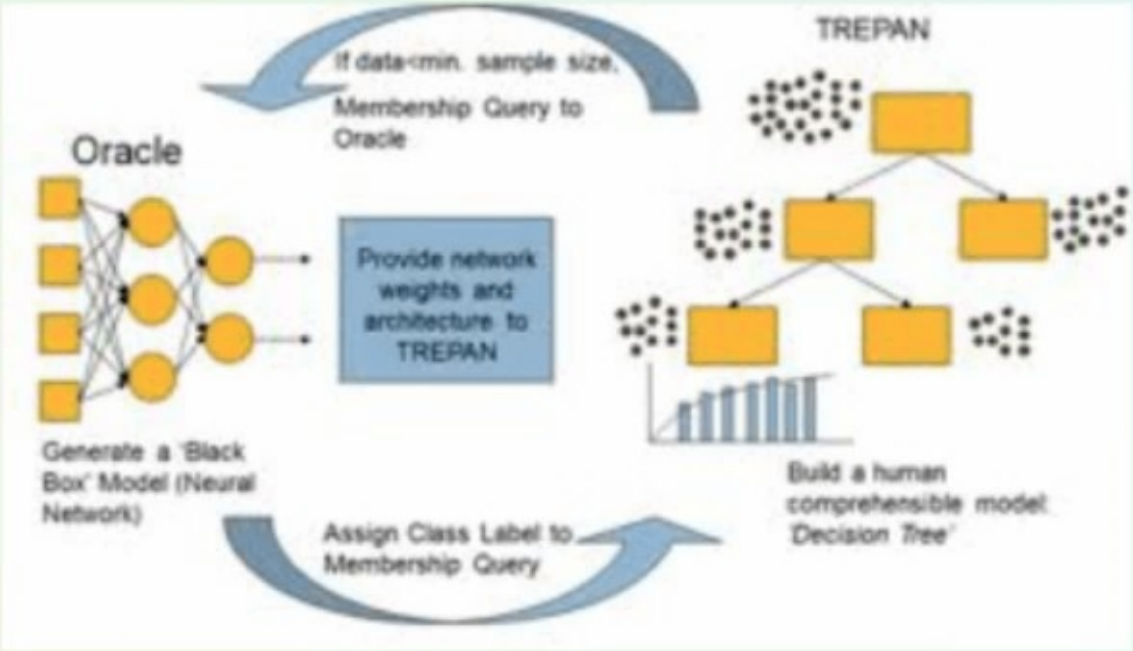
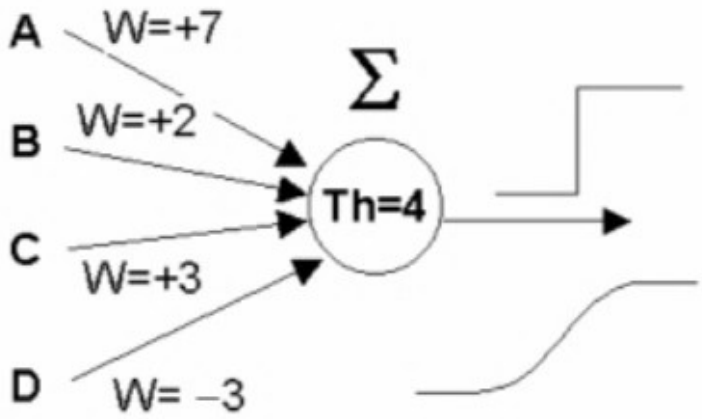
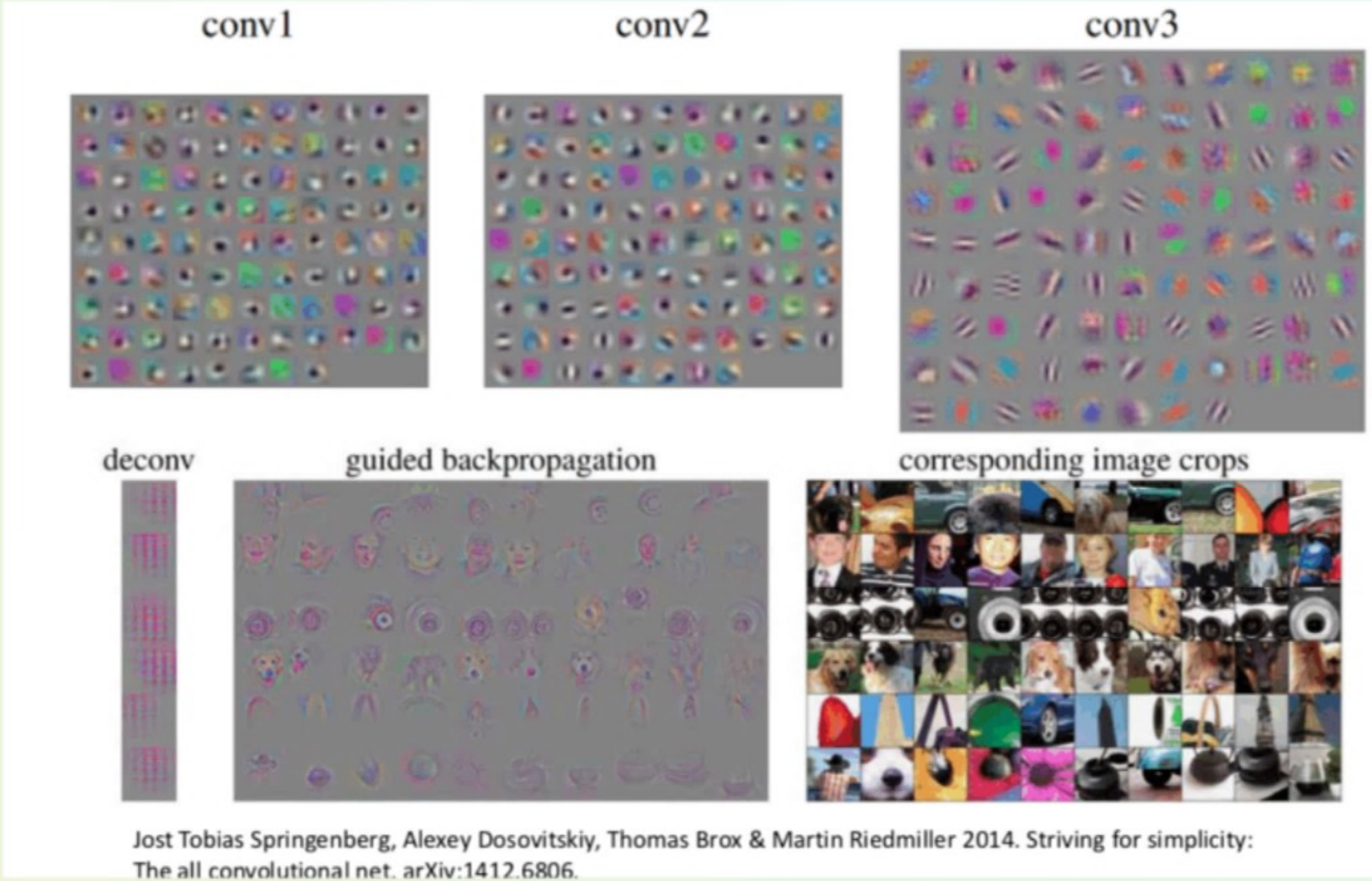
Tworzenie interpretowalnych reprezentacji zastępczych dla złożonych modeli (ANN)



IF{(A and B) or (A and C) or
(A and not D) and
(A and B and C) and
(B and C and not D)}
THEN True
ELSE False

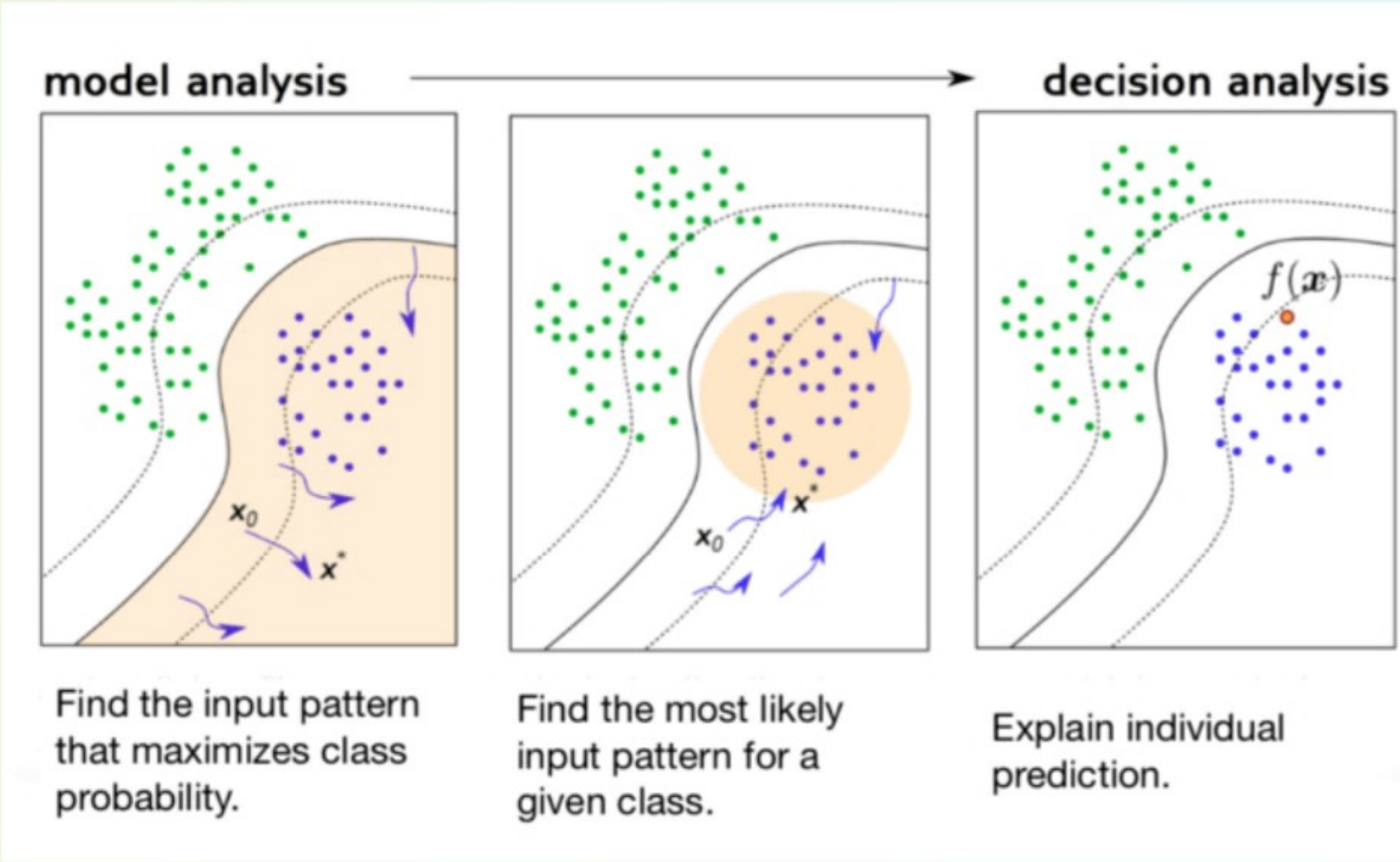


Deconvolution of CNN and other visualizations of internal layers



Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox & Martin Riedmiller 2014. Striving for simplicity: The all convolutional net. arXiv:1412.6806.

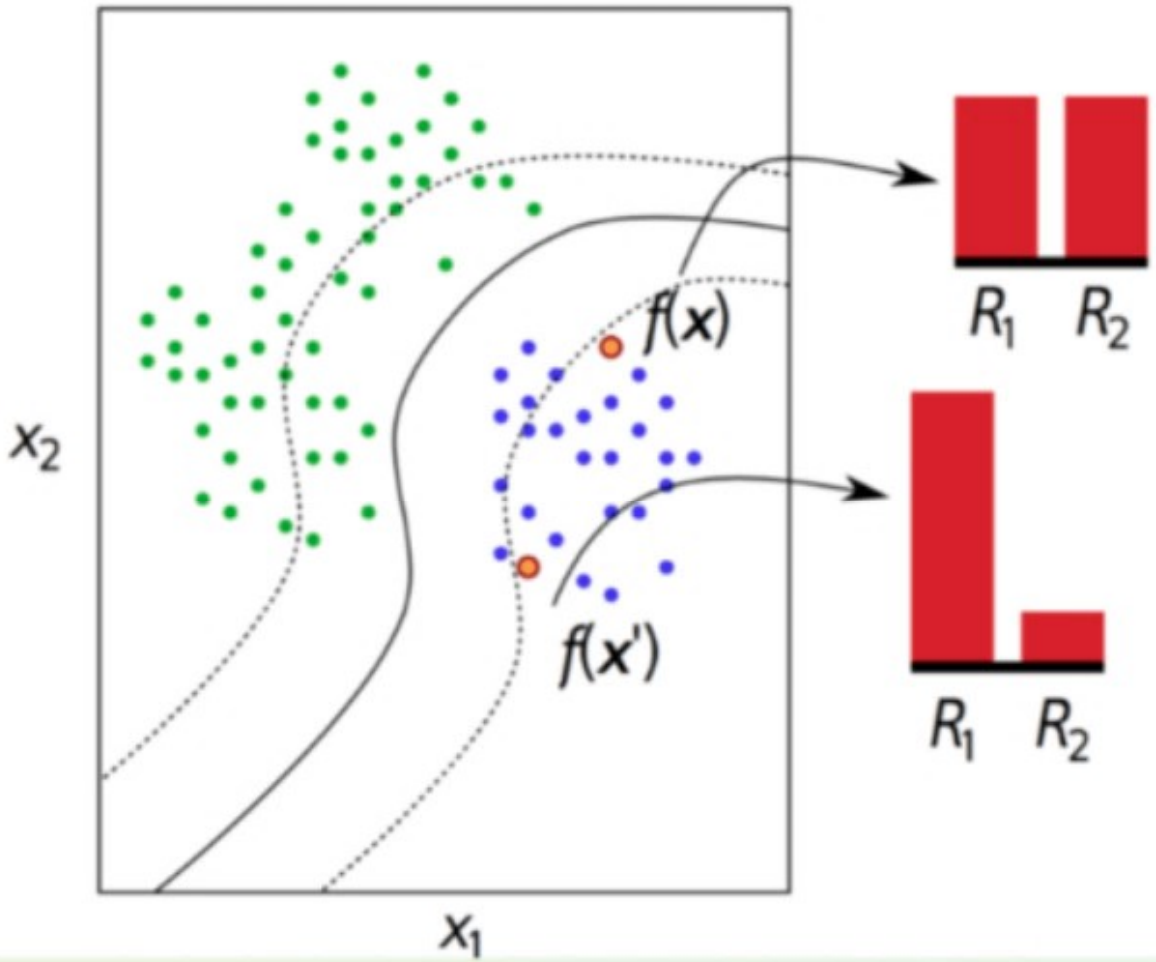
Skupienie uwagi na przykładach



Różne możliwości / za wykladem W.Samek

Explaining Individual Decisions

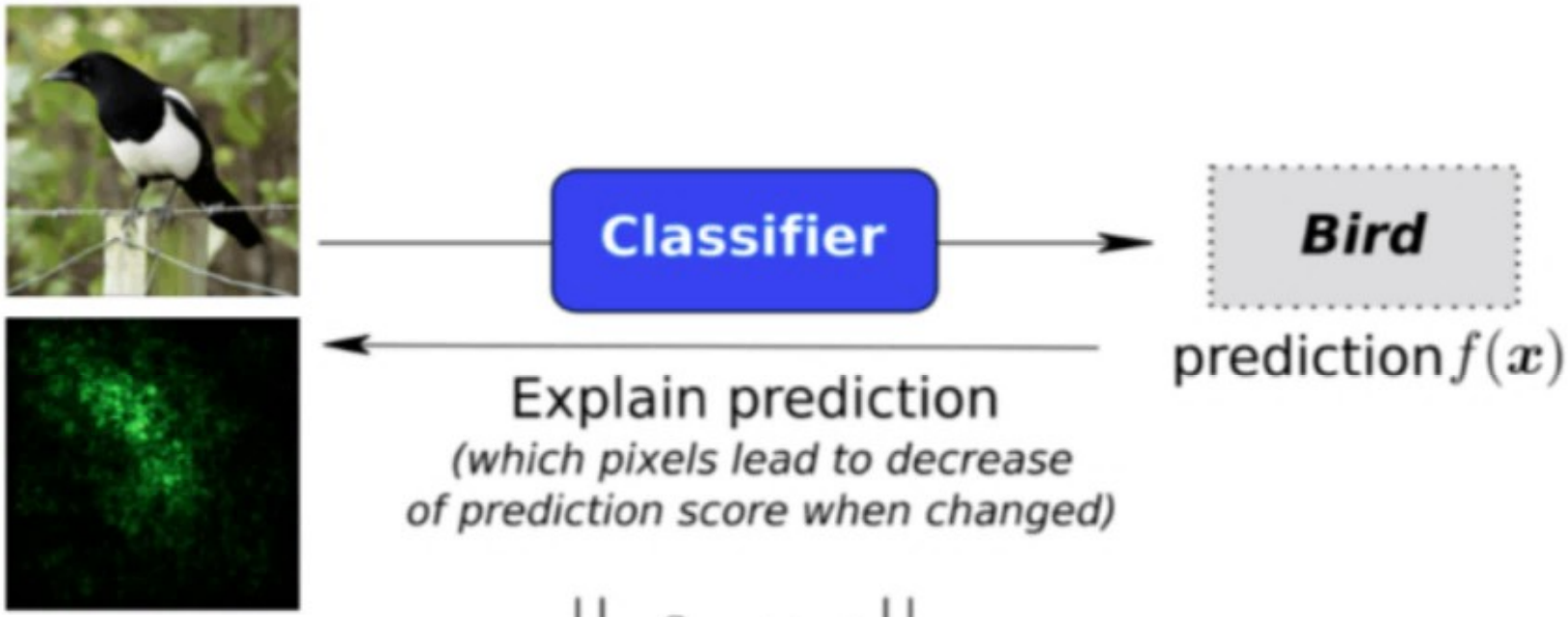
Goal: Determine the relevance of each input variable for a given decision $f(x_1, x_2, \ldots, x_d)$, by assigning to these variables *relevance* scores $R_1, R_2, \ldots, R_d$.



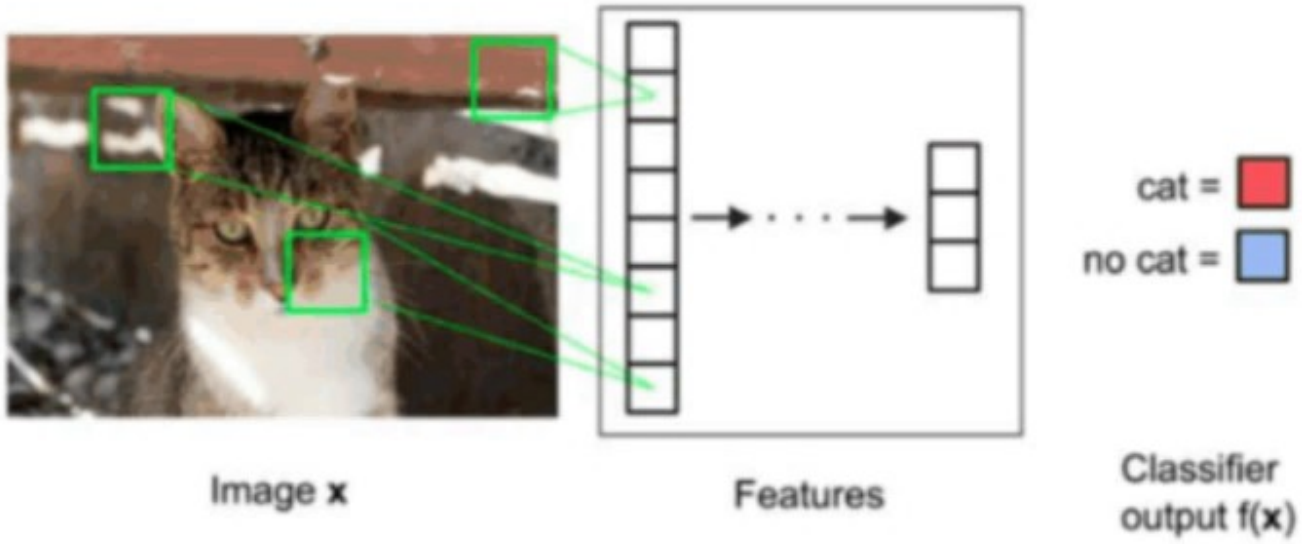
Poszukiwanie krytycznych elementów w danych

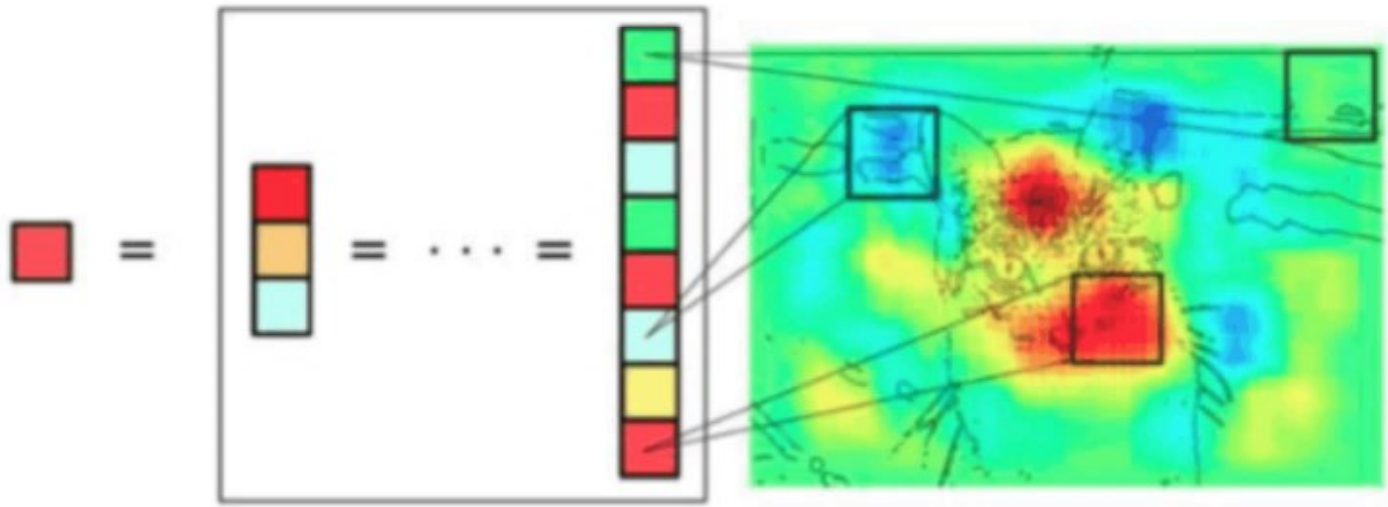
Techniques of Interpretation

Sensitivity Analysis
(Simonyan et al. 2014)



Specyfika wyjaśnień obrazowych





Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek 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.

Techniques of Interpretation

Layer-wise Relevance Propagation (LRP)
(Bach et al. 2015)

input $\boldsymbol{x}$

Classifier

Rooster

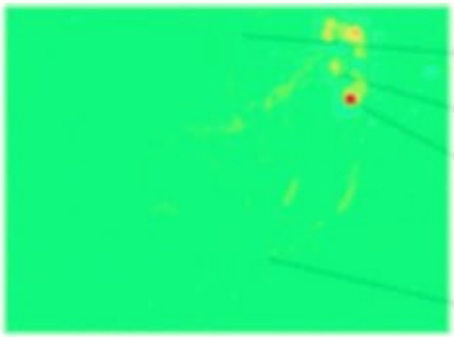
prediction $f(\boldsymbol{x})$

Explain prediction

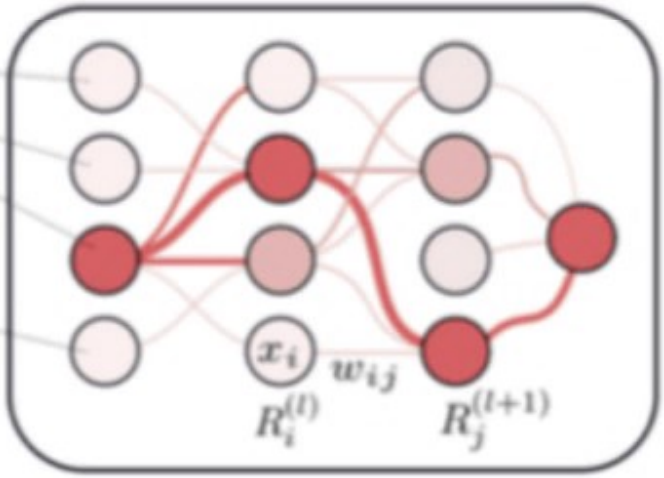
(how much each pixel contributes to prediction)

Idea: Decompose function

$$\sum_i R_i = f(x)$$



heatmap



"every neuron gets it's share of relevance depending on activation and strength of connection."

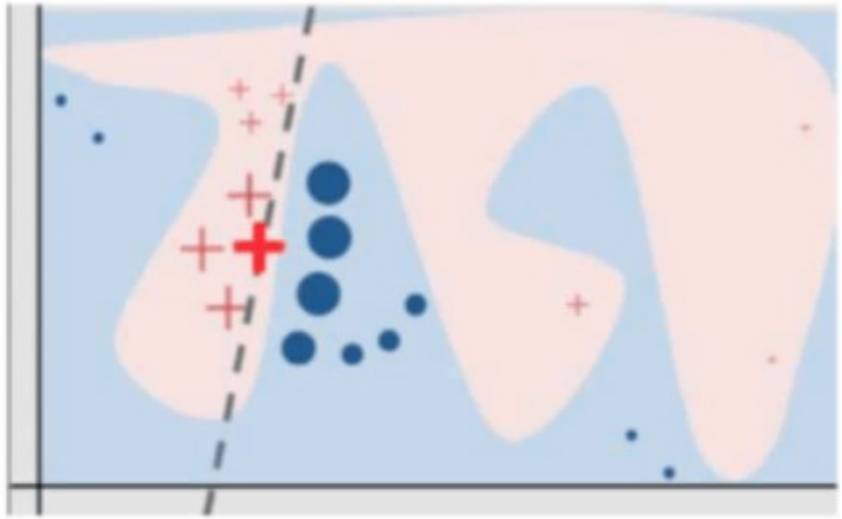
redistribute $f(\boldsymbol{x})$

LIME

What are explanations in the sense of LIME ?

HCAI

- Explanation := local linear approximation of the model's behaviour. While the model may be very complex globally, it is easier to approximate it around the vicinity of a particular instance. While treating the model as a black box, we perturb the instance we want to explain and learn a sparse linear model around it -> used as explanation.

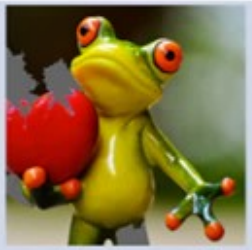


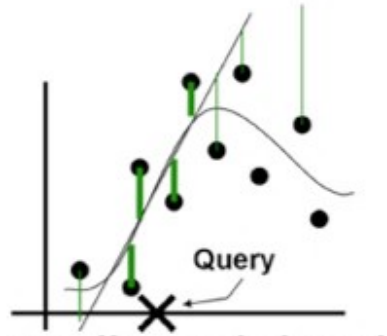
- Look at the image: The model's decision function is represented by the blue/pink background = clearly nonlinear. The bright red cross is the instance being explained (let's call it X). We sample instances around X, and weight them according to their proximity to X (weight here is indicated by size). We then learn a linear model (dashed line) that approximates the model well in the vicinity of X, but not necessarily globally!

https://github.com/marcotcr/lime

LIME – rozpoznawanie obrazów

Original Image
P(tree frog) = 0.54

Perturbed Instances	P(tree frog)
	0.85
	0.00001
	0.52

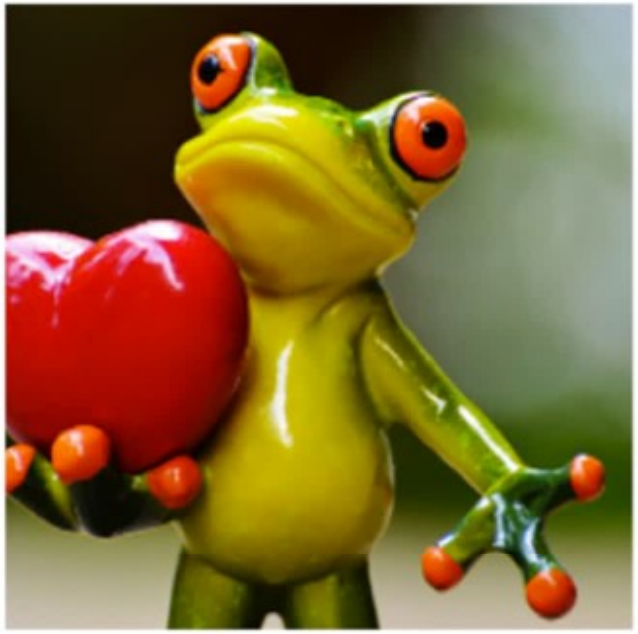


Locally weighted regression

Explanation

https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime

LIME – rozpoznawanie obrazów



Original Image

Interpretable Components

https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime

LIME – rozpoznawanie obrazów

Źródło: Explaining Black-Box Machine Learning Predictions – Sameer Singh

LIME – rozpoznawanie obrazów

Predicted: wolf
True: wolf

Predicted: husky
True: husky

Predicted: wolf
True: wolf

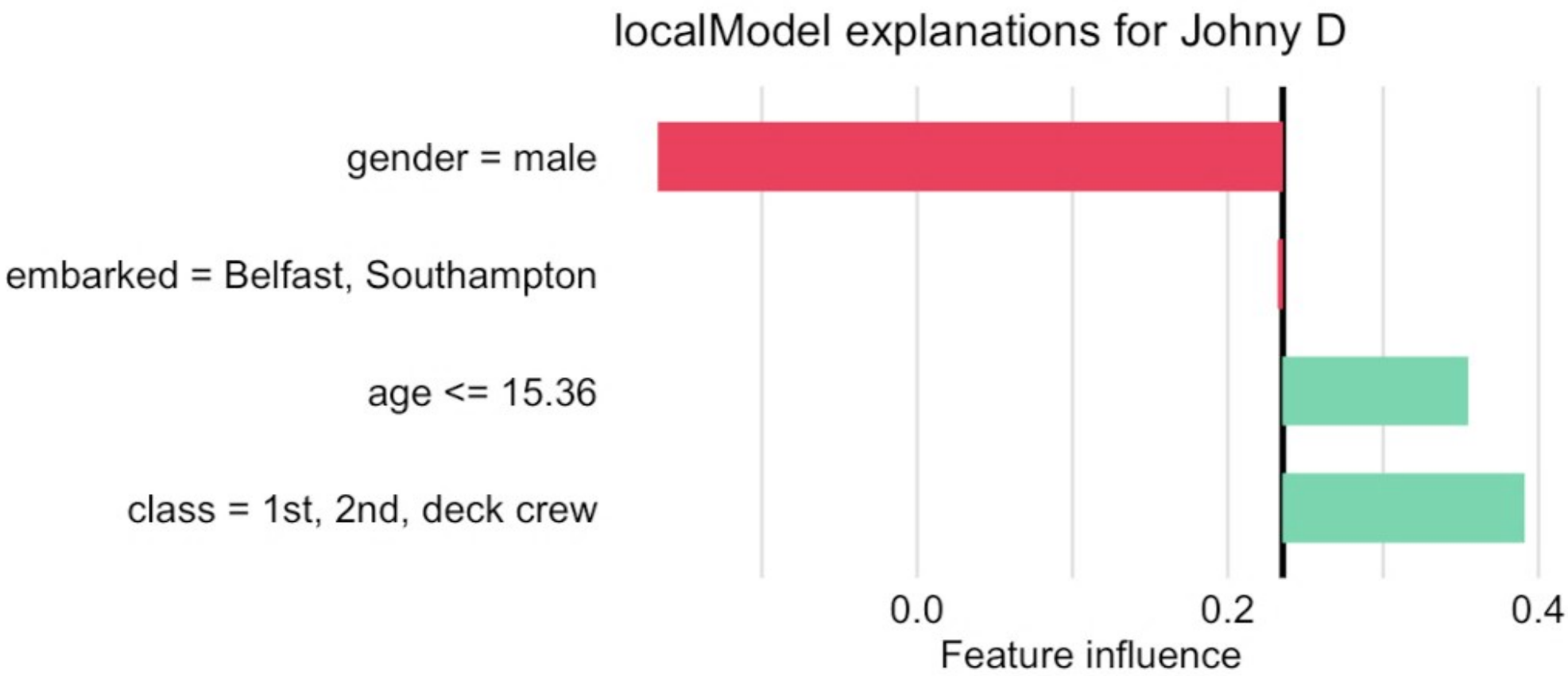
Predicted: wolf
True: husky

Predicted: husky
True: husky

Predicted: wolf
True: wolf

Źródło: Explaining Black-Box Machine Learning Predictions – Sameer Singh

LIME



Źródło: https://ema.drwhy.ai/LIME.html