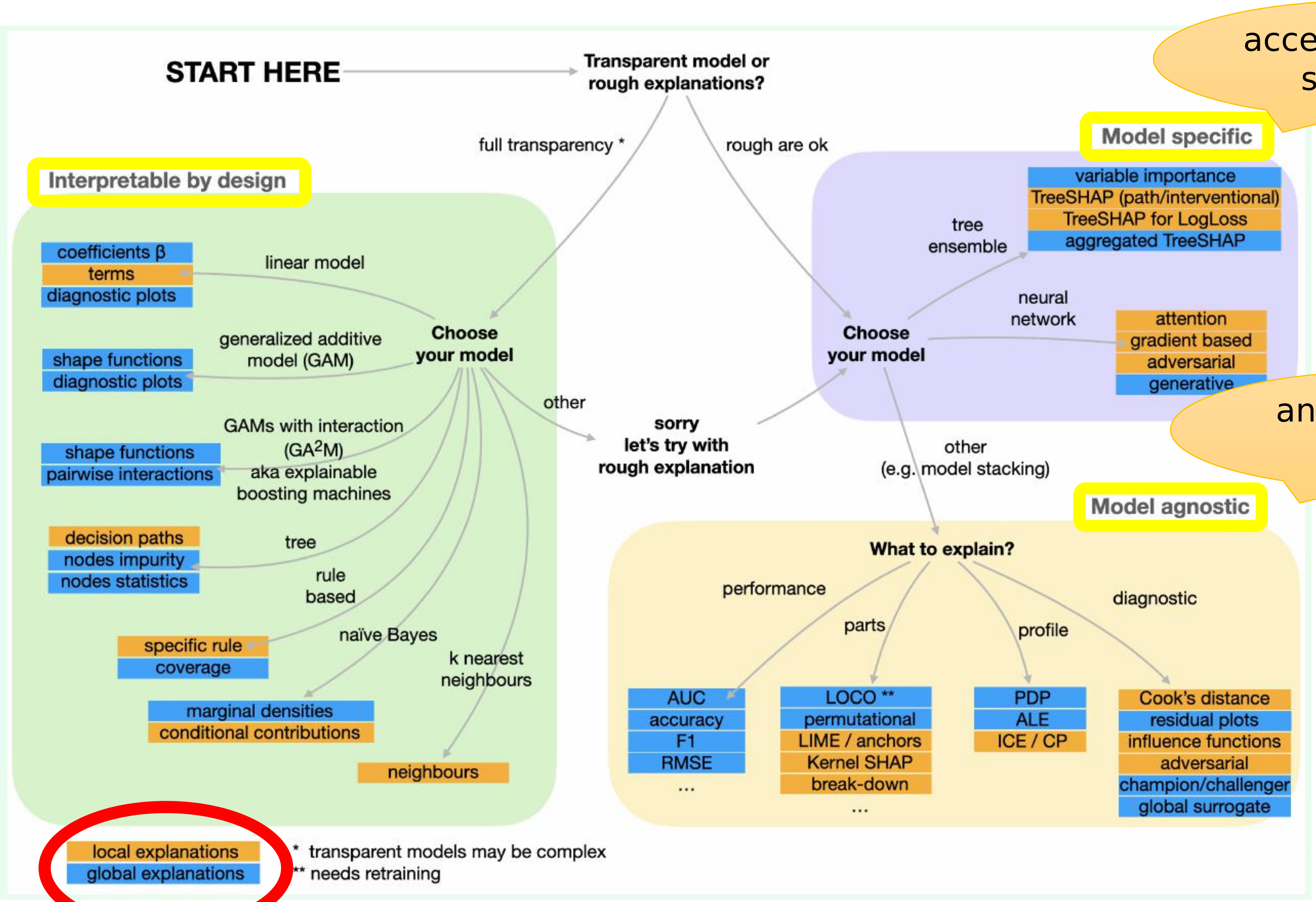
explainable AI

Interpretable ML, Responsible ML



access of model structure

any predictive model

2nd way of grouping

local explanation: covariates at a sample point
global explanation: covariates across all sample points

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

    shuffle values of one variable and check change in model performance
    - 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_perf ~ env_covariate_range. if important, there should be a peak
    - model profile based ('flatness' of the PDP)

these models don't take interactions into account, but there are workarounds by doing permutations in X2 for each value of X1. if interaction, model performance profile won't change across values of X1

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 expansion