Pattern Recognition
and Applications Lab

Explainable AI

Battista Biggio, Maura Pintor, Ambra Demontis

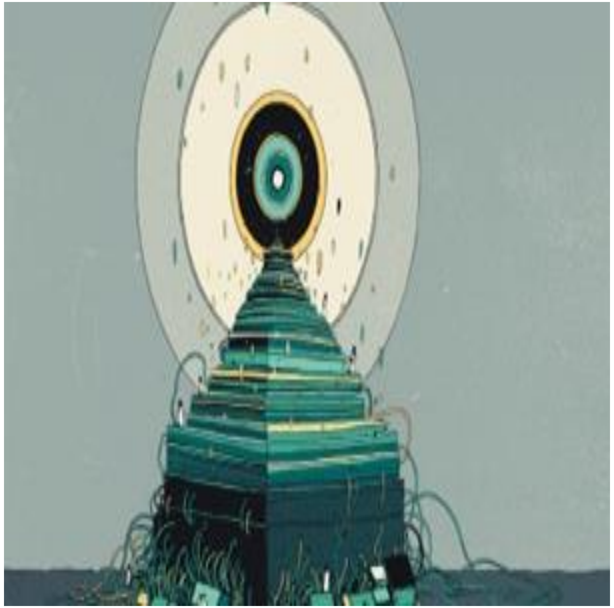
Department of Electrical and Electronic Engineering
University of Cagliari, Italy

When and Why Model Understanding?

ML is increasingly being employed in complex high-stakes settings.
According with the AI EU Act the high risk applications includes health and recruitment.

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

Causality: Check that only causal relationships are picked up

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

Explainability and Why It Is Important

It is not simple to understand if a model has learned something useful looking at the accuracy.

Classifier: Logistic Regression

Training - 20 images

Test - 10 images:

- 2 error;
- 8 correct;

Would you trust this AI only looking at the results?

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

Explainability and Why It Is Important

It is not simple to understand if a model has learned something useful looking at the accuracy.

Classifier: Logistic Regression

Training - 20 images

Test - 10 images:

- 2 error;
- 8 correct;

(a) Husky classified as wolf (b) Explanation

Would you trust this AI only looking at the results?

Would you still trust it if I show you to what the classifier is giving importance to decide?

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

Explainability and Why It Is Important

It is not simple to understand if a model has learned something useful looking at the accuracy.

Classifier: Logistic Regression

Training - 20 images (husky + snow, wolf + not snow)

Test - 10 images:

- 1 wolf not on snow (error);
- 1 husky on snow (error);
- 8 wolf on snow, husky not on snow (correct)


(a) Husky classified as wolf


(b) Explanation

Would you trust this AI only looking at the results?

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

Explainability and Why It Is Important

Fairness: Ensuring that predictions are unbiased

Causality: Check that only causal relationships are picked up

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

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

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)

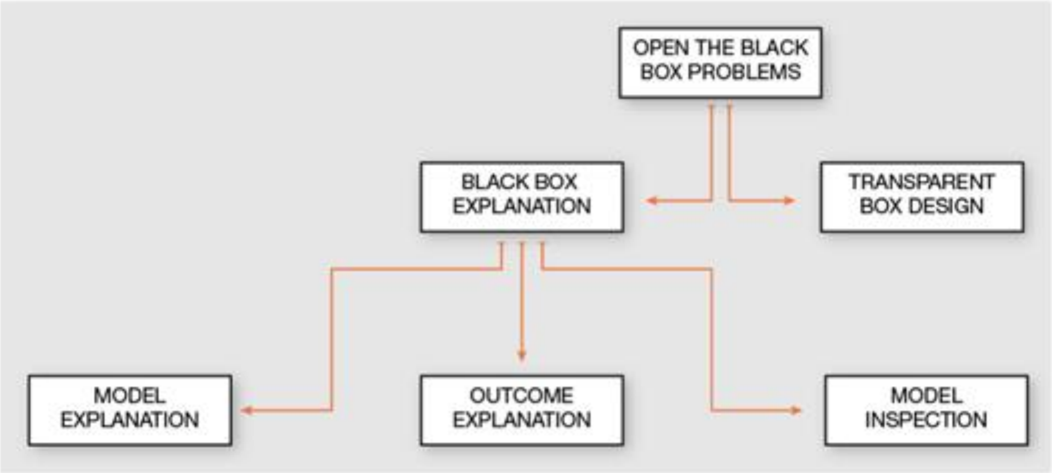
Regulatory agencies (e.g., FDA, European commission)

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



The goals of

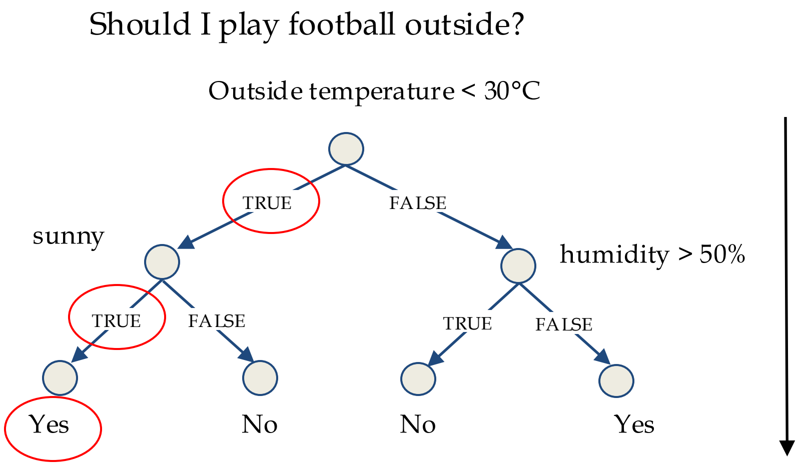
Model explanation: understanding the whole model logic

Outcome explanation: understanding the decision for a specific sample

Model inspection: understanding how the internal model behavior changes when the sample is modified

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

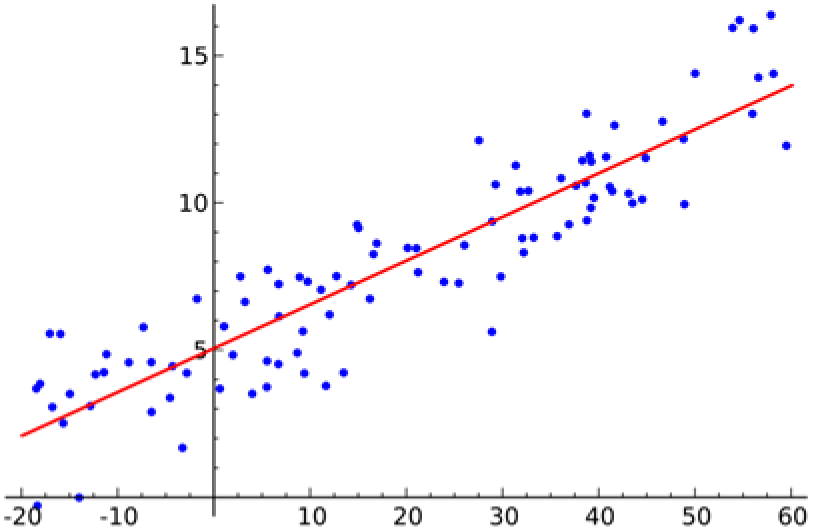
Interpretable-by-Design (Transparent) Models



Depth = how many levels of decision

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

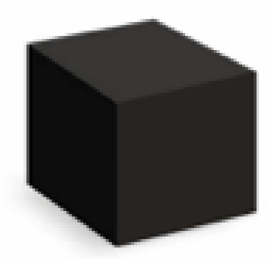
Interpretable-by-Design (Transparent) Models

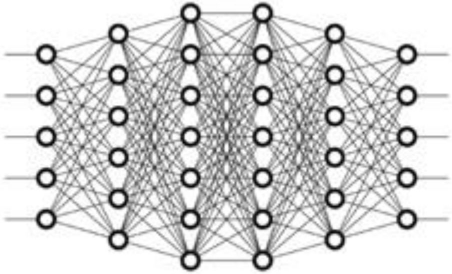


Even **linear** classifiers may be hard to interpret when dealing with high-dimensional problems

Black-box Explanation

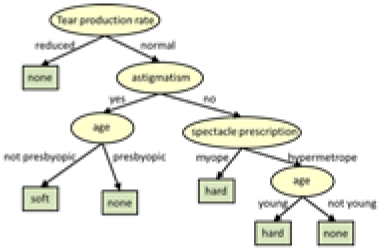
Explain pre-built models *in a post-hoc manner*





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

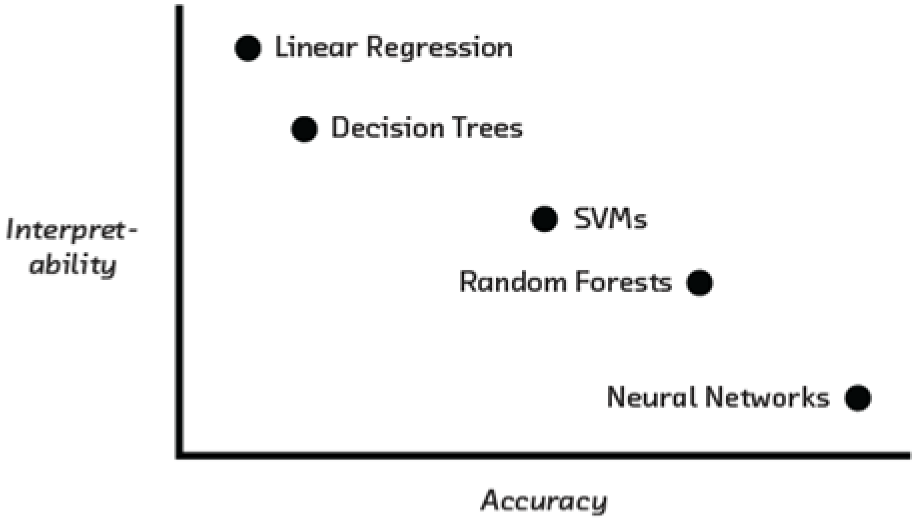


Ribeiro et. al. 2016, Ribeiro et al. 2018; Lakkaraju et. al. 2019

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

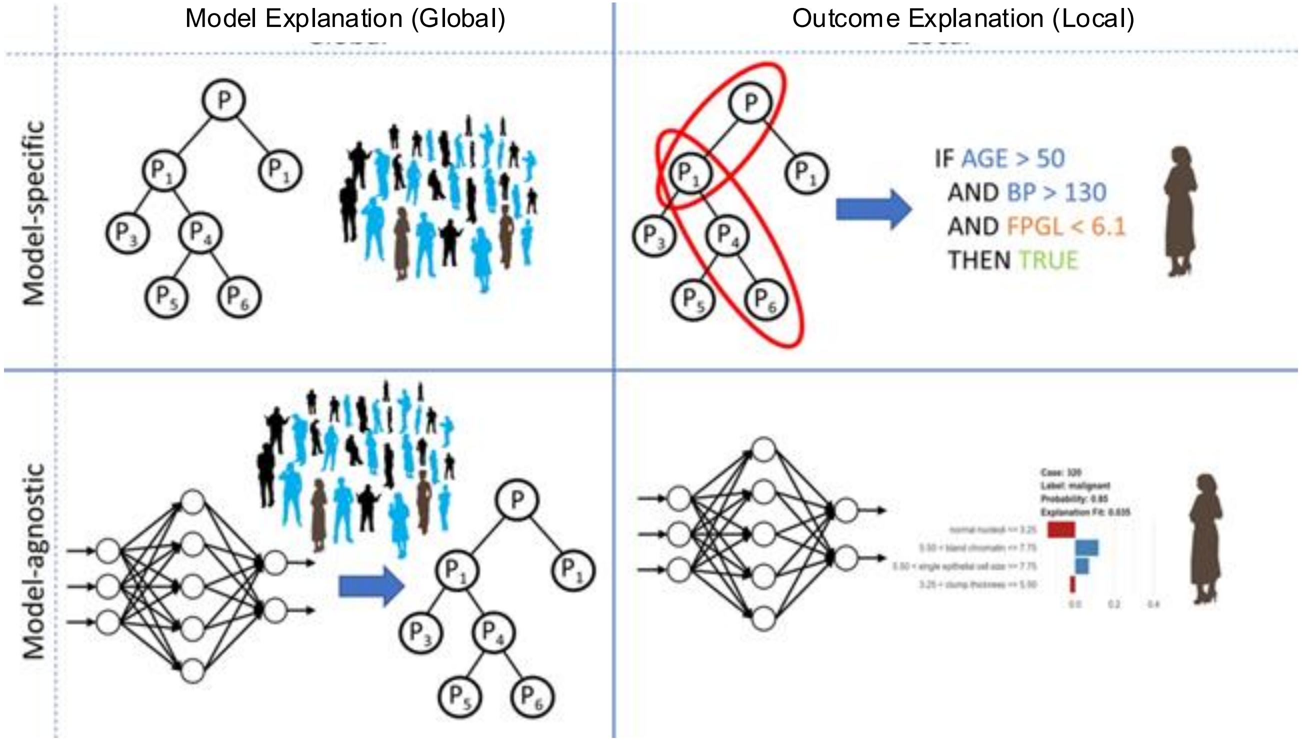


Myth or Reality?

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

Slide from: H. Lakkaraju, Interpreting Machine Learning Models: State-of-the-art, Challenges, Opportunities - 2022
Cynthia discussion: https://www.youtube.com/watch?v=I0yrJz8uc5Q&t=329s

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

Local Explanations	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

Blue/pink background = decision function f of the black-box model

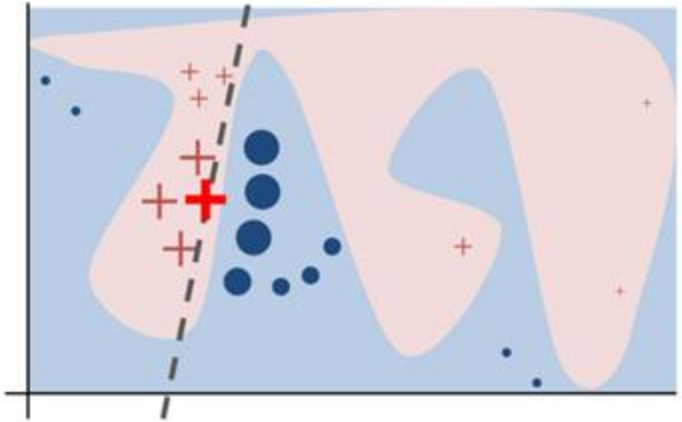
Bold red cross = instance being explained

cross and dot = instances that Lime:

1. samples
2. get predictions for them
3. weight them by the proximity (π) to the instance being explained (x).

dashed line = learned explainable model (g) that is locally (but not globally) faithful.

Problem: works well only if the decision region is locally-linear near to x.



$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

family of interpretable models

regularization

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

SHAP

Trains a model with and without subsets of features, compares the difference between the scores (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**.

Problem: Slow eg. for cybersecurity applications!

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \boxed{\frac{|z'|!\,(M - |z'| - 1)!}{M!}} \overbrace{\left[f_x(z') - f_x(z' \setminus i)\right]}^{\text{difference in outcome when the subset of features is removed}}$$

subset of features

weighting term
(how many features are in the subset)

M is the number of simplified features (total).

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

LEMNA

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

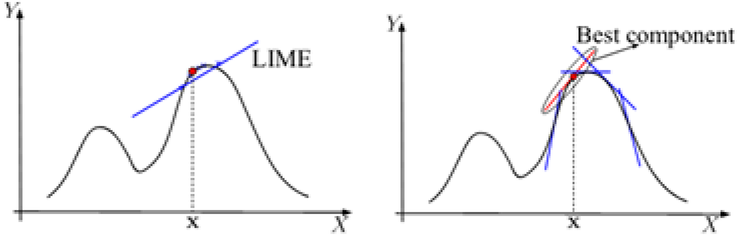
Fused lasso (penalty that forces relevant/adjacent features to be grouped together to give meaningful explanations)

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

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

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

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.

White-box Methods

Explaining using Gradients

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

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

Problem:

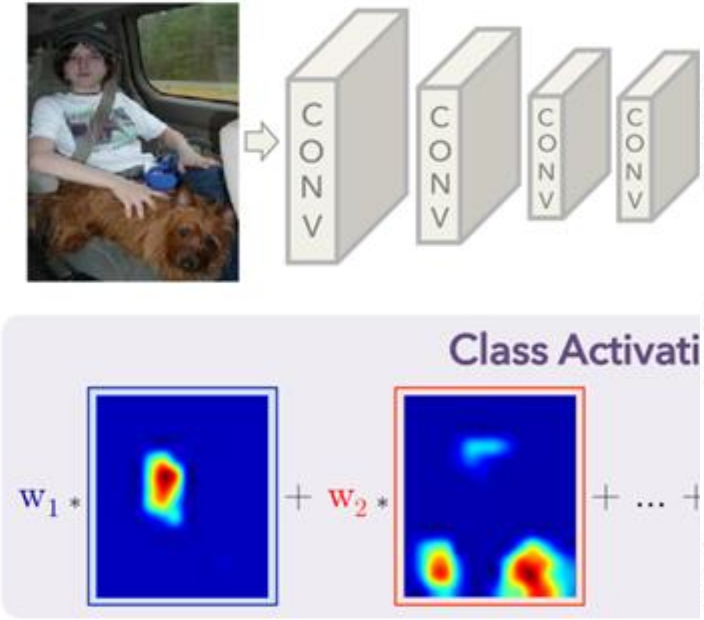
Suppose you want to explain the decision of a linear classifier with many features trained to distinguish between legitimate and spam emails.

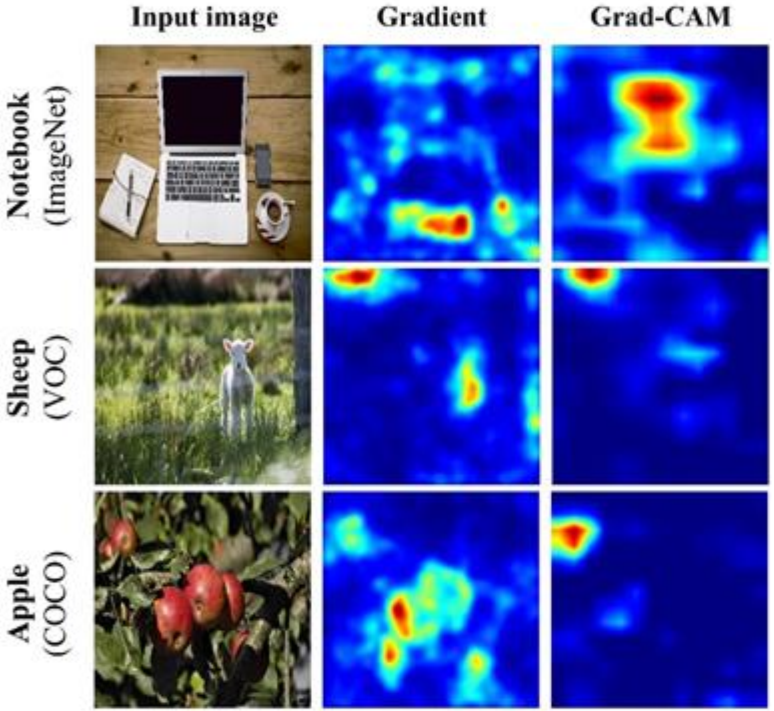
Suppose each feature represents the presence or absence of a feature.

If you employ just the gradient, **you may obtain a high relevance for features not present in the email you are trying to explain.**

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

Model Inspection: Class Activation Maps (CAM)





Problem: Not theoretically principled, thus m
highlight region that are not really meaningful.

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

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

Multiplies the gradient for the input.

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

Problem: it breaks a desired property called **Sensitivity** desired e.g., for images:
"Every input and baseline that differ in one feature but have different predictions then the differing feature should be given a non-zero attribution."

NB: Different application domain have different required properties!

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

Integrated Gradients

Considers the straightline path from a **baseline** x′ to the input x, and compute the gradients at all points along the path.

Integrated gradients are obtained by cumulating these gradients.

Specifically, are defined as the intergral of the gradients along the straightline path from the baseline x′ to the input x.

Problem: There can be noise due to essentially meaningless local variation in partial derivatives.

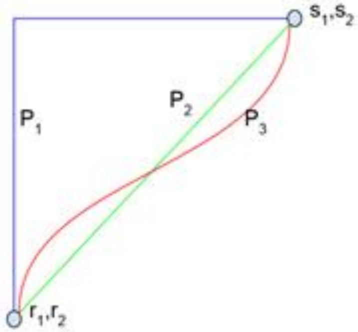


Figure 1. Three paths between an a baseline (r_1, r_2) and an input (s_1, s_2). Each path corresponds to a different attribution method. The path P_2 corresponds to the path used by integrated gradients.

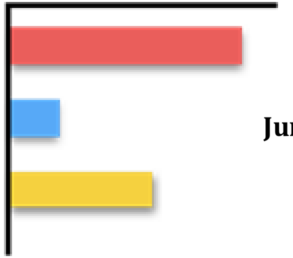
Sundararajan et al. "*Axiomatic attribution for deep networks.*" *ICML*, 2017.

SmoothGrad Smilkov et. al. 2017

Input

Model

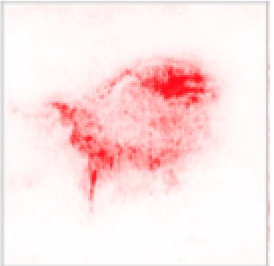
Predictions



Junco Bird

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

Gaussian noise



Average Input-gradient of ‘noisy’ inputs

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

Prototype-based Methods

Prototype-based methods

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

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

Computes how the network parameters θ changes when we upweight a sample z.

Using the chain rule computes how the parameters change influence the loss on a test point z_{test}.

Problem: tested only for linear classifier trained on top of a feature extractor

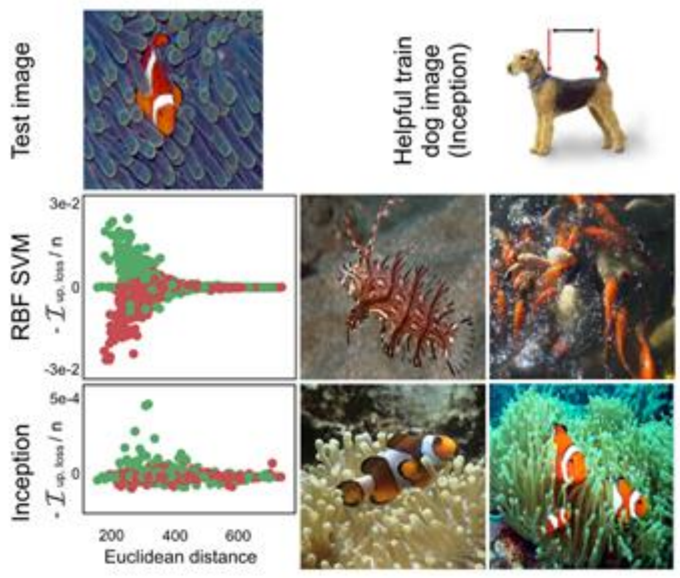


Figure 4. **Inception vs. RBF SVM. Bottom left:** $-\mathcal{I}_{\text{up,loss}}(z, z_{\text{test}})$ vs. $\|z - z_{\text{test}}\|_2^2$. Green dots are fish and red dots are dogs. **Bottom right:** The two most helpful training images, for each model, on the test. **Top right:** An image of a dog in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

Found with adversarial techniques promoting:

- **feasibility** of the counterfactual actions given user context and constraints

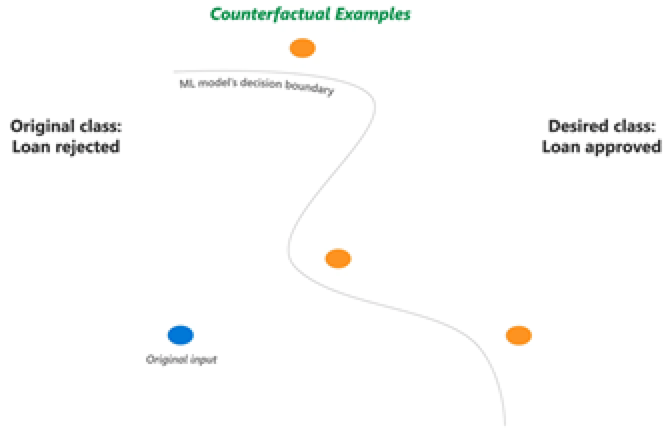
- **diversity** among the counterfactuals presented (different solutions)

$$\arg\min_{x'} \lambda(f_w(x') - y')^2 + d(x_i, x')$$

w are the weights

y’ is the target label != original;

λ is a constant that is increased until we find the counterfactual.



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.

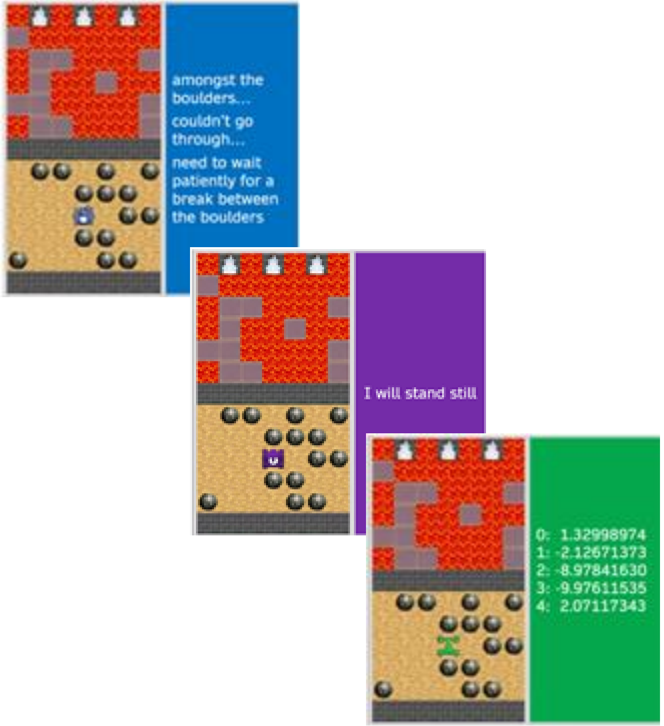
Final Remarks

Summary

- There is a great variety of explainability methods
- They have been tested **predominantly on images and text**
- Different domains have **different desired properties**
- 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

- The sample can be manipulated in a way that creates an **arbitrary explanation**

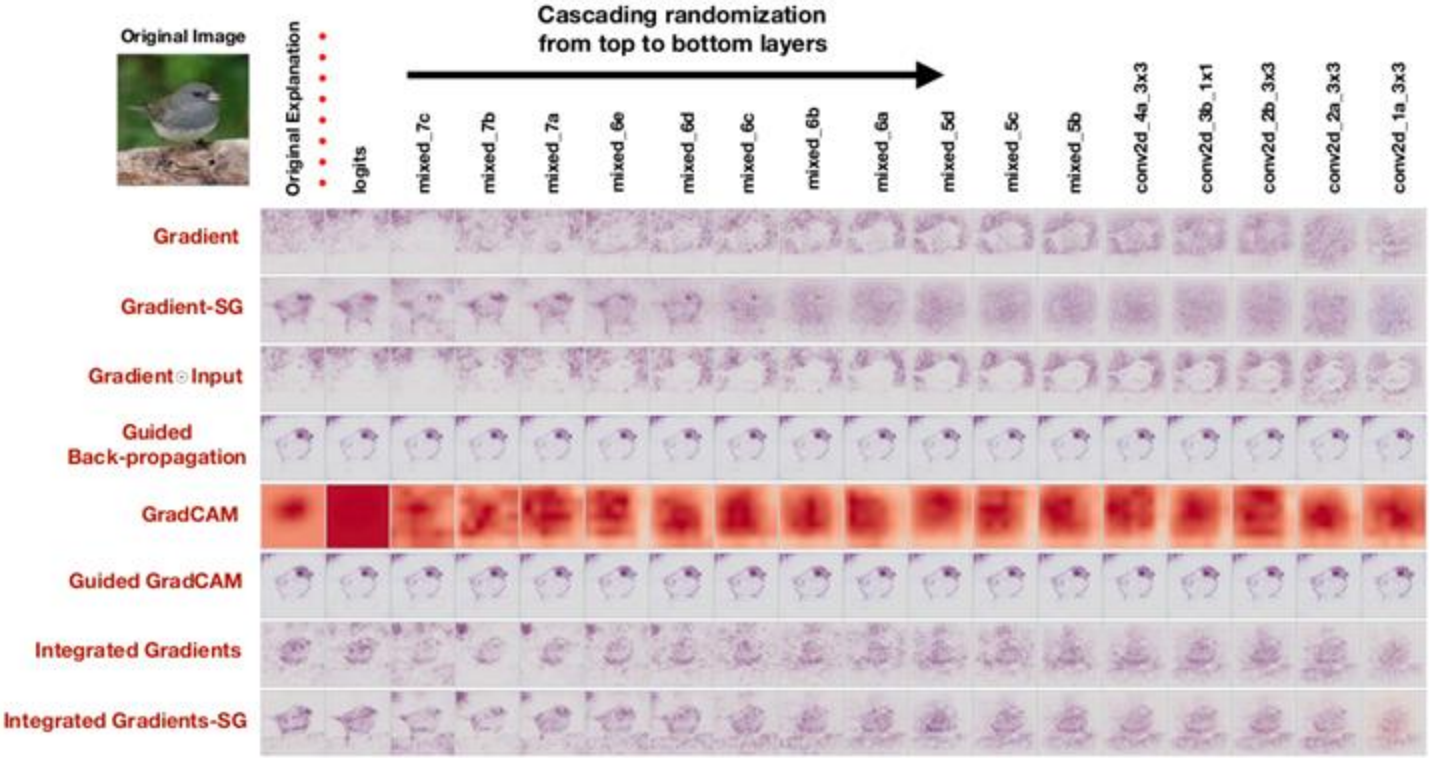
$$\mathcal{L} = \left\|h(x_{\text{adv}}) - h^t\right\|^2 + \gamma \left\|g(x_{\text{adv}}) - g(x)\right\|^2$$

- L is the loss optimized with respect to x_{adv} (the manipulated image)

- g is the classification function
- h^t is the target explanation
- h is the explanation function

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

Limitations: Yet Another Sanity Check…



Adebayo et al., NeurIPS 2018