

# Explainable Artificial Intelligence (XAI)

**Francesca Grisoni**, Assistant Professor

Institute for Complex Molecular Systems & Eindhoven AI Systems Institute  
Department of Biomedical Engineering

f.grisoni@tue.nl

# Machine learning for drug discovery



## PhD Candidates

Derek van Tilborg

Riza Özçelik

Yves Nana (w/ IBM Zurich)

Cristina Izquierdo Lozano (w/ L. Albertazzi)

Emanuele Criscuolo (Guest PhD)

## MSc Students

Luke Rossen

Meilina Reksoprodjo

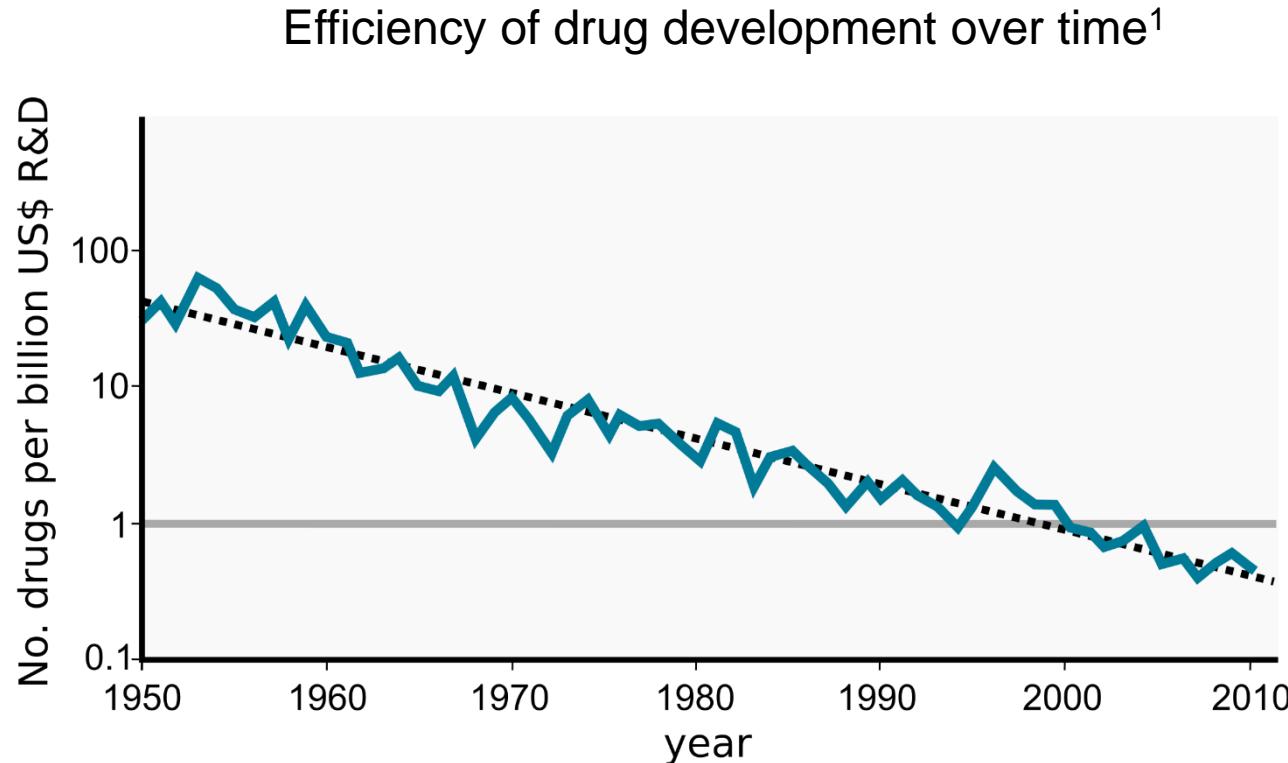
## BEP Students

Lisa Nooren

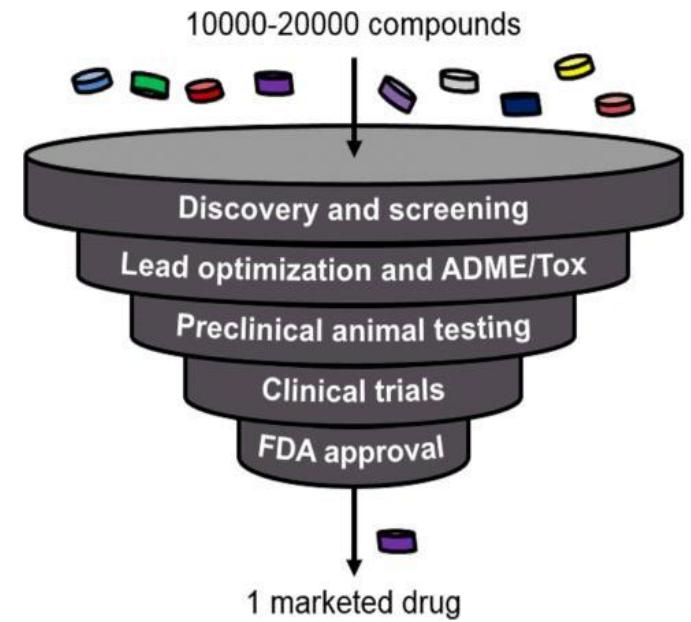
Sarah de Ruiter



# Molecular machine learning for drug discovery



Drug development pipeline<sup>2</sup>



<sup>1</sup>Data from: Scannell *et al.* (2012). *Nature Reviews Drug Discovery* **11**, 191.

<sup>2</sup>Figure adapted from: Ware and Khetani (2017). *Trends in Biotechnology* **35**, 172.

# Molecular machine learning for drug discovery

Chemical Universe<sup>1,2</sup>



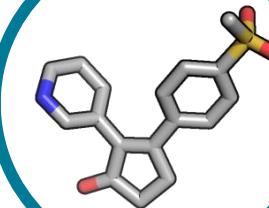
$10^{23} - 10^{100}$

Stars in the Milky Way

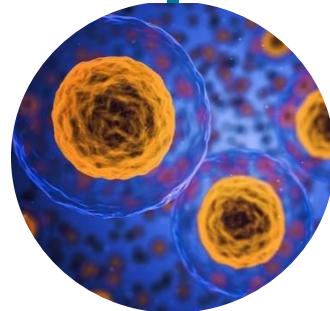
$10^8 - 10^9$



Known small molecule drugs



$10^4$



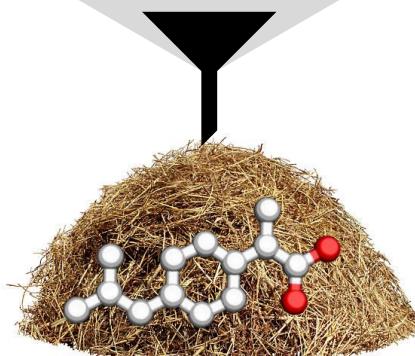
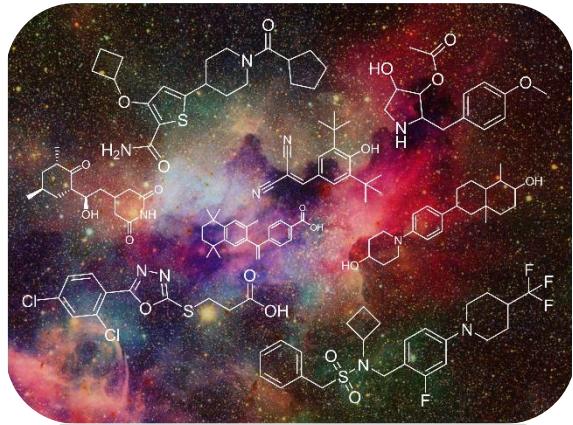
Cells in a human body  
 $10^{13} - 10^{14}$

<sup>1</sup>Ertl (2002) *Journal of Chemical Information and Computer Sciences* **43**, 374.

<sup>2</sup>Walters et al. (1998). *Drug Discovery Today* **3**, 160.

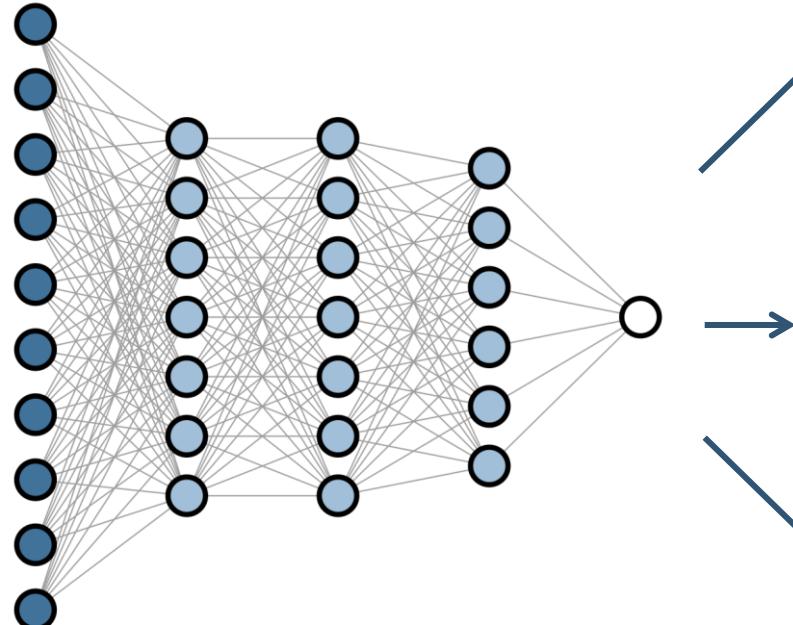
# AI for drug discovery

Chemical universe  
 $10^{23} - 10^{100}$  molecules

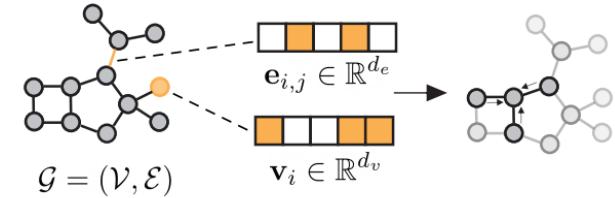


**Drug discovery**  
Finding a *needle in a haystack*

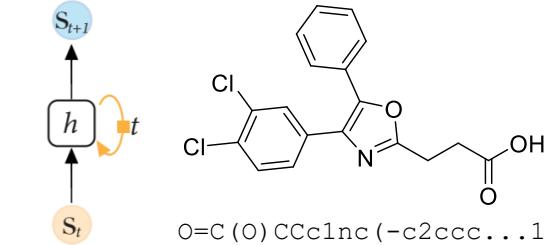
## Deep learning



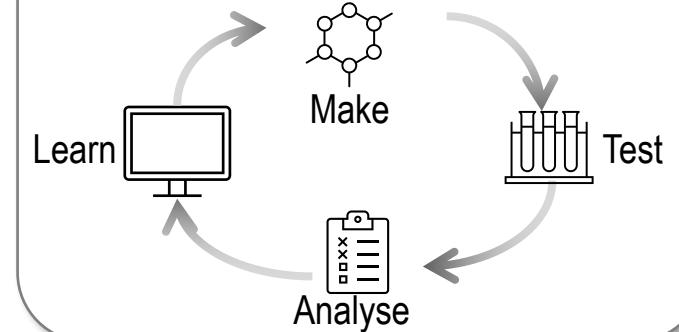
Molecular property prediction



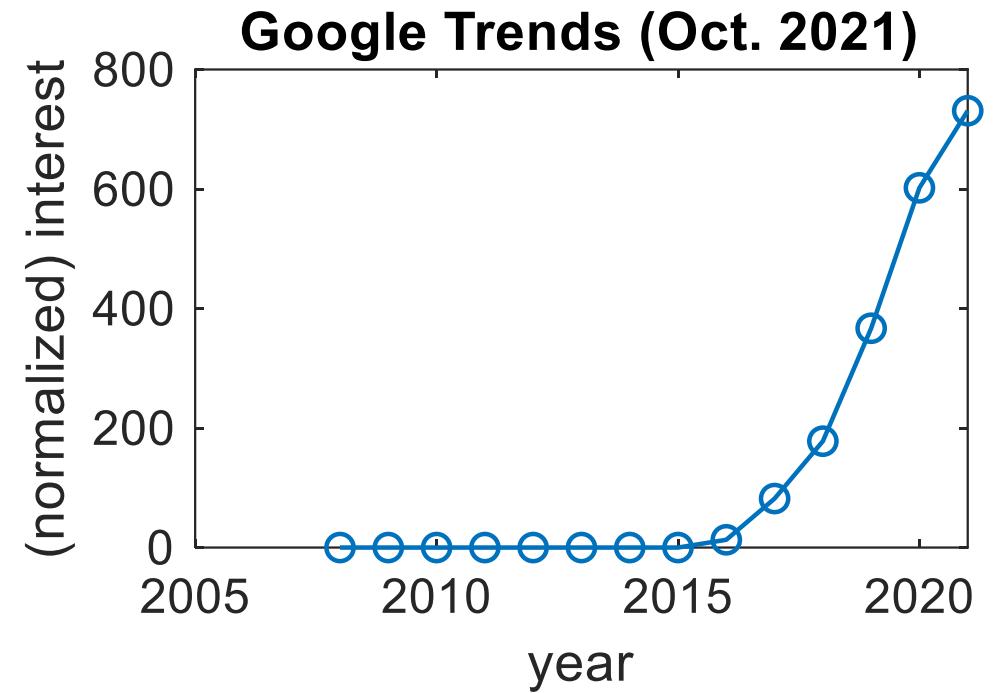
De novo molecule design



Active learning



# Explainable AI (XAI)



# Explainable AI (XAI)



Extraction of **relevant knowledge** from a machine learning model concerning **relationships contained in data or learned by the model**.



- **Transparency:** *how did the system reached an answer?*
- **Justification:** *is the answer acceptable?*
- **Informativeness:** *what can I learn from it?*
- **Uncertainty estimation:** *how reliable is a prediction?*

# Disclaimer: Explainability ≠ interpretability

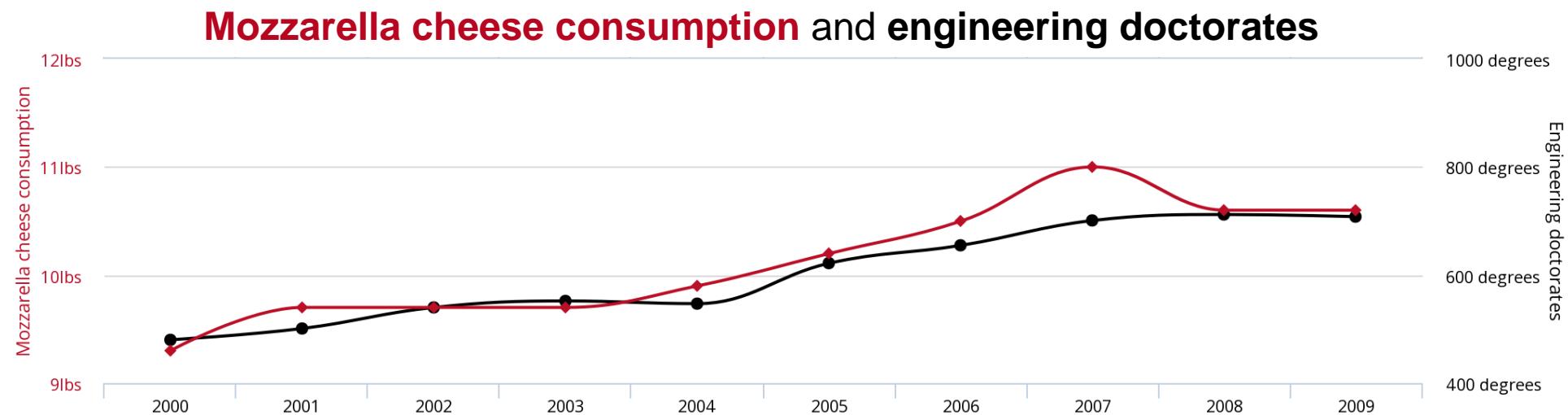
Explainability

Relationships within your data  
or in the model predictions

Interpretability

Determination of cause-effect  
relationships

≠



# Explainability-performance trade-off

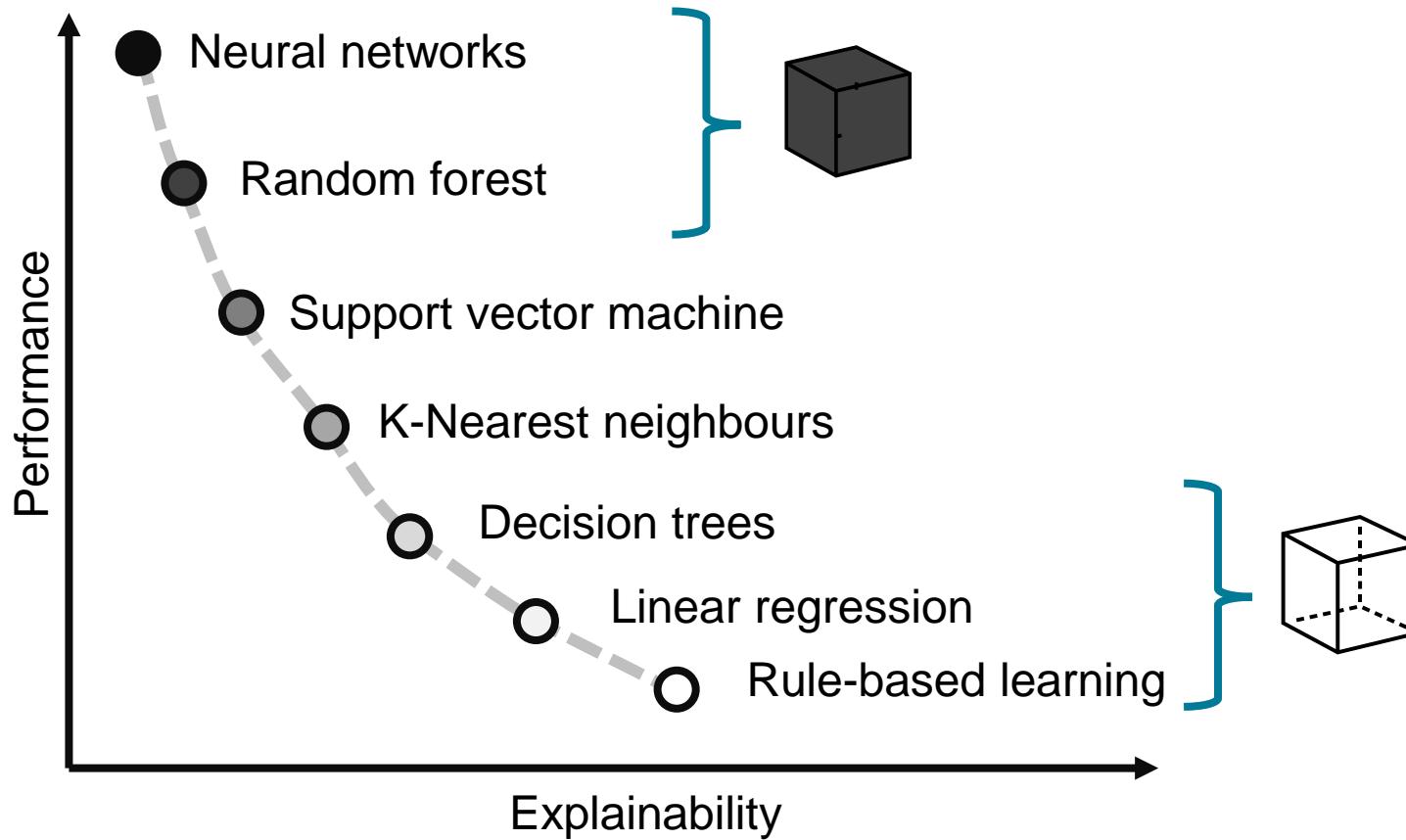


Figure inspired by: Morocho-Cayamcela et al. (2019) *IEEE Access* 7, 137184.

# Explainability-performance trade-off

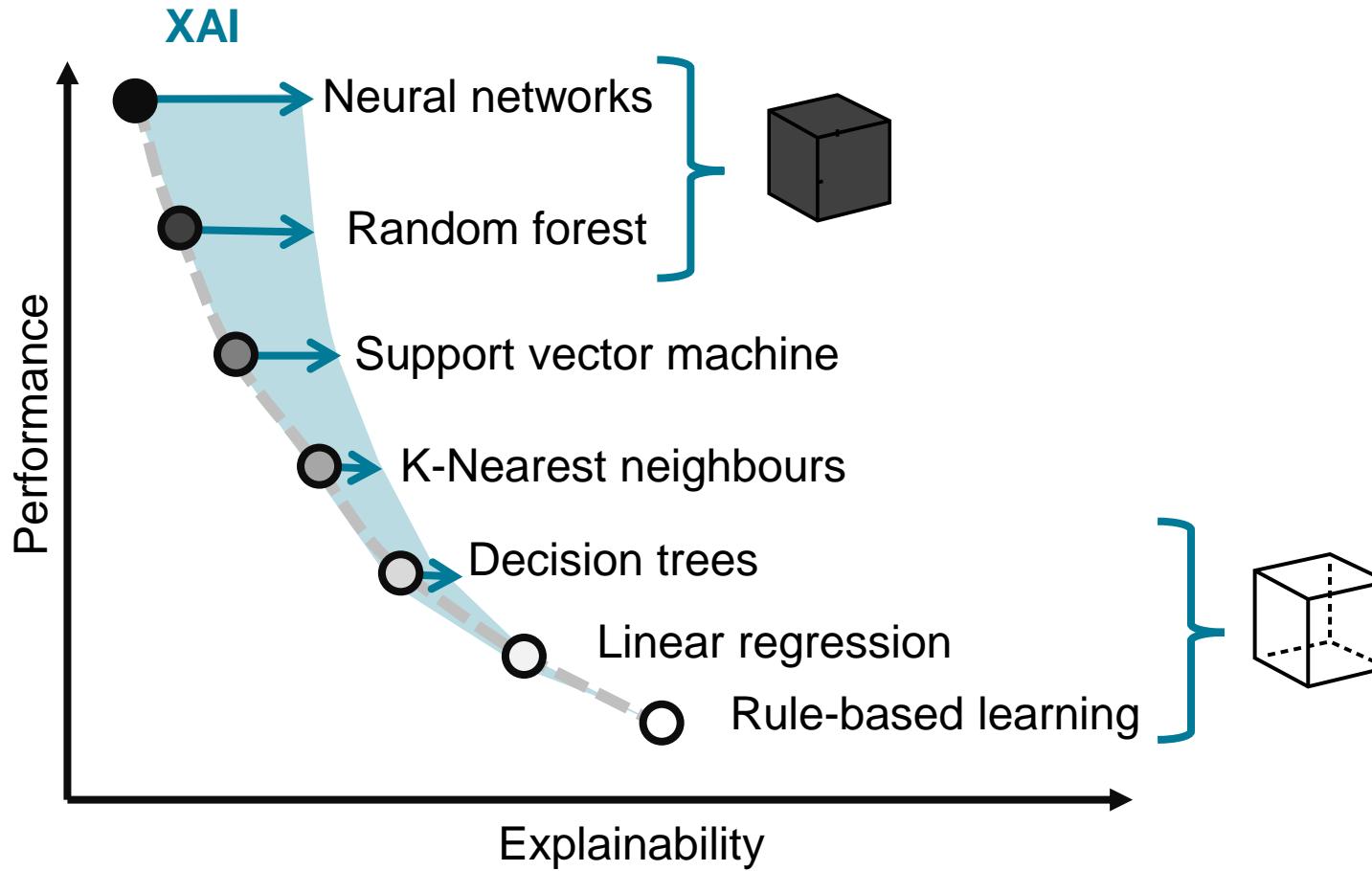


Figure inspired by: Morocho-Cayamcela et al. (2019) *IEEE Access* 7, 137184.

# Clever Hans (“der kluge Hans”)



- Berlin, 1900.
- Horse claimed to perform arithmetic (hoof tapping).
- The horse was responding directly to involuntary cues in the body language of the human trainer.

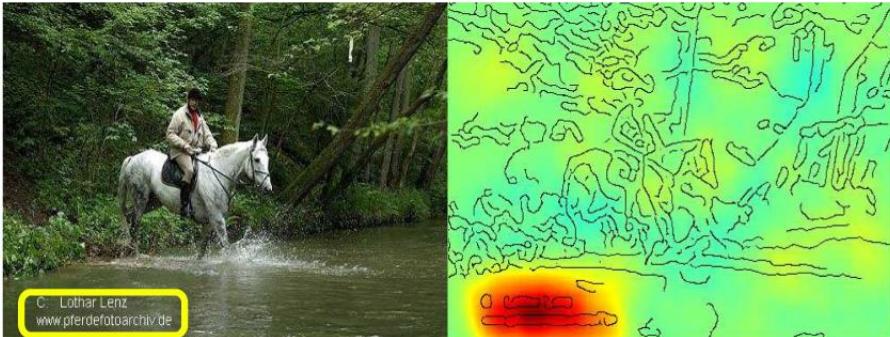


**Clever Hans effect**  
Correct outcome for the wrong reasons

Sebeok and Rosenthal (1981). *Annals of the New York Academy of Sciences*.

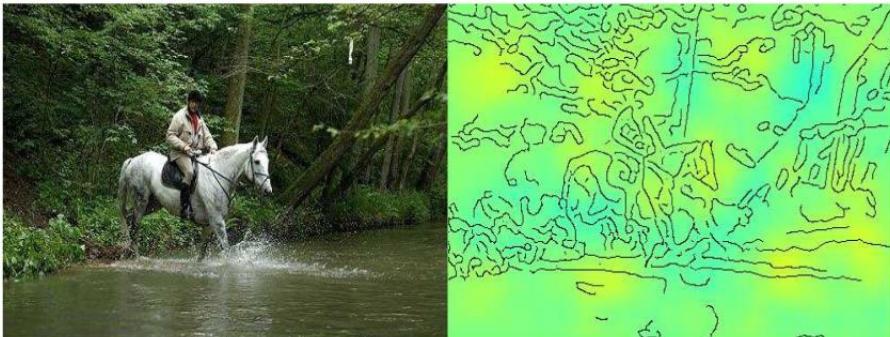
# “Clever Hans” in AI (shortcut learning)

Horse-picture from Pascal VOC data set



Source tag present  
↓

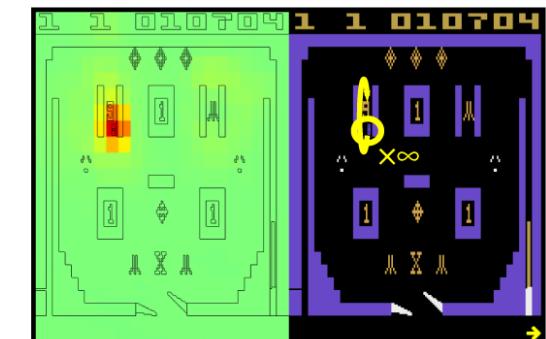
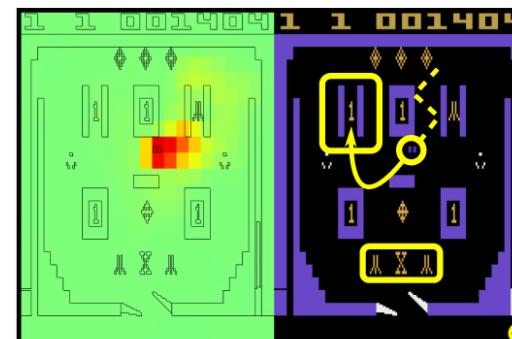
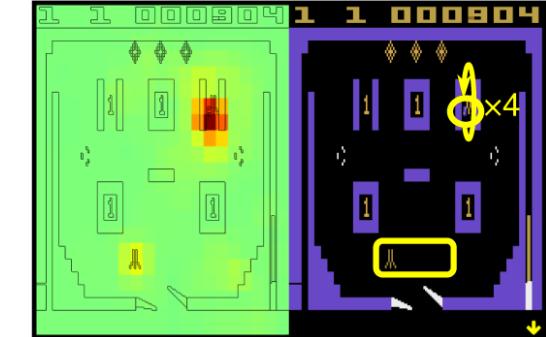
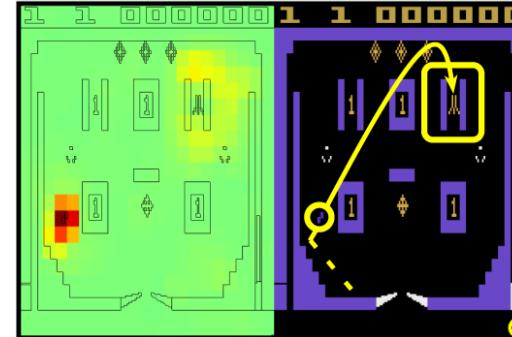
Classified as horse



No source tag present  
↓

Not classified as horse

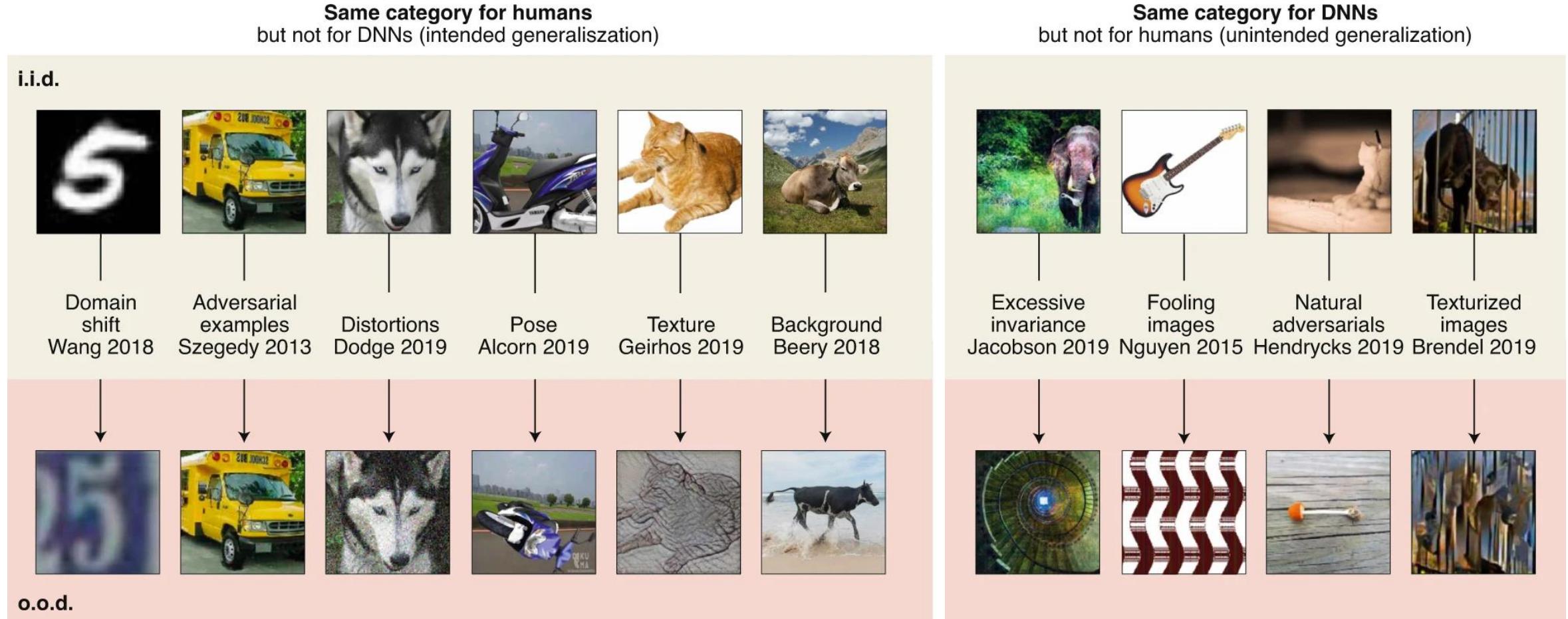
Pinball - relevance during game play



Identify valid vs invalid problem-solving behaviours

Lapuschkin et al. (2019) *Nature Communications* **10**, 1096.

# Surface learning and unintended generalization



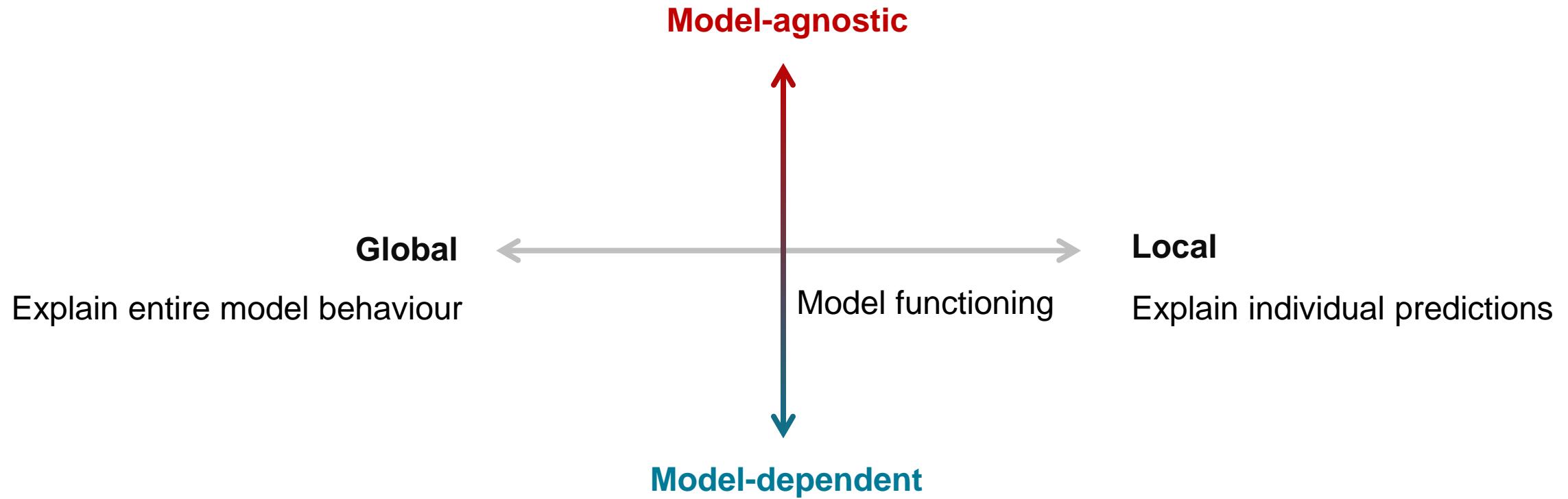
Performance metrics might just not be enough!

Geirhos et al. (2020). *Nature Machine Intelligence* 2, 665.

# A roadmap of XAI: pt. 1

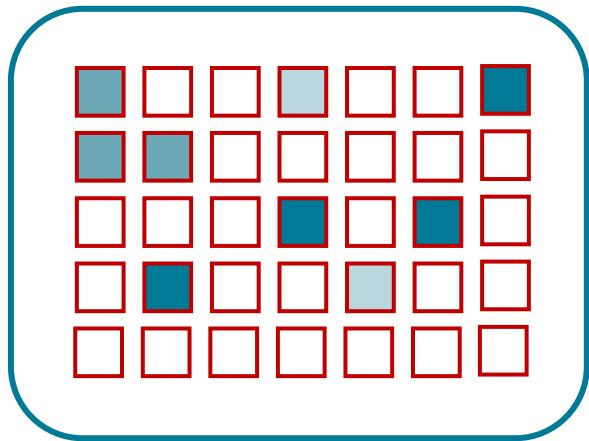


# A roadmap of XAI: pt. 1

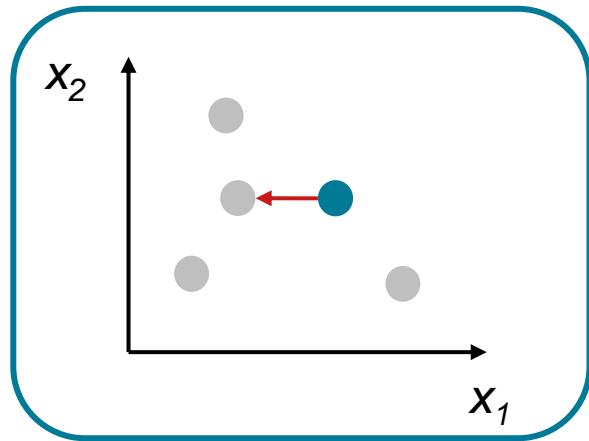


# A roadmap of XAI: pt. 2

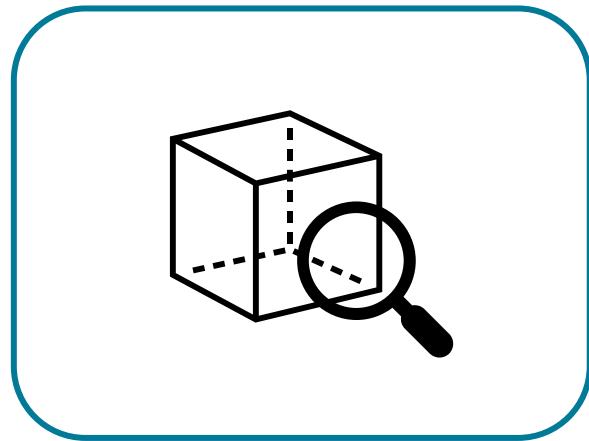
Feature attribution



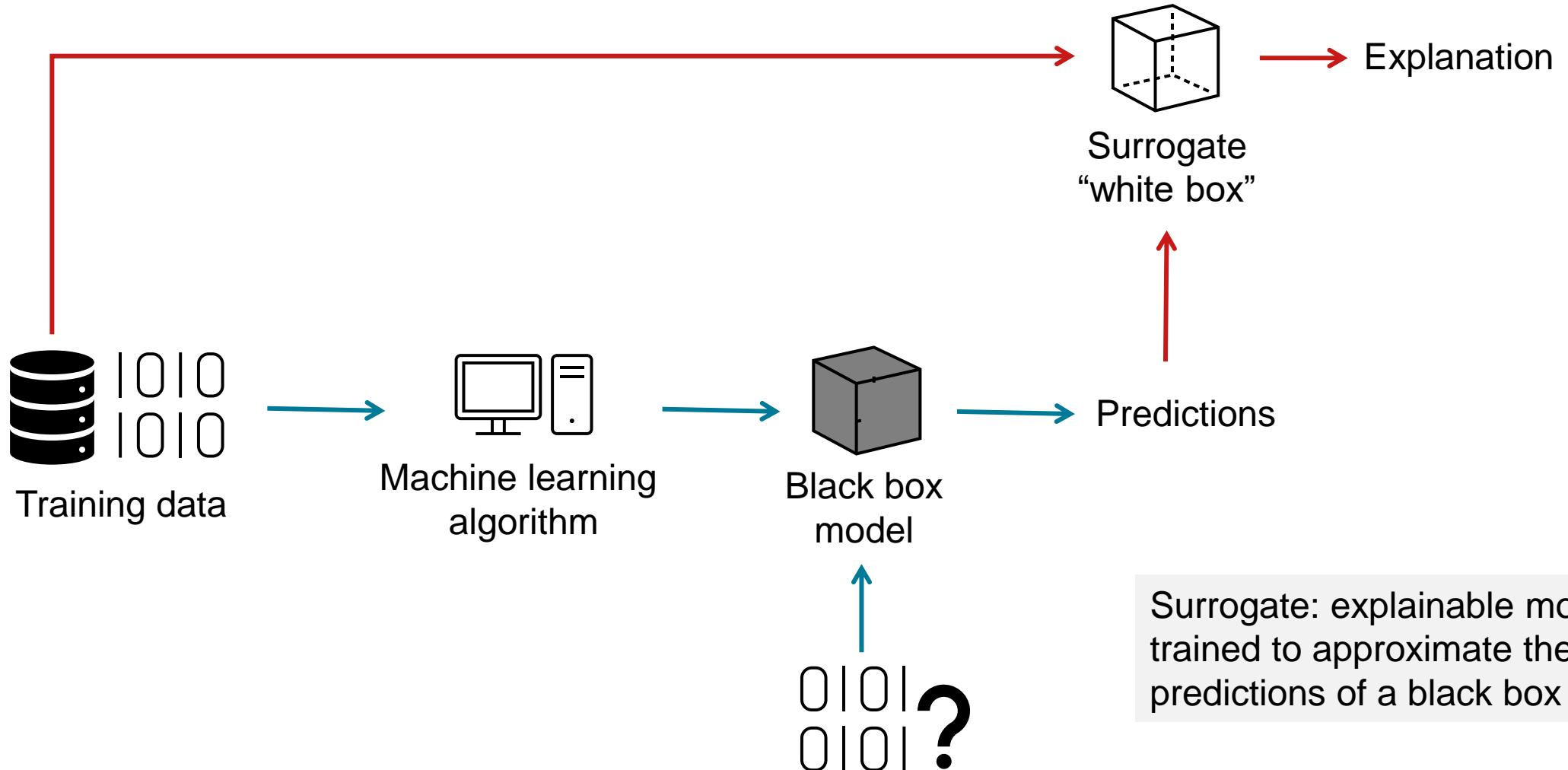
Instance-based



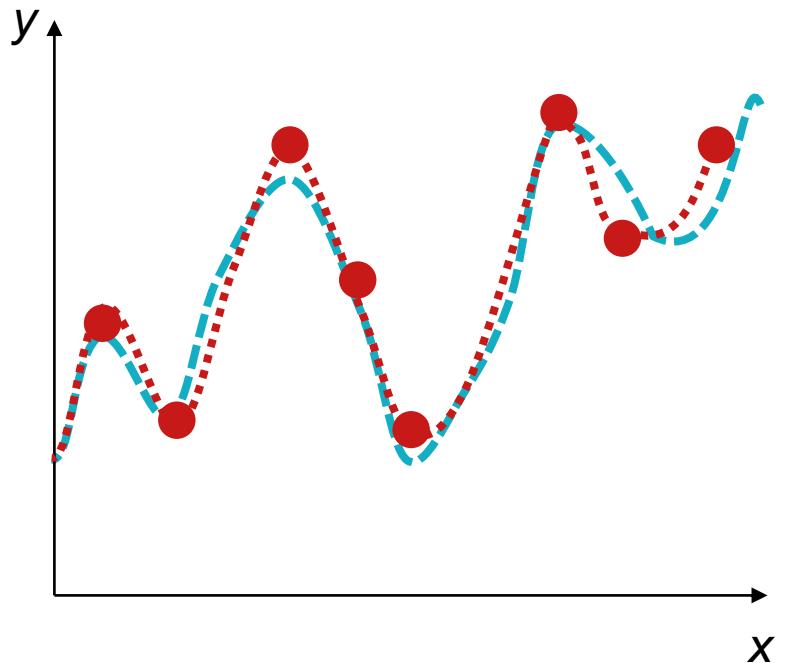
Self-explaining



# Surrogate models (Feature attribution)



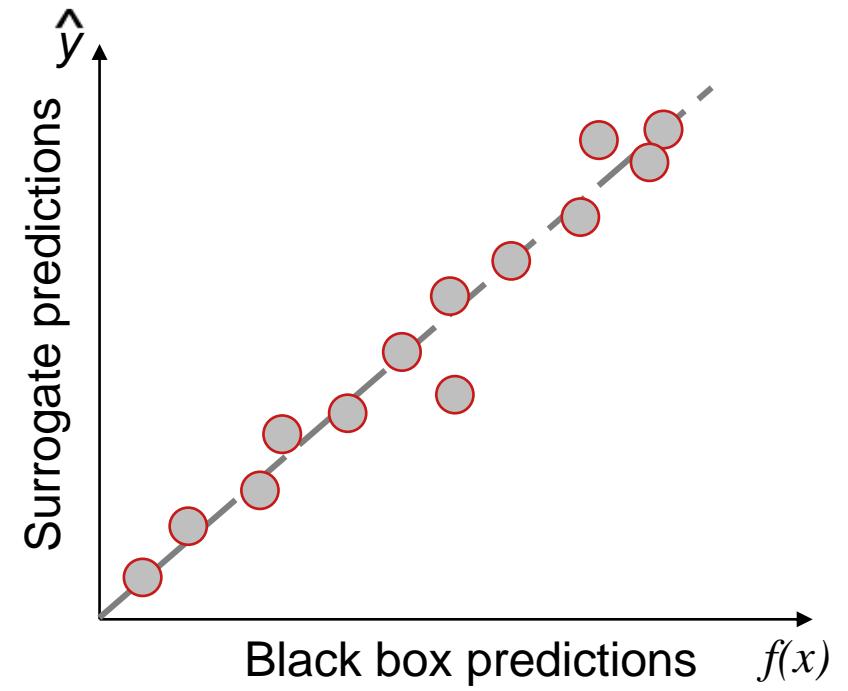
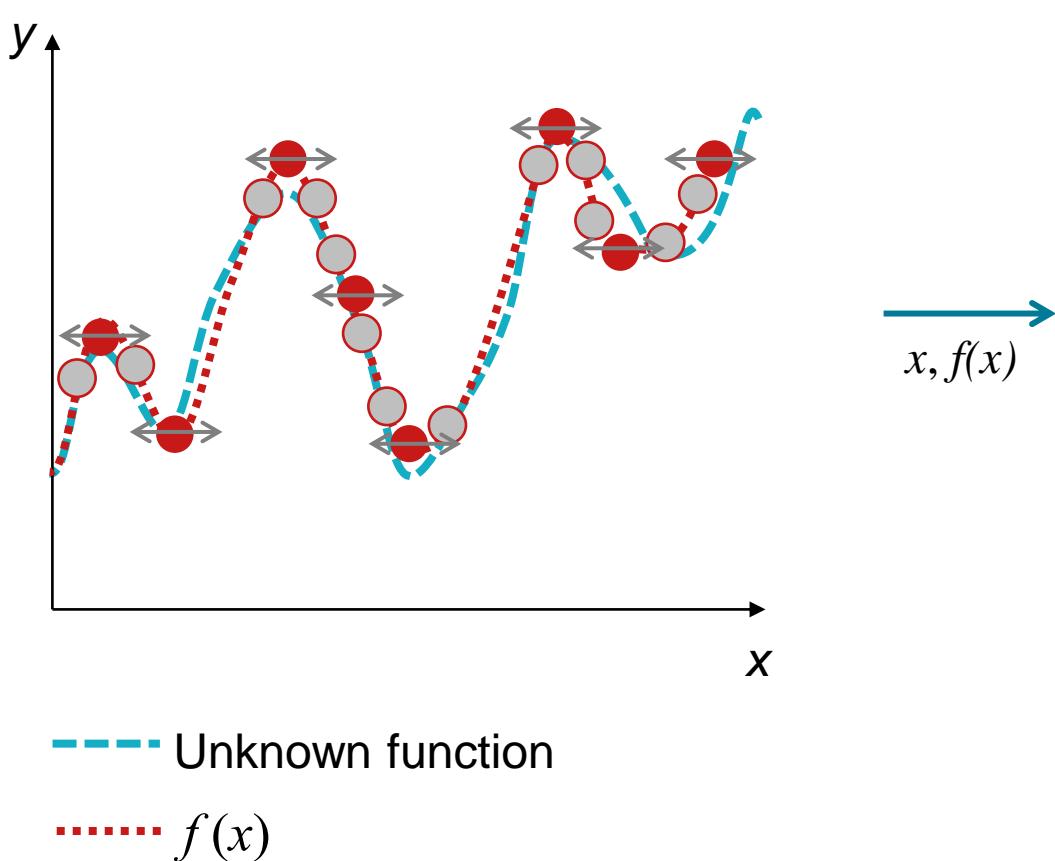
# Global surrogates



— Unknown function

···  $f(x)$

# Global surrogates



# Local interpretable model-agnostic explanations (LIME)<sup>1</sup>



## “Why Should I Trust You?” Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro  
University of Washington  
Seattle, WA 98105, USA  
marcotcr@cs.uw.edu

Sameer Singh  
University of Washington  
Seattle, WA 98105, USA  
sameer@cs.uw.edu

Carlos Guestrin  
University of Washington  
Seattle, WA 98105, USA  
guestrin@cs.uw.edu

### ABSTRACT

Despite widespread adoption, machine learning models remain mostly black boxes. Understanding the reasons behind predictions is, however, quite important in assessing *trust*, which is fundamental if one plans to take action based on a prediction, or when choosing whether to deploy a new model. Such understanding also provides insights into the model, which can be used to transform an untrustworthy model or prediction into a trustworthy one.

In this work, we propose LIME, a novel explanation technique that explains the predictions of *any* classifier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction. We also propose a method to explain models by presenting representative individual predictions and their explanations in a non-redundant way, framing the task as a submodular optimization problem. We demonstrate the flexibility of these methods by explaining different models for text (e.g. random forests) and image classification (e.g. neural networks). We show the utility of explanations via novel experiments, both simulated and with human subjects, on various scenarios that require trust: deciding if one should trust a prediction, choosing between models, improving an untrustworthy classifier, and identifying why a classifier should not be trusted.

how much the human understands a model's behaviour, as opposed to seeing it as a black box.

Determining trust in individual predictions is an important problem when the model is used for decision making. When using machine learning for medical diagnosis [6] or terrorism detection, for example, predictions cannot be acted upon on blind faith, as the consequences may be catastrophic.

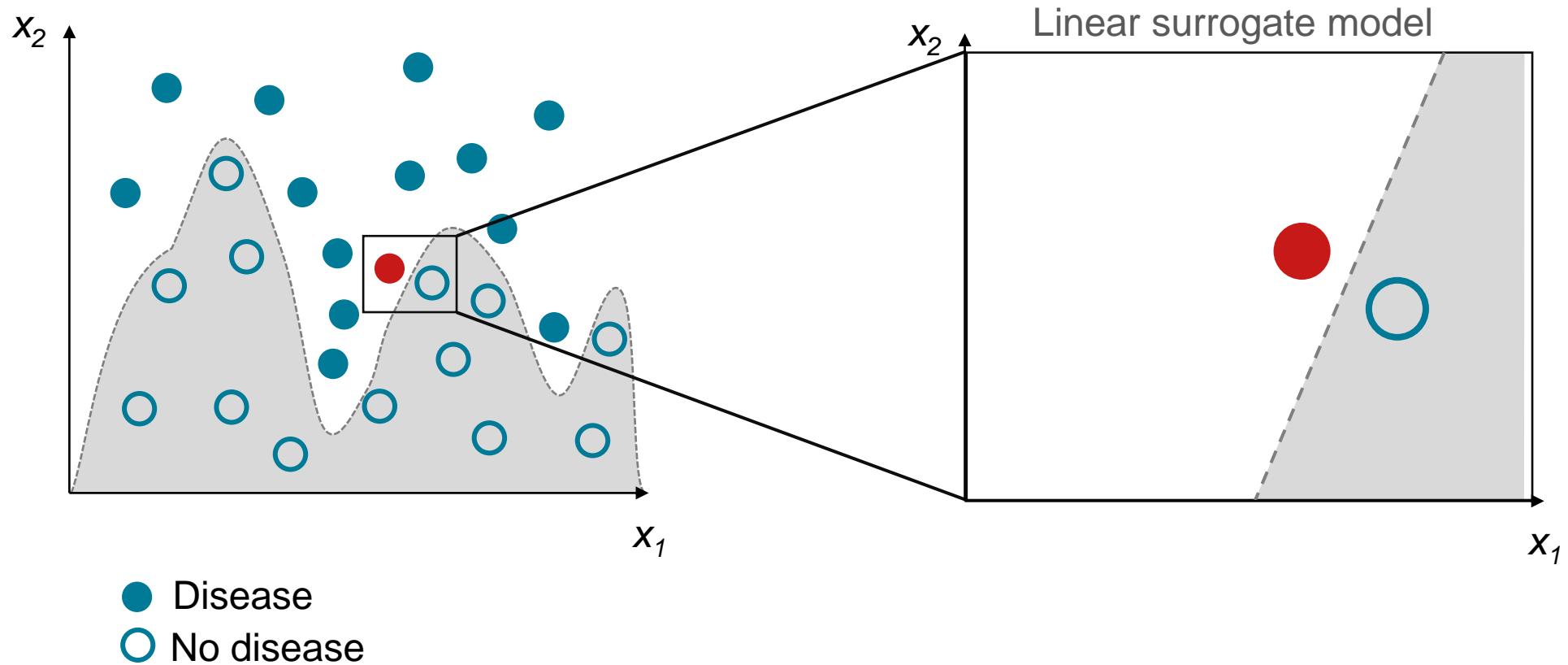
Apart from trusting individual predictions, there is also a need to evaluate the model as a whole before deploying it “in the wild”. To make this decision, users need to be confident that the model will perform well on real-world data, according to the metrics of interest. Currently, models are evaluated using accuracy metrics on an available validation dataset. However, real-world data is often significantly different, and further, the evaluation metric may not be indicative of the product’s goal. Inspecting individual predictions and their explanations is a worthwhile solution, in addition to such metrics. In this case, it is important to aid users by suggesting which instances to inspect, especially for large datasets.

In this paper, we propose providing explanations for individual predictions as a solution to the “trusting a prediction” problem, and selecting multiple such predictions (and explanations) as a solution to the “trusting the model” problem. Our main contributions are summarized as follows.

- Works on any “black-box” model
- Does not address the model internal functioning
- “Locally-faithful” explanations

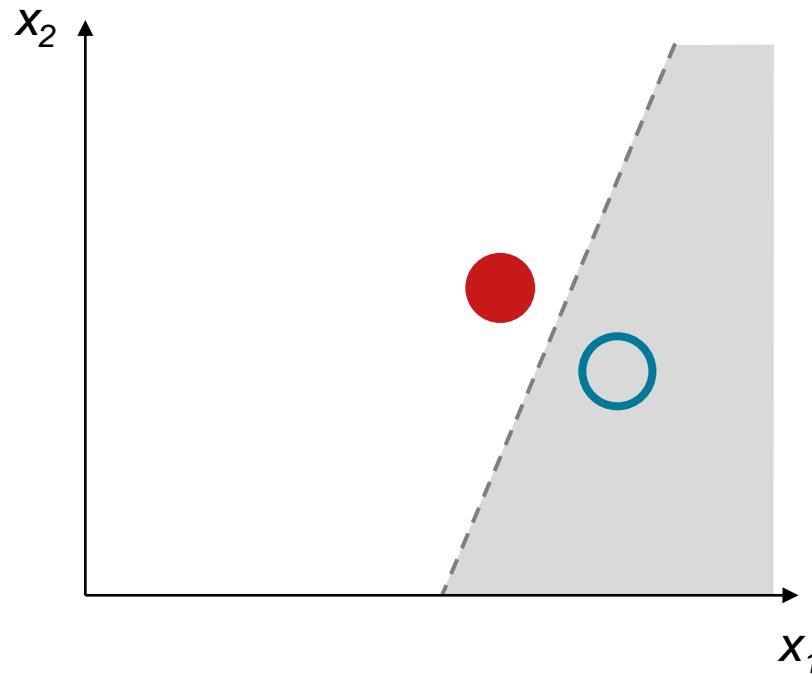
<sup>1</sup>Ribeiro et al. (2016). Proc. ACM SIGKDD Int, 1135.

# Local interpretable model-agnostic explanations (LIME)<sup>1</sup>



<sup>1</sup>Ribeiro et al. (2016). Proc. ACM SIGKDD Int, 1135.

# Local interpretable model-agnostic explanations (LIME)<sup>1</sup>

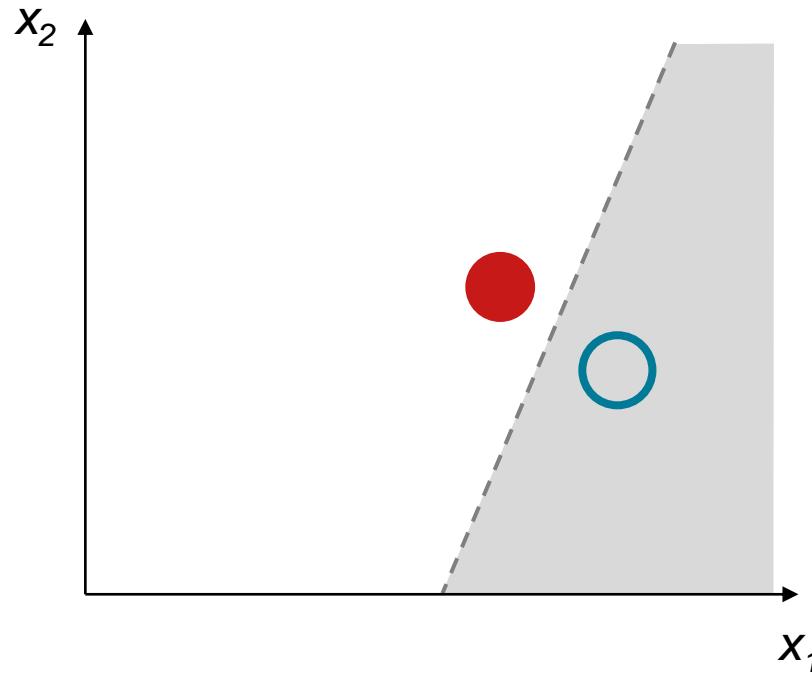


$$\xi(x) = \arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

Approximation   Complexity

<sup>1</sup>Ribeiro et al. (2016). Proc. ACM SIGKDD Int, 1135.

# Local interpretable model-agnostic explanations (LIME)<sup>1</sup>



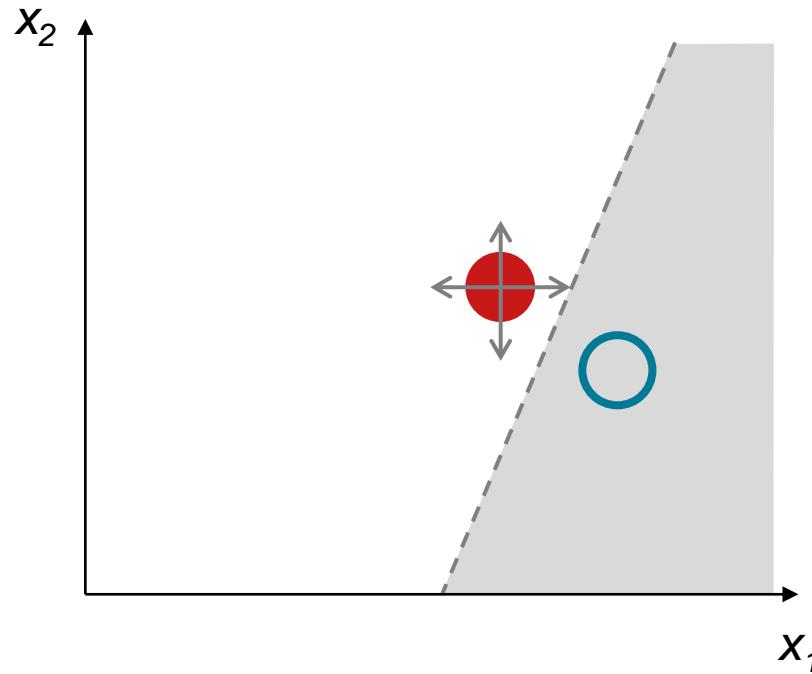
$$\xi(x) = \arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

Approximation Complexity  
Complex model Simple model Proximity

The equation  $\xi(x) = \arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g)$  is displayed, with a light blue box highlighting the term  $L(f, g, \pi_x)$ . Three arrows point from the text labels "Complexity", "Proximity", and "Simple model" to the terms  $\Omega(g)$ ,  $L(f, g, \pi_x)$ , and  $\Omega(g)$  respectively.

<sup>1</sup>Ribeiro et al. (2016). Proc. ACM SIGKDD Int, 1135.

# Local interpretable model-agnostic explanations (LIME)<sup>1</sup>



$$\xi(x) = \arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

Complex model

Simple model

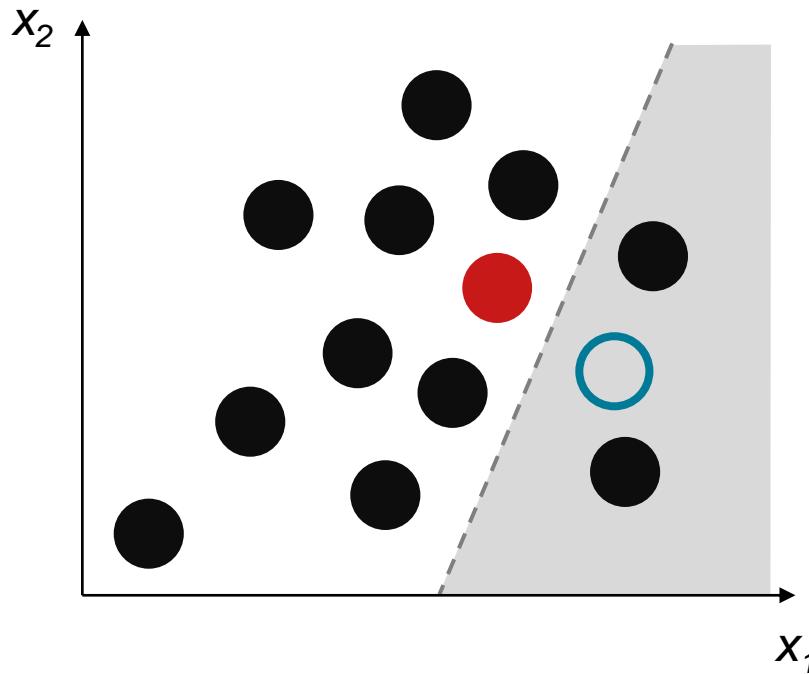
Approximation Complexity

Proximity

Random perturbations of  $x$   
(e.g., normal distribution)

<sup>1</sup>Ribeiro et al. (2016). Proc. ACM SIGKDD Int, 1135.

# Local interpretable model-agnostic explanations (LIME)<sup>1</sup>



$$\xi(x) = \arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

Complex model

Simple model

Approximation

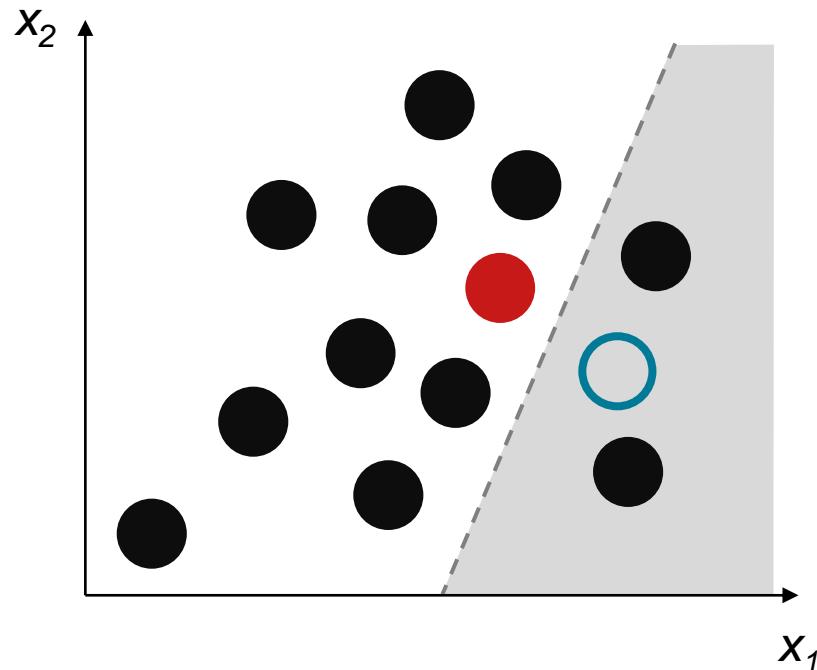
Complexity

Proximity

Random perturbations of  $x$   
(e.g., normal distribution)

**New dataset!**

# Local interpretable model-agnostic explanations (LIME)<sup>1</sup>



$$\xi(x) = \arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

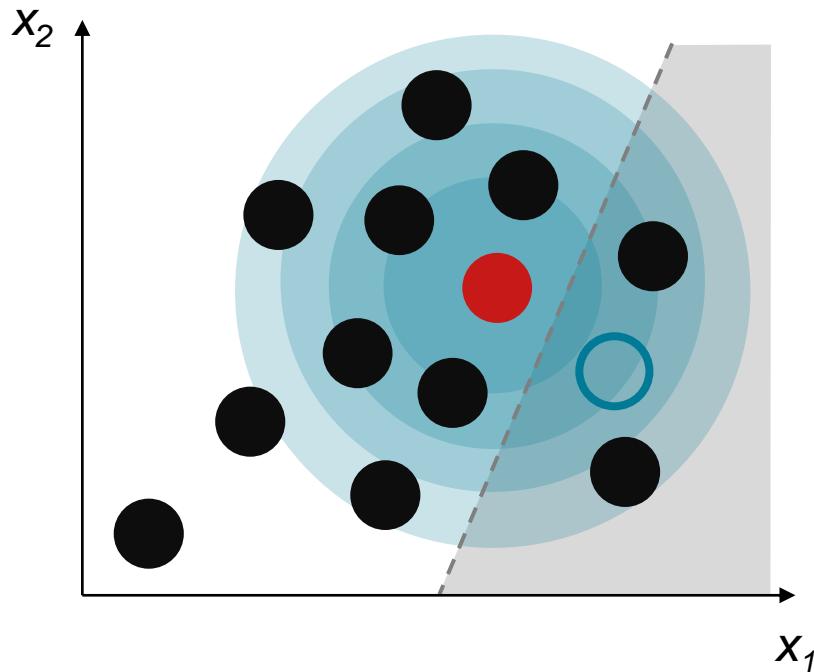
Approximation Complexity  
Complex model Simple model Proximity

Sum of Squared distances

$$L(f, g, \pi_x) = \sum_z \pi_x(z)(f(z) - g(z))^2$$

<sup>1</sup>Ribeiro et al. (2016). Proc. ACM SIGKDD Int, 1135.

# Local interpretable model-agnostic explanations (LIME)<sup>1</sup>



$$\xi(x) = \arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

Approximation Complexity  
Complex model Simple model Proximity

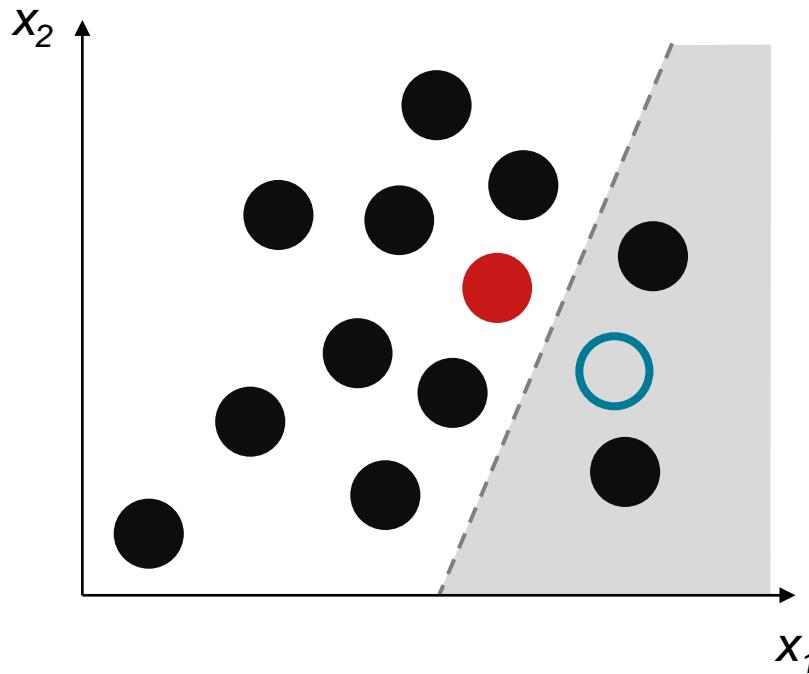
Sum of Squared distances

$$L(f, g, \pi_x) = \sum_z \pi_x(z)(f(z) - g(z))^2$$

Exponential kernel

<sup>1</sup>Ribeiro et al. (2016). Proc. ACM SIGKDD Int, 1135.

# Local interpretable model-agnostic explanations (LIME)<sup>1</sup>



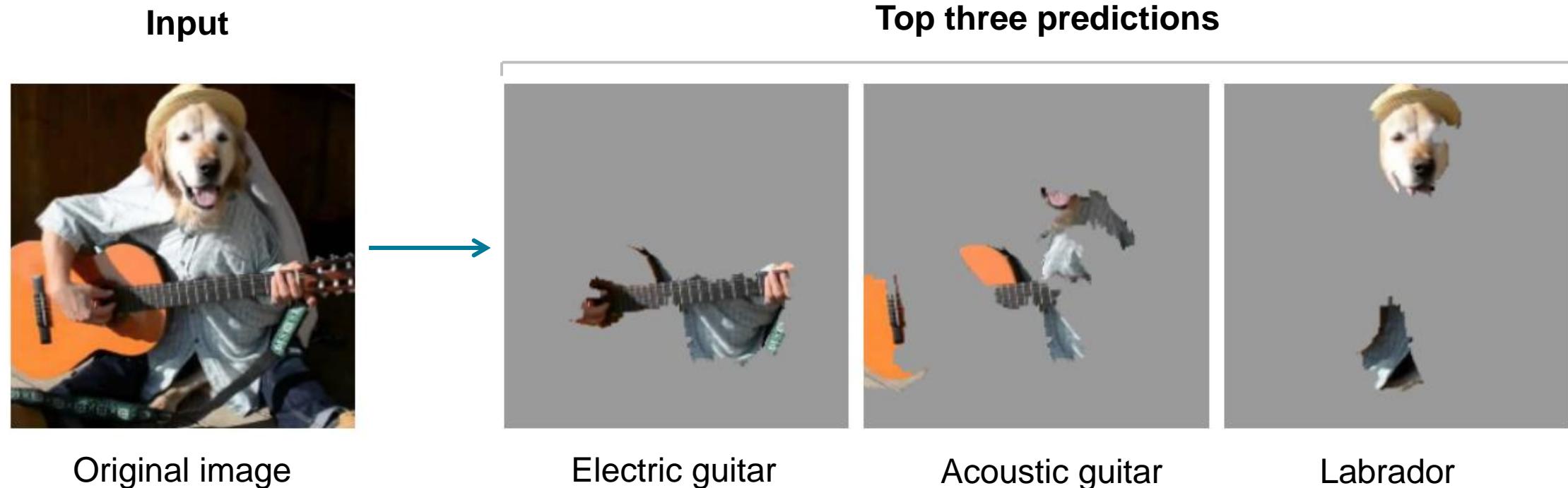
$$\xi(x) = \arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

Approximation      Complexity

**Regularization**  
Sparse linear models  
(e.g., LASSO)

<sup>1</sup>Ribeiro et al. (2016). Proc. ACM SIGKDD Int, 1135.

# Local interpretable model-agnostic explanations (LIME)<sup>1</sup>



<sup>1</sup>Ribeiro et al. (2016). Proc. ACM SIGKDD Int, 1135.

# Surrogate models

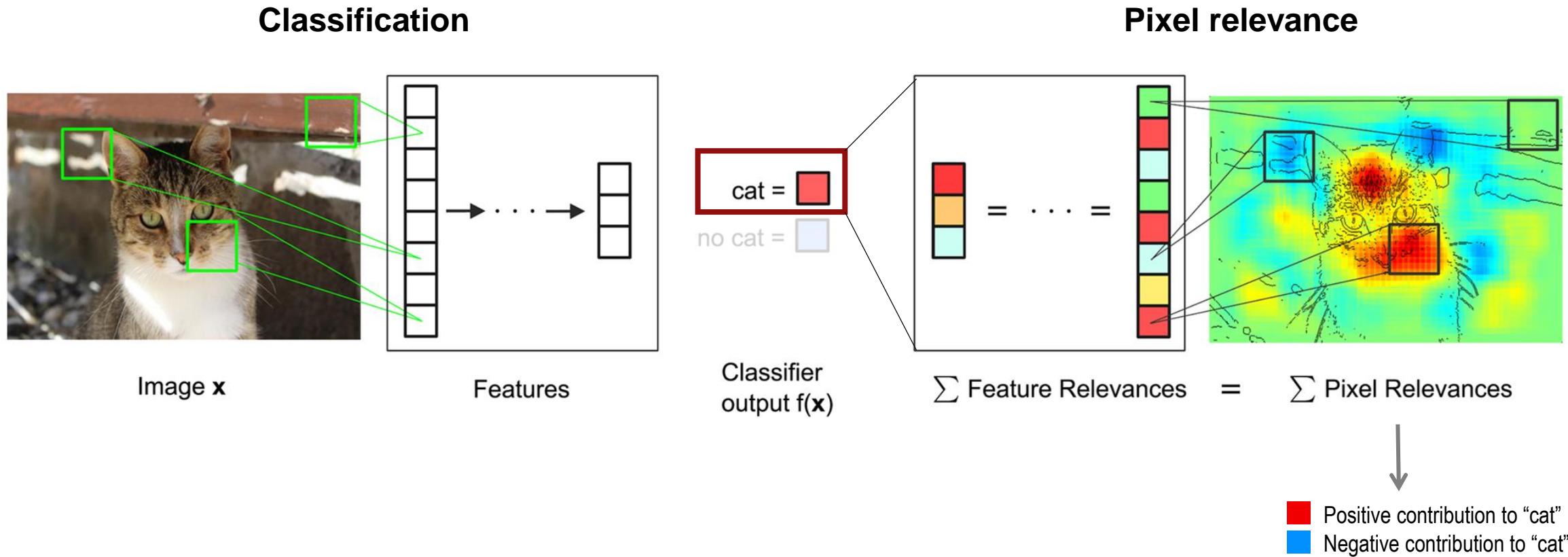
## Pros

- Easy to understand
- Can use *some* known information
- Model agnostic

## Cons

- The surrogate model is not the same as our original model
- Approximated explanations
- Conclusions about the black box model, not the data

# Pixel attribution methods (Feature attribution)



Bach et al. (2015). PLoS one **10**, e0130140.

# Layer-wise relevance propagation (LRP)



RESEARCH ARTICLE

## On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation

Sebastian Bach<sup>1,2\*</sup>, Alexander Binder<sup>2,5\*</sup>, Grégoire Montavon<sup>2</sup>, Frederick Klauschen<sup>3</sup>, Klaus-Robert Müller<sup>2,4\*</sup>, Wojciech Samek<sup>1,2\*</sup>

**1** Machine Learning Group, Fraunhofer Heinrich Hertz Institute, Berlin, Germany, **2** Machine Learning Group, Technische Universität Berlin, Berlin, Germany, **3** Charité University Hospital, Berlin, Germany, **4** Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea, **5** ISTD Pillar, Singapore University of Technology and Design (SUTD), Singapore

\* These authors contributed equally to this work.

\* [sebastian.bach@hhi.fraunhofer.de](mailto:sebastian.bach@hhi.fraunhofer.de) (SB), [klaus-robert.mueller@tu-berlin.de](mailto:klaus-robert.mueller@tu-berlin.de) (KM), [wojciech.samek@hhi.fraunhofer.de](mailto:wojciech.samek@hhi.fraunhofer.de) (WS)



### OPEN ACCESS

**Citation:** Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W (2015) On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLoS ONE 10(7): e0130140. doi:10.1371/journal.pone.0130140

**Editor:** Oscar Deniz Suárez, Universidad de Castilla-La Mancha, SPAIN

**Received:** May 19, 2014

**Accepted:** May 15, 2015

**Published:** July 10, 2015

**Copyright:** © 2015 Bach et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

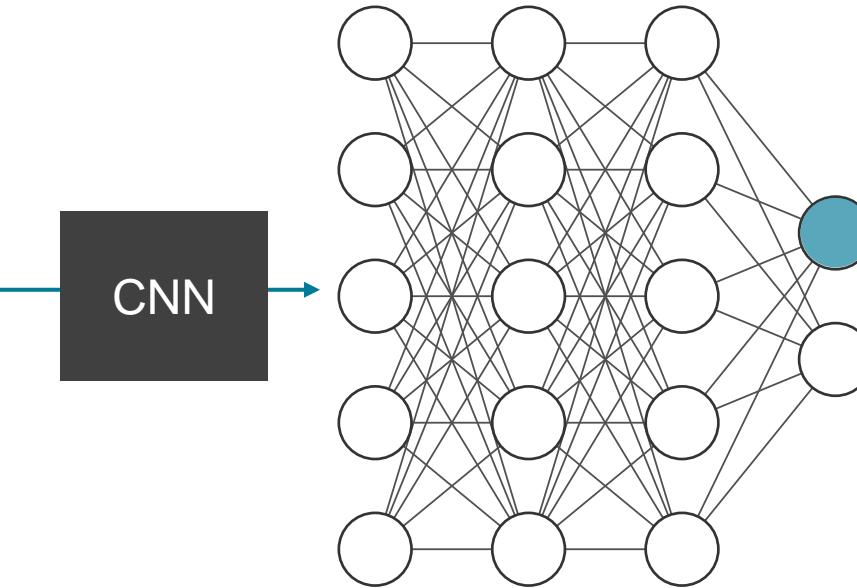
### Abstract

Understanding and interpreting classification decisions of automated image classification systems is of high value in many applications, as it allows to verify the reasoning of the system and provides additional information to the human expert. Although machine learning methods are solving very successfully a plethora of tasks, they have in most cases the disadvantage of acting as a black box, not providing any information about what made them arrive at a particular decision. This work proposes a general solution to the problem of understanding classification decisions by pixel-wise decomposition of nonlinear classifiers. We introduce a methodology that allows to visualize the contributions of single pixels to predictions for kernel-based classifiers over Bag of Words features and for multilayered neural networks. These pixel contributions can be visualized as heatmaps and are provided to a human expert who can intuitively not only verify the validity of the classification decision, but also focus further analysis on regions of potential interest. We evaluate our method for classifiers trained on PASCAL VOC 2009 images, synthetic image data containing geometric shapes, the MNIST handwritten digits data set and for the pre-trained ImageNet model available as part of the Caffe open source package.

- Which pixels contributed to a certain prediction
- Model-specific (neural networks)
- Computes the relevance via pixel-wise decomposition

Bach et al. (2015). *PloS one* **10**, e0130140.

# Layer-wise relevance propagation (LRP)

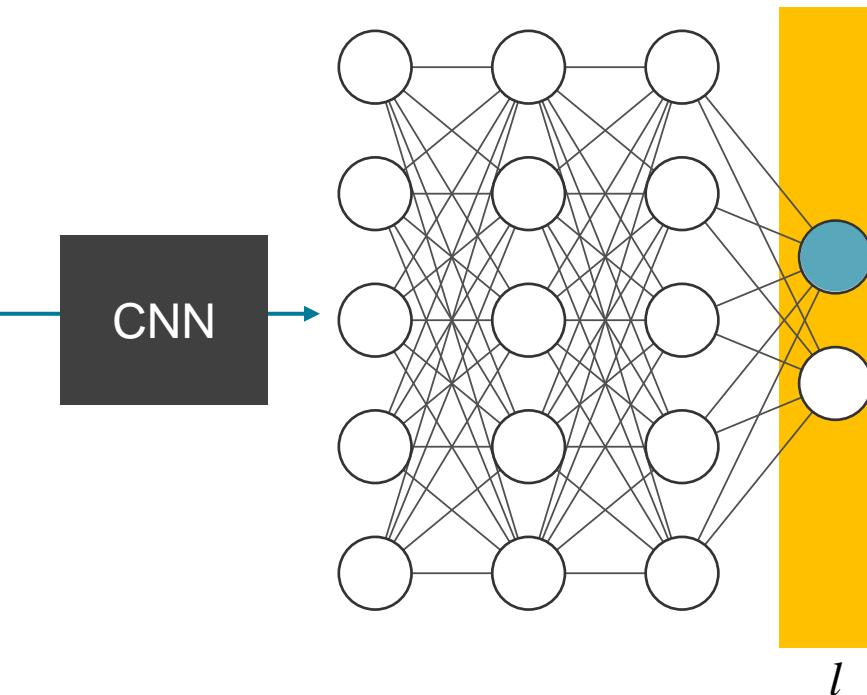


“Cat”  $R_1^{(4)} = f(x) = 0.8$   
80%

“Non-cat”  
20%

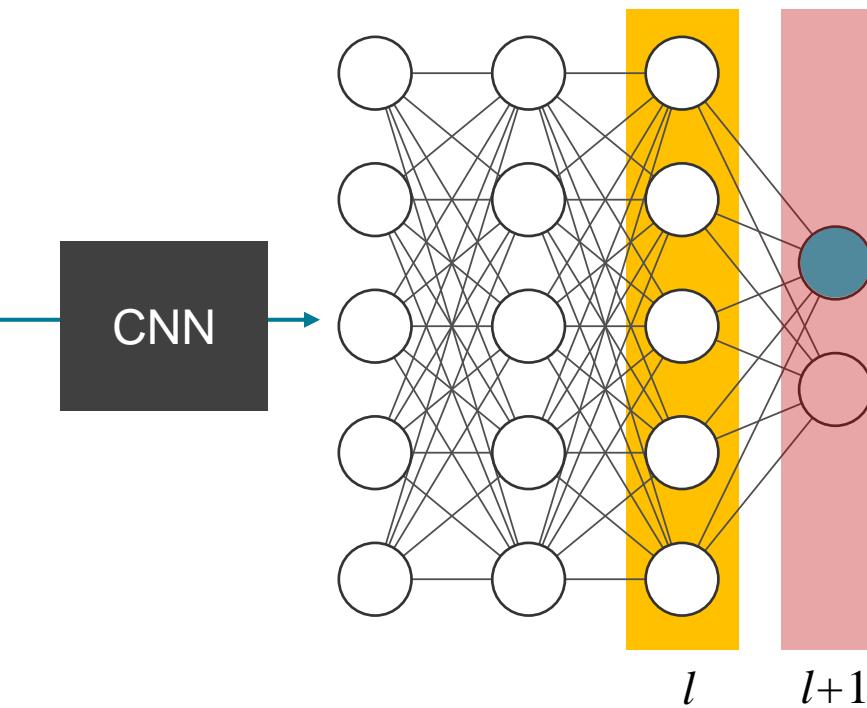
Example from: <https://www.youtube.com/watch?v=PDRewtcqmal>  
Bach et al. (2015). PloS one 10, e0130140.

# Layer-wise relevance propagation (LRP)



Example from: <https://www.youtube.com/watch?v=PDRewtcqmal>  
Bach et al. (2015). PloS one 10, e0130140.

# Layer-wise relevance propagation (LRP)



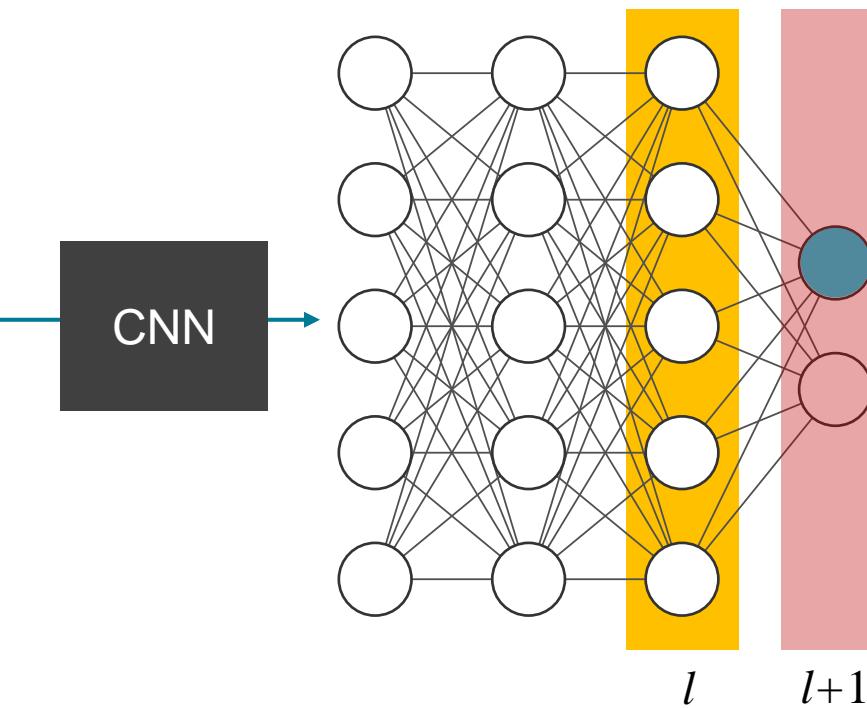
“Cat”  $R_1^{(4)} = f(x) = 0.8$   
“Non-cat”

$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_k z_{kj}} R_j^{(l+1)} \quad \text{with} \quad z_{ik} = x_i^{(l)} w_{ij}^{(l,l+1)}$$

Neurons in layer  $l$   
Neurons in layer  $l+1$

Example from: <https://www.youtube.com/watch?v=PDRewtcqmal>  
Bach et al. (2015). PLoS one 10, e0130140.

# Layer-wise relevance propagation (LRP)



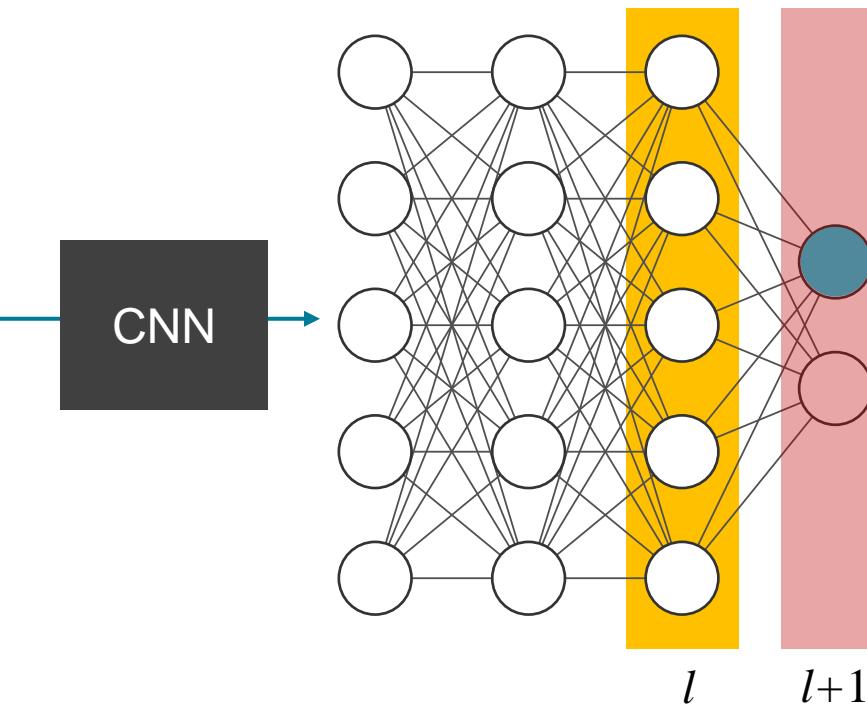
$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_k z_{kj}} R_j^{(l+1)} \quad \text{with} \quad z_{ik} = x_i^{(l)} w_{ij}^{(l,l+1)}$$

Relative activation

“Cat”  $R_1^{(4)} = f(x) = 0.8$   
“Non-cat”

Example from: <https://www.youtube.com/watch?v=PDRewtcqmal>  
Bach et al. (2015). PLoS one 10, e0130140.

# Layer-wise relevance propagation (LRP)



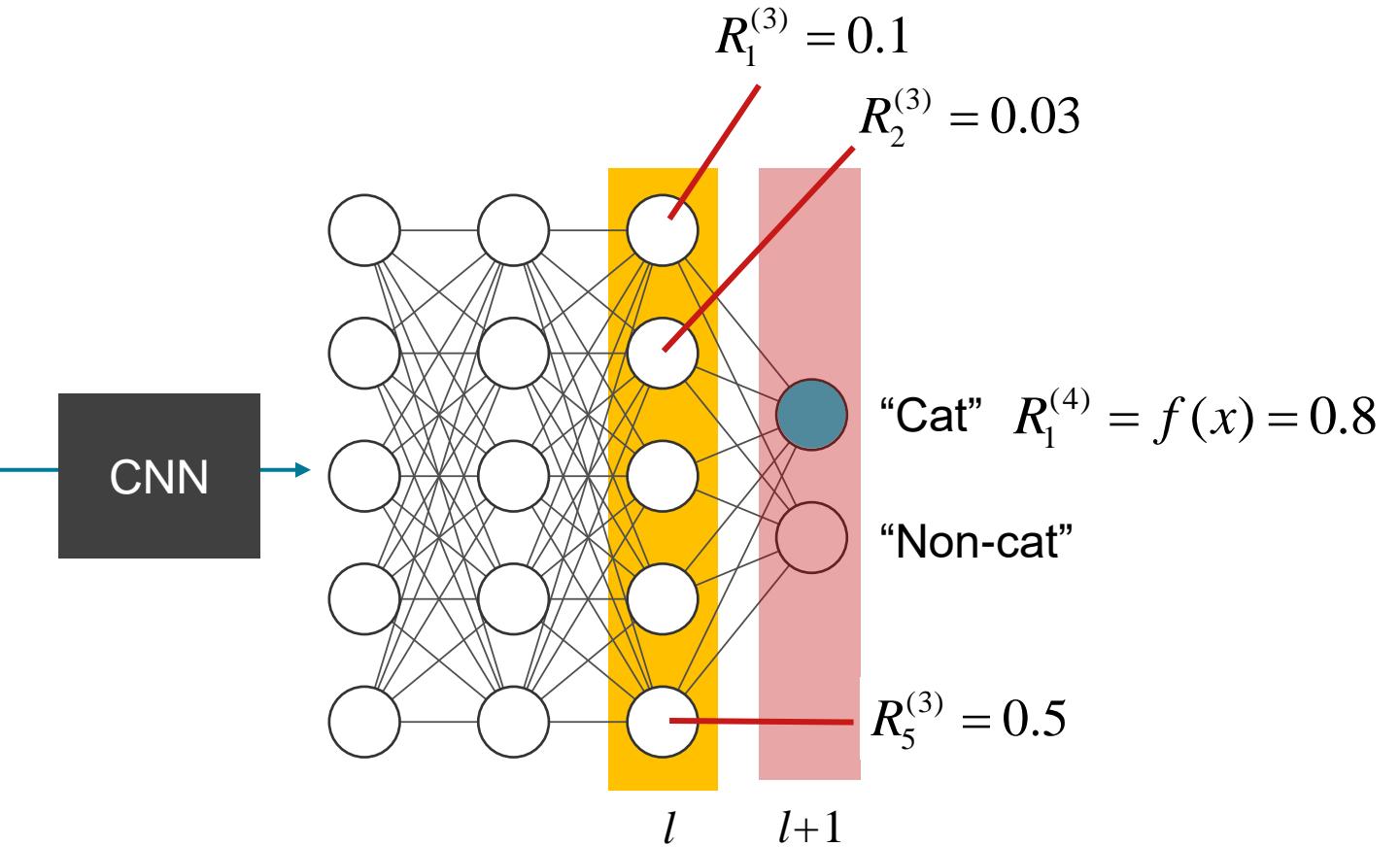
$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_k z_{kj}} R_j^{(l+1)}$$
 with  $z_{ik} = x_i^{(l)} w_{ij}^{(l,l+1)}$

Propagate the previous relevance backwards

"Cat"  $R_1^{(4)} = f(x) = 0.8$   
"Non-cat"

Example from: <https://www.youtube.com/watch?v=PDRewtcqmal>  
Bach et al. (2015). PLoS one 10, e0130140.

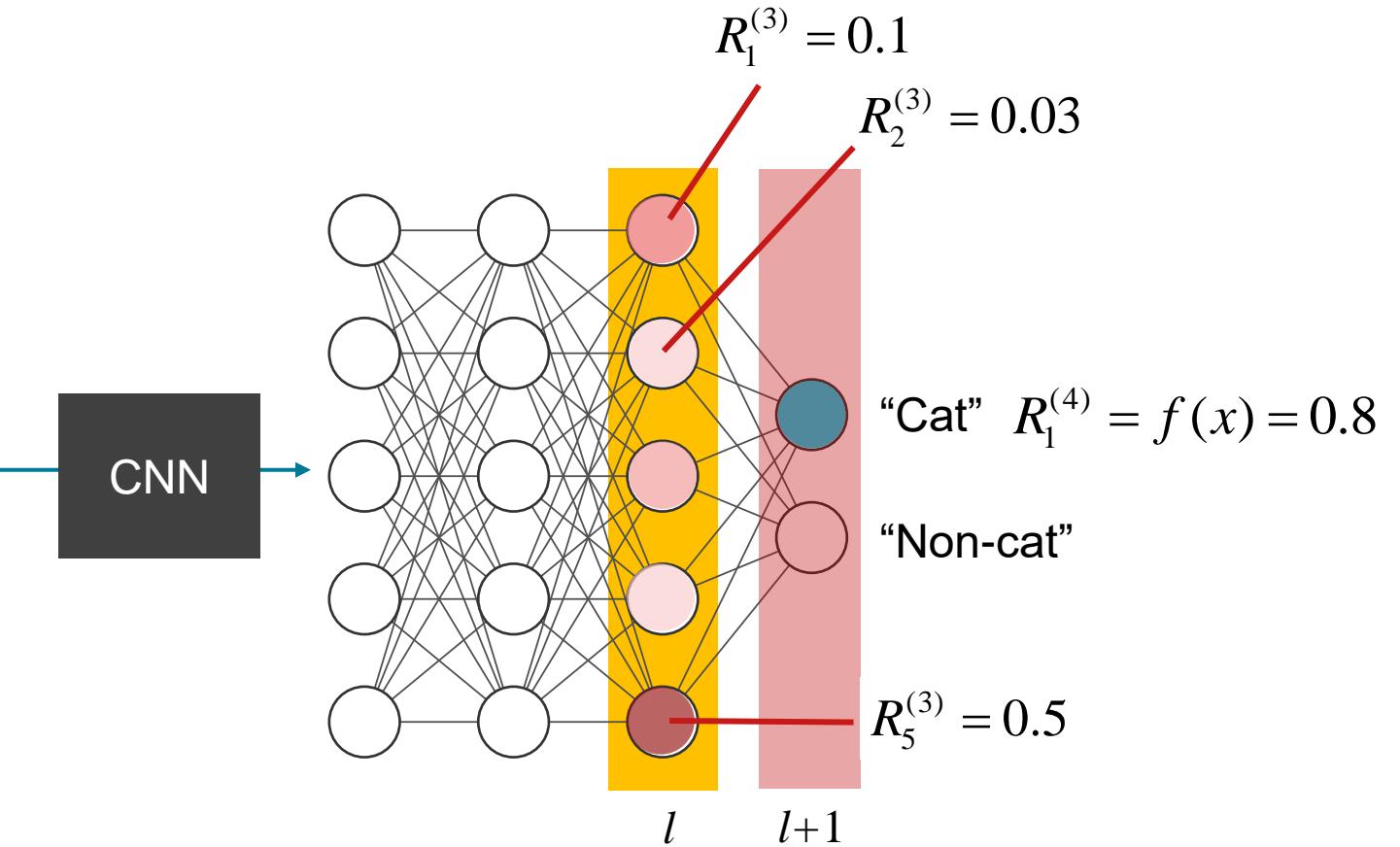
# Layer-wise relevance propagation (LRP)



$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_k z_{kj}} R_j^{(l+1)} \quad \text{with} \quad z_{ik} = x_i^{(l)} w_{ij}^{(l,l+1)}$$

Example from: <https://www.youtube.com/watch?v=PDRewtcqmal>  
Bach et al. (2015). PloS one 10, e0130140.

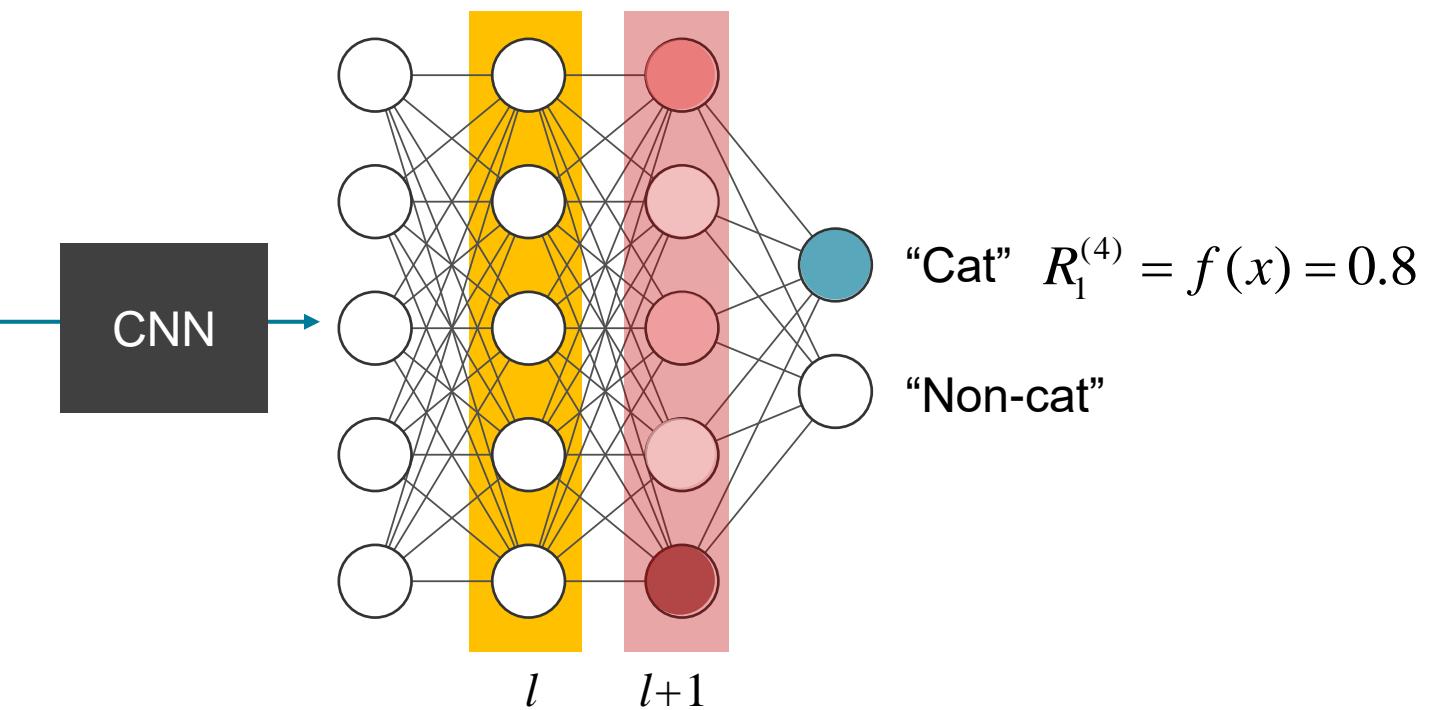
# Layer-wise relevance propagation (LRP)



$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_k z_{kj}} R_j^{(l+1)} \quad \text{with} \quad z_{ik} = x_i^{(l)} w_{ij}^{(l,l+1)}$$

Example from: <https://www.youtube.com/watch?v=PDRewtcqmal>  
Bach et al. (2015). PloS one 10, e0130140.

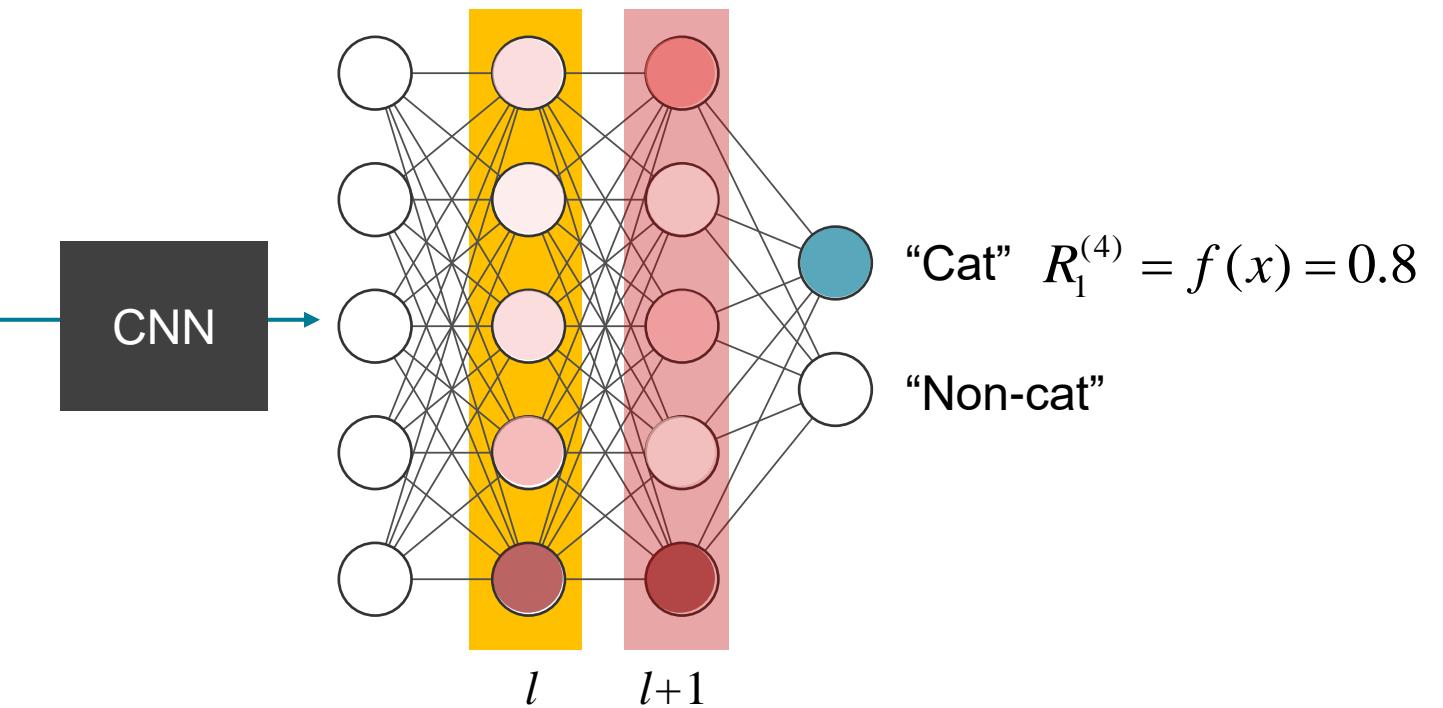
# Layer-wise relevance propagation (LRP)



$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_k z_{kj}} R_j^{(l+1)} \quad \text{with} \quad z_{ik} = x_i^{(l)} w_{ij}^{(l,l+1)}$$

Example from: <https://www.youtube.com/watch?v=PDRewtcqmal>  
Bach et al. (2015). PloS one 10, e0130140.

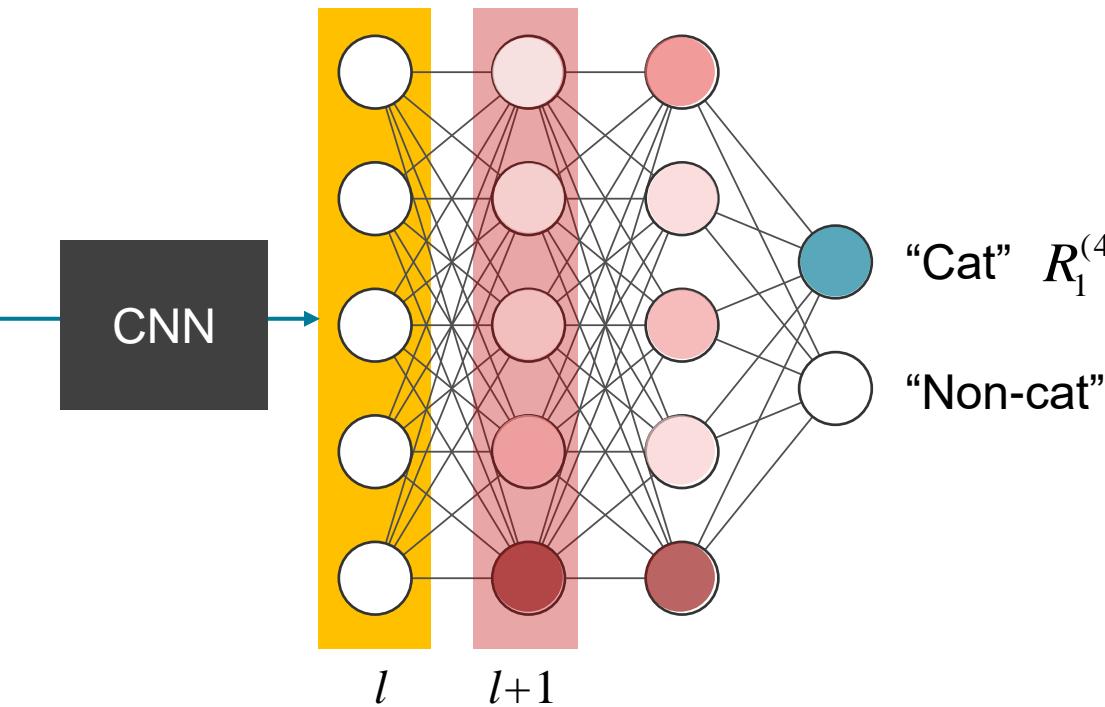
# Layer-wise relevance propagation (LRP)



$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_k z_{kj}} R_j^{(l+1)} \quad \text{with} \quad z_{ik} = x_i^{(l)} w_{ij}^{(l,l+1)}$$

Example from: <https://www.youtube.com/watch?v=PDRewtcqmal>  
Bach et al. (2015). PloS one 10, e0130140.

# Layer-wise relevance propagation (LRP)

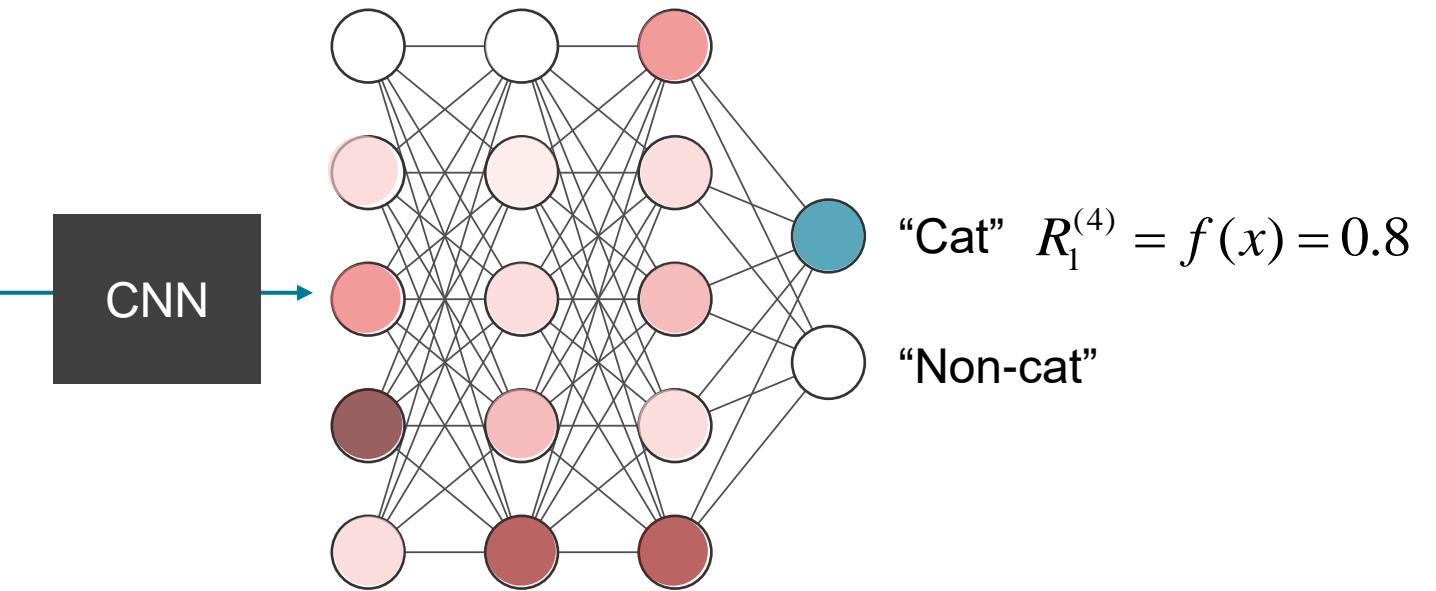


$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_k z_{kj}} R_j^{(l+1)} \quad \text{with} \quad z_{ik} = x_i^{(l)} w_{ij}^{(l,l+1)}$$

"Cat"  $R_1^{(4)} = f(x) = 0.8$   
"Non-cat"

Example from: <https://www.youtube.com/watch?v=PDRewtcqmal>  
Bach et al. (2015). PloS one 10, e0130140.

# Layer-wise relevance propagation (LRP)



$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_k z_{kj}} R_j^{(l+1)} \quad \text{with} \quad z_{ik} = x_i^{(l)} w_{ij}^{(l,l+1)}$$

Example from: <https://www.youtube.com/watch?v=PDRewtcqmal>  
Bach et al. (2015). *PLoS one* **10**, e0130140.

# Layer-wise relevance propagation (LRP)

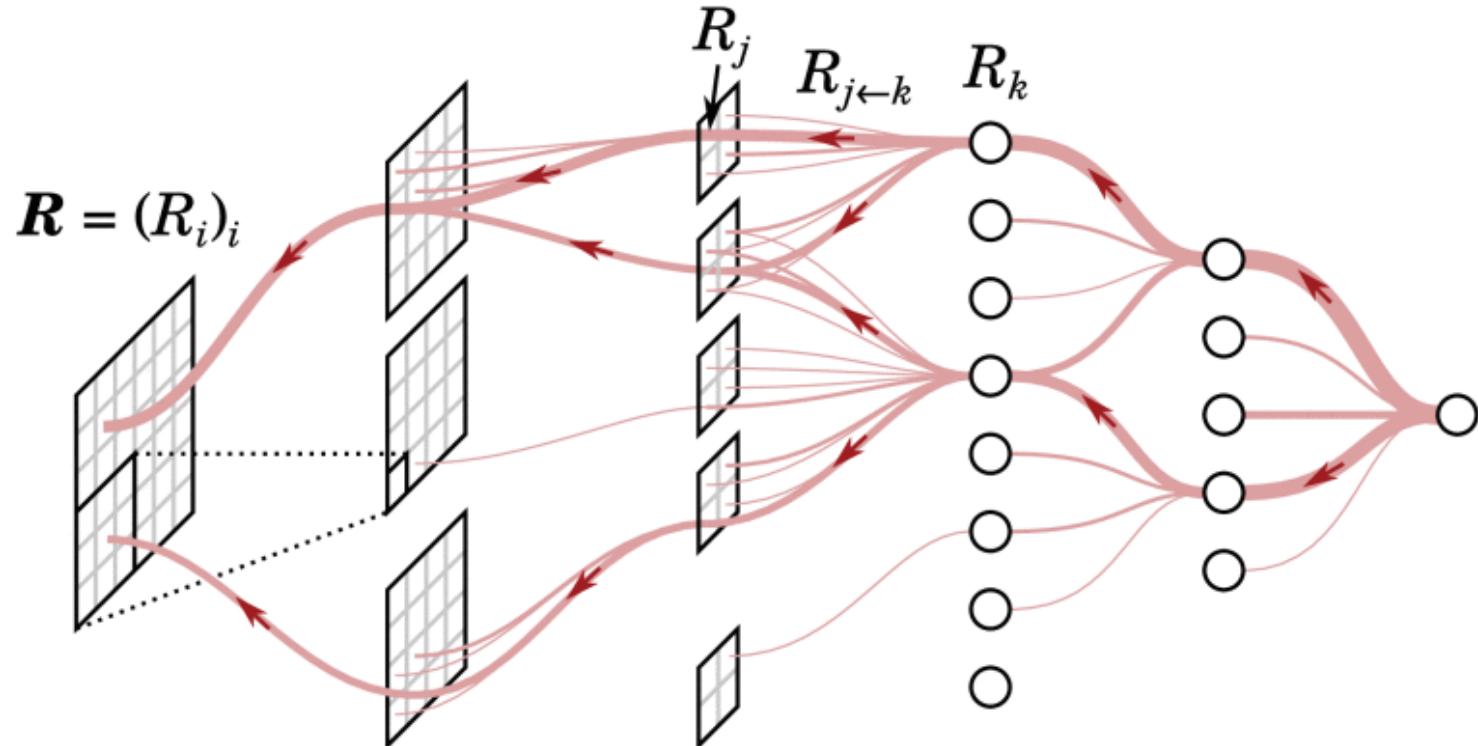
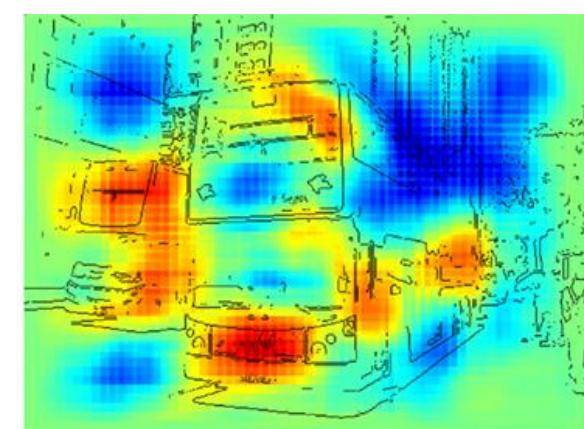
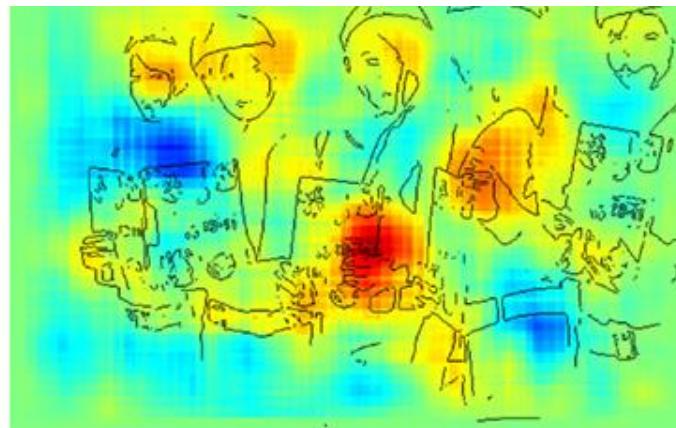
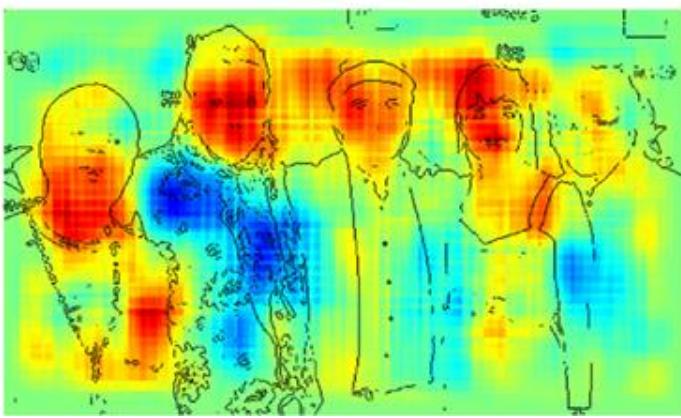


Figure from: Samek et al. (2021). *IEEE Proceedings* **109**, 247.  
Bach et al. (2015). *PLoS one* **10**, e0130140.

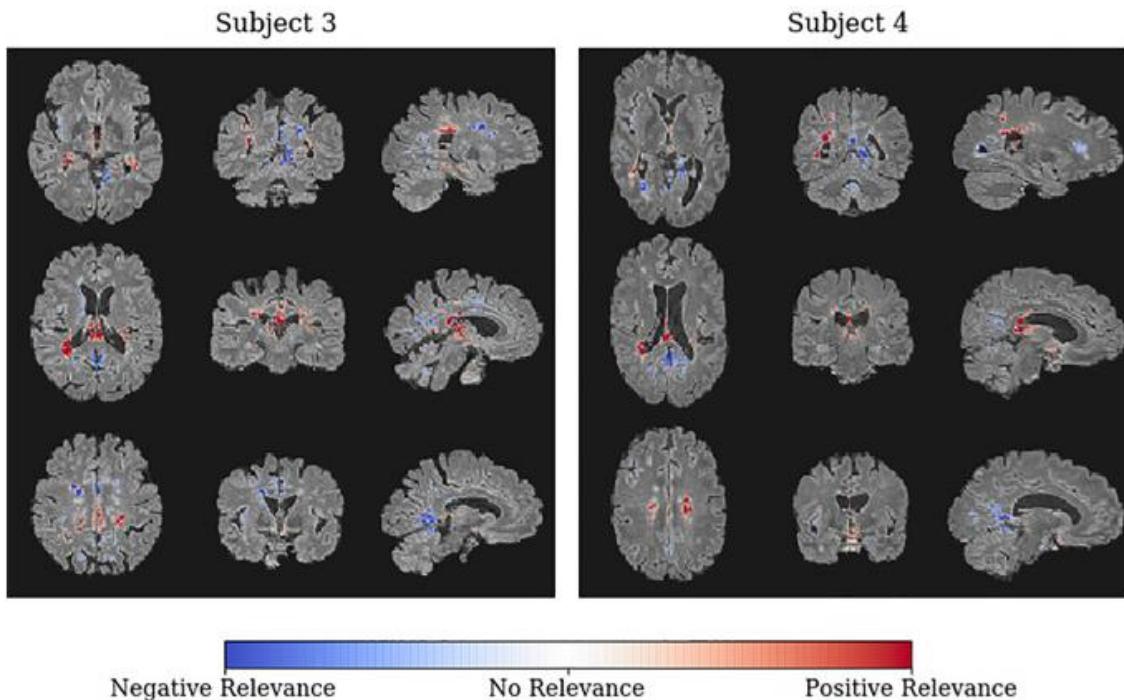
# Layer-wise relevance propagation (LRP)



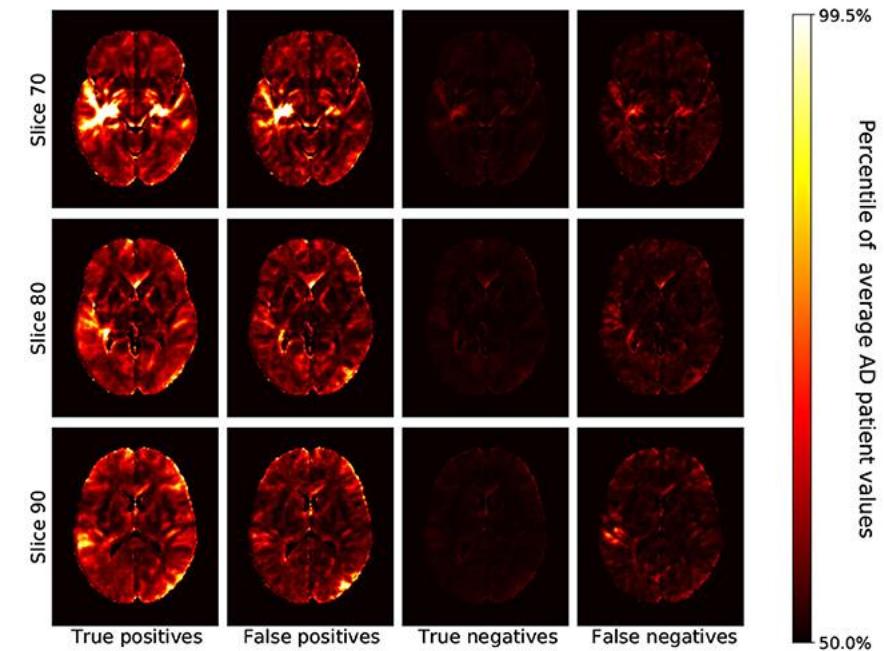
Bach et al. (2015). *PLoS one* **10**, e0130140.

# LRP in diagnostic imaging

# **Diagnosis of multiple sclerosis<sup>1</sup>**



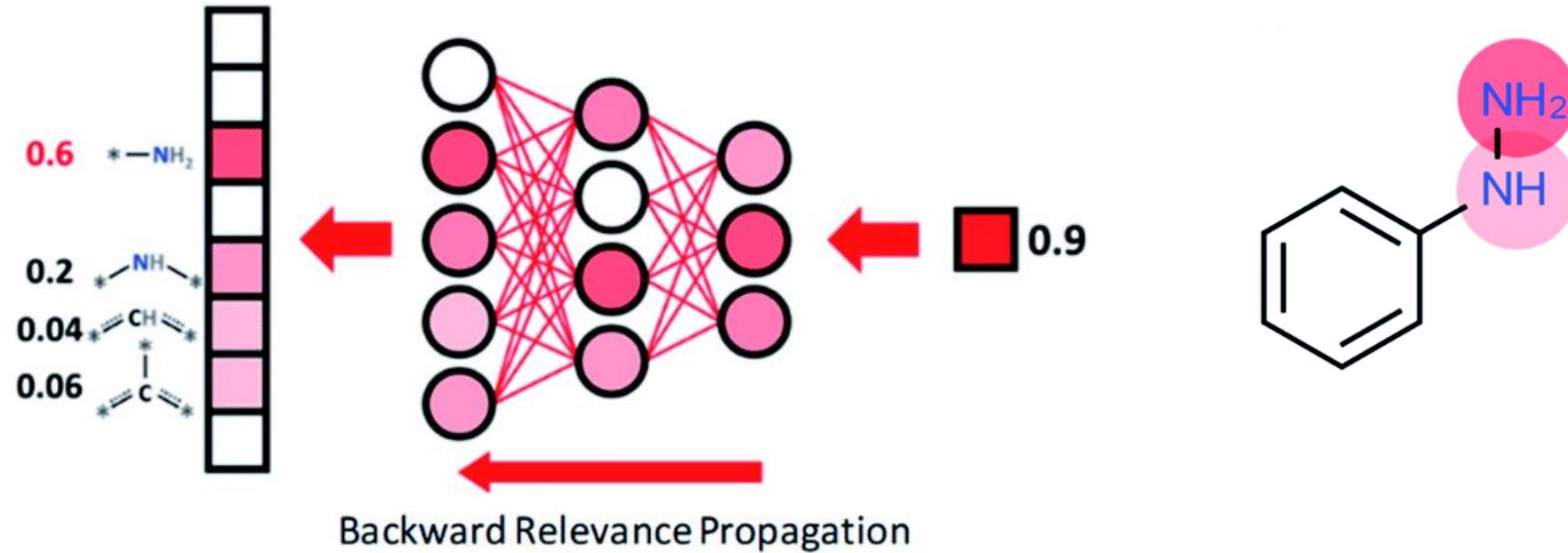
## Alzheimer's disease classification<sup>2</sup>



<sup>1</sup>Eitel et al. (2019). *NeuroImage: Clinical* 24, 102003.

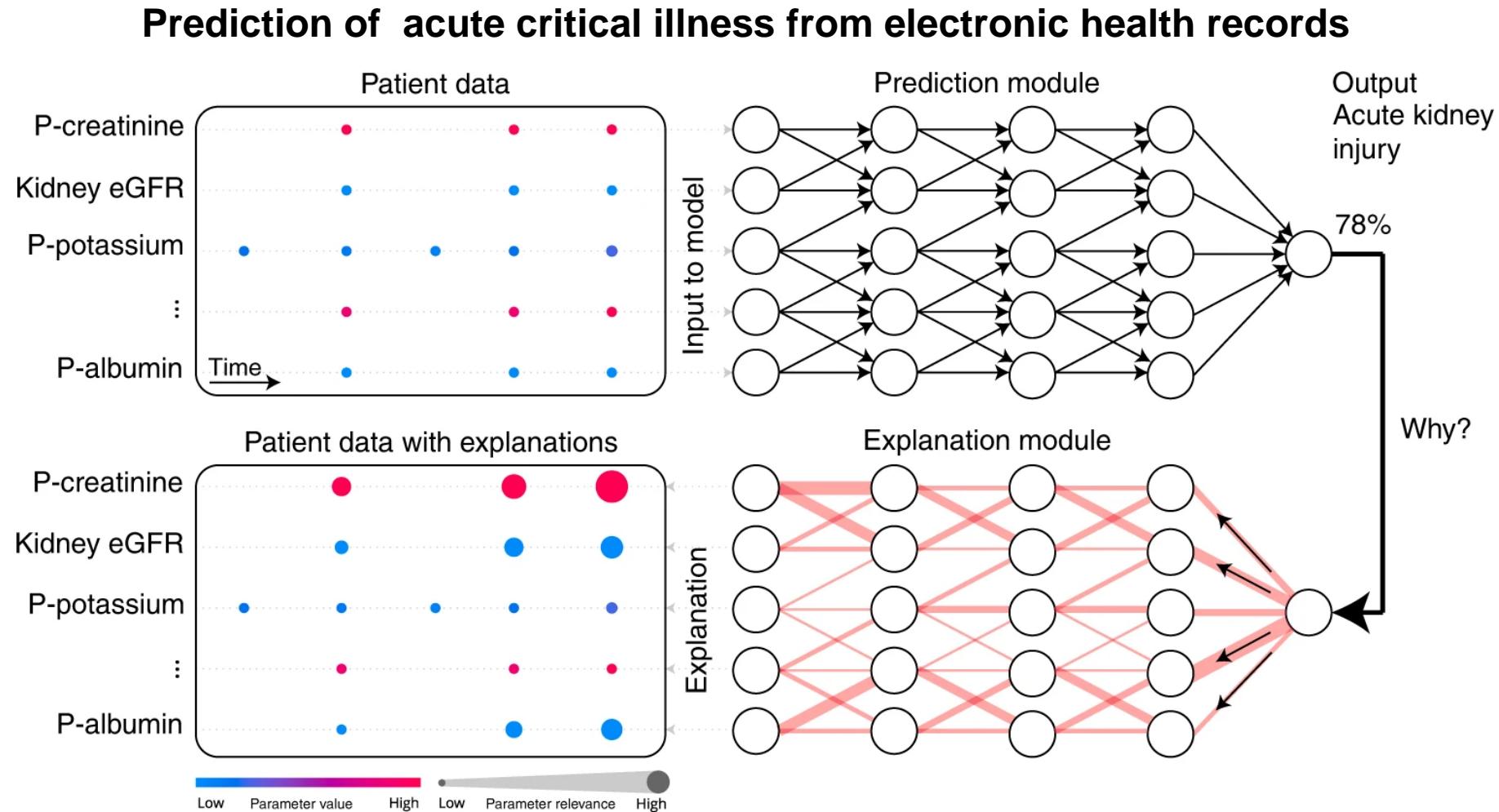
<sup>2</sup>Böhle et al. (2019). Frontiers Aging Neurosci. 11, 194.

# Other usages of LRP



Kim et al. (2021) *Chem. Sci.* **12**, 11028.

# Other usages of LRP



Lauritsen et al. (2020). *Nature Communications* 11, 11.

# Pixel (feature) attribution methods

## Pros

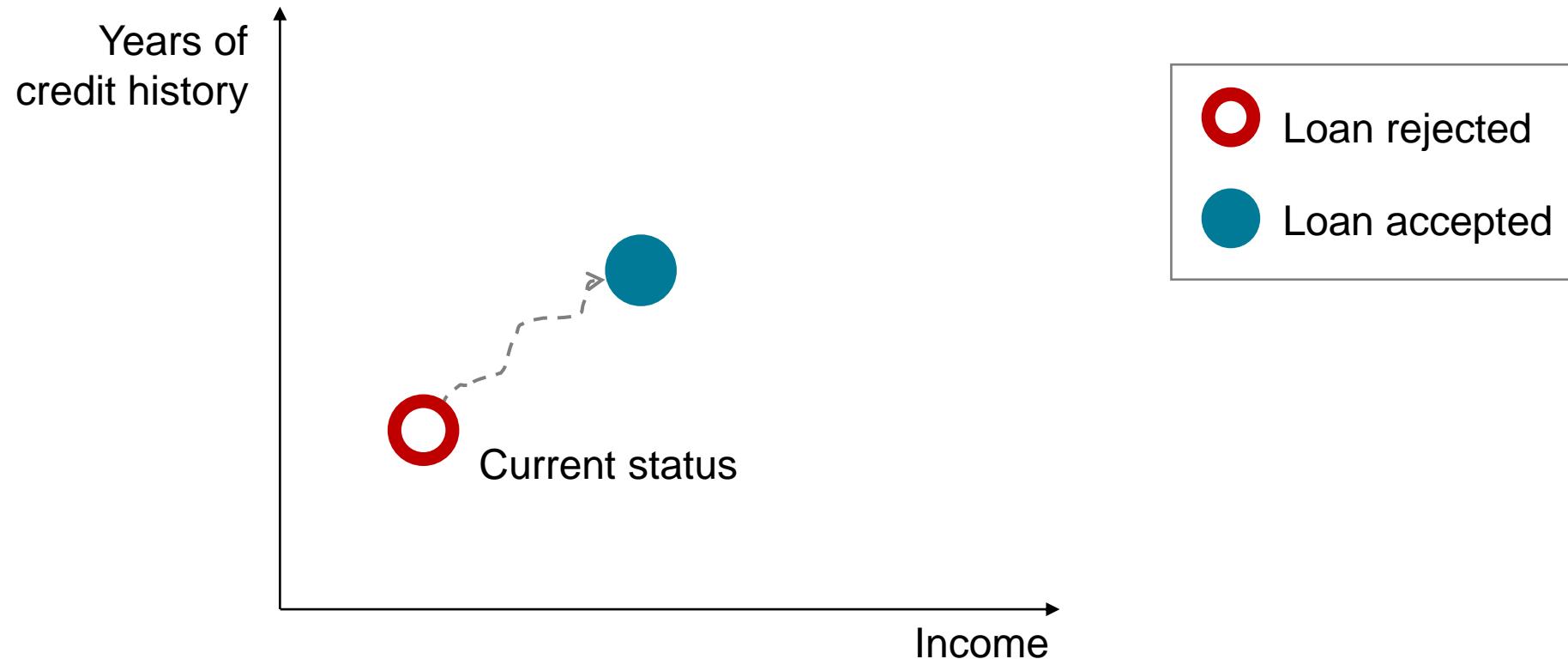
- Visual and quick to inspect
- Important regions in the image
- Allow to detect certain shortcomings

## Cons

- Difficult to know whether an explanation is correct
- Small perturbations to an image can lead to very different pixels highlighted
- Often very qualitative

# Counterfactual explanations (Instance based)

**Counterfactual:** smallest change in the input features that changes the predictions to another output.



# Adversarial attacks

**Adversarial examples:** generate false predictions by leveraging the shortages of the algorithm.

- AI safety
- Make machine learning models more secure against manipulations

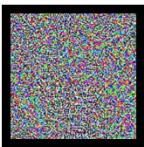


Minimal change in the input to generate a different output



"panda"

Adversarial Noise



+

=

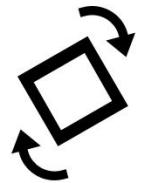


"gibbon"



"vulture"

Adversarial Rotation



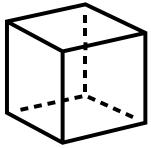
+

=

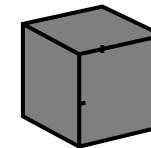
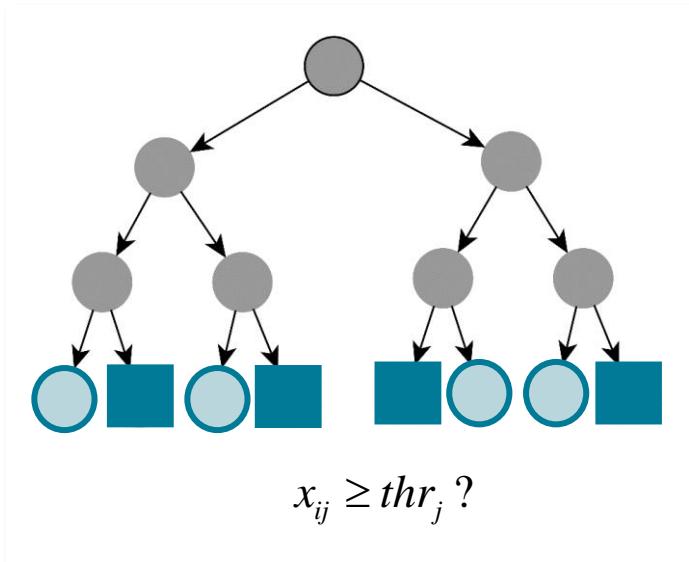


"orangutan"

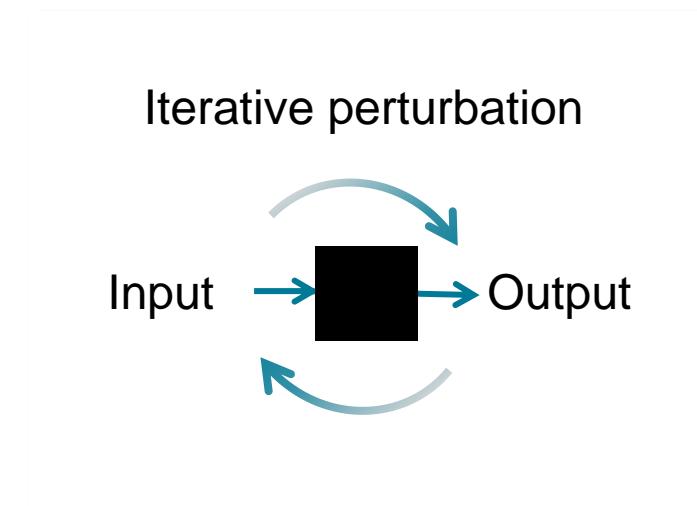
# Counterfactuals



White box → model-dependent

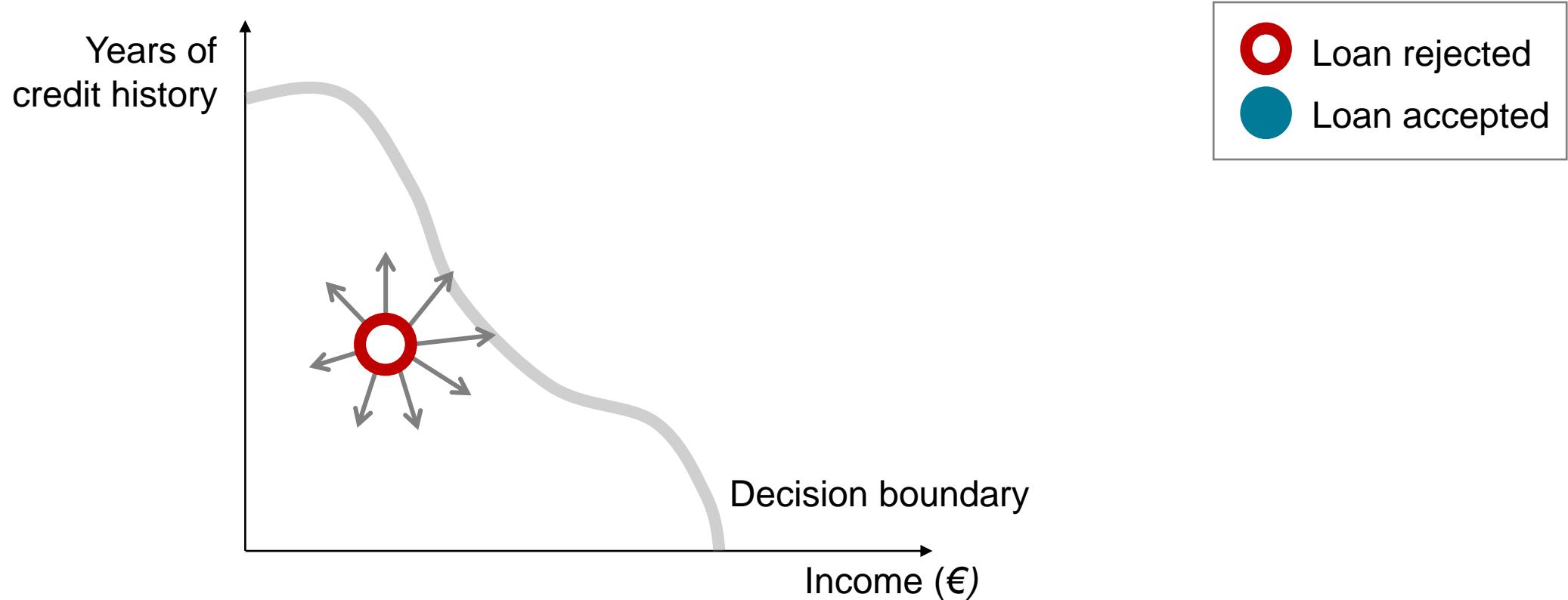


Black box → model-agnostic



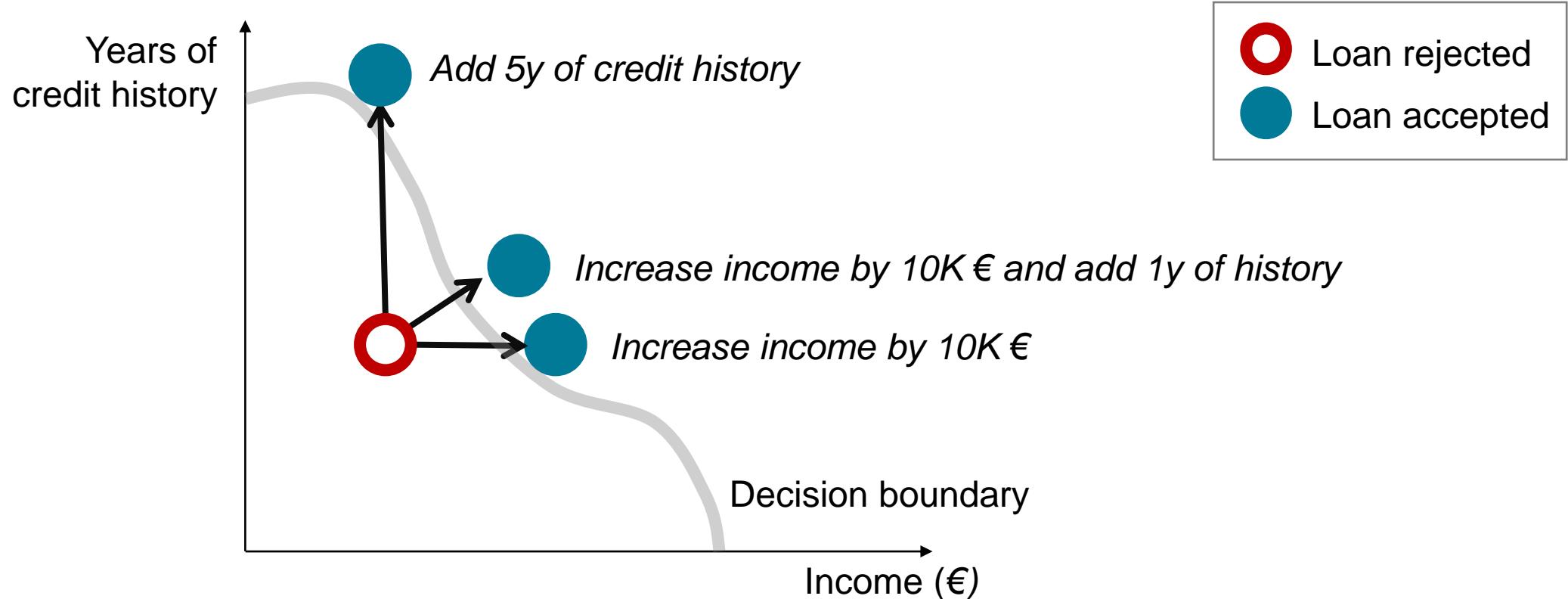
# Counterfactual explanations

**Counterfactual:** smallest change in the input features that changes the predictions to another output.

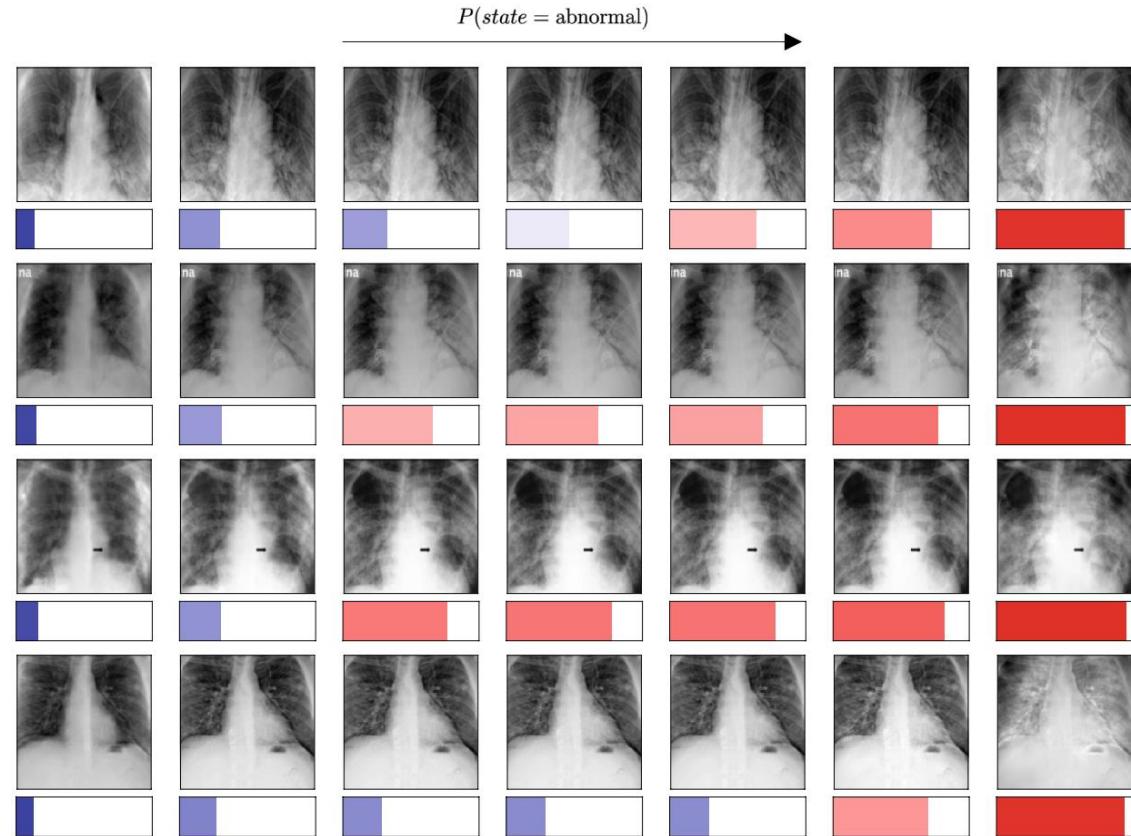
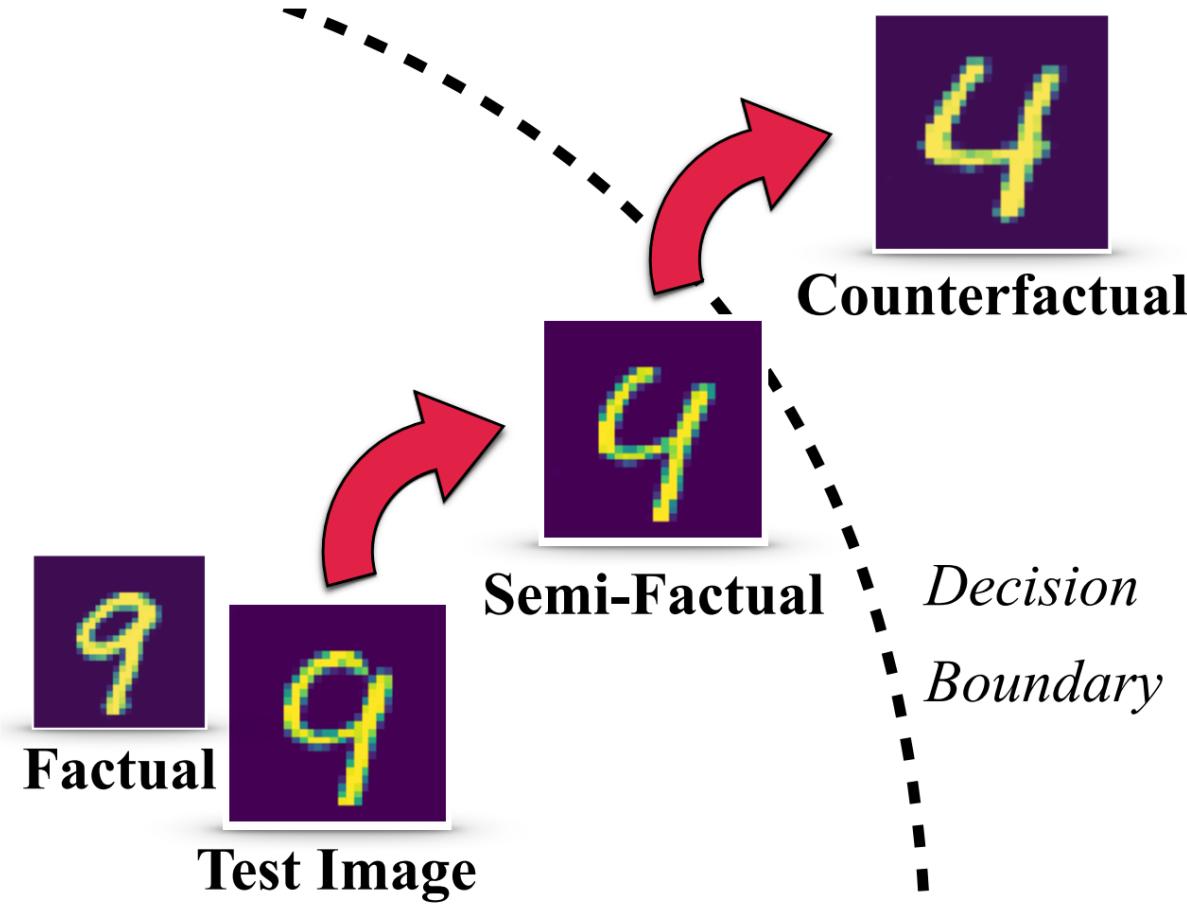


# Counterfactual explanations

**Counterfactual:** smallest change in the input features that changes the predictions to another output.



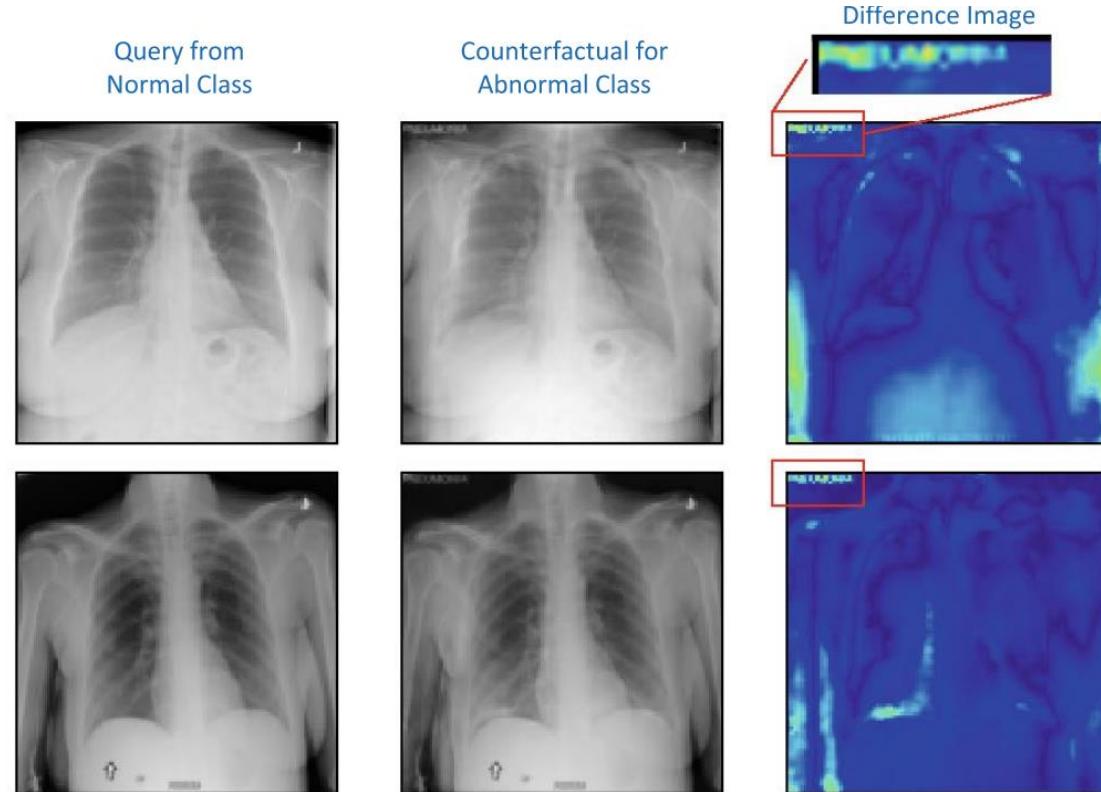
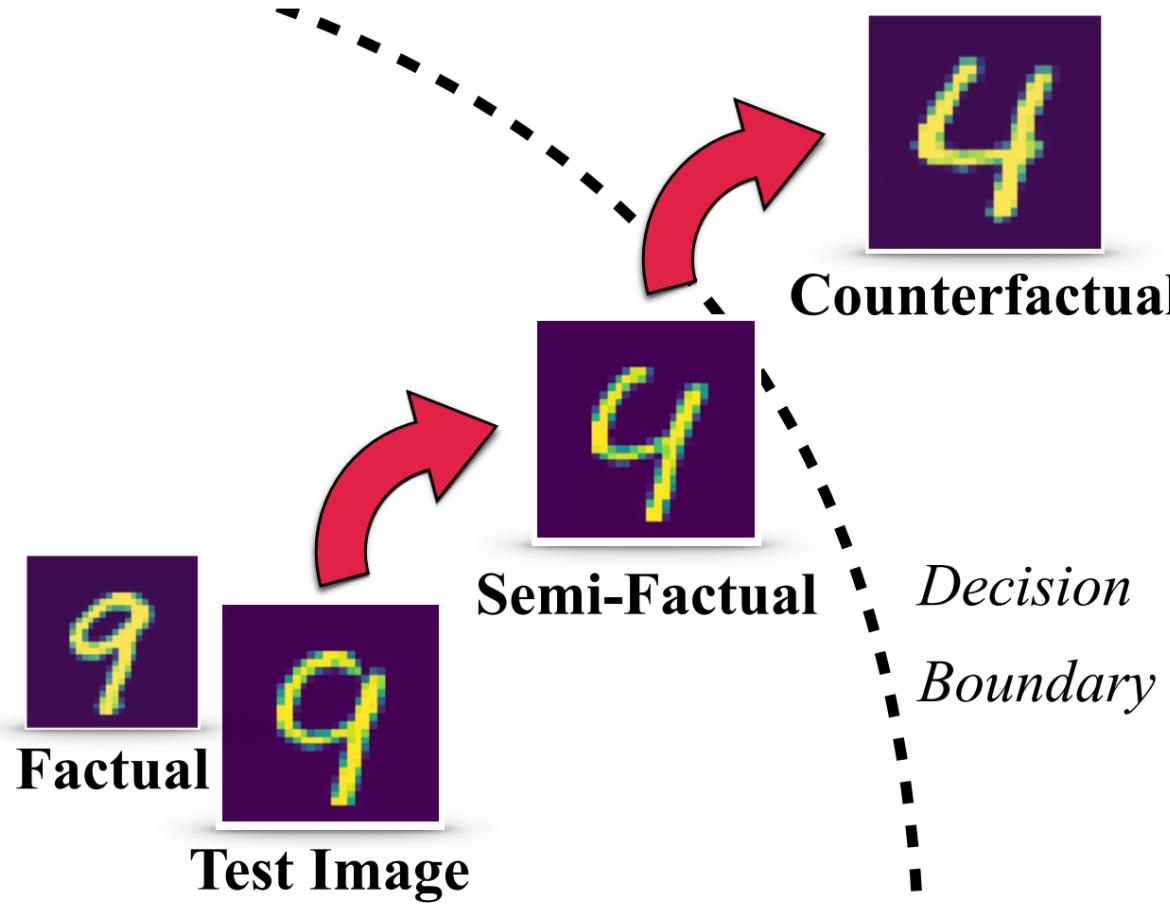
# Counterfactual explanations



Kenn and Keane (2021), Proceedings of the AAAI Conference on Artificial Intelligence 35, 11575.

Thiagarajan et al. (2022) Sci Rep 12, 597.

# Counterfactual explanations



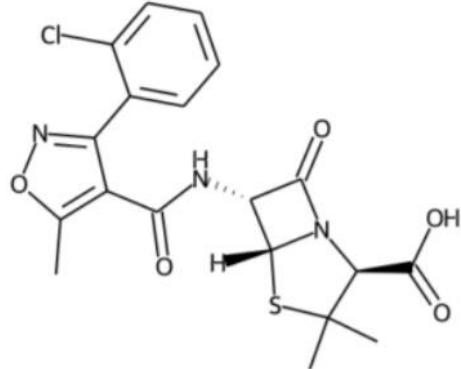
Kenn and Keane (2021, *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 11575).

Thiagarajan et al. (2022) *Sci Rep* 12, 597.

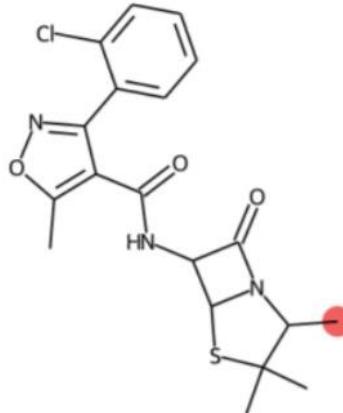
# Counterfactual explanations for molecules

Blood-brain barrier permeating molecules

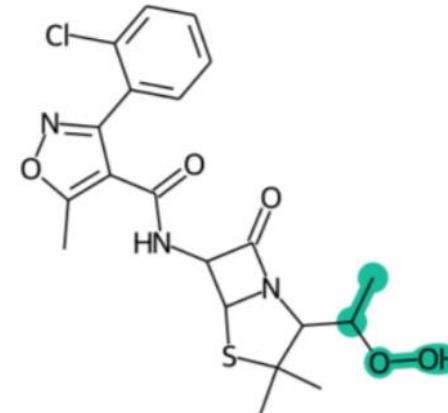
Base  
 $f(x) = 0.000$



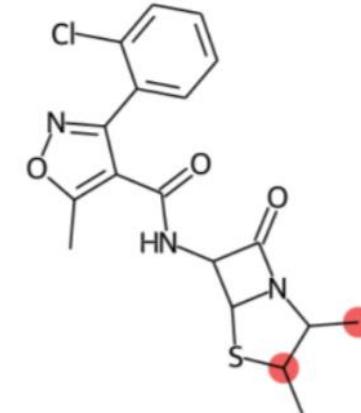
Similarity = 0.80  
Counterfactual 1  
 $f(x) = 1.000$



Similarity = 0.75  
Counterfactual 2  
 $f(x) = 1.000$



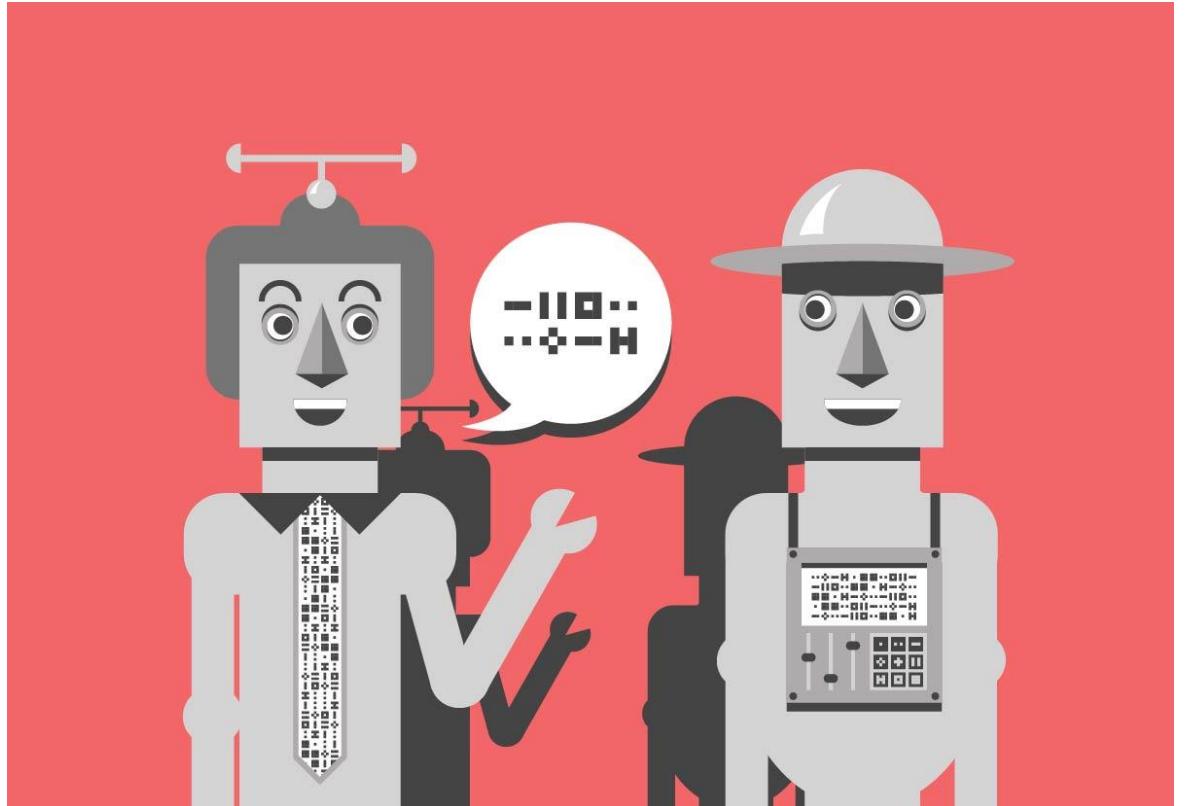
Similarity = 0.68  
Counterfactual 3  
 $f(x) = 1.000$



Wellawatte et al. (2021). ChemRxiv.10.33774/chemrxiv-2021-4qkg8-v2.

# Self-explaining approaches

- Intrinsic explanation.
- Glimpse into the inner functioning of the model.
- Promote verification and error analysis, and be directly linkable with domain knowledge.



<https://byrds.ch/en/ai-algorithms-talking/>

# Feature attribution vs feature visualization (Self-explaining)

**Feature attribution**

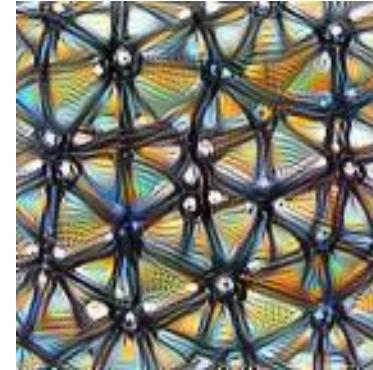


What part of the example is  
responsible for the prediction?



*Image in*

**Feature visualization**



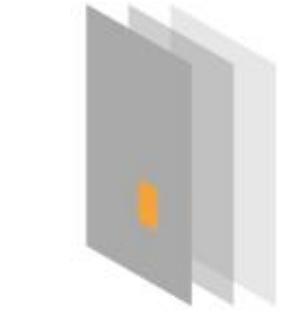
What is a network looking for?



*Image out*

Images from: <https://distill.pub/2017/feature-visualization/>

# Feature visualization



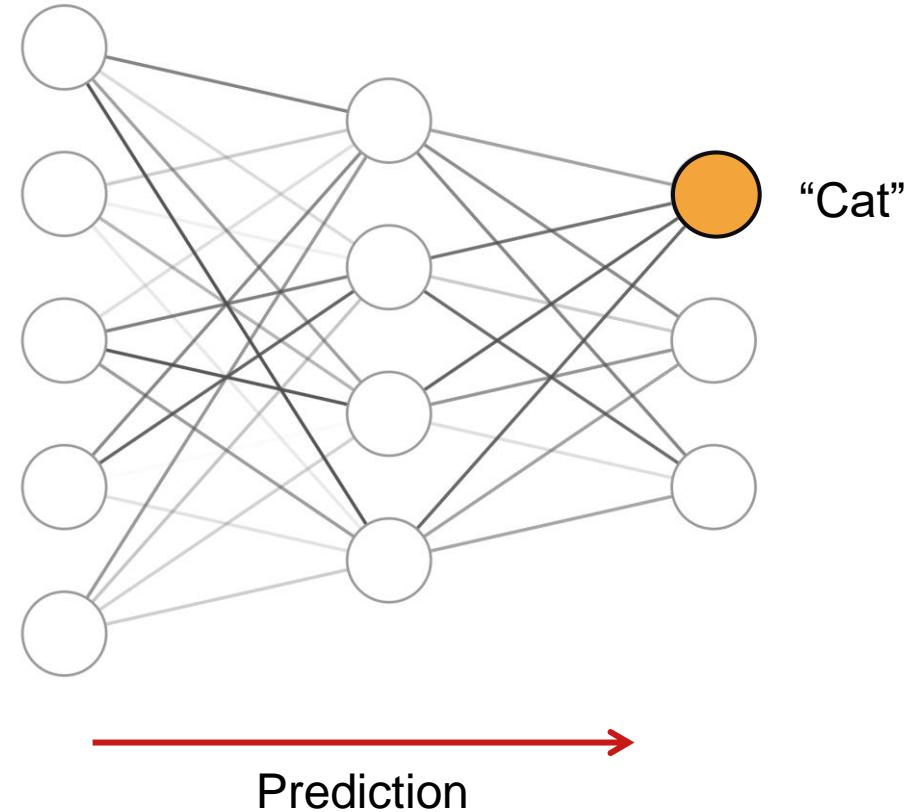
Neuron



Layer/DeepDream

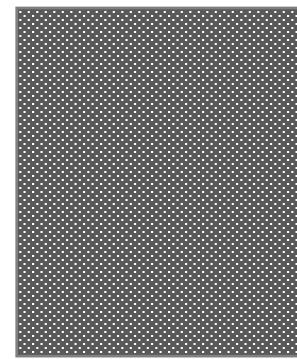


Real picture

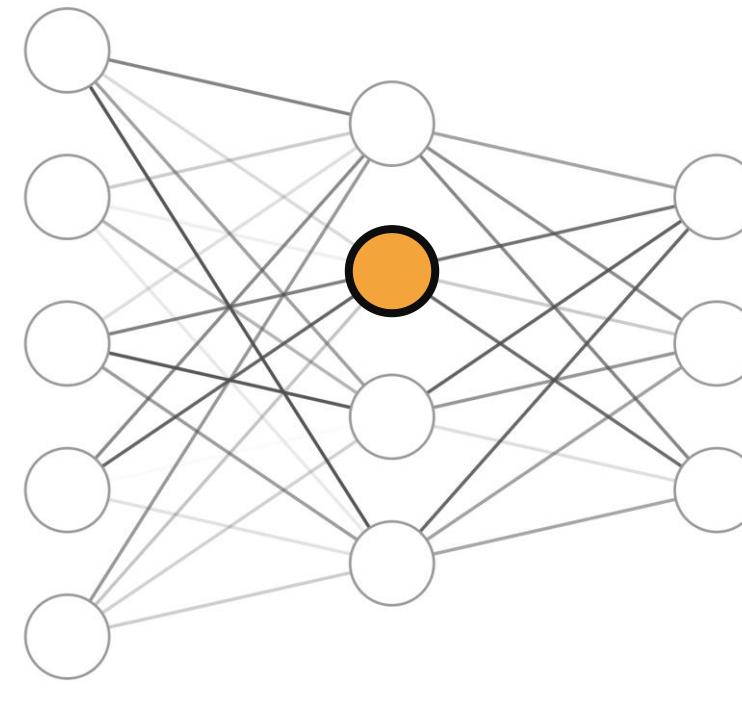


Source: <https://ai.googleblog.com/2018/03/the-building-blocks-of-interpretability.html>.

# Feature visualization



Random noise



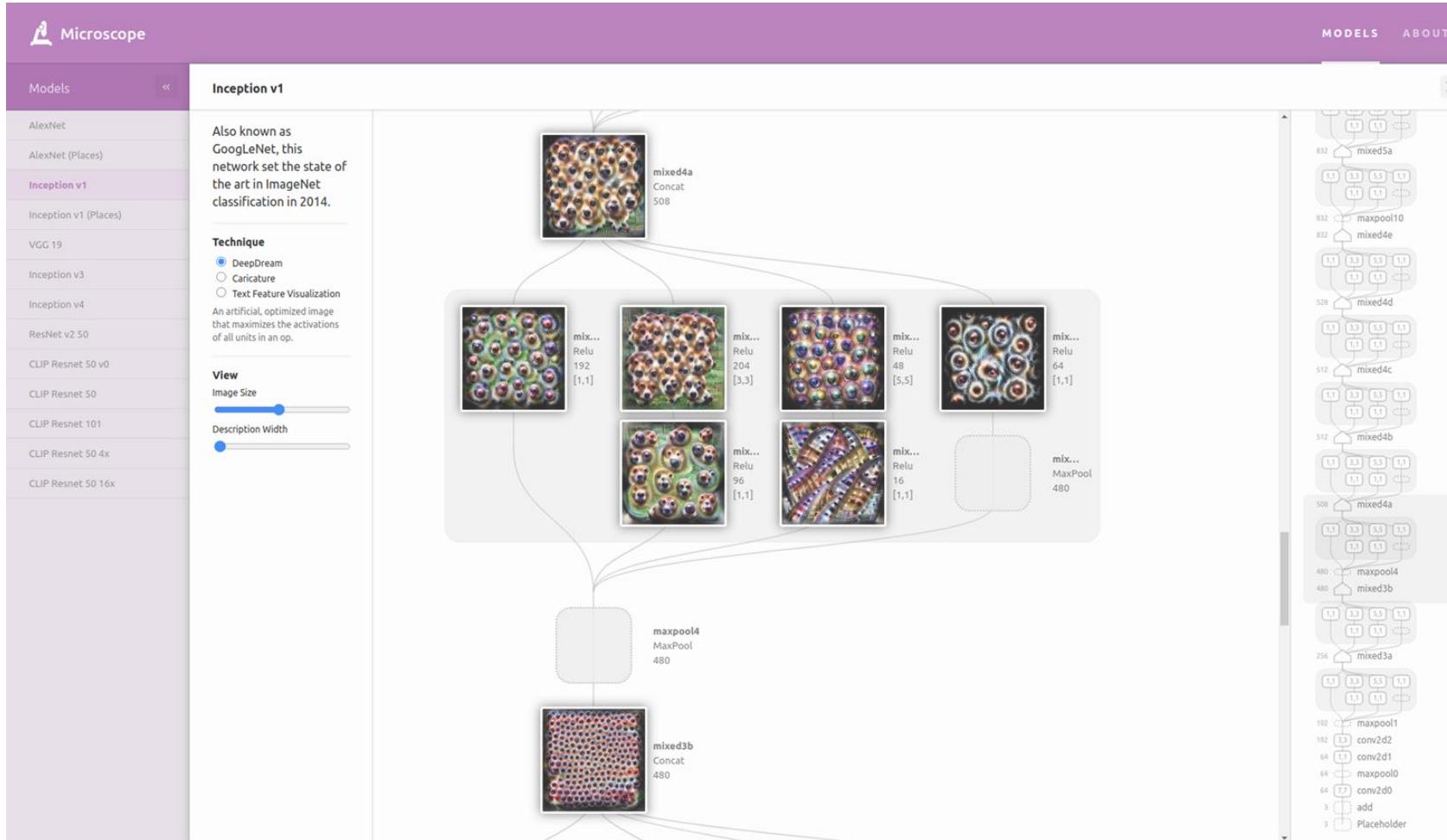
Source: <https://ai.googleblog.com/2018/03/the-building-blocks-of-interpretability.html>.

# Visualizing layers and neurons



Images from: <https://distill.pub/2017/feature-visualization/>

# OpenAI microscope



[https://microscope.openai.com/models/inceptionv1?models.technique=deep\\_dream](https://microscope.openai.com/models/inceptionv1?models.technique=deep_dream)

# Feature visualization

## Pros

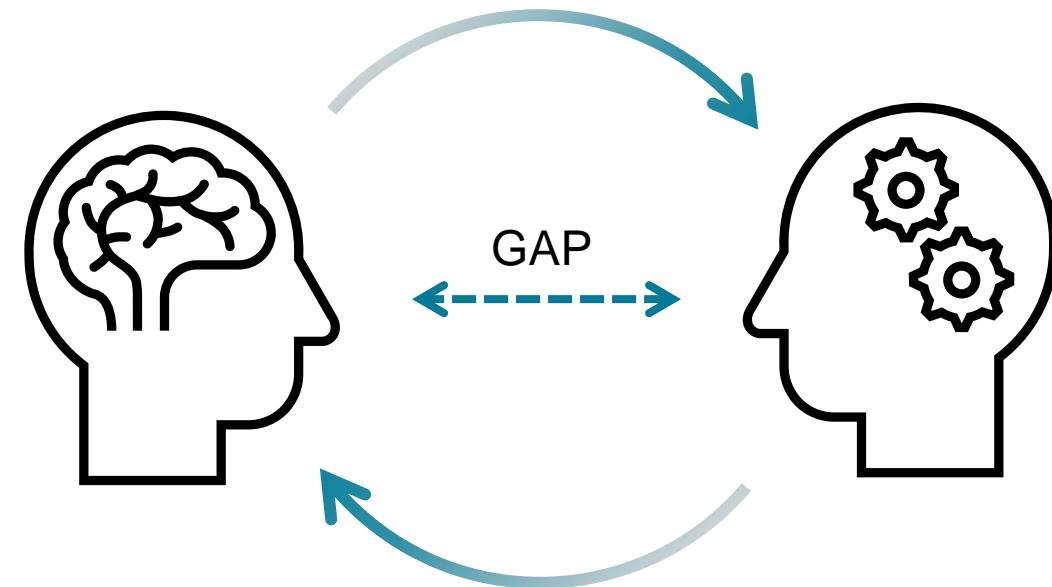
- Automatically link units to concepts
- Non-technical communication
- Detect concepts beyond the class labels
- Make great t-shirts/posters!

## Cons

- Some images are not interpretable at all
- Too many units to look at
- Illusion of interpretability

# (Personal) reflections on Explainable AI

- Increasing applications
- Innumerable approaches
- High flexibility

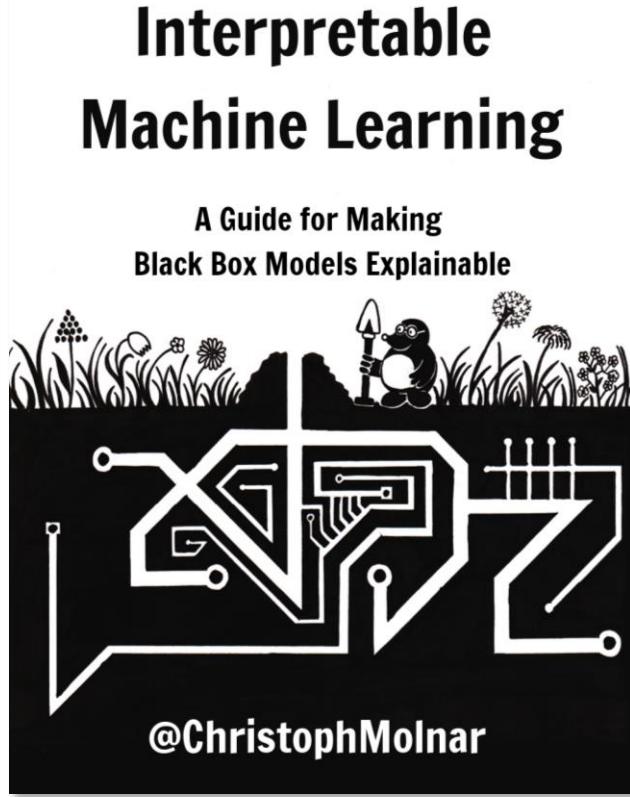


- How to interpret?
- Partial information
- Lack of systematic evaluation

# Summing it up

- XAI: information on relationships within the data and/or predictions using a model.
- Local or global, model-agnostic or model-dependent.
- The choice depends on what you want to achieve, and on the model/input data.
- Explanation is not interpretation.
- Lack of systematic evaluation of explainability vs interpretability/causality.

# Want to know more?



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

This image is a screenshot of a scientific article from 'nature machine intelligence'. The title is 'Drug discovery with explainable artificial intelligence'. The authors listed are José Jiménez-Luna, Francesca Grisoni, and Gisbert Schneider. The abstract discusses the promise of deep learning for drug discovery and the need for explainability. The page includes a 'Check for updates' button and a 'REVIEW ARTICLE' section with a DOI link.

**Drug discovery with explainable artificial intelligence**

José Jiménez-Luna<sup>1,2</sup>, Francesca Grisoni<sup>1,2</sup> and Gisbert Schneider<sup>1,2</sup>

Deep learning bears promise for drug discovery, including advanced image analysis, prediction of molecular structure and function, and automated generation of innovative chemical entities with bespoke properties. Despite the growing number of successful predictive applications, the underlying mathematical models often remain elusive to interpretation by the human mind. This lack of 'explainability' is a major challenge to address the gap between AI and the language of the molecular sciences. This Review summarizes the most promising algorithmic concepts of explainable artificial intelligence, and forecasts future opportunities, potential applications as well as several remaining challenges. We also hope it encourages additional efforts towards the development and acceptance of explainable artificial intelligence techniques.

<https://www.nature.com/articles/s42256-020-00236-4>

This image shows a YouTube channel interface for 'EXPLAINABLE AI EXPLAINED!'. The main video thumbnail features a purple and orange abstract background with a brain icon. Below the video, the title 'EXPLAINABLE AI EXPLAINED!' is displayed. The channel has 1,044 subscribers and 104,112 views. The video duration is 1:04. The channel description reads: 'Explaining AI models for drug discovery and beyond'. The video player shows the first few frames of the video. Below the video, there is a list of six video thumbnails, each with a title, duration, and a 'DeepFindr' logo.

**EXPLAINABLE AI EXPLAINED!**

Introduction

Explaining AI models for drug discovery and beyond

- Explainable AI explained! | #3 LIME
- Explainable AI explained! | #4 SHAP
- Explainable AI explained! | #5 Counterfactual explanations and adversarial attacks
- Explainable AI explained! | #6 Layerwise Relevance Propagation with MRI data
- How to explain Graph Neural Networks (with XAI)

<https://www.youtube.com/playlist?list=PLV8yxwGOxvvovp-j6ztxhF3QcKXT6vORU>