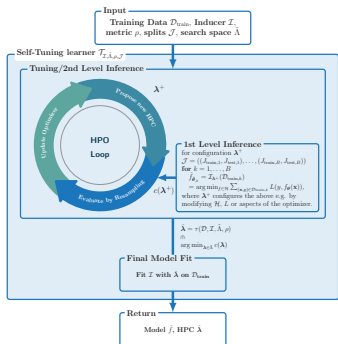


# Introduction to Machine Learning

## Hyperparameter Tuning

### Practical Aspects

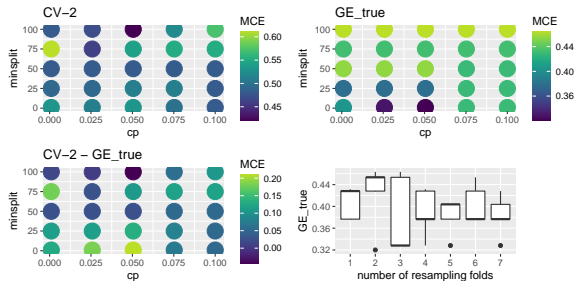


### Learning goals

- Understand the possible design choices for HPO
- Know termination criteria of HPO

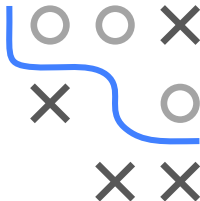
# PRACTICAL ASPECTS OF HPO

- Choosing resampling
  - Nr of observations, i.i.d assumption for data sampling process
  - Higher resampling rates likely result in a better model; however they are computationally more expensive



## PRACTICAL ASPECTS OF $\text{HPO}_2$

Tuning a CART on the `spirals` data with a k-fold CV (k=1 means here a 2/3 holdout split) using grid search and estimating the true GE with a very large test set (5 repetitions)

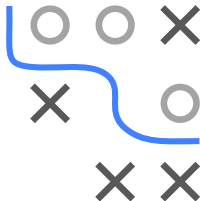


A 3x3 grid with a blue path starting at the top-left cell (0,0) and ending at the bottom-right cell (2,2). The path consists of the following cells: (0,0), (0,1), (1,1), (1,2), and (2,2). The cells (0,1), (0,2), (1,0), and (2,0) are marked with a grey 'X', while the cells (1,0) and (2,1) are marked with a grey 'O'.

- 
- The figure consists of two plots. The left plot is a heatmap showing the relationship between 'misspelt' (y-axis, 0 to 100) and 'cd' (x-axis, 0.00 to 0.05). The background is a heatmap where color represents 'accuracy' (ranging from 0.76 to 0.80). Overlaid on this are points colored by 'type' (red for  $\hat{\lambda}^+$ , grey for  $\hat{\lambda}^-$ ) and shaped by 'type' (triangles for  $\hat{\lambda}^+$ , circles for  $\hat{\lambda}^-$ ). The right plot is a line graph showing 'accuracy' (y-axis, 0.78 to 0.81) versus 'area' (x-axis, 1 to 5). It contains two lines: a solid line for  $\hat{\lambda}^+$  and a dashed line for  $\hat{\lambda}^-$ , both showing an increasing trend in accuracy as area increases.

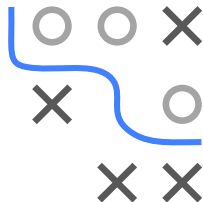
## PRACTICAL ASPECTS OF $\text{HPO}_{4}^{2-}$

Tuning `cp` and `minsplit` for a CART on the `titanic` data over 3 increasing rectangular search spaces with random search (candidates number fixed) and comparing the result with the optimal model (found with exhaustive grid search)



# PRACTICAL ASPECTS OF HPO

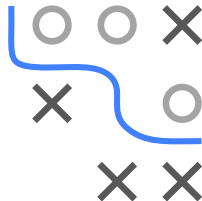
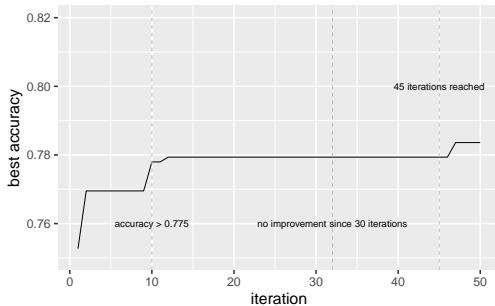
- Choosing HPO algorithm
  - For few HPS (1-3), grid search can be used
  - BO with GPs for upto 10 numeric HPs
  - BO with RFs handle mixed HP spaces
  - Random search and Hyperband work well as long as the “effective” dimension is low
  - EAs are somewhat in-between BO and RS, can handle very complex spaces, but less sample efficient than BO
  - **Also: use something that’s stable and robust! More an aspect of the implementation than the algo!**



# PRACTICAL ASPECTS OF HPO

When to terminate HPO

- Specify a certain amount of runtime/budget beforehand
- Set a lower bound regarding  $\widehat{GE}$
- Terminate if performance improvement stagnates



Different stopping points while tuning CART on the `titanic` data depending on which termination criterion is used

# PRACTICAL ASPECTS OF HPO

- Warm starts
  - Evaluations (e.g., weight sharing of neural networks)
  - Optimization (initializing with HPCs that worked well before)
- Control of execution
  - Parallelizability of HPO algorithms differs strongly
  - HPO execution can be parallelized at different levels (outer resampling, iteration, evaluation, inner resampling, model fit)

