Explainable and trustworthy Deep Learning

Danilo Numeroso

University of Pisa

What is Explainability?

It is the "ability to explain or to present in understandable terms to a human how a deep learning model makes prediction."
[Doshi-Velez and Kim, 2017]

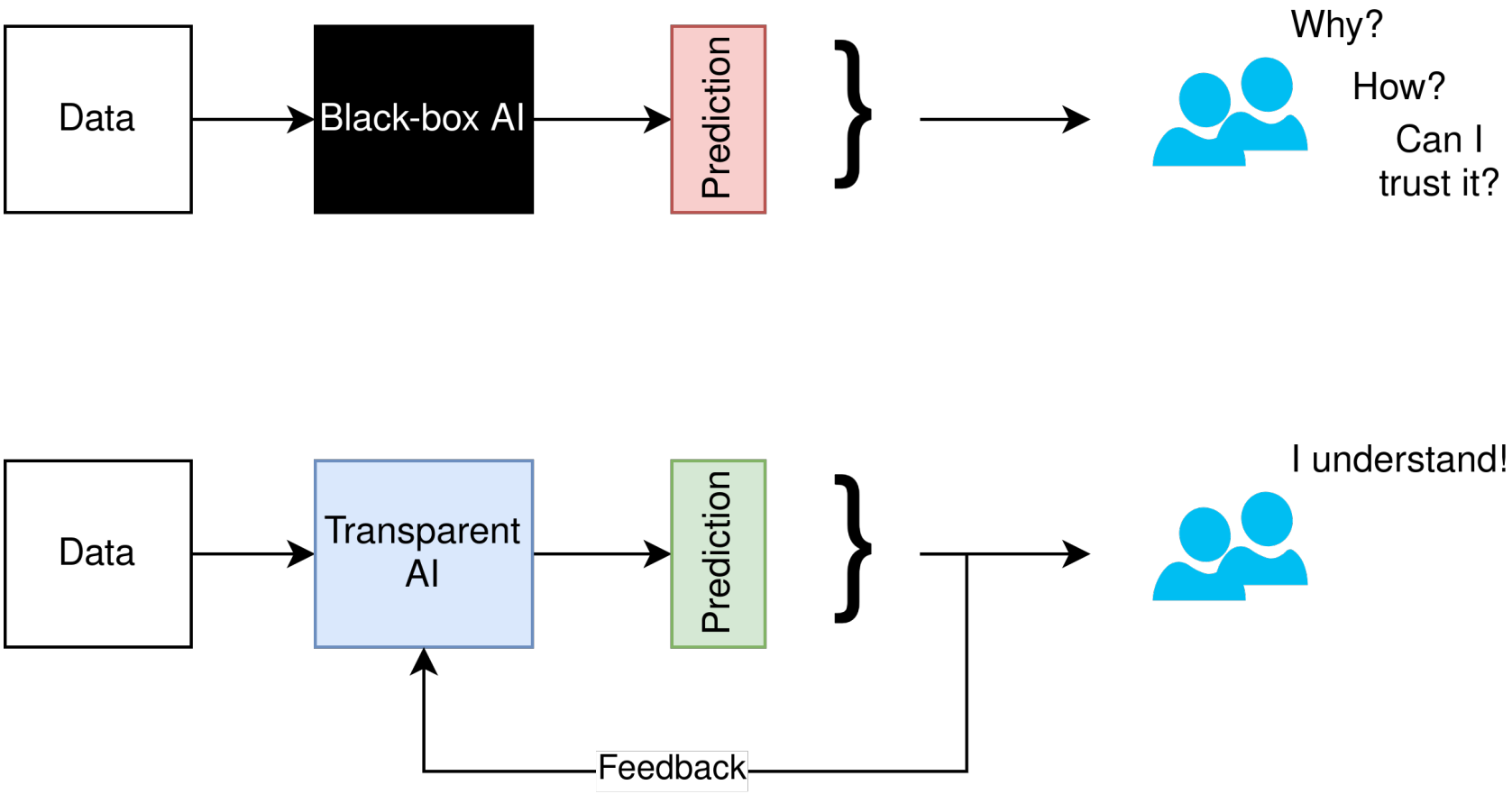
In literature many different shades (i.e, Accountability, Transparency, Explainability, Interpretability, ...).

What is the goal?

Explainable AI (XAI for shorts) aims to develop methods, methodologies and models to increase trustworthiness of AI systems.

The main objective is to produce interpretable explanations for certain predictions made by a black box models.

Black-box AI vs Transparent AI



Motivation (i)

Why should we care?

Motivation (ii)

Motivation (iii)

Researchers say use of artificial intelligence in medicine raises ethical questions

In a perspective piece, Stanford researchers discuss the ethical implications of using machine-learning tools in making health care decisions for patients.

MAR 14
2018

Artificial intelligence is hard at work crunching health data to improve diagnostics and help doctors make better decisions for their patients. But researchers at the Stanford University School of Medicine say the furious pace of growth in the development of machine-learning tools calls for physicians and scientists to carefully examine the ethical risks of incorporating them into decision-making.

Danton Char

In a perspective piece published March 15 in *The New England Journal of Medicine*, the authors acknowledged the tremendous benefit that machine learning can have on patient health. But they cautioned that the full benefit of using this type of tool to make predictions and take alternative actions can't be realized without careful consideration of the accompanying ethical pitfalls.

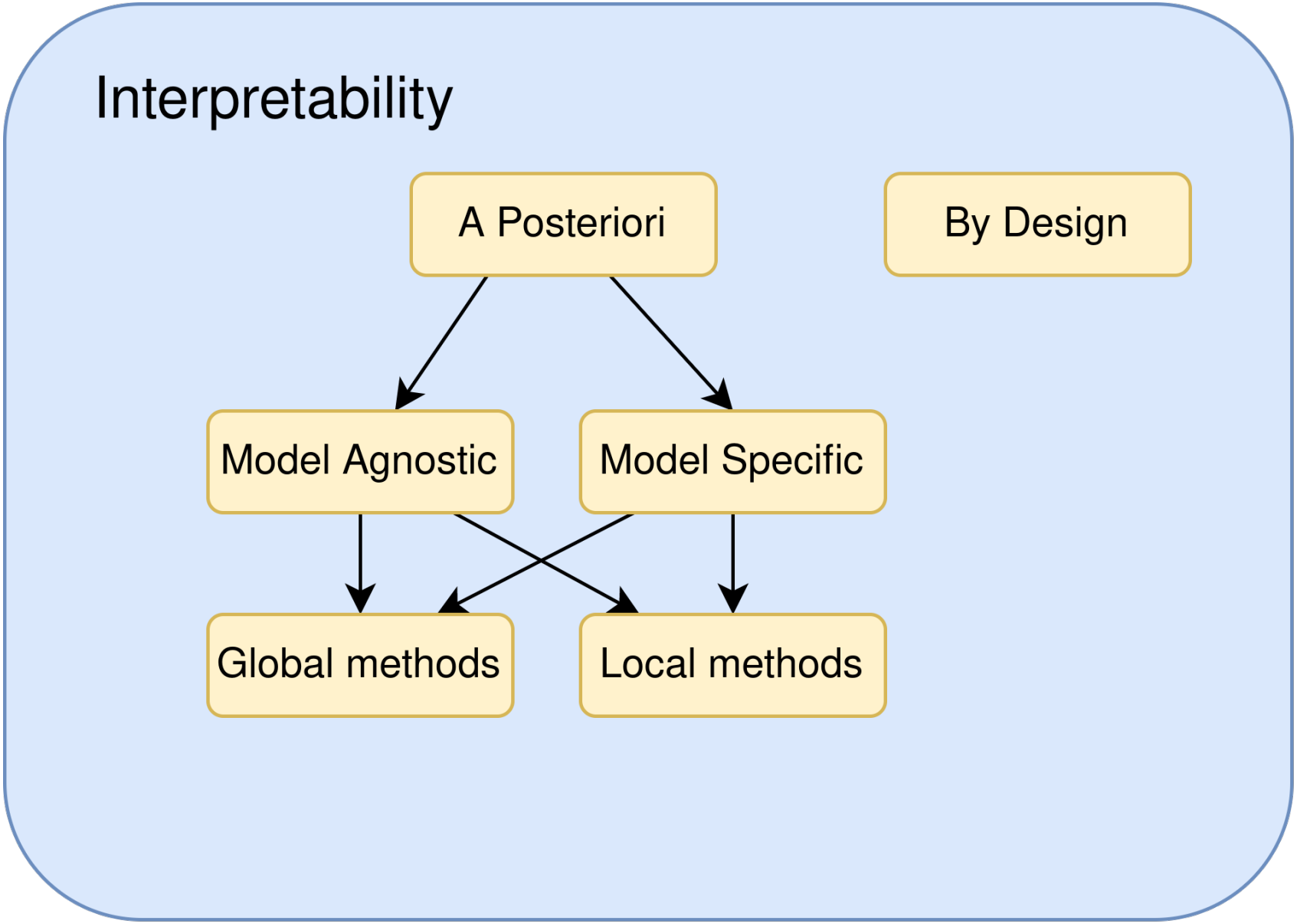
"Because of the many potential benefits, there's a strong desire in society to have these tools piloted and implemented into health care," said the lead author, Danton Char, MD, assistant professor of

Motivation (iv)

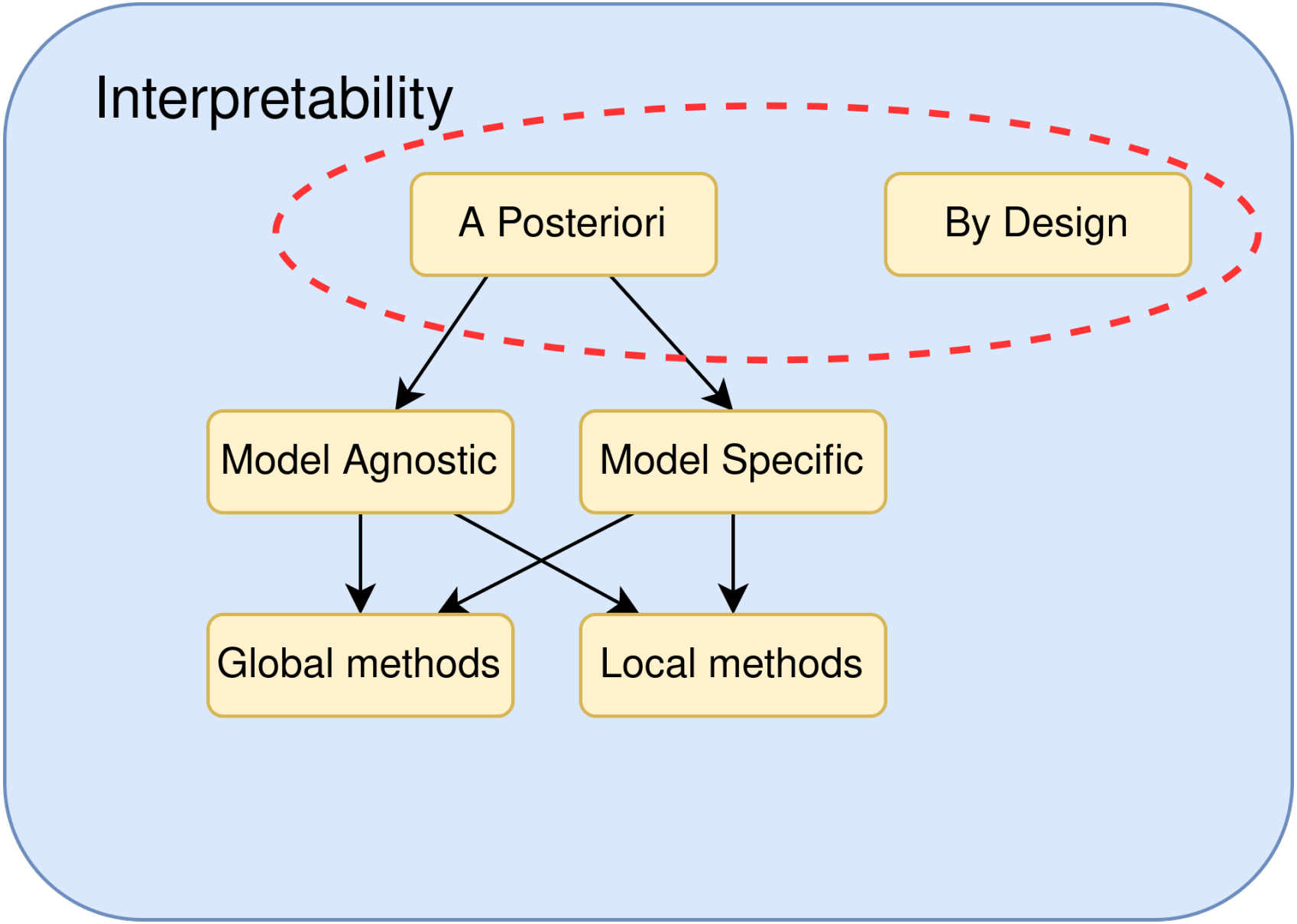
The ‘right to an explanation’ under EU data protection law

EUROPEAN DATA PROTECTION SUPERVISOR

Problem taxonomy

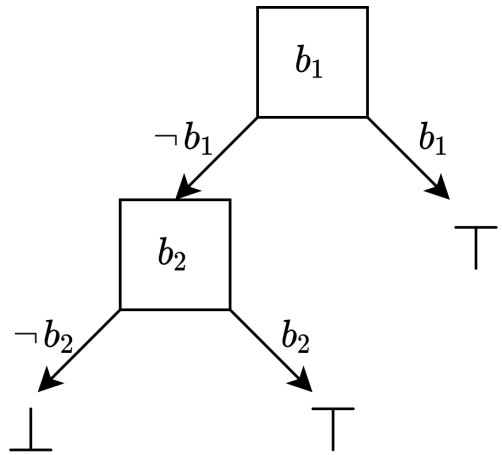


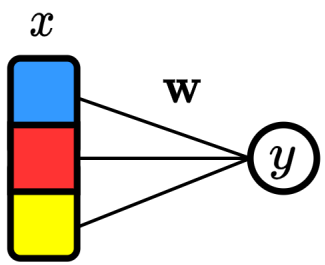
Problem taxonomy



White-box AI and Interpretability by-design

White-box (or transparent) models are a class of models that are inherently interpretable, i.e. linear regression, decision tree, k-nn classifier.





$$y = \square \cdot w_1 + \square \cdot w_2 + \square \cdot w_3$$

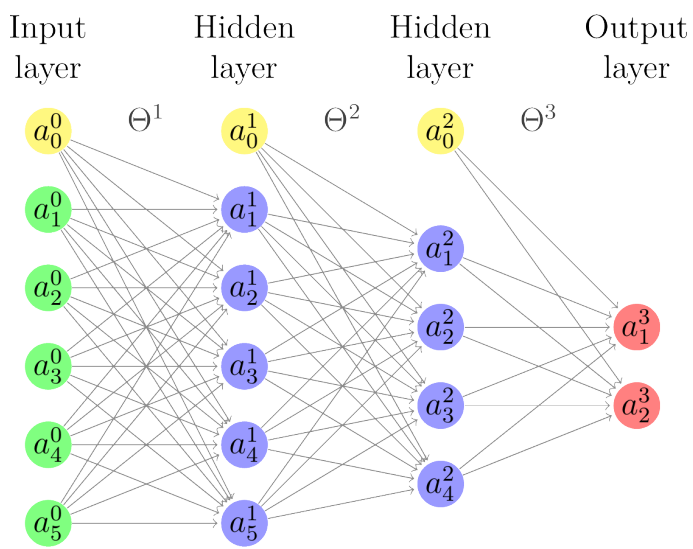
$$\bot = \neg b_1 \wedge \neg b_2$$

$$\top = b_1 \vee \neg b_1 \wedge b_2$$

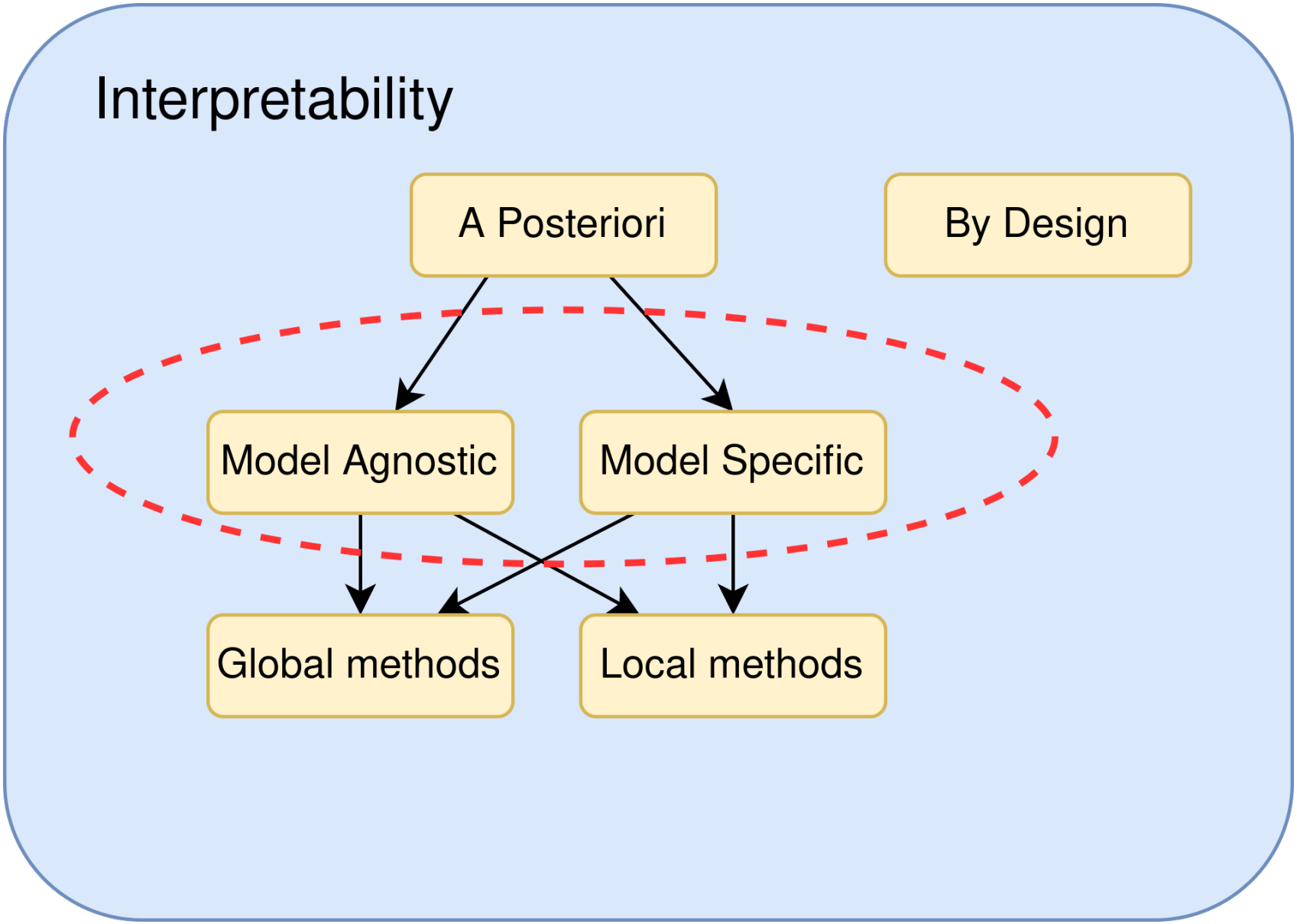
Black-box AI

As opposed to white-box models, a black box is a model, whose internal mechanisms are either unobservable or uninterpretable by humans, i.e. deep neural networks.

Such models usually require a posteriori explanations for their predictions.



Problem taxonomy



Model agnostic vs Model specific

Explainers may be:

1. Model-agnostic: the explanation process is independent from the model being explained, i.e. it can explain any model.
2. Model-specific: the explainer makes assumptions on the model topology, i.e. investigation of hidden representations.

Model agnostic vs Model specific

Explainers may be:

1. Model-agnostic: the explanation process is independent from the model being explained, i.e. it can explain any model.
2. Model-specific: the explainer makes assumptions on the model topology, i.e. investigation of hidden representations.

Model agnostic vs Model specific

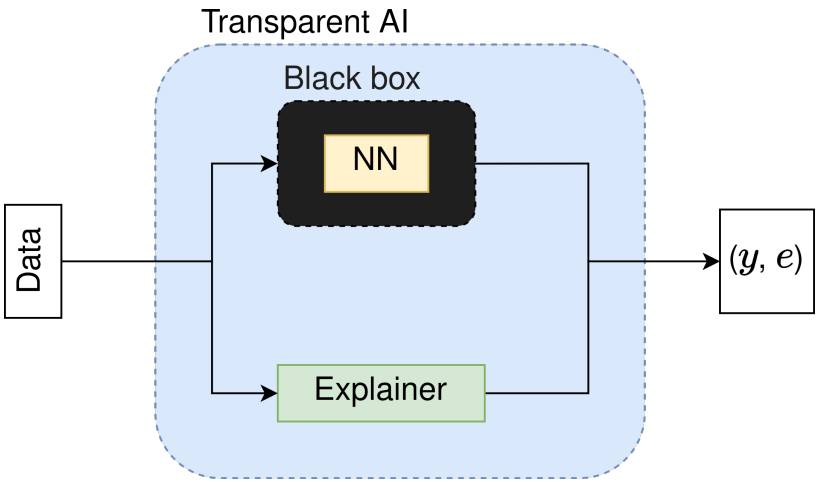


Figure 1: Model agnostic

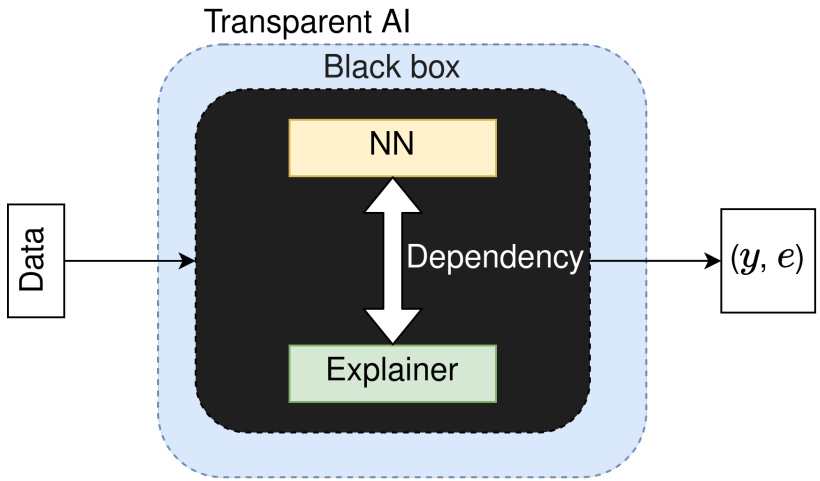
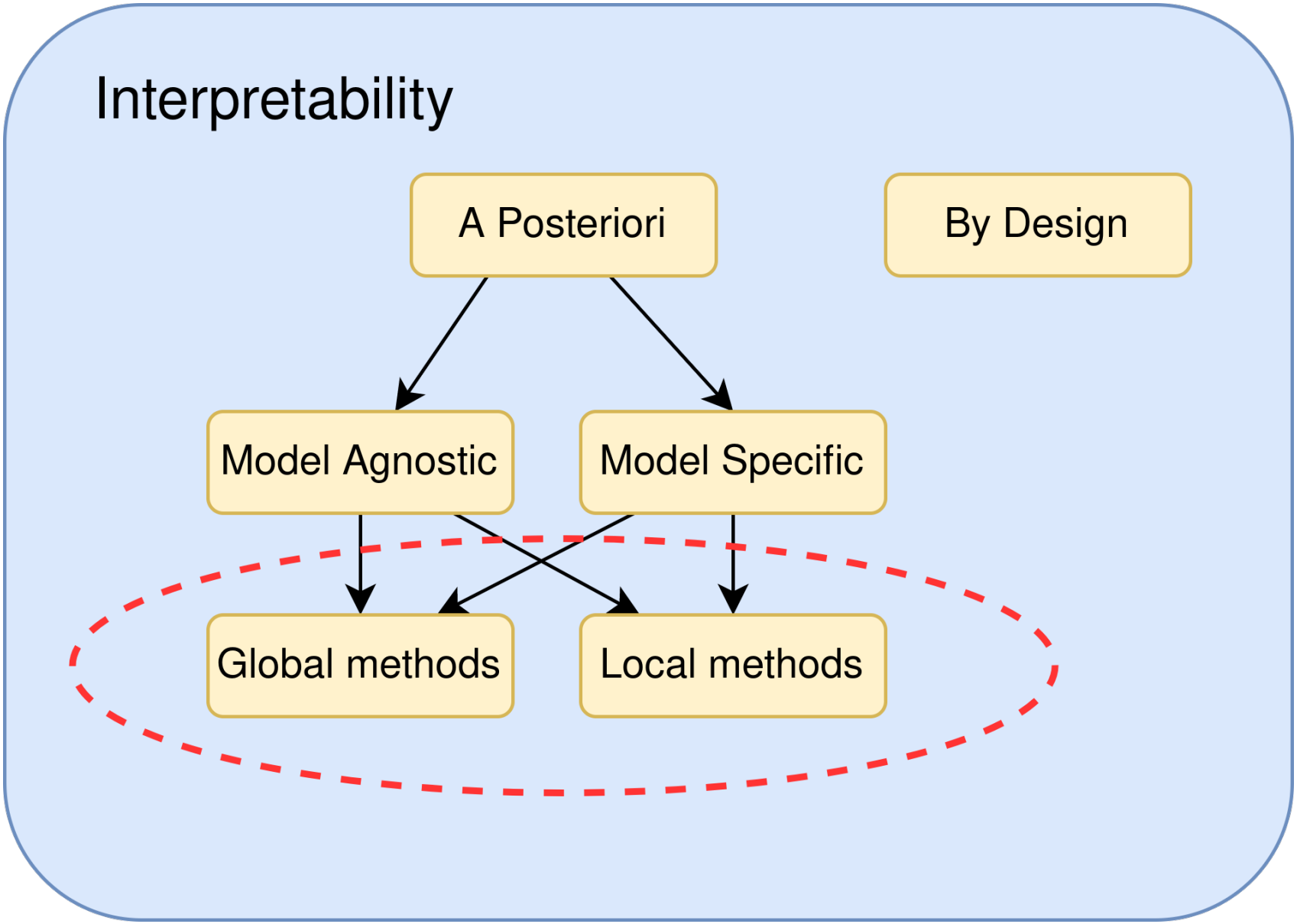


Figure 2: Model specific

Problem taxonomy



Locally interpretable vs Globally interpretable models

Explanations may be:

1. Model-level: explaining a complex model global behaviour thoroughly.
2. Local-level: emitting explanations for a particular (set of) prediction(s) made by the complex model under study.

Locally interpretable vs Globally interpretable models

Explanations may be:

1. Model-level: explaining a complex model global behaviour thoroughly.
2. Local-level: emitting explanations for a particular (set of) prediction(s) made by the complex model under study.

Category

Any a posteriori explainer can be categorised in:

1. Global Model agnostic
2. Global Model specific
3. Local Model agnostic
4. Local Model specific

How to get explanations?

Many methods:

- Sensitivity analysis
- Relevance Propagation
- Contrastive explanations
- Perturbation methods
- Counterfactual explanations
- Mimic learning
- Input Optimisation

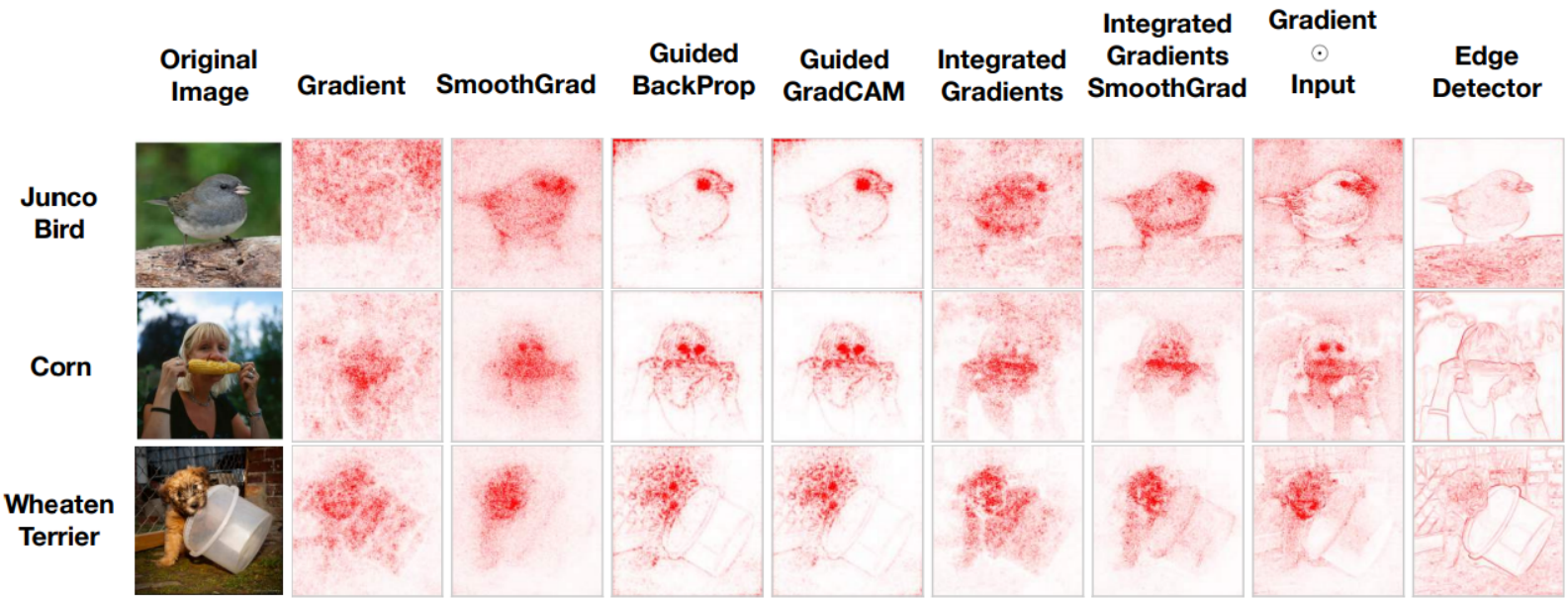
Sensitivity analysis (SA)

Main idea: creating feature heat-maps by exploiting the model gradients.

Given a predictor $f : \mathrm{R}^d \to \mathcal{Y}$ and an input x, take the derivative $\frac{df}{dx}$.

Several implementations: Saliency Maps [Simonyan et al., 2013], CAM [Zhou et al., 2015], Grad-CAM [Selvaraju et al., 2016], ...

Sensitivity analysis (SA)



[Adebayo et al., 2020]

LIME

Local Interpretable Model-agnostic Explanations (LIME) [Ribeiro et al., 2016] generates random noises for a given input x and train interpretable models g on the perturbed neighbourhood π_x.

Solves:

$$\text{argmin}_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

where, $\mathcal{L}$ is a measure of faithfulness and Ω a measure of complexity.

LIME

The explainer $g \in G$ can be any white-box model, e.g. linear regression.

Ω acts as regularisation for the explanation interpretability, e.g. number of non-zero weights in linear models.

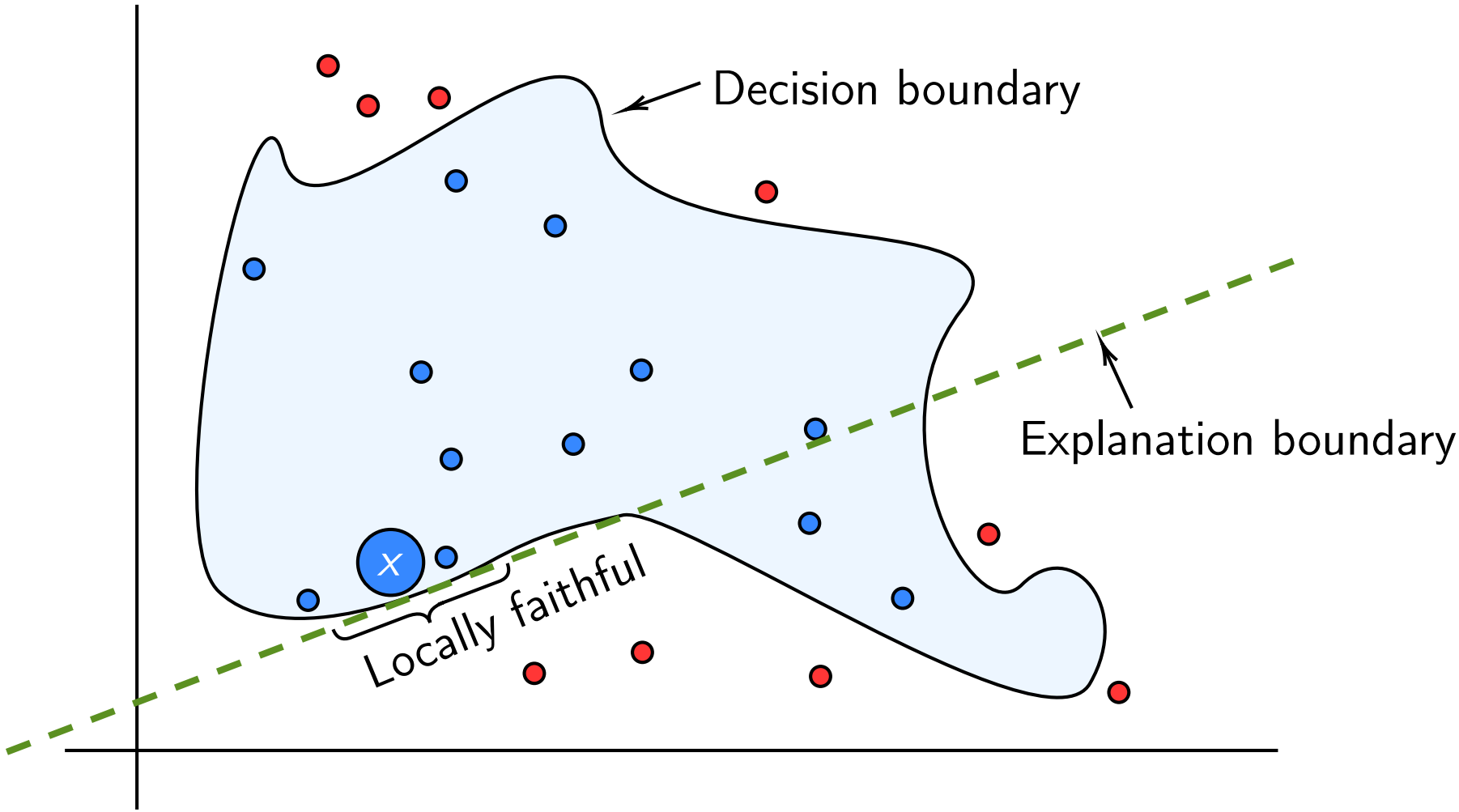
The main objective is to approximate the complex global behaviour in a locality of an input x.

LIME: Optimisation Problem

To learn explanations, LIME solve an optimisation problem.
Starting from an input x, generate perturbations $z_i \sim x \in \mathcal{Z}$.
Then, we plug the following loss in the previous objective function:

$$\mathcal{L}(f, g, \pi_x) = \sum_{z \in \mathcal{Z}} \pi_x(z)(f(z) - g(z))^2$$

LIME Example



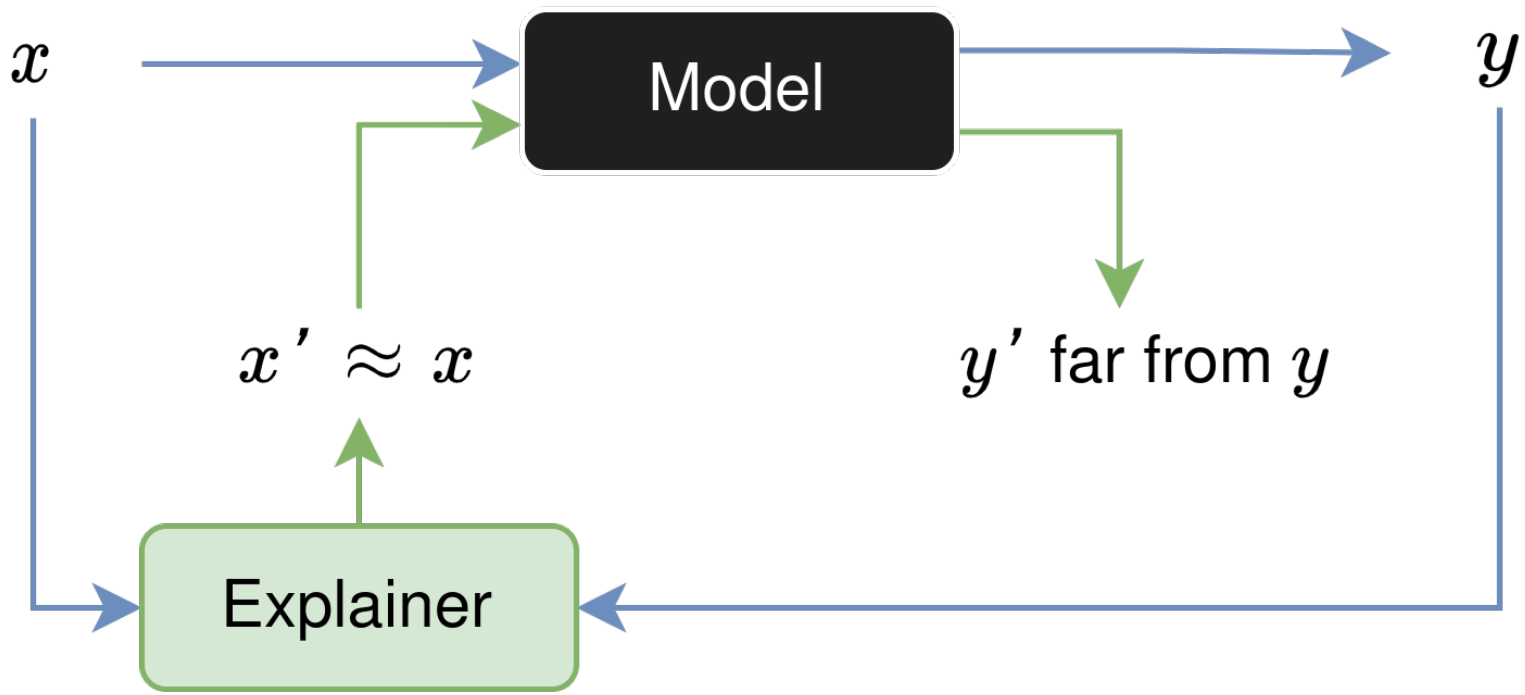
LIME

Drawbacks:

- What if the learnt decision boundary is highly non-linear?
- Explanations are dependent from hyperparameters.

Counterfactual Explanations

Counterfactual Explanations highlights minimal changes, i.e. perturbation, needed to change the decision of the predictor radically.

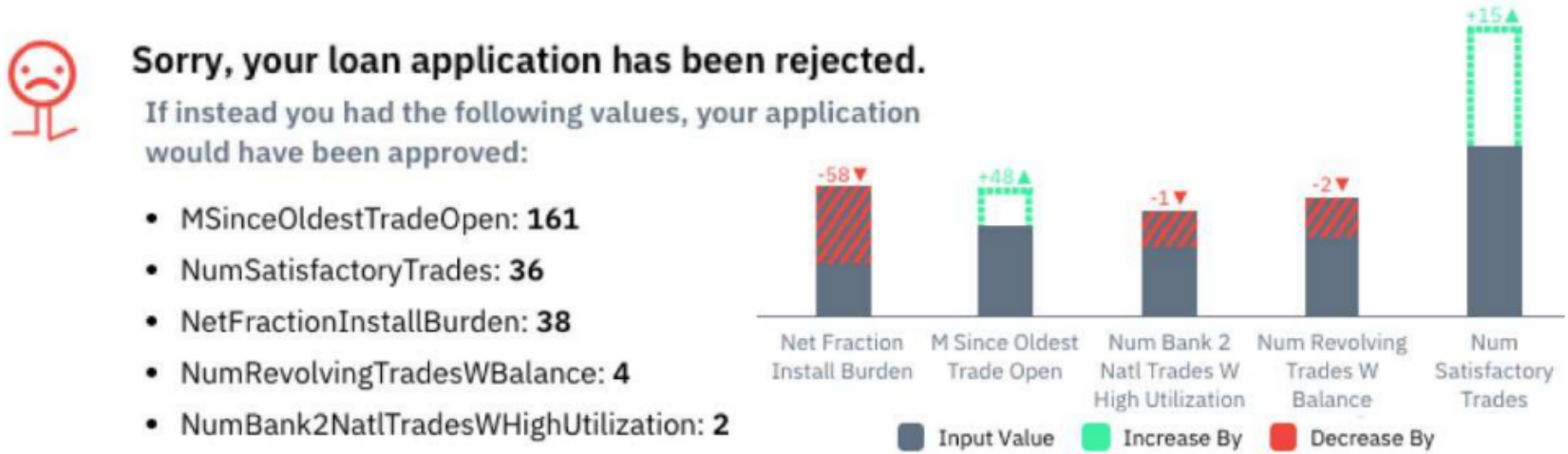


Counterfactual Explanations: Example

Settings: Deep Learning model deciding whether to accept or not loan applications.

Outcome: You have been rejected.

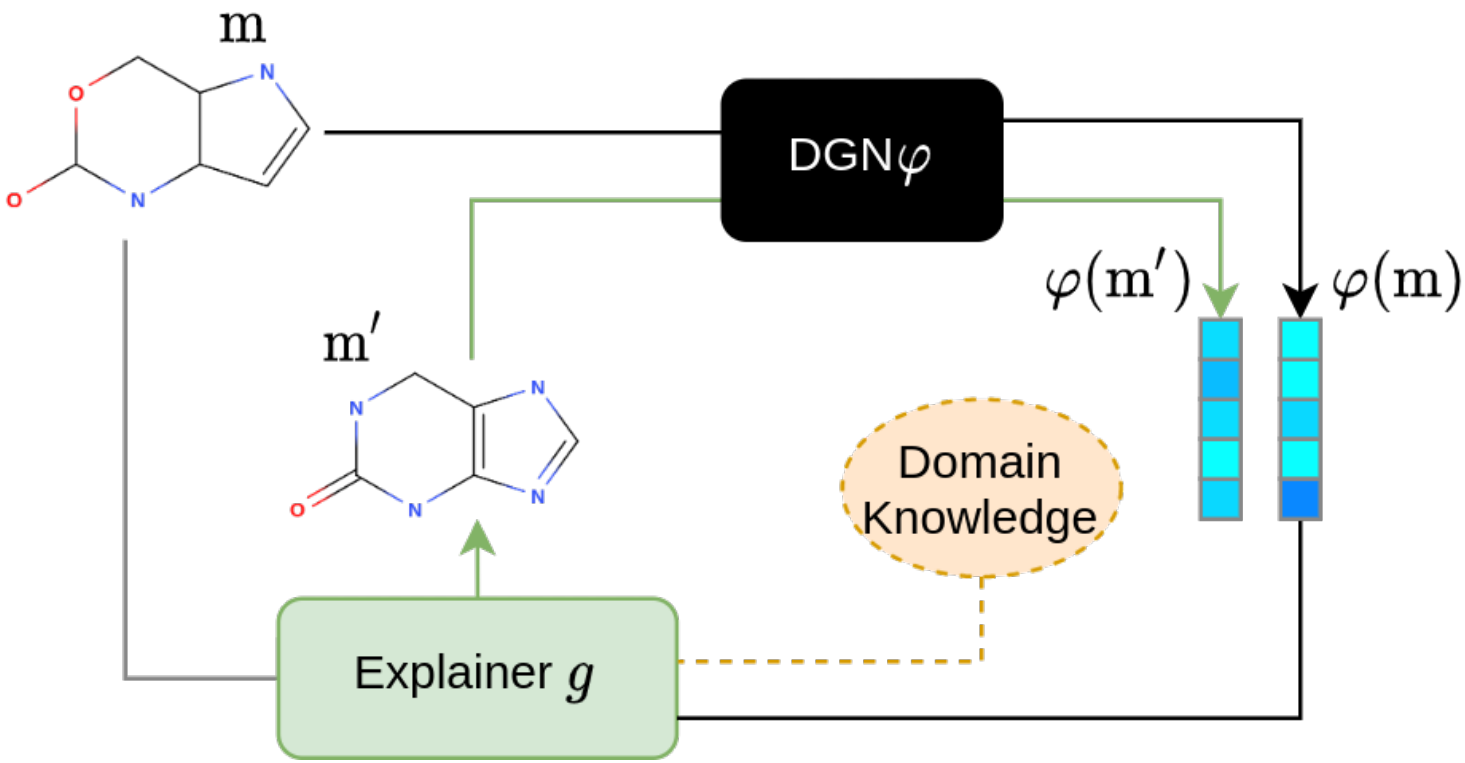
Counterfactuals make suggestions on how to increase your score.



[Grath et al., 2018]

MEG: Molecular Explanation Generator

MEG [Numeroso and Bacciu, 2020] is a model-agnostic, local explainer based on Reinforcement Learning (RL) that produces counterfactual explanations in the molecular domain.



MEG: Molecular Explanation Generator

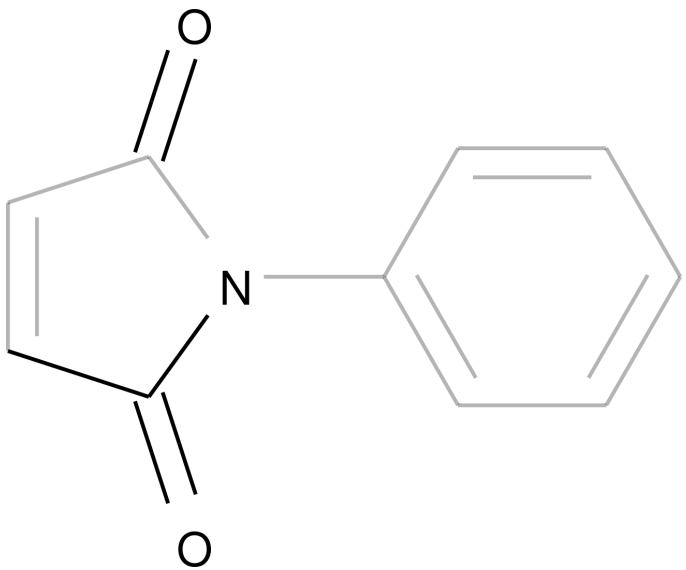
The learning problem for the RL agent takes the form of a MDP($\mathcal{S}$, $\mathcal{A}$, $\mathcal{Q}$, π, $\mathcal{R}$).

Briefly, the explainer is trained to output counterfactual explanations maximising the following reward function $\mathcal{R}$:

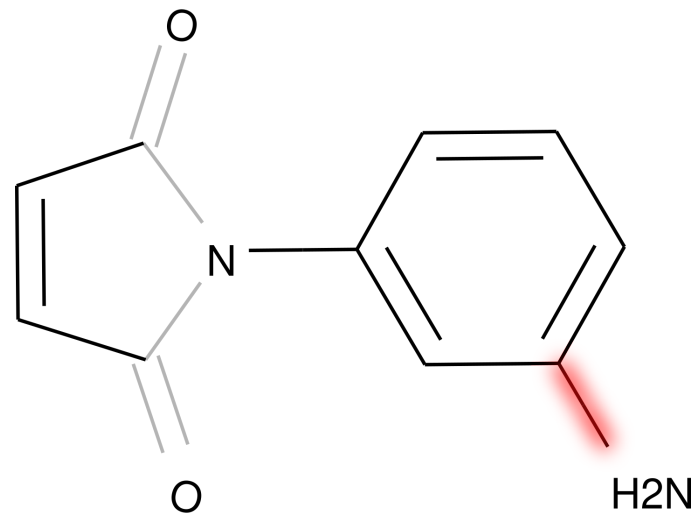
$$\text{argmax}_{m'} \mathcal{L}\big(\varphi(m), \varphi(m')\big) + \mathcal{K}\big[m, m'\big].$$

The leftmost part accounts for the difference in prediction, whereas $\mathcal{K}$ is a measure of similarity between molecules.

MEG: Example



(a) A0 → NoTox 70%



(b) A1 → Tox 90% Sim 0.76

Global explainers

Global explanations ultimate goal is to find a concise (mathematical) description of the model under study.

For this reason, generating global explanations is the most difficult task in this field.

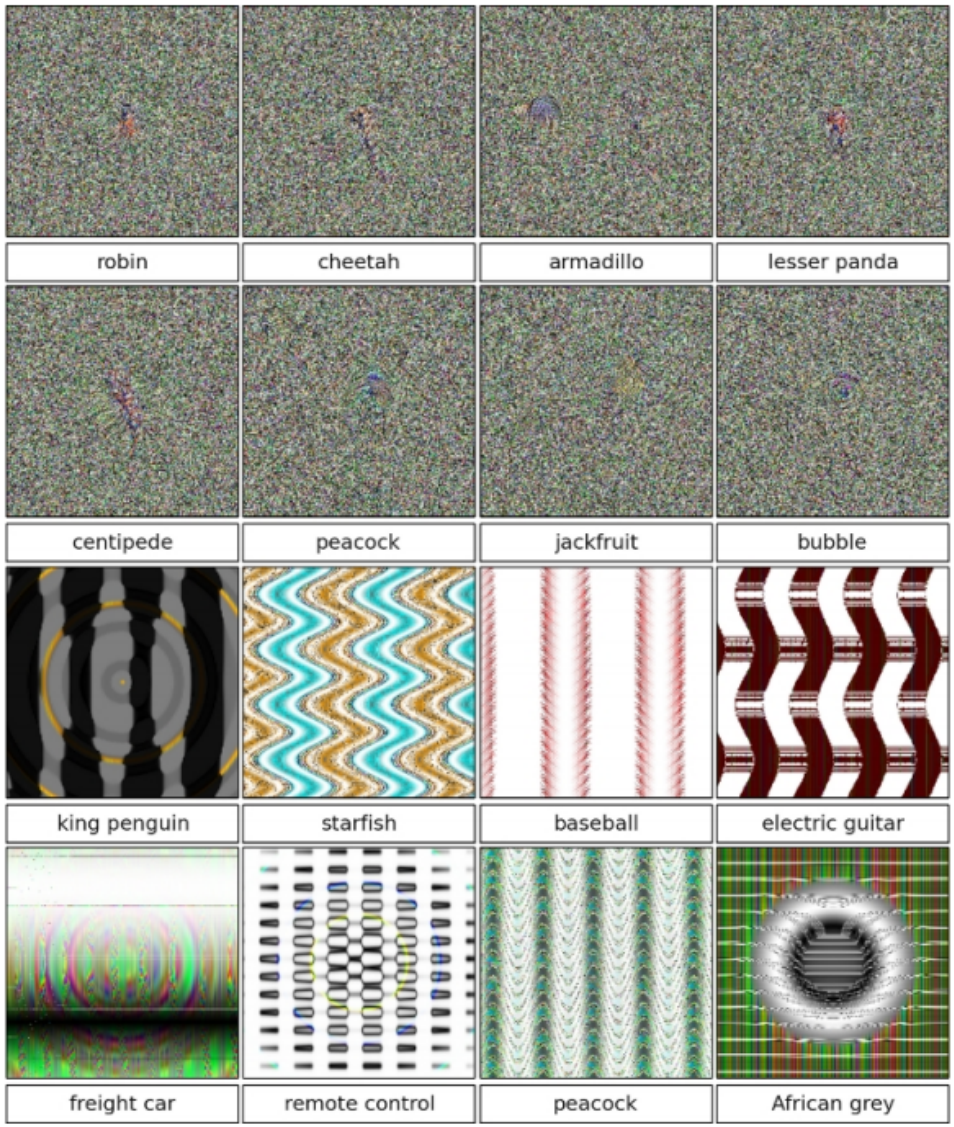
Input optimisation

Let $f : \mathcal{X} \to \mathcal{C}$ be a classifier, where $\mathcal{C}$ represent the set of targets (classes).

Input Optimisation aims to generate samples x_i that maximally activate one class of prediction $c \in \mathcal{C}$.

This way, one may investigate the patterns that the model seeks for when making predictions.

Input optimisation: Example on images



Take-home messages

- Explainability is needed, especially in safety critical contexts.
- There is no an actual formal definition of what an explanation should be.
- Therefore, validate the generated explanations is difficult. Does my explanation really reflect the model behaviour?

Thanks for your attention!

References I

- Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. (2020).
  Sanity checks for saliency maps.
- Doshi-Velez, F. and Kim, B. (2017).
  Towards a rigorous science of interpretable machine learning.
- Grath, R. M., Costabello, L., Van, C. L., Sweeney, P., Kamiab, F., Shen, Z., and Lecue, F. (2018).
  Interpretable credit application predictions with counterfactual explanations.
- Numeroso, D. and Bacciu, D. (2020).
  Explaining deep graph networks with molecular counterfactuals.
- Ribeiro, M. T., Singh, S., and Guestrin, C. (2016).
  "why should i trust you?": Explaining the predictions of any classifier.

References II


Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2016).
Grad-cam: Visual explanations from deep networks via gradient-based localization.


Simonyan, K., Vedaldi, A., and Zisserman, A. (2013).
Deep inside convolutional networks: Visualising image classification models and saliency maps.


Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2015).
Learning deep features for discriminative localization.