IDSIA

SUPSI

Explainable AI (XAI) Interpreting, Explaining and Visualising Machine and Reinforcement Learning

Alessandro Antonucci (alessandro@idsia.ch)

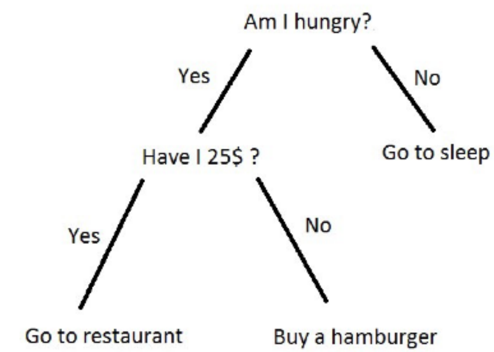
Senior Lecturer-Researcher @IDSIA USI-SUPSI

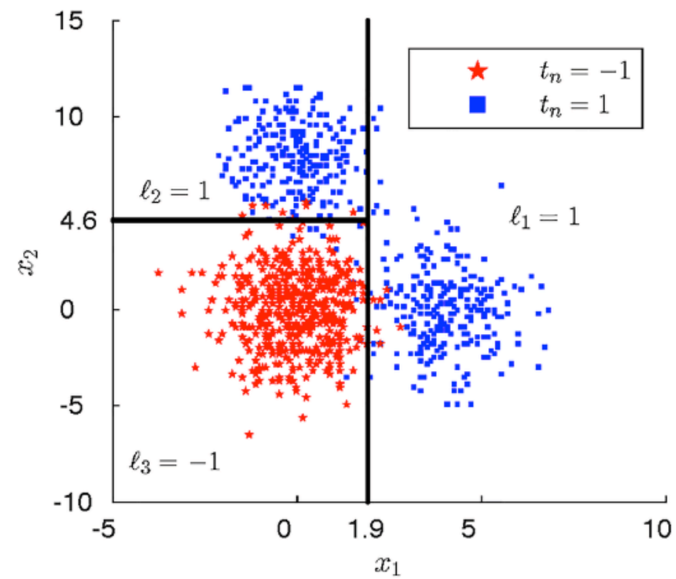
AIDD Spring School - Advanced Machine Learning for Drug Discovery

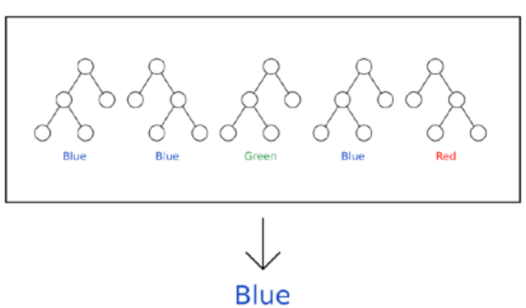
Lugano, May 12, 2022

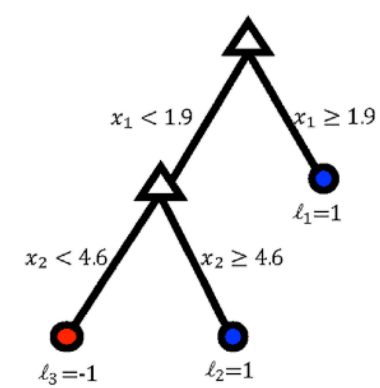
Do we really need XAI? Decision Trees

- "Self-explanatory" ML algorithms?
- Decision Tree Classifiers = recursive splits of data by purity measures giving rule-based classifiers (for discrete/continuous)
- Simple-but-powerful idea: Random Forests & Gradient Boosted Trees are DTC sophistications

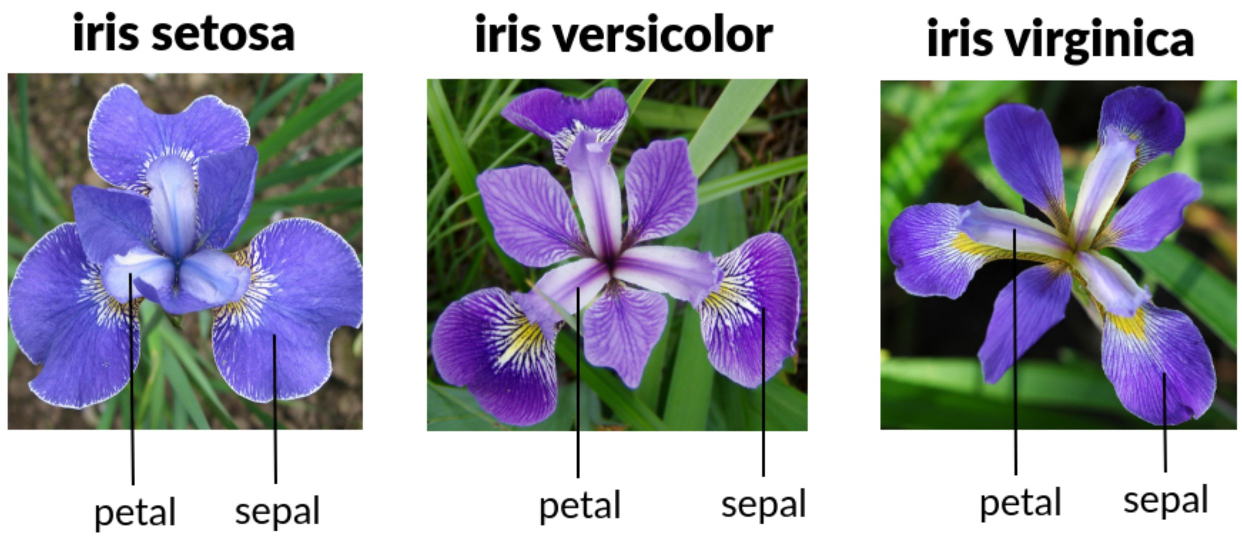








Explaining the Iris Data Set with Decision Trees?



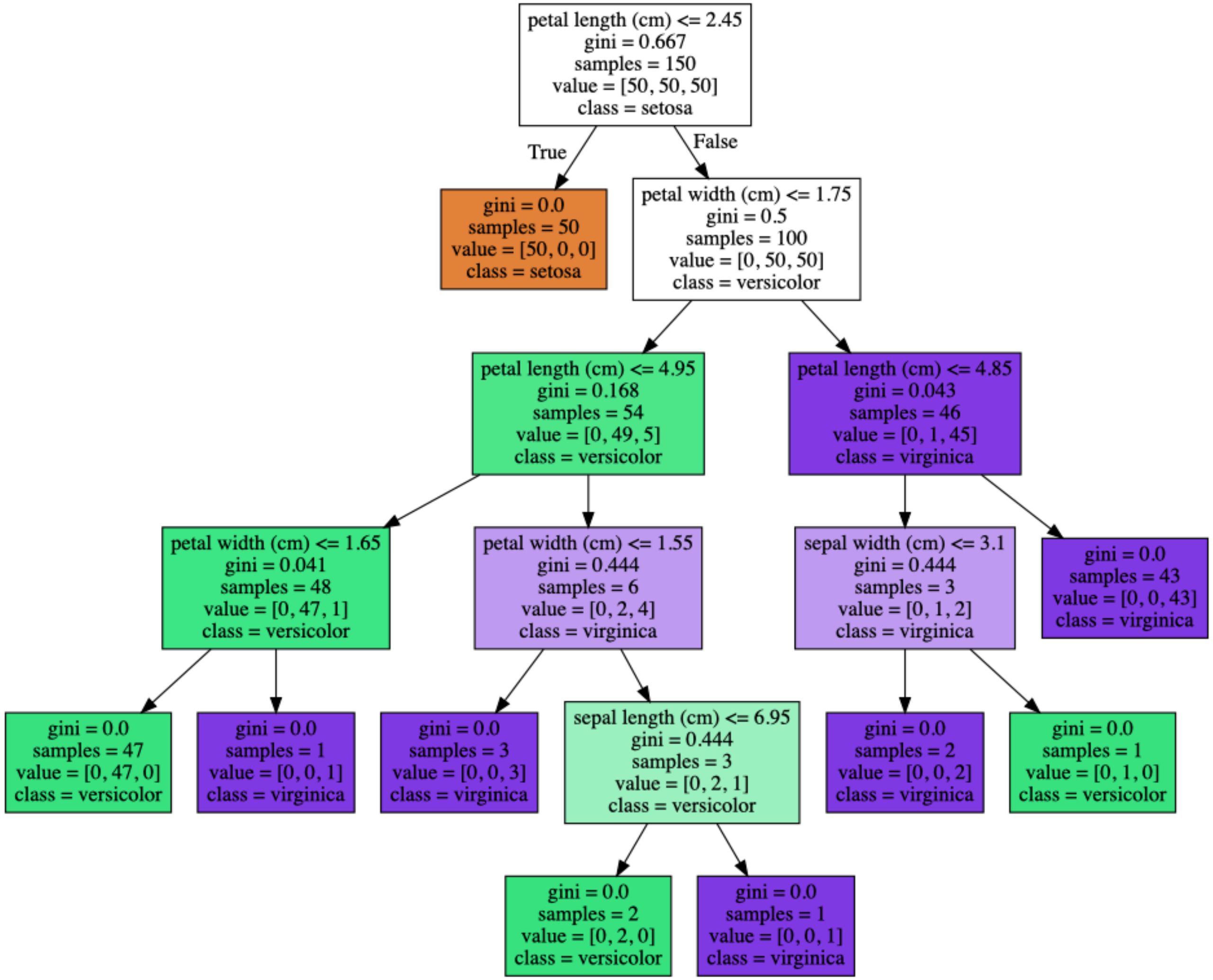
ID	PetalLength	PetalWidth	SepalLength	SepalWidth	Species
1	1.4	0.2	5.1	3.5	Iris-setosa
2	1.4	0.2	4.9	3.0	Iris-setosa
3	1.3	0.2	4.7	3.2	Iris-setosa
4	1.5	0.2	4.6	3.1	Iris-setosa
5	1.4	0.2	5.0	3.6	Iris-setosa
6	1.7	0.4	5.4	3.9	Iris-setosa
7	1.4	0.3	4.6	3.4	Iris-setosa
8	1.5	0.2	5.0	3.4	Iris-setosa
9	1.4	0.2	4.4	2.9	Iris-setosa
10	1.5	0.1	4.9	3.1	Iris-setosa
11	1.5	0.2	5.4	3.7	Iris-setosa
12	1.6	0.2	4.8	3.4	Iris-setosa
13	1.4	0.1	4.8	3.0	Iris-setosa
14	1.1	0.1	4.3	3.0	Iris-setosa
15	1.2	0.2	5.8	4.0	Iris-setosa

Classification task

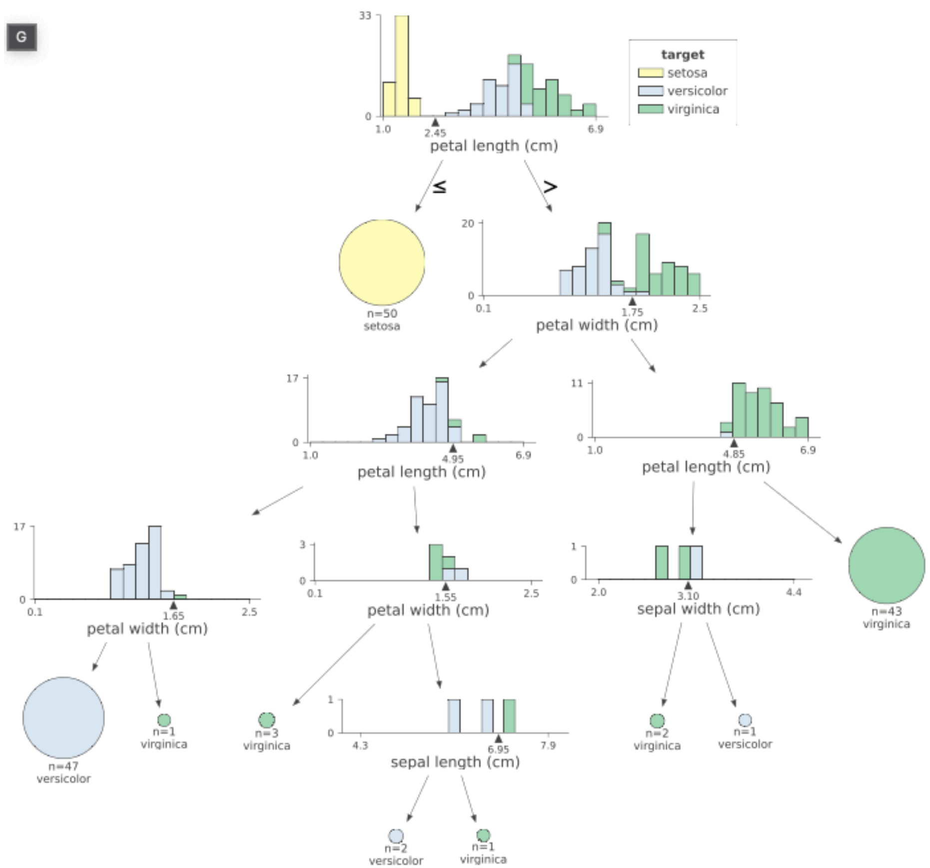
3 possible classes

4 numerical features

A (Complicated) Explanation of the Iris Data Set with DTs



A (Still Complicated) Explanation of the Iris Data Set with DTs

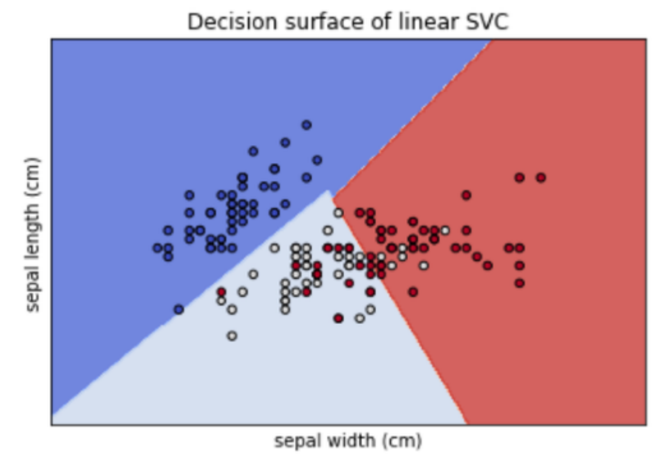


Informative, but only with "small" trees

Small tree = simple"rules" but ML is mostly used when rules are complex ...

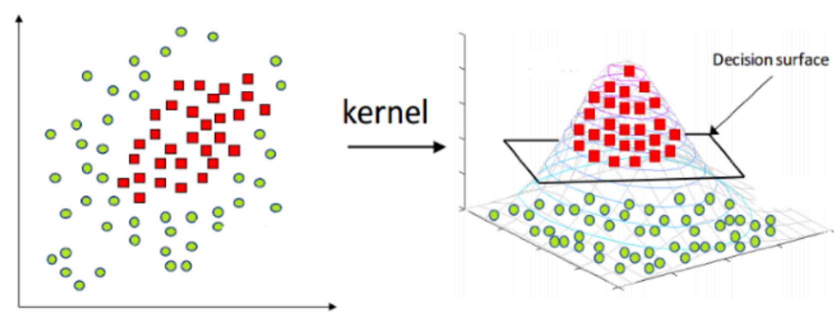
(Already) Interpretable Models

- Sparse or Low Dimensional **Linear** Models
  - Support Vector Machines (SVMs) line DTs with non-orthogonal separators
  - Regressors, NBC, etc., ...
- Rules can be obtained
- Linearity is often unrealistic, but kernel transformations leading to non-linear models can be used



	Conditions		Probability	Support
IF	IrregularShape AND Age ≥ 60	THEN malignancy risk is	85.22%	230
ELSE IF	SpiculatedMargin AND Age ≥ 45	THEN malignancy risk is	78.13%	64
ELSE IF	IllDefinedMargin AND Age ≥ 60	THEN malignancy risk is	69.23%	39
ELSE IF	IrregularShape	THEN malignancy risk is	63.40%	153
ELSE IF	LobularShape AND Density ≥ 2	THEN malignancy risk is	39.68%	63
ELSE IF	RoundShape AND Age ≥ 60	THEN malignancy risk is	26.09%	46
ELSE		THEN malignancy risk is	10.38%	366

Table 1: Falling rule list for mammographic mass dataset.



Generative Models (Bayesian Nets)

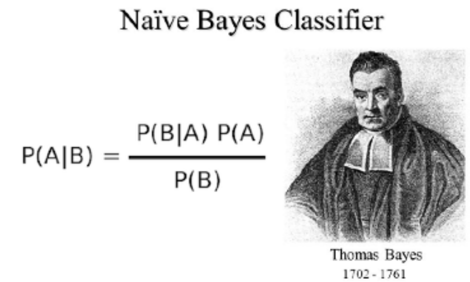
- Generative Modelling?

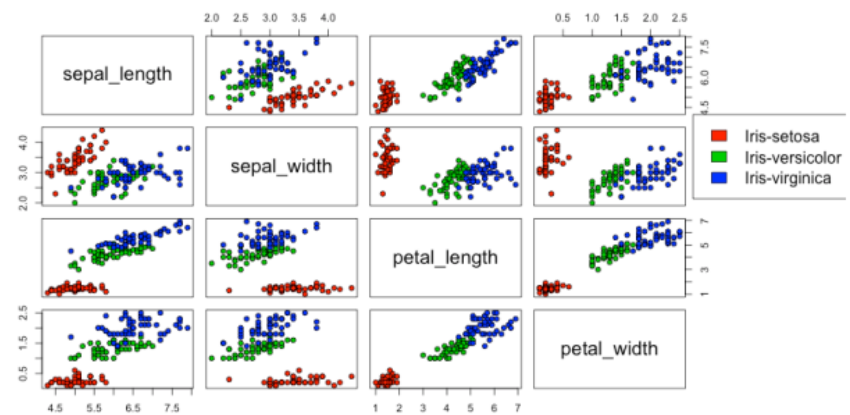
  Joint Probability $P(C, F_1, F_2, F_3, F_4)$
- Classification?

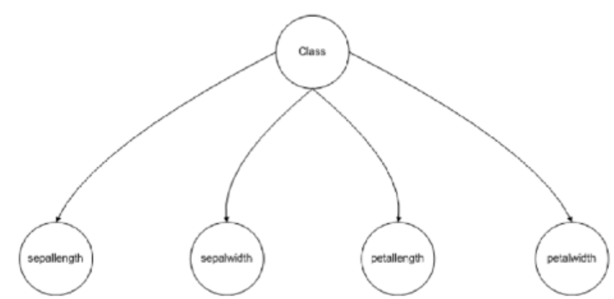
  $\arg\max_C P(C \,|\, F_1, F_2, F_3, F_4)$
- Explanations? $P(F_j \,|\, C)$
- Joint Elicitation?

  Decomposition by independence
- E.g., Naive Bayes Classifier

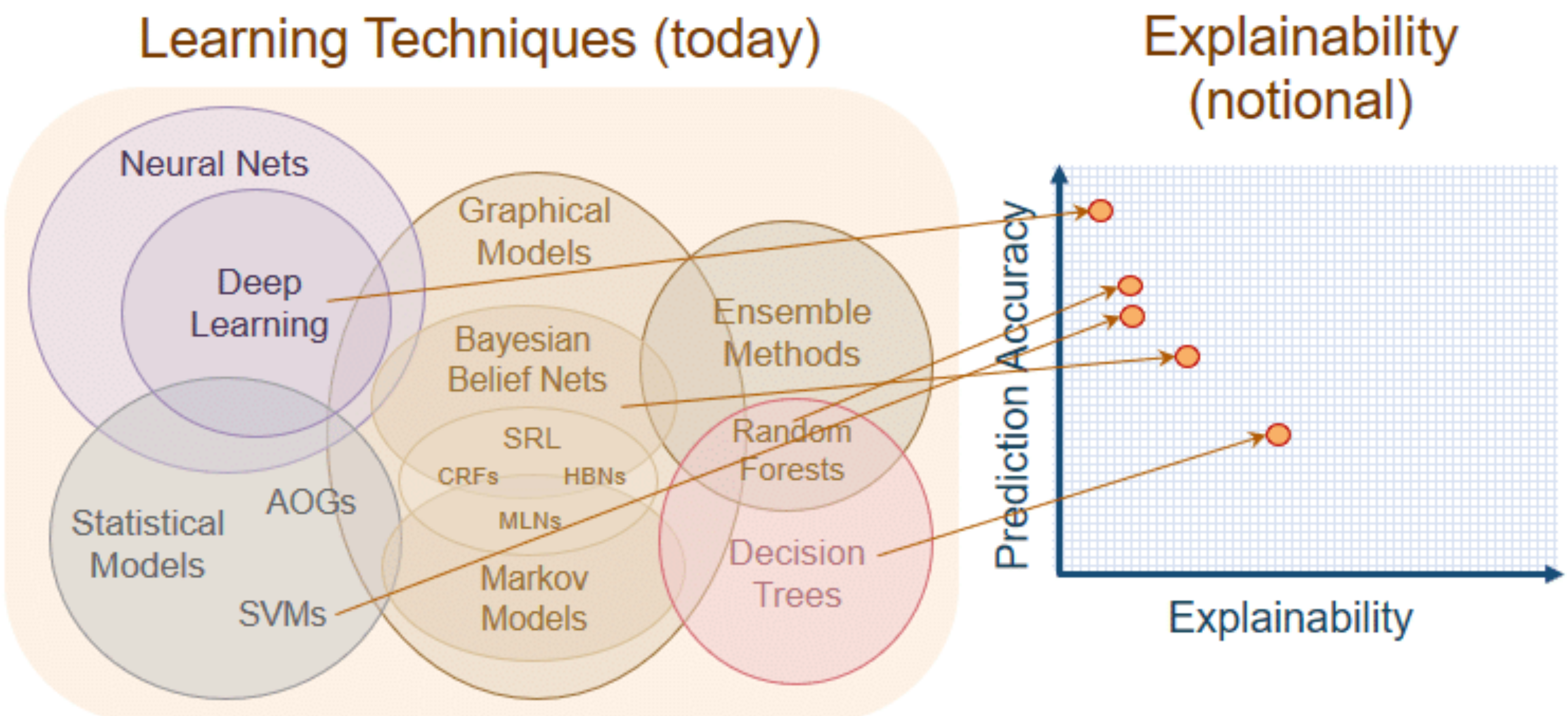
$$P(C, F_1, F_2, F_3, F_4) = P(C) \prod_{i=1}^{4} P(F_i \,|\, C)$$





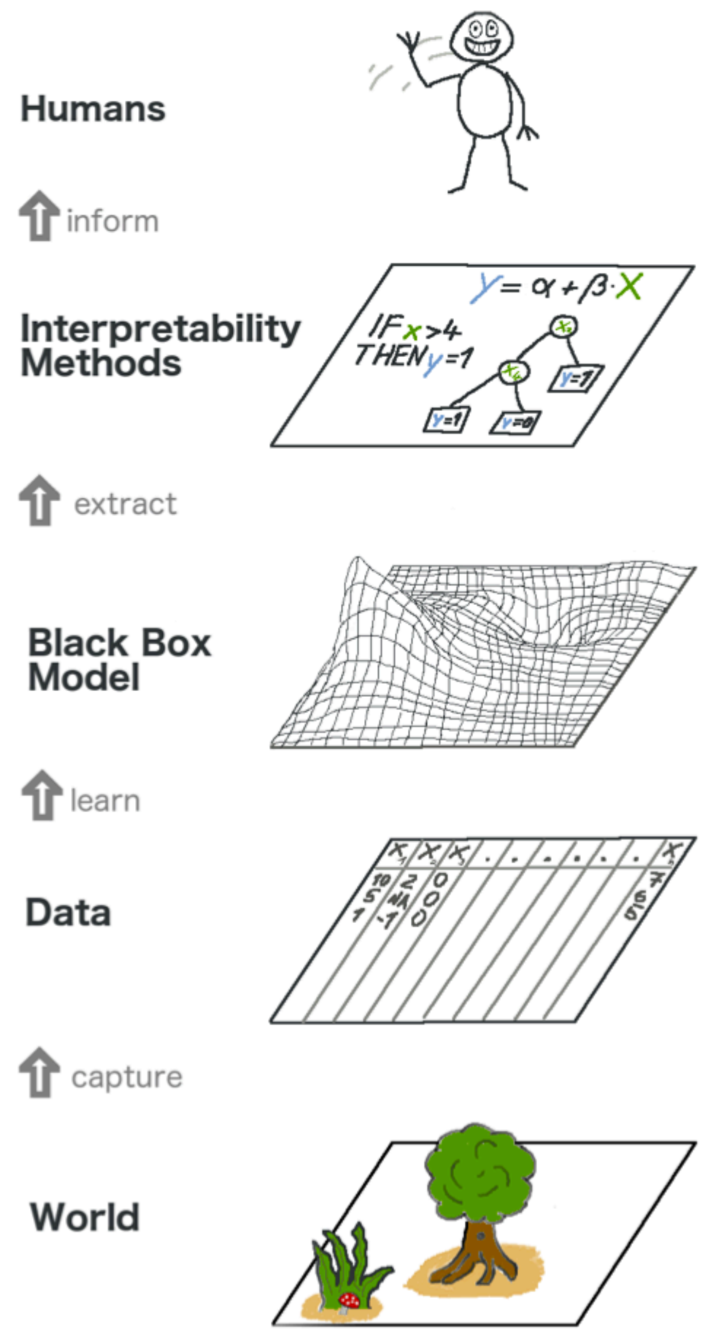


But Accuracy Matters ...



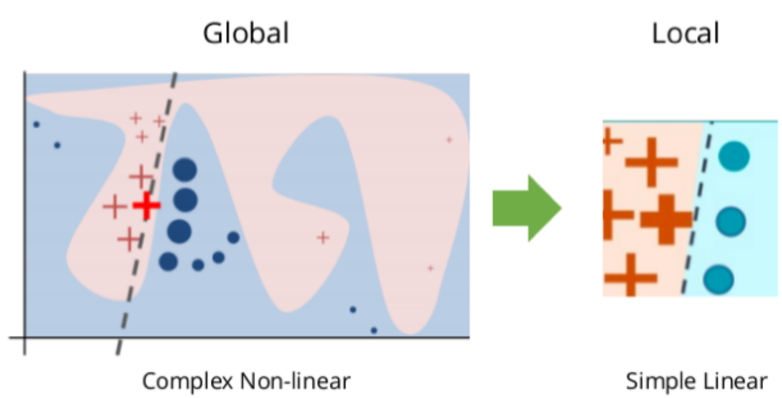
Towards Model-Agnostic Approaches

- Why model-agnostic?
- We want to use more powerful methods (ex. Deep Learning)
- Producing better explanations independently of the model

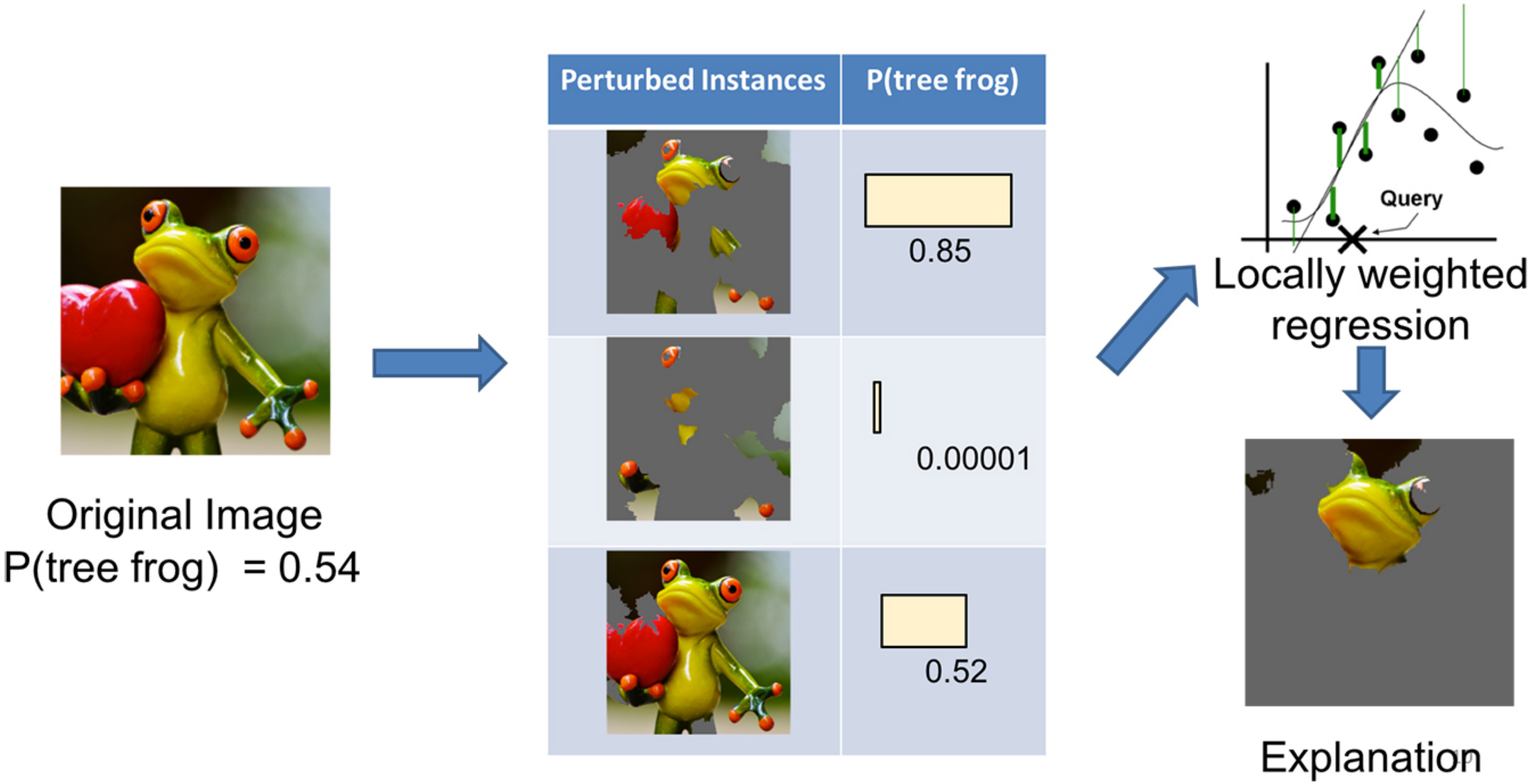


Local Interpretable Model-agnostic Explanations (LIME)

- First, but still popular, MA-XAI algorithm (Ribeiro et al., 2016)
- Simple and flexible idea working with categorical or continuous data, text, images (and open-source library)
- Given instance x, train surrogate model on a neighbourhood of x
- The ML algorithm annotates the neighbour instances
- Locality (neighbours) makes linearity a tenable assumption

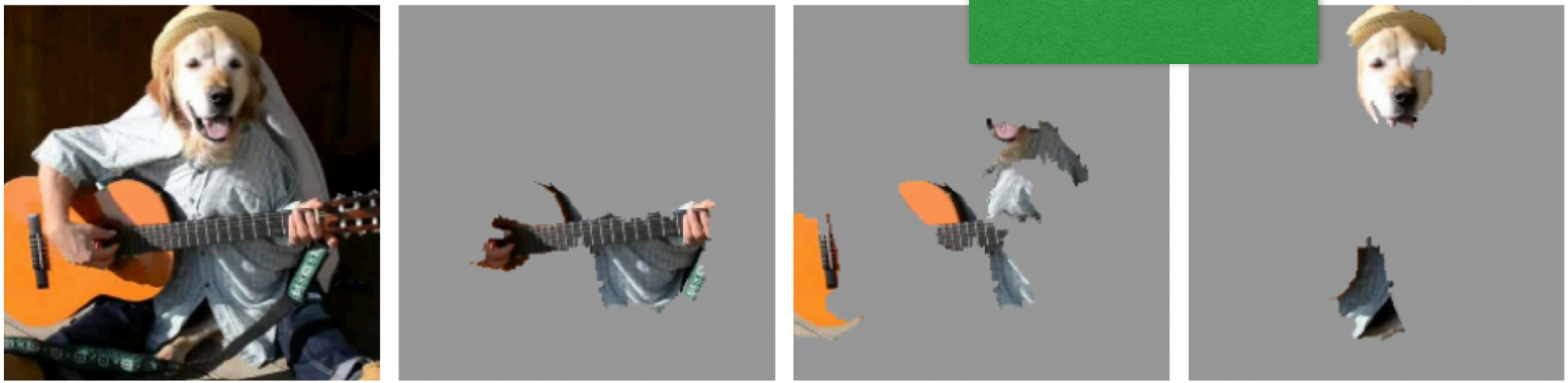


How LIME works



LIME Examples: Image Recognition

- Google Inception Network for Image Recognition
- Best classification outputs:
  - P(Electric Guitar) = 0.32 (Why? Lime? Fretboard)
  - P(Acoustic Guitar) = 0.24
  - P(Labrador) = 0.21



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

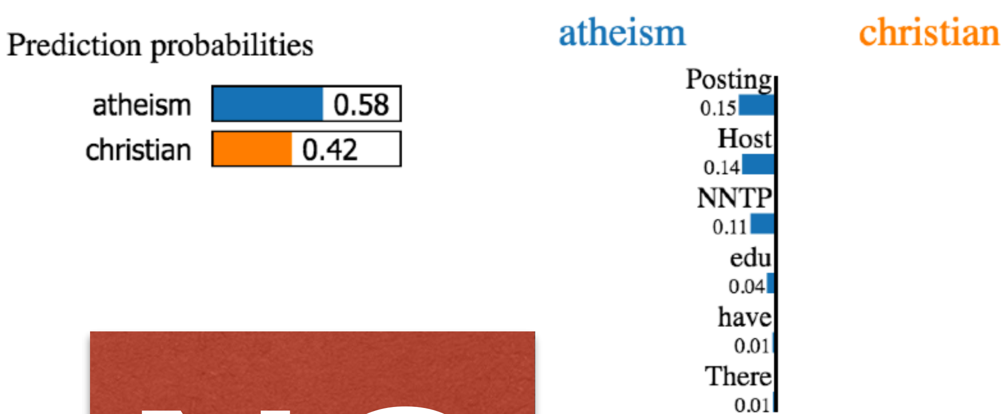
LIME Examples: Image Recognition

- Husky vs Wolf ?
- IF Snow THEN Wolf ...

(a) Husky classified as wolf (b) Explanation

LIME Examples: Text Recognition

- Newsgroups (Old) Dataset
- Christian vs. Atheist …



Text with highlighted words

From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

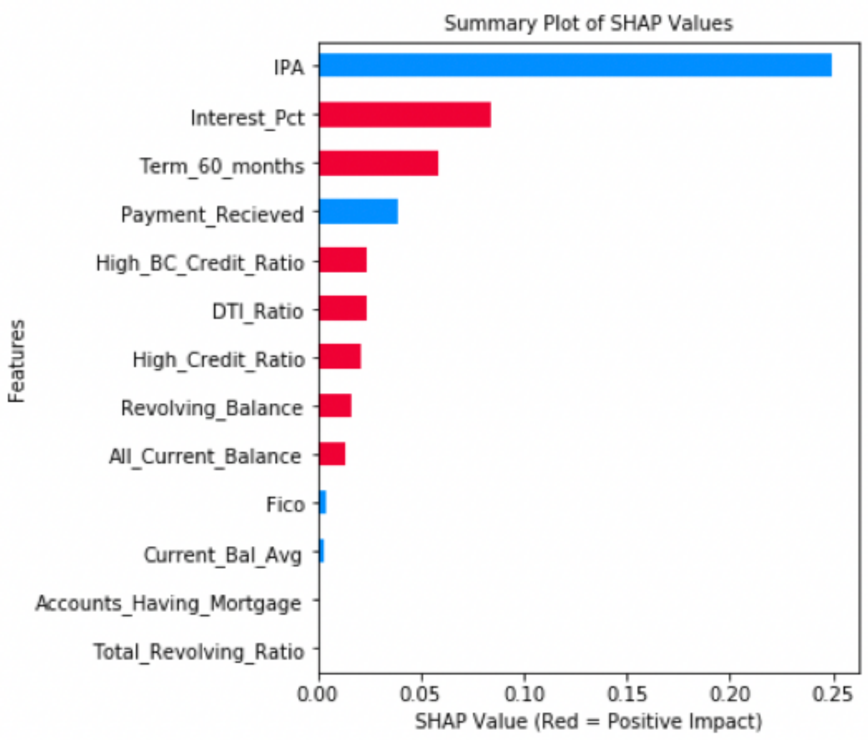
Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish.
This is the same question I have and I have not seen an answer on the
net. If anyone has a contact please post on the net or email me.

SHapley Additive exPlanations (SHAP) (Lundberg & Lee 2017)

- Another Model Agnostic Method
- Any kind of data and open source
- Shapley values? Game-theoretic concept: each actor gains as much or more as they would have from acting independently
- Let's explain the (explanation) model by a simple example ...

SUMMARY PLOT



IDSIA

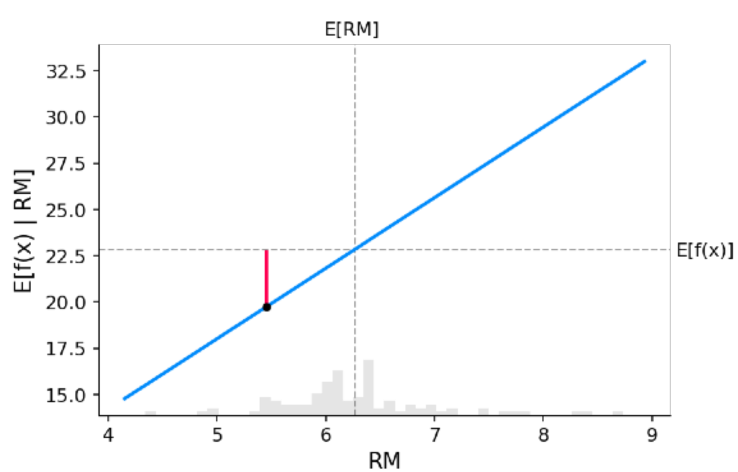
SUPSI

```
The Boston housing prices dataset has an ethical problem. You can refer to
the documentation of this function for further details.

The scikit-learn maintainers therefore strongly discourage the use of this
dataset unless the purpose of the code is to study and educate about
ethical issues in data science and machine learning.
```

Boston Hosing Data to understand SHAP

- Simple Linear Regression on the Boston Housing Data Set (506 regions, 14 features, median home price as target value)
- Coefficients are not so transparent (different scales)
- Partial dependence more informative, Shapley value is just the gap

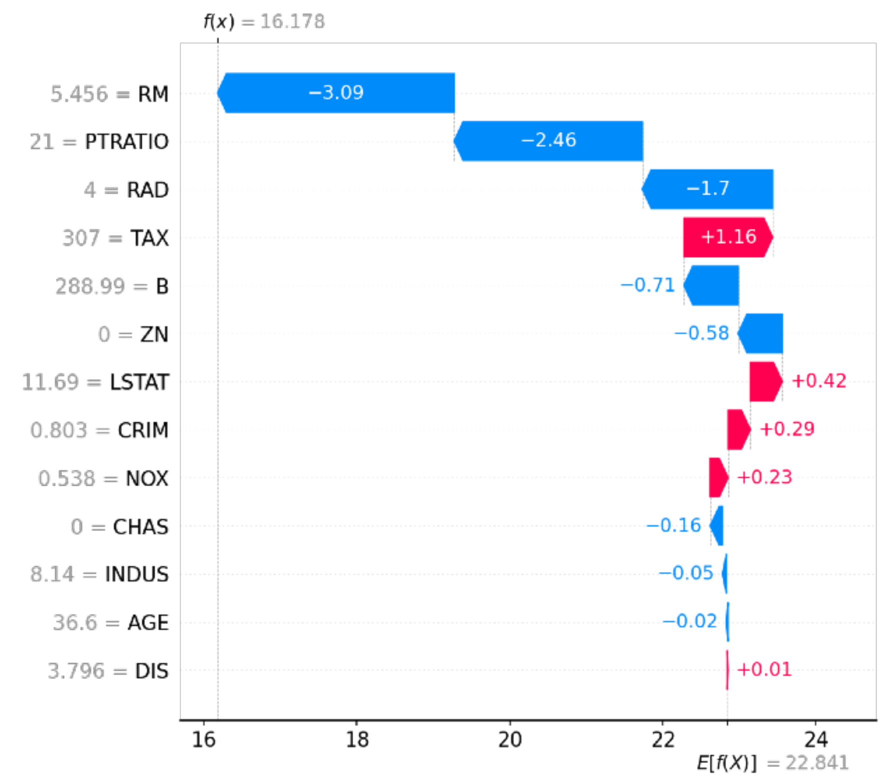
1. CRIM - per capita crime rate by town
2. ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS - proportion of non-retail business acres per town.
4. CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
5. NOX - nitric oxides concentration (parts per 10 million)
6. RM - average number of rooms per dwelling
7. AGE - proportion of owner-occupied units built prior to 1940
8. DIS - weighted distances to five Boston employment centres
9. RAD - index of accessibility to radial highways
10. TAX - full-value property-tax rate per $10,000
11. PTRATIO - pupil-teacher ratio by town
12. B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13. LSTAT - % lower status of the population
14. MEDV - Median value of owner-occupied homes in $1000's



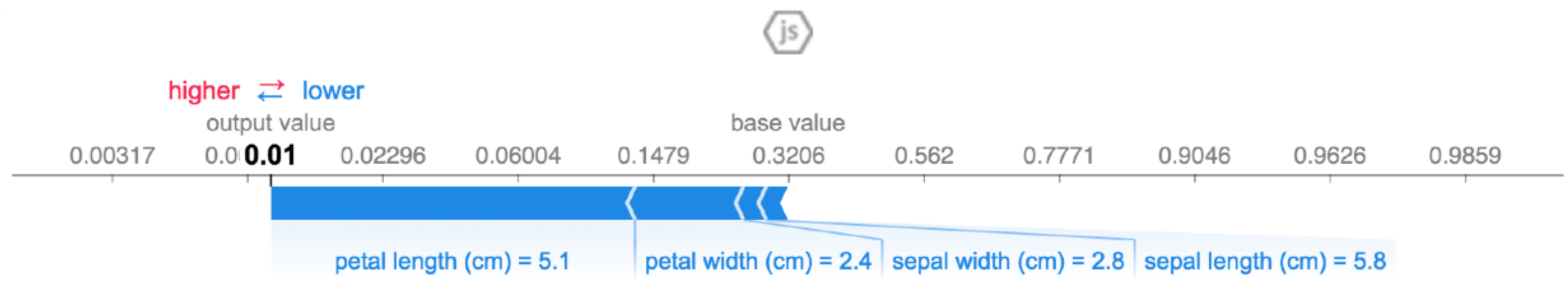

```
Model coefficients:

CRIM = -0.108
ZN = 0.0464
INDUS = 0.0206
CHAS = 2.6867
NOX = -17.7666
RM = 3.8099
AGE = 0.0007
DIS = -1.4756
RAD = 0.306
TAX = -0.0123
PTRATIO = -0.9527
B = 0.0093
LSTAT = -0.5248
```

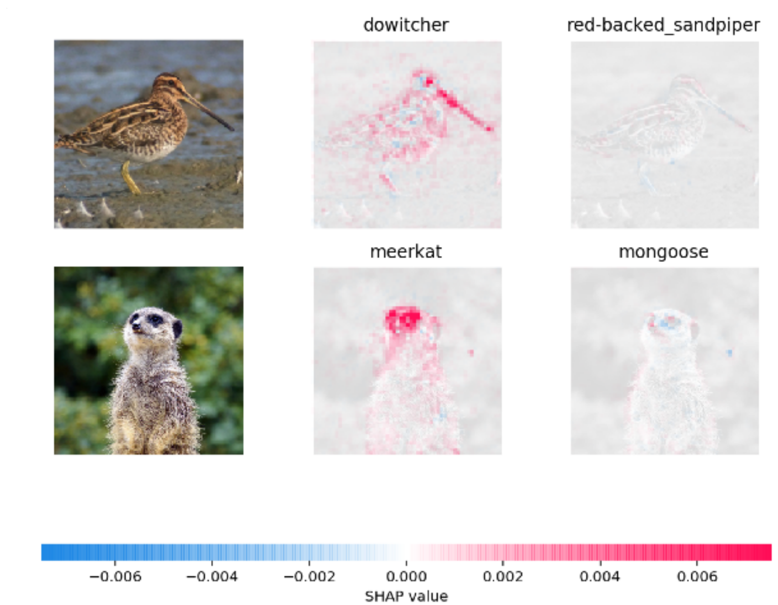
Understanding Shapley Values

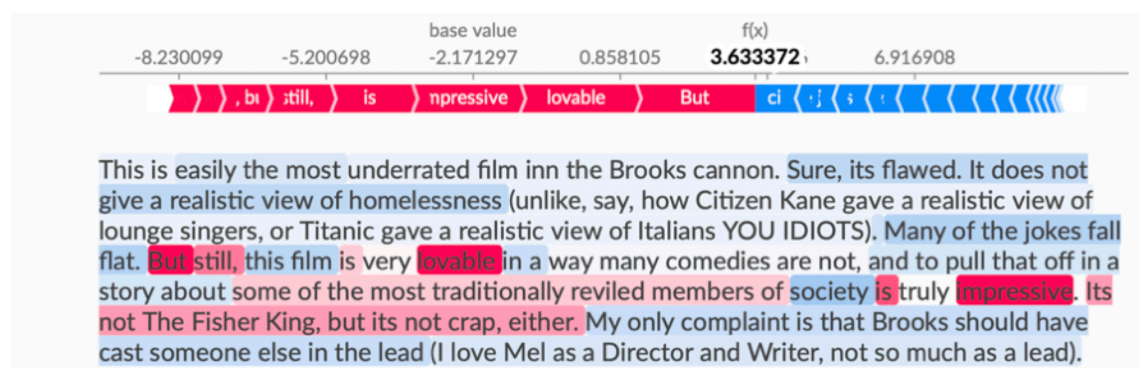
- Shapley value computed for each feature (less trivial but possible for general ML algorithms)
- Their sum is the difference between the baseline expected model output and the current model output
- This allows to explain the impact on a particular result
- Same additive property can be kept on non-linear models



Some SHAP Examples

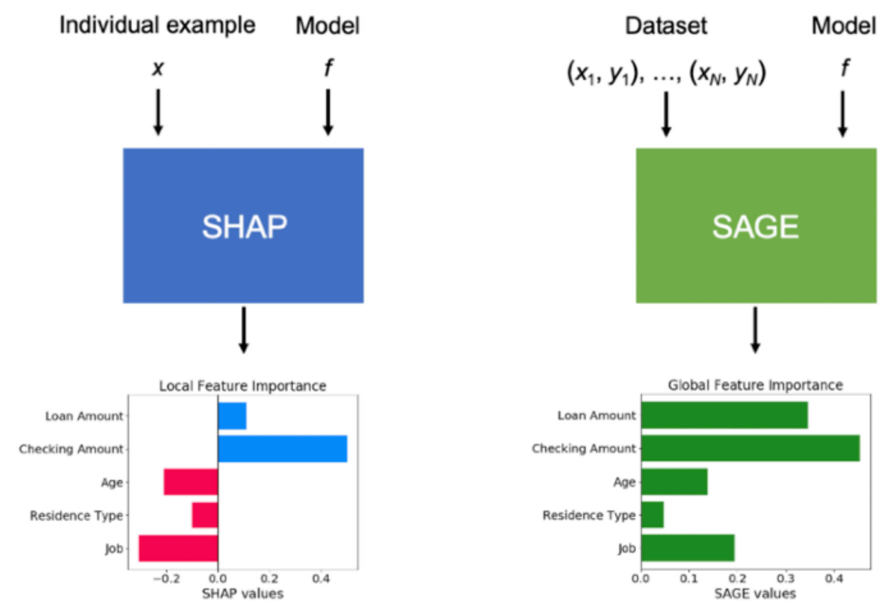






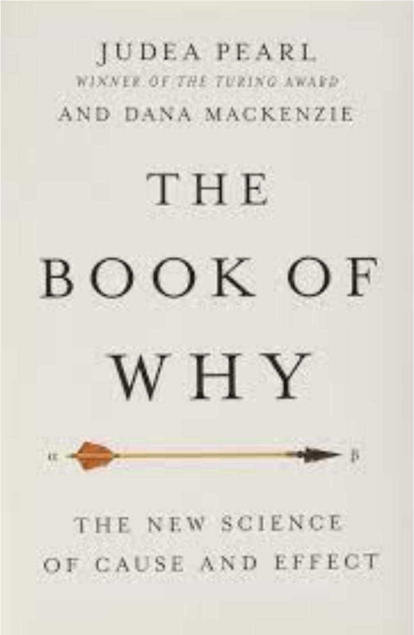
Shapley Additive Global Importance (SAGE) (2020)

- Very recent (Covert et al., 2020) extension of SHAP towards global explanations
- Still based on Shapley values
- SHAP answers the question how much does each feature contribute to this individual prediction?
- SAGE answers the question how much does the model depend on each feature overall?
- Local (SHAP) vs. Global (SAGE)



Counterfactual Explanations

- Causal analysis distinguishes between observations and interventions
  $P(X|y) \neq P(X|\mathrm{do}(y))$
- This allows for WHAT-IF reasoning: if an input datapoint was x instead of x', then a ML output would be y instead of y'
- Counterfactual Probabilities
  $P(y_x|y', x') := P(y|x', y', \mathrm{do}(x))$
- Pearl's Causal Models allow to compute CFs (in general only partially identifiable)



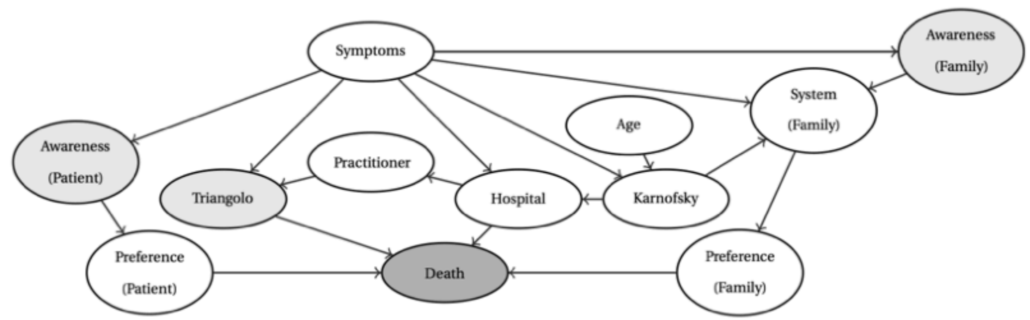
Credici
Credal Inference for Causal Inference

arXiv:2011.02912 (cs)
[Submitted on 4 Nov 2020 (v1), last revised 22 Nov 2021 (this version, v3)]
Causal Expectation-Maximisation
Marco Zaffalon, Alessandro Antonucci, Rafael Cabañas

arXiv:2008.00463 (cs)
[Submitted on 2 Aug 2020]
Structural Causal Models Are (Solvable by) Credal Networks
Marco Zaffalon, Alessandro Antonucci, Rafael Cabañas

A Counterfactual Analysis in Palliative Care

Study of terminally ill cancer patients’ preferences wrt their place of death (home or hospital)



A causal model (BN) based on expert knowledge and data

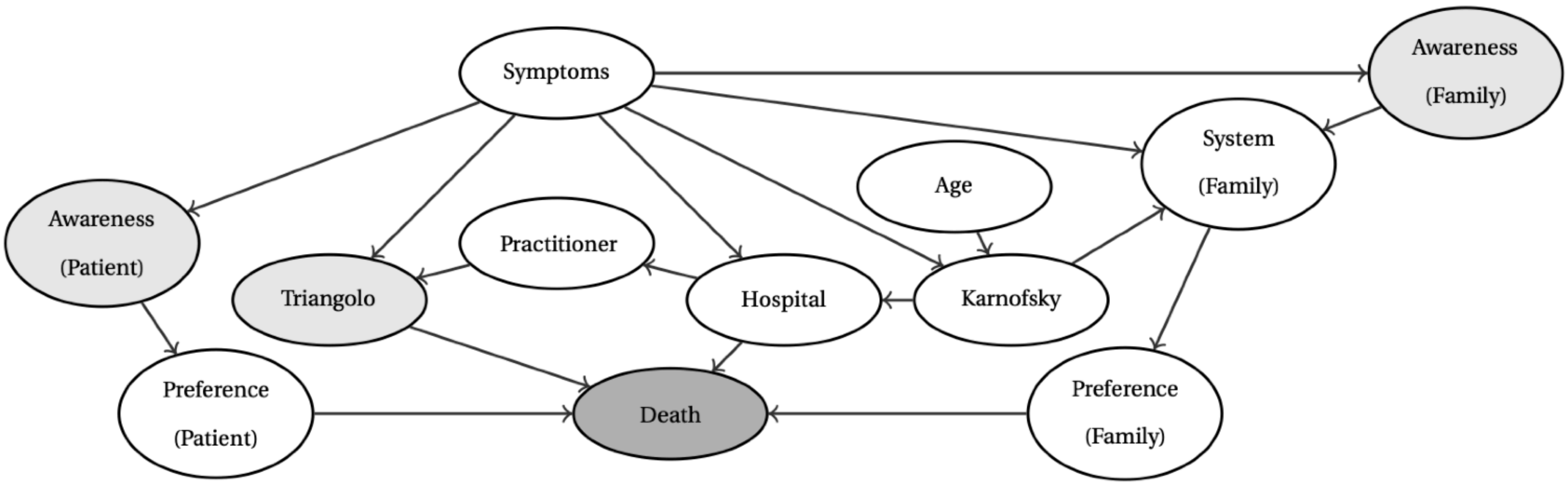
ASSOCIAZIONE TRIANGOLO
qualitépalliative marchio di qualità "in cure palliative"
servizio cure palliative e domiciliari

Impact on place of death in cancer patients: a causal exploration in southern Switzerland

Heidi Kern[1], Giorgio Corani[2], David Huber[2], Nicola Vermes[2], Marco Zaffalon[2], Marco Varini[3], Claudia Wenzel[4], André Fringer[5]

A Counterfactual Analysis in Palliative Care

most patients prefer to die at home,
but a majority actually die in institutional settings
interventions by health care professionals that can facilitate dying at home?



A Counterfactual Analysis in Palliative Care

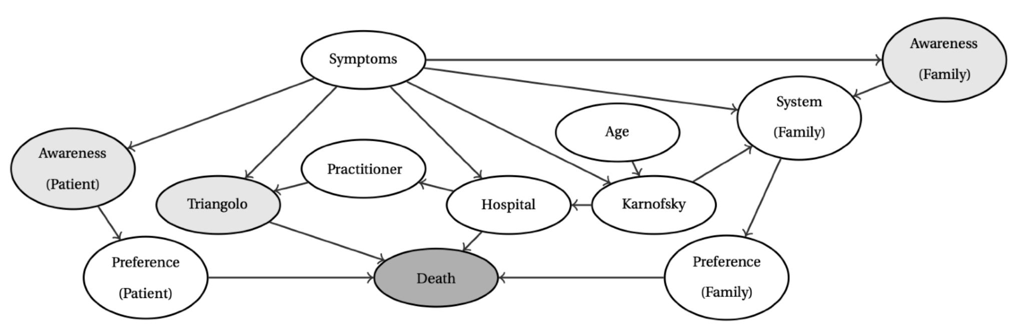
- Finding the most important variable on which to act
- Importance by probability of necessity and sufficiency

$$PNS := P(Y_{X=1} = 1, Y_{X=0} = 0)$$

$\text{PNS(Triangolo)} \in [0.30, 0.31]$

$\text{PNS(Patient\_Awareness)} \in [0.03, 0.10]$

$\text{PNS(Family\_Awareness)} \in [0.06, 0.10]$



A Cou

- Fir
  va
- Im
  ne

PNS

0,0.31]

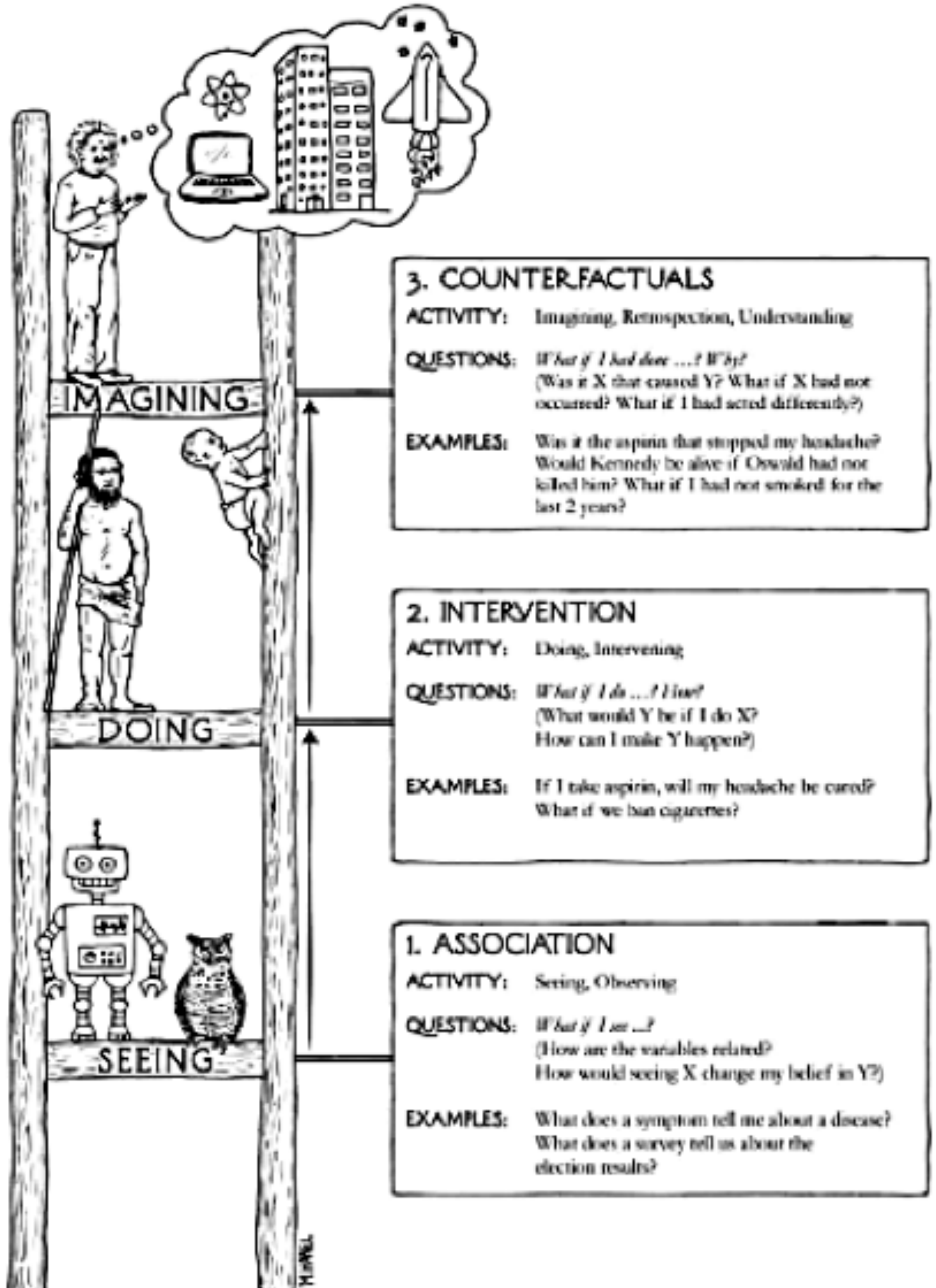
0.03,0.10]

0.06,0.10]

One should act on Triangolo first: for instance, by making Triangolo available to all patients, we should expect a reduction of people at the hospital by 30%

This would save money too, and would allow politicians to do economic considerations as to which amount it is even economically profitable to fund Triangolo, and have patients die at home, rather than spending more to have patients die at the hospital

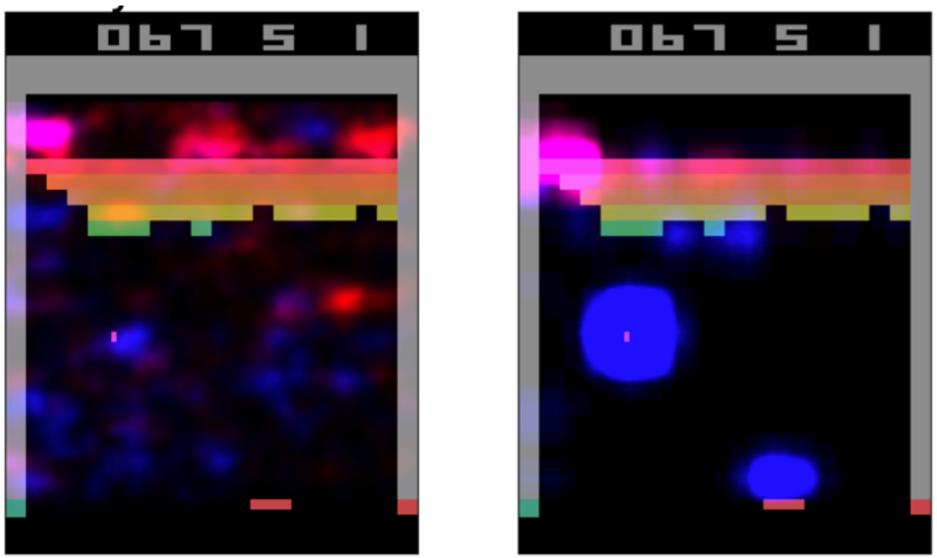
Pearl's Ladder of Causation



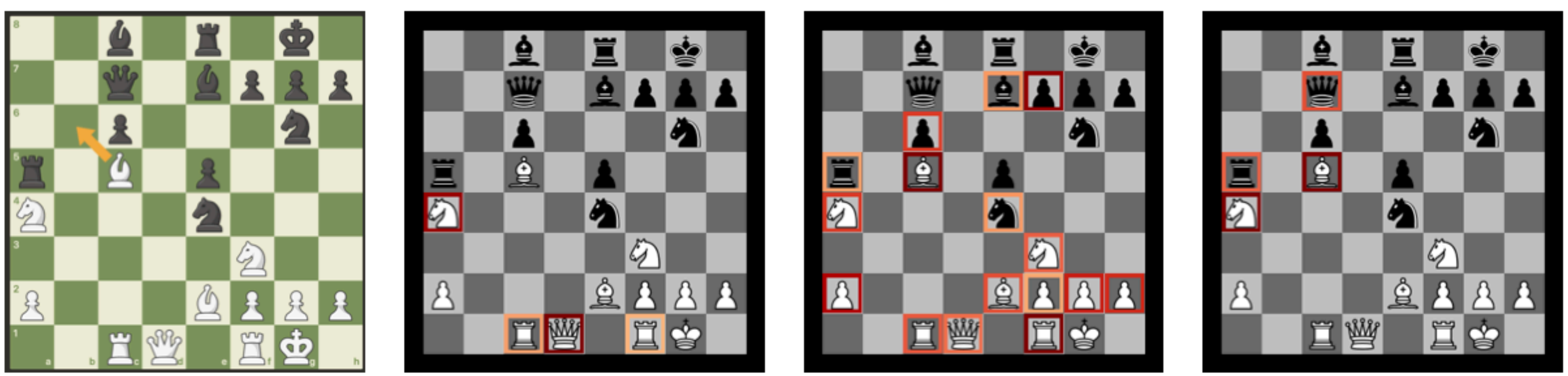
Explaining Reinforcement Learning

- Agent operating in state space $\mathcal{S}$
- Set of actions $\mathcal{A}_s$
- Q-value function $Q(s, a)$ available for each $s \in \mathcal{S}$ and $a \in \mathcal{A}_s$
- Greedy agent $\hat{a} = \arg\max_a Q(a, s)$
- For each feature f compute its saliency $S[f]$
- s' perturbation of s obtained by changing the value of f
- $S[f]$ corresponds to the Q-value change
- E.g., Iyer (2018): $S[f] = Q(s, \hat{a}) - Q(s', \hat{a})$
- Alternatives have been proposed

Explainable Reinforcement Learning by Saliency Maps



- Saliency maps can be created by means of the computed saliency levels



(a) Original Position (b) Iyer et al. (2018) (c) Greydanus et al. (2018) (d) SARFA

Strategic Training by XRL

(a) Chess board

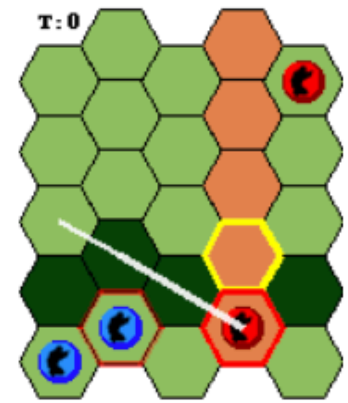
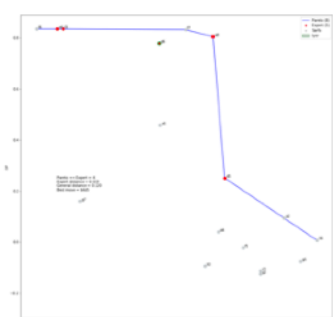
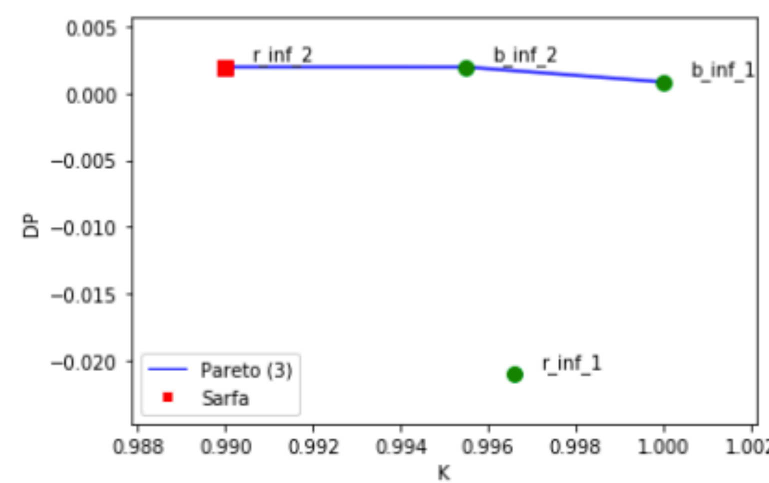


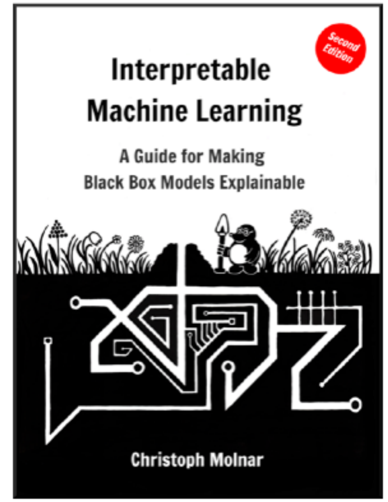
Figure 5.6: Saliency map for NTW



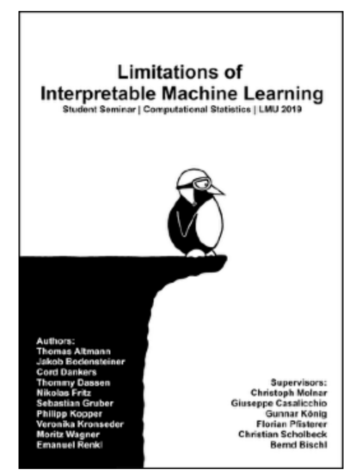


Concluding Remarks

- Interpretable Machine (and Reinforcement) Learning is an important open challenge for contemporary AI
- Tools/libraries are already available
- Explaining your ML model does not require to trade-off your accuracy
- Causal analysis as the ultimate



christophm.github.io/interpretable-ml-book/



slds-lmu.github.io/iml_methods_limitations/

Python Notebook with Some Examples

https://colab.research.google.com/drive/1Oyk7W94fjb_b_W3A8JjKEdduFIETJh5j

Questions?

alessandro@idsia.ch