# explanation

## Authors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

## Abstract

An explanation accompanies a prediction delivered by a machine learning (ML) method and says what about the data point drove it. It can be text, a score per feature, a simple hypothesis that approximates the learned one near the data point, or a heat map over the regions of an image. Two requirements pull against each other: an explanation must be faithful, reflecting the computation the learned hypothesis carries out, and effective, letting the user it is written for anticipate the predictions it accompanies. An explanation that agreed with the learned hypothesis everywhere would be that hypothesis again, so a simpler one gives up faithfulness somewhere.

## Definition

One approach to enhance the transparency of a machine learning (ML) method for its human user is to provide an explanation alongside the predictions delivered by the method. Explanations can take different forms. For instance, they may consist of human-readable text or quantitative indicators, such as feature importance scores for the individual features of a given data point (Molnar, 2025). Fig. 1 illustrates two types of explanations. The first is a local linear approximation $g(\mathbf{x})$ of a nonlinear learned hypothesis $\hat{h}(\mathbf{x})$ around a specific feature vector $\mathbf{x}'$, as used in the method local interpretable model-agnostic explanations (LIME). The second form of explanation depicted in the figure is a sparse set of predictions $\hat{h}(\mathbf{x}^{(1)}), \hat{h}(\mathbf{x}^{(2)}), \hat{h}(\mathbf{x}^{(3)})$ at selected feature vectors, offering concrete reference points for the user. For a differentiable $\hat{h}$, the local linear approximation is the one determined by the gradient $\nabla \hat{h}(\mathbf{x}')$, and the reference points are values of the same function.

A widely used form of explanation is a heat map: an intensity map that scores each region of an image by how much it drove the prediction, drawn over the image itself (Selvaraju et al., 2017). For a convolutional neural network (CNN), those scores are read off the network's own activations, which is what a class activation map (CAM) does.

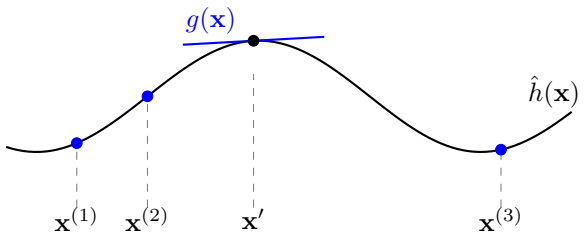


Figure 1: A learned hypothesis $\hat{h}(\mathbf{x})$ explained locally at some point $\mathbf{x}'$ by a linear approximation $g(\mathbf{x})$, and by the function values $\hat{h}\big(\mathbf{x}^{(r)}\big)$ for $r = 1, 2, 3$

Whatever form it takes, an explanation carries two requirements: it must be (i) understandable and (ii) faithful. Understandable means that the user can work with the explanation on their own. When the explanation is a local linear approximation $g$, this is concrete: the user must be able to evaluate $g$ at a feature vector of their choosing and read off what it delivers for the data points of a test set. An explanation the user cannot evaluate leaves the predictions as unanticipated as no explanation at all, which is what explainability measures.

Faithful means that the explanation reflects the computation the learned hypothesis actually carries out, rather than one that is easier to present. A map over the pixels of an image is faithful when manipulating a few of the pixels it highlights changes the prediction while changing as many dark ones does not (see explainable artificial intelligence (XAI)). The two requirements are separate, and neither follows from the other. An unfaithful explanation can still be understandable and still let the user anticipate well, whenever it tracks the prediction without carrying anything about the computation: maps have been found that are independent of both the model parameters and the training set, yet look like the ones that are not (Adebayo et al., 2018). Such an explanation predicts by correlation, so it fails when the correlation does, and it cannot support a user who acts on the features it highlights.

The two requirements also pull against each other. An explanation that agreed with the learned hypothesis everywhere would be that hypothesis again, so a simpler explanation gives up faithfulness somewhere; where the hypothesis is itself simple enough to follow, it serves as its own explanation and none has to be constructed (see interpretable machine learning (interpretable ML)).

Fig. 2 shows both requirements at work on a prediction made from weather radar. The features are the hourly precipitation over a 48 km box around Krems an der Donau, and the prediction answers whether it will rain at Krems two hours later. A CNN fitted to these images is explained by a CAM, which scores each cell of the image by how much it contributed to the prediction (Selvaraju et al., 2017). For the hour drawn, the network is certain of rain, and the map puts its weight on the band of precipitation south-east of the town rather than on the rain already overhead. Setting the precipitation to zero in the 160 cells the map scores highest moves the predicted score by 3.32, against 0.64 for the

160 it scores lowest, and the map-guided cells move it further on 58 of the 72 held-out hours.

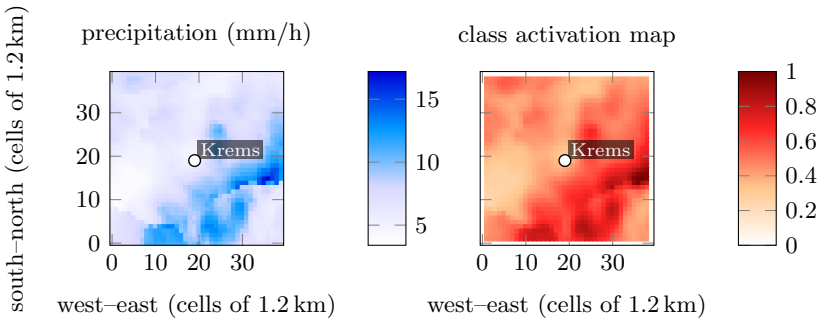


Figure 2: An explanation for a prediction made from weather radar on 15 September 2024 at 08:00 UTC. Left: the observed precipitation over a 48 km box around Krems an der Donau (circle), from which a CNN predicts rain at Krems two hours later. Right: the CAM explaining that prediction, scoring each cell by its contribution. Data generated by `pythondemos/explanation.py`

## Cross References

ML, prediction, feature, data point, classification, explainability, XAI, interpretable ML, LIME, CAM

## References

1. Adebayo et al. (2018). Sanity Checks for Saliency Maps. In: Adv. Neural Inf. Process. Syst. (NeurIPS).

2. Selvaraju et al. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: 2017 IEEE Int. Conf. Comput. Vis., pp. 618–626.

3. Molnar (2025). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 3rd ed., Ebook.