Explainable AI

Getting the *why* out of the *how*

Mattia Setzu

Why use Machine Learning?

1. Don’t know how to solve a problem
2. Can solve a problem, but can’t put it down on paper

Why use Machine Learning?

1. Don’t know how to solve a problem

but…

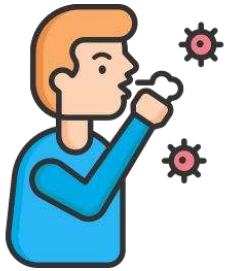
   a. Knowing how gives insight
   b. Want to know diagnose the how

Why use Machine Learning?

2. Can solve a problem, but can't put it down on paper

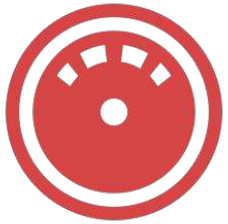
but...

a. Could learn how to solve the problem
b. Want to know diagnose the how

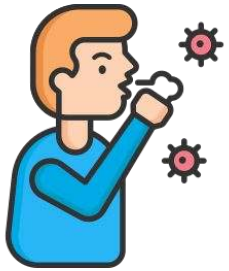


Patient
I have a
weird cough!

Doctor
Let me check!

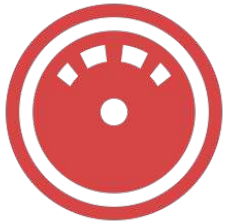


HospitAL 9000

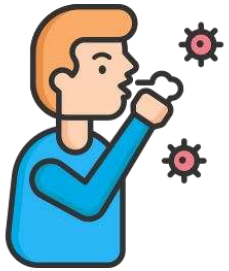


Patient
I have a
weird cough!

Doctor
You may suffer
from D, take pill X.

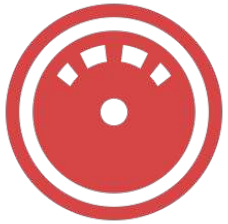


HospitAL 9000
Likely disease D.

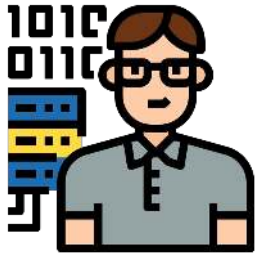


Patient
I have a
weird cough!

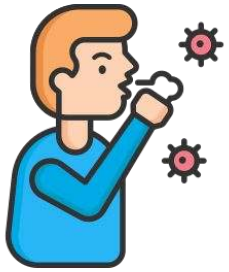
Doctor
You may suffer
from D, take pill X.



HospitAL 9000
Likely disease D.

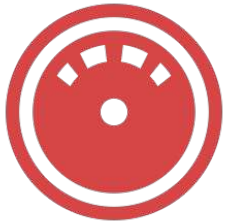


Model developer
Is my model
working?

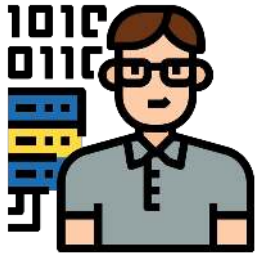


Patient
I have a
weird cough!

Doctor
You may suffer
from D, take pill X.

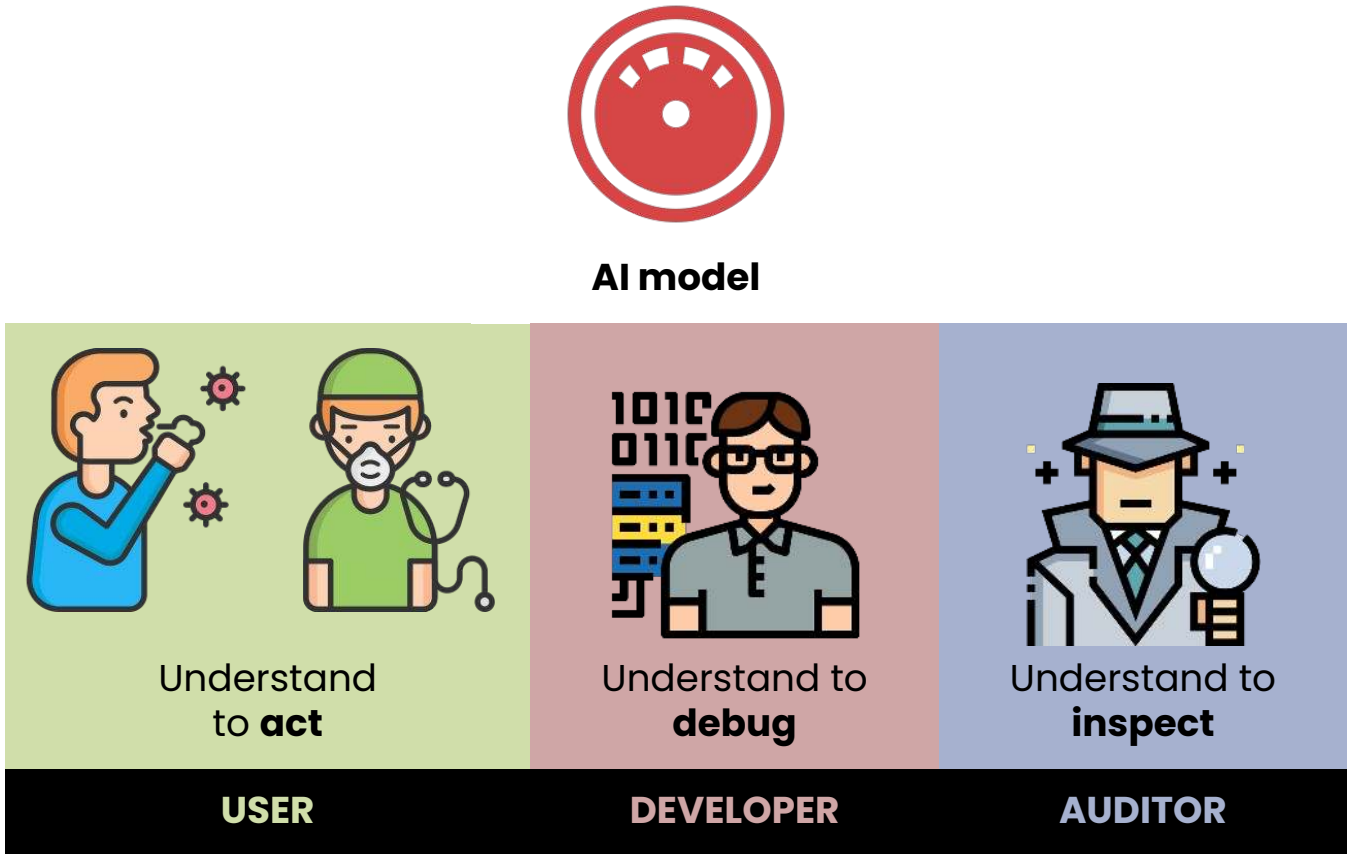


HospitAL 9000
Likely disease D.



Model developer
Is my model
working?

Model auditor
How is the
model working?


AI model
Understand
to act
Understand to
debug
Understand to
inspect
USER
DEVELOPER
AUDITOR

Asking the Ws

- What if...?
- What is important...?
- When...?
- Where...?

A Field with (too) Many variables

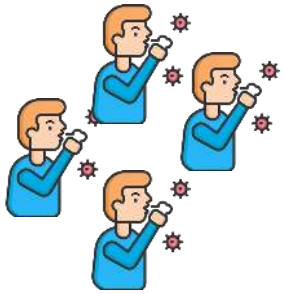
- Model type
- Data type and access
- User type
- Scope demand
- Time demand

Model: what are we explaining?

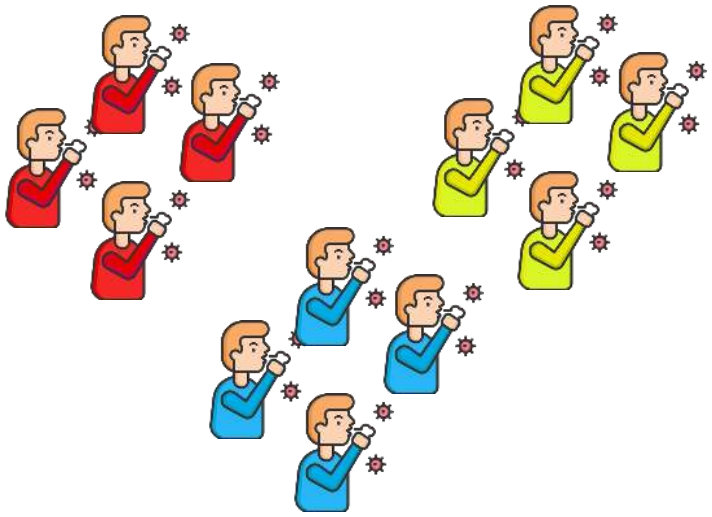
- By design
- Any model (agnostic)
- Neural model
- Ensemble Tree model
- …

Scope: what are we explaining?

Local
Single instance



Subglobal
Group of instances

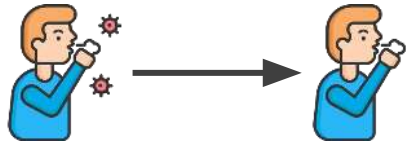


Global
All instances

Explanation: what type of explanation are we providing?

- Feature Importance (FI)
- Data (DT)
  - Prototypes (PRT)
  - Influential Records (IR)
- Model Importance (MI)
- Decision Rules (DR)
- Natural Language (NL)

What if...?
Counterfactual!



Influential Records

Decision Rule

Local

What do I need to change to get a different outcome?

- Gradient-based[1]
- Model agnostic[2]
- Editor-based[3]

[1] FACE: Feasible and Actionable Counterfactual Explanations, Poyiadzi et al.
[2] Decisions, Counterfactual Explanations and Strategic Behavior, Tsirtsis and Gomez-Rodriguez
[3] Explaining NLP Models via Minimal Contrastive Editing (MICE), Ross et al.

What is... important?
Feature Importance

Feature Importance

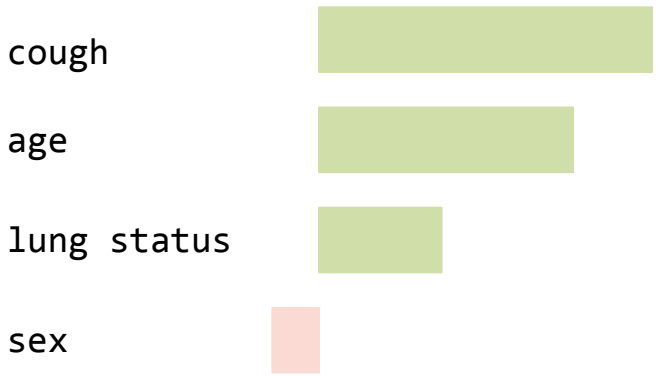
Prototypes

Influential Samples

Local

Global

The higher the importance, the higher the impact on prediction.



[4] A unified approach to interpreting model predictions, Lundberg and Lee
[5] "Why should I trust you?" Explaining the predictions of any classifier, Ribeiro et al.

What is... important?
Prototypes

Feature Importance

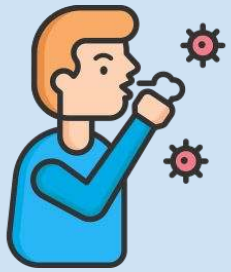
Prototypes

Influential Samples

Local

Global

What is a prototypical record for this decision?



Sex: Male
Age: 50
Smokes: Yes
Lifestyle: Sedentary

[6] This Looks Like That: Deep Learning for Interpretable Image Recognition, Chen et al.
[7] Explaining image classifiers generating exemplars and counter-exemplars from latent representations, Guidotti et al.

What is... important?
Influential records

Feature Importance

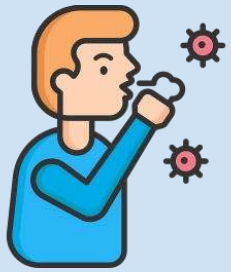
Prototypes

Influential Samples

Local

Global

What training data lead the model to act like this?



Sex: Male
Age: 50
Smokes: Yes
Lifestyle: Sedentary

[8] Tracing Knowledge in Language Models Back to the Training Data, Akyurek et al.
[9] Understanding Black-box Predictions via Influence Functions, Koh & Liang

When… does this happen?
Decision rules

Decision Rules

Local

Global

Can we **describe** the behavior of the model?

```
{Age >= 50 Smokes: Yes,
Respiratory history: Yes},
{Age in [18, 30],
Sex: Male, Lifestyle:
Sedentary},
…
```

```
{Smokes: No, Respiratory
history: No,
Lifestyle: Sporty},
{Age in [2, 24], BMI: Low},
…
```

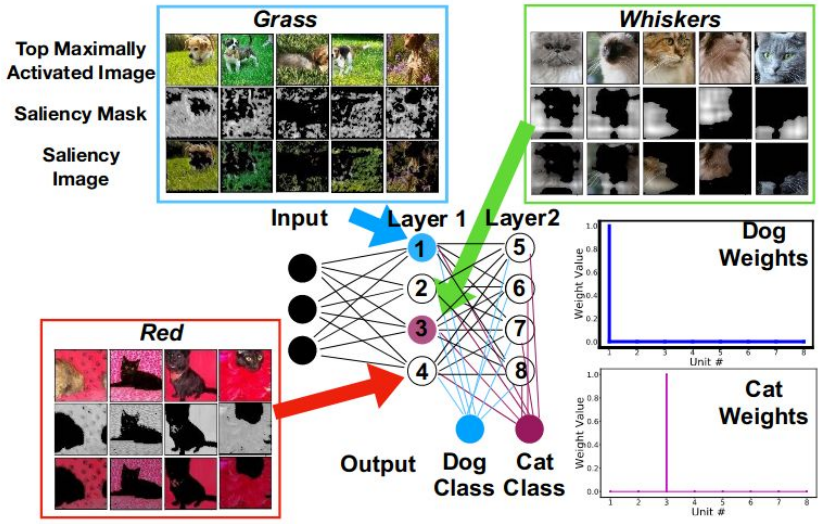
[10] Anchors: High Precision Model-Agnostic Explanations, Ribeiro et al.
[11] Factual and counterfactual explanations for black box decision making, Guidotti et al.

Where... does this happen?
Skill neurons and Neural circuits

Model Importance

Global

Can we identify where a decision is taken?



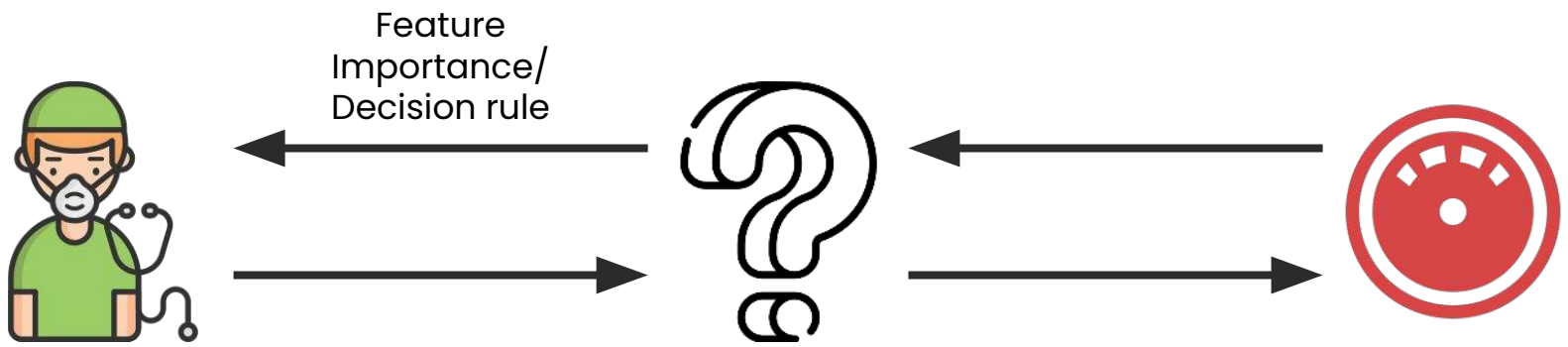
[12] NeuroView: Explainable Deep Network Decision Making, Barberan et al.
[13] Neuron Shapley: Discovering the Responsible Neurons, Ghorbani & Zou

Practical XAI

Q&A

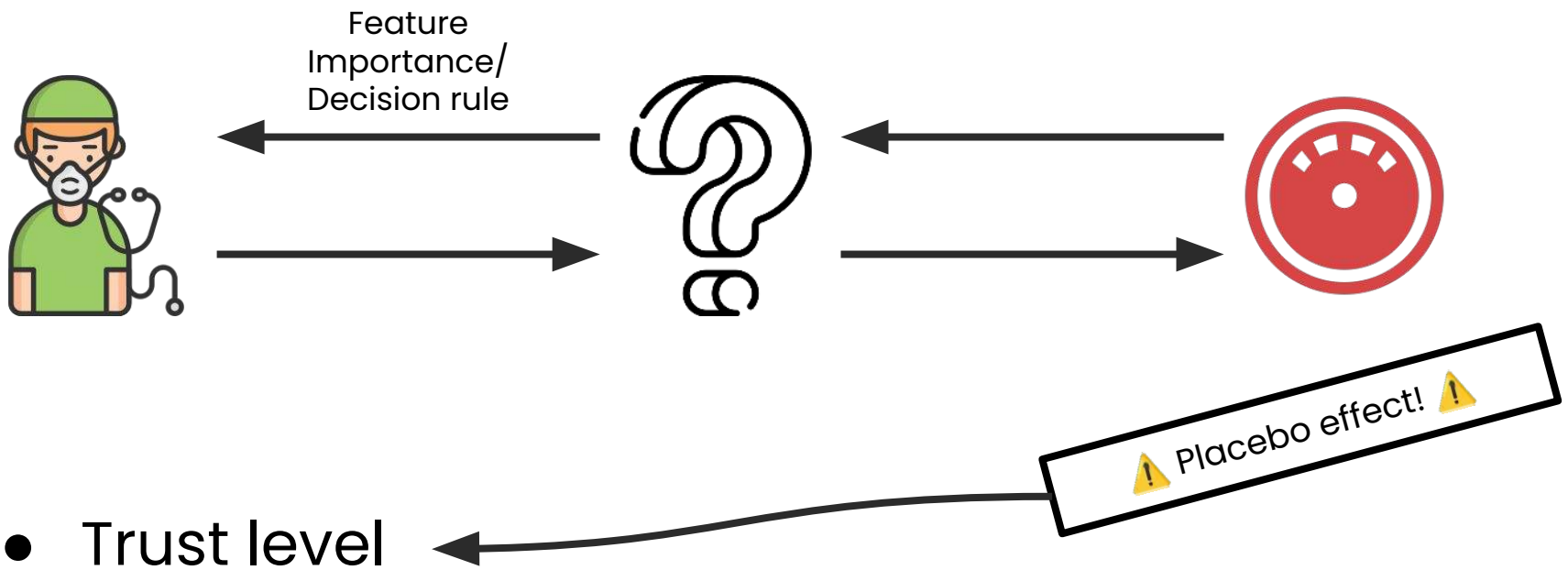
What do I do with the explanation?

What do I do with the explanation?



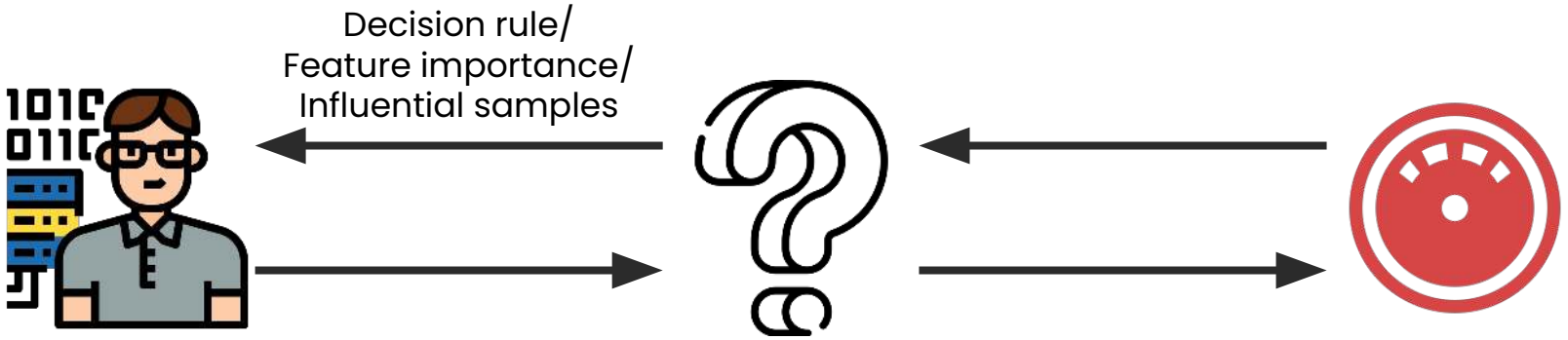
- Trust level
- Adjust your decision

What do I do with the explanation?

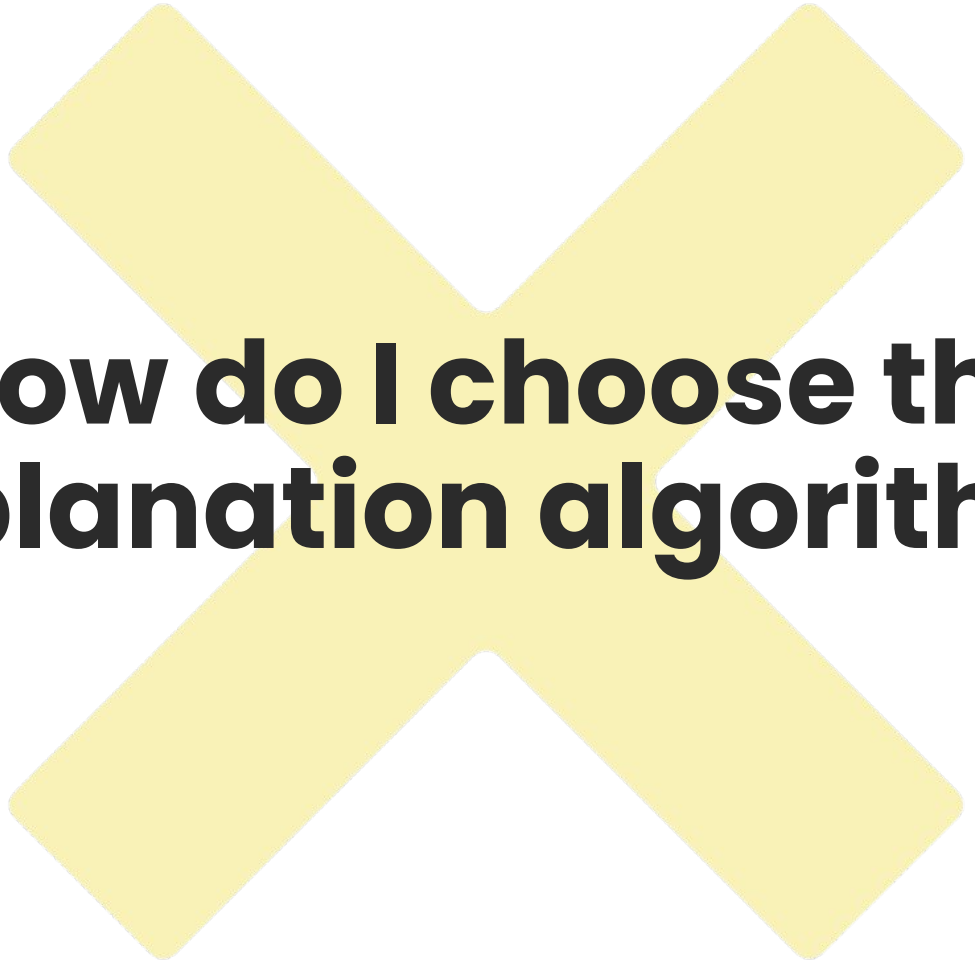


- Trust level
- Adjust your decision

What do I do with the explanation?



- Retrain your model
- Gather data



How do I choose the explanation algorithm?

Should I use a black box or go explainable by-design?

- Dataset type & size
- Task complexity

What explanation type should I choose?

- Question type
- Dataset type & size

What explanation algorithm should I use?

- Dataset size (#features)
- Data density
- Data sensibility to perturbations
- Complexity of the black box

Take-home messages

Trusting models only goes so far.

Different explanations to different questions.

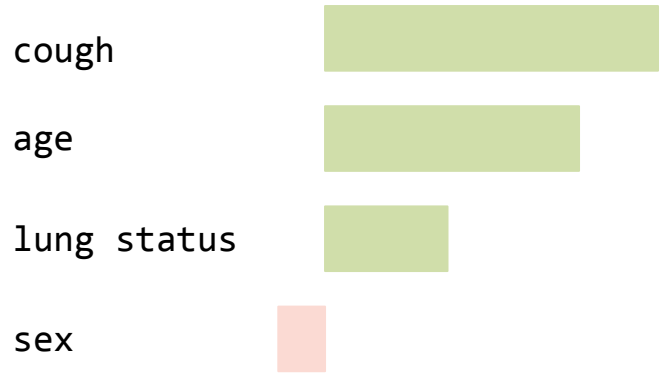
Explanations can enhance both the model and the human.

mattia.setzu@.unipi.it

Feature importance w/ SHAP

SHAP (SHAPley approximations) is an algorithm for local feature importance computation.

SHAP importances are *local* scores that indicate the local contribution of a feature towards a prediction. Twist? They can be aggregated *globally*!



SHAP: the algorithm

1. Select reference set C
2. Compute marginals on C across feature subsets
3. Sample from marginals
4. Sensitivity analysis on the marginals
5. Aggregate sensitivity scores

Validating SHAP

Globally

- How coherent are explanations?

Locally

- Is the model sensitive to high-importance features perturbations?
- Is the model robust to low-importance features perturbations?

Interpreting SHAP

Globally

- What is high importance?
- Are instances w/ high importance different from the ones w/ low importance?