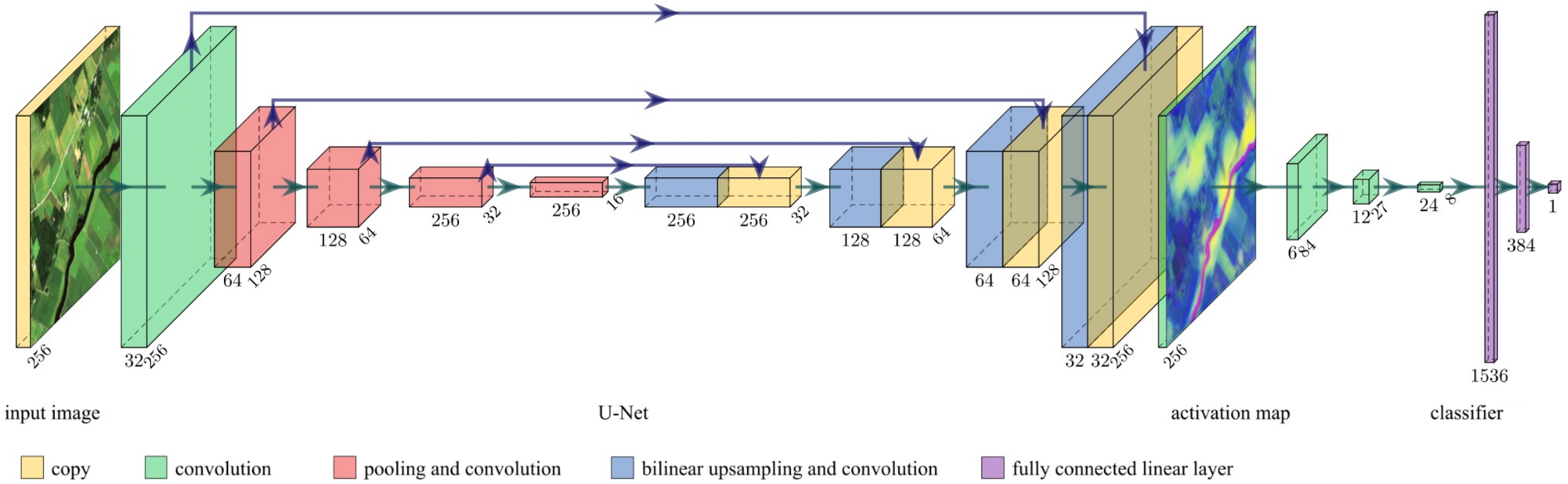
Accuracies

- train: 99,70%
- validation: 99,96%
- test: 99,70%

➢ No uncertainties in the decision process

➢ Uncertainties in the decision process would lead to uncertainties in the interpretations and explanations

Activation maps

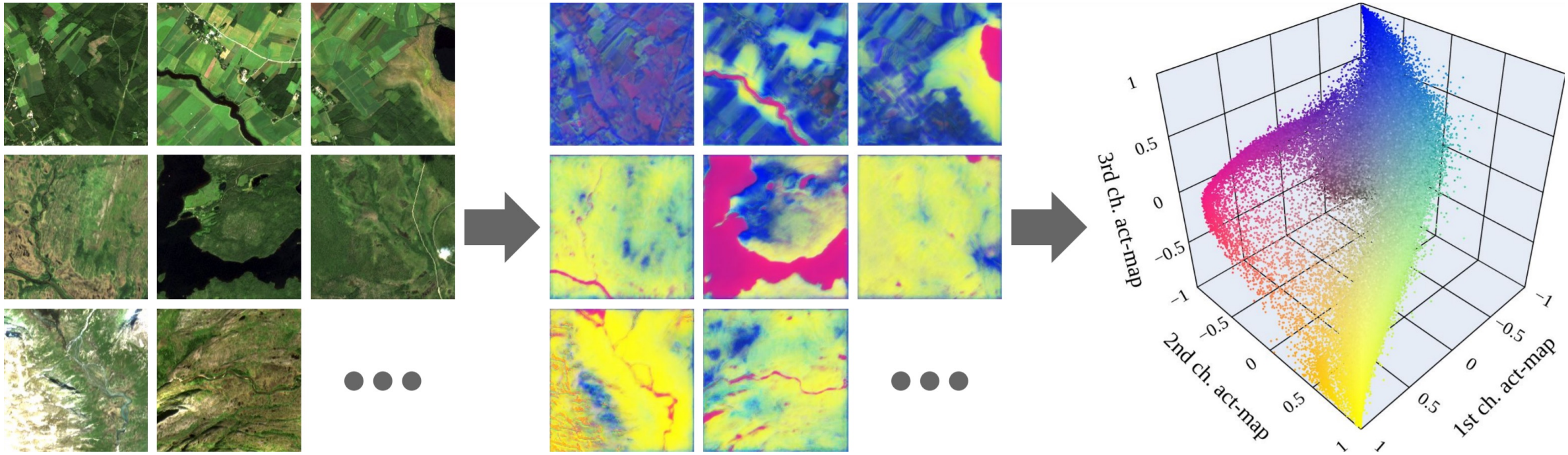


Activation space

19.123 training samples

activation maps (3 channels)

activation space

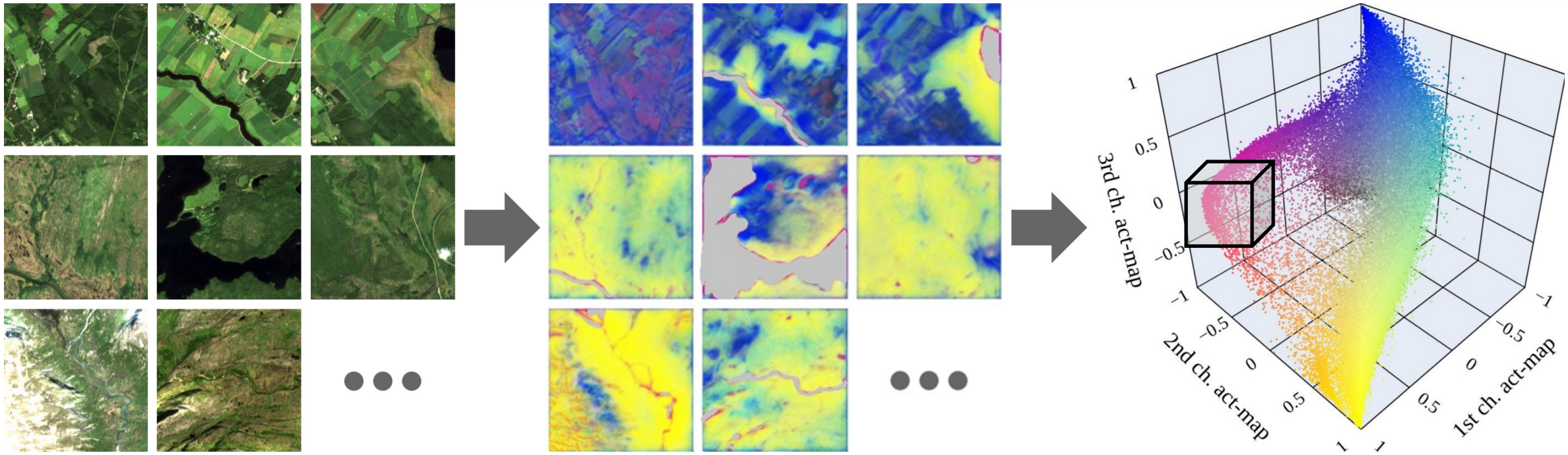


Activation space occlusion sensitivity

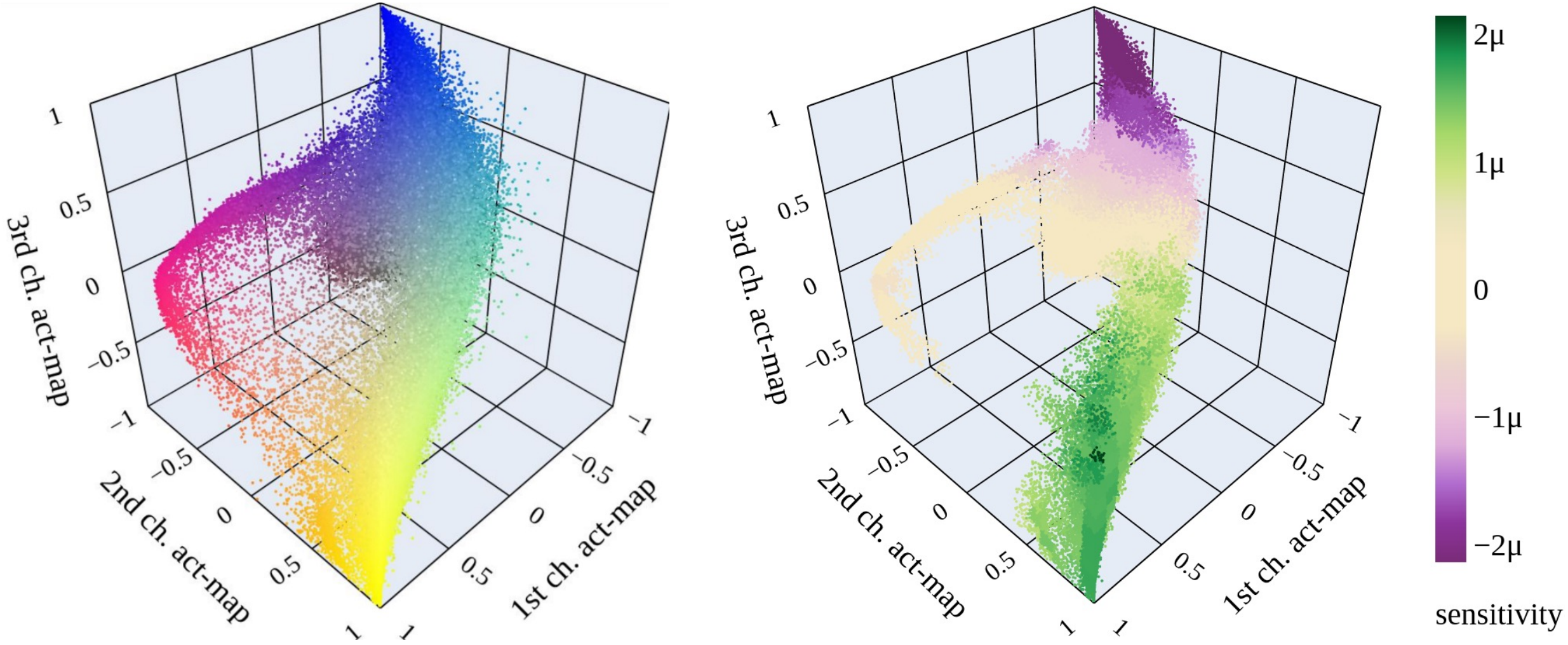
19.123 training samples

activation maps (3 channels)

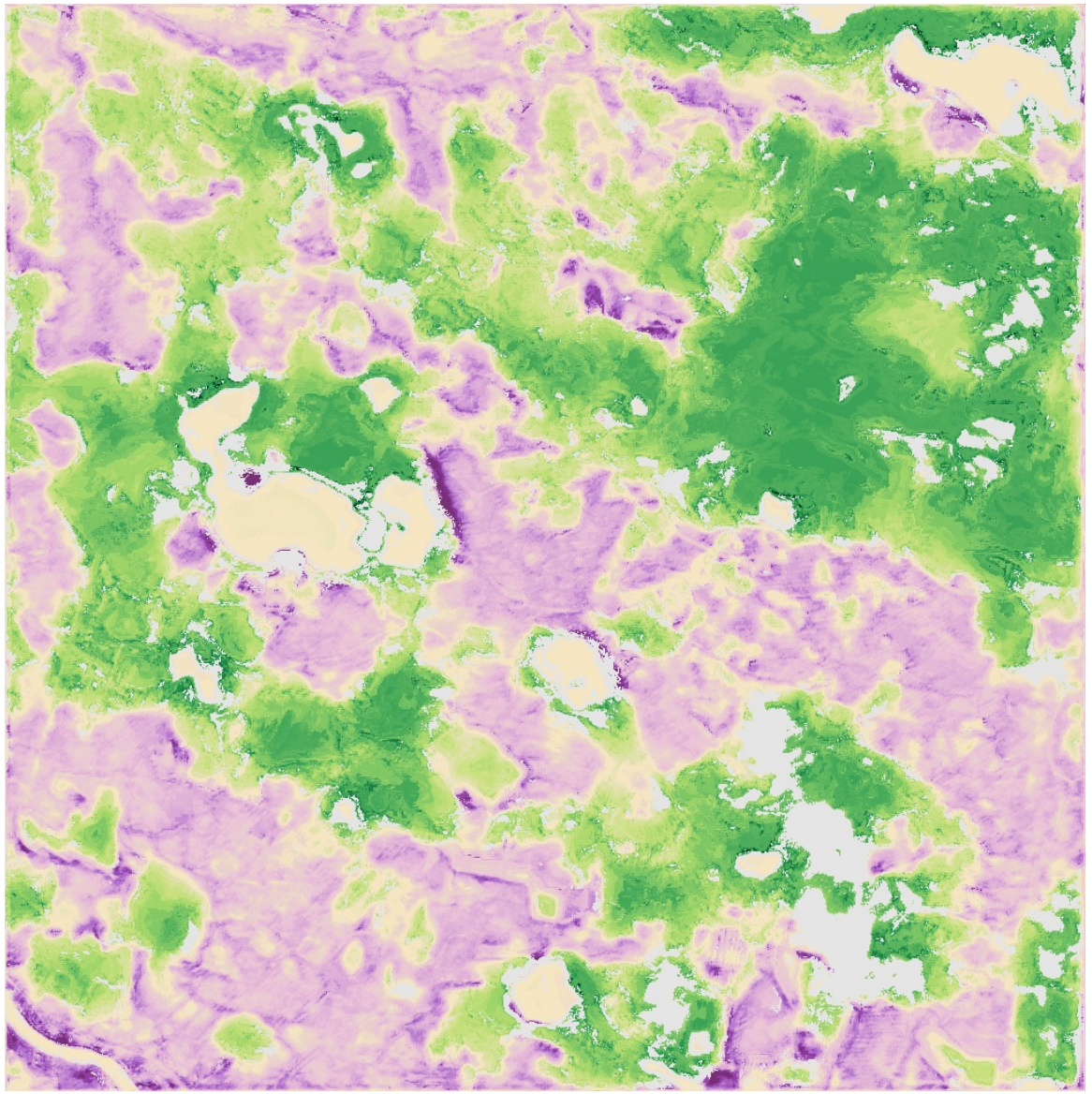
activation space



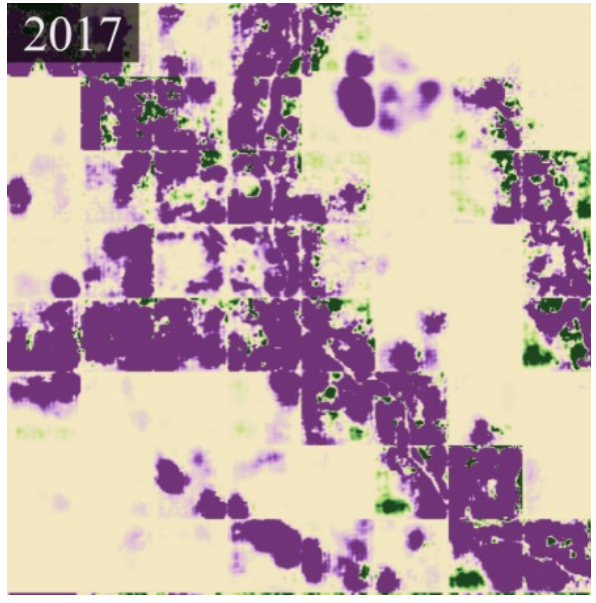
Activation space occlusion sensitivity



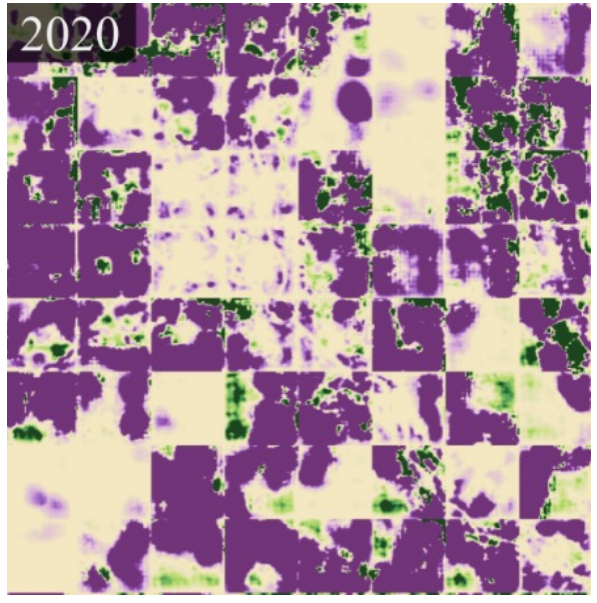
Results: North Ostrobothnia, Finland

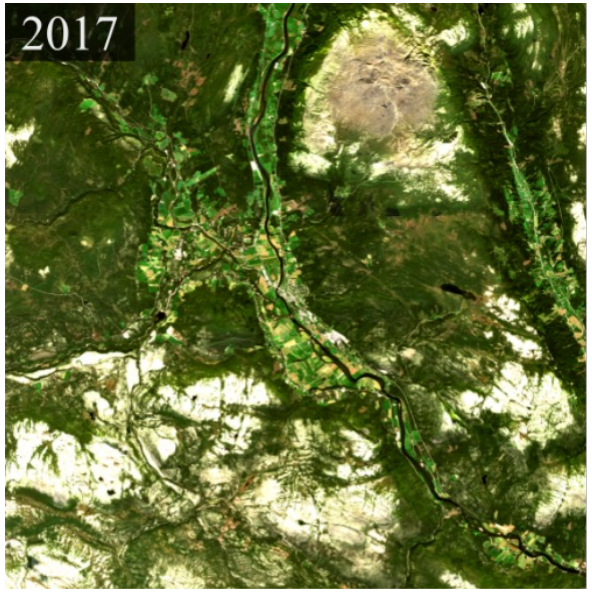


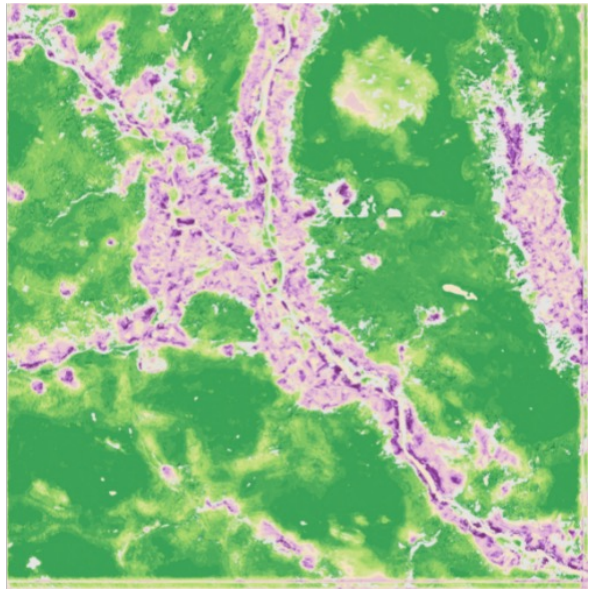
Results: Alvdal, Norway

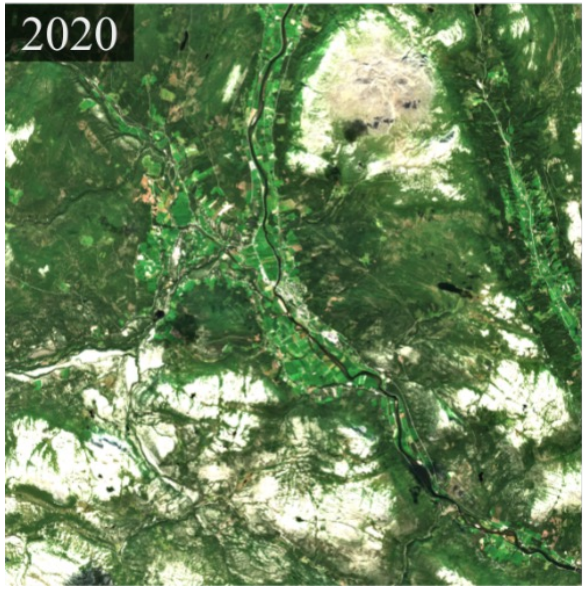


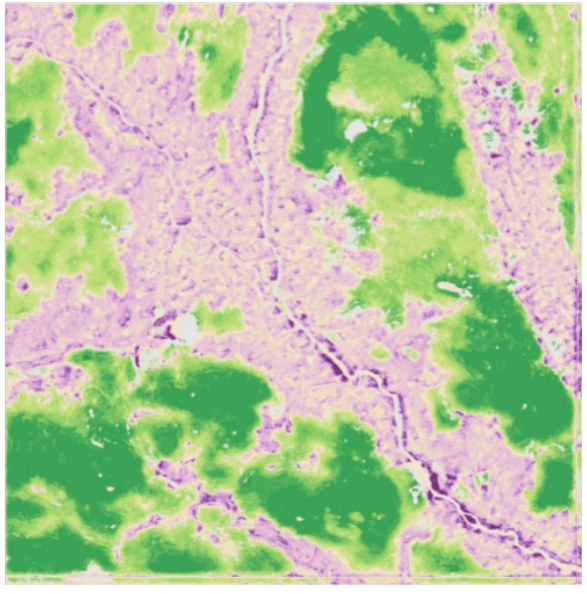
input image occlusion sensitivity



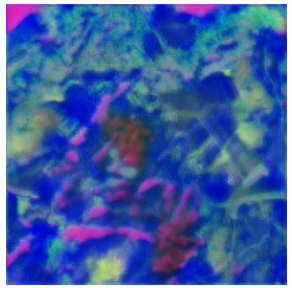


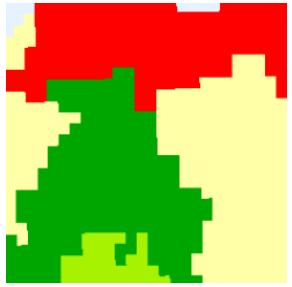


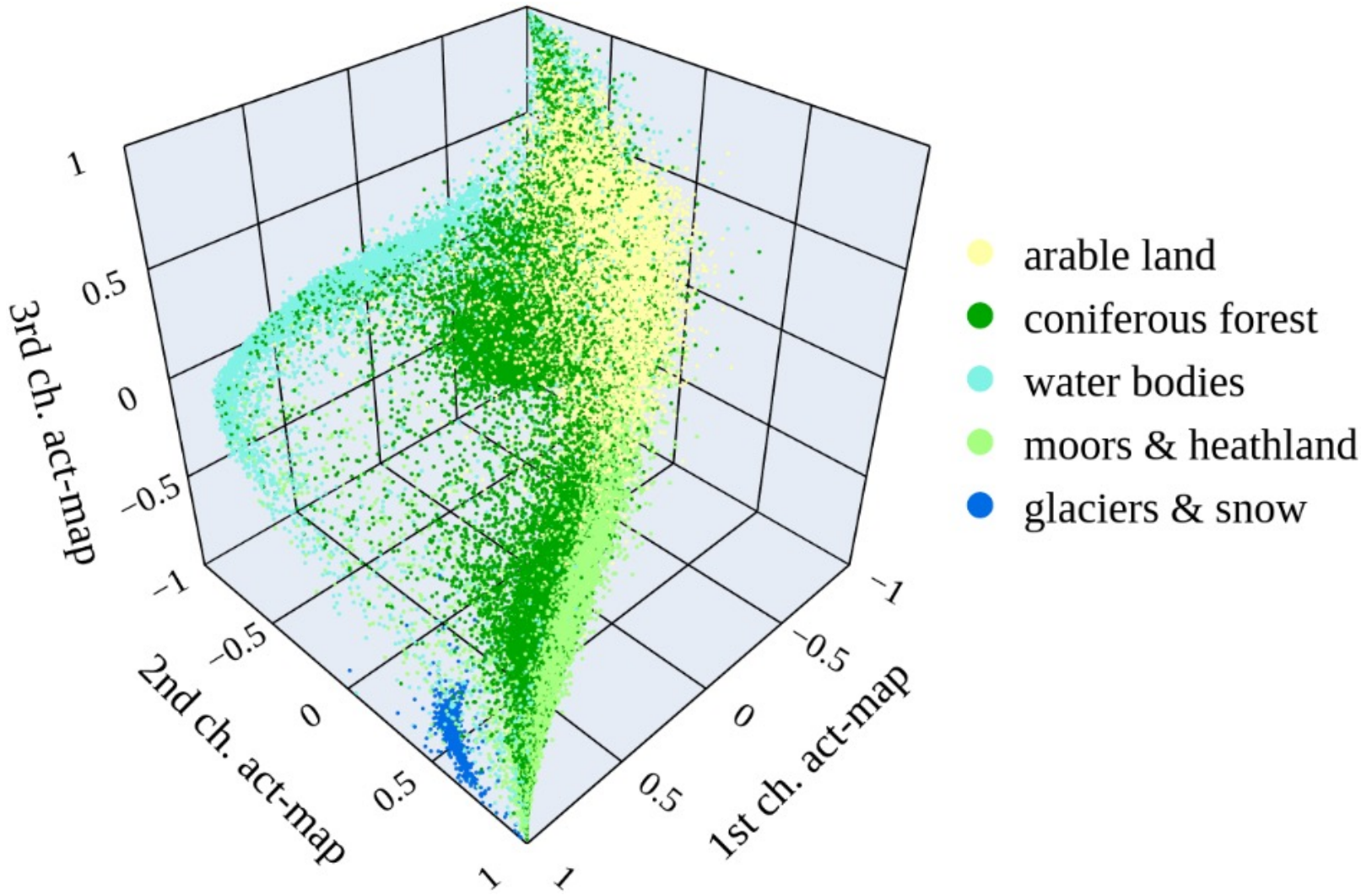




CORINE land cover in activation space







Evaluation of interpretations (for models aligned with existing domain knowledge)

Application-level evaluation

Quantify how easy an expert can explain the same decision as derived by the model/application

Human-level evaluation

Present different interpretations/explanations to a laypersons and let them choose the best one

Function-level evaluation

How easy is the interpretation based on known preferences

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Take away: Explainable machine learning…

…is not new

…offers a lot of methods which need to be chosen carefully based on your analysis goal

…connected to uncertainty quantification

…goes beyond explaining models which are aligned with our given knowledge

…needs domain experts

Further literature

- “Interpretable machine learning” by Christoph Molnar: https://christophm.github.io/interpretable-ml-book
- Heatmapping.org
- https://github.com/adebayoj/sanity_checks_saliency