# Explainability in Machine Learning

**Nicola Picchiotti**

**6 July 2020**

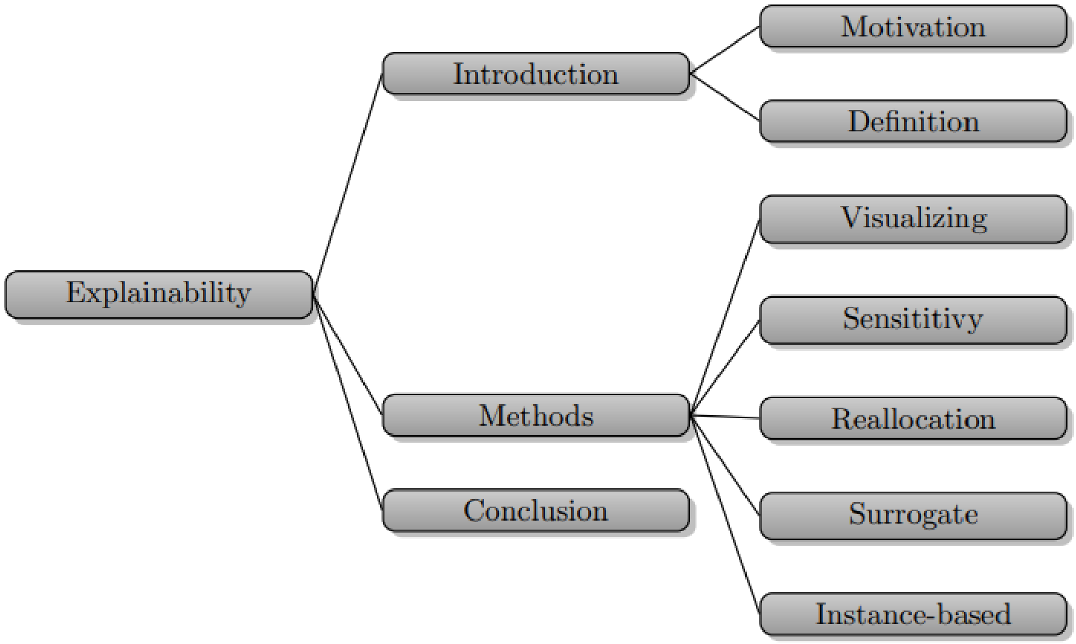
UNIVERSITÀ DI SIENA
1240

# Disclaimer

The views, thoughts and opinions expressed in this talk are those of the authors in their individual capacity and should not be attributed to Banco BPM or to the authors as representatives or employees of Banco BPM.

# Outline

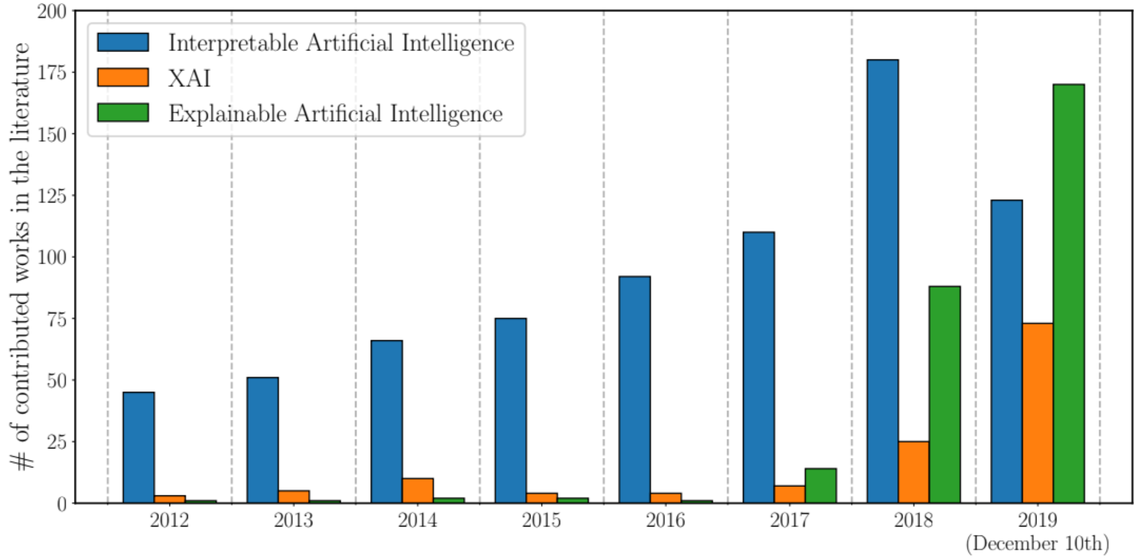
1. Introduction
2. Visualizing
3. Sensitivity
4. Reallocation
5. Surrogate
6. Instance - based
7. Conclusion



# Introduction

## State of Art

- **Molnar, Christoph “*Interpretable Machine Learning*”, 2020.**
- **Guidotti, Riccardo et al. ”*A Survey of Methods for Explaining Black Box Models*”, 2018.**
- **Arrieta Alejandro, “*Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI*”, 2019.**
- **Adadi, Amina et al. “*Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI)*" , 2018**




*The term explainability denotes the **ability to translate** something e.g. a **model**, a piece of the model, or a prediction of the model in an **understandable manner to human***

# Motivation

*“The human wants something that mectric doesn’t”*

Incompleteness in the problem formalization:


- Mismatch goal-objective



- Safety for high risk application

- Ethics, non-discriminative, right of explanation

- Knowledge discovering / Feature Engineering


- **Lipton, Zachary C "The mythos of model interpretability" 2018,**
- **Ribeiro, Marco Tulio et al. " Why should i trust you? Explaining the predictions of any classifier“, 2016.**
- **Doshi-Velez, Finale et al. "Towards a rigorous science of interpretable machine learning.“ , 2017.**

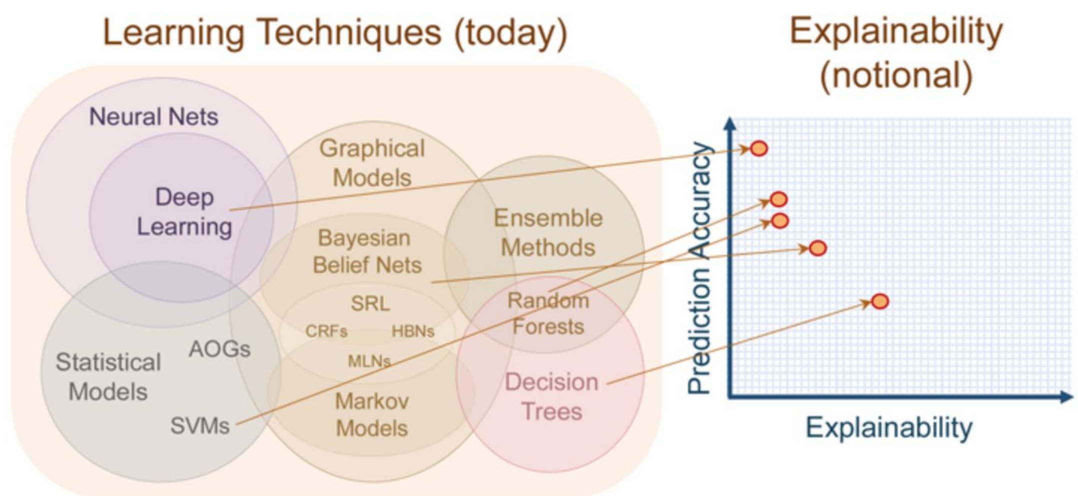
# White/Gray/Black box



- Linear regression
- Rule based

- Decision Tree
- Logistic Regression
- Linear SVM

- Ensemble Methods
- Neural Network
- SVM




**Explainability techniques**

| | |
|---|---|
| Global | Local |
| Model agnostic | Model Specific |

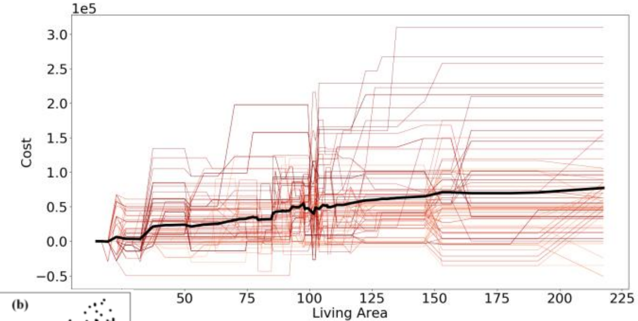
Tabular | Text | Image

- **Gunning, David " Explainable Artificial Intelligence (XAI).“ , DARPA, 2016.**

# Visualizing

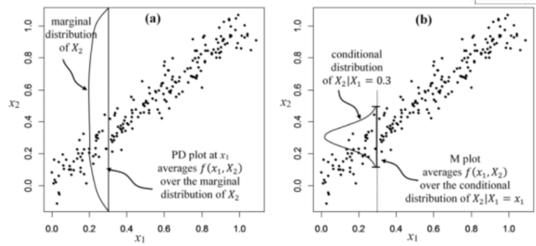
- Individual Conditional Expectation (ICE)

$$ICE^{(n)}(\boldsymbol{x}_S) = \hat{f}([\boldsymbol{x}_S, \boldsymbol{x}^{(n)}_{D/S}])$$



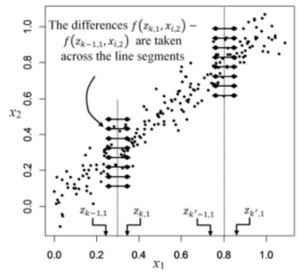

- Partial Dependence Plot (PDP)

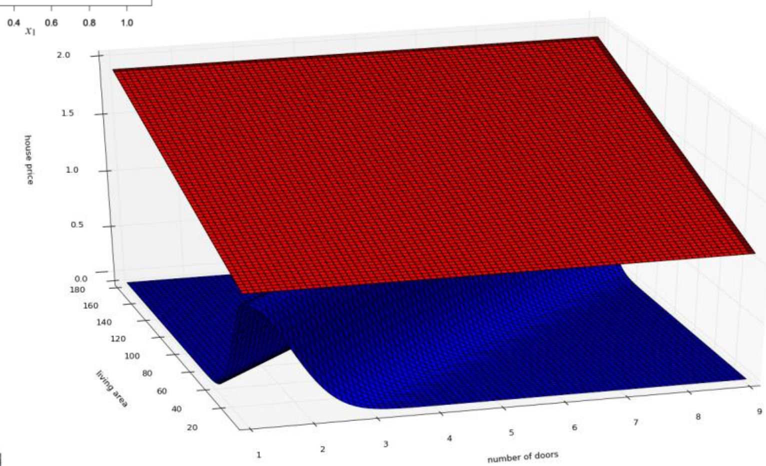
$$f_S(\boldsymbol{x}_S) = \mathbb{E}_{\boldsymbol{x}_C}[\hat{f}(\boldsymbol{x}_S, \boldsymbol{x}_C)] = \int_{\boldsymbol{x}_C} f(\boldsymbol{x}_S, \boldsymbol{x}_C) d\mathbb{P}\boldsymbol{x}_C$$




- M plots

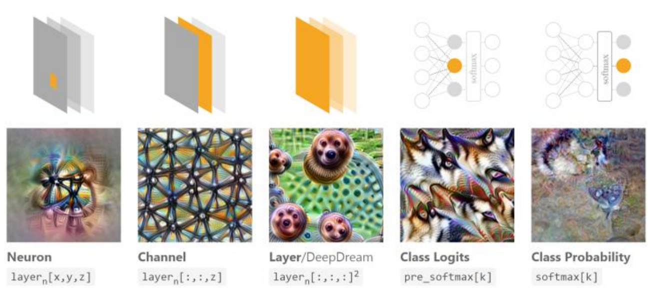
$$f_S(\boldsymbol{x}_S) = \mathbb{E}_{\boldsymbol{x}}[f(\boldsymbol{x})|\boldsymbol{x}_S] = \int_{\boldsymbol{x}} f(\boldsymbol{x}) d\mathbb{P}(\boldsymbol{x}_C|\boldsymbol{x}_S)$$




- Accumulated local effects (ALE)

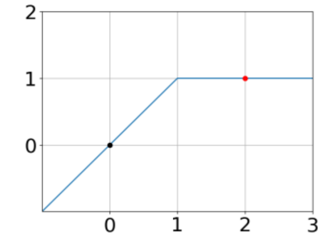
- Feature Visualization

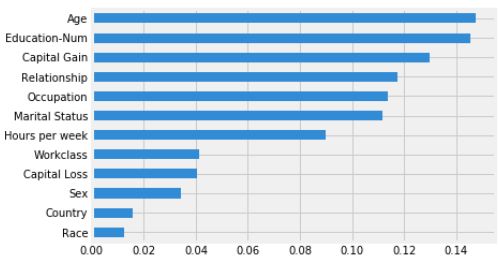



- Apley, Daniel et al. “Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models”, 2019.
- Goldstein, Alex et al. " Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation " , 2014.

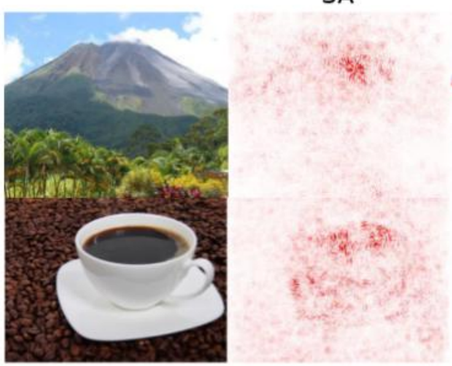
# Sensitivity

- Perturbation-based $\boxed{L}$
  - Permuted feature importance $$F_i = L(Y, f(X_{\text{permuted}, i})) - L(Y, f(X))$$
- Gradient-based $\boxed{f}$
  - Activation Maximization (maximization penalized)
  - Gradient Norm
  - Integrated Gradients (IG)
  - Deep Learning Important FeaTures (Deep Lift)

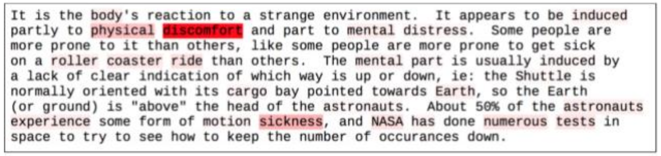



SA



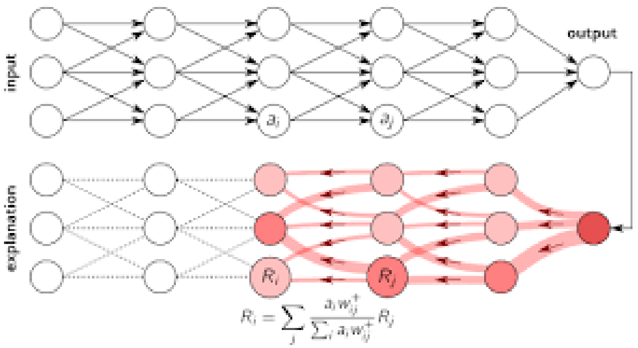
SA



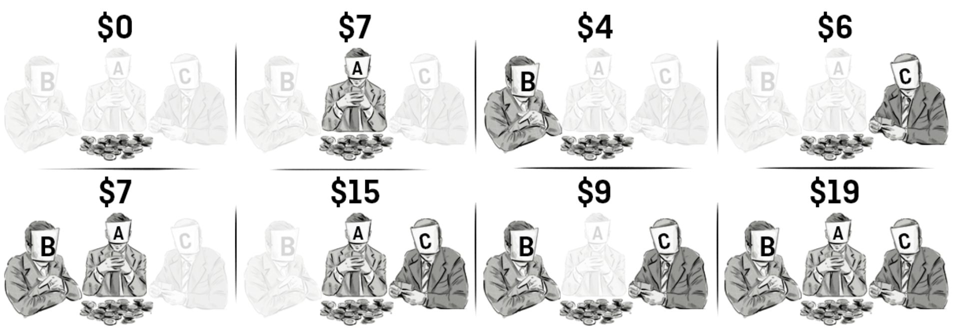

- **Samek , Wojciech et al. Explainabile Artificial Intelligence: understanding, visualizing and interpreting deep learning models**
- **Shrikumar , Avanti et al. Learning Important Features Through Propagating Activation Differences, 2019**
- **Sundararajan, Mukund et al. “Axiomatic Attribution for Deep Networks”, 2017**

# Reallocation (1)

- Layer-wise relevance propagation (LRP)




- Shapley Value




$$v(S \cup \{i\}) - v(S)$$

- ○ Players: $D$
- ○ Overall Payoff: $v(D)$
- ○ Characteristic function: $\nu : \mathcal{P}(D) \rightarrow \mathbb{R}$

$$\sum_{S \subseteq D/\{i\}\ with |S|=k} \frac{v(S \cup \{i\}) - v(S)}{\frac{|D-1|!}{k!(|D|-1-k)!}}$$

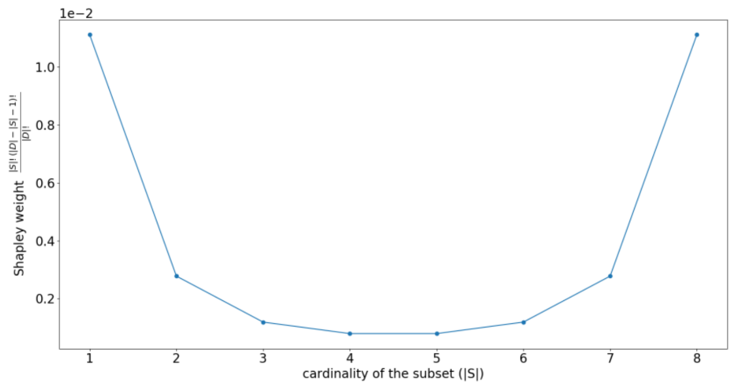
$$\sum_{k \in K} \left( \sum_{S \subseteq D/\{i\} with |S|=k} \frac{(v(S \cup \{i\}) - v(S))}{\frac{|D-1|!}{k!(|D|-1-k)!}} \right) \frac{1}{|D|}$$



- **Mantovan, Gregoire et al. "Methods for interpreting and understanding deep neural networks", 2018**

# Reallocation (2)

$$I_i(v) = \sum_{S\subseteq D/\{i\}} \frac{|S|!(|D|-|S|-1)!}{|D|!}\left(v(S\cup\{i\}) - v(S)\right)$$



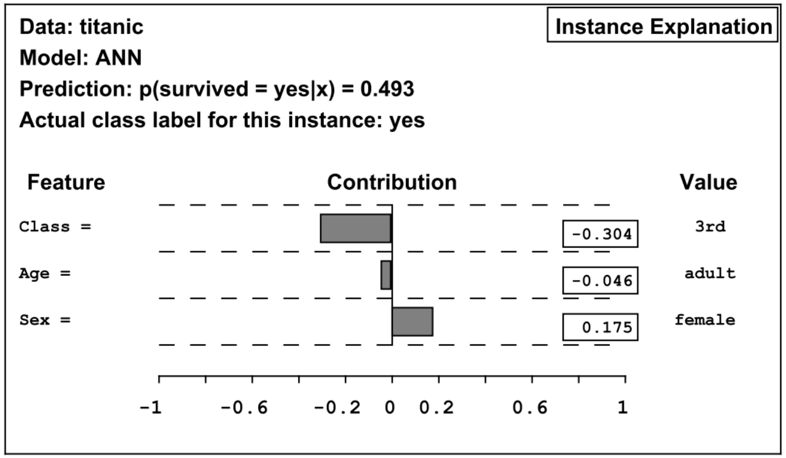

Efficency property

$$\sum_{i\in D} I_i(v) = v(D)$$

- Players -> Features of a particular instance
- Overall Payoff (prediction gain) -> $\hat{f}(\boldsymbol{x}) - \mathbb{E}_X[\hat{f}(X)]$
- Characteristic function $v(Q) = \mathbb{E}_{D/Q}[\hat{f}(D/Q)|Q] - \mathbb{E}_D[\hat{f}(D)]$

$$I_i(\boldsymbol{x}, v) = \frac{1}{|D|!}\sum_{\mathcal{O}\in\pi(|D|)}\left(v(Pre^i(\mathcal{O})\cup\{i\}) - v(Pre^i(\mathcal{O}))\right)$$

$$I_i(v) = \frac{1}{|D|!}\sum_{\mathcal{O}\in\pi(|D|)}\sum_{\boldsymbol{w}\in\mathcal{X}} p(\boldsymbol{w})\left(\hat{f}(\boldsymbol{w}_{[\forall i\in Pre^i(\mathcal{O})\cup\{i\}:\, w_i=x_i]}) - (\hat{f}(\boldsymbol{w}_{[\forall i\in Pre^i(\mathcal{O}):\, w_i=x_i]})\right)$$




- **Strumbelj, Erik et al. “Explaining prediction models and individual predictions with feature contributions”, 2014**

# Surrogate (1)

- Local Interpretable Model-Agnostic Explanations (LIME)





$$\zeta(x) = arg \min_{g \in G} \mathcal{L}(f, g, \pi_{\boldsymbol{x}}) + \Omega(g)$$

$$\mathcal{L}(f, g, \pi_{\boldsymbol{x}}) = \sum_{\boldsymbol{z}, \boldsymbol{z}' \in Z} \pi_{\boldsymbol{x}}(\boldsymbol{z}) \left(f(\boldsymbol{z}) - g(\boldsymbol{z}')\right)^2$$

x

| F 1 | F 2 | F 3 | F 4 |
|---|---|---|---|
| 0.2 | 0.8 | 0.1 | 0.9 |

z

| F 1 | F 2 | F 3 | F 4 |
|---|---|---|---|
| 0.2 | 0.8 | 0.5 | 0.9 |

| F 1 | F 2 | F 3 | F 4 |
|---|---|---|---|
| 0.2 | 0.5 | 0.5 | 0.9 |

| F 1 | F 2 | F 3 | F 4 |
|---|---|---|---|
| 0.2 | 0.8 | 0.5 | 0.5 |

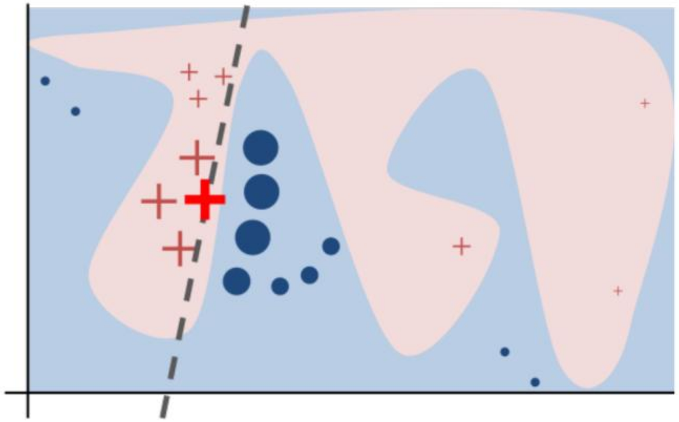
z’

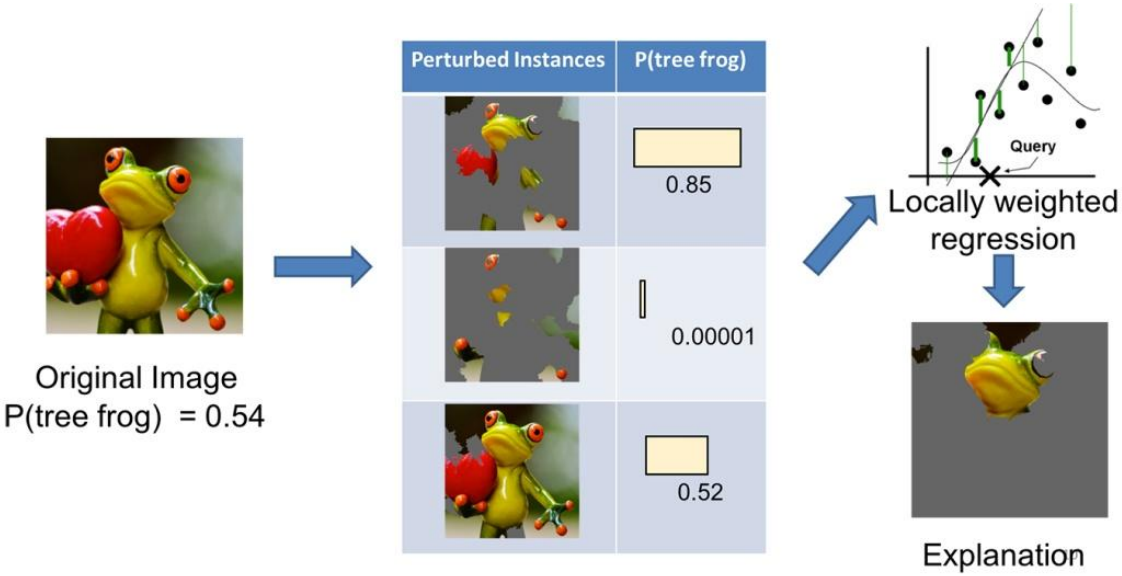
| F 1 | F 2 | F 3 | F 4 |
|---|---|---|---|
| 1 | 1 | 0 | 1 |

| F 1 | F 2 | F 3 | F 4 |
|---|---|---|---|
| 1 | 0 | 0 | 1 |

| F 1 | F 2 | F 3 | F 4 |
|---|---|---|---|
| 1 | 1 | 0 | 0 |

# Surrogate (2)






https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime

- Trepan: tree based surrogate models

- **Craven, Market al. "Extracting Tree-Structured Representations of Trained Networks "**

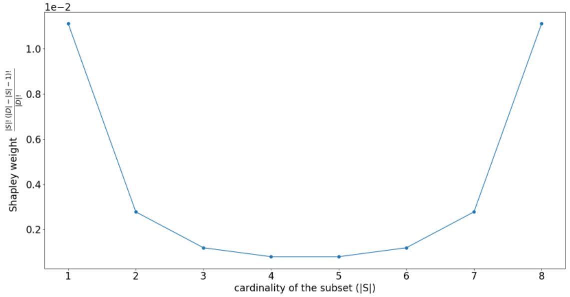
# Surrogate (3)

- SHAP





$$\min_{\boldsymbol{x}} \sum_{S\subset D} k(S)\left(v(S)-\sum_{i\in S}x_i\right)^2$$




| | $v(S)$ | $\sum_{i\in S} x_i$ |
|---|---|---|
| F1 | 2 | 3 |
| F2 | 3 | 3 |
| F3 | 1 | 2 |
| F1 + F2 | 7 | 6 |
| F1 + F3 | 6 | 5 |
| F2 + F3 | 6 | 5 |
| F1 + F2 + F3 | 8 | 8 |

Shapley Values


- **Lundberg, Scott et al. "A unified approach to interpreting model predictions", 2017**
- **Charnes, A et al. "Extremal principle solutions of games in characteristic function form: core, Chebychev and Shapley value generalizations", 1988**

# Instance-Based

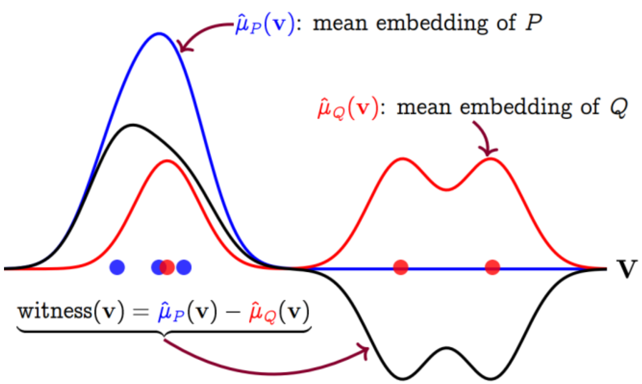
- Counterfactual explaination

$$L(\boldsymbol{x}, \boldsymbol{x}', y_c, \lambda) = \lambda \cdot (\hat{f}(\boldsymbol{x}') - y_c)^2 + d(\boldsymbol{x}, \boldsymbol{x}')$$

$$\arg\min_{\boldsymbol{x}'}$$

- Prototype selection

$$MMD(F, P, Q) = \sup_{f \in \mathcal{F}} (\mathbb{E}_{X \sim P}[f(X)] - \mathbb{E}_{Y \sim Q}[f(Y)])$$



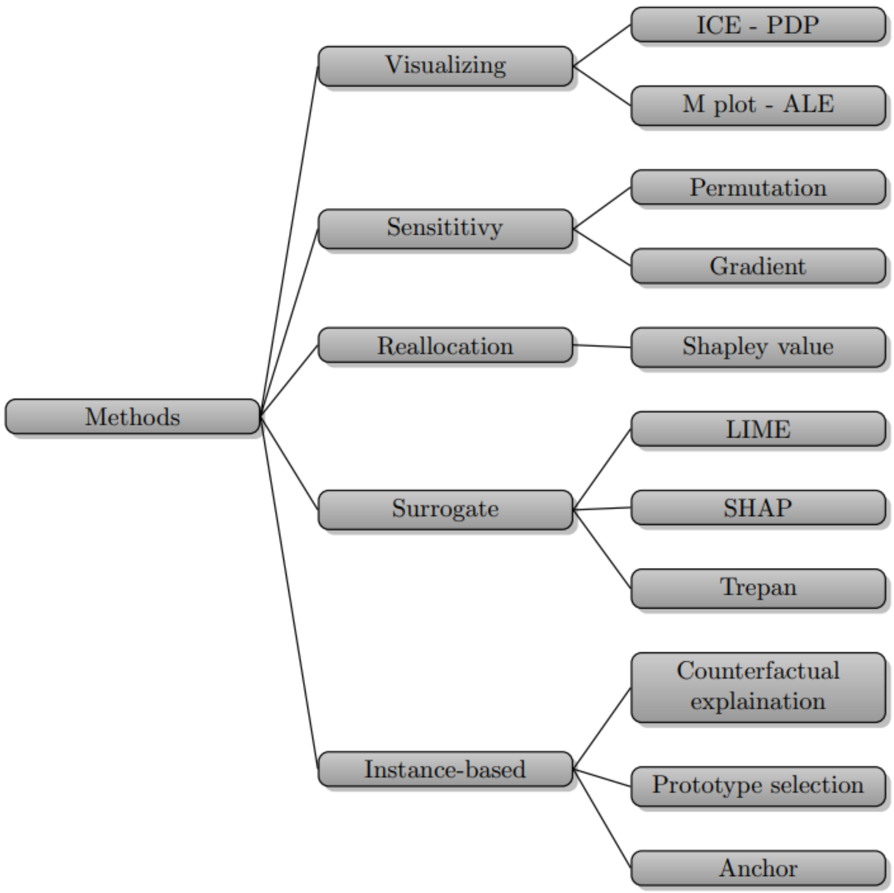


- Anchor

- **Been, Kim et al. "Examples are not Enough, Learn to Criticize! Criticism for Interpretability", 2016**
- **Zhu , Xiaojin et al. "Machine Teaching: An Inverse Problem to Machine Learning and an Approach Toward Optimal Education", 2015**

# Conclusion

- Evolving, broad, etherogeneous field;
- Visualizing / Sensitivity /Local Linearize / Game theory;
- Explainability by design;
- Logic constraints with feature importances.



# Thanks for your attention

nicola.picchiotti@gmail.com