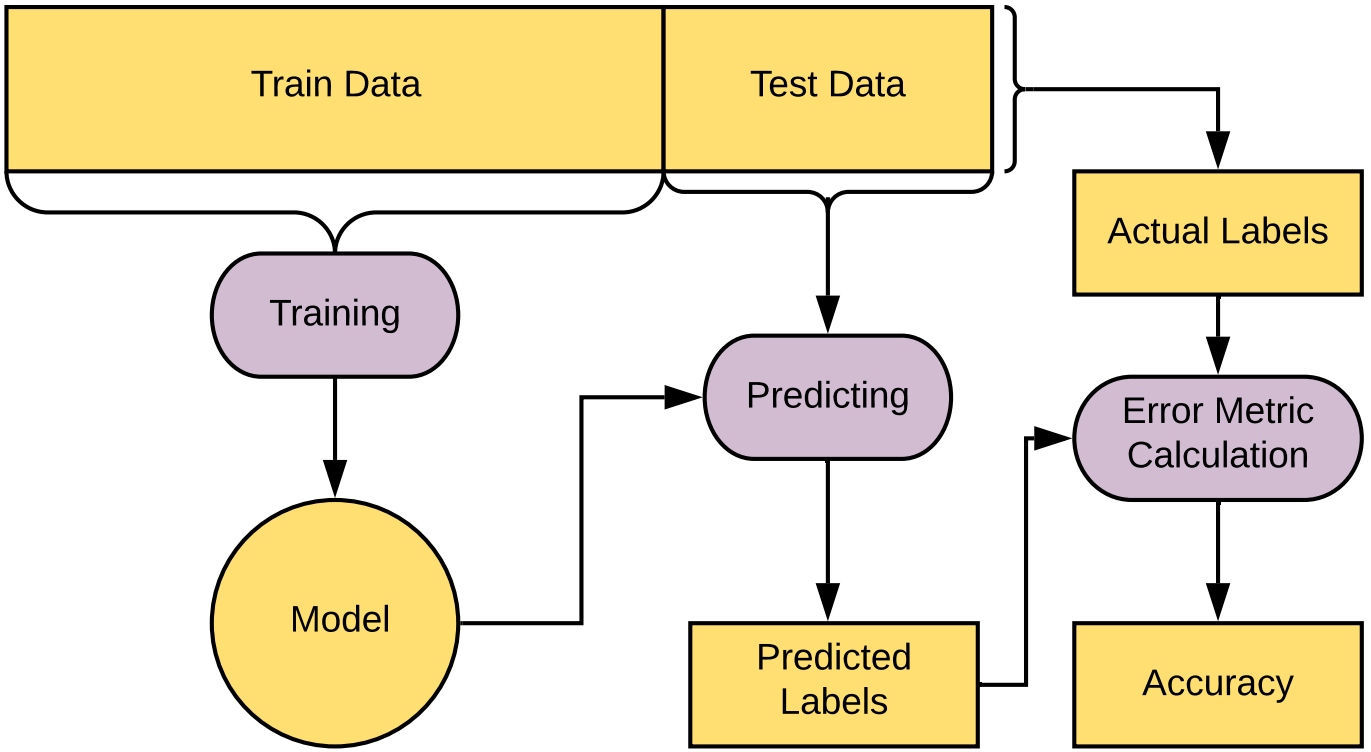
Cross-Validation

Nipun Batra and teaching staff

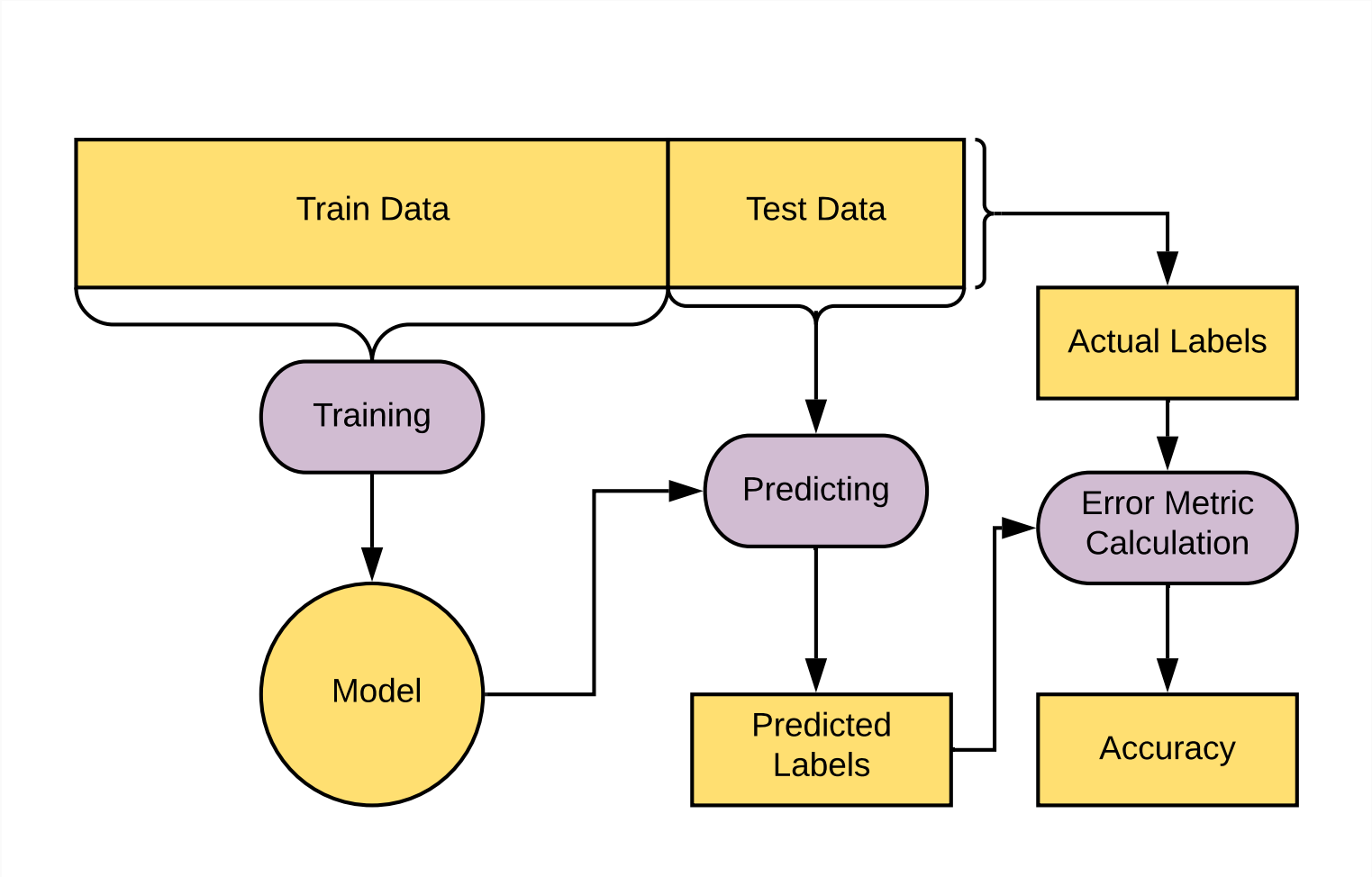
January 12, 2024

IIT Gandhinagar

Our General Training Flow

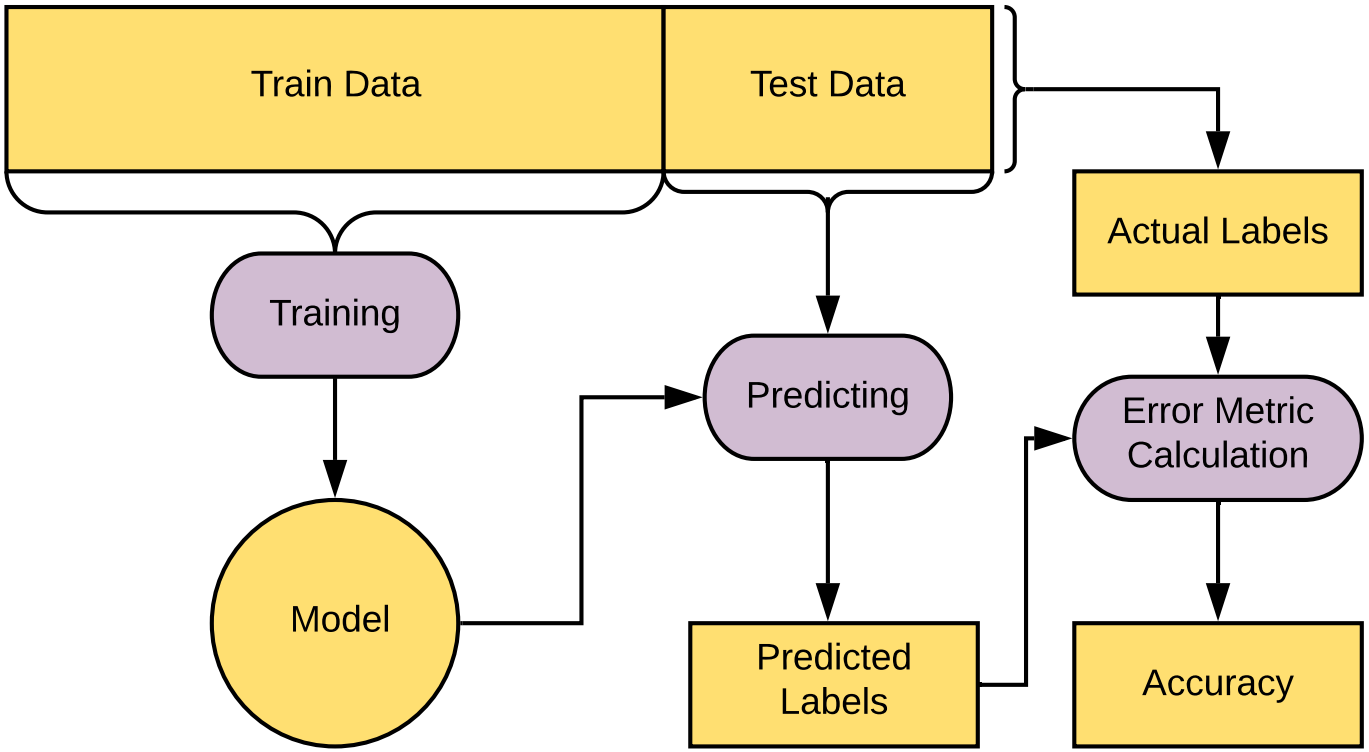


Our General Training Flow



- Does not use the full dataset for training and does not test on the full dataset

Our General Training Flow



- Does not use the full dataset for training and does not test on the full dataset
- No way to optimise hyperparameters

How to use the full dataset for training?

How to use the full dataset for training?

- Over multiple iterations, use different parts of the dataset for training and testing.

How to use the full dataset for training?

- Over multiple iterations, use different parts of the dataset for training and testing.
- Typically done via different random splits of the dataset.

How to use the full dataset for training?

- Over multiple iterations, use different parts of the dataset for training and testing.
- Typically done via different random splits of the dataset.
- Challenge?

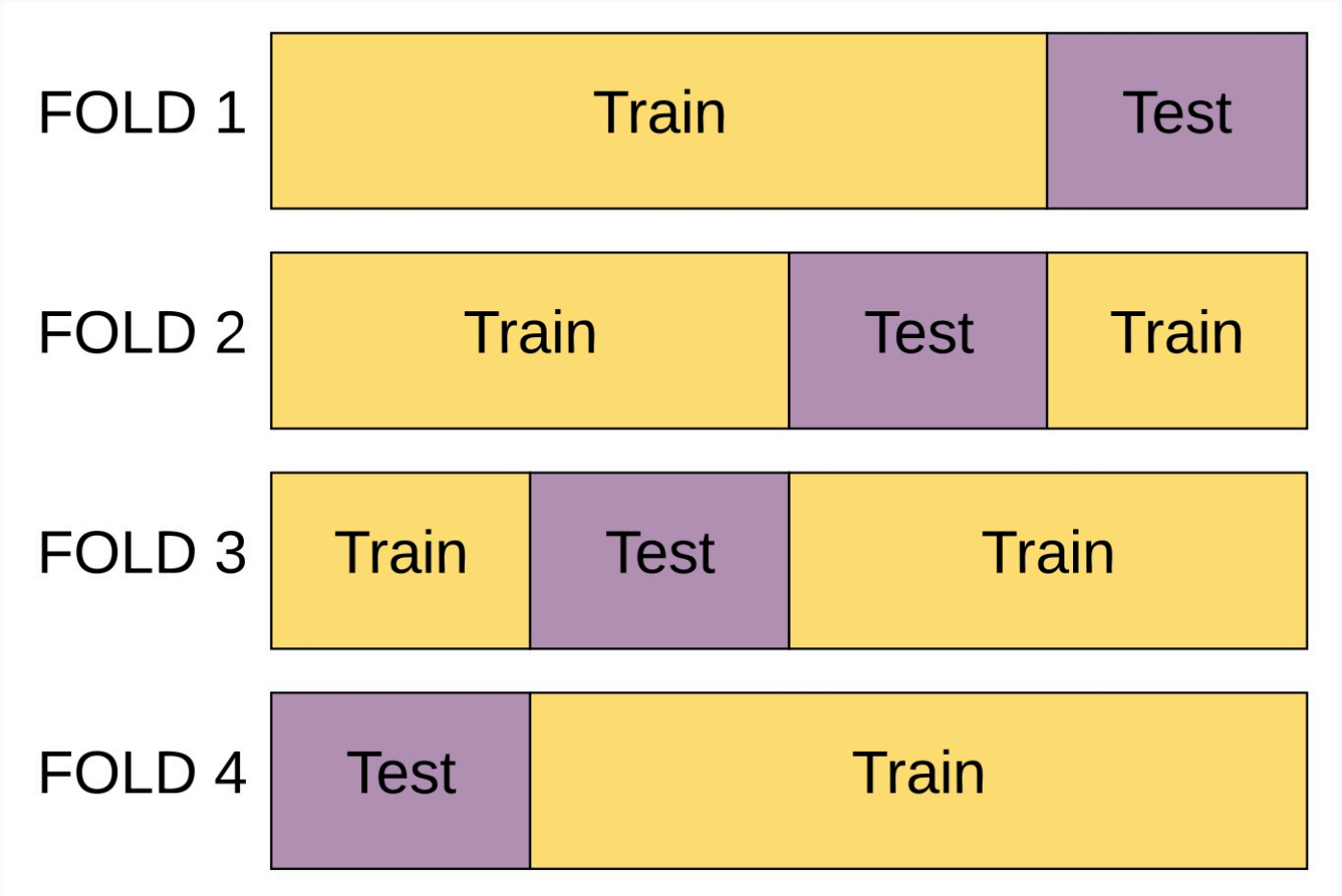
How to use the full dataset for training?

- Over multiple iterations, use different parts of the dataset for training and testing.
- Typically done via different random splits of the dataset.
- Challenge?
- May not use every data point for training or testing

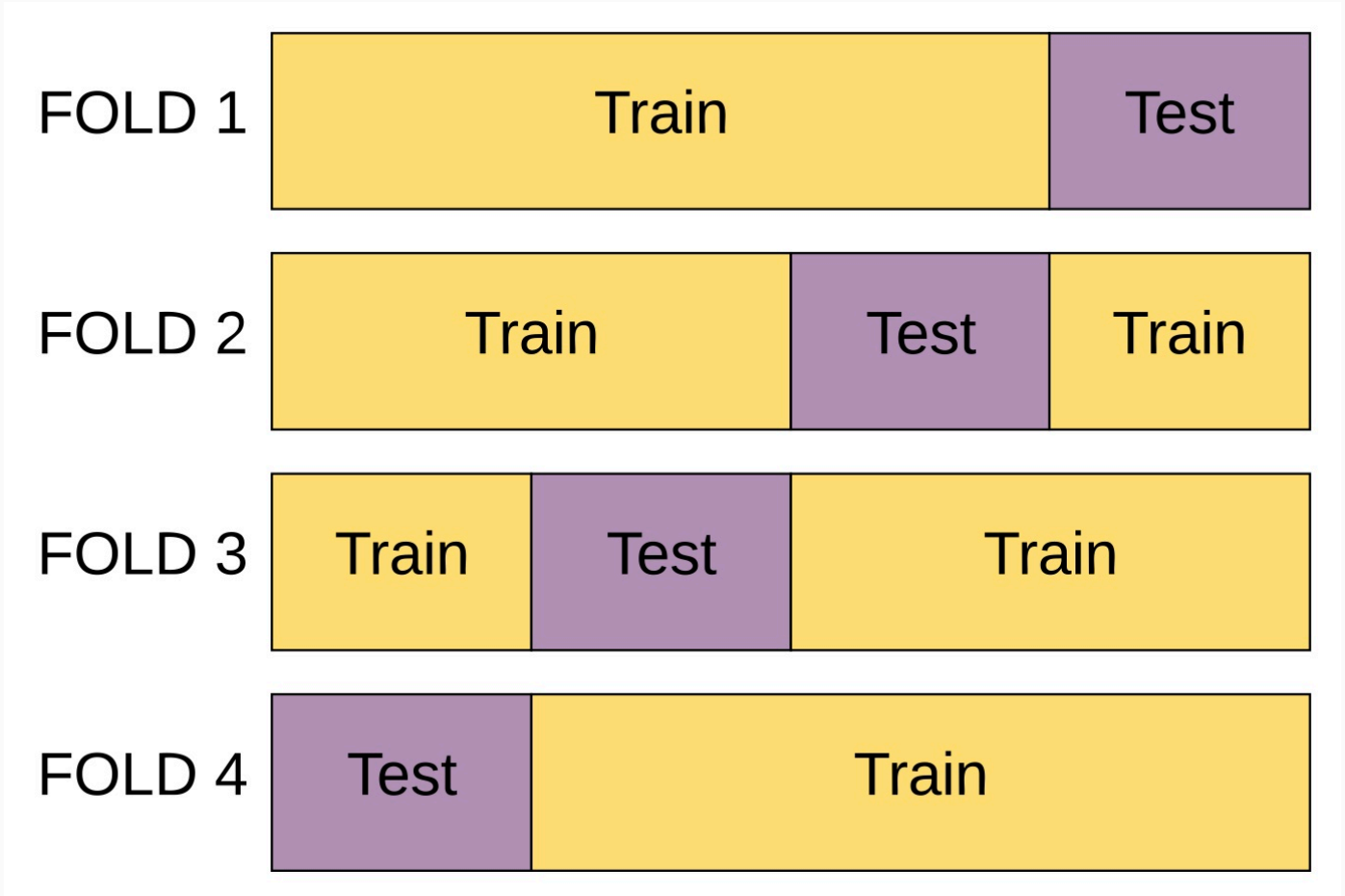
How to use the full dataset for training?

- Over multiple iterations, use different parts of the dataset for training and testing.
- Typically done via different random splits of the dataset.
- Challenge?
- May not use every data point for training or testing
- May be computationally expensive

K-Fold cross-validation: Utilise full dataset for testing

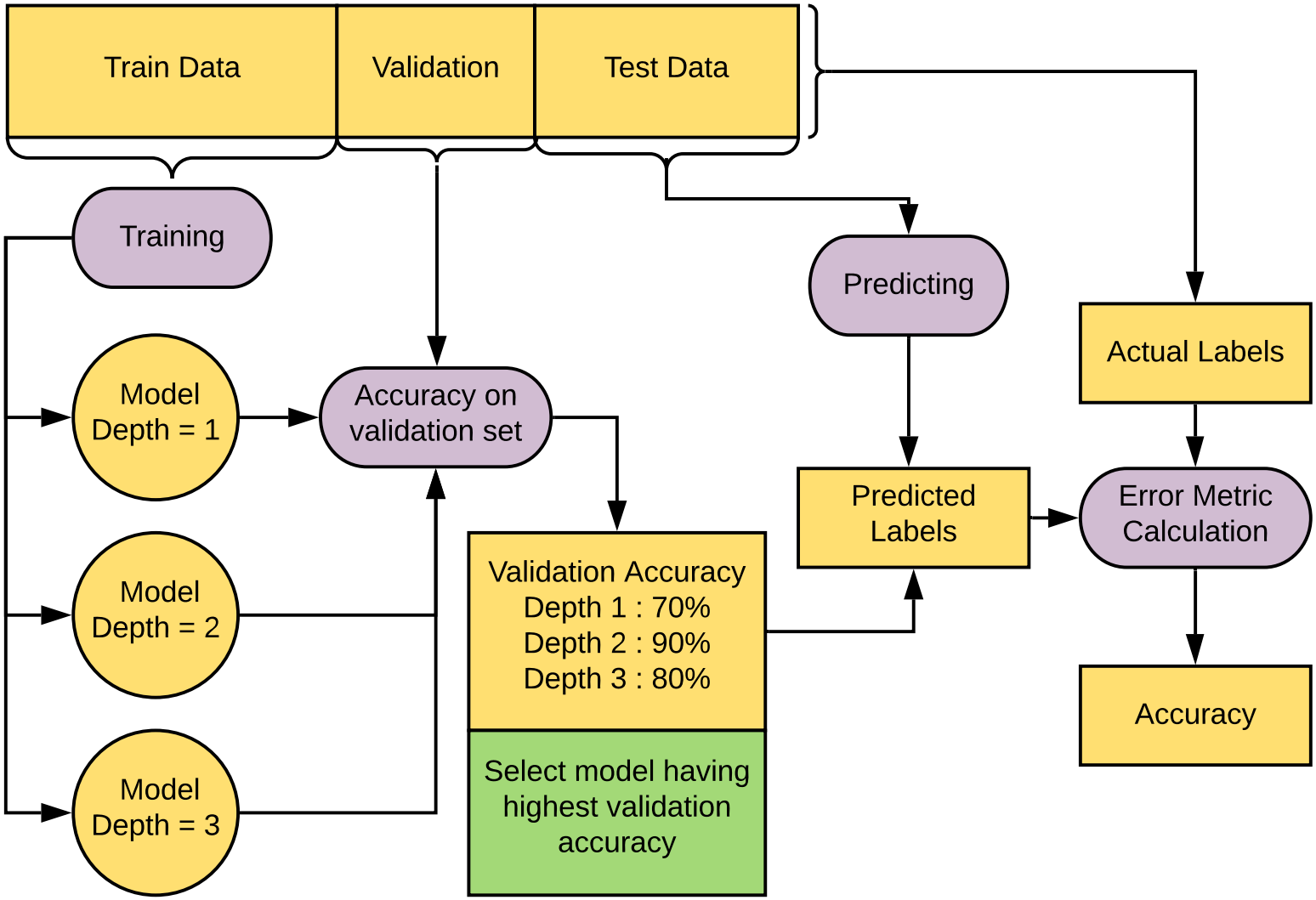


K-Fold cross-validation: Utilise full dataset for testing



Each data point is used for testing exactly once.

Optimizing hyperparameters via the Validation Set

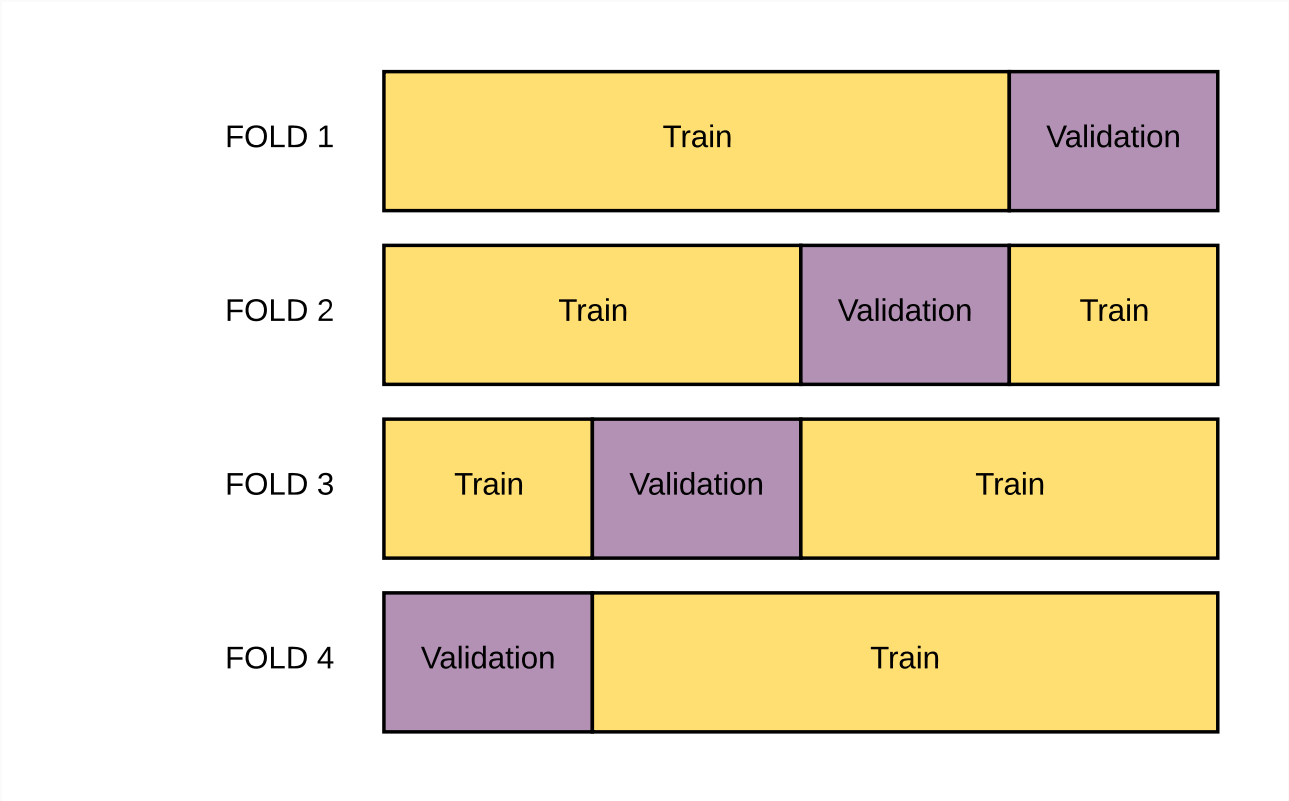


Nested Cross Validation

Divide your training set into K equal parts.

Cyclically use 1 part as "validation set" and the rest for training.

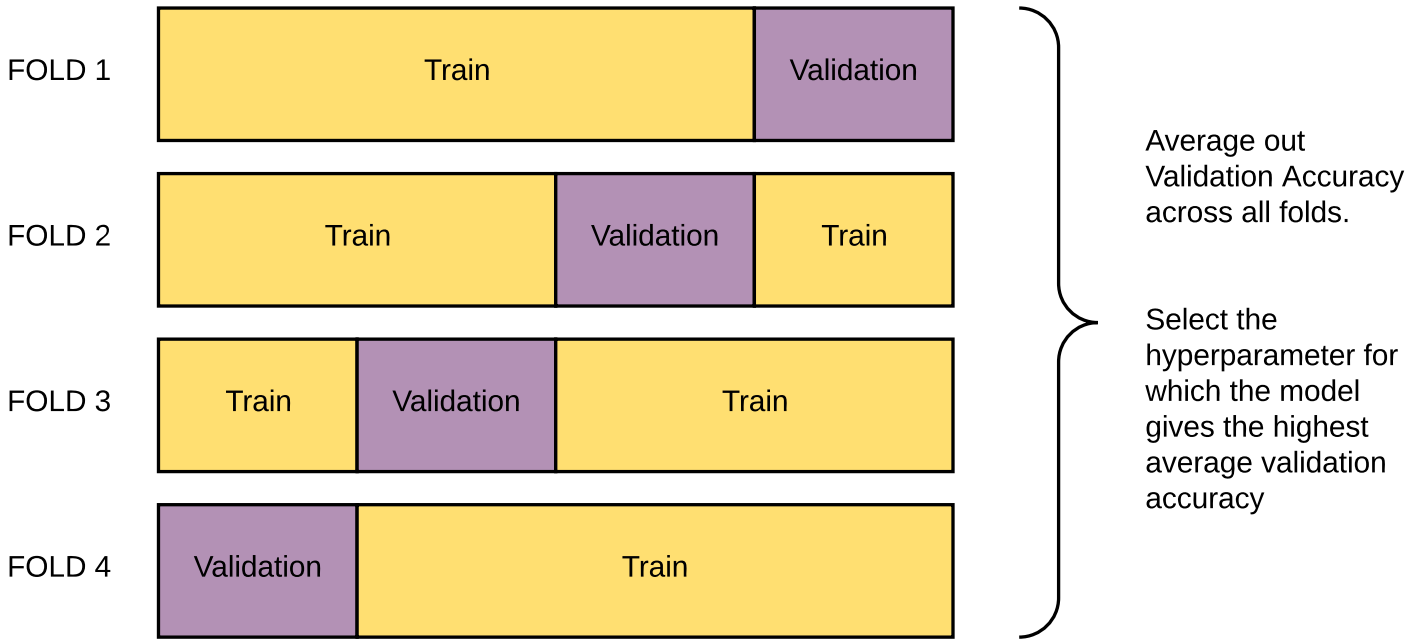
Here $K = 4$



Nested Cross Validation

Average out the validation accuracy across all the folds

Use the model with highest validation accuracy



Next time: Ensemble Learning

- How to combine various models?
- Why to combine multiple models?
- How can we reduce bias?
- How can we reduce variance?