ALGOSOC

UOB

DATA SCIENCE & MACHINE LEARNING WORKSHOP

Week 5 - Evaluating Models

GOOGLE GEMINI X ALGOSOC TALK ON AI

WE WILL HAVE SOME EXTERNAL SPEAKERS THAT WILL BE DELIVERING AN ENGAGING WORKSHOP FOR STUDENTS ON HOW TO LEVERAGE GEMINI, GOOGLE'S AI TOOL, TO PRACTICALLY AND ETHICALLY ENHANCE—NOT REPLACE—THEIR ACADEMIC STUDIES.

THERE WILL BE 2 SLOTS:

1. 2:30PM - 3:30PM
2. 4:00PM - 5:00PM.

YOU ARE FREE TO ATTEND EITHER ONE - THEY WILL BE THE SAME TALK.

LIMITED AVAILIBILITY!!!
SCAN THE QR CODE TO SECURE YOUR SPOT!!!

LOCATION - Y3-G34
DATE - 20TH NOV 2025

AGENDA

Week 5 topic:

- Motivation for Validation
- Classification Metrics: Accuracy, precision, recall, F1-score
- Overfitting vs. underfitting
- Why we need model selection
- Model selection and Hyperparameter tuning

Full agenda this semester:
https://bit.ly/DataScienceAlgosoc

Repository:
https://github.com/AlgoSoc/Data-Science

MOTIVATION

Motivation for Validation

Model Selection

How do we select the “optimal” parameter(s) or model for a given classification problem?

- Number of neighbours in KNN
- Depths of a decision tree, pruning criteria
- Number of layers and architecture in MLP

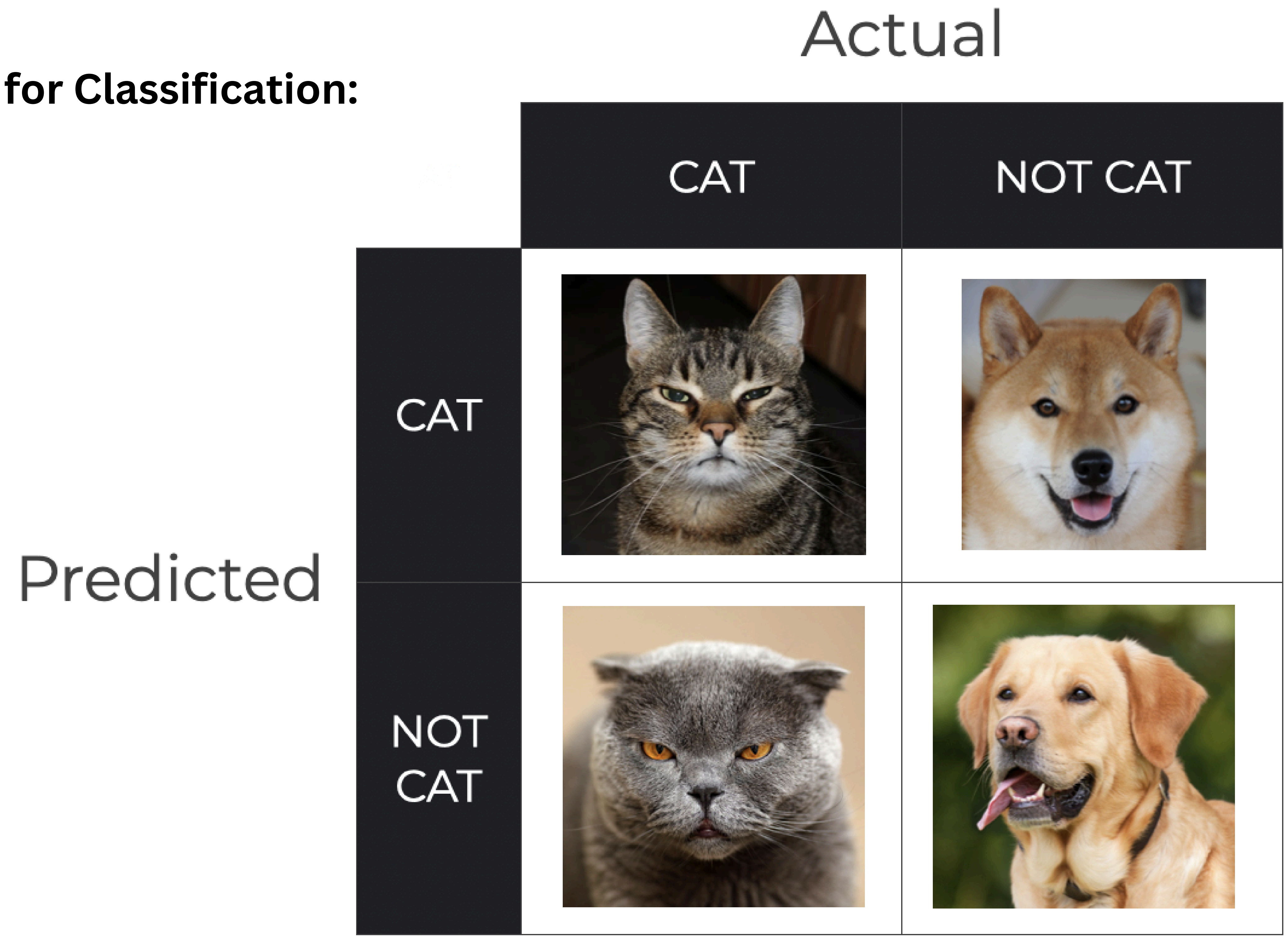
Performance Evaluation

Once we train the model, how do we measure its performance?

We use various evaluation metrics, quantifying performance

CLASSIFICATION METRICS

Previously for Classification:



https://sharpsight.ai/blog/confusion-matrix-explained/

CLASSIFICATION METRICS

Confusion matrix can also be used for multi-class classification (more than two classes) and we can quantify the model performance based on a certain class

?

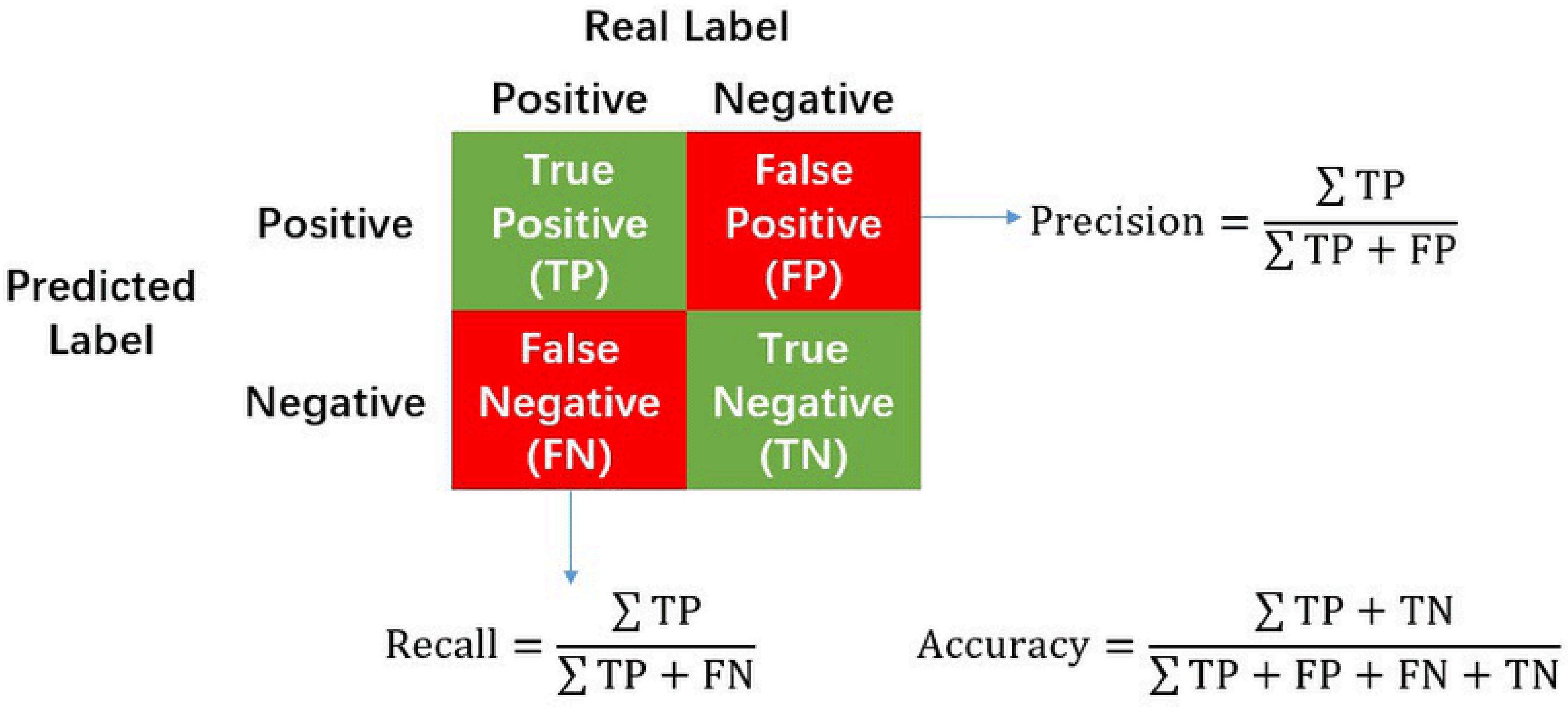
		PREDICTED classification			
	Classes	a	b	c	d
ACTUAL classification	a	TN	FP	TN	TN
	b	FN	TP	FN	FN
	c	TN	FP	TN	TN
	d	TN	FP	TN	TN

In this case, we are looking into the model's performance on **class b**

https://stats.stackexchange.com/questions/559636/interpreting-a-confusion-matrix-for-a-multiclass-classification

CLASSIFICATION METRICS

Previously:



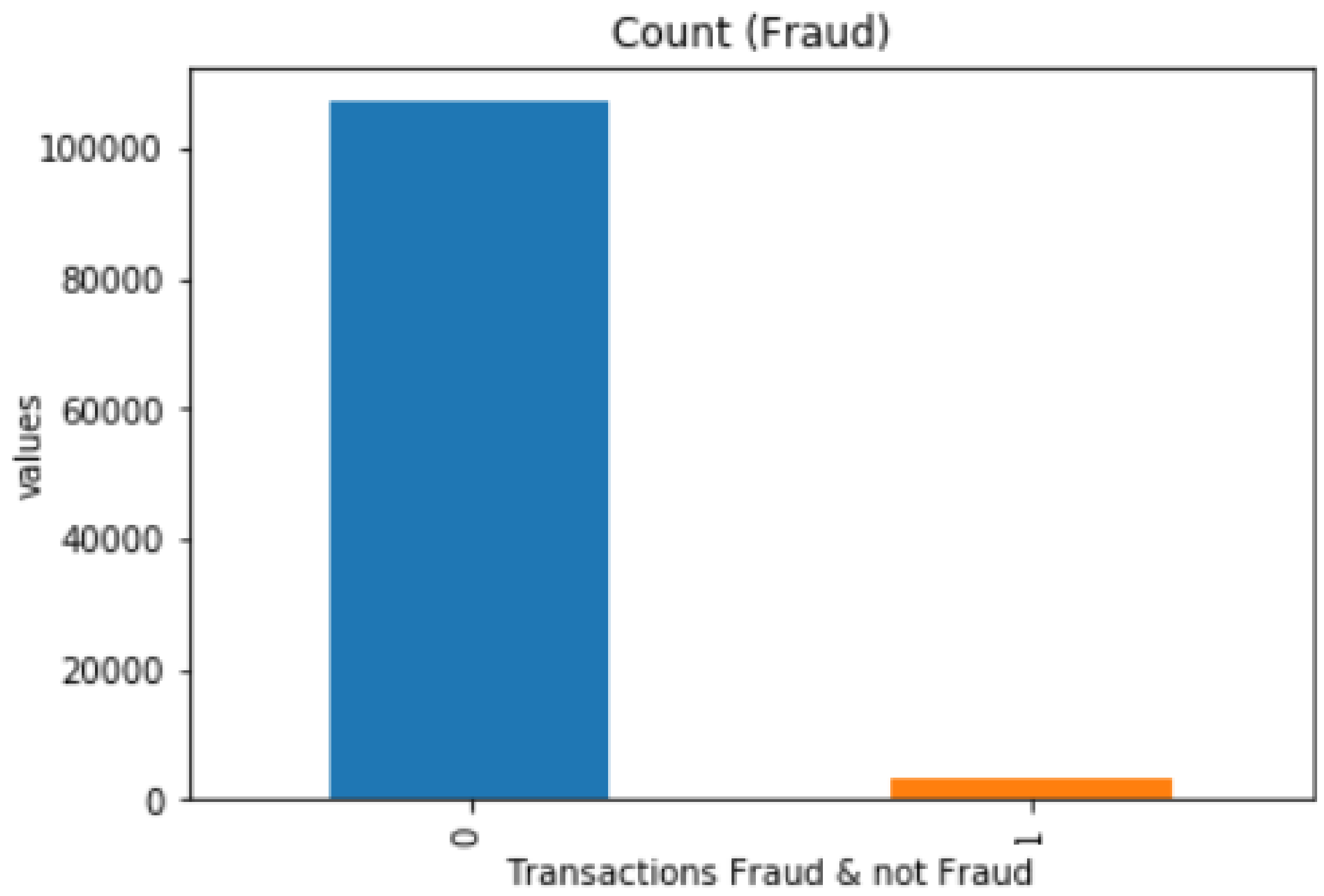
https://www.researchgate.net/figure/Calculation-of-Precision-Recall-and-Accuracy-in-the-confusion-matrix_fig3_336402347

CLASSIFICATION METRICS

ALGOSOC UOB

Problem with accuracy

- **Accuracy is not suitable in some applications.**
- In text mining, we may only be interested in the documents of a particular topic, which are only a small portion of a big document collection.
- In classification involving skewed or highly imbalanced data, e.g., network intrusion and financial fraud detections, we are interested only in the minority class.
  - High accuracy does not mean any intrusion is detected.
  - E.g., 1% intrusion. Achieve 99% accuracy by doing nothing.
- The class of interest is commonly called the **positive class**, and the rest **negative classes**.



https://www.researchgate.net/figure/Class-distribution-before-sampling_fig1_364111446

CLASSIFICATION METRICS

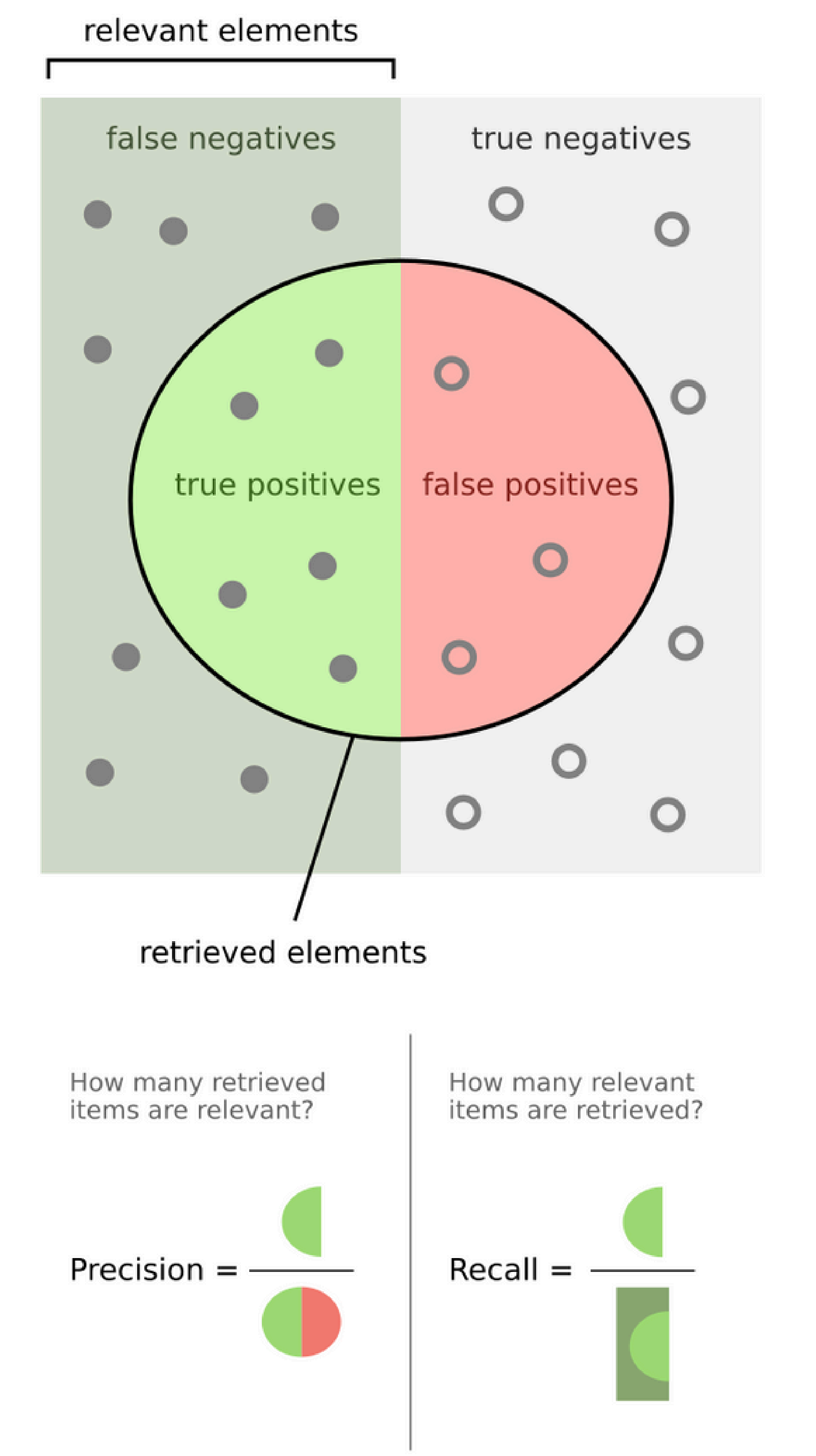
In multi-class scenario especially, we would be concerned in using precision and recall

- Used in information retrieval and text classification.

$$p = \frac{TP}{TP + FP}. \qquad r = \frac{TP}{TP + FN}.$$

- Precision p is the number of correctly classified positive examples divided by the total number of examples that are classified as positive.
- Recall r is the number of correctly classified positive examples divided by the total number of actual positive examples in the test set.

CS583, Bing Liu, UIC 27



https://en.wikipedia.org/wiki/Precision_and_recall

CLASSIFICATION METRICS

Example

	Classified Positive	Classified Negative
Actual Positive	1	99
Actual Negative	0	1000

- This confusion matrix gives
  - precision p = 100% and
  - recall r = 1%

    because we only classified one positive example correctly and no negative examples wrongly.
- Note: precision and recall only measure classification on the positive class.

CLASSIFICATION METRICS

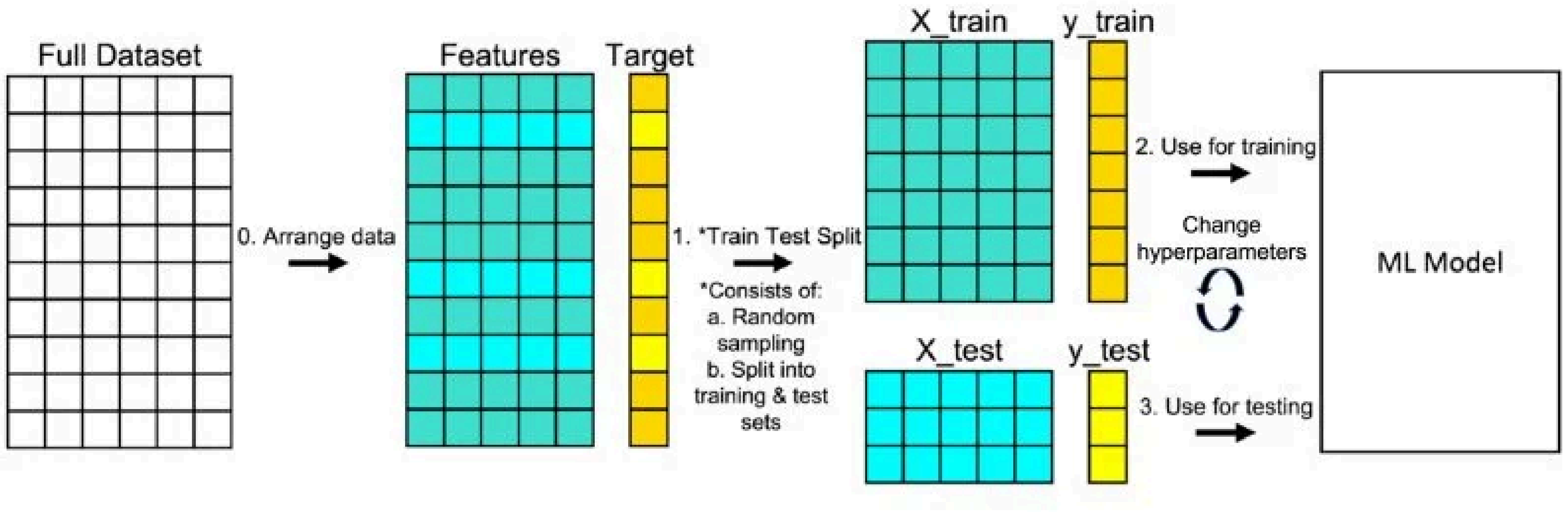
Measuring precision and recall at the same time is hard

Therefore, we use F Values or (F1-Score) to represent precision and recall in equal manner.

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$$

TRAIN AND TEST

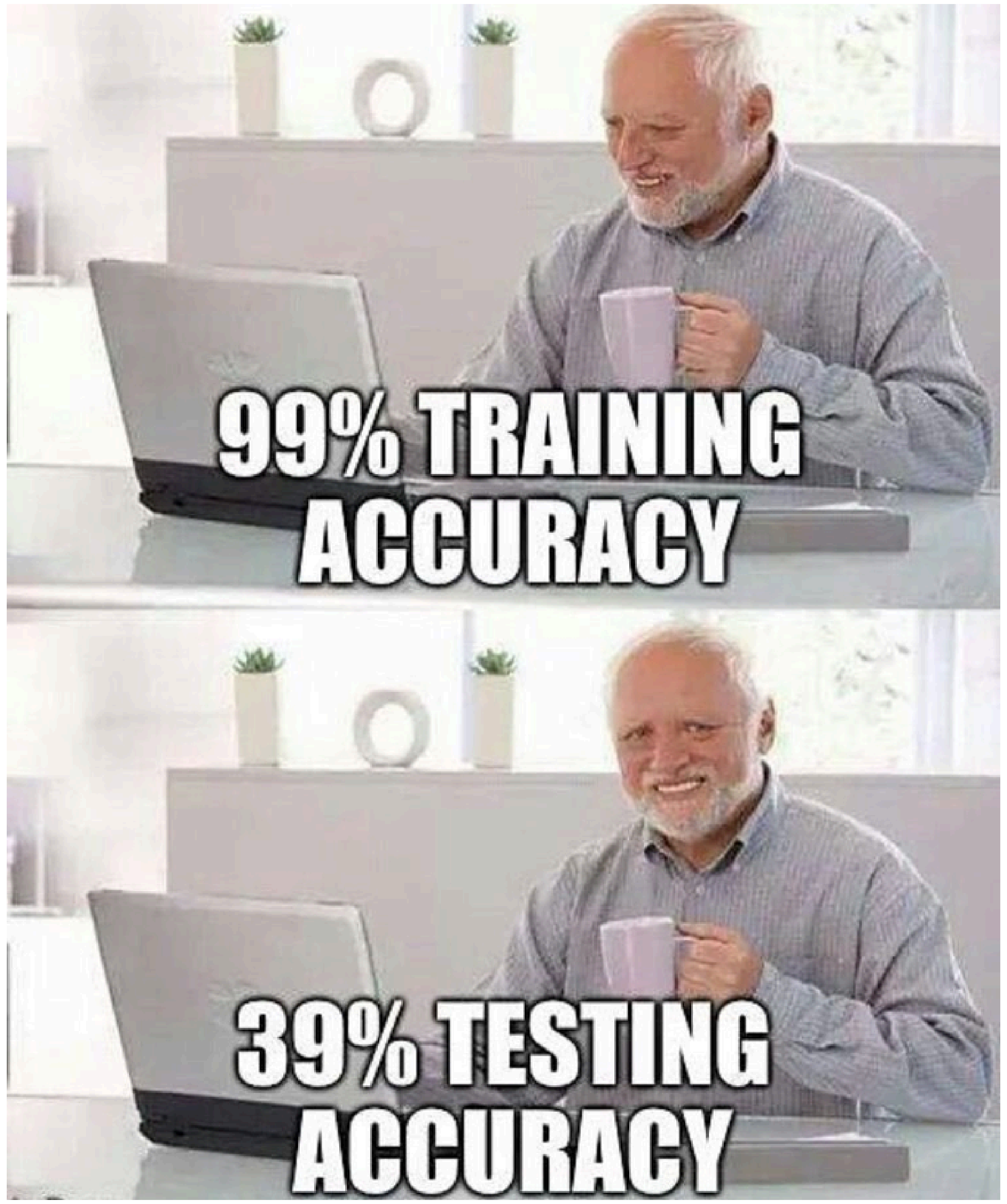
Remember:
We want to measure the model's performance on **unseen data (i.e., data not used on training)**



https://builtin.com/data-science/train-test-split

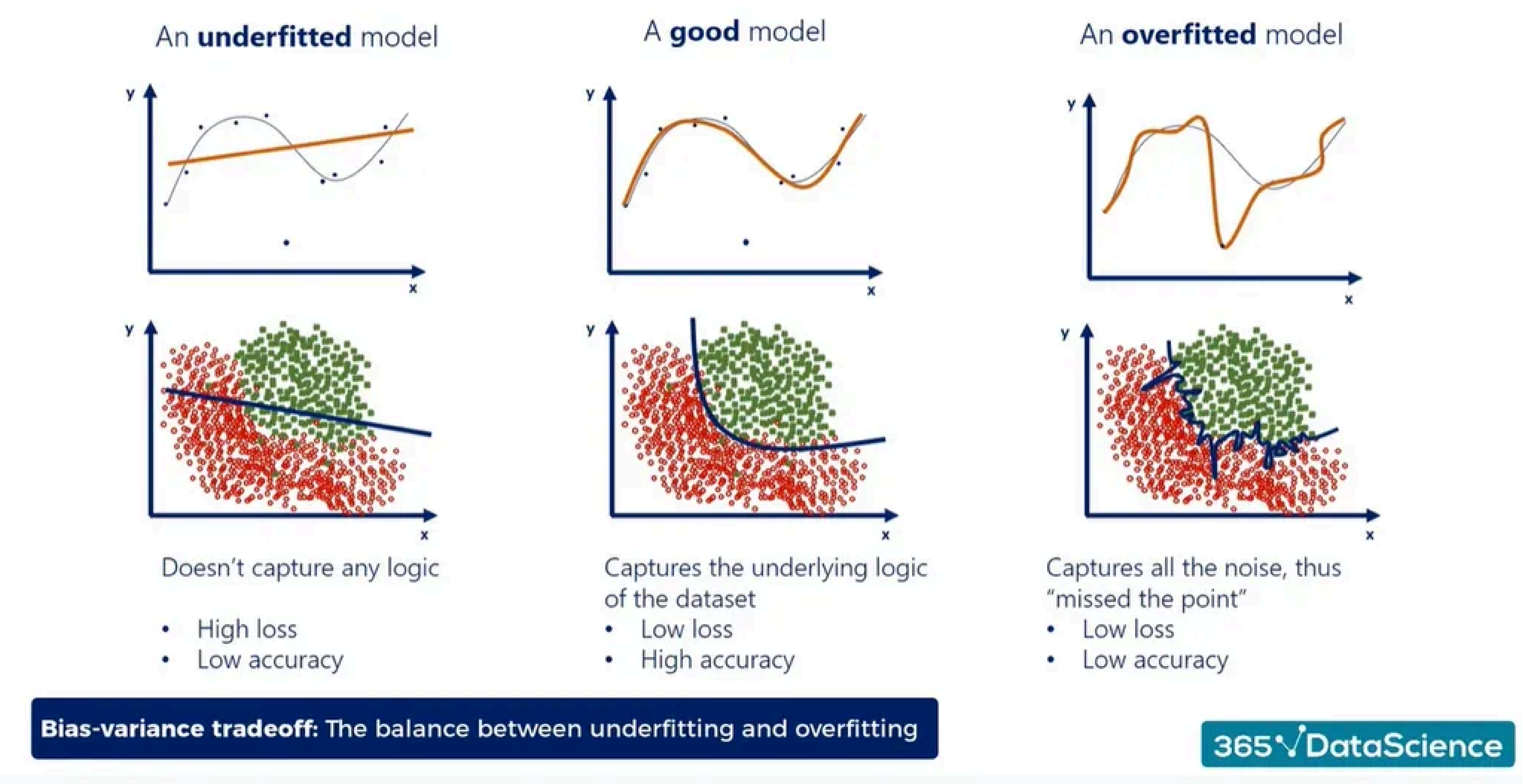
The uncertainty measurment usually would be evaluated against the model's performance with the test data, **NOT** the training data

OVERFITTING AND UNDERFITTING



https://x.com/MaartenvSmeden/status/1522230905468862464

OVERFITTING AND UNDERFITTING



https://365datascience.com/tutorials/machine-learning-tutorials/overfitting-underfitting/

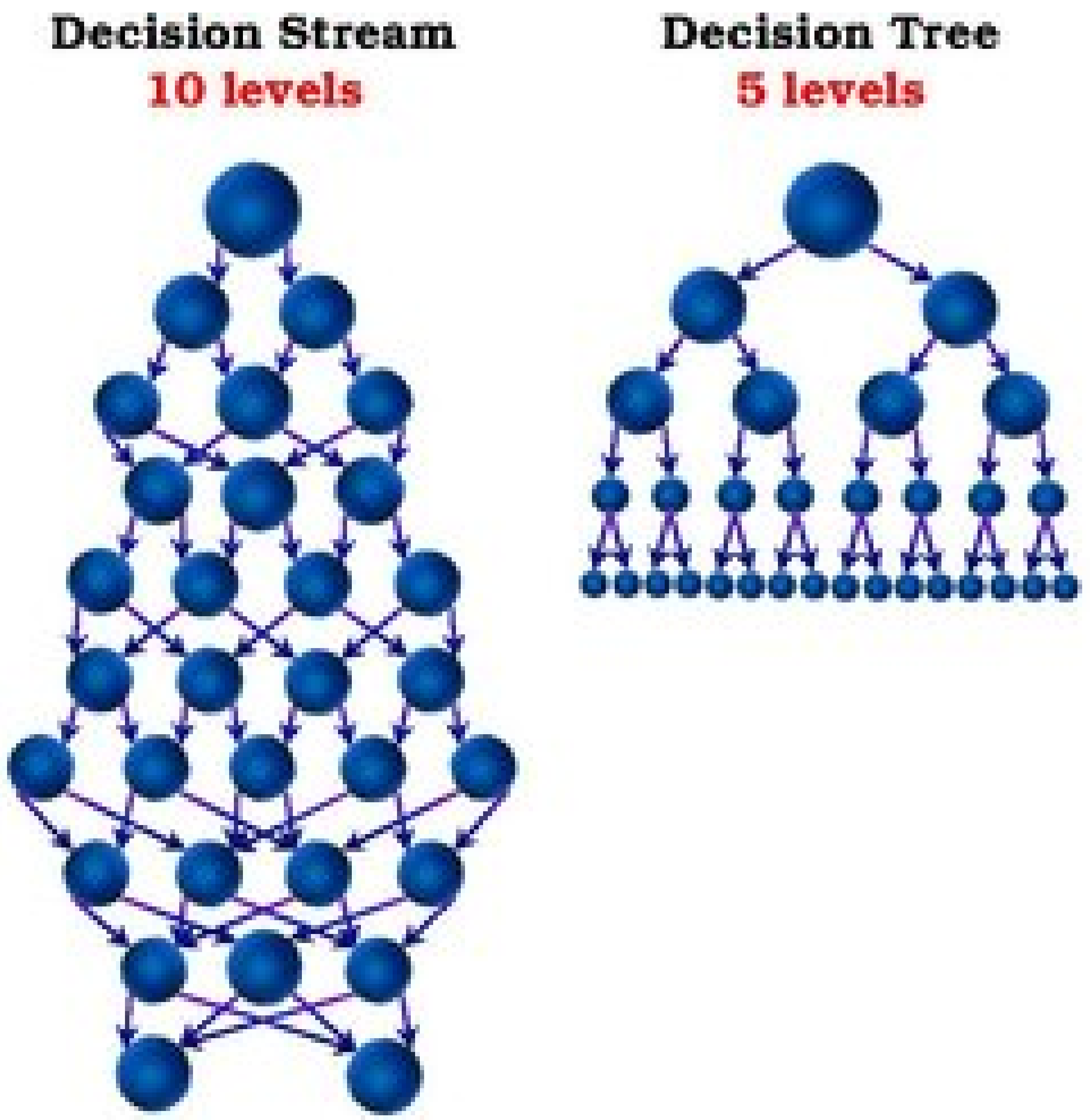
WHY MODEL SELECTION

Sometimes, we may achieve better result with different model settings. For example, we might be able to fit the training data well into a decision tree with more depth.

However, how does that actually perform?

We don’t want the model to overfit on the training example and cannot generalise on the testing example

We need model selection procedures for validation and testing



MODEL SELECTION

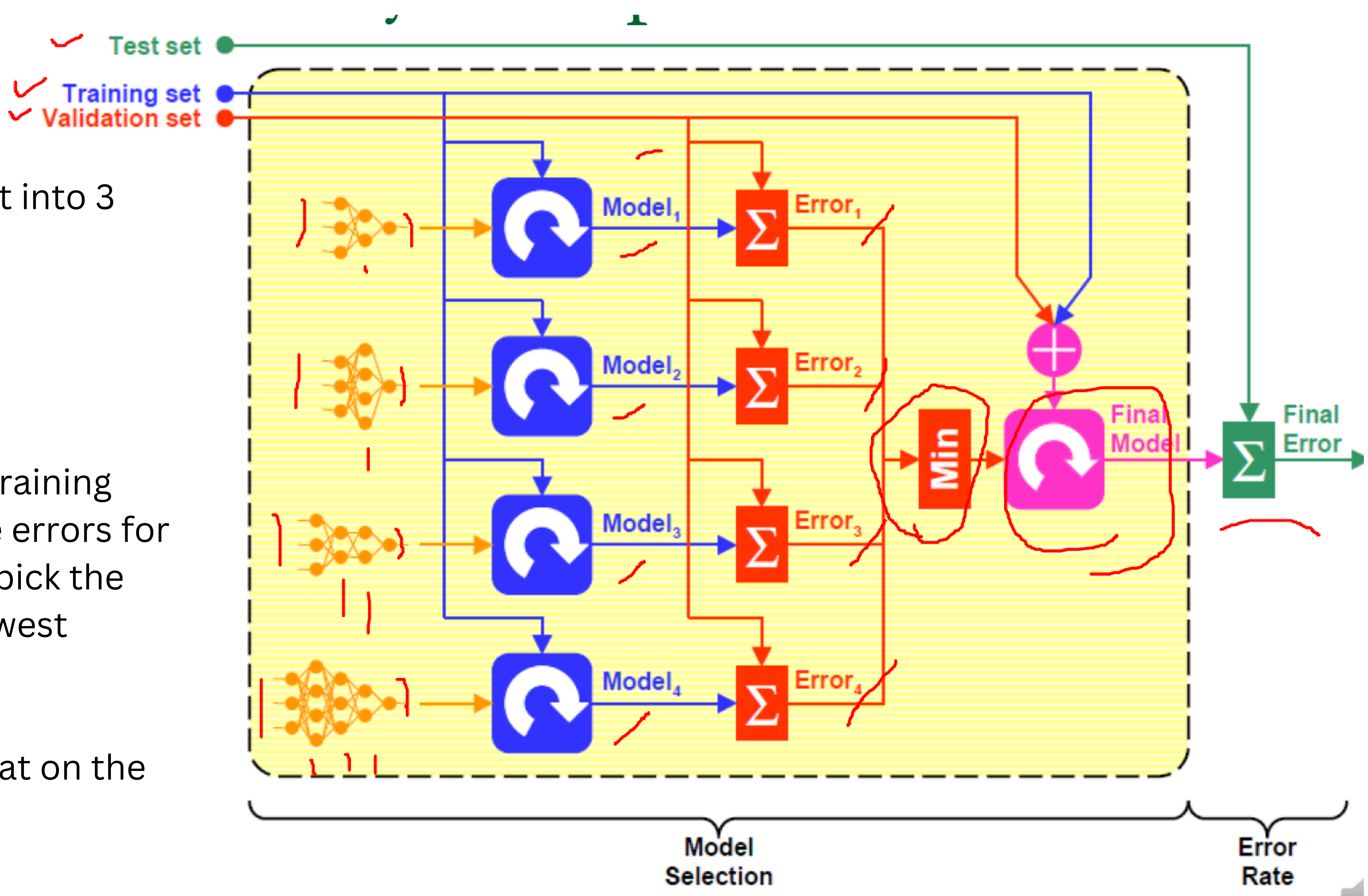
ALGOSOC UOB

We can partition our dataset into 3 sets:

- Training
- Validation
- Testing

Validation sets are unseen training sets, being used to estimate errors for model selection. We would pick the best model based on the lowest validation set error

Furthermore, we will test that on the test set.



MODEL SELECTION

There are several methods on hyperparamter tuning:

- Grid Search: We create a grid of selectable parameter configurations and use exhaustive search to find the best-performing model
- Random Search: Using the same grid, we use a randomised non-exhaustive method. Might be suboptimal, but would still let us search for a better configuration
- Bayesian Optimisation: The hyperparameters are treated based on probabilistic values, making us have a set of belief on how likely they would come up.
- And many more

Recommended reading: https://en.wikipedia.org/wiki/Hyperparameter_optimization

HANDS ON

Link
https://bit.ly/AlgosocWk5

Colab