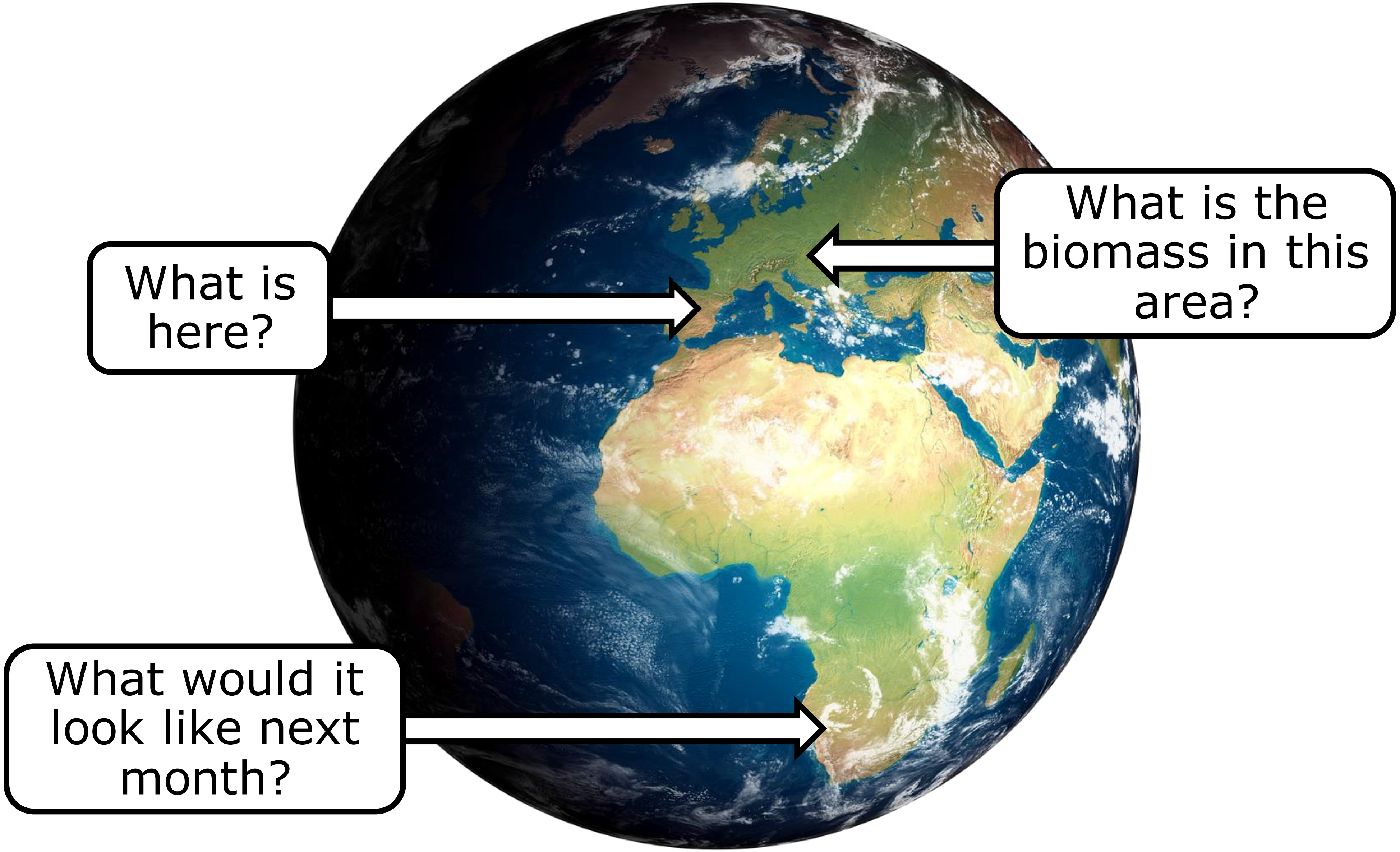
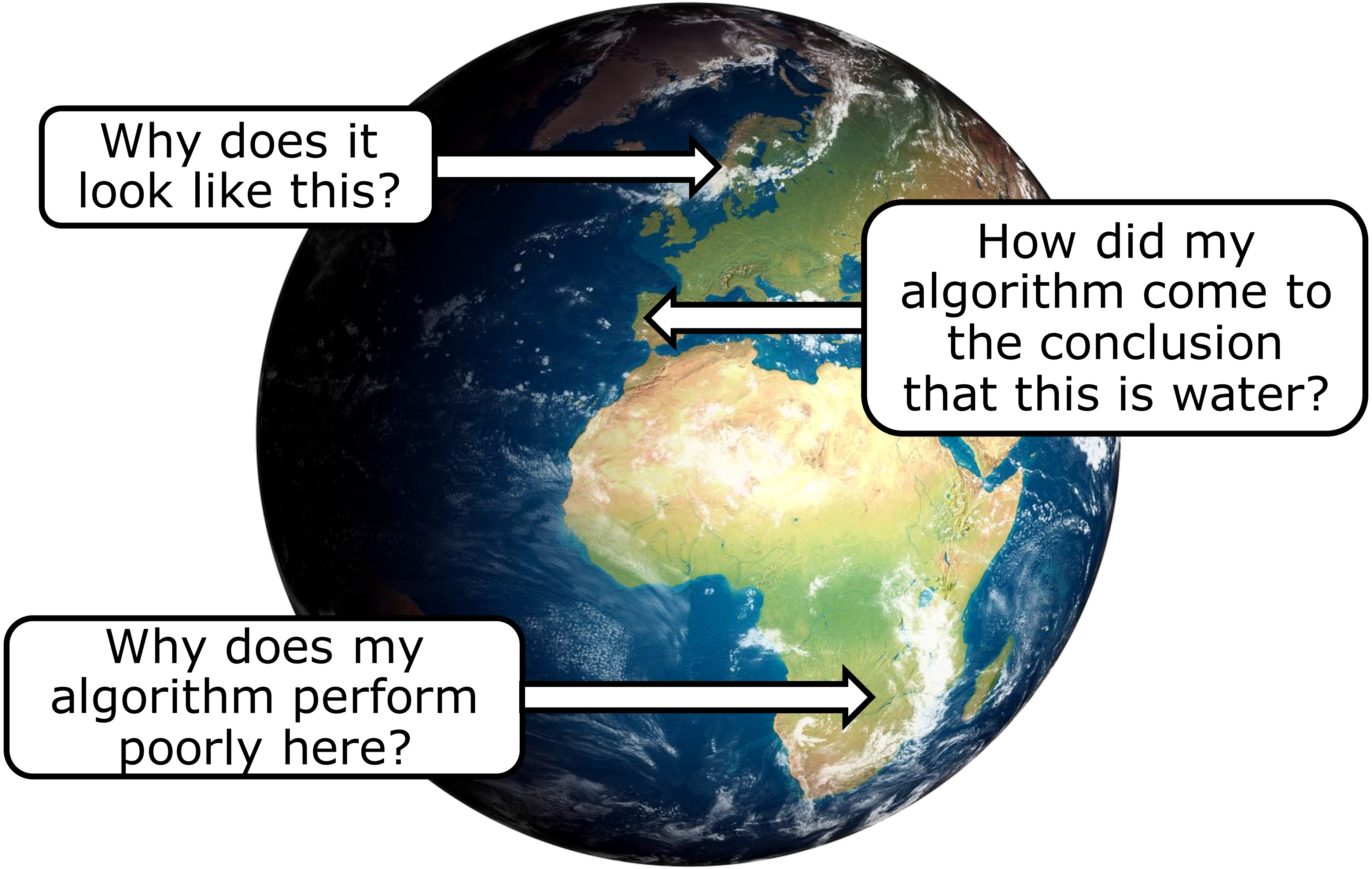
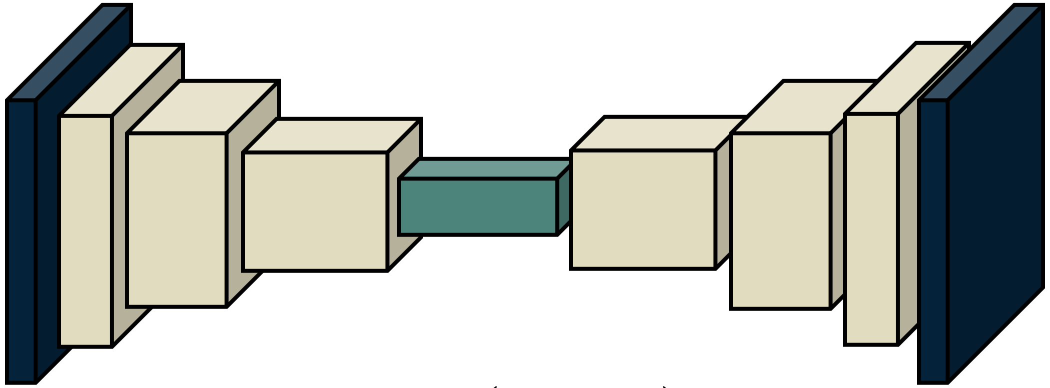
Introduction to explainable machine learning

Ribana Roscher


What is here?
What is the biomass in this area?
What would it look like next month?


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



$$\tilde{y} = f(X; \boldsymbol{w})$$

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: An image time-series dataset for GROWth analysis of cauLIFLOWER. *Journal of Field Robotics*, *40*(2), 173-192.

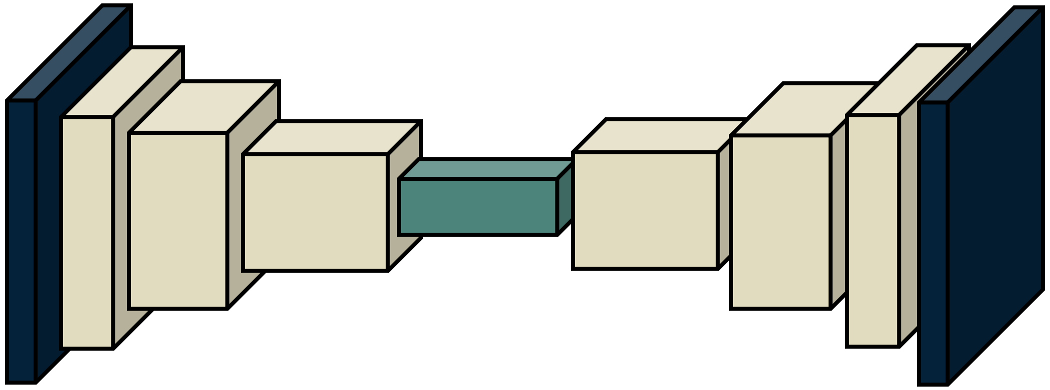
Core elements

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

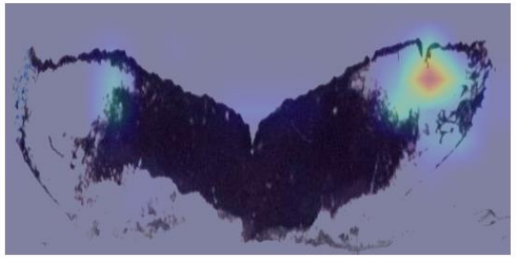
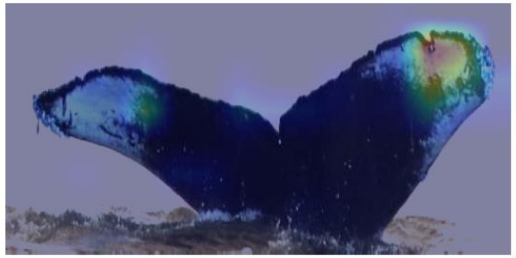
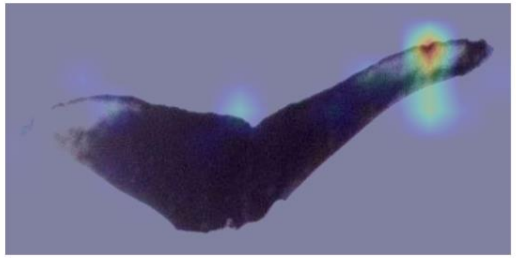
Example: heatmaps

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

- Also known as XAI, intelligible intelligence, etc.
- Combination of interpretable entities with **domain knowledge** (and an analysis goal)

Interpretability vs. explainability

Interpretability

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

Explainability

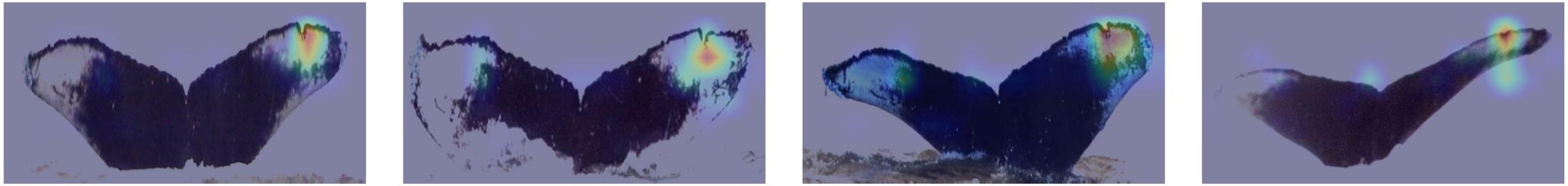
Combine interpretable entities with domain knowledge (and an analysis goal)

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 […].

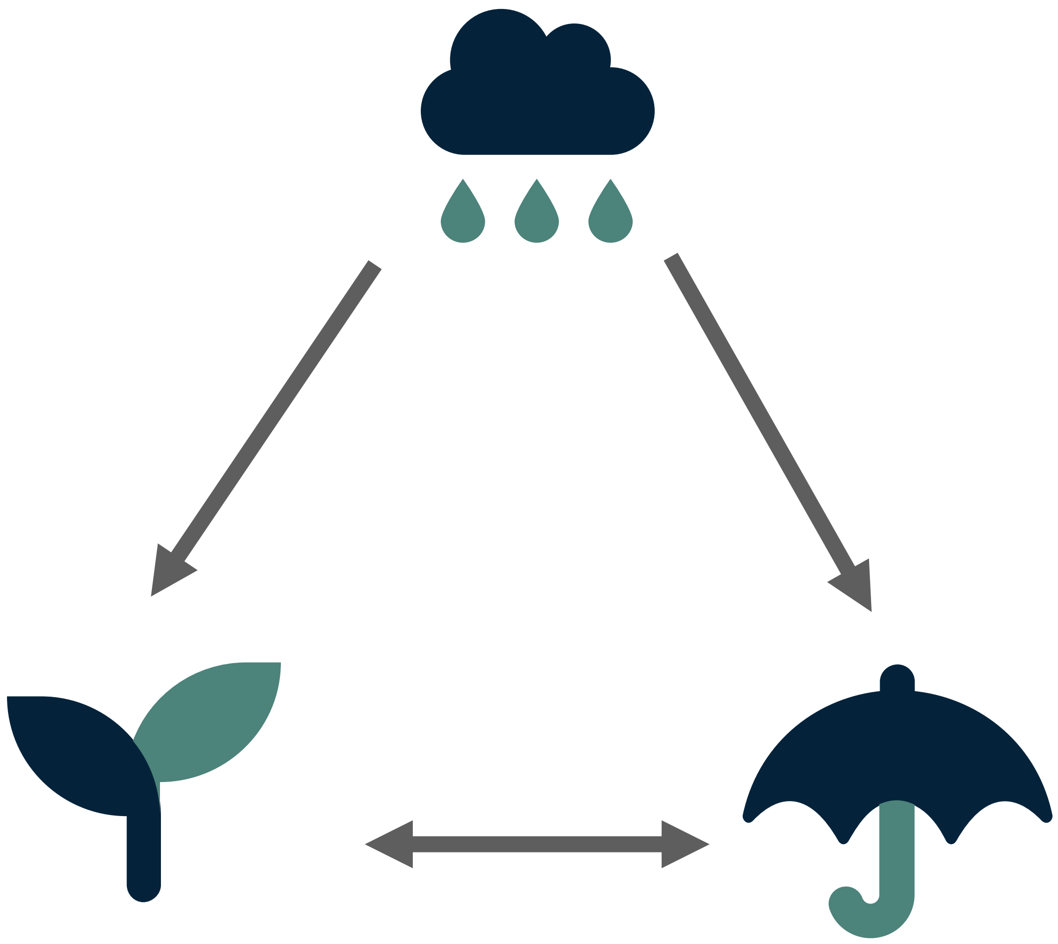
Explanation
The notch in the fluke of the whale with ID […] is a relevant fluke pattern for identifying this specific whale.

Connection to correlation and causation

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

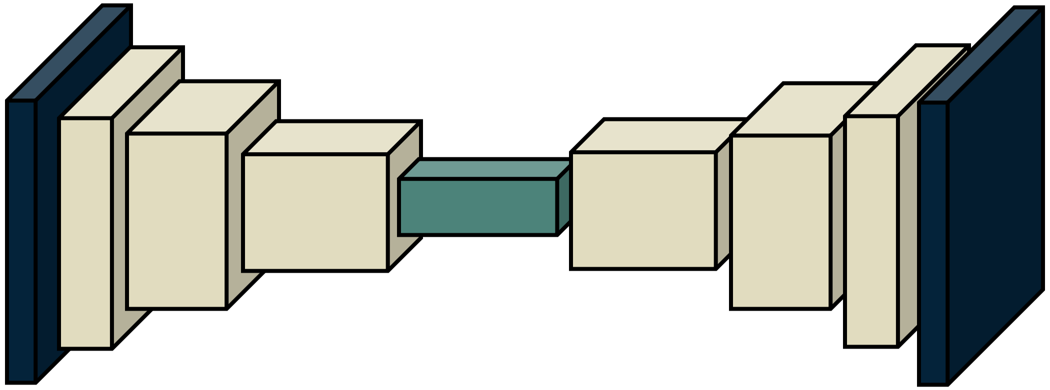
Correlation measures the relationship between input and output

- ➢ does not imply causation



Connection to correlation and causation

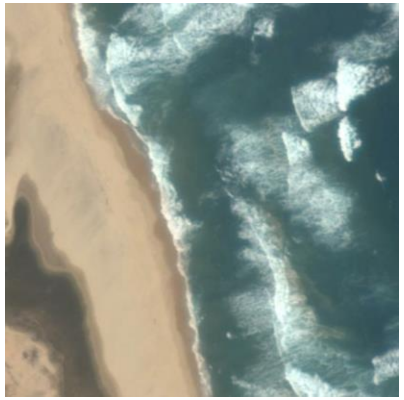
Interpretation tools present properties of a machine learning model and generally build on correlation

input →  → output

Confirmation bias

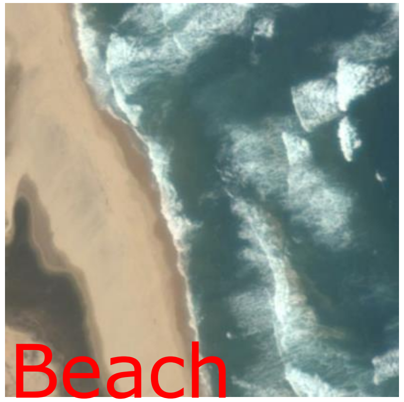
Underlying tendency to search for explanations which are in line with our existing knowledge

Clever Hans effect



no text is present

- classified correctly



text is present

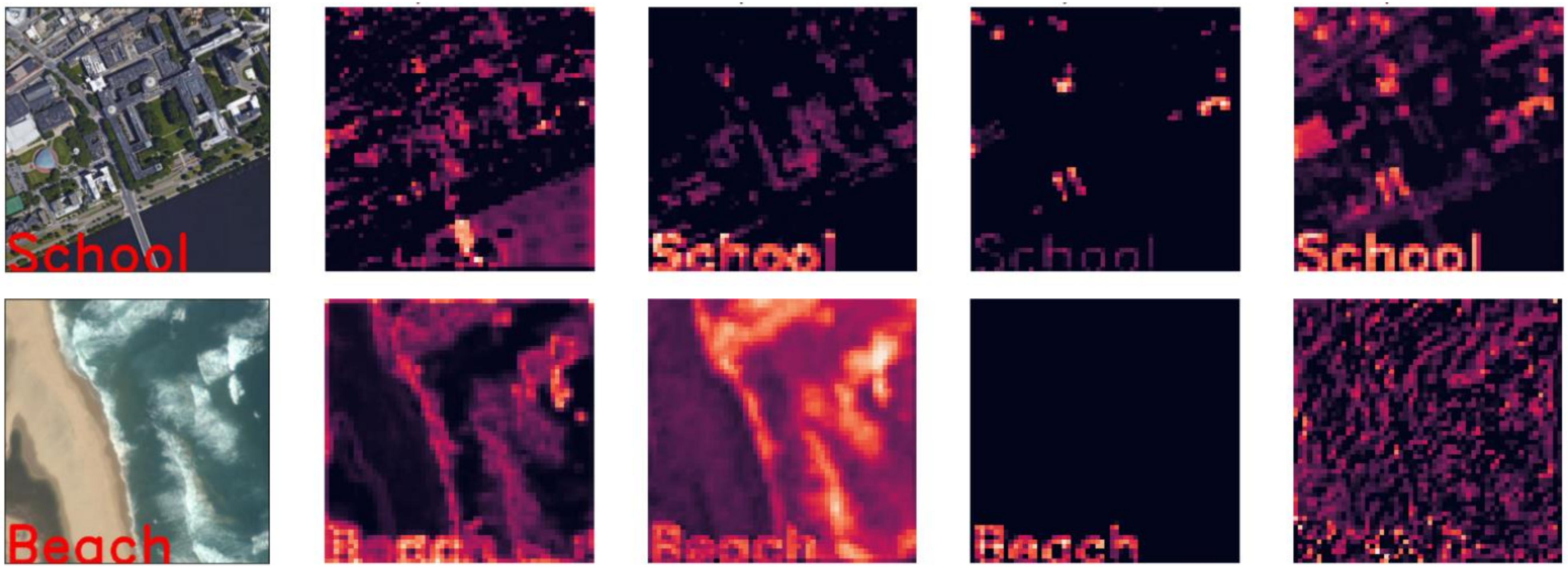
- classified correctly if text is correct
- classified incorrectly if text is wrong

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.

Xia, G. S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., ... & Lu, X. (2017). AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, *55*(7), 3965-3981.

Clever Hans effect

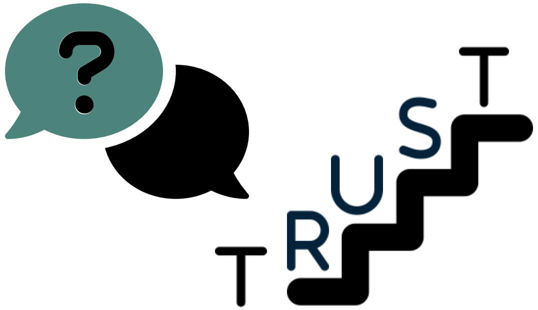
Text is highly activated in the feature maps instead of the objects representing the relevant class



Reasons to seek explanations

Justify decisions

- Explain the particular outcome rather than describing the inner workings of a machine learning model
- Especially important when decision is unexpected
- Should defend the outcome to be fair and reasonable
- Increases trust



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

Reasons to seek explanations

(Enhance) control

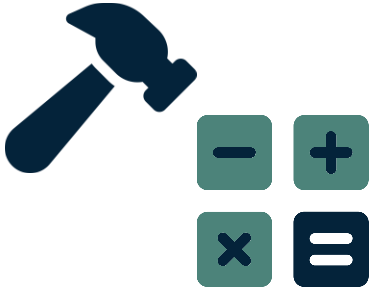
- To prevent that things go wrong
- Insights into model behavior ensures visibility of flaws
- Enables a fast reaction time

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

Reasons to seek explanations

Improve models

- A better understanding of the inner workings and the behavior of a model enables a targeted improvement
- Improvement can concern data and model



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

Reasons to seek explanations

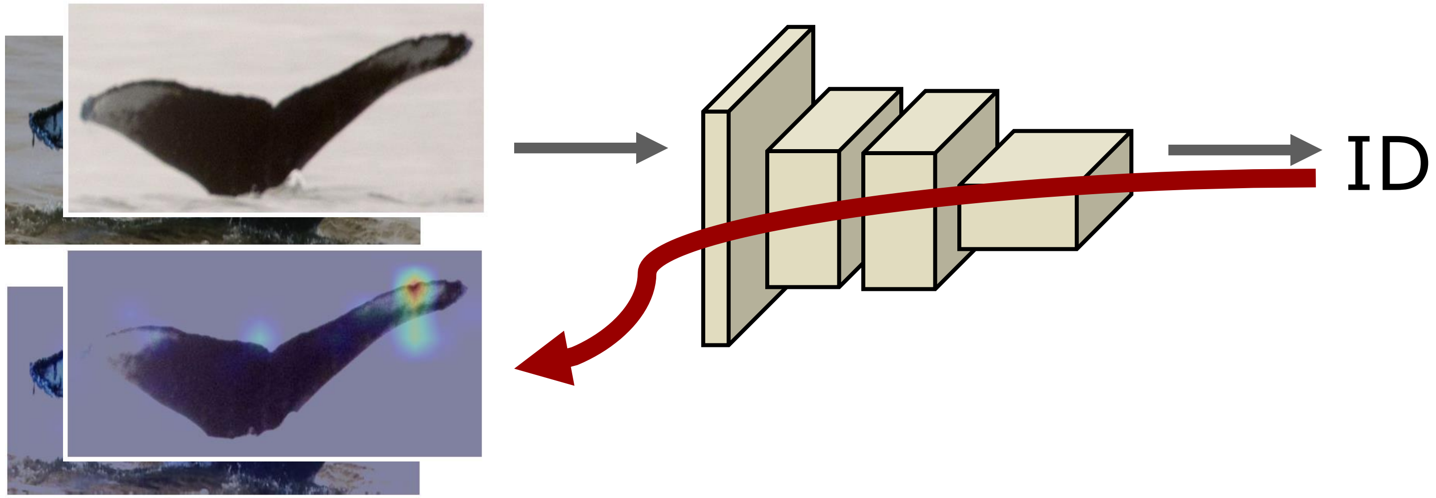
Discover new knowledge

- Patterns in decision process or new insights in the data can teach us new things
- If model performs better than human, understanding the model can help to improve and correct our previous knowledges

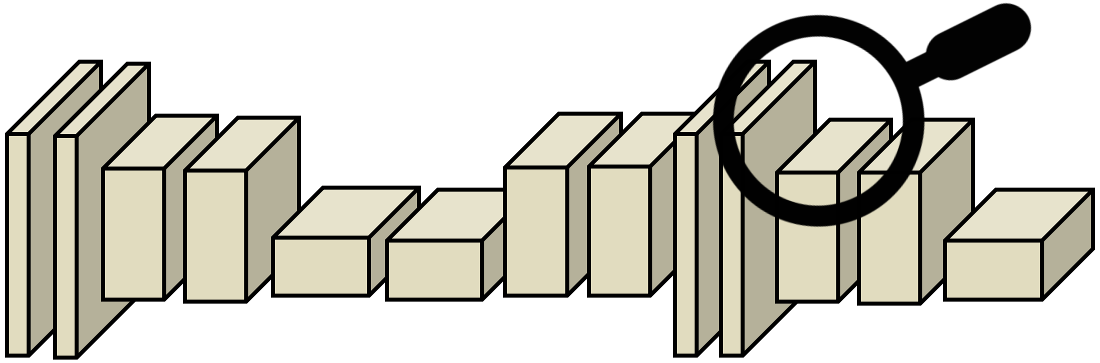
Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138-52160.

Approaches categorized by specificity

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

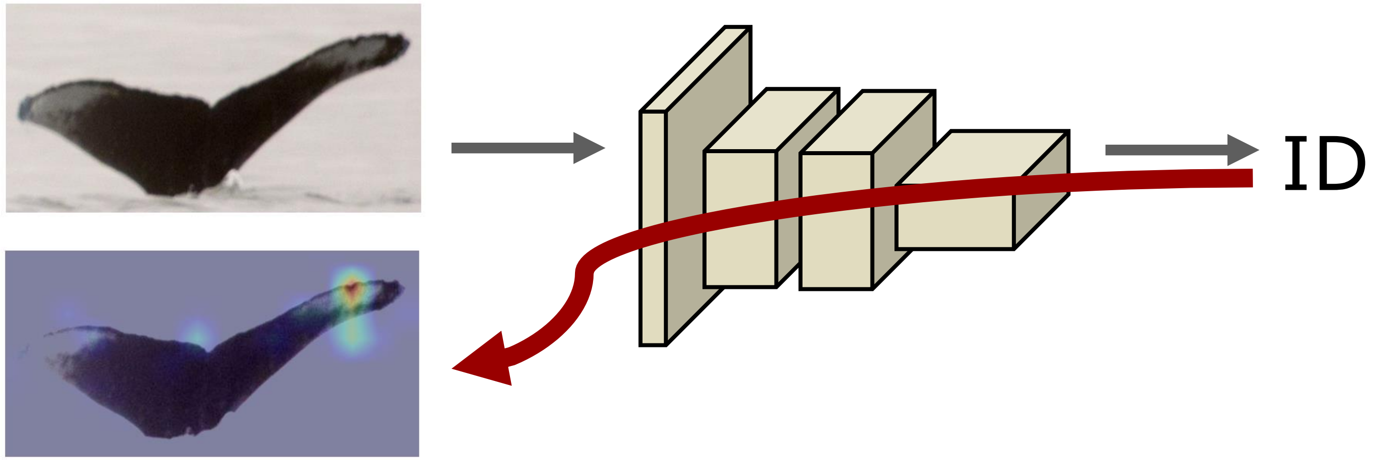


Explaining the whole model or parts (model-specific)

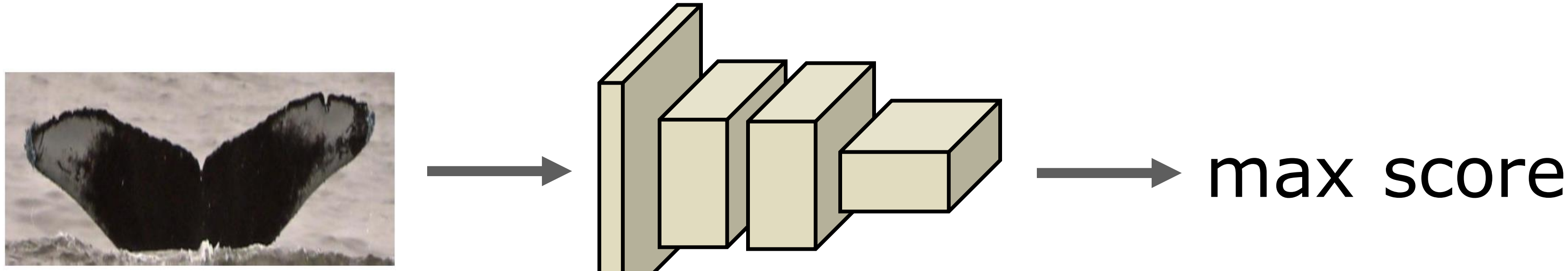


Approaches categorized by locality

Explain locally (individual output)



Explain globally (entire model)



prototype

Properties of interpretation (techniques)

Expressive power
Language of extracted explanations (if-then rules, histograms, natural language, etc.)

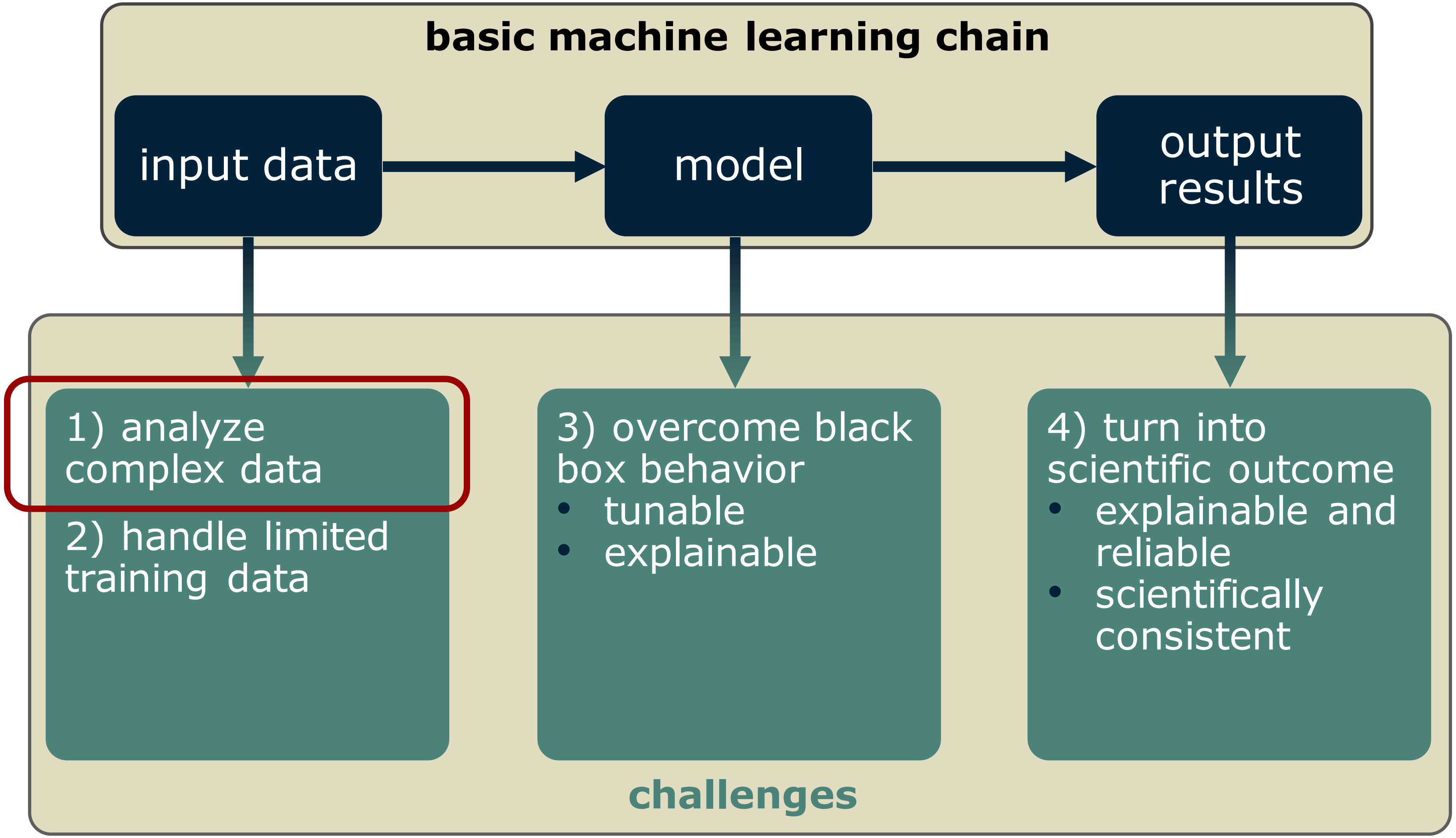
Translucency
Degree to which a method looks into model

Portability
Range of methods the technique can be applied to/combined with

Algorithmic complexity
Computational complexity to produce an explanation

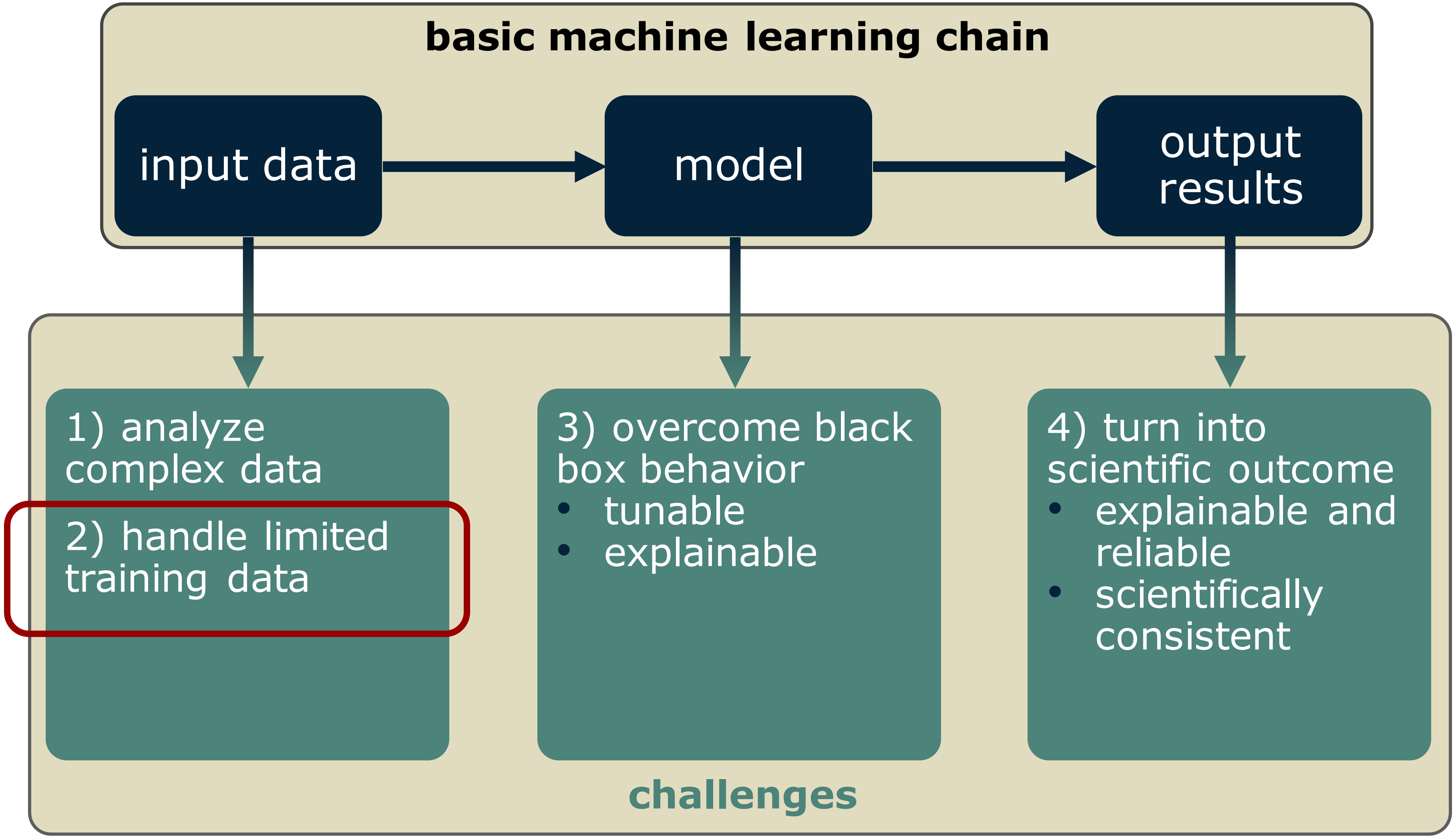
Robnik-Šikonja, M., & Bohanec, M. (2018). Perturbation-based explanations of prediction models. In *Human and machine learning* (pp. 159-175). Springer, Cham.

Can explainable ML be useful in the sciences?



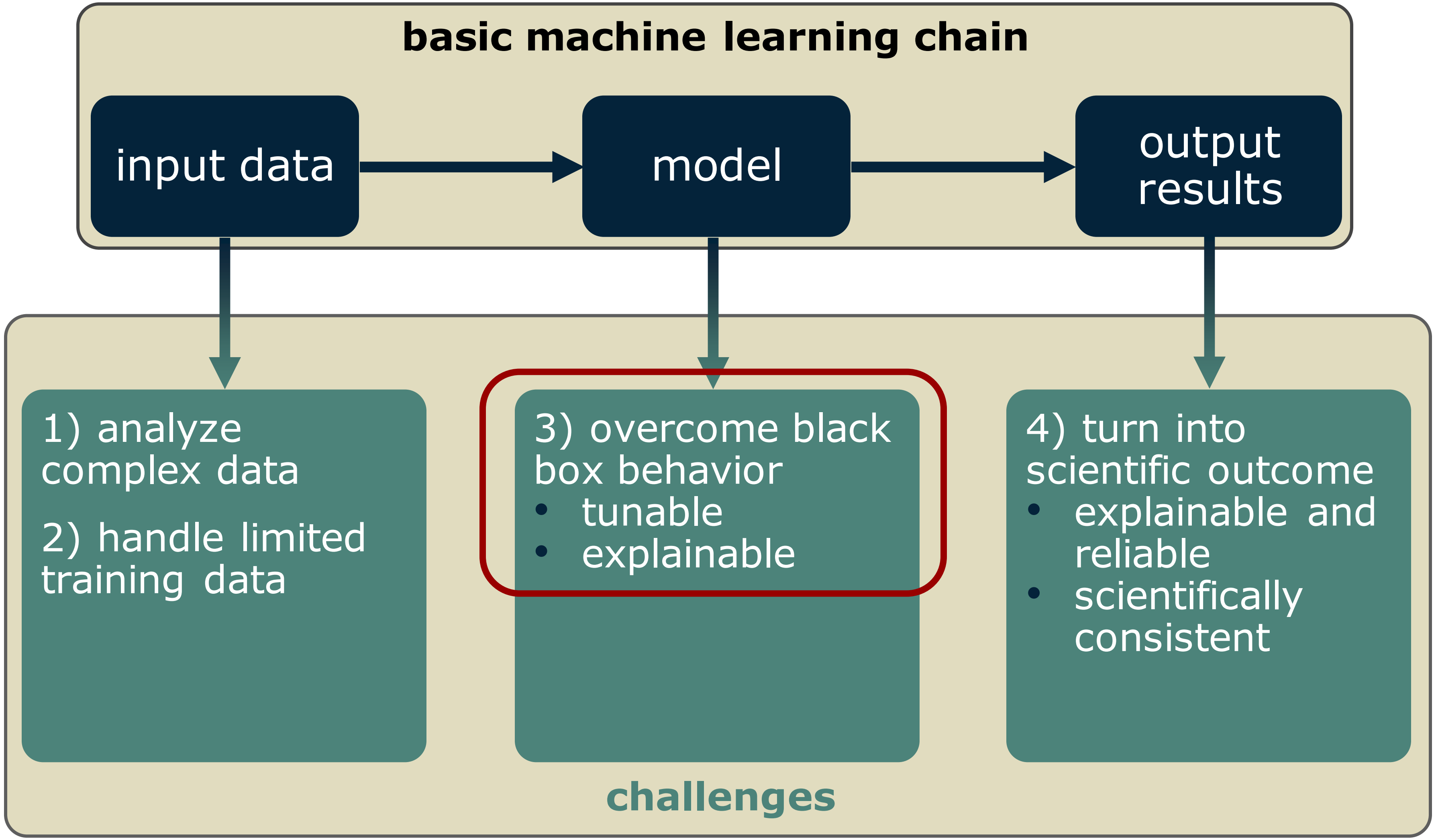
Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explain it to me - facing remote sensing challenges in the bio- and geosciences with explainable machine learning. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *5*, 817-824.

Can explainable ML be useful in the sciences?



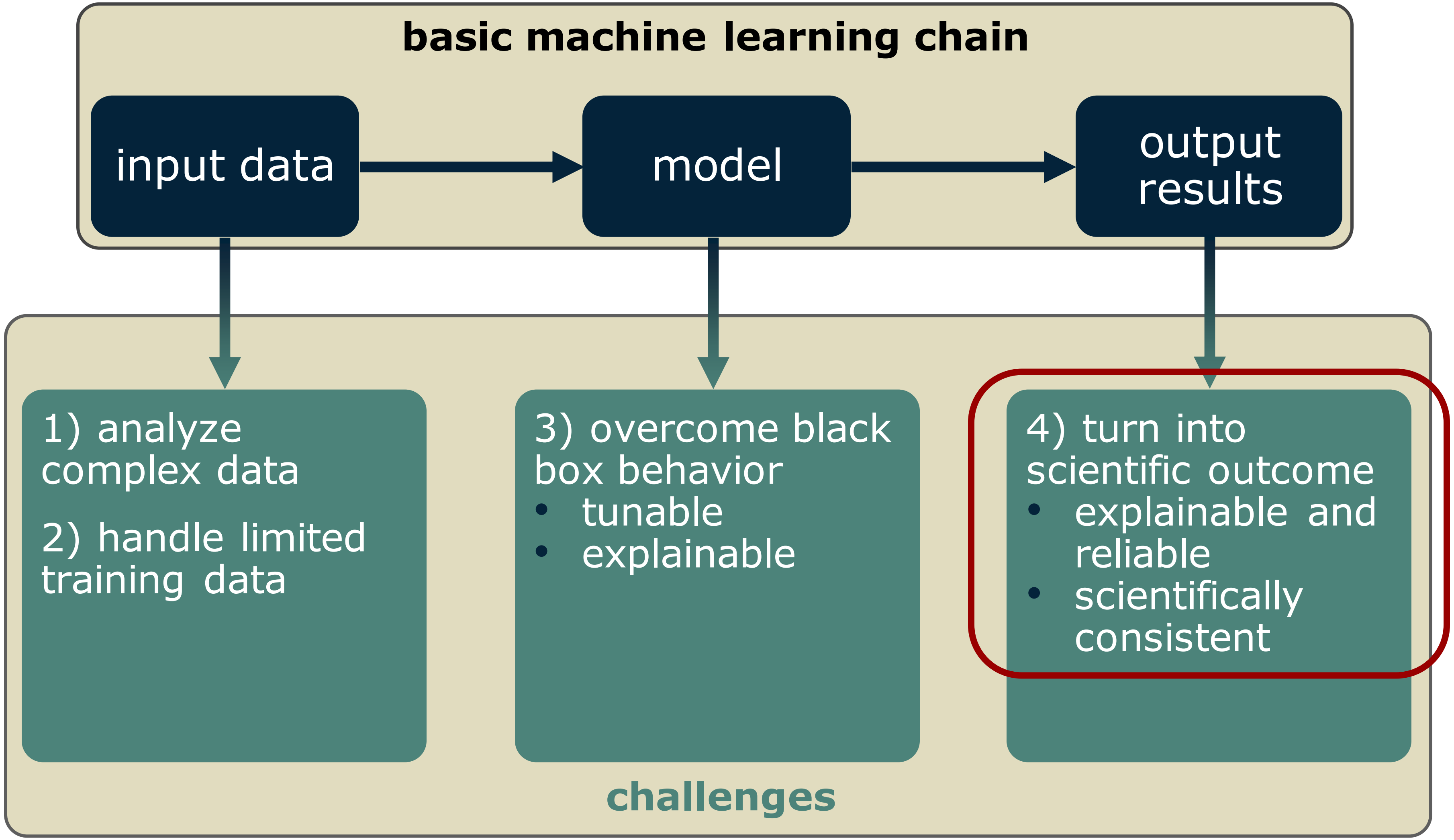
Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explain it to me - facing remote sensing challenges in the bio- and geosciences with explainable machine learning. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *5*, 817-824.

Can explainable ML be useful in the sciences?



Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explain it to me - facing remote sensing challenges in the bio- 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 the sciences?



Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explain it to me - facing remote sensing challenges in the bio- and geosciences with explainable machine learning. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *5*, 817-824.

Terms connected to interpretability

impact

contribution

feature attribution

influence

importance

relevance

saliency

sensitivity

value

summary statistic

pixel attribution

Inherently interpretable models

Notation

Given **feature vectors**

$$\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_N$$

and some **target** (response) outputs

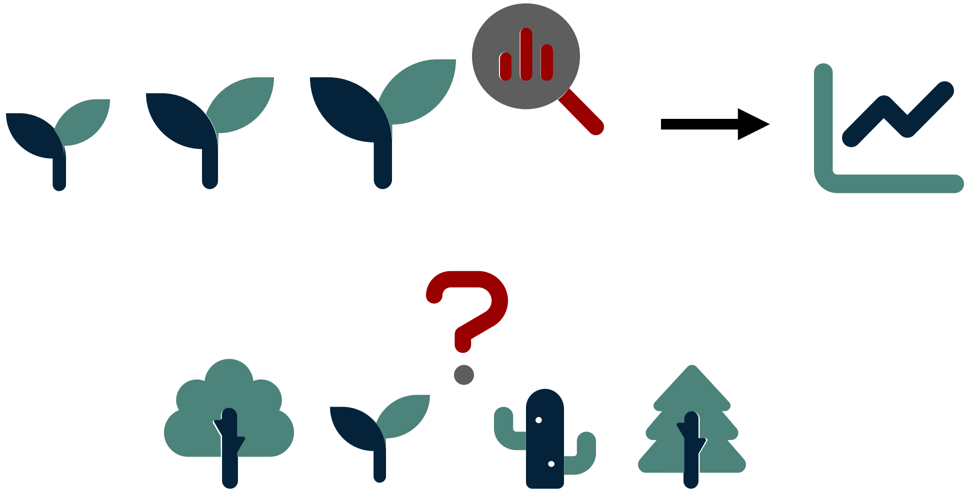
$$\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N$$

the goal is to predict the **output** given new inputs

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

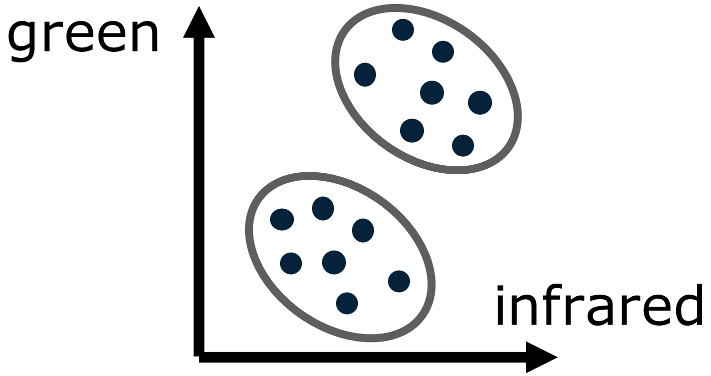
Notation

Supervised learning goal:
map inputs to output



$$\mathcal{T} = \{(\phi_n, \boldsymbol{y}_n)\}_N$$

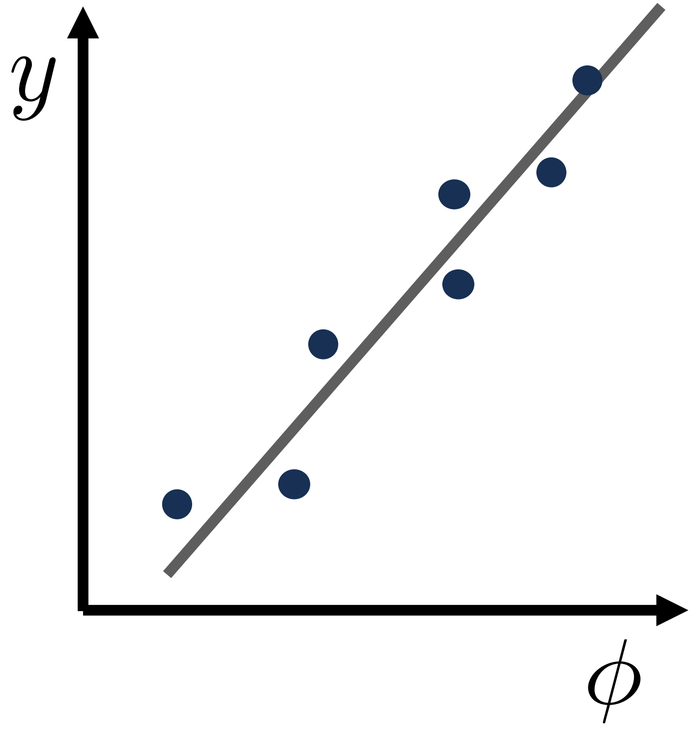
Unsupervised learning goal:
learn the underlying structure
of the data **without targets**



$$\mathcal{T} = \{\phi_n\}_N$$

Linear regression

ϕ: input features
y: output, response



bias term (intercept)

weights (direction)

$$f(\boldsymbol{\phi}, \boldsymbol{w}) = \tilde{y} = w_0 + \sum_d w_d \phi_d + \epsilon$$

Interpretations

- Coefficients/weights and intercept
- Feature importance
- Feature effect

Interpretation of coefficents

- Interpretation in combination with feature
- Intercept is interpreted with feature $\phi_0 = 1$

$$\tilde{y} = w_0 + \sum_d w_d \phi_d + \epsilon$$

Numerical

feature ϕ_d by $\Delta\phi_d = 1$ outcome $\tilde{y}$ by w_d

Binary {0,1}

feature ϕ_d by $\Delta\phi_d = 1$ outcome $\tilde{y}$ by w_d

Interpretation of intercept

- The value of the intercept is the outcome of a sample with all features at their mean value
- Requirement: all features need to be normalized to zero mean and standard deviation of one

$$\tilde{y} = w_0\phi_0 + \boxed{\sum_d w_d\phi_d + \epsilon}$$

$= 0$, when all features are at their mean value

Feature importance

Depends on value of associated weight and the weight's variance (standard error)

t-statistic

$$t_{\widetilde{w}_d} = \frac{\widetilde{w}_d}{\epsilon_{\mathrm{st}}(w_d)}$$

- Indicates whether the weight is significantly different from zero

Effect

- Value of weights depend on the range of the features
- Effect: contribution of weight-feature combination to the actual outcome

$$e_d = w_d \phi_d$$

- Can be performed for one feature and for the whole dataset

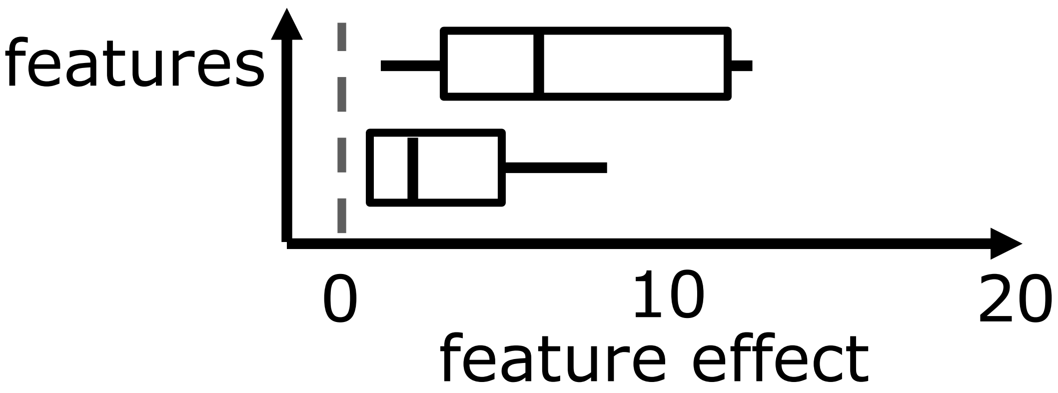
Weights vs. effect visualization

Weight plot

- Weight values with confidence intervals
- Weights difficult to compare

Effect plot

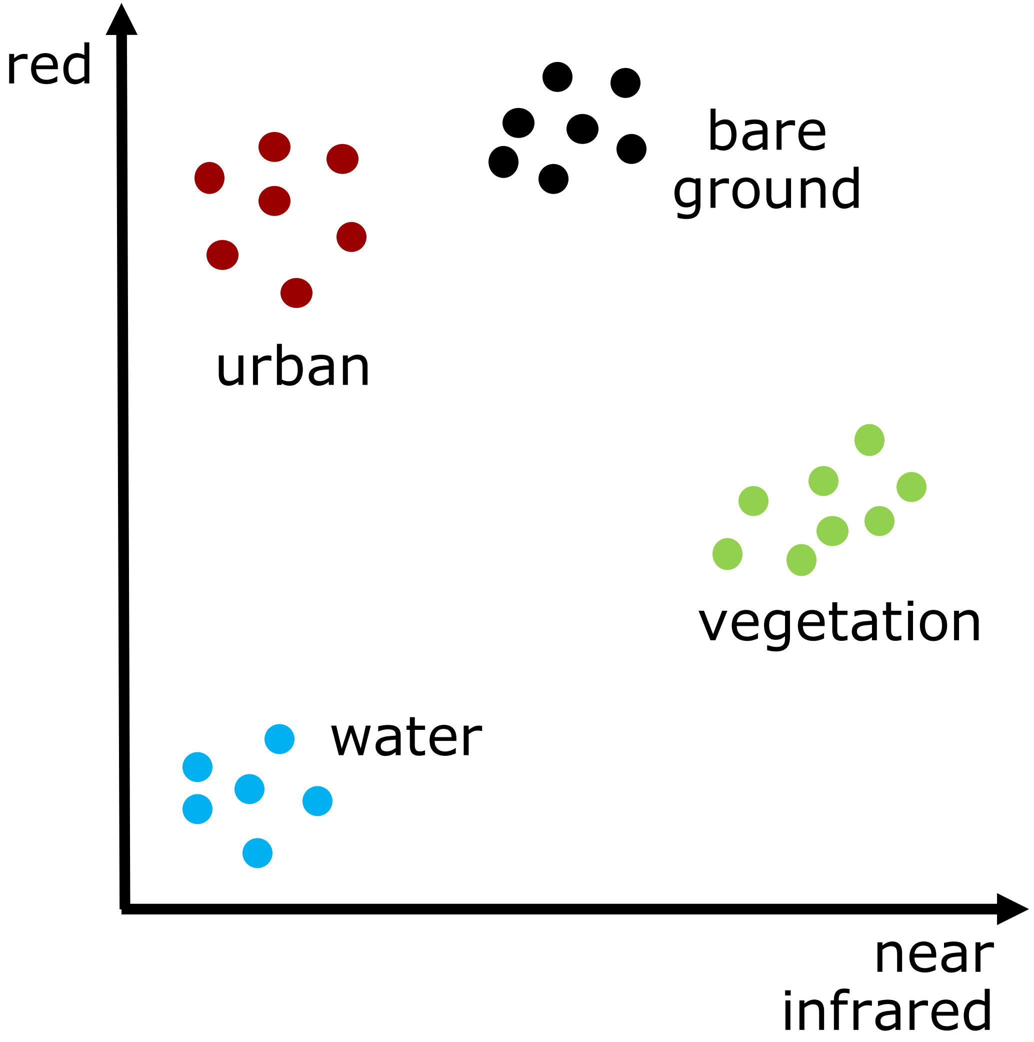
- Boxplot with quantiles, median effect, max and min values (without outliers)
- Distribution of effects per feature



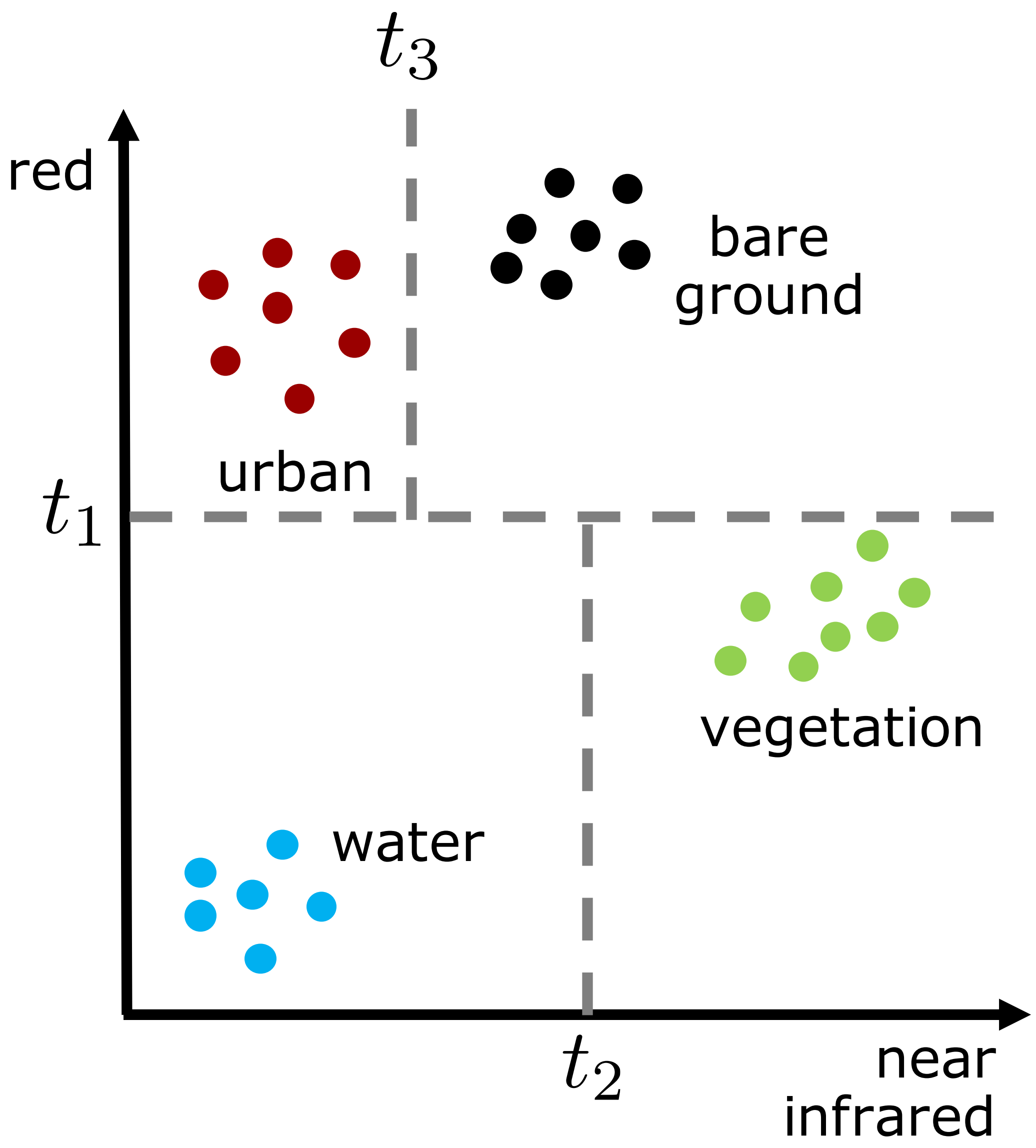
Advantages and disadvantages

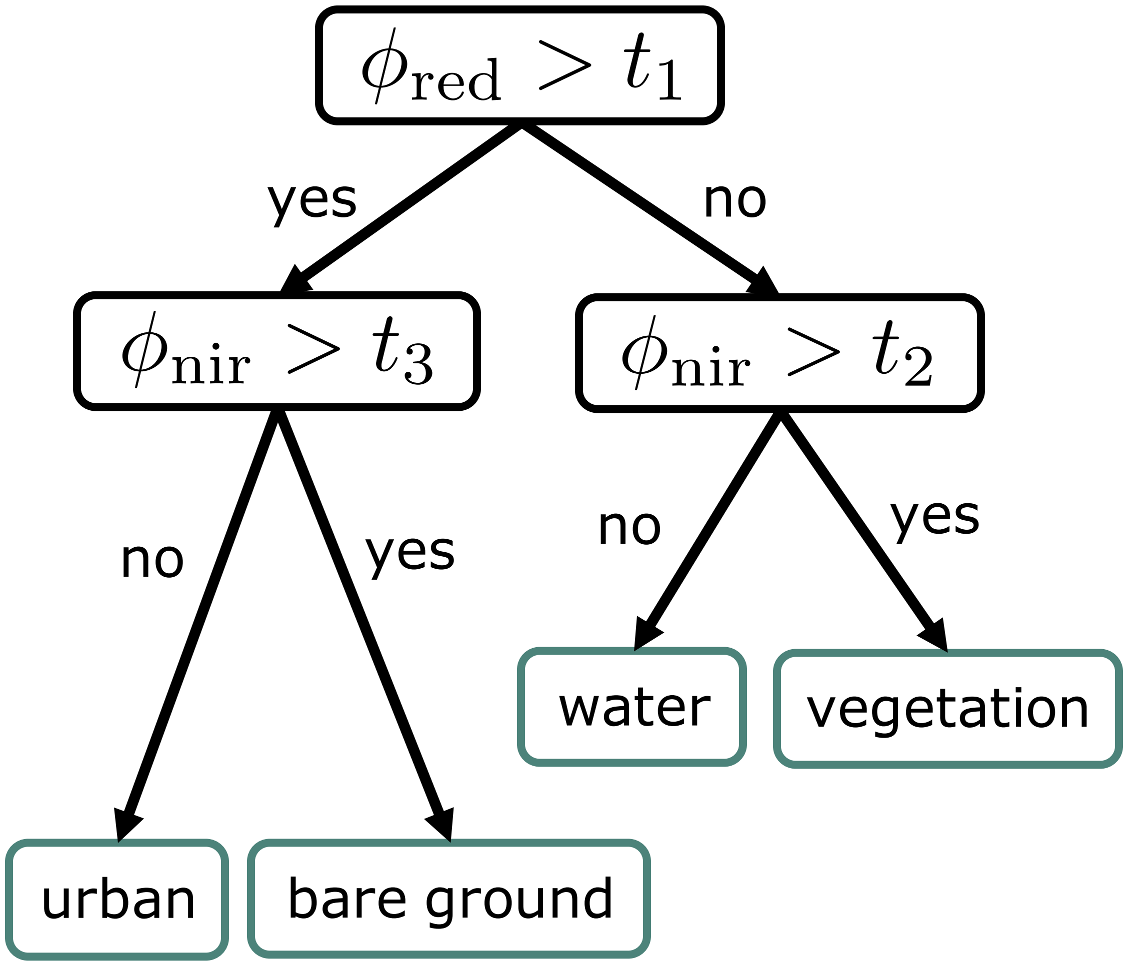
- Easy to understand and analyze
- Linear regression is well understood and common tools for interpretation exist
- Linear relationship is a strong assumption
- Due to linearity, predictive performance is mostly low
- Correlated features might lead to unintuitive interpretations

Decision tree for land cover classification

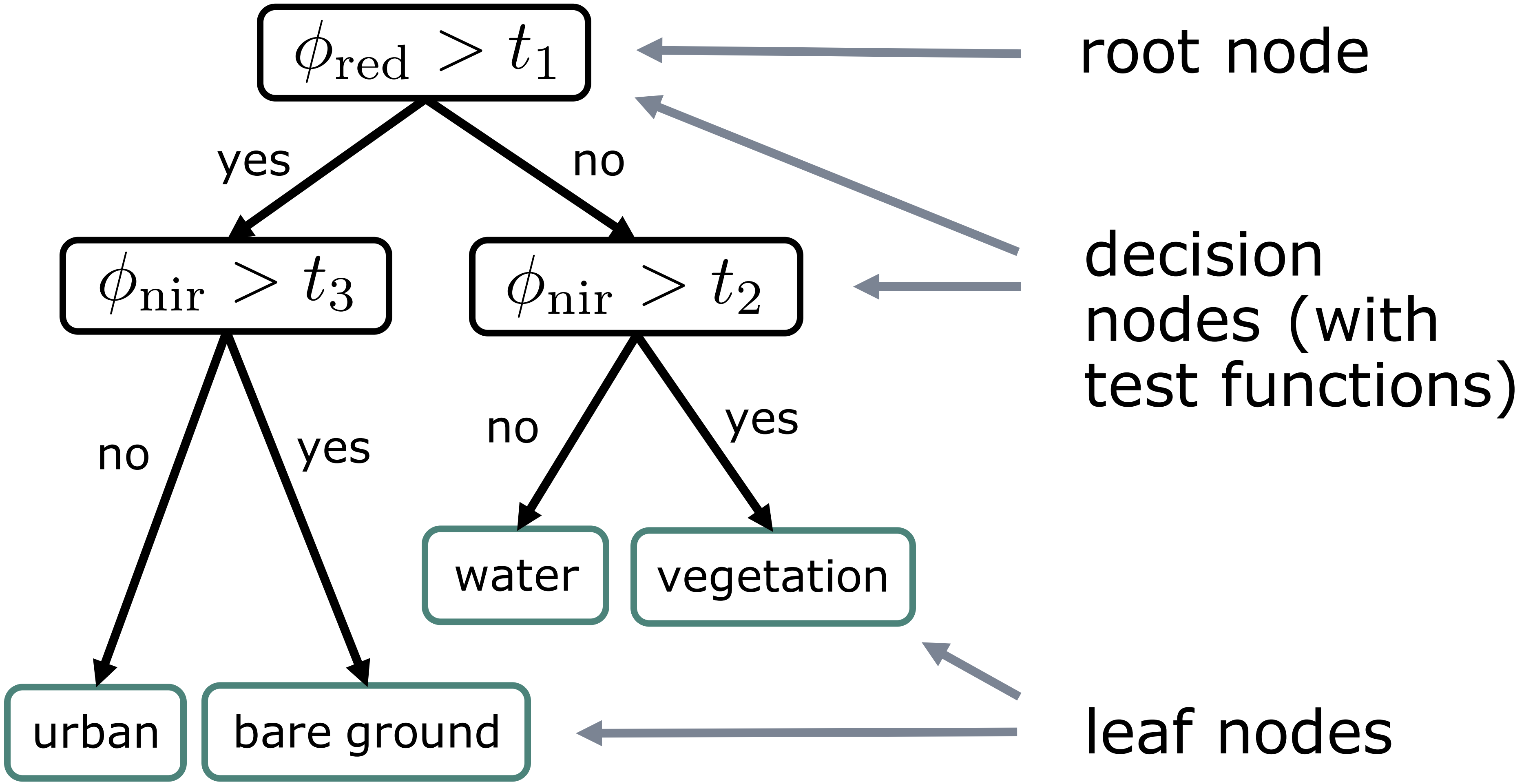


Decision tree for land cover classification



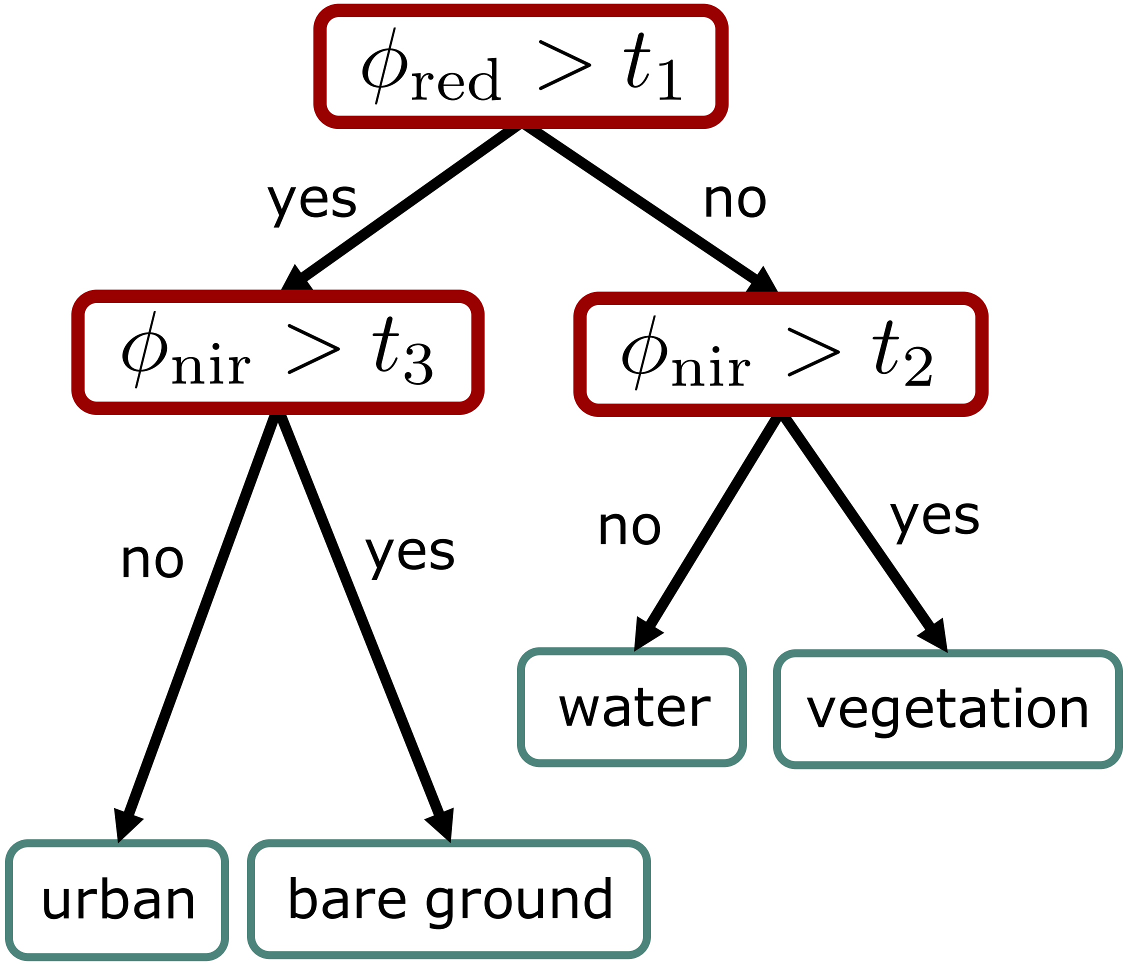


Elements of a Decision Tree



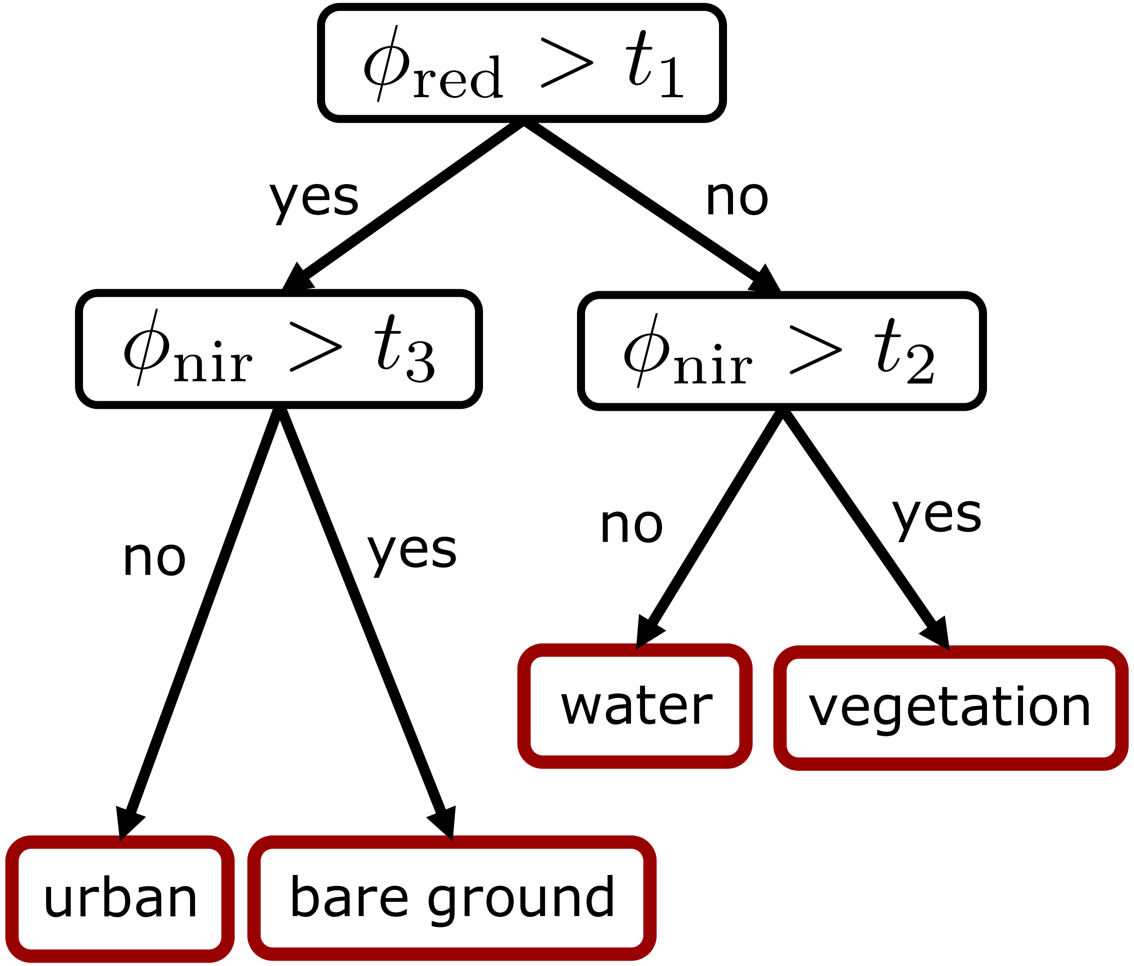
Decision nodes

- Each **decision node** (split node) implements a **test function with discrete outcomes**
- The test function of each decision node splits the input space into regions $\mathcal{R}_l$



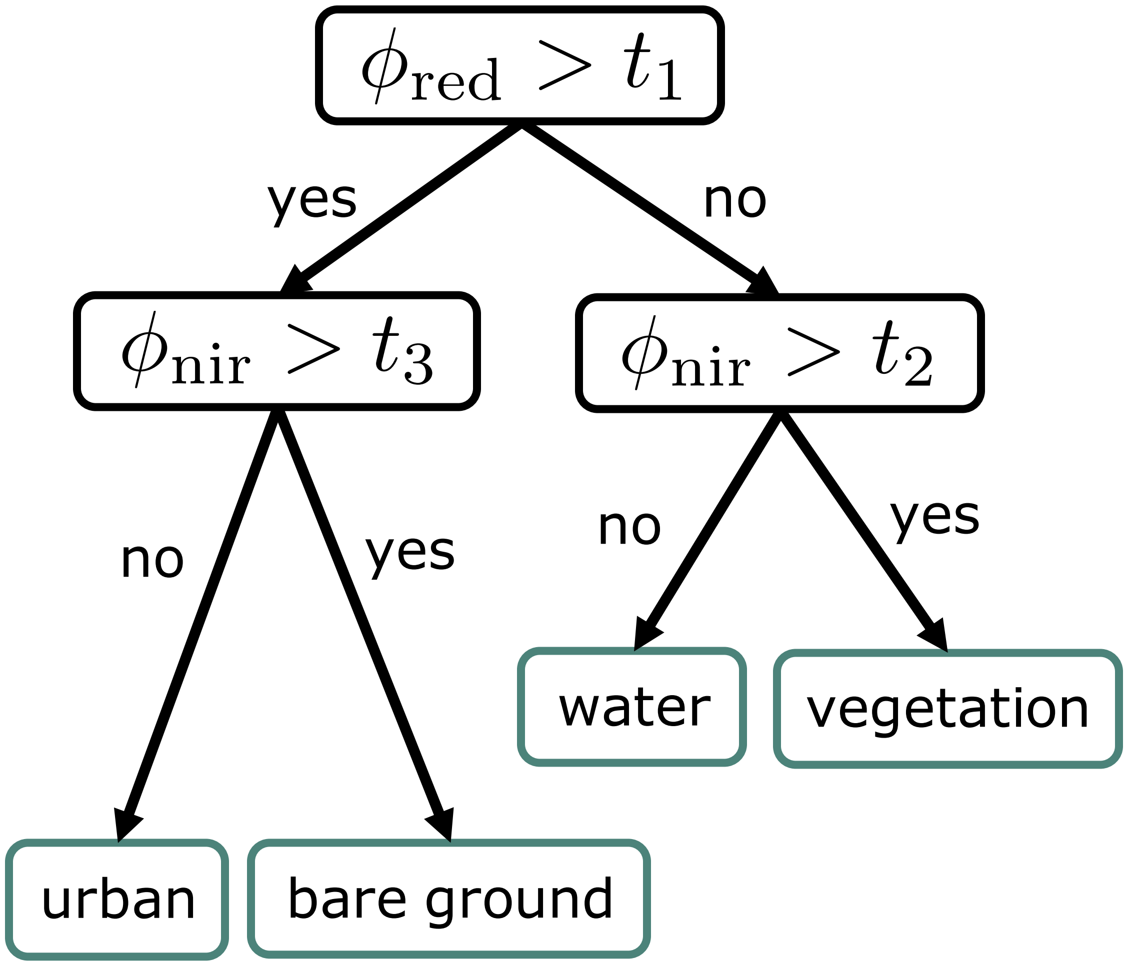
Leaf nodes

- A **leaf** node symbolizes the **end of a sequence of decisions**
- A single (output) class is associated to each leaf node
- A leaf node defines a localized region in the input space where samples falling in this region have the same label



Classify a sample with a given decision tree

1. Start at the root node
2. If current node is a leaf node, return class label
3. Perform the test of the current decision node and follow the corresponding branch
4. Goto 2



Interpretations

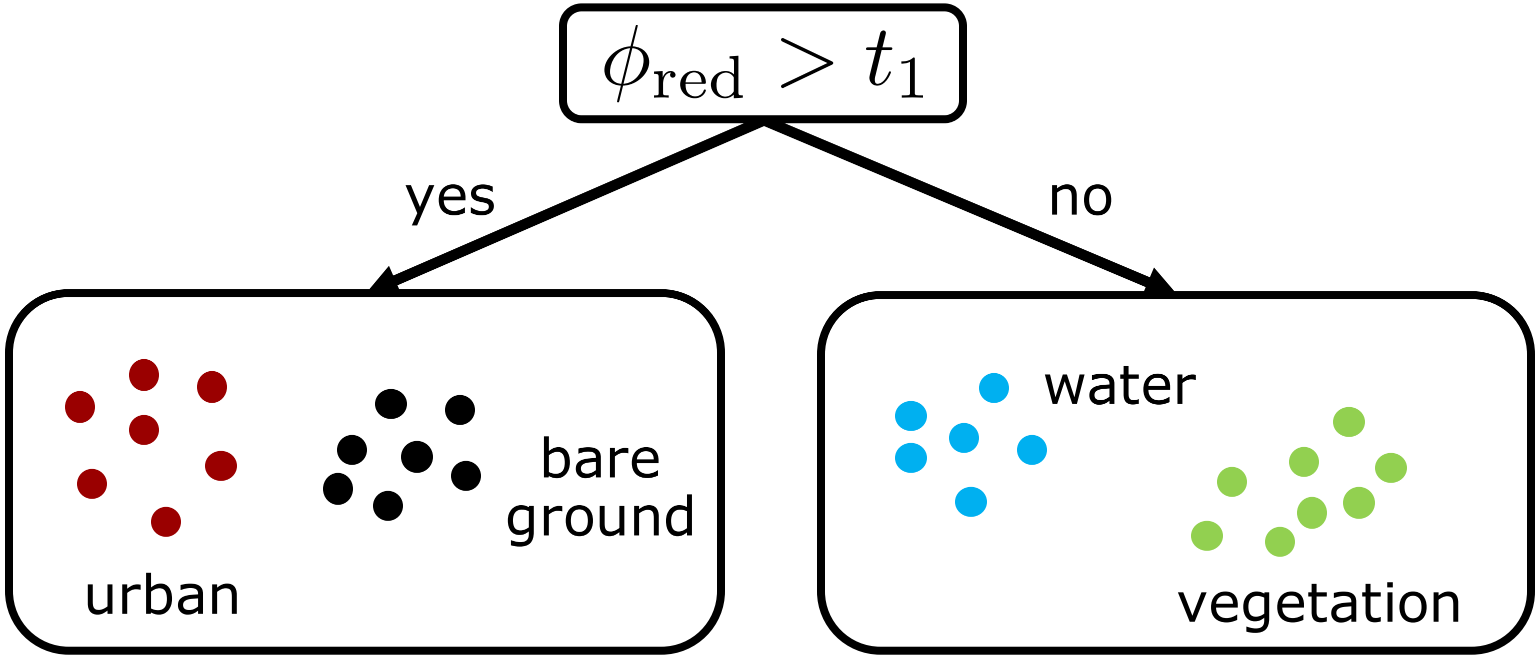
Whole decision process
Formulate each path in the decision tree as a chain of decisions

Feature importance
Calculate how much each feature contributes to the overall model performance

Individual predictions
Determine how much each feature contributed to a single prediction

Purity of a node

Measure of homogeneity with respect to class labels



Minimization of impurity is used to find splits in the tree (the "purer" a node, the better was previous split)

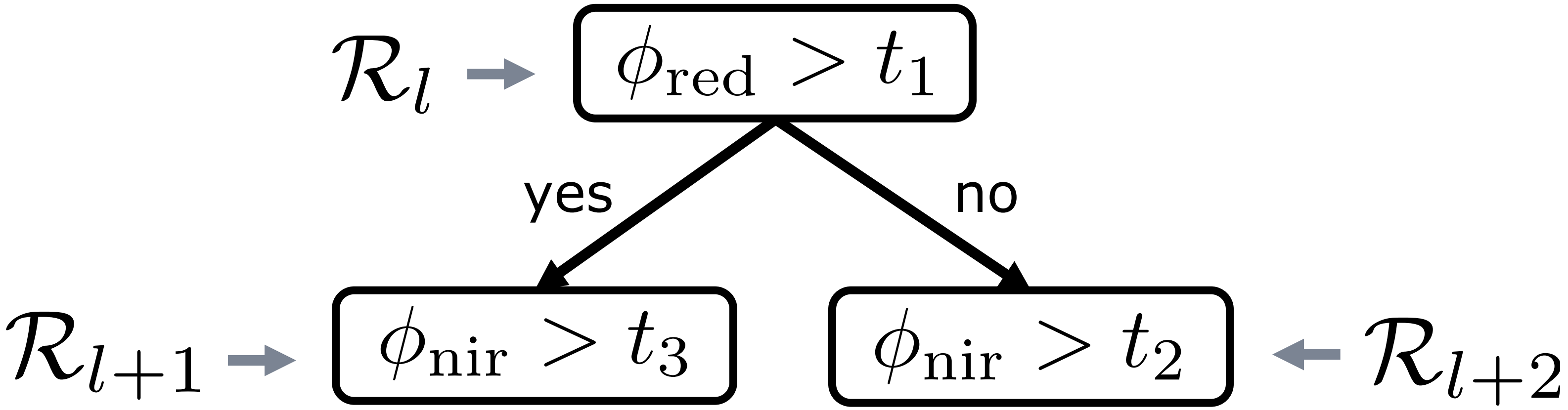
Feature importance

Calculated based on a measure of purity, e.g., entropy

$$h_l = -\sum_k \tilde{p}_{lk} \log \tilde{p}_{lk}$$

$$\tilde{p}_{lk} = \frac{1}{|\mathcal{R}_l|} \sum_{\phi_n \in \mathcal{R}_l} I(c_n = k)$$

Feature importance

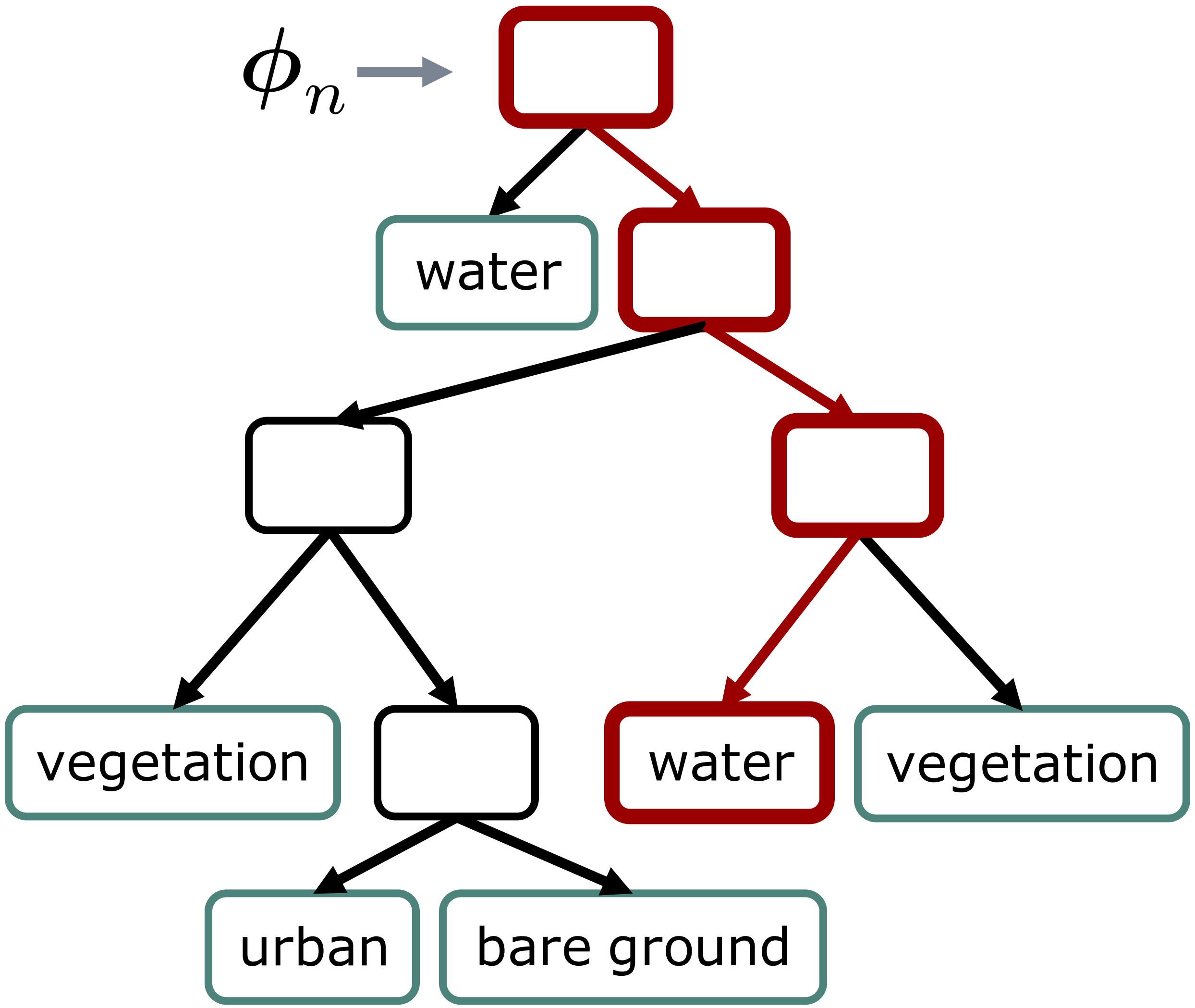


$$\Delta h_l = h_l - \left(\frac{|\mathcal{R}_{l+1}|}{|\mathcal{R}_l|} h_{l+1} + \frac{|\mathcal{R}_{l+2}|}{|\mathcal{R}_l|} h_{l+2} \right)$$

- Sum over all Δh_l that include a specific feature
- Normalize so that all features sum to 1

Contribution of a single prediction

- Feature importance for a specific decision path
- **Feature-wise** calculation by summing up the contributions of each feature from the root to the leaf node
- Contribution can be computed by impurity or other measures



Advantages and disadvantages

- Decision trees can model non-linear relationships (but are inefficient with linear relations)
- Create easy to understand human-friendly interpretations
- Small changes in the input can cause aprupt changes in the outcome
- Small changes in the dataset can cause big changes in the tree architecture

Conclusion

- Inherently interpretable models are only easy to interpret as long as they are small
- High interpretability usually comes at the expense of predictive performance
- No additional tools are necessary

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

- See references in the bottom of the slides
- “Interpretable machine learning” by Christoph Molnar: https://christophm.github.io/interpretable-ml-book