Dividing your data set

- Do I need to divide my data set?
- How to choose the size of the fractions
- Randomisation

Do I need to divide my data set

YES

Data: $\{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n\}$ n: # observations (# women)

$$\underline{x} := \begin{bmatrix} ge1 \\ ge2 \\ \vdots \\ ge7018 \\ coffee1 \\ coffee2 \\ coffee3 \end{bmatrix}$$

Research question:

Which genes are related to coffee consumption?

Research question

⇓

Working hypotheses

⇓

Final hypotheses

Before accessing the data

After accessing the data

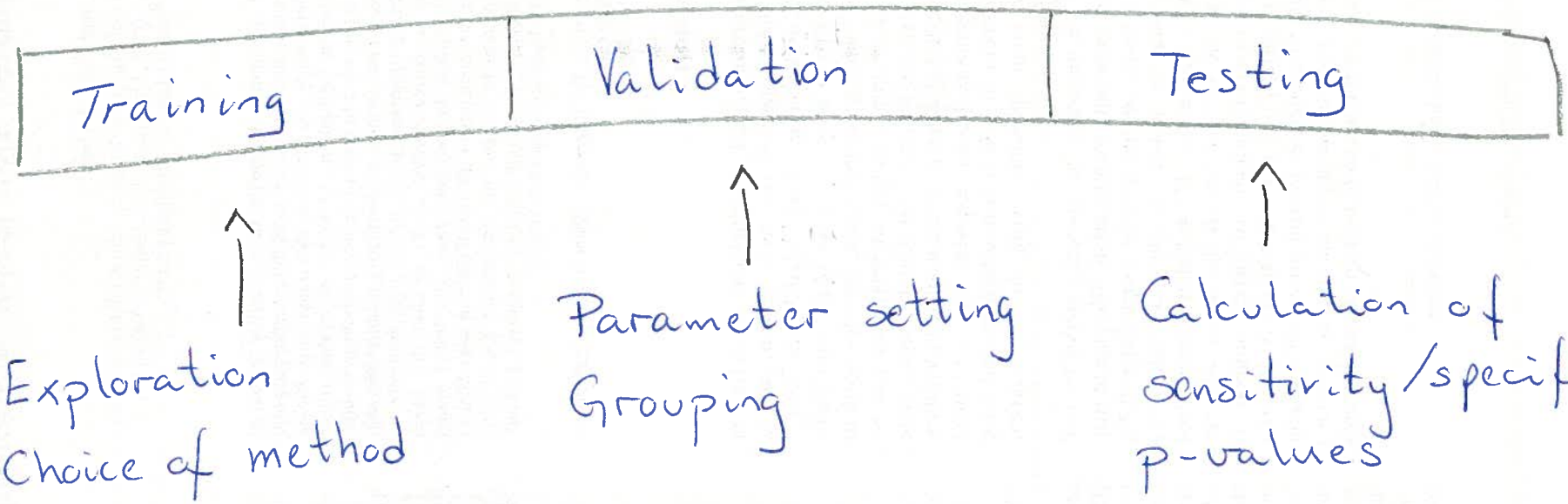
ANY adjustment of your method after accessing the data gives a HIGH risk of wrong conclusions due to overfitting.

ANY adjustment:

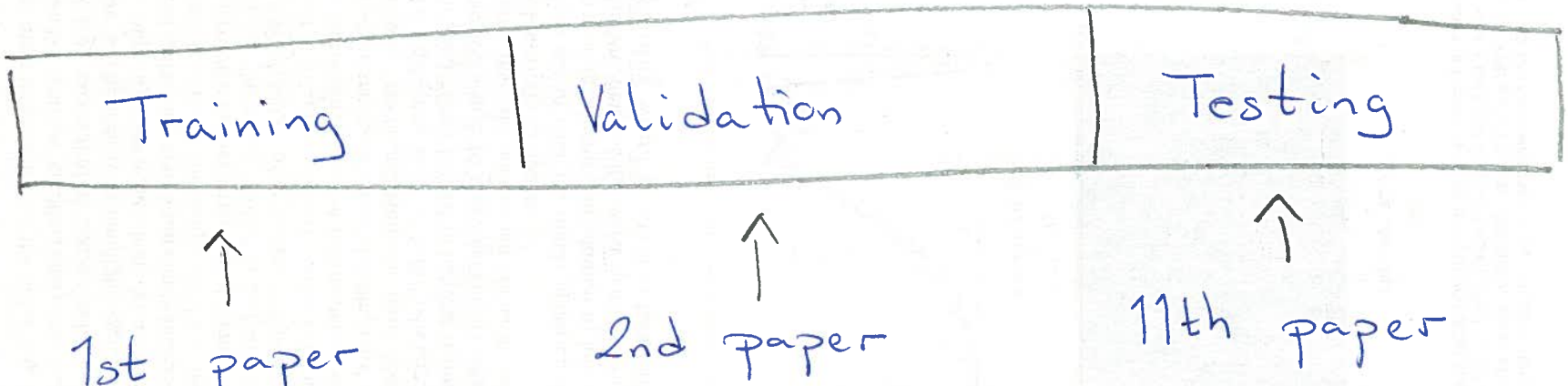
outlier removal, pre-processing, exclusion criteria, p-value versus false discovery rate, grouping, mean versus median, ...

HIGH risk:

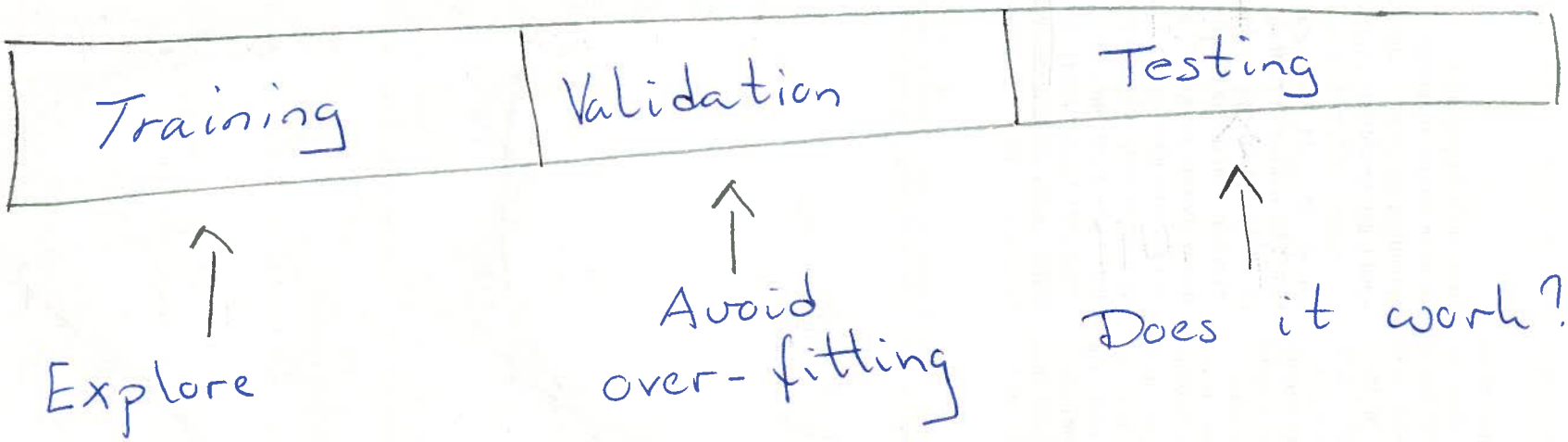
The probability of making a wrong conclusion is inaccessible.


Training
Validation
Testing
Exploration
Choice of method
Parameter setting
Grouping
Calculation of
sensitivity/specificity,
p-values

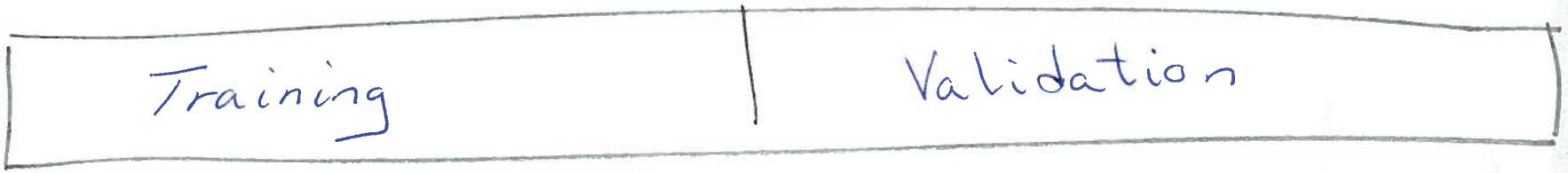
Exploratory data analysis



Exploratory data analysis



Exploratory data analysis



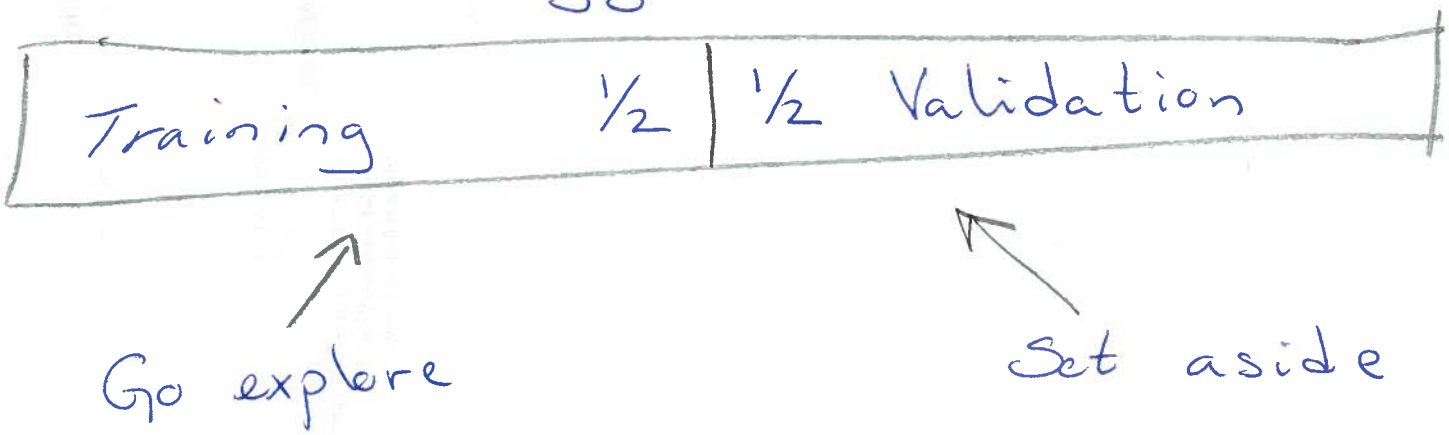
$n_{train} \longrightarrow n$: better model

$n_{validate} \longrightarrow n$: more confidence in the result

$n_{train} \longrightarrow n$: model cannot be validated

$n_{validate} \longrightarrow n$: confident that the model is wrong

Suggestion



Cross-validation: Estimate variance

POSITIVELY BIASED estimate of the performance of your models.

Increase training set?


μ_1
μ_2


μ_1
μ_2

Limited by the number of observations in the smallest class

Training 1/1

Report methods, not results

↑

all

Cross-validation

"Wrong use of cross-validation is the single most costly error in medicine, biology, chemistry, physics, and related fields of research."

Correct use of cross-validation:

Any adjustment must be done independently for each repetition.

Exploratory research: Erase the memory of the researcher.

How to use Cross-Validation and Bootstrapping in exploratory research?

- Overly optimistic estimate of how your method works on an independent data set!
- Choose method, estimate variation in outcome

Randomisation

- Random not consecutive
- Stratified or not

Stratify if smallest class is very small.
Stratification gives positive bias.

Summary

- Change of plans (or no plan)

→ Validation

- Too small data set

→ Methods, not results