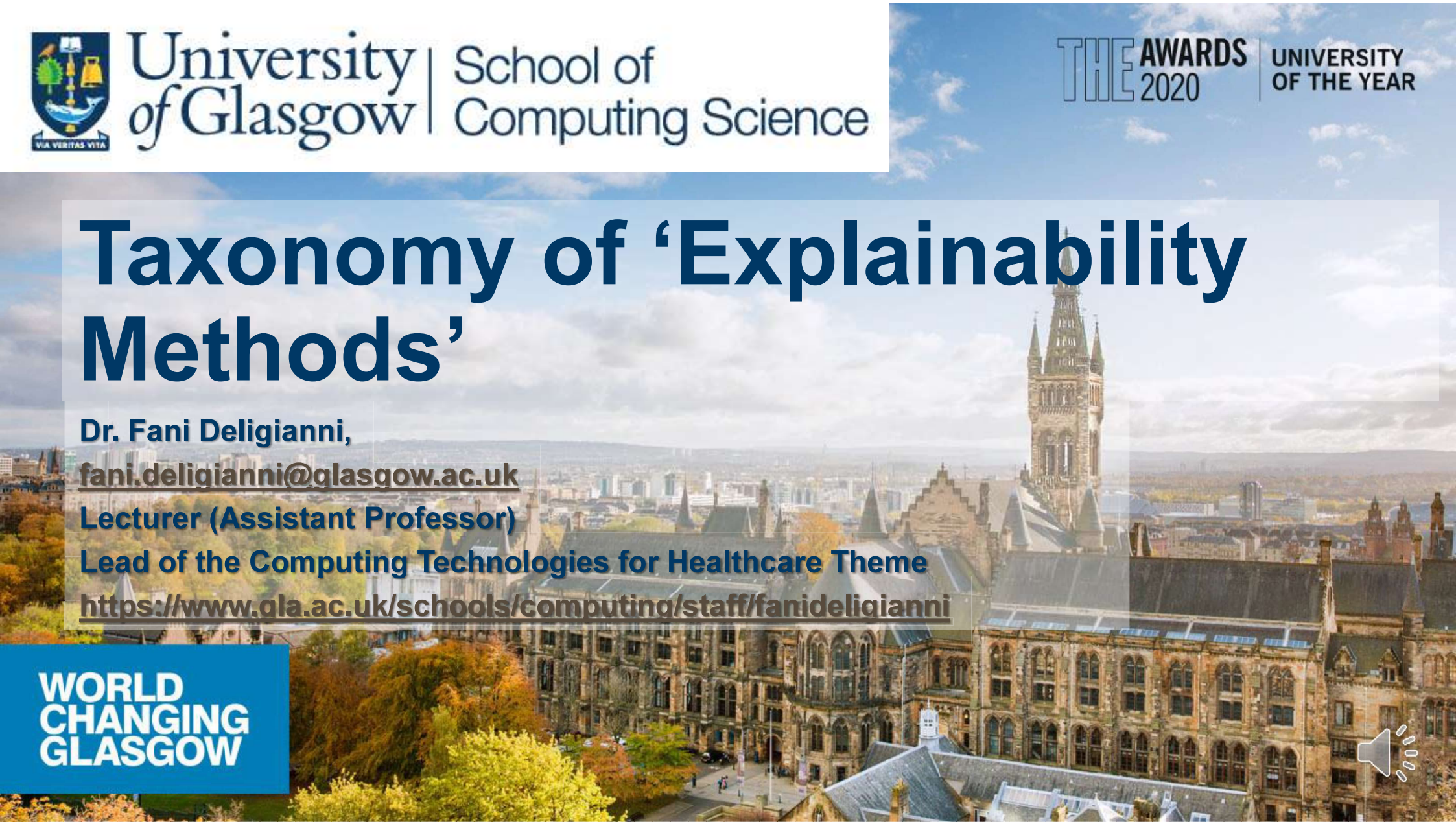


Taxonomy of 'Explainability Methods'

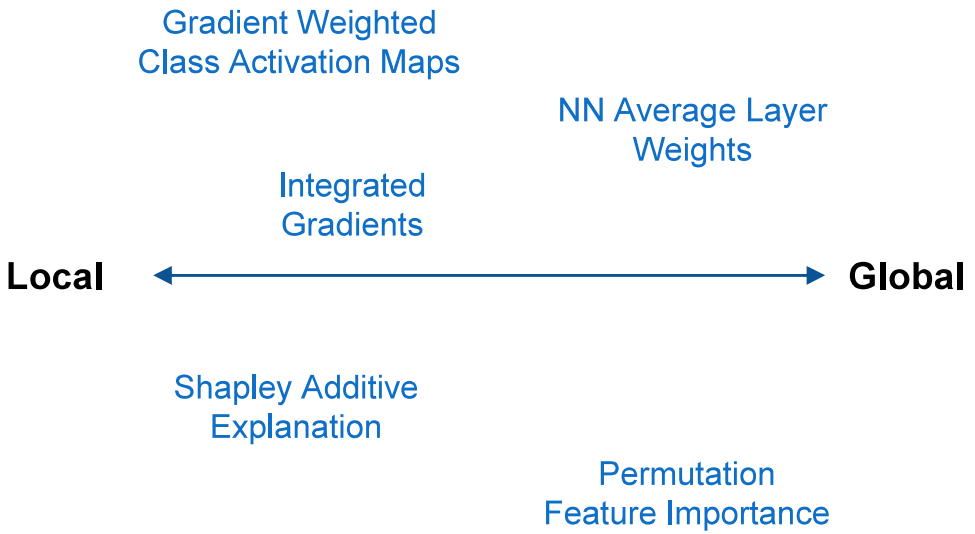
Dr. Fani Deligianni,
fani.deligianni@glasgow.ac.uk
Lecturer (Assistant Professor)
Lead of the Computing Technologies for Healthcare Theme
https://www.gla.ac.uk/schools/computing/staff/fanideligianni

Taxonomy

- Local vs Global Explanations
- Model Agnostic vs Model Specific Explanations
- Data Modality Specific vs Data Modality Agnostic
- Ad-Hoc vs Post-Hoc Explanations

Local vs Global Explanations

- Local methods provide explanations for individual samples
- Global methods provide explanations for the entire model or group of samples
- Local methods results per sample can be averaged across samples
- Global methods will weight input parameters the same way regardless the individual prediction



Global Interpretability

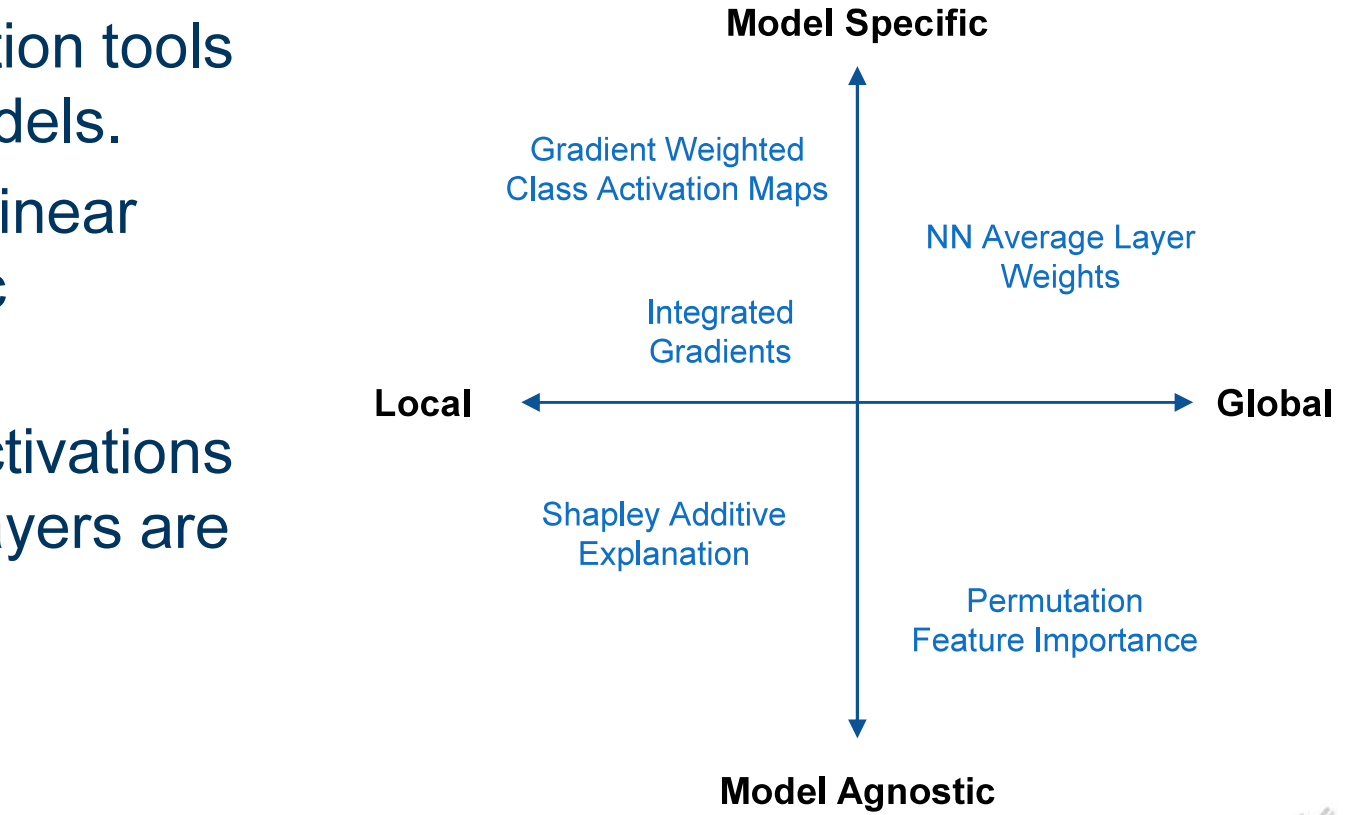
- Overall view of the model, along with data predictions and explanations.
- The **data exploration**, which displays an overview of the data set along with the prediction values.
- The **global importance**, these aggregates, features, importance values of individual data points, to show the model's overall top key.
- The explanation demonstrates how a feature affects the change in the model prediction values

Local Interpretability

- The **local importance** highlights the important features for any individual prediction.
- It illustrates the local behavior, of the underlying model, at a specific data point.
- The **data collection exploration**, is a sort of what if analysis.
- These observations allows changes to feature values of the selected data points, and observers outing changes to the prediction value.
- Local explanations can illustrate how a prediction would change when a feature changes.

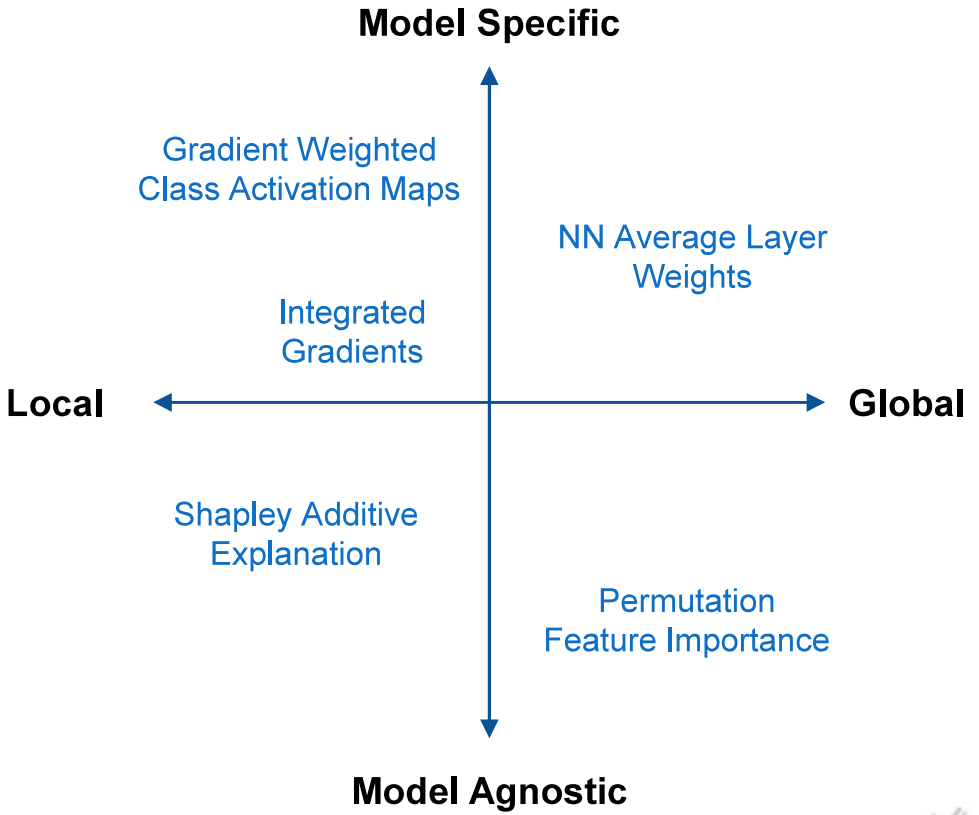
Model Specific Explanations

- Model-specific interpretation tools are limited to specific models.
- Regression weights in a linear model is a model-specific explanation
- Methods based on the activations of deep neural network layers are model-specific



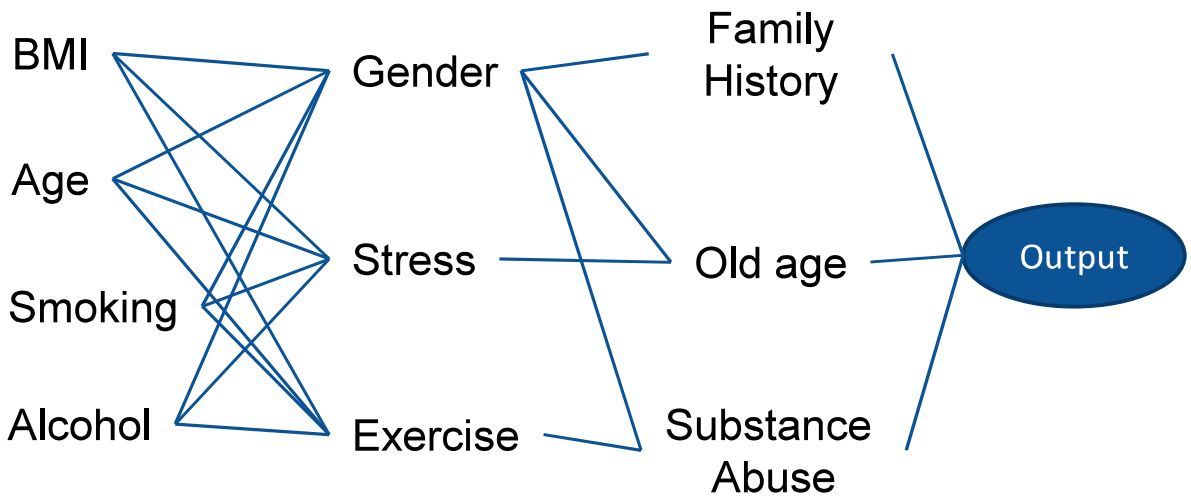
Model Agnostic Explanations

- Model-agnostic tools can be used on any machine learning model
- Agnostic methods usually work by analyzing feature input and output pairs.
- These methods cannot have access to model architecture such as layer weights or structural information.
- Apply surrogate models
- Permutation based approches



Ad-Hoc Explanations

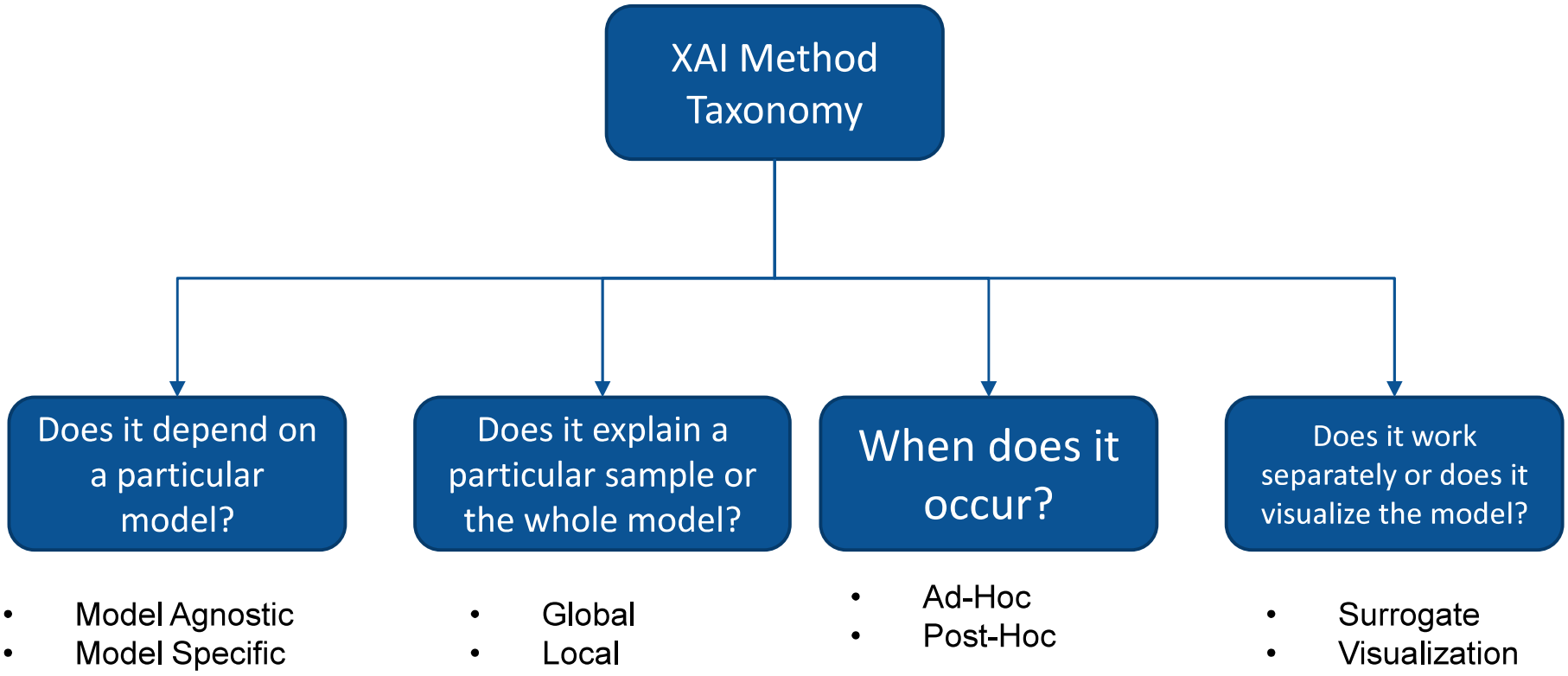
- In **ad-hoc explanations** the model has been designed to be intrinsically explainable
- Representation learning can result in ad-hoc explanations
- Identify latent factor and disentangle their influence on the outcome



Other type of taxonomies

- Data Modality Specific vs Data Modality Agnostic
  - Tabular data vs Time-Series Data
  - Time-Series data vs imaging data
  - Data Modality agnostic are related to Model Agnostic explanations.
- Surrogate Models vs Attribution/Visualisation Methods
  - Employ an interpretable model that approximates the black box
  - Attempt to visualize certain aspects of the model to allow explanation on why and how the model reach a decision

Overview


XAI Method Taxonomy
Does it depend on a particular model?
Does it explain a particular sample or the whole model?
When does it occur?
Does it work separately or does it visualize the model?
• Model Agnostic
• Model Specific
• Global
• Local
• Ad-Hoc
• Post-Hoc
• Surrogate
• Visualization

Summary

- ‘Explainability’ is defined in several levels and this result in a large diversification of ‘explainability’ approaches
- These approaches can be fundamentally different and result in completely different kind of explanations
- Sometimes more than one explainability method might be required to provide a better insight of a model in a diverse set of users

References

- Arrieta et al. ‘Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI’, Information Fusion, 2020.
- Molnar ‘Interpretable Machine Learning - A Guide for Making Black Box Models Explainable’ https://christophm.github.io/interpretable-ml-book/