K-Fold Cross Validation

-MUKESH KUMAR

- K-fold cross-validation is a robust technique used in machine learning to evaluate the performance of a model.

- It helps ensure that your model generalizes well to unseen data by mitigating overfitting or underfitting.

What is K-fold Cross-Validation?

- K-fold cross-validation involves splitting your dataset into k smaller subsets, or "folds."
- The model is trained k times, each time using a different fold as the validation set while the remaining k-1 folds are used for training.
- The final performance metric is the average of the metrics from each fold.

Steps in K-fold Cross-Validation

Step1: Split the Dataset:

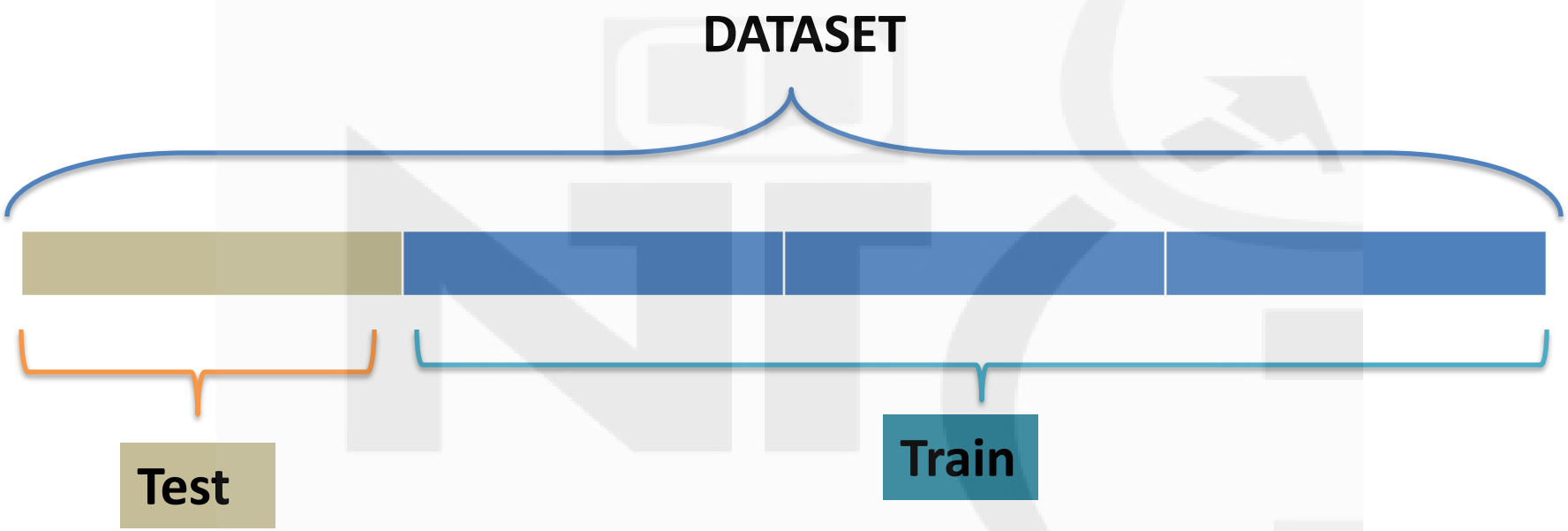
- Divide your dataset into k equally (or nearly equally) sized folds.
- For example, if k=5, the data is split into 5 parts.

Step2:Train and Validate the Model:

- For each fold:
  - Use the first fold as the validation set.
  - Use the remaining k-1 folds as the training set.
  - Train the model on the training set.
  - Evaluate the model on the validation set and record the performance metric (e.g., accuracy, precision, recall).

Step3: Calculate the Average Performance:

- After all k iterations, calculate the mean of the recorded performance metrics.
- This average performance metric is used as the final estimate of the model's performance.



Once the model is trained its evaluated on test set

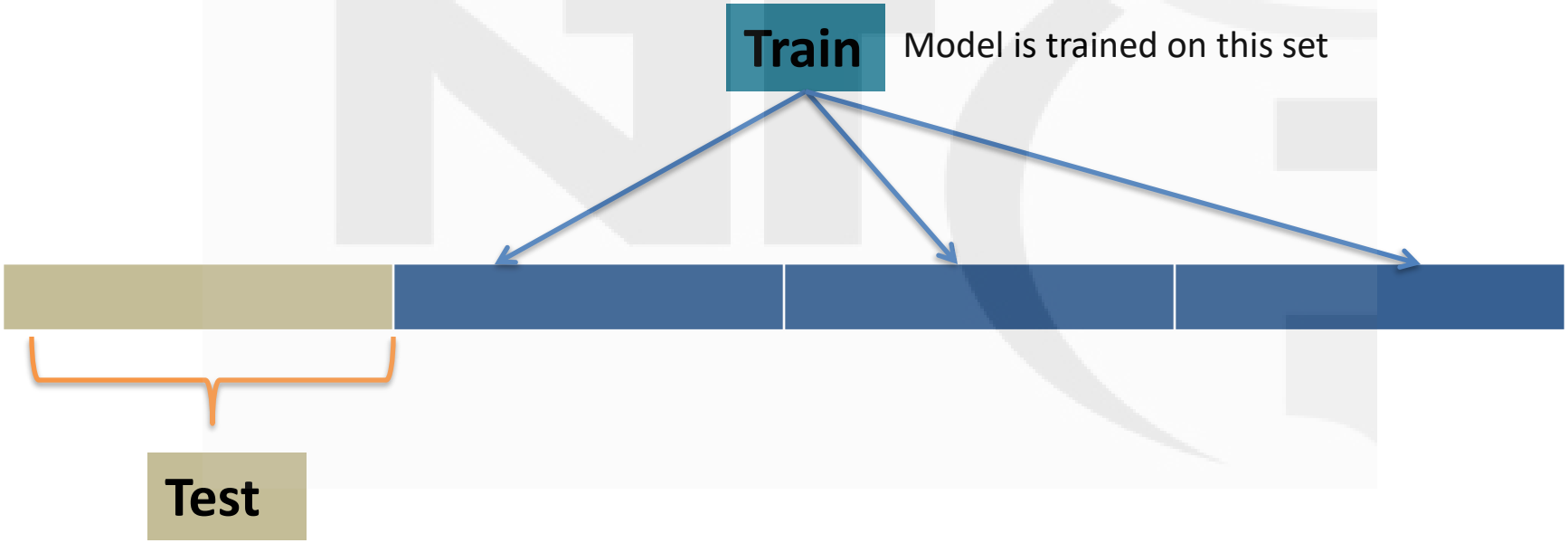
Model is trained on training set

- In K-fold cross validation data is split into k-folds
- Model is trained k-times, each time on different folds

Iteration1:

K=4

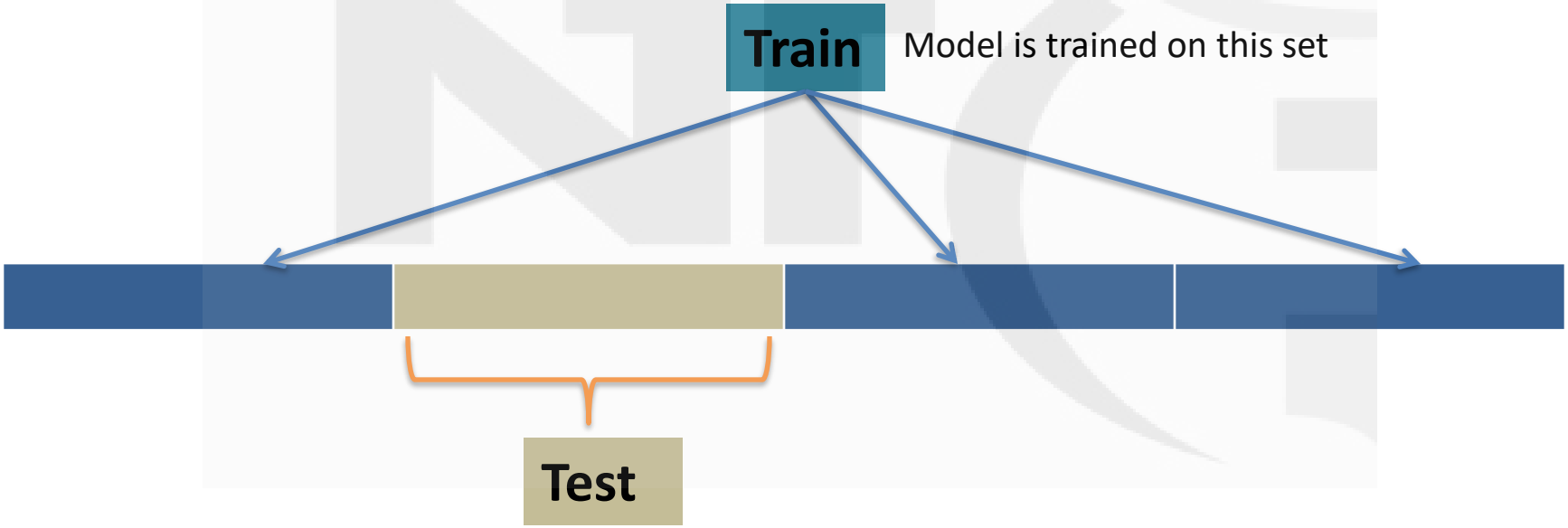
- first fold – test set
- remaining 3 – training set



Iteration2:

K=4

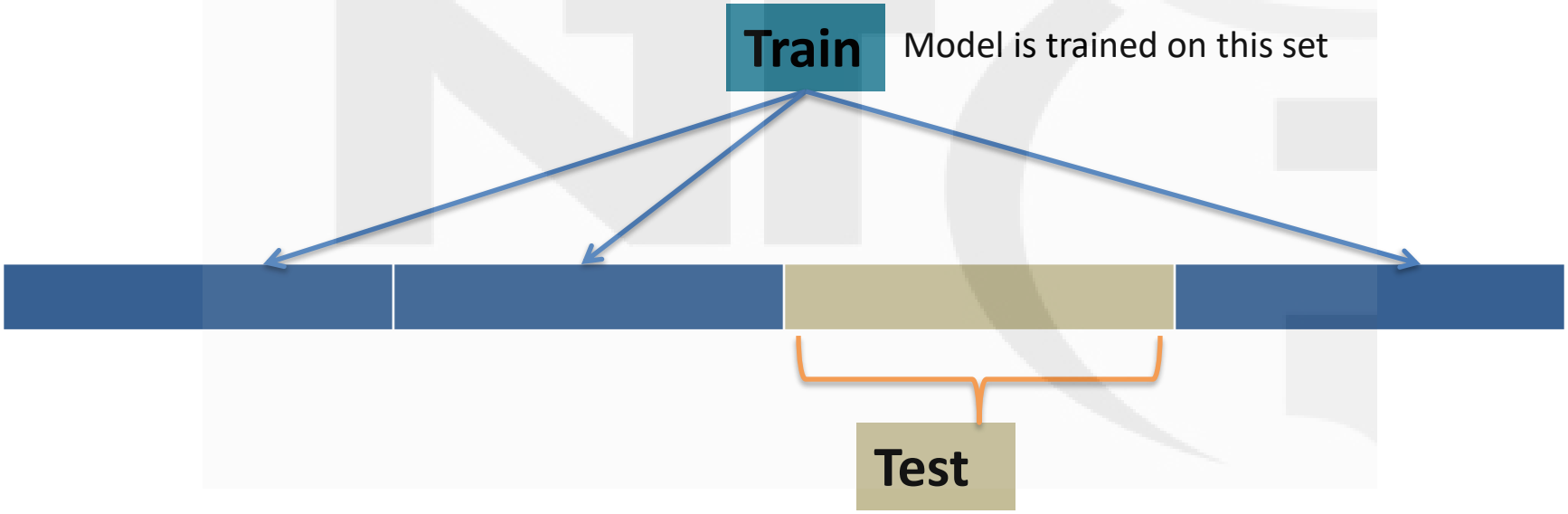
- second fold – test set
- remaining 3 – training set



Iteration3:

K=4

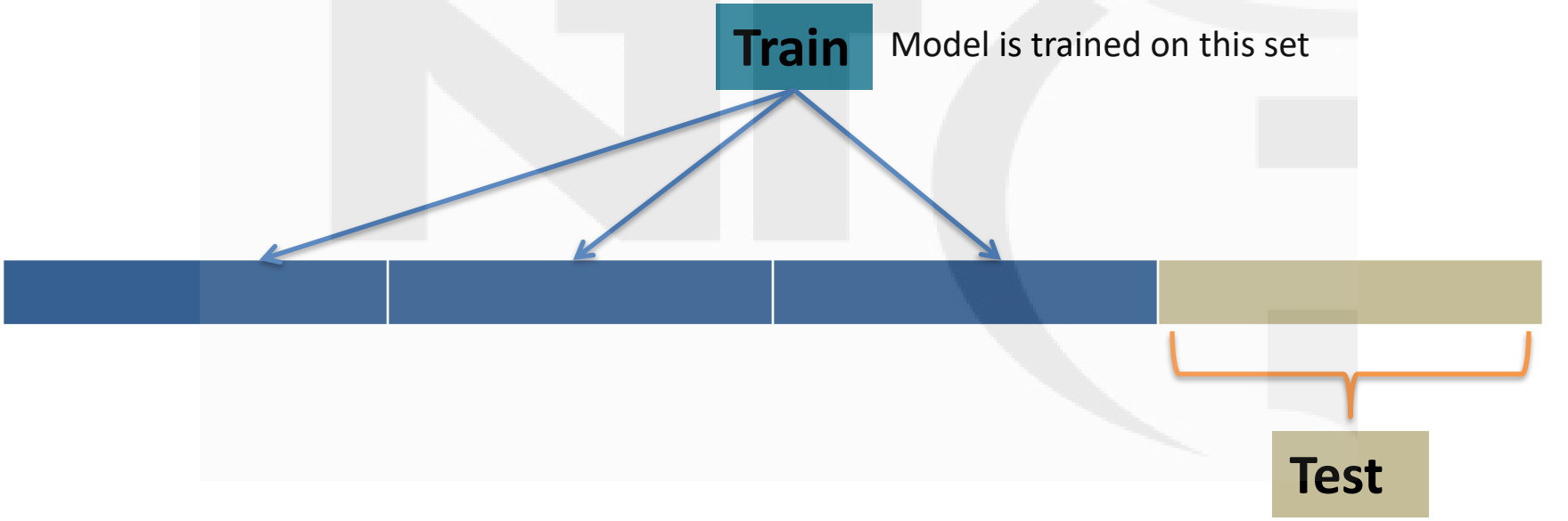
- third fold – test set
- remaining 3 – training set



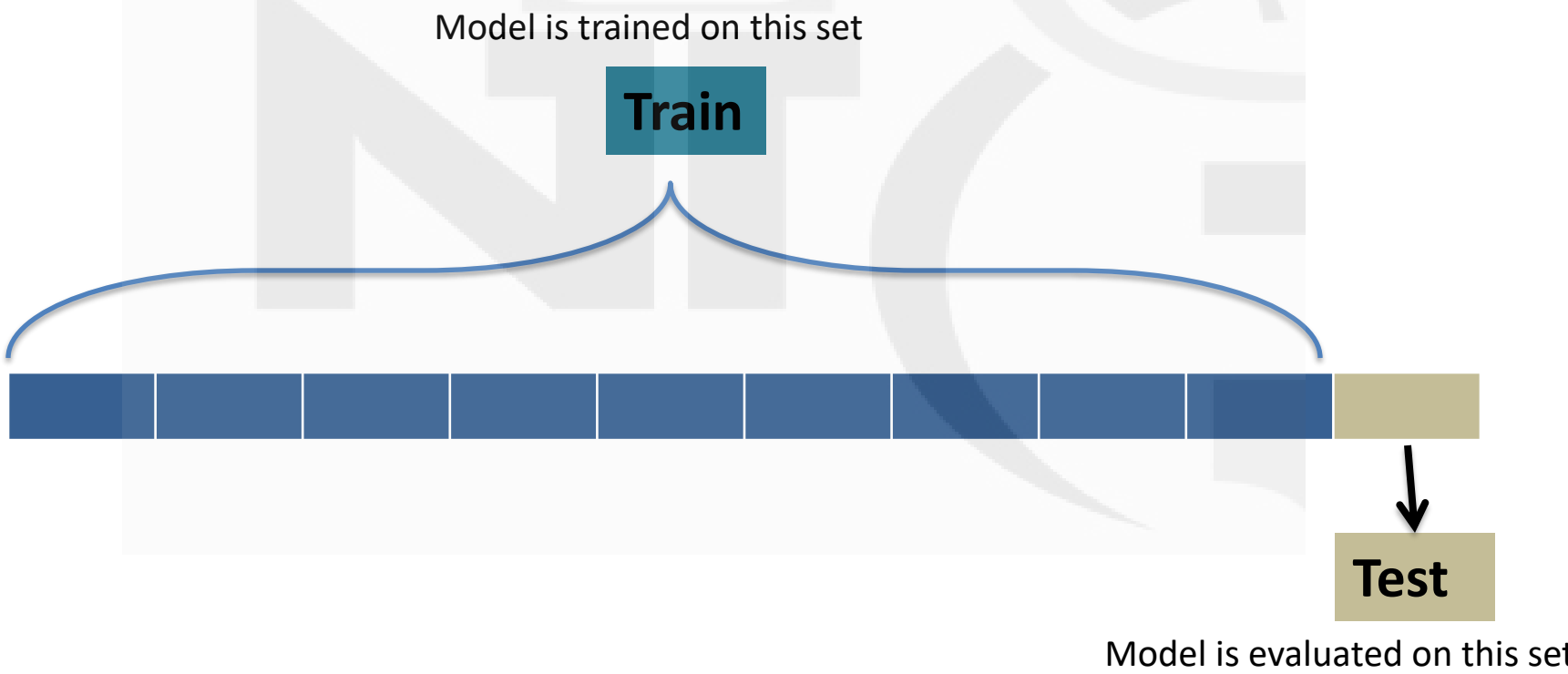
Iteration4:

K=4

- fourth fold – test set
- remaining 3 – training set



For K=10 there will be 10 iterations



Advantages of K-fold Cross-Validation

- **More Reliable Performance Metric**: By evaluating the model on multiple subsets of data, it provides a more reliable estimate of model performance.

- **Efficient Use of Data**: All data points are used for both training and validation, maximizing the amount of data available for model learning.

Choosing k

- Common choices for k are 5 or 10.

- A smaller k (like 5) results in a larger training set per iteration, while a larger k (like 10) gives a better estimate of model performance but is computationally more expensive.

Stratified K-Fold Cross-Validation

- If you have imbalanced classes, consider using **Stratified K-Fold Cross-Validation**.
- It ensures that each fold has a similar distribution of classes, which helps in producing a more accurate performance estimate

Summary

- K-fold cross-validation is a powerful tool to assess how well a model will perform on unseen data.
- By repeating the training and validation process multiple times, it helps to identify models that generalize well, avoiding the pitfalls of overfitting.