Explainable AI

ML Training Course

Ana Prieto Nemesio & Mariana Clare

ECMWF

Outline

1. What is XAI and why do we care about it?
2. How is explainability measured?
3. Different XAI methods
    I. Model agnostic
    II. Model specific
4. Assessment of XAI Metrics
5. Conclusions

What is XAI and why do we care about it?

eXplainable AI techniques

Explainable

A machine learning method is **explainable** if the reason why it predicted the result can be understood by experts in that field (in our case domain weather and climate scientists)

Motivation

Machine learning methods suffer as decision-making tools because they lack the ability to explain how they reach their prediction, making them potentially untrustworthy.

A method is trustworthy if its results are explainable and interpretable

In the context of weather, in a changing climate, the underlying physics of a problem may alter and it is important to understand whether methods trained on historical data are still fit-for-purpose on future data

Furthermore new laws passed in EU and US say that AI methods must be explainable when used for decision making.

XAI techniques

Suppose we have a neural network which correctly predicts the image below is a horse

Original Image

XAI techniques explain **WHY** the neural network has predicted this image is a horse

Standard LRP

The neural network incorrectly finds the text helpful

Bykov, K., et al. (2020). How Much Can I Trust You?--Quantifying Uncertainties in Explaining Neural Networks. arXiv preprint arXiv:2006.09000.

How is explainability measured?

ECMWF

How is explainability measured?

Explainability Evaluation

Quantitative

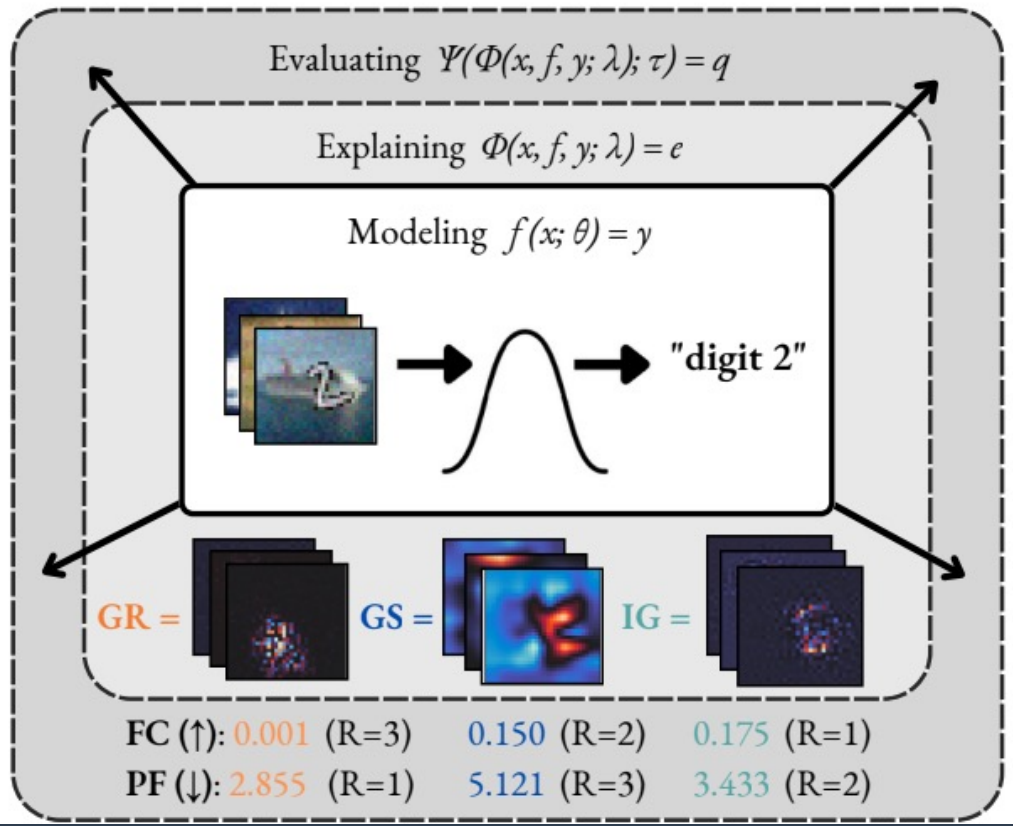
- Faithfulness
- Robustness
- Localisation
- Complexity
- Axiomatic
- Randomisation

Qualitative

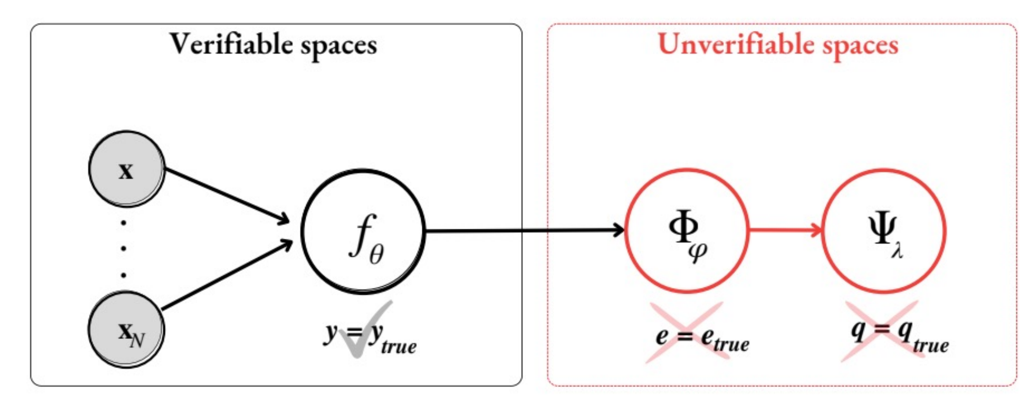
- Visual Inspection

How is explainability measured?

Quantitative → **Challenging**



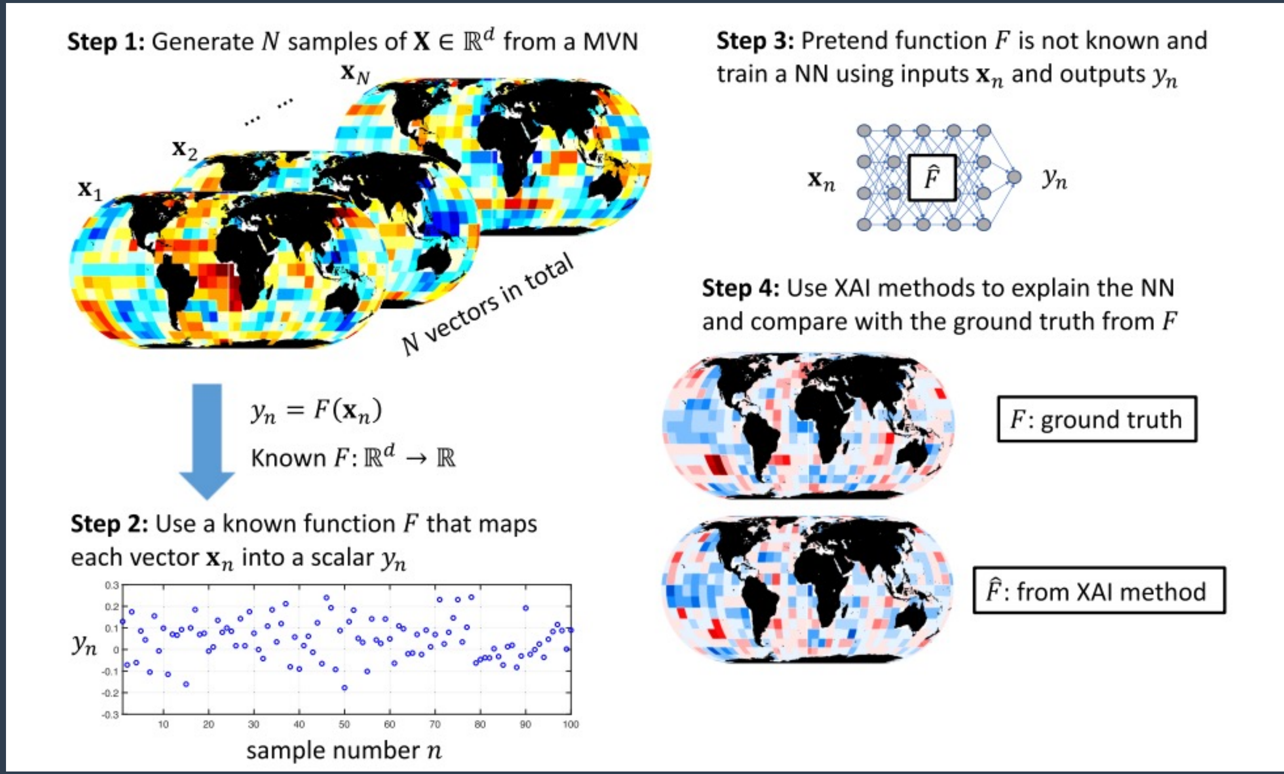
The Meta-Evaluation Problem in Explainable AI: Identifying Reliable Estimators with MetaQuantus, Hedström.A , Bommer.P et. al



ECMWF

How is explainability measured?

Quantitative → **Attribution benchmark datasets**



Neural Network Attribution Methods for Problems in Geoscience: A Novel Synthetic Benchmark Dataset. Mamalakis A., Ebert-Uphoff I, Barnes E.A.

ECMWF

Different XAI Methods

ECMWF

XAI Methods

Stage

- **Post-hoc**: Model analysed after being trained
- **Intrinsic**: Model is explainable by itself

Applicability

- **Model-Agnostic**: Applicable to any model
- **Model-Specific**: Applicable to a specific model

Scope

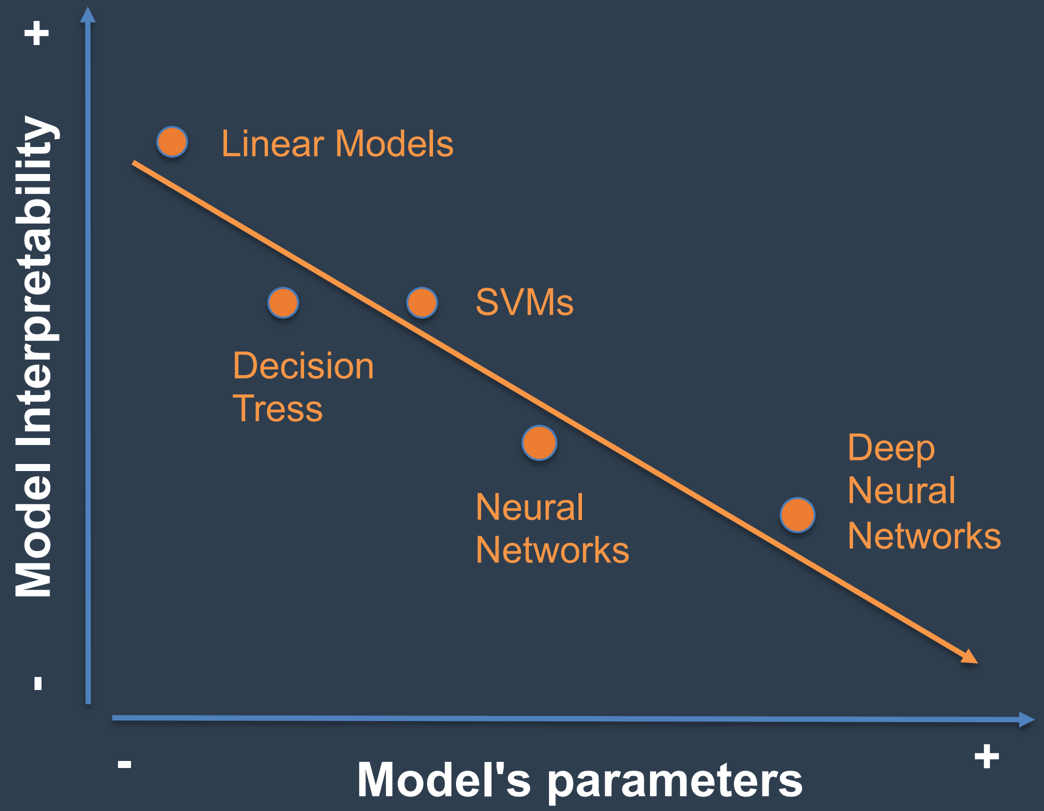
- **Global**: Focus on the model – understanding of the decision process
- **Local**: Focus on the data– individual explanations

XAI Methods Taxonomy

Stage

Post-hoc → "Black-box" models

Intrinsic → Linear Models
Shallow Decision Trees



ECMWF

XAI Methods Taxonomy

Post-Hoc

"Black-box" models

Intrinsic

Linear/Logistic Regression
Shallow Decision Trees

Model-Agnostic

Global

Activation-
Maxisimisation
PDP
Permutation
Feature
Importance

Local

SHAP
LIME
Occlusion
Sensitivity

Model-Specific

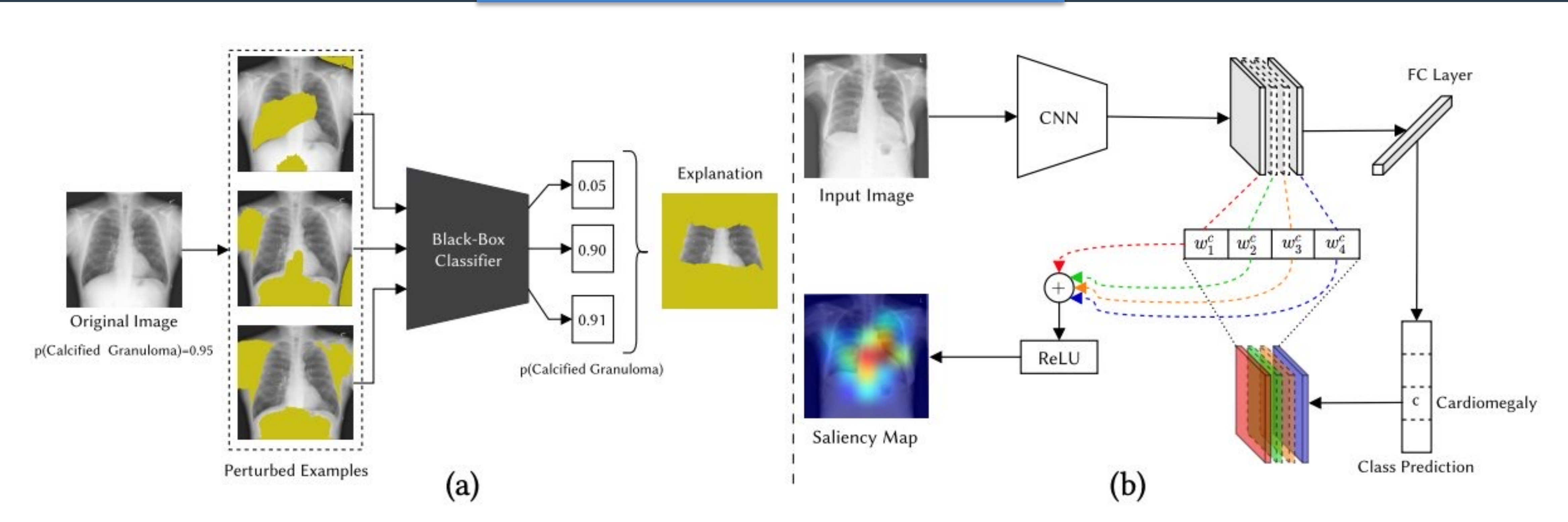
Saliency-Maps
Attention Visualization
LRP

Local

https://courses.minnalearn.com/en/courses/trustworthy-ai/preview/explainability/types-of-explainable-ai/

ECMWF

Attribution-based Methods



Perturbation-based Method (Occlusion)

Gradient-Based Methods

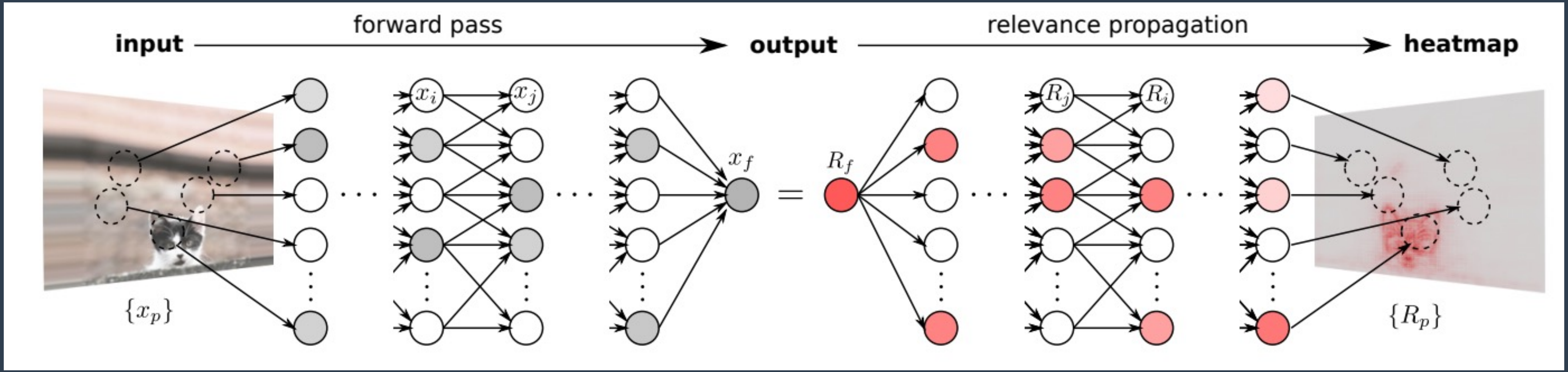
Explainable Deep Learning Methods in Medical Image Classification: A Survey
arXiv:2205.04766

Model Specific

ECMWF

https://github.com/albermax/innvestigate

LRP (Layerwise Relevance Propagation)



Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K. R. (2017). Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern recognition*, *65*, 211-222.

Using LRP

There is more than one LRP rule to compute the relevance on each node e.g.

Relevance for new layer

Relevance for previous layer

LRP-0

$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k$$

LRP-ε

$$R_j = \sum_k \frac{a_j w_{jk}}{\epsilon + \sum_{0,j} a_j w_{jk}} R_k$$

LRP-γ

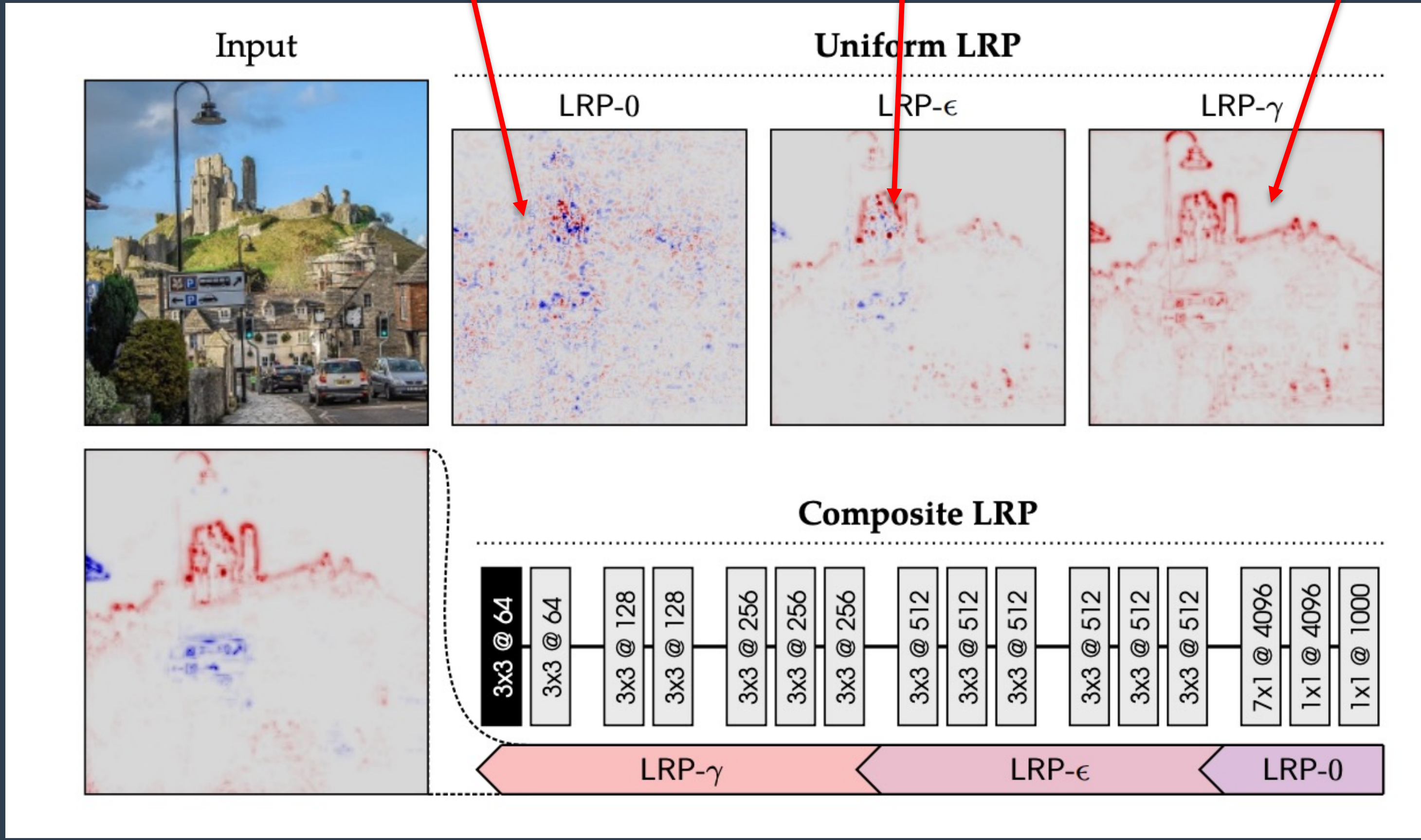
$$R_j = \sum_k \frac{a_j \cdot (w_{jk} + \gamma w_{jk}^{+})}{\sum_{0,j} a_j \cdot (w_{jk} + \gamma w_{jk}^{+})} R_k$$

Using LRP

Gradient Shattering

Removes noise and faithful explanation but too sparse

Densely highlights features but picks unrelated concepts



LRP composite combines approaches and overcomes these disadvantages

Montavon, Grégoire, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. "Layer-wise relevance propagation: an overview." *Explainable AI: interpreting, explaining and visualizing deep learning* (2019): 193-209.

Advantages and Disadvantages of LRP

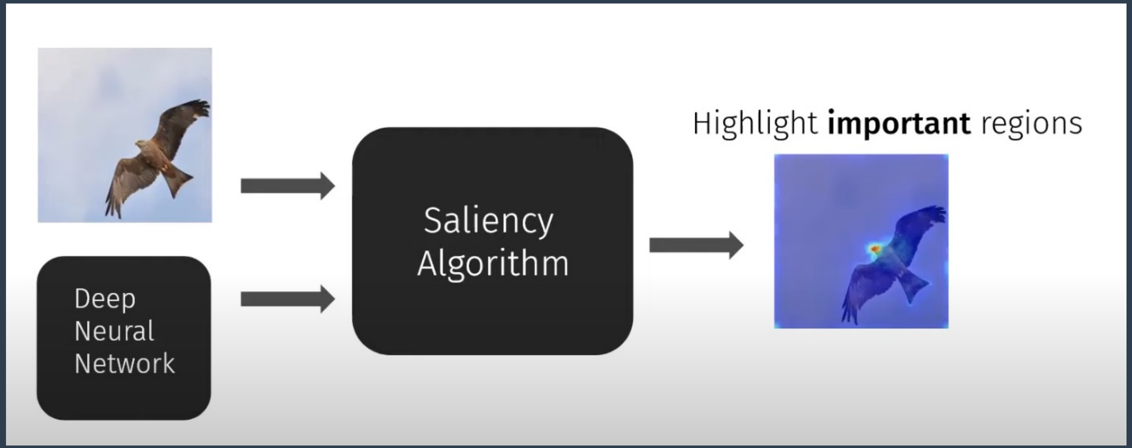
Advantages

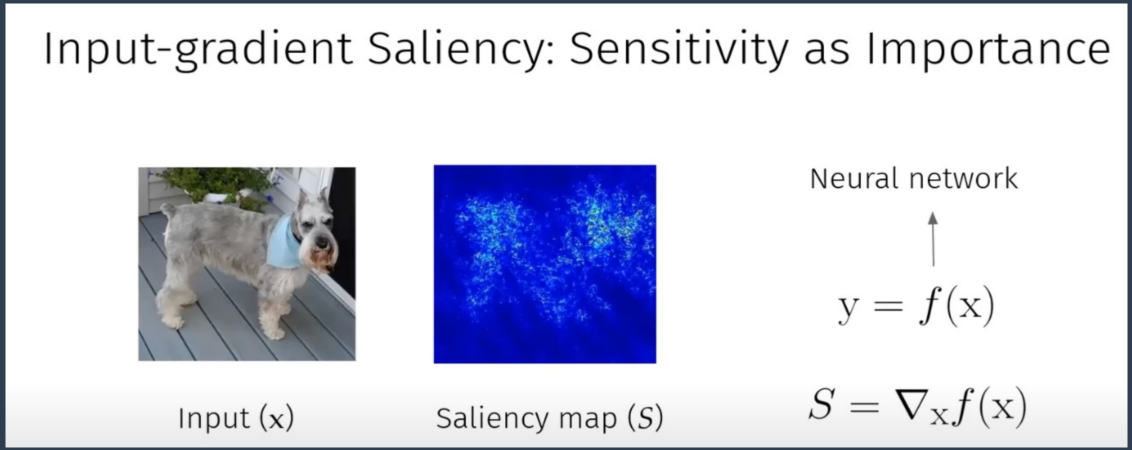
- Calculates relevance for all outputs jointly meaning relatively inexpensive
- Creates easily interpretable maps of relevance

Disadvantages

- Can be tricky to apply as the right combination of relevance formulae must be found and there are hyperparameters to tune
- Can only be applied to neural networks

Saliency Maps (Gradient-based Methods)

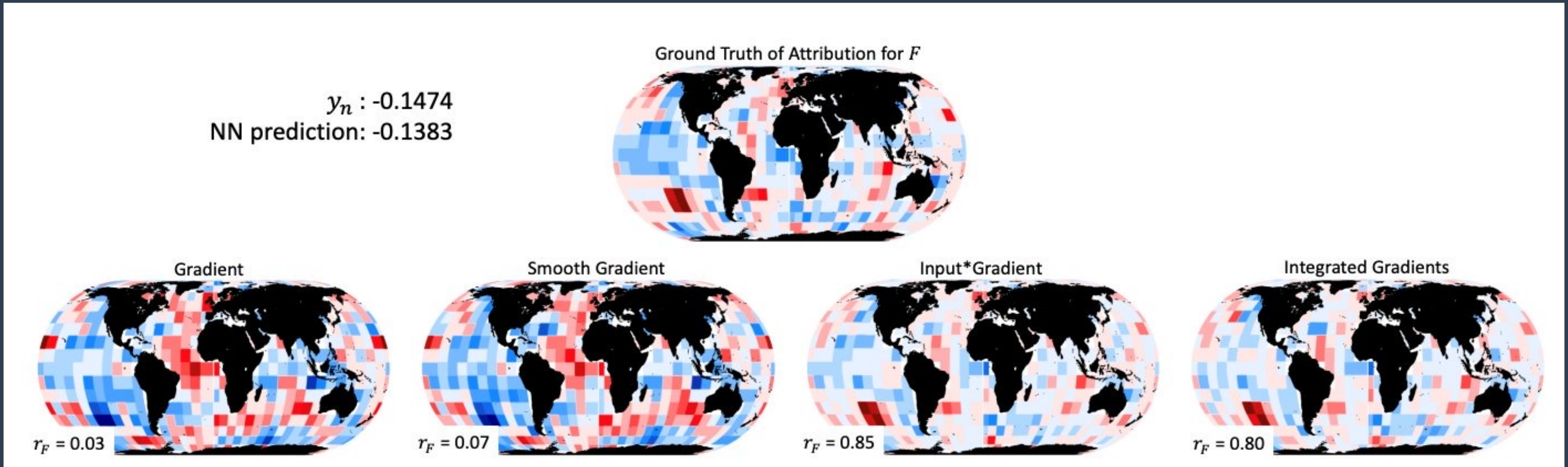




Pitfalls of Saliency Map Interpretation in Deep Neural Networks - Suraj Srinivas

Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps

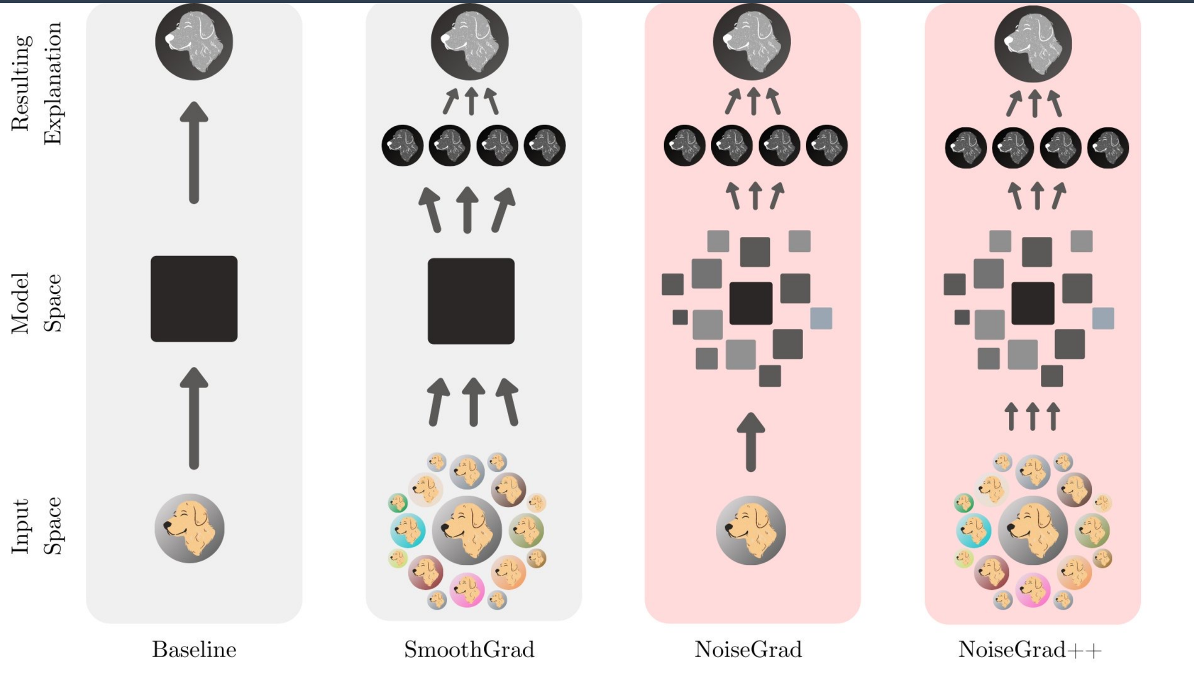
Saliency Maps



Neural Network Attribution Methods for Problems in Geoscience: A Novel Synthetic Benchmark Dataset. Mamalakis A., Ebert-Uphoff I, Barnes E.A.

Saliency Maps

Smooth Gradients



$$\Phi_{SG}(f_c, x) = \mathbb{E}_{\varepsilon \sim \mathcal{N}(0,\sigma^2 \boldsymbol{I})}\left[\Phi(f_c, x+\varepsilon)\right]$$

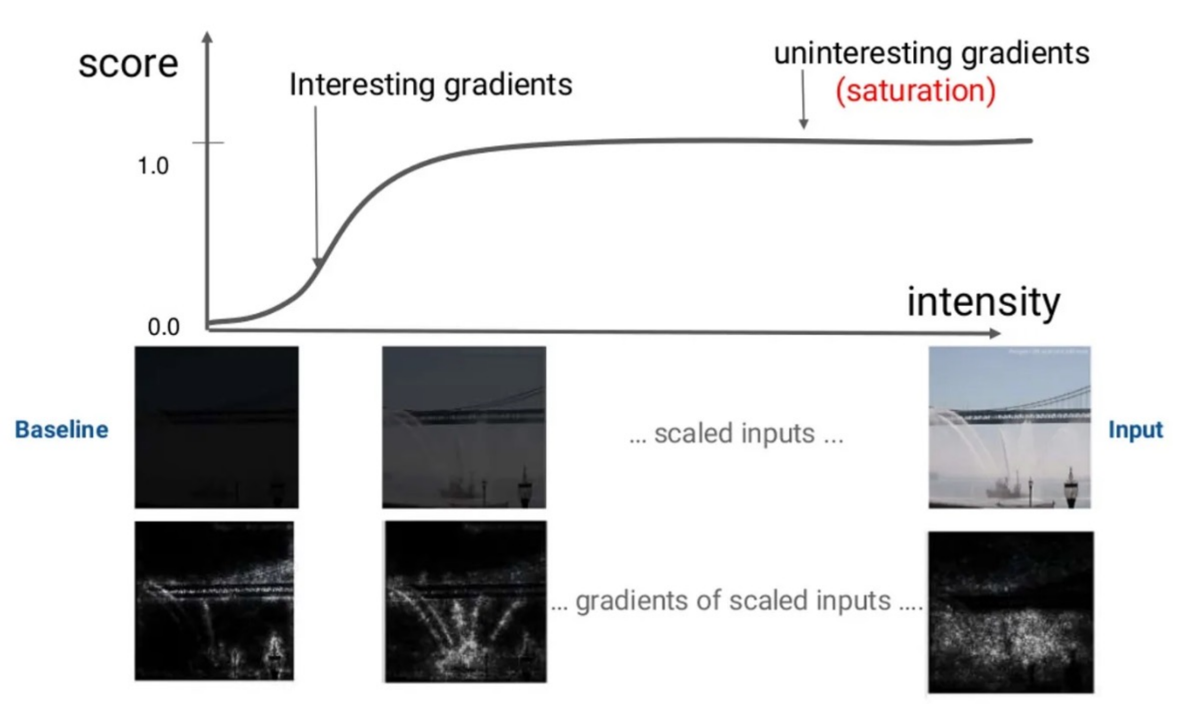
NoiseGrad — Enhancing Explanations by Introducing Stochasticity to Model Weights
arXiv:2106.10185v3

Saliency Maps

Input x Gradient

$$\Phi_{InputXGrad}\left(f_c, x\right) = x \odot \nabla f_c\left(x\right)$$

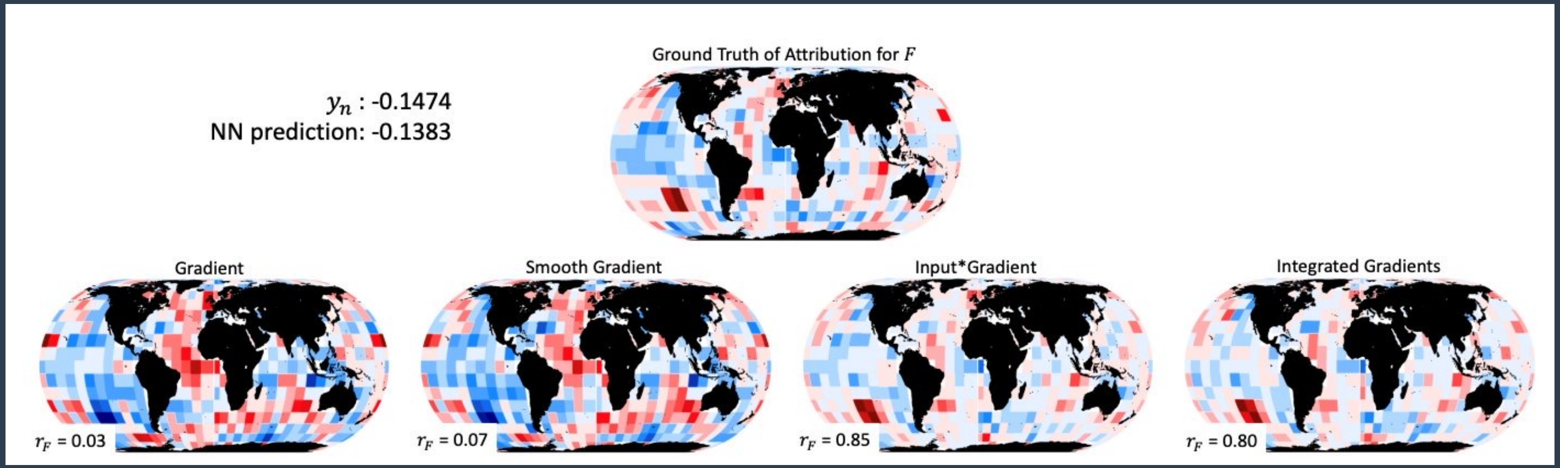
Integrated Gradients



$$\Phi^d_{IG}(f_c, x) = (x_d - x'_d) \times \int_0^1 \frac{\partial f_c(x)}{\partial x_d}\Big|_{x=x'+a(x-x')} da \quad \forall\, d \in \{1, \cdots, D\}$$

Axiomatic Attribution for Deep Networks
arxiv.org/abs/1703.01365

Saliency Maps



Sensitivity

Vanilla Gradients

Smooth Gradients

Attribution

Integrated Gradients

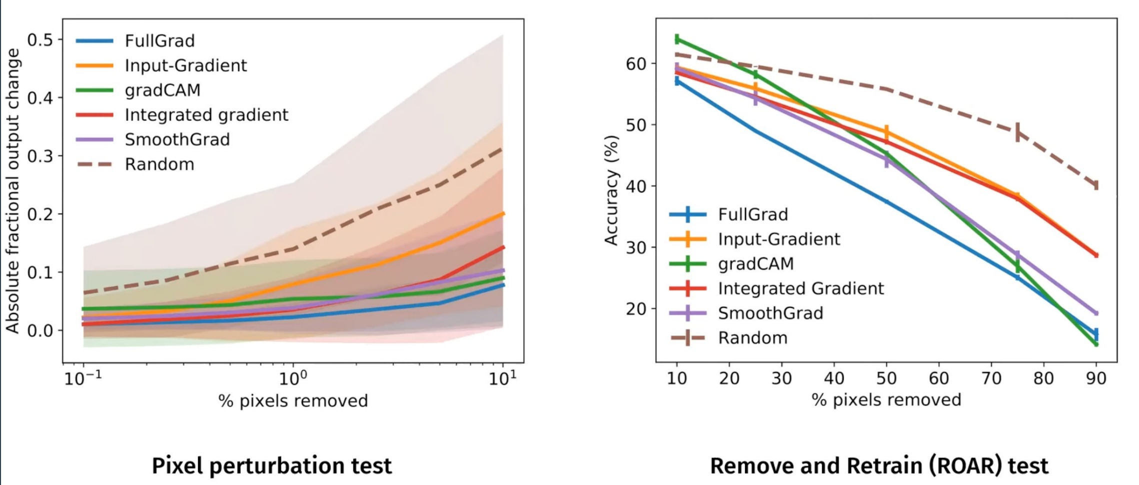
Input x Gradient

Neural Network Attribution Methods for Problems in Geoscience: A Novel Synthetic Benchmark Dataset. Mamalakis A., Ebert-Uphoff I, Barnes E.A.

Advantages and Disadvantages of Gradient-Based Methods

Advantages

- Computationally fast
- Generated explanation maps are robust in terms of input perturbation



Pitfalls of Saliency Map Interpretation in Deep Neural Networks - Suraj Srinivas

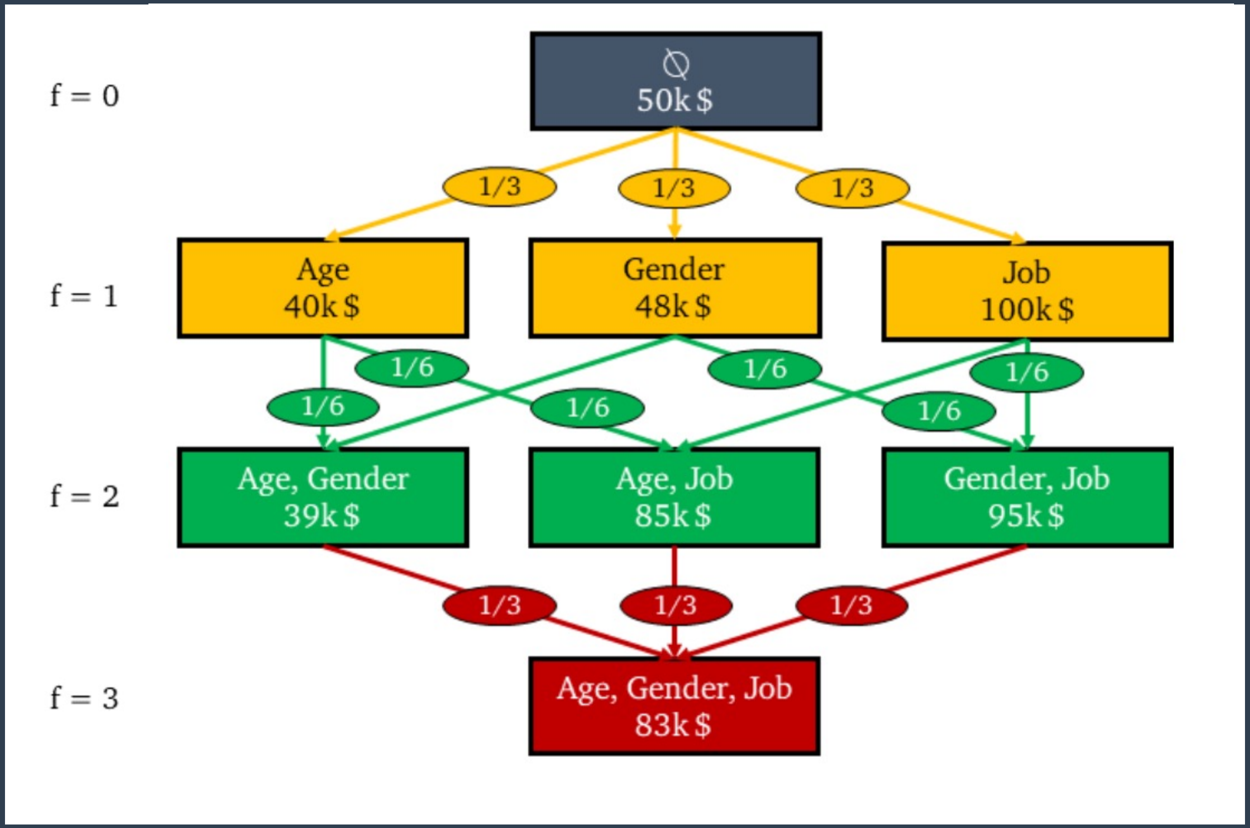
Disadvantages

- Applied to just a single example or few examples – results obtained may be too brittle and could lead to a false conclusion about the performance of the model
- There is no 'one size fits all' gradient-method (class invariant, input transformation, etc)
- Difficult to quantitatively evaluate

Model Agnostic

SHAP (SHapley Additive explanation) values

Suppose we want to calculate the SHAP value of the input age for a given prediction of salary



All possible combinations of input features to be included in the model

Marginal contribution for this output from adding age only

$$MC_{Age,\{Age\}}(x_0) = Predict_{\{Age\}}(x_0) - Predict_{\varnothing}(x_0) = 40k\$ - 50k\$ = -10k\$$$

SHAP value – sum of weighted marginal contributions from adding age in each feature combination

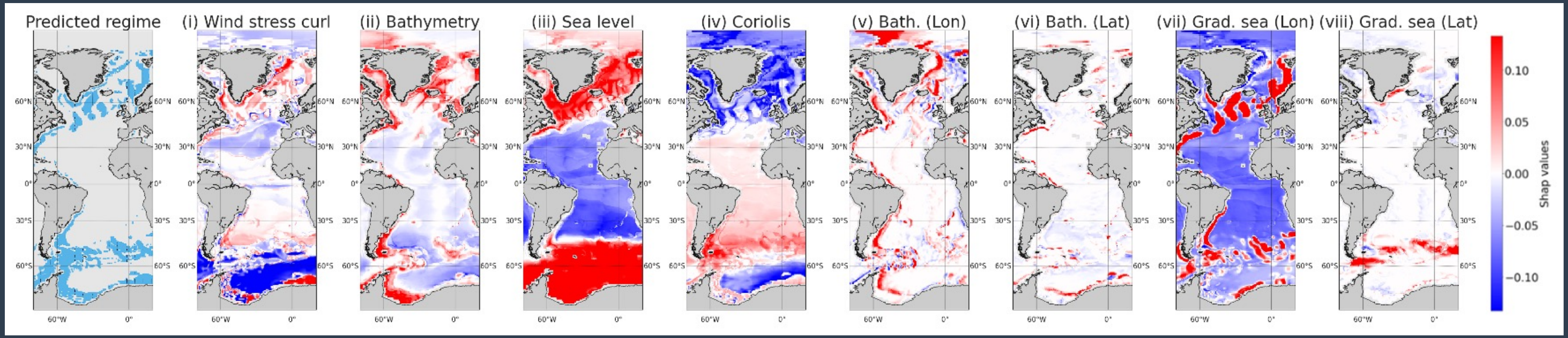
$$\begin{aligned} SHAP_{Age}(x_0) = {} & [(1\times\binom{3}{1})]^{-1}\times MC_{Age,\{Age\}}(x_0) \; + \\ & [(2\times\binom{3}{2})]^{-1}\times MC_{Age,\{Age,Gender\}}(x_0) \; + \\ & [(2\times\binom{3}{2})]^{-1}\times MC_{Age,\{Age,Job\}}(x_0) \; + \\ & [(3\times\binom{3}{3})]^{-1}\times MC_{Age,\{Age,Gender,Job\}}(x_0) \; + \\ = {} & \frac{1}{3}\times(-10k\$) + \frac{1}{6}\times(-9k\$) + \frac{1}{6}\times(-15k\$) + \frac{1}{3}\times(-12k\$) \\ = {} & -11.33k\$ \end{aligned}$$

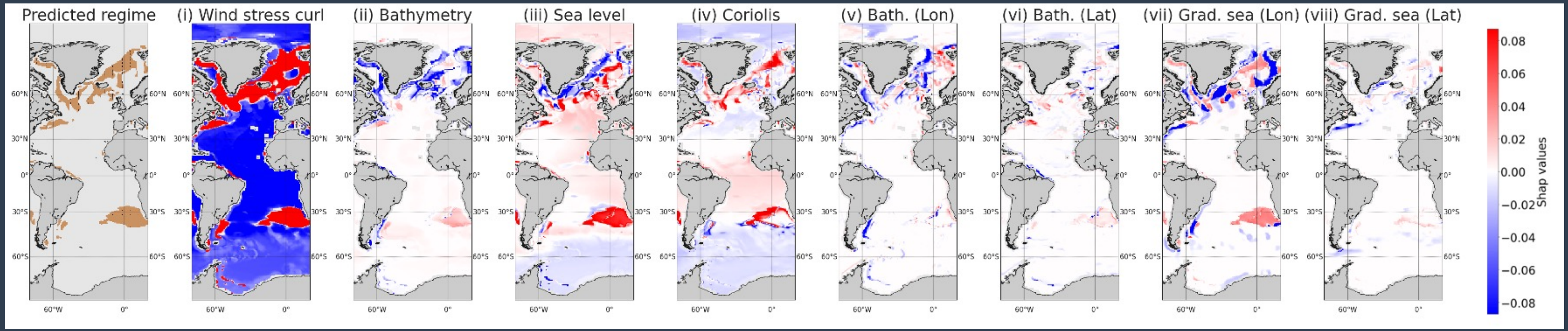
https://towardsdatascience.com/shap-explained-the-way-i-wish-someone-explained-it-to-me-ab81cc69ef30

https://github.com/slundberg/shap

SHAP values

SHAP sees problem as binary for each output: including a feature either increases the probability of the specific output being considered there or decreases it.
Therefore have many more values than e.g. LRP and takes much longer to compute

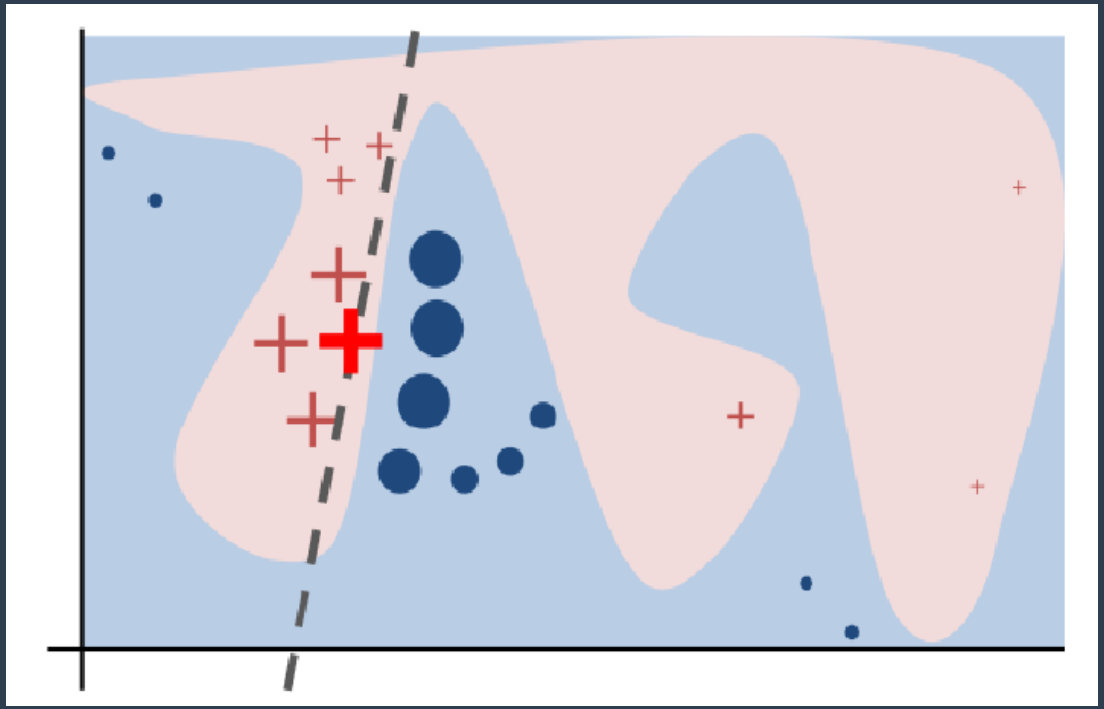




Clare, M. C., Sonnewald, M., Lguensat, R., Deshayes, J., & Balaji, V. (2022). Explainable artificial intelligence for Bayesian neural networks: Toward trustworthy predictions of ocean dynamics. *Journal of Advances in Modeling Earth Systems*, *14*(11), e2022MS003162.

Kernel SHAP

LIME + SHAP



LIME

- Create a set of coalition vectors based on the features. If features have a corresponding value of 1 in the vector, they are replaced in the vector by their actual values, and if they have a corresponding value of 0, replaced by different feature values.
- Weight of each feature is calculated and a linear model is trained (LIME).
- Coefficient values of the linear model correspond to Shapley Values for each feature.

Advantages and Disadvantages of SHAP

Advantages

- Measures the impact of each feature on model predictions which is helpful for feature engineering and model optimisation, as well as showing potential biases
- Can be applied to any type of model

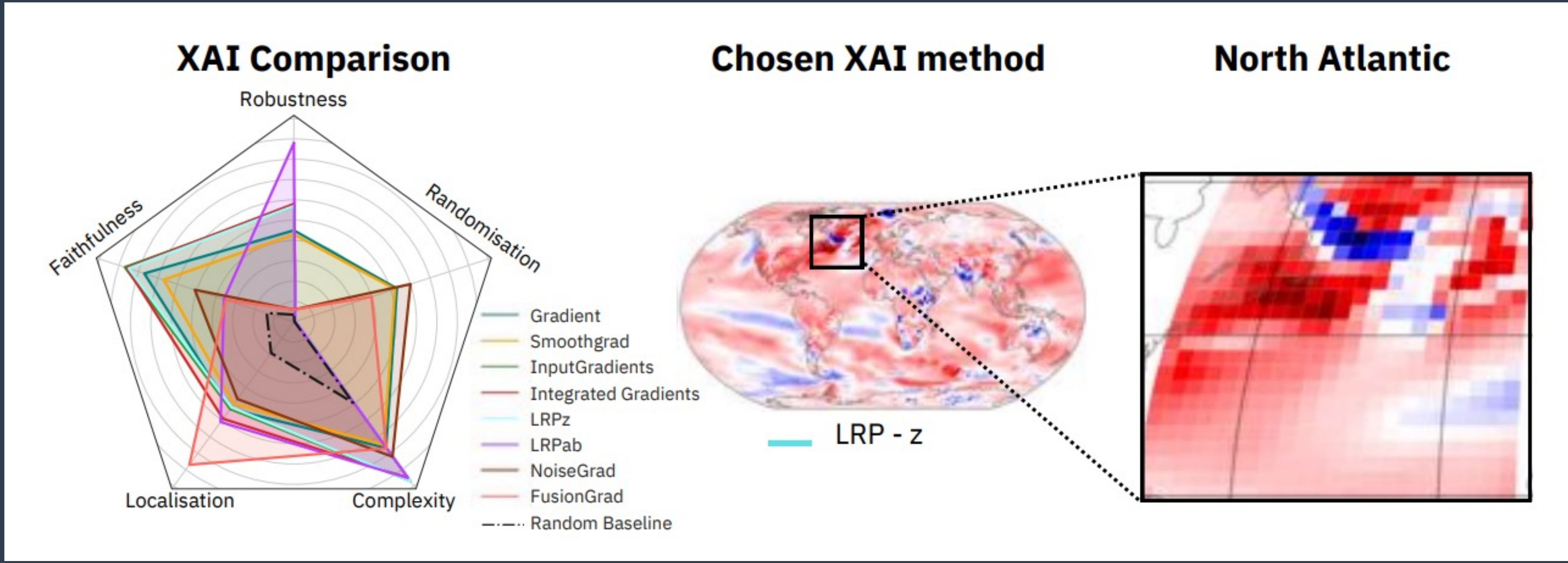
Disadvantages

- Computationally expensive to compute and provides a lot of information to the user which can be difficult to digest
- *Correlation does not imply causation:* SHAP shows relationships between variables but does not explain their causal nature

Assessment of XAI metrics

Key components to measure explainability

Faithfulness | Robustness | Localisation | Complexity | Randomisation

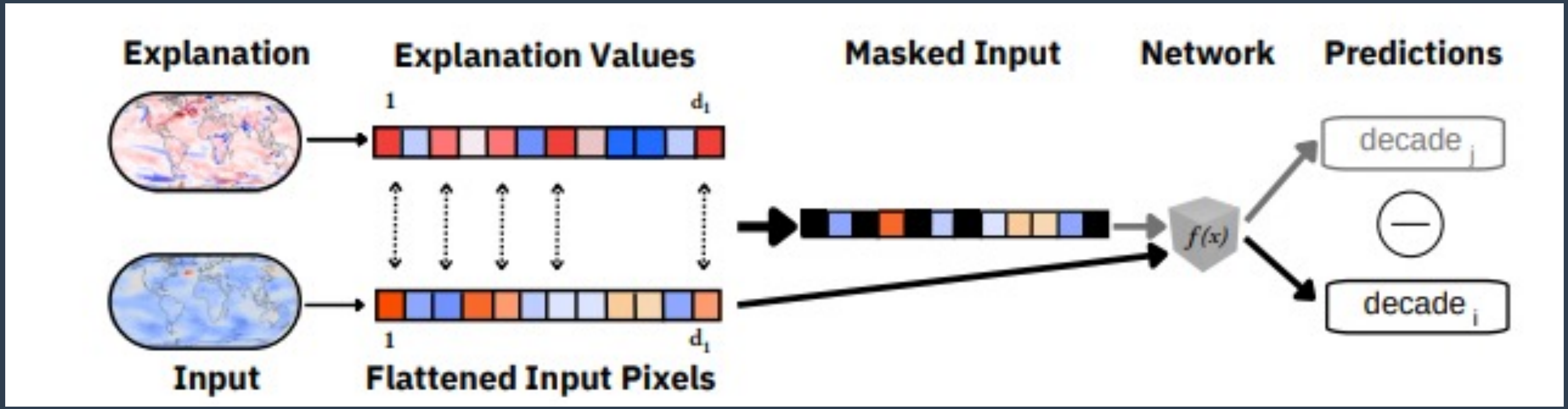


Throughout this section following from:
Bommer, P., Kretschmer, M., Hedström, A., Bareeva, D., & Höhne, M. M. C. (2023). Finding the right XAI method--A Guide for the Evaluation and Ranking of Explainable AI Methods in Climate Science. *arXiv preprint arXiv:2303.00652*.

Key components to measure explainability

Faithfulness [redacted] [redacted] [redacted] [redacted]

Measures whether a feature that the XAI method assigned high relevance actually changes the prediction

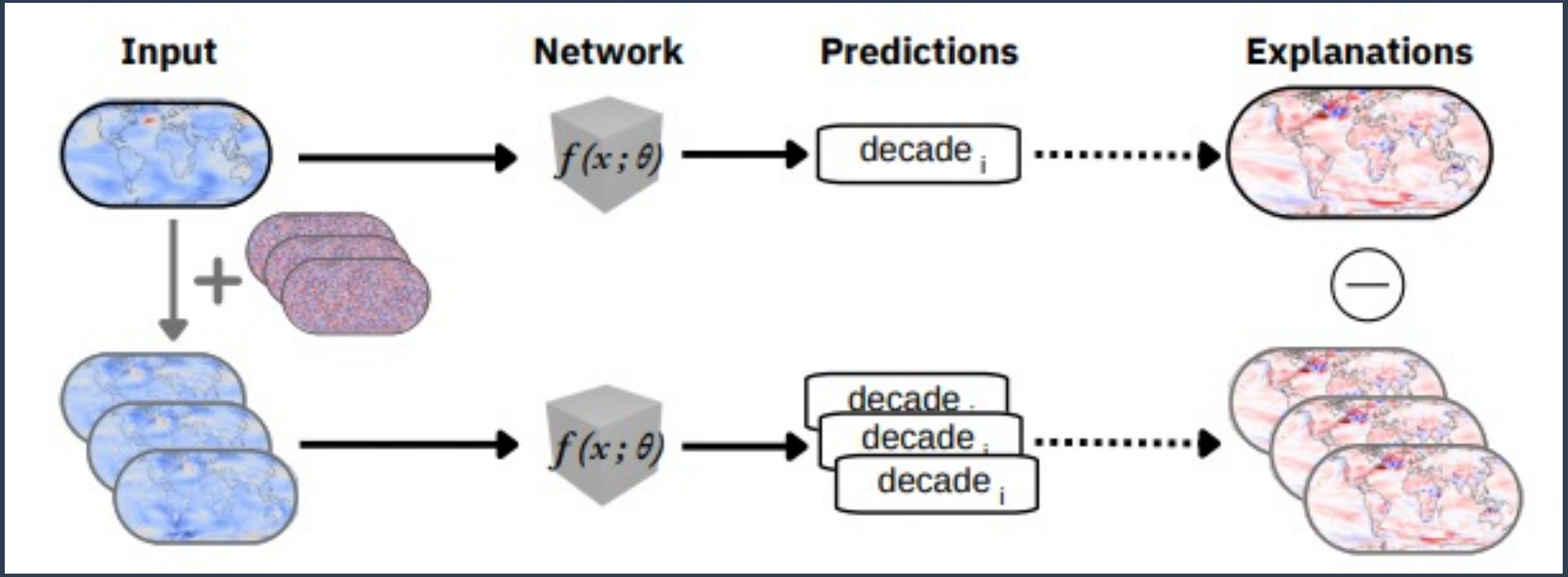


If the masking is based on a faithful feature, then prediction of masked input should be different to prediction of full input

Key components to measure explainability

[redacted] | **Robustness** | [redacted] | [redacted] | [redacted]

Measure stability of an explanation with respect to small changes in the input x + δ
i.e. perturb inputs and compare explanation maps

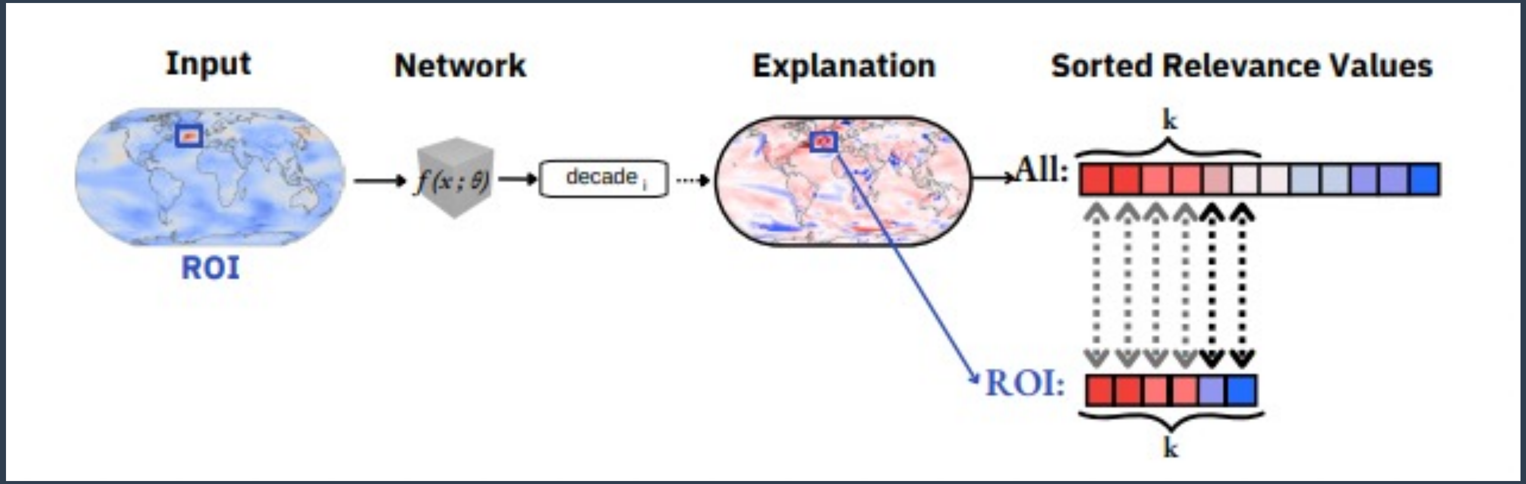


If the method is robust, the difference between the perturbed and unperturbed explanations should be small

Key components to measure explainability

[redacted] [redacted] **Localisation** [redacted] [redacted]

Measures quality of an explanation based on user-defined region of interest

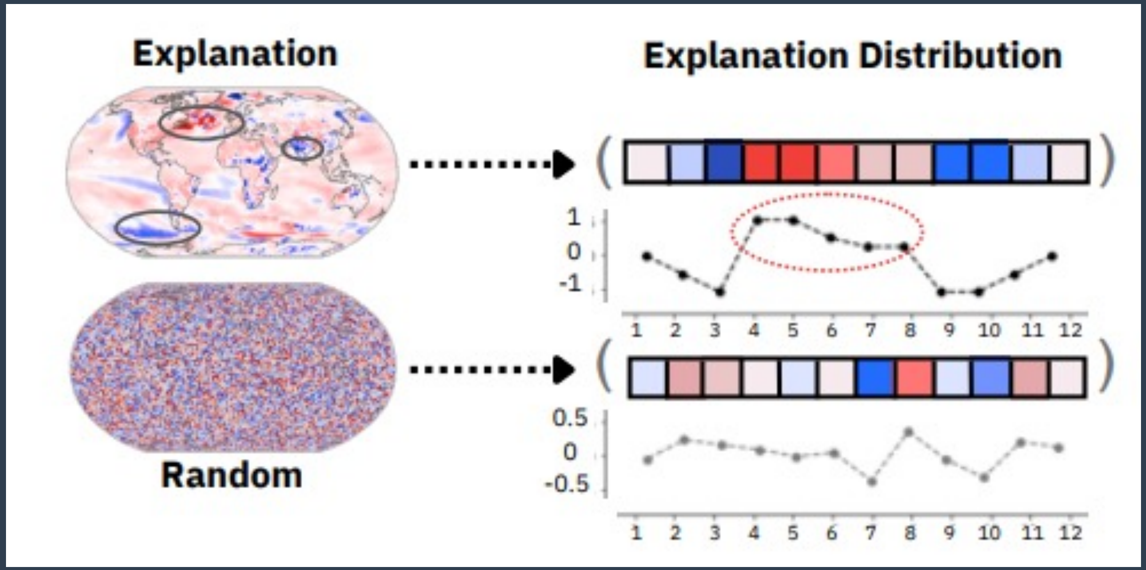


If the method has strong localisation, the explanation values for the ROI should be the highest values of the sorted explanation values across all pixels

Key components to measure explainability

[redacted] [redacted] [redacted] **Complexity** [redacted]

Measures conciseness of an explanation i.e. should consist only of a few strong features

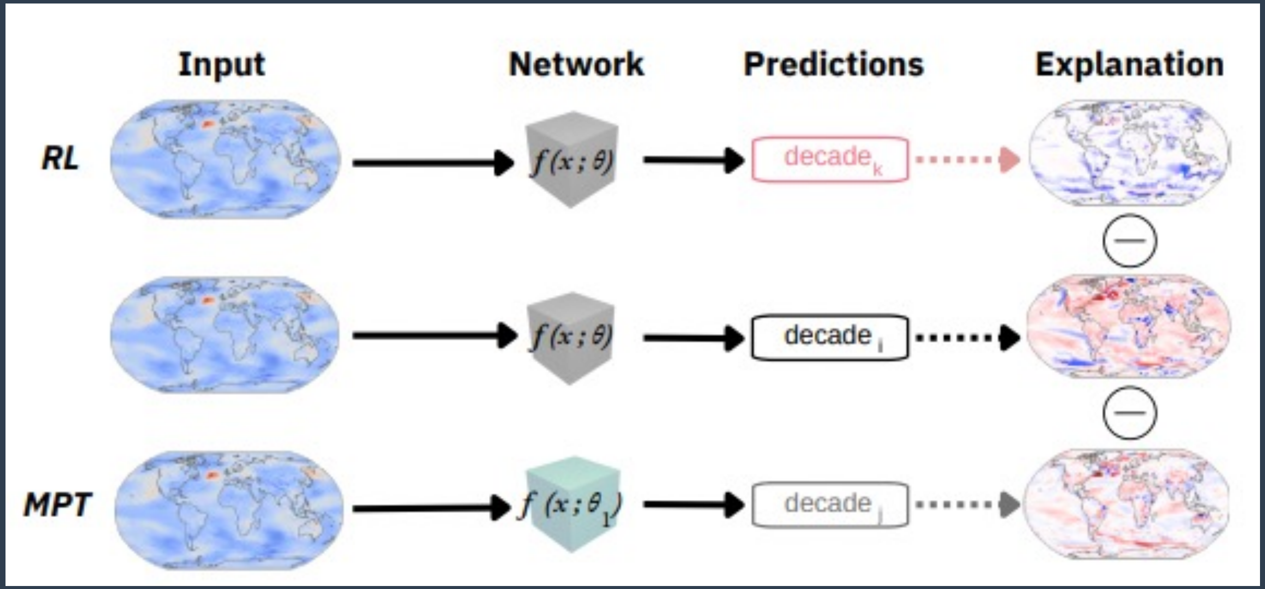


Explanation distribution should have clear max/min compared to sampling from uniform distribution

Key components to measure explainability

[redacted] [redacted] [redacted] [redacted] Randomisation

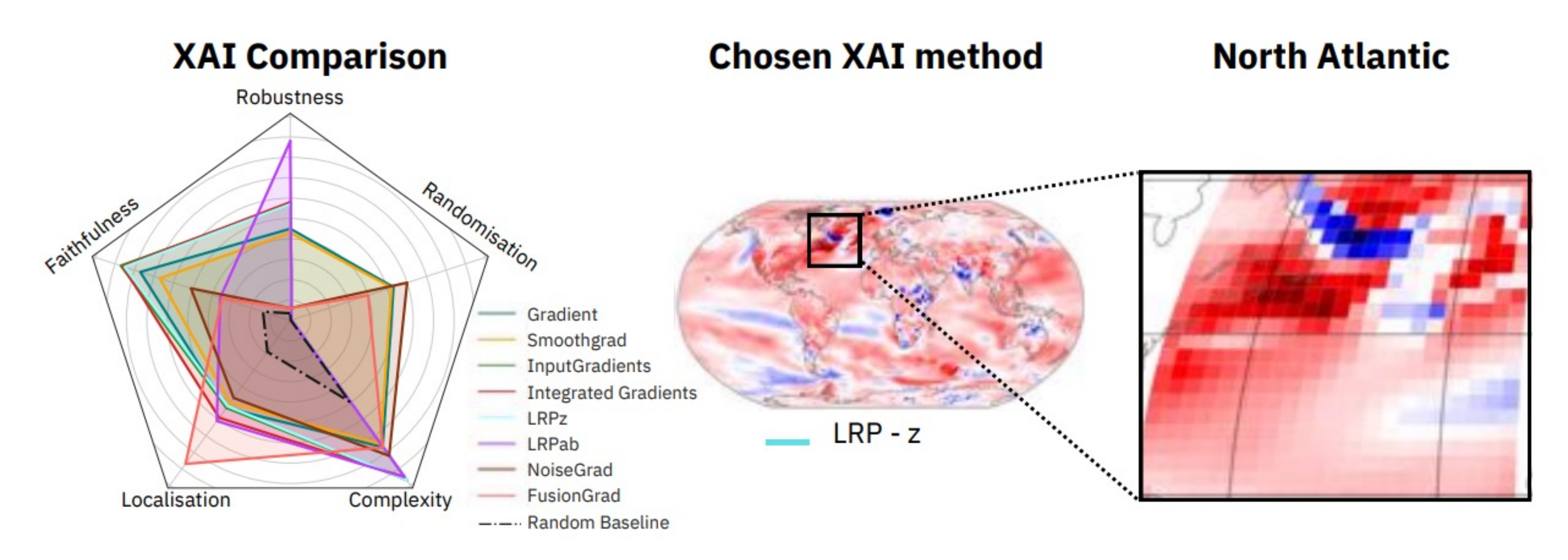
Measures effect on the explanation of a random perturbation scenario



Explanation should differ when random parameters are perturbed or noise is added

Key components to measure explainability

Faithfulness | Robustness | Localisation | Complexity | Randomisation



Bommer, P., Kretschmer, M., Hedström, A., Bareeva, D., & Höhne, M. M. C. (2023). Finding the right XAI method--A Guide for the Evaluation and Ranking of Explainable AI Methods in Climate Science. *arXiv preprint arXiv:2303.00652*.

ECMWF

How is explainability measured?

Quantitative

Faithfulness	Robustness	Localisation	Complexity	Randomisation	Axiomatic
• Faithfulness Correlation • Faithfulness Estimate • Pixel-Flipping • Region segmentation • Monotonic-Arya • Monotonic-Nguyen • Selectivity • Sensitivity • IROF • Infidelity • ROAD • Sufficiency	• Local Lipschtiz Estimate • Max-Sensitivity • Avg-Sensitivity • Continuity • Input independence Rate • Consistency • Relative Input Stability • Relative Output Stability • Relative Representation Stability	• Pointing Game • Attribution Localisation • TKI • Relevance Rank Accuracy • Relevance Mass Accuracy • AUC	• Sparseness • Complexity • Effective Complexity	• Model parameter Randomisation • Random Logit	• Completeness • Non-sensitivity • Input variance

Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond
Hedström.A , Bommer.P et. al

ECMWF

Conclusions "XAI is the answer"

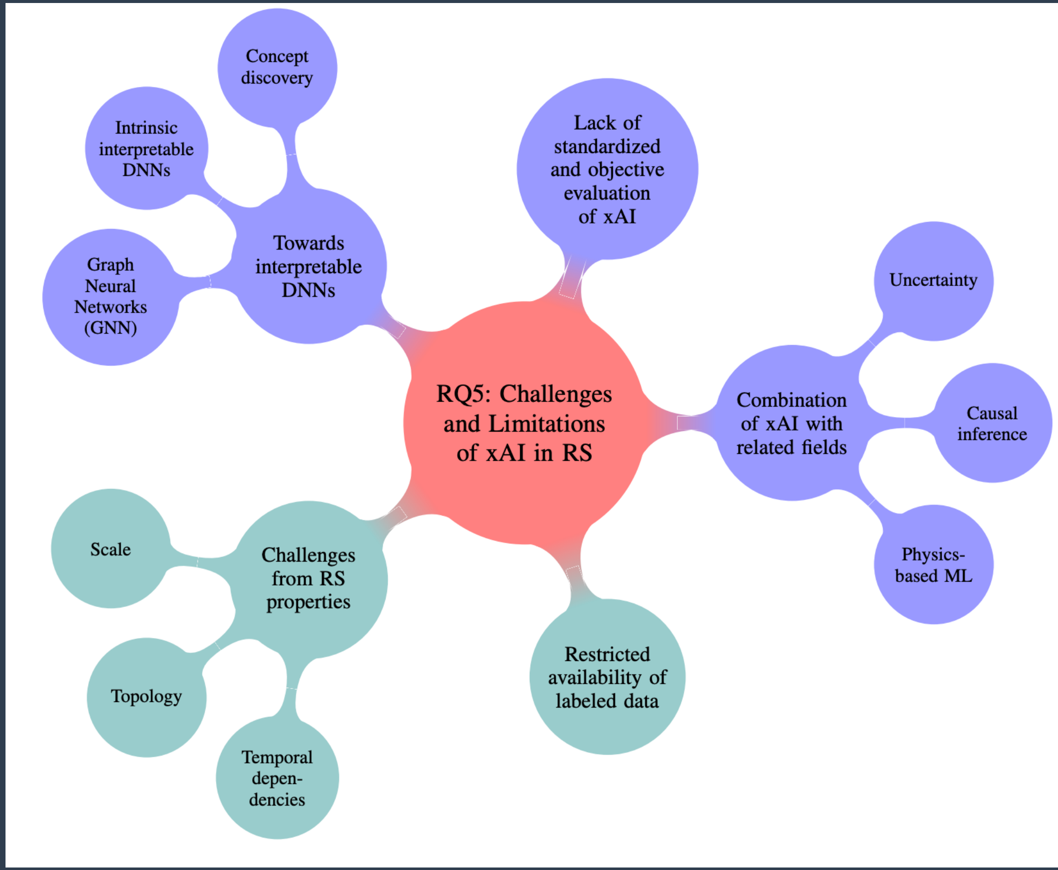
Conclusions

- Active research field –> new models being developed and frameworks that facilitate its usage (SHAP, CAPTUM, QUANTUS, iNNvestigate,IntepretDL)

- XAI quantification and evaluation –> more complex due to the lack of "ground truth" in the explanation space

- Applicability of XAI methods – more commonly seen in classification problems – further development needed to understand how to apply it to regression problems

- Model specific methods – new methods being developed to 'catch-up' with latest DL architectures based on transformers or graphs. Also worth pointing out – methods applicable also to other types of data like NLP



Opening the Black-Box: A Systematic Review on Explainable AI in Remote Sensing
arXiv:2402.13791

ECMWF

XAI for Regression (XAIR)

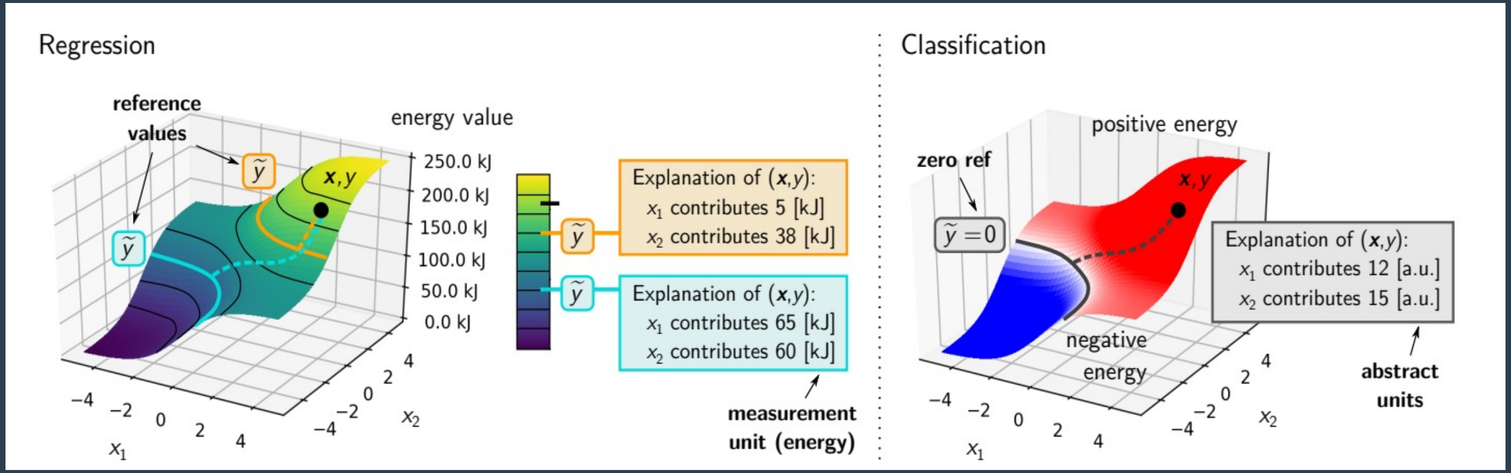
Challenges

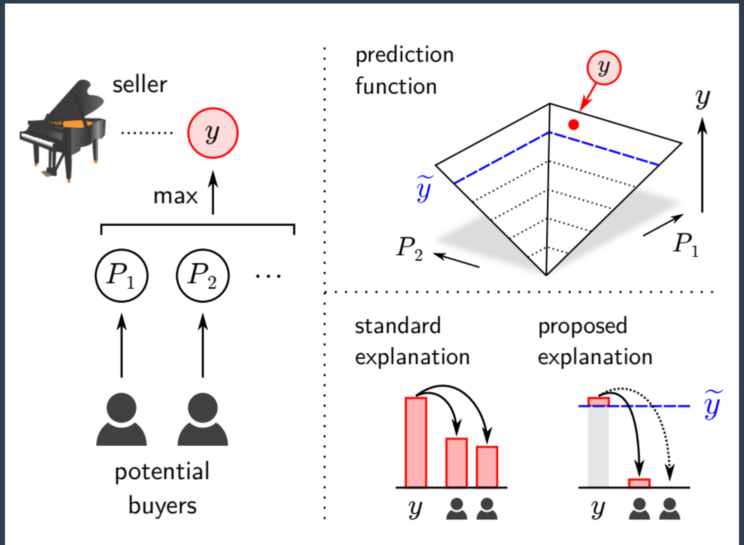
Regression → Quantities with Units (Physical Meaning); Fixed Baseline/Reference Scenario

Proposed Solutions

user-provided reference values

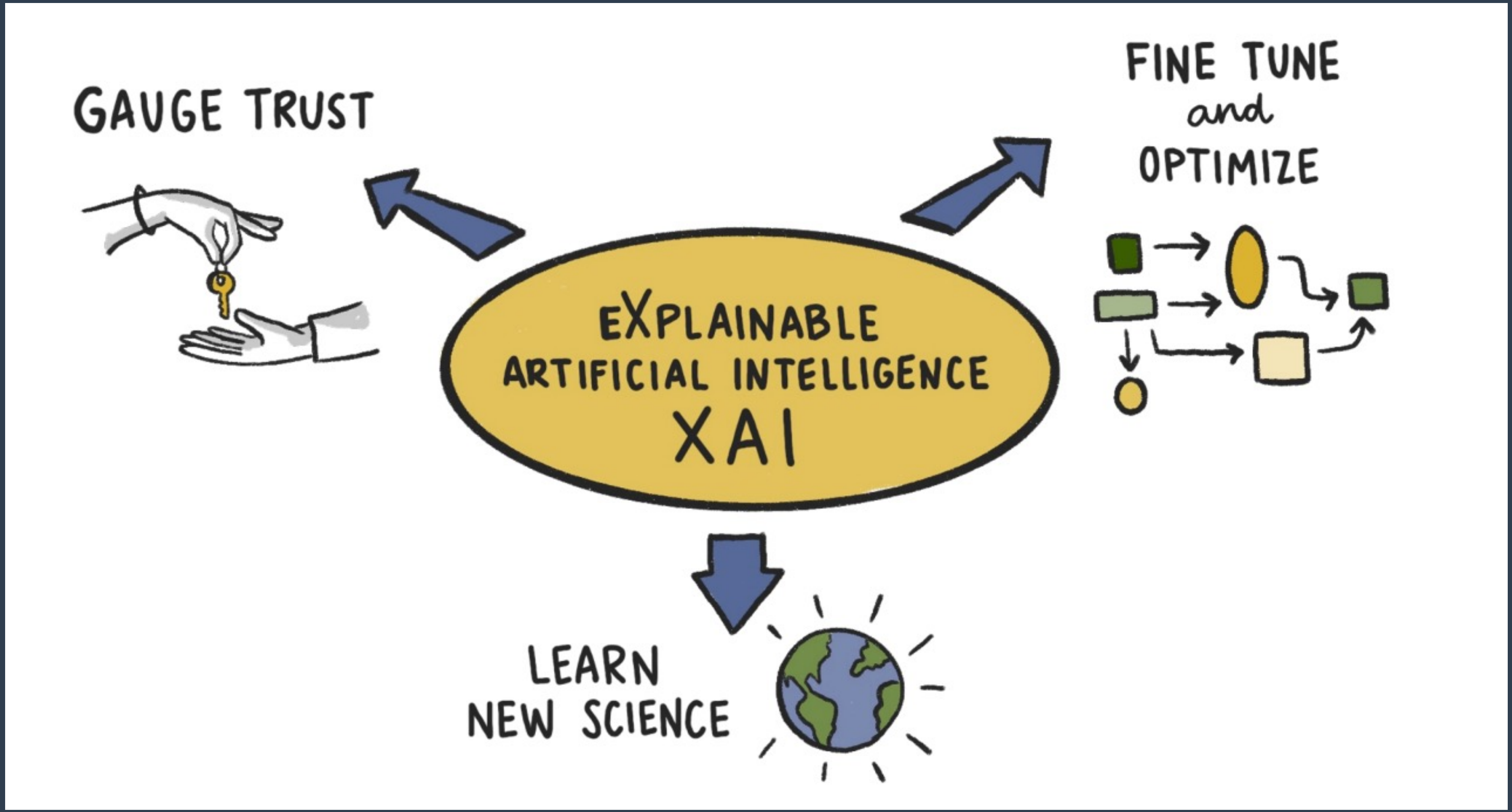
$$g(\boldsymbol{x}) = f(\boldsymbol{x}) - \tilde{y}$$





Toward Explainable AI for Regression Models
arXiv:2112.11407

Take home message



References

- *A Höhl, Dengel.A, Zhu X.X., Opening the Black-Box: A Systematic Review on Explainable AI in Remote Sensing, arXiv:2402.13791v1, 2023*

- *H. R. Tamaddon-Jahromi, N. K. Chakshu, I. Sazonov, L. M. Evans, H. Thomas, and P. Nithiarasu. Data-driven inverse modelling through neural network (deep learning) and computational heat transfer. Computer Methods in Applied Mechanics and Engineering, 369:113217, 2020.*

- F. Kratzert, M. Herrnegger, D. Klotz, S. Hochreiter, and G. Klambauer. Neuralhydrology - interpreting lstms in hydrology. In Explainable AI, volume 11700 of Lecture Notes in Computer Science, pages 347–362. Springer, 2019.

- A. Mamalakis, I. Ebert-Uphoff, and E. Barnes, “Explainable artificial intelligence in meteorology and climate science: Model fine-tuning, calibrating trust and learning new science,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 13200 LNAI, pp. 315

- P. W. Keys, E. A. Barnes, and N. H. Carter, “A machine-learning approach to human footprint index estimation with applications to sustainable development,” Environmental Research Letters, vol. 16, no. 4, p. 044 061, Apr. 2021, ISSN: 1748-9326. DOI: 10.1088/1748-9326/abe00a.

- Z. Labe and E. Barnes, “Predicting slowdowns in decadal climate warming trends with explainable neural networks,” Geophysical Research Letters, vol. 49, no. 9, 2022. DOI: 10.1029/2022GL098173.

- http://www.heatmapping.org/