MULTILINEAR REGRESSION

WHY MULTIVARIATE MODELS?

CONCEPTS

INPUT
Also *predictors*, *independent variables*, *features*

OUTPUT
Also *response, dependent variable*

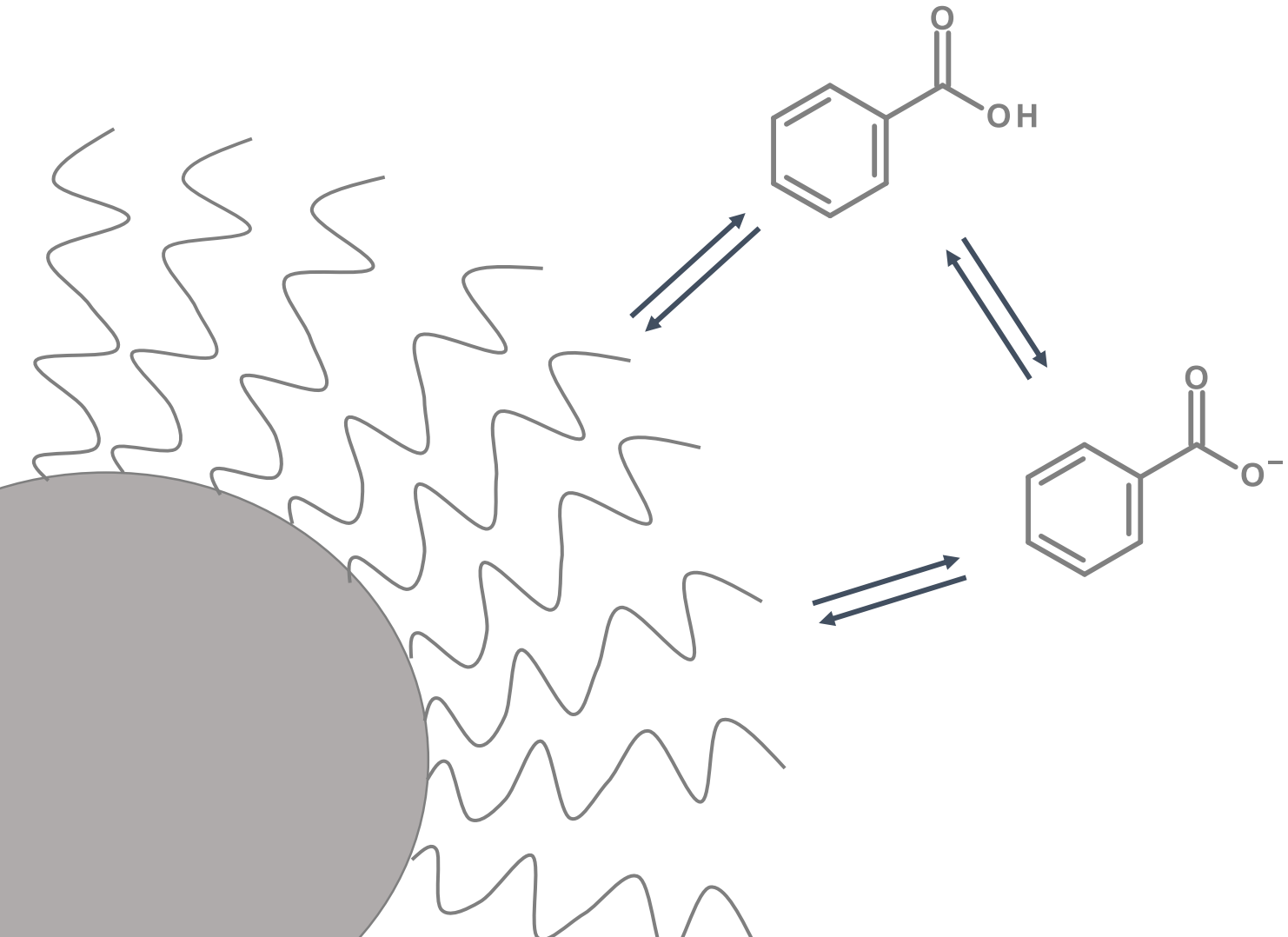
FUNCTIONAL FORM
The conceptual relation between *input* and *output*

EXAMPLES OF MLR

IN CHEMISTRY

OUTPUT	INPUT
Retention time	Analyte: log$P$, pK_a, polarizability, … System: solvent, pH, gradient speed, column, …
Kaolinite content in paper	Full IR spectrum
Response of the compound in MS	Analyte: log$P$, pK_a, polarizability, … System: solvent, pH, buffer type, …
pH of water/organic mixture	Water phase pH, organic solvent, solvent ratio, buffer type, …
Water content in building materials	NIR spectrum
Solubility	Analyte properties Solvent properties
Collision cross-section (ion mobility)	#C, #H, volume, area, …

RETENTION BEHAVIOUR



IDEA

Suppose that we observe a quantitative **response Y** and ***p* different predictors, X_1, X_2, ..., X_p**.

We assume that there is **some relationship** between Y and $X = (X_1, X_2, \ldots, X_p)$, which can be written in the very general form:

$Y = f(X) + \varepsilon$.

f represents the *systematic* information that X provide about Y.

AIM

We can either aim at

- Making predictions about the future state

or

- Understanding the processes better

UNDERSTANDING

Which predictors are associated with the response?
What is the relationship between the response and each predictor?

Simpler models are simpler to interpret

EXAMPLES

- Optimizing reaction yield
- Assessing quality of a product based on several parameters
- Design a drug

PREDICTING

$Y = f(X)$,

We can use complicated models and should first and foremost care about prediction accuracy

EXAMPLES

- Choosing the most potent drug out of candidates
- Predict retention time as conformation point in non-targeted screening
- Estimate the concentration of the compounds detected with LC/ESI/HRMS without standards
- Evaluate the fibre content of a textile

RESEARCH QUESTIONS

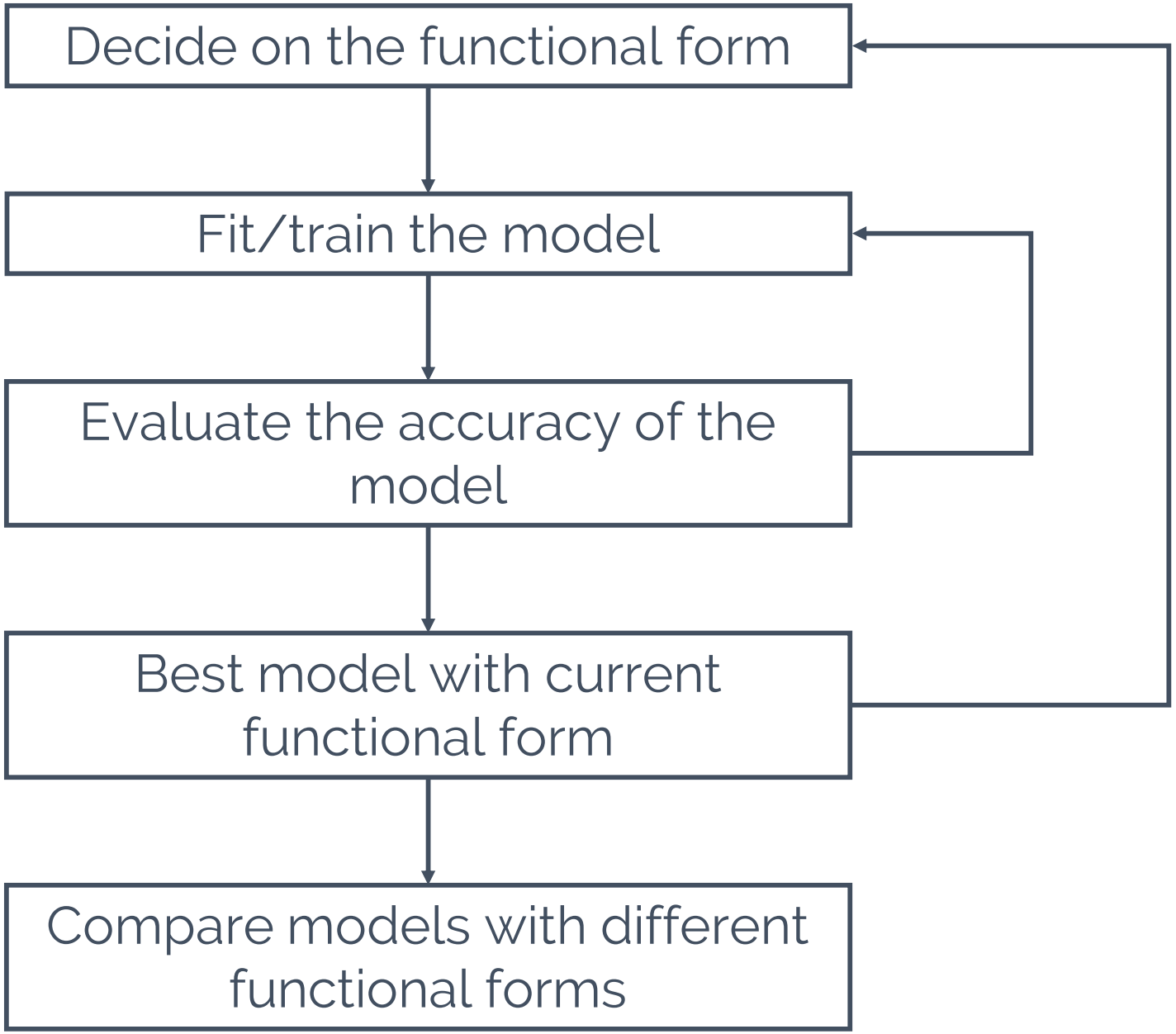
1. Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?

2. Do all the predictors help to explain Y, or is only a subset of the predictors useful?

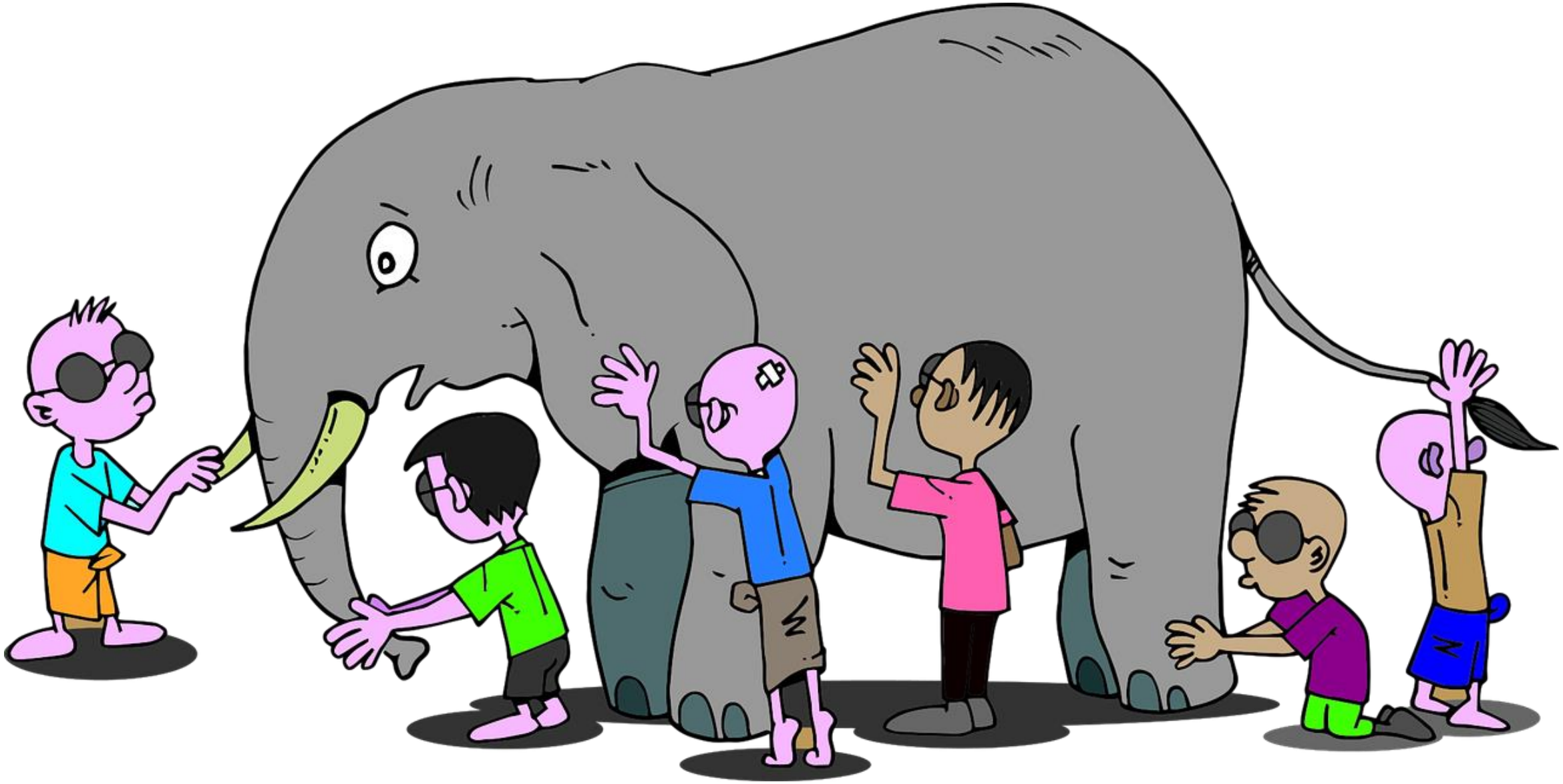
3. How well does the model fit the data?

4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

WORKFLOW


Decide on the functional form
Fit/train the model
Evaluate the accuracy of the model
Best model with current functional form
Compare models with different functional forms

HOW TO ESTIMATE *f*?



Anneli Kruve

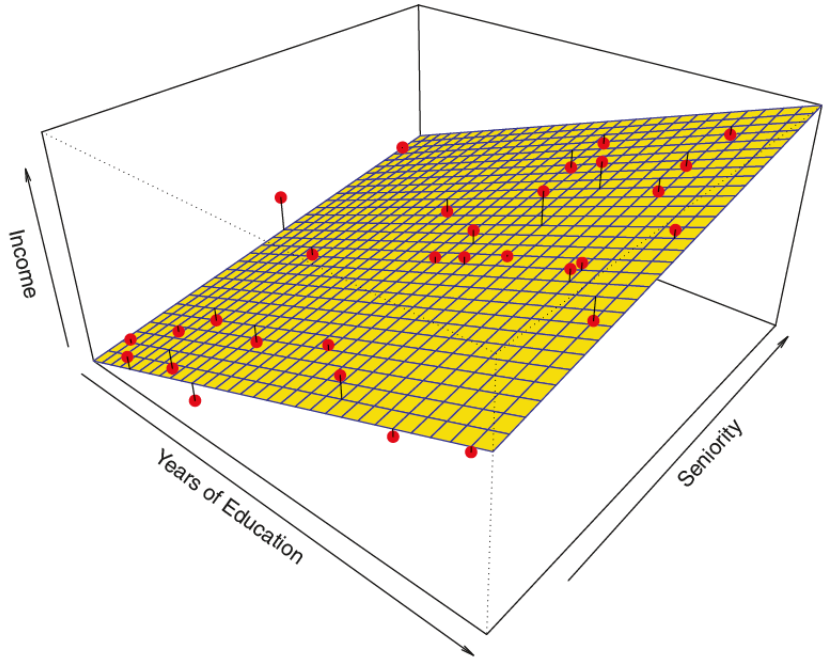
MODEL FLEXIBILITY

Parametric – model based approach
Disadvantage
- will usually not match the true unknown form of *f*

Advantage
- Easy to interpret

Flexible models
Advantage
- can fit many different possible functional forms for *f*

Disadvantage
- complex models can lead to *overfitting*

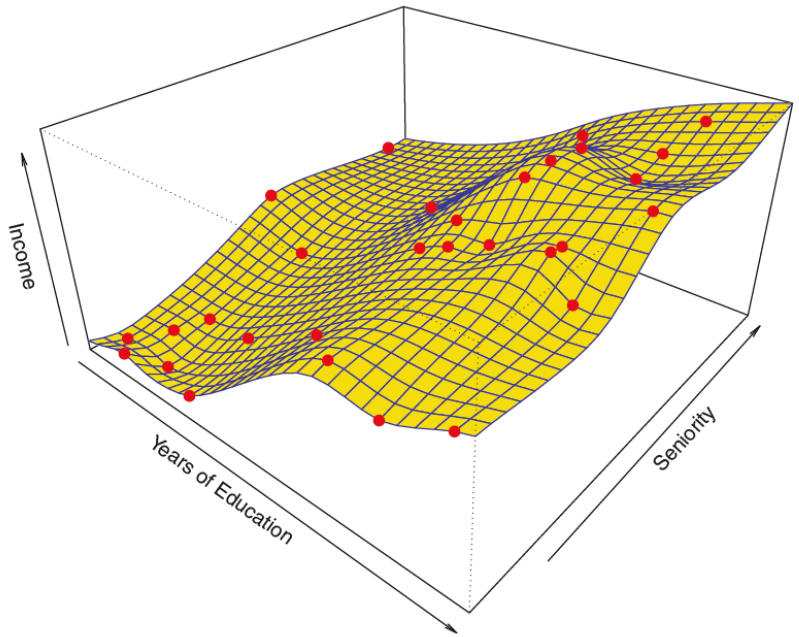


NON-PARAMETRIC METHODS

Do not make explicit assumptions about the functional form of *f*

More flexible if:
- More input parameters
- More complicated model type

Disadvantage:
- number of observations required
- are hard/impossible to interpret

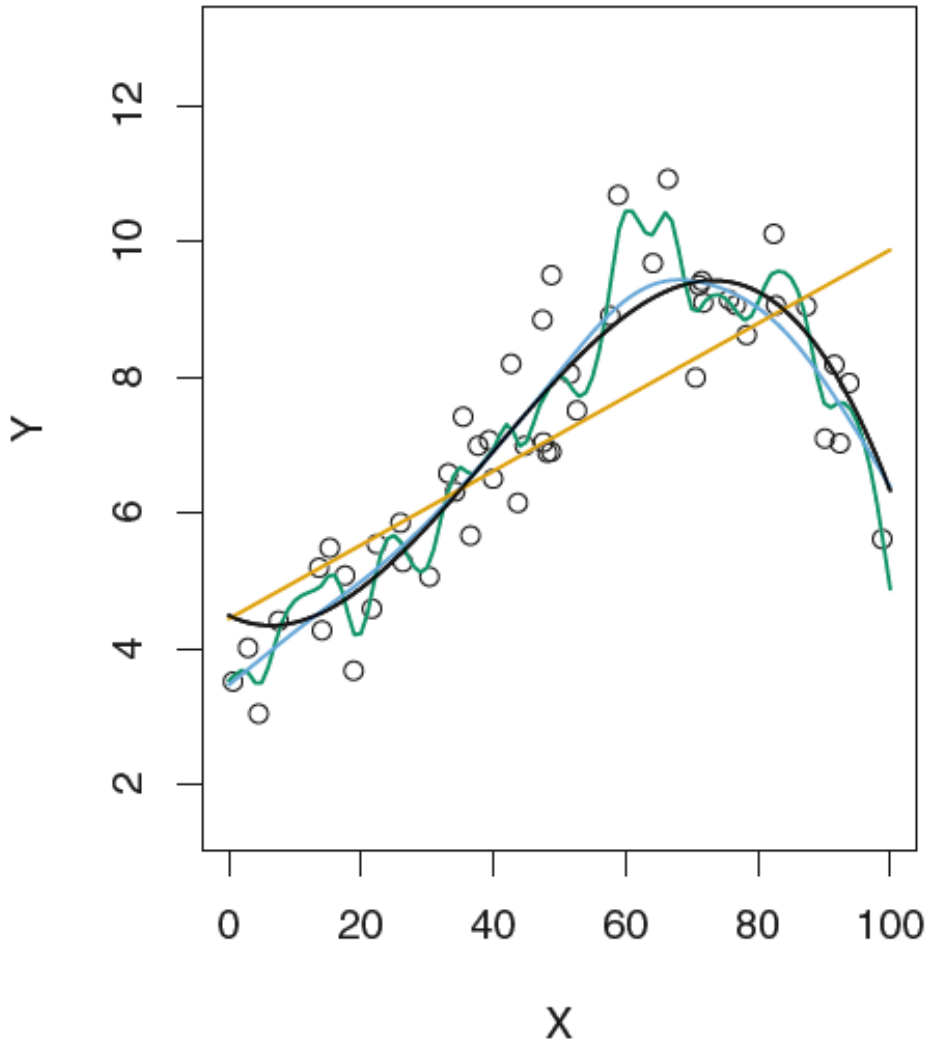


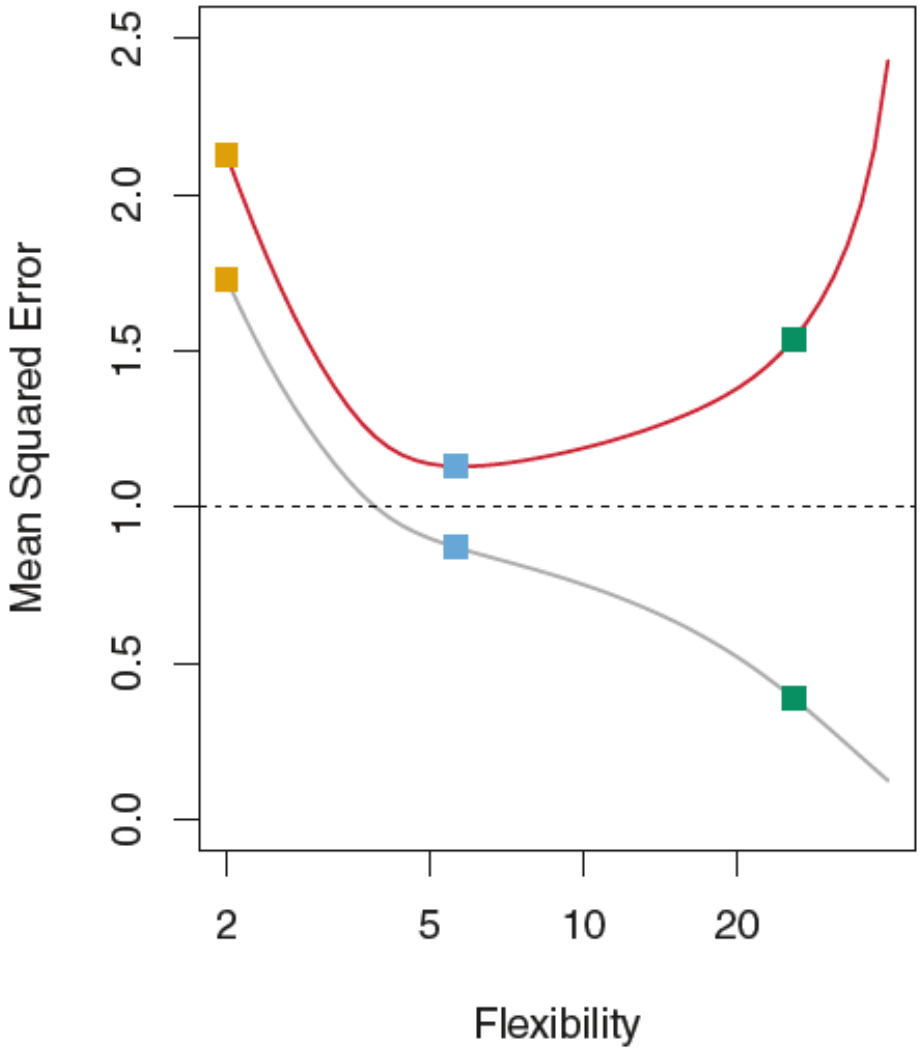
OVERFITTING

A developed model:

- has high prediction accuracy on data used for developing the model
- has poor prediction accuracy for data not “seen” by the model

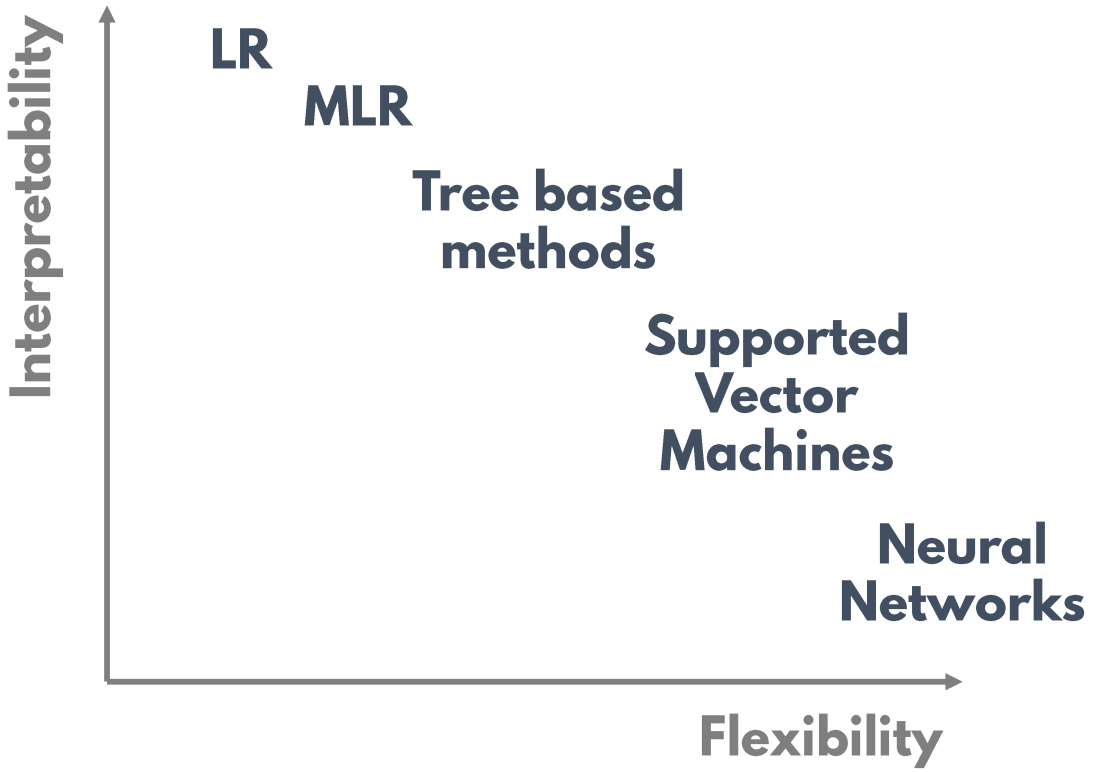
EXAMPLE

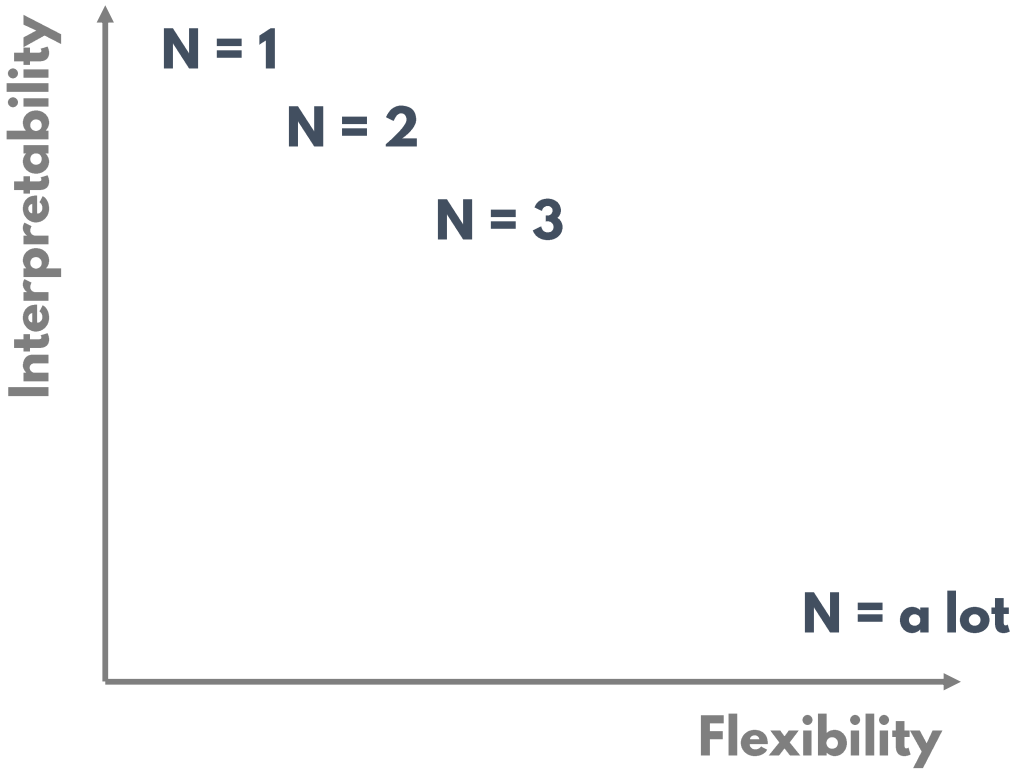

Y
2
4
6
8
10
12
0
20
40
60
80
100
X


Mean Squared Error
0.0
0.5
1.0
1.5
2.0
2.5
2
5
10
20
Flexibility

WHY TO PREFER A MORE RESTRICTIVE METHOD?

INTERFERENCE & OWERFITTING





FITTING MLR MODEL

1. FUNCTION FORM

MAKE THE ASSUMPTION ABOUT FUNCTION FORM OR SHAPE

In case of MLR

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p.$$

E.g. In case of predicting retention time for LC

$t_R = a \cdot \log P + b \cdot pK_a$

ERROR TYPES

reducible error and the irreducible error

ε random error term catches all that we miss with the model

- The true relationship is probably not linear
- Missing input variables associated with output
- Measurement error

The total model prediction error consists of:

$$
\begin{aligned}
E(Y-\hat{Y})^2 &= E[f(X)+\epsilon-\hat{f}(X)]^2 \\
&= \underbrace{[f(X)-\hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}} ,
\end{aligned}
$$

Irreducible error provides an upper bound to the method accuracy. And it is unknown.

“Garbage in garbage out”

$$
\begin{aligned}
E(Y-\hat{Y})^2 &= E[f(X)+\epsilon-\hat{f}(X)]^2 \\
&= \underbrace{[f(X)-\hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}} ,
\end{aligned}
$$

You and your model do not know what is the proportion or size of the reducible and irreducible error.
Try to keep irreducible error as low as possible.
E.g. find the best data.

BUT keep the data collection conditions close to the intended use.

DUMMY VARIABLES

Some variables are categorical

- Solvent (MeOH, MeCN, …)
- Buffer type
- Column

Model can not accommodate with it.

Column type
C18
C18
C8
HILIC
HILIC

C18	C8
1	0
1	0
0	1
0	0
0	0

2. FIT OR TRAIN THE MODEL

Usually we train a number of models in parallel

Along the road we might figure out that additional input parameters are required

We need to estimate the parameters $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$, …, $\boldsymbol{\beta}_p$

Ordinary least squares is commonly used

3. ASSESSING MODEL ACCURACY

What is the suitable model?

- No one model fits all solutions
- Keep as simple as possible

Are the input parameters sufficiently descriptive of the output parameter?

- Do we need to add anything?
- Are all of the input parameters required?

Parameters

- RMSE residual standard error
- R^2
- F-statistic

RMSE

ROOT MEAN SQUARE ERROR

$$RMSE = \sqrt{\frac{1}{n-2} RSS} = \sqrt{\frac{1}{n-2} \sum (y_i - \hat{y}_i)^2}$$

R^2

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

This is not a calibration graph! R^2 values can be much lower.
In certain applications 0.5 or lower may still be useful!

F-statistic

Explained vs unexplained variation

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)},$$

Less complicated vs more complicated model

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS})/q}{\mathrm{RSS}/(n-p-1)}.$$

VARIABLE SELECTION

1. Test all possible combinations
2. Forward selection
3. Backward selection
4. Mixed selection

TESTING all possible combinations

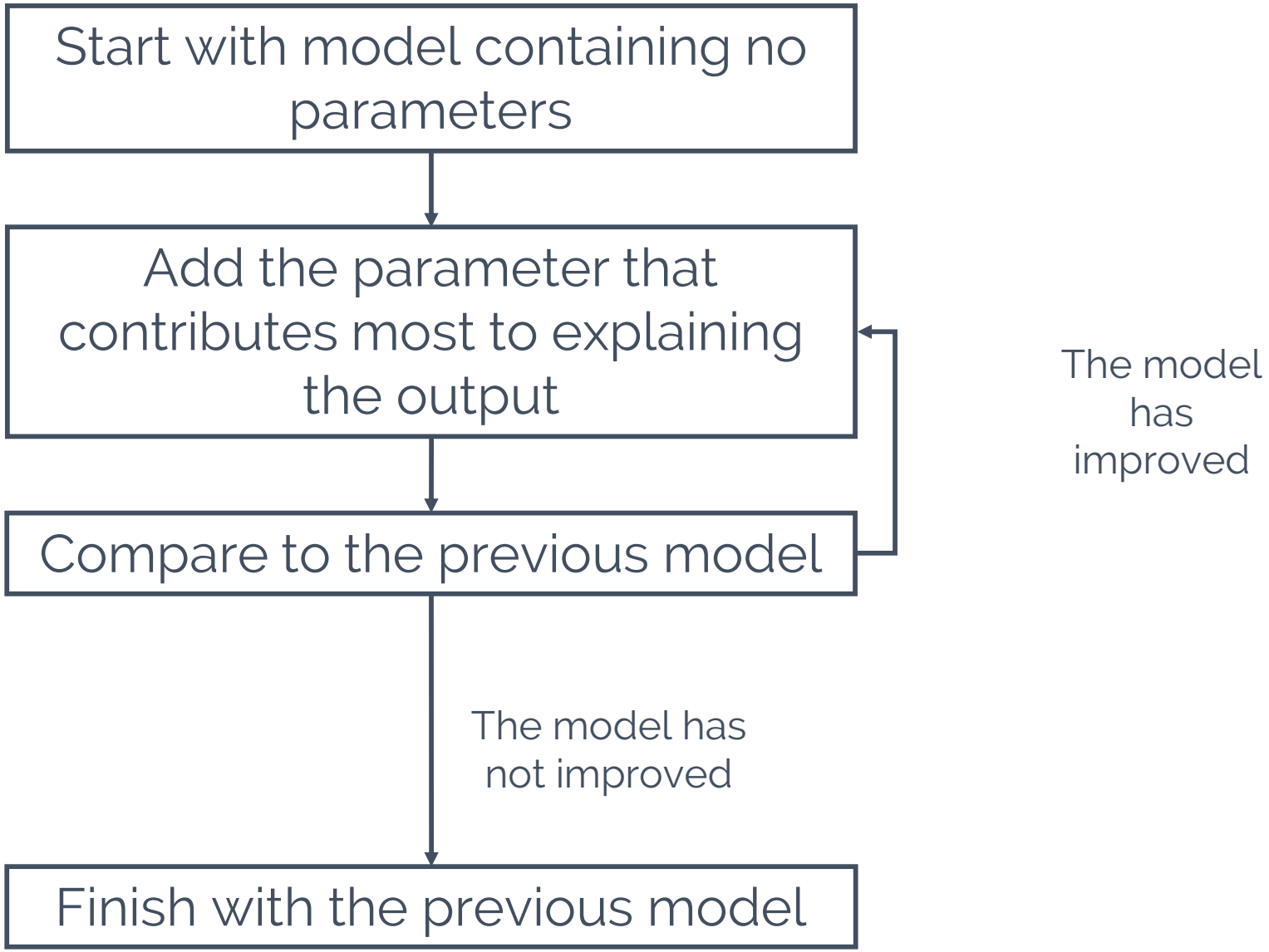
1. All one parameter models, p models
2. All 2 parameter models, p • (p-1) models
3. ….
4. One model with all parameters

Altogether 2^p models

p = 30 this is 1 073 741 824 models

Computationally expensive and time-consuming!

FORWARD SELECTION



FORWARD SELECTION

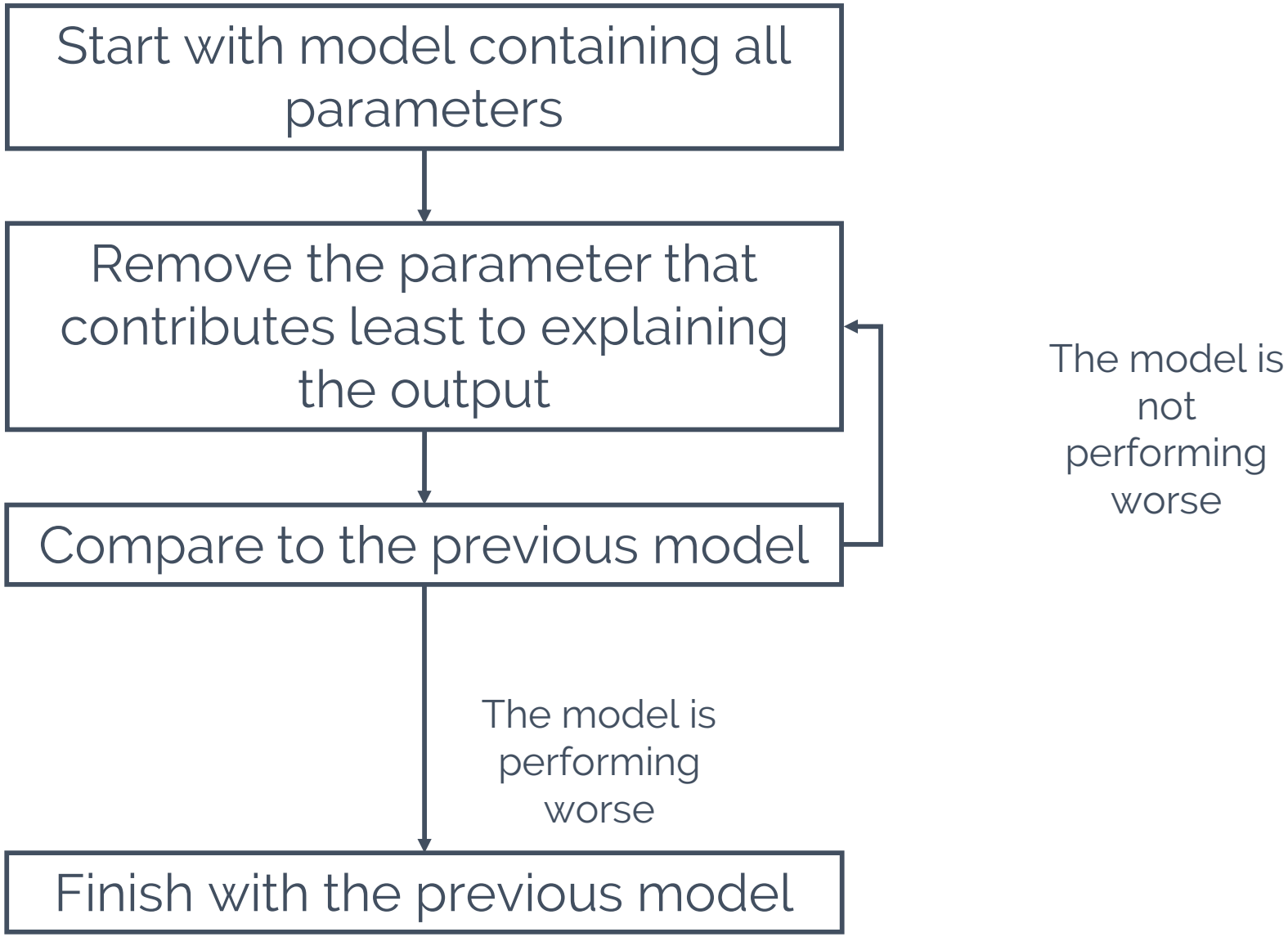
Advantages:

- Can be used if $p > n$
- Fast

Disadvantage:

- The model may be stuck with the parameters inserted early on

BACKWART SELECTION



BACKWART SELECTION

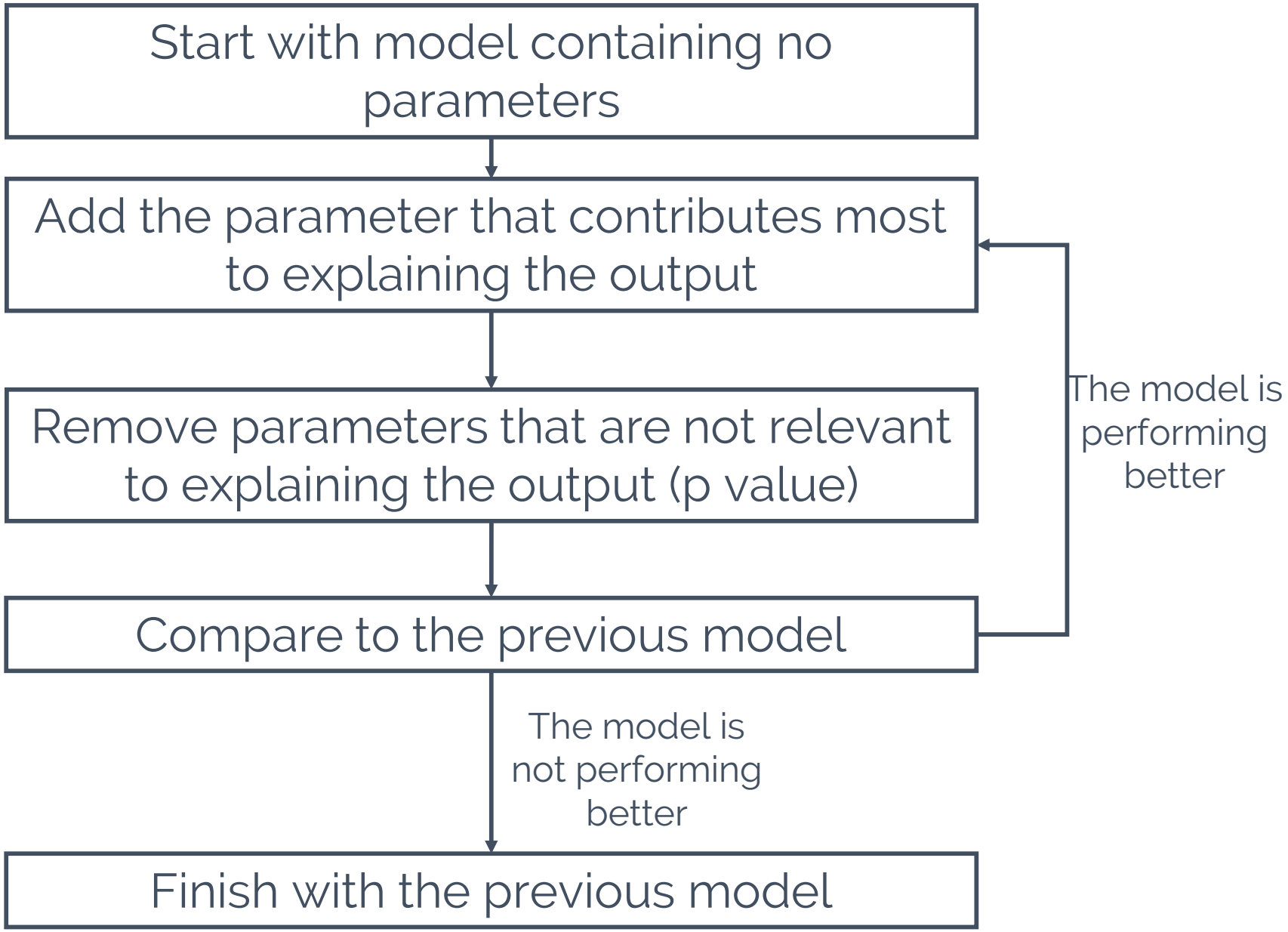
Advantages:

- Fast

Disadvantage:

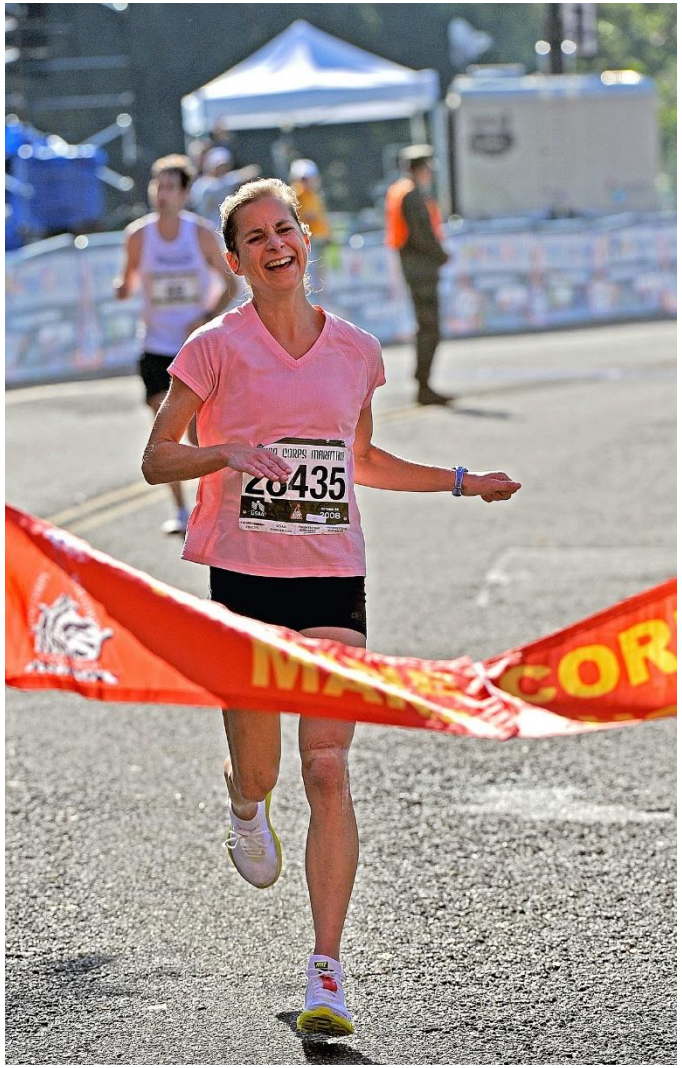
- Can not be used if $p > n$

MIXED SELECTION



Anneli Kruve

FINISH, RIGHT?



Anneli Kruve

NO, NOT QUITE!

Anneli Kruve

VALIDATION

WILL MY MODEL ALSO WORK ON NEW DATA?

RESAMPLING METHODS

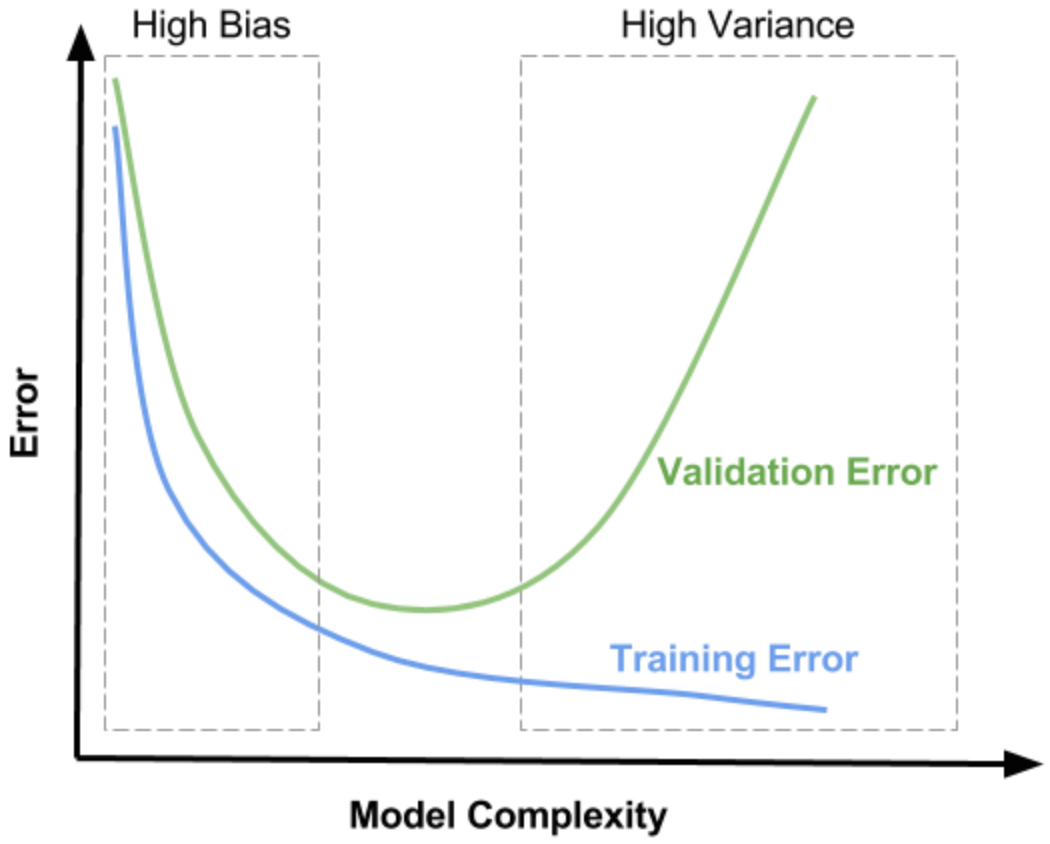
1. Cross-validation
    1. **Leave-One-Out**
    2. **The Validation Set Approach**
    3. **k-fold Cross-Validation**

- Based on the data you have at hand while developing the model

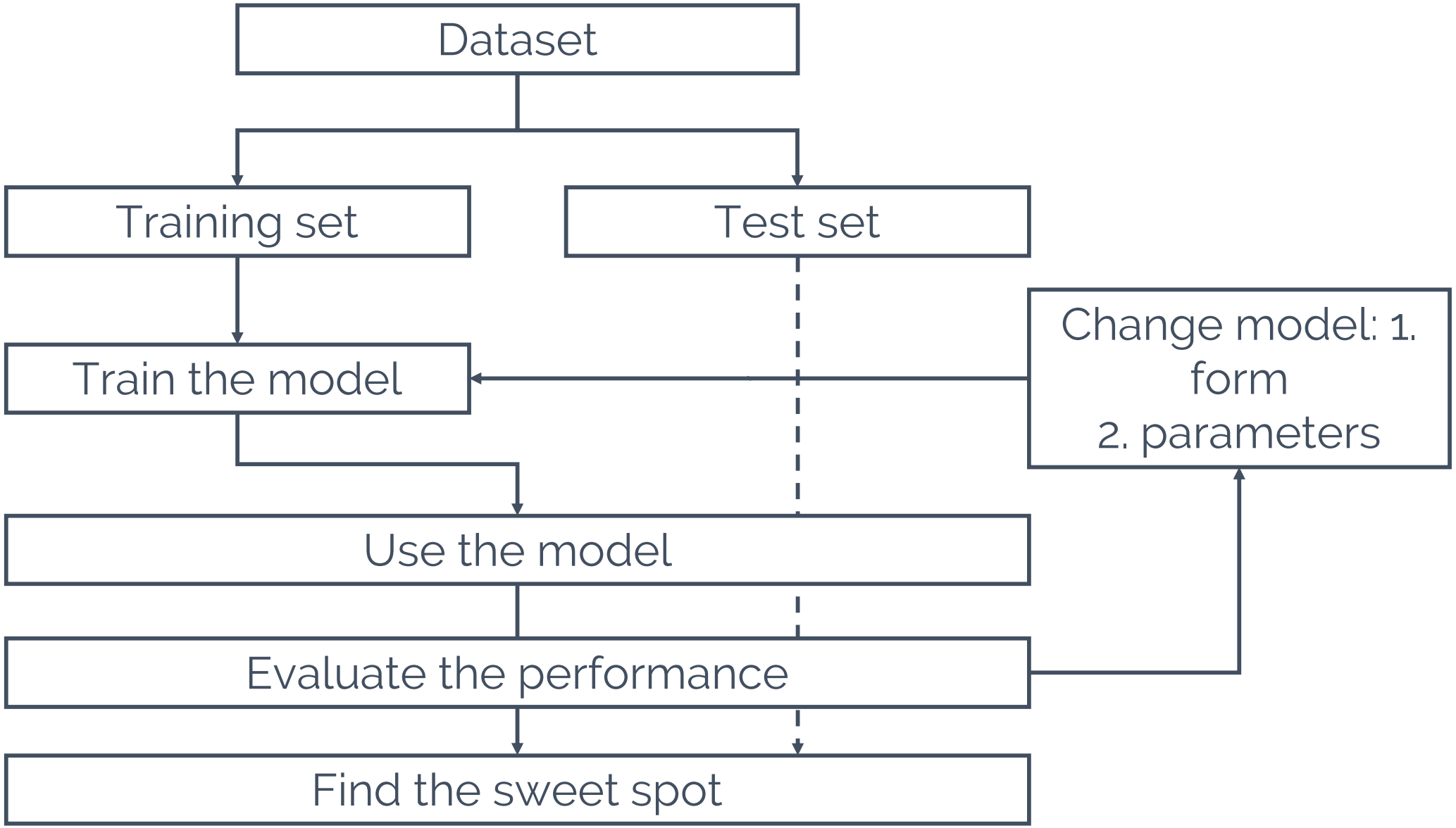
2. External validation

- Based on independent set of data (within application scope)
- Especially important in chemistry

TRAINING DATA ALWAYS LIE

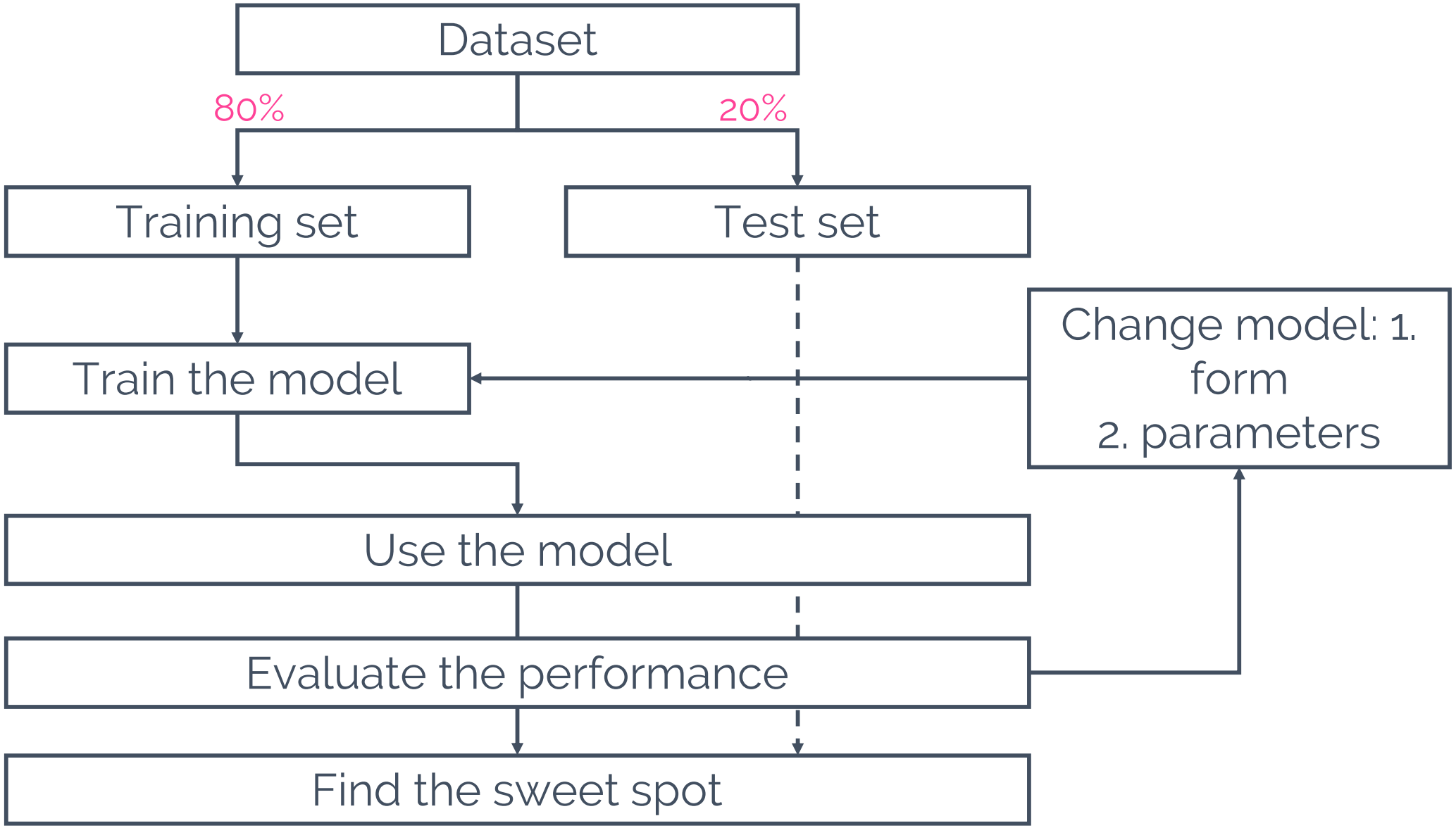


CROSS-VALIDATION



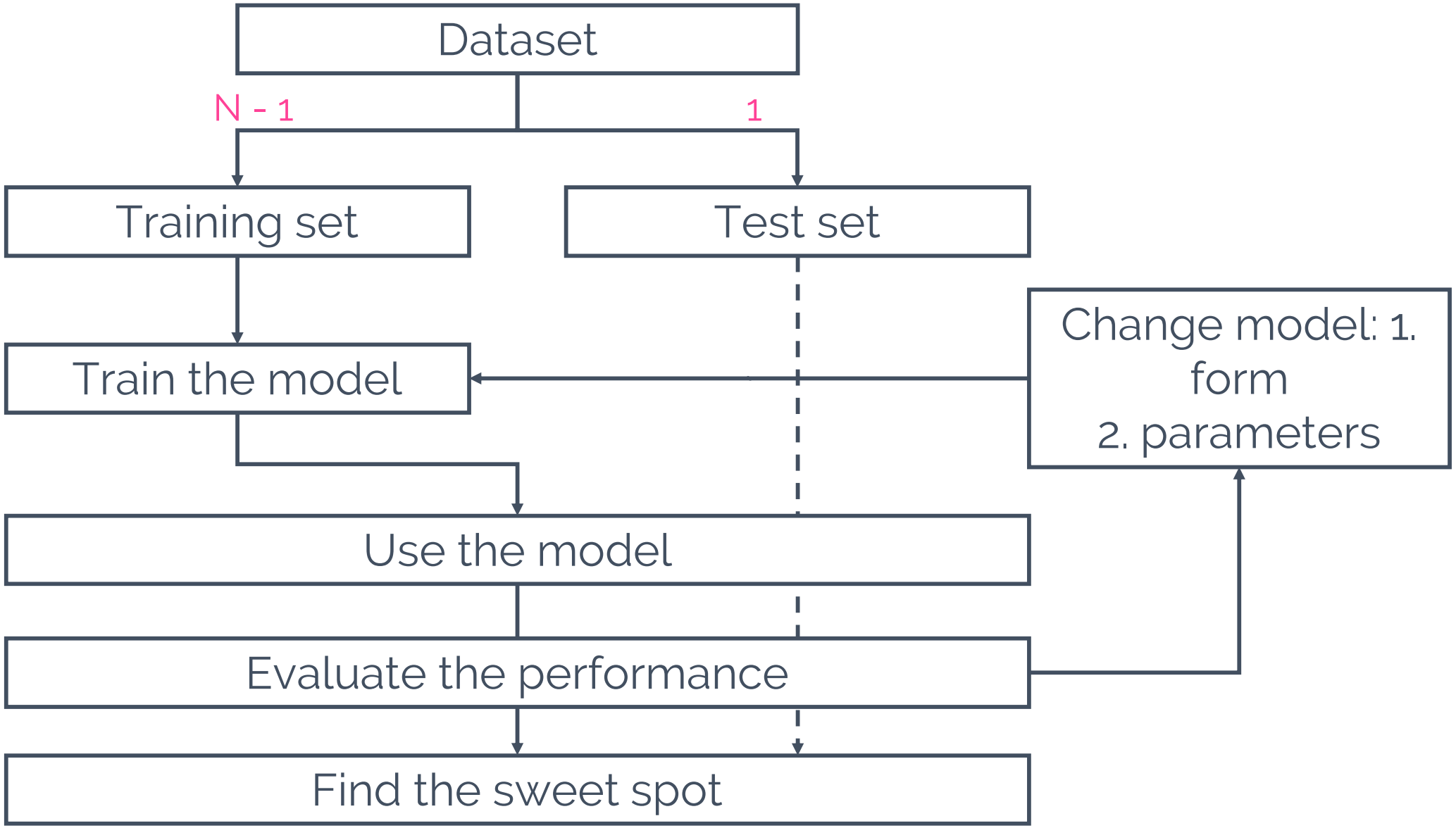
Anneli Kruve

VALIDATION SET APPROACH

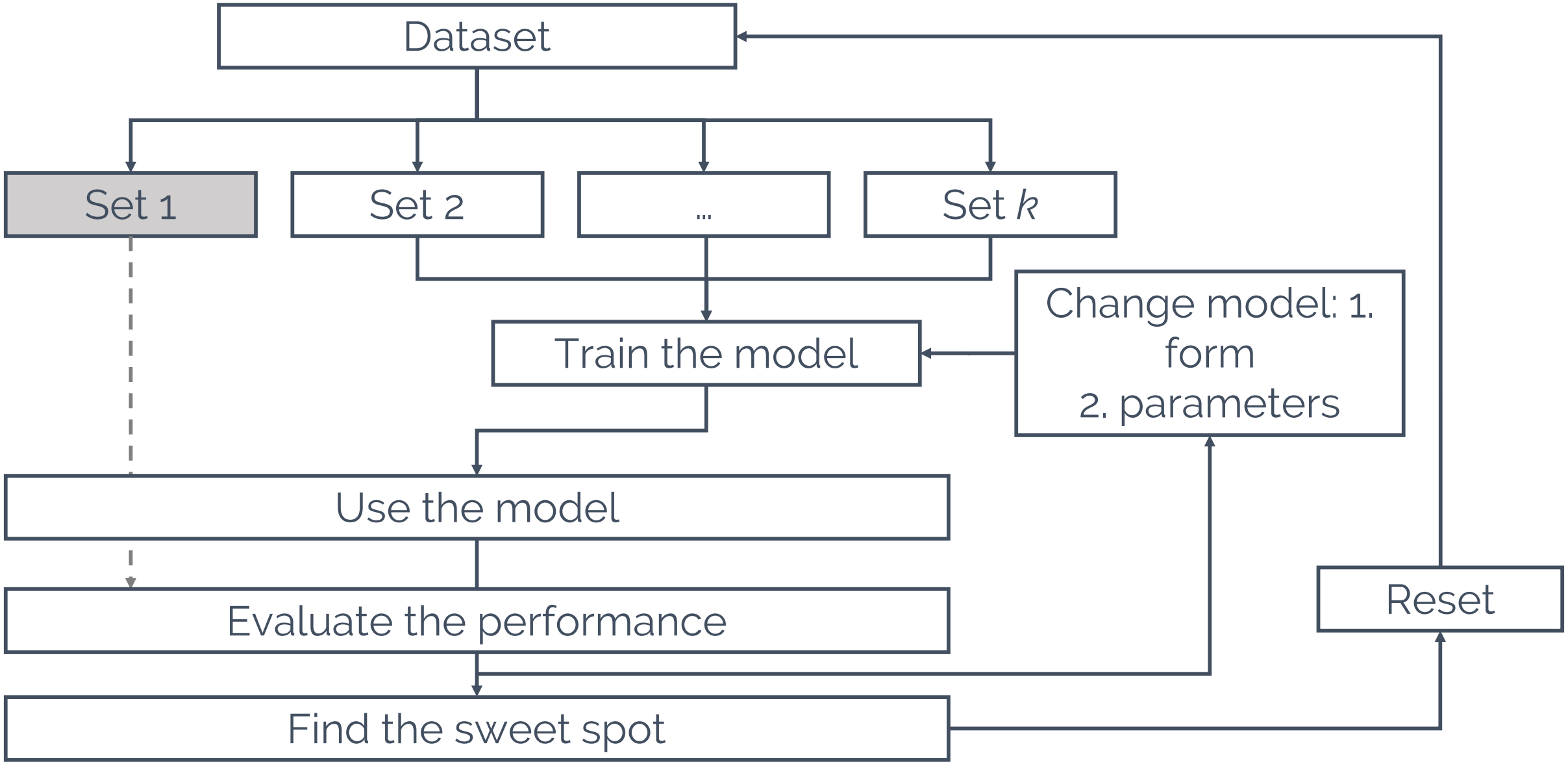


Anneli Kruve

LEAVE-ONE-OUT

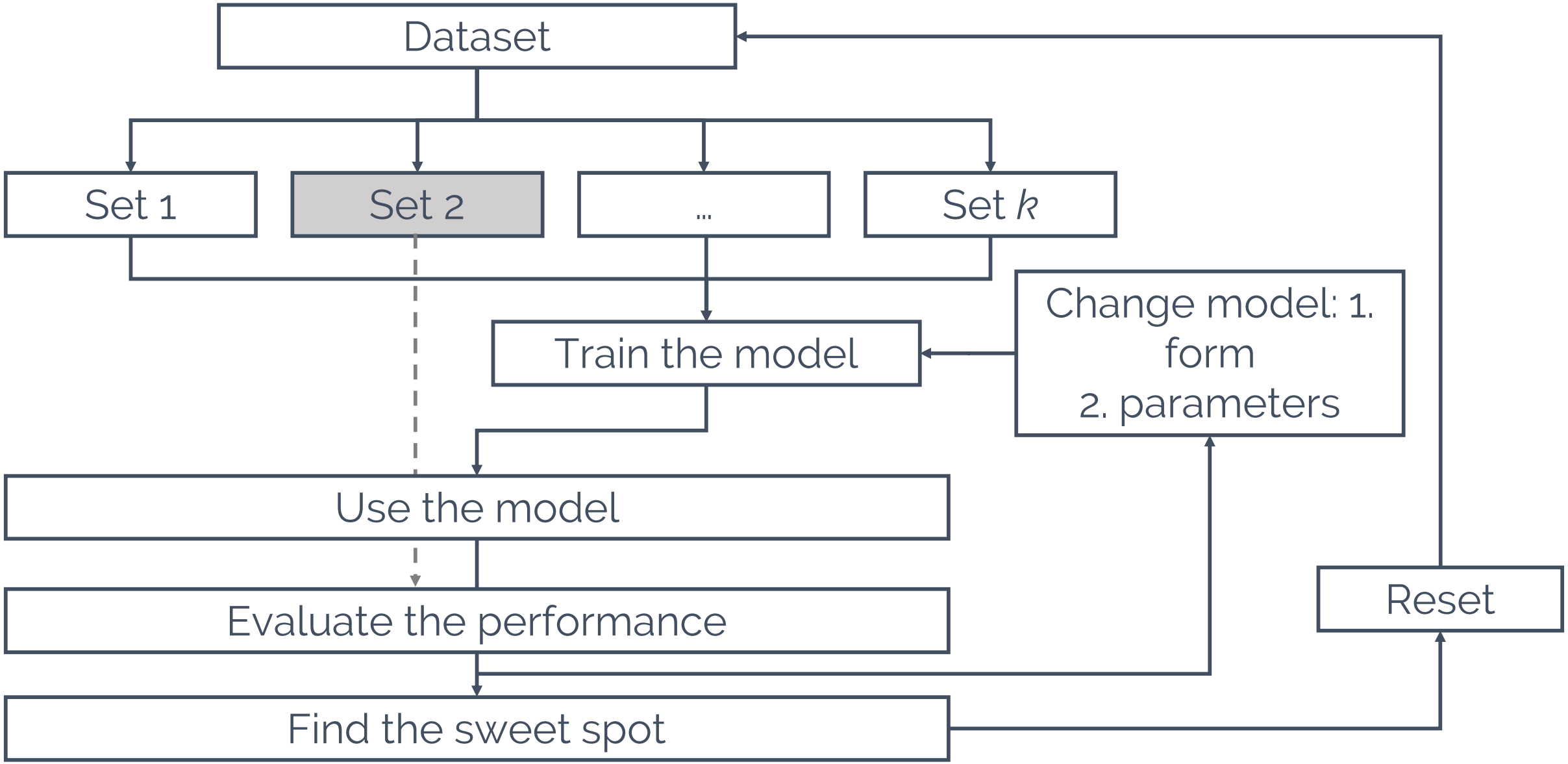


k-FOLD CROSS-VALIDATION



Anneli Kruve

k-FOLD CROSS-VALIDATION



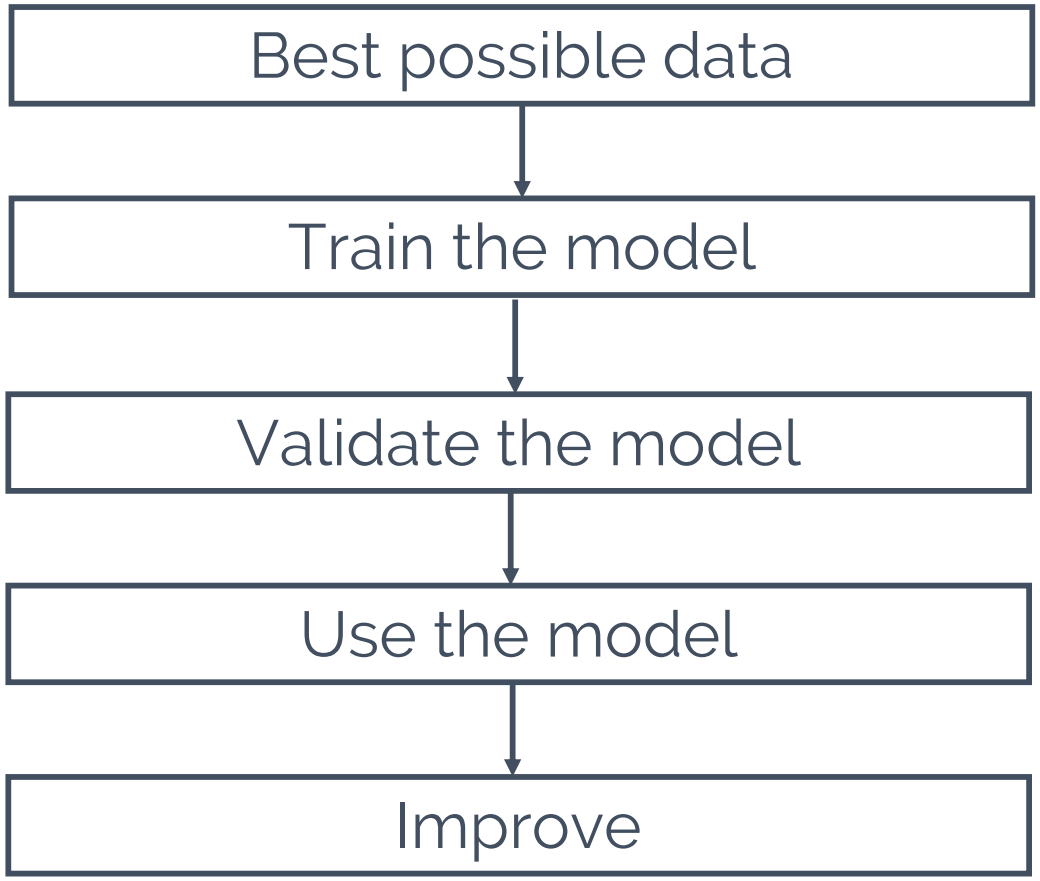
Anneli Kruve

NB!

Random splitting only works if you have homogenous data
Heterogenous data

- Data from several labs
- Several compounds measured under different conditions
- Time split data

MLR IN ACTION



AND NOW

….Interpretation

- It is very important not to trust models blindly
- …and to understand if models represent the real chemistry
- …and to learn from the models to make new chemistry