UiT The Arctic University of Norway

Explainable machine learning

From scalars to vectors

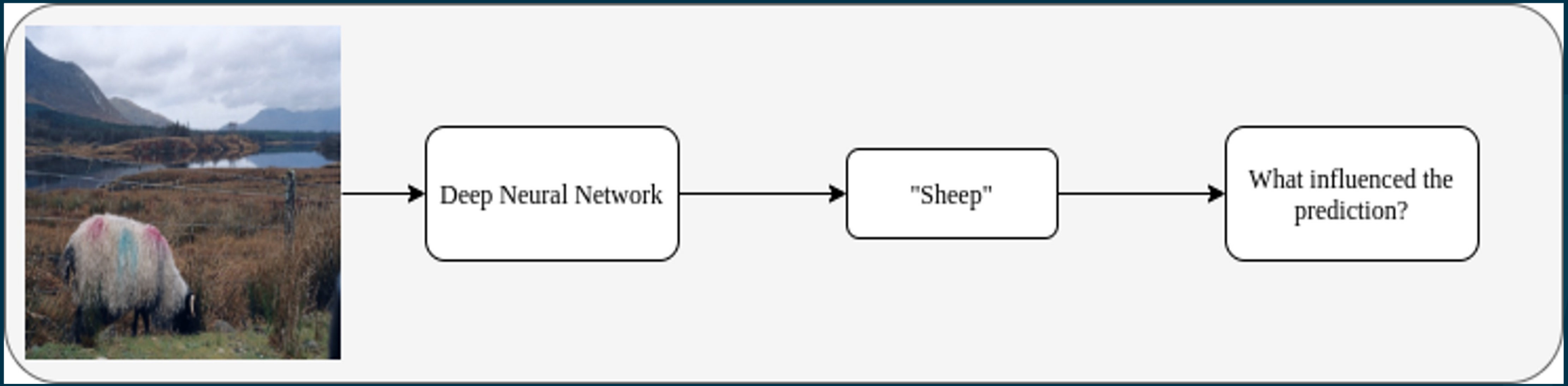
Kristoffer Knutsen Wickstrøm
UiT Machine Learning Group and Visual Intelligence

Schedule

- First lecture – Introduction to explainable artificial intelligence (XAI)
  - Why do we need explainability?
  - How do we get explainability?
  - Challenges in XAI
- Second lecture – XAI in representation learning
  - How to explain vectorial representations of data?
  - Why are standard XAI techniques not suitable.
  - Representation learning explainability with RELAX

What is explainability?

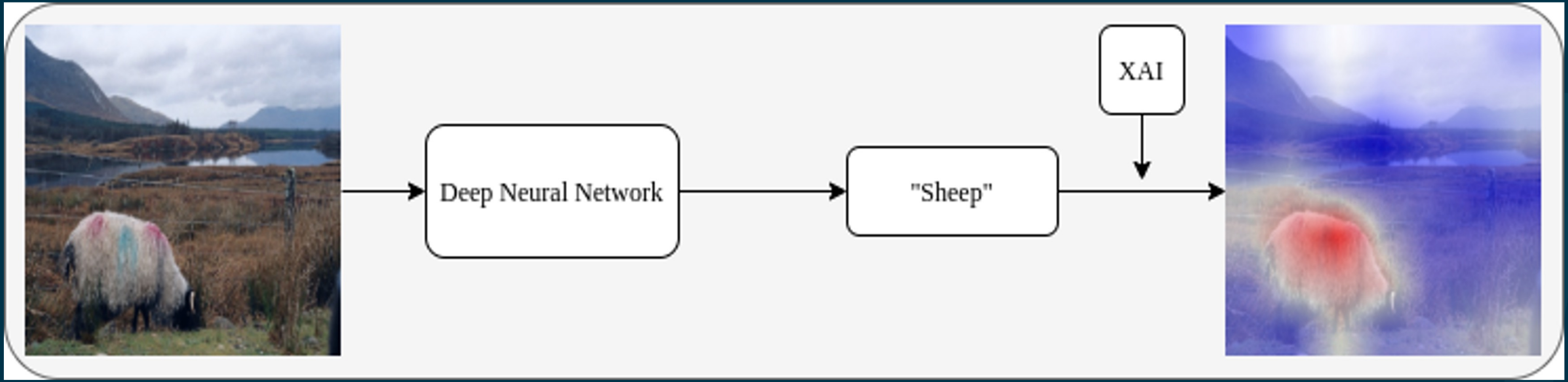
- A tool for answering the question “why?”[1]



[1]*A. Holzinger et al., “Explainable AI methods - a brief overview”. xxAI - Beyond Explainable AI, 2022*

What is explainability?

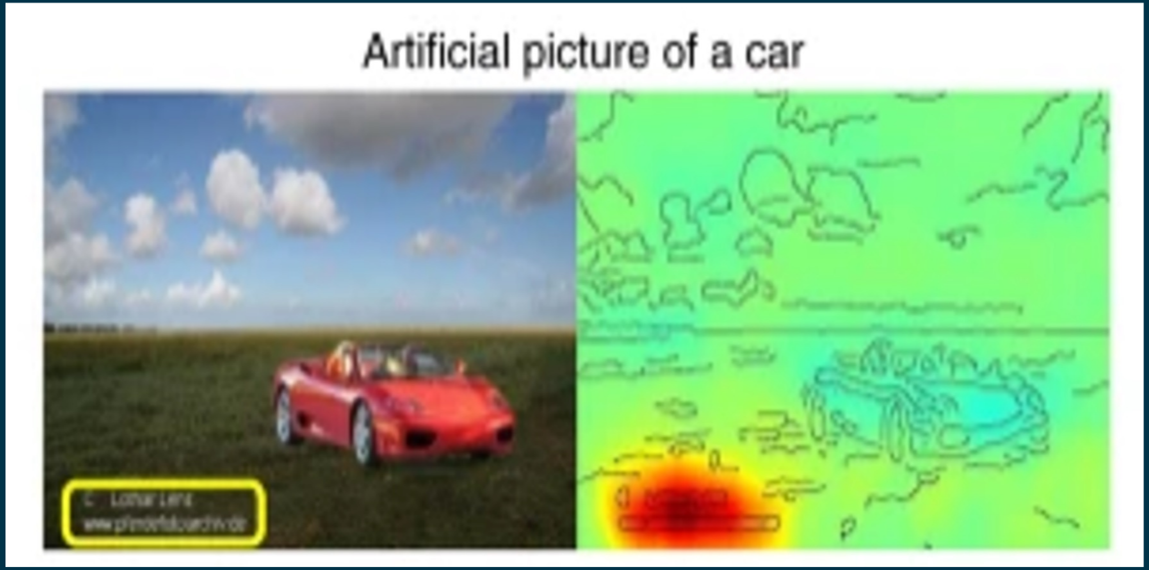
- A tool for answering the question "why?"[1]



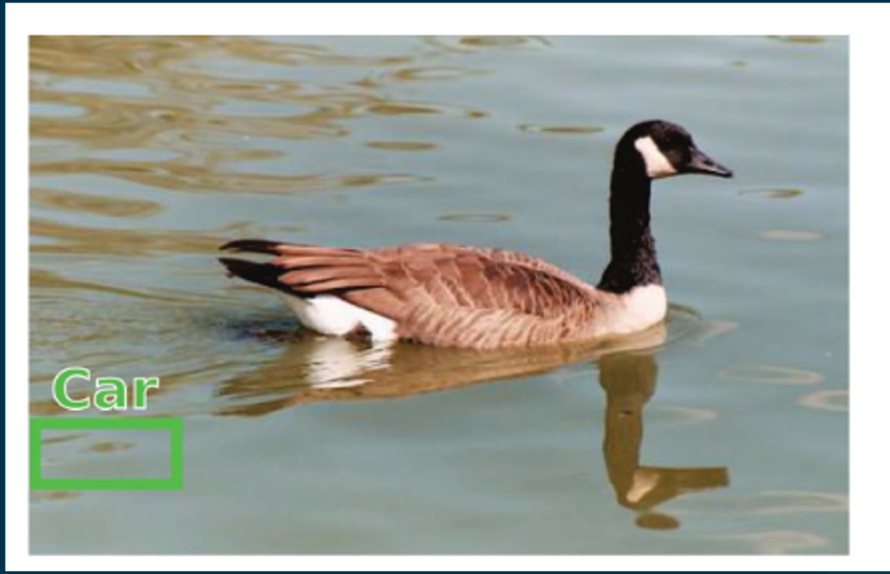
[1]A. Holzinger et al., "Explainable AI methods - a brief overview". xxAI - Beyond Explainable AI, 2022

What is explainability?

- A tool for answering the question “why?”[1]



S. Lapuschkin et.al., 2019[1]

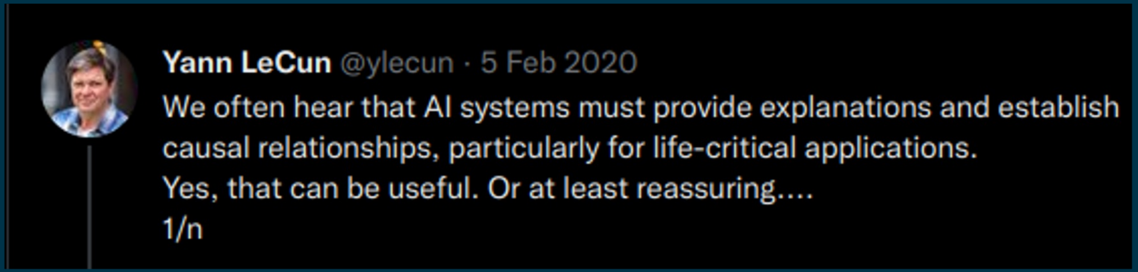


C. Vondrick et.al., 2013[2]

[1]S. Lapuschkin, et al., “*Unmasking Clever Hans predictors and assessing what machines really learn*”. Nature Communications, 2019.
[2]C. Vondrick, et al., “*HOGgles: Visualizing Object Detection Features*”. ICCV, 2013

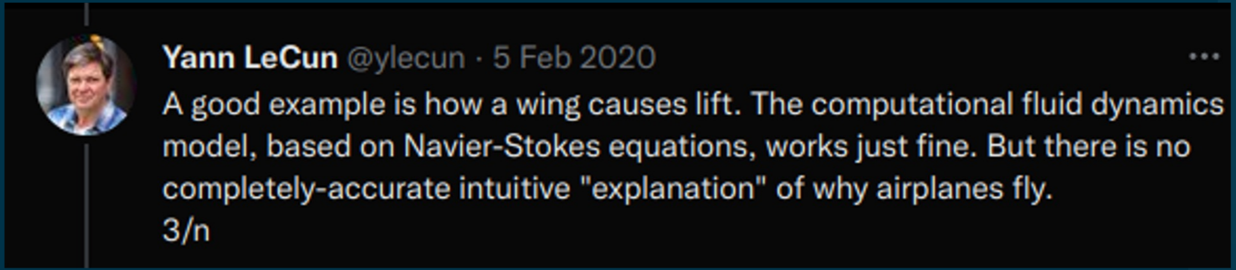
Why do we need explainability?

- Do we need it?
- Many motivating factors[1]:
  - Trust
  - Causality
  - Informativeness
  - Fair and ethical decision making



Yann LeCun @ylecun · 5 Feb 2020

We often hear that AI systems must provide explanations and establish causal relationships, particularly for life-critical applications.
Yes, that can be useful. Or at least reassuring....
1/n

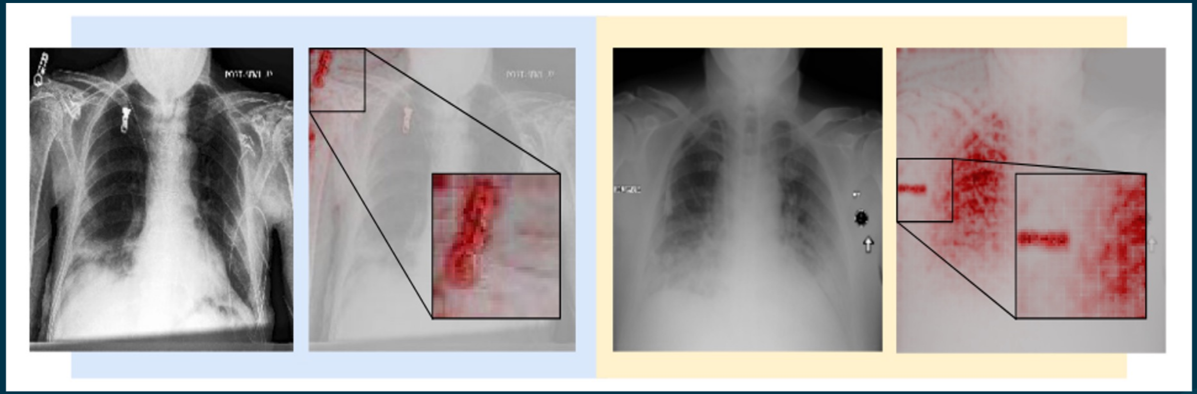


Yann LeCun @ylecun · 5 Feb 2020

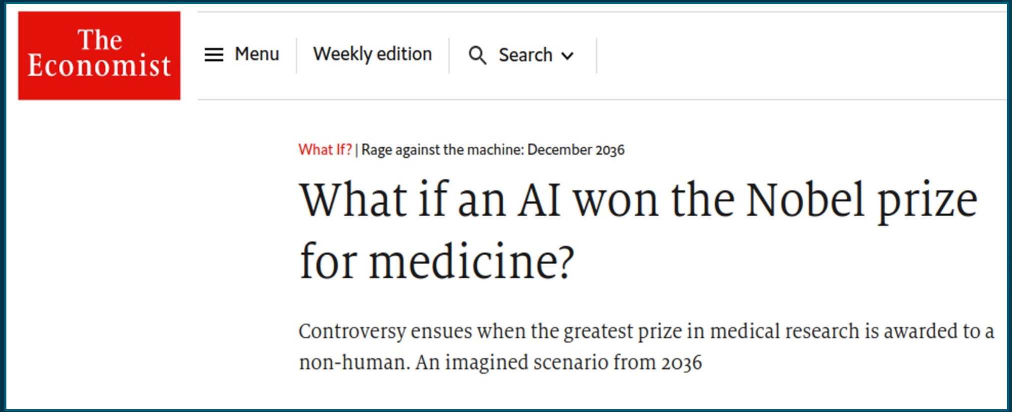
A good example is how a wing causes lift. The computational fluid dynamics model, based on Navier-Stokes equations, works just fine. But there is no completely-accurate intuitive "explanation" of why airplanes fly.
3/n

Why do we need explainability?

- Many motivating factors:
  - Trust
  - Causality
  - Informativeness
  - Fair and ethical decision making



S. Gautam et.al., 2022[2]



What If? | Rage against the machine: December 2036

What if an AI won the Nobel prize for medicine?

Controversy ensues when the greatest prize in medical research is awarded to a non-human. An imagined scenario from 2036

The Economist, 2022

[1]Z. Lipton, "The Mythos of model intepretability". ICML Workshop, 2016.
[2]S. Gautam et al., "*Demonstrating the risk of imbalanced datasets in chest x-ray image-based diagnostics by prototypical relevance propagation*". ISBI, 2022.

Why do we not have explainability?

- Deep learning is the dominant force in contemporary machine learning.
  - Black box models.
- A known issue, but has become more pressing with recent success[1]
- Not limited to deep learning[2]

[1]N.J.S. Morch et al., “Visualization of neural networks using saliency maps”. International Conference on Neural Networks, 1995.
[2]D. Baehrens et al., “How to explain individual classification decisions, JMLR, 2010.

How do we get explainability?

- A vast number of competing methods!
- No clearly superior method
  - More on that later

Table 3 Glossary of Interpretability Methods With Abbreviations Referenced Throughout Our Review

Method		Abbrv.	Method		Abbrv.
Anchors	[148]	ANCH	Layer-wise Relevance Propagation (full)	[13]	LRP
ApproShapley (Shapley Value Sampling)	[29]	AS	LRP (composite strategy)	[103], [126]	LRP-CMP
Class Activation Mapping	[206]	CAM	LRP (specific variants)	[13], [126]	LRP-$*$
Contextual Prediction Difference Analysis	[56]	CPDA	Local Interpretable Model-agnostic Explanations	[147]	LIME
DeconvNet	[201]	DCN	Meaningful Perturbation	[47]	MP
DeepLIFT	[170]	DL	NeuronConductance	[36]	NC
DeepLIFT (Rescale)	[170]	DLR	NeuronGuidedBackprop	[178]	NGB
DeepLIFT SHAP	[116]	DLSHAP	NeuronIntegratedGradients	[172]	NIG
Deep Taylor Decomposition	[127]	DTD	Occlusion Analysis	[201]	OCC
ExcitationBackprop	[202]	EB	PatternAttribution	[90]	PA
ExtremalPerturbation	[46]	EP	PatternNet	[90]	PN
GNNExplainer	[198]	GNNEXP	Prediction Difference Analysis	[208]	PDA
GNN-LRP	[162]	GLRP	Randomized Input Sampling for Explanation	[142]	RISE
GradCAM	[167]	GC	Saliency Analysis / Gradient	[14], [174]	SA
Gradient SHAP	[116]	GSHAP	SHapley Additive exPlanations	[116]	SHAP
Gradient $\times$ Input	[170]	GI	SHAP Interaction Index	[115]	SHAPIDX
GuidedBackprop	[178]	GB	SmoothGrad	[176]	SG
Guided GradCam	[167]	GGC	SmoothGrad^2	[76]	SG-SQ
Integrated Gradients	[183]	IG	Spectral Relevance Analysis	[104]	SpRAy
Internal Influence	[110]	II	TreeExplainer	[115]	TEXP
Kernel SHAP	[116]	KSHAP	VarGrad	[1]	VG
LayerConductance	[172]	LC	Testing with Concept Activation Vectors	[89]	TCAV
Local Rule-based Explanations	[58]	LORE	TotalConductance	[36]	TC

W. Samek et.al., 2021[1]

[1]W. Samek et al., "*Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications*". Proceedings of the IEEE, 2021

Important distinctions in XAI

- Local versus global explanations
  - Explain prediction versus explain model
- Model-aware versus model-agnostic
  - With or without access to inner-workings of model
- Explainable versus non-explainable methods
  - Inherently explainable models:
    - Linear models and decision tress.
    - The Occam dilemma[1]

Jellyffish

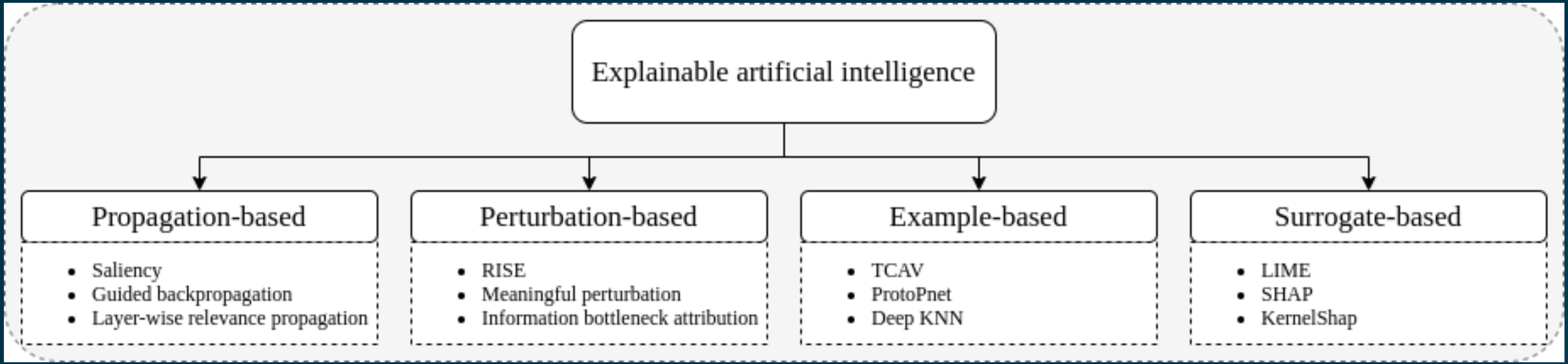
K. Bykov, 2022[1]

- Accuracy generally requires more complex prediction methods. Simple and interpretable functions do not make the most accurate predictors.

L. Breiman, 2001[2]

[1]K. Bykov et al., "*NoiseGrad — Enhancing Explanations by Introducing Stochasticity to Model Weights*". AAAI, 2022.
[2]L. Breiman, "*Statistical modelling: The Two Cultures*". Statistical Science, 2001.

A taxonomy of XAI methods



Propagation-based explainability

- One way to think about explainability:
  - How does a change in the input affect the prediction for a class?
  - Just the gradient!: $\frac{dy_c}{dx} = \boldsymbol{g}$
- Local and model-aware
- Simple, fast, and intuitive
- Numerous variants

Limitations of propagation-based methods

- Nosiy due to gradient shattering[1]
- Can sometimes feel “unintuitive”

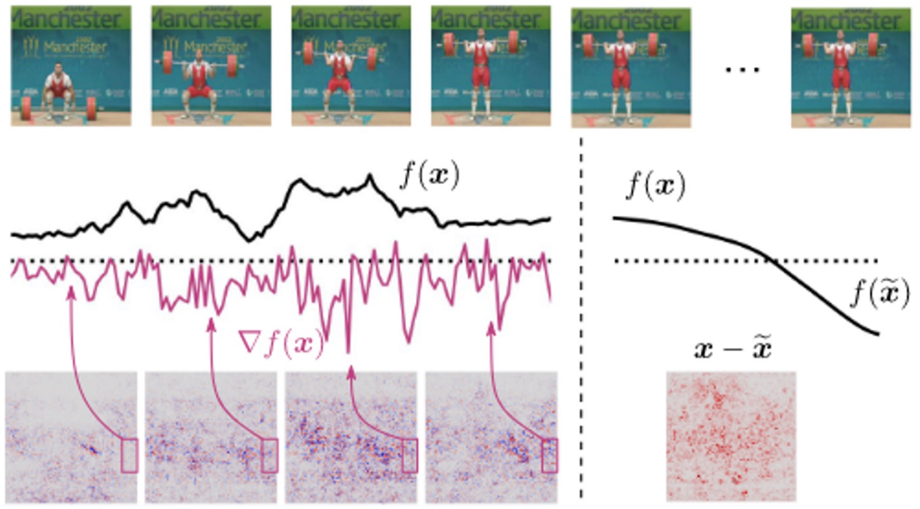


Fig. 3. ***Two difficulties encountered when explaining DNNs. Left: shattered gradient effect causing gradients to be highly varying and too noisy to be used for explanation. Right: pathological minima in the function, making it difficult to search for meaningful reference points.***

[1]W. Samek et al., “*Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications*”. Proceedings of the IEEE, 2021

Permutation-based explainability

- Perturb the input and measure change in prediction score
- Local and model agnostic

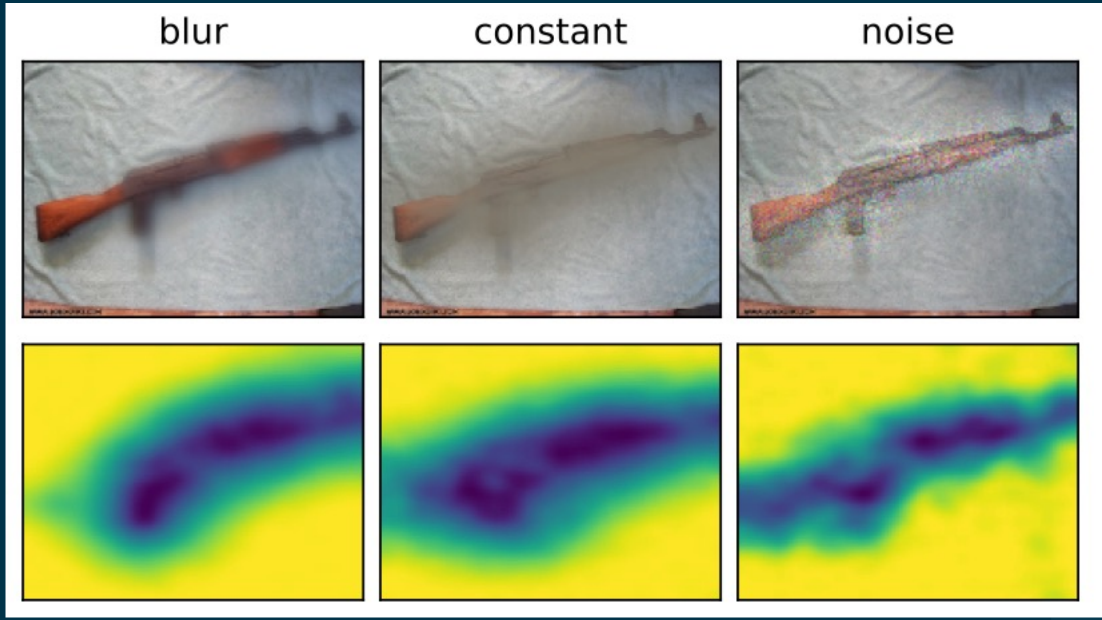
Figure 1. An example of a mask learned (right) by blurring an image (middle) to suppress the softmax probability of its target class (left: original image; softmax scores above images).

R. Fong et.al., 2017

[1]R. Fong et al., "*Interpretable Explanations of Black Boxes by Meaningful Perturbation*". ICCV, 2017

Limitations of perturbation-based methods

- Requires optimizing or sampling per sample:
  - Can be slow
- How to replace input parts is non-trivial



R. Fong et.al., 2017

[1]R. Fong et al., "*Interpretable Explanations of Black Boxes by Meaningful Perturbation*". ICCV, 2017

Surrogate-based explainability

- Train a simple interpretable model to explain the black box model
- Very versatile!

The recipe for training local surrogate models:

- Select your instance of interest for which you want to have an explanation of its black box prediction.
- Perturb your dataset and get the black box predictions for these new points.
- Weight the new samples according to their proximity to the instance of interest.
- Train a weighted, interpretable model on the dataset with the variations.
- Explain the prediction by interpreting the local model.

C. Molnar[1]

[1]C. Molnar , https://christophm.github.io/interpretable-ml-book/lime.html

Limitations of surrogate-based explainablity

- Need to train the simple classifier:
  - Can give different explanation for same sample due to optimization or the Rashomon effect[1]
- Lacks robustness[2]

[1] L. Breiman, "*Statistical modelling: The Two Cultures*". Statistical Science, 2001.
[2] D. Alvarez-Melis et.al., "*On the robustness of interpretablity methods*". ICML workshop, 2018.

Challenges for contemporary XAI

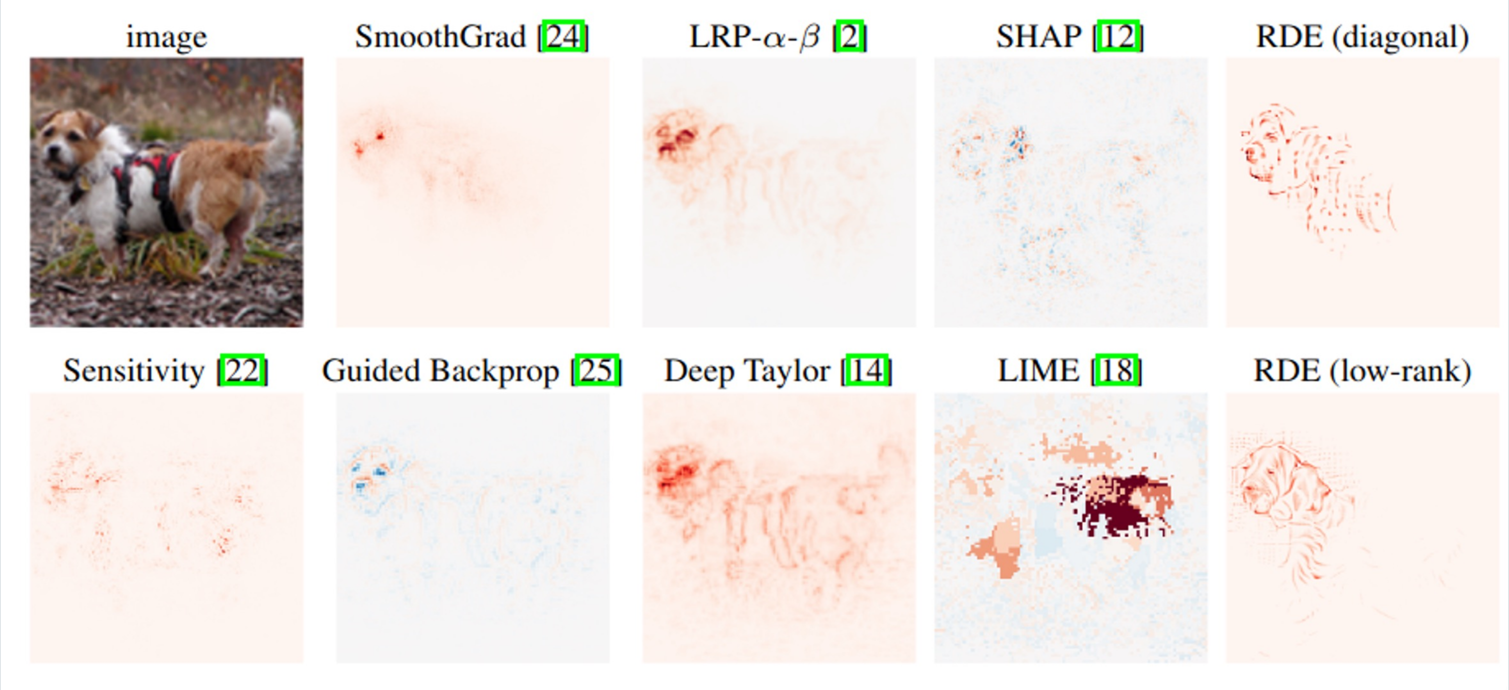
- Numerous recent advances in XAI
- However, many challenges on the horizon:
  - Disagreement among methods or what makes a “good explanation?
  - How to go beyond explaining “just” scalar predictions
  - How to model uncertainty in explanations?
  - How to explain highly complex input data?

Challenges for contemporary XAI

- Numerous recent advances in XAI
- However, many challenges on the horizon:
  - Disagreement among methods or what makes a “good explanation?
  - How to go beyond explaining “just” scalar predictions.
  - How to model uncertainty in explanations?
  - How to explain highly complex input data?

What makes a good explanation?

- What is the best explanation in the following example?



J. Macdonald et.al., 2019[1]

[1]J. Macdonald et al., "*A Rate-Distortion Framework for Explaining Neural Network Decisions*". Arxiv, 2019

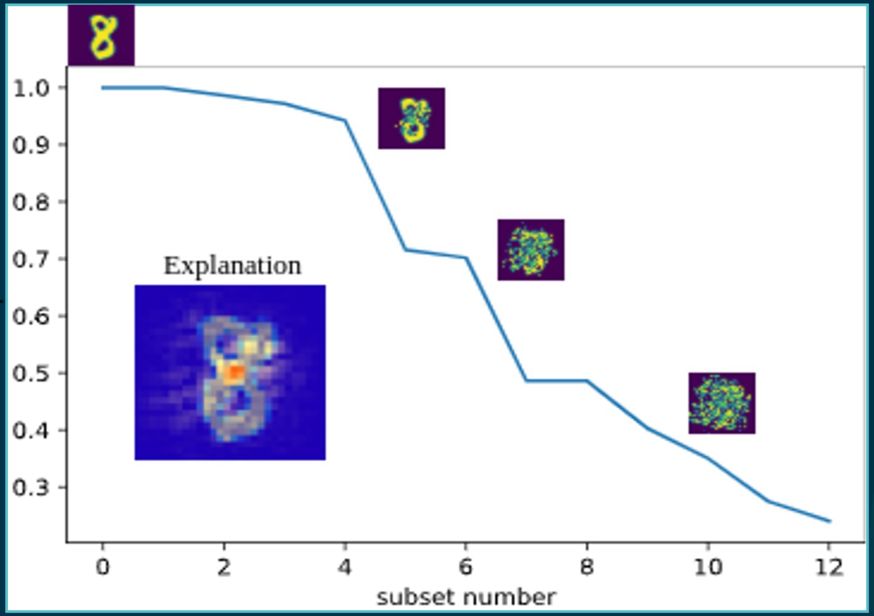
What makes a good explanation?

- XAI lacks ground truth explanation for verification
- A result of the challenge of unverifiability[1]
- Also known as the disagreement problem[2]
- Two directions to tackle this challenge:
  - Quantitative analysis
  - Self-explainable models

[1] A. Hedström et al., "*The Meta-Evaluation Problem in Explainable AI: Identifying Reliable Estimators with MetaQuantus*". TMLR, 2023
[1] S. Krishna et al., "*The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective* ". ArXiv, 2022

Route 1: quantitative analysis

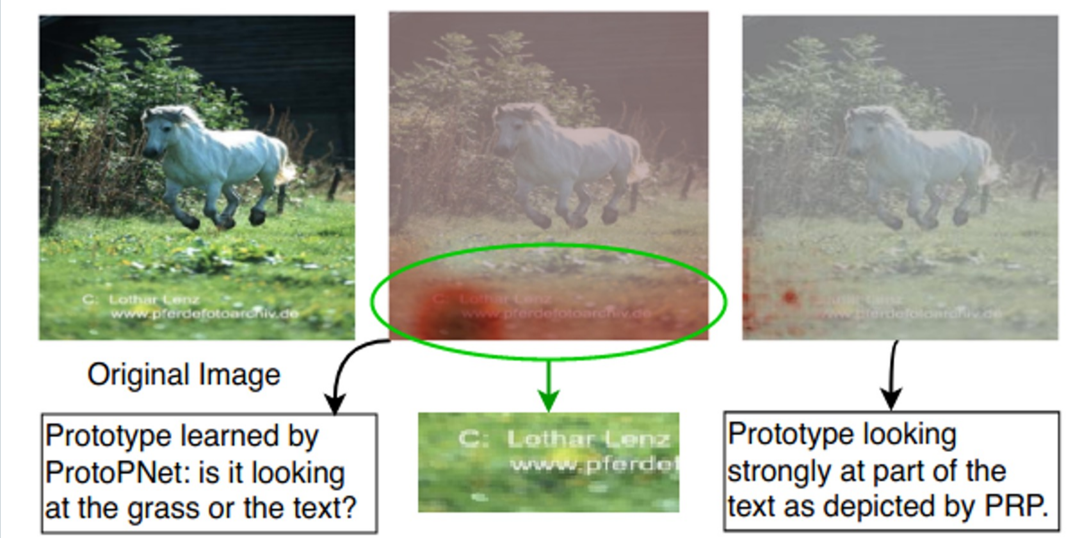
- Define desirable properties for the explanation to fulfil
- Quantus: recent toolbox for quantitative analysis[1]
- Quantitative analysis categories:
  - Localization
  - Faithfulness
  - Robustness
  - Complexity
- Faithfulness example ->



[1]A. Hedström et al., "Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations". JMLR, 2022

Route 2: Self-explainable models

- Do not explain the model, build explainability into the model[1]
- ProtoPnet[2]:
  - Add prototypes to network
  - Prototypes are related to concepts
- Drawback:
  - Additional complexity
  - Can hurt performance



S. Gautam et.al., 2022[3]

[1]C. Rudin, "*Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*". Nature Machine Intelligence, 2019
[2]C. Chen et al., "This Looks Like That: Deep Learning for Interpretable Image Recognition". NeurIPS, 2019
[3]S. Gautam et al., "This looks more like that: Enhancing Self-Explaining Models by Prototypical Relevance Propagation". Pattern Recognition, 2022.

Break before next seminar

- In next seminar:
  - XAI in representation learning
  - How to explain representations?