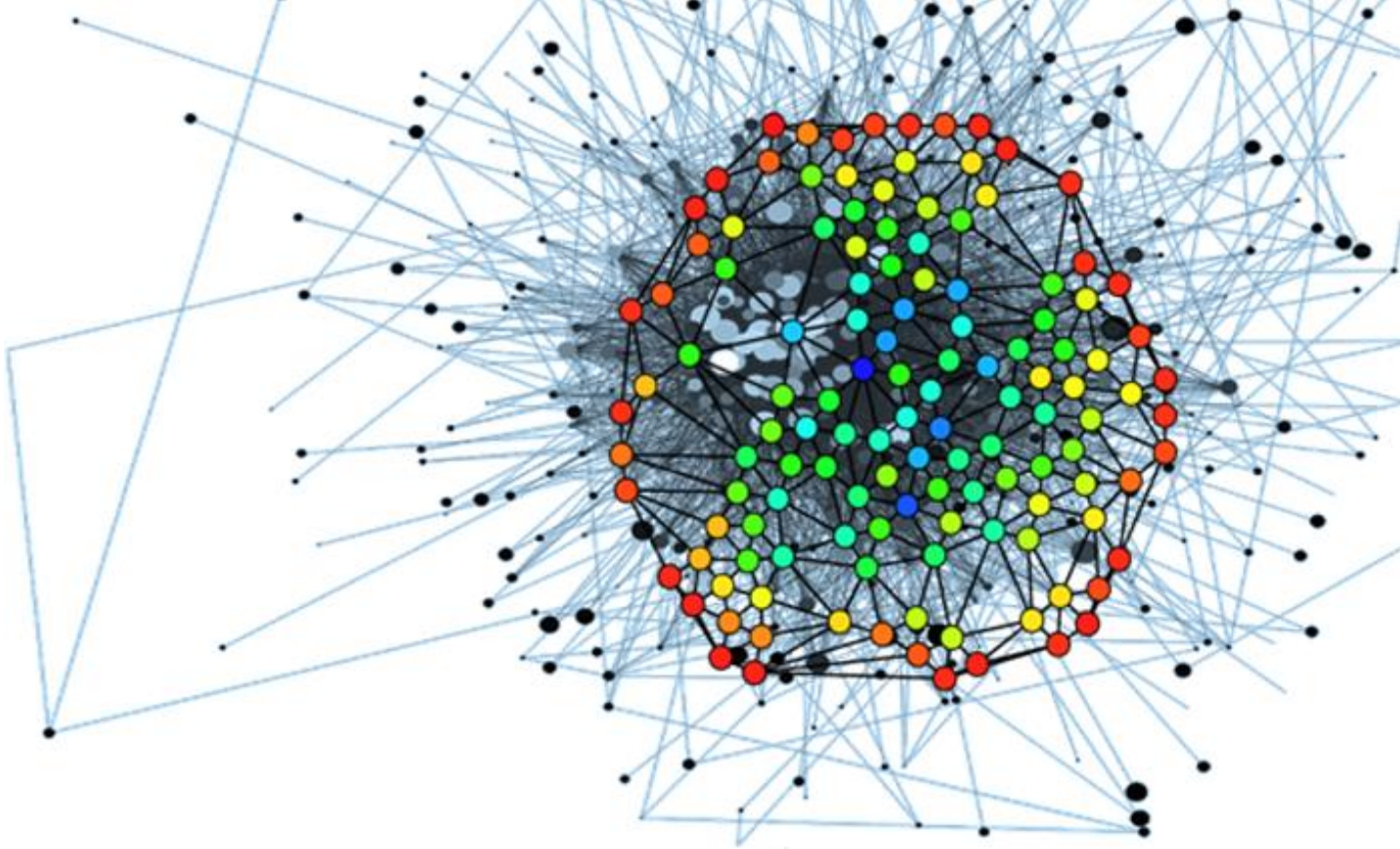


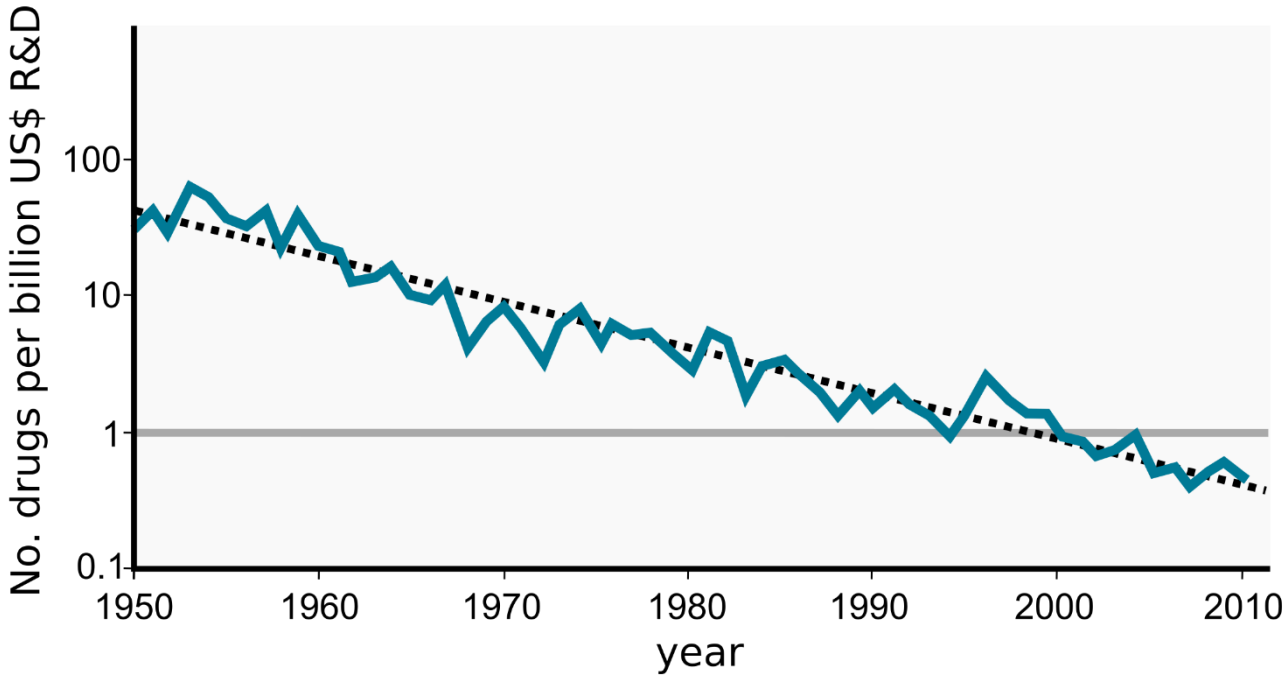
Explainable Artificial Intelligence

Francesca Grisoni, Asst. Prof.
Institute for Complex Molecular Systems
Department of Biomedical Engineering

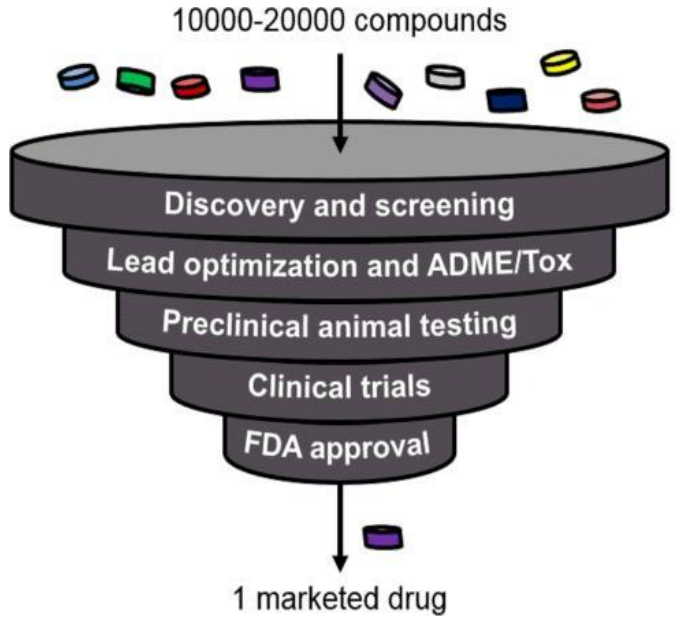
f.grisoni@tue.nl

Molecular machine learning for drug discovery

Efficiency of drug development over time[1]



Drug development pipeline[2]



"Fail early, fail cheap"

[1]Data from: Scannell *et al.* (2012). *Nature Reviews Drug Discovery* **11**, 191.
[2]Figure adapted from: Ware and Khetani (2017). *Trends in Biotechnology* **35**, 172.

TU/e

Molecular machine learning for drug discovery

Chemical Universe[1,2]

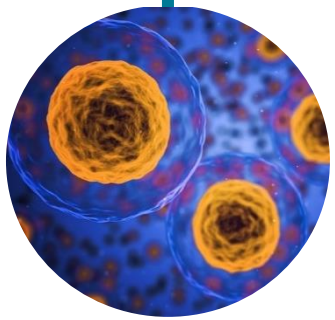
$10^{23} – 10^{100}$

Stars in the Milky Way

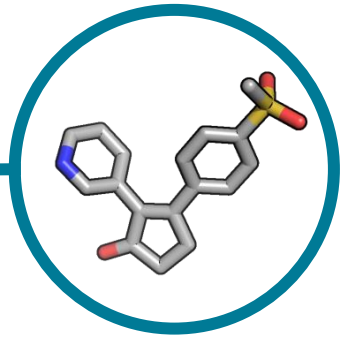
$10^{8} – 10^{9}$

Cells in a human body

$10^{13} – 10^{14}$



Known
small molecule drugs



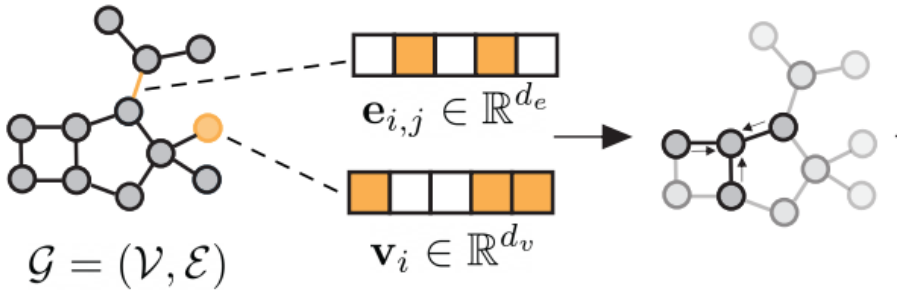
10^{4}

[1]Ertl (2002) *Journal of Chemical Information and Computer Sciences* **43**, 374.
[2]Walters et al. (1998). *Drug Discovery Today* **3**, 160.

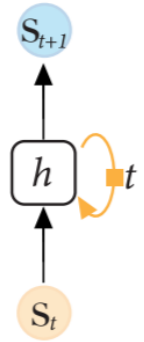
TU/e

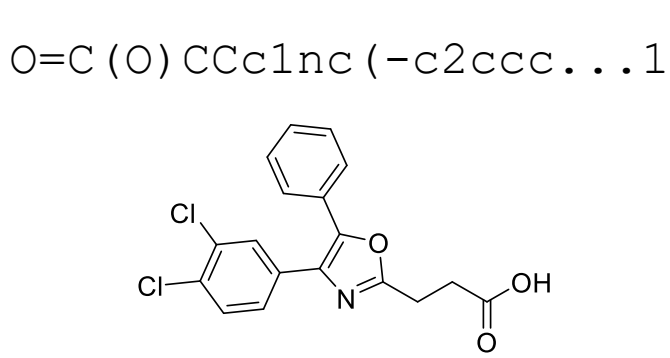
Molecular machine learning for drug discovery

Molecular property prediction with machine/deep learning

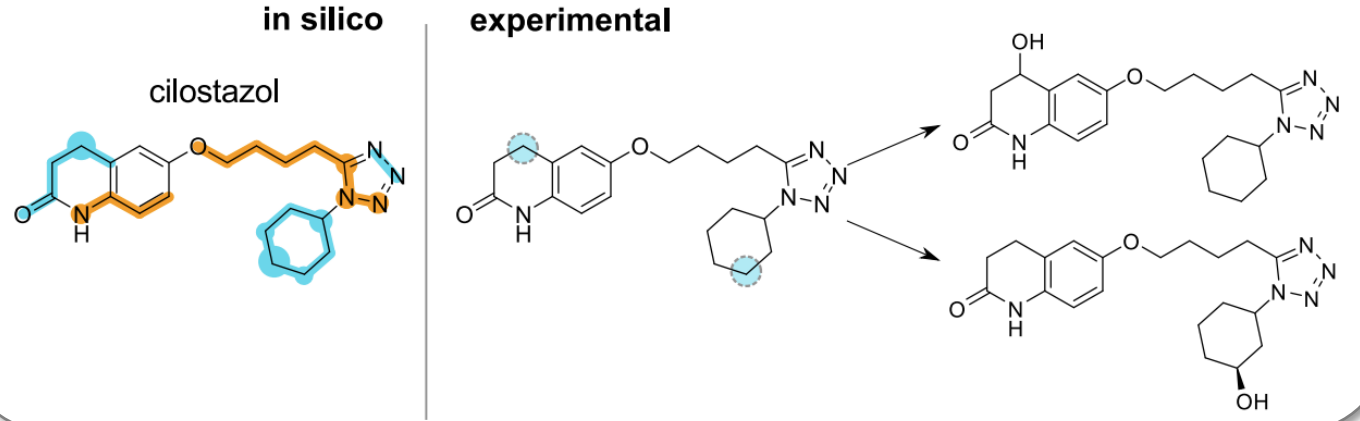


De novo molecule design with generative AI



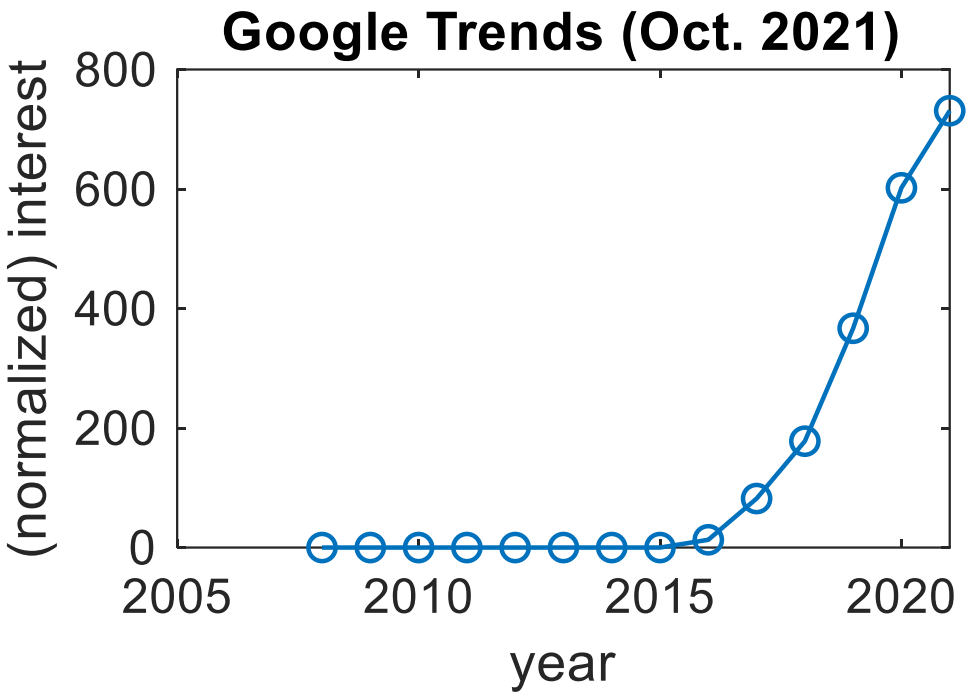


Explaining molecule predictions



TU/e

Explainable AI (XAI)

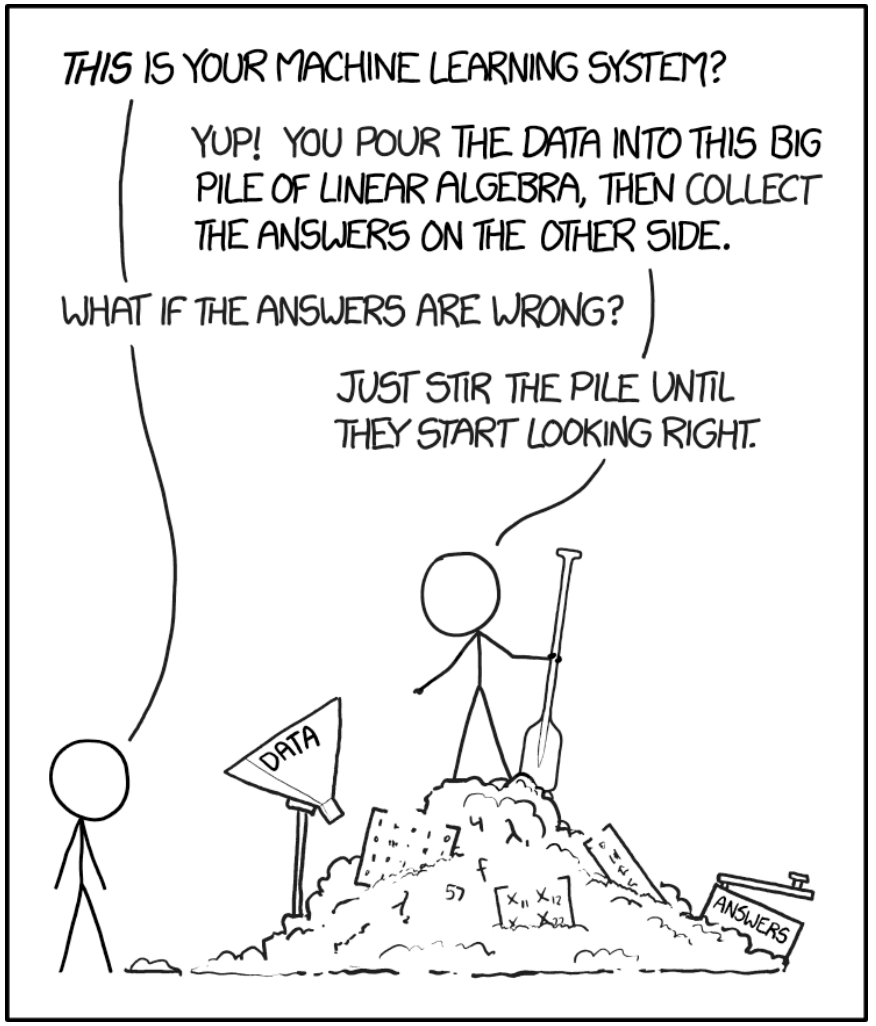


TU/e

Explainable AI (XAI)

Extraction of **relevant knowledge** from a machine learning model concerning **relationships** either 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*?



TU/e

Explainability vs interpretability (!)

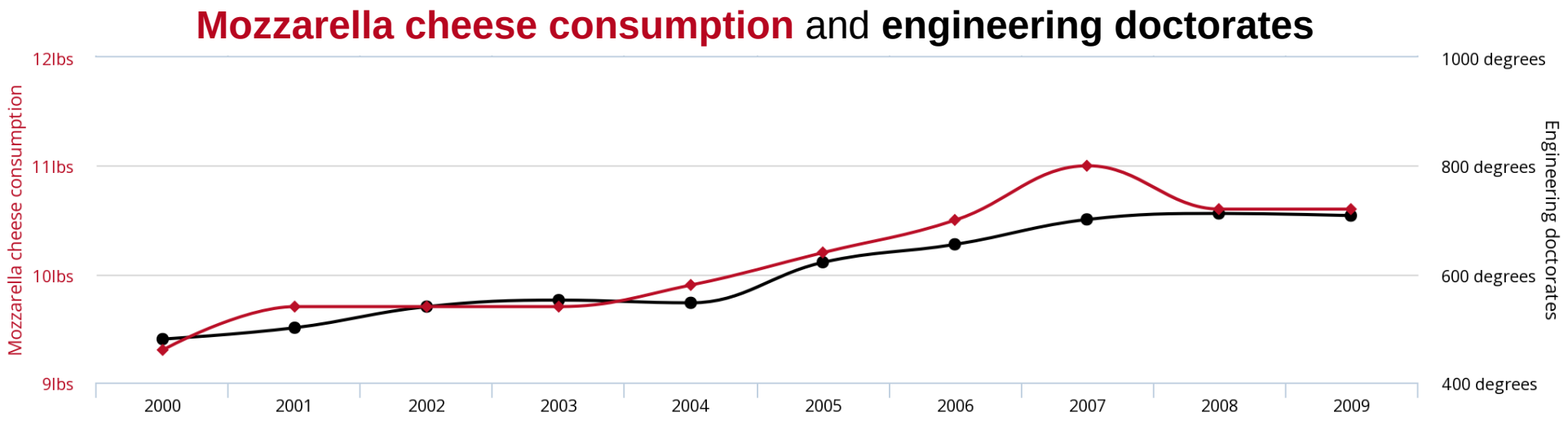
Explainability

Relationships within your data or in the model predictions

≠

Interpretability

Determination of cause-effect relationships



TU/e

Explainability-performance trade-off

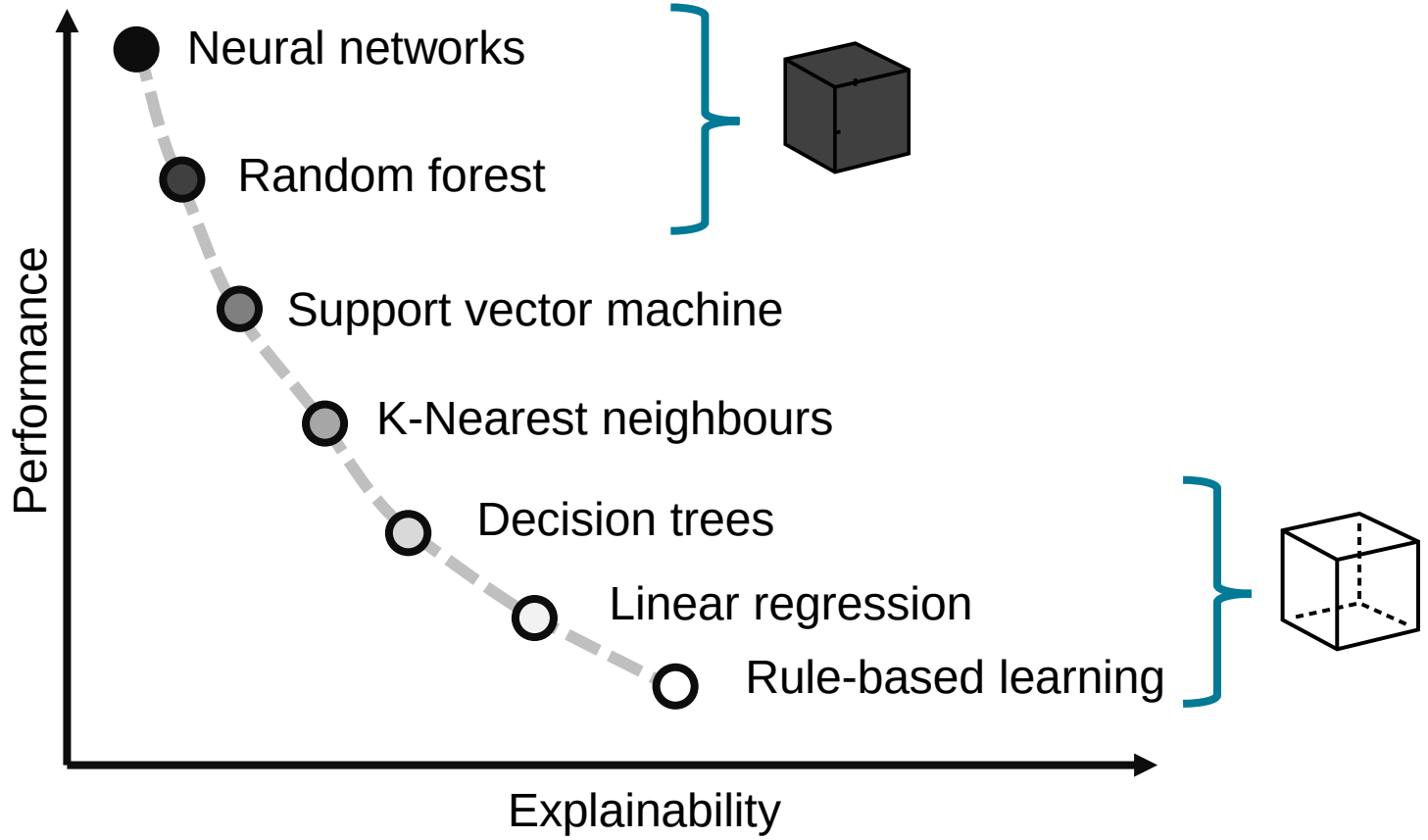


Figure inspired by: Morocho-Cayamcela *et al.* (2019) *IEEE Access* **7**, 137184.

TU/e

Explainability-performance trade-off

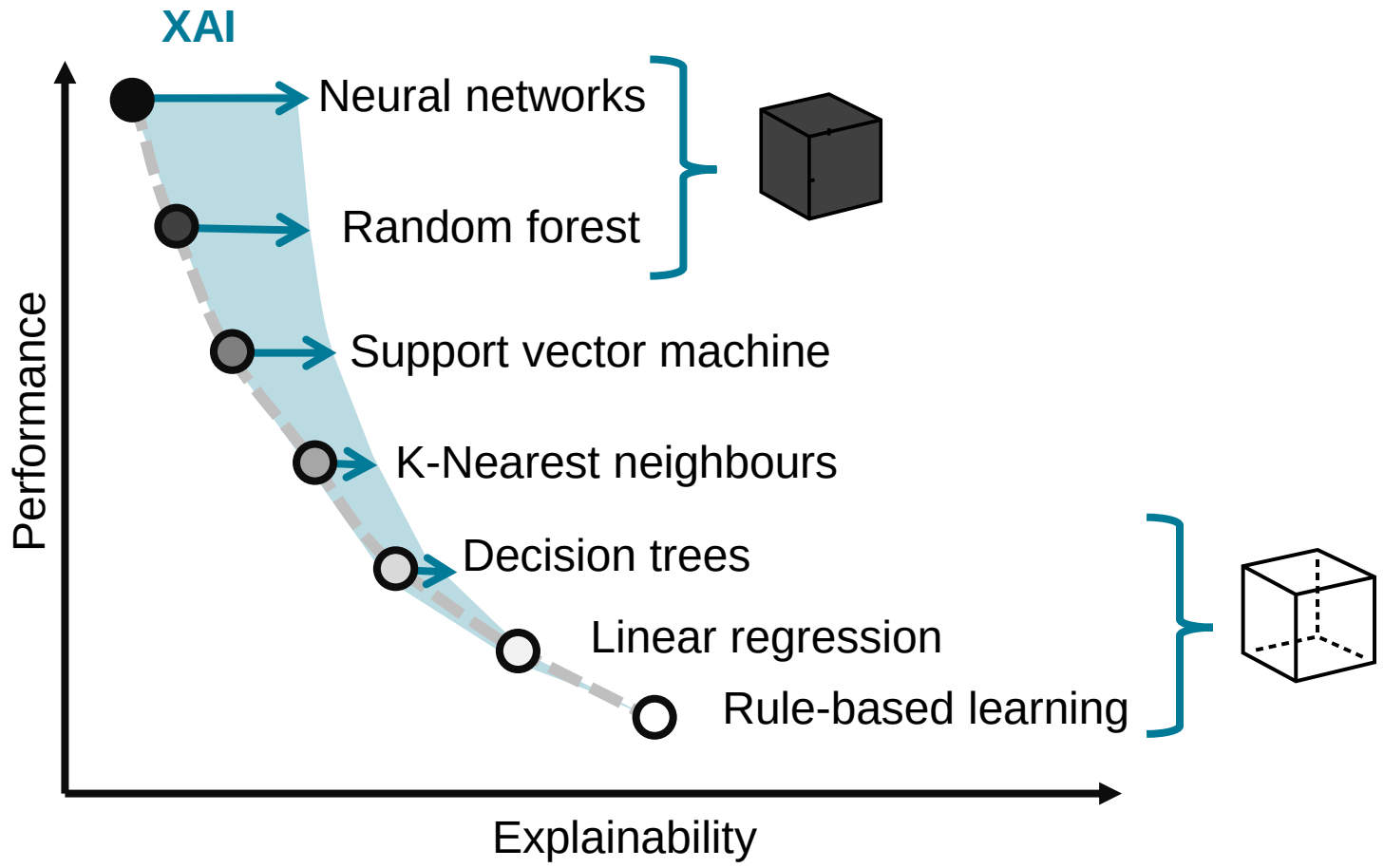


Figure inspired by: Morocho-Cayamcela *et al.* (2019) *IEEE Access* **7**, 137184.

TU/e

Clever Hans (der kluge Hans)

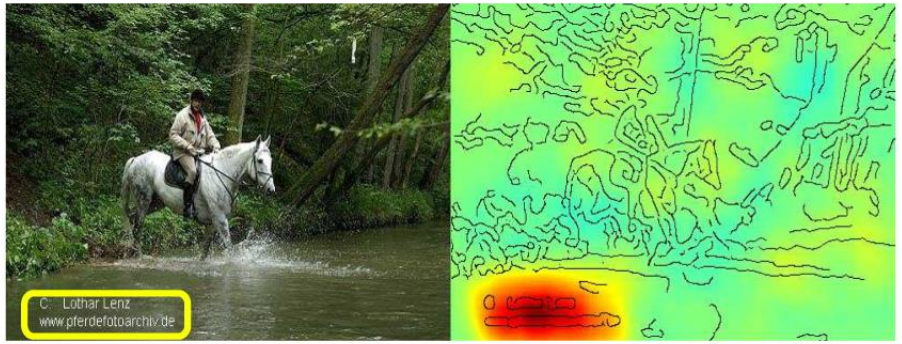
- Berlin, 1900.
- Horse claimed to perform arithmetic (hoof tapping).
- The horse was responding directly to involuntary cues in the body language of the human trainer.

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)

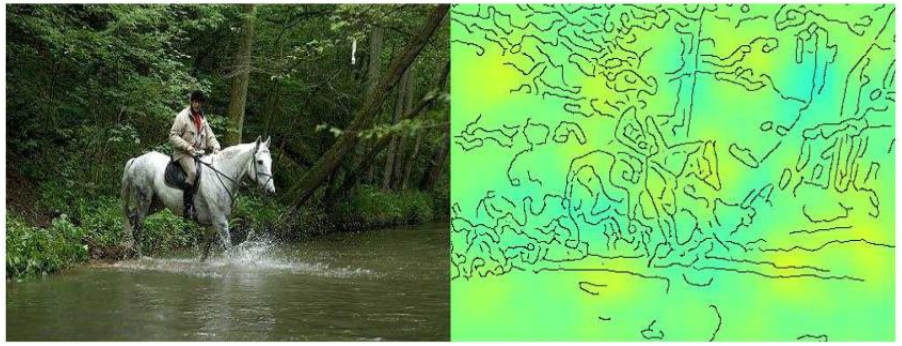
Horse-picture from Pascal VOC data set



Source tag present

↓

Classified as horse

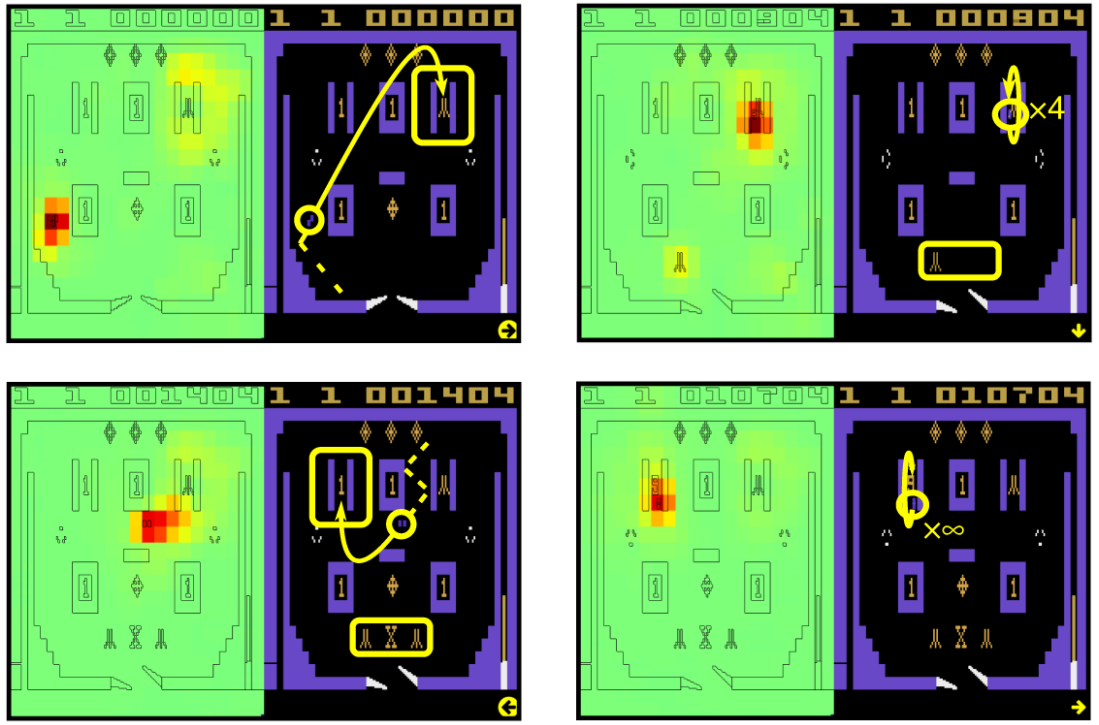


No source tag present

↓

Not classified as horse

Pinball - relevance during game play

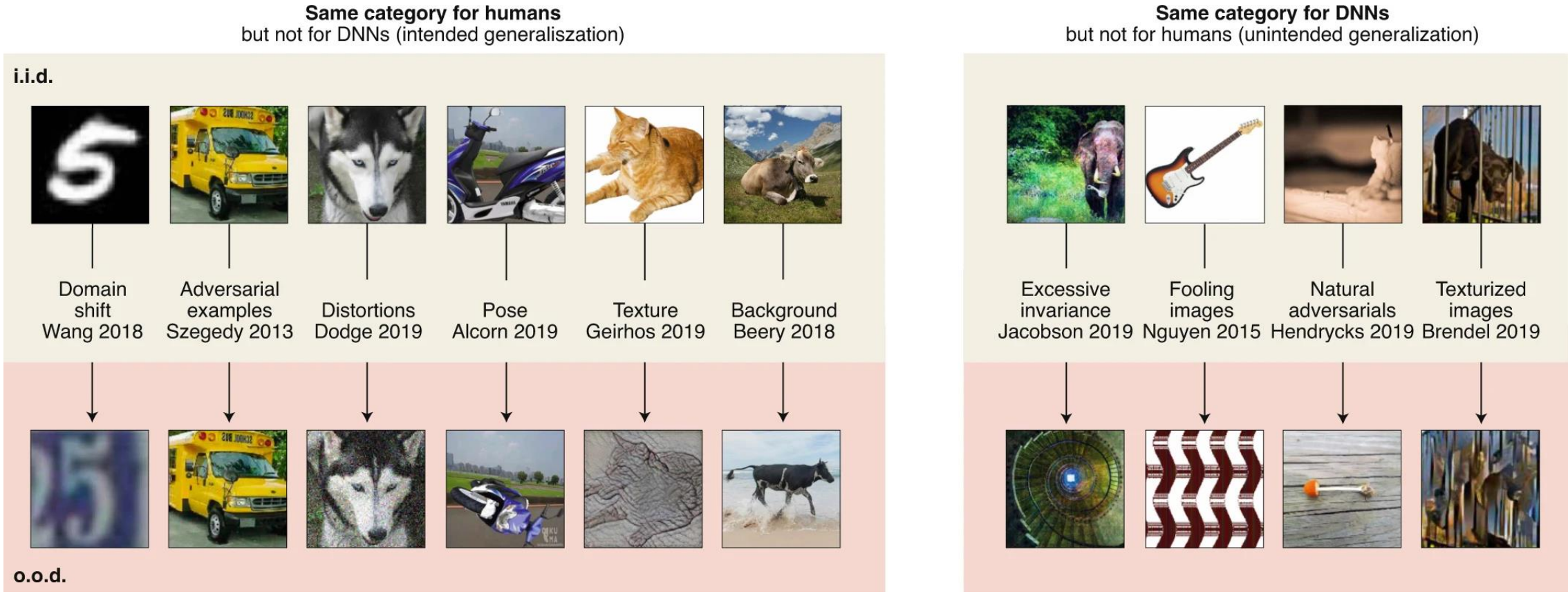


Identify valid vs invalid problem-solving behaviours

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

TU/e

Surface learning and unintended generalization



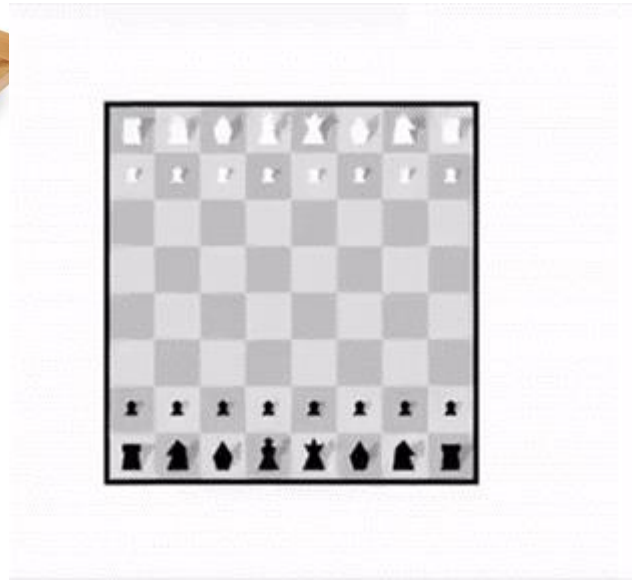
Performance metrics might just not be enough!

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

TU/e

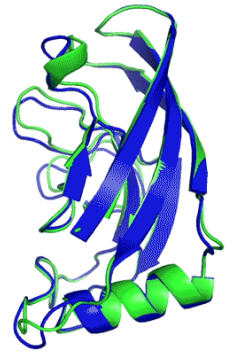
Imagine if we could learn from AI

Mastering complex games[1,2]



Model natural phenomena[3,4]

T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)



T1049 / 6y4f
93.3 GDT
(adhesin tip)

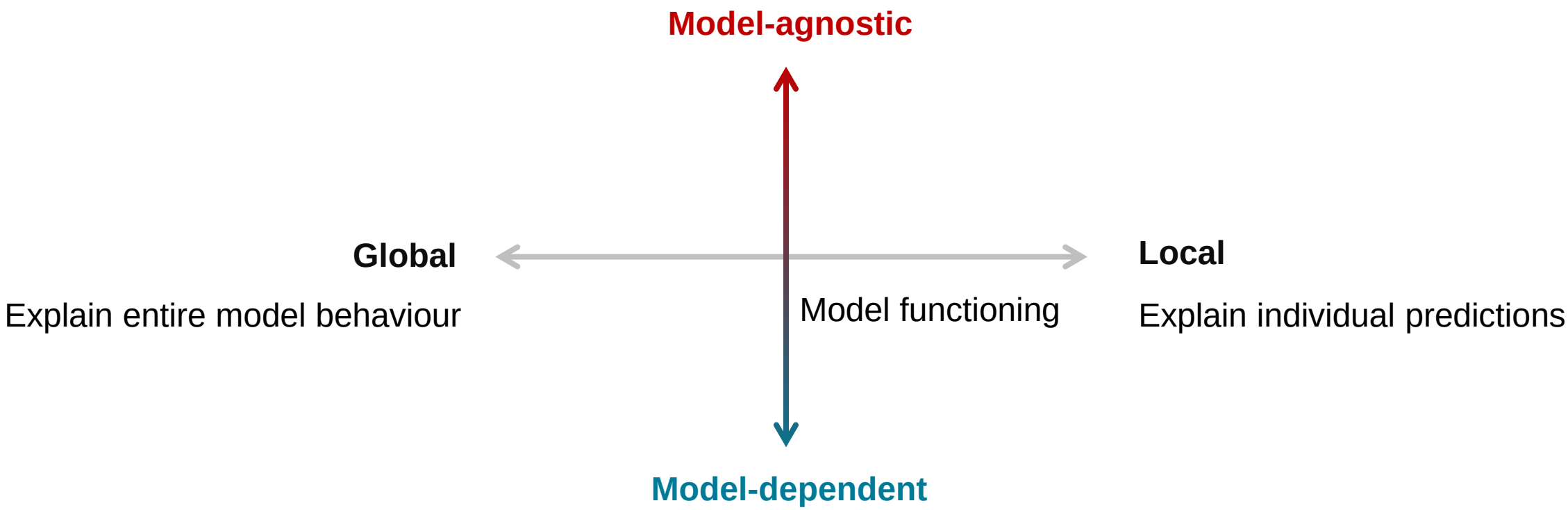
Experimental result
Computational prediction

[1]Silver *et al.* (2017) *Nature* **550**, 254; [2]Schrittwieser *et al.* (2020) *Nature* **588**, 604.
[3]Ravuri *et al.* (2021) *Nature* **597**, 672; [4]Jumper *et al.* (2021) *Nature* **596**, 583.

TU/e

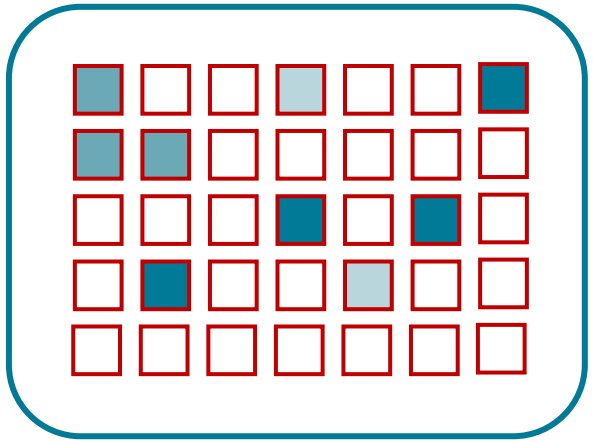
A roadmap of XAI: pt. 1

A roadmap of XAI: pt. 1

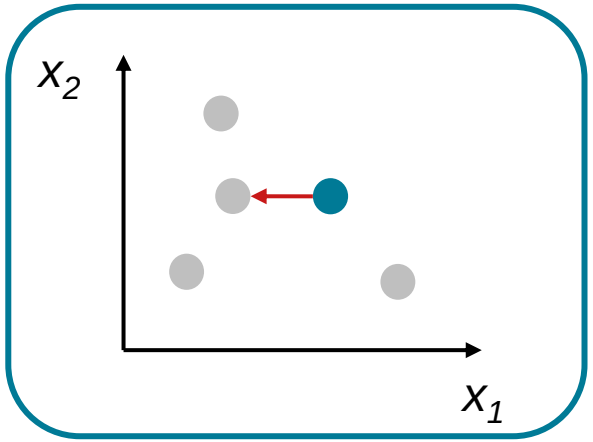


A roadmap of XAI: pt. 2

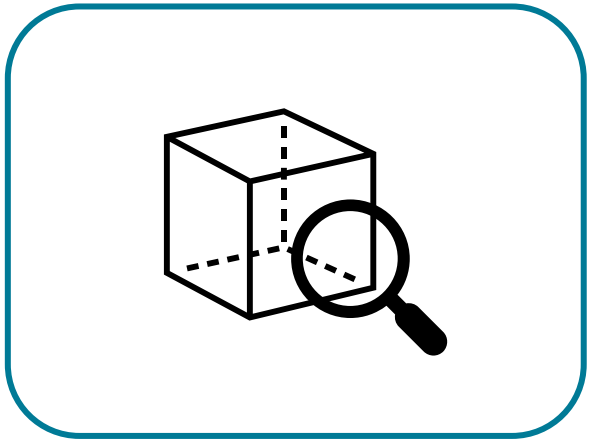
Feature attribution



Instance-based

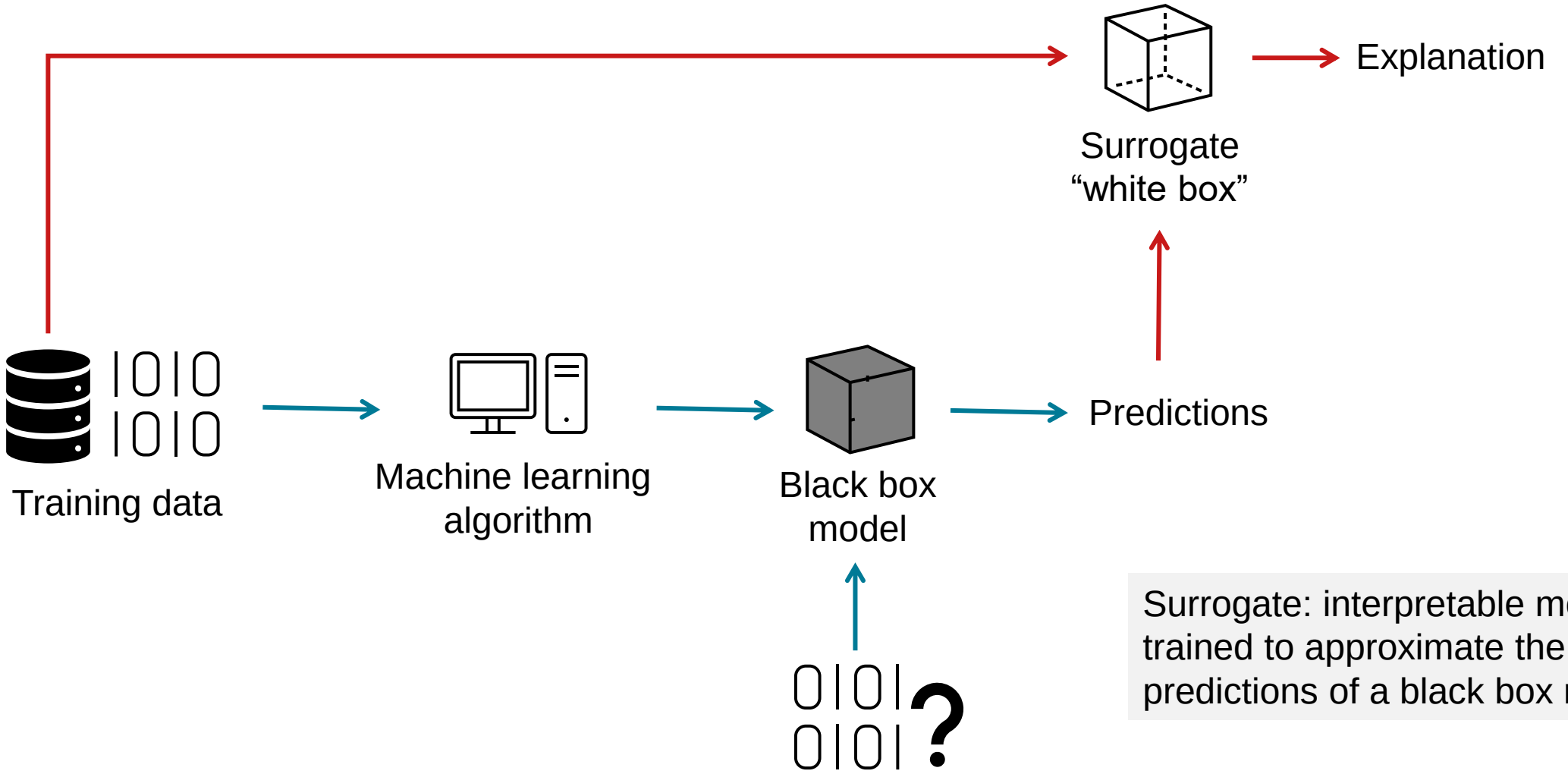


Self-explaining



TU/e

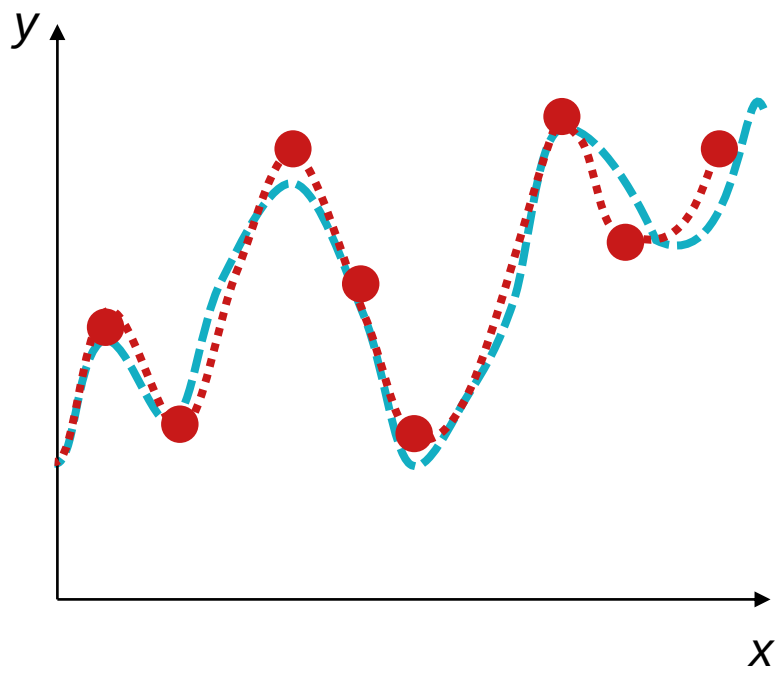
Surrogate models



Surrogate: interpretable model trained to approximate the predictions of a black box model.

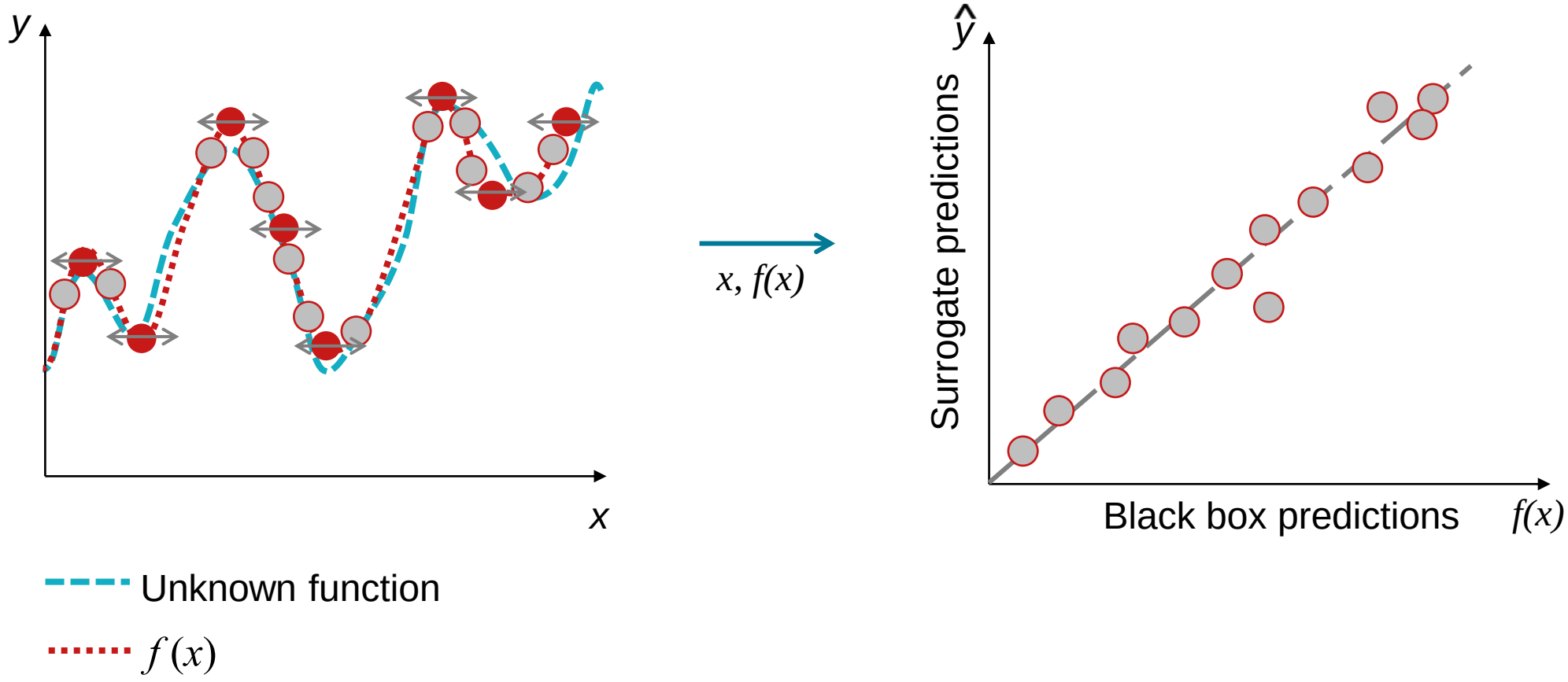
TU/e

Global surrogates



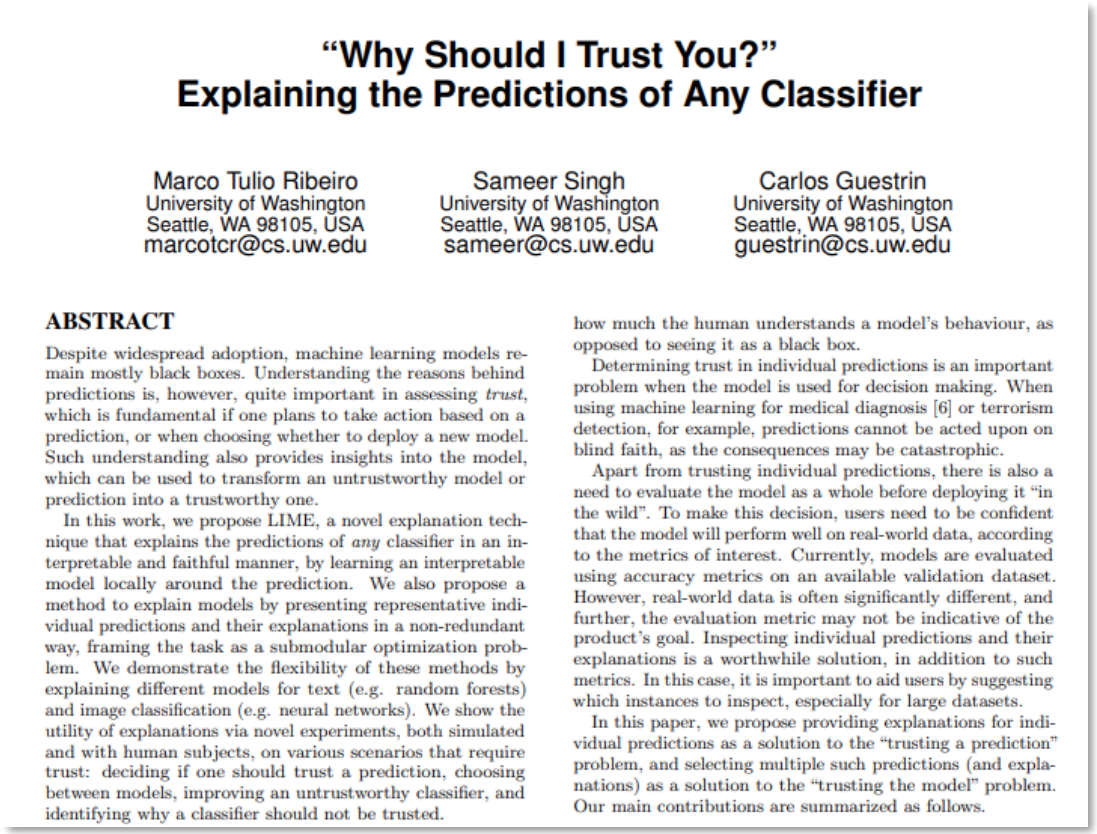
TU/e

Global surrogates



TU/e

Local interpretable model-agnostic explanations (LIME)[1]



"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.

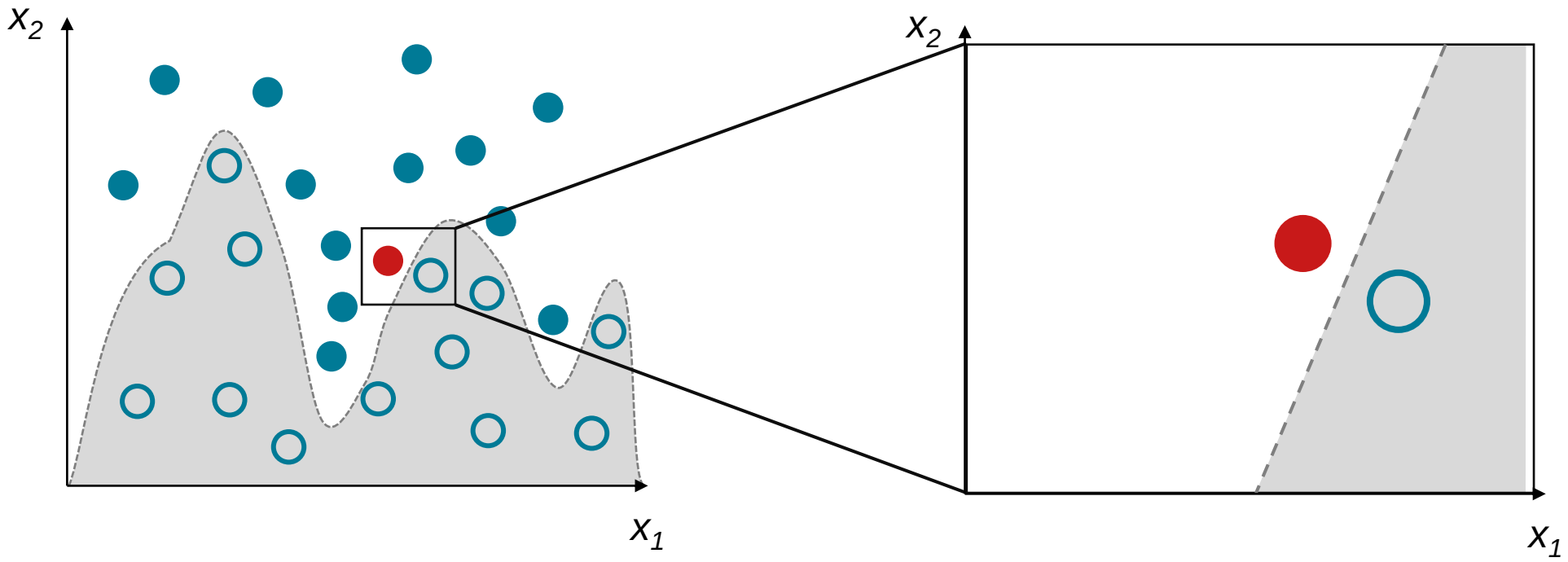
In this paper, we propose providing explanations for individual predictions as a solution to the "trusting a prediction" problem, and selecting multiple such predictions (and explanations) as a solution to the "trusting the model" problem. Our main contributions are summarized as follows.

- Works on any "black-box" model
- Does not address the model internal functioning
- Aimed to create trust using domain expertise on the functioning of the process
- "Locally-faithful" explanations

[1]Ribeiro e*t al.* (2016). Proc. ACM SIGKDD Int, 1135.

TU/e

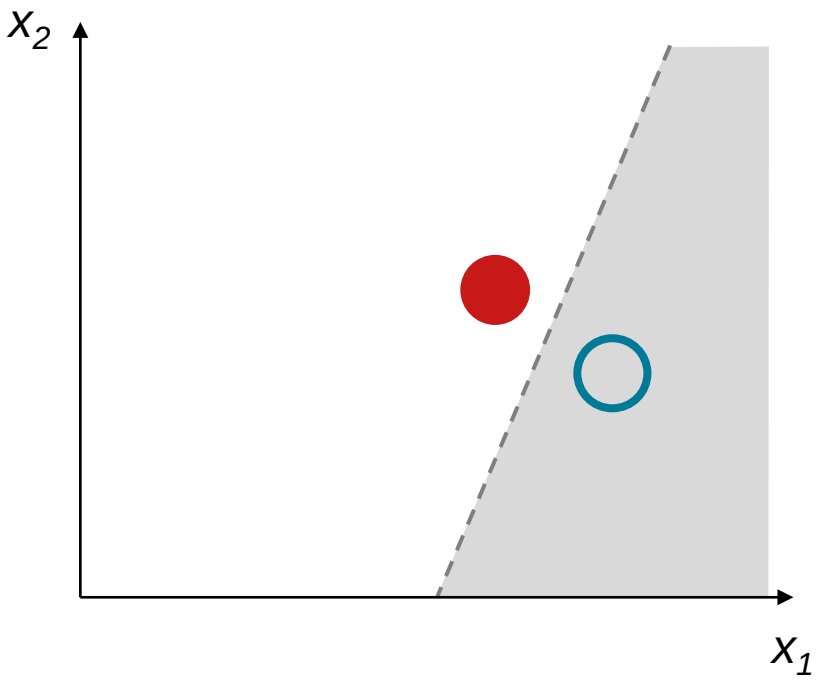
Local interpretable model-agnostic explanations (LIME)[1]



[1]Ribeiro e*t al.* (2016). Proc. ACM SIGKDD Int, 1135.

TU/e

Local interpretable model-agnostic explanations (LIME)[1]



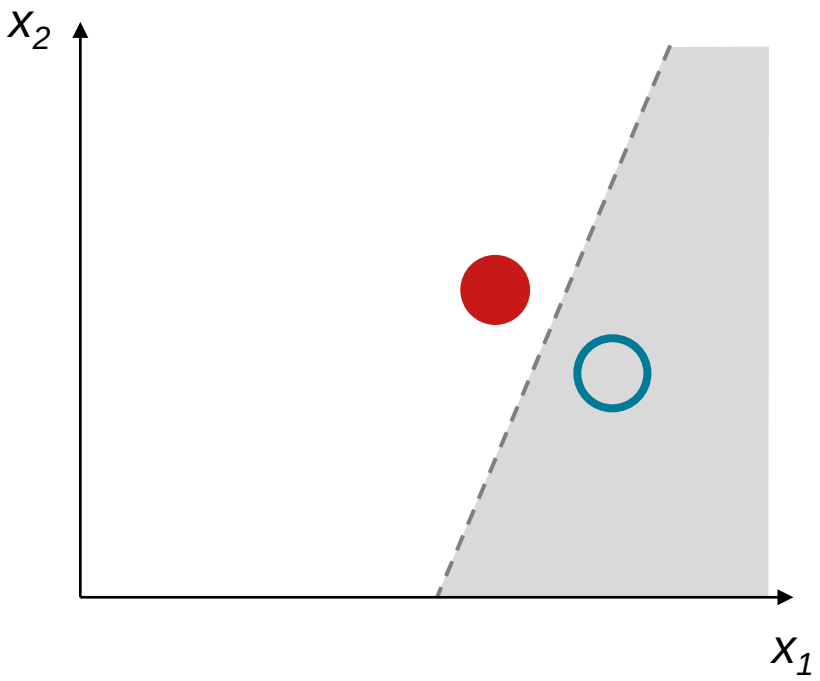
Approximation Complexity

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

[1]Ribeiro e*t al.* (2016). Proc. ACM SIGKDD Int, 1135.

TU/e

Local interpretable model-agnostic explanations (LIME)[1]



Approximation Complexity

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

Complex model

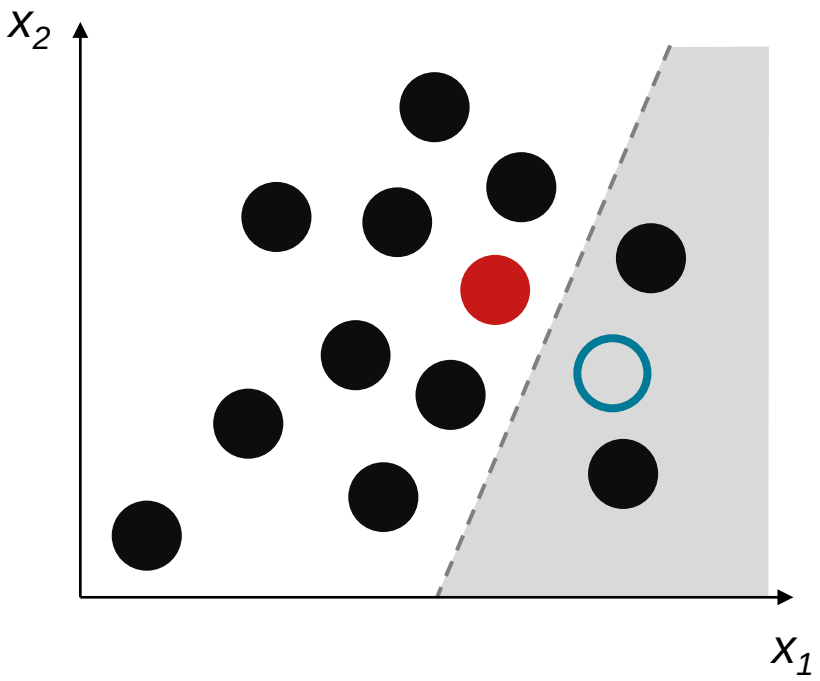
Simple model

Proximity

[1]Ribeiro e*t al.* (2016). Proc. ACM SIGKDD Int, 1135.

TU/e

Local interpretable model-agnostic explanations (LIME)[1]



Approximation Complexity

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

Complex model

Proximity

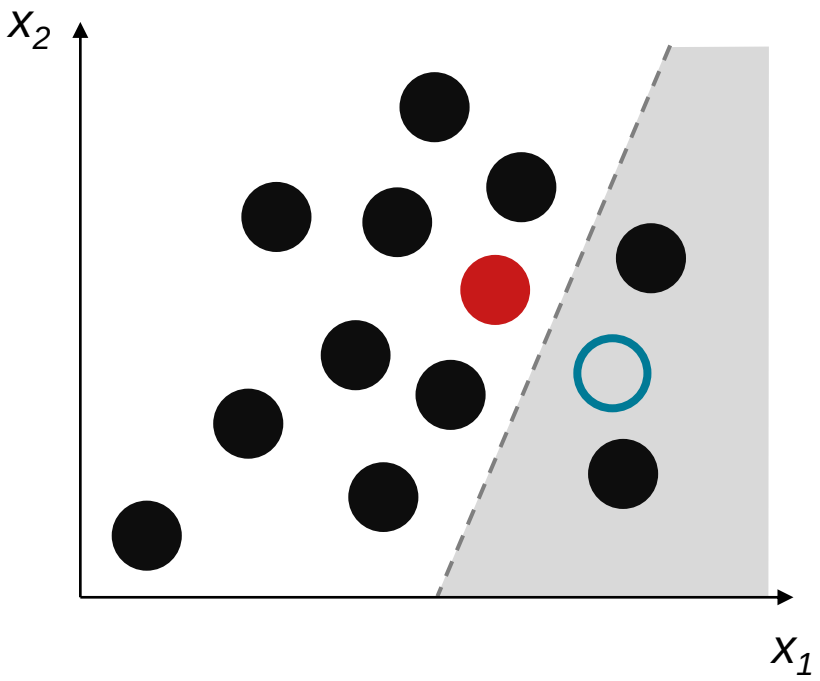
Simple model

New dataset!

[1]Ribeiro e*t al.* (2016). Proc. ACM SIGKDD Int, 1135.

TU/e

Local interpretable model-agnostic explanations (LIME)[1]



Approximation Complexity

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

Complex model

Simple model

Proximity

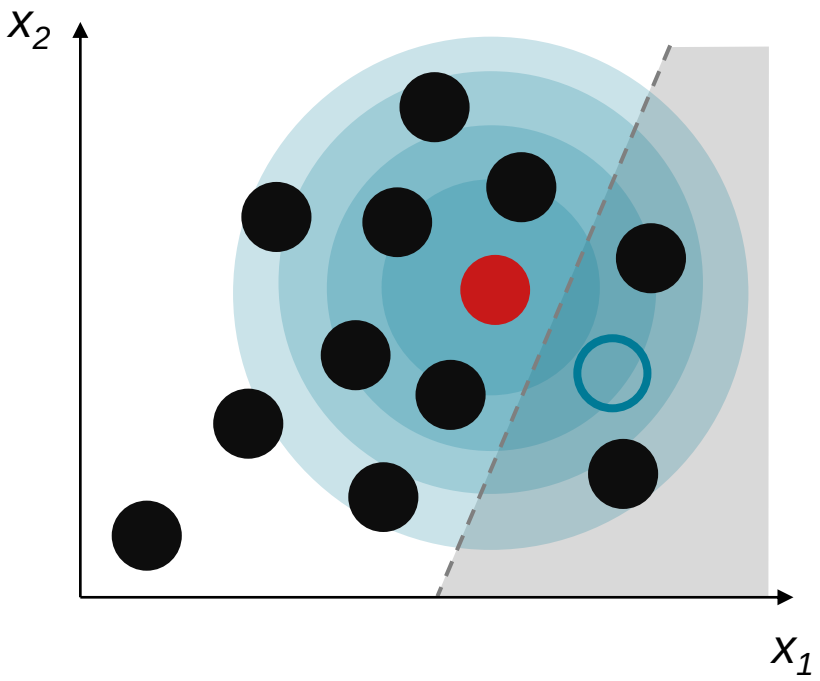
Sum of Squared distances

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

[1]Ribeiro e*t al.* (2016). Proc. ACM SIGKDD Int, 1135.

TU/e

Local interpretable model-agnostic explanations (LIME)[1]



Approximation Complexity

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

Complex model

Simple model

Proximity

Sum of Squared distances

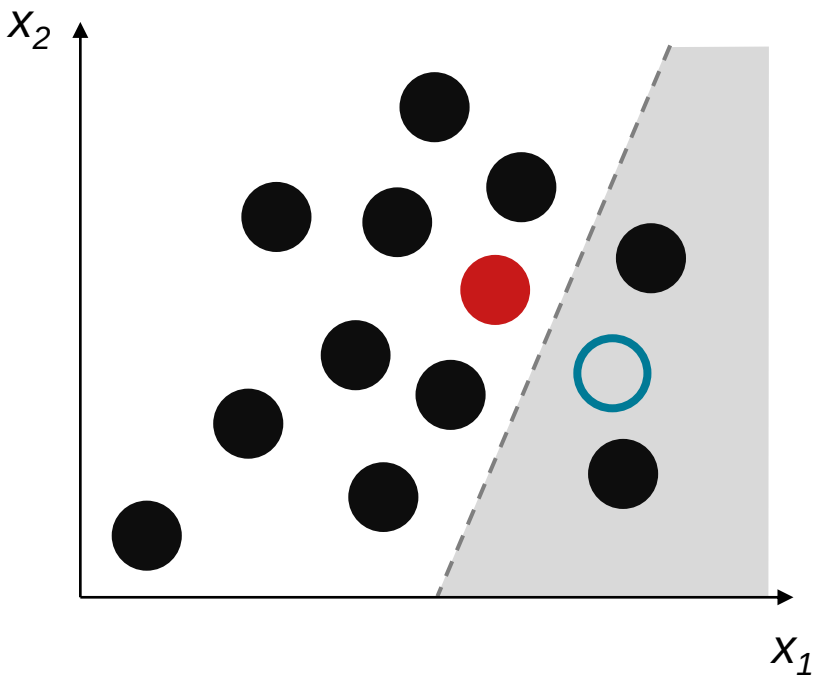
$$L(f, g, \pi_x) = \sum_z \pi_x(z)(f(z) - g(z))^2$$

Exponential kernel

[1]Ribeiro e*t al.* (2016). Proc. ACM SIGKDD Int, 1135.

TU/e

Local interpretable model-agnostic explanations (LIME)[1]



Approximation Complexity

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

Regularization
(sparse linear models)

[1]Ribeiro e*t al.* (2016). Proc. ACM SIGKDD Int, 1135.

TU/e

Local interpretable model-agnostic explanations (LIME)[1]

Input **Predictions**

Original image | Electric guitar | Acoustic guitar | Labrador

[1]Ribeiro e*t al.* (2016). Proc. ACM SIGKDD Int, 1135.

TU/e

Surrogate models

Pros

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

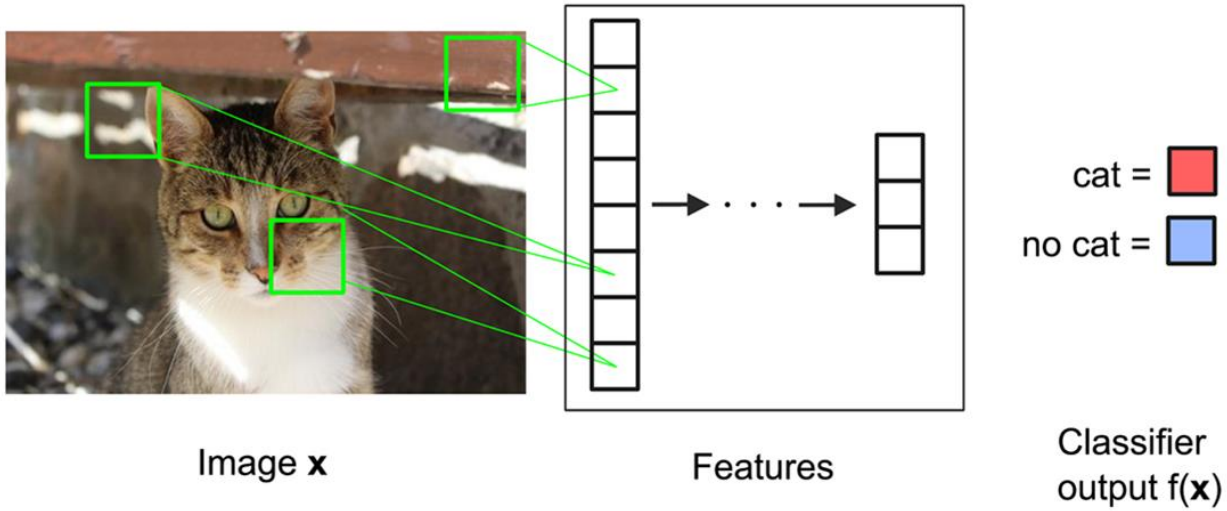
Cons

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

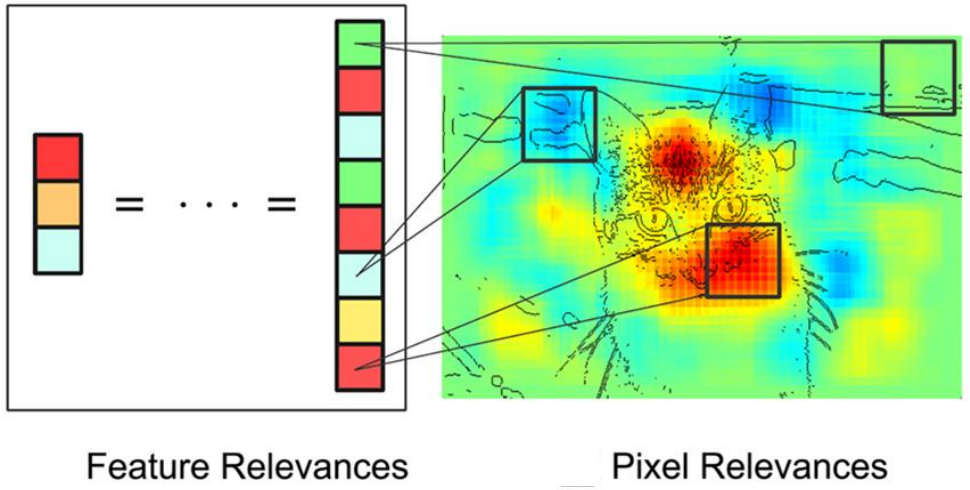
TU/e

Pixel attribution methods

Classification



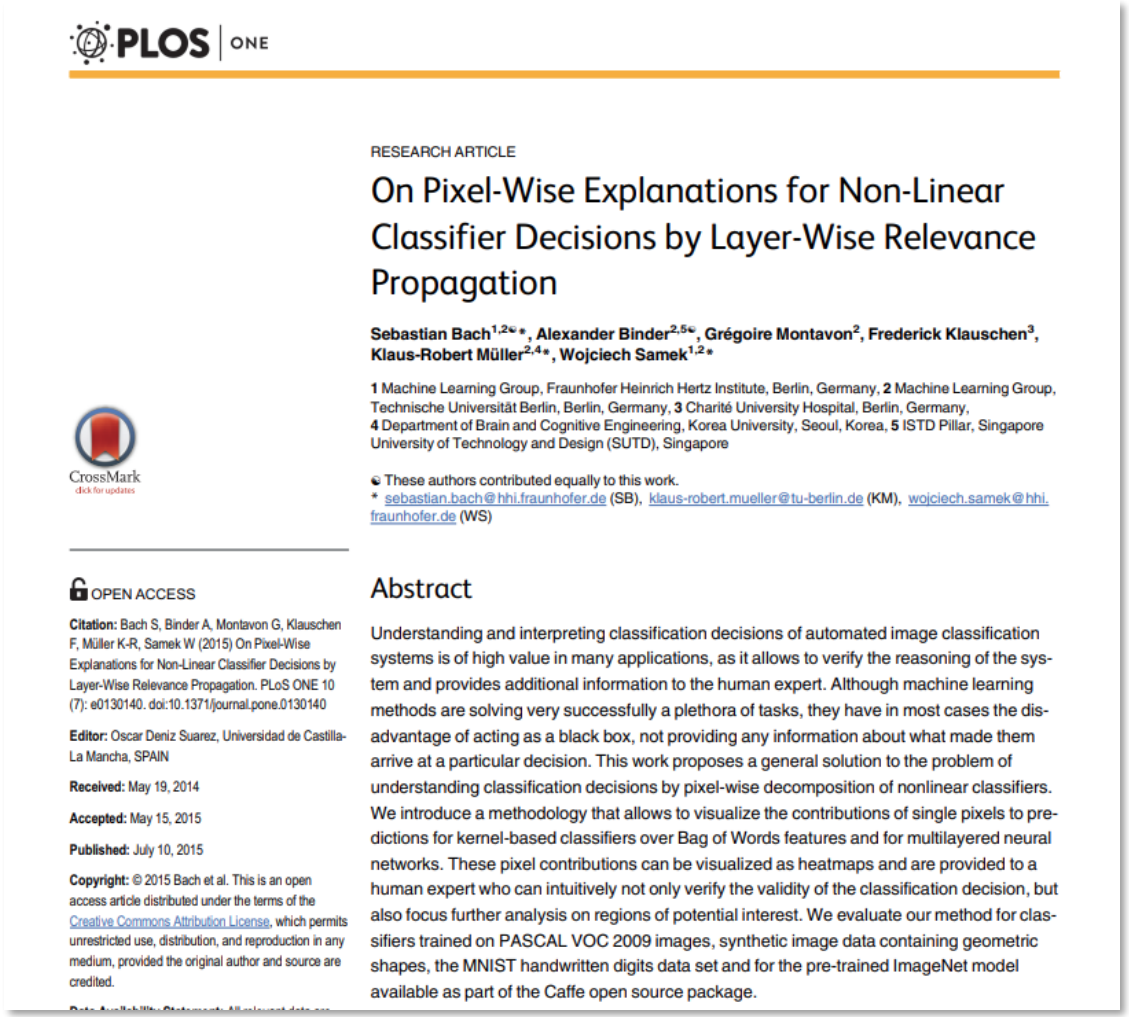
Pixel relevance



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

TU/e

Layer-wise relevance propagation (LRP)



PLOS ONE

RESEARCH ARTICLE

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

Sebastian Bach[1,2☯*], Alexander Binder[2,5☯], Grégoire Montavon[2], Frederick Klauschen[3], Klaus-Robert Müller[2,4*], Wojciech Samek[1,2*]

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 (SB), klaus-robert.mueller@tu-berlin.de (KM), 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 Suarez, Universidad de Castilla-La Mancha, SPAIN

Received: May 19, 2014

Accepted: May 15, 2015

Published: July 10, 2015

Abstract

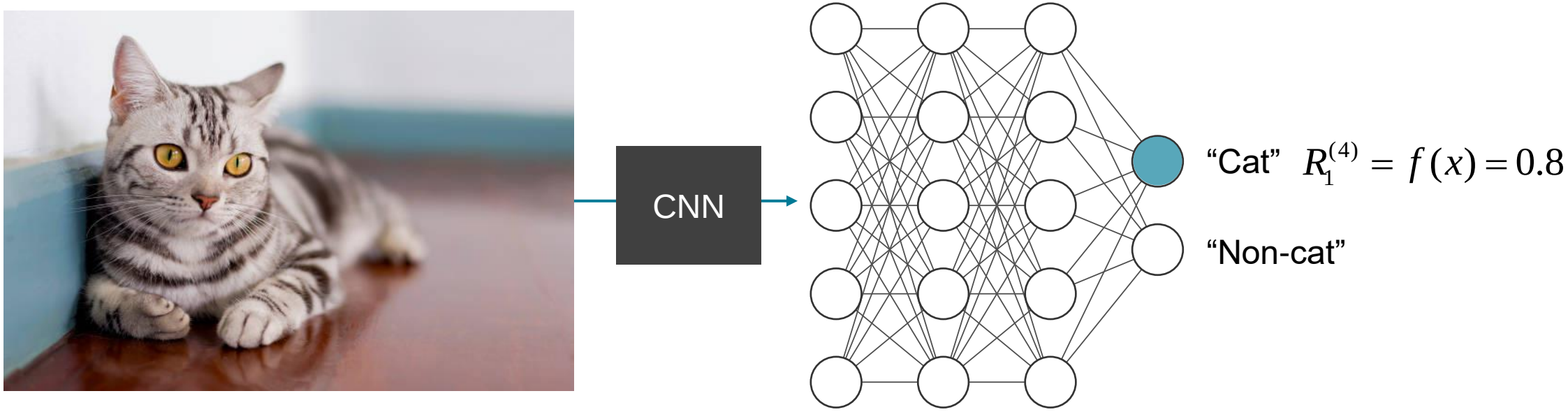
Understanding and interpreting classification decisions of automated image classification systems is of high value in many applications, as it allows to verify the reasoning of the system and provides additional information to the human expert. Although machine learning methods are solving very successfully 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.

TU/e

Layer-wise relevance propagation (LRP)

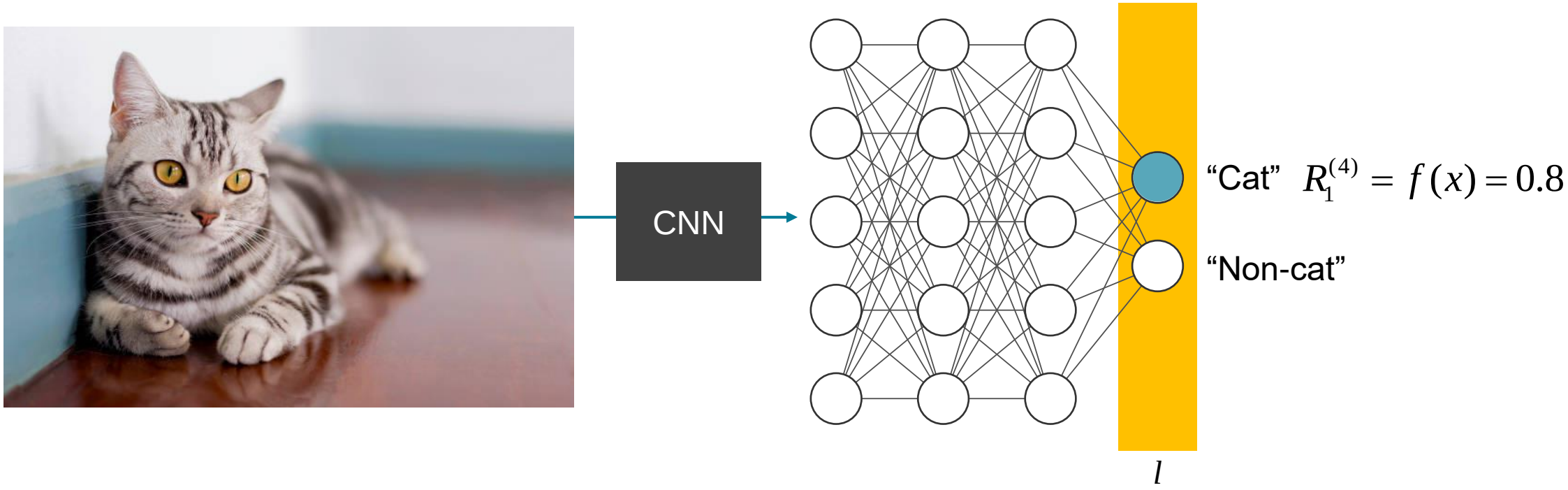


$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_k z_{kj}} R_j^{(l+1)} \quad \text{with } z_{ik} = x_i^{(l)} w_{ij}^{(l,l+1)}$$

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

TU/e

Layer-wise relevance propagation (LRP)



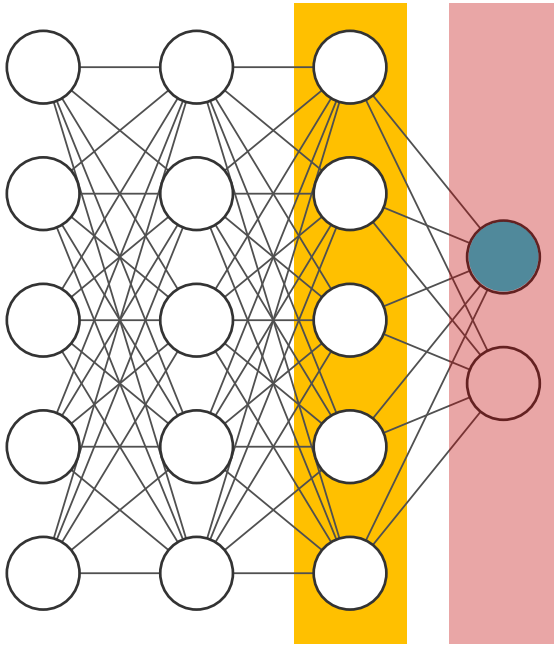
$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_k z_{kj}} R_j^{(l+1)} \quad \text{with } z_{ik} = x_i^{(l)} w_{ij}^{(l,l+1)}$$

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

TU/e

Layer-wise relevance propagation (LRP)

CNN



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

"Non-cat"

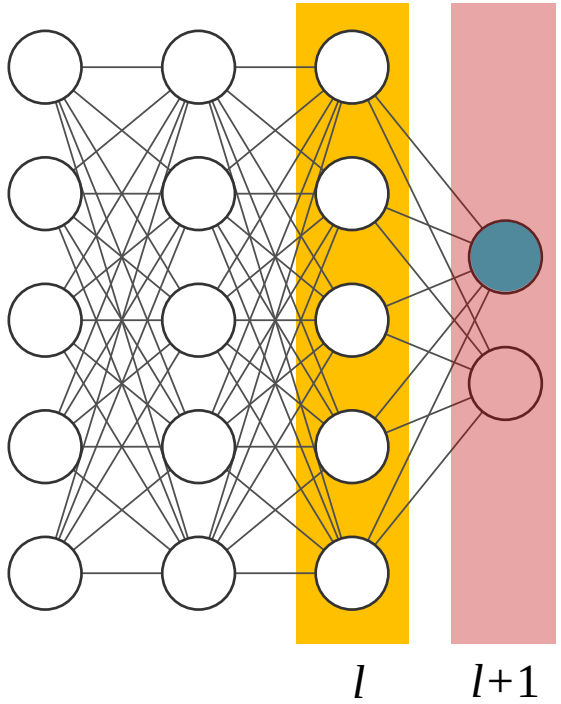
$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_k z_{kj}} R_j^{(l+1)} \quad \text{with } z_{ik} = x_i^{(l)} w_{ij}^{(l,l+1)}$$

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

TU/e

Layer-wise relevance propagation (LRP)

CNN



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

"Non-cat"

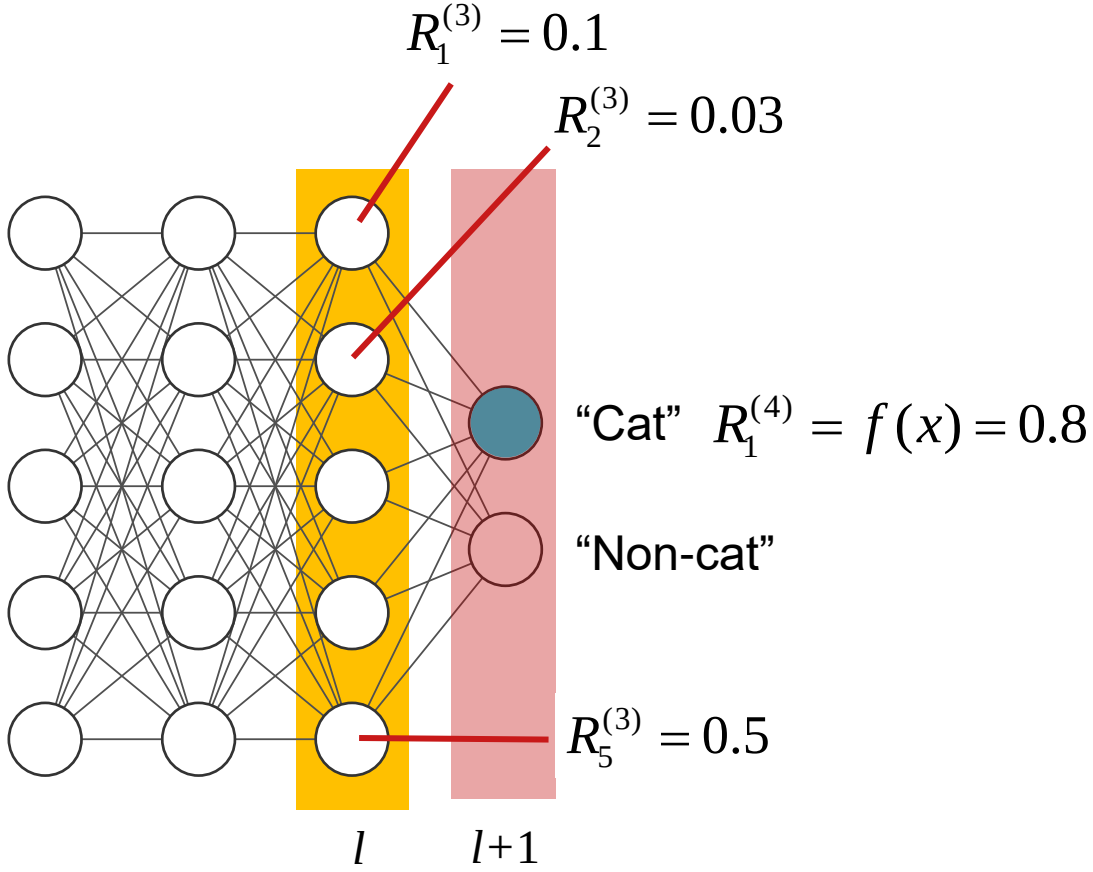
$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_k z_{kj}} R_j^{(l+1)} \quad \text{with} \quad z_{ik} = x_i^{(l)} w_{ij}^{(l,l+1)}$$

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

TU/e

Layer-wise relevance propagation (LRP)

CNN



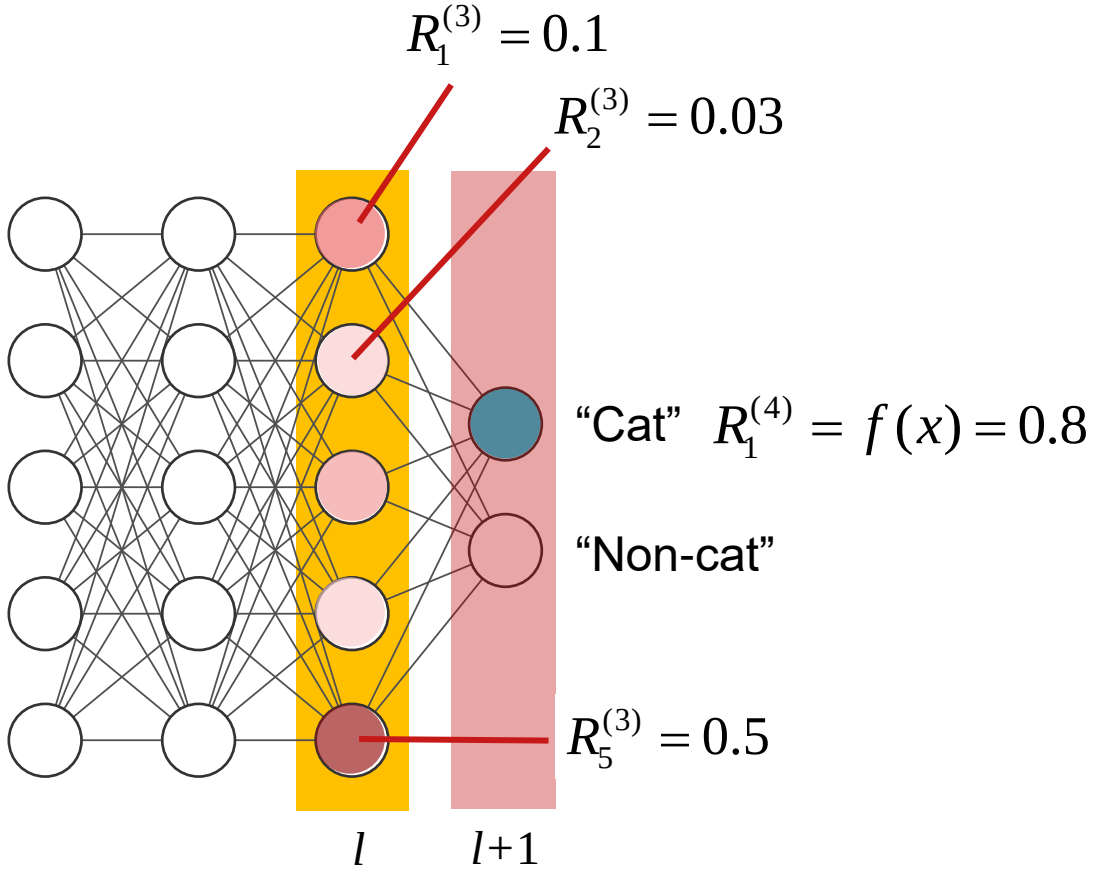
$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_k z_{kj}} R_j^{(l+1)} \quad \text{with } z_{ik} = x_i^{(l)} w_{ij}^{(l,l+1)}$$

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

TU/e

Layer-wise relevance propagation (LRP)

CNN



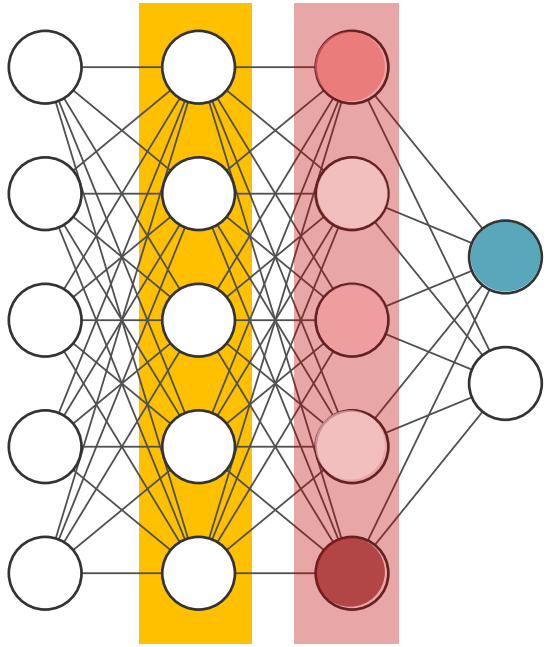
$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_k z_{kj}} R_j^{(l+1)} \quad \text{with } z_{ik} = x_i^{(l)} w_{ij}^{(l,l+1)}$$

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

TU/e

Layer-wise relevance propagation (LRP)

CNN



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

"Non-cat"

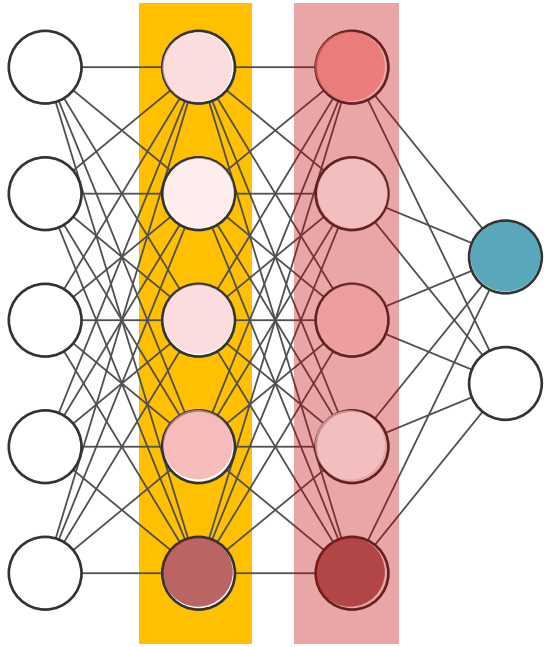
$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_k z_{kj}} R_j^{(l+1)} \quad \text{with} \quad z_{ik} = x_i^{(l)} w_{ij}^{(l,l+1)}$$

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

TU/e

Layer-wise relevance propagation (LRP)

CNN



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

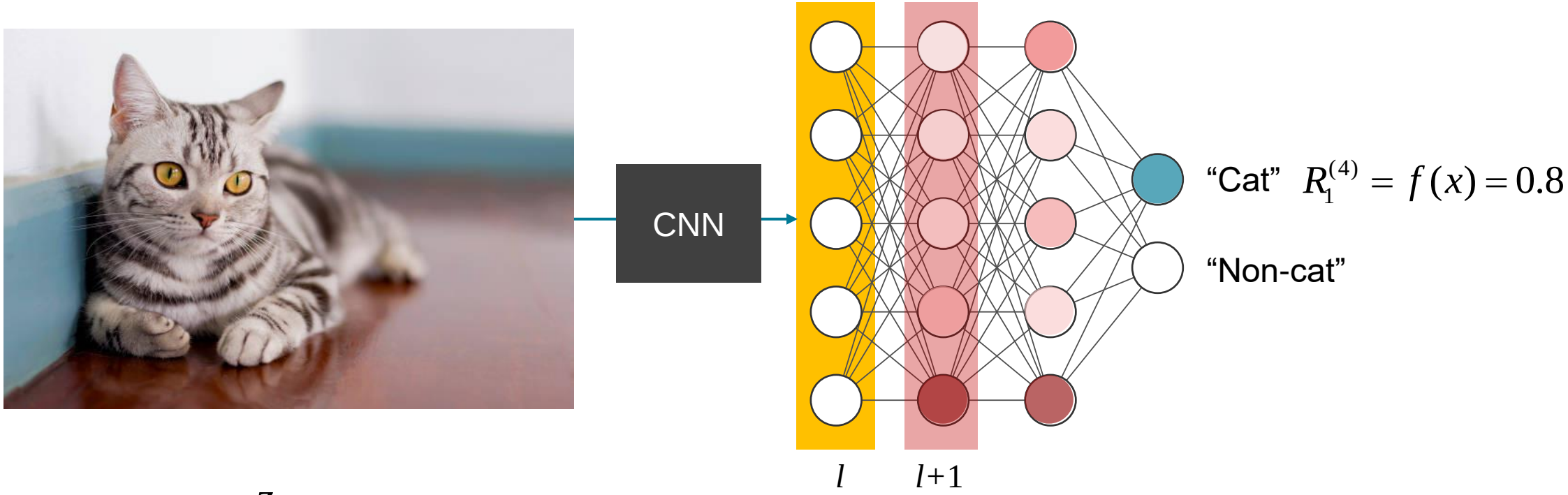
"Non-cat"

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

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

TU/e

Layer-wise relevance propagation (LRP)

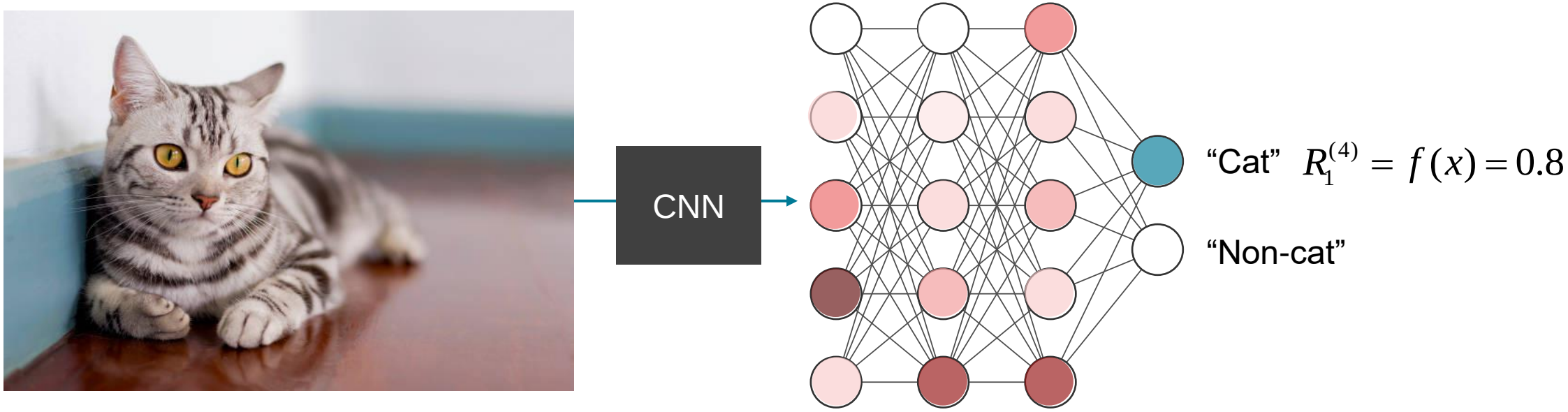


$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_k z_{kj}} R_j^{(l+1)} \quad \text{with } z_{ik} = x_i^{(l)} w_{ij}^{(l,l+1)}$$

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

TU/e

Layer-wise relevance propagation (LRP)



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

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

TU/e

Layer-wise relevance propagation (LRP)

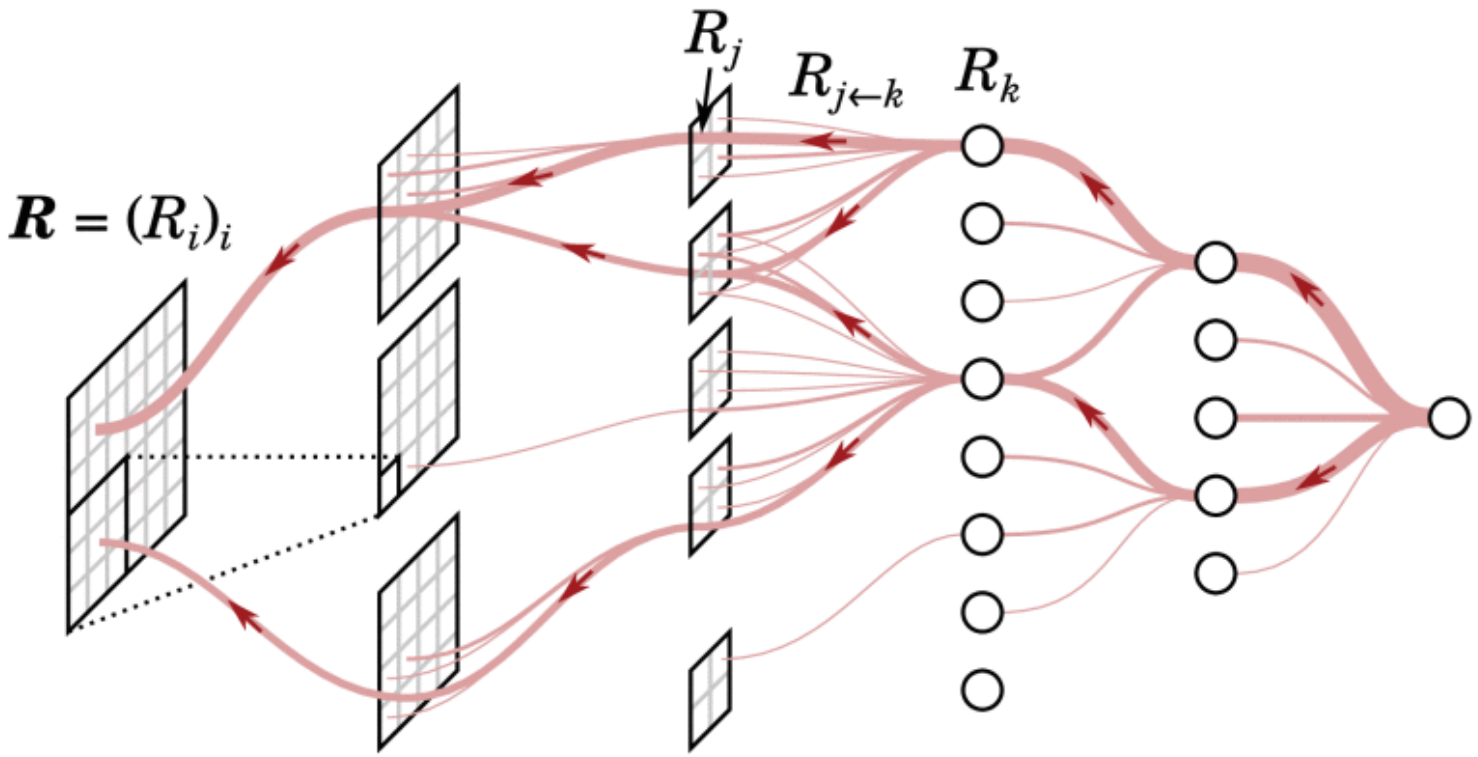
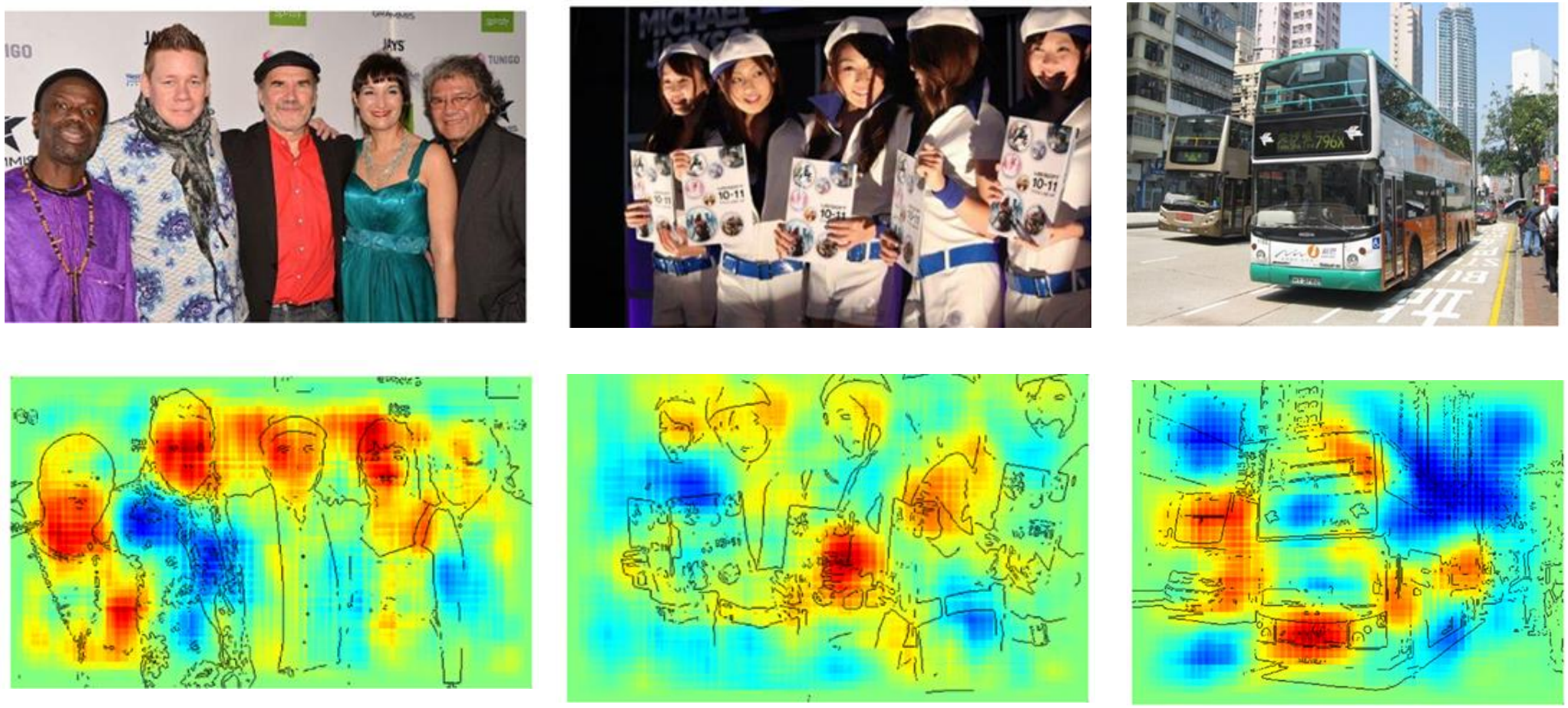


Figure from: Samek et al. (2021). *IEEE Proceedings* **109**, 247.
Bach *et al.* (2015). *PloS one* **10**, e0130140.

TU/e

Layer-wise relevance propagation (LRP)

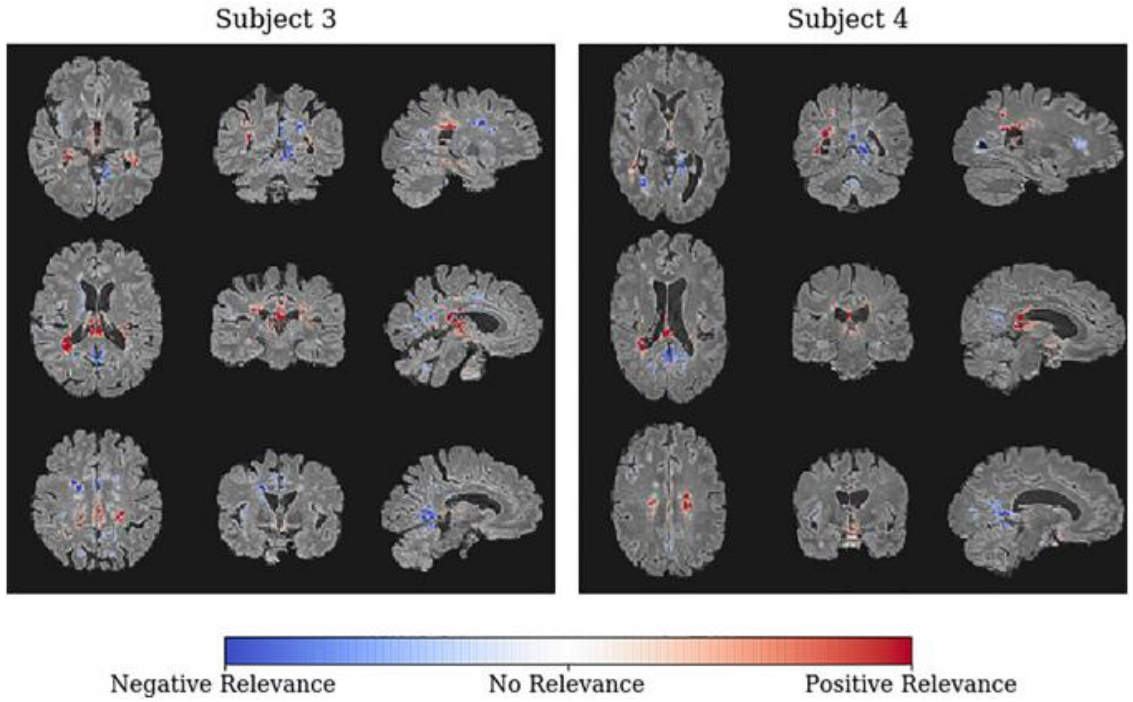


Bach *et al.* (2015). *PloS one* **10**, e0130140.

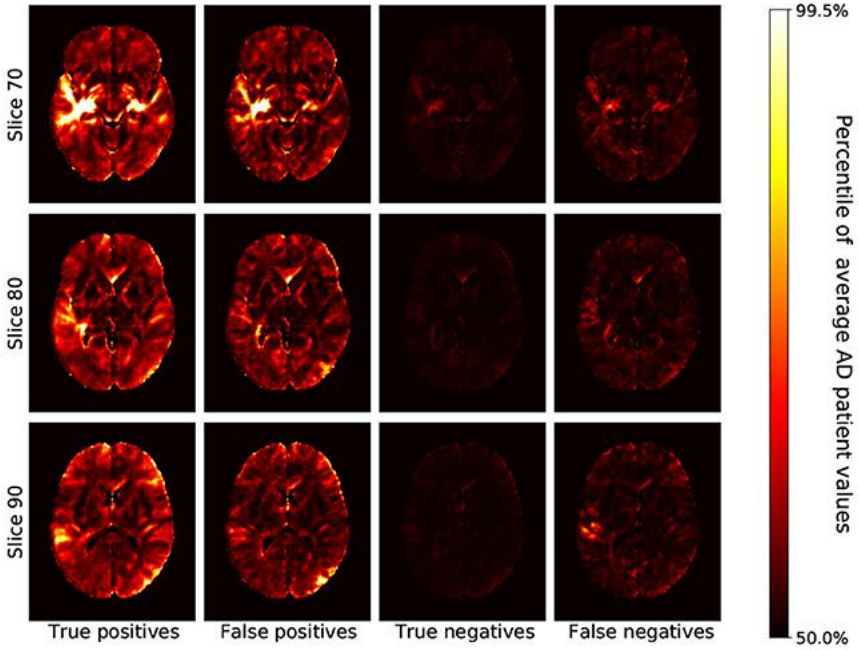
TU/e

LRP in diagnostic imaging

Diagnosis of multiple sclerosis[1]



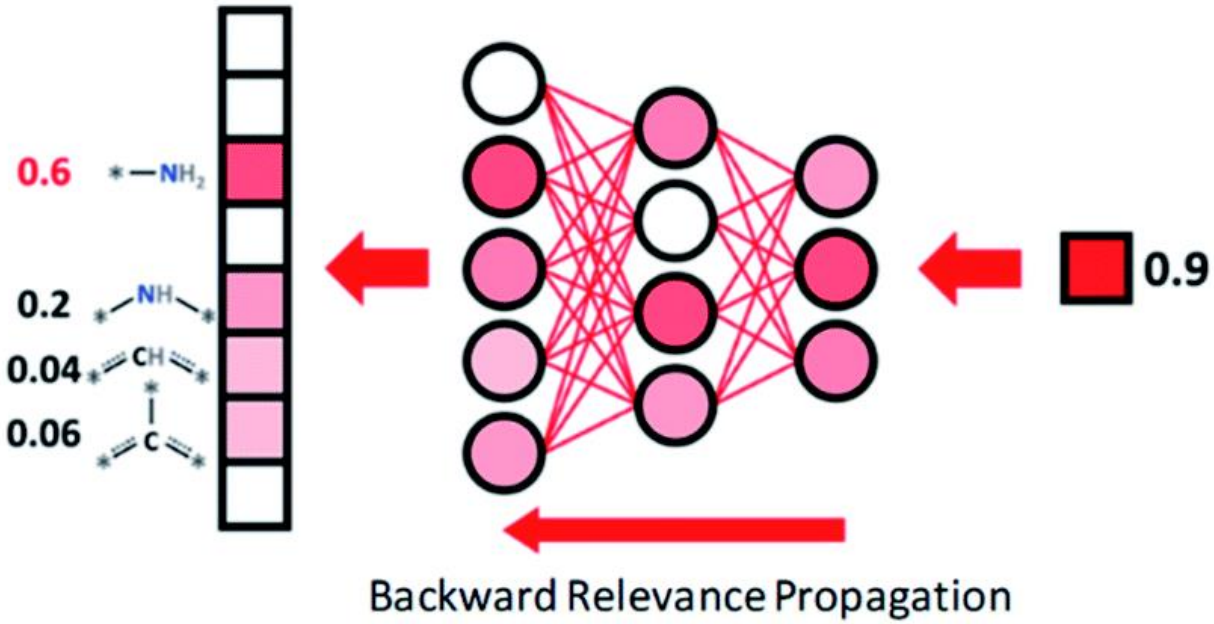
Alzheimer's disease classification[2]

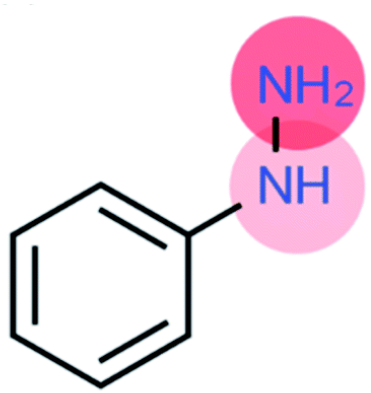


[1]Eitel *et al.* (2019). *NeuroImage: Clinical* 24, 102003.
[2]Böhle *et al.* (2019). Frontiers Aging Neurosc. 11, 194.

TU/e

Other usages of LRP

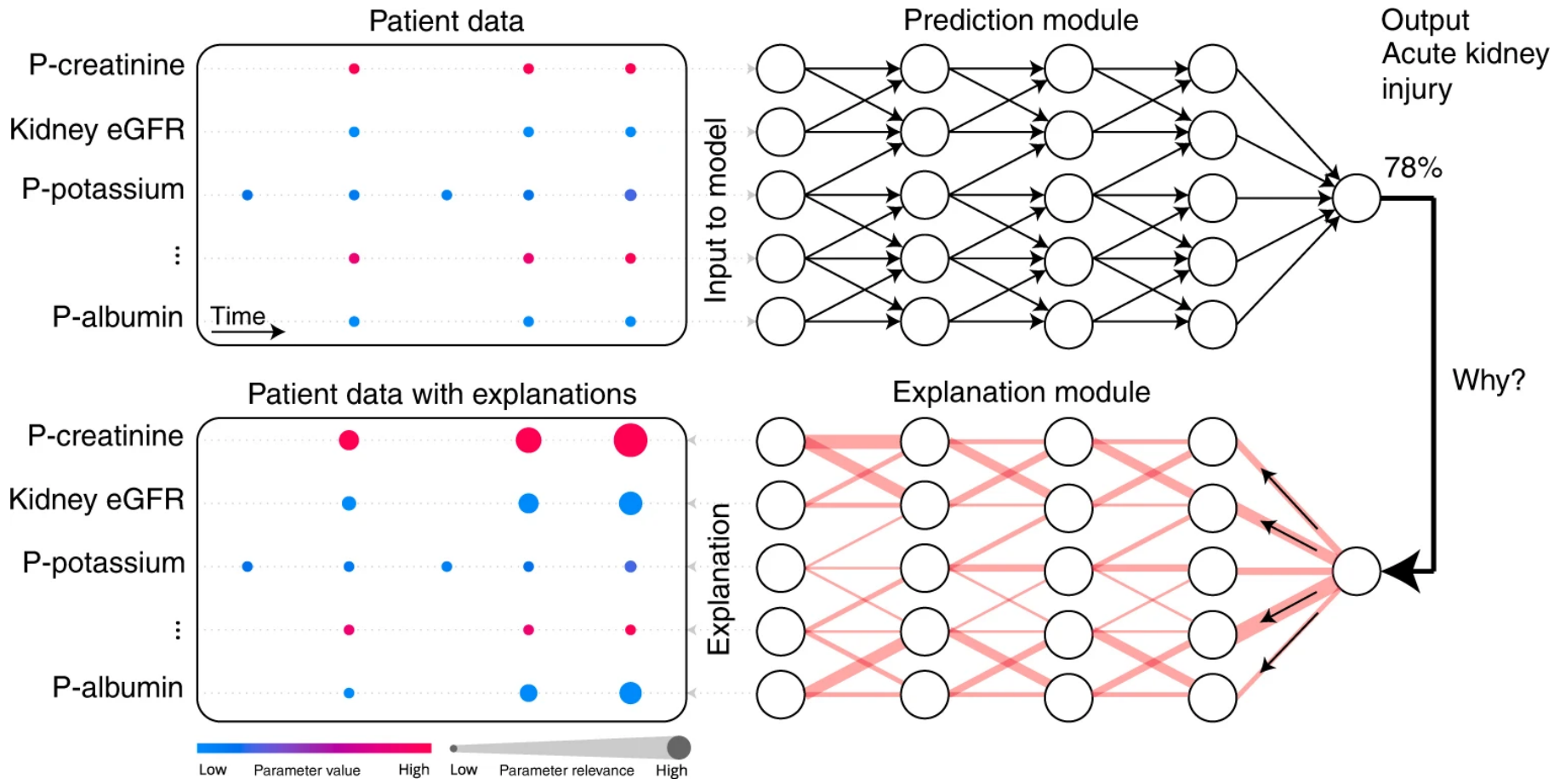




Kim *et al.* (2021) *Chem. Sci.* **12**, 11028.

TU/e

Other usages of LRP



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

Pixel (feature) attribution methods

Pros

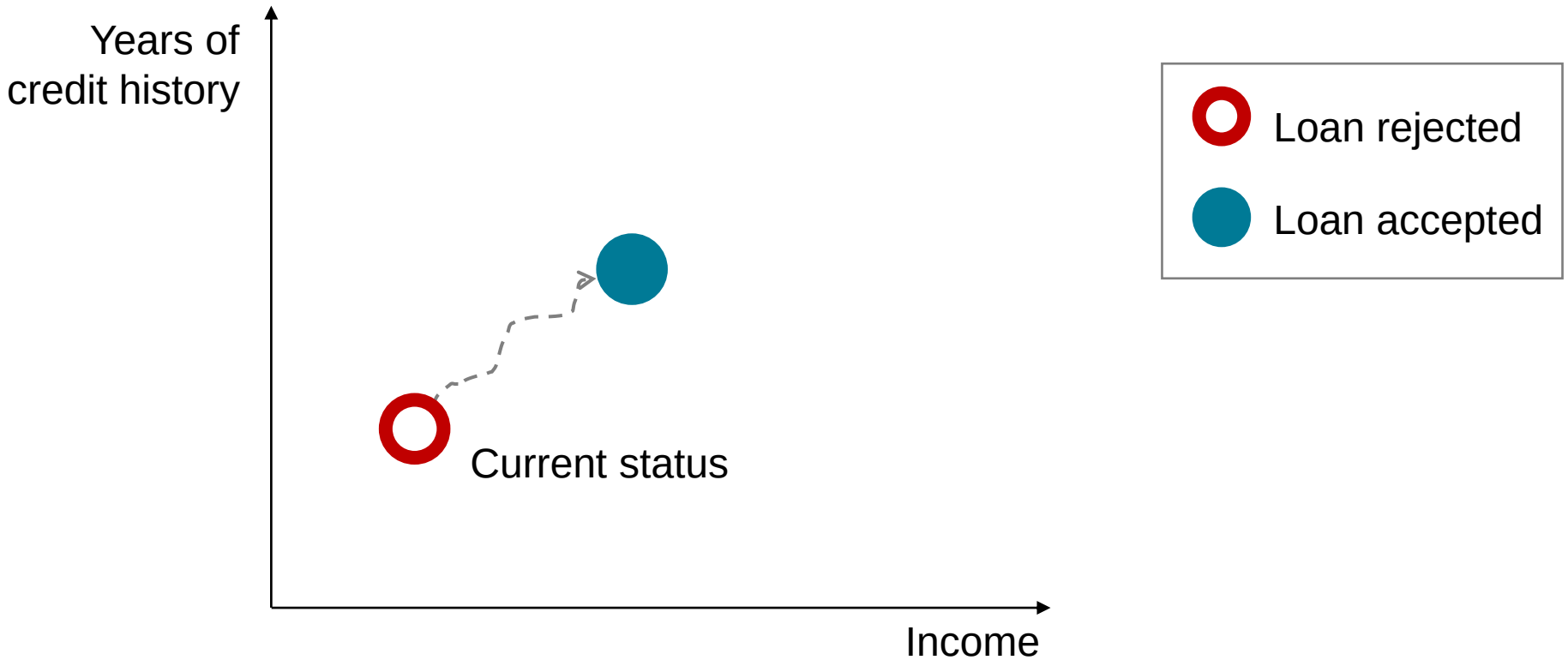
- Visual and quick to inspect
- Important regions in the image
- Allow to detect certain shortcomings

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

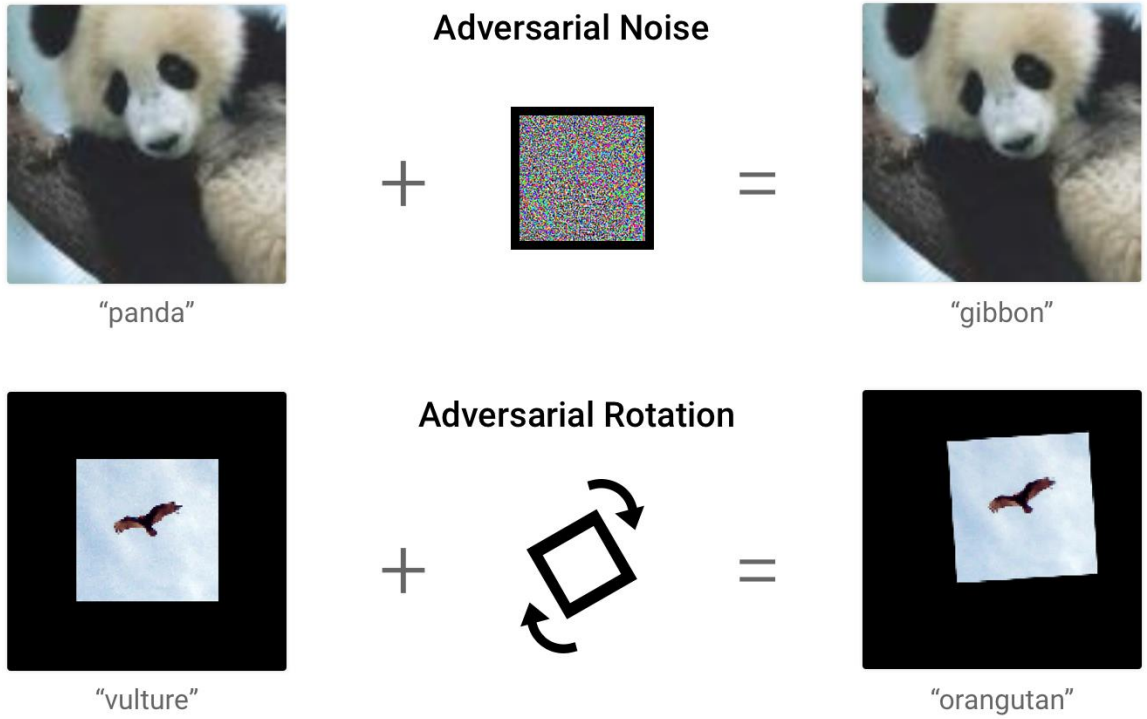
Counterfactual: smallest change in the input features that changes the predictions to another output.



Adversarial attacks

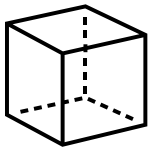
Adversarial examples: generate false predictions by leveraging the shortages of the algorithm.

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

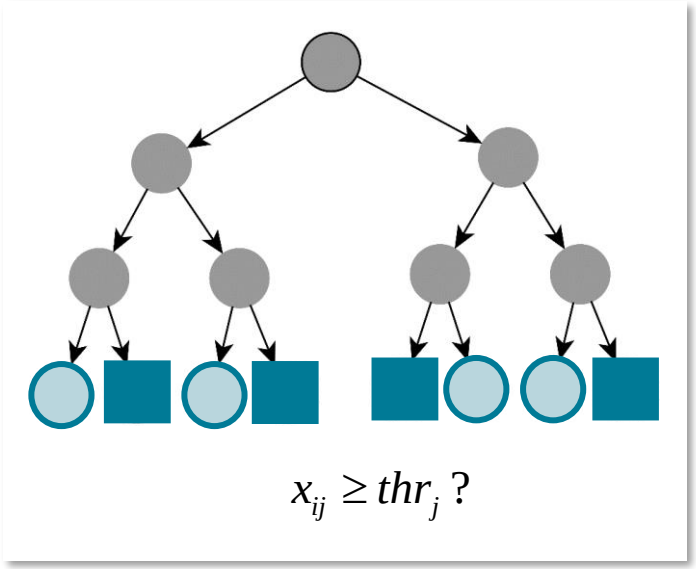


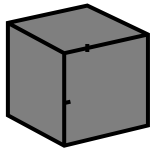
Minimal change in the input to generate a different output

Counterfactuals

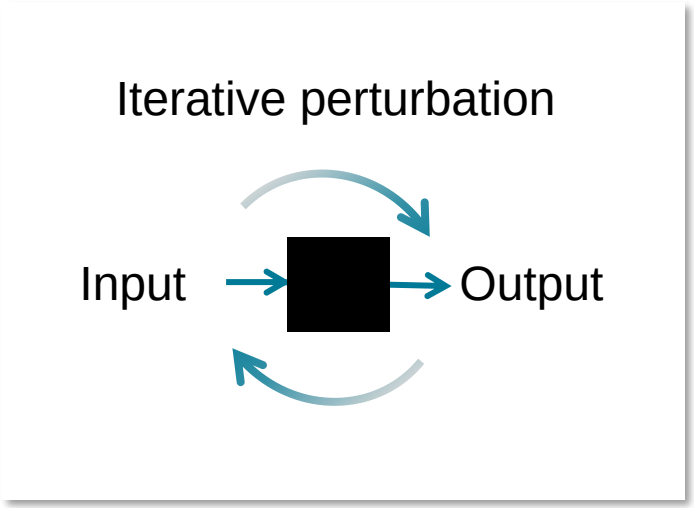


White box → model-dependent





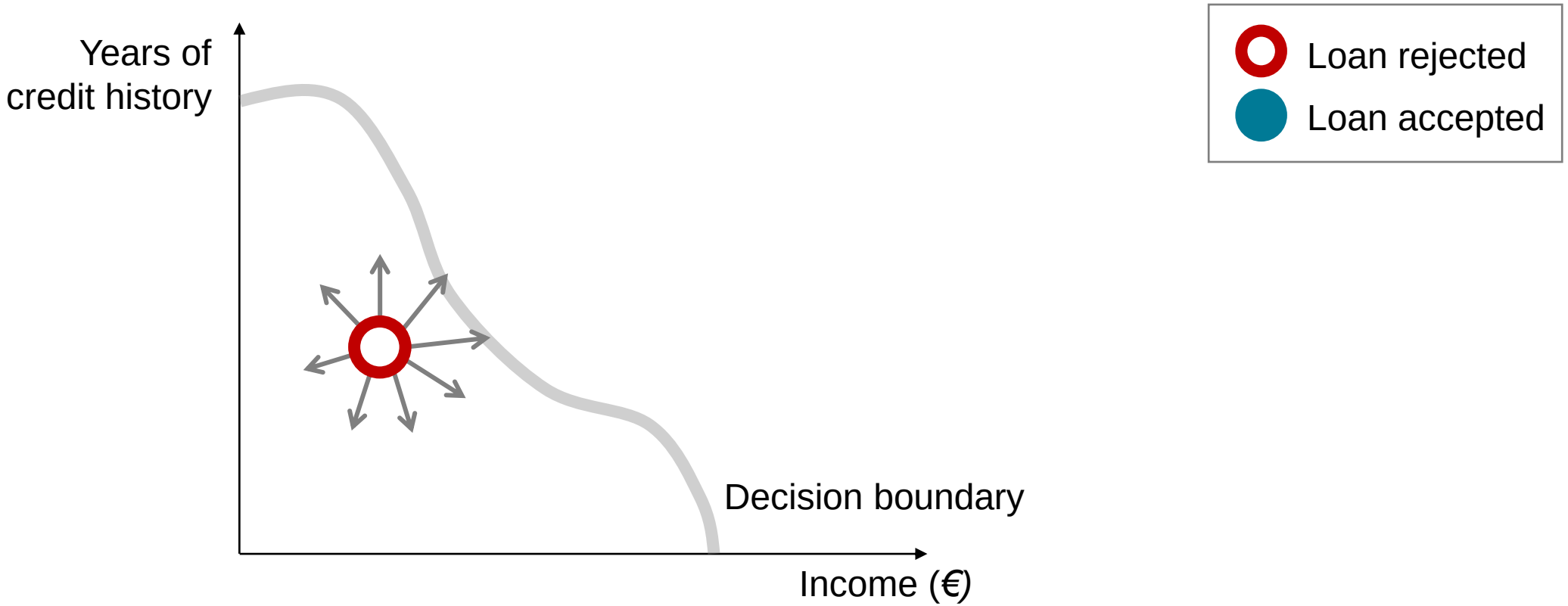
Black box → model-agnostic



TU/e

Counterfactual explanations

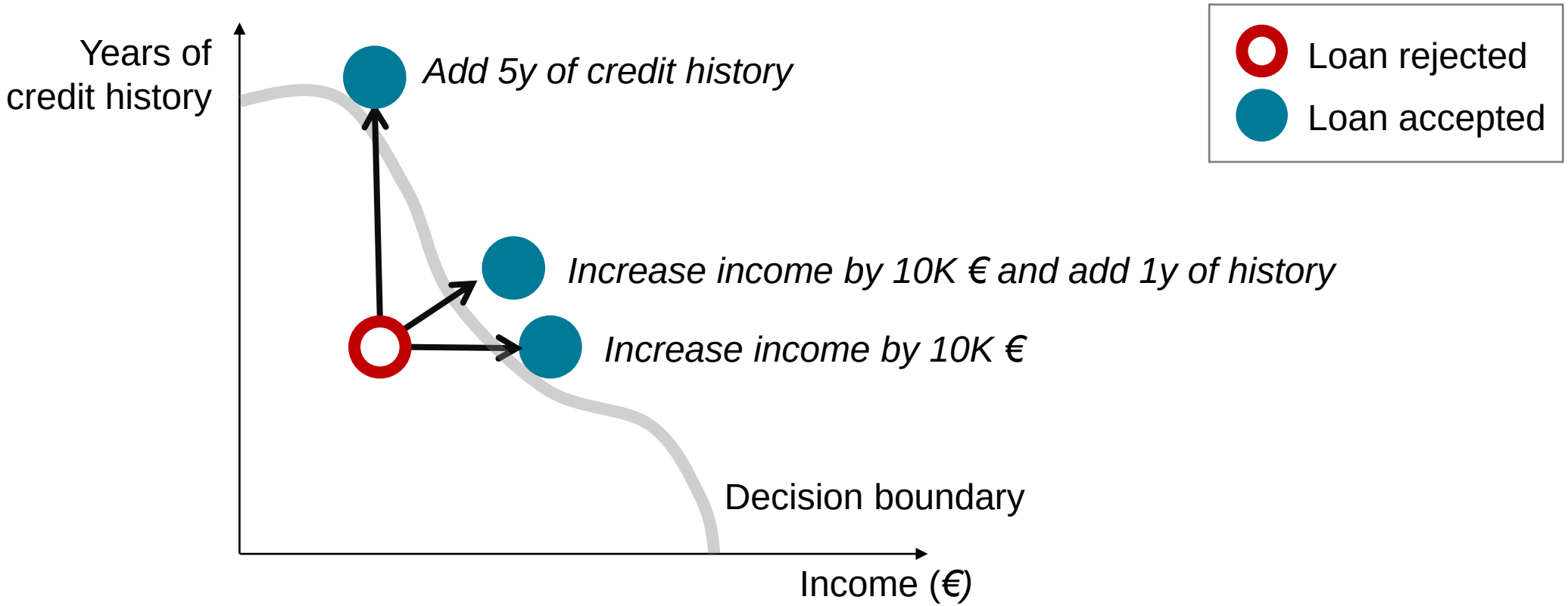
Counterfactual: smallest change in the input features that changes the predictions to another output.



TU/e

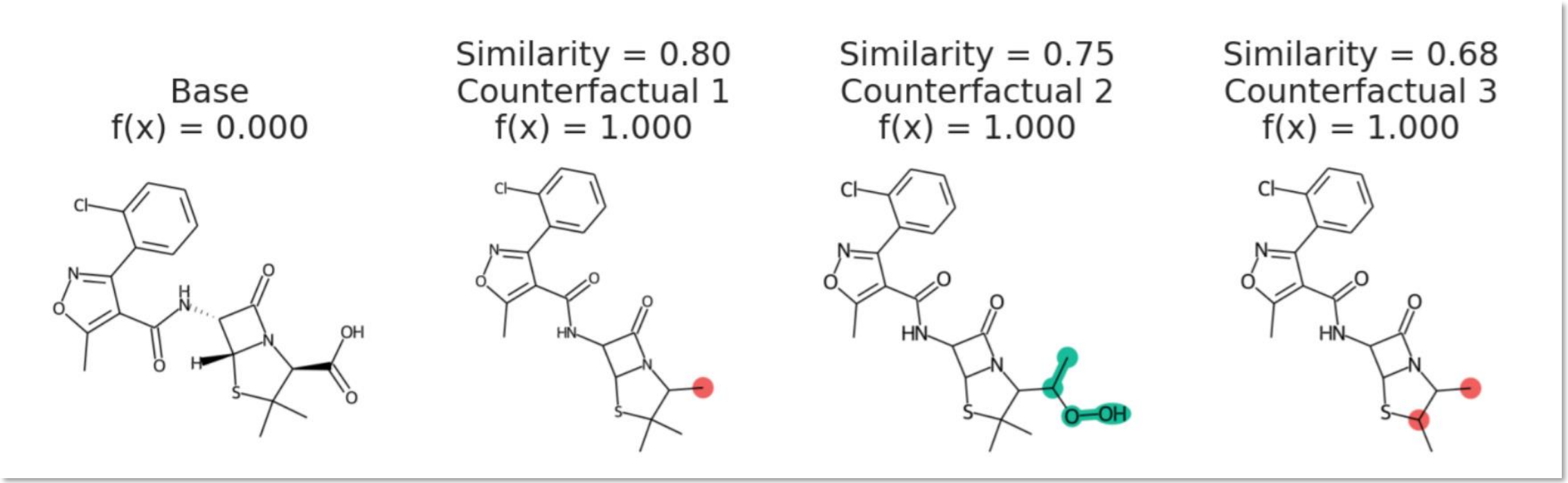
Counterfactual explanations

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



Counterfactual explanations for molecules

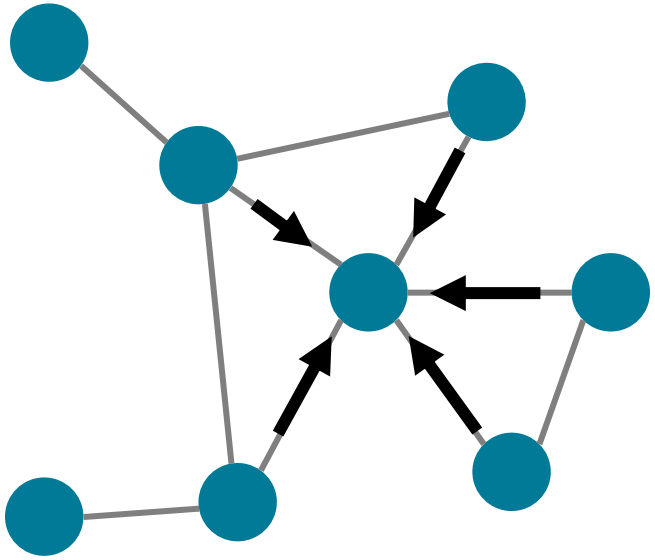
Blood-brain barrier permeating molecules

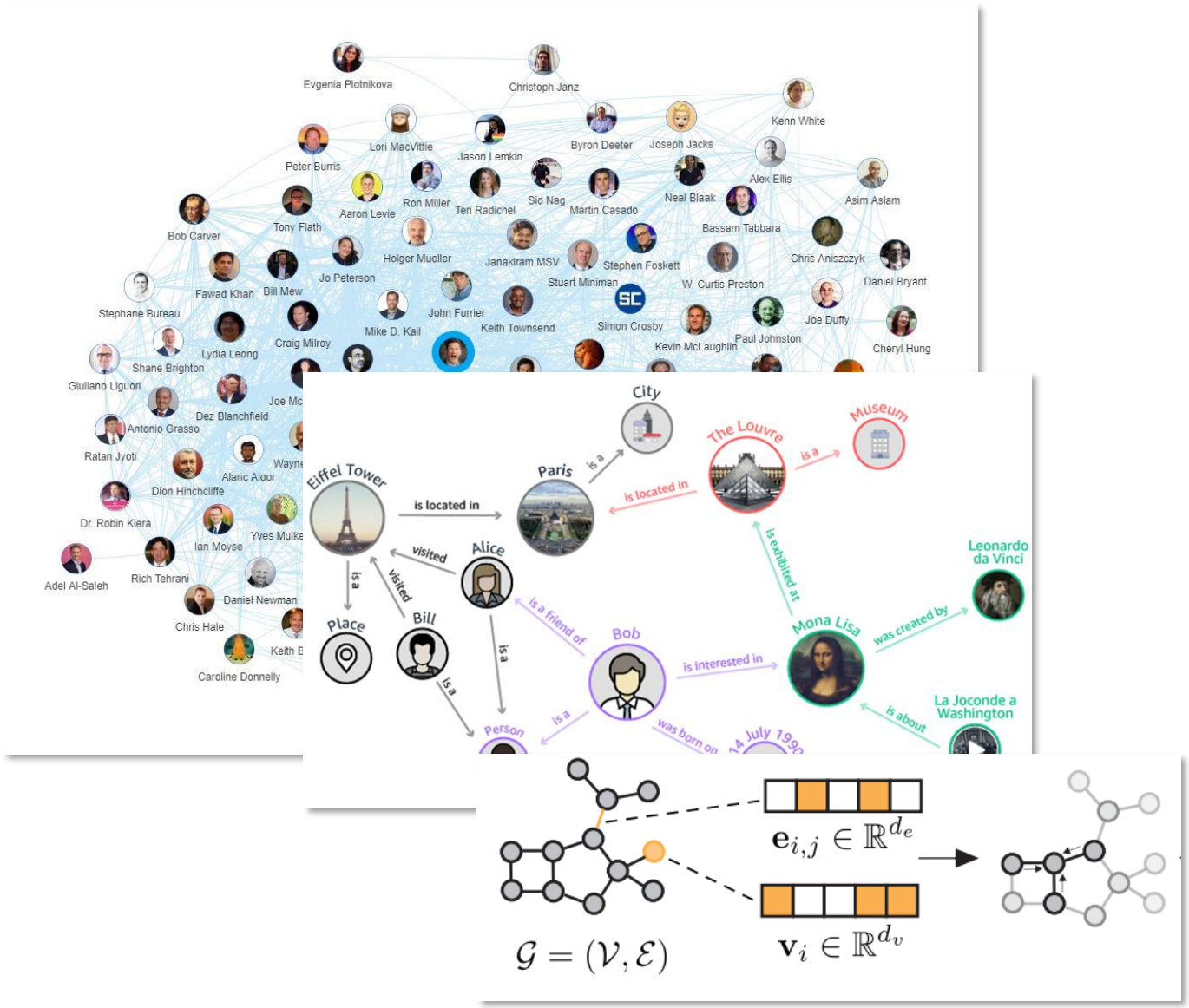


Wellawatte *et al.* (2021). ChemRviX.10.33774/chemrxiv-2021-4qkg8-v2.

TU/e

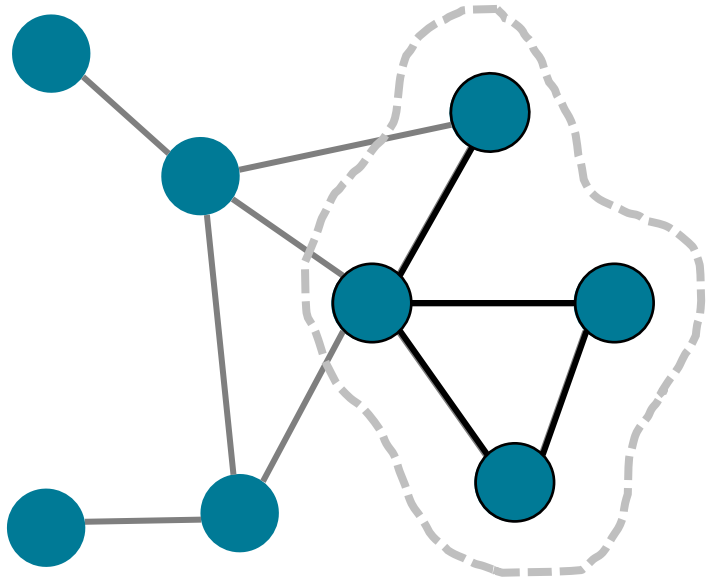
Graph neural networks





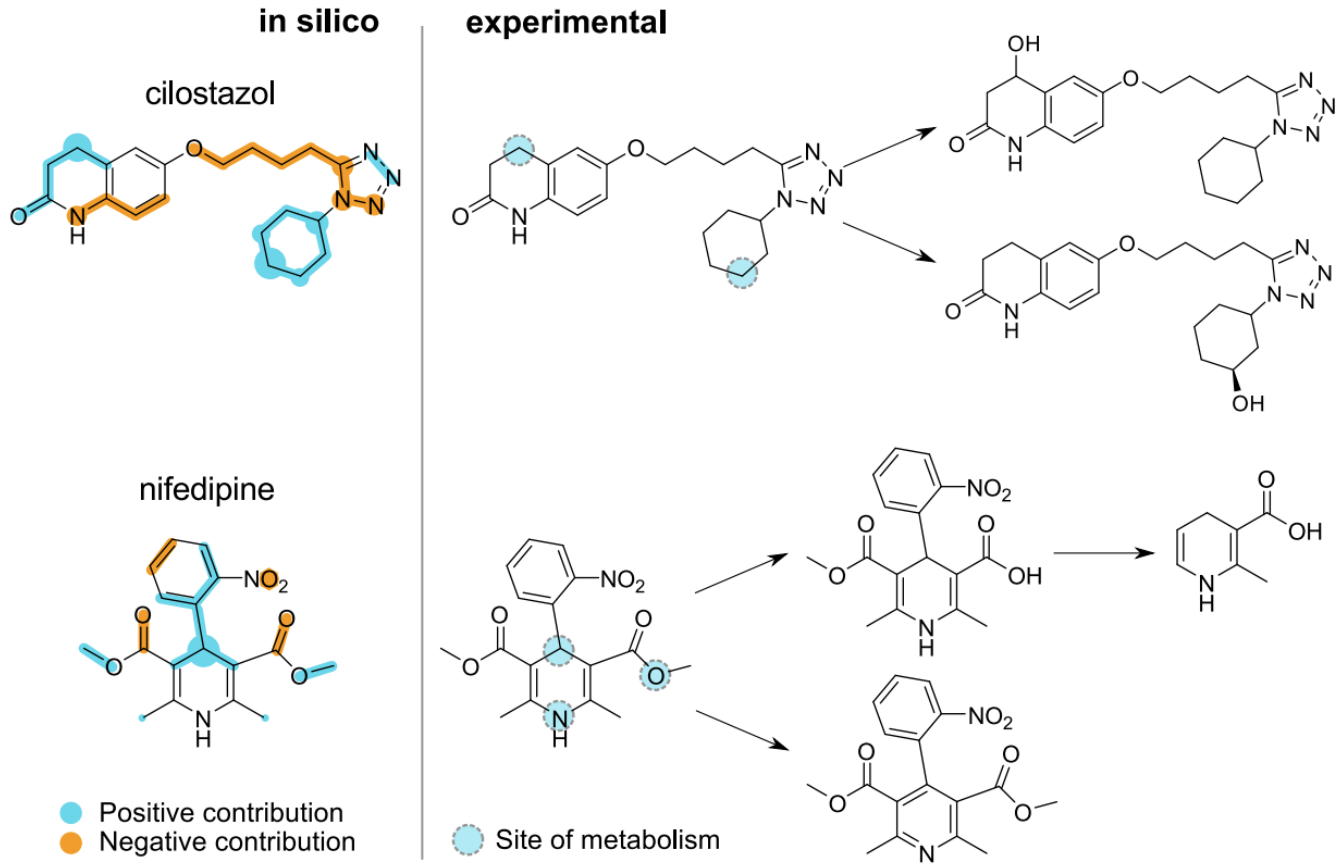
Graph neural networks

GNNExplainer[1]



Nodes and edges that maintain the outputs but are subsets of the graph

GNNs drug metabolism[2]

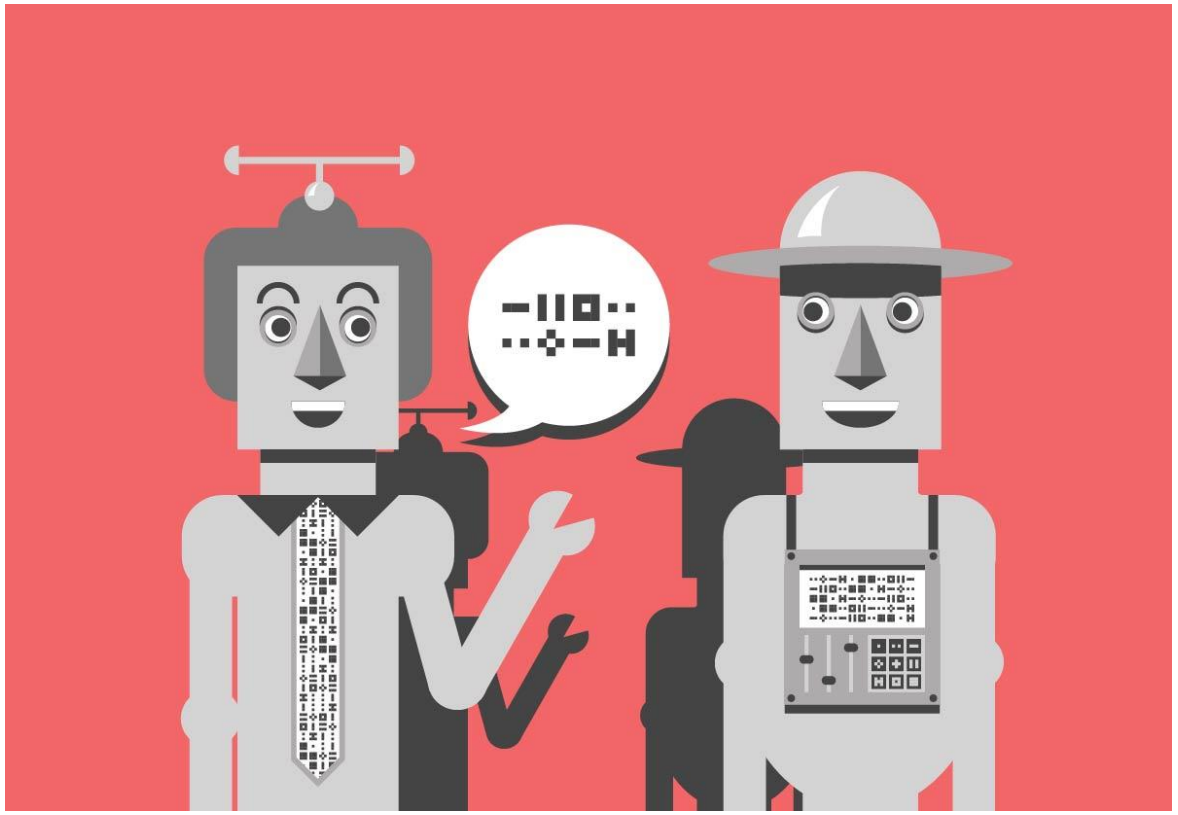


[1]Ying (2019). *Advances in neural information processing systems* **32**, 9240.
[2]Jiménez-Luna *et al.* (2020). *Nature Machine Intelligence* **2**, 573.

TU/e

Self-explaining approaches

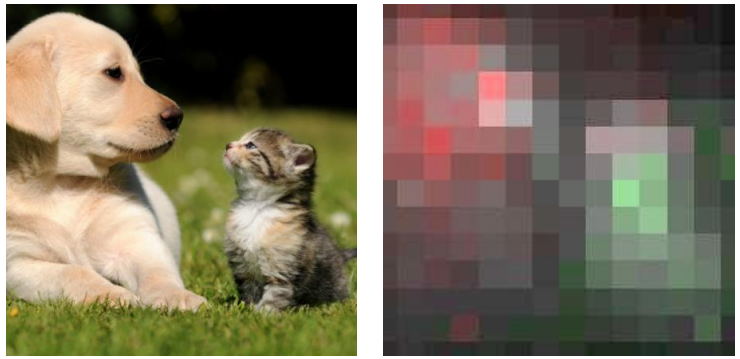
- Intrinsic explanation.
- Glimpse into the inner functioning of the model.
- Promote verification and error analysis, and be directly linkable with domain knowledge.



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

Feature attribution vs feature visualization

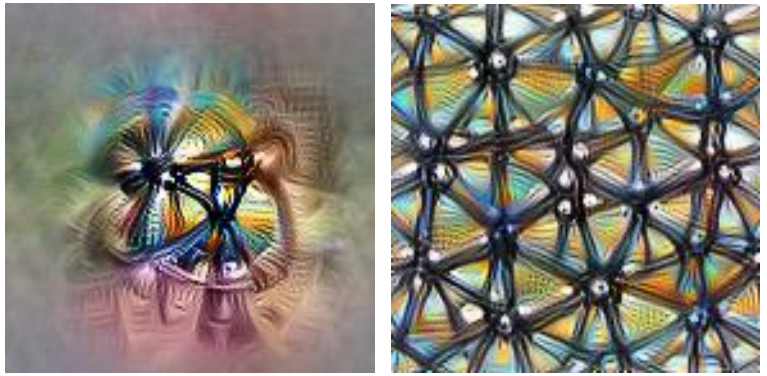
Feature attribution



What part of the example is responsible for the prediction?

Image *in*

Feature visualization



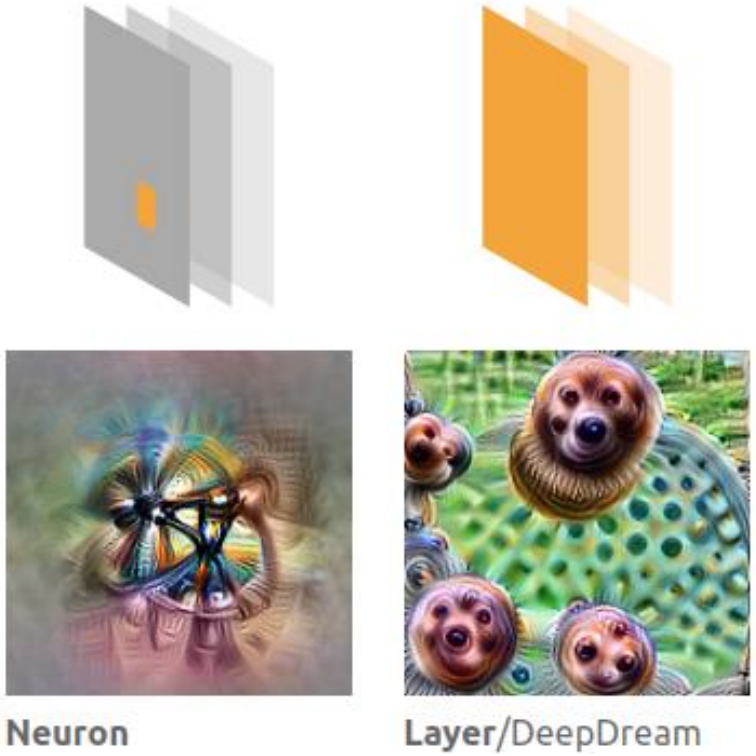
What is a network looking for?

Image *out*

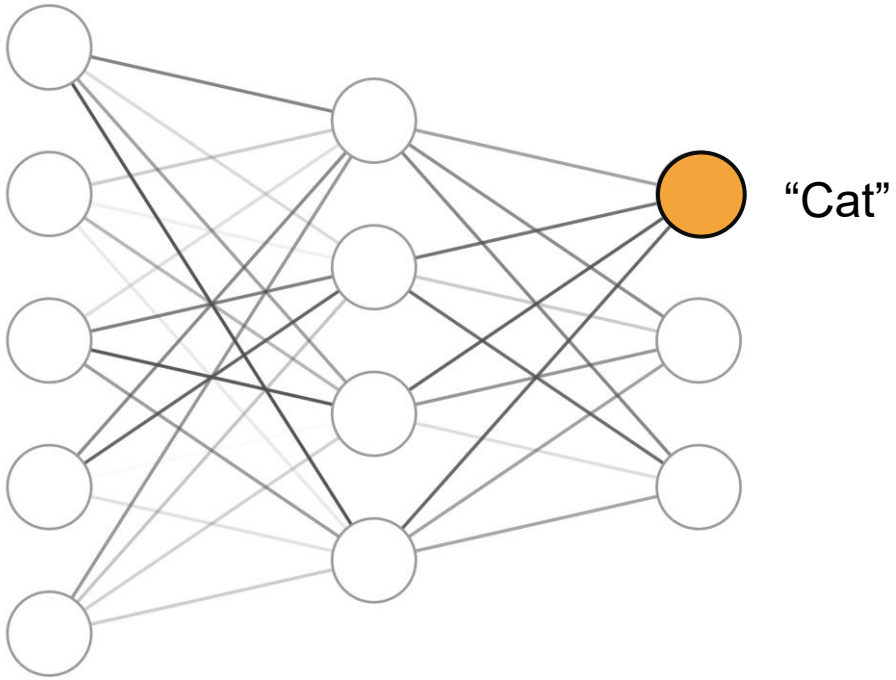
Images from: https://distill.pub/2017/feature-visualization/

TU/e

Feature visualization



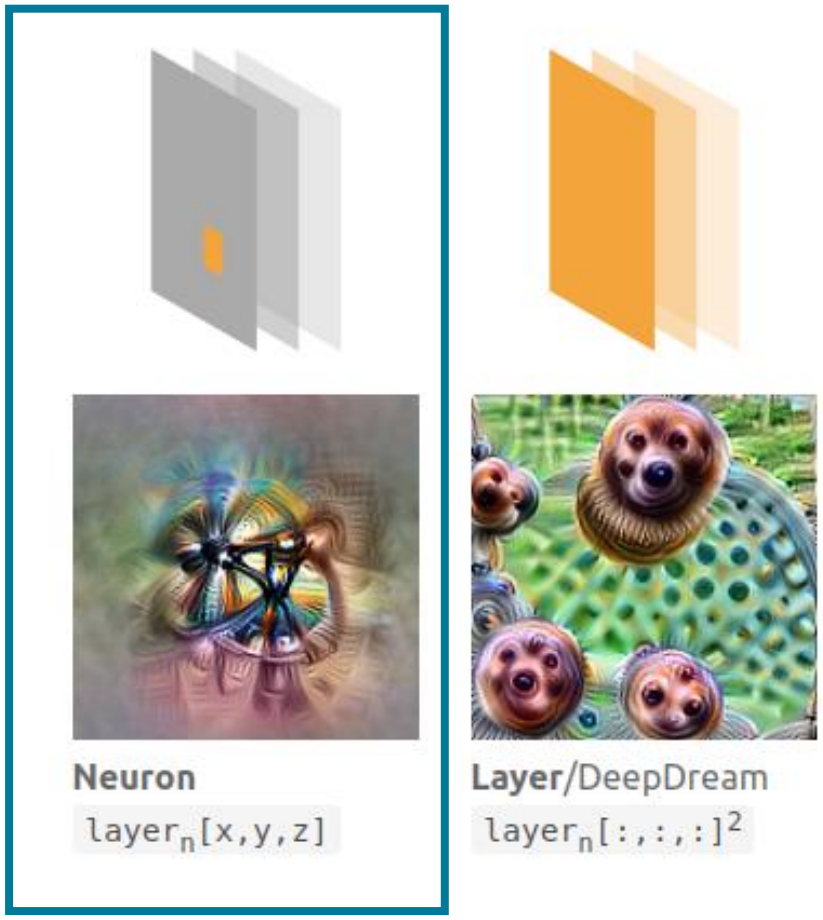
Real picture

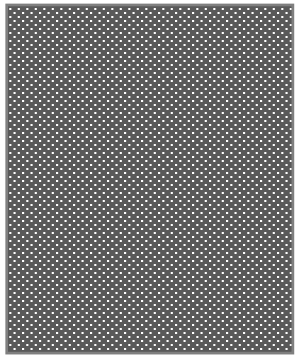


Source: https://ai.googleblog.com/2018/03/the-building-blocks-of-interpretability.html.

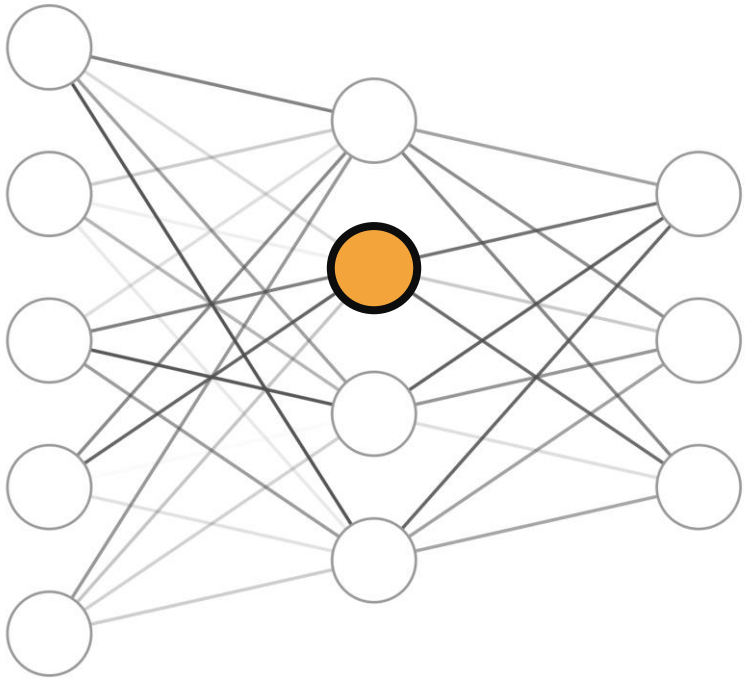
TU/e

Feature visualization





Random noise



Backpropagation

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

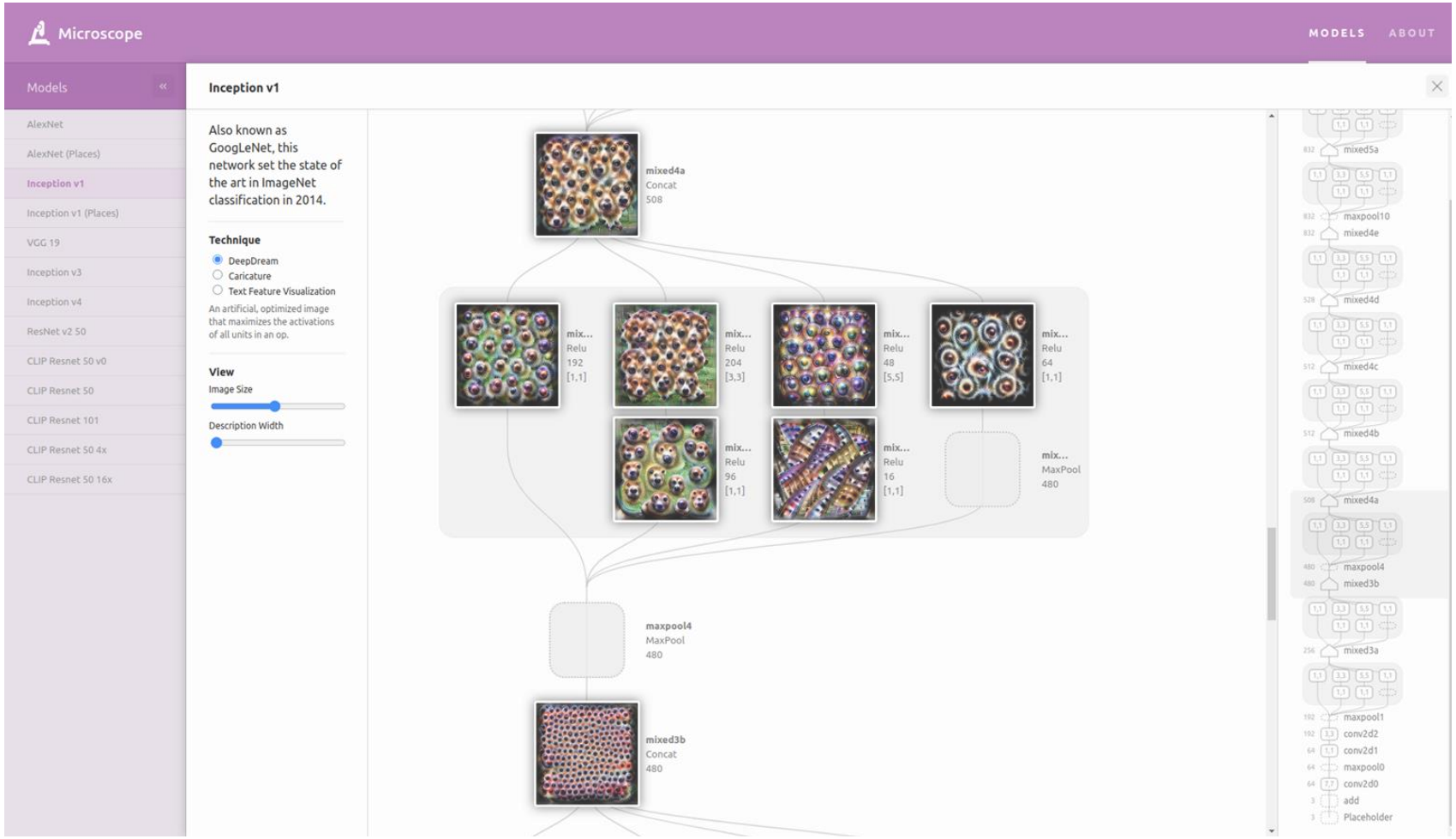
TU/e

Visualizing neurons

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

TU/e

OpenAI microscope



https://microscope.openai.com/models/inceptionv1?models.technique=deep_dream

TU/e

Feature visualization

Pros

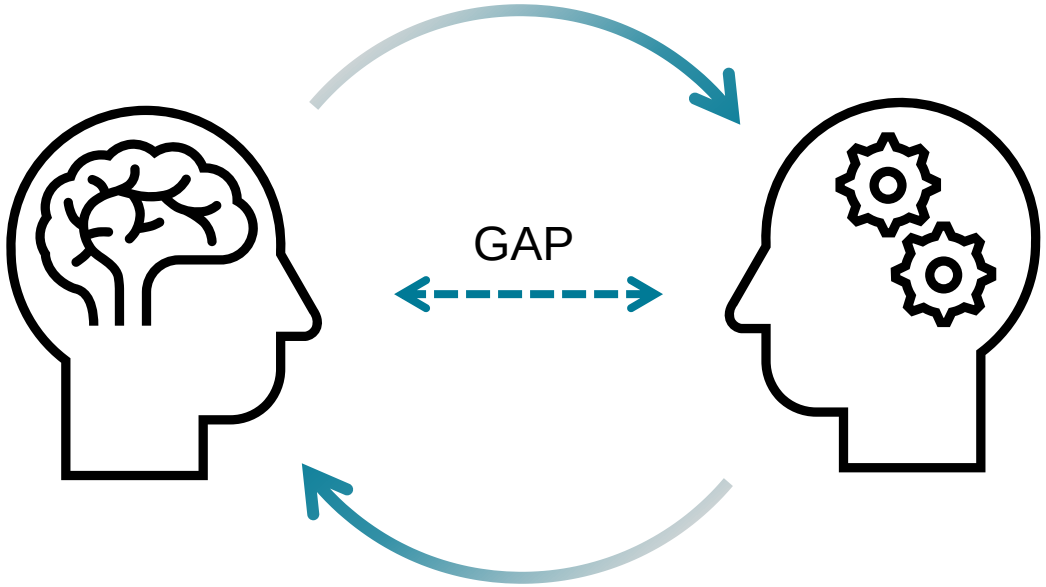
- Automatically link units to concepts
- Non-technical communication
- Detect concepts beyond the class labels
- Make great t-shirts/posters!

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



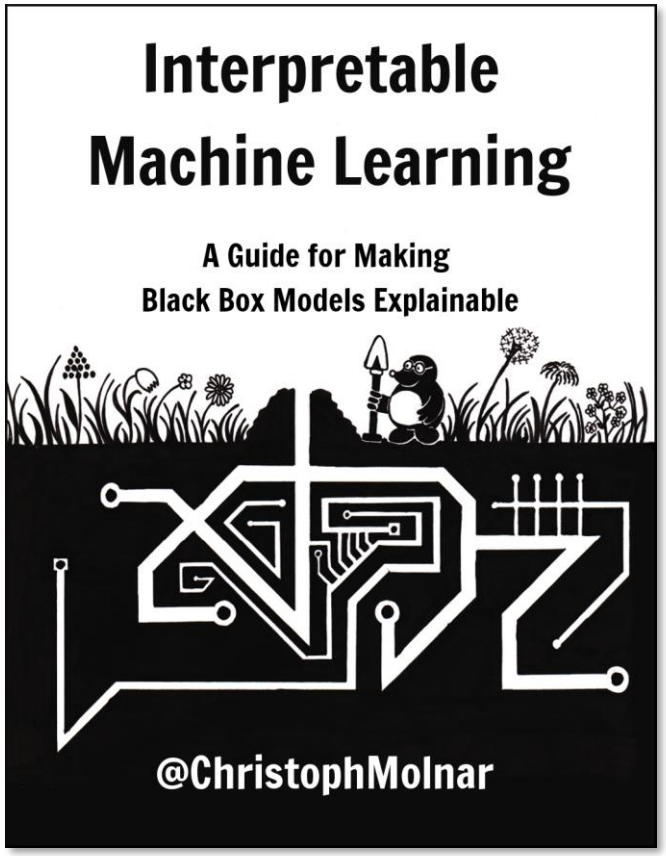
- How to interpret?
- Partial information
- Lack of systematic evaluation

TU/e

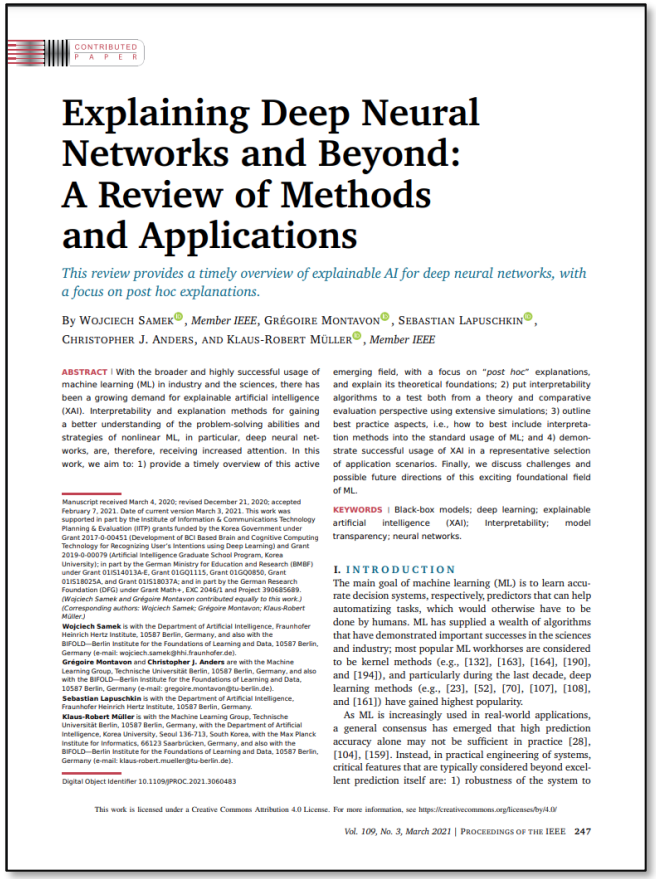
Summing it up

- 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 input data.
- Explanation is not interpretation.
- Lack of systematic evaluation of explainability vs interpretability.

Want to know more?



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



CONTRIBUTED PAPER

Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications

This review provides a timely overview of explainable AI for deep neural networks, with a focus on post hoc explanations.

By Wojciech Samek, *Member IEEE*, Grégoire Montavon, Sebastian Lapuschkin, Christopher J. Anders, and Klaus-Robert Müller, *Member IEEE*

ABSTRACT | With the broader and highly successful usage of machine learning (ML) in industry and the sciences, there has been a growing demand for explainable artificial intelligence (XAI). Interpretability and explanation methods for gaining a better understanding of the problem-solving abilities and strategies of nonlinear ML, in particular, deep neural networks, are, therefore, receiving increased attention. In this work, we aim to: 1) provide a timely overview of this active emerging field, with a focus on "*post hoc*" explanations, and explain its theoretical foundations; 2) put interpretability algorithms to a test both from a theory and comparative evaluation perspective using extensive simulations; 3) outline best practice aspects, i.e., how to best include interpretation methods into the standard usage of ML; and 4) demonstrate successful usage of XAI in a representative selection of application scenarios. Finally, we discuss challenges and possible future directions of this exciting foundational field of ML.

KEYWORDS | Black-box models; deep learning; explainable artificial intelligence (XAI); Interpretability; model transparency; neural networks.

Manuscript received March 4, 2020; revised December 21, 2020; accepted February 7, 2021. Date of current version March 3, 2021. This work was supported in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea Government under Grant 2017-0-00451 (Development of BCI Based Brain and Cognitive Computing Technology for Recognizing User's Intentions using Deep Learning) and Grant 2019-0-00079 (Artificial Intelligence Graduate School Program, Korea University); in part by the German Ministry for Education and Research (BMBF) under Grant 01IS14013A-E, Grant 01GQ1115, Grant 01GQ0850, Grant 01IS18025A, and Grant 01IS18037A; and in part by the German Research Foundation (DFG) under Grant Math+, EXC 2046/1 and Project 390685689. (*Wojciech Samek and Grégoire Montavon contributed equally to this work.*) (*Corresponding authors: Wojciech Samek; Grégoire Montavon; Klaus-Robert Müller.*)

Wojciech Samek is with the Department of Artificial Intelligence, Fraunhofer Heinrich Hertz Institute, 10587 Berlin, Germany, and also with the BIFOLD—Berlin Institute for the Foundations of Learning and Data, 10587 Berlin, Germany (e-mail: wojciech.samek@hhi.fraunhofer.de).

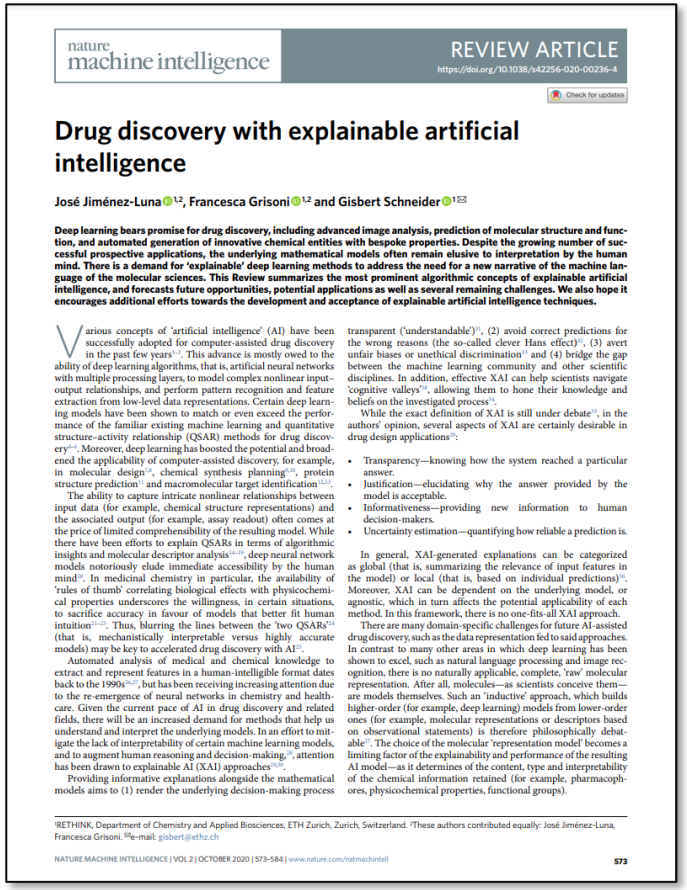
Grégoire Montavon and **Christopher J. Anders** are with the Machine Learning Group, Technische Universität Berlin, 10587 Berlin, Germany, and also with the BIFOLD—Berlin Institute for the Foundations of Learning and Data, 10587 Berlin, Germany (e-mail: gregoire.montavon@tu-berlin.de).

Sebastian Lapuschkin is with the Department of Artificial Intelligence, Fraunhofer Heinrich Hertz Institute, 10587 Berlin, Germany.

Klaus-Robert Müller is with the Machine Learning Group, Technische Universität Berlin, 10587 Berlin, Germany, with the Department of Artificial Intelligence, Korea University, Seoul 136-713, South Korea, with the Max Planck Institute for Informatics, 66123 Saarbrücken, Germany, and also with the BIFOLD—Berlin Institute for the Foundations of Learning and Data, 10587 Berlin, Germany (e-mail: klaus-robert.mueller@tu-berlin.de).

Digital Object Identifier 10.1109/JPROC.2021.3060483

https://ieeexplore.ieee.org/document/9369420



nature machine intelligence

REVIEW ARTICLE

https://doi.org/10.1038/s42256-020-00236-4

Check for updates

Drug discovery with explainable artificial intelligence

José Jiménez-Luna[1,2], Francesca Grisoni[1,2] and Gisbert Schneider[1]

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 prospective applications, the underlying mathematical models often remain elusive to interpretation by the human mind. There is a demand for 'explainable' deep learning methods to address the need for a new narrative of the machine language of the molecular sciences. This Review summarizes the most prominent 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.

Various concepts of 'artificial intelligence' (AI) have been successfully adopted for computer-assisted drug discovery in the past few years[1–7]. This advance is mostly owed to the ability of deep learning algorithms, that is, artificial neural networks with multiple processing layers, to model complex nonlinear input–output relationships, and perform pattern recognition and feature extraction from low-level data representations. Certain deep learning models have been shown to match or even exceed the performance of the familiar existing machine learning and quantitative structure–activity relationship (QSAR) methods for drug discovery[8–16]. Moreover, deep learning has boosted the potential and broadened the applicability of computer-assisted discovery, for example, in molecular design[17,18], chemical synthesis planning[19,20], protein structure prediction[21] and macromolecular target identification[22,23].

The ability to capture intricate nonlinear relationships between input data (for example, chemical structure representations) and the associated output (for example, assay readout) often comes at the price of limited comprehensibility of the resulting model. While there have been efforts to explain QSARs in terms of algorithmic insights and molecular descriptor analysis[24–26], deep neural network models notoriously elude immediate accessibility by the human mind[27]. In medicinal chemistry in particular, the availability of 'rules of thumb' correlating biological effects with physicochemical properties underscores the willingness, in certain situations, to sacrifice accuracy in favour of models that better fit human intuition[27–29]. Thus, blurring the lines between the 'two QSARs'[24] (that is, mechanistically interpretable versus highly accurate models) may be key to accelerated drug discovery with AI[25].

Automated analysis of medical and chemical knowledge to extract and represent features in a human-intelligible format dates back to the 1990s[30–37], but has been receiving increasing attention due to the re-emergence of neural networks in chemistry and healthcare. Given the current pace of AI in drug discovery and related fields, there will be an increased demand for methods that help us understand and interpret the underlying models. In an effort to mitigate the lack of interpretability of certain machine learning models, and to augment human reasoning and decision-making[38], attention has been drawn to explainable AI (XAI) approaches[39,40].

Providing informative explanations alongside the mathematical models aims to (1) render the underlying decision-making process transparent ('understandable')[41], (2) avoid correct predictions for the wrong reasons (the so-called clever Hans effect)[42], (3) avert unfair biases or unethical discrimination[43] and (4) bridge the gap between the machine learning community and other scientific disciplines. In addition, effective XAI can help scientists navigate 'cognitive valleys'[39], allowing them to hone their knowledge and beliefs on the investigated process[44].

While the exact definition of XAI is still under debate[45], in the authors' opinion, several aspects of XAI are certainly desirable in drug design applications[46]:

- Transparency—knowing how the system reached a particular answer.
- Justification—elucidating why the answer provided by the model is acceptable.
- Informativeness—providing new information to human decision-makers.
- Uncertainty estimation—quantifying how reliable a prediction is.

In general, XAI-generated explanations can be categorized as global (that is, summarizing the relevance of input features in the model) or local (that is, based on individual predictions)[36]. Moreover, XAI can be dependent on the underlying model, or agnostic, which in turn affects the potential applicability of each method. In this framework, there is no one-fits-all XAI approach.

There are many domain-specific challenges for future AI-assisted drug discovery, such as the data representation fed to said approaches. In contrast to many other areas in which deep learning has been shown to excel, such as natural language processing and image recognition, there is no naturally applicable, complete, 'raw' molecular representation. After all, molecules—as scientists conceive them—are models themselves. Such an 'inductive' approach, which builds higher-order (for example, deep learning) models from lower-order ones (for example, molecular representations or descriptors based on observational statements) is therefore philosophically debatable[47]. The choice of the molecular 'representation model' becomes a limiting factor of the explainability and performance of the resulting AI model—as it determines of the content, type and interpretability of the chemical information retained (for example, pharmacophores, physicochemical properties, functional groups).

[1]RETHINK, Department of Chemistry and Applied Biosciences, ETH Zurich, Zurich, Switzerland. [2]These authors contributed equally: José Jiménez-Luna, Francesca Grisoni. ✉e-mail: gisbert@ethz.ch

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

TU/e