Cross Validation

If a student keeps practicing only **one set of practice questions and scores full marks**, are they really well-prepared for the exam?

Answer: **No**

The student needs to practice on various types/sets of question.

That's exactly what happens when a model is tested on the same data it learned from.

Cross Validation

Cross-validation is a **model evaluation technique** used in machine learning to check how well your model will perform on *new, unseen data*.

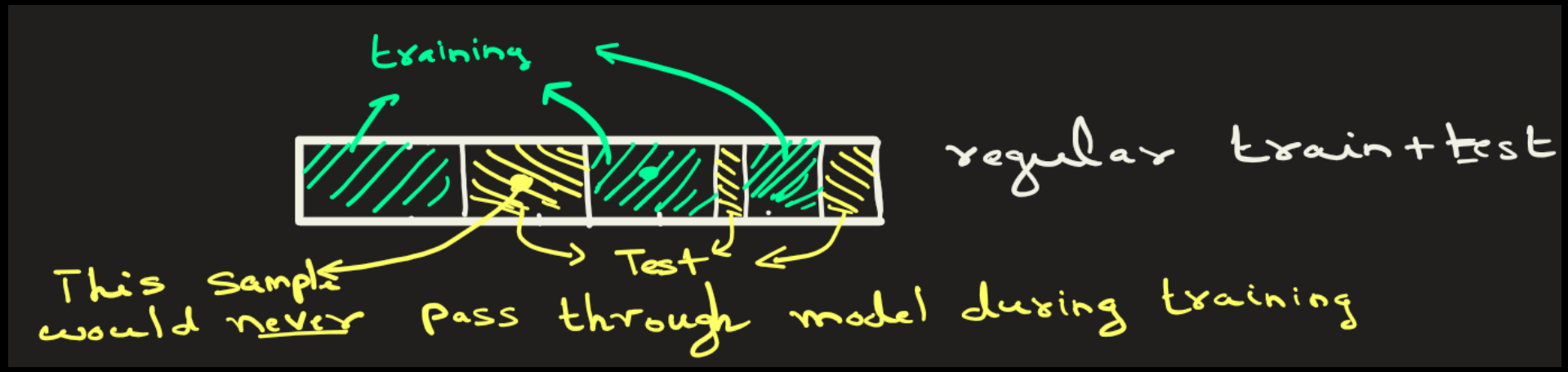
Instead of testing the model on just one fixed train–test split, cross-validation **tests the model multiple times on different data splits**, giving a more reliable performance estimate.

Cross Validation: Algorithm

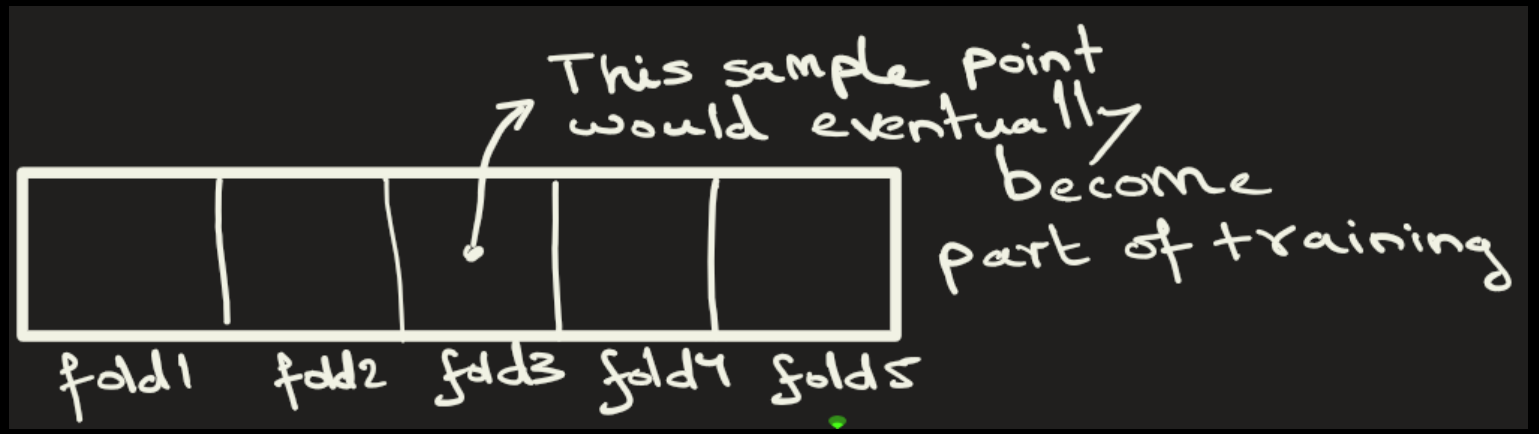
1. The dataset is divided into multiple parts (called **folds**).
2. The model is trained on some folds.
3. It is tested on the remaining fold. The score is noted
4. This process is repeated multiple times until it is evaluated on the entire dataset
5. Final performance = **average of all test scores**

Cross Validation

In traditional splitting, there would be points in test dataset that would model would never see during training.



In CV, each data point is used **for training** and also, **for testing exactly once.** The model sees all the points.



Cross Validation: Types

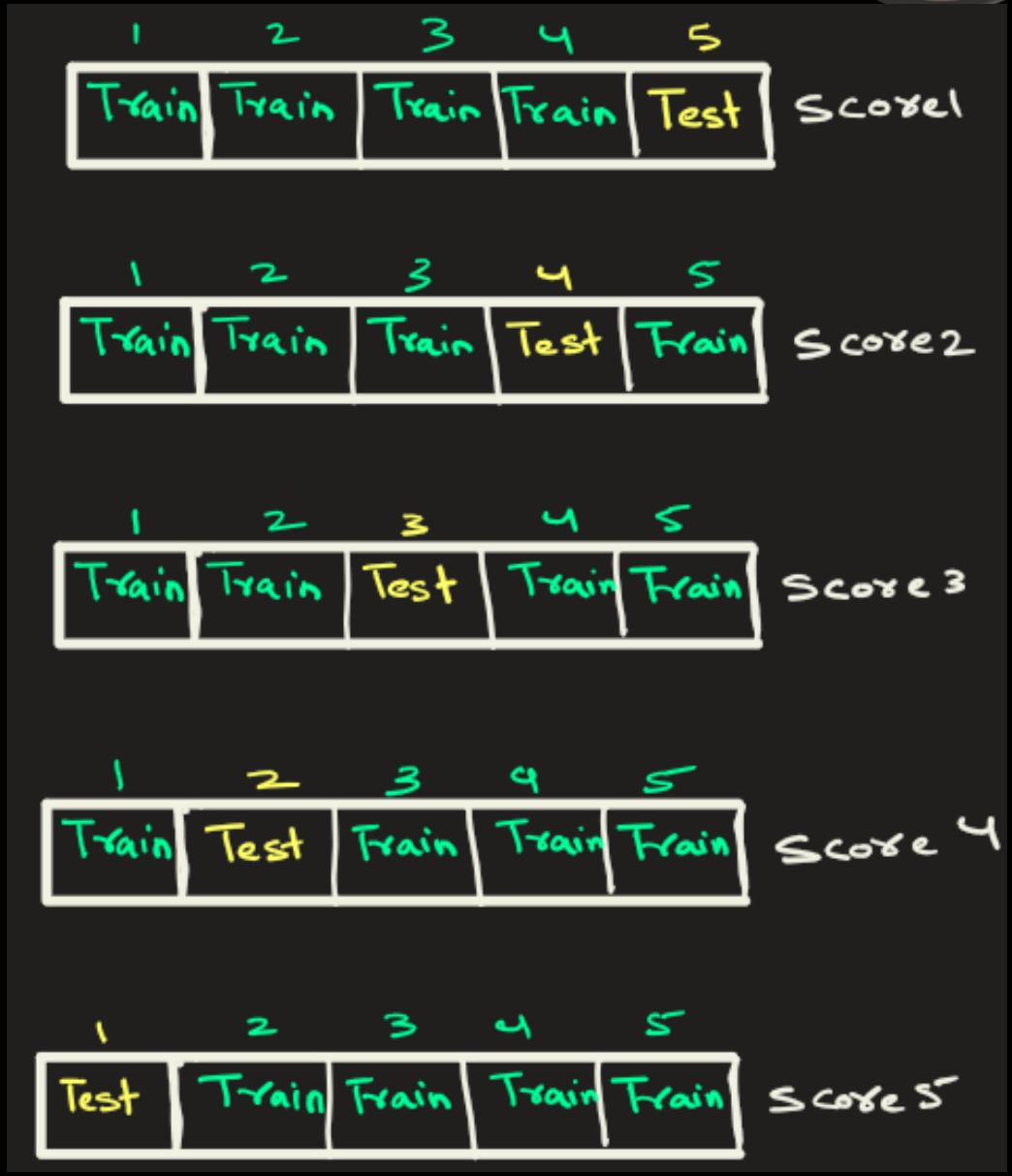
1. K-Fold Cross-Validation

Most widely used because it works for most ML problems
Typical values: k = 5 or k = 10

Example: 5-Fold Cross-Validation
Your dataset is split into **5 equal parts**:

Iteration	Training Data	Testing Data
1	Folds 1-4	Fold 5
2	Folds 1,2,3,5	Fold 4
3	Folds 1,2,4,5	Fold 3
4	Folds 1,3,4,5	Fold 2
5	Folds 2-5	Fold 1

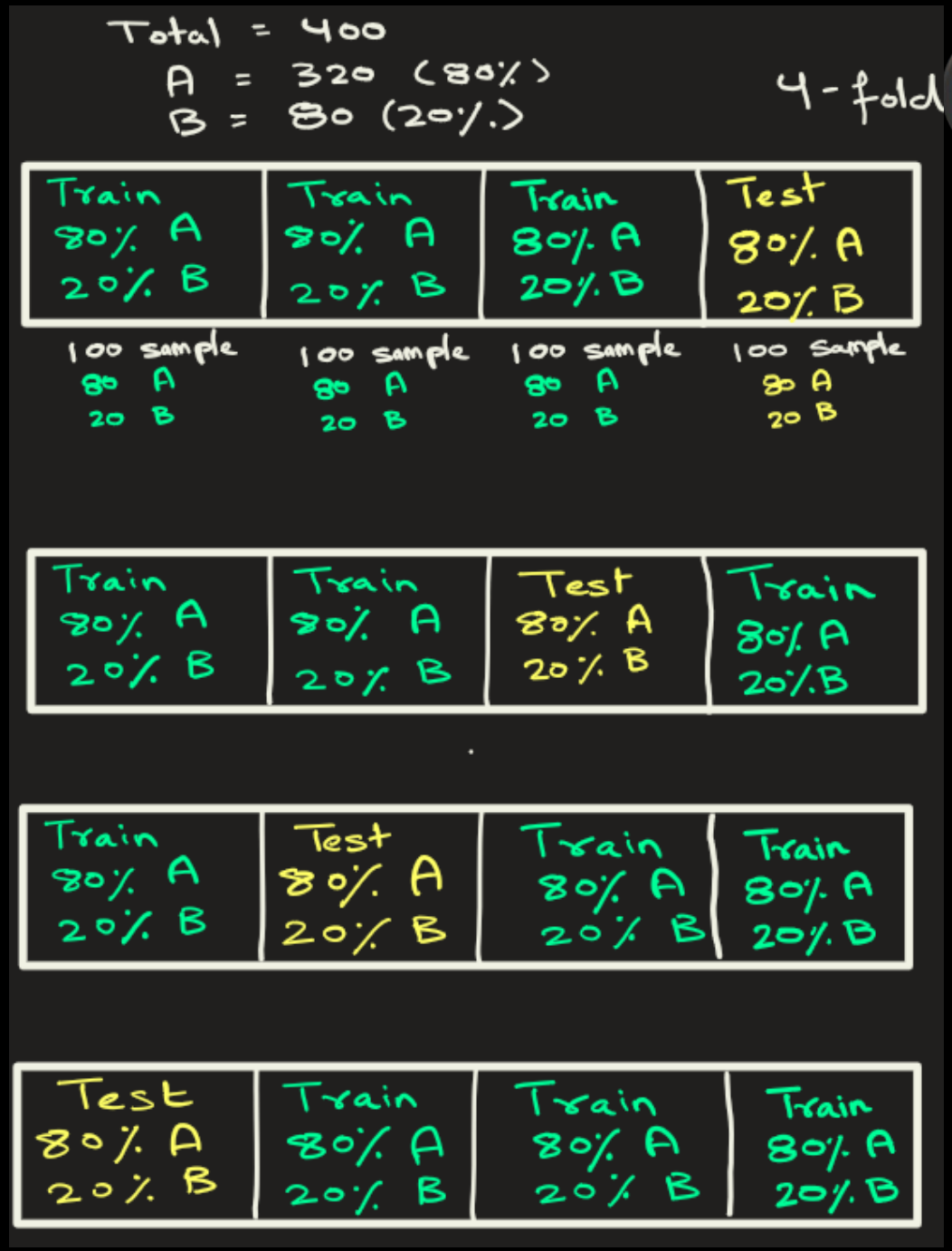
Final Accuracy = Average of all 5 test accuracies.



2. Stratified K-Fold (For Classification)

Maintains **class proportions** in each fold.
Very important for **imbalanced datasets**

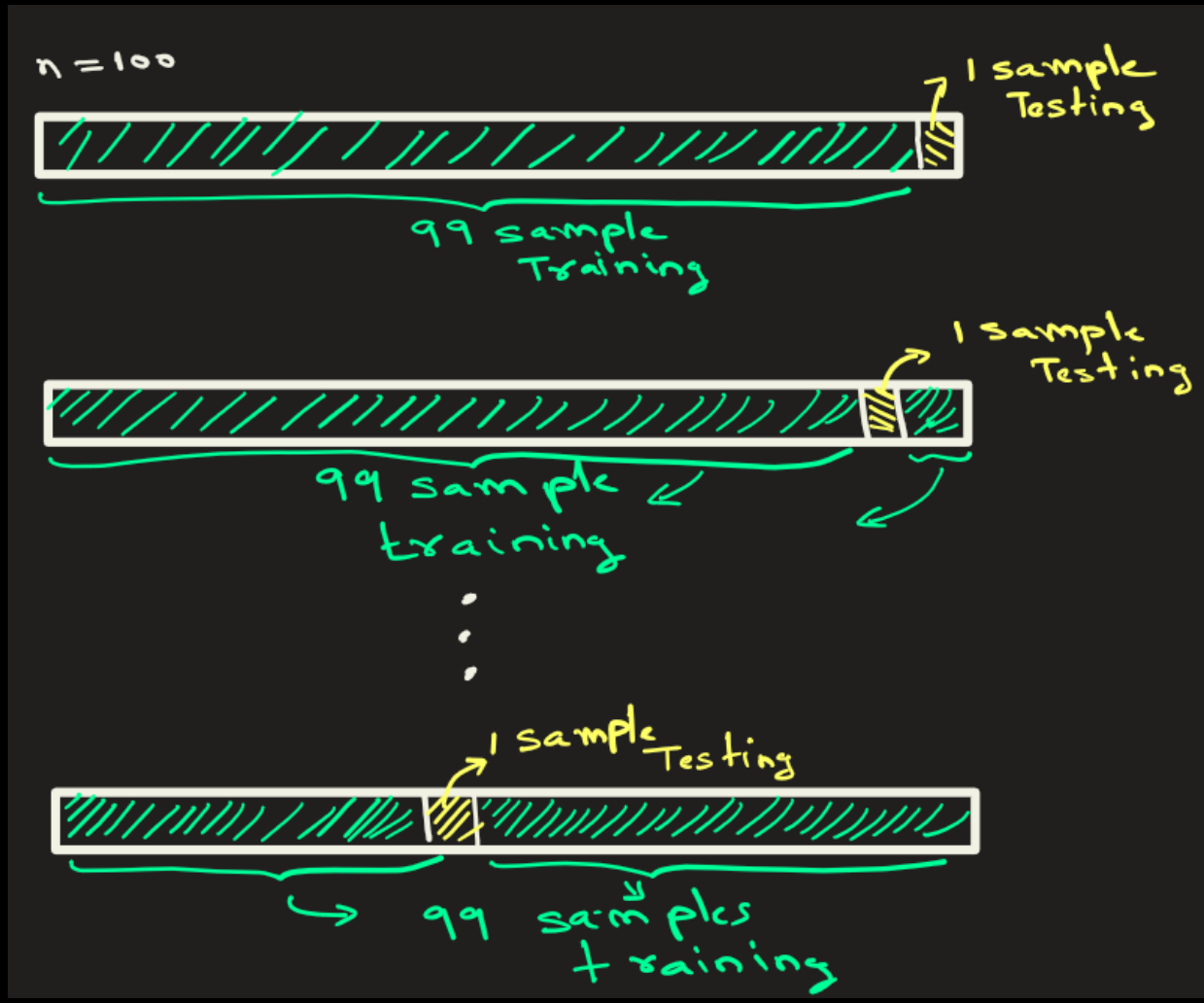
Example: Cancer detection, fraud detection



Cross Validation

3. Leave-One-Out Cross-Validation (LOOCV)

Each sample is used once as test data
Very accurate but **very slow**
Used only for **very small datasets**



Example:
Consider a dataset with N=100 points. 100-1=99 will be the training set and 1 point will be the testing set.
Another point will be chosen as the testing data and the rest of the points will be training.
This will repeat for the rest of the dataset, i.e.: N=100 times.
The **final performance measure will be the average of the measures for all N=100 iterations**

Cross Validation

4. Time Series Cross-Validation

Used when data has a **time order**
NO random shuffling
Trains on **past** → tests on **future**

Used in: Stock prediction, sensor data, weather forecasting

Cross Validation

Why Do We Need Cross-Validation?

If you train on one dataset split and test on only one test set:

- Your accuracy may be **too optimistic** and **m**odel may be **overfitting**
- Result depends heavily on **how the data was split**

Cross-validation solves this by using multiple train–test combinations.

Cross Validation

Disadvantages

- More computationally expensive
- Slower for large datasets
- Not suitable for real-time training

Tip:
Avoid CV for large datasets and live streaming data

EXTRA

Cross Validation

Advantages of Cross-Validation

- Better estimate of model performance
  Reduces overfitting risk
  Uses the full dataset efficiently
  Helps compare multiple models fairly

Heading Goes Here

Fhdsklf
Fjdsklf
Fjskldf