Explainable Artificial Intelligence

01 – Introduction

MSc in Artificial Intelligence

MSc in Computer Engineering

Marco Lippi

marco.lippi@unifi.it

UNIVERSITÀ
DEGLI STUDI
FIRENZE

UNIVERSITÀ DEGLI STUDI FIRENZE

Table of Contents

Credits

These slides are largely an adaptation of (a lot of!) existing material, mostly:

- Explainable Artificial Intelligence (F. Lecue et al., 2023)
- Explainable and Trustworthy Artificial Intelligence (Politecnico di Torino, 2023)
- Explainable AI (Harvard University, 2023)

What is XAI?

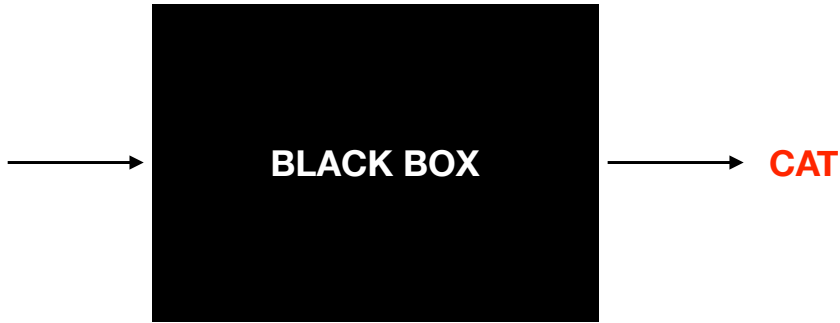
1 General concepts

[Image from Wikipedia]

What is XAI?

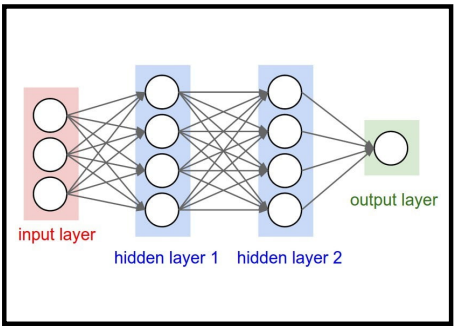
Explainable AI (XAI) explores and investigates **methods** to produce or complement AI models to make **accessible** and **interpretable** the internal logic and the outcome of algorithms, making such process **understandable by humans**

Black-box models



Black-box models

1 General concepts

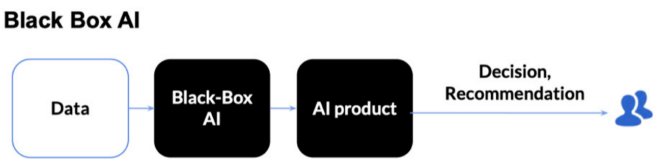


CAT

Why do we need XAI?

1 General concepts

1. To understand AI decisions



Confusion with Today's AI Black Box

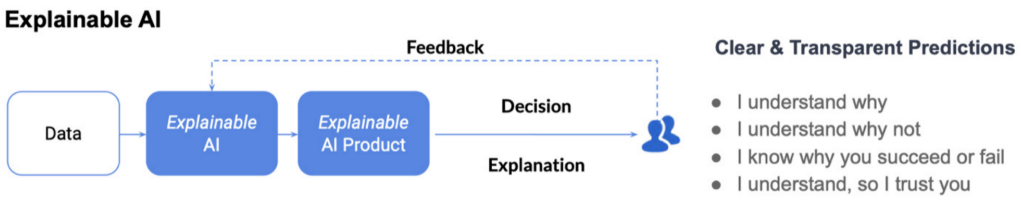
- Why did you do that?
- Why did you not do that?
- When do you succeed or fail?
- How do I correct an error?

[Image from Marc Plantevit]

Why do we need XAI?

1 General concepts

2. To improve AI models

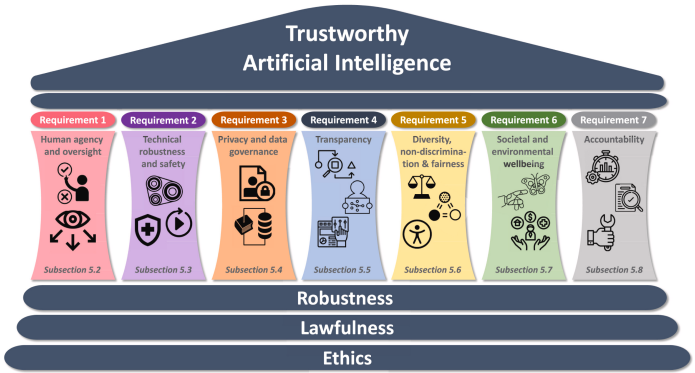


[Image from Marc Plantevit]

Why do we need XAI?

1 General concepts

3. To build reliable and trustworthy AI systems



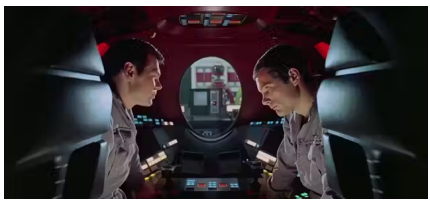
[Image from Diaz-Rodriguez et al., 2023]

The alignment problem

1 General concepts

"If we use, to achieve our purposes, a mechanical agency with whose operation we cannot interfere effectively… We had better be quite sure that the purpose put into the machine is the purpose which we really desire."

Norbert Wiener, 1960



[Source: 2001: A space odyssey (1968)]

The alignment problem

1 General concepts

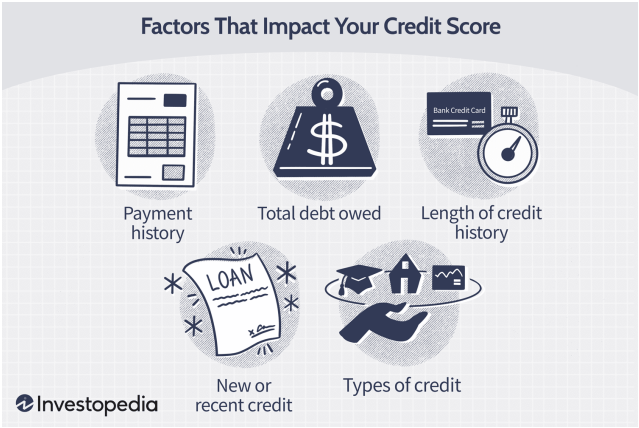
The Three Laws of Robotics (Isaac Asimov, 1942)

1. The First Law: A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. The Second Law: A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. The Third Law: A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

[Source: Io, robot (2004)]

Risks and perils in the (mis)use of AI

1 General concepts



Ethics and Justice, Machine Learning

How Flawed Data Aggravates Inequality in Credit

AI offers new tools for calculating credit risk. But it can be tripped up by noisy data, leading to disadvantages for low-income and minority borrowers.

Aug 6, 2021 | Edmund L. Andrews

[Image from HAI Stanford]

Risks and perils in the (mis)use of AI

1 General concepts

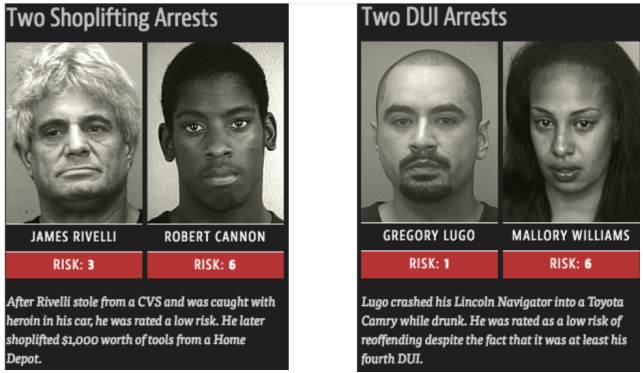


Image source: Propublica

Risks and perils in the (mis)use of AI

Imagine that an AI system is designed to predict the **risk of a patient** affected by pneumonia. If we learn a rule like the following one...

$$\text{HasAsthma}(x) \Rightarrow \text{LowerRisk}(x)$$

...one could think that patients with a history of asthma have **lower risk** of dying from pneumonia. This is counterintuitive, but **reflected** in the data!

- Asthmatic patients with pneumonia typically admitted directly to Intensive Care Unit
- This is the reason of a lower risk of complications!
- Trained diagnostic models would incorrectly **learn that asthma lowers risk**
- Asthmatic patients are actually much higher risk (if not hospitalized)
- Beware of trusting accuracy on **retrospective studies** only!

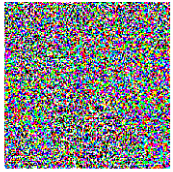
Black-box models can be fooled

$\boldsymbol{x}$

$y =$ “panda”

w/ 57.7% confidence

$+ .007 \times$



$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

“nematode”

w/ 8.2% confidence

$=$

$\boldsymbol{x} + \epsilon \cdot \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

“gibbon”

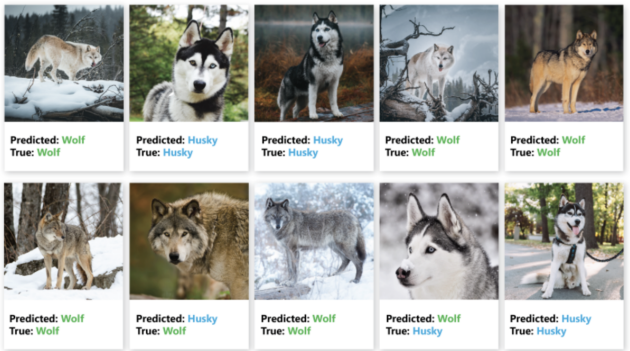
w/ 99.3 % confidence

Image source: Goodfellow et al., 2016

Bias in data

1 General concepts

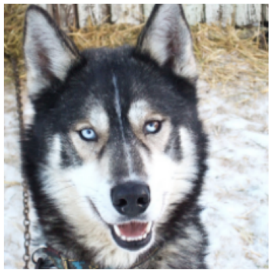
Ribeiro et al., “Why Should I Trust You?” Explaining the Predictions of Any Classifier



Bias in data

1 General concepts

Ribeiro et al., “Why Should I Trust You?” Explaining the Predictions of Any Classifier



(a) Husky classified as wolf

(b) Explanation

Bias in data

1 General concepts

Wedding Ceremony

Wedding Ceremony

People

[Source: Marya Raifer]

Bias in data

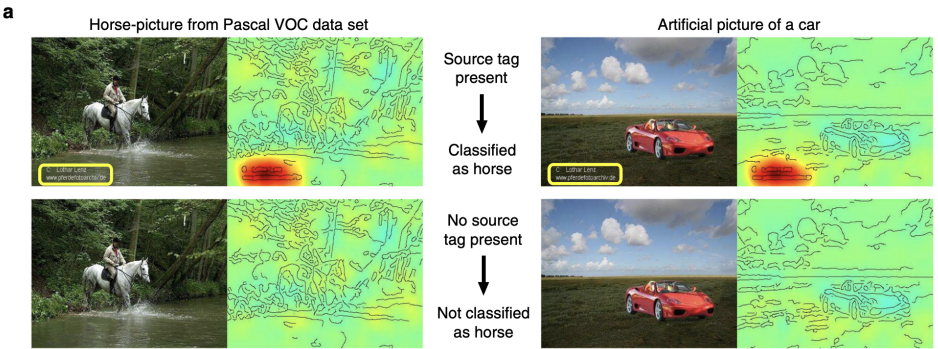
1 General concepts

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi[1], **Kai-Wei Chang**[2], **James Zou**[2], **Venkatesh Saligrama**[1,2], **Adam Kalai**[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

Bias in data

1 General concepts



[Source: Lapuschkin et al., 2023]

General Data Protection Regulation

1 General concepts

Recital 71
EU GDPR

(71) The data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention.

Such processing includes 'profiling' that consists of any form of automated processing of personal data evaluating the personal aspects relating to a natural person, in particular to analyse or predict aspects concerning the data subject's performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, where it produces legal effects concerning him or her or similarly significantly affects him or her.

However, decision-making based on such processing, including profiling, should be allowed where expressly authorised by Union or Member State law to which the controller is subject, including for fraud and tax-evasion monitoring and prevention purposes conducted in accordance with the regulations, standards and recommendations of Union institutions or national oversight bodies and to ensure the security and reliability of a service provided by the controller, or necessary for the entering or performance of a contract between the data subject and a controller, or when the data subject has given his or her explicit consent.

In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision.

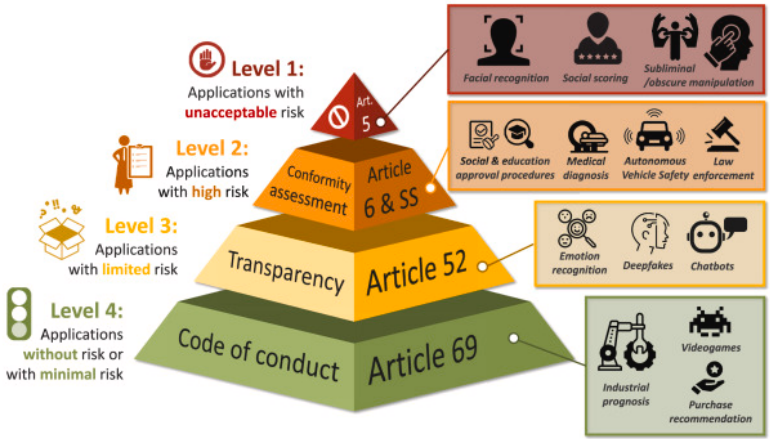
Article 13 of the EU AI Act

1. High-risk AI systems shall be designed and developed in such a way as to ensure that their operation is sufficiently transparent to enable deployers to interpret a system's output and use it appropriately. An appropriate type and degree of transparency shall be ensured with a view to achieving compliance with the relevant obligations of the provider and deployer set out in Section 3.

2. High-risk AI systems shall be accompanied by instructions for use in an appropriate digital format or otherwise that include concise, complete, correct and clear information that is relevant, accessible and comprehensible to deployers.

AI Act

1 General concepts



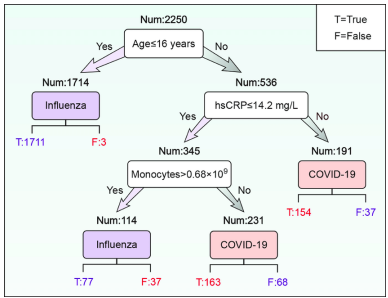
[Image from Diaz-Rodriguez et al., 2023]

Interpretability vs. Explainability

1 General concepts

Some models do not need to be **explained**: they are **inherently interpretable** by design

- Decision trees
- Linear models
- Rule-based systems
- Inductive logic programming
- …

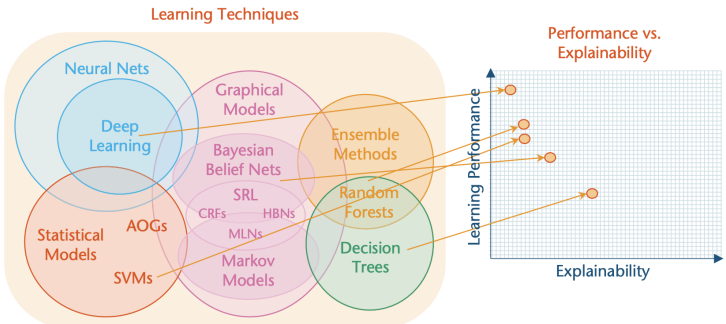


[Image from Zhou et al., 2021]

Interpretability vs. Accuracy

1 General concepts

Some argue that there is a **trade-off** between interpretability and accuracy
Is it **really** the case?



[Figure from DARPA]

Interpretability vs. Accuracy

Sometimes interpretable models are **also** accurate!

- In that case, they should be preferred!
- Use a black-box model only if there is **need** to do it...
- ...or if it is the only available model (e.g., a proprietary system)
- Otherwise, try to build **post-hoc explanations**

A taxonomy of approaches

Interpretable-by-design

- Decision trees
- Rule-based systems (lists, rules)
- Generalized linear models
- Counterfactual reasoning
- Inductive logic programming
- ...

A taxonomy of approaches

Black-box explanations

- **Model explanation**: global explanation of an opaque AI system through an interpretable and transparent model that fully captures its logic
- **Model inspection**: a representation that makes it possible to understand some specific properties of an opaque model or of its predictions
- **Outcome explanation**: do not aim to reconstruct the opaque AI system, but to build an explanator that provides explanations for any specific instance

UNIVERSITÀ DEGLI STUDI FIRENZE

Table of Contents

2 Bias and fairness

Bias

2 Bias and fairness

"Bias is defined as a systematic error in decision-making processes that results in unfair outcomes" (definition by E. Ferrara, 2023)

A machine learning **learns to replicate bias patterns** present in the training data. Identifying and addressing bias is crucial to ensure fair and equitable decision-making

Bias in data

- **Sampling bias**: occurs when the training data is not collected as a random sample of the population, that is some individuals have a larger probability to be sampled (e.g. Literary Digest Poll for 1936 US Presidential Elections)
- **Representation bias**: similar to sampling bias, happens when a dataset does not accurately represent the population it is meant to model, e.g., by missing or under-representing sub-populations (e.g., non-Western cultures in ImageNet)
- **Artefact bias**: occurs when spurious correlations exist between an artefact and labels
- **Aggregation bias**: occurs when conclusions about individuals are drawn from considerations about the entire population (e.g., Simpson's paradox)

Bias in data: Simpson's paradox

NUMBER OF HEALED CASES	Treatment A	Treatment B
Total	273/350 (78.0%)	289/350 (82.6%)

Bias in data: Simpson's paradox

NUMBER OF HEALED CASES	Treatment A	Treatment B
Small stones	81/87 (93.1%)	234/270 (86.7%)
Large stones	192/263 (73.0%)	55/80 (68.8%)
Total	**273/350 (78.0%)**	**289/350 (82.6%)**

Bias in data: Simpson's paradox

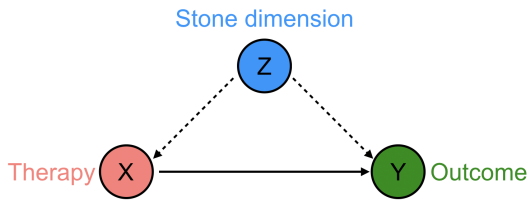
Which is the best treatment for the disease?

- A is best according to the partial measurement
- B is best according to the global measurements

We speak of **Simpson's paradox** in statistics when a trend can be observed in some groups of data, but it disappears when such groups are combined.

Bias in data: Simpson's paradox

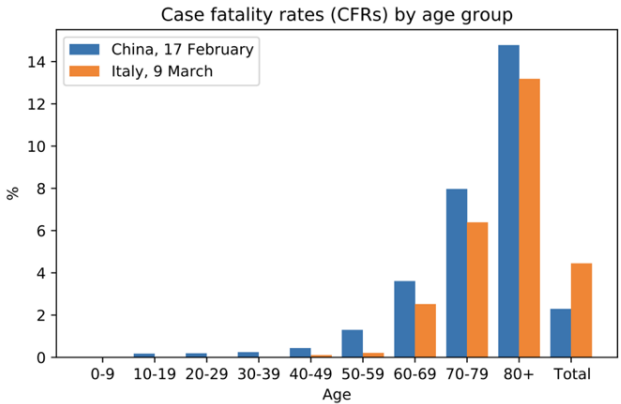
A **confounding variable** (or confounding factor) is a variable which influences both the independent and the dependent variable in our model, and may cause the paradox



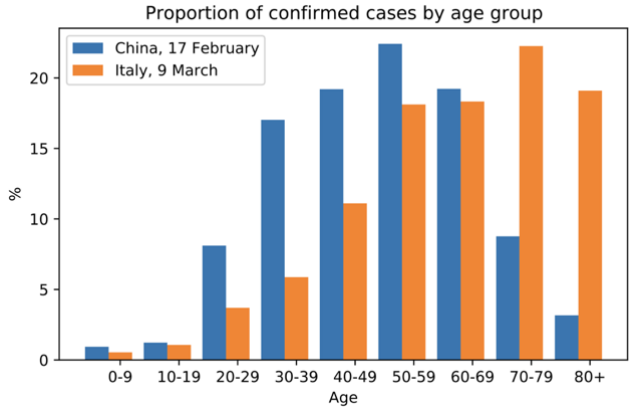
In the kidney study, the stone dimension (i.e., severity) has influenced the decisions regarding the choice of treatment (simpler cases more frequently treated with B)

Bias in data: Simpson's paradox

2 Bias and fairness



Bias in data: Simpson's paradox



Bias in data: Simpson's paradox

Sure-thing principle

An action A that **increases** the probability of an event E in each **subpopulation** must also increase the probability of E in the **population** as a whole, **provided only** that the action does not change the **distribution** of the subpopulations

Bias in algorithms

2 Bias and fairness

- **Algorithmic bias**: occurs when bias is not present in the data, but is introduced by the algorithm (e.g., software for plagiarism comparing longer text sequences is more likely to identify plagiarists among non-English speakers)

- **Interaction bias**: when interaction with humans is biased, resulting in unfair treatment (e.g., chatbot responding differently to men and women)

Bias in algorithms

- **Confirmation bias**: when an AI system is used to confirm pre-existing biases or beliefs held by creators or users (e.g., predict success of job candidates based on biases held by hiring manager)

- **Generative bias**: occurs when the output of a generative model disproportionately reflects specific attributes, perspectives, or patterns present in training data, leading to skewed or unbalanced representations in generated content (e.g., stereotypes)

Fairness

2 Bias and fairness

Bias can easily lead to **unfair decisions**, but how to define fairness?

Fairness is the absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics

There exist many quantities to measure **fairness**. Most of them are based on the existence of a machine learning system that models the probability P of a positive class, and its relation with some **protected attribute** A

Equalized Odds

A predictor satisfies equalized odds with respect to the protected attribute A and outcome Y, if Y and A are independent conditional on $\hat{Y}$:

$$P(\hat{Y} = 1|A = 0, Y = y) = P(\hat{Y} = 1|A = 1, Y = y) \quad , y \in \{0, 1\}$$

The probability of a positive example being correctly classified as positive (TP) and the probability of a negative example being incorrectly classified as positive (FP) should be the same for both protected and unprotected group members

Protected and unprotected groups should have **equal rates** for TP and FP

Equal Opportunity

A binary predictor Y satisfies equal opportunity with respect to A and Y if:

$$P(\hat{Y} = 1|A = 0, Y = 1) = P(\hat{Y} = 1|A = 1, Y = 1)$$

The probability of a positive example being correctly classified as positive should be equal for both protected and unprotected group members

Protected and unprotected groups should have equal TP rates

Demographic Parity, also known as **statistical parity**

A predictor Y satisfies demographic parity if:

$$P(\hat{Y}|A = 0) = P(\hat{Y}|A = 1)$$

The likelihood of a positive outcome should be the same regardless of whether the example belongs to the protected group or not

Fairness

- **Fairness through awareness**
  An algorithm is fair if it gives similar predictions to similar individuals" [48, 87]. In other words, any two individuals who are similar with respect to a similarity (inverse distance) metric defined for a particular task should receive a similar outcome

- **Fairness through unawareness**
  An algorithm is fair as long as any protected attributes A are not explicitly used in the decision-making process

Fairness

More definitions exist...

- Treatment equality
- Test fairness
- Counterfactual fairness
- Fairness in relational domains
- Conditional statistical parity

How to enforce fairness

How to handle bias?

1. Pre-processing: transform the data to remove the underlying discrimination
2. In-processing: modify learning algorithms to remove discrimination during training (e.g., by changing the objective function or by imposing a constraint)
3. Post-processing: performed after training by using a validation set

UNIVERSITÀ DEGLI STUDI FIRENZE

Table of Contents

3 Evaluation

Evaluating explainability

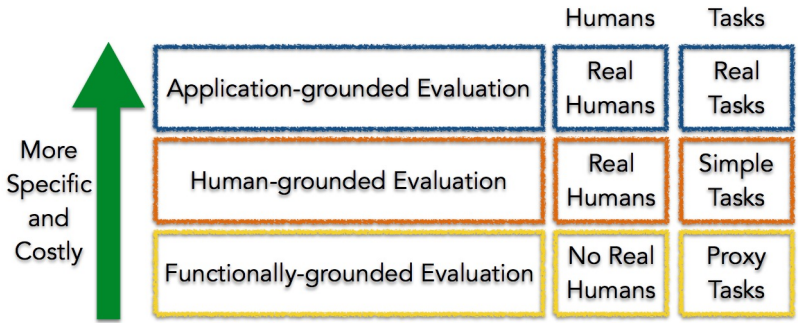
3 Evaluation

Evaluating explainability and interpretability is not straightforward!
There are several ways in which the problem can be addressed...

- Is the system working as designed?
- Are system users treated fairly?
- Is the system compliant to the law?
- Shall we evaluate explanations in the context of the application?
- Shall we evaluate explanations via a quantifiable proxy?
- Are some explanations better than others?

Evaluating explainability

3 Evaluation



[Figure by H. Lakkaraju]

Evaluating explainability

Application-grounded evaluation

- Real humans, real tasks
- Experiment with domain experts on the exact application task
- Most reliable and costly solution as well

Evaluating explainability

Human-grounded evaluation

- Real humans, simplified tasks
- Experiment with non-experts on tasks close to the target application
- Larger pool, less expensive (e.g., pairwise comparisons)

Evaluating explainability

Functionally-grounded evaluation

- Human experiments not needed, maybe unethical, not replicable
- Computational proxy metrics (e.g., feature importance by perturbation)
- Sometimes appropriate when model already validated by humans
- A method is not yet mature
- What proxies are best for what real world applications?
- What factors to consider when designing simpler tasks in place of real world tasks?

Evaluating explainability

3 Evaluation

Quantitative performance metrics (with human contribution)

- **Completeness**: whether an explanation is complete for a user
- **Semplicity**: simpler explanations should be preferred (Occam's Razor)
- **Complexity**: needed time for a human being to understand the explanation
- **Plausibility**: whether an explanation is persuasive for a user
- **Simulability**: whether an explanation can be used on new data
- **Rilevance**: specific metric for specific domain, such as clinical or juridical

Evaluating explainability

Auxiliary or proxy performance metrics

- **Sensitivity**: to what extent a model is sensible to the value of a feature
- **Continuity**: similar examples in input space should have similar explanations
- **Consistency**: to what extant en explanation is consistent across different models
- **Computational cost**: time required to compute the explanation
- **Correctness**: comparison with some ground-truth

UNIVERSITÀ DEGLI STUDI FIRENZE

Table of Contents

4 Exploratory Data Analysis

Data visualization

4 Exploratory Data Analysis

- Univariate analysis: descriptive statistics, visualization, etc.
- Multivariate analysis: correlations, pairwise interaction, etc.
- Data distribution: feature-by-feature analysis, missing values, etc.
- Identify outliers, hints for possible bias, etc.

Data visualization

4 Exploratory Data Analysis

Beware of descriptive statistics: e.g., Anscombe's quartet

A set of four artificially designed datasets with (almost) identical statistics

1		2		3		4	
X1	Y1	X2	Y2	X3	Y3	X4	Y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

[Source: Wikipedia]

Data visualization

4 Exploratory Data Analysis

Beware of descriptive statistics: e.g., Anscombe's quartet

A set of four artificially designed datasets with (almost) identical statistics

Property	Value	Accuracy
Mean of x	9	Exact
Variance of x	11	Exact
Mean of y	7.50	Up to 2 decimal places
Variance of y	4.125	Plus/Minus 0.003
Correlation between x and y	0.816	Up to 3 decimal places
Linear regression line	y = 5x + 3	Up to 2/3 decimal places

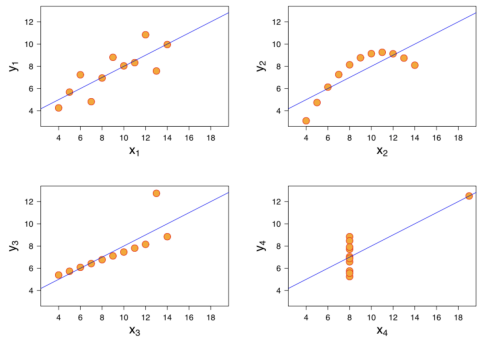
[Source: Wikipedia]

Data visualization

4 Exploratory Data Analysis

Beware of descriptive statistics: e.g., Anscombe's quartet

A set of four artificially designed datasets with (almost) identical statistics

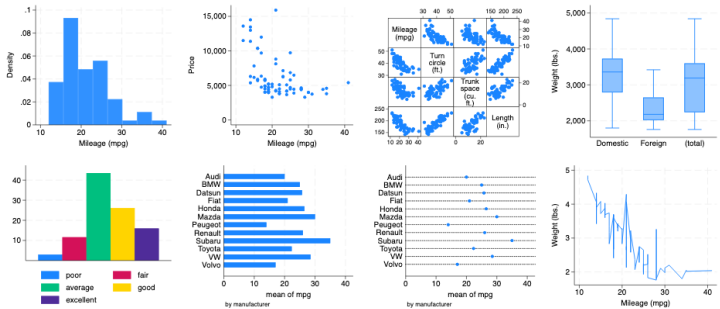


[Source: Wikipedia]

Data visualization

4 Exploratory Data Analysis

Many chart types…



[Source: stata.com]

Data visualization

4 Exploratory Data Analysis

Python libraries and tools

- Classic: pandas, numpy, matplotlib
- More recent: seaborn, plotly, vega-altair
- Highly specific: ydata-profiling, FACETS, KNIME, etc.
- Interactive tools (e.g., see this post)

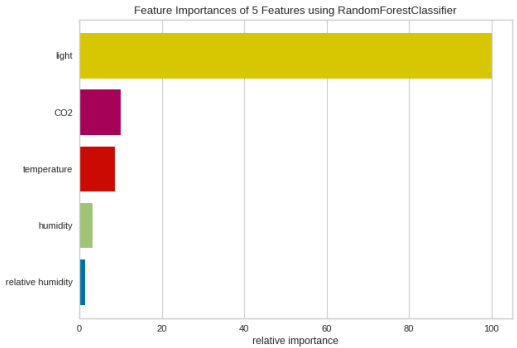
Table of Contents

5 Examples

UNIVERSITÀ DEGLI STUDI FIRENZE

Feature importance

5 Examples



Attention

5 Examples

Task: Hotel location

you get what you pay for . not the cleanest rooms but bed was clean and so was bathroom . bring your own towels though as very thin . service was excellent , let us book in at 8:30am ! **for location and price , this ca n't be beaten** , but it is cheap for a reason . if you come expecting the hilton , then book the hilton ! for uk travellers , think of a blackpool b&b.

Task: Hotel cleanliness

you get what you pay for . **not the cleanest rooms but bed was clean and so was bathroom** . bring your own towels though as very thin . service was excellent , let us book in at 8:30am ! for location and price , this ca n't be beaten , but it is cheap for a reason . if you come expecting the hilton , then book the hilton ! for uk travellers , think of a blackpool b&b.

Task: Hotel service

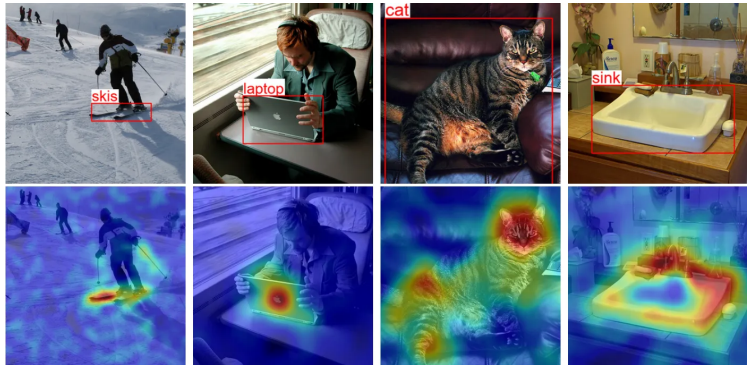
you get what you pay for . not the cleanest rooms but bed was clean and so was bathroom . bring your own towels though as very thin . **service was excellent** , let us book in at 8:30am ! for location and price , this ca n't be beaten , but it is cheap for a reason . if you come expecting the hilton , then book the hilton ! for uk travellers , think of a blackpool b&b.

[Source: Bao et al., 2018]

UNIVERSITÀ DEGLI STUDI FIRENZE

Saliency Maps

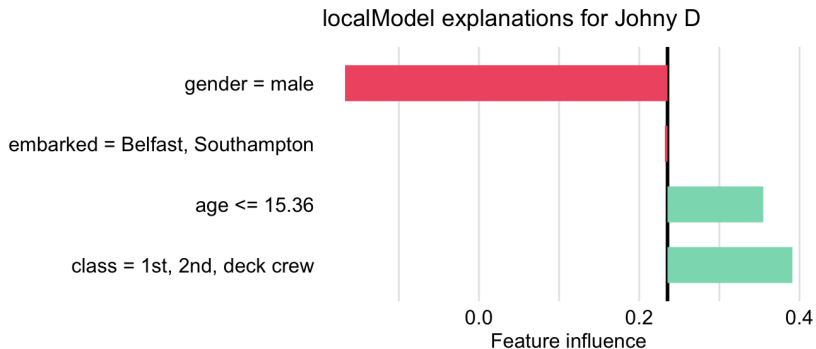
5 Examples



[Source: Petsiuk et al., 2023]

Local Interpretable Model-agnostic Explanations (LIME)

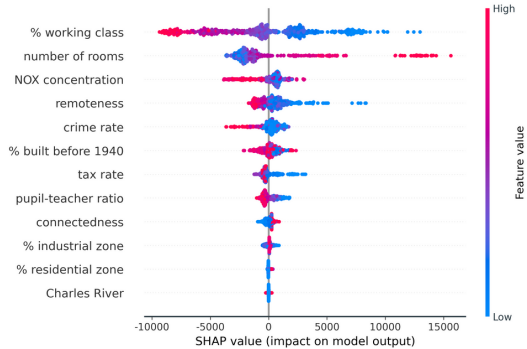
5 Examples



[Source: Biecek and Burzykowski, 2022]

SHapley Additive exPlanations (SHAP)

5 Examples



[Source: Kaggle]