cbs

Big Data in official statistics
Using Machine Learning as Statistical Methods.

Marco Puts

Quality of Official Statistics

- Relevance
- Accuracy
- Accessibility
- Clarity
- Coherence
- Comparability

Quality of Official Statistics

- **Relevance**
- **Accuracy**
- **Accessibility**
- **Clarity**
- **Coherence**
- **Comparability**

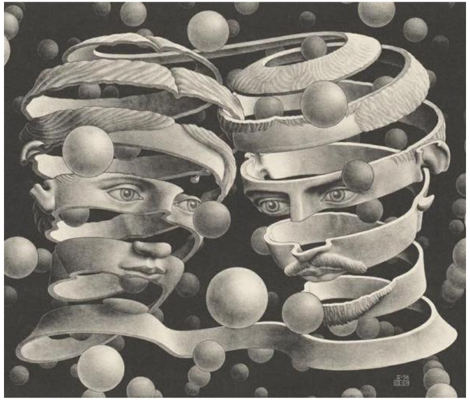
Methods used in Official Statistics also have to meet these quality standards!

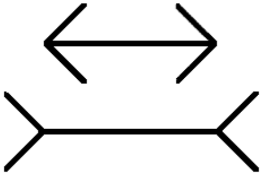
Using Machine Learning in Official Statistics

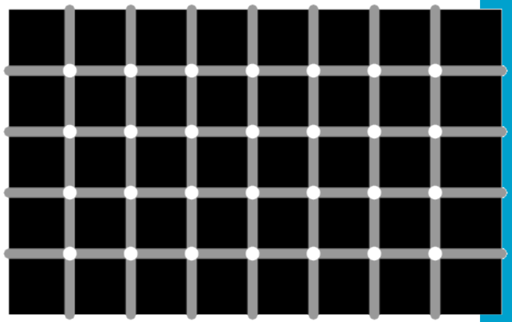
- Induction, Deduction and Abduction
- Machine Learning
- Classification
- The asymptotical behavior towards annotated data
- Representativity of training sets
- Explainable AI

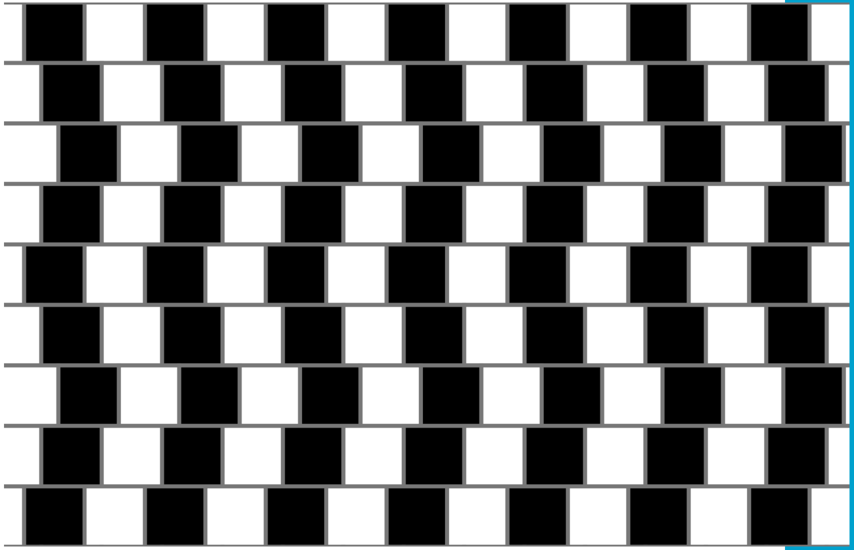
Induction, Deduction and Abduction

Inductive Research









Theory vs. Data driven

- Inductive vs. deductive

- **Deductive**
  - Theory
    - Hypothesis
      - Observation
        - Interpretation

- **Inductive**
  - Observation(Data)
    - Pattern
      - Possible hypothesis
        - Theory

Machine learning

Machine Learning

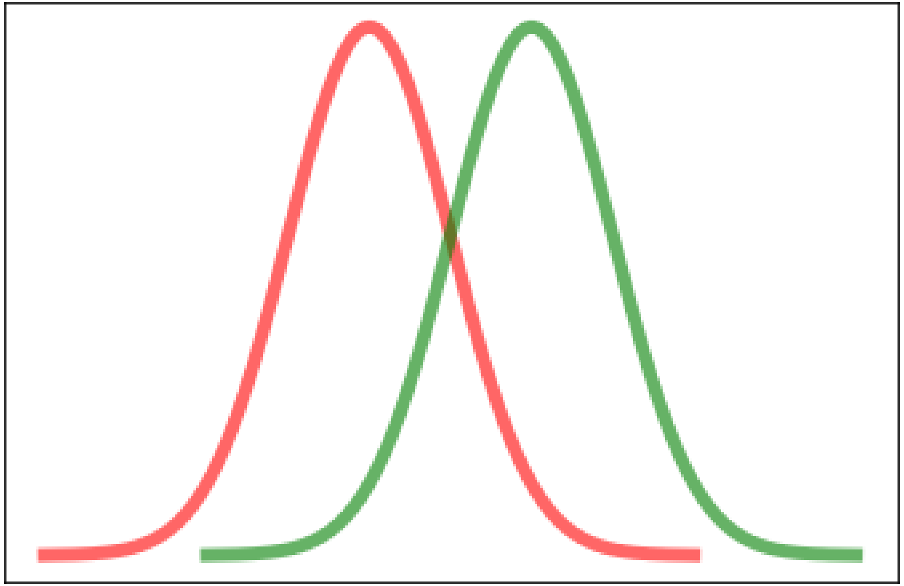
Subfield of Artificial Intelligence

“Learning strategies for Computers”

- Unsupervised learning
  - Clustering
- Supervised learning
  - Classification
  - Regression

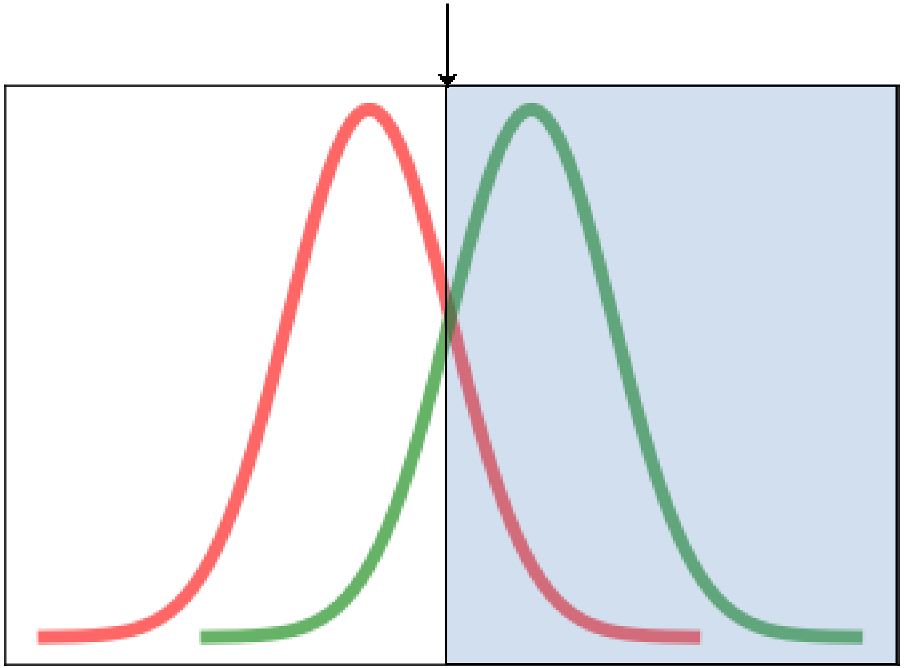
Machine Learning

Classification Explained



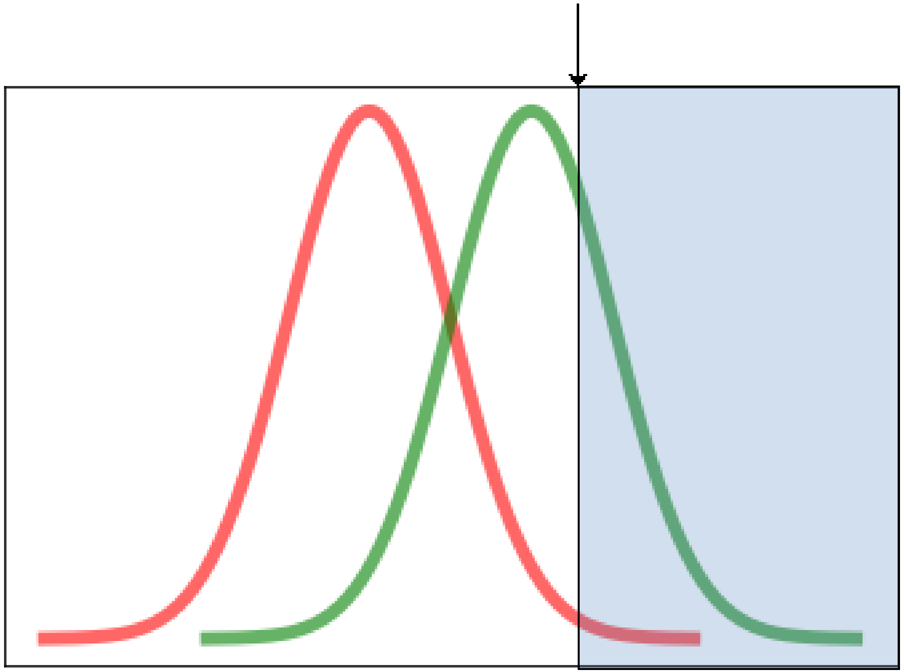
Machine Learning

Threshold



Machine Learning

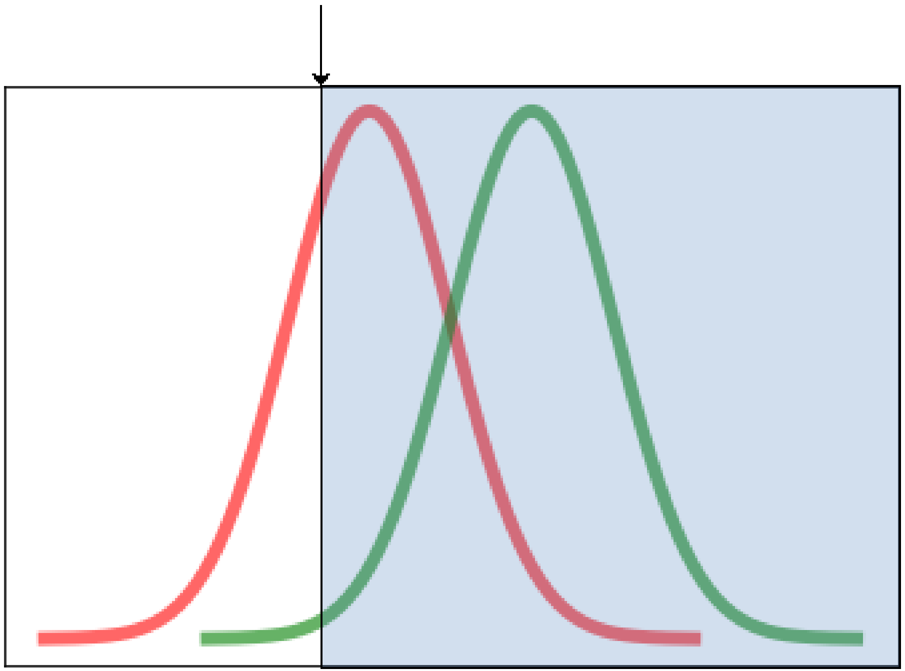
Threshold



High Precision
Low Recall

Machine Learning

Threshold

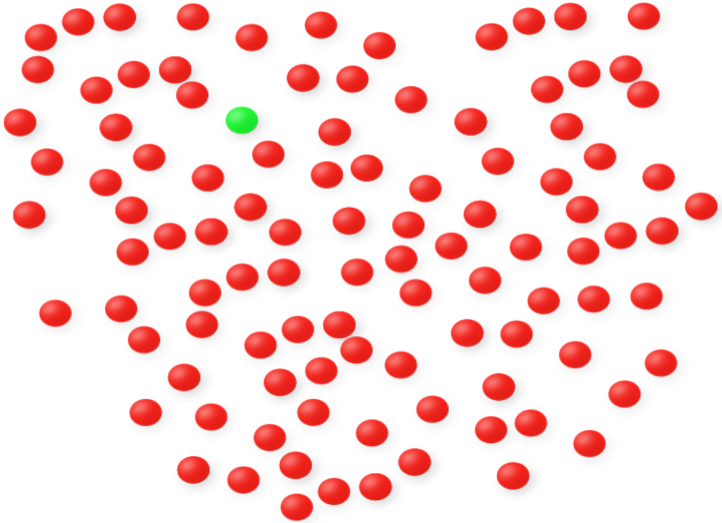


Low Precision
High Recall

Bias in Classifications

Thought experiment

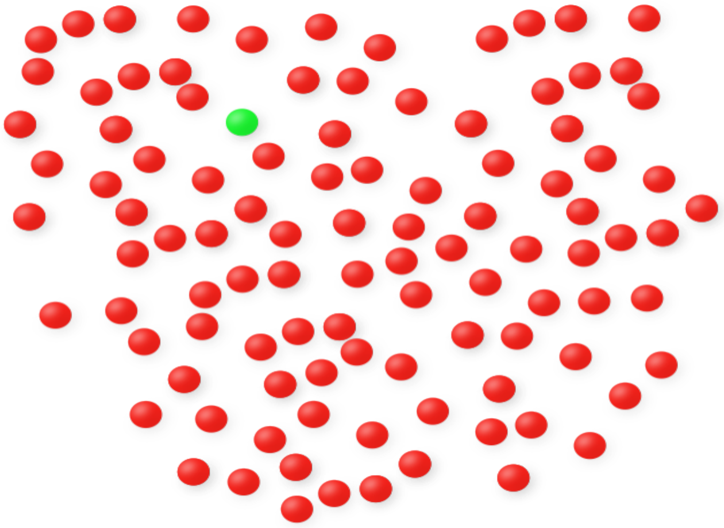
- 99 red marbles and 1 green marble
- Which model can predict the color correctly in 99% of the cases?



Bias in Classifications

Thought experiment

- 99 red marbles and 1 green marble
- Which model can predict the color correctly in 99% of the cases?



Best Model:

Always predict that the marble is RED

Bayesian Ideal Observer Model

A Bayesian view on Classifiers

$$P(c = C|\overline{x}) = \frac{P(\overline{x}|c = C)P(c = C)}{P(\overline{x})}$$

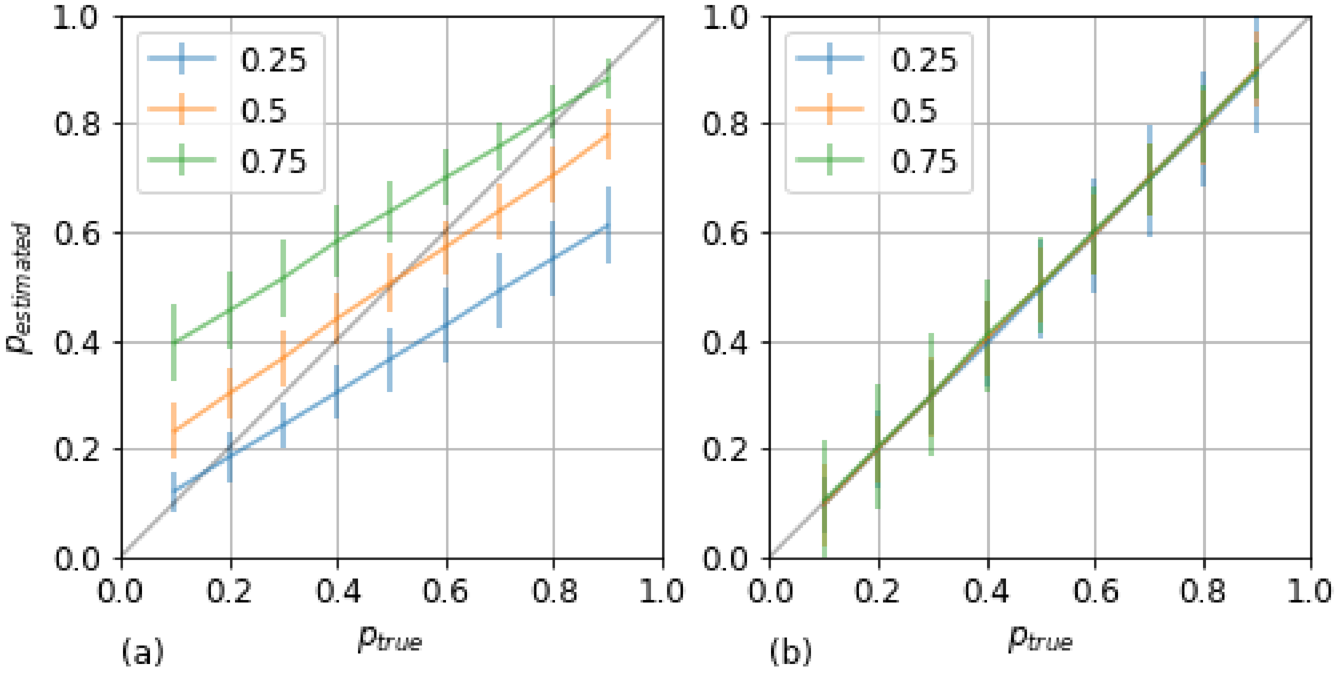
- The prior introduces a Bias!
- $P(\overline{x}) = \sum_{e \in \overline{C}} P(\overline{x}|e)P(e)$

Bias in Classifications

Simulated dataset

without ML correction

with ML correction



Classification vs. Quantification

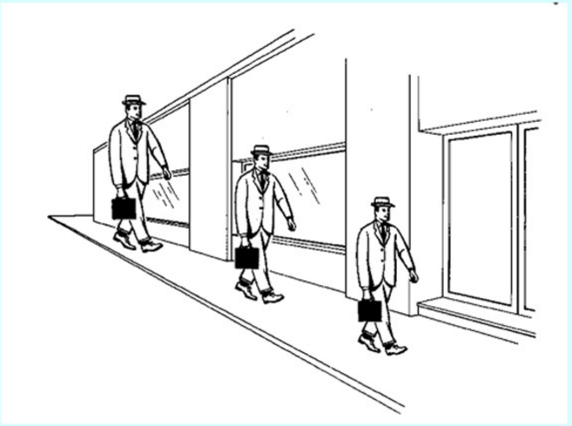
Two ways of using a classifier:

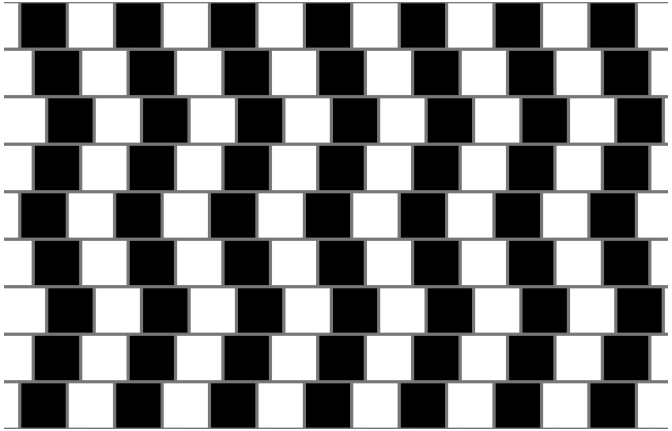
- Threshold/Argmax (classification)
- Expected value by adding up probabilities (quantification)

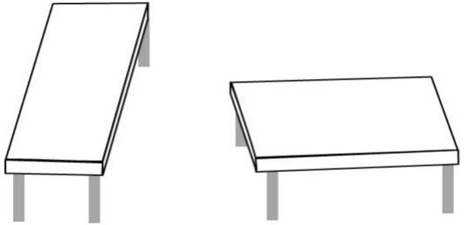
The asymptotical behavior towards annotated data

The asymptotical behavior towards annotated data


Do you see a face or an Eskimo?

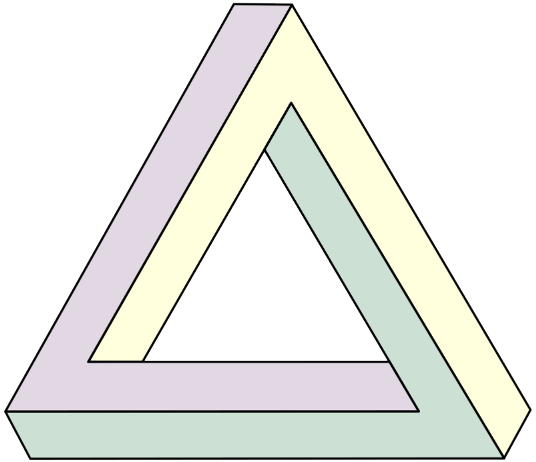






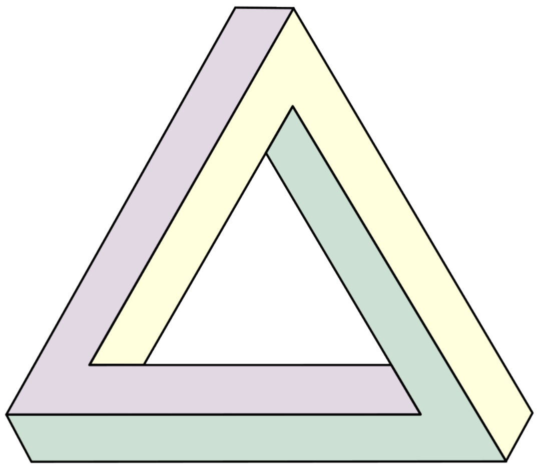
The asymptotical behavior towards annotated data

To what extend is the “observed Ground Truth” real?



The asymptotical behavior towards annotated data

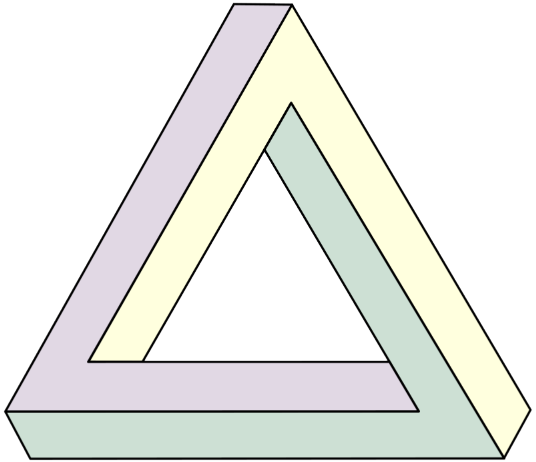
The ML algorithm can never outperform the annotator, since it will learn the mistakes of the annotator.



The asymptotical behavior towards annotated data

Mistakes are present in:

- Training set
- Test set
- Validation set



So how to detect these errors?

Representativity of training sets

Representativity of training sets

get the right set of features

Hard to find the correct set of features

- Rare cases
- Minor classes

Sampling methodology is a valid way to overcome this:

- (Stratified) Random Sampling in the population

Representativity of training sets

get the right set of features

Finding strata:

- Clustering features
- Using background information

Apply stratification:

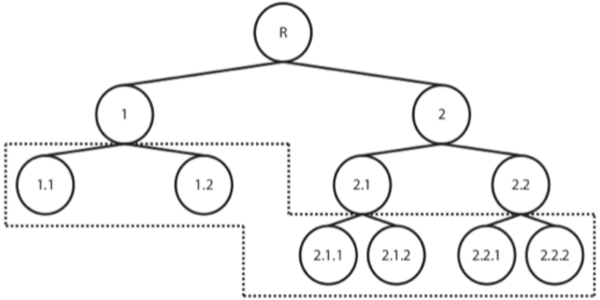
- Weighing
- multiple models

Representativity of training sets

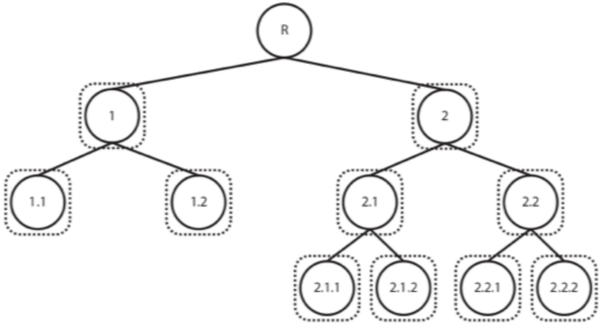
Multiple models

Related model:

- Hierarchical Classification



vs.



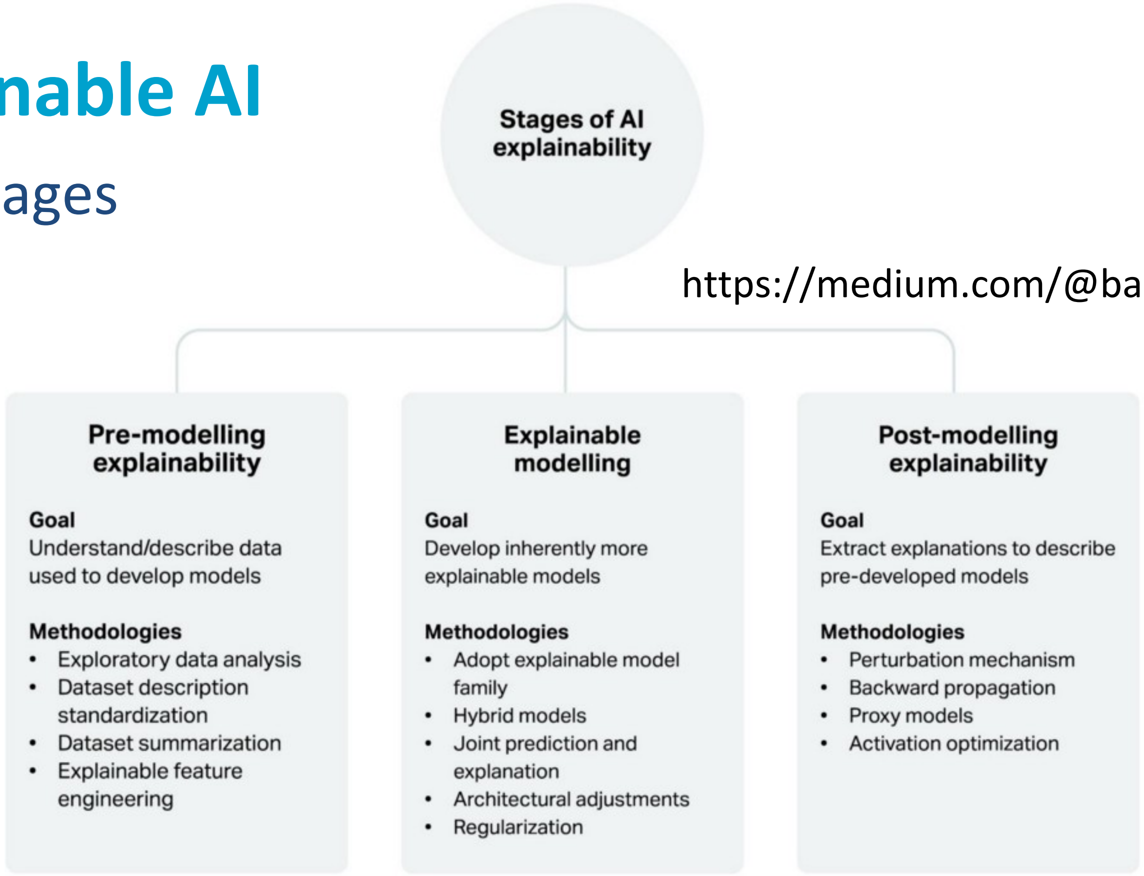
Pedro Chaves, Hierarchical Classification – a useful approach for predicting thousands of possible categories, KDNuggets

eXplainable AI

eXplainable AI

Three Stages

https://medium.com/@bahador.khaleghi



The How of Explainable AI: Pre-modelling Explainability

The How of Explainable AI: Explainable modelling

The How of Explainable AI: Post-modelling Explainability

eXplainable AI

Validation

- The best way to validate a model is by understanding
- Marr (1982): Three levels at which an information -processing device should be described to be fully understood:
  - Computational Theory (How does the model relate to the reality?)
    - What is the goal?
    - Why is it appropriate?
    - Logic of the strategy?
  - Representation and algorithm (Design Pattern)
    - Input/output
    - Algorithm
  - (hardware) Implementation (How is it realized?)

eXplainable AI

Computational Theory (cf. Marr)

- ”..., trying to understand perception by studying only neurons is like trying to understand bird flying by studying only feathers.”
- In AI, the *how* question is often confused with the *why* question.
- “Why are these features selected” vs. “How are these features selected”
- Does it matter how complex the model is when we use this strategy?

Conclusion

Quality of Official Statistics

- Relevance
- Accuracy
- Accessibility
- Clarity
- Coherence
- Comparability

- New research topics within machine learning appear due to applications in Official Statistics!

Facts that matter