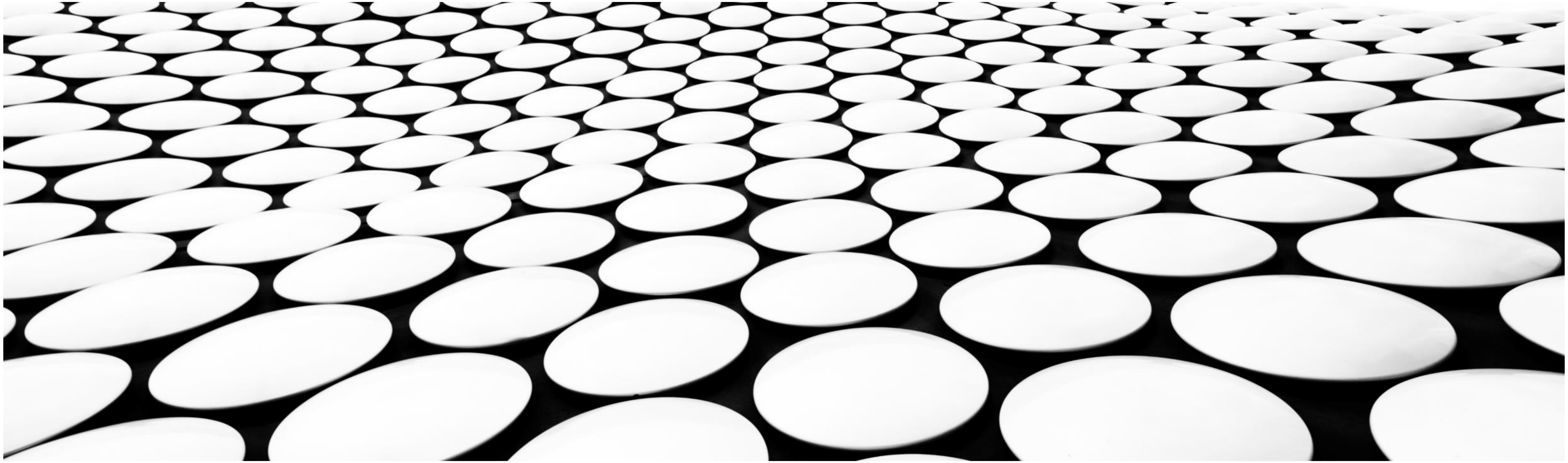
CAROLINE BARRIÈRE, CSI4106, FALL 2020

SUPERVISED MACHINE LEARNING

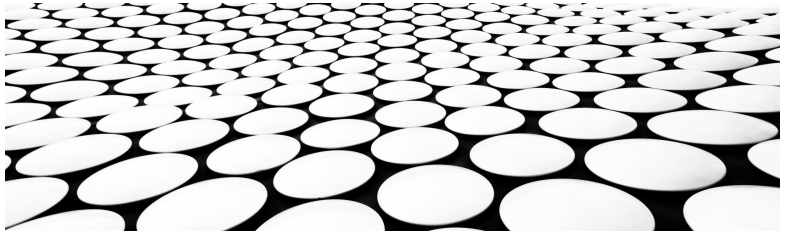
PART 4 OF 4 –EVALUATION

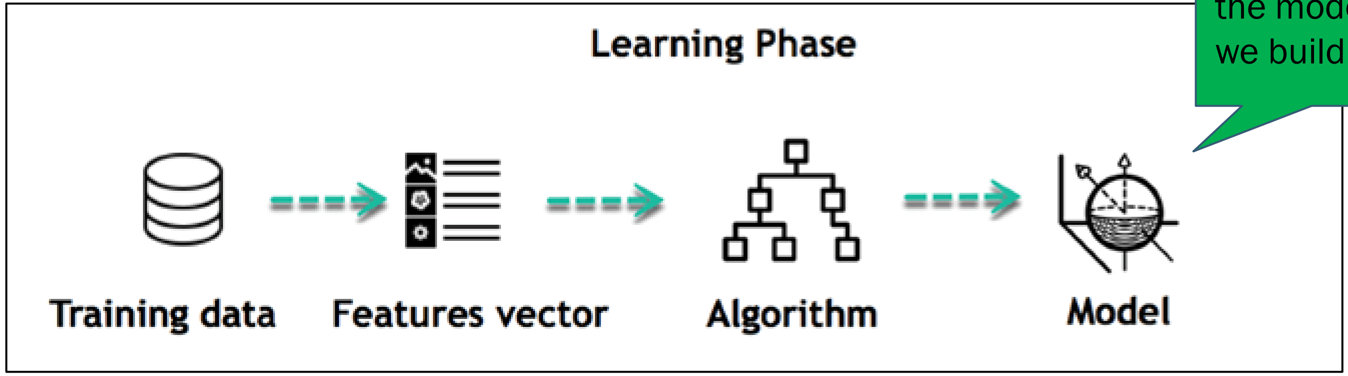


GOALS

- How to set up an experiment?
- How to evaluate?

LEARNING / PREDICTION

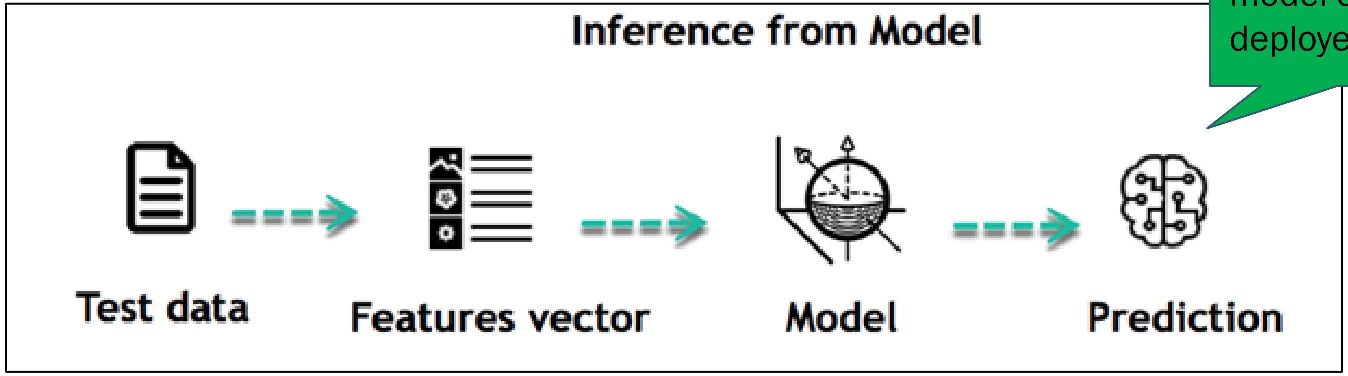




Learning Phase

Training data → Features vector → Algorithm → Model

How good is the model as we build it?

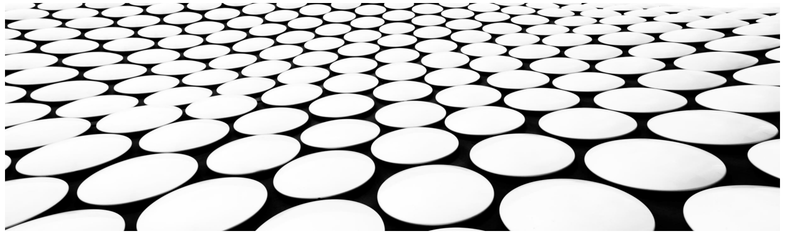


Inference from Model

Test data → Features vector → Model → Prediction

How good is the model once deployed?

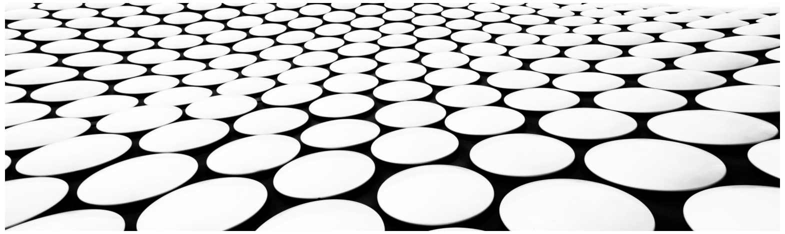
FACTORS TO CONSIDER



What influences the evaluation:

1. What do we evaluate on?
2. Is the training data representative of the test data ?
3. What is the performance measure?
4. Is the Gold Standard unanimously agreed on?

(1) WHAT DO WE EVALUATE ON?



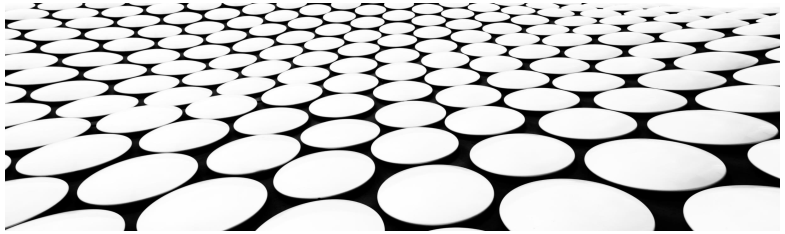
We should ALWAYS evaluate on unseen data.

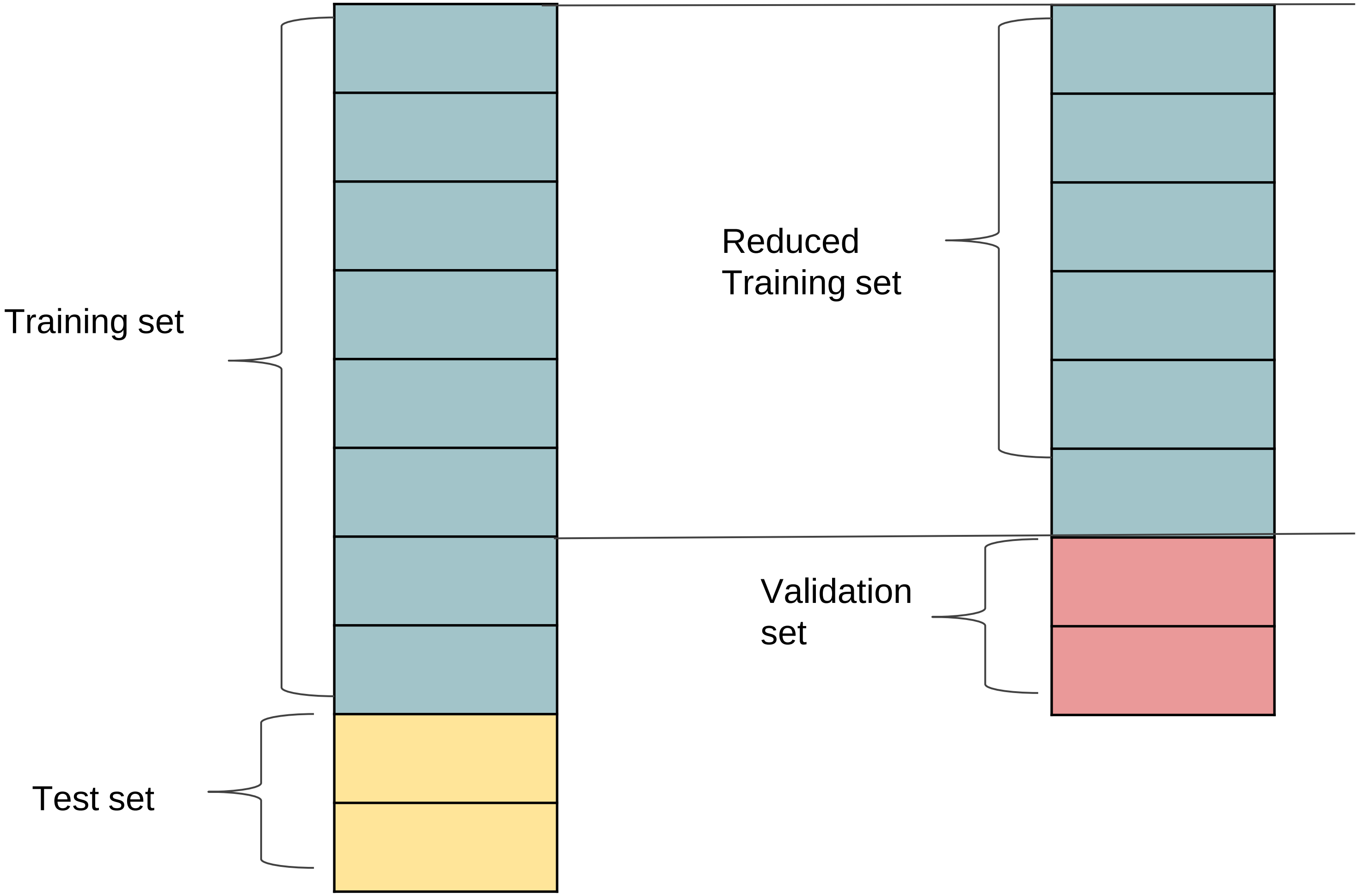
Training set

Test set

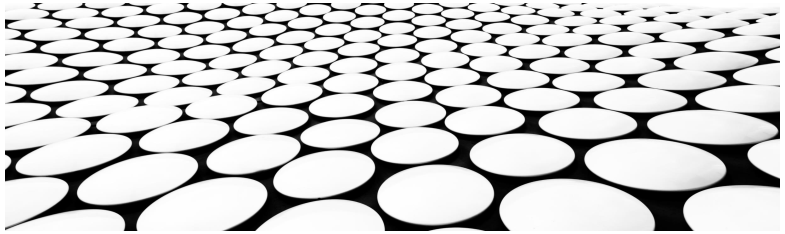
	Prediction	Target
M1	Drama	Drama
M2	Comedy	Drama
M3	Comedy	Comedy
M4	Comedy	Comedy
M5	Drama	Drama
M6	Drama	Comedy
M7	Comedy	Comedy
	Drama	?
	Comedy	?

VALIDATION SET





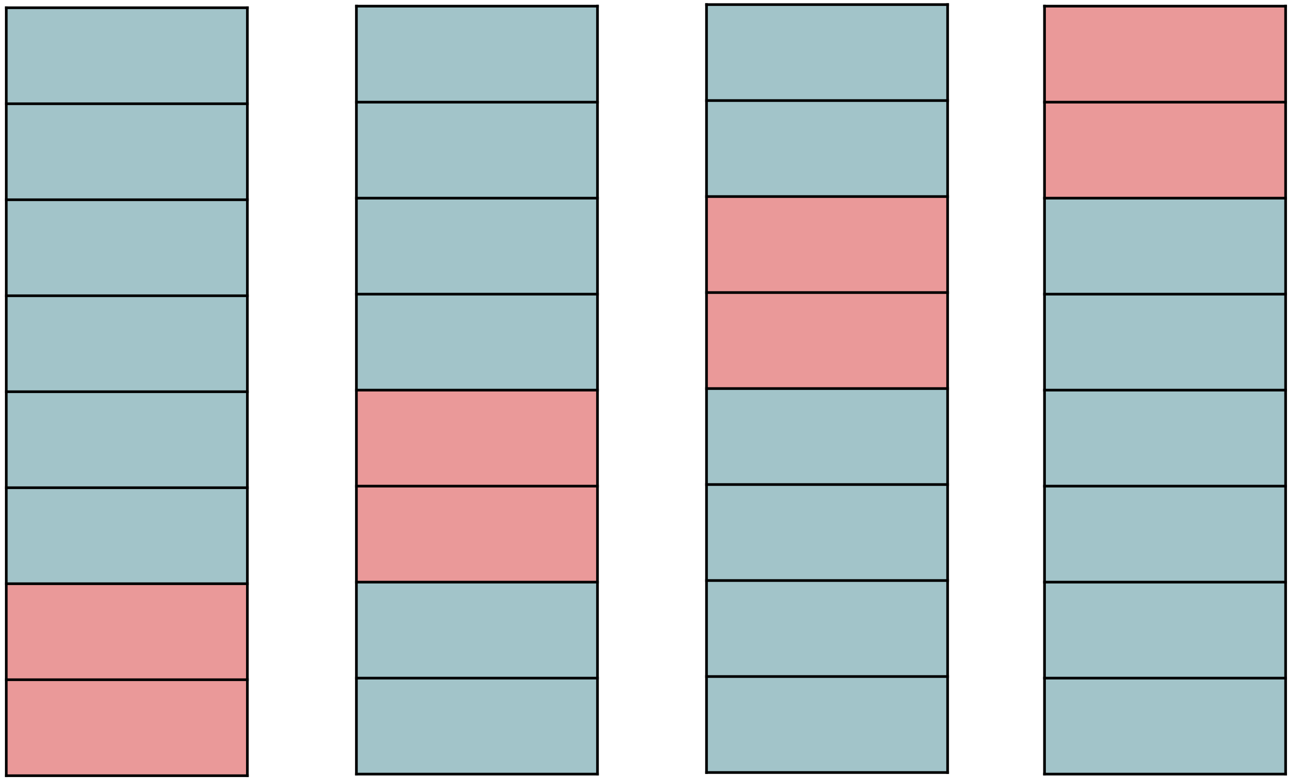
CROSS-VALIDATION



🟦 Training set

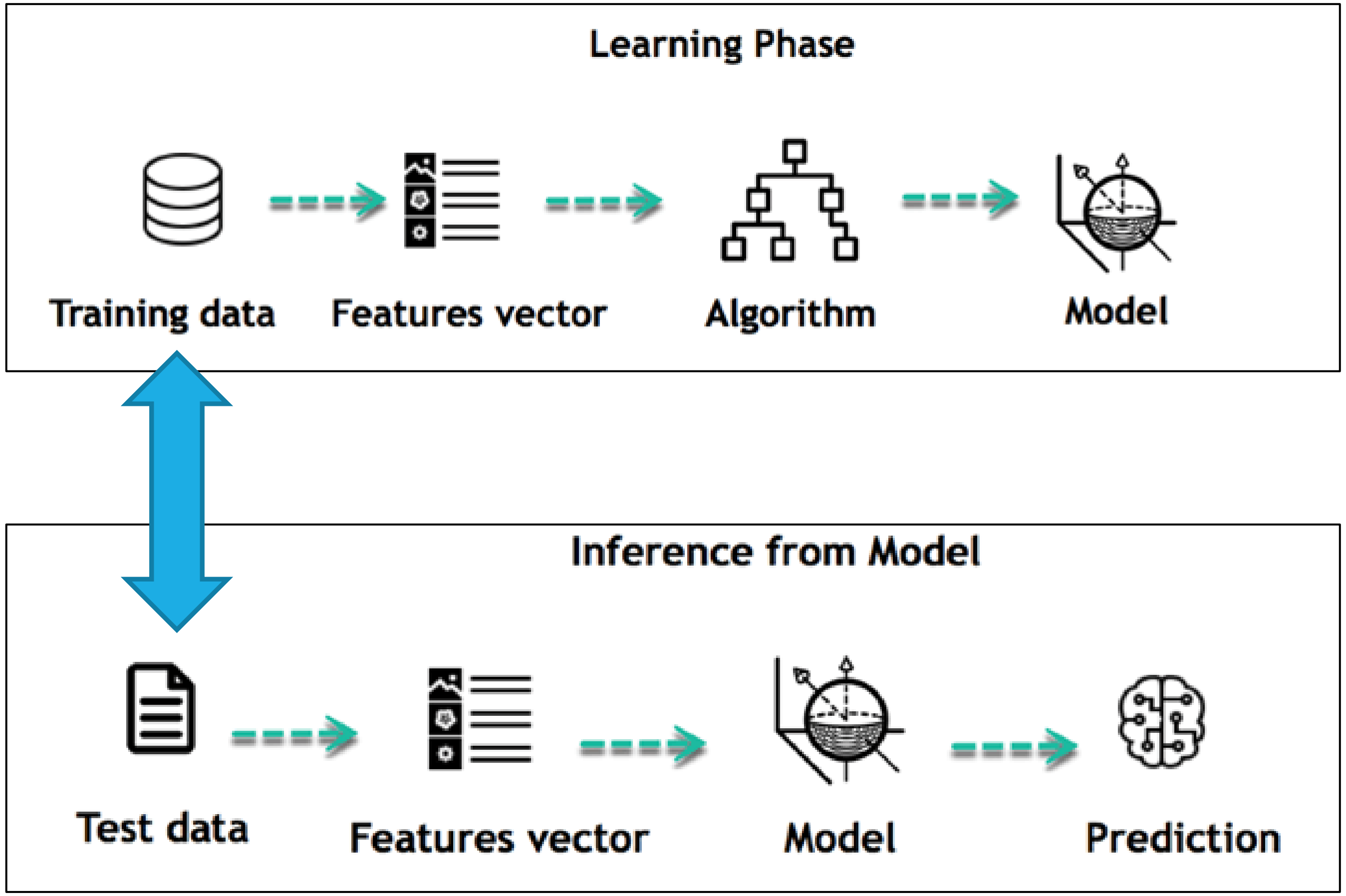
🟥 Validation set

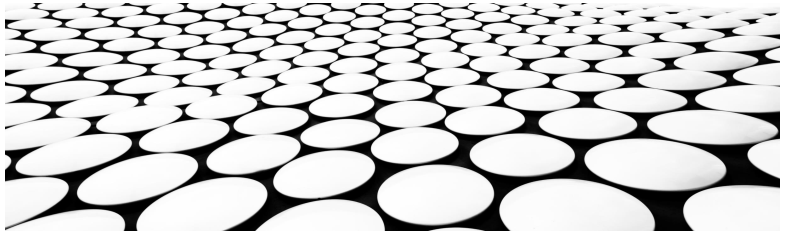
4-fold cross-validation



🟦	🟦	🟦	🟥
🟦	🟦	🟦	🟥
🟦	🟦	🟥	🟦
🟦	🟦	🟥	🟦
🟦	🟥	🟦	🟦
🟦	🟥	🟦	🟦
🟥	🟦	🟦	🟦
🟥	🟦	🟦	🟦

(2) IS THE TRAINING DATA REPRESENTATIVE OF TEST DATA?



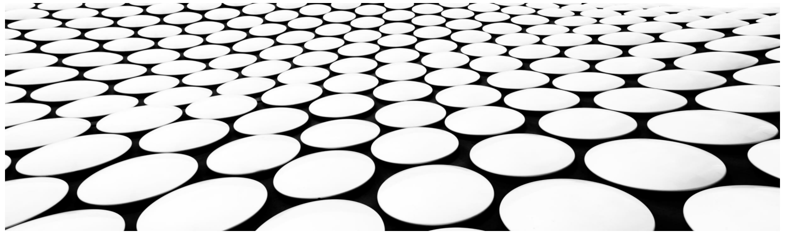


(3) WHAT IS THE PERFORMANCE MEASURE?

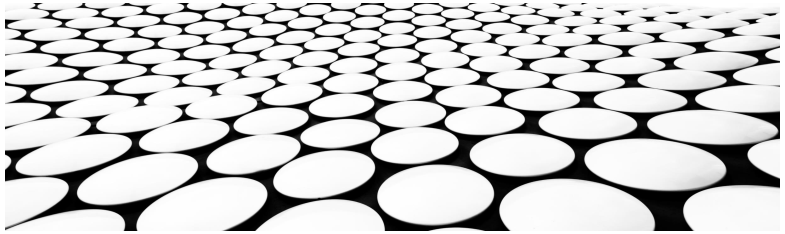
Confusion matrix

		Predicted		
		Bike	Not Bike	
Gold Standard	Bike	Tp = 3	Fn = 3	6
	Not bike	Fp = 1	Tn = 4	5
		4	7	11

Test	Gold Standard	Prediction
1	Bike	Drive
2	Drive	Drive
3	Drive	Drive
4	Bike	Drive
5	Bike	Bike
6	Drive	Drive
7	Bike	Bike
8	Drive	Drive
9	Bike	Drive
10	Bike	Bike
11	Drive	Bike



(3) WHAT IS THE PERFORMANCE MEASURE?



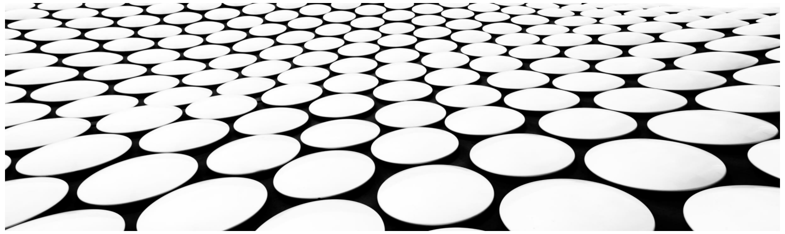
Precision = Tp / (Tp + Fp)
= 3 / (3 + 1) = 0.75

Recall = Tp / (Tp + Fn)
= 3 / (3 + 3) = 0.5

		Predicted		
		Bike	Not Bike	
Gold Standard	Bike	Tp = 3	Fn = 3	6
	Not bike	Fp = 1	Tn = 4	5
		4	7	11

Do the same precision/recall evaluation on the Drive class.

(3) WHAT IS THE PERFORMANCE MEASURE?



Per class precisions

	Bike	Drive
System	3/4 = 0.75	4/7 = 0.57

Macro-average : Average on the results per class

Macro-average on precisions: $(\text{Prec}_{C1} + \text{Prec}_{C2})/ 2$

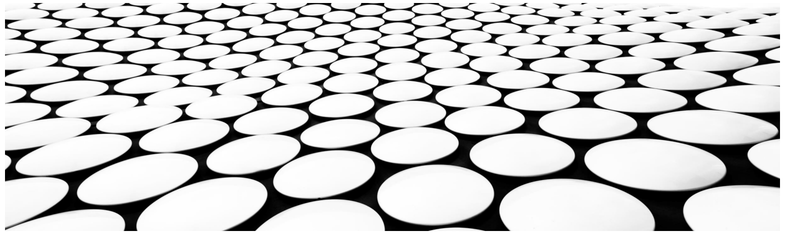
$(0.75 + 0.57)/2 = 0.66$

Micro-average : Average when putting all the data together.

Micro-average of precision:: $(TP_{C1} + TP_{C2}) / (TP_{C1} + FP_{C1} + TP_{C2\,+} + FP_{C2})$

$(3 + 4) / (4 + 7) = 7 / 11 = 0.64$

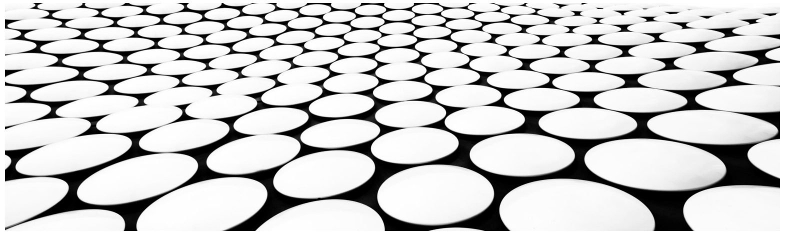
(3) WHAT IS THE PERFORMANCE MEASURE?



Comparative evaluation

	Bike		Drive	
	Precision	Recall	Precision	Recall
System 1	3/4 = 0.75	3/6 = 0.5	4/7 = 0.57	4/4 = 1.0
System 2	3/5 = 0.6	3/6 = 0.5	3/6 = 0.5	3/4 = 0.75

(4) IS THE GOLD STANDARD UNANIMOUSLY AGREED ON?

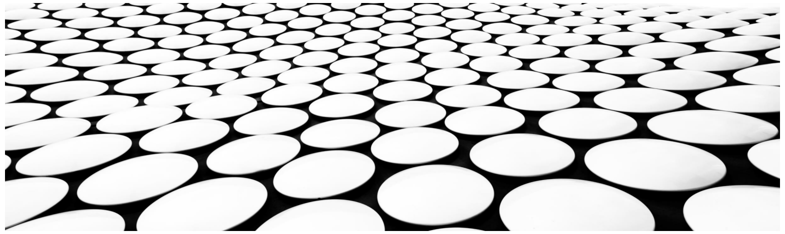


We assume that the annotation (classification by humans) is the "Gold Standard", but do humans all say the same thing?

What is your annotation? Rotten or Fresh

	Review	Rotten / Fresh
1	In action, the film is breathtaking, but as a whole it suffers from a relative lack of ambition.	
2	After the setup, the air leaks out of the movie, flattening its momentum with about an hour to go.	
3	This film is not a groundbreaking film by any means, but at least it's fun	
4	A warm and fun crowd pleaser	
5	This is a tedious tale badly told.	

FACTORS TO CONSIDER



What influences the evaluation:

1. What do we evaluate on?
2. Is the training data representative of the test data ?
3. What is the performance measure?
4. Is the Gold Standard unanimously agreed on?

IN SUMMARY

- Supervised Machine Learning
  - Components of a SML system (part 1)
  - Features (part 2)
  - Generative vs Discriminative Models (part 3)
  - Evaluation (part 4)