MEDNARODNA
PODIPLOMSKA ŠOLA
JOŽEFA STEFANA

INFORMATION AND COMMUNICATION TECHNOLOGIES
PhD study programme

Data Mining and Knowledge Discovery

Petra Kralj Novak

January 22, 2020

http://kt.ijs.si/petra_kralj/dmkd3.html

So far …

- Nov. 11, 2019
  - Basic classification
  - Orange hands on data visualization and classification
- Dec. 11, 2019
  - Fitting and overfitting
  - Data leakage
  - Decision boundary
  - Evaluation methods
  - Classification evaluation metrics: confusion matrix, TP, FP, TN, FN, accuracy, precision, recall, F1, ROC
  - Imbalanced data and unequal misclassification costs
  - Probabilistic classification
  - Naïve Bayes classifier

So far …

- Dec. 18 2019
  - Naive Bayes classifier
  - Laplace estimate
  - Regression (numeric prediction) and its evaluation
- Jan. 13, 2020
  - Association rules
- Jan. 15, 2020
  - Neural networks
- Jan. 21, 2020
  - Clustering: K-means, Hierarchical, DBSCAN

Artificial intelligence

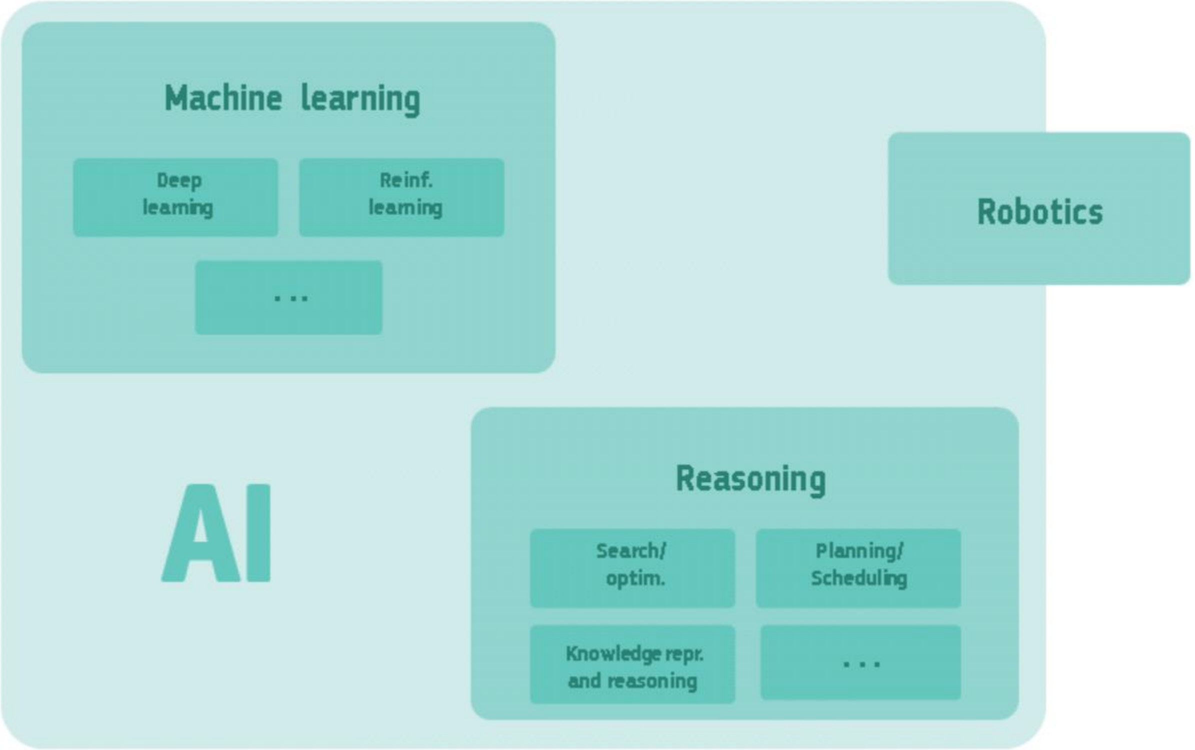
- Artificial intelligence (AI) refers to systems that display intelligent behavior by analyzing their environment and taking actions – with some degree of autonomy – to achieve specific goals.
- AI-based systems can be **purely software-based**, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be **embedded in hardware devices** (e.g. advanced robots, autonomous cars, drones or Internet of Things applications).

High-Level Expert Group on Artificial Intelligence: A definition of AI: main capabilities and scientific disciplines
https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60651

A simplified overview of AI’s sub-disciplines

Both machine learning and reasoning include many other techniques, and robotics includes techniques that are outside AI.

The whole of AI falls within the computer science discipline.



High-Level Expert Group on Artificial Intelligence: A definition of AI: main capabilities and scientific disciplines
https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60651

AI as a scientific discipline

- As a scientific discipline, AI includes several approaches and techniques, such as
  - machine learning (of which deep learning and reinforcement learning are specific examples),
  - machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization),
  - and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems).

High-Level Expert Group on Artificial Intelligence: A definition of AI: main capabilities and scientific disciplines
https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60651

AI system

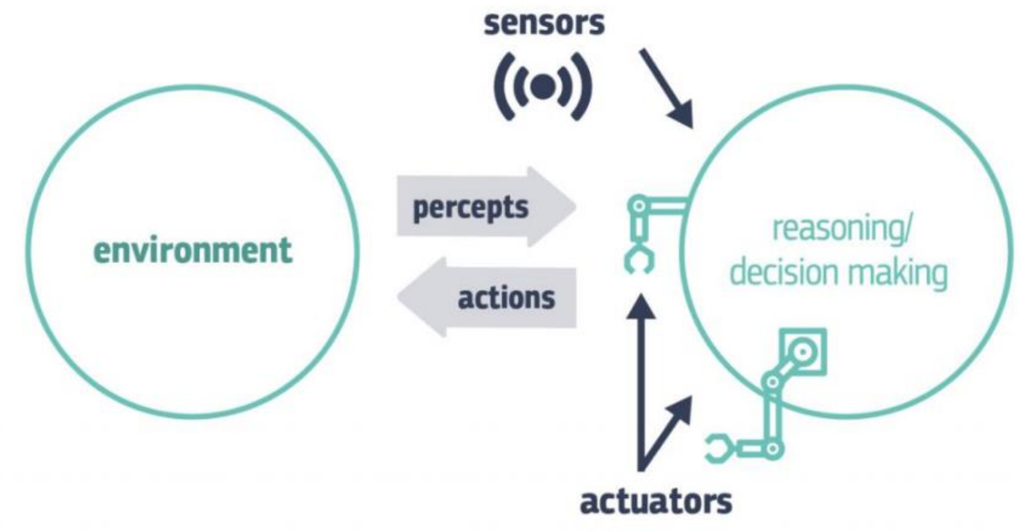


Figure 1: Schematic depiction of an AI system.

- AI system: any AI-based component, software and/or hardware.
- AI systems are (usually) embedded as components of larger systems, rather than standalone systems.

High-Level Expert Group on Artificial Intelligence: A definition of AI: main capabilities and scientific disciplines
https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60651

AI system

- “Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by
  - perceiving their environment through data acquisition,
  - interpreting the collected structured or unstructured data,
  - reasoning on the knowledge, or processing the information, derived from this data
  - and deciding the best action(s) to take to achieve the given goal.
- AI systems can either
  - use symbolic rules or
  - learn a numeric model, and
  - they can also adapt their behavior by analyzing how the environment is affected by their previous actions.

High-Level Expert Group on Artificial Intelligence: A definition of AI: main capabilities and scientific disciplines
https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60651

Trustworthy AI

- According to the *High-Level Expert Group on AI's* **Ethics Guidelines for Trustworthy Artificial Intelligence**, trustworthy AI should be:
- (1) lawful - respecting all applicable laws and regulations
- (2) ethical - respecting ethical principles and values
- (3) robust - both from a technical perspective while taking into account its social environment

High-Level Expert Group on Artificial Intelligence: Ethics guidelines for trustworthy AI." *B-1049 Brussels* (2019). https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

An example of automated discrimination

- A computer program for screening job applicants was used during the 1970s and 1980s in St. George’s Hospital Medical School, London, UK.
- The program used information from applicants’ forms, without any reference to ethnicity.
- However, the program was found to **unfairly discriminate against ethnic minorities and women** by inferring this information from surnames and place of birth, and lowering their chances of being selected for interview.

Stella Lowry and Gordon Macpherson. 1988. A blot on the profession. Brit. Med. J. Clin. Res. 296, 6623 (1988), 657

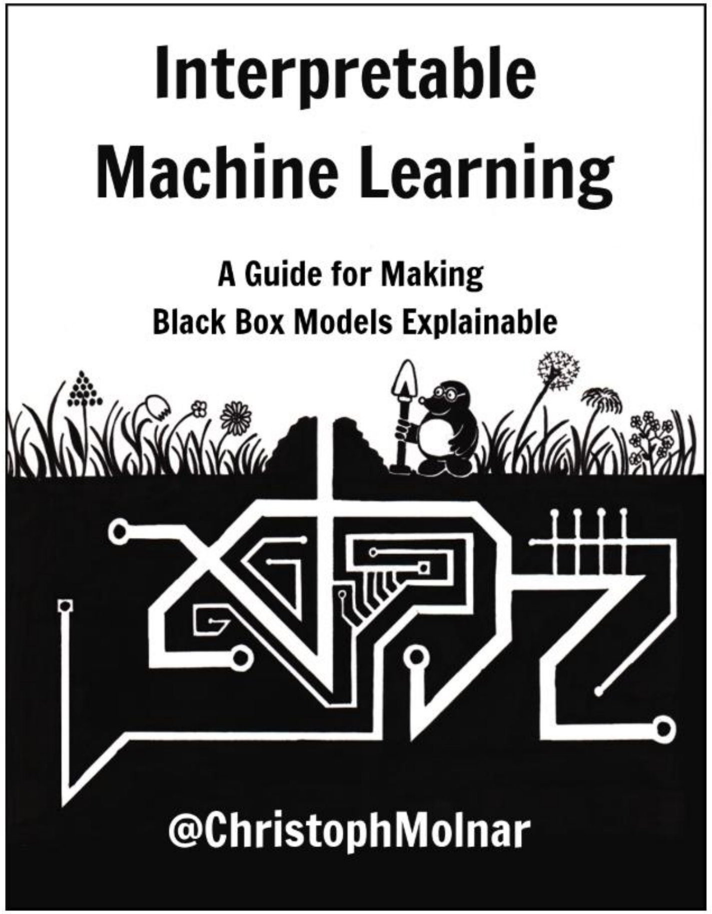
Guidance on Trustworthy AI

Ensure that the development, deployment and use of AI systems meets the seven key requirements for Trustworthy AI:

1. human agency and oversight,
2. technical robustness and safety,
3. privacy and data governance,
4. **transparency**,
5. diversity, non-discrimination and fairness,
6. environmental and societal well-being and
7. accountability.

Consider technical and non-technical methods to ensure the implementation of those requirements.

Transparency of ML



Interpretable Machine Learning

A Guide for Making Black Box Models Explainable

@ChristophMolnar

Molnar, Christoph. **Interpretable machine learning.** Lulu. com, 2019. https://christophm.github.io/interpretable-ml-book/

Black box

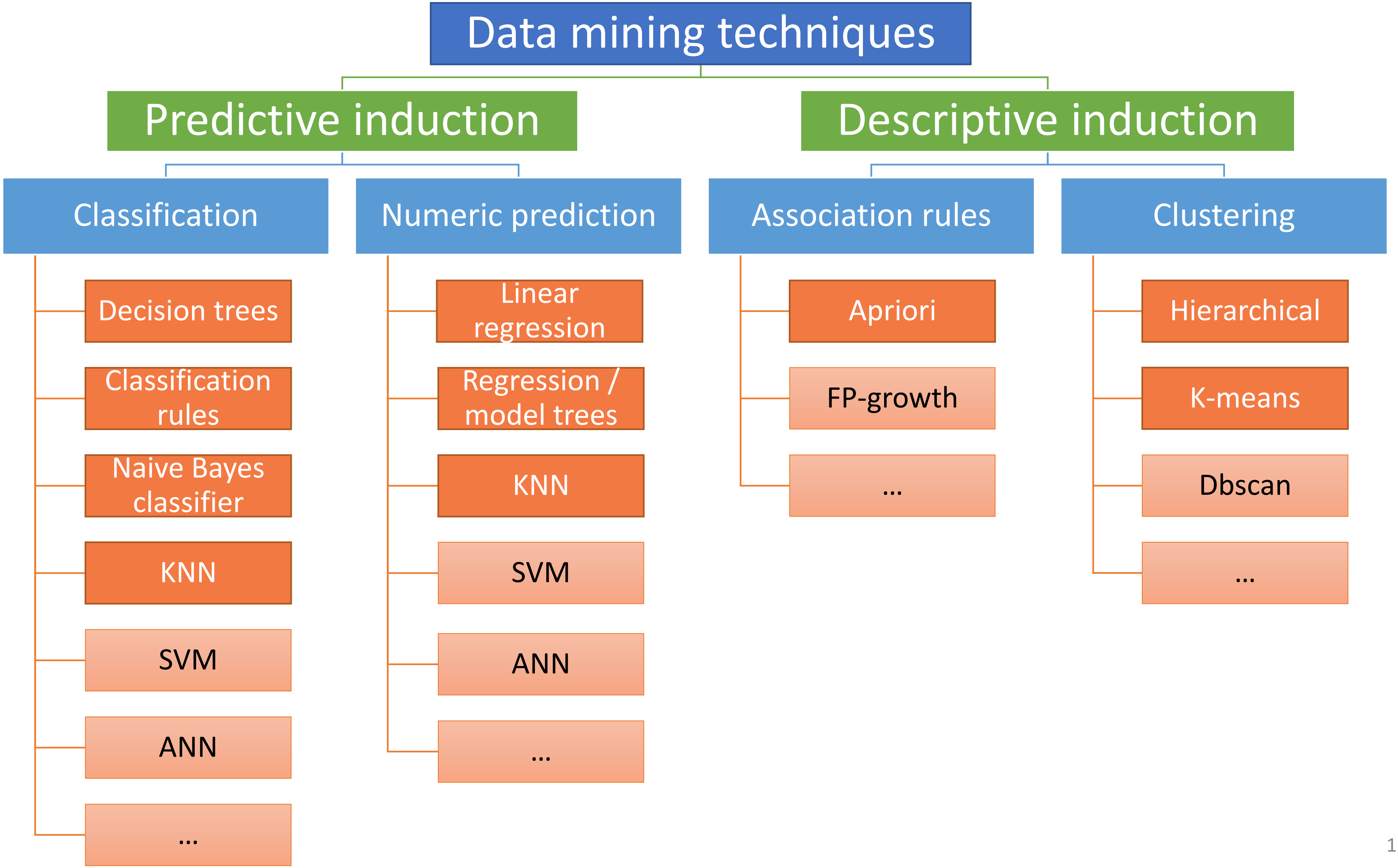
- A black box system can be viewed in terms of its inputs and outputs, **without any knowledge of its internal workings**.

- The opposite of a black box is a system where the inner components or logic are available for inspection, which is most commonly referred to as a **white box** (which can also come be called a "clear box" or a "glass box").

Interpreting black box systems

- What does it mean that a model is interpretable or transparent?
- What is an explanation?
- When is a model or an explanation comprehensible?
- Which is the best way to provide an explanation?
- Which are the problems requiring interpretable models/predictions?

- How much are we willing to lose in prediction accuracy to gain any form of interpretability?



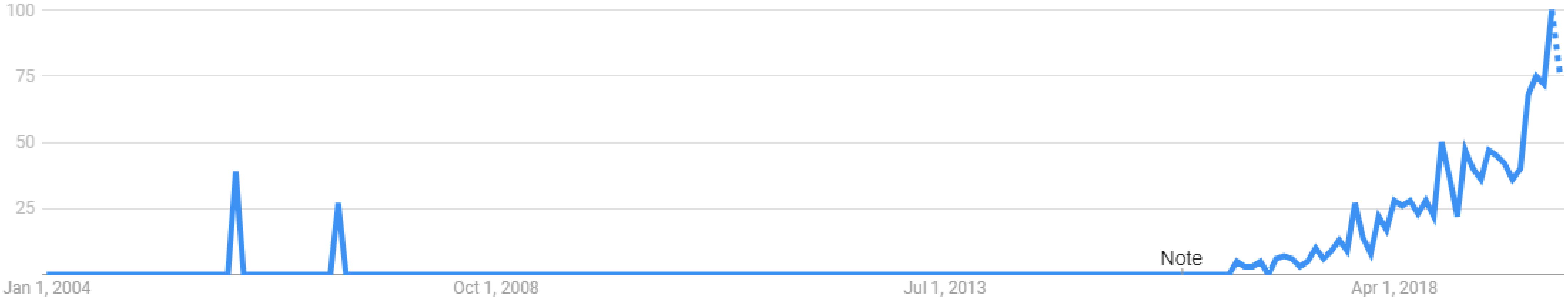
Data mining techniques

- Predictive induction
  - Classification
    - Decision trees
    - Classification rules
    - Naive Bayes classifier
    - KNN
    - SVM
    - ANN
    - …
  - Numeric prediction
    - Linear regression
    - Regression / model trees
    - KNN
    - SVM
    - ANN
    - …
- Descriptive induction
  - Association rules
    - Apriori
    - FP-growth
    - …
  - Clustering
    - Hierarchical
    - K-means
    - Dbscan
    - …

XAI = Explainable AI

- XAI tends to refer to the movement, initiatives, and efforts made in response to AI transparency and trust concerns, more than to a formal technical concept

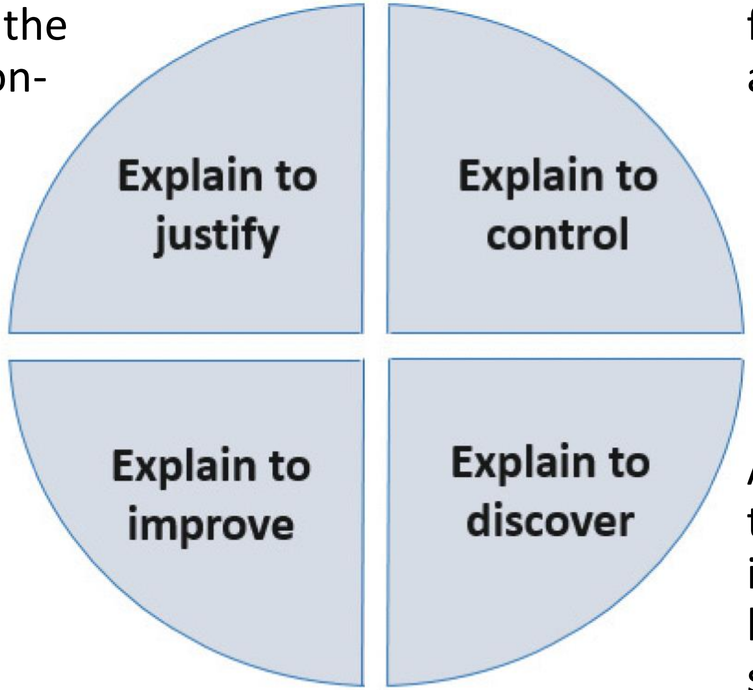
Google trends for “Explainable AI”



XAI: Need and Application Opportunities

Explanation for a decision: the need for reasons or **justifications for** that particular **outcome**, rather than a description of the inner workings or the logic of reasoning behind the decision-making process in general.

Understanding more about system behavior provides greater visibility over unknown vulnerabilities and flaws, and helps to rapidly identify and correct errors.



A model that can be explained and understood is one that can be more easily improved.

Asking for explanations is a helpful tool to learn new facts, to gather information and thus to gain knowledge. Only explainable systems can be useful for that.

Adadi, A., & Berrada, M. (2018). **Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI)**. IEEE Access, 6, 52138-52160.

XAI methods

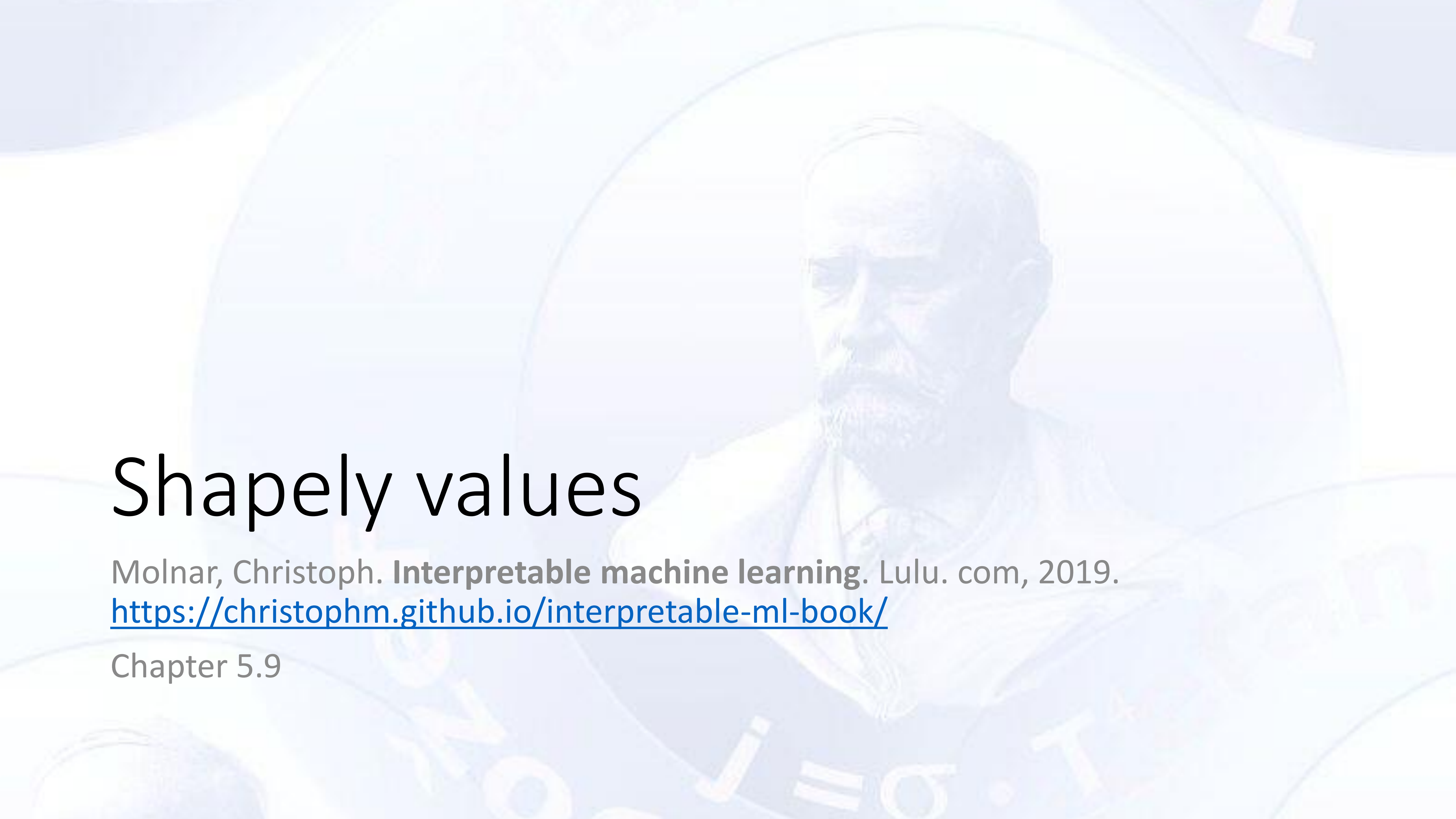
- **Intrinsic or post hoc?**

- **Model-specific or model-agnostic?**
  - Model-specific interpretation tools are limited to specific model classes.
  - Agnostic methods usually work by analyzing feature input and output pairs.

- **Local or global?**
  - Does the interpretation method explain an individual prediction or the entire model behavior?

Result of the interpretation method

- **Feature summary statistic**: feature importance, the pairwise feature interaction strengths…
- **Feature summary visualization**: Partial dependence plots are curves that show a feature and the average predicted outcome.
- **Model internals:** learned weights, tree, rules, …
- **Data point**: counterfactual explanations,...
- **Intrinsically interpretable model**: approximate black box models (either globally or locally) with an interpretable model.



Shapely values

Molnar, Christoph. **Interpretable machine learning**. Lulu. com, 2019. https://christophm.github.io/interpretable-ml-book/

Chapter 5.9

Shapely values

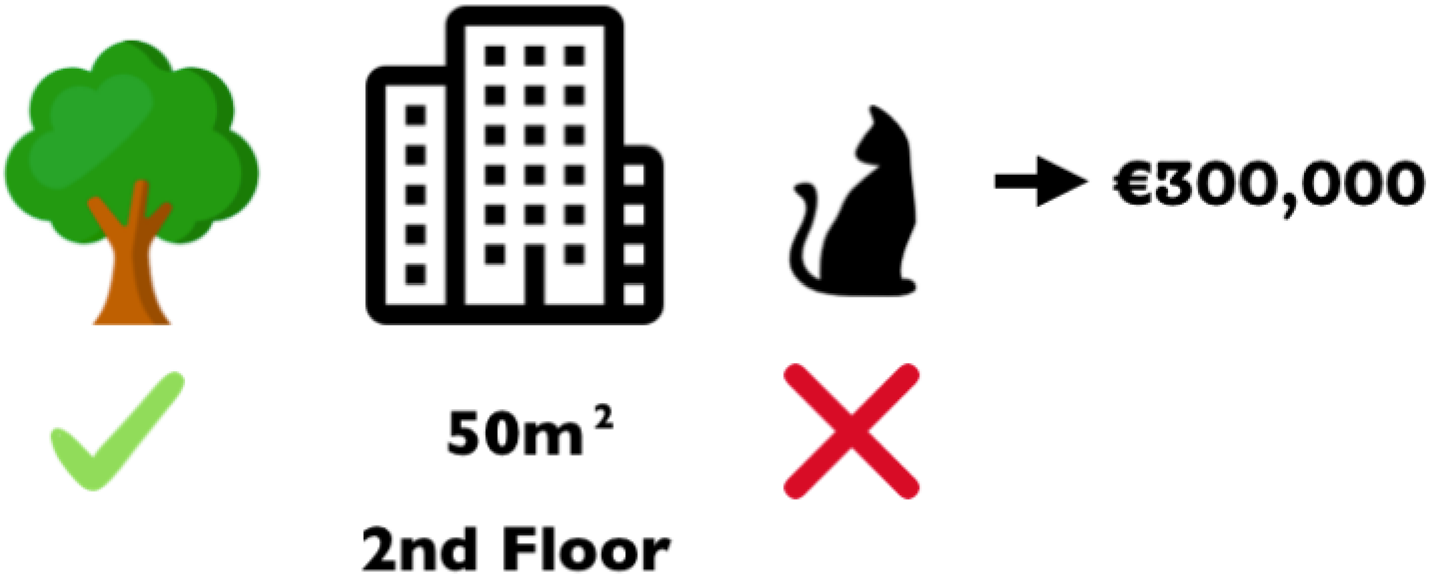
instance

Given the current set of feature values, the contribution of a feature value to the difference between the actual prediction and the mean prediction.

- Based on a method from coalitional game theory:

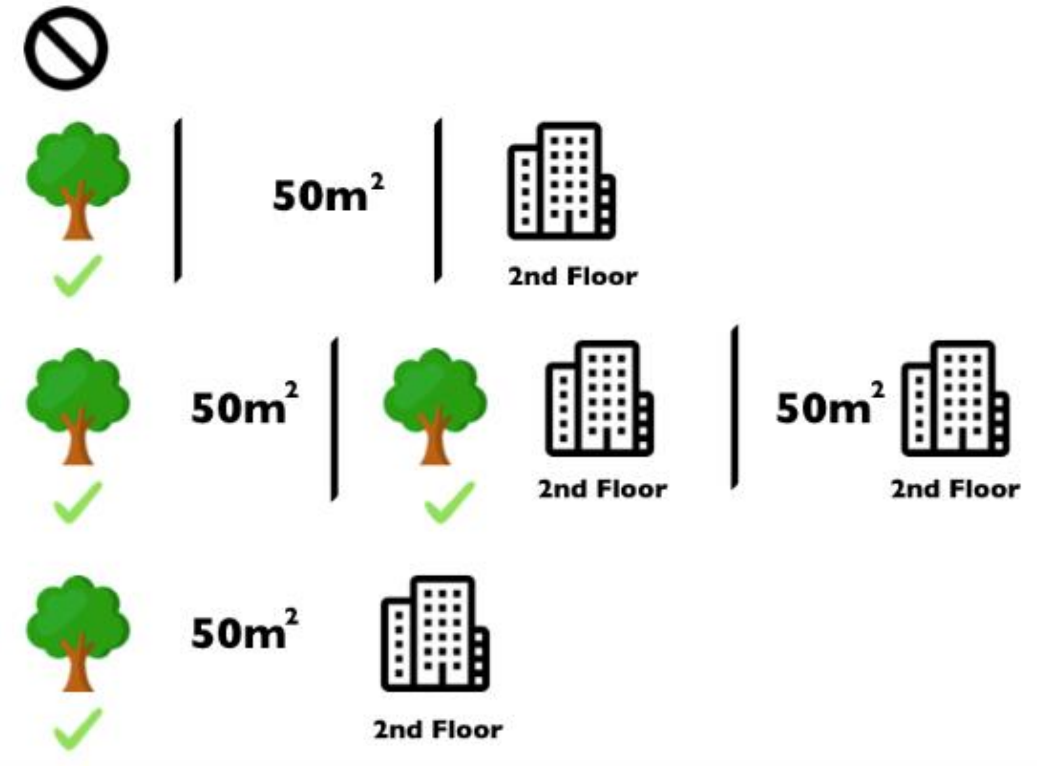
A group of differently skilled participants are all cooperating together for a collective reward. How should the reward be ***fairly*** *divided amongst the group?*

Shapely explained



The average prediction for all apartments is €310,000. How much has the feature value “cats not allowed” contributed to the prediction (compared to the average prediction)?

Computing Shapely: Coalitions



- 8 coalitions composed of the features in the instance "cats not allowed"
- For each coalition, compute the difference between
- Model output of the coalition and model output of the coalition with the added feature value "cats not allowed"

Shapely values

The Shapley value is the average marginal contribution of a feature value across all possible coalitions.

$$\phi_j(val) = \sum_{S \subseteq \{x_1,\ldots,x_p\} \setminus \{x_j\}} \frac{|S|!\,(p-|S|-1)!}{p!} \left(val\left(S \cup \{x_j\}\right) - val(S)\right)$$

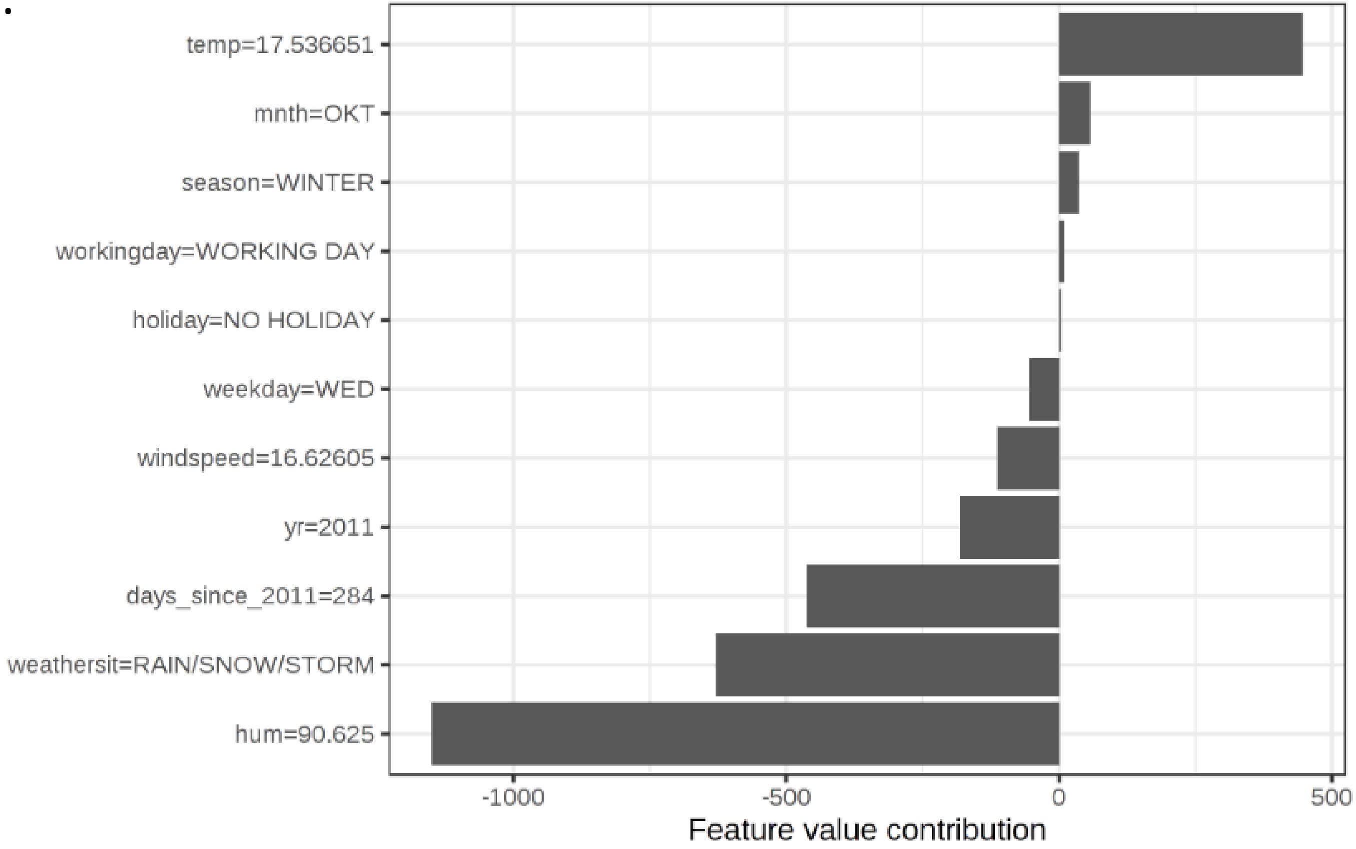
	Coalitional game	Machine learning
{x1,….,xp}	set of all players	Set of all attribute-value pairs = instance
i	The i-th player	The i-th feature-value pair
val	the function *val* gives the value (or payout) for any subset of those players	Predictive ML model

Shapely example

The interpretation of the Shapley value for feature value j is:

The value of the j-th feature contributed φj to the prediction of this particular instance compared to the average prediction for the dataset.

bike rental dataset



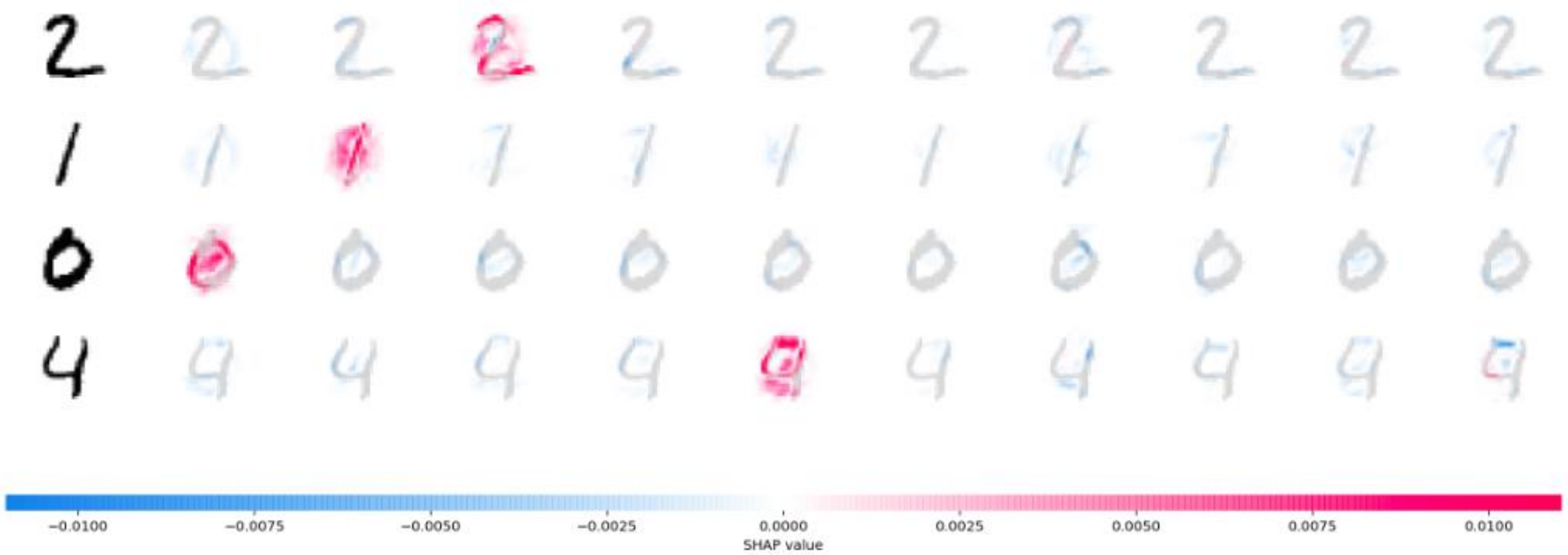
Shapley value summary

- The Shapley value is the average contribution of a feature value to the prediction in different coalitions.
- The Shapley value is NOT the difference in prediction when we would remove the feature from the model.
- The Shapley value works for both classification (if we are dealing with probabilities) and regression.
- **A lot of computing time**. In 99.9% of real-world problems, only the approximate solution is feasible.

____

- SHAP (SHapley Additive exPlanations) has a fast implementation for tree-based models (random forest,…)
- shap Python package

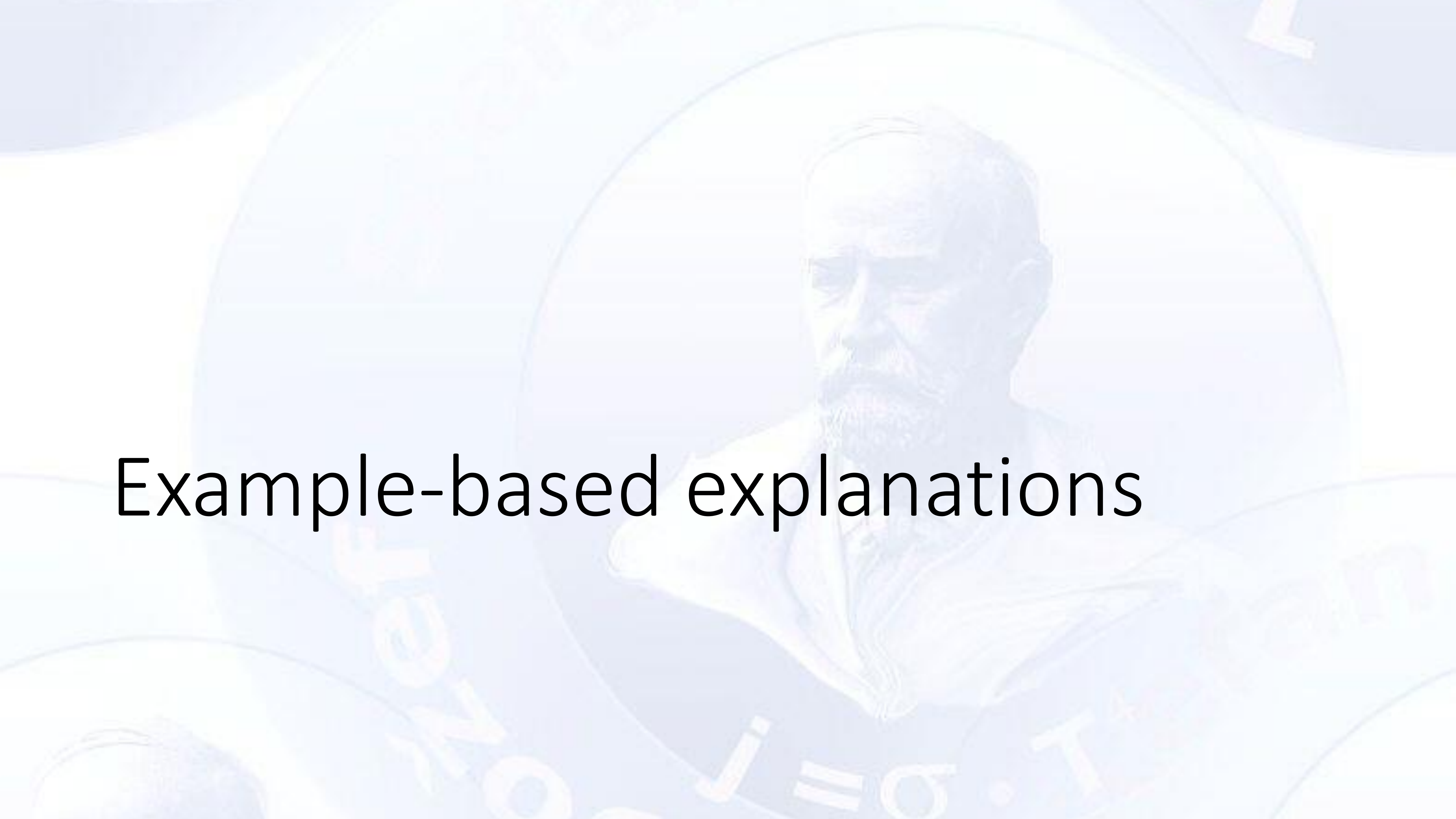
SHAP result example



The plot above explains ten outputs (digits 0-9) for four different images. Red pixels increase the model's output while blue pixels decrease the output. The input images are shown on the left, and as nearly transparent grayscale backings behind each of the explanations. The sum of the SHAP values equals the difference between the expected model output (averaged over the background dataset) and the current model output. Note that for the 'zero' image the blank middle is important, while for the 'four' image the lack of a connection on top makes it a four instead of a nine.

https://github.com/slundberg/shap

Example-based explanations



Counterfactual Explanations

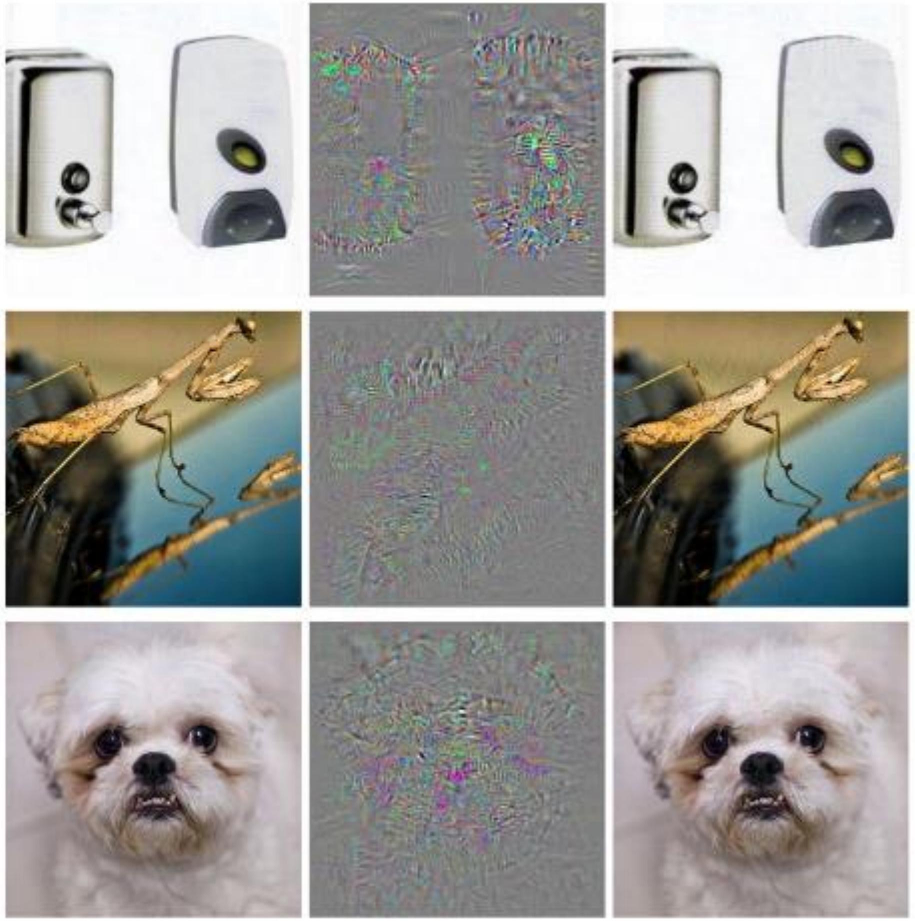
- A counterfactual explanation of a prediction describes **the smallest change to the feature values that changes the prediction** to a predefined output.

- “If X had not occurred, Y would not have occurred”

- Model-agnostic (it only works with the model inputs and output)
- Explain predictions of individual instances

Adversarial Examples

An adversarial example is an instance with small, intentional feature perturbations that cause a machine learning model to make a false prediction.



Adversarial examples for AlexNet by Szegedy et. al (2013). All images in the left column are correctly classified. The middle column shows the (magnified) error added to the images to produce the images in the right column all categorized (incorrectly) as ‘Ostrich’.

Literature

- Molnar, Christoph. **Interpretable machine learning**. Lulu. com, 2019. https://christophm.github.io/interpretable-ml-book/
- Adadi, A., & Berrada, M. (2018). **Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI)**. IEEE Access, 6, 52138-52160.
- 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.
- High-Level Expert Group on Artificial Intelligence: **A definition of AI: main capabilities and scientific disciplines** https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60651
- High-Level Expert Group on Artificial Intelligence: **Ethics guidelines for trustworthy AI**." B-1049 Brussels (2019). https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Exam & seminar

- Instructions on the web page
- http://kt.ijs.si/petra_kralj/dmkd3.html