DATA MINING 2
Explainability

Riccardo Guidotti

a.a. 2024/2025

UNIVERSITÀ DI PISA

Definitions

- To ***interpret*** means to give or provide the meaning or to explain and present in understandable terms some concepts.

- In AI, and in data mining and machine learning, interpretability is the ***ability to explain*** or to provide the meaning ***in understandable terms to a human***.

- https://www.merriam-webster.com/
- Finale Doshi-Velez and Been Kim. 2017. ***Towards a rigorous science of interpretable machine learning***. arXiv:1702.08608v2.

What is a Black Box Model?

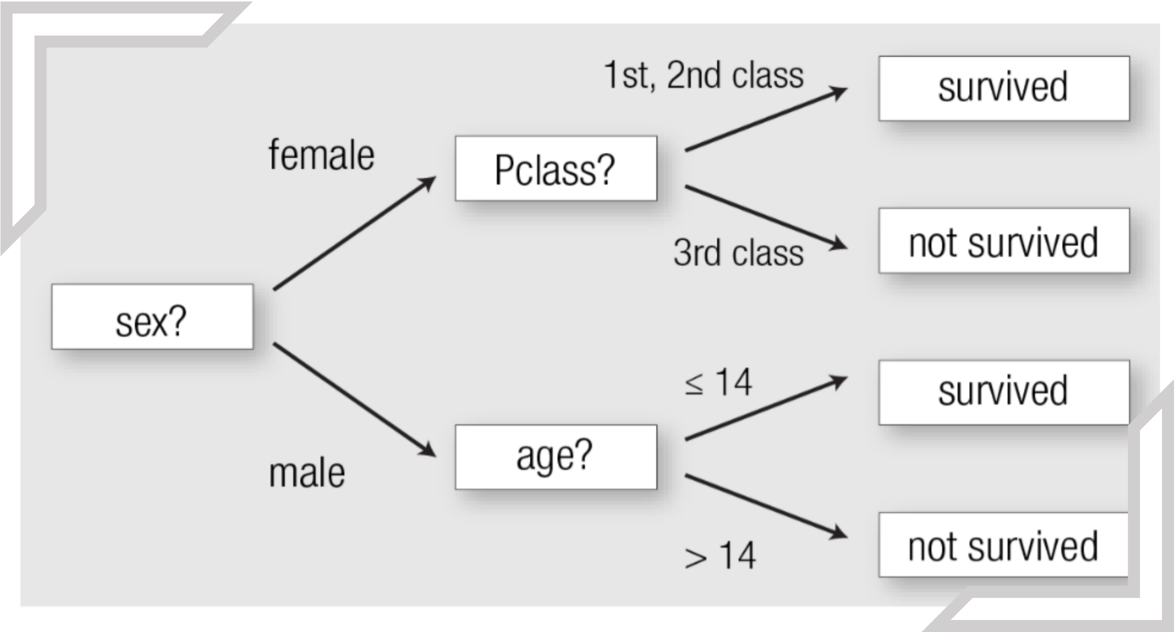
A ***black box*** is a model, whose internals are either unknown to the observer or they are known but uninterpretable by humans.

Example:

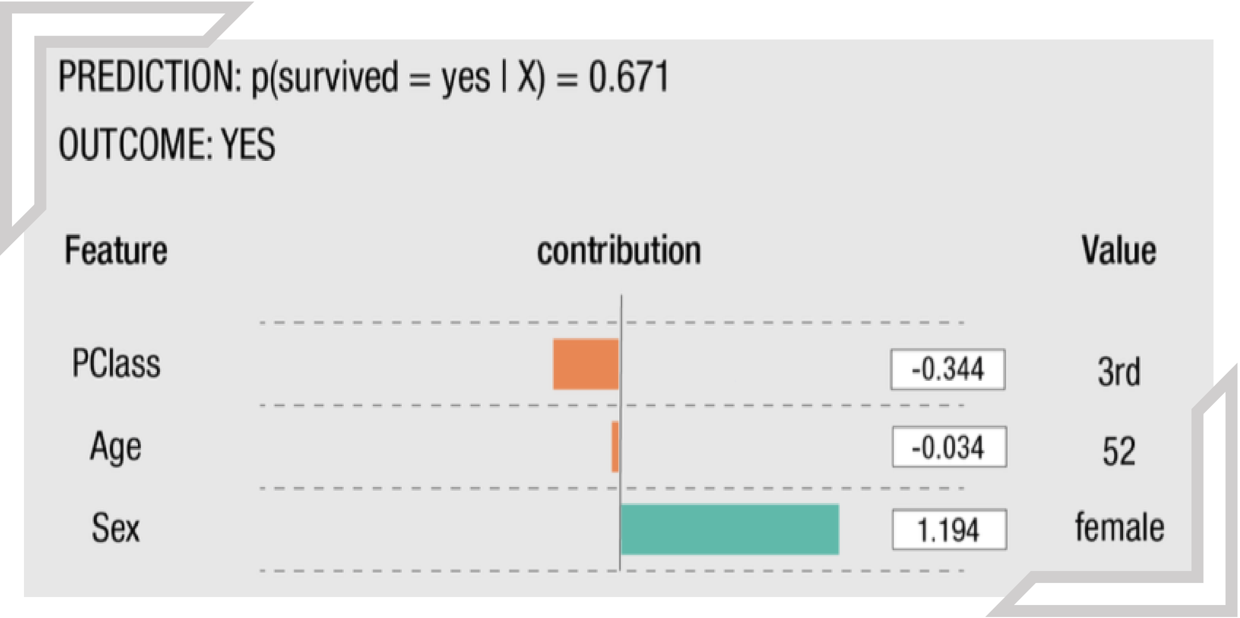
- DNN
- SVM
- Ensemble

- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). ***A survey of methods for explaining black box models***. *ACM Computing Surveys (CSUR)*, *51*(5), 93.

Interpretable Models



Decision Tree



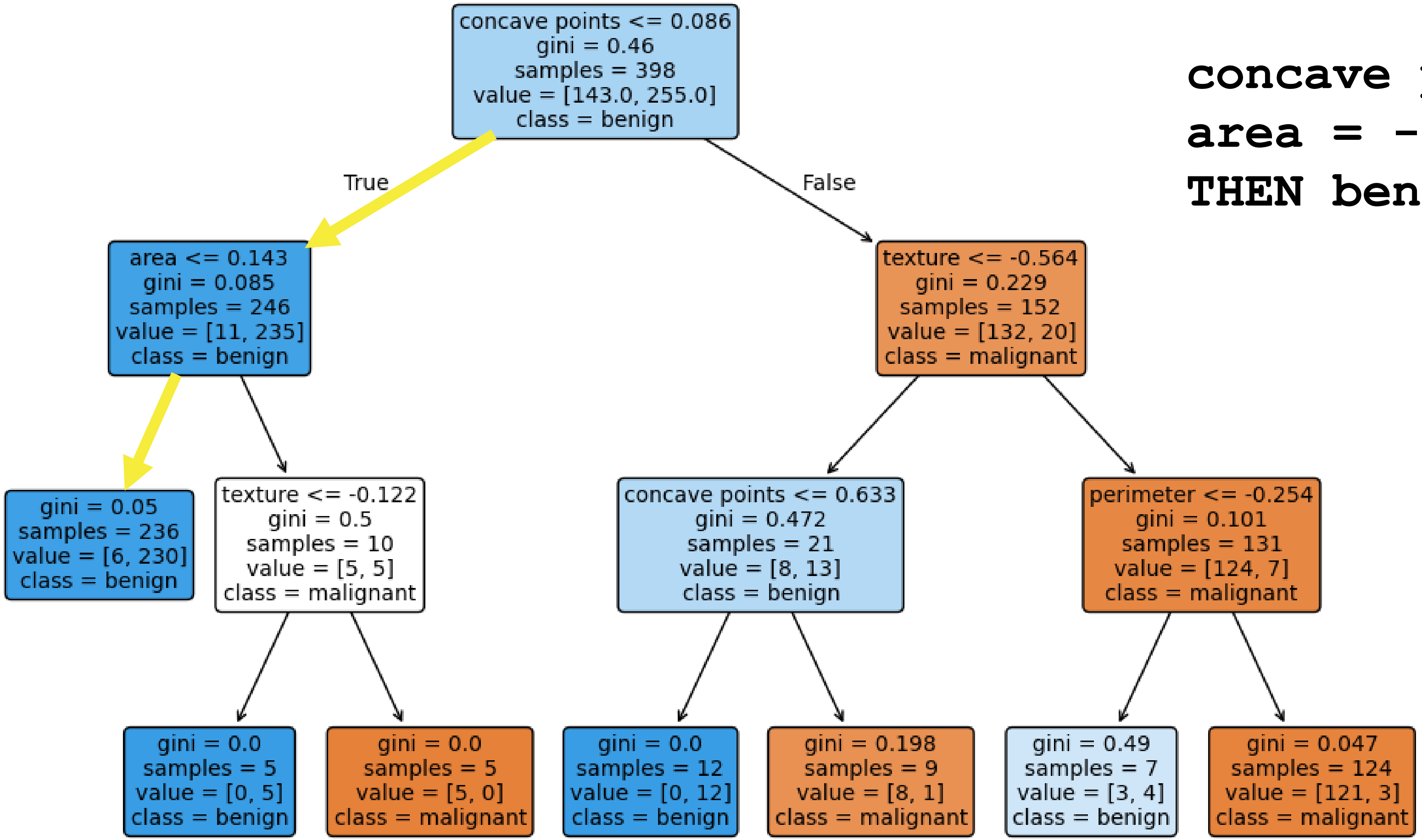
Linear Model

$$\text{if } condition_1 \wedge condition_2 \wedge condition_3 \text{ then } outcome$$

Rules

Rule-based

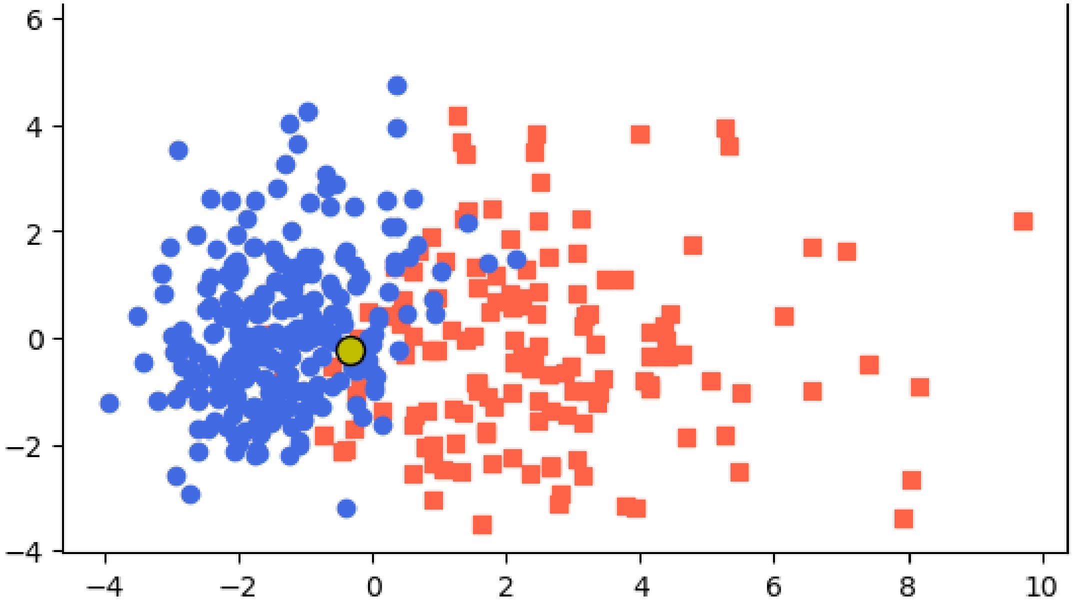
	radius	texture	perimeter	area	smoothness	compactness	concavity	concave points	symmetry	fractal dimension
x	0.09	-1.34	0.05	-0.01	0.21	-0.53	-0.38	-5.59e-02	1.04	-1.22

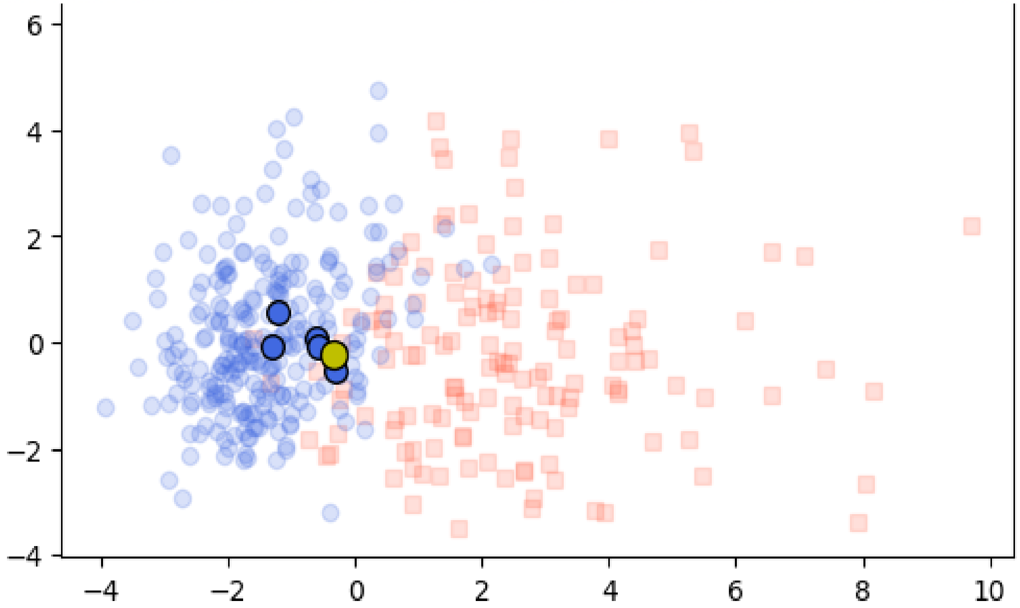


concave points = -0.056 <= 0.086 AND
area = -0.015 <= 0.143
THEN benign

Instance-based Models

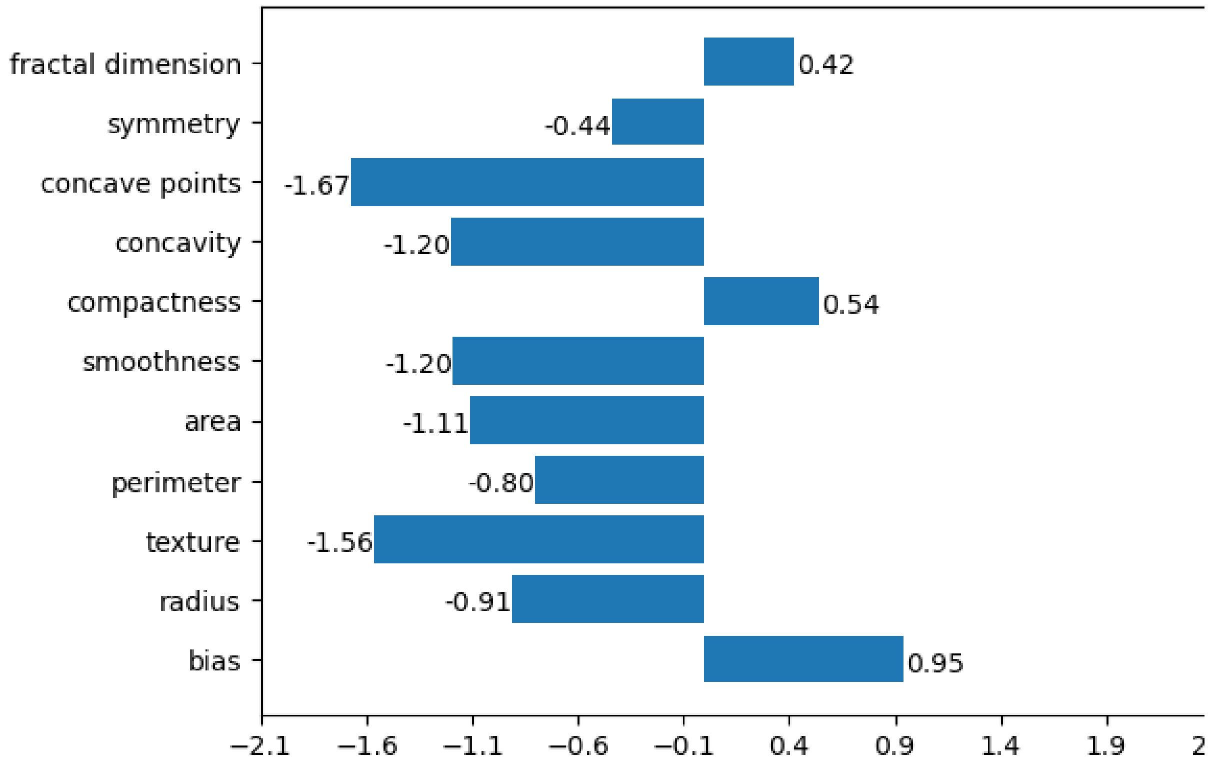
	radius	texture	perimeter	area	smoothness	compactness	concavity	concave points	symmetry	fractal dimension	target	dist
x	0.09	-1.34	0.05	-0.01	0.21	-0.53	-0.38	-5.59e-02	1.04	-1.22	benign	0.00
x1	-0.05	-0.94	-0.10	-0.17	-0.07	-0.52	-0.55	-2.89e-01	1.17	-0.62	benign	0.87
x2	-0.14	-1.13	-0.16	-0.23	0.12	-0.43	-0.26	-9.39e-03	0.32	-0.75	benign	0.98
x3	-0.49	-1.60	-0.51	-0.52	0.49	-0.66	-0.61	-2.92e-01	0.60	-0.47	benign	1.38
x4	0.15	-1.30	0.12	0.04	0.16	-0.27	-0.05	-1.22e-01	-0.24	-0.94	benign	1.38
x5	-0.33	-0.70	-0.37	-0.38	-0.13	-0.80	-0.62	-8.37e-01	0.79	-0.75	benign	1.43

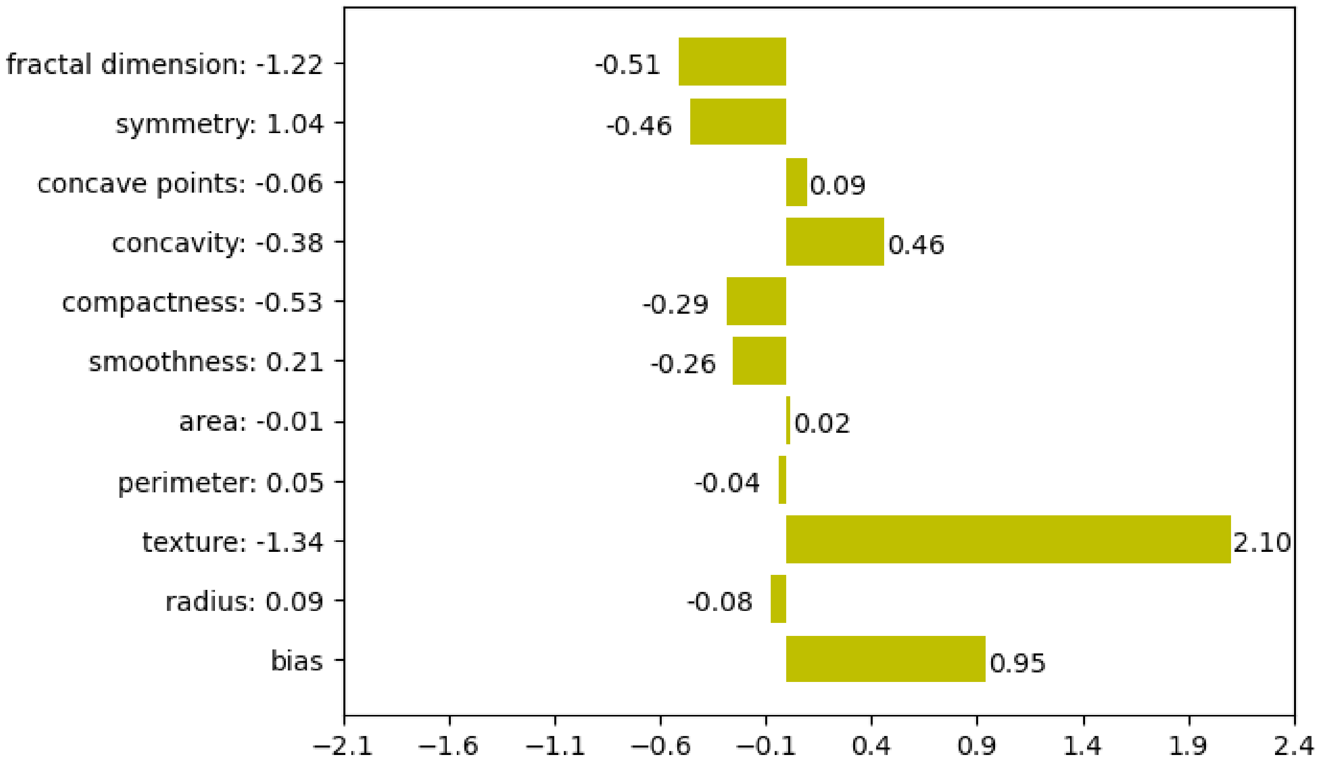




Linear Models

	radius	texture	perimeter	area	smoothness	compactness	concavity	concave points	symmetry	fractal dimension
x	0.09	-1.34	0.05	-0.01	0.21	-0.53	-0.38	-5.59e-02	1.04	-1.22



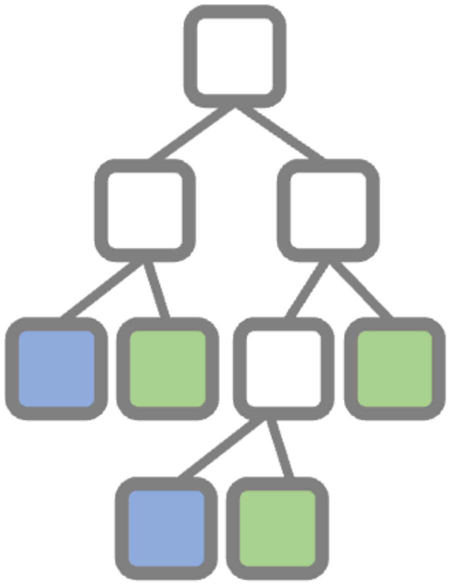


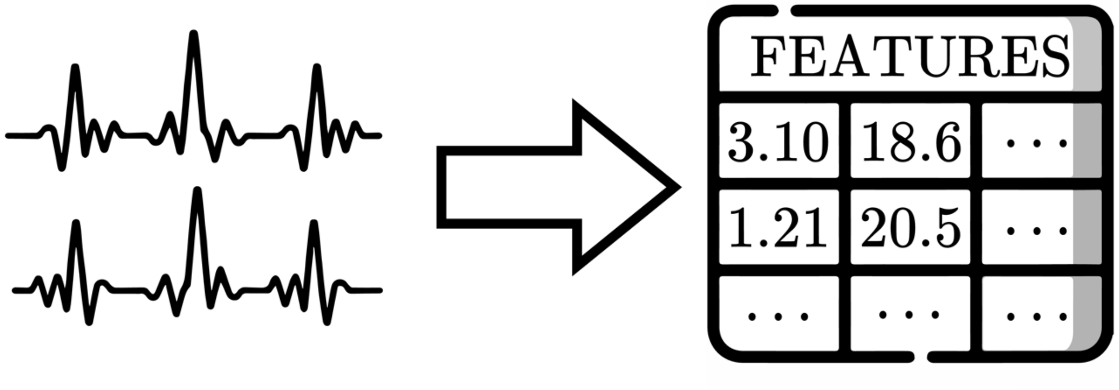
1/(1+e^-1.98) = 0.87

→ **benign**

Considerations

- The interpretability of a model depends on the domain in which decisions are made and how explanations are conveyed to the user.
- Consider data transformations that can be utilized by interpretable models without requiring external knowledge is important to foster interpretability.

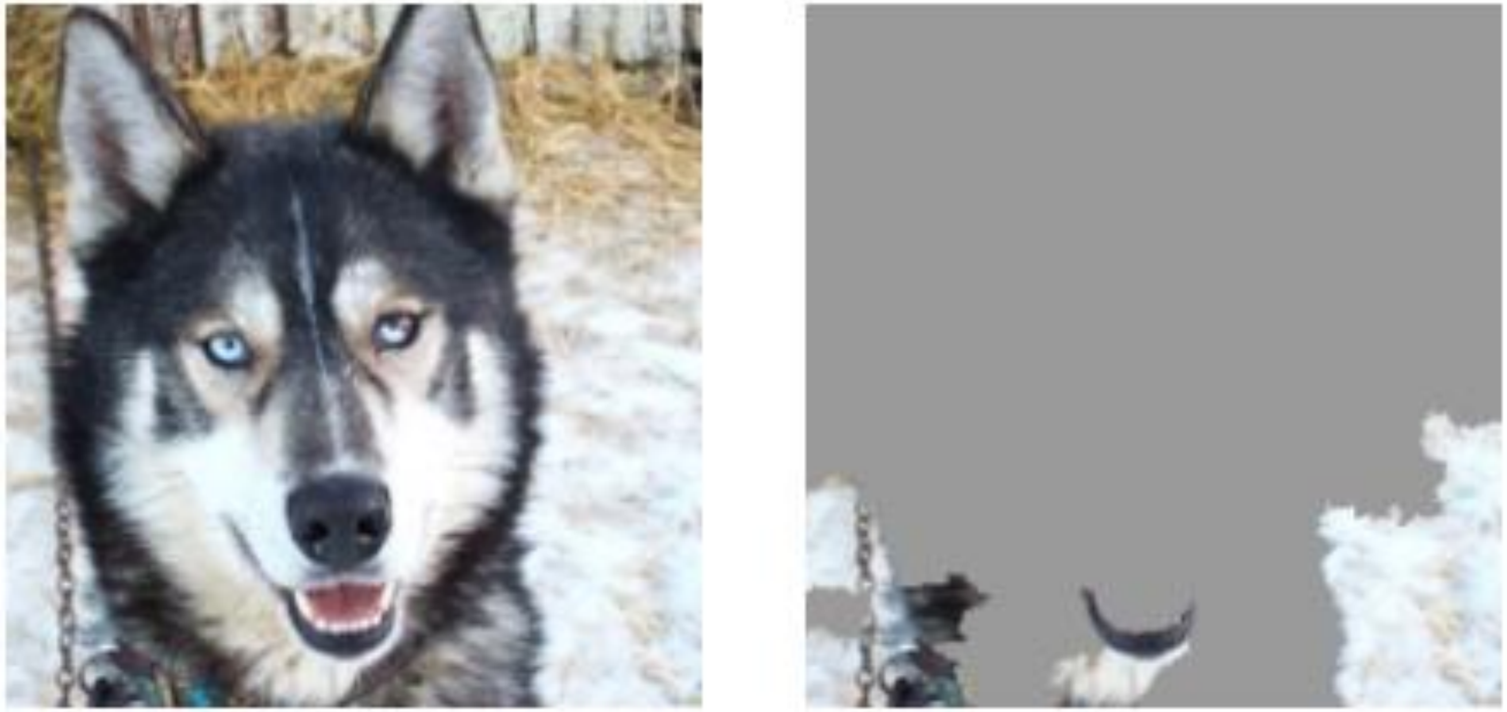




Motivations For Explanation Methods

COMPAS Recidivism

The Wolf and the Husky



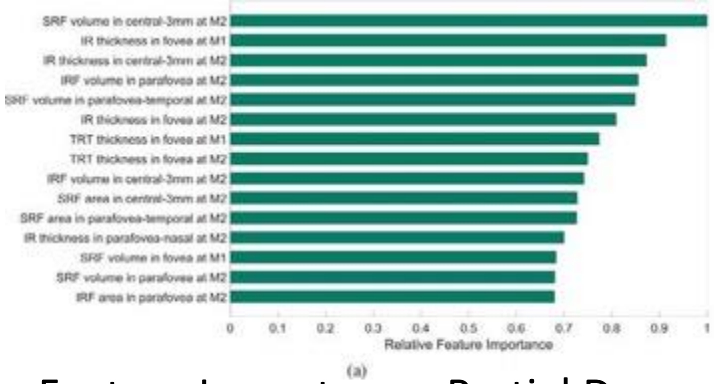
(a) Husky classified as wolf (b) Explanation

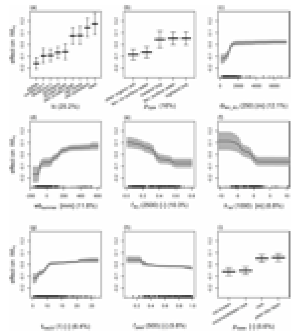
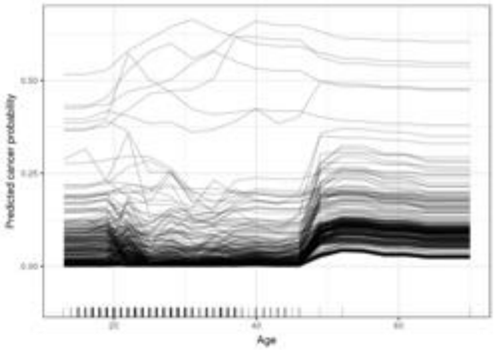
Right of Explanation

Since 25 May 2018, GDPR establishes a right for all individuals to obtain “meaningful explanations of the logic involved” when “automated (algorithmic) individual decision-making”, including profiling, takes place.

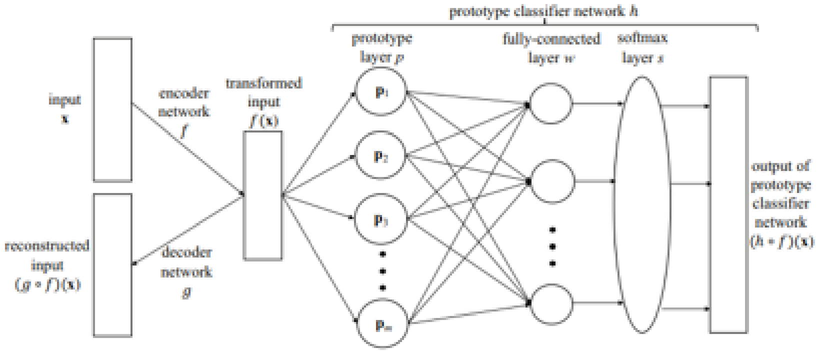
Explanation in different AI fields

- Machine Learning



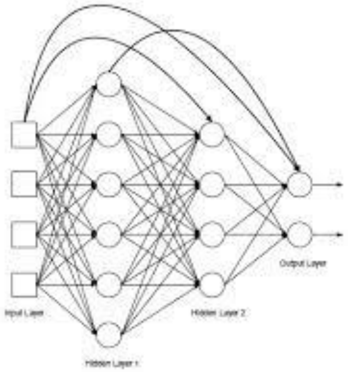
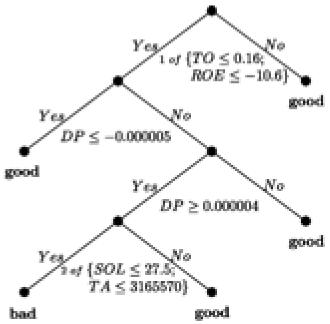



Feature Importance, Partial Dependence Plot, Individual Conditional Expectation



Auto-encoder

Oscar Li, Hao Liu, Chaofan Chen, Cynthia Rudin: Deep Learning for Case-Based Reasoning Through Prototypes: A Neural Network That Explains Its Predictions. AAAI 2018: 3530-3537

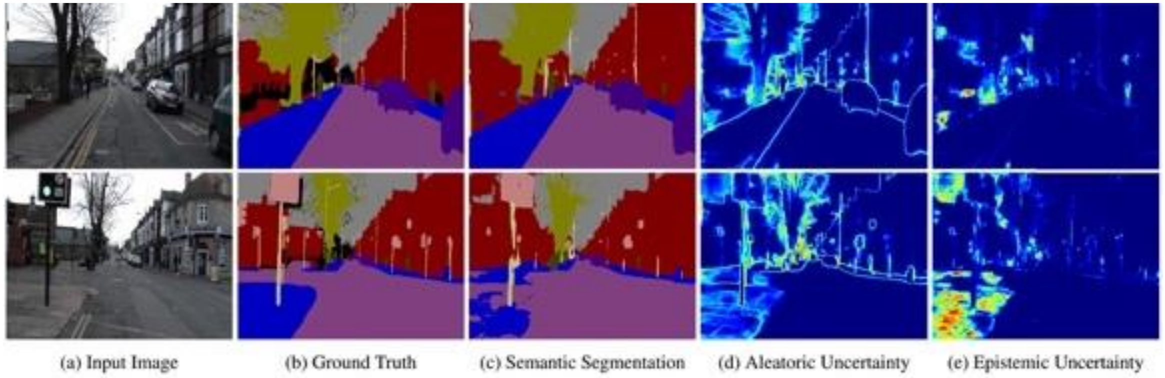



Surogate Model

Mark Craven, Jude W. Shavlik: Extracting Tree-Structured Representations of Trained Networks. NIPS 1995: 24-30

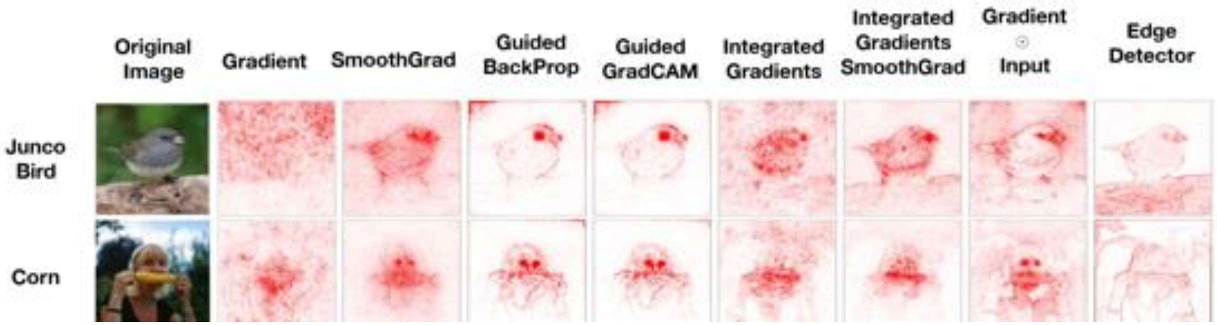
Explanation in different AI fields

- Machine Learning
- Computer Vision



Uncertainty Map

Alex Kendall, Yarin Gal: What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? NIPS 2017: 5580-5590

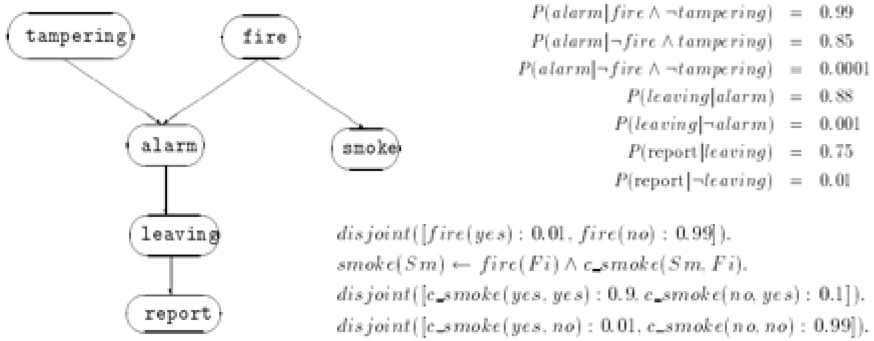


Saliency Map

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, Been Kim: Sanity Checks for Saliency Maps. NeurIPS 2018: 9525-9536

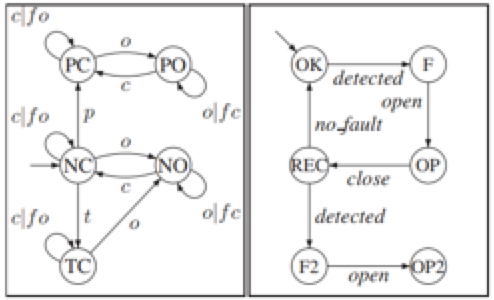
Explanation in different AI fields

- Machine Learning
- Computer Vision
- Knowledge Representation and Reasoning



Abduction Reasoning (in Bayesian Network)

David Poole: Probabilistic Horn Abduction and Bayesian Networks. Artif. Intell. 64(1): 81-129 (1993)

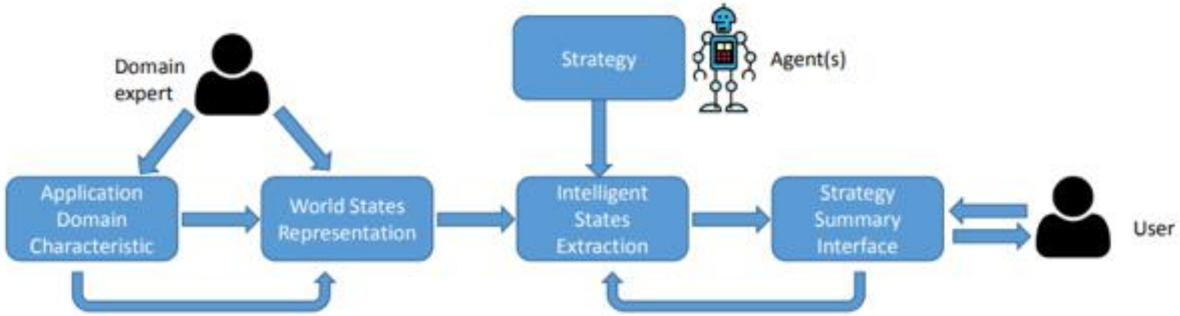


Diagnosis Inference

Alban Grastien, Patrik Haslum, Sylvie Thiébaux: Conflict-Based Diagnosis of Discrete Event Systems: Theory and Practice. KR 2012

Explanation in different AI fields

- Machine Learning
- Computer Vision
- Knowledge Representation and Reasoning
- Multi-agent Systems



Agent Strategy Summarization

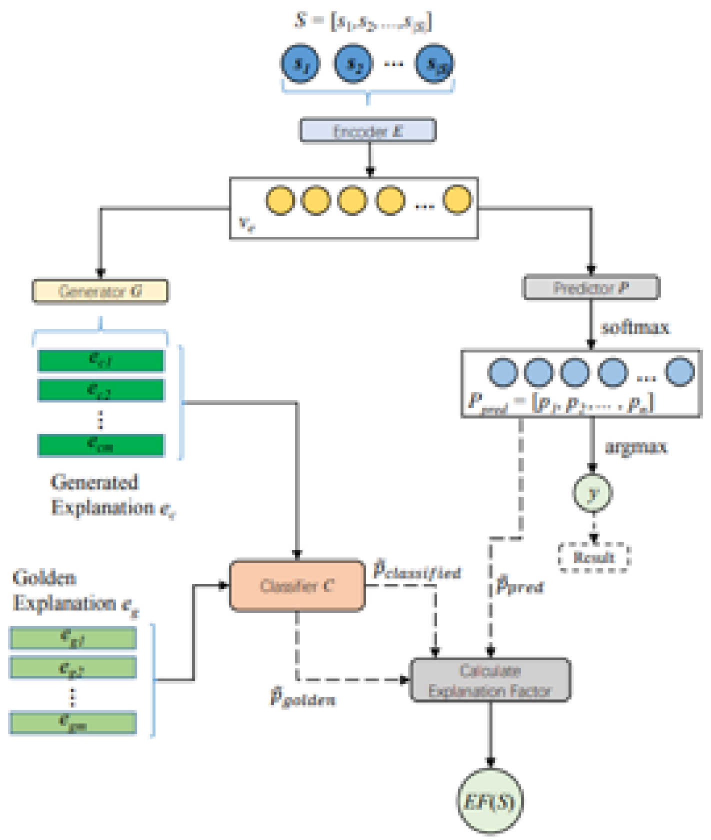
Ofra Amir, Finale Doshi-Velez, David Sarne: Agent Strategy Summarization. AAMAS 2018: 1203-1207

Explainable Agents

Joost Broekens, Maaike Harbers, Koen V. Hindriks, Karel van den Bosch, Catholijn M. Jonker, John-Jules Ch. Meyer: Do You Get It? User-Evaluated Explainable BDI Agents. MATES 2010: 28-39

Explanation in different AI fields

- Machine Learning
- Computer Vision
- Knowledge Representation and Reasoning
- Multi-agent Systems
- NLP

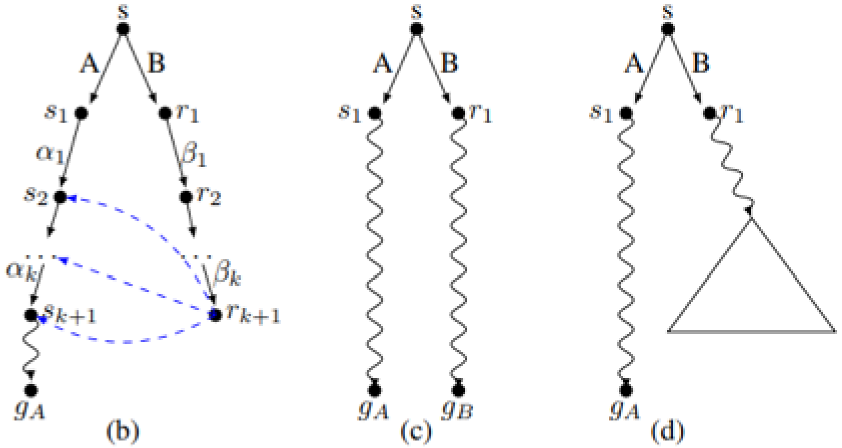


Explainable NLP

Hui Liu, Qingyu Yin, William Yang Wang: Towards Explainable NLP: A Generative Explanation Framework for Text Classification. CoRR abs/1811.00196 (2018)

Explanation in different AI fields

- Machine Learning
- Computer Vision
- Knowledge Representation and Reasoning
- Multi-agent Systems
- NLP
- Planning and Scheduling



Human-in-the-loop Planning

Maria Fox, Derek Long, Daniele Magazzeni: Explainable Planning. CoRR abs/1709.10256 (2017)

Explanation in different AI fields

- Machine Learning
- Computer Vision
- Knowledge Representation and Reasoning
- Multi-agent Systems
- NLP
- Planning and Scheduling
- Robotics

Robot: I have decided to turn left.

Human: Why did you do that?

Robot: I believe that the correct action is to turn left
BECAUSE:
I'm being asked to go forward
AND This area in front of me was 20 cm higher than me *highlights area*
AND the area to the left has maximum protrusions of less than 5 cm *highlights area*
AND I'm tilted to the right by more than 5 degrees.
Here is a display of the path through the tree that lead to this decision. *displays tree*

Human: How confident are you in this decision?

Robot: The distribution of actions that reached this leaf node is shown in this histogram. *displays histogram* This action is predicted to be correct 67% of the time.

Human: Where did the threshold for the area in front come from?

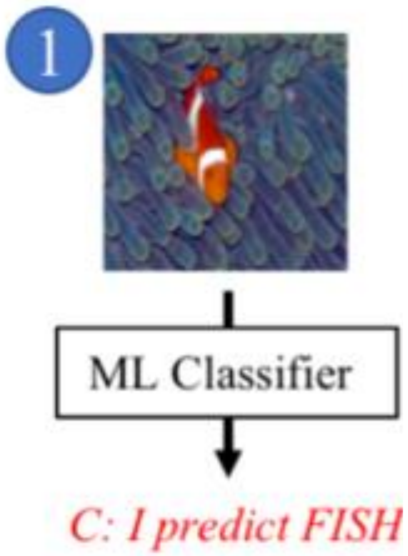
Robot: Here is the histogram of all training examples that reached this leaf. 80% of examples where this area was above 20 cm predicted the appropriate action to be "drive forward".

From Decision Tree to human-friendly information

Raymond Ka-Man Sheh: "Why Did You Do That?" Explainable Intelligent Robots. AAAI Workshops 2017

Explanation as *Machine-Human Conversation*

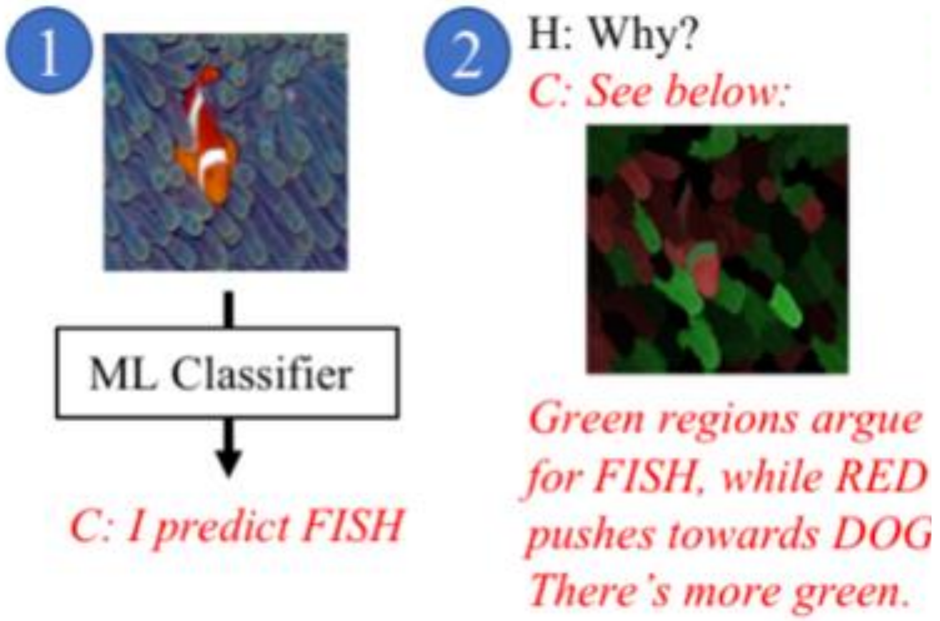
[Weld and Bansal 2018]



- Humans may have follow-up questions
- Explanations cannot answer all users' concerns

Explanation as *Machine-Human Conversation*

[Weld and Bansal 2018]



- Humans may have follow-up questions
- Explanations cannot answer all users' concerns

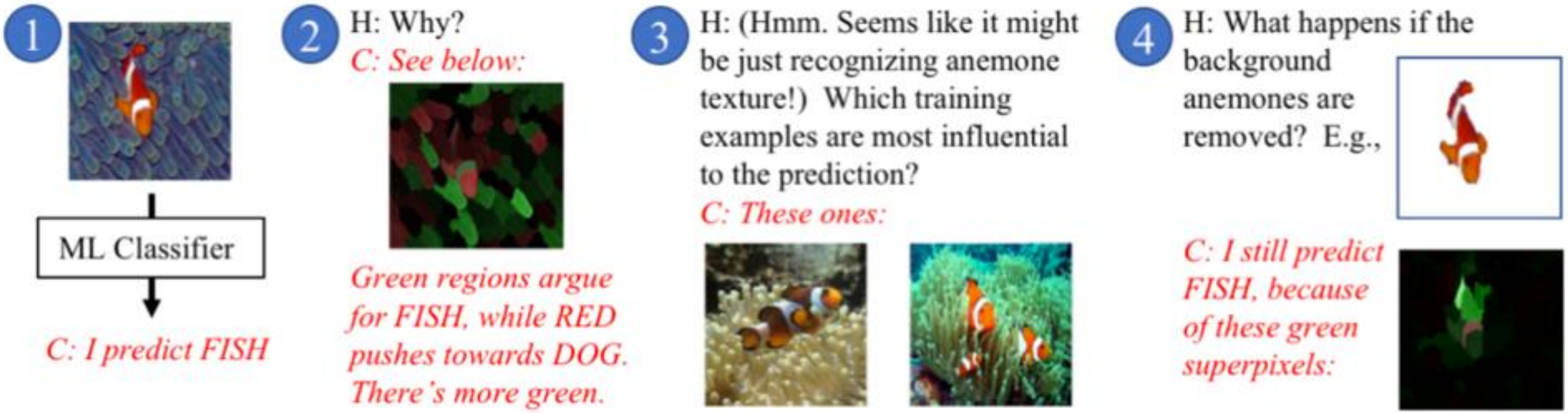
Explanation as *Machine-Human Conversation*

[Weld and Bansal 2018]

- Humans may have follow-up questions
- Explanations cannot answer all users' concerns

Explanation as *Machine-Human Conversation*

[Weld and Bansal 2018]



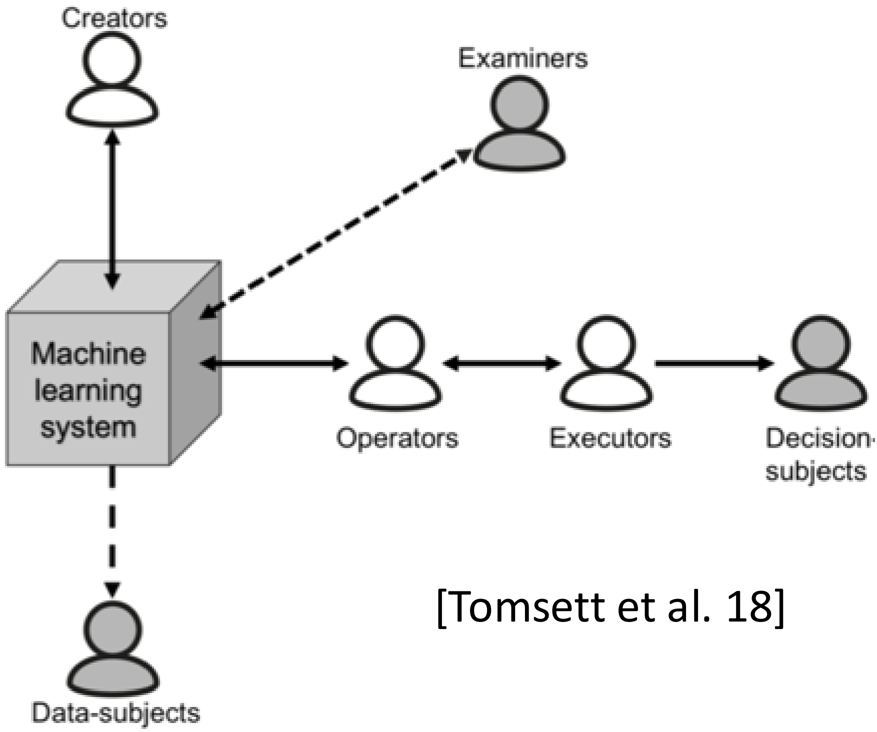
- Humans may have follow-up questions
- Explanations cannot answer all users' concerns

Role-based Interpretability

"~~Is the explanation interpretable~~?" → "*To whom* is the explanation interpretable?"

No Universally Interpretable Explanations!

- **End users** "Am I being treated fairly?"
  "Can I contest the decision?"
  "What could I do differently to get a positive outcome?"
- **Engineers, data scientists**: "Is my system working as designed?"
- **Regulators** " Is it compliant?"



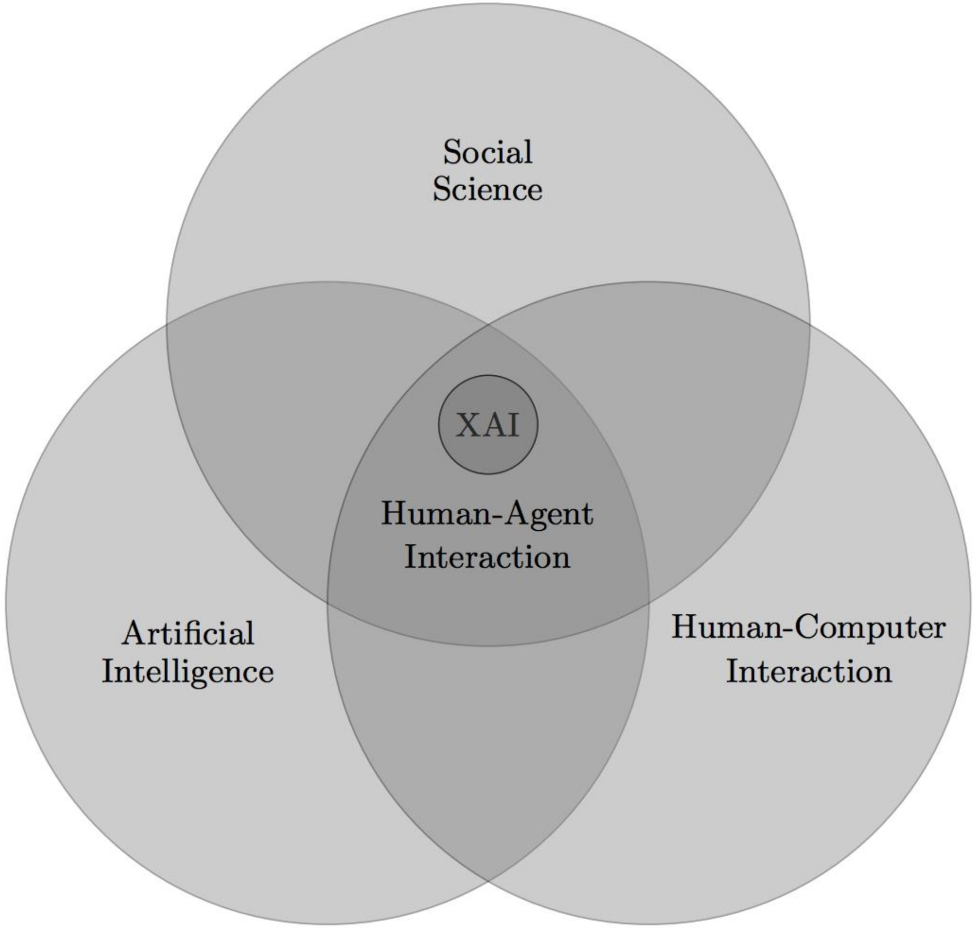
[Tomsett et al. 18]

An ideal explainer should model the *user background.*

[Tomsett et al. 2018, Weld and Bansal 2018, Poursabzi-Sangdeh 2018, Mittelstadt et al. 2019]

XAI is Interdisciplinary

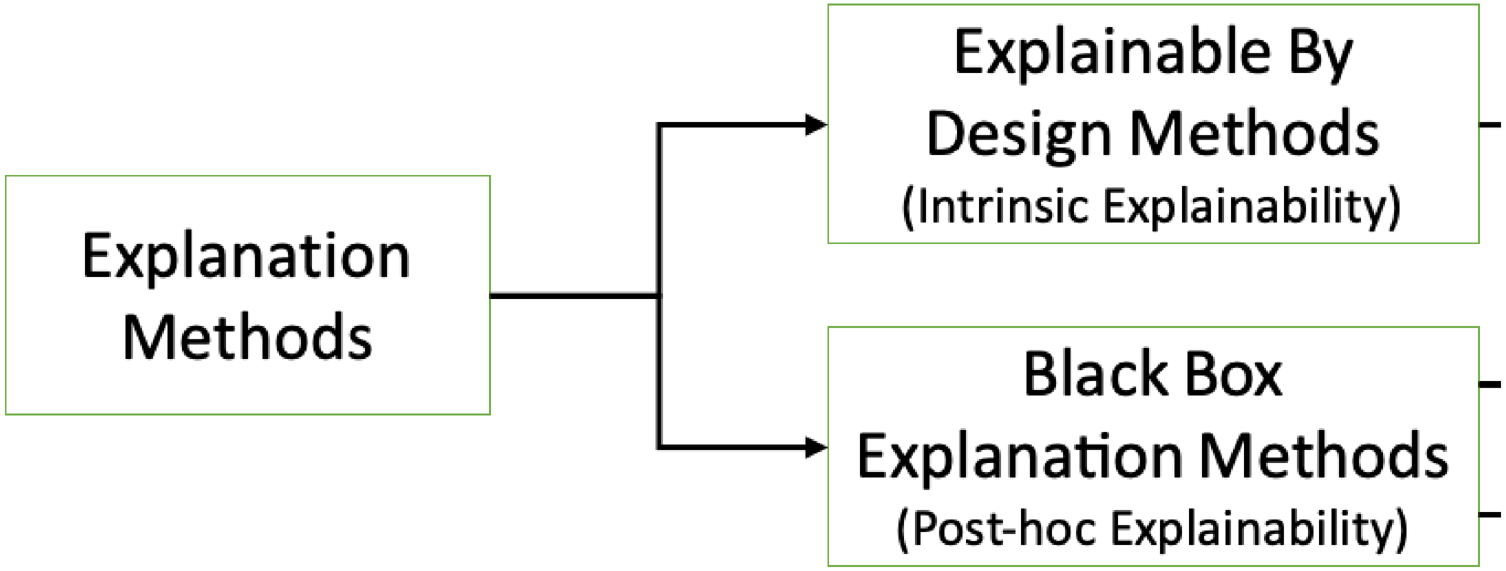
- For millennia, philosophers have asked the questions about what constitutes an explanation, what is the function of explanations, and what are their structure
- **[Tim Miller 2018]**



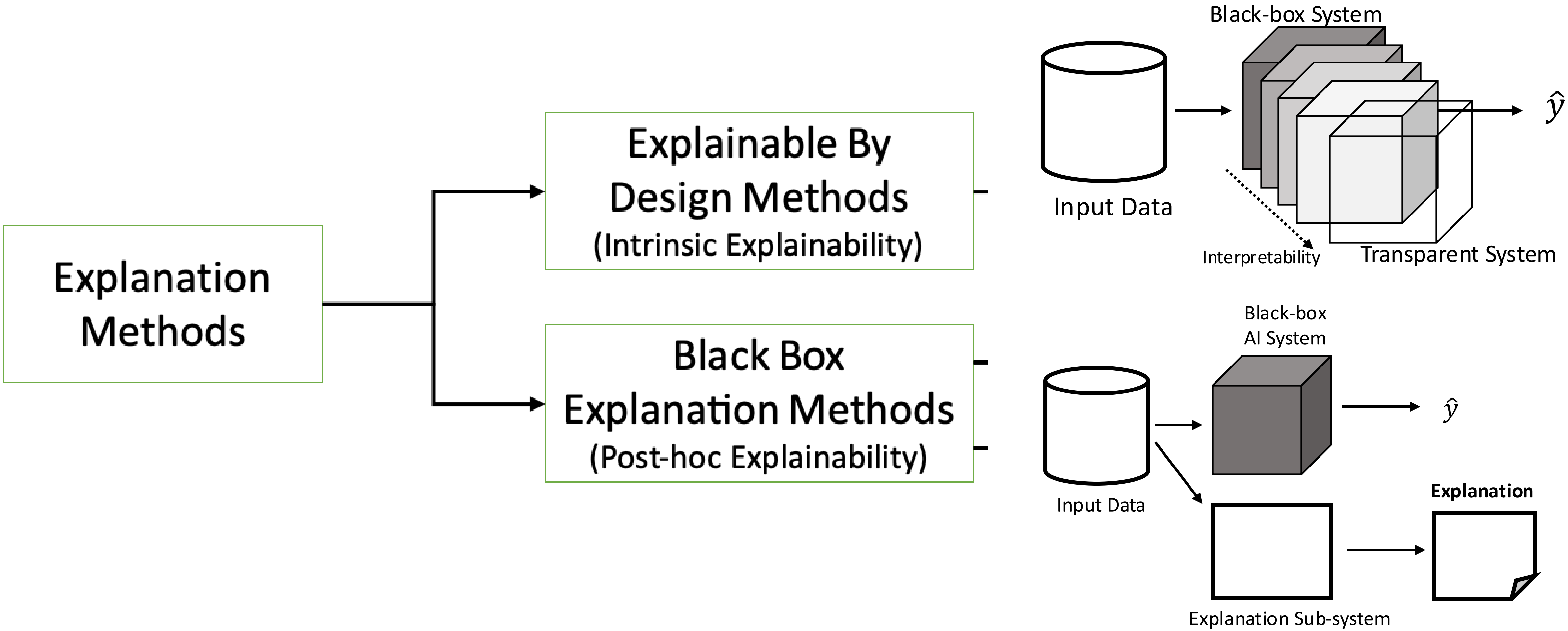

0
90
80
70
60
How to Open the Black Box

XAI Taxonomy of Explanation Methods

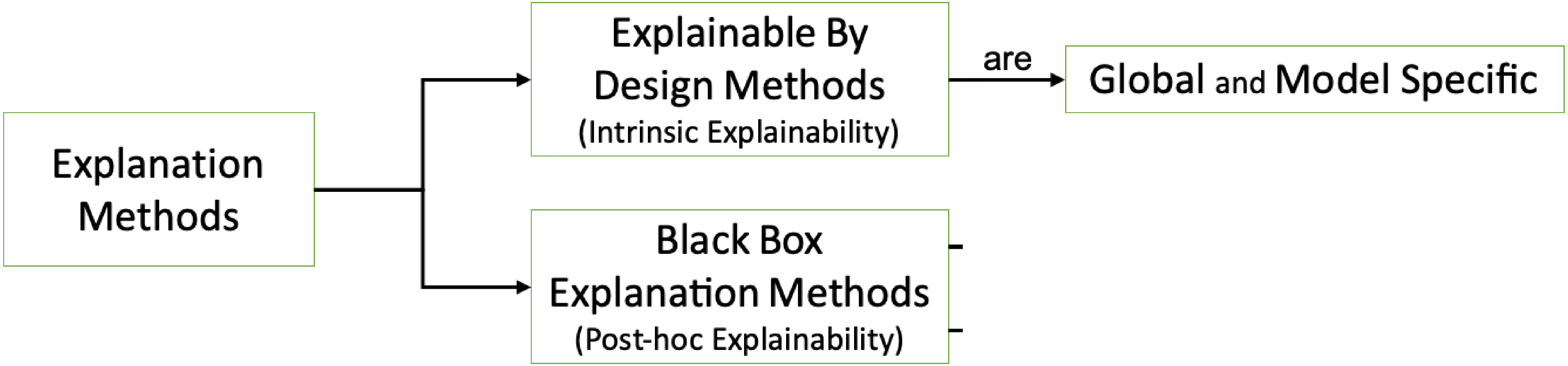
XAI Taxonomy of Explanation Methods



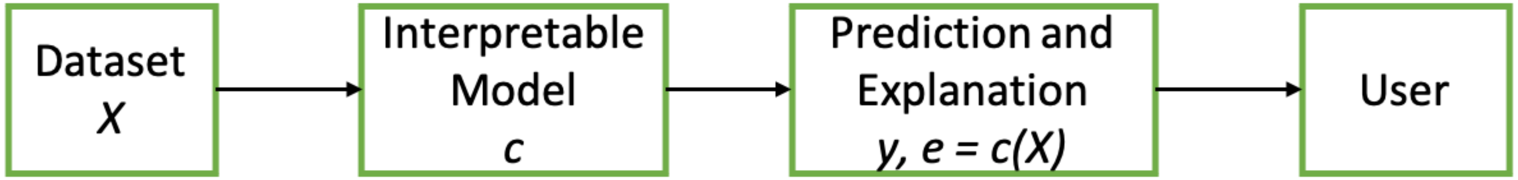
XAI Taxonomy of Explanation Methods

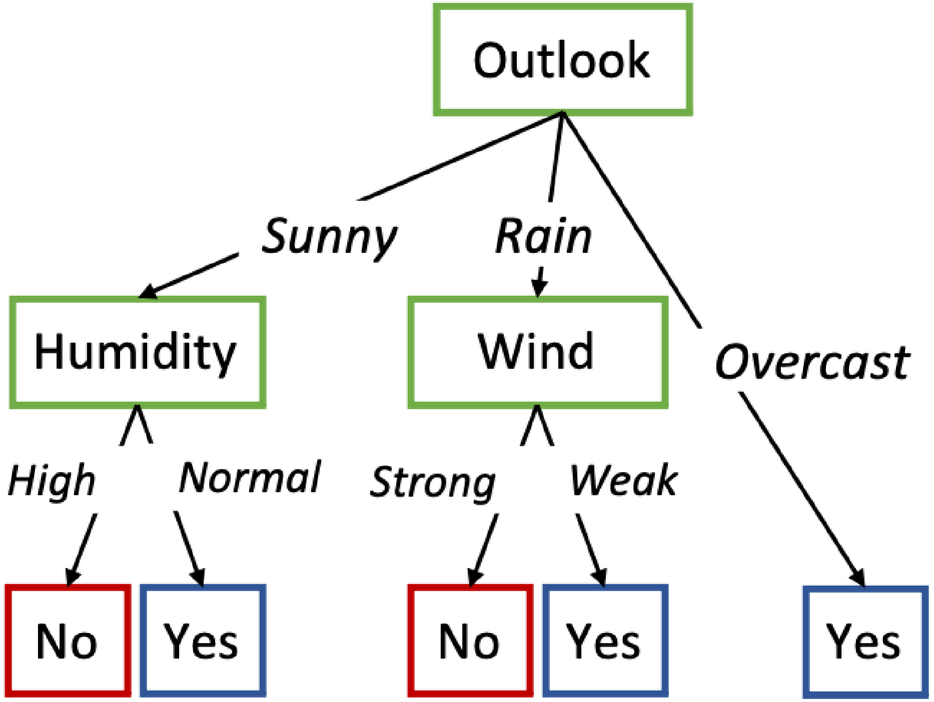


XAI Taxonomy of Explanation Methods

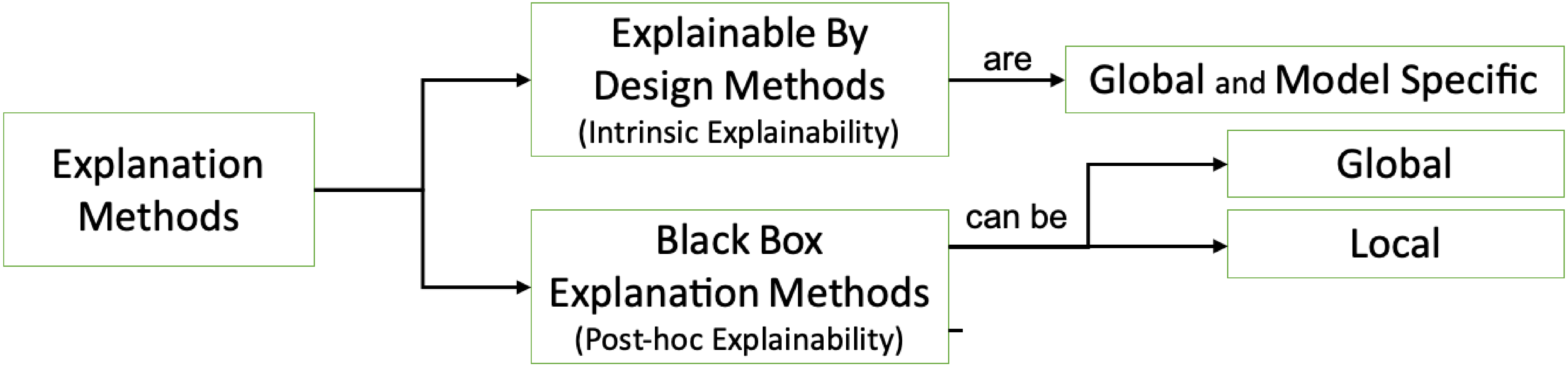


Explainable by Design Method

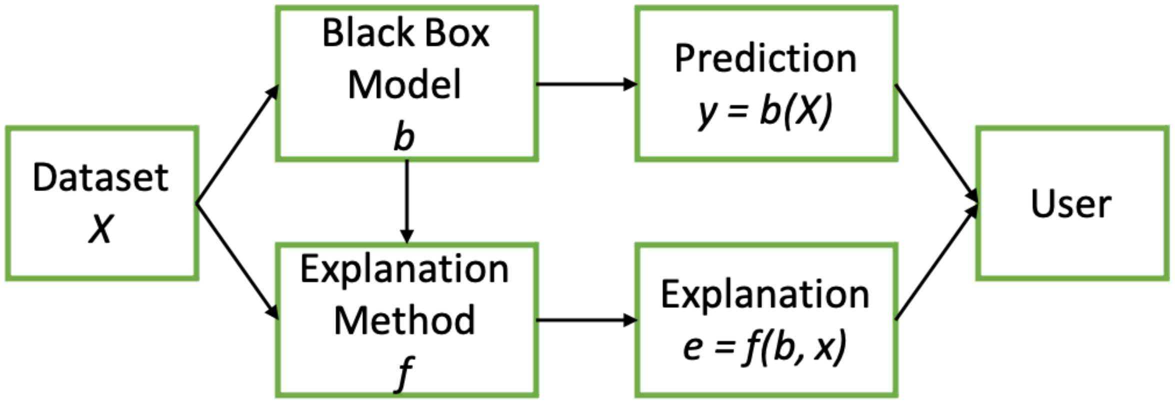


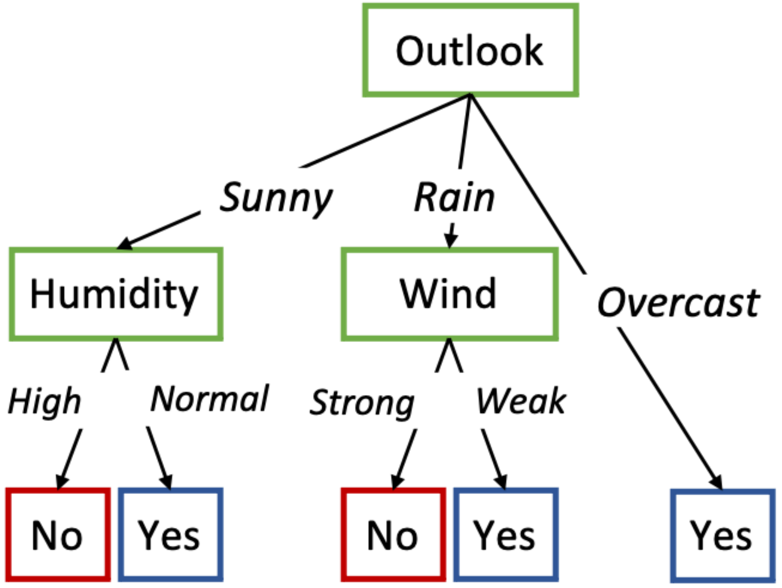


XAI Taxonomy of Explanation Methods



Black Box Explanations: Global vs Local





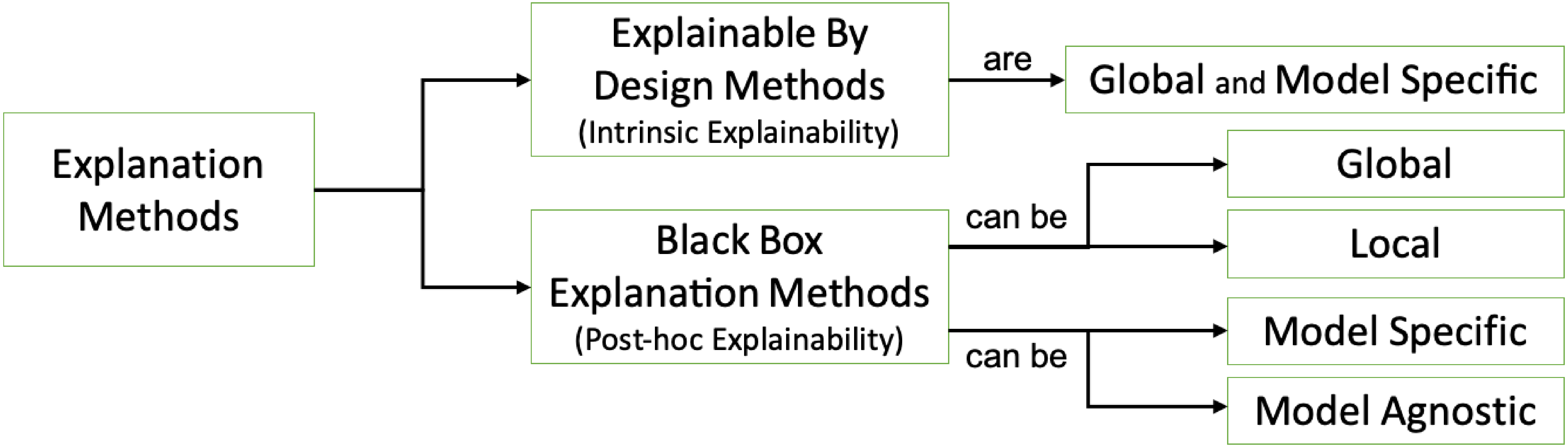
Global Explanation

If Outlook = *Sunny* and Humidity = *Normal* **then** Play Tennis = ***Yes***

- Outlook: *0.7*
- Humidity: *-0.4*
- Wind: *0.0*

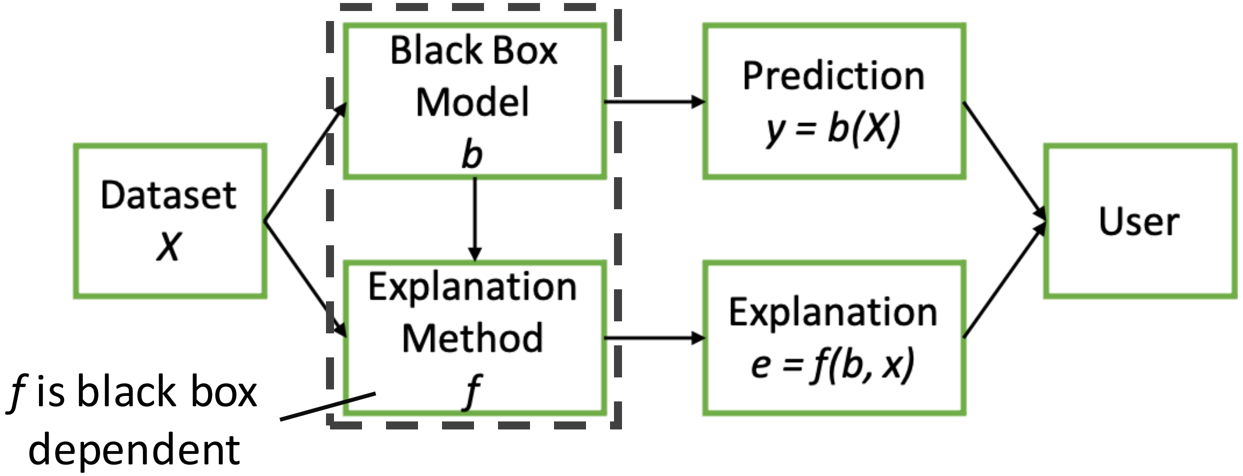
Local Explanations

XAI Taxonomy of Explanation Methods

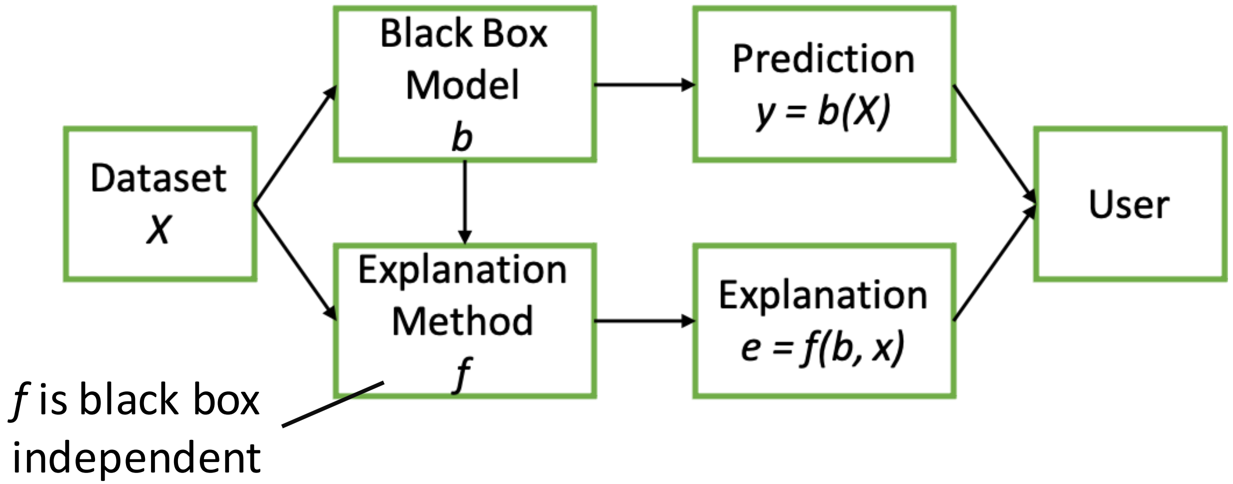


Black Box Explanations: Specific vs Agnostic

Model Specific



Model Agnostic



Types of Data

Table of baby-name data
(baby-2010.csv)

name	rank	gender	year
Jacob	1	boy	2010
Isabella	1	girl	2010
Ethan	2	boy	2010
Sophia	2	girl	2010
Michael	3	boy	2010

Field names

One row (4 fields)

2000 rows all told

Tabular

(TAB)

Images

(IMG)

Text

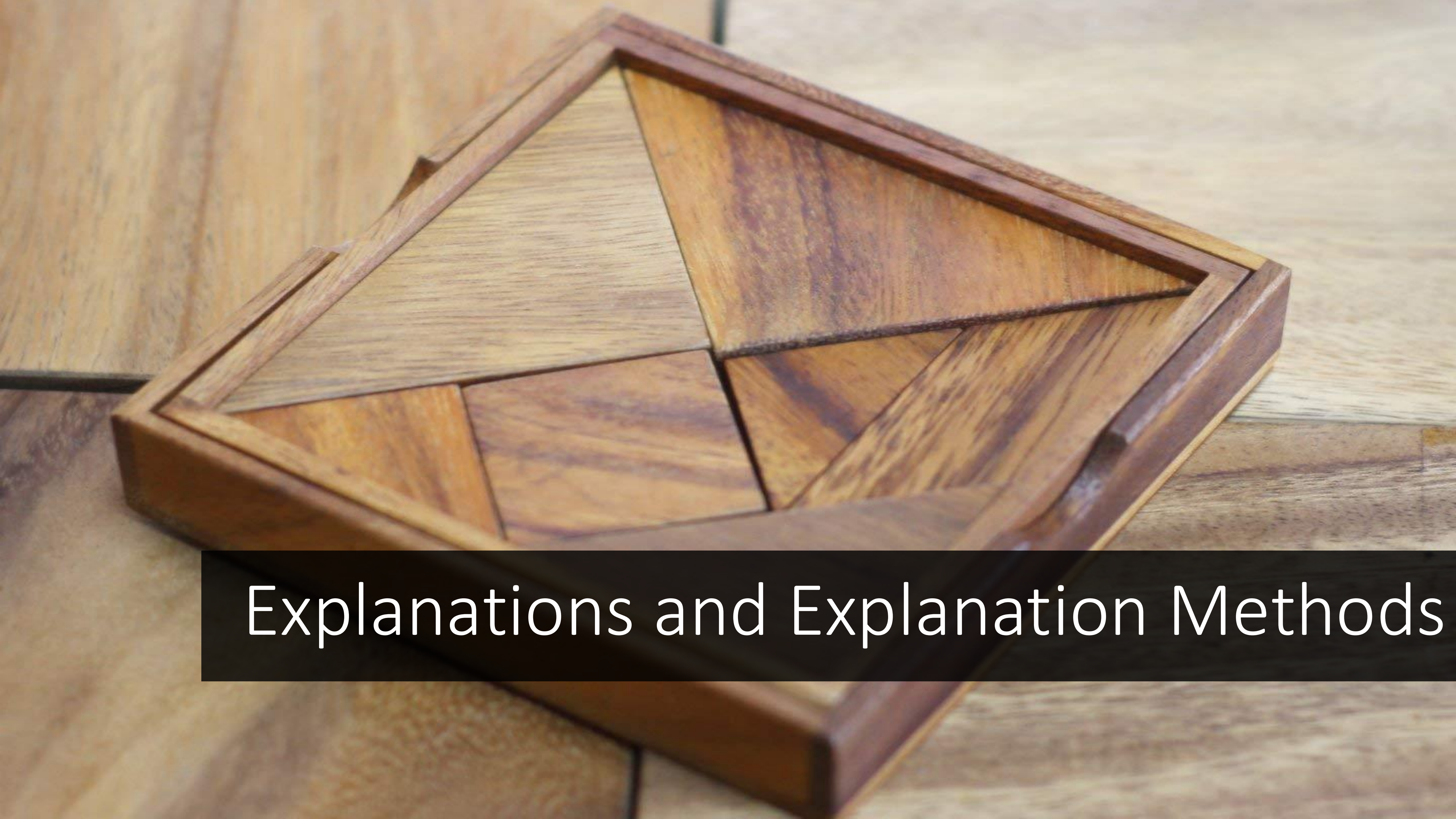
(TXT)

Types of Explanations

- Tabular Data
  - Rule-based
  - Decision Tree
  - Features Importance
  - Prototypes
  - Counter-exemplars
- Images
  - Saliency Maps
  - Concept Attributions
  - Prototypes
  - Counter-exemplars
- Text
  - Sentence Highlighting
  - Attention-based
  - Prototypes
  - Counter-exemplars

If Outlook = *Sunny* and Humidity = *Normal*
then Play Tennis = ***Yes***

- Outlook: *0.7*
- Humidity: *-0.4*
- Wind: *0.0*

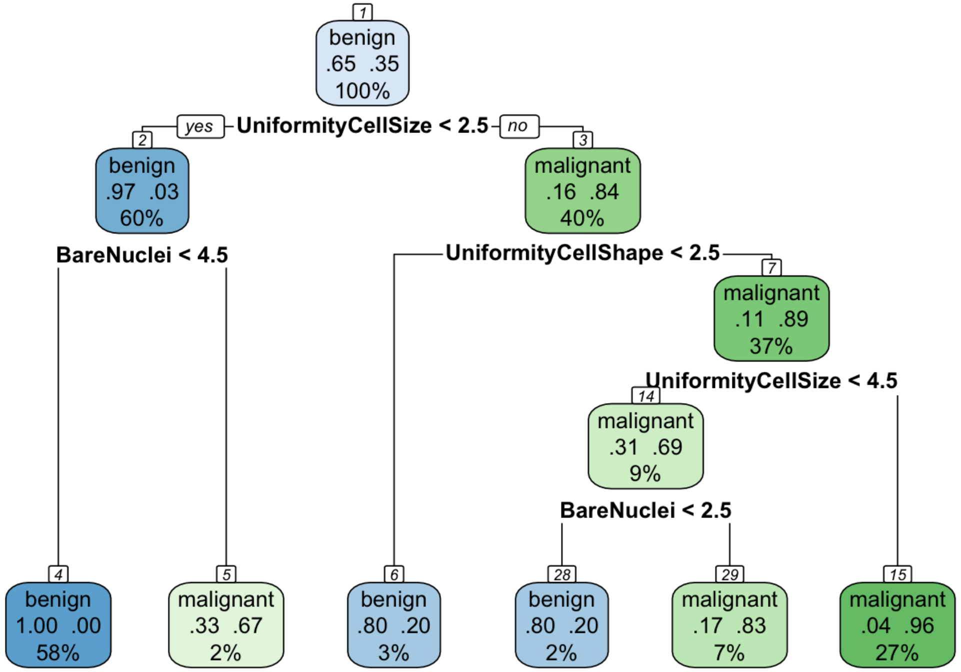


Explanations and Explanation Methods

TREPAN

Trepan

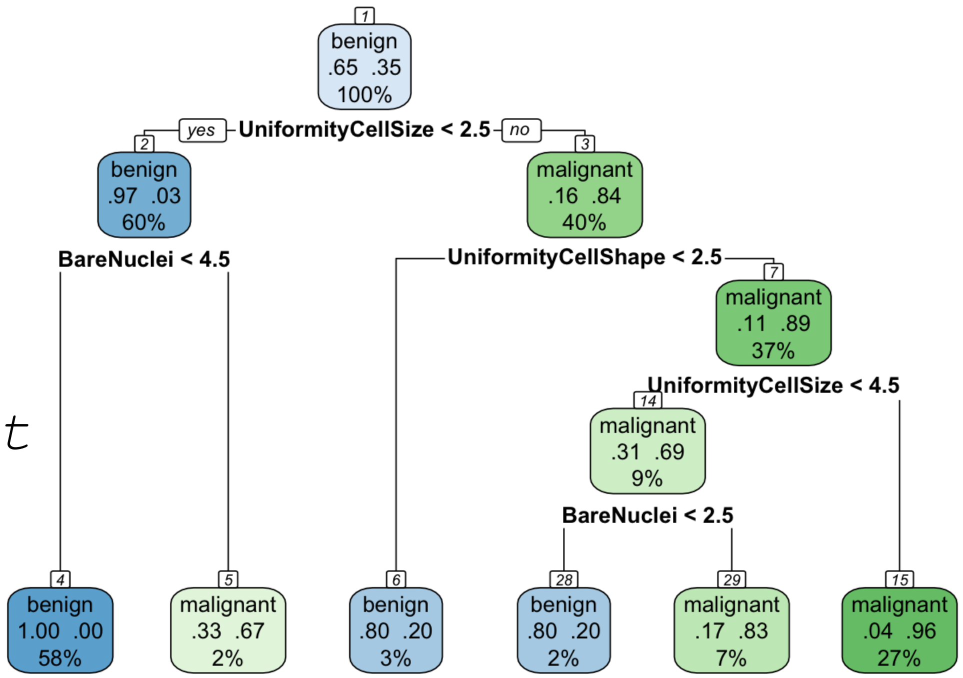
- Global explainer designed to explain NN but usable for any type of black box.
- It aims at approximating a NN with a DT classifier using best-m-of-n rules.
- At each node split the feature to split is selected on the original data extended with random samples respecting the current path.
- It learns to predict the label returned by the black box, not the original one.



Trepan

```
T = root_of_the_tree()
Q = <T, X, {}>
while Q not empty & size(T) < limit
      N, X_N, C_N  = pop(Q)
      Z_N  = random(X_N, C_N)
      y_Z  = b(Z), y = b(X_N)
      if same_class(y U y_Z)
            continue
      S = best_split(X_N U Z_N, y U y_Z)
      S'= best_m-of-n_split(S)
      N = update_with_split(N, S')
      for each condition c in S'
            C = new_child_of(N)
            C_C = C_N U {c}
            X_C = select_with_constraints(X_N, C_N)
            put(Q, <C, X_C, C_C>)
```

black box auditing → (line 06: y_Z = b(Z), y = b(X_N))

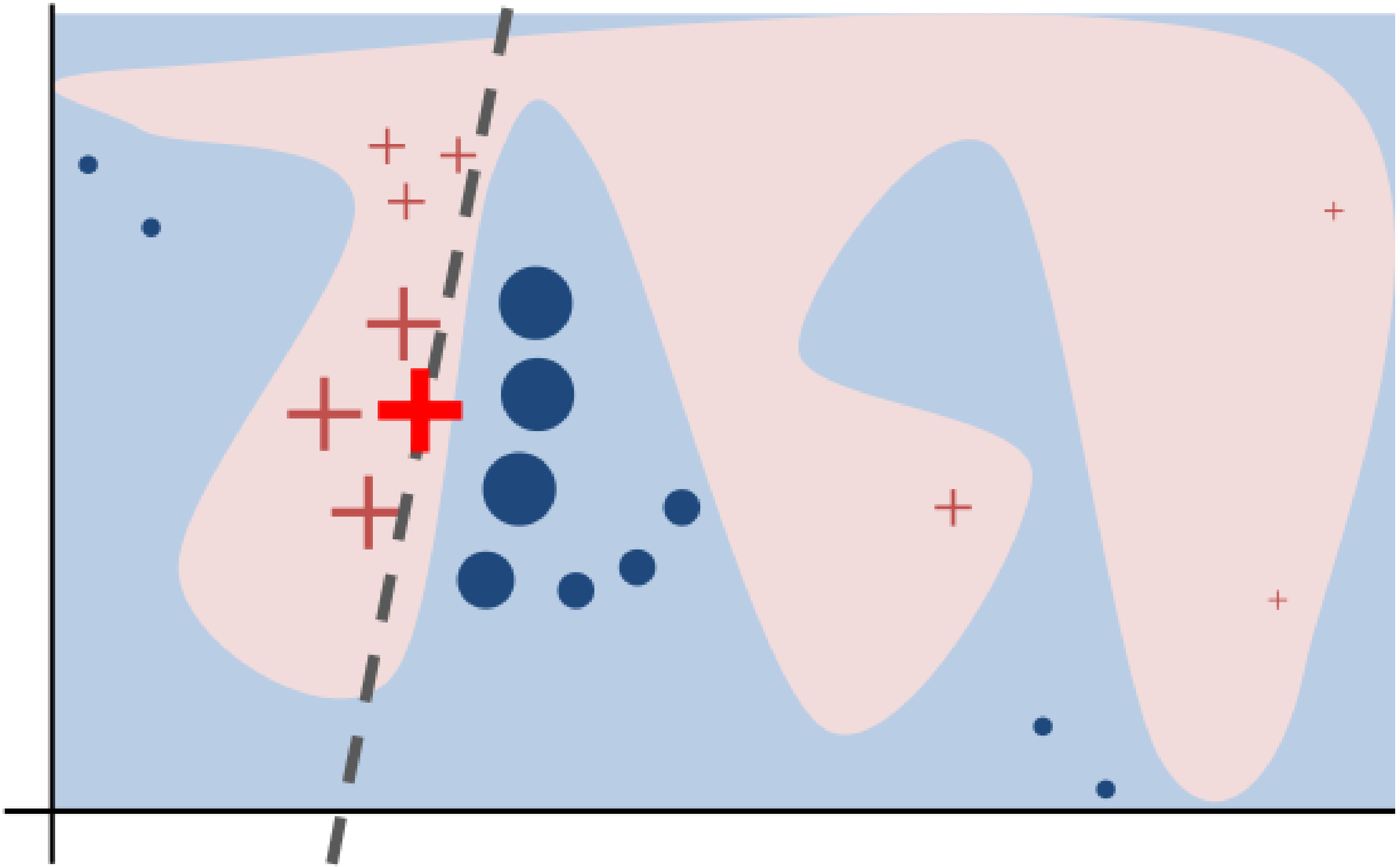


- Mark Craven and JudeW. Shavlik. 1996. ***Extracting tree-structured representations of trained networks***. NIPS.

LIME

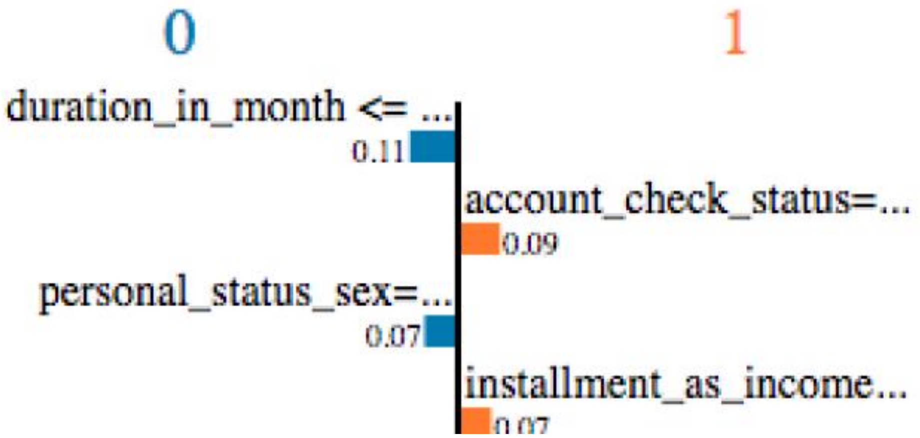
Local Explanation

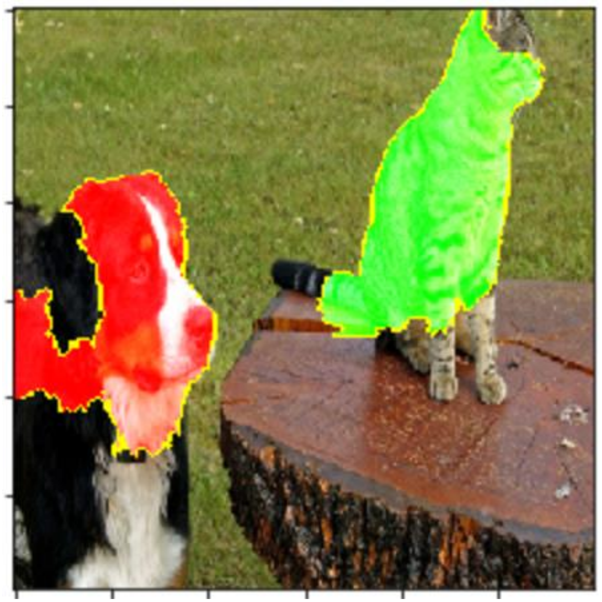
- The overall decision boundary is complex
- In the neighborhood of a single decision, the boundary is simple
- A single decision can be explained by auditing the black box around the given instance and learning a ***local*** decision.



Local Interpretable Model-agnostic Explanations

- Local model-agnostic explainer that reveals the black box decisions through features importance/saliency maps.
- It locally approximates the behavior of a black box with a local surrogate expressed as a liner regressor (with Lasso or Ridge penalization).
- Synthetic neighbors are weighted w.r.t. the distance with the instance to explain.





LIME

Sepal length	Sepal width	Petal length	Petal width	b(setosa)	b(versic)	b(virgi)
3	4	3	6	0.1	0.7	0.2

LIME

Sepal length	Sepal width	Petal length	Petal width	b(setosa)	b(versic)	b(virgi)
3	4	3	6	0.1	0.7	0.2
3	4	5	6	0.4	0.4	0.6

LIME

Sepal length	Sepal width	Petal length	Petal width	b(setosa)	b(versic)	b(virgi)
3	4	3	6	0.1	0.7	0.2
3	4	5	6	0.4	0.4	0.6
3	2	3	8	0.3	0.6	0.1

LIME

Sepal length	Sepal width	Petal length	Petal width	b(setosa)	b(versic)	b(virgi)
3	4	3	6	0.1	0.7	0.2
3	4	5	6	0.4	0.4	0.6
3	2	3	8	0.3	0.6	0.1
5	2	3	6	0.0	0.3	0.7
2	4	4	7	0.0	0.8	0.2

LIME

Sepal length	Sepal width	Petal length	Petal width	b(setosa)	b(versic)	b(virgi)
3	4	3	6	0.1	0.7	0.2
3	4	5	6	0.4	0.4	0.6
3	2	3	8	0.3	0.6	0.1
5	2	3	6	0.0	0.3	0.7
2	4	4	7	0.0	0.8	0.2

Train a Linear Regressor

LIME

Sepal length	Sepal width	Petal length	Petal width	b(setosa)	b(versic)	b(virgi)
3	4	3	6	0.1	0.7	0.2
3	4	5	6	0.4	0.4	0.6
3	2	3	8	0.3	0.6	0.1
5	2	3	6	0.0	0.3	0.7
2	4	4	7	0.0	0.8	0.2

Train a Linear Regressor

Returns the coefficients as Explanation

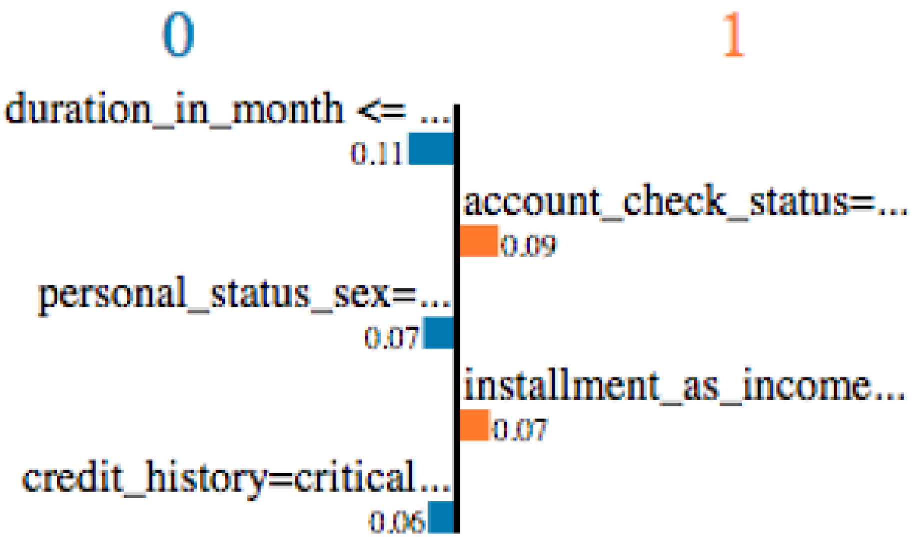
LIME

```
Z = {}
x instance to explain
x' = real2interpretable(x)
for i in {1, 2, …, N}
        z_i= sample_around(x')
        z = interpretabel2real(z')
        Z = Z U {<z_i, b(z_i), d(x, z)>}
w = solve_Lasso(Z, k)
return w
```

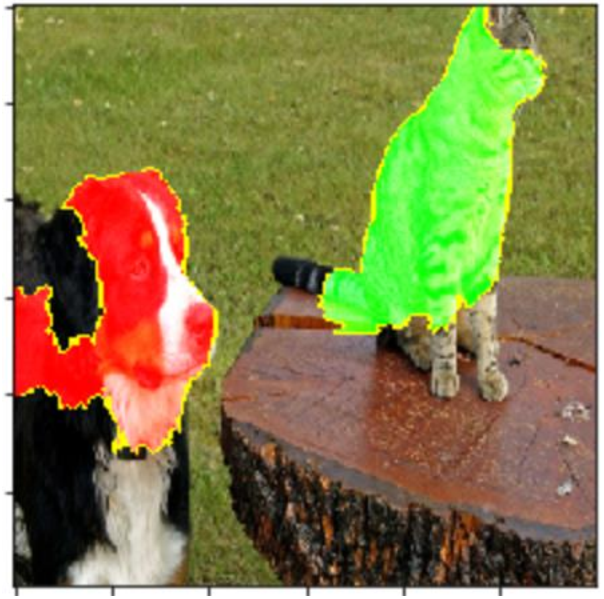
black box auditing

- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. KDD.

Features Importance

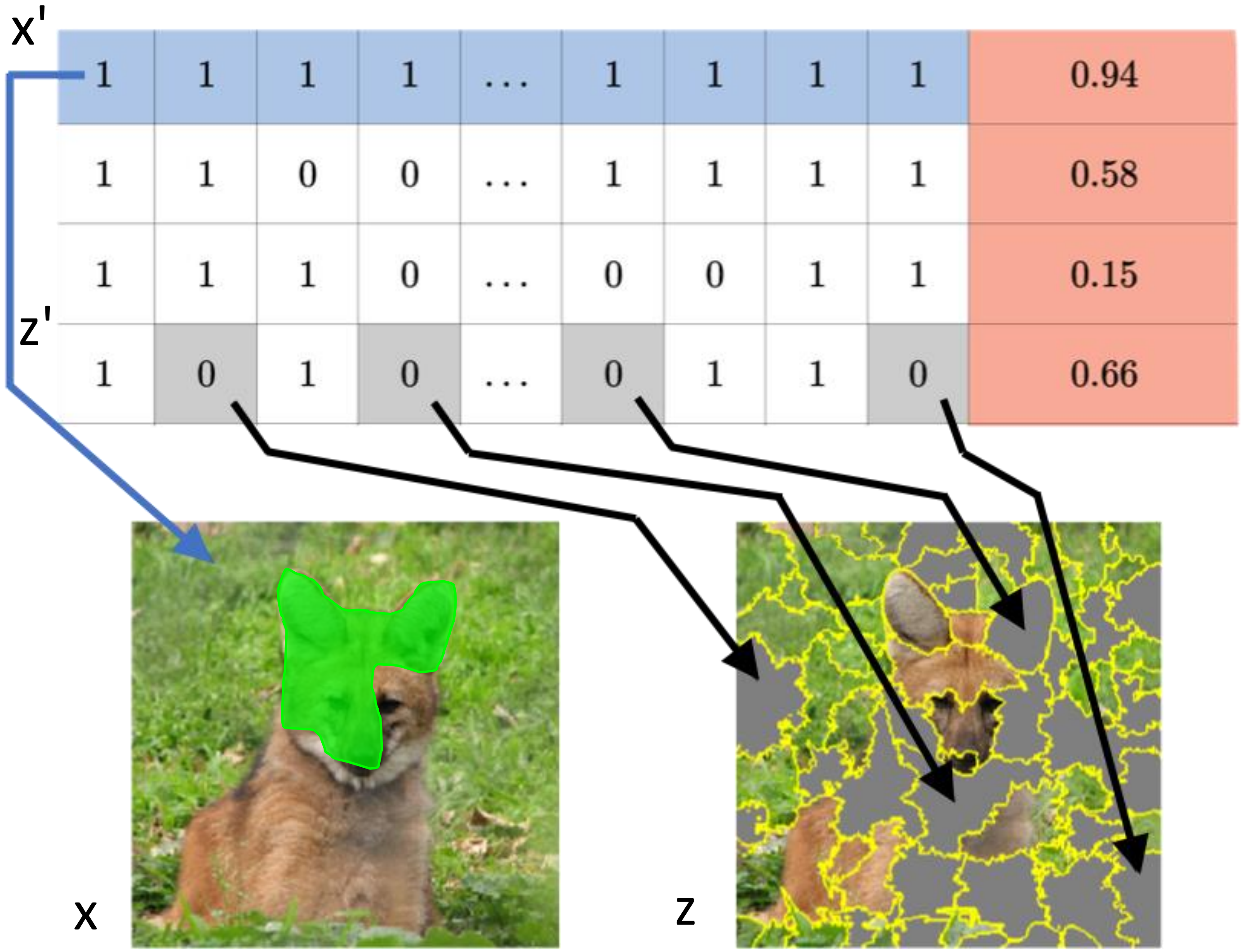


Saliency Map



LIME

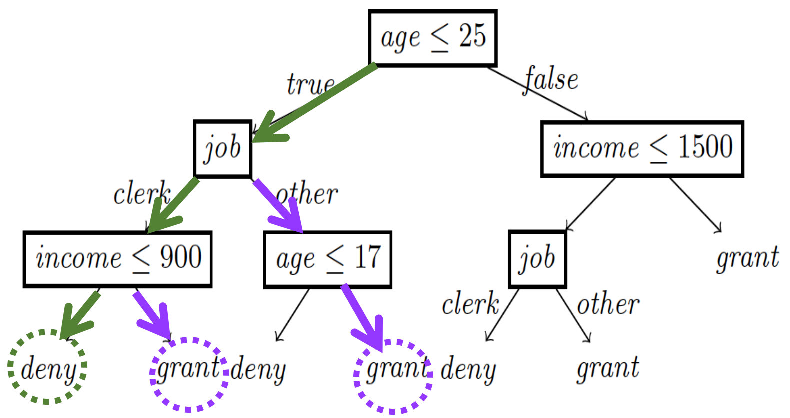
- LIME ***turns*** an image x to a vector x’ of interpretable superpixels expressing presence/absence.
- It ***generates*** a synthetic neighborhood Z by randomly perturbing x’ and labels them with the black box.
- It ***trains*** a linear regression model (interpretable and locally faithful) and assigns a weight to each superpixel.



LORE

*LO*cal *R*ule-based *E*xplainer

- LORE extends LIME adopting as local surrogate a decision tree classifier and by generating synthetic instances through a genetic procedure that accounts for both instances with the same labels and different ones.
- It can be generalized to work on images and text using the same data representation adopted by LIME.



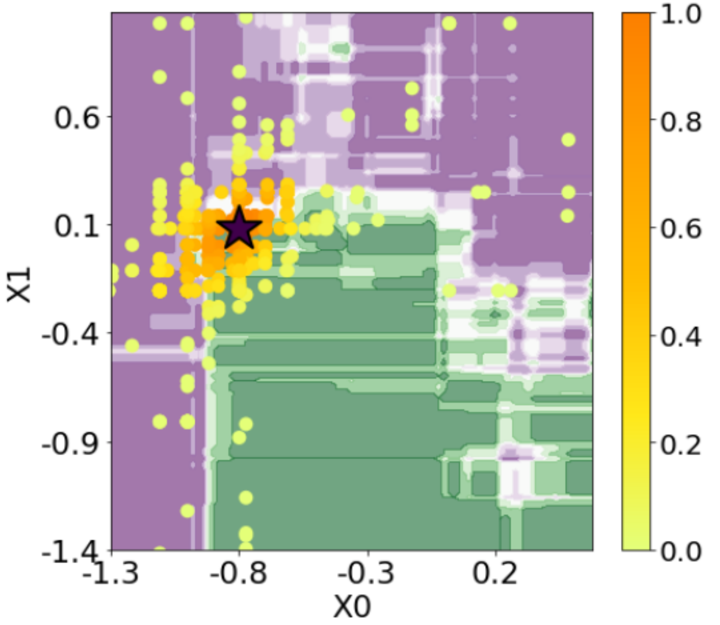
LORE

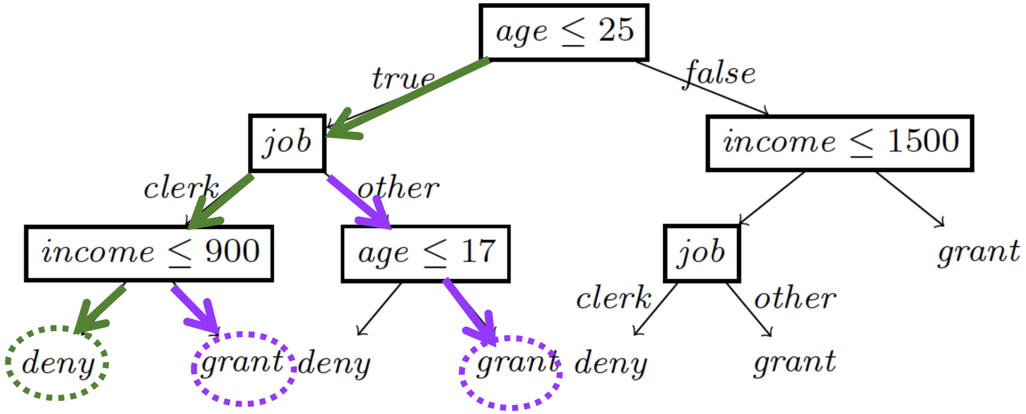
parent 1	25	clerk	10k	yes
parent 2	30	other	5k	no
children 1	25	other	5k	yes
children 2	30	clerk	10k	no

parent	25	clerk	10k	yes
children	27	clerk	7k	yes

```
x instance to explain
Z= = geneticNeighborhood(x, fitness=, N/2)
Z≠ = geneticNeighborhood(x, fitness≠, N/2)
Z = Z= ∪ Z≠
c = buildTree(Z, b(Z))
r = (p -> y) = extractRule(c, x)
ϕ = extractCounterfactual(c, r, x)
return e = <r, ϕ>
```

black box auditing





r = {age ≤ 25, job = clerk, income ≤ 900} -> deny

Φ = {({income > 900} -> grant),
({17 ≤ age < 25, job = other} -> grant)}

LORE

$$x_1 = \{ \text{Education} = \text{Bachelors}, \text{Occupation} = \text{Prof-specialty}, \text{Sex} = \text{Male}, \text{NativeCountry} = \text{Vietnam}, \text{Age} = 35, \text{Workclass} = 3, \text{HoursWeek} = 40, \text{Race} = \text{Asian-Pac-Islander}, \text{MaritialStatus} = \text{Married-civ}, \text{Relationship} = \text{Husband}, \text{CapitalGain} = 0, \text{CapitalLoss} = 0 \}, > 50k$$

$$r_{lore} = \{ \text{Education} > \text{5-6th}, \text{Race} > 0.86, \text{WorkClass} \leq 3.41, \text{CapitalGain} \leq 20000, \text{CapitalLoss} \leq 1306 \} \rightarrow > 50k$$

$$c_{lore} = \{ \text{CapitalLoss} \geq 436 \} \rightarrow \leq 50k$$

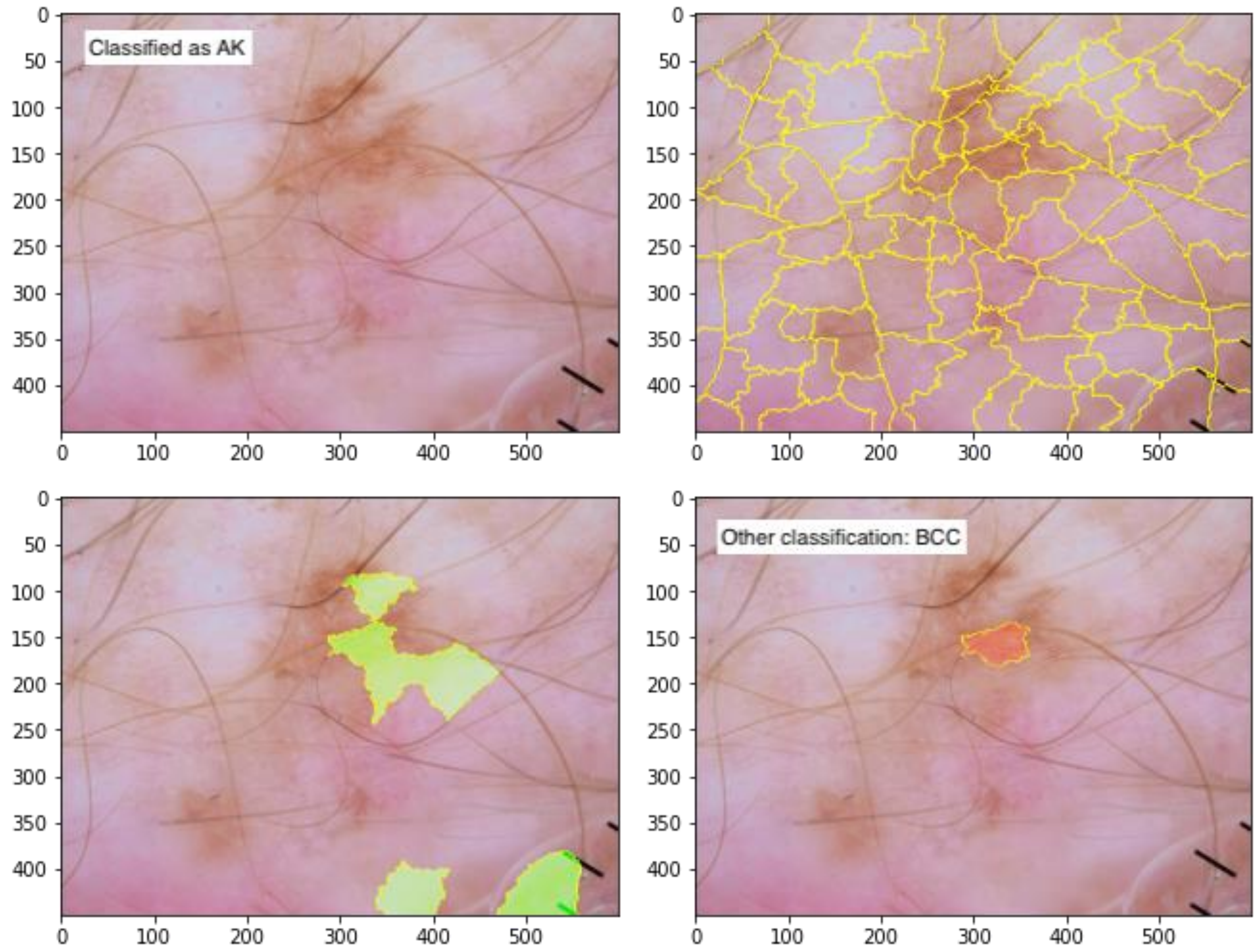
$$x_2 = \{ \text{Education} = \text{College}, \text{Occupation} = \text{Sales}, \text{Sex} = \text{Male}, \text{NativeCountry} = \text{US}, \text{Age} = 19, \text{Workclass} = 2, \text{HoursWeek} = 15, \text{Race} = \text{White}, \text{MaritialStatus} = \text{Married-civ}, \text{Relationship} = \text{Husband}, \text{CapitalGain} = 2880, \text{CapitalLoss} = 0 \}, \leq 50k$$

$$r_{lore} = \{ \text{Education} \leq \text{Masters}, \text{Occupation} > -0.34, \text{HoursWeek} \leq 40, \text{WorkClass} \leq 3.50, \text{CapitalGain} \leq 10000, \text{Age} \leq 34 \} \rightarrow \leq 50k$$

$$c_{lore} = \{ \text{Education} > \text{Masters} \} \rightarrow > 50k$$
$$\{ \text{CapitalGain} > 20000 \} \rightarrow > 50k$$
$$\{ \text{Occupation} \leq -0.34 \} \rightarrow > 50k$$

LORE on Medical Images

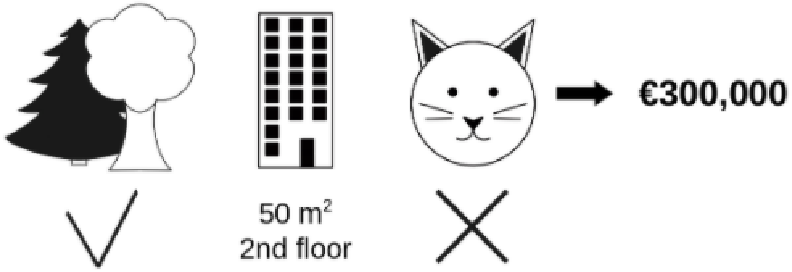
- The goal is to classify dermoscopic images among categories such as: Melanoma (MEL), Melanocytic Nevus (NV); Basal Cell Carcinoma (BCC), Actinic Keratosis (AK), etc.
- The original is classified as AK
- The counterfactual as BCC.



SHAP

Shapely Values

- A prediction can be explained by assuming that each feature value of the instance is a "player" in a game where the prediction is the payout. Shapley values -- a method from coalitional game theory -- tells us how to fairly distribute the "payout" among the features.
- Example: A black box predicts apartment prices. For a certain apartment it predicts €300,000 and you need to explain this prediction. The apartment has an area of 50 m^2, is located on the 2nd floor, has a park nearby and cats are banned.

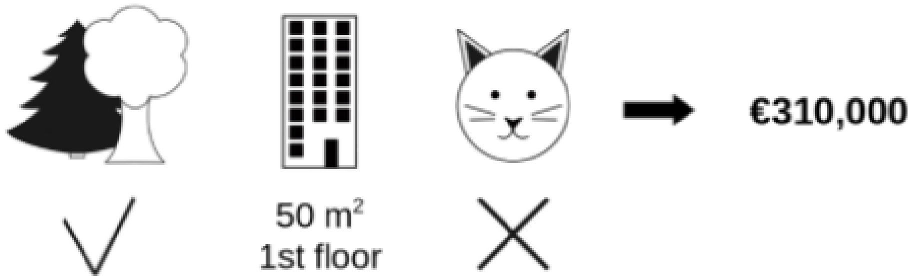


Shapely Values and Game Theory

- The average prediction is €310,000. How much has each feature value contributed to the prediction compared to the average prediction?
- The "game" is the prediction task for a single instance of the dataset.
- The "gain" is the actual prediction for this instance minus the average prediction for all instances.
- The "players" are the feature values of the instance that collaborate to receive the gain (= predict a certain value).
- The explanation could be: The park-nearby contributed €30,000; area-50 contributed €10,000; floor-2nd contributed €0; cat-banned contributed -€50,000. The contributions add up to -€10,000, the final prediction minus the average predicted apartment price.

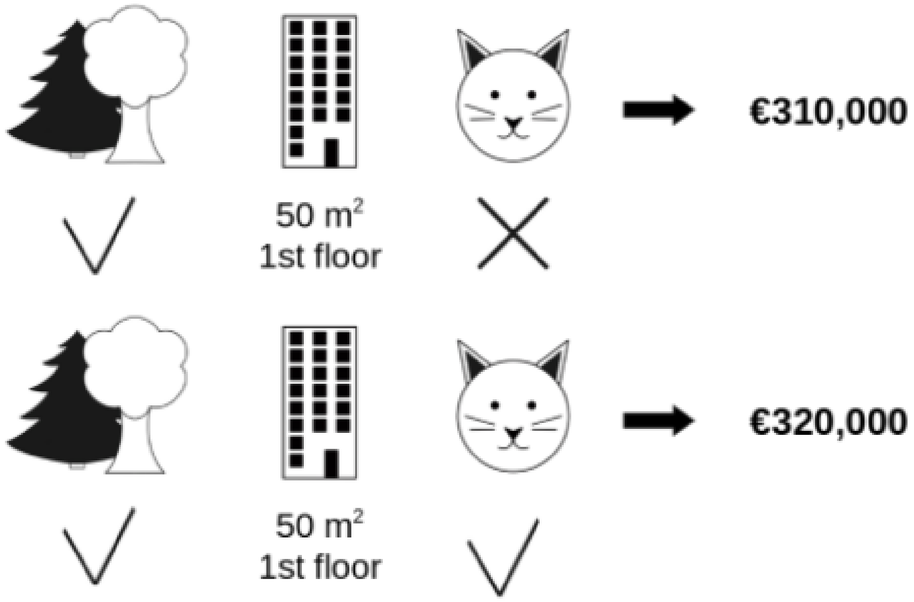
Shapely Values Example

- The Shapley value is the average marginal contribution of a feature value across all possible *coalitions* (combination of fixed feature values).
- We evaluate the contribution of *cat-banned when it is added to a coalition of park-nearby and area-50*.
- We simulate that only park-nearby, cat-banned and area-50 are in a coalition by randomly drawing another apartment from the data and using its value for the floor feature.
- The floor-2nd is replaced by the randomly drawn floor-1st.
- Then we predict the price of the apartment with this combination (€310,000).



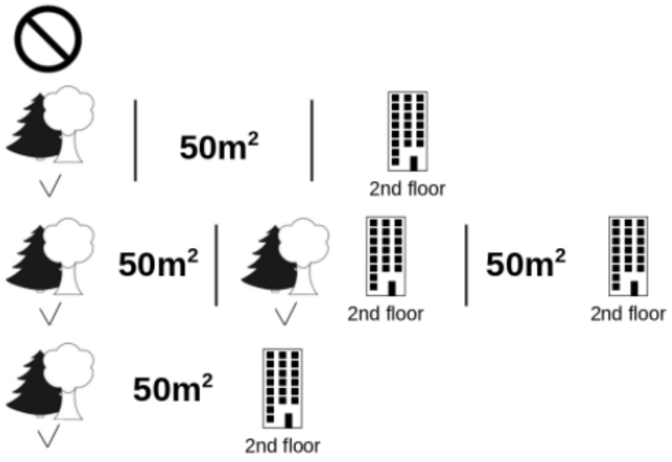
Shapely Values Example

- In a second step, we remove cat-banned from the coalition by replacing it with a random value of the cat allowed/banned from the randomly drawn apartment. In the example it was cat-allowed, but it could have been cat-banned again.
- We predict the apartment price for the coalition of park-nearby and area-50 (€320,000).
- The contribution of cat-banned was €310,000 - €320,000 = -€10,000. This estimate depends on the values of the randomly drawn apartment that served as a "donor" for the cat and floor feature values.
- We get better estimates if we repeat this sampling step and average the contributions.

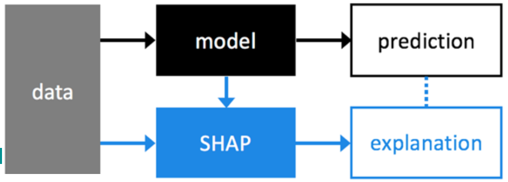


Shapely Values Example

- We repeat this computation for all possible coalitions.
- The Shapley value is the average of all the marginal contributions to all possible coalitions.
- The computation time increases exponentially with the number of features.
- For each of these coalitions we compute the predicted apartment price with and without the feature value cat-banned and take the difference to get the marginal contribution.
- We replace the feature values of features that are not in a coalition with random feature values from the apartment dataset to get a prediction from the black box.
- If we estimate the Shapley values for all feature values, we get the complete distribution of the prediction (minus the average) among the feature values.



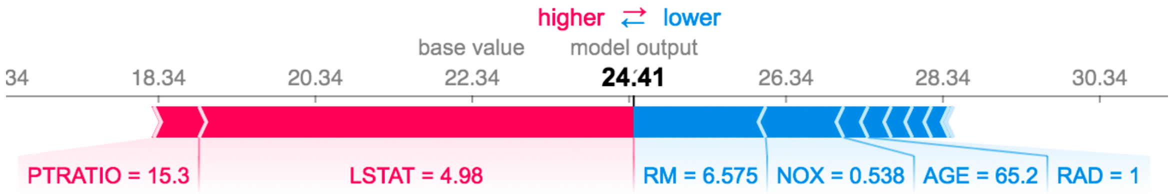
SHAP

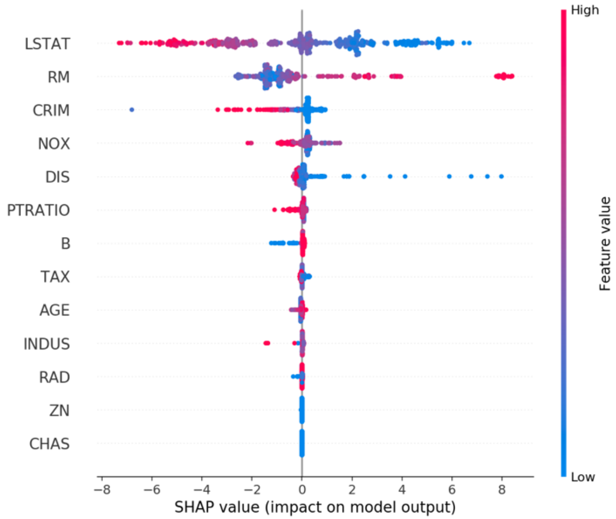


- SHAP (SHapley Additive exPlanations) assigns each feature an importance value for a particular prediction by means of an additive feature attribution method.
- It assigns an importance value to each feature that represents the effect on the model prediction of including that feature

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i,$$

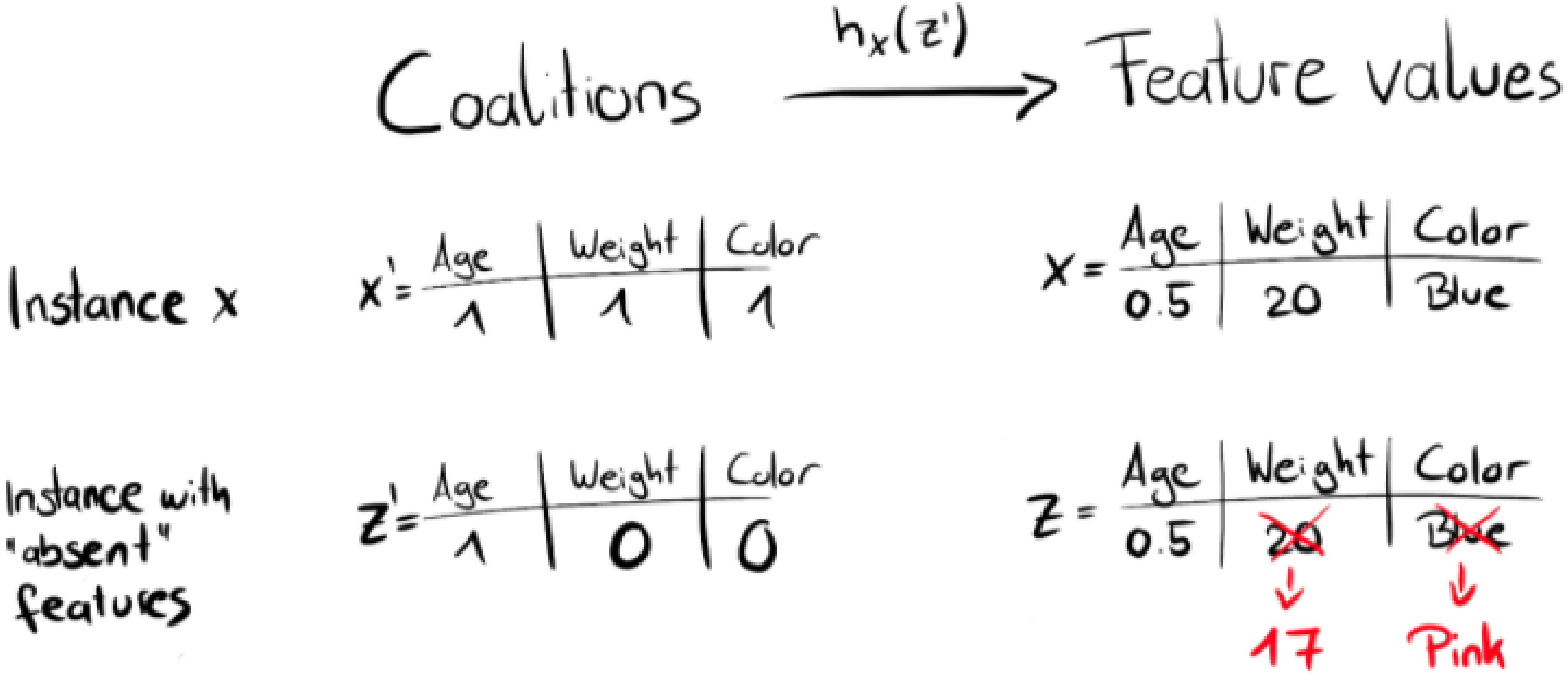
$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right]$$



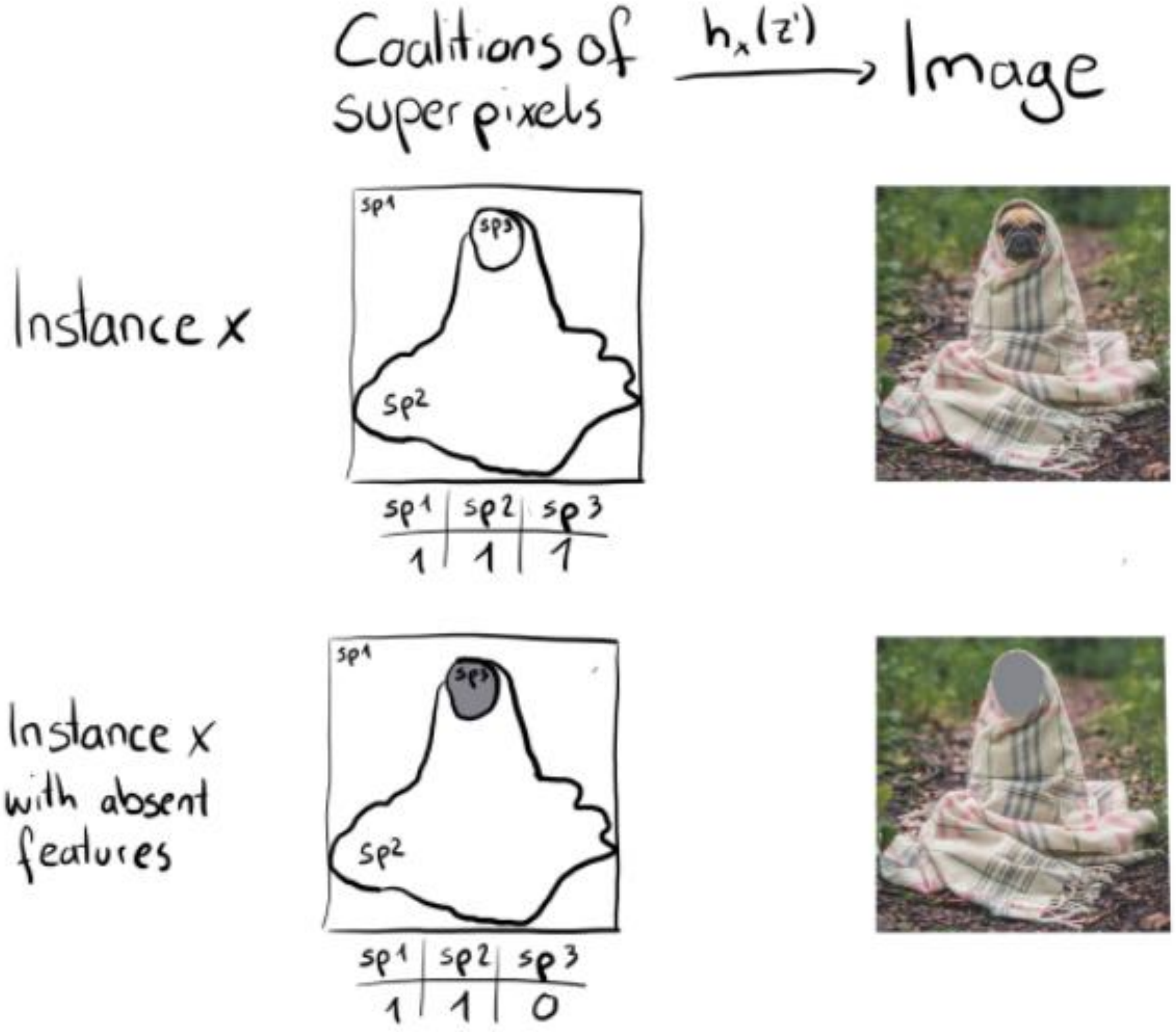


- Lundberg, Scott M., and Su-In Lee. ***A unified approach to interpreting model predictions***. *Advances in Neural Information Processing Systems*. 2017.

SHAP on Tabular Data



SHAP on Images

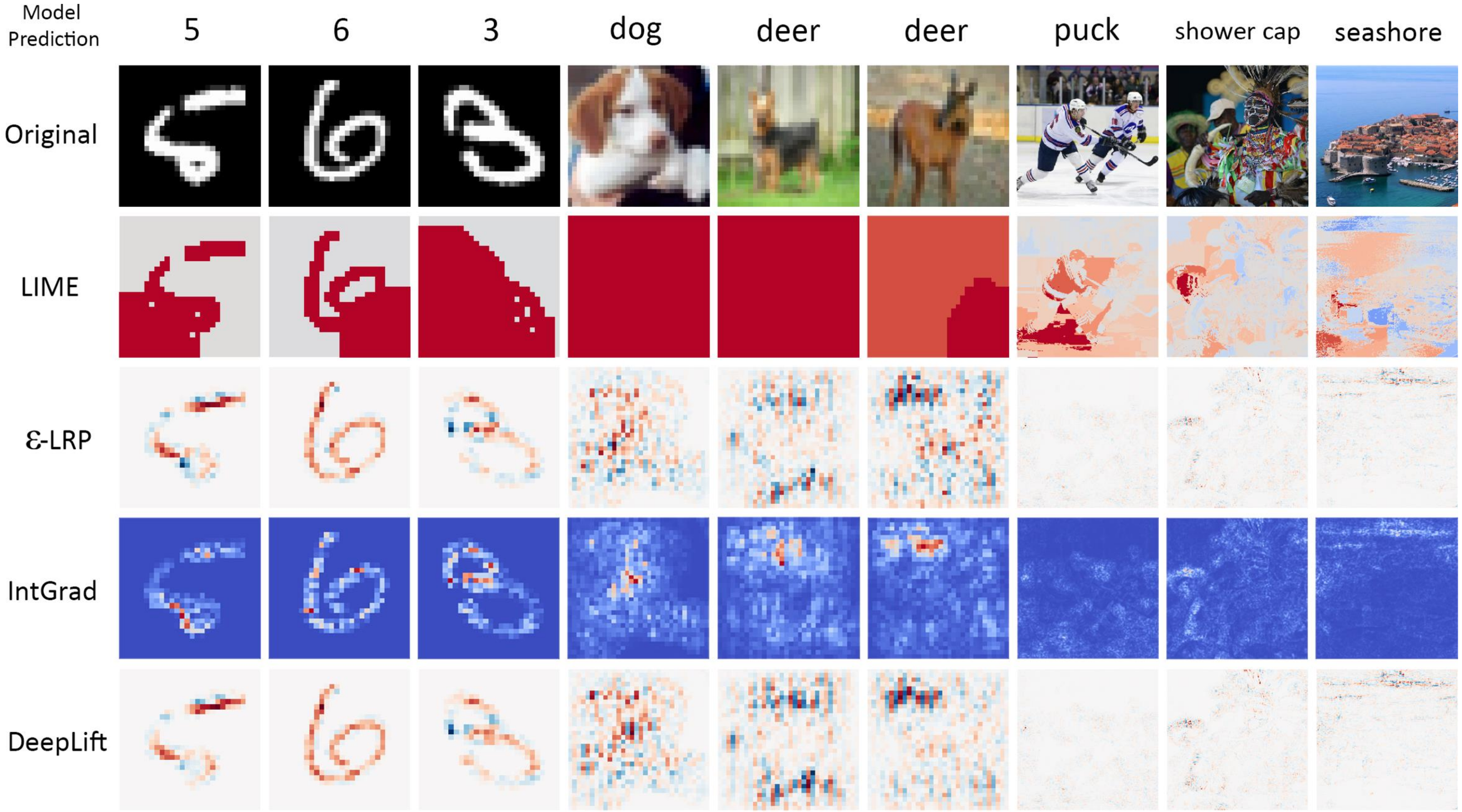


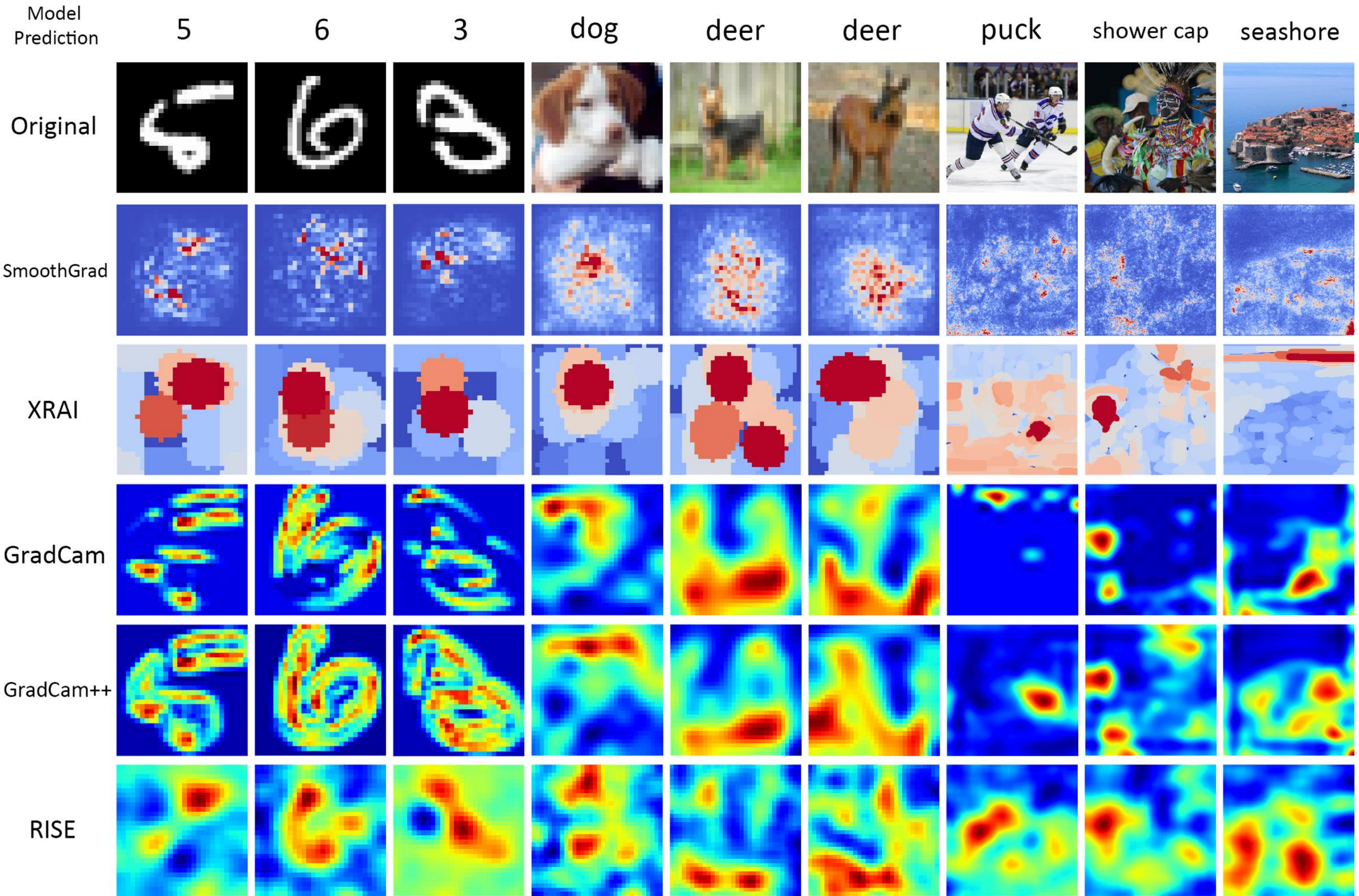
Saliency Maps

Saliency Maps

- A saliency map is an image in which a pixel's brightness represents how salient the pixel is. A positive value (red) means that the pixel has contributed positively to the classification, while a negative one (blue) means that has contributed negatively.
- There are two methods for creating SMs.
  1. Assign to ***every pixel*** a saliency value.
  2. Segment the image into different ***pixel groups (superpixels or segments)*** and then assign a saliency value for each group.

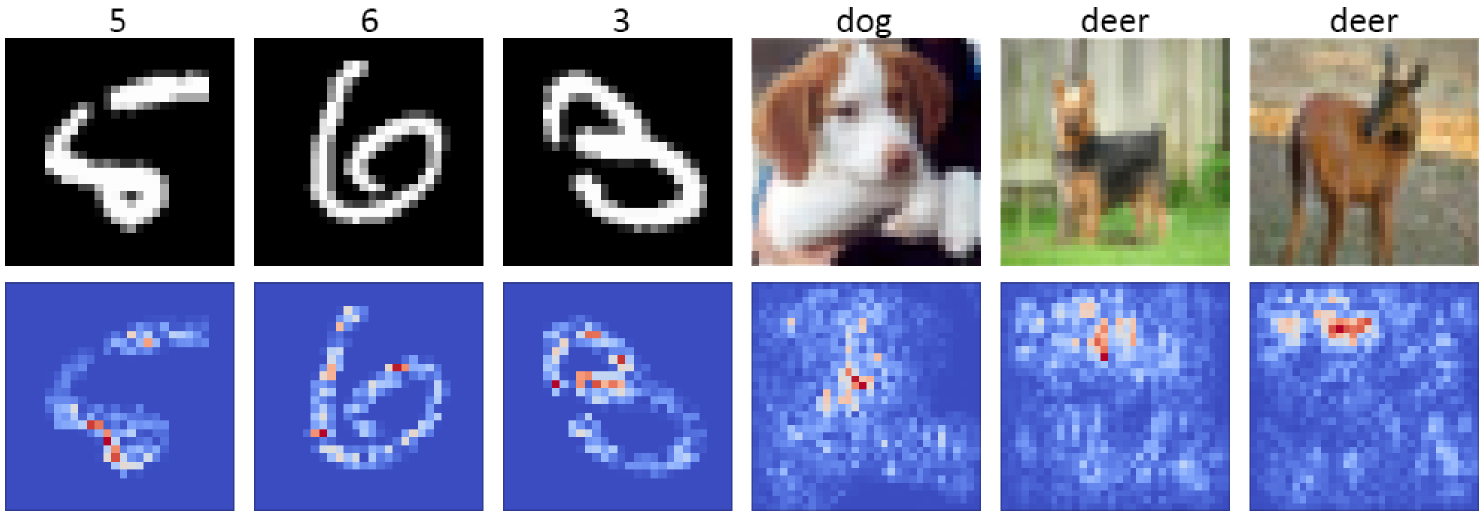
Saliency Maps




Model Prediction
5
6
3
dog
deer
deer
puck
shower cap
seashore
Original
SmoothGrad
XRAI
GradCam
GradCam++
RISE

Integrated Gradient

- INTGRAD can only be applied to differentiable models.
- INTGRAD constructs a path from the baseline image x' to the input x and computes the gradients of points along the path.
- The points are taken by overlapping x with x', and gradually modifying the opacity of x. Saliency maps are obtained by cumulating the gradients of these points.



- Mukund Sundararajan, Ankur Taly, Qiqi Yan. ***Axiomatic Attribution for Deep Networks***. arXiv preprint arXiv:1703.01365. 2017

MASK

01 x instance to explain

02 ***varying*** x into x′ maximizing b(x)**~**b(x′) ← ***black box auditing***

03 the variation runs replacing a region R of x with:

constant value, noise, blurred image

04 reformulation: find ***smallest*** R such that $b(x_R) \ll b(x)$

- Ruth Fong and Andrea Vedaldi. 2017. ***Interpretable explanations of black boxes by meaningful perturbation***. arXiv:1704.03296 (2017).

Sentence Highlighting

Instance-based Explanations

Instance-based Explanations

- Instance-based explanation methods select particular instances of the dataset or generate synthetic instances to explain black box behaviors.
- Instance-based explainers are mainly local explainers.
- Instance-based explanations only make sense if we can represent an instance of the data in a humanly understandable way.
- This works well for:
  - images
  - tabular data with not many features
  - short texts

Instance-based Explanations

- We mainly recognize the following example-based explanations:

  - ***Prototypes***: a selection of representative instances having the same class of the instance under analysis. Among prototypes we also recognize:
    - ***Criticisms:*** instances that are not well represented by prototypes.
    - ***Influential Instances:*** training points that were the most influential for the training of the black-box or for the prediction itself.

  - **Counterfactuals:** a selection of representative instances having a different class w.r.t. the instance under analysis.

Counterfactual Explanations

- A counterfactual explanation describes a causal situation in the form: "If X had not occurred, Y would not have occurred".
- Thinking in counterfactual terms requires imagining a hypothetical reality that contradicts the observed facts.
- Even if the relationship between the inputs and the outcome to be predicted might not be causal, we can see the inputs of a model as the cause of the prediction.
- ***A counterfactual explanation of a prediction describes the smallest change to the feature values that changes the prediction to a predefined output.***

Counterfactual Explan

- Counterfactuals answer why a decision has been made by highlighting what changes in the input would lead to a different outcome.
- CF are not generalizations!!!

income: 1200$
car owner: no
other debts: yes

Denied!

income: 1200$
car owner: yes
other debts: yes

income: 1500$
car owner: no
other debts: yes

Accepted!

Generating Counterfactual Explanations

- A simple and naive approach to generating counterfactual explanations is ***searching by trial and error:*** randomly changing feature values of the instance of interest and stopping when the desired output is predicted.
- As an alternative we can define *a loss function* that consider the instance of interest, a counterfactual and the desired (counterfactual) outcome. Then, we can find the ***counterfactual explanation that minimizes this loss using an optimization algorithm***.
- Many methods proceed in this way but differ in their definition of the loss function and optimization method.

Counterfactuals with a Brute Force Procedure

age	income	other debts	car owner
25	1200$	yes	no

age	income	other debts	car owner
25	500$	yes	no

age	income	other debts	car owner
25	10000$	yes	no

age	income	other debts	car owner
25	1200$	no	no

age	income	other debts	car owner
25	1200$	yes	yes

age	income	other debts	car owner
25	500$	no	no

age	income	other debts	car owner
25	500$	yes	yes

Counterfactuals by Optimization Problems

- Most of the counterfactual explainers return counterfactuals by solving an optimization problem.
- The problem is typically designed through the *definition of a loss function* aimed at guaranteeing a set of desired properties.
- The objective is to find a counterfactual instance that minimizes this loss using an optimization (OPT) algorithm.

Optimized CF Search

Wachter et al. suggest minimizing the following loss:

$$L(x, x', y', \lambda) = \lambda \cdot (\hat{f}(x') - y')^2 + d(x, x')$$

$$d(x, x') = \sum_{j=1}^{p} \frac{|x_j - x'_j|}{MAD_j}$$

balance the prediction

$$MAD_j = \text{median}_{i \in \{1,...,n\}}(|x_{i,j} - \text{median}_{l \in \{1,...,n\}}(x_{l,j})|)$$

1. Sample a random CF x'
2. Optimize the loss L
3. If not $|\hat{f}(x') - y'| \leq \epsilon$
4. Increase Lambda. Go to 2.
5. Return the CF x' that minimizes the loss.

- Wachter, Sandra and Mittelstadt, Brent and Russell, Chris. ***Counterfactual explanations without opening the black box: Automated decisions and the GDPR.*** 2017. Harv. JL & Tech

Optimized CF Search

- The loss function minimized by Wachter et al. is

$$\lambda(b(x') - y')^2 + d(x, x')$$

- where the first term is the quadratic distance between the desired outcome ***y'*** and the classifier prediction on ***x'***, and the second term is the distance between ***x*** and ***x'***.
- Lambda balances the contribution of the first term against the second term.

Wachter S, Mittelstadt BD, Russell C (2017) Counterfactual explanations without opening the black box: Automated decisions and the GDPR. HarvJL & Tech 31:841

Distance Functions

- Manhattan distance weighed with the inverse median absolute deviation MAD (used by Wachter)

$$d(x, x') = \sum_{j=1}^{p} \frac{|x_j - x'_j|}{MAD_j} \quad MAD_j = \text{median}_{i \in \{1,...,n\}}(|x_{i,j} - \text{median}_{l \in \{1,...,n\}}(x_{l,j})|)$$

- Mixed Distance (used by Mothilal)

$$d(a, b) = \frac{m_{con}}{n_{con}} \sum_{i \in con} \frac{|a_i - b_i|}{MAD_i} + \frac{m_{cat}}{n_{cat}} \sum_{i \in cat} \mathbb{1}_{a_i \neq b_i}$$

DICE - Diverse Counterfactual Explanations

- DICE solves an optimization problem with penalization terms to ensure plausibility by similarity and diversity.
- It returns a set of ***k*** plausible and different counterfactuals for ***x***.

$$C(\boldsymbol{x}) = \underset{\boldsymbol{c}_1,\ldots,\boldsymbol{c}_k}{\arg\min} \frac{1}{k}\sum_{i=1}^{k} \mathrm{yloss}(f(\boldsymbol{c}_i), y) + \frac{\lambda_1}{k}\sum_{i=1}^{k} dist(\boldsymbol{c}_i, \boldsymbol{x}) - \lambda_2\, \mathrm{dpp\_diversity}(\boldsymbol{c}_1, \ldots, \boldsymbol{c}_k)$$

Mothilal RK, Sharma A, Tan C (2020) Explaining machine learning classifiers through diverse counterfactual explanations. In: FAT*, ACM, pp 607–617
Mothilal RK, Mahajan D, Tan C, Sharma A (2021) Towards unifying feature attribution and counterfactual explanations: Different means to the sameend. In: AIES, ACM, pp 652–663

Counterfactuals through Heuristic Strategies

- Heuristic strategies are typically much more efficient than optimization algorithms.
- Efficiency is paid with solutions that are not necessarily optimal.
- The search strategy is typically designed such that at each iteration, $\boldsymbol{x}'$ is updated with the objective of *minimizing a cost function*.
- The cost function is based on a local and heuristic choice aiming for a valid counterfactual similar to $\boldsymbol{x}$.

SEDC - Search for Explanations for Document Classification

- The search is guided by local improvements via best-first search with pruning.

- b("the quick brown fox jumps over the lazy dog") = y (0.8) Prob. of y

Input

- b("~~the~~ quick brown fox jumps over the lazy dog") = y (0.8)
- b("the ~~quick~~ brown fox jumps over the lazy dog") ≠ y'(0.3)
- b("the quick ~~brown~~ fox jumps over the lazy dog") = y (0.7)

Iter 1

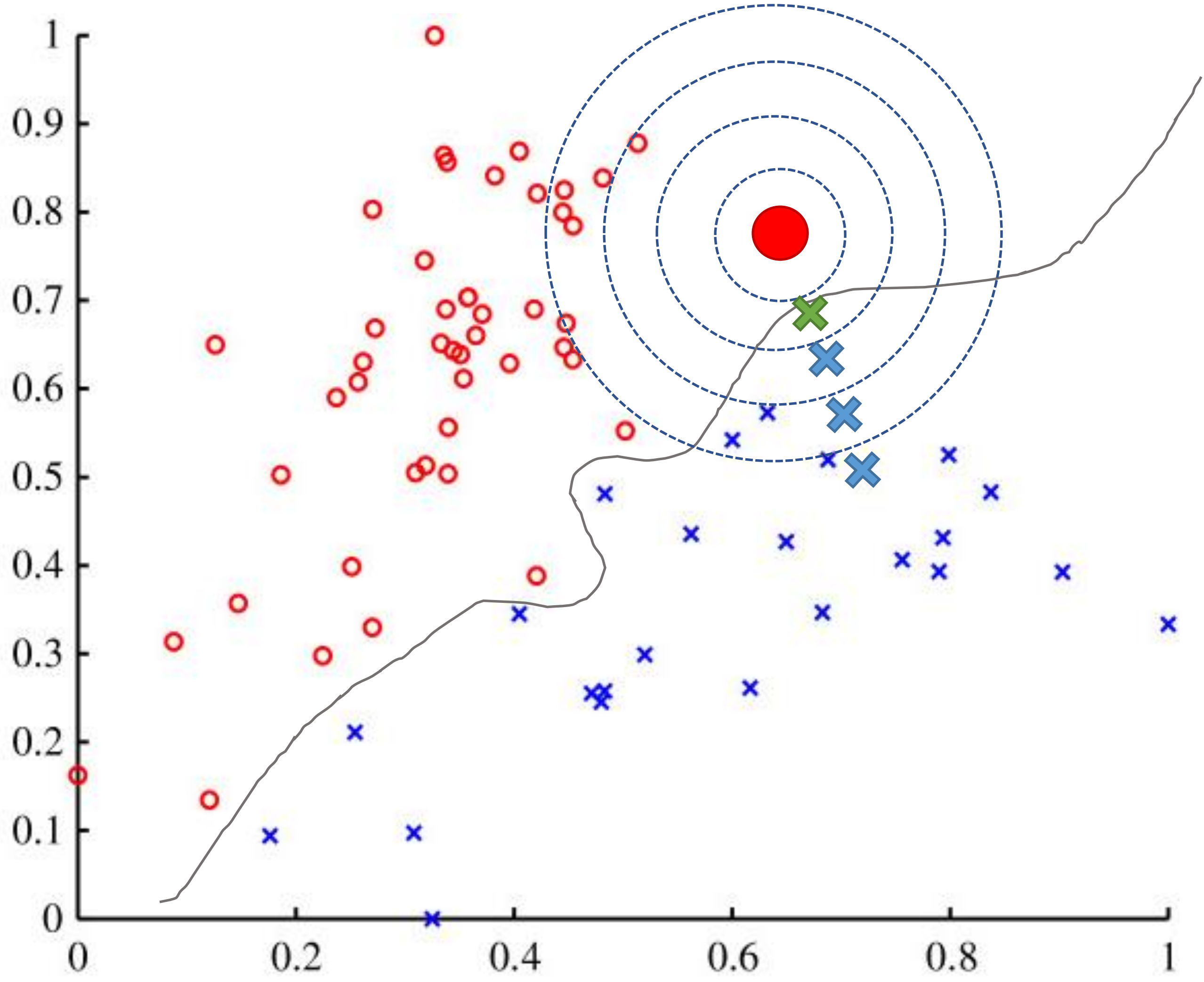
- …
- b("~~the~~ quick ~~brown~~ fox jumps over the lazy dog") = y (0.6)
- b("the quick ~~brown~~ ~~fox~~ jumps over the lazy dog") = y (0.6)
- b("the quick ~~brown~~ fox ~~jumps~~ over the lazy dog") ≠ y'(0.4)

Iter 2

Martens D, Provost FJ (2014) Explaining data-driven document classifications. MIS Q 38(1):73–99

GSG - Growing Spheres Generation

- GSG relies on a generative approach growing a *sphere* of synthetic instances around ***x*** to find the closest counterfactual ***x'***.
- GSG ignores in which direction the closest classification boundary might be.



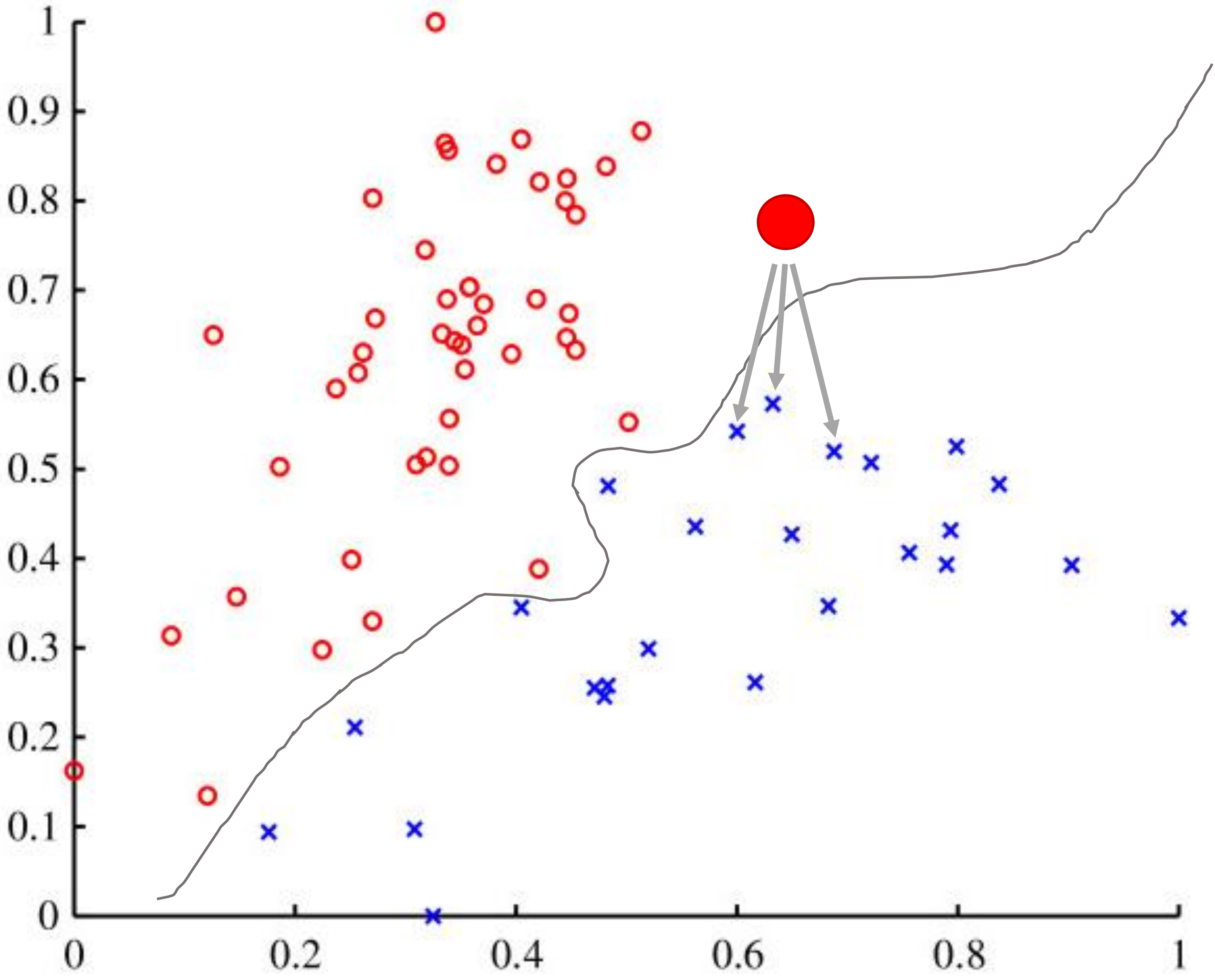
Laugel T, Lesot M, Marsala C, Renard X, Detyniecki M (2018) Comparison-based inverse classification for interpretability in machine learning. In:IPMU (1), Springer, Communications in Computer and Information Sci-ence, vol 853, pp 100–111

Counterfactuals with Instance-Based Strategies

- The very simple but effective idea of instance-based (or case-based) approaches for counterfactual explanation is to search into a reference population instances to be used as counterfactuals.

NNCE - Nearest-Neighbor Counterfactual Explainer

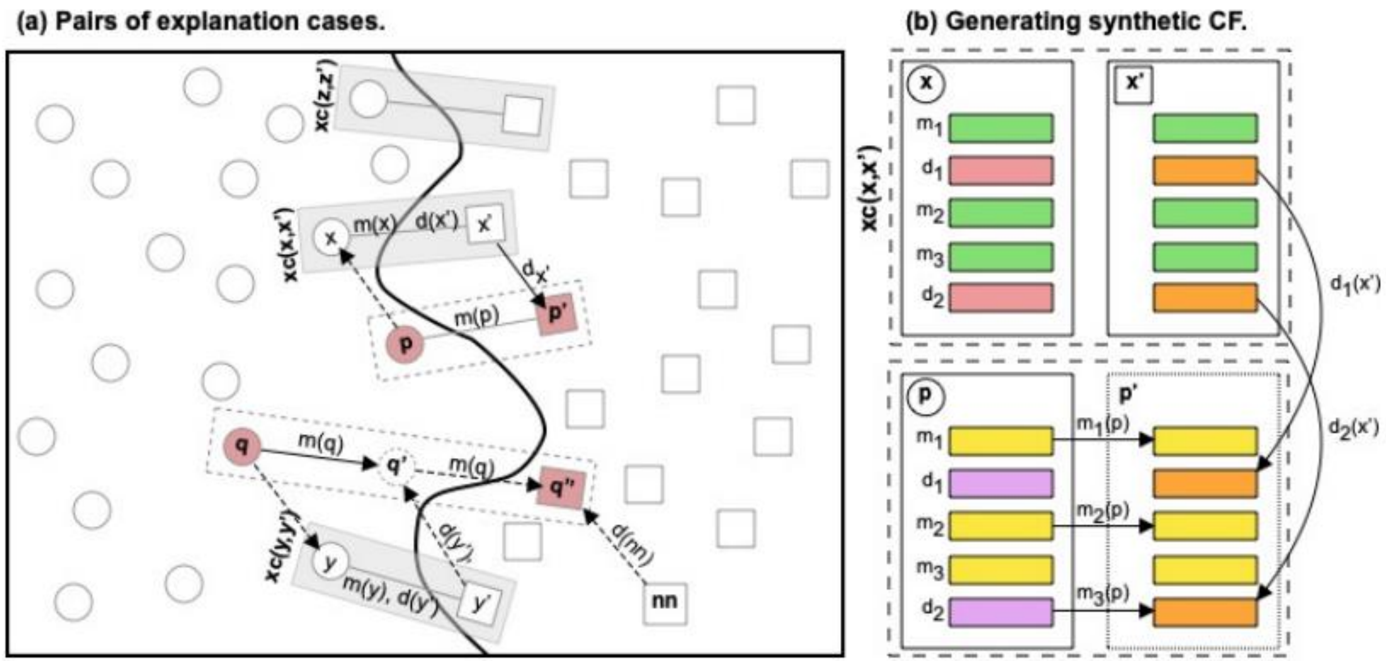
- NNCE is an endogenous counterfactual explainer inspired by kNN classifiers that select as counterfactual(s) the instances in ***x'∈X*** most similar to ***x*** and with a different label, i.e., ***b(x') ≠ b(x)***.
- Candidate counterfactuals are sorted with respect to the distance between ***x,*** and the ***k*** most similar ones are selected.



Shakhnarovich G, Darrell T, Indyk P (2008) Nearest-neighbor methods inlearn

CBCE - Case-Based Counterfactual Explainer

- CBCE refines NNCE.
- It adopts the notion of *explanation case (**xc**).*
- *G*iven ***X***, an ***xc*** is a couple of instances ***(x,x')*** such that ***(x,x')*** are the two most similar instances in ***X*** and ***b(x') ≠ b(x)***.



Keane MT, Smyth B (2020) Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable AI(XAI). In: ICCBR, Springer, Lecture Notes in Computer Science, vol 12311,pp 163–178

Take Home Message

Open The Black Box!

- ***To empower*** individual against undesired effects of automated decision making
- ***To reveal*** and protect new vulnerabilities
- ***To implement*** the “right of explanation”
- ***To improve*** industrial standards for developing AI-powered products, increasing the trust of companies and consumers
- ***To help*** people make better decisions
- ***To align*** algorithms with human values
- ***To preserve*** (and expand) human autonomy

Open Research Questions

- There is ***no agreement*** on ***what an explanation is***
- There is ***not a formalism*** for ***explanations***
- How to evaluate the ***goodness of explanations***?
- There is ***no work*** that seriously addresses the problem of ***quantifying*** the grade of ***comprehensibility*** of an explanation for humans
- What if there is a ***cost*** for querying a black box?

References

- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM computing surveys (CSUR), 51(5), 1-42.
- Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D., & Rinzivillo, S. (2021). Benchmarking and Survey of Explanation Methods for Black Box Models. arXiv preprint arXiv:2102.13076.
- Guidotti, R. (2022). Counterfactual explanations and how to find them: literature review and benchmarking. Data Mining and Knowledge Discovery, 1-55.
- Molnar, C. (2020). Interpretable machine learning. Lulu. com.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence, 267, 1-38.
- Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEWEE. 2018.
- Artelt, A., & Hammer, B. (2019). On the computation of counterfactual explanations--A survey. arXiv preprint arXiv:1911.07749.

- Artelt, A., & Hammer, B. (2019). On the computation of counterfactual explanations--A survey. arXiv preprint arXiv:1911.07749.

- Zhang, Y., & Chen, X. (2018). Explainable recommendation: A survey and new perspectives. arXiv preprint

References

- Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018, October). Explaining explanations: An overview of interpretability of machine learning. In 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA) (pp. 80-89). IEEE.
- Ribeiro, M. T., et al. " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD. 2016
- Lundberg, S., & Lee, S. I. A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874. 2017
- Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., & Turini, F. (2019). Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems*, *34*(6), 14-23.
- Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Ruggieri, S., & Turini, F. (2019, July). Meaningful explanations of Black Box AI decision systems. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 9780-9784).
- Guidotti, R., Monreale, A., Matwin, S., & Pedreschi, D. (2019, September). Black box explanation by learning image exemplars in the latent feature space. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 189-205). Springer, Cham.
- Setzu, M., Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., & Giannotti, F. (2021). GLocalX-From Local to Global Explanations of Black Box AI Models. Artificial Intelligence, 294, 103457.
- Guidotti, R. (2021). Evaluating local explanation methods on ground truth. Artificial Intelligence, 291, 103428.

References

- Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, *31*, 841.
- Mittelstadt, B., Russell, C., & Wachter, S. (2019, January). Explaining explanations in AI. In Proceedings of the conference on fairness, accountability, and transparency (pp. 279-288).
- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
- Freitas, A. A. (2014). Comprehensible classification models: a position paper. ACM SIGKDD explorations newsletter, 15(1), 1-10.
- Romei, A., & Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. The Knowledge Engineering Review, 29(5), 582-638.
- Craven, M. W., & Shavlik, J. W. (1996). Extracting tree-structured representations of trained networks. Advances in neural information processing systems, 24-30.
- Augasta, M. G., & Kathirvalavakumar, T. (2012). Reverse engineering the neural networks for rule extraction in classification problems. Neural processing letters, 35(2), 131-150.
- Fong, R. C., & Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. In Proceedings of the IEEE International Conference on Computer Vision (pp. 3429-3437).
- Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., & Flach, P. (2020, February). FACE: feasible and actionable counterfactual explanations. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (pp. 344-350).

References

- Cortez, P., & Embrechts, M. J. (2011, April). Opening black box data mining models using sensitivity analysis. In 2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM) (pp. 341-348). IEEE.
- Kim, B., Gilmer, J., Wattenberg, M., & Viégas, F. (2018). Tcav: Relative concept importance testing with linear concept activation
- Mothilal, R. K., Sharma, A., & Tan, C. (2020, January). Explaining machine learning classifiers through diverse counterfactual explanations. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 607-617). vectors.
- Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. (2017). Learning certifiably optimal rule lists for categorical data. arXiv preprint arXiv:1704.01701.
- Sundararajan, M., Taly, A., & Yan, Q. (2017, July). Axiomatic attribution for deep networks. In International Conference on Machine Learning (pp. 3319-3328). PMLR.
- Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825.
- Chen, J., Song, L., Wainwright, M., & Jordan, M. (2018, July). Learning to explain: An information-theoretic perspective on model interpretation. In International Conference on Machine Learning (pp. 883-892). PMLR.
- Dhurandhar, A., Chen, P. Y., Luss, R., Tu, C. C., Ting, P., Shanmugam, K., & Das, P. (2018). Explanations based on the missing: Towards contrastive explanations with pertinent negatives. arXiv preprint arXiv:1802.07623.
- Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences, 116(44), 22071-22080.

Explanation Toolboxes and Repositories

- https://github.com/jphall663/awesome-machine-learning-interpretability
- https://github.com/pbiecek/xai_resources
- https://github.com/ModelOriented/DrWhy
- https://fat-forensics.org/
- https://github.com/Trusted-AI/AIX360
- https://captum.ai/
- https://github.com/interpretml/interpret
- https://github.com/SeldonIO/alibi
- https://github.com/pair-code/what-if-tool