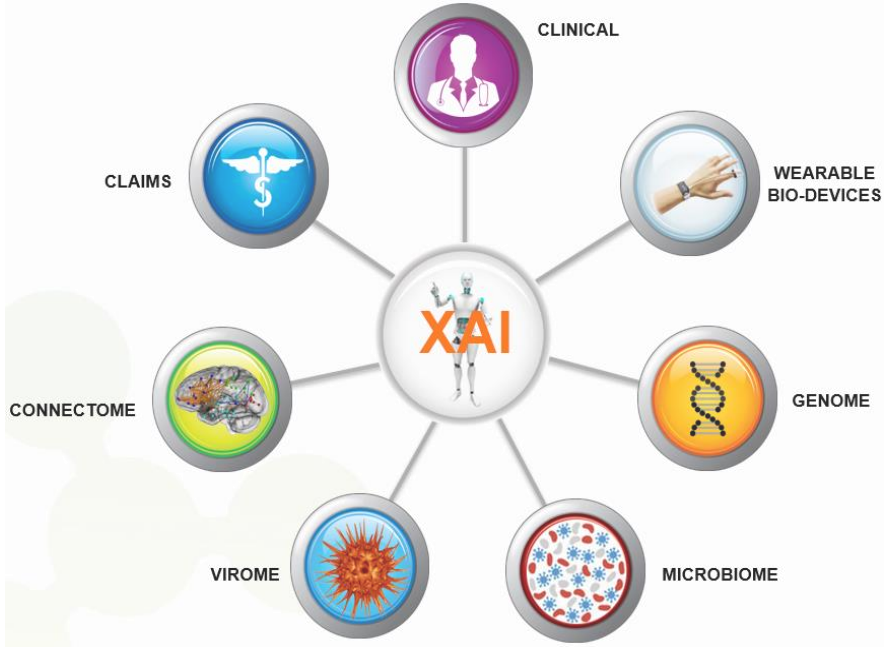
AI@INFN workshop, Bologna 2-3 May 2022

Trustworthy AI in medical applications



Angela Lombardi, PhD

Dipartimento Interateneo di Fisica – Università degli Studi di Bari e

Istituto Nazionale di Fisica Nucleare – Sezione di Bari

Towards a **Trustworthy AI** in medicine

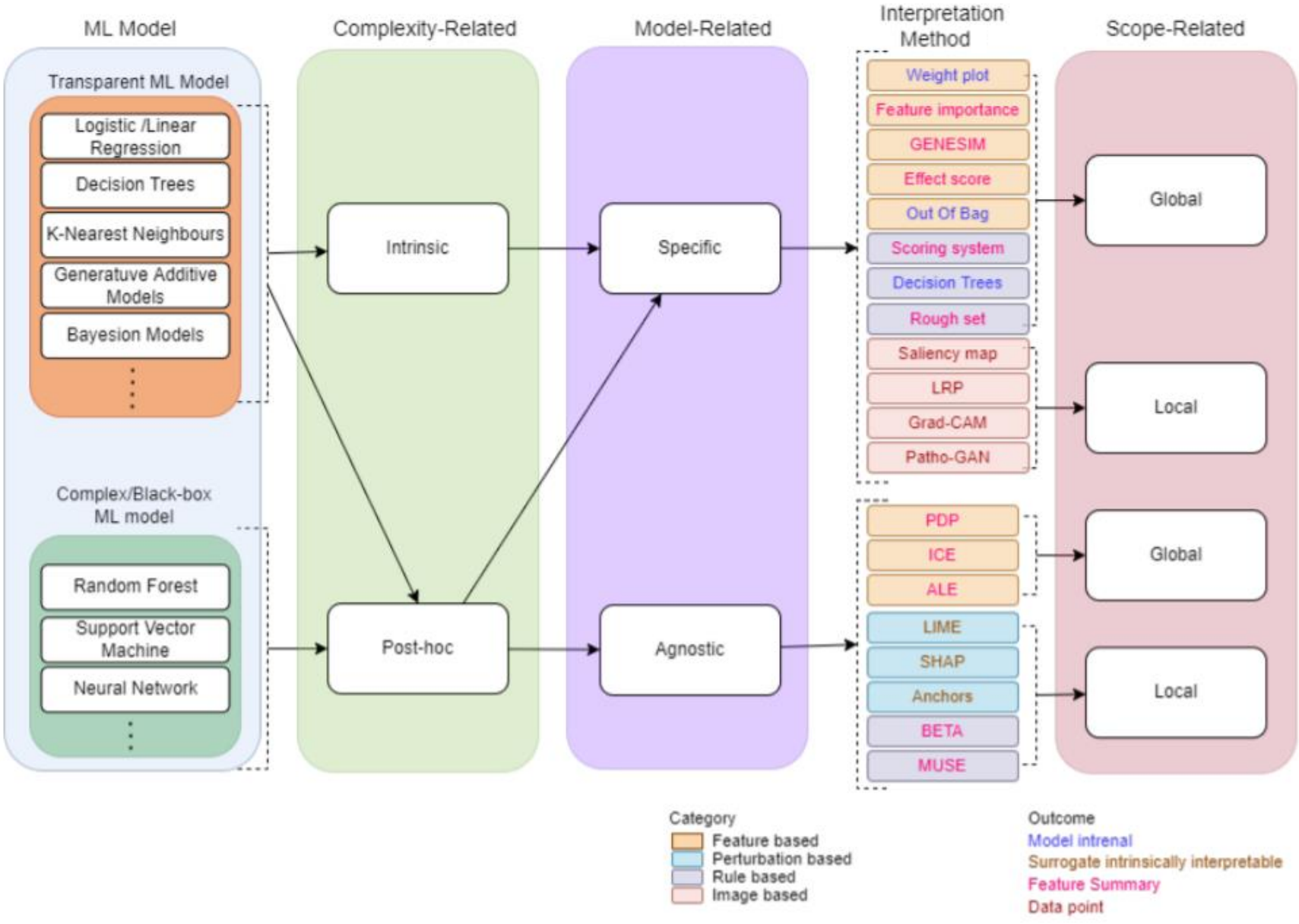
G20 AI five complementary value-based principles:

- inclusive growth, sustainable development and well-being;
- **human-centred** values and fairness;
- **transparency** and **explainability**;
- **robustness**, security and safety;
- **accountability**.

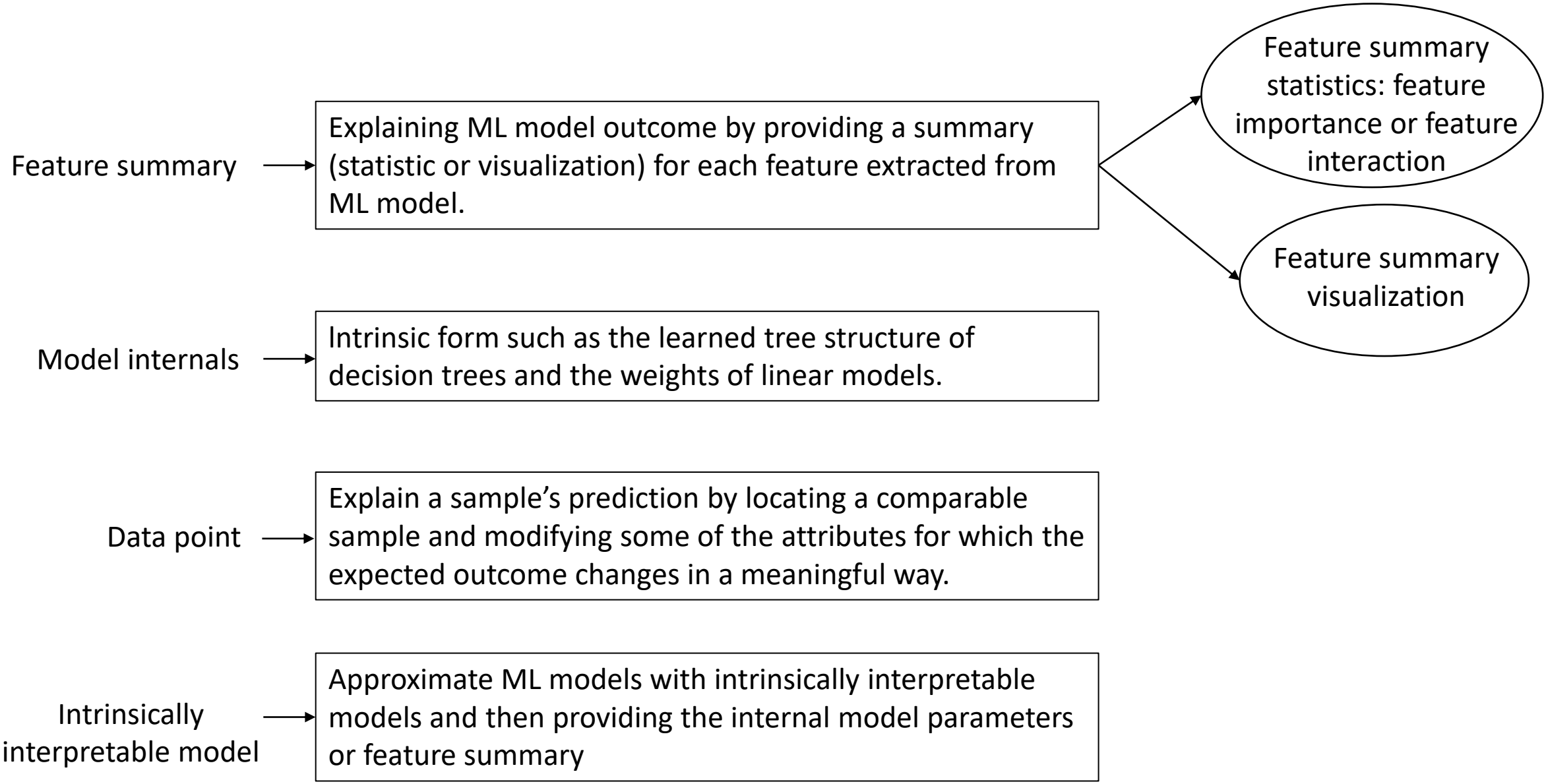
National Academy of Medicine's recommendations for AI in health

- Seek out **robust evaluations of model performance**, utility, vulnerabilities, and bias.
- There should be a deliberate effort to identify, mitigate, and **correct biases in AI tools**.
- Demand **transparency in data collection** and **algorithm evaluation** processes.
- Develop **AI systems with adversaries** (bad actors) in mind.
- Use **AI systems to engage**, rather than stifle, uniquely **human abilities**.
- Use **automated systems to reach patients** where existing health systems do not.

eXplainable models for trustworthy AI: XAI



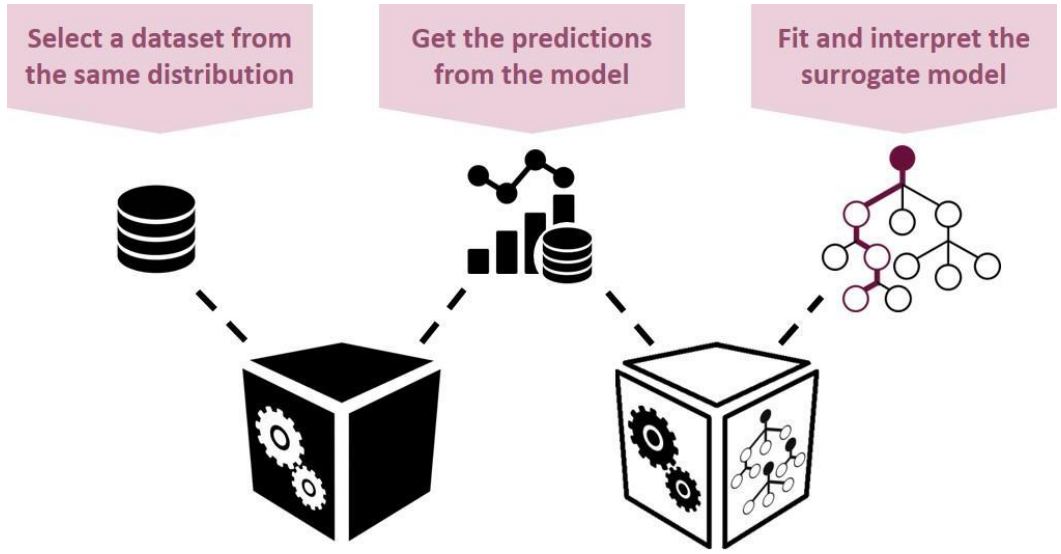
Outcomes of XAI



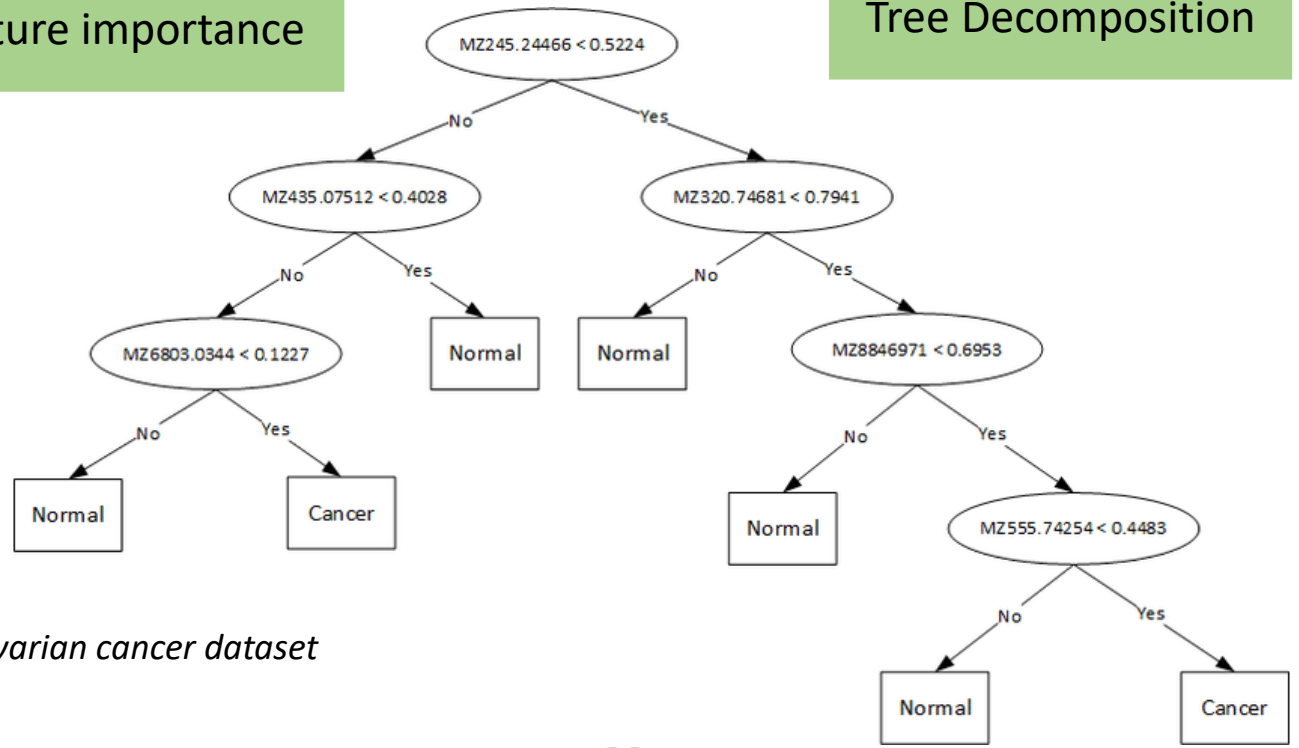
Simple human-friendly explanations

Feature importance

Tree Decomposition



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



Ovarian cancer dataset

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

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

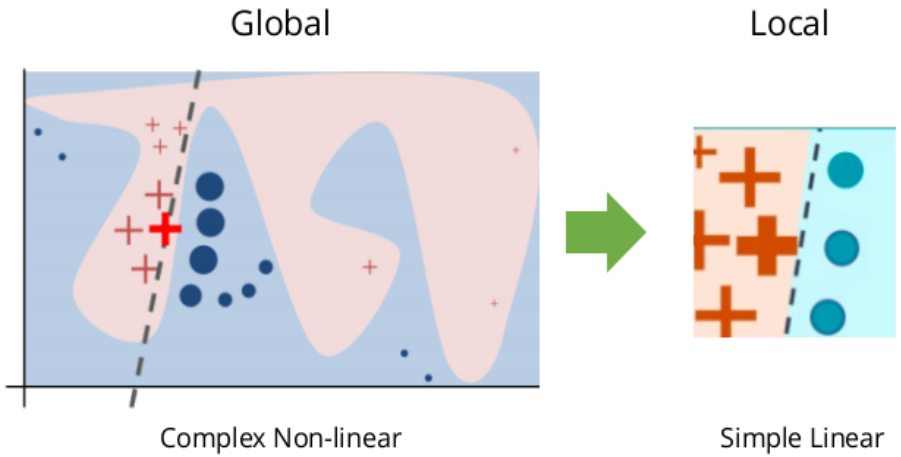
If an instance falls into a leaf node R_l, the predicted outcome is $\hat{y} = c_l$ where c_l is the average of all training instances in leaf node R_l.

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

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:

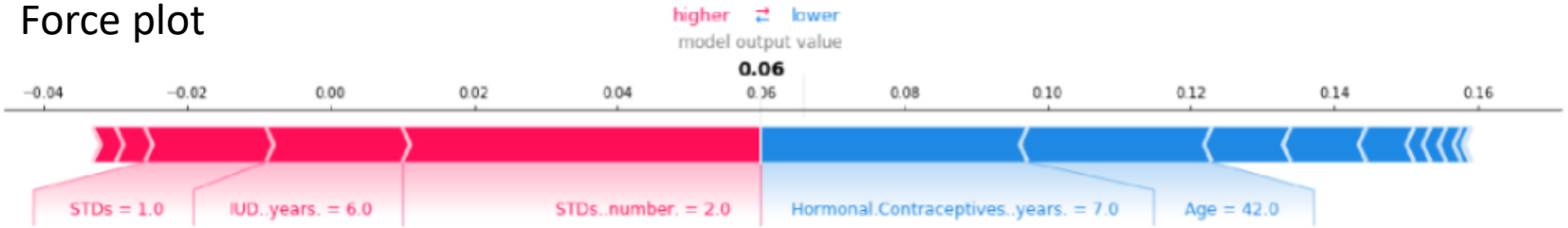
$$\phi(i) = \sum_{S \subseteq \mathcal{F} \setminus \{i\}} \frac{|S|! \cdot (F - |S| - 1)!}{F!} \left(f_{S \cup \{i\}}\left(x_{S \cup \{i\}}\right) - f_S\left(x_S\right) \right)$$

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

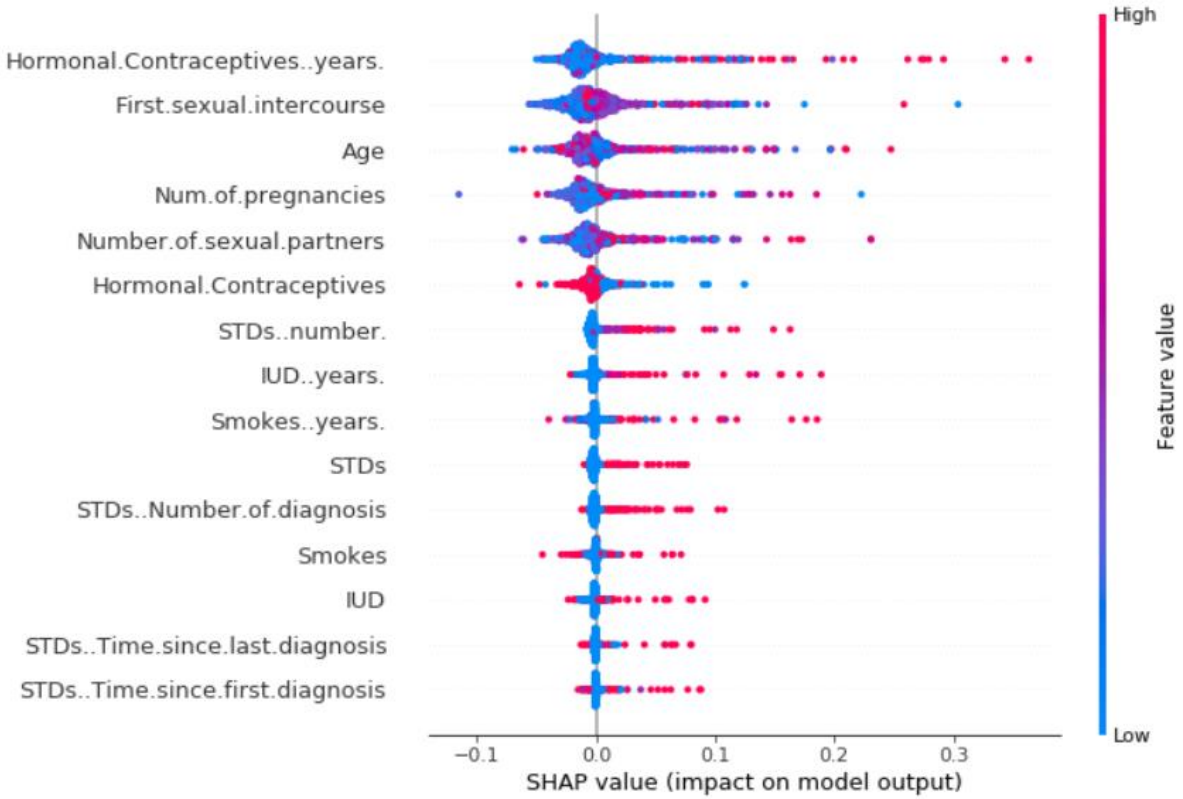
$$g(x) = w_0 + \sum_{i=1}^{F} w_i \cdot x_i \qquad w_0 = \phi_0, \qquad w_i = \frac{\phi_i}{x_i - \mu_i}, \qquad 1 \leq i \leq F, \mu_i \in \mathbb{B}_{\{i\}}$$

SHAP: visualization

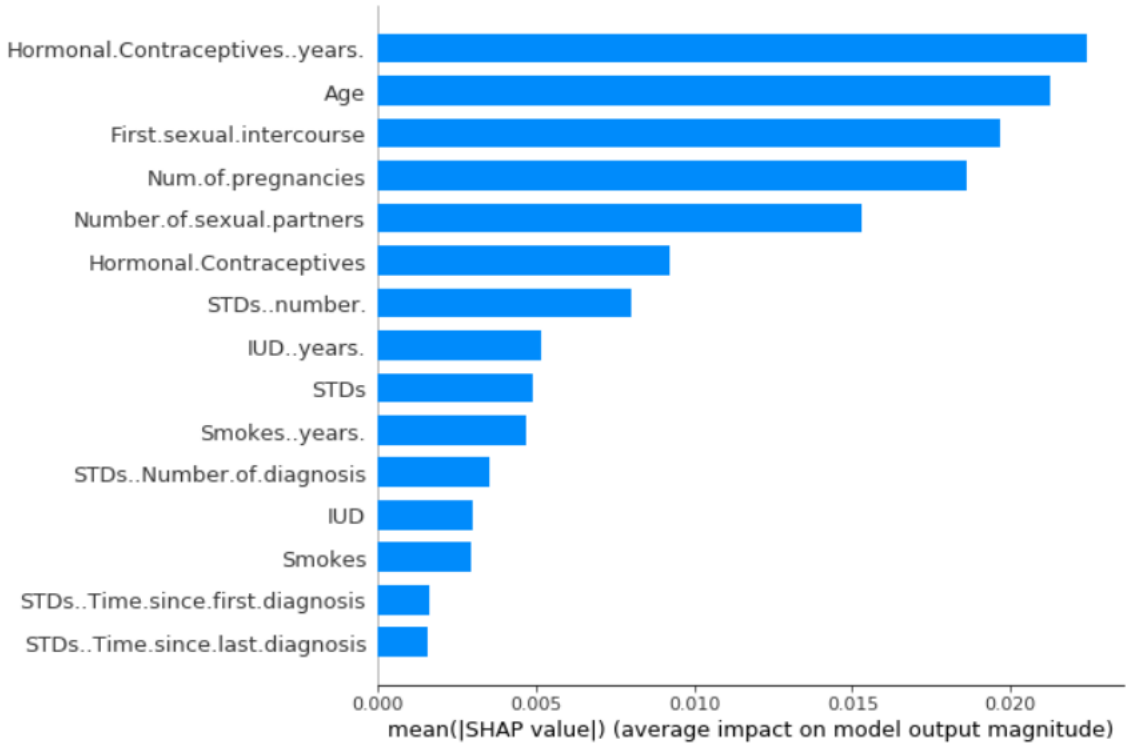
Force plot



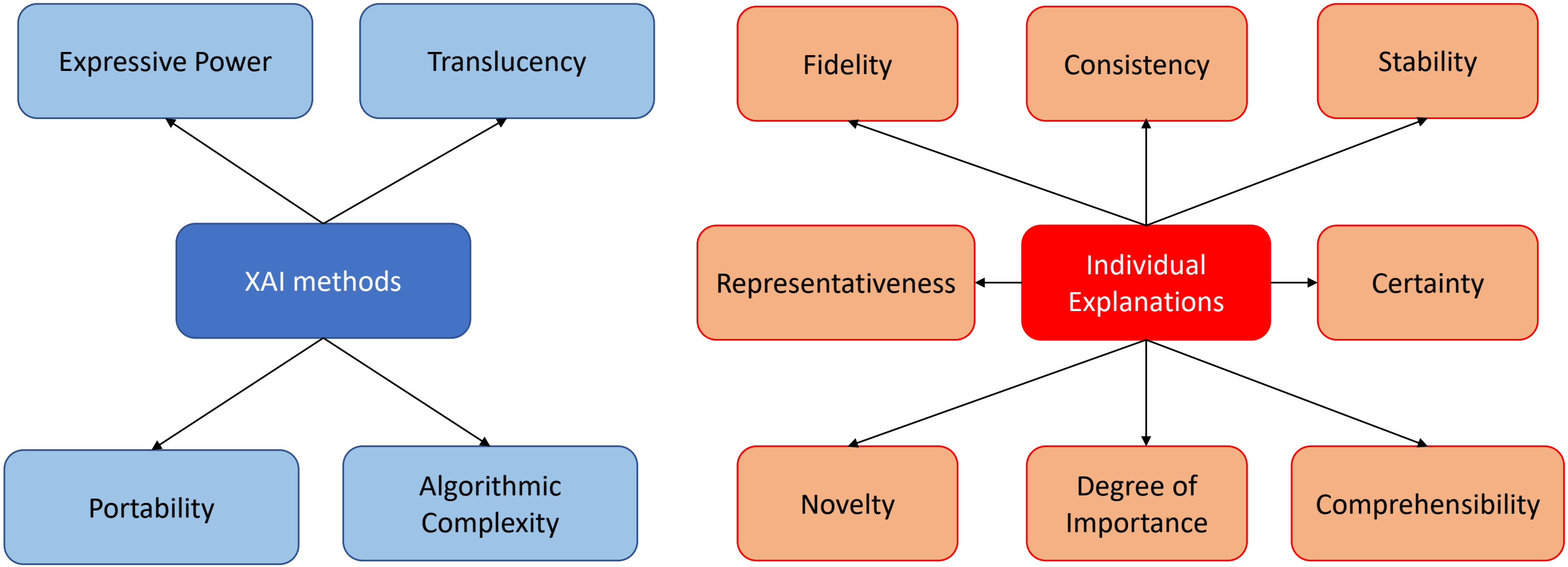
Summary plot



Global feature importance



TO DO: formalization of properties of XAI for medical applications

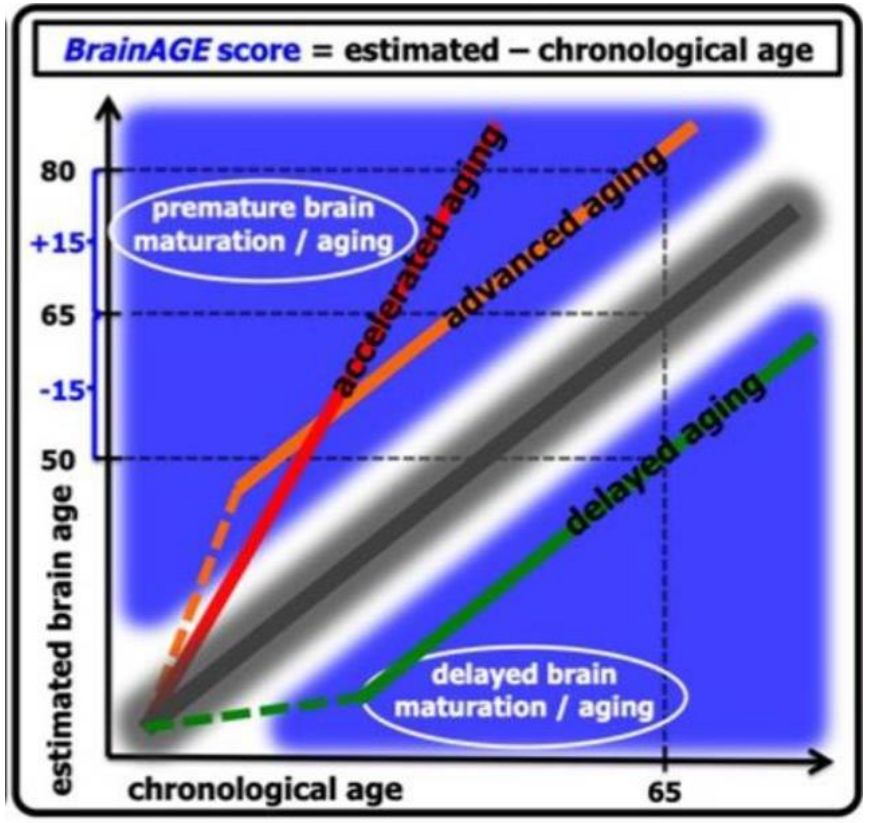


Case study: predicting brain age with ML/DL

- 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.

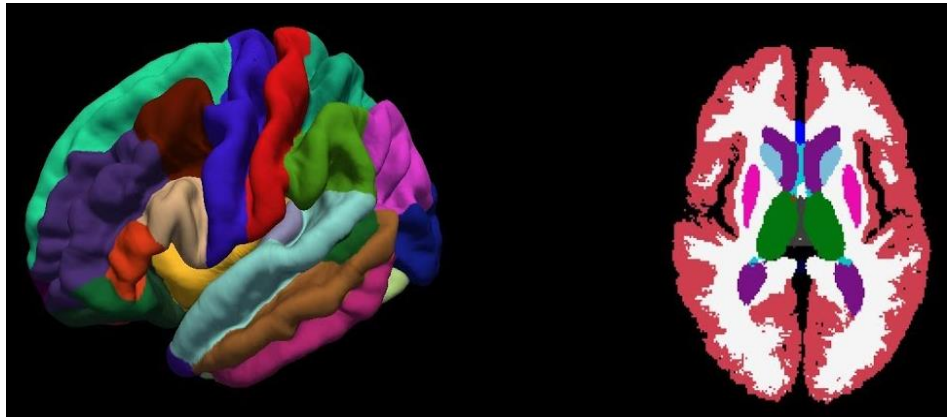


Franke, Katja, and Christian Gaser. "Ten years of brain age as a neuroimaging biomarker of brain aging: what insights have we gained?." *Frontiers in neurology* 10 (2019): 789.

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;

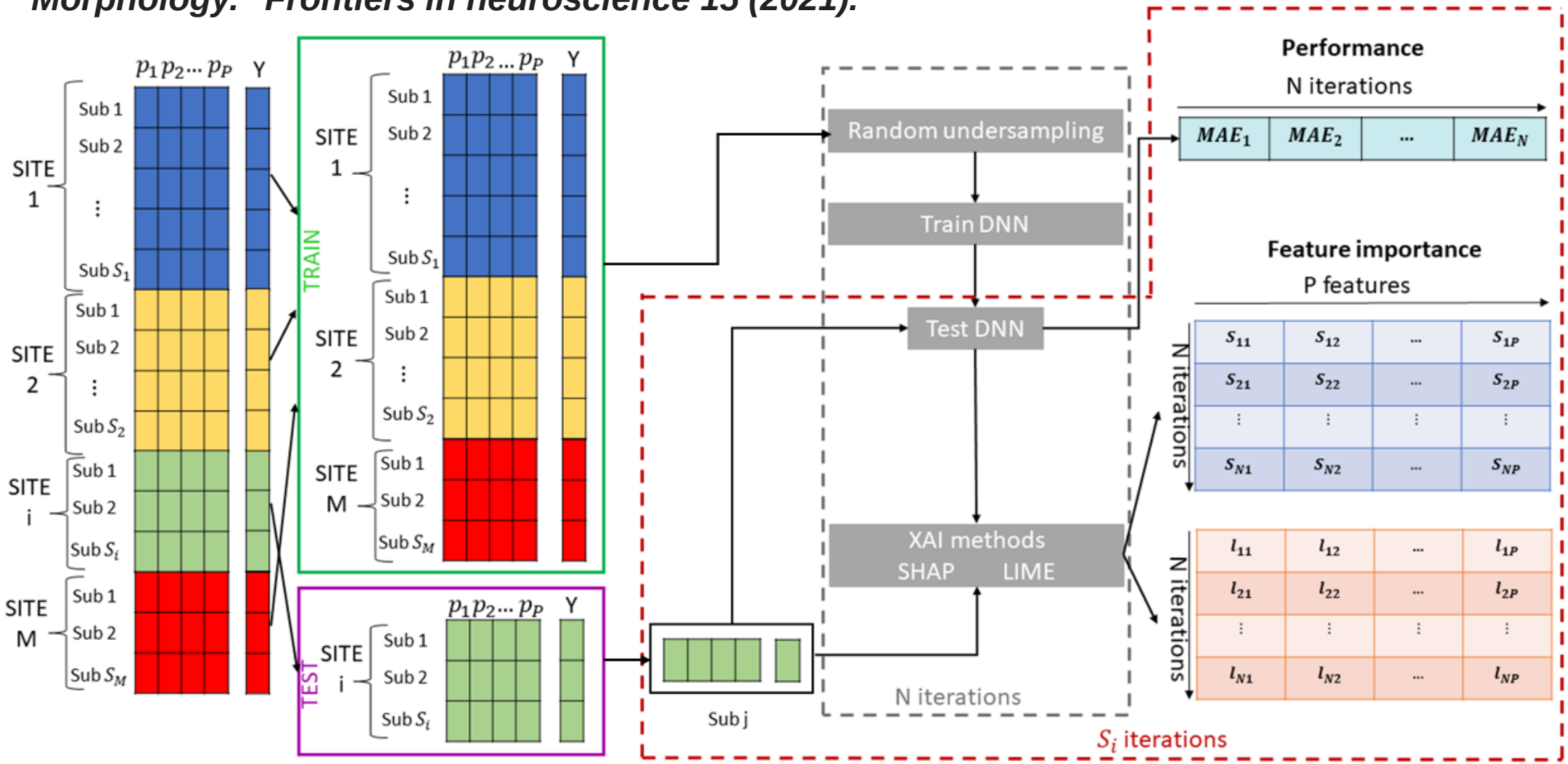
•**volume, surface area, Gaussian curvature, mean curvature, curvature index, folding index, thickness mean, and thickness standard deviation** for the **34 cortical brain regions of each hemisphere**;

•**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.

Example: Brain age prediction

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

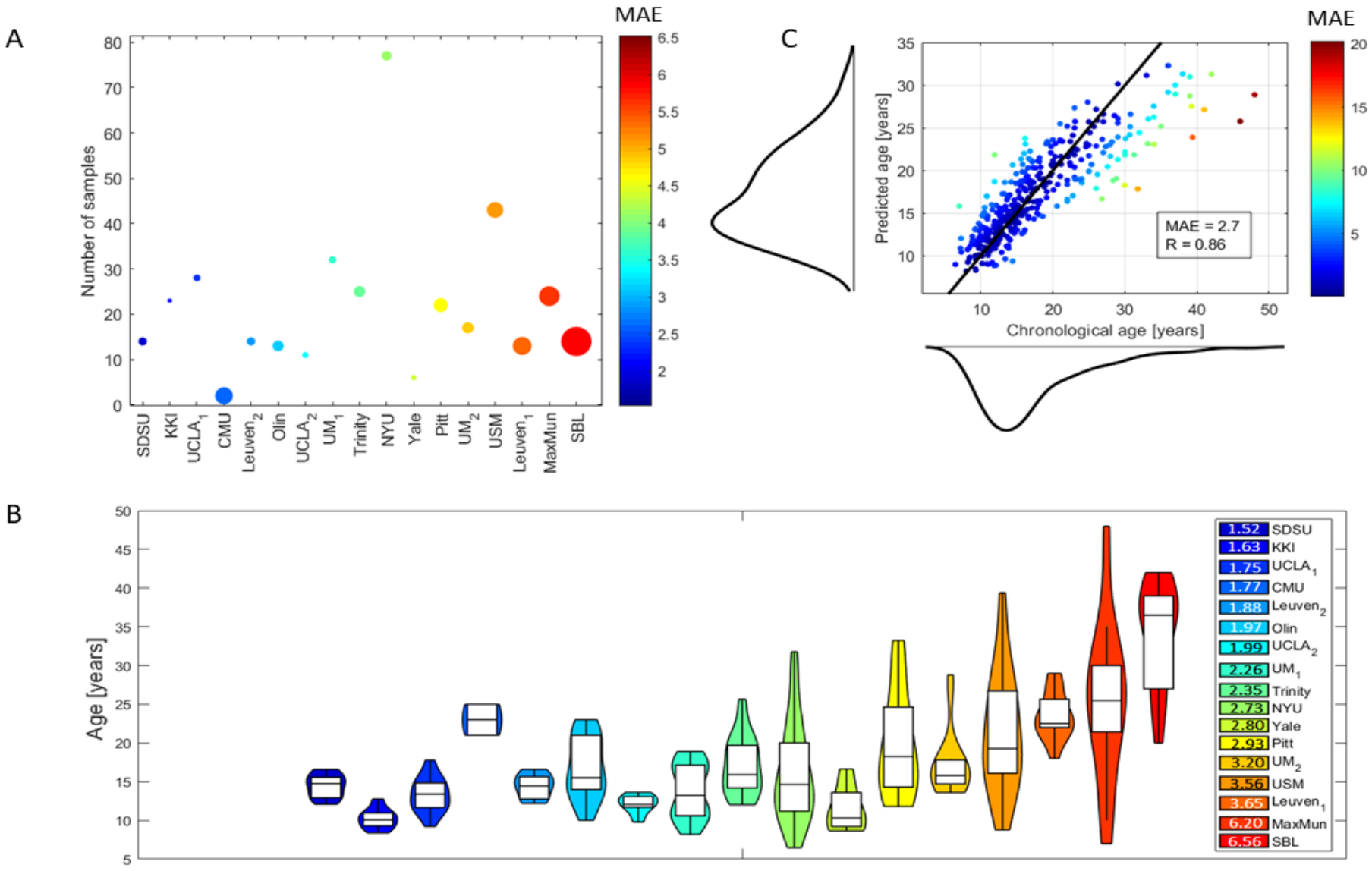


Quantify the variability of XAI scores

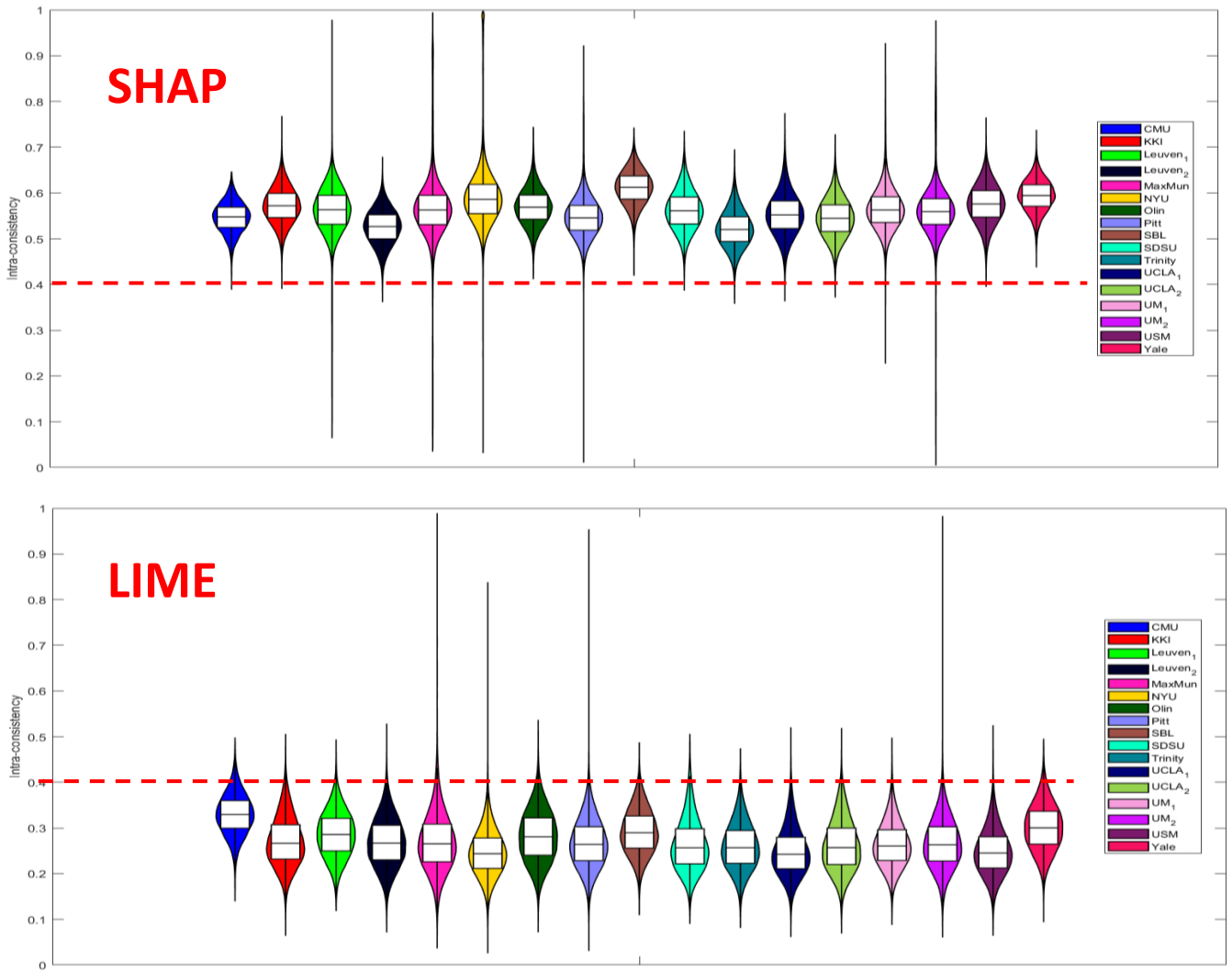
Reliability of XAI models to explain local «subject-level» decisions:

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?

Results: explain performance



Results: stability of XAI methods

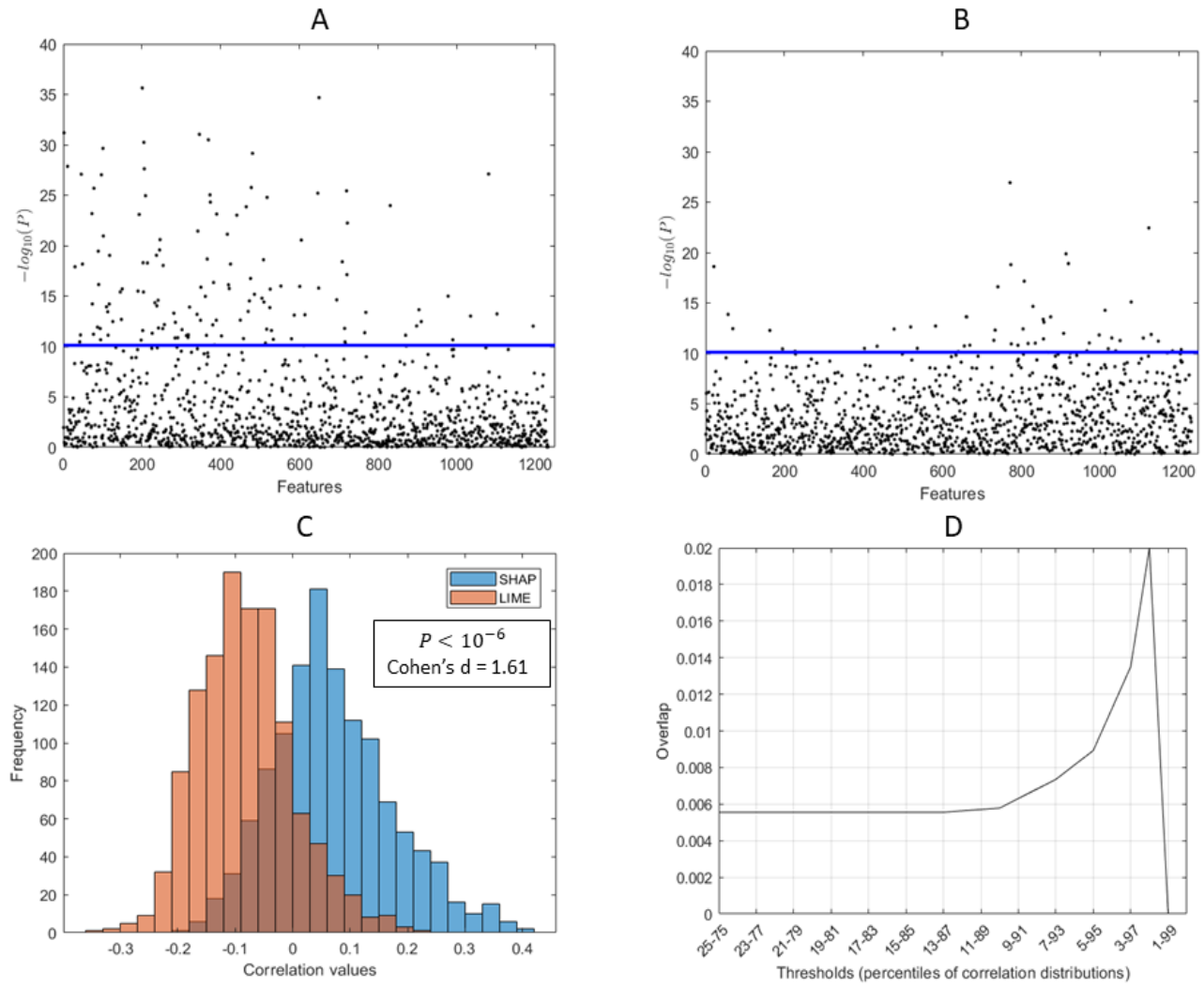


Intra-consistency = 0.4

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!**

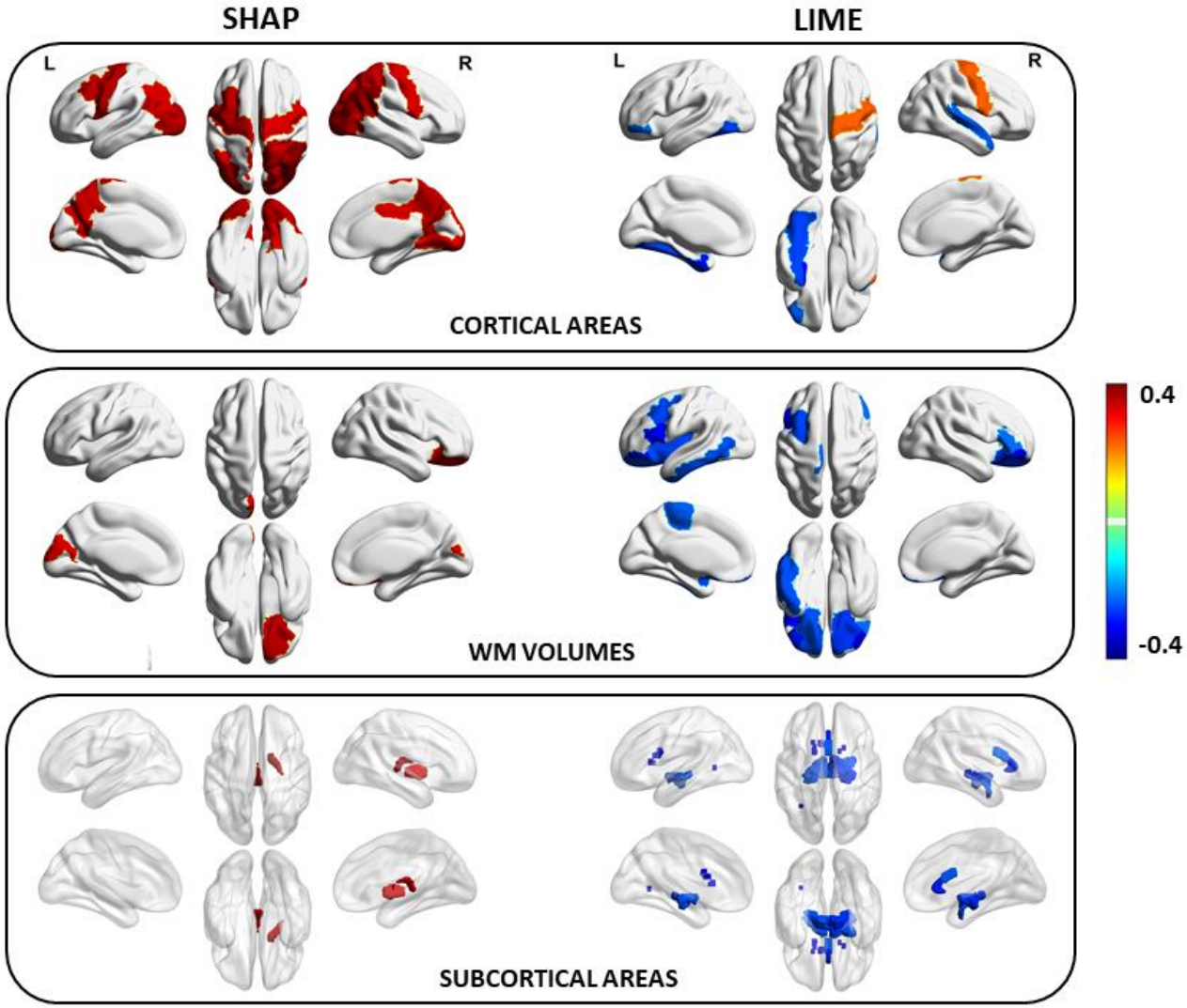
Results: global XAI



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).

Remarks

- It is significant to use **XAI models in healthcare domains** to help healthcare professionals make wise and interpretable decisions.
- 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