Elements of Machine Learning & Data Science

Winter semester 2023/24

Evaluation of Supervised Learning (1)

Prof. Holger Hoos (partially based on material from Wil van der Aalst)

Learning Goals

At the end of this module, students should be able to

- assess the quality of a model obtained from a supervised machine learning method
  using widely accepted methods, including standard performance metrics, confusion matrices, ROC curves
- demonstrate understanding and working knowledge of the problems that can occur
  when using supervised learning procedures and the models obtained from them
- explain when and why it is important to distinguish between training, validation and testing data
- explain standard validation techniques, including *k*-fold and leave-one-out cross-validation
- assess performance differences using appropriate statistical techniques
- explain the problems that can arise from unbalanced data sets and demonstrate understanding
  as well as working knowledge of methods for addressing these problems

Key questions:

- How good is an ML model?
- How good could an ML model be?

Key questions:

- How good is an ML model?
- How good could an ML model be?

You have used supervised ML to train a predictive model.

Question: How do you assess the quality of the model?

Motivation: Predicting delayed flights

ID	Origin	Destination	Precipitation	...	Traffic	Target
1	Frankfurt	Cologne	139	...	152	On Time
2	Madrid	Paris	349	...	55	On Time
3	La Paz	Madrid	702	...	76	Delayed
4	Hanoi	Singapore	251	...	169	On Time
5	Dubai	Frankfurt	615	...	117	Delayed
6	Cologne	Madrid	400	...	89	On Time
7	Bergen	Paris	698	...	28	Delayed
8	Rome	Barcelona	322	...	9	On Time
9	Berlin	Rome	221	...	5	On Time
10	Paris	Paris	132	...	165	On Time
11	Toronto	Frankfurt	730	...	220	Delayed
...	...	...	...	...	...	...

You have used supervised ML to train a predictive model.

Question: How do you assess the quality of the model?

TPS = Think, Pair, Share Exercises

1. Be ready to take some notes.
2. Think about the problem/question.
3. Jot down your answers (bullet points).

4. Pair up with your neighbour, explain/discuss your answers.
5. Modify your answers based on your discussion.

6. Volunteer to give/explain your answer to everyone.

TPS Exercise:

You have used supervised ML to train a predictive model.

Question: How do you assess the quality of the model?

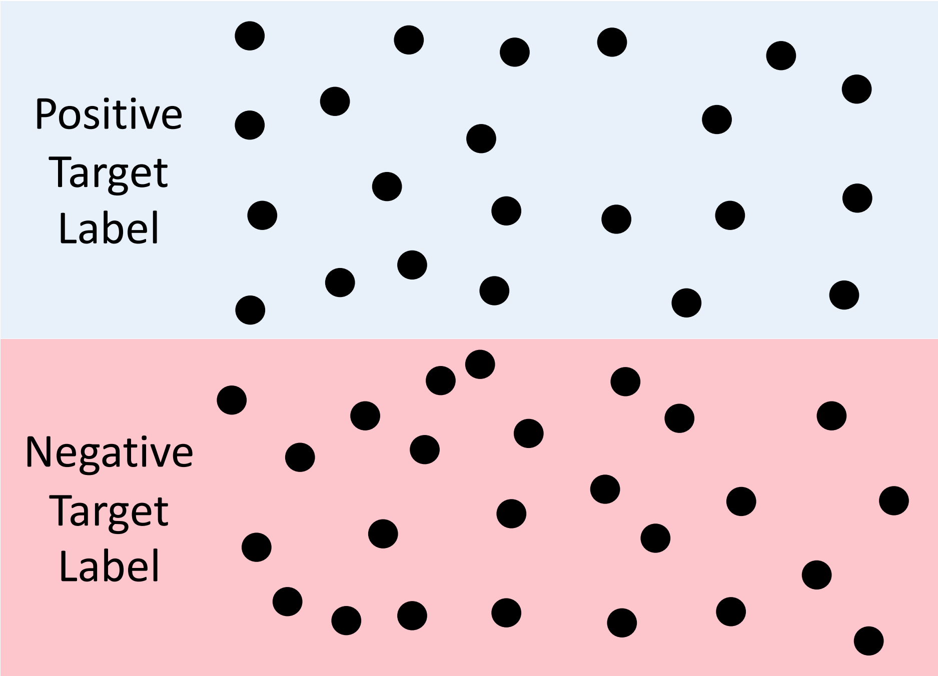
Running Example

Predicting delayed flights
(set of 20 instances)

- Target Feature:
  On Time (positive),
  Delayed (negative)

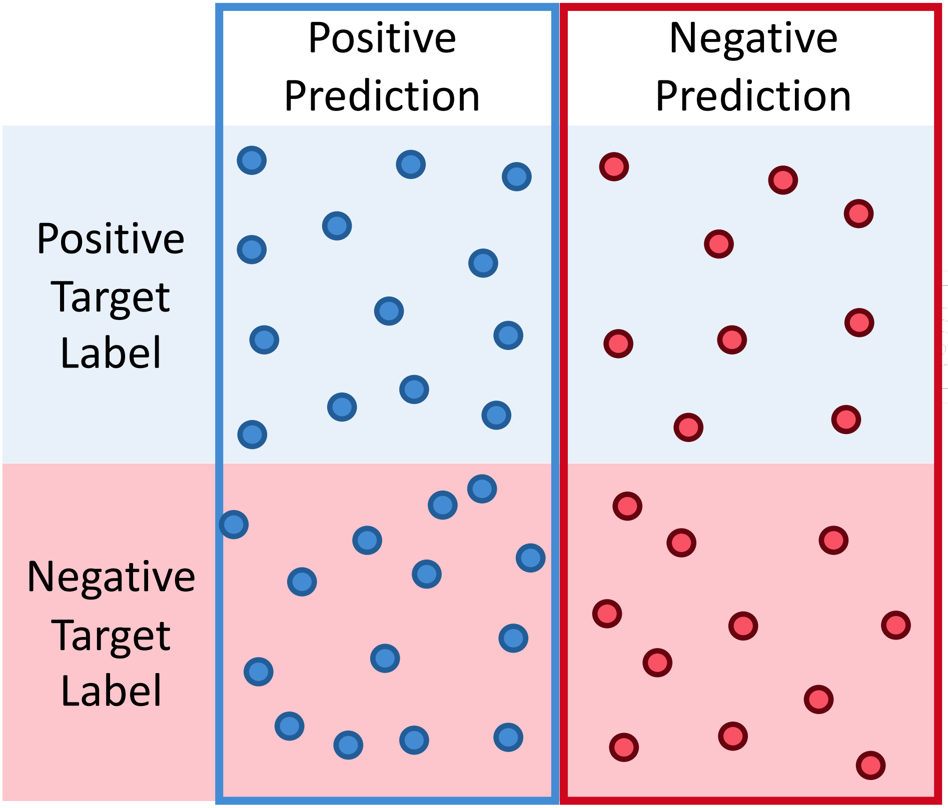
ID	Target Label	Prediction
1	On Time	Delayed
2	On Time	Delayed
3	Delayed	Delayed
4	On Time	On Time
5	Delayed	Delayed
6	On Time	On Time
7	Delayed	Delayed
8	On Time	On Time
9	On Time	On Time
10	On Time	On Time

ID	Target Label	Prediction
11	Delayed	Delayed
12	On Time	Delayed
13	Delayed	Delayed
14	Delayed	Delayed
15	Delayed	Delayed
16	Delayed	Delayed
17	Delayed	On Time
18	On Time	On Time
19	Delayed	Delayed
20	Delayed	**On Time**



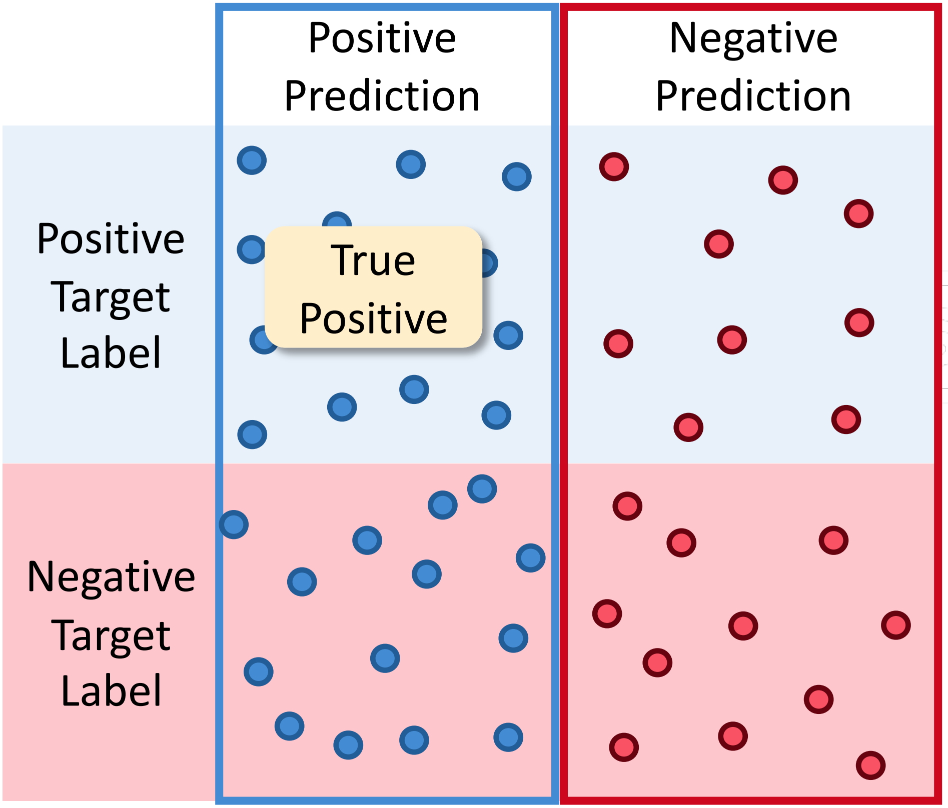
ID	Target Label	Prediction	
1	On Time	Delayed	
2	On Time	Delayed	
3	Delayed	Delayed	
4	On Time	On Time	
5	Delayed	Delayed	
6	On Time	On Time	
7	Delayed	Delayed	
8	On Time	On Time	
9	On Time	On Time	
10	On Time	On Time	

ID	Target Label	Prediction	
11	Delayed	Delayed	
12	On Time	Delayed	
13	Delayed	Delayed	
14	Delayed	Delayed	
15	Delayed	Delayed	
16	Delayed	Delayed	
17	Delayed	On Time	
18	On Time	On Time	
19	Delayed	Delayed	
20	Delayed	On Time	



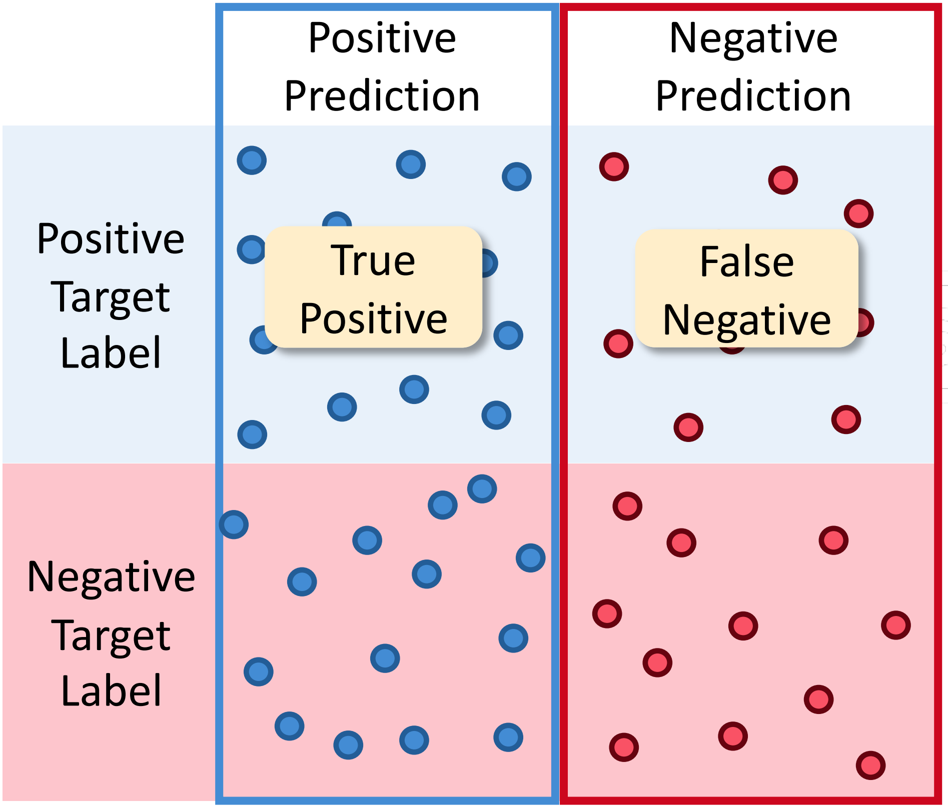
ID	Target Label	Prediction	
1	On Time	Delayed	
2	On Time	Delayed	
3	Delayed	Delayed	
4	On Time	On Time	
5	Delayed	Delayed	
6	On Time	On Time	
7	Delayed	Delayed	
8	On Time	On Time	
9	On Time	On Time	
10	On Time	On Time	

ID	Target Label	Prediction	
11	Delayed	Delayed	
12	On Time	Delayed	
13	Delayed	Delayed	
14	Delayed	Delayed	
15	Delayed	Delayed	
16	Delayed	Delayed	
17	Delayed	On Time	
18	On Time	On Time	
19	Delayed	Delayed	
20	Delayed	On Time	



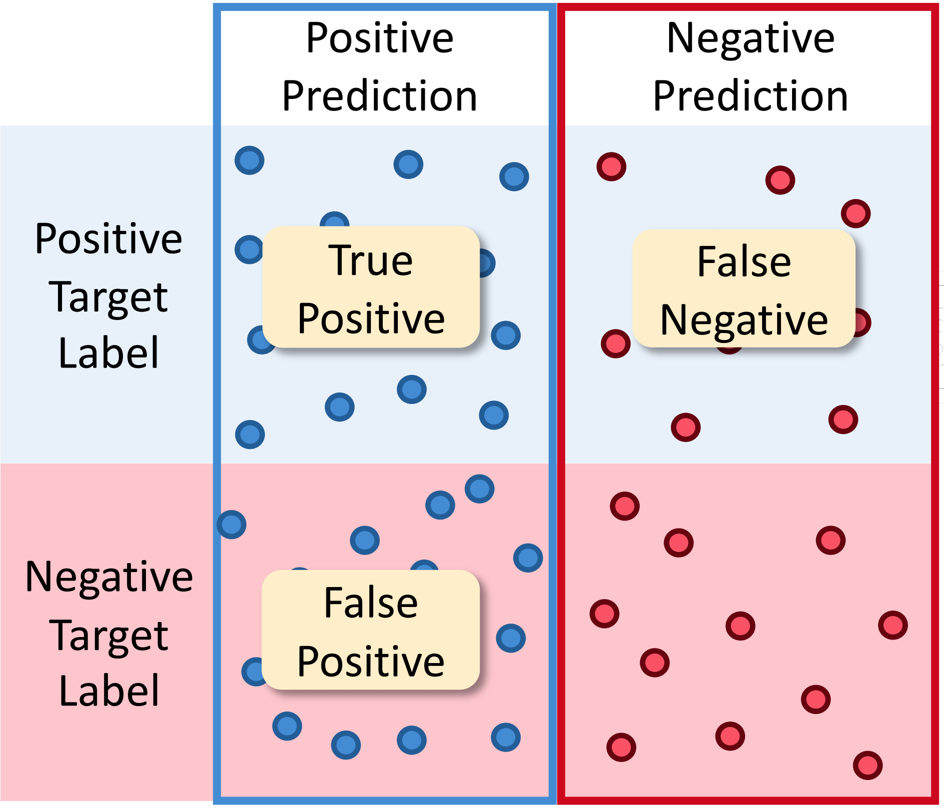
ID	Target Label	Prediction	
1	On Time	Delayed	
2	On Time	Delayed	
3	Delayed	Delayed	
4	On Time	On Time	**TP**
5	Delayed	Delayed	
6	On Time	On Time	**TP**
7	Delayed	Delayed	
8	On Time	On Time	**TP**
9	On Time	On Time	**TP**
10	On Time	On Time	**TP**

ID	Target Label	Prediction	
11	Delayed	Delayed	
12	On Time	Delayed	
13	Delayed	Delayed	
14	Delayed	Delayed	
15	Delayed	Delayed	
16	Delayed	Delayed	
17	Delayed	On Time	
18	On Time	On Time	**TP**
19	Delayed	Delayed	
20	Delayed	On Time	



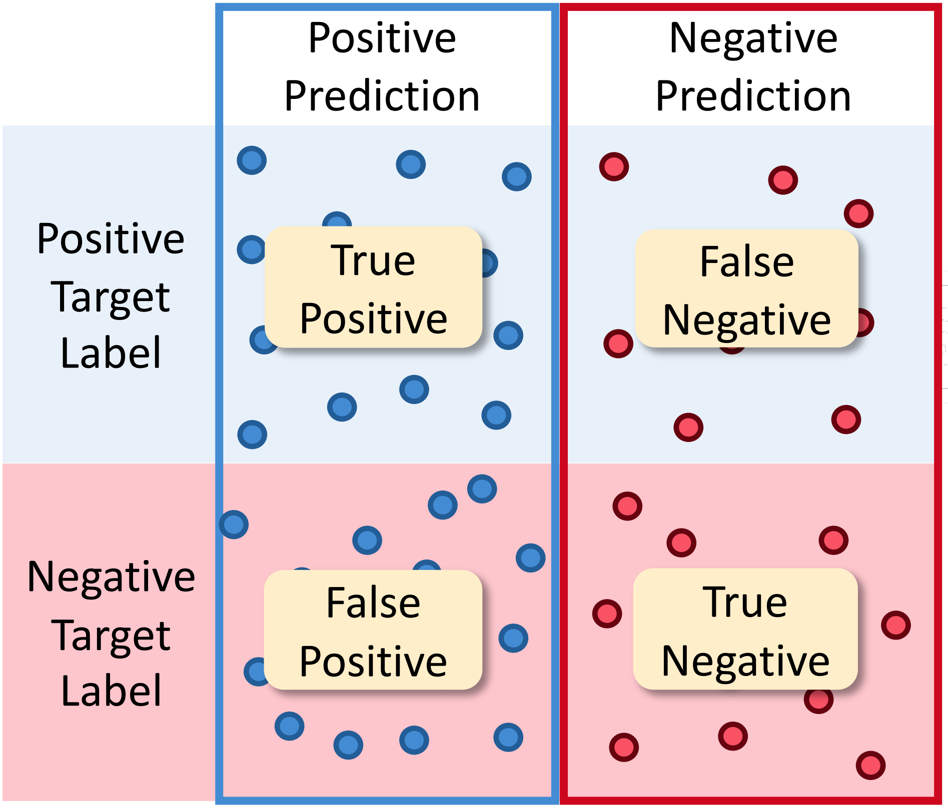
ID	Target Label	Prediction	
1	On Time	Delayed	**FN**
2	On Time	Delayed	**FN**
3	Delayed	Delayed	
4	On Time	On Time	**TP**
5	Delayed	Delayed	
6	On Time	On Time	**TP**
7	Delayed	Delayed	
8	On Time	On Time	**TP**
9	On Time	On Time	**TP**
10	On Time	On Time	**TP**

ID	Target Label	Prediction	
11	Delayed	Delayed	
12	On Time	Delayed	**FN**
13	Delayed	Delayed	
14	Delayed	Delayed	
15	Delayed	Delayed	
16	Delayed	Delayed	
17	Delayed	On Time	
18	On Time	On Time	**TP**
19	Delayed	Delayed	
20	Delayed	On Time	



ID	Target Label	Prediction	
1	On Time	Delayed	**FN**
2	On Time	Delayed	**FN**
3	Delayed	Delayed	
4	On Time	On Time	**TP**
5	Delayed	Delayed	
6	On Time	On Time	**TP**
7	Delayed	Delayed	
8	On Time	On Time	**TP**
9	On Time	On Time	**TP**
10	On Time	On Time	**TP**

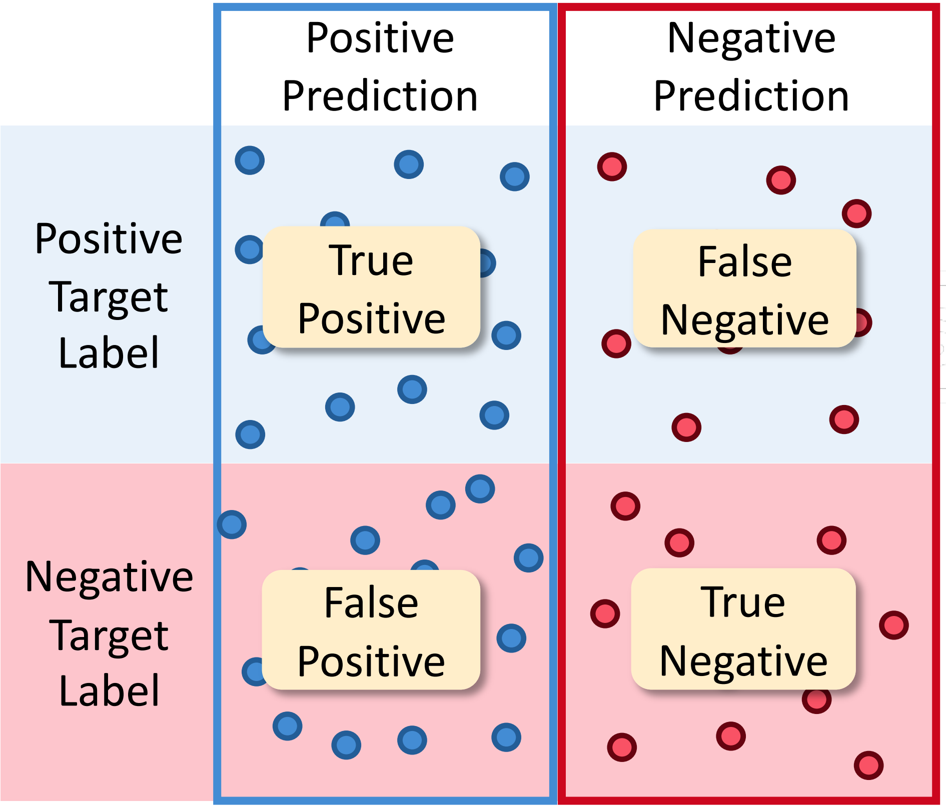
ID	Target Label	Prediction	
11	Delayed	Delayed	
12	On Time	Delayed	**FN**
13	Delayed	Delayed	
14	Delayed	Delayed	
15	Delayed	Delayed	
16	Delayed	Delayed	
17	Delayed	On Time	**FP**
18	On Time	On Time	**TP**
19	Delayed	Delayed	
20	Delayed	On Time	**FP**



ID	Target Label	Prediction	
1	On Time	Delayed	**FN**
2	On Time	Delayed	**FN**
3	Delayed	Delayed	**TN**
4	On Time	On Time	**TP**
5	Delayed	Delayed	**TN**
6	On Time	On Time	**TP**
7	Delayed	Delayed	**TN**
8	On Time	On Time	**TP**
9	On Time	On Time	**TP**
10	On Time	On Time	**TP**

ID	Target Label	Prediction	
11	Delayed	Delayed	**TN**
12	On Time	Delayed	**FN**
13	Delayed	Delayed	**TN**
14	Delayed	Delayed	**TN**
15	Delayed	Delayed	**TN**
16	Delayed	Delayed	**TN**
17	Delayed	On Time	**FP**
18	On Time	On Time	**TP**
19	Delayed	Delayed	**TN**
20	Delayed	On Time	**FP**

Confusion Matrix



ID	Target Label	Prediction	
1	On Time	Delayed	**FN**
2	On Time	Delayed	**FN**
3	Delayed	Delayed	**TN**
4	On Time	On Time	**TP**
5	Delayed	Delayed	**TN**
6	On Time	On Time	**TP**
7	Delayed	Delayed	**TN**
8	On Time	On Time	**TP**
9	On Time	On Time	**TP**
10	On Time	On Time	**TP**

ID	Target Label	Prediction	
11	Delayed	Delayed	**TN**
12	On Time	Delayed	**FN**
13	Delayed	Delayed	**TN**
14	Delayed	Delayed	**TN**
15	Delayed	Delayed	**TN**
16	Delayed	Delayed	**TN**
17	Delayed	On Time	**FP**
18	On Time	On Time	**TP**
19	Delayed	Delayed	**TN**
20	Delayed	On Time	**FP**

Confusion Matrix

	Positive Prediction	Negative Prediction
Positive Target Label	**TP** (number of true positives)	**FN** (number of false negatives)
Negative Target Label	**FP** (number of false positives)	**TN** (number of true negatives)

ID	Target Label	Prediction	
1	On Time	Delayed	**FN**
2	On Time	Delayed	**FN**
3	Delayed	Delayed	**TN**
4	On Time	On Time	**TP**
5	Delayed	Delayed	**TN**
6	On Time	On Time	**TP**
7	Delayed	Delayed	**TN**
8	On Time	On Time	**TP**
9	On Time	On Time	**TP**
10	On Time	On Time	**TP**

ID	Target Label	Prediction	
11	Delayed	Delayed	**TN**
12	On Time	Delayed	**FN**
13	Delayed	Delayed	**TN**
14	Delayed	Delayed	**TN**
15	Delayed	Delayed	**TN**
16	Delayed	Delayed	**TN**
17	Delayed	On Time	**FP**
18	On Time	On Time	**TP**
19	Delayed	Delayed	**TN**
20	Delayed	On Time	**FP**

Confusion Matrix

	Positive Prediction	Negative Prediction
Positive Target Label	**6**	**3**
Negative Target Label	**2**	**9**

ID	Target Label	Prediction	
1	On Time	Delayed	**FN**
2	On Time	Delayed	**FN**
3	Delayed	Delayed	**TN**
4	On Time	On Time	**TP**
5	Delayed	Delayed	**TN**
6	On Time	On Time	**TP**
7	Delayed	Delayed	**TN**
8	On Time	On Time	**TP**
9	On Time	On Time	**TP**
10	On Time	On Time	**TP**

ID	Target Label	Prediction	
11	Delayed	Delayed	**TN**
12	On Time	Delayed	**FN**
13	Delayed	Delayed	**TN**
14	Delayed	Delayed	**TN**
15	Delayed	Delayed	**TN**
16	Delayed	Delayed	**TN**
17	Delayed	On Time	**FP**
18	On Time	On Time	**TP**
19	Delayed	Delayed	**TN**
20	Delayed	On Time	**FP**

TPS Exercise:

We have trained a predictive model using supervised learning and computed a confusion matrix based on predictions on a given set of data.

Question:
How can we assess performance with a single number?

Confusion Matrix ➡️ Performance Measures

	Positive Prediction	Negative Prediction
Positive Target Label	**TP=6**	**FN=3**
Negative Target Label	**FP=2**	**TN=9**

True Positive Rate: $TPR = \frac{TP}{TP+FN}$

False Negative Rate: $FNR = \frac{FN}{TP+FN}$

False Positive Rate: $FPR = \frac{FP}{FP+TN}$

True Negative Rate: $TNR = \frac{TN}{FP+TN}$

Classification Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$

Misclassification Rate: $\frac{FP+FN}{TP+TN+FP+FN}$

Confusion Matrix ➡️ Performance Measures

	Positive Prediction	Negative Prediction
Positive Target Label	**TP=6**	**FN=3**
Negative Target Label	**FP=2**	**TN=9**

True Positive Rate: $TPR = \frac{TP}{TP+FN}$

False Negative Rate: $FNR = \frac{FN}{TP+FN}$

False Positive Rate: $FPR = \frac{FP}{FP+TN}$

True Negative Rate: $TNR = \frac{TN}{FP+TN}$

Recall: $recall = \frac{TP}{TP+FN} = TPR$

Precision: $precision = \frac{TP}{TP+FP}$

F_1: $F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$

Confusion Matrix ➡️ Performance Measures

	Positive Prediction	Negative Prediction
Positive Target Label	**TP=6**	**FN=3**
Negative Target Label	**FP=2**	**TN=9**

$TPR + FNR = 1$

True Positive Rate: $TPR = \frac{TP}{TP+FN} = \frac{6}{6+3} = \frac{2}{3}$

False Negative Rate: $FNR = \frac{FN}{TP+FN} = \frac{3}{6+3} = \frac{1}{3}$

False Positive Rate: $FPR = \frac{FP}{FP+TN} = \frac{2}{2+9} = \frac{2}{11}$

True Negative Rate: $TNR = \frac{TN}{FP+TN} = \frac{9}{2+9} = \frac{9}{11}$

$FPR + TNR = 1$

Classification Accuracy:

Misclassification Rate:

Recall:

Precision:

F_1:

Confusion Matrix ➡️ Performance Measures

	Positive Prediction	Negative Prediction
Positive Target Label	**TP=6**	**FN=3**
Negative Target Label	**FP=2**	**TN=9**

True Positive Rate: $TPR = \frac{TP}{TP+FN} = \frac{6}{6+3} = \frac{2}{3}$

False Negative Rate: $FNR = \frac{FN}{TP+FN} = \frac{3}{6+3} = \frac{1}{3}$

False Positive Rate: $FPR = \frac{FP}{FP+TN} = \frac{2}{2+9} = \frac{2}{11}$

True Negative Rate: $TNR = \frac{TN}{FP+TN} = \frac{9}{2+9} = \frac{9}{11}$

Classification Accuracy: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{6+9}{6+9+2+3} = \frac{15}{20}$

Misclassification Rate: $\frac{FP+FN}{TP+TN+FP+FN} = \frac{2+3}{6+9+2+3} = \frac{5}{20}$

$$\text{Classification Accuracy} + \text{Misclassification Rate} = 1$$

Recall:

Precision:

F_1:

Confusion Matrix ➡️ Performance Measures

	Positive Prediction	Negative Prediction
Positive Target Label	**TP=6**	**FN=3**
Negative Target Label	**FP=2**	**TN=9**

True Positive Rate: $TPR = \frac{TP}{TP+FN} = \frac{6}{6+3} = \frac{2}{3}$

False Negative Rate: $FNR = \frac{FN}{TP+FN} = \frac{3}{6+3} = \frac{1}{3}$

False Positive Rate: $FPR = \frac{FP}{FP+TN} = \frac{2}{2+9} = \frac{2}{11}$

True Negative Rate: $TNR = \frac{TN}{FP+TN} = \frac{9}{2+9} = \frac{9}{11}$

Classification Accuracy: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{6+9}{6+9+2+3} = \frac{15}{20}$

Misclassification Rate: $\frac{FP+FN}{TP+TN+FP+FN} = \frac{2+3}{6+9+2+3} = \frac{5}{20}$

Recall: $recall = \frac{TP}{TP+FN} = TPR = \frac{2}{3} \approx 0.67$

Precision: $precision = \frac{TP}{TP+FP} = \frac{6}{6+2} = \frac{3}{4} = 0.75$

F_1: $F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} = \frac{2 \cdot \frac{3}{4} \cdot \frac{2}{3}}{\frac{3}{4} + \frac{2}{3}} = \frac{12}{17} \approx 0.71$

TPS Exercise:

We have trained a predictive model using supervised learning and computed a confusion matrix based on predictions on a given set of data.

Question:
Which measure should we use to assess performance?

TPS Exercise:

We have trained a predictive model using supervised learning and computed a confusion matrix based on predictions on a given set of data.

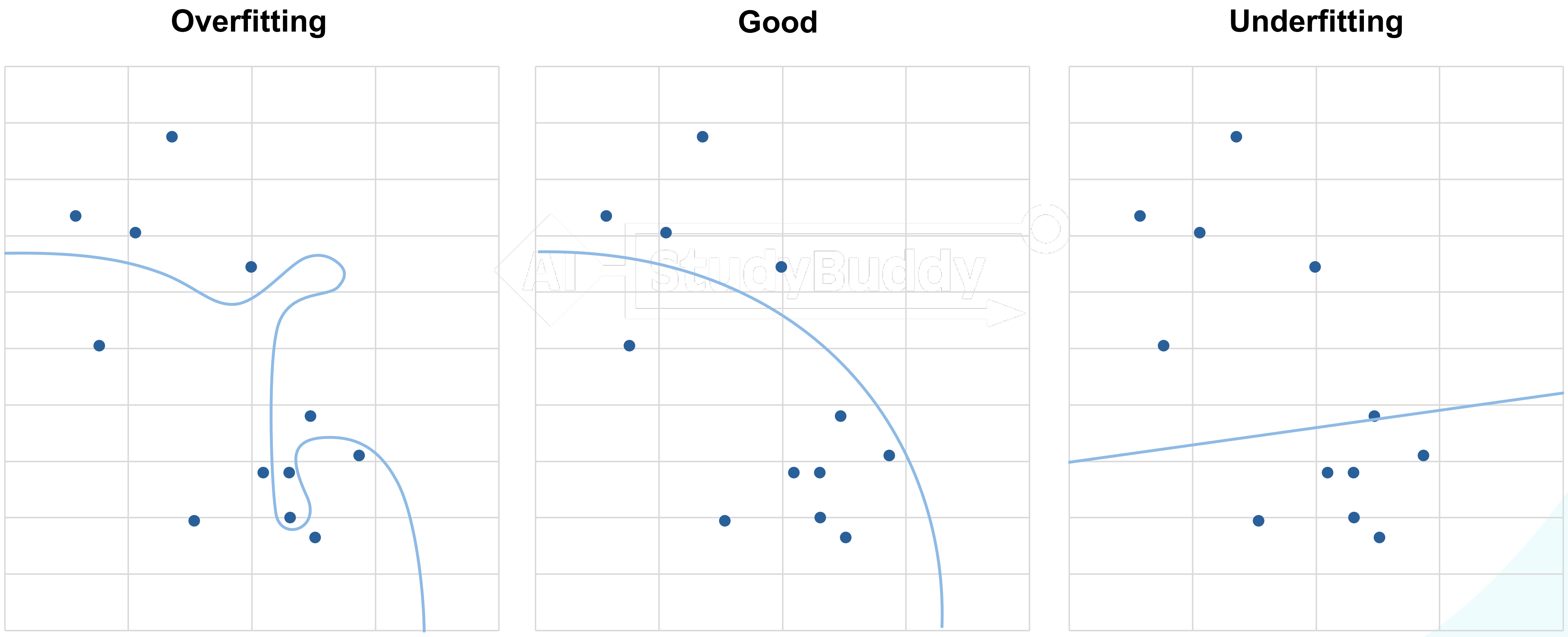
Question:
Which measure should we use to assess performance?

It depends ...
Often, a single measure is not enough.

What set of instances do we use as the basis for assessing performance?

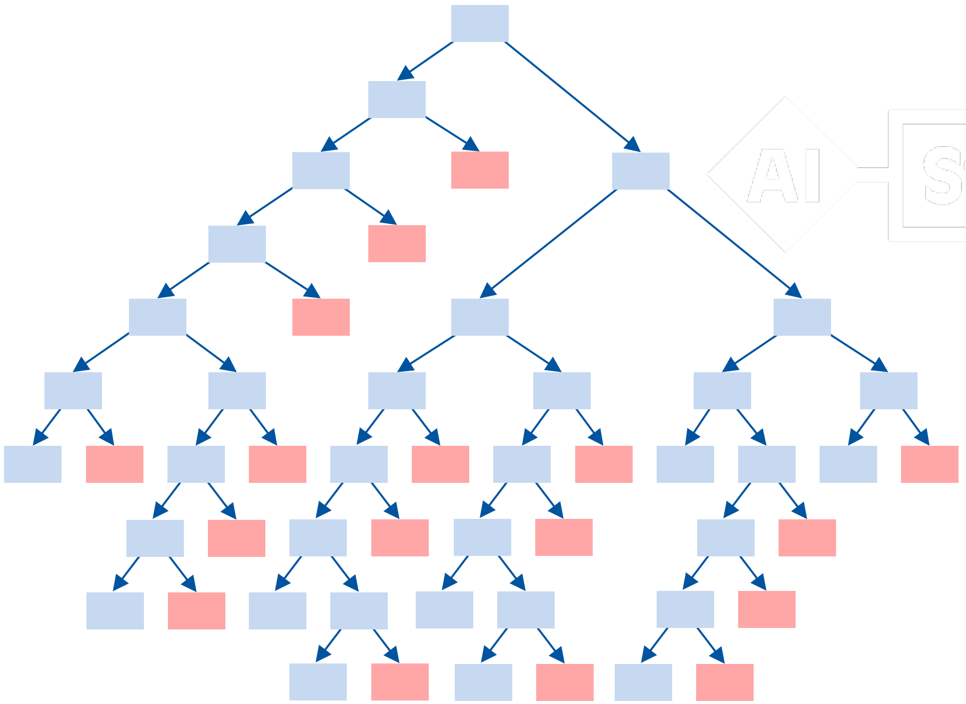
What set of instances do we use as the basis for assessing performance?

Let’s use training instances. What could go wrong?

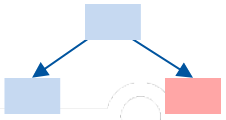

Overfitting
Good
Underfitting

Flight Classification (Running Example)

Overfitting



Underfitting

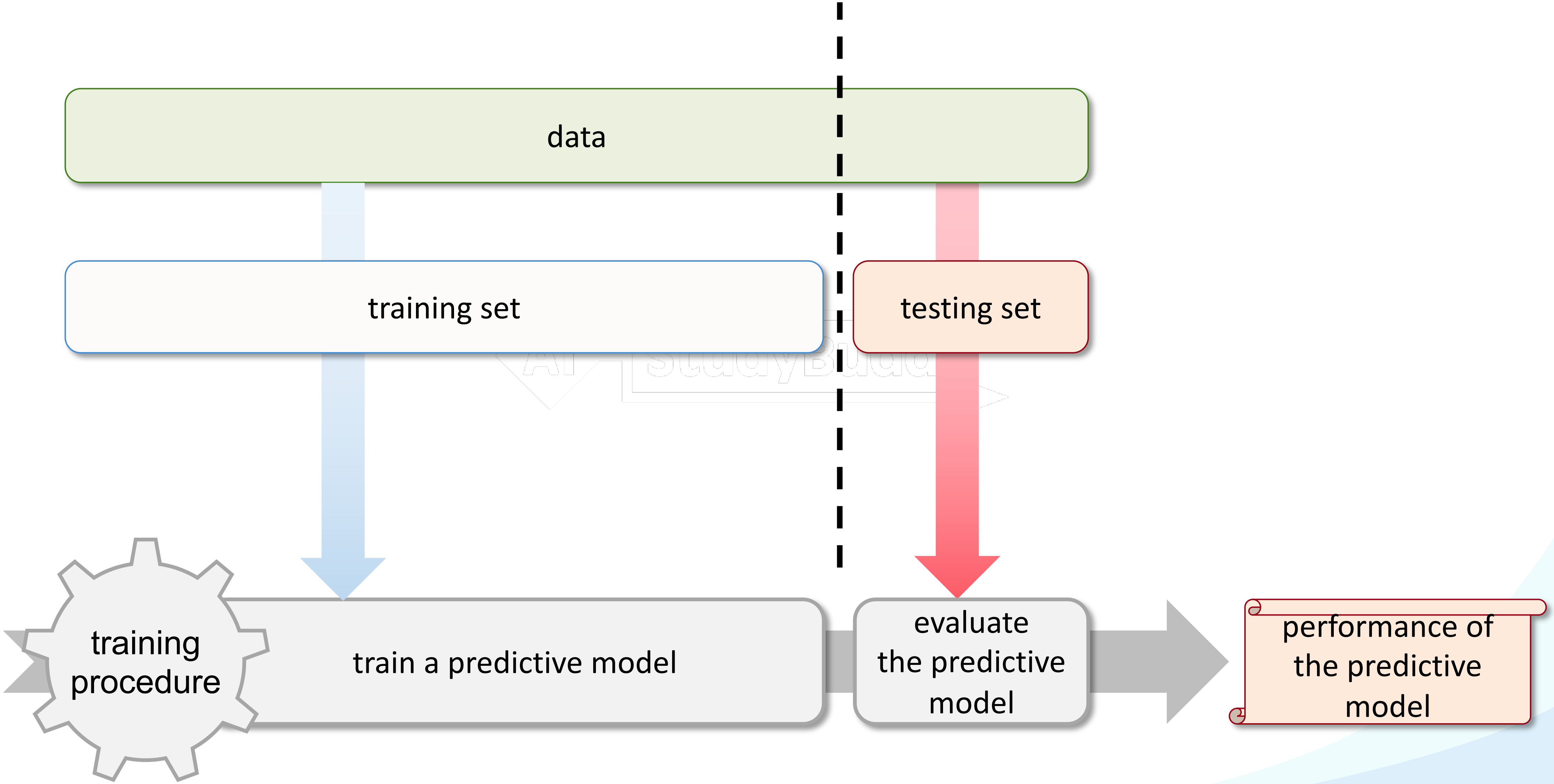


ID	Target
1	On Time
2	On Time
3	Delayed
4	On Time
5	Delayed
6	On Time
7	Delayed
8	On Time
9	On Time
10	On Time
11	Delayed
...	...

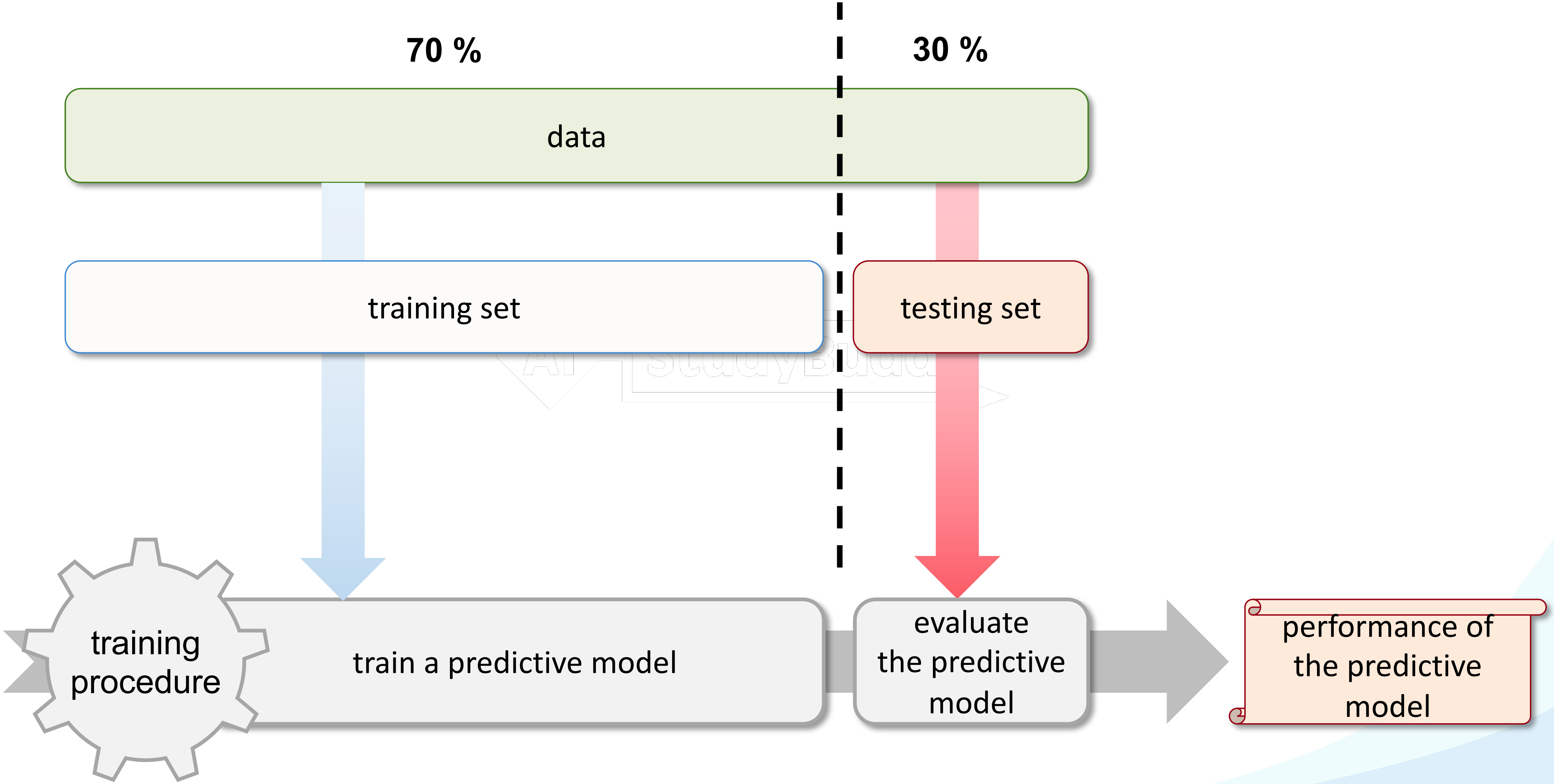
What set of instances do we use as the basis for assessing performance?

Key issue: generalisation to new data
➡️ don't assess performance based on training data!

Training & Testing Data

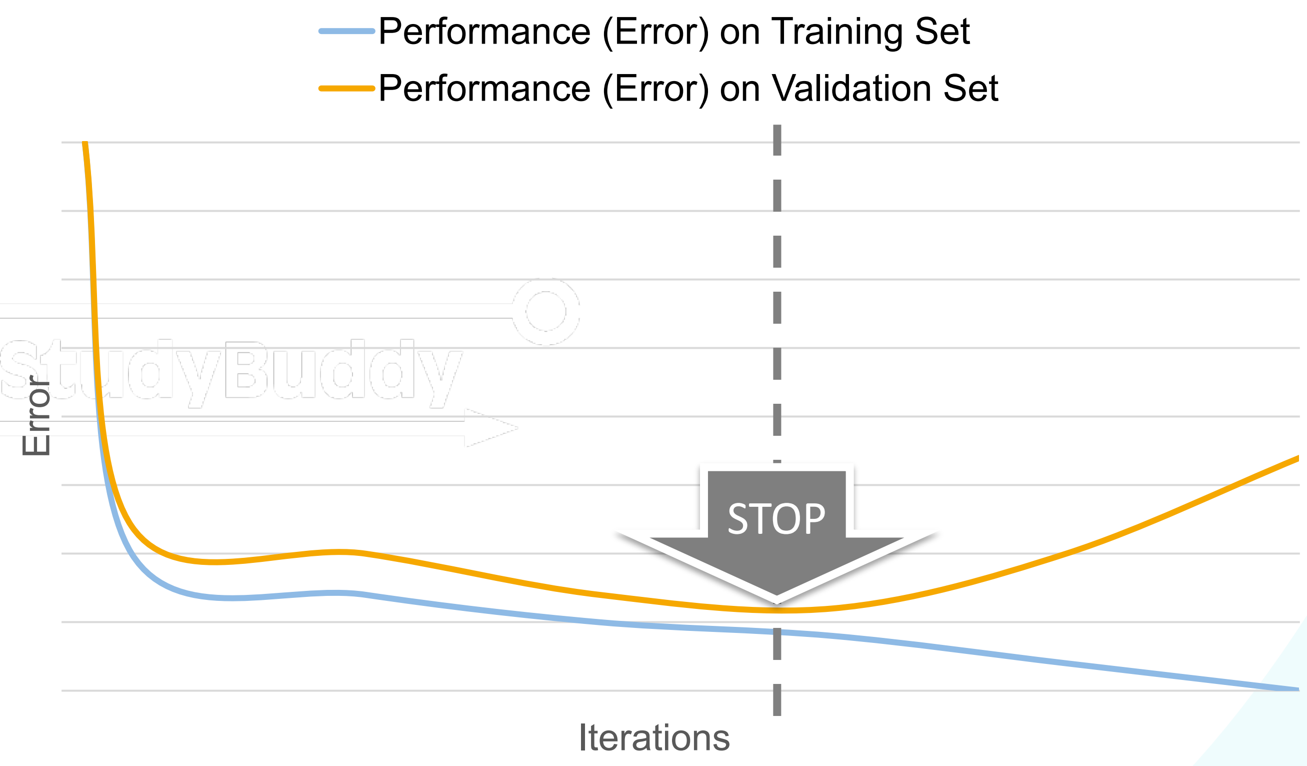


Training & Testing Data

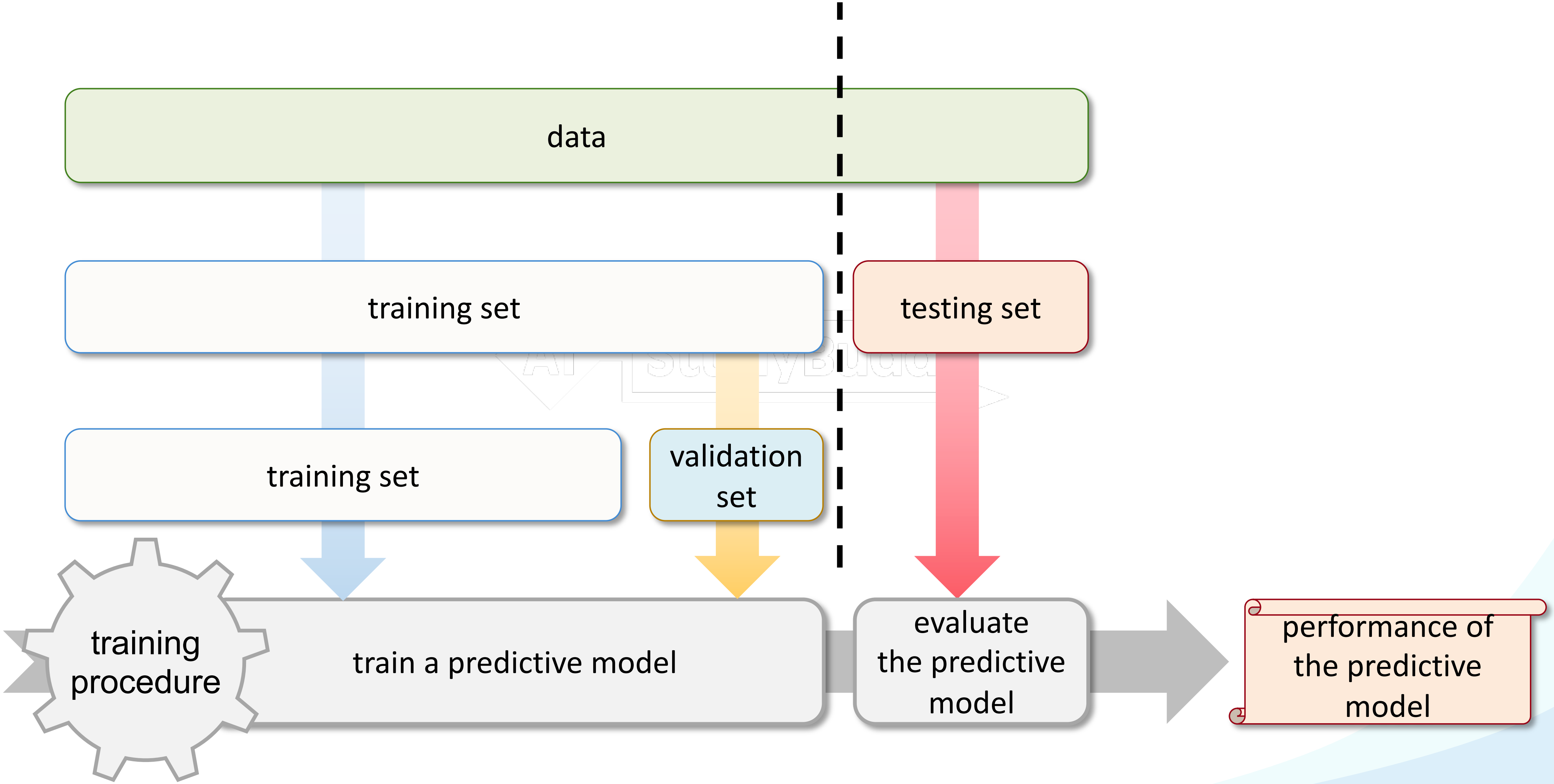


Validation Set

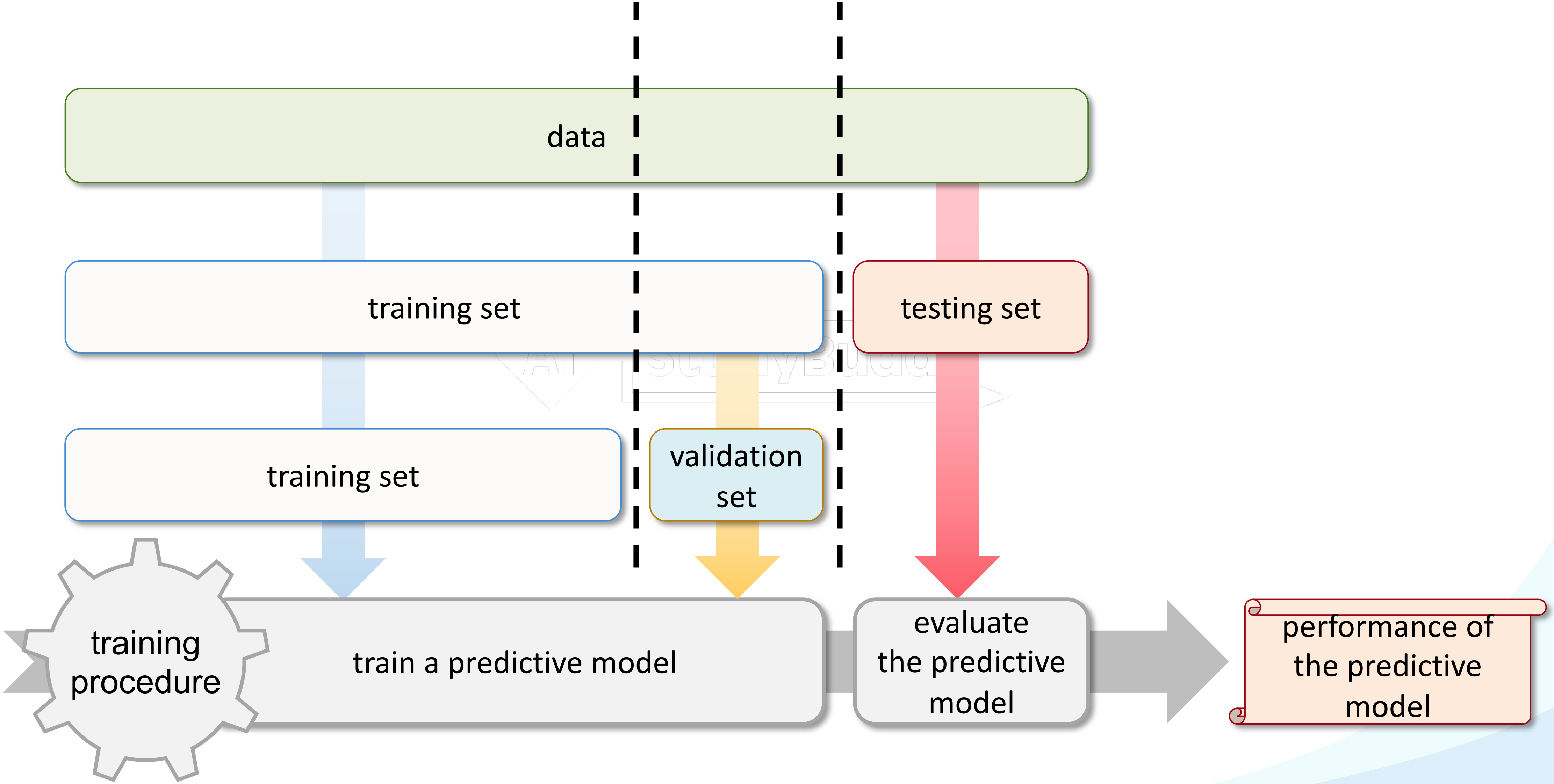
- Training a predictive model is often done iteratively (e.g., Regression, Neural Networks)
- The model is fitted closer and closer to the training data
- The validation set can be used to avoid overfitting the training data
- Often used for parameter selection or hyperparameter tuning



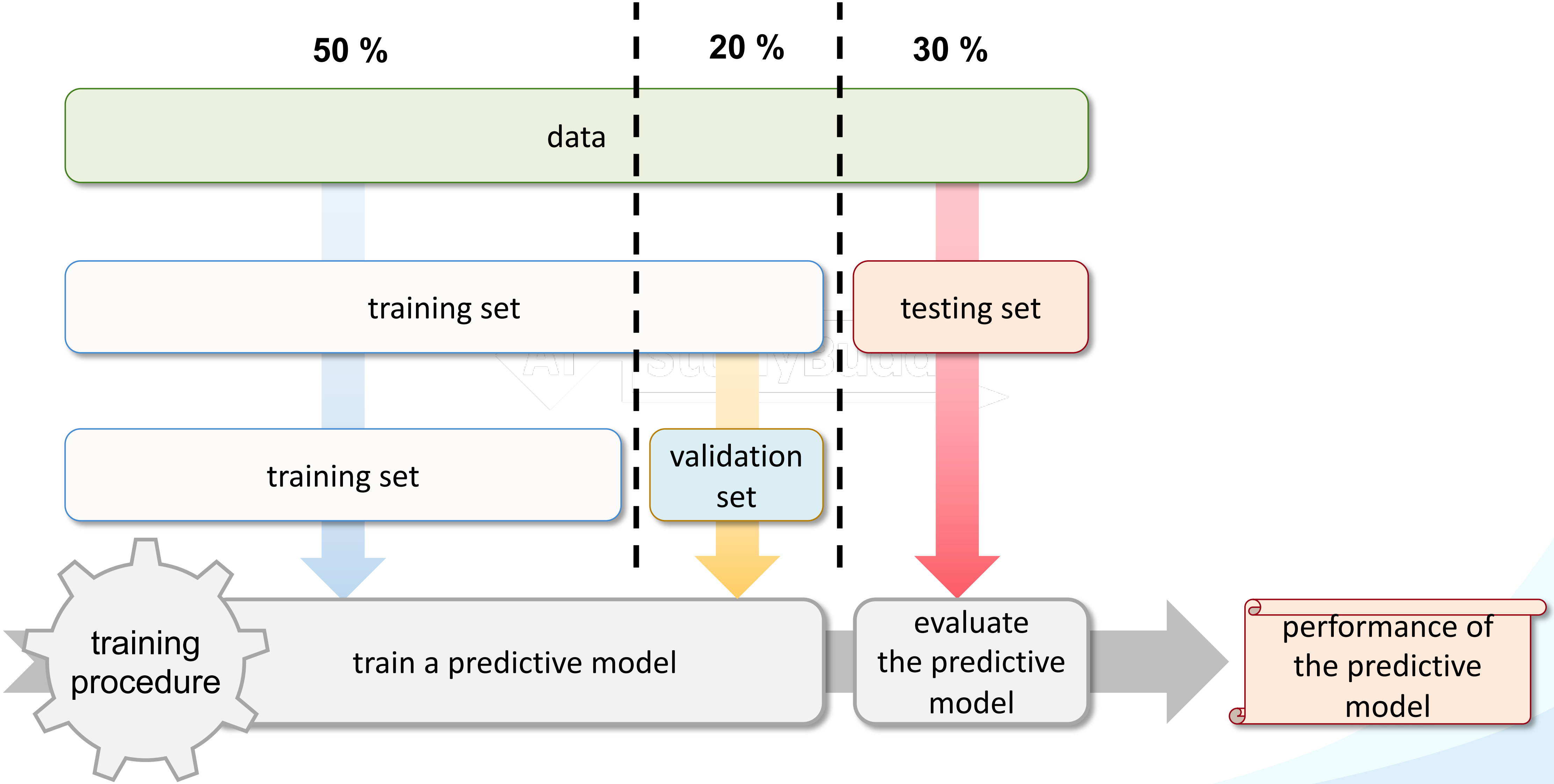
Training & Testing Data



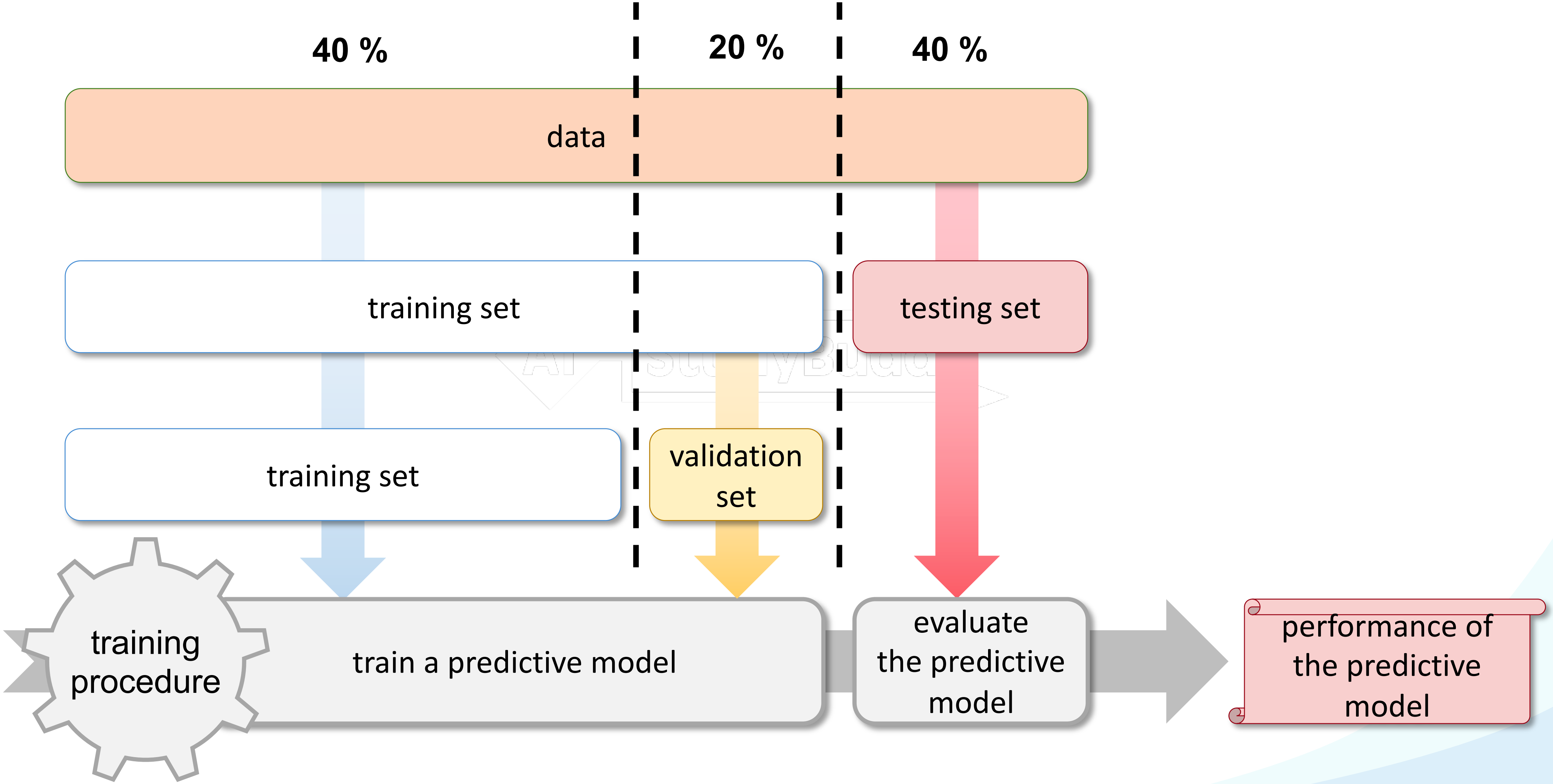
Training & Testing Data



Training & Testing Data



Training & Testing Data



How to split into training, validation and testing sets if there are only 20 instances?

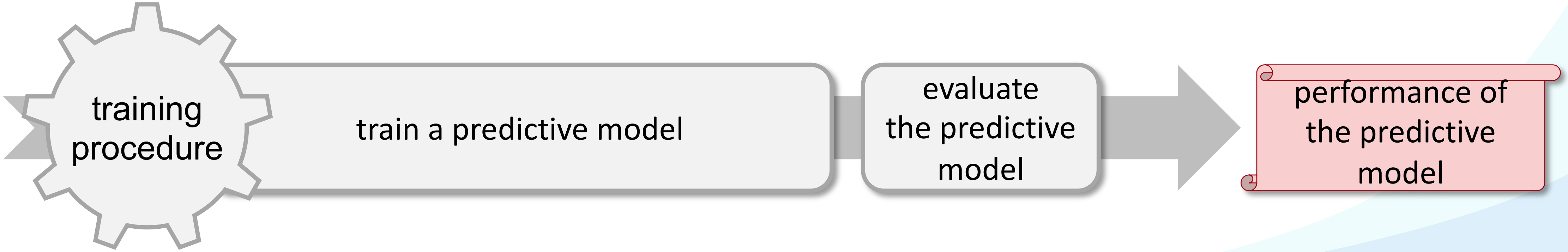
TPS Exercise:

You are using supervised learning to obtain a predictive model, training on a dataset with only 20 instances.

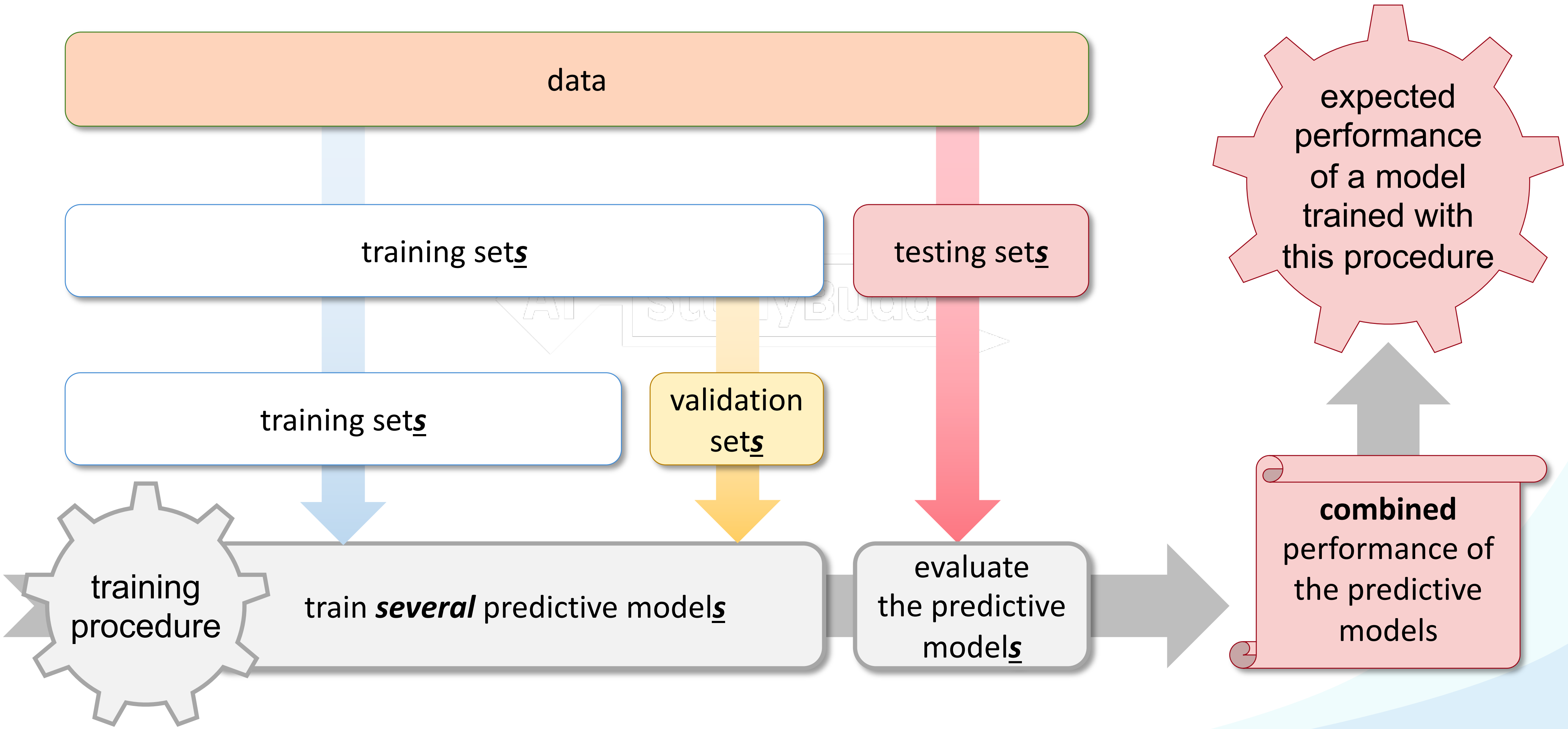
Question: How do you assess the quality of the model?

Dealing with small datasets:

- Splitting into **one** training and **one** testing set is reliable only for sufficiently large data sets
- On small data sets the training, validation or testing set become too small
- Small data set increases danger of a ‘lucky split’ (with most easy instances in the testing set)

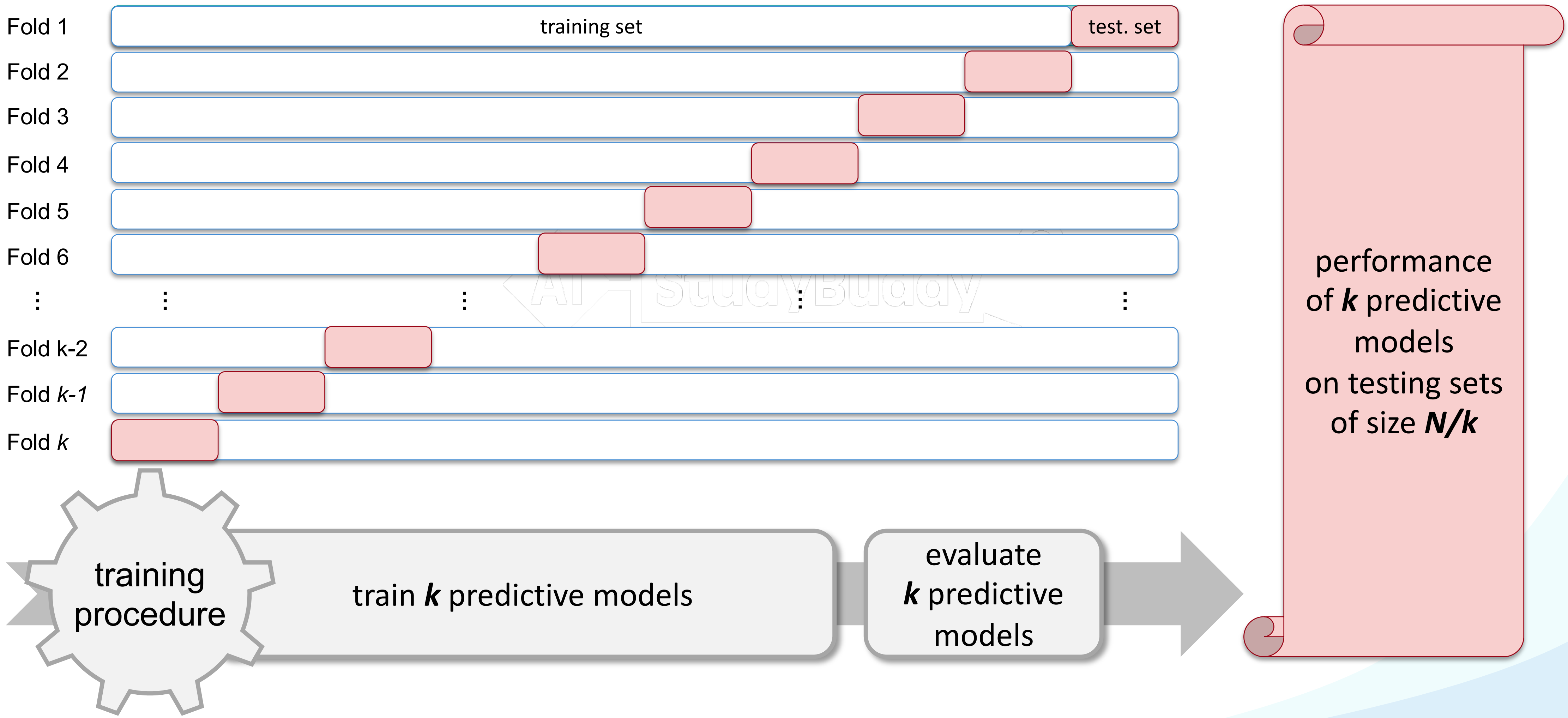


Motivation

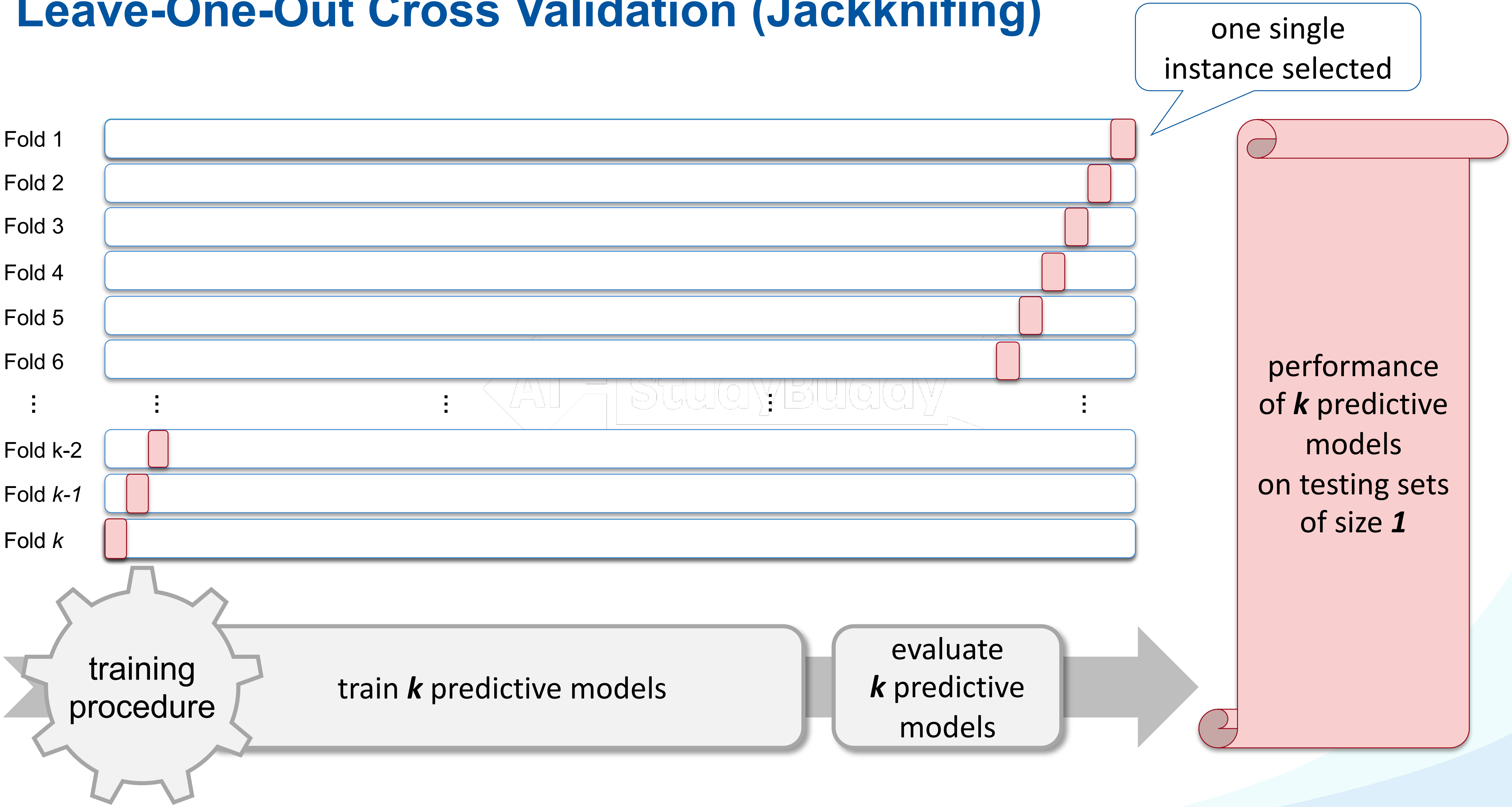


data
training sets
testing sets
training sets
validation sets
training procedure
train several predictive models
evaluate the predictive models
combined performance of the predictive models
expected performance of a model trained with this procedure

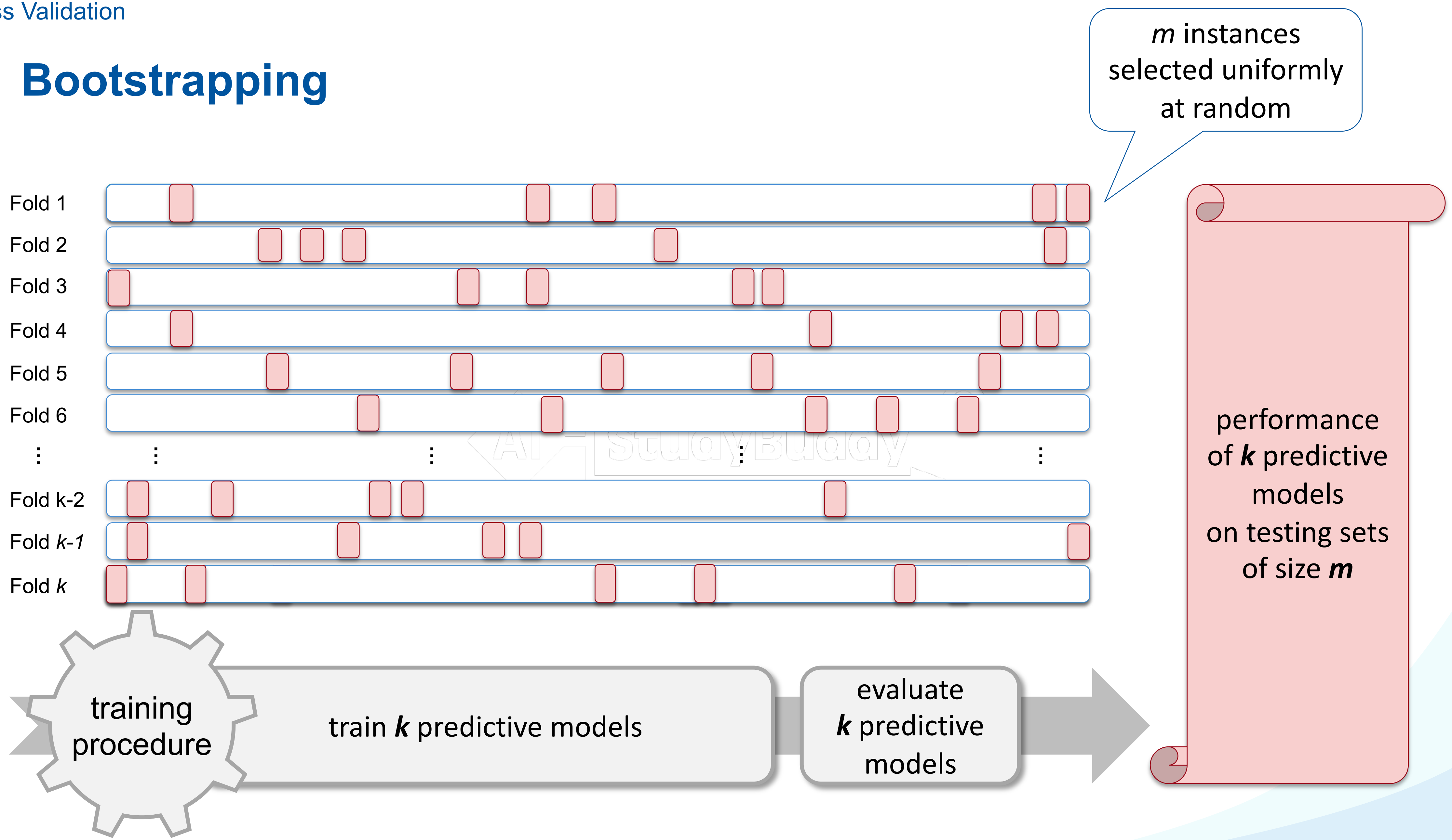
k-Fold Cross Validation



Leave-One-Out Cross Validation (Jackknifing)



Bootstrapping



What problem could arise?

ID	Target Label	Prediction	ID	Target Label	Prediction
1	On Time	On Time	11	On Time	On Time
2	On Time	On Time	12	On Time	On Time
3	On Time	On Time	13	On Time	On Time
4	On Time	On Time	14	On Time	On Time
5	On Time	On Time	15	On Time	On Time
6	On Time	On Time	16	On Time	On Time
7	On Time	On Time	17	On Time	On Time
8	On Time	On Time	18	On Time	On Time
9	On Time	On Time	19	Delayed	On Time
10	On Time	**On Time**	**20**	Delayed	**On Time**

Imbalanced Data

	On Time Prediction	Delayed Prediction
On Time Target Label	**18**	**0**
Delayed Target Label	**2**	**0**

Motivational Example

- A test set with many (18) positive instances and few (2) negative instances
- A model that always predicts positive

ID	Target Label	Prediction	ID	Target Label	Prediction
1	On Time	On Time	11	On Time	On Time
2	On Time	On Time	12	On Time	On Time
3	On Time	On Time	13	On Time	On Time
4	On Time	On Time	14	On Time	On Time
5	On Time	On Time	15	On Time	On Time
6	On Time	On Time	16	On Time	On Time
7	On Time	On Time	17	On Time	On Time
8	On Time	On Time	18	On Time	On Time
9	On Time	On Time	19	Delayed	On Time
10	On Time	**On Time**	**20**	Delayed	**On Time**

Imbalanced Data

	On Time Prediction	Delayed Prediction
On Time Target Label	**18**	**0**
Delayed Target Label	**2**	**0**

Motivational Example

- A test set with many (18) positive instances and few (2) negative instances
- A model that always predicts positive (= On Time)

Recall:

$$recall = \frac{TP}{TP+FN} = \frac{18}{18+0} = 1.0$$

Precision:

$$precision = \frac{TP}{TP+FP} = \frac{18}{18+2} = \frac{18}{20} = 0.9$$

ID	Target Label	Prediction	ID	Target Label	Prediction
1	On Time	On Time	11	On Time	On Time
2	On Time	On Time	12	On Time	On Time
3	On Time	On Time	13	On Time	On Time
4	On Time	On Time	14	On Time	On Time
5	On Time	On Time	15	On Time	On Time
6	On Time	On Time	16	On Time	On Time
7	On Time	On Time	17	On Time	On Time
8	On Time	On Time	18	On Time	On Time
9	On Time	On Time	19	Delayed	On Time
10	On Time	**On Time**	**20**	Delayed	**On Time**

Imbalanced Data

	On Time Prediction	Delayed Prediction
On Time Target Label	**18**	**0**
Delayed Target Label	**2**	**0**

Average Class Accuracy

Average recall over the elements in the set of possible target feature values $C = \{\text{Delayed}, \text{On Time}\}$

ID	Target Label	Prediction	ID	Target Label	Prediction
1	On Time	On Time	11	On Time	On Time
2	On Time	On Time	12	On Time	On Time
3	On Time	On Time	13	On Time	On Time
4	On Time	On Time	14	On Time	On Time
5	On Time	On Time	15	On Time	On Time
6	On Time	On Time	16	On Time	On Time
7	On Time	On Time	17	On Time	On Time
8	On Time	On Time	18	On Time	On Time
9	On Time	On Time	19	Delayed	On Time
10	On Time	**On Time**	**20**	Delayed	**On Time**

Imbalanced Data

	On Time Prediction	Delayed Prediction
On Time Target Label	**18**	**0**
Delayed Target Label	**2**	**0**

$$recall = \frac{TP}{TP+FN} = \frac{18}{18+0} = 1.0$$

	Delayed Prediction	On Time Prediction
Delayed Target Label	**0**	**2**
On Time Target Label	**0**	**18**

$$recall = \frac{TP}{TP+FN} = \frac{0}{0+2} = 0.0$$

Average Class Accuracy

Average recall over the elements in the set of possible target feature values $C = \{\text{Delayed, On Time}\}$

ID	Target Label	Prediction	ID	Target Label	Prediction
1	On Time	On Time	11	On Time	On Time
2	On Time	On Time	12	On Time	On Time
3	On Time	On Time	13	On Time	On Time
4	On Time	On Time	14	On Time	On Time
5	On Time	On Time	15	On Time	On Time
6	On Time	On Time	16	On Time	On Time
7	On Time	On Time	17	On Time	On Time
8	On Time	On Time	18	On Time	On Time
9	On Time	On Time	19	Delayed	On Time
10	On Time	**On Time**	**20**	Delayed	**On Time**

Imbalanced Data

	On Time Prediction	Delayed Prediction
On Time Target Label	**18**	**0**
Delayed Target Label	**2**	**0**

$$recall = \frac{TP}{TP+FN} = \frac{18}{18+0} = 1.0$$

	Delayed Prediction	On Time Prediction
Delayed Target Label	**0**	**2**
On Time Target Label	**0**	**18**

$$recall = \frac{TP}{TP+FN} = \frac{0}{0+2} = 0.0$$

Average Class Accuracy

Average recall over the elements in the set of possible target feature values $C = \{\text{Delayed}, \text{On Time}\}$

- arithmetic mean:
$$\frac{1}{|C|}\sum_{c \in C} recall_c$$
- harmonic mean:
$$\frac{1}{\frac{1}{|C|}\sum_{c \in C}\frac{1}{recall_c}}$$

ID	Target Label	Prediction	ID	Target Label	Prediction
1	On Time	On Time	11	On Time	On Time
2	On Time	On Time	12	On Time	On Time
3	On Time	On Time	13	On Time	On Time
4	On Time	On Time	14	On Time	On Time
5	On Time	On Time	15	On Time	On Time
6	On Time	On Time	16	On Time	On Time
7	On Time	On Time	17	On Time	On Time
8	On Time	On Time	18	On Time	On Time
9	On Time	On Time	19	Delayed	On Time
10	On Time	**On Time**	**20**	Delayed	**On Time**

Imbalanced Data

	On Time Prediction	Delayed Prediction
On Time Target Label	18	0
Delayed Target Label	2	0

$$recall = \frac{TP}{TP+FN} = \frac{18}{18+0} = 1.0$$

	Delayed Prediction	On Time Prediction
Delayed Target Label	0	2
On Time Target Label	0	18

$$recall = \frac{TP}{TP+FN} = \frac{0}{0+2} = 0.0$$

Average Class Accuracy

Average recall over the elements in the set of possible target feature values

$$C = \{\text{Delayed, On Time}\}$$

- arithmetic mean:

$$\frac{1}{|C|}\sum_{c \in C} recall_c = \frac{1}{2}(1+0) = 0.5$$

- harmonic mean:

$$\frac{1}{\frac{1}{|C|}\sum_{c \in C} \frac{1}{recall_c}} = \frac{1}{\frac{1}{2}(\frac{1}{1}+\frac{1}{0})} = 0.0$$

$\frac{1}{0} = \infty$ in the limit

ID	Target Label	Prediction	ID	Target Label	Prediction
1	On Time	On Time	11	On Time	On Time
2	On Time	On Time	12	On Time	On Time
3	On Time	On Time	13	On Time	On Time
4	On Time	On Time	14	On Time	On Time
5	On Time	On Time	15	On Time	On Time
6	On Time	On Time	16	On Time	On Time
7	On Time	On Time	17	On Time	On Time
8	On Time	On Time	18	On Time	On Time
9	On Time	On Time	19	Delayed	On Time
10	On Time	**On Time**	**20**	Delayed	**On Time**

What’s worse:

Predicting a flight to be delayed and having it arrive on time, or predicting it to be on time and find it to be delayed?

- Does the self-driving car need to stop?
- Should the patient be tested for a severe disease?

➡️ **FP**s and **FN**s can have (very) different cost!

[2]

Profit Matrix

Example Flight Classification

- Correctly inform customers about a delay:
  - Customers can plan to arrive later
  - ***A little*** 'profit' from less unhappy customers
- Incorrectly inform customers about a delay:
  - Customers arrive too late
  - ***Huge*** loss of 'profit' by unnecessarily delayed flight

- Incorrectly predicting 'Delayed' (FN) costs more than incorrectly predicting 'On Time' (FP)

Profit Matrix		Prediction: On Time	Prediction: Delay
Target Label	On Time	**0**	**-80**
	Delay	**-10**	**20**

Profit (Utility) Matrix

M_1

		Prediction	
		On Time	Delay
Target Label	On Time	6	3
	Delay	2	9

M_2

		Prediction	
		On Time	Delay
Target Label	On Time	5	0
	Delay	9	6

Profit Matrix

		Prediction	
		On Time	Delay
Target Label	On Time	0	-80
	Delay	-10	20

Profit Matrix

M_1

		Prediction: On Time	Prediction: Delay
Target Label	On Time	6	3
	Delay	2	9

M_2

		Prediction: On Time	Prediction: Delay
Target Label	On Time	5	0
	Delay	9	6

Profit Matrix

		Prediction: On Time	Prediction: Delay
Target Label	On Time	0	-80
	Delay	-10	20

M_1

		Prediction: On Time	Prediction: Delay
Target Label	On Time	0	-240
	Delay	-20	180
	Profit	-80	

M_2

		Prediction: On Time	Prediction: Delay
Target Label	On Time	0	0
	Delay	-90	120
	Profit	30	

Profit Matrix

M_1

Target Label \ Prediction	On Time	Delay
On Time	6	3
Delay	2	9

M_2

Target Label \ Prediction	On Time	Delay
On Time	5	0
Delay	9	6

M_1

Target Label \ Prediction	On Time	Delay
On Time	0	-240
Delay	-20	180
Profit	-80	

M_2

Target Label \ Prediction	On Time	Delay
On Time	0	0
Delay	-90	120
Profit	30	

Profit Matrix

Target Label \ Prediction	On Time	Delay
On Time	0	-80
Delay	-10	20

$$\begin{aligned} \textit{profit} &= \mathbf{FP} \cdot \mathbf{FP}_{\text{profit}} + \mathbf{TP} \cdot \mathbf{TP}_{\text{profit}} \\ &+ \mathbf{FN} \cdot \mathbf{FN}_{\text{profit}} + \mathbf{TN} \cdot \mathbf{TN}_{\text{profit}} \end{aligned}$$

Key concepts covered today:

- confusion matrix
- performance measures for binary classification
- training, testing and validation sets
- *k*-fold cross validation
- leave-one-out cross validation (jackknife)
- bootstrap sampling validation
- imbalanced data, average class accuracy
- profit (utility) matrix

Preparation for Tuesday:

Investigate the following questions:

- **How to assess predictive models for multi-class classification?**
  (> 2 target classes, *e.g.*, on time, mildly delayed, severely delayed)

- **How to assess predictive models for regression tasks?**
  (predictions = numbers, *e.g.*, minutes of delay)

(We will use this for TPS exercises with the T part done before class.)

Sources

- [1] Erik Heddema on Unsplash, Unsplash License, (https://unsplash.com/de/fotos/k_kz0jmyOmE)
- [2] Matt C on Unsplash, Unsplash License, (https://unsplash.com/de/fotos/ubHRHM37ddE)