

# **INTRODUCTION TO MACHINE LEARNING EXPLAINABILITY**

Part II

Kacper Sokol

# TOPICS

- Classification of Explanations
- Explanation Modalities
- Examples of Explanations
- Data Explainability
- Transparent Modelling
- Post-hoc Explainability

# **CLASSIFICATION OF EXPLANATIONS**

## O1 Explanation Family

- associations between antecedent and consequent
- contrasts and differences
- causal mechanisms

## ASSOCIATIONS BETWEEN ANTECEDENT AND CONSEQUENT

- feature importance
- feature attribution / influence
- rules
- exemplars (prototypes & criticisms)

## **CONTRASTS AND DIFFERENCES**

- (non-causal) counterfactuals  
i.e., contrastive statements
- prototypes & criticisms

## **CAUSAL MECHANISMS**

- causal counterfactuals
- causal chains
- full causal model

# EXPLANATION MODALITIES



## O2 Explanatory Medium

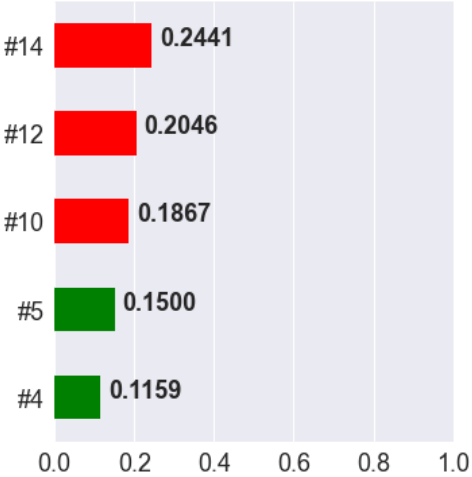
- (statistical / numerical) summarisation
- visualisation
- textualisation
- formal argumentation

# O4 Explanation Domain

## Original domain



## Transformed domain



(**O3** *System Interaction* & **U4** *Interactiveness*)


Provided within a static or interactive protocol

- ~~interactive interface~~
- interactive explanation

# EXAMPLES OF EXPLANATIONS

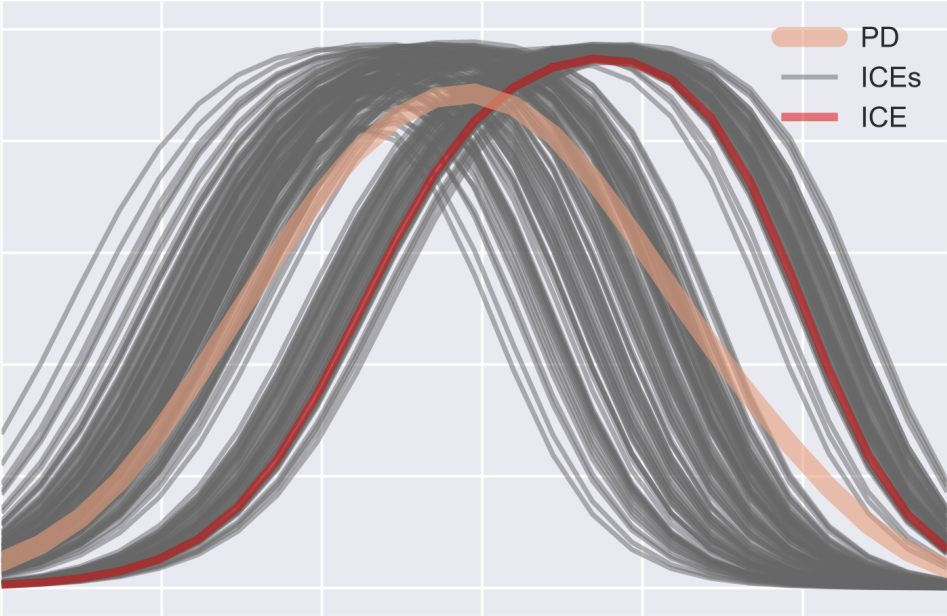
# PERMUTATION FEATURE IMPORTANCE

Height at age 20 (cm)	Height at age 10 (cm)	...	Socks owned at age 10
182	155	...	20
175	147	...	10
...	...	...	...
156	142	...	8
153	130	...	24

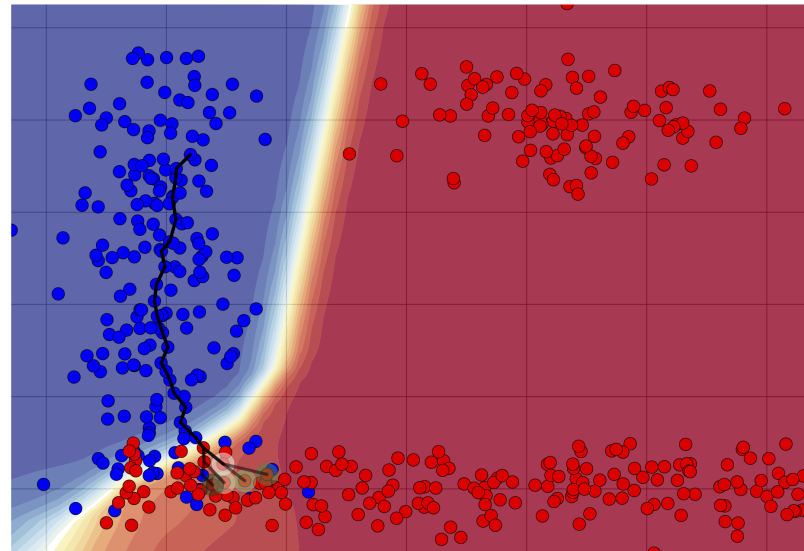


<https://www.kaggle.com/code/dansbecker/permutation-importance>

# INDIVIDUAL CONDITIONAL EXPECTATION & PARTIAL DEPENDENCE

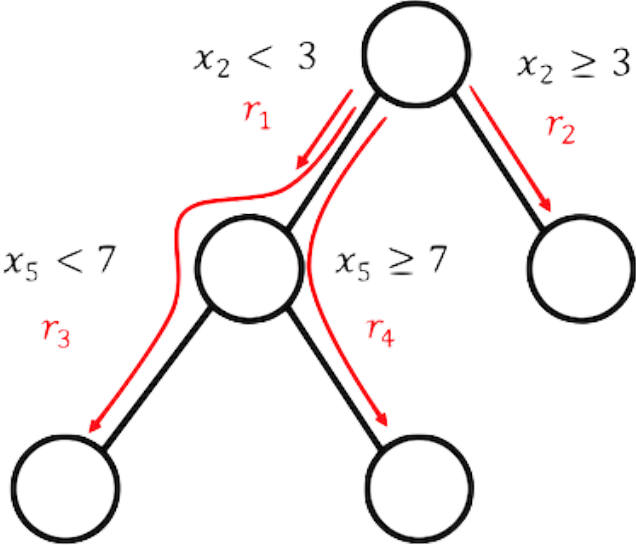


# FACE COUNTERFACTUALS



Poyiadzi, Sokol, Santos-Rodriguez, De Bie and Flach, 2020. FACE: Feasible and actionable counterfactual explanations

# RULEFIT



<https://christophm.github.io/interpretable-ml-book/rulefit.html>



# DATA EXPLAINABILITY

- Data as an (implicit) model
- Data summarisation and description
- ~~Exemplars, prototypes and criticisms~~
- ~~Dimensionality reduction (e.g., t-SNE)~~

# TRANSPARENT MODELLING

- Rule lists and sets
- Linear models
- Decision trees
- $k$ -nearest neighbours and  $k$ -means

# POST-HOC EXPLAINABILITY

Understandable model of the relation between *inputs* and *outputs*

- SHAP
- LIME

