ε.δε.μ²
ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ & ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

# STOCHASTIC PROCESSES & OPTIMIZATION IN MACHINE LEARNING

## eXplainable AI (XAI)

## 1. Definitions, Intrinsic & Model-Agnostic XAI Methods

## 2. PI (Permutation Feature Importance)

## 3. SHAP (Shapley Additive exPlanations)

## 4. LIME (Local Interpretable Model Agnostic Explanation)

Dr. Nikos Kostopoulos

nkostopoulos@netmode.ntua.gr

www.netmode.ntua.gr

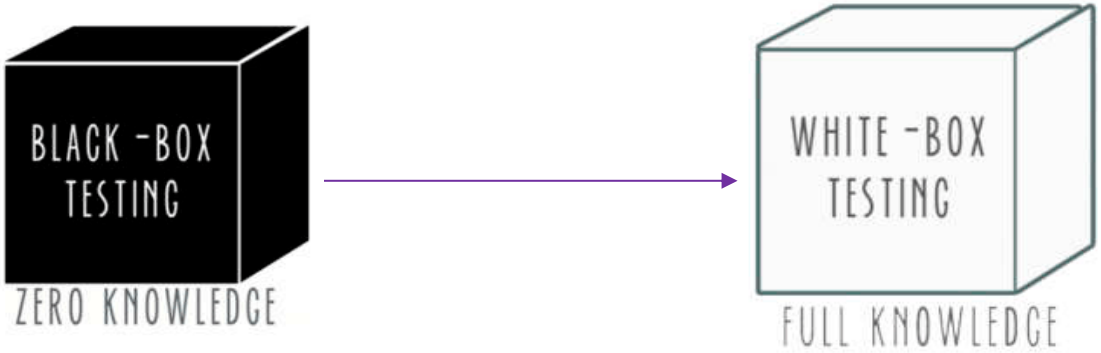
Room 002, New ECE Building

Wednesday May 21, 2025

# STOCHASTIC PROCESSES & OPTIMIZATION IN MACHINE LEARNING

## Reasons for eXplainable Artificial Intelligence (XAI) (1/2)

https://link.springer.com/book/10.1007/978-3-030-68640-6



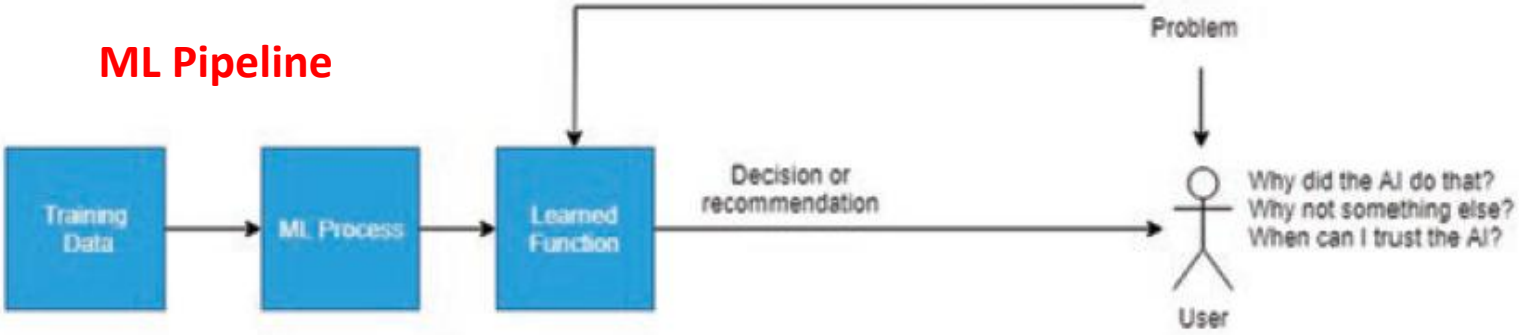

- The ***accuracy*** may not be a sufficient criterion to select an ***Artificial Intelligence*** (**AI**) - **Machine Learning** (**ML**) model
- Distrust of users - developers - analysts – regulators in uninterpretable ***black-box*** decision systems
- Need for ***eXplainable Artificial Intelligence*** (**XAI**) methods to justify decision making criteria and parameter tuning algorithms to designers and users of **AI** - **ML** models
- Requirement for ***reasoning*** of **AI** - **ML** replies to users – clients on personalized matters (***local interpretations***)
- Need for ***comparative evaluation*** of input ***features*** and ***justification*** of model selection, parameter/hyperparameter ***tuning*** etc.
- Feature engineering tools that assess the relative ***importance*** and detect potential ***correlations*** of ***sample data,*** with user-friendly graphics add-ons

# STOCHASTIC PROCESSES & OPTIMIZATION IN MACHINE LEARNING

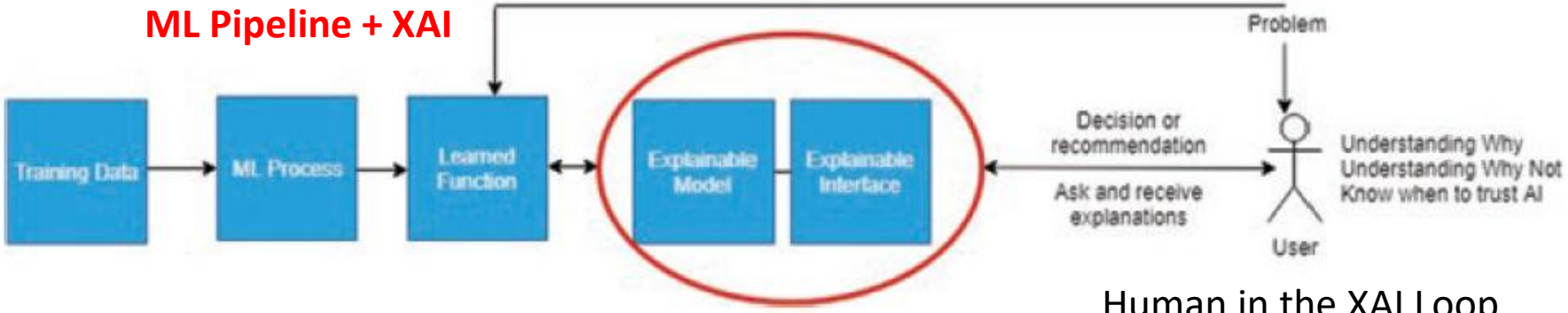
## Reasons for eXplainable Artificial Intelligence (XAI) (2/2)

https://link.springer.com/book/10.1007/978-3-030-68640-6

**ML Pipeline**




**ML Pipeline + XAI**




Human in the XAI Loop

**XAI**

# STOCHASTIC PROCESSES & OPTIMIZATION IN MACHINE LEARNING

## Definitions of Interpretability & Explainability

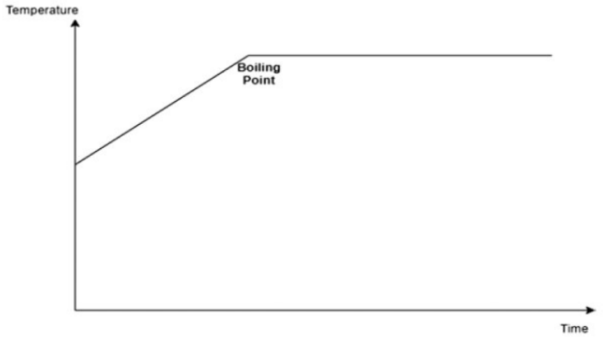
https://link.springer.com/book/10.1007/978-3-030-68640-6

- **Interpretability:** Possibility of understanding the mechanics of an ML model, but not necessarily knowing why
- **Explainability:** Understanding why

| Question | Interpretability | Explainability |
|---|---|---|
| Which are the most important features that are adopted to generate the prediction or classification? | ✓ | ✓ |
| How much the output depends on small changes in the input? | ✓ | ✓ |
| Is the model relying on a good range of data to select the most important features? | ✓ | ✓ |
| What are the criteria adopted to come across the decision? | ✓ | ✓ |
| How would the output change if we put different values in a feature not present in the data? | ✗ | ✓ |
| What would happen to the output if some feature or data had not occurred? | ✗ | ✓ |

**Explainability ⟶ Interpretability**
**(but not the opposite)**

- Model describing the process of boiling water
- Task: Predict temperature given time




Temperature can be predicted, but the physics of the boiling point cannot be directly explained

**Example of Context Difference:**
Boiling Water Model
- Interpretable: YES
- Explainable: NO

# STOCHASTIC PROCESSES & OPTIMIZATION IN MACHINE LEARNING

## Taxonomy of XAI Techniques

https://link.springer.com/book/10.1007/978-3-030-68640-6

- **Intrinsic vs. Post-hoc Models**
  - **Intrinsic**: Interpretations are **integrated** within the ***learning*** phase (e.g. weight determination in Linear Regression, Gini Index used to configure a Decision Trees)
  - **Post-hoc**: Explanations follow learning and parameter tuning, within the test phase
- **Model Agnostic vs. Model Specific Post-hoc Models**
  - **Model Agnostic**: Explanations rely on the observed input - output elements of the model and not on its specific structure
  - **Model Specific**: Explanations refer to specific parameters of the model
- **Global vs. Local Explainability**
  - **Global**: Interpretations for the **whole** sample (all sample elements)
  - **Local**: Interpretations for a **specific** sample element

# STOCHASTIC PROCESSES & OPTIMIZATION IN MACHINE LEARNING

## Taxonomy of XAI Techniques

https://link.springer.com/book/10.1007/978-3-030-68640-6

- **Intrinsic vs. Post-hoc Models**
  - **Intrinsic**: Interpretations are **integrated** within the ***learning*** phase (e.g. weight determination in Linear Regression, Gini Index used to configure a Decision Trees)
  - **Post-hoc**: Explanations follow learning and parameter tuning, within the test phase
- **Model Agnostic vs. Model Specific Post-hoc Models**
  - **Model Agnostic**: Explanations rely on the observed input - output elements of the model and not on its specific structure
  - **Model Specific**: Explanations refer to specific parameters of the model
- **Global vs. Local Explainability**
  - **Global**: Interpretations for the **whole** sample (all sample elements)
  - **Local**: Interpretations for a **specific** sample element

**Example of Intrinsic XAI Model**: Predication of Smartphones Sales / Buyers Age

**LINEAR REGRESSION MODEL**

**Simple Interpretation, Problematic Accuracy**

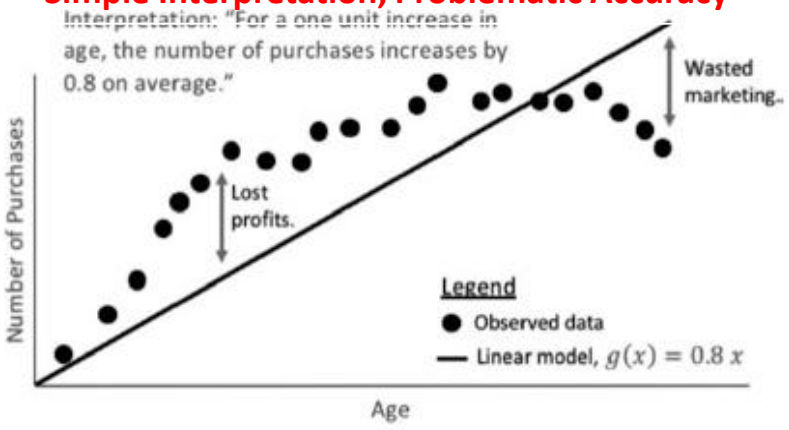



Fig. 2.6 A linear monotonic function gives simple, ready-to-go explanations with one global characteristic, the variation of purchases for one unit of age

**NON-LNEAR REGRESSION MODEL**

**Complex Interpretation – Requires Local Models**

**Improved Accuracy – Risk of Overfitting**

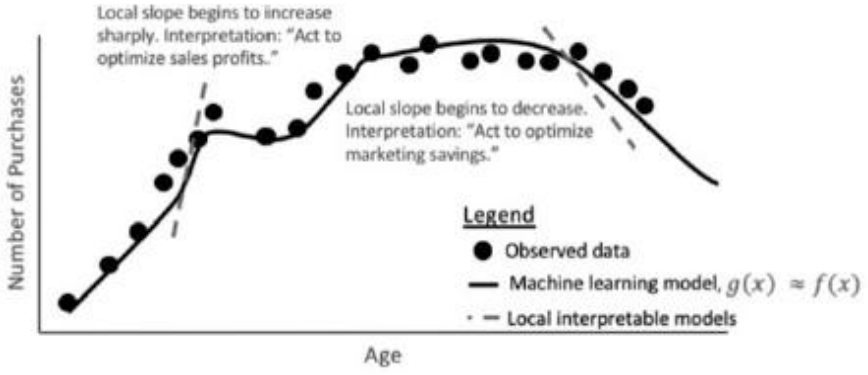



Fig. 2.7 With nonlinear non-monotonic function, we lose a global easily explainable model in favor of an accuracy improvement

# STOCHASTIC PROCESSES & OPTIMIZATION IN MACHINE LEARNING

## Model-Agnostic XAI Methods

https://link.springer.com/book/10.1007/978-3-030-68640-6

- **PI** (Permutation Importance)
- **SHAP** (Shapley Additive exPlanations)
- **LIME** (Local Interpretable Model agnostic Explanation)

# STOCHASTIC PROCESSES & OPTIMIZATION IN MACHINE LEARNING

## Permutation Importance - PI

https://link.springer.com/book/10.1007/978-3-030-68640-6

- Returns an ordered list of the most important ***features*** of the sample used to ***predict*** its behavior by an **AI** - **ML** model resulting from successive ***permutations*** of ***features*** of input sample elements

- It classifies sample ***features*** in ***Test*** sample elements (***Post-Hoc,* Global Explainability**) (not in elements of the ***Training Dataset***)

- It performs comparisons of the model ***output*** by ***reshuffling*** the order of ***features*** in an ***input vector*** of the ***test*** dataset. If a significant deviation is observed, the feature is deemed as ***significant***

- It does not compare ***feature values***, it just classifies them depending on their impact to the model output (***prediction***)

- It does not reveal ***correlations*** of features

- The influence of a ***feature*** classified as important in the quest of a model prediction can be illustrated in ***Partial Dependence Plot*** - ***PDP***

**NOTE**: The prompt classification of ***features*** **ALSO** in the ***Training Dataset*** can expedite training of **ML** κmodels and reduce the ***overfitting risks*** by deleting ***non-essential*** characteristics (or with negative values) during ***regularization*** of ***datasets***

# STOCHASTIC PROCESSES & OPTIMIZATION IN MACHINE LEARNING

## Shapley Additive exPlanations – SHAP (1/4)

https://link.springer.com/book/10.1007/978-3-030-68640-6

- The **SHAP** method aims at explaining the ***post-hoc*** influence of ***features*** of ***input sample points*** to the output of an **AI** - **ML** model. As such it can also be used as a preprocessor in ***feature engineering*** methods

- It refers to specific input sample points (***local prediction***) after the ***training*** process (***post-hoc***), based on input- output data and not relying on the model structure (***model agnostic***)

- It depends on ***Shapley*** values, introduced by ***Lloyd S. Shapley*** in 1953 https://en.wikipedia.org/wiki/Lloyd_Shapley to analyze the contributions of $M$ players in ***stochastic cooperative games***

# STOCHASTIC PROCESSES & OPTIMIZATION IN MACHINE LEARNING

## Shapley Additive exPlanations – SHAP (2/4)

https://link.springer.com/book/10.1007/978-3-030-68640-6

- The $M$ ***Players*** correspond to $M$ ***features*** contributing to the ***prediction*** of the output $y$ of an **ML** system

- Contributions are evaluated based on $y = f_X(S)$ for input elements $\mathrm{x} \in X$ with ***features*** in the subset $S$. The functions $f_X(S)$ are estimated via a ***simulated*** training process for the sample subset $X$ (***background set***) in a pre-tuned **ML** model (***post-hoc***)

- If $\varphi_i$ is the contribution of ***feature*** $i$ (***Shapley Value***), the total contribution of all $\sum_{i=0}^{M} \varphi_i$ (with $\varphi_0 = constant$) needs to be apportioned in the $M$ ***features*** proportionally to ***Shapley Values*** (we assume that $\varphi_i$ are additive and monotonic regarding their contribution)

- Evaluation of ***Shapley Values*** for contribution of ***feature*** $i$ in an element of the ***test dataset***:

$$\varphi_i = \sum_{S \sqsubseteq N\{i\}} \frac{(M - |S| - 1)!\,|S|!}{M!} [f_X(S \cup \{i\}) - f_X(S)]$$

The sum is evaluated for all subsets $S$ of ***features*** that do not contain $i$, $f_X(S)$ is the model output with inputs the subset $S$, while $f_X(S \cup \{i\}) - f_X(S)$ is the contribution of ***feature*** $i$

The combinatorial factor $\frac{(M-|S|-1)!|S|!}{M!}$ denotes the number of combinations of ***feature*** subsets

# STOCHASTIC PROCESSES & OPTIMIZATION IN MACHINE LEARNING

## Shapley Additive exPlanations – SHAP (3/4)

https://link.springer.com/book/10.1007/978-3-030-68640-6

### SHAP Force Plot: Prediction of Team with Best Player in Uruguay – Russia Match

- Using the **ML** model ***Random Forest*** (post-hoc), predict the probability that the best player belongs to ***Uruguay*** (***local XAI***)
- Investigate the ***impact*** on the prediction of input ***features***, e.g. "**number of *Uruguay* goals**" of specific input element (***local XAI***)
- In **red** denote ***features*** with positive contribution, e.g. "**number of *Uruguay* goals = *3***"
- In **blue** ***features*** with negative contribution
- The ***size*** of positive or negative contribution is proportional to the ***length*** of the corresponding segment in the diagram
- The output (**predictive**) value $\mathbf{0.52}$ (probability that the best player belongs to ***Uruguay***) is the sum *of* **red** segments minus the sum of **blue** (a little higher than $\mathbf{0.50}$ of the absolute unpredictability!)

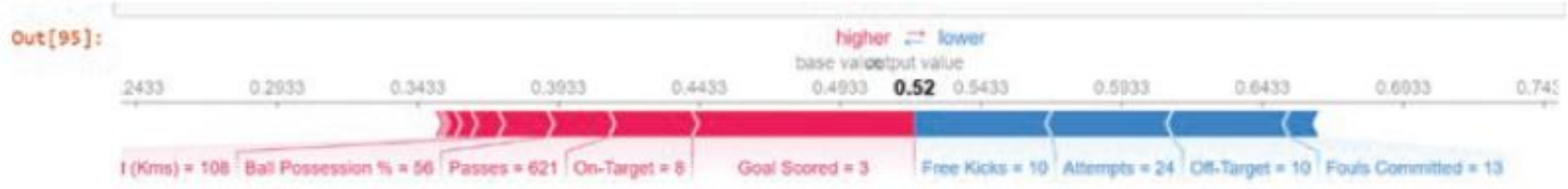



**Fig. 4.7** SHAP diagram that shows how the features impact on the match Uruguay-Russia. A force diagram representing how much the features change the final value. For example we see that "Goal Scored = 3" has the most impact for it pushes the final value to the right with the biggest interval (Becker 2020)

# STOCHASTIC PROCESSES & OPTIMIZATION IN MACHINE LEARNING

## Shapley Additive exPlanations – SHAP (4/4)

https://link.springer.com/book/10.1007/978-3-030-68640-6

**SHAP Summary Plot: Prediction of Team with Best Player in Uruguay – Russia Match**

**SUMMARY PLOT**

- Investigate the ***impact*** of ***features*** of multiple input elements in a tuned **ML** model via repeated ***local*** evaluations of ***SHAP Values*** (leading to ***global explainability***), with mean ***SHAP Value*** **0.0**
- Number of dots in the diagram: Number of sample elements in the ***test dataset***
- The **red** dots correspond to elements with ***large feature values.*** With **blue** elements of **small** ***feature values***

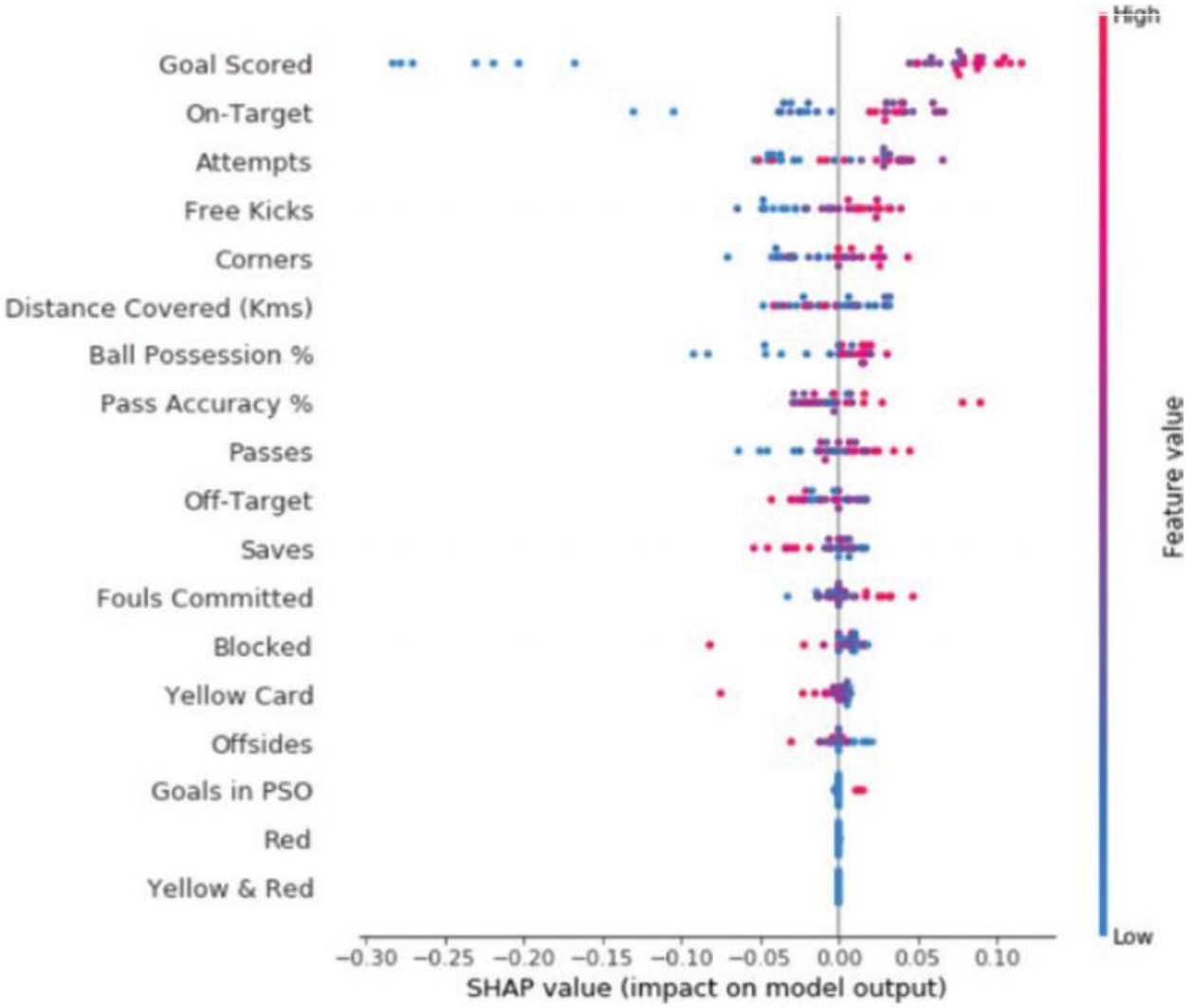



**Fig. 4.8** SHAP diagram that shows the features' ranking and the related impact on the match prediction (Becker 2020)

# STOCHASTIC PROCESSES & OPTIMIZATION IN MACHINE LEARNING

## Local Interpretable Model-agnostic Explanation - LIME

https://link.springer.com/book/10.1007/978-3-030-68640-6

- Modify the ***training*** sample by adding ***surrogate*** elements (e.g. ***Gaussian*** noise) for ***model agnostic*** equivalent classification systems, easily ***interpretable*** for a new ***test*** sample element
- The goal is to reduce dimensionality (number of ***features***) by deleting non-significant features and construct ***linearly separable*** prediction models (e.g. ***linear regression***) that can be ***intrinsically interpretable***
- The modified ***surrogate elements*** are assigned new weights that favor concentrations around ***new test*** elements, thus enforcing ***local interpretability*** and enabling easy ***linear classification*** algorithms, while reducing the influence of ***outliers***

In summary, the modified sample enables ***local interpretability*** with fewer ***features*** and satisfactory prediction ***accuracy***, provided that the ***surrogate*** elements do not exhibit significant deviations from the original sample elements (e.g. in terms of ***Mean Square Error***)

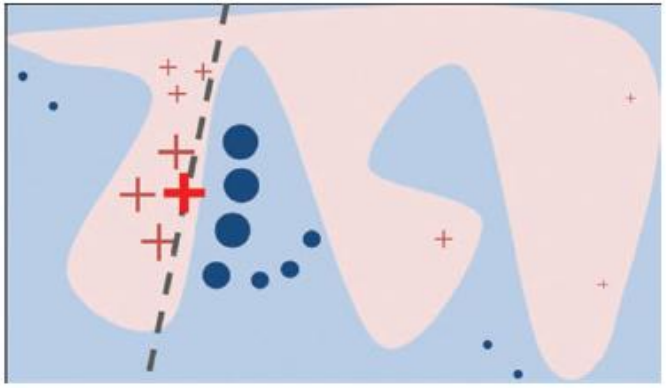


**Fig. 4.9** Schematics for LIME. The class outputs of the model are circles or crosses, and the dimension reminds us of the weight, so distant points are weighted less (Ribeiro et al. 2016)

- Add to the proximity of a ***test*** element (+ bold red cross) ***surrogate*** elements, with distance proportional weights, so as to reach ***linear separability*** in its neighborhood between two classes: **Crosses** and **Circles**
- The ***surrogate model*** is ***intrinsically explainable***, at least in the neighborhood of + (***local* XAI**)