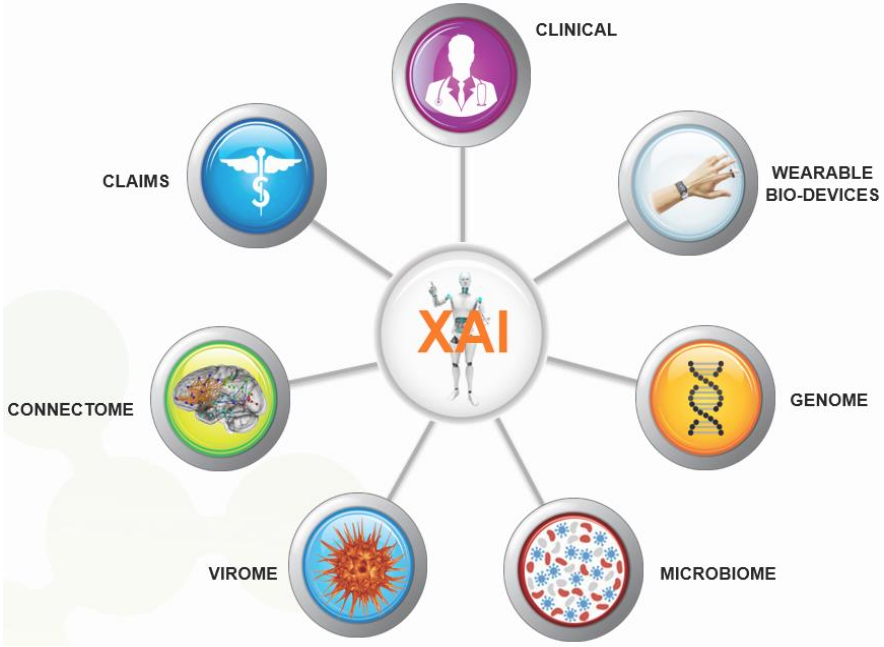
Introduction to XAI methods for medical data analysis

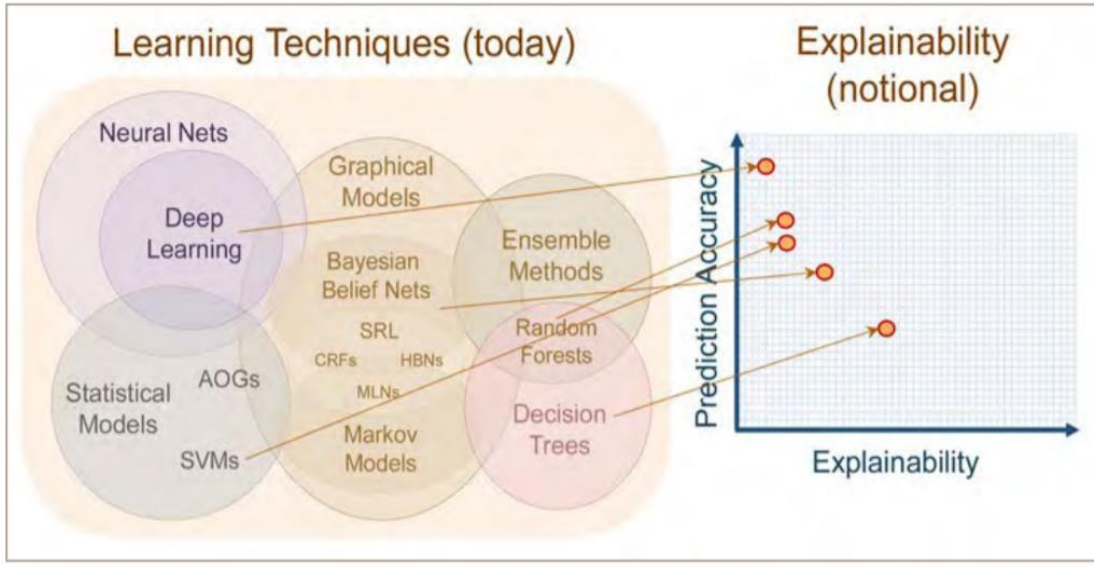


Angela Lombardi, PhD

Dipartimento Interateneo di Fisica – Università degli Studi di Bari e Istituto Nazionale di Fisica Nucleare – Sezione di Bari

Next-AIM Kickoff Meeting, February 2022

We need to explain AI/ML models



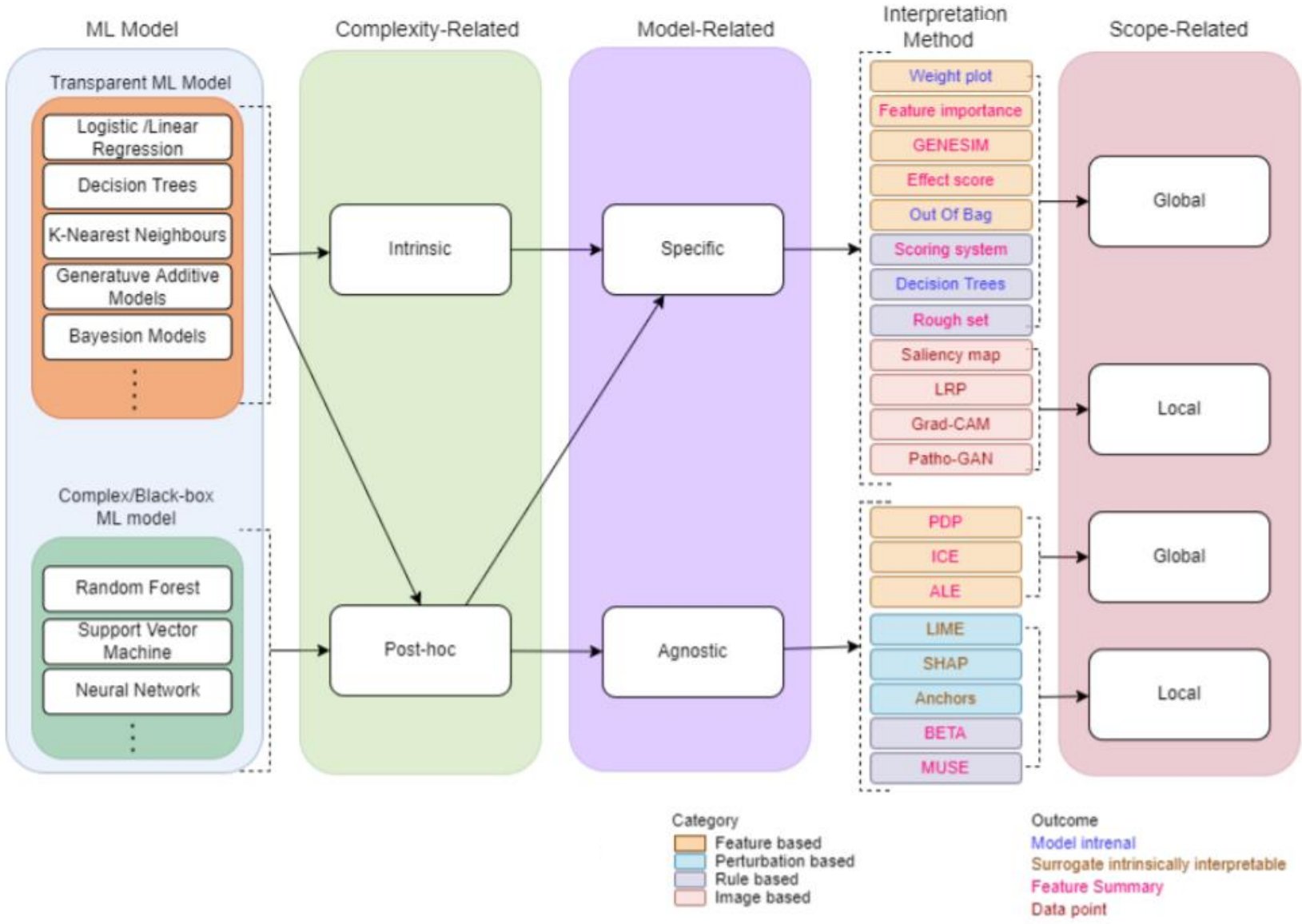
Source: DARPA

- The revolution in industrial technology using AI/ML proves the **great success of ML** and its applications in analyzing complex patterns, which are presented in a variety of applications in a wide range of sectors, including healthcare.
- The **best performance models belong to very complex or ensemble models** that are very difficult to explain.
- A huge and increasing number of issues are being addressed, including the **negative aspects of automated applications such as possible biases and failures**.
- In April 2019, the European Commission High-Level Expert Group on AI presented "**Ethics Guidelines for Trustworthy Artificial Intelligence**" and three of the guidelines directly refer to explainability.

Explainability (or interpretability)

- "Interpretability in machine learning is **a degree to which a human can understand the cause of a decision from an ML model**".
- It can also be defined as "**the ability to explain the model outcome in understandable ways to a human**".
- "The **use of machine learning models to extract specific data-contained information of domain relationships**".
- Gilpin et al. describe the primary purpose of interpretability as **being to effectively explain the model structure to users.**

Taxonomy of XAI



How to select a XAI algorithm?

- **Intrinsic or post hoc?** Interpretability is achieved by restricting the complexity of the machine learning model (intrinsic) or by applying methods that analyze the model after training (post hoc).
- **Results of the interpretation method** :

A) **Feature summary statistic**

B) **Feature summary visualization**

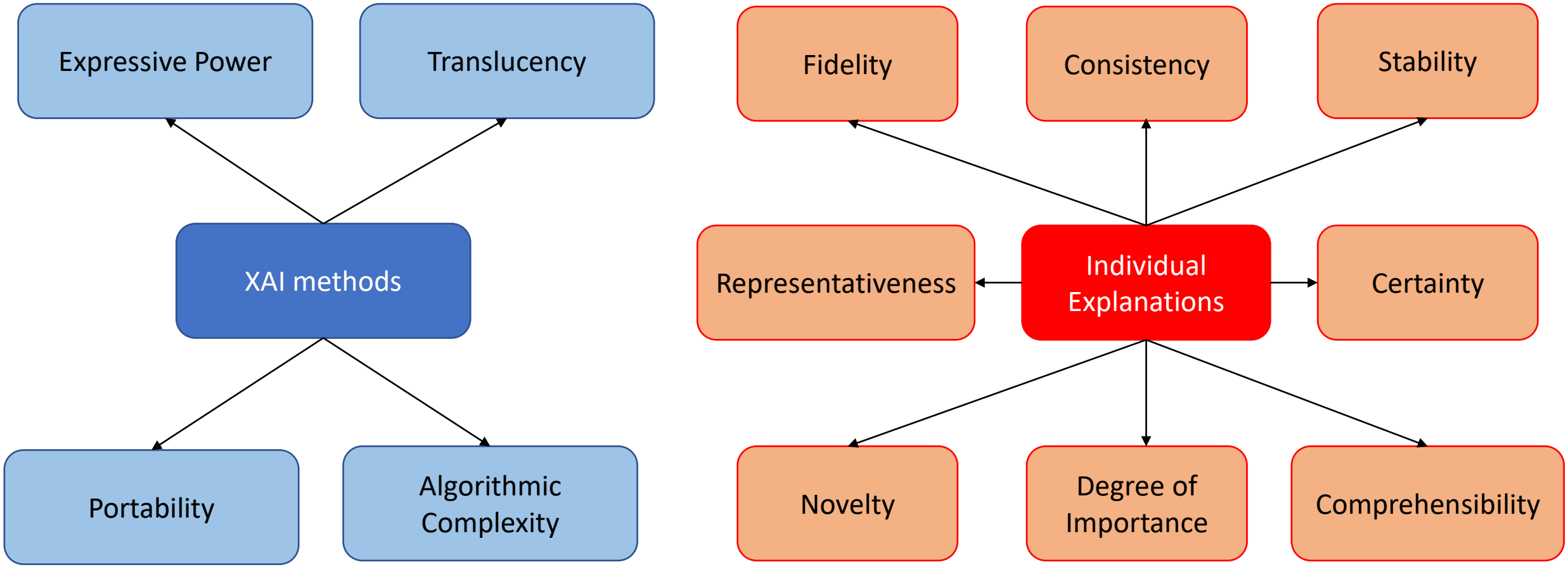
C) **Model internals** (e.g. learned weights)

D) **Data point** (for images and texts)

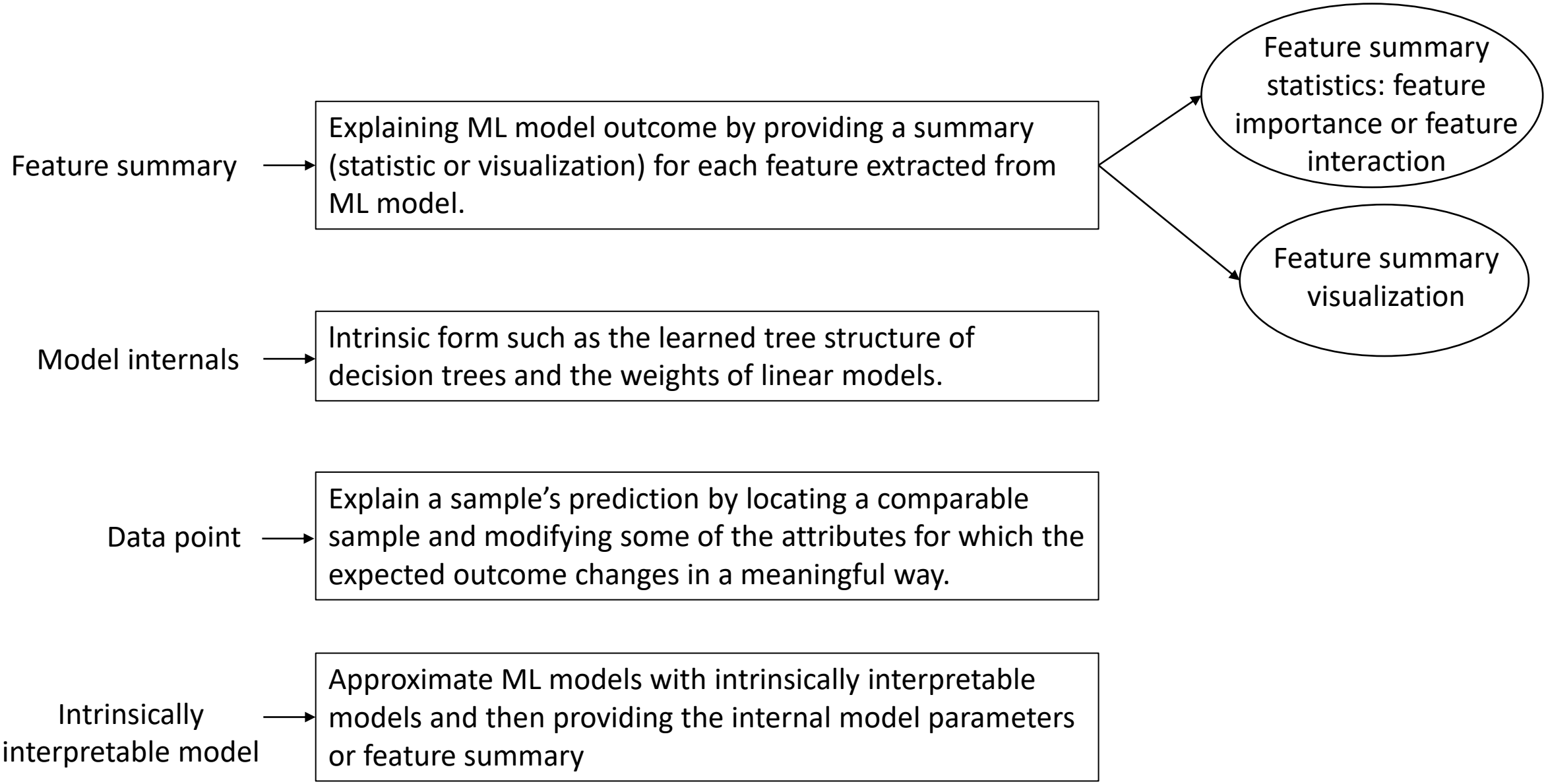
E) **Intrinsically interpretable model**: approximate black boxes (either globally or locally) with an interpretable model.

- **Model-specific or model-agnostic?**
- **Local or global?**

Properties of XAI



Outcomes of XAI



Interpretable ML models

Linear: if the association between features and target is modelled linearly.

Monotonicity constraints: the relationship between a feature and the target outcome always goes in the same direction over the entire range of the feature (an increase in the feature value either always leads to an increase or always to a decrease in the target outcome).

Interactions between features: by manually creating interaction features to predict the target outcome.

Algorithm	Linear	Monotone	Interaction	Task
Linear regression	Yes	Yes	No	regr
Logistic regression	No	Yes	No	class
Decision trees	No	Some	Yes	class,regr
RuleFit	Yes	No	Yes	class,regr
Naive Bayes	No	Yes	No	class
k-nearest neighbors	No	No	No	class,regr

Simple interpretable linear models

Linear regression

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$$

Interpretation of numerical Features

$x_k + 1 \rightarrow y + \beta_k$

Interpretation of a categorical features

x_k from 0 to 1 $\rightarrow$ $y + \beta_k$

Feature Importance

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

Assumptions:
Linearity, Normality, Homoscedasticity, Independence, Absence of multicollinearity

Non-Gaussian Outcomes

Generalized Linear Models (GLMs)

$$g(E_Y(y|x)) = \beta_0 + \beta_1 x_1 + \ldots \beta_p x_p$$

Interpretation of GLM weights

Depending on the assumed distribution E_Y together with the link function g

Interactions

Nonlinear Effects

Generalized Additive Models (GAMs)

$$g(E_Y(y|x)) = \beta_0 + f_1(x_1) + f_2(x_2) + \ldots + f_p(x_p)$$

Model functions as splines

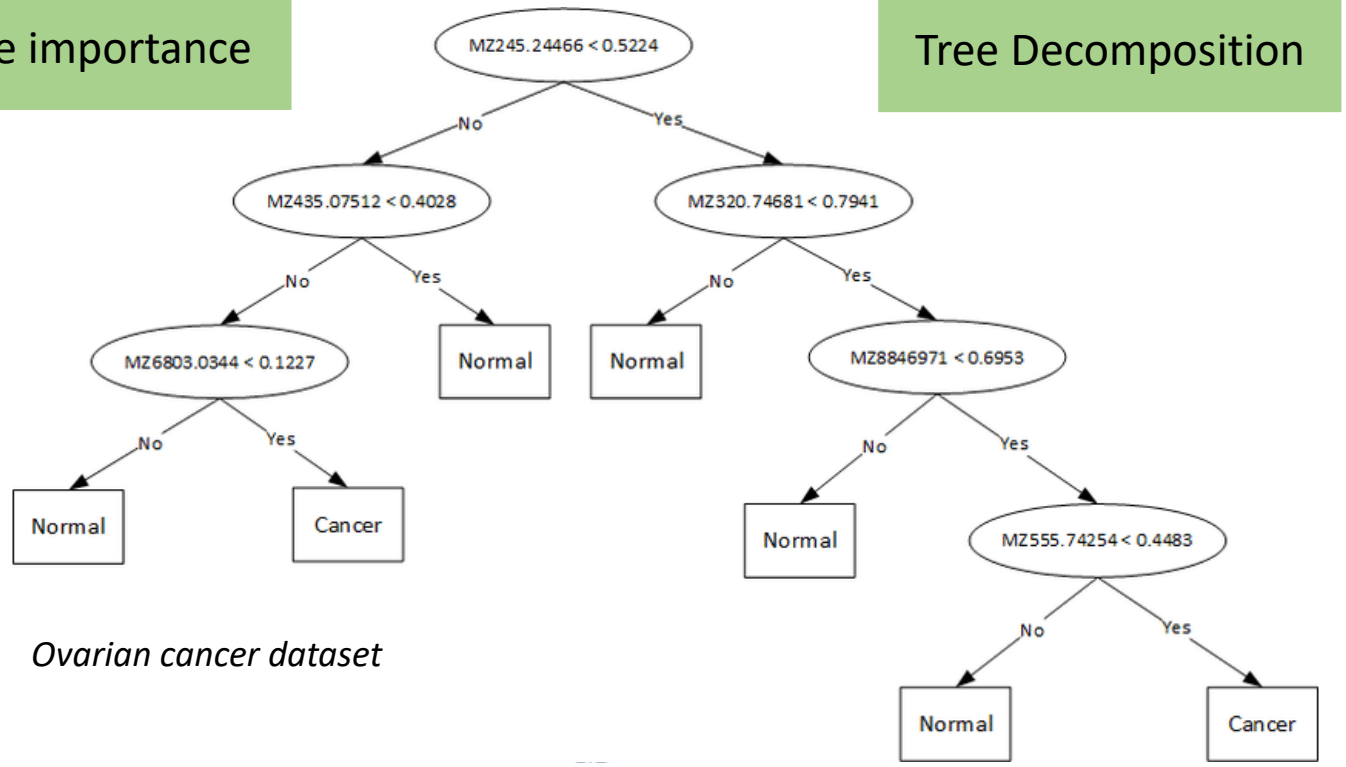
Estimates the spline weights

Decision trees – human-friendly explanations

1. Tree based models split the data multiple times according to certain cutoff values in the features.
2. Through splitting, different subsets of the dataset are created, with each instance belonging to one subset.
3. The final subsets are called terminal or leaf nodes and the intermediate subsets are called internal nodes or split nodes.
4. To predict the outcome in each leaf node, the average outcome of the training data in this node is used. Trees can be used for classification and regression.

Feature importance

Tree Decomposition



Ovarian cancer dataset

$$\hat{y} = \hat{f}(x) = \sum_{m=1} c_m I\{x \in R_m\}$$

$I_{\{x \in R_m\}}$ is the identity function

If an instance falls into a leaf node R_l, the predicted outcome is $\hat{y} = c_l$ where c_l is the average of all training instances in leaf node R_l.

Feature-based model-agnostic global interpretation techniques

- Global methods describe the **average behavior** of a machine learning model.
- Global methods are often expressed as **expected values** based on the distribution of the data.
- Since global interpretation methods describe average behavior, they are particularly useful when the modeler wants to understand the **general mechanisms in the data** or debug a model.

❖**A partial dependence plot** is a feature effect method.

❖**Accumulated local effect plots** is another feature effect method that works when features are dependent.

❖**Feature interaction** (H-statistic) quantifies to what extent the prediction is the result of joint effects of the features.

❖**Functional decomposition** is a central idea of interpretability and a technique that decomposes the complex prediction function into smaller parts.

❖**Permutation feature importance** measures the importance of a feature as an increase in loss when the feature is permuted.

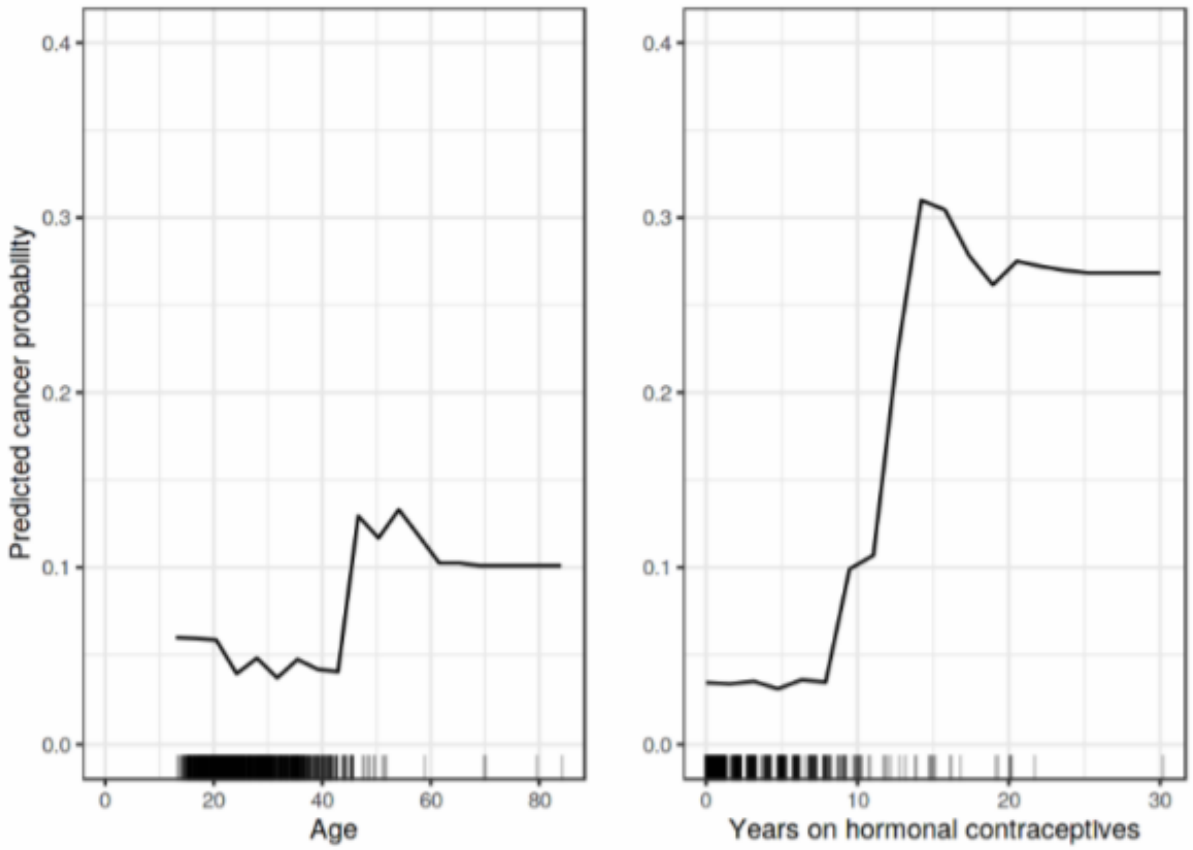
❖**Global surrogate models** replaces the original model with a simpler model for interpretation.

Partial Dependence Plot

The Partial Dependence Plot (PDP) is graphical representation that indicates th **marginal effect** of input variables b visualizing the **average partial correlatio between one or more features on an M model prediction outcome**. The PDP ca estimate if the relationship between th output and the feature is linea monotonous, or more complex.

$$\hat{f}_{x_S}(x_S) = E_{x_C}\left[\hat{f}(x_S, x_C)\right] = \int \hat{f}(x_S, x_C) d\mathbb{P}(x_C)$$

where x_S are the plotted features, x_C are th other features used in ML model $\hat{f}$



Accumulated Local Effects Plot

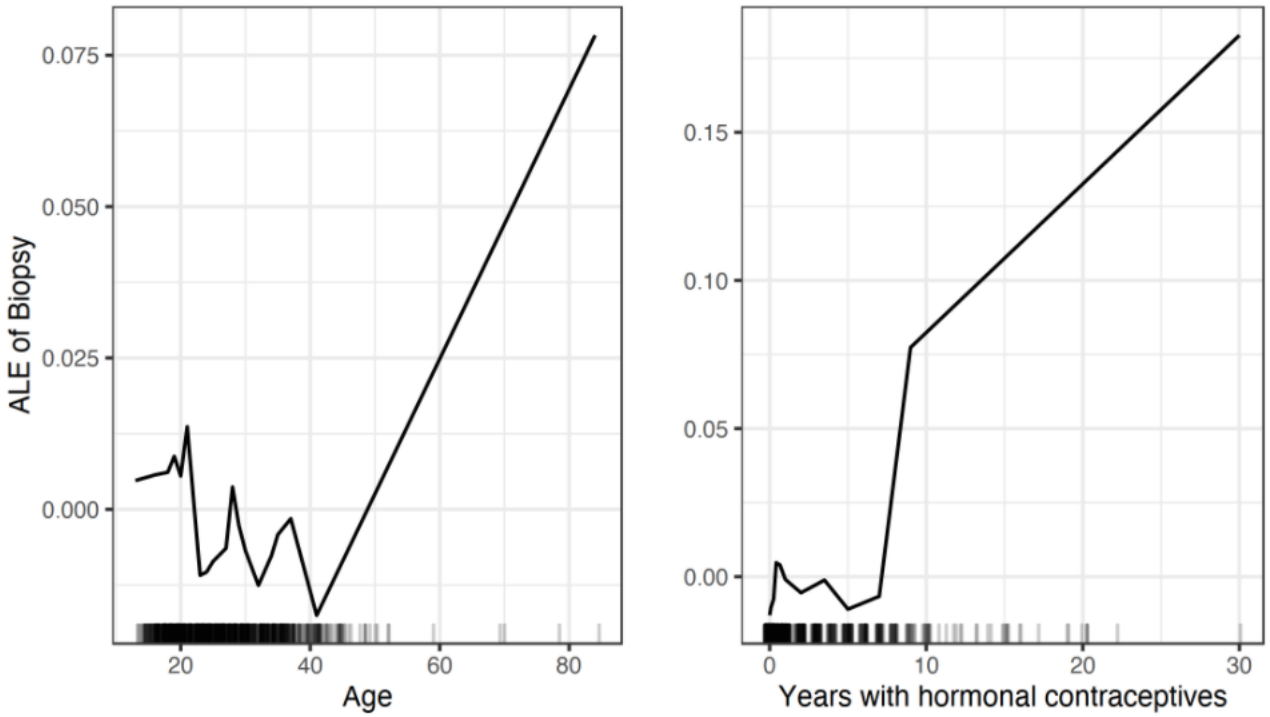
-Accumulated Local Effects plot (ALE) aims at **evaluating the interaction effects of predictors** in black-box ML models, which avoids the prior issues with PDP plots.

-ALE discusses **how characteristics impact the average prediction** of a machine learning model.

-It is a **more efficient and unbiased** alternative to partial dependency diagrams (PDPs).

-ALE plots average the changes in the prediction and accumulate them over the grid

$$\hat{f}_{x_S,ALE}(x_S) = \int_{z_{0,1}}^{x_S} E_{X_C|X_S}\left[\hat{f}^S(X_s, X_c) \mid X_S = z_S\right] dz_S - C$$



Permutation feature importance

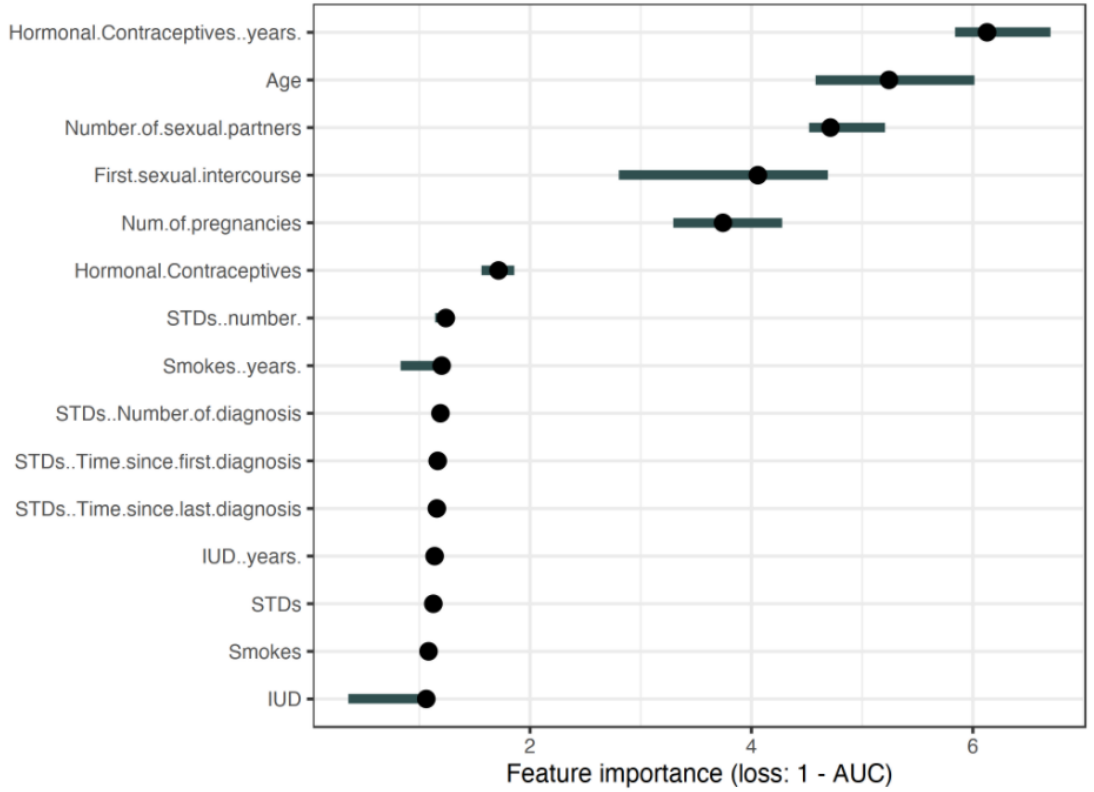
We measure the importance of a feature by calculating **the increase in the model's prediction error after permuting the feature**. A feature is "important" if shuffling its values increases the model error, because in this case the model relied on the feature for the prediction.

Input: Trained model $\hat{f}$, feature matrix X, target vector y, error measure $\mathrm{L}(y, \hat{f})$

1.Estimate the original model error $e_{origin} = L\left(y, \hat{f}(X)\right)$ (e.g. mean squared error)

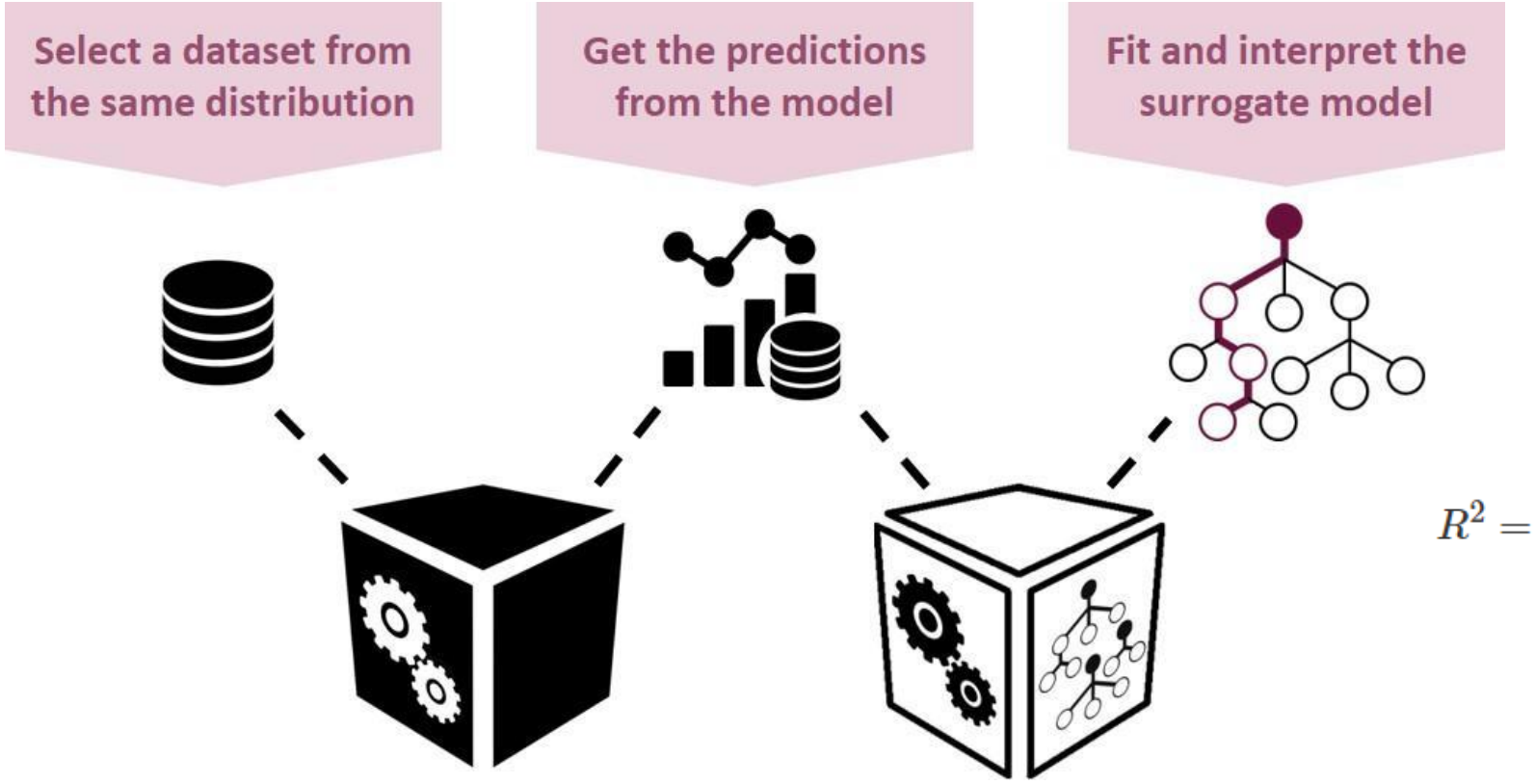
2.For each feature j∈{1,...,p} do:

1. Generate feature matrix X_{perm} by permuting feature j in the data X. This breaks the association between feature j and true outcome y.
2. Estimate $e_{perm} = L\left(y, \hat{f}\left(X_{perm}\right)\right)$ based on the predictions of the permuted data.
3. Calculate permutation feature importance as quotient $FI_j = e_{perm}/e_{origin}$ or difference $FI_j = e_{perm} - e_{origin}$.
4. Sort features by descending FI.



Global surrogate models

A global surrogate model is an interpretable model that is trained to approximate the predictions of a black box model. We can draw conclusions about the black box model by interpreting the surrogate model.



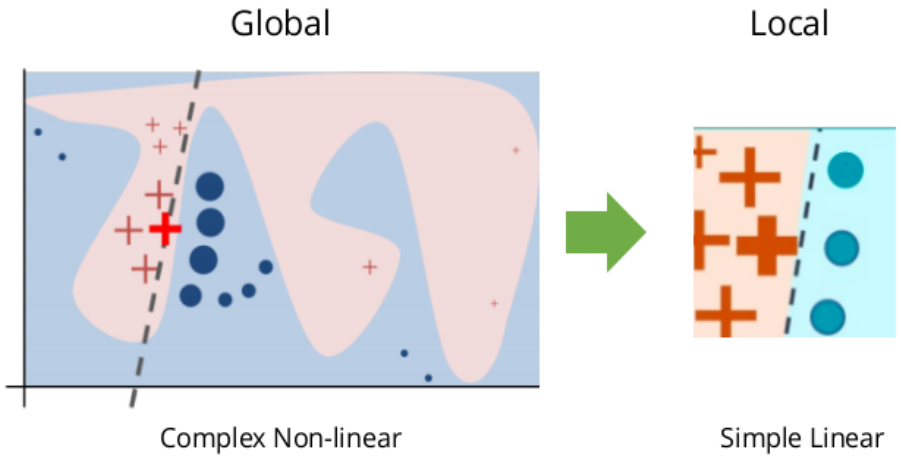
How well the surrogate model replicates the predictions of the black box model?

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_*^{(i)} - \hat{y}^{(i)})^2}{\sum_{i=1}^{n}(\hat{y}^{(i)} - \bar{\hat{y}})^2}$$

LIME: local interpretable model-agnostic explanations

Surrogate models are trained to approximate the predictions of the underlying black box model. Instead of training a global surrogate model, LIME focuses on training local surrogate models to explain individual predictions.

1.LIME generates a **new dataset consisting of perturbed samples** and the corresponding predictions of the black box model.
2. **On this new dataset LIME then trains an interpretable model**, which is weighted by the proximity of the sampled instances to the instance of interest.
3. **The learned model should be a good approximation of the machine learning model predictions locally**, but it does not have to be a good global approximation. This kind of accuracy is also called **local fidelity**.



$$\text{explanation}(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

SHAP: shapley additive explanations

- The SHAP method (Lundberg & Lee, 2017) derives local explanation models using the concept of **Shapley values from cooperative game theory**
- A SHAP explanation is a vector $\phi = (\phi_0, \phi_1 \ldots \phi_F)$ that assigns a feature importance ϕ_i to each input feature. Intuitively, the **input features** of a classifier are akin to **players cooperating to win a game** (the model prediction). The more important a player *i* is to the cooperation, the higher is its Shapley value $\phi(i)$. Features are grouped into ***coalitional sets***, corresponding to the power set of the set of features *F*.
- For a feature i∈F, its Shapley value ϕ_i is defined as follows:

$$\phi(i) = \sum_{S \subseteq \mathcal{F} \setminus \{i\}} \frac{|S|! \cdot (F - |S| - 1)!}{F!} \left(f_{S \cup \{i\}}\left(x_{S \cup \{i\}}\right) - f_S\left(x_S\right) \right)$$

- A linear local model g is computed as a linear regressor:

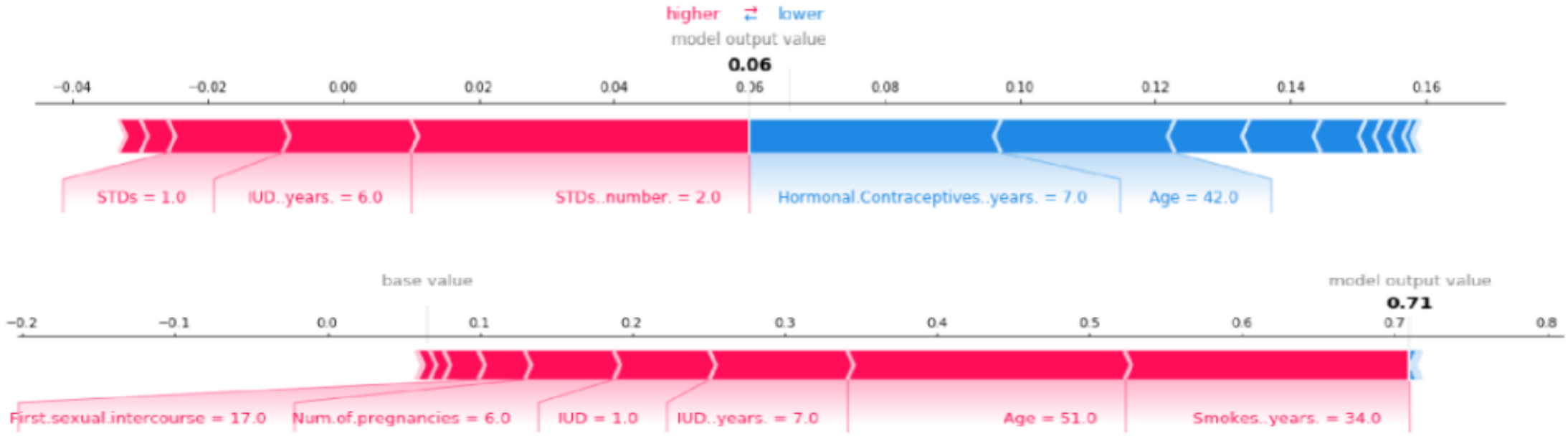
$$g(x) = w_0 + \sum_{i=1}^{F} w_i \cdot x_i \qquad w_0 = \phi_0, \qquad w_i = \frac{\phi_i}{x_i - \mu_i}, \qquad 1 \leq i \leq F, \mu_i \in \mathbb{B}_{\{i\}}$$

SHAP: force plots

-The Shapley values can be viewed as "**forces**": each feature value is a force that either increases or decreases the prediction.

-The prediction starts from the baseline. **The baseline for Shapley values is the average of all predictions.**

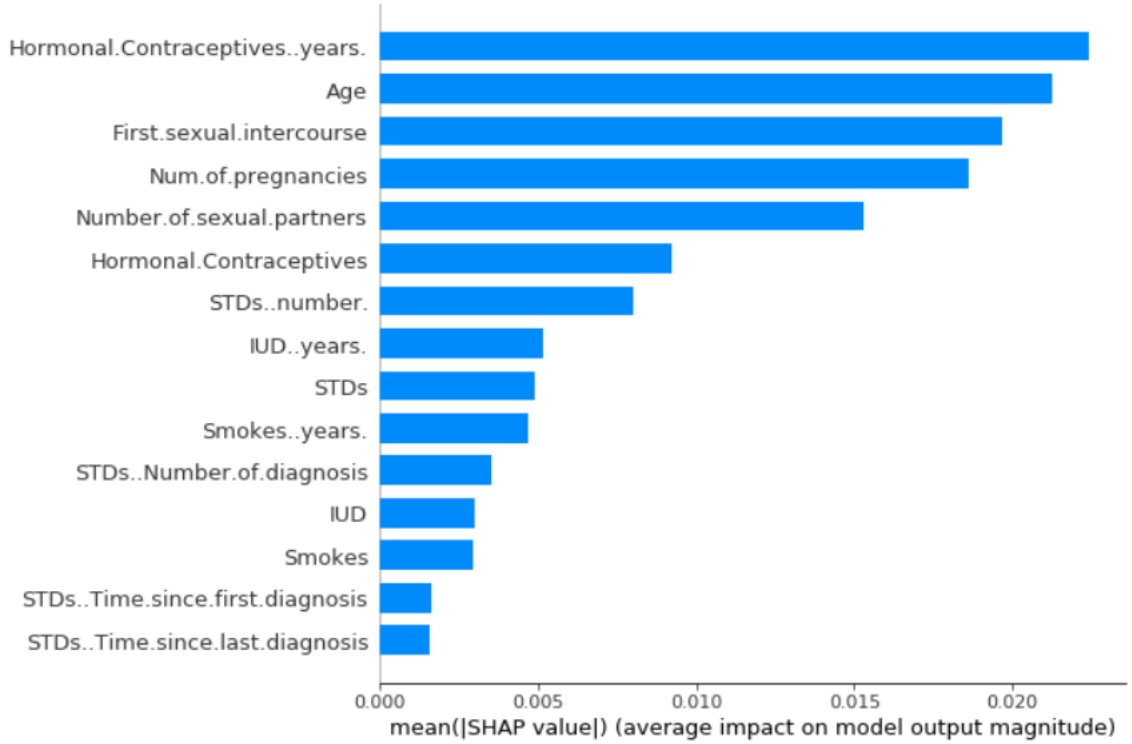
-In the plot, **each Shapley value is an arrow that pushes to increase (positive value) or decrease (negative value) the prediction.** These forces balance each other out at the actual prediction of the data instance.



SHAP: feature importance

The idea behind SHAP feature importance is simple: features with large absolute Shapley values are important. Since we want the global importance, we average the **absolute** Shapley values per feature across the data:

$$\frac{1}{n} I_j = \sum_{i=1}^{n} |\phi_j^{(i)}|$$

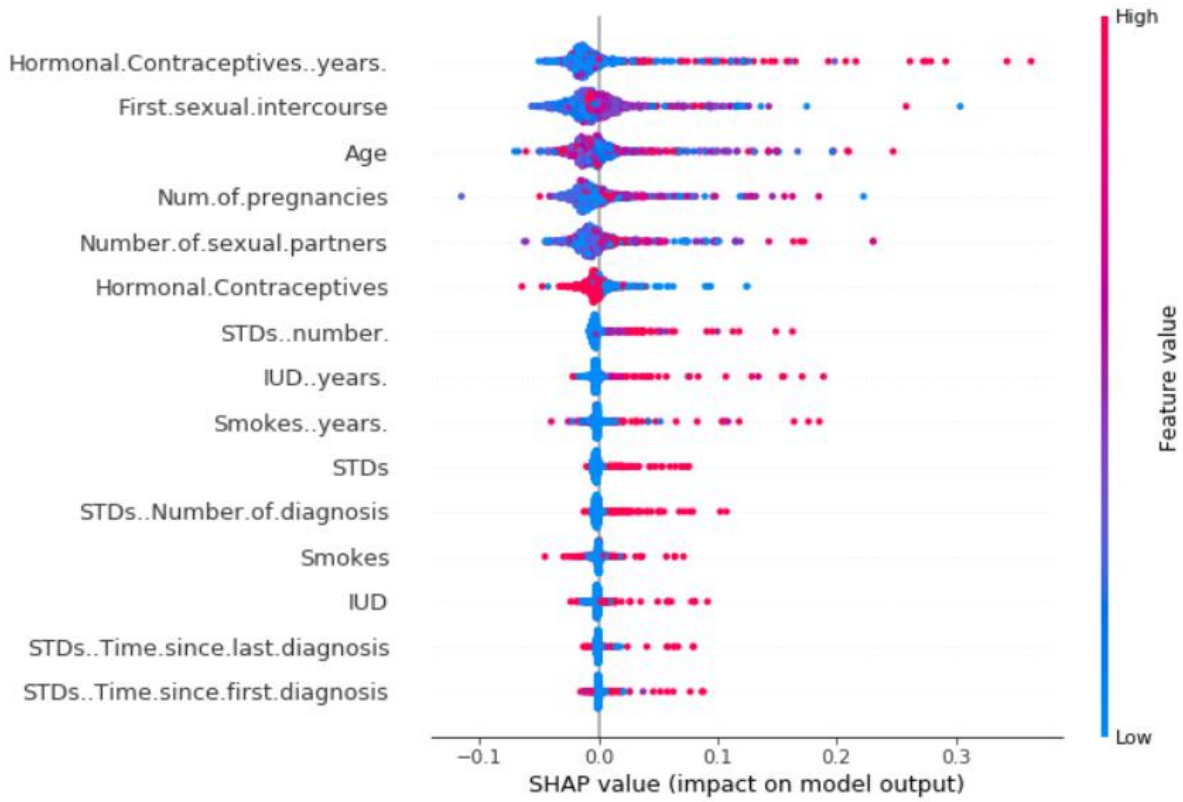


SHAP: Summary Plot

-The summary plot combines feature **importance with feature effects**.

-Each point on the summary plot is a **Shapley value for a feature and an instance**. The position on the y-axis is determined by the feature and on the x-axis by the Shapley value.

-**The color represents the value of the feature from low to high**. Overlapping points are jittered in y-axis direction, so we get a sense of the distribution of the Shapley values per feature. The features are ordered according to their importance.



SHAP: Dependence Plot

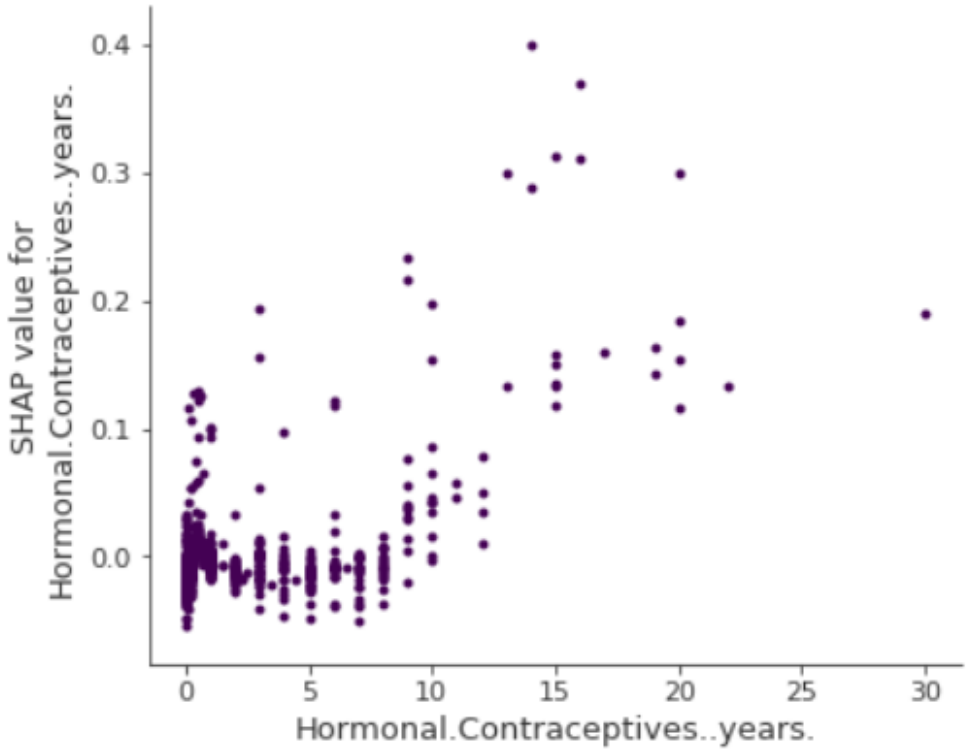
SHAP feature dependence might be the simplest global interpretation plot:

1) Pick a feature.

2) For each data instance, plot a point with the feature value on the x-axis and the corresponding Shapley value on the y-axis.

Mathematically, the plot contains the following points:

$$\{(x_j^{(i)}, \phi_j^{(i)})\}_{i=1}^{n}$$



Case study: predicting brain age with ML/DL

- The last few decades have seen significant advances in neuroimaging methodologies and machine learning (ML) techniques focused on identifying structural and functional features of the brain associated with the age.

- Age prediction is typically performed using **a multivariate set of features derived from one or multiple imaging modalities.** A dataset is then specified by including the characteristics of different subjects and their chronological ages.

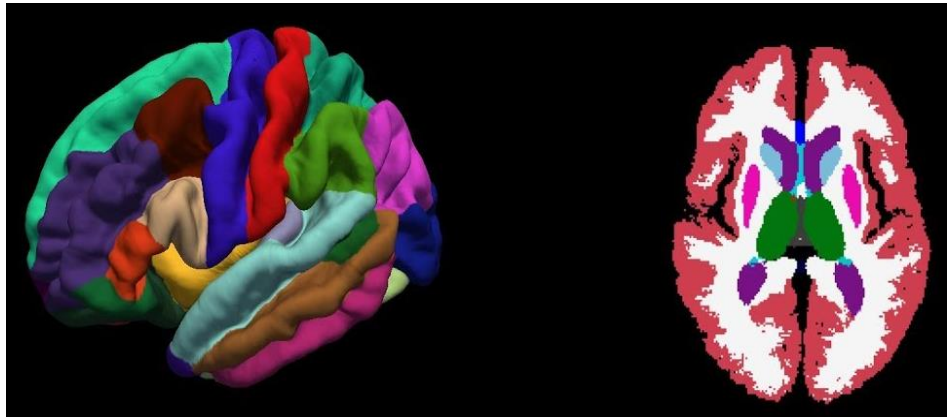
- The dataset is employed to train one or more **supervised machine learning** algorithms which attempt to **predict a given subject's brain age by using the brain imaging features** while minimizing the difference from the true age and preventing overfitting.

A. Lombardi, et al. "Explainable Deep Learning for Personalized Age Prediction With Brain Morphology." Frontiers in neuroscience 15 (2021).

Dataset

378 MALE CONTROL subjects from 17 sites (ABIDE I DATASET) Age range 6-48; mean=17; std=7;

P=1213 morphological features resulting from recon-all FreeSurfer pipeline:



DESIKAN ATLAS
34 ROIs for hemisphere

ASEG ATLAS
40 ROIs

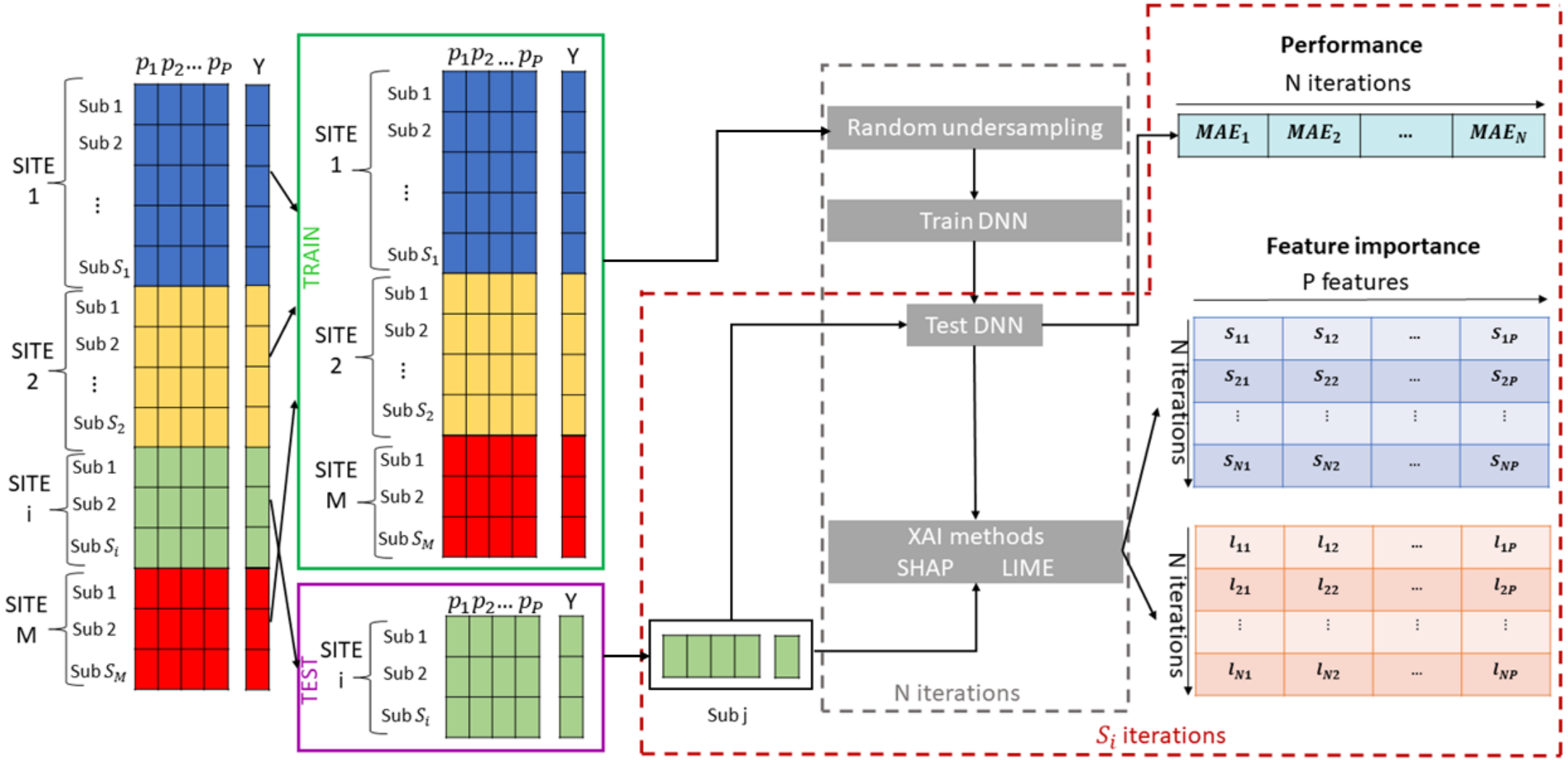
•**Volume, intensity mean, standard deviation, minimum, maximum**, and range of **40 sub-cortical brain structures** and white matter parcellation of brain cortex;

•**volume, surface area, Gaussian curvature, mean curvature, curvature index, folding index, thickness mean, and thickness standard deviation** for the **34 cortical brain regions of each hemisphere**;

•**global brain metrics**, including surface and volume statistics of each hemisphere; total cerebellar gray and white matter volume, brainstem volume, corpus callosum volume, and white matter hypointensities.

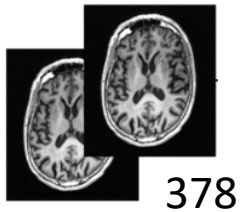
Di Martino, Adriana, et al. "The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism." *Molecular psychiatry* 19.6 (2014): 659.

Brain age prediction



Age Prediction – Computational Framework

FreeSurfer



T1

378 suggetti

preprocessing_pipeline.sh

24 h per subject

specs

4 CPU
8 Gb RAM
subject_name

Preprocessing_pipeline.specs

HTCondor

sub_1.job
sub_2.job

sub_378.job

Age	f_1	f_2	...	f_{1213}
13				
46				
23				
8				

features

DNN_XAI_pipeline.sh

6 h per subject
Train 100 models
for each subject

specs

4 CPU
16 Gb RAM
subject_name

DNN_XAI_pipeline.specs

HTCondor

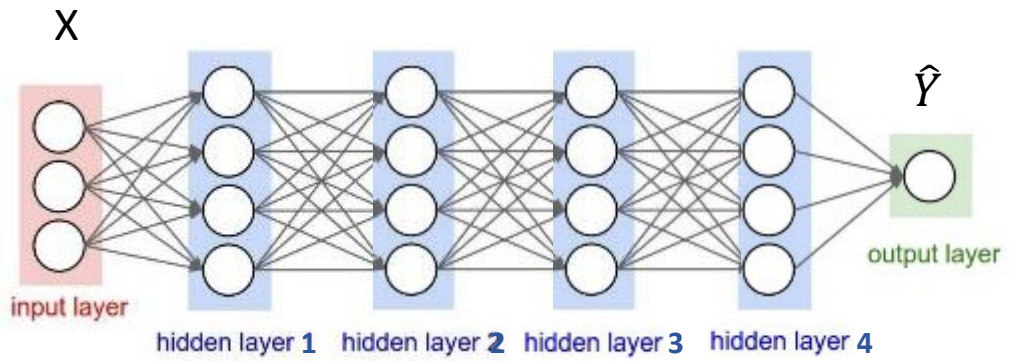
sub_1.job
sub_2.job

sub_378.job

100 predictions per subject
100 XAI scores per subject

Deep Learning models

- A fully connected Deep Neural Networks models to predict the brain age Y of a healthy cohort of subjects by using their morphological features X;
- each model with 10-fold Grid Search cross validations on training sets, using the left out site as a completely independent test set.
- A final configuration with **4-layers with 512 units per layer**, **RELU as activation function**, the **SGD optimizer** with learning rate $5e^{-5}$ and momentum 0.9, the **loss function Huber** and dropout 0 was obtained at the end of the Grid Search.



Performance metrics

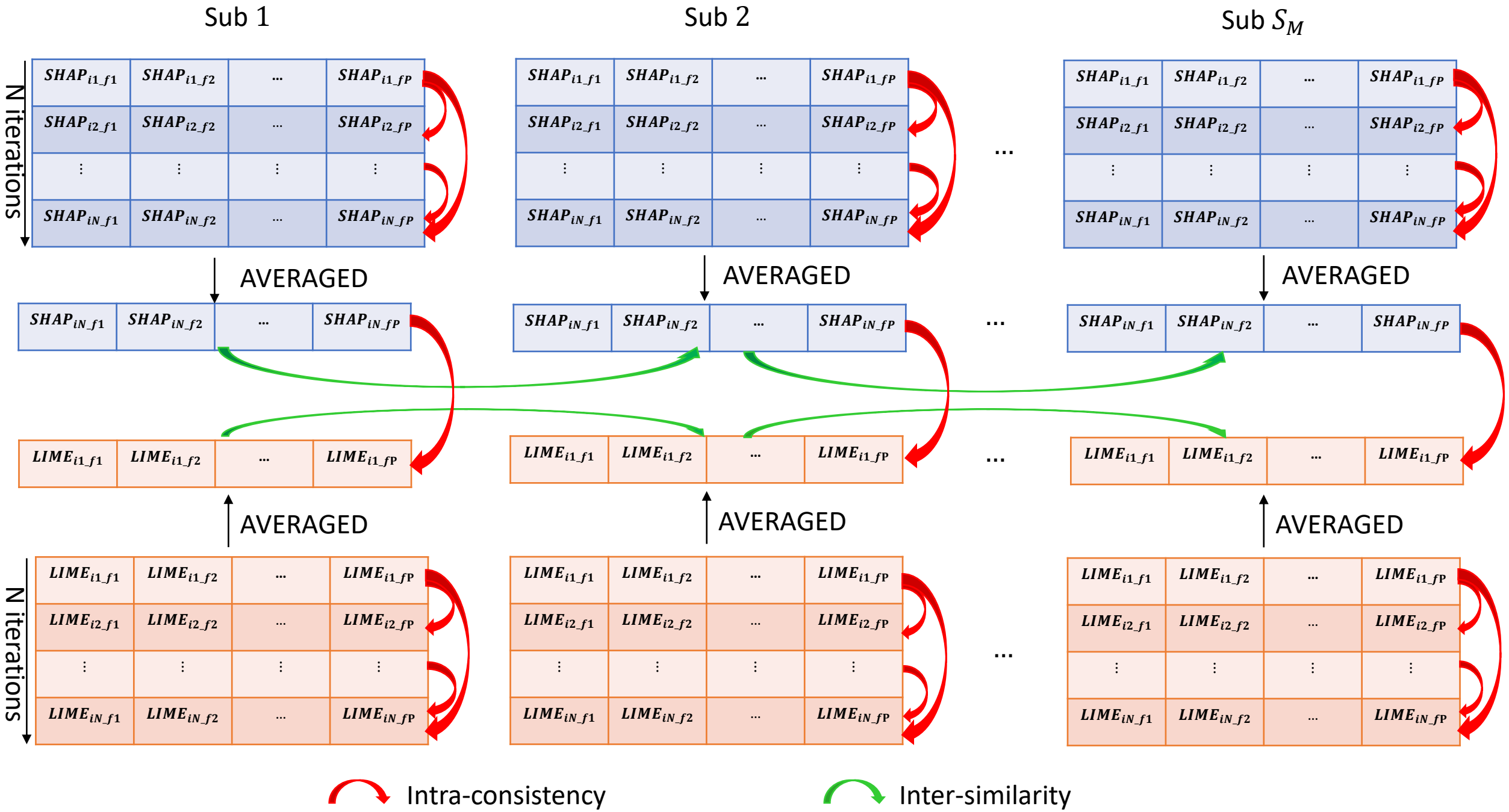
$$MAE = \frac{1}{t}\sum_{i=1}^{t}|y_i - \hat{y}_i|$$

$$R = \frac{\sum_{i=1}^{T}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{T}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{T}(\hat{y}_i - \bar{\hat{y}})^2}}.$$

Quantify the variability of XAI scores

Reliability of XAI models to explain local «subject-level» decisions:

intra-consistency: by varing the training set, how do the local scores concerning the individual subject vary?
inter-similarity: by varing the training set, how do the local scores vary across subjects?


Sub 1
Sub 2
Sub S_M
N iterations
AVERAGED
AVERAGED
AVERAGED
AVERAGED
AVERAGED
AVERAGED
N iterations
Intra-consistency
Inter-similarity

Stability of XAI scores

- Local XAI methods produce a feature importance vector for each test sample. A stability analysis of XAI scores is required to quantify the variation of the score values by slightly varying the training set.
- We applied both SHAP and LIME algorithms to extract the age-related feature importance vector for each subject collecting the two matrix S and L of dimension [NXP] whose generic element s_{nk} (l_{nk})indicates the SHAP (LIME) value for the k feature within the n iteration.
- We investigated the reliability of both SHAP and LIME values by computing the **intra-consistency coefficient** of the scores, i.e., the correlation between each couple of score vectors:

$$IC_{kz} = \frac{\sum_{p=1}^{P}(s_{kp} - \bar{s_k})(s_{zp} - \bar{s_z})}{\sqrt{\sum_{p=1}^{P}(s_{kp} - \bar{s_k})^2}\sqrt{\sum_{p=1}^{P}(s_{zp} - \bar{s_z})^2}}$$

- The intra-consistency coefficient varies between 0 (zero) and 1 (one), hence we compared the IC distributions between the two XAI methods by using the Wilcoxon rank-sum test and Cohen's d coefficient in order **to choose the most reliable and stable algorithm**.

Variability of XAI scores across subjects

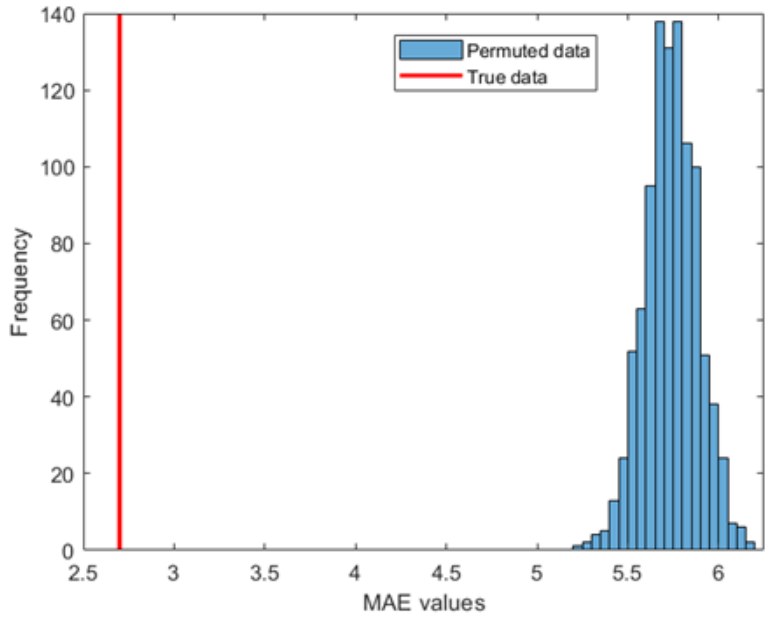
- For the best algorithm, we averaged the N= 100 realizations of both values to obtain a single representative XAI vector;
- The **inter-subject similarity** was firstly computed as the correlation between the SHAP (LIME) score vectors s_t and s_u (l_t and l_u) for each couple of subjects t and u, with t,u=1,…,T:

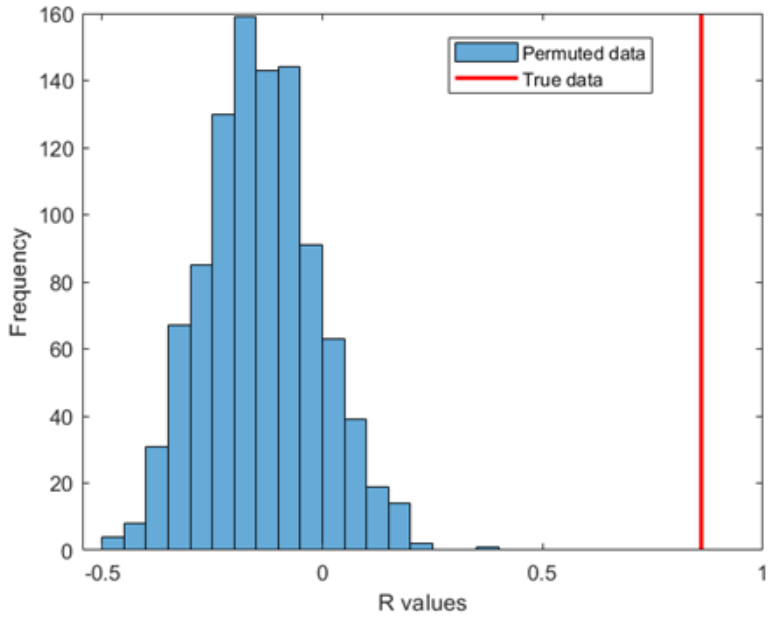
$$IS_{ut} = \frac{\sum_{p=1}^{P}(s_{u,Ap} - \bar{S_u})(s_{t,Ap} - \bar{S_t})}{\sqrt{\sum_{p=1}^{P}(s_{u,Ap} - \bar{S_u})^2}\sqrt{\sum_{p=1}^{P}(s_{t,Ap} - \bar{S_t})^2}},$$

- Then, the **inter-similarity matrix IS** was obtained for the best XAI method, where the entry (u,t)=$IS_{u,t}$ indicates the similarity value between the scores of subjects u and t.
- We applied the **stability-based k-medoid criterion** to find the best partition into **clusters of the inter-similarity matrix.** This criterion assesses the **clusterwise stability of a dataset** by resampling it several times with different methods such as bootstrap or subsetting and by identifying the most stable clusters across the iterations.

Results: DL models for brain age prediction

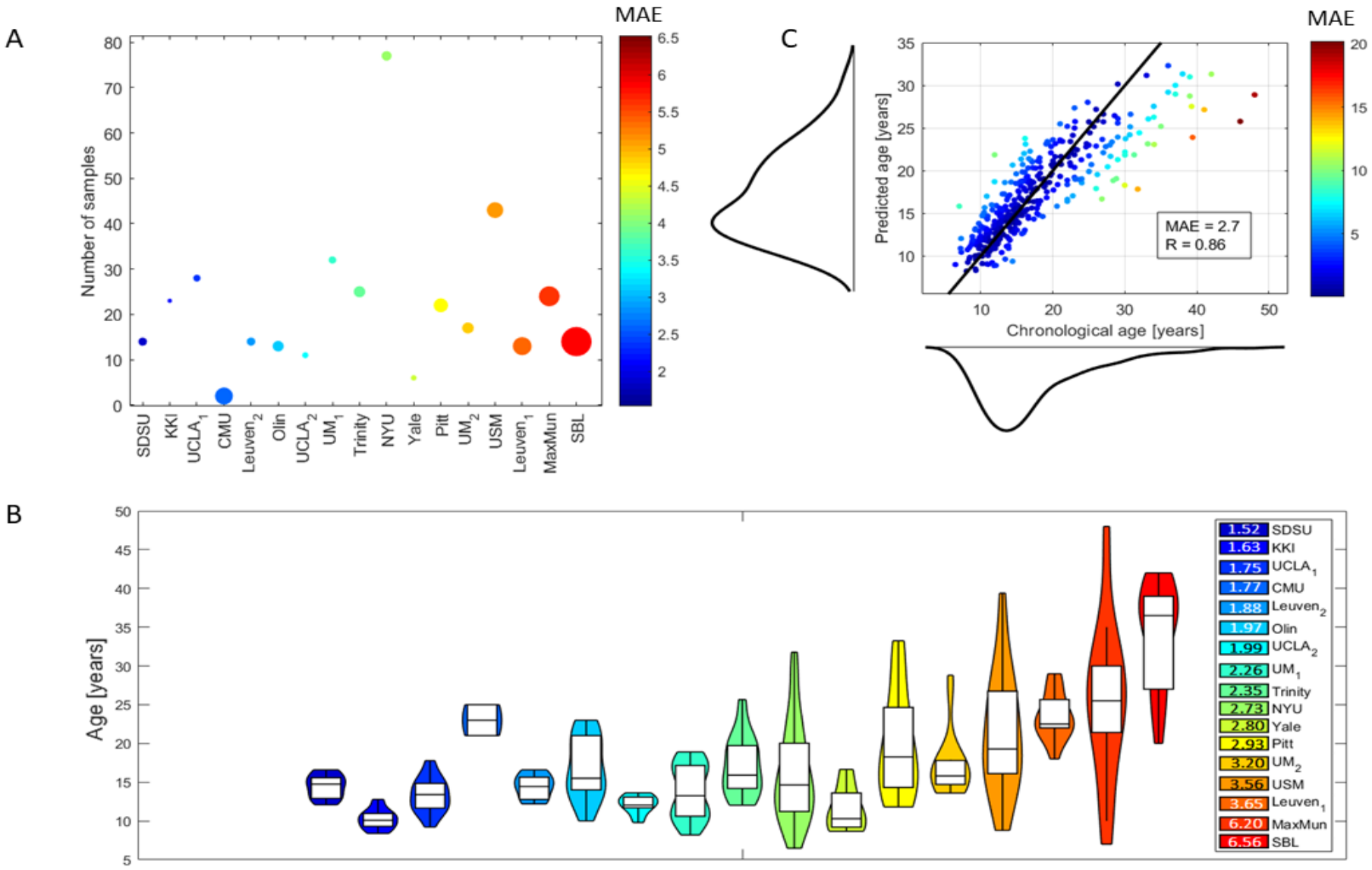
Predictive performance (MAE and R) based on permuted data (1.000 permutations) in relation to the predictive performance based on the true non-permuted data (red vertical line).



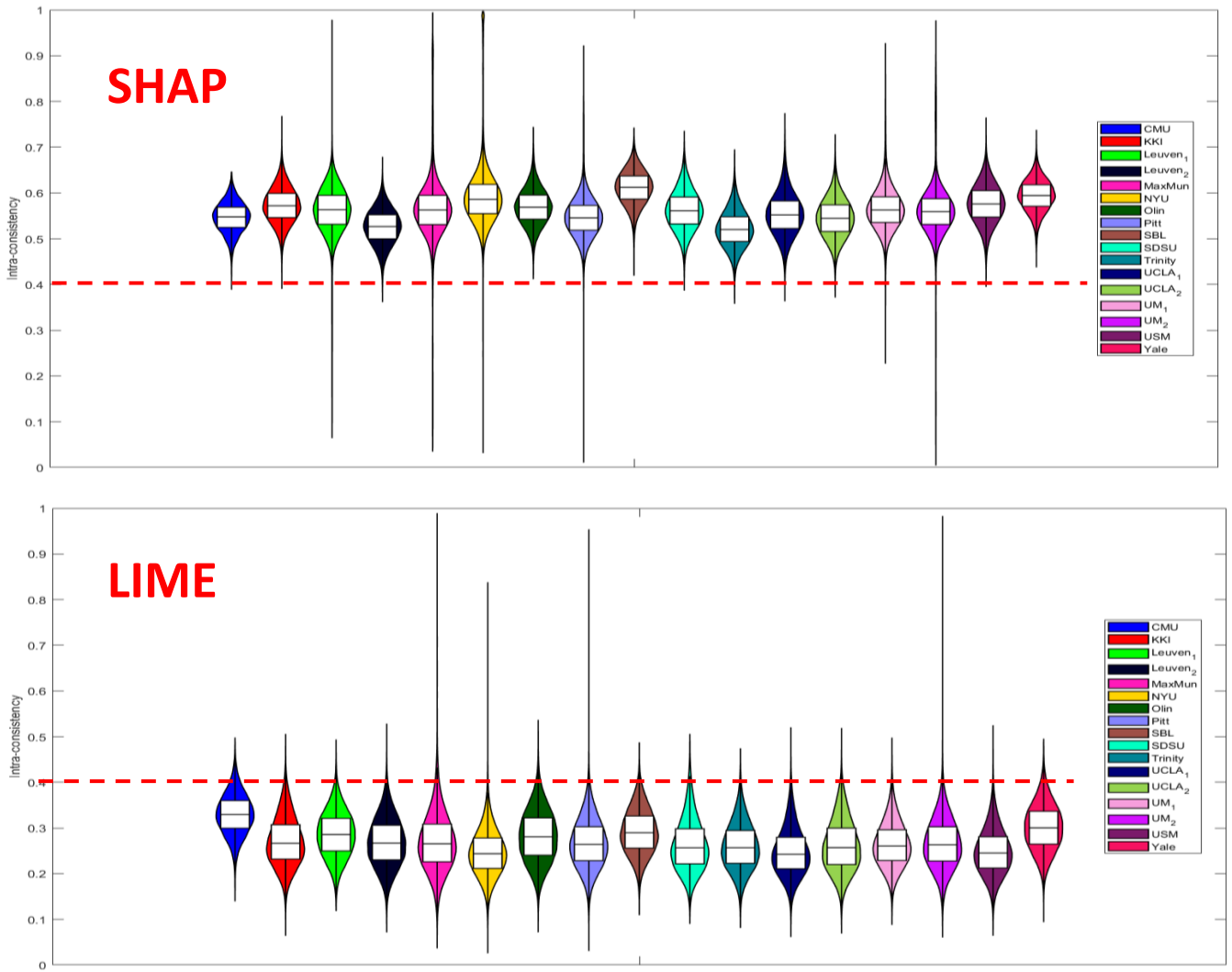


For the whole dataset, the proposed **DL models achieved MAE and R values that compare favourably with the literature showing the overall performance MAE = 2.7 and R = 0.86**. Both performance metrics were found to be significantly different from the chance level, resulting p = 0 from the non-parametric permutation test.

Results: explain performance



Results: stability of XAI methods

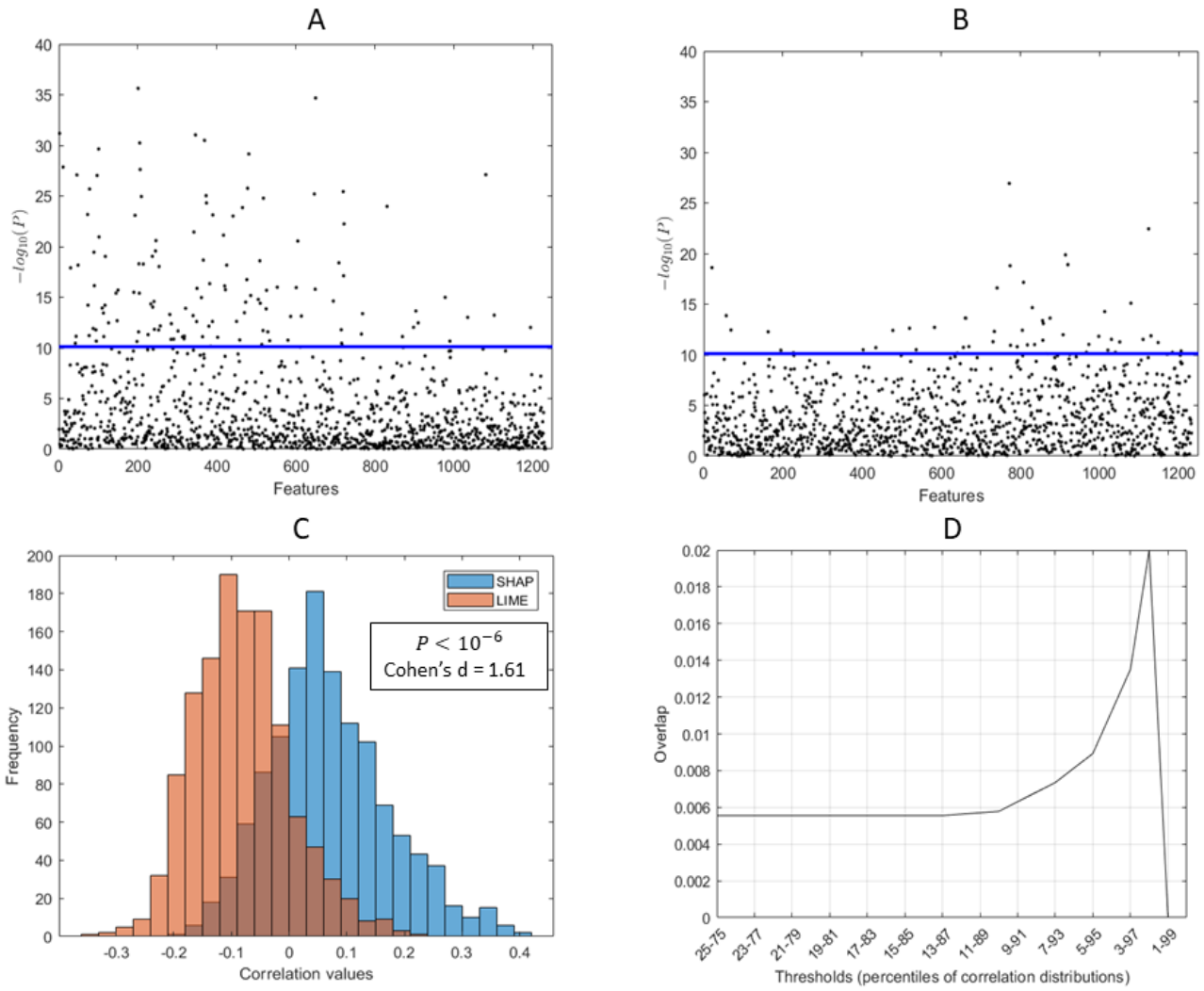


Intra-consistency = 0.4

Apart from a slight difference between the different sites for both scores, the **LIME scores** show consistently lower intra-consistency values (**lower than 0.4 for all the sites**) than those exhibited by the **SHAP scores (greater than 0.5 for all the sites**).

The **SHAP algorithm has been selected has the most reliable!**

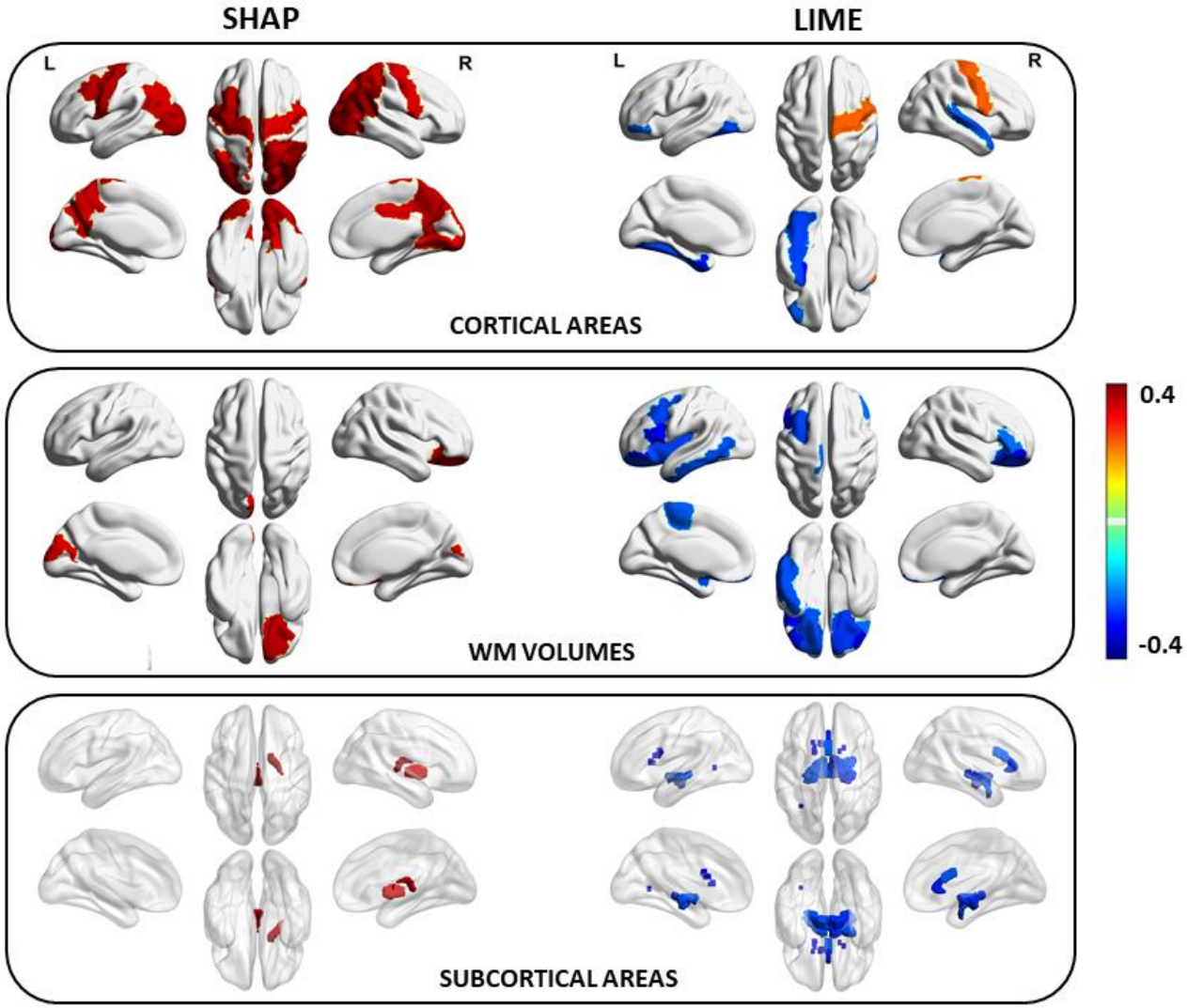
Results: global XAI



A **correlation analysis** between each **feature score vector and the age of the subjects** was performed to yield a set of morphometric descriptors whose relevance for age prediction is most variable with age.

This step of the framework provides **global explanations of the DNN models** since a set of age-related scores is extracted from the whole population under investigation.

Results: biological interpretation



The brain regions corresponding to the most age-related features for both XAI methods are shown in figure.

Notably, only the SHAP method showed a significant correlation between the importance of the cortical thickness of both hemispheres and age (R = 0.38 for left and R = 0.36 for right).

Conclusions

- It is significant to use **XAI models in healthcare domains** to help healthcare professionals make wise and interpretable decisions.
- However, before rushing into very complex ML models, it is always better to **train different simple methods** and evaluate the performance.
- The **correct measurement of XAI properties** is one of the biggest challenges of XAI.
- **ML interpretability is domain-specific**: different users require different types of explanations!
- **Model-agnostic methods** have gained researchers' attention due to their flexibility: try different ML models, select the most accurate and explain it.

Thank you
for your attention

Questions?

angela.lombardi@ba.infn.it