Pattern Recognition
and Applications Lab

Explainable AI

Battista Biggio, Maura Pintor

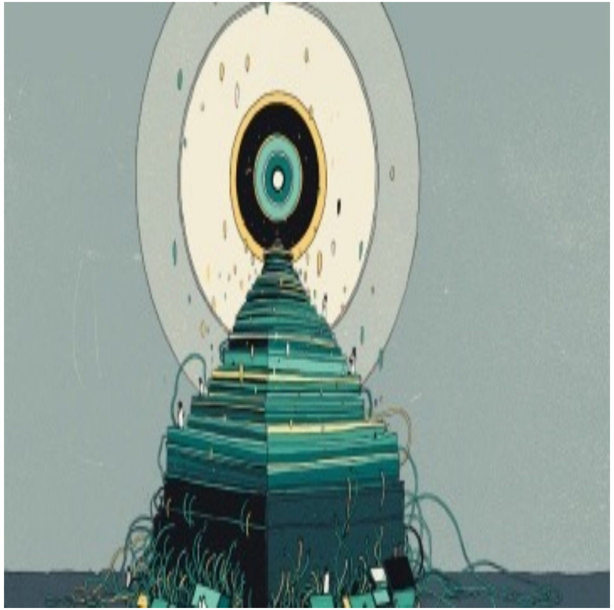
Department of Electrical and Electronic Engineering
University of Cagliari, Italy

When and Why Model Understanding?

ML is increasingly being employed in complex high-stakes settings

Slide from: H. Lakkaraju, Interpreting Machine Learning Models: State-of-the-art, Challenges, Opportunities - 2022

Safety to the Fore…



The black box of AI
D. Castelvecchi, Nature, Vol. 538, 20, Oct 2016

Machine learning is becoming ubiquitous in basic research as well as in industry. But for scientists to trust it, they first need to understand what the machines are doing.

Ellie Dobson, director of data science at the big-data firm Arundo Analytics in Oslo:

- If something were to go wrong as a result of setting the UK interest rates, she says, “the Bank of England can’t say, the black box made me do it”.

Explainability and Why It Is Important

Fairness: Ensuring that predictions are unbiased

Privacy: Ensuring that sensitive information in the data is protected

Safety and Robustness: Ensuring that small changes in the input do not lead to large changes in the prediction

Causality: Check that only causal relationships are picked up

Trust: It is easier for humans to trust a system that explains its decisions compared to a black box

(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

M. T. Ribeiro et al.,"Why should I trust you?": …, KDD 2016

Summary: Why Model Understanding?

Utility

Debugging

Bias Detection

Recourse

If and when to trust model predictions

Vet models to assess suitability for deployment

Stakeholders

End users (e.g., loan applicants)

Decision makers (e.g., doctors, judges)

Regulatory agencies (e.g., FDA, European commission)

Researchers and engineers

Slide from: H. Lakkaraju, Interpreting Machine Learning Models: State-of-the-art, Challenges, Opportunities - 2022

Explainability Methods

A Survey of Methods for Explaining Black-box Models

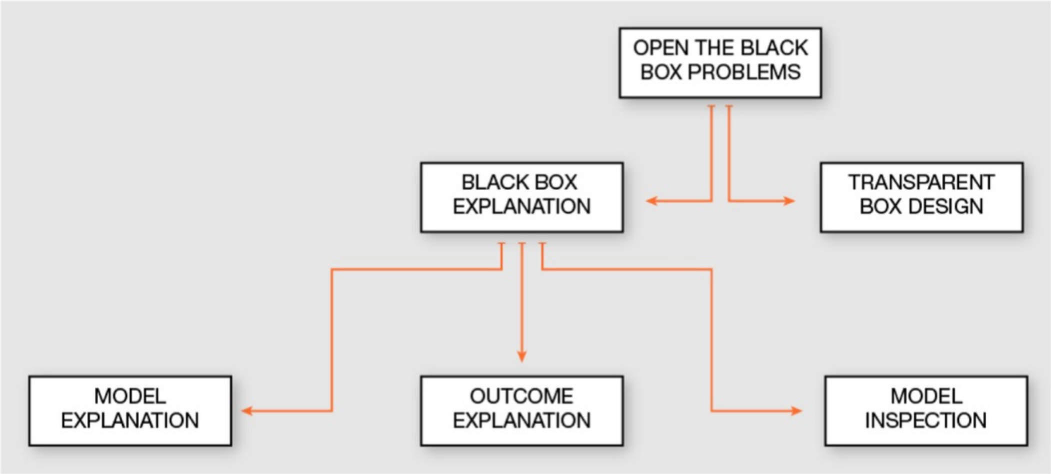
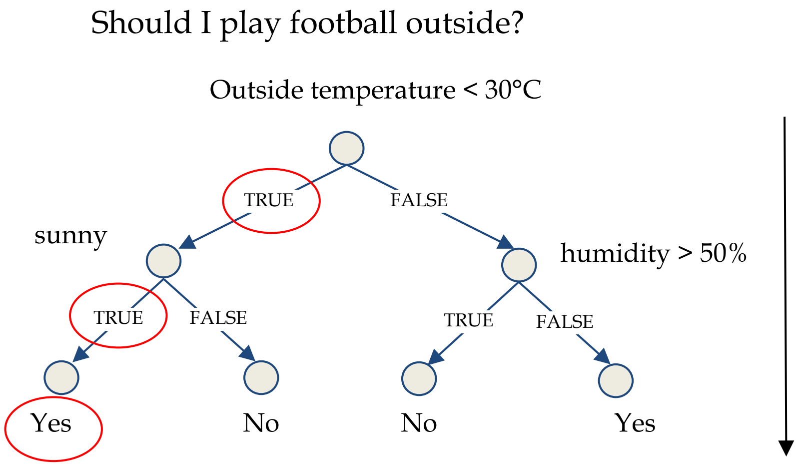


Fig. 4. Open the black box problems taxonomy. The *Open the Black Box Problems* for understanding how a black box works can be separated from one side as the problem of *explaining* how the decision system returned certain outcomes (*Black Box Explanation*) and on the other side as the problem of directly designing a *transparent* classifier that solves the same classification problem (*Transparent Box Design*). Moreover, the Black Box Explanation problem can be further divided among *Model Explanation* when the explanation involves the whole logic of the obscure classifier, *Outcome Explanation* when the target is to understand the reasons for the decisions on a given object, and *Model Inspection* when the target to understand how internally the black box behaves changing the input.

R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. ACM Comput. Surv., 2019

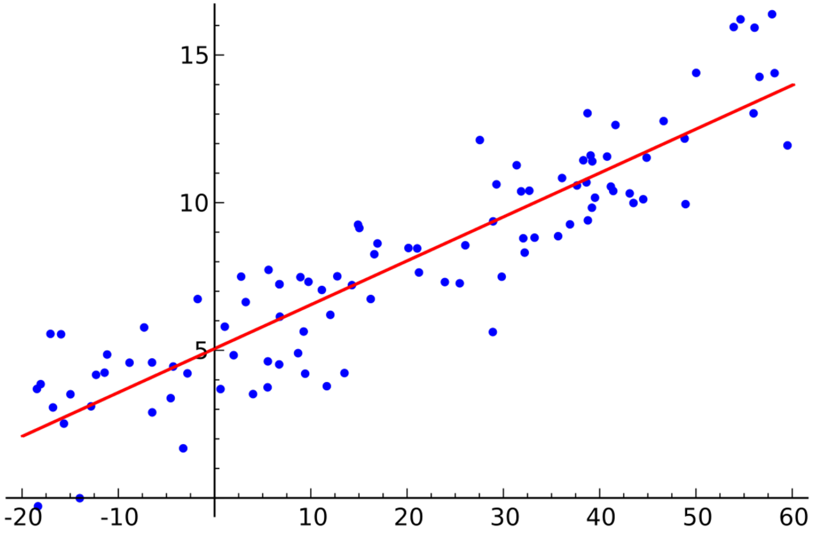
Interpretable-by-Design (Transparent) Models



Depth = how many levels of decision

Too much depth makes the model **not interpretable**

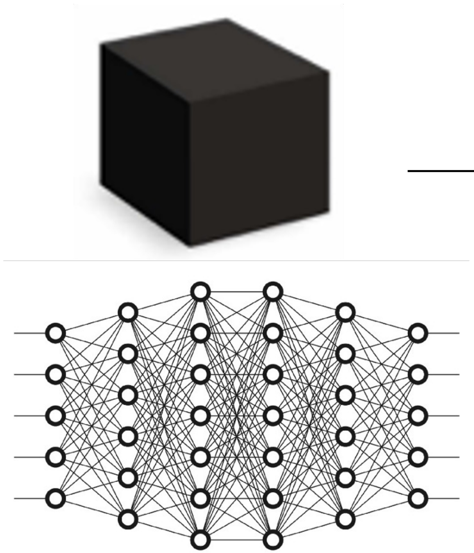
Interpretable-by-Design (Transparent) Models



Even **linear** classifiers may be hard to interpret when dealing with high-dimensional problems

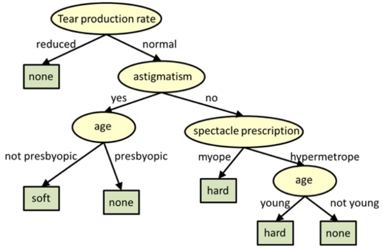
Black-box Explanation

Explain pre-built models *in a post-hoc manner*



Explainer

if $(age = 18 - 20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21 - 23)$ **and** $(priors = 2 - 3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no

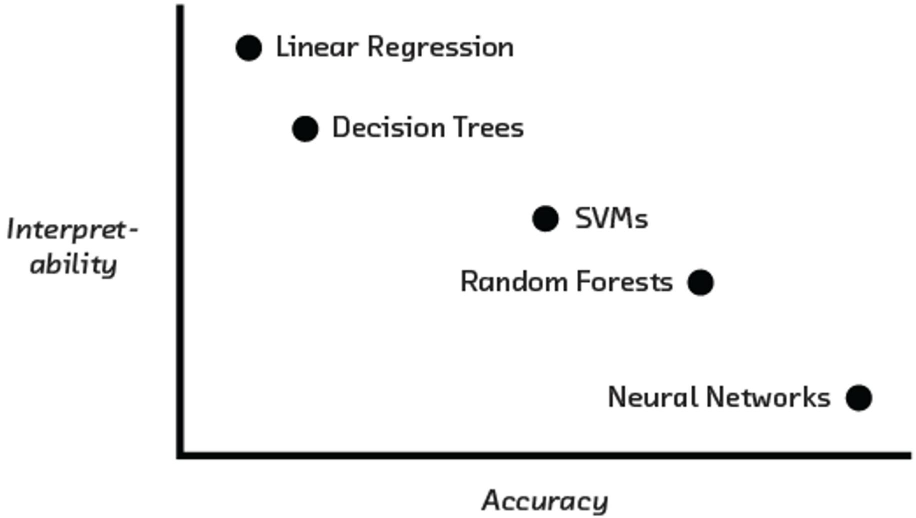


Ribeiro et. al. 2016, Ribeiro et al. 2018; Lakkaraju et. al. 2019

Slide from: H. Lakkaraju, Interpreting Machine Learning Models: State-of-the-art, Challenges, Opportunities - 2022

Interpretable-by-Design Models vs. Post-hoc Explanations

- In **certain** settings, *accuracy-interpretability trade offs* may exist



Myth or Reality?

Cynthia Rudin: **Please stop doing Explainable ML!**

Slide from: H. Lakkaraju, Interpreting Machine Learning Models: State-of-the-art, Challenges, Opportunities - 2022

A Survey of Methods for Explaining Black-box Models

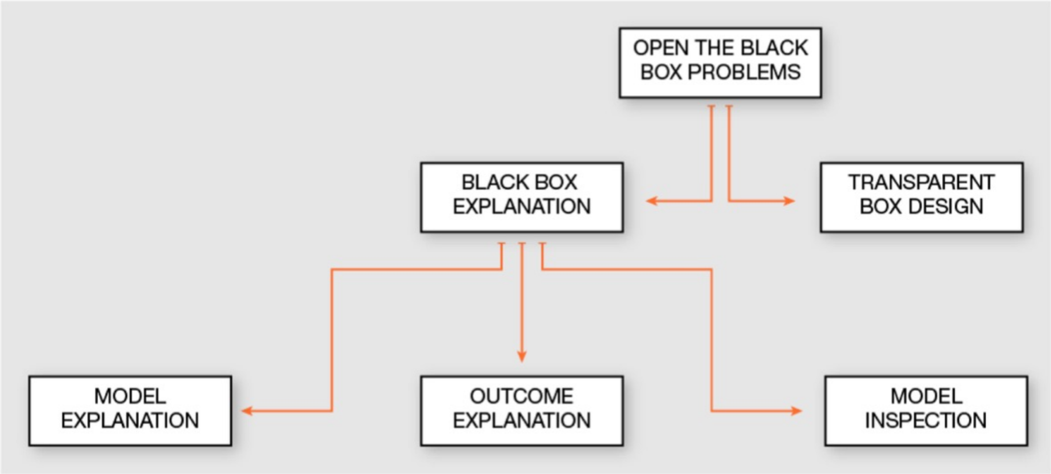
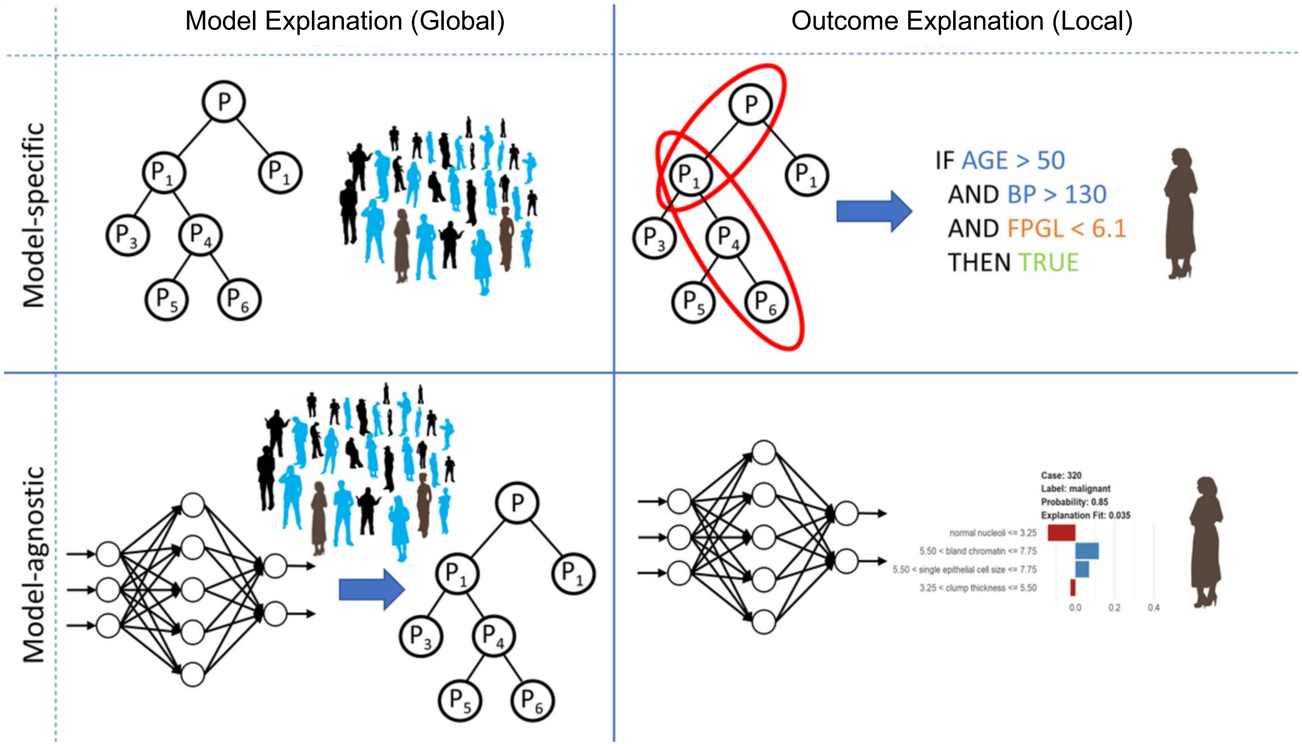


Fig. 4. Open the black box problems taxonomy. The *Open the Black Box Problems* for understanding how a black box works can be separated from one side as the problem of *explaining* how the decision system returned certain outcomes (*Black Box Explanation*) and on the other side as the problem of directly designing a *transparent* classifier that solves the same classification problem (*Transparent Box Design*). Moreover, the Black Box Explanation problem can be further divided among *Model Explanation* when the explanation involves the whole logic of the obscure classifier, *Outcome Explanation* when the target is to understand the reasons for the decisions on a given object, and *Model Inspection* when the target to understand how internally the black box behaves changing the input.

R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. ACM Comput. Surv., 2019

Taxonomy of Explainability Methods



Stiglic et al. "*Interpretability of machine learning-based prediction models in healthcare*". *WIREs Data Mining Knowl. Discov*, 2020.

Local Explanations vs. Global Explanations

Explain individual predictions	Explain complete behavior of the model
Help unearth biases in the *local neighborhood* of a given instance	Help shed light on *big picture biases* affecting larger subgroups
Help vet if individual predictions are being made for the right reasons	Help vet if the model, at a high level, is suitable for deployment

Slide from: H. Lakkaraju, Interpreting Machine Learning Models: State-of-the-art, Challenges, Opportunities - 2022

Approaches for Post hoc Explainability

Local Explanations

- Feature Importances
- Rule Based
- Saliency Maps
- Prototypes/Example Based
- Counterfactuals

Global Explanations

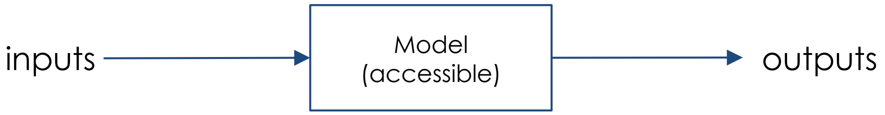
- Collection of Local Explanations
- Representation Based
- Model Distillation
- Summaries of Counterfactuals

Slide from: H. Lakkaraju, Interpreting Machine Learning Models: State-of-the-art, Challenges, Opportunities - 2022

Model-agnostic Methods

- **Black-box**: work by observing only input-output pairs

- **White-box**: access to model’s internals (usually gradients)



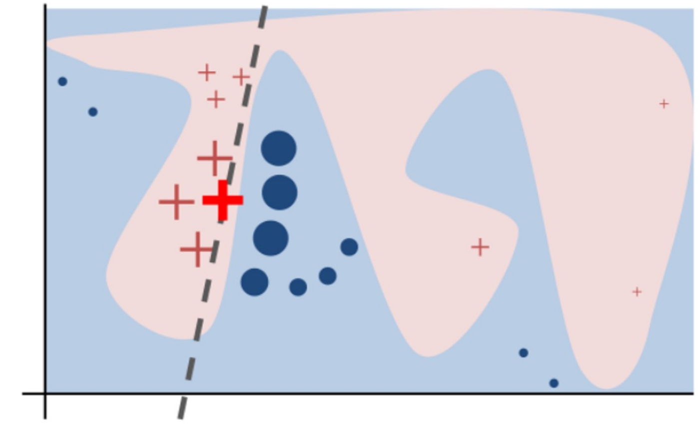
Black-box Methods

LIME

Local linear approximation, weighting perturbed points by **proximity**

Additive attribution (each feature contributes additively to the outcome)

Local fidelity, i.e. explained features might differ from one sample to the other (as opposed to global explanations)



$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \underbrace{\mathcal{L}\left(f, g, \pi_x\right)}_{\text{local approximation}} + \Omega(g)$$

family of interpretable models

regularization

Ribeiro et al. "" *Why should i trust you?" Explaining the predictions of any classifier*." ACM SIGKDD 2016.

LEMNA

Fused lasso (penalty that forces relevant/adjacent features to be grouped together to give meaningful explanations)

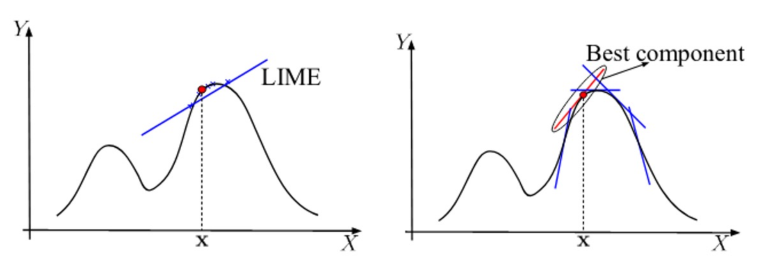
Mixture regression model (combines different linear models to approximate more complex functions)

$$L(f(\mathbf{x}), y) = \sum_{i=1}^{N} \| f(\mathbf{x}_i) - y_i \|$$

$$\text{subject to } \underbrace{\sum_{j=2}^{M} \left\| \beta_{kj} - \beta_{k(j-1)} \right\|}_{\text{fused lasso regularization}} \leq S, k = 1, \ldots, K$$

$$f(x) = \sum_{j=1}^{K} \pi_j \left(\beta_j \cdot x + \epsilon_j \right)$$

weighted sum of K linear models



(a) Linear regression model. (b) Mixture regression model.

Guo et al. "Lemna: Explaining deep learning based security applications." ACM SIGSAC 2018.

SHAP

Additive attribution method (like LIME)

Trains a model with and without subsets of features, compares the difference in performance (and then weight features based on all differences observed)

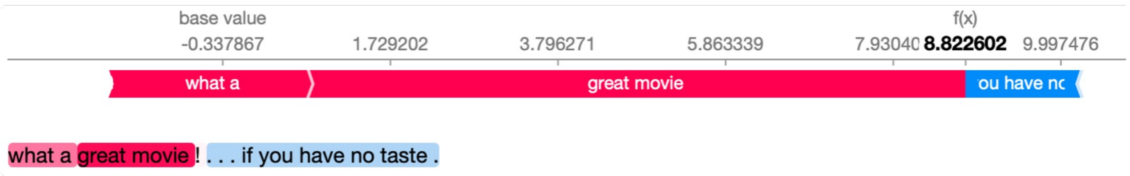
Finds out the **marginal contribution** of each feature and feature sets

Weights the features by the **information they contain**, rather than the proximity

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \boxed{\frac{|z'|!\,(M - |z'| - 1)!}{M!}} \overbrace{\left[f_x(z') - f_x(z' \backslash i)\right]}^{\text{difference in outcome}}$$

subset of features

weighting term
(how many features are in the subset)



Lundberg et al. "*A unified approach to interpreting model predictions.*" *NeurIPS* 2017.

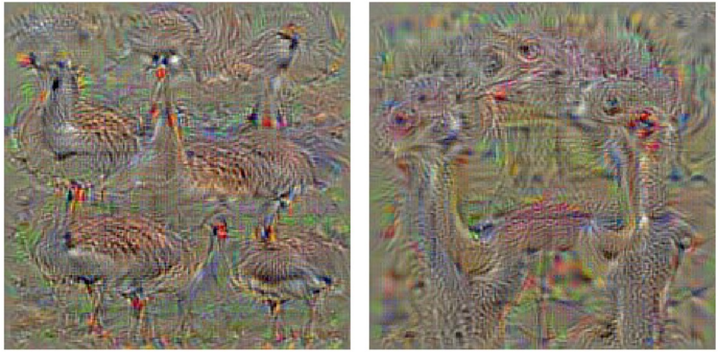
White-box Methods

Explaining using Gradients

Compute **gradients of the output class** w.r.t. the input

- Can be unstable/not very informative!

$$r_i = \frac{\partial y}{\partial x_i}$$



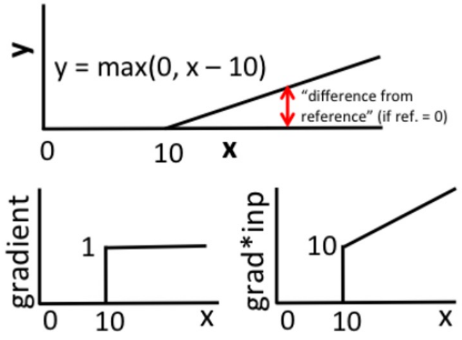
goose ostrich

Simonyan et al. "*Deep inside convolutional networks: Visualising image classification models and saliency maps*." *preprint arXiv:1312.6034* (2013).

Gradients x input, a.k.a. Linear Approximation

Decomposes the output on a specific input by backpropagating the contributions of all neurons to every feature

$$r_i = \frac{\partial y}{\partial x_i} x_i$$



Shrikumar et al. "*Learning important features through propagating activation differences*." *ICML*, 2017.

Integrated Gradients

Improves the linear approximation by referring to a **counterfactual baseline input**

Accumulates the gradients along the path

$$r_i = (x_i - x_i') \int_0^1 \frac{\partial f_N \left(x' + \alpha \left(x - x'\right)\right)}{\partial x_i} \, \mathrm{d}\alpha$$

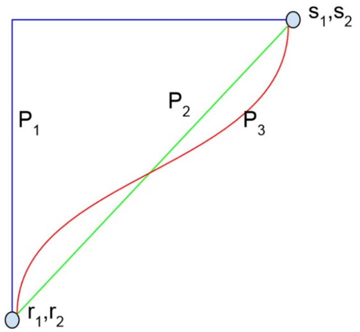
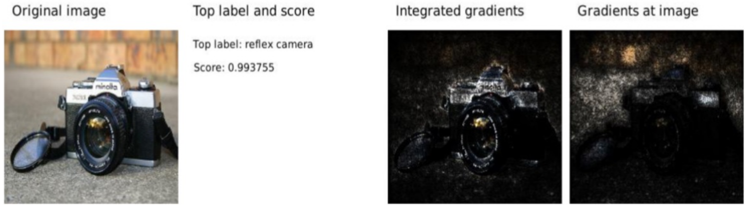


Figure 1. Three paths between an a baseline (r_1, r_2) and an input (s_1, s_2). Each path corresponds to a different attribution method. The path P_2 corresponds to the path used by integrated gradients.



Sundararajan et al. "*Axiomatic attribution for deep networks.*" *ICML*, 2017.

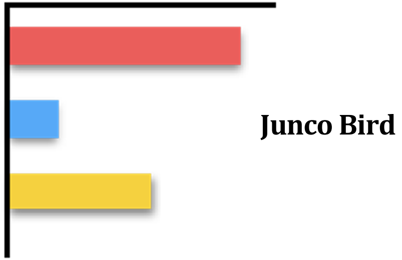
SmoothGrad
Smilkov et. al. 2017

Input

Model

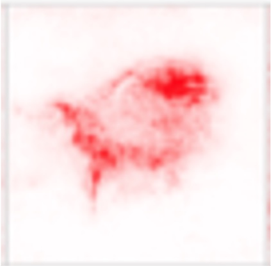
F_θ

Predictions



$$\frac{1}{N}\sum_{i}^{N}\nabla_{(x+\epsilon)}F_i(x+\epsilon)$$

Gaussian noise

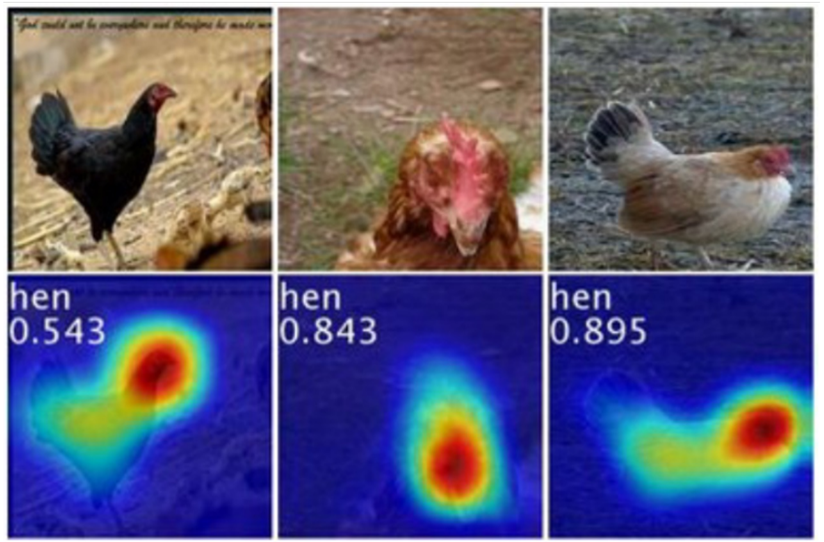


Average Input-gradient of 'noisy' inputs

Slide from: H. Lakkaraju, Interpreting Machine Learning Models: State-of-the-art, Challenges, Opportunities - 2022

Model Inspection: Class Activation Maps (CAM)

- Scale features from the last hidden layer with the weight connecting them to the desired output node
- Simple method, but often saturates and creates useless maps



Zhou et al. "*Learning Deep Features for Discriminative Localization*". *CVPR* 2016

Model Inspection: Layer-wise Relevance Propagation (LRP)

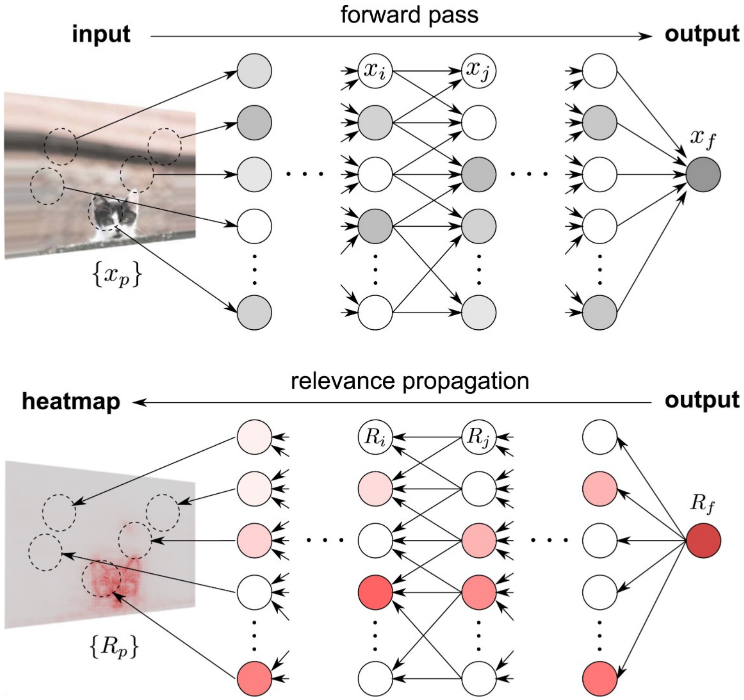
- A map that assigns a value to each feature, representing the effect of that input being set to a reference value (usually zero), as opposed to its original value

$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_{i'} z_{i'j}} R_j^{(l+1)} \quad \text{with} \quad \mathrm{z_{ij}} = \mathrm{x_i^{(l)} w_{ij}^{(l,l+1)}}$$

relevance at next layer

influence of the single neuron
w.r.t. the sum of layer neurons

relevance at current layer



Bach et al. "*On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation.*" *PloS one* 2015.
Image source: *Montavon et al. "Explaining nonlinear classification decisions with deep taylor decomposition." Pattern Recognition*, 2017.

Prototype-based Methods

Prototype-based methods

Goal: to identify training points most responsible for a given prediction

Influence function: how would the model's predictions change if we did not have this training point?

$$\hat{\theta}_{\epsilon,z} \stackrel{\text{def}}{=} \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} L(z_i,\theta) + \epsilon L(z,\theta)$$

$$\begin{aligned}\mathcal{I}_{\text{up,loss}}(z, z_{\text{test}}) &\stackrel{\text{def}}{=} \left.\frac{dL(z_{\text{test}}, \hat{\theta}_{\epsilon,z})}{d\epsilon}\right|_{\epsilon=0} \\ &= \nabla_\theta L(z_{\text{test}}, \hat{\theta})^\top \left.\frac{d\hat{\theta}_{\epsilon,z}}{d\epsilon}\right|_{\epsilon=0} \\ &= -\nabla_\theta L(z_{\text{test}}, \hat{\theta})^\top H_{\hat{\theta}}^{-1} \nabla_\theta L(z, \hat{\theta}).\end{aligned}$$

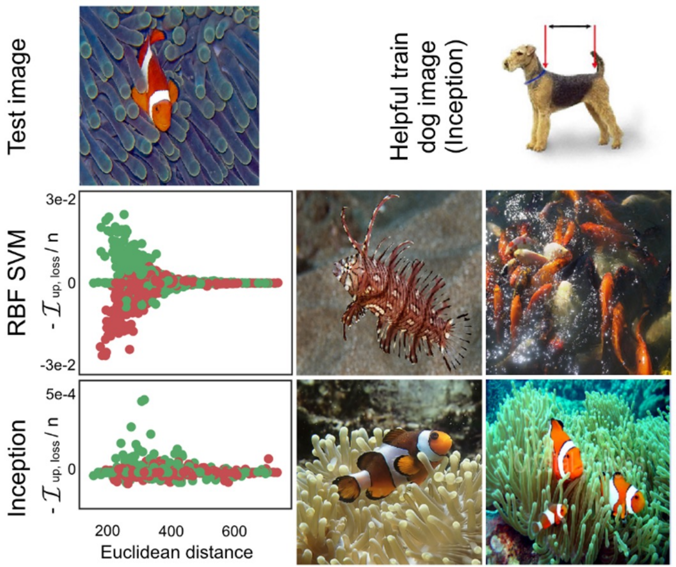


Figure 4. **Inception vs. RBF SVM. Bottom left:** $-\mathcal{I}_{\text{up,loss}}(z, z_{\text{test}})$ vs. $\|z - z_{\text{test}}\|_2^2$. Green dots are fish and red dots are dogs. **Bottom right:** The two most helpful training images, for each model, on the test. **Top right:** An image of a dog in the training set that helped the Inception model correctly classify the test image as a fish.

Koh et al. "*Understanding black-box predictions via influence functions.*" *ICML,* 2017.

Counterfactual Explanations

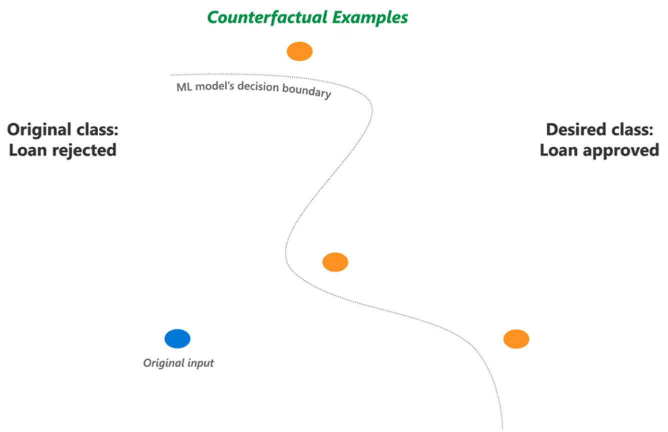
Hypothetical examples that show how to obtain a different prediction (using "*intervention*")

Found with adversarial techniques

Feasibility of the counterfactual actions given user context and constraints

Diversity among the counterfactuals presented (different solutions)

$$\boldsymbol{c} = \arg\min_{\boldsymbol{c}} y \, \mathrm{loss}(f(\boldsymbol{c}), y) + |\boldsymbol{x} - \boldsymbol{c}|$$



Wachter et al. "*Counterfactual explanations without opening the black box: Automated decisions and the GDPR*".
Image source: Mothilal et al. "*Explaining machine learning classifiers through diverse counterfactual explanations.*" ACM FaccT. 2020.

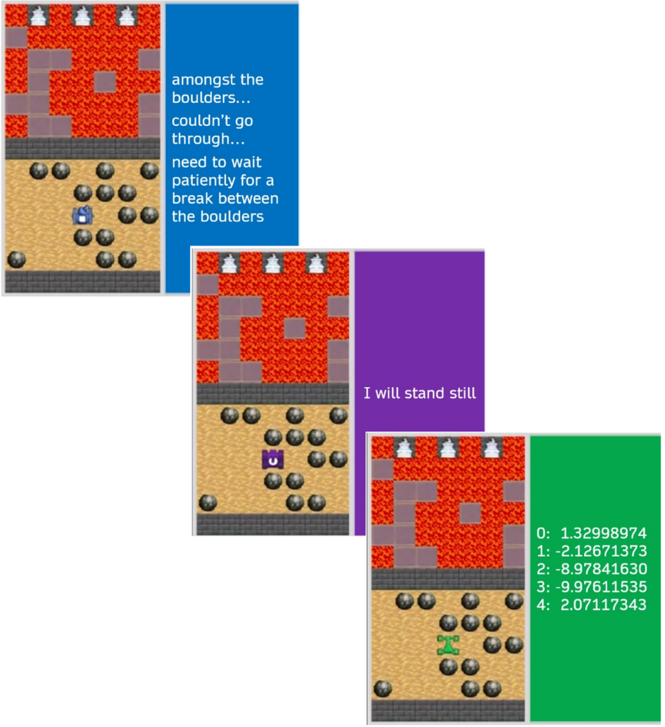
Final Remarks

Summary

- There is a great variety of explainability methods
- They have been tested **predominantly on images and text**
- but… there is no clear definition of what **explainability** is and how to measure it
  - How do you quantify if a method is "explainable"?
- Cynthia Rudin: **Please stop doing Explainable ML!**
  - https://www.youtube.com/watch?v=I0yrJz8uc5Q

Human-centric xAI

- Study on how the explanations provided by AI are perceived by who opens the "black box"
- Studies how two different groups, with and without background in AI, **perceive** the explanations
- Aims towards **tailoring** the explanations to the public that is using them



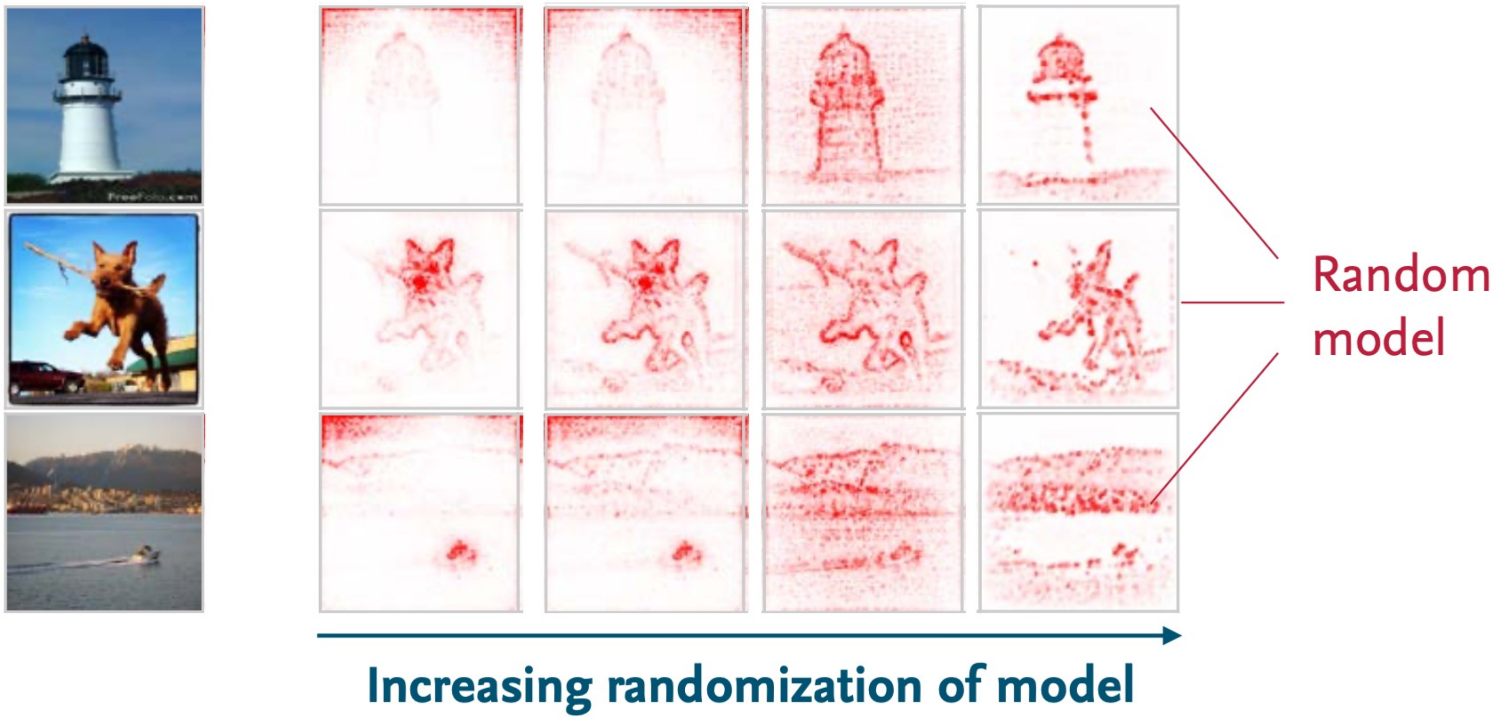
Ehsan et al. "*The who in explainable ai: How ai background shapes perceptions of AI explanations.*" *arXiv:2107.13509*, 2021.

Limitations: Adversarial Attacks against Explanations

- Explanations are not robust to adversarial attacks
- The sample can be manipulated in a way that creates an **arbitrary explanation**

Dombrowski et al. "*Explanations can be manipulated and geometry is to blame.*" *NeurIPS 2019*.

Limitations: Yet Another Sanity Check…



Adebayo et al., NeurIPS 2018