TECHNISCHE UNIVERSITÄT DRESDEN
Faculty of Electrical and Computer Engineering, Chair of Fundamentals of Electronics

AWI - AI user group meeting

Unmasking Clever Hans Predictors
An introduction to explainable AI-assisted human fault diagnosis

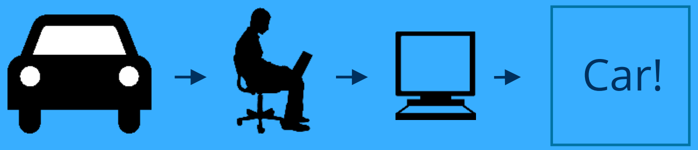
Steffen Seitz

Prof. Ronald Tetzlaff

Dresden, 06.12.22

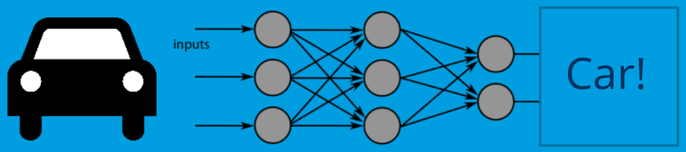
Classic Machine Learning vs. Deep Learning

Classic Machine Learning

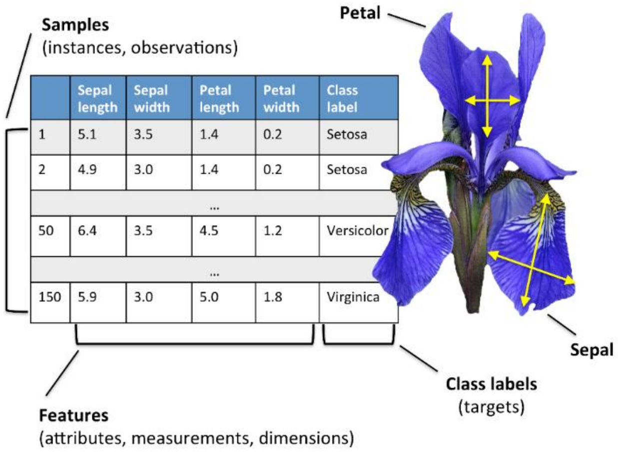


Input → Feature extraction → Classifi cation → Output

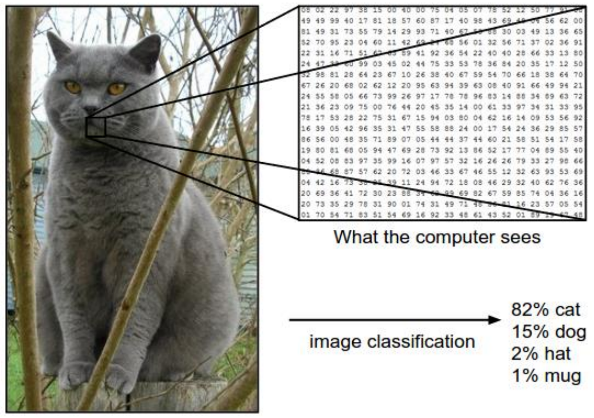
Deep Learning (Neural networks)



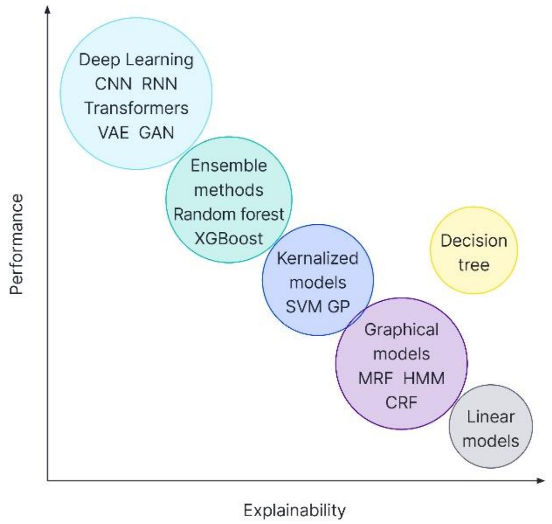
Input → Feature extraction + Classification → Output

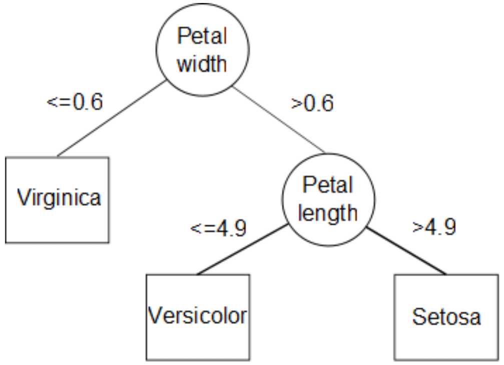


	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			...		
50	6.4	3.5	4.5	1.2	Versicolor
			...		
150	5.9	3.0	5.0	1.8	Virginica



Explainability vs. Performance

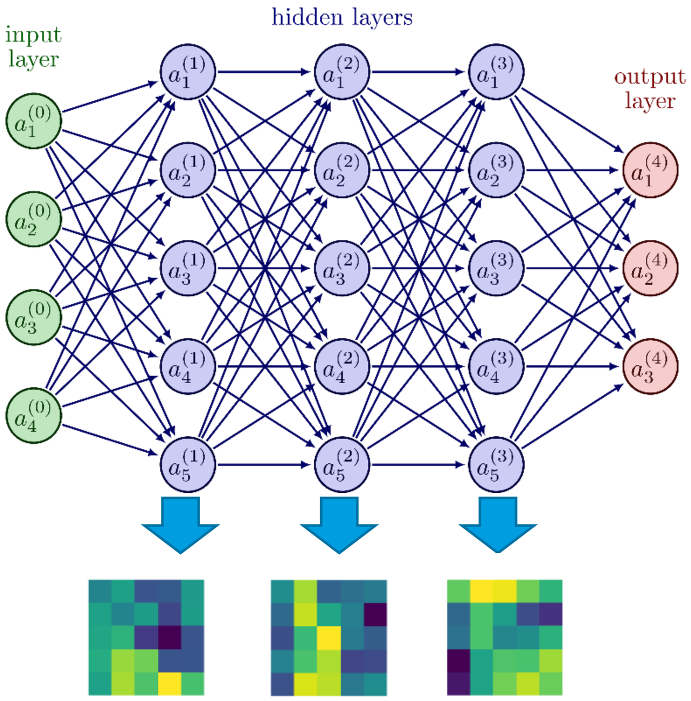




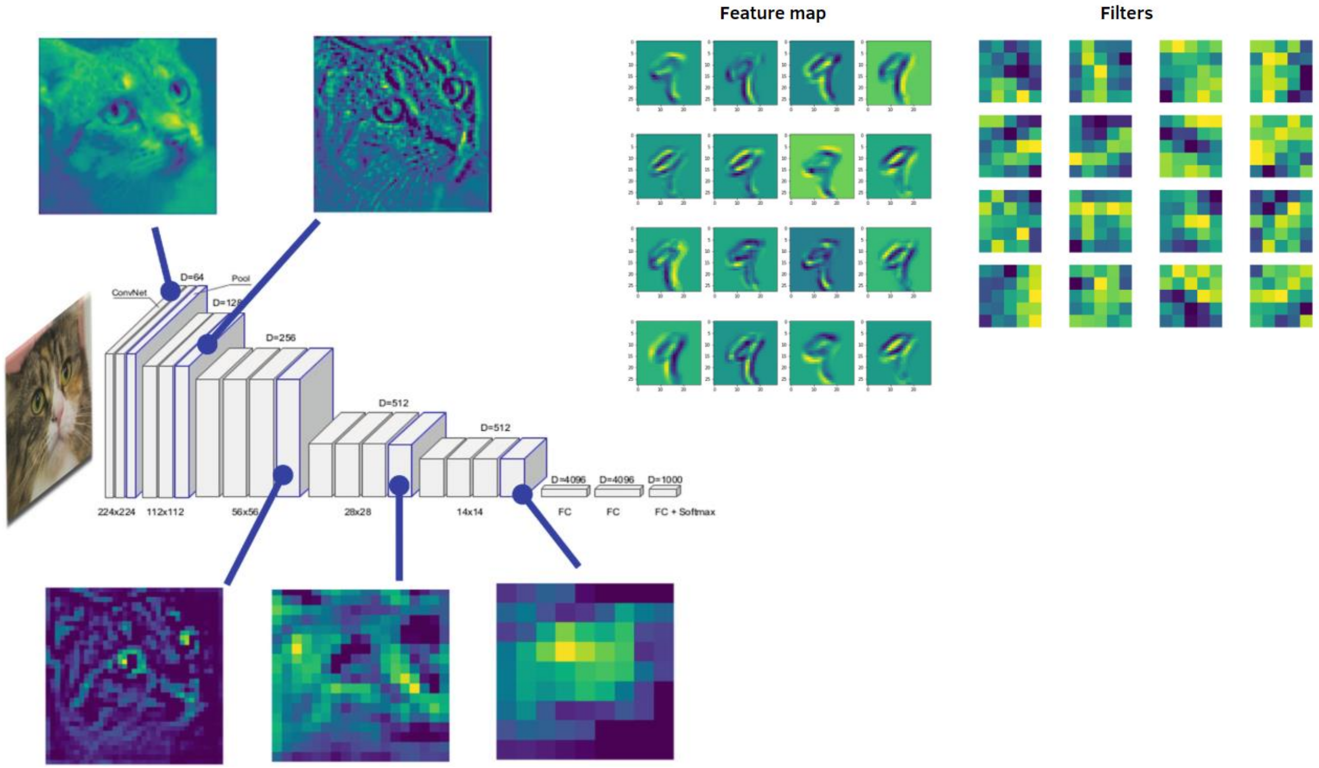
Classic Decision Tree

Classic **machine learning** classification approaches where fundamentally **explainable** in some sense since these model algorithms were already understandable by default. Unfortunately, these methods lack on the **performance** side compared to modern **Deep Learning** based approaches.

Deep Learning Features



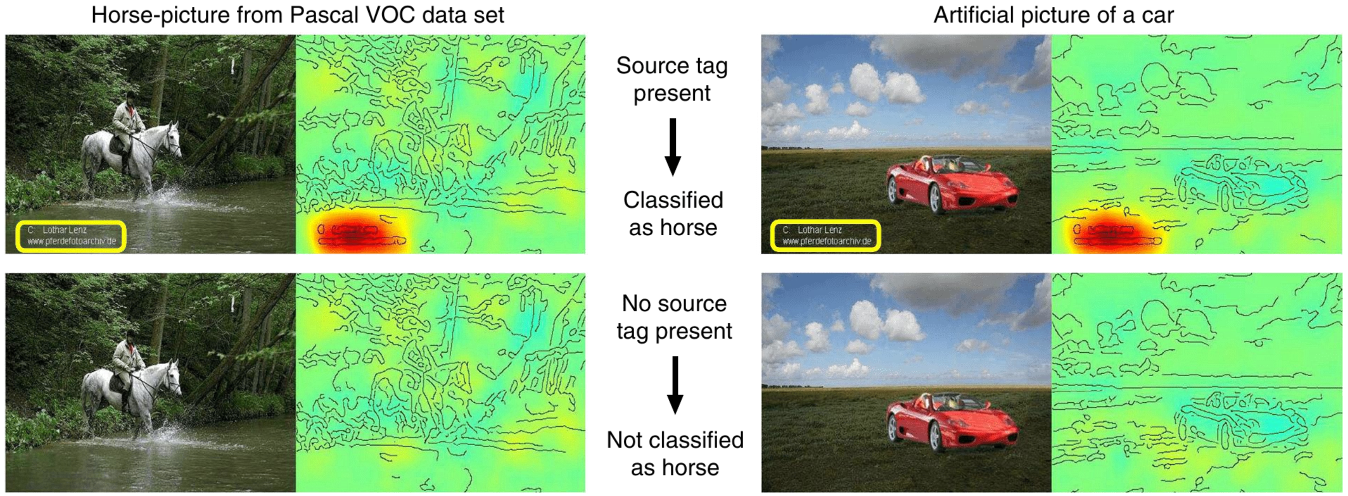
In MLP the extracted **features** are stored across the **weights.**



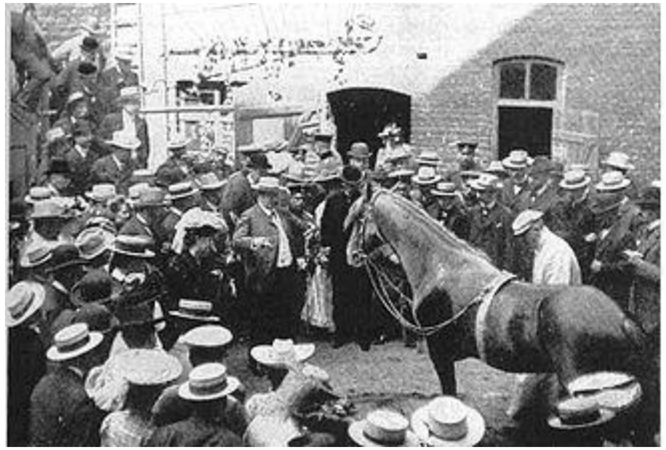
In CNN the extracted **features** stored across the **filters.** Similar to the MLP case this leads to layer intermediate outputs (feature maps) of different shape.

Why do we need Explainable AI? (XAI)

Unmasking Clever Hans Predictors



Trivia: The clever Hans



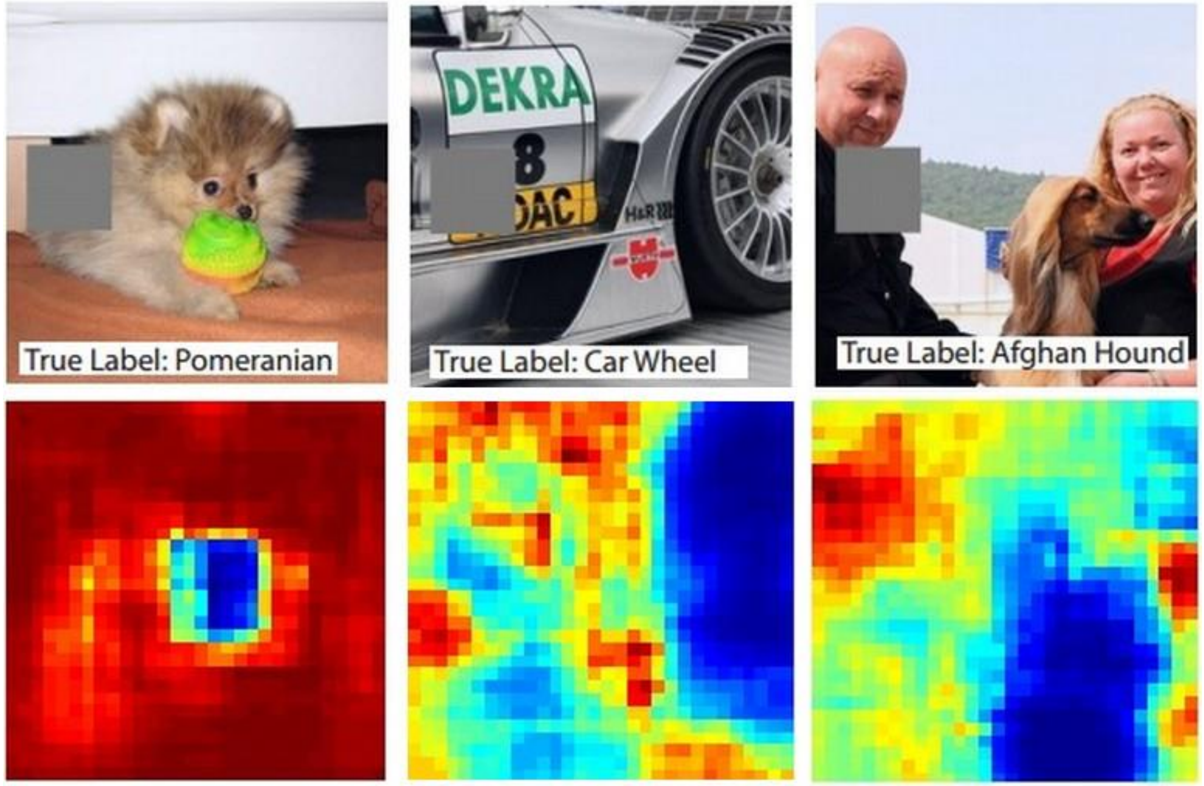
Hans was a horse that was claimed to have performed arithmetic and other intellectual tasks was actually responding directly to involuntary cues in the body language of the human trainer

DL-Algorithms are supposed to find the **easiest solution** to a given **problem**. In the case of a **Clever Hans predictor**, the algorithm utilizes information **given by mistake** to skip learning the underlying (hard) problem. Instead it develops a surprisingly trivial solution. This approach is considered as **“cheating”** since if the information is taken away the models **true performance** is still close to **random guessing**.

*Lapuschkin, S., Wäldchen, S., Binder, A. *et al.* Unmasking Clever Hans predictors and assessing what machines really learn. *Nat Commun* **10,** 1096 (2019). https://doi.org/10.1038/s41467-019-08987-4

Perturbation based Methods
Feature Visualization by Occlusion

Matthew Zeiler:
Visualizing and Understanding Convolutional Networks (2013)



What part of an Image is **relevant** to a Neural Networks decision?

A native approach to this is the **occlusion** of specific **pixels** or **regions** in an Image.

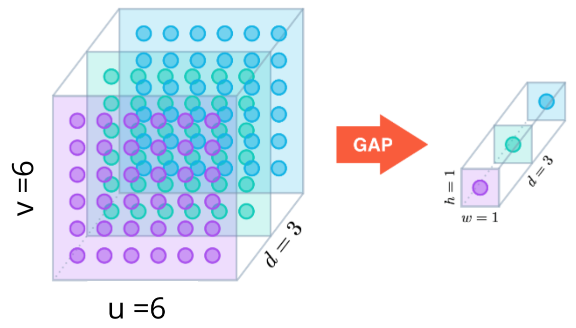
Here we **iterate** over regions of the image, set a patch of the image to be all **zero**, and look at the **probability** of the class.

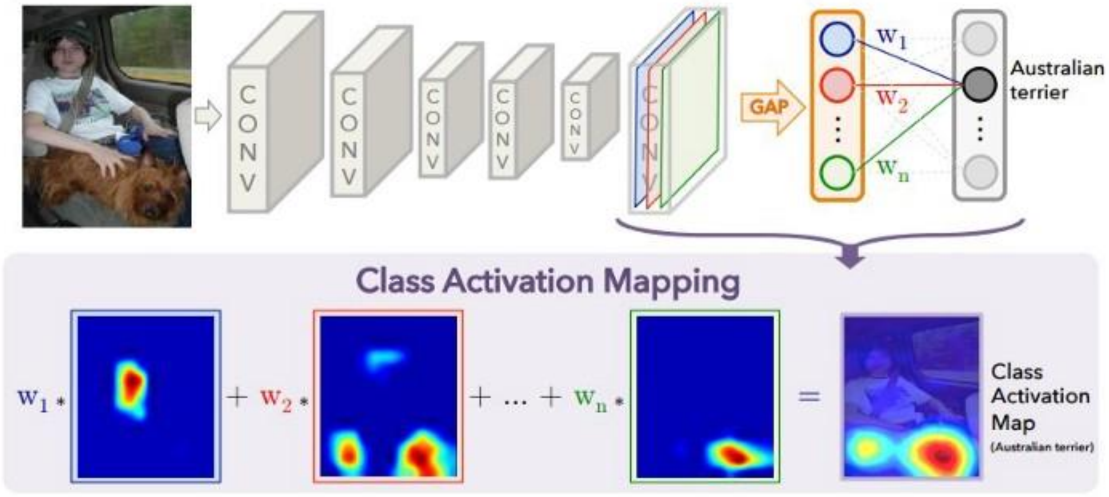
Backpropagation based Methods
CAM and GradCam

ML-Research has shown that convolutional feature maps **retain** spatial information, which is **lost** in fully-connected layers (MLP). **Last Conv Layer** can be thought as the important **features** for the classification.

→ Classificate from these feature Maps after using **GAP** and **sum** the **weighted positive** feature Maps = CAM (make them positive by ReLu(x))
→ Compute the GAP pooled gradient of the last layer

Global Average Pooling (GAP)





Number of FMaps

Weight (Cam)

$$L^c_{Grad-CAM} \sim= \sum_{k=1}^{K} \alpha^c_k A^k$$

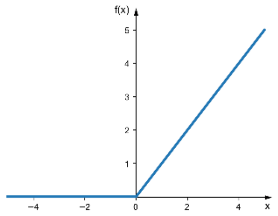
= **CAM**

Feature Map

$$\alpha^c_k = \frac{1}{uv} \sum_{i=1}^{u} \sum_{j=1}^{v} \frac{dy^c}{dA^k_{i,j}}$$

GAP of Gradient (GradCam)

ReLu

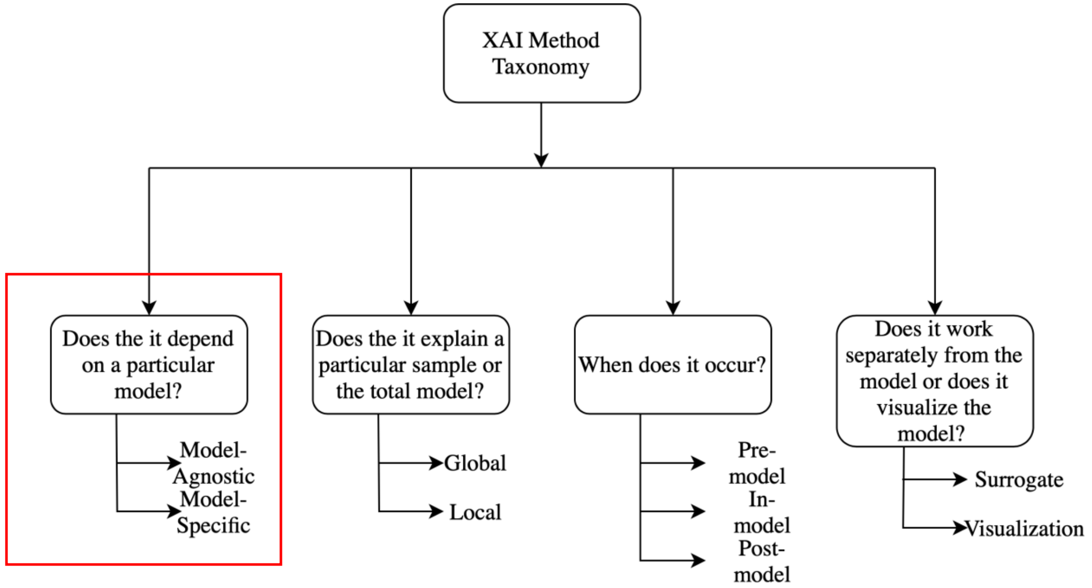


$$L^c_{Grad-CAM} = ReLU\left(\sum_{k=1}^{K} \alpha^c_k A^k\right)$$

= **GradCAM**

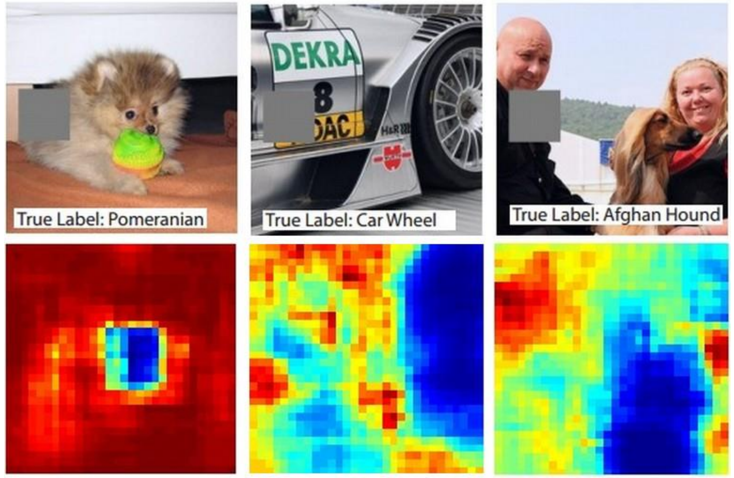
ReLu because we are only interested in the **features** that have a **positive influence** on the class of interest

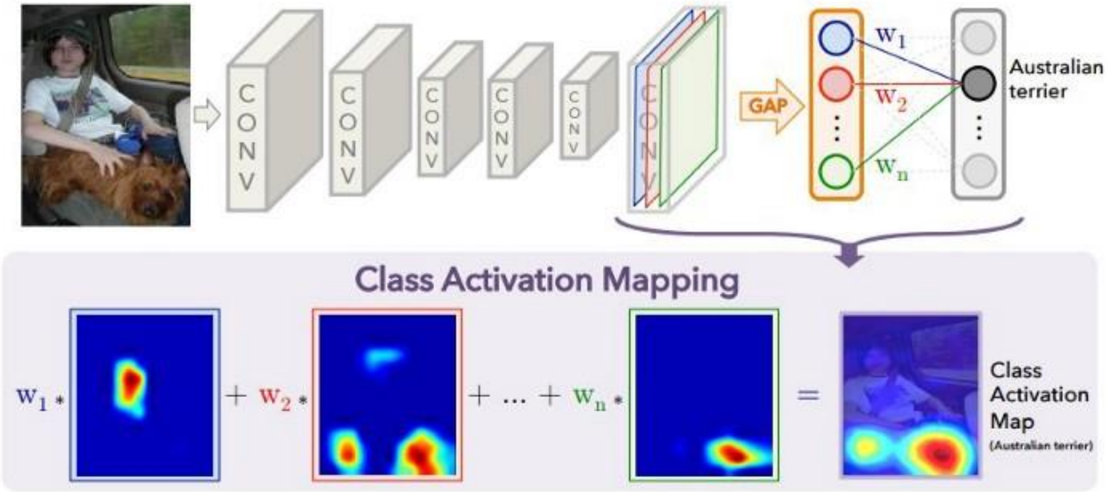
Taxonomy
Agnostic vs. Specific



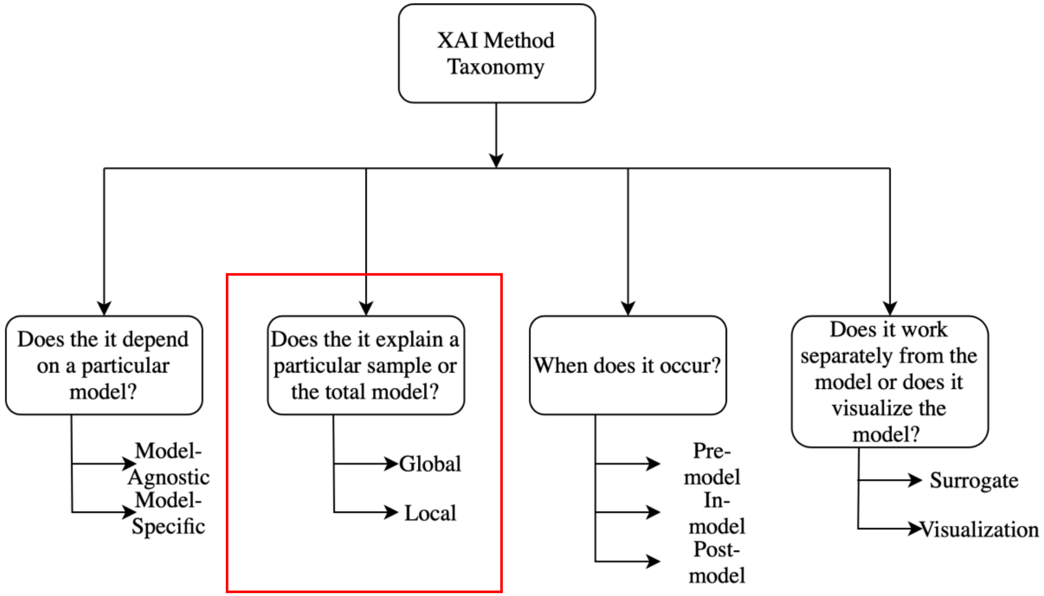
Model **agnostic** methods like the previously seen **occlusion** method by Zeiler and Fergus do **not require** a certain **model type** to work. These methods do **not** have direct access to the **internal model weights** or structural parameters.

Model **specific** interpretation methods (e.g. **GradCAM**) are based on the parameters of the individual models.





Taxonomy
Global vs. Local

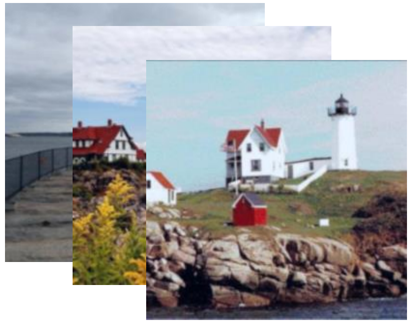


Local explainability

Model Output:

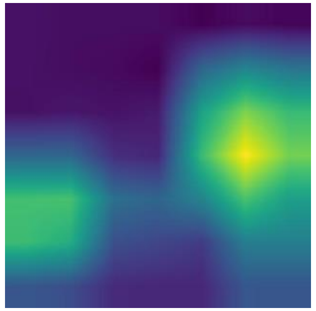
„For the classification of this lighthouse I used the following input features“

Global Explainability



Model Output:

„For the classification of all of the lighthouses I mainly used the following input features“



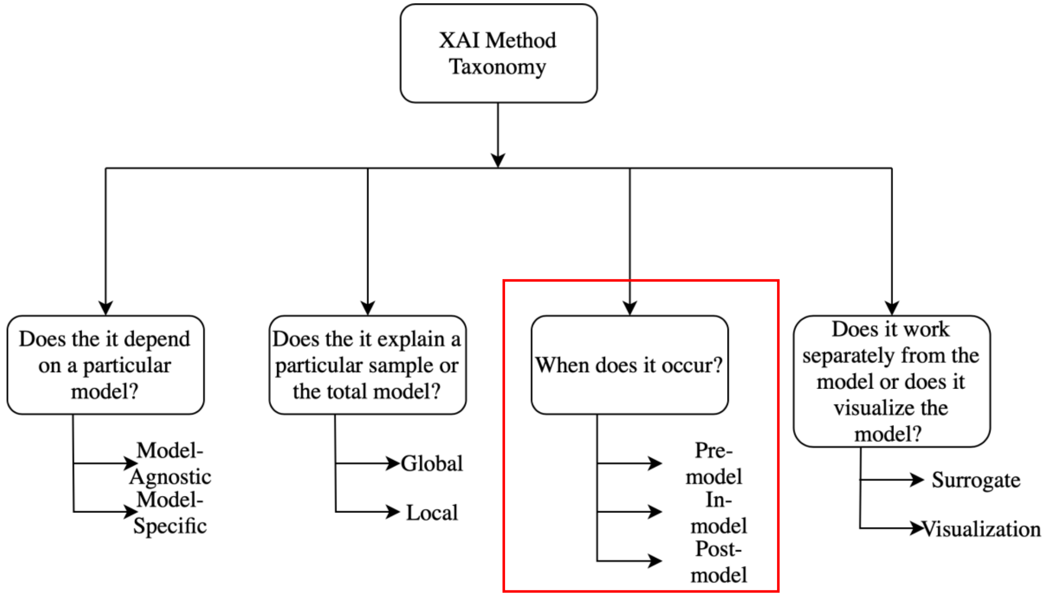
Taxonomy
Pre- vs In- vs Post-hoc

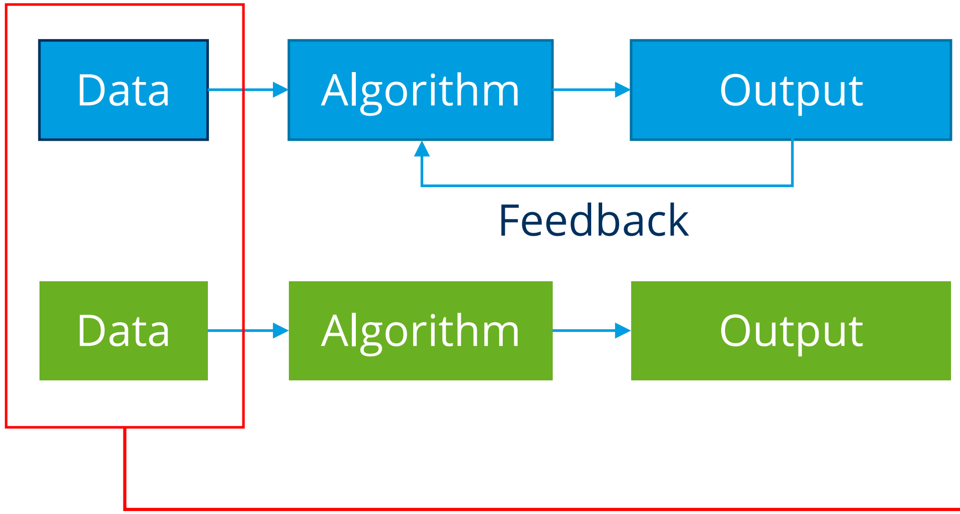
Train

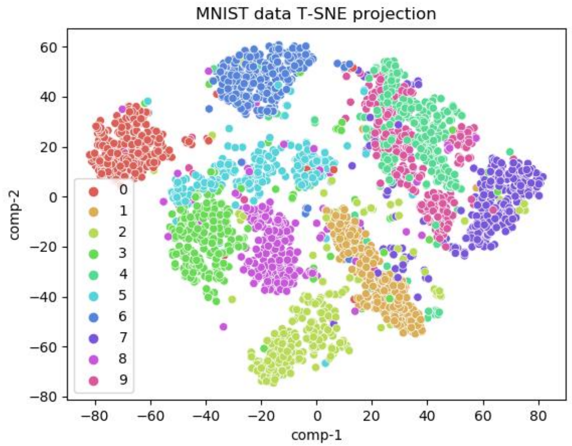
Test

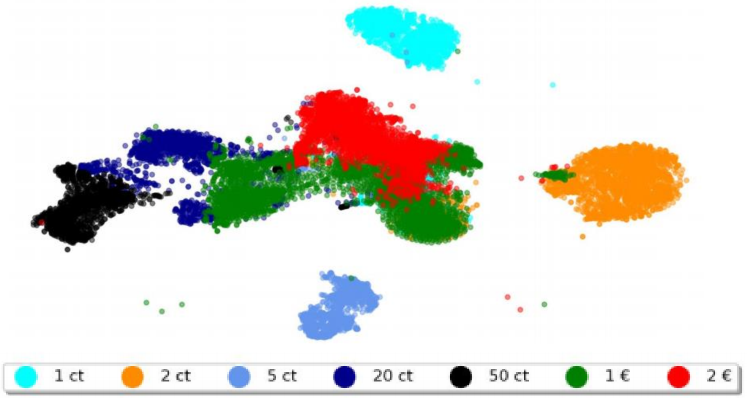
Pre-model methods

Pre-model methods are independent and does not depend on a particular model architecture to use it on. They are applied **pre training** to **explain** more the **data** then the actual model itself.

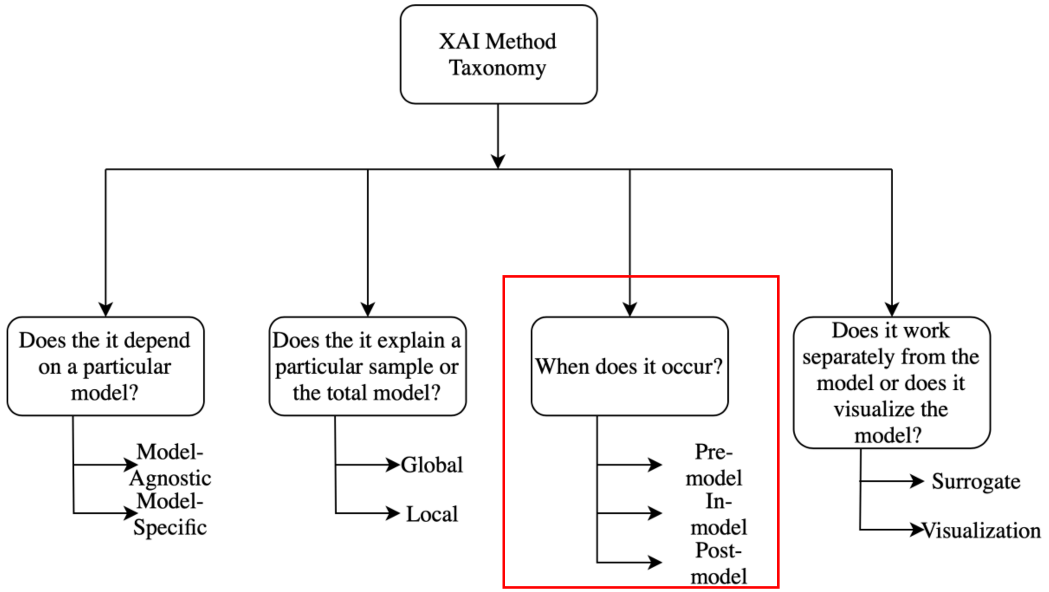






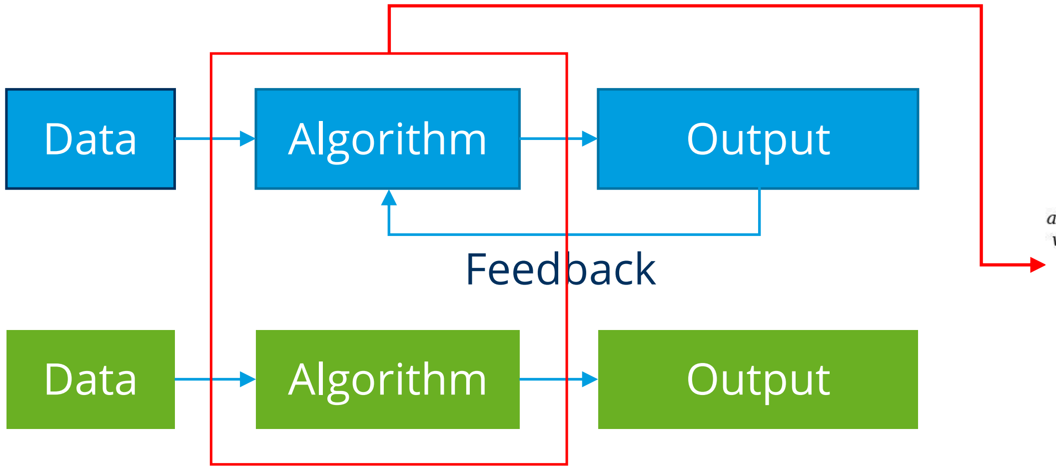


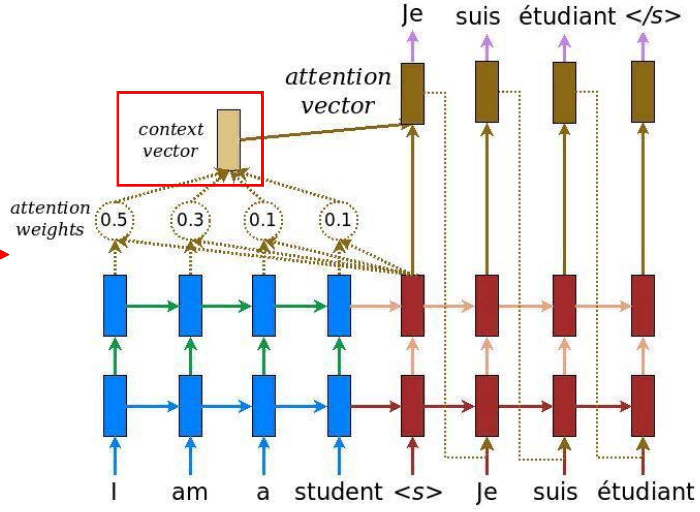
Taxonomy
Pre- vs In- vs Post-hoc



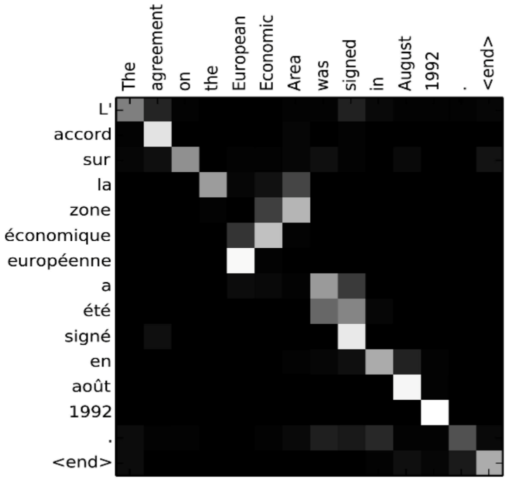
In-model methods

In model methods are created **while training** the algorithm itself. They are often a **side product** of the **models structure**. They can be accessed while inference.





Model



Plotted Context Vector

Taxonomy
Pre- vs In- vs Post-hoc

Train

Test

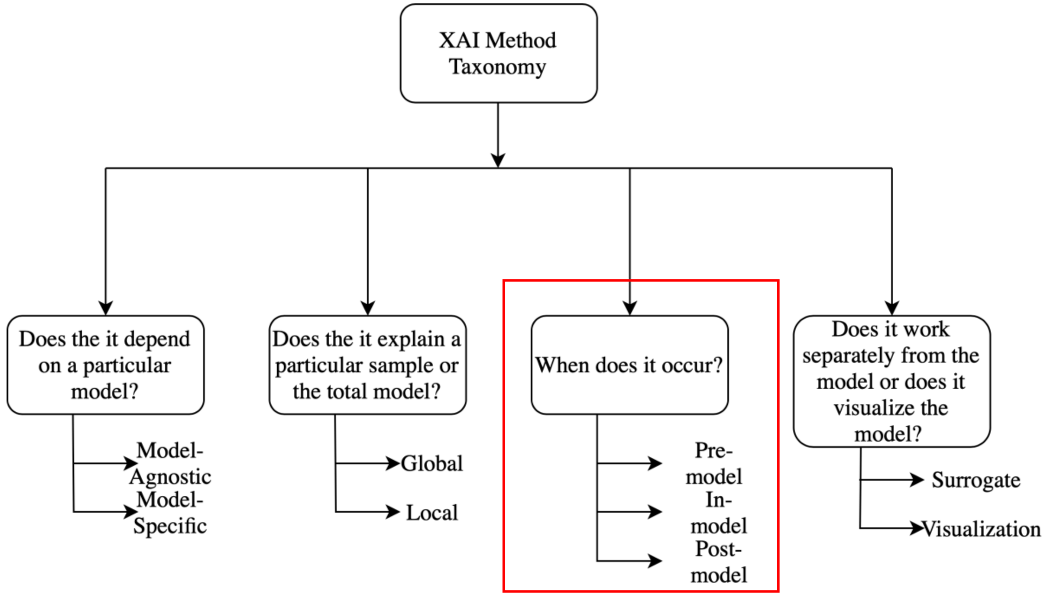
Post-hoc methods

Post-hoc methods are applied after the training process of the model. They can use the model at inference to create meaningful insights about what it might has learned.

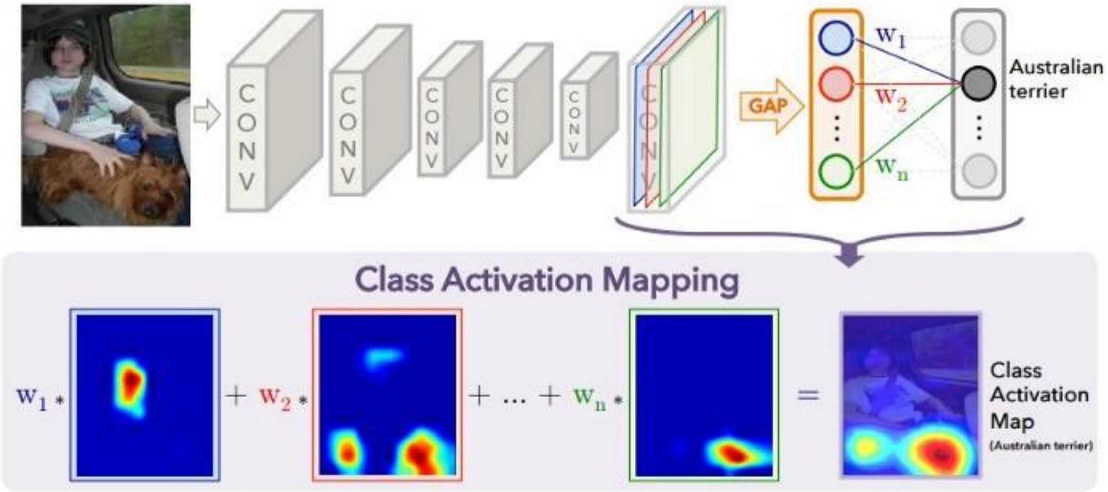
Data → Algorithm → Output

Feedback

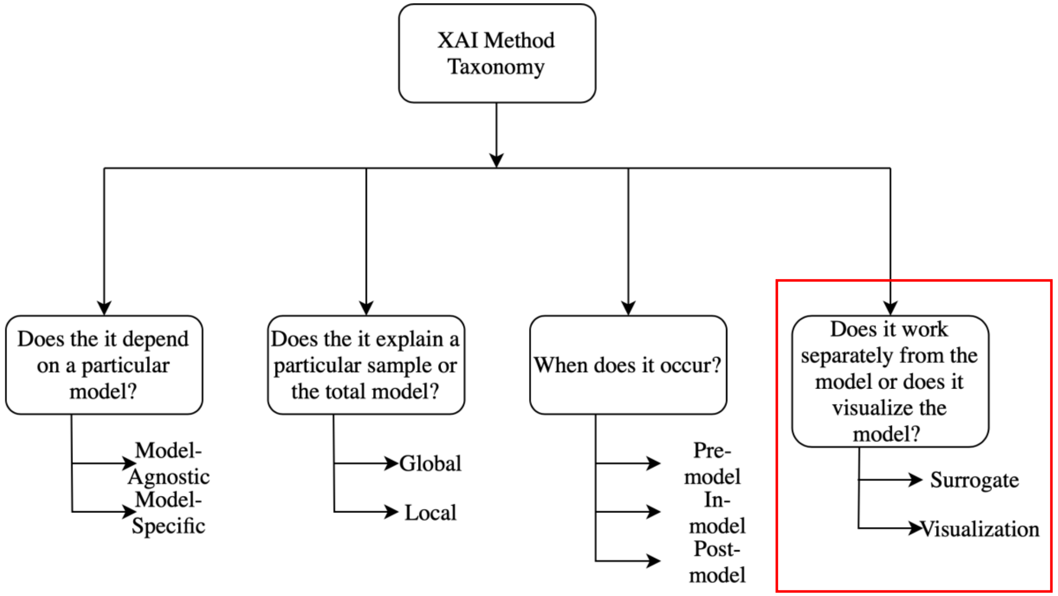
Data → Algorithm → Output



Example: GradCAM



Taxonomy
Surrogate vs. Visualization

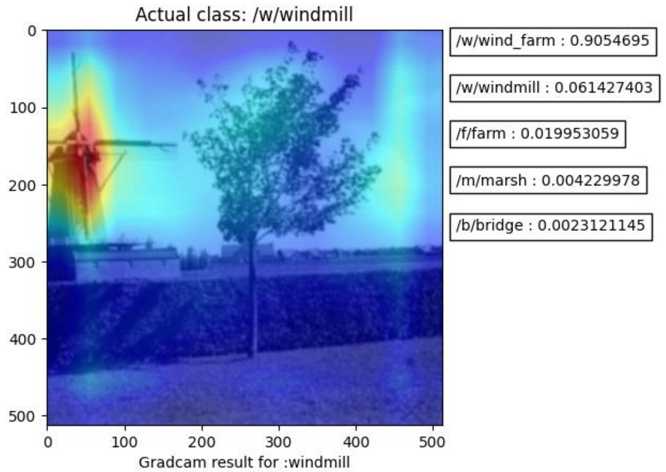


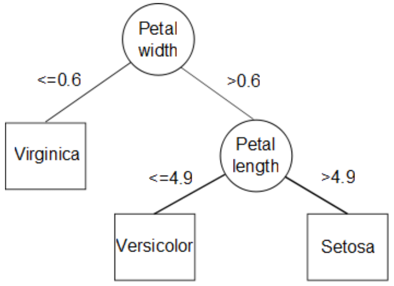
Surrogate

The idea is that we take our “black box” model and **create predictions** using it. Then we train a **transparent surrogate** model on the predictions **produced** by the “black box” model and **compare** the black-box model’s decision and surrogate model’s decision.

Visualization

Visualization methods are not a different model, but it helps to explain some parts of the models by visual understanding like activation maps or GradCAM images.





Classic Decision Tree

XAI State of the Art

How do we measure the quality of any XAI method?

There are **many XAI methods** out there e.g. Lime, GradCAM, LRP, SHAP… just to name a few. But which is the **best**?

This is what I see quite regularly. →

Authors often **evaluate** XAI methods based on a **“visual proof”.** A **non-subjective** comparison method similar to the accuracy for image classification is needed!

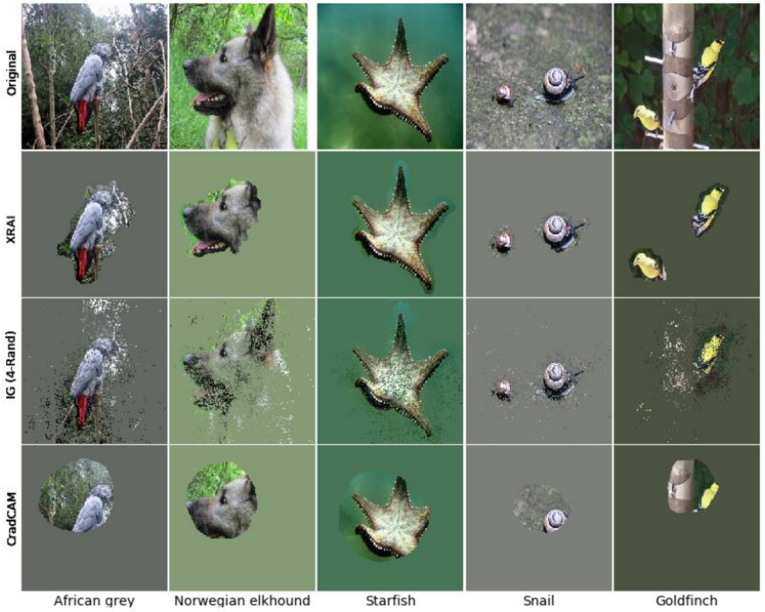


Figure 8. XRAI (2nd row) compared to Integrated Gradients with random baselines (3rd row) and GradCam (bottom row). GradCam can produce blobby regions, whereas XRAI tend to create regions tightly bound around identified objects.

Alternative: Occlude unimportant regions based on XAI

Data

Algorithm

Lighthouse!
(Acc 99%)

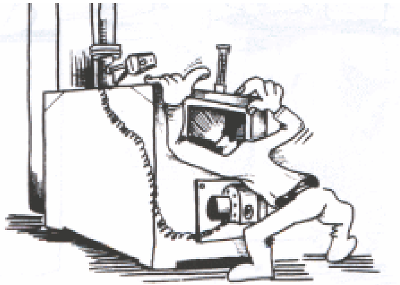
After XAI

Algorithm

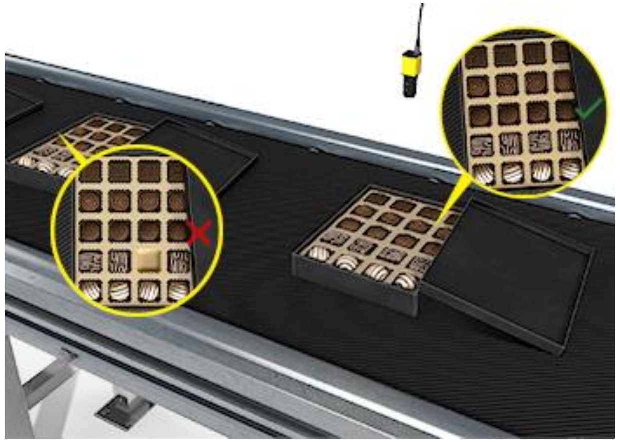
Lighthouse!
(Acc 97%)

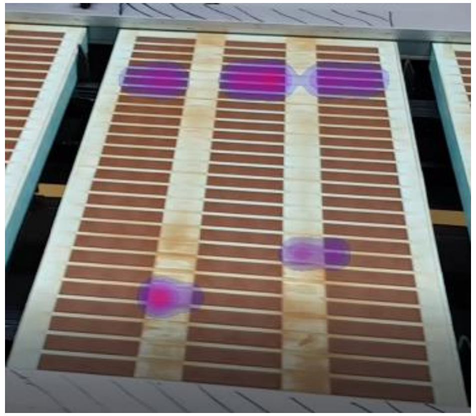
Downside: It’s an evaluation based on the same model that led to the XAI input. It is therefore **not** independent truly **independent**.
→ External grading would be useful

Explainable artificial intelligence for fault diagnosis:
Impacts on human diagnostic processes and performance

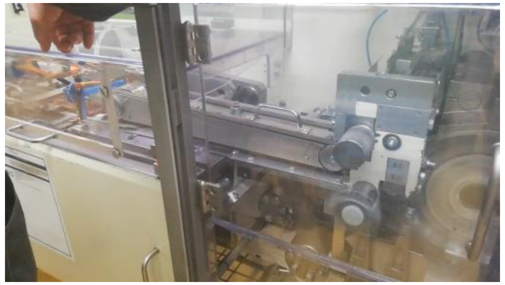


Problem context: chocolate moulding





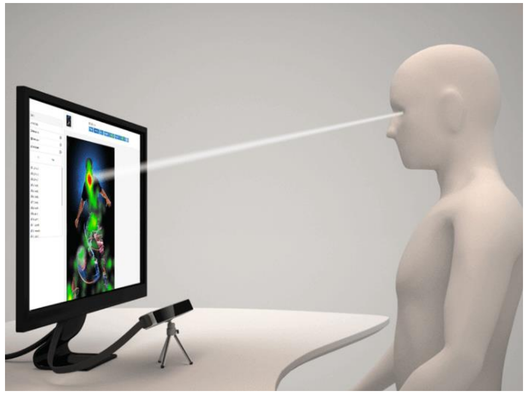
Task: implement XAI algorithms to **detect** process **deviations** and **explain** network **decisions** to the operator. In this context we will conduct studies on the human troubleshooting **speed** and the **desicion acceptability** related to the **operator XAI interaction.**

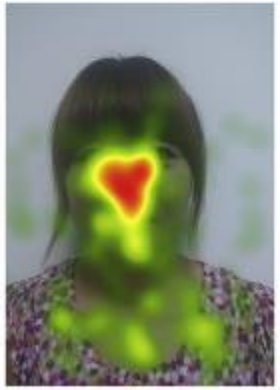


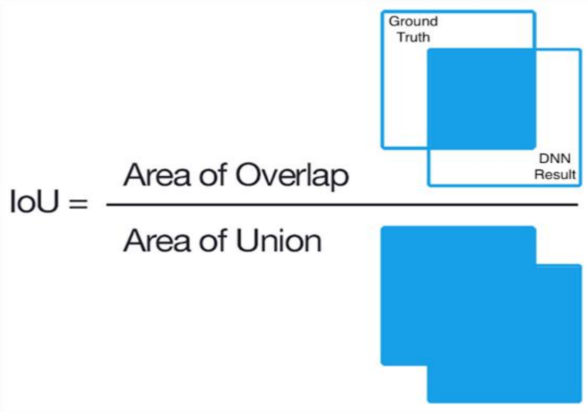
XAI-Dia Experiments

We will conduct studies to **compare human affection** and **XAI** based explainations. We aim to validate different XAI methods based on the intersection (e.g. IoU) between **human gaze** based **heatmaps** and **XAI visualisations**.

Unfortunately, we still need to **wait** for the labeled chocolate data. Thus we will perform pre-studies on the **Places365 Dataset**. (e.g. study different XAI behaviour).



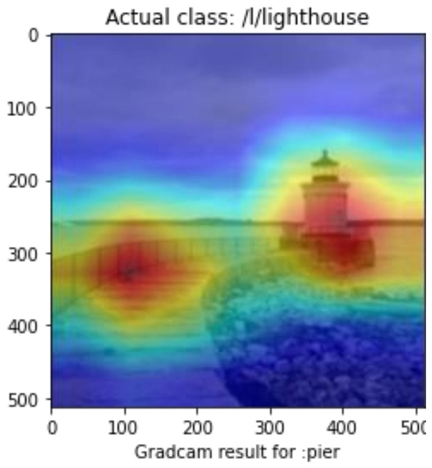




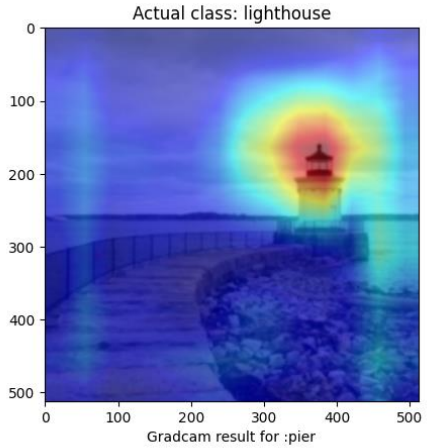
Model Specific Performance
Example: GradCAM

Model dependent algorithms can have **different results** even if asked for the same explaination.

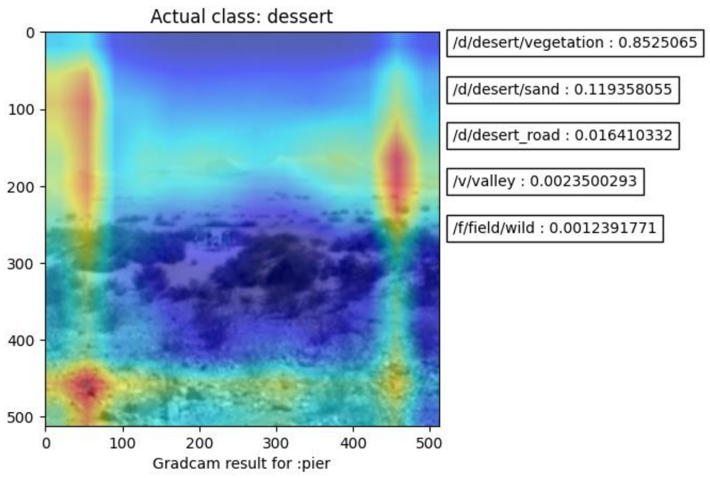
Resnet152



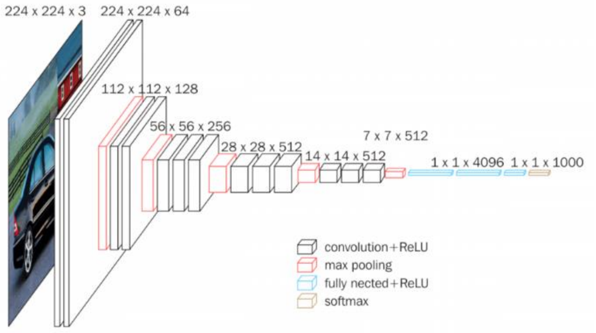
VGG 16



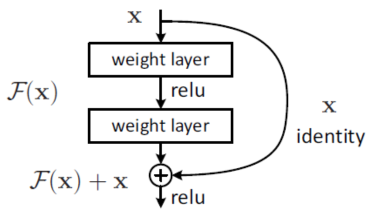
Can suffer from heavy artefacts

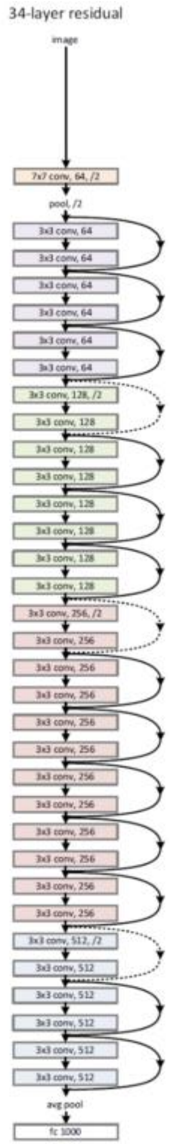


VGG 16



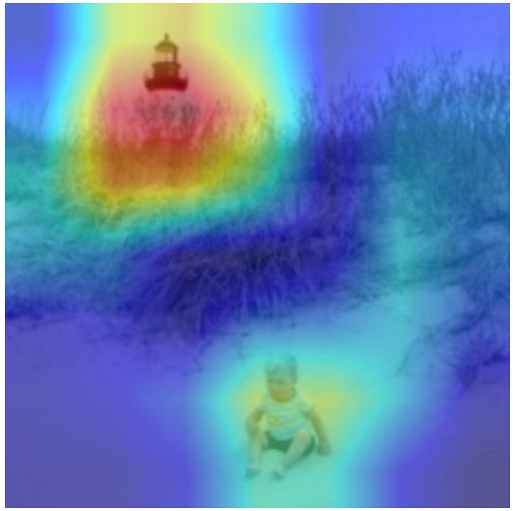
Resnet152

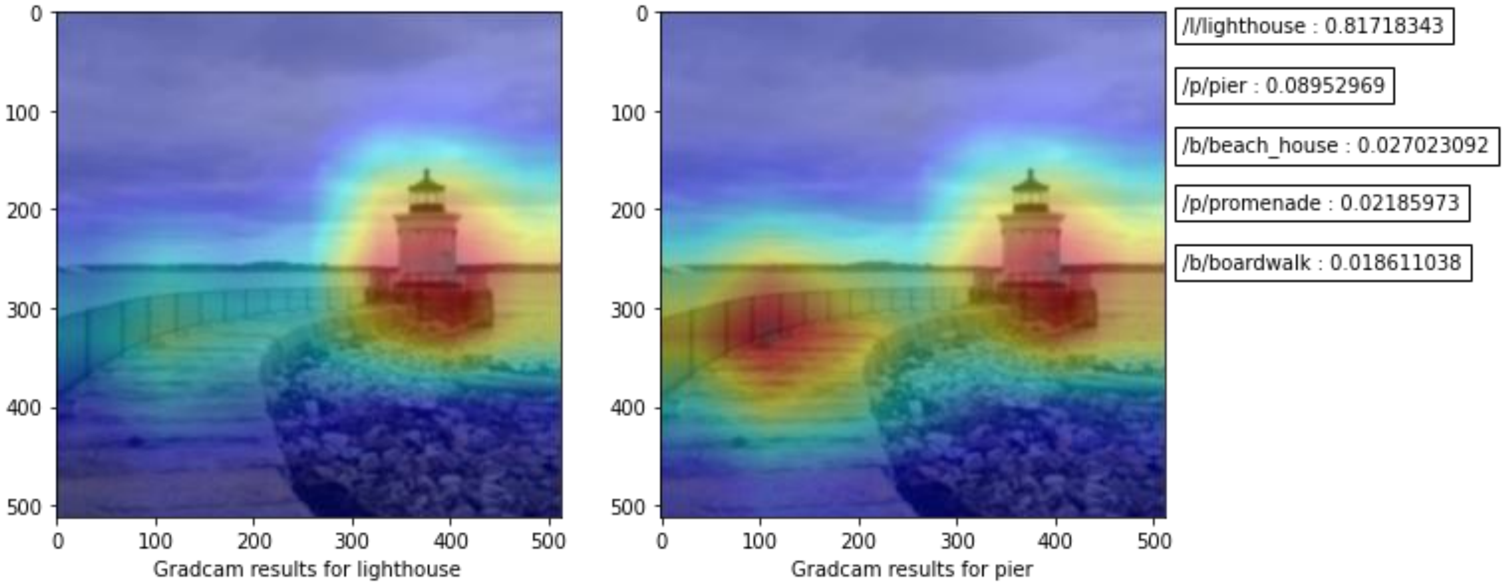




Combinations of model & XAI method can **suffer from artefacts** that can overrule XAI based explanations **despite** of **high accuracy** in the models prediction.

Salient Object Focus



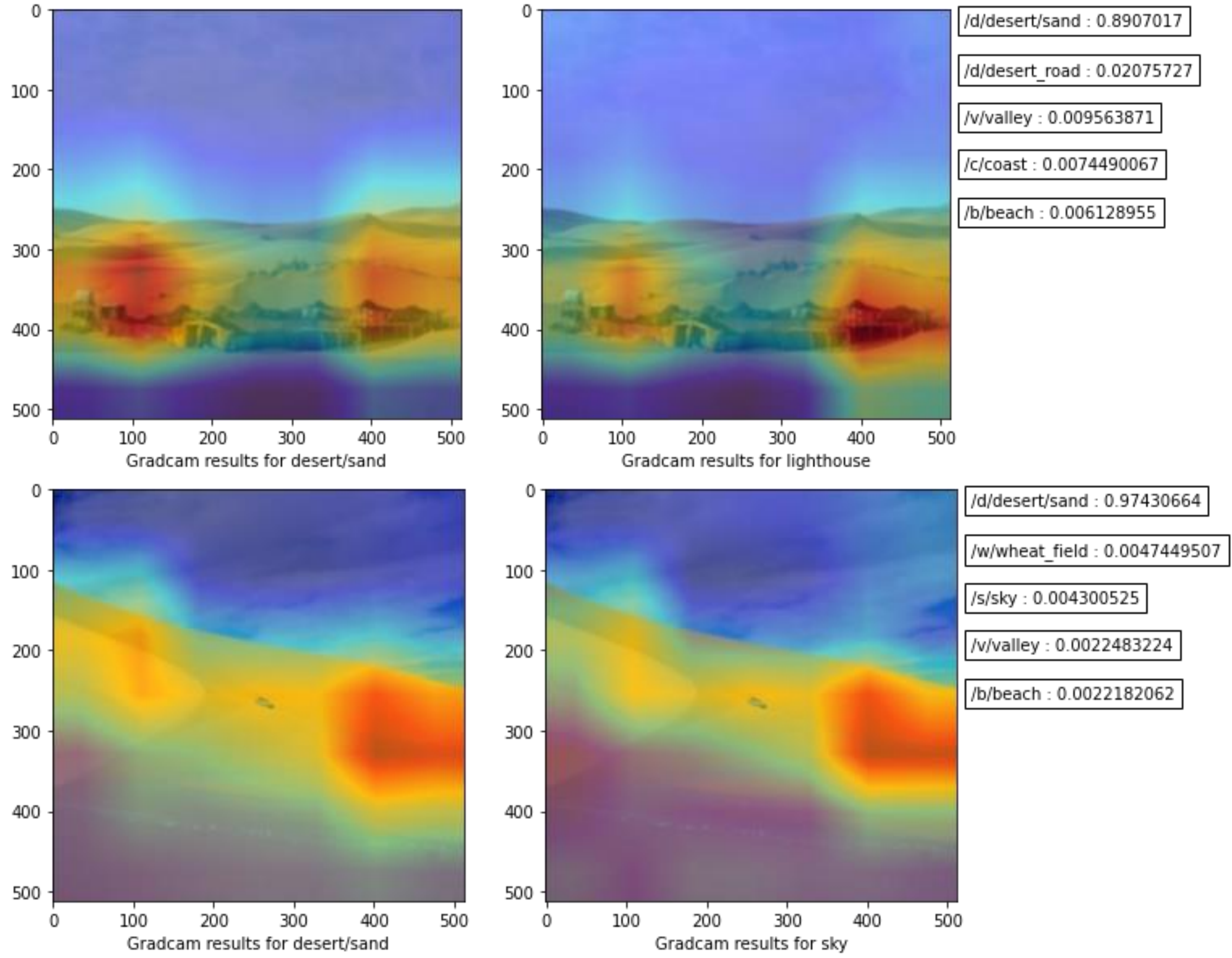


The results and the explanations for some classes **focus** on so called **salient objects** even if they are **not asked** for the **salient class explanation** context. This is a known **problem** also in **human gaze** based heatmapping.

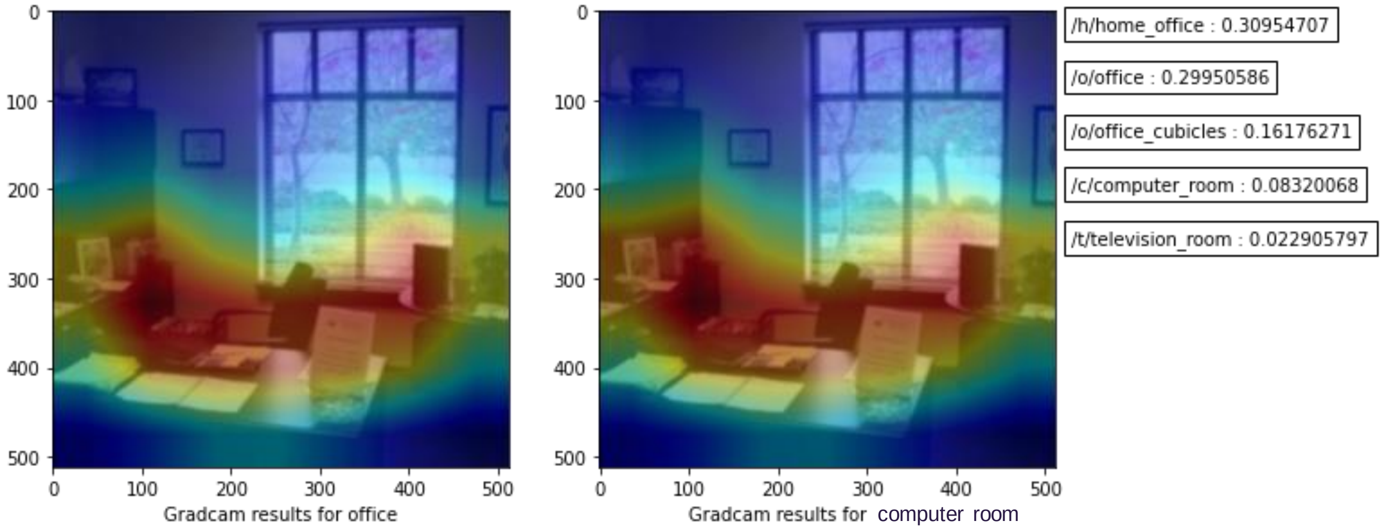
Class Dominance for Images without Salient Objects

The explanations for some classes **focus** on **specific regions** of the dominant class. This happens if the class is not present in the picture.

This also seems to happen if the class (here it is sky) is also **present but non-dominant** in the image.



Similiar Class Overlap



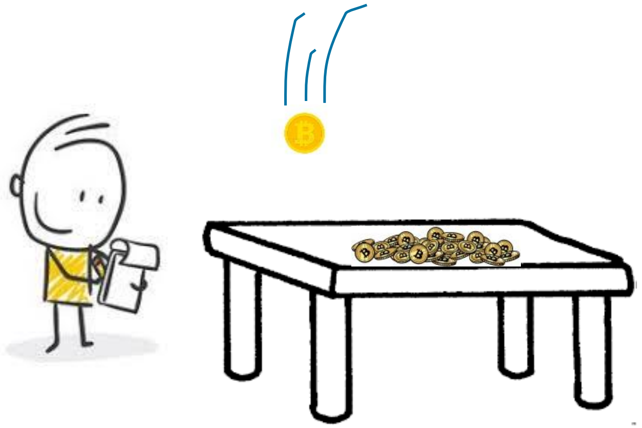
If **classes** are too **similar**, the XAI visualizations of different objectives merge. In these cases the model explanations seems to see no difference between these classes despite a **computer room** can look **very different** from an **office**. (This could also be a problem of a dominant class, but we are not sure yet)

Thank you for your attention!

I am happy to answer questions.

Backup

Sensor Raw Data Monitoring
Classification of Ultrasound

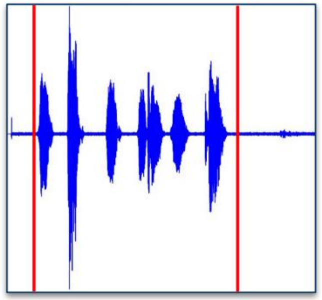


In **2016**, I conducted an **entry level study** for myself**:**

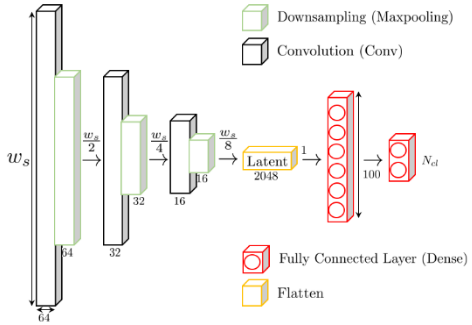
Recognize the **coin** based on **raw** sound data!

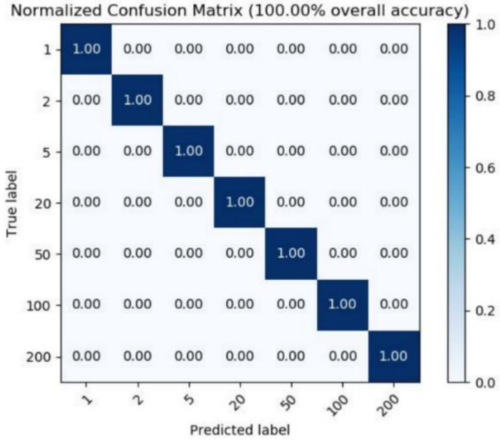
Euro Coins

Ultrasound **signal**



(Basic) 1D-CNN **classifier**





The classifier has been **integrated** into the Sonotec device and **presented at CeBit** in 2018 (advertisement)

<u>**Triva:**</u> All of my students have to **re-do** it today … instead of MNIST.

Erklärbarkeit von Klassifikationen
Erste Versuche - Münzklassifikation

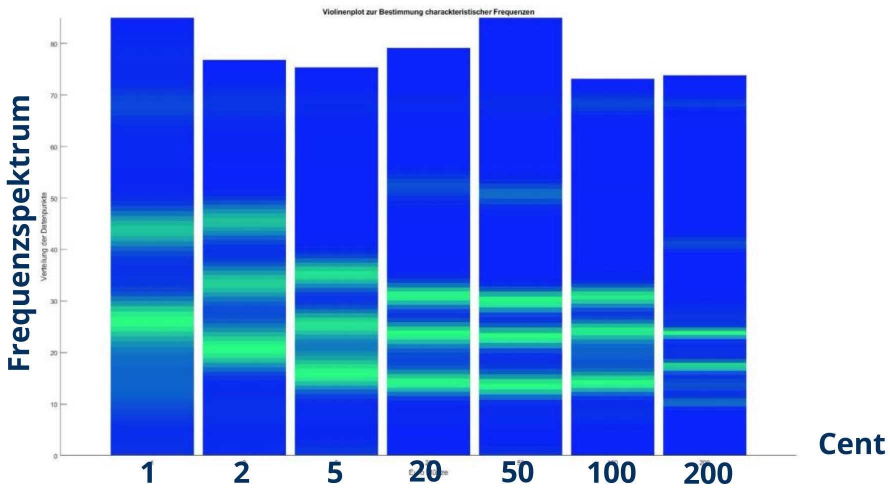
Die bereits erwähnte Münzklassifikation lässt sich auch über ein klassisches Merkmal realisieren das auf der Erkennung der **charackteristischen Frequenzen** beim Aufprall beruht.

Zur Analyse was das Netzwerk aus den Sensordaten (ohne dieses Vorwissen) **„gelernt"** hat wurde **t-SNE** (t-distributed stochastic neighbor embedding), ein Verfahren zur Dimensionsreduktion (ähnlich PCA) eingesetzt.

Die Überlappung von Bereichen deuten auf entdeckte **Gemeinsamkeiten** bei der Klassifikation verschiedener Münzen hin. Diese Gemeinsamkeiten sind **identisch** zu denen der charackteristischen Frequenzen.

→ Das Netzwerk **muss** die charackteristischen Frequenzen bestimmt haben.

Charackteristische Frequenzen



t-SNE des „latent Layers" des Netzwerks