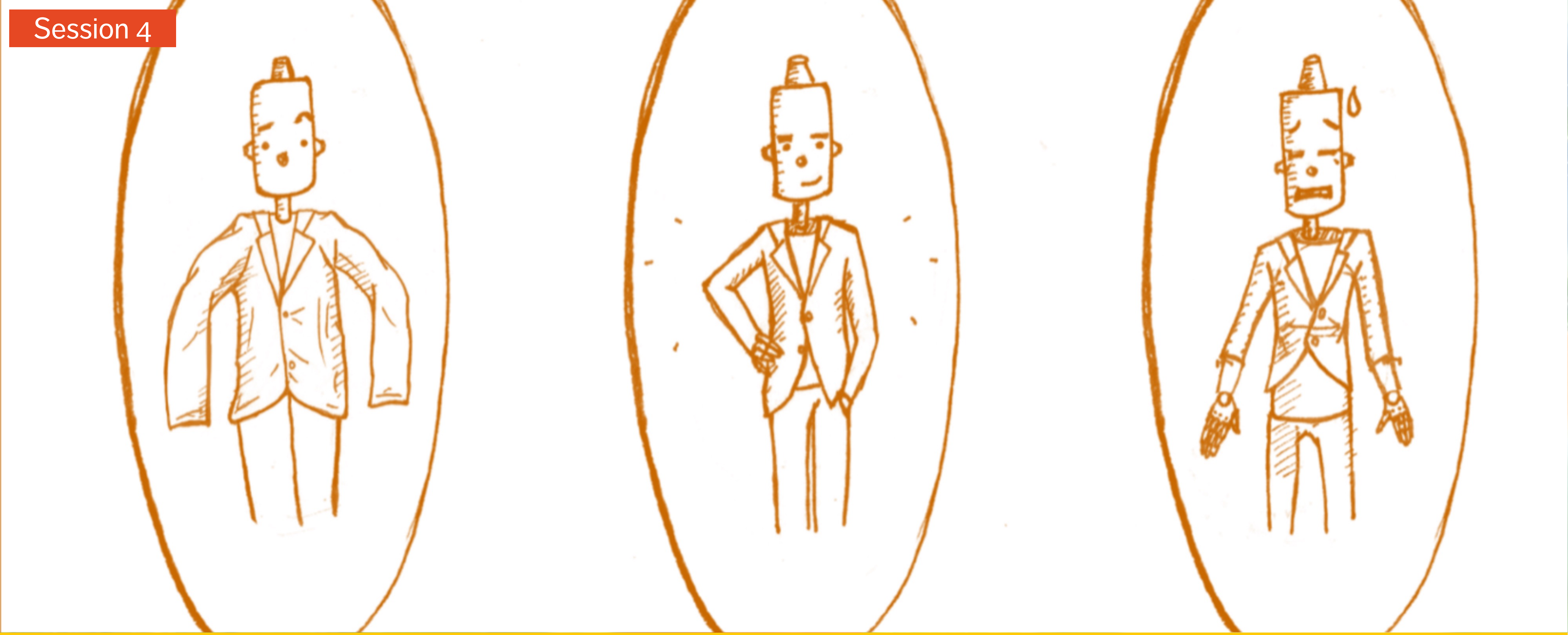
Session 4



11. Underfitting and Overfitting/Transferability

Fundamentals

Rules or models generated by any technique on a **training set** have to be generalizable to **new data** (or **validation/ testing sets**) to be useful.

Problems arise when knowledge that is gained from **supervised learning** does not generalize properly to the data.

Unsupervised learning can also be affected.

Ironically, this may occur if the rules or models fit the training set **too well** – the results are **too specific to the training set**.

Example of Rules

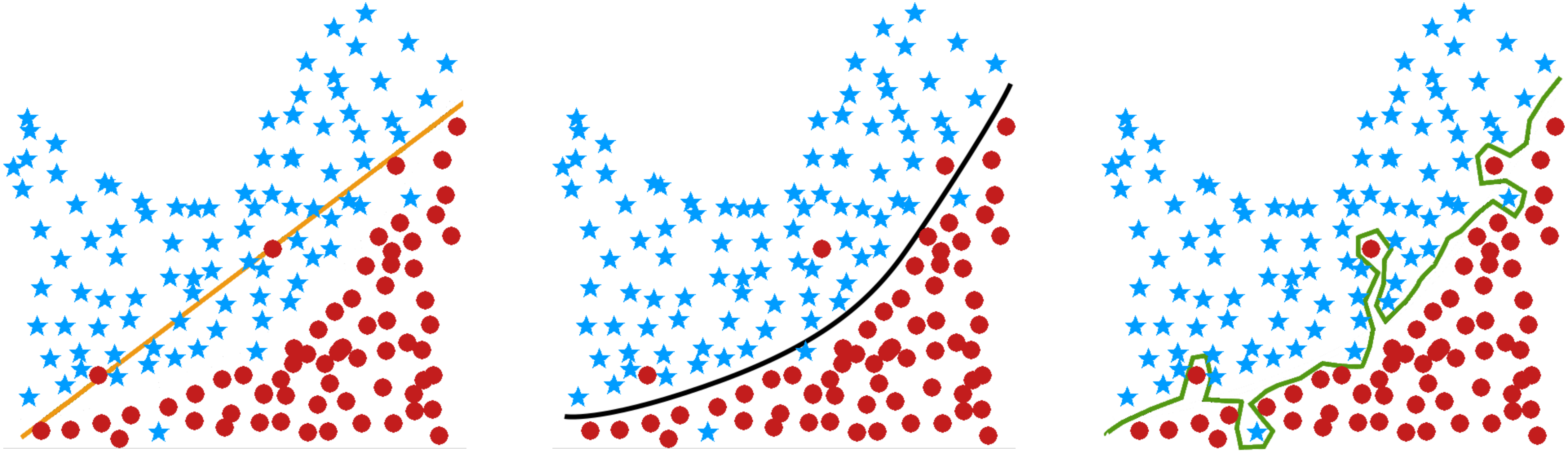
Rule I: based on a survey of 400 Germans, we infer that 43.75% of the world's population has black hair, 37.5% have brown hair, 9% have blond hair, 0.25% have red hair, and 9.5% grey hair.

Rule II: humans' hair colour is either black, brown, blond, red, or grey.

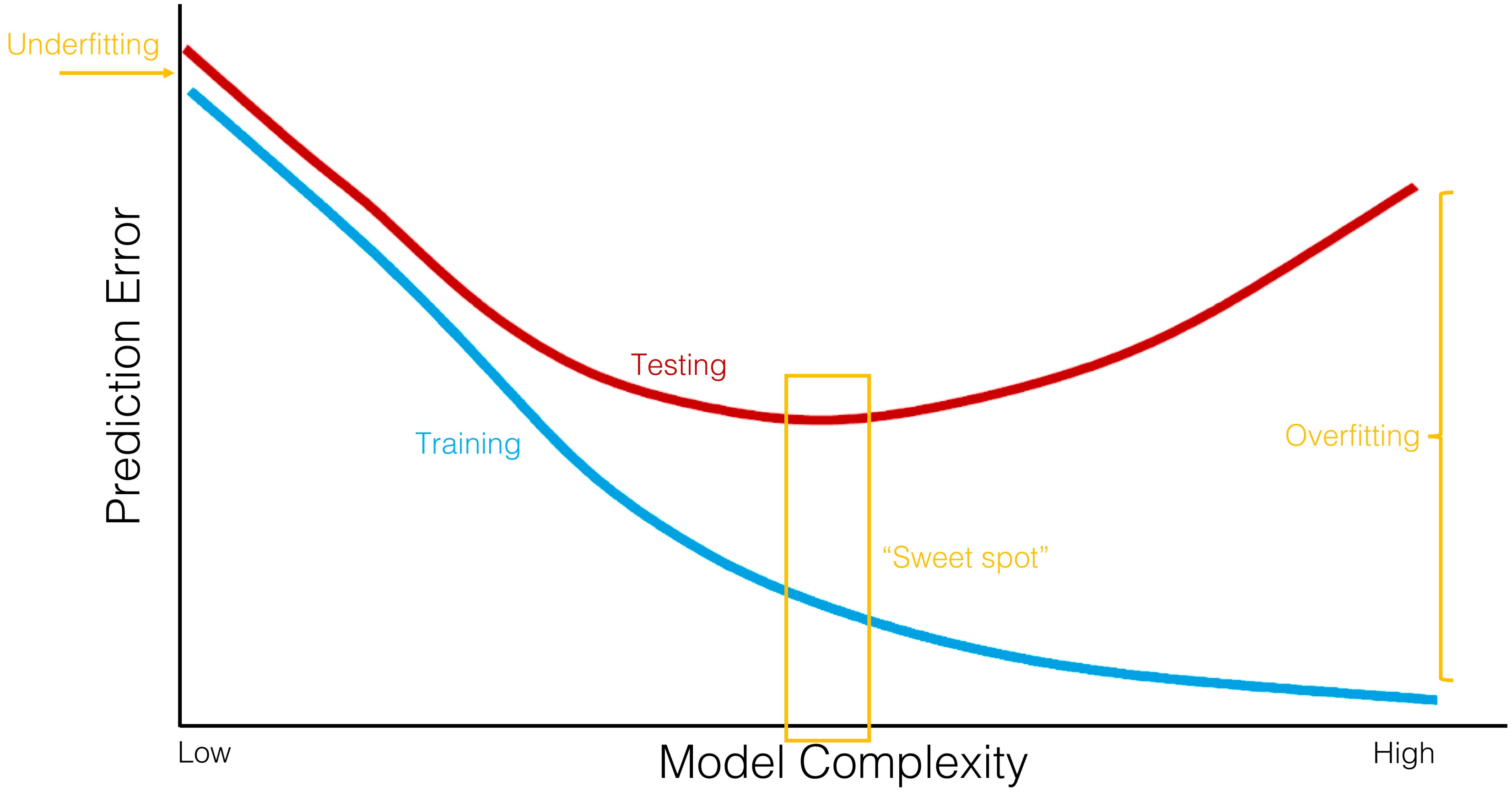
Rule III: approx. 40% of humans have black hair, 40% have brown hair, 5% blond, 2% red and 13% grey.

Which of the 3 rules is most useful? The most vague? Which is overly specific?

Goldilocks and the Three Models



Bias-Variance Trade-Off



ALWAYS evaluate models on unseen (testing) data!

Bias-Variance Trade-Off

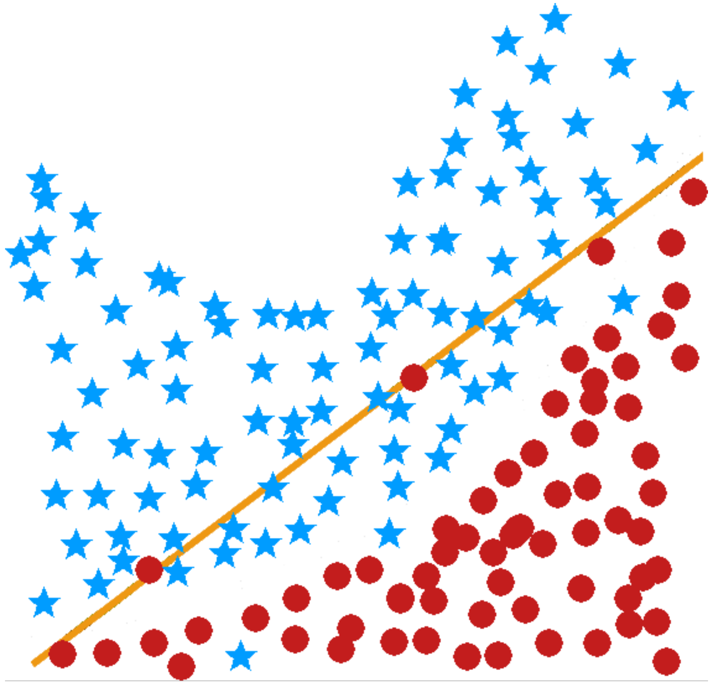
We **build** a model on **historical data** and **evaluate** its performance on **new data**.

Let $\mathrm{Error}_{\mathrm{Test}}$ be the model's performance on test data:

$$\mathrm{Error}_{\mathrm{Test}} = \mathrm{Bias}^2_{\mathrm{Model}} + \mathrm{Variance}_{\mathrm{Model}}$$

The **bias** measures the model's prediction **accuracy**; the **variance**, its **sensitivity** to (small) changes in the data.

Goldilocks and the Three Models



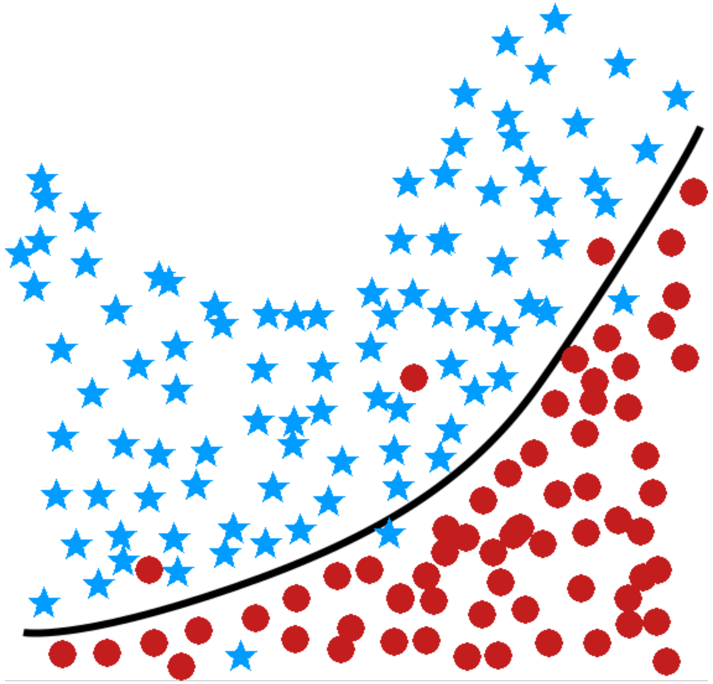
underfit

bias = high

variance = low

error = **high**

predictions are not very accurate

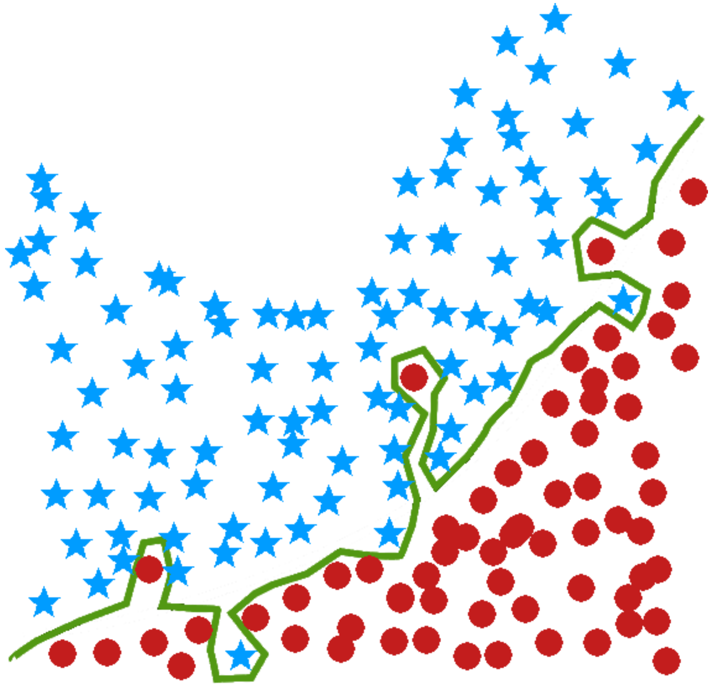


just right

bias = medium

variance = medium

error = **medium**



overfit

bias = low

variance = high

error = **high**

model is too specific to the data

Possible Solutions

Underfitting can be overcome by considering models that are more complex.

Overfitting can be overcome in several ways:

- **using multiple training sets**
  overlap is allowed (or not: see cross-validation)
- **using larger training sets**
  70% - 30% split is suggested
- **optimizing the data instead of the model**
  models are only as good as the data

Recommended Procedures

Small datasets (less than a few hundred observations)

- use 100-200 repetitions of a **bootstrap** procedure

Average-sized datasets (less than a few thousand observations)

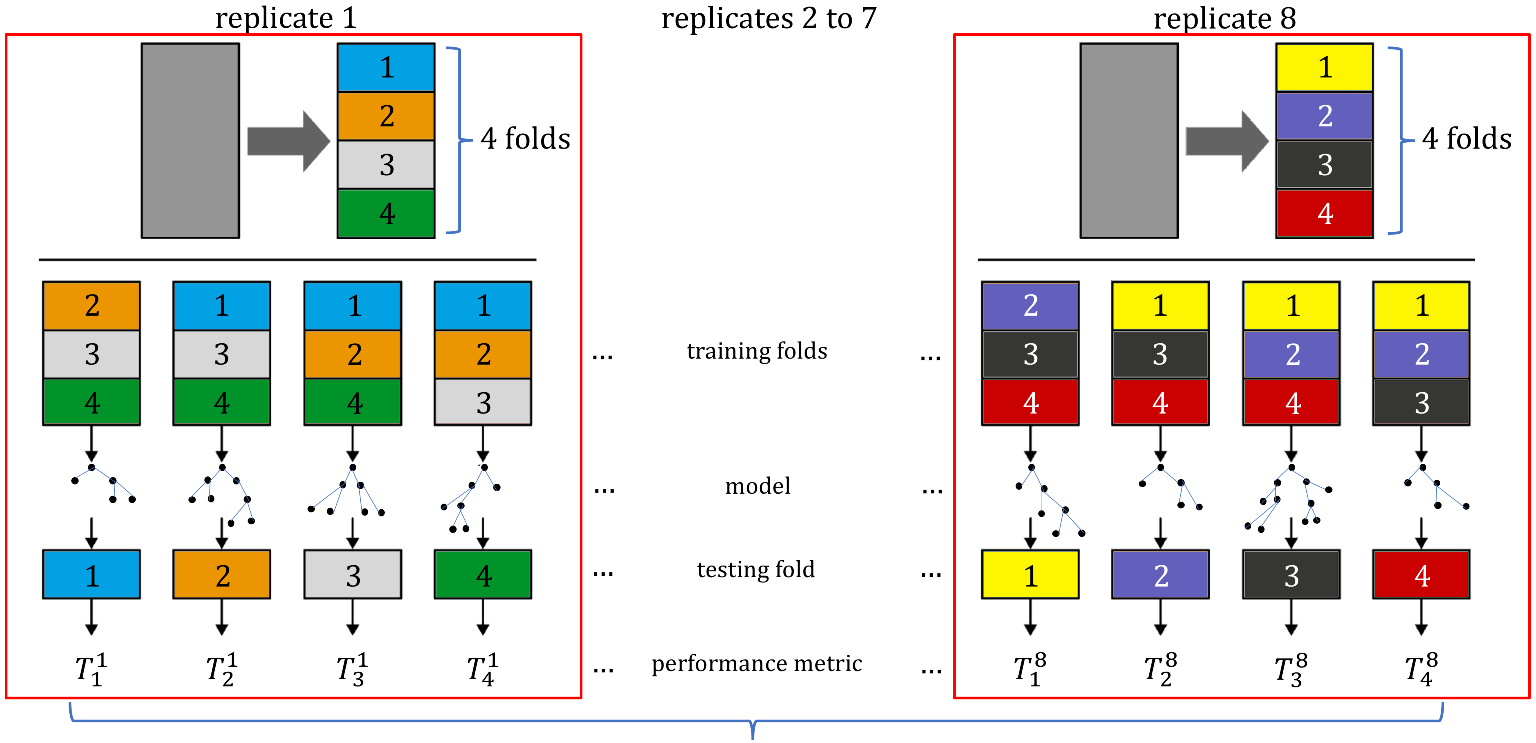
- use a few repetitions of 10-fold **cross-validation** on the training set (see next slide)

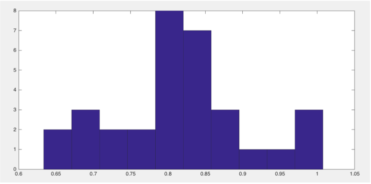
Large datasets

- use a few repetitions of **holdout** (70%-30%) split

The **decision boundaries** depend on computing power and the number of tasks in the workflows.

Cross-Validation





mean accuracy = 0.81
standard dev = 0.09

Appropriateness and Transferability

Data science and machine learning models will continue to be used heavily in the coming years.

We have discussed pros and cons of some of the applications on ethical and other non-technical grounds, but there are also **technical challenges**.

DS/ML methods are **not appropriate**:

- If you must absolutely use an existing (**legacy**) datasets instead of an **ideal** dataset ("it's the best data we have!")

Appropriateness and Transferability

DS/ML methods are **not appropriate** (cont.):

- if the dataset has attributes that usefully predict a value of interest, but which are **not available** when a prediction is required

  Example: the total time spent on a website may be predictive of a visitor's future purchases, but the prediction must be made before the total time spent on the website is known.

- if you attempt to predict **class membership** using an **unsupervised** learning algorithm

  Example: clustering loan default data might lead to a cluster that contains multiple defaulters. If new instances get added to this cluster, should they automatically be viewed as loan defaulters? (no)

Non-Transferable Assumptions

Every model makes certain assumptions about what is and is not **relevant** to its workings, but there is a tendency to only gather data which is **assumed** to be relevant to a particular situation.

If data is used in other contexts, or to make predictions depending on attributes without data, validating the results may prove impossible.

- **Example:** can we use a model that predicts mortgage defaulters to also predict car loan defaulters? A car is not a house: they play different roles in an individual's life; the values are of different magnitudes; and so on...
- That being said, is there truly no link between mortgage defaults and car loan defaults?

Suggested Reading

Underfitting and Overfitting/ Transferability

Data Understanding, Data Analysis, Data Science
Machine Learning 101

Issues and Challenges

- Overfitting/Underfitting
- Appropriateness and Transferability

Regression and Value Estimation

*Statistical Learning (advanced)

- Model Evaluation
- Bias-Variance Trade-Off

*Resampling Methods (advanced)

- Cross-Validation

*Model Selection (advanced)

- Selecting the Optimal Model (Validation and Cross-Validation Reprise)

Exercises

Underfitting and Overfitting/ Transferability

1. This exercise illustrates overfitting/underfitting.
   a. Randomly generate $n = 150$ values in $[0,10]$ for the predictor x.
   b. Randomly generate $n = 150$ response values according to $y = 10 + x - 2x^2 + 17x^3 + \varepsilon$, where ε is a random error term of your choice.
   c. Fit a linear model, a quadratic model, a cubic model, and a polynomial model of degree 10 to the data.
   d. Add 3 observations to the data as in steps a. and b. Repeat step c. Do the models change much?
   e. Which model(s) would you trust to make predictions on new data?

2. Modify the Gapminder example from Cross-Validation to select a model in the previous question.