TRAIN-TEST LEAKAGE

- Ultimately we are interested in predicting on future data!
- We can not use ‘knowledge’ about future (test) data during training.
- Requires clearly separating train and test data in order to avoid overoptimistic conclusions:
- **Examples:**
  - Using knowledge about which features are relevant in test data
  - Scaling based on statistics of the test data
  - Missing patterns in the test data not present in training data

TRAIN-TEST LEAKAGE