Università degli Studi di Padova

Machine Learning
2024/2025

AMCO
Artificial Intelligence, Machine Learning and Control Research Group

Lecture #33 eXplainable Artificial Intelligence (XAI)

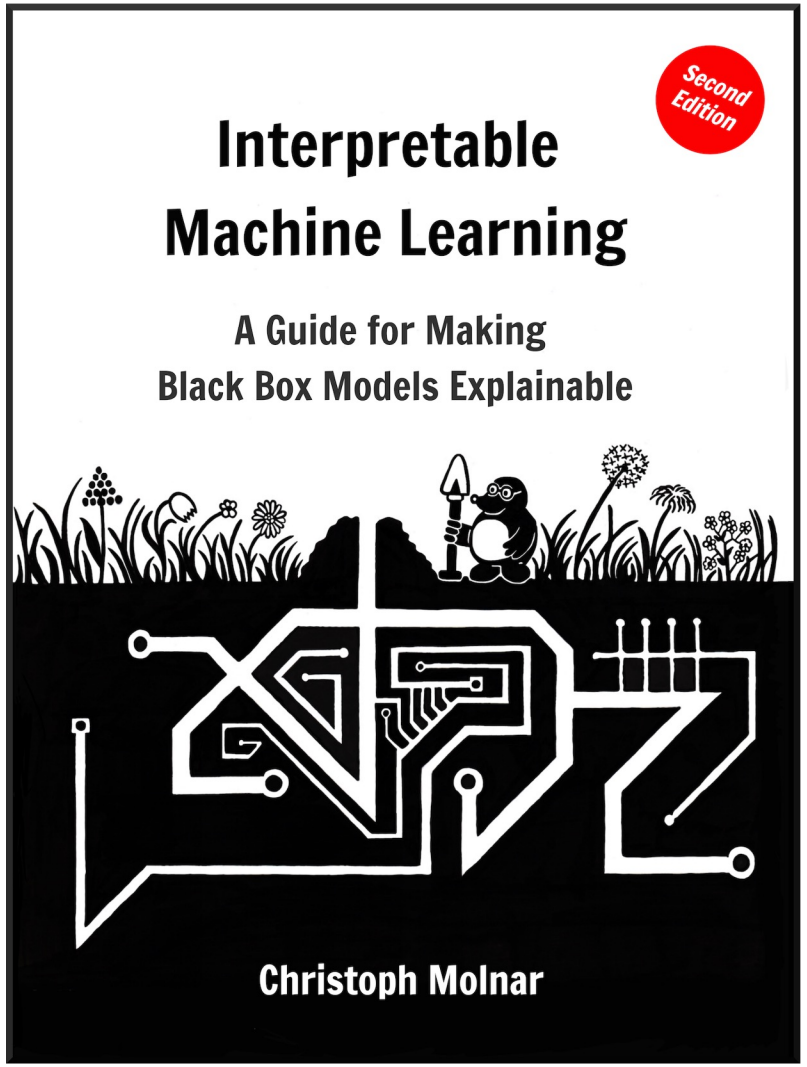
Gian Antonio Susto

Main Reference

‘Interpretable Machine Learning. A Guide for Making Black Box Models Explainable’

by Christoph Molnar

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



Before explainability, let's talk about Machine Learning (ML)

DEVELOPERS
- ML used to develop new technologies

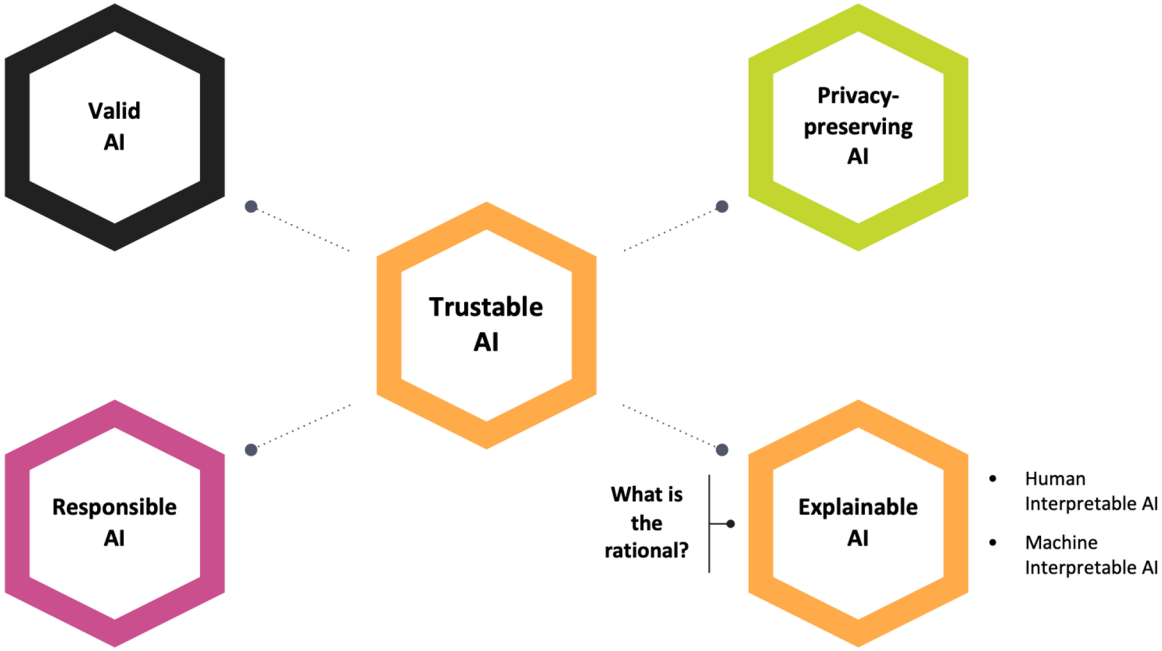
USERS (for example in the medical domain)
- ML used for handling multiple information collected during experiments and for extracting new knowledge

It doesn't matter in which team you are in, if you are dealing with Machine Learning then you probably should be interested in explainability

Let’s start with:

- Introduction to interpretability
  - Motivation
  - Examples
  - Discussion

- Key concepts
  - Definitions
  - Desiderata of ML models and relationship with Interpretability



F. Lecue et al. On Explainable AI: From Theory to Motivation, Industrial Applications and Coding Practices, AAAI2023
https://xaitutorial2023.github.io/

Disclaimer: the terms ‘ML Interpretability’ and ‘ML Explainability’ will be used as synonyms in the first part of the course, however some researchers will tell you differently… for the meantime let’s assume they both refer to the same concept

Introduction

QUESTION

Which word would you is closely connected to the problem of explainability?

Introduction

QUESTION

Which word would yo oblem of explainability?

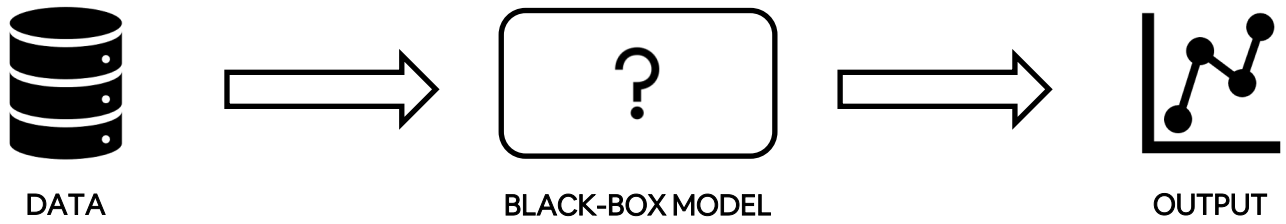
Introduction

QUESTION

Guess who's not affiliated with the "WHY Fan Club"?

Machine Learning "black-box" models!

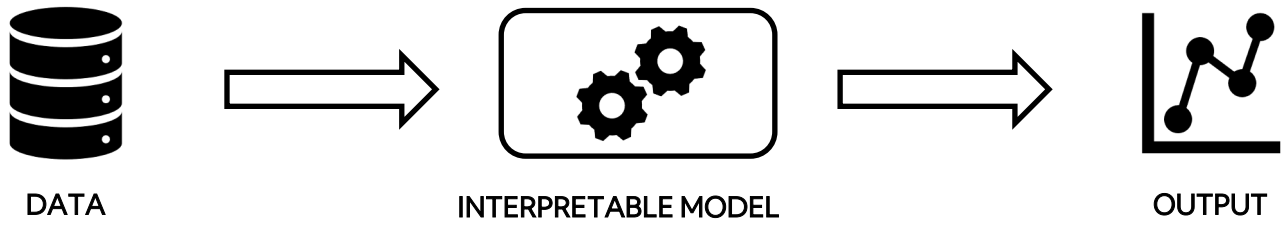
Introduction



The model provide us outputs, but we have no clue about its inner workings (think for example about the highly multivariate/non-linear mechanisms of some ML approaches)

The logic governing the model's behavior is not understandable by humans

Introduction



The elaborations performed by the model in order to generate the output aim at being simple enough to be understood by humans

Introduction

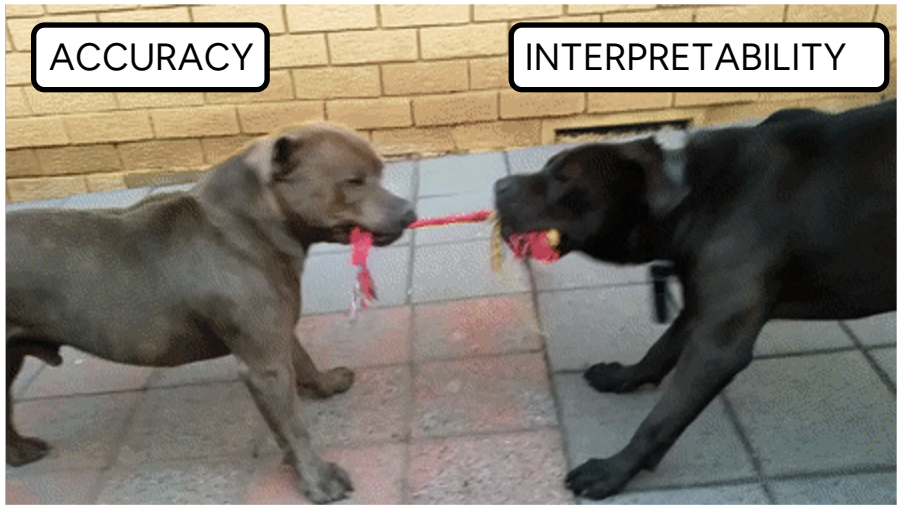
QUESTION

So why don't we use only interpretable models?

Introduction

QUESTION

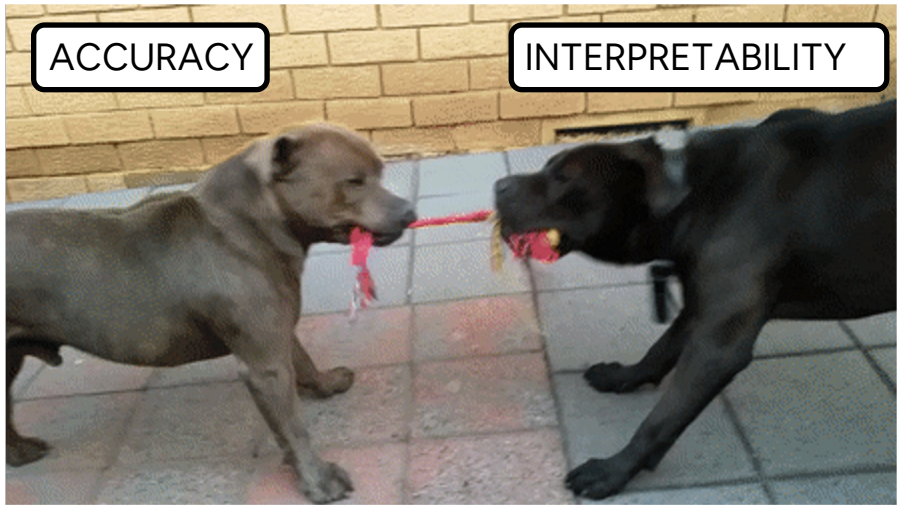
So why don't we use only interpretable models?



Introduction

QUESTION

So why don't we use only interpretable models?

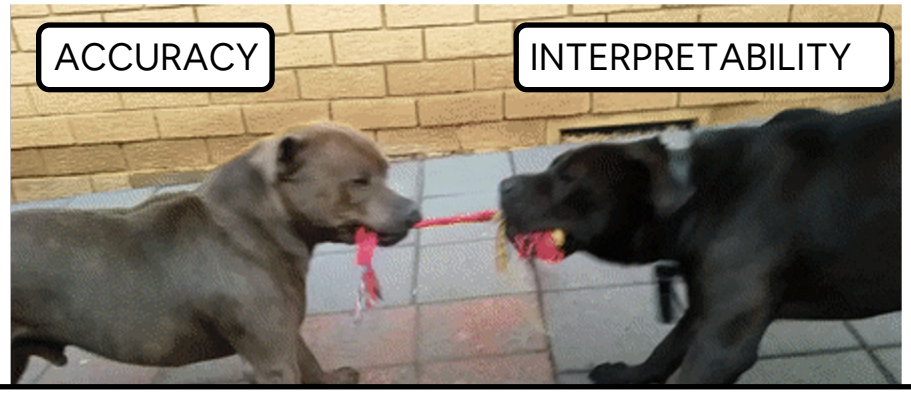


Not the only reason...

Introduction

QUESTION

So why don't we use only interpretable models?



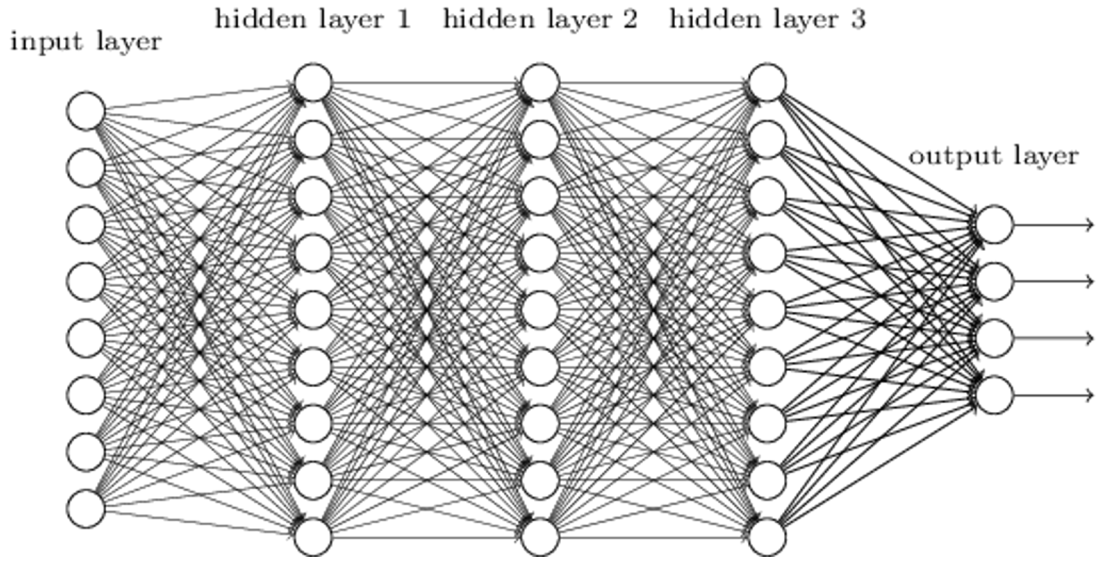
Not the only reason...

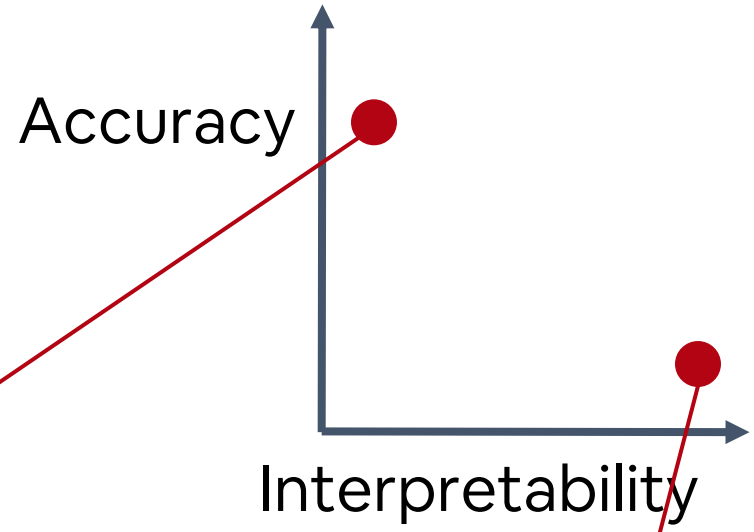
SPOILER

Models commonly considered as interpretable are actually not always easy to interpret!

Introduction

Deep Neural Networks





Linear Regression

$$y = a_1 \cdot x_1 + a_2 \cdot x_2 + \varepsilon$$

Introduction

Let's assume for a moment that we can choose between a highly accurate model and a (truly) interpretable model

Introduction

Let's assume for a moment that we can choose between a highly accurate model and a (truly) interpretable model

QUESTION

Which one would you choose?

Introduction

Let's assume for a moment that we can choose between a highly accurate model and a (truly) interpretable model

QUESTION

Which one would you choose?

Same old story: it depends!

Introduction

Let's assume for a moment that we can choose between a highly accurate model and a (truly) interpretable model

QUESTION

Which one would you choose?

Same old story: it depends!

following a project requirement

Introduction

Let's assume for a moment that we can choose between a highly accurate model and a (truly) interpretable model

QUESTION

Which one would you choose?

Same old story: it depends!

on the problem formalization

following a project requirement

Introduction

Let's assume for a moment that we can choose between a highly accurate model and a (truly) interpretable model

QUESTION

Which one would you choose?

Same old story: it depends!

to ensure technology acceptance

on the problem formalization

following a project requirement

Introduction

Let's assume for a moment that we can choose between a highly accurate model and a (truly) interpretable model

QUESTION

Which one would you choose?

Same old story: it depends!

- on the application
- to ensure technology acceptance
- on the problem formalization
- following a project requirement

Introduction

EXAMPLE 1 Stock market forecasting

From the web

Introduction

EXAMPLE 1 Stock market forecasting

We just need predictions!

From the web

Introduction

EXAMPLE 1 Stock market forecasting

From the web

We just need predictions!

The more accurate the predictions, the more profit we can make

Introduction

EXAMPLE 1 Stock market forecasting

From the web

We just need predictions!

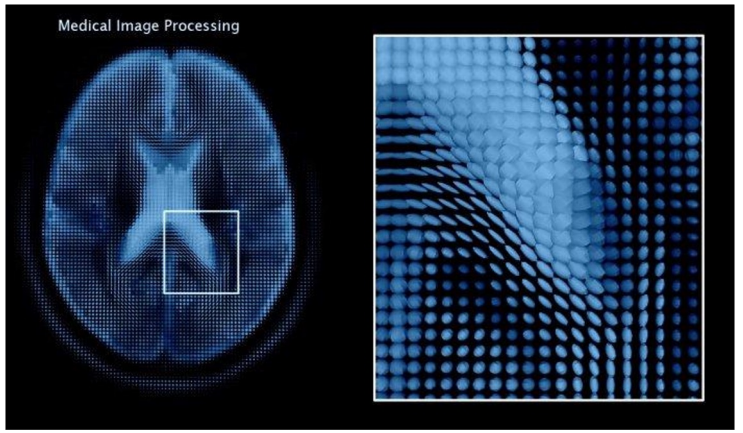
The more accurate the predictions, the more profit we can make

→ use the most accurate model, even though it is not interpretable!

Introduction

EXAMPLE 2

ML for automated medical diagnosis

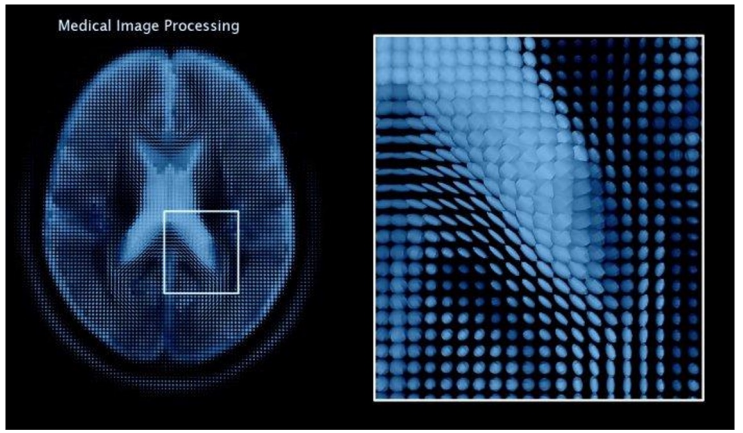


From the web

Introduction

EXAMPLE 2

ML for automated medical diagnosis



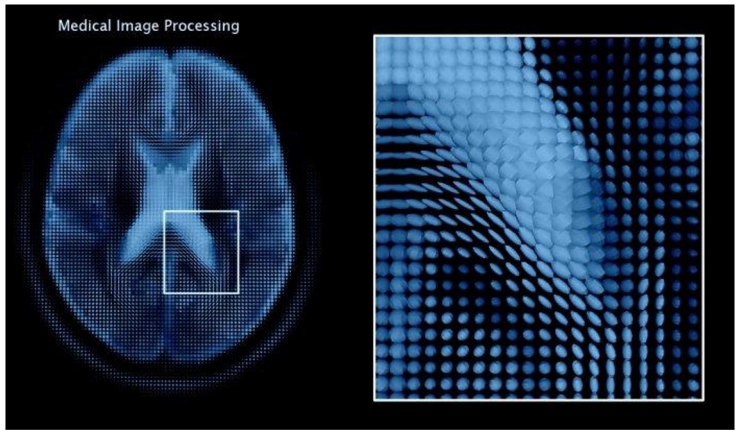
From the web

Doctors may want to verify whether they can trust or not the model's predictions, since they have to make hard decisions about other people's health

Introduction

EXAMPLE 2

ML for automated medical diagnosis



From the web

Doctors may want to verify whether they can trust or not the model's predictions, since they have to make hard decisions about other people's health

→ trade some accuracy for higher interpretability

Introduction

DISCUSSION Human surgeon VS robotic surgeon*

*inspired by "The Great AI Debate -NIPS2017" (https://www.youtube.com/watch?v=93Xv8vJ2acI)

Introduction

DISCUSSION Human surgeon VS robotic surgeon*

You have a disease and you need surgery. You can choose between:

*inspired by "The Great AI Debate -NIPS2017" (https://www.youtube.com/watch?v=93Xv8vJ2acI)

Introduction

DISCUSSION Human surgeon VS robotic surgeon*

You have a disease and you need surgery. You can choose between:

HUMAN SURGEON

- 10% mortality rate
- Fully interpretable
- Profound knowledge of human body and functioning based on years of study and experience

*inspired by "The Great AI Debate -NIPS2017" (https://www.youtube.com/watch?v=93Xv8vJ2acI)

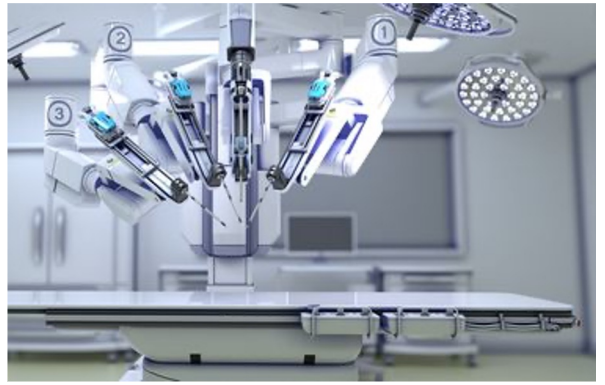
Introduction

DISCUSSION Human surgeon VS robotic surgeon*

You have a disease and you need surgery. You can choose between:

HUMAN SURGEON
- 10% mortality rate
- Fully interpretable
- Profound knowledge of human body and functioning based on years of study and experience

ROBOTIC SURGEON
- 1% mortality rate
- Black-box
- Trained on examples for a single task



*inspired by "The Great AI Debate -NIPS2017" (https://www.youtube.com/watch?v=93Xv8vJ2acI)

Introduction

DISCUSSION Human surgeon VS robotic surgeon*

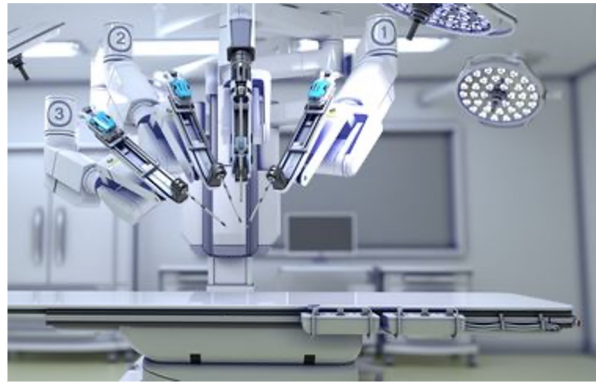
You have a disease and you need surgery. You can choose between:

HUMAN SURGEON
- 10% mortality ra
- Fully interpretab
- Profound knowledge of human body and functioning based on years of study and experience

QUESTION

Which one would you pick?

RGEON
- y rate
- Trained on examples for a single task



*inspired by "The Great AI Debate -NIPS2017" (https://www.youtube.com/watch?v=93Xv8vJ2acI)

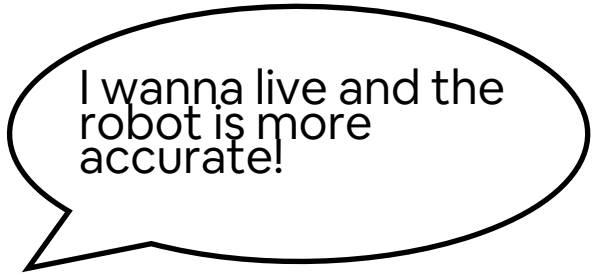
Introduction

TEAM ROBOT

TEAM HUMAN

Introduction

TEAM ROBOT



TEAM HUMAN

Introduction

TEAM ROBOT

I wanna live and the robot is more accurate!

TEAM HUMAN

Was the testing procedure conducted properly? Is the test set representative of the real-world scenario?

Introduction

TEAM ROBOT

I wanna live and the robot is more accurate!

TEAM HUMAN

Was the testing procedure conducted properly? Is the test set representative of the real-world scenario?

We cannot just deploy the model and see whether people die or not...

Introduction

TEAM ROBOT

I wanna live and the robot is more accurate!

That's exactly how drugs are tested!

TEAM HUMAN

Was the testing procedure conducted properly? Is the test set representative of the real-world scenario?

We cannot just deploy the model and see whether people die or not...

Introduction

TEAM ROBOT

I wanna live and the robot is more accurate!

That's exactly how drugs are tested!

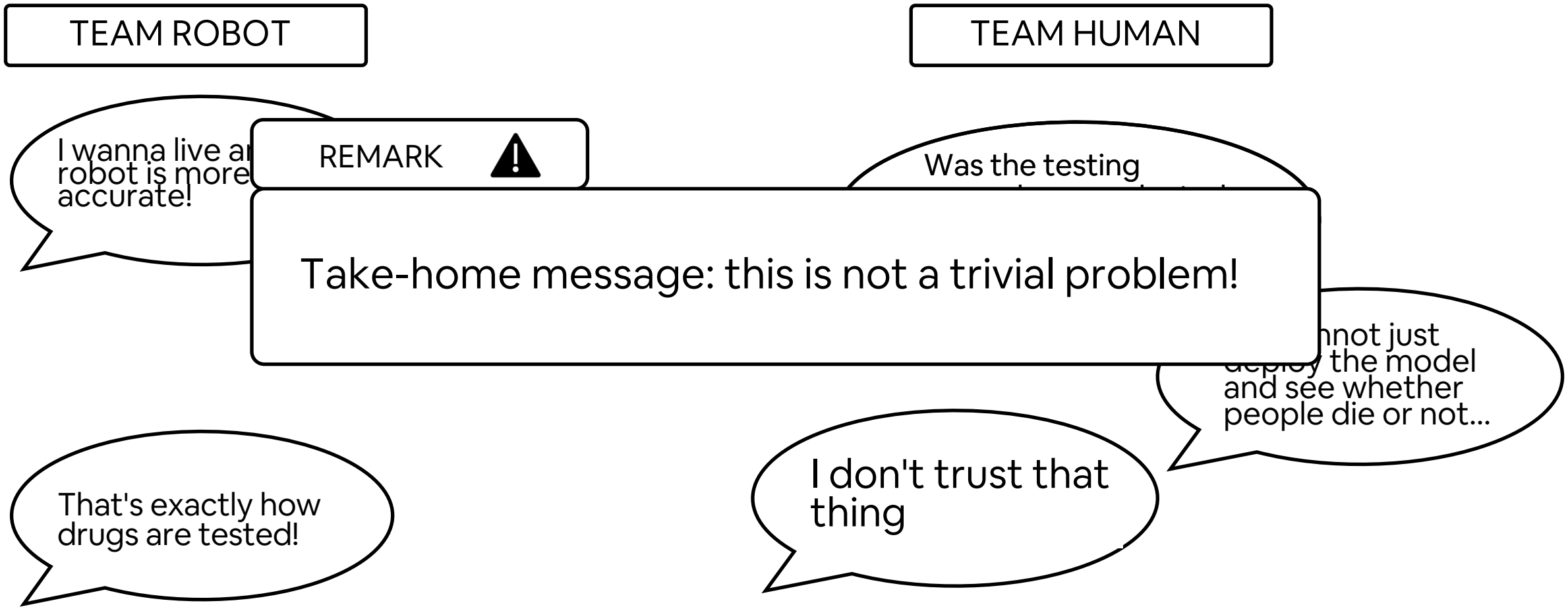
TEAM HUMAN

Was the testing procedure conducted properly? Is the test set representative of the real-world scenario?

We cannot just deploy the model and see whether people die or not...

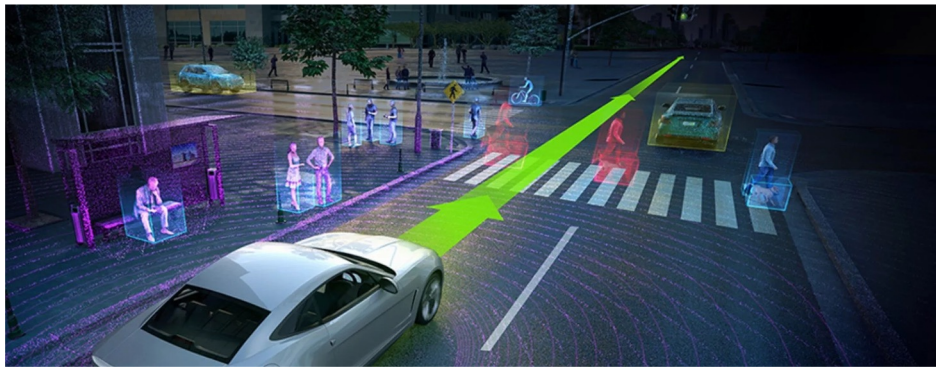
I don't trust that thing

Introduction



Introduction

EXAMPLE 3 Image recognition for self-driving cars

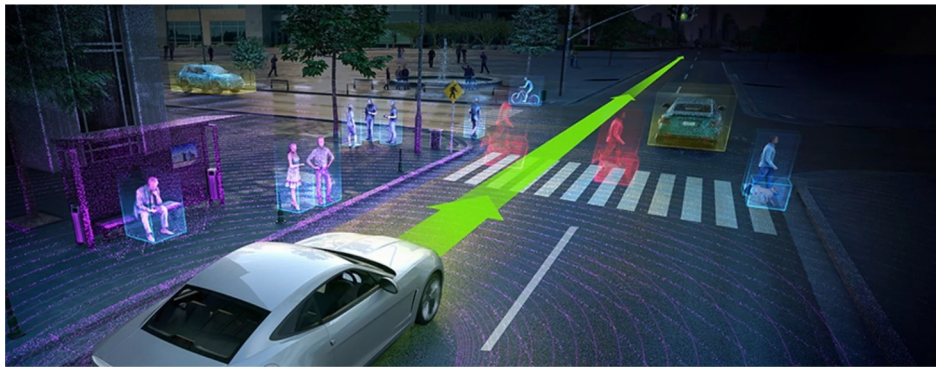


From the web

Introduction

EXAMPLE 3

Image recognition for self-driving cars



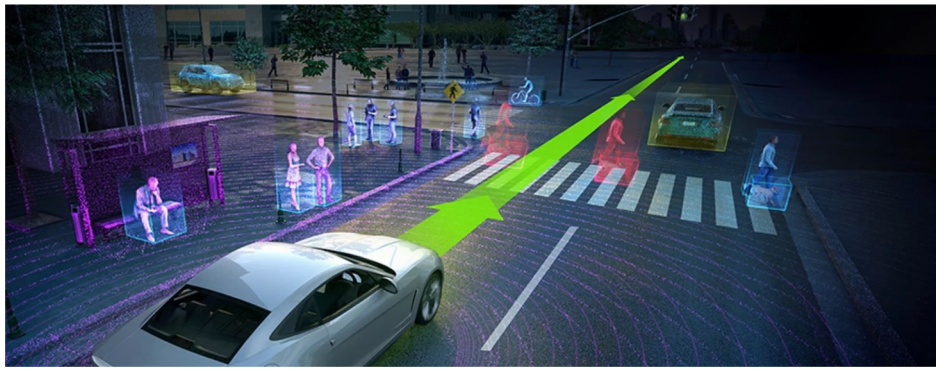
From the web

We want to be 100% sure that our model performs well in every possible real scenario

Introduction

EXAMPLE 3

Image recognition for self-driving cars



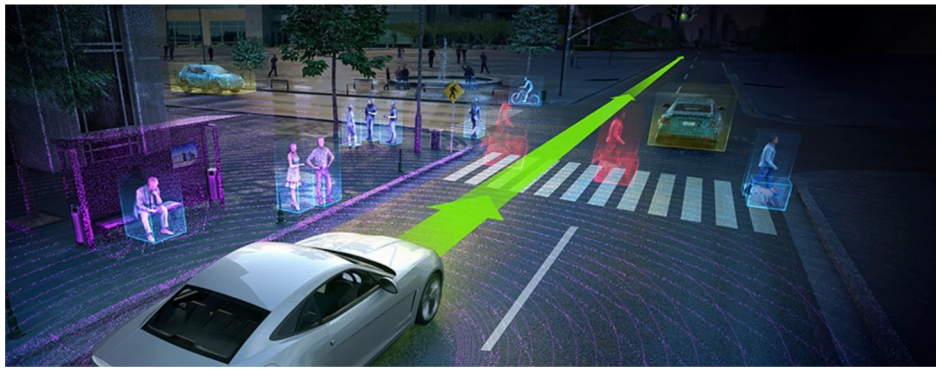
From the web

We want to be 100% sure that our model performs well in every possible real scenario

Consequences of poor performance in real applications may be catastrophic!

Introduction

EXAMPLE 3 Image recognition for self-driving cars



From the web

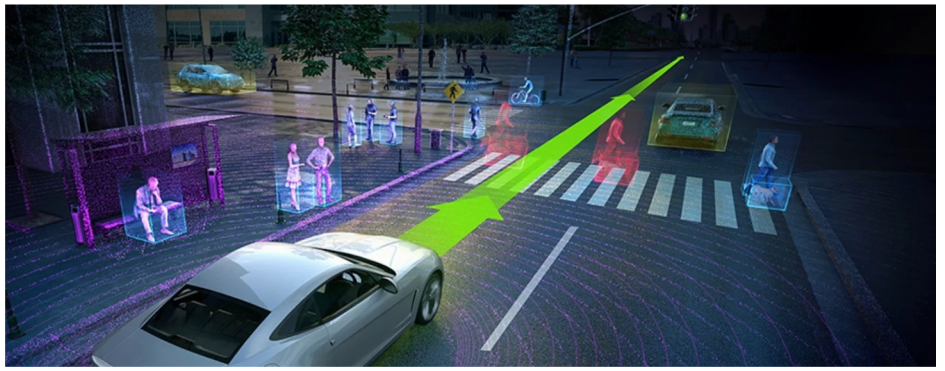
We want to be 100% sure that our model performs well in every possible real scenario

Consequences of poor performance in real applications may be catastrophic!

PROBLEM: we may not be able to test the model simulating all possible scenarios

Introduction

EXAMPLE 3 Image recognition for self-driving cars



From the web

We want to be 100% sure that our model performs well in every possible real scenario

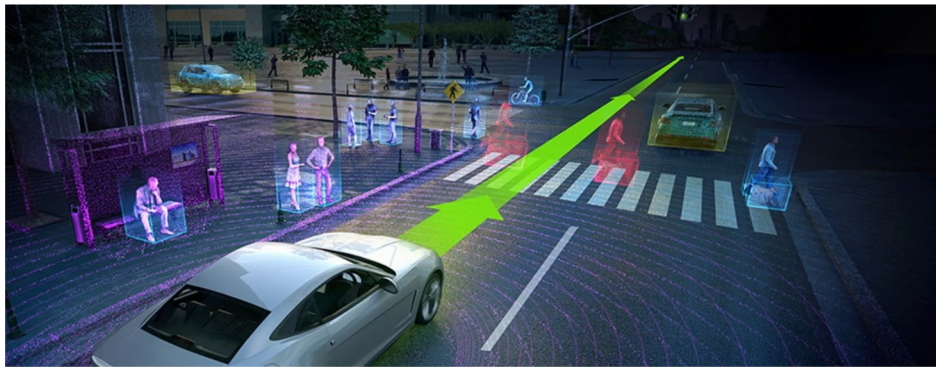
Consequences of poor performance in real applications may be catastrophic!

PROBLEM: we may not be able to test the model simulating all possible scenarios

Interpretability can help in understanding whether the model has learnt reasonable representations for the task at hand

Introduction

EXAMPLE 3 Image recognition for self-driving cars



From the web

We want to be 100% sure that our model performs well in every possible real scenario

Consequences of poor performance in rea

Incompleteness in the problem formalization (evaluation procedure)!

PROBLEM: we may not be able to test the model simulating all possible scenarios

Interpretability can help in understanding whether the model has learnt reasonable representations for the task at hand

Introduction

EXAMPLE 4

Automatic Risk Modeling in Insurance

From the web

Insurance companies are using Machine Learning approaches that consider many factors to perform risk modeling.

Insurance premiums can be based on computed risks and people can be denied an insurance coverage based on such risks.

While companies may only be interested in maximizing their profits, they need to care about interpretability...

Introduction

From the web

EXAMPLE 4

Automatic Risk Modeling in Insurance

Insurance companies are using Machine Learning approaches that consider many factors to perform risk modeling.

Insurance premiums can be based on computed risks and people can be denied an insurance coverage based on such risks.

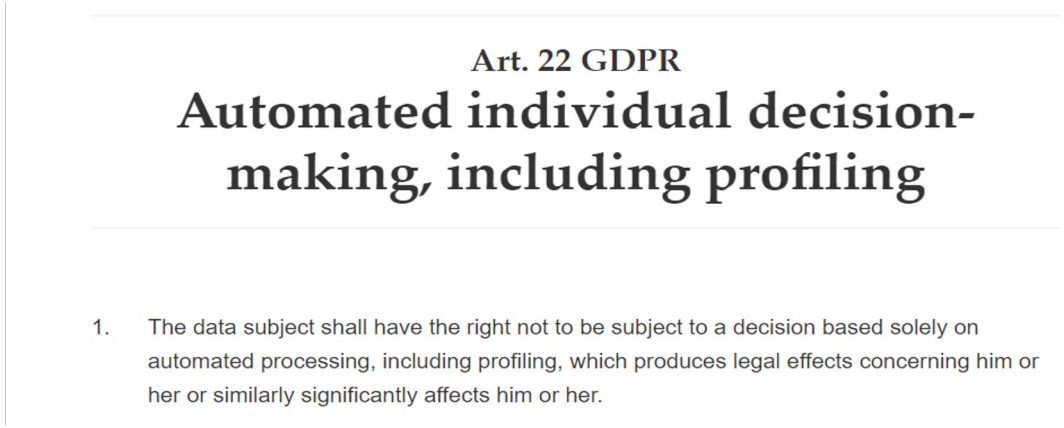
While companies may only be interested in maximizing their profits, they need to care about interpretability...

(For example, we found that in Italy car insurance companies are not properly dealing with fairness issues in insurance costs https://arxiv.org/abs/2105.10174)

Introduction

General Data Protection Regulation (GDPR)

So-called "Right to explanation": in critical applications (like in the medical and legal domain), any decision involving human beings based on automated processing should be adequately justified



Art. 22 GDPR

Automated individual decision-making, including profiling

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

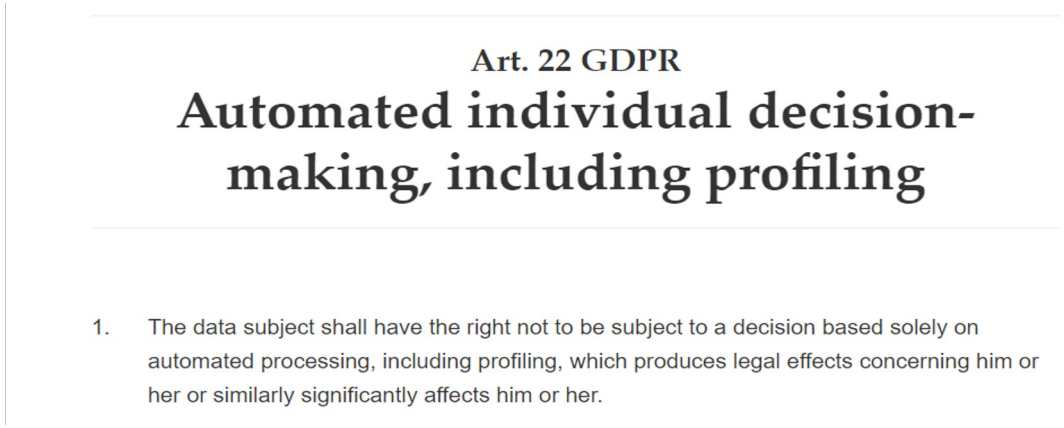
https://gdpr-info.eu/art-22-gdpr/

Introduction

Interpretability as a constraint, required in the development of the ML technology

General Data Protection Regulation (GDPR)

So-called "Right to explanation": in critical applications (like in the medical and legal domain), any decision involving human beings based on automated processing should be adequately justified



Art. 22 GDPR

Automated individual decision-making, including profiling

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

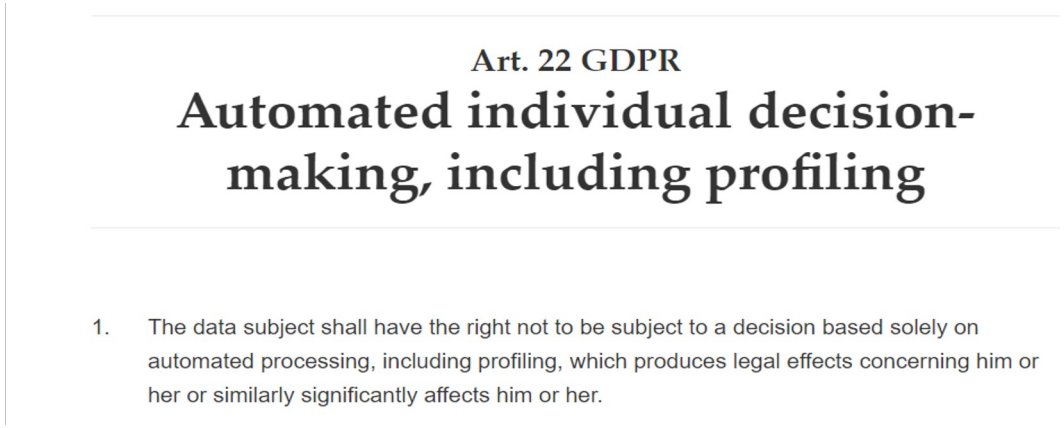
https://gdpr-info.eu/art-22-gdpr/

Introduction

Interpretability as a constraint, required in the development of the ML technology

General Data Protection Regulation (GDPR)

So-called "Right to explanation": in critical applications (like in the medical and legal domain), any decision involving human beings based on automated processing should be adequately justified



Art. 22 GDPR

Automated individual decision-making, including profiling

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

https://gdpr-info.eu/art-22-gdpr/

AI Act is coming!

https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence

Introduction

QUESTION

How does ML and ML interpretability relate to you and your work?

Introduction

QUESTION

How does ML and ML interpretability relate to you and your work?

TEAM DEVELOPERS

- ML experts
- Aware of ML limitations and guarantees
- Desire to get additional information from the ML model (besides predictions)

Introduction

QUESTION

How does ML and ML interpretability relate to you and your work?

TEAM DEVELOPERS

- ML experts
- Aware of ML limitations and guarantees
- Desire to get additional information from the ML model (besides predictions)

TEAM USERS

- ML end-users
- Need to use ML to get better results (or reduce the effort)
- Need to soften your colleagues' skepticism towards opaque ML systems

Concepts: Definitions

Most of the key concepts in the field of explainable AI have no clear definition

The name of the research field itself is not unique and many expressions are being used interchangeably:

- Explainable AI (XAI)
- Explainable Machine Learning
- Interpretable Machine Learning
- Transparent Machine Learning
- …

Concepts: Definitions

Although the problem is not new, only recently some researchers and scientists put a big effort in organizing concepts, procedures and solutions

Throughout this course we mainly rely on the following works:

- Lipton, Z. C. (2018). The mythos of model interpretability. Queue, 16(3), 31-57.
- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
- 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.
- 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.
- Molnar, C. (2019). Interpretable machine learning. Lulu. com.

Concepts: Definitions

"*Interpretability*" is an ill-defined term, which might be used in the ML literature to refer to slightly different ideas

Generally speaking, we can define "interpretability" as the science/art of producing descriptions simple enough to be easily understood by humans

While representing a concrete attempt to increase trust in black-box models, it is just the first step towards a more ambitious goal...

Concepts: Definitions

As claimed by Gilpin et al., interpretability might not suffice to provide a comprehensive solution to the problem of opaque models

⟶ we need "*explainability*"!

Explainability = *Interpretability* **AND** *Completeness*

where an explanation is considered as "complete" when it allows humans to anticipate the model's prediction

Concepts: Definitions

QUESTION

Would you prefer an interpretable explanation or a complete one?

Concepts: Definitions

```
list(model.parameters())
```

```
torch.Size([10, 784]) torch.Size([10])

[Parameter containing:
 tensor([[ 4.8673e-03,  2.5654e-02,  1.4312e-02,  ..., -1.4949e-02,
          -1.1675e-02,  1.8740e-02],
         [ 1.2544e-03, -1.2904e-02,  1.0140e-02,  ..., -1.5626e-02,
           2.9115e-02,  3.5050e-03],
         [-3.0447e-02,  1.6315e-02, -8.7722e-03,  ..., -6.3704e-03,
          -1.2951e-02, -3.4346e-02],
         ...,
         [ 2.4438e-02,  2.6935e-02,  4.3357e-03,  ...,  1.2128e-03,
          -3.4761e-02,  2.4345e-03],
         [-3.5191e-03,  3.3461e-02, -9.0063e-03,  ...,  2.0578e-02,
           1.8074e-02,  2.5010e-03],
         [ 3.1909e-02, -6.9384e-03,  1.4326e-03,  ..., -5.1625e-05,
          -7.3041e-03, -2.7546e-02]], requires_grad=True),
 Parameter containing:
 tensor([ 0.0299,  0.0349, -0.0304,  0.0285,  0.0297,  0.0052,  0.012
7,  0.0190,
          0.0332, -0.0139], requires_grad=True)]
```

Concepts: Definitions

```
list(model.parameters())
```

```
torch.Size([10, 784]) torch.Size([10])

[Parameter containing:
 tensor([[ 4.8673e-03,  2.5654e-02,  1.4312e-02,  ..., -1.4949e-02,
          -1.1675e-02,  1.8740e-02],
         [ 1.2544e-03, -1.2904e-02,  1.0140e-02,  ..., -1.5626e-02,
           2.9115e-02,  3.5050e-03],
         [-3.0447e-02,  1.6315e-02, -8.7722e-03,  ..., -6.3704e-03,
          -1.2951e-02, -3.4346e-02],
         ...,
         [ 2.4438e-02,  2.6935e-02,  4.3357e-03,  ...,  1.2128e-03,
          -3.4761e-02,  2.4345e-03],
         [-3.5191e-03,  3.3461e-02, -9.0063e-03,  ...,  2.0578e-02,
           1.8074e-02,  2.5010e-03],
         [ 3.1909e-02, -6.9384e-03,  1.4326e-03,  ..., -5.1625e-05,
          -7.3041e-03, -2.7546e-02]], requires_grad=True),
 Parameter containing:
 tensor([ 0.0299,  0.0349, -0.0304,  0.0285,  0.0297,  0.0052,  0.012
7,  0.0190,
          0.0332, -0.0139], requires_grad=True)]
```

Complete explanation of a very simple Neural Net... Is it useful?

Concepts: Definitions

As you might imagine, since we cannot have both interpretability and completeness, we end up with another trade-off!

- Interpretable explanations may be too simple to catch the whole logic behind the model's predictions

- Complete explanations may lose communication power due to their overwhelming level of detail

Concepts

> “ *The need for interpretability stems from an incompleteness in the problem formalization.*”
>
> [*Towards A Rigorous Science of Interpretable Machine Learning*,
>
> Finale Doshi-Velez and Been Kim]

ML models are optimized to minimize the error, but in real-world applications we usually require additional features which cannot be translated into an optimization problem

Example: how can we force a ML model to be “*ethical*”? Can we quantify ethical traits/make such characteristic measurable?

Concepts

Desidered Qualities of AI systems:

- Trust
- Causality
- Transferability
- Informativeness
- Robustness
- Fairness
- ... and many others!

Concepts

QUESTION

Ok, but... how does interpretability relate to this?

Concepts

QUESTION

Ok, but... how does interpretability relate to this?

Interpretability is used to verify whether the mentioned desiderata are met or not!

Of course, it does not provide the solution to the problem, but rather represents a reasonable tool to flag the existence of *incompleteness* in the problem formalization

Concepts

Desidered Qualities of AI systems:

- Trust
- Causality
- Transferability
- Informativeness
- Robustness
- Fairness
- … and many others!

Concepts

Desidered Qualities of AI systems - Trust

Interpretability as a prerequisite for trust (in ML systems)

What is trust?

- Is it guaranteed that the model will perform well when deployed? Will the model be robust to perturbations (in the data/parameters)?

- Is it a subjective notion (a personal preference for a specific model)?

Concepts

Desidered Qualities of AI systems - Trust (2)

We may also consider how the model behaves compared to humans

- Does the model fail on the same examples on which also humans fail?
- In many applications, ML-based approaches have to outperform humans in order to be trusted!

This is a crucial question we need to answer when deciding whether to maintain or not human supervision

Concepts

Desidered Qualities of AI systems:

- Trust
- Causality
- Transferability
- Informativeness
- Robustness
- Fairness
- ... and many others!

Concepts

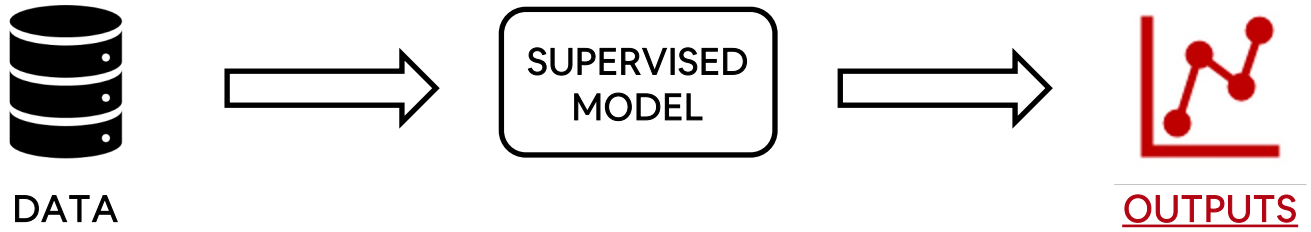
Desidered Qualities of AI systems:

- Trust
- Causality
- Transferability
- Informativeness
- Robustness
- Fairness
- … and many others!

Concepts

Desidered Qualities of AI systems – Informativeness

A ML model provides information most commonly through its outputs

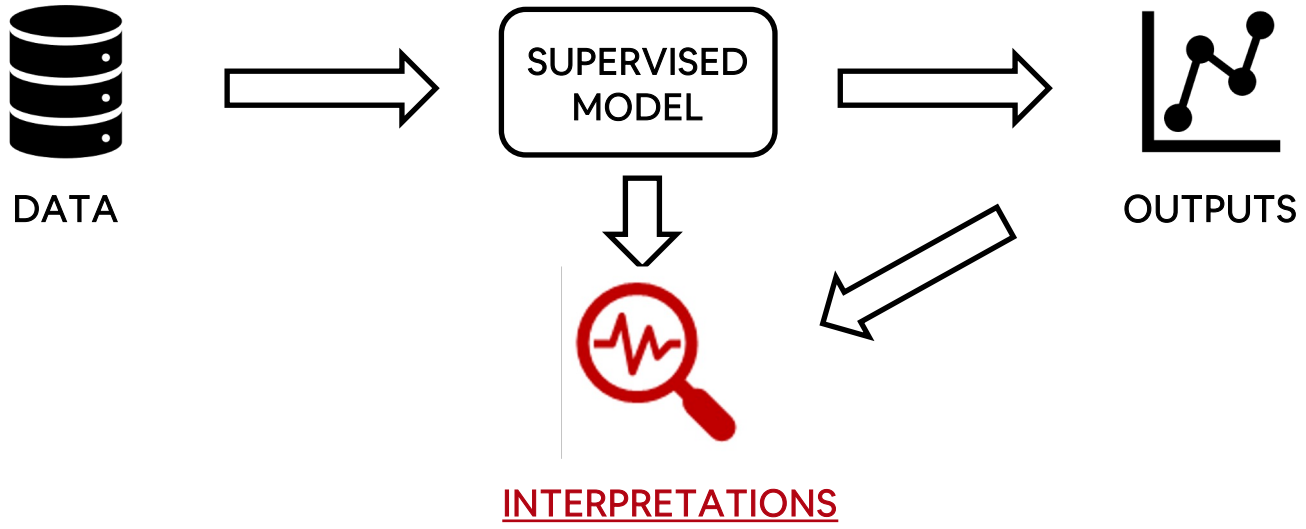


This paradigm is suitable when we are mainly interested in the outputs

Concepts

Desidered Qualities of AI systems – Informativeness (2)

An alternative way a ML model could provide information is through the interpretation of its outputs or structure



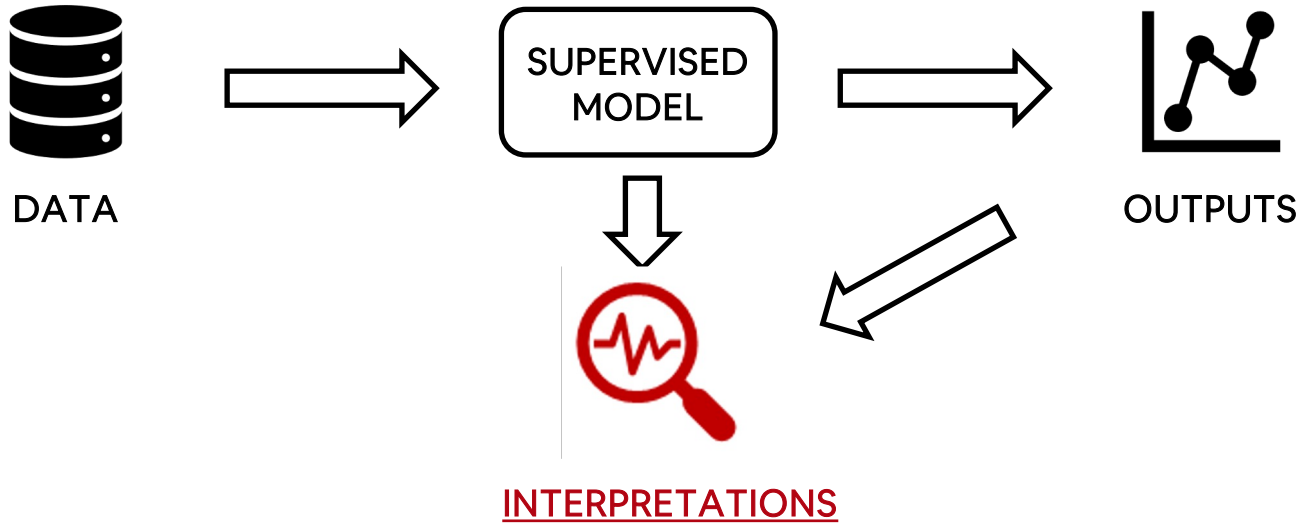
Concepts

Desidered Qualities

REMARK ⚠

In this context, the formulation as a ML problem may be just a proxy to gain additional knowledge about the problem at hand!

An alternative way a
the interpretation of its outputs or structure



Concepts

Desidered Qualities of AI systems:

- Trust
- Causality
- Transferability
- Informativeness
- Robustness
- Fairness
- ... and many others!

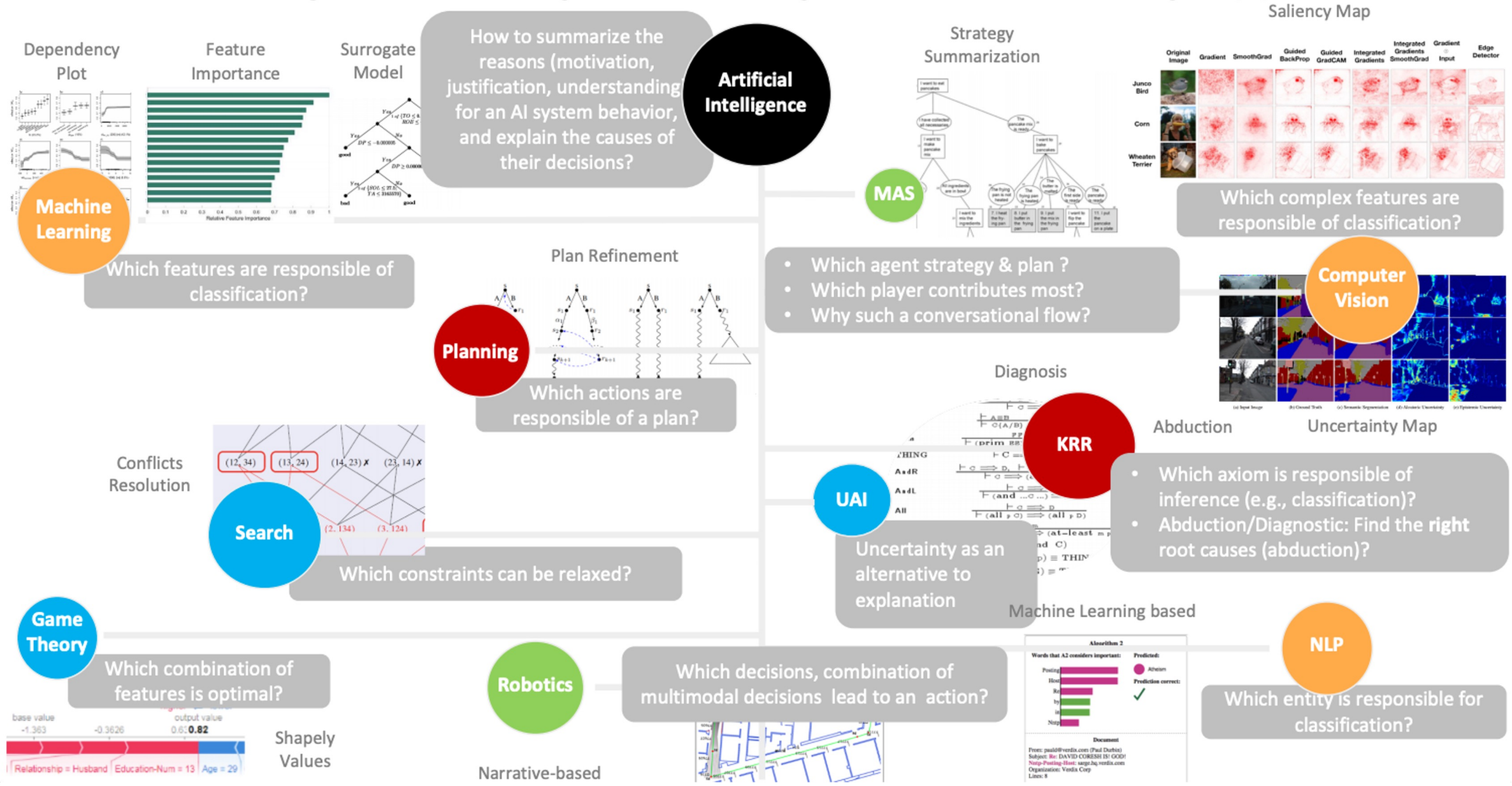
Concepts

Desidered Qualities of AI systems – Fairness

We must (not just “should”!) make sure that predictions made by a ML model do not discriminate protected groups

Example: if we train a model to predict the risk of recidivism (Compas https://www.psychologytoday.com/us/blog/psyched/201801/law-enforcement-ai-is-no-more-or-less-biased-people), we need to ensure that the predictions do not rely on the ethnicity

XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches



F. Lecue et al. On Explainable AI: From Theory to Motivation, Industrial Applications and Coding Practices, AAAI2021

Taxonomy

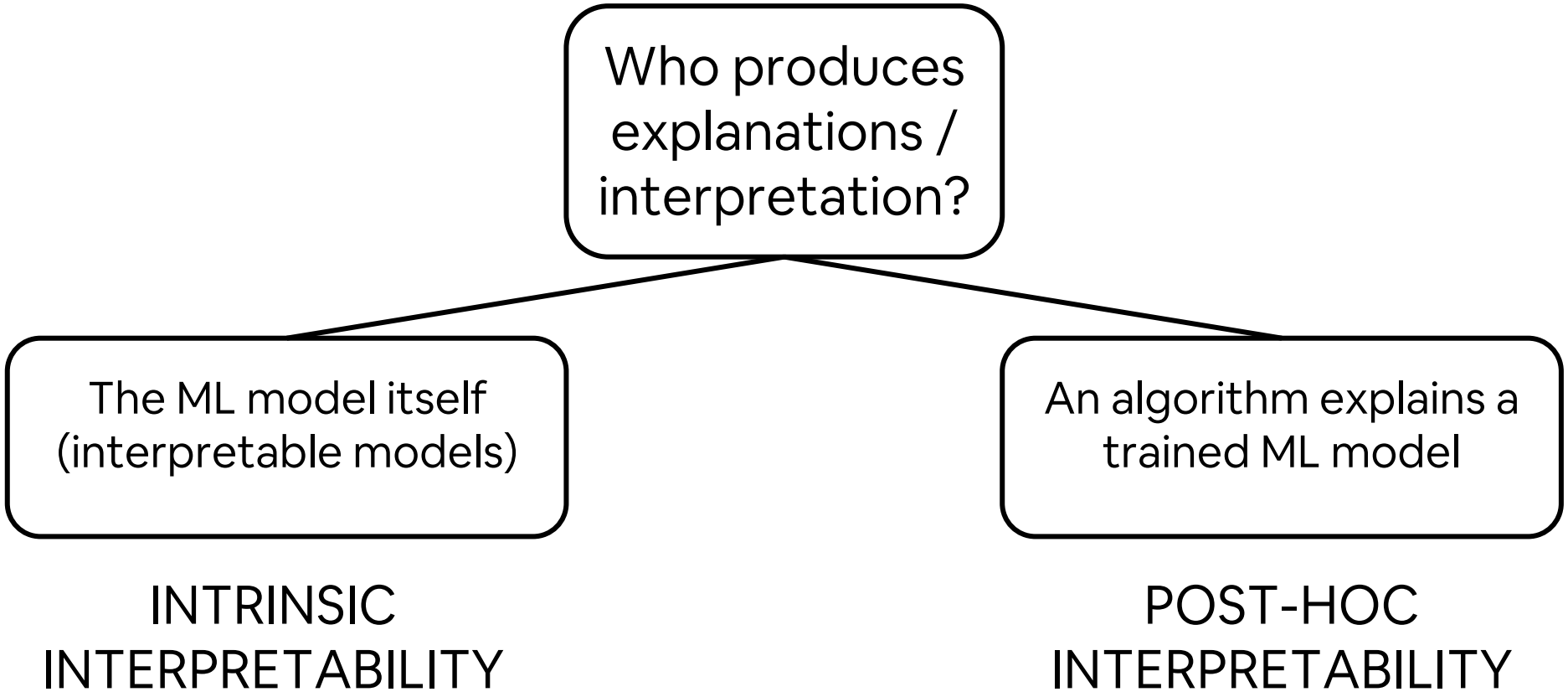
Interpretability methods can be categorized according to several criteria, depending on the specific aspect we want to highlight

In other words, each criterion characterize the problem of interpretability as viewed from a particular angle

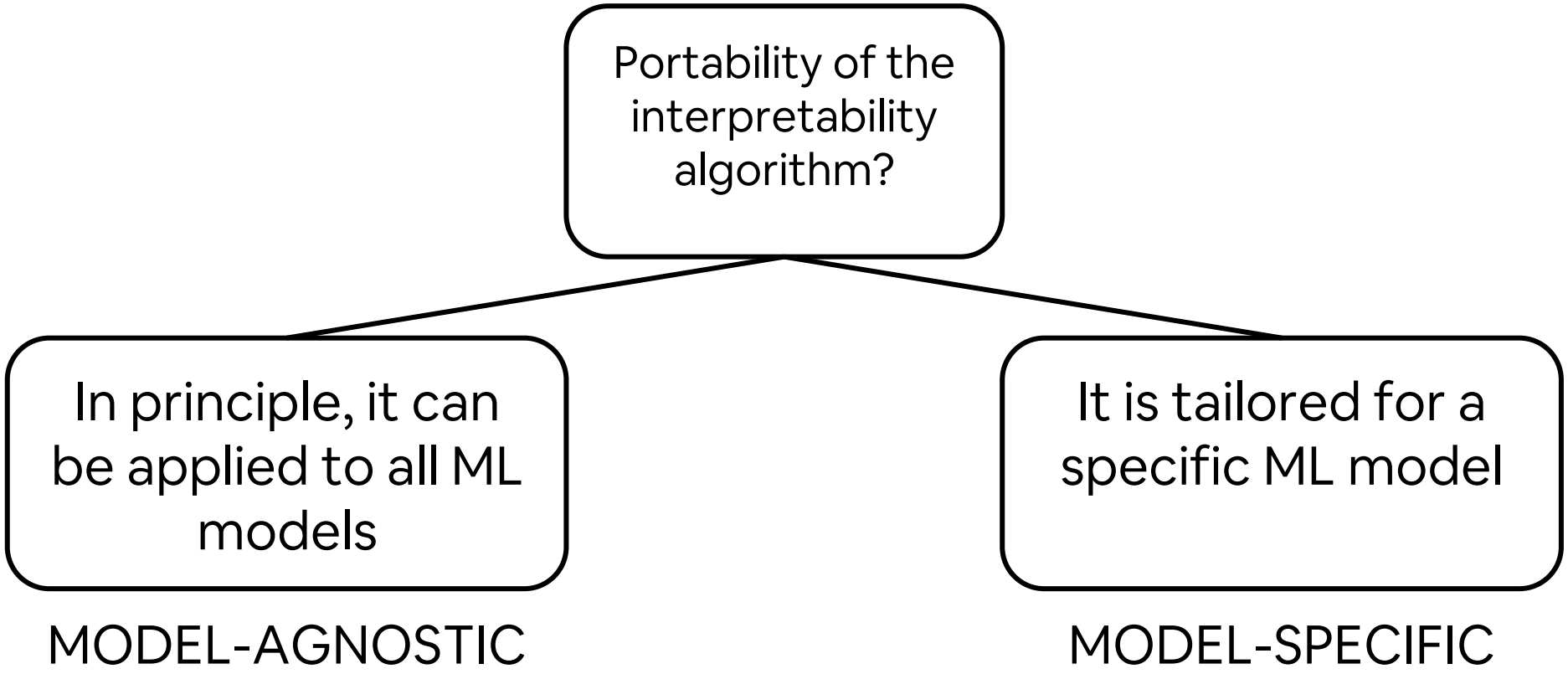
In this lecture we focus on 3 dichotomies:

- intrinsic vs post-hoc interpretability
- model-agnostic vs model-specific methods
- global vs local methods

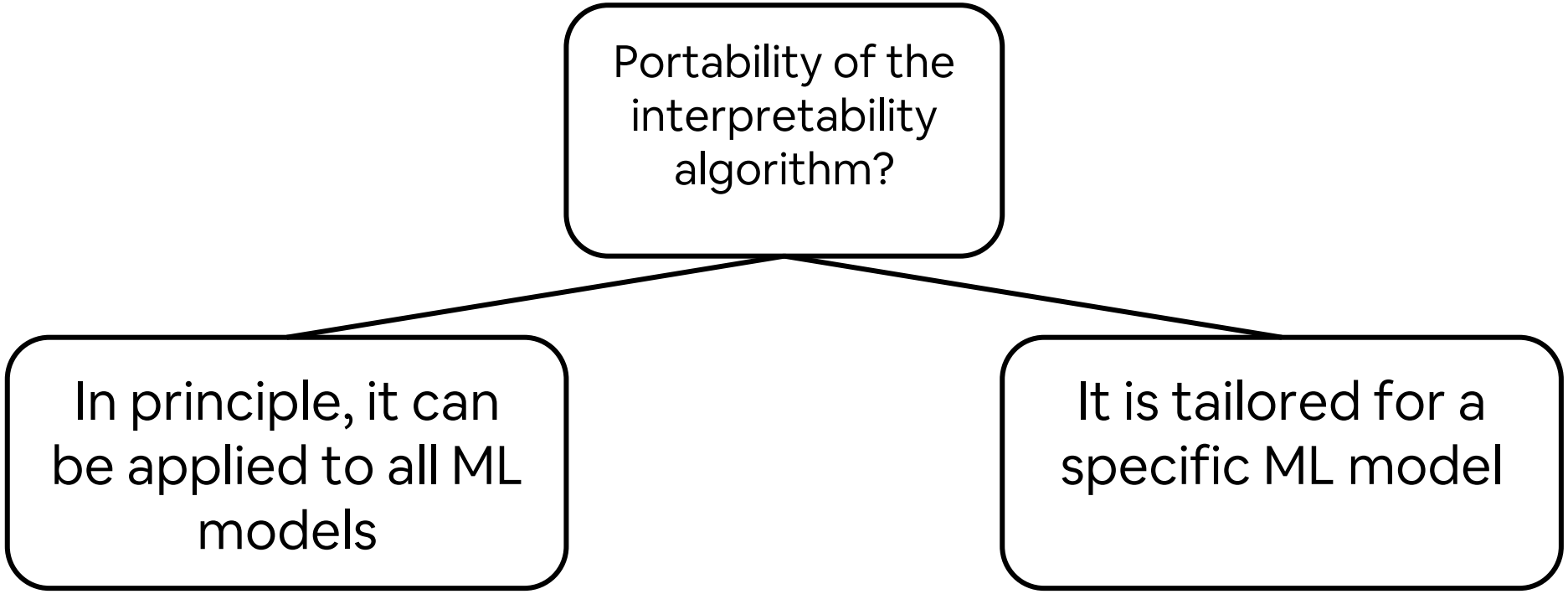
Taxonomy: intrinsic vs post-hoc



Taxonomy: model-agnostic vs model-specific



Taxonomy: model-agnostic vs model-specific



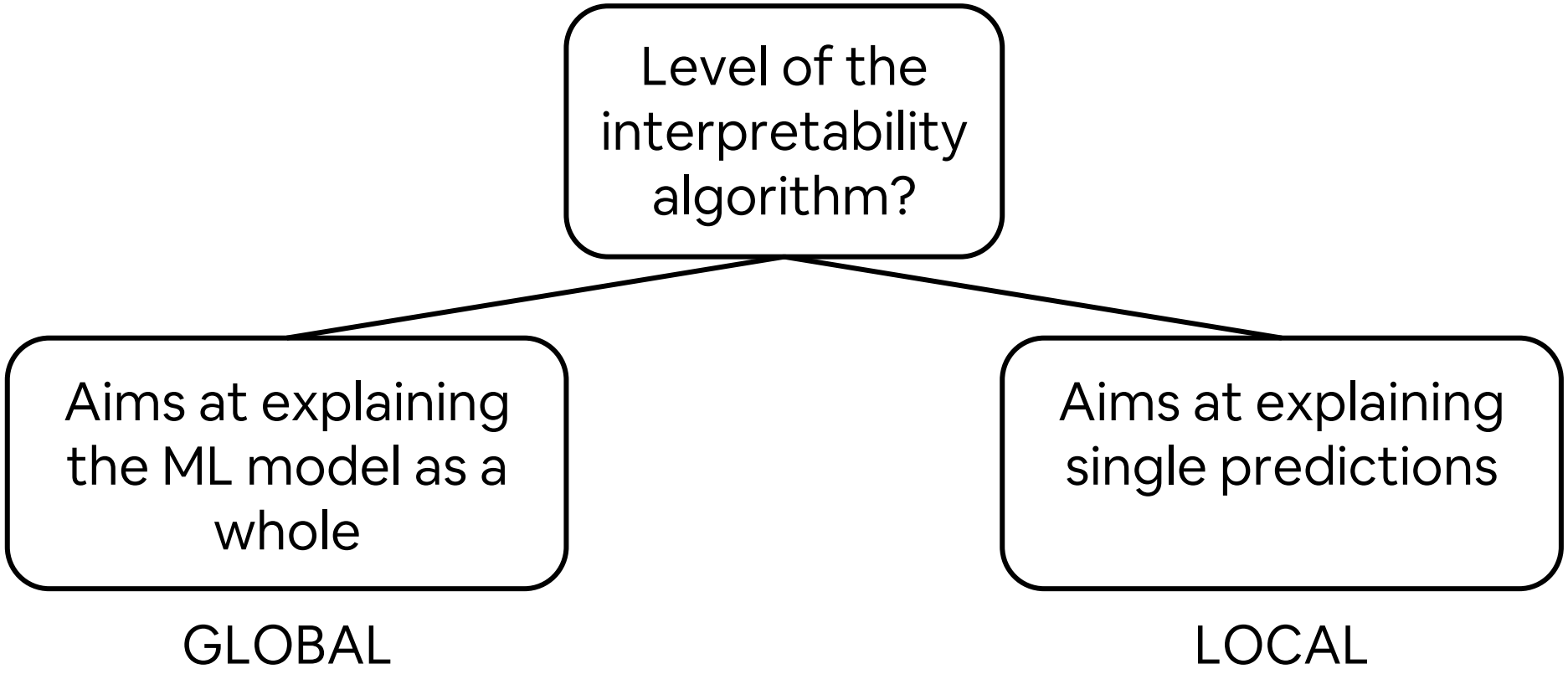
MODEL-AGNOSTIC

+ Portability
+ Easier comparison with different models

MODEL-SPECIFIC

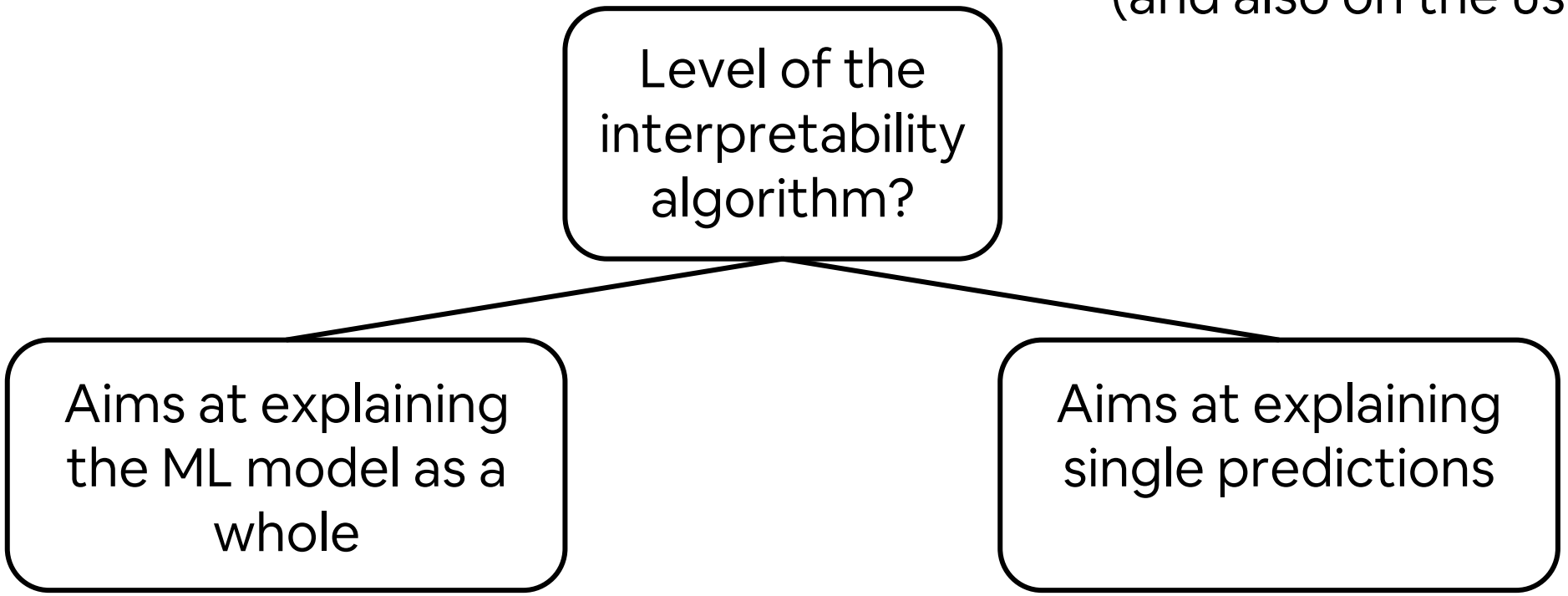
+ Accuracy
+ Easier/Faster to be computed

Taxonomy: global vs local



Taxonomy: global vs local

Which one to choose?
It depends on the application
(and also on the user)



GLOBAL

Interesting for example for Knowledge Extraction (life science team) and for developers

LOCAL

Interesting for example to users

Output of interpretability methods

Preliminaries

We recall that, in supervised settings with tabular data, a *data point* is composed of

- input variables (or features) $x_1, x_2, \dots, x_p$
- target variable y (what we want to predict)

Case	Factor				Typical output	
	Mesh method	Element type	Boundary condition	Constitutive model	Computational cost (s)	Computa-tional error
1	1	1	1	1	191	167.3
2	1	2	2	2	193	758.2
3	1	3	3	3	198	27.5
4	2	1	2	3	312	18.8
5	2	2	3	1	241	64.4
6	2	3	1	2	354	769.1
7	3	1	3	2	534	755.9
8	3	2	1	3	780	55.2
9	3	3	2	1	674	62.2

Output of interpretability methods

Interpretability methods' outputs can be:

- **simple feature summary statistics:** for each feature, a quantity representing the corresponding importance is provided; advanced methods also provide importance for each pair of features to explain how different features interact with each other
- **feature summary visualizations:** more complex feature summary statistics which can be effectively provided in the form of curves/plots
- **data points:** to explain a single prediction, we can provide the user with a set of similar data points obtained by slightly changing some features and whose predictions are significantly dissimilar w.r.t. the point we want to explain (so-called "*counterfactual examples*")
- **...**

(adapted from Molnar)

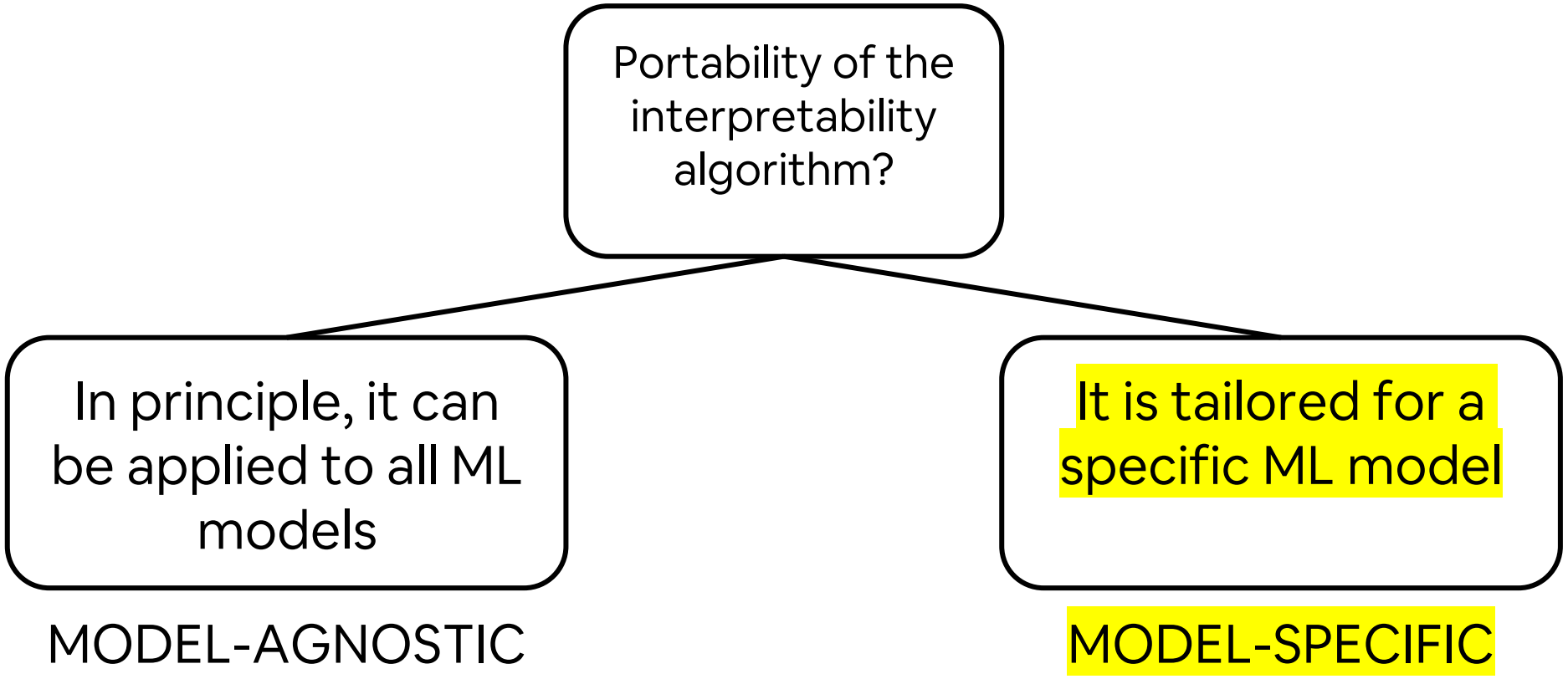
Good explanations

Good explanations should be:

- **contrastive:** instead of answering the question “why class A?”, we should answer (when possible) “why not class B?”
- **short and intuitive:** as we have seen in the first lecture, *complete* explanations are usually not very useful; a good explanations should focus on few very important features
- **tailored on the audience:** explanations for technical audience should be different from explanations for non-experts
- **…**

(adapted from Molnar)

Taxonomy: model-agnostic vs model-specific



Feature statistics, model specific - Linear Regression

Linear models, while typically not the preferred choice if we are aiming at high accuracy, are however an obvious choice when it came to interpretability

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$$

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta_0,\ldots,\beta_p} \sum_{i=1}^{n} \left(y^{(i)} - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j^{(i)} \right) \right)^2$$

Feature statistics, model specific - Linear Regression

Linear models, while typically [illegible] ning at high accuracy, are however an ob[illegible] tability

WEIGHTS are the simplest describers of feature importance

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$$

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta_0,\ldots,\beta_p} \sum_{i=1}^{n} \left(y^{(i)} - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j^{(i)} \right) \right)^2$$

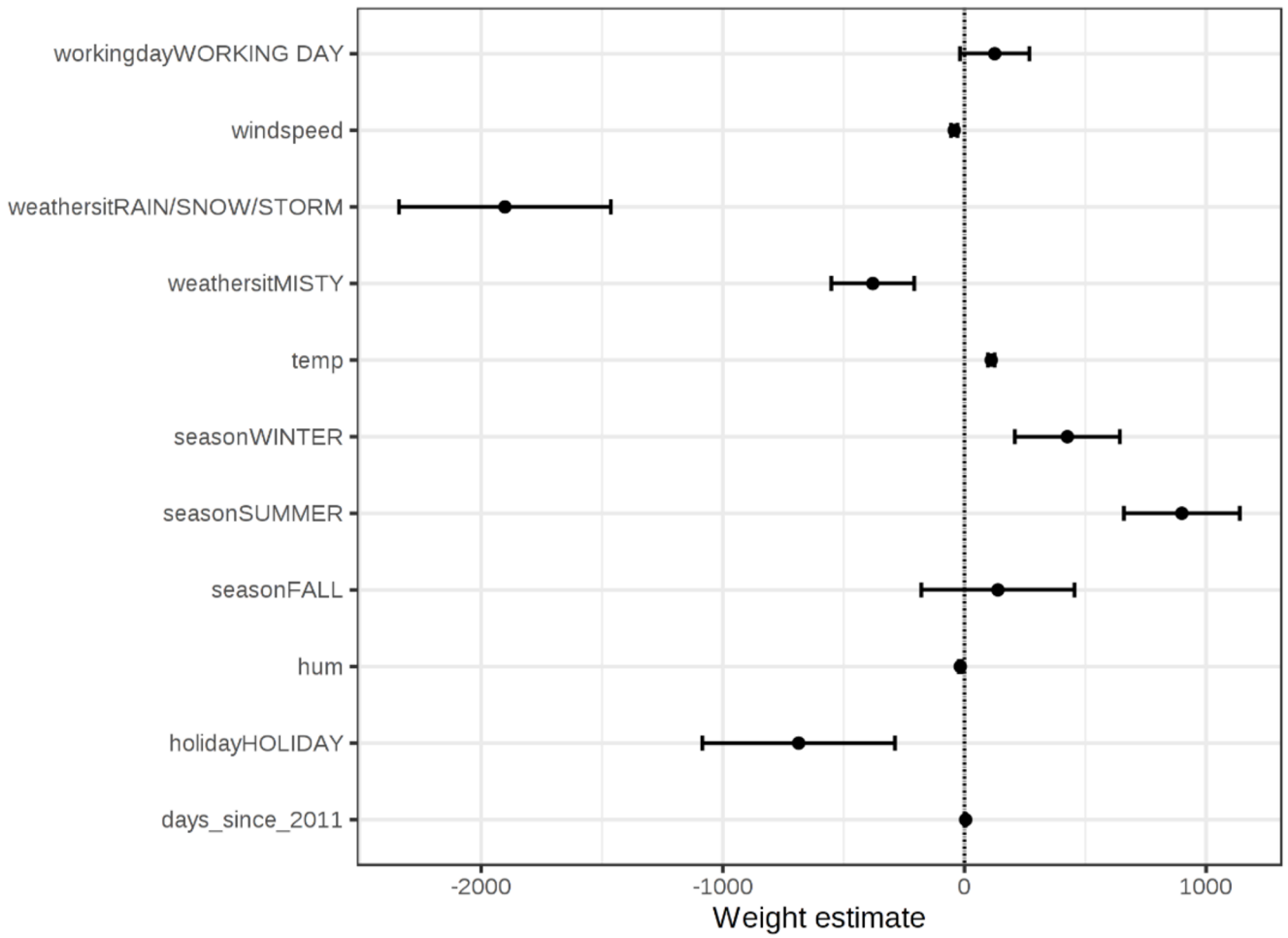
Feature statistics, model specific - Linear Regression

	Weight
(Intercept)	2399.4
seasonSUMMER	899.3
seasonFALL	138.2
seasonWINTER	425.6
holidayHOLIDAY	-686.1
workingdayWORKING DAY	124.9
weathersitMISTY	-379.4
weathersitRAIN/SNOW/STORM	-1901.5
temp	110.7
hum	-17.4
windspeed	-42.5
days_since_2011	4.9

Let's consider the example of prediction of rented bikes given weather and calendar information.

For each feature the table shows the estimated weight.

Feature statistics, model specific - Linear Regression



- Weights also come with confidence intervals (in this case 95%): a range for the weight estimate that covers the “true” weight with a certain confidence.
- Confidence interval (CI): if we repeated the estimation 100 times with newly sampled data, the CI would include the true weight in 95 out of 100 cases, given that the linear regression model is the correct model for the data.

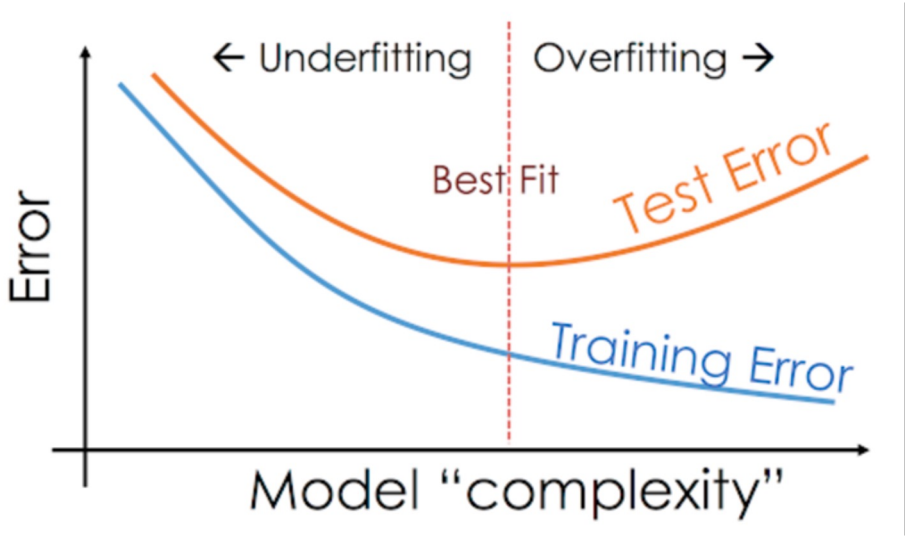
In the context of linear regression: the LASSO

- LASSO (least absolute shrinkage and selection operator) is a popular approach that by design provide sparsity: ie. some of the weights are forced to be equal to zero
- This make feature statistics analysis simpler

L1 penalty / Penalty Term / Regularisation Term

$$RSS_{LASSO}(\beta_i, \beta_0) = \underset{\beta}{argmin}\Big[\sum_{i=1}^{n}(y_i - (\beta_i x_i + \beta_0))^2 + \alpha \sum_{j=1}^{k} |\beta_j|\Big]$$

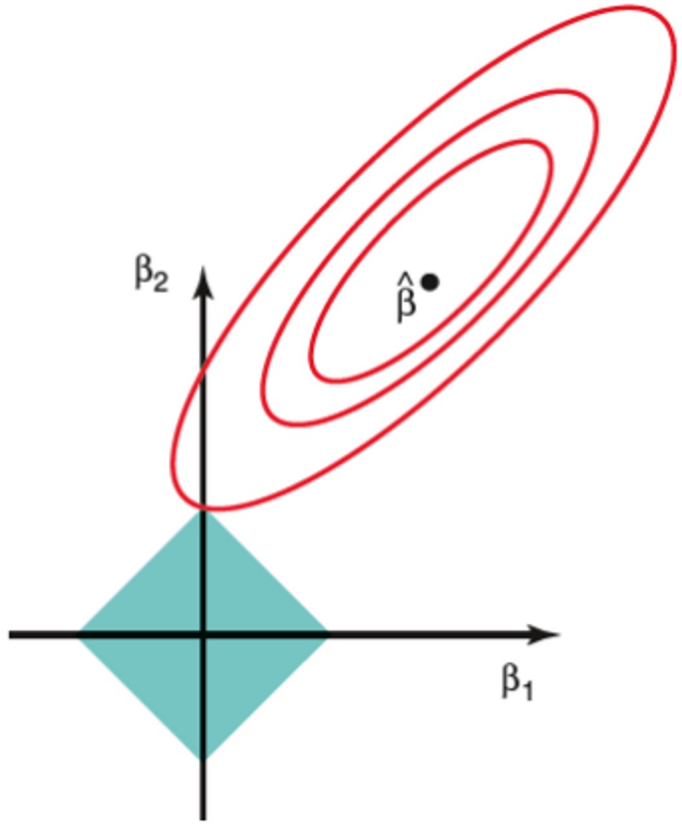
Fit training data well (OLS) Keep parameters small



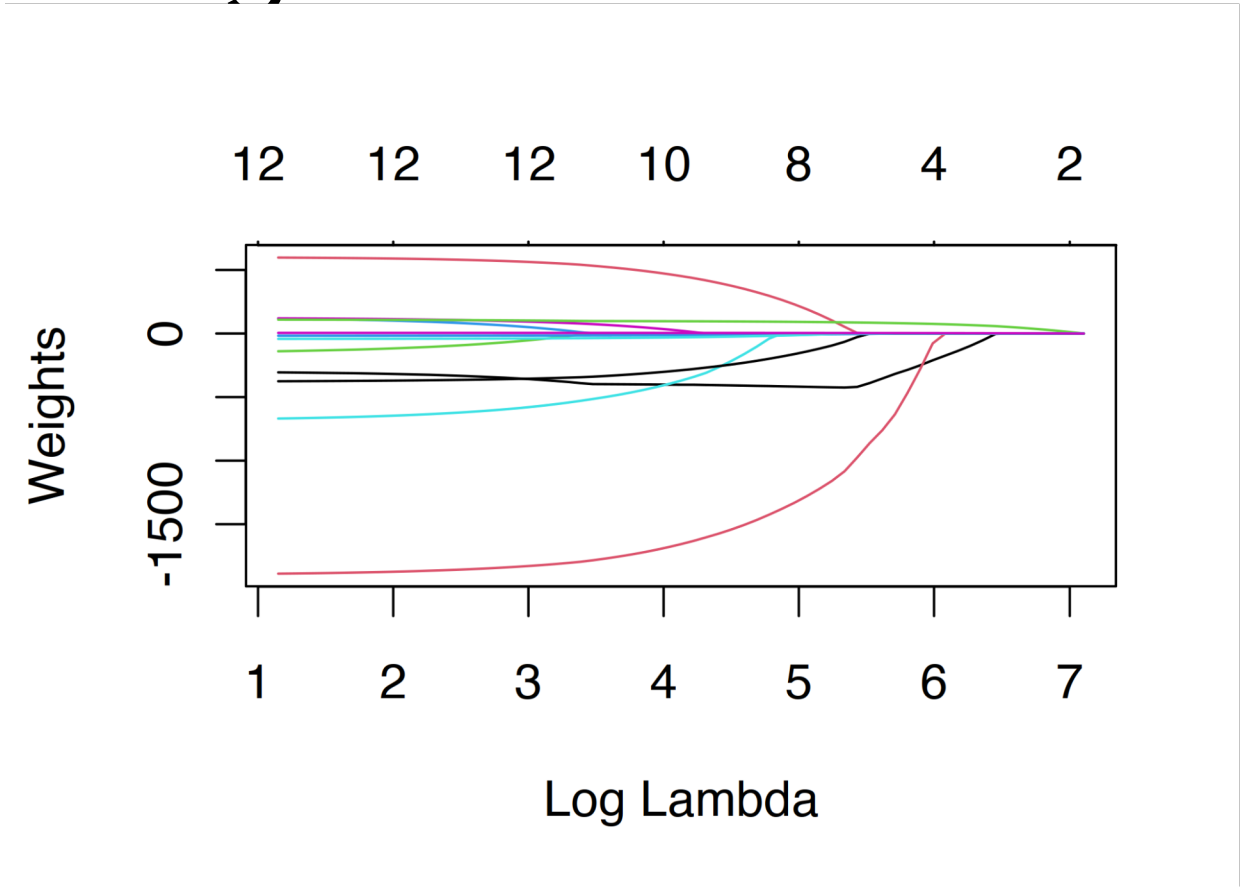
In the context of linear regression: the LASSO

- LASSO (least absolute shrinkage and selection operator) is a popular approach that by design provide sparsity: ie. some of the weights are forced to be equal to zero
- This make feature statistics analysis simpler

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$$



Feature statistics, model specific - Linear Regression



- A decreased number of variables ‘entering’ the model could lead to more intuitive interpretations
- Nevertheless, the are drawbacks, for example the so-called ‘grouping effect’: variables highly correlated to each other can alternatively entered in the model, but that do not means that a variable is not relevant for solving a task

Since we are talking about linear models...

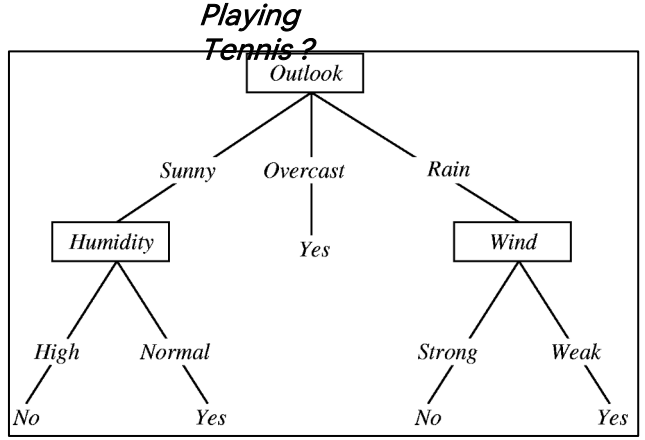
Model	Form	Intelligibility	Accuracy
Linear Model	$y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$	+++	+
Generalized Linear Model	$g(y) = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$	+++	+
Additive Model	$y = f_1(x_1) + ... + f_n(x_n)$	++	++
Generalized Additive Model	$g(y) = f_1(x_1) + ... + f_n(x_n)$	++	++
Full Complexity Model	$y = f(x_1, ..., x_n)$	+	+++

Lou, Caruana, Gehrke, Hooker, Accurate Intelligeble Models with Pairwise Interactions. KDD2013

Sometimes the trade-off ‘Accuracy vs Interpretability’ is clearly there...

Feature statistics, Model specific - Tree-based methods

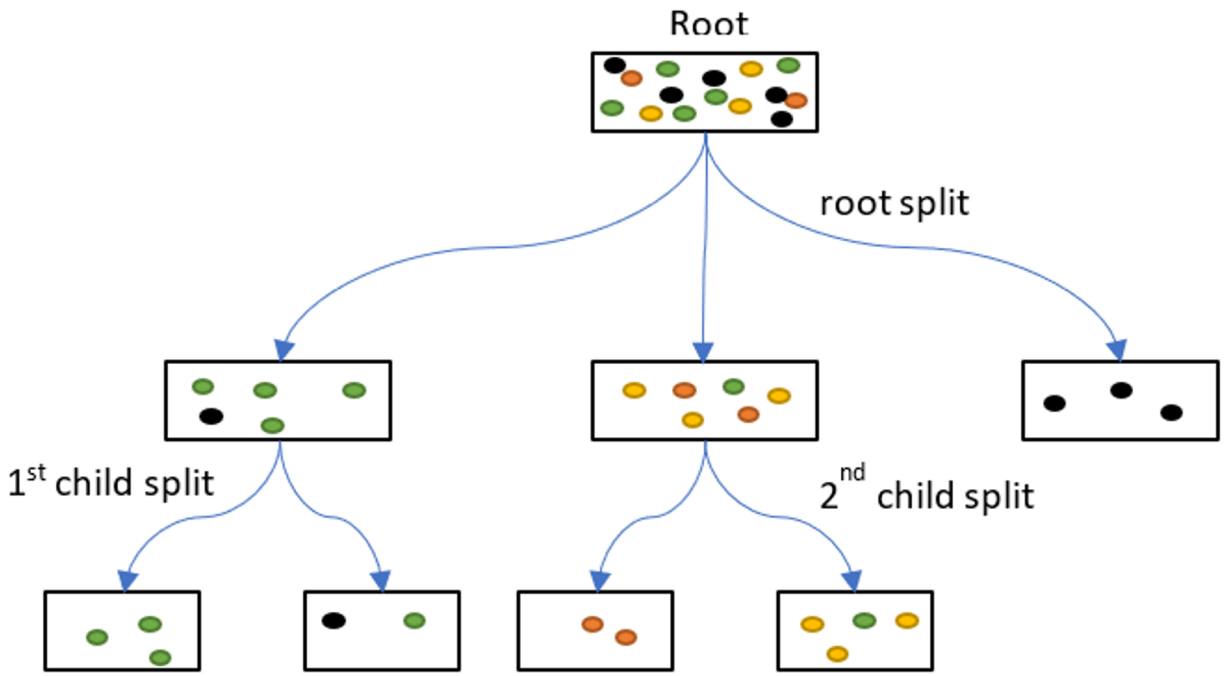
- Decision Trees are ‘classical’ solutions to supervised tasks
- The classification is done by following a tree-structure:
  - each interior node is a input variable (and there are edges to children for each possible value of that variable)
  - each leaf is a class
- Advantages
  - ‘Easily interpretable’
  - They require no data normalization
  - The outcome computation is almost immediate



From: ‘Machine Learning’, Mitchell 1997

Feature statistics, Model specific - Tree-based methods

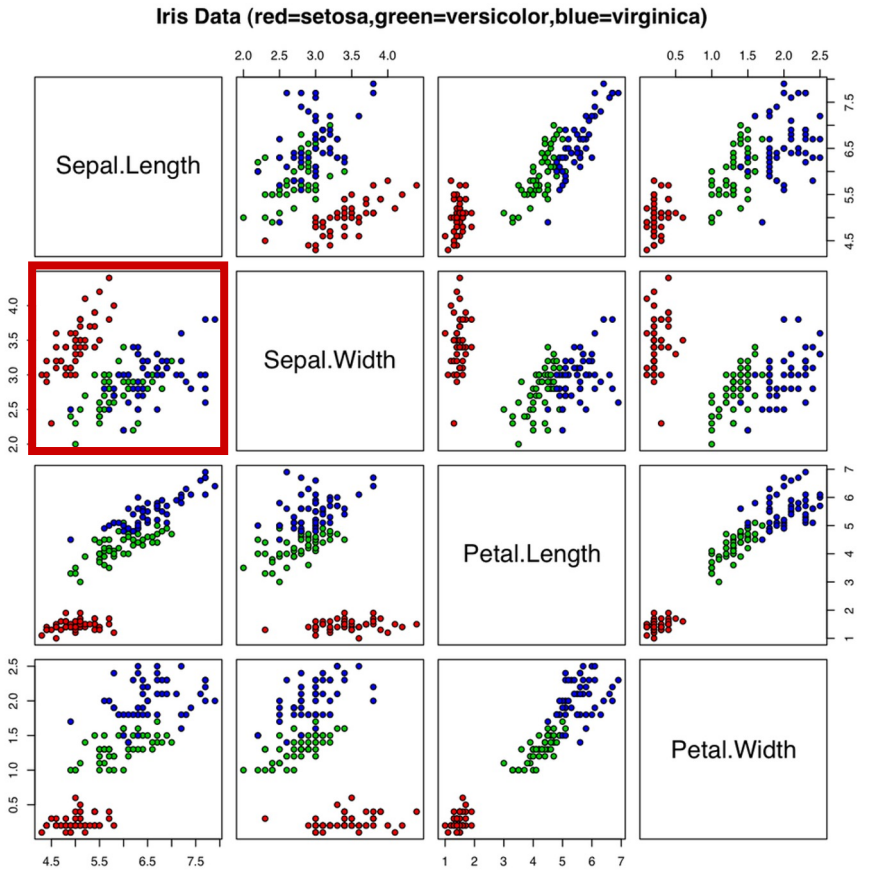
- DTs are constructed with *top-down* approaches: at each step of the algorithm is to choose a variable that 'best' splits the set of observations (recursive partitioning)
- Many criteria:
  - entropy and information gain
  - Gini impurity / Mean Decrease in impurity
  - Variance reduction



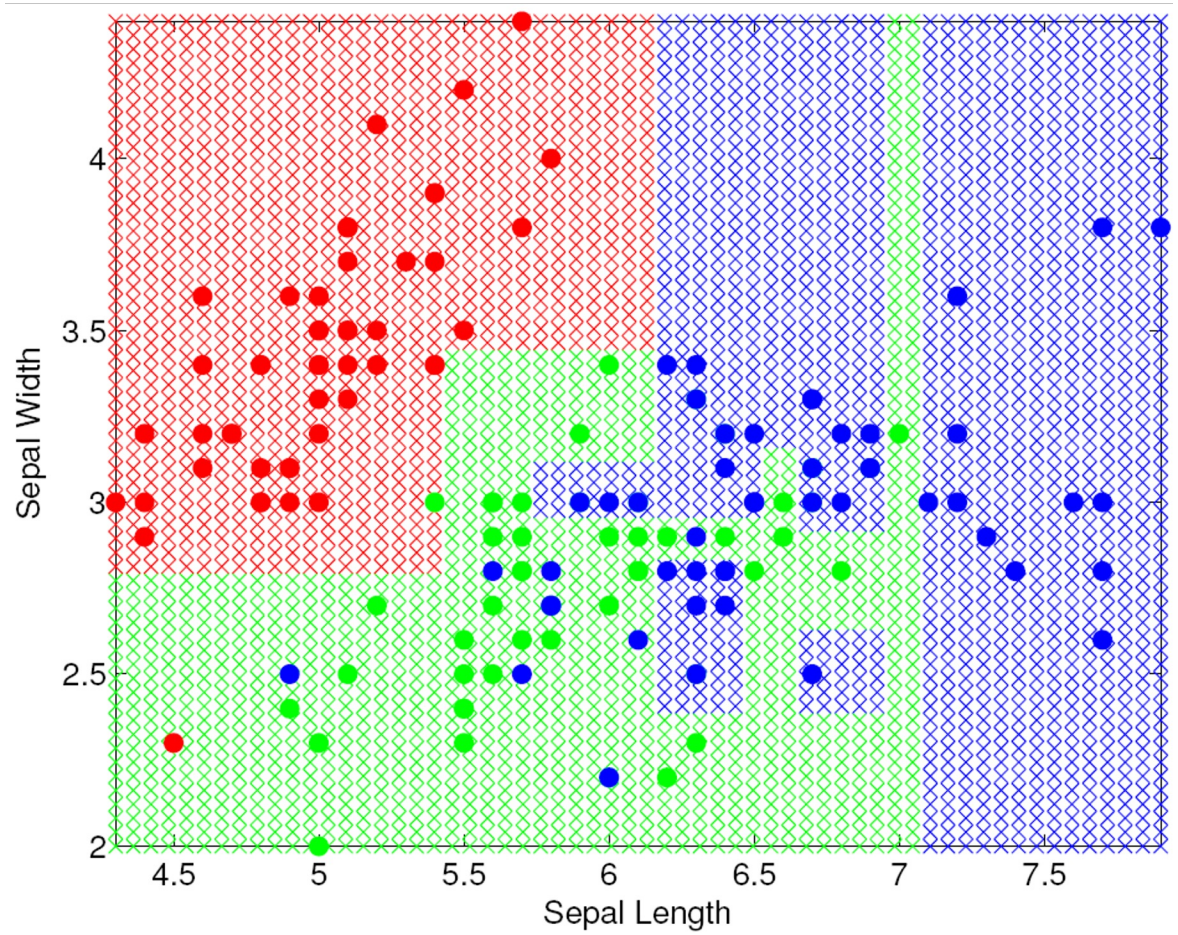
Feature statistics, Model specific - Tree-based methods

Example: ‘Iris Classification’ dataset, Ronald Fisher (1936) - UCI ML Repository

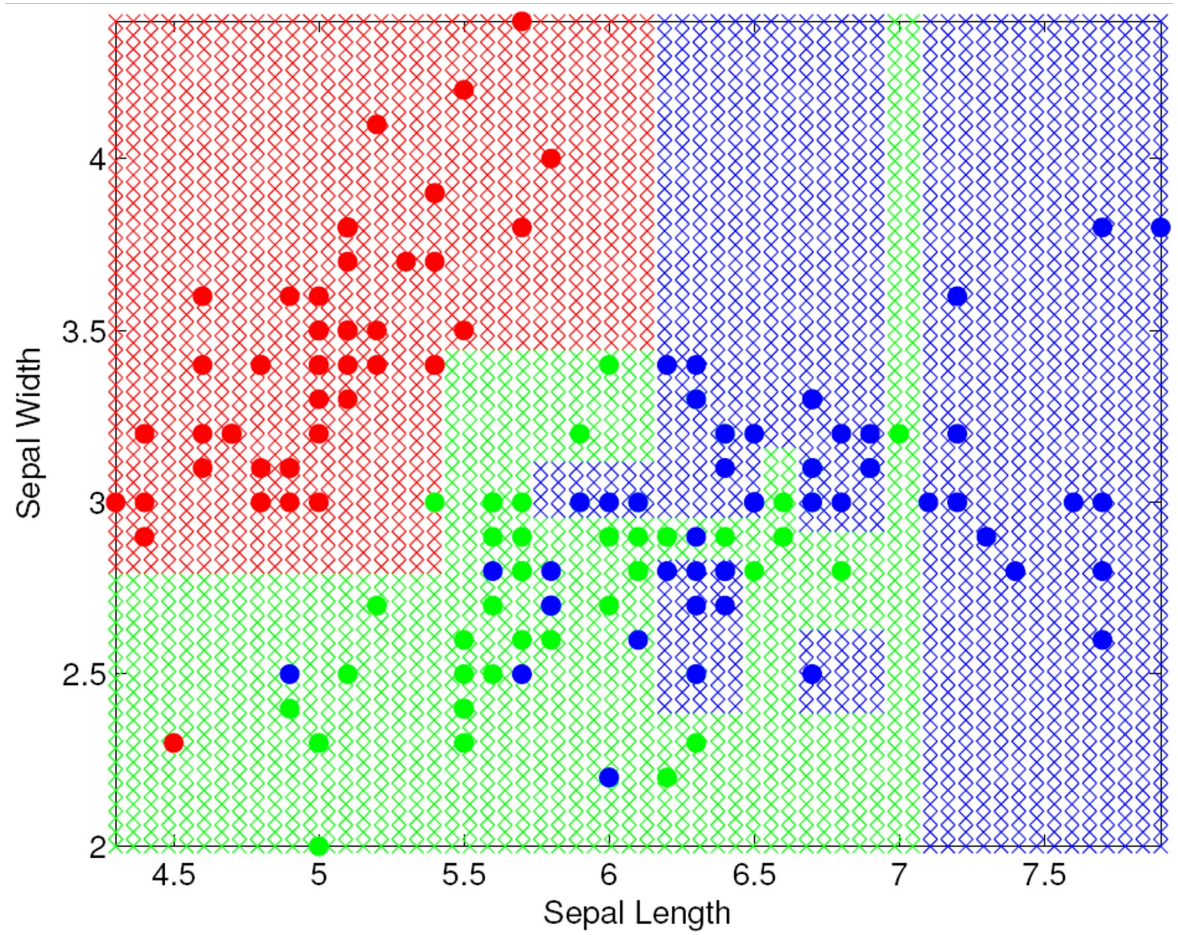
L = 3 classes problem: classify Setosa, Versicolour and Virginica iris from data containing sepal and petal width and length – n = 150 samples, p = 4 variables

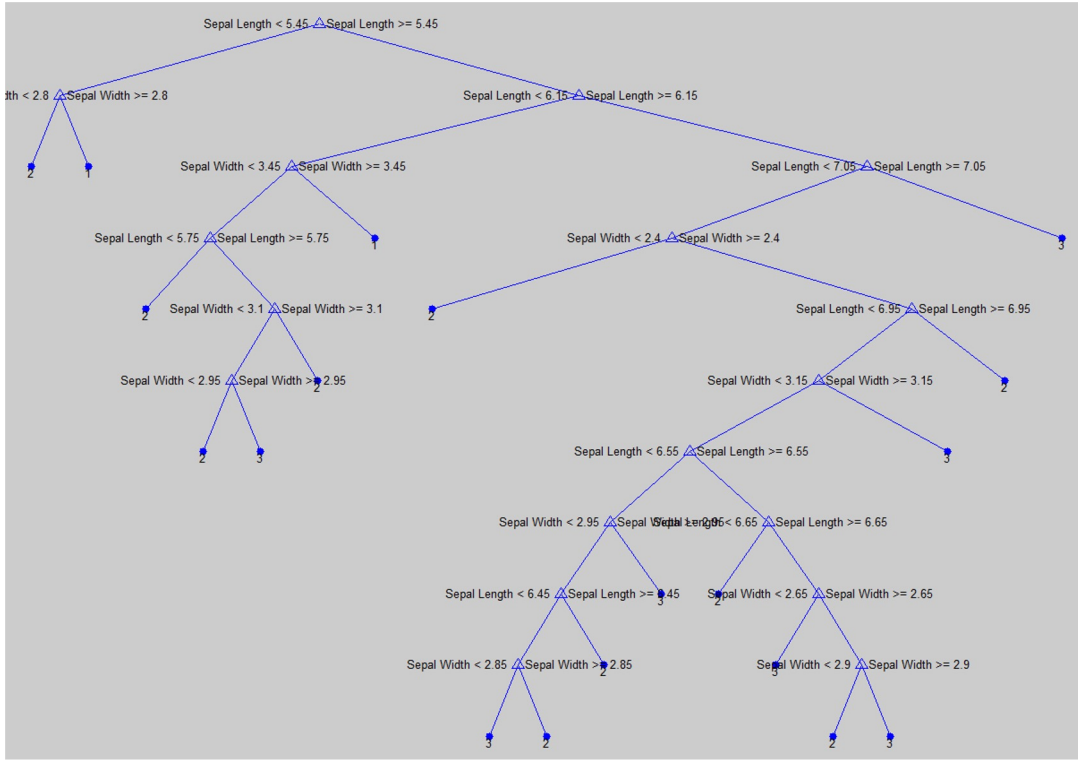


Feature statistics, Model specific - Tree-based methods



Feature statistics, Model specific - Tree-based methods

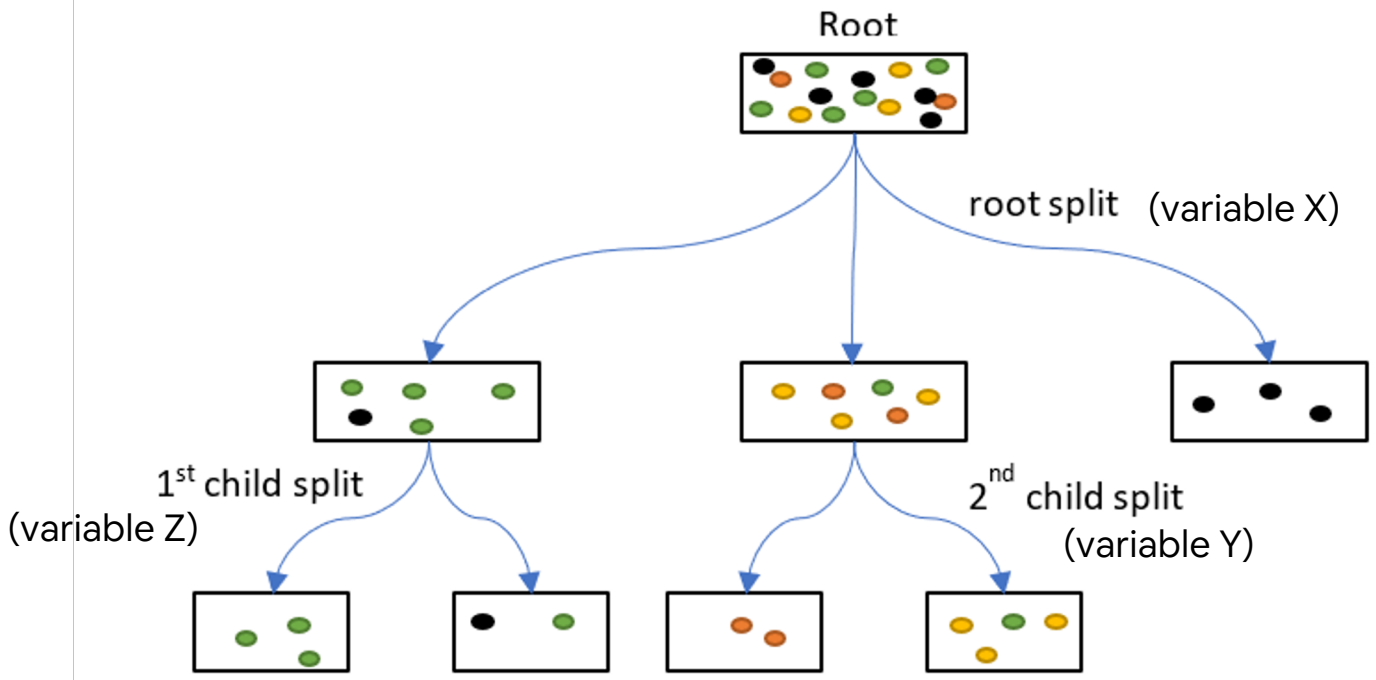




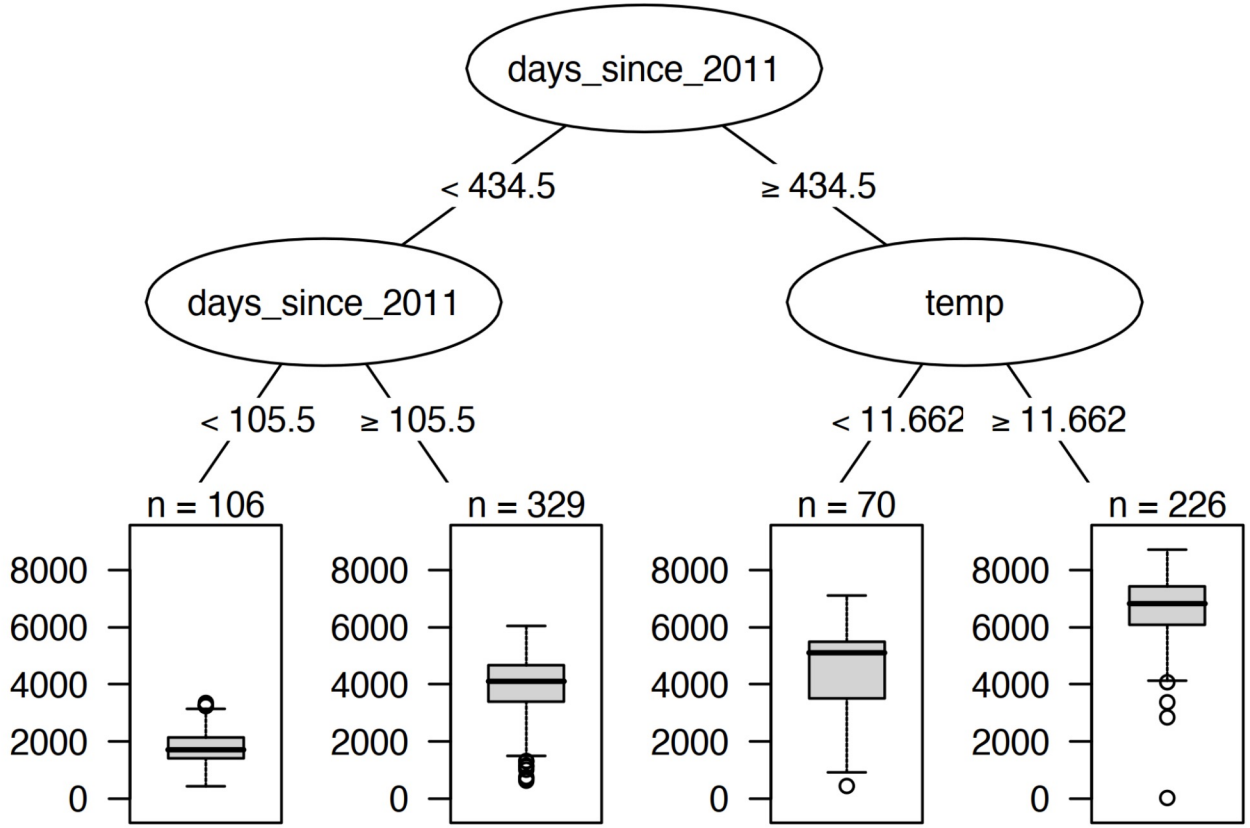
19 splits!

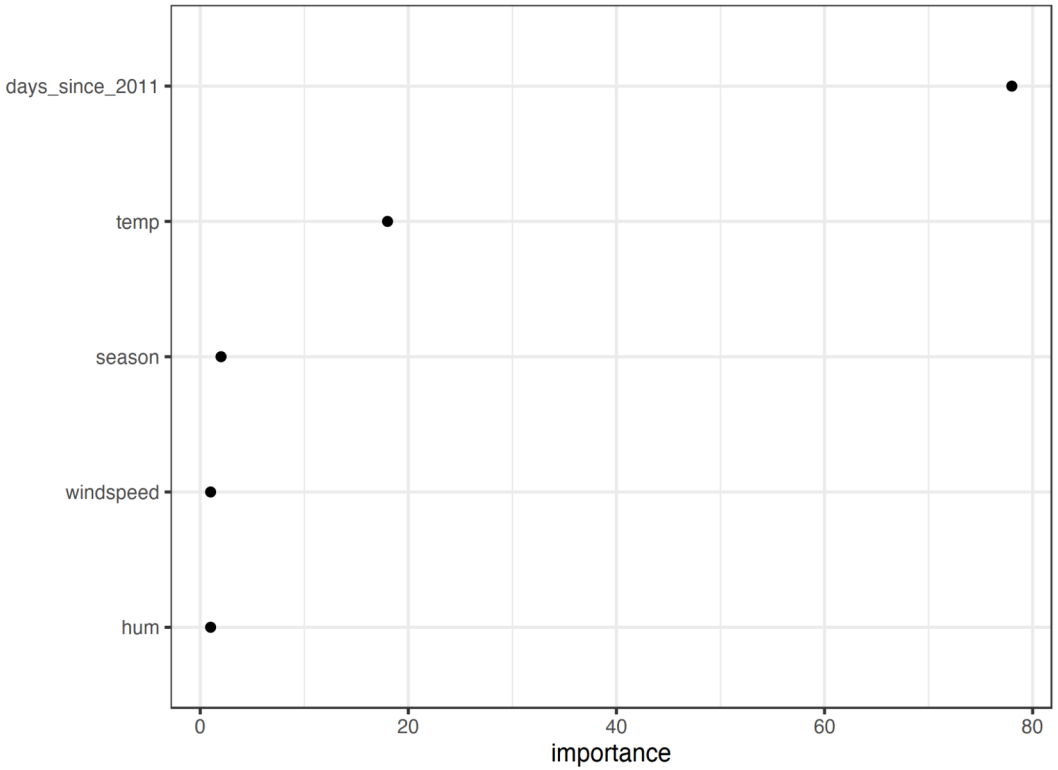
Feature statistics, Model specific - Tree-based methods

- Also in this case we would like to provide feature statistics summary: what are the most important variables?
- **Gini Importance** or **Mean Decrease in Impurity (MDI)** calculates each feature importance as the sum over the number of splits that include the feature, proportionally to the number of samples it splits



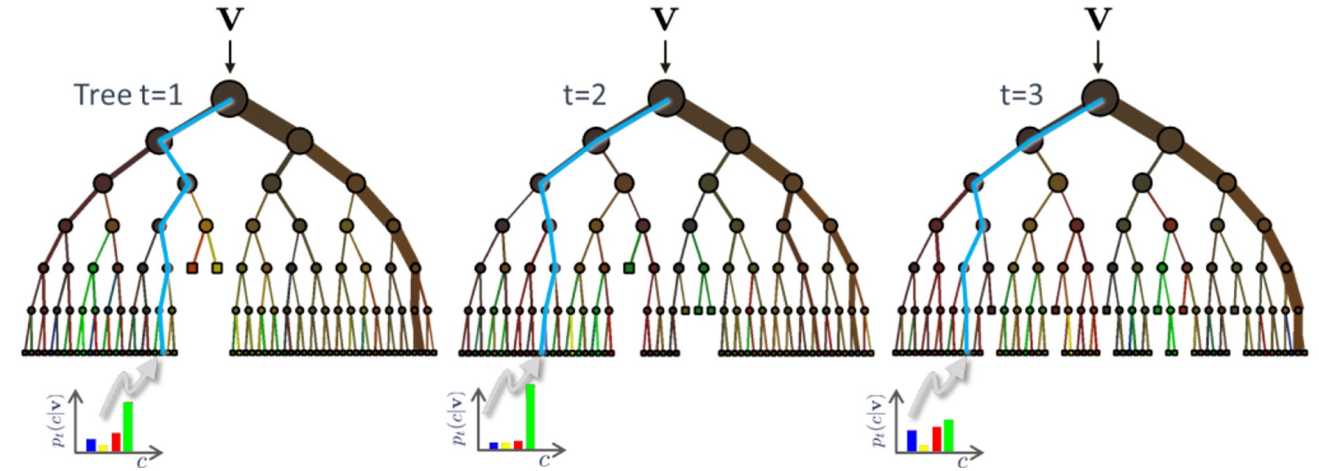
Feature statistics, Model specific - Tree-based methods





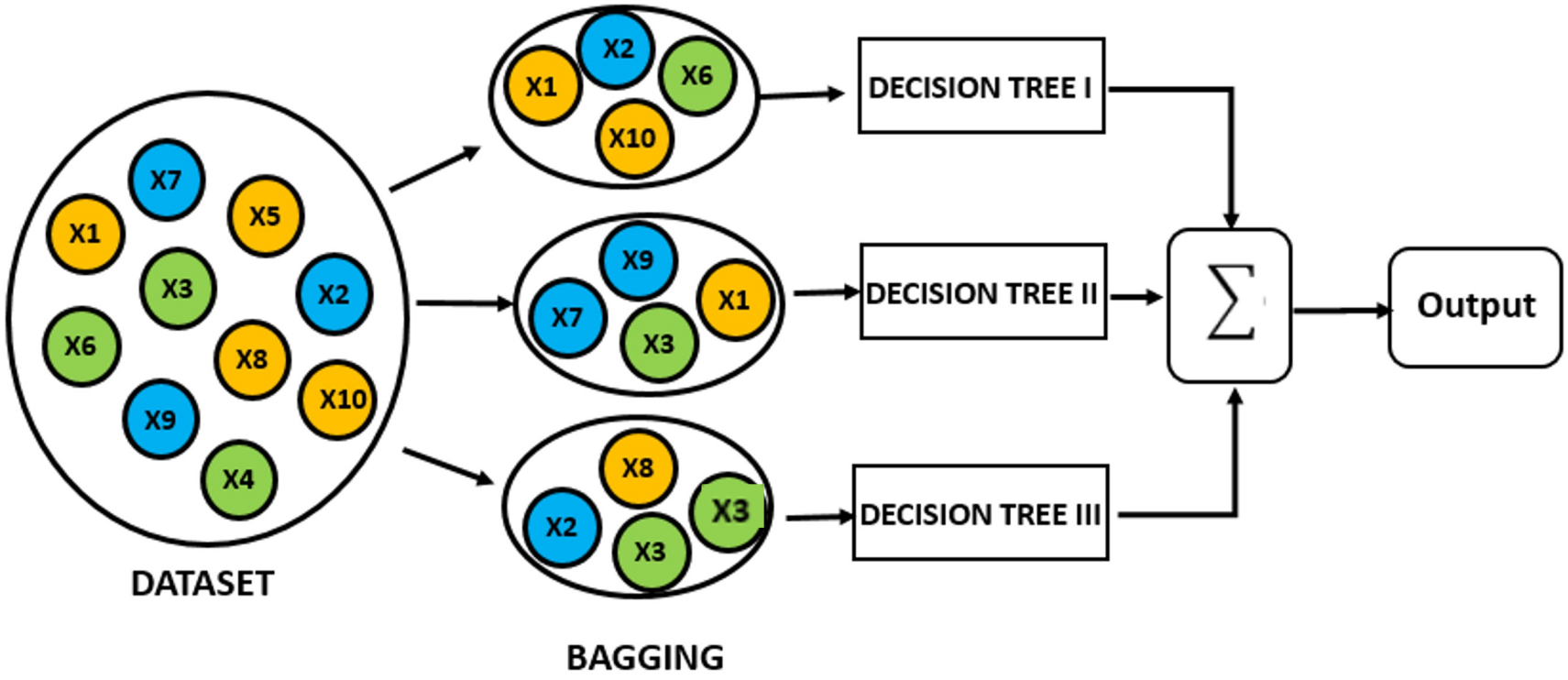
Feature statistics, Model specific - Tree-based methods

- Decision Trees are high variance models: they are really sensible to small fluctuations in the training set and they tend to overfit
- The variance can be reduced with Random Forest: one of the most powerful supervised approach, adopted in several areas
- RF are an *ensemble approach*: it is a set of several DTs, where the classification is done by a majority vote / regression
- Ensemble tree-based methods (RF, CatBoost, XGBoost) are among the most accurate ML models

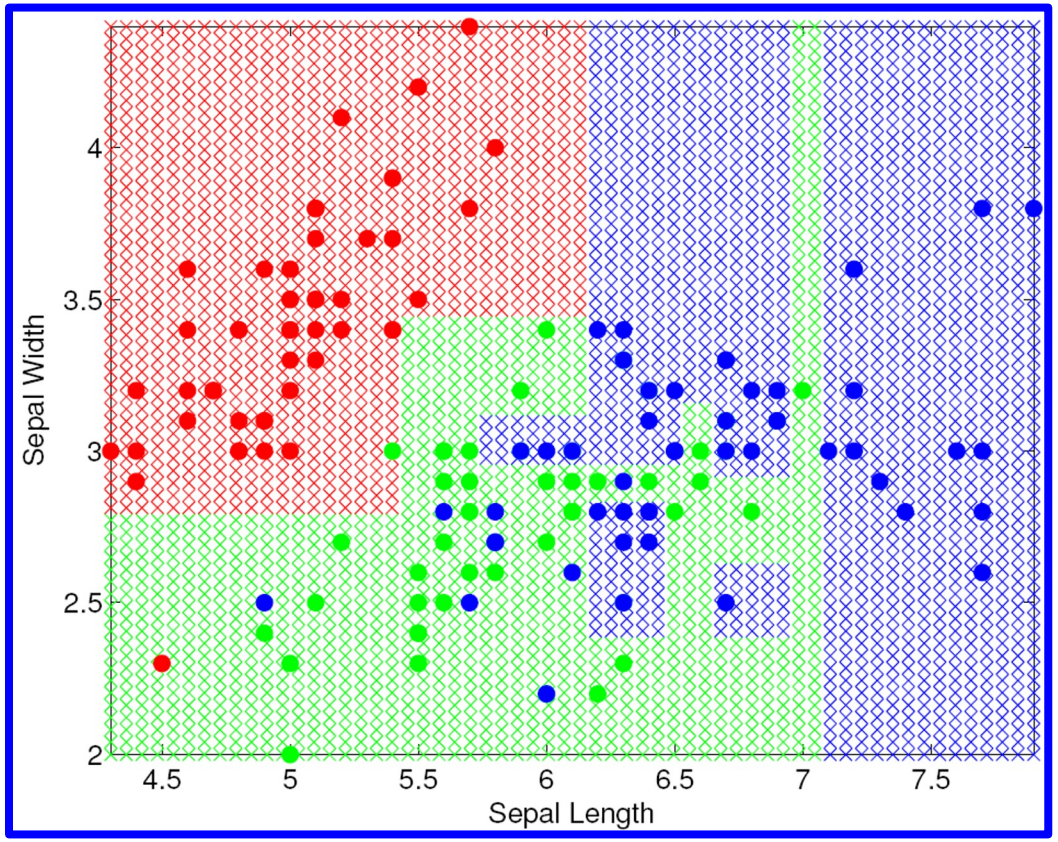


Feature statistics, Model specific - Tree-based methods

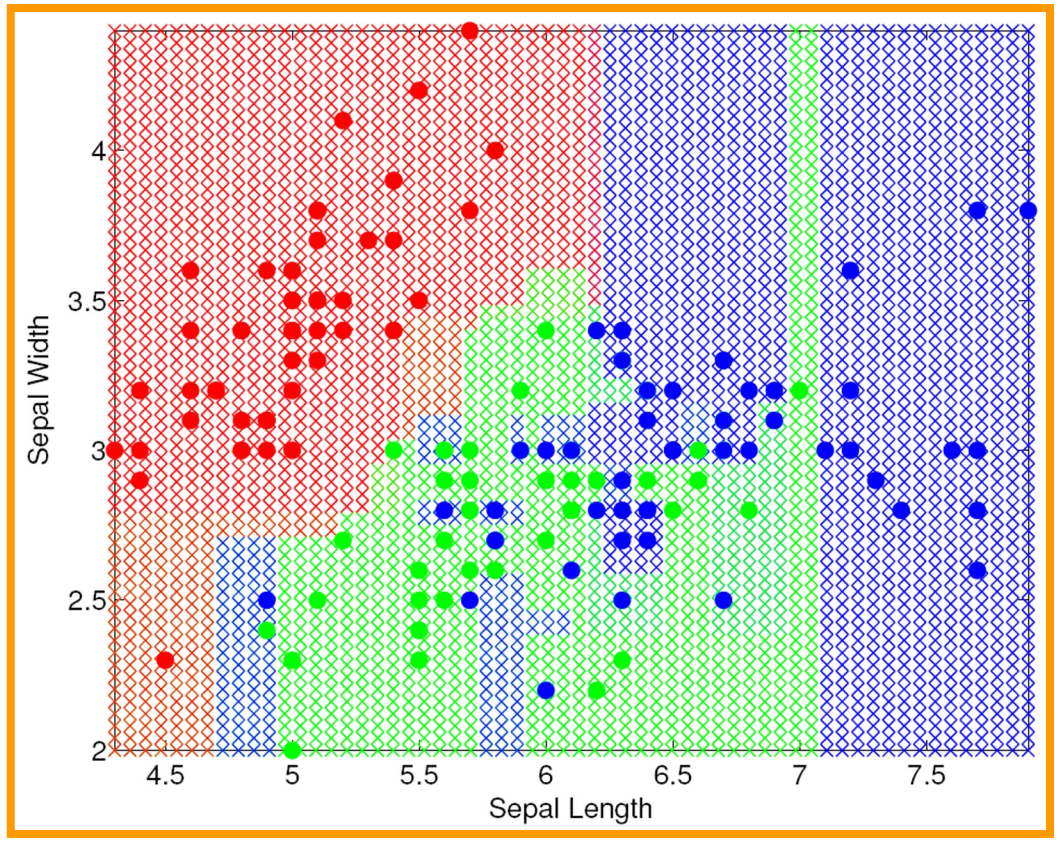
- Random Forest are generally based on bagging (bootstrap aggregating): the creation of several dataset by uniformly sampling with replacement from the original dataset



Feature statistics, Model specific - Tree-based methods



Decision tree



Random Forest

Feature statistics, Model specific - Tree-based methods

Even though RF consists of a collection of Decision Trees (which are recognized as interpretable models), its interpretation isn’t as trivial as it may seem

The most widely used feature importance measure in this context is again the Mean Decrease Impurity (MDI): think about averaging MDI of the individual Decision Trees

Feature statistics, Model specific - Tree-based methods

Even though RF consists of a collection of Decision Trees (which are recognized as interpretable models), its interpretation isn't as trivial as it may seem

The most widely used feature importance measure in this context is again the Mean Decrease Impurity (MDI): think about averaging MDI of the individual Decision Trees

REMARK ⚠

Problem: MDI measure suffers from so-called "*feature selection bias*", i.e. it may erroneously assign high MDI values to features that are not highly correlated to the output

Feature statistics, Model specific - Tree-based methods

Ev sists of a collection of Decision Trees (which are re t as trivial as it ma

REMARK ⚠

Solution: "*A Debiased MDI Feature Importance Measure for Random Forests*", by Li et al.

Th context is again the Mean Decrease Impurity (MDI): think about averaging MDI of the individual Decision Trees

REMARK ⚠

Problem: MDI measure suffers from so-called "*feature selection bias*", i.e. it may erroneously assign high MDI values to features that are not highly correlated to the output

Feature statistics, Model specific - Tree-based methods

So, we have a robust model-specific method to compute feature importance for RF... are we done?

Not really... in several applications we may need to detect high-order interactions between features!

Feature statistics, Model specific - Tree-based methods

So, we have a robust model-specific method to compute feature importance for RF... are we done?

Not really... in several applications we may need to detect high-order interactions between features!

REMARK ⚠

Solution: "*iterative Random Forests to discover predictive and stable high-order interactions*", by Basu et al. (THIS IS A 'NEW' INTERPRETABLE-ORIENTED MODEL)

Feature statistics, Model specific - Tree-based methods

For other ensemble tree-based methods, similar approaches can be used.

The are other approaches, for example:

Boruta implements a different feature selection algorithm. It randomly permutes variables like Permutation Importance (next slides) does, but performs on all variables at the same time and concatenates the shuffled features with the original ones. The concatenated result is used to fit the model.

Miron B. Kursa, Witold R. Rudnicki (2010). Feature Selection with the Boruta Package.

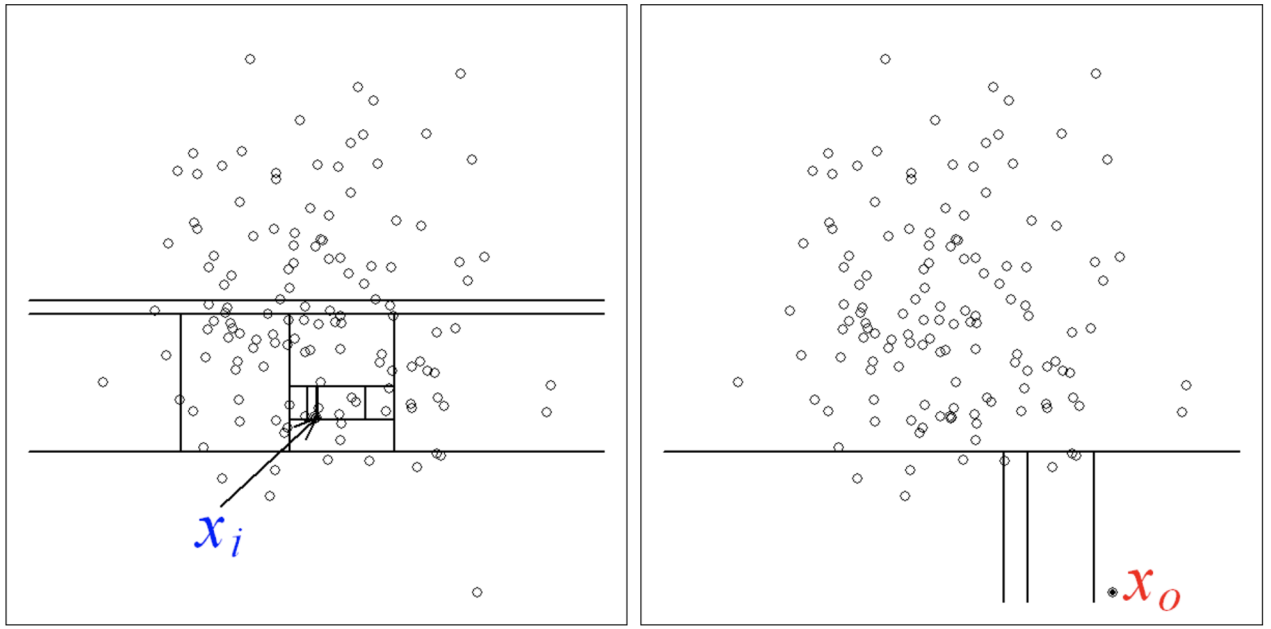
Journal of Statistical Software, 36(11) , p. 1–13.

Tree-based Approaches for Anomaly Detection

There are tree-based approaches for anomaly/outlier detection that are quite powerful and widely adopted

Isolation Forest is based on recursive partitioning

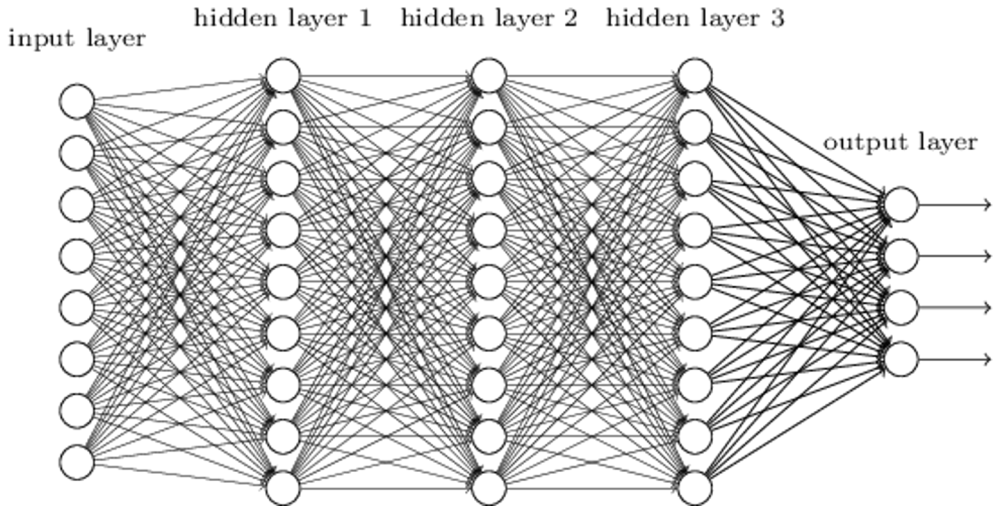
Some fellows developed methods for providing interpretability (feature summary, model-specific) in Isolation Forests



M. Carletti, M. Terzi, G.A. Susto. **Interpretable Anomaly Detection with DIFFI: Depth-based Feature Importance for the Isolation Forest**
https://arxiv.org/abs/2007.11117

Model-specific methods for DNNs

Deep Neural Networks (DNNs) are arguably the hardest ML models to be interpreted by human beings



Despite their amazing performance on a wide variety of applications (e.g. Computer Vision, Natural Language Processing, ...), interpretation of DNNs and produced outputs is still an open research problem

In this lecture we just give a brief overview of the research works in this field and refer the curious readers to the work of Gilpin et al. for further details and analyses

Model-specific methods for DNNs

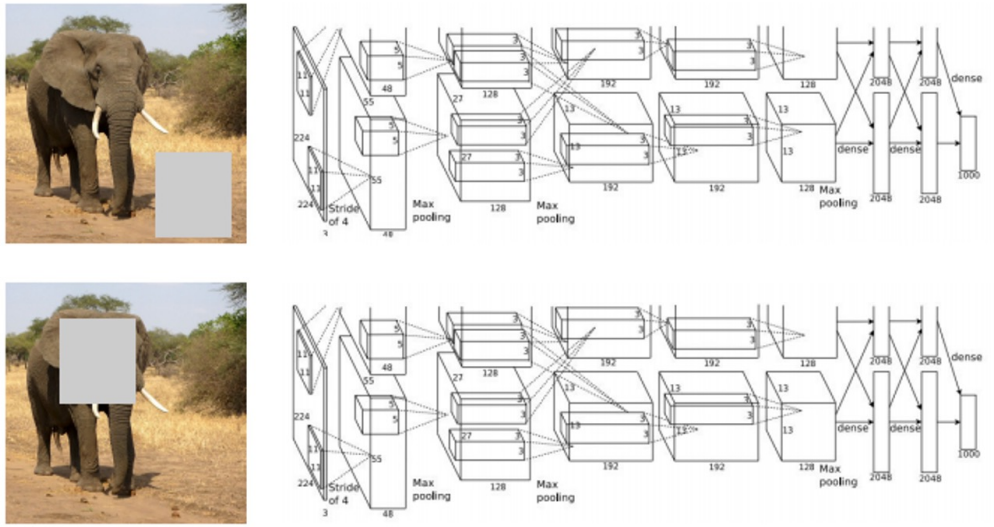
As described in Gilpin et al., interpretability methods for DNNs can be roughly divided into three main categories:

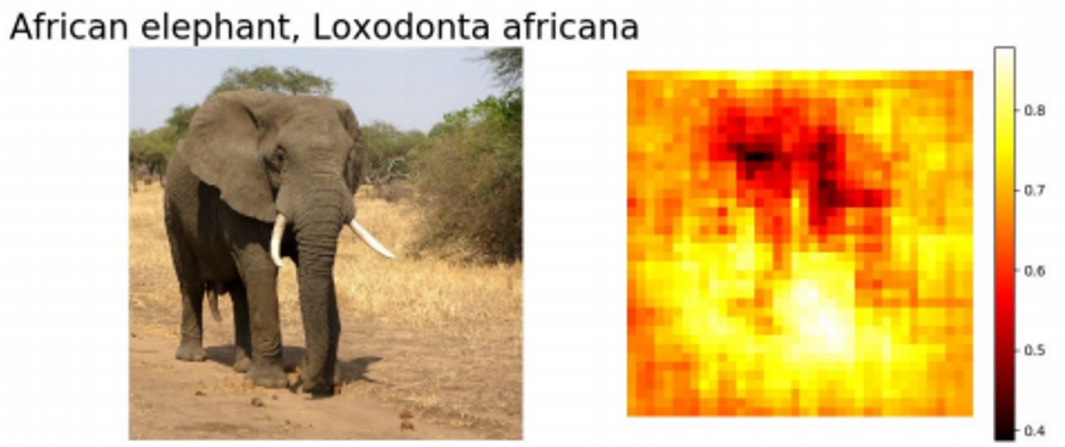
- methods focused on the explanation of the processing of the data by a DNNs
- methods focused on the explanation of the representations generated within the DNNs
- methods focused on the design of architectures that facilitate interpretations of the network's behavior

Model-specific methods for DNNs

Explanation of the processing

Saliency maps:





Zeiler and Fergus, “Visualizing and Understanding Convolutional Networks”, ECCV 2014

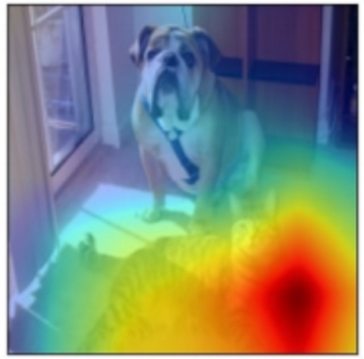
Model-specific methods for DNNs

Explanation of the processing

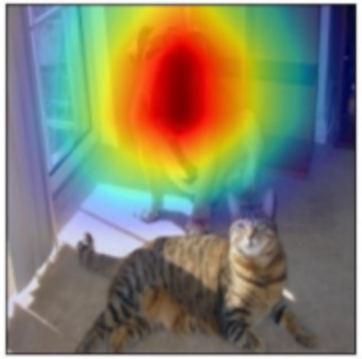
Some examples:

- surrogate models specifically tailored for DNNs (such as DeepRED, ANN-DT)
- saliency mapping (such as DeepLIFT, Grad-CAM)

(a) Original Image



(f) ResNet Grad-CAM 'Cat'



(l)ResNet Grad-CAM 'Dog'

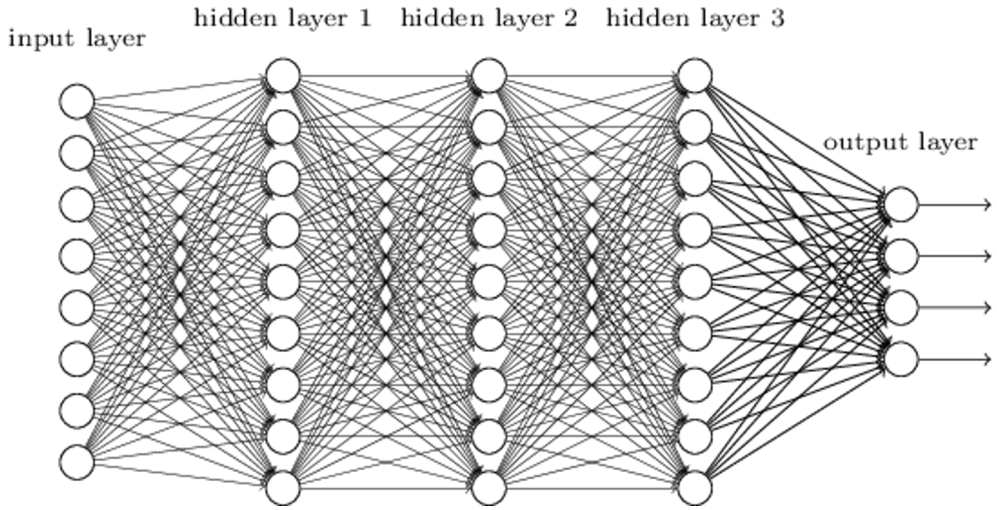
From "*Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization*", Selvaraju et al.

Model-specific methods for DNNs

Explanation of the representations

Some examples:

- role of layers - example Razavian et al.
- role of individual units, both units or filters (like in CNN) - example Network dissection)
- role of other representation vectors - example Concept Activation Vectors

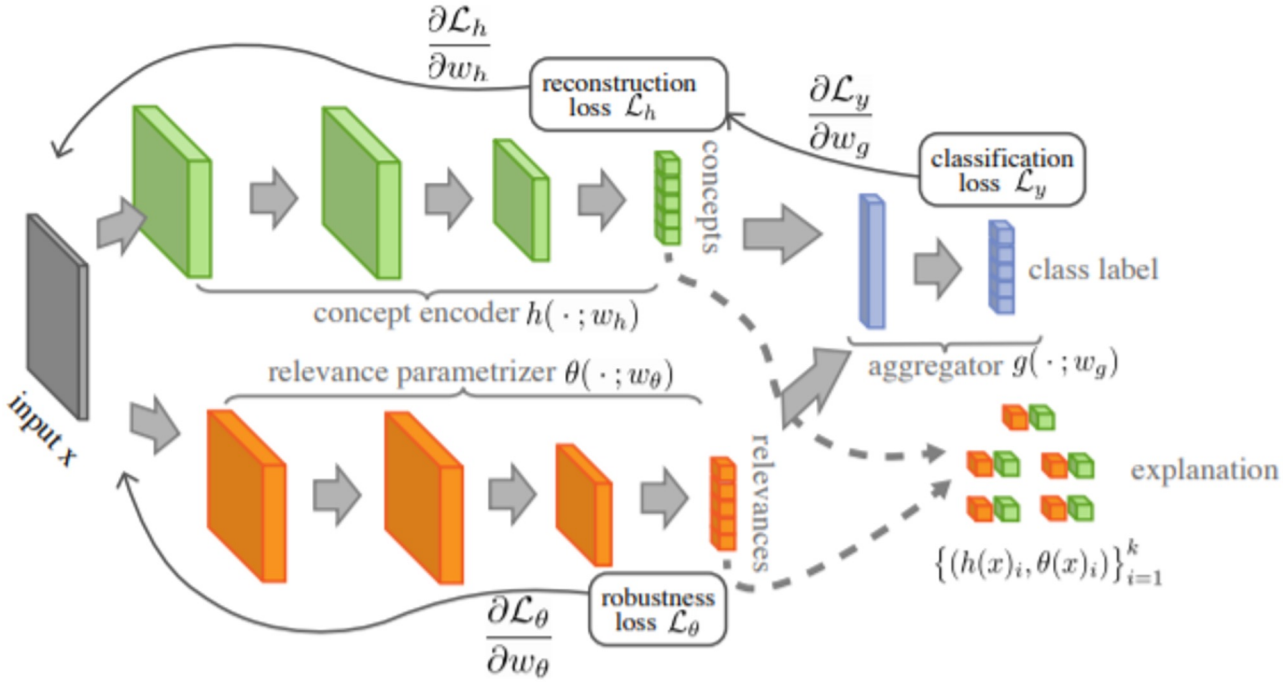


Model-specific methods for DNNs

Explanation-producing systems

Some examples:

- attention networks (such as Xiao et al.)
- Self-Explaining Neural Networks (Alvarez-Melis et al.)



From "*Towards Robust Interpretability with Self-Explaining Neural Networks*", Alvarez-Melis et al.

Thank you!

Gian Antonio Susto