Evaluating Machine Learning Models

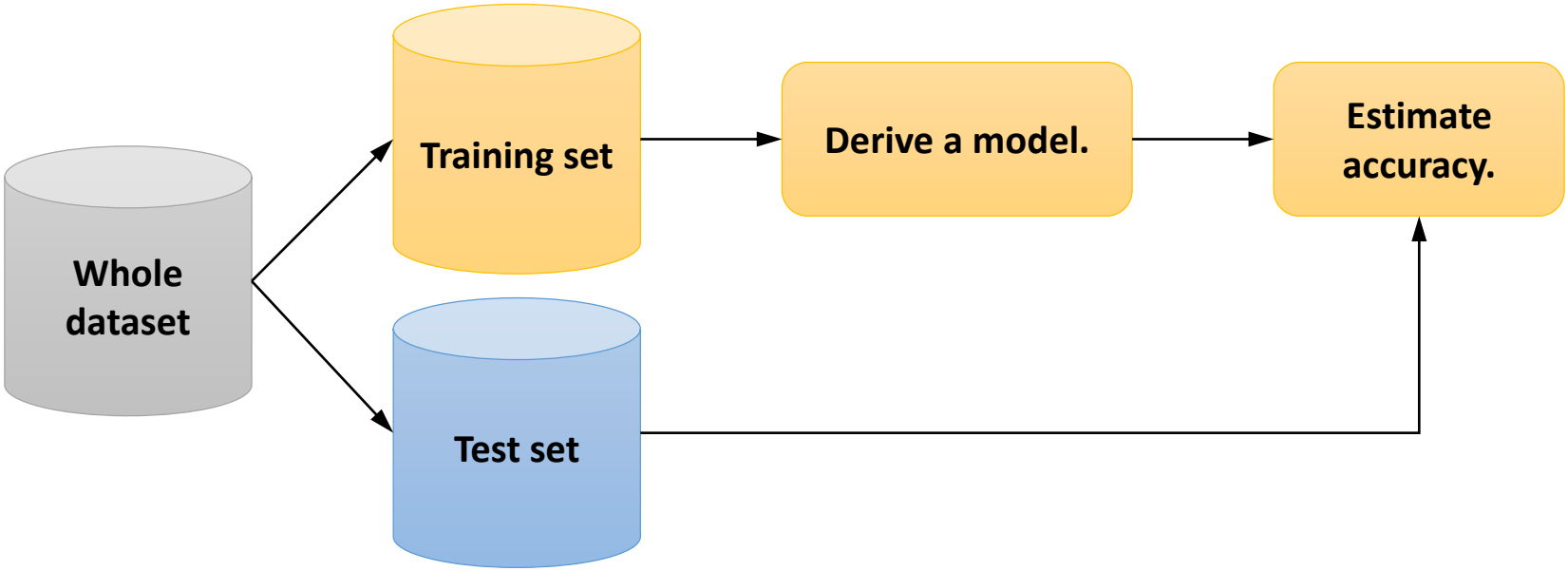
Data Intelligence and Learning (DIAL) Lab

Prof. Jongwuk Lee

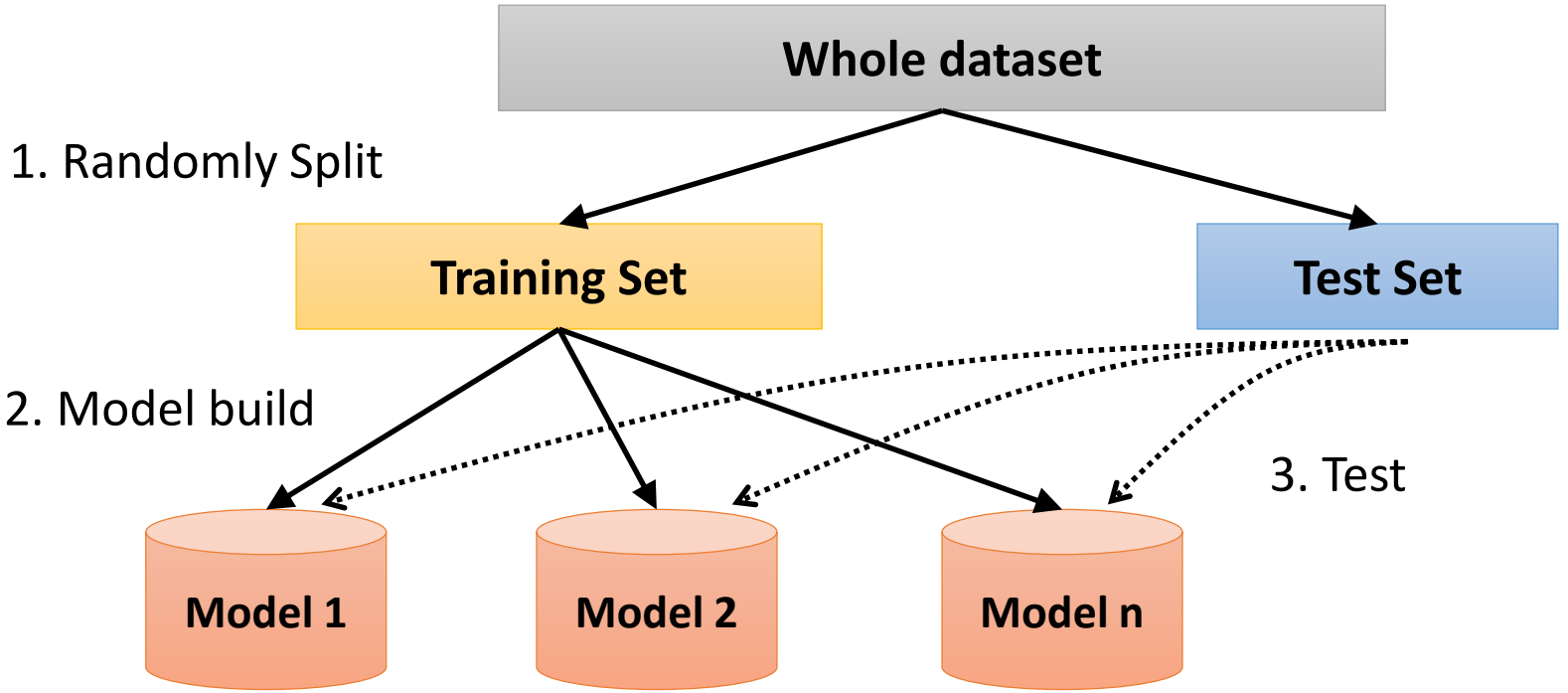
Cross Validation

Hold-out Method

- **Divide the given data into a training set and a test set.**
  - The training set and the test set **should NOT overlap** each other.
- **How to choose a good model?**
  - With the training set, build various models.
  - With the test set, evaluate each model.
  - Choose the model which shows the best performance with test set.



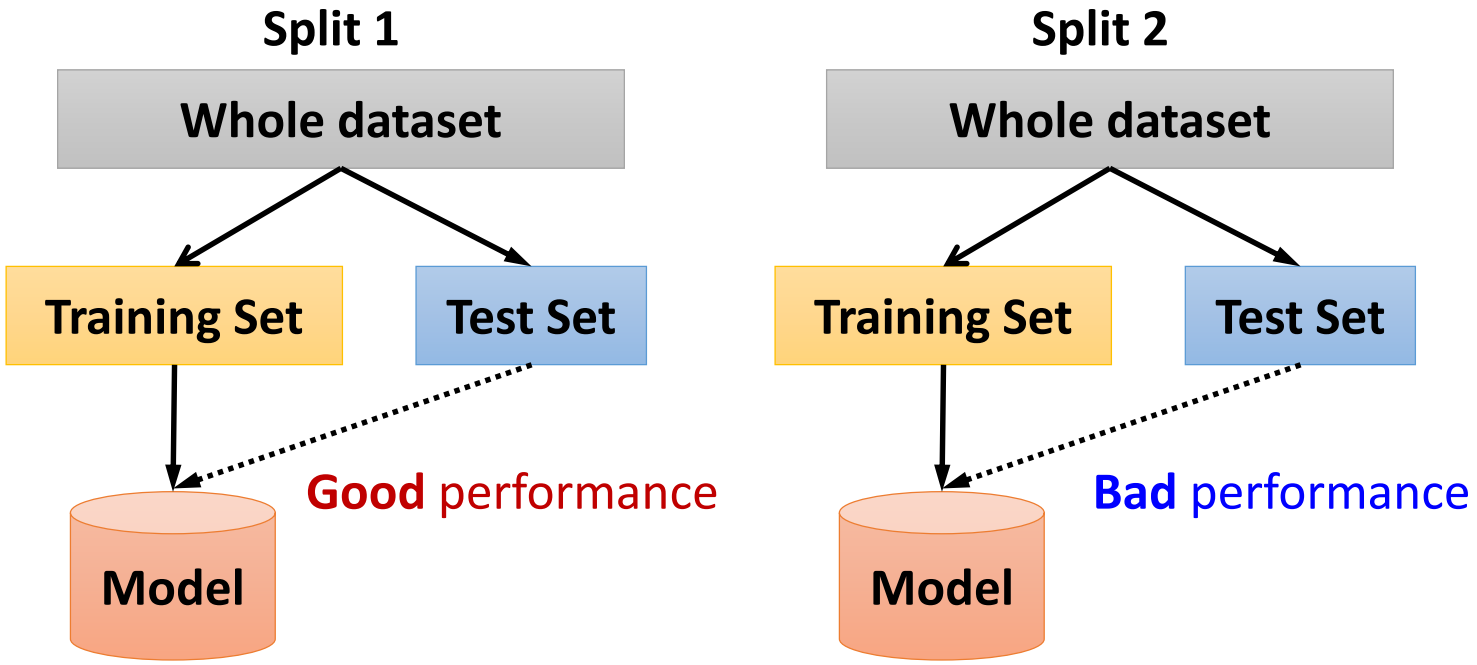
Hold-out Method



4. Choose the best model.

Hold-out Method

- **Advantage: Simple and easy**
- **Disadvantage**
  - **Waste of data**: The test set is not used for modeling building.
  - **Random split**: Evaluation can be different depending on data split.



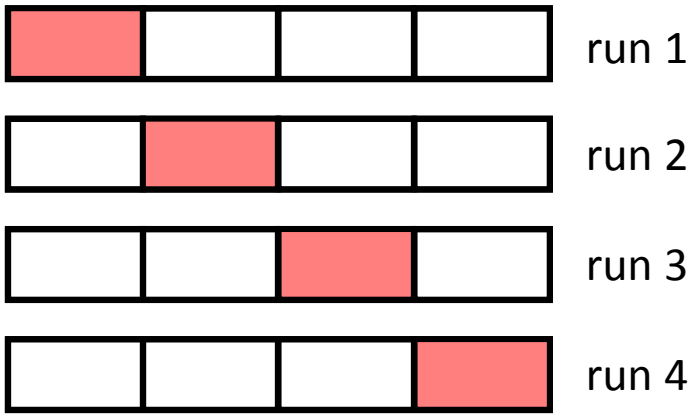
Cross Validation

Cross-validation (k-fold)

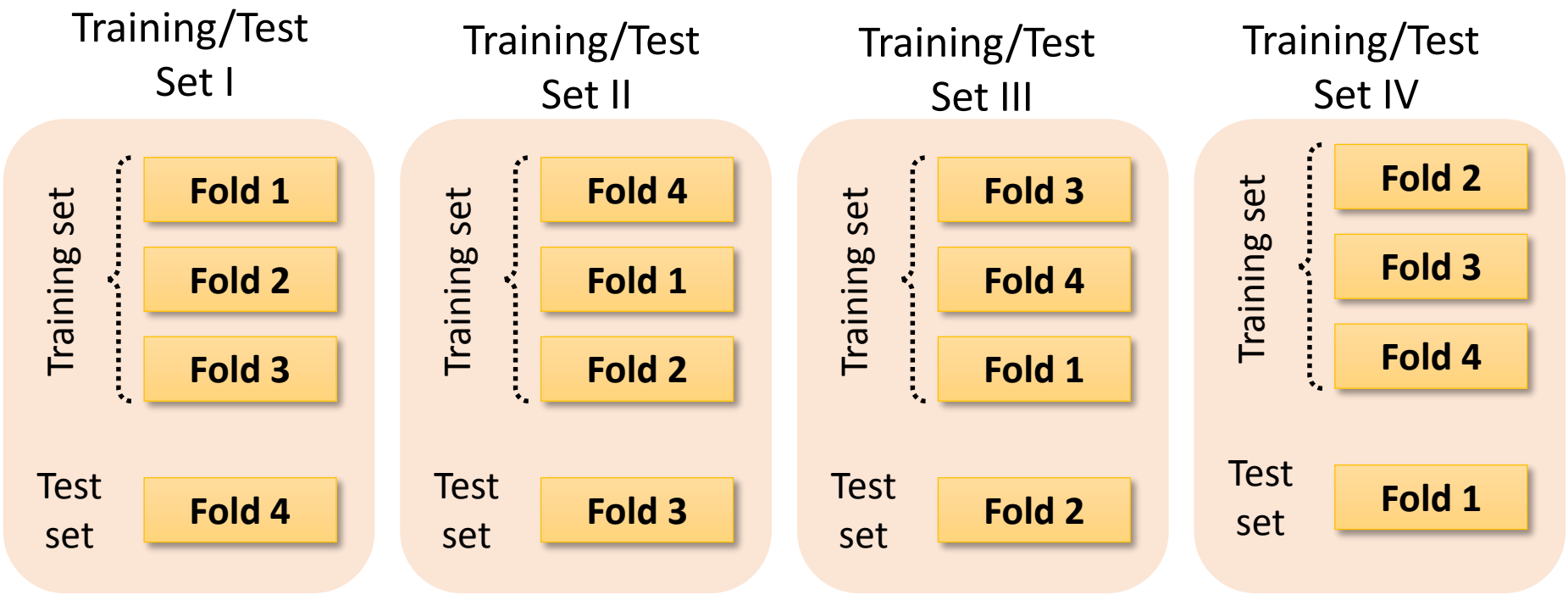
- Data D is randomly partitioned into **k mutually exclusive subsets** $\{D_1, \dots, D_k\}$, each approximately equal size.
- $k = 10$ is most popular.

Overall procedure

- The data is partitioned into k groups.
  - $k - 1$ of the groups are used for training the model.
  - One remaining group is used for evaluating the model.
- Repeat procedure for all k choices.
- Performance from the k runs are averaged.

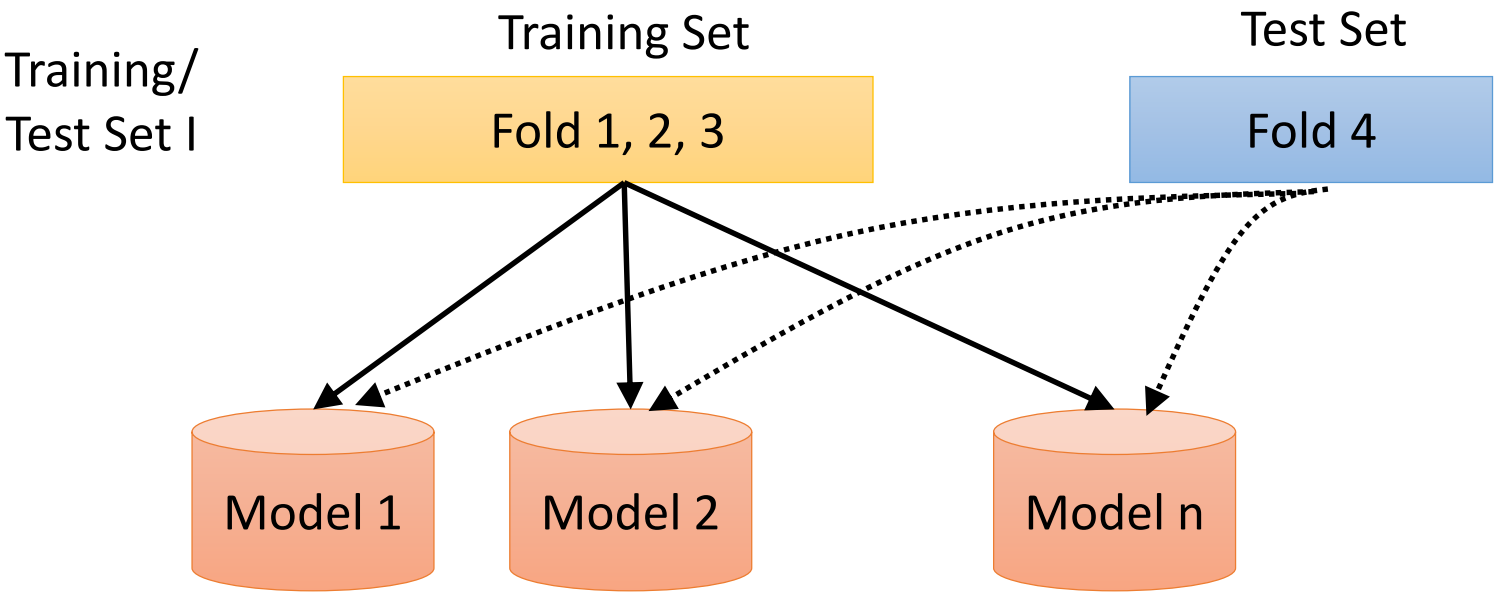


Example: 4-Fold Cross Validation



Choose a model by the average performance of four sets.

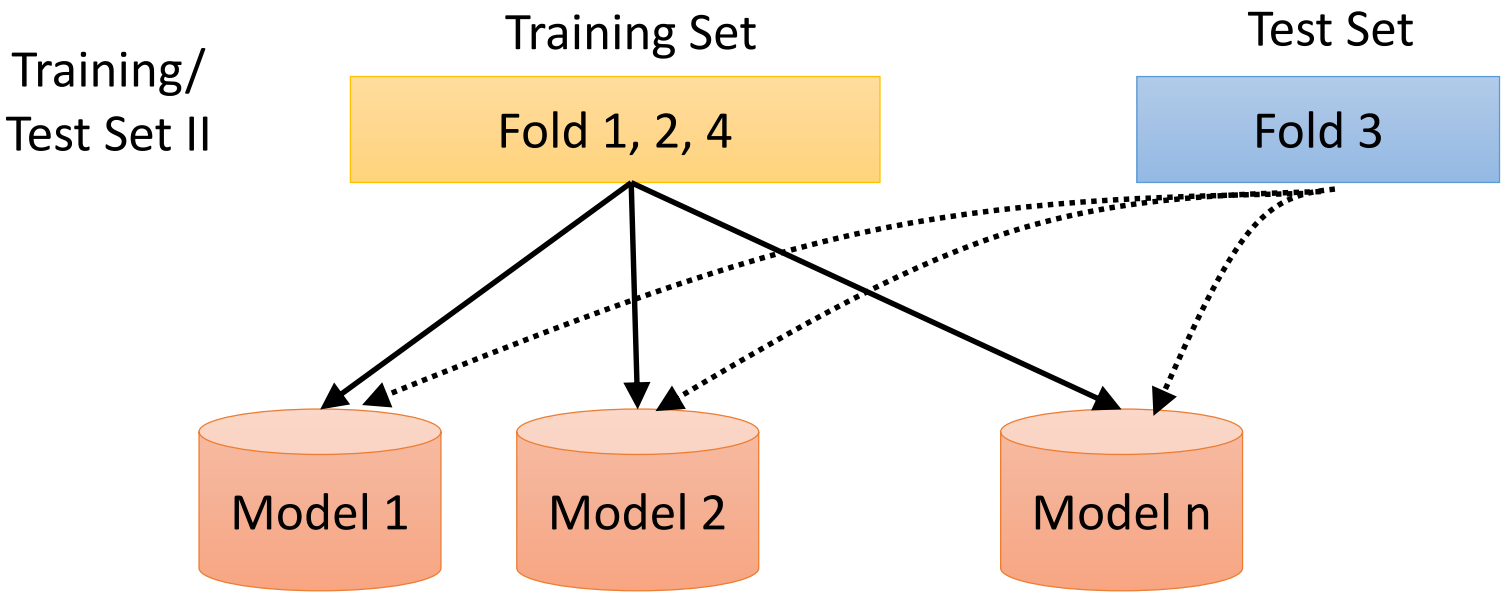
Example: 4-Fold Cross Validation



Performance

Set I	0.75	0.80	…	0.72
Set II				
Set III			…	
Set IV			…	

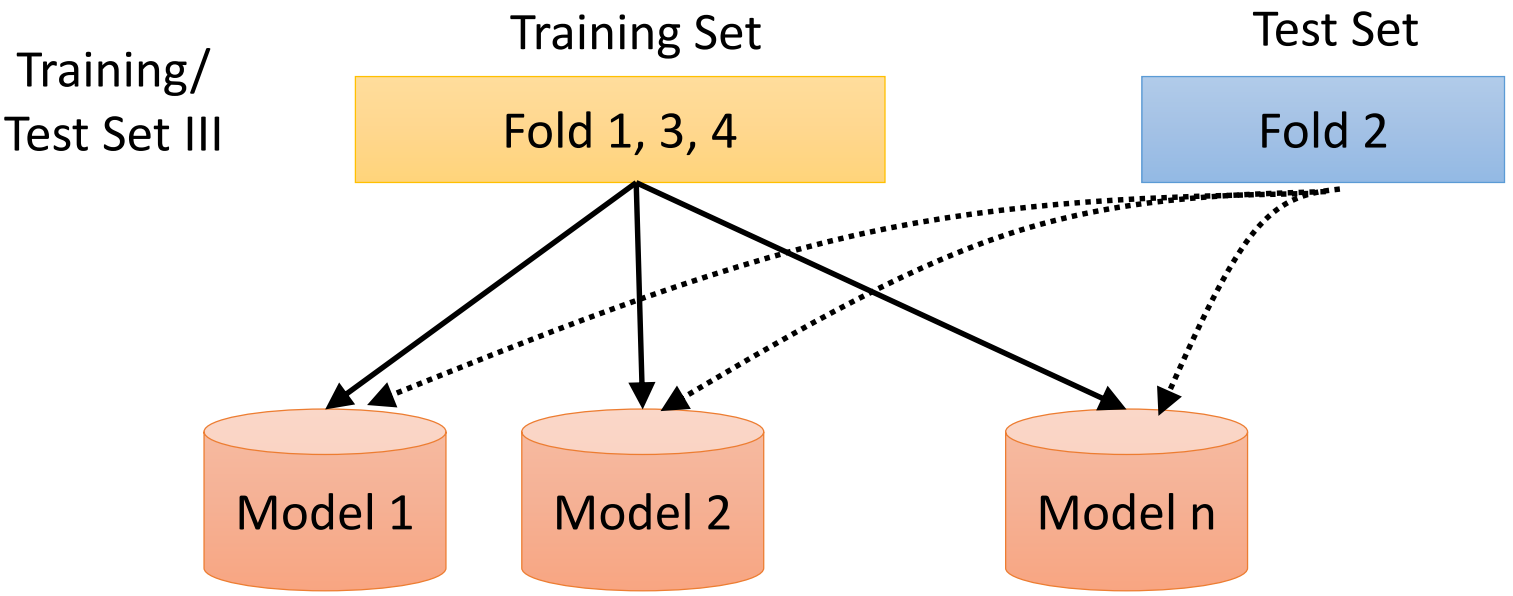
Example: 4-Fold Cross Validation



Performance

Set I	0.75	0.80	…	0.72
Set II	0.80	0.85	…	0.65
Set III			…	
Set IV			…	

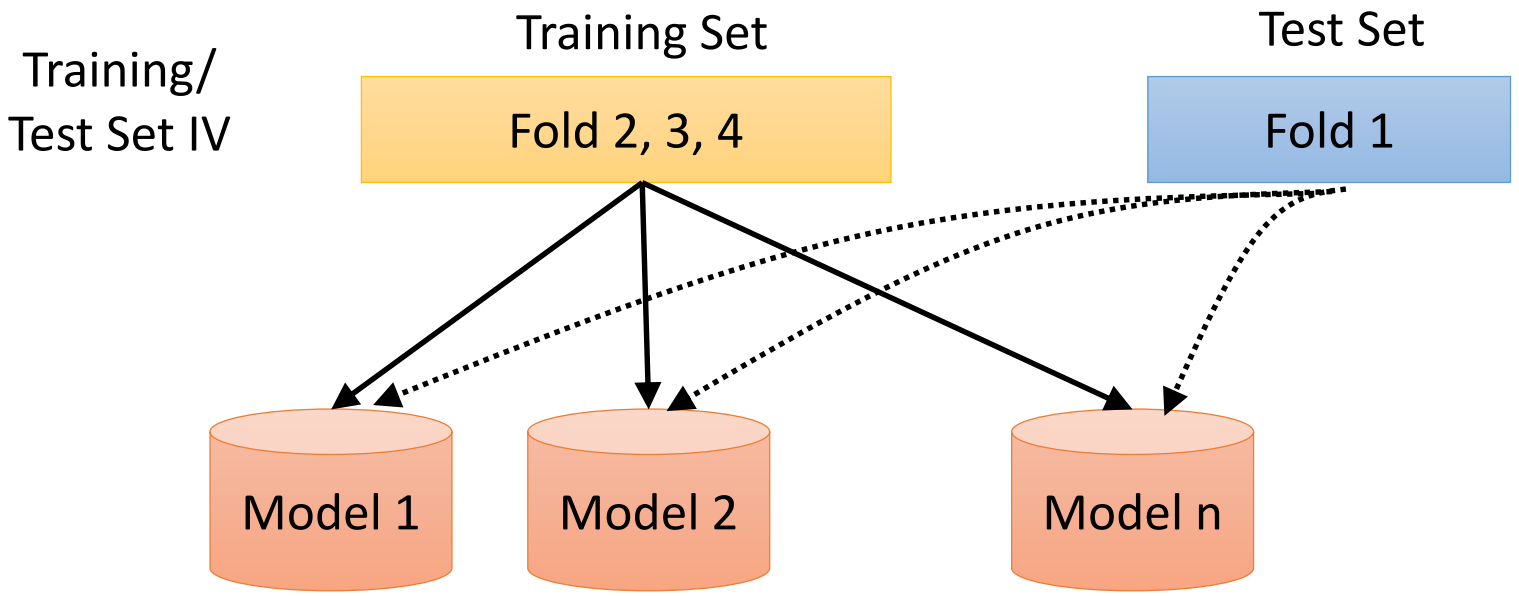
Example: 4-Fold Cross Validation



Performance

Set I	0.75	0.80	…	0.72
Set II	0.80	0.85	…	0.65
Set III	0.72	0.70	…	0.75
Set IV			…	

Example: 4-Fold Cross Validation



Performance

Set I	0.75	0.80	…	0.72
Set II	0.80	0.85	…	0.65
Set III	0.72	0.70	…	0.75
Set IV	0.75	0.69	…	0.72

Cross Validation

Summary

- The data set is divided into k subsets, and the hold-out method is repeated k times.
- Each time, one of the k subsets is used as the test set and the other $k-1$ subsets are put together to form a training set.
- The average error across all k trials is computed.
- The variance is reduced as k is increased.

Advantage

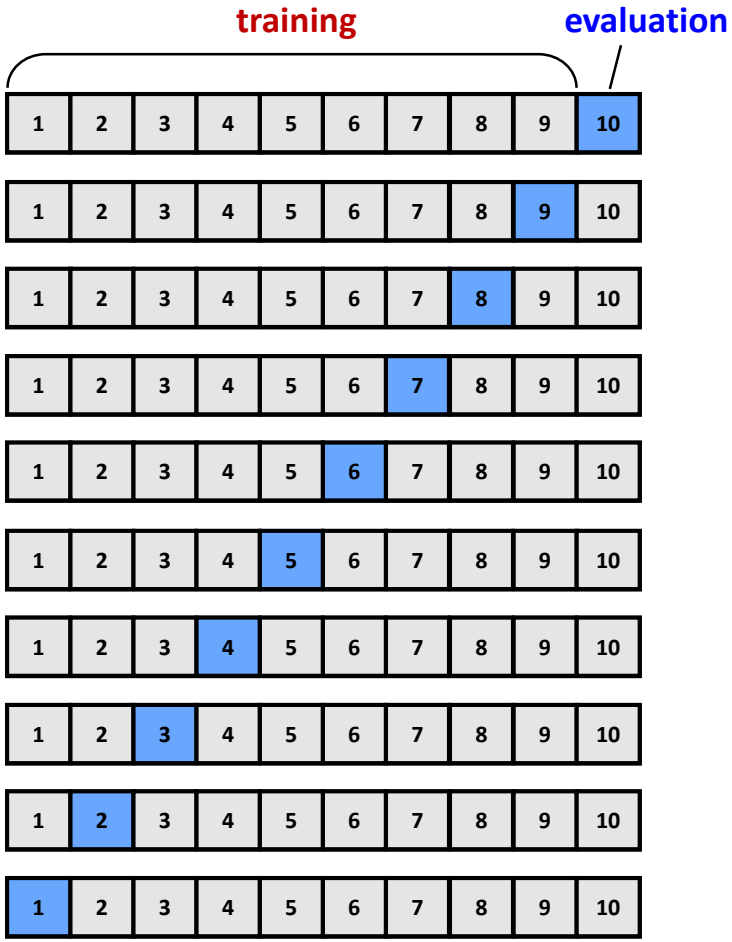
- Less dependent on how the data gets divided.
- Every data point gets to be in a test set exactly once and gets to be in a training set $k-1$ times.

Disadvantage

- Time!

Leave-one-out Cross Validation

- **Extreme case of k-fold cross validation**
  - **If data size is n, set $k = n$.**
  - Every data samples except one is used for training and the remaining one is used for testing.
  - Repeat this n times.



Evaluation Metrics

Evaluation in Regression Models

➢ Mean absolute error (MAE) and mean squared error (MSE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|f\left(\mathbf{x}^{(i)}\right) - y^{(i)}\right|$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\left|f\left(\mathbf{x}^{(i)}\right) - y^{(i)}\right|\right)^2$$

➢ Root mean squared error (RMSE)

- MSE is more popular than MAE because MSE **punishes** larger errors.
- But, RMSE is even **more popular** than MSE because RMSE is interpretable in the "y" axis.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\left|f\left(\mathbf{x}^{(i)}\right) - y^{(i)}\right|\right)^2}$$

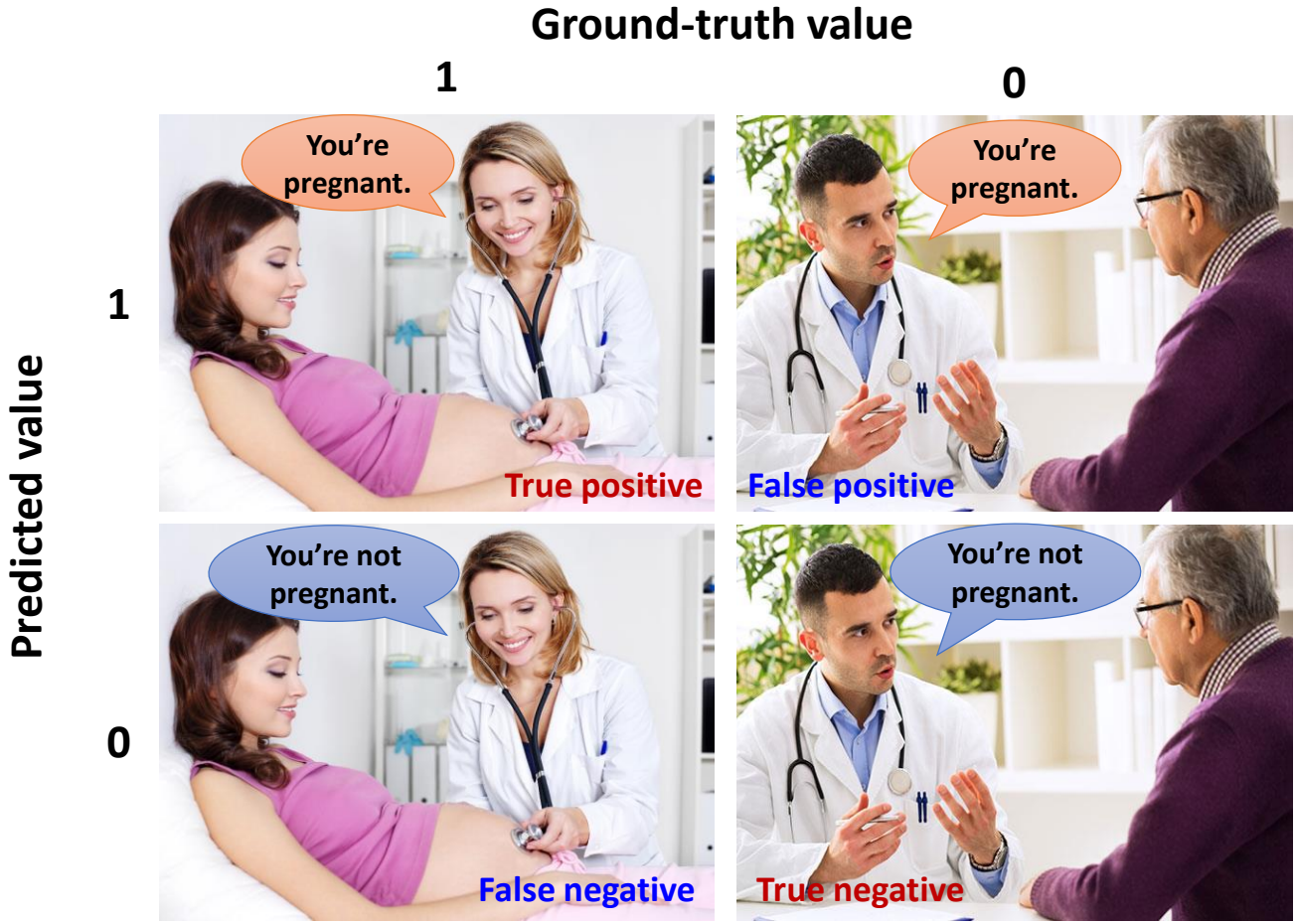
Confusion Matrix

➢ Building a confusion matrix for actual and predicted results

		Ground-truth value: Positive (1)	Ground-truth value: Negative (0)
Predicted value	Positive (1)	True Positive (TP)	False Positive (FP)
Predicted value	Negative (0)	False Negative (FN)	True Negative (TN)

- This can be used to measure **accuracy**, **precision**, **recall**, and so on.

Confusion Matrix



Accuracy, Error Rates

- **The fraction of these classifications that are correct**
  - Given an image, classify it into “cat” or “No cat”.

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$error\ rate = 1 - accuracy$$

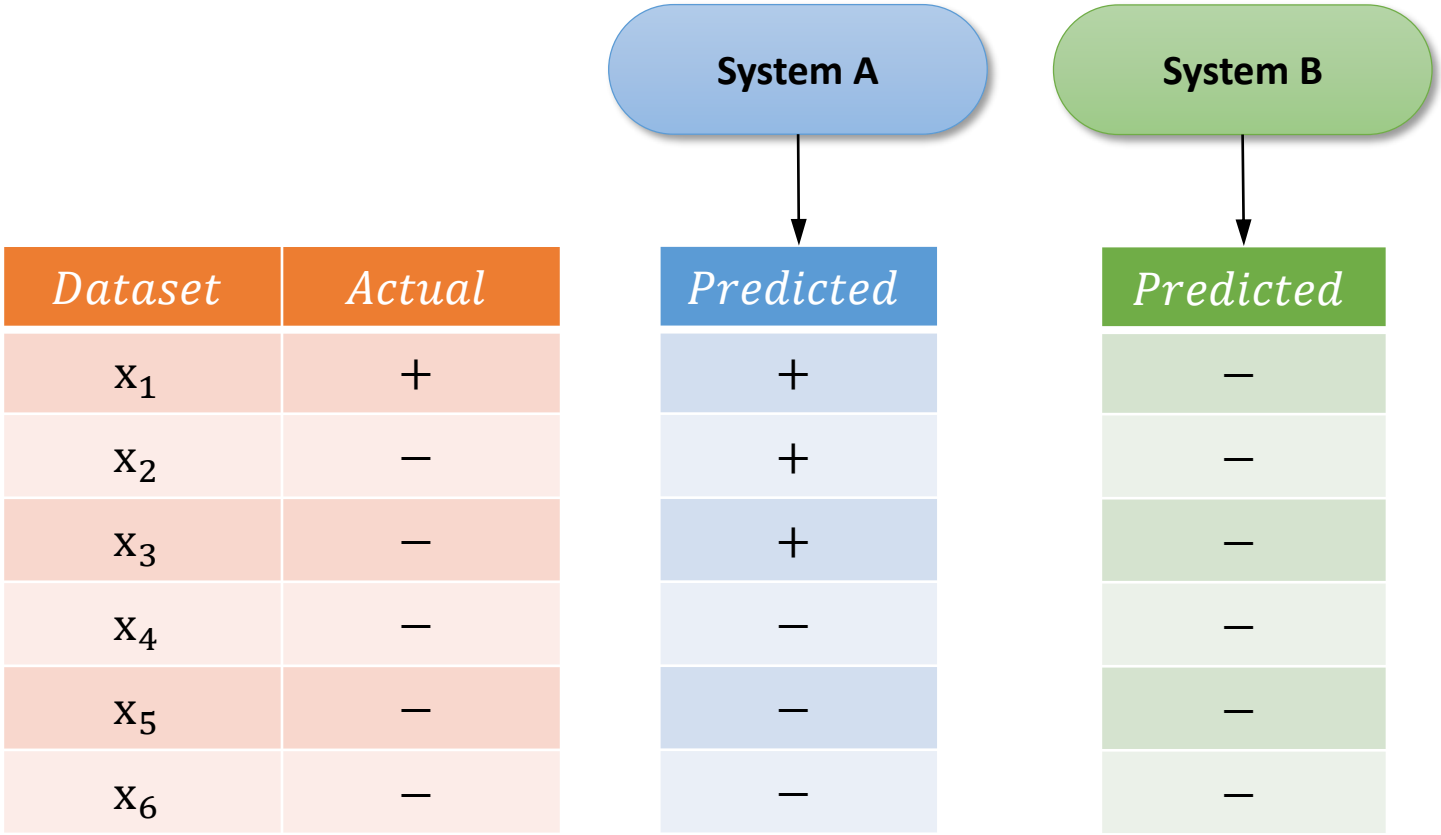
Ground-truth value

Predicted value		Positive (1)	Negative (0)
	Positive (1)	True Positive (TP)	False Positive (FP)
	Negative (0)	False Negative (FN)	True Negative (TN)

- **Why is not a useful evaluation measure in some domains?**

Example: Accuracy

- Which system is better in terms of accuracy?



Dataset	*Actual*	*Predicted*	*Predicted*
x_1	+	+	−
x_2	−	+	−
x_3	−	+	−
x_4	−	−	−
x_5	−	−	−
x_6	−	−	−

Example: Why not just Use Accuracy?

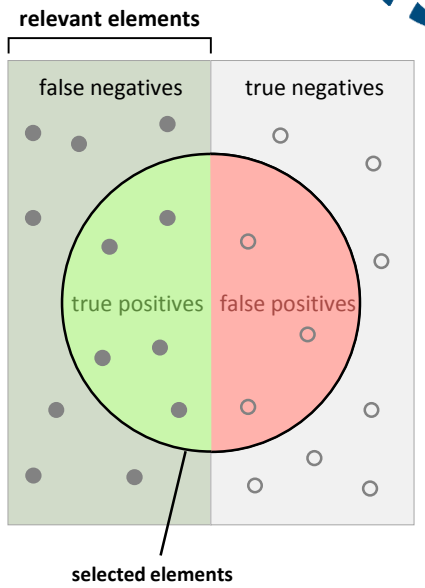
- **99.9% of documents are irrelevant in most of the cases**
  - Labeling every document as **irrelevant** has **high accuracy**, but it is useless in the Web search engine.

snoogle.com

Search for:

0 matching results found.

Precision and Recall



➢ Precision

- **Exactness:** how many selected items are relevant?

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

Precision =

➢ Recall

- **Completeness**: how many relevant items are selected?

$$Recall = \frac{TP}{TP + FN}$$

Recall =

➢ The perfect score of both measures is 1.0.

- In general, the **inverse** relationship between precision and recall

F-Measure (F-Score)

F-measure (F_1 or F-score)

- The weighted **harmonic mean** of precision and recall

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R} \quad \text{where } \beta^2 = \frac{1-\alpha}{\alpha}$$

- When $\alpha = 0.5$ $(i.e., \beta = 1.0)$

$$F = \frac{2PR}{P+R}$$

Why harmonic mean?

- The harmonic mean is **always less than or equal to** the arithmetic mean and the geometric mean.
- When P and R differ greatly, **the harmonic mean is closer to their minimum** than to their arithmetic mean.

Example: Precision and Recall

Dataset	*Actual*	*Predicted* (System A)	*Predicted* (System B)
x_1	+	+	+
x_2	−	+	−
x_3	−	−	−
x_4	+	+	−
x_5	−	+	−
x_6	−	+	−

Precision:
Recall:
F-score:

Precision:
Recall:
F-score:

Extending Multi-class Classification

- Categorizing each sample into 1 of N different classes

Ground-truth value

Predicted value	Cat	Dog	Fish
Cat	4	6	3
Dog	1	2	0
Fish	1	1	6

Example: Accuracy

- The accuracy is $(3 + 2 + 5)/25 = 10/25$.

Predicted value \ Ground-truth value	Cat	Dog	Fish
Cat	3	6	5
Dog	2	2	0
Fish	1	1	5

Example: Precision and Recall

- For Cat, the precision is $3/(3+6+5) = 3/14$.

Predicted value \ Ground-truth value	Cat	Dog	Fish
Cat	3	6	5
Dog	2	2	0
Fish	1	1	5

Example: Precision and Recall

- For Cat, the recall is $3/(3+2+1) = 3/6$.

Ground-truth value

Predicted value	Cat	Dog	Fish
Cat	3	6	5
Dog	2	2	0
Fish	1	1	5

Example: Precision and Recall

- For Dog and Fish, the precision is $2/4$ and $5/7$.
- For Dog and Fish, the recall is $2/9$ and $5/10$.

Ground-truth value

Predicted value	Cat	Dog	Fish
Cat	3	6	5
Dog	2	2	0
Fish	1	1	5

Macro-Precision and Macro-Recall

- **For three classes, the average of precision and recall is called macro-precision and macro-recall.**
  - Macro-precision: $(3/14 + 2/4 + 5/7)/3 = 0.476$
  - Macro-recall: $(3/6 + 2/9 + 5/10)/3 = 0.407$

Ground-truth value

Predicted value	Cat	Dog	Fish	
Cat	3	6	5	$3/14$
Dog	2	2	0	$2/4$
Fish	1	1	5	$5/7$
	$3/6$	$2/9$	$5/10$	

Q&A