AI for Good, Trustworthy AI series

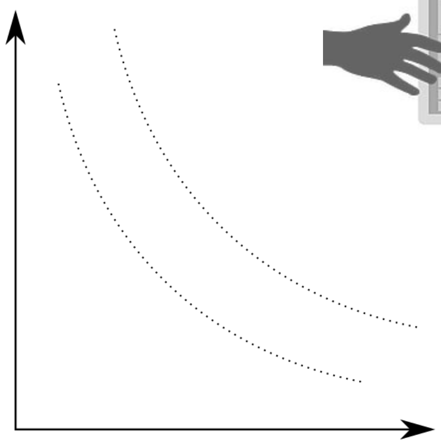
Explainable AI (XAI) and trust

Grégoire Montavon **et al.**

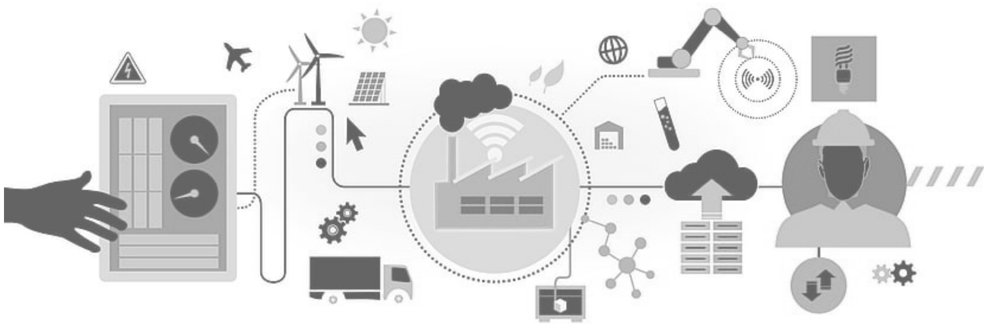
Thursday, 27 May 2021

The Need for Trustworthy AI

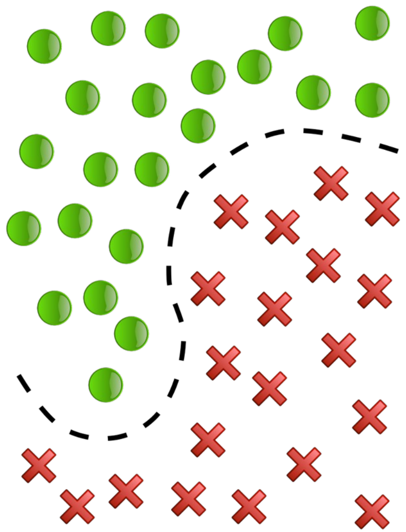
High-Stake Decisions



Autonomous Decisions



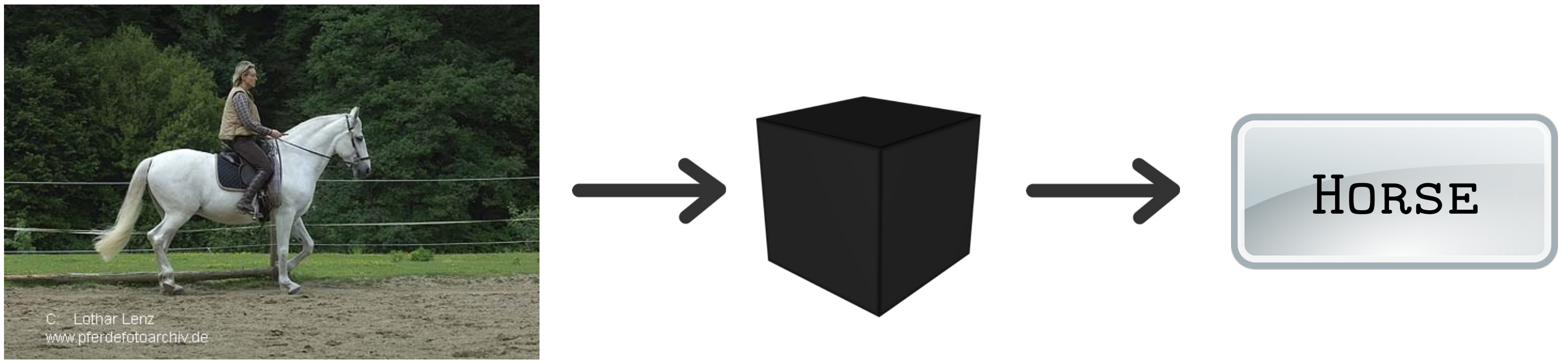
Machine Learning Decisions



Machine learning puts the focus on collecting the **data** that the decision function has to correctly predict rather than specifying the function by hand.

Question: Can we trust machine learning models?

Example: Detecting Horses



input image

ML Blackbox
(BoW classifier)

prediction

Observation of the predicting behavior of the ML model: Images of horses are being correctly classified as "horses".

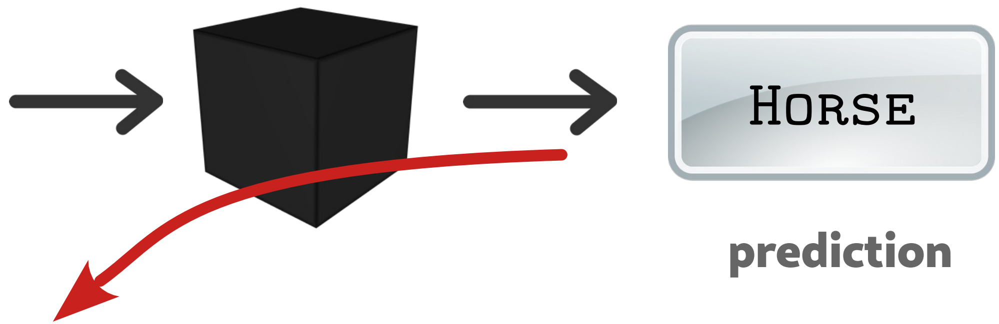
Example: Detecting Horses

average precision of the Fisher Vector model on the PascalVOC dataset

aer 79.08	bic 66.44	bir 45.90	boa 70.88	bot 27.64
bus 69.67	car 80.96	cat 59.92	cha 51.92	cow 47.60
din 58.06	dog 42.28	hor 80.45	mot 69.34	per 85.10
pot 28.62	she 49.58	sof 49.31	tra 82.71	tvm 54.33

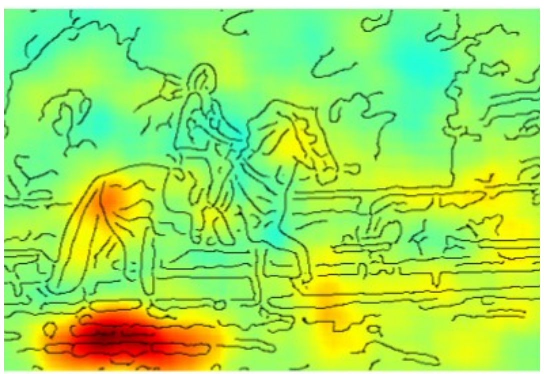
The accuracy of horse detection is high on average on the available test data.

Example: Detecting Horses



prediction

+ explanation



Unexpected: The classifier predicts correctly based on an **artifact** in the data (aka. '**Clever Hans**').

Example: Detecting Horses

Reason: This strategy works on the current data (many horses images have a copyright tag) → **spurious correlation**.

Example: Detecting Horses

Because the classifier relies on a non-informative feature (the copyright tag), it can be easily fooled.

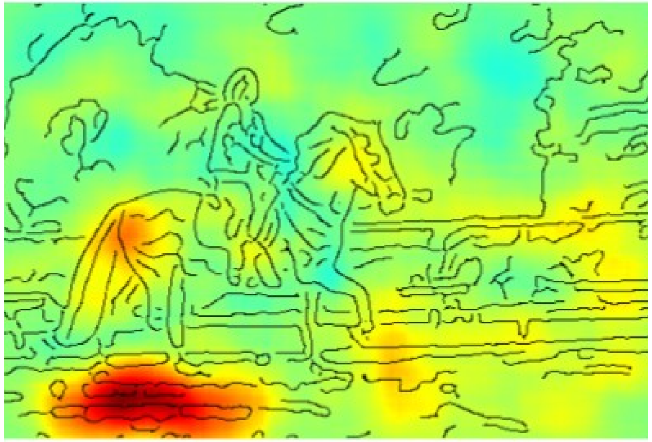
Examples:

→ NOT HORSE

→ HORSE

Clever Hans models are unlikely to perform well on **future data**.

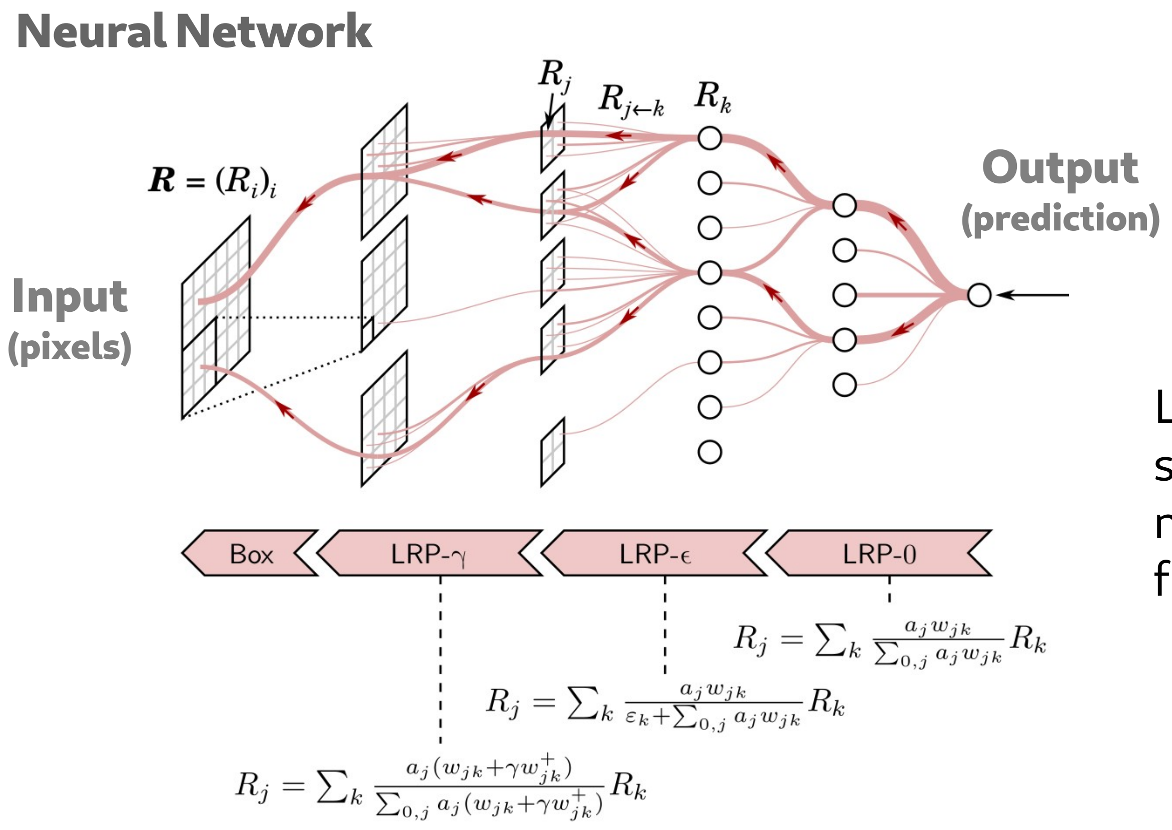
But how do we get these Heatmaps?



Computing reliable explanations of the prediction is a **non-trivial task** (the ML model only outputs a prediction, but has no intrinsic self-explainability).

Fast progress has been made on explaining ML predictions. A technique we developed for this is **Layer-wise Relevance Propagation (LRP)**.

Layer-wise Relevance Propagation (LRP)



LRP runs in the order of a single backward pass (no need to evaluate the function multiple times).

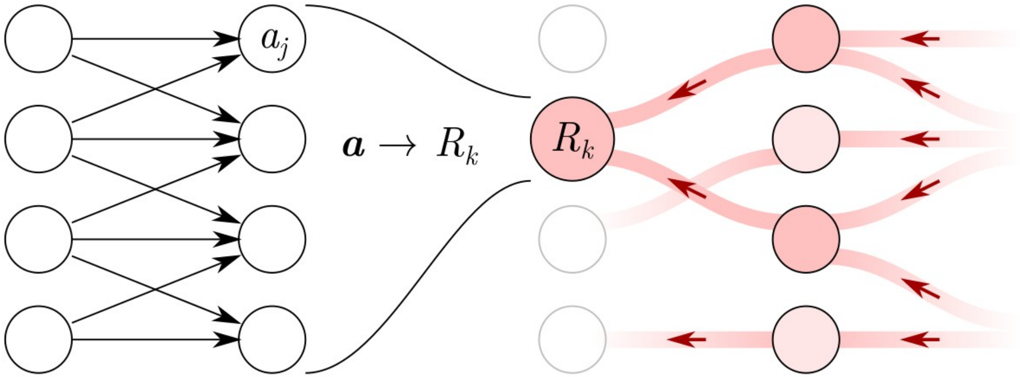
Bach et al. (2015) On Pixel-wise Explanations for Non-Linear Classifier Decisions by Layer-wise Relevance Propagation

Can LRP be Justified Theoretically?

$$R_j = \sum_k \frac{a_j \cdot \rho(w_{jk})}{\epsilon + \sum_{0,j} a_j \cdot \rho(w_{jk})} R_k$$

Answer: Yes, using the deep Taylor decomposition framework.

Deep Taylor Decomposition



Key idea: Taylor expansions at each layer

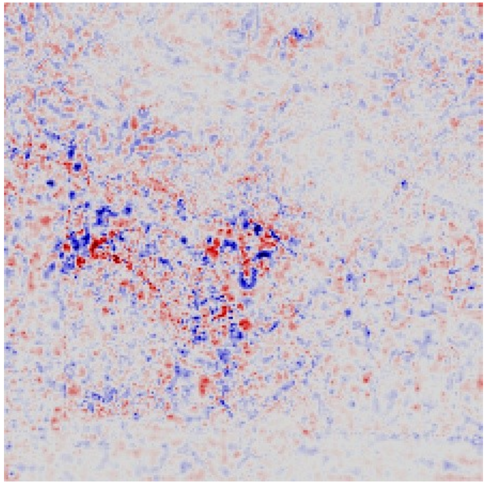
$$R_k(\boldsymbol{a}) \approx \widehat{R}_k(\widetilde{\boldsymbol{a}}) + \underbrace{\textstyle\sum_j [\nabla \widehat{R}_k(\widetilde{\boldsymbol{a}})]_j \cdot (a_j - \widetilde{a}_j)}_{\textbf{LRP}} + \ldots$$

Montavon et al. (2017) Explaining nonlinear classification decisions with deep Taylor decomposition

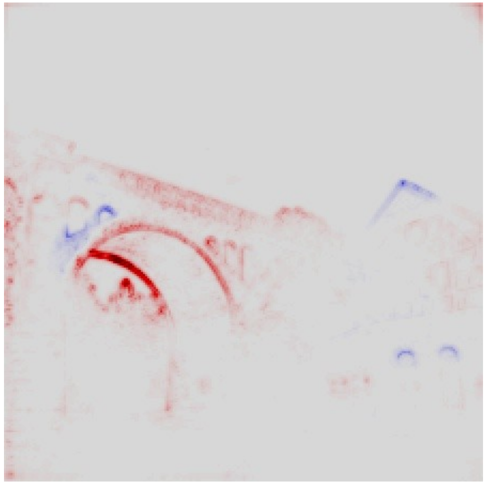
LRP is More Stable than Gradient

Image classified by a DNN as a viaduct.

Gradient explanation

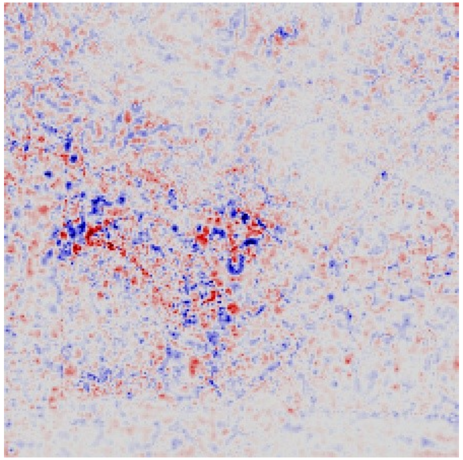


LRP explanation

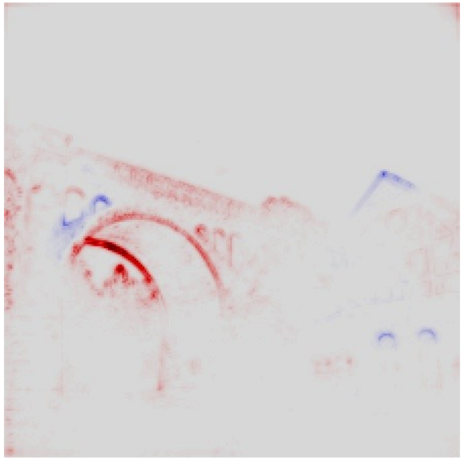


LRP is More Stable than Gradient

Gradient explanation

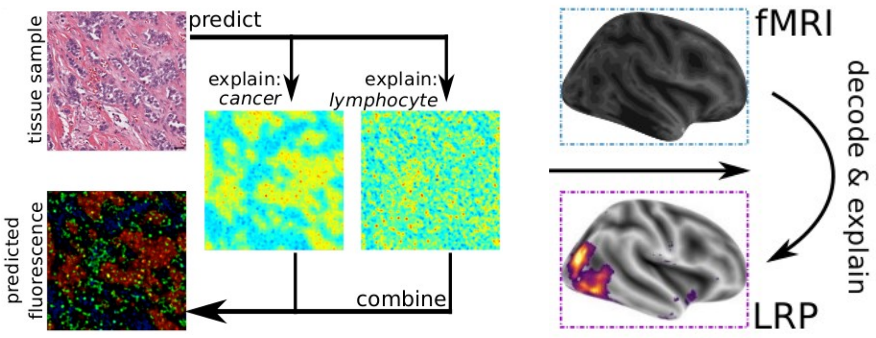


LRP explanation

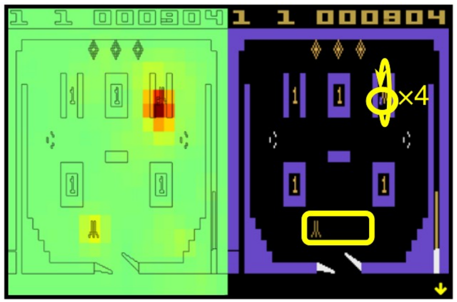


LRP on Different Types of Data

Medical data (images/FMRI/EEG/...)



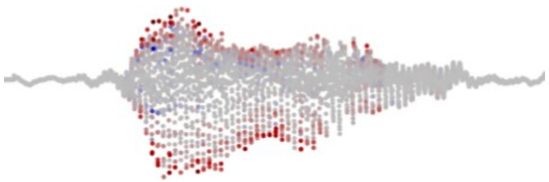
Arcade games



Natural language

```
on a roller coaster ride than others.  The mental part is usually induced by
a lack of clear indication of which way is up or down, ie: the Shuttle is
normally oriented with its cargo bay pointed towards Earth, so the Earth
(or ground) is "above" the head of the astronauts.  About 50% of the astronauts
experience some form of motion sickness, and NASA has done numerous tests in
```

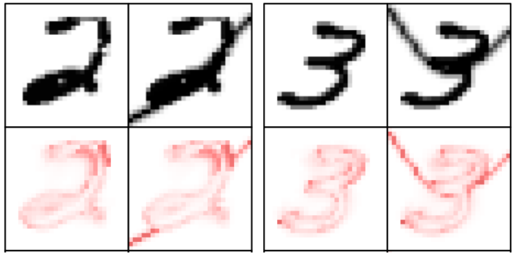
Speech



LRP for Different Types of Models

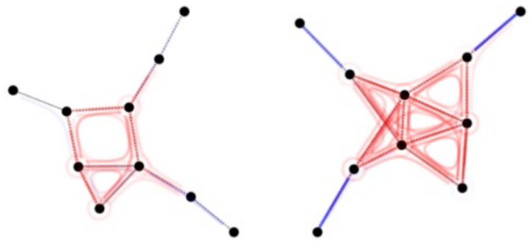
DNN Classifiers

Anomaly models

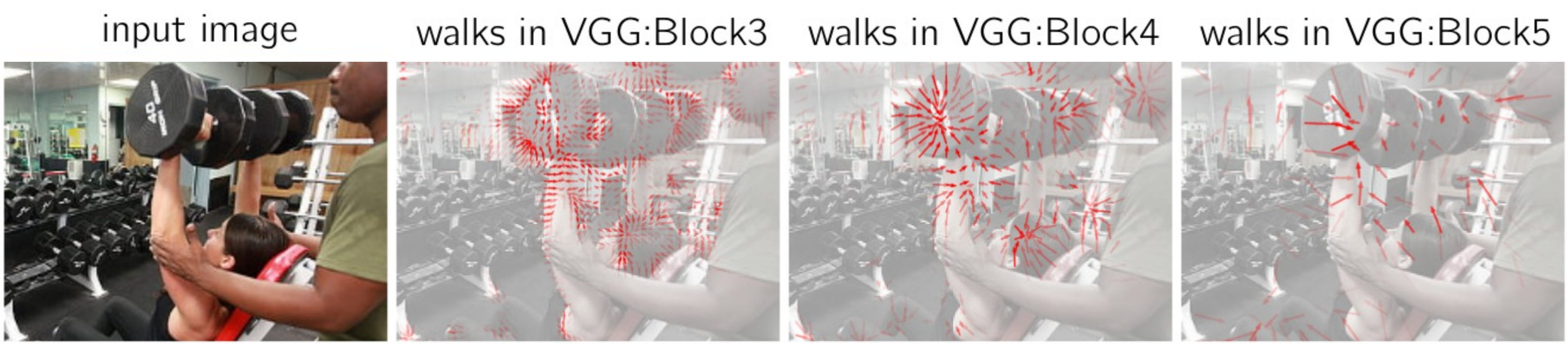


Similarity models (BiLRP)

Graph neural networks (GNN-LRP)



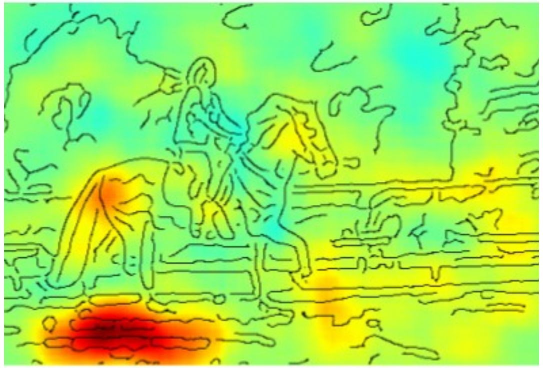
Advanced Explanation with GNN-LRP



VGG-16 performs edge/corner detection	VGG-16 detects independent objects	VGG-16 merges the arm and the dumbell

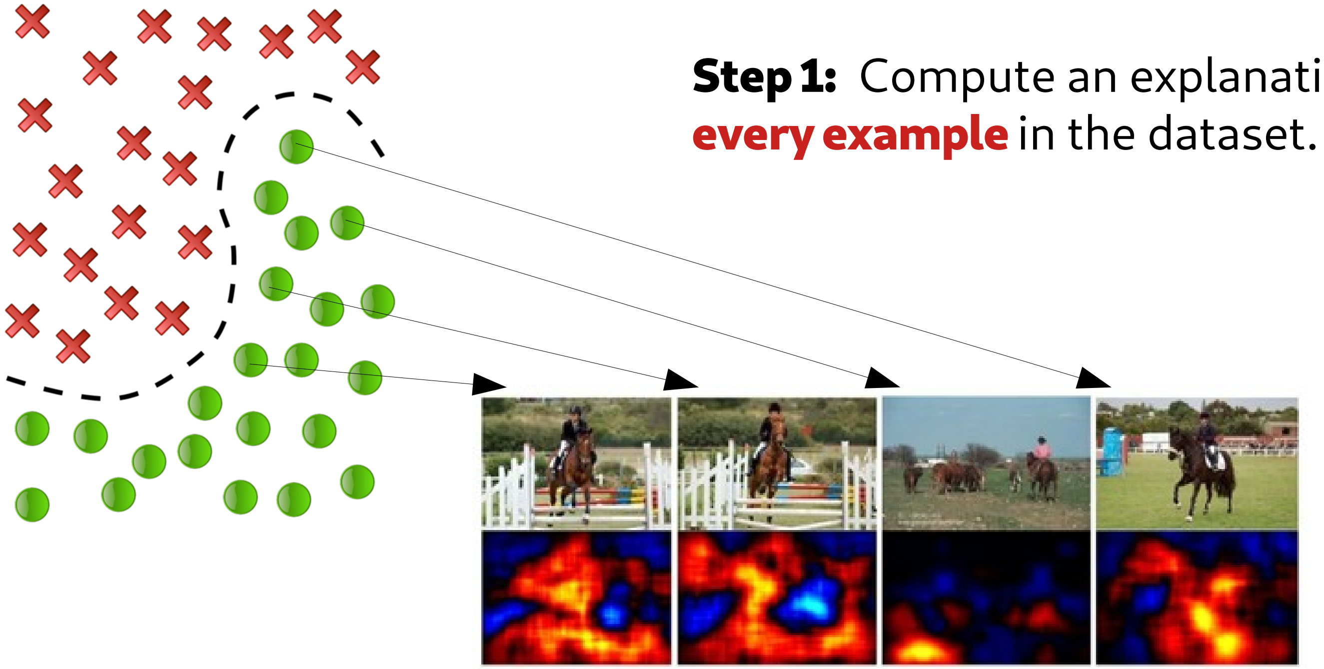
Schnake et al. (2020)
Higher-Order Explanations of Graph Neural Networks via Relevant Walks

Systematically Finding Clever Hans



The decision artefact has been found occasionally by having the user look at an explanation for some image of the class horse. But can we achieve a **broader** and more systematic inspection of the model ?

Idea: Spectral Relevance Analysis (SpRAy)

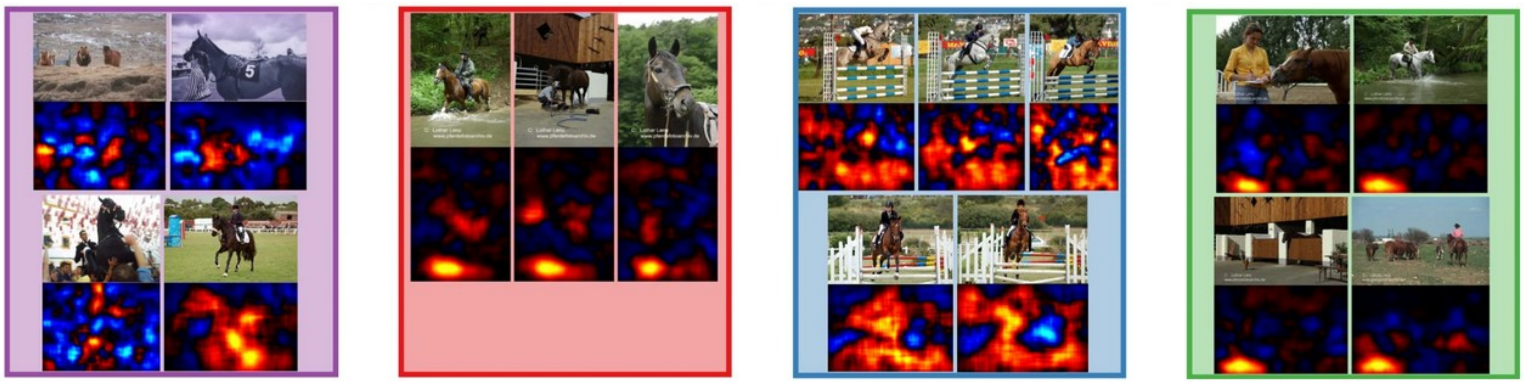


Step 1: Compute an explanation for **every example** in the dataset.

Lapuschkin et al. (2019)
Unmasking Clever Hans Predictors and Assessing What Machines Really Learn

Idea: Spectral Relevance Analysis (SpRAy)

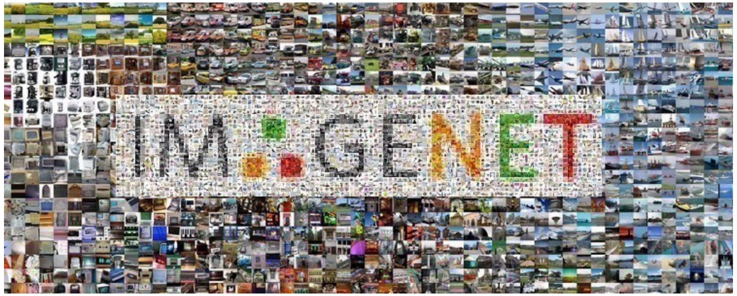
Step 2: Organize explanations into **clusters**.



Clever Hans effects are now obtained **systematically**.

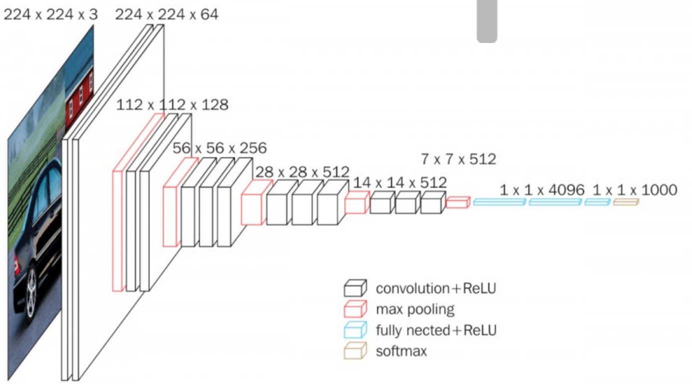
Lapuschkin et al. (2019)
Unmasking Clever Hans Predictors and Assessing What Machines Really Learn

The Revolution of Depth (2012-...)



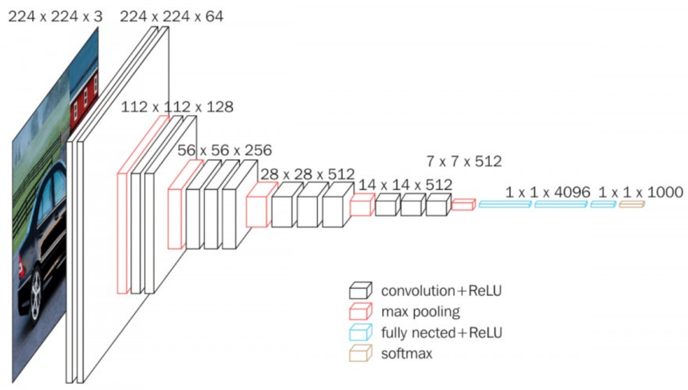
millions of labeled images

+



deep neural networks (trained on GPUs)

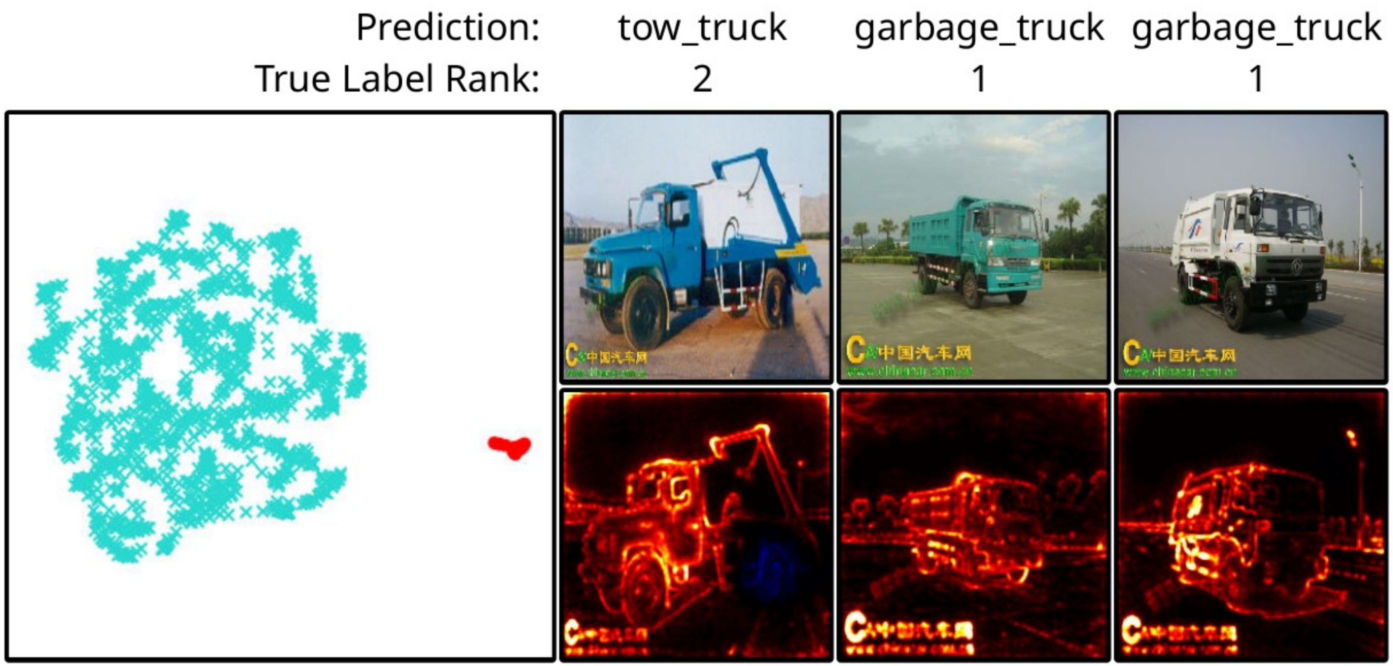
Clever Hans on Large Models



Question:

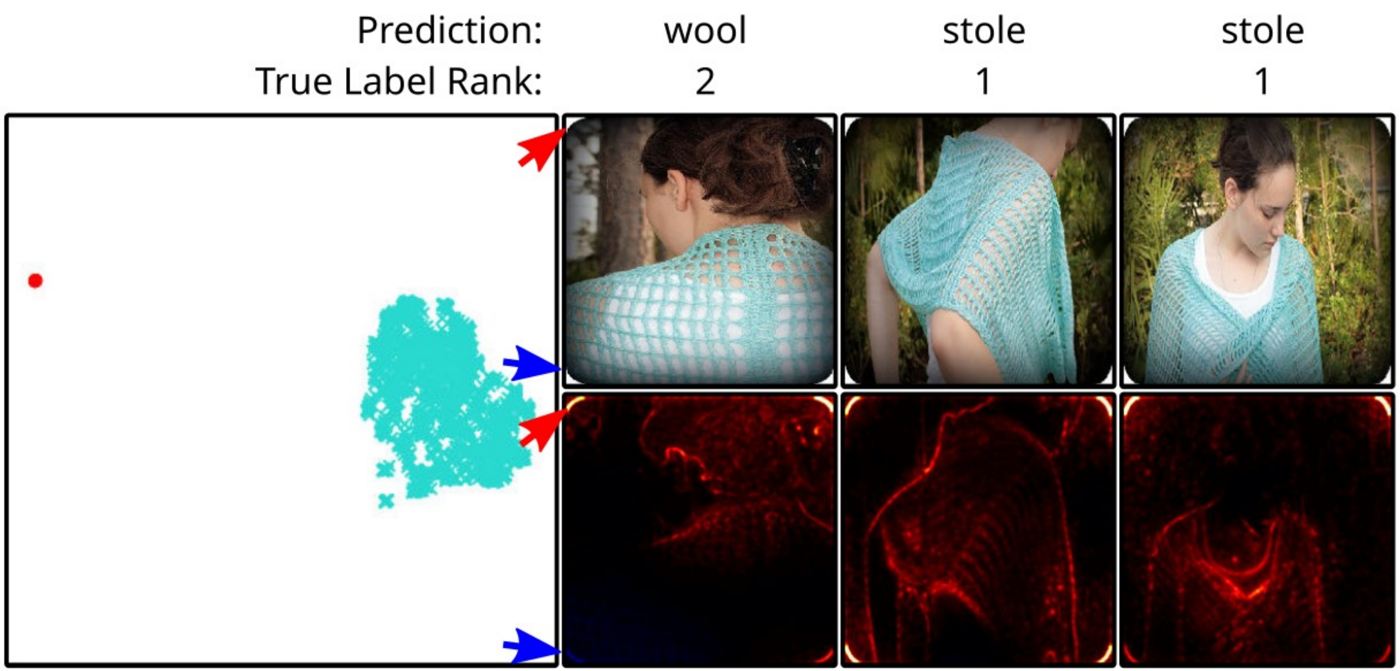
Are large deep neural networks trained on **millions** of data points also subject to the **Clever Hans** effect?

Clever Hans on the VGG-16 Image Classifier



Anders et al. (2020) Finding and Removing Clever Hans: Using Explanation Methods to Debug and Improve Deep Models

Clever Hans on the VGG-16 Image Classifier



Anders et al. (2020) Finding and Removing Clever Hans: Using Explanation Methods to Debug and Improve Deep Models

XAI Current Challenges

Explanation Fidelity: Explanation must accurately capture the decision strategy of the model. Accurately evaluating explanation fidelity is still an open question.

XAI Current Challenges

Explanation Fidelity: Explanation must accurately capture the decision strategy of the model. Accurately evaluating explanation fidelity is still an open question.

Explanation Understandability: When the decision strategy is complex, the user may not be able to distinguish between a correct and a flawed decision strategy, even if the explanation is correct.

XAI Current Challenges

Explanation Fidelity: Explanation must accurately capture the decision strategy of the model. Accurately evaluating explanation fidelity is still an open question.

Explanation Understandability: When the decision strategy is complex, the user may not be able to distinguish between a correct and a flawed decision strategy, even if the explanation is correct.

Explanation for Validating a ML Model: Even after applying SpRAy, there may in theory still be “hidden” Clever Hanses in the model (especially for models with strong ability to generalize).

XAI Current Challenges

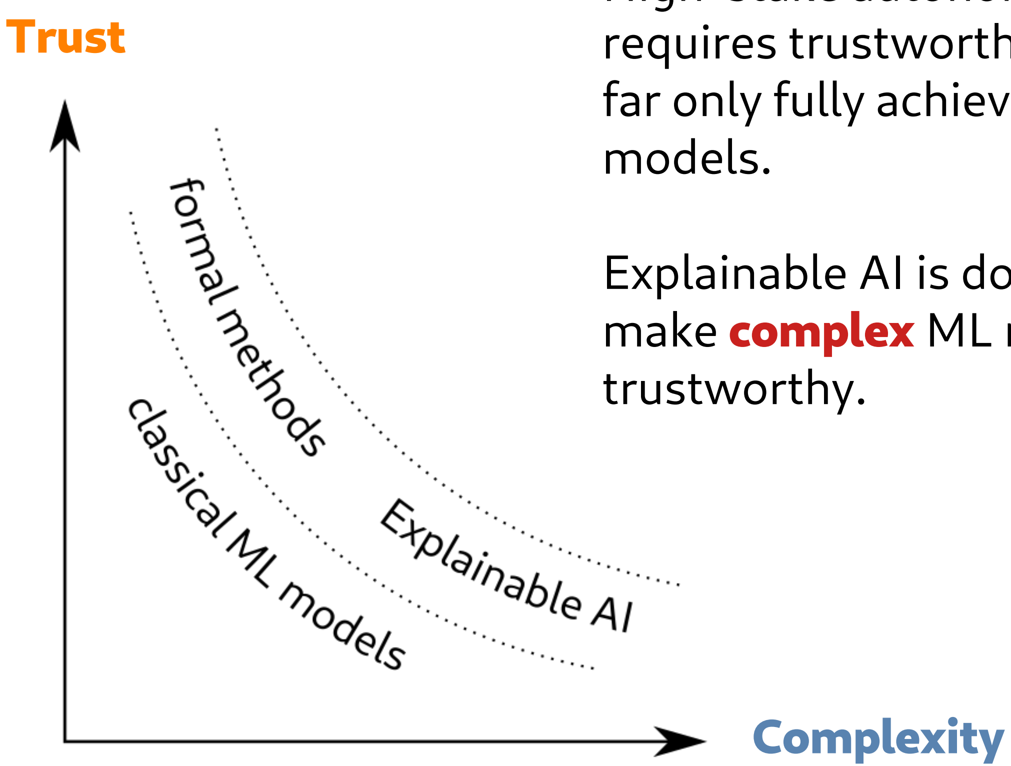
Explanation Fidelity: Explanation must accurately capture the decision strategy of the model. Accurately evaluating explanation fidelity is still an open question.

Explanation Understandability: When the decision strategy is complex, the user may not be able to distinguish between a correct and a flawed decision strategy, even if the explanation is correct.

Explanation for Validating a ML Model: Even after applying SpRAy, there may in theory still be "hidden" Clever Hanses in the model (especially for models with strong ability to generalize).

Explanation Robustness: XAI is potentially vulnerable to adversarial attacks (e.g. crafting input and models that produce wrong explanations).

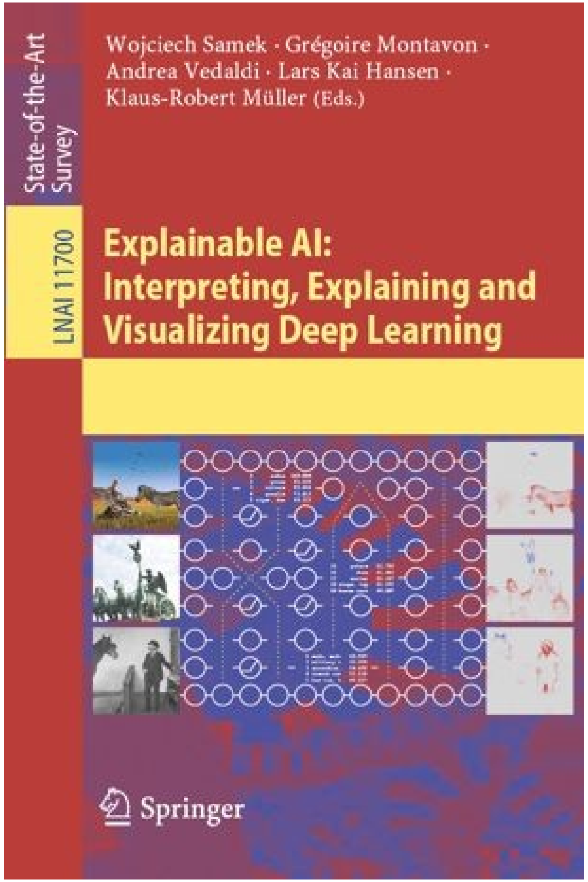
Towards Trustworthy AI



High-stake autonomous decisions requires trustworthy models. This is so far only fully achievable for **simple** models.

Explainable AI is doing rapid progress to make **complex** ML models more trustworthy.

Our Book on Explainable AI



Our Explainable AI Website

Online demos, tutorials, code examples, software, etc.

And our recent review paper:

W Samek, G Montavon, S Lapuschkin, C Anders, KR Müller. Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. Proceedings of the IEEE, 109(3):247-278, 2021

References

[1] S Bach, A Binder, G Montavon, F Klauschen, KR Müller, W Samek: On Pixel-wise Explanations for Non-Linear Classifier Decisions by Layer-wise Relevance Propagation. PLOS ONE, 10(7):e0130140 (2015)

[2] G Montavon, S Lapuschkin, A Binder, W Samek, KR Müller: Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recognit. 65: 211-222 (2017)

[3] S Lapuschkin, S Wäldchen, A Binder, G Montavon, W Samek, KR Müller. Unmasking Clever Hans Predictors and Assessing What Machines Really Learn, Nature Communications, 10:1096, 2019

[4] J Kauffmann, KR Müller, G Montavon. Towards Explaining Anomalies: A Deep Taylor Decomposition of One-Class Models, Pattern Recognition, 107198, 2020

[5] O Eberle, J Büttner, F Kräutli, KR Müller, M Valleriani, G Montavon. Building and Interpreting Deep Similarity Models, IEEE Transactions on Pattern Analysis and Machine Intelligence, Early Access, 2020

[6] T Schnake, O Eberle, J Lederer, S Nakajima, K T. Schütt, KR Müller, G Montavon. Higher-Order Explanations of Graph Neural Networks via Relevant Walks, arXiv:2006.03589, 2020

[7] W Samek, G Montavon, S Lapuschkin, C Anders, KR Müller. Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. Proceedings of the IEEE, 109(3):247-278, 2021