Schmidt Sciences

National Oceanography Centre

CALIPSO.

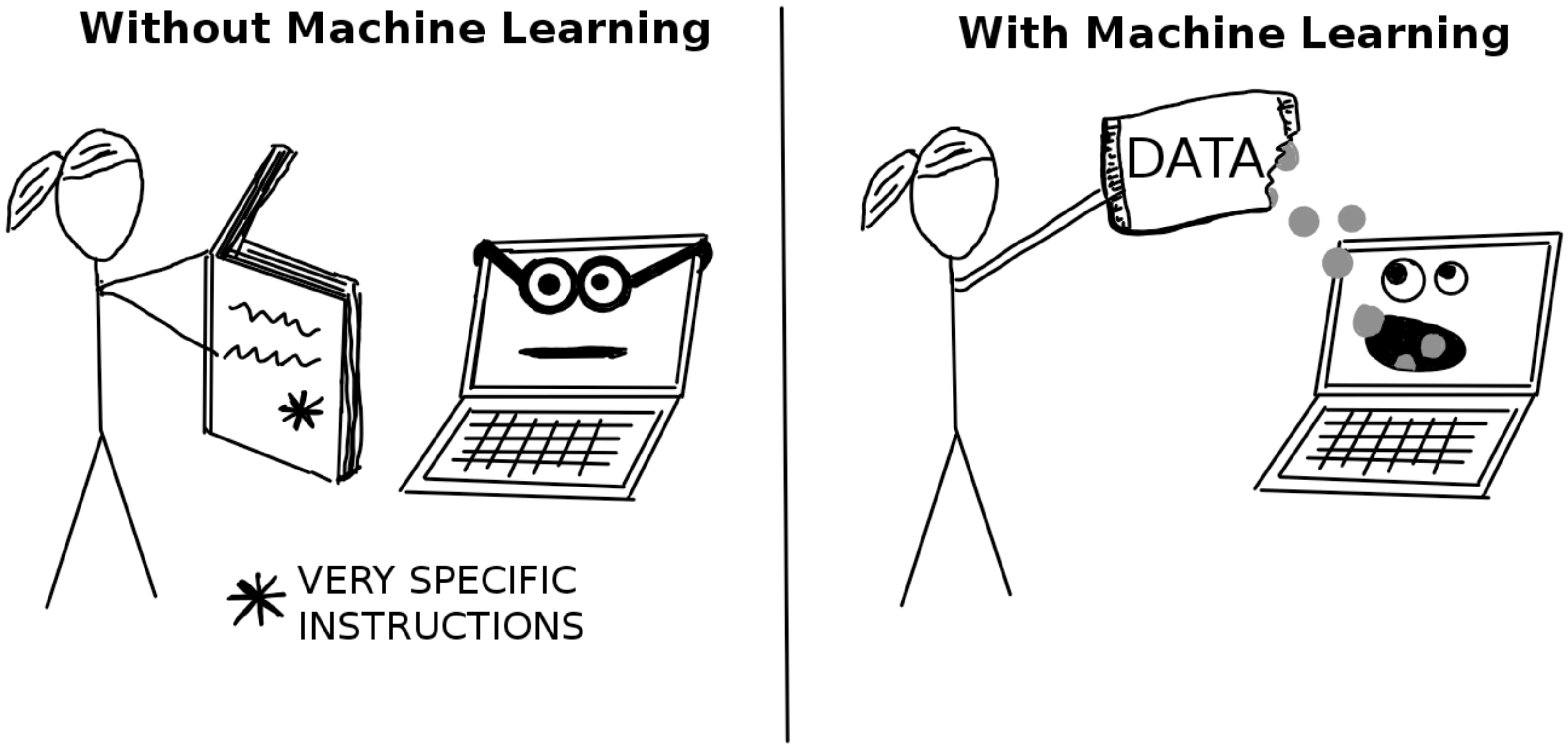
Machine Learning for feature detection

Linking zooplankton diversity and particulate organic carbon (POC)

Thelma Panaïotis
thelma.panaiotis@noc.ac.uk

11/07/2024

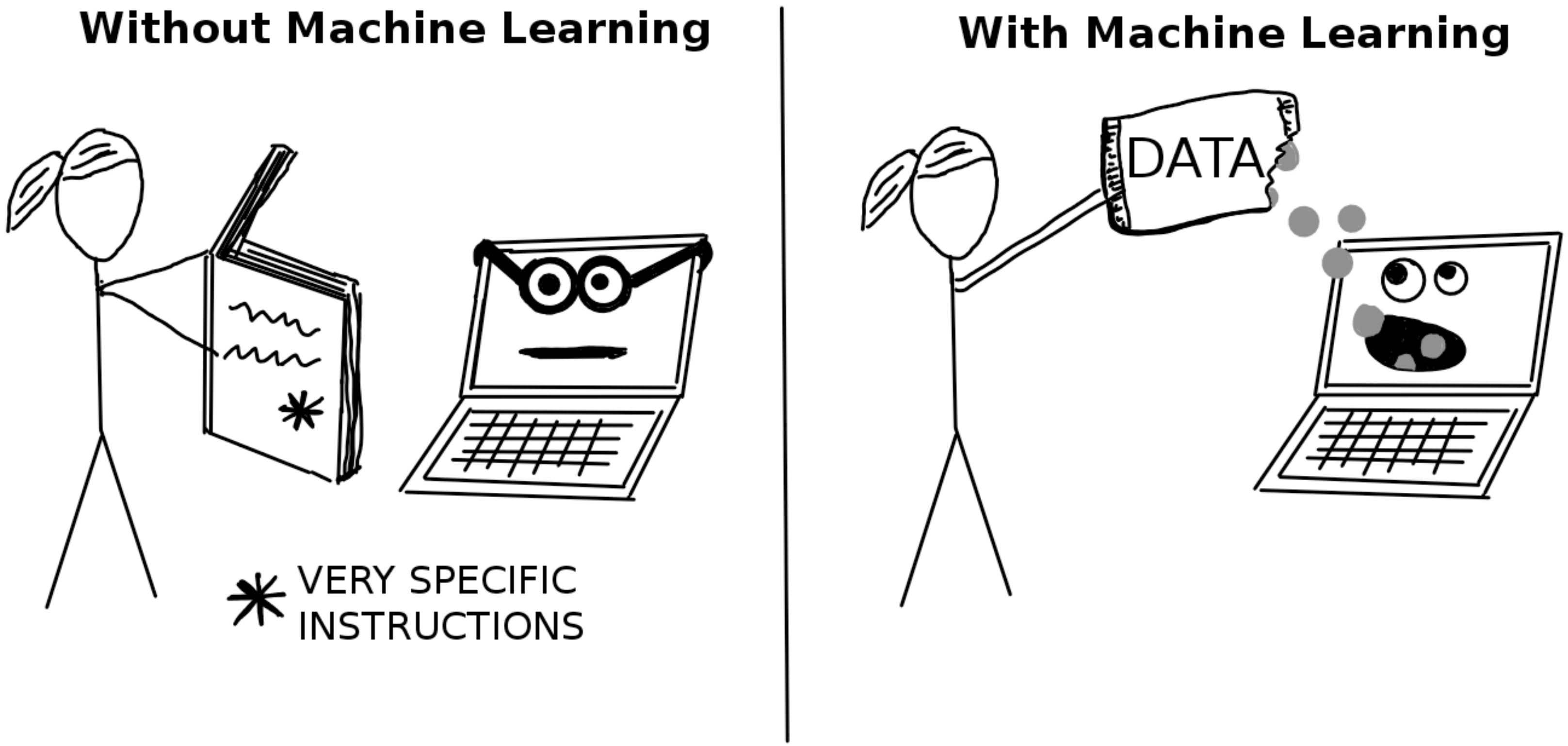
A Machine Learning definition



Christoph Molnar
https://christophmolnar.com/

A Machine Learning definition

ML: finding patterns in data, without specific instruction, and possibly predicting outcome for new data



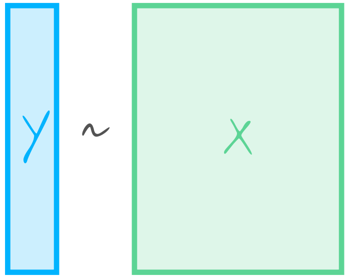
Christoph Molnar
https://christophmolnar.com/

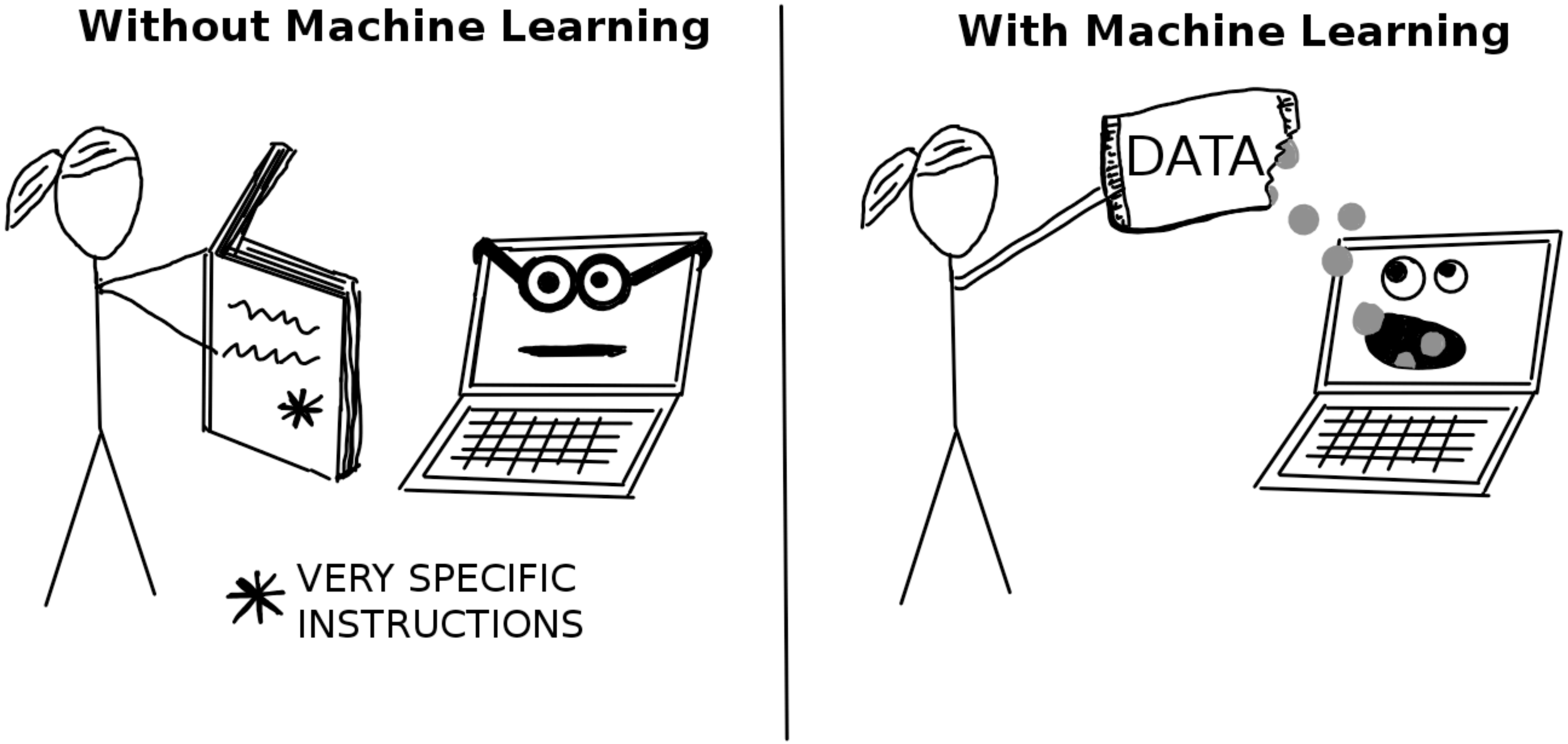
A Machine Learning definition

National Oceanography Centre

ML: finding patterns in data, without specific instruction, and possibly predicting outcome for new data

Supervised ML: relating response variable(s) (target) to predictors (features)





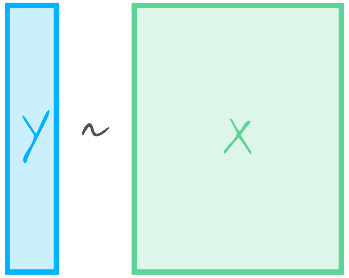
Christoph Molnar
https://christophmolnar.com/

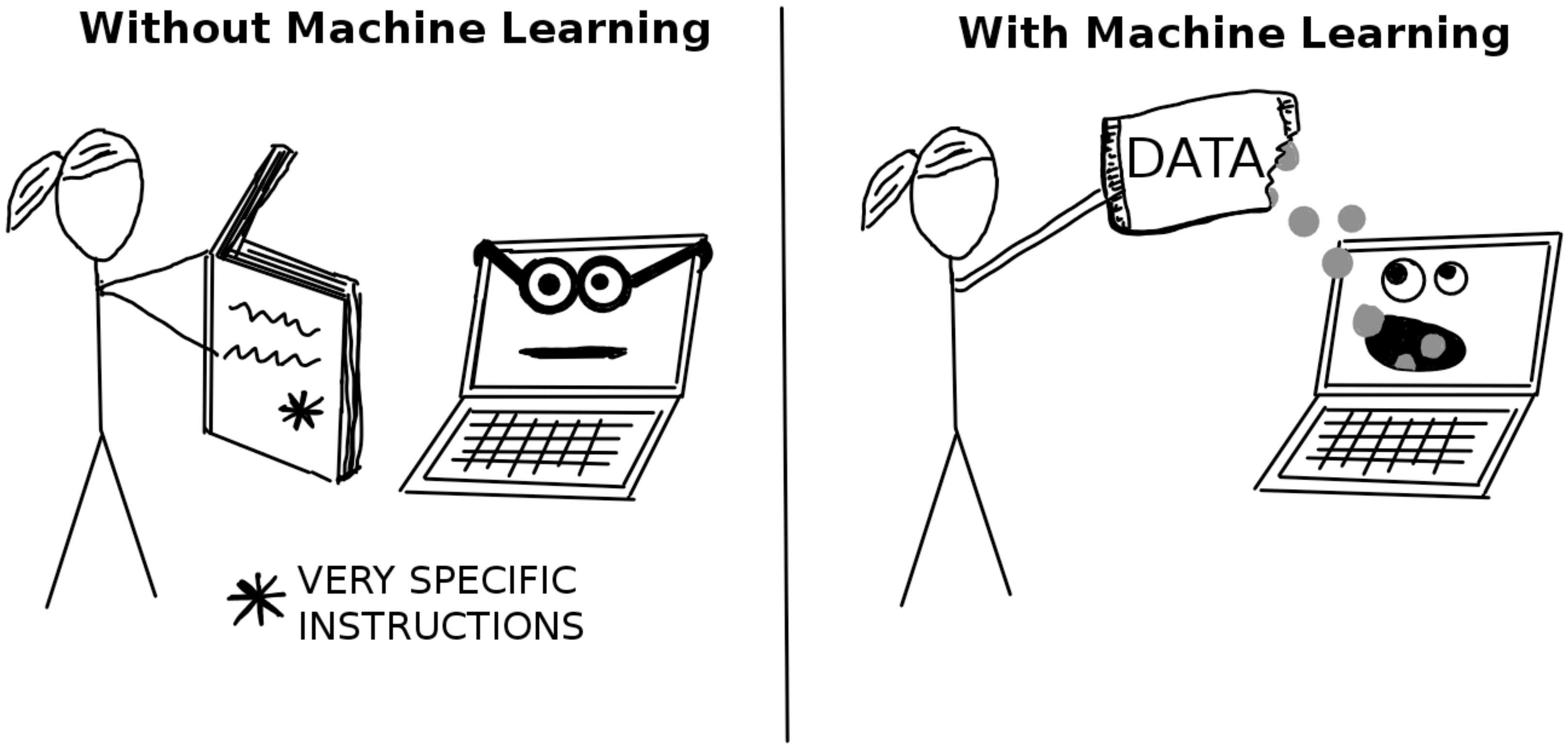
A Machine Learning definition

ML: finding patterns in data, <u>without specific instruction</u>, and possibly predicting outcome for new data

Supervised ML: relating response variable(s) (target) to predictors (features)

importance?



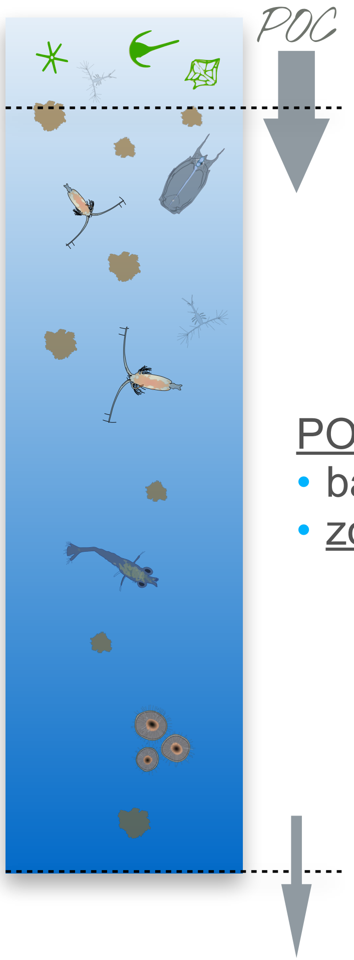


Christoph Molnar
https://christophmolnar.com/

Response *Predictors*

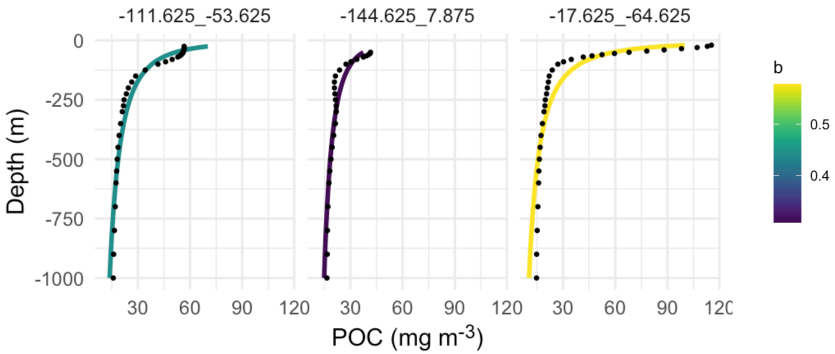
Context: relating POC attenuation to zooplankton diversity

National Oceanography Centre



POC attenuation
- bacteria
- zooplankton

Martin's b



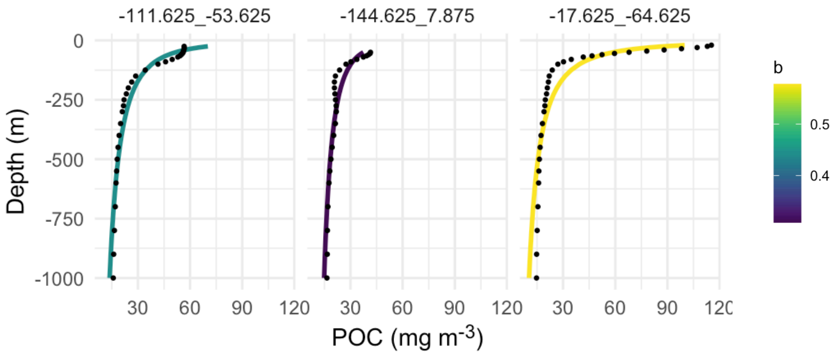
Response Predictors

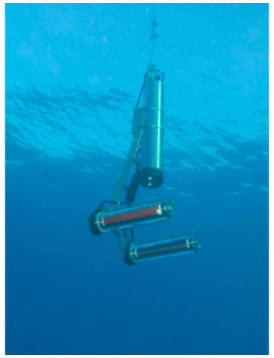
Context: relating POC attenuation to zooplankton diversity

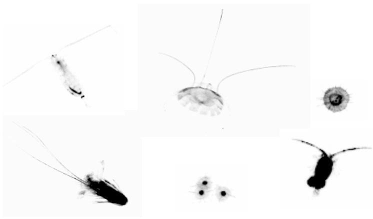
National Oceanography Centre

POC attenuation
- bacteria
- zooplankton

Martin's b







~30 descriptors:
- taxonomy
- morphology
- size spectra

Response *Predictors*

Context: relating POC attenuation to zooplankton diversity

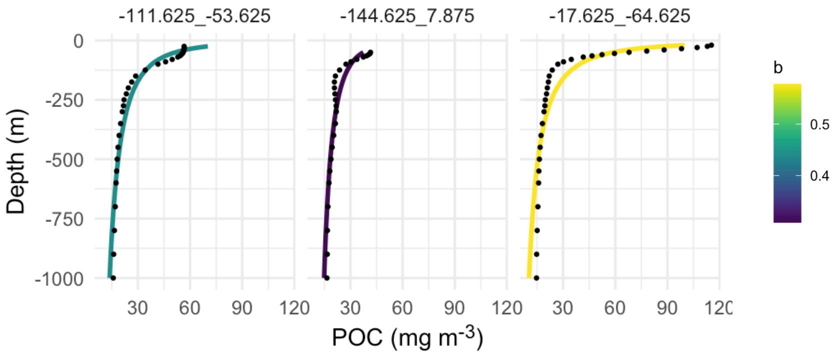
National Oceanography Centre

Euphotic depth

1000 m

POC attenuation
- bacteria
- zooplankton

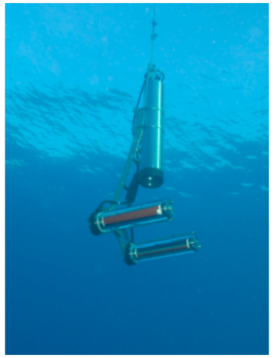
Martin's b

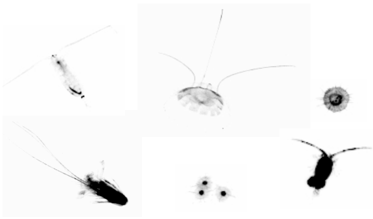


Predict

Explained variance?

Importance?





~30 descriptors:
- taxonomy
- morphology
- size spectra

Response Predictors

Context: relating POC attenuation to zooplankton diversity

National Oceanography Centre

Boosted regression trees

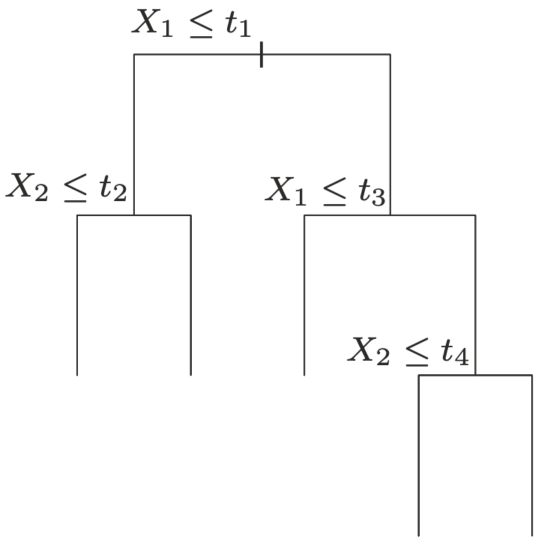
Response   Predictors

Context: relating POC attenuation to zooplankton diversity

National Oceanography Centre

Boosted regression trees

Binary splits to relate response to predictors



Response Predictors

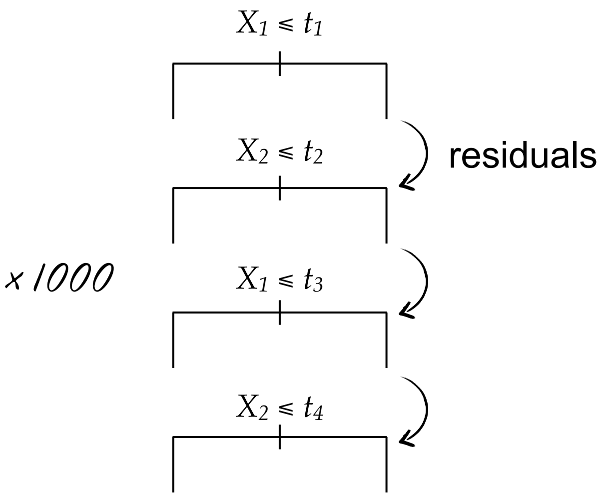
Context: relating POC attenuation to zooplankton diversity

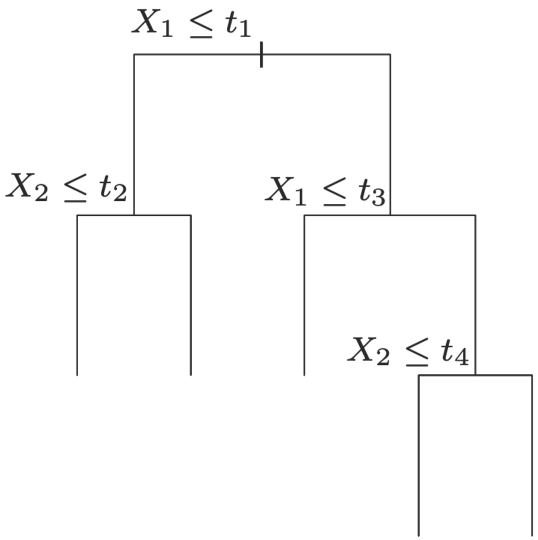
National Oceanography Centre

Boosted regression trees

Combining many small models

Binary splits to relate response to predictors





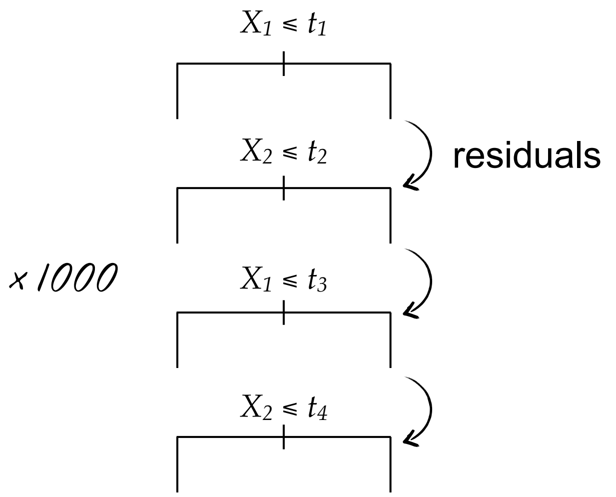
Response *Predictors*

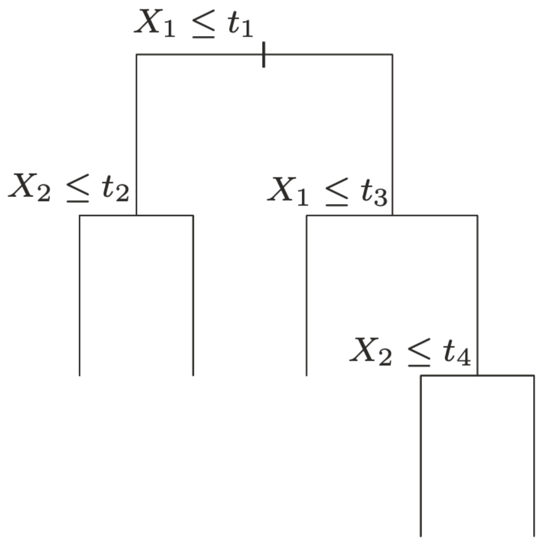
Context: relating POC attenuation to zooplankton diversity

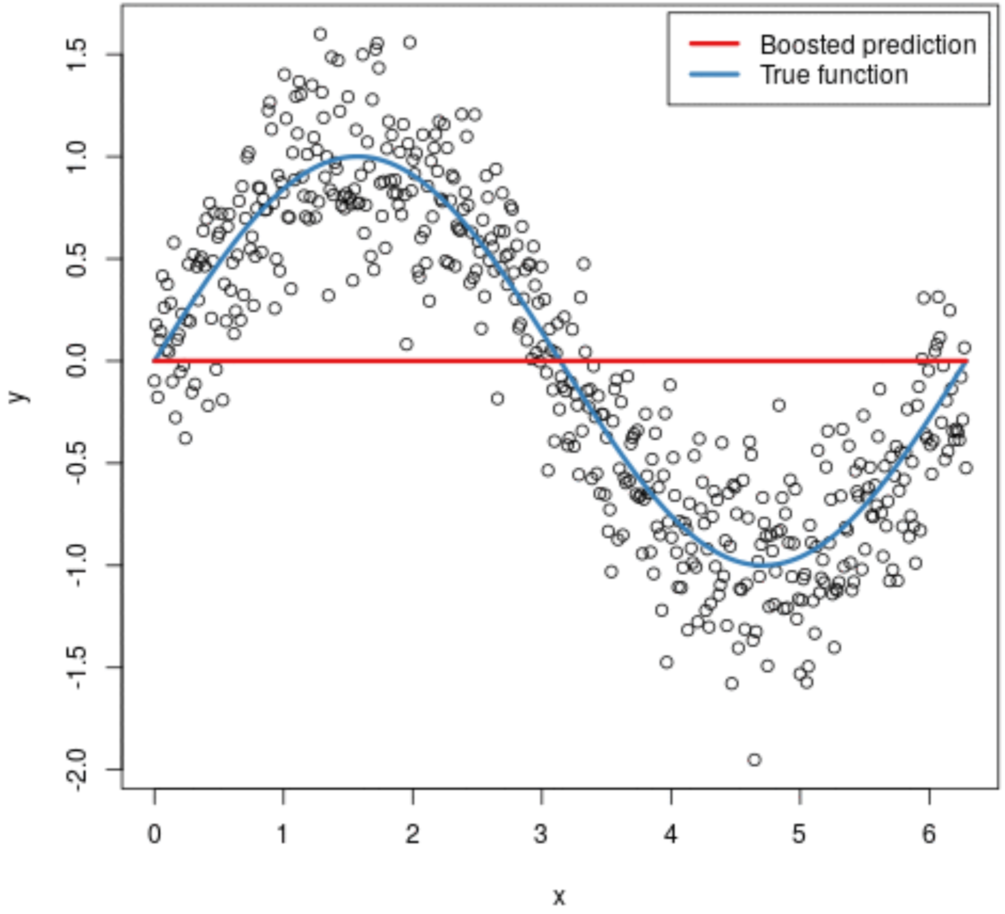
Boosted regression trees

Combining many small models

Binary splits to relate response to predictors







https://github.com/bgreenwell

Response Predictors

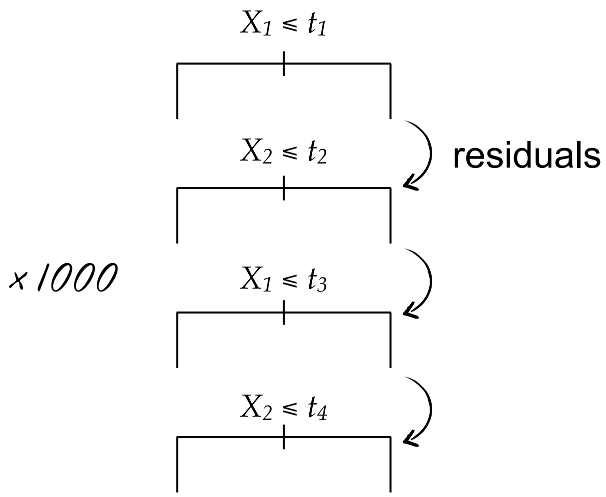
Context: relating POC attenuation to zooplankton diversity

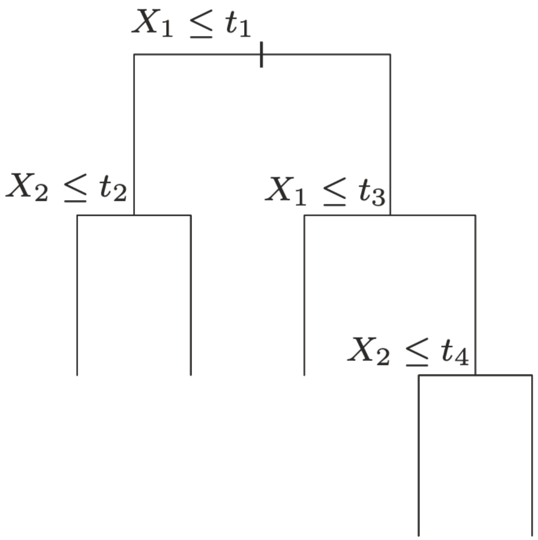
National Oceanography Centre

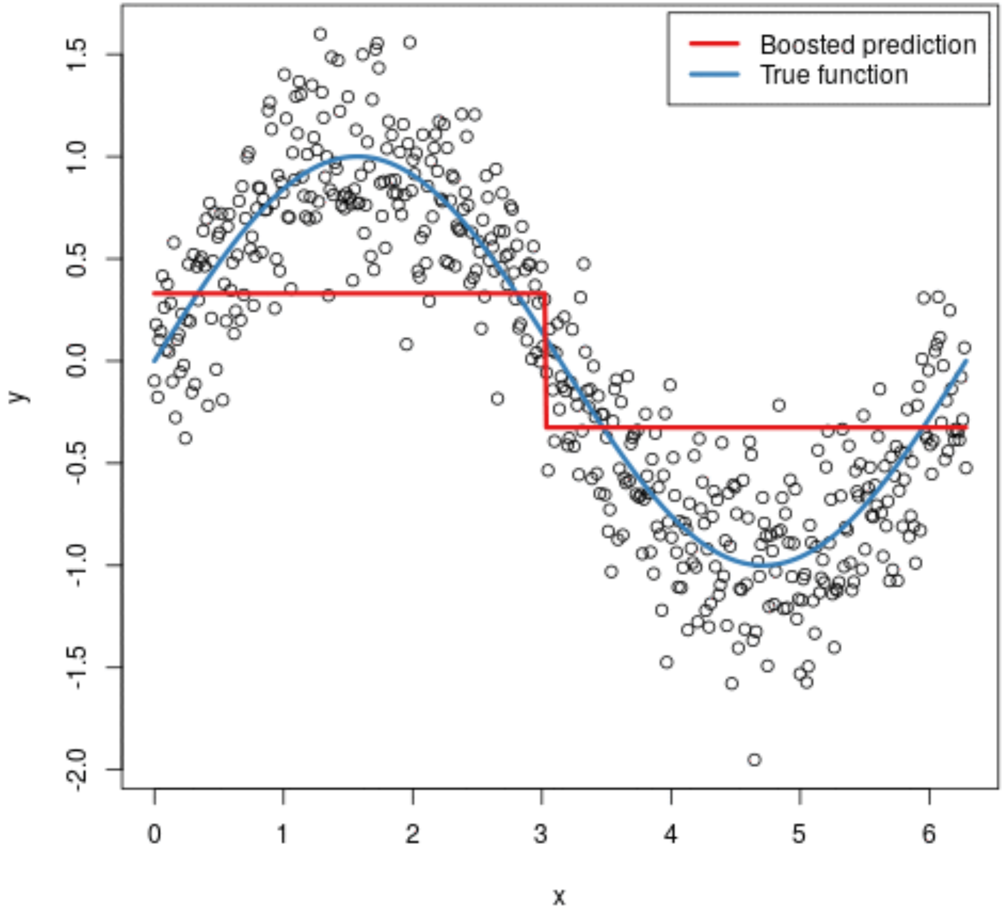
Boosted regression trees

Combining many small models

Binary splits to relate response to predictors







Context: relating POC attenuation to zooplankton diversity

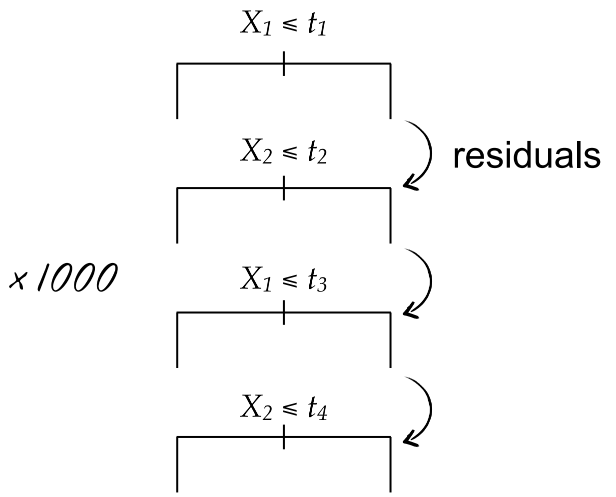
Response *Predictors*

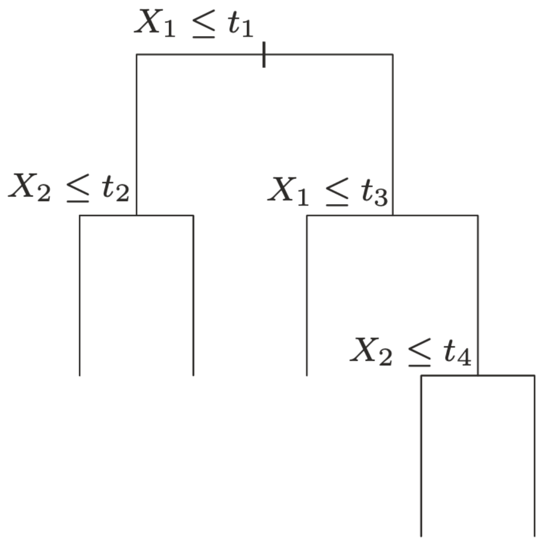
National Oceanography Centre

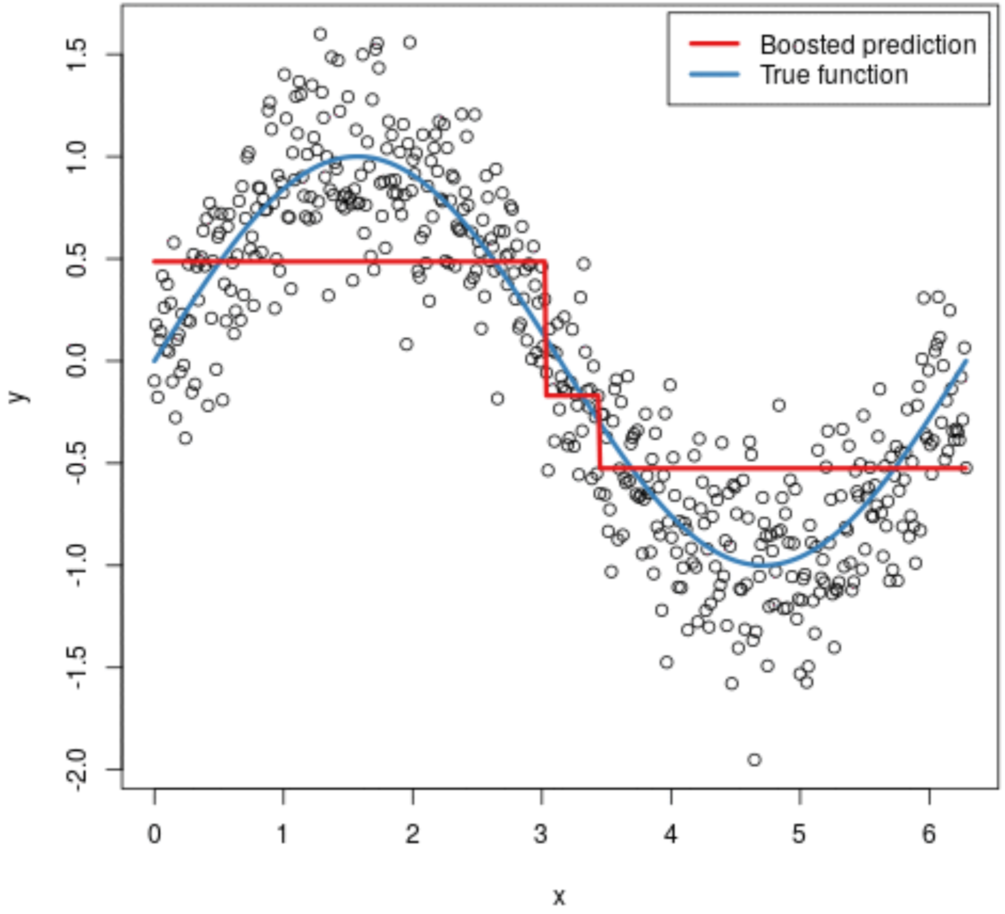
Boosted regression trees

Combining many small models

Binary splits to relate response to predictors







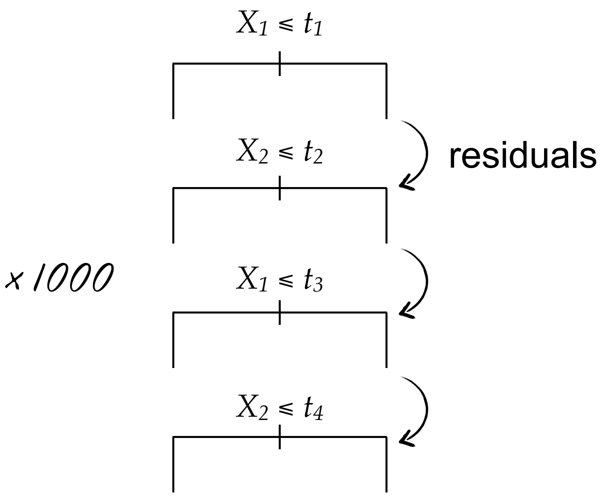
Response Predictors

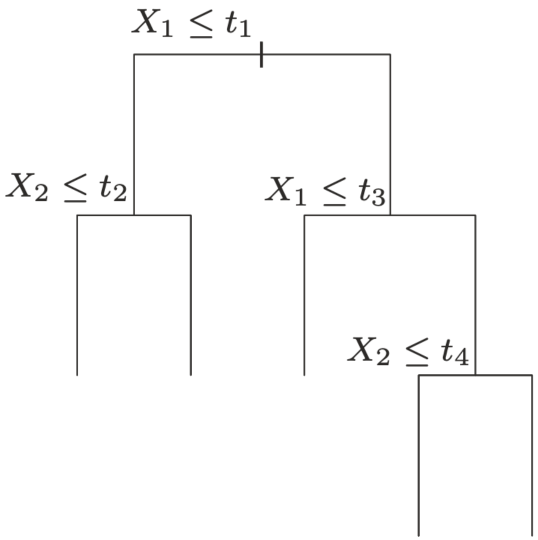
Context: relating POC attenuation to zooplankton diversity

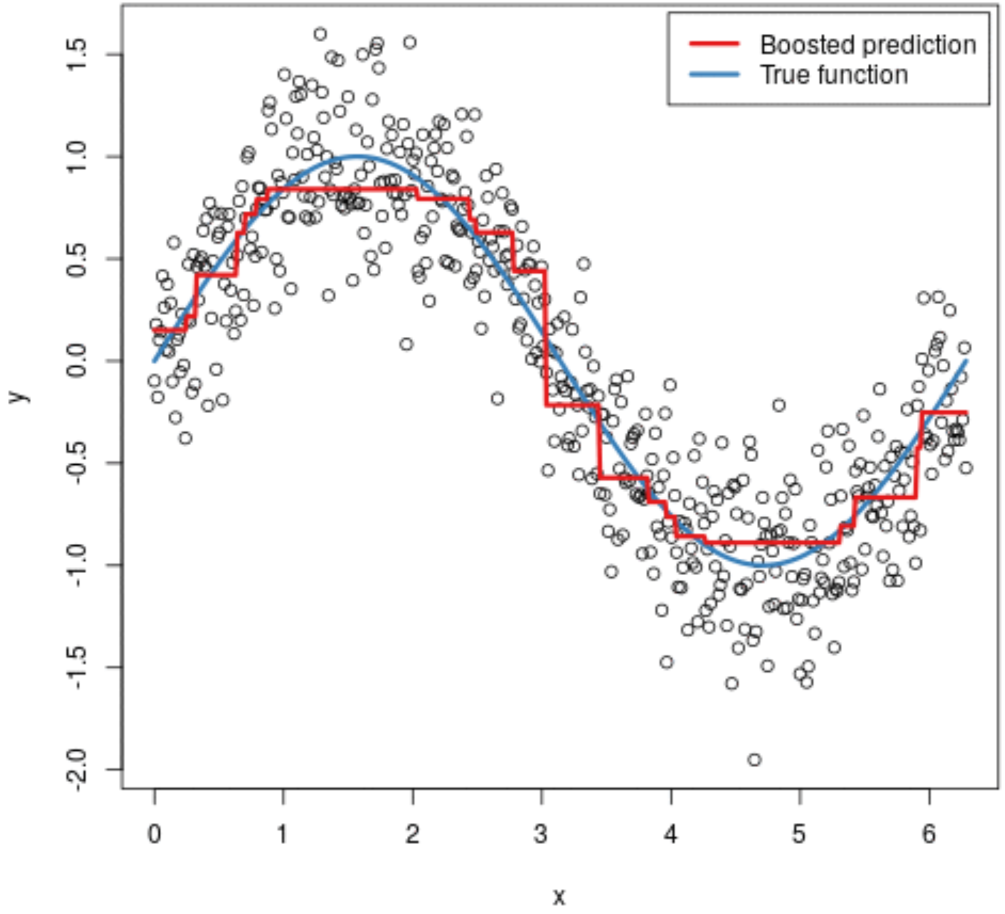
Boosted regression trees

Combining many small models

Binary splits to relate response to predictors







Response *Predictors*

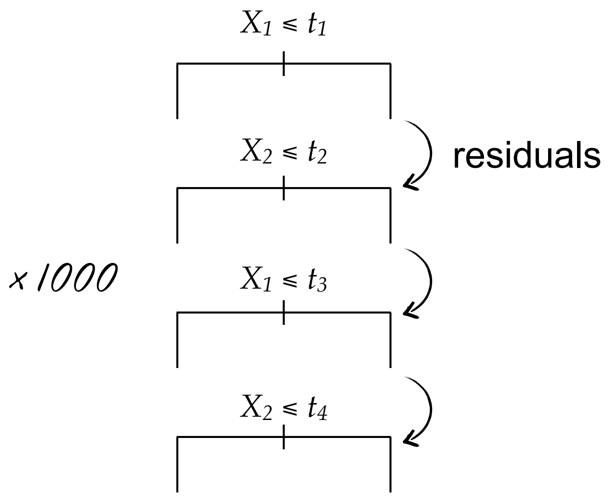
Context: relating POC attenuation to zooplankton diversity

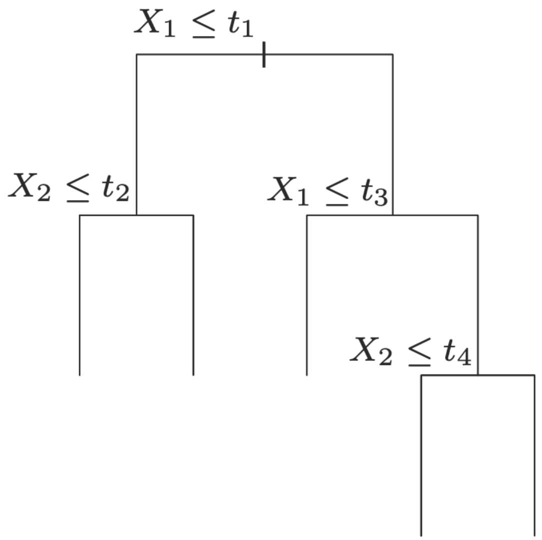
National Oceanography Centre

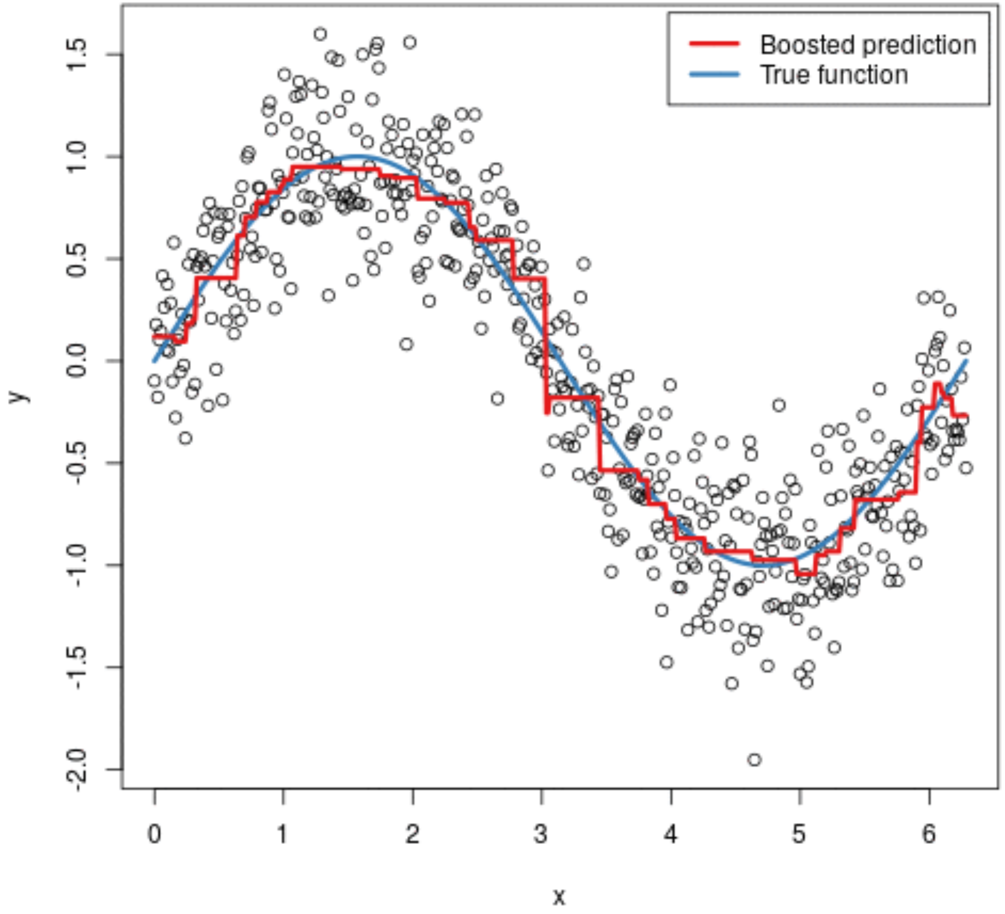
Boosted regression trees

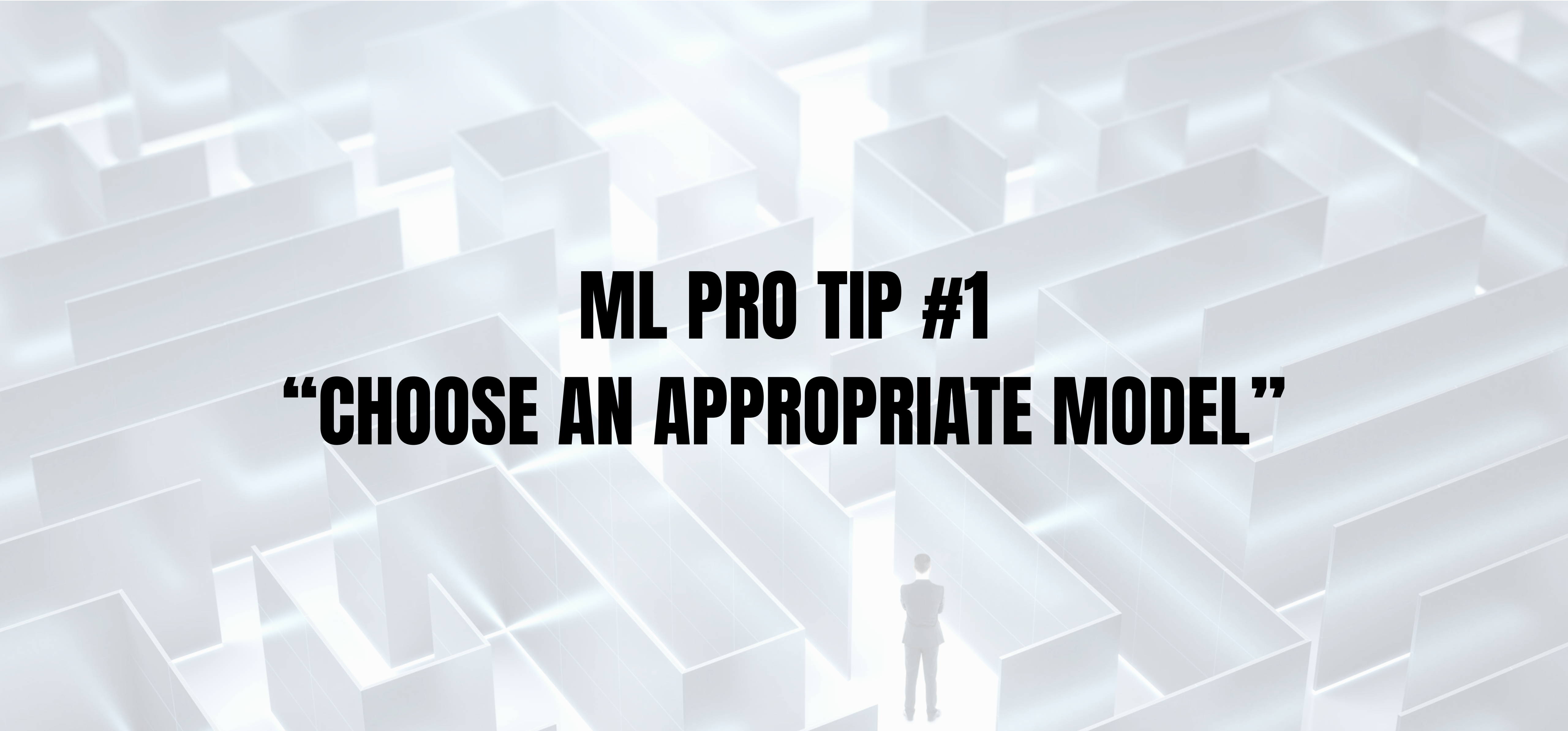
Combining many small models

Binary splits to relate response to predictors







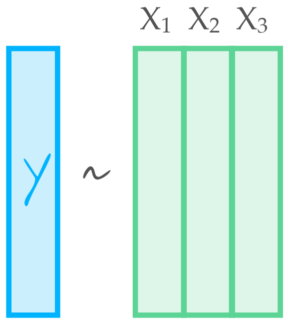


ML PRO TIP #1

“CHOOSE AN APPROPRIATE MODEL”

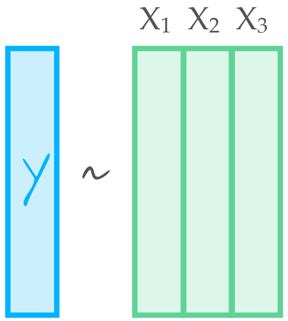
Tree ensembles are fantastic*

*if input is numeric



Tree ensembles are fantastic*

*if input is numeric



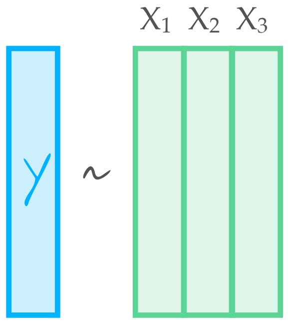
Many advantages:

- input flexibility (type, distribution, missing values, relevance)
- complex non-linear relationships + interactions
- good predictive power
- interpretable
- many implementations (R, Python)

Elith et al., 2008; Hastie et al., 2009

Tree ensembles are fantastic*

*if input is numeric



Many advantages:

- input flexibility (type, distribution, missing values, relevance)
- complex non-linear relationships + interactions
- good predictive power
- interpretable
- many implementations (R, Python)

Elith et al., 2008; Hastie et al., 2009

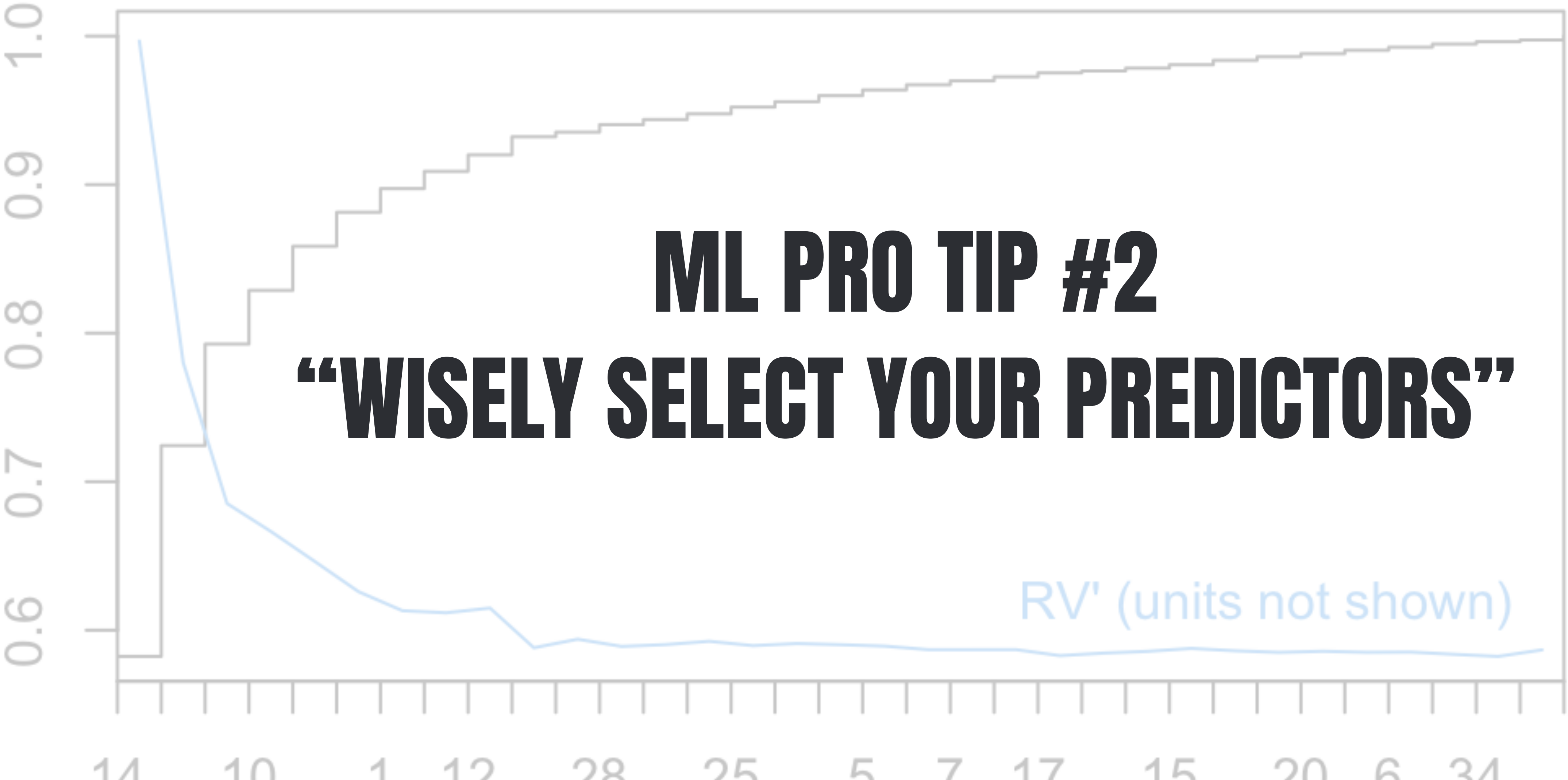
Classification: tree ensembles (RF) > neural network (ANN/MLP)

Regression?

Fernández-Delgado et al., 2014

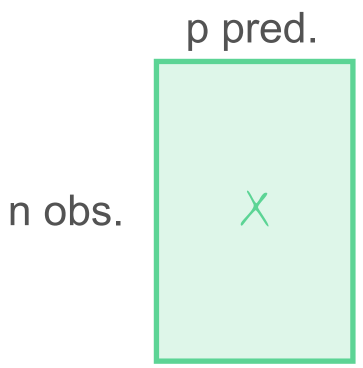
ML PRO TIP #2

“WISELY SELECT YOUR PREDICTORS”



Less is more

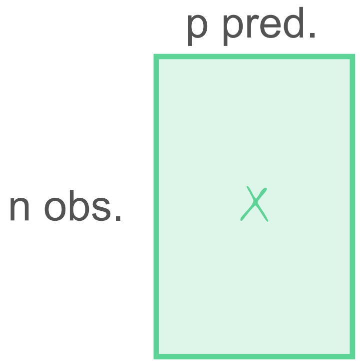
Number of predictors/features VS number of observations.



$n \gg p$

Less is more

Number of predictors/features VS number of observations.

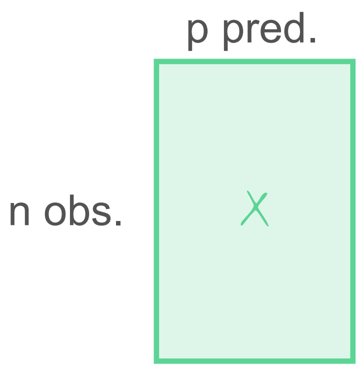


$\boxed{n \gg p}$

Trees can ignore non-relevant predictors.

Less is more

Number of predictors/features VS number of observations.



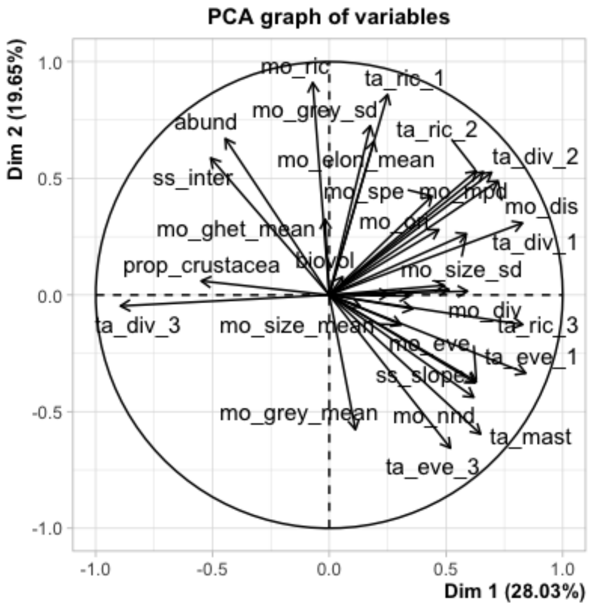
$n \gg p$

Trees can ignore non-relevant predictors.

Parsimony

Feature selection
- PCA
- Escoufier's equivalent vectors
- VIF

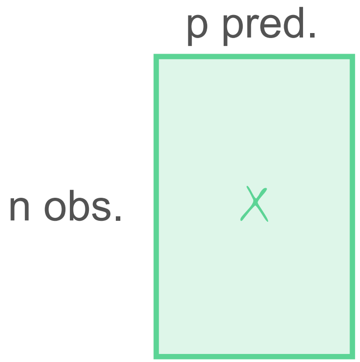
Feature engineering
- PCA: use PCs as predictors



Less is more

National Oceanography Centre

Number of predictors/features VS number of observations.



$n \gg p$

Trees can ignore non-relevant predictors.

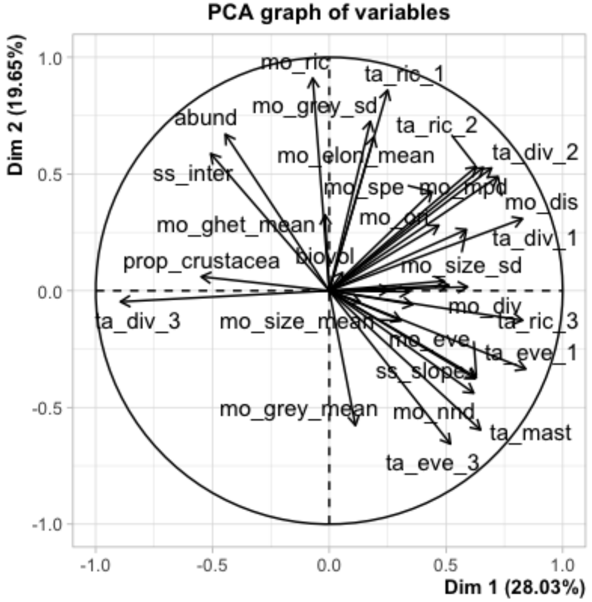
Parsimony

Feature selection

- PCA
- Escoufier's equivalent vectors
- VIF

Feature engineering

- PCA: use PCs as predictors



Correlated features?
Depends on your model. Tree ensembles are fairly robust.

Hastie et al., 2009

ML PRO TIP #3
“MANAGE YOUR DATA BUDGET”

Need to spend the data

National Oceanography Centre

~80% Train VS Test ~20%

Fit the model

Evaluate model performance, **at the very end**, **single use**.

How the model will perform with new data?

Regression
- R^2
- RMSE

Classification
- Accuracy
- Precision
- Recall

Need to spend the data

~80% **Train VS Test** ~20%

Fit the model

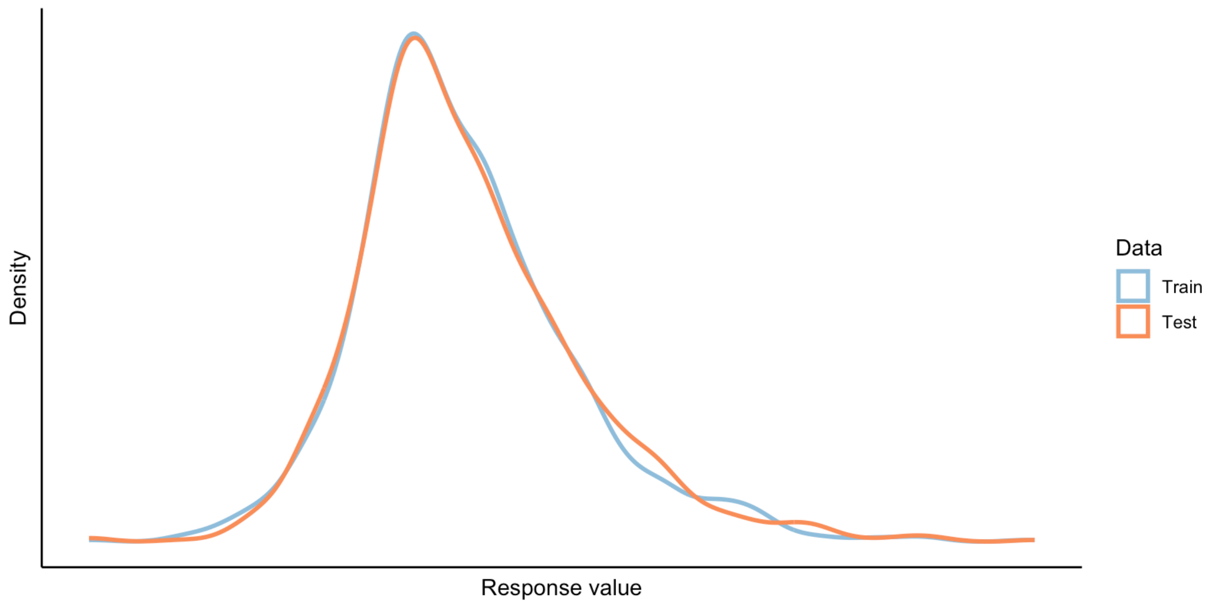
Evaluate model performance, **at the very end**, **single use**.

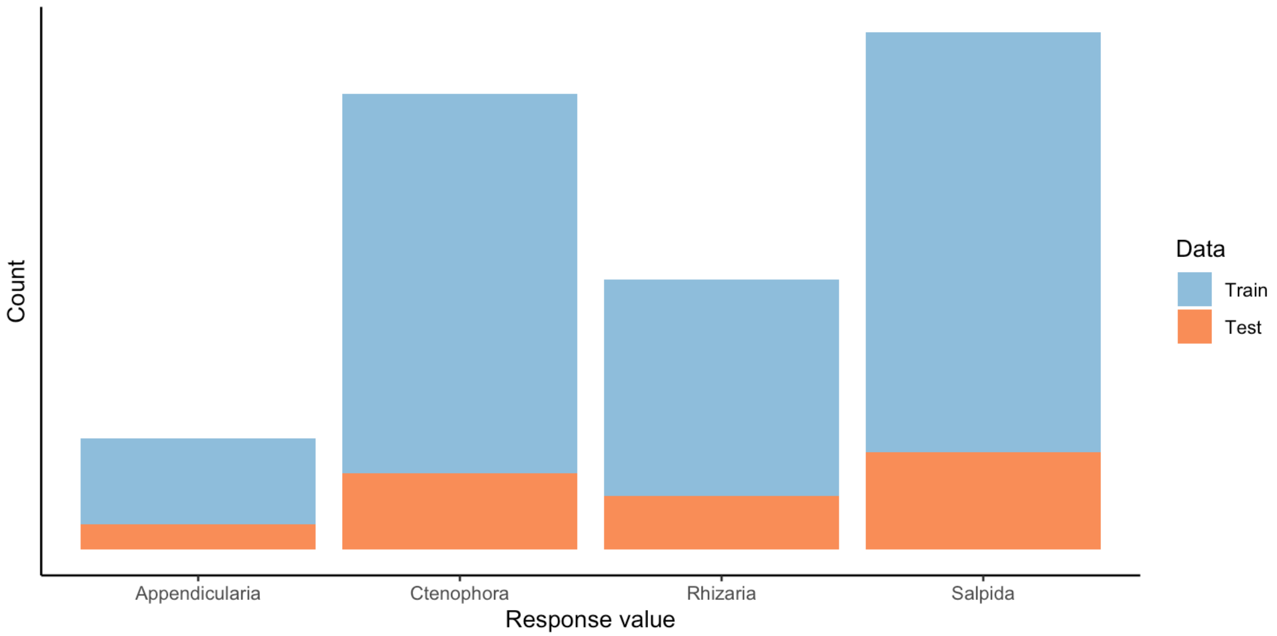
How the model will perform with new data?

Regression
- R^2
- RMSE

Classification
- Accuracy
- Precision
- Recall

Test set representative of the training set.





Need to spend the data

~80% **Train VS Test** ~20%

Fit the model

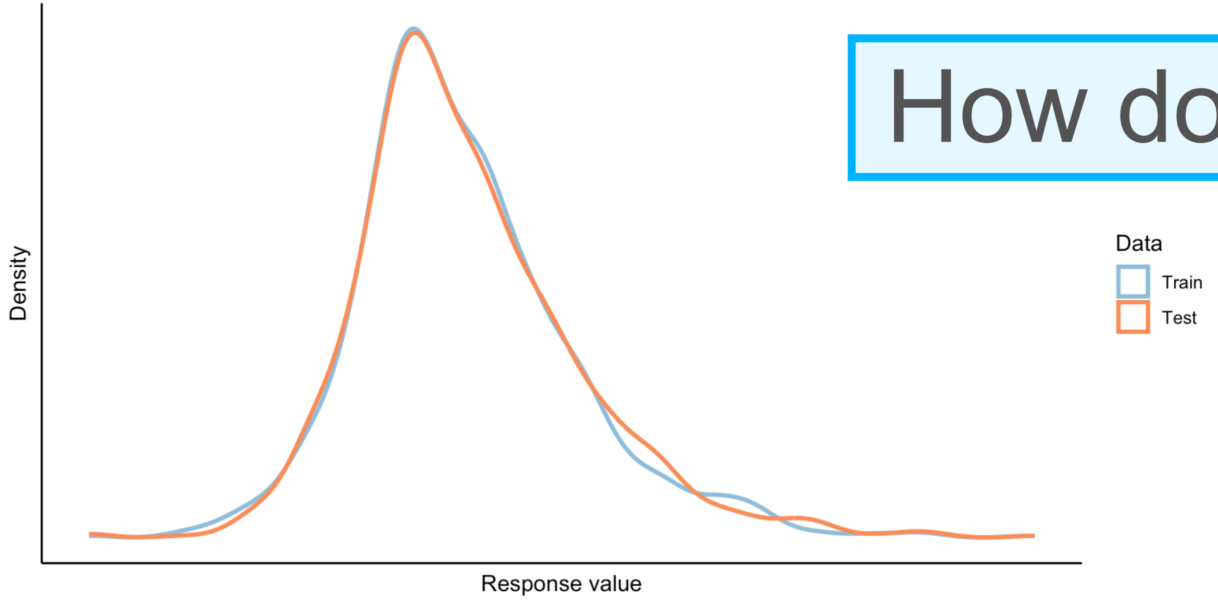
Evaluate model performance, **at the very end**, **single use**.

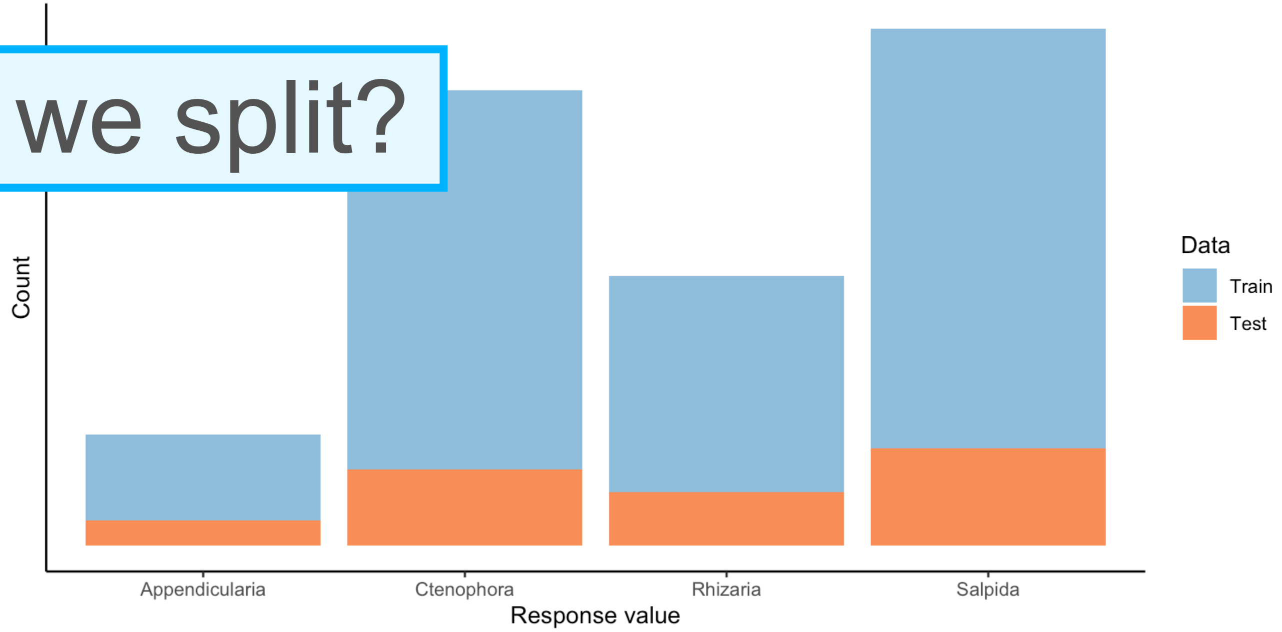
How the model will perform with new data?

Regression
- R^2
- RMSE

Classification
- Accuracy
- Precision
- Recall

Test set representative of the training set.

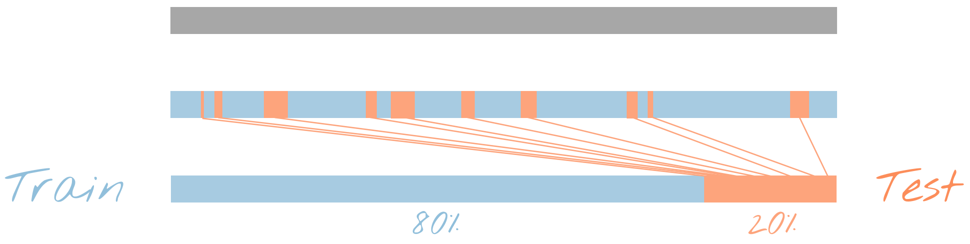




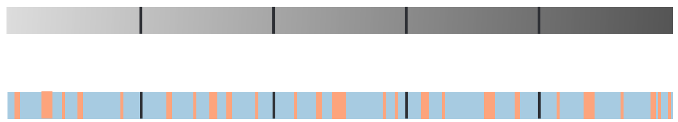
How do we split?

How to split your data

Random

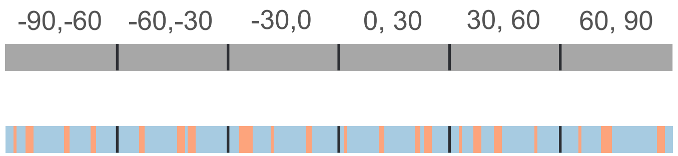


Quantiles of response variable

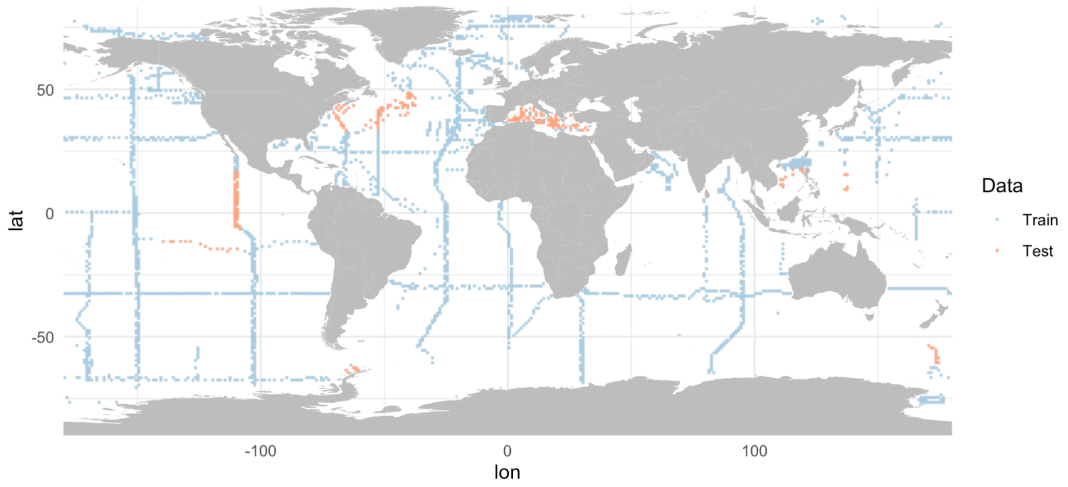


Groups

(e.g. latitude bands)

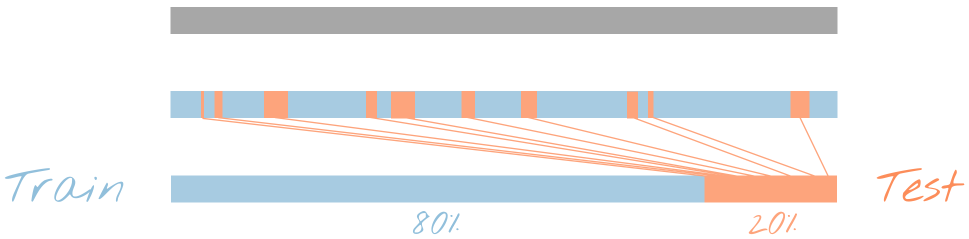


Spatial / temporal

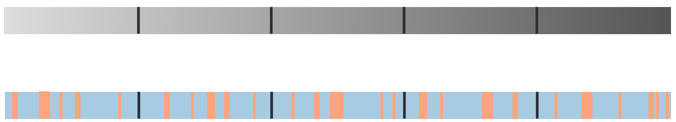


How to split your data

Random



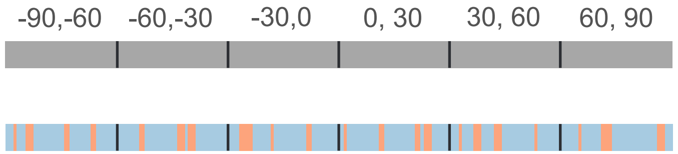
Quantiles of response variable



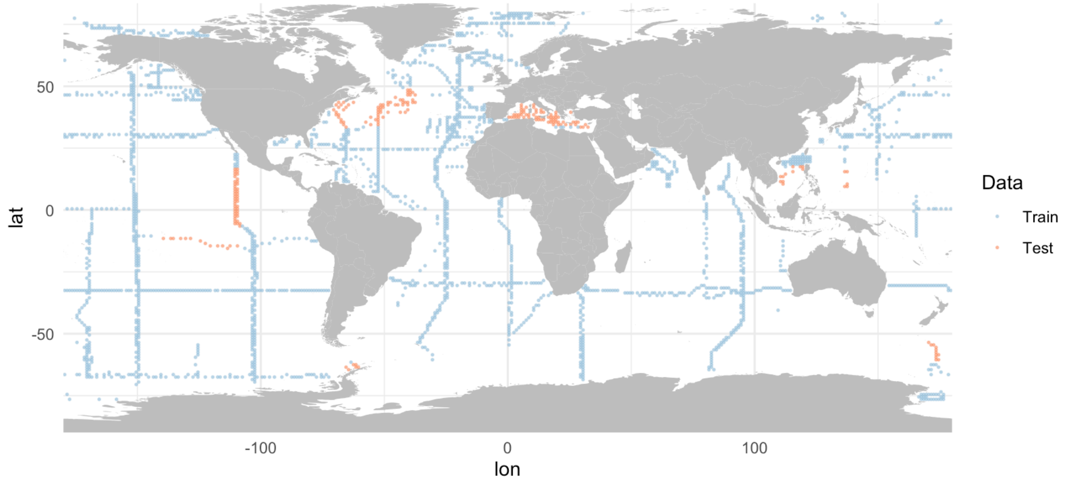
Depends on your data!

Groups

(e.g. latitude bands)



Spatial / temporal



ML PRO TIP #4

“SOME ML MODELS NEED TO BE TUNED”

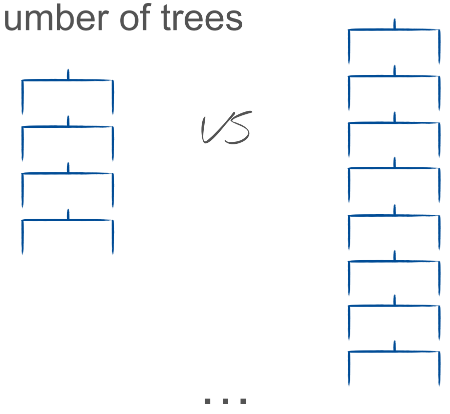
Optimising your model hyperparameters

Tree ensembles

Boosted trees

Depth of trees

Number of trees



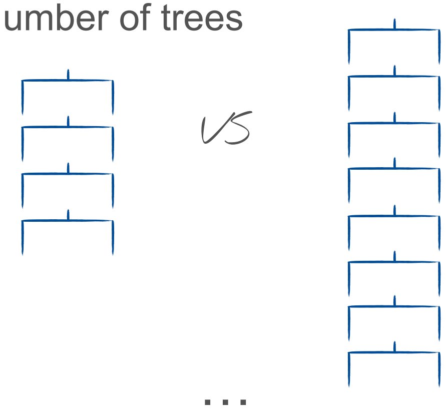
Optimising your model hyperparameters

Tree ensembles

Boosted trees

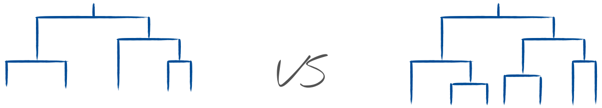
Depth of trees

Number of trees

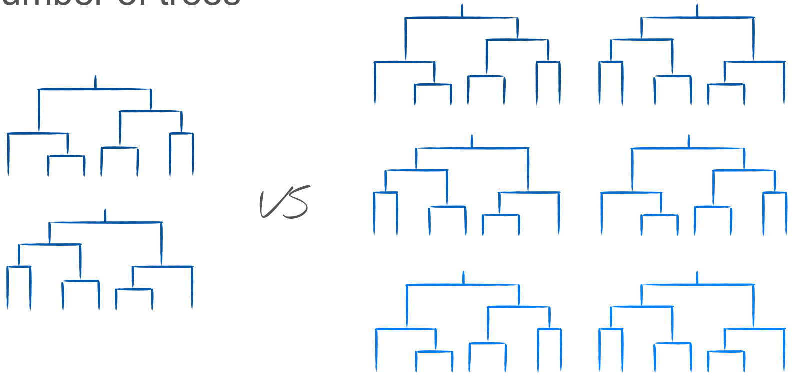


Random Forest

Depth of trees



Number of trees



Optimising your model hyperparameters

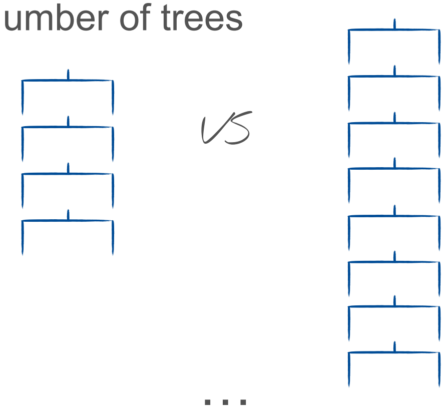
National Oceanography Centre

Tree ensembles

Boosted trees

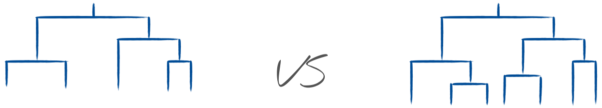
Depth of trees

Number of trees

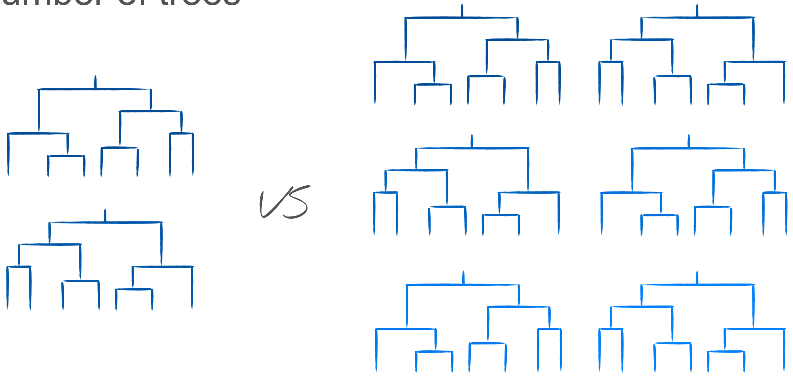


Random Forest

Depth of trees



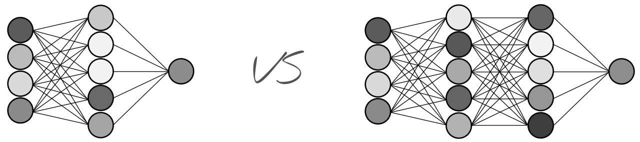
Number of trees



Neural networks (MLP/ANN)

Size of layers

Number of layers



Optimising your model hyperparameters

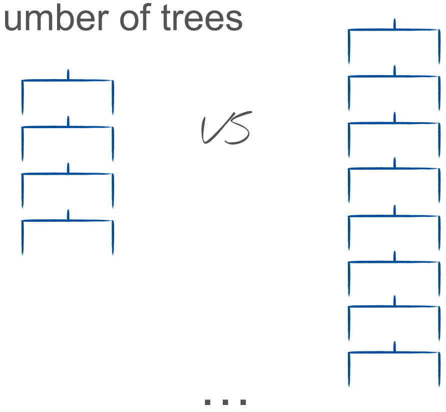
National Oceanography Centre

Tree ensembles

Boosted trees

Depth of trees

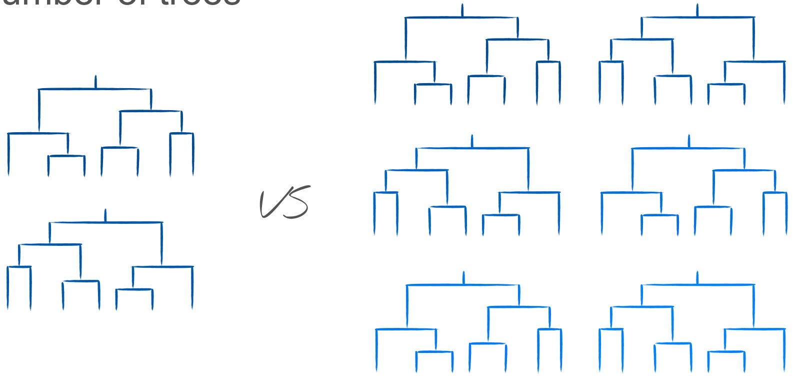
Number of trees



Random Forest

Depth of trees

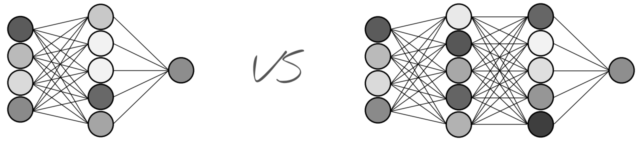
Number of trees



Neural networks (MLP/ANN)

Size of layers

Number of layers



→ to be tuned: model tuning / gridsearch

That's what the validation set is for

Example for boosted regression trees

Tuning the tree depth

That’s what the validation set is for

Example for boosted regression trees

Tuning the tree depth

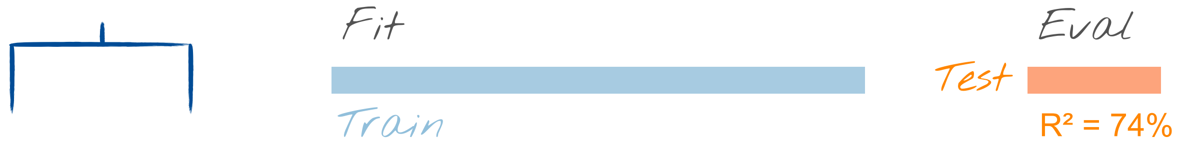
shallow

That's what the validation set is for

Example for boosted regression trees

Tuning the tree depth

shallow



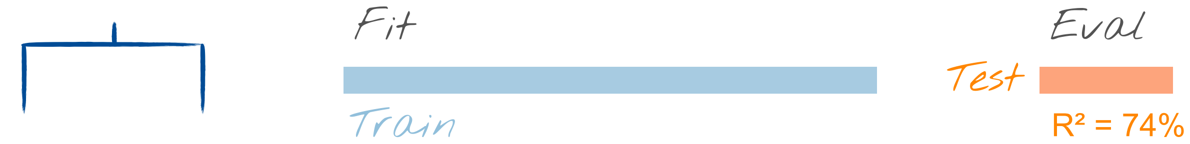
That's what the validation set is for

National Oceanography Centre

Example for boosted regression trees

Tuning the tree depth

shallow

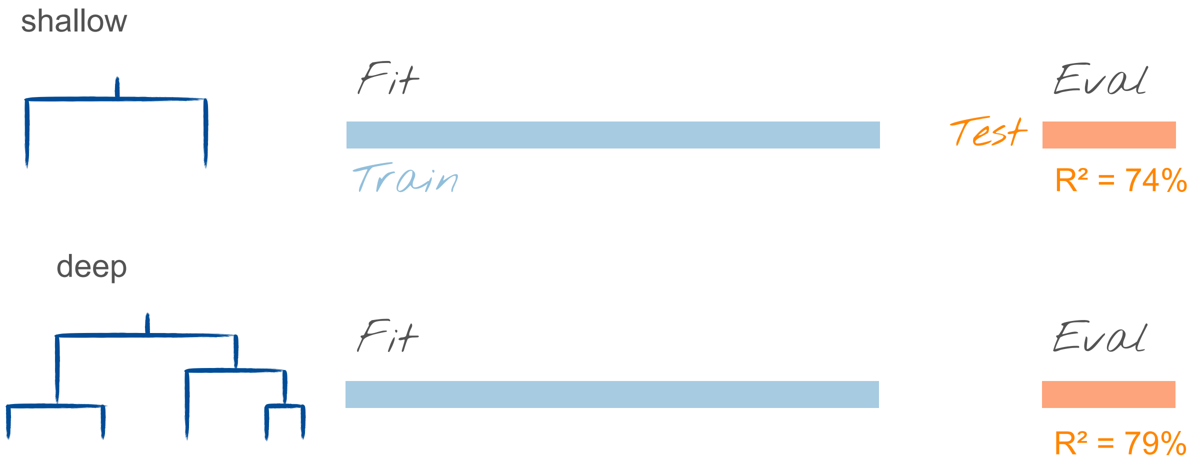


deep

That’s what the validation set is for

Example for boosted regression trees

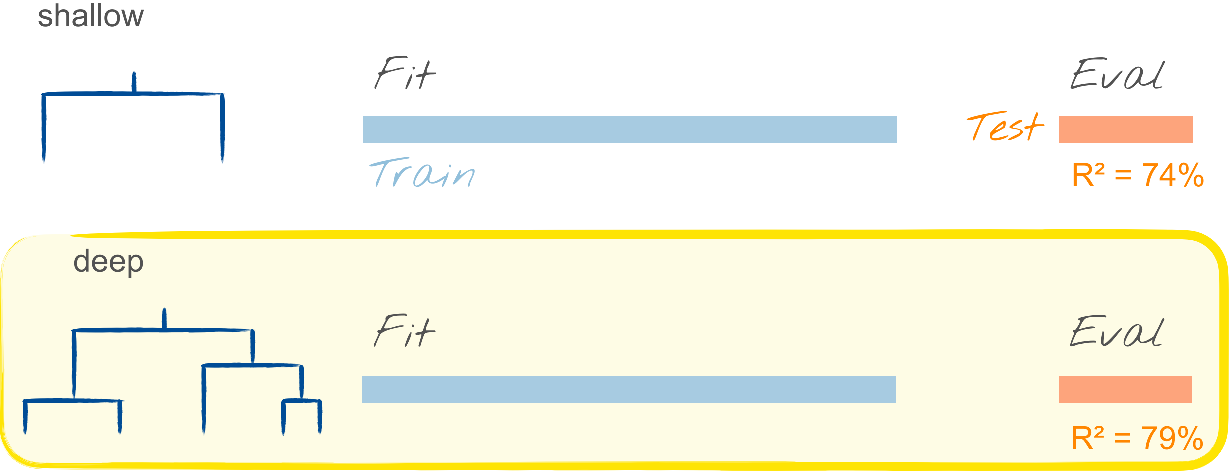
Tuning the tree depth



That’s what the validation set is for

Example for boosted regression trees

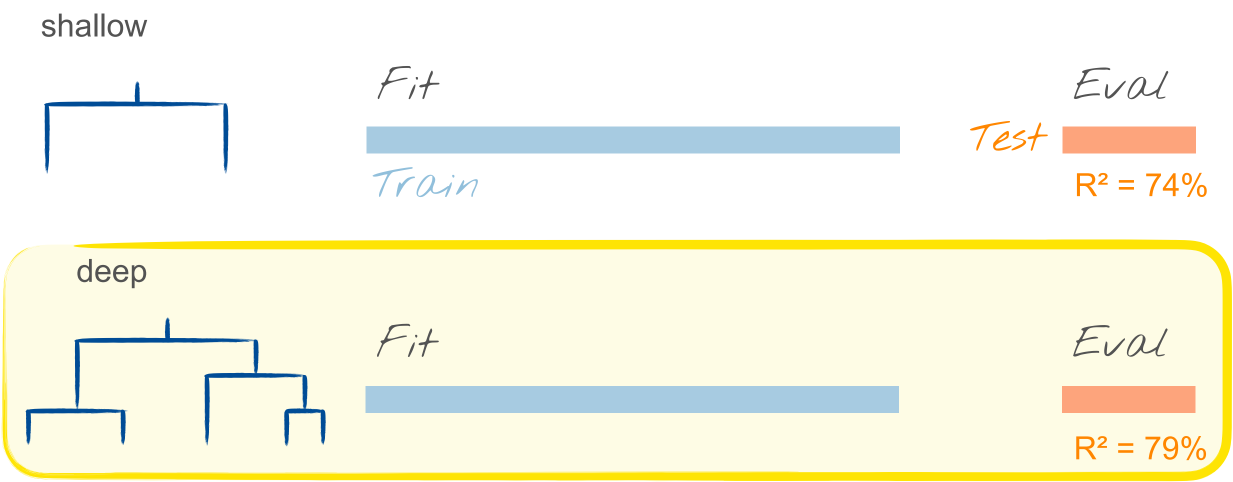
Tuning the tree depth



That's what the validation set is for

Example for boosted regression trees

Tuning the tree depth



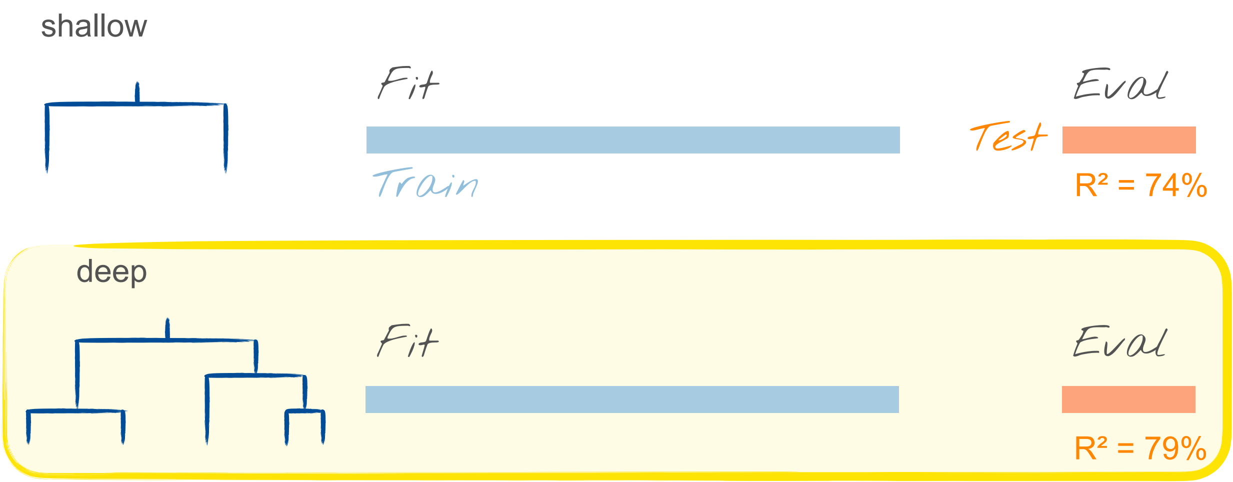
Test

optimising hyperparameters + estimating generalisation error

That's what the validation set is for

Example for boosted regression trees

Tuning the tree depth



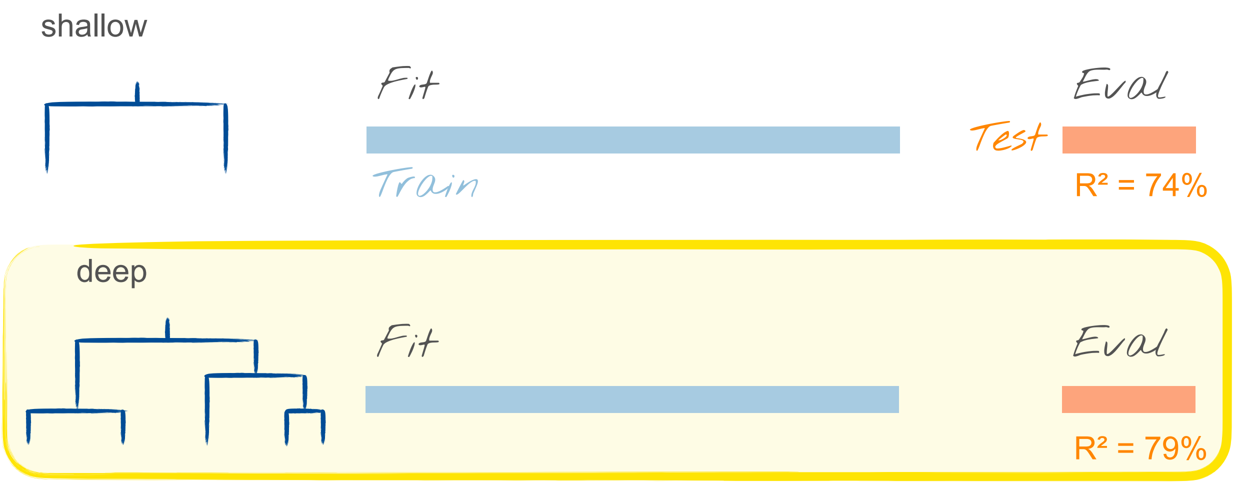
Tes

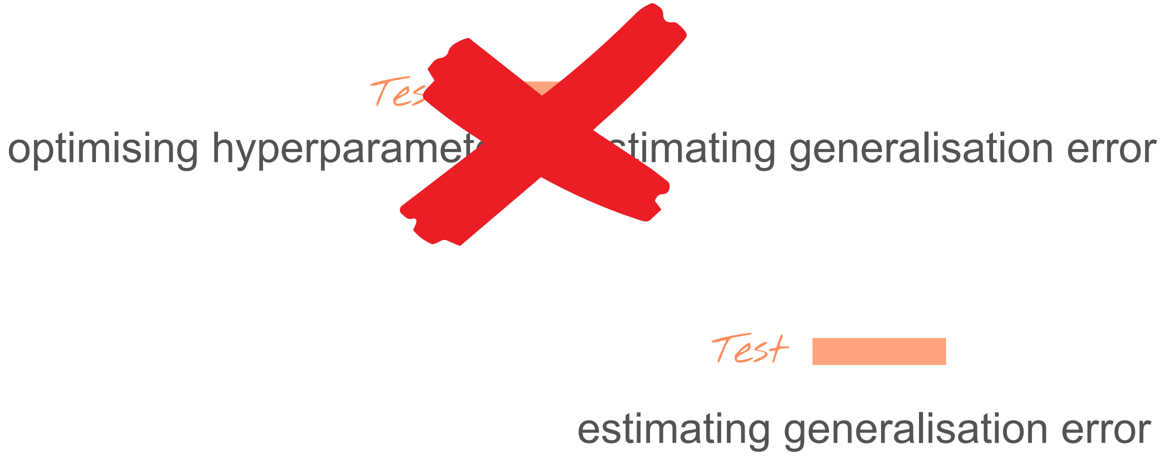
optimising hyperparamet timating generalisation error

That's what the validation set is for

Example for boosted regression trees

Tuning the tree depth



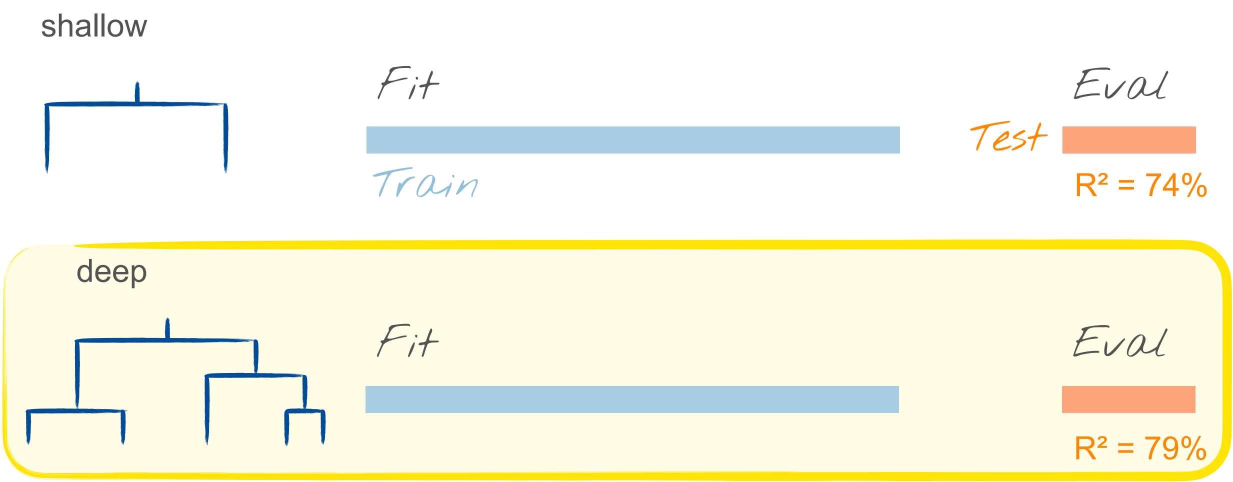


That’s what the validation set is for

National Oceanography Centre

Example for boosted regression trees

Tuning the tree depth



~~Tes~~

~~optimising hyperparamet~~ ~~stimating generalisation error~~

Test

estimating generalisation error

Validation

optimising hyperparameters

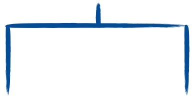
That's what the validation set is for

National Oceanography Centre

Example for boosted regression trees

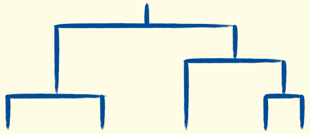
Tuning the tree depth

shallow



Fit — Train — Eval — Test — $R^2 = 74\%$

deep



Fit — Eval — $R^2 = 79\%$

~~Tes~~ ~~optimising hyperparamet~~ ~~timating generalisation error~~

Test

estimating generalisation error

Validation

optimising hyperparameters

Need 3 splits — Train — Valid — Test

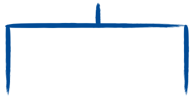
That's what the validation set is for

National Oceanography Centre

Example for boosted regression trees

Tuning the tree depth

shallow



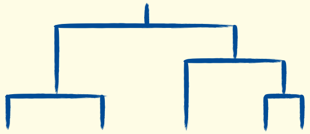
Fit Eval

Test

Train

R^2 = 74%

deep



Fit Eval

R^2 = 79%

~~Tes~~

~~optimising hyperparamet~~ ~~stimating generalisation error~~

Test

estimating generalisation error

Validation

optimising hyperparameters

Need 3 splits

Train Valid Test

Validation set used to tune the model

ML PRO TIP #5
“CROSS-VALIDATION CAN BE GREAT”

What about cross validation?

Overcome the effect of randomness in your splits

What about cross validation?

Overcome the effect of randomness in your splits

Dataset

5 folds CV

What about cross validation?

Overcome the effect of randomness in your splits

5 folds CV

What about cross validation?

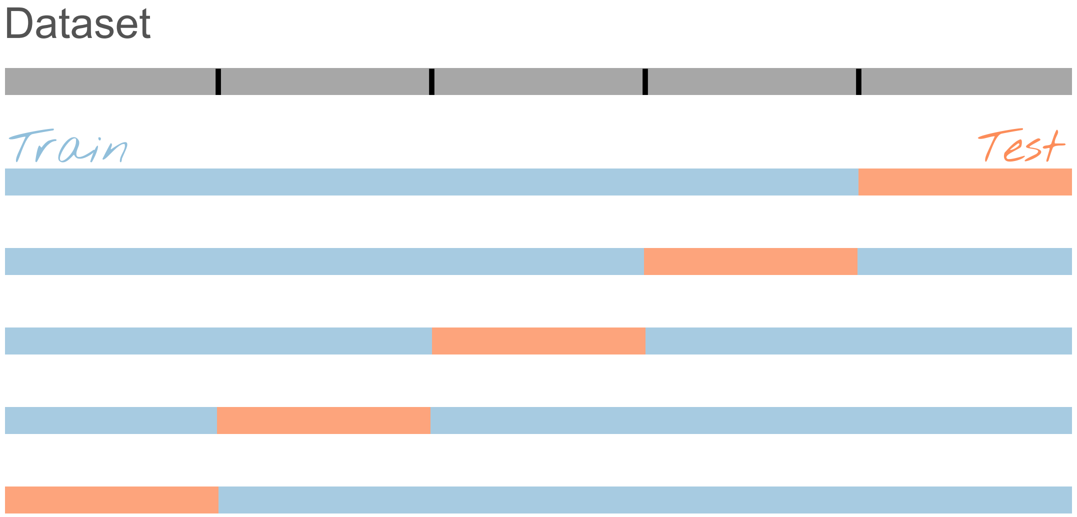
Overcome the effect of randomness in your splits

5 folds CV

→ 5 iterations

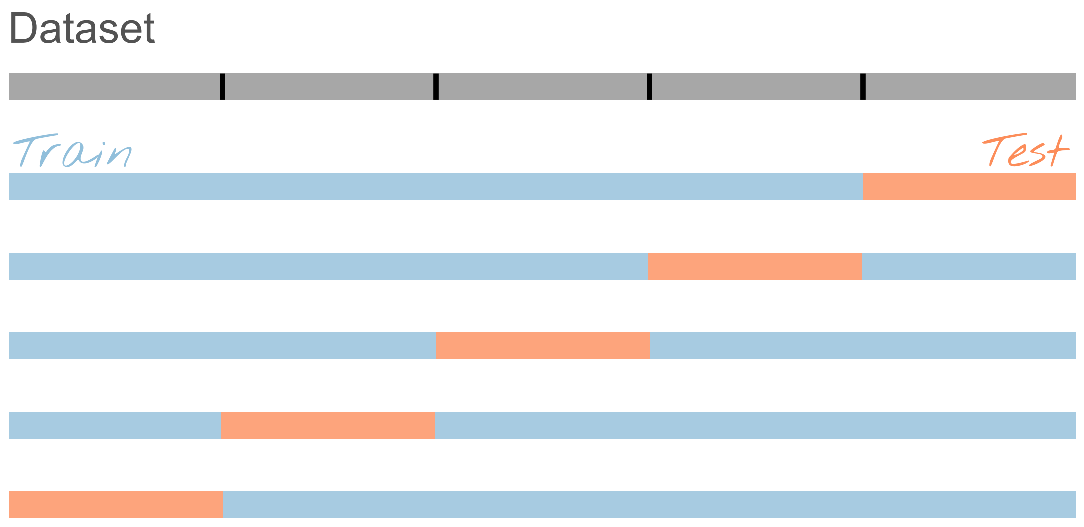
→ 5 models fitted

→ 5 performance estimates



What about cross validation?

Overcome the effect of randomness in your splits



5 folds CV

→ 5 iterations

→ 5 models fitted

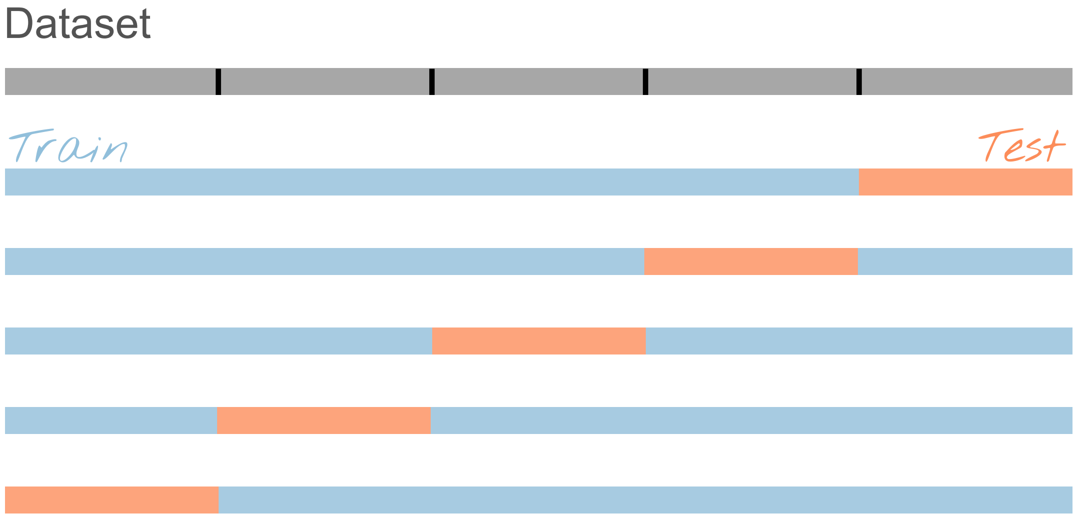
→ 5 performance estimates

Validation and model tuning?

National Oceanography Centre

What about cross validation?

Overcome the effect of randomness in your splits



5 folds CV

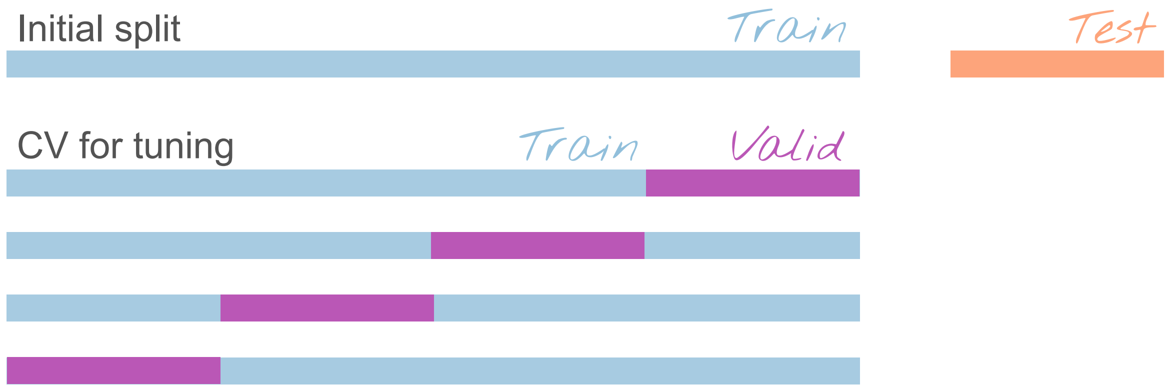
→ 5 iterations

→ 5 models fitted

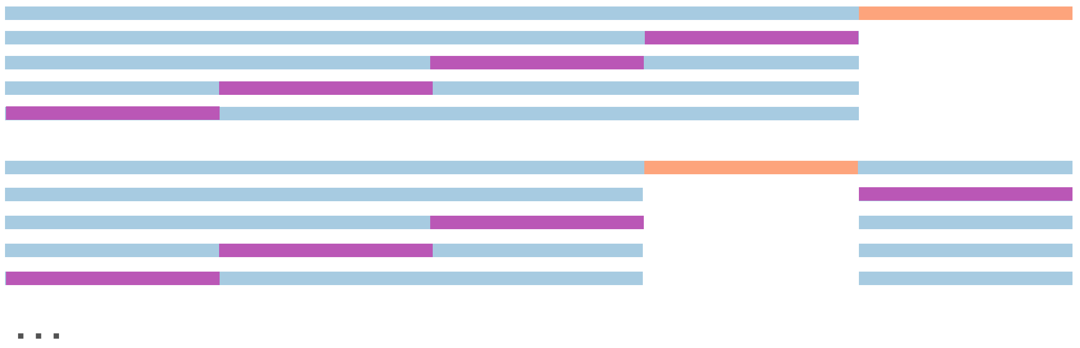
→ 5 performance estimates

Validation and model tuning?

Initial split + CV



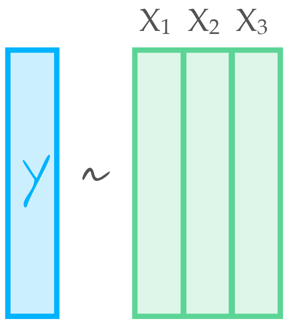
Nested CV



National Oceanography Centre

ML PRO TIP #6
"ML MODELS CAN BE INTERPRETED"

ML models are not black boxes



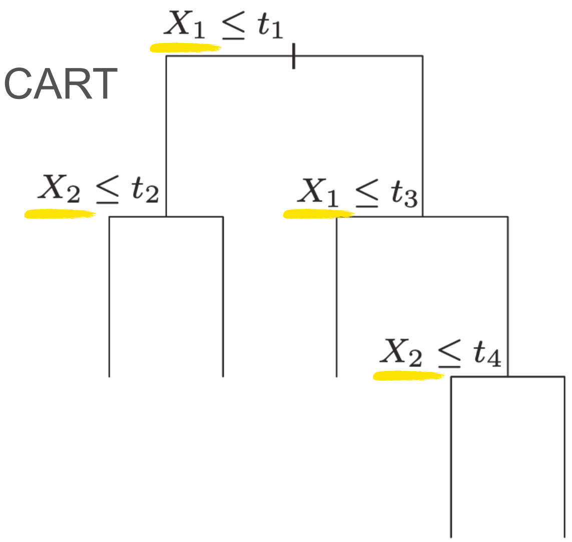
ML models are not black boxes

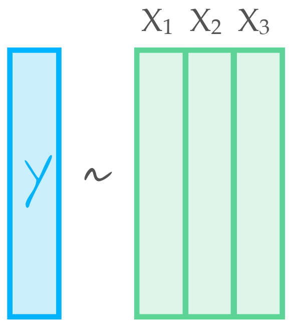
- Some models are easy to interpret

linear regression

$$Y = b + 3.7 \times X_1 + 0.01 \times X_2 + 1.6 \times X_3$$

CART





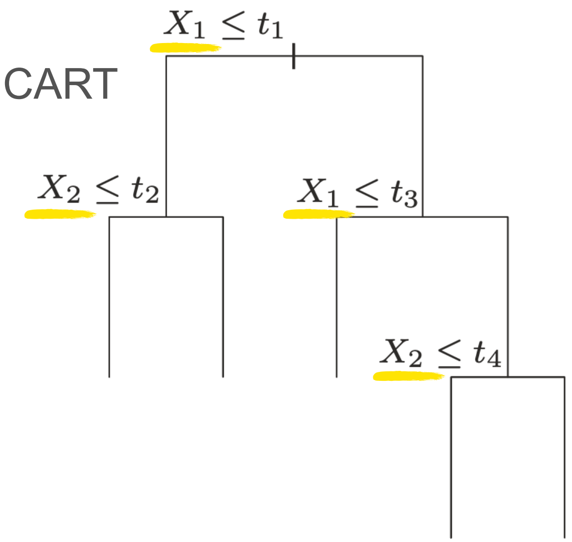
ML models are not black boxes

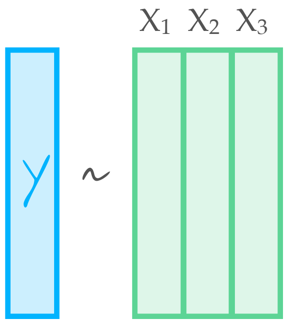
- Some models are easy to interpret

linear regression

$$Y = b + 3.7 \times X_1 + 0.01 \times X_2 + 1.6 \times X_3$$

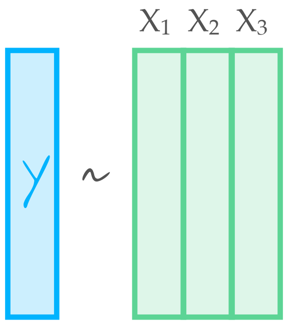
CART





- For other ones, there are workarounds

Orig. model



Orig. performance

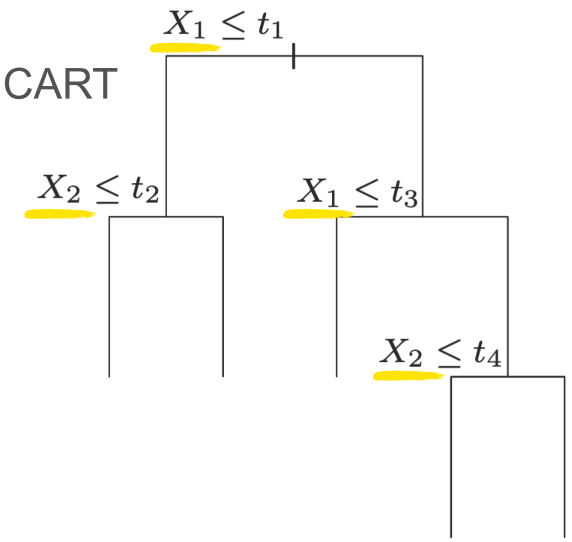
ML models are not black boxes

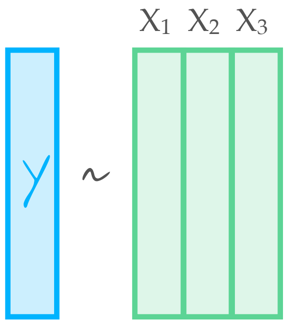
- Some models are easy to interpret

linear regression

$$Y = b + 3.7 \times X_1 + 0.01 \times X_2 + 1.6 \times X_3$$

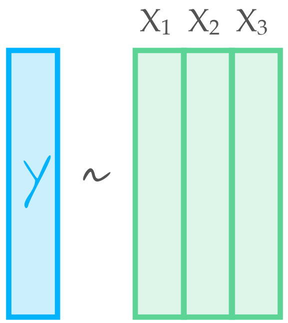
CART





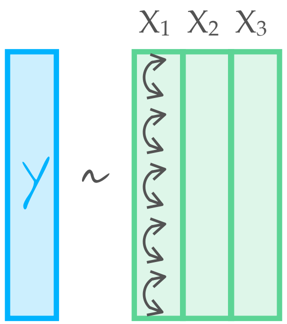
- For other ones, there are workarounds

Orig. model



Orig. performance

Shuffle X1



~ same performance

→ X1 is not important

ML models are not black boxes

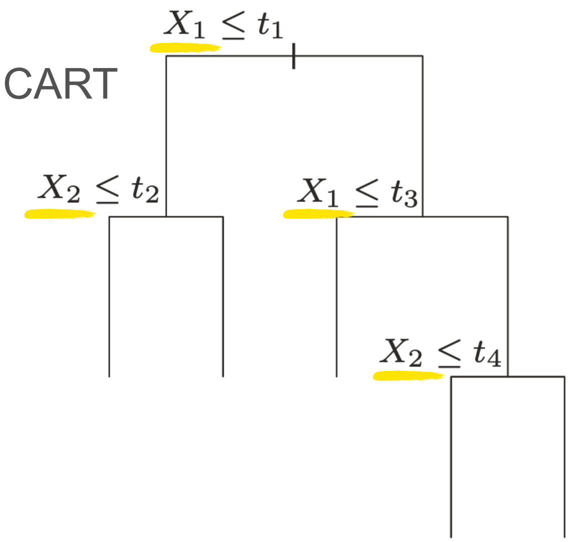
National Oceanography Centre

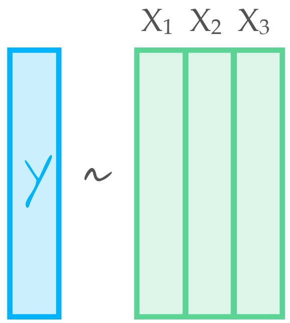
- Some models are easy to interpret

linear regression

$$Y = b + 3.7 \times X_1 + 0.01 \times X_2 + 1.6 \times X_3$$

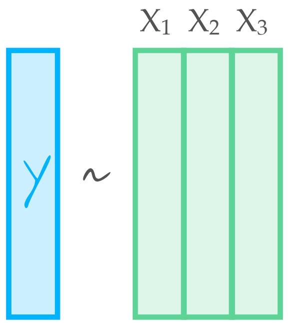
CART





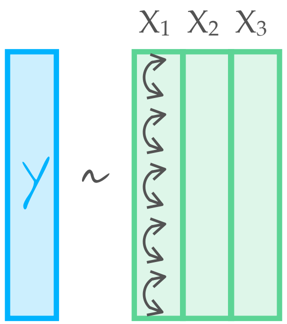
- For other ones, there are workarounds

Orig. model



Orig. performance

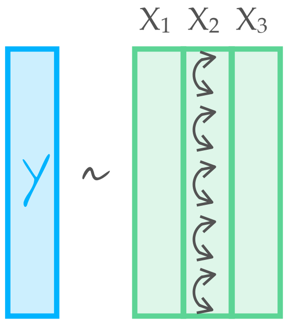
Shuffle X_1



~ same performance

→ X1 is not important

Shuffle X_2



Lower performance

→ X2 is important

…

Importance of each predictor

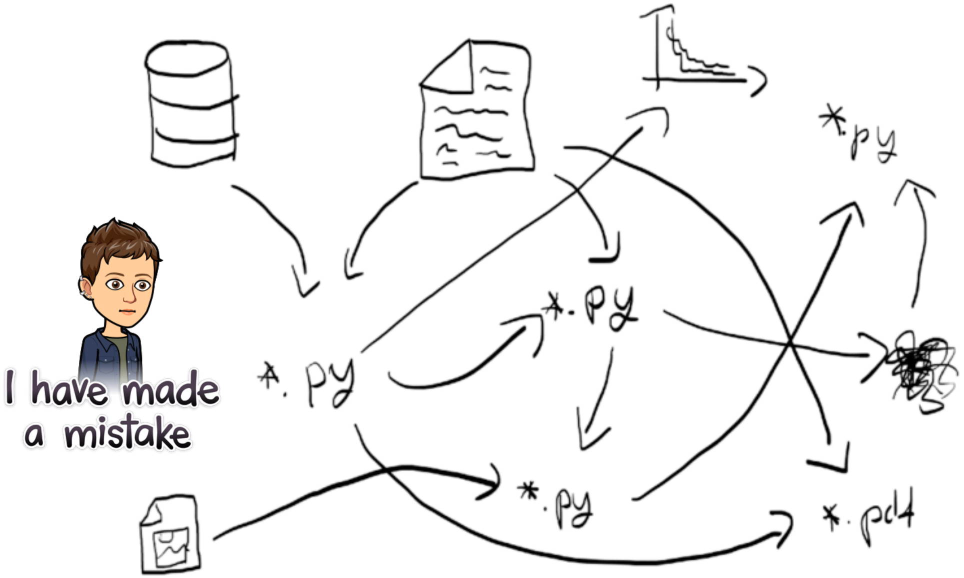
Breiman 2001

National Oceanography Centre

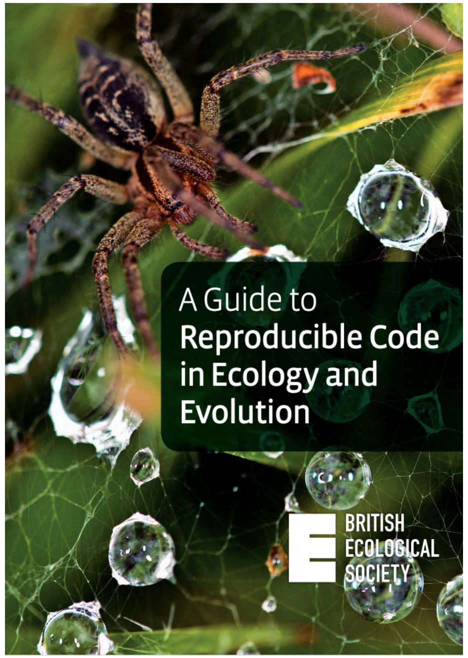
ML PRO TIP #7
"SHARE WITH OTHERS, LEARN FROM OTHERS"

The importance of version control and sharing

- Version control
  - no more mess
  - go back in time



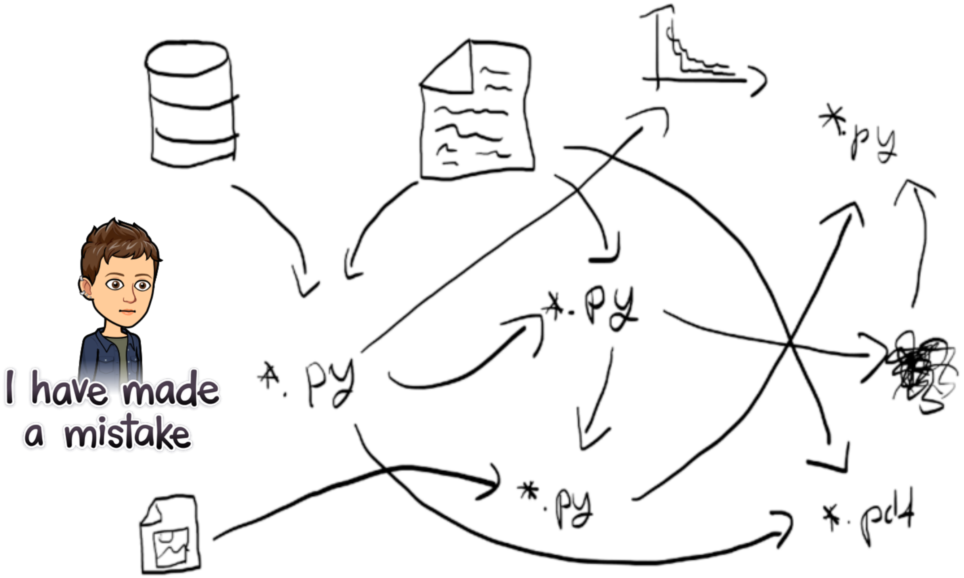
Broad Research Communication Lab



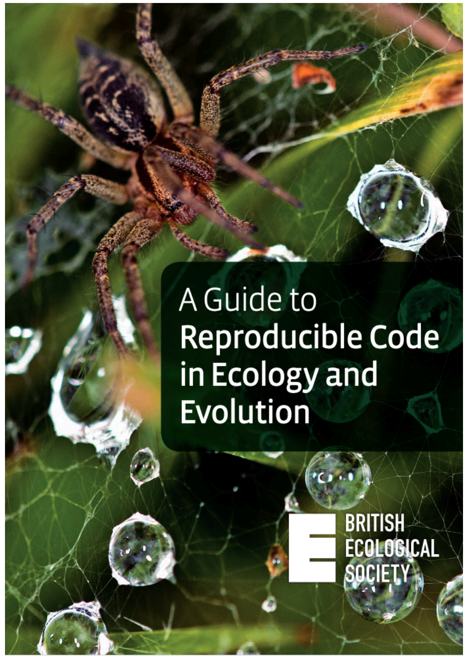
BES & Cooper, N. A Guide to Reproducible Code in Ecology and Evolution. (2017).

The importance of version control and sharing

- Version control
  - no more mess
  - go back in time



Broad Research Communication Lab



BES & Cooper, N. A Guide to Reproducible Code in Ecology and Evolution. (2017).

- Sharing
  - improve yourself
  - reproducibility

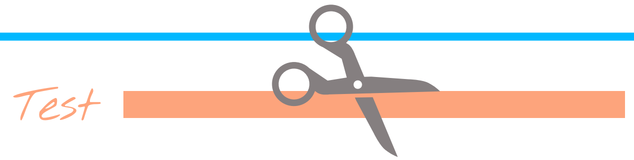
Conclusions

National Oceanography Centre

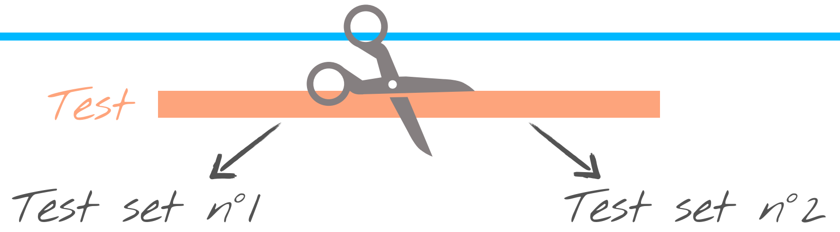
- Supervised ML relates response variable(s) to predictors
- ML is neither black magic, nor a black box
- **ML models can be interpreted**
- A few checks are essential to do it ~~right~~ not wrong
- Multiple choices are possible, depends on your data

Thank you

On the importance of the validation set

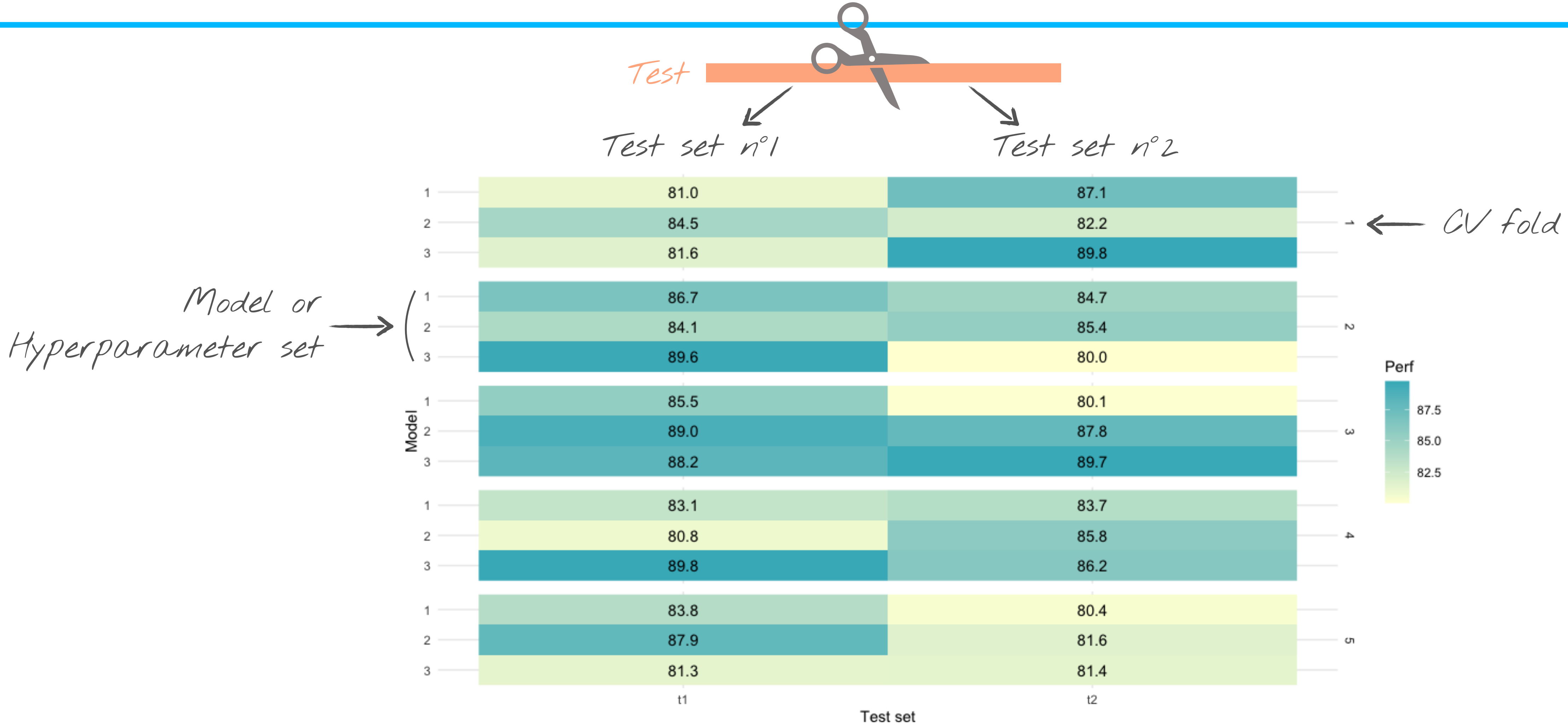


On the importance of the validation set

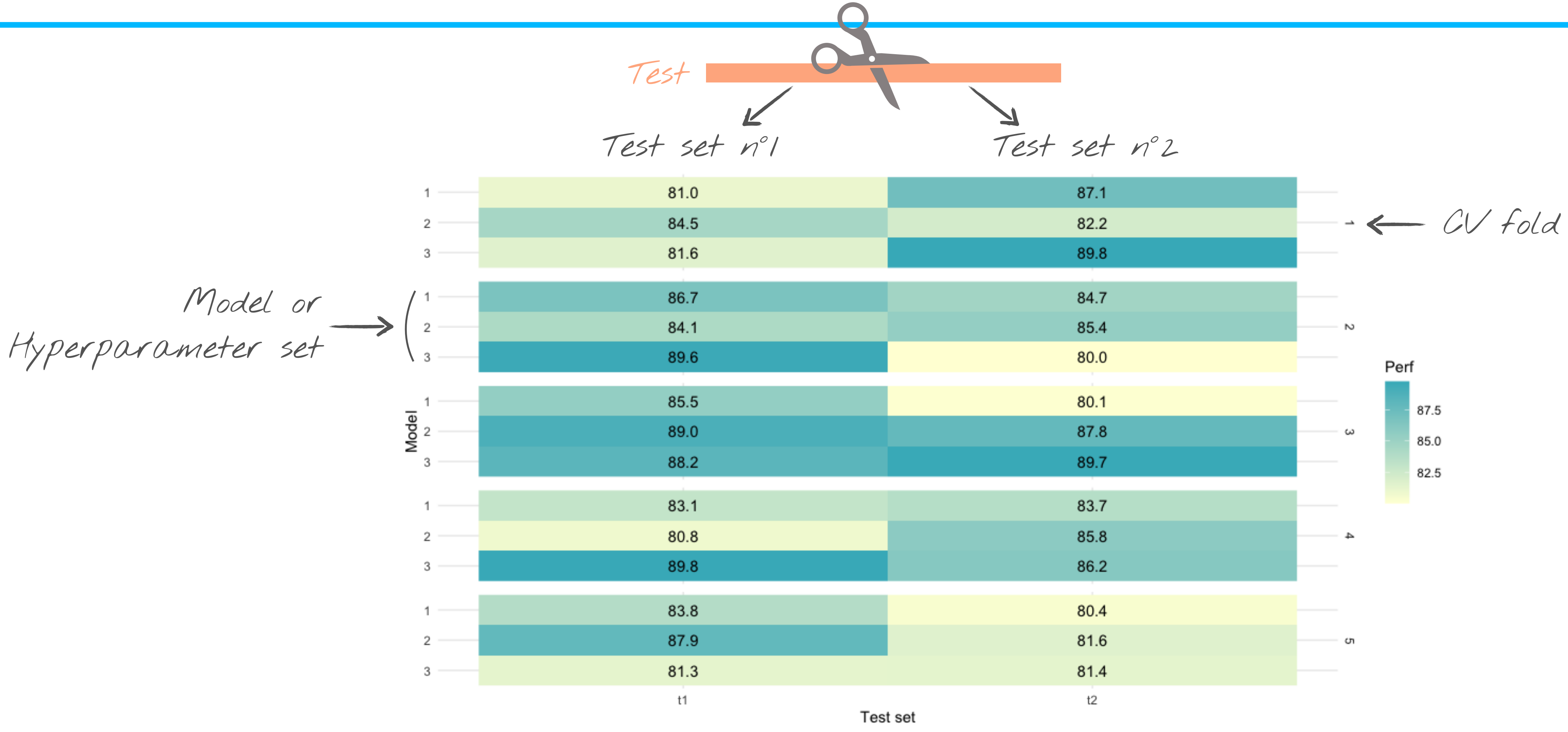


On the importance of the validation set

National Oceanography Centre

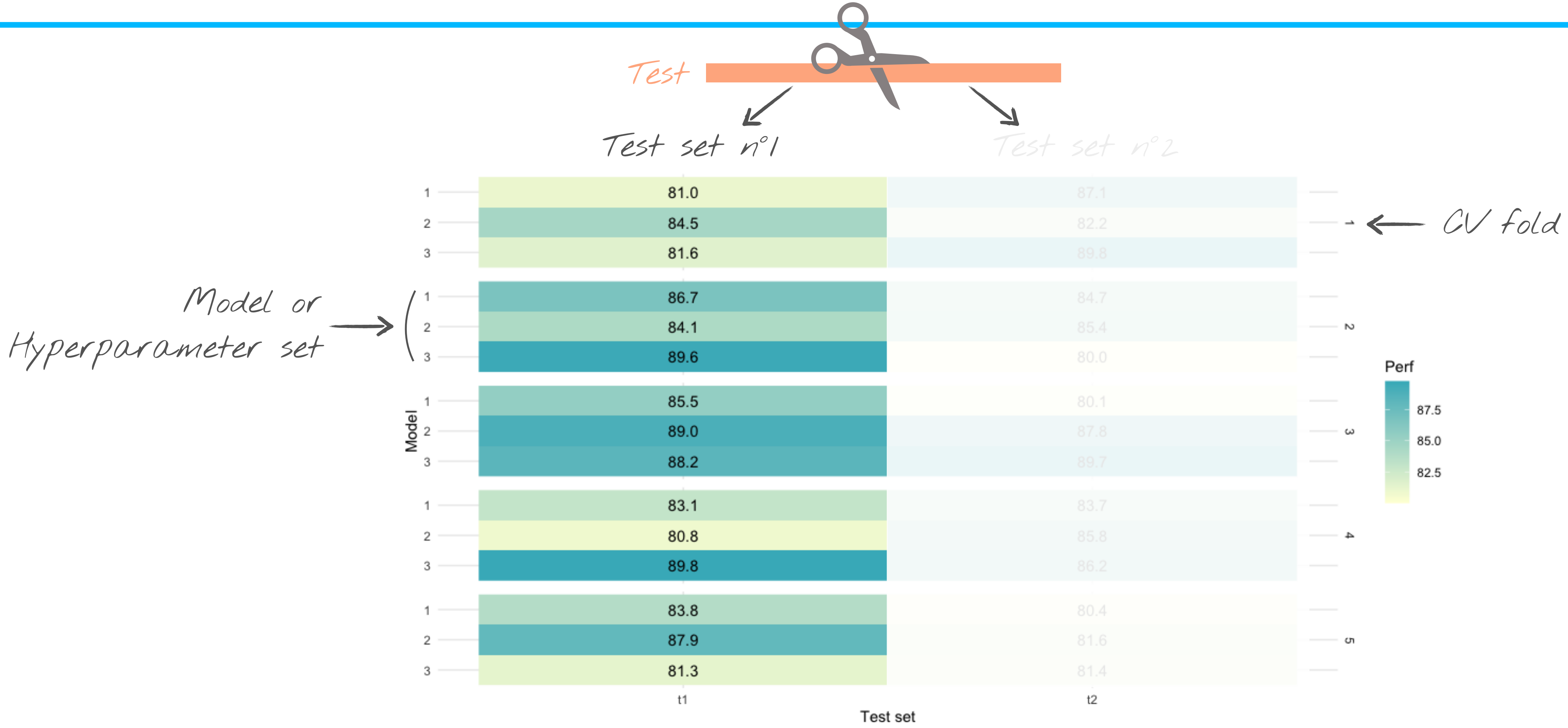


On the importance of the validation set



Overall avg 85.1% 84.4%

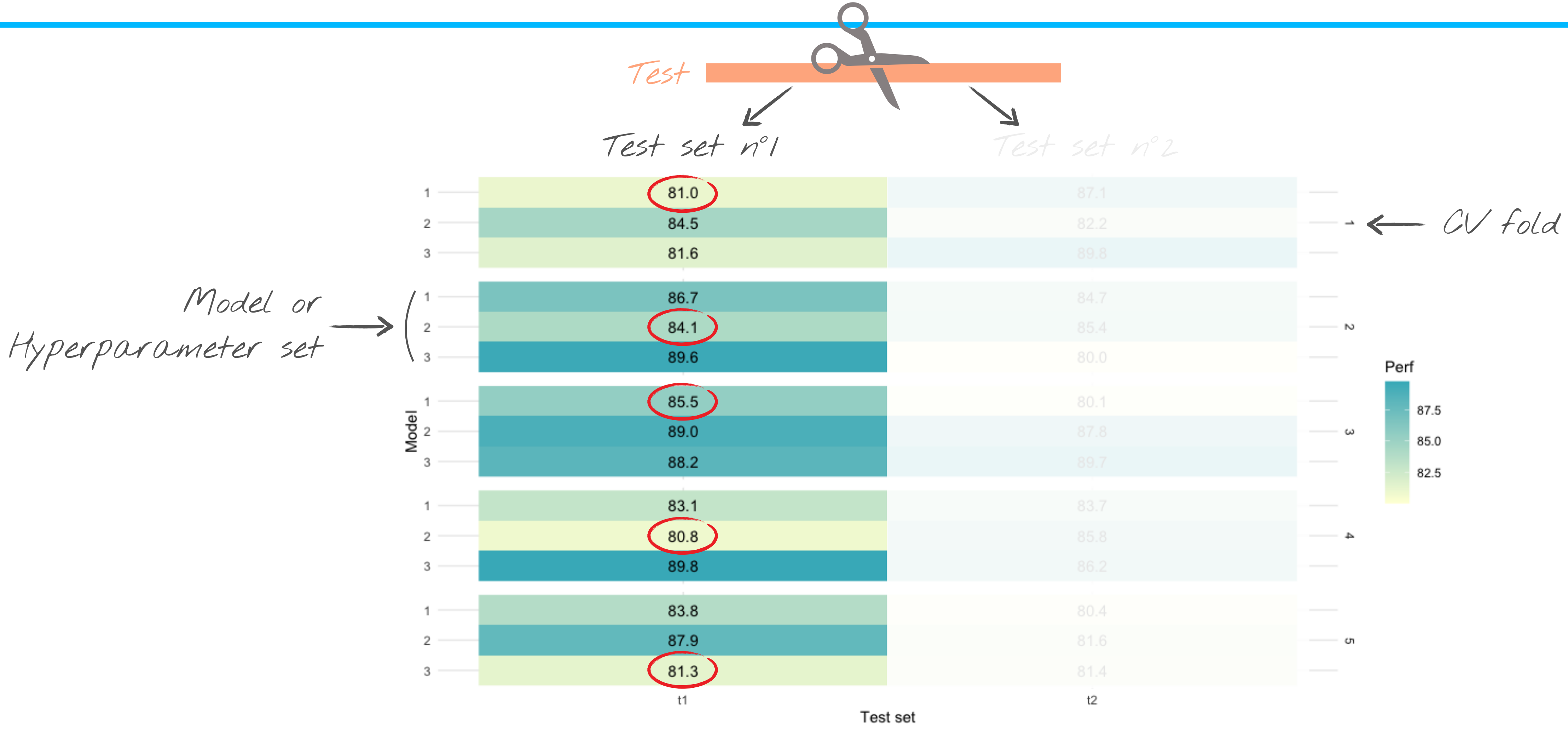
On the importance of the validation set



Overall avg 85.1% 84.4%

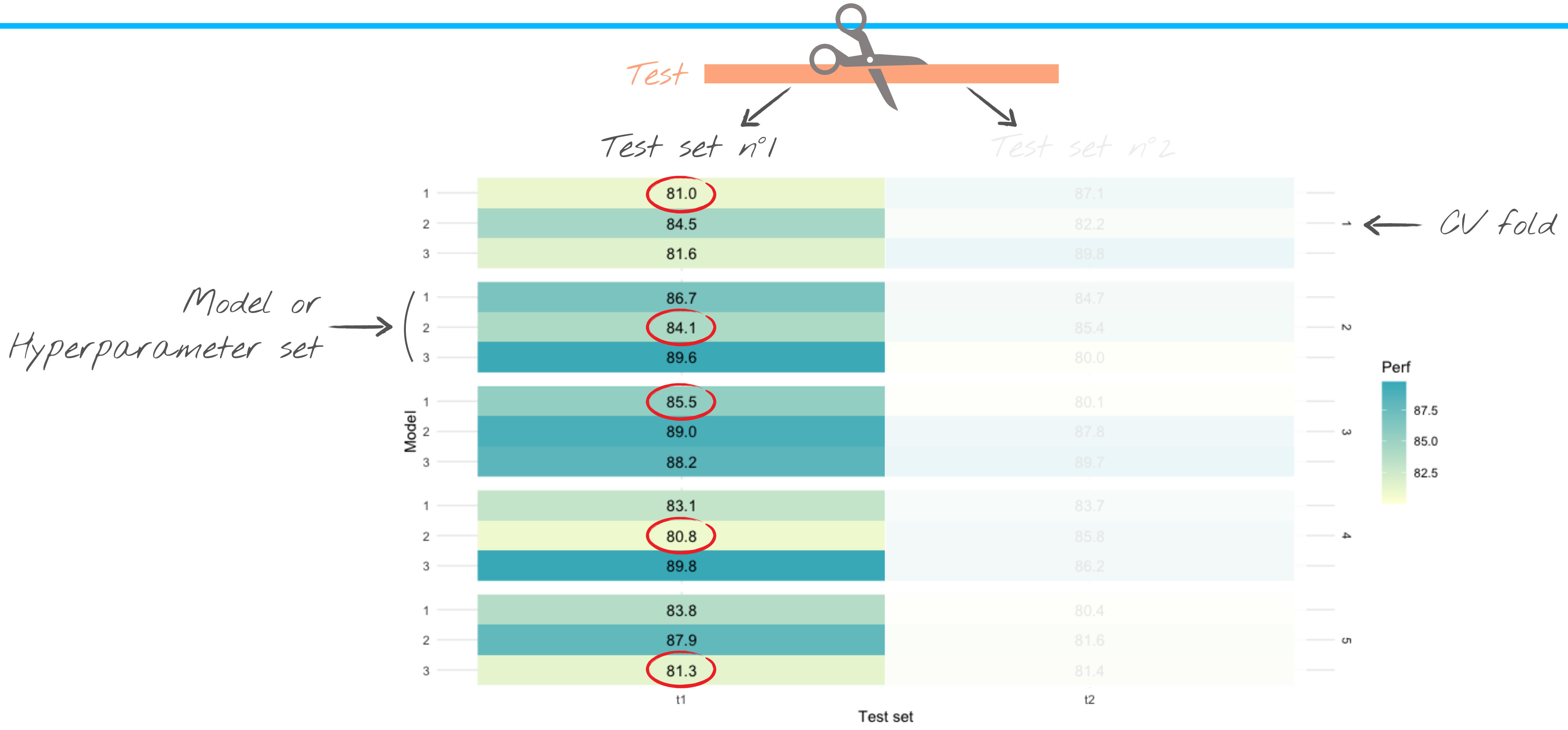
On the importance of the validation set

National Oceanography Centre



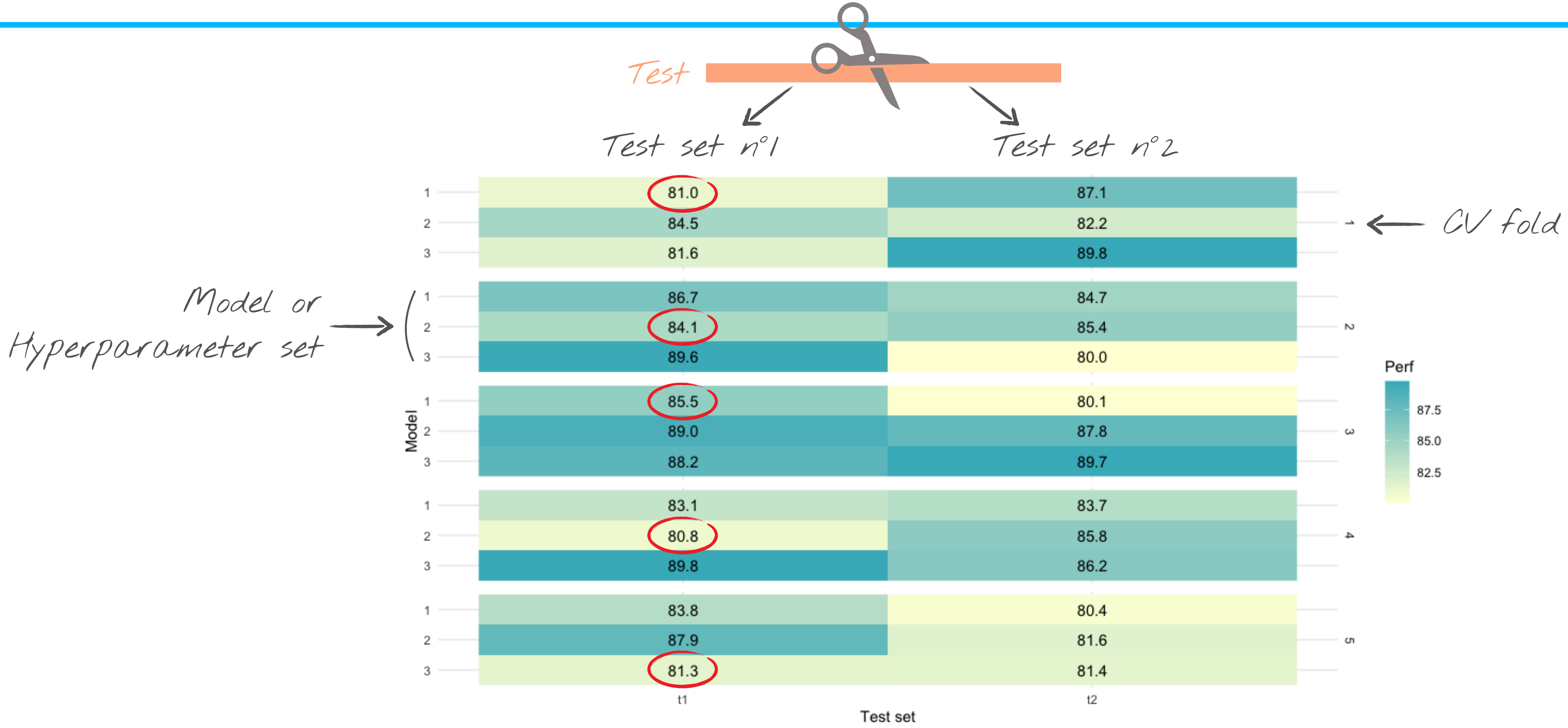
Overall avg 85.1% 84.4%

On the importance of the validation set



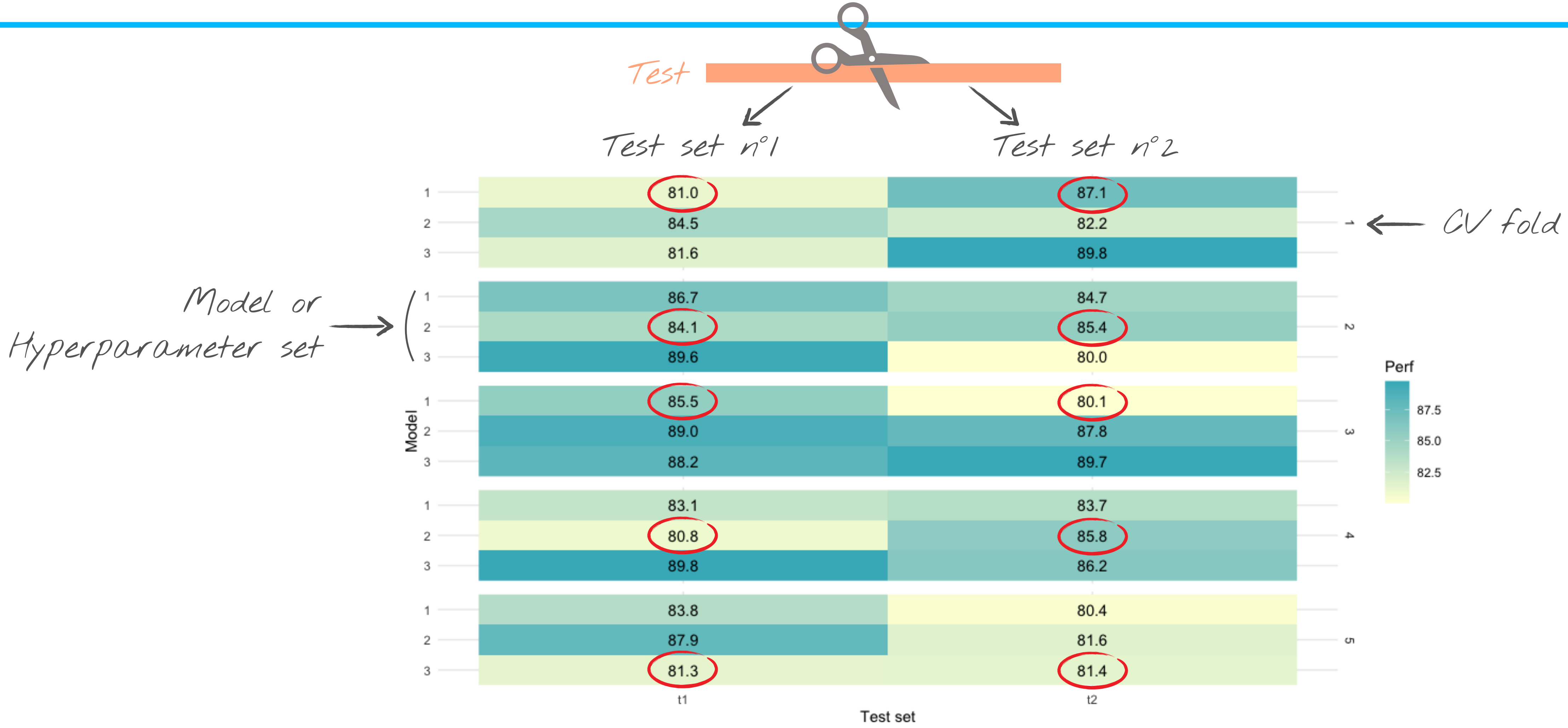
Overall avg	85.1%	84.4%
Best model	82.5%	

On the importance of the validation set



Overall avg	85.1%	84.4%
Best model	82.5%	

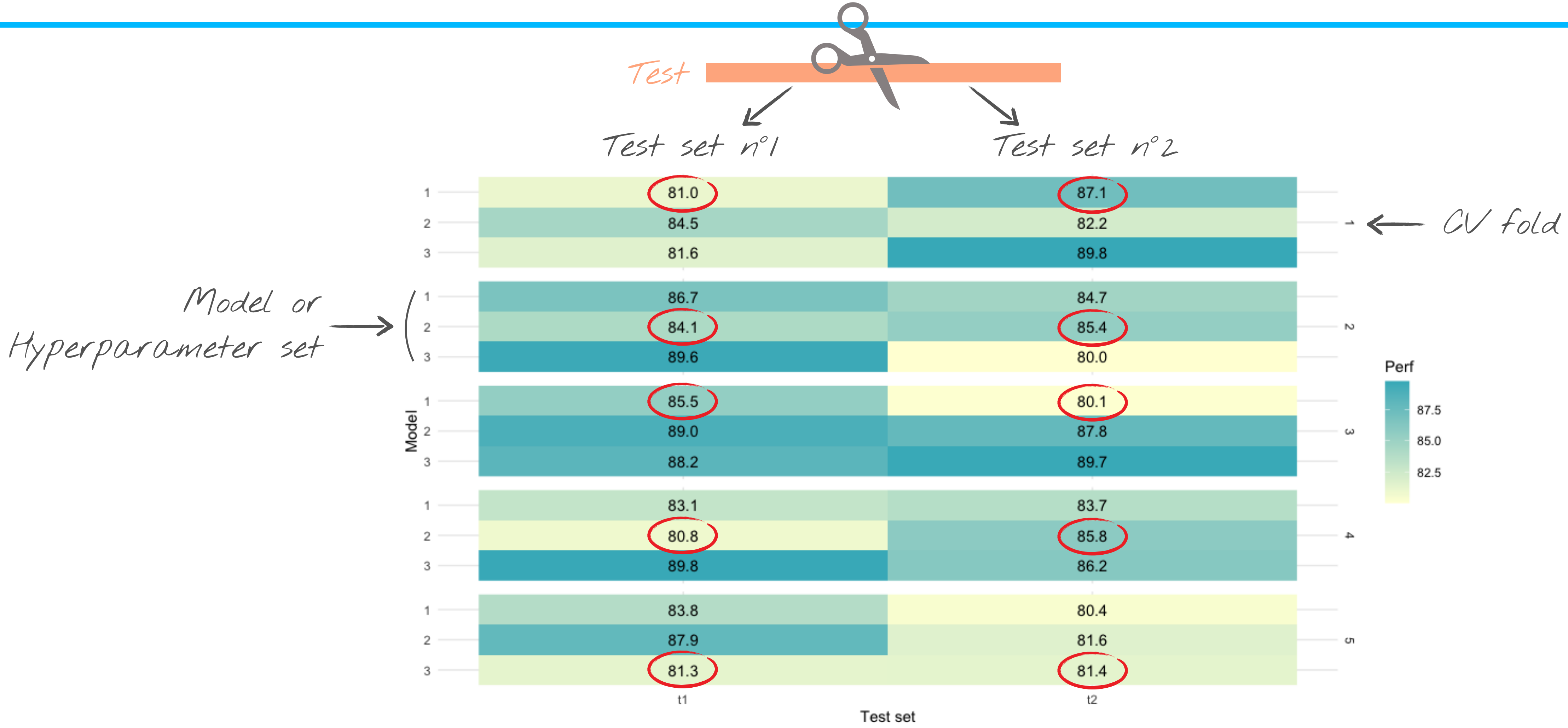
On the importance of the validation set



Overall avg	85.1%	84.4%
Best model	82.5%	

On the importance of the validation set

National Oceanography Centre



	Test set n°1	Test set n°2
Overall avg	85.1%	84.4%
Best model	82.5%	84.0%

On the importance of the validation set

National Oceanography Centre

Validation

optimising hyperparameters

Test

Test set n°1

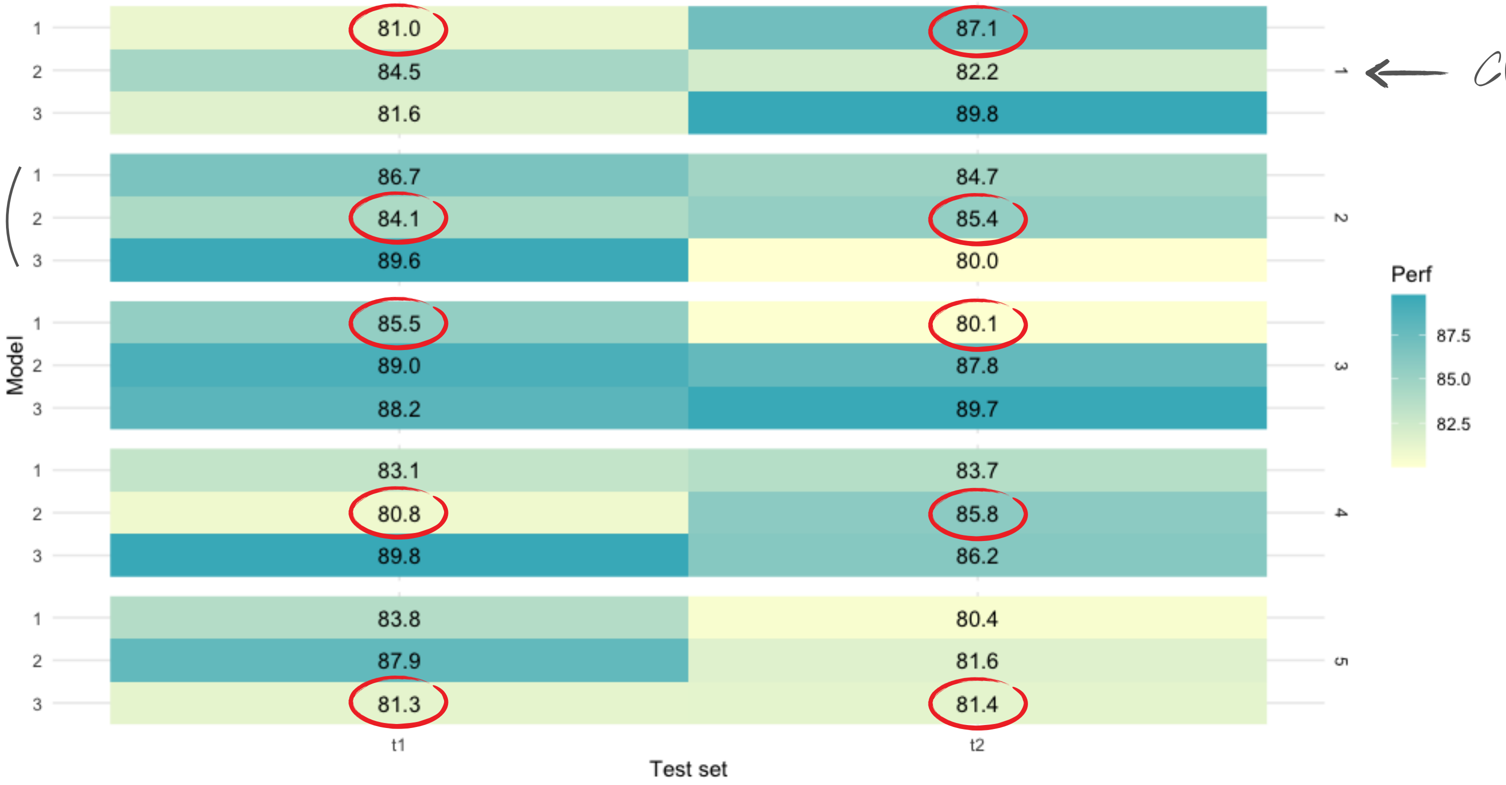
Test set n°2

Test

estimating generalisation error

CV fold

Model or Hyperparameter set



Overall avg	85.1%	84.4%
Best model	82.5%	84.0%