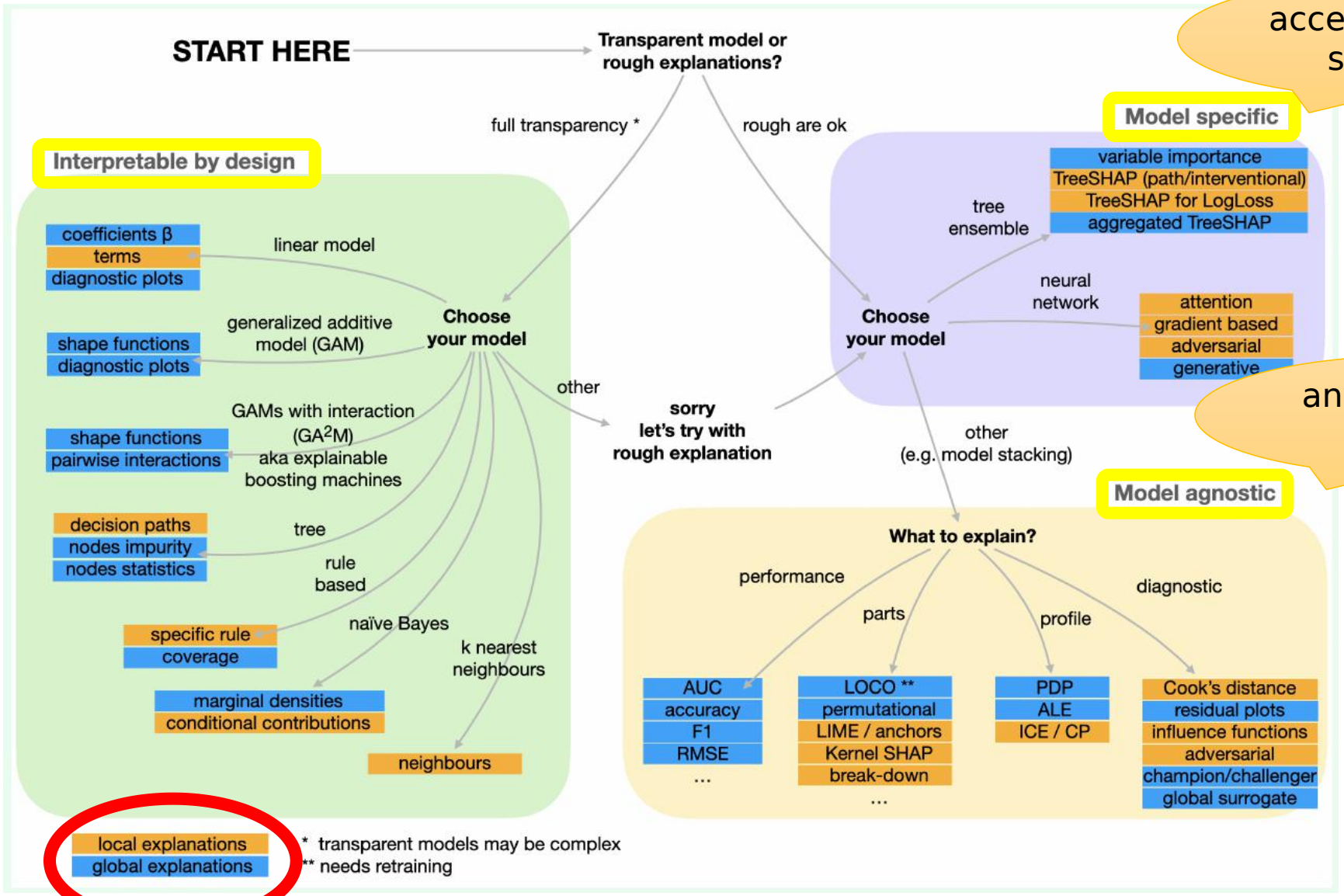
explainable AI

Interpretable ML, Responsible ML



2nd way of grouping

access of model structure

any predictive model

To understand:

- importance of variables
- profile of variables
- model performance

Types of model agnostics

without touching the structure of DNN, a generic analysis

- model profile
  - variable vs. model response (partial dependence plot, individual conditional expectation (ICE), accumulated local effect (ALE))
- variable importance
  - permutational-based
    - masking one variable & check model performance (Leave One Covariate Out (LOCO), surrogate tree)
  - Shapley-based
    - local explanation, based on coalitional game theory
  - variance-based
    - model profile based ('flatness' of the PDP)

VI for (Deep) Neural Network

- NN
  - Garson algorithm
    - all weighted connection between nodes of interest
  - Olden algorithm
    - sum of product of raw connection weights
- DNN
  - Gedeon
    - weights of first two hidden layers
  - Layer-wise Relevance Propagation
    - Deep Taylor's expantion