ORTA DOĞU TEKNİK ÜNİVERSİTESİ
MIDDLE EAST TECHNICAL UNIVERSITY

CENG 463: Introduction to Natural Language Processing
Evaluation in NLP

Asst. Prof. Cagri Toraman
Computer Engineering Department
ctoraman@ceng.metu.edu.tr

21.10.2025

NLP Evaluation

How to assess the performance of your trained model?

<u>Dimension-1: Automatic vs. Manual evaluation</u>

Automatic: Comparison between model output and ground truth (gold label/standard)

Manual: Human experts to evaluate performance

What are the pros/cons of each approach?

NLP Evaluation

How to assess the performance of your trained model?

<u>Dimension-2: Intrinsic vs. Extrinsic Evaluation</u>

Intrinsic: Focusing on internal performance/capabilities of NLP model

Extrinsic: Focusing on external NLP tasks while evaluating NLP model

What are the pros/cons of each approach?

NLP Evaluation

How to assess the performance of your NLP model?

Automatic and Intrinsic Evaluation

$$\begin{aligned}\text{perplexity}(W) &= P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}} \\ &= \sqrt[N]{\frac{1}{P(w_1 w_2 \ldots w_N)}}\end{aligned}$$

By using the Chain Rule:

$$\text{perplexity}(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_1 \ldots w_{i-1})}}$$

The lower the perplexity of a model on the data, the better the model!

NLP Evaluation

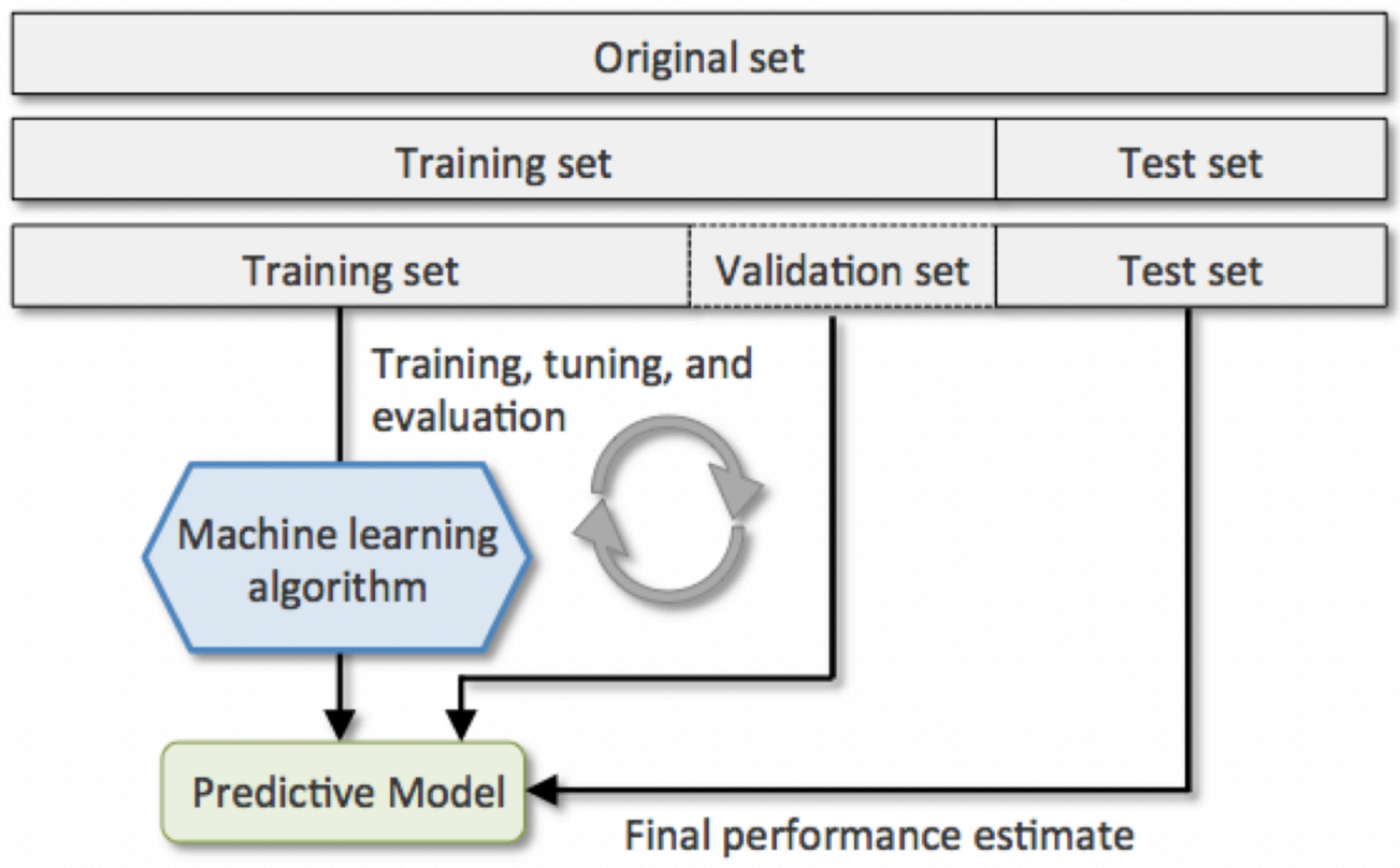
How to assess the performance of your NLP model?

Automatic and Extrinsic Evaluation

In-sample data: Data used in training phase

Out-of-sample data: Data not used in training (also called held-out set)

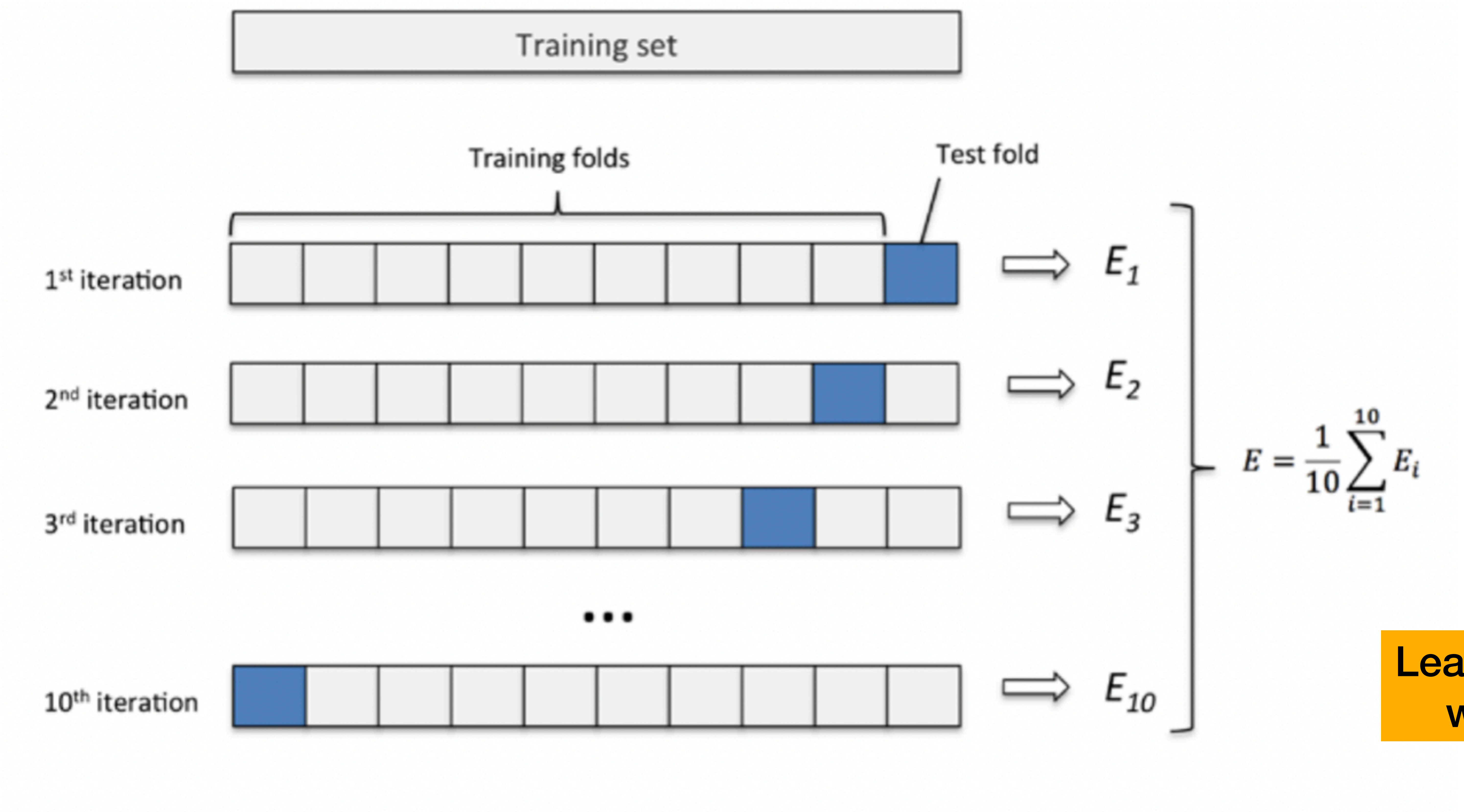
Model Evaluation



How to choose train/ validation/test size?

Model Evaluation

K-fold Cross-Validation:



How to choose *k*?

Leave-one-out cross-validation: when k equals to data size

Model Evaluation

Some common mistakes during splitting data:

- Ground Truth Errors: Data labeling/annotation is very important (garbage in —> garbage out)
- Duplicates: Keeping same instances in both train and test

  (e.g. due to duplicates in data)
- Type Dependency: Keeping same type of instances in both train and test

  (e.g. due to multiple instances from the same person when person is target class)
- Time Dependency: Data has a time order but test split has same/previous time period with train

  (e.g. when target is to predict tweet engagement)

Evaluation Metrics

How to assess the performance of your NLP model?

Consider different NLP tasks:

Sentiment Analysis
Named Entity Recognition
Morphological Analysis
Language Modeling
Generative NLP

How to choose evaluation metrics?

Evaluation Metrics

How to assess the performance of your trained model?

Some important evaluation metrics for supervised NLP tasks:

- Accuracy
- Precision
- Recall
- F1-Score
- ROC Curve

Evaluation Metrics

Accuracy:

$$\frac{\text{Number of correctly classified instances}}{\text{Total number of instances}}$$

In which scenario, accuracy becomes a poor metric?

Evaluation Metrics

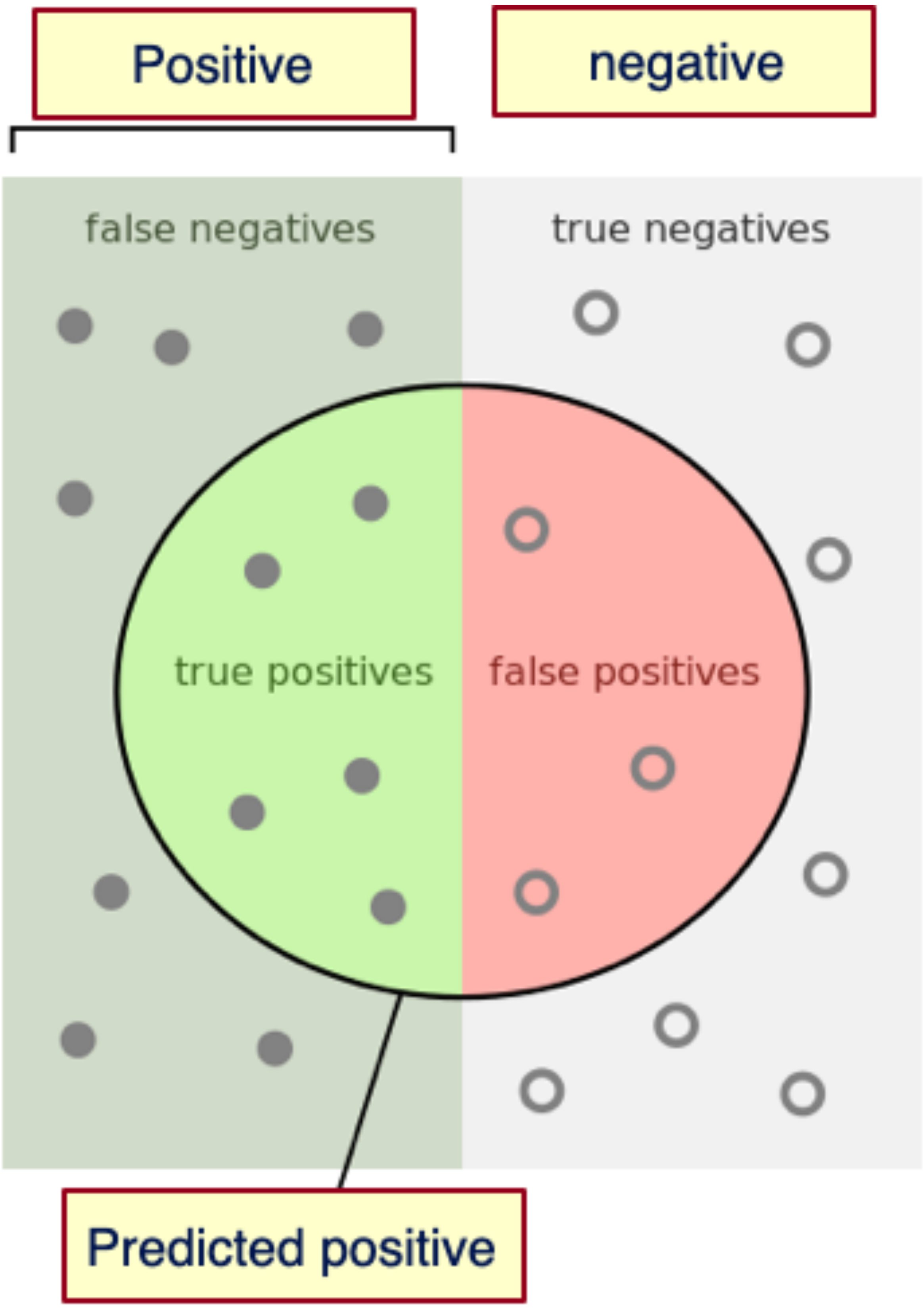
We need more than Accuracy!

True Positive (TP): Ground truth is *c* and the model predicts *c*

False Positive (FP): Ground truth is not *c* but the model predicts *c*

True Negative (TN): Ground truth is not *c* and the model does not predict *c*

False Negative (FN): Ground truth is *c* but the model does not predict *c*

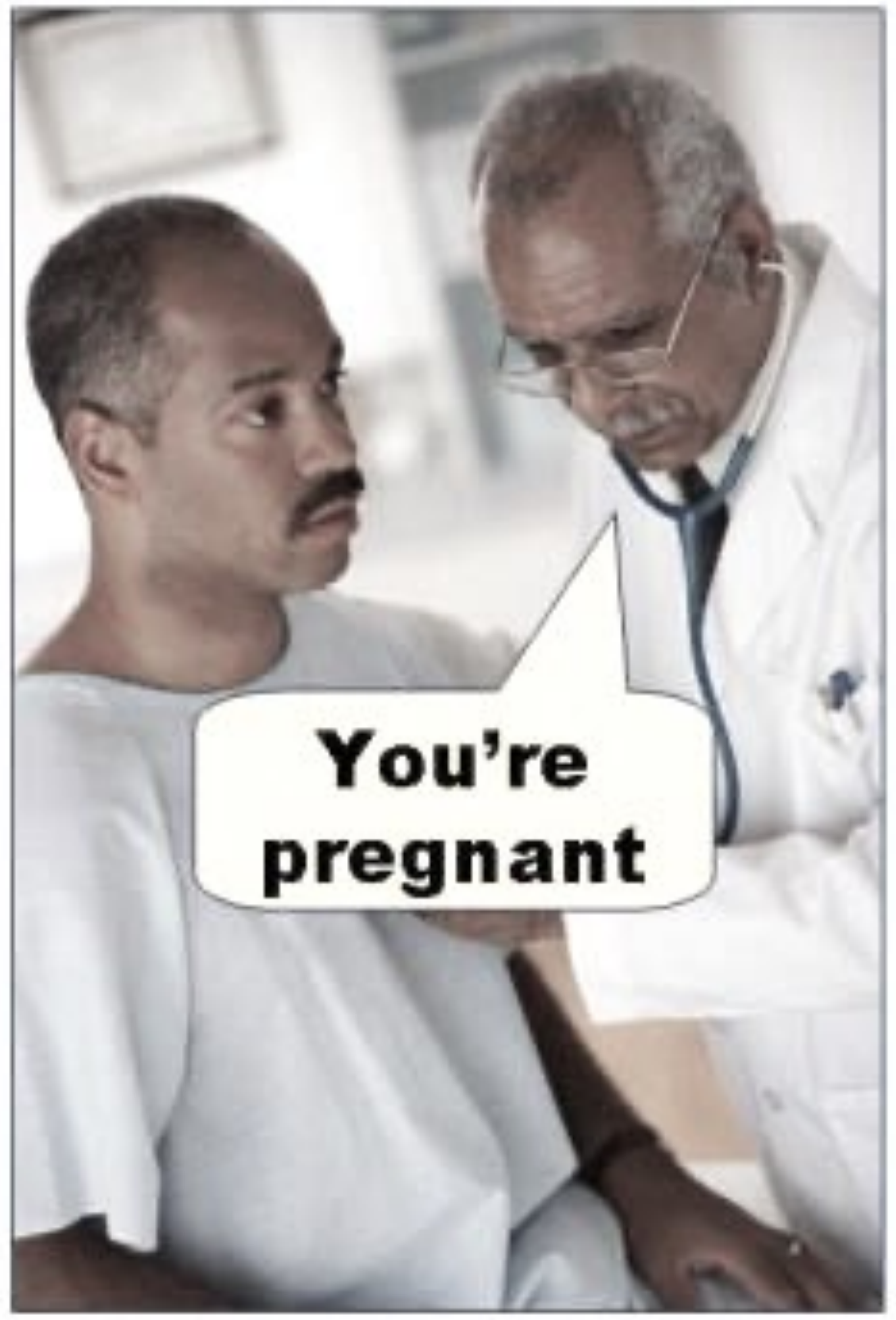


Evaluation Metrics

False Positive (FP): Type I Error

False Negative (FN): Type II Error

Which one is important?



Evaluation Metrics

Precision:

$$\frac{\text{Number of correctly classified positive instances}}{\text{Total number of positive predictions}} = \frac{TP}{TP+FP}$$

Recall:

$$\frac{\text{Number of correctly classified positive instances}}{\text{Total number of positive instances}} = \frac{TP}{FN+TP}$$

In which scenarios, precision and recall are important?

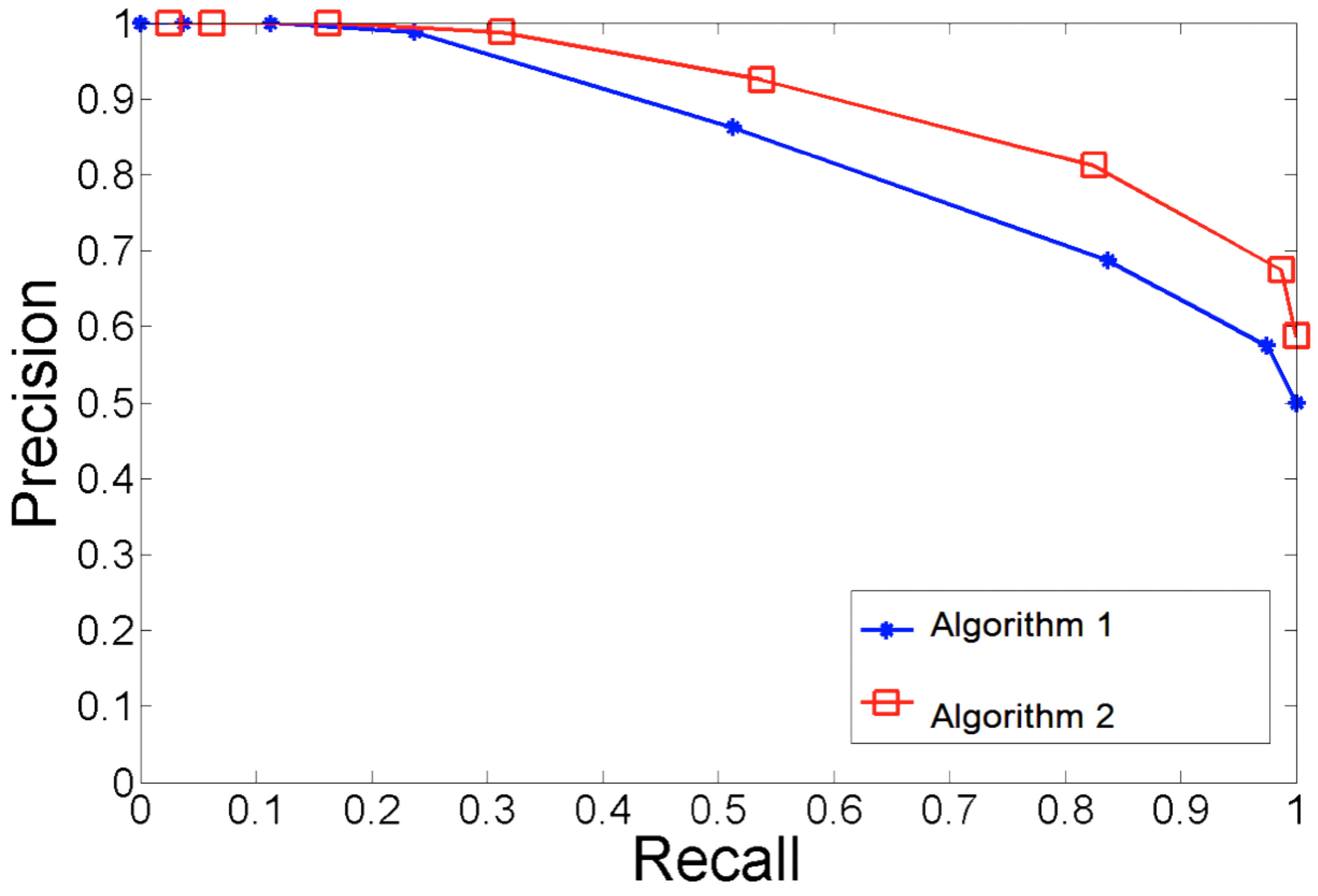
Evaluation Metrics

How to use precision and recall together?

F1-Score:
Harmonic mean of precision and recall

$$F1 = 2\frac{PRE \times REC}{PRE + REC}$$

Precision-Recall Curve:



Evaluation Metrics

How to report overall evaluation metric for all classes?

Micro Average (e.g. Precision):

$$\frac{\text{Sum of the number of correctly classified positive instances}}{\text{Total number of positive predictions}} = PRE_{micro} = \frac{TP_1 + \cdots + TP_k}{TP_1 + \cdots + TP_k + FP_1 + \cdots + FP_k}$$

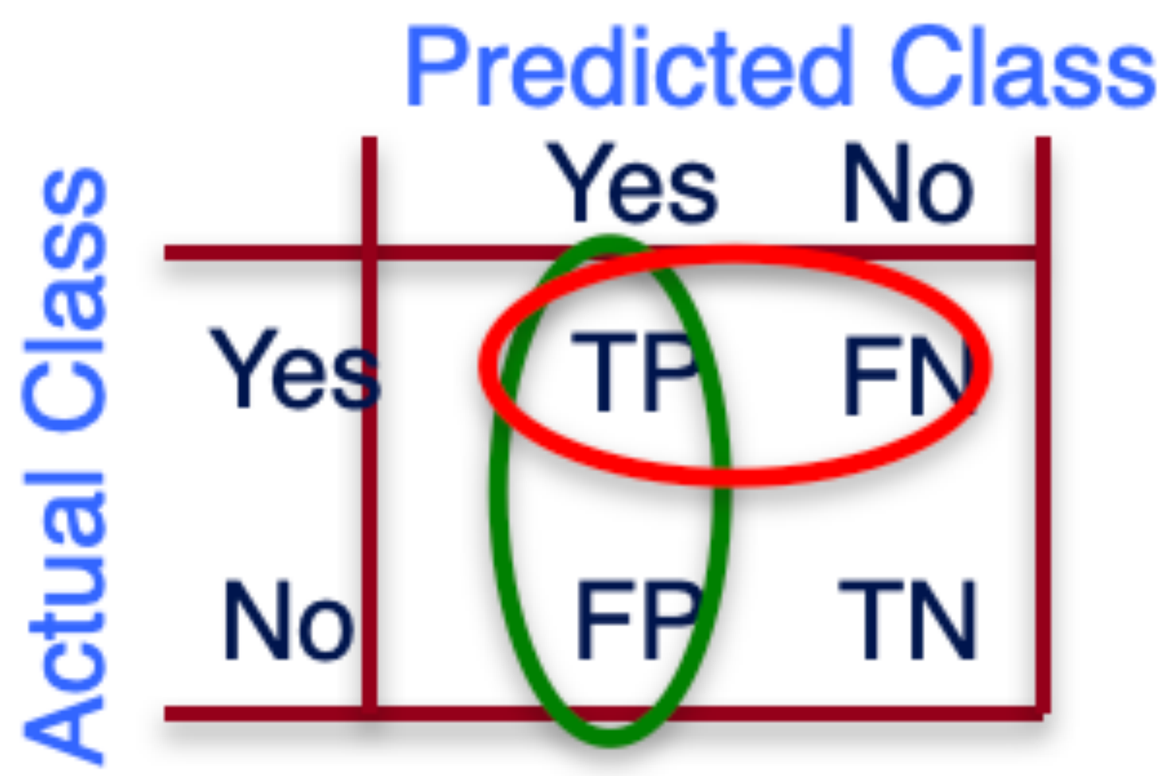
Macro Average (e.g. Precision):

$$\frac{\text{Sum of the precision scores of each class}}{\text{Total number of positive instances}} = PRE_{macro} = \frac{PRE_1 + \cdots + PRE_k}{k}$$

Error Analysis

Confusion Matrix:

A table that reports the number of test instances with ground truth label and predicted label.



Evaluation Metrics

An example:

test data	prediction	ground truth	TP	FP	FN	TN
X1	+	-		1		
X2	+	+	1			
X3	-	+			1	
X4	-	-				1
X5	+	-		1		

Precision
= TP / (TP+FP)
= 1 / (1+2)
= 0.33

Recall
= TP / (TP+FN)
= 1 / (1+1)
= 0.5

F1-Score
= 2*Recall*Precision / (Recall+Precision)
= (2 * 1/3 * 1/2) / (1/3 + 1/2)
= 0.4

Evaluation Metrics

Another example:

		predicted labels 1	predicted labels 0
true labels	1	10	10
	0	20	160

Baseline (majority accuracy): 90%

Accuracy: 85%

Precision: 0.333

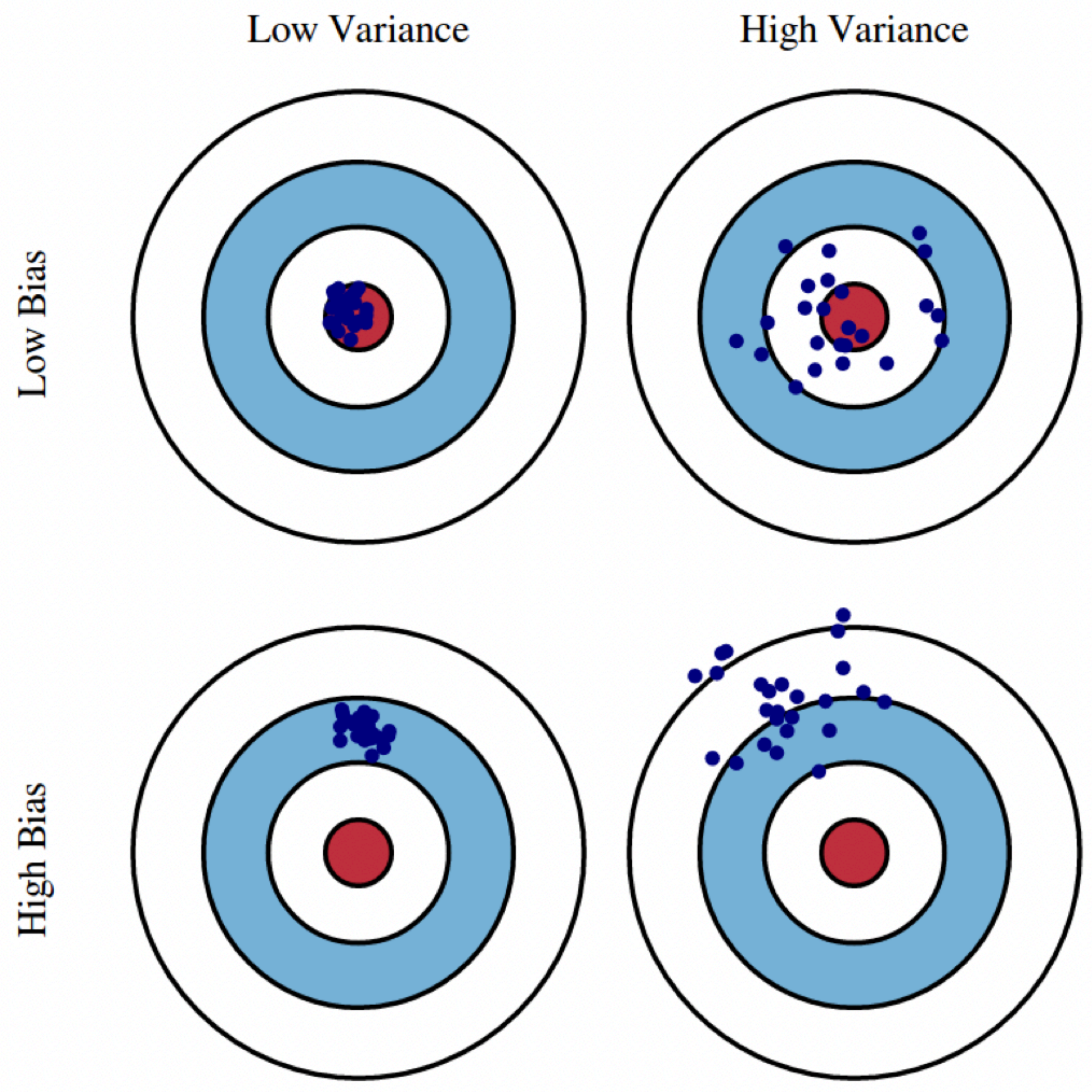
Recall: 0.5

F1: 0.4

Bias and Variance

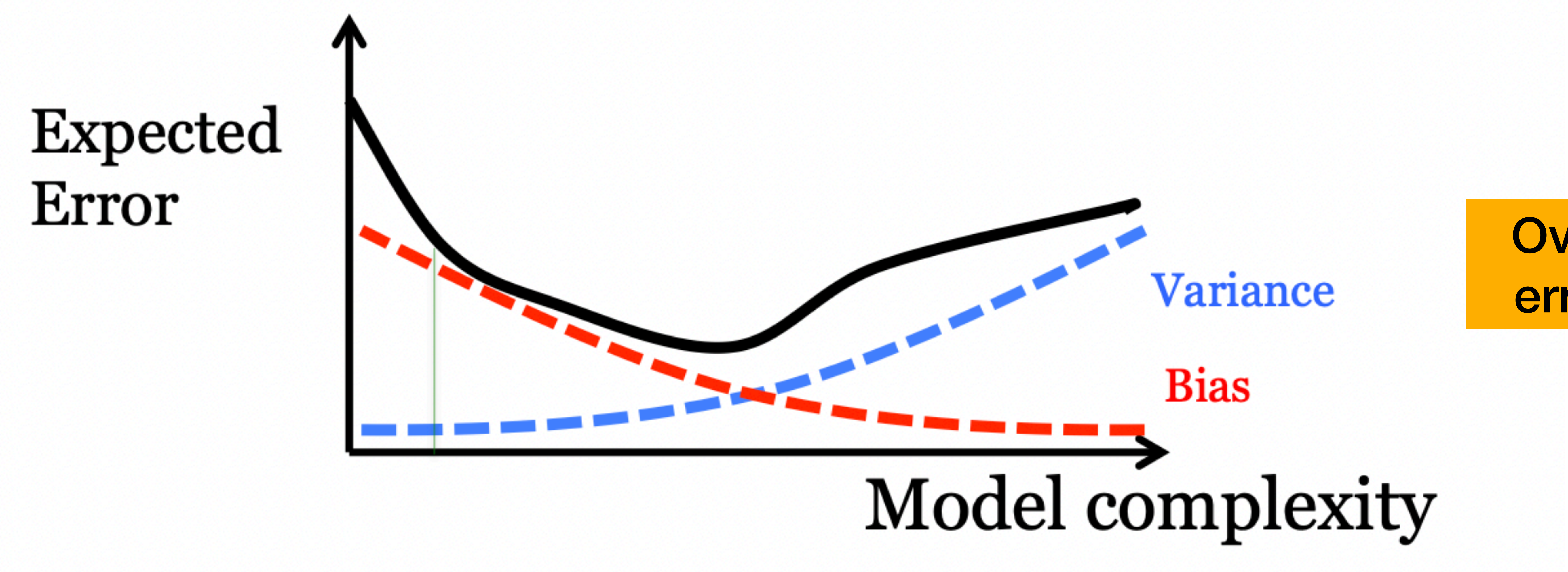
<u>Bias</u>:
Error due to incorrect assumptions
(learning algorithm can not represent the concept)

<u>Variance</u>:
Error due to variance of training samples
(learning algorithm overreacts to noise in the training data)



Bias vs. Variance Trade-off

Underfitting: High training error and high test error



Overfitting: Low training error and high test error

More complex models overfit while the simplest models underfit.

Baseline Model

Baseline is a successful method for the given task.

Baseline is mostly a state-of-the-art (SOTA) model from the literature, which exhibits well-known performance in the related domain and task.

NLP Evaluation

How to assess the performance of your NLP model?

<u>Manual and Extrinsic Evaluation</u>

Human annotations/labels

LLM-as-a-Judge

NLP Evaluation

How to assess the performance of your NLP model?

<u>Manual and Extrinsic Evaluation</u>

Human annotations/labels:

Need to find domain expert(s)

Need to find multiple annotators to have consistent results

Need to instruct all annotator(s) with careful guidelines

What could possible go wrong
for human annotations?

NLP Evaluation

Inter-annotator Agreement:

A method to measure the reliability of manual human annotations.

What if there are two human annotators for two labeling classes (Positive and Negative):

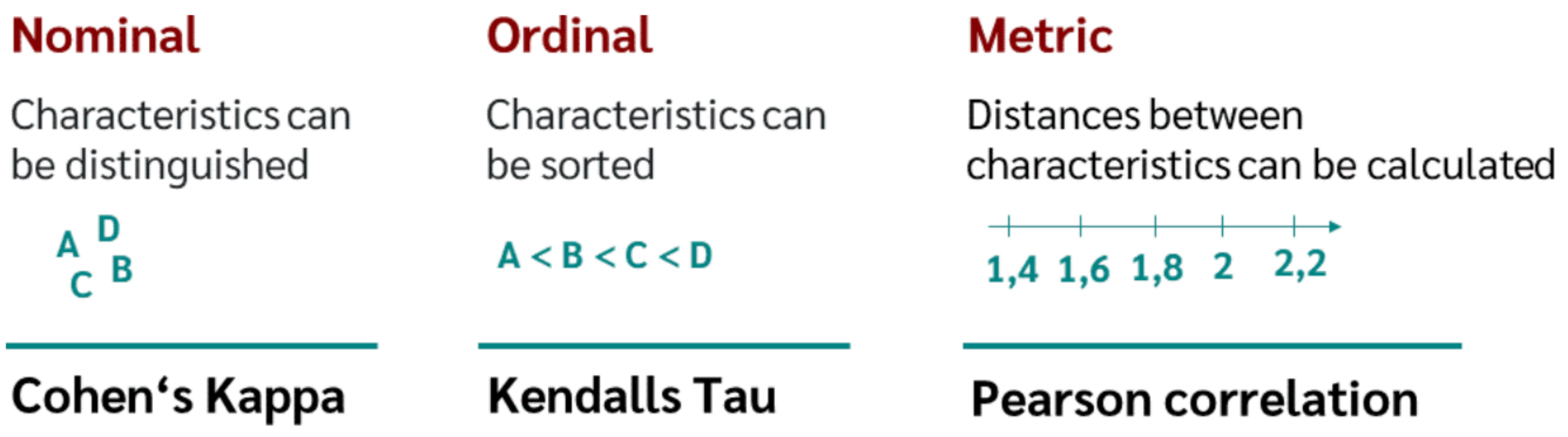
Annotator-1: 90% Positive, 10% Negative
Annotator-2: 30% Positive, 70% Negative

Possibility that two annotators have agreement by chance:
$0.9*0.3 + 0.1*0.7 = 0.34$

NLP Evaluation

Inter-annotator Agreement:

Cohen’s Kappa and variations



NLP Evaluation

Inter-annotator Agreement:

Cohen's Kappa

the observed agreement (total agreements divided by total number of items)

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

an estimate of chance agreement varying according to the specific measure

NLP Evaluation

<u>Inter-annotator Agreement:</u>

Cohen's Kappa

Annotator-2

Annotator-1	🙂	☹️
🙂	**17**	**8**
☹️	**6**	**19**

$$p_o = \frac{36}{50} = 72\%$$

$$\kappa = \frac{p_o - p_e}{1 - p_e} = \frac{0.72 - 0.5}{1 - 0.5} = 0.44$$

	🙂	☹️	
🙂	17	8	25
☹️	6	19	25
	23	27	

$$p_e = \underbrace{\underbrace{\frac{25}{50} \cdot \frac{23}{50}}_{0.5 \cdot 0.46}}_{0.23} + \underbrace{\underbrace{\frac{25}{50} \cdot \frac{27}{50}}_{0.5 \cdot 0.54}}_{0.27} = 0.5$$

probability that both annotators would say "negative" by chance

NLP Evaluation

How to assess the performance of your NLP model?

<u>Manual and Extrinsic Evaluation</u>

LLM-as-a-Judge:

Use Generative LLM/AI instead of human experts

Need to employ multiple LLM/AI models

Need to prepare optimal prompts for evaluation

What could possible go wrong for LLM/AI annotations?

ORTA DOĞU TEKNİK ÜNİVERSİTESİ
MIDDLE EAST TECHNICAL UNIVERSITY

Thanks for your participation!

Çağrı Toraman
21.10.2025